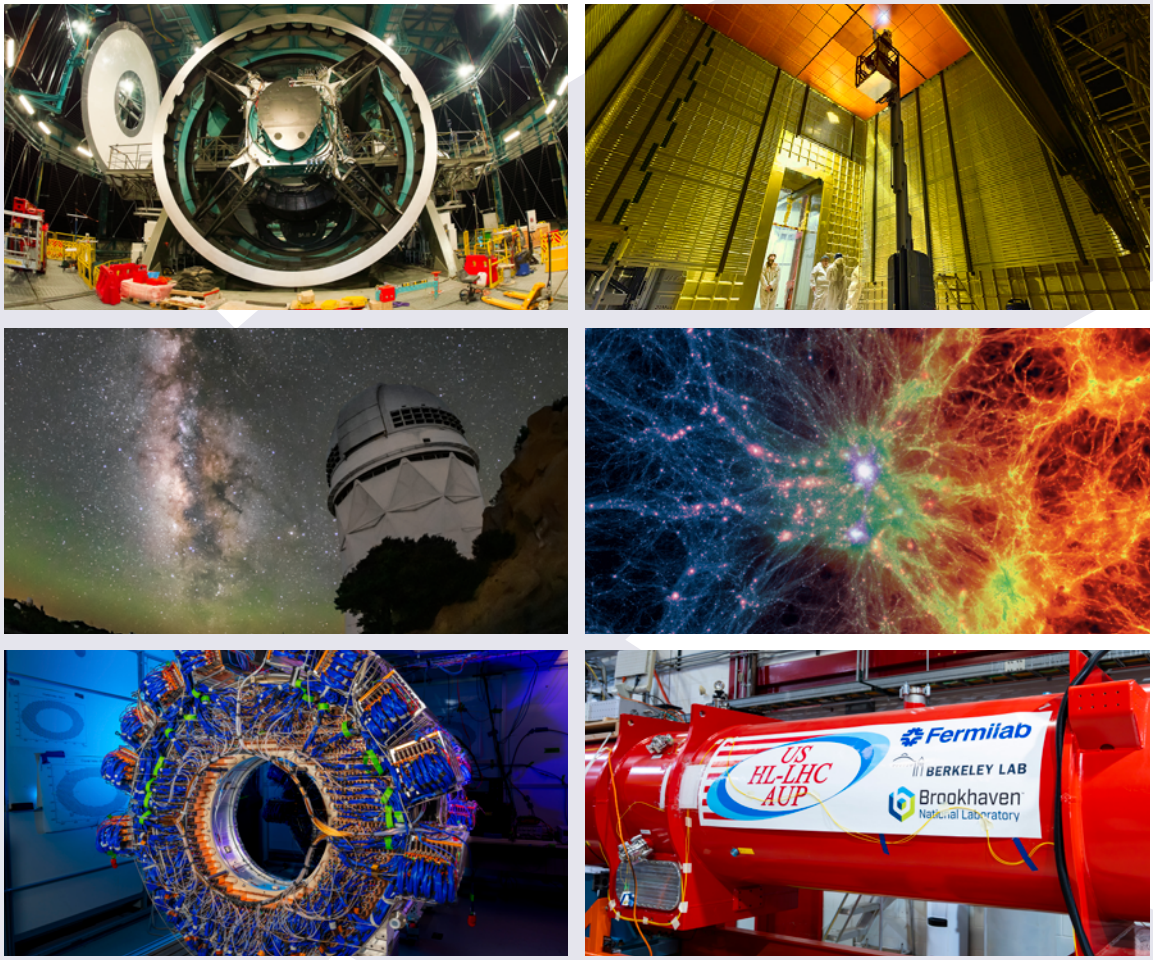
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

ESnet
ENERGY SCIENCES NETWORK

High Energy Physics Network Requirements Review: Final Report

July 2024–December 2024



BERKELEY LAB

U.S. DEPARTMENT of ENERGY
Office of Science

ESnet
ENERGY SCIENCES NETWORK

High Energy Physics Network Requirements Review: Final Update

July 2024–December 2024

Office of High Energy Physics, DOE Office of Science
Energy Sciences Network (ESnet)

ESnet is funded by the US DOE, Office of Science, Office of Advanced Scientific Computing Research. Carol Hawk is the ESnet Program Manager.
ESnet is operated by Lawrence Berkeley National Laboratory (Berkeley Lab), which is operated by the University of California for the US Department of Energy under contract DE-AC02-05CH11231. This work was supported by the Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of High Energy Physics.

This is LBNL report number LBNL-2001696*

Disclaimer
This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Cover Images:
(Top left) The Rubin Observatory's LSST Camera. Courtesy of Rubin Observatory/NSF/AURA, Will O'Mullane
(Top right) DUNE-LBNF excavation at Sanford Lab, South Dakota. Dual-phase ProtoDUNE. Courtesy of Fermi National Accelerator Laboratory/Ryan Postel
(Middle left) DESI observes the sky from the Mayall Telescope, shown here beneath the Milky Way. Courtesy of KPNO/NOIRLab/NSF/AURA/R.T. Sparks
(Middle right) This composite image from the Illustris simulation is centered on the most massive galaxy cluster existing today. It morphs from concentrations of dark matter (at left in blue and purple) to normal matter made mostly of hydrogen and helium gas (at right in red, orange and yellow). Courtesy of Illustris Collaboration.
(Bottom left) The first of two calorimeter disks for the Mu2e experiment. Courtesy of Fermi National Accelerator Laboratory/Ryan Postel
(Bottom right) The AUP team completed testing of the CA01 Cryomodule for use in the LHC. Courtesy of Fermi National Accelerator Laboratory/Dan Svoboda

* https://escholarship.org/uc/item/1dc9j42v

Participants and Contributors

William Allcock, *Argonne National Laboratory (ANL) & Argonne Leadership Computing Facility (ALCF)*

Elias Lopez Asamar, *Universidad Autónoma de Madrid*

Stephen Bailey, *Lawrence Berkeley National Laboratory (LBNL)*

Justas Balcas, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Debbie Bard, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Amanda Bauer, *Vera C. Rubin Observatory*

Keith Bechtol, *University of Wisconsin-Madison*

Matthew Becker, *Argonne National Laboratory (ANL)*

Chris Bee, *Stonybrook University*

Doug Benjamin, *Brookhaven National Laboratory (BNL)*

Wahid Bhimji, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Veronica Bindi, *University of Hawaii at Manoa*

Bob Blum, *Vera C. Rubin Observatory*

Andrey Bobyshev, *Fermi National Accelerator Laboratory (Fermilab)*

Adam Bolton, *SLAC National Accelerator Laboratory (SLAC)*

Vincent Bonafede, *Brookhaven National Laboratory (BNL)*

Julian Borrill, *Lawrence Berkeley National Laboratory (LBNL) & University of California, Berkeley*

Steve Brice, *Fermi National Accelerator Laboratory (Fermilab)*

Paolo Calafiura, *Lawrence Berkeley National Laboratory (LBNL)*

Simone Campana, *The European Organization for Nuclear Research (CERN)*

Shane Canon, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Dale Carder, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

John Carlstrom, *University of Chicago & Argonne National Laboratory (ANL)*

Concetta Cartaro, *SLAC National Accelerator Laboratory (SLAC)*

Giuseppe Cerati, *Fermi National Accelerator Laboratory (Fermilab)*

Evangelos Chaniotakis, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Jim Chiang, *Vera C. Rubin Observatory*

Elisa Chisari, *Utrecht University*

Rafael Coelho Lopes de Sa, *University of Massachusetts Amherst*

John Corlett, *Lawrence Berkeley National Laboratory (LBNL)*

Tom Costello, *Argonne National Laboratory (ANL)*

Travis Croxdale, *Oak Ridge National Laboratory (ORNL)*

Eli Dart, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Cian Dawson, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Kaushik De, *The University of Texas at Arlington*

Phil DeMar, *Fermi National Accelerator Laboratory (Fermilab)*

Jack Deslippe, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Seth Digel, *SLAC National Accelerator Laboratory (SLAC)*

Pengfei Ding, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Daniel Eisenstein, *Harvard University*

Johannes Elmsheuser, *Brookhaven National Laboratory (BNL)*

Brian Etz, *Oak Ridge National Laboratory (ORNL) & Oak Ridge Leadership Computing Facility (OLCF)*

Rafael Ferreira Da Silva, *Oak Ridge National Laboratory (ORNL) & Oak Ridge Leadership Computing Facility (OLCF)*

Mark Foster, *SLAC National Accelerator Laboratory (SLAC)*

Steven Gardiner, *Fermi National Accelerator Laboratory (Fermilab)*

Robert Gardner, *University of Chicago*

Richard Gerber, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Gil Gilchriese, *Lawrence Berkeley National Laboratory (LBNL)*

Salman Habib, *Argonne National Laboratory (ANL)*

Carol Hawk, *DOE Office of Science*

Ryan Heath, *Fermi National Accelerator Laboratory (Fermilab)*

Katrin Heitmann, *Argonne National Laboratory (ANL)*

Ken Herner, *Fermi National Accelerator Laboratory (Fermilab)*

Renée Hložek, *University of Toronto*

Dirk Hufnagel, *Fermi National Accelerator Laboratory (Fermilab)*

Hironori Ito, *Brookhaven National Laboratory (BNL)*

David Jaffe, *Brookhaven National Laboratory (BNL)*

Jeff Kantor, *Vera C. Rubin Observatory*

Gizella Kapus, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Heather Kelly, *SLAC National Accelerator Laboratory (SLAC)*

Liz Kennedy, *Fermi National Accelerator Laboratory (Fermilab)*

Reijo Keskitalo, *Lawrence Berkeley National Laboratory (LBNL) & University of California, Berkeley*

Wes Ketchum, *Fermi National Accelerator Laboratory (Fermilab)*

Michael Kirby, *Brookhaven National Laboratory (BNL)*

Ted Kisner, *Lawrence Berkeley National Laboratory (LBNL)*

Alexei Klimentov, *Brookhaven National Laboratory (BNL)*

Andrei Kounine, *The European Organization for Nuclear Research (CERN)*

Robert Kutschke, *Fermi National Accelerator Laboratory (Fermilab)*

David Lange, *Princeton University*

Tom Lehman, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Giovanna Lehmann, *The European Organization for Nuclear Research (CERN)*

James Letts, *University of California, San Diego*

Jeremy Love, *DOE Office of Science*

Frederick Luehring, *Indiana University*

Mark Lukasczyk, *Brookhaven National Laboratory (BNL)*

Adam Lyon, *Fermi National Accelerator Laboratory (Fermilab)*

Phil Marshall, *SLAC National Accelerator Laboratory (SLAC)*

Edoardo Martelli, *The European Organization for Nuclear Research (CERN)*

Verena Martinez Outschoorn, *University of Massachusetts Amherst*

Shawn McKee, *University of Michigan*

Andrew Melo, *Vanderbilt University*

Inder Monga, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Omar Moreno, *SLAC National Accelerator Laboratory (SLAC)*

Gautham Narayan, *University of Illinois Urbana-Champaign*

Nikolay Nikonov, *University of Hawaii at Manoa*

Will O'Mullane, *Vera C. Rubin Observatory*

Drew Paine, *Lawrence Berkeley National Laboratory (LBNL)*

Ray Pasetes, *Fermi National Accelerator Laboratory (Fermilab)*

Abid Patwa, *DOE Office of Science*

Daniel Pelfrey, *Oak Ridge National Laboratory (ORNL) & Oak Ridge Leadership Computing Facility (OLCF)*

Ryan Prout, *Oak Ridge National Laboratory (ORNL) & Oak Ridge Leadership Computing Facility (OLCF)*

Srini Rajagopalan, *Brookhaven National Laboratory (BNL)*

Katherine Riley, *Argonne National Laboratory (ANL) & Argonne Leadership Computing Facility (ALCF)*

Kate Robinson, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

David M. Rogers, *Oak Ridge National Laboratory (ORNL) & Oak Ridge Leadership Computing Facility (OLCF)*

Heidi Schellman, *Oregon State University*

Kate Scholberg, *Duke University*

Uros Seljak, *Lawrence Berkeley National Laboratory (LBNL) & University of California, Berkeley*

Cedric Serfon, *Brookhaven National Laboratory (BNL)*

Ignacio Sevilla, *Centre for Energy, Environmental and Technological Research (CIEMAT)*

Syed Asif Shah, *Fermi National Accelerator Laboratory (Fermilab)*

Baosong Shan, *Beihang University*

Brandon Siegel, *Argonne National Laboratory (ANL)*

Cristián Silva, *Vera C. Rubin Observatory*

Richard Simon, *Lawrence Berkeley National Laboratory (LBNL)*

Nick Smith, *Fermi National Accelerator Laboratory (Fermilab)*

Eric Smith , *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Cory Snavely, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Tavia Stone Gibbins, *Lawrence Berkeley National Laboratory (LBNL) & National Energy Research Scientific Computing Center (NERSC)*

Rune Stromsness, *Lawrence Berkeley National Laboratory (LBNL)*

Alessandro Thea, *The European Organization for Nuclear Research (CERN)*

Jordan Thomas, *DOE Office of Science*

Jiachuan Tian, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Steve Timm, *Fermi National Accelerator Laboratory (Fermilab)*

David Toback, *Texas A&M University*

Chris Tracy, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Tom Uram, *Argonne National Laboratory (ANL) & Argonne Leadership Computing Facility (ALCF)*

Venkat Vishwanath, *Argonne National Laboratory (ANL) & Argonne Leadership Computing Facility (ALCF)*

Paul Wefel, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Andrew Wiedlea, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Linda Winkler, *Argonne National Laboratory (ANL)*

Wei Yang, *SLAC National Accelerator Laboratory (SLAC)*

Xi Yang, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet)*

Jim Yeck, *Brookhaven National Laboratory (BNL)*

Stefan Zatschler, *SNOLAB*

Andrea Zonca, *San Diego Supercomputer Center (SDSC) and the University of California San Diego*

Jason Zurawski, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet*

Report Editors

Jason Zurawski, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* zurawski@es.net

Dale Carder, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* dwcarder@es.net

Evangelos Chaniotakis, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* haniotak@es.net

Cian Dawson, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* cbdawson@es.net

Eli Dart, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* dart@es.net

Carol Hawk, *DOE Office of Science:* carol.hawk@science.doe.gov

Jeremy Love, *DOE Office of Science:* jeremy.love@science.doe.gov

Drew Paine, *Lawrence Berkeley National Laboratory (LBNL) & High Performance Data Facility (HPDF):* pained@lbl.gov

Abid Patwa, *DOE Office of Science:* abid.patwa@science.doe.gov

Kate Robinson, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* katerobinson@es.net

Jiachuan Tian, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* jiachuantian@es.net

Chris Tracy, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* ctracy@es.net

Andrew Wiedlea, *Lawrence Berkeley National Laboratory (LBNL) & Energy Sciences Network (ESnet):* awiedlea@es.net

Table of Contents

Disclaimer II

Participants and Contributors III

Report Editors VI

1 Executive Summary 1

1.1 Summary of Review Findings 3
1.1.1 Facility Management and Readiness 3
1.1.2 Scientific Data Management 3
1.1.3 Scientific Workflow 3
1.1.4 Computational and Storage Requirements 4
1.1.5 Remote Collaboration and Operational Requirements 4
1.1.6 Multifacility Computational Workflows 4
1.1.7 Networking for Local and Wide-Area Data Mobility 5
1.1.8 Emerging Needs 5
1.1.9 IRI Responsiveness 6
1.2 Summary of Review Actions 6
1.2.1 Facility Management and Readiness 6
1.2.2 Scientific Data Management 6
1.2.3 Scientific Workflow 7
1.2.4 Computational and Storage Requirements 7
1.2.5 Remote Collaboration and Operational Requirements 7
1.2.6 Multifacility Computational Workflows 7
1.2.7 Networking for Local and Wide-Area Data Mobility 8
1.2.8 Emerging Needs 8
1.2.9 IRI Responsiveness 8

2 Requirements Review Overview 9

2.1 Purpose and Process 9
2.2 Structure 9
2.2.1 Background 10
2.2.2 Case Study Methodology 10
2.3 ESnet 12
2.4 About ASCR 12
2.5 About the Office of HEP 12

3 Assessment of 2020 HEP Requirement Review 14

3.1 Prior Action Items 14
3.2 Assessment 14
3.2.1 ESnet Actions 14
3.2.2 DOE SC Actions 15

4 Review Findings **17**

4.1 Facility Management and Readiness 17
4.2 Scientific Data Management 21
4.3 Scientific Workflow 26
4.4 Computational and Storage Requirements 32
4.5 Remote Collaboration and Operational Requirements 36
4.6 Multifacility Computational Workflows 38
4.7 Networking for Local and Wide-Area Data Mobility 40
4.8 Emerging Needs 44
4.9 IRI Responsiveness 49

5 Review Actions **53**

5.1 Facility Management and Readiness 53
5.2 Scientific Data Management 54
5.3 Scientific Workflow 54
5.4 Computational and Storage Requirements 55
5.5 Remote Collaboration and Operational Requirements 56
5.6 Multifacility Computational Workflows 57
5.7 Networking for Local and Wide-Area Data Mobility 57
5.8 Emerging Needs 62
5.9 IRI Responsiveness 64

6 HEP Case Studies **66**

6.1 Cosmological Simulation Research **66**
6.1.1 Discussion Summary 67
6.1.2 Cosmological Simulation Research: Experiment Profile 70
6.1.2.1 Science Background 70
6.1.2.2 Collaborators 71
6.1.2.3 IRI Readiness 72
6.1.2.4 Instruments and Facilities 72
6.1.2.5 Generalized Process of Science 73
6.1.2.6 Remote Science Activities 73
6.1.2.7 Software Infrastructure 74
6.1.2.8 Network and Data Architecture 74
6.1.2.9 Emerging Technology Use 74
6.1.2.9.1 Cloud Services 74
6.1.2.9.2 Artificial Intelligence and ML 75
6.1.2.9.3 Wireless or Distributed Location Science Activities 75
6.1.2.10 Data-Related Resource Constraints 75
6.1.2.11 Data Mobility Endpoints 75
6.1.2.12 Outstanding Issues 75
6.1.2.13 Facility Profile Contributors 76

6.2 DESI **77**
6.2.1 Discussion Summary 77
6.2.2 DESI: Experiment Profile 80
6.2.2.1 Science Background 80
6.2.2.2 Collaborators 81
6.2.2.3 IRI Readiness 82
6.2.2.4 Instruments and Facilities 82
6.2.2.5 Generalized Process of Science 83
6.2.2.6 Remote Science Activities 83
6.2.2.7 Software Infrastructure 83
6.2.2.8 Network and Data Architecture 84
6.2.2.9 Emerging Technology Use 84
6.2.2.9.1 Cloud Services 84
6.2.2.9.2 AI and ML 85
6.2.2.9.3 Wireless or Distributed Location Science Activities 85
6.2.2.10 Data-Related Resource Constraints 85
6.2.2.11 Data Mobility Endpoints 85
6.2.2.12 Outstanding Issues 85
6.2.2.13 Facility Profile Contributors 86
6.3 Vera C. Rubin Observatory **87**
6.3.1 Discussion Summary 87
6.3.2 Vera C. Rubin Observatory: Facility Profile 91
6.3.2.1 Science Background 91
6.3.2.2 Collaborators 92
6.3.2.2.1 North-South Networking 93
6.3.2.2.2 National and International Networking 94
6.3.2.3 IRI Readiness 94
6.3.2.4 Instruments and Facilities 95
6.3.2.5 Generalized Process of Science 96
6.3.2.6 Remote Science Activities 96
6.3.2.7 Software Infrastructure 97
6.3.2.7.1.1 Data and Software Products 97
6.3.2.7.1.1.1 Prompt Products Production 97
6.3.2.7.1.1.2 Data-Release Production 99
6.3.2.7.1.1.3 Calibration and Ancillary Data 101
6.3.2.7.1.1.4 Education and Public Outreach (EPO) Data Products 101
6.3.2.7.1.1.5 Software and Documentation 101
6.3.2.7.1.2 Data Rights 102
6.3.2.7.1.3 Science Platform 102
6.3.2.7.1.3.1 Science Platform Resource Allocation 103
6.3.2.7.1.4 Interfaces Between Rubin Observatory and the Community 104
6.3.2.7.1.5 Beyond Rubin Operations: Community-Based Data Analysis 106
6.3.2.8 Network and Data Architecture 109
6.3.2.8.1 Data Management System 109
6.3.2.8.1.1 Rubin Observatory DMS Architecture 109

6.3.2.8.1.2 Compute and Storage Sizing 110
6.3.2.8.1.2.1 Storage Requirements 111
6.3.2.8.1.2.2 Compute Requirements 113
6.3.2.8.1.3 Chilean Data Center 114
6.3.2.8.1.3.1 Prompt Base 114
6.3.2.8.1.3.2 Archive Base 115
6.3.2.8.1.3.3 Commissioning Cluster 115
6.3.2.8.1.3.4 Chilean DAC 115
6.3.2.8.1.4 US Data Facility 116
6.3.2.8.1.4.1 Prompt USDF 116
6.3.2.8.1.4.2 Archive USDF 117
6.3.2.8.1.4.3 Offline Production USDF 117
6.3.2.8.1.4.4 US Data Access Center 118
6.3.2.8.1.5 Non-US Data Facilities 119
6.3.2.8.1.5.1 IN2P3 Data Facility 119
6.3.2.8.1.5.2 UKDF 120
6.3.2.8.1.5.3 Offline Processing Satellite 120
6.3.2.8.1.6 Backbone Services 120
6.3.2.8.1.7 Wide Area Networks 120
6.3.2.8.1.7.1 Network Requirements 120
6.3.2.8.1.7.2 Network Architecture and Design 122
6.3.2.8.1.8 Configuration Control, Test, and Release Procedures 125
6.3.2.9 Emerging Technology Use 126
6.3.2.9.1 Cloud Services 126
6.3.2.9.2 AI and ML 126
6.3.2.9.3 Wireless or Distributed Location Science Activities 126
6.3.2.10 Data-Related Resource Constraints 127
6.3.2.11 Data Mobility Endpoints 127
6.3.2.12 Outstanding Issues 127
6.3.2.13 Facility Profile Contributors 127
6.4 DESC **128**
6.4.1 Discussion Summary 128
6.4.2 DESC: Experiment Profile 132
6.4.2.1 Science Background 132
6.4.2.2 Collaborators 132
6.4.2.3 IRI Readiness 133
6.4.2.4 Instruments and Facilities 134
6.4.2.5 Generalized Process of Science 135
6.4.2.6 Remote Science Activities 136
6.4.2.7 Software Infrastructure 137
6.4.2.8 Network and Data Architecture 137
6.4.2.9 Emerging Technology Use 137
6.4.2.9.1 Cloud Services 137
6.4.2.9.2 Artificial Intelligence and ML 137
6.4.2.9.3 Wireless or Distributed Location Science Activities 137

6.4.2.10 Data-Related Resource Constraints 137
6.4.2.11 Data Mobility Endpoints 138
6.4.2.12 Outstanding Issues 138
6.4.2.13 Facility Profile Contributors 139
6.5 CMB 140
6.5.1 Discussion Summary 140
6.5.2 CMB: Experiment Profile 143
6.5.2.1 Science Background 143
6.5.2.2 Collaborators 144
6.5.2.3 IRI Readiness 145
6.5.2.4 Instruments and Facilities 145
6.5.2.4.1 Observatory in the Chilean Atacama 145
6.5.2.4.2 Computing Resources 146
6.5.2.5 Generalized Process of Science 147
6.5.2.6 Remote Science Activities 147
6.5.2.7 Software Infrastructure 147
6.5.2.7.1 Data Movement 147
6.5.2.7.2 Data Simulation and Reduction 148
6.5.2.8 Network and Data Architecture 148
6.5.2.9 Emerging Technology Use 149
6.5.2.9.1 Cloud Services 149
6.5.2.9.2 AI and ML 149
6.5.2.9.3 Wireless or Distributed Location Science Activities 150
6.5.2.10 Data-Related Resource Constraints 150
6.5.2.11 Data Mobility Endpoints 150
6.5.2.12 Outstanding Issues 150
6.5.2.13 Facility Profile Contributors 151
6.6 Muon Experimentation at Fermilab: Muon g-2 and Mu2e 152
6.6.1 Discussion Summary 152
6.6.2 Muon Experimentation at Fermilab: Experiment Profile 154
6.6.2.1 Science Background 154
6.6.2.1.1 Muon g-2 Science Background 155
6.6.2.1.2 Mu2e Science Background 156
6.6.2.2 Collaborators 157
6.6.2.2.1 Muon g-2 Collaborators 157
6.6.2.2.2 Mu2e Collaborators 157
6.6.2.3 IRI Readiness 159
6.6.2.3.1 Muon g-2 IRI Readiness 159
6.6.2.3.2 Mu2e IRI Readiness 159
6.6.2.4 Instruments and Facilities 159
6.6.2.4.1 Muon g-2 Instruments and Facilities 159
6.6.2.4.2 Mu2e Instruments and Facilities 159
6.6.2.5 Generalized Process of Science 161
6.6.2.5.1 Muon g-2 Process of Science 161
6.6.2.5.2 Mu2e Process of Science 161

6.6.2.6 Remote Science Activities 163
6.6.2.6.1 Muon g-2 Remote Science Activities 163
6.6.2.6.2 Mu2e Remote Science Activities 163
6.6.2.7 Software Infrastructure 163
6.6.2.8 Network and Data Architecture 164
6.6.2.9 Emerging Technology Use 164
6.6.2.9.1 Cloud Services 164
6.6.2.9.2 AI and ML 164
6.6.2.9.3 Wireless or Distributed Location Science Activities 164
6.6.2.10 Data-Related Resource Constraints 164
6.6.2.11 Data Mobility Endpoints 164
6.6.2.12 Outstanding Issues 164
6.6.2.13 Facility Profile Contributors 165
6.7 Belle II Experiment 166
6.7.1 Discussion Summary 166
6.7.2 Belle II: Experiment Profile 168
6.7.2.1 Science Background 168
6.7.2.2 Collaborators 168
6.7.2.3 IRI Readiness 170
6.7.2.4 Instruments and Facilities 170
6.7.2.5 Generalized Process of Science 171
6.7.2.6 Remote Science Activities 171
6.7.2.7 Software Infrastructure 173
6.7.2.7.1. Rucio 173
6.7.2.7.2. DIRAC 174
6.7.2.7.3. Conditions Database 174
6.7.2.7.4. CVMFS 174
6.7.2.8 Network and Data Architecture 174
6.7.2.8.1 Domestic Connectivity 174
6.7.2.8.2 International Connectivity 178
6.7.2.9 Emerging Technology Use 179
6.7.2.9.1 Cloud Services 179
6.7.2.9.2 AI and ML 179
6.7.2.9.3 Wireless or Distributed Location Science Activities 179
6.7.2.10 Data-Related Resource Constraints 179
6.7.2.11 Data Mobility Endpoints 179
6.7.2.12 Outstanding Issues 179
6.7.2.13 Facility Profile Contributors 179
6.8 Neutrino Experiments at Fermilab: SBN and DUNE 180
6.8.1 Discussion Summary 180
6.8.2 Neutrino Experiments at Fermilab: Experiment Profile 185
6.8.2.1 Science Background 185
6.8.2.1.1 DUNE Science Background 185
6.8.2.1.2 SBN Science Background 186

6.8.2.2 Collaborators 187
6.8.2.2.1 DUNE Collaborators 187
6.8.2.2.2 SBN Collaborators 188
6.8.2.3 IRI Readiness 189
6.8.2.3.1 DUNE IRI Readiness 189
6.8.2.3.2 SBN IRI Readiness 189
6.8.2.4 Instruments and Facilities 190
6.8.2.4.1 DUNE Instruments and Facilities 190
6.8.2.4.2 SBN Instruments and Facilities 191
6.8.2.5 Generalized Process of Science 192
6.8.2.5.1 DUNE Process of Science 192
6.8.2.5.2 SBN Process of Science 193
6.8.2.6 Remote Science Activities 195
6.8.2.7 Software Infrastructure 195
6.8.2.7.1 DUNE Software Infrastructure 195
6.8.2.7.2 SBN Software Infrastructure 196
6.8.2.8 Network and Data Architecture 196
6.8.2.8.1 DUNE Network and Data Architecture 196
6.8.2.8.2 SBN Network and Data Architecture 196
6.8.2.9 Emerging Technology Use 196
6.8.2.9.1 Cloud Services 196
6.8.2.9.2 Artificial Intelligence and ML 196
6.8.2.9.2.1 DUNE AI & ML 196
6.8.2.9.2.2 SBN AI & ML 196
6.8.2.9.3 Wireless or Distributed Location Science Activities 197
6.8.2.10 Data-Related Resource Constraints 197
6.8.2.10.1 DUNE Data-Related Resource Constraints 197
6.8.2.10.2 SBN Data-Related Resource Constraints 197
6.8.2.11 Data Mobility Endpoints 197
6.8.2.11.1 DUNE Data Mobility Endpoints 197
6.8.2.11.2 SBN Data Mobility Endpoints 198
6.8.2.12 Outstanding Issues 198
6.8.2.13 Facility Profile Contributors 198
6.9 LHC: ATLAS Experiment 199
6.9.1 Discussion Summary 199
6.9.2 ATLAS: Experiment Profile 202
6.9.2.1 Science Background 203
6.9.2.2 Collaborators 203
6.9.2.3 IRI Readiness 214
6.9.2.4 Instruments and Facilities 214
6.9.2.4.1 LHC 214
6.9.2.4.2 ATLAS 215
6.9.2.4.3 WLCG 216
6.9.2.4.4 US ATLAS T1 at BNL 216
6.9.2.4.5 US ATLAS T2 Infrastructure 218

6.9.2.4.6 US ATLAS Analysis Facilities 219
6.9.2.5 Generalized Process of Science 219
6.9.2.5.1 Computation 219
6.9.2.5.1.1 High-Performance Computing 223
6.9.2.5.1.2 Distributed Computing 223
6.9.2.5.2 Storage 223
6.9.2.5.3 Network Use Cases & Data Flow 226
6.9.2.6 Remote Science Activities 231
6.9.2.7 Software Infrastructure 231
6.9.2.7.1 Rucio 231
6.9.2.7.2 PanDA 231
6.9.2.7.3 Frontier 232
6.9.2.7.4 CVMFS 232
6.9.2.7.5 HTCondor 232
6.9.2.7.6 ROOT 232
6.9.2.8 Network and Data Architecture 232
6.9.2.8.1 US ATLAS T1 at BNL 232
6.9.2.8.2 US ATLAS T2 Infrastructure 236
6.9.2.9 Emerging Technology Use 241
6.9.2.9.1 Cloud Services 241
6.9.2.9.2 Artificial Intelligence and ML 241
6.9.2.9.3 Wireless or Distributed Location Science Activities 241
6.9.2.10 Data-Related Resource Constraints 241
6.9.2.11 Data Mobility Endpoints 242
6.9.2.12 Outstanding Issues 242
6.9.2.13 Facility Profile Contributors 242
6.10 LHC: CMS Experiment 243
6.10.1 Discussion Summary 243
6.10.2 CMS: Experiment Profile 248
6.10.2.1 Science Background 248
6.10.2.2 Collaborators 249
6.10.2.3 IRI Readiness 252
6.10.2.4 Instruments and Facilities 252
6.10.2.4.1 HLT 253
6.10.2.4.2 Tier-0 253
6.10.2.4.3 Tier-1 253
6.10.2.4.4 Tier-2 253
6.10.2.4.5 CMS Computing Capacity and Planning 253
6.10.2.4.6 Tier-3 254
6.10.2.4.7 HPC facilities 254
6.10.2.5 Generalized Process of Science 256
6.10.2.5.1 Collision Data 256
6.10.2.5.2 Simulation Data 258
6.10.2.5.3 User Data 259

6.10.2.5.4 Event Reconstruction and Analysis Data Creation 260
6.10.2.5.5 Analysis Data Formats 261
6.10.2.5.6 Data Access 263
6.10.2.5.7 Analysis Group and Researcher Workflows 264
6.10.2.6 Remote Science Activities 264
6.10.2.7 Software Infrastructure 265
6.10.2.8 Network and Data Architecture 266
6.10.2.8.1 The US CMS Tier-1 266
6.10.2.8.2 The US CMS Tier-2s 267
6.10.2.9 Emerging Technology Use 269
6.10.2.9.1 Cloud Services 269
6.10.2.9.2 AI and ML 269
6.10.2.9.3 Wireless or Distributed Location Science Activities 270
6.10.2.10 Data-Related Resource Constraints 270
6.10.2.11 Data Mobility Endpoints 270
6.10.2.12 Outstanding Issues 270
6.10.2.13 Facility Profile Contributors 271
6.11 LHC: LHC Operations 272
6.11.1 Discussion Summary 272
6.11.2 LHC Operations: Profile 276
6.11.2.1 Science Background 276
6.11.2.2 Collaborators 277
6.11.2.3 IRI Readiness 277
6.11.2.4 Instruments and Facilities 277
6.11.2.4.1 LHC 277
6.11.2.4.2 Globally Distributed Computing Sites 278
6.11.2.5 Generalized Process of Science 278
6.11.2.6 Remote Science Activities 278
6.11.2.7 Software Infrastructure 279
6.11.2.7.1 Data Management Tools 279
6.11.2.7.2 Monitoring Tools 281
6.11.2.7.3 Network Management Tools 282
6.11.2.7.3.1 SENSE/Rucio Architecture Overview 282
6.11.2.7.3.2 Key Components and Functionalities 283
6.11.2.7.3.3 Testbeds and Site Upgrades 283
6.11.2.7.3.4 Benefits Targeted 283
6.11.2.7.3.5 Future Plans and Expansion 283
6.11.2.8 Network and Data Architecture 283
6.11.2.9 Emerging Technology Use 285
6.11.2.9.1 Cloud Services 285
6.11.2.9.2 AI and ML 285
6.11.2.9.3 Wireless or Distributed Location Science Activities 285
6.11.2.9.4 Data-Centre Interconnect and Spectrum Sharing 285
6.11.2.10 Data-Related Resource Constraints 286

6.11.2.11 Data Mobility Endpoints 286
6.11.2.12 Outstanding Issues 286
6.11.2.13 Facility Profile Contributors 287
6.12 LHC: HL-LHC & Network Research 288
6.12.1 Discussion Summary 288
6.12.2 HL-LHC: Network Research Profile 293
6.12.2.1 Science Background 293
6.12.2.2 Collaborators 294
6.12.2.3 IRI Readiness 295
6.12.2.4 Instruments and Facilities 296
6.12.2.4.1 ATLAS Computing Infrastructure 296
6.12.2.4.2 CMS Computing Infrastructure 298
6.12.2.5 Generalized Process of Science 299
6.12.2.5.1 Analysis Facilities 300
6.12.2.5.2 Inference As a Service 301
6.12.2.6 Remote Science Activities 302
6.12.2.7 Software Infrastructure 302
6.12.2.7.1 ATLAS 302
6.12.2.7.2 CMS 302
6.12.2.8 Network and Data Architecture 303
6.12.2.9 Emerging Technology Use 304
6.12.2.9.1 Cloud Services 304
6.12.2.9.2 AI and ML 304
6.12.2.9.3 Wireless or Distributed Location Science Activities 304
6.12.2.10 Data-Related Resource Constraints 305
6.12.2.11 Data Mobility Endpoints 305
6.12.2.12 Outstanding Issues 306
6.12.2.13 Facility Profile Contributors 306
6.13 Super Cryogenic Dark Matter Search (SuperCDMS SNOLAB) 307
6.13.1 Discussion Summary 307
6.13.2 SuperCDMS: Experiment Profile 309
6.13.2.1 Science Background 309
6.13.2.2 Collaborators 310
6.13.2.3 IRI Readiness 311
6.13.2.4 Instruments and Facilities 311
6.13.2.5 Generalized Process of Science 312
6.13.2.6 Remote Science Activities 313
6.13.2.7 Software Infrastructure 313
6.13.2.8 Network and Data Architecture 313
6.13.2.9 Emerging Technology Use 313
6.13.2.9.1 Cloud Services 313
6.13.2.9.2 AI and ML 314
6.13.2.9.3 Wireless or Distributed Location Science Activities 314
6.13.2.10 Data-Related Resource Constraints 314

6.13.2.11 Data Mobility Endpoints 314
6.13.2.12 Outstanding Issues 314
6.13.2.13 Facility Profile Contributors 315
6.14 Alpha Magnetic Spectrometer (AMS) Experiment 316
6.14.1 Discussion Summary 316
6.14.2 AMS: Experiment Profile 318
6.14.2.1 Science Background 318
6.14.2.2 Collaborators 318
6.14.2.3 IRI Readiness 321
6.14.2.4 Instruments and Facilities 322
6.14.2.5 Generalized Process of Science 323
6.14.2.6 Remote Science Activities 324
6.14.2.7 Software Infrastructure 325
6.14.2.7.1 Data Resource Management 325
6.14.2.7.2 Data mobility 325
6.14.2.7.3 Data Process 325
6.14.2.8 Network and Data Architecture 325
6.14.2.8.1 The LAN Capabilities 325
6.14.2.8.2 The Local Network Configuration 325
6.14.2.8.3 Network Bandwidth Connection Speed 325
6.14.2.9 Emerging Technology Use 325
6.14.2.9.1 Cloud Services 325
6.14.2.9.2 AI and ML 326
6.14.2.9.3 Wireless or Distributed Location Science Activities 326
6.14.2.10 Data-Related Resource Constraints 326
6.14.2.11 Data Mobility Endpoints 326
6.14.2.12 Outstanding Issues 326
6.14.2.13 Facility Profile Contributors 326

7 Technology Profiles 327

7.1 NERSC 327
7.1.1 Discussion Summary 327
7.1.2 NERSC: Technology Profile 328
7.1.2.1 Facilities Operated and Supported 328
7.1.2.1.1 Computing 328
7.1.2.1.2 Storage 328
7.1.2.1.3 Networking 328
7.1.2.1.4 Planned Computing 329
7.1.2.1.5 HEP Projects Supported by NERSC 330
7.1.2.1.5.1 Top 10 HEP Compute Projects 330
7.1.2.2 Network and Data Architecture 332
7.1.2.2.1 Current NERSC Network 332
7.1.2.2.2 Next (Two to Five Year) Network Design 333
7.1.2.3 Computational and Storage Infrastructure 333

7.1.2.4 Cloud Services 333
7.1.2.5 AI and ML 334
7.1.2.6 Wireless Capabilities 334
7.1.2.7 Data-Related Resource Constraints 334
7.1.2.8 Data Mobility Endpoints 334
7.1.2.8.1 Configuration 334
7.1.2.8.2 Access 334
7.1.2.8.3 Available File Systems 335
7.1.2.8.4 File Transfer Software 335
7.1.2.8.5 Restrictions 335
7.1.2.9 Outstanding Issues 335
7.1.2.10 Technology Profile Contributors 335
7.2 ALCF 336
7.2.1 Discussion Summary 336
7.2.2 ALCF: Technology Profile 336
7.2.2.1 Facilities Operated and Supported 336
7.2.2.2 Network and Data Architecture 338
7.2.2.3 Computational and Storage Infrastructure 339
7.2.2.4 Cloud Services 339
7.2.2.5 AI and ML 340
7.2.2.6 Wireless Capabilities 340
7.2.2.7 Data-Related Resource Constraints 341
7.2.2.8 Data Mobility Endpoints 341
7.2.2.9 Outstanding Issues 341
7.2.2.10 Technology Profile Contributors 341
7.3 OLCF 342
7.3.1 Discussion Summary 342
7.3.2 OLCF: Technology Profile 343
7.3.2.1 Facilities Operated and Supported 343
7.3.2.2 Network and Data Architecture 344
7.3.2.3 Computational and Storage Infrastructure 344
7.3.2.4 Cloud Services 346
7.3.2.5 AI and ML 346
7.3.2.6 Wireless Capabilities 346
7.3.2.7 Data-Related Resource Constraints 347
7.3.2.8 Data Mobility Endpoints 347
7.3.2.8.1 Interactive Access 347
7.3.2.8.2 Access From Globus Online 347
7.3.2.9 Outstanding Issues 347
7.3.2.10 Technology Profile Contributors 347
7.4 Fermilab 348
7.4.1 Discussion Summary 348
7.4.2 Fermilab: Technology Profile 349
7.4.2.1 Facilities Operated and Supported 349

7.4.2.2 Network and Data Architecture 351
7.4.2.2.1 Network Perimeter 352
7.4.2.2.2 General Datacenter for Multiple Experiments 352
7.4.2.2.3 The US CMS-Tier-1 Network 353
7.4.2.2.3.1 US CMS-Tier-2 Locations 354
7.4.2.2.4 DUNE Network Architecture 354
7.4.2.2.4.1 DUNE WAN 356
7.4.2.2.4.2 SURF -> Fermilab Data Movement 356
7.4.2.2.4.3 Fermilab Distribution of DUNE Data To Collaboration Sites 357
7.4.2.2.5 NOvA (NuMI Off-axis **ν**e Appearance) 358
7.4.2.3 Computational and Storage Infrastructure 358
7.4.2.3.1 Storage 358
7.4.2.3.2 Compute 359
7.4.2.3.3 Future Storage/Compute Resource Upgrades/Expansion 359
7.4.2.4 Cloud Services 359
7.4.2.5 Artificial Intelligence and ML 359
7.4.2.6 Wireless Capabilities 360
7.4.2.7 Data-Related Resource Constraints 360
7.4.2.8 Data Mobility Endpoints 360
7.4.2.9 Outstanding Issues 360
7.4.2.10 Technology Profile Contributors 360
7.5 BNL **361**
7.5.1 Discussion Summary 361
7.5.2 BNL: Technology Profile 362
7.5.2.1 Facilities Operated and Supported 362
7.5.2.2 Network and Data Architecture 362
7.5.2.2.1 BNL Network Architectures 362
7.5.2.2.2 BNL Network Perimeter 363
7.5.2.2.2.1 BNL WAN and Internet Circuits 364
7.5.2.2.2.2 Connectivity to Science DMZ(s) and Campus Network 365
7.5.2.2.2.3 BNL's Support for LHCOPN and LHCONE Services 365
7.5.2.2.2.4 Network Perimeter Security Features 366
7.5.2.2.3 BNL Campus Network 366
7.5.2.2.3.1 BNL Campus Architecture 367
7.5.2.2.4 HTSN 368
7.5.2.2.5 NSLS2 Network Infrastructure 370
7.5.2.2.5.1 NSLS2 Science Network (N2SN) Components 370
7.5.2.2.5.2 NSLS2 Accelerator/Controls Network Components 371
7.5.2.2.5.3 N2SN Core 371
7.5.2.2.5.4 N2SN Beamline Network 372
7.5.2.2.5.5 NSLS2 Data Services Block 372
7.5.2.2.5.6 NSLS2 DMZ 373
7.5.2.2.5.7 NSLS2 Accelerator/Controls Infrastructure 373
7.5.2.2.5.8 NSLS2 Network Management 373
7.5.2.3 Computational and Storage Infrastructure 374

7.5.2.4 Cloud Services 374
7.5.2.5 Artificial Intelligence and ML 374
7.5.2.6 Wireless Capabilities 374
7.5.2.6.1 Enterprise Wireless 375
7.5.2.6.2 LoRaWAN Remote Monitoring 375
7.5.2.6.3 Site Cellular Network Connectivity 376
7.5.2.6.4 NSLS2 Wireless Network 376
7.5.2.7 Data-Related Resource Constraints 377
7.5.2.8 Data Mobility Endpoints 377
7.5.2.9 Outstanding Issues 377
7.5.2.10 Technology Profile Contributors 377
7.6 SLAC National Accelerator Laboratory 378
7.6.1 Discussion Summary 378
7.6.2 SLAC: Technology Profile 378
7.6.2.1 Facilities Operated and Supported 378
7.6.2.2 Network and Data Architecture 379
7.6.2.3 Computational and Storage Infrastructure 380
7.6.2.3.1 Interactive Computing Resources 381
7.6.2.3.2 Batch Computing Resources 381
7.6.2.3.3 Service Computing Resources 381
7.6.2.4 Cloud Services 382
7.6.2.5 AI and ML 382
7.6.2.6 Wireless Capabilities 382
7.6.2.7 Data-Related Resource Constraints 382
7.6.2.8 Data Mobility Endpoints 382
7.6.2.8.1 bbcp 382
7.6.2.8.2 Globus 383
7.6.2.9 Outstanding Issues 383
7.6.2.10 Technology Profile Contributors 383
7.7 ANL 384
7.7.1 Discussion Summary 384
7.7.2 ANL: Technology Profile 384
7.7.2.1 Facilities Operated and Supported 385
7.7.2.2 Network and Data Architecture 386
7.7.2.2.1 Overview 386
7.7.2.2.2 WAN 387
7.7.2.2.3 Science DMZ 387
7.7.2.2.4 Core 387
7.7.2.2.5 LAN 387
7.7.2.3 Computational and Storage Infrastructure 388
7.7.2.4 Cloud Services 388
7.7.2.5 AI and ML 388
7.7.2.6 Wireless Capabilities 389
7.7.2.7 Data-Related Resource Constraints 389
7.7.2.8 Data Mobility Endpoints 389

7.7.2.9 Outstanding Issues 389
7.7.2.10 Technology Profile Contributors 389

8 Group Discussion 390

8.1 Topic Summary 390
8.1.1 IRI and the HEP Community 390
8.1.2 HEP R&D Efforts 391
8.1.3 Data Challenges 392
8.1.4 LHCONE Support and International Networking 393
8.2 Outcomes 393
8.2.1 Facility Management and Readiness 393
8.2.2 Scientific Data Management 394
8.2.3 Scientific Workflow 394
8.2.4 Computational and Storage Requirements 394
8.2.5 Remote Collaboration and Operational Requirements 394
8.2.6 Multifacility Computational Workflows 395
8.2.7 Networking for Local and Wide-area Data Mobility 395
8.2.8 Emerging Needs 396
8.2.9 IRI Responsiveness 397

List of Abbreviations 398

1 Executive Summary

The world-class research infrastructure at the US Department of Energy (DOE) Office of Science (SC) provides the research community with premier observational, experimental, computational, and network capabilities. Each user facility is designed to provide unique capabilities to advance the core DOE mission in science and technology for its SC program to stimulate rich scientific discoveries and enhance its innovation ecosystem. Research communities gather and flourish around each user facility, bringing together new and enhanced perspectives. The continual reinvention of the practice of science — as users and staff forge novel approaches expressed in research workflows — unlocks new discoveries and propels scientific progress.

Within this research ecosystem, the high-performance computing (HPC) and networking user facilities stewarded by the SC's Advanced Scientific Computing Research (ASCR) program play a dynamic cross-cutting role, enabling complex workflows demanding high-performance data, networking, and computing solutions. The ASCR facilities enterprise seeks to understand and meet the needs and requirements across SC and DOE domain science programs and priority efforts, highlighted by the formal requirements review methodology.

Between July and December 2024, the Energy Sciences Network (ESnet) and the Office of High Energy Physics (HEP) of the DOE-SC organized an ESnet requirements review of HEP-supported program activities. Preparation for these events included identification of key stakeholders: program and facility management, research groups, and technology providers. Each stakeholder group was asked to prepare formal case study documents about its relationship to the HEP program to build a complete understanding of the current, near-term, and long-term status, expectations, and processes that will support the science going forward.

This review includes case studies from the following HEP user facilities, experiments, and joint collaborative efforts:

- Cosmological Simulation Research
- Dark Energy Spectroscopic Instrument (DESI)
- Vera C. Rubin Observatory
- Dark Energy Science Collaboration (DESC)
- Cosmic Microwave Background (CMB)
- Muon Experimentation at Fermilab: Muon g-2 and Mu2e
- Belle-II Experiment
- Neutrino experiments at Fermilab: Short-Baseline Neutrino (SBN) and Deep Underground Neutrino Experiment (DUNE)
- Large Hadron Collider (LHC): A Toroidal LHC Apparatus (ATLAS) Experiment
- LHC: Compact Muon Solenoid (CMS) Experiment
- LHC: LHC Operations
- LHC: HL-LHC and Network Research
- Super Cryogenic Dark Matter Search (SuperCDMS)
- Alpha Magnetic Spectrometer (AMS) Experiment

Additionally, technical profiles are presented in this document reflecting the current state of support offered by DOE SC-funded laboratories and user facilities in collaboration with HEP-supported activities. The laboratories and user facilities profiled include:

- National Energy Research Scientific Computing Center (NERSC)
- Argonne Leadership Computing Facility (ALCF)
- Oak Ridge Leadership Computing Facility (OLCF)
- Fermi National Accelerator Laboratory (Fermilab)
- Brookhaven National Laboratory (BNL)
- SLAC National Accelerator Laboratory (SLAC)
- Argonne National Laboratory (ANL)

A summary of each case study and profile is presented prior to the case study document, along with a "Discussion Summary" that highlights key areas of conversation from authors and attendees. These brief write-ups are not meant to replace a full review of the case study, but provide a snapshot of the discussion and focus during the in-person review.

Review participants spanned the following roles:

- Subject-matter experts from the HEP program activities listed previously;
- ESnet Site Coordinators Committee (ESCC) members from HEP activity host institutions, including the following DOE national laboratories and facilities: ALCF, ANL, BNL, Fermilab, High-Performance Data Facility (HPDF), Lawrence Berkeley National Laboratory (LBNL), NERSC, OLCF, Oak Ridge National Laboratory (ORNL), and SLAC;
- DOE-SC staff spanning ASCR and HEP; and
- ESnet staff.

In recent years, the research communities around the DOE-SC user facilities have begun experimenting with and demanding solutions directly integrated with HPC and data infrastructure. This rise of integrated-science approaches is well documented, and there is a broad need for integrated computational, data, and networking solutions. In response to these drivers, DOE has developed a vision for an Integrated Research Infrastructure (IRI)[1] to empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation.

The IRI vision is fundamentally about establishing new data-management and computational paradigms. Within these, DOE SC user facilities and their research communities build bridges across traditional silos to improve existing capabilities and create new possibilities. Implementation of IRI solutions will give researchers simple and powerful tools with which to implement multifacility research data workflows. This work will also extend analysis done on IRI patterns[2] and discuss ways future HEP workflows can benefit from the approach.

1 "Integrated Research Infrastructure Architecture Blueprint Activity (Final Report 2023)", https://www.osti.gov/biblio/1984466/
2 "ESnet Requirements Review Program Through the IRI Lens: A Meta-Analysis of Workflow Patterns Across DOE Office of Science Programs", https://www.osti.gov/biblio/2205078/

1.1 Summary of Review Findings

The requirements review produced several important findings from the case studies and subsequent virtual conversations.

1.1.1 Facility Management and Readiness

- A common theme of the HEP case studies are challenges of data access throughout a collaboration. Some collaborations feature highly structured ecosystems that facilitate access to groups and individuals via a number of tools that manage identity and permissions. Other collaborations that have less structure may struggle with the available tools and are investigating other solutions (e.g., neutral site hosting through cloud providers, or adopting the policies of the strictest member of a collaboration). [Section 8]
- Removing the need to worry about data locality allows researchers to invest more time on the science, versus spending energy in activities to stage data and allocate technology resources. It also prevents half-measures being used today to create mechanisms to perform analysis on partial data sets, or adopt streaming approaches that are not always successful on some of the larger computational resources not designed to support that mode of operation. [Section 6.1], [Section 8]

1.1.2 Scientific Data Management

- The Rubin Observatory expects ~30 TB of continuous image and metadata to flow from Chile to the US each night. To support the Rubin Observatory, the US Data Facility (USDF) at SLAC will curate 5 PB/yr of data, with ancillary and other image data products that could add up to 500 PB by the end of survey in 2035. A very demanding application is Data Release Processing and Alert Production, where 12.7 GB of pixel data need to be transferred in seven-second bursts from the observatory to the USDF. Two such bursts occur approximately every minute of nighttime observing. The bursts need to be sustained for approximately 355 days each year. [Section 6.4]
- DUNE will produce ~45 PB of data per year at full operation. The first full DUNE detector is slated to begin full operation in mid-2029, and a second module will come online in 2030–2031. Transfer of the raw data produced from the DUNE detectors is likely to be the largest single network task for the experiment. In addition to beam and calibration data, the signal of neutrinos produced from a nearby supernova burst presents a significant challenge due to the large volume of data collected in a very short period (i.e., up to 600 TB in 100 seconds) and the time-critical nature of analyzing that data for optical astronomical observations. [Section 6.8]

1.1.3 Scientific Workflow

- All LHC Tier-1 and Tier-2 computing infrastructure sites are expected to have perfSONAR nodes deployed to monitor network performance. perfSONAR allows network administrators to identify potential network issues that could impact data-transfer rates, both intersite and with CERN. It plays a vital role in the global transfer of data across the Worldwide LHC Computing Grid (WLCG). [Section 6.11]
- Given the volume of data involved, making all Rubin releases available on fast storage for the duration of the project is not feasible. Rubin will keep the two most-recent releases available for rapid access, and archive older releases to tape. Users will not be able to perform database queries against these archived releases. However, they would be made available for bulk download. [Section 6.3]

1.1.4 Computational and Storage Requirements

- Cosmological simulation continues to be substantially limited by the availability of storage, particularly when considering how to distribute the simulation data to collaborators. This issue has several associated dimensions to consider: [Section 6.1]
  - Sizes and numbers of simulations will increase.
  - Locations producing simulations are expected to increase.
  - Support spans different funding agencies.
  - Usefulness over time, and to different communities, is increasing.
- The impact of HL-LHC on storage and compute resources is significant. The corresponding increase in LHC luminosity not only generates significantly more data, but also significantly more complex events, which require more processing to resolve. [Section 6.12]

1.1.5 Remote Collaboration and Operational Requirements

- Data movement of cosmological simulation data between facilities is not a critical problem, but broad access for the user community remains challenging. Simplifying access with updated portal software, or tools like Globus, has improved data mobility. The next step is to combine this with dynamic access to computing resources. [Section 6.1]
- All the Rubin collaborations, as well as independent scientists, will access and analyze data through either the hybrid cloud-based US Data Access Centers (DACs; cloud-based Rubin Science Platform (RSP) and USDF back-end data services) or Independent Data Access Centers (IDACs) operated by in-kind contributor communities. The RSP is a set of web applications and services that will be used to access, visualize, interact with, and analyze data. Data will be exposed to users through the web portal (interactive use cases), the Jupyter Notebook interface (custom user use cases), and machine-accessible application program interfaces (APIs; remote tools/analyses use cases). [Section 6.3]

1.1.6 Multifacility Computational Workflows

- The network speeds between major DOE HPC centers are not a limitation, and transferring PB-sized data sets is effectively routine. Data set sizes are expected to increase in the coming years, but will be within the capabilities of ESnet and the DOE HPC resources. [Section 6.1]
- Having reliable mechanisms to transfer data into HPC facilities from the LHC Tier-1 and Tier-2 computing centers is of growing importance. HPCs are different from other resources used by the LHC primarily because they provide significant CPU resources but very limited disk resources, and because network access to or from HPC compute nodes is typically limited or not allowed. The broader LHC community has considered how to best utilize these resources within operational constraints. Unlike most HPC users, the LHC community cannot concentrate its efforts on using one or a handful of HPC facilities, but instead is motivated to use HPC facilities worldwide. Network access and data management are important aspects of this. [Section 6.9], [Section 6.10]

1.1.7 Networking for Local and Wide-Area Data Mobility

- Data volumes at the two large LHC experiments, ATLAS and CMS, can average 5 PB a day, with volumes exceeding this level during busy periods when the LHC is in full operation. The Tier-1 and Tier-2s within the US can see data flows of tens to hundreds of Gbps on a routine basis. Most are connected via 100 Gbps or 400 Gbps diverse connections to ESnet, or regional networking providers. All US Tier-2 sites are strongly encouraged to provide at least 100 Gbps connectivity, preferably with resilience. The intent is to have multiple 400 Gbps connectivity for each Tier-2 by the time HL-LHC starts. [Section 6.9], [Section 6.10], [Section 8]
- ESnet will investigate ways to integrate resources into LHCONE at non-LHC-affiliated sites. DOE HPC centers (e.g., NERSC) and AFs (e.g., SLAC) could benefit from being connected to better identify and prioritize traffic. [Section 6.9], [Section 8]
- Networking has been fundamental to the success of LHC computing to date, enabling the exploitation of globally distributed resources for computationally limited science. This will remain the case to meet the budget-constrained computing challenges of HL-LHC. Strategies for HL-LHC computing are based on extensive use of powerful networks to reduce data replication by streaming over the network, and consolidating distributed resources into cohesive virtual federations such as data lakes. HL-LHC data and processing scales are large enough that optimizing for efficient bandwidth use will be essential. [Section 6.12]

1.1.8 Emerging Needs

- Commercial cloud resources are an intriguing addition to HEP simulation workflow structures. While the cost of using these resources remains high, augmenting existing resources during periods of heavy use and justifying the increased expenditure may be possible. One could imagine a situation where the use of artificial intelligence (AI) and machine learning (ML) frameworks becomes an instantaneous source of large-scale data; if challenges exist in allocating local computational and storage resources to meet the demand, these workflows may need to be integrated with commercial cloud infrastructure that can expand as needed to fit the needs. [Section 8]
- DESI has used emerging Low Earth Orbit (LEO) satellites at the remote telescope site as a back-up connection for when the primary Internet is down. This was especially critical in 2022 after a wildfire resulted in a multi-month outage of the fiber optic connection to the mountain. The LEO connection was sufficient for command and control of the telescope, but was not sufficient for transferring observational data, which required driving hard disks up and down the mountain. [Section 6.2]
- SENSE (SDN for End-to-End Networked Science at the Exascale) uses Software-Defined Network (SDN) to dynamically manage network paths, prioritize traffic, and provide optimized routing. This setup is ideal for data-intensive tasks, as it can allocate resources in real time based on network conditions and experimental priorities. Rucio, developed by CERN in part for the LHC experiments, complements SENSE by managing large-scale data distribution across the network. [Section 6.11]

1.1.9 IRI Responsiveness

- Long-term storage of data remains a critical problem throughout HEP. The IRI program and HPDF facility should present a possible solution to this problem, providing a deterministic way for long-running efforts can be cataloged, stored, and found by the global community. If a single location cannot become the home for data sets, efforts must be made to unify a cataloging system to enable locality searches, despite the distributed nature of the data. [Section 6.1], [Section 8]
- An area of friction for a number of HEP experiments involves the mechanisms used to manage identity and permissions for the large population of scientific users. A common problem is the management of credentials to access computation and storage across facilities, and how these may differ from accounts used to manage other aspects of the scientific workflow (e.g., home institutional accounts). [Section 8]
- HEP experiments and facilities can benefit from IRI, especially if the DOE HPC centers and the newly awarded HPDF make resources available for use: [Section 6.4], [Section 6.9], [Section 6.10], [Section 6.12]
    - IRI would provide the opportunity to implement near real-time workflows for critical payloads.
    - Incorporating HPC resources typically come with some, potentially substantial, work to set up given facility specific configurations and constraints. A common interface (or access requirements) across DOE HPC sites and models for sustained access are potential IRI benefits.
    - Workflows that are more compute intensive could possibly branch out across multiple sites in the future without having to invest heavily in software and hardware integration steps.

1.2 Summary of Review Actions

Lastly, ESnet plans to follow up with review participants on a number of high-level actions identified. These items are listed as guidance for future collaboration, and do not reflect formal project timelines. ESnet will review these with HEP participations on a yearly basis, until the next requirements review process begins.

1.2.1 Facility Management and Readiness

- The HL-LHC predictions show a mismatch between the computing and storage resources the experiments can afford versus the resources needed to reach the HL-LHC science goals. In response to this gap, the experiments are exploring alternatives in how to utilize storage, computing, and network infrastructure. The network baseline currently being planned is composed of terabit-scale (1–2 Tbps) backbone networks with the largest resource sites connected at multiple 400G scale (800 Gbps+). Network use will be at least a factor of 10 larger than those in the LHC Run 3 program. [Section 6.12]

1.2.2 Scientific Data Management

- The cosmological community is interested in advancing a data management model that would help reduce the number of times and locations that need to be involved to archive long-term data products, and is looking for assistance in scoping and realizing this design goal. Establishing permanent storage facilities would allow the community to establish a more regular pattern of collaboration: a smaller set of data portals that can be used to catalog and retrieve long-term campaigns of products along with a simplified way to support analysis activities, potentially leveraging interactive computing approaches, at a well-known number of sites. [Section 6.1], [Section 8], [Section 5.2]

1.2.3 Scientific Workflow

- DESI uses Globus for transfers of collaboration data, and users report positive experiences with Globus when they want to download an entire directory tree. But they have found the interface lacking for the more common use case of selecting a subset of data from a complex directory tree (e.g., all files matching a given regex regardless of where they appear in the directory structure). Discussions with Globus may be useful to help facilitate ways to address this use case. [Section 6.2]

1.2.4 Computational and Storage Requirements

- The barrier to using DOE HPC resources within the LHC community remains, particularly in integrating the systems as resources into the software components, and normalizing the way the resources could be accounted for (expected versus ad hoc or opportunistic). More specifically, there is a desire to create agreements or other forms of arrangements, similar to an Service Level Agreement (SLA), with DOE HPC facilities such that resources can be considered pledged, like how Tier-1s and Tier-2s operate. This helps with planning for both the HEP community and the ASCR resources. Some of this may require discussion at the programmatic level. [Section 6.12], [Section 8]
- The LHC community experiences sub-optimal performance when supporting a streaming-based workflow at large HPC facilities. Workarounds have been established to support this, either by allowing the worker nodes to stream directly or by leveraging caches and data lakes that are deployed close topologically. All these design choices come with trade-offs, which implies a set of R&D activities that will extend in the coming years. [Section 6.12], [Section 8]

1.2.5 Remote Collaboration and Operational Requirements

- Several facilities are not connected to LHC Open Network Environment (LHCONE) currently, despite being heavily involved in LHC-affiliated computation and storage activities. These include major computational sites, AFs, and potentially commercial clouds. No formal agreement with any of these facilities exists now to be part of the LHC collaboration in a structured way, but investigating their ability to connect to LHCONE to support several important use cases is recommended. [Section 8]

1.2.6 Multifacility Computational Workflows

- Using a single centralized HPC facility to support a given HEP experiment is beneficial, since there's a single ecosystem to learn and it requires minimal overhead to share results with collaborators from other institutions. At the same time, it can also become a single point of failure and requires mitigations to be developed on how to proceed with operations during downtime. The limiting factors are not networking for data transfer, but rather the tools for making that data sharing automated and transparent for hundreds of users accessing PBs of data stored in millions of files. [Section 8]

1.2.7 Networking for Local and Wide-Area Data Mobility

- The DOE HEP and DOE ASCR program offices are working out the specifics of the connectivity to SURF that ESnet will need to complete for the proper connection. The efforts are planned to converge by the end of fiscal year 2026. [Section 6.8], [Section 8]
- LHCONE runs a hybrid approach to supporting Internet Protocol version 6 (IPv6) currently. The upcoming government mandate (e.g., M-21-07) will imply that Fermilab and BNL must move toward IPv6 capabilities, and subsequently the associated LHC Tier-2s will have to move toward dual stack support. One critical issue that cannot be solved by the US community will be the ability of CERN to deliver IPv6 capabilities from its service infrastructure. ESnet will continue to work with US collaborators on this to ease the transition to IPv6 operation. [Section 8]
- Several international networking partners factored heavily into the requirements review discussions; connectivity to Europe, South America, and the Asia-Pacific region is accomplished through relationships with several R&E partners. ESnet maintains peering relations with most of these partners, and their connectivity carries significant amounts of HEP traffic daily. ESnet will continue to work with these partners to ensure networking access to international science locations. [Section 8]

1.2.8 Emerging Needs

- Rucio's adoption throughout the HEP community for distributed data management provides an opportunity for ESnet and other Research and Education (R&E) networks to explore data-movement intentions and required quality of service (QoS) or deadline requirements. ESnet will continue to work with the LHC community on co-design efforts such as this, along with SENSE, SciTags, and other potential opportunities. [Section 6.11], [Section 6.12], [Section 8]
- Several HEP case studies outlined their interest in commercial cloud resources. ESnet is identifying additional ways to assist, and will continue to help with connectivity and peering when applicable. Networking can be a bottleneck in certain situations: mainly when traffic remained within a cloud versus being allowed to egress and use more predicable R&E routing and transit. ESnet can engage with experiments, sites, and cloud-providers to suggest potential solutions. [Section 8]

1.2.9 IRI Responsiveness

- The discussions within the HEP community touched on several factors that should be considered by the IRI effort. A first step will be to establish a hierarchy of needs: an example being that accessing a set of resources or tools which span administrative boundaries is very different than the ability to use different tools and resources within the same location. A similar policy language and control structure could solve both, if ways to differentiate the needs and articulate the required control structures exist. [Section 8]

2 Requirements Review Overview

ESnet and ASCR use requirements reviews to discuss and analyze current and planned science use cases and anticipated data output of a particular program, user facility, or project to inform ESnet's strategic planning, including network operations, capacity upgrades, and other service investments.

2.1 Purpose and Process

The requirements review process, when performed regularly and comprehensively, surveys major science stakeholders' plans and processes to investigate data-management requirements over the next 5–10 years. Questions crafted to explore this space include the following:

- How, and where, will new data be analyzed and used?
- How will the process of doing science change over the next 5–10 years?
- How will changes to the underlying hardware and software technologies influence scientific discovery?

Requirements reviews help ensure that key stakeholders have a common understanding of the issues and the actions that ESnet may need to undertake to offer solutions. The ESnet Science Engagement team leads the effort and relies on collaboration from other ESnet teams. ESnet meets with each individual program office within the DOE SC every three years, with an intermediate virtual update scheduled between the full review. ESnet collaborates with the relevant program managers to identify the appropriate principal investigators, and their information technology partners, to participate in the review process. ESnet organizes, convenes, executes, and shares the outcomes of the review with all stakeholders.

Requirements reviews are a critical part of a process to understand and analyze current and planned science use cases across the DOE SC. This is done by eliciting and documenting the anticipated data outputs and workflows of a particular program, user facility, or project to better inform strategic planning activities. These include, but are not limited to, network operations, capacity upgrades, and other service investments for ESnet, as well as a complete and holistic understanding of science drivers and requirements for the program offices.

We achieve these goals through reviewing the case study documents, discussions with authors, and general analysis of the materials. The resulting output is a set of review findings and actions that will guide future interactions between HEP, ASCR, and ESnet. These terms are defined as follows:

- **Findings:** key facts or observations gleaned from the entire review process that highlight specific challenges, particularly those shared among multiple case studies.
- **Actions:** potential strategic or tactical activities, investments, or opportunities that can be evaluated and potentially pursued to address the challenges laid out in the findings.

2.2 Structure

The requirements review process is hybrid, and relies on a combination of asynchronous and synchronous activities to understand specific facility and experimental use cases. The review is a highly conversational process through which all participants gain shared insight into the salient data-management challenges of the subject program/facility/project. Requirements reviews help ensure that key stakeholders have a common understanding of the issues and the potential actions that can be implemented in the coming years.

2.2.1 Background

Through a case study methodology, the review provides ESnet with information about the following:

- Existing and planned data-intensive science experiments and/or user facilities, including the geographical locations of experimental site(s), computing resource(s), data storage, and research collaborator(s).
- For each experiment/facility project, a description of the "process of science," including the goals of the project and how experiments are performed and/or how the facility is used. This description includes information on the systems and tools used to analyze, transfer, and store the data produced.
- Current and anticipated data output on near- and long-term timescales.
- Timeline(s) for building, operating, and decommissioning of experiments, to the degree these are known.
- Existing and planned network resources, usage, and "pain points" or bottlenecks in transferring or productively using the data produced by the science.

2.2.2 Case Study Methodology

The case study template and methodology are designed to provide stakeholders with the following information:

- Identification and analysis of any data-management gaps and/or network bottlenecks that are barriers to achieving the scientific goals.
- A forecast of network capacity requirements by area of science, particularly in geographic regions where data production/consumption is anticipated to increase or decrease.
- A survey of the data-management needs, challenges, and capability gaps that could inform strategic investments in solutions.

The case study format seeks a network-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the network services needed; and how the network will be used over three timescales: the near term (immediately and up to two years in the future); the medium term (two to five years in the future); and the long term (greater than five years in the future).

The case studies address the following sections with review participants:

Science Background: a brief description of the scientific research performed or supported, the high-level context, goals, stakeholders, and outcomes. The section includes a brief overview of the data life cycle and how scientific components from the target use case are involved.

Collaborators: aims to capture the breadth of the science collaborations involved in an experiment or facility focusing on geographic locations and how datasets are created, shared, computed, and stored.

IRI Readiness: Research communities that use DOE SC user facilities are experimenting with and demanding solutions integrated with HPC and data infrastructure. The Integrated Research Infrastructure Architecture Blueprint Activity (IRI-ABA) brought together domain experts from all DOE SC Programs to look for common patterns within diverse workflows across a range of scientific disciplines. Participants discovered three common patterns:

- Time-sensitive pattern.
- Data integration–intensive pattern.
- Long-term campaign pattern.

Instruments and Facilities: description of the instruments and facilities used, including any plans for major upgrades, new facilities, or similar changes. When applicable, descriptions of the instrument or facility's compute, storage, and network capabilities are included. An overview of the composition of the datasets produced by the instrument or facility (e.g., file size, number of files, number of directories, total dataset size) is also included.

Process of Science: documentation on the way in which the instruments and facilities are and will be used for knowledge discovery, emphasizing the role of networking in enabling the science (where applicable). This should include descriptions of the science workflows, methods for data analysis and data reduction, and the integration of experimental data with simulation data or other use cases.

Remote Science Activities: use of any remote instruments or resources for the process of science and how this work affects or may affect the network. This could include any connections to or between instruments, facilities, people, or data at different sites.

Software Infrastructure: discussion of the tools that perform tasks, such as data-source management (local and remote), data-sharing infrastructure, data-movement tools, processing pipelines, collaboration software, etc.

Network and Data Architecture: the network architecture and bandwidth for the facility and/or laboratory and/or campus. The section includes detailed descriptions of the various network layers (local-area network [LAN], metropolitan-area network [MAN], and wide-area network [WAN]) capabilities that connect the science experiment/facility/data source to external resources and collaborators.

Emerging Technology Use: The use of new technologies in the scientific process is a constant evolution. Cloud services, artificial intelligence (AI) / machine learning (ML) technologies, and the use of nonterrestrial networking are growing in popularity to support research activities around the world. This section asks questions about the role these technologies, and potentially others, are playing in the overall trajectory of the scientific use case.

Cloud Services: if applicable, cloud services that are in use or planned for use in data analysis, storage, computing, or other purposes.

Artificial Intelligence and Machine Learning (AI/ML): Outlining how the laboratory or facilities have invested in AI/ML technologies. This could be the purchase and operation of on-site capabilities available for use or partnerships with other entities (research and education [R&E], commercial, etc.) to use remote resources.

Remote or Field Research: Discussing aspects of a scientific workflow that depend on resources that may be located in an isolated area, e.g., a sensor array or instrument deployed far from terrestrial networking that may need access to cellular or satellite connectivity.

Data-Related Resource Constraints: any current or anticipated future constraints that affect productivity, such as insufficient data-transfer performance, insufficient storage system space or performance, difficulty finding or accessing data in community data repositories, or unmet computing needs.

Data Mobility Endpoints: If a facility or experiment has dedicated infrastructure to facilitate data sharing, ESnet is interested in learning more about how it is constructed and maintained. ESnet maintains a set of well-tuned test endpoints, and recommends regular testing to evaluate data-transfer capabilities.

Outstanding Issues: an open-ended section where any relevant challenges, barriers, or concerns not discussed elsewhere in the case study can be addressed by ESnet.

2.3 ESnet

ESnet is the high-performance network user facility for the US DOE SC and delivers highly reliable data transport capabilities optimized for the requirements of data-intensive science. In essence, ESnet is the data circulatory system that enables the DOE science mission by connecting all its laboratories and facilities in the US and abroad. ESnet is funded and stewarded by the ASCR program and managed and operated by the Scientific Networking Division at LBNL. ESnet is widely regarded as a global leader in the R&E networking community.

ESnet interconnects DOE national laboratories, user facilities, and major experiments so that scientists can use remote instruments and computing resources as well as share data with collaborators, transfer large datasets, and access distributed data repositories. ESnet is specifically built to provide a range of network services tailored to meet the unique requirements of the DOE's data-intensive science.

In short, ESnet's mission is to enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools. ESnet's vision is summarized by these three points:

1. Scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data.
2. Collaborations at every scale, in every domain, will have the information and tools they need to achieve maximum benefit from scientific facilities, global networks, and emerging network capabilities.
3. ESnet will foster the partnerships and pioneer the technologies necessary to ensure that these transformations occur.

2.4 About ASCR

The mission of the ASCR program is to discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to the DOE. A particular challenge of this program is fulfilling the science potential of emerging computing systems and other novel computing architectures, which will require numerous significant modifications to today's tools and techniques to deliver on the promise of exascale science.

To accomplish its mission and address the challenges described previously, the ASCR program is organized into three subprograms:

- The Mathematical, Computational, and Computer Sciences Research subprogram develops mathematical descriptions, models, methods, and algorithms to describe and understand complex systems, often involving processes that span a wide range of time and/or length scales.
- The HPC and Network Facilities subprogram delivers forefront computational and networking capabilities and contributes to developing next-generation capabilities through support of prototypes and test beds.
- The Advanced Computing Technology Division supports research focused on the development of emerging computing technologies such as quantum information systems (QIS) and neuromorphic efforts as well as investments in microelectronics in partnership with the other Office of Science program offices, the Exascale Computing Project (ECP), Research and Evaluation Prototypes (REP), and ASCR-specific investments in cybersecurity.

2.5 About the Office of HEP

HEP explores what the world is made of and how it works at the smallest and largest scales, seeking new discoveries from the tiniest particles to the outer reaches of space. This quest inspires young minds, trains an expert workforce, and drives innovation that improves the nation's health, wealth, and security.

Our research is inspired by some of the biggest questions about our universe. What is it made of? What forces govern it? How did it become the way it is today? Finding these answers requires the combined efforts of some of the largest scientific collaborations in the world, using some of the most sensitive detectors in the world, at some of the largest scientific machines in the world.

HEP supports US researchers who play leading roles in these international efforts and world-leading facilities at our national laboratories that make this science possible. We also develop new accelerator, detector, and computational tools to open new doors to discovery science, and through Accelerator Stewardship work to make transformational accelerator technology widely available to science and industry.

HEP is organized into six subprograms:

- **Energy Frontier:** Researchers at the Energy Frontier use the world's largest and highest-energy particle accelerator to recreate the universe as it was a billionth of a second after the big bang.
- **Intensity Frontier:** Researchers at the Intensity Frontier investigate some of the rarest particle interactions in nature and subtle effects that require large data sets to observe and measure.
- **Cosmic Frontier:** Researchers at the Cosmic Frontier use naturally occurring cosmic particles and phenomena to reveal the nature of dark matter, cosmic acceleration, and more.
- **Theoretical, Computational, and Interdisciplinary Physics:** Theoretical and computational physics provide the vision and the framework for extending our knowledge of particles and the universe.
- **Advanced Technology R&D:** Cutting-edge research in the physics of particle accelerators, particle beams, and particle detection enables scientists to stay on the threshold of discovery.
- **Accelerator Stewardship:** Supporting use-inspired basic research in accelerator science and technology to make particle accelerator technology widely available to science and industry.

3 Assessment of 2020 HEP Requirement Review

ESnet recorded a set of action items from the 2020 HEP Network Requirements Review, with the goal of continuing support of collaborations funded by the HEP program beyond the publication of the report. ESnet performed regular check-ins with HEP between the reviews (holding webinars with the community in 2022 and 2023) to ensure progress was made on these identified areas of mutual interest. These sections outline the previous actions and discuss progress made since the publication of the report.

3.1 Prior Action Items

The actions identified during the 2020 HEP Network Requirements Review included:

- ESnet will work with laboratories, user facilities, experiments, domestic networking peers, universities, partners, and international peers in the coming years to evaluate and upgrade network connections to support essential science missions.
- ESnet will continue to provide leadership in the development, implementation, and operation of cyberinfrastructure components (e.g., the Science DMZ, Data Transfer Nodes (DTNs), perfSONAR, and the Modern Research Data Portal) used by the scientific user community. Emerging testbeds and technologies, including AutoGOLE, Adaptive ProgrammaBle Research Infrastructure for Computer Science and Science Applications (FABRIC), and Software-Defined Networking (SDN) for End-to-End Networked Science at the Exascale (SENSE), will see increased adoption as they are deployed more widely in the coming years.
- ESnet will continue to evaluate and expand international connectivity options in support of international scientific requirements.
- ESnet will continue to work closely with LHC R&D efforts to develop and deploy new capabilities and services that address scientific use cases. These efforts include methods to better measure and track network traffic, manage and provision networks, and improve in-network caching capabilities.
- ESnet will remain active in the design, implementation, and operation of the Large Hadron Collider Open Network Environment (LHCONE) effort as it expands scope and usage to support the LHC and affiliated scientific experiments.
- ESnet will collaborate with ASCR computing facilities to ensure connectivity matches scientific data requirements at all stages of operation (intake, dissemination, etc.).

3.2 Assessment

The 2020 HEP Network Requirements Review action items were assessed and reviewed with the HEP community. This section does not offer a line-by-line response for each, but instead provides an update on several major initiatives undertaken by the DOE SC programs, along with individual actions by ASCR, HEP, and ESnet.

3.2.1 ESnet Actions

Since the 2020 review, ESnet completed the design, procurement, and installation of ESnet6, the next-generation networking platform that is currently running to deliver on DOE SC mission requirements. This was a multiyear project that involved complete replacement of the underlying optical and routed network that ESnet operates and resulted in upgraded capacities across the backbone and to connected laboratories and facilities.

Supporting HEP through transatlantic (TA) connectivity continues through the current time, with many upgrades and changes implemented since the previous review cycle. ESnet continues to collaborate with international partners (e.g., GÉANT) to establish networking spectrum on subsea cable resources. Efforts are underway to

share this resource across several diverse cable resources on a long-term basis (e.g., 10- to 15-year contracts). This will result in approximately 10 Tbps of aggregate connectivity, across multiple connections, that will be available to support the US and EU science missions, addressing the requirements of the LHC community projections for Run 4 data production. By the 2027 time frame, 3.2 Tbps of capacity is expected to be available.

ESnet has also initiated the following efforts and programs that address some of the actions from the 2020 HEP Network Requirements Review:

- Upgrading the connectivity to support the LHC T1 sites in the US to support increased data volumes, and planning for more upgrades when technology and procurement of equipment become available:
  - BNL now supports 1.6 Tbps.
  - Fermilab now supports 2.4 Tbps.
- Discussing connectivity upgrade options with regional networking partners that support LHC T2 facilities, and setting a schedule to upgrade connection capacities, peering locations, and downstream connections to campuses. All regional networks now support at least 100 Gbps connections, with many also upgrading to 400 Gbps, or slating to do so within the 2025 to 2026 time frame. Downstream campuses are also connected using multiple connections of at least 100 Gbps, or supporting 400 Gbps.
- Fully adopting and supporting the LHCONE networking overlay, and the ability of other affiliated scientific efforts (e.g., DUNE, Belle II, etc.) to utilize this for network traffic that traverses similar computational sites and network paths.
- Creating a program (i.e., the Fasterdata DTN framework) as a mechanism to support data mobility needs across the backbone and regional networking infrastructure, as well as touchpoints in campus environments.
- Deploying the Stardust measurement and monitoring platform, which allows ESnet connectors to visualize and categorize their scientific network patterns.
- Upgrading to all perfSONAR and DTN hardware to at least 100 Gbps-capable speeds.
- Participating (continued) in LHC-affiliated R&D efforts including SENSE and Rucio integration, Scientific Network Tags (SciTags), deployment of networking caching, exploration of data lake performance, and others.
- Making an organizational commitment to participate in "co-design" efforts. These joint projects between ESnet engineering staff, along with scientific experts at other facilities and projects, facilitate access to technical expertise and build strong relationships throughout the scientific community.

3.2.2 DOE SC Actions

A major focus for ASCR, along with other DOE SC programs like HEP, in the time between the 2020 and 2024 HEP Network Requirements Review was the convening of groups to discuss ways that the future generation of technology provided on the network, and within computational facilities, will grow to meet the demands of scientific experimentation.

The complexity of scientific pursuits is increasing rapidly with aspects that require dynamic integration of experiment, observation, theory, modeling, simulation, visualization, ML, AI, and analysis. Research projects across the DOE are increasingly data and compute intensive. Innovative research teams are accelerating the pace of discovery by using high-performance computational and data tools in their research workflows and leveraging multiple research infrastructures. DOE SC has embraced what can be possible with this integration and has

adopted a vision for IRI[3]: to empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation.

IRI defined three core patterns during its early work:

- **Time-sensitive pattern: r**equires real-time, or near real-time, response across more than one facility or resource for timely decision-making and experiment steering.
- **Data integration–intensive pattern:** requires combining and analyzing data from multiple sources, e.g., sites, experiments, and/or computational runs, to deepen and expand context.
- **Long-term campaign pattern:** requires sustained access to more than one facility or resource, often centered around rare or unique scientific instruments, over years or decades to accomplish a well-defined objective.

These activities address some of the actions from the 2020 HEP Network Requirements Review, namely in trying to define a sensible set of requirements that DOE HPC facilities can deliver to assist the future needs of experimental workflows and better scale service offerings for emerging efforts. These observations come from the broader work to classify ESnet Requirements Reviews through the IRI lens.[4]

IRI acknowledges that data movement, be it in the form of bulk or streaming requirements, is a core goal and must be supported fully by all experimental facilities that connect to ESnet. ESnet is committed to working with HEP, and other DOE SC programs, to execute on the vision of IRI starting with the items identified in the 2024 HEP Network Requirements Review.

3 https://www.osti.gov/biblio/1984466
4 https://www.osti.gov/biblio/2008205

4 Review Findings

The requirements review process helps to identify important facts and opportunities from the programs and user facilities that are profiled. The following sections outline a set of findings from the HEP and ESnet requirements review. These points summarize important information gathered during the review discussions surrounding case studies and the HEP-managed user program in general. These findings are organized by topic area for simplicity and by common themes:

- Facility management and readiness.
- Scientific data management.
- Scientific workflow.
- Computational and storage requirements.
- Remote collaboration and operational requirements.
- Multifacility computational workflows.
- Networking for local and wide-area data mobility.
- Emerging needs.
- IRI responsiveness.

4.1 Facility Management and Readiness

- Cosmological simulations provide theoretical predictions to understanding physical phenomena. These data products help to analyze and interpret results from cosmological surveys, and aid in survey design, optimization, and error estimates. [Section 6.1]
- DESI is a DOE-led cosmology collaboration that will operate through at least 2026 and create the world's largest 3D map of the universe. DESI utilizes the Mayall 4-meter telescope, located at Kitt Peak National Observatory (KPNO). NERSC is used as the primary scientific analysis center for DESI. [Section 6.2]
- The Rubin Observatory, an NSF-DOE funded partnership operated by NOIRLab and SLAC, will carry out the Legacy Survey of Space and Time (LSST), using the Simonyi Survey Telescope and the Rubin Observatory LSST Camera. It will take repeated images of the southern sky, and the resulting dataset will provide a static census of approximately 40 billion objects. The primary deliverables for Rubin Observatory include a real-time stream of "alerts" of transient events, prompt data products, annual data-release data products, and a science platform. [Section 6.3]
- Rubin Observatory pre-operations ran early versions of all the data production elements between 2021 and 2023 in a commercial cloud environment (i.e., the "Interim Data Facility" [IDF]) using precursor data sets. Transition from the IDF to the permanent US Data Facility (USDF) was completed in FY24. Pre-operations will culminate with data production at full-scale testing in FY25. The full LSST will run for 10 years, and final data processing will take place after that for two years. Survey data taking ends in FY35. [Section 6.3]
- The DESC international science collaboration will make high-accuracy measurements of fundamental cosmological parameters using data gathered from the LSST produced by the Rubin Observatory, and investigate the nature of dark energy and dark matter over 10 years. The DESC collaboration currently comprises some 1,270 members, originating from 23 countries and more than 200 institutions. Of the members, about 50% are from the United States, 9% from the United Kingdom, and 7% from France. [Section 6.4]

- The CMB, or relic radiation, consists of photons created in the Big Bang that have been propagating through the universe continuously until detected in current times. CMB-S4, a joint project between the National Science Foundation (NSF) and the DOE, is the "Stage 4" ground-based CMB experiment, bringing the entire US community together to build a single experiment capable of detecting the faintest CMB signals. The CMB-S4 collaboration currently consists of 520 members at 142 institutions in 24 countries and 26 US states. The institutions include DOE, NASA, and Department of Commerce laboratories, as well as a host of universities. [Section 6.5]
- In mid-2024, the NSF was not able to commit to funding for portions of the project that involved experimentation to be conducted at the South Pole.[5] Thus, the information provided by CMB-S4 in this report assumes the entirety of the observatory will be based in Chile. The experimentation will produce three coordinated surveys from Chile between 2028 to 2035: [Section 6.5]
  - A wide and deep survey of ~70% of the sky using 2 x 6m telescopes that operate with a daily cadence.
  - An ultra-deep survey of ~3% of the sky using 18 x 0.5m telescopes.
  - An ultra-deep survey of the same ~3% of the sky using a different 5m telescope.
- The Muon g-2 experiment at Fermilab collected data from 2018 until summer of 2023 and had two major releases of results: the first in 2021 based on Run 1 data and the second in 2023 based on Runs 2+3 data. The collaboration is analyzing the remaining data (Runs 4–6) and expects to make a final measurement during FY25. [Section 6.6]
- The Mu2e collaboration features a total of 239 scientists from 39 institutions in six countries: China, Germany, Italy, Russia, the UK, and the US. Mu2e will start taking cosmic-ray commissioning data early in mid CY 2025 and will begin commissioning with beam operations late in calendar 2026. The first physics run will begin early in CY 2027 and will continue to the end of CY 2027. During 2028 and 2029 Fermilab accelerator operations will pause for upgrades to the accelerator complex in preparation for the Long-Baseline Neutrino Facility (LBNF). Mu2e will resume running in 2030 and continue until the 2033 summer shutdown. [Section 6.6]
- The Belle II experiment collects data at the SuperKEKB asymmetric electron-positron collider at the Japanese High-Energy Accelerator Research Organization KEK in Tsukuba, Japan. Belle II's goal is to accumulate about 50 times more data than Belle. Operations began in 2019, and about half as much data as Belle collected has been collected. The plan is to increase the amount of data collected per year to reach the goal by the 2030s. The Belle II collaboration comprises 1,183 members from 125 institutions and 28 countries around the world. There are 118 US collaborators from 16 institutions, including BNL. [Section 6.7]
- Belle II recently resumed data taking in March 2024 after a 19-month shutdown to replace the innermost vertex detector and a portion of the particle-identification detector as well as complete an upgrade of the data acquisition (DAQ). A SuperKEKB upgrade with a corresponding ~1.5 year shutdown, tentatively scheduled for 2029, will be required to reach design goals. [Section 6.7]
- The Short-Baseline Neutrino (SBN) program consists of three detectors — the Short-Baseline Near Detector (SBND), MicroBooNE, and the short-baseline far detector Imaging Cosmic And Rare Underground Signals (ICARUS T600) — each detecting neutrinos from Fermilab's Booster Neutrino Beamline (BNB). The design of the SBN detectors will prepare for the future

5 Scientific research using CMB data continues to be a priority for DOE and NSF, an initiative that provides crucial information on the fundamental nature of matter and energy in the universe. However, DOE and NSF jointly decided in calendar year 2025 that they can no longer support the P5-envisioned CMB-S4 project. Nevertheless, DOE and NSF will continue to partner with the CMB science community to explore the potential science that can be achieved with limited upgrades to existing experiments to further this important U.S.-led research.

long-baseline neutrino experiment: DUNE. The SBN international collaborations include approximately 250 scientists from more than 50 institutions across the US, CERN, Italy, the UK, Switzerland, Spain, Turkey, and Brazil. [Section 6.8]

- DUNE will study the fundamental nature of neutrinos through measurements of neutrino oscillations from both neutrino beam and other natural neutrino sources. In addition to studying neutrino oscillations, DUNE has a broader physics program that includes searches for proton decay, detection of neutrinos from supernova collapse, and other exotic phenomena. The DUNE Collaboration comprises approximately 1,500 scientists from more than 200 institutions in 39 countries around the world. These institutions include both national laboratories and universities. [Section 6.8]
- The core facilities of DUNE are the near and far detectors, located on-site at Fermilab (Illinois) and in the Sanford Underground Research Facility (SURF) in South Dakota respectively, and will interact with the beam of the future LBNF at Fermilab. The network that connects SURF to the R&E network infrastructure will be critical to the success of DUNE's operation. [Section 6.8]
- DUNE has built two prototypes at CERN: ProtoDUNE Horizontal Drift (ProtoDUNE-HD) and ProtoDUNE Vertical Drift (ProtoDUNE-VD). ProtoDUNE-HD just completed a three-month beam run at CERN in the fall of 2024, and ProtoDUNE-VD is anticipated to start a roughly yearlong test period in summer 2025. [Section 6.8]
- The work of SBN and the ProtoDUNE experiments at CERN will prepare for DUNE, which is scheduled to start in several years' time. Both experiments will use grid-computing approaches provided by Open Science Grid (OSG), Worldwide LHC Computing Grid (WLCG), and HEP-community software for data movement, cataloging, simulation, and analysis. [Section 6.8]
- The LHC at CERN is the most powerful particle accelerator in the world. The LHC collides protons more than a billion times every second, out of which the detector experiments select interesting collisions for recording at a rate of a thousand times every second. Over 2,000 hours of data collection every year when the LHC is running means a huge data sample for physicists to analyze. [Section 6.9], [Section 6.10]
- ATLAS is the largest of four particle detectors that measure and record the particle collisions at the LHC, and consists of a global collaboration of physicists, engineers, technicians, students, and support staff from around the world. It is one of the largest collaborative efforts, with over 5,500 members and almost 3,000 scientific authors. [Section 6.9]
- The CMS experiment at the LHC is designed to probe new phenomena at the energy frontier. The CMS collaboration is made up of more than 3,000 members from more than 50 countries. Researchers at US institutions comprise about 30% of the collaboration. [Section 6.10]
- CERN facilities are used to provide the ATLAS and CMS data reconstruction, to support detector commissioning work as well as physics analysis work. Then globally distributed Tier-1 and Tier-2 facilities are responsible for data archiving, simulated data generation, analysis data storage, and physics analysis activities. [Section 6.9], [Section 6.10]
- The LHC schedule has operating periods (called "runs") interspersed with maintenance and upgrade periods (called "long shutdowns"). The first run of the LHC was 2009–2012 followed by long shutdown 1 (LS1) from 2012–2014. The second run (Run 2) was from 2014–2018, followed by LS2 from 2019–2021. Run 3 started in 2022 and will continue until 2026. The LHC will be upgraded to increase its luminosity during LS3, from 2026–2029. The new accelerator, HL-LHC, will start producing collisions in Run 4 in 2030. ATLAS and CMS are expected to produce 10 times more data in Run 4, compared with what will be produced in Run 3. [Section 6.11]

- The collaborators involved with LHC operations are the four main LHC experiments (ALICE, ATLAS, CMS, and LHCb) as well as the technical teams running the accelerator and the globally distributed middleware (OSG, European Grid Infrastructure [EGI], Rucio, perfSONAR, File Transfer Service [FTS], etc.) and infrastructures (the various Tier-x sites) supporting the activities of the experiments. [Section 6.11]
- SuperCDMS SNOLAB is a next-generation experiment aiming to detect sub-giga-electronvolt (GeV) dark matter particles funded and operated under the DOE SC HEP program in the Cosmic Frontier, and co-funded by NSF. Non-US funding is provided through the Canadian Foundation for Innovation (CFI). The SuperCDMS collaboration is spread between many institutions in the US, Canada, Germany, India, Spain, United Arab Emirates, and United Kingdom. The approximate count is about 130 collaborators in 28 institutions. [Section 6.13]
- The SuperCDMS SNOLAB experiment is currently in the installation phase at SNOLAB (Canada), 2 km underground. Commissioning is expected to start in fall 2025, and science operations are expected to start at the beginning of 2026 and to last for at least three years, leaving open the possibility of upgrades (more detectors and/or new technology to leverage the cryogenic facility). [Section 6.13]
- AMS is a precision particle physics detector installed on the International Space Station (ISS) in May 2011, conducting a unique, long-duration mission of fundamental physics research in space that is not replicable on earth. AMS is the largest instrument, and is scheduled to run until the ISS project ends, estimated to be after 2030. Since 2011, AMS has collected more than 200 billion events. The AMS collaboration is composed of about 600 physicists and 44 institutes and organizations from 14 countries. [Section 6.14]
- NERSC supplies supercomputing, data storage, networking, auxiliary equipment, and technical consulting to projects supported by the DOE SC. NERSC supports approximately 100 HEP projects, both simulation and data focused. NERSC-10 is expected in the 2027 time frame, and is anticipated to have 10 times the aggregate performance of Perlmutter. [Section 7.1]
- The ALCF operates multiple supercomputing systems and enclaves that support all communities, including HEP. The ALCF-4 system is in the early planning stages, and expected in the three- to four-year time frame. [Section 7.2]
- The OLCF operates Frontier, and other systems, to support DOE SC communities such as HEP. OLCF is preparing for the commissioning of the next HPC system, OLCF-6 and titled "Discovery." The next system is expected in the 2027–2028 time frame. [Section 7.3]
- Fermilab is the DOE's particle physics and accelerator laboratory. It is a single-purpose laboratory, with a mission to support DOE-sponsored HEP experiments and activities. Fermilab is the US Tier-1 center for the LHC/CMS experiment, hosting roughly 40% of CMS experiment data. The laboratory is also the host institution for the DUNE experiment and supports the Tier-0 center for the experiment. Other active Fermilab-based experiments include SBN and SBD, NuMI Off-axis νe Appearance (NOvA), Mu2e, Muon g-2, ICARUS, and MicroBoone. [Section 7.4]
- The Scientific Data and Computer Center (SDCC) at BNL began in 1997 when the Relativistic Heavy Ion Collider (RHIC) and ATLAS Computing Facility were established. SDCC provides dedicated data processing, storage, and analysis resources for a set of diverse, expansive experiments with general computing capabilities and support for users. [Section 7.5]
- The primary scientific computation and storage facilities at SLAC are provisioned in the SLAC Shared Scientific Data Facility (S3DF). The S3DF is a compute, storage, and network architecture designed to support massive-scale analytics required by all SLAC experimental

facilities and programs. The S3DF infrastructure is optimized for data analytics and is characterized by large, massive throughput, high-concurrency storage systems. [Section 7.6]

4.2 Scientific Data Management

- Cosmology data sets are typically owned by the team that generated them, unless the work was specifically part of, or commissioned by, a collaborative activity. The transition toward open science, along with increased use for new domains, has led to more interest in placing simulation results in the public domain. Funding agencies often require datasets to be publicly available; the barrier to compliance is a lack of technical resources (e.g., computing, storage) to make this possible. [Section 6.1], [Section 8]
- Cosmology simulations tend not to lose value over time. Older simulation catalogs continue to have high value and are being used in emerging work that utilizes AI and ML to construct generative models. Simulations can never truly be reproduced, even if the original codes and inputs are retained: differences in the underlying behavior of the systems that are used to generate are common, resulting in almost entirely new simulations. Older simulations will always have an intrinsic value, since they are unique and nonreproducible. However, the community understands that it is infeasible to keep every data set around forever, and mechanisms exist to archive sets that may not receive as much use as others. [Section 6.1], [Section 8]
- Data storage needs for cosmological simulations post mid-2020s are likely to be in the 300 petabyte (PB) range of spinning disk equivalent, requiring transfer rates of >300 Gbps on an intermittent basis when data mobility is required. [Section 6.1]
- The network requirements for DESI must support transfer of an entire night's observation (e.g., 100 GB or more). Tools including rsync are used help to keep a sliding window of 40 terabyte (TB) of previous observations up to date. In a worst-case scenario, transportable media could be used to take an evening's worth of data manually down to the base station, or through a slower network path such as a low Earth orbit (LEO) Satellite (e.g., Starlink). [Section 6.2], [Section 8]
- At NERSC, the DESI data volume is projected to grow at a rate of ~1 PB/year from 2020 to 2025, reaching ~10 PB by the end of 2026. Approximately 1–2 PB originating from the Rubin Observatory with each yearly data release will be in this calculation, along with simulations produced at the DOE HPC facilities, which are expected to be several 100 TB, and data catalogs produced during the analysis process, which are estimated to be 20 TB. Extensions to the DESI effort have been proposed; if approved, these would continue the data growth at a similar rate through the mid-2030s. [Section 6.2]
- Public DESI datasets will be curated by the Astro Data Lab at NOIRLab to be available beyond project completion. The intention is to also host the public DESI dataset via the Cosmology Data Repository at NERSC, though long-term funding commitment for that is not guaranteed. [Section 6.2]
- The Rubin Observatory expects ~30 TB of continuous image and metadata to flow from Chile to the US each night. Real-time transfer over dedicated high-bandwidth segments is required to analyze each image at the USDF (hosted at SLAC) and broadcast alerts to the community. To support the Rubin Observatory, the USDF at SLAC will curate 5 PB/yr of data, with ancillary and other image data products that could add up to 500 PB by end of survey in 2035. [Section 6.3]

- Once in full operations, major Rubin Observatory data releases for static cosmology analyses will be generated yearly. The bulk of data will originate with Rubin and be of the order of 1–2 PB to be transferred every two to three years, and, assuming some data reduction strategies, could eventually be yearly. [Section 6.4]
- DESC has tested Globus transfers of tens and hundreds of TB between NERSC and international sites, with transfer rates of 400 MB/s and 180 MB/s respectively. [Section 6.4]
- DESC simulation activities will require data movement from other computing centers to NERSC. Data transfers will be on the order of 100 TB initially. When the Rubin Observatory Data Previews are present at SLAC, an additional 10–100 TB will be transferred to NERSC; this is expected in May 2025 and early 2026. The goal is to support PB-scale transfers between the facilities, and potentially to international collaborators to prepare for full production. [Section 6.4]
- In operations, CMB-S4 will produce a data rate of approximately 2 Gbps. Data will be transferred in real time over fiber networks to the primary data center at the DOE's NERSC, and copied to the secondary data center at the DOE's ALCF. 20 TB of data are expected daily during operations (~7 PB of data annually). Along with the raw data files CMB-S4 will create, the primary data products will be the frequency maps: [Section 6.5]
  - Daily (7 years x 365 days): 150 TB in 51,000 files.
  - Monte Carlo (MC) (1,000 realizations x 7 years): 450 TB in 150,000 files.
- CMB-S4 will deploy several telescope resources starting in 2030. These resources have the potential to produce 8 PB/year of data, sent to the US nightly in 20 TB sets. Local computation and storage resources can support storing a buffer of data for up to a month if necessary due to network degradation or downtime. [Section 6.5]
- CMB-S4 will conduct a series of data challenges, starting in 2025 and running yearly through 2029, to help prepare for the operational challenge of data mobility from Chile to the US, and the analysis pipeline being constructed using DOE- and NSF-funded computing resources. [Section 6.5]
- The Muon g-2 experiment will complete its studies with nearly 7 PB of raw experimental data, and an additional 1.5 PB of processed results. All data are planned to remain on tape at Fermilab, and at the current time no backup is planned or required. Once the processing is complete, the collaboration does not anticipate reprocessing will happen at this time. [Section 6.6], [Section 8]
- The first aspect of operational Mu2e science will be to handle cosmic-ray events, which are expected to produce up to 500 TB of event data per year. Most of this effort will be computed through Fermilab, but there is a possibility that some computation could occur at other WLCG sites; a project goal is to consider up to half of computing needs occurring off-site. Regular operational data from Mu2e will top 7 PB of raw data a year. All data will be stored at Fermilab, with back-up copies stored at partner sites in Europe (locations are being determined). [Section 6.6], [Section 8]
- Mu2e expects that many institutions will request copies of analysis-format datasets and that these will be stored locally at each institution. At this time, it is estimated that one copy of the analysis-format dataset for the full experiment will be, at most, on the scale of 50 TB. [Section 6.6]

- Belle II was an active participant in the LHC's DC24 activities, and showed expected results for operations during this period; observation showed the experiment reached 250 TB/day, which is more than six times the maximum throughput expected at the end of the decade. A primary reason was testing of new software components shared with the LHC experiments. At the current time, Belle II does not anticipate participating in DC27. [Section 6.7], [Section 8]
- Two copies of the raw data coming out of the Belle II DAQ at KEK are permanently stored: one copy at KEK and a second copy distributed between BNL (30%), Canada (15%), France (15%), Germany (20%), and Italy (20%). All raw data centers are connected to LHCONE. BNL houses approximately 3 PB of Belle II data, and anticipates having 5 PB of tape and 5 PB of disk storage available approaching 2028. [Section 6.7]
- The SBN detectors (ICARUS and SBND) will produce approximately 5 PB of data per year that will be stored at Fermilab, and at backups sites. The SBN ICARUS detector has been taking data since 2021. The SBND detector is completing the commissioning phase and is expected to take physics-quality data in fall 2024. [Section 6.8]
- ProtoDUNE efforts have produced approximately 5 PB of data per year at CERN during the initial run, and will exhibit a similar pattern for a second run between 2025 and 2026. [Section 6.8]
- DUNE will produce 45 PB of data per year at full operation. This includes the near and far detectors, and 10 PB of processed data. The first full DUNE Far Detector, FD-VD, is slated to begin full operation at SURF in Lead, South Dakota, in mid-2029. A second module will come online in 2030–2031. These systems will require significant testing and calibration. They are projected to produce 10–15 PB of data per year per module to start and will be capped at 30 PB/year when in full operation. This includes tests of supernova burst sensitivity that will produce around 320 TB over a few hours. Steady rates of 100 TB/day are anticipated. [Section 6.8]
- Transfer of the raw data produced from the DUNE detectors is likely to be the largest single network task for the experiment. It is anticipated that along with initial processing, annual reprocessing of all data accumulated to date will occur. In addition to beam and calibration data, the signal of neutrinos produced from a nearby supernova burst presents a significant challenge due to the large volume of data collected in a very short period (up to 600 TB in 100 seconds) and the time-critical nature of analyzing that data for optical astronomical observations. [Section 6.8]
- DUNE participated in the DC24 activities with the LHC sites, and learned a lot about their approach to software and data management. DUNE anticipates participating in future DCs as well, due to the overlap with other WLCG computing resources. DUNE will also conduct its own DC activities on an annual basis. Being able to handle production traffic from the detector, as well as a supernova activity, will be an area of focus. This part of the DC will rely on the connection between SURF and Fermilab when the network is provisioned (expected in 2026 or 2027). For the DC activities prior to this, BNL will currently serve as a stand-in for SURF. [Section 6.8], [Section 8]
- LHC analyses are done by small and large groups that share data at a Tier-2 center, at a local Tier-3, or on other local computing facilities. An "average" event can be processed 50–100 times by analysts. This means that LHC analysts are processing hundreds of PBs each year just to carry out the initial analysis steps, which process data created by the production system. [Section 6.9], [Section 6.10]

- Datasets in ATLAS are collections of files organized by category/workflow. These datasets are the fundamental units in ATLAS, fully integrated into the Rucio data management system, and used for workflow, transfers, and user analysis. Individual datasets vary greatly in size: [Section 6.9]
    - Raw datasets are in the range of 1 TB to about 250 TB.
    - Analysis Object Data (AOD) datasets are in the range of 1 GB to about 100 TB.
    - DAOD datasets are in the range of 1 GB to about several TB.
    - HITS datasets are in the order of several TB.
- The Data Challenges (the last being DC24) were incredibly useful for the ATLAS experiment as well as the entire LHC community. Several issues were identified, including with the FTS tool, which limited productivity because there was no way to convey network capacity information to help influence performance. Simply stated, poor performance of some transfers limited the deployment of more to better-connected sites. Additional work to move from Virtual Organization Membership Service (VOMS), which is based on x.509 certificates, to HTTP-based tokens is also underway. FTS (and Rucio) are developed at CERN, and have limited resources to implement fixes. The timing of the next DC is still being discussed, and is tentatively scheduled for 2027. [Section 6.9], [Section 8]
- The LHC provides data to CMS in roughly 10-hour data-taking periods with minimal downtime (typically a few hours) between periods. A roughly constant stream of data is therefore available during operation. Event rates of around 1 kHz of physics events are typical today and for Run 3, and 7.5 kHz is expected for HL-LHC. This translates into raw data volumes of: [Section 6.10]
    - Run 2: 45 PB written during four years (2015–2018).
    - Run 3: approximately 120 PB (2022–2024).
    - Run 4: 270 PB/yr (estimated in 2022).
- CMS has produced simulations with output that scales in size with the luminosity collected. CMS has about 70 PB (170 PB) of Run 3 (Run 2) MC simulation datasets stored on tape, representing around 75 billion (250 billion) MC events. [Section 6.10]
- A paper was published after DC24 with several findings. The LHC collaborators' DC24 activity revealed several weaknesses in the software stack, namely with the way that FTS performed globally when faced with bottlenecks at certain locations. The discussion in CMS about the use of alternative data formats, and the use of other R&D efforts such as Caching, may also influence how much bandwidth is needed between facilities, with the TA links being a main focus area. ESnet's TA strategy has not changed substantially since the previous HEP review, and will continue to add capacity (mainly through using more flexible spectrum agreements, and sharing resources with GÉANT across multiple connections). [Section 6.11], [Section 8]
- LHC Run 3 will be completed by the end of 2026, and the network traffic is expected to keep growing at the usual rate until then; the LHC community sees an annual increase of 20–30% of traffic in Large Hadron Collider Optical Private Network (LHCOPN) and LHCONE, the two networks dedicated to transfers between WLCG sites. The current networking capacities are typically 100 Gbps WAN links, with many sites already having resilience (multiple) 100 Gbps connections. This will need to evolve as the LHC community prepares for Run 4. The expected traffic during Run 4 beyond 2030, when the HL-LHC comes into operation, will significantly increase. ATLAS and CMS have estimated this increase in network traffic to be a factor of 10 compared with Run3. [Section 6.11]

- In preparation for HL-LHC, the LHC experiments have defined a calendar of "data challenges" to progressively test the capabilities of the WLCG software and hardware infrastructure to meet the requirements. Here is the way they have estimated the network bandwidth: [Section 6.11]
  - ATLAS & CMS T0 to T1 per experiment:
    - 350 PB raw data per year; average of 50 Gbps or 400 Gbps during LHC running time.
    - Another 100 Gbps estimated for prompt reconstruction data tiers (AOD, other derived output).
    - Estimated 1 Tbps for CMS and ATLAS summed.
- ALICE & LHCb T0 Export
  - 100 Gbps per experiment estimated from Run 3 rates.
- The CMS experiment anticipates the following annual data volume for each data type during HL-LHC operation: [Section 6.12]
  - Raw data: 250 PB.
  - AOD data: 210 PB.
  - MiniAOD data: 27 PB.
  - NanoAOD data: 0.6 PB.
- The ATLAS experiment anticipates the following annual data volume for each data type during HL-LHC operation: [Section 6.12]
  - Raw data: 168 PB.
  - AOD data: 200 PB.
  - DAOD data: 76 PB.
  - DAOD_PHYSLITE Data: 2.6 PB.
- The estimated data volume of SuperCDMS SNOLAB is 350 TB/year, including calibration data and simulated data. [Section 6.13]
- AMS produces ~40 TB of raw data per year, which are transmitted to CERN, and then reconstructed to 170 TB data per year available for analysis by collaboration members. The production is performed incrementally every six months. Reprocessing is performed only when changes to reconstruction software occur. Each version of full reproduction is called a "pass." AMS has acquired over 500 TB of raw data, from which 2 PB of pass 8 data have been produced and copied to collaborators' institutions for further analyses. [Section 6.14]
- AMS data analysis requires massive MC simulation, performed by computing centers on a regular time basis. Currently, 23 PB of MC data have been produced since the start of the project in 2011. [Section 6.14]
- AMS has more than 300K readout channels and 1,500 temperature sensors. More than 500 microprocessors are onboard, which reduce the data rate from 7 Gbps to 10 Mb/s. Data taken on board of the ISS by AMS consist of raw events, and these are logically grouped in sequences called runs. Events inside one run are numbered in sequence. Every quarter of the orbit, a scientific run starts. The duration of each run corresponds to 23 minutes, and every run contains on average 700k events. The average size of a raw event is 2 KB; thus yearly raw data yield is 40 TB. [Section 6.14]

- By the end of 2025, AMS will undergo a significant upgrade on the ISS; it will involve the addition of a new tracker layer at the top of the instrument. This advancement is essential for progressing dark matter searches and improving precision measurements. The new configuration of AMS will require a complete overhaul of simulations and data analysis tools, which will be developed over the next three to five years. The total volume of data is expected to increase to 50 PB over the next five years. [Section 6.14]

4.3 Scientific Workflow

- Cosmological simulation data streams typically involve moving between 10–100 TB of bulk data, when an entire data set is needed. Smaller downloads that may target a subset of a data set for individual analysis may involve <1 TB chunks of data. Progress in remote visualization methods is allowing real-time visualization of many large data sets. [Section 6.1]
- Over the next two to five years, an evolution in cosmological simulation workflow tools to manage local and remote simulation data analysis will occur. This effort to design more approachable ways to manage download sizes, and ease of location, will allow users more power over their analysis behaviors. [Section 6.1]
- Every ~15 minutes, DESI obtains spectra of 5,000 astronomical objects, resulting in 1–2 GB of raw data per exposure. Data are then processed at NERSC in semi-real time (minutes to hours after capture), which results in a processed file size of ~10 GB per exposure. DESI projects that the total data volume will grow to O(10 PB) by the end of the survey. Every 1–2 years, DESI reprocesses all the data taken to date using the latest data-processing code and calibrations. [Section 6.2]
- Science analyses using DESI data are primarily performed at NERSC. The scale of analyses varies from single-core Jupyter Notebook use cases, to large parallel-batch jobs consuming many cores and machines. Many of the analyses make use of simulated datasets generated at NERSC, while others are generated at other DOE HPC facilities and transferred to NERSC. [Section 6.2]
- Transfers of DESI's processed data from NERSC to the NOIRLab and University of Utah mirrors use Globus for the initial bulk transfer, then use rsync to pick up any later differences. [Section 6.2]
- DESI experiences friction with the management of data mobility outside of NERSC. Data can be egressed via approaches like Globus, but this typically is an option only for users who are familiar with the tool and have the necessary credentials to manage it. An efficient method like commercial tools (e.g., Dropbox, iPhoto, etc.) may appeal to users less sophisticated with data mobility. Users often do not want to be bothered with the functional details of a data mobility tool and just want things to work efficiently and simply. That DESI would adopt some of the other tools the HEP community has championed (e.g., Rucio) is unlikely, as the cultural shift required to do so is not very strong. [Section 6.2], [Section 8]
- Rubin alerts are required to be broadcast via at least five community-developed brokers in 60 seconds from the time the camera shutter closes on the mountain top in Chile. Other "prompt data products" are made available to the community after an 80-hour embargo period. The embargo period is implemented to prevent prompt detection of artificial satellites in the LSST data stream. In 2021, Rubin completed selection of the community alert brokers. These include seven international teams: ANTARES (NOIRLab), ALeRCE (Chile), Lasair (UK), Pitt-Google (University of Pittsburgh), Fink (France), AMPEL (Germany), and Babamul (California Institute of Technology). Two downstream brokers that operate off one of the primary brokers were also endorsed by the Science Advisory Committee (SAC): SNAPS (Northern Arizona University) and POI (Vanderbilt University). [Section 6.3]

- All image data will be carried to the project's Archive Center at the USDF, hosted within the S3DF. The total northbound data from Rubin to the Archive Center is approximately 30 TB/night. A very demanding application using these flows is Data Release Processing and Alert Production, where 12.7 GB (uncompressed) of pixel data need to be transferred from the Base Facility in La Serena Chile to the Archive Center in seven-second bursts. Two such bursts occur approximately every minute of nighttime observing. The bursts need to be sustained for approximately 355 days each year. The required reliability level for these flows is very high and is seen to require a path-diverse network with nearly lossless networking. [Section 6.3]
- Over its 10 years of planned operations, Rubin Observatory will produce annual data releases. Each data release will include reprocessing of all data from the start of the survey up to the cutoff date for that release. [Section 6.3]
- The DESC project is investigating the use of Rucio and FTS, as is the Rubin Observatory, to see if these would be a good fit for the data management aspects of the project. The use of any software comes with risks: the project strives to ensure that individual users can easily interact with the systems they create, and will gravitate toward things that are easy to install, use, and support long term. The latter observation holds true for managing access for scientific users: the lack of a cohesive multi-institutional way to manage credentials is also a challenge for this collaboration. [Section 6.4], [Section 8]
- DESC is exploring ways to perform multimodal analysis using other streams of research data. Some possibilities include the European Space Agency's Euclid mission[6] and the Nancy Grace Roman Space Telescope.[7] The effort is early, and questions remain about ways to get access to the data and how the data can be integrated through software and computational hardware. One question is that with the portal system for these three data products (and in general, all individual collaborations that produce a data set that others may want to use), no uniformity exists in how the data can be retrieved. This causes a challenge for the greater scientific community. [Section 6.4], [Section 8]
- DESC is exploring the use of Independent Data Access Centers (IDACs) for the Rubin Observatory, which are being developed as in-kind contributions to the observatory by institutions around the world. Some of the IDAC resources will be made available to DESC and other LSST science collaborations for science analyses and could potentially provide specialized capabilities. [Section 6.4]
- When CMB-S4 is operating, each day of data will be immediately reduced to a map of the observed sky, using US agency computing resources (DOE HPC, NSF High-Throughput Computing [HTC], and potentially FABRIC). These maps will be analyzed to identify transient sources, with alerts issued to the entire scientific community via the Community Alert Broker, and for data quality and telescope health checks. [Section 6.5]
- Given the size and complexity of the CMB-S4 data set (~100 PB), collecting, transporting, and reducing the raw data from time streams of detector samples to well-characterized sky maps is the responsibility of the CMB-S4 project. Given the extraordinary range of science these maps support and their relatively small size (~100 TB), analyzing these maps is the responsibility of the CMB-S4 collaboration and the scientific community. [Section 6.5]

6 https://www.esa.int/Science_Exploration/Space_Science/Euclid
7 https://roman.gsfc.nasa.gov/

- Current and near-term work in data movement to support CMB-S4 will focus on building a test stand for both data and metadata into the baseline data movement tool (Rucio). These tests will primarily involve local operations within NERSC: running Rucio as a service on the Spin container-based service and ingesting simulated data into the Community File System (CFS). CMB-S4 will also practice moving data between local filesystems using Rucio. By the end of this time frame, transfers between NERSC and ALCF, or another large computing center, will be tested to validate performance expectations. [Section 6.5]
- Mu2e will be adopting Rucio and other tools that are well-known to the HEP community to manage workflow and data mobility. The project is limited in exploration of alternative software and computational platforms due to funding and staff. Adopting new workflows is not something Mu2e has the time or resources to explore currently, but Mu2e has been able to make its simulation and analysis code work at NERSC and ALCF. During long shutdowns, or extended technical stops, it may be possible to explore other use cases. [Section 6.6], [Section 8]
- Simulations for Mu2e will be run on Fermigrid resources, OSG resources, at NERSC, and at the ALCF. The simulated events and reconstructed simulated events will be stored at Fermilab. Mu2e expects that many institutions will request copies of the analysis-format datasets derived from the simulated events; these datasets are likely to be about a few times the volume of the corresponding experimental datasets. [Section 6.6]
- All Mu2e data can be processed on Fermigrid, but for planning purposes it is assumed that about half will be processed on Fermigrid and half on opportunistic OSG. On average, Mu2e expects to process each event twice, leading to an estimate of 4 TB/year of outbound data and 2.5 TB/year of inbound data. Discussions have taken place that may lead to some calibration work being done in the UK or at the Italian Institute for Nuclear Physics (INFN), using resources provided by their funding agencies. For this exercise, Mu2e assumes that about 0.5 PB/year will be transferred to each location. [Section 6.6]
- Belle II has a grid-computing model with a three-level hierarchical structure ("tiers") of computing sites: raw data centers (Tier-1), regional data centers (Tier-2), and MC (simulated data) production centers (Tier-3). Many of the Tier-1 and Tier-2 sites in Belle II are part of the WLCG collaboration and shared with LHC experiments. This also includes the network connecting these sites, particularly the ones connected by LHCOne. The sites without LHCOne connection are small and can be used without LHCOne. [Section 6.7]
- Belle II reprocessing campaigns based on improved calibration run typically once per year or biyearly. Reprocessing will be every other year beginning 2025. [Section 6.7]
- Belle II chose to migrate from its bespoke data management solution to Rucio in January 2021. Since then, Rucio, with FTS as back-end transfer system, has managed all production and analysis data production on the grid. Rucio ensures the replication of raw data from KEK to the other data centers. Rucio also ensures the replication of the dataset output generated by the production system and makes a second copy of user or group data to other sites. Since the migration to Rucio in 2021, no network bottlenecks have been observed. [Section 6.7]
- In addition to detector data, DUNE will produce large quantities of simulation that in some cases will rival the size of the detector data. These data will be produced at distributed compute facilities and be cataloged and stored in storage elements around the globe. DUNE is currently using more than 20 compute facilities and more than 10 storage facilities to achieve its computing tasks. This number is expected to grow between now and late 2028, when the first far detector module will become operational. [Section 6.8]

- Three DUNE prototype detectors will be active over the next two years: ProtoDUNE-HD, ProtoDUNE-VD, and ND-LAr 2x2 demonstrator. The ProtoDUNE detectors are at CERN, while the demonstrator is at Fermilab. The raw data arrive from the ProtoDUNE detectors at a steady rate of approximately 1 TB/hr (~300 MB/s), and from the 2x2 detector at approximately 42 GB/hr (~15 MB/s). The ProtoDUNE data are delivered to CERN on site, and a second copy of the ProtoDUNE data is also transferred to Fermilab. The raw data are made available for immediate processing, which is accessed via XRootD to run on the WLCG. For the 2x2 demonstrator, the data are automatically stored at Fermilab, and then transferred in batches to a DUNE allocation at NERSC for processing. [Section 6.8]
- DUNE data acquisition for the first far detector module will commence post 2029, and the output of the module will be roughly 30 PB/year going to archival storage. Offline processing and data flows will look very similar to prototype operations but with slightly different endpoints. (e.g., instead of data coming from CERN, data will be created at SURF). There is no significant change in the software or process of science between the activities of the prototypes and the actual DUNE detectors by design. It is anticipated that the current design of the DUNE computing model implemented with the prototypes will scale and be sufficient to achieve the science goals of DUNE. [Section 6.8]
- For data management software, DUNE uses the Rucio replica management service, which has wide support in the high energy physics community. This is combined with the MetaCat data catalog which tracks the provenance of the data and can also interface with other databases, such as the run conditions database. Rucio can work with the FTS system, which in turn uses primarily https-based transfers to move the data sets to and from remote data sets. It can also use Globus as a transport mechanism, and DUNE expects to get data to and from the leadership facilities. There is a distributed global disk model in which current data sets are hosted at a set of 15 storage elements in the US and Europe. [Section 6.8]
- The ATLAS grid infrastructure consists of the Tier-0 computing site at CERN, 9 Tier-1s, 70 Tier-2s, and about 50 Tier-3 sites distributed worldwide. Any workflow can be executed at any location with memorandums of understanding (MOUs) established for the Tier-0, Tier-1, and Tier-2 sites specifically. Tape storage to store raw and AOD files is available at the Tier-0 and Tier-1 sites and a small subset of Tier-2 sites. ATLAS Tier-3 sites at most universities are usually smaller local central processing unit (CPU) clusters with several TBs of disk storage attached. At these sites individual end-user analysis happens without any production activity. [Section 6.9]
- Data from the ATLAS detector are calibrated, reconstructed, and automatically distributed all around the world by the ATLAS Data Management system. The ATLAS Production System then filters through these events and selects the ones needed for a particular type of analysis. This brings the data set down to a manageable size for someone doing an analysis on a laptop. All members of the ATLAS Collaboration have equal access possibilities to all ATLAS data, independently of their geographical location, thanks to the WLCG. [Section 6.9]
- The open-source software framework Rucio is used to organize, manage, and access the ATLAS data. The data are distributed across heterogeneous data centers at widely distributed locations worldwide. The Production and Distributed Analysis System (PanDA) ecosystem manages all workflows and workloads in ATLAS. PanDA uses the HTCondor system for job submission to most computing resources. [Section 6.9]
- The data rates that CMS originally estimated for Run 3 (i.e., in 2020, just prior to the run beginning) were smaller than what was witnessed by a factor of two to three times beyond the 2018 values, and four to six times beyond the 2017 values. This increase in size has not

translated directly into increased traffic on network connections; however, some of the R&E efforts to improve data mobility tools and workflow management have helped mitigate big spikes in traffic. [Section 6.10], [Section 8]

- CMS globally has seven Tier-1s, including Fermilab in the US. The collection of Tier-1s is responsible for operating a distributed archive of all centrally produced data, including a second copy of the raw. The copy of the raw across the Tier-1s is considered the "active copy"; i.e., the raw copy at CERN is considered strictly a backup. In addition, the Tier-1s provide computing resources to the experiment that are primarily used for centrally organized processing. The US CMS Tier-1 Center at Fermilab is the largest of CMS Tier-1 facilities, currently providing 40% of the total CMS Tier-1 capacity. CMS has roughly 50 Tier-2s globally. The collection of the Tier-2s is responsible for all data analysis and some simulations. Tier-2s also provide disk space to host data for analysis, and a moderate amount of disk space for staging in data for processing. [Section 6.10]
- CMS Tier-3 sites are typically university-hosted clusters smaller than Tier-2 sites and typically without dedicated support to ensure high resource availability. Those in the US are not funded through the US CMS operations program. Tier-3 sites can be part of the data management system of CMS and receive official datasets, or can access data through streaming. These sites can also provide disk space to hold user data as required by the local community. [Section 6.10]
- CMS uses compute resources beyond the tiered computing centers. These include HPC centers in the US. During 2023, the CMS usage of shared HPC facilities funded by DOE and NSF amounted to about 10% of the overall computing budget globally. This increased substantially in the recent past, by around four times in relative fraction of CMS compute since 2020 (and a substantially greater increase in absolute terms), and substantially increased again in 2024. CMS expects that this model will change for the HL-LHC to include larger allocations on HPC facilities. [Section 6.10]
- HPC centers funded by DOE and NSF as shared facilities across multiple science domains do not generally provide the full functionality required from an LHC Tier-2 or Tier-1 center (e.g., they normally do not provide any substantial amount of storage under the control of the experiment). Data access at HPC centers relies on either temporarily staging in input data on storage at the HPC site or streaming access, if the site setup allows access to the wider network. In general, HPC centers run only compute-dominated applications that have modest IO/CPU requirements (e.g., the central production workflows involved in central processing and simulations). Output is archived back to the Tier-1 sites, either directly from the application or through other mechanisms. [Section 6.9], [Section 6.10]
- The centrally produced data from CMS comes in multiple formats ranging from the most versatile and complete (raw and AOD) to the easiest, smallest, and fastest to use (MiniAOD and NanoAOD). CMS has been moving toward the use of NanoAOD, a data format designed for interactive end-user analysis, and produced centrally. CMS adoption of this format has already grown in Run 3 to around 50% of analyses using the NanoAOD as their primary data tier. [Section 6.10]
- CMS currently uses Rucio for bulk-data transfer activities, having fully switched from its previous system based on PhEDEx and Dynamo tools, to manage dataset storage and dataset transfers. Rucio has become a community solution for data management. Rucio uses the FTS to manage scheduling and file transfer. For bulk transfers via Rucio and executed by FTS, US CMS has moved to Third Party Copy (TPC) https implemented via XRootD servers and away from GridFTP. [Section 6.10]

- Failures in data streaming are a large source of job failures in CMS. This has also become an issue for analysis-facing activities, where a greater variety of tools are used to access CMS data. While CMS can instrument its centrally maintained software with retry mechanisms and other fallback logic, extending this to community-developed tools is more difficult. Reliable and high-capacity streaming of input data, either raw or pileup simulation, considerably reduces the disk requirements of CMS at HPC and other nondedicated computing facilities. [Section 6.10]
- For LHC operations, the use of either pledged or opportunistic HPC resources located away from traditional tiered sites is currently very difficult because each HPC facility is "special" in terms of bespoke hardware, network access, storage policies, and job submission. This means each experiment using such facilities needs to dedicate significant resources to operating and maintaining the HPC for its use case. HEP experiments have developed in-house mechanisms, like HEPCloud and Harvester, to manage access to disparate resources. Therefore, the LHC community would greatly benefit from a unified security model across DOE (and NSF) HPC facilities and a unified resource brokering API via efforts such as IRI. [Section 6.11]
- The LHC community maintains some common efforts to provide software used throughout the ecosystem: [Section 6.11]
    - Rucio, a distributed data management system.
    - FTS, responsible for reliably transferring data while not overloading end-systems.
    - Monitoring:
        - SciTags, to label flows and/or IPv6 packets by experiment and activity.
        - WLCG Site Network Monitoring to track each site's IN/OUT traffic.
    - Networking orchestration:
        - SENSE/Rucio to integrate SDN with distributed data management.
        - NOTED, a tool to identify network congestion and move large flows to other paths dynamically.
- All LHC Tier-1 and Tier-2 sites are expected to have perfSONAR nodes deployed to monitor network performance. These nodes are integrated into a mesh configuration, enabling regular tests such as latency and bandwidth checks between sites. perfSONAR allows network administrators to identify potential network issues that could impact data-transfer rates, both intersite and with CERN. It plays a vital role in the global transfer of data across the LHC Computing Grid (LCG). [Section 6.11]
- The exact computing model for the HL-LHC has not been finalized, given that the LHC Run 4 is not expected to start till 2030. It is assumed that the current computing models will be the baseline model, with some small improvements. No fixed hierarchy of computing sites will exist for most data processing and data access services. To improve usage efficiency, sites will be primarily categorized by size, service level, and capability. Hence the location of datasets and users is not deterministic. Tier-1s will need additional capacity to handle worldwide traffic, and Tier-2s will be scoped larger for national traffic needs. The network capacity should be provisioned to match the scale of the available resources at each site. [Section 6.12]

- ATLAS has chosen to evaluate two scenarios based on how aggressively it is anticipated that the HL-LHC storage and computing R&D program will deliver improvements: [Section 6.12]
    - "Conservative" R&D is a program where work can be completed with the existing person-power. The R&D activities currently underway to support Run 3 will be put into place, and will encourage an almost universal adoption of the DAOD_PHYSLITE format.
    - A full program of "aggressive" higher-risk R&D can be completed only with new person-power. ATLAS implements new developments that very significantly improve the speed and storage volumes of workflows that currently are heavy consumers of resources.
- SuperCDMS SNOLAB data will be acquired underground at SNOLAB, transferred to the surface, and then transferred to SLAC where they will be processed. To ensure access to data to all members of the SuperCDMS collaboration (access to SLAC resources is restricted for some collaborators), processed data will also be copied on the Open Storage Network (OSN) and the Digital Research Alliance of Canada (DRAC). [Section 6.13]
- SuperCDMS will not rely on SNOLAB for data analysis. Data analysis will happen primarily at SLAC, DRAC, the OSN, NSF ACCESS facilities, and at other smaller institutions. SNOLAB is minimally equipped with software, and each experiment has a dedicated subnetwork with limited bandwidth. Given the small amount of data to be moved from SNOLAB to SLAC, automated transfer tools based on rsync (single stream), with the option to move to bbcp (multiple streams, four by default), will be used. The limited outgoing bandwidth and the single destination make the use of such simple approaches ideal for this scenario. [Section 6.13]
- Due to limitations in the available bandwidth that Tracking and Data Relay Satellite (TDRS) can deliver, and the number of use cases that exist on the ISS, AMS is able to access only about 14 Mbps of space-to-ground transmission. This limitation has not forced the adoption of high-speed data migration tools, as the data volumes will always remain small for this aspect of the workflow. NASA has been evaluating whether the use of other satellite technologies (e.g., Starlink) could serve as an augmented communication mechanism, or a replacement for TDRS, but this would not be something AMS has control over. [Section 6.14], [Section 8]
- The AMS project creates many simulations during its process of science; over 20 PB have been produced since the start of the project. With the pending upgrade to AMS hardware in the coming years, it is anticipated that all produced MC will have to be reprocessed. [Section 6.14], [Section 8]
- Recorded data are packed to fixed length packets and transmitted via the TDRS satellites to ground at White Sands, New Mexico, and then relayed to NASA's Marshall Space Flight Center (MSFC) in Alabama. At MSFC data are first transferred to the AMS relay computers using the User Datagram Protocol (UDP) and finally to the AMS Science Operation Center (SOC) at CERN, in Geneva, Switzerland, with bbftp software. Sustained transferring speed from MSFC to SOC up to 50 Mbit/sec is achieved. [Section 6.14]
- AMS previously used Globus for cross-site data moving, but as CERN dropped Globus support, AMS now uses XRDCP to transfer data across sites. [Section 6.14]

4.4 Computational and Storage Requirements

- As increased computational power becomes available, large-volume cosmological simulations with gas dynamics and subgrid modeling of astrophysical mechanisms have become possible, which were previously the domain of gravity-only simulations. Increased computational power will enable large-scale ensemble simulations that can be used directly for inference or to construct generative models using AI/ML techniques that can be then used for this purpose.

The level of resolution and volume required of cosmological simulations will continue to increase in the coming years, driving data volumes significantly upwards due to these factors. [Section 6.1]

- Cosmological simulations can generate far more data than optical and microwave observations. This is expected to be the case when data become available from the Rubin Observatory, the largest of the imaging surveys. Estimation of statistical errors relies on hundreds to tens of thousands of simulated examples, and cosmologists often need to simulate results obtained from dozens to hundreds of varying simulation input conditions. While the on-sky pixel-level data can in some cases be larger, the simulated catalogs are already at the level of PBs for large simulation suites. [Section 6.1]
- The size, sophistication, and realism of simulated sky catalogs has improved significantly, which puts further stress on storage and data-transfer capabilities. Typical simulation products can be a TB in size, approaching a PB for some of the wide-ranging simulations or those that contain more detail due to improvements in the modeling techniques. [Section 6.1]
- Cosmological simulation continues to be substantially limited by the availability of storage, particularly when considering how to distribute the simulation data beyond the group that performed the simulation. Further, the need for storage beyond the initial computing project is critical; the lifetime of the need is often set by the life cycle of the analysis, which is happening in larger collaborations, and as the projects become more complex, over longer timelines. Some projects are more storage-limited than they are compute-limited. This issue has several associated dimensions to consider: [Section 6.1]
  - Sizes and numbers of simulations will increase.
  - Locations producing simulations are expected to increase.
  - Funding spans different agencies.
  - Usefulness over time, and to different communities, is increasing.
- The DESI collaboration space at NERSC will eventually host 10 PB of experimental data. These data are processed nightly, and analyzed by users that access NERSC from remote locations. Backups are performed to High-Performance Storage System (HPSS) at NERSC, and mirrored to partner facilities (NOIRLab in Tucson, AZ, and the University of Utah). [Section 6.2]
- Given the volume of data involved, making all Rubin releases available on fast storage for the duration of the project is not feasible. Instead, Rubin will keep the two most-recent releases available for rapid access, and archive older releases to tape. Users will not be able to perform database queries against these archived releases. However, they would be made available for bulk download. [Section 6.3]
- Rubin requires a constant 1,188 CPU cores throughout the survey period to perform prompt processing. These cores will be used during the night for alert generation, and during the day for catch-up (where necessary) and solar system processing. Each core will have access to 5 GB of memory, and all these cores are located at the USDF. [Section 6.3]
- In FY24 the Rubin Observatory IDF transitioned to the hybrid-cloud US Data Access Center (US DAC) model, with the user-facing Rubin Science Platform (RSP) remaining in Google Cloud and the back-end data holdings and data services hosted at the USDF at SLAC. The RSP is implemented in a virtualization framework that is portable across cloud providers or Kubernetes-based on-premises infrastructure. [Section 6.3]

- DESC analyses will be carried out primarily at NERSC, and SLAC is the primary data source for Rubin Observatory data. ESnet current connectivity between these sites is sufficient for DESC's early needs, and planned upgrades in future years should ensure the data volumes are managed efficiently. During this period, DESC foresees testing and carrying out workflows across multiple sites as supplementary computing, trying to minimize data movement. In the next two to five years, NERSC will be migrating to NERSC-10, and DESC will be testing the incorporation of IDACs and other IRI platforms to these analyses. [Section 6.4]
- The current computing needs of Mu2e are dominated by simulation campaigns to provide datasets of simulated events for the purpose of value engineering the design, validating the design, and providing input for the development of trigger, reconstruction, calibration, and analysis algorithms. At first the simulations were done on OSG, including Fermilab, but in recent years resources on Theta at ALCF and on Cori and Perlmutter at NERSC have been used. The output of each of these simulation campaigns is on the scale of 250 TB. [Section 6.6]
- Mu2e has had preliminary discussions with groups in Italy and in the UK to provide computing resources for the calibration of subsystems. If these plans are realized, outbound datasets will be on the scale of 0.5 PB/year to each location, and inbound datasets will accumulate to tens of TB/year. If additional computing resources are available in collaborating institutions, Mu2e plans to use them for simulation campaigns; in this case there will be little outbound traffic and inbound traffic of up to a few hundred TB/year. [Section 6.6]
- Belle II uses a grid model initially developed to support the computing needs of the LHC experiments. This model has a hierarchical structure with a Tier-0 site (e.g., KEK) that produces and hosts a primary copy of the raw data produced by the experiment Tier-1 sites that host a secondary copy of the raw data, and that are used for prompt processing and reprocessing. and Tier-2 sites mainly dedicated to MC production and user analysis. [Section 6.7]
- In addition to use of Fermilab central computing, SBN will onboard computing at the national center of INFN (CNAF), CERN, and the UK to further facilitate general production and analysis work. SBN also plans to make use of HPC facilities like Polaris at ALCF and Perlmutter at NERSC to further enable general production processing. [Section 6.8]
- For reconstruction processing of data, DUNE needs to have an amount of outbound network traffic from the disk elements equal to the output of the detectors so that processing can keep up with the production of raw data. Data need to be delivered to wherever that production is running: we expect this to be a combination of computing resources from around the world. DUNE expects data input/output rates to be tape/disk input/output (I/O) limited for the foreseeable future. [Section 6.8]
- For SBN's ICARUS reconstruction processing, computing jobs need to have data delivered at a consistent average rate of up to 2 TB/hr to satisfy demands during data production campaigns (~60 days in length). Data will be written back to Fermilab for storage at a rate of ~1 TB/hr. Data need to be delivered to wherever that production is running: SBN expects this to be largely Fermilab computing resources, but it will include some fraction of jobs to available OSG sites across the US and Europe, and to CNAF computing in Italy. For SBND, these numbers will peak at approximately 1 TB/hr for data reading and ~0.8 TB/hr for data write-back in this similar two-year period. [Section 6.8]
- The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. The mission of the WLCG is to provide global computing resources to store, distribute, and analyze the ~50 to 70 PB of data expected every year of operations from the LHC at CERN. WLCG is coordinated by CERN. It is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS, and LHCb) and the participating computer centers. [Section 6.9], [Section 6.10]

- ATLAS has a long history of successfully using HPC resources during Run 2 and Run 3 at the LHC. From 2016 to 2020, US-based HPC resources supported 10–25% of ATLAS simulation production. This share decreased after several HPC resources were commissioned in Europe. In the first half of 2024, 21% of the worldwide ATLAS CPU resources came from HPC resources with 70% of them from the European HPC Vega in Slovenia, 8% from Perlmutter at NERSC, and 5% from CSCS (the Swiss National Supercomputing Centre) in Switzerland. European HPC resources are mostly used through grid interfaces. US resources at NERSC and Texas Advanced Computing Center (TACC) require special edge services to run ATLAS jobs. Only simulations were carried out, since they are not data intensive. In the future, ATLAS plans to adapt all workloads to run on HPC resources. This will put much higher demands on networking. [Section 6.9]
- ATLAS computing is fully distributed. All processing and reprocessing of the data is managed by PanDA, and Rucio is used for data cataloging and data transfer. All computing activities are free to occur at any site, irrespective of their tier, and based on intelligent brokering of tasks and jobs. Distributed analysis jobs are also brokered by site capability; users are discouraged from choosing a specific site. Four distinct storage types are used by ATLAS computing environments: disk, tape, cloud, and cache storage. [Section 6.9]
    - Disk storage is the main bulk of the storage: 1042M files with the volume of 412 PB are in disk storage. In the US, 198M files with the volume of 98.5 PB are in disk storage.
    - Tape storage is used as a main data archival system: 370M files with the volume of 639 PB are stored in the tape archive system. In the US, 63M files with the volume of 107 PB are stored in the BNL Tier-1 HPSS system.
    - S3 type storage is used for temporary output of event service jobs.
    - XRootD XCache storage is used for unmanaged disk data cache for input of user analysis jobs.
- Changes made to CMS analysis formats have also had a large impact on network traffic. The NanoAOD format, which was introduced late in Run 2, is becoming more widely used. The format is smaller, and thus occupies less network bandwidth and can run faster on CPUs. More than 50% of all physics analysis activities are now based on NanoAOD, with another 40% or more to be based on MiniAOD. The use of the larger AOD format is now minimal, despite it occupying a lot of disk space. At future HL-LHC scales, CMS expects to keep AOD on tape archives only. [Section 6.10], [Section 8]
- The LHC computing and storage infrastructure is organized in a tiered computing model including four tiers, 0–3. Originally, when the connectivity among the sites was extremely limited, the tiers were hierarchical in that Tier-N+1 was connected only to an associated Tier-N above and associated Tier-N+2 below. Data were not allowed to flow from Tier-N+2 to Tier-N. This rigid concept has been eliminated over the last decade as network capacity and capability have progressed. [Section 6.10]
- CMS acknowledges that the data volume that can be handled by the networks within and coming out of the CMS detector facility far exceeds what can be handled offline within current CPU, storage, and networking infrastructures. This bottleneck will remain for the foreseeable future, as network growth patterns move faster than storage and computation. [Section 6.10]
- WLCG's hardware choices for the next major cycles will have a huge influence on the performance of the overall computation and storage system, as witnessed during the LHC collaborators' DC24 activity. Since funding requests must be made in advance, and hardware choice and prices fluctuate, uncertainty exists about what capabilities may exist. As a result

of this, a series of mini-DC activities will occur on a regular basis to evaluate aspects of the ecosystem: computing, storage, software, and networking, culminating in the larger complete DC activities that focus on major goal milestones. [Section 6.11], [Section 8]

- The impact of HL-LHC on ATLAS and CMS storage and compute resources is significant. The increase in luminosity not only generates significantly more data but also significantly more complex events, which require more processing to resolve. [Section 6.12]
- The expanded use of High-Performance Computing (HPC) will also have an impact on the compute resources (storage and networking). These HPC centers are increasing in computing power, and several exaflop-scale machines will be operational during the start of the HL-LHC. These centers are more capable of processing larger and larger volumes of data; this means the workloads that will run on these machines will evolve. HPCs will be capable of producing a large volume of simulated data. [Section 6.12]
- SuperCDMS SNOLAB's raw data, in compressed Maximum Integrated Data Acquisition System (MIDAS) format, will be processed at SLAC, and the reconstructed output data, in ROOT format, can be directly used for data analyses. Collaborators will access and analyze processed data stored at SLAC using local computing resources. The data will also be transferred from SLAC to Fermilab, where they will be stored on tape for additional backup. Simulated data are generated at Texas A&M, transferred to SLAC, and then transferred to Fermilab for tape backup and to other sites as needed. [Section 6.13]
- Currently, a new data analysis and computing center is being developed at the University of Hawaii at Mānoa to support AMS. This will be the first computing and analysis center on US soil dedicated exclusively to AMS data. About 20,000 CPU cores and 10 PB of data storage will be employed for AMS data production, analysis, MC simulations, and archiving. [Section 6.14]

4.5 Remote Collaboration and Operational Requirements

- Removing the need to worry about data locality allows researchers focused on developing cosmological simulation to invest more time on the science, versus spending energy in activities to stage data and allocate technology resources. It also prevents half-measures being used today to create mechanisms to perform analysis on partial data sets, or adopt streaming approaches that are not always successful on some of the larger computational resources not designed to support that mode of operation. [Section 6.1], [Section 8]
- Data movement of cosmological simulation data between facilities is not a critical problem, but broad access for the user community remains challenging. Simplifying access with updated portal software, or tools like Globus, has improved data mobility. The next step is to combine this with dynamic access to computing resources. [Section 6.1]
- The ratio of the cosmological simulation user community to the size of the simulation groups will increase as the role of simulations advances; dataflow patterns will likely become more diverse, and pressure will increase on data-transfer facilities that serve the resulting simulation outputs. Although the main focus is the continental US, the geographical locations of the endpoints are worldwide. The diversity of endpoints and the number of geographical locations is expected to grow significantly. [Section 6.1]
- DESI computing is done using interactive computing methods at NERSC, namely Jupyter Notebooks. DESI has considered moving to more widely available computing options not tied to a single facility, such as a cloud platform (like those that the Rubin Observatory has adopted); a core concern lies in the user community, however, who are slow to adopt new approaches to computing and are making do with legacy tools for a lot of the analysis process. Investment into a new system halfway through an experiment is not feasible. [Section 6.2], [Section 8]

- The primary data access mode for DESI is for collaborators to login to NERSC directly, and perform work at the location of the data. Collaborators can download a subset of data for processing at local institutions for analysis. Public (non-collaborator) access is only through downloads from NERSC, or through Jupyter servers at the NOIRLab Astro Data Lab, which host a subset of the public DESI data. Many collaborators have expressed a preference for downloading subsets of data to use locally; however, DESI lacks developed tools for easy management of that remote analysis workflow. This is currently achieved through ad hoc downloads with Globus, scp, or https. [Section 6.2]
- The USDF will provide raw and processed data to Rubin Observatory authorized users, including DESC, and other science collaborations investigating dark energy in other participating countries. The USDF will distribute data to one or more data access centers in Europe and elsewhere. The scope of these centers and their reliance on the USDF is currently being worked out. It is worth noting that the data volume of an annual release is very significant, with just the catalogs of detected objects beginning at PB-scale. [Section 6.3]
- All the Rubin collaborations, as well as independent scientists, will access and analyze data through either the hybrid cloud-based US DAC (cloud-based RSP and USDF back-end data services) or IDACs operated by in-kind contributor communities. [Section 6.3]
- The RSP is a set of web applications and services that will be used by the data rights community to access, visualize, interact with, and analyze Rubin Observatory data. The RSP is being designed to be simple enough to engage general users and flexible enough to meet the needs of experienced users. Data will be exposed to users through the web portal (interactive use cases), the Jupyter Notebook interface (custom user use cases), and machine-accessible application program interfaces (APIs; remote tools/analyses use cases). [Section 6.3]
- DESC's current analysis model involves bringing collaborators to deploy their analyses at NERSC to run at scale and access collaboration data products. DESC is planning to expand these capabilities and has had successful tests and large campaigns at the UK and France data facilities for Rubin, though details for securing these resources for DESC science purposes are being settled. These resources could potentially be used for DESC workflows that require Rubin imaging products available at these sites. [Section 6.4]
- DESC transfers have largely been via HTTP, rsync, and Globus for data sets moving from NERSC to and from collaboration sites. DESC will explore in this period the use of Rucio over FTS following the steps of the Rubin project for SLAC-to-NERSC transfers, and possibly for its own purposes in interactions with IDACs. DESC is currently using Parsl for workflow management of image reprocessing tasks as well as certain simulation and analysis workflows. [Section 6.4]
- CMS applications and the infrastructure that distributes and launches them support multiple types of data access: [Section 6.10]
    - Traditional top-down data placement at Tier-1 and Tier-2 centers combined with applications specifying the dataset they need, and being automatically routed and executed at the sites that have it.
    - CMS supports streaming data access to any data on disk across the WLCG facilities from any location with an internet connection at any time. This is called AAA, for "any data, anytime, anywhere" within CMS.
    - Bottom-up data placement, as is implicit in caching. Here the applications are routed to sites with caches, applications access the cache locally, and cache misses are handled by the CMS XRootD data federation.

4.6 Multifacility Computational Workflows

- The data products generated by simulation groups can be created at any number of DOE-or NSF-funded HPC facilities, or other institutions, with sufficient resources to execute and store simulation codes. These data products are used "downstream" by many users, who are members of science working groups in cosmological surveys and experiments. A common use pattern is to build a mock survey from the simulation data, then perform a clustering analysis on that survey, potentially using multiple facilities for computation and storage. [Section 6.1]
- DESI data from KPNO are transferred in 10-minute intervals to NERSC using rsync. This rsync tool is not known for high performance; however, the cadence of the experiment dictates that nightly data collection is less than 100 GB and must reach NERSC via network that is limited to 1 Gbps between the telescopes at KPNO and the base station; use of more advanced tools will not improve the overall process of science. [Section 6.2]
- DESI is fundamentally a multisite experiment, with data from a telescope in Arizona (KPNO), a primary computing center in California (NERSC), and a worldwide collaboration of scientists. DESI is currently limited in its ability to utilize non-NERSC DOE computing facilities for a variety of reasons, including: [Section 6.2]
    - The ability to manage accounts for hundreds of users across different sites .
    - Creating a system to synchronize PB-scale datasets of millions of files between multiple sites.
- Nightly data from the Rubin Observatory are accumulated and processed with science pipelines at the USDF as well as partner data facilities in France and the UK, producing deep stacked images and source catalogs. These "annual data-release" products are served to the user community via an RSP hosted in Google Cloud and backed by data-access services hosted at the USDF. Rubin expects a significant amount of data will be moved in bulk to other analysis facilities. Data will be moved in conjunction with the distributed data-release processing efforts. [Section 6.3]
- Rubin includes a program of in-kind contributions from international participant communities, some of which take the form of IDACs. Large amounts of data will be moved to other computing facilities to support custom analysis by large science collaborations (e.g., DESC); this will likely amount to the mirroring of a significant fraction of each annual data release from SLAC to NERSC. Additional mirroring efforts to support science analysis may be initiated for centers such as the ALCF and the TACC. Rubin anticipates a key role for ESnet in supporting these data transfers. [Section 6.3]
- DESC's primary data source is the Rubin USDF at SLAC. Workflows will primarily involve heterogeneous analysis operations over either the processed data products from the alert stream of transient events, or the data set extracted from the accumulated images and catalogs. Nominally, most of DESC's computing will take place at NERSC. As the survey proceeds, DESC will likely need to obtain additional resources elsewhere through specific programs or through international collaborations. [Section 6.4]
- The anticipated start of operations at the Rubin Observatory is FY25Q4. In this period DESC expects to bulk download from the primary data source (USDF at SLAC) to NERSC the released datasets corresponding to the commissioning and science verification periods. The specifics of the data set sizes, files, etc. are uncertain due to the nature of these pre-operations' activities and final scope. DESC expects the resulting data size will be of a similar size as Rubin's recent data preview, and of the order of 10 TB of catalogs, with the possibility of transferring up to 1–2 PB of imaging data. DESC's main computing resource will continue to be NERSC. [Section 6.4]

- CMB-S4 raw data will be archived at NERSC and ALCF, and a subset of recent data will be stored on disk. Daily data-quality statistics and maps will be produced at NERSC, ALCF, NSF ACCESS centers, and possibly FABRIC. Bulk reprocessing will be performed at NERSC, ALCF, and NSF ACCESS centers. Simulated data will be generated at NERSC and ALCF and typically reduced to maps to avoid I/O and storage overheads. In both cases, the resulting maps will be stored on disk and archived at NERSC and ALCF. Map-level data will be analyzed at NERSC and on local compute resources across the collaboration. Published data (primarily maps and derived scientific data) will be made available to the scientific community at NERSC and Legacy Archive for Microwave Background Data Analysis (LAMBDA) located at the NASA Goddard Space Flight Center (GSFC). The raw data and the software used to characterize and reduce them will also be made available to the scientific community at NERSC. [Section 6.5]
- During the CMB-S4 construction phase, all the significant data will be simulated. Simulations will be run at NERSC or ALCF, and the data copied between them, peaking at O(10) PB in 2029. Reductions to maps will be run at NERSC, ALCF, NSF ACCESS centers, and possibly FABRIC. Maps will be made available to the collaboration at NERSC, with the bulk of the analyses performed there. Small amounts of data may be moved to collaborators' local resources for analysis. [Section 6.5]
- Mu2e will utilize most of the existing computational and storage resources at Fermilab during production, but has used several DOE HPC facilities (i.e., ALCF and NERSC in particular) during simulation construction. Currently the project produces on the order of hundreds of TB of simulation output per year using OSG-connected resources at Fermilab, the HPC facilities, and other collaboration sites. [Section 6.6], [Section 8]
- Mu2e expects to use resources at several DOE-supported sites. At Fermilab Mu2e will use local interactive machines provided by the lab, Fermigrid, the Fermilab Elastic Analysis Facility, and the data storage resources at Fermilab. Mu2e has demonstrated the ability to use big CPU cycles at HPC centers, and does not foresee a need for significant GPU resources currently. [Section 6.6]
- SBN currently uses the Polaris supercomputer facility at ANL in addition to computing resources at Fermilab. Globus-based transfers of SBND neutrino data to ANL are up and running, and sustained access to Polaris over the next few years for ongoing data processing will be needed to ensure that physics results can be obtained with maximum efficiency. SBN is also investigating whether other DOE SC user facilities, such as NERSC, may provide unique advantages that could complement the present workflows mentioned previously. [Section 6.8]
- DUNE expects to incorporate the use of HPC centers, like Perlmutter at NERSC, for performing simulation and data reconstruction general processing on the demonstrator data, as well as for executing dedicated AI/ML processing workflows. [Section 6.8]
- During the DUNE Prototyping phase, prototype detectors running at CERN and Fermilab send their data to DUNE storage sites worldwide. As more and more services come online at the DUNE Far site (SURF), data will begin to flow from SURF in South Dakota to Fermilab and onward to DUNE collaboration sites worldwide for additional processing. When the first far detector module is operational in mid-2029, supernova data will need to be transferred in real-time from SURF to US HPCs for fast turnaround processing. [Section 6.8]
- The primary data-related resource constraint for DUNE deals with the supernova use case; the data from South Dakota must be transmitted to DOE HPC facilities in a matter of hours, and then be in a place where they can be easily accessed by the job processing software. Once the first far detector of DUNE is online, currently estimated to be 2029, supernova triggers are expected regularly, albeit some may be false alarms; the estimate is currently one per month.

DUNE will require a set of machines fast enough to source and sink the data, and a distributed file system to make it accessible to the computational jobs. [Section 6.8]

- The US ATLAS Tier-1 is hosted at BNL's Scientific Data and Computing Center (SDDC). The ATLAS connection to ESnet is shared with other programs hosted at the SDCC. The US Tier-1 is the largest of the ATLAS experiment: it represents about 25% of the computing resources of ATLAS at Tier-1s. [Section 6.9]
- Four ATLAS Tier-2 centers are in the US: Northeast Tier-2 (NET2), Great Lakes Tier-2 (AGLT2), Midwest Tier-2 (MWT2), and Southwest Tier-2 (SWT2). These centers are used for all distributed production and user analysis workloads. Each Tier-2 center consists of multiple university-based clusters. In 2023 to 2024 ATLAS also added a new Tier-2 outside the US that contributes to the US ATLAS cloud. This site (TW-FTT) is derived from the former Tier-1 site in Taiwan. ESnet has helped to provide connectivity for this new site into LHCONE. However, this site has limited bandwidth (3 Gbps) and is only running tasks that are not data intensive. [Section 6.9]
- Two ATLAS Analysis Facilities (AFs) are currently located at DOE national laboratories (BNL and SLAC). A third AF has been set up and is operational at University of Chicago. The data accessing model at these AFs includes both Rucio-managed data transfers to the AFs and using XCache to access ATLAS data remotely. [Section 6.9]
- Having reliable mechanisms to transfer data into HPC facilities from the CMS Tier-1 and Tier-2 centers is of growing importance. HPCs are different from other resources used by CMS primarily because they provide significant compute (CPU and GPU) resources but very limited disk resources, and because network access to or from HPC compute nodes is typically limited or not allowed. The broader LHC community has considered how to best utilize these resources within operational constraints. Unlike most HPC users, CMS cannot concentrate its efforts on using one or a handful of HPC facilities, but is instead asked to use HPC facilities worldwide. Network access and data management are important aspects of this. [Section 6.10]

4.7 Networking for Local and Wide-Area Data Mobility

- The network speeds between major DOE HPC centers are not a limitation for cosmological simulation, and transferring PB-sized data sets is effectively routine. The data set sizes are expected to increase in the coming years, but will be within the capabilities of ESnet and the DOE HPC resources. [Section 6.1]
- DESI's infrastructure relies on two major components linked via network infrastructure: KPNO and NERSC. A number of other instruments are at the KPNO facility, and all share a common 1 Gbps-capable network connection from the summit to a base station. A larger 10 Gbps network then links the base station to the rest of the R&E community. Each of these connections has a different operator, and each has experienced trouble in the past. Natural disasters (e.g., wildfire) have severed connectivity, and bandwidth limitations imply that all scientific operations must share the common connection fairly. This can limit productivity, given there is no back-up option, which can impact real-time requirements. DESI has ways to buffer data, but is otherwise stalled without alternative paths. LEO satellites have been used for some back-up capability in the past, but this is not a complete solution to support DESI. [Section 6.2], [Section 8]
- International networking has a role in supporting Rubin Observatory production processes. Annual data releases will be computed at the USDF, and at the satellite computing facilities in France and the UK. This generates a need for networking to support continuous data production, in addition to distribution of data for scientific data analysis. By agreement, CC-IN2P3 is the responsible institution for arranging connectivity between CC-IN2P3 and the US. [Section 6.3]

- A memorandum of agreement (MOA) has been established between IN2P3, the LSST Corporation (LSSTC), the Rubin Observatory, and the University of Illinois for the IN2P3 Computing Center (CC-IN2P3) in Lyon, France to provide 40% of the computing capacity needed for annual data-release processing during LSST operations and to host a complete backup of the Rubin Observatory data archive. The MOA requires that France will provide the TA network connectivity to ESnet needed to support data transport to and from CC-IN2P3. [Section 6.3]
- The Rubin Observatory UK Data Facility (UKDF) will be a distributed entity. Its computing hardware will be provisioned through Institute for Research and Innovation in Software for High Energy Physics (IRIS), a national organization providing computing resources for projects funded by the Science and Technology Facilities Council (STFC). RIS sites are well integrated into national and international networks, for the transfer of LHC and other data, and have a wealth of experience supporting a wide range of international science collaborations. The UKDF will be hosted by the University of Edinburgh. [Section 6.3]
- The required peak bandwidth from the Rubin Observatory Base Facility in La Serena to the USDF is established by the need to transfer the images for prompt processing. For the USDF to release alerts of transient and variable sources within 60 seconds of image readout from the camera electronics, seven seconds of this time budget is allocated to data transfer, including compression/decompression. Two seconds of that budget is in the Summit to Base Facility transfer, and the other five seconds is in the Base Facility to Archive Transfer and compression/decompression. This equates to a minimum of 40 Gbps bandwidth. [Section 6.3]
- DESC's major data movement activity, occurring on a yearly basis, is a bulk data download from SLAC to NERSC. The network infrastructure in place between the facilities (via ESnet) is sufficient to handle the expected volume of data, and SLAC is in parallel moving to larger 400 Gbps capabilities to support other efforts. Since this is not a real-time operation, there is no time limit (or performance expectation) for transfer. [Section 6.4], [Section 8]
- DESC is open to the idea of working on structured data challenges, as other communities have done in the past. This could involve simulation data to start, and characterize performance between SLAC, NERSC, and the IDACs. Poor performance to some sites in the EU from NERSC has been observed, but given this is not a part of the critical path of science, it is a lower priority for DESC to address. [Section 6.4], [Section 8]
- CMB-S4 will leverage networking resources that exist in Chile today. The Simons Observatory (SO)[8] and Atacama Large Millimeter/submillimeter Array (ALMA)[9] invested heavily into the networking infrastructure over prior years, and have a clean fast path to national and international computing resources. Networking to support experimentation includes multiple strands of fiber optic cables, as well as a perfSONAR monitoring infrastructure. A 10 Gbps dedicated fiber link from the site, to the ALMA, is being installed by Simons Observatory currently and already receives support from ESnet to ensure global connectivity. ESnet will continue to work with CMB-S4 on networking and the overlap with the existing projects. [Section 6.5], [Section 8]
- Mu2e networking needs are relatively straightforward; a maximum of 10 Gbps may be required to handle intensive times of activity to support external computing needs. Given that most of the computation and storage is allocated to Fermilab, and connectivity to other WLCG sites will follow the same patterns as the LHC computation, additional networking support is not currently required. [Section 6.6], [Section 8]

8 https://simonsobservatory.org/
9 https://www.almaobservatory.org/en/about-alma/

- Belle II, hosted by KEK Japan, relies on a similar software infrastructure and grid-computing model to that of the LHC experiments, and has converted to use Rucio as a data management platform. Because of this sharing of infrastructure, networking and data movement work well for the current volumes. International connectivity is provided by partners Science Information NETwork (SINET)[10] and TransPAC[11], via multiple direct and indirect paths. Belle II also participates in the LHCONE network to enable consistent performance between facilities. [Section 6.7], [Section 8]
- The BNL network perimeter has two diverse 800 Gbps (2 x 400 Gbps) circuits that peer with ESnet routers. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits. The BNL Network Perimeter on average ingests and egresses between 10 and 15 PB of data monthly. [Section 6.7]
- DUNE relies heavily on Fermilab networking for a variety of tasks, including data archiving, replication, and batch processing. Two geographically diverse 100 Gbps links between the far detector complex at SURF and Fermilab are in the process of being acquired to provide WAN connectivity for the far detector. The target deployment date for these links is 2027. DUNE is approved for use of LHCONE, and has been making use of LHCONE and LHCOPN at certain grid sites that also have an LHC presence. [Section 6.8]
- ATLAS data volumes can average 5 PB a day, with volumes exceeding this during busy periods when the LHC is in full operation. The Tier-1 and Tier-2s within the US can see data flows of tens to hundreds of Gbps on a routine basis. Most are connected via 100 Gbps or 400 Gbps diverse connections to ESnet, or regional networking providers. BNL (the US ATLAS Tier-1) sees most of this traffic coming directly from CERN, then distributing outward toward other Tier-2s at US universities. All US Tier-2 sites are strongly encouraged to provide at least 100 Gbps connectivity, preferably with resilience. The intent is to have multiple 400 Gbps connectivity for each Tier-2 by the time HL-LHC starts. [Section 6.9], [Section 8]
- ESnet will investigate ways to integrate resources into LHCONE at non-LHC-affiliated sites. DOE HPC centers (e.g., NERSC) and AFs (e.g., SLAC) could benefit from being connected to better identify and prioritize traffic. [Section 6.9], [Section 8]
- Fermilab's WAN architecture utilizes LHCOPN connectivity to support movement of raw data from the T0 (CERN), as well as production data movement with some of the other CMS T1s. In terms of aggregate WAN capacity out of the site, Fermilab currently has provisioned 2.4 Tb/s of link capacity to ESnet, via geographically diverse connections. A link aggregation group (LAG) of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports Fermilab's general internet traffic, which includes some portion of science traffic. [Section 6.10]
- Much of the LAN and WAN networking infrastructure to and within a CMS Tier-2 site has been an institutional contribution to the US CMS Operations Program, i.e., it has not for the most part been directly funded by the program. All CMS Tier-2s connect to the LHCONE at least 100 Gbps, through partnership with regional networking infrastructure. In some cases, a Tier-2 facility may have multiple connections either as a backup, or for active use. [Section 6.10]
- The LHC community is currently a major user of the ESnet TA network links. Activities including the archival of raw data, initial copies of analysis data samples, and user-derived data sets using CERN or other non-US facilities are clear use cases for TA network usage. LHC tools do not prioritize site proximity (in the networking sense) when scheduling data transfers. Streaming data across the TA link is allowed (if discouraged). [Section 6.9], [Section 6.10]

10 https://www.sinet.ad.jp/en/
11 https://internationalnetworks.iu.edu/projects/TransPAC/index.html

- Networking has been fundamental to the success of ATLAS and CMS computing to date, enabling the exploitation of globally distributed resources for computationally limited science. This will remain the case to meet the budget-constrained computing challenges of HL-LHC. Strategies for HL-LHC computing are based on extensive use of powerful networks to reduce data replication by streaming over the network, and consolidating distributed resources into cohesive virtual federations such as data lakes. HL-LHC data and processing scales are large enough that optimizing for efficient bandwidth use will be essential. The ability to mark network packets to identify sources of traffic, and to route traffic to control speed and cost, may become vital. [Section 6.12]
- As a rough guide as to the network bandwidth that facilities in US CMS will require in the HL-LHC era for the purpose of reconstructing the raw data and producing corresponding simulation datasets, one can take the following estimate. From the 250 PB of raw data collected per year, about 40% (e.g., 100 PB) is expected to be reconstructed in the US. That these large-scale reconstruction campaigns take approximately 10 weeks is desired, implying a data input rate to processing facilities of 10 PB/week, and an output rate of 4 PB/week. In parallel, simulation production campaigns will be executed continuously throughout the year, producing another 40% of the total 240 PB of output data (e.g., 96 PB) to be persisted on tape and/or disk storage within the US. In total, these workflows will require a sustained rate of up to 130 Gbps between processing and storage facilities. To account for burst traffic, and to support data analysis workloads at AFs discussed later, it is desired that all major processing facilities, namely the Fermilab Tier-1 and US CMS Tier-2 centers, are provisioned to 400 Gbps. [Section 6.12]
- At present, it is assumed that the general LHC Tier-2 network requirements as specified previously will be sufficient to support AFs at the Tier-2s. A user at an AF may expect to process a year's worth of MiniAOD data and simulation (27 PB) in a timescale of a week. If the data are not cached at the AF, that amounts to a sustained 400 Gbps network load from the data origin site. In practice, it is unlikely a user would read the entirety of the data catalog, though the exact fraction will vary significantly depending on the analysis goals. [Section 6.12]
- Global R&E network use will have different characteristics on the timescale of the HL-LHC, compared to the situation the WLCG experienced in Runs 1, 2, and 3. Specifically, the experiments foresee other science domains (astronomy, biology, engineering, and others) becoming global network users at scales equaling or exceeding the LHC users. In such an environment, it is not wise to assume the experiments can continue to treat the network as an infinite resource. The LHC community must explore options for effectively operating within a bandwidth constrained environment. [Section 6.12]
- The network link that supports SNOLAB from the underground portion to the surface is made of 2 x 10 Gbps aggregate links shared by about 10 experiments. Only ~0.2% is expected to be needed by SuperCDMS during peak usage. From the surface, the data are copied to SLAC through the SNOLAB network link to the Ontario Research and Innovation Optical Network (ORION) of 1 Gbps. From there, ESnet routes become available. SNOLAB is looking into upgrading the outgoing bandwidth to 10 Gbps. [Section 6.13]
- NERSC features 1 Tbps aggregate bandwidth to ESNET via 2 x 400 Gbps links and 2 x 100 Gbps links. NERSC will be expanding to a 4 x 400 Gbps architecture in future years, in addition to considering a migration to 800 Gbps technology. [Section 7.1]
- ORNL has two 400 Gbit connections to ESnet and OLCF will have two 400 Gbps connections to the Information Technology Services Directorate (ITSD) of ORNL. Additionally, a 100 Gbps dedicated WAN connection is between Defiant, an HPC cluster on the Advanced Computing Ecosystem (ACE) testbed, and the ESnet testbed for IRI activities. [Section 7.3]

- The WLCG recommendation for large Tier-1s, such as Fermilab, at HL-LHC startup is a minimum of 1 Tb/s. Fermilab is currently provisioned at 1.6 Tbps for science traffic and 800 Gbps for general-purpose traffic. The latter does include the fraction of Fermilab's science not carried via the science networks. This configuration is expected to be adequate to satisfy the currently known requirements for 2024 to 2028. If needed, additional 400G connections can be added to the existing LAG interfaces. Switching to 800G Ethernet technology might become justifiable in 2028 to 2030. [Section 7.4]
- The DUNE experiment extends across two facilities ~1,300km apart. One site is Fermilab in Batavia, Illinois, and the second site, the far detector (FD) location, is SURF, in Lead, South Dakota. The Fermilab-managed perimeter network will provide connectivity for the DUNE FD site to Fermilab. Sufficient capacity to move ~30 PB of science data annually is required. The perimeter network equipment will reside in two geographically diverse buildings at SURF: the Ross Dry Building and the Yates Administration Building. Each building has a shaft that provides access to the underground facilities at the 4,850 ft level (4850L). WAN connectivity is to be provided and managed by ESnet. [Section 7.4]
- The BNL Network Perimeter consists of two redundant routers that support interface speeds of 10 GbE, 40 GbE, 100 GbE, and 400 GbE. These routers are split between BNL's two data centers. Each BNL router connects to a local ESnet router at 800 Gbps (2 x 400 GbE) totaling 1.6 Tbps of Internet/WAN connectivity. Each 800 Gbps circuit is utilized by all scientific and administrative communities at BNL: all data traffic to and from BNL flows through either of these circuits. To support the large data flows to and from A Toroidal LHC ApparatuS (US ATLAS) resources, BNL offsets the LHCOPN and LHCONE primary circuits so each service in theory could transfer 800 Gbps of data. [Section 7.5]
- The SLAC networking core currently operates at multiple 100 GbE for connected enclaves and groups. The networking core of S3DF operates at 100 GbE, with a pair of 2 x 100 GbE links to the SLAC core. SLAC connects to ESnet via two pairs of 100 GbE links; these link pairs typically operate as primary/secondary. No general load balancing or dual use of the links occurs. One pair of the SLAC/ESnet connections is dedicated for use by the Rubin Observatory Virtual Routed Network (Rubin Long-Haul Network, LHN) over ESnet. Within the next two years, SLAC expects to augment the SLAC/ESnet connections with a pair of 400 GbE links. [Section 7.6]
- External network connectivity to ANL is provided by ESnet, Metropolitan Research and Education Network (MREN), Internet2, and Internet2 Peering Exchange (I2PX). ESnet has optical transport and routers deployed on the ANL campus in a redundant configuration supporting up to 400 gigabits per second (Gbps) connections. The connections to MREN are 2 x 100 Gbps. The border routers are connected to each other with 4 x 100 Gbps connections. The connections to the campus core network are 2 x 40 Gbps and 40 Gbps to the perimeter firewall. By the end of FY25, the perimeter firewall will be replaced with a Palo Alto PA-5450 with 100 Gbps connectivity. [Section 7.7]

4.8 Emerging Needs

- Commercial cloud resources are an intriguing addition to the possible cosmology simulation workflow structure. While the cost of using these resources remains high, augmenting existing resources during periods of heavy use and justifying the increased expenditure may be possible. One could imagine a situation where the use of AI/ML frameworks becomes an instantaneous source of large-scale data; if difficulty exists in allocating local computational and storage resources to meet the demand, these workflows may need to be integrated with commercial cloud infrastructure that can expand as needed. [Section 6.1], [Section 8]

- Future cosmological simulation data sets may be too large to store locally by individual research groups, necessitating the use of streaming workflows developed to facilitate on-the-fly analysis; use of that new paradigm may lead to repeated downloads and analyses. This situation may be exacerbated as AI/ML frameworks can be essentially instantaneous sources of large-scale data; at the same time this difficulty could be ameliorated if only the models were sent over and local computational and storage resources were sufficient to meet the demand. In the future, these workflows could be integrated with commercial cloud infrastructure. [Section 6.1]
- Cosmological simulation resource usage becomes a question of latency, bandwidth, amount of associated computing, and cost for computing, data transfer, and storage. Primarily due to cost issues with commercial clouds, the cosmological simulation community does not see this option as being viable currently, but the situation could change very quickly. Initial tests with cloud vendors are expected in the coming years to address some aspects of the experimental workflow. An anticipated advantage is the ability to make data available, with server-side compute, while placing the cost of using the data onto the end user, who may be supported by a different funding agency/source or country. Hybrid-cloud solutions that bridge to DOE facilities could be very beneficial, as other HEP experiments like Rubin have explored. [Section 6.1]
- A possible disruptive element for cosmological simulation is the role of AI/ML in low-level simulation and analysis code, as well as at the catalog-analysis level. Since such techniques are typically very data intensive, the associated network requirements could conceivably be significant, especially if compute resources are not local to where the data are available. [Section 6.1]
- DESI does not currently use cloud services for data processing or serving core data products. Exploratory work for sharing DESI data through the Amazon Open Data program has occurred, though this is not used by the collaboration itself. There has been interest in deploying the data-processing pipeline on Amazon Web Services (AWS) as an alternate site to be used during outages of the primary computing resource; data transfer and hosting costs make this option prohibitive. [Section 6.2]
- DESI makes some use of AI/ML for its data-processing pipeline through standard methods that do not require any special resources. More active cutting-edge work is happening on using AI/ML for DESI science analyses, particularly using generative AI to create simulated datasets faster than traditional simulations. This experimentation uses standard toolkits available via NERSC's Perlmutter, and does not require additional hardware or software support. [Section 6.2]
- DESI has used emerging LEO satellites, such as Starlink, at the remote telescope site as a back-up connection for when the primary Internet is down. This was especially critical in 2022 after a wildfire resulted in a multi-month outage of the fiber optic connection to the mountain. The LEO connection was sufficient for command and control of the telescope, but was not sufficient for transferring observational data, which required driving hard disks up and down the mountain. [Section 6.2]
- Currently the Rubin Data Management system does not make heavy use of AI/ML for data production or distribution. Members of the scientific community are expected to deploy their own AI/ML techniques in analysis of Rubin's scientific data products for modeling, inference, and anomaly detection. This expectation is a major driver for bulk download capabilities for Rubin data products, and may also influence the development roadmap for Rubin-supported analysis platforms on five-year timescales. Multiple community alert brokers currently provide users with AI/ML capabilities for classification and anomaly detection. [Section 6.3]
- DESC has no concrete plans for using any DESC-specific cloud applications. Initial engagement by DESC scientists will happen in the US DAC, through the RSP. [Section 6.4]

- Due to the large variety of science applications and multiple approaches and stages of DESC's analyses, ML will be applied in many instances of large-scale computation (CPU and GPU) at NERSC. [Section 6.4]
- CMB-S4 has flexibility in how it approaches aspects of computing and networking. The project will explore all available options for the software used in data management, as well as available computational resources (HPC, grid, or cloud). [Section 6.5], [Section 8]
- CMB-S4 has no plans to purchase dedicated AI/ML hardware or services. Possible applications of ML exist in areas of low-level data cleaning (detection of classes of "glitches" in the data), and for improvements in observing strategies / planning. In both cases the evaluation of such models is cheap, and model training can be performed at existing centers such as NERSC. [Section 6.5]
- The CMB-S4 project has no intention of using WiFi, cellular, or satellite connectivity at the site in Chile. Some of this equipment can generate a request for information (RFI) that impacts science measurements. CMB-S4 may make use of an existing dedicated point-to-point microwave link to the nearby town. This link has been shown not to interfere with existing experiments and is a useful back-up connection for remote access. [Section 6.5]
- New features for DUNE in two to five years related to networking are related to further use of HPC centers, like Polaris at ANL or Perlmutter at NERSC, for performing simulation and data reconstruction general processing, as well as for executing dedicated AI/ML processing workflows. This could amount to a sustained average rate of >1 TB/hr to and from that facility over a production campaign period (~60 days). Similarly, transfers of dedicated data formats for AI/ML processing to SLAC (temporary location) and then the DOE HPC centers (where the workflow is run) will also require approximately a rate of 1 TB/hr. [Section 6.8]
- DUNE has followed the strategic direction of Fermilab in using cloud approaches only for those new hardware types which DUNE cannot quickly acquire. Examples include tests of inference servers on Triton (e.g., a commercial entity), NERSC Perlmutter as part of the ongoing SONIC project, and Google Cloud Platform (GCP). With one of the above tests with compute workers at Fermilab and the inference server at Google cloud in Iowa, DUNE was able to saturate the Fermilab general off-site network bandwidth, which at the time was limited to 100 Gbps. [Section 6.8]
- DUNE has several scientists who are actively exploring ML use in reconstruction analysis. Currently, most of them run locally on Fermilab's Elastic Analysis Facility. [Section 6.8]
- Google Cloud and AWS computing services have been successfully integrated within the ATLAS workflow and data management systems PanDA and Rucio. The solutions may be considered in the implementation of AFs for HL-LHC, with all the added values and services offered by the commercial cloud providers. In future work, data exports from cloud data centers to WLCG sites will have to be optimized, involving research network providers. In 2024 the Google Cloud is used for MC production processing as an extension of a US Tier-2 center. [Section 6.9]
- ML is used in a wide range of applications in ATLAS, including physics analysis, simulation, and physics object identification. Models trained for classification and regression, as well as generative models, are core components of the software and analysis chain. Most models used by the collaboration are trained on private resources due to the modest resource requirements and satisfactory availability of institutional GPU resources. ATLAS does not expect any significant changes in network load or the distributed computing software stack design during the ongoing analysis of Run 3 data. [Section 6.9]

- Given the size, complexity, and fully distributed nature of ATLAS computing, all workflow and data distribution could eventually be optimized and managed with AI. [Section 6.9]
- The use of in-network caching has had some impact on network use for the LHC community. Much of the R&D efforts in the recent years have been determining where caches may be beneficial for placement, how large they need to be to be effective, and what strategy should be used to store and replace data sets. Questions remain about deployment strategies to work out for both CMS and ATLAS. Regional caches were found to not be as effective (since the overlap of data use was not high enough), but caches within a facility, and ensuring that the workflow management tools know to check the caches, produced the most reliable results. The adoption and more widespread use of the NanoAOD format is anticipated to have some impact on cache performance over time. [Section 6.10], [Section 8]
- CMS does not currently use cloud resources to any significant extent. Previous studies have shown them to be a more costly model than the owned-resource model that CMS currently relies on. A notable exception would be the case of needed resource bursts for either CPU or network. The resources available for either CPU or TA networking in the cloud far exceed those available to CMS. At this point, CMS tools are generally able to use cloud services, typically via infrastructure at one of the tiered sites, but CMS does not have plans to use cloud services extensively in the near or longer term. [Section 6.10]
- CMS has an increasing use of AI/ML in its production workflows. Analysis use cases have evolved even more quickly with greater freedom to adopt new toolkits and approaches. Both the Fermilab Tier-1 and Tier-2 facilities have GPU capabilities for CMS use, and the CMS community also relies on national resources. CMS is also researching a dedicated AI/ML training facility (at Vanderbilt) in recognition that AI/ML brings two distinct use cases, training and inference, that both require substantial resources (but in differing access patterns). CMS expects to predominantly use on-site and HPC resources rather than commercial cloud. [Section 6.10]
- Reliable networking can be used to reduce disk replica requirements either by using tape recall or caching. By the end of Run 4, a copy of the entire CMS MiniAOD will be around 100 PB. If 10% of this is used during any given month in a caching system, one can estimate the need for 10 PB/month of transfers to keep the cache up to date with the most recently used data. Cache network needs will be typically bursty because users read full data sets rather than just a file or two from one. Understanding and expanding caching use cases and needs are part of ongoing R&D. It is expected that investments into disk space at Tier-2s against network utilization will become more critical in the long term. Other forces may also push higher network requirements, such as how spinning disk technology evolves in terms of cost per terabyte and I/O capabilities relative to CMS needs. [Section 6.10]
- ATLAS and CMS are exploring advanced techniques for improving data transfer and monitoring capabilities across the WLCG. New technologies such as SciTags and packet marking are being tested and deployed to enhance network monitoring, traffic differentiation, and optimized data flow across high-speed networks. The deployment of SciTags and packet marking is in progress across several sites, which will lead to better traffic differentiation and resource management across the network. [Section 6.11]
- SENSE (SDN for End-to-End Networked Science at the Exascale) uses SDN to dynamically manage network paths, prioritize traffic, and provide optimized routing. This setup is ideal for data-intensive tasks, as it can allocate resources in real time based on network conditions and experimental priorities. Rucio, developed by CERN, complements SENSE by managing large-scale data distribution across the network. Its features include automated data replication,

strategic data placement, and support for geographically distributed resources, which streamline the movement and availability of critical datasets. [Section 6.11]

- Many of the LHC experiments have explored cloud services and are using them as part of their global set of resources. Two primary modes of operation are envisioned: a particular site contracts with a cloud provider and resources are (transparently) added as part of the site from their experiment's perspective, and an alternative method where a new site is created within a cloud provider, utilizing the various provisioning, management, and resilient infrastructure available. While the first option has been used with varying degrees of success, the second is not yet feasible for several reasons: [Section 6.11]
    - The entity setting up a site within a cloud provider needs to be able to control costs to make it feasible, but policies and tools are not usually available to allow them to do so.
    - Egress charges from cloud providers are usually expensive and may vary depending upon where the egress goes.
    - Connecting to resources like LHCONE is very challenging because of the acceptable use policy (AUP) requirement that all traffic from the advertised subnets for LHCONE must be LHC related.
- During HL-LHC the tracking algorithms used in event reconstruction are expected to see increased use of GPUs. Parts of the events may need to be shipped to data centers (HPCs, cloud vendors, WLCG data centers) for GPU processing, with the results sent elsewhere for recombination with the data processed on CPU-only resources. Future DOE HPC machines are focusing on support for AI/ML, with large numbers of GPUs and other accelerators. ML training requires large amounts of data transferred into the HPCs for use during the training. This will have an impact on the WAN network, as data are transferred from WLCG data centers to HPCs in the US for ML training and inference. [Section 6.12]
- The global infrastructure is currently relatively uniform, with approximately 70 compute and storage elements presenting an identical interface to CMS. The data distribution model proposed by WLCG for the HL-LHC is commonly referred to as the "data lake model." This model envisions fewer regional entry points, allowing internal optimizations and potentially reducing the operational burden. Data transfer between two lakes is a top-down-controlled activity governed by Rucio and executed via FTS. Access to data from processing elements inside the lake is mediated either via caches or streaming directly from one of the data lake origins. [Section 6.12]
- ESnet and the LHC experiments (ATLAS and CMS) have started an R&D project that makes use of in-network data caches operated by ESnet using XRootD software. The pilot project focuses on caching of PHYSLITE data products to see if the time to reach physics results can be shorted, and the overall network traffic reduced for data analysis (only used parts of data files cached and reused). [Section 6.12]
- SuperCDMS SNOLAB will make minimal or no use of cloud storage. Only a few services will be cloud-based, including Slack and GitLab. [Section 6.13]
- SLAC is building support for the execution of AI/ML workflows with substantial investment in GPU. The SuperCDMS SNOLAB reconstruction working group (WG) is looking into taking advantage of the GPUs for event reconstruction. TensorFlow is also used in some analyses. One use case is for datasets severely limited by noise coming from vibrations. With ML we can remove the noise effectively while saving the low-energy signal events, which are important for low-mass data management (DM) search. The same method can be used to predict the energy deposited in the crystals, which works well in low energies where other techniques do not perform well. [Section 6.13]

- AMS does not leverage any commercial cloud services for its computing or storage needs. [Section 6.14]
- AMS does not require any connectivity to be provided by wireless or satellite connectivity, beyond its core networking use case that leverages TDRS. [Section 6.14]
- For NERSC-10 and beyond, NERSC will consider incorporating "cloud" technologies that add significant value to the NERSC user community and are affordable. This approach to provide cloud-like services includes NERSC Spin, a container-based platform where users can build their own websites, API endpoints, databases, and other network services that can integrate with center networks and data storage systems. In the next two to five years, NERSC is exploring deeper hybridization of Spin with the HPC system itself, which will open possibilities for performing computational tasks in a more cloud-like environment. [Section 7.1]
- OLCF anticipates more integration with outside clouds for certain applications. OLCF has been testing integration with things such as weights and biases; this is an ML/distribution layer (DL) analytics dashboard. The heavy computations happen on Frontier, but the analytics dashboard is off-site and integrated through API mechanisms. OLCF is surveying cloud resource performance for HPC workloads, particularly for large-scale leadership-class systems. Current constraints around data in the cloud play a large role in limiting its use, particularly cost. With IRI, the High-Performance Data Facility (HPDF),[12] and Frontiers in Artificial Intelligence for Science, Security and Technology (FASST),[13] things will continue to change and become more distributed. [Section 7.3]

4.9 IRI Responsiveness

- All the factors that exacerbate the issue of long-term storage for cosmological simulation may not have a single solution, and may require a variety of approaches. If a single location cannot become the home for data sets, efforts must be made to unify a cataloging system to enable locality searches, despite the distributed nature of the data. [Section 6.1]
- Long-term storage of large-scale cosmological simulation data remains a critical problem to address, as many of these data sets are connected to ASCR Leadership Computing Challenge (ALCC) and the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) programs,[14] which have finite lifetimes. The advent of the IRI program and HPDF facility should present a global solution to this problem, to be deterministic about the way that long-running simulation efforts can be cataloged, stored, and found by the global community. [Section 6.1]
- An area of friction for a number of scientific efforts (e.g. DESC, and DESI in particular) remains the mechanisms used to manage identity and permissions for the large population of scientific users. A common problem is the management of credentials to access computation and storage across facilities (e.g., DOE HPC centers), and how these may differ from accounts used to manage other aspects of the scientific workflow (e.g., home institutional accounts). [Section 6.2], [Section 8]

12 https://www.hpdf.science/
13 https://www.energy.gov/fasst/
14 https://doeleadershipcomputing.org/about/

- The DESI data lifecycle involves all three IRI patterns: [Section 6.2]
  - Time-sensitive analysis to detect operations problems and update the survey plan.
  - Data integration–intensive work to combine PB-scale data from multiple other collaborations and funding agencies for analysis.
  - Long-term campaigns to process data for the five-year DESI survey, with proposed extensions and follow-up programs through at least 2042.
- The Rubin-LSST science enterprise exhibits all three IRI patterns to some degree: [Section 6.3]
  - Time sensitive: Rubin images are transferred in near-real time from the Summit to the USDF, where they are promptly processed to generate alerts.
  - Data integration intensive: Many Rubin science analyses will require extensive integration between multiple data sets both internal to Rubin (images, catalogs, light curves) and external (data from complementary surveys at other wavelengths, spectroscopic facilities, and space-based telescopes). This data integration is likely to involve a range of different AFs both within the DOE system (e.g., SLAC, NERSC, ALCF) and beyond.
  - Long-term campaign: Rubin has a 10-year primary survey mission and a potentially longer horizon in an extended mission mode. Successful scientific operation over this horizon requires sustained access and integration between Rubin's experimental-support computing facility at SLAC and the DESC analysis facility at NERSC.
- Rubin operations employs a model of distributed data management and processing across multiple sites. The current baseline plan calls for using a combination of the Rucio software system to coordinate multisite data movement and the PanDA software system for orchestrating distributed processing jobs. This use case is highly aligned with the vision for multisite interoperability under IRI. [Section 6.3]
- DESC sees a potential for certain workflows that are more compute intensive to branch out across multiple sites in the future. In the past, working across different facilities usually meant specific workflow setups tailored to the infrastructure/hardware available at each facility. The IRI vision will help overcome this hurdle and provide users with a very similar environment across facilities to ease moving across platforms. [Section 6.4]
- DESC workflows are heterogeneous but primarily fall under two categories: [Section 6.4]
  - The great majority (>90%) both in terms of compute and storage will correspond to long-term campaigns. These will include the processing and reprocessing of LSST data releases to extract data vectors, analyze systematics, model the signal, and infer cosmological parameters. These require sustained access to resources at scale over a longer time to accomplish a well-defined objective.
  - Some of our workflows are time sensitive. These rely on the Rubin "alert stream," or intermediary brokers, for astrophysical events and generate actions or outputs that need to be processed within some period.
- CMB-S4 will leverage aspects of IRI including federated login and access across facilities, as well as tools for job submission and management that can span domains of operation. Ideally, these integrated tools would allow job management not only on HPC centers but also on HTC or Grid-computing resources. [Section 6.5]
- Mu2e has one example of the time-sensitive IRI pattern. Mu2e plans to perform initial processing of data from the experiment within several hours of the data being taken. The time constraint is driven by the need to produce high-precision, high-statistics information sensitive to subtle effects below the sensitivity of the instruments themselves. [Section 6.6]

- DUNE exhibits all the IRI patterns in design and operation. The full DUNE near detector simulation and reconstruction stack has been run at NERSC Perlmutter, as well as many simulations for the DUNE FD. Patterns exhibited include: [Section 6.8]
    - Time sensitive: The supernova pointing workflow will be time sensitive and requires the analysis of as much of 320 TB from the first two detector modules in the span of 4 hours or less.
    - Data integration intensive: It will be necessary to bring data from the DUNE near and far detectors together into a common analysis format. This is an extremely I/O intensive job.
    - Long-term campaign: DUNE intends to be taking data for at least two decades, and during that time our computing model calls for reconstructing all the data we have taken to date at least once a year.
- The LHC experiments can benefit from IRI, especially if the DOE HPC centers and the newly awarded HPDF are seamlessly integrated into the existing HTC (grid) infrastructure. IRI would provide the opportunity to implement near real-time workflows for critical payloads, such as High-Level Trigger, which should produce results in a timely manner. IRI would benefit all HEP by providing a uniform interface to disparate resources in an easier to integrate with way, and to better fit the variety of HEP workflows. [Section 6.9], [Section 6.10]
- The LHC experiments have proactively pursued opportunities at US HPC sites, and currently have many allocations available for use. Typical workflows run on these allocations are like the IRI concept of a long-term campaign, which relies on sustained access at scale. New HPC resources typically come with some, potentially substantial, work to set up given facility specific configurations and constraints. A common interface (or access requirements) across DOE HPC sites and models for sustained access are potential IRI benefits. [Section 6.9], [Section 6.10]
- Globally, many funding agencies are responsible for both HPCs and the LHC's WLCG experiments, and a natural outcome is to try to support WLCG needs via HPCs. As an example, Germany has consolidated many of its Tier-2 sites into a new HPC allocation. For the US, the DOE HPC facilities becoming "standardized" would be very beneficial, allowing collaborations to utilize resources without having to dedicate staff to operate on a bespoke basis. Without something like IRI driving this uniform interface, the experiments are also forced to engage with each HPC as it is designed to help ensure it is usable for that experiment. [Section 6.11]
- From the ATLAS and CMS point of view, the IRI long-term campaign pattern best fits the LHC use case. Historically, the collaboration expects a full year's worth of reconstructed data and simulation to be produced on a few-month timescale, so there is relatively little time sensitivity. Online data-quality monitoring and time-sensitive calibration procedures are handled by a separate coordination area of the ATLAS and CMS experiments and have dedicated resources at CERN, the host laboratory. In addition, since each particle collision event is an independent identically distributed quantity, the HL-LHC science workflow is trivially parallelizable and does not need large-scale data shuffle operations to execute. [Section 6.12]
- If, under the IRI, HPC centers could provide some of the services found at the WLCG data centers, both ATLAS and CMS will be able to make better use of these resources. [Section 6.12]
- Many of the workflows that the LHC experiments will use during HL-LHC under the IRI category of data-intensive workflows (event reconstruction and ML) or long-term campaign (large-scale MC simulations). IRI will also provide the opportunity to implement some near-real-time workflows for critical payloads. [Section 6.12]

- For the LHC experiments, the US provides ~30–40% of the global computing and storage capacity. If this community is to effectively engage with IRI, it will be necessary to ensure that the interface to computing and storage resources made available by IRI is compatible with the global infrastructure, including university and international partners. The current WLCG interface is expected to evolve in preparation for HL-LHC, though the exact nature of this evolution is not yet determined. [Section 6.12]
- SuperCDMS SNOLAB is not equipped, so far, to work in an IRI workflow. Its data structure, data taking, and data-processing patterns are simple enough that the need for advanced solutions has not appeared. It would instead require additional personnel and resources that are not accessible with the current funding level. [Section 6.13]

5 Review Actions

ESnet recorded a set of high-level actions from the HEP user facilities and ESnet requirements review that extend ESnet's ongoing support of HEP-funded collaborations. Based on the key findings, the review identified several actions for HEP, ASCR, and ESnet to jointly pursue. These items are listed as guidance for future collaboration, and do not reflect formal project timelines. ESnet will review these with HEP participations on a yearly basis, until the next requirements review process begins.

These are also organized by topic area for simplicity and follow common themes:

- Facility management and readiness.
- Scientific data management.
- Scientific workflow.
- Computational and storage requirements.
- Remote collaboration and operational requirements.
- Multifacility computational workflows.
- Networking for local and wide-area data mobility.
- Emerging needs.
- IRI responsiveness.

5.1 Facility Management and Readiness

- The HL-LHC will produce over 2,000 hours of data collection every year when the HL-LHC starts running in 2030. The HL-LHC program is expected to last for a decade or more (through the 2040s). Large improvements in networking, throughout the R&E community, will be required to enable the ambitious physics goals of the HL-LHC. [Section 6.12]
- The current members of the LHC collaborations are not expected to encounter major changes during HL-LHC operation. The collaborators will continue to be distributed worldwide, which puts strong requirements on global networking, and cooperation between entities that operate R&E networks. Computing resources will continue to be globally distributed, with no centralized data or computing location, implying that physicists will need access to the distributed resources. [Section 6.12]
- In the HL-LHC era, it is expected that the US will continue to provide two Tier-1s and more than a dozen Tier-2 facilities, funded via DOE and NSF respectively, as resources for ATLAS and CMS. Both experiments will make use of any HPC centers funded by either DOE or NSF that will grant allocations to community science projects. Commercial cloud providers are not expected to be a major component of computing infrastructure, but may provide some capacity. R&E networking providers are critical to connecting all these resources. [Section 6.12]
- HL-LHC predictions show a mismatch between the computing and storage resources the experiments can afford, versus the resources needed to reach science goals. In response to this gap, the experiments are exploring alternatives in how to utilize storage, computing, and network infrastructure. The network baseline currently being planned is composed of terabit-scale (1–2 Tbps) backbone networks with the largest resource sites connected at multiple 400G scale (800 Gbps+). Network use will be at least a factor of 10 larger than Run 3. [Section 6.12]

5.2 Scientific Data Management

- The cosmological community is interested in advancing a model that would help reduce the number of times and locations that need to be involved to archive long-term data products, and is looking for assistance in scoping and realizing this design goal. Establishing permanent storage facilities would allow the community to establish a more regular pattern of collaboration: a smaller set of data portals that can be used to catalog and retrieve long-term campaigns of products along with a simplified way to support analysis activities, potentially leveraging interactive computing approaches, at a well-known number of sites. [Section 6.1], [Section 8], [Section 5.2]
- DC24 was widely successful for the LHC community, and several important outcomes came from the exercise including understanding bottlenecks in the network configuration at the Tier-1 centers and the ways many of the tools were improved and replaced since Run 2. The LHC community is interested in developing a program to test the effectiveness of transfers both to benchmark methods at increased capacity, and integrate new functionality into existing tools. Such tests should be done in collaboration with ESnet. For example, one lesson learned from the most-recent data challenge organized by WLCG (DC24) was that it is necessary to utilize a given link at full capacity to determine whether it has any performance abnormalities. It is desirable to develop a process of regular testing of network capacities (e.g., relying on perfSONAR) to test connections to the maximum and identify problems before they become a bottleneck in production. [Section 6.10], [Section 8]
- The LHC community will conduct a full analysis of the observed results of DC24, and discuss ways to both capture and act on observational data in near-real time for DC27. An in-depth data analysis will be needed, and this may lead to a process of evaluating each networking link in the ecosystem, and ways that both LHC monitoring data, and that from ESnet, can be captured and used to evaluate overall system performance. [Section 6.11], [Section 8]
- A common theme across many of the HEP case studies was the surrounding challenges of data access as collaboration space grows larger. The LHC collaboration has a highly structured ecosystem that facilitates access to groups, and individuals, through a number of tools that manage identity and permissions across a federated grid infrastructure. Other collaborations that have less structure may struggle with the available tools to manage collaboration size long term and are attempting other solutions (e.g., neutral site hosting through cloud providers, or adopting the policies of the strictest member of a collaboration). [Section 8]
- The DC24 exercise provided critical insight into the behaviors of all layers of the technology stack for the LHC communities: data management and ability software, computation and storage, as well as local- and wide-area networking. The DC27 is expected to be a crucial preparation step for the HL-LHC operation. Beyond the major milestones of a full DC, the mini-DC activities in the run-up to a full DC will be useful ways to validate specific portions of technology support structure. ESnet plans to engage with the LHC community on ways to participate in all these. [Section 8]

5.3 Scientific Workflow

- DESI uses Globus for transfers of collaboration data, and users report good experiences with Globus when they want to download an entire directory tree. But they have found the interface lacking for the more common use case of selecting a subset of data from a complex directory tree (e.g., all files matching a given regex regardless of where they appear in the directory structure). Discussions with Globus may be useful to help facilitate ways to address this use case. [Section 6.2]

- The Rubin Observatory faces several challenges, some of which are outside the control of ESnet and DOE. Like other large facilities, it must take cybersecurity into heavy consideration. NOIRLab has been responsive to recent threats and has put in place controls that can make collaboration harder. In addition, the theme of federated identity management is also present for Rubin, and represents a barrier to the use of some of the tools that were selected during design. As a result, Rubin is looking to adopt tools used elsewhere in HEP for data mobility and workflow management, including Rucio. The integration steps to have this work in Rubin's own environment have been challenging, and Rubin will be looking for help and advice. [Section 6.3], [Section 8]
- Known or potential data-related resource constraints for Rubin as of FY24 are: [Section 6.3]
    - Insufficient computing resources available to meet the full analysis demand from the Rubin scientific user community, and/or difficulty in co-locating Rubin data products with available computing resources. This includes unbounded or unscoped demand for AI- and ML-based analysis of Rubin data.
    - Changing cybersecurity threats and policy landscapes that can present barriers to planned data-access and data-sharing services, and/or can introduce new costs, technical requirements, and schedule impacts.
    - Challenges associated with implementing and routinely operating the distributed processing model for annual data-release processing.
    - Challenges and constraints of operating in a shared data facility with other experiments.
    - Network and I/O performance impacts of transferring a very large number of relatively small files.
- DUNE's software stack will leverage many of the tools already in use within the HEP community, including Rucio, FTS, and XrootD. DUNE is also looking into facilitating Globus as a transport mechanism with the workflow managers, such as Rucio and XrootD. DUNE has converted to use containers, which has greatly increased the ability to make the analysis work portable across available resources, particularly at DOE HPC centers like NERSC. [Section 6.8], [Section 8]
- Limitations are imposed by the various SuperCDMS collaborator sites depending on firewall rules, use of virtual private networks (VPNs), and multifactor authentication (MFA) requirements that often limit the amount of automation of the transfer tools. The lack of uniformity in the availability of tools (e.g., not all sites have Globus endpoints) and the different network protection stances also imply the need to develop data-transfer tools tailored for each site. For a computing group of a small experiment like SuperCDMS, the lack of standardization creates additional overhead. [Section 6.13]

5.4 Computational and Storage Requirements

- Within the NERSC computing center, DESI is limited by I/O performance stability of the "community file system": median performance is fine, but semi-rare outlier events can cause multi-minute delays or even lead to job hangs or timeouts. The Perlmutter scratch disk performance is better, but the difficulties of constantly syncing datasets between the scratch filesystem and CFS has been human-error prone and effort-prohibitive. [Section 6.2]
- The inability to pragmatically synchronize constantly evolving PB-scale datasets between multiple locations is a limiting factor for DESI. That includes internal to a data center (e.g., between CFS and the scratch filesystem) and external (e.g., between NERSC and another DOE HPC facility). [Section 6.2]

- For the full-fledged Rubin Observatory data releases, expected in the two- to five-year time frame, DESC will see a limitation in some of the science cases that require full access to the complete single-visit image data set (only available at SLAC or CC-IN2P3, but not NERSC as currently planned) rather than just the galaxy catalogs. DESC is currently looking for mitigations to this problem, such as limiting the size of transferred data to NERSC but enabling the analyses nonetheless, or shipping software containers to run remotely at these facilities instead. In addition, large-scale image simulations produced at NERSC or elsewhere would possibly stress DESC's small-latency storage allocation at NERSC, as well as compete with computing allocation for its analyses in parallel to real data analyses. The transition to NERSC-10 in this period and its unknown impact on established pipelines also introduces some uncertainty. [Section 6.4]
- The SBN program anticipates using more DOE HPC resources for its simulation workflows. SBN does anticipate needing more disk space during this process, which limits output. [Section 6.8], [Section 8]
- SBN's primary concern is not network bandwidth, but rather I/O performance. The project relies heavily on tape-based storage, which causes a bottleneck when reading out older data. Fixes are being explored to rely less on this approach, and may increase expected data volumes. [Section 6.8], [Section 8]
- The barrier to using DOE HPC resources within the LHC community remains, particularly in integrating the systems as resources into the software components, and normalizing the way the resources could be accounted for (expected versus ad hoc or opportunistic). More specifically, there is a desire to create agreements or other forms of arrangements, similar to an SLA, with DOE HPC facilities such that resources can be considered pledged, like how Tier-1s and Tier-2s operate. This helps with planning for both the HEP community and the ASCR resources. Some of this may require discussion at the programmatic level. [Section 6.12], [Section 8]
- The LHC community experiences friction with supporting a streaming-based workflow at large HPC facilities. Workarounds have been established to support this, either by allowing the worker nodes to stream directly or by leveraging caches and data lakes that are deployed close topologically. All these design choices come with trade-offs, which implies a set of R&D activities that will extend in the coming years. [Section 6.12], [Section 8]
- The changes to hardware at LHC Tier-1 and Tier-2 sites will impact some of the performance aspects of future DC activities, but the schedule of hardware upgrades is hard to predict. Some sites plan on upgrading hardware yearly, so their upgrades appear more "rolling"; when a DC happens they will go with what is available. Others may be able to schedule larger hardware purchases, and plan to install and test prior to the DC27 activity. Many sites will not get funding until closer to DC27 (e.g., end of 2026, start of 2027); thus they may not be able to upgrade anything before the next major challenge. [Section 8]

5.5 Remote Collaboration and Operational Requirements

- Broadly speaking, DESI is not limited by raw bandwidth but by the convenience of tools for managing cross-site data flow. Globus is efficient for one-off data transfers, but issues such as expiring certificates, multi-factor authentication, data discovery, and data filtering (i.e., selecting what needs to be downloaded) limit convenience for day-to-day random access of huge datasets. It is desirable to consider if computing centers and end users would be able to subscribe to datasets from each other, with automatic efficient syncing of the subsets of the data actively being used, with the toolkits managing purging unused data to stay within the local quota limits.

The faster the network is, the more viable this becomes, since it reduces the barrier to having to re-download data. [Section 6.2]

5.6 Multifacility Computational Workflows

- Using a single centralized facility to support DESI provides benefits of a single ecosystem to learn, and minimal overhead for sharing results with collaborators from other institutions. At the same time, it becomes a single point of failure during NERSC outages, leading to ongoing interest in multisite computing operations in the future. The limiting factors are not networking for data transfer, but rather the tools for making that data sharing automated and transparent for hundreds of users accessing PBs of data stored in millions of files. DESI could transfer its major data release to another site, but that is of little use unless the hundreds of DESI users at NERSC can also get accounts at that site, and the outputs of their daily work can be transparently synced back and forth between the two sites. [Section 6.2]
- SuperCDMS is evaluating options for network outages and how to proceed with operations. In case of extensive, site-wide outages at SLAC, it would also be important to have another data path from SNOLAB to an alternative computation site, namely DRAC and/or Texas A&M, to not interrupt the processing flow. The criteria needed to temporarily switch from SLAC to another site for storage and processing needs to be carefully evaluated and the process tested. [Section 6.13]

5.7 Networking for Local and Wide-Area Data Mobility

- ESnet will follow up with international partners to include up-to-date networking maps of the international paths, namely TransPAC,[15] which is operated by International Networks at Indiana University, and Americas Pathway (AMPATH),[16] which is operated by Florida International University, for updated connectivity diagrams and information regarding support for networking in the region with other partners. [Section 6.7], [Section 8]
- The DOE HEP and DOE ASCR program offices are working out the specifics of the connectivity to SURF that ESnet will need to complete for the proper connection. The efforts are planned to converge by the end of fiscal year 2026. [Section 6.8], [Section 8]
- The LHC experiments anticipate network needs will grow gradually. Increasing network capacity and performance will be needed at Tier-1, Tier-2, and Tier-3 facilities. The exact magnitude of future network provisioning is expected to be determined jointly by the experiments, ESnet, the R&E networking community, facilities and campuses, funding agencies, and other involved parties. [Section 6.9], [Section 6.10]
- CMS would like to develop a program to test the effectiveness of transfers both to benchmark methods at increased capacity and integrate new functionality into existing tools. Such tests should be done in collaboration with ESnet. For example, one lesson learned from the most-recent data challenge organized by WLCG (DC24) was that it is necessary to utilize a link at full capacity to learn if a given link has any performance abnormalities. It is desirable to develop a process of regular testing of network capacities (e.g., relying on perfSONAR) to test connections to the maximum and identify problems before they become a bottleneck in production. [Section 6.10]

15 https://internationalnetworks.iu.edu/projects/TransPAC/index.html
16 https://ampath.net/

- Networking support at BNL and Fermilab, to support the ATLAS and CMS Tier-1 computing facilities, respectively, has increased and is approaching an aggregate of 1 Tbps through all connectivity. Both sites are exploring ways to move toward 800 Gbps native connections in the coming years, when the technology is ready to be supported. As of 2024, multiple 400G connections have been installed to the Fermilab and Brookhaven national labs as well as HPC facilities used by HEP, and the upgrades to 400G (or N X 400G) at some US CMS campuses are either in place or imminent. [Section 6.11]
- One place that R&E networks could help would be in jointly negotiating with various cloud providers to create a template contract that would support no egress charges if the R&E providers bring high-bandwidth links to the cloud providers, as well as defining policies and associated tools to allow LHC site creators to manage costs by, for example, controlling how data flow to only use the zero (or low cost) egress connections globally. [Section 6.11]
- The HL-LHC requirements for LHCOPN links approach 1 Tbps and have pushed sites to explore more affordable technologies to achieve it. IT-INFN-CNAF, the Italian Tier-1, has deployed a pilot connection using Data Center Interconnect (DCI) equipment at the Tier-1 (CNAF) and Tier-0 (CERN) sites, which use four light channels on the dark-fibers of two different providers, GÉANT and GARR. The solution can provide 4 x 400 Gbps link for a total of 1.6 Tbps. Spectrum sharing is achieved by interconnecting the reconfigurable optical add-drop multiplexer (ROADM) device of the two providers. It saves on the cost of more expensive router interfaces, since most of the existing LHCOPN links are provided as Multiprotocol Label Switching (MPLS) Layer 2 virtual circuits. [Section 6.11]
- During the 2020–2021 studies of LHC network requirements, the LHC experiments estimated that Tier-1 centers would need terabit/sec connections and Tier-2 sites would require 400G connections. That was driven by estimates of some (but not all) data-related operations of CMS and ATLAS, leading to an estimated bandwidth need of 4.8 terabits/sec (Tbps) across the Atlantic for CMS and ATLAS alone in a so-called "minimal model" and 9.6 Tbps in a more "flexible" and realistic model. [Section 6.11]
- The LHC experiments have defined two models to predict bandwidth needs into the era of HL-LHC: [Section 6.11]
    - Minimal model:
        - Sum (ATLAS,ALICE,CMS,LHCb)*2(for bursts)*2(safety-margin) = 4.8 Tbps expected HL-LHC bandwidth.
    - Flexible model:
        - Experiments may need to reprocess and reconstruct the collected data during the year.
        - This requires doubling the bandwidth of the minimal model: 9.6 Tbps expected HL-LHC bandwidth.
- All the major Tier-1s are expected to have at least 800 Gbps LHCOPN links at the start of Run 4 with the ability to grow to 1.2 Tbps or even 1.6 Tbps during the Run. Tier-1s should also have an equivalent amount of bandwidth to LHCONE to serve Tier-2s. The LHC experiments predict the following bandwidth estimate for sites during HL-LHC: [Section 6.11]
    - Each major Tier-1:
        - 1 Tbps to the Tier-0 (LHCOPN).
        - 1 Tbps to the Tier-2s (aggregated, LHCONE).
    - Each major Tier-2:
        - >400 Gbps (LHCONE) and resilient.

- The individual LHC experiments are interested in discussing the set of features that must be present in future network designs. If identifying a minimum set of capabilities in switches and routers is possible to ensure the next generation of network components can meet the needs of the community, it would simplify the procurement and installation process. The sooner those capabilities are in place, the sooner prototyping and testing future capabilities will be possible. [Section 6.12]
- For HL-LHC networking, four main requirements, beyond connectivity, have been identified: [Section 6.12]
    - Understanding the impact of new features in networking (SDN/NFV) by testing, prototyping, and evaluating impact is necessary. The experiments will need to evolve applications, facilities, and computing models to meet the HL-LHC challenges; this will take time.
    - The ability to understand WAN network flows is limited. New methods to mark and monitor network use are needed. Packet marking is viewed as a high priority so that everyone (experiments, sites, and R&E networks) can understand the origin and intent of network flows at any point along their path.
    - Developing, prototyping, and testing network features at suitable scale will be needed. Having at-scale network testbeds will be very important to understand the potential impact changes might have on operations.
- The LAN connectivity for SNOLAB is 2 x 10 Gbps, and the WAN connectivity consists of a 1 Gbps link that is also shared with 10 other experiments. This limitation implies that SuperCDMS must plan to work within a budget of no more than 10% of either of these connections during operation. This link allows SNOLAB to reach Laurentian University, which is then connected to ORION (the provincial R&E network of Ontario), and Canadian Network for the Advancement of Research, Industry and Education (CANARIE, the national R&E network of Canada). CANARIE peers with ESnet, which facilitates the connectivity to reach SLAC. The network limitations were explored during a project review in 2022. At the time, the cost was considered too high to upgrade the 1 Gbps WAN link, which will cause the experiments that share this resource to be more austere in their use. SuperCDMS will consider this in the design of their trigger mechanism and computing and storage strategy. ESnet can discuss options with SuperCDMS regarding the use of other connectivity options. [Section 6.13], [Section 8]
- The network connectivity between SNOLAB and SLAC is a critical component, given that all workflows and all scientific analysis happen off-site. And the bandwidth bottleneck between SNOLAB and the ORION network could present a challenge in the future if no upgrades to the bandwidth occur, given the tendency of data volumes from current and future experiments to grow rather than shrink. SNOLAB and SLAC could explore alternative network paths (e.g., those that do not rely on a terrestrial link) to address the critical need to keep networking active during experimentation. [Section 6.13]
- Strategic planning (five+ years) for OLCF shows LAN traffic will be increased to 800 GbE for Science DMZ and data-movement networks with individual 1.6 Tbit links for network interconnects. This will also imply that DMZ traffic (e.g., related to data-movement) will scale to 6–12Tbit/sec. All other LAN traffic will have multiple 800 Gbit connections. Firewalls in this time frame will be 4–6 Tbit/second in throughput. WAN capabilities at ORNL could have multiple 800 Gbit connections to ESnet, and OLCF could have multiple 800 Gbit connections to ITSD. The connections between the ESnet testbed and OLCF resources could be further expanded to multiple 400 Gbit interfaces. [Section 7.3]

- ESnet and Fermilab network engineering teams have drafted a plan for connectivity back to Fermilab to support DUNE. The plan calls for redundant, geographically diverse, vendor-diverse, 100 Gbps circuits from SURF to ESnet points-of-presence (PoPs) in Chicago and Denver, respectively. ESnet backbone services (On-demand Secure Circuits and Reservation System [OSCARS]) would then extend those L2 paths back to Fermilab. Deployment of both 100 Gbps circuits should occur by the 2026–2027 time frame. Both pathways will be connected to ESnet in a production capacity. Currently, a dedicated 10 Gbps circuit between SURF and ESnet's Kansas City PoP provides DUNE FD connectivity back to the Fermilab campus. This circuit, provided by South Dakota's Research Education and Economic Development Network (REED), will revert to a tertiary back-up role once the two 100 Gbps circuits are procured and deployed. [Section 7.4]
- Fermilab is concerned about not having the capability to adequately stress test the WAN infrastructure as network technology step-increases push bandwidth capacities to terabit scale. The ability to generate traffic at 400 GE, 800 GE, and 1.6 TE is challenging. Fermilab desires to be able to routinely stress test the WAN infrastructure to ensure it is performing as expected. The complexity of having multiple types of virtual network services sharing the same physical infrastructure adds another level of difficulty to this. It is desirable to consider this as a capability to concurrently generate high traffic loads across various virtual network services (LHCOPN, LHCONE, default routed Internet Protocol [IP] path, etc.). This capability is a critical tool in Fermilab's ability to troubleshoot aggregate data bottlenecks where multiple virtual network services are involved. Fermilab proposes that this should be developed as a service that ESnet provides to all its sites. [Section 7.4]
- To adhere to the LHCONE and LHCOPN AUP, various complex routing policies and features must be implemented on BNL Network Perimeter. This AUP states only specific HEP resources at BNL (e.g., US ATLAS, Belle II, DUNE, etc.) should have access to the LHCONE and LHCOPN services, which leaves all other resources at BNL able only to traverse the general Internet services. To provide this exclusivity for LHCONE and LHCOPN, separate campus Virtual Routing and Forwarding (VRF) instances have been created on both perimeter routers. Policies are then applied to only import default, black hole, and local BNL routes into the campus VRFs. Then policy-based routing (PBR) policies are configured on all perimeter downstream ingress interfaces. In short, these PBR policies state that all private IPv4 and IPv6 and HEP subnets forward to the default routing table, which has full visibility to all WAN circuits, while all other BNL subnets should forward to the campus VRF instance, which can only see the general Internet services. [Section 7.5]
- Within the next five years, SLAC expects to upgrade the 2 x 400 GbE links to 2x or 4x 800 GbE links. LCLS-II has a requirement for 1–2 Tbps capacity between SLAC and NERSC. This requirement is expected to be met with the N x 800 GbE upgrades. The SLAC core network will also be upgraded to N x 800 GbE. [Section 7.6]

- Several data-related resource constraints were identified for ANL's support of scientific efforts: [Section 7.7]
  - Within the Computing, Environment, and Life Sciences (CELS) directorate, the recent upgrades to 400G WAN & Science DMZ paths, which very seldom exceed 25% sustained utilization, have resulted in very few concerns on data-related constraints.
  - Outside of the Science DMZ, the replacement of 40G firewall paths with 100G should provide room for future growth of non-CELS scientific divisional network traffic.
  - The perfSONAR infrastructure, used for network monitoring and testing, has not been upgraded in a number of years and in some cases, it was decommissioned without immediate replacement due to operating system support factors.
  - Rebuilding the fleet of perfSONAR nodes throughout campus is necessary for us to proactively identify data-related constraints. However, this has been a challenge to prioritize due to other initiatives.
- During DC24, ESnet directly assisted several sites in "full capacity" testing of their network links. This was done using multiple perfSONAR servers to generate traffic, and changing routing policy on both the network and the site end. This testing helped to reveal configuration bottlenecks, and was invaluable in gaining an understanding of the maximum expected performance. Future DC activities can take a similar approach, but there are also capabilities that ESnet's High Touch servers (that feature Field Programmable Gate Arrays (FPGAs)) could assist with. Several High Touch servers could conceivably be programmed to create network traffic and direct it at resources located within a lab or facility perimeter. ESnet will follow up with interested sites on this prototype concept. [Section 8]
- The LHC community and ESnet are collecting an enormous amount of data on the performance of data mobility tools, the networks that are a part of data transfer, and the individual hardware components (routers, switches, servers, storage). These data were incredibly useful during the DC24 activity, and helped to identify transfer bottlenecks and other gaps to delivering the requirements for HL-LHC operation. Despite having visibility, it is not possible to review and act on all the collected telemetry data. Cultivating or hiring a data analyst to review the data and suggest ways they can be better incorporated in future tools would benefit both efforts. [Section 8]
- Several facilities are not connected to LHCONE currently, despite being heavily involved in LHC-affiliated computation and storage activities. These include major computational sites (e.g., NERSC and TACC) used for simulation workflows, SLAC (which has operated Afs in the past), and when applicable, commercial clouds. No formal agreement with any of these facilities exists now to be part of the LHC collaboration in a structured way, but investigating their ability to connect to LHCONE to support several important use cases is recommended. [Section 8]
- LHCONE runs a hybrid approach to supporting IPv6 currently. The upcoming government mandate (e.g., M-21-07) will imply that Fermilab and BNL will move toward IPv6 capabilities, and subsequently the Tier-2s will have to move toward dual stack support at a minimum. One critical issue that cannot be solved by the US community will be the ability of CERN to deliver IPv6 capabilities from its service infrastructure; some capabilities may be IPv4 only in the coming years. ESnet will continue to work with US collaborators on this to ease the transition to IPv6 operation. [Section 8]

- Several international networking partners factored heavily into the discussions this year; connectivity to Europe, South America, and the Asia-Pacific region is accomplished through relationships with several R&E partners. ESnet maintains peering relations with most of these partners, and their connectivity carries significant amounts of HEP traffic daily (e.g., LHC and ProtoDUNE traffic from Europe, Rubin Observatory and CMB-S4 traffic from South America, Belle II traffic from Japan, etc.). ESnet will include current network capability diagrams into this report to document the current state of international networking. [Section 8]

5.8 Emerging Needs

- ATLAS has raised a concern regarding a lack of sufficiently fine-grained network monitoring and performance information to help debug and automate data transfers in real time between various sites. Better visibility is being worked on through various entities including WLCG Site Network Monitoring, the SciTags initiative (labeling network flows and IPv6 packets with owner/purpose), and improved monitoring from the various tools and infrastructures, but ATLAS has significant gaps in monitoring. [Section 6.9]
- CMS has treated the global network of sites as a mesh with identical links when it comes to bulk transfers. The XrootD data federation was designed from the beginning to be cognizant of the TA links being limited, but treated links within the US as identical. The data lake model currently discussed in WLCG makes clear regional distinctions. The concept of creating a data lake has progressed slowly due to challenges around how increased latency due to the longer international links will affect application performance. [Section 6.10]
- CMS would like to continue exploring with ESnet how to contribute towards global R&D projects such as the Global P4 Lab testbed, use of ESnet High Touch services, and evolving technologies such as Smart NICs. [Section 6.10]
- The next several years are critical to transition managed network R&D by CMS, including results of the SENSE project, into production for the HL-LHC. The SENSE, Caltech, and University of California, San Diego (UCSD) teams are close to 400G throughput with production infrastructure using XrootD, FTS, and XrootD. [Section 6.10]
- ESnet will continue to work with the LHC community on co-design efforts like SENSE and Rucio, SciTags, and other potential opportunities. The experiments need to be specific in what they are hoping to achieve, however. The aim of joint projects could be to "co-design" an effective system that makes best use of the available network resources, coordinated with the site resources, and manage joint site and network operations cooperatively. This needs to be spelled out in more detail and discussed with ESnet leadership. [Section 6.11], [Section 8]
- For LHC experiments in general, the resource requirements are growing beyond what flat budgets can support. This leads the experiments to work very hard at optimizing their data and analysis processes to maximize what the available resources can do. Various trade-offs are continually being explored (network use vs multiple replicas of data, data format changes vs number of required datasets to meet various needs, exploitation of nonstandard computing capacity, reconstructing data on demands vs storing them), to try to achieve the time to science target with constrained resources. [Section 6.11]
- Because of recent changes to technology drivers, the evolution of ATLAS and CMS's "process of science," and the outlook for 2030 when the HL-LHC era will begin, it appears that the actual networking requirements may be somewhat larger than foreseen during the 2020 HEP Requirements Review. Detailed projections are difficult at this point, but should be tracked and updated every two years or so to make quantitative estimates and (re-)establish the requirements in both network capacity and capability. [Section 6.11]

- One interesting option to explore on the timescale of HL-LHC is the joint ATLAS and CMS use of Rucio for distributed data management. This could provide a mechanism to interact with ESnet (and other R&E networks), communicating near-term data-movement intents and perhaps negotiating for any required quality of service (QoS) or deadline requirements. [Section 6.12]
- GPU use to support some aspects of AI and ML within the LHC community is expected to continue; unknown currently is how this may impact WAN usage. Experiments like DUNE have shown that bursting to a commercial cloud to utilize GPUs is possible, but leveraging DOE HPC resource would allow better control over the WAN components. Opportunities could exist to schedule network use (e.g., using mechanics like SENSE and Rucio) to coincide with the availability of resources at an HPC facility. [Section 6.12], [Section 8]
- The LHC work on data lakes will continue, and LHC has learned several things from the years of R&D and use during activities like DC24. An effort to reduce the number of "entry points" into storage resources has begun. The XrootD tool right now is not bound by factors such as network capacity or distance/latency, and may find nonoptimal answers to requests for data from clients. The overall goal of a data lake should be to simplify the way data are accessed, along with providing the most performant ways to access and retrieve them. Data lakes can also leverage some of the work on caching, and allow the caches (which often do have more intelligence) to assist in making optimal choices about retrieval. [Section 6.12], [Section 8]
- The use of in-network caching was discussed at length to identify ways that ESnet and the LHC community can work together more efficiently to design and operate an effective service. ESnet has some previous work in attempting to create a Content Delivery Network (CDN) for science with others in the DOE, but the usage patterns never amounted to a measurable increase in performance. A core R&D question to address remains how large the caches should be, and what policies on data retrieval and retention will be valuable to implement to keep cache hit rates high. ESnet would like to gain better monitoring insight into the caching work, either by having access to LHC data (e.g., MONIT) or by having more network layer visibility integrated into Stardust. [Section 6.12], [Section 8]
- As in the discussion after the LHC Operations use case, ESnet will continue to work with the LHC community to advance R&D efforts that support the HL-LHC operations era. This includes the already mentioned efforts (e.g., SENSE and Rucio, SciTags), but also ways that HPC resources can be integrated and utilized more efficiently. The use of ESnet High Touch data, and integration with Stardust, may be critical in passing along network-level telemetry to the storage and computational components of the LHC community. [Section 6.12], [Section 8]
- Fermilab has a strong interest in making use of ESnet High Touch Services to generate data for AI/ML-based analysis of network traffic patterns. Fermilab is currently working cooperatively with ESnet on distinguishing between streaming-based CMS data movement and file-transfer-based CMS data movement. It is desirable not just to continue with this effort but also to look toward additional opportunities to expand use of ESnet High Touch Service capabilities in analyzing WAN traffic. [Section 7.4]
- ANL has interest in learning how to leverage the MetrANOVA platform, like how perfSONAR has been adopted. [Section 7.7]
- Authentication, authorization, and accounting approaches have limitations. Typically, it is possible to specify who can do what and keep a record of what has happened. Within an organizational structure this is not too much of a challenge, but across organizations with different policies it can be complicated, as different implementations and validation methods are employed. [Section 8]

- Several HEP case studies outlined their production, proto-duction, and experimental use of commercial cloud resources. ESnet, as a network carrier, is identifying ways to assist with the execution of the scientific mission, and will continue to help with connectivity and peering when applicable. During some HEP evaluations of cloud solutions for portions of analysis workflows, the network was identified as a bottleneck in certain situations: mainly when traffic remained within a cloud versus being allowed to egress and use more predicable R&E routing and transit. In situations like this, ESnet could serve as a valuable partner between connected end sites and cloud resources, to try to force routing choices that take advantage of R&E WAN paths versus the internal networking a cloud may use. [Section 8]

5.9 IRI Responsiveness

- The cosmological simulation community expects that future workflows, especially in the IRI context, may see the dataflow controlled from a central location, or distributed to a few "trunk" sites. As simulations become more integral to the work of individual researchers, potential exists to investigate: [Section 6.1]
  - The concept of distributed analysis sites that will better adjust to the workflow demands.
  - On-demand data-intensive analysis capabilities at the DOE computing centers as an IRI/HPDF development.
- DESC analyses will eventually employ the galaxy catalogs, but also the full imaging data of approximately 10 PB, in the later years of the project. Replicating even the high-level imaging data at NERSC is inefficient and unfeasible, due to the scale of the expected allocation. DESC expects to embrace a fully distributed analysis at different facilities. Workflows will be adapted, preferably through IRI approaches, to a more flexible computing environment. [Section 6.4]
- Future DESC needs, more than five years out, will likely require that the analysis process employ not only the complete galaxy catalogs, but also the full imaging data of O(10s) PB. Replicating even the high-level imaging data at NERSC is inefficient and unfeasible, due to the scale of the expected allocation. Thus, embracing a fully distributed analysis at different IRI and international sites, through grid solutions such as Distributed Infrastructure with Remote Agent Control (DIRAC), could be an option to offset some of these allocation difficulties. [Section 6.4]
- DESC has two outstanding issues to work through in the coming years: [Section 6.4]
  - DESC needs technical support in implementing data distribution solutions (Rucio, DIRAC). Standing up these web services requires close interaction with the facilities (e.g., NERSC). Built-in facility support for this software (e.g., just like the facilities provide built-in support for Globus) would be very beneficial.
  - DESC faces complex issues with identity and account management across various services/sites. Two broad classes of sites exist. First are sites like NERSC, where a DESC member has a NERSC account that lets them access both NERSC and Globus. Second are sites like in the UK, France (CC-IN2P3), and Rubin IDACs, where the DESC needs to deploy a proper Virtual Organization (VO) and have a proper grid certificate. These items are needed to enable interactions with grid software like Rucio and DIRAC. Ideally, the VO and certificates would be linked to each member's NERSC account. Further, we would ideally want to be able to target the NERSC queues using grid software like DIRAC so that we can run jobs, move data, etc., across multiple sites transparently.
- DUNE can take advantage of some aspects of the IRI design approaches, and the collaboration is currently exploring using NERSC and ALCF for aspects of simulation and analysis workflows. GPU resources required for parts of the workflow may need to be used at the DOE HPC

facilities. DUNE has not fully identified a long-term location for the storage of its public data archive. This may also fall into the realm of an IRI workflow, or an emerging use case for HPDF to assist with. [Section 6.8], [Section 8]

- The DOE IRI effort has identified "cybersecurity and federated access practice" as one of the six core practice areas for future work. This practice area will focus on creating novel solutions that enable seamless scientific collaboration within a secure and trusted IRI ecosystem. The discussions within this effort are nascent, but a core understanding will be recognizing that the security posture and policies of participating facilities may differ, and there must be a way to determine and agree on appropriate mechanisms to implement identity and access control. [Section 8]
- The discussions within the HEP community touched on several factors that should be considered by the IRI effort. A first step will be to establish a hierarchy of needs: an example being that accessing a set of resources or tools that spans administrative boundaries is very different than the ability to use different tools and resources within the same location. A similar policy language and control structure could solve both, if ways to differentiate the needs and articulate the required control structures exist. [Section 8]

6 HEP Case Studies

The case studies presented in this document are a written record of the current state of scientific process, and technology integration, for a subset of the projects, facilities, and principal investigators (Pis) funded by the Office of HEP of the DOE SC. These case studies were discussed between July 2024 and December 2024.

The case studies were presented, and are organized in this report, in a deliberate format to present an overview based on individual experiments, larger facilities, and in some cases the encompassing laboratory environments that provide critical resources for operation. The case studies profiled include:

- Cosmological Simulation Research
- DESI
- Vera C. Rubin Observatory
- DESC
- CMB
- Muon Experimentation at Fermilab: Muon g-2 and Mu2e
- Belle II Experiment
- Neutrino Experiments at Fermilab: SBN and DUNE
- LHC: ATLAS Experiment
- LHC: CMS Experiment
- LHC: LHC Operations
- LHC: HL-LHC & Network Research
- SuperCDMS
- AMS Experiment

Each of these documents contains a complete set of answers to the questions posed by the organizers:

- How, and where, will new data be analyzed and used?
- How will the process of doing science change over the next 5–10 years?
- How will changes to the underlying hardware and software technologies influence scientific discovery?

A summary of each will be presented prior to the case study document, along with a "Discussion Summary" that highlights key areas of conversation from authors and attendees. These brief write-ups are not meant to replace a full review of the case study, but will provide a snapshot of the discussion and focus during the in-person review.

6.1 Cosmological Simulation Research

Cosmological simulations are used to provide detailed theoretical predictions to understand dark matter and dark energy, cosmological constraints on neutrino physics, and the nature of primordial fluctuations. These data products are essential for analyzing and interpreting results from physical cosmological surveys, as well as aiding in survey design and optimization, and in the estimation and control of statistical errors. The level of resolution and volume required of the simulations has been, and will continue, to increase in the coming years, driving data volumes significantly upwards. Additionally, as increased computational power has become available in the exascale era, large-volume simulations with gas dynamics and subgrid modeling of astrophysical mechanisms

have become possible in regimes which were previously the domain of gravity-only simulations. Similarly, the increased computational power will also enable large-scale ensemble simulations that can be used directly for inference or to construct generative models using AI/ML techniques that can be then used for this purpose.

6.1.1 Discussion Summary

The cosmological community is interested in advancing a model that would help reduce the number of times and locations that need to be involved to archive long-term data products. Establishing permanent storage facilities would allow the community to establish a more regular pattern of collaboration; a smaller set of data portals that can be used to catalog and retrieve long-term campaigns of products along with a simplified way to support analysis activities, potentially leveraging interactive computing approaches, at a well-known number of sites.

Removing the need to worry about where the data resides allows researchers focused on developing cosmological simulation to invest more time on the science, versus spending energy in activities to stage data and allocate technology resources. It also prevents half-measures that are being used today to create mechanisms to perform analysis on partial data sets, or adopt streaming approaches which are not always successful on some of the larger computational resources that are not designed to support that mode of operation.

Cosmology data sets are typically owned by the team that generated them, unless the work was specifically part of a collaboration and/or commissioned by the relevant collaboration. But the increasing requirements toward open science — as well as a sense that such simulations are useful for a wide range of applications — has led to more interest in placing simulation results in the public domain. Funding agencies often have requirements for subsets to be made publicly available. The primary barrier in making simulation data sets public is the lack of available storage and/or server-side public-access compute capability. This is an excellent opportunity for an interaction with the IRI effort. Simulation datasets would typically exercise the Data integration–Intensive and Long-Term Campaign patterns.

Cosmology simulations tend not to lose value over time, in fact older simulation catalogs continue to have high value and are being used in emerging work that utilizes AI and ML to construct generative models. Simulations can never truly be reproduced, even if the original codes and inputs are retained: differences in the underlying behavior of the systems that are used to generate are common, resulting in almost entirely new simulations. Thus, older simulations will always have an intrinsic value, since they are unique and nonreproducible. This being said, the community understands that it is infeasible to keep every data set around forever, and there are mechanisms to archive sets that may not receive as much use as others.

Cloud resources are an intriguing addition to the possible cosmology simulation workflow structure. While the cost of using these potential locations remains high now, they may become a way to augment resources during periods of heavy use. One could imagine a situation where the use of AI/ML frameworks become an instantaneous source of large-scale data; if there is difficulty in allocating local computational and storage resources to meet the demand, it becomes possible that these workflows will need to be integrated with commercial cloud infrastructure that can expand as needed to fit the needs.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- Cosmological simulations are used to provide theoretical predictions to understand dark matter and dark energy, cosmological constraints on neutrino physics, and the nature of primordial fluctuations. These data products are essential for analyzing and interpreting results from physical cosmological surveys, as well as aiding in survey design and optimization, and in the estimation and control of statistical errors. [Section 4.1]
- Results from major cosmological simulation runs can be useful for many years, up to a decade or more in some cases. [Section 4.2]

- Data storage needs for cosmological simulations post mid-2020s are likely to be in the 300PB range of spinning disk equivalent, requiring transfer rates of >300 Gbps on an intermittent basis when data mobility is required. [Section 4.2]
- Simulation data streams typically involve moving between 10–100 TB of bulk data when entire data sets are needed. The much smaller "user" downloads for individual analysis of a smaller subset to local storage typically involve <1 TB chunks of data. Progress in remote visualization methods is allowing real-time visualization of many large data sets. [Section 4.3]
- Over the next two to five years, there will be an evolution in cosmological simulation workflow tools to manage local and remote simulation data analysis. This area is in flux, and community desires are only slowly being captured in a design process, mostly because of the complex nature of the requirements and the dynamic nature of the use cases. [Section 4.3]
- Typically, the cosmological simulation dataflow consists of two stages, with the understanding that the data motion may be staged in a scheduled manner or can be highly bursty, depending on the use case: [Section 4.3]
    - Within the facility where the simulation is run (supercomputer, storage, analysis system) and
    - From the host facility to a remote analysis/archive site.
- As increased computational power becomes available in the exascale era of HPC, large-volume cosmological simulations with gas dynamics and subgrid modeling of astrophysical mechanisms have become possible, which were previously the domain of gravity-only simulations. Similarly, the increased computational power will also enable large-scale ensemble simulations that can be used directly for inference or to construct generative models using AI/ML techniques that can be then used for this purpose. The level of resolution and volume required of cosmological simulations has been, and will continue, to increase in the coming years, driving data volumes significantly upwards due to these factors. [Section 4.4]
- Cosmological simulations can, in principle, easily generate far more data than optical and microwave observations. This is expected to be the case even when data becomes available from the Rubin Observatory, the largest of the imaging surveys, which is expected to begin operations in 2026. Estimation of statistical errors relies on hundreds to tens of thousands of simulated examples, and cosmologists often need to simulate results obtained from dozens to hundreds of varying simulation input conditions. While the on-sky pixel-level data can in some cases be larger, the simulated catalogs are already at the level of PBs for large simulation suites. [Section 4.4]
- The size, sophistication, and realism of simulated sky catalogs has improved significantly since the 2020 ESnet Requirements Review on the HEP program, and this puts further stress on storage and data-transfer capabilities. Typical simulation products can be a TB in size, approaching a PB for some of the wide-ranging simulations or those that contain more detail due to improvements in the modeling techniques. [Section 4.4]
- Cosmological simulation continues to be substantially limited by the availability of storage, particularly when considering how to distribute the simulation data beyond the group that performed the simulation. Further, the need for storage beyond the initial computing project is critical; the lifetime of the need is often set by the life cycle of the analysis, which is happening in larger collaborations, and as the projects become more complex, over longer timelines. [Section 4.4]

- Storage for cosmological simulation will continue to be a critical issue over time. Some projects are more storage-limited than they are compute-limited. This issue has several associated dimensions to consider: [Section 4.4]
    - Sizes and numbers of simulations will increase.
    - Locations producing simulations are expected to increase.
    - Funding spans different agencies.
    - Usefulness over time, and to different communities, is increasing.
- Data movement of cosmological simulation data between facilities is not a critical problem, but broad access for the user community remains challenging. Simplifying access with updated portal software, or tools like Globus, has improved data mobility. The next step is to combine this with dynamic access to computing resources. [Section 4.5]
- The ratio of the simulation user community, to the size of the simulation groups, will increase as the role of simulations continues to increase; dataflow patterns will likely become more diverse and there will be increased pressure on data-transfer facilities that serve the resulting simulation outputs. Although the main focus is the continental US, the geographical locations of the endpoints are worldwide. The diversity of endpoints and the number of geographical locations is expected to grow significantly. [Section 4.5]
- The data products generated by simulation groups can be created at any number of DOE- or NSF-funded HPC facilities or other institutions with sufficient resources to execute and store simulation codes. These include all major DOE HPC facilities and a number of other HEP-affiliated laboratories. These data products generated are used "downstream" by a large number of users who are members of science working groups in cosmological surveys and experiments. A common use pattern is to build a mock survey from the simulation data, then perform a clustering analysis on that survey. [Section 4.6]
- The network speeds between major centers are not a limitation for cosmological simulation, and transferring PB-sized data sets is effectively routine. This size is expected to increase in the coming years, but within the capabilities of ESnet and the DOE HPC resources. [Section 4.7]
- It is possible that future cosmological simulation data sets will be too large to store locally by individual research groups, necessitating streaming workflows the data are developed to facilitate on-the-fly analysis, which can then lead to repeated downloads and analyses coordinated with the data flow. This situation may be exacerbated as AI/ML frameworks can be essentially instantaneous sources of large-scale data; at the same time this difficulty can be ameliorated if only the models were sent over and local computational and storage resources were sufficient to meet the demand. In future, these particular workflows could be integrated with commercial cloud infrastructure. [Section 4.8]
- Cosmological simulation resource usage becomes a question of latency, bandwidth, amount of associated computing, and cost for computing, data transfer, and storage. Primarily due to cost issues with commercial clouds, the cosmological simulation community does not see this option as being viable currently, but this situation could change very quickly. Some initial tests with cloud vendors are expected in the coming years to address some aspects of the experimental workflow. An anticipated advantage is the ability to make data available, with server-side compute, while placing the cost of using the data onto the end user, who may be supported by a different funding agency/source or country. On-premise cloud availability at the DOE facilities would be very beneficial, as other HEP experiments like Rubin have explored. [Section 4.8]

- A possible disruptive element for cosmological simulation is the role of AI/ML in low-level simulation and analysis code, as well as at the catalog analysis level. Since such techniques are typically very data intensive, it is conceivable that the associated network requirements might be significant, especially if compute resources are not local to where the data are available. [Section 4.8]
- Long-term storage of large-scale cosmological simulation data remains a critical problem to address, since much of it is often connected to ALCC and INCITE programs that have finite lifetimes. IRI and HPDF should present a global solution to this problem, to be deterministic about the way that long-running simulation efforts can be cataloged, stored, and found by the global community. [Section 4.9]
- All of the factors that exacerbate the issue of long-term storage for cosmological simulation may not have a single unified solution, and may require a variety of approaches. If a single location cannot become the home for data sets, there must be efforts to unify a cataloging system to enable locality searches, despite the distributed nature of the data. This would be an excellent IRI project candidate. [Section 4.9]
- The cosmological simulation community expects that future workflows, especially in the IRI context, may see the dataflow controlled from a central location, or distributed to a few "trunk" sites. As simulations become more integral to the work of individual researchers, there is a potential to investigate: [Section 5.9]
    - The concept of distributed analysis sites that will better adjust to the workflow demands
    - On-demand data-intensive analysis capabilities at the DOE computing centers as an IRI/HPDF development.

6.1.2 Cosmological Simulation Research: Experiment Profile

For the purpose of this case study, supercomputers function as data-generating instruments running flagship-scale simulations. Indeed, cosmological simulations can, in principle, easily generate far more data than optical and microwave observations. This is expected to be the case even when data becomes available from the Rubin Observatory, the largest of the imaging surveys, which is expected to begin operations in 2026. Halo/galaxy catalogs from simulations are typically far larger than their on-sky counterparts: estimation of statistical errors relies on hundreds to tens of thousands of simulated examples (or more), and cosmologists often need to simulate results obtained from dozens to hundreds of varying simulation input conditions. While the on-sky pixel-level data can in some cases be larger, the simulated catalogs are already at the level of PBs for large simulation suites — in fact they could be larger, but storage constraints have been a continuing limitation. The size, sophistication, and realism of simulated sky catalogs has improved significantly since the last report and this puts further stress on storage and data-transfer capabilities. As mentioned above, many simulations are also needed whenever one wants to match the simulations to the actual data, by reconstruction of initial conditions. In this inverse problem application, one iterates on the initial guess, and after convergence one wants to sample from the posterior, where each operation is a separate simulation.

6.1.2.1 Science Background

Typically, the dataflow consists of two stages: (1) within the facility where the simulation is run (supercomputer, storage, analysis system) and (2) from the host facility to a remote analysis/archive site. The data motion may be staged in a scheduled manner or can be highly bursty, depending on the use case. In the future, especially in the IRI context, control of the dataflow may be centralized or distributed to a few "trunk" sites. As simulations become more integral to the work of individual researchers, there is a potential to investigate 1) the concept of distributed analysis sites that will better adjust to the workflow demands, or 2) on-demand data-intensive analysis capabilities at the DOE computing centers as an IRI/HPDF development. Further, if the data sets are too large to store locally by individual research groups, then it may be necessary to stream the data for on-the-fly analysis,

which can then lead to repeated downloads and analyses coordinated with the data flow. This situation may be exacerbated as AI/ML frameworks can be essentially instantaneous sources of large-scale data; at the same time this difficulty can be ameliorated if only the models were sent over and local computational and storage resources were sufficient to meet the demand. In future, it is possible that these workflows will need to be integrated with commercial cloud infrastructure.

The data generated from the simulations exist at multiple levels, from the basic representation used in the codes (particles, fields), to science-level information (e.g., dark matter density and velocity fields, gas distribution, velocities and temperatures, halo and galaxy information), to catalog-level information (properties of simulated galaxies). The usage pattern varies from the group that generated the data analyzing them, to working within distributed collaborations, and to making the results publicly available. Results from major simulation runs can be useful for many years, up to a decade or more in some cases.

Data sets are typically owned by the team that generated them, unless the work was specifically part of a collaboration and/or commissioned by the collaboration. But the increasing requirements toward open science — as well as a sense that such simulations are useful for a wide range of applications — has led to more interest in placing simulation results in the public domain. Funding agencies can put requirements on the data sets corresponding to making subsets publicly available. The primary barrier to making simulation data sets public is the lack of available storage and/or server-side public-access compute capability. This is an excellent opportunity for an interaction with the IRI effort.

6.1.2.2 Collaborators

All the HEP Labs (Argonne, BNL, Fermilab, LBNL, and SLAC) participate in this work, as well as other labs (Los Alamos National Laboratory, Lawrence Livermore National Laboratory, ORNL) and a large number of collaborating universities, including the University of California, Berkeley, the University of Chicago, Harvard University, the University of Illinois Urbana-Champaign, the University of Pennsylvania, Stanford University, the University of Washington, Yale University, and many others. Many of the collaborations involve directly working with the associated experiments (e.g., Cosmic Microwave Background Stage IV (CMB-S4), Dark Energy Survey (DES), (DESI, Rubin Observatory LSST — Dark Energy Science Collaboration (LSST DESC), South Pole Telescope (SPT). Computational facilities used include ALCF, NERSC, and OLCF as well as NSF computing facilities, along with institutional resources available at the participating institutions.

The data products generated by simulation groups are used downstream by a large number of users who are members of science working groups in cosmological surveys and experiments. A common use pattern is to build a mock survey from the simulation data, then perform a clustering analysis on that survey. In the future, the ratio of the user community to the size of the simulation groups will only increase as the role of simulations continues to become more important. Thus, the dataflow pattern will likely become more diverse and there will be increased pressure on data-transfer facilities. Although the main focus is the continental US, the geographical locations of the endpoints are worldwide. The diversity of endpoints and the number of geographical locations is expected to grow significantly.

The data sources are major computing centers and the major analysis endpoints are typically also the same centers. Large DOE collaborations have accounts at these centers, especially at NERSC. Data access for local analysis at secondary sites is provided by the primary data-holding facilities. Public data sets from other efforts can also be made available through these sites.

The number of users can vary considerably depending on the use cases. Analysis of individual simulation outputs may involve small teams, but curation and use of synthetic catalogs by survey teams may involve multiple projects with hundreds of people across the projects. Cosmological survey collaborations are growing in size and are almost at the same scale as high energy physics experiments (Rubin's LSST Dark Energy Science Collaboration has, for example, more than a thousand members). Another trend is the use of multiple data sources (cross-correlation) in cosmology. Simulations are now designed to provide outputs for multiple types of observations so that they can be of use to multiple surveys.

User/ Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
ARGONNE (ALCF)	Primary/ secondary	Data transfer via Globus (portal-enabled)	Variable, TBs to PBs	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage
HARVARD UNIVERSITY	Primary/ secondary	Data transfer via Globus (portal-enabled)	Variable, TBs to PBs	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage
LBNL (NERSC)	Primary/ secondary	Data transfer via Globus (portal-enabled)	Variable, TBs to PBs	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage
ORNL (OLCF)	Primary/ secondary	Data transfer via Globus	Variable, TBs to PBs	Intermittent, week/month	Rarely, via Globus	Stable long-term storage provided by Kronos (HPSS replacement)

Table 6.1.1: Cosmological Simulation Data Mobility

There are too many potential endpoints to list individually in Table 6.1.1. The science collaborations typically have a large number of members (hundreds), from a large number of institutions (~100) all of whom have the right to specific simulation data products. Each consumer endpoint also varies significantly in how much data throughput and volume it can support (from individual desktops to institutional clusters, to cloud usage). Another thing to keep in mind is that in future there may be the need to support a much larger number of users, for example, as part of Rubin Observatory's user base, but this remains to be negotiated. Automation of data transfer with analysis workflows will be an important feature of future applications, which is not properly captured in Table 6.1.1.

6.1.2.3 IRI Readiness

The cosmological simulation community is extremely interested in working with the IRI given the multiple DOE facilities that are used, and have major interests in expanding the coordination across sites that is potentially possible. There are a number of areas, such as integrated workflows (including computation and wide-area data movement), federated identity management, and single-point data management that can be significant early use case examples for IRI. In these scenarios, simulation datasets would typically exercise the Data Integration–Intensive and Long-Term Campaign patterns.

6.1.2.4 Instruments and Facilities

The work performed is primarily computational — the major facilities used are primarily those at the ALCF (Polaris and Frontier, once accepted), NERSC (Perlmutter), and OLCF (Frontier). The use case consists of running large simulations on the primary systems, carrying out analyses on these as well as smaller satellite machines, and making the data available for downstream work. Smaller local resources are available at a number of participating institutions. Institutional storage ranges from ~100 TB to a few PB, and internal data-transfer rates are typically hundreds of Gbps bandwidths, but can sometimes be lower. This level of storage is significantly smaller than the total storage available at the HPC centers, which, it is important to note, is increasing relatively slowly. Additionally, long-term storage of large-scale simulation data remains a problem since much of it is often connected to ALCC and INCITE programs that have finite lifetimes. An exception is the availability of Kronos at OLCF; eventually IRI/HPDF should present a global solution to this problem.

The cosmological simulation community does not foresee a change in the modus operandi over the next five years (and even beyond, unless cloud facilities — which are otherwise very capable and currently dwarf DOE HPC resources — become cost-competitive). Transfer rates are expected to improve as infrastructure upgrades are made. Data storage needs post mid-2020s are likely to be in the ~300 PB range of spinning disk equivalent (although given how slowly the accessible storage has grown, this may be unrealistic), requiring transfer rates of >300 Gbps (intermittent).

6.1.2.5 Generalized Process of Science

The simulation results can be viewed as corresponding to three levels of data that are used for science. Level 1 is the raw data from the simulations, level 2 is intermediate-level analysis data, and level 3 is level 2 data reduced to the level of catalogs/databases (when possible), including large-scale sky maps covering a wide range of observables. Level 1 analysis can be performed both in in situ or post-processing modes, while level 2 and level 3 analyses are primarily in post-processing mode, although this is changing to in situ modes in some cases. All three levels are part of the local and remote "process of science." Level 1 and 2 data analyses involve batch processing, while level 3 analyses can have a significant level of interactivity, so the data access patterns can be quite different (in the future, this last case may have some interesting IRI-related consequences).

In situ analysis uses the supercomputer's own network, or co-scheduled computational resources, whereas level 2 and level 3 analyses can be conducted locally or remotely and may involve moving data from file systems back to the host supercomputer or to analysis resources. This is where the networking support is most important. Level 2/3 data may also be moved over in batches to remote sites where they can be locally analyzed. Note that the actual mass of data at the three levels is roughly similar, except that the granularity increases significantly from level 1 to level 3.

Future activity will be along the lines mentioned previously, and this effort should not change significantly over the next five years.

6.1.2.6 Remote Science Activities

As already stated, the major computational facilities used are primarily those at the ALCF, NERSC, and OLCF. Storage is provided at the facilities at the level of hundreds of TB to low PB per sub-project of disk (potentially larger, in some instances) and substantially more on tape. The larger allocations have included special dispensations for the ALCC and INCITE program projects. Some pre-IRI activities might also be given relatively large allocations. The major data sources are the supercomputers and other associated compute resources (analysis and visualization clusters). The use of cloud resources may also be considered in the near future, both commercial and institutional. WAN is typically via Internet2 and ESnet. This mode of operation is expected to remain unchanged over the next five years.

The current data stream involves typically moving ~10–100 TB of bulk data, with larger transfers carried out on occasion. The much smaller "user" downloads to local storage typically involve <1 TB chunks of data. Progress in remote visualization methods should allow (almost) real-time visualization of many large data sets. Examples of this capability already exist but are not being used in production.

As to the supercomputers themselves, significant architectural changes are unlikely over the next five years, but in any case, these will not be disruptive with respect to the networks.

Over this time span, a few major simulation data archive/analysis centers will emerge (e.g., increase in size of the community file systems at the leadership computing facilities (LCFs) and NERSC, potential on-pre/remote cloud services). Most of the data will eventually be hosted there; they should also have substantial local analysis computing available (since computing must follow the data) — this is essentially an argument for the HPDF and would be a good example case for an HPDF pilot (of course, the HPDF timeline is uncertain).

The data stream will consist of two major types:

- "Feeders" to the data centers, moving ~PB, and
- Data center to "user" links which would typically move <100 TB data chunks (possibly interactive and part of a wider workflow).

Remote visualization methods should allow real-time visualization of many large data sets.

6.1.2.7 Software Infrastructure

There will be some changes to simulation software, but these are unlikely to be very significant (at least partly due to inertia, because of the size of the current software base). Most of the changes that were expected beyond the early 2020s (e.g., ECP-related) have been made already. Additionally, architecture changes in next-generation supercomputers are expected to be evolutionary, with a corresponding situation for developing tool infrastructure.

Software infrastructure for data management varies widely. The use of code repositories is now widespread. Higher-level workflow tools to manage simulations and simulation analysis (including remote analysis) are slowly emerging, although most of this work is still performed by hand-written scripts. Software containers are seeing more use for complex software stacks. Data transfer is primarily through Globus, which has proven very effective both within institutions and externally as a graphical tool or when coupled with a portal. A number of tools are used to process data sets, including embedded capabilities (e.g., data compression within I/O). The use of data containers (HDF5, PnetCDF) is sometimes limited by their reliance on message passing interface (MPI) -IO; HDF5 is also considered to be too complex by some users. Native I/O tools written for simulation codes still obtain the best performance.

Over the next two to five years, there will be an evolution in workflow tools to manage local and remote simulation data analysis. This area is still in flux, and community desires are only slowly being captured in a design process, mostly because of the complex nature of the requirements and the dynamic nature of the use cases.

6.1.2.8 Network and Data Architecture

These factors are highly specific to individual programs and facilities. Much of the simulation activity involves the use of DOE HPC centers that ESnet has already constructed sufficient network capabilities. For purposes of this review, networking and data architecture features are not foreseen to be a major source of requirements versus future HPC center/university compute and storage requirements.

6.1.2.9 Emerging Technology Use

6.1.2.9.1 Cloud Services

As far as cloud computing and storage are concerned, usage is primarily a question of latency, bandwidth, amount of associated computing, and cost for computing, data transfer, and storage. Primarily due to cost issues with commercial clouds, the cosmological simulation community does not see this option as being viable currently, but this situation could change very quickly. Some initial tests with cloud vendors are expected shortly.

There are several attractive features of the cloud model, including virtualization and resource elasticity; there is an expectation that these will become ever more important with time and to drive increased use of cloud resources. Another advantage is the ability to make data available, with server-side compute, while placing the cost of using the data onto the end user, who may be supported by a different funding agency/source or country. On-premise cloud availability at the DOE facilities would be very beneficial.

More recently, cloud hosting of supercomputing (CPU/GPU) at massive scales and specialized AI/ML-optimized hardware, such as Tensor Processing Units (TPUs), has added a new layer of capability that has potential benefits for cosmological simulations and the associated data analysis, especially since the AI/ML capabilities can be easily

exploited with productivity languages such as Python. Cost concerns remain the key issue given the scale of the operational requirements.

6.1.2.9.2 Artificial Intelligence and ML

The cosmological simulation community uses a broad set of ML tools, such as classic approaches embodied by principal component analysis (PCA) and Gaussian Processes. Additionally, there has been an explosion of interest in modern AI/ML for a variety of applications, especially in the area of generative models, inference, and speeding up simulations by replacing traditional compute kernels by AI/ML "look-up" methods. There is also considerable interest in applying these techniques for analyses that go beyond conventional summary statistics. Since DOE supercomputers are all based (mostly) on GPUs, they are highly suited to large-scale AI/ML training and inference — it is easy for the simulation community to utilize these resources for AI/ML applications.

6.1.2.9.3 Wireless or Distributed Location Science Activities

There are no use cases that involve these types of activities.

6.1.2.10 Data-Related Resource Constraints

This work continues to be substantially limited by the availability of storage, particularly when considering how to distribute the simulation data beyond the group that performed the simulation. Further, the need for storage beyond the initial computing project is critical; the lifetime of the need is often set by the life cycle of the analysis, which is happening in larger collaborations, and as the projects become more complex, over longer timelines.

The network speeds between major centers are not a limitation, and transferring PB-sized data sets is effectively routine. This size is expected to increase in the coming years but not by a large factor. User-level transfer remains smaller, as downloading entire catalogs is still not a typical use case.

6.1.2.11 Data Mobility Endpoints

The cosmological simulation community is using the DTNs at ALCF, NERSC, and OLCF for most of the large-scale traffic.

6.1.2.12 Outstanding Issues

The following factors are considered to be high priority for the long-term success of this community:

- A possible disruptive element is the role of AI/ML in low-level simulation and analysis code, as well as at the catalog analysis level. Since such techniques are typically very data intensive, it is conceivable that the associated network requirements might be significant, especially if compute resources are not local to where the data are available.
- Storage for simulations will continue to be a critical issue over time. Some projects are more storage-limited than they are compute-limited. This issue has several associated dimensions to consider:
    - Sizes and numbers of simulations will increase.
    - Locations producing simulations are expected to increase.
    - Funding spans different agencies.
    - Usefulness over time, and to different communities, is increasing.

- All of the factors that exacerbate the issue of long-term storage may not have a single unified solution, and may require a variety of approaches. If a single location cannot become the home for data sets, there must be efforts to unify a cataloging system to enable locality searches, despite the distributed nature of the data. This would be an excellent IRI project candidate.
- Data movement between facilities is not a critical problem, but broad access to the user community remains challenging. Simplifying the access method to use updated portal software, or tools like Globus, has improved data mobility. The next step is to combine this dynamic access to computing resources.

6.1.2.13 Facility Profile Contributors

Cosmological Simulation Representation

- Daniel Eisenstein, Harvard University, deisenstein@cfa.harvard.edu
- Salman Habib, ANL, habib@anl.gov
- Katrin Heitmann, ANL, heitmann@anl.gov
- Uros Seljak, UC Berkeley and LBNL, useljak@berkeley.edu

6.2 DESI

DESI is a DOE-led international cosmology collaboration which is creating the world's largest 3D map of the universe. A five-year survey started in 2021 to observe the spectra of tens of millions of galaxies, quasars, and stars. These data are used to construct the map to study the origins of the accelerated expansion of the universe and test models of gravity.

6.2.1 Discussion Summary

DESI's infrastructure relies on two major components that are linked via network infrastructure: KPNO and NERSC. There are a number of other instruments at the KPNO facility, and all share a common 1 Gbps-capable network connection from the summit to a base station. A larger 10 Gbps network then links the base station to the rest of the R&E community. These connections have experienced trouble in the past; natural disasters (e.g., wildfire) have severed connectivity, and bandwidth limitations imply that all scientific operations must share the common connection fairly. Beyond the network, computing at a single site (e.g., NERSC) means that all associated operations can be brought to a halt should there be planned or unplanned maintenance. This can limit productivity, given there is no back-up option, which can impact real-time requirements. DESI has ways to buffer data when NERSC is unavailable, but is otherwise stalled without alternative computing mechanisms.

Despite the challenges of network availability, the network requirements are relatively modest. An entire night's observation can reach 100 GB or more (and the tooling understands to rsync the last 40 TB of collected data, in the event of network interruption), which is typically sent in bursts every 10 minutes via tools like rsync. In a worst-case scenario, use of transportable media could be used to take an evening's worth of data manually down to the base station, or through a slower network path such as a LEO Satellite (e.g., Starlink). This could potentially disrupt the nightly calibration. ESnet will work with DESI, and the partners of KNPO to investigate the network connectivity, and offer pathways to increase the reliability, capacity, or develop alternative methods.

DESI computing is done using interactive computing methods at NERSC, namely Jupyter Notebooks. DESI has considered moving to more widely available computing options not tied to a single facility, such as a cloud platform (similar to those that the Rubin Observatory has adopted); a core concern lies in the user community however, who are slow to adopt new approaches to computing and are making do with legacy tools for a lot of the analysis process. Investment into a new system halfway through an experiment is not feasible to consider.

One area of friction that could be addressed is an easier method to manage data mobility outside of the NERSC facilities. Data can be egressed via approaches like Globus, but this typically only is an option for users who are familiar with the tool, and have the necessary credentials to manage it. An efficient tool that works like more commercial tools (e.g., Dropbox, iPhoto, etc.) may appeal to the users. Typically, users do not want to be bothered with the functional details of a data mobility tool, and just want things to work efficiently and simply; rsync is easy enough to understand wherein data appears in the correct folders without much thought. For this reason, it is unlikely that DESI would adopt some of the other tools that the HEP community has championed (e.g., Rucio), as the cultural shift required to do so is not very strong.

Another related area of friction remains the mechanisms that are used to manage identity and permissions for users. A common problem is the management of credentials to access computation and storage (e.g., NERSC) and how these may differ from accounts used to manage other aspects of the scientific workflow (e.g., home institutional accounts, NOIRLab accounts, etc.). DESI is a firm supporter of the work IRI is investigating to simplify this aspect of the scientific workflow.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- DESI is a DOE-led international cosmology collaboration that started in 2021, and will operate for five years. The project will create the world's largest 3D map of the universe. DESI utilizes the Mayall 4-meter telescope, located at KPNO outside of Tucson, AZ. [Section 4.1]

- NERSC in Berkeley, CA, is used as the primary scientific analysis center for DESI, and the facility hosts the ~1000-person international collaboration of scientists, who access the data the cosmology analyses. [Section 4.1]
- At NERSC, the data volume is projected to grow at a rate of ~1 PB/year from 2020-2025, reaching ~10 PB by the end of 2026. The data are stored in a hierarchy of thousands of directories with millions of files. Extensions to the DESI have been proposed and received positive reviews; if approved these would continue the data growth at a similar rate through the mid-2030s. [Section 4.2]
- Public datasets will be curated beyond the timescale of DESI by the Astro Data Lab at NOIRLab. The intention is to also host the public DESI data via the Cosmology Data Repository at NERSC, though long-term funding commitment for that is not guaranteed. [Section 4.2]
- Science analyses using DESI data are primarily performed at NERSC, with the scale of analyses varying from single-core Jupyter Notebooks to large parallel-batch jobs. Many of the analyses make use of simulated datasets generated at NERSC, while some others are generated at their DOE HPC facilities and transferred to NERSC. [Section 4.3]
- Every ~15 minutes, DESI obtains spectra of 5000 astronomical objects, resulting in 1–2 GB of raw data per exposure. Data is processed at NERSC in semi-real time (minutes to hours after capture), and results in a processed file size of ~10 GB per exposure. DESI projects that the total data volume will grow to O(10 PB) by the end of the survey. [Section 4.3]
- Every one to two years, DESI reprocesses all the data taken to date using the latest data-processing code and calibrations. These serve as reference datasets for cosmology papers and are publicly released six months after the key cosmology projects have been published. [Section 4.3]
- Transfers of processed data from NERSC to the NOIRLab and University of Utah mirrors use Globus for the initial bulk transfer, then rsync to pick up any later differences. [Section 4.3]
- The DESI collaboration space at NERSC will eventually host 10PB of experimental data. These data are processed nightly, and analyzed by users that access NERSC from remote locations. Backups are performed to HPSS at NERSC, and mirrored to partner facilities (NOIRLab in Tucson, AZ, and the University of Utah). [Section 4.4]
- Collaborators primarily access the data through direct login to NERSC, or through NERSC's Jupyter server. Many collaborators have expressed a preference for downloading subsets of data to use locally, though DESI lacks developed tools for easy management of that remote analysis workflow. This is currently achieved through ad hoc downloads with Globus, scp, or https. [Section 4.5]
- The primary data access mode for DESI is for collaborators to directly login to NERSC directly, and perform work at the location of the data. It is possible for collaborators to download a subset of data, for processing at local institutions for analysis. Public (noncollaborator) access is only through downloads from NERSC, or through Jupyter servers at the NOIRLab AstroData Lab, who host a subset of the public DESI data. [Section 4.5]
- DESI data from KNPO is transferred in 10-minute intervals to NERSC using rsync. This rsync tool is not known for high performance, however the cadence of the experiment dictates that nightly data collection is less than 100 GB and must reach NERSC via network that is limited to 1 Gbps between the telescopes at KPNO and the base station; use of more advanced tools will not improve the overall process of science. [Section 4.6]

- DESI is fundamentally a multisite experiment, with data from a telescope in Arizona (KPNO), a primary computing center in California (NERSC), and a worldwide collaboration of scientists. DESI is currently limited in its ability to utilize non-NERSC DOE computing facilities for a variety of reasons, including: [Section 4.6]
    - The ability to manage accounts for hundreds of users across different sites
    - Creating a system to synchronize PB-scale datasets of millions of files between multiple sites
- Raw data are directly downloaded from KPNO to NERSC, which routes via the University of Arizona in Tucson where it picks up Internet2. The bandwidth limited step is the KPNO to Tucson 1 Gbps fiber link leased from the Tohono O'odham Utility Authority. [Section 4.7]
- As a remote mountain site, KPNO has occasional network outages so a key operational requirement is that observations can continue without network connectivity. As an extreme case, in 2022 a wildfire melted the internet fiber optic cables to the telescope. Operations proceeded for several months with a minimal Starlink connection for instrument command and control, and external hard drives with data shuttled up and down the mountain by car. The hard drives were connected to computers in Tucson each day to initiate the data transfer to NERSC. [Section 4.7]
- DESI does not currently use cloud services for data processing or serving core data products. There has been exploratory work for sharing DESI data through the Amazon Open Data program, though this is not used by the collaboration itself. There has been interest in deploying the data-processing pipeline on AWS as an alternate site during NERSC outages, though data transfer and hosting costs make this prohibitive as a primary computing solution. [Section 4.8]
- DESI makes some use of AI/ML for its data-processing pipeline through fairly standard methods that do not require any special resources. There is more active cutting-edge work on using AI/ML for DESI science analyses, in particular using generative AI to create simulated/mock datasets faster than traditional simulations. At the same time, these use standard toolkits and NERSC Perlmutter computing and do not require novel hardware or tools. [Section 4.8]
- DESI has used Starlink at the remote telescope site as a back-up connection solution for when the primary internet is down. This was especially critical in 2022 after a wildfire resulted in a multi-month outage of the fiber optic connection to the mountain. The Starlink bandwidth was sufficient for command and control of the telescope, but was not sufficient for transferring observational data, which required driving hard disks up and down the mountain. [Section 4.8]
- The DESI data lifecycle involves all three of the IRI patterns: [Section 4.9]
    - Time-Sensitive analysis to detect operations problems and update the survey plan;
    - Data Integration–Intensive work to combine PB-scale data from multiple other collaborations and funding agencies for analysis;
    - Long-Term Campaigns to process data for the five-year DESI survey, with proposed extensions and follow-up programs through at least 2042.
- DESI uses Globus for transfers of collaboration data. DESI users report good experiences with Globus when they want to download an entire directory tree, but have found the interface lacking for the more common use case of selecting a subset of data from a complex directory tree, e.g., all files matching a given regex regardless of where they appear in the directory structure. [Section 5.3]

- Within the NERSC computing center, DESI is limited by I/O performance stability of the "community file system" — median performance is fine, but semi-rare outlier events can cause multi-minute delays or even lead to job hangs/timeouts. The Perlmutter scratch disk performance is better, but the difficulties of constantly syncing datasets between the scratch filesystem and CFS has been human-error prone and effort-prohibitive. [Section 5.4]
- The inability to pragmatically synchronize constantly evolving PB-scale datasets between multiple locations is a limiting factor for DESI. That includes internal to a data center (e.g., between CFS and the scratch filesystem) and external (e.g., between NERSC and another DOE HPC facility). [Section 5.4]
- Broadly speaking, DESI is not limited by raw bandwidth, but is limited by the convenience of tools for managing cross-site data flow. Globus is efficient for one-off data transfers, but issues such as expiring certificates, multi-factor-authentication, data discovery, and data filtering (i.e., selecting what needs to be downloaded) limit convenience for day-to-day random access of huge datasets. It is desirable to consider if computing centers and end users would be able to subscribe to datasets from each other, with automatic efficient syncing of the subsets of the data that are actively being used, with the toolkits managing purging unused data to stay within the local quota limits. The faster the network is, the more viable this becomes since it reduces the barrier to having to re-download data. [Section 5.5]
- Using a single centralized HPC facility to support a given HEP experiment is beneficial, since there's a single ecosystem to learn and it requires minimal overhead to share results with collaborators from other institutions. At the same time, it becomes a single point of failure during NERSC outages leading to ongoing interest in multisite computing operations in the future. The limiting factors are not networking for data transfer, but rather the tools for making that data sharing automated and transparent for hundreds of users accessing PBs of data stored in millions of files. DESI could transfer its major data release to another site, but that is of little use unless the hundreds of DESI users at NERSC can also get accounts at that site, and the outputs of their daily work can be transparently synced back and forth between the two sites. [Section 5.6]

6.2.2 DESI: Experiment Profile

Observations begin at the Mayall 4-meter telescope at KPNO outside of Tucson, AZ. Every ~15 minutes, DESI obtains spectra of 5000 astronomical objects, resulting in 1–2 GB of data per exposure. These data are transferred in semi-real time to the NERSC in Berkeley, CA, where they are processed in a real-time queue such that results from each night are available to collaboration scientists each morning. Processed data are ~10 GB/exposure. Results from each night are used to update the observing plan for the following night. Real-time quality assurance (QA) analysis is performed at both KPNO and NERSC to guide observers throughout the night. Hosting the QA at KPNO allows DESI to be robust to network and NERSC outages, while simultaneously processing at NERSC enables read-only access for all collaborators without requiring secure access to the mission-critical operations at KPNO.

6.2.2.1 Science Background

NERSC is also used as the primary scientific analysis center for the ~1000-person international collaboration of scientists, who access the data at NERSC for the cosmology analyses. Every one to two years, DESI reprocesses all the data taken to date using the latest data-processing code and calibrations. These serve as reference datasets for cosmology papers and are publicly released six months after the key cosmology projects have been published. DESI projects that the total data volume will grow to O(10 PB) by the end of the survey. The primary data access mode is for collaborators to directly login to NERSC and perform their work there. Some collaborators download a subset of the data to their local institutions for analysis. Public (noncollaborator) access is only through downloads from NERSC, or through Jupyter servers at the NOIRLab AstroData Lab, who host a subset of the public DESI data.

6.2.2.2 Collaborators

DESI is an international collaboration with O(1000) collaborators from over 70 institutions in 10 countries spanning North and South America, Europe, Asia, and Australia. The primary data access and sharing occurs through direct logins to NERSC. This is motivated both by the scale of the data (PBs), and through the need for a centralized location to share results between collaborators from many institutions.

Some DESI collaborators also download subsets of the data for analysis at their local institutions. Small volumes of data are typically accessed via scp or https, while more expert users download larger datasets using Globus.

Public (noncollaborator) access is either through direct download from NERSC (Globus or https), or through Jupyter Notebooks at the NOIRLab AstroData Lab, who host a subset of the DESI public data.

A mirror of the raw data is kept at the Astro Data Lab at the NSF's National Optical Infrared Astronomy Research Laboratory (NSF's NOIRLab) in Tucson, AZ, and a mirror of a subset of the processed data is kept at the University of Utah. These provide a geographically separated copy of the data for safety in case of a catastrophic event at NERSC, but they are not used for primary analyses.

Most of DESI's data are still proprietary, though an "Early Data Release" in 2023 provided pre-survey data to the scientific community. Future data releases (approximately every two years) will expand the user base of DESI data.

User/ Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
KPNO (NEAR TUCSON, AZ)	Raw data origin	rsync	40 TB of most-recent data	10-minute intervals	minimal, via subversion (svn) and rsync	Incompatible multisite security policies (VPN, MFA, firewalls) required backdoor exceptions
DESI COLLABORATION NERSC (BERKELEY, CA)	Primary	Direct login, Jupyter, Globus, https	Growing to ~10 PB by 2026	Raw data received every 10 minutes then processed at NERSC	Minimal, via svn and rsync	Collaborators from certain countries are not allowed to have accounts at NERSC, thus limiting their data access to the primary data site. Downtimes and short mean- time-to-failure affect productivity of collaborators using data directly at this site
DESI COLLABORATION HOME INSTITUTIONS (WORLDWIDE)	Secondary	scp, rsync, https, Globus	Unknown, ad hoc	Ad hoc	No	Need better tools
DESI OPERATIONS NERSC HPSS TAPE (BERKELEY, CA)	Secondary	HPSS TAR (HTAR)	Growing to ~10 PB by 2026	Daily backups of raw and processed data	No	
NOIRLAB OPERATIONS (TUCSON, AZ)	Secondary	rsync	~100 TB	Daily backup of raw data	No	

DESI OPERATIONS PROCESSED DATA MIRROR (UNIVERSITY OF UTAH)	Secondary	rsync, https	200 TB; may increase to 1 PB	Daily backup of processed data	No	
NON-DESI PUBLIC ACCESS NERSC DOWNLOADS (BERKELEY, CA)	Primary	https, Globus	ad hoc access	daily	No	Need better tools
NON-DESI PUBLIC ACCESS NOIRLAB ASTRO DATA LAB (TUCSON, AZ)	Secondary	Data portal; Jupyter	~100 TB	Every 1–2 years for public releases	No	

Table 6.2.1: DESI Data Mobility

6.2.2.3 IRI Readiness

The DESI data lifecycle involves all three of these patterns: Time-Sensitive analysis to detect operations problems and update the survey plan; Data Integration–Intensive work to combine PB-scale data from multiple other collaborations and funding agencies for analysis; and Long-Term Campaigns to process data for the five-year DESI survey, with proposed extensions and follow-up programs through at least 2042.

DESI is fundamentally a multisite experiment, with data from a telescope in Arizona, a primary computing center in California (NERSC), and a worldwide collaboration of scientists. DESI is currently limited in its ability to utilize non-NERSC DOE computing facilities for a variety of reasons including

- Account management for hundreds of users across different sites (ideally this would be a single approval + account, allowing access to all sites with the same credentials)
- How to real-time sync PB-scale datasets of millions of files between multiple sites
- Ecosystem differences in queue policy, web servers, Jupyter access, etc., i.e., a workflow (or even a single Jupyter Notebook) that works at one site requires significant overhead to port to another DOE facility.

6.2.2.4 Instruments and Facilities

Data originates from the DESI spectrographs at the Mayall 4-meter telescope at the KPNO near Tucson, AZ. Raw data are transferred to NERSC for processing and science analysis, with subsets of the processed data downloaded to individual institutions for further customized/ad hoc analysis. Raw and processed data are backed up to NERSC's HPSS tape system. Raw data are also mirrored to NOIRLab in Tucson, AZ and a subset of processed data are mirrored to the University of Utah as geographically separated sites from NERSC. Raw data are retained at KPNO until they are verified to be at NERSC disk, NERSC tape, and NOIRLab disk.

Raw data are directly downloaded from KPNO to NERSC, which routes via the University of Arizona in Tucson where it picks up Internet2. The bandwidth limited step is the KPNO to Tucson 1 Gbps fiber link leased from the Tohono O'odham Utility Authority, though this is sufficient for the needs of DESI and other telescopes operating at KPNO. A small cluster of computers at KPNO supports observations, but this is not provided as a general resource to the collaboration for data analysis. As a remote mountain site, KPNO has occasional network outages so a key operational requirement is that observations can continue without network connectivity. As an extreme case, in 2022 a wildfire melted the internet fiber optic cables to the telescope. Operations proceeded for several months with a minimal Starlink connection for instrument command and control, and external hard drives with data shuttled up and down the mountain by car. The hard drives were connected to computers in Tucson each day to initiate the data transfer to NERSC.

At NERSC, the data volume is projected to grow at a rate of ~1 PB/year from 2020 to 2025, reaching ~10 PB by the end of 2026. The data are stored in a hierarchy of thousands of directories with millions of files. Extensions to the DESI have been proposed and received positive reviews; if approved these would continue the data growth at a similar rate through the mid-2030s.

Public datasets will be curated beyond the timescale of DESI by the Astro Data Lab at NOIRLab. The intention is to also host the public DESI data via the Cosmology Data Repository at NERSC, though long-term funding commitment for that is not guaranteed.

DESI uses the full ecosystem of services available at NERSC: data transfer nodes; the real-time queue for rapid processing; the interactive queue for data analysis by collaborators; the regular queue for large reprocessing runs; database servers; webservers for data distribution and visualization; and a Jupyter server for data exploration.

6.2.2.5 Generalized Process of Science

Science analyses using DESI data are primarily performed at NERSC, with the scale of analyses varying from single-core Jupyter Notebooks to large parallel-batch jobs. Networking is used to get raw data to NERSC, to make off-site backups of processed data, and for some users to download subsets of the data to their local institutions. Many of the analyses also make use of simulated datasets. Most of these are directly generated at NERSC, while some others are generated elsewhere (e.g., Oakridge) and transferred to NERSC.

Using a single centralized facility provides benefits of a single ecosystem to learn, and minimal overhead for sharing results with collaborators from other institutions. At the same time, it becomes a single point of failure during NERSC outages leading to ongoing interest in multisite computing operations in the future. The limiting factors are not network bandwidth for data transfer, but rather the tools for making that data sharing automated and transparent for hundreds of users accessing PBs of data stored in millions of files. E.g., DESI could transfer its major data release to another site, but that is of little use unless the hundreds of DESI users at NERSC can also get accounts at that site, and the outputs of their daily work can be transparently synced back and forth between the two sites.

6.2.2.6 Remote Science Activities

The data are obtained at KPNO (near Tucson, AZ), processed at NERSC (in Berkeley, CA), and accessed by a worldwide collaboration, so the workflow is fundamentally and permanently multisite. Although the total data volume will grow to be rather large at NERSC (~10 PB), the real-time bandwidth requirements to support the operations are relatively modest (e.g., transferring a burst of 1 to 2 GB every 15 minutes from KPNO to NERSC throughout the night). Remote collaboration access to the data is more limited by the convenience of the tools than raw network bandwidth.

6.2.2.7 Software Infrastructure

Data transfer from KPNO to NERSC is performed with rsync spawned by a custom daemon-like python script that checks for new data every 10 minutes. Backups to NERSC HPSS tape use HPSS Tape Archiver (HTAR). Transfers of processed data from NERSC to the NOIRLab and University of Utah mirrors use Globus for the initial bulk transfer, then rsync to pick up any later differences. Transfers of data within the NERSC center use a combination of Globus, rsync, and Unix cp.

Collaborators primarily access the data through direct login to NERSC, or through NERSC's Jupyter server. Many collaborators have expressed a preference for downloading subsets of data to use locally, though DESI lacks developed tools for easy management of that remote analysis workflow. This is currently achieved through ad hoc downloads with Globus, scp, or https.

Data processing is performed at NERSC using Python scripts, leveraging numpy, scipy, cupy, numba, and mpi4py for efficiency and parallelism. The workflow is entirely file-based and does not rely upon a centralized database or workflow server for job coordination.

6.2.2.8 Network and Data Architecture

Figure 6.2.1 shows the network diagram connecting the DESI instrument (lower left) at the Mayall telescope to the NERSC computing center (upper right). DESI instrument operations and Mayall telescope operations are isolated in secure critical zones with very limited access. These route off the mountain via 1 Gb connections to the Tucson Data Center, where they pickup 10 Gb connections to NERSC.

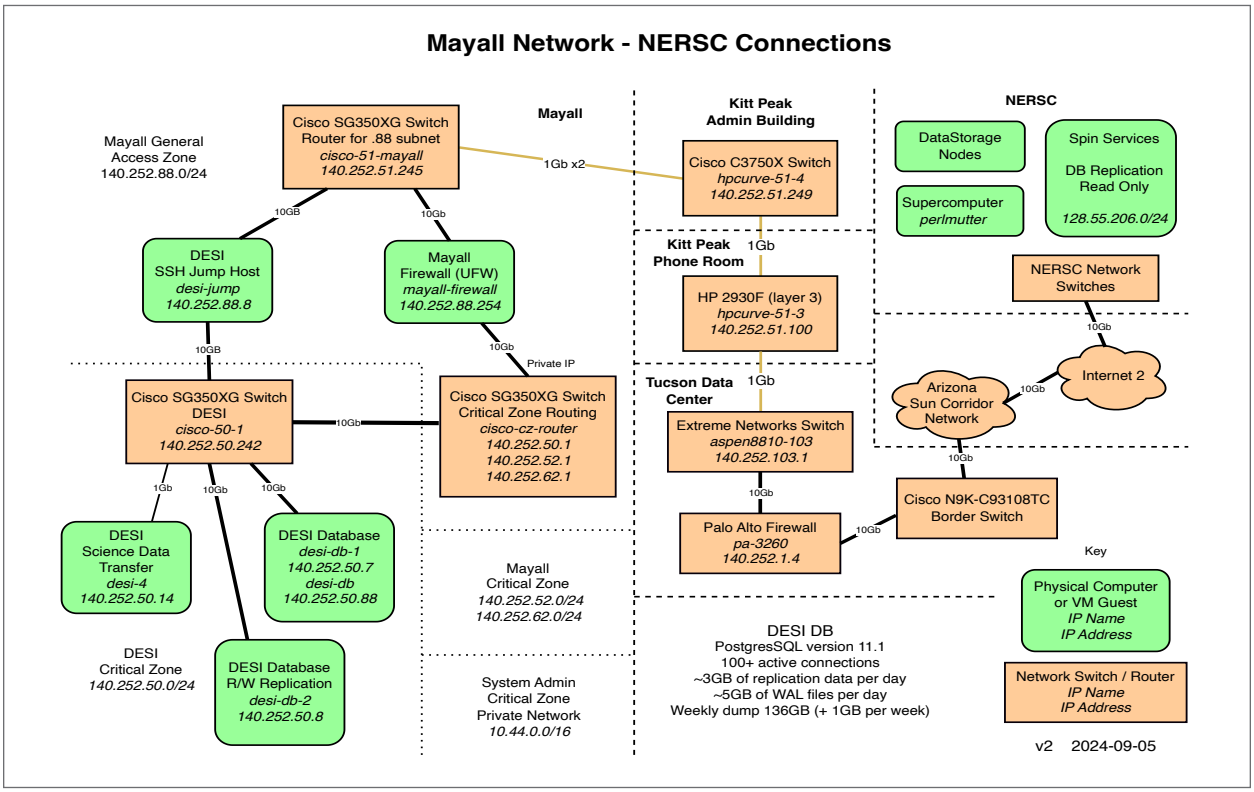


Figure 6.2.1: DESI Network Diagram

6.2.2.9 Emerging Technology Use

6.2.2.9.1 Cloud Services

DESI uses AWS for hosting their wiki, svn repository, mailing lists, document repository, publication database, desi.lbl.gov website, and a public-facing version of the imaging data browser.[17]

DESI does not currently use cloud services for data processing or serving core data products. There has been exploratory work for sharing DESI data through the Amazon Open Data program, though this is not used by the collaboration itself.

There has been interest in deploying the data-processing pipeline on AWS as an alternate site during NERSC outages, though data transfer and hosting costs make this prohibitive as a primary computing solution.

17 http://viewer.legacysurvey.org

6.2.2.9.2 AI and ML

DESI makes some use of AI/ML for its data-processing pipeline through fairly standard methods of random forests and convolutional neural networks that do not require any special resources. The most custom aspect is that DESI wrote its own pure-numpy evaluators of these models that do not rely upon external libraries like scikitLearn or TensorFlow. This provides long term (5–10+ years) reproducibility in a way that is not possible with the rapidly evolving landscape of AI/ML libraries that do not guarantee decade-level backwards compatibility.

There is more active cutting-edge work on using AI/ML for DESI science analyses, in particular using generative AI to create simulated/mock datasets faster than traditional simulations. At the same time, these use standard toolkits and NERSC Perlmutter computing and do not require novel hardware or tools.

6.2.2.9.3 Wireless or Distributed Location Science Activities

DESI has used Starlink at the remote telescope site as a back-up connection solution for when the primary internet is down. This was especially critical in 2022 after a wildfire resulted in a multi-month outage of the fiber optic connection to the mountain. The Starlink bandwidth was sufficient for command and control of the telescope, but was not sufficient for transferring observational data, which required driving hard disks up and down the mountain.

6.2.2.10 Data-Related Resource Constraints

Within the NERSC computing center, DESI is limited by I/O performance stability of the "community file system" — median performance is fine, but semi-rare outlier events can cause multi-minute delays or even lead to job hangs/timeouts. The Perlmutter scratch disk performance is better, but the difficulties of constantly syncing datasets between the scratch filesystem and CFS has been human-error prone and effort-prohibitive.

More generally, the inability to pragmatically sync constantly evolving PB-scale datasets between multiple locations is a limiting factor for DESI. That includes internal to a data center (e.g., between CFS and the scratch filesystem) and external (e.g., between NERSC and another DOE computing facility).

Although astronomy datasets are made publicly available, this is done through https servers providing individual files for download, not through systems designed for accessing and jointly processing entire datasets from different sites, and it is rare for data portals to provide any significant amount of computing resources to process their data by external users. This leaves users on their own to download and process data at different data centers in an uncoordinated manner.

6.2.2.11 Data Mobility Endpoints

DESI uses the "NERSC DTN" Globus endpoint for transfers of proprietary collaboration data, and the "DESI Public Data" Globus endpoint (also served from NERSC) for public data.

Overall users report good experiences with Globus when they want to download an entire directory tree, but have found the interface lacking for the more common use case (for DESI users) of wanting to download a subset of data from a complex directory tree, e.g., all files matching a given regex regardless of where they appear in the directory structure.

6.2.2.12 Outstanding Issues

Broadly speaking, DESI is not limited by raw bandwidth — organizations like ESnet and Internet2 have stayed ahead of the curve such that bandwidth is not a problem — Thanks! DESI is limited by the convenience of tools for managing cross-site data flow. Globus was a huge step forward for efficient one-off data transfers, but issues such as expiring certificates, multi-factor-authentication, data discovery, and data filtering (i.e., selecting what needs to be downloaded) limit its convenience for day-to-day random access of huge datasets.

Ideal would be something like Apple's iCloud Photos — devices have access to a library of photos much larger than can fit on the device, and all syncing back and forth between the device and the iCloud storage happens transparently to the user. Ideally computing centers and end users would be able to subscribe to datasets from each other, with automatic efficient syncing of the subsets of the data that are actively being used, with the toolkits managing purging unused data to stay within the local quota limits. The faster the network is, the more viable this becomes since it reduces the barrier to having to re-download data.

6.2.2.13 Facility Profile Contributors

DESI Representation

Stephen Bailey, LBNL, StephenBailey@lbl.gov

ESCC Representation

Richard Simon, LBNL, rsimon@lbl.gov

Rune Stromsness, LBNL, rstrom@lbl.gov

6.3 Vera C. Rubin Observatory

Rubin Observatory will carry out the LSST, using the Simonyi Survey Telescope and the Rubin Observatory LSST Camera. It will take repeated images over more than 18,000 deg^2 of the southern sky in six broad-band optical filters (ugrizy). The resulting dataset will provide a static census of approximately 40 billion objects as well as a dynamic time-domain census over an unprecedented range of time scales and flux limits. The primary deliverables for Rubin Observatory include a real-time stream of "alerts" of transient events, prompt data products, annual data-release data products, and a science platform specifically developed to reduce the barrier of entry to Rubin Observatory data and to shorten the path to science for the user community.

6.3.1 Discussion Summary

Rubin does face several challenges, some of which are outside of the control of ESnet and DOE. Like other large facilities, they must take cybersecurity into consideration. NOIRLab has been responsive to recent threats, and has put in place controls that sometimes make collaboration harder. In addition, the theme of federated identity management is also present for Rubin, and represents a barrier to the use of some tools. Rubin is looking to adopt some of the tools used elsewhere in HEP for data mobility and workflow management, including Rucio. The steps to integrate this work in their own environment have been challenging, and Rubin will be looking for additional assistance and advice.

ESnet will follow up with international partners to include up-to-date networking maps of the corresponding international paths.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- Rubin Observatory will carry out the LSST, using the Simonyi Survey Telescope and the Rubin Observatory LSST Camera. It will take repeated images of the southern sky, and the resulting dataset will provide a static census of approximately 40 billion objects. The primary deliverables for Rubin Observatory include a real-time stream of "alerts" of transient events, prompt data products, annual data-release data products, and a science platform. [Section 4.1]
- Rubin Observatory is an NSF-DOE funded partnership. Operations partners are NSF's NOIRLab and SLAC, whose managing organizations are Association of Universities for Research in Astronomy, Inc. (AURA) and Stanford University, respectively. There are four operations affiliate partners: [Section 4.1]
    - Chile, host country of the telescope facility
    - Brazil, providing high-bandwidth network connectivity from Santiago to Saõ Paulo
    - CC-IN2P3 physics institute in France, providing 40% of the annual data-processing capacity
    - IRIS Network in the United Kingdom, providing 25% of the annual data-processing capacity
- The operations of Rubin Observatory are broken into three phases: [Section 4.1]
    - FY19–FY25 (pre-operations)
    - FY26–FY35 (main survey)
    - FY36–FY37 (post operations)

- Rubin pre-operations ran early versions of all the data production elements, and between 2021–2023 some of this was done in a commercial cloud (e.g., the IDF) deployment using precursor data sets. Transition from the IDF to the USDF was completed in FY24. Pre-operations will culminate with data production at full-scale testing in FY25. The full LSST will run for 10 years and final data processing will take place after that for two years. Survey data taking ends in FY35. [Section 4.1]
- Rubin expects ~30 TB of continuous image and metadata to flow from Chile to the US each night. Real-time transfer over dedicated high-bandwidth segments is required to analyze each image at the USDF (hosted at SLAC National Accelerator Laboratory) and broadcast alerts to the community. [Section 4.2]
- The USDF at SLAC will curate 5 PB/yr of data, with ancillary and other image data products that could add up to 500 PB by end of survey in 2035. [Section 4.2]
- Rubin Observatory data are collected from the 3.2 gigapixel LSSTCam mounted on the 8.4m Simonyi Survey Telescope located on the summit of Cerro Pachón in Chile. Data are transported via an encrypted tunnel over a long-haul network from the observatory summit to the USDF at SLAC. Data are routed at modest bandwidth over time to the French Data Facility at IN2P3 in Lyon for geographically distributed backup. [Section 4.3]
- Rubin alerts are required to be broadcast via community-developed brokers in 60 seconds from the time the camera shutter closes on the mountain top in Chile. Other "prompt data products" are made available to the community after an 80-hour embargo period. The embargo period is implemented to prevent prompt detection of artificial satellites in the LSST data stream. [Section 4.3]
- All image data will be carried to the project's Archive Center at the USDF, hosted within the SLAC Shared Science Data Facility (S3DF). The total northbound data from Rubin to the Archive Center is approximately 30 TB/night. A very demanding application using these flows is Alert Production, where 12.7 GB (uncompressed) of pixel data need to be transferred from the Base Facility in La Serena Chile to the Archive Center in 7 seconds bursts. There are two such bursts approximately every minute of nighttime observing. The bursts need to be sustained for approximately 355 days in each year. The required reliability level for these flows is very high, and is seen to require a path-diverse network with nearly lossless networking. [Section 4.3]
- The Rubin prompt processing will generate a nightly stream of sources which vary in brightness. This stream will be made available to at least seven community brokers. These brokers allow (outside Rubin) access to the alert stream including the tools needed to filter the stream for specific science cases. At least some of the brokers will be open to any member of the worldwide science community. [Section 4.3]
- Rubin community-provided alert brokers will be the worldwide community access avenue for alerts. At least one community alert broker will provide access to the worldwide community. In 2021, Rubin completed selection of the community alert brokers, these include seven international teams: ANTARES (NOIRLab), ALeRCE (Chile), Lasair (UK), Pitt-Google (U of Pittsburgh), Fink (France), AMPEL (Germany), BABAMUL (CALTECH). Two downstream brokers that operate off one of the primary brokers were also endorsed by the SAC: SNAPS (Northern Arizona University) and POI (Vanderbilt). [Section 4.3]
- The Rubin Data-Release Production system will be executed approximately annually , and will ingest and process all data which has been taken to date. By co-adding and combining data taken at multiple epochs, it will provide a more detailed characterization of fainter astrophysical objects than is provided by the Prompt Data Products. [Section 4.3]

- Over its 10 years of planned operations, Rubin Observatory will produce annual data releases. Each data release will include reprocessing of all data from the start of the survey up to the cutoff date for that release. [Section 4.3]
- Given the volume of data involved, it is not feasible to make all Rubin releases available on fast storage for the duration of the project. Instead, Rubin will keep the two most-recent releases available for rapid access, and archive older releases to tape. Users will not be able to perform database queries against these archived releases. However, they would be made available for bulk download. [Section 4.4]
- In FY24 the IDF transitioned to the hybrid-cloud US DAC model, with the user-facing RSP remaining in Google Cloud and the back-end data holdings and data services hosted at the USDF at SLAC. The RSP is implemented in a virtualization framework that is portable across cloud providers or Kubernetes-based on-premises infrastructure. Multiple other scientific services, applications, and databases (including a staff-facing instance of the RSP) run within a Kubernetes cluster environment at the USDF. [Section 4.4]
- Rubin requires a constant 1,188 CPU cores throughout the survey period to perform Prompt Processing. These cores will be used during the night for alert generation, and during the day for catch-up (where necessary) and solar system processing. Each core will have access to 5 GB of memory, and all these cores are located at the USDF. [Section 4.4]
- The RSP is a set of web applications and services that will be used by the data rights community to access, visualize, interact with, and analyze Rubin Observatory data. The RSP is being designed to be simple enough to engage general users and flexible enough to meet the needs of experienced users. Data will be exposed to users through the web portal (interactive use cases), the Jupyter Notebook interface (custom user use cases), and machine-accessible application program interfaces (APIs; remote tools/analyses use cases). [Section 4.5]
- All the Rubin collaborations, as well as independent scientists, will access and analyze data through either the hybrid cloud-based US DAC (cloud-based RSP and USDF back-end data services) or IDACs operated by in-kind contributor communities. There will also be a data access center (DAC) at the base facility in Chile. Chileans will have priority access to this DAC, but others can use it as well. [Section 4.5]
- The USDF will provide raw and processed data to Rubin Observatory authorized users, including the DOE Dark Energy Science Collaboration and their international collaborators, including science collaborations investigating Dark Energy in other participating countries. The USDF will distribute data to one or more data access centers in Europe and elsewhere. The scope of these centers and their reliance on the USDF is currently being worked out. It is worth noting that the data volume of an annual release is very significant, with just the catalogs of detected objects beginning at PB-scale. [Section 4.5]
- Nightly data from the Rubin Observatory are accumulated and processed with science pipelines at the USDF as well as partner data facilities in France and the UK, producing deep stacked images and source catalogs. These "annual data-release" products are served to the user community via a RSP hosted in Google Cloud and backed by data-access services hosted at the USDF. [Section 4.6]
- Rubin expects a significant amount of data will be moved in bulk to other Afs. Data will be moved back and forth between the USDF and the FrDF+ UK Data Facility (UKDF) in conjunction with the distributed data-release processing efforts. [Section 4.6]

- Rubin also includes a program of in-kind contributions from international participant communities, some of which take the form of IDACs. Additionally, large amounts of data will be moved to other computing facilities to support custom analysis by large science collaborations. In particular, for the DESC this will likely amount to the mirroring of key data products from each annual data release from SLAC to NERSC. Additional mirroring efforts to support science analysis may be initiated for centers such as ALCF, OLCF and TACC. Rubin anticipates a key role for ESnet in supporting these data transfers. [Section 4.6]
- International networking has a role in supporting Rubin Observatory LSST production processes. LSST annual releases will be computed at the USDF and at the satellite computing facilities in France and the UK. This generates a need for networking to support continuous data production, in addition to distribution of data for scientific data analysis. By agreement, CC-IN2P3 is the responsible institution for arranging connectivity between CC-IN2P3 and the US. [Section 4.7]
- A memorandum of agreement (MOA) has been established between IN2P3, the LSST Corporation (LSSTC), the Rubin Observatory, and the University of Illinois for the IN2P3 Computing Center (CC-IN2P3) in Lyon, France to provide 40% of the computing capacity needed for annual data-release processing during LSST operations and to host a complete backup of the Rubin Observatory data archive. The MOA requires that France will provide the TA network connectivity to ESnet needed to support data transport to and from CC-IN2P3. [Section 4.7]
- The UKDF will be a distributed entity. Its computing hardware will be provisioned through IRIS, a national organization providing computing resources for projects funded by the STFC. RIS sites are well integrated into national and international networks, for the transfer of LHC and other data, and have a wealth of experience supporting a wide range of international science collaborations. The UKDF will be hosted by the University of Edinburgh. [Section 4.7]
- The required peak bandwidth from the Base Facility in La Serena to the USDF is established by the need to transfer the images for prompt processing. In order for the USDF to release alerts of transient and variable sources within 60 seconds of image readout from the camera electronics, seven seconds of this time budget is allocated to data transfer, including compression/decompression. Two seconds of that budget is in the Summit to Base transfer, and the other five seconds is in the Base to Archive Transfer and compression/decompression. This equates to a minimum of 40 Gbps bandwidth. [Section 4.7]
- Currently the Rubin Data Management system does not make heavy use of AI/ML for data production or distribution. The scientific community is expected to deploy their own AI/ML techniques in analysis of Rubin's scientific data products for modeling, inference, and anomaly detection. This expectation is a major driver for bulk download capabilities for Rubin data products, and may also influence the development roadmap for Rubin-supported analysis platforms on five-year timescales. Multiple community alert brokers currently provide users with AI/ML capabilities for classification and anomaly detection. [Section 4.8]
- The Rubin-LSST science enterprise exhibits all three IRI patterns to some degree: [Section 4.9]
    - Time Sensitive: Rubin images are transferred in near-real time from the Summit to the USDF, where they are promptly processed to generate alerts on objects that have changed in brightness, moved, or newly appeared.
    - Data Integration Intensive: Many Rubin science analyses will require extensive integration between multiple data sets both internal to Rubin (images, catalogs, light curves) and external (data from complementary surveys at other wavelengths, spectroscopic facilities,

and space-based telescopes). This data integration is likely to involve a range of different Afs both within the DOE system (e.g., SLAC, NERSC, ALCF) and beyond (e.g., partner data-processing facilities in France and the UK, NSF-funded computing centers like TACC, NASA archive and science processing centers hosting data from the Roman Space Telescope, and Rubin IDACs in other partner countries).

    - Long-term Campaign: Rubin has a 10-year primary survey mission, and a potentially longer horizon in an extended mission mode. Successful scientific operation over this horizon requires sustained access and integration between Rubin's experimental-support computing facility at SLAC and the DESC analysis facility at NERSC.

- Rubin operations employs a model of distributed data management and processing across three main sites (USDF, FrDF, UKDF). The current baseline plan calls for using a combination of the Rucio software system to coordinate multisite data movement and the PanDA software system for orchestrating distributed processing jobs. This use case is highly aligned with the vision for multisite interoperability under IRI, and Rubin expects to follow IRI development both as an exemplar experiment and as a future beneficiary of new infrastructure in the future. [Section 4.9]
- Known or potential data-related resource constraints for Rubin as of FY24 are: [Section 5.3]
    - Insufficient computing resources available to meet the full analysis demand from the Rubin scientific user community, and/or difficulty in co-locating Rubin data products with available computing resources. Includes unbounded/unscoped demand for AI/ML-based analysis of Rubin data.
    - Changing cybersecurity threat and policy landscape can present barriers to planned data-access and data-sharing services, and/or can introduce new costs, technical requirements, and schedule impacts.
    - Challenges associated with implementing and routinely operating the distributed processing model for annual data-release processing.
    - Challenges and constraints of operating in a shared data facility with other experiments.
    - Network and I/O performance impacts of transferring a very large number of relatively small files.

6.3.2 Vera C. Rubin Observatory: Facility Profile

Rubin expects ~30 TB of image and metadata to flow from Chile to the US each night. Real-time transfer over dedicated high-bandwidth segments is required to analyze each image at the USDF (hosted at SLAC National Accelerator Laboratory) and broadcast alerts to the community of all sources in the image that have varied in brightness by more than 5 sigma compared to a reference image. Alerts are required to be broadcast via community-developed brokers that receive a stream from Rubin in 60 seconds from the time the camera shutter closes on the mountain top in Chile. Other "prompt data products" are made available to the community after an 80-hour embargo period. The embargo period is implemented to prevent prompt detection of artificial satellites in the LSST data stream.

6.3.2.1 Science Background

Nightly data are accumulated and processed with science pipelines at the USDF as well as partner data facilities in France and the UK (FrDF and UKDF respectively), producing deep stacked images and source catalogs. These "annual data-release" products are served to the user community via a RSP hosted in Google Cloud and backed by data-access services hosted at the USDF.

Rubin expects a significant amount of data will be moved in bulk to other Afs. First, large amounts of data will be moved back and forth between the USDF and the FrDF+UKDF in conjunction with the distributed data-release

processing efforts. Second, Rubin includes a program of in-kind contributions from international participant communities, some of which take the form of IDACs in those countries serving copies of either subsets or full copies of LSST data releases. Third, large amounts of data will be moved to other computing facilities to support custom analysis by large science collaborations. In particular, for the DESC this will likely amount to the mirroring of key data products from each annual data release from SLAC to NERSC. Additional mirroring efforts to support science analysis may be initiated for centers such as the ALCF and the TACC. Rubin anticipates a key role for ESnet in supporting these data transfers.

All Rubin data are proprietary for two years except the alert stream which is world public. After two years, the data are shareable with any one, but no specific mechanism for making these data public and serving them is yet defined.

6.3.2.2 Collaborators

Rubin Observatory is an NSF-DOE funded partnership. Operations partners are NSF's NOIRLab and SLAC whose managing organizations are AURA and Stanford University, respectively. Rubin counts with four operations affiliate partners:

1. Chile, host country of the telescope facility
2. Brazil, providing high-bandwidth network connectivity from Santiago to Saõ Paulo
3. CC-IN2P3 physics institute in France, providing 40% of the annual data-processing capacity
4. IRIS Network in the United Kingdom, providing 25% of the annual data-processing capacity

The primary science community includes data scientists, astronomers and physicists in the USA and Chile. International participants in Europe, Asia, and North America who have approved in-kind contributions to the Rubin science enterprise will also be collaborators with equal access and rights to work with Rubin data. All these collaborators can work independently or through self-governed science collaborations. Presently, Rubin engages with eight such science collaborations who largely drive the discussion about the science needs of the community.

Depending on the size of the in-kind contribution, International members may have agreements with DOE or NSF directly or with AURA or SLAC.

The total worldwide Rubin science user community is estimated to comprise of order 10,000 individuals.

User/ Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
SUMMIT / LSSTCAM	Yes — primary initial copy, retained for ~30 days	Quick-look capabilities for observing staff	30 TB / night	Continuous transfer during the night to SLAC, after shutter closes on each exposure	No	Per data security requirements, data must be transferred over IPSec encrypted tunnel to SLAC
USDF (LOCATED AT SLAC)	Yes — primary archival copy	Cloud-based portal for all users backed by on-prem data holdings; real-time broadcast of alert packets to selected brokers	5 PB/yr. With ancillary and other image data products, up to 500 PB by end of survey in 2035	Nightly raw images stored. Weekly/ monthly calibration data. Annual data-release data products	Processed data sent to data center in Chile; calibration and QA data sent to Summit	Large number of small files may impact network transfer performance

CHILE	Yes	Data Portal for Chilean users	5 PB/yr	Nightly raw images stored. Weekly/monthly calibration data. Annual data-release data products	No	No
BRAZIL	No	None, data only transits	5 PB/yr	Same as Chile (same data transits Brazil)	Same as USDF (same data transits Brazil)	No
CC-IN2P3 / FRDF	Yes — secondary copy	Portal (possibly), default is USDF	5 PB/yr	Nightly raw images sent at low or modest rate after embargo period, plus additional data as part of distributed annual data-release processing	40% of annual data-release products generated here and returned to USDF	No
IRIS / UKDF	No	Full IDAC user interface	few PB/yr	Data transferred from USDF to UKDF to enable (a) 25% of annual data-release processing and (b) full IDAC for annual data-release	25% of annual data-release products generated here and returned to USDF	UK serves as both processing partner and full IDAC
OTHER "LITE" IDACS (IN-KIND): ARGENTINA, AUSTRALIA, BRAZIL, CANADA, CROATIA, DENMARK, JAPAN, MEXICO, POLAND, SLOVENIA, SPAIN, SOUTH KOREA	No	Varies, but all provide some form of user interface or login	Typically, few 100's of TB/yr each	Annually, subset of each data release ("lite" rather than "full" IDAC in most cases)	No	Large number of destination sites with heterogeneous local resources and environments
DESC / NERSC	No	Not Rubin supported	few PB/yr	Modest transfer rate for fraction of annual images	Some custom data products could be sent back to USDF for users generally.	Possible policy incompatibility (certificates, auth) with Rucio data transfer tool
OTHER ANALYSIS FACILITIES	No	Varies by facility (not Rubin supported)	Up to few PB/yr	Roughly annual transfer in conjunction with data releases	No	Possible policy incompatibility (certificates, auth) with Rucio data transfer tool
RUBIN OBSERVATORY HQ TUCSON	No	No general data access, observatory science and management activity	few TB/week	Few TB/week	No	No

Table 6.3.1: Vera C. Rubin Observatory Data Mobility

6.3.2.2.1 North-South Networking

All image data will be carried to the project's Archive Center at the USDF, hosted within the SLAC Shared Science Data Facility (S3DF). The total northbound data from the observatory to the Archive Center is

approximately 30 TB/night. A very demanding application using these flows is Alert Production, where 12.7 GB (uncompressed) of pixel data need to be transferred from the Base Facility in La Serena Chile to the Archive Center in 7 seconds bursts. There are two such bursts approximately every minute of nighttime observing. The bursts need to be sustained for approximately 355 days in each year. The required reliability level for these flows is very high, and is seen to require a path-diverse network with nearly lossless networking.

This same network carries substantial traffic from the USDF to the Rubin Observatory Base Facility La Serena, in the form of Calibration data and annual Data Releases, but the aggregate volume is not as large as the South-North traffic, nor is there a near-real-time latency requirement for that traffic. Establishing duplex service at the level required for the South–North traffic should more than suffice for the North–South traffic. Lesser, but operationally significant volumes of data will flow from the Base Facility to the Rubin Observatory Headquarters site in Tucson and to other sites involved in Rubin operations.

Collaboration among the participating network operators supporting the Rubin North-South network connection is convened under the auspices of the Americas Lightpaths Express & Protect (AmLight-ExP) consortium and the South American–African Astronomy Coordination Committee (SA3CC).

6.3.2.2.2 National and International Networking

The USDF will provide raw and processed data to Rubin Observatory authorized users, including the DOE DESC and their international collaborators, including science collaborations investigating Dark Energy in other participating countries. The USDF will distribute data to one or more data access centers in Europe and elsewhere. The scope of these centers and their reliance on the USDF is currently being worked out. It is worth noting that the data volume of an annual release is very significant, with just the catalogs of detected objects beginning at PB-scale.

International networking has a role in supporting Rubin Observatory LSST production processes. LSST annual releases will be computed at the USDF and at the satellite computing facilities in France and the UK. This generates a need for networking to support continuous data production, in addition to distribution of data for scientific data analysis. By agreement, CC-IN2P3 is the responsible institution for arranging connectivity between CC-IN2P3 and the US.

6.3.2.3 IRI Readiness

The Rubin-LSST science enterprise exhibits all three IRI patterns to some degree:

- Time Sensitive: As soon as they are collected, Rubin images are transferred in near-real time from the Summit to the USDF, where they are promptly processed to generate alerts on objects that have changed in brightness, moved, or newly appeared. These alert packets are broadcast, also in near-real-time, to a network of "alert broker" software platforms in the community. These broker systems enable scientists to implement their own filters on the alert stream to find objects of greatest interest, and to trigger follow-up observing with other telescopes during the same night.
- Data Integration Intensive: Many Rubin science analyses will require extensive integration between multiple data sets both internal to Rubin (images, catalogs, light curves) and external (data from complementary surveys at other wavelengths, spectroscopic facilities, and space-based telescopes). This data integration is likely to involve a range of different Afs both within the DOE system (e.g., SLAC, NERSC, ALCF) and beyond (e.g., partner data-processing facilities in France and the UK, NSF-funded computing centers like TACC, NASA archive and science processing centers hosting data from the Roman Space Telescope, and Rubin IDACs in other partner countries). Rubin will also deliver a cornerstone dataset for the coming generations of multimodal AI foundation models in astronomy and cosmology.

- Long-term Campaign: Rubin has a 10-year primary survey mission, and a potentially longer horizon in an extended mission mode. Successful scientific operation over this horizon requires sustained access and integration between Rubin's experimental-support computing facility at SLAC and the DESC analysis facility at NERSC, as well as a long-term effort to evolve and maintain software and workflow tools to take advantage of changing technology platforms and opportunities. Rubin data itself must be curated and preserved over this same time frame, as each successive analysis campaign draws on the cumulative data set since the start of the survey.
- Rubin operations employs a model of distributed data management and processing across three main sites (USDF, FrDF, UKDF). The current baseline plan calls for using a combination of the Rucio software system to coordinate multisite data movement and the PanDA software system for orchestrating distributed processing jobs (see Rubin DMTN-213[18] for details). This use case is highly aligned with the vision for multisite interoperability under IRI, and Rubin expects to follow IRI development both as an exemplar experiment and as a future beneficiary of new infrastructure in the future.

In the near term, time-critical aspects of Rubin commissioning and early operations may limit the extent to which current project data management activities can engage with IRI prototype capabilities. Rubin should nevertheless serve as an important IRI use case, and major aspects of Rubin data management could be candidates for migration into IRI frameworks on the five-year timescale of Rubin's planned mid-survey hardware refresh.

6.3.2.4 Instruments and Facilities

Rubin Observatory data are collected from the 3.2 gigapixel LSSTCam mounted on the 8.4m Simonyi Survey Telescope located on the summit of Cerro Pachón in Chile.

Data are transported via an encrypted tunnel over a long-haul network from the observatory summit to the USDF at SLAC. Data are routed at modest bandwidth over time to the French Data Facility at IN2P3 in Lyon for geographically distributed backup.

The Rubin Observatory Data Management System and Facilities are described in detail in Sections 6.3.2.7 and 6.3.2.8 below, via excerpted content from the current detailed 10-year Operations Plan. To preserve subsection reference correctness, the original Chapter and section numbering is preserved in the excerpt content:

- Case Study Section 6.3.2.7 includes Operations Plan Chapter 18
  - Data and software products are described in Sec. 6.3.2.7.1.1
  - The RSP (primary science user interface) is described in Sec. 6.3.2.7.1.3
- Case Study Section 6.3.2.8 includes Operations Plan Chapter 17
  - Overall architecture is described in Sec. 6.3.2.8.1.1
  - The compute and storage sizing model is described in Sec. 6.3.2.8.1.2
  - Data Facilities in the US, Chile, and Europe are described in Secs 6.3.2.8.1.3–6.3.2.8.1.5
  - The Data Backbone is described in Sec. 6.3.2.8.1.6
  - Wide-area / long-haul networking is described in Sec. 6.3.2.8.1.1

The operations of Rubin Observatory are broken into three phases:

- FY19–FY25 (pre-operations)
- FY26–FY35 (main survey)
- FY36–FY37 (post operations)

18 https://dmtn-213.lsst.io/

Pre-operations will run early versions of all the data production elements described above. In 2021–2023 some of this was done in a commercial cloud IDF deployment using precursor data sets (simulated or from other telescopes). Transition from the IDF to the USDF was completed in FY24. Pre-operations will culminate with data production at full-scale testing in FY25, ideally using science validation survey data from the telescope and LSSTCam obtained in the last part of commissioning.

The full LSST will run for 10 years and final data processing will take place after that for two years. Survey data taking ends in FY35.

6.3.2.5 Generalized Process of Science

The Rubin LSST is a public survey. The Rubin team itself is not responsible for the delivered science. The DOE is chiefly involved in Rubin for dark matter and dark energy science. The DOE supports the DESC and their science is described in a separate case study.

For NSF, the science goals are based on studies of the nature and distribution of Solar System bodies, the halo of the Milky Way, galaxy formation and evolution, and the transient and variable universe. Each of these science areas has an associated science collaboration. These are self-governed and only DESC has a direct funding connection to one of the agencies (DOE).

All the collaborations as well as independent scientists will access and analyze data through either the hybrid cloud-based US DAC (cloud-based RSP and USDF back-end data services) or IDACs operated by in-kind contributor communities. There will also be a DAC at the base facility in Chile. Chileans will have priority access to this DAC, but others can use it as well.

All scientists will have compute and storage resources at the US DAC or Chilean DAC. Tools available will allow complex queries of the object and source catalogs as well as statistical and other common analyses.

The prompt processing will generate a nightly stream of sources which vary in brightness. This stream will be made available to seven community brokers. These brokers allow (outside Rubin) access to the alert stream including the tools needed to filter the stream for specific science cases. At least some of the brokers will be open to any member of the worldwide science community.

Solar System objects will be identified in the nightly processing and new and repeat observations of these bodies will be sent to the Minor Planet Center (MPC)[19] for further orbit determination and cataloging. Scientists can access the latest information from the MPC or the RSP.

A subset of Rubin Data will be managed by the Rubin and NOIRLab education teams. This cloud-based resource will support public engagement and citizen science.

6.3.2.6 Remote Science Activities

All observational data are collected at the Rubin Observatory summit facility on Cerro Pachón in Chile, which is remote from the main data archiving and processing site.

Most science users will do their data exploration and analysis through the RSP, hosted in Google Cloud and backed by data services through the USDF at SLAC. Users will access the RSP from their own local computers through an online portal.

Other science analysis campaigns will be conducted at IDACs in partner countries, or at other computing centers in the US such as NERSC.

The model of an observatory that is remote from the data facility is foundational to the Rubin experimental design and is not expected to change in the foreseeable future. The role of IDACs and other Afs can be expected to evolve and may be significantly different over strategic planning horizons (five years and beyond).

19 http://www.minorplanetcenter.org

6.3.2.7 Software Infrastructure

This section excerpts Chapter 18 of the current version of the 10-year Rubin Observatory Operations plan, describing the Observatory's data and software products. The content and internal references have been lightly edited to make this Case Study document self-contained, however the numbering of subsections and figures is preserved from the original document.

6.3.2.7.1 Data Products and Community Support

Rubin Observatory Data Management is required to deliver science-ready, calibrated data products to the Rubin Observatory data rights community (i.e., scientists at US and Chilean institutions and named international member scientists who receive access in return for contributions to Rubin Observatory operations). The Rubin Observatory data rights community will have access to all data products with no proprietary period. Anyone who is not a member of this data rights community will have limited rights to Rubin Observatory's data products, as described in Section 6.3.2.7.1.2. Rubin Observatory is also required to deliver the software used to generate the data products and the documentation necessary to understand and reproduce the generation of Rubin Observatory data products.

The data products to be produced by the Rubin Observatory Data Management System (DMS) are defined in the Rubin Observatory Data Products Definition Document (LSE-163[20]; the DPDD). Broadly, these are categorized into prompt data products, which are designed to enable rapid identification of transient and variable astrophysical objects, and detection and tracking of moving objects within the solar system, and data-release data products, which are generated annually with the goal of deep detection and precise characterization of astrophysical phenomena. Note that all prompt data products (with the exception of real-time alerts) are regenerated during data-release processing: it is expected that most scientific use cases, with the key exception of those which are dependent on low-latency notifications, will be best served by data-release products.

The fundamental inputs to both the prompt and data-release processing systems are single-visit images — which may be either "standard visits," consisting of a pair of successive 15 second exposures called snaps, or "alternate standard visits," consisting of a single 30 second exposure — delivered by Rubin Observatory LSST Camera to the DMS. Details of how these inputs are processed are provided in the subsequent sections.

After the Single Frame Processing step is completed, the pixel values and their variances are not modified any more. These processed visit images (PVIs) are used downstream to produce coadded and difference images. The rest of the processing is essentially a model-based interpretation of imaging observations that includes numerous astrophysical and other assumptions.

6.3.2.7.1.1 Data and Software Products

6.3.2.7.1.1.1 Prompt Products Production

6.3.2.7.1.1.1.1 Prompt Products Processing Sketch

Prompt Products Processing is designed to provide time-critical data products to the community at low latency. As such, it is designed to work on a per charge-coupled device (CCD) basis: as each image is received from the LSSTCam, it is rapidly transferred to the USDF and immediately processed. Processing of each image proceeds as follows:

- The image is processed by the Single Frame Processing pipeline. Single Frame Processing consists of an instrument signature removal (ISR) step, in which calibration reference data are used to remove artifacts resulting from the instrument or detector electronics, followed by determination of the image's background, point spread function, and photometric and astrometric calibration. The result of the Single Frame Processing pipeline is a PVI.
- The PVI is differenced against a deep "template," constructed from previous observations of the area being observed.

20 https://docushare.lsst.org/docushare/dsweb/Get/LSE-163

- Source detection is performed on the difference image, resulting in a collection of DIASources. The flux and shape of each DIASource is measured.
- DIASources are positionally associated with previous detections of both astrophysical (DIAObject) and solar system (SSObject) objects. If no association can be made, a new DIAObject is created.
- If a DIASource is associated with an SSObject it is flagged for further processing during the following day. At this point, artificial satellites are also largely filtered out, with any remaining confounding contaminants excluded from the alert stream by the Alert Verification System (AVS).
- Otherwise, the DIAObject — which provides aggregated and derived quantities describing the astrophysical object — is updated based on measurements made on the DIASource.
- An alert is issued, which includes the complete DIAObject or SSObject record, the DIASource record, and all previous measurements of that object from the last twelve months of observing.
- Within 24 hours of observing, precovery forced photometry will be performed on any difference image overlapping the position of new DIAObjects. Alerts will not be issued with precovery photometry information but the resulting DIAForcedSource measurements will be included in future alerts from this DIAObject.

Solar System Processing proceeds during the day. In brief, the daytime processing will:

- Analyze all DIASources seen during the previous night which are not associated with previously known objects in an attempt to fit tracks between them.
- Designate tracks which are consistent with Keplerian orbit around the sun as new SSObjects and submit them to the IAU MPC.
- Retrieve an updated orbit catalog for all known Solar System bodies from the MPC and recompute properties of all SSObjects.

6.3.2.7.1.1.1.2 Prompt Data Products: Alert Stream, Filtering Service, Images, Catalogs

The Prompt Products Processing system, sketched in the previous section, will identify transient, variable and moving objects observed by the Rubin Observatory. Examples of source types that will be detected by this process are (i) moving objects in the Solar System and (ii) supernovae superimposed on bright extended galaxies. In total, Rubin expects to detect around 10,000,000 such sources per night.

To enable follow-up observations, each of these 10,000,000 sources will be transmitted to seven Community Brokers via an "alert stream," where they will be accessible by the worldwide astronomical community with no proprietary period. Two specialized downstream brokers receive subsets of alerts from other brokers. Each alert is designed to provide as complete a description as possible of the observation, so as to enable informed decisions on follow-up. This will include the time of detection, position, flux, size and shape. The shape parameters will include weighted second moments and fits to trailed and dipole source models. The alert stream also will include prior variability information and data from the same night, if available, including low order light-curve moments and the probability that the object is variable. Additionally, alert packets will provide the IDs for nearby Objects in the latest LSST Data Release catalog (i.e., potential host galaxies for transients).

Community-provided alert brokers (e.g., ANTARES, Section 7.1.5) will be the worldwide community access avenue for alerts. At least one community alert broker will provide access to the worldwide community. More information on the current plans for alert production and community brokers is given in LDM-612.[21] In 2021, Rubin completed selection of the community alert brokers through a formal review led by a sub-committee of the SAC. These include seven international teams: Arizona–NOIRLab Temporal Analysis and Response to

21 https://docushare.lsst.org/docushare/dsweb/Get/LDM-612

Events System (ANTARES), ALeRCE (Chile), Lasair (UK), Pitt-Google (U of Pittsburgh), Fink (France), AMPEL (Germany), BABAMUL (CALTECH). Two downstream brokers that operate off one of the primary brokers were also endorsed by the SAC: SNAPS (Northern Arizona University) and POI (Vanderbilt).

In addition to the alert stream, the images (PVIs and difference images) and catalogs (of DIASources, DIAObjects, and SSObjects) generated during Prompt Processing will be made available through the RSP to members of the data rights community within a specified period from the time the corresponding observation was completed.

6.3.2.7.1.1.2 Data-Release Production

6.3.2.7.1.1.2.1 Data-Release Processing Sketch

In contrast to the Prompt Products Processing system, described above, the Data-Release Production system will be executed approximately annually,[22] and will ingest and process all data which has been taken to date. By co-adding and combining data taken at multiple epochs, it will provide a more detailed characterization of fainter astrophysical objects than is provided by the Prompt Data Products. Data-Release Processing proceeds as follows:

- The Image Characterization and Calibration system is used to generate PVIs for each input image. There are two parts to this system:
  - An equivalent of the Single Frame Processing system described in Prompt Processing is used to perform ISR and to estimate the point spread function (PSF) and astrometric and photometric calibration for each image.
  - The global simultaneous calibration system combines source catalogs derived from multiple images to arrive at a refined photometric and astrometric calibration.
- These steps are iterated: the refined calibration is passed back to Single Frame Processing, where it is used to improve the PSF estimate and generate new source catalogs, which are, in turn, used in global calibration to further refine the calibration. Depending on available resources, this cycle may either be iterated to convergence, or for a set number of cycles.
- The PVIs are combined to produce coadded images. A suite of coadds are produced, including deep, seeing-optimized, short-period, and multi-band. Great care is taken to properly set the background levels and to remove image artifacts during coaddition.
- A difference image is generated for each PVI, using an appropriate coadd as the template.
- Source detection is performed on coadds and difference images, generating CoaddSources and DIASources, respectively. The source detection algorithm identifies regions of connected pixels, known as footprints, above the signal-to-noise threshold in the PSF-likelihood image of the visit. An additional technique will be used to detect extended low surface brightness objects.
- The complete set of sources detected to date (coadds and difference images) is associated and deblended to synthesize a list of unique astrophysical objects. This process makes use of all information available at this stage, including, for example, the knowledge of peak positions, bands, time, time variability, and galactic longitude and latitude.
- Objects are characterized by measuring their properties — fluxes, adaptive moments, etc.; refer to the DPDD (LSE-163) for a complete list — on the coadded images.
- Detailed measurements of the shape of each object, appropriate for weak lensing shear analysis, are made. The details of this procedure are still being refined in response to ongoing work in the community on algorithms such as meta calibration (Zuntz et al., 2018). The plan is to use a third-party shear estimation algorithm that runs on small-cell, edge-free coadds.

22 In fact, eleven times over the course of the survey

- Source fluxes will be measured on every visit (PVIs and image differences) with the position and motion parameters characterized in previous steps kept fixed. This process of "forced photometry" results in the ForcedSource catalog, which is used to measure the light-curve of each object in the survey.

6.3.2.7.1.1.2.2 Catalog and Image Data Releases

Over its 10 years of planned operations, Rubin Observatory will produce annual data releases. Each data release will include reprocessing of all data from the start of the survey up to the cutoff date for that release.

Data releases include catalogs of Sources, Objects, DIASources, DIAObjects, and ForcedSources, as well as the raw image data, generated and retained image data (PVIs, coadded images),[23] and all ancillary data necessary to regenerate the release (e.g., calibration frames, telemetry, configuration metadata). Some information is provided in more than one format: for example, catalog data are provided as both Apache Parquet[24] tables, which are designed to facilitate next-to-data processing and map-reduce style operations, and loaded into the Qserv database (Section 6.3.2.8.1.4.2.2, excerpted in the next major section of this case study), which provides a relational query capability. The expected growth of the total data volume of each release (including this duplication) with time is shown in 6.3.1.

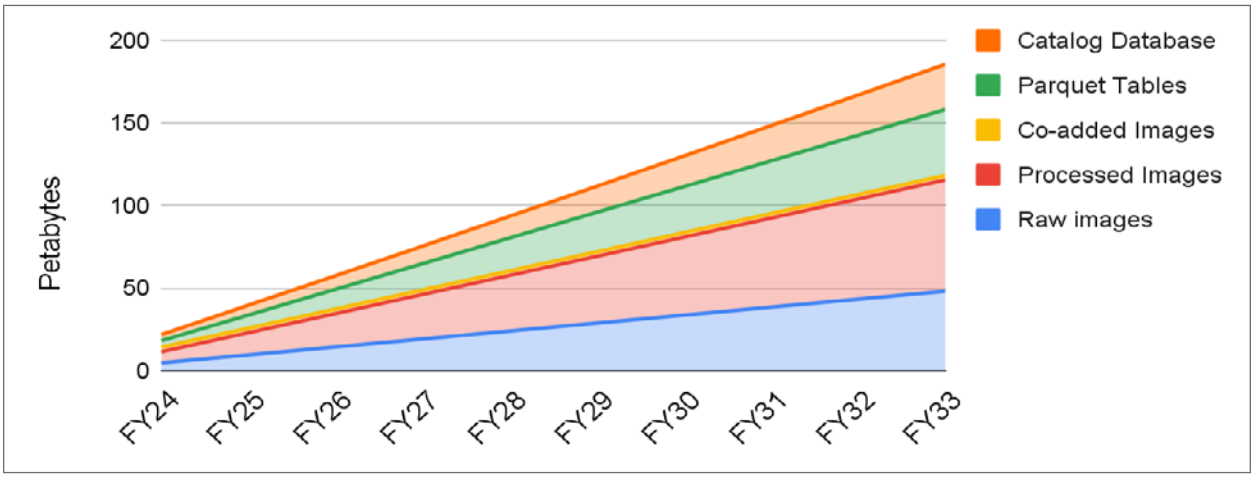


Figure 6.3.1: Total size of a Rubin Observatory data release, as a function of survey year.

Given the volume of data involved, it is not feasible to make all releases available on fast storage ("spinning disk") for the duration of the project. Instead, Rubin will keep the two most-recent releases available for rapid access, and archive older releases to tape. Users will not be able to perform database queries against these archived releases. However, they would be made available for bulk download through appropriate interfaces (see Section 7.1.4 below). Database software and data loading scripts will be provided for users who wish to set up a running copy of an older (or current) data-release database on their systems.

Rubin anticipates keeping small, overlapping samples of data from older data releases loaded in the database to assist with data-quality monitoring and assessment. A sample of a few percent of the older data releases, with larger samples early on, would be sufficient for characterization of the same data for monitoring variations from release to release.

23 Members of the data rights community will be able to access all image data generated over the course of a data release, but it may not all be materialized on physical storage as part of the release; rather, some data products may be regenerated on demand.
24 https://parquet.apache.org

The Operations Team recognizes that the strong scientific drivers maximize the amount of data stored in online databases, and the feasibility of increasing the amount of data available on fast storage is currently under study by the construction DM team. Given the relatively small size of early data releases, storing additional early data releases would not bear much cost impact. In later years of operations, armed with real-time knowledge of fast storage costs in the mid-2020s, the Operations Team will be able to revisit this analysis.

Data products will only be made available after thorough science data-quality assessment has been performed by the System Performance Department, and with the approval of the directorate. Data products will be distributed using common community standards where available and appropriate. In particular, Rubin Observatory's data access services make extensive use of International Virtual Observatory Alliance standards, including Table Access Protocol (TAP)[25] and Astronomical Data Query Language (ADQL)[26] for catalog access, and ObsCore[27], SIA[28], and Hierarchical Progressive Survey (HiPS)[29] for image access.

6.3.2.7.1.1.3 Calibration and Ancillary Data

Calibration data and engineering telemetry (available from the Engineering Facility Database, Engineering & Facility Database [EFD] through the RSP) will be available to the Rubin Observatory community. Calibration data products include the raw calibration images such as flat fields, bias frames, spectra of standard stars obtained using the spectrograph on the Auxiliary Telescope, and images from the all-sky camera. Substantial volumes of low-level telemetry will be obtained and archived from the guiders as well as instruments such as the weather station, anemometers, and temperature sensors used throughout the observatory. Although these data are generally used for engineering purposes and to support operations, these data also will be stored and made available to the community.

6.3.2.7.1.1.4 Education and Public Outreach (EPO) Data Products

Rubin Observatory DMS will produce data products that will populate the EPO program (outside the scope of this ESnet case study.)

6.3.2.7.1.1.5 Software and Documentation

Rubin Observatory regards reproducibility and transparency as key to achieving the best possible scientific results. As such, it is a requirement on the construction project, and a policy Rubin intends to continue into operations, that all data-processing software is to be written in publicly accessible repositories and be released under an open-source license. At present, all the Science Pipelines software (see the science pipelines design document, LDM-151[30]) which is being written to perform the data processing described above is available on GitHub[31] and licensed under Open-Source Initiative-approved licenses (generally the GNU General Public License version 3, with a few exceptions for administrative reasons). Note that all relevant project-authored tooling will come preinstalled and ready for use within Rubin Observatory Science Platform's Notebook Environment. It is expected that the details of distribution and licensing may vary over the course of the operational period, but the principles will remain constant.

Given the extent and complexity of the deliverables — both in software and in data products — produced by Rubin Observatory, documentation is regarded as essential to enable to effectively and responsibly analyze and interpret the Observatory's output. All software and data releases will be accompanied by extensive and appropriate documentation to make this possible.

25 Table Access Protocol; http://www.ivoa.net/documents/TAP/
26 Astronomical Data Query Language; http://www.ivoa.net/documents/latest/ADQL.html
27 http://www.ivoa.net/documents/ObsCore/
28 Simple Image Access; http://ivoa.net/Documents/SIA/
29 Hierarchical Progressive Survey; http://ivoa.net/Documents/SIA/
30 https://docushare.lsst.org/docushare/dsweb/Get/LDM-151
31 https://github.com/

6.3.2.7.1.2 Data Rights

Rubin Observatory data policy is fully described in the document RDO-013[32]. RDO-013 describes a number of scientist communities with rights to access and use Rubin Observatory proprietary data. The communities include US and Chilean scientists, so-called builders from Rubin Observatory construction project and pre-operations, international members with agreements to provide in-kind contributions, and the public. See section 4 of RDO-013. It is expected that Rubin Observatory will make public data (i.e., data after the two-year proprietary period) accessible to the worldwide community. Such access may require non-data rights communities to support the costs of data transfer.

6.3.2.7.1.3 Science Platform

The RSP is a set of web applications and services that will be used by the data rights community to access, visualize, interact with, and analyze Rubin Observatory data. The RSP is being designed to be simple enough to engage general users and flexible enough to meet the needs of experienced users. The RSP also needs to enable users to access contextual knowledge of Rubin Observatory survey and data, facilitating their understanding of the relationships among the data products.

The platform consists of three primary user-facing elements: the web Portal, the Notebook analysis environment, and a set of network-addressable Application Programming Interfaces (APIs).

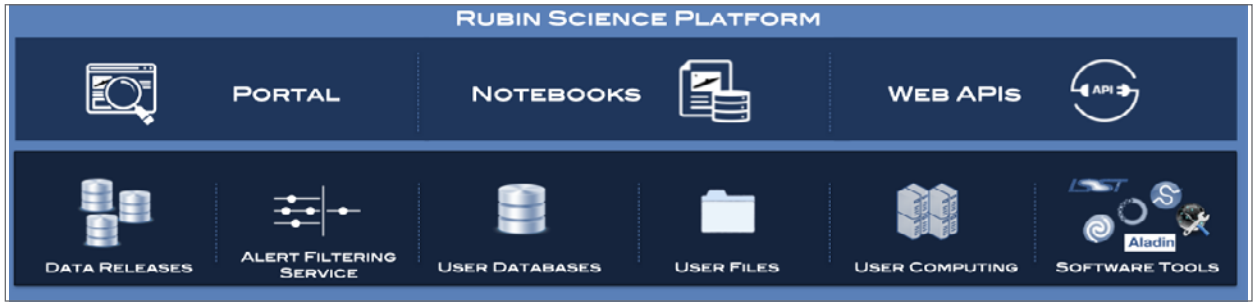


Figure 6.3.2: A layered view of the RSP.

As illustrated in Figure 6.3.2, Rubin Observatory data will be exposed to users through the web portal (interactive use cases), the Jupyter Notebook interface (custom user use cases), and machine-accessible application program interfaces (APIs; remote tools/analyses use cases). The Portal component will provide the essential data access and visualization services common to present-day archives, e.g., Sloan Digital Sky Survey (SDSS) and NOIRLab Data Lab holdings. The Notebook component, based on the Jupyter family of technologies (JupyterHub and JupyterLab) will enable more sophisticated next-to-the-data analysis. These user-visible services will be supported by the user database, storage, computing, and tools components, enabling the users to access, sub-select, and perform added-value processing of the data products.

The Portal component is a web portal designed to provide the essential data access and visualization services through a simple-to-use website. Through the Portal, users will be able to view Rubin Observatory images, request subsets of data via simple web forms or SQL queries, store the results of such queries to their personal workspaces, construct and customize plots, and generally explore the LSST dataset in a way that allows them to identify and access (subsets of) data required by their science case. It will enable browsing and visualization of the available datasets in ways the users are accustomed to at archives such as NOIRLab Data Lab, IRSA, MAST, or SDSS with an enhanced level of interactivity in line with expectations for then-contemporary archive portals. The database serving the data products will support asynchronous queries.

The Notebook environment, based (at least) initially on the Jupyter family of technologies (e.g., JupyterHub and JupyterLab), will enable sophisticated data selection, analysis, and creation of added-value user-generated data products. The Notebook environment will come preinstalled with a library of commonly used and useful software

32 https://docushare.lsst.org/docushare/dsweb/Get/RDO-013

tools such as AstroPy, the Rubin Observatory Science Pipelines software stack, Anaconda Scientific Python Distribution, and others. Users will be able to upload and install their own tools as well.

The Notebook user experience will be nearly identical to working with Jupyter Notebooks locally, except that computation and analysis will occur at resources provided at the USDF. This is an implementation of the "bringing computation to the data" paradigm. Rather than imposing the burden of downloading, storing, and processing (large) subsets of Rubin Observatory data at their home institutions, the RSP will enable users to bring their codes and perform their analysis at the USDF. This will reduce the barrier to entry by users of Rubin Observatory data and shorten the path to science for the Rubin Observatory science community.

The RSP is being designed to facilitate collaborative work. The capabilities ranging from sharing of derived datasets within smaller groups, collaborations, or with the broader Rubin Observatory community, to collaborative visualization and editing capabilities that are expected to become available within the Jupyter ecosystem. Scientists will be able to create user-generated data products from nightly and/or annual data products in order to support particular scientific goals. Once created and with the concurrence of the creators, user-generated data products may be associated with nightly and annual data products through database federation. However, there is no requirement that external groups ever make their products publicly available (see RDO-013). In rare cases, and only with initiation by the creators, community-developed algorithms used to create user-generated data products may be incorporated into the regular Rubin Observatory data processing.

Analyses also can be performed on a variety of supported external systems. Common personal Unix flavors such as Linux and macOS will be supported as resources allow, as well as commonly used cluster and HPC environments.

6.3.2.7.1.3.1 Science Platform Resource Allocation

Queries, visualizations, and analysis performed through the Portal and Notebooks will be served by a shared computing cluster, storage, and database resource. There is a high-level requirement that, in order to enable data exploration and user-generated data product creation, Rubin Observatory provides its user community with software and computing capability equivalent to at least 10% of total Rubin Observatory computing and data storage capacity. The current sizing model for Rubin Observatory compute and storage is summarized in Section 8.1.2 (excerpted in the next major section of this ESnet Case Study), and described in detail in DMTN-135[33]. User resources — shown in table 22 of that document — amount to 500 cores for the USDF Science Platform and 100 cores in the Chilean DAC during the first year of operations. These correspond to 10% and 20% of the total Data-Release processing capacity respectively, and are expected to expand as the Data Facility expands. In year one, 2 PB of storage (corresponding to 0.4 TB for each of 5000 users) is available to the Science Platform in the USDF; this will expand to 10 PB over the course of operations.

Each RSP user account will be provided with a baseline level of resources, to be determined based on the initial size of the user base and the average number of concurrent users the RSP experiences, in practice. The current plan is for 500 cores for the US DAC in 2023 (increasing annually), dedicated preliminary end-user science analyses (e.g., working on small numbers of images), and the creation of user-generated data products. Qserv has its own resources for catalog queries. For the RSP the minimum resource allocation, per concurrent user, is half a core. A good analogy is one of being given a server with a few TB of disk, a few TB of database storage that is co-located next to Rubin Observatory data, and with a chance to use tens to hundreds of cores for analysis (depending on system load). Cluster resources beyond this baseline allocation will be requested from and approved by a Rubin Observatory Resource Allocation Committee—analogous to a Telescope Allocation Committee.

For larger endeavors (e.g., pixel-level reprocessing of the entire LSST dataset), the users will be steered toward resources beyond Rubin Observatory DF (e.g., national supercomputing centers, university computing centers, or the public cloud). Clear, but evolving, guidelines will be published to familiarize scientists with the limits of Rubin Observatory-provided computing and the information necessary to apply for non-Rubin Observatory resources.

33 https://dmtn-135.lsst.io/

Back-end RSP services, including access to databases, images, and other files, will be exposed through machine-accessible web APIs serving community-accepted formats and protocols. This will make it easy to initiate remote computations operating on Rubin Observatory data. Virtual Observatory interfaces will allow connection to other archives and enable the use of standard tools such as Tool for oPerations on Catalogues And Tables (TOPCAT) or Aladin. Rubin Observatory also will be capable of exporting images and catalogs to external computing sites for projects with requirements that exceed the capacity constraints of Rubin Observatory DACs as provided, for example, through in-kind contributions.

6.3.2.7.1.4 Interfaces Between Rubin Observatory and the Community

As a facility, Rubin Observatory will provide science-ready data to the community of scientists to use in their wide-ranging astronomical investigations. These investigations will lead to new understanding of the data, beyond that which an Operations team could ever develop alone. Therefore, the opportunity exists for the Operations team to capture this community experience in order to feed it back into the schedule of the telescope and the processing of the data. This section highlights the mechanisms in place to facilitate the flow of information and knowledge into and out of the Rubin Observatory system and give some illustrative examples of how this feedback process will work.

Interactions between the System Performance Community Science team and the Rubin Observatory LSST Science Collaborations will be a particularly valuable interface because this is a) where active community investigation of the Rubin LSST data is likely to take place and b) where staff are already established in good relationships with the astronomy community. However, ensuring that the feedback from researchers who do not have the time or resources required to sustain membership in these large collaborations is solicited, ingested, and acted on is essential for promoting diversity and inclusion among the LSST science community. The paragraphs below describe a range of interfaces to serve the broad community.

Community Scientists: The individual Community Scientists in the System Performance department Community Science Team will be the primary human interface to the Science Collaborations and the wider scientific community. They will work with the community to intake, understand, and triage the issues identified and then activate the appropriate investigative and corrective activities within the System Performance and other relevant departments. To best enable this work, the cumulative astronomical expertise of the Community Scientists will cover the four LSST science pillars (at minimum).

The LSST Science Collaborations: Since 2006, the SCs have evolved and grown into eight active, self-managed teams that represent a diverse and geographically distributed network of scientists collaboratively preparing for science with the LSST data set. They also provide scientific expertise to Rubin Observatory, and help to guide design and operations choices (e.g., survey strategy, in-kind contributions, algorithms). In 2021, a Federation Charter document[34] was established to recognize the contributions of the Science Collaborations to Rubin Observatory; to establish mutually agreed-upon principles, rules, rights, and privileges of the Science Collaborations; and to define the obligations of Rubin Observatory to the Science Collaborations to support their activities which benefit the entire Rubin Community. During Construction, the DMS Science Team had staff serve as liaisons to each of the Science Collaborations by, e.g., attending their regular telecons and/or face-to-face meetings. During Operations, these liaison roles will be filled by staff with appropriate expertise from across the Rubin departments, and the liaison program itself will be facilitated by the Community Science team.

Documentation: The collection and public dissemination of information about Rubin Observatory, the LSST Science Pipelines, and the data products will begin during commissioning as the camera and full operational system is first characterized as-built. Documents describing key features and software evolution specific to each annual data release will be published approximately following the SDSS model, and made freely available (including to non-data rights holders). During Operations, the Community Scientists will monitor and respond to community needs for additional documentation. Document-writing will be a shared responsibility among many,

34 https://docushare.lsst.org/docushare/dsweb/Get/Agreement-241

coordinated mainly by the Community Documentation positions in the System Performance department. In addition to formal documentation, the Community Science team will curate a suite of tutorials to illustrate usage of the LSST data products and science pipelines.

Q&A Support: An open community forum will be provided for community members to ask questions or request support (e.g., Science Platform tools, coding bugs) and receive assistance from project staff within a few days. In other projects this type of support is referred to as a "Help Desk" and is implemented as a standalone web form (or is an email address), but for Rubin Observatory this is a Support category within the Rubin Community Forum (Community.lsst.org; see below). The Community Scientists will take primary responsibility for monitoring Support requests and connecting users with project expertise. Embedding a "Help Desk"-style functionality in an open forum allows users to help other users, and project staff to help users, without a bottleneck of request processing by a "Help Desk" team.

Rubin Community Forum (Community.lsst.org[35]; see below): Providing an interface with which the users can ask questions and find answers (e.g., like "Stack Overflow") or post interesting results for discussion is of great benefit to both the wider community and to the Rubin Observatory. During Construction the Project has fostered a healthy multithreaded conversation about a wide range of matters at community.lsst.org, an initiative which has enabled collaboration within the community, and between the community and Project staff. This is a public forum; anyone anywhere can obtain an account and view or participate in discussions, including non-data rights holders. There are also private topics restricted to Science Collaboration members and/or Project staff. This online forum is an ever-growing archive of information that enables the community to benefit from their own cumulative expertise, and from interactions with Operations staff.

Users Committee: The Rubin Users Committee is the formal interface between the science community and Rubin Observatory. The User's Committee is charged with actively using the LSST data and the RSP, soliciting feedback from the community, and regularly reporting to the Rubin Lead Community Scientist and to the Rubin Observatory Director with recommendations for science-driven improvements to the LSST data products and the RSP tools and services. The Users Committee Charge (rdo-051.lsst.io[36]) and website[37] provides more details about membership, timescales, and responsibilities.

Survey Strategy Initiatives: Community investigations of the observing schedule and its impact on the science are already underway and being documented in an open "Observing Strategy White Paper,"[38] which was begun as a joint effort between the Science Collaborations and the Rubin Project team. It is anticipated that evolving into an ongoing public conversation between the science community and the System Performance department's Survey Scheduling Team. Detailed cadence simulations are currently underway by the Rubin construction team that will be evaluated in 2021. The input to this cadence analysis was initially a set of 40 detailed white papers submitted to The Rubin Observatory SAC in 2019 after an open call to the worldwide community (i.e., including non-data rights holders). The Survey Scheduling team supports community use of Metric Analysis Framework and interacts with both the Science Collaborations and the broader community.

Following the initial cadence white papers, the scheduling team advanced optimization of the baseline cadence for the full survey start. Then in 2020 a Survey Cadence Optimization Committee (SCOC) was established. The committee solicited further cadence "notes" from the community to further advance the initial cadence or survey plan. As of this writing the SCOC and scheduling team are actively working on the draft initial survey cadence for full survey operations. A set of recommendations from the SCOC is scheduled to be received by the Rubin Director in December 2021. Further simulations will be completed and a revised baseline for the survey cadence will be presented in December 2022.

35 https://community.lsst.org/
36 https://rdo-051.lsst.io/
37 https://lsst.org/scientists/users-committee
38 https://github.com/LSSTScienceCollaborations/ObservingStrategy/tree/master/whitepaper/figs

Table 6.3.2 shows the main examples of Rubin Observatory's approach to community interaction and collaboration that are presented above in detail, and/or discussed elsewhere in this Operations Proposal.

Bulk Downloads: Rubin expects to support bulk downloads of certain data products to affiliated collaborations. The details of this service are not yet worked out, but will be in advance of DR1. Limited support may require resource allocation by the Resource Allocation Committee or be externally supported.

Interface	Function/description
COMMUNITY SCIENCE TEAM	The Community Science team in the System Performance department will be the main human interface between the community and the Operations team. Their mandate is to ensure diversity and inclusion in access to the LSST data products, services, and documentation, and to identify and respond to any issues that might inhibit the scientific productivity of the LSST user community. This team will be responsible for most of the interfaces on this list.
Q&A SUPPORT (COMMUNITY SCIENCE TEAM)	Q&A Support —– The project will provide a venue for Q&A support for users, such as an online forum which will be moderated to ensure answers are posted within a few days. The Community Science Team will be responsible for monitoring this forum and sourcing the appropriate project expertise to assist users.
PUBLIC WEBSITES (EPO/CEE)	Public websites will be used to host documentation, tutorials, and discussion forums for the broad user community. These websites will be archived and searchable to enable the community to make use of an ever-growing repository of information. The Community Science team will be responsible for the contents of these websites.
SCIENCE ADVISORY COMMITTEE	This is the channel for strategic input from the community to LSST. This committee is advisory to the Director, and is comprised of scientists external to LSST that represent the diversity of LSST-enabled science. This committee aims to maximize LSST science and tackles both policy questions and technical topics of interest to LSST and the science community. Community members at large may reach out directly to the SAC chair, or other SAC members, with comments or concerns.
SURVEY CADENCE OPTIMIZATION COMMITTEE	This sub-committee of the SAC will be advisory to the Director on the efficacy of the LSST observing strategy on achieving the key science goals of the community. This committee will be the formal avenue for input on the LSST Observing Strategy during operations.
USERS COMMITTEE	A sub-committee of the SAC, the Users Committee will advise the Lead Community Scientist and the Director on the usability of LSST data products and services, and advocate for science-driven improvements and changes to them. The Users Committee's activities will be coordinated and facilitated by the Community Science Team.
RESOURCE ALLOCATION COMMITTEE	The default allocations of computing, storage, and database queries at the Data Access Centers will be limited. Requests for additional resources needed for large processing tasks will be evaluated on a six month basis by a committee of community members (like a telescope allocation committee). The Director will delegate running this committee to the Lead Community Scientist.
LIMITED COMMUNITY WORKSHOPS	There will be 1–2 workshops a year on LSST-specific science topics, aimed to communicate key developments in the LSST survey implementation, data products, and user resources, as well as provide training and an opportunity for science community collaboration.

Table 6.3.2: A Summary of Interfaces for Community Interaction

6.3.2.7.1.5 Beyond Rubin Operations: Community-Based Data Analysis

In fulfilling its core operational mission, Rubin Observatory will release both real-time and static-sky data at a scale unprecedented in optical astronomy, both in terms of rate and volume and in terms of the diversity of scientific interests served.

The analyses of these data products will be the responsibility of the user community. In order to maximize the scientific return of the data sets, Rubin Observatory user community will need to develop software tools targeted toward specific research problems. A description of some high-priority capabilities needed to maximize Rubin Observatory science can be found in the National Research Council report on Optimizing the US Ground-based Optical and Infrared Astronomy System (NRC, 2015[39]). A subsequent NSF-endorsed community-based study

39 https://nap.nationalacademies.org/catalog/21722/optimizing-the-us-ground-based-optical-and-infrared-astronomy-system

and workshop, sponsored by the Kavli Foundation and jointly led by Rubin Observatory and National Optical Astronomy Observatory (NOAO, now part of NOIRLab), has further detailed and prioritized these capabilities in a report that was released in 2016.

Some of the computational resources and software capabilities highlighted in these reports are:

- Robust and scalable brokering systems to filter, characterize, classify, and annotate time-domain alerts (there are seven full brokers in development, several of which are operating now on Zwicky Transient Facility [ZTF]);
- Curated catalogs of data from surveys complementary to LSST, served through interfaces allowing flexible and large-scale joining with Rubin Observatory catalog data;
- Networks, resources, and workshops for training in and exchange of the data-intensive astronomy techniques required by Rubin Observatory science;
- Computational resources for large-scale, high-fidelity numerical simulations to support the astrophysical interpretation of Rubin Observatory data products; and
- Computational resources to permit development and running of modified Rubin Observatory data-processing code for the generation and distribution of data products that go beyond those provided by Rubin Observatory.

Examples of relevant ongoing work, outside of the scope of this operations plan, that is relevant to the community's use of LSST data, include:

ANTARES, the Arizona-NOAO Temporal Analysis and Response to Events System — NOIRLab and the University of Arizona Computer Science and Mathematics Departments have collaborated to build a software infrastructure system to process alerts generated by astronomical time-domain survey programs. ANTARES will serve as a broker, taking in alerts, annotating them with catalog information from surveys across the electromagnetic spectrum, diverting objects that do not require rapid follow up to a database, and ranking the remaining sources by rarity. These alerts then can be distributed to other brokers who are working on specific projects for more detailed characterization or classification. The goal of ANTARES is to build a system that is scalable to the very large number of alerts that will be generated by Rubin Observatory and to serve as a resource for the general astronomical community. ANTARES is currently running every night at ZTF scale and is under active development to scale to Rubin Observatory size. In 2021, Rubin Observatory selected ANTARES to be one of the seven "community brokers" which will receive and disseminate the Rubin Observatory alert stream (LDM-612[40]). The other six teams include the ALeRCE broker developed by members of the Chilean community, University of Pittsburgh+Google, Caltech, a French community broker, a German community broker, and Lasair, a broker from the UK community.

DESC — This DOE-funded Science Collaboration will require capabilities and services tuned specifically to their domain. The DESC is preparing its own DOE-funded operations plan that complements the activities described in Rubin Observatory's operations plan. Control and understanding of systematics are essential for analyses of the probes of dark energy, and the DESC already has working groups focused on the analysis of weak lensing, large-scale structure, supernovae, and clusters, as well as the calibration of photometric redshifts. In addition, there are working groups on cosmological and survey simulations.

LSSTC Enabling Science Initiatives — LSSTC has sponsored a number of activities designed to prepare scientists to work effectively with Rubin Observatory data as soon as it begins to flow. The Data Science Fellowship Program is a training program designed to teach graduate students skills that are required for working with the large Rubin Observatory databases — skills that are not usually addressed by current astrophysics programs. Currently in its fourth year, the program is in high demand. Recent topics covered by LSSTC workshops include the 2018 Cadence Hack-a-thon which paid stipends totaling 75K USD to 42 white paper

40 https://docushare.lsst.org/docushare/dsweb/Get/LDM-612

authors. These white papers are at the core of the current Rubin Observatory process to simulate and choose the ultimate cadence for the survey. Even more recently, LSSTC supported the "Managing Follow-Up, 2019" workshop giving seed funding and/or observing time to 15 community-based teams. In 2021, LSSTC announced two significant community-based efforts to maximize LSST science. These are the LINCC program (LSST Interdisciplinary Network for Collaboration and Computing) and the Catalyst Fellowship program. Both these programs add resources and effort to LSST research within the Rubin community with particularly strong connections to LSSTC member institutions. LSSTC and its member institutions are vital Rubin Observatory supporters in the development of non-federal funding sources to support Rubin science in the community as well as highly motivated members of the Rubin enterprise in their own right.

NOIRLab Community Science and Data Center —One of NOIRLab's four programs is the Community Science and Data Center (CSDC, the other three are Rubin, Gemini, and the Mid-Scale Observatories). The CSDC will have two core missions: (i) facilitate and support data-intensive astronomical research and (ii) facilitate and support O/IR System optimization. This program will focus on the skills and workflow necessary for users of the public data holdings curated and served by NOIRLab and its constituent programs, in particular Rubin Observatory. This role is highly complementary to the role to be delivered by Rubin Observatory's Data Management and System Performance departments. The requirements that these departments are structured to meet include the generation, assurance, archiving, and delivery of science-ready products to Rubin Observatory data rights community. Accordingly, the community scientist staffing includes community support to (i) ingest and react to community-identified issues with the data or observing strategy and (ii) to ensure the accessibility and usability of Rubin Observatory's science products. Rubin Observatory itself, for example, will not provide support for the additional time-domain infrastructure necessary to follow up on the alert stream. Rubin Observatory also will not provide support for user training in basic programming or data analysis techniques.

Ongoing efforts at NOIRLab are laying the groundwork for this value-added support of Rubin Observatory community during operations. In particular NOIRLab programs Gemini, SOAR, and Cerro Tololo Inter-American Observatory (CTIO) are actively collaborating with Las Cumbres Observatory to develop a dynamically scheduled follow-up network called AEON (Alert Event Observatory Network) and associated end-to-end tools to maximally take advantage of limited US federal resources for follow-up of Rubin Observatory and other transient events.

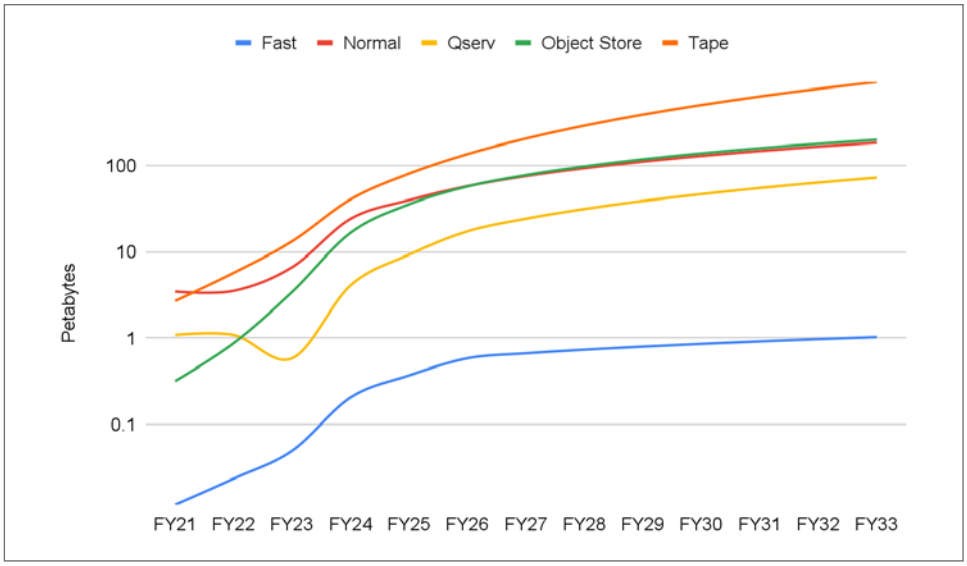


Figure 6.3.4: Total storage required at the USDF as a function of time.

To enable broad-based community science in collaboration with Rubin Observatory, NOIRLab will build on the software capabilities and data products to be delivered by Rubin Observatory Operations, as well as build on current programs within NOIRLab's CSDC, including the Data Lab project, to enable flexible data-intensive science within large archival survey data sets. A goal of Rubin, CSDC, and NOIRLab is to evolve the combined technologies to common solutions and approaches through the Rubin operations phase, specifically after Rubin is through Data Release 1. AEON[41] will incorporate the ANTARES project to develop a scalable public event broker for Rubin Observatory alert stream. It is anticipated that the NOIRLab will interact with Rubin Observatory Operational activities through the LSST Community Scientist (among other channels), who will lead the LSST Community Science Team.

In the special case of dark energy science analysis, SLAC expects to provide specialized capabilities and services to DESC. Wherever these capabilities and services overlap significantly with the needs of the broader US astronomical community, NOIRLab will explore the feasibility of leveraging DOE investment for the benefit of the NSF-supported Rubin Observatory community.

6.3.2.8 Network and Data Architecture

This section excerpts Chapter 17 of the current version of the 10-year Rubin Observatory Operations plan, describing the Observatory's data systems and network architecture. The content and internal references have been lightly edited to make this Case Study document self-contained, however the numbering of subsections and figures is preserved from the original document.

6.3.2.8.1 Data Management System

6.3.2.8.1.1 Rubin Observatory DMS Architecture

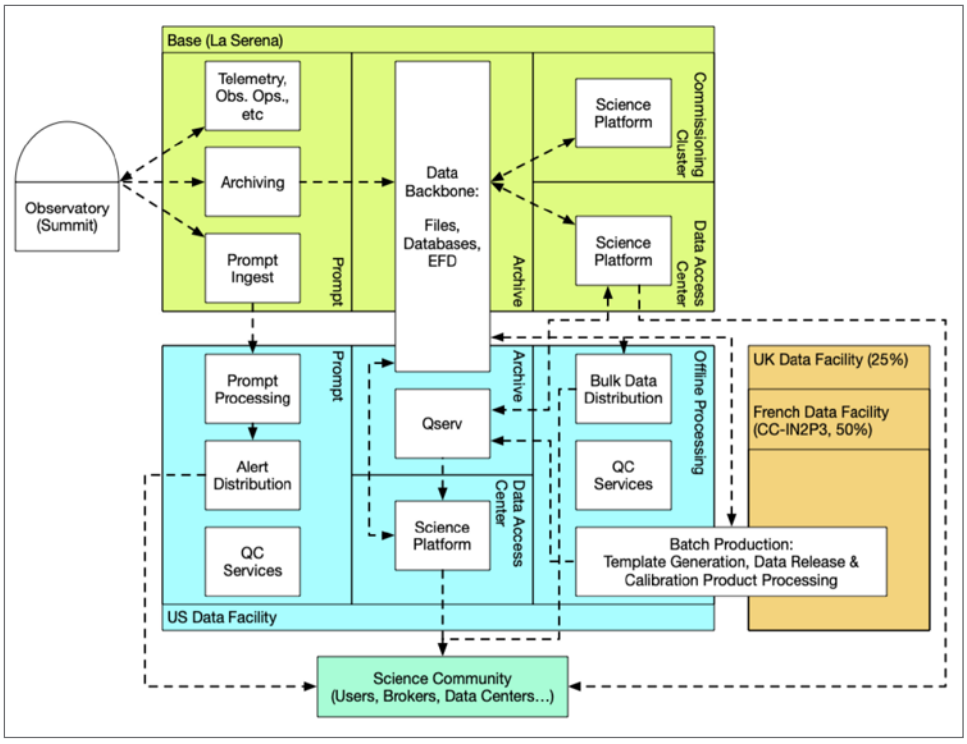


Figure 6.3.3: Architectural overview of the Rubin Observatory DMS. Colored boxes represent the sites over which the DMS is deployed; divisions within them show the deployment enclaves; white boxes indicate particular services.

41 https://lco.global/aeon/

The Rubin Observatory DMS is the combined hardware and software infrastructure which will collect data from the Rubin Observatory LSSTCam and other ancillary instrumentation, transport, archive, process it into science-ready data products, and make the raw data and derived data products available to the community.

The DMS is structured around a number of services, which are deployed and managed within enclaves, and hosted over three separate sites — the Chilean Data Center, the USDF, and CC-IN2P3 — connected by WANs. This architecture is shown schematically in Figure 6.3.3.

Each site is separated into functional enclaves based on requirements for availability, access, and change management. Enclaves are distinguished by having different users, operational timescales, interfaces, and services. The Chilean Data Center supports four enclaves: Prompt Base, Archive Base, Commissioning Cluster, and Chilean DAC. The US Data Facility provides Prompt USDF, Archive USDF, US DAC, and Offline Processing USDF enclaves. A single Offline Processing Satellite enclave is deployed at IN2P3.

Each enclave supports one or more services. These services can be considered as consisting of four tiers of components:

1. The Science Platform, which provides a user interface and data aFs for use by both Observatory staff and the wider scientific community.
2. Applications Software, which provides libraries of data structures and algorithmic code for representing and processing Rubin Observatory data. These libraries are available for use through the Science Platform, and are used to construct the Science Pipelines which are executed to generate the standard Rubin Observatory data products.
3. Middleware is used to provide the environment within which the Science Pipelines execute. It provides a layer of isolation and abstraction between the Pipeline "payloads" and the underlying infrastructure on which they execute, and provides data access services to both science users and Observatory staff.
4. Infrastructure — the compute hardware, storage, networking, and the low-level services and software which are necessary to run and manage a data center — underpins and supports all the above component tiers.

The various enclaves, and the services which are hosted within them, are described below.

6.3.2.8.1.2 Compute and Storage Sizing

The DMS construction team has completed an extensive evaluation of the compute and storage infrastructure which is necessary to process and release all the data collected during the Rubin Observatory LSST (the "sizing model"), and have further used that to estimate the total non-labor cost of the system (the "costing model"). The full details of this evaluation are available in technical note DMTN-135[42]; this section presents a brief overview of the estimation process.

The sizing and costing models described in this section have been initially developed based on the prototype Data Facility built by the Construction Project at the National Center for Supercomputing Applications at University of Illinois at Urbana-Champaign (NCSA). The ultimate location of the operational Data Facility is SLAC and this has introduced a slightly different costing. This should be clear in DMTN-135.

42 https://dmtn-135.lsst.io/

6.3.2.8.1.2.1 Storage Requirements

The storage sizing model is derived on the basis of six fundamental considerations:

- LSE-81[43], which provides estimates of the total number of sources which will be observed by the Rubin Observatory based on the LSST Science Book[44], the Science Requirements Document (LPM-17[45]), and the Observatory System Specifications (LSE-30[46]);
- LSE-163[47], the Data Products Definition Document, which summarizes the various types of image which will be made available, and describes the information which will be included in Rubin Observatory catalogs; and LDM-153[48] which maps those catalogs onto a detailed database schema;
- Practical experience of processing precursor datasets with prototype versions of the Rubin Observatory Science Pipelines, which helps to establish intermediate data products which must be stored temporarily; provides a basis for estimating how effective compression is when applied to Rubin Observatory data; and establishes the size of ancillary data products like representations of the PSF;
- Science Platform users will have access to personal storage space for generating and storing their work;
- Access requirements for data products, including ensuring that products where low-latency access is required are stored on fast storage; that all data releases are archived to tape; and that the two most-recent data releases are available on spinning disk (LPM-151[49]);
- The approach taken to disaster recovery, which determines how much storage redundancy is required.

Based on these considerations, there are five separate classes of storage:

- Fast storage, likely comprised of solid-state devices, is used for the Alert Production Database (APDB; for which low-latency access is required as part of the Prompt Products Production system described Section 7.1.1.1 in the Case Study section above) and the client-facing query management and indexing components of the distributed Qserv database (Section 8.1.4.2.2).
- Normal storage, corresponding to networked filesystems hosted on enterprise-grade spinning magnetic disks, is used for initial collection of raw data, and intermediate and final data products which are being generated or modified during the production of a data release, and for user storage.
- Object stores, hosted on consumer-grade spinning disks, are used for providing cheap and scalable access to read-only data. In general, it is expected data will migrate from filesystem-based "normal" storage to object stores when processing is completed.
- Qserv storage is dedicated to the distributed database system which is used to manage data-release catalogs. The Qserv system is intrinsically fault-tolerant, so consumer-grade spinning disks are used.
- Tape storage is used for archiving and back-up all data products which are hosted on other storage media.

43 https://ls.st/lse-81
44 https://www.lsst.org/scientists/scibook
45 https://docushare.lsst.org/docushare/dsweb/Get/LPM-17
46 https://docushare.lsst.org/docushare/dsweb/Get/LSE-30
47 https://docushare.lsst.org/docushare/dsweb/Get/LSE-163
48 https://docushare.lsst.org/docushare/dsweb/Get/LDM-153
49 https://docushare.lsst.org/docushare/dsweb/Get/LPM-151

At any given time, the two most-recent data releases are made available on the "live" Qserv and object store system; older data releases are only available from tape, by special arrangement and at substantially greater latency.

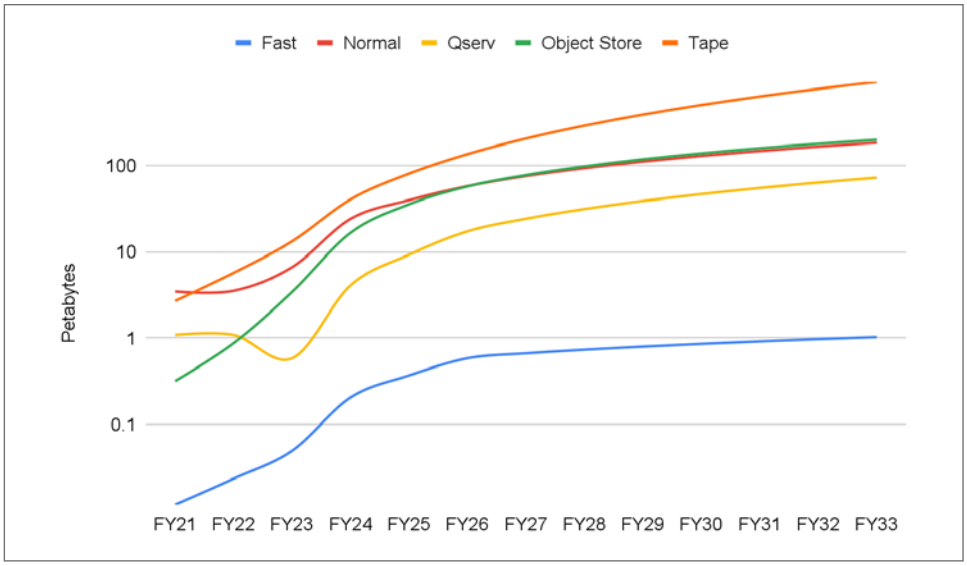


Figure 6.3.4: Total storage required at the USDF as a function of time.

The USDF will store and make available to the data rights community all Rubin Observatory LSST data products. It will also archive old data releases, and maintain back-up copies of the data. Finally, it provides personal storage for Science Platform users[50]. Taken together, these imply the total storage requirements for the USDF shown in 6.3.4. Note that this figure also shows storage required in support of construction, commissioning, and pre-operations activities before the start of full operations. Storage used by the construction project will be purchased by the construction budget, and will (unless retired) transition into the operations project. However, the bulk of storage required in operations will be purchased by the operations budget.

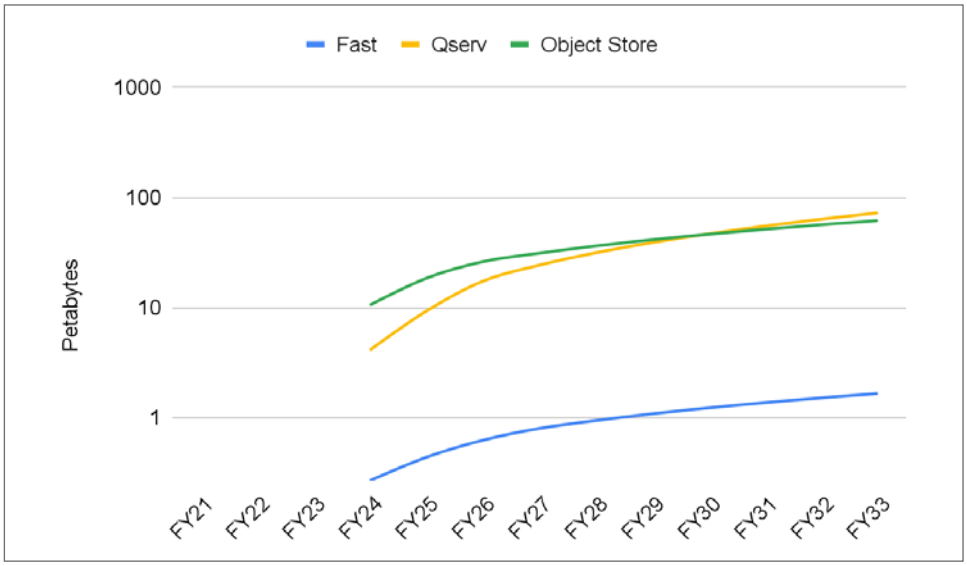


Figure 6.3.5: Total storage required at the Chilean DAC as a function of time.

50 This assumes 5000 users each having access to 0.4 TB in the first year of operations, scaling to 7500 users having access to 1.3 TB by the end of the survey.

Storage is also required to support the activities of the Chilean DAC (Section 6.3.2.8.1.3.2.2). The Chilean DAC is not required to support any activities before the start of full operations; no storage is required before the first data release is made available at the start of FY23. Further, the model assumes that all data in Chile is hosted on fast storage (Qserv query management and index), Qserv storage (the bulk Qserv database), or object storage (for image and file-based data products). The expected evolution of storage at the Chilean DAC is shown in Figure 6.3.5.

6.3.2.8.1.2.2 Compute Requirements

Compute requirements are estimated separately for Prompt Processing (Section 8.1.1.1), Data-Release Processing (Section 7.1.1.2, excerpted in the preceding major section of this Case Study), and the Science Platform (Section 7.1.3, excerpted in the preceding major section of this Case Study). The Science Platform is made available for users at the US DAC, the Chilean DAC, and for internal staff use. This analysis does not consider other processes, such as periodic template generation or calibration products processing, which are assumed to be negligible.

Estimated compute requirements for the Science Pipelines are derived from processing precursor data — primarily taken from Hyper Suprime-Cam on the Subaru Telescope and DECam on the Victor M. Blanco Telescope[51] — through existing pipeline prototypes, and scaling the results based on the source counts, derived from scientific considerations, documented in LSE-81[52]. Compute times are measured on existing hardware, and converted[53] into core-hours on fiducial hardware[54]. This estimation methodology incorporates all I/O, memory bandwidth, cache miss, and other overheads into the core-hour measurement. Note that the fiducial hardware is not assumed to evolve with time; in practice, it is expected that continued technology evolution will result in a reduction of the number of core-hours required late in the survey relative to the estimates presented here.

In some cases, Rubin Observatory algorithms are still under active development, and are not yet appropriate for making informed estimates of actual compute cost. In these cases, core-hour figures obtained for comparable jobs from the DES[55] have been used in their place. Where equivalent DES and Rubin codes are available (e.g., in single-visit processing), run-times have been shown to be well matched, which gives us confidence in this basis of estimate.

Based on the above considerations, Rubin requires a constant 1,188 CPU cores throughout the survey period to perform Prompt Processing. These cores will be used during the night for alert generation, and during the day for catch-up (where necessary) and solar system processing. Each core will have access to 5 GB of memory, and all these cores are located at the USDF.

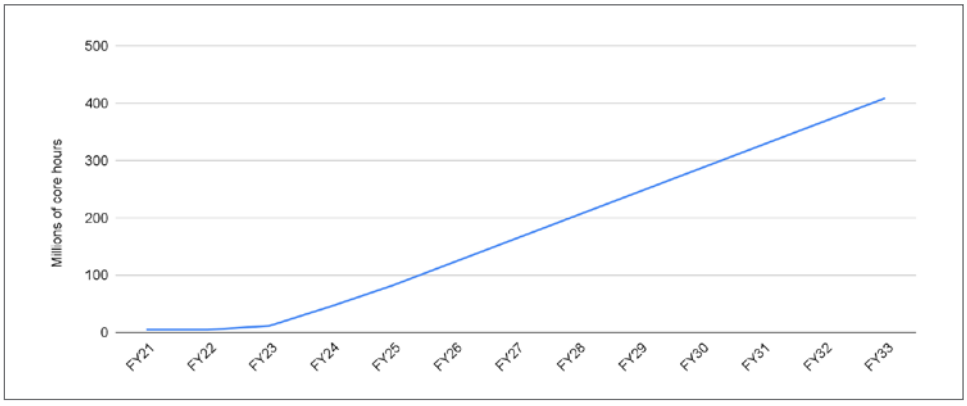


Figure 6.3.6: Fiducial core-hours required for data-release processing as a function of time.

51 https://www.darkenergysurvey.org/the-des-project/instrument/
52 https://ls.st/lse-81
53 For conversion between different architectures, the ratio of industry-reported achievable FLOPS is used.
54 Intel Xeon E5-2680v3 at 2.50 GHz
55 https://www.darkenergysurvey.org

The core-hours required for data-release processing increases as the volume of data collected grows, as shown in Figure 6.3.6. One third of these cores will have access to 5 GB of memory; the remainder, to 20 GB; jobs will be allocated to high- or low-memory core depending on their requirements (e.g., operations on a single CCD will generally be executed on a low-memory system, while combining multiple detectors or deblending complex scenes will require a high-memory system). These cores will be divided between the USDF and the Satellite Data Processing (Section 6.3.2.8.1.5) to enable split-site processing.

Cores are allocated to user-driven processing in the Science Platform as a ratio of the total available compute system, following the construction Science Requirements Document (LPM-17[56]). On this basis:

- USDF user computing is sized as 10% of the total Data-Release Processing compute, amounting to over 500 cores at the start of operations;
- Chilean DAC computing as sized as 20% of USDF user computing;
- Project staff computing is sized as 10% of USDF user computing.

It will be possible to dynamically reallocate CPU cores between services at the USDF. In particular, it will be possible to reallocate cores from Data-Release Processing to the Science Platform at times of high demand, assuming that the long-term average level of Data-Release compute is adequate to meet the release schedule.

6.3.2.8.1.3 Chilean Data Center

The Chilean Data Center is hosted at the La Serena Base Computing Center. This center is a megawatt facility with redundant power service provided by on-site gas generators. The Chilean Data Center has 6 x 100 Gbps network links to the Summit Facility — to receive streamed data from the telescope and to enable remote control — and will have dedicated dual PoP and path-diverse 100 Gbps networking to the USDF as well as connectivity to dedicated R&E networks and the commodity (or open commercial) Internet.

Through an agreement with the other AURA centers on the NOIRLab recinto, the Computing Center will be used to consolidate computing and network servers from CTIO, SOAR, and AURA operations (and likely Gemini in the future).

The Chilean Data Center hosts four DMS enclaves, as described below. In addition, it will host a backup of the entire Rubin Observatory data set.

The Chilean Data Center will be operated by the Rubin DevOps team, under the supervision of the DevOps Manager. This team is part of the Rubin Data Management Department. Most services deployed within the various enclaves will operate software which is developed and maintained by the Data Management Department (primarily, the Pipeline Middleware and Science Platform and Reliability Engineering groups). The Chilean Data Center is hosted within the physical Data Facility in La Serena which is managed by NOIRLab facilities and ITOps. The shared space was developed under leadership of Rubin Observatory but is available for use by all Programs in NOIRLab in Chile.

6.3.2.8.1.3.1 Prompt Base

The Prompt Base enclave exists to support services which must interact with the Observatory Control System (OCS) or the Camera Data Acquisition System (DAQ). In several cases, these services are closely related to counterparts in the Prompt USDF enclave (Section 6.3.2.8.1.4.1). Services in this enclave are, where possible, designed to operate autonomously and with a high degree of fault tolerance; the enclave is designed to be functional whenever the Observatory is taking data.

The Prompt Base enclave supports the following services.

56 https://docushare.lsst.org/docushare/dsweb/Get/LPM-17

6.3.2.8.1.3.1.1 Archiving

The Archiving Service captures raw images taken by the Observatory's cameras (both the main LSSTCam and the Auxiliary Telescope spectrograph), including data from wavefront and guide sensors. These are augmented by metadata and telemetry captured by the Header Service, which forms part of the Archiving Service. Image data and metadata are then supplied to the Observatory Operations Data Service (OODS; Section 6.3.2.8.1.3.1.4) and staged for ingestion into the Data Backbone (Section 6.3.2.8.1.6).

The Archiving Service also includes a "catch-up" capability, which can stage any data to the OODS and Data Backbone that was missed by the primary archiving system due to network or other outages.

6.3.2.8.1.3.1.2 Planned Observation Publication

The Planned Observation Publication Service retrieves telemetry from the OCS describing the location of the next visit, and the scheduler's predictions of future visits. It makes these available as a globally accessible web page for human inspection, and via machine readable web APIs.

6.3.2.8.1.3.1.3 Prompt Processing Ingest

The Prompt Processing Ingest Service captures images from the LSSTCam data system together with selected metadata from the OCS and forwards it directly to the Prompt Processing Service in the Prompt USDF enclave (Section 6.3.2.8.1.4.1.1).

6.3.2.8.1.3.1.4 Observatory Operations Data

The OODS provides access to files and metadata to Observatory systems. It is designed to provide lower-latency than the Data Backbone (Section 6.3.2.8.1.6) to facilitate fast turnaround analyses by the team in Chile.

6.3.2.8.1.3.1.5 Observatory Control System Driven Batch

The OCS Driven Batch Service provides a batch computing service capable of executing science payloads. It is used for processing daily calibration data. Data is read from and written to the Data Backbone (Section 6.3.2.8.1.6).

6.3.2.8.1.3.2 Archive Base

The Archive Base enclave provides a single service: an endpoint for the Data Backbone. The Data Backbone is described in Section 6.3.2.8.1.6.

6.3.2.8.1.3.3 Commissioning Cluster

The Commissioning Cluster enclave provides a single service: an interactive computing environment for rapid turnaround of human-driven ad hoc analysis of data during any recommissioning of the system. In addition, this enclave supports human-driven quality control activities undertaken in Chile.

The Commissioning Cluster runs an instance of the LSST Science Platform (Section 6.3.2.7.1.3, excerpted in the preceding major section of this Case Study) with low-latency access to the OODS (Section 6.3.2.8.1.3.1.4).

6.3.2.8.1.3.4 Chilean DAC

The Chilean DAC is responsible for providing science-user-facing data access and analysis services to the Chilean community. It will contain all raw data on disk and probably a cache of other data. See LDM-572[57] for details. The primary user access and interface to the DAC is through an instance of the LSST Science Platform (Section 6.3.2.7.1.3). The DAC will enforce authorization and authentication policies set by the Rubin Observatory for access to compute and data resources.

57 https://ldm-572.lsst.io/

6.3.2.8.1.4 US Data Facility

The USDF, sited at SLAC, acts as the primary processing and distribution site for all Rubin Observatory data products. This includes responsibility for Prompt Processing and alert distribution (Section 6.3.2.7.1.1.1, Case Study section above), 35% of Data-Release Processing (Section 6.3.2.7.1.1.2; the remainder of the processing is carried out at CC-IN2P3 (40%) and in the UK (25%), as described in Section 6.3.2.8.1.5), archiving the results of processing, and making them available to the data rights community through a DAC. The USDF will also provide for user-generated data product production. .

The USDF is configured with a high-reliability IT service model that ensures professional, secure, and consistent operation. Systems at the USDF have been designed to allow for dynamic reallocation of resources between Observatory operations, batch processing, and DAC subsystems in response to shifting peak demand.

The USDF will be operated by the Infrastructure and Support team within the Data Management Department, under the supervision of the USDF Lead. Generation of data products using Data Facility infrastructure will be carried out by the Data Management Department's Execution group, under the leadership of the Lead Production Scientist. These groups will work closely with the Data Abstraction and Algorithms and Pipelines teams, which will provide the execution middleware and the science payloads, which will be used to generate data products, and with the Science Platform and Reliability Engineering team, which will develop and support the Science Platform which provides the primary mechanism by which Observatory staff and members of the data rights community will access and work with data stored in the facility.

The USDF hosts four DMS enclaves.

6.3.2.8.1.4.1 Prompt USDF

The Prompt USDF enclave is responsible for compute-intensive processing for all near-real time operations and other operations closely tied with the Observatory. The services provided by this enclave feedback status information to the OCS through the Telemetry Gateway (Section 6.3.2.8.1.3.1.6); as such, they form an important part of the overall Rubin Observatory system, and are generally expected to be available whenever observing is ongoing.

6.3.2.8.1.4.1.1 Prompt Processing

The Prompt Processing Service receives images and accompanying metadata from the Prompt Processing Ingest Service (Section 6.3.2.8.1.3.1.3), executes one of a number of possible algorithmic payloads upon them, and stores resulting data products to the Data Backbone (Section 6.3.2.8.1.6).

Algorithmic payloads include:

- Alert Production pipelines, as described in Section 6.3.2.7.1.1.1.1 (Use Case section above). Alerts are generated for point sources detected in the difference images, and also from certain "streak" features consistent with Near Earth Objects. (Artificial satellites are largely filtered out by the Rubin Science Pipelines. Any remaining confounding contaminants are excluded from the alert stream by Rubin's AVS). Science alerts are passed to the Alert Distribution Service (Section 6.3.2.8.1.4.1.2) for distribution. In this mode, data from two visits is processed simultaneously to meet the throughput and latency requirements resulting from the duration of each visit being 30 seconds, and alert distribution taking place within 60 seconds of visit completion.
- Calibration Product Production pipelines. These data taken by the LSSTCam and/or ancillary information to generate calibration products, such as bias, flat, and dark frames, which are used to calibrate the telescope during Alert Production processing.
- Exposures may be grouped and processed by specially configured pipelines based on the above when deep-drilling fields are being observed.

6.3.2.8.1.4.1.2 Alert Distribution

The Alert Distribution Service receives alert packets generated by the Prompt Processing Service (Section 6.3.2.8.1.4.1.1) and distributes them to consumers. A number — at least five — of community-provided "brokers" will receive all alerts generated (expected to amount to around 10,000,000 per night). In addition, the Alert Distribution Service provides a limited capacity and capability Alert Filtering Service, which will allow individual members of the data rights community to receive alerts directly from the Observatory.

6.3.2.8.1.4.1.3 Prompt Quality Control

The Prompt Quality Control Service will monitor the execution of the Prompt Processing Service (Section 6.3.2.8.1.4.1.1) and will post-process data products written to the Data Backbone (Section 6.3.2.8.1.6) to generate additional measurements. Warnings are provided to the responsible Observing Specialists and Data Facility Production Scientists when these measurements cross predefined thresholds.

6.3.2.8.1.4.2 Archive USDF

The Archive USDF enclave provides bulk storage for file and catalog data which are used by other enclaves within the US Data Facility. It provides the following services.

6.3.2.8.1.4.2.1 Data Backbone Endpoint

The Archive USDF enclave provides an endpoint for the Data Backbone. The Data Backbone is described in Section 6.3.2.8.1.6.

6.3.2.8.1.4.2.2 Qserv Distributed Database

Over the course of the operational period, Data-Release Production (Section 6.3.2.7.1.1.2, Use Case section above) will generate catalogs consisting of tens of trillions of rows and tens of PBs of data (Section 6.3.2.7.1.1.2.2). These will be made available for data rights holders to access and query through the Science Platform (Section 6.3.2.7.1.3). Catalogs will be stored and queried using Qserv, a bespoke massively parallel relational database system (LDM-135[58]). Qserv makes intelligent use of replication, chunking, vertical partitioning, and shared scans to provide a system which is fault tolerant and capable of scaling to meet the variety and complexity of anticipated queries, which range from simple object lookups to complex full-sky correlations over billions of elements. Although Qserv is being developed directly to meet Rubin Observatory's needs, it builds upon the mature open-source technologies MariaDB[59] and xRootD[60].

The Qserv instance is hosted in the Archive USDF and provides the primary source-of-truth for released data — this is where catalogs are ingested when Data-Release processing is complete — and provides query capabilities to users of the US DAC (Section 6.3.2.8.1.4.4). Additional Qserv instances may be deployed locally by other DACs as required.

6.3.2.8.1.4.3 Offline Production USDF

The Offline Production USDF enclave is responsible for all long-period data processing carried out at the US Data Facility. This includes the largest and most complex science payloads executed by the Rubin Observatory system, used to generate calibration products and data releases. This enclave also includes systems for monitoring and quality control of ongoing processing jobs, and for bulk distribution of data to partner institutions.

Services in the Offline Production USDF enclave are expected to manage long-running processing jobs— spanning weeks or months of real time — efficiently and reliably. They track the execution of millions or billions of individual tasks, and ensure output data are collected and saved to the Archive USDF enclave. Services are designed for autonomous operation where possible, although provisions are made for manual intervention where appropriate.

58 https://docushare.lsst.org/docushare/dsweb/Get/LDM-135
59 https://mariadb.org
60 http://xrootd.org

In general, low-latency processing is not critical in this enclave. However, because the enclave is sized to provide appropriate capacity for generating data releases on schedule, downtime must be minimized to ensure data are released on time.

The Offline Production USDF enclave provides the following services.

6.3.2.8.1.4.3.1 Batch Production

The Batch Production Service executes processing "campaigns," where each campaign consists of a given science pipeline, configuration, and set of inputs and outputs. Various different campaigns can be executed on different cadences, including:

- Periodic calibration product generation;
- Annual data-release processing;
- Daily solar system processing;
- As needed alert catch-up processing and special programs processing.

More than one campaign can be executed simultaneously. In particular, data-release processing will largely be scheduled as a single campaign which runs for much of the year, while other, smaller, campaigns execute in parallel with it.

This service can coordinate with its peer at the IN2P3 Satellite Data Processing Center (Section 6.3.2.8.1.5; also, potentially, with other facilities) to enable split-site processing of data releases.

6.3.2.8.1.4.3.2 Offline Quality Control

The Prompt Quality Control Service will monitor the execution of the Batch Production Service (Section 6.3.2.8.1.4.3.1) and will post-process data products written to the Data Backbone (Section 6.3.2.8.1.6) to generate additional measurements. Warnings are provided to the responsible Observing Specialists and Data Facility Production Scientists when these measurements cross predefined thresholds.

6.3.2.8.1.4.3.3 Bulk Distribution

The Bulk Distribution Service is used to provide Prompt and Data-Release products to major facilities and partners such as Science Collaborations. It extracts data products from the USDF Archive and transmits them over high-bandwidth connections to designated recipients. It is not available for direct access by science users, who will, instead, access data through the DAC (Section 6.3.2.8.1.4.4).

6.3.2.8.1.4.4 US Data Access Center

The US DAC is responsible for providing science-user-facing data access and analysis services to the worldwide Rubin Observatory data rights community. It will be hosted on the Cloud while all raw data, all file-based data-release products, and all databases will be held at SLAC. The primary user access and interface to the DAC is through an instance of the RSP, Section 6.3.2.7.1.3, Case Study section above). The DAC will enforce authorization and authentication policies set by the Rubin Observatory for access to compute and data resources. The RSP has been shown to be portable, having been operated in multiple situations, including the Google Cloud. The success of the IDF has motivated hosting the RSP in the Google cloud moving forward.

6.3.2.8.1.5 Non-US Data Facilities

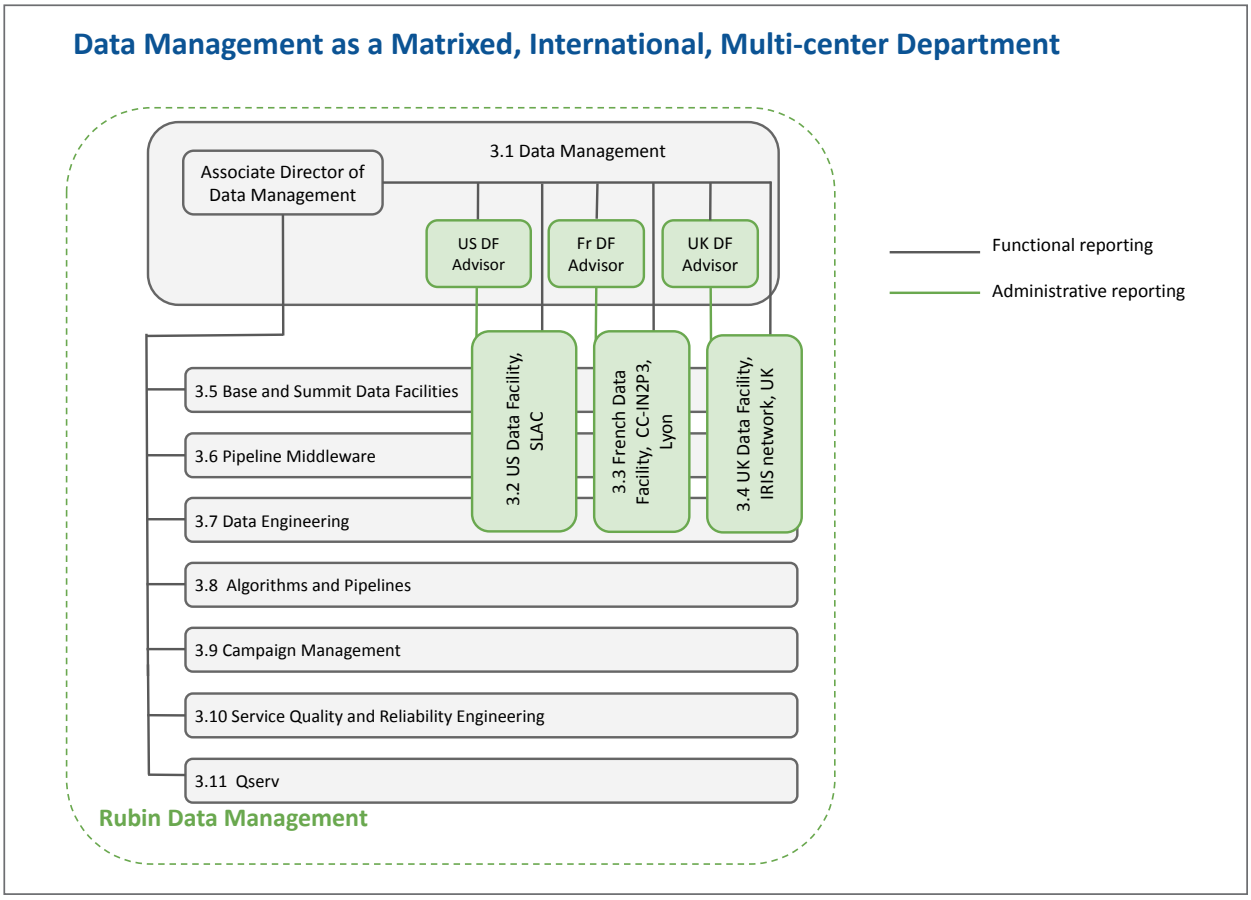


Figure 6.3.7: Data Management organization. The USDF, and the French and UK Satellite Data Processing Centers are matrixed in the Data Management Department. An advisory council advises the DP AD.

In operations, Rubin envisions CC-IN2P3 and the UK as integrated parts of the Data Management Department. Computing and storage will be managed locally, but an advisory council with membership from IN2P3, the UK and the USDF will meet regularly with the Data Management AD to ensure smooth co-processing of data. This structure is illustrated in 6.3.7. Further, staff members from IN2P3 and the UK will be integrated with the Data Management Work Breakdown Structure (WBS), providing excellent communication and awareness throughout the department.

6.3.2.8.1.5.1 IN2P3 Data Facility

A memorandum of agreement has been established by IN2P3, LSSTC, Rubin Observatory, and the University of Illinois for the IN2P3 Computing Center (CC-IN2P3) in Lyon, France to provide 40% of the computing capacity needed for annual data-release processing during LSST operations, as estimated at the time of the MOA, and to host a complete backup of the Rubin Observatory data archive. CC-IN2P3 is a highly data-intensive supercomputer center with an established history of data management for high energy physics experiments, including the LHC at CERN, and in which large-scale astronomy experiments are a future strategic priority. The MOA requires that France will provide the TA network connectivity to ESnet needed to support data transport to and from CC-IN2P3.

During the current construction phase, this collaboration is managed by a Joint Coordination Committee chaired by NCSA. This committee is charged with preparing for combined processing which will take place during survey operations.

6.3.2.8.1.5.2 UKDF

The UKDF will be a distributed entity. Its computing hardware will be provisioned through IRIS[61], a national organization providing computing resources for projects funded by the STFC with large-scale computing requirements. At the moment, IRIS provides the UK's contribution to the WLCG, as well as computational support for preparatory work for a number of future astronomy projects, including the Square Kilometer Array and Euclid. IRIS sites are already well integrated into national and international networks, for the transfer of LHC and other data, and have a wealth of experience supporting a wide range of international science collaborations.

The UKDF will be hosted by the University of Edinburgh, with hardware at its Advanced Computing Facility, which is home to the UK's national academic supercomputing service, and staff working with the Wide-Field Astronomy Unit, located at the Royal Observatory Edinburgh, which hosts a number of existing sky survey archives. The location of the UKDF staff contributing to Data-Release Processing will be determined by the range of roles required to support the UK's quarter share of the Data-Release Processing workload, which is currently being discussed by Rubin operations management and the LSST:UK Consortium.

6.3.2.8.1.5.3 Offline Processing Satellite

The Offline Processing Satellite enclave provides a single service.

6.3.2.8.1.5.3.1 Batch Production

The Batch Production Service offered at the Satellite Data Processing Center offers a similar range of capabilities to that at the USDF (Section 6.3.2.8.1.4.3.1). It will be used to perform 40% of the annual data-release processing.

6.3.2.8.1.6 Backbone Services

The Data Backbone is responsible for data storage, transport, and replication, including transferring files between the constituent facilities of the Rubin Observatory system. Specifically, the Data Backbone is responsible for:

1. Management of files and data resident in designated database management systems in Chile and at the USDF;
2. Enforcing data retention and redundancy policies set by Rubin Observatory for data products at all levels;
3. Providing disaster recovery capabilities for data;
4. Transporting data as needed between the observatory sites; and
5. Meeting data service levels specified for all users (e.g., members of the scientific community) and other consumers of Rubin Observatory data services (e.g., alert brokers).

To meet these responsibilities, the Data Backbone provides policy-based replication of files (including images) and databases (including metadata about files as well as other miscellaneous databases but not including the Qserv data-release database) across multiple physical locations. It provides for caching of data at each endpoint, and automatic persistence of data to long-term archival storage when appropriate.

The Data Backbone has "endpoints," where data may be stored or retrieved, at the USDF (Section 6.3.2.8.1.4.2.1) and the Chilean Data Center (Section 6.3.2.8.1.3.2).

6.3.2.8.1.7 Wide Area Networks

6.3.2.8.1.7.1 Network Requirements

The Rubin Observatory's WANs (6.3.2.8) are responsible for transferring data and control information among the distributed observatory sites, and also provide connectivity to data centers for scientists and non-scientists alike. As such, the requirements are complex, being drawn not just from the Observatory System

61 http://www.iris.ac.uk/

Specifications (LSE-30[62]) as flowed down through the DMS Requirements (LSE-61[63]), but also from memorandums of agreement between the Observatory and key partners. These requirements are summarized in the Network Design Document (LSE-78[64]) for convenience.

These requirements necessitate data transfers on a wide range of timescales, including:

- Order of milliseconds within the OCS and the control systems of each subsystem (the Telescope, Camera, and DMS);
- Order of seconds in acquiring and processing the raw image stream from the Camera through Data Production to create transient alerts;
- Order of days in archiving the raw image stream and other metadata for subsequent use in the production of astronomical catalogs and other data-release products;
- Order of weeks to months in producing and deploying data necessary for calibration of the Telescope and Camera; and
- Order of years in producing and deploying Data Releases of processed images and catalogs.

In addition, there are large geographical distances between the various sites, which imply non-trivial latency in certain data transfers. There are also data-transfer requirements associated with supporting user access in DACs in Chile and the United States, and in the Education and Public Outreach Data Center (EDC) in the United States. In the case of Chile, these requirements are derived from MOAs between Rubin Observatory and AURA (Document-11647), and between AURA and Chile (Document-11622) that require providing a DAC in Chile in return for authorization to site Rubin Observatory in that country.

Considering continuity and criticality, Rubin has explicitly planned for certain levels of outages and failures, and has replaced or augmented existing systems to enhance reliability. For example, before the start of construction a microwave link was the only connection between the mountain and the base site. This did not meet the reliability requirements for Rubin Observatory operations, and so the project has installed fiber optic-based networks while retaining the microwave system as a backup. The system also has redundancy and "diverse" paths in every link from La Serena to the USDF.

62 https://docushare.lsst.org/docushare/dsweb/Get/LSE-30
63 https://docushare.lsst.org/docushare/dsweb/Get/LSE-61
64 https://docushare.lsst.org/docushare/dsweb/Get/LSE-78

6.3.2.8.1.7.2 Network Architecture and Design

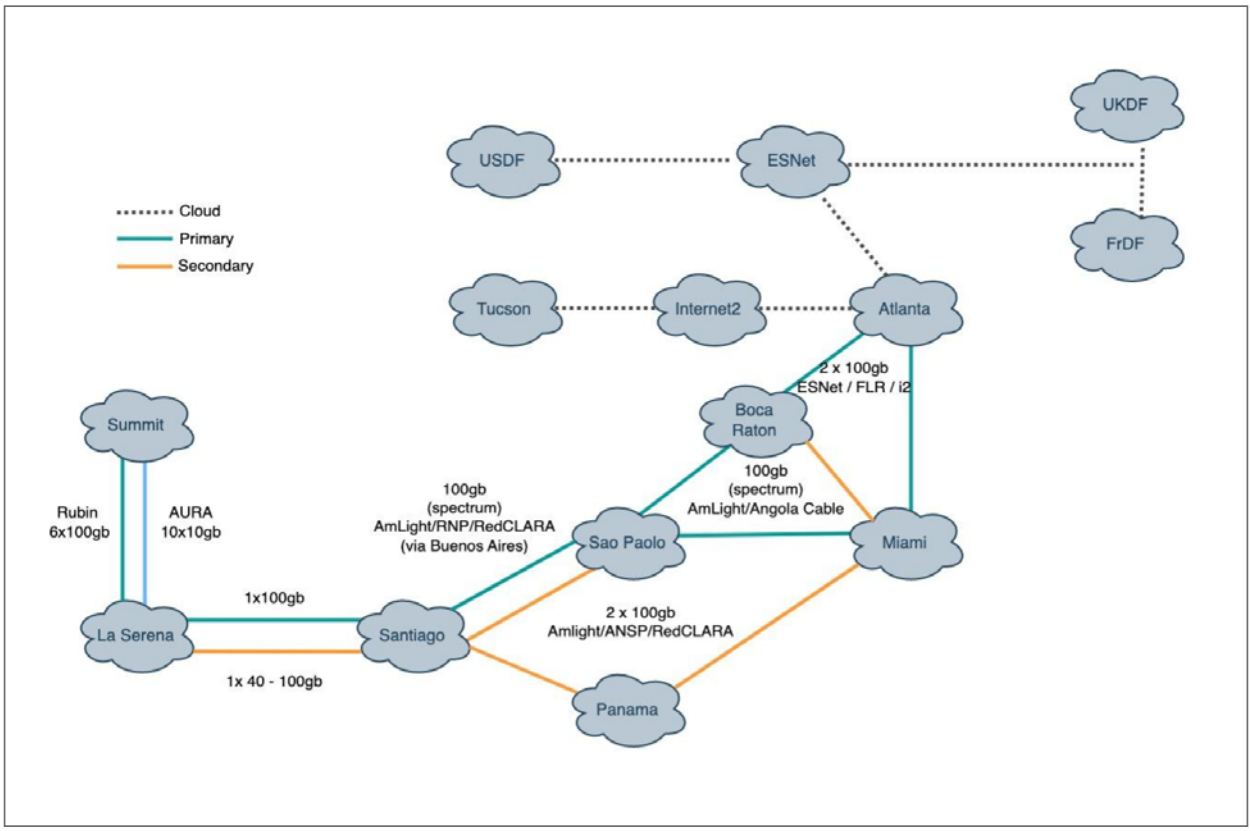


Figure 6.3.8: Rubin Observatory network paths and capacities.

6.3.2.8.1.7.2.1 Summit to Base Link

The key driving requirements for the Summit to Base communications are the bandwidth and reliability required to transfer the image data and associated EFD metadata for alert processing, to transfer the raw image data to the Base Facility, and to handle OCS commands and control traffic (see Figure 6.3.9 below).

Rubin Observatory has installed an optical fiber pair between the Summit and Base Facility utilizing dense wavelength division multiplexing technology to provide 6 x 100 Gbps lambdas with end nodes owned and operated for Rubin Observatory by NOIRLab COS-ITOps.

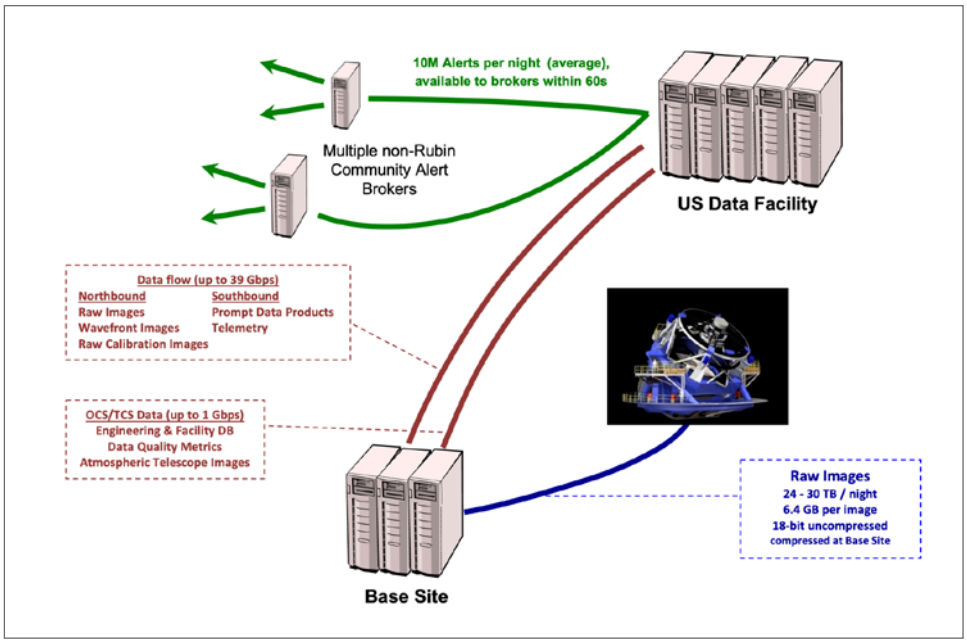


Figure 6.3.9: Nightly flow of data to support prompt processing. Data products are described in detail in Section 18 (Case Study section above).

This capacity is divided into a minimum of 100 Gbps for image data transfer and 100 Gbps for OCS command and control data. A second pair of fibers carrying non-Rubin AURA traffic will be available for failover.

The historical data for fibers in north-central Chilean mountainous areas since 2018 suggest a mean time between failures of 340 days and a mean time to repair of 2 days. The availability estimate on this segment is 98%. Therefore, the summit facility is capable of storing a week's worth of data to mitigate network downtime events between the summit and the base. During such downtime, the existing microwave backup link will be used as a survival link to check on the status of core network devices. Rubin has deployed a wireless link to the summit to provide basic internet services (no data link) as an out of bounds service independent of the fiber.

6.3.2.8.1.7.2.2 Base to USDF Link

The required peak bandwidth from the Base Facility in La Serena to the USDF is established by the need to transfer the images for prompt processing (see Figure 6.3.9).

In order for the USDF to release alerts of transient and variable sources within 60 seconds of image readout from the camera electronics, seven seconds of this time budget is allocated to data transfer, including compression/decompression. Two seconds of that budget is in the Summit to Base transfer, and the other five seconds is in the Base to Archive Transfer and compression/decompression. This equates to a minimum of 40 Gbps bandwidth.

Current agreements between NOIRLab and other South American observatories have secured dual, PoP- and path-diverse leased 100 Gbps links between Chile and the US and an additional 100 Gbps spectrum link. The latter is expandable with purchased equipment upgrades at Rubin Observatory's discretion. The non-spectrum links are shared and have a guaranteed minimum for Rubin of 2 x 40 Gbps during observing and 2 x 20 Gbps during non-observing times. These agreements have been achieved at the equivalent cost of a single 10 Gbps link if acquired via non-partner commercial leased services.

Networking between the Base and the USDF rests on provisioning initially provided by the Major Research Equipment and Facilities Construction construction project and is sustained and enhanced as needed by the operations project. This networking includes

- One 100 Gbps lambda and one 40–100 Gbps lambda on physically distinct footprints from the Base Facility to Santiago, Chile;
- One 100 Gbps lambda running up the west coast of South America, landing in a nuclear- and hurricane-hardened facility in Miami, Florida;
- One 100 Gbps lambda from Santiago up the eastern side of South America, landing in a nuclear- and hurricane-hardened facility in Miami, Florida;
- One 100 Gbps spectrum[65] from Santiago up the eastern side of South America, landing in Boca Raton, Florida;
- Networking provided by Florida LambdaRail to a demarcation point for national backbone research networks;
- Networking provided by ESnet national research backbone network to SLAC, via Atlanta; and to CC-IN2P3;
- The networks will be hardware encrypted on each end (Summit, USDF).

All the network links are being tested by Rubin Observatory and Network Partners at multiple levels during construction/commissioning, including Installer tests supervised by Rubin and Partners (test reports available), demonstrations and end-to-end tests (Rubin Observatory Network End-to-End Test Plan, Document-14789), Verification tests (Rubin Observatory Network Verification Plan, LDM-732), and Commissioning tests (Rubin Observatory Commissioning Plan, LSE-79).

This infrastructure provides general research connectivity to the US for Rubin Observatory. The networking supports creation of bandwidth-protected channels that provide a guaranteed minimum rate with extra idle bandwidth available to the channel. Networking to and from the computing center in France is provided by ESnet in the US, and by both GÉANT and Renater in Europe.

65 Spectrum refers to the right to use a given frequency band on a fiber, using end equipment of one's choosing, such that the capacity of the link can be increased by the deployment of more capable equipment. This choice is up to the "owner" of the spectrum.

6.3.2.8.1.7.2.3 Archive to DAC and EPO Center to User Links

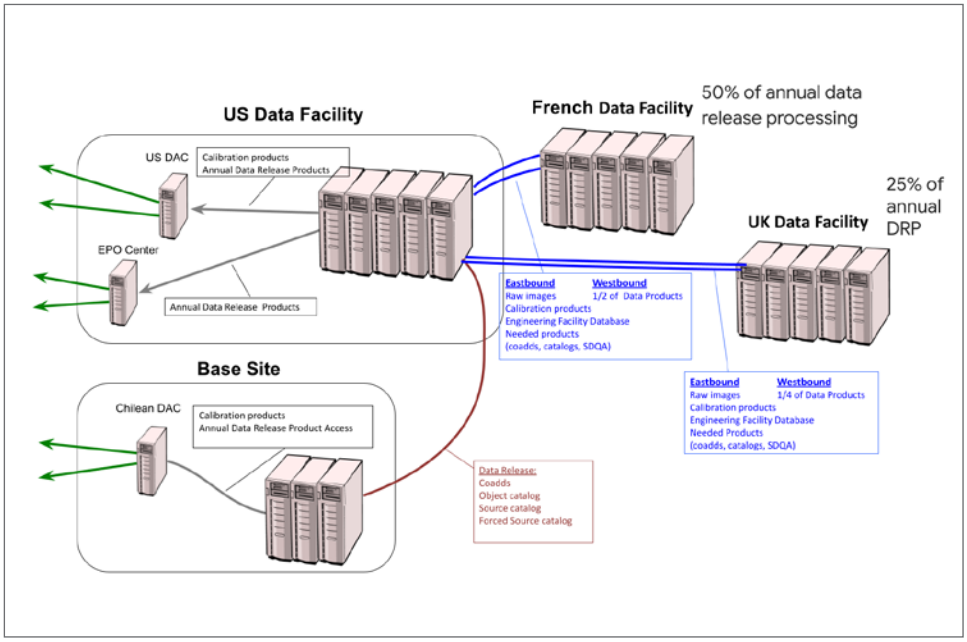


Figure 6.3.10: Data flows during processing and distribution of calibration and data-release products.

Rubin Observatory must provide data access to the US, Chilean, and international contributor scientific communities and to a worldwide community of non-specialist users whose curiosity may be triggered by Rubin Observatory education and public outreach (see Figure 6.3.10). User access to the DACs and EPO Data Center will be via public and R&E network connections (e.g., Internet and Internet2, NSF ACCESS, ESnet), and the aggregate bandwidth will be limited only by the connectivity of the hosting and using institutions. In cases of standalone DACs or science centers funded outside the project, the entity developing and operating the center will be responsible for providing network connectivity to Rubin Observatory USDF to enable data transfer.

Rubin Observatory network architecture will include identity management, cybersecurity management, and privilege limitations on usage of any resources (e.g., computing, storage, networks, software, and data) in a resource management layer on top of the network. Rubin Observatory network management will include "taps," firewalls, and other security mechanisms to enable this resource management.

6.3.2.8.1.8 Configuration Control, Test, and Release Procedures

The Data Management Configuration Control Board (CoCB), which will be chaired by the AD of Data Management or their designate, is responsible for change control and establishing testing and release procedures across the Data Management department.

The CoCB is responsible for managing changes to the Data Management configuration baseline, which specifies which technologies and which versions of software are deployed within the DMS. Any changes to the baseline, including deploying new releases of scientific processing pipelines, must be approved by the CoCB. The CoCB will maintain the software release schedule, and will evaluate bug and problem reports, determining if fixes must be deployed immediately, or if they can be deferred until a future scheduled software release or maintenance window.

All changes will go through the full set of automated testing. Each data facility will have an efficient (and, where possible, self-service) method for deployment of any released software to test systems in the data facility. These will support at least two types of test:

1. Full regression testing for normal releases—a process that may take many days; and
2. Bug testing for specific quick fixes with a limited regression.

In exceptional circumstances, "quick fix" deployments may be made which bypass regular CoCB procedures and/or avoid full regression testing; these would always require explicit approval from the AD of Data Management.

6.3.2.9 Emerging Technology Use

6.3.2.9.1 Cloud Services

As mentioned above, in pre-operations, Rubin Operations deployed a commercial cloud-based IDF on the GCP. Rubin Data Production software and precursor data sets were deployed to the IDF to allow Rubin staff to develop processes and train for operations readiness.

The three-year IDF program was also used to deploy pre-operations data sets known as Data Previews: early previews of what operations annual data releases will look like, including data products, science analysis tools, user support, and documentation. The first Data Previews have been based on simulated "Data Challenge 2" (DC2) data from the DESC. Subsequent previews will include data obtained in commissioning and science verification.

In FY24 the IDF transitioned to the hybrid-cloud US DAC model, with the user-facing RSP; see Sec. 7.1.3 in Case Study Section 7 above) remaining in Google Cloud and the back-end data holdings and data services hosted at the USDF at SLAC (as described in Sec. 8.1.4.4 under Case Study Section 8 above.) The RSP is implemented in a virtualization framework that is portable across cloud providers or Kubernetes-based on-premises infrastructure. Multiple other scientific services, applications, and databases (including a staff-facing instance of the RSP) run within a Kubernetes cluster environment at the USDF.

6.3.2.9.2 AI and ML

Currently the Rubin Data Management system does not make heavy use of AI/ML for data production or distribution. The scientific community is expected to deploy their own AI/ML techniques in analysis of Rubin's scientific data products for modeling, inference, and anomaly detection. This expectation is a major driver for bulk download capabilities for Rubin data products, and may also influence the development roadmap for Rubin-supported analysis platforms on five-year timescales. Multiple community alert brokers currently provide users with AI/ML capabilities for classification and anomaly detection.

It is possible that some elements of the Rubin Science Pipelines may be migrated in the future to use AI/ML techniques and/or hardware acceleration such as with GPUs.

6.3.2.9.3 Wireless or Distributed Location Science Activities

Rubin staff use wireless networking extensively for routine productivity and communications purposes. In addition to wireless networks in standard laboratory / office spaces, Rubin requires wireless networking at the (remote) summit facility. This networking is currently provided by the Rubin project itself, in conjunction with NSF NOIRLab.

6.3.2.10 Data-Related Resource Constraints

Known or potential data-related resource constraints as of FY24 are:

- Insufficient computing resources available to meet the full analysis demand from the Rubin scientific user community, and/or difficulty in co-locating Rubin data products with available computing resources. Includes unbounded/unscoped demand for AI/ML-based analysis of Rubin data.
- Changing cybersecurity threat and policy landscape can present barriers to planned data-access and data-sharing services, and/or can introduce new costs, technical requirements, and schedule impacts.
- Challenges associated with implementing and routinely operating the distributed processing model for annual data-release processing.
- Challenges and constraints of operating in a shared data facility with other experiments.
- Network and I/O performance impacts of transferring a very large number of relatively small files.

6.3.2.11 Data Mobility Endpoints

Information about data mobility and sharing is given above in Section 2 (Collaborators), Section 3 (IRI Readiness), Section 7 (Software Infrastructure), and Section 8 (Network and Data Architecture).

Plans for bulk download of Rubin data are currently under development. The baseline expectation is to use the Rucio software framework to distribute science data from the USDF to other data-access centers and aFs. A draft plan for supporting bulk downloads is available as RTN-086.[66]

6.3.2.12 Outstanding Issues

Rubin will be in a commissioning phase during FY25, with multiple major milestones still remaining in advance of regular survey operations. In addition, "unknown unknowns" will continue to arise.

A high-level view of commissioning milestones can be found in DMTN-232[67]. This document is regularly updated.

6.3.2.13 Facility Profile Contributors

Vera C. Rubin Observatory Representation

Jeff Kantor, Rubin Observatory, jkantor@lsst.org

Will O'Mullane, Rubin Observatory, womullan@lsst.org

Bob Blum, Rubin Observatory, rblum@lsst.org

Phil Marshall, SLAC, pjm@slac.stanford.edu

Amanda Bauer, Rubin Observatory, abauer@lsst.org

Adam Bolton, SLAC, adam.bolton@slac.stanford.edu

ESCC Representation

Mark Foster, SLAC, fosterm@slac.stanford.edu

Phil DeMar, Fermilab, demar@fnal.gov

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

66 https://rtn-086.lsst.io/v/u-abolton-bulkdownload/index.html
67 https://dmtn-232.lsst.io/

6.4 DESC

DESC is the international science collaboration that will make high-accuracy measurements of fundamental cosmological parameters using data gathered from the LSST produced by Vera C. Rubin Observatory, and investigate the nature of dark energy and dark matter. Over ten years, the observatory will image the southern sky (18,000 deg^2) hundreds of times in six broad-band filters, accumulating a time-lapse 'movie' of the sky, as well as obtaining extremely sensitive measurements of billions of galaxies in the coadded image data. The observatory, with an 8.2 m primary mirror, a remarkably large field of view (approximately 10 square degrees), and the largest-ever digital camera (3.2 gigapixel, with a two-second readout), is designed to enable Rubin through its LSST to revolutionize optical cosmology and many fields of astronomy. DESC will make high-accuracy measurements of fundamental cosmological parameters, and derive cosmological constraints, using data from the LSST.

6.4.1 Discussion Summary

The DESC project is investigating the use of Rucio and FTS, as is the Rubin Observatory, to see if it would be a good fit for the data management aspects of the project. The use of any software comes with risks: the project is striving to ensure that individual users can easily interact with the systems they create, and will gravitate toward things that are easy to install, use, and support long term. The latter observation holds true for managing access for scientific users: the lack of a cohesive multi-institutional method to manage credentials is also a challenge for this collaboration. DESC hopes this is an outcome of IRI, as they stand to have numerous users that will be impacted by this issue as they use computing and storage resources that could span facilities.

DESC is exploring ways to perform multimodal analysis using other streams of research data. Some possibilities include the European Space Agency's Euclid mission, and the Nancy Grace Roman Space Telescope. The effort is still in its early stages, and there are open questions about ways to get access to the data, and how it can be integrated through software and computational hardware. A general question that was raised was that the portal system for these three data products (and in general, all individual collaborations that produce a data set that others may want to use) is that there will be no uniformity in how the data can be retrieved. This causes a challenge for the greater scientific community.

DESC's major data movement activity, occurring on a yearly basis, is a bulk data download from SLAC to NERSC. The network infrastructure in place between the facilities (via ESnet) is sufficient to handle the expected volume of data, and SLAC is in parallel moving to larger 400 Gbps capabilities to support other efforts. Since this is not a real-time operation, there is no time limit (or performance expectation) for transfer. DESC is open to the idea of working on structured data challenges, as other communities have done in the past. This could involve simulation data to start, and characterize performance between SLAC, NERSC, and the IDACs. There has been observed poor performance to some sites in the EU from NERSC, but given this is not a part of the critical path of science, it is a lower priority for DESC to address.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The DESC is the international science collaboration that will make high-accuracy measurements of fundamental cosmological parameters using data gathered from the LSST produced by Vera C. Rubin Observatory, and investigate the nature of dark energy and dark matter over ten years. DESC will make high-accuracy measurements of fundamental cosmological parameters, and derive cosmological constraints, using data from the LSST. [Section 4.1]
- The DESC collaboration currently comprises some 1,270 members, of whom about 250 are full members. These originate in 23 countries and more than 200 institutions. Of the members, about 50% are from the United States, 9% from the United Kingdom and 7% from France. About 520 of the members are graduate students or postdocs. [Section 4.1]

- DESC data volumes are expected to reach more than 10PB by the end of the project, with approximately 1–2 PB originating from the Rubin Observatory with each yearly data-release. Additional data volume will come from simulations produced at the DOE HPC facilities, which are expected to be several 100 TB, and data catalogs produced during the analysis process, which are estimated to be 20 TB. [Section 4.2]
- DESC has tested Globus transfers of tens and hundreds of TB between NERSC and international sites, with transfer rates of 400 MB/s and 180 MB/s respectively. [Section 4.2]
- DESC simulation activities will require data movement from other computing centers to NERSC. Data transfers will be on the order of 100 TB initially. When the Rubin Observatory Data Previews are present at SLAC, an additional 10–100 TB will be transferred to NERSC; this is expected in May 2025 and early 2026. The goal is to support PB-scale transfers between the facilities, and potentially to international collaborators to prepare for full production. [Section 4.2]
- Once in full operations, major Rubin data releases for static cosmology analyses will be generated yearly. The bulk of data will still originate with Rubin, and will be of the order of 1–2 PB to be transferred every two to three years, and, assuming some data reduction strategies, could eventually be yearly. [Section 4.2]
- DESC is exploring the use of IDACs for the Rubin observatory, which are being developed as in-kind contributions to Rubin Observatory by a number of institutions around the world. Some of the IDAC resources will be made available to DESC and other LSST science collaborations for science analyses and could potentially provide specialized capabilities. [Section 4.3]
- Simulated data sets have been produced at the US and France Data Facilities and transferred through http, rsync and Globus. DESC foresees changing the data management model to employ Rucio over FTS to integrate better with the Rubin Project's systems. [Section 4.3]
- DESC analyses will be carried out primarily at NERSC. During this period, DESC foresees testing and carrying out workflows across multiple sites as supplementary computing, trying to minimize data movement. Specific tasks might be run completely outside of NERSC, such as those related to imaging, details still to be defined (and summary data products or results would then need to be moved to NERSC). In the next two to five years, NERSC will be migrating to NERSC-10, and DESC will be testing the incorporation of IDACs and other IRI platforms to these analyses. Large efforts in simulations in parallel to data analyses will be required using all these resources. [Section 4.4]
- NERSC is the main computing platform for DESC, and SLAC is the primary data source for Rubin Observatory data. The most critical component of this activity is an efficient transfer between these sites. ESnet current connectivity between these sites is sufficient for DESC's early needs, and planned upgrades in future years should ensure the data volumes are managed efficiently. [Section 4.4]
- DESC transfers have largely been via HTTP, rsync, and Globus for data sets moving from NERSC to and from collaboration sites. DESC will explore in this period the use of Rucio over FTS following the steps of the Rubin project for SLAC-to-NERSC transfers, and possibly for its own purposes in interactions with IDACs. DESC is currently using Parsl for workflow management of image reprocessing tasks as well as certain simulation and analysis workflows, and we have funded support for this. DESC has good experience with other workflow tools and has been able to work closely with the Parsl team on desired features to run distributed processing with minimal human oversight. [Section 4.5]

- All DESC science activities are intrinsically 'remote' as the project is reliant on data obtained by the Rubin Observatory, processed and distributed from the USDF at SLAC, with subsidiary copies in other points of the world (France, Chile, UK). In the general case DESC will be pulling data from SLAC to NERSC with a yearly cadence; a smaller volume alert stream will be monitored coming from diverse brokers that cull the original kHz stream of transient source alerts from the Observatory. [Section 4.5]
- DESC's current analysis model involves bringing collaborators to deploy their analyses at NERSC to run at scale and access collaboration data products. DESC is planning to expand these capabilities and has had successful tests and large campaigns at the UK and France Data Facilities for Rubin, though details for securing these resources for DESC science purposes are being settled. These resources could potentially be used for DESC workflows that require Rubin imaging products available at these sites. [Section 4.5]
- DESC's primary data source is the Rubin USDF at SLAC. Workflows will primarily involve heterogeneous analysis operations over either the processed data products from the alert stream of transient events, or the data set extracted from the accumulated images and catalogs. Nominally, most of DESC's computing will take place at NERSC; moving selected datasets from the Rubin USDF at SLAC to NERSC is a priority. As the survey proceeds, DESC will likely need to obtain additional resources elsewhere through specific programs or through international collaborations. [Section 4.6]
- The anticipated start of operations at the Rubin Observatory is FY25Q4. In this period DESC expects to bulk download from the primary data source (USDF at SLAC) to NERSC the released datasets corresponding to the commissioning and science verification periods. The specifics of the data set sizes, files, etc. are uncertain due to the nature of these pre-operations' activities and final scope. DESC expects the resulting data size will be of a similar size as Rubin's recent data preview, and of the order of 10 TB of catalogs, with the possibility of transferring up to 1–2 PB of imaging data. DESC's main computing resource will continue to be NERSC Perlmutter, as well as the Spin platform as a database to store selected streamed data from the alerts brokers. [Section 4.6]
- DESC has no concrete plans for using any DESC-specific Cloud applications. It is worth noting that a good part of the initial engagement with the data by DESC scientists, in terms of image visualization, exploratory analysis, will happen in the US DAC, through an interface called the RSP. [Section 4.8]
- Due to the large variety of science applications and multiple approaches and stages of DESC's analyses, ML will be applied in many instances of large-scale computation (CPU and GPU) at NERSC. [Section 4.8]
- DESC sees a potential for certain workflows that are more compute intensive (such as image simulations and Markov Chain MC analyses of cosmology parameter space) to branch out across multiple sites in the future. In the past, working across different facilities usually meant specific workflow setups tailored to the infrastructure/hardware available at each facility. The IRI vision will help overcome this hurdle and provide users with a very similar environment across facilities to ease moving across platforms. [Section 4.9]
- DESC workflows are heterogeneous but primarily fall under two categories: [Section 4.9]
    - The great majority (>90%) both in terms of compute and storage will correspond to long-term campaigns. These will include the processing and reprocessing of LSST data releases to extract data vectors, analyze systematics, model the signal, and infer cosmological parameters. They will be complemented by cosmological and image simulations in a similar amount. Therefore, these require "sustained access to resources at scale over a longer time

to accomplish a well-defined objective." Reliability and reproducibility of results at our computing platforms are key for the success of DESC. In this regard, compute-intensive multisite workflows would be an enormous benefit given possible allocation constraints to reach enhanced goals for DESC science.

    - Some of our workflows are time sensitive. These rely on the Rubin 'alert stream', or intermediary brokers, for astrophysical events and generate actions or outputs that need to be processed within some time period (typically few minutes to few days, depending on the nature of the source). If the sources are not monitored reliably through our selected computing platform(s) in the best case the result will be a delay or degradation of the information obtained about the source, and in the worst case, the loss of a relevant or unique astronomical event.

- For the full-fledged Rubin Observatory data releases, expected in the two- to five-year time frame, DESC will see a limitation in some of the science cases which require full access to the complete single-visit image data set (only available at SLAC or CC-IN2P3, but not NERSC as currently planned) rather than just the galaxy catalogs. DESC is currently looking for mitigations to this problem such as limiting the size of transferred data to NERSC but enabling the analyses nonetheless or shipping software containers to run remotely at these facilities instead. In addition, large-scale image simulations produced at NERSC or elsewhere would possibly stress DESC's small-latency storage allocation at NERSC, as well as compete with computing allocation for its analyses in parallel to real data analyses. The transition to NERSC-10 in this period and its unknown impact on established pipelines also introduces some uncertainty. [Section 5.4]
- DESC has two outstanding issues to work through in the coming years: [Section 5.9]
    - DESC needs technical support in implementing data distribution solutions (Rucio, DIRAC). Standing up these web services requires close interaction with the facilities (e.g., NERSC). Built-in facility support for this software (e.g., just like the facilities provide built-in support for Globus) would be very beneficial.
    - DESC faces complex issues with identity and account management across various services/sites. There are two broad classes of sites. First, there are sites like NERSC where a DESC member has a NERSC account that lets them access both NERSC and Globus. Second, there are sites like in the UK, France (CC-IN2P3), and Rubin IDACs where the DESC needs to deploy a proper VO and have a proper grid certificate. These items are needed to enable interactions with grid software like Rucio and DIRAC. Ideally, the VO and certificates would be linked to each member's NERSC account. Further, we would ideally want to be able to target the NERSC queues using grid software like DIRAC so that we can run jobs, move data, etc., across multiple sites transparently.
- DESC analyses will eventually employ the galaxy catalogs, but also the full imaging data of approximately 10PB, in the later years of the project. Replicating even the high-level imaging data at NERSC is inefficient and unfeasible, due to the scale of the expected allocation. DESC expects to embrace a fully distributed analysis at different facilities. Workflows will be adapted, preferably through IRI approaches, to adapt to a more flexible computing environment. [Section 5.9]
- Future DESC needs, more than five years out, will likely require that the analysis process employ not only the complete galaxy catalogs, but also the full imaging data of O(10s) PB. Replicating even the high-level imaging data at NERSC is inefficient and unfeasible, due to the scale of the expected allocation. Thus, embracing a fully distributed analysis at different IRI and international sites, through grid solutions such as DIRAC, could be an option to offset some of these allocation difficulties. [Section 5.9]

6.4.2 DESC: Experiment Profile

Key planned DESC analyses of LSST include cosmological constraints on the matter content and amplitude of clustering from the combination of correlation functions of the weak lensing and clustering observables (the so-called '3x2pt' analysis), and constraints from only the weak lensing shear and clustering 2-point correlation function; cosmological constraints on the equation of state of dark energy and the matter density from the large LSST supernova sample; and parameter constraints from the abundance of clusters and cluster lensing. DESC will also test for tensions in inferred parameters from LSST to those derived from other experiments (such as those seen today in estimates of the Hubble parameter and clustering amplitude derived from different probes).

6.4.2.1 Science Background

As was the case with precursor photometric surveys, the image dataset is extremely rich and offers various avenues for exploring the core science cases. In particular, the Rubin LSST will provide DESC the ability to do cosmology in the time domain on an unprecedented scale. Millions of transient events are generated and selected per night of observation to provide $O(10^6)$ of cosmologically relevant objects at the end of the survey. In addition, the analysis of wide and deep accumulated images will lead to static catalogs of galaxies that will be released yearly by the Rubin Observatory, probing large volumes and enabling cross-correlation science with other current and planned surveys in different wavelengths.

The primary data source is the Rubin USDF at SLAC. Workflows will primarily involve heterogeneous analysis operations over either the processed data products from the alert stream of transient events, or the data set extracted from the accumulated images and catalogs. Nominally, most of the computing will take place at NERSC, so moving selected datasets from the Rubin USDF at SLAC to NERSC is a priority. However, as the survey proceeds, DESC will likely need to obtain additional resources elsewhere through specific programs or through international collaborations.

6.4.2.2 Collaborators

The collaboration currently comprises some 1,270 members, of whom about 250 are full members. These originate in 23 countries and more than 200 institutions. Of the members, about 50% are from the United States, 9% from the United Kingdom and 7% from France. About 520 of the members are graduate students or postdocs.

DESC's current analysis model involves bringing collaborators to deploy their analyses at NERSC to run at scale and access collaboration data products. DESC is planning to expand these capabilities and has had successful tests and large campaigns at the UK and France Data Facilities (CC-IN2P3) for Rubin, though details for securing these resources for DESC science purposes are still being settled. These resources could potentially be used for DESC workflows that require Rubin imaging products available at these sites. Some of the computing and storage allocation at NERSC could then be offloaded. Additional potential resources for DESC are the IDACs for the Rubin observatory, which are being developed as in-kind contributions to Rubin Observatory by a number of institutions around the world. Some of the IDAC resources will be made available to DESC and other LSST science collaborations for science analyses and could potentially provide specialized capabilities (e.g., complementary databases, Hadoop platforms, opportunistic resources for peak times at NERSC).

Until now, data sets (generally simulations) have been produced at the US and France Data Facilities (CC-IN2P3) and transferred through http, rsync and Globus. As detailed later, DESC foresees changing the model to use Rucio over FTS to integrate better with the Rubin Project's systems. DESC has begun working directly with Rubin on this, capitalizing on their expertise.

DESC has detailed the current approach to the IRI in the IRI-specific section below. Of particular note, DESC utilizes DOE LCFs (both ALCF and Oak Ridge Leadership Facility, OLCF) for large simulation campaigns. The simulation products are currently transferred to NERSC via Globus.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
NERSC, MAIN COMPUTING SITE	Partial secondary	Data portal and data transfer	Coadd catalogs (20 TB), a fraction of the coadded images (O(1–2 PB))	Ad hoc	No	Rucio deployment needs admin permissions
SLAC (RUBIN USDF), MAIN DATA SITE	Primary	Data portal and data transfer	DESC defers to Rubin section	Yearly	Yes (20 TB) through FTS or Globus	
IDACS, INCLUDING UK AND FRANCE (CC-IN2P3) COMPUTING RESOURCES	Partial secondary	Data portal and data transfer	Coadd catalogs (20 TB), possibly coadd images in France/UK	Yearly	Possibly	Establishing a DESC VO for authentication and multisite processing will be complex as it requires coordination with Rubin (data rights).
ALERT BROKERS AT SEVERAL SITES ACROSS THE WORLD	Partial primary	Data transfer, data portal, and Kafka stream	N/A	Continuous	No	Requires reliable platform to receive the data stream, database infrastructure to store a subset of the data at NERSC
ALCF	No	Data transfer	O(100TB)	Ad hoc	No	
OLCF	No	Data transfer	O(100TB)	Ad hoc	No	

Table 6.4.1: DESC Data Mobility

All data volume quantities are estimates of what is expected for Year 1 dataset (in 2027).

The Rubin Observatory will be supported by a network of IDACs which will replicate and serve Rubin LSST data products and broaden access in the worldwide data rights community.

Alert brokers are streaming data sources around the world that have culled the primary alert stream from the USDF.

ALCF to date has been used to generate simulation products that have been transported to NERSC on a one-off basis. Eventually, in the long term, certain workflows with relatively small data movement (< a few TB), could be processed through a multisite approach.

6.4.2.3 IRI Readiness

NERSC is the primary computing resource for DESC. However, as the needed computing time and expected available NERSC allocation are anticipated to diverge in a few years' time, DESC is exploring available complementary resources. Some are from the international community — in the past DESC has undertaken simulations and image processing in France at CC-IN2P3. In the past, DESC has also used other DOE computing resources for specific, well-suited needs, such as the ALCF Theta (and other) platforms for large-scale simulation efforts.

As for IRI, DESC sees a potential for certain workflows that are more compute intensive (such as image simulations and Markov Chain MC analyses of cosmology parameter space) to branch out across multiple sites. In the past, working across different facilities usually meant specific workflow setups tailored to the infrastructure/

hardware available at each facility. IRI is supposed to help overcome this hurdle and provide users with a very similar environment across facilities to ease moving across platforms. With regard to data movement, Globus has helped immensely across US facilities to provide an easy-to-use interface that is the same at all facilities. DESC very much hopes that this also will be the case in the future (though only a small number of DESC-relevant facilities support Globus currently). With regard to workflow implementation, there is a lot of work ahead. DESC will follow IRI developments closely and make use of infrastructure developed wherever possible.

DESC workflows are quite heterogeneous but primarily fall under two categories:

- The great majority (>90%) both in terms of compute and storage will correspond to long-term campaigns. These will include the processing and reprocessing of LSST data releases to extract data vectors, analyze systematics, model the signal, and infer cosmological parameters. They will be complemented by cosmological and image simulations in a similar amount. Therefore, these require "sustained access to resources at scale over a longer time to accomplish a well-defined objective." Reliability and reproducibility of results at computing platforms are key for the success of DESC. In this regard, compute-intensive multisite workflows would be an enormous benefit given possible allocation constraints to reach enhanced goals for DESC science.
- Some of the workflows are time sensitive. These rely on the Rubin 'alert stream', or intermediary brokers, for astrophysical events and generate actions or outputs that need to be processed within some time period (typically few minutes to few days, depending on the nature of the source). If the sources are not monitored reliably through selected computing platform(s) in the best case the result will be a delay or degradation of the information obtained about the source, and in the worst case, the loss of a relevant or unique astronomical event.

6.4.2.4 Instruments and Facilities

DESC will be responsible only for analyzing data collected by the Rubin Observatory with possible enhancements coming from the addition of legacy data or concurrent missions through agreements to perform joint processing. DESC is not directly involved in the development of the instrument or in the data-taking process. DESC also has no influence on the facility's computing capabilities. DESC defers to the Vera C. Rubin observatory case study for some of the requested details.

Present-FY26Q4: The anticipated start of the survey is FY26Q1. In this period DESC expects to bulk download from the primary data source (USDF at SLAC) to NERSC the released datasets corresponding to the commissioning and science verification periods (also called Data Previews). The specifics of the data set sizes, files, etc. are uncertain due to the nature of these pre-operations' activities and final scope, but DESC can estimate it will be of a similar size as Rubin's recent Data Preview 0.2 (DP0.2), of the order of 10 TB of catalogs, with the possibility of transferring up to 1–2 PB of imaging data to understand how much DESC will eventually use. In this period, the main computing resource will continue to be NERSC Perlmutter. DESC will be using the Spin platform at NERSC as a database to store selected streamed data from the alerts brokers, the yearly data volume expected to be stored this way is 1–2 TB.

FY27-FY31: Major Rubin data releases for static cosmology analyses will be generated in this period. DESC may incorporate other complementary data sets to enhance the science (both legacy, with small impact in relative data volume, and concurrent data sets, from the Euclid and Roman space observatories, with likewise minor impact in data volume as they will be limited to relatively small areas of the sky). The bulk of data will still originate with Rubin, and will be of the order of 1–2 PB to be transferred every two to three years, and, assuming some data reduction strategies, could eventually be yearly.

FY32 onward: Eventually, analyses would employ fully not only the galaxy catalogs, but the full imaging data of O(10s) PB. Replicating even the high-level imaging data at NERSC is inefficient and unfeasible, due to the scale of the expected allocation. Thus, embracing a fully distributed analysis at different IRI and international sites, through grid solutions such as DIRAC, could be an option to offset some of these allocation difficulties. This

would enable enhanced goals for analyses, such as more thorough exploration of cosmological parameter space through Markov Chain MC processes, or forward modeling of pixel-level data of supernovae and host galaxies — scene modeling photometry — to produce more accurate uncertainties and reduce the systematic error budget on supernova analysis. Another enhanced goal would combine fully the Rubin and Euclid data sets at the imaging level assuming an agreement is reached for this processing.

6.4.2.5 Generalized Process of Science

Present–FY26Q4: This period will see mainly preparatory work for cosmology analyses in DESC, as the first relevant Data Release from the Rubin Observatory for cosmology (key projects) will not be delivered until at least FY27Q4. However, plenty of scientific publications will precede this milestone, e.g., comparative work with previous surveys, calibration work relevant for cosmology, validations of data and methods. These analyses will be carried out at NERSC primarily.

Most of the computational work will still be in the form of simulations and analyses on simulated or precursor datasets (in most cases, without need of massive data movement). An example is forecasting for analysis choices: the analysis of the data will require us making choices on the scales to be analyzed and the impact of marginalization over survey and astrophysical systematics. Soon, DESC expect that the collaboration will run a series of simulated analyses to inform these choices. This is computationally intensive because it requires testing of many different scenarios through realistic likelihood evaluation for many of the probes.

Some very early science cases will run at the end of this period over science verification data, as precursor analysis runs, mainly constrained to NERSC with data-intensive and compute-intensive workflows. For example, the analysis of the alert stream will be used to find faint nearby supernovae very early in Rubin data (ideally within hours of explosion). At these phases, the supernova ejecta will plough through any circumstellar material or binary companion, and the resulting emission can contain signatures of the progenitor system. Identifying these systems rapidly and scheduling target-of-opportunity follow-up will yield insights into the explosion physics, and can potentially refine the models used to infer supernova distances at high redshift, as well as shed light into other modes of stellar death. Targeted spectroscopic follow-up of a sample of low-to-intermediate redshift supernovae can be used to build a small sample of objects with reduced systematic error. These small samples can be used to validate the entire supernova analysis pipeline for Year 1 and Year 2 data.

FY27-FY31: A high-level overview of plans for science with the first full-blown data releases follows:

1. A key project analysis process primarily starts from the annual object catalogs from the Data Release (starting with DR2) or database information built up by DESC from the culled transient stream (known as the FAST DB).
2. For weak lensing analyses, galaxy shapes are the main observable used, combined with redshift measurements inferred from multi-band photometry. Three summary statistics are currently envisaged which correlate galaxy shapes with each other; correlate galaxy shapes with positions; and correlate galaxy positions with each other. These are combined into a set of three 2-pt correlations (data-intensive). DESC also anticipates higher-order statistics being needed to capture non-Gaussian information. This analysis task includes running large-scale image and source injection simulation studies to estimate and account for systematic effects.
3. Analysis of supernovae is based on identifying SNIa from the transients' stream, followed by using the apparent brightness and redshift to determine a recession velocity.
4. Strong lensing is based on finding lensed quasars and supernova systems and then measuring the time delays for various paths of the lensed object using multiyear light curves.

5. Finally, clusters of galaxies are used to predict cosmological parameters based on the relationships between cluster properties and halo mass. These parameters are obtained via a likelihood analysis using mass function predictions obtained from simulations. Many of these analyses are followed up by compute-intensive analysis to explore a multi-dimensional parameter space with Bayesian statistics.

These analyses will be carried out primarily at NERSC. As mentioned in other sections, during this period DESC foresees testing and carrying out workflows across multiple sites as supplementary computing, trying to minimize data movement. Specific tasks might be run completely outside of NERSC, such as those related to imaging, details still to be defined (and summary data products or results would then need to be moved to NERSC).

During this period, NERSC will be migrating to NERSC-10, and DESC will be testing the incorporation of IDACs and other IRI platforms to these analyses. Large efforts in simulations in parallel to data analyses will be required using all these resources.

FY32 onward: DESC expects workflows to eventually be distributed across multiple sites, with the exact technology to be determined (see also next section). Major cosmology analyses will be undertaken with LSST year 4 and year 10 data releases (starting in approximately 2032 and 2038, respectively and extending for two or three years). DESC also anticipates an increasing level of joint analyses of LSST data with survey data from space-based observatories Euclid and Roman Space Telescope during this period, the scope of which remains to be determined.

6.4.2.6 Remote Science Activities

As mentioned above, all DESC science activities are intrinsically 'remote' as DESC is reliant on data obtained by the Rubin Observatory, processed and distributed from the USDF at SLAC, with subsidiary copies in other points of the world (France, Chile, UK). In the general case DESC will be pulling data from SLAC to NERSC with a yearly cadence approximately (keeping in general only the latest data release). In addition, an alert stream will be monitored coming from diverse brokers that cull the original kHz stream of transient source alerts from the Observatory.

Present-FY26Q4: Most of the current work in DESC is being done with simulated data which already exist at NERSC or are being created at NERSC or at other computing facilities. In general, the simulation data are already curated where the simulation is carried out to reduce transferred data volumes considerably. However, in this time frame, simulation campaigns will need to be created for DESC to explore cosmological parameters and to generate covariance estimates. This will mean that data will have to be moved from other computing centers to NERSC. Data transfers will be O(100) TB. During FY25 and FY26, data from the Rubin Observatory will start appearing at SLAC and DESC will perform test data transfers at first (Data Preview 1, O(10) TB, expected May 2025) and fully run the data transfer system with Data Preview 2 (O (100 TB), expected in early 2026). In this period, DESC would exercise small transfers (O(10 TB) in 2025, progressively towards 1–2 PB scale) and possibly coordinated workflows (2026) with international collaborators, in preparation for full production.

FY27-FY31: In this period DESC will transfer the first dataset relevant for the DESC Key Projects as described in the Science Plan. This data set is expected to be of the order of 1–2 PB, after some optimization. These data transfers will happen more or less yearly. Most of the primary work will happen at NERSC but supplemented with information transferred from other experiments. In most cases, these will be O(10) TB sized catalogs (real data and simulations) and on occasion smaller imaging datasets. As for analysis workflows, though the primary work will happen at NERSC, DESC will explore grid-like approaches and multisite workflow management with previous, smaller data sets, and effectively deploy them in this period.

FY32 onward: Full implementation of multisite workflows, for image processing, Markov Chain MC to benefit from the Rubin/IRI computing ecosystems. For example, DESC conceives using DIRAC as a grid-based option for this. IDACs will be used to offload cycles for debugging or complementing NERSC jobs, which would still be the primary facility.

6.4.2.7 Software Infrastructure

Present-FY26Q4: With simulated data totaling a few hundred TB, transfer has largely been via HTTP, rsync, and Globus for data sets moving from NERSC to and from CC-IN2P3 (France), ALCF, OLCF, and SLAC. DESC will explore in this period the use of Rucio over FTS following the steps of the Rubin project for SLAC-to-NERSC transfers, and possibly for its own purposes in interactions with IDACs. DESC is currently using Parsl for workflow management of image reprocessing tasks as well as certain simulation and analysis workflows, and DESC has funded support for this. DESC has good experience with other workflow tools and has been able to work closely with the Parsl team on desired features to run distributed processing with minimal human oversight.

FY27-onward: DESC has no concrete plans for workflow management tools beyond the ad hoc tools DESC has built and Parsl, which will need to be tested at scale with realistic data sets and use cases.

6.4.2.8 Network and Data Architecture

NERSC is the main computing platform and SLAC the data source for Rubin Observatory data. The most critical component is an efficient transfer between these sites. Today, there are at least a pair of 100 Gbps connections between SLAC and NERSC, via ESnet. NERSC has a sum of 1Tb/s connection to ESnet, so if SLAC upgrades their site connection then a higher bandwidth is possible. These upgrades are expected at SLAC in the coming five-year period, as detailed in the SLAC Site Technology profile.

As complementary information, DESC has tested Globus transfers of tens and hundreds of TB between NERSC and international sites in France (CC-IN2P3) and Spain, with transfer rates of 400 MB/s and 180 MB/s respectively.

6.4.2.9 Emerging Technology Use

6.4.2.9.1 Cloud Services

DESC has no concrete plans for using any DESC-specific Cloud applications. It is worth noting that a good part of the initial engagement with the data by DESC scientists, in terms of image visualization, exploratory analysis, will happen in the US DAC, through an interface called the RSP. DESC defers to the Rubin Observatory case study for technical details.

6.4.2.9.2 Artificial Intelligence and ML

Due to the large variety of science applications and multiple approaches and stages of DESC's analyses, there are many instances where ML will be applied. For large-scale computation at NERSC as the primary computing facility, DESC benefits from its large investment in GPU platforms (1792 GPU-accelerated nodes in the current platform Perlmutter). Many of the most widely used packages and libraries are available (TensorFlow, PyTorch and others).

6.4.2.9.3 Wireless or Distributed Location Science Activities

DESC has no wireless networking requirements.

6.4.2.10 Data-Related Resource Constraints

Present-FY26Q4: In this period DESC does not expect major bottlenecks in either the data-transfer throughput or available computing time. In the maximal case, DESC might expect up to O(1–2) PB of data transfer between SLAC and NERSC (Rubin's Data Previews 1 and 2, corresponding to commissioning and science verification data, including images and intermediate products) and a moderate increase in compute allocation to deal with first exploratory and proof of principle analyses.

FY27-FY31: For the full-fledged data releases, DESC will see a limitation in some of the science cases which require full access to the complete single-visit image data set (only available at SLAC or CC-IN2P3, but not NERSC as currently planned) rather than just the galaxy catalogs. DESC is currently looking for mitigations

to this problem such as limiting the size of transferred data to NERSC but enabling the analyses nonetheless or shipping containers to run at these facilities instead. In addition, large-scale image simulations produced at NERSC or elsewhere would possibly stress small-latency storage allocation at NERSC, as well as compete with the computing allocation for its analyses in parallel to real data analyses. The transition to NERSC-10 in this period and its unknown impact on established pipelines also introduces some uncertainty.

FY32 onward: A full-fledged implementation of multisite processing and possibly increased computing allocations at the USDF or France (CC-IN2P3) would be the final goal to realize the full scale of DESC analyses.

6.4.2.11 Data Mobility Endpoints

Major transfers from data sources (currently, mostly simulated data) has happened through https from the Google Cloud (where the Rubin USDF was being hosted temporarily), and through rsync and Globus from CC-IN2P3 in France. The current plan is to adopt Rucio and FTS in a way similar to what the Vera Rubin Observatory has done to manage its data sets, for two purposes:

- Transfer data from the USDF to NERSC after each yearly data release.
- Synchronize subsets of data or value-added data sets DESC generates across sites with which DESC has an agreement to cooperate in analysis or host copies of data for DESC scientists.

The back-up plan, however, is to use Globus. DESC would use NERSC DTN as the main endpoint which has access to the filesystems of interest.

6.4.2.12 Outstanding Issues

There are two outstanding issues.

- DESC needs technical support in implementing some of data distribution solutions (Rucio, DIRAC). Standing up these web services requires close interaction with the facilities (e.g., NERSC). Built-in facility support for this software (e.g., just like the facilities provide built-in support for Globus) would be very beneficial.
- Second, the DESC faces complex issues with identity and account management across various services/sites. There are two broad classes of sites. First, there are sites like NERSC where a DESC member has a NERSC account that lets them access both NERSC and Globus. Second, there are sites like in the UK, France (CC-IN2P3), and Rubin IDACs where the DESC needs to deploy a proper VO and have a proper grid certificate. These items are needed to enable interactions with grid software like Rucio and DIRAC. Ideally, the VO and certificates would be linked to each member's NERSC account. Further, we would ideally want to be able to target the NERSC queues using grid software like DIRAC so that DESC can run jobs, move data, etc., across multiple sites transparently.

6.4.2.13 Facility Profile Contributors

DESC Representation

Debbie Bard, NERSC, djbard@lbl.gov

Keith Bechtol, University of Wisconsin-Madison, kbechtol@lsst.org

Matthew Becker, ANL, mrbecker@anl.gov

Elisa Chisari, Utrecht University, n.e.chisari@uu.nl

Seth Digel, SLAC, digel@slac.stanford.edu

Katrin Heitmann ANL, heitmann@anl.gov

Renée Hložek, Toronto, hlozek@dunlap.utoronto.ca

Heather Kelly, SLAC, heather@slac.stanford.edu

Gautham Narayan, UIUC, gsnarayan@gmail.com

Ignacio Sevilla, CIEMAT, ignacio.sevilla@ciemat.es

ESCC Representation

Mark Foster, SLAC, fosterm@slac.stanford.edu

6.5 CMB

The CMB consists of the photons created in the Big Bang, propagated through the universe until detected today. As a consequence, the photons that make up the CMB trace the entire history of the universe and provide a unique window on fundamental physics and cosmology. In addition, CMB data sets are extraordinarily sensitive mm-wave surveys, incidentally supporting a wealth of astrophysics and astronomy. The full scope and scale of CMB science is laid out in the CMB-S4 Decadal Survey Report.[68]

CMB-S4 is the "Stage 4" ground-based CMB experiment, bringing the entire US community together to build a single experiment capable of detecting the faintest CMB signals.[69] The four design-driving science cases are:

1. To detect (or very tightly constrain) primordial gravitational waves to test theories of inflation.
2. To detect (or very tightly constrain) additional species of light relic particles.
3. To map the matter in the universe and elucidate the formation and evolution of galaxy clusters.
4. To detect mm-wave transients and add mm-wave data to multi-messenger astronomy.
5. Note that meeting these will necessarily also enable the full scope of other CMB science.

6.5.1 Discussion Summary

CMB-S4 will leverage resources that exist in Chile today. The Simons Observatory (SO) and Atacama Large Millimeter/submillimeter Array (ALMA) invested heavily into the networking infrastructure over prior years, and has a clean fast path to national and international computing resources. Networking to support experimentation includes multiple strands of fiber optic cables, as well as a perfSONAR monitoring infrastructure. ESnet will continue to work with CMB-S4 on networking capabilities, as well as any overlap with the existing projects.

CMB-S4 is still early in its design and development, and has flexibility in the way it approaches aspects of computing and networking. The project plans to attempt data challenges in the coming years (with ESnet support), and will also explore all available options for the software used in data management, as well as available computational resources (HPC, Grid, or Cloud). They are interested in aspects of IRI, in particular how federated identity may ease the transition between multiple ways to compute and store research data for a distributed set of users.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The CMB consists of the photons created in the Big Bang, propagated through the universe until detected today. As a consequence, the photons that make up the CMB trace the entire history of the universe and provide a unique window on fundamental physics and cosmology. CMB-S4, a joint project between the NSF and the DOE, is the "Stage 4" ground-based CMB experiment, bringing the entire US community together to build a single experiment capable of detecting the faintest CMB signals. [Section 4.1]

68 https://arxiv.org/abs/1907.04473

69 Scientific research using CMB data continues to be a priority for DOE and NSF, an initiative that provides crucial information on the fundamental nature of matter and energy in the universe. However, DOE and NSF jointly decided in calendar year 2025 that they can no longer support the P5-envisioned CMB-S4 project. Nevertheless, DOE and NSF will continue to partner with the CMB science community to explore the potential science that can be achieved with limited upgrades to existing experiments to further this important U.S.-led research.

- In mid-2024, the NSF was not able to commit to funding for portions of the project that involved experimentation to be conducted at the South Pole. Thus, the information provided by CMB-S4 in this report assumes the entirety of the observatory will be based in Chile. The experimentation will produce three coordinated surveys from Chile between 2028 to 2035. [Section 4.1]
    - A wide and deep survey of ~70% of the sky using 2 x 6m telescopes that operates with a with a daily cadence
    - An ultra-deep survey of ~3% of the sky using 18 x 0.5m telescopes
    - An ultra-deep survey of the same ~3% of the sky using a different 5m telescope.
- The CMB-S4 collaboration currently consists of 520 members at 142 institutions in 24 countries and 26 US states. The institutions include DOE, NASA, and Department of Commerce laboratories, as well as a host of universities. [Section 4.1]
- In operations, CMB-S4 will gather 1.9 x 10 ^ 8 samples per second, corresponding to a compressed data rate of 2.0 Gbps. Data will be transferred in real time over fiber networks to the primary data center at the DOE's NERSC and copied to the secondary data center at the DOE's ALCF. All data are archived at both NERSC and ALCF. 20 TB of data are expected daily during operations (~7PB of data annually). [Section 4.2]
- CMB-S4 will deploy a number of resources (2 x 6m telescopes, 1 x 5m + 18 x 0.5m telescopes) starting in 2030. These resources have the potential to produce 8 PB/year of data, sent to the United States nightly in 20 TB sets. Local computation and storage resources (670TB) can support storing a buffer of data for up to a month due to network degradation or downtime. [Section 4.2]
- Along with the raw data files CMB-S4 will create, the primary data products will be the frequency maps: [Section 4.2]
    - Daily (7 years x 365 days): 150 TB in 51,000 files.
    - MC (1,000 realizations x 7 years): 450 TB in 150,000 files.
- CMB-S4 will conduct a series of data challenges, starting in 2025 and running yearly through 2029, to help prepare for the operational challenge of data mobility from Chile to the US, and the analysis pipeline being constructed using DOE- and NSF-funded computing resources. Outcomes include: [Section 4.2]
    - Validate and verify the experiment design at the degree needed at each DOE/NSF review gate.
    - Validate and verify the project data management and collaboration data analysis pipelines.
    - Demonstrate scaling the data volume the simulation/reduction tools can process to achieve 100% by 2030.
    - Demonstrate porting and sufficient optimization of the data management software stack to each new generation of HPC and HTC architecture deployed by the DOE/NSF.
- Current and near-term work in data movement will focus on building a test stand for both data and metadata into the baseline data-movement tool (Rucio). These tests will primarily involve local operations within NERSC: running Rucio as a service on the Spin container-based service and ingesting simulated data into CFS. CMB-S4 will also practice moving data between local filesystems using Rucio. By the end of this time frame, transfers between NERSC and ALCF, or another large computing center, will be tested to validate performance expectations. [Section 4.3]

- Each day's data are immediately reduced to a map of the observed sky, using US agency computing resources (DOE HPC, NSF HTC, and potentially FABRIC). These maps are analyzed to identify transient sources, with alerts issued to the entire scientific community via the Community Alert Broker, and for data quality and telescope health checks. [Section 4.3]
- Given the size and complexity of the CMB-S4 data set (~100 PB), collecting, transporting, and reducing the raw data from timestreams of detector samples to well-characterized sky maps is the responsibility of the CMB-S4 project. Given the extraordinary range of science these maps support and their relatively small size (~100 TB), analyzing these maps is the responsibility of the CMB-S4 collaboration and the scientific community at large. [Section 4.3]
- CMB-S4 raw data will be archived at NERSC and ALCF, and a subset of recent data will be stored on disk at those centers. Daily data-quality statistics and maps will be produced at NERSC, ALCF, NSF ACCESS, and possibly FABRIC. Bulk reprocessing will be performed at NERSC, ALCF, and NSF ACCESS. Simulated data will be generated at NERSC and ALCF and typically reduced to maps on the fly to avoid I/O and storage overheads. In both cases, the resulting maps will be stored on disk and archived at NERSC and ALCF. Map-level data will be analyzed at NERSC and on local compute resources across the collaboration. Published data (primarily maps and derived scientific data) will be made available to the scientific community at NERSC and LAMBDA located at the NASA GSFC. The raw data and the software used to characterize and reduce them will also be made available to the scientific community at NERSC. [Section 4.6]
- During the construction phase, all the significant data will be simulated. Simulations will be run at NERSC or ALCF, and the data copied between them, peaking at O(10) PB in 2029. Reductions to maps will be run at NERSC, ALCF, NSF ACCESS centers, and possibly FABRIC. Maps will be made available to the collaboration at NERSC, with the bulk of the analyses performed there. Small amounts of data may be moved to collaborators' local resources for analysis. [Section 4.6]
- CMB-S4 will build upon existing infrastructure established by the Simons Observatory pathfinder experiment, which produces approximately 10% of the expected CMB-S4 data volume. A 10 Gbps dedicated fiber link from the site, to the ALMA, is being installed by Simons Observatory currently and already receives support from ESnet to ensure global connectivity. [Section 4.7]
- The fiber connection from the CMB-S4 experiment site, to the broader network, consists of a fiber pair from the Simons Observatory to the Red Universitaria Nacional (REUNA) PoP at the ALMA observatory. The fiber pair has an eight-channel passive optical multiplexer installed at each end, providing eight optical channels. Each of those channels can support 1 Gbps or 10 Gbps Ethernet with off-the-shelf pluggable optics. The Simons Observatory is using one channel today. CMB-S4 intends to use two of those channels for production operations. [Section 4.7]
- CMB-S4 has no plans to purchase dedicated AI/ML hardware or services. There are possible applications of ML in areas of low-level data cleaning (detection of classes of "glitche" in the data), and for improvements in observing strategies / planning. In both cases the evaluation of such models is cheap and model training can be performed at existing centers such as NERSC. [Section 4.8]
- The CMB-S4 project has no intention of using WiFi / Cellular / Satellite connectivity at the site in Chile. Some of this equipment can generate RFI that impacts science measurements. CMB-S4 may make use of an existing dedicated point-to-point microwave link to the nearby town. This link has been shown not to interfere with existing experiments and is a useful back-up connection for remote access / debugging if the nominal fiber connection goes down. [Section 4.8]

- By design, the CMB-S4 project plans to leverage multiple DOE SC user facilities and is a natural fit for consideration as an IRI workflow. Data will be replicated and analyzed across multiple DOE HPC facilities, with real-time requirements on results to help guide experimental outcomes. The results are broadly interesting to a wide community of users across the HEP community, and must be available decades into the future. [Section 4.9]
- CMB-S4 will leverage aspects of IRI including federated login and access across facilities, as well as tools for job submission and management that can span domains of operation. Ideally, these integrated tools would allow job management not only on HPC centers, but also on HTC or Grid-computing resources. HTC resources are ideally suited for rapid data-quality testing. [Section 4.9]

6.5.2 CMB: Experiment Profile

CMB-S4 was case-studied in 2020 with South Pole and Chile Atacama observatory sites. In the time after this initial case study, the experiment design is in the process of being revisited to site the observatory wholly in Chile. Since the process is incomplete, a conservative approach to the design is being taken, where all the 2020 assets will be deployed in Chile for the sake of this report. Revised data rates and requirements are anticipated by the midterm review.

6.5.2.1 Science Background

CMB-S4 is a joint DOE and NSF project, recommended by the Particle-Physics Project Prioritization Panel (P5)[70] in 2014 and 2023 and the Astronomy and Astrophysics Advisory Committee[71] in 2017. It adds the capacities and capabilities of ANL, Fermilab, LBNL, and SLAC under the DOE Office of HEP to the longstanding NSF university program spanning the Division of Astronomical Sciences (MPS/AST) and the Division of Physics (MPS/PHY) in the Directorate of Mathematical and Physical Sciences and the Office of Polar Programs in the Directorate of Geosciences. Other partners include NASA's Jet Propulsion Laboratory and GSFC, the Department of Commerce's National Institute of Standards and Technology, Associated Universities Incorporated, and the Smithsonian Astrophysical Observatory. Collaborative discussions are underway with the Cerro Chajnantor Atacama Telescope (CCAT-prime) and the Simons Observatory and its supporting Simons and Heising-Simons foundations. CMB-S4 also anticipates international partners and in-kind contributions, especially from Australia, Canada, Europe, and Japan.

Here and throughout the document, CMB-S4 is discussing a temporary project plan that will change as the mission design is optimized for the Chilean observing site. Note that while CMB-S4 anticipate their availability, none of the off-project network and computing resources that are planned for use have yet been secured as writing of this report (November, 2024)

To meet the CMB-S4 science goals (and all the other science that this will be enable) three coordinated surveys will be conducted from 2028 to 2035 at the high altitude Chilean site in the Atacama desert:

1. A wide/deep survey of ~70% of the sky with O(270,000) detectors spanning sixmm-wave frequencies on 2 x 6m telescopes with a daily cadence.
2. An ultra-deep survey of ~3% of the sky with O(150,000) detectors spanning 8 mm-wave frequencies on 18 x 0.5m telescopes.
3. An ultra-deep survey of the same ~3% of the sky with O(120,000) detectors spanning 7 mm-wave frequencies on 1 x 5m telescope.

70 https://www.usparticlephysics.org
71 https://www.nsf.gov/events/event_summ.jsp?cntn_id=299907&org=NSF

In operations, CMB-S4 will gather 1.9 x 10 ^ 8 samples per second, corresponding to a compressed data rate of 2.0 Gbps. Data will be transferred in real time over fiber networks to the primary data center at the DOE's NERSC and copied to the secondary data center at the DOE's ALCF. All data are archived at both NERSC and ALCF.

Each day's data are immediately reduced to a map of the observed sky, using US agency computing resources (DOE HPC, NSF HTC, and potentially FABRIC[72]). These maps are analyzed to identify transient sources, with alerts issued to the entire scientific community via the Community Alert Broker, and for data quality and telescope health checks.

The bulk data are analyzed on US agency resources to:

1. Characterize the instruments (e.g., determine their exact pointing, measure their beams and band-passes, etc.).
2. Identify systematic effects in and develop software mitigations sufficient for their residuals to be less than the tiny CMB signals.
3. Reduce them to maps at each observing frequency for a multitude of different data cuts over time and/or detectors, together with a statistical characterization of each map (e.g., a noise correlation matrix or a MC suite of simulated raw data sets reduced to equivalent sets of maps).

These well-characterized frequency maps are then passed to the various collaboration analysis working groups for the full range of scientific analyses.

Given the size and complexity of the CMB-S4 data set (~100 PB), collecting, transporting, and reducing the raw data from timestreams of detector samples to well-characterized sky maps is the responsibility of the CMB-S4 project. Given the extraordinary range of science these maps support and their relatively small size (~100 TB), analyzing these maps is the responsibility of the CMB-S4 collaboration and the scientific community at large.

6.5.2.2 Collaborators

The CMB-S4 collaboration currently consists of 520 members at 142 institutions in 24 countries and 26 US states. The institutions include DOE, NASA, and Department of Commerce laboratories, as well as a host of universities.

Currently all data are simulated, and the primary site for creation, storage, sharing, and analysis of these is NERSC, although this will expand to ALCF and NSF ACCESS as the construction project proceeds.

CMB-S4 raw data will be archived at NERSC and ALCF, and a subset of recent data will be stored on disk at those centers. Daily data-quality statistics and maps will be produced at NERSC, ALCF, NSF ACCESS and possibly FABRIC. Bulk reprocessing will be performed at NERSC, ALCF, and NSF ACCESS. Simulated data will be generated at NERSC and ALCF and typically reduced to maps on the fly to avoid I/O and storage overheads. In both cases, the resulting maps will be stored on disk and archived at NERSC and ALCF. Map-level data will be analyzed at NERSC and on local compute resources across the collaboration. Published data (primarily maps and derived scientific data) will be made available to the scientific community at NERSC and LAMBDA[73] located at the NASA GSFC[74]. The raw data and the software used to characterize and reduce them will also be made available to the scientific community at NERSC.

72 https://fabric-testbed.net
73 https://lambda.gsfc.nasa.gov
74 https://www.nasa.gov/goddard

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
TELESCOPE SITE: CHILEAN ATACAMA	Primary	Data transfer	20 TB	Daily (continuous)	No	No
PRIMARY DATA CENTER: NERSC AT LBNL	Primary	Data portal, data transfer	20 TB (daily), 8 PB (annually)	Daily and annually	No	No
SECONDARY DATA CENTER: ALCF AT ANL	Secondary	Data transfer	20 TB (daily), 8 PB (annually)	Daily and annually	Data transfer	No
GRID COMPUTING: NSF ACCESS (DISTRIBUTED)	Secondary (transient)	Data transfer	20 TB (daily), 8 PB (annually)	Daily and annually	Data transfer	To be evaluated
ARCHIVE: LAMBDA AT GSFC	Tertiary	Data transfer	600 TB	Single use	No	To be evaluated
COLLABORATION MEMBERS (SEE ABOVE)	Tertiary	Data portal, data transfer	< 600 TB	Intermittent	Data transfer	To be evaluated

Table 6.5.1: CMB-S4 Data Mobility

6.5.2.3 IRI Readiness

During operations, the CMB-S4 project plans to leverage multiple DOE SC facilities in the following ways:

1. A copy of all raw data will be replicated across at least two computing facilities (e.g., NERSC and ALCF).
2. Near real-time QA tests will be run on raw data. To avoid interruptions due to facility maintenance, it must be possible to run these tests on any HPC or HTC compute resources available to the project. (time-sensitive pattern)
3. Prior to data releases, a large volume of archival data may be reprocessed. This requires substantial compute resources and jobs may be distributed across multiple HPC centers. (long-term campaign pattern)

For points (2) and (3) above, CMB-S4 will leverage aspects of IRI including federated login / access across facilities and also tools for job submission and management. Ideally, these integrated tools would allow job management not only on HPC centers (i.e., running slurm), but also on HTC / Grid-computing resources. HTC resources are ideally suited for rapid data-quality testing.

6.5.2.4 Instruments and Facilities

6.5.2.4.1 Observatory in the Chilean Atacama

Present / Next Two to Five Years: Simons Observatory pathfinder experiment with ~10% of the CMB-S4 data volume.

10 Gbps dedicated fiber link from the site to the ALMA being installed by Simons Observatory now, with ESnet support.

Beyond: 2 x 6m telescopes, 1 x 5m + 18 x 0.5m telescopes to be deployed starting in 2030:

- Gathering 8 PB/year.
- File size/count/distribution TBD.

On-site computing resources to be deployed in 2030 sufficient to:

- Store one month of data (670 TB)

6.5.2.4.2 Computing Resources

Present / Next Two to Five Years: Sufficient cycles and storage for experiment design studies, data management subsystem development, data analysis pipeline validation, and verification.

NERSC disk and data portal for data distribution to the collaboration (Tera-Scale).

The exact distribution of cycle/storage resources between these systems will depend on the agencies and their allocations. The overall driver in the construction project will be a series of four data challenges intended to:

- Validate and verify the experiment design at the degree needed at each DOE/NSF review gate.
- Validate and verify the project data management and collaboration data analysis pipelines.
- Demonstrate scaling the data volume the simulation/reduction tools can process to achieve 100% by 2030.
- Demonstrate porting and sufficient optimization of the data management software stack to each new generation of HPC and HTC architecture deployed by the DOE/NSF.

An estimate of the resources required for each challenge and the target HPC system (to be augmented by HTC resources) is in the following table.

Challenge	DC1	DC2	DC3	DC4
YEAR	2025	2026	2027	2029
DATA FRACTION	12.50%	25%	50%	100%
TOTAL CYCLES (EFLOP)	3	6	12	24
PEAK MEMORY (PB)	2.5	4.9	9.9	19.8
PEAK SCRATCH (PB)	1.3	2.7	5.4	10.7
SYSTEM	PERLMUTTER	AURORA	NERSC-10	NERSC-10
PEAK CYCLES (PFLOP/S)	150	1000	1000	1000
TOTAL MEMORY (PB)	3	5	5	5
TOTAL SCRATCH (PB)	150	200	200	200
FULL-SYSTEM RUNTIME (HRS)	11.12	3.33	6.67	6.67
PEAK MEMORY FRACTION	165%	198%	395%	395%
PEAK SCRATCH FRACTION	2%	3%	5%	5%

Table 6.5.1: CMB-S4 computational challenge summary

Note that CMB-S4 exceeds the peak memory available in all cases, limiting us to annual maps (1/7 of the requirement) rather than the full data maps at this stage.

Beyond: Sufficient cycles and storage for operations data management and data analysis. NERSC disk and data portal for data distribution to the collaboration (Tera-Scale). In operations, CMB-S4 anticipates a 10–1000x increase in the integrated computing requirements over DC4, progressively growing between 2030 and 2040, but

no increase in the peak requirements. These requirements — which will include a mixture of very many small analyses and a small number of very large analyses — will be distributed across all the available systems, although CMB-S4 anticipate that the small analyses will primarily be run on HTC, and the large on HPC, systems.

Along with the raw data files listed under the sites, the primary data products will be the frequency maps:

- Daily (7 years x 365 days): 150 TB in 51,000 files.
- MC (1,000 realizations x 7 years): 450 TB in 150,000 files.

The full CMB-S4 computational requirements can be found online, and were last reviewed in September 2020.

6.5.2.5 Generalized Process of Science

Present / Next Two to Five Years: During the construction phase, all the significant data will be simulated (although small amounts of laboratory data will be generated, primarily at the national laboratories, some of which may be transferred to NERSC).

Simulations of raw data will be run at NERSC or ALCF and the data copied between them, peaking at O(10) PB in 2029. Reductions to maps will be run at NERSC, ALCF, NSF ACCESS, and possibly FABRIC. Maps will be made available to the collaboration at NERSC, with the bulk of the analyses performed there. Small amounts of data may be moved to collaborators' local resources for analysis.

Beyond: During the operations phase, data will be transferred to NERSC and ALCF over the network, with the exact mechanics (including failsafe processes for system downtimes) still to be determined.

Daily reductions to maps will be run at NERSC, ALCF, NSF ACCESS, and possibly FABRIC. Bulk-data reductions to maps will be run at NERSC, ALCF, and NSF ACCESS.

MC simulations will be used for uncertainty quantification and debiasing of the instrument data. Such simulations will be generated at NERSC and ALCF and reduced to maps on the fly. The maps will be archived at both sites and made available to the collaboration at NERSC.

6.5.2.6 Remote Science Activities

The observing site (Chilean Atacama) is remote.

Present / Next Two to Five Years: Precursor experiment (Simons Observatory in Chile) will exercise many of the systems CMB-S4 plans to use, albeit at ~10% of the data volume.

Beyond: In operations, the Chilean telescopes will generate 2.0 Gbps, and will be connected with a robust 10 Gbps network to the United States.

6.5.2.7 Software Infrastructure

6.5.2.7.1 Data Movement

Present / Next Two to Five Years: Current and near-term work in data movement will focus on building a test stand for ingesting data / metadata into the baseline data movement tool (Rucio). These tests will primarily involve local operations within NERSC: running Rucio as a service on the Spin container-based service and ingesting simulated data into CFS. CMB-S4 will also practice moving data between local filesystems using Rucio. By the end of this time frame, transfers between NERSC and ALCF, or another large computing center, will be tested to validate performance expectations.

Separate from data-movement tests internal to the project, CMB-S4 will also increase the number of datasets (both simulations and legacy public datasets) that are hosted through the portal hosted at NERSC. This portal supports data downloads using globus.

Beyond Five Years: Prior to operations, a review of the Rucio baseline selection will occur. Even if a new baseline software tool is chosen, the same basic transfer patterns would persist. During operations, continuous transfer of data from Chile to NERSC would occur using Rucio over globus. Data would then be replicated with the same tools to ALCF or another computing facility.

6.5.2.7.2 Data Simulation and Reduction

CMB-S4 currently uses (and plan to continue using) two software frameworks for simulation and analysis of experimental data. The Time Ordered Astrophysics Scalable Tools (TOAST) framework[75] is focused on use in an HPC context. The SPT-3G framework[76] is focused on HTC computing environments. Both are open source and interoperability work is ongoing.

6.5.2.8 Network and Data Architecture

In the context of the CMB-S4 project, the "local" experiment / facility network describes the infrastructure in place at the site in Chile and the interfaces from this location to the broader internet. The raw experimental data consists of compressed "data frames" that are packed into binary files (with a tunable size) prior to transmission to North America. These binary files have historically been around a Gigabyte in size, which strikes a balance between file size and number of files.

Present / Next Two to Five Years: During the construction phase of the project, minimal CMB-S4 data will be transferred from the experiment site back to North America. One of the proposed sites in Chile (Cerro Toco) has established infrastructure and a currently operating experiment (the Simons Observatory). CMB-S4 will leverage that existing connection to run some tests with perfSONAR and also transfer some small simulated data files. Connectivity to the broader Internet is via REUNA, the Chilean NREN. CMB-S4 has an excellent collaborative relationship with REUNA.

Beyond Five Years: The fiber connection from the experiment site to the broader Internet consists of a fiber pair from the Simons Observatory site to the REUNA PoP at the ALMA observatory. The fiber pair has an eight-channel passive optical multiplexer installed at each end, providing 8 optical channels. Each of those channels is capable of supporting 1 Gbps or 10 Gbps Ethernet with off-the-shelf pluggable optics. The Simons Observatory is using one channel today. CMB-S4 intends to use two of those channels for production operations.

Although the exact telescope configuration is currently being re-baselined, if a maximal set of 27 small aperture telescopes and 4 large aperture telescopes is assumed, the sustained data rate is approximately 2 Gbps. CMB-S4 plans to have a 5x margin to allow for "catch-up" transfers in the event of an outage. Although this could fit into a single optical channel on the multiplexer, there is a plan to use two of the eight channels to allow for service activities on one channel at a time if needed (and/or reduce catch-up time).

75 https://github.com/hpc4cmb/toast
76 https://github.com/CMB-S4/spt3g_software

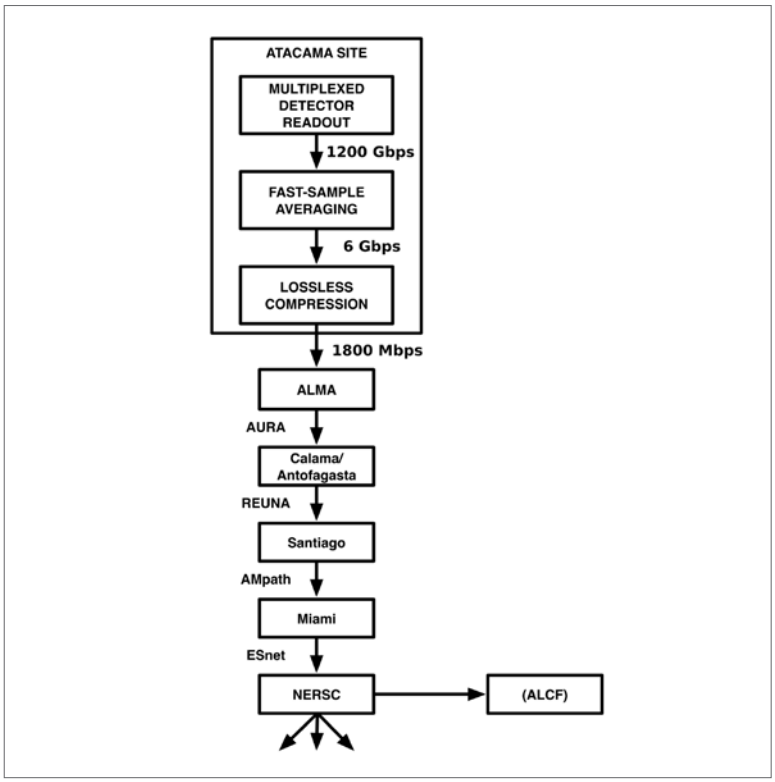


Figure 6.5.1: CMB-S4 Network Diagram

6.5.2.9 Emerging Technology Use

6.5.2.9.1 Cloud Services

Although CMB-S4 has no plans to pursue traditional cloud compute resources, cloud-based services for software development and deployment (github / github actions, conda-forge, docker hub, etc.) will be utilized. CMB-S4 has also explored the use of the FABRIC nearline testbed (described below).

Present / Next Two to Five Years: CMB-S4 has explored the use of FABRIC for several time-critical tasks. These nodes have reasonable core counts and memory for tasks such as detecting transient astrophysical sources in small maps made of the raw data. The appeal of using a platform like this is the immediate access to a small amount of compute cycles and the ability to stream data to these nodes in near real time.

Beyond: Although the baseline plan for real-time computing needs is a combination of HPC centers and HTC resources, CMB-S4 will explore other possibilities for the redesigned, all-Chile configuration. This might include limited computing resources on-site, novel solutions like FABRIC, or small dedicated compute nodes in North America. Commercial cloud solutions are not likely to play a future role, simply due to the cost of transferring / streaming substantial data volumes to those providers.

6.5.2.9.2 AI and ML

CMB-S4 has no plans to purchase dedicated AI/ML hardware or services. There are possible applications of ML in areas of low-level data cleaning (detection of classes of "glitches" in the data), and for improvements in observing strategies / planning. In both cases the evaluation of such models is cheap and model training can be performed at existing centers such as NERSC.

6.5.2.9.3 Wireless or Distributed Location Science Activities

The CMB-S4 project has no intention of using WiFi / Cellular / Satellite connectivity at the site in Chile. Some of this equipment can generate RFI that impacts science measurements. CMB-S4 may make use of an existing dedicated point-to-point microwave link to the nearby town. This link has been shown not to interfere with existing experiments and is a useful back-up connection for remote access / debugging if the nominal fiber connection goes down.

6.5.2.10 Data-Related Resource Constraints

The biggest overall data-related challenge is the capacity and capability of HPC and HTC systems, especially in their total system memory and their available cycles and the efficiency with which CMB-S4 can use them (especially in the post Moore's law epoch of energy-constrained computing).

6.5.2.11 Data Mobility Endpoints

CMB-S4 maintains a data portal at data.cmb-s4.org which has the capability of delivering data both to the CMB-S4 collaboration and to the wider CMB community through Globus.

The portal is just a statically generated website that contains metadata and links to documentation for data, mostly in Flexible Image Transport System (FITS) files, hosted on Globus. Data is accessible either through HTTPS, convenient for quickly downloading individual files to personal machines, or through the Globus application itself, suitable for batch downloading of data to Supercomputers.

Data is hosted on the NERSC Community File System accessible from all NERSC Supercomputers in specific folders which are made available on Globus through Globus Guest Collections. The CMB-S4 team is admin of the Guest Collections but the entire infrastructure is maintained by NERSC.

The NERSC Share Endpoint has id 5dbafbb3-578b-4fb5-aa2f-2b60ad18072d and the two Guest Collections CMB-S4 maintains are for CMB-S4 specific data, and for public datasets.

CMB-S4 is currently hosting about 30 TB of data mostly as 2-GB-sized uncompressed FITS files.

6.5.2.12 Outstanding Issues

There are no outstanding issues to report at this time beyond continuing to work with ESnet on wide-area/international bandwidth strategies that will become a factor in the future.

6.5.2.13 Facility Profile Contributors

CMB-S4 Representation

Reijo Keskitalo, LBNL and University of California Berkeley, rtkeskitalo@lbl.gov

Ted Kisner, LBNL and University of California Berkeley, tskisner@lbl.gov

Andrea Zonca, SDSC and UCSD, zonca@sdsc.edu

Julian Borrill, LBNL and University of California, Berkeley, jdborrill@lbl.gov

John Carlstrom, University of Chicago and ANL, jc@kicp.uchicago.edu

Jim Yeck, University of Wisconsin, Madison

John Corlett, LBNL, jncorlett@lbl.gov

Gil Gilchriese, LBNL, mggilchriese@lbl.gov

ESCC Representation

Rune Stromsness, LBNL, rstrom@lbl.gov

Richard Simon, LBNL, rsimon@lbl.gov

Tavia Stone Gibbins, LBNL and NERSC, tavia@lbl.gov

6.6 Muon Experimentation at Fermilab: Muon g-2 and Mu2e

Muons are a potential window to new physics. Extreme precision measurements and searches for extraordinarily rare phenomena offer comparisons to the Standard Model (SM) that may reveal discrepancies that will narrow down the models beyond the Standard Model (BSM) that are viable. In contrast to many LHC studies that search for new BSM particles, these studies are indirect searches for new physics and require very intense beams.

6.6.1 Discussion Summary

The muon g-2 experiment will complete its studies with nearly 7 PB of raw experimental data, and an additional 1.5 PB of processed results. All data are planned to remain on tape at Fermilab, and at current time there is no backup planned or required. Once the aforementioned processing is complete, the collaboration does not anticipate reprocessing will happen at this time.

Mu2e will utilize most of the existing computational and storage resources at Fermilab during production, but has used a number of DOE HPC facilities (i.e., ALCF and NERSC in particular) during simulation construction. Currently the project produces on the order of hundreds of TB of simulation output per year using OSG-connected resources at Fermilab, the HPC facilities, and other collaboration sites.

The first aspect of operational Mu2e science will be to handle cosmic-ray events, which are expected to produce up to 500 TB of event data per year. The majority of this effort will be computed through Fermilab, but there is possibility that some computation will occur at other WLCG sites; a project goal is to consider up to half of computing needs occurring off-site. Regular operational data from Mu2e will top 7PB or raw data a year. All data will be stored at Fermilab, with back-up copies stored at partner sites in Europe (locations are still being determined).

Mu2e networking needs are relatively straightforward; a maximum of 10 Gbps may be required to handle intensive times of activity to support external computing needs. Given that a majority of computation and storage is allocated to Fermilab, and connectivity to other WLCG sites will follow the same patterns as the LHC computation, additional networking support is not required at this time.

Mu2e will be adopting Rucio and other tools that are well-known to the HEP community to manage workflow and data mobility. The project is limited in exploration of alternative software and computational platforms namely due to funding and staff. Adopting new workflows is not something the collaboration has the time or resources do explore currently, but they have been able to make their simulation and analysis code work at locations like NERSC and ALCF. During long shutdowns, or extended technical stops, it may be possible to explore other use cases.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The Muon g-2 experiment at Fermilab collected data from 2018 until summer of 2023 and has two major releases of results: the first in 2021 based on Run 1 data and the second in 2023 based on Runs 2+3 data. The collaboration is analyzing the remaining data (Runs 4–6) and expects to make a final measurement during FY25. [Section 4.1]
- The Muon to Electron Conversion Experiment at Fermilab (Mu2e) will search for the neutrinoless decay of a muon into an electron in the Coulomb field of an atomic nucleus. Mu2e will start taking cosmic-ray commissioning data early in mid CY 2025 and will begin commissioning with beam operations late in calendar 2026. The first physics run will begin early in CY 2027 and will continue to the end of CY 2027. During 2028 and 2029 Fermilab accelerator operations will pause for upgrades to the accelerator complex in preparation for LBNF. Mu2e will resume running in 2030 and continue until the 2033 summer shutdown. [Section 4.1]

- The Muon g-2 collaboration comprises 181 collaborators from 33 institutions located in 7 countries: China, Italy, Russia, South Korea, Germany, United Kingdom, and the United States. [Section 4.1]
- The Mu2e collaboration features a total of 239 scientists from 39 institutions in 6 countries: China, Germany, Italy, Russia, the UK and the US. [Section 4.1]
- The full Muon g-2 analysis data set, is approximately 1.3 PB. These data are processed at Fermilab to generate much smaller summary files (e.g., histograms in files on order of GB) that are transferred to collaborators for analysis at their institutions. [Section 4.2]
- Mu2e expects that many institutions will request copies of analysis-format datasets and that these will be stored locally at each institution. At this time, it is estimated that one copy of the analysis-format dataset for the full experiment will be, at most, on the scale of 50 TB. [Section 4.2]
- At design rate the Mu2e experiment will produce about 7 PB/year of raw data, including physics data, calibration data and special runs. The current projection is that Mu2e will achieve the physics goals of the experiment taking about 4 PB/year of raw data. For most of these data two copies will be stored on tape, but a subset will have only one copy. All tape resident copies will be stored at Fermilab. The details of how many events will be selected for calibration, how many events will be written out by reconstruction and the size per event of these datasets are not well understood at this time. The current estimate is that an additional 5 PB/year will be needed for intermediate data sets, repeat processing and for storage of simulated events. This brings the total tape needs to about 13 PB/year. [Section 4.2]
- All Mu2e data can be processed on Fermigrid, but for planning purposes it is assumed that about half of it will be processed on Fermigrid and half on opportunistic OSG. On average Mu2e expects to process each event twice, leading to an estimate of 4 TB/year of outbound data and 2.5 TB/year of inbound data. Discussions have taken place that may lead to some calibration work being done in the UK or at INFN, using resources provided by their funding agencies. For this exercise Mu2e assumes that about 0.5 PB/year will be transferred to each location. [Section 4.3]
- As Mu2e will be under construction in CY 2025 and 2026, the project will be commissioned with cosmic rays. During this period, MC data volumes are likely to remain constant and experimental data volumes will be small, likely much less than 0.5 PB/year. On the time scale of three to five years Mu2e will commission with beam and have a one-year run. This will be followed by the two-year-long shutdown. During the run there will be nominal data volumes. During the two-year-long shutdown reprocessing of the full dataset with the latest calibrations and algorithms is expected. On a timescale of more than five years Mu2e will restart data taking following the long shutdown, complete data taking and begin final analysis. A significant increase in network needs compared to the three- to five-year forecast is not expected. [Section 4.3]
- Simulations for Mu2e will be run on Fermigrid, OSG, at NERSC and ALCF. The simulated events, and reconstructed simulated events will be stored at Fermilab. Mu2e expects that many institutions will request copies of the analysis-format datasets derived from the simulated events; these datasets are likely to be about a few times the volume of the corresponding experimental datasets. [Section 4.3]
- The primary data store for raw Mu2e data and intermediate datasets will be in the computing centers on the Fermilab site. Most of the calibration jobs, and all the reconstruction jobs, will be run on Fermigrid or use opportunistic cycles on OSG. Mu2e currently has no collaborators with access to dedicated cycles on OSG. [Section 4.3]

- The current computing needs of Mu2e are dominated by simulation campaigns to provide datasets of simulated events for the purpose of value engineering the design, validating the design, and providing input for the development of trigger, reconstruction, calibration and analysis algorithms. At first the simulations were done on OSG, including Fermilab, but in recent years resources on Theta at ALCF and on Cori and Perlmutter at NERSC have been used. The output of each of these simulation campaigns is on the scale of 250 TB. [Section 4.4]
- Mu2e has had preliminary discussions with groups in Italy and in the UK to provide computing resources for the calibration of subsystems. If these plans are realized, outbound datasets will be on the scale of 0.5 PB/year to each location, and inbound datasets would accumulate to tens of TB/year. If additional computing resources are available in collaborating institutions, Mu2e plans to use them for simulation campaigns; in this case there will be little outbound traffic and inbound traffic of up to a few hundred TB/year. [Section 4.4]
- Mu2e expects to use resources at several DOE supported sites. At Fermilab Mu2e will use local interactive machines provided by the lab, Fermigrid, the Fermilab Elastic Analysis Facility, and the data storage resources at Fermilab. Mu2e has demonstrated the ability to use big CPU cycles at HPC centers, and does not foresee a need for significant GPU resources at this time. [Section 4.6]
- Mu2e has one example of the time-sensitive IRI pattern. Mu2e plans to perform initial processing of data from the experiment within several hours of the data being taken. This will be used to select events for several different calibration streams and to produce Data-Quality Monitoring (DQM) information. The time constraint is driven by the need to produce high-precision, high-statistics DQM information that is sensitive to subtle effects that are below the sensitivity of online DQM. [Section 4.9]

6.6.2 Muon Experimentation at Fermilab: Experiment Profile

6.6.2.1 Science Background

Measuring the anomalous magnetic moment of the muon (g-2) has been an enterprise for the last seven decades, with each experiment achieving better precision and probing the SM more deeply.

The Muon to Electron Conversion Experiment at Fermilab (Mu2e) will search for the neutrinoless decay of a muon into an electron in the Coulomb field of an atomic nucleus. Observation of this process would provide unambiguous evidence for physics BSM. The projected single event sensitivity of Mu2e is 2.87 x 10-17, approximately four orders of magnitude more sensitive than the previous best experiment, SINDRUM-II. For the quoted sensitivity the projected background level is about 0.4+/-0.2 events. More information[77] may be found in the Mu2e Technical Design Report (TDR)[78].

77 https://mu2e.fnal.gov
78 https://arxiv.org/abs/1501.05241

6.6.2.1.1 Muon g-2 Science Background

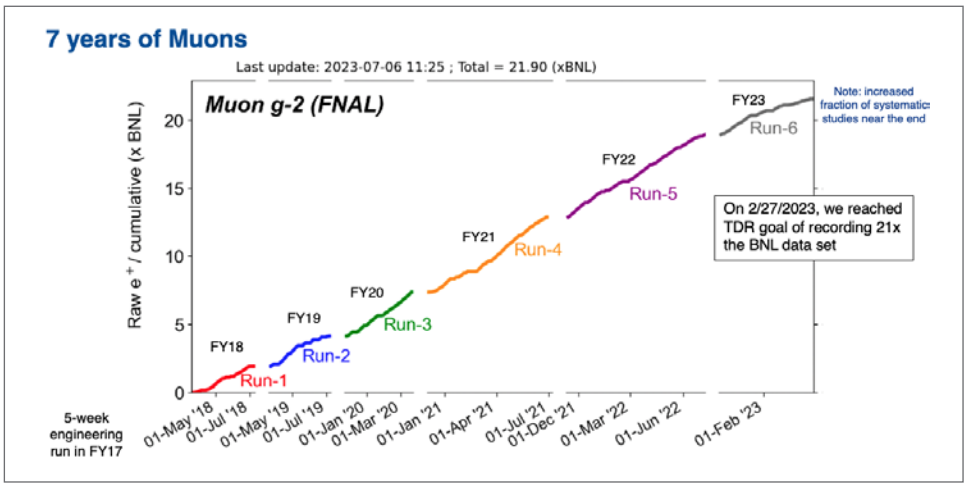


Figure 6.6.1: Data taking for the Muon g-2 experiment at Fermilab

The Muon g-2 experiment at Fermilab collected data, as shown in Fig. 6.6.1, from 2018 until summer of 2023 and has two major releases of results: the first in 2021 based on Run 1[79] data and the second in 2023 based on Runs 2+3 data.[80] The Runs 2+3 analysis achieved a systematic uncertainty of 70 ppb, exceeding the design goal of the entire experiment. Fig. 6.6.2 shows the latest results, including the Fermilab 2023 release and the world average, the latter dominated by the Fermilab result. The discrepancy between the experiment measurement and the SM prediction[81] exceeds five-sigma.

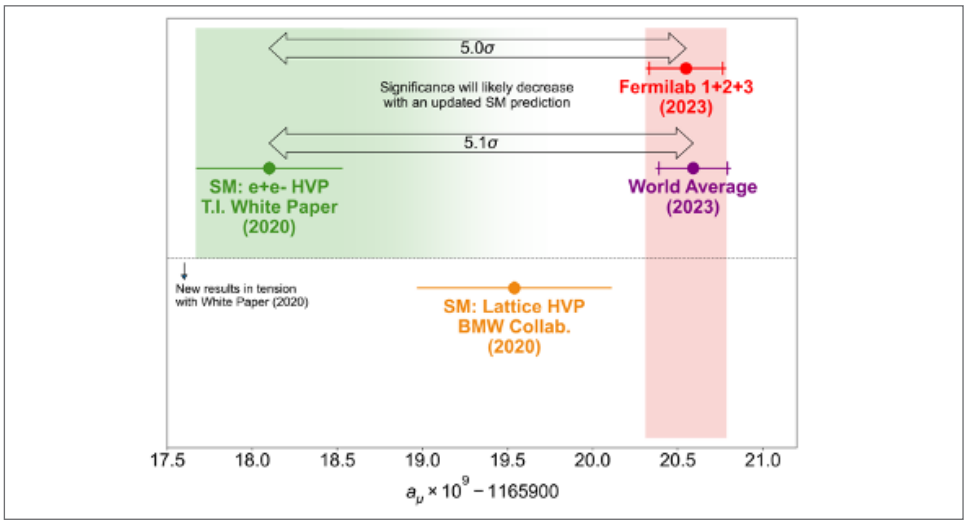


Figure 6.6.2: Latest results from Muon g-2 experiments

However, there are new predictions from the lattice quantum chromodynamics (QCD) community[82] that are much less discrepant compared to the measurements. The Muon g-2 Theory Initiative is in the process of evaluating these predictions and the possibility of updating their consensus.

79 B. Abi et al., (Muon g−2 Collaboration), Phys. Rev. Lett. 126, 141801 (2021)
80 D. P. Aguillard et al., (Muon g−2 Collaboration), Phys. Rev. Lett. 131, 161802 (2023)
81 T. Aoyama et al., Phys. Rep. 887, 1 (2020)
82 Sz. Borsanyi et al., Nature (London) 593, 51 (2021)

The collaboration is analyzing the remaining data (Runs 4–6) and expects to make a final measurement during FY25. Production efforts (large-scale processing) are nearly complete and will not be discussed here. Other results that require the full dataset, including the electric dipole moment of the muon and tests of Lorentz invariance, will likely follow shortly thereafter.

6.6.2.1.2 Mu2e Science Background

Mu2e raw data will flow from sensors in five subsystems, through the data acquisition system (DAQ) and trigger to storage on tape in the Fermilab central computing facility. The raw data will be processed into analysis-format datasets in a multistep process with a latency of a few weeks. The analysis-format data sets will be small and widely distributed to collaborating institutions; the intermediate datasets will be used by members of the collaboration to improve calibrations and algorithms. From time to time Mu2e will reprocess all data using the most up-to-date calibrations and algorithms; when this is done, intermediate datasets older than a few iterations will be retired[83].

Mu2e is still under construction. Mu2e will start taking cosmic-ray commissioning data early in mid CY 2025 and will begin commissioning with beam operations late in calendar 2026. The first physics run will begin early in CY 2027 and will continue to the end of CY 2027. During 2028 and 2029 Fermilab accelerator operations will pause for upgrades to the accelerator complex in preparation for LBNF. Mu2e will resume running in 2030 and continue until the 2033 summer shutdown. This called the Long Shutdown. Additional information is available in section 3, below.

In the past years, the computing needs of Mu2e were dominated by simulation campaigns to provide datasets of simulated events for the purpose of value engineering the design, validating the design, and providing input for the development of trigger, reconstruction, calibration and analysis algorithms. A typical scale is 25 million core hours once every few years plus 10 million hours per year for developing the code to produce these datasets and extracting results from these datasets. This will remain the computing driver until the start of beam data in late calendar 2026.

At first the simulations were done on OSG, including Fermilab, but in recent years resources on Theta at ALCF and on Cori and Perlmutter at NERSC have been used. The output of each of these simulation campaigns is on the scale of 250 TB.

The "story of the data" is similar for both Fermilab muon experiments (and, for that matter, most other Fermilab experiments) and is depicted in Fig. 6.6.3. The main components relevant to this case study are the Mass Storage and Compute systems and especially the network connecting them. These components are described in subsequent sections and elsewhere. The Fermilab network is described in Section 7.4.

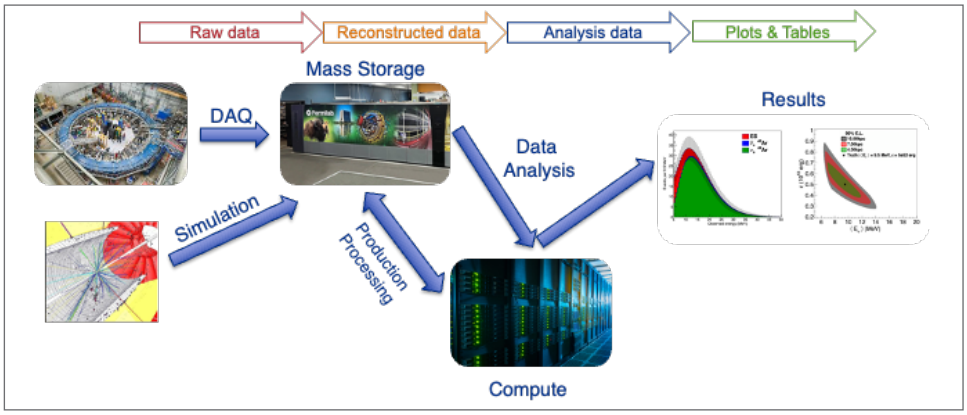


Figure 6.6.3: The flow of data from raw data to analysis results.

83 https://publicdocs.fnal.gov/cgi-bin/ShowDocument?docid=515

6.6.2.2 Collaborators

6.6.2.2.1 Muon g-2 Collaborators

The Muon g-2 collaboration comprises 181 collaborators from 33 institutions located in 7 countries: China, Italy, Russia (very limited access to data), South Korea, Germany, United Kingdom, United States. As mentioned, large-scale data processing is nearly complete. The full analysis data set, even with data reduced, is about 1.3 PB. These data are processed at Fermilab to generate much smaller summary files (e.g., histograms in files on order of GB) that are transferred to collaborators for analysis at their institutions. Given that the experiment is no longer taking data and data analysis is nearing the end, the collaboration is not expected to grow in size.

Muon g-2 has no contribution to the table below. The collaboration is no longer executing large data transfers, as currently only histogram files are needed for final analysis. These files are small (~GB scale).

6.6.2.2.2 Mu2e Collaborators

The Mu2e Collaboration list may be found here[84]. There are a total of 239 scientists from 39 institutions in six countries: China, Germany, Italy, Russia, the UK and the US.

The primary data store for raw data and intermediate datasets will be in the computing centers on the Fermilab site. Most of the calibration jobs, and all the reconstruction jobs, will be run on Fermigrid or use opportunistic cycles on OSG. Mu2e currently has no collaborators with access to dedicated cycles on OSG.

There have been preliminary discussions with groups at INFN in Italy and in the UK for them to provide computing resources at the level of a few million CPU hours per year that would be used for the calibration of subsystems for which they have the primary responsibility. Mu2e expects firm commitments from the Italian groups for CY 2025. Should these plans be realized, Mu2e estimate that outbound datasets will be on the scale of 0.5 PB/year to each location and that inbound datasets will be small, at most a few tens of TB/year; the inbound datasets will be calibration sets and DQM information. If additional computing resources are available in collaborating institutions Mu2e plans to use them for simulation campaigns. In this case there will be little outbound traffic and there will be inbound traffic of up to a few hundred TB/year.

Mu2e expects that many institutions will request copies of analysis-format datasets and that these will be stored locally at each institution. At this time, it is estimated that one copy of the analysis-format dataset for the full experiment will be, at most, on the scale of 50 TB; it may be considerably smaller.

Simulations will be run on Fermigrid, OSG, at NERSC and ALCF. The simulated events, and reconstructed simulated events will be stored at Fermilab. Mu2e expects that many institutions will request copies of the analysis-format datasets derived from the simulated events; these datasets are likely to be about a few times the volume of the corresponding experimental datasets.

Mu2e will run four classes of jobs:

- **Data Production:** this is a multi-stage process: select events for calibration, perform calibrations, main reconstruction pass, optional secondary reconstruction passes and creation of analysis-format data sets. From time to time the secondary reconstruction will be repeated and perhaps once every few years the full dataset will be reprocessed; in both cases the goal is to take advantage of improved calibrations and algorithms. Mu2e plans to run these jobs using dedicated Fermigrid resources and opportunistic OSG or HPC resources. The breakdown is not known at this time. For purposes of this study, it is assumed that 50% of the work is done on Fermigrid and that other 50% is done on OSG or at HPC centers.
- **End-user Analysis:** The input to this work is the analysis-format data sets that are the final product of data production. This work will likely be distributed across Fermigrid, opportunistic OSG and resources at remote institutions.

84 https://mu2e.fnal.gov/mu2e_collaboration_list.shtml

- **MC Production pre-data:** These jobs create simulated events and process them through the same reconstruction chain as for experimental data. These jobs have no data inputs, only configuration input. An unusual feature of Mu2e compared to other HEP experiments is that the output of these datasets is small per unit CPU. A simulation campaign of 10 million hours will produce somewhere between 50 TB and 250 TB, depending on what processes are simulated, and what level of intermediate results are retained. The reason is that simulation of the signal is trivial and does not require significant resources. What takes time is the simulation of pileup processes; Mu2e has exploited all the opportunities to sculpt these processes at the generator level but the reality is that many corners of generator phase space must be included that produce very rare, but still important backgrounds. Therefore, most simulated events fail the selection criteria and are not written to the output file. The dominant workload in these jobs is Geant4 and Mu2e has demonstrated the ability to use multithreaded Geant4 on "Knights Landing" architecture for Intel CPUs (KNL) based HPC resources. Mu2e has run these jobs on Fermigrid, OSG, on Theta and BeBop at ALCF, and on Cori and Perlmutter at NERSC. The analysis-format data sets generated from these simulations are on the scale of 2 TB per 10 million hours of computing.
- **MC Production with data:** once Mu2e has data, random triggers will be used as the pileup model. The project will overlay random trigger events on simulated signal and signal-like background tracks, and will also select signal-like tracks from off-spill running. Mu2e will take the hits from these tracks and overlay random triggers over them to study reconstruction efficiency and resolution tails in the presence of pileup. This will greatly reduce the need for simulation of pileup, which is the driver for the simulation CPU needs. Mu2e anticipates still needing to simulate pileup in order to understand features that are observed in the random trigger events; MC production CPU will remain the same during data taking was needed pre-data.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
OSG (ANALYSIS)	N	xRootD	~50 TB/y in ~0.5 TB/y out	On demand	Y xRootD	No
OSG (PRODUCTION)	N	xRootD	~4 PB/y in ~2.5 PB/y out	On demand	Y xRootD	No
NERSC (PROPOSED MC PRODUCTION)	Y	Rucio transfer to NERSC disk	0 input ~50 TB/y out	On demand	Rucio	No
ALCF (MC PRODUCTION)	N	Rucio	0 input ~50 TB/y out			
DEDICATED INFN (ITALY) & UK RESOURCES	Y Secondary	Rucio transfer to local disk	1 PB/y in 50 TB/y out	On demand	Rucio	No

Table 6.6.1: Mu2e Data Mobility

Projected Mu2e data-movement requirements. "In" refers to inbound to the named resource and "out" refers to transfers back to Fermilab. For production, network usage will distribute throughout the year with spikes to a few times nominal. The driver is OSG production and the second largest use is transfer to European collaborators.

6.6.2.3 IRI Readiness

6.6.2.3.1 Muon g-2 IRI Readiness

The Muon g-2 collaboration is currently performing the final analysis on local machines, the Fermilab Elastic Analysis Facility, and laptops and only involves simple workflows evaluating histograms. An IRI notebook based analysis facility may be an interesting resource, though the timing is not right for use.

6.6.2.3.2 Mu2e IRI Readiness

Mu2e expects to use resources at several DOE supported sites. At Fermilab Mu2e will use local interactive machines provided by the lab, Fermigrid, the Fermilab Elastic Analysis Facility and the data storage resources at Fermilab. Mu2e has demonstrated the ability to use big CPU cycles at HPC centers, and does not foresee a need for significant GPU resources at this time.

Mu2e has one example of the time-sensitive pattern. Mu2e plans to perform initial processing of data from the experiment within several hours of the data being taken. This will be used to select events for several different calibration streams and to produce DQM information. The time constraint is driven by the need to produce high-precision, high-statistics DQM information that is sensitive to subtle effects that are below the sensitivity of online DQM. This will inform the calibration work.

At this time Mu2e does not foresee examples of the data integration intensive pattern.

Regarding the long-term campaign pattern. Including commissioning and the planned Fermilab accelerator shutdown Mu2e expects to take data from 2025 to 2033. This will be followed by several more years of final analysis where continuous access to the resources discussed above will be required.

6.6.2.4 Instruments and Facilities

6.6.2.4.1 Muon g-2 Instruments and Facilities

The experiment ceased collecting data in July 2023[85], and does not plan any further simulation campaigns. Therefore, no data are expected to be generated except for small files associated with analyses. the final result from the experiment are expected in FY25, and some ancillary results to appear soon thereafter. By FY26 or FY27, all Muon g-2 computing activity will end.

6.6.2.4.2 Mu2e Instruments and Facilities

For full information about the Mu2e apparatus see the Mu2e TDR[86].

The parts of the apparatus that generate data are a straw tube tracking system, a caesium iodide calorimeter, a plastic scintillator based Cosmic Ray Veto System, a pixel telescope-based Beam Extinction Monitor System and a Stopping Target Monitor comprising two crystals, one high purity Germanium and the other LaBr. The data generated by these systems are processed by the Trigger and Data Acquisition System (TDAQ).

Other major subsystems include the proton delivery system, including an extinction insert, and the Muon Beamline comprising the three large superconducting magnets, the production target, a collimator system, the stopping target and the muon beam stop. One can also consider the detailed arrangement of shielding materials as a subsystem in its own right.

All these subsystems are located in the Muon Campus at Fermilab.

The operational cycle of Mu2e is 1.4s, divided into an on-spill period of approximately 0.4s and on off-spill period of approximately 1.0s. During the on-spill period a proton pulse hits the production target every 1695 ns and the apparatus is configured to record one event for every proton pulse. During the off-spill period the apparatus will

85 B. Abi et al., (Muon g−2 Collaboration), Phys. Rev. Lett. 126, 141801 (2021)
86 https://doi.org/10.48550/arXiv.1501.05241

be configured to record cosmic rays and to perform in-band calibration tasks; off-spill events will have a duration of 100 μs, which is a tradeoff between several competing requirements. Data from the five data-generating subsystems flows into a streaming, deadtime-less Data Acquisition system and from there into a software based Trigger system that will inspect every event. During on-spill data taking, the trigger will accept approximately 1 in 400 events. During off-spill data taking, the trigger will select cosmic-ray events in the tracker and calorimeter at a rate of order 10 Hz and will prescale other events to collect unbiased samples that are needed to optimize the time-dependent trade-offs between efficiency and fake rates.

At the design rate of protons on target, Mu2e TDAQ system will produce 7 PB/year of raw data, which will be copied to the Fermilab computer center for processing and long-term storage.

Mu2e is currently under construction. Mu2e will start taking cosmic-ray commissioning data early in mid calendar 2025 and first commissioning beam data late in calendar 2026. When the proton beam data begins, the rate of protons on target (POT), will be about 55% of the design rate and the experiment will ramp up towards the design POT rate during early calendar 2027. Starting in January 2028 there will be a two-year shutdown of Fermilab accelerator operations to allow building the LBNF beamline. Mu2e will resume operations at full design POT rate in the fall of calendar 2030 and run until summer 2033.

Accelerator operations for Mu2e will be managed by the Fermilab Accelerator Division while the operation of the Mu2e apparatus will be managed by staff from the collaboration and two Fermilab directorates, Particle Physics Directorate and Applied Physics and Superconduction Technologies Directorate. Operation of computing for Mu2e will be managed by the collaboration, drawing on resources, services and staff supplied by the Fermilab Computational Science and Artificial Intelligence Directorate (CSAID).

For the Snowmass 2021 planning exercise, members of Mu2e performed preliminary design studies for an upgrade of the apparatus, named Mu2e-II. The work is ongoing, The goal is to identify the minimal set of modifications to the apparatus that will improve the physics reach by a factor of 10, using the proton beam that will be available from the Fermilab PIP-II Linac. The technically limited schedule would be to begin commissioning Mu2e-II about two years after the end of Mu2e data taking.

Throughout all these periods Mu2e expects to use computing resources supplied by CSAID. On the time scale of now to five years from now it is expected that the primary resource for data production will be Fermigrid and OSG. For MC production Mu2e will also use HPC facilities that are available, and resources provided by members of the collaboration.

The comments in the above paragraph also apply to the Mu2e-II design work that will be done during this period.

Mu2e can also run reconstruction code on the big core HPC resources. The code is not thread safe and is not designed to run on GPUs. Mu2e has no short-term plans to address these issues because it would divert limited human resources from critical preparations for data taking. Once operations are established, the project will have effort available for making the code thread-safe, porting selected components to GPU based architectures and reimplementing algorithms to exploit resources that will then be available or in the pipeline. This effort could begin in calendar 2027 and is likely to be well underway during the LBNF shutdown of fiscal 2028-2029. Once Mu2e has this capability, a large fraction of reconstruction work will be well suited to GPU based HPC centers.

One could also imagine upgrading the trigger hardware to a new architecture. The return on this investment would be a greater physics reach accompanied by a reduction in the resources required for offline reconstruction, including network resources. It is very likely such a trigger system could be in place for Mu2e-II and one could imagine testing components of that system during the later stages of Mu2e. At this time, it seems unlikely that this work would converge on time for an upgrade of the Mu2e trigger compute servers during the LBNF shutdown.

In summary, on the 2 year time horizon Mu2e's computing needs will be dominated by simulations with needs similar to recent years. Towards the end of that period the project will begin to take cosmic-ray data, which will

mean a modest increase in data processing needs. In the period two to five years from now, Mu2e will complete commissioning and will have one-year physics data taking. Towards the end of that period Mu2e will begin the effort of porting code to run on new architectures. On the time scale of 5 to 10 years from now Mu2e will have 3 year run at nominal POT and, if Mu2e-II is funded, construction of Mu2e-II will begin. The Mu2e-II trigger will be in an excellent position to exploit then-available technologies.

6.6.2.5 Generalized Process of Science

6.6.2.5.1 Muon g-2 Process of Science

There is no substantial large-scale computing expected, as the collaboration is in the final analysis stage. There is a possibility that it may be necessary to reprocess or re-evaluate a systematic uncertainty, and that could revive part of the production workflow. This activity could mean the following,

- Re-running the reconstruction program on some of the raw data, involving staging potentially several PBs of data from tape and processing. The collaboration already has workflows in place centered around the local grid facility at Fermilab (FermiGrid). A small fraction of jobs may use nodes on the OSG, which could entail moving TBs of files.
- Running a deep analysis to determine a systematic uncertainty. The workflow is similar to the above, but running over the reduced reconstructed dataset, which is much smaller than the raw data (1.3 PB).
- Running a simulation campaign to understand a detector effect or systematic. A large-scale simulation campaign may produce ~1 PB of data. Simulation workflows would likely be run on FermiGrid but could potentially run at NERSC, as has been performed in the past. The latter would potentially involve moving that data back to Fermilab.

At this time, none of the above activities are anticipated.

6.6.2.5.2 Mu2e Process of Science

At design POT rate the Mu2e experiment will produce about 7 PB/year of raw data, including physics data, calibration data and special runs. The current projection is that Mu2e will achieve the physics goals of the experiment taking about 4 PB/year of raw data. For most of these data two copies will be stored on tape but a subset will have only one copy. All tape resident copies will be stored at Fermilab. The details of how many events will be selected for calibration, how many events will be written out by reconstruction and the size per event of these datasets are not well understood at this time. The current estimate is that an additional 5 PB/year will be needed for intermediate data sets, repeat processing and for storage of simulated events. This brings the total tape needs to about 13 PB/year.

When data arrives at the computer center the goal is to complete the first processing step with a turnaround of less than six hours 95% of the time. This step will select events to be used for calibration and will produce nearline monitoring of data quality. It will also split the event stream into several sub-streams based on trigger information. The datasets used as input to calibration will be quite small. First, most events will not be interesting for calibration. Second, the output events will be much smaller than the input events! In a typical Mu2e on-spill event, the duration of the live window is about 1 μs, during which time the tracker will record about 3000 hits; most triggered events will contain a single reconstructable track and only small fraction will contain more than one track. A typical reconstructed track will have 30 to 50 hits on it and the time that a track is "live" in the apparatus is about 100 ns. The vast majority of the hits are not associated with any reconstructable track and are not relevant for most calibration purposes. Therefore, the calibration datasets will be stripped of uninteresting hits, making them much smaller per event than the raw data. Mu2e can perform a similar compression on the hits in the calorimeter and CRV. A subset of the events will be written out without this compression so that they can serve as input to other studies.

Once the calibration workflows are complete, the raw data will then be processed through the main reconstruction pass. There are many open questions here: will the full raw data be read, or one of the sub-streams of the raw data? How will prescale events that have good reconstructable tracks that are of no value for physics but have value for monitoring be treated? For what fraction of the dataset will uninteresting hits be removed? The net result is that Mu2e expects the size of the output of the main reconstruction pass to be smaller than the size of the raw data.

For the first few months of data taking, the POT rate will be about 55% of nominal. For some of this time Mu2e may open up some prescales to study events that would normally be discarded but it is believed that for most of this period the computing needs will scale with the POT rate.

Not all intermediate datasets will be written to tape. For example, the datasets produced by iterations over the same data several times in order to derive a calibration set will likely be disk resident only and be discarded once the calibration information is extracted.

As discussed above, the present plan is that it would be possible to process all these data on Fermigrid, but for planning purposes it is assumed that about half of it will be processed on Fermigrid and half on opportunistic OSG. On average Mu2e expects to process each event twice, leading to an estimate of 4 TB/year of outbound data and 2.5 TB/year of inbound data.

Discussions have taken place that may lead to some calibration work being done in the UK or at INFN, using resources provided by their funding agencies. For this exercise Mu2e assumes that about 0.5 PB/year will be transferred to each location.

Figure 6.6.4 shows an estimate of the time dependence of Mu2e network bandwidth needs from now until the end of analysis.

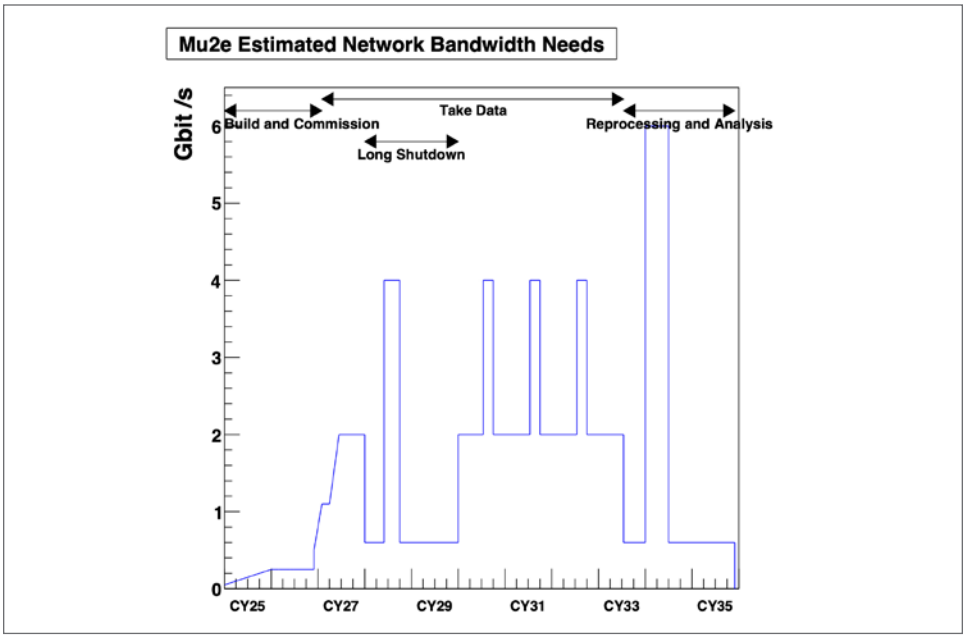


Figure 6.6.4: The blue curve shows the estimated bandwidth needs for Mu2e for all purposes listed in the table above; the notations at the top of the figure show the main periods in the lifecycle of the experiment. The spike in bandwidth during the long shutdown represents reprocessing needs during this shutdown. Similarly, the spikes each summer and in CY34 represent planned reprocessing with improved conditions and algorithms. The values shown represent averages over periods of order 1 month; within a month spikes to a few times the average may occur for periods of hours. The assumptions going into this plot are discussed in the text.

6.6.2.6 Remote Science Activities

6.6.2.6.1 Muon g-2 Remote Science Activities

The collaboration is performing the final analysis at this time. Large-scale use of remote resources is not anticipated. Possibilities that could occur were covered in the previous section.

6.6.2.6.2 Mu2e Remote Science Activities

As discussed in section 2.2 Mu2e plans to use remote resources in the following ways:

1. MC generation on OSG, HPC centers and at collaborating institutions, definitely at INFN and possibly in the UK.
2. Data reconstruction and calibration on FermiGrid, OSG and possibly at HPC centers, particularly NERSC and ALCF.
3. Calibration activities at INFN and possibly in the UK.
4. All collaborating institutes are likely to have their own copies of the analysis-format datasets of both experimental data and simulated events.

During the next two years Mu2e will be under construction and will be commissioned with cosmic rays. During this period, MC data volumes are likely to remain constant and experimental data volumes will be small, likely much less than 0.5 PB/year.

On the time scale of three to five years Mu2e will commission with beam and have a one-year run. This will be followed by the two-yearlong shutdown. During the run there will be nominal data volumes. During the two-yearlong shutdown reprocessing of the full dataset with the latest calibrations and algorithms is expected.

On a timescale of more than five years Mu2e will restart data taking following the long shutdown, complete data taking and begin final analysis. A significant increase in network needs compared to the three- to five-year forecast is not expected.

6.6.2.7 Software Infrastructure

All of the data movement and management for the Intensity Frontier Fermilab experiments, which includes Muon g-2 and Mu2e, are handled by the Fermilab FabrIc for Frontier Experiments (FIFE) project. Components include Security Account Manager (SAM) for data management, transport, and cataloging; Jobsub/HEPCloud for job submission; and Fermi-FTS for automated and robust file transfer management. (Warning: Some of these names, like SAM and FTS, are used in other projects and may lead to confusion. See the FIFE web page for the meanings here). All of these components are open source and use standard open-source tools such as xRootD.

FIFE and subcomponents are significant projects at Fermilab and are managed by the Fermilab CSAID. Note that Mu2e is in the process of migrating to Rucio.

Both experiments use Geant4 to simulate the interaction of particles with materials and are following the development of improvements to Geant4 and the development of GeantV. Mu2e has already demonstrated the ability to use multithreaded Geant4, thereby opening up running on memory poor machines such as KNL-based HPC machines.

Both experiments write their own C++ code for processing raw data into processed products (by the reconstruction program). Muon g-2 and Mu2e write their code within the art event processing framework[87]. At this time, both experiments write out data as ROOT Tree files. Many experiments are moving to Python based analysis codes. This choice is generally left to the individual analyzers.

87 https://art.fnal.gov

The general evolution of these software tools is an important open discussion for the intensity frontier experiments and especially DUNE. Muon g-2 is likely to remain with its current system of art and C++ code given the lifetime of the experiment. Mu2e will likely take advantage of progress spearheaded or in collaboration with DUNE. Fermilab has been tasked with writing a new software framework for DUNE; that project is in the requirements gathering phase.

6.6.2.8 Network and Data Architecture

The network and data architecture supporting Mu2e and Muon g-2 can be found in Section 7.4.

6.6.2.9 Emerging Technology Use

6.6.2.9.1 Cloud Services

In general, use of Cloud services is not a decision made by the experiment, but rather by Fermilab CSAID if a proposed processing task is deemed suitable and other resources are not available or not compatible. Currently, and for the foreseeable future, Mu2e and Muon g-2 will not use Cloud resources explicitly. A future feature of HEPCloud is that it may choose to use Cloud resources if the pricing is competitive and other resources are full.

6.6.2.9.2 AI and ML

Fermilab CSAID has been slowly building a local GPU resource attached to the Elastic Analysis Facility for use in AI efforts.

Muon g-2 never found a need for large-scale AI in reconstruction nor analysis. There were attempts to use ML to improve tracking efficiencies, showing some promise. Unfortunately, the effort needed to evaluate the gains further and integrate the techniques into the production workflows could not be sustained.

Mu2e uses AI in several places in the reconstruction workflow. All of it is small scale and would not significantly benefit from GPU resources. At present an AI based classifier is used to decide whether time clusters of hits are signal-like or background-like. Pattern recognition proceeds only on signal-like time clusters. Mu2e also uses AI to produce a track quality metric that decides if a given track is well reconstructed and is suitable for use in analysis, and more small scale uses are expected in the future. A graduate student is working on an AI based algorithm to replace the existing Mu2e pattern recognition algorithms. That work has just begun and it is much too early to tell if that effort will produce results that improve on the existing algorithms.

6.6.2.9.3 Wireless or Distributed Location Science Activities

These activities do not apply to Muon g-2 or to Mu2e.

6.6.2.10 Data-Related Resource Constraints

Muon g-2 does not anticipate constraints on performing the final analyses beyond sustaining the needed human effort.

Mu2e expects to be able to complete its full scientific program with currently available resources.

6.6.2.11 Data Mobility Endpoints

Data mobility endpoints are managed by the appropriate operations team in CSAID and not by the experiments themselves.

6.6.2.12 Outstanding Issues

This section does not apply to Muon g-2 or to Mu2e.

6.6.2.13 Facility Profile Contributors

Muon g-2 and Mu2e Representation

Adam L. Lyon, Fermilab, lyon@fnal.gov

Robert K. Kutschke, Fermilab, kutschke@fnal.gov

ESCC Representation

Phil DeMar, Fermilab, demar@fnal.gov

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

6.7 Belle II Experiment

The Belle II experiment collects data at the SuperKEKB asymmetric electron-positron collider at KEK in Tsukuba, Japan. SuperKEKB operations at and near the Y(4S) resonance to produce pairs of B mesons. Other heavy flavor particles, such as charm mesons and tau leptons, are also copiously produced. Belle II precision measurements of rare decays and charge conjugation parity violation (matter-antimatter asymmetry) in heavy quarks and leptons provide a unique probe of physics beyond the SM.

6.7.1 Discussion Summary

Belle II, hosted by KEK Japan, relies on a similar software infrastructure and grid-computing model as that of the LHC experiments, and has converted to use Rucio as a data management platform. Because of this sharing of infrastructure, networking and data movement work well for the current volumes. International connectivity is provided by partners Science Information NETwork (SINET) and TransPAC , via multiple direct and indirect paths. Belle II also participates in the LHCONE network to enable consistent performance between facilities.

Belle II was an active participant in the LHC DC24 activities, and presented expected results for its operations during this period. A primary reason for their participation was to test new software components that are common and shared with the LHC experiments. At the current time, the experiment does not anticipate participating in DC26, but may still monitor the situation to ensure that the increased data volumes are not disruptive to Belle II operations.

ESnet will follow up with International networking partners, namely TransPAC which is operated by International Networks at Indiana University, for updated connectivity diagrams and information regarding support for networking in the region with other partners.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The Belle II experiment collects data at the SuperKEKB asymmetric electron-positron collider at KEK in Tsukuba, Japan. Belle II's goal is to accumulate about 50 times more data than Belle, the predecessor experiment. Operations began in 2019 and about half as much data as Belle has been collected. The plan is to increase the amount of data collected per year to reach the goal by the 2030s. [Section 4.1]
- Belle II recently resumed data taking in March 2024 after a 19-month shutdown to replace the innermost vertex detector and a portion of the particle-identification detector as well as complete an upgrade of the DAQ. A SuperKEKB upgrade with a corresponding ~1.5 year shutdown, tentatively scheduled for 2029, will be required to reach the design goal. [Section 4.1]
- The Belle II collaboration comprises 1183 members from 125 institutions and 28 countries around the world. There are 118 US collaborators from 16 institutions, including BNL. [Section 4.1]
- Two copies of the raw data coming out of the Belle II DAQ at KEK are permanently stored; one copy at KEK and a second copy distributed between BNL (30%), Canada (15%), France (15%), Germany (20%) and Italy (20%). All raw data centers are connected to LHCONE. BNL houses approximately 3PB of Belle II data, and anticipates having 5PB of tape and 5PB of disk storage available approaching 2028. [Section 4.2]
- The transfers between sites are monitored and classified. Belle II participated in the DC24 LHC Data Challenge and managed to reach 250 TB/day which is more than six times the maximum throughput expected at the end of the decade. [Section 4.2]
- There is no current or future network or data-related constraint expected for Belle II. The Data Challenge performed in February 2024 demonstrated that the current network will be able to handle the Belle II needs, including safety factors, until the end of the decade. [Section 4.2]

- Belle II chose to migrate from its bespoke data management solution to Rucio in January 2021. Since then, Rucio, with FTS as back-end transfer system, has managed all production and analysis data production on the grid. Rucio ensures the replication of raw data from KEK to the other raw Data Centers according to the defined share detailed in the previous section. It also ensures the replication of the dataset output generated by the production system and makes a second copy of user or group data to other sites. Since the migration to Rucio in 2021, no network bottlenecks have been observed. [Section 4.3]
- Many of the Tier-1 and Tier-2 sites in Belle II are part of the WLCG collaboration and shared with LHC experiments. This also includes the network connecting these sites, in particular the ones connected by LHCOne. The sites without LHCOne connection are small and can be used without LHCOne. [Section 4.3]
- Belle II has a grid-computing model with a three-level hierarchical structure ("tiers") of computing sites: raw data centers (Tier-1), regional data centers(Tier-2) and MC(simulated data) production centers(Tier-3). [Section 4.3]
- The fully reconstructed events, coming from the raw data-processing step, are stored in the miniDST (mDST) format. MC events are simulated and reconstructed using the same software used to process detector events and then also stored in mDST format. [Section 4.3]
- Belle II Reprocessing campaigns based on improved calibration run typically once per year or biyearly. Reprocessing will be every other year beginning 2025. [Section 4.3]
- Belle II uses a grid model initially developed to support the computing needs of the LHC experiments. This model has a hierarchical structure with a Tier-0 site that produces and hosts a primary copy of the raw data produced by the experiment (KEK in the Belle II case), Tier-1 sites (also known as raw Data Centers) that host a secondary copy of the raw data and that are used for prompt processing and reprocessing and Tier-2 sites that are mainly dedicated to MC production and user analysis. [Section 4.4]
- The processing of the MC or data files in Belle II is always done on the site that hosts the data (no WAN access) except for the small sites that do not host a Storage Element. The jobs always go to the data; data are not moved to sites with empty computing sites. The output of the jobs are also stored locally, except for the sites without storage elements or if the storage element is temporarily inaccessible. Additional copies are created based on the replication policies using the Rucio subscription mechanism. [Section 4.4]
- The BNL network perimeter has two diverse 800 Gbps (2 x 400 Gbps) circuits that peer with ESnet routers. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits. The BNL Network Perimeter on average ingests and egresses between 10–15 PB of data monthly. [Section 4.7]
- The success of Belle II relies on transpacific networking capacity provided by a number of partners that participate in the LHCONE overlay network. Belle II relies heavily on network access, thus the partnerships that underpin LHCONE are critical for ongoing operations. [Section 4.7]
- Belle II is connected to SINET , a Japanese academic backbone network that works in collaboration with Asia Pacific Advanced Network (APAN). SINET is connected to the United States via multiple direct and indirect paths; 400 Gbps via partners PacificWave (PacWave), Indiana University International Networks, and Japan Gigabit Network (JGN) [Section 4.7]

6.7.2 Belle II: Experiment Profile

Belle II's goal is to accumulate about 50 times more data than Belle, the predecessor experiment. Operations began in 2019 and about half as much data as Belle has been collected. The plan is to increase the amount of data collected per year to reach the goal by the 2030s. The Belle II collaboration comprises 1183 members from 125 institutions and 28 countries around the world. There are 118 US collaborators from 16 institutions, including BNL. The US contribution to Belle II is funded by DOE OS HEP.

6.7.2.1 Science Background

Belle II has a grid-computing model with a three-level hierarchical structure ("tiers") of computing sites: raw data centers (Tier-1), regional data centers(Tier-2) and MC(simulated data) production centers(Tier-3). Raw data centers are also regional data centers and MC production centers. Regional data centers are also MC production centers. The raw data coming out of the Belle II DAQ at KEK are permanently stored, calibrated and processed at raw data centers. Two copies of the raw data are stored. One copy at KEK and a second copy distributed between BNL (30%), Canada (15%), France (15%), Germany (20%) and Italy (20%). All raw data centers are connected to LHCONE.

The fully reconstructed events, coming from the raw data-processing step, are stored in the miniDST (mDST) format. MC events are simulated and reconstructed using the same software used to process detector events and then also stored in mDST format. Detector and MC mDST are stored in regional data centers. User analysis is generally performed on filtered subsets of the mDSTs dubbed uDSTs. Generally, two replicas of mDSTs and uDSTs are stored. In addition, Belle II retains replicas of the results of processing with a previous major software release, which are expected on a yearly basis.

The use and allocation of Belle II computing resources is re-evaluated annually, taking into account past performance and experience as well as expected changes in data rates. The computing model is periodically updated to ensure efficient usage of resources.

Belle II recently resumed data taking in March 2024 after a 19-month shutdown to replace the innermost vertex detector and a portion of the particle-identification detector as well as complete an upgrade of the DAQ. A SuperKEKB upgrade with a corresponding ~1.5 year shutdown, tentatively scheduled for 2029, will be required to reach the design goal. The data volume and resource estimates take these events into account.

6.7.2.2 Collaborators

The resource levels for 2024 are provided in Table 1. Note that secondary and primary copy storage is listed separately (e.g., Italy holds both a primary and secondary copy and thus appears twice). Transfer to primary storage is constant during data taking (up to nine months per year) and minimal otherwise. Transfer to secondary storage occurs during production of simulated data or processing of raw data which nominally occurs throughout the year with an 83% duty factor. Distributed data management techniques, discussed later in this case study, enable "grid" access.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
CANADA	Primary	Grid	0.34 PB	Constant during data taking	N	None
FRANCE	Primary	Grid	0.34 PB	Constant during data taking	N	None
GERMANY	Primary	Grid	0.46 PB	Constant during data taking	N	None
ITALY	Primary	Grid	0.46 PB	Constant during data taking	N	None
KEK	Primary	Grid	4.65 PB	Constant during data taking	N	None
USA	Primary	Grid	3.05 PB	Constant during data taking	N	None
ARMENIA	Secondary	Grid	0.09 PB	Constant during production	N	None
AUSTRALIA	Secondary	Grid	0.35 PB	Constant during production	N	None
AUSTRIA	Secondary	Grid	0.31 PB	Constant during production	N	None
CANADA	Secondary	Grid	0.54 PB	Constant during production	N	None
CHINA	Secondary	Grid	1.39 PB	Constant during production	N	None
CZECH	Secondary	Grid	0.22 PB	Constant during production	N	None
FRANCE	Secondary	Grid	0.8 PB	Constant during production	N	None
GERMANY	Secondary	Grid	2.7 PB	Constant during production	N	None
INDIA	Secondary	Grid	0.54 PB	Constant during production	N	None
ISRAEL	Secondary	Grid	0.18 PB	Constant during production	N	None
ITALY	Secondary	Grid	2.52 PB	Constant during production	N	None
KEK	Secondary	Grid	4.96 PB	Constant during production	N	None
KOREA	Secondary	Grid	0.81 PB	Constant during production	N	None
MALAYSIA	Secondary	Grid	0.09 PB	Constant during production	N	None
MEXICO	Secondary	Grid	0.27 PB	Constant during production	N	None
POLAND	Secondary	Grid	0.18 PB	Constant during production	N	None
RUSSIA	Secondary	Grid	1.39 PB	Constant during production	N	None

SAUDI ARABIA	Secondary	Grid	0.13 PB	Constant during production	N	None
SLOVENIA	Secondary	Grid	0.49 PB	Constant during production	N	None
SPAIN	Secondary	Grid	0.04 PB	Constant during production	N	None
SWEDEN	Secondary	Grid	0 PB	Constant during production	N	None
TAIWAN	Secondary	Grid	0.31 PB	Constant during production	N	None
THAILAND	Secondary	Grid	0.18 PB	Constant during production	N	None
TURKEY	Secondary	Grid	0.09 PB	Constant during production	N	None
UK	Secondary	Grid	0 PB	Constant during production	N	None
UKRAINE	Secondary	Grid	0.18 PB	Constant during production	N	None
USA	Secondary	Grid	2.66 PB	Constant during production	N	None
VIETNAM	Secondary	Grid	0.09 PB	Constant during production	N	None

Table 6.7.1: Belle II Data Mobility

6.7.2.3 IRI Readiness

Not applicable to Belle II.

6.7.2.4 Instruments and Facilities

Figure 6.7.1 shows the amount of disk and tape resources provided by the Belle II Tier-1 site at BNL. Values up to 2024 are as provided; values from 2025 are from the most up-to-date projection of August 2020. The BNL raw data archive at the Tier-1 site was required to house a 100% copy of the raw data through 2021.

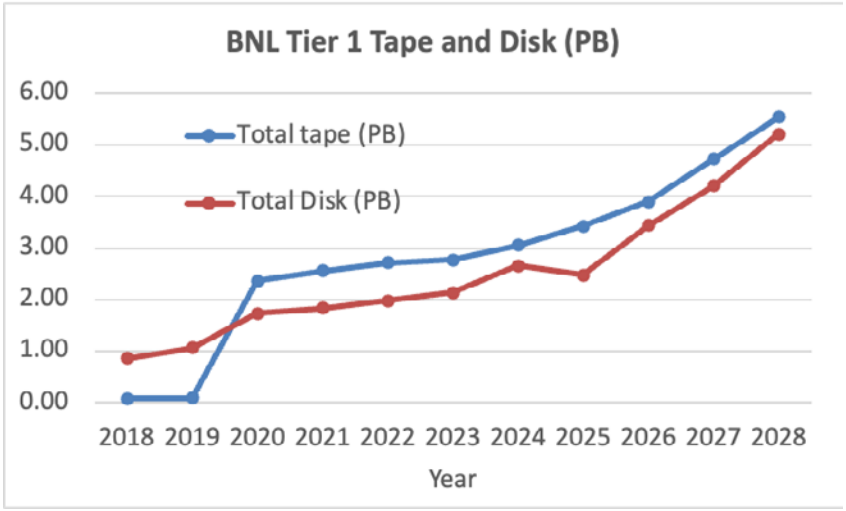


Figure 6.7.1: Belle II tape and disk resources at BNL.

6.7.2.5 Generalized Process of Science

The different workflows used by Belle II are briefly explained in this section. Details are provided in the next section.

All raw data produced at KEK are archived at KEK and one of the Tier-1 (raw data centers) sites (Section 6.7.2.1) to ensure that these data, that cannot be reproduced, can be recovered in case one of the copies is corrupted or lost. A subset of the raw data is used for detector calibration. The calibration is applied to process the data (aka prompt reconstruction) to produce files that can be used by the end users to conduct their analyses. In addition to real data, MC data are also constantly generated.

Reprocessing campaigns based on improved calibration run typically once per year or biyearly. Reprocessing will be every other year beginning 2025.

The final important workflow is the user analysis where end users use the reconstructed data coming from the detector as well as the MC data to extract physics results.

6.7.2.6 Remote Science Activities

Belle II uses a grid model initially developed at the turn of the millennium to support the computing needs of the LHC experiments. This model has a hierarchical structure with a Tier-0 site that produces and hosts a primary copy of the raw data produced by the experiment (KEK in the Belle II case), Tier-1 sites (also known as raw Data Centers) that host a secondary copy of the raw data and that are used for prompt processing and reprocessing and Tier-2 sites (Table 1) that are mainly dedicated to MC production and user analysis.

Many of the Tier-1 and Tier-2 sites in Belle II are part of the WLCG collaboration and shared with LHC experiments. This also includes the network connecting these sites, in particular the ones connected by LHCOne. The sites without LHCOne connection are small and can be used without LHCOne. Because of these resources sharing, Belle II is an associate member of WLCG. This grid-based computing model is not foreseen to change for the lifetime of the experiment.

In this grid model framework, the raw data produced at KEK are exported from KEK to the raw data centers according to the shares in Table 6.7.2. The sharing is implemented by Rucio.

Site	BNL (US)	CNAF (IT)	IN2P3CC (FR)	uVic (CA)	DESY (DE)	KIT (DE)
FRACTION OF RAW DATA	30%	20%	15%	15%	10%	10%
ESTIMATED VOLUME (TB/DAY) EXPORTED FROM KEK IN 2029	12	8	6	6	4	4

Table 6.5.1: CMB-S4 computational challenge summary

As described in the previous section, a fraction of raw data (calibration skims, raw data format) is, in addition, exported and processed at BNL to derive calibration constants. The calibration proceeds in several stages with an important reusable intermediate data product, the cDST, being a byproduct: many re-calibrations can reuse this data format. The raw data are approximately 70 kB/event, while a cDST is approximately 120 kB/event but is produced on only a fraction of the data. Calibration constants are uploaded to the Conditions database at BNL.

Once calibration constants are available, the raw data can be reconstructed to produce the mDST that can, in principle, be used for physics analysis; it has a size of approximately 15 kB/event. As the retention rate of most analyses is on the order of 1%, a subsequent processing step produces uDSTs for several event selections. The event size of this data format is approximately 20 kB/event as information is added to the mDST information. Physicists process the uDST files with a grid-based analysis framework (gbasf2) to produce ntuples that they download to their local resources for subsequent analysis.

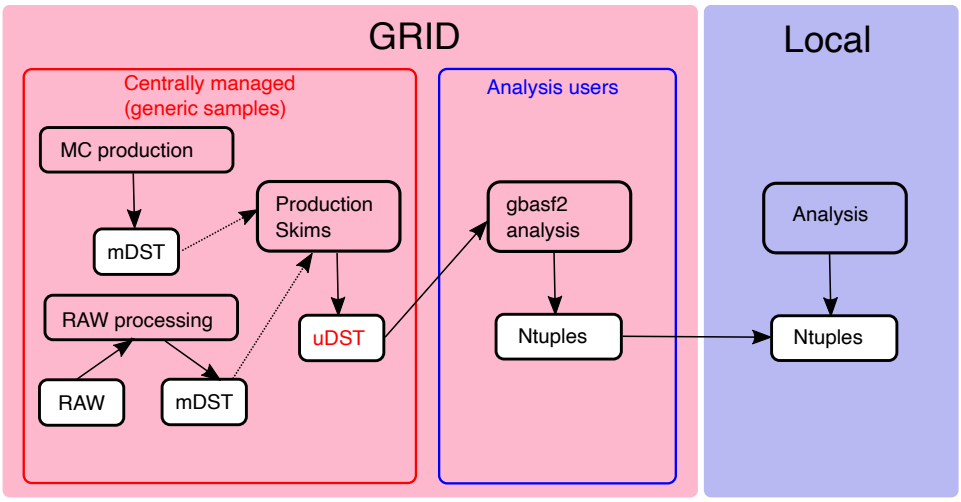


Figure 6.7.2: Belle II computing model

In the Belle II computing model (see Figure 2), raw data processing only occurs at raw data centers. The output mDST is distributed across seven regional data centers. uDST files are distributed across many more sites. For simulation, the picture is similar but simpler: mDST is produced directly across the more numerous MC production sites, and uDST distribution follows. A key input to simulation are samples of background data taken from real data. These background files must be distributed across the MC production sites to respect the MC production policy.

The processing of the MC or data files is always done on the site that hosts the data (no WAN access) except for the small sites that do not host a Storage Element. The jobs always go to the data; data are not moved to sites with empty computing sites. The output of the jobs are also stored locally, except for the sites without storage elements or if the storage element is temporarily inaccessible. Additional copies are created based on the replication policies using the Rucio subscription mechanism.

The transfers between sites are monitored and classified into different "activity" as shown in Figure 3. As shown in this figure, Belle II participated in the DC24 Data Challenge and managed to reach 250 TB/day which is more than six times the maximum throughput expected at the end of the decade. The other activities have much lower throughput. The second highest throughput is the staging activity that only occurs during staging campaigns (once per year before 2025, then once every two years after) which do not involve WAN transfers since the data are only staged at the Tier-1 sites. Since the migration to Rucio in 2021, no network bottlenecks have been observed. Although DC24 showed satisfactory results for Belle II for the rest of the decade, Belle II plans to participate in future Data Challenges because many Belle II sites are co-located with LHC sites, which might affect throughput in the future.

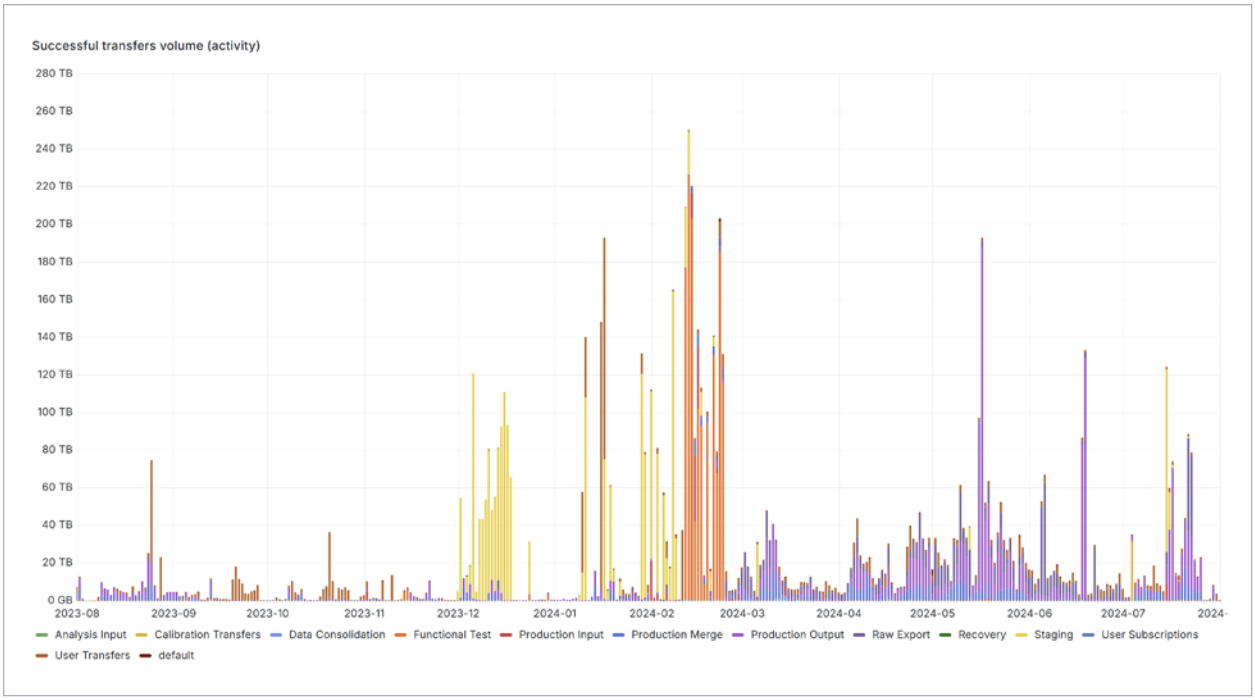


Figure 6.7.3: The figure shows the daily volume transferred by Rucio over the last year, split by activity. The maximum activity observed peaking at 250 TB/day corresponds to the WLCG Data Challenge DC24. The other activities have much lower throughput. Staging activity is always done within the same site and does not use WAN.

Other remote science activities include :

- Downloading of job output by the end users. This activity fluctuates significantly with time day/night, weekend, confrence, etc. However, the volume for these transfers is negligible with respect to the intersite transfers.
- Access to condition database. The condition database service (see next section) is hosted at BNL and accessed by each job, but direct access to the service is limited by local and remote caching.

6.7.2.7 Software Infrastructure

6.7.2.7.1. Rucio

Rucio, initially developed by and for the ATLAS collaboration, is now a common Distributed Data Management solution within the HEP community and beyond. It is used in production by many big collaborations (e.g., ATLAS or CMS) at scale well above the one from Belle II (e.g., ATLAS currently manages 1 EB which is much more than the 25-30 PB stored by Belle II). Rucio now enjoys large community support.

Belle II chose to migrate from its bespoke data management solution to Rucio in January 2021. Since then, Rucio, with FTS as back-end transfer system, has managed all production and analysis data production on the grid. Rucio ensures the replication of raw data from KEK to the other raw Data Centers according to the defined share detailed in the previous section. It also ensures the replication of the dataset output generated by the production system, again according to predefined share, takes care of the staging for the reprocessing campaigns, and makes a second copy of user or group data to other sites.

In January 2024, it was decided that Rucio will also become the official metadata service for Belle II to replace the old metadata service called ARDA Metadata Grid Application (AMGA) hosted and operated at KEK. This transition is expected to occur in 2025.

This critical service is operated at BNL.

6.7.2.7.2. DIRAC

Belle II uses DIRAC as a Workflow and Workload Management system. It allows users to interact with all grid resources to schedule and run production and analysis jobs, etc. DIRAC was initially developed by LHCb and, similarly to Rucio, it is now used by a large community. In order to fit Belle II needs, an extension to DIRAC, called BelleDIRAC, was developed. In particular, this extension implements a Production system that enables large-scale MC production as well as an interface to Rucio.

Most of the BelleDIRAC services are operated at KEK. A single service for Data Management certification is operated at BNL.

6.7.2.7.3. Conditions Database

BNL operates the Belle II Conditions Database (CDB) service. The CDB supports data processing on the Belle II computing grid. In addition, as part of the detector's calibration cycle, the CDB serves as persistent storage for selected data collected by the Belle II DAQ at KEK which is properly formatted for CDB injection. The CDB system is composed of two sub-services: the payload service and the metadata service. The payload service is accessed if the client-side software cannot find a copy of the payload on CERN Virtual Machine File System (CVMFS)(payloads are migrated to CVMFS with a latency of hours) or on the client's local caches such that the load on the CDB payload service at BNL for reading is typically modest. Writing new calibration data payloads, which can only proceed via the CDB service at BNL, is also typically modest as the CDB service is a mostly read service. The CDB metadata service provides the list of which conditions data payloads needed for a given data-processing job. This service also has a selective backup on CVMFS but, unlike for payloads, the metadata service is the primary source. The data transfer to or from this (metadata) service per query is also modest. Nevertheless, bursts of requests to the CDB system have been observed and successfully supported. Squid caches have been set up at the major data centers to intercept part of the load going to the CDB service, reducing even further the traffic to the service.

A migration to a new CDB service based on the recommendations of the HEP Software Foundation is foreseen in the next few years. This new service should not significantly change the network needs for this service.

6.7.2.7.4. CVMFS

CernVM Files System(CVMFS) is a well-known service that enables distribution of software. In addition to software distribution, CVMFS is also used to distribute the CDB payload as explained in the previous subsection.

6.7.2.8 Network and Data Architecture

6.7.2.8.1 Domestic Connectivity

The SDCC at BNL has implemented a vendor agnostic, resilient, scalable, and modular Tbps High-Throughput Science Network (HTSN) which serves as the primary network transport for all data-intensive collaborations at BNL (Figures 4, 5, 6, 7 and 8). It provides high-throughput connectivity to all HPC and HTC collaborations and supports the timely transfer of large amounts of scientific data via the internet.

The HTSN has five key components:

1. Network Perimeter
    - Two diverse 800 Gbps (2 x 400 GbE) circuits that peer with ESnet routers within the BNL 515 and 725 data centers. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits.
    - The BNL Network Perimeter on average ingests and egresses between 10–15 PB of data monthly.

2. Science DMZ
    - Supports open, high-speed WAN/Internet) access for all scientific collaborations throughout the BNL campus.
3. Science Core
    - A Tbps Science and Data Center Interconnect for data intensive collaborations at BNL. This Science Interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, sPHENIX, CAD, Center for Functional Nanomaterials (CFN), NSLS-II, High-Performance Computing Clusters and SDCC.
    - Intelligence and routing policies are applied within the Science Core to restrict or grant access to specific resources within the SDCC.
4. Spine
    - A Tbps network Spine that interconnects all Leaf switches. Leaf switches can consist of top of rack (ToR) or chassis-based switches that connect compute, storage or general infrastructure service servers.
    - The responsibility of the Spine is fast packet forwarding and flexibility, not policy insertion or server termination.
    - The External Border Gateway Protocol (eBGP) is utilized throughout the HTSN, and was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each Spine group shares the same Autonomous System Number (ASN) but does not have Internal Border Gateway Protocol (IBGP) peering between them. Each Leaf or pair of Leaves will require their own ASN.
5. Storage Core
    - A redundant terabit per second switching block that aggregates high-performance storage services.

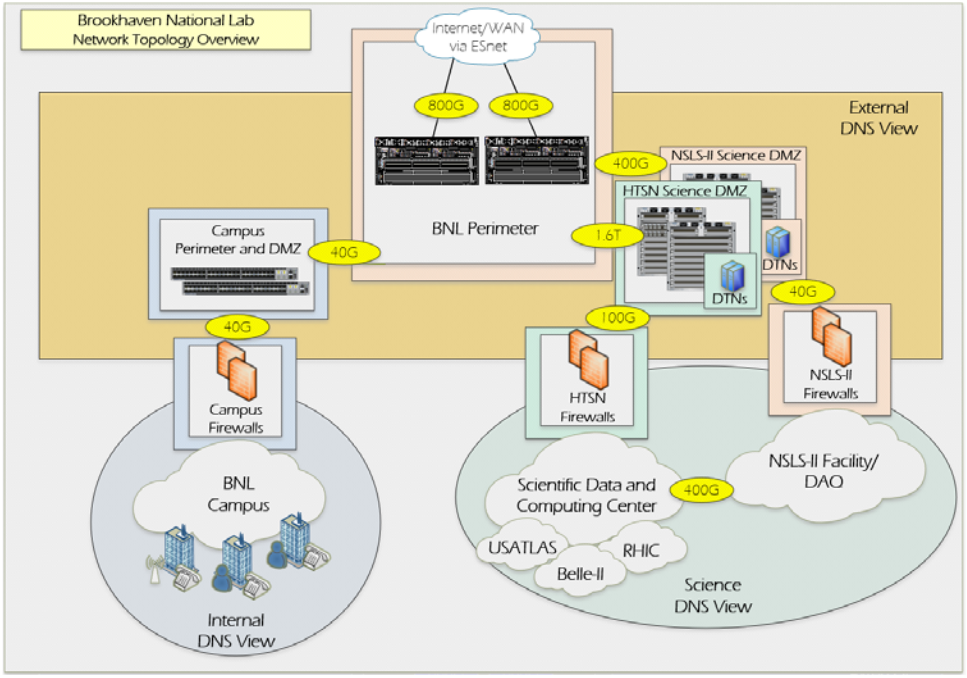


Figure 6.7.4: High-level overview of the BNL network perimeter and domain name service (DNS) architecture.

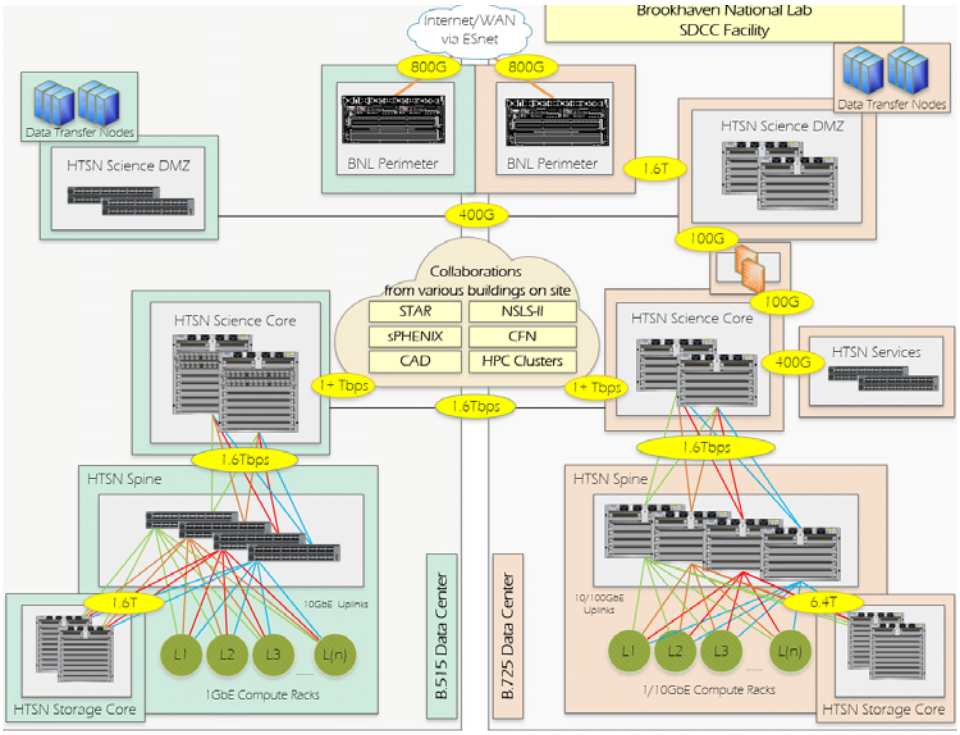


Figure 6.7.5: High-level overview of the BNL HTSN in FY24 (includes HTSN expansion into new data center, right hand side).

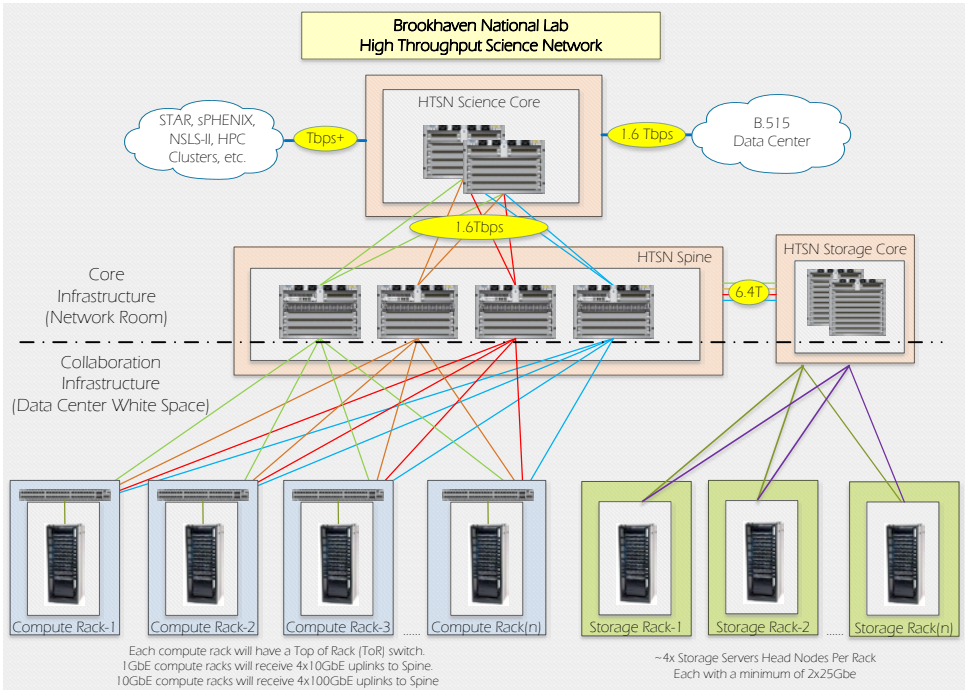


Figure 6.7.6: High-level overview of the BNL HTSN, demarcation points between core and collaboration infrastructures.

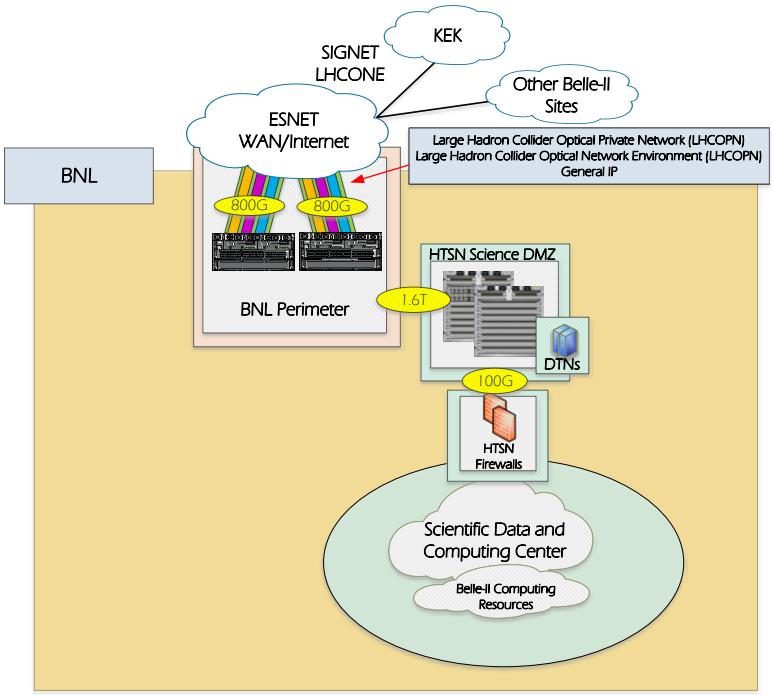


Figure 6.7.7: Network diagram Belle II Tier-1 site at BNL as of Aug 2024.

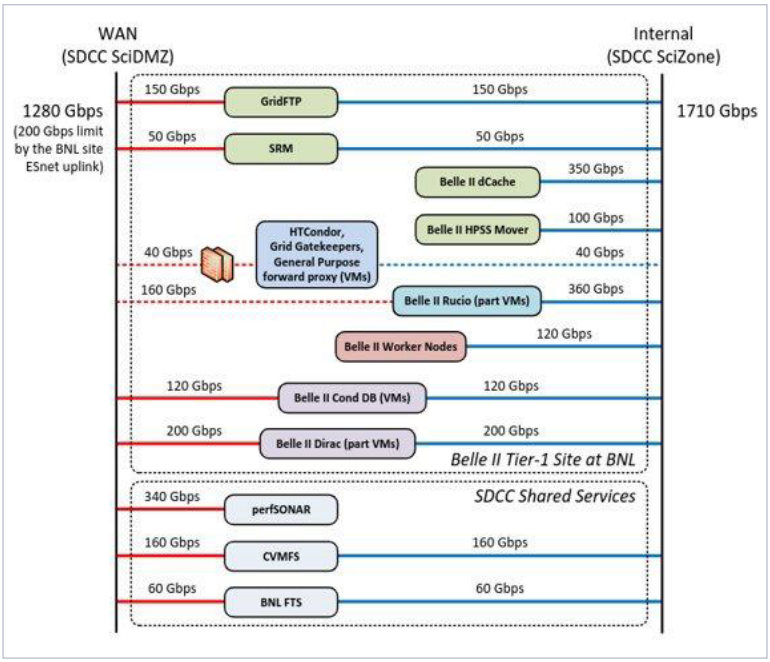


Figure 6.7.8: Belle II service capacity (LAN and WAN)

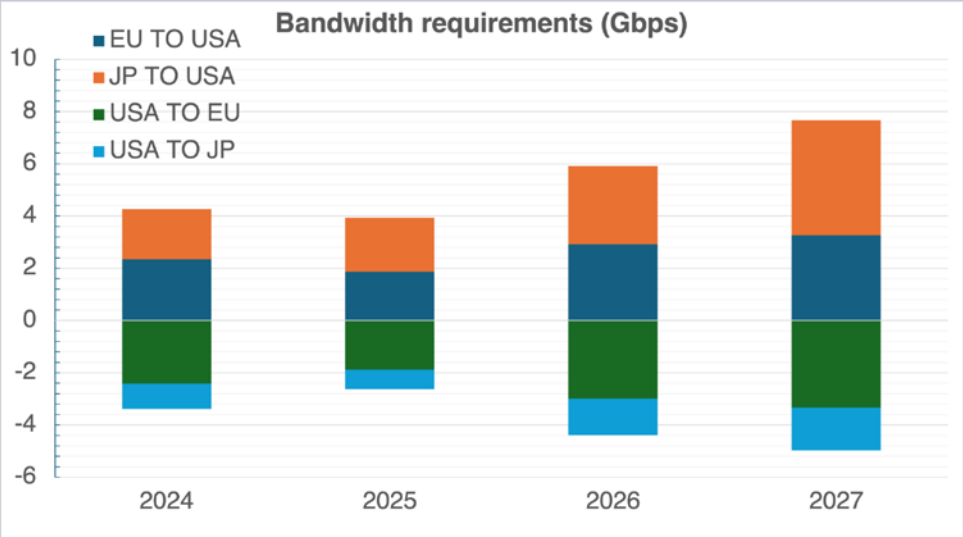


Figure 6.7.9: International bandwidth requirements for Belle II.

Reliability and performance of the network between main data centers are monitored with perfSONAR.[88]

6.7.2.8.2 International Connectivity

The success of Belle II relies on transpacific networking capacity provided by a number of partners that participate in the LHCONE overlay network. Belle II relies heavily on network access, thus the partnerships that underpin LHCONE are critical for ongoing operations. LHCONE delivers service using physical resources operated by a number of different entities:

- Belle II is connected to SINET,[89] a Japanese academic backbone network that works in collaboration with APAN.[90]
- SINET is connected to the United States via multiple direct and indirect paths:
  - 100 Gbps is available via the NSF-funded PacificWave (PacWave)[91] project to Los Angeles, California.
  - 100 Gbps is available via the NSF-funded TransPAC link,[92] operated by Indiana University International Networks to Seattle, Washington.
  - 100 Gbps via the National Institute of Communications Technology R&D testbed network: JGN[93] that connects to the United States via Hong Kong and Guam. This connectivity is operated in partnership with the Pacific Islands Research (PIREN) project[94] and TransPac.
  - 100 Gbps to Singapore operated with JGN, where alternate peering arrangements can function as backups to carry traffic to the United States via Guam or Australia.
- The PacWave exchange points in the United States are located in Seattle ,Washington; Los Angeles, California; and Sunnyvale, California. Domestic networks such as Internet2, ESnet, and CENIC,[95] as well as other international partners, maintain peering here.

Figure 6.7.9 shows the international bandwidth requirements for US Belle II.

88 http://maddash.aglt2.org/maddash-webui/index.cgi?dashboard=Belle%20II%20Mesh%20Config
89 https://www.sinet.ad.jp
90 https://apan.net/about
91 http://pacificwave.net
92 https://internationalnetworks.iu.edu/projects/TransPAC/index.html
93 https://testbed.nict.go.jp/jgn/english/
94 https://www.hawaii.edu/piren/
95 https://cenic.org

6.7.2.9 Emerging Technology Use

6.7.2.9.1 Cloud Services

In 2022 BNL moved to using Microsoft Office 365 for email, calendar service, Sharepoint, and OneDrive cloud storage. Although, bandwidth to and from these services is much lower than that required for scientific data, the reliance of the laboratory on cloud services for day-to-day business has made the resilience of Internet connectivity critical. During periods of peak scientific data movement, cloud data for day-to-day work has been affected, and QoS and other traffic engineering controls were put in place to protect the business traffic from being affected by the scientific data flows.

6.7.2.9.2 AI and ML

Belle II makes use of ML applications for triggering, track reconstruction, particle identification, etc., but Belle II does not expect significant network impact or noteworthy changes to the distributed computing software stack.

6.7.2.9.3 Wireless or Distributed Location Science Activities

Not applicable to Belle II.

6.7.2.10 Data-Related Resource Constraints

There is no current or future network or data-related constraint expected. The Data Challenge performed in February 2024 demonstrated that the current network will be able to handle the Belle II needs, including safety factors, until the end of the decade.

6.7.2.11 Data Mobility Endpoints

Not applicable to Belle II.

6.7.2.12 Outstanding Issues

Not applicable to Belle II.

6.7.2.13 Facility Profile Contributors

Belle II Representation

Hironori Ito, BNL, hito@bnl.gov

David Jaffe, BNL, djaffe@bnl.gov

Alexei Klimentov, BNL, aak@bnl.gov

Cedric Serfon, BNL, cedric.serfon@cern.ch

ESCC Representation

Vincent Bonafede, BNL, bonafede@bnl.gov

Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

6.8 Neutrino Experiments at Fermilab: SBN and DUNE

Fermilab features a number of experiments focused on neutrinos, often hard to detect particles that will help to understand the origin of matter as well as the unification of forces. This detection can be done over both short and long distances. The two experiments profiled are examples of this: SBN and DUNE. Both focus on the study of neutrino oscillations, and use a similar set of scientific technology for observation. It is also expected that both experiments will share similar approaches to their implementations of computational and storage infrastructures at Fermilab, and their approach to wide-area networking. The work of SBN and the ProtoDUNE experiments at CERN will prepare for DUNE, which is scheduled to start in several years' time. Both experiments will utilize grid-computing approaches provided by OSG, WLCG, and HEP-community software for data movement, cataloging, simulation, and analysis.

6.8.1 Discussion Summary

The SBN program anticipates using more DOE HPC resources for their simulation workflows. The program does anticipate needing more disk space during this process, which limits their output.

SBN's primary concern is not network bandwidth, but rather I/O performance. The project relies heavily on tape-based storage, which causes a bottleneck when reading out older data. Fixes are being explored to rely less on this approach, but which may increase the expected data volumes.

DUNE can take advantage of some aspects of the IRI design approaches, and the collaboration is currently using NERSC and exploring using ALCF for some aspects of their simulation and analysis workflows. In particular, it is expected that GPU resources required for some parts of the workflow may need to be used at the DOE HPC facilities. DUNE has not fully identified a long-term location for the storage of their public data archive. This may also fall into the realm of an IRI workflow, or an emerging use case for HPDF to assist with.

DUNE's software stack will leverage many of the tools that are already in use within the HEP community including Rucio, FTS, and xRootD. DUNE is also looking into facilitating Globus as a transport mechanism with the workflow managers such as Rucio and xRootD. DUNE has converted to use containers, which has greatly increased the ability to make the analysis work portable across available resources, particularly at DOE HPC centers like NERSC.

DUNE participated in the DC24 activities with the LHC sites, and learned a lot about their approach to software and data management. They anticipate participating in future DCs as well, due to the overlap they have with other WLCG computing resources. DUNE will also conduct their own DC activities on an annual basis. In particular, being able to handle production traffic from the detector, as well as a supernova activity, will be an area of focus. This particular part of the DC will rely on the connection between SURF and Fermilab when the network is provisioned (expected 2026 or 2027). For the DC activities prior to this, BNL will currently serve as a stand-in for SURF.

DOE's HEP and ASCR program offices are still working out the specifics of the connectivity to SURF that ESnet will need to complete for the proper connection. The efforts are planned to converge by the end of fiscal year 2026.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The SBN program studies properties of neutrinos and neutrino interactions using beams of accelerator-generated neutrinos at Fermilab. In total three detectors—- the SBND, MicroBooNE, and the short baseline far detector (ICARUS T600)—- detect neutrinos from Fermilab's BNB. The SBN detectors can measure important properties of neutrino interactions on argon that will be critical for the future long-baseline neutrino experiment DUNE. [Section 4.1]

- The three international SBN collaborations comprise of approximately 250 scientists from more than 50 institutions, and include members from across the US, CERN, Italy, the UK, Switzerland, Spain, Turkey, and Brazil. [Section 4.1]
- The work of SBN and the ProtoDUNE experiments at CERN will prepare for DUNE, which is scheduled to start in several years' time. Both experiments will utilize grid-computing approaches provided by OSG, WLCG, and HEP-community software for data movement, cataloging, simulation, and analysis. [Section 4.1]
- DUNE will study the fundamental nature of neutrinos through measurements of neutrino oscillations from both neutrino beam and other natural neutrino sources. In addition to studying neutrino oscillations, DUNE has a broader physics program that includes searches for proton decay, detection of neutrinos from supernova collapse, and other exotic phenomena. The core facilities of DUNE are the near and far detectors, located on-site at Fermilab (Illinois) and in SURF (South Dakota). The detectors are placed such that they are on-axis with the beam of the future LBNF at Fermilab. The separation of the DUNE near and far detectors means that the experiment is relying on a WAN that originates at SURF in South Dakota, and must transfer all data back to Fermilab: this emphasis on near-constant network connectivity is shaping the choices made for buffering, storage, and analysis at both locations. [Section 4.1]
- The DUNE Collaboration comprises approximately 1500 scientists from more than 200 institutions in 39 countries around the world. These institutions include both national laboratories and universities. [Section 4.1]
- DUNE has built two prototype cryostats at CERN for testing of final far detector technology designs, ProtoDUNE-HD and ProtoDUNE-VD. These two cryostats provide an opportunity to test both the detector technology but also the offline computing resources and software. ProtoDUNE-HD has just completed a three-month beam run at CERN, and ProtoDUNE-VD is anticipated to start a roughly yearlong test period in late 2024. The ProtoDUNE-VD test run will comprise both beam and cosmic-ray data taking. [Section 4.1]
- The SBN detectors (ICARUS and SBND) will produce approximately 5 PB of data per year that will be stored at Fermilab, and at backups sites. [Section 4.2]
- The SBN ICARUS detector has been taking data since 2021. The raw data rate from the detector is 2 PB, with raw data being stored at Fermilab and a backup for a significant portion of ICARUS data stored at CNAF. The SBND detector is completing the commissioning phase and is expected to take physics-quality data in fall 2024. The current expected raw data rate from the detector is 3 PB. Data management and processing will follow similar strategies as ICARUS, complementing the primary Fermilab resources with additional off-site data centers. SBN is expected to end operations at the end of FY27. [Section 4.2]
- ICARUS data are expected to be approximately 100 MB per event at the raw data stage, but then only 40 MB/event after first stage reconstruction. The second stage of reconstruction will increase that size by approximately a factor of 2, but then high-level reconstruction objects can be summarized to then create a final size of approximately 50 kB per event. [Section 4.2]
- For normal data taking from the ICARUS detector, raw data arrives at a steady rate of approximately 370 GB/hr (~100 MB/s), and from SBND at approximately 320 GB/hr (~90 MB/s). These data are delivered to Fermilab, and is made available for immediate processing via xRootD to run on-site at Fermigrid, but can be accessed and run on OSG sites as available. Raw data backups are ~75 GB/hr (~20 MB/s) to CNAF for ICARUS, and 130 GB/hr (~35 MB/s) for SBND (Polaris at ANL). [Section 4.2]

- When operating, the DUNE FD will produce approximately 30PB of data per year, and the near detector will produce approximately 5PB of data per year that will be stored at Fermilab, and back-up sites. ProtoDUNE efforts have produced approximately 5PB of data per year at CERN during the initial run, and will exhibit a similar pattern for a second run between 2025 and 2026. [Section 4.2]
- Transfer of the raw data produced from the DUNE detectors is likely to be the largest single network task for the experiment. The data must be transferred from the detectors to archival storage at Fermilab and at least one additional archival location geographically separate from storage at Fermilab. It is anticipated that along with initial processing, there will be annual reprocessing of all data accumulated to date. In addition to beam and calibration data, the signal of neutrinos produced from a nearby supernova burst present a significant challenge due to the large volume of data collected in a very short period of time (up to 600 TB in 100 seconds) and the time-critical nature of analyzing that data for optical astronomical observations. [Section 4.2]
- The first full DUNE Far Detector, FD-VD, is slated to begin full operation at SURF in Lead, SD in mid-2029. A second module will come online in 2030-31. These systems will require significant testing and calibration and are projected to produce 10–15 PB of data per year per module to start, and will be capped at 30 PB/year when in full operation. This includes tests of supernova burst sensitivity that will produce around 320 TB over a few hours (they are actually produced over 100s but are constrained by local network speeds). Early calibration runs will also produce significant data volumes over short time periods. Steady rates of 100TB/day are anticipated. [Section 4.2]
- DUNE, as a whole, will produce 45PB of data per year at full operation. This includes the near and far detectors, and 10PB of processed data. [Section 4.2]
- DUNE data are expected to be approximately 8 GB per event at the raw data stage, but then only 20 MB/event after first stage reconstruction. The second stage of reconstruction will increase that size by approximately a factor of 2, but then high-level reconstruction objects can be summarized to then create a final size of approximately 20kB per event. [Section 4.2]
- There are three DUNE prototype detectors that will be active during the next two years: ProtoDUNE-HD, ProtoDUNE-VD, and ND-lAr 2x2 demonstrator. The ProtoDUNE detectors are located at CERN, while the demonstrator is located at Fermilab. The raw data arrive from the ProtoDUNE detectors at a steady rate of approximately 1 TB/hr (~300 MB/s), and from the 2x2 detector at approximately 42 GB/hr (~15 MB/s). The ProtoDUNE data are delivered to CERN on-site, and a second copy of the ProtoDUNE data is also transferred to Fermilab. The raw data are made available for immediate processing, which is accessed via xRootD to run on the WLCG. For the 2x2 demonstrator, the data are automatically stored at Fermilab, and then transferred in batches to a DUNE allocation at NERSC for processing. [Section 4.3]
- In addition to detector data, DUNE will produce large quantities of simulation that in some cases will rival the size of the detector data. This data will be produced at distributed compute facilities and be cataloged and stored in storage elements around the globe. DUNE is currently using more than 50 compute facilities, and more than 10 storage facilities to achieve its computing tasks. It is expected that this number will grow between now and late 2028 when the first far detector module will become operational. [Section 4.3]
- DUNE data acquisition for the first far detector module will commence post 2029, and the output of the module will be roughly 30 PB/year going to archival storage. Offline processing and data flows will look very similar to prototype operations but with slightly different endpoints. (e.g., instead of data coming from CERN, data will be created at SURF). There is no significant change in the software or process of science between the activities of the prototypes

and the actual DUNE detectors, and this is by design. It is anticipated that the current design of the DUNE computing model implemented with the prototypes will scale and be sufficient to achieve the science goals of DUNE. [Section 4.3]

- For data management software DUNE uses the Rucio replica management service which has wide support in the high energy physics community. This is combined with the MetaCat data catalog which tracks the provenance of the data and can also interface to other databases such as the run conditions database. Rucio can work with the FTS system which in turn uses primarily https-based transfers to actually move the data sets to and from remote data sets. It can also use Globus as a transport mechanism and DUNE expects to get data to and from the leadership facilities. There is a distributed global disk model in which current data sets are hosted at a set of 15 storage elements in the US and Europe. [Section 4.3]
- For ICARUS reconstruction processing, computing jobs need to have data delivered at a consistent average rate of up to 2 TB/hr to satisfy demands during data production campaigns (~60 days in length). Data write rates back to Fermilab for storage is ~1 TB/hr. Data need to be delivered to wherever that production is running: SBN expects this to be largely Fermilab computing resources, but it will include some fraction of jobs to available OSG sites across the US and Europe, and to CNAF computing in Italy. For SBND, these numbers will peak at approximately 1 TB/hr for data reading and ~0.8 TB/hr for data write-back in this similar two-year period. [Section 4.4]
- ICARUS simulation will use off-beam data events for base simulation, which will require a data delivery rate of ~2 TB/hr to computing nodes over a production campaign period (~60 days), with a similar rate for data to write back to send back to Fermilab storage. [Section 4.4]
- In addition to use of Fermilab central computing, SBN will onboard computing at CNAF, CERN, and the UK to further facilitate general production and analysis work. SBN also plans to make use of HPC facilities like Polaris at ANL and Perlmutter at NERSC/LBNL to further enable general production processing (e.g., event simulation, signal processing, and event reconstruction at ANL) [Section 4.4]
- For reconstruction processing of data, DUNE needs to have an amount of outbound network traffic from the disk elements equal to the output of the detectors so that processing can keep up with the production of raw data. Data write rates back to disk storage elements from production processing are roughly 1/10 of the raw data rates. Data need to be delivered to wherever that production is running: we expect this to be a combination of computing resources from around the world, but they will be located close to storage elements. DUNE expects data input/output rates to be tape/disk I/O limited for the foreseeable future. [Section 4.4]
- SBN currently uses the Polaris supercomputer facility at Argonne in addition to computing resources at Fermilab. Globus-based transfers of SBND neutrino data to Argonne are up and running, and sustained access to Polaris over the next few years for ongoing data processing will be needed to ensure that physics results can be obtained with maximum efficiency. SBN is also investigating whether other DOE SC user facilities, such as NERSC, may provide unique advantages that could complement the present workflows mentioned above. [Section 4.6]
- DUNE expects to continue to incorporate the use of HPC centers, like Perlmutter at NERSC, for performing simulation and data reconstruction general processing on the 2x2 demonstrator data, as well as for executing dedicated AI/ML processing workflows. [Section 4.6]
- During the DUNE Prototyping phase, prototype detectors running at CERN and Fermilab send their data to DUNE storage sites worldwide. As more and more services come on line at the DUNE Far site–- SURF, data will begin to flow from SURF in South Dakota to Fermilab

and onward to DUNE collaboration sites worldwide for additional processing. When the first far detector Module is operational in mid-2029 supernova data will need to be transferred real-time from SURF to US HPC's for fast turnaround processing. [Section 4.6]

- The primary data-related resource constraint for DUNE deals with the supernova use case; the data from South Dakota must be transmitted to DOE HPC facilities in a matter of hours, and then have it in a place where it can be easily accessed by the job processing software. Once the first FD of DUNE is online, currently estimated to be 2029, supernova triggers are expected regularly, albeit some may be false alarms; the estimate is currently one per month. DUNE will require a set of machines fast enough to source and sink the data, and a distributed file system to make it accessible to the computational jobs. [Section 4.6]
- DUNE relies heavily on Fermilab networking for a variety of tasks, including data archiving, replication, and batch processing. Two geographically diverse 100 Gbps links between the FD complex at SURF and Fermilab are in the process of being acquired to provide WAN connectivity for the FD. The target deployment date for these links is 2027. DUNE is approved for use of LHCONE, and has been making use of LHCONE and LHCOPN at certain grid sites that also have an LHC presence. [Section 4.7]
- DUNE has followed the strategic direction of Fermilab in using cloud only for those new hardware types which DUNE cannot quickly acquired. Examples include tests of inference servers on Triton (e.g., a commercial entity), NERSC Perlmutter as part of the ongoing SONIC project, and GCP. With one of the above tests with compute workers at Fermilab and the inference server at Google cloud in Iowa, DUNE was able to saturate the Fermilab general off-site network bandwidth which at the time was limited to 100 Gbps. [Section 4.8]
- DUNE has several scientists who are actively exploring ML use in reconstruction, analysis. At the moment most of them run locally on Fermilab's Elastic Analysis Facility. [Section 4.8]
- New features for DUNE in the two- to five-year time period related to networking are related to further use of HPC centers, like Polaris at ANL or Perlmutter at NERSC, for performing simulation and data reconstruction general processing, as well as for executing dedicated AI/ML processing workflows. This could amount to a sustained average rate of >1 TB/hr to and from that facility over a production campaign period (~60 days). Similarly, transfers of dedicated data formats for AI/ML processing to SLAC (temporary location) and then the DOE HPC centers (where the workflow is actually run) will also require approximately a rate of 1 TB/hr. [Section 4.8]
- DUNE exhibits all the IRI patterns in design and operation. The full DUNE near detector simulation and reconstruction stack has been run at NERSC Perlmutter, as well as many simulations for the DUNE FD. The various near detector software has also been run at ALCF at least at the proof of principle status. Short-Baseline software has also been run at ALCF. Patterns exhibited include: [Section 4.9]
    - Time-sensitive: The Supernova pointing workflow will be time sensitive and requires the analysis of as much of 320 TB from the first two detector modules in the span of 4 hours or less.
    - Data Integration Intensive: It will be necessary to bring data from the DUNE near and far detectors together into a common analysis format. This is an extremely I/O intensive job.
    - Long-term Campaign: DUNE intends to be taking data for at least two decades, and during that time our computing model calls for reconstructing all the data we have taken to date at least once a year.

6.8.2 Neutrino Experiments at Fermilab: Experiment Profile

DUNE will study the fundamental nature of neutrinos through measurements of neutrino oscillations from both neutrino beam and other natural neutrino sources. In addition to studying neutrino oscillations, DUNE has a broader physics program that includes searches for proton decay, detection of neutrinos from supernova collapse, and other exotic phenomena. The core facilities of DUNE are the near and far detectors, located on-site at Fermilab (Illinois) and in SURF (South Dakota), respectively. The detectors are placed such that they are on-axis with the beam of the future LBNF at Fermilab. The separation of the DUNE near and far detectors means that the experiment is relying on a WAN that originates at SURF in South Dakota, and must transfer all data back to Fermilab: this emphasis on near-constant network connectivity is shaping the choices made for buffering, storage, and analysis at both locations.

The SBN program studies properties of neutrinos and neutrino interactions using beams of accelerator-generated neutrinos at Fermilab. In total three detectors — the SBND, MicroBooNE, and the short baseline far detector (ICARUS T600) — detect neutrinos from Fermilab's BNB, with the major goal of conclusively identifying or ruling out anomalous neutrino oscillations hinted at by many other measurements. The detectors are all Liquid Argon Time Projection Chambers (LArTPCs), which combine great position and calorimetric resolution to enable precise event reconstruction and particle identification. In addition to searching or ruling out 'new physics' in neutrino oscillations, the SBN detectors can measure important properties of neutrino interactions on argon that will be critical for the future long-baseline neutrino experiment DUNE.

6.8.2.1 Science Background

6.8.2.1.1 DUNE Science Background

Transfer of the raw data produced from the DUNE detectors is likely to be the largest single network task for the experiment. The LBNF beam will produce pulses of neutrinos at roughly a 1 Hz repetition rate, but there are additional triggers for data recording beyond beam interactions produced from the detectors. Estimates of those data volumes are presented in Section 6.8.2.4. The data must be transferred from the detectors to archival storage at Fermilab and at least one additional archival location geographically separate from storage at Fermilab. This data must then be processed through production reconstruction software packages before it can be analyzed by collaboration members and utilized for measurements. The distribution of the data to compute facilities and transfer back of the resulting datasets is the next largest network activity. It is anticipated that along with initial processing, there will be annual reprocessing of all data accumulated to date. In addition to beam and calibration data, the signal of neutrinos produced from a nearby supernova burst present a significant challenge due to the large volume of data collected in a very short period of time (up to 600 TB in 100 seconds) and the time-critical nature of analyzing that data for optical astronomical observations.

DUNE has built two prototype cryostats at CERN for testing of final far detector technology designs, ProtoDUNE-HD and ProtoDUNE-VD. These two cryostats are 1:20 scale of the final far detector designs, and provide an opportunity to test both the detector technology but also the offline computing resources and software. ProtoDUNE-HD has just completed a 3 month beam run at CERN, and ProtoDUNE-VD is anticipated to start a roughly yearlong test period in late 2024. The ProtoDUNE-VD test run will comprise both beam and cosmic-ray data taking. These detectors produce significant amounts of data for archival storage at CERN and Fermilab, along with considerable need for distributed computing in order to analyze the data.

In addition to detector data, DUNE will produce large quantities of simulation that in some cases will rival the size of the detector data. This data will be produced at distributed compute facilities and be cataloged and stored in storage elements around the globe. DUNE is currently using more than 20 compute facilities, and more than 10 storage facilities to achieve its computing tasks. It is expected that this number will grow between now and late 2028 when the first far detector module will become operational.

6.8.2.1.2 SBN Science Background

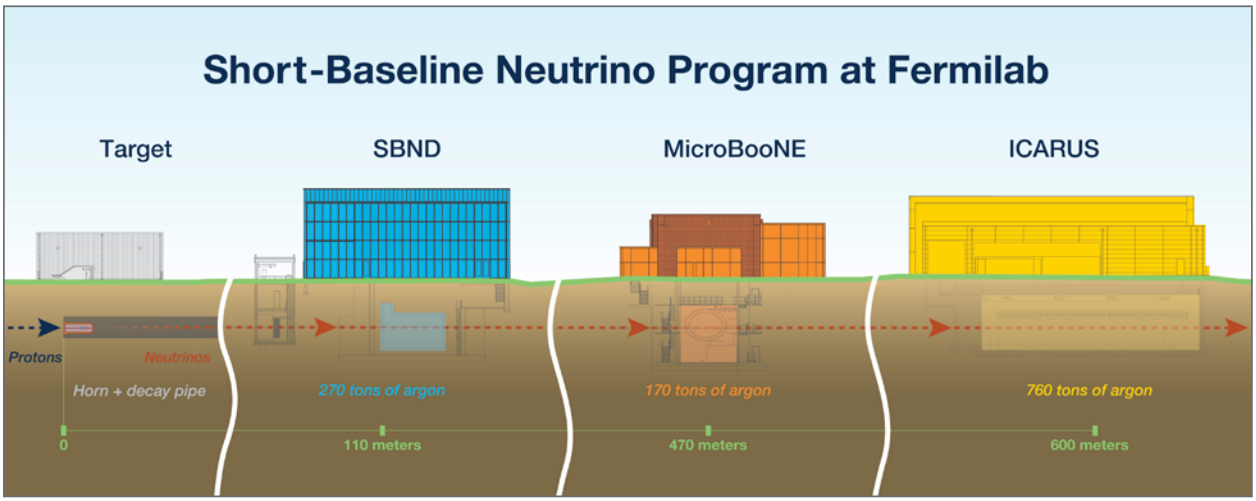


Figure 6.8.1: The layout of the SBN Program detectors along the BNB at Fermilab.

All three detectors are located on Fermilab's campus at distances of a few hundred meters apart, and record interaction events at an approximately 1 Hz rate. Event data are dominated by the data from the LArTPCs themselves: each detector has thousands to tens of thousands of sensing wires that sample and record charge signals over a period of milliseconds for each interaction event. In each detector, the LArTPC data are combined by the data acquisition system (DAQ) with data from other detector systems that record scintillation light signals and the passage of cosmic rays through the detectors, and then written to a local computing cluster at each detector building site. The complete 'raw data' are then transferred to Fermilab's central computing, where it is staged and then written to archival tape.

Raw data must then be processed for analysis. A simple view of the processing chain is as follows: raw data goes through a signal processing and hit-finding stage to filter out noise, deconvolve electronics and electric field response, and identify areas of detected charge; then, higher-level reconstruction algorithms perform pattern recognition and image analysis combining data across all views of the detector to reconstruct 3D charged particle tracks and electromagnetic showers and perform calorimetric energy measurements; and, finally, those higher-level objects are used in analysis to identify neutrino interactions and reconstruct their topologies and final-state particles.

Collected data are typically compared to simulations to assess reconstruction efficiencies, systematic uncertainties, and ultimately to test predicted models of new physics or of neutrino interactions. Simulated data goes through the same processing chain as above, but first must go through a generation process to produce equivalent raw data. Neutrino interactions are modeled using calculations available in software packages, particle interactions in matter are then modeled using GEANT4[96], and then detector modeling is applied through custom-written software. Because each detector is located near the surface of the earth, cosmic-ray backgrounds are a constant presence: they can be simulated through software, but for final analyses the detectors will take data events taken out-of-time with the beam which offers a random snapshot of activity induced by cosmic rays in the detector, and then neutrino interaction simulations are 'overlaid' on top of this raw data. A library of 'out-of-time' data events is collected to allow for this overlay simulation.

The SBN[97] detectors detect millions of neutrino interactions, collect many more events that are simply cosmic-ray-induced background interactions, and also need to simulate tens of millions of neutrino interactions to achieve its physics goals. Typically, simulation and the two stages of reconstruction (signal processing and

96 https://geant4.web.cern.ch/
97 https://arxiv.org/abs/1503.01520

3D reconstruction) are run centrally by the experiment data production teams, and produce files containing 3D reconstruction information with a size on the order of 50 kB per event. These files are stored centrally at Fermilab, then accessed by analyzers across the collaboration to perform the final physics measurements. Alternative workflows for 3D reconstruction based on AI/ML methods run at HPC centers are being integrated in the SBN operations.

As improvements in the simulation and reconstruction are always ongoing, the SBN experiments plan a roughly yearly data production cycle. Raw data are collected from the detectors throughout the year and stored permanently, but derived data will be reproduced each year. Simulation and signal-processing outputs will be produced roughly once per year, while 3D reconstruction outputs will be produced roughly twice per year (using the previous stages as inputs). The output of 3D reconstruction will then be accessed through the year for the performance of physics analysis. SBN currently has a data lifetime model of saving the bulk of derived data for two years, and the much smaller analysis data used for publications permanently.

6.8.2.2 Collaborators

6.8.2.2.1 DUNE Collaborators

The DUNE Collaboration comprises approximately 1500 scientists from more than 200 institutions in 39 countries around the world. These institutions include both national laboratories (Fermilab, CERN, BNL, LBNL, IN2P3, CNAF, NIKHEF, RAL, etc.) and universities. The center of activity is Fermilab, but there are clusters of both scientists and resources located at national labs that are being utilized for both compute, storage, and collaboration. Some examples of these clusters are located across the US, CERN, France, Brazil, the UK, Spain, and Italy.

Currently, there are two major endpoints for raw data archival storage (tape): Fermilab and CERN. There are currently test instances for archival storage in the UK and France that are expected to ramp up capacity as activity at the near and far detector increases. DUNE extensively uses disk storage facilities as part of the distributed computing model to make sure that compute facilities are efficiently utilized. For disk storage, DUNE currently has close to 30 active facilities spread across the US, the UK, France, the Netherlands, Spain, CERN, Czechia, and India. The addition of more disk storage endpoints is expected as DUNE expands its compute resources in the next 5–10 years.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
FAR DETECTOR	N	Create/transfer data to Fermilab	30 PB/year	Continuous	N	Network reliability and bandwidth for supernova events
NEAR DETECTOR	N	Create/transfer data to Fermilab	5 PB/year	Continuous	N	Network reliability
FERMILAB STORAGE	Primary copy from 1 & 2, secondary copy of ProtoDUNE data, some production output	Transfer data to secondary storage, ingest data from ProtoDUNE, ingest data from production processing	60 PB/year	Continuous	Y	

CERN	Primary copy of ProtoDUNE data, secondary copy of some FD/ND data	Transfer data to secondary storage, ingest data from FD/ND, ingest data from production processing	10 PB/year	Continuous	Y	
PROTODUNE-HD/VD	N	Create/transfer data to CERN	5 PB/year	Continuous	N	
NERSC	N	Transfer in raw ND data, transfer out production output	6 PB/year	Continuous	Y	Ingress and egress to HPC facilities requires staging instead of streaming
HPC FACILITIES	N	Transfer in/out of facilities	1 PB/year	2-3 times per year (uncertain)	Y	
NON-US ARCHIVAL STORAGE	Y	Transfer from fermilab/transfer out to compute	30 PB/year	Continuous	Y	
US DISK STORAGE	N	Transfer from fermilab/transfer out to compute	5 PB/year	Continuous	Y	
NON-US DISK STORAGE	N	Transfer from non-us archive/transfer out to compute	5 PB/year	Continuous	Y	

Table 6.8.1: DUNE Data Mobility

6.8.2.2.2 SBN Collaborators

The SBN international collaborations comprise of approximately 250 scientists from more than 50 institutions, and include members from across the US, CERN, Italy, the UK, Switzerland, Spain, Turkey, and Brazil. Member institutions critically include both scientific laboratories and universities. While the center of activity is typically at Fermilab, where many collaborators (particularly young researchers) locate themselves for some period of time, analysis and data activities span across these institutions, with significant/large clusters of researchers and activity centered around the UK, CERN, and Italy.

Typically, the collaborations have relied on the computing infrastructure at the host lab (Fermilab) for data storage and the bulk of computing. Dedicated backups of raw data are planned to be sent and stored at the INFN-CNAF facility in Bologna, Italy, and investigations on a possible additional back-up location will be underway soon. Computing resources at other institutions have been and will continue to be added to be used in central processing much like sites are added to the OSG: those sites will be used to augment computing resources as much as possible, and will likely be used for dedicated purposes on top of general computing: for example, the generation of dedicated or specialized datasets.

The ICARUS data backups to CNAF are about 1 PB per year, collected and sent roughly continuously during beam data-taking periods (which is generally consistently over 9 months of the year). Similarly, SBND data will be transferred to Polaris at ANL. Transfers will be managed via Rucio or Globus subscriptions, and be validated through checksums. While data should be transferred in a timely fashion (to avoid a pileup of data needing to be backed up), there is not currently a strict latency requirement.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
ICARUS (SBN FAR DETECTOR) DETECTOR	No	Create/transfer to fermilab storage	~2 PB/year	Continuous	No	
SBND (SBN NEAR DETECTOR)	No	Create/transfer to and from Fermilab	~ 3 PB/year	Continuous	No	
FERMILAB DATA CENTER	Yes	Ingest from 1–2 and streaming transfer out for prompt processing on Fermigrid/OSG/WLCG	Up to 5 PB/year (most done on Fermigrid)	Continuous	Yes	Ingest
FERMILAB DATA CENTER	Yes	Transfer out to secondary tape/disk archives (e.g., INFN-CNAF)	~2 PB/year	Continuous	No	Egress from fermilab
FERMILAB DATA CENTER / US HPC SITES	No	Transfer to/from HPC facilities	Up to 5 PB (still under development)	~2–4 Times per year	Yes	
FERMILAB DATA CENTER OSG/ COMPUTING SITES	No	Transfer to/from OSG sites in large-scale production	~2.5 PB	~Twice yearly	Yes	
COLLABORATING INSTITUTIONS	No	Generally small datasets received data streaming/transfer, largely for analysis development work	Up to ~4 PB/ year, roughly ~1–100 GB at a time	Ad hoc	Yes	

Table 6.8.2: SBN Data Mobility

6.8.2.3 IRI Readiness

6.8.2.3.1 DUNE IRI Readiness

The full DUNE near detector simulation and reconstruction stack has been run at NERSC Perlmutter, as well as many simulations for the DUNE FD. The various near detector software has also been run at ALCF at least at the proof of principle status. Short-Baseline software has also been run at ALCF. These are workflows that make extensive use of GPU and thus DUNE will need to make use of multiple DOE SC resources.

Time-sensitive: The supernova pointing workflow, described in more detail below, will be very time sensitive, requiring the analysis of as much of 320 TB from the first two detector modules in the span of 4 hours or less. Thus, DUNE is planning for robust network links from the DUNE FD to the DOE supercomputing centers.

Data Integration-Intensive: It will be necessary to bring data from the DUNE near and far detectors together into a common analysis format. This is an extremely I/O intensive job.

Long-term Campaign: DUNE intends to be taking data for at least two decades and during that time the computing model calls for reconstructing all the data taken to date at least once a year.

6.8.2.3.2 SBN IRI Readiness

SBN currently uses the Polaris supercomputer facility at Argonne in addition to computing resources at Fermilab. Globus-based transfers of SBND neutrino data to Argonne are up and running, and sustained access to Polaris

over the next few years for ongoing data processing (i.e., a "long-term campaign") will be needed to ensure that physics results can be obtained with maximum efficiency. GPU-accelerated machine-learning algorithms are actively being developed for SBN physics analysis, and the availability of both GPUs and CPUs at Polaris makes it an ideal site for related processing to take place.

SBN is also investigating whether other DOE SC user facilities, such as NERSC, may provide unique advantages that could complement the present workflows mentioned above.

6.8.2.4 Instruments and Facilities

6.8.2.4.1 DUNE Instruments and Facilities

Assets that generate data

DUNE consists of several related components located across the US and in Europe.

Offline data processing is done worldwide, including multiple sites in Europe and sites in India and Brazil.

Present: Prototypes - Running now

Prototypes located at Fermilab, CERN and collaborating institutes. These are currently running. A 700-ton ProtoDUNE-HD detector running at CERN has already generated close to 5 PB of data in 2024 and DUNE will perform another run of a different prototype of similar scale in 2025-2026.

Processing of these data is done worldwide using conventional grid resources. File sizes are kept between 1 and 8 GB and there are several million such files.

A smaller prototype of a novel pixel detector (the 2x2) is running at Fermilab and generated ~ 15 TB in its first 1 week run. It will run again in 2025.

Up to 2029: Analysis of Prototype data and start of DUNE commissioning and calibration

On the two- to five-year time scale, DUNE will be processing and analyzing the data from the prototypes.

The first full DUNE far detector, FD-VD, is slated to begin full operation at SURF in Lead, SD in mid-2029 with significant testing of DAQ and data-movement systems starting one to two years earlier. A second module will come online in 2030-31. These systems will require significant testing and calibration and are projected to produce 10–15 PB of data per year per module. This includes tests of supernova burst sensitivity that will produce around 320 TB over a few hours (they are actually produced over 100s but are constrained by local network speeds). Early calibration runs will also produce significant data volumes over short time periods. Steady rates of 100TB/day are anticipated. Files will be kept to 1–8 GB that implies ~100,000 files produced in a supernova candidate or large calibration run.

Longer term 2029–2040

In 2030-31 the second far detector module and the near detectors located at Fermilab will come on board. Near detector data volumes are expected to be in the few PB/year range, while DUNE will plan to cap the total raw data from the far detectors at 30 PB/year. At this time, the plan is to store data in Europe and in the US near processing resources.

The near detector groups are already developing code optimized for reconstruction and simulation using GPU/accelerators and will likely make significant use of national class compute facilities. Efforts are also underway to optimize algorithms to process FD data using parallel compute facilities.

	CERN	Fermilab	SURF	Processing	Total
NEAR TERM	ProtoDUNE	2x2 prototype		Worldwide	
	10 PB/yr	1 PB/yr		10 PB/yr	20 PB/Yr
MID TERM		prototypes	FD commissioning		
		1 PB/yr	30 PB/year	5 PB/yr	35 PB/Yr
LONG TERM		Near Detector	Far Detector		
		5 PB/yr	30 PB/yr	10 PB/yr	45 PB/yr

Table 6.8.3: Summary of yearly data production from DUNE and prototypes. This does not include multiple copies or transfer buffers.

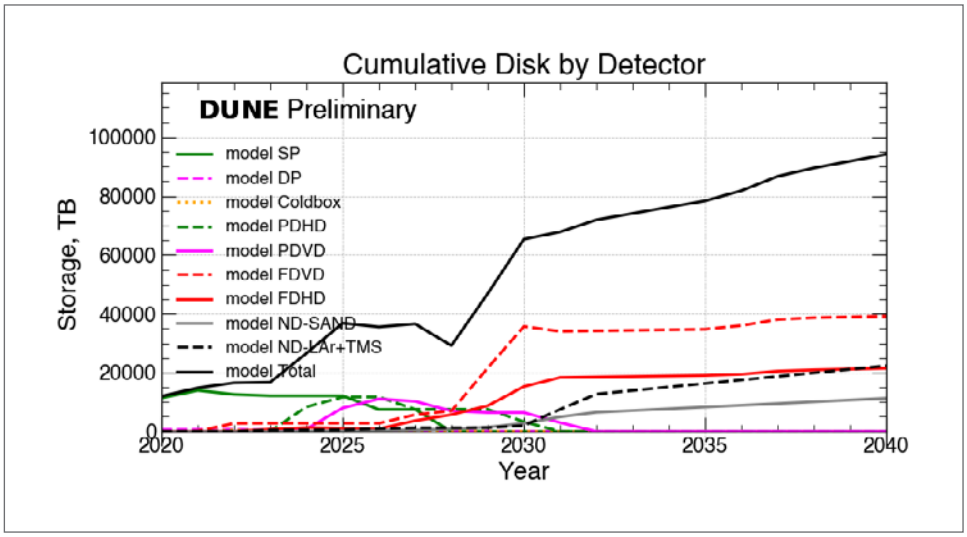


Figure 6.8.2: Estimates for total disk needs by detector until 2040. This assumes that data samples have one to two copies and lifetimes on disk of one to three years.

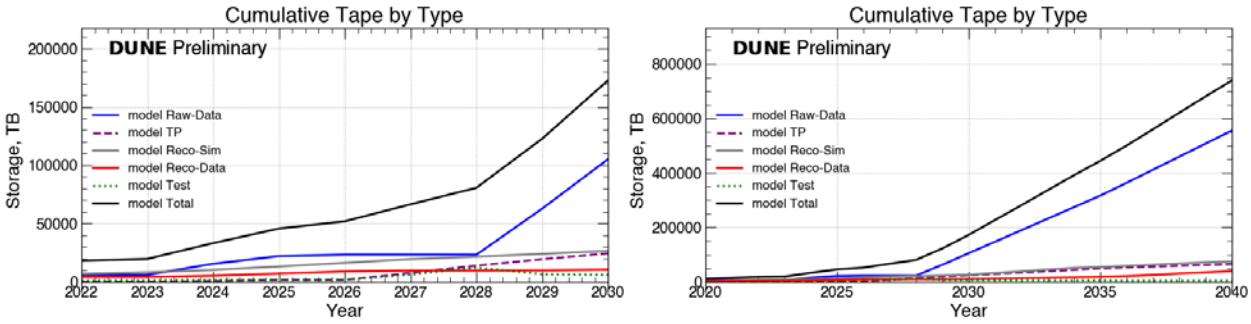


Figure 6.8.3: Estimate for total archival tape storage needed until 2030 and until 2040. After 2030, the data volumes increase by 80 PB/year, mainly driven by the need to store two copies of raw data.

6.8.2.4.2 SBN Instruments and Facilities

Present-2 years:

The MicroBooNE detector took data from 2015 until 2020. The data are stored centrally at Fermilab and is actively being analyzed by the Collaboration. Further processing of the data is being planned for the upcoming years, and will support the legacy results of the experiment.

The ICARUS detector has been taking data for three years. The raw data rate from the detector is 2 PB. Raw data are stored at Fermilab, with a backup for a significant portion of ICARUS data stored at CNAF. For large-scale processing, Fermilab computing and OSG sites will be used.

The SBND detector is completing the commissioning phase and is expected to take physics-quality data in the fall. The current expected raw data rate from the detector is 3 PB. Data management and processing will follow similar strategies as ICARUS, complementing the primary Fermilab resources with additional off-site data centers.

Two to five years:

The data taking period will be completed within this time frame, with an expected end of operations at the end of FY27. DUNE expects significant activities related to storage and processing of the data, as well as simulation. In addition to use of Fermilab central computing, SBN will onboard computing at CNAF, CERN, and the UK to further facilitate general production and analysis work. SBN also plans to make use of HPC facilities like Polaris at ANL and Perlmutter at NERSC/LBNL to further enable general production processing (e.g., event simulation, signal processing, and event reconstruction at ANL) and final-fits/sensitivity estimation (e.g., Feldman-Cousins analysis at NERSC). There is no major change expected in the detector infrastructure or basic computing envisioned, though DUNE hopes to continue to evolve the ability to rely less on tape for data storage, and have a greater availability of disk for data that has an expected lifetime less than a few years (e.g., much of the bulky derived data from simulation and first-stage reconstruction output).

Beyond five years:

The SBN program will be finished with data taking at this timeline, and will likely be in the final stages of physics exploitation. This will hopefully reduce significant computing load for taking new data, but larger scale production and analysis are likely to continue for some time in this period. It is very unlikely that significant changes in data sources or workflows will occur in this time.

6.8.2.5 Generalized Process of Science

6.8.2.5.1 DUNE Process of Science

Data is generated either from the detector ("raw data") or is simulated. The former is managed by the online DAQ and data management team in the collaboration, while the latter is managed by the offline production team. Data then proceeds through two stages of reconstruction (signal processing, and 3D reconstruction), managed by the offline production team. The output of the reconstruction includes slimmed 'analysis tuples' that will be read by analyzers across the collaboration.

Data reduction is largely accomplished through the reconstruction processing. After signal processing, algorithms to identify regions of interest are employed to reduce the search region for hit reconstruction in waveforms. For example, DUNE data are expected to be ~8 GB per event at the raw data stage, but then only 20 MB / event after first stage reconstruction. The second stage of reconstruction will increase that size by approximately a factor of 2 (largely to hold detailed charged particle track trajectories), but then high-level reconstruction objects can be summarized to then create a final analysis tuple with the goal of a size of ~20 kB per event.

Present to two years:

There are at least three prototype detectors that will be active during the next two years, ProtoDUNE-HD, ProtoDUNE-VD, and ND-lAr 2x2 demonstrator. The ProtoDUNE detectors are located at the Neutrino Platform in the North area of CERN, while the 2x2 demonstrator is located in the main injector neutrino oscillation search (MINOS) hall at Fermilab. The raw data arrive from the ProtoDUNE detectors at a steady rate of about 1 TB/hr (~300 MB/s), and from the 2x2 detector at about 42 GB/hr (~15 MB/s). The ProtoDUNE data are delivered to CERN Exabyte Object Storage (EOS) (disk) and Cloud Tiering Appliance (CTA) (tape) on-site, and a second copy of the ProtoDUNE data are also transferred to Fermilab for archival storage. The raw data are made

available for immediate processing which is accessed via xRootD to run on OSG and WLCG compute nodes. For immediate processing, data files are still in disk cache and so do not need to be accessed from tape. For the 2x2 demonstrator, the data are automatically transferred to Fermilab Enstore (tape) on-site, and then transferred in batches to a DUNE allocation at NERSC for processing on HPC.

For reconstruction processing of data, DUNE needs to have an amount of outbound network traffic from the disk elements equal to the output of the detectors so that processing can keep up with the production of raw data. Data write rates back to disk storage elements from production processing are roughly 1/10 of the raw data rates. Data need to be delivered to wherever that production is running: DUNE expects this to be a combination of computing resources from around the world, but they will be located close to storage elements.

While the collaboration pushes for most analysis to use the slimmed analysis tuples (hosted at Fermilab, but small and able to be copied easily elsewhere), there will be a need for calibration and event selection tasks to access reconstruction datasets. These will be several 100 TB datasets staged from central storage locations and delivered to computing nodes at Fermilab or OSG sites.

New features in this period related to networking are related to further use of HPC centers, like Perlmutter at NERSC, for performing simulation and data reconstruction general processing on the 2x2 demonstrator data,, as well as for executing dedicated AI/ML processing workflows.

Two to five years:

Generally, the situation for the present to two years is repeated for the two- to five-year case, with similar operations of the prototype detectors during that time. The first far detector module is expected to operate in late 2028 and so is included in the final section.

General needs for production processing will increase year-to-year in a generally stable way: data reconstruction and simulation (and its corresponding reconstruction) scale with the total amount of data taken/neutrino interactions collected. DUNE expects data input/output rates to not be network bound, but rather tape/disk I/O limited, and so the peak rates above should not increase by more than a factor of 2, and can be offset by longer production campaign timelines (e.g., moving from 60 to 90 days)

In this period, slimmed analysis tuples should be almost exclusively the files that most collaboration data analyzers will access (necessary to keep total data I/O rates under control).

Beyond five years:

At this time period, data acquisition for the first far detector module will commence and the output of the module will be roughly 30 PB/year going to archival storage. It is anticipated that the offline processing and data flows will look very similar to prototype operations but with slightly different endpoints. (e.g., instead of data coming from CERN, data will be created at SURF). There is no significant change in the software or process of science between the activities of the prototypes and the actual DUNE detectors, and this is by design. It is anticipated that the current design of the DUNE computing model[98] implemented with the prototypes will scale and be sufficient to achieve the science goals of DUNE. For details of the updated rates during DUNE operations, please see Section 6.8.2.4.1

6.8.2.5.2 SBN Process of Science

Data is generated either from the detector ("raw data") or is simulated. The former is managed by the online DAQ and data management team in the collaboration, while the latter is managed by the offline production team. Data then proceeds through two stages of reconstruction (signal processing, and 3D reconstruction), managed by the offline production team. The output of the reconstruction includes slimmed 'analysis tuples' that will be read by analyzers across the collaboration.

98 https://lss.fnal.gov/archive/design/fermilab-design-2022-01.pdf

Data reduction is largely accomplished through the reconstruction processing. After signal processing, algorithms to identify regions of interest are employed to reduce the search region for hit reconstruction in waveforms. For example, ICARUS data are expected to be ~100 MB per event at the raw data stage, but then only 40 MB / event after first stage reconstruction. The second stage of reconstruction will increase that size by approximately a factor of 2 (largely to hold detailed charged particle track trajectories), but then high-level reconstruction objects can be summarized to then create a final analysis tuple with the goal of a size of ~50 kB per event.

Present-2 years:

For normal data taking from the detector, raw data arrive from ICARUS at a steady rate of about 370 GB/hr (~100 MB/s), and from SBND at about 320 GB/hr (~90 MB/s). These data are delivered to Fermilab dCache and Enstore tape storage on-site, and is made available for immediate processing which, for the most part, is accessed data via xRootD to run on-site at Fermigrid, but can be accessed and run on OSG sites as available. For immediate processing, data files are still in disk cache and so do not need to be accessed from tape. Raw data backups are ~75 GB/hr (~20 MB/s) to CNAF for ICARUS, and 130 GB/hr (~35 MB/s) for SBND (Polaris at ANL).

For reconstruction processing of data, ICARUS computing jobs need to have data delivered at a consistent average rate of up to 2 TB/hr to satisfy demands during data production campaigns (~60 days in length). Data write rates back to Fermilab for storage is ~1 TB/hr. Data need to be delivered to wherever that production is running: SBN expects this to be largely Fermilab computing resources, but it will include some fraction of jobs to available OSG sites across the US and Europe, and to CNAF computing in Italy. For SBND, these numbers will peak at approximately 1 TB/hr for data reading and ~0.8 TB/hr for data write-back in this similar two-year period.

In this two-year period, ICARUS simulation should begin using off-beam data events for its base simulation (see Section 6.8.2.1.2), which will require a data delivery rate of ~2 TB/hr to computing nodes over a production campaign period (~60 days), with a similar rate for data to write back to send back to Fermilab storage. SBND may not be doing these 'overlay' simulations as a default at this point, and so would see a much reduced need for reading data when starting simulation generation; however, data write-back rates will still be large, approximately 1.25 TB/hr during a production campaign period.

ICARUS and SBND production campaigns overlap with each other in time, and typically include data reconstruction processing alongside simulation. Thus, total data delivery rates coming out from Fermilab data storage at a sustained rate of 5 TB/hr over a period of 60 days (if possible), and with data back to data storage at a rate of ~3.8 TB/hr over a similar period.

While the collaboration pushes for most analysis to use the slimmed analysis tuples (hosted at Fermilab, but small and able to be copied easily elsewhere), there will be a need for calibration and event selection tasks to access reconstruction datasets. These will be several 100 TB datasets staged from central storage locations and delivered to computing nodes at Fermilab or OSG sites.

New features in this period related to networking are related to further use of HPC centers, like Polaris at ANL or Perlmutter at NERSC, for performing simulation and data reconstruction general processing, as well as for executing dedicated AI/ML processing workflows. Significant fractions of the SBND data inputs for simulation and reconstruction could be sent to Polaris for more efficient processing. Following the numbers above, this could amount to a sustained average rate of >1 TB/hr to and from that facility over a production campaign period (~60 days). Similarly, transfers of dedicated data formats for AI/ML processing to SLAC (temporary location) and then the HPC centers (where the workflow is actually run) will also require approximately a rate of 1 TB/hr.

Two to five years:

Generally, the situation for the present to two years is repeated for the two- to five-year case. Detector operations/ data acquisition will continue at the steady rates presented there until the end of FY27.

General needs for production processing will increase year-to-year in a generally stable way: data reconstruction and simulation (and its corresponding reconstruction) scale with the total amount of data taken/neutrino interactions collected. SBN expects data input/output rates to not be network bound, but rather tape/disk I/O limited, and so the peak rates above should not increase by more than a factor of 2, and can be offset by longer production campaign timelines (e.g., moving from 60 to 90 days)

In this period, slimmed analysis tuples should be almost exclusively the files that most collaboration data analyzers will access (necessary to keep total data I/O rates under control).

Beyond five years:

At this time period, data acquisition will be over, and this time period will see general production processing slow down as the reconstruction and simulation reaches its final state. Most activity will be centered around analysis, which will depend on many analyzers accessing small analysis tuples.

6.8.2.6 Remote Science Activities

The DUNE computing infrastructure is inherently distributed. Data centers are classified as Computing Site (High Throughput Computing (HTC) or HPC), Data site (disks), Archive site (tapes). DUNE uses data centers spread globally. Some DUNE sites only provide compute, others compute and storage and a few provide compute, disk and tape to DUNE. During the DUNE Prototyping phase, prototype detectors running at CERN and Fermilab send their data to DUNE storage sites worldwide. As more and more services come on line at the DUNE Far site - SURF, data will begin to flow from SURF in South Dakota to Fermilab and onward to DUNE collaboration sites worldwide for additional processing. When the first far detector Module is operational in mid-2029 supernova data will need to be transferred real-time from SURF to US HPC's for fast turnaround processing.

In 2030-31, the second far detector module will come online along with the near detectors located at Fermilab, their data will be sent worldwide for processing. A copy of all processed data will come back to Fermilab for archival storage. The near detector data will be sent to national class compute facilities for processing using the GPU/accelerators already there.

6.8.2.7 Software Infrastructure

6.8.2.7.1 DUNE Software Infrastructure

For data management software DUNE uses the Rucio replica management service which has wide support in the high energy physics community. This is combined with the MetaCat data catalog which tracks the provenance of the data and can also interface to other databases such as the run conditions database. Rucio can work with the FTS system which in turn uses primarily https-based transfers to actually move the data sets to and from remote data sets. It can also use Globus as a transport mechanism and DUNE expects to get data to and from the leadership facilities. There is a distributed global disk model in which current data sets are hosted at a set of 15 storage elements in the US and Europe.

The task of moving raw data into final and intermediate formats is accomplished by the DUNE offline software, for which a new framework is currently being written.

Present-2 years: DUNE continues using Rucio and Metacat for data management and movement, and the legacy framework (ART) for data reduction

Two to five years: DUNE continues using Rucio and Metacat, adding the Globus transport, and the new offline framework

Beyond five years: DUNE will continue using Rucio and Metacat, adding the Globus transport, and the new offline framework

6.8.2.7.2 SBN Software Infrastructure

SBN experiments use much of the resources offered through FIFE at Fermilab. SBN largely follows the software infrastructure being developed for general experiments and for DUNE (including SAM, Rucio, CVMFS, xRootD, dCache).

6.8.2.8 Network and Data Architecture

The network and data architecture supporting DUNE and SBN can be found in Section 7.4.

6.8.2.8.1 DUNE Network and Data Architecture

DUNE relies heavily on Fermilab networking for a variety of tasks, including data archiving, replication, and batch processing. Two geographically diverse 100 Gbps links between the FD complex at SURF and Fermilab are in the process of being acquired to provide WAN connectivity for the FD. The target deployment date for these links is 2027. DUNE is approved for use of LHCONE, and has been making use of LHCONE and LHCOPN at certain grid sites that also have an LHC presence.

6.8.2.8.2 SBN Network and Data Architecture

SBN has no additional information to provide for this section.

6.8.2.9 Emerging Technology Use

6.8.2.9.1 Cloud Services

DUNE has used minimal cloud services, mostly in test of inference servers such as Triton in which the bulk of the computing is done on commodity resources but the GPU, FPGA or tensor unit is located in the cloud, and is accessed by a call-out over the network. DUNE has for the most part followed the strategic direction of Fermilab in using cloud only for those new hardware types which DUNE cannot quickly acquire. Even for this application, tests of running the inference server on NERSC Perlmutter as part of the ongoing Services for Optimized Network Inference on Coprocessors (SONIC) project are being performed.

With one of the above tests with compute workers at Fermilab and the inference server at Google cloud in Iowa, DUNE was able to saturate the Fermilab general off-site network bandwidth which at the time was 100 gBit/s.

DUNE expects to continue to have just a minimal or token use of the cloud going forward in the above three time frames.

6.8.2.9.2 Artificial Intelligence and ML

6.8.2.9.2.1 DUNE AI & ML

DUNE has several scientists who are actively exploring ML use in reconstruction, analysis. At the moment most of them run locally on Fermilab's Elastic Analysis Facility.

6.8.2.9.2.2 SBN AI & ML

SBN collaborators developed an end-to-end AI/ML data reconstruction workflow, currently outperforming more traditional reconstruction approaches [arXiv:2102.01033]. This workflow is based on data presented in an image format, after signal-processing steps. The input data are stored in "larcv" files [arXiv:2209.04023], which take 35 MB/event in ICARUS. These files are transferred to SLAC directly from the OSG or Fermigrid worker nodes

using Rucio. Once stored at SLAC, they are further transferred to HPC centers where they are processed using the massive GPU resources available. The choice of HPC centers is driven by the allocations obtained by the PI leading the effort, or by the Collaboration as a whole. Results of the workflow are stored in hdf5 files, typically of size 1 MB/event. These hdf5 files are transferred back to Fermilab where their content is merged into the analysis data format files which are already used by the Collaboration. These data transfers are included in the rate estimates provided in Section 6.8.2.5.2.

6.8.2.9.3 Wireless or Distributed Location Science Activities

No specific needs for increased wireless capabilities for SBN or DUNE.

6.8.2.10 Data-Related Resource Constraints

6.8.2.10.1 DUNE Data-Related Resource Constraints

DUNE has run significant amounts of simulation of the far detector at NERSC. The analysis and simulation for the DUNE near detector Liquid Argon 2x2 prototype as well as for the full near detector at the moment requires GPU capabilities which are currently supplied at NERSC Perlmutter. This MLReco code has also been run at ANL Polaris.

The main data-related constraint deals with the supernova use case in which DUNE must get the data from South Dakota to one or the other of the computing centers in a matter of hours, and then have it in a place where it can be easily accessed by the jobs.

For the present two years DUNE will only be doing tests with data of that size and shape.

Once the first FD of DUNE is full of liquid argon and functional, currently estimated to be 2029 (five years from now) then DUNE will be having supernova triggers to deal with regularly, albeit mostly false alarms. Estimate is currently one per month.

By that time DUNE will need a set of machines fast enough to source and sink the data that fast, and a distributed file system to make it accessible to the workers.

6.8.2.10.2 SBN Data-Related Resource Constraints

SBN's major concern, based on MicroBooNE production experiences, is the performance of data storage I/O, rather than networking. SBN currently relies on tape storage for the bulk of data storage needs, and staging back and forth from tape can generally be the main obstacle for data production workflows. SBN is working to update the computing model to better handle this: if so, this may change some calculations presented in previous sections and put more strain on networking needs. SBN and Fermilab computing experts will need to keep a close eye on how modifications in production workflows have impacts on the whole system.

6.8.2.11 Data Mobility Endpoints

6.8.2.11.1 DUNE Data Mobility Endpoints

Fermilab currently has a proof of principle Globus end point set up for SBN program. (if-globusdtn-01.fnal.gov) Planning for a more permanent and scalable Globus end point in future. The main DUNE data are accessed via dCache, a.k.a. fndcadoor.fnal.gov. Fermilab is the archival point for all DUNE data, i.e., at least one copy ends up on tape there for all raw and processed data.

6.8.2.11.2 SBN Data Mobility Endpoints

A Globus endpoint has been established for transferring SBN files from Fermilab to Argonne for processing on Polaris. At present, the endpoint is only used by SBND, but it will likely be used by ICARUS as well in the future. The endpoint's name is currently "Fermilab-SBND" with Unique Identifier (UID) d868b0a3-3883-44ff-bd0e-7e1d2110d6dd. Technical details about how the setup compares with others are available upon request.

6.8.2.12 Outstanding Issues

SBN and DUNE do not have any to report at this time.

6.8.2.13 Facility Profile Contributors

SBN & DUNE Representation

Michael Kirby, BNL, mkirby@bnl.gov

Steve Timm, Fermilab, timm@fnal.gov

Heidi Schellman, Oregon State University, heidi.schellman@oregonstate.edu

Ken Herner, Fermilab, kherner@fnal.gov

Doug Benjamin, BNL, dbenjamin@bnl.gov

Ryan M Heath, Fermilab, rheath@fnal.gov

Giuseppe Cerati, Fermilab, cerati@fnal.gov

Steven Gardiner, Fermilab, gardiner@fnal.gov

ESCC Representation

Phil DeMar, Fermilab, demar@fnal.gov

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

6.9 LHC: ATLAS Experiment

The LHC at CERN[99] is the most powerful particle accelerator in the world. Highly energetic protons, traveling almost at the speed of light around a 27 kilometer long ring in both directions, are steered to collide head-on, creating new particles and new interactions to probe fundamental natural laws.

6.9.1 Discussion Summary

The Data Challenges (the last being DC24) were incredibly useful for the ATLAS experiment, as well as the entire LHC community. A number of issues were identified, including with the FTS tool, which was limiting productivity because it was not aware of throughput capabilities. Simply stated, poor performance of some transfers limited the deployment of more to better-connected sites. Additional work to move from VOMS, which is based on x.509 certificates, to HTTP-based tokens is also underway. FTS (and Rucio) are developed at CERN, and have limited resources to implement fixes. The timing of the next DC is currently being discussed, and tentatively is scheduled for 2027.

ATLAS data volumes can average 5PB a day, with volumes exceeding this during busy periods when the LHC is in full operation. The Tier-1 and Tier-2s within the United States can see data flows of tens to hundreds of Gbps on a routine basis. Most are connected via 100 Gbps or 400 Gbps diverse connections to ESnet, or regional networking providers. BNL (the US ATLAS Tier-1) sees a majority of this traffic coming directly from CERN), then distributing outward toward other Tier-2s at US Universities.

ESnet will investigate ways to integrate resources into LHCONE at non-LHC-affiliated sites. In particular DOE HPC centers (e.g., NERSC) and aFs (e.g., SLAC) could benefit from being connected to better identify and prioritize traffic.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The LHC at CERN is the most powerful particle accelerator in the world. Highly energetic protons, traveling almost at the speed of light around a 27 kilometer long ring in both directions, are steered to collide head-on, creating new particles and new interactions to probe fundamental natural laws. [Section 4.1]
- ATLAS is the largest of four particle detectors that measure and record the particle collisions at the LHC. The primary scientific goal is to quantitatively measure and discover properties of the SM of particle physics. [Section 4.1]
- ATLAS is a global collaboration of physicists, engineers, technicians, students and support staff from around the world. It is one of the largest collaborative efforts ever attempted in science, with over 5500 members and almost 3000 scientific authors. [Section 4.1]
- The LHC collides protons more than a billion times every second, out of which the detector experiments select interesting collisions for recording at a rate of a thousand times every second. With over two thousand hours of data collection every year when the LHC is running, a huge data sample for physicists to analyze from the completed Runs (1 and 2) is already available. Run 3 started data collection in 2022 and is planned to continue until the end of 2025. Run 4 (High Luminosity) is anticipated in 2030. [Section 4.1]
- ATLAS is designed to observe up to 1.7 billion proton-proton collisions per second, with a combined data volume of more than 60 million megabytes per second (480 Tbps). However, only some of these events will contain interesting characteristics that might lead to new discoveries. [Section 4.2]

99 https://home.cern/

- Datasets in ATLAS are collections of files organized by category/workflow. These datasets are the fundamental units in ATLAS, fully integrated into the Rucio DMS, and used for workflow, transfers and user analysis. Individual datasets vary largely in size: [Section 4.2]
  - raw datasets are in the range of 1 TB to about 250 TB
  - AOD datasets are in the range of 1 GB to about 100 TB
  - DAOD datasets are in the range of 1 GB to about several TB
  - HITS datasets are in the order of several TBs
- The ATLAS grid infrastructure consists of the Tier-0 computing site at CERN, 9 Tier-1's, 70 Tier-2's and about 50 Tier-3 sites distributed worldwide. All workflows are executed at all Tiers with MOUs: the Tier-0, Tier-1 and Tier-2 sites. Tape storage to store raw and AOD files is available at the Tier-0 and Tier-1 sites and a small subset of Tier-2 sites. [Section 4.3]
- ATLAS Tier-3 sites at most universities are usually smaller local CPU clusters with several TBs of disk storage attached. At these sites individual end-user analysis happens without any production activity. [Section 4.3]
- Data from the ATLAS detector is calibrated, reconstructed and automatically distributed all around the world by the ATLAS DMS. The ATLAS Production System then filters through these events and selects the ones needed for a particular type of analysis. This brings the data set down to a manageable size for someone doing an analysis on their laptop. All members of the ATLAS Collaboration have equal access possibilities to all ATLAS data, independently of their geographical location, thanks to the WLCG. [Section 4.3]
- The open-source software framework Rucio is used to organize, manage, and access the ATLAS data. The data are distributed across heterogeneous data centers at widely distributed locations worldwide. [Section 4.3]
- The PanDA ecosystem manages all workflows and workloads in ATLAS. It is designed to handle complex multistep workflows, running over thousands of files, using many different application workloads, and optimization to process over a large number of distributed sites deployed worldwide. PanDA uses the HTCondor system for job submission to a majority of computing resources. [Section 4.3]
- ATLAS has a long history of successfully using HPC resources during Run 2 and Run 3 at the LHC. From 2016 to 2020, US-based HPC resources supported 10–25% of ATLAS simulation production. This share decreased after several HPC resources were commissioned in Europe. In the first half of 2024, 21% of the worldwide ATLAS CPU resources came from HPC resources with 70% of them from the European HPC Vega in Slovenia, 8% from Perlmutter at NERSC, and 5% from CSCS (the Swiss National Supercomputing Centre) in Switzerland. European HPC resources are mostly used through grid interfaces. US resources at NERSC and Texas Advanced Computing Center (TACC) require special edge services to run ATLAS jobs. Only simulations were carried out, since they are not data intensive. In the future, ATLAS plans to adapt all workloads to run on HPC resources. This will put much higher demands on networking. [Section 4.4]
- ATLAS computing is fully distributed. All processing and reprocessing of the data is managed by PanDA, and Rucio is used for data cataloging and data transfer. All computing activities are free to occur at any site, irrespective of their Tier, and based on intelligent brokering of tasks and jobs. Distributed analysis jobs are also brokered by site capability—- users are discouraged from choosing a specific site. The distributed nature of ATLAS computing drives the network performance requirements between ATLAS sites. All ATLAS workloads and workflows may be run on demand at any time. [Section 4.4]

- There are four distinct storage types used by ATLAS computing environments; disk, tape, cloud and cache storage. [Section 4.4]
    - Disk storage is the main bulk of the storage: 1042M files with the volume of 412 PB are stored in the disk storage. In the US, 198M files with the volume of 98.5 PB are stored in the disk storage.
    - Tape storage is used as a main data archival system: 370M files with the volume of 639 PB are stored in the tape archive system. In the US, 63M files with the volume of 107 PB are stored in the BNL Tier-1 HPSS system.
    - S3 type storage is used for temporary output of event service jobs.
    - xRootD xCache storage is used for unmanaged disk data cache for input of user analysis jobs.
- BNL has implemented a vendor agnostic, resilient, scalable and modular terabit-per-second (Tbps) HTSN which serves as the primary network transport for all data-intensive collaborations at BNL. It provides high-throughput connectivity to all High-Performance Computing (HPC) and HTC collaborations and supports the timely transfer of large amounts of scientific data via the Internet through two diverse 800 Gbps (2 x 400 GbE) circuits that peer with ESnet routers. [Section 4.4]
- The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. The mission of the WLCG is to provide global computing resources to store, distribute and analyze the ~50-70 PB of data expected every year of operations from the LHC at CERN. WLCG is coordinated by CERN. It is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS and LHCb) and the participating computer centers. [Section 4.4]
- Depending on the size of the site (irrespective of the Tier label) there are 100 to 200k CPUs and 0.5 to 50 PB of disk storage available. In total ATLAS has a pledged disk space of 378 PB and a tape space of 625 PB at all sites worldwide combined in summer 2024. [Section 4.4]
- The US ATLAS Tier-1 is hosted at BNL's SDDC. The ATLAS connection to ESnet is shared with other programs hosted at the SDCC. The US Tier-1 is the largest of the ATLAS experiment, it represents about 25% of the computing resources of ATLAS at Tier-1s. [Section 4.6]
- There are four ATLAS Tier-2 centers in the US—- NorthEast Tier-2 (NET2), Great Lakes Tier-2 (AGLT2), MidWest Tier-2 (MWT2), and SouthWest Tier-2 (SWT2). These centers are used for all distributed production and user analysis workloads. Each Tier-2 center consists of multiple university-based clusters. [Section 4.6]
- In 2023-2024 ATLAS also added a new Tier-2 outside the US that contributes to the US ATLAS cloud. This site (TW-FTT) is derived from the former Tier-1 site in Taiwan and has a relatively small number of job slots (2200) and about 3 PB of storage. ESnet has helped to provide connectivity for this new site into LHCONE. However, this site has limited bandwidth (3 Gbps) and is only running tasks that are not data intensive. [Section 4.6]
- There are two ATLAS aFs currently located at DOE National Laboratories (BNL and SLAC). A third AF has been set up and is operational at University of Chicago (UC). Each AF consists of a few PB of storage and about 1000 cores provisioned as a local batch system. They also support Jupyter based interactive analysis, and provide some GPU or/and Acorn RISC Machine (ARM) resources. In general, more than 500 individual users are using AF. The data accessing model at these aFs includes both Rucio-managed data transfers to the aFs, and using xCache to access ATLAS data remotely. The popular Jupyter interactive analysis model presents a different kind of need for high bandwidth, low latency, and on demand networking. [Section 4.6]

- All US Tier-2 sites are strongly encouraged to provide at least 100 Gbps connectivity, preferably with resilience. The intent is to have multiple 400 Gbps connectivity for each Tier-2 by the time HL-LHC starts. [Section 4.7]
- Google Cloud and AWS computing services have been successfully integrated within the ATLAS workflow and data management systems PanDA and Rucio. Scaling to more than 100k concurrently running vCPUs in PanDA has been demonstrated. The solutions may be considered for example in the implementation of aFs for HL-LHC, with all the added values and services offered by the commercial cloud providers. In future work, data exports from cloud data centers to WLCG sites will have to be optimized, involving research network providers. In 2024 the Google Cloud is used for MC production processing as extension of a US Tier-2 center, and was able to scale of up to 15k job slots. [Section 4.8]
- ML is used in a wide range of applications in ATLAS, including physics analysis, simulation and physics object identification. Models trained for classification and regression, as well as generative models, are core components of the software and analysis chain. Most models used by the collaboration are trained on private resources due to the modest resource requirements and satisfactory availability of institutional GPU resources. ATLAS does not expect any significant changes in network load or the distributed computing software stack design during the ongoing analysis of Run 3 data. [Section 4.8]
- Given the size, complexity, and fully distributed nature of ATLAS computing, all workflow and data distribution could eventually be optimized and managed with AI. [Section 4.8]
- The LHC experiments can benefit from IRI, especially if the DOE HPC centers and the newly awarded HPDF are seamlessly integrated into the existing HTC (grid) infrastructure. IRI would provide the opportunity to implement near real-time workflows for critical payloads, such as High-Level Trigger, which should produce results in a timely manner. IRI would benefit all HEP by providing a uniform interface to disparate resources in a way that is easier to integrate with, and to better fit the variety of HEP workflows. [Section 4.9]
- ATLAS anticipates network needs will grow gradually. Increasing network capacity and performance will be needed at US ATLAS Tier-1, Tier-3 and Tier-3 aFs. The exact magnitude of future network provisioning is expected to be determined jointly by US ATLAS, ESnet and other involved parties. [Section 5.7]
- The primary concern for ATLAS is lack of sufficiently fine-grained network monitoring and performance information to help us debug and automate data transfers in real time between various sites. Better visibility is being worked on through various entities including WLCG Site Network Monitoring, the SciTags initiative (labeling network flows and IPv6 packets with owner/purpose) and improved monitoring from the various tools and infrastructures, but ATLAS still have significant gaps in monitoring. [Section 5.8]

6.9.2 ATLAS: Experiment Profile

ATLAS[100] is the largest of four particle detectors that measure and record the particle collisions at the LHC. The primary scientific goal is to quantitatively measure and discover properties of the SM of particle physics. ATLAS took a giant step in this with the discovery of the Higgs Boson in 2012, which led to the 2013 Nobel Prize in Physics for Peter Higgs and Francois Englert.

100 http://atlas.cern/

6.9.2.1 Science Background

In addition to the paper announcing the discovery of the Higgs Boson, ATLAS has published over a thousand new results in particle physics. A major emphasis of the ongoing Run 3 at the LHC will be the precision measurement of Higgs properties.

The SM of particle physics is considered to be an incomplete theory. In order to explain many observations and measurements like dark matter and Higgs mass hierarchy, new phenomenology remains to be discovered experimentally. ATLAS physics goals combine a strong program of SM measurements with search for new phenomena like Supersymmetry.

6.9.2.2 Collaborators

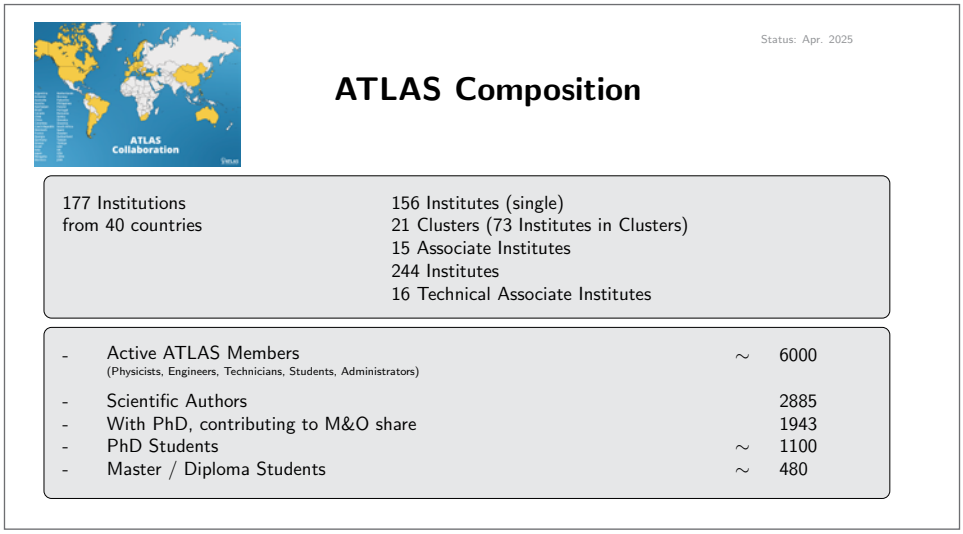


Figure 6.9.1: ATLAS Collaboration

ATLAS Institutions

- Argentina
  - Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, y National Scientific and Technical Research Council (CONICET)
  - Instituto de Física de Buenos Aires (IFIBA), Buenos Aires
  - Instituto de Física La Plata, Universidad Nacional de La Plata and CONICET, La Plata
- Armenia
  - Yerevan Physics Institute, Yerevan
- Australia
  - Department of Physics, University of Adelaide, Adelaide School of Physics, University of Sydney, Sydney
  - School of Physics, University of Melbourne, Victoria
- Austria
  - Universität Innsbruck, Department of Astro and Particle Physics, Innsbruck
  - Fachhochschule Wiener Neustadt, Wiener Neustadt

- Azerbaijan
  - Institute of Physics, Azerbaijan Academy of Sciences, Baku
- Brazil
  - Brazil Cluster: Departamento de Engenharia Elétrica, Universidade Federal de Juiz de Fora (UFJF), Juiz de Fora;Federal University of Bahia, Bahia;Instituto de Física, Universidade de São Paulo, São Paulo;Rio de Janeiro State University, Rio de Janeiro;Universidade Federal do Rio De Janeiro COPPE/EE/IF, Rio de Janeiro;
- Canada
  - Department of Physics, Simon Fraser University, Burnaby BC
  - Department of Physics, University of Alberta, Edmonton AB
  - Department of Physics, McGill University, Montreal QC
  - Group of Particle Physics, University of Montreal, Montreal QC
  - Department of Physics, Carleton University, Ottawa ON
  - Department of Physics and Astronomy, York University, Toronto ON;TRIUMF, Vancouver BC
  - Department of Physics, University of Toronto, Toronto ON
  - Department of Physics, University of British Columbia, Vancouver BC
  - Department of Physics and Astronomy, University of Victoria, Victoria BC
- CERN
  - European Organization for Nuclear Research, Geneva, Switzerland
- Chile
  - Chile Cluster: Departamento de Física, Pontificia Universidad Católica de Chile, Santiago;Departamento de Física, Universidad Técnica Federico Santa María, Valparaíso;Instituto de Alta Investigación, Universidad de Tarapacá, Arica;Instituto de Investigación Multidisciplinario en Ciencia y Tecnología, y Departamento de Física, Universidad de La Serena;Millennium Institute for Subatomic physics at high energy frontier (SAPHIR), Santiago;Universidad Andres Bello, Department of Physics, Santiago;
- China
  - Institute of High Energy Physics, Chinese Academy of Sciences, Beijing
  - Physics Department, Tsinghua University, Beijing
  - China NJU-SYSU cluster: Department of Physics, Nanjing University, Nanjing;School of Science, Shenzhen Campus of Sun Yat-sen University;
  - China USTC-SDU-SJTU-ZZU Cluster: Department of Modern Physics and State Key Laboratory of Particle Detection and Electronics, University of Science and Technology of China, Hefei;Institute of Frontier and Interdisciplinary Science and Key Laboratory of Particle Physics and Particle Irradiation (Ministry of Education, MOE), Shandong University, Qingdao;School of Physics and Astronomy, Shanghai Jiao Tong University, Key Laboratory for Particle Astrophysics and Cosmology (MOE), SKLPPC, Shanghai;School of Physics and Microelectronics, Zhengzhou University;Tsung-Dao Lee Institute, Shanghai;
  - Hong Kong Cluster: Department of Physics and Institute for Advanced Study, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong;Department of Physics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong;Department of Physics, University of Hong Kong, Hong Kong;

- Colombia
  - Colombia Cluster: Departamento de Física, Universidad Nacional de Colombia, Bogotá;Facultad de Ciencias y Centro de Investigaciónes, Universidad Antonio Nariño, Bogotá;
- Czech Republic
  - Palacký University, Joint Laboratory of Optics, Olomouc
  - Charles University, Faculty of Mathematics and Physics, Prague
  - Czech Technical University in Prague, Prague
  - Institute of Physics of the Czech Academy of Sciences, Prague
- Denmark
  - Niels Bohr Institute, University of Copenhagen, Copenhagen
- France
  - LAPP, Université Savoie Mont Blanc, CNRS/IN2P3, Annecy
  - LPC, Université Clermont Auvergne, CNRS/IN2P3, Clermont-Ferrand
  - IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette
  - LPSC, Université Grenoble Alpes, CNRS/IN2P3, Grenoble INP, Grenoble
  - Centre de Physique des Particules de Marseille (CPPM), Aix-Marseille Université, CNRS/IN2P3, Marseille
  - IJCLab, Université Paris-Saclay, CNRS/IN2P3, 91405, Orsay
  - APC, Université Paris Cité, CNRS/IN2P3, Paris
  - LPNHE, Sorbonne Université, Université Paris Cité, CNRS/IN2P3, Paris
  - L2IT, Université de Toulouse, CNRS/IN2P3, UPS, Toulouse
- Georgia
  - Georgia Cluster: E. Andronikashvili Institute of Physics, Iv. Javakhishvili Tbilisi State University, Tbilisi; High Energy Physics Institute, Tbilisi State University, Tbilisi; University of Georgia, Tbilisi;
- Germany
  - Institut für Physik, Humboldt Universität zu Berlin, Berlin
  - Physikalisches Institut, Universität Bonn, Bonn
  - Fakultät Physik , Technische Universität Dortmund, Dortmund
  - Institut für Kern- und Teilchenphysik, Technische Universität Dresden, Dresden
  - Physikalisches Institut, Albert-Ludwigs-Universität Freiburg, Freiburg
  - II. Physikalisches Institut, Justus-Liebig-Universität Giessen, Giessen
  - II. Physikalisches Institut, Georg-August-Universität Göttingen, Göttingen
  - Deutsches Elektronen-Synchrotron DESY, Hamburg and Zeuthen
  - Kirchhoff-Institut für Physik, Ruprecht-Karls-Universität Heidelberg, Heidelberg;Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg
  - Institut für Physik, Universität Mainz, Mainz

- Fakultät für Physik, Ludwig-Maximilians-Universität München, München
- Max-Planck-Institut für Physik (Werner-Heisenberg-Institut), München
- Department Physik, Universität Siegen, Siegen
- Fakultät für Physik und Astronomie, Julius-Maximilians-Universität Würzburg, Würzburg
- Fakultät für Mathematik und Naturwissenschaften, Fachgruppe Physik, Bergische Universität Wuppertal, Wuppertal

- Greece
  - National Centre for Scientific Research "Demokritos," Agia Paraskevi
  - Physics Department, National and Kapodistrian University of Athens, Athens
  - Department of Physics, Aristotle University of Thessaloniki, Thessaloniki
  - Physics Department, National Technical University of Athens, Zografou
- Israel
  - Department of Physics, Technion, Israel Institute of Technology, Haifa
  - Department of Particle Physics and Astrophysics, Weizmann Institute of Science, Rehovot
  - Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv
- Italy
  - INFN Sezione di Bologna; Dipartimento di Fisica e Astronomia A. Righi, Università di Bologna, Bologna
  - INFN e Laboratori Nazionali di Frascati, Frascati
  - INFN Sezione di Genova; Dipartimento di Fisica, Università di Genova, Genova
  - INFN Sezione di Lecce; Dipartimento di Matematica e Fisica, Università del Salento, Lecce
  - INFN Sezione di Milano; Dipartimento di Fisica, Università di Milano, Milano
  - INFN Sezione di Napoli; Dipartimento di Fisica, Università di Napoli, Napoli
  - INFN Sezione di Pavia; Dipartimento di Fisica, Università di Pavia, Pavia
  - INFN Sezione di Pisa; Dipartimento di Fisica E. Fermi, Università di Pisa, Pisa
  - INFN Gruppo Collegato di Cosenza, Laboratori Nazionali di Frascati; Dipartimento di Fisica, Università della Calabria, Rende
  - INFN Sezione di Roma; Dipartimento di Fisica, Sapienza Università di Roma, Roma
  - INFN Sezione di Roma Tor Vergata; Dipartimento di Fisica, Università di Roma Tor Vergata, Roma
  - INFN Sezione di Roma Tre; Dipartimento di Matematica e Fisica, Università Roma Tre, Roma
  - INFN-TIFPA; Università degli Studi di Trento, Trento
  - INFN Gruppo Collegato di Udine, Sezione di Trieste; Dipartimento Politecnico di Ingegneria e Architettura, Università di Udine, Udine; INFN Gruppo Collegato di Udine, Sezione di Trieste; Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste

- Japan
  - Research Center for Advanced Particle Physics and Department of Physics, Kyushu University, Fukuoka
  - Graduate School of Science, Kobe University, Kobe
  - Faculty of Science, Kyoto University, Kyoto
  - Department of Physics, Shinshu University, Nagano
  - Graduate School of Science and Kobayashi-Maskawa Institute, Nagoya University, Nagoya
  - Graduate School of Science, Osaka University, Osaka
  - Department of Physics, Tokyo Institute of Technology, Tokyo
  - Graduate School of Science and Technology, Tokyo Metropolitan University, Tokyo
  - International Center for Elementary Particle Physics and Department of Physics, University of Tokyo, Tokyo
  - Ochanomizu University, Otsuka, Bunkyo-ku, Tokyo
  - Waseda University, Tokyo
  - Division of Physics and Tomonaga Center for the History of the Universe, Faculty of Pure and Applied Sciences, University of Tsukuba, Tsukuba
  - KEK, High Energy Accelerator Research Organization, Tsukuba
- Morocco
  - Morocco Cluster: Faculté des Sciences Ain Chock, Université Hassan II de Casablanca; Faculté des Sciences Semlalia, Université Cadi Ayyad, LPHEA-Marrakech; Faculté des sciences, Université Mohammed V, Rabat; Faculté des Sciences, Université Ibn-Tofail, Kénitra;Institute of Applied Physics, Mohammed VI Polytechnic University, Ben Guerir; LPMR, Faculté des Sciences, Université Mohamed Premier, Oujda;
- Netherlands
  - Nikhef National Institute for Subatomic Physics and University of Amsterdam, Amsterdam
  - Institute for Mathematics, Astrophysics and Particle Physics, Radboud University/Nikhef, Nijmegen
- Norway
  - Department for Physics and Technology, University of Bergen, Bergen
  - Department of Physics, University of Oslo, Oslo
- Poland
  - University of Krakow, Faculty of Physics and Applied Computer Science, Krakow; Marian Smoluchowski Institute of Physics, Jagiellonian University, Krakow;
  - Institute of Nuclear Physics Polish Academy of Sciences, Krakow
- Portugal
  - Portugal Cluster: Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Lisboa; Departamento de Física, Universidade de Coimbra, Coimbra; Departamento de Física, Universidade do Minho, Braga; Departamento de Física, Instituto Superior Técnico, Universidade de Lisboa, Lisboa; Laboratório de Instrumentação e Física Experimental de Partículas - LIP, Lisboa; Departamento de Física Teórica y del Cosmos, Universidad de Granada, Granada (Spain);

- Romania
  - Romania Cluster: Department of Physics, Alexandru Ioan Cuza University of Iasi, Iasi;Faculty of Physics, University of Bucharest, Bucharest; Horia Hulubei National Institute of Physics and Nuclear Engineering, Bucharest; National Institute for Research and Development of Isotopic and Molecular Technologies, Physics Department, Cluj-Napoca; National University of Science and Technology Politechnica, Bucharest; Transilvania University of Brasov, Brasov; West University in Timisoara, Timisoara
- Russia
  - D.V. Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, Moscow
  - Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of National Research Centre "Kurchatov Institute," Moscow
  - National Research Nuclear University MEPhI, Moscow
  - P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Moscow
  - Novosibirsk State University Novosibirsk; Budker Institute of Nuclear Physics and NSU, SB RAS, Novosibirsk
  - Institute for High Energy Physics of the National Research Centre Kurchatov Institute, Protvino
  - Konstantinov Nuclear Physics Institute of National Research Centre "Kurchatov Institute," PNPI, St. Petersburg
  - Tomsk State University, Tomsk
- JINR
  - Joint Institute for Nuclear Research, Dubna, Russia
- Serbia
  - Institute of Physics, University of Belgrade, Belgrade
  - Slovak Republic
  - Slovak Republic Cluster: Department of Subnuclear Physics, Institute of Experimental Physics of the Slovak Academy of Sciences, Kosice; Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava;
- Slovenia
  - Department of Experimental Particle Physics, Jožef Stefan Institute and Department of Physics, University of Ljubljana, Ljubljana
- South Africa
  - South Africa Cluster: Department of Mechanical Engineering Science, University of Johannesburg, Johannesburg; Department of Physics, University of Cape Town, Cape Town; School of Physics, University of the Witwatersrand, Johannesburg; University of South Africa, Department of Physics, Pretoria; University of Zululand, KwaDlangezwa; iThemba Labs, Western Cape

- Spain
  - Institut de Física d'Altes Energies (IFAE), Barcelona Institute of Science and Technology, Barcelona
  - Departamento de Física Teorica C-15 and Centro de Inteligencia de las Fuerzas Armadas, Universidad Autónoma de Madrid, Madrid
  - Instituto de Física Corpuscular (IFIC), Centro Mixto Universidad de Valencia - CSIC, Valencia
- Sweden
  - Fysiska institutionen, Lunds universitet, Lund
  - Department of Physics, Royal Institute of Technology, Stockholm
  - Department of Physics, Stockholm University; Oskar Klein Centre, Stockholm
  - Department of Physics and Astronomy, University of Uppsala, Uppsala
- Switzerland
  - Albert Einstein Center for Fundamental Physics and Laboratory for High Energy Physics, University of Bern, Bern
  - Département de Physique Nucléaire et Corpusculaire, Université de Genève, Genève
- Türkiye
  - Ankara Cluster: Department of Physics, Ankara University, Ankara;Division of Physics, TOBB University of Economics and Technology, Ankara
  - Bogazici Cluster: Department of Physics Engineering, Gaziantep University, Gaziantep;Department of Physics, Bogazici University, Istanbul;Department of Physics, Istanbul University, Istanbul
  - Istinye University, Sariyer, Istanbul
- Taiwan
  - Department of Physics, National Tsing Hua University, Hsinchu
  - Institute of Physics, Academia Sinica, Taipei
- United Arab Emirates
  - UAE Cluster: New York University Abu Dhabi, Abu Dhabi;United Arab Emirates University, Al Ain;University of Sharjah, Sharjah;
- United Kingdom
  - School of Physics and Astronomy, University of Birmingham, Birmingham
  - Department of Physics and Astronomy, University of Sussex, Brighton
  - Cavendish Laboratory, University of Cambridge, Cambridge
  - Department of Physics, University of Warwick, Coventry
  - Particle Physics Department, Rutherford Appleton Laboratory, Didcot
  - SUPA - School of Physics and Astronomy, University of Edinburgh, Edinburgh
  - Department of Physics, Royal Holloway University of London, Egham
  - SUPA - School of Physics and Astronomy, University of Glasgow, Glasgow
  - Physics Department, Lancaster University, Lancaster

- Oliver Lodge Laboratory, University of Liverpool, Liverpool
- Department of Physics and Astronomy, University College London, London
- School of Physics and Astronomy, Queen Mary University of London, London
- School of Physics and Astronomy, University of Manchester, Manchester
- Department of Physics, Oxford University, Oxford
- Department of Physics and Astronomy, University of Sheffield, Sheffield

- United States of America
  - Department of Physics and Astronomy, University of New Mexico, Albuquerque NM
  - Department of Physics and Astronomy, Iowa State University, Ames IA
  - Department of Physics, University of Massachusetts, Amherst MA
  - Department of Physics, University of Michigan, Ann Arbor MI
  - High Energy Physics Division, ANL, Argonne IL
  - Department of Physics, University of Texas at Arlington, Arlington TX
  - Department of Physics, University of Texas at Austin, Austin TX
  - Berkeley cluster: Physics Division, LBNL, Berkeley CA; University of California, Berkeley CA
  - Department of Physics, Indiana University, Bloomington IN
  - Department of Physics, Boston University, Boston MA
  - California State University, CA
  - Laboratory for Particle Physics and Cosmology, Harvard University, Cambridge MA
  - Enrico Fermi Institute, University of Chicago, Chicago IL
  - Ohio State University, Columbus OH
  - Physics Department, Southern Methodist University, Dallas TX
  - Department of Physics, Northern Illinois University, DeKalb IL
  - Department of Physics, Duke University, Durham NC
  - Department of Physics and Astronomy, Michigan State University, East Lansing MI
  - University of Iowa, Iowa City IA
  - Department of Physics and Astronomy, University of California Irvine, Irvine CA
  - Nevis Laboratory, Columbia University, Irvington NY
  - Department of Physics, University of Wisconsin, Madison WI
  - Department of Physics and Astronomy, Tufts University, Medford MA
  - Department of Physics, Yale University, New Haven CT
  - Department of Physics, New York University, New York NY
  - Homer L. Dodge Department of Physics and Astronomy, University of Oklahoma, Norman OK
  - Institute for Fundamental Science, University of Oregon, Eugene, OR
  - Department of Physics, University of Pennsylvania, Philadelphia PA
  - Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh PA

- Physics Department, University of Texas at Dallas, Richardson TX
- Louisiana Tech University, Ruston LA
- Santa Cruz Institute for Particle Physics, University of California Santa Cruz, Santa Cruz CA
- Department of Physics, University of Washington, Seattle WA
- SLAC National Accelerator Laboratory, Stanford CA
- Department of Physics, Oklahoma State University, Stillwater OK
- Departments of Physics and Astronomy, Stony Brook University, Stony Brook NY
- Department of Physics, University of Arizona, Tucson AZ
- Physics Department, BNL, Upton NY
- Department of Physics, University of Illinois, Urbana IL
- Department of Physics, Brandeis University, Waltham MA

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
INTERNATIONAL (OVER 100 SITES IN MORE THAN 30 COUNTRIES)	Yes	Data transfer	In range of 10 gb to 250 tb	Continuously	Yes, see note below	Long tails from distributed transfers and access
US - BNL TIER-1 BNL UPTON	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US - AGLT2 (U MICHIGAN ANN ARBOR, MICHIGAN STATE LANSING)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US - MWT2 (U CHICAGO, U INDIANA BLOOMINGTON, U ILLINOIS URBANA)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US - NET2 (U MASSACHUSETTS AMHERST)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US - SWT2 (UT ARLINGTON, OKLAHOMA STATE)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US - SLAC STANFORD (SHARED TIER-3 ANALYSIS FACILITY)	Secondary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US - U NEW MEXICO ALBUQUERQUE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - IOWA STATE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - ANL, ARGONNE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - UT AUSTIN	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - LBNL, BERKELEY	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - BOSTON U	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - CALIFORNIA STATE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	

US - HARVARD U	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - OHIO STATE, COLUMBUS	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - SOUTHERN METHODIST DALLAS	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - NORTHERN ILLINOIS, DEKALB	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - DUKE, DURHAM	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - UC IRVINE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - COLUMBIA, NY	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - WISCONSIN MADISON	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - TUFTS, MEDFORD	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - YALE NEW HAVEN	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - NYU NEW YORK	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - U OKLAHOMA NORMAN	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - OREGON EUGENE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - U PENNSYLVANIA PHILADELPHIA	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - U PITTSBURGH	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - U TEXAS DALLAS	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - LOUISIANA TECH RUSTON	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - UC SANTA CRUZ	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - U WASHINGTON SEATTLE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - OKLAHOMA STATE STILLWATER	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - SUNY STONY BROOK	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - U ARIZONA TUSCON	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US - BRANDEIS U WALTHAM	Tertiary	Data transfer (pull)	Same	Occasionally	Same	

Table 6.9.1: ATLAS Data Mobility

Datasets in ATLAS are collections of files organized by category/workflow. In the table above, the range of ATLAS dataset sizes (not the size of the full data sample since those are never stored at any single site) are shown. Datasets are the fundamental units in ATLAS, fully integrated into the Rucio DMS, and used for workflow, transfers and user analysis. Individual datasets vary largely in size:

- raw datasets are in the range of 1 TB to about 250 TB
- AOD datasets are in the range of 1 GB to about 100 TB
- DAOD datasets are in the range of 1 GB to about several TB
- HITS datasets are in the order of several TBs

The ATLAS grid infrastructure consists of the Tier-0 computing site at CERN, 9 Tier-1's, 70 Tier-2's and about 50 Tier-3 sites distributed worldwide. Basically, all workflows are executed at all Tiers with MOUs: the Tier-0, Tier-1 and Tier-2 sites. Tape storage to store raw and AOD files is available at the Tier-0 and Tier-1 sites and selected Tier-2 sites (DESY and NET2) . The US Tier-1 and Tier-2 sites are described in Section 6.9.2.3. Plots of computing usage broken down by activity/workflow (data processing, MC, analysis etc.) are described in Section 6.9.2.4.1.

The setup and scope of Tier-3 sites vary significantly. There are dedicated AFs with all Grid components at MWT2, SLAC and BNL (denoted as either Shared Tier-3s or AFs). Here, similar activities at a reasonable fraction of a Tier-2 site can be expected. The role and number of AF sites across BNL and ATLAS will be defined over the next few years, and it is foreseen that bulk user access to the AF will be needed in the future, perhaps with shared storage between all the AFs. This may require a review of some networking parameters. Tier-3 sites at most universities are usually smaller local CPU clusters with several TBs of disk storage attached. At these sites individual end-user analysis happens without any production activity.

Depending on the size of the site (irrespective of the Tier label) there are 100 to 200k CPUs and 0.5 to 50 PB of disk storage available. In total ATLAS has a pledged disk space of 378 PB and a tape space of 625 PB at all sites worldwide combined in summer 2024. There is a CPU pledge of approximately 4300 kHS23 (about 500k batch cores) for the full year 2024.

ATLAS requests additional CPU resources "beyond pledge" from all sites to generate simulated events which are crucial for physics studies. On average, a substantial amount (~3700 kHS23) is delivered as a beyond pledge resource. About 2/3 of this came from HPC site Vega in the past year. In terms of number of cores, the baseline is around 500k CPU cores constantly used worldwide. Further details about the breakdown of the number of cores used concurrently by different workflow at US facilities is presented in Section 6.9.2.4.

The Figures 6.9.2 and 6.9.3 show the total dataset volume on disk and tape by data type.

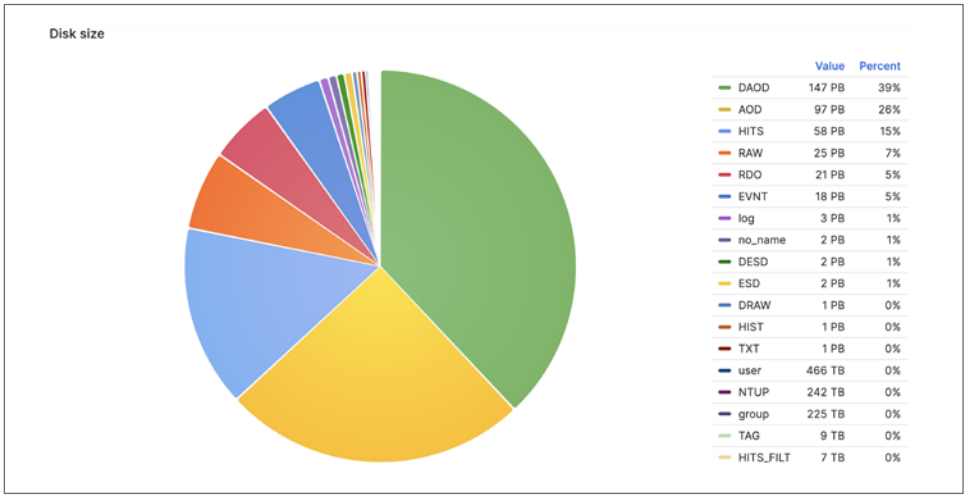


Figure 6.9.2: Total Dataset Volume — Disk

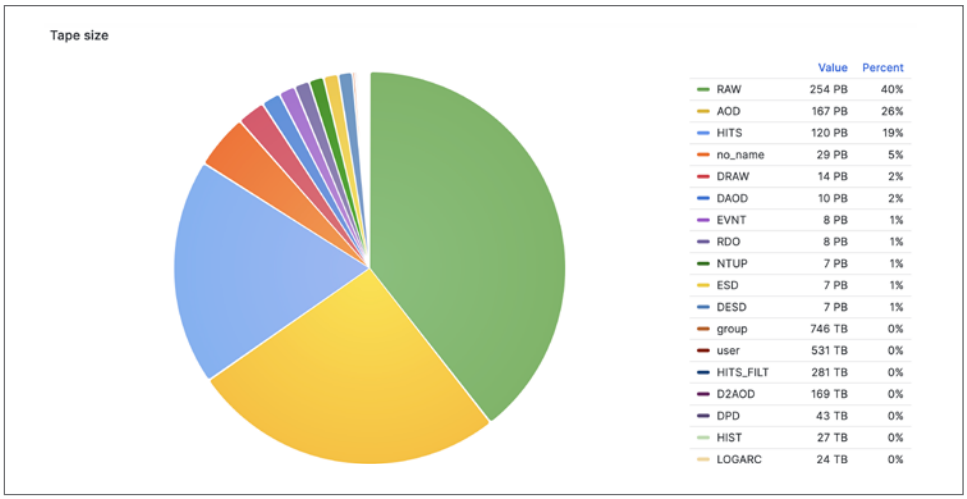


Figure 6.9.3: Total Dataset Volume — Tape

6.9.2.3 IRI Readiness

ATLAS and HEP experiments in general can benefit from IRI, especially if LCF and HPDF are seamlessly integrated into the existing HTC (grid) infrastructure (including OSG). IRI will also provide the opportunity to implement near real-time workflows (WF) for critical payloads such as High-Level Trigger WF, which should produce results in a timely (near real time) manner. The BNL network infrastructure has been designed with IRI patterns in mind. This is shown in the network design of the external Science DMZ network for data transfer and a firewalled science portals network for services. IRI benefits HEP by providing a uniform interface to disparate resources in a way that is easier to integrate with, better fit to HEP workflows. The ATLAS workflows that benefit most from an IRI approach are those currently running at LCF and HPCs.

6.9.2.4 Instruments and Facilities

6.9.2.4.1 LHC

The LHC collides protons more than a billion times every second, out of which ATLAS selects interesting collisions for recording at a rate of a thousand times every second. With over two thousand hours of data collection every year when the LHC is running, ATLAS has a huge data sample for physicists to analyze from the completed Run 1 and Run 2. Run 3 started data collection in 2022 and is planned to continue until the end of 2025. Table 6.9.2 shows typical LHC parameters.

Quantity	Number
CIRCUMFERENCE	26659 m
DIPOLE OPERATING TEMPERATURE	1.9 K (-271.3 °C)
NUMBER OF MAGNETS	9593
NUMBER OF MAIN DIPOLES	1232
NUMBER OF MAIN QUADRUPOLES	392
NUMBER OF RADIO FREQUENCY (RF) CAVITIES	8 per beam
NOMINAL ENERGY, PROTONS	6.8 TeV
NOMINAL ENERGY, IONS	2.76 TeV/u (energy per nucleon)
NOMINAL ENERGY, PROTONS COLLISIONS	13.6 TeV
NUMBER OF BUNCHES PER PROTON BEAM	2340

NUMBER OF PROTONS PER BUNCH (AT START)	1.6 x 1011
NUMBER OF TURNS PER SECOND	11245
NUMBER OF COLLISIONS PER SECOND	1 billion

Table 6.9.2: LHC Parameters

All of the equipment needed for the High-Luminosity LHC, the LHC's successor, and its experiments will be installed during Long Shutdown 3, between 2026 and the end of 2028. The High-Luminosity LHC is scheduled to come into operation at the end of 2029, though this could shift to 2030 with the outcome decided in October 2024.

Stable beam time for 2024 (with a duty cycle of 50%) will continue with an average pileup of $<\mu>=55$ and luminosity leveling at $\mu=64$, leading to a projected 110 fb-1 of delivered integrated luminosity. Additionally, about 4 weeks of heavy ion collisions are expected. In 2025 a similar data taking setup is planned.

6.9.2.4.2 ATLAS

ATLAS is the largest of four particle detectors that measure and record the particle collisions at the LHC. Selected highlights from this web page are shown below. The primary scientific goal is to combine a strong program of SM measurements with search for new phenomena like Supersymmetry.

The four major components of the ATLAS detector are the Inner Detector, the Calorimeter, the Muon Spectrometer and the Magnet System. Integrated with the detector components are: the Trigger and Data Acquisition System, a specialized multi-level computing system, which selects physics events with distinguishing characteristics; and the Computing System, which develops and improves computing software used to store, process and analyze vast amounts of collision data at 130 computing centers worldwide.

ATLAS is designed to observe up to 1.7 billion proton-proton collisions per second, with a combined data volume of more than 60 million megabytes per second. However, only some of these events will contain interesting characteristics that might lead to new discoveries. To reduce the flow of data to manageable levels, ATLAS uses a specialized two-level online event selection system — the Trigger System — which selects events with distinguishing characteristics that make them interesting for physics analyzes.

The ATLAS trigger system carries out the selection process in two stages. The Level-1 hardware trigger, constructed with custom-made electronics, works on a subset of information from the calorimeter and muon detectors. The decision to keep the data from an event is made less than two-and-half microseconds after the event occurs, and the event is then retrieved from pipelined storage buffers. The Level-1 trigger can save at most 100,000 events each second for the High-Level Trigger (HLT).

The HLT is a large farm of CPUs — i.e., a software-based trigger — which refines the analysis of the hardware-based Level-1 trigger. It conducts a very detailed analysis either by performing overall examination of the whole event for selected layers of the detector (for example calorimeters, trackers, muon detectors), or, by utilizing the data in smaller and isolated regions of the detector. About 1000 events per second are selected by the HLT analysis and are fully assembled into an event record. These events are passed on to a data storage system for offline analysis.

The ATLAS Computing System analyzes the data produced by the ATLAS detector, developing and improving computing software used to store, process and analyze vast amounts of collision data.

Data from the ATLAS detector is calibrated, reconstructed and automatically distributed all around the world by the ATLAS DMS. The ATLAS Production System then filters through these events and selects the ones needed for a particular type of analysis. This brings the data set down to a manageable size for someone doing an analysis on their laptop.

All members of the ATLAS Collaboration have equal access possibilities to all ATLAS data, independently of their geographical location, thanks to the WLCG. ATLAS computing infrastructure and software are constantly evolving — with the help of members of the Collaboration.

ATLAS has over 130 computing centers worldwide — located on every inhabited continent — nursed around the clock by members of the Collaboration.

6.9.2.4.3 WLCG

The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. The mission of the WLCG project is to provide global computing resources to store, distribute and analyze the ~50-70 PB of data expected every year of operations from the LHC at CERN on the Franco-Swiss border.

The scale and complexity of data from the LHC is unprecedented. These data need to be stored, easily retrieved and analyzed by physicists all over the world. This requires massive storage facilities, global networking, immense computing power, and, of course, funding.

CERN does not have the computing or financial resources to crunch all the data on-site, so in 2002 it turned to grid computing to share the burden with computer centers around the world. The WLCG builds on the ideas of grid technology initially proposed in 1999 by Ian Foster and Carl Kesselman.

WLCG is coordinated by CERN. It is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS and LHCb) and the participating computer centers. It is reviewed by a board of delegates from partner country funding agencies, and scientifically reviewed by the LHC Experiments Committee. The WLCG is partnered with EGI, OSG, and NeIC (Nordic e-Infrastructure Collaboration).

6.9.2.4.4 US ATLAS T1 at BNL

The US ATLAS Tier-1 is hosted at BNL's SDDC. The ATLAS connection to ESnet is shared with other programs hosted at the SDCC. The US Tier-1 is the largest of the ATLAS experiment, it represents about 25% of the computing resources of ATLAS at Tier-1s.

The infrastructure of the US ATLAS Tier-1 site at BNL consists of the following three large blocks, currently physically located in the existing SDCC datacenter based on B725 building at BNL site:

- ATLAS Tier-1 Linux Farm including HTCondor and Grid CE / Gatekeeper infrastructure (CPU resource),
- ATLAS Tier-1 dCache storage system, including a Ceph testbed and GridFTP / SRM / XRootD DTNs (disk resource),
- Oracle SL8500 based part of the SDCC HPSS complex devoted to ATLAS (tape resource).
    - some tape libraries with LHC Run1 data are based in the 'old' computing center in B515

The network architecture and the high-level service layout of the US ATLAS Tier-1 site is described in the Section 6.9.2.8.1.

Figures 6.9.4, 6.9.5, and 6.9.6 show the amount of usable CPU, disk and tape resources provided by the ATLAS Tier-1 site at BNL (values in FY18–20 range are actual delivered values, and values in the range FY21–27 are from the most up-to-date projection as of Aug 2020).

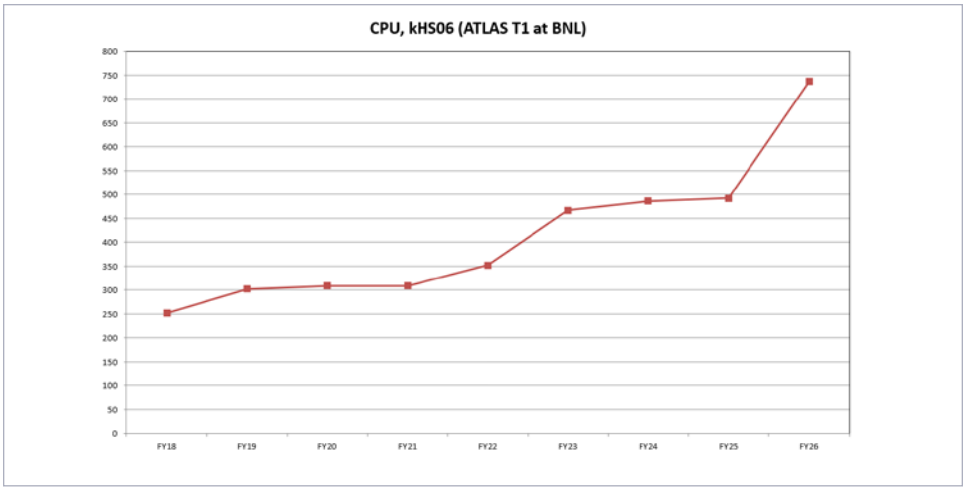


Figure 6.9.4: CPU resources at ATLAS Tier-1 site at BNL

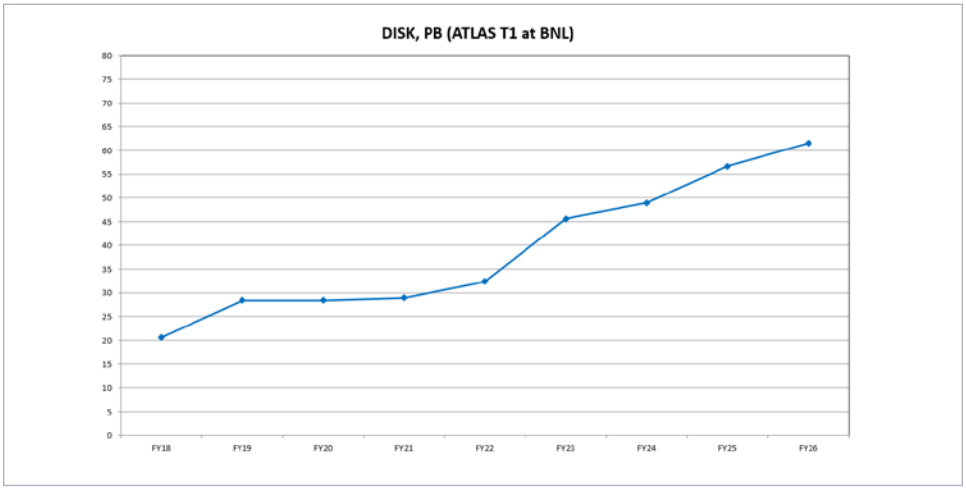


Figure 6.9.5: Disk Storage at ATLAS Tier-1 site at BNL

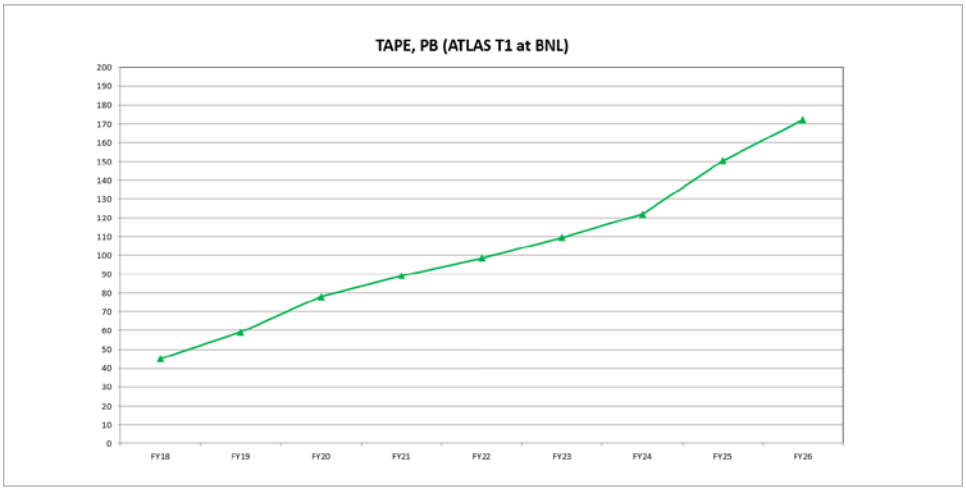


Figure 6.9.6: Tape Storage ATLAS Tier-1 site at BNL

6.9.2.4.5 US ATLAS T2 Infrastructure

There are four ATLAS Tier-2 centers in the US: NorthEast Tier-2 (NET2), Great Lakes Tier-2 (AGLT2), MidWest Tier-2 (MWT2), and SouthWest Tier-2 (SWT2). These centers are used for all distributed production and user analysis workloads. Each Tier-2 center consists of multiple university-based clusters. The ATLAS Great Lakes Tier-2 clusters are located at the University of Michigan and Michigan State University. The MidWest Tier-2 clusters are at University of Chicago, Indiana University and the University of Illinois at Urbana-Champaign. The NorthEast Tier-2 is run by the University of Massachusetts at Amherst (which collaborates with Harvard for the storage operations) and is located in the Massachusetts Green High Performance Computing Center (MGHPCC), in Holyoke, Massachusetts. Finally, the SouthWest Tier-2 center has clusters at the University of Texas at Arlington, and Oklahoma University. The MWT2 is funded at a level 50% higher than the other Tier-2 sites in the US, and therefore provides 50% more resources.

Note that in 2023-2024 ATLAS also added a new Tier-2 outside the US that contributes to the US ATLAS cloud. This site (TW-FTT) is derived from the former Tier-1 site in Taiwan and has a relatively small number of job slots (2200) and about 3 PB of storage. ESnet has helped to provide connectivity for this new site into LHCONE. However, this site has limited bandwidth (3 Gbps) and is only running tasks that are not data intensive.

The resources provisioned at each US Tier-2 center is based on the WLCG pledges, which are set by ATLAS/ WLCG Resource Review Board and Computing Resource Scrutiny Group annually. These pledges provide the baseline for the minimal CPU and Storage resources to be made available by each Tier-2 center in ATLAS. The pledges are set in the April of every year, and are tracked monthly.

In addition, ATLAS officially requests additional CPU cycles from all sites, in order to accommodate MC simulations in support of physics publications, equivalent to the pledged amount. Finally, ATLAS adds 20% to both compute and storage pledges to accommodate the physics analysis of US users — dedicating a higher share to analysis queues than the ATLAS average. Combining all these factors, the US Tier-2 sites aim to provide 220% of the CPU pledge, and 120% of the storage pledge. The extra amounts for US physicists are subject to changes based upon funding and needs.

Table 6.9.3 shows how the Tier-2s have performed at the end of FY24 procurement versus the targets.

AFTER FY24 PROCUREMENT	CPU		Disk	
Site	US Target CPU (kHS23)	Total CPU (kHS23)	US Target (PB)	Total Disk (PB)
AGLT2	227	246	12.3	18.5
MWT2	340	679	18.4	21.2
NET2	227	239	12.3	12.4
SWT2	227	341	12.3	17.3
TOTAL	1021	1,505	55.2	69.4

Table 6.9.2: LHC Parameters

All US Tier-2 sites are strongly encouraged to provide at least 100 Gbps connectivity, preferably with resilience. In Section 6.9.2.8 below ATLAS will provide some details for Tier-2 networking. The intent is to have multiple 400 Gbps connectivity for each Tier-2 by the time HL-LHC starts.

Figure 6.9.7 shows the number of HEPscore23 (HS23) weighted jobs slots used at all ATLAS Tier-2 sites during this two-year period from October 2022 to September, 2024. Over this period, MWT2 and SWT2 provided the highest number of slots to ATLAS among all Tier-2 sites worldwide, with AGLT2 at number 4. The data are binned by weeks and only considers ATLAS Grid resources.

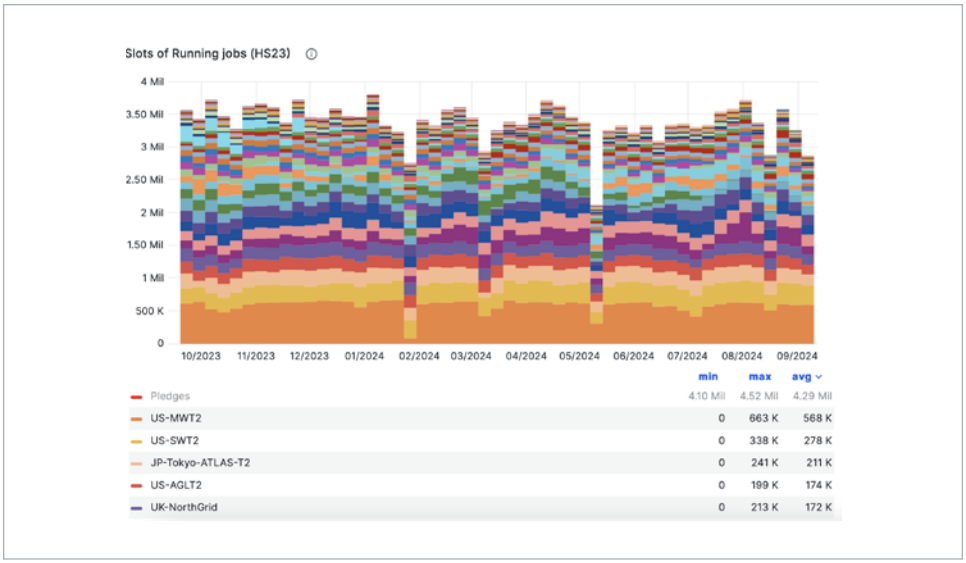


Figure 6.9.7: HEPscore23 (HS23) weighted jobs slots used at all ATLAS Tier-2 sites

6.9.2.4.6 US ATLAS Analysis Facilities

Final stages of analysis (statistical inference, weighting, signal/background calculations, plots etc.) are usually not done on the Tier-1 and Tier-2 sites using grid tools. US physicists typically copy the selected smaller data products from the grid to local facilities for final analysis leading to publications. These local analysis sites are called Tier-3 (T3) or/and AFs.

US ATLAS consolidated the user analysis resource and supports, and currently maintains three common and shared AFs, available for all US ATLAS physicists. These sites require good network bandwidth to WLCG sites, including US Tier-1 and Tier-2 sites.

There are two AFs currently located at DOE National Laboratories (BNL and SLAC). A third AF has been set up and is operational at University of Chicago (UC). Each AF consists of a few PB of storage and about 1000 cores provisioned as a local batch system. They also support Jupyter based interactive analysis, and provide some GPU or/and ARM resources. In general, more than 500 individual users are using AF. The data accessing model at these AFs includes both Rucio-managed data transfers to the AFs, and using XCache to access ATLAS data remotely. The popular Jupyter interactive analysis model presents a different kind of need for high bandwidth, low latency, and on demand networking.

The future development of the AF and number of AF in US ATLAS depends very much on the WLCG-wide policy. In 2024 the LHC Coordinating Committee (LHCC) defined a list of questions to be answered by all LHC experiments.

6.9.2.5 Generalized Process of Science

6.9.2.5.1 Computation

The main activities in terms of CPU usage are:

- MC event generation. Different event generators are used to repeatedly generate proton-proton or heavy ion collision events and particularly interesting particle decays. There are usually no or very small input file sizes in the order of GBs required and the output file sizes in the order of a few MBs per job.
- MC data simulation. The outputs of the event generation are used to simulate the particle decays in a detailed or parameterized ATLAS detector simulation. These are the most CPU

demanding tasks with event processing times of in the range of 1 to 30 minutes. The output file sizes are in the order of several 100 MBs to a few GBs per job.

- MC simulation reconstruction. Individual physics objects like electrons, muons, photons etc. of the previously simulated particle decays in the collision events are reconstructed for a later analysis and stored in AOD files. The processing times per event are in the range of 10-30 seconds and the output file sizes are in the range of 1–10 GB per job.
- Detector data reconstruction. Physics objects are reconstructed from raw detector data and stored in AOD files. The processing times per event are in the range of 10-30 seconds. The input file sizes are in the range of 2 to 5 GB and the output file sizes are in the range of 1 to 10 GB per jobs.
- Production of derived analysis formats (derivation production) as inputs for physics and performance analysis. The physics object files are augmented with additional calibrated information and events are filtered based on interesting event patterns and stored in DAOD files. About 80 different individual DAOD formats are written. The input and output file sizes are in the range of 100 MB to 10 GB per job.
- Individual and group analysis. Individual physicists are processing the filtered physics objects files to filter out highly specialized information. The input and output files are in the range of 100 MB to 10 GB per job.

All of the above activities process datasets with sizes between 10 GB and 250 TB consisting of 10–100k event data files and file sizes up to 15-25 GBs each. The processing of these datasets is broken down into individual jobs by the Workflow Management System PanDA and each job processes only a fraction of the datasets in each job.

During Run 3 the so-called data carousel came into production that synchronizes data staging from tape and its delivery with job submission through the PanDA workflow management system. This allows a more efficient usage of disk storage at the grid sites.

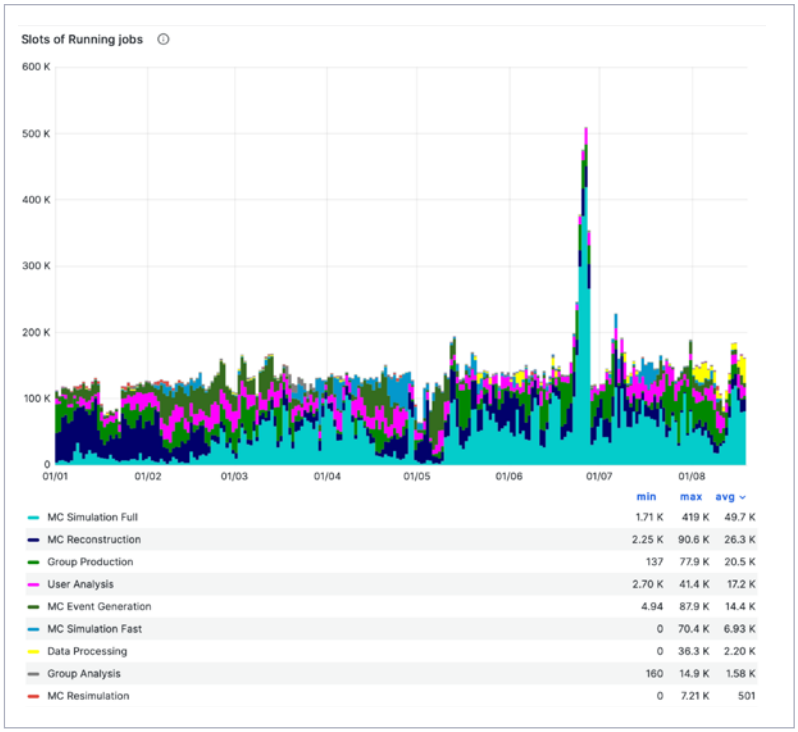


Figure 6.9.8: Number of running job slots (cores accessible through PanDA) at all US grid sites in the year 2024 so far, as a function of the activity type

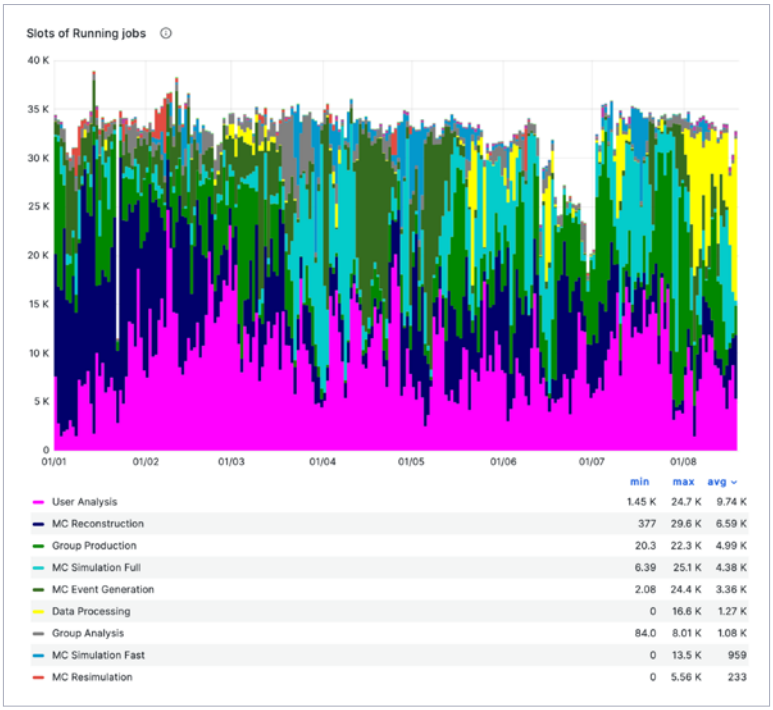


Figure 6.9.9: Number of running job slots in the year 2024 so far at the BNL Tier-1 as a function of activity type

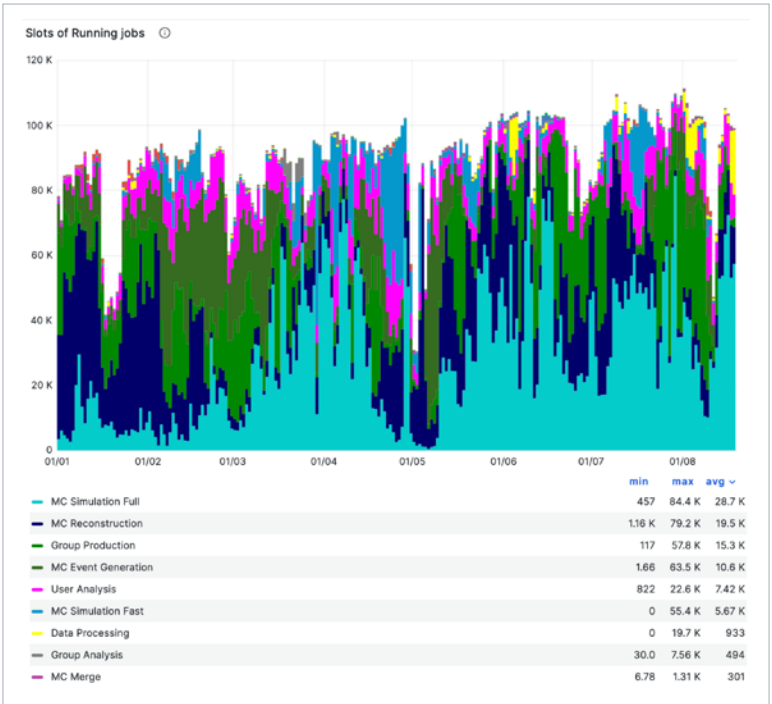


Figure 6.9.10: Number of running job slots in the year 2024 so far at the US Tier-2 sites AGLT2, MWT2, NET2 and SWT2 as a function of activity type

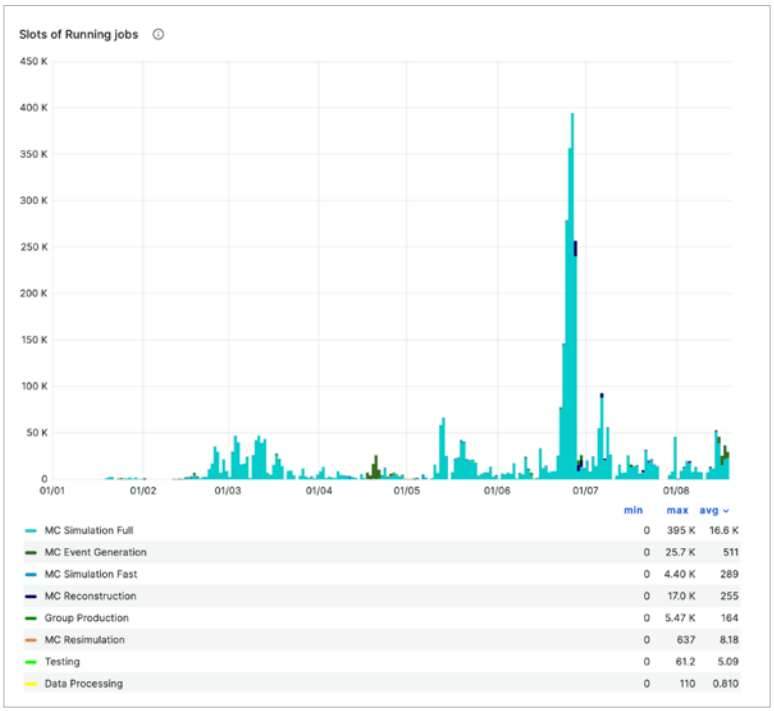


Figure 6.9.11: Number of running job slots in 2024 so far with the HPC Perlmutter at NERSC and the Frontera HPC at TACC Note that the HPC sites are mainly used for MC simulations

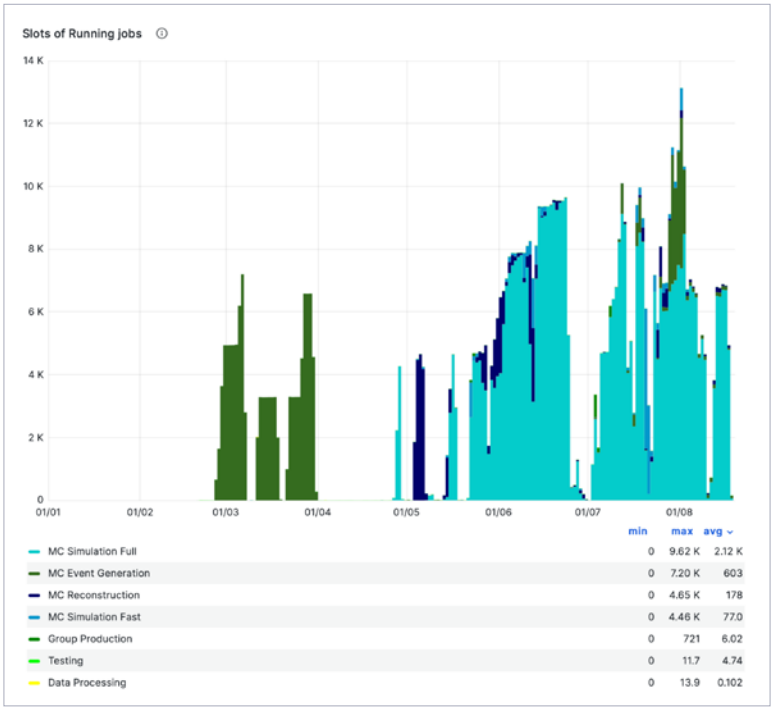


Figure 6.9.12: Number of running job slots in 2024 so far with the Google Cloud as an extension of the SWT2

6.9.2.5.1.1 High-Performance Computing

ATLAS has a long history of successfully using HPC resources during Run 2 and Run 3 at the LHC. From 2016-2020, US-based HPC resources supported 10-25% of ATLAS simulation production. This share decreased after several HPCs were commissioned in Europe (Slovenia, Spain and Czechia) as part of the EU Exascale project. In the first half of 2024 21% of the worldwide ATLAS CPU resources came from HPCs, with 70% of them from the European HPC Vega in Slovenia, 8% from Perlmutter at NERSC and 5% from CSCS in Switzerland.

European HPC resources are mostly used through grid interfaces. US resources at NERSC and TACC require special edge services through Harvester to run ATLAS jobs. Only simulations were carried out, since they are not data intensive. In the future, ATLAS plans to run all workloads at HPCs. This will put much higher demands on networking.

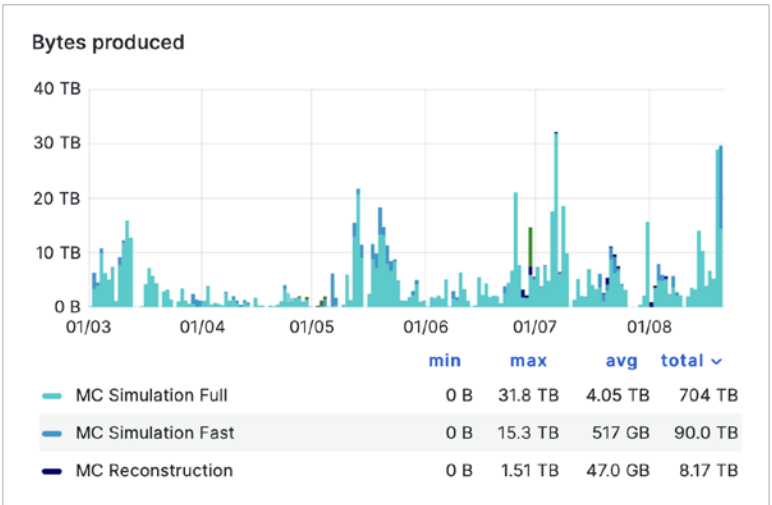


Figure 6.9.13: Amount of data produced per day at NERSC since March 2024 that is directly transferred from the jobs from NERSC to BNL storage

6.9.2.5.1.2 Distributed Computing

ATLAS computing is fully distributed, as soon as the raw data are transferred out of CERN. The first step in raw data distribution is planned to the 10 Tier-1 sites, according to MOU pledges. This allows for a complete second copy of the raw data. After this archival step, all processing and reprocessing of the data is managed by PanDA, based on the current availability of resources. Rucio is used for data cataloging and data transfer. All computing activities are free to occur at any site, irrespective of their Tier, and based on intelligent brokering of tasks and jobs. Distributed analysis jobs are also brokered by site capability: users are discouraged from choosing a specific site. The distributed nature of ATLAS computing drives the network performance requirements between ATLAS sites. All ATLAS workloads and workflows may be run on demand at any time.

6.9.2.5.2 Storage

ATLAS storage usage

- There are four distinct storage types used by ATLAS computing environments; disk, tape, cloud and cache storage.
    - Disk storage is the main bulk of the storage used by ATLAS. Currently, 1042M files with the volume of 412 PB are stored in the disk storage. In the US, 198M files with the volume of 98.5 PB are stored in the disk storage.

- Tape storage is used as a main data archival system. Currently, 370M files with the volume of 639 PB are stored in the tape archive system. In the US, 63M files with the volume of 107 PB are stored in the BNL Tier-1 HPSS system.
- S3 type storage is used for temporary output of event service jobs. The files in S3 are small and will be merged to produce the large files before they are written to the disk storage described above.
- XRootD XCache storage is used for unmanaged disk data cache for input of user analysis jobs. These small cached data are created on demand by user jobs. And, their data are copied from the disk storage described above and stay in the cache for the duration of the user analysis jobs.

- The main disk storage is split into a few different areas. The largest fraction (~90%) of the disk storage is assigned to DATADISK space where the production system uses it for inputs and outputs of the jobs. GROUPDISK area is used for Physics group production while SCRATCHDISK area is used mainly by inputs and outputs of user analysis jobs.
- Files are categorized as primary or secondary depending on if they are the main copy of the data or duplicate copy of the data. Roughly, 75% of the data in the disk storage are primary while 25% of the data are secondary. In addition, files are classified by their data types such as raw, HITS, AOD and DAOD, etc. raw data are unprocessed data from the detectors. They are stored mainly in tape storage for archiving purposes. But they can be brought to disk storage for production of AOD. AOD are further processed, and resulted in DAOD, Derived AOD. AOD and DAOD data types occupy the largest fraction of the disk storage with ~40% and ~25% respectively.
- Storage usage at US T1 and T2 sites are shown in Figures 6.9.14 through 6.9.18.
- Changes in Run 3: Due to the usage of the data carousel less AODs remain persistent on disk storage and are staged back from tape on demand.

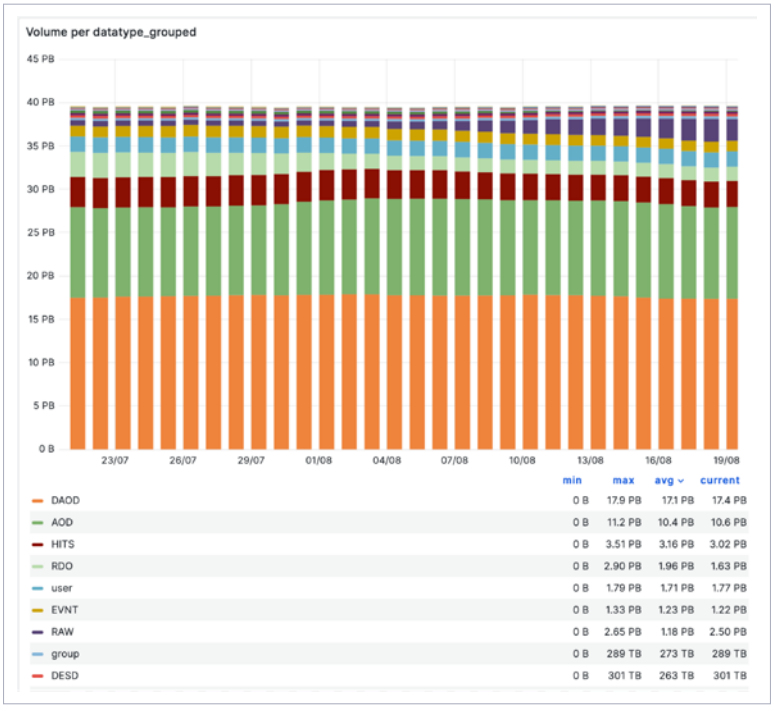


Figure 6.9.14: Disk storage usage in US T1/BNL categorized by data type

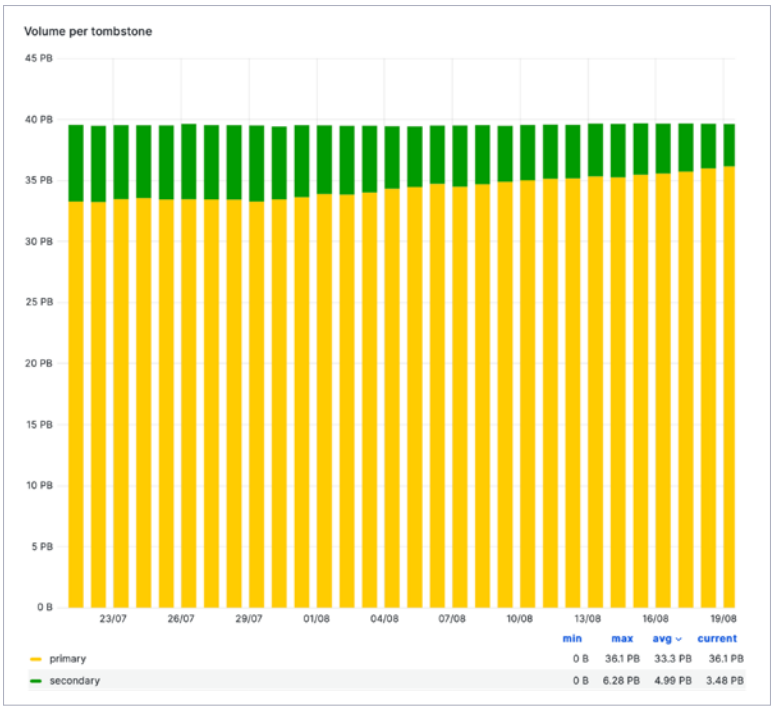


Figure 6.9.15: Disk storage usage in US T1/BNL categorized by primary and secondary data

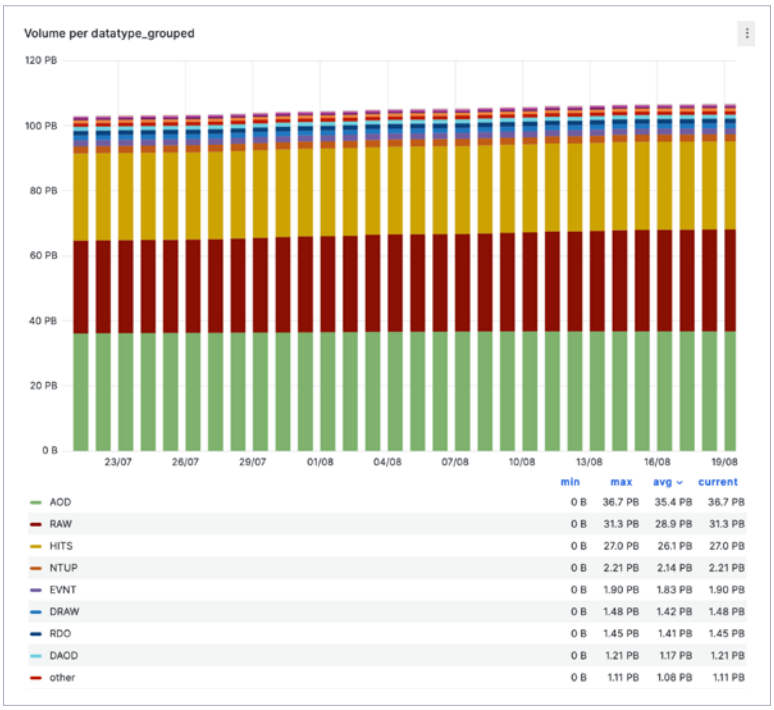


Figure 6.9.16: Tape storage usage in US T1/BNL categorized by data type

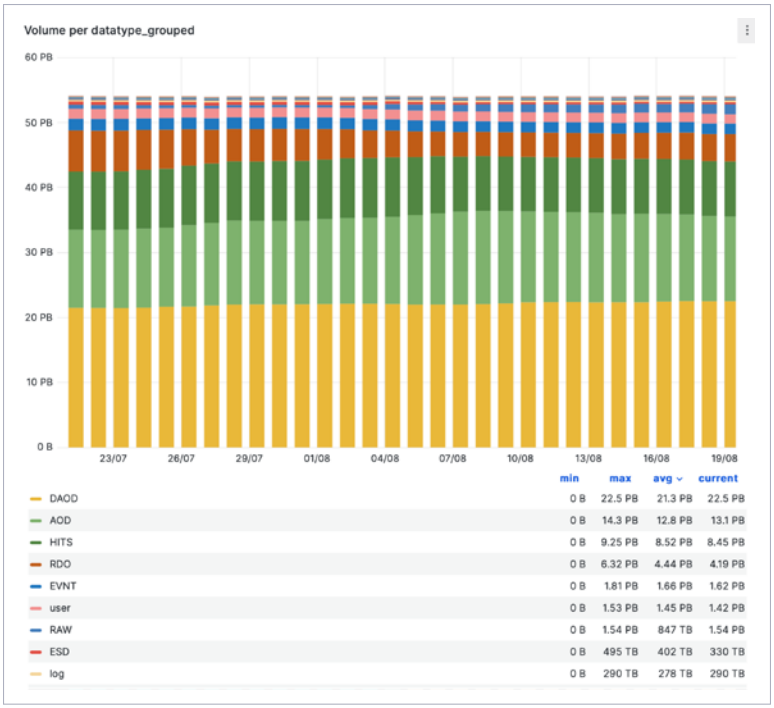


Figure 6.9.17: Storage usage in US T2s categorized by data type

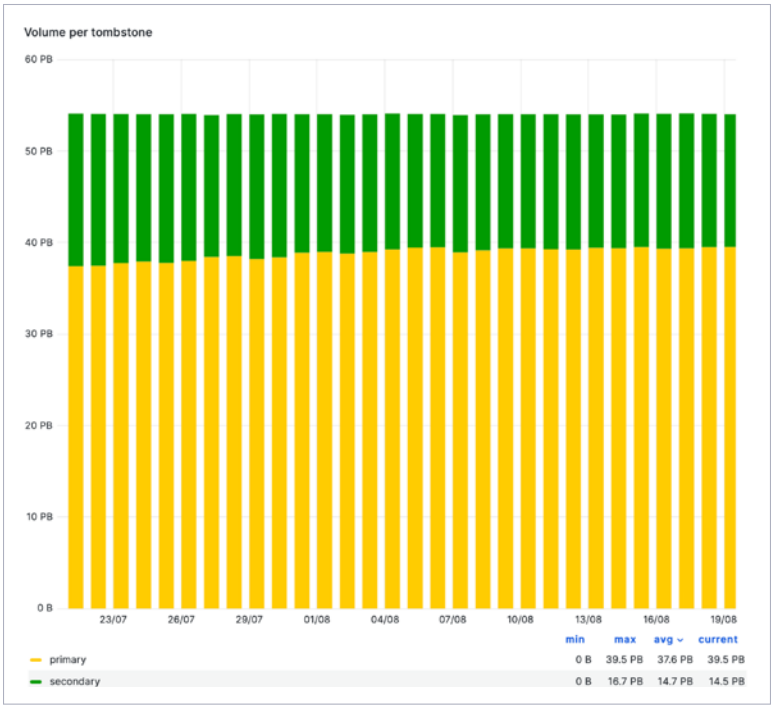


Figure 6.9.18: Storage usage in US T2s categorized by primary and secondary data type

6.9.2.5.3 Network Use Cases & Data Flow

The network use cases and data flows are already described in Sections 6.9.2.3 and 6.9.2.4. ATLAS provides some additional plots (see Figures 6.9.19 to 6.9.26) to quantify data transfer rates and volumes.

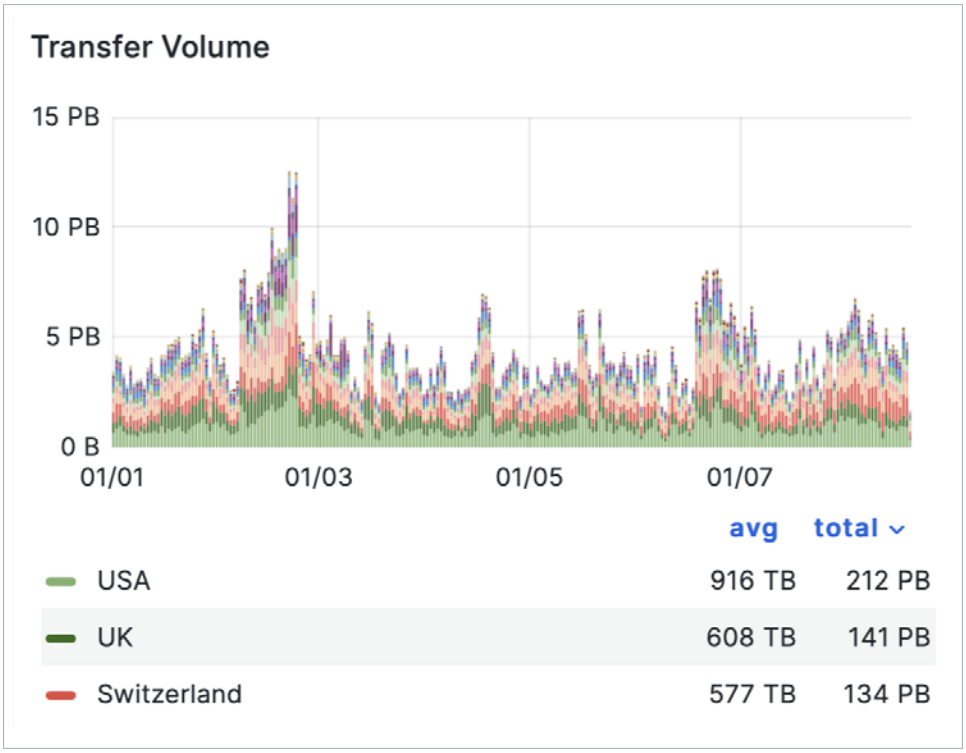


Figure 6.9.19: Transfer volume per day worldwide in 2024 so far

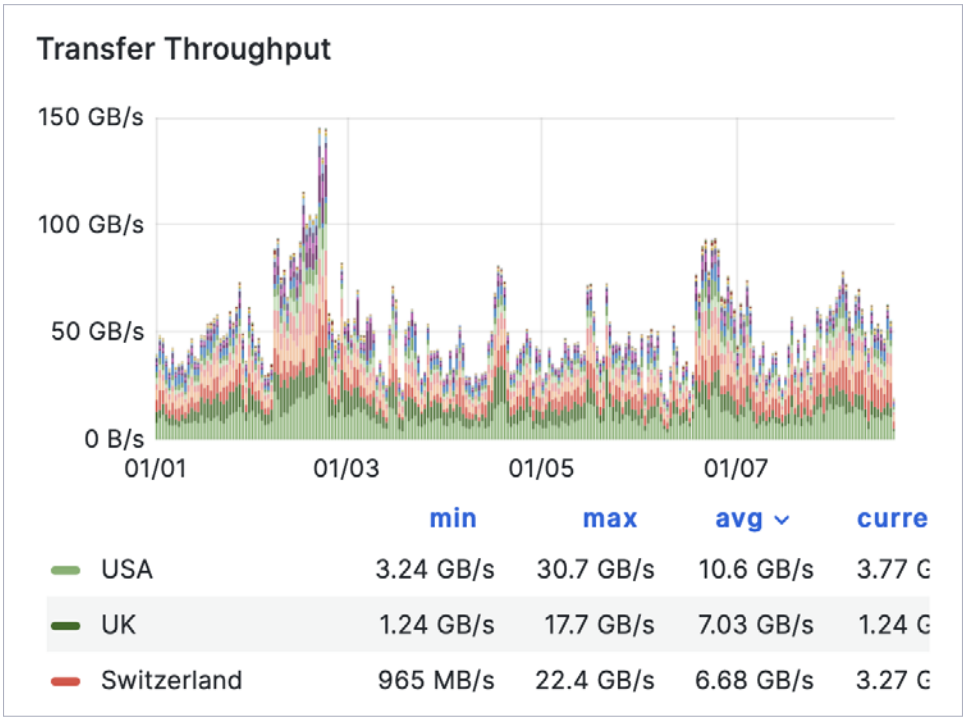


Figure 6.9.20: Transfer rate per day worldwide in 2024 so far

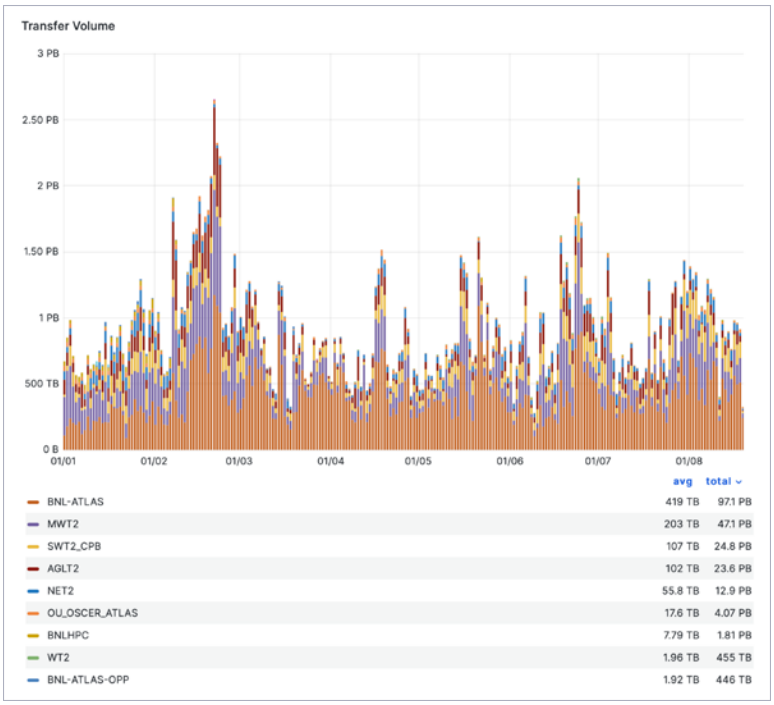


Figure 6.9.21: Transfer volume in 2024 so far by US grid site

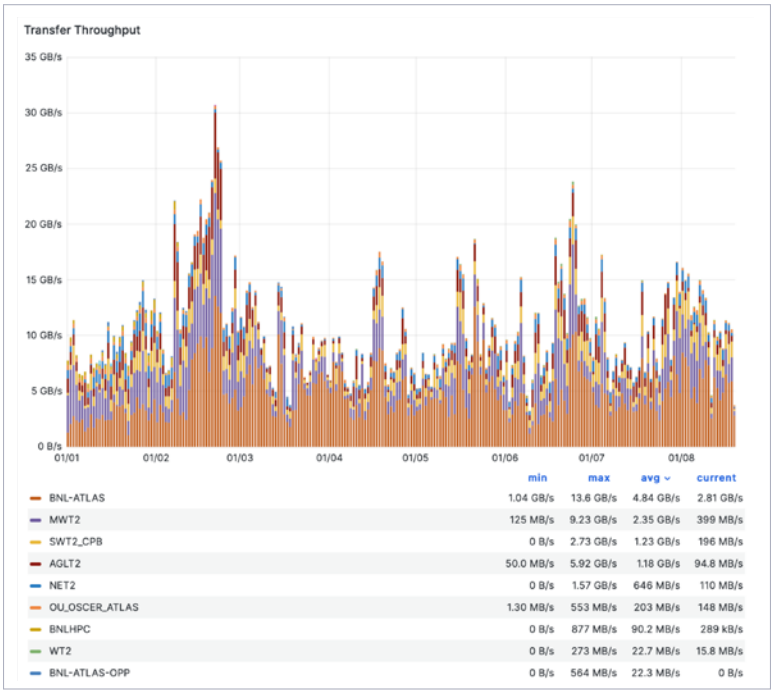


Figure 6.9.22: Transfer rate in 2024 so far by US grid site

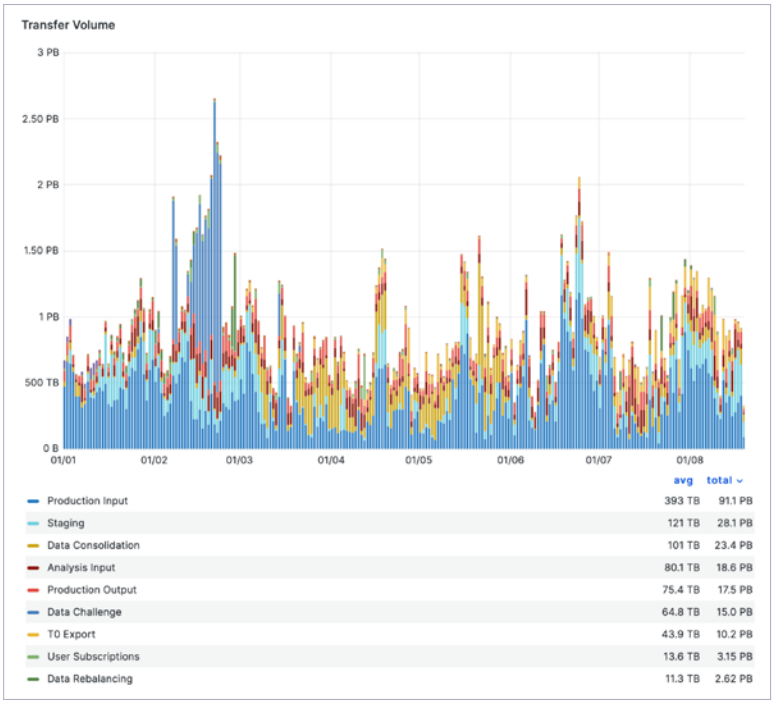


Figure 6.9.23: Transfer volume in 2024 so far in the US by activity

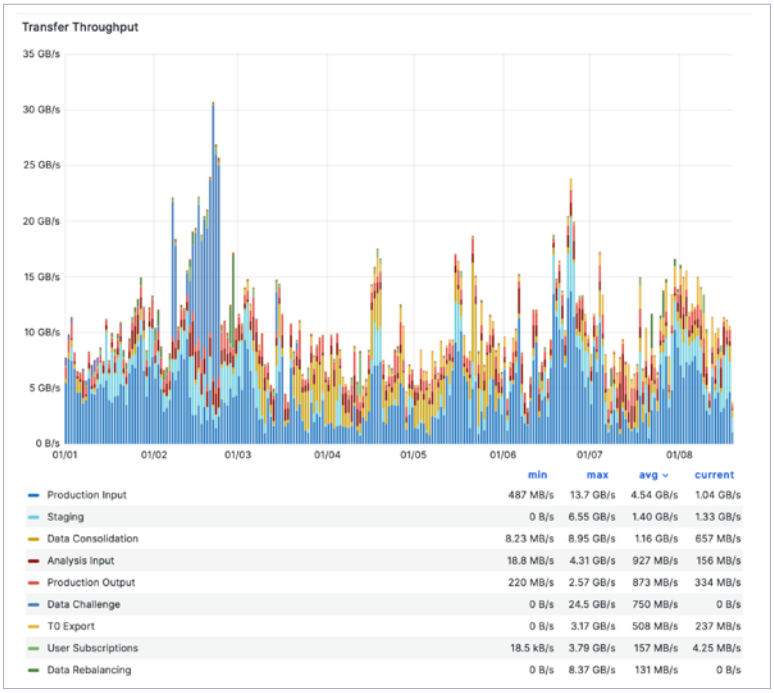


Figure 6.9.24: Transfer rates in 2024 so far in the US by activity

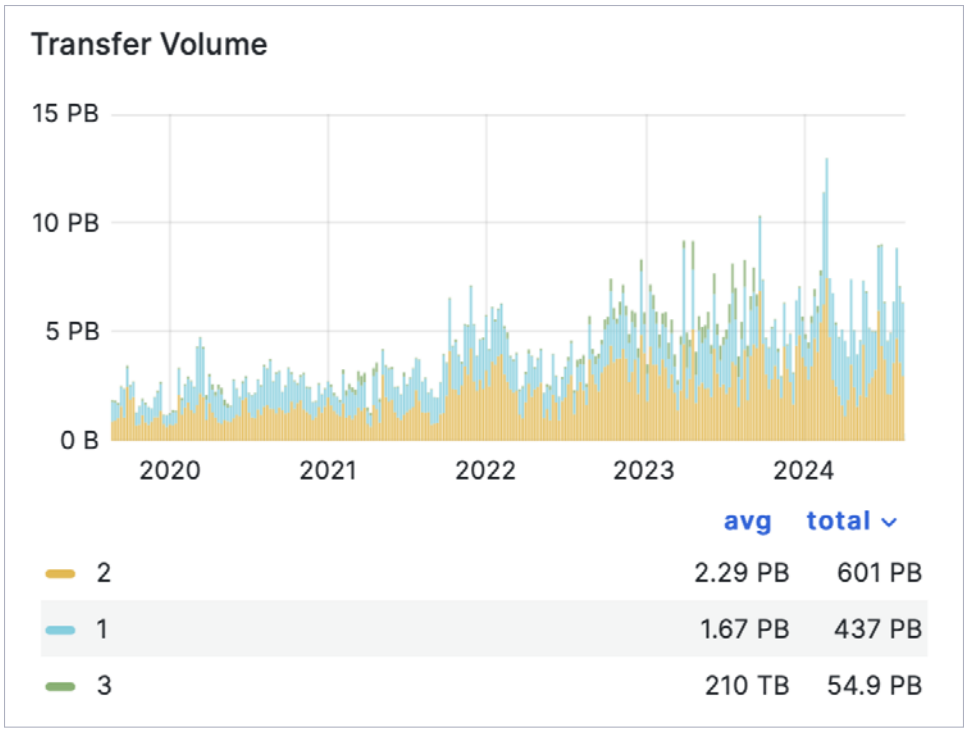


Figure 6.9.25: Data transfer (PB per week) over the last five years with US ATLAS grid sites as destination for each site category (Tier-1, Tier-2s, Tier-3).

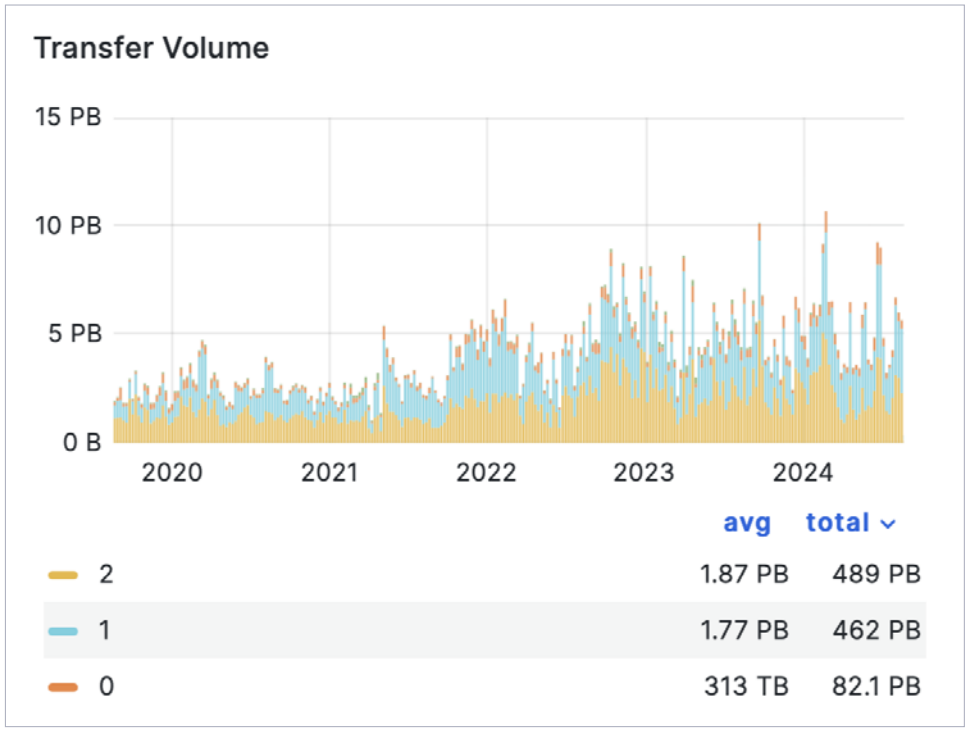


Figure 6.9.26: Data transfer (PB per week) over the last five years with US ATLAS grid sites as origin for each site category (Tier-1, Tier-2s, Tier-3)

6.9.2.6 Remote Science Activities

As described in previous sections, ATLAS computing is fully distributed, and therefore remote from CERN. The raw data are collected exclusively at CERN. All other activities leading to the final published products are done remotely.

6.9.2.7 Software Infrastructure

6.9.2.7.1 Rucio

The open-source software framework Rucio is used to organize, manage, and access the ATLAS data. The data are distributed across heterogeneous data centers at widely distributed locations worldwide. Rucio is built in different components:

A central database hosted at CERN contains the so-called dataset catalog. This catalog groups individual files of a detector run or MC simulation process into datasets together with their file location information. These datasets are stored on disk or tape storages at more than 100 sites worldwide.

Dataset files are not accessed from remote, but are moved between storages and sites asynchronously before a processing job request. Currently ATLAS is moving approximately 4–5 PB per day aggregated between sites worldwide.

The middleware to establish direct storage to storage transfers over the network, commonly called third-party copy, is provided by the FTS. FTS establishes connections between storage systems using the required protocols and ensures that the files are correctly transferred over the networks. Rucio decides which files to move, groups them in transfer requests, submits the transfer requests to FTS, monitors the progress of the transfers, retries in case of errors, and notifies the clients upon completion. If there are multiple FTS servers available, Rucio is able to orchestrate transfers among them for improved parallelism and reliability.

The input files for individual jobs are accessed through the LAN from a worker node to the local disk storage.

Remote access to analysis job inputs with network latency hiding is used at several ATLAS grid sites using XRootD and file caching though XCache at the sites.

The outputs of individual user analysis jobs in the order of 1 GB - 50 TB is downloaded to university clusters.

6.9.2.7.2 PanDA

The PanDA ecosystem manages all workflows and workloads in ATLAS. It is designed to handle complex multistep workflows, running over thousands of files, using many different application workloads, and optimization to process over a large number of distributed sites deployed worldwide. The main components of PanDA include ProdSys which translates the physics workflows to executable tasks. Deft provides an intelligent user interface to create workflows. Job Execution and Definition Interface (JEDI) transforms tasks into jobs that can run on single cores, single nodes or a single HPC. The execution steps managed by JEDI range from simulations, data processing, to distributed ML. Harvester provides an integrated interface to computing resources, with edge services that can transparently manage execution of workflows across all types of resources - grids, clouds and HPCs. Finally, the most important element of job execution is the PanDA pilot system, which manages job executions that are orchestrated by PanDA. All of these components are deeply integrated with the ATLAS DMS, Rucio.

The highly customizable PanDA system not only manages clusters and storage, but also optimizes workflow over existing networks. Over the years, many capabilities to optimize network performance have been built into PanDA and Rucio. However, direct integration with network layers have remained elusive over the past decade.

6.9.2.7.3 Frontier

The ATLAS model for remote access to database resident information relies upon a limited set of dedicated and distributed Oracle database repositories complemented with the deployment of Frontier system infrastructure on the WLCG. ATLAS clients with network access can get the database information they need dynamically by submitting requests to a squid server in the Frontier network which provides results from its cache or passes new requests along the network to the launchpads located at CERN since 2023. Since the beginning of LHC Run 1, the system has evolved in terms of client, squid, and launchpad optimizations but the distribution model has remained fundamentally unchanged.

6.9.2.7.4 CVMFS

The ATLAS software for simulation and reconstruction is distributed via the CernVM files system (CVMFS). CVMFS is implemented as a Portable Operating System Interface (POSIX) read-only file system in user space (a Filesystem in Userspace [FUSE] module). Files and directories are hosted on standard web servers and mounted in the universal namespace /cvmfs. Internally, CVMFS uses content- addressable storage and Merkle trees in order to maintain file data and metadata. CVMFS uses outgoing HTTP connections only. It transfers data and metadata on demand and verifies data integrity by cryptographic hashes.

In addition, ATLAS is currently exploring the usage of docker containers for software distribution. These containers are distributed via the CVMFS files system.

6.9.2.7.5 HTCondor

PanDA uses the HTCondor system for job submission to a majority of computing resources. This allows the Harvester to use a reliable and well-defined HEP standard. While Advanced Resource Connector (ARC), native cloud interfaces, or direct interface to batch systems are sometimes used, HTCondor is used for the vast majority of job submissions.

6.9.2.7.6 ROOT

ROOT is a software framework with building blocks for:

- Data processing
- Data analysis
- Data visualization
- Data storage

ROOT is written mainly in C++ (C++20 standard) and has bindings for Python available as well. It is highly adopted in high energy physics and other sciences but also in industry. More than 1 EB of data are stored in ROOT format.

6.9.2.8 Network and Data Architecture

6.9.2.8.1 US ATLAS T1 at BNL

BNL has implemented a vendor agnostic, resilient, scalable and modular terabit per second (Tbps) HTSN which serves as the primary network transport for all data-intensive collaborations at BNL. It provides high-throughput connectivity to all High-Performance Computing (HPC) and HTC collaborations and supports the timely transfer of large amounts of scientific data via the Internet.

The HTSN has five key components:

1. Network Perimeter
    a. Two diverse 800 Gbps (2 x 400 GbE) circuits that peer with ESnet routers within the BNL 515 and 725 data centers. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits.
    b. The BNL Network Perimeter on average ingests and egresses between 10-15 PB of data monthly.
2. Science DMZ
    a. Supports open, high-speed WAN/Internet) access for all scientific collaborations throughout the BNL campus.
3. Science Core
    a. A Tbps Science and Data Center Interconnect for data intensive collaborations at BNL. This Science Interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, sPHENIX, CAD, CFN, NSLS-II, High-Performance Computing Clusters and SDCC.
    b. Intelligence and routing policies are applied within the Science Core to restrict or grant access to specific resources within the SDCC.
4. Spine
    a. A Tbps network Spine that interconnects all Leaf switches. Leaf switches can consist of ToR or chassis-based switches that connect compute, storage or general infrastructure service servers.
    b. The responsibility of the Spine is fast packet forwarding and flexibility, not policy insertion or server termination.
    c. External Border Gateway Protocol (eBGP) is utilized throughout the HTSN. EBGP was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each Spine group shares the same ASN but does not have Internal BGP (iBGP) peering between them. Each Leaf or pair of Leaves will require their own ASN.
5. Storage Core
    a. A redundant terabit per second switching block that aggregates high-performance storage services.

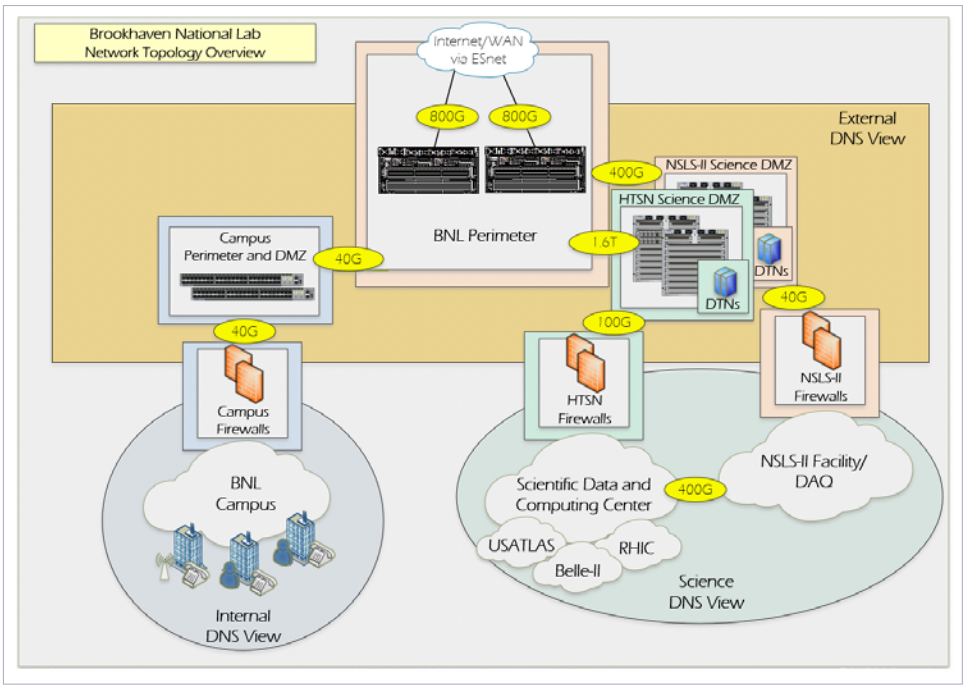


Figure 6.9.27: High-level overview of the BNL network perimeter and DNS architecture.

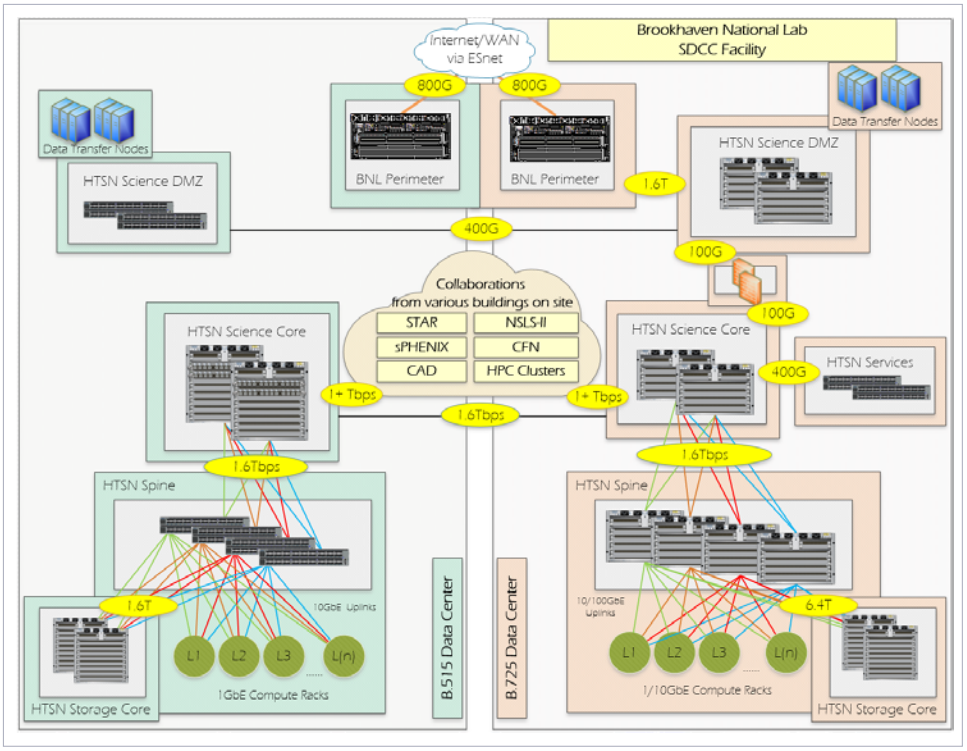


Figure 6.9.28: High-level overview of the BNL HTSN in FY24 (includes HTSN expansion into new data center, right hand side).

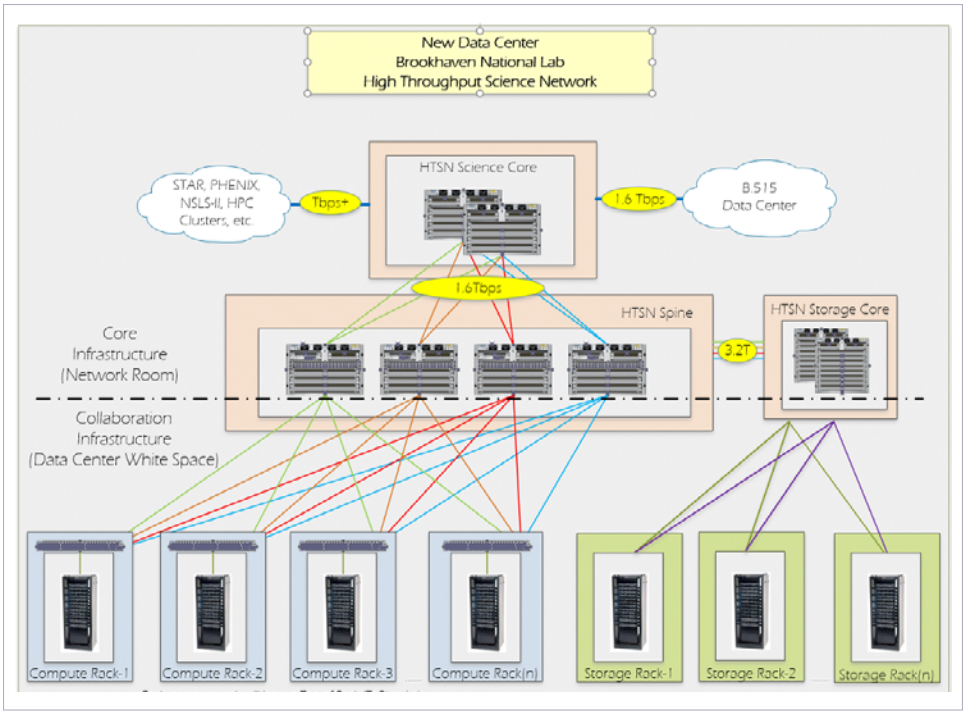


Figure 6.9.29: High-level overview of the BNL HTSN, demarcation points between core and collaboration infrastructures.

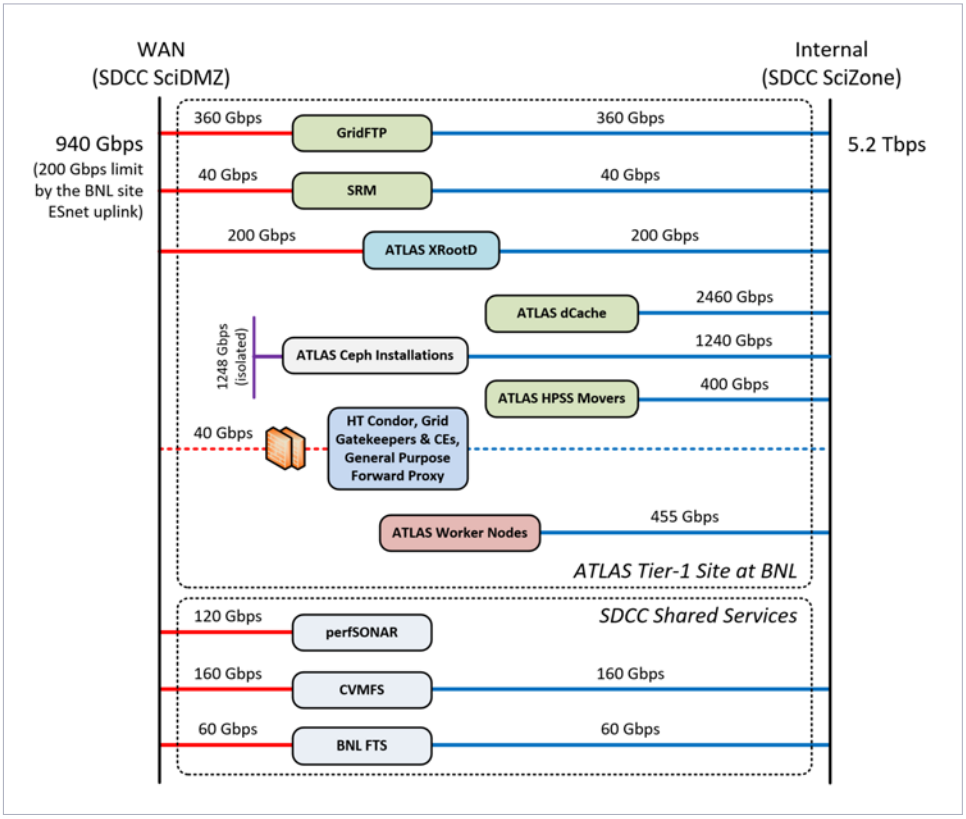


Figure 6.9.30: The high-level network layout of the ATLAS Tier-1 site at BNL site is shown below.

Reliability and performance of the network between main data centers are monitored with perfSONAR.

6.9.2.8.2 US ATLAS T2 Infrastructure

Each Tier-2 site has unique LAN/WAN architecture developed in coordination with local and regional network managers. ATLAS show some of the representative architectures below. The first three network diagrams below are for the various sites in the MWT2 starting with the UChicago site. MWT2 is the largest Tier-2 in US ATLAS, about 1.5 times the size of the other Tier-2's. The next three diagrams show the configuration of the network for AGLT2.

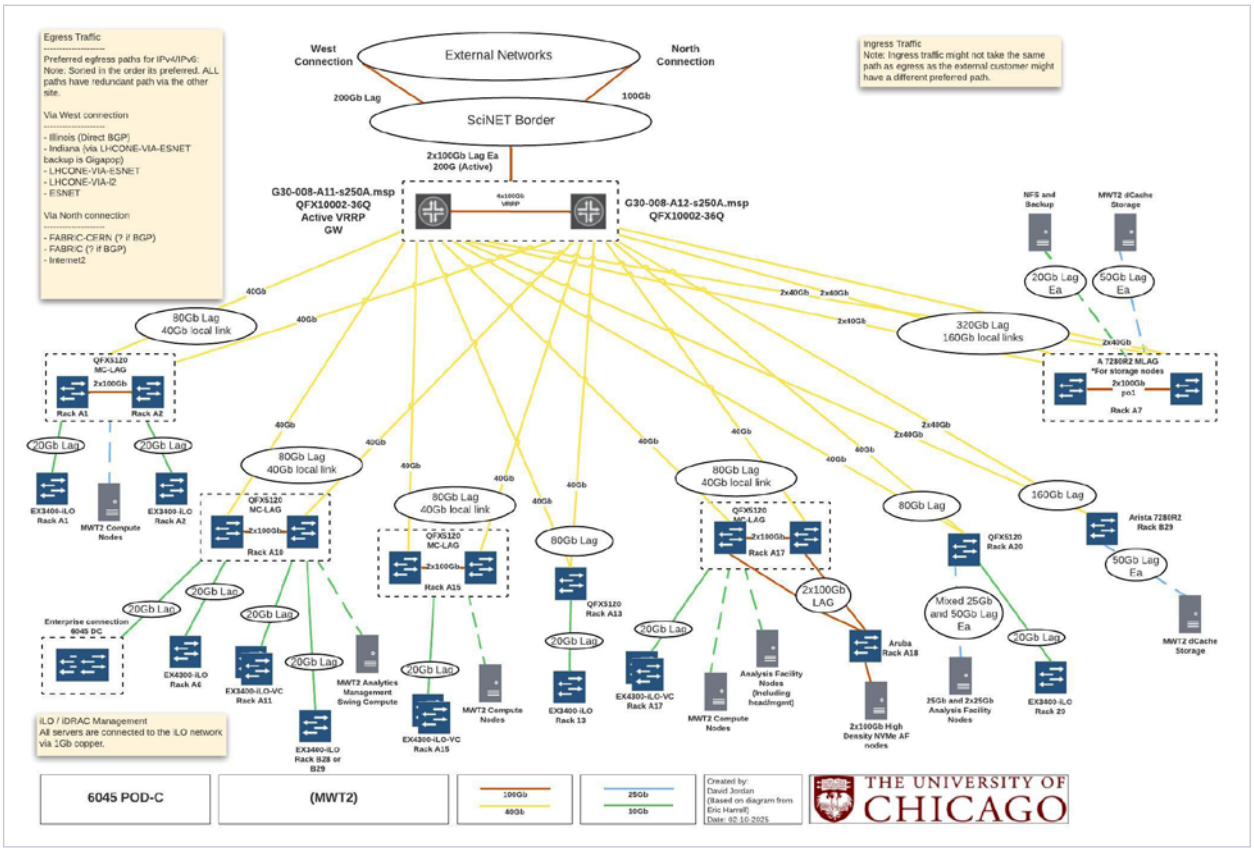


Figure 6.9.31: MWT2 UC Network Diagram

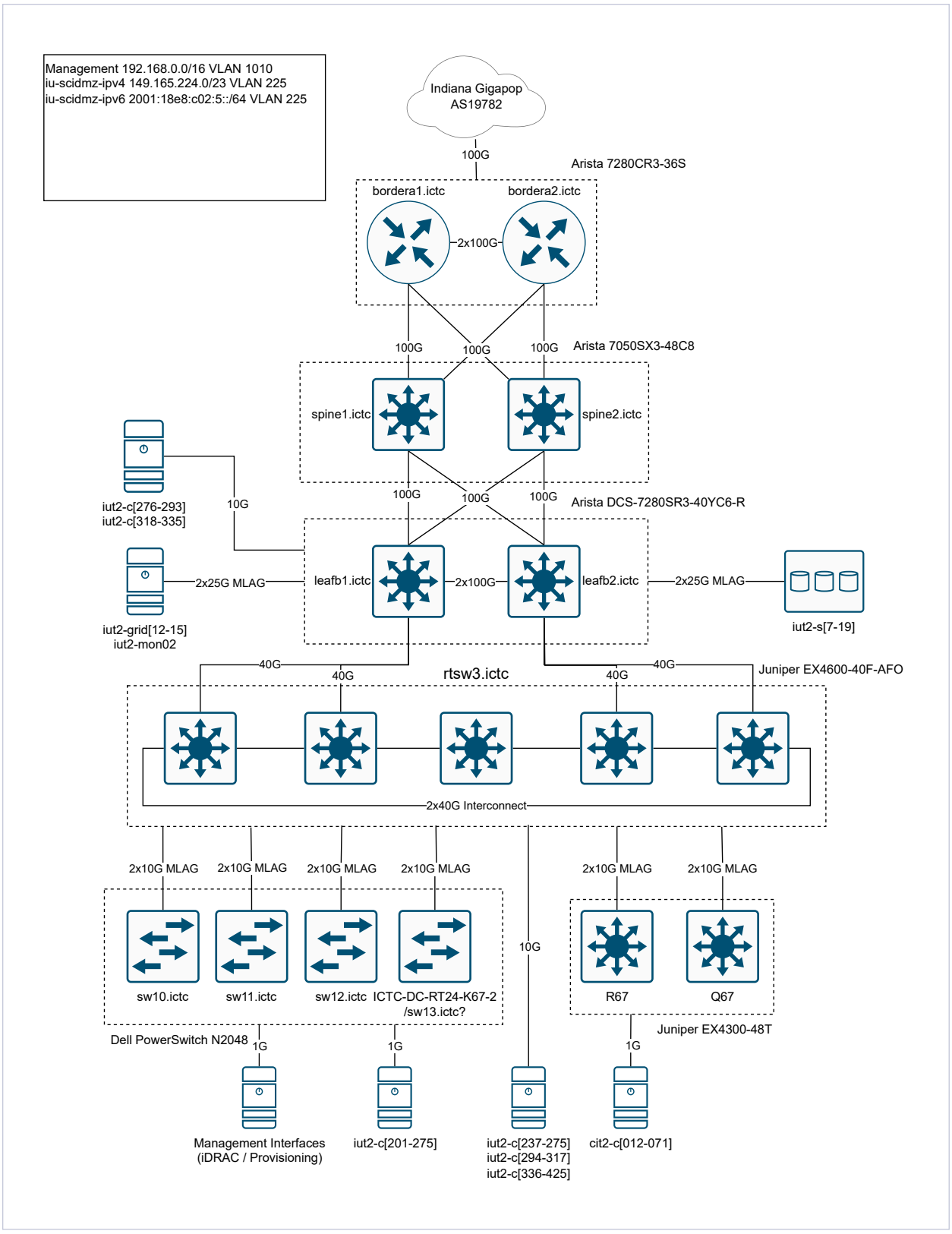


Figure 6.9.32: MWT2 IU Network Diagram

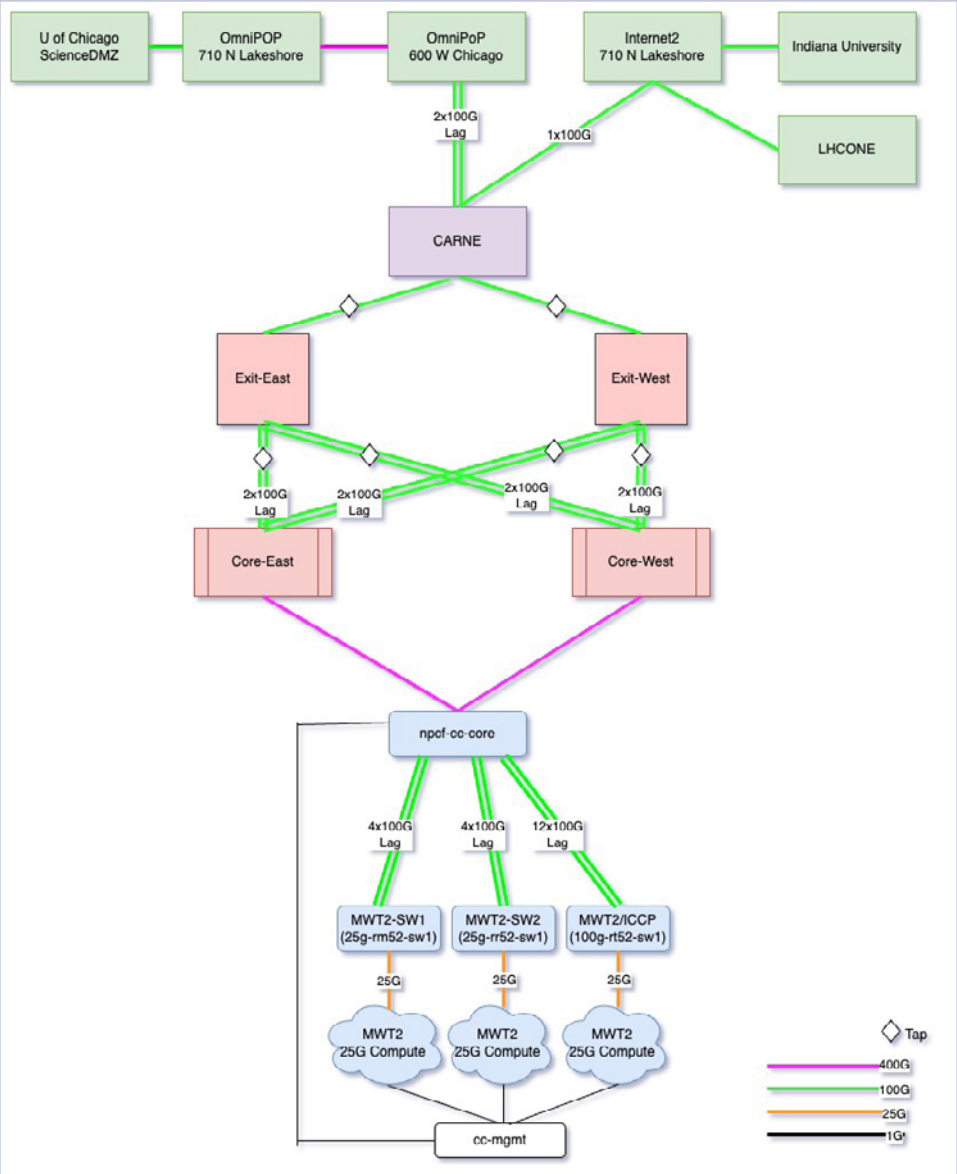


Figure 6.9.33: MWT2 UIUC Network Diagram

Similar diagrams from the ATLAS Great Lakes Tier-2 (AGLT2) are shown below.

This first diagram shows a high-level logical view of AGLT2 networking with the University of Michigan (UM) site shown in the top left, the Michigan State University (MSU) data center shown on the right and the Chicago WAN connection point shown in the bottom left.

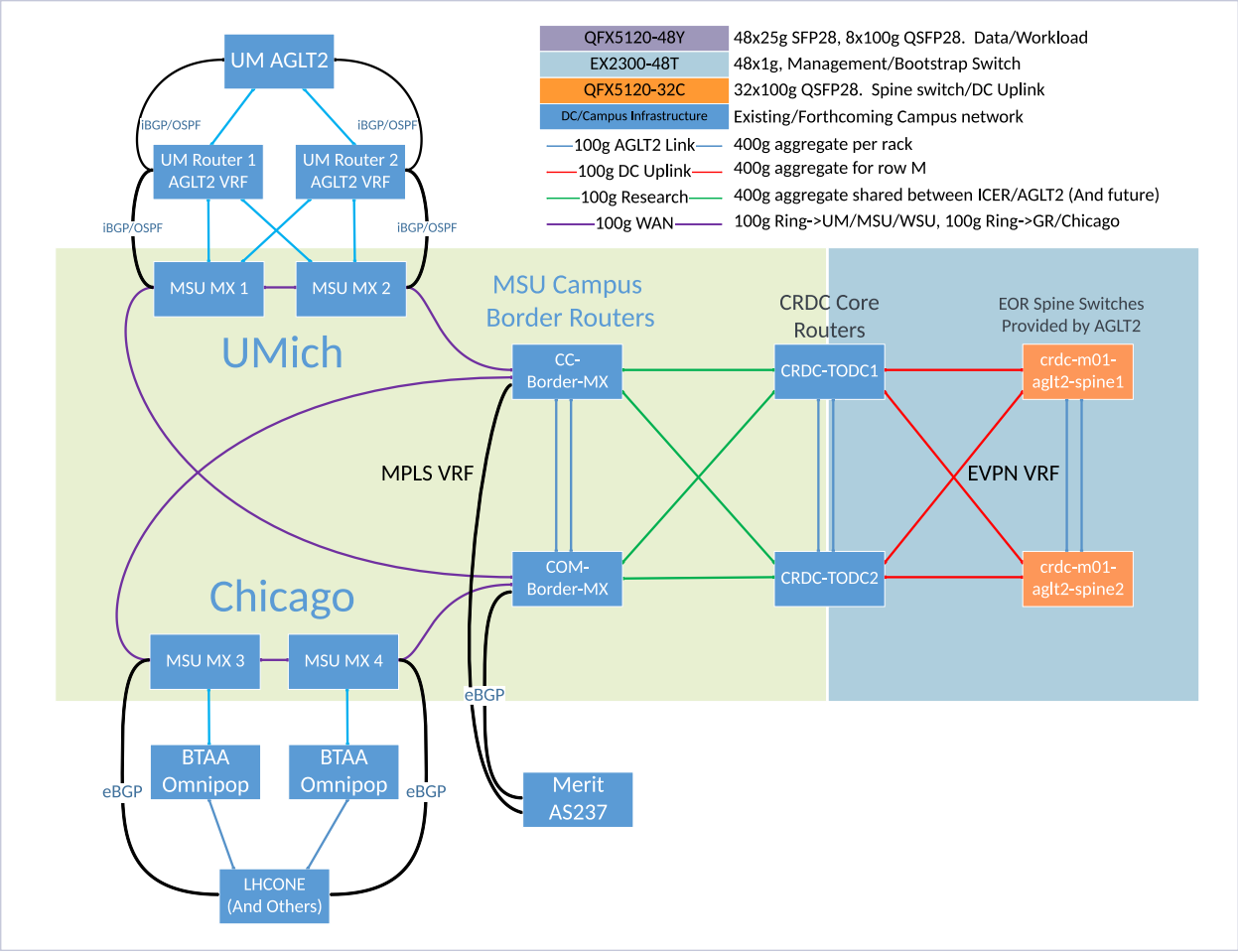


Figure 6.9.34: AGLT2 Network Diagram

A more detailed diagram of the MSU site LAN is shown below, with the AGLT2 equipment connected to the various Crescent Road Data Center (CRDC) routers in the center of the diagram.

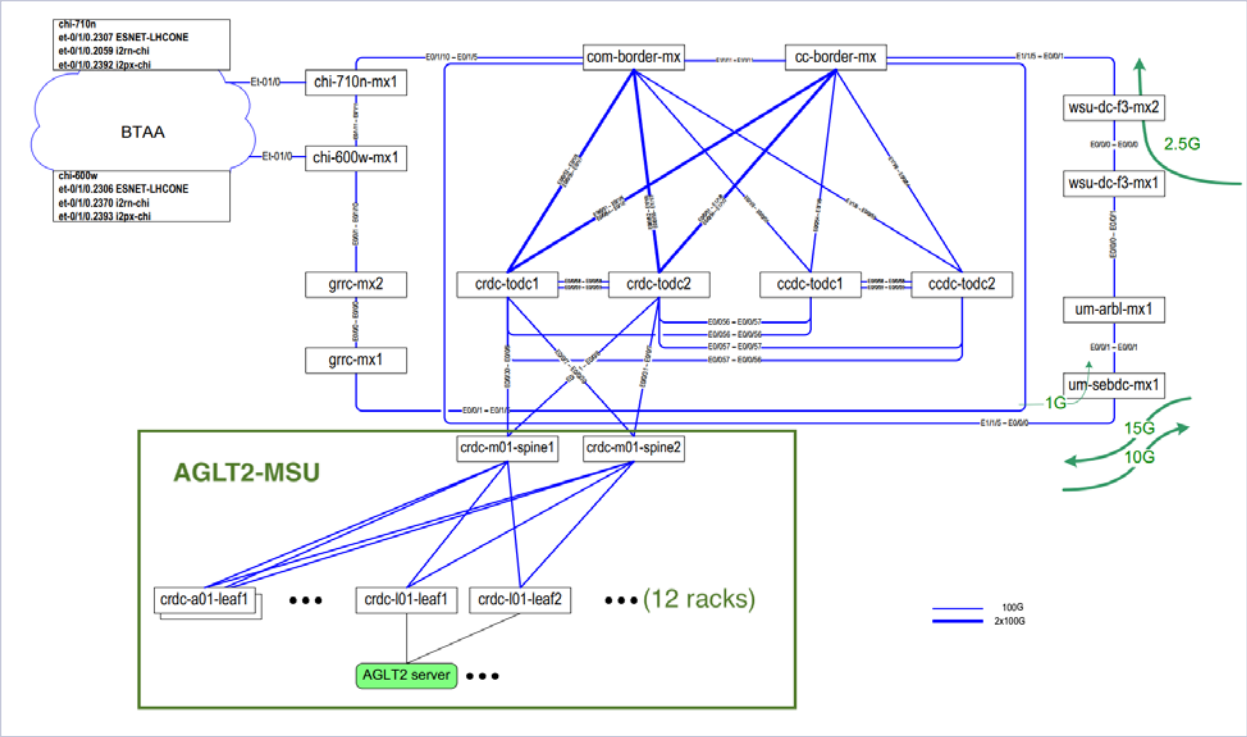


Figure 6.9.35: MSU Network Diagram

The UM AGLT2 site uses a set of Cisco router/switches providing WAN connectivity and each rack uses a set of Dell data and management switches to connect equipment.

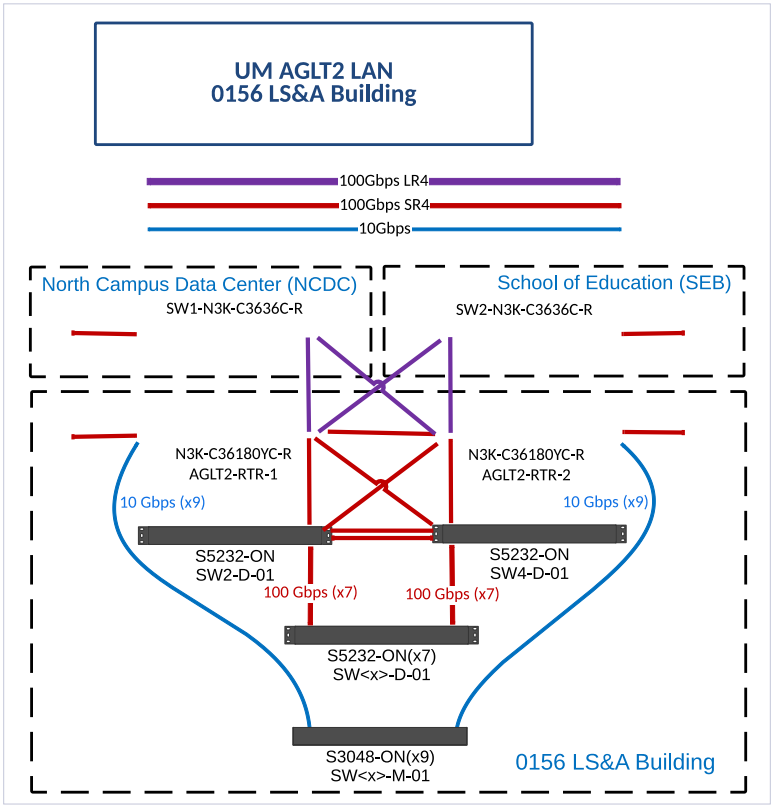


Figure 6.9.36: UM Network Diagram

6.9.2.9 Emerging Technology Use

6.9.2.9.1 Cloud Services

Google Cloud and AWS computing services have been successfully integrated within the ATLAS workflow and data management systems PanDA and Rucio. State-of-the-art technologies like Kubernetes, nonstandard processors like ARM or GPUs and cloud storage are used and have been operated at the scale of a current medium-sized Grid site. Scaling to more than 100k concurrently running vCPUs in PanDA has been demonstrated. The solutions may be considered for example in the implementation of AFs for HL-LHC, with all the added values

and services offered by the commercial cloud providers. In future work, data exports from cloud data centers to WLCG sites will have to be optimized, involving research network providers.

In 2024 the Google Cloud is used for MC production processing as extension of the SWT2 at the scale of up to 15k job slots as shown in Section 6.9.2.5.1[101].

In 2022 BNL (as some other labs) moved to using Microsoft Office 365 for email, calendar service, Sharepoint, and OneDrive cloud storage. Although, bandwidth to and from these services is much lower than that required for scientific data, the reliance of the laboratory on cloud services for day-to-day business has made the resilience of Internet connectivity critical. During periods of peak scientific data movement, cloud data for day-to-day work has been affected, and QoS and other traffic engineering controls were put in place to protect the business traffic from being affected by the scientific data flows.

6.9.2.9.2 Artificial Intelligence and ML

From the ATLAS Run3 software and computing paper: ML is used in a wide range of applications in ATLAS, including physics analysis, simulation and physics object identification within the reconstruction and trigger system. Models trained for classification and regression, as well as generative models, are core components of the software and analysis chain.

Although deep neural networks are reliant on GPUs for training, inference is performed almost exclusively on CPUs. In addition to supporting inference of modern ML models, training infrastructure is available through WLCG with support for job submissions using PanDA, including access to many GPUs. However, most models used by the collaboration are trained on private resources due to the modest resource requirements and satisfactory availability of institutional GPU resources.

ATLAS does not expect any significant changes in network load or the distributed computing software stack design during the ongoing analysis of Run 3 data.

6.9.2.9.3 Wireless or Distributed Location Science Activities

No wireless capability studies or capability upgrades are planned at present. Some users may do their work from systems connected via wireless, but this is not currently an area of concern for ATLAS because most such users use wireless to access remote systems that are doing the actual data transfers or access.

6.9.2.10 Data-Related Resource Constraints

The primary concern for ATLAS is lack of sufficiently fine-grained network monitoring and performance information to help us debug and automate data transfers in real time between various sites. Better visibility is being worked on through various entities including WLCG Site Network Monitoring, the SciTags initiative (labeling network flows and IPv6 packets with owner/purpose) and improved monitoring from the various tools and infrastructures, but ATLAS still have significant gaps in monitoring.

101 https://link.springer.com/article/10.1007/s41781-024-00128-x

6.9.2.11 Data Mobility Endpoints

BNL SDCC operates Globus and FTS endpoints. These File transfer services have access to the HEP storage systems such as dCache which is connected to the ESnet network at the bandwidth of 1.6 Tbps. Together with ESnet and various site network and storage administrators, FTS service was used to generate the network load during Data Challenge 24 and various mini-Data Challenges in the preparation for the Data Challenge.

A related note is that the idea of DTNs were not the basis for how ATLAS distributed data management was originally architected. Instead, the focus was on using many flows from many servers at each end of a site-to-site transfer. The distributed data management infrastructure could benefit from incorporating tools and architectures targeting higher performance, but the cost-benefit analysis needs to be continually evaluated.

6.9.2.12 Outstanding Issues

The primary wish list consists of capabilities to monitor and manage data transfers automatically. Given the size, complexity, and fully distributed nature of ATLAS computing, all workflow and data distribution needs to be optimized and managed with AI.

ATLAS anticipates network needs will grow gradually, following the patterns described in this document. Increasing network capacity and performance will be needed at US ATLAS Tier-1, Tier-3, and Tier-3 AFs. The exact magnitude of future network provisioning is expected to be determined jointly by US ATLAS, ESNET and other involved parties.

6.9.2.13 Facility Profile Contributors

ATLAS Representation

Johannes Elmsheuser, BNL, jelmsheuser@bnl.gov

Kaushik De, University of Texas, Arlington, kaushik@uta.edu

Alexei Klimentov, BNL, aak@bnl.gov

Shawn McKee, UM, smckee@umich.edu

Frederick Luehring, Indiana University, luehring@indiana.edu

ESCC Representation

Vincent Bonafede, BNL, bonafede@bnl.gov

Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

6.10 LHC: CMS Experiment

This case study functions as a high-level introduction describing the process of science within the larger context of the global CMS collaboration: how data are produced, how science is derived from these data, its computing facilities, and its requirements process. This case study also provides more details about the networking implications of performing LHC particle physics specifically at the Tier-1 site at Fermilab, the seven Tier-2 sites at US universities, and the numerous other "sites" used to process data and to perform analysis. CMS researchers also utilize campus or research group facilities at about 50 US institutions, the largest of which are Tier-3 sites, as well as individually managed computing resources of each of the approximately 700 researchers within US CMS.

6.10.1 Discussion Summary

DC24 was successful for the CMS experiment, and several important outcomes came from the exercise. Understanding bottlenecks in the network configuration at Fermilab were important, along with the way that many of the tools that were improved and replaced since LHC Run 2. CMS would be interested in developing a program to test the effectiveness of transfers both to benchmark methods at increased capacity, and integrate new functionality into existing tools. Such tests should be done in collaboration with ESnet. For example, one lesson learned from the most-recent data challenge organized by WLCG (DC24) was that it is necessary to utilize a link at full capacity, to learn if a given link has any performance abnormalities. It is desirable to develop a process of regular testing of network capacities (e.g., relying on perfSONAR) to test connections to the maximum and identify problems before they become a bottleneck in production.

Changes made to analysis formats have also had a large impact on network traffic. The NanoAOD format, which was introduced late in Run 2, is becoming more widely used. The format is smaller, and thus occupies less network bandwidth and can run faster on CPUs. More than 50% of all physics analysis activities are now based on NanoAOD, with another 40% or more to be based on MiniAOD. The use of the larger AOD format is now minimal, despite occupying a lot of disco space. At future HL-LHC scales, CMS expects not be able to afford to keep AOD on disk any more. These may be relegated to tape archives.

The data rates that CMS originally estimated for Run 3 (i.e., in 2020 just prior to run beginning) were smaller than what was witnessed by a factor of two to three times beyond the 2018 values, and four to six times beyond the 2017 values. This increase in size has not translated directly into increased traffic on network connections; however, some of the R&E efforts to improve data mobility tools and workflow management has helped to mitigate big spikes in traffic.

The use of Caching has had some impact on network use, and CMS credits ESnet's role in the development and deployment as being critical to understanding what is possible and how it can be leveraged for Run 3 and the latter run comprising the HL-LHC operating period. Much of the R&D efforts in the recent years have been determining where caches may be beneficial for placement, how large they need to be to be effective, and what strategy should be used to store and replace data sets. Nevertheless, there are still open questions about deployment strategies that need to be worked out for both CMS and ATLAS. Regional caches were found to not be as effective (since the overlap of data use was not high enough), but caches within a facility, and ensuring that the workflow management tools know to check the caches, produced the most reliable results. The adoption and more widespread use of the NanoAOD format is anticipated to have some impact on cache performance over time.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The CMS experiment at the LHC is designed to probe new phenomena at the energy frontier. The CMS collaboration is made up of more than 3000 members from more than 50 countries. Researchers at US institutions comprise about 30% of the collaboration. [Section 4.1]

- CERN facilities are used to provide the initial CMS data reconstruction, to support detector commissioning work as well as physics analysis work. Then globally distributed Tier-1 and Tier-2 facilities are responsible for data archiving (at the Tier-1 only), simulated data generation, analysis data storage, and physics analysis activities (primarily at the Tier-2s). The US operates one Tier-1 facility (Fermilab) and seven Tier-2 facilities (Univ. of Florida, UC-San Diego, Caltech, Massachusetts Institute of Technology [MIT], Univ. of Wisconsin, Univ. of Nebraska, and Purdue University). Data also moves from these facilities to other universities as analysis data sets are reduced and refined during the analysis process. [Section 4.1]
- Analyses are done by small and large groups that share data either at a CMS TIer-2 center, at a local Tier-3 or on other local computing facilities. An "average" event collected by CMS is processed 50–100 times by analysts. That means that CMS analysts are processing hundreds of PB each year just to carry out the initial analysis steps which process data created by the production system. [Section 4.2]
- CMS devotes 30% of its disk storage to derived data sets, or around 100 PB. These data are typically used much more frequently as the size for any given analysis is smaller, and can be more quickly processed as the analysis process evolves. [Section 4.2]
- CMS distinguish three types of data: Data from collisions in the accelerator, data from simulations, and user produced data. The first two are centrally produced, while the third is produced by the physicists themselves as part of their analysis workflow. [Section 4.2]
- The LHC provides data to CMS in roughly ten-hour data taking periods with minimal downtime (typically a few hours) between periods. As such, a roughly constant stream of data is available during operation. Event rates of around 1 kHz of physics events are typical today and for Run 3, and 7.5 kHz is expected for HL-LHC. This translates into raw data volumes of: [Section 4.2]
    - Run 2: 45 PB written during 4 years (2015-2018)
    - Run 3: approximately 120 PB (2022-2024)
    - Run 4: 270 PB/yr (estimated in 2022)
- CMS has produced simulations that scale in size with the luminosity collected. CMS has about 70 PB (170 PB) of Run 3 (Run 2) MC simulation datasets stored on tape, representing around 75 billion (250 billion) MC events. [Section 4.2]
- The centrally produced data from CMS comes in multiple formats ranging from the most versatile and complete (raw and AOD) to the easiest, smallest, and fastest to use ones (MiniAOD and NanoAOD). CMS has been moving to toward the use of NanoAOD, a data format designed for interactive end-user analysis, and produced centrally. CMS adoption of this format has already grown in Run 3 to around 50% of analyses using the NanoAOD as their primary data tier. [Section 4.3]
- The CMS Tier-0 is at CERN, and has two primary functions. It is responsible for "prompt calibration and reconstruction" and archiving of all the raw data at CERN. When the LHC is not running, the Tier-0 compute cluster may be used for other processing by CMS. [Section 4.3]
- CMS globally has 7 Tier-1s, including Fermilab in the US The collection of the Tier-1s are responsible for operating a distributed archive of all centrally produced data, including a second copy of the raw. The copy of the raw across the Tier-1s is considered the "active copy," i.e., the raw copy at CERN is considered strictly a backup. In addition, the Tier-1s provide computing resources to the experiment that are primarily used for centrally organized processing. The US CMS Tier-1 Center at Fermilab is the largest of CMS Tier-1 facilities, currently providing 40% of the total CMS Tier-1 capacity. [Section 4.3]

- CMS has roughly 50 Tier-2s globally. The collection of the Tier-2s is responsible for all data analysis, and some simulations. Tier-2s also provide disk space to host data for analysis, and a moderate amount of disk space for staging in data for processing. [Section 4.3]
- Tier-3 sites are typically university-hosted clusters that are smaller than Tier-2 sites and are typically without dedicated support to ensure high resource availability. Those in the US are not funded through the Operations Program. Tier-3 sites can be part of the DMS of CMS and receive official datasets, or can access data through streaming. These sites can also provide disk space to hold user data as required by the local community. [Section 4.3]
- CMS uses compute resources beyond the tiered computing centers. These include HPC centers in the US During 2023, the CMS usage of shared HPC facilities funded by DOE and NSF amount to about 10% of the overall computing budget globally. This has increased substantially in the recent past: by around 4x in relative fraction of CMS compute since 2020 (and a substantially greater increase in absolute terms), and has substantially increased again in 2024. CMS expects that this model will change for the HL-LHC to include larger allocations on HPC facilities. The size and locations of future facilities is not fully predictable; however, CMS would rely on excellent network connectivity to any HPC facility from its Tier-1 and Tier-2 facilities. [Section 4.3]
- HPC centers funded by DOE and NSF as shared facilities across multiple science domains do not generally provide the full functionality required from a Tier-2 or Tier-1 center, e.g., they normally do not provide any substantial amount of storage under the control of the experiment. Data access at HPC centers relies on either temporarily staging in input data on storage at the HPC site, or streaming access if the site setup allows for access to the wider network. In general, HPC centers only run compute dominated applications that have modest IO/CPU requirements, e.g., the central production workflows involved in central processing and simulations. Output is archived back to the Tier-1 sites, either directly from the application or through other mechanisms. [Section 4.3]
- CMS currently uses Rucio for bulk-data transfer activities, having fully switched from its previous system based on PhEDEx and Dynamo tools, to manage dataset storage and dataset transfers. Rucio has become a community solution for data management. Rucio uses the FTS to manage scheduling and file transfer. For bulk transfers via Rucio and executed by FTS, US CMS has moved to TPC https implemented via XRootD servers and away from gridFTP. [Section 4.3]
- Failures in data streaming are a large source of job failures in CMS. This has also become an issue for analysis-facing activities, where a greater variety of tools are used to access CMS data. While CMS can instrument its centrally maintained software with retry mechanisms and other fallback logic, it is more difficult to extend this to community-developed tools. Reliable and high-capacity streaming of input data, either raw or pileup simulation, considerably reduces the disk requirements of CMS at HPC and other nondedicated computing facilities. [Section 4.3]
- CMS acknowledges that the data volume that can be handled by the networks within and coming out of the CMS detector facility far exceeds what can be handled offline within current CPU, storage and networking infrastructures. This bottleneck will remain for the foreseeable future, as network growth patterns move faster than storage and computation. [Section 4.4]
- The LHC computing and storage infrastructure is organized in a tiered computing model including four tiers, 0-3. Originally, when the connectivity among the sites was extremely limited, the tiers were hierarchical in that Tier-N+1 was connected only to an associated Tier-N above and associated Tier-N+2 below. Data was not allowed to flow from Tier-N+2 to Tier-N. This rigid concept has been eliminated over the last decade as network capacity and capability progressed. [Section 4.4]

- The global CMS collaboration, together with WLCG and OSG, will define services that sites perform. Today the service profile defines the site to belong to a certain tier. In the future, centers of a given tier today may no longer provide all the services that today would be expected from that tier. In addition, it is likely that the HL-LHC data and processing infrastructure will no longer support the full global mesh of data flows among all tiers. These changes are part of an ongoing R&D program. [Section 4.4]
- CMS applications and the infrastructure that distributes and launches them support multiple types of data access: [Section 4.5]
    - Traditional top-down data placement at Tier-1 and Tier-2 centers combined with applications specifying the dataset they need, and being automatically routed and executed at the sites that have it.
    - CMS supports streaming data access to any data on disk across its the WLCG facilities from any location with an internet connection at any time. This is called AAA, for "Any Data, Anytime, Anywhere" within CMS.
    - Bottom-up data placement, as is implicit in caching. Here the applications are routed to sites with caches, applications access the cache locally, and cache misses are handled by the CMS XRootD Data Federation.
- Having reliable mechanisms to transfer data into HPC facilities from the CMS Tier-1 and Tier-2 centers is of growing importance. HPCs are different from other resources used by CMS primarily because they provide significant compute (CPU and GPU) resources but very limited disk resources, and because network access to or from HPC compute nodes is typically limited or not allowed. The broader LHC community has considered how to best utilize these resources within operational constraints. Unlike most HPC users, CMS cannot concentrate its efforts on using one or a handful of HPC facilities, but is instead asked to use HPC facilities worldwide. Network access and data management are important aspects of this. [Section 4.6]
- CMS is currently a major user of the TA network link. Activities including the archival of raw data, initial copies of analysis data samples, and user-derived data sets using CERN or other non-US facilities are clear use cases for TA network usage. CMS tools do not prioritize site proximity (in the networking sense) when scheduling data transfers. Streaming data across the TA link is allowed (even if discouraged). [Section 4.7]
- The US CMS Tier-1 facility at Fermilab has local network infrastructure consists of fully redundant, high-performance switching fabric distributed across multiple Fermilab data centers. That switching fabric is based on 400 GE network technology for inter-switch connectivity, with 10 GE and 100 GE connectivity available for host system connectivity. The LAN provides approximately 10 Tb/s of network capacity in total. [Section 4.7]
- Fermilab's WAN architecture utilizes LHCOPN connectivity to support movement of raw data from the T0 (CERN), as well as production data movement with some of the other CMS T1s. In terms of aggregate WAN capacity out of the site, Fermilab currently has provisioned 2.4Tb/s of link capacity to ESnet, via geographically diverse connections. A LAG of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports the Laboratory's general internet traffic, which includes some portion of science traffic. [Section 4.7]
- Much of the LAN and WAN networking infrastructure to and within a CMS Tier-2 site has been an institutional contribution to the US CMS Operations Program, i.e., it has not for the most part been directly funded by the program. [Section 4.7]

- All CMS Tier-2s connect to the LHCONE at least 100 Gbps, through partnership with regional networking infrastructure. In some cases, a Tier-2 facility may have multiple connections either as a backup, or for active use. [Section 4.7]
- CMS does not currently use cloud resources to any significant extent. Previous studies have shown them to be a more costly model than the owned-resource model that CMS currently relies on. A notable exception would be the case of needed resource bursts for either CPU or network. The resources available for either CPU or TA networking in the cloud far exceed those available to CMS. At this point, CMS tools are generally able to use cloud services, typically via infrastructure at one of the tiered sites, but CMS does not have plans to use cloud services extensively in the near or longer term. This can evolve depending on largely external factors (e.g., cost evolution, or government policies). [Section 4.8]
- CMS has an increasing use of AI/ML in its production workflows. Analysis use cases have evolved even more quickly as there is greater freedom to adopt new toolkits and approaches. Both the Fermilab Tier-1 and Tier-2 facilities have GPU capabilities for CMS use, and the CMS community also relies on national resources. CMS is also researching a dedicated AI/ML training facility (at Vanderbilt) in recognition that AI/ML brings two distinct use cases, training and inference, that both require substantial resources (but in differing access patterns). CMS expects to predominantly use on-site and HPC resources rather than commercial cloud. [Section 4.8]
- Reliable networking can be used to reduce disk replica requirements either by the use of tape recall or caching. By the end of Run 4, a copy of the entire CMS MiniAOD will be around 100 PB. If 10% of this is used during any given month in a caching system, one can estimate the need for 10 PB/month of transfers to keep the cache up to date with the most recently used data. Cache network needs will be typically bursty because users read full data sets rather than just a file or two from one. Understanding and expanding caching use cases and needs are part of ongoing R&D. It is expected that investments into disk space at Tier-2s against network utilization will become more critical in the long term. Other forces may also push in the direction of higher network requirements, such as how spinning disk technology evolves in terms of cost per terabyte and I/O capabilities relative to CMS needs. [Section 4.8]
- US CMS has proactively pursued opportunities at US HPC sites, and currently has many allocations available for use. Typical workflows run on these allocations are similar to the IRI concept of a "long-term campaign," that relies on sustained access at scale. New HPC resources typically come with some, potentially substantial, work to set up given facility specific configurations and constraints. A common interface (or access requirements) across DOE HPC sites, as well as models for sustained access, are potential IRI benefits. [Section 4.9]
- CMS would like to develop a program to test the effectiveness of transfers both to benchmark methods at increased capacity, and integrate new functionality into existing tools. Such tests should be done in collaboration with ESnet. For example, one lesson learned from the most-recent data challenge organized by WLCG (DC24) was that it is necessary to utilize a link at full capacity, to learn if a given link has any performance abnormalities. It is desirable to develop a process of regular testing of network capacities (e.g., relying on perfSONAR) to test connections to the maximum and identify problems before they become a bottleneck in production. [Section 5.7]
- CMS has treated the global network of sites as a mesh with identical links when it comes to bulk transfers. The XRootD data federation was designed from the beginning to be cognizant of the TA links being limited, but treated links within the US as identical. The data lake model currently discussed in WLCG makes clear regional distinctions. The concept of creating a data lake has progressed slowly due to challenges around how increased latency due to the longer international links will affect application performance. [Section 5.8]

- CMS would like to continue exploring with ESnet how to contribute towards global R&D projects such as the Global P4 Lab testbed, use of ESnet High Touch services, and evolving technologies such as Smart NICs. [Section 5.8]
- The next several years are critical to transition managed network R&D by CMS, including results of the Software-Defined Network for End-to-end Networked Science at the Exascale (SENSE) project, into production for the HL-LHC. The SENSE, Caltech, and UCSD teams are close to 400G throughput with production infrastructure using XRootD, FTS and XRootD. [Section 5.8]

6.10.2 CMS: Experiment Profile

The content presented in this case study is written from the perspective of networking and data. CMS identifies two primary periods: Run 3 and Run 4. These map roughly on the time periods 2022-2025 and 2029-2031. The time period from 2026 through 2028 is a transition period where the detector, together with the LHC accelerator complex, is undergoing a major upgrade that is presently hard to predict in terms of computing activities. It will probably include some large-scale data challenges that are discussed in the third case study.

The CMS experiment at the LHC is designed to probe new phenomena at the energy frontier. Since starting operations in 2010, CMS has collected nearly 200 fb-1 of luminosity at center-of-mass energies between 7 and 13 TeV.

6.10.2.1 Science Background

The CMS collaboration is made up of more than 3000 members from more than 50 countries. Researchers at US institutions comprise about 30% of the collaboration. The collaboration has published more than 1000 papers with scientific findings across a broad physics program enabled by all collaborators having access to the entire data sample for their work. No previous HEP experiment has produced this many publications. The success of the physics program of CMS depends on the availability of sufficient computing resources to store, process and analyze the data in an efficient fashion.

The LHC provides proton-proton bunch crossings (where each crossing produces many proton-proton interactions) in CMS at a rate of nearly 40 MHz at a duty cycle of about 50% during 8 months of a typical running year. CMS commonly uses the word "event" to refer to all the data that corresponds to a single bunch crossing, both for collision data and simulated data. The interactions that result from each beam crossing are independent of those from other beam crossings. This means that CMS can reduce, model and analyze each event as being independent from all other events.

40 MHz worth of detector data (called "raw" data in this document) cannot all be saved for analyst use, in part because raw events are nearly 1 MB in size. CMS uses a two-step filtering system ("Level-1" and HLT) to reduce the data rate down to approximately 7.5 kHz of events to be processed and made available to the collaboration for analysis. This event rate has increased considerably since Run 2, and during Run 3, driven by physics opportunities and available computing resources.

Part of this online data reduction process is to broadly categorize events into "datasets" according to the physics signatures observed during the filtering process. Datasets are also separated according to the data-taking period to ensure that any large changes in LHC conditions, CMS detector operations or software used can be treated separately by analysts. A typical CMS publication requires analysis of a few datasets of detector data, and a few dozen simulation datasets for each year of data taking it includes. Larger endeavors may require up to 50% of the entire detector data, and corresponding simulations. Analysts nearly always aim to analyze as much of the available integrated luminosity as possible (e.g., using as much run time as possible). There are more than a hundred analyses ongoing in CMS at any given time, each engaging between a handful and up to dozens of researchers who are actively involved in the analysis of data (either directly from production, or using samples derived from it).

CMS executes a variety of tasks on its distributed computing infrastructure, for reconstructing collision data recorded by the detector, simulating collision data, and analyzing both. The CMS computing system relies on infrastructure distributed worldwide, and as such relies heavily on excellent network connectivity among its dedicated computing centers, and increasingly its connections to shared facilities at universities, high-performance computing (HPC) centers and potentially cloud resources when economically viable. Beyond the substantial real-time facility for data taking that is co-located with the CMS detector, CERN facilities are used to provide the initial data reconstruction, to support detector commissioning work as well as physics analysis work. Then globally distributed Tier-1 and Tier-2 facilities are responsible for data archiving (at the Tier-1 only), simulated data generation, analysis data storage, and physics analysis activities (primarily at the Tier-2s). The US operates one Tier-1 facility (Fermilab) and seven Tier-2 facilities (Univ. of Florida, UC-San Diego, Caltech, MIT, Univ. of Wisconsin, Univ. of Nebraska, and Purdue University). Data also moves from these facilities to other universities as analysis data sets are reduced and refined during the analysis process.

In a nutshell, one may think of the science publication process of the CMS experiment as starting with detector operations to collect data, then central production to process and do initial data reduction and validation, followed by a decentralized (but coordinated) data analysis process that derives results by using progressively improved reconstruction and analysis methods, as well as new approaches and ideas that come from across the collaboration. This is followed by a convergence of those results through multiple stages of discussion and sometimes refinements of the analysis, leading to publications via a centrally organized internal review process. Publications are then passed onwards to the peer review processes of the respective journals.

The CMS production and analysis process is quite storage and data-movement intensive. Data sets are replicated, moved or deleted according to their usage patterns, site and resource availability and experiment priorities. Individual analysts or analysis groups create and manage further derived data sets for their analysis workflows. Analyses are done by small and large groups that share data either at a CMS TIer-2 center, at a local Tier-3 or on other local computing facilities. An "average" event collected by CMS is processed 50–100 times by analysts. That means that CMS analysts are processing hundreds of PB each year just to carry out the initial analysis steps which process data created by the production system.

CMS devotes 30% of its disk storage to derived data sets, or around 100 PB. These data are typically used much more frequently as the size for any given analysis is smaller, and can be more quickly processed as the analysis process evolves. In addition to these resources, final analysis steps are done locally on the laptops, desktops or computing clusters of individual researchers or groups. Reliable distribution of data to these researchers is essential to the CMS analysis process. It enables researchers to work outside of a centrally managed and maintained system, and it thus reduces the complexity and management difficulty that the central system would otherwise have to face.

The infrastructure to support this process is largely developed and operated by CMS experts, including a significant number of computing professionals as well as physicists that have specialized in software and computing technologies. However, the experiment strives to be sufficiently open in the infrastructure such that computing and storage resources at individual universities can be integrated and used in both the central production and decentralized analysis parts of this process.

As the entire CMS data flow and analysis process relies on collaboration and data sharing, the CMS research program and methods are discussed in the next section.

6.10.2.2 Collaborators

The CMS experiment is designed, built, and operated by a collaboration of close to 200 institutions across more than 50 countries, and comprises roughly 3,000 members, of which close to ⅔ are physicists with authorship privileges on all CMS physics papers. The US makes up about 30% of the authors at about 50 institutions.

The collaboration as a collective produces official data samples via a central production team. Applications composed of algorithms developed by the collaboration convert and reduce the raw detector data into derived formats ready for analysis. These data are distributed across the CMS computing facilities on reliable disk, with topical and popular data sets having the greatest accessibility (e.g., largest number of replicas across the distributed system). Tools and support are provided to facilitate the use of these data. All members of the collaboration have access to all centrally produced data, and are free to pursue any kind of physics analysis that is supported by these data. Researchers use everything from local computing facilities to laptops to interact with these organized resources typically via custom approaches developed according to research group interests, abilities and prior experience.

Typically, physicists self-organize into small groups of a few to a few dozen researchers to collaborate on one or more physics publications. Here again reduced data samples are shared amongst members of these groups, or often between groups with related interests. Due to the sheer size of the data CMS, the processing of these reduced data samples often require substantial resources both to produce and to store. CMS nominally devotes 30% of its disk resources to data samples derived by data analysis groups.

Researchers target conference results and subsequently journal publications for discussion and peer review of their results. Conferences do bring a set of seasonal deadlines that impact the way resources are used in the collaboration. Most researchers have a peak of activity for the main spring (i.e., March) and summer (i.e., July) conferences. As such, production campaigns may be scheduled to make data available with the best available algorithms and calibrations in advance of these deadlines. Given the already very substantial Run 2 data set, this practice has reduced during Run 3 and analysis practices have become more balanced. This is simply driven by the scientific reach of most analyses, scaling only with the square root of the luminosity used. In addition to conference deadlines, CMS schedules are now driven by software and calibration readiness and other experimental aspects.

Results are peer reviewed via a process that is agreed upon by the entire collaboration, and in which every member of the collaboration is asked to participate. Typically, there are multiple stages of work and review. Analyses are initially discussed and reviewed by physicists who are expert in related analyses (e.g., analysis working groups), experts in various reconstruction, event selection and signal-vs-background separation methods, experts specialized in "statistics" (hypothesis testing), and subsequently more broadly by the CMS analysis community. The resulting publications that come out of this process carry the names of all members with authorship privileges. There is a well-defined process for somebody new to the collaboration to obtain and maintain authorship privileges.

CMS researchers collaborate broadly on technical and physics research topics. For example, CMS works closely with software authors, infrastructure providers and the theoretical particle-physics community. However, as CMS data are largely internal to the collaboration, the impact on network use of these collaborations is minimal. The vast majority of network use is thus networks that connect the close to 200 collaborating institutions within CMS.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
CERN TIER-0 FACILITY (GENEVA, SWITZERLAND) SUPPORTS ALL CMS RESEARCHERS	Partial primary copy (includes all raw data)	Data transfers Data streaming Access via on-site CPU	54 PB active disk for CMS	Constant (7/24 support)	No	N/A
FERMILAB TIER-1 FACILITY SUPPORT ALL US CMS RESEARCHERS. ALL CMS RESEARCHERS CAN ACCESS STORAGE VIA STREAMING OR BY GRID ANALYSIS JOBS (5% OF CPU DEVOTED TO USER REQUESTS)	Partial primary copy (raw and analysis data)	Data transfers Data streaming Access via on-site CPU	49 PB active disk for CMS + US user facility resources	Constant (7/24 support)	Yes	N/A
OTHER CMS TIER-1 FACILITIES: CCIN2P3 (LYON, FRANCE) RAL (OXFORD, UK) KIT (KARLSRUHE, GERMANY) JINR (DUBNA, RUSSIA, USED ONLY AS OPPORTUNISTIC RESOURCE) CNAF (BOLOGNA, ITALY) PIC (BARCELONA, SPAIN) ALL CMS RESEARCHERS CAN ACCESS STORAGE VIA STREAMING OR BY GRID ANALYSIS JOBS (5% OF CPU DEVOTED TO USER REQUESTS)	Partial primary copy (raw and analysis data)	Data transfers Data streaming Access via on-site CPU	5-16 PB active disk for CMS	Constant (7/24 support)	Yes	N/A
US CMS TIER-2 FACILITIES: UNIVERSITY OF NEBRASKA (LINCOLN, NE,USA) UNIVERSITY OF WISCONSIN (MADISON, WI, USA) PURDUE UNIVERSITY (WEST LAFAYETTE, IN, USA) UNIVERSITY OF FLORIDA (JACKSONVILLE, FL, USA) MIT (BATES LABORATORY) (MIDDLETON, MA, USA) CALIFORNIA INSTITUTE OF TECHNOLOGY (PASADENA, CA, USA) UNIVERSITY OF CALIFORNIA, SAN DIEGO (LA JOLLA, CA, USA) EACH SUPPORTS ABOUT 100 RESEARCHERS WITH LOCAL ACCOUNT ACCESS WITH COMPUTE AND STORAGE. ALL CMS RESEARCHERS CAN ACCESS STORAGE VIA STREAMING OR BY GRID ANALYSIS JOBS.	Partial primary copy (Analysis data)	Data transfers Data streaming Access via on-site CPU	Each site has 5 PB active disk for CMS + additional resources for US community (~5 PB in addition)	Constant (5/8 or better support)	Yes	N/A
OTHER CMS TIER-2 FACILITIES EACH SUPPORTS REGIONAL RESEARCHERS WITH LOCAL ACCOUNT ACCESS WITH COMPUTE AND STORAGE. ALL CMS RESEARCHERS CAN ACCESS STORAGE VIA STREAMING OR BY GRID ANALYSIS JOBS.	Partial primary copy (Analysis data)	Data transfers Data streaming Access via on-site CPU	Storage varies from 100 TB to 10 PB	Constant (5/8 or better support)	Yes	N/A
LPC TIER-3 FACILITY AT FERMILAB SUPPORTS ANALYSIS ACTIVITIES OF LOCAL RESEARCHERS, COLLABORATORS AND US-BASED RESEARCHERS IN GENERAL	Partial secondary copy	Data streaming Access to Fermilab Tier-1 disk			Yes	N/A

US TIER-3 FACILITIES (ABOUT 50% OF US UNIVERSITIES COLLABORATING IN CMS) SUPPORTS LOCAL RESEARCHERS AND COLLABORATORS (SIZE VARIES FROM FEW TO O(100)). SOME SITES ALSO PROVIDE ALL CMS RESEARCHERS CAN ACCESS STORAGE VIA STREAMING OR BY GRID ANALYSIS JOBS.	Partial secondary copy	Data transfers Data streaming Access via on-site CPU	Storage varies from 10 TB to 1 PB	Ad hoc, but some facilities have constant data transfers when running production	Yes	N/A
OTHER US UNIVERSITIES (ABOUT 30% OF US UNIVERSITIES COLLABORATING IN CMS) SUPPORTS LOCAL UNIVERSITY GROUP (SIZE VARIES FROM FEW TO DOZENS)	Primary copy of user-derived data	None	Variable active storage (< 1PB)	Ad hoc	Yes	N/A
CMS RESEARCHERS (AT CMS COLLABORATING INSTITUTIONS WORLDWIDE INCLUDING ~50 INSTITUTIONS AND ~700 RESEARCHES IN USA)	Primary or secondary copy of user-derived data	None	Variable active storage (<100TB)	Ad hoc	Yes	N/A

Table 6.10.1: CMS Data Mobility

The term "data streaming" is defined to mean remote file open, and direct reads to remotely opened files. The term "data transfer" is defined to mean transfer of entire files, most often in bulk.

6.10.2.3 IRI Readiness

US CMS has proactively pursued opportunities at US HPC sites and currently has allocations on numerous sites as described in the previous section. Typical workflows run on these allocations are typically long-term campaigns that rely on sustained access at scale. New HPC resources typically come with some, potentially substantial, work to set up given facility specific configurations and constraints. A common interface (or access requirements) across US HPC sites as well as models for sustained access at HPCs are potential IRI benefits. This is discussed in more detail in Section 6.12.

6.10.2.4 Instruments and Facilities

The LHC computing and storage infrastructure is organized in a tiered computing model including four tiers, 0-3. Originally, when the connectivity among the sites was extremely limited, the tiers were hierarchical in that Tier-N+1 was connected only to an associated Tier-N above and associated Tier-N+2 below. Data was not allowed to flow from Tier-N+2 to Tier-N. This rigid concept has been eliminated over the last decade as network capacity and capability progressed.

Data now flows across the full mesh, among all sites of all tiers. Today, and in the future, the global CMS collaboration together with WLCG and OSG will define services that sites perform. Today the service profile defines the site to belong to a certain tier. For the HL-LHC, even that might change. In the future, centers of a given tier today may no longer provide all the services that today would be expected from that tier. In addition, it is likely that the HL-LHC data and processing infrastructure will no longer support the full global mesh of data flows among all tiers. These changes are part of an ongoing R&D program towards an updated computing model for HL-LHC, and will be discussed in the HL-LHC document in more detail.

In the following, CMS will describe the current services profile for each of the tier, starting with the High-Level Trigger compute cluster (HLT).

6.10.2.4.1 HLT

The Run 3 HLT computing system is a roughly 30,000-core x86 plus GPU cluster. There are 200 nodes each with two Advanced Micro Devices (AMD) EPYC "Milan" 7763 processors and two Nvidia T4 GPUs, as well as 18 nodes each with 2 AMD EPYC "Bergamo" 9754 processors and three Nvidia L4 GPUs. Its primary purpose is to perform the last stage of the online filtering of events. All events that pass this filter will be stored forever in the archive as raw data. It is located at the site of the CMS detector roughly 10 miles away from the main CERN site, on the other side of the LHC ring. The HLT is part of LHC's trigger and data-acquisition system (TriDAS), the CMS trigger and data acquisition system that is responsible for all the real-time data reduction and collection for CMS. Its size, architecture and components are reviewed annually, independently of the annual offline and computing requirements process discussed further below.

CMS also continues to run the HLT computing system used during Run 2 as a permanent offline "cloud" resource for processing MC simulations and prompt reconstruction. On average this contributes around 25,000 x86 cores to the CMS offline processing capability at CERN.

6.10.2.4.2 Tier-0

The Tier-0 is at CERN, and has two primary functions. It is responsible for "prompt calibration and reconstruction" and archiving of all the raw data at CERN. When the LHC is not running, the Tier-0 compute cluster may be used for other processing by CMS.

6.10.2.4.3 Tier-1

CMS globally has seven Tier-1s, including Fermilab in the US The collection of the Tier-1s are responsible for operating a distributed archive of all centrally produced data, including a second copy of the raw. The copy of the raw across the Tier-1s is considered the "active copy," i.e., the raw copy at CERN is considered strictly a backup. In addition, the Tier-1s provide computing resources to the experiment that are primarily used for centrally organized processing. Historically, each Tier-1's computing was responsible for all the reprocessing of the data it archives. The computing and storage capacity of a given Tier-1 center was thus directly related to the amount of data it pledged to archive. More information on pledges follows below.

One change in the Tier-1 facilities since 2020 is that given the ongoing sanctions that impact network and other aspects of infrastructure in Russia, the JINR Tier-1 is now used as a purely opportunistic CPU resource for CMS. This is in agreement with WLCG and CERN policies with respect to JINR. This means that the substantial tape resource at JINR is no longer used.

CMS is also exploring the possibility of adding additional Tier-1 centers. Serbia has signed an MOU with CERN to provide a Tier-1 facility for CMS and is expected to be operational in the coming year or two.

6.10.2.4.4 Tier-2

CMS has roughly 50 Tier-2s globally. The collection of the Tier-2s is responsible for all data analysis, and some simulations. Tier-2s also provide disk space to host data for analysis, and a moderate amount of disk space for staging in data for processing.

6.10.2.4.5 CMS Computing Capacity and Planning

The Tier-0, -1, and -2 resources are subject to an annual needs planning process. CMS offline and computing management makes a needs assessment every year that projects out two years into the future. After internal review, the annual needs are reviewed externally by the LHCC twice a year. After sign-off by the LHCC, the needs determined in this way are used to guide the funding agencies in the various member countries, leading to annual computing resource pledges. In the US, the LHC operations programs, jointly funded by DOE and NSF, are responsible for the US pledges, and coordinate them annually with the funding agencies.

Throughout the year, the actual delivered computing capacity is measured from all Tier-1s and Tier-2s via the WLCG accounting system. For the US, the propagation of usage into this accounting system is the responsibility of OSG. This will be described in Section 6.12.

6.10.2.4.6 Tier-3

Tier-3 sites are typically university-hosted clusters that are smaller than Tier-2 sites and are typically without dedicated support to ensure high resource availability. Those in the US are not funded through the Operations Program. Tier-3 sites can be part of the DMS of CMS and receive official datasets, or can access data through streaming. These sites can also provide disk space to hold user data as required by the local community. However, they are not part of the annual requirements review process. As such, the total Tier-3 resources of CMS are presently modest compared to the other tiers.

6.10.2.4.7 HPC facilities

CMS also uses compute resources beyond the dedicated resources located in one of the tiered computing centers. These include HPC centers in the US During 2023, the CMS usage of shared HPC facilities funded by DOE and NSF amount to about 10% of the overall computing budget globally, and are not part of any pledges. This has increased substantially in the recent past: by around 4x in relative fraction of CMS compute since 2020 (and a substantially greater increase in absolute terms), and has substantially increased again in 2024. Recent allocations include 1.2 million node-hours at NERSC, 600k node-hours on TACC Frontera, 229k node-hours on Pittsburgh Supercomputing Center Bridges-2, 236k node-hours on San Diego Supercomputing Center (SDSC) Expanse and 260k node-hours on Purdue Anvil. CMS expects that this model will change for the HL-LHC to include larger allocations on HPC facilities. The size and locations of future facilities is not fully predictable; however, CMS would rely on excellent network connectivity to any HPC facility from its Tier-1 and Tier-2 facilities. Current DOE HPC facilities are well-connected to ESnet, while NSF-supported HPC facilities have excellent connectivity to Internet2. For example, NSF facilities typically take just one or two years from proposal acceptance to being in production. For more details on this see the HL-LHC case study.

Today, HPC centers funded by DOE and NSF as shared facilities across multiple science domains do not generally provide the full functionality required from a Tier-2 or Tier-1 center, e.g., they normally do not provide any substantial amount of storage under the control of the experiment. Data access at HPC centers relies on either temporarily staging in input data on storage at the HPC site, or streaming access if the site setup allows for access to the wider network. In general, HPC centers only run compute dominated applications that have modest IO/CPU requirements, e.g., the central production workflows involved in central processing and simulations. Output is archived back to the Tier-1 sites, either directly from the application or through other mechanisms.

This infrastructure setup defines the data flows. raw and official datasets flow from the Tier-0 to the Tier-1s for archival storage. Simulations are produced on all tiers and are transferred from everywhere to the Tier-1s for archival. Datasets that are needed for input to analysis are distributed across the Tier-1 and Tier-2 disk resources. The distribution is automatic and follows rules defined by the collaboration. The automatic distribution has the ability to increase the number of replicas of datasets to allow for more processing resources to have access to the data, as well as to decrease replication when demand for datasets is decreasing. All disk resources in CMS are accessible through the AAA federated data access system, implemented via XRootD, and can stream data to remote applications. CMS production jobs have a typical output rate of 1 Megabyte/sec per processing hyperthread; compute nodes with 128 Hyperthreads (HT) or more will require more than just a 1 Gbps LAN connection to the worker nodes. Accordingly, most worker nodes at Tier-1 and Tier-2 are connected at 1 Gbps or more to their LAN.

Table 6.10.2 shows the most-recent LHCC-reviewed needs and pledges for the CMS Tier-0, Tier-1 and Tier-2 facilities. The LHC has its own performance unit to standardize performance of different CPU processors: HS23, which has been recently updated from the previous HS06 standard. HS23 is built from LHC workflows to provide a more reliable standard of reference. For reference, a Dual AMD EPYC "Bergamo" 9754 system (2x128

cores per node) provides roughly 7500 HS23, and a US CMS Tier-2 provides upwards of 400 kHS23 in aggregate. As is probably obvious from the description, it is possible to have more or less pledged than needed, and more or less consumed than pledged. In addition, resource providers commonly provide resources to their local researchers beyond what is pledged to CMS. Both the US Tier-2s and the Fermilab LPC provide US researchers with CPU and disk resources beyond the US share of the CMS resource pledges. The documentation that is prepared for the LHCC review does include detailed quantitative discussion about all this. It is the primary reconciliation process of what is deemed needed, what was pledged, and what was actually consumed. The US contributes approximately its authorship share in Tier-1 and Tier-2 pledged resources to CMS.

AFTER FY24 PROCUREMENT	CPU		Disk	
Site	US Target CPU (kHS23)	Total CPU (kHS23)	US Target (PB)	Total Disk (PB)
AGLT2	227	246	12.3	18.5
MWT2	340	679	18.4	21.2
NET2	227	239	12.3	12.4
SWT2	227	341	12.3	17.3
TOTAL	1021	1,505	55.2	69.4

Table 6.10.2: LHCC-reviewed needs and pledges for the CMS Tier-0, Tier-1 and Tier-2 facilities

CMS also projects these needs into the future, including the first running years of HL-LHC, as shown in the figures below for CPU and disk. These projections were made for a LHCC review during 2021 and are detailed in the CMS Phase-2 Computing Model: Update Document[102]. These projections are made assuming little change in the CMS computing or analysis model. The ongoing R&D described in the HL-LHC case study will have a big impact on these projections as illustrated by the dashed line. By now these estimates are three years out of date. They will be updated during the ongoing offline and computing conceptual design report process in CMS. It is currently expected to conclude during 2025 and is intended to capture a more complete picture of the R&D towards HL-LHC operations and to update the overall computing model that the experiment envisions for HL-LHC.

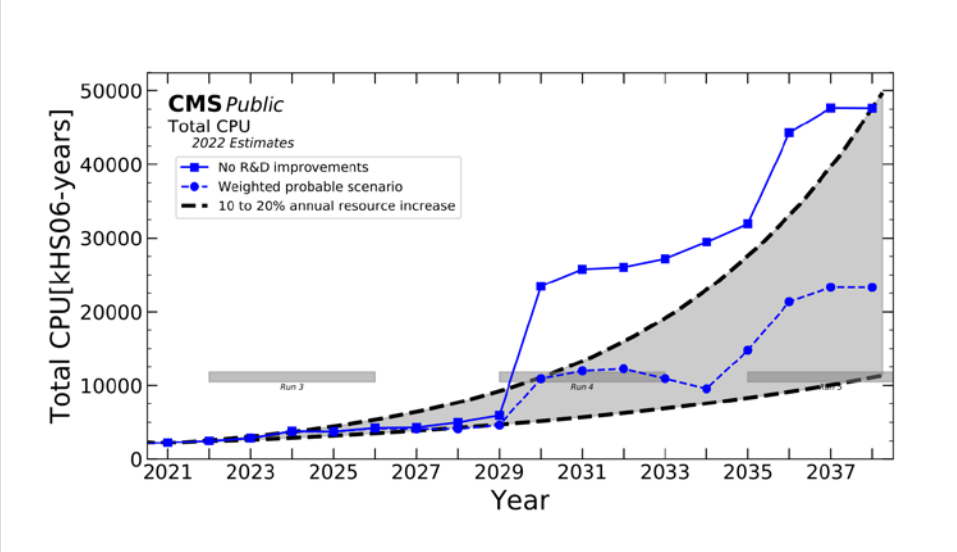


Figure 6.10.1: CMS Total CPU

102 https://cds.cern.ch/record/2815292?ln=en

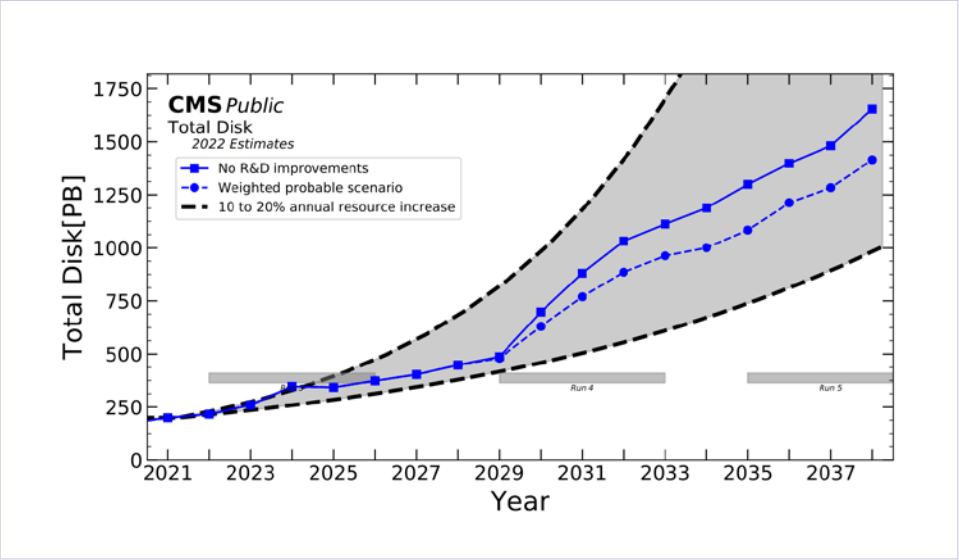


Figure 6.10.2: CMS Total Disk

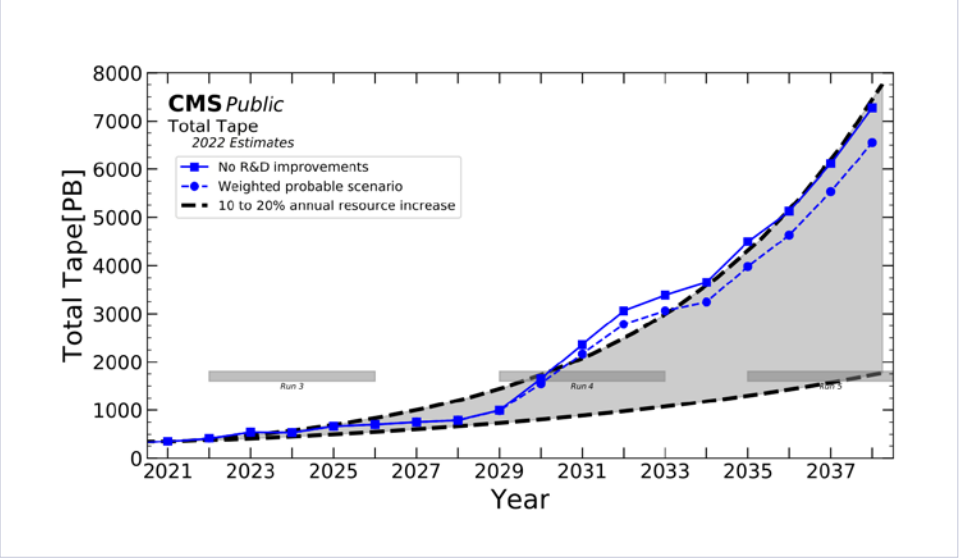


Figure 6.10.3: CMS Total Tape

6.10.2.5 Generalized Process of Science

For the purpose of this discussion, CMS distinguish three types of data: Data from collisions in the accelerator, data from simulations, and user produced data. The first two are centrally produced, while the third is produced by the physicists themselves as part of their analysis workflow. In the following the three types of data to illuminate their fundamental differences are described.

6.10.2.5.1 Collision Data

The LHC operates at a typical cadence of three running years, followed by two shutdown years. Within a running year, collisions start around March to May, and end in November or December. Software releases are prepared for the beginning of data taking each year. Changes from one year to the next within a running period are typically more modest than changes from one multiyear running period to the next.

The LHC provides data to CMS in roughly ten-hour data taking periods with minimal downtime (typically a few hours) between periods. As such, a roughly constant stream of data comes out of the TriDAS, the combined Trigger and Data Acquisition system. Event rates of around 1 kHz of physics events are typical today and for Run 3, and 7.5 kHz is expected for HL-LHC. Detector level correction ("calibrations") processes are run within 48 hours to allow for a "prompt reconstruction" to start. Beyond the baseline program, there are also targeted data taking periods with higher data taking rates. This translates into raw data volumes of:

- Run 2: 45 PB written during 4 years (2015-2018)
- Run 3: ~120 PB (2022-2024, as of September 2024)
- Run 4: 270 PB/yr (Only baseline program as estimated in 2022)

It is important to note that to capture the total CMS data volume or flow, these figures are to be multiplied by several factors. For example, data are typically processed by multiple software versions, multiple replicas of important data are kept on disk, and multiple formats for analysis data are provided to researchers.

During the annual running period, bug fixes to the software are introduced into both the online and the prompt reconstruction. In some years, these changes are significant due to the discovery of features of the CMS detector that were not expected nor foreseen (i.e., things break occasionally requiring adjustments in software because the instrument is inaccessible until shutdowns).

Towards the end of the annual running period, a final software release with the best possible calibrations, alignments and reconstruction methods is prepared, and the entire years' worth of data are processed a second time. During the shutdown between two running periods the entire data from that period is then processed one more time to arrive at a dataset that is as consistent as possible across the three-year running period.

CMS is working to evolve its data processing pipelines to avoid or at least minimize the need for either an end-of-year reprocessing or a final processing during shutdowns. This has so far been achieved with moderate success during Run 3, and needs for a full Run 3 (or Run 2 plus Run 3) processing during the long shutdown after Run 3 have not yet been determined.

Data analysis may start at any moment in this cycle, and thus may be performed on any of the three versions of the collision data. A given version of the data thus has a "lifecycle" from the time it is produced until it is retired. All centrally produced data are archived in a distributed tape archive across the Tier-0 and the Tier-1 Centers from the time it becomes available until it is retired. Only raw data and the most-recent reprocessing are archived forever, and raw data are archived at the Tier-0 and the Tier-1 sites to achieve two physically separated copies. CMS refers to the last reprocessing as the "legacy" version of the data.

The retirement date is subject to negotiations within the global collaboration with a tension between wanting to retire data to free tape resources, and wanting to retain versions necessary to support ongoing analyses until they are accepted for publication.

There are about 50 "data streams" coming out of the TriDAS. These streams are designed to take into account online trigger decisions and minimize overlaps among streams. The target is to allow at most 10% overall overlap among all streams, i.e., the sum of all collisions in these streams is no more than 10% larger than the sum of all collisions exiting the TriDAS. These data streams vary in size by about a factor of 70 from the largest to the smallest and are sometimes limited in size by offline constraints on per file sizes and luminosity accounting (which is done in approximately 23 second intervals that must be kept together as one file per data stream to ensure proper accounting).

In addition, depending on the stability of the detector throughout the annual running period, there may be 5–10 distinct "epochs" within a year. Epochs are typically similar in duration, one or two months, varying by at most a factor of five in duration. During an epoch, the raw data from a given data stream is transferred consistently to a particular Tier-1 site for archival purposes. Rebalancing of archival sites happens only at epoch breakpoints.

Multiplying the 50 trigger streams by the 3–5 epochs by the three versions, CMS observes that there are O(500) unique datasets from collisions for a given year's data collection. Each of these is managed separately across the CMS computing infrastructure.

The collaboration has developed two additional data collection modes that allow it to increase the effective trigger rate without the need for proportionately scaling up the complete computing infrastructure.

1. The "scouting" stream consists of only properties calculated during the trigger decision and does not contain the full information content of all detector hits. It is thus much smaller in size per event, allowing many more events to be saved for the same total bandwidth and saved volume of data. Therefore, the scouting stream can be written at much higher rates than the physics streams. The scouting stream can be stored on minimal disk and tape resources and analyzed directly by the collaboration. Reprocessing of scouting stream data is not possible as not enough information per event is kept. CMS is commissioning a second type of scouting, in this case based on the first level trigger information. This scouting provides data with no trigger selection aside from a minimal prescale. Thus, it gives analysts a view at the full 40 MHz data rate of LHC collisions.
2. In addition to the scouting streams, the collaboration can "park" data produced by the TriDAS directly, while skipping the following processing steps. In its design, the raw data are stored on tape and recalled at a later time, most commonly in between the running periods when the accelerator and detectors are maintained and upgraded. At that point, the full processing is done on these data and the output is provided to the collaboration for analysis just like any other data produced by the CMS detector. In practice, CMS has used the CERN and Tier-1 resources to process the parked data during the prompt reconstruction during Run 3. Parked data significantly increases the total tape archiving requirements but has less of an impact on the total computing and storage needs until it is processed. In addition, CMS keeps the raw parked data only on tape at the CERN Tier-0 in order to minimize the impact it has on the distributed computing system.

6.10.2.5.2 Simulation Data

In addition to data from the detector, CMS has produced simulations that scale in size with the luminosity collected. The planning assumption is 400 million events per fb-1, however given HPC and other compute resources available to CMS that are beyond its request, more events are typically produced. Around 100-150 billion events have been produced per year in the run 3 era. For the HL-LHC, it is currently envisioned to have roughly the same number of simulated events per unit luminosity as currently planned. This means the number of simulated events will increase considerably. The compute needed to produce a simulated event will also grow substantially. For this reason, CMS is investigating parameterized and AI/ML-based approaches to replace some or even most of the simulation production. Simulated data serves all aspects of the CMS scientific process. Simulations are needed to commission detectors, develop detector calibration and alignment procedures, design algorithms that define higher-level physics objects from lower-level instrument readings, determine their efficiency and resolution, and to develop data analysis strategies for obtaining results. A mix of simulation and carefully designed measurements based on detector data are used to determine the backgrounds to the expected or hypothesized signal for any given physics analysis. Simulations are thus of central importance to the physics program of CMS.

In total, CMS has about 70 PB (170 PB) of Run 3 (Run 2) MC simulation datasets stored on tape, representing around 75 billion (250 billion) MC events. While some intermediate clean up campaigns have taken place, this is representative of the total four-year production of simulation. Below the parameters that drive the evolution of data volumes from Run 2 to Run 3 to the HL-LHC are discussed.

As the nature of the known, yet to be detected, and hypothesized new physics processes explored at the LHC is very diverse, spanning more than ten orders of magnitude of production rate between common and rare processes, the simulations are correspondingly diverse. This is a fundamental difference between the collision

data described above and simulations. For collision data the TriDAS decides how datasets are put together, while for simulations it is the configuration of the physics generator, the software that produces the simulated energy-momentum four vectors of the particles in an event at or near their point of origin, that determines a dataset. As a result, there are 10,000 to 20,000 distinct datasets in a typical annual simulation campaign compared to about 50 distinct datasets from the detector. The size of the simulation samples varies by orders of magnitude from roughly 10,000 to 100 million simulated collisions per sample. While 70% of all simulation datasets created for the 2016 data taking era (a typical data taking year during Run2) had fewer than 100,000 collision events in them, 80% of all the simulated collision events were in datasets with more than 10 million events per dataset.

Simulation data have the same data lifecycle issues for the same reasons as detector data. In fact, CMS typically produces mostly the same set of physics processes for every software release used in detector data processing. Simulated samples that correspond to the legacy processing are kept in the archive.

CMS has changed the way it produces most simulation samples in recent years. The simulation process is a multistep one, where separate applications run different simulation aspects (e.g., physics generator, detector simulation, digitization, reconstruction). Instead of running each of these on separate worker nodes (or separate condor jobs) and staging out the results, most CMS simulation is now performed by running a chain of applications and staging out only the results to be saved for analysis. This has reduced the local and WAN usage as well as disk requirements for MC production.

6.10.2.5.3 User Data

As they are the starting point of all physics publications, the data collections or "streams" within the centrally produced CMS data are defined to ease and increase the efficiency of data access for the analyses. A specific analysis typically uses only a fraction of the events collected by the experiment; it uses only certain data streams, and only a subset of the objects that characterize the events within those streams that are available from central production. The first step of most analysis efforts is thus the extraction and formation of user defined datasets that will become the basis of further analysis. This is most often a massive down select of PBs of official data into Terabytes of user data. It also often involves a format conversion from formats that are relatively slow to process, e.g., an event processing rate of a few to 10 Hz per CPU core, to a "custom Ntuple" that is afterwards processed at multiple kHz. This supports the nature of end-user analysis which is interactive in nature and needs rapid repetition of making plots and tables.

Until recently, it has been up to the user community to produce and manage their custom "Ntuples." CMS has made a first step towards simplifying this process by introducing the NanoAOD, a data format designed for interactive end-user analysis, and produced centrally. CMS adoption of this format has already grown in Run 3 to around 50% of analyses using the NanoAOD as their primary data tier. For the HL-LHC era, additional R&D is being pursued, especially within the context of IRIS-HEP[103], to further simplify and streamline this process. For more details see the HL-LHC use case document.

Managing the disk space to host the user data is only partly within the scope of the global CMS collaboration. 20% of the disk resources pledged to global CMS are for user data, under the assumption that national entities provide more user disk space to enable the local community. The US exceeds this as detailed below in the section on Tier-2 facilities. User data are not necessarily registered in the central CMS file catalog, or replicable via the CMS DMS. Now that CMS has moved to Rucio[104] for its data management, it does have the needed infrastructure support to include user data. However, this has so far been used in only a handful of cases and is not available for widespread use.

While CERN provides some user analysis CPU resources, member countries are expected to provide the bulk of these resources for their national communities. The US CMS collaboration chose to place extra resources at the US CMS Tier-1 facility and Tier-2 centers to support user data and user analysis as described in the Tier-1

103 https://iris-hep.org
104 https://rucio.readthedocs.io/en/latest/

and Tier-2 sections below. Each US CMS member institution is nominally assigned to one of the US CMS Tier-2 centers, or the Fermilab LPC (a Tier-3 co-located with the Fermilab Tier-1), for both CPU and disk resources.

6.10.2.5.4 Event Reconstruction and Analysis Data Creation

Figure 6.10.4 summarizes the typical process for going from raw data through to analysis data in CMS (i.e., all the centrally run production processing steps that are performed).

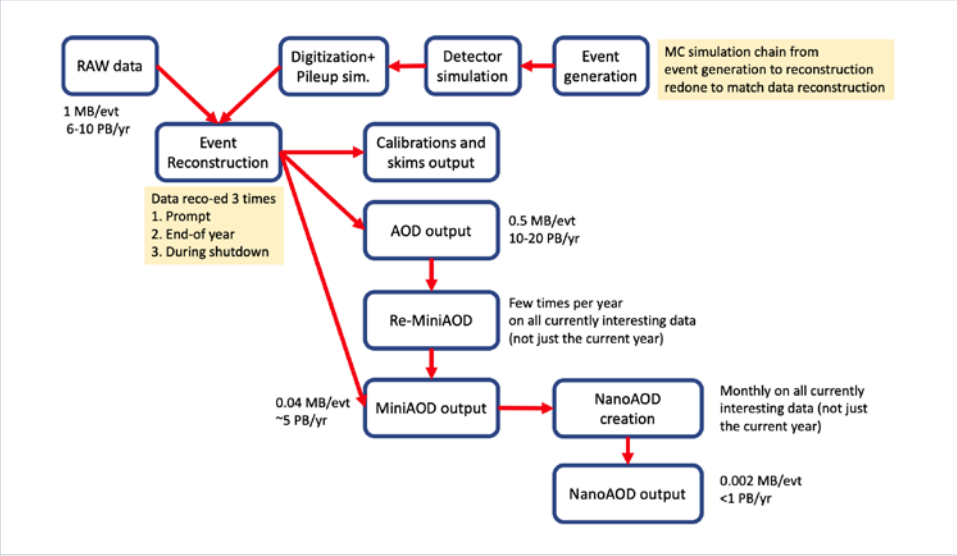


Figure 6.10.4: CMS Data Processing Steps

This multistep process differs slightly for data and simulation processing:

- Prompt reconstruction: 48 hours after a data taking fill, a period used for deriving calibrations, CMS makes a first pass through the raw data to derive physics objects from the detailed detector information. This process is called reconstruction and happens at CERN on the Tier-0 facility. The reconstruction produces analysis data which is distributed to Tier-1 and Tier-2 disk and Tier-1 tape.
- Simulation production: Simulated samples are created across the CMS Tier-1 and Tier-2 facilities as well as through HPC allocations. There are four processing steps: event generation, detector simulation, detector digitization and pileup simulation, and then event reconstruction. The way this has run in CMS has varied with time. Currently the generator step (which models the physics process and produces the particles to be simulated) is run first and saved (a few kB per event). The remaining steps are then run on a single compute node sequentially. This node reads the generator input, typically by streaming it from the site on which it was generated, and produces the analysis data formats. In addition to the event generator input, this process reads a second input file for pileup simulation. The pileup samples are pre-generated and then read via streaming as part of the digitization process. Currently these samples are 3.7 (2.7) MB/event for Run 3 (Run 2), and are expected to grow to 13 MB/event for HL-LHC. The pileup samples are each 1 PB or more and are thus located only at Fermilab and CERN. The event generation, simulation, digitization and reconstruction production chain takes about one minute on a typical CPU, so the streaming I/O need is around 60 MB/second/1000 processing threads. For HL-LHC, this rate is currently estimated to be reduced given the large increase in reconstruction time per event. The simulation production process produces analysis data which is distributed to Tier-1 and Tier-2 disk and Tier-1 tape.

- Re-reconstruction: After final software and detector conditions are derived, CMS reperforms the event reconstruction of all data collected during a year. This processing is much like the prompt reconstruction; however, it relies on the Tier-1 and Tier-2 facilities instead of the Tier-0. raw data must be recalled from tape for this process (currently done by pre-staging campaigns to recall the data files from tape and distribute them to appropriate sites). It can also be streamed as input (1 MB/event for Run 2 and Run 3 and 6 MB/event for HL-LHC). The reconstruction produces analysis data which is distributed to Tier-1 and Tier-2 disk and Tier-1 tape.
- Analysis-format reduction ("MiniAOD" and "NanoAOD" production): As the full event reconstruction takes significant resources, CMS has also developed ways to revise its smaller analysis formats (MiniAOD and NanoAOD), described below, using information saved in its larger analysis formats (as described below). In practice this processing behaves a lot like a re-reconstruction in that it requires its input data to be distributed on Tier-1 and Tier-2 disk, and analysis data formats are produced as output (however, only a subset of them are produced). Outputs are distributed to Tier-1 and Tier-2 disk pools and saved to Tier-1 tape.

Table 6.10.3 shows high-level estimates for a typical (or idealized) way of doing event processing in CMS.

Task	Events to process	Input data volume	Inputs streamed?	Inputs staged from tape?	Output data volume
PROMPT RECONSTRUCTION	19 billion/yr	23 PB (1.2 MB/evt)	No	No	(1 MB/evt)
SIMULATION PRODUCTION	45 billion/yr	175 PB (3.7 MB/evt)	Yes	No	(0.5 MB/evt)
RE-RECONSTRUCTION	50 billion/yr	60 PB (1.2 MB/evt)	Sometimes	Yes	(0.5 MB/evt)
RE-MINIAOD	190 billion/yr	80 PB (0.4 MB/evt)	Sometimes	Yes	(0.05 MB/evt)
RE-NANOAOD	190 billion/yr	8 PB (0.05 MB/evt)	Sometimes	No	(0.002 MB/evt)
CENTRALLY RUN ANALYSIS (FOR EACH TYPICAL USER)	5 billion/yr	0.25 PB (0.05 MB/evt)	Sometimes	No	Varies

Table 6.10.3: CMS Event Processing

6.10.2.5.5 Analysis Data Formats

The centrally produced data from CMS comes in multiple formats ranging from the most versatile and complete (raw and AOD) to the easiest, smallest, and fastest to use ones (MiniAOD and NanoAOD).

Data resulting from central processing is represented in memory (aside from the NanoAOD format) by objects, which are made persistent in files for analysis use. Data is represented in a two-dimensional structure, where one dimension is the collision number (recorded or simulated) and the other dimension are the objects that represent the information content for each collision. The latter can include the raw data, the simulation output, the lower-level reconstructed objects such as tracks and jets, and the higher-level reconstructed objects such as electrons, muons, quark-jets, etc. Persistence of data in CMS is based on the ROOT persistency implementation. CMS makes extensive use of ROOT's compression capabilities on storage, and thus data need to be uncompressed during retrieval.

Data formats differ in the level of detail stored per collision. The table below shows that accordingly, the average size of the data stored per collision differs by multiple orders of magnitude between raw and NanoAOD. The HL-LHC numbers are presently largely event size targets, as the actual formats are not yet finalized. The Table also shows some values for the typical processing time per event.

	Run 2	Run 3	HL-LHC
LHC ENERGY (TEV)	13	13.6	14
AVERAGE PU	35	55	200
INTEGRATED LUMI (FB-1/YR)	~50	100	350[105]
LIVETIME/YR (106 SEC)	6.5	6.5	6.5
PROMPT HLT RATE (KHZ)	1	2.6	7.5
PARKED HLT RATE (KHZ)	0.4	4.9	0
COLLECTED EVENTS/YR (X109)	9	53	53
MC EVENTS/YEAR (X109)	22	49	184
GENERATION/SIMULATION/ DIGITIZATION HS06-SECONDS/EVENT	600	625	1900
RECONSTRUCTION HS23-SEC/EVENT	250	370	5100
MINIAOD CREATION HS23-SEC/EVENT	20	30	80
NANOAOD CREATION HS23-SEC/EVENT	2	2	5
EVENT SIZE FOR RAW	0.9 MB	1.2 MB	5.9MB
EVENT SIZE FOR AOD	350 kB	500 kB	2 MB
EVENT SIZE FOR MINIAOD	35 kB	80 kB	200 kB
EVENT SIZE FOR NANOAOD	1 kB	2 kB	4 kB

Table 6.10.4: LHC parameters for a standard production year in different Runs.

The use of the various data tiers has evolved considerably over the lifetime of CMS, and is expected to continue doing so. It has become possible to use more compact event forms for an increasing fraction of the analyses, as the experiment's software base and the objects used in analysis matured. Run 1 analyses were primarily based on AOD samples. MiniAOD (10x smaller) was introduced for Run 2, and is by now well established as the primary analysis tier, used in about 90% of all analysis activity. MiniAOD includes information on physics objects sufficient to perform various analysis-specific optimizations, and even development of physics object selections, but is not sufficiently detailed to redo CPU intensive aspects of the event reconstruction, for example. Physics analyses for long-lived objects that require specialized tracking algorithms thus need to fall back to AOD. The standardized physics objects that most analyses use are derived from information in MiniAOD rather than being stored directly in the MiniAOD format, and may require some fair amount of processing. Typical processing rates for MiniIAOD today range from a few Hz to a few tens of Hz. The MiniAOD format is optimized to support remote reading, as well as some partial file access. An analysis executable typically accesses O(10)% of the data in a MiniAOD file.

NanoAOD was introduced late in Run 2 and has become the most popular data tier for analysis over the last several years. As shown in the table above is more than an order of magnitude smaller than MiniAOD. Standard physics objects are pre-computed and stored in a flat columnar format. Processing speeds are thus orders of magnitudes faster than MiniAOD. This is achieved via a combination of pre-computed standard objects, smaller event sizes, and more aggressive support of partial file reads.

More than 50% of all physics analysis activities are now based on NanoAOD, with another 40% or more to be based on MiniAOD. The use of AOD for analysis is relatively minimal given its large footprint, however it continues to occupy a large part of the CMS disk. At HL-LHC scales, CMS expects not be able to afford to keep AOD on disk any more, given its size, except as a transit version that is staged from tape for a dedicated processing application and then removed from disk. In that scenario, access to AOD would require retrieval from archival storage.

105 For Run 4, this number is scaled down to 270 fb^{-1}/year, according to models with a slower start of HL-LHC operations.

6.10.2.5.6 Data Access

CMS applications and the infrastructure that distributes and launches them support multiple types of data access:

- Traditional top-down data placement at Tier-1 and Tier-2 centers combined with applications specifying the dataset they need, and being automatically routed and executed at the sites that have it. In this mode, all data access is local to the site via the site's LAN. The figure below shows the global transfer rate for data set placement by CMS in 2019. Levels of 5 to 6 Gbpsecond were typical during this non-data taking year.
- CMS supports streaming data access to any data on disk across its grid facilities from any location with an internet connection at any time. This is called AAA, for "Any Data, Anytime, Anywhere" within CMS. CMS considers this a reasonable access modality when the application requires very little IO per CPU. It is used as part of the centrally organized production at some HPC centers simply because simulation, digitization, and reconstruction are all very CPU intensive, leading to small IO/CPU ratios. It is also used as a failover when the storage at a site is down while jobs are still running.
- Bottom-up data placement, as is implicit in caching. Here the applications are routed to sites with caches, applications access the cache locally, and cache misses are handled by the CMS XRootD Data Federation (also referred to as the AAA federation). CMS allows access in this fashion for all the MiniAOD and NanoAOD formats, but neither AOD nor raw. raw is only on tape, and AOD is only accessible via top-down placement. Caching is expected to become the dominant data access for end users of MiniAOD and NanoAOD formats. Use of caches has also proven to be of benefit to overall CMS application performance when compared to the otherwise mixed system of on-prem data and remote access patterns for CMS applications.

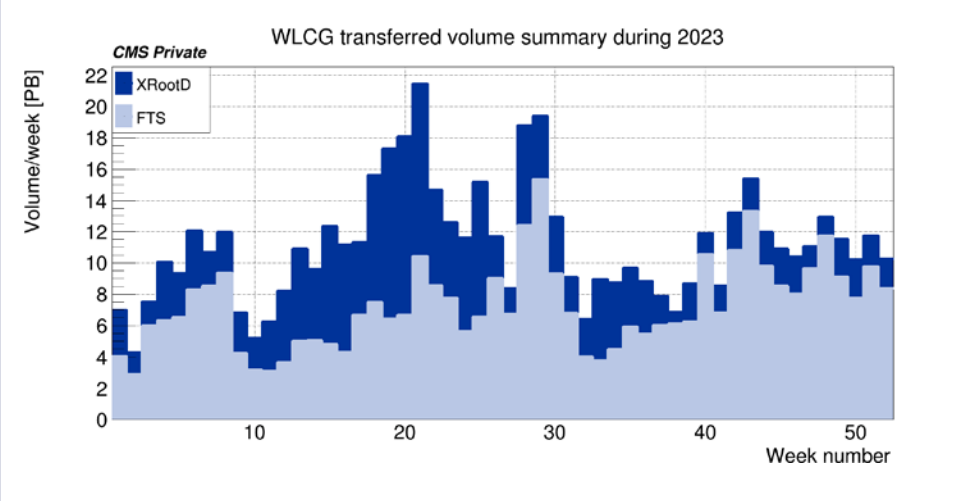


Figure 6.10.5: 2023 statistics for data set transfers in CMS for all sites globally

As a consequence, central production workflows use top-down data access and streaming access for workflows that have a very small data to CPU ratio. Caching access is not (yet) used in central production except in a few sites (specifically Caltech and UCSD in the US).

Managing data access and data distribution for a distributed storage system is different from a system with one or a handful of sites. One challenge faced by CMS is the need to keep the available disk of each of the O(50) Tier-2 facilities full of useful data that analysts want. CMS has adopted Rucio to help manage these data. Rucio allows both production teams and users to define rules for retaining data on disk, and on specific sites, for specified time periods. Data not retained by any specific rule can be removed from disk as needed to make room for new data.

Unlike 2020, CMS does not have a dynamic DMS that attempts to replicate data as needed for running jobs. With Rucio such a system has proven not to be needed. As a result, 2023 data transfers are globally less than those in 2020 despite a considerable increase in Run 3 data rates.

User data may be accessed through any of these modalities, although as user data are most often Ntuples that are by design very fast to analyze, it is unlikely to be streamed. In addition, some users reduce the size of their ntuples to something small enough such that it can be taken outside the CMS computing infrastructure, onto solid-state drives (SSDs) in laptops, or single standalone servers in people's offices or departmental server rooms.

6.10.2.5.7 Analysis Group and Researcher Workflows

So far this section has focused primarily on the extensive central production system of CMS. In addition to this, an average of about 50,000 cores across Tier-1 and Tier-2 facilities (e.g., 25% of the CMS CPU resource) are used by individual researchers or research groups for data analysis and other compute-intensive work. CMS does not place significant restrictions on the methods and tools used for analysis; and as such, groups have different ways of doing their work. One workflow that will be common for Run 3 analysis work is shown in this Figure 6.10.6:

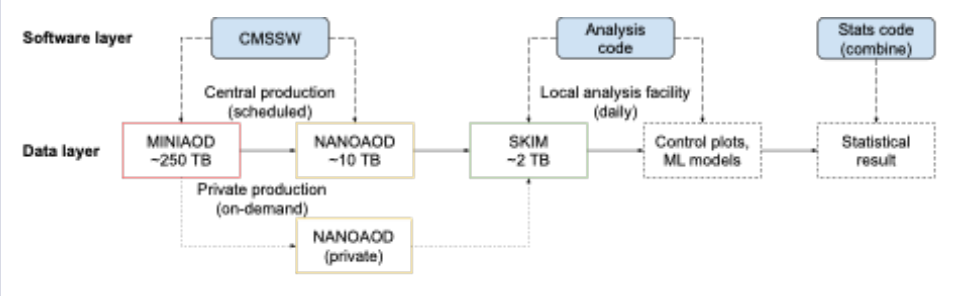


Figure 6.10.6: CMS Analysis Workflow Example

Here, data in either MiniAOD or NanoAOD format from the production system are used as input to the analysis workflow. Analysts then reduce these data by removing (aka skimming) events that are deemed uninteresting for the specific physics processes being studied, and by saving only the information in the original data format that is needed for further processing steps. These further processing steps are then performed outside of the distributed computing system, in many cases by transferring data sets to local clusters or other nearby computing systems where interactive access is possible.

Local compute access and smaller data formats provide a more interactive environment for researchers to evaluate different approaches, create proper graphical representations of their data, perform systematic studies, and to derive results. This is typically a many step process, with different codes being used through the process. A challenge for CMS through the HL-LHC is to provide researchers with tools that both enable this interactive research at a much larger volume of data, but also to reduce the need for intermediate processing steps that currently consume significant storage and compute resources. R&D on this style of "Analysis Facility" is currently quite advanced. Several US-based facility prototypes are available for users, including collocated with the Nebraska and Purdue Tier-2 and at Fermilab (EAF). One essential aspect of AFs will be excellent data I/O achieved by caching, fast disks and other approaches.

6.10.2.6 Remote Science Activities

CMS relies on a fully distributed computing system. In that sense, nearly all scientific activities are remote in nature. The use and evolution of these resources are described in the previous sections. As all science in CMS is remote, the use of HPC is by no means special in its remoteness.

Having reliable mechanisms to transfer data into HPC facilities from the CMS Tier-1 and Tier-2 centers is of growing importance. HPCs are different from other resources used by CMS primarily because they provide significant compute (CPU and GPU) resources but very limited disk resources, and because network access to

or from HPC compute nodes is typically limited or not allowed. CMS[106] and the broader LHC community[107] have considered how to best utilize these resources within operational constraints. Unlike most HPC users, CMS cannot concentrate its efforts on using one or a handful of HPC facilities, but is instead asked to use HPC facilities worldwide. Network access and data management are important aspects of this.

Today, the common solution found by CMS to use US HPCs is to stream any input data via XRootD/AAA. For MC production, the biggest piece of this is the premixing library data which is 3.7 MB for about 60 CPU seconds of processing per hyperthread, or about 0.5 MB/s for a typical eight-core MC job. This implies about 4 Gbps per 10000 cores on average at full application efficiency.

6.10.2.7 Software Infrastructure

The CMS data management software tools are more completely described in case study 12 on infrastructure and tools.

To first approximation, there are four sources of data transfers of CMS data:

- Bulk transfers due to top-down placement.
- Output handling from analysis or central processing workflows.
- XCache fetching data to handle cache misses.
- Streaming data via XRootD protocol from remote servers to applications.

The first two use cases use CMS-developed tools that sit on top of FTS[108], which is standard across the CERN community. The last two use cases use the XRootD protocol.[109]

FTS to manage scheduling and file transfer. For bulk transfers, CMS currently uses Rucio[110], having fully switched from its previous system based on PhEDEx and Dynamo tools, to manage dataset storage and dataset transfers (while still relying on FTS underneath). Rucio has become a community solution for data management.

For bulk transfers via Rucio and executed by FTS, US CMS has moved to TPC https implemented via XRootD servers and away from gridFTP. Sites typically have multiple such servers that each provide 10 Gbps, and all have access to the same filesystem. Large bandwidth transfers are thus accomplished by orchestrating very many flows across many servers.

Data streaming uses the XRootD protocol to optimally support partial file reads. XCache handles immediate data transfers, and also uses the XRootD protocol, to refresh the target cache following a cache miss. This is handled so as to minimize latency and thus minimize idle CPU when cache misses occur. The XCache server invokes an XRootD client that initially fetches only the vector of bytes requested by the application, and then later fills in the rest of the file when the server is not too busy. While of course limited by cache size, the CMS experience is that it is better to use even a cold XCache than to do remote reads in many cases (in part due to reliability of reading).

New concepts of caching are currently being established in CMS. A cache provides access to all or a subset of the official CMS data or user data without the need for organized data movement. The AAA federation is providing data access for the caches. Caches at the level of hundreds of TB up to a few PB are currently in place at Caltech and UCSD. These have remained approximately the same size during Run 3 as was anticipated given the planned Run 3 CMS data volumes.

106 https://cds.cern.ch/record/2707936/files/NOTE2020_002.pdf
107 https://zenodo.org/record/3647548
108 https://fts.web.cern.ch
109 https://xrootd.slac.stanford.edu
110 https://rucio.readthedocs.io/en/latest

To account for the complete set of all transfers it is thus sufficient to instrument and monitor XRootD and FTS within the overall transfer software stack. CMS has recently completed an analysis of the accuracy of the monitoring tools for both of these, and have found issues with both. CMS has started a program of work to improve the monitoring capabilities of both in order to arrive at reliable accounting of data transfers. It is expected this will be completed, including deployment on at least the US Tier-1 and Tier-2 sites, towards the end of 2021.

Fully instrumented XRootD software for both the HTTPS and XRootD protocols provides some redundancy in accounting as FTS uses HTTPS as implemented via XRootD servers at all US CMS Tier-2s. It is expected to be useful to achieve long-term reliable accounting of transfers together with the perfSONAR updates done at all sites leading up to the recent DC24 challenge. Moreover, FTS accounting allows transfer accounting per end-to-end link, while XRootD servers at both source and destination will have accounting for their outgoing and incoming traffic. CMS believes that continuing joint work with ESnet staff to regularly exchange network usage accounting information will continue to be important now that these monitoring improvements have been put in place.

CMS also relies on high availability for software distribution and distribution of detector conditions and calibration data to production and analysis applications. To avoid redundant data transfers, CMS uses a squid-caching system, built around Frontier[111] for distributed database caching, and CVMFS[112] for efficient global distribution of software.

6.10.2.8 Network and Data Architecture

6.10.2.8.1 The US CMS Tier-1

The US CMS Tier-1 Center at Fermilab is the largest of CMS Tier-1 facilities, currently providing 40% of the total CMS Tier-1 capacity. The Fermilab Tier-1 consists of dedicated computing infrastructure, which includes:

- 370 kHS23 growing to 520 kHS23 in 2025 (approximately 40k CPU cores) of compute nodes
- 48.8 PB growing to 63 PB of dCache distributed disk storage
- 152 PB growing to 178 PB of archival tape storage capacity that will soon be run by CTA (in addition there is a separate tape archive for the CMS Heavy Ion data).

The CMS facility also supports the LHC Physics Center analysis facility, supporting several hundred physicists with interactive computing nodes, an additional 5000 cores of batch compute, and about 9 PB of distributed EOS storage.

The US CMS Tier-1's local network infrastructure consists of fully redundant, high-performance switching fabric distributed across multiple Fermilab data centers. That switching fabric has the following characteristics:

- Currently based on 400GE network technology for inter-switch connectivity, with 10GE and 100GE connectivity available for host system connectivity.
- Extensive use of link aggregation (LAG) to scale bandwidth for inter-switch connectivity. At the present time, the US CMS Tier-1 LAN provides approximately 10Tb/s of network capacity in total.
- PBR techniques to route CMS high volume traffic over special purpose networks (such as LHCOPN, LHCONE) utilized for WAN data movement.

Fermilab's WAN architecture is based on separating its high-impact science data traffic from its general internet traffic. Conceptually, this design is analogous to a Science DMZ architecture. Most traffic into and out of the US CMS T1 is via the science data path(s). For CMS, those science data paths mean the LHCOPN, LHCONE,

111 http://frontier.cern.ch
112 https://cernvm.cern.ch/portal/filesystem

and end-to-end OSCARS circuits. Fermilab's LHCOPN connectivity supports movement of raw data from the T0 (CERN), as well as production data movement with some of the other CMS T1s. Fermilab's LHCOPN configuration consists of three OSCARS circuits (primary, secondary, and tertiary) to CERN, which provide levels of redundancy with differing bandwidth guarantees for that traffic. LHCONE supports production data movement between the Fermilab Tier-1 and most CMS Tier-2s, as well as CMS T1s that do not use LHCOPN for T1-T1 data movement. Like the LHCOPN, connectivity to LHCONE is via geographically redundant (primary/failover) paths. US CMS Tier-1 WAN traffic that does not utilize either the LHCOPN or LHCONE paths traverses the Laboratory's general internet path.

In terms of aggregate WAN capacity out of the site, Fermilab currently has provisioned 2.4Tb/s of link capacity to ESnet, via geographically diverse connections. A LAG of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports the Laboratory's general internet traffic, which includes some portion of science traffic, for example, CMS traffic to/from sites not connected to LHCONE.

Future network enhancements:

- Upgrade of the US CMS Tier-1 LAN infrastructure to 800GE technology is expected for 2028–2030 time frame;
- Upgrade of the Laboratory network perimeter infrastructure to 800GE technology is expected also in 2028 - 2030 time frame, likely to be aligned with availability of 800GE services from ESnet;
- In 2026–2028 time frame, additional WAN capacity to ESnet will be implemented in the form of additional 400GE links. The 2 x 400GE general internet connection is expected to be upgraded to 4 x 400GE prior to the start of Run 4.

6.10.2.8.2 The US CMS Tier-2s

Current experience at US Tier-2 facilities suggests the following level of network use for various infrastructure components:

- FTS transfers to and from the site: 3–5 Gbps spread across a number of servers (e.g., 10 at a typical site).
- I/O from compute nodes: 1.5 Mbps per thread, corresponding to an average of 40 Gbit/s at a current Tier-2 site.
- XCache servers: 2 Gbps (in or out) per server. As an example, Caltech maintains 2 PB over 2 cache servers.
- Squids (for CVMFS content): 100 Mbps/squid. Each site runs at least two squids.
- HTCondor gatekeepers: 2M bps/s per gatekeeper. Sites typically have three gatekeepers.
- Other traffic: up to 1 Gbit/s for user interactive access, monitoring, sync with user's desktops and laptops, etc.

CMS has a LAN requirement of 1 Megabyte/HT, i.e., nodes with 128 Hyperthreads (HT) and more will require more than just a single 1 Gbps LAN connection to the worker nodes. Accordingly, most worker nodes at Tier-1 and Tier-2 are connected at 1 Gbps or more to their LAN. Different centers have networking hardware from different vendors, and LAN aggregations, ranging from cluster-wide large Cisco gear to ToR switches that are aggregated into cluster switches via multiple 10 Gbps links. Generally, much of the LAN and WAN networking infrastructure to and within a Tier-2 site has been an institutional contribution to the US CMS Operations Program, i.e., it has not for the most part been directly funded by the program.

The Tier-2 LAN is then typically connected up to the University border switch routers via either one or more 100 Gbps links or with 400 Gbps links. Depending on location, this campus network connectivity may be shared with others. Most sites then share their outgoing connection with other Science DMZ customers on campus. Most sites have a regional network that must be traversed before connecting to ESnet. In all cases, that regional network is shared with others. Thus, actual usable bandwidth to get to ESnet is generally less, and will vary over time depending on other activities.

Figure 6.10.7 shows the ESnet network map as available on the ESnet website with the US CMS HEP Tier-1 and Tier-2 sites overlaid. Not shown is the Vanderbilt Tier-2 as it is dedicated to Heavy Ion physics, and participated in the nuclear physics ESnet requirements process in 2024.

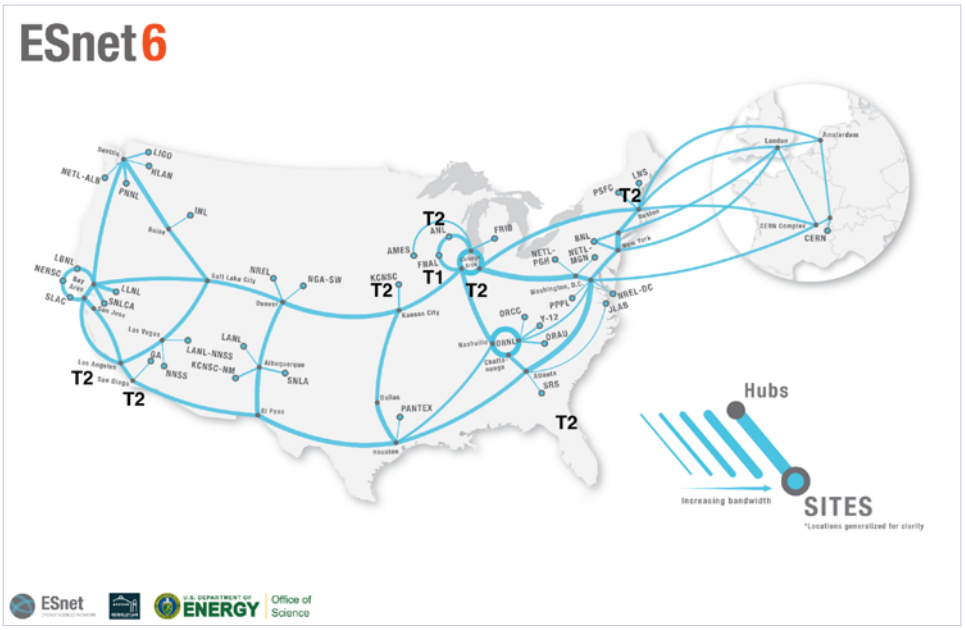


Figure 6.10.7: ESnet network map with US CMS Tier-1 ("T1") and Tier-2 ("T2") centers overlaid.

All Tier-2 sites are roughly the same in terms of CPU and storage capacity, and should thus be thought of as having roughly the same networking requirements. Based on this map, CMS can identify specific links that could be used to estimate the growth in networking as a function of time:

- **SALT - ECHO**: Covers the ESnet bandwidth from UCSD and Caltech to all other Tier-2s and Fermilab, assuming that traffic from Southern California to the East Coast and Midwest is routed via this link.
- **KANS - STLO**: Adds traffic from University of Nebraska-Lincoln (UNL) to Caltech and UCSD on its way east.
- **BOST - ALBA**: Traffic to and from MIT west should travel this route.
- **JACK - ATLA**: Traffic from and to University of Florida Tier-2 should travel this route.

Wisconsin and Purdue connect to ESnet at Starlight at 100 Gbps without traversing ESnet. There are thus no obvious ESnet links to characterize traffic from these two sites. The networking for the US Tier-2 facilities is summarized in Table 6.10.5:

US CMS Site	WAN Bandwidth	Regional Provider	LHCONE Peering Pt
CALTECH (T2)	100 Gbps	CENIC	Sunnyvale
FLORIDA (T2)	200 Gbps	FLR	Atlanta (SoX)
MIT (T2)	100 Gbps	NoX	NY (MAN-LAN)
NEBRASKA (T2)	200 Gbps	GPN	Kansas City
PURDUE (T2)	100 Gbps	I-Light	Chicago
UCSD (T2)	100 Gbps	CENIC	Sunnyvale
WISCONSIN (T2)	100 Gbps	BRAA	Chicago
VANDERBILT (HI T1)	100 Gbps	---	Atlanta (SoX)

Table 6.10.5: Networking for the US Tier-2 facilities

US CMS delegates network performance measurement collection to OSG, and expects to continue to do so. At present, Caltech, UCSD, and UNL have 100 Gbps perfSONAR hosts in a dashboard. Other locations within this mesh include various Internet2 backbone nodes in Chicago, Manhattan, Kansas City, etc. A corresponding mesh is starting up as part of the OSG-LHC networking activities. Long term, it is expected to have all Tier-2s and the Tier-1 to keep up their perfSONAR instrumentation with the bandwidth requirements for the sites.

All sites participate in perfSONAR measurements which are archived by OSG. A perfSONAR upgrade campaign was recently carried out as part of the recent DC24 data challenge. Sites have 100 Gbps, and most sites updated their perfSONAR and DTNs.

6.10.2.9 Emerging Technology Use

6.10.2.9.1 Cloud Services

CMS does not currently use cloud resources to any significant extent. Previous studies have shown them to be a more costly model than the owned-resource model that CMS currently relies on. A notable exception would be the case of needed resource bursts for either CPU or network. The resources available for either CPU or TA networking in the cloud far exceed those available to CMS. At this point, CMS tools are generally able to use cloud services, typically via infrastructure at one of the tiered sites, but CMS does not have plans to use cloud services extensively in the near or longer term. This can evolve depending on largely external factors (e.g., cost evolution, or government policies).

One specific use case is the SONIC approach which enables the use of a remote (on-prem or off-prem) accelerator resource by CMS applications. SONIC is an interesting way to adopt GPUs as it allows workers to share a set of GPUs which can then load-balance ML models (and other GPU kernels). Data access is a clear issue for models with large inputs (e.g., tracking applications).

6.10.2.9.2 AI and ML

As elsewhere in scientific pursuits, CMS has an increasing use of AI/ML in its production workflows. Analysis use cases have evolved even more quickly as there is greater freedom to adopt new toolkits and approaches. Both the Fermilab Tier-1 and Tier-2 facilities have GPU capabilities for CMS use, and the CMS community also relies on national resources such as NRP. CMS is also researching a dedicated AI/ML training facility (at Vanderbilt) in recognition that AI/ML brings two distinct use cases, training and inference, that both require substantial resources (but in differing access patterns).

CMS expects to predominantly use on-site and HPC resources rather than commercial cloud. As described just above, SONIC is one important access pattern that CMS expects for AI/ML. SONIC facilitates both on-prem and remote GPU usage, potentially with considerable LAN/WAN requirements. As of today, these are difficult to estimate as CMS is still in the process of R&D towards production use cases for SONIC.

6.10.2.9.3 Wireless or Distributed Location Science Activities

CMS does not have any work or plans in this area.

6.10.2.10 Data-Related Resource Constraints

Several examples of resource constraints or perceived resource constraints for CMS researchers and their science are present:

- CMS data volume limitations: The data volume that can be handled by the networks within and coming out of the CMS detector facility far exceeds what can be handled offline within current CPU, storage and networking infrastructures.
- TA link: CMS is currently a major user of the TA network link. Activities including the archival of raw data, initial copies of analysis data samples, and user-derived data sets using CERN or other non-US facilities are clear use cases for TA network usage. The raw data transferred to Fermilab alone are expected to average more than 10 Gbpsecond during HL-LHC operations. Today the raw data are a small part of the TA network usage by CMS. Indeed, while data rates have increased considerably during Run 3, the TA link traffic from FTS transfers inbound to the US CMS sites has not changed substantially from the end of Run 2 data taking (contrary to the concerns from the 2020 ESnet case study). CMS tools do not prioritize site proximity (in the networking sense) when scheduling data transfers. Streaming data across the TA link is allowed (even if discouraged).
- Streaming reliability: Failures in data streaming are a large source of job failures in CMS. This has also become an issue for analysis-facing activities, where a greater variety of tools are used to access CMS data. While CMS can instrument its centrally maintained software with retry mechanisms and other fallback logic, it is more difficult to extend this to community-developed tools.
- Streaming to sites beyond the CMS infrastructure: Reliable and high-capacity streaming of input data, either raw or pileup simulation considerably reduces the disk requirements of CMS at HPC and other nondedicated computing facilities.
- Disk vs. network trade-offs: As is the case with HPCs, reliable networking can be used to reduce disk replica requirements either by the use of tape recall or caching. By the end of Run 4, a copy of the entire CMS MiniAOD will be around 100 PB. If 10% of this is used during any given month in a caching system, one can estimate the need for 10 PB/month of transfers to keep the cache up to date with the most recently used data. Cache network needs will be typically bursty because users read full data sets rather than just a file or two from one. This suggests that aggregate network measures are a poor metric in this use case. Understanding and expanding caching use cases and needs are part of ongoing R&D. Recent focus has been on analysis facility use cases that have become increasingly important.

6.10.2.11 Data Mobility Endpoints

CMS does not currently use Globus or other endpoints for data sharing. There is ongoing R&D in the Pelican project to allow direct transfers from XRootD Rucio Storage Element (RSE) (the software CMS sites utilize) to Globus RSE. This would imply CMS would benefit without needing doing anything with Globus directly, and without having the operational overheads of multi-hop transfers.

6.10.2.12 Outstanding Issues

Traditionally, the LHC community has treated the network as a free and infinite resource. As the community prepares for the "Exabyte per year" era of the HL-LHC, it is envisioned to both make more aggressive use of the network, and account and manage usage much more carefully:

- Via the R&D activities in caching, it is expected to be in a position to trade investments into disk space at Tier-2s against network utilization in the long term. Other forces may also push in the direction of higher network requirements, such as how spinning disk technology evolves in terms of cost per terabyte and I/O capabilities relative to CMS needs.
- CMS would like to develop a program to test the effectiveness of transfers both to benchmark methods at increased capacity, and integrate new functionality into existing tools. Such tests should be done in collaboration with ESnet. For example, one lesson learned from the most-recent data challenge organized by WLCG (DC24) was that it is necessary to utilize a link at full capacity, to learn if a given link has any performance abnormalities. It is desirable to develop a process of regular testing of network capacities (e.g., relying on perfSONAR) to test connections to the maximum and identify problems before they become a bottleneck in production.
- Traditionally, CMS has treated the global network of CMS sites as a mesh with identical links when it comes to bulk transfers. The XRootD data federation was designed from the beginning to be cognizant of the TA link being limited, but treated links within the US as identical. The Data Lake model currently discussed in WLCG makes clear regional distinctions. Data Lake concepts have progressed slowly due to challenges around how increased latency affects application performance. However, it is expected that at least the existence of the Atlantic is likely to become an architectural feature of the data distribution architecture.
- CMS would like to continue exploring with ESnet how to contribute towards global R&D projects such as the Global P4 Lab testbed, use of ESnet High Touch services, and evolving technologies such as Smart NICs.
- The next several years are critical to transition managed network R&D, including results of the SENSE project, into production for CMS in time for HL-LHC. The SENSE and Caltech + UCSD teams are close to 400G throughput with production infrastructure using XRootD, FTS and XRootD.
- CMS believes that national and international collaboration bringing together researchers, data management experts and networking experts is important for making better use of network resources as usage levels of research networks increase. In HEP, forums for these collaborations include the WLCG Networking Throughput WG, or more broadly groups including the Global Network Advancement Group.

6.10.2.13 Facility Profile Contributors

CMS Representation

David Lange, Princeton University, David.Lange@cern.ch

Dirk Hufnagel, Fermilab, hufnagel@fnal.gov

ESCC Representation

Phil DeMar, Fermilab, demar@fnal.gov

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

6.11 LHC: LHC Operations

The experiments running at the LHC are exploring the fundamental structure of matter and the forces that govern their interaction. The two main general-purpose collaborations at the LHC are ATLAS and CMS and both collaborations have thousands of collaborators distributed around the globe who require access to the data generated by their detectors at CERN and to simulated data generated at sites around the world.

6.11.1 Discussion Summary

A paper was published after DC24[113] with a number of findings. The LHC collaborators' DC24 activity revealed several weaknesses in the software stack, namely with the way that FTS performed globally when faced with bottlenecks at certain locations. The discussion in CMS about the use of alternative data formats, and the use of other R&D efforts such as Caching may also influence how much bandwidth is needed between facilities, with the TA links in particular being a main focus area. ESnet's TA strategy has not changed substantially since the previous HEP review, and will continue to add capacity (mainly through the use of more flexible spectrum agreements, and sharing resources with others partners like GÉANT across multiple connections).

A final observation from the LHC collaborators' DC24 activity was that WLCG's hardware choices for the next major cycles will have a huge influence on the performance of the overall computation and storage system. Since funding requests have to be made in advance, and hardware choice and prices fluctuate, there is uncertainty about what capabilities may exist. As a result of this, there will be a series of mini-DC activities that occur on a regular basis to evaluate aspects of the ecosystem: computing, storage, software, and networking, culminating in the larger complete DC activities that focus on major goal milestones. ESnet can assist with mini-DC activities when needed, specifically by making resources like perfSONAR and ESnet High Touch available to assist with traffic generation and results gathering.

Networking support at BNL and Fermilab, to support the ATLAS and CMS Tier-1 computing facilities, respectively, has increased and is approaching an aggregate of 1Tbps through all connectivity. Both sites are exploring ways to move toward 800Gbbps native connections in the coming years, when the technology is ready to be supported.

One last positive impact of the DC24 activity, that is not fully planned for within the next DC, will be a full analysis of the results and ways to both capture and act on these data in near-real time. A data analysis will be needed, and this effort should begin the process of evaluating each link in the ecosystem and ways that both LHC monitoring data, and that from ESnet, can be captured and used to evaluate overall system performance.

ESnet will continue to work with the LHC community on co-design efforts like SENSE and Rucio, SciTags, and other potential opportunities. The experiments do need to be specific in what they are hoping to achieve, however. The aim of joint projects could be to "co-design" an effective system that makes best use of the available network resources, coordinated with the site resources, and manage joint site and network operations cooperatively. This needs to be spelled out in more detail, and discussed with ESnet leadership.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The LHC schedule has operating periods (called "runs") interspersed with maintenance and upgrade periods (called "long shutdowns"). The first run of the LHC was 2009-2012 followed by long shutdown 1 (LS1) from 2012-2014. The second run (Run 2) was from 2014-2018, followed by LS2 from 2019-2021. Run 3 has started in 2022 and will continue until 2026. The LHC will be upgraded to increase its luminosity during LS3, from 2026 to 2029. The new accelerator, HL-LHC will start producing collisions in Run 4 in 2030. ATLAS and CMS are expected to produce 10 times more data compared to Run 3. [Section 4.1]

113 https://zenodo.org/records/11444180

- The collaborators involved with LHC operations are the four main LHC experiments (ALICE, ATLAS, CMS and LHCb) as well as the technical teams running the accelerator and the globally distributed middleware (OSG, EGI, Rucio, perfSONAR, FTS, etc.) and infrastructures (the various Tier-x sites) supporting the activities of the experiments. [Section 4.1]
- LHC Run 3 will be completed by the end of 2026 and it is expected that the network traffic will keep growing at the usual rate until then. In the latest years, the LHC community has seen an annual increase of 20-30% of traffic in LHCOPN and LHCONE, the two networks dedicated to transfers between WLCG sites. The current networking capacities are typically 100 Gbps WAN links, with many sites already having resilience (multiple) 100 Gbps connections. This will need to evolve as the LHC community prepares for Run 4. The expected traffic during Run 4 beyond 2030, when the HL-LHC will come into operation, will significantly increase. ATLAS and CMS have estimated this increase in network traffic to be a factor of 10 compared to Run3. [Section 4.2]
- In preparation of HL-LHC the LHC experiments have defined a calendar of "data challenges" to progressively test the capabilities of the WLCG software and hardware infrastructure to meet the requirements. Here is the way they have estimated the network bandwidth: [Section 4.2]
    - ATLAS & CMS T0 to T1 per experiment
        - 350PB raw data per year; average of 50 Gbps or 400 Gbps during LHC running time
        - Another 100 Gbps estimated for prompt reconstruction data tiers (AOD, other derived output)
        - estimated 1 Tbps for CMS and ATLAS summed
    - ALICE & LHCb T0 Export
        - 100 Gbps per experiment estimated from Run 3 rates
- For LHC operations, the use of either pledged or opportunistic HPC resources located away from traditional tiered sites is currently very difficult because each HPC facility is "special" in terms of bespoke hardware, network access, storage policies and job submission. This means each experiment using such facilities needs to dedicate significant resources to operating and maintaining the HPC for their use case. HEP experiments have developed in-house mechanisms, like HEPCloud and Harvester, to manage access to disparate resources. Therefore, the LHC community would greatly benefit from a unified security model across DOE (and NSF) HPC facilities and a unified resource brokering API via efforts such as IRI. [Section 4.3]
- The LHC community maintains some common efforts to provide software that is used throughout the ecosystem: [Section 4.3]
    - Rucio, a distributed DMS.
    - FTS, responsible for reliably transferring data while not overloading end-systems.
    - Monitoring.
        - SciTags, to label flows and/or IPv6 packets by experiment and activity
        - WLCG Site Network Monitoring to track each site's IN/OUT traffic
    - Networking orchestration
        - SENSE/Rucio to integrate SDN with distributed data management
        - NOTED a tool to identify network congestion and move large flows to other paths dynamically

- All Tier-1 and Tier-2 sites are expected to have perfSONAR nodes deployed to monitor network performance. These nodes are integrated into a mesh configuration, enabling regular tests such as latency and bandwidth checks between sites. perfSONAR allows network administrators to identify potential network issues that could impact data transfer rates, both intersite and with CERN. It plays a vital role in the global transfer of data across the LCG. [Section 4.3]
- Monitoring for CERN IT Data Centre and WLCG Infrastructure (MONIT) is CERN's scalable, central monitoring system that aggregates, stores, and analyzes monitoring data from multiple sources across the LHC experiments. It provides a unified interface for monitoring the infrastructure, including network performance, storage, and computing resources. MONIT gathers data from various monitoring tools (e.g., perfSONAR, FTS, XRootD, and custom storage monitoring systems) deployed across all Tier-1 and Tier-2 sites. [Section 4.3]
- The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. The mission of the WLCG project is to provide global computing resources to store, distribute and analyze the ~50-70 PB of data expected every year of operations from the LHC at CERN. WLCG is coordinated by CERN. It is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS and LHCb) and the participating computer centers. [Section 4.4]
- In terms of aggregate WAN capacity out of the site, CMS Tier-1 (Fermilab) currently has provisioned 2.4Tb/s of link capacity to ESnet, via geographically diverse connections. A LAG of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports the Laboratory's general internet traffic, which includes some portion of science traffic, for example, CMS traffic to/from sites not connected to LHCONE. [Section 4.7]
- BNL has two path-diverse 800 Gbps (2 x 400 GbE) circuits that peer with ESnet routers within their 515 and 725 data centers. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits. [Section 4.7]
- SENSE (SDN for End-to-End Networked Science at the Exascale) uses SDN to dynamically manage network paths, prioritize traffic, and provide optimized routing. This setup is ideal for data-intensive tasks, as it can allocate resources in real-time based on network conditions and experimental priorities. Rucio, developed by CERN, complements SENSE by managing large-scale data distribution across the network. Its features include automated data replication, strategic data placement, and support for geographically distributed resources, which streamline the movement and availability of critical datasets. The CMS & ATLAS experiments, in collaboration with ESnet, are integrating the SENSE/Rucio architecture to improve its network management and data transfer capabilities. [Section 4.8]
- ATLAS and CMS are exploring advanced techniques for improving data transfer and monitoring capabilities across the WLCG. New technologies such as SciTags and packet marking are being tested and deployed to enhance network monitoring, traffic differentiation, and optimized data flow across high-speed networks. The deployment of SciTags and packet marking is in progress across several sites, which will lead to better traffic differentiation and resource management across the network. [Section 4.8]
- Many of the LHC experiments have explored cloud services and are using them as part of their global set of resources. There are two primary modes of operation envisioned: [Section 4.8]
    - A particular site contracts with a cloud provider and resources are (transparently) added as part of the site from their experiment's perspective.

- A new site is created within a cloud provider, utilizing the various provisioning, management and resilient infrastructure available. However, this option not feasible for the following reasons:
  - The entity setting up a site within a cloud provider needs to be able to control costs to make it feasible, but there are not usually policies and tools to allow them to do so.
  - Egress charges from cloud providers are usually expensive and may vary depending upon where the egress goes to.
  - Connecting to resources like LHCONE is very challenging because of the AUP requirement that all traffic from the advertised subnets for LHCONE must be LHC related.

- Globally, many funding agencies are responsible for both HPCs and WLCG experiments and a natural outcome is to try to support WLCG needs via HPCs. As an example, Germany has consolidated many of its Tier-2 sites into a new HPC allocation. For the US, it would be very beneficial for the DOE HPC facilities to become "standardized," allowing collaborations to utilize resources without having to dedicate approximately 0.25 FTE to operate things on a bespoke basis. Without something like IRI driving this uniform interface, the experiments are also forced to engage with each HPC as it is designed to help ensure it is usable for that experiment. [Section 4.9]
- The LHC experiments have defined two models to predict bandwidth needs into the era of HL-LHC: [Section 5.7]
  - Minimal Model
    - Sum (ATLAS,ALICE,CMS,LHCb)*2(for bursts)*2(safety-margin) = 4.8 Tbps expected HL-LHC bandwidth
  - Flexible Model
    - Experiments may need to reprocess and reconstruct the collected data during the year
    - This requires doubling the bandwidth of the Minimal model: 9.6 Tbps expected HL-LHC bandwidth
- The LHC experiments predict the following bandwidth estimate for sites: [Section 5.7]
  - Each Major Tier-1s:
    - 1 Tbps to the Tier-0 (LHCOPN)
    - 1 Tbps to the Tier-2s (aggregated, LHCONE)
  - Each Major Tier-2s:
    - >400 Gbps (LHCONE) and resilient
- One place that R&E networks could help would be in jointly negotiating with various cloud providers to create a template contract they would support no egress charges if the R&E providers bring high-bandwidth links to the cloud providers, as well as defining policies and associated tools to allow LHC site creators to manage costs by, for example, controlling how data flows to only use the zero (or low cost) egress connections globally. [Section 5.7]
- The HL-LHC requirements for LHCOPN links approach 1 Tbps and has pushed sites to explore more affordable technologies to achieve it. IT-INFN-CNAF, the Italian Tier-1, has deployed a pilot connection using DCI equipment at the Tier-1 (CNAF) and Tier-0 (CERN) sites, which use four light channels on the dark-fibers of two different providers, GÉANT and GARR. The solution can provide 4 x 400 Gbps link for a total of 1.6 Tbps. Spectrum Sharing

is achieved by interconnecting the ROADM device of the two providers. It saves on the cost of more expensive router interfaces, since most of the existing LHCOPN links are provided as MPLS Layer 2 virtual circuits. [Section 5.7]

- During the 2020-2021 studies of LHC network requirements, the LHC experiments estimated that Tier-1 centers would need terabit/sec connections and Tier-2 sites would require 400G connections. That was driven by estimates of some (but not all) data-related operations of CMS and ATLAS, leading to an estimated bandwidth need of 4.8 terabits/sec (Tbps) across the Atlantic for CMS and ATLAS alone in a so-called "minimal model" and 9.6 Tbps in a more "flexible" and realistic model. [Section 5.7]
- As of 2024, multiple 400G connections have been installed to the Fermilab and Brookhaven national labs as well as HPC facilities used by HEP, and the upgrades to 400G (or N X 400G) at some US CMS campuses are either in place or imminent. As mentioned above, 400G-scale operations in production mode are already being demonstrated, and are expected to be available in the mass market before HL-LHC startup in 2030. [Section 5.7]
- It is expected that all the major Tier-1s to have at least 800 Gbps LHCOPN links at the start of Run 4 with the ability to grow to 1.2 Tbps or even 1.6 Tbps during the Run. Tier-1s should also have an equivalent amount of bandwidth to LHCONE to serve Tier-2s. Details of the networking for ATLAS and CMS are in the experiment specific case studies (#9 and #10 respectively). [Section 5.7]
- For LHC experiments in general the resource requirements are growing beyond what flat budgets can support. This leads the experiments to work very hard at optimizing their data and analysis processes to maximize what the available resources can do. Various trade-offs are continually being explored (network use vs multiple replicas of data, data format changes vs number of required datasets to meet various needs, exploitation of nonstandard computing capacity, reconstructing data on demands vs storing it), to try to achieve the time to science target with constrained resources. [Section 5.8]
- Because of recent changes to technology drivers, the evolution of ATLAS and CMS's "process of science", and the outlook for 2030 when the HL-LHC era will begin, it appears that the actual networking requirements may be somewhat larger than foreseen during the 2020 HEP Requirements Review. Detailed projections are difficult at this point, but should be tracked and updated every two years or so to make quantitative estimates and (re-)establish the requirements in both network capacity and capability. [Section 5.8]

6.11.2 LHC Operations: Profile

The LHC schedule has operating periods (called "runs") interspersed with maintenance and upgrade periods (called "long shutdowns"). The first run of the LHC was 2009-2012 followed by long shutdown 1 (LS1) from 2012-2014. The second run (Run 2) was from 2014-2018, followed by LS2 from 2019-2021. Run 3 has started in 2022 and will continue until 2026.

6.11.2.1 Science Background

The LHC will be upgraded to increase its luminosity during LS3, from 2026 to 2029. The new accelerator, HL-LHC will start producing collisions in Run 4 in 2030. ATLAS and CMS are expected to produce 10 times more data compared to Run 3. It is important to note that each succeeding run significantly increases the amount and complexity of the data generated, requiring increased storage, network and compute capacity to successfully exploit.

6.11.2.2 Collaborators

The collaborators involved with LHC operations are the four main LHC experiments (ALICE, ATLAS, CMS and LHCb) as well as the technical teams running the accelerator and the globally distributed middleware (OSG, EGI, Rucio, perfSONAR, FTS, etc.) and infrastructures (the various Tier-x sites) supporting the activities of the experiments. Sections 6.9 and 6.10 should provide the details for each collaboration's process of science, specifics of data volumes, and internalized hardware and software infrastructure.

6.11.2.3 IRI Readiness

For LHC operations, the use of either pledged or opportunistic HPC resources is currently very difficult because each HPC is "special" in terms of hardware, network access, storage policies and job submission. This means each experiment using such facilities needs to dedicate significant resources to operating and maintaining the HPC for their use case.

However, ATLAS and CMS are the top two users of the NERSC CPU partition, on track to use over 2 million node-hours between them. Since each HPC is special, HEP experiments have developed in-house mechanisms like HEPCloud and Harvester to manage access to disparate resources. Therefore, the LHC community would greatly benefit from a unified security model across DOE (and NSF) HPC facilities and a unified resource brokering API.

Globally, many funding agencies are responsible for both HPCs and WLCG experiments and a natural outcome is to try to support WLCG needs via HPCs. As an example, Germany has consolidated many of its Tier-2 sites into a new HPC allocation. For the US, it would be very beneficial for the DOE HPC facilities to become "standardized," allowing collaborations to utilize resources without having to dedicate approximately 0.25 FTE to operate things on a bespoke basis. Without something like IRI driving this uniform interface, the experiments are also forced to engage with each HPC as it is designed to help ensure it is usable for that experiment.

6.11.2.4 Instruments and Facilities

6.11.2.4.1 LHC

The LHC collides protons more than a billion times every second, out of which each experiment selects interesting collisions for recording at a variable rate, depending upon the triggering and capacity of the experiment's data acquisition system. Typically experiments have about two thousand hours of data collection every year when the LHC is running, and the four main experiments have a huge data sample for physicists to analyze from the completed Run 1 and Run 2 periods. Run 3 started data collection in 2022 and is planned to continue to 2026. Table 6.11.1 shows typical LHC parameters.

Quantity	Number
CIRCUMFERENCE	26659 m
DIPOLE OPERATING TEMPERATURE	1.9 K (-271.3 °C)
NUMBER OF MAGNETS	9593
NUMBER OF MAIN DIPOLES	1232
NUMBER OF MAIN QUADRUPOLES	392
NUMBER OF RF CAVITIES	8 per beam
NOMINAL ENERGY, PROTONS	6.8 TeV
NOMINAL ENERGY, IONS	2.76 TeV/u (energy per nucleon)
NOMINAL ENERGY, PROTONS COLLISIONS	13.6 TeV
NUMBER OF BUNCHES PER PROTON BEAM	2340
NUMBER OF PROTONS PER BUNCH (AT START)	1.6×10^{11}
NUMBER OF TURNS PER SECOND	11245
NUMBER OF COLLISIONS PER SECOND	1 billion

Table 6.11.1: LHC parameters

All of the equipment needed for the High-Luminosity LHC, the LHC's successor, and its experiments will be installed during Long Shutdown 3, between 2026 and the end of 2028. The High-Luminosity LHC is scheduled to come into operation at the end of 2029, though this could shift to 2030 with the outcome decided in October 2024.

Stable beam time for 2024 (with a duty cycle of 50%) will continue with an average pileup of $<\mu>=55$ and luminosity leveling at $\mu=64$, leading to a projected 110 fb-1 of delivered integrated luminosity. Additionally, about 4 weeks of heavy ion collisions are expected. In 2025 a similar data taking setup is planned.

6.11.2.4.2 Globally Distributed Computing Sites

Both ATLAS and CMS have organized their resource sites following a hierarchical tiered model with their Tier-0 at CERN, national scale Tier-1 sites and regional Tier-2 sites distributed in various countries.

6.11.2.5 Generalized Process of Science

Each of the LHC experiments has its own unique requirements for its scientific processes. See Sections 6.9 and 6.10 for more details.

Generally, the experiments gather data from the LHC collisions occurring within their detectors with real-time event selection via various hardware and software triggers to capture events of interest. These data are initially processed at the Tier-0 (CERN) and then globally distributed to various computing centers. The further processing and analysis become more experiment specific but generally include categorizing, thinning and skimming data to make it more usable for physicists. Various tools are used to track data, manage workflows and produce suitable data formats and simulated results, all leading to scientific publications for the areas of high energy physics each experiment is exploring.

6.11.2.6 Remote Science Activities

LHC computing is fully distributed, and therefore remote from CERN. The raw data are collected exclusively at CERN - all other activities leading to the final published products are done remotely.

6.11.2.7 Software Infrastructure

The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. The mission of the WLCG project is to provide global computing resources to store, distribute and analyze the ~50-70 PB of data expected every year of operations from the LHC at CERN on the Franco-Swiss border.

The scale and complexity of data from the LHC is unprecedented. These data need to be stored, easily retrieved and analyzed by physicists all over the world. This requires massive storage facilities, global networking, immense computing power, and, of course, funding.

CERN does not have the computing or financial resources to crunch all the data on-site, so in 2002 it turned to grid computing to share the burden with computer centers around the world. The WLCG builds on the ideas of grid technology initially proposed in 1999 by Ian Foster and Carl Kesselman.

WLCG is coordinated by CERN. It is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS and LHCb) and the participating computer centers. It is reviewed by a board of delegates from partner country funding agencies, and scientifically reviewed by the LHC Experiments Committee. The WLCG is partnered with EGI, OSG, and NeIC.

There are some common efforts to provide software that is used by more than one LHC experiment. Some examples are summarized here and will be presented in more detail below:

- Rucio, a distributed DMS.
- FTS, responsible for reliably transferring data while not overloading end-systems.
- Monitoring
    - SciTags, to label flows and/or IPv6 packets by experiment and activity.
    - WLCG Site Network Monitoring to track each site's IN/OUT traffic.
- Networking orchestration
    - SENSE/Rucio to integrate SDN with distributed data management.
    - NOTED a tool to identify network congestion and move large flows to other paths dynamically.

6.11.2.7.1 Data Management Tools

For ATLAS & CMS, the majority of data transfers over wide-area networks are facilitated through either the FTS or XRootD. Rucio now handles all large-scale data transfers over wide-area networks. It coordinates file movements and replication across distributed sites using FTS for third-party transfers and HTTPS for secure, efficient data transfer. The adoption of HTTPS in Rucio marks a significant upgrade in security compared to the old GridFTP system.

XRootD, which is still used for direct data access, works in conjunction with Rucio to ensure smooth file transfers. Rucio uses Data Identifiers to manage and organize files, datasets, and containers, making data tracking and management more efficient. This system ensures that data can be easily moved between different storage locations without errors.

For both ATLAS and CMS, most data transfers over wide-area networks are managed through the FTS or XRootD. However, CMS exhibits a slightly different data access pattern compared to ATLAS, particularly when it comes to data streaming using XRootD, where CMS jobs read data remotely. To effectively support these streaming access scenarios, deploying a robust network infrastructure, such as a CDN, is crucial. A CDN can significantly enhance data streaming performance, improving the efficiency and reliability of remote data access for CMS batch processing.

A list of common ATLAS & CMS current and future tools designed to handle the high throughput and complexity of LHC data processing and storage includes:

1. **Rucio:** Rucio is a data management tool used by ATLAS and CMS (and other LHC experiments) to manage large volumes of distributed data. It provides automated data replication, transfer management, and dataset tracking across multiple storage sites. Rucio allows efficient and reliable data movement between Tier-1, Tier-2, and Tier-3 sites. Rucio is integrated with the experiments' Workflow Management Systems for production and analysis workflows. The Rucio framework is expected to evolve to accommodate higher data rates, improved automation, and dynamic resource management. Enhancements may include better integration and improved support for real-time data analytics.
2. **XRootD:** XRootD is a high-performance, scalable data access framework. It supports fast, efficient remote data access, making it critical for distributed LHC data processing. Both ATLAS and CMS use XRootD to provide federated access to data, depending upon the site involved, enabling users to access files from remote locations without needing local replicas.
3. **FTS:** FTS is a multiprotocol data transfer tool. It manages and schedules large-scale data transfers between WLCG sites. FTS plays a significant role in ensuring that the CMS & ATLAS can efficiently transfer data to and from CERN and other sites.
4. **dCache:** A distributed storage system designed to manage large amounts of data with efficient I/O operations. dCache provides a hierarchical storage solution that integrates with disk, tape, and SSD, optimizing data access for different tiers. It is heavily used by both ATLAS and CMS, particularly for managing cached data between hot, warm, and cold storage.
5. **CRIC (Computing Resource Information Catalog):** CRIC is a resource information system used to manage and describe the experiment's computing infrastructure. ATLAS and CMS use CRIC to manage their resource details (e.g., storage, computing nodes, network capacity), which are crucial for data processing and analysis workflows.
6. **CVMFS:** CVMFS is used to distribute software to computing nodes in a fast and scalable way. ATLAS and CMS rely on CVMFS to access software environments on both local and remote computing nodes. It ensures that up-to-date software versions are available globally, which is essential for consistent analysis and data processing across distributed resources.
7. **Frontier (Squid):** Frontier is a system for retrieving condition data from databases, using a squid-caching layer to reduce database load. Both ATLAS and CMS use Frontier/Squid to cache and serve condition data used by their jobs, reducing latency and database load during analysis or reconstruction tasks.
8. **HTTP-TPC (used by FTS and other transfer tools):** HTTP-TPC is a third-party copy mechanism for transferring data between storage systems using HTTP. ATLAS and CMS use HTTP-TPC for data transfers between different storage elements, often in conjunction with FTS, allowing more efficient data movements across their infrastructure.
9. **HTCondor:** HTCondor is a job scheduling and management system for distributed computing used by both ATLAS and CMS. It targets HTC systems.
10. **Simple Linux Utility for Resource Management (SLURM):** SLURM is a job scheduling and management system used for distributed computing by both ATLAS and CMS, typically for HPC systems.
11. **perfSONAR tool:** perfSONAR is a network measurement tool used to monitor and troubleshoot network performance. It helps identify bottlenecks and optimize the paths between the experiment sites, ensuring efficient data flows and enabling easier identification of network problems.

6.11.2.7.2 Monitoring Tools

In the CMS and ATLAS experiments, efficient network and storage monitoring across Tier-1 and Tier-2 sites is critical for ensuring reliable, high-performance distribution of large datasets. These sites are globally distributed and connected through high-speed scientific networks, including the WLCG infrastructure, to facilitate data transfer and analysis. Here is an overview of various network and storage monitoring systems in place, such as perfSONAR and CERN's MONIT (which includes monitoring for the FTS, XRootD, and storage systems).

- Tier-1 and Tier-2 Monitoring Using perfSONAR:
  - perfSONAR (Performance Service-Oriented Network Architecture) is an open-source network monitoring system widely used across the LHC Tier-1 and Tier-2 sites for measuring network performance. It helps identify and resolve network bottlenecks by continuously monitoring network latency, packet loss, jitter, and throughput.
  - All Tier-1 and Tier-2 sites are expected to have perfSONAR nodes deployed to monitor network performance. These nodes are integrated into a mesh configuration, enabling regular tests such as latency and bandwidth checks between sites. perfSONAR allows network administrators to identify potential network issues that could impact data transfer rates, both intersite and with CERN. It plays a vital role in the global transfer of data across the LCG.
- SciTags and Packet Marking Deployment in US
  - ATLAS and CMS are exploring advanced techniques for improving data transfer and monitoring capabilities across the WLCG. As the demand for large-scale data processing grows, driven by the High Luminosity LHC (HL-LHC) era, new technologies such as SciTags and packet marking are being tested and deployed to enhance network monitoring, traffic differentiation, and optimized data flow across high-speed networks.
  - The deployment of SciTags and packet marking is in progress across several US CMS sites. The UNL and UCSD have already successfully implemented SciTags as part of their network monitoring infrastructure. Other US Tier-2 sites are currently in the process of deploying these technologies, which will lead to better traffic differentiation and resource management across the network.
  - UNL's deployment has been a crucial step in demonstrating the feasibility of SciTags and packet marking for large-scale scientific data transfers. The feedback and results from this deployment are helping to refine the implementation at other US sites.
- CERN's MONIT System:
  - MONIT is CERN's scalable, central monitoring system that aggregates, stores, and analyzes monitoring data from multiple sources across the LHC experiments. It provides a unified interface for monitoring the infrastructure, including network performance, storage, and computing resources.
  - MONIT gathers data from various monitoring tools (e.g., perfSONAR, FTS, XRootD, and custom storage monitoring systems) deployed across all Tier-1 and Tier-2 sites.
  - MONIT provides dashboards for real-time data visualization and analysis. It uses tools like Grafana to present a comprehensive overview of the network and storage infrastructure.

The Tier-1 and Tier-2 sites use a combination of network and storage monitoring tools, to ensure the reliable and efficient transfer and storage of experimental data. The experiment's monitoring efforts are tightly integrated with CERN's MONIT platform, enabling centralized monitoring and rapid issue resolution. This comprehensive monitoring infrastructure is crucial for supporting the high-performance computing needs of the experiments.

This monitoring setup ensures that the experiments can operate efficiently, even with the increased data demands during the High-Luminosity LHC (HL-LHC) era. Recommendations include:

1. Continuous Improvement of perfSONAR Testing: Regularly updating perfSONAR configurations and running extensive tests will ensure network issues are detected early.
2. Storage System Expansion: As data grows, expanding storage monitoring capabilities will become increasingly important.
3. Increased Use of MONIT Dashboards: LHC sites should leverage MONIT's full potential by creating custom dashboards for different types of network and storage performance indicators.

6.11.2.7.3 Network Management Tools

The CMS & ATLAS experiments, in collaboration with ESnet, are integrating the SENSE/Rucio architecture to improve its network management and data transfer capabilities. This system is designed to address the growing data demands and transfer needs of CMS & ATLAS, ensuring faster, more reliable, and prioritized data flows across its global network of data centers.

6.11.2.7.3.1 SENSE/Rucio Architecture Overview

SENSE (SDN for End-to-End Networked Science at the Exascale) uses SDN to dynamically manage network paths, prioritize traffic, and provide optimized routing. This setup is ideal for data-intensive tasks, as it can allocate resources in real-time based on network conditions and CMS & ATLAS priorities. Rucio, developed by CERN, complements SENSE by managing large-scale data distribution across the network. Its features include automated data replication, strategic data placement, and support for geographically distributed resources, which streamline the movement and availability of critical datasets.

As illustrated in Figure 1, the integration of SENSE with Rucio is based on identifying data flows that require high-priority handling. For example, if Rucio marks specific transfers as critical, SENSE allocates network paths to prioritize them, reducing latency and enhancing throughput. This orchestration involves automated, real-time adjustments across all parts of the data transfer stack, including FTS and XRootD systems, ensuring end-to-end efficiency.

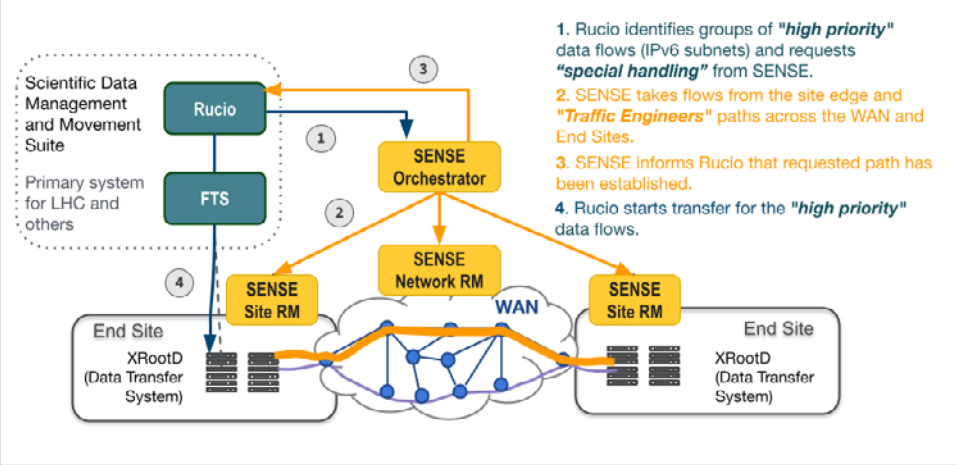


Figure 6.11.1: The SENSE architecture and capabilities: to traffic engineer end-to-end flows with deterministic bandwidth guarantees across a multidomain footprint.

6.11.2.7.3.2 Key Components and Functionalities

The SENSE/AutoGOLE stack incorporates Site and Network Resource Managers (RM) and an Orchestrator to deploy Layer 2 virtual circuits globally. Layer 3 topology support also enables managing multi-subnet flows, which enhances flexibility in accessing various overlays and managing data across complex network structures. The architecture uses automation at the Site RM level, allowing rapid deployment of network topologies, virtual circuit overlays, and server configurations. A Data Movement Manager further enhances automation by responding to multiple requests and allocating bandwidth according to data transfer priority levels.

6.11.2.7.3.3 Testbeds and Site Upgrades

Testbeds at Fermilab (Fermilab), Caltech, UCSD, and the UNL play a crucial role in assessing SENSE/Rucio's performance. These sites are part of the CMS Tier-1 and Tier-2 data center infrastructure, handling large data volumes from the CMS experiment. Notable upgrades, such as Fermilab's 400 Gbps link via LHCONE and Caltech's 4 x 400 Gbps connections, facilitate the evaluation of SENSE/Rucio under high-throughput conditions, especially during peak data collection and processing periods.

On the ATLAS side, the Northeast Tier-2 (NET2), operated by the University of Massachusetts at Amherst and located in the MGHPCC, has started testing SENSE/Rucio from the ATLAS perspective in fall 2024. NET2 is in the process of getting high-performance network connectivity to ESnet and should upgrade its 10 Gbps link to 400 Gbps by mid-November 2024. This upgrade will allow them to test SENSE/Rucio for ATLAS at relevant scale.

6.11.2.7.3.4 Benefits Targeted

The SENSE/Rucio architecture is working towards these advantages:

- **Optimized Data Transfers:** SENSE dynamically assigned network paths with low latency and high bandwidth for priority data transfers, reducing transfer times significantly.
- **Traffic Prioritization:** SDN-based controls allowed critical data to be prioritized, ensuring timely delivery of high-importance datasets for real-time analysis.
- **Enhanced Resource Utilization:** By adapting network paths based on real-time conditions, SENSE improved bandwidth utilization, avoiding static routing limitations.
- **Resilience and Reliability:** The architecture dynamically rerouted traffic around network congestion and hardware failures, which is essential for uninterrupted data transfers during intensive data taking.

6.11.2.7.3.5 Future Plans and Expansion

Based upon demonstrated improvements to existing capabilities, both CMS & ATLAS are planning to expand SENSE/Rucio integration to all Tier-2 sites, transitioning from experimental testbeds to full-scale production. As a part of this expansion, network monitoring and telemetry tools will be deployed to track performance, identify bottlenecks, and optimize configurations.

By enhancing network management capabilities, SENSE/Rucio helps CMS & ATLAS to maintain a scalable, resilient infrastructure that meets the evolving data needs of its experiments. This integration not only ensures faster data transfers but also enables the CMS & ATLAS computing system to offer more predictable, priority-based data workflows essential for timely scientific discoveries.

6.11.2.8 Network and Data Architecture

LHC Run 3 will be completed by the end of 2026 and it is expected that the network traffic will keep growing at the usual rate until then. In the latest years, the LHC community has seen an annual increase of 20-30% of traffic in LHCOPN and LHCONE, the two networks dedicated to transfers between WLCG sites.

The current networking capacities are typically 100 Gbps WAN links with many sites already having resilience (multiple) 100 Gbps connections. This will need to evolve as the LHC community prepares for Run 4. The expected traffic during Run 4 (2029/2030-2032/2033), when the HL-LHC will come into operation, will significantly increase. ATLAS and CMS have estimated this increase in network traffic to be a factor of 10 compared to Run3.

In preparation of HL-LHC the LHC experiments have defined a calendar of "data challenges" to progressively test the capabilities of the WLCG software and hardware infrastructure to meet the requirements. Here is the way they have estimated the network bandwidth:

- ATLAS & CMS T0 to T1 per experiment
  - 350 PB raw data per year; average of 50 Gbps or 400 Gbps during LHC running time
  - Another 100 Gbps estimated for prompt reconstruction data tiers (AOD, other derived output) estimated 1 Tbps for CMS and ATLAS summed
- ALICE & LHCb T0 Export
  - 100 Gbps per experiment estimated from Run 3 rates

From these numbers they have defined two models:

- Minimal Model
  - Sum (ATLAS,ALICE,CMS,LHCb)*2(for bursts)*2(safety-margin) = 4.8 Tbps expected HL-LHC bandwidth
- Flexible Model
  - Experiments may need to reprocess and reconstruct the collected data during the year
  - This requires doubling the bandwidth of the Minimal model: 9.6 Tbps expected HL-LHC bandwidth

Which gives this estimate for sites:

- Each Major Tier-1s:
  - 1 Tbps to the Tier-0 (LHCOPN)
  - 1 Tbps to the Tier-2s (aggregated, LHCONE)
- Each Major Tier-2s:
  - >400 Gbps (LHCONE) and resilient

It is expected that all the major Tier-1s to have at least 800 Gbps LHCOPN links at the start of Run 4 with the ability to grow to 1.2 Tbps or even 1.6 Tbps during the Run. Tier-1s should also have an equivalent amount of bandwidth to LHCONE to serve Tier-2s. Details of the networking for ATLAS and CMS are in the experiment specific case studies (#9 and #10 respectively).

In terms of aggregate WAN capacity out of the site, CMS Tier-1 (Fermilab) currently has provisioned 2.4Tb/s of link capacity to ESnet, via geographically diverse connections. A LAG of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports the Laboratory's general internet traffic, which includes some portion of science traffic, for example, CMS traffic to/from sites not connected to LHCONE.

BNL has two path-diverse 800 Gbps (2 x 400 GbE) circuits that peer with ESnet routers within their 515 and 725 data centers. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits.

A report on the results of the last WLCG Data Challenge can be found online.[114]

6.11.2.9 Emerging Technology Use

6.11.2.9.1 Cloud Services

Many of the LHC experiments have explored cloud services and are using them as part of their global set of resources. There are two primary modes of operation envisioned:

1. A particular site contracts with a cloud provider and resources are (transparently) added as part of the site from their experiment's perspective.
2. A new site is created within a cloud provider, utilizing the various provisioning, management and resilient infrastructure available.

While option 1 above has been used with varying degrees of success, option 2 is not yet feasible for a number of reasons.

- The entity setting up a site within a cloud provider needs to be able to control costs to make it feasible, but there are not usually policies and tools to allow them to do so.
- Egress charges from cloud providers are usually expensive and may vary depending upon where the egress goes to.
- Connecting to resources like LHCONE is very challenging because of the AUP requirement that all traffic from the advertised subnets for LHCONE must be LHC related.

One place that R&E networks could help would be in jointly negotiating with various cloud providers to create a template contract they would support no egress charges if the R&E providers bring high-bandwidth links to the cloud providers, as well as defining policies and associated tools to allow LHC site creators to manage costs by, for example, controlling how data flows to only use the zero (or low cost) egress connections globally.

6.11.2.9.2 AI and ML

This is an area of very active research and development using ML frameworks like but sci-kit-learn, PyTorch, and, to a lesser extent, Keras/Tensorflow and JAX. The experiments rely upon the same Nvidia Cuda platform. However, training tasks are also very experiment specific. See the individual case studies for details.

6.11.2.9.3 Wireless or Distributed Location Science Activities

No wireless capability studies or capability upgrades are planned at present. Some users may do their work from systems connected via wireless, but this is not currently an area of concern for LHC, because most such users use wireless to access remote systems that are doing the actual data transfers or access.

6.11.2.9.4 Data-Centre Interconnect and Spectrum Sharing

The HL-LHC requirements for LHCOPN links approach 1 Tbps and has pushed sites to explore more affordable technologies to achieve it. IT-INFN-CNAF, the Italian Tier-1, has deployed a pilot connection using DCI equipment at the Tier-1 (CNAF) and Tier-0 (CERN) sites, which use four light channels on the dark-fibers of two different providers, GEANT and GARR. The solution can provide 4 x 400 Gbps link for a total of 1.6 Tbps.

114 https://zenodo.org/doi/10.5281/zenodo.11401877

Spectrum Sharing is achieved by interconnecting the ROADM device of the two providers. It saves on the cost of more expensive router interfaces, since most of the existing LHCOPN links are provided as MPLS Layer 2 virtual circuits. The solution is described in detail online.[115]

6.11.2.10 Data-Related Resource Constraints

For LHC experiments in general, but ATLAS and CMS in particular, the resource requirements are growing beyond what flat budgets can support. This leads the experiments to work very hard at optimizing their data and analysis processes to maximize what the available resources can do. Various trade-offs are continually being explored (network use vs multiple replicas of data, data format changes vs number of required datasets to meet various needs, exploitation of nonstandard computing capacity (GPUs, FPGAs, etc.), reconstructing data on demands vs storing it, etc.), to try to achieve the time to science target with constrained resources.

During the 2020-2021 studies of LHC network requirements, the LHC experiments estimated that Tier-1 centers would need terabit/sec connections and Tier-2 sites would require 400G connections. That was driven by estimates of some (but not all) data-related operations of CMS and ATLAS, leading to an estimated bandwidth need of 4.8 terabits/sec (Tbps) across the Atlantic for CMS and ATLAS alone in a so-called "minimal model" and 9.6 Tbps in a more "flexible" and realistic model.

As of 2024, multiple 400G connections have been installed to the Fermilab and Brookhaven national labs as well as HPC facilities used by HEP, and the upgrades to 400G (or N X 400G) at some US CMS campuses are either in place or imminent. As mentioned above, 400G-scale operations in production mode are already being demonstrated, and the standalone capability of individual "Gen5" (PCIe 5.0) inexpensive data servers with multiple 400G and 200G to reach several hundred Gbps was demonstrated at SC23, with Tbps capability per data server on the horizon. Next-generation servers (PCIe 6.0 with Compute Express Link [CXL] memory interconnects) are emerging between now and approximately 2026, and a further technology generation is expected to be available in the mass market before HL-LHC startup in 2030.

Given the above factors: technology drivers and the evolution of ATLAS and CMS's "process of science," the outlook for 2030, the currently foreseen start of the HL-LHC era, it appears that the actual requirements may be somewhat larger than foreseen in 2021. Detailed projections are difficult at this point, but should be tracked and updated every two years or so to make quantitative estimates and (re-)establish the requirements in both network capacity and capability (as in the programmability aspects summarized above).

6.11.2.11 Data Mobility Endpoints

Some experiments utilize Globus endpoints or, more generically, DTNs but this was not the typical design pattern for the LHC experiments. Instead, most of them rely upon the Storage Element design which various storage technologies in use (dCache, XRootD, STORM, EOS, etc.) support. Transfers fill capacity in the WAN by using a many-to-many model with sites typically having ~50 storage servers that serve as either sources or destinations. Many small transfers run in parallel to try to efficiently utilize the available network and storage IO capacity.

6.11.2.12 Outstanding Issues

None to report.

115 https://events.geant.org/event/1582/contributions/1748/attachments/1010/1497/2024-04-09_sig_ngn_SpectrumSharingDCI_Bolletta.pdf

6.11.2.13 Facility Profile Contributors

LHC Operations Representation

Edoardo Martelli, CERN, edoardo.martelli@cern.ch

Shawn McKee, UM, smckee@umich.edu

Syed Asif Shah, Fermilab, sshah@fnal.gov

ESCC Representation

Vincent Bonafede, BNL, bonafede@bnl.gov

Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

Phil DeMar, Fermilab, demar@fnal.gov

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

6.12 LHC: HL-LHC & Network Research

The High-Luminosity LHC (HL-LHC) will collide protons more than a billion times every second, out of which ATLAS and CMS will select interesting collisions for recording at a rate of ~ten thousand times every second.

6.12.1 Discussion Summary

The barrier to using DOE HPC resources still remains, particularly in integrating the systems as resources into the software components, and normalizing the way the resources could be accounted for (expected versus ad hoc or opportunistic). More specifically, there is a desire to create agreements or other forms of arrangements, similar to an SLA, with DOE HPC facilities such that resources can be considered "pledged," similar to how Tier-1s and Tier-2s operate. This helps with planning for both the HEP community and the ASCR resources. Some of this may require discussion at the programmatic level.

An additional operations concern is related to the way that the LHC resources rely on a streaming workflow, the ability of large HPC facilities to support this remains important: either by allowing the worker notes stream directly, or by leveraging caches and data lakes that are deployed close topologically. All of these design choices come with trade-offs, which implies a set of R&D activities that will extend in the coming years.

WLCG's hardware choices for the next major cycles will have a huge influence on the performance of the overall computation and storage system. Hardware choice, and how this relates to overall cost, can fluctuate leading to uncertainty about what capabilities may exist. Some of this leads to speculation about analysis output, data volumes, and the roles of other technology (e.g., networking).

It is expected that GPU use to support some aspects of AI and ML within the LHC community will continue; the challenging parts that are unknown currently are how this may impact WAN usage. Experiments like DUNE have shown that bursting to a commercial cloud to utilize GPUs is possible, but leveraging DOE HPC resource would allow better control over the WAN components. There could be opportunities to schedule network use (e.g., using mechanics like SENSE and Rucio) to coincide with the availability of resources at an HPC facility.

The LHC work on Data Lakes will continue, and has learned several things from the years of R&D and use during activities like DC24. In particular, there is now an effort to reduce the number of "entry points" into storage resources. The XRootD tool right now is not bound by factors such as network capacity or distance/latency, and may find nonoptimal answers to requests for data from clients. The overall goal of a data lake should be to simplify the way data are accessed, along with providing the most performant ways to access and retrieve it. Data lakes can also leverage some of the work on caching, and allow the caches (which often do have more intelligence) to assist in making optimal choices about retrieval.

The use of in-network caching was discussed at length to identify ways that ESnet and the LHC community can work together more efficiently to design and operate an effective service. ESnet has some previous work in attempting to create a "CDN for Science" with others in the DOE, but the usage patterns never amounted to a measurable increase in performance. A core R&D question to address remains: how large should the caches be, and what policies on data retrieval and retention will be valuable to implement to keep cache hit rates high. To date there are not many good answers due to the sometimes chaotic nature of the uses of the cache. For example, user analysis between facilities may focus on completely different portions of a larger data set; within a facility it may be more relevant and can take advantage of caching. Workflow tools that are integrated to use a cache get a small uptick in performance in these cases, but may also use nonoptimal reasoning when retrieving data sets. ESnet would like to gain better monitoring insight into the caching work, either by having access to LHC data (e.g., MONIT) or by having more network layer visibility integrated into Stardust.

As in the discussion after the LHC Operations use case, ESnet will continue to work with the LHC community to advance R&D efforts that support the HL-LHC operations era. This includes the already mentioned efforts (e.g.,

SENSE and Rucio, SciTags), but also ways that HPC resources can be integrated and utilized more efficiently. The use of ESnet High Touch data, and integration with Stardust, may be critical in passing along network-level telemetry to the storage and computational components of the LHC community.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The CMS experiment anticipates the following annual data volume for each data type during HL-LHC operation: [Section 4.2]
  - raw Data: 250 PB
  - AOD Data: 210 PB
  - MiniAOD Data: 27 PB
  - NanoAOD Data: 0.6 PB
- The ATLAS experiment anticipates the following annual data volume for each data type during HL-LHC operation: [Section 4.2]
  - raw Data: 168 PB
  - AOD Data: 200 PB
  - Derived AOD (DAOD) Data: 76 PB
  - DAOD_PHYSLITE Data: 2.6 PB
- The exact computing model for the HL-LHC has not been finalized, given that the LHC Run is not expected to start till 2030. It is assumed that the current computing models will be the baseline model, with some small improvements. There will be no fixed hierarchy of computing sites for most data processing and data access services. In order to improve usage efficiency, sites will be primarily categorized by size, service level, and capability. Hence the location of datasets and users is not deterministic. Tier-1s will need additional capacity to handle worldwide traffic, and Tier-2s will be scoped larger for national traffic needs. The network capacity should be provisioned to match the scale of the available resources at each site. [Section 4.3]
- PanDA, ProdSys and Rucio are being extended for the demands of HL-LHC. Rucio has been adopted as the LHC and WLCG standard. A notable difference between the two collaborations is that CMS more heavily relies on streaming data from storage endpoints to remote processing centers via the XRootD protocol. CMS is currently re-evaluating its workflow management infrastructure solution, which may lead to changes in network usage patterns if a new system is adopted. [Section 4.3]
- ATLAS has chosen to evaluate two scenarios based on how aggressively it is anticipated that the HL-LHC storage and computing R&D program will deliver improvements: [Section 4.3]
  - "Conservative" R&D is a program where work can be completed with the existing person-power. The R&D activities currently underway to support Run 3 will be put into place, and will encourage an almost universal adoption of the DAOD_PHYSLITE format.
  - A full program of "aggressive" higher-risk R&D can only be completed with new person-power. ATLAS implements new developments that very significantly improve the speed and storage volumes of workflows that currently are heavy consumers of resources.
- There have been significant efforts made to understand the computing and storage requirements for HL-LHC. The process began by extrapolating current methods forward, accounting for the data increases from HL-LHC and the additional complexity associated with high-luminosity operations of the collider. Further assumptions about the initially conservative

estimates of computing and storage capacities would be increased for fixed cost and additionally assumed the notion of a "flat budget" for the foreseeable future. To decrease the gap, significant efforts were made to better optimize how less resources could be utilized to do the same amount of science. This has helped decrease the difference between what is needed and what is affordable. While there are potentially some optimistic assumptions in this estimate, it is known that the planning process must change to ensure resource availability on the time scale of HL-LHC. At this point there is still a significant gap in the amount of storage believed to be needed versus what can be provisioned with current budget projections. [Section 4.4]

- The impact of HL-LHC on ATLAS and CMS storage and compute resources is significant. As previously noted, the increase in luminosity not only generates significantly more data but also significantly more complex events which require more processing to resolve. [Section 4.4]
- The expanded use of High-Performance Computers (HPC) will also have an impact on the compute resources (storage and networking). These HPC centers are increasing in computing power and several exaflop-scale machines will be operational during the start of the HL-LHC. These centers are more capable of processing larger and larger volumes of data; this means that the workloads that will run on these machines will evolve. HPC's will be capable of producing a large volume of simulated data. [Section 4.4]
- Global R&E network use will have different characteristics on the timescale of the HL-LHC, compared to the situation the WLCG experienced in Runs 1, 2, and 3. Specifically, the experiments foresee other science domains (astronomy, biology, engineering, and others) becoming global network users at scales equaling or exceeding the LHC users. In such an environment, it is not wise to assume the experiments can continue to treat the network as an infinite resource . The LHC community must explore options for effectively operating within a bandwidth constrained environment. [Section 4.7]
- Networking has been fundamental to the success of ATLAS and CMS computing to date, enabling the exploitation of globally distributed resources for computationally limited science. This will remain the case to meet the budget-constrained computing challenges of HL-LHC. Strategies for HL-LHC computing are based on extensive use of powerful networks to reduce data replication by streaming over the network, and consolidating distributed resources into cohesive virtual federations such as data lakes. HL-LHC data and processing scales are large enough that optimizing for efficient bandwidth use will be essential. The ability to mark network packets to identify sources of traffic, and to route traffic to control speed and cost may become vital. [Section 4.7]
- As a rough guide as to the network bandwidth that facilities in US CMS will require in the HL-LHC era for the purpose of reconstructing the raw data and producing corresponding simulation datasets, one can take the following estimate: from the 250 PB of raw data collected per year, about 40% (e.g., 100 PB) is expected to be reconstructed in the US. It is desired that these large-scale reconstruction campaigns take approximately 10 weeks, implying a data input rate to processing facilities of 10 PB/week, and an output rate of 4 PB/week. In parallel, simulation production campaigns will be executed continuously throughout the year, producing another 40% of the total 240 PB of output data (e.g., 96 PB) to be persisted on tape and/or disk storage within the US. In total, these workflows will require a sustained rate of up to 130 Gbps between processing and storage facilities. To account for burst traffic, and to support data analysis workloads at AFs discussed below, it is desired that all major processing facilities, namely the Fermilab Tier-1 and US CMS Tier-2 centers, are provisioned to 400 Gbps. [Section 4.7]

- At present, it is assumed that the general LHC Tier-2 network requirements as specified above will be sufficient to support AFs at the Tier-2s. A user at an AF may expect to process a year's worth of MiniAOD data and simulation (27 PB) in a timescale of a week. If the data are not cached at the AF, that amounts to a sustained 400 Gbps network load from the data origin site. In practice, it is unlikely a user would read the entirety of the data catalog, though the exact fraction will vary significantly depending on the analysis goals. [Section 4.7]
- The global infrastructure is currently relatively uniform, with approximately 70 compute and storage elements presenting an identical interface to CMS. The data distribution model proposed by WLCG for the HL-LHC is commonly referred to as the "Data Lake Model." The data lake model envisions fewer regional entry points than the 70 storage elements currently available, allowing internal optimizations and potentially reducing the operational burden on international CMS. Data transfer between two lakes is a top-down-controlled activity governed by Rucio and executed via FTS using third-party copy HTTPS transfer protocols with capability token authentication. Access to data from processing elements inside the lake is mediated either via caches or streaming directly from one of the data lake origins. [Section 4.8]
- ESnet and the LHC experiments (ATLAS and CMS) have started an R&D project that makes use of in-network data caches operated by ESnet using XRootD software. The pilot project focuses on caching of PHYSLITE data products to see if the time to reach physics results can be shorted, and the overall network traffic reduced for data analysis (only used parts of data files cached and reused). [Section 4.8]
- During HL-LHC the tracking algorithms used in event reconstruction are expected to see increased use of GPUs. It is conceivable parts of the events will need to be shipped to data centers (HPC's, Cloud vendors, WLCG data centers) for GPU processing with the results sent elsewhere for recombination with the data processed on CPU-only resources. Future DOE HPC machines are focusing on support for AI/ML, with large numbers of GPUs and other accelerators. ML training requires large amounts of data transferred into the HPC's for use during the training. This will have an impact on the WAN network as data are transferred from WLCG data centers to HPCs in the US for ML training and inference. [Section 4.8]
- From the ATLAS and CMS point of view, the IRI long-term campaign pattern best fits the LHC use case. Historically, the collaboration expects a full year's worth of reconstructed data and simulation to be produced on a few-month timescale, so there is relatively little time sensitivity. Online DQM and time-sensitive calibration procedures are handled by a separate coordination area of the ATLAS and CMS experiments and have dedicated resources at CERN, the host laboratory. In addition, since each particle collision event is an independent identically distributed quantity, the HL-LHC science workflow is trivially parallelizable and does not need large-scale data shuffle operations to execute. [Section 4.9]
- If, under the IRI, HPC centers could provide some of the services found at the WLCG data centers (for example proxy caches, CVMFS outbound connectivity from the worker nodes - through a reverse proxy), both ATLAS and CMS will be able to make better use of these resources. [Section 4.9]
- Most of the workflows that ATLAS will use during HL-LHC will fall under the IRI category of data-intensive workflows (event reconstruction and ML) or long-term campaign (large-scale MC simulations). IRI will also provide the opportunity to implement some near real-time workflows for critical payloads. [Section 4.9]
- For CMS, the US provides, depending on the resource type, 30-40% of the global computing and storage capacity. Since the international CMS collaboration is responsible for operating a common computing infrastructure, if CMS is to effectively use IRI, it will be necessary

to ensure that the interface to computing and storage resources made available by IRI is compatible with the global infrastructure, including university and international partners. It is expected that the current WLCG interface (Compute and Storage Elements) that CMS is presented with, will evolve in preparation for HL-LHC, though the exact nature of this evolution is not yet determined. [Section 4.9]

- Given the compatibility of CMS workflow pattern with IRI plans, previous experience with integrating HPC facilities, and the need to provide a homogenous computing interface to an international collaboration, CMS would be an excellent choice for a long-term campaign pathfinder project for the IRI.[Section 4.9]
- The current members of the collaborations are not expected to encounter major changes to the membership for the HL-LHC. The collaborators will continue to be distributed worldwide, which puts strong requirements on global networking. Specifically, it is expected that computing resources will continue to be globally distributed, with no centralized data or computing location. Physicists will need access to the distributed resources from all corners of the world. [Section 5.1]
- In the HL-LHC era, it is expected that, in terms of the current computing model, the US will continue to provide one of the Tier-1 and several Tier-2 facilities, fully funded via DOE and NSF respectively, as resources for CMS. It is expected that US CMS will make use of any HPC centers funded by either DOE or NSF that will grant allocations. Commercial cloud providers are not expected to be a major component of the CMS computing infrastructure. [Section 5.1]
- The High-Luminosity LHC (HL-LHC) will collide protons more than a billion times every second, out of which ATLAS and CMS will select interesting collisions for recording at a rate of ~ten thousand times every second. With over two thousand hours of data collection every year when the HL-LHC starts running in 2030, ATLAS and CMS will have a huge data sample for physicists to analyze worldwide. The HL-LHC program is expected to last for a decade. Large improvements in networking will be required to enable the ambitious physics goals of the HL-LHC. The entire HL-LHC era will last for 10 years of data taking, with 12-24 month maintenance periods interspersed roughly every three years. [Section 5.1]
- For the HL-LHC era, the predictions show a mismatch between the computing and storage resources the experiments can afford versus the resources needed to reach science goals. In response to this gap, the experiments are exploring alternatives in how to utilize storage, computing and network infrastructure. The network baseline currently being planned is composed of terabit-scale (1–2 Tbps) backbone networks with the largest resource sites connected at multiple 400 G scale (800 Gbps+). Network use will be at least a factor of 10 larger than Run 3. [Section 5.1]
- The individual LHC experiments are interested in discussing the set of features that must be present in future network designs. If it is possible to identify a minimum set of capabilities in switches and routers, to ensure the next generation of network components can meet the needs of the community, it would simplify the procurement and installation process. The sooner those capabilities are in place, the sooner prototyping and testing future capabilities will be possible. [Section 5.7]
- For HL-LHC networking, four main requirements have been identified: [Section 5.7]
    - Run 3 is moving to multiple 100G links for large sites, while Run 4 (HL-LHC) is targeting Tbps links, perhaps based upon either 400G or 800G waves.

- It is necessary to understand the impact of new features in networking (SDN/NFV) by testing, prototyping and evaluating impact. The experiments will need to evolve applications, facilities and computing models to meet the HL-LHC challenges; it will take time.
- The ability to understand WAN network flows is limited. New methods to mark and monitor network use are needed. Packet marking is viewed as a high priority so that everyone (experiments, sites and R&E networks) can understand the origin and intent of network flows at any point along their path.
- Developing, prototyping, and testing network features at suitable scale will be needed. Having at-scale network testbeds will be very important to understand the potential impact changes might have on operations.

- One interesting option to explore on the timescale of HL-LHC is the joint ATLAS and CMS use of Rucio for distributed data management. This could provide a mechanism to interact with ESnet (and other R&E networks), communicating near-term data-movement intents and perhaps negotiating for any required QoS or deadline requirements. [Section 5.8]

6.12.2 HL-LHC: Network Research Profile

With over two thousand hours of data collection every year when the HL-LHC starts running in 2030, ATLAS and CMS will have a huge data sample for physicists to analyze worldwide. The HL-LHC program is expected to last for a decade. Large improvements in networking will be required to enable the ambitious physics goals of the HL-LHC.

6.12.2.1 Science Background

As shown in Figure 6.12.1, the HL-LHC era starting in 2030 will accumulate roughly the same integrated luminosity of data in three years of LHC running as the entire period of running of the LHC has produced up to then. All of that data will be accumulated at a center of mass energy of 14 TeV. This implies that the science capabilities of the first years of HL-LHC data taking at design luminosity, in 2030-2032, is expected to be roughly equivalent to the data taking from 2010-2025, or Runs 1, 2, and 3 combined.

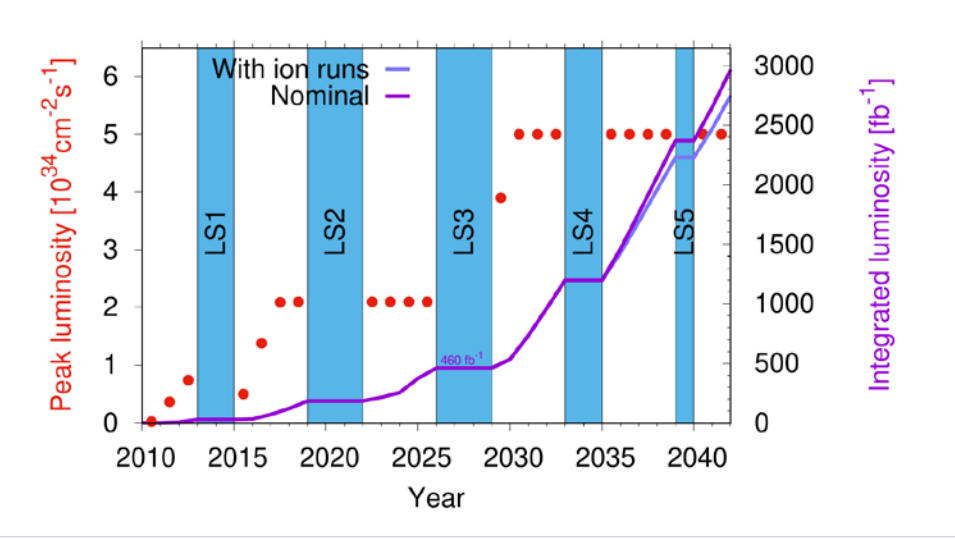


Figure 6.12.1: The expected maximum luminosity and integrated luminosity for the LHC as a function of the calendar year. Note this figure will be updated to account for the recent schedule change with HL-LHC starting in May 2030.

The entire HL-LHC era will last for 10 years of data taking, with 12-24 month maintenance periods interspersed roughly every three years[116]. Each data taking year, the accelerator expects to deliver 250 fb-1 of data[117] to the ATLAS and CMS experiments.

The ATLAS and CMS estimated annual data volumes are shown in Table 1 below. More details are provided in the ATLAS and CMS case studies (Section 6.9 and 6.10 respectively). Both experiments make the same assumptions around how often data must be processed and reprocessed over the course of a year. To understand totals as well as differences, it is thus sufficient to just look at the current expectations for the first year of full HL-LHC luminosity, in 2030. Section 6.11 (LHC Operations) includes figures that overlay the CPU, disk, and tape needs versus time computed from the current assumptions. Differences in some of the parameters between the two experiments result from different choices of analysis procedures and are largely driven by sociology.

CMS Name	CMS Annual Volume	ATLAS Name	ATLAS Volume
RAW	250 PB	raw	168 PB
AOD	210 PB	AOD	200 PB
MINI	27 PB	DAOD	76 PB
NANO	0.6 PB	DAOD_PHYSLITE	2.6 PB

Table 6.12.1: Expected annual unique data volumes for ATLAS and CMS in 2030-2032. The "raw" data are only from the detector only while all others include detector data and simulations. For raw data, the assumptions going into the figures include a 10kHz rate of events, 6e6 seconds live time per year, and a 4.3 MB (CMS) or 2.8 MB (ATLAS) event size.

6.12.2.2 Collaborators

The current members of the ATLAS collaboration are described in Section 6.9. It is not expected that major changes to the membership will occur for the HL-LHC. The collaborators will continue to be distributed worldwide, which puts strong requirements on global networking. Specifically, it is expected that computing resources will continue to be globally distributed, with no centralized data or computing location. Physicists will need access to the distributed resources from all corners of the world.

The international nature of CMS, location of its facilities, and location of any shared HPC centers is essentially unchanged over time. Refer to Section 6.10 for more details.

The exact computing model for the HL-LHC has not been finalized, given that the LHC Run is not expected to start till 2030. It is assumed that the current ATLAS computing model will be the baseline model, with some small improvements. There will be no fixed hierarchy of computing sites for most data processing and data access services. In order to improve usage efficiency, sites will be primarily categorized by size, service level, and capability. Hence the location of datasets and users is not deterministic - ATLAS expects worldwide distribution of all resources and users. ATLAS also expects about seven large sites in the US, which will all be required to have a full range of distributed computing capabilities. They will store both primary and secondary data, will provide access to hundreds of users, and will participate in continuous data transfers. These sites will include the BNL Tier-1, the current Tier-2's (Great Lakes, Midwest, Northeast, and Southwest), SLAC, and the larger DOE and NSF HPC's. It is expected that a few 400 Gbps links will be needed between them when the HL-LHC starts. BNL Tier-1 will need additional capacity to handle worldwide traffic. The network capacity should be provisioned to match the scale of the available resources at each site.

116 https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm

117 Integrated luminosity, measured in inverse femtobarn, equivalent to 1039 cm-2, relates theoretical predictions for scattering probabilities to expected number of (interesting) particle collision events in the detector.

6.12.2.3 IRI Readiness

The IRI envisions three primary workflow patterns: time-sensitive; data-integration intensive; and long-term campaign. From the ATLAS and CMS point of view, the long-term campaign pattern best fits the LHC use case. Historically, the collaboration expects a full year's worth of reconstructed data and simulation to be produced on a few-month timescale, so there is relatively little time sensitivity. Online DQM and time-sensitive calibration procedures are handled by a separate coordination area of the ATLAS and CMS experiments and have dedicated resources at CERN, the host laboratory. In addition, since each particle collision event is an independent identically distributed quantity, the HL-LHC science workflow is trivially (or, pleasantly) parallelizable and does not need large-scale data shuffle operations to execute.

HEP experiments in general can benefit from IRI, especially if LCF and HPDF are seamlessly integrated into the existing HTC (grid) infrastructure (including OSG). The IRI will provide a unified programmatic interface to both compute and data flow. This standardization of interfaces across the DOE HPC centers will make it possible to use these resources by the time of HL-LHC.

If, under the IRI, HPC centers could provide some of the services found at the WLCG data centers (for example proxy caches, CVMFS outbound connectivity from the worker nodes - through a reverse proxy), both ATLAS and CMS will be able to make better use of these resources.

Most of the workflows that ATLAS will use during HL-LHC will fall under the category of data-intensive workflows (event reconstruction and ML) or long-term campaign (large-scale MC simulations). IRI will also provide the opportunity to implement some near real-time workflows (WF) for critical payloads such as High-Level Trigger WF, which should produce results in a near-real-time manner. As an example, the BNL network infrastructure has been designed with IRI patterns in mind. This is shown in the network design of the external Science DMZ network for data transfer and a firewalled science portals network for services. IRI benefits HEP by providing a uniform interface to disparate resources in a way that is easier to integrate with, better fit to HEP workflows. The ATLAS workflows that benefit most from an IRI approach are those currently running at LCF and HPCs.

CMS has extensive experience integrating HPC sites within the US In particular, the Fermilab Tier-1 computing site uses the HEPCloud interface to extend its computing capacity provided to CMS by roughly double through use of the NERSC Perlmutter and TACC Frontera supercomputers. CMS expects that, in IRI, long-term campaign workflows will be at risk of pre-emption to make room for time-sensitive workflows, and prefers a service level agreement characterized by the integrated core-hours made available over some time window. For example, an allocation might provide a total of 100M core-hours with a minimum 20M core-hours over a one-month window. This allows for reliable long-term campaign planning. R&D to increase the resilience of CMS workflows to pre-emption will be needed to fully exploit such an infrastructure.

For CMS, the US provides, depending on the resource type, 30–40% of the global computing and storage capacity. Since the international CMS collaboration is responsible for operating a common computing infrastructure, if CMS is to effectively use IRI, it will be necessary to ensure that the interface to computing and storage resources made available by IRI is compatible with the global infrastructure, including university and international partners. It is expected that the current WLCG interface (Compute and Storage Elements) that CMS is presented with, will evolve in preparation for HL-LHC, though the exact nature of this evolution is not yet determined.

Given the compatibility of CMS workflow pattern with IRI plans, previous experience with integrating HPC facilities, and the need to provide a homogenous computing interface to an international collaboration, CMS would be an excellent choice for a long-term campaign pathfinder project for the IRI.

6.12.2.4 Instruments and Facilities

The impact of HL-LHC on ATLAS and CMS storage and compute resources is significant. As previously noted, the increase in luminosity not only generates significantly more data but also significantly more complex events which require more processing to resolve.

The expanded use of High-Performance Computers (HPC) will also have an impact on the compute resources (storage and networking). These HPC centers are increasing in computing power and several exaflop-scale machines will be operational during the start of the HL-LHC. These centers are also becoming more capable of processing larger and larger volumes of data than they did in the past. This means that the workloads that will run on these machines will evolve. HPC's will be capable of producing a large volume of simulated data. This simulated data will be reconstructed at the HPC where it was produced. The output files from the simulation step and the reconstruction step will need to be quickly transferred to ATLAS and CMS data centers for subsequent processing. During HL-LHC the tracking algorithms used in event reconstruction are expected to see increased use of GPUs. The foreseen reconstruction processing will be a mix of CPU and GPU codes. It is conceivable parts of the events will need to be shipped to data centers (HPC's, Cloud vendors, WLCG data centers) for GPU processing with the results sent elsewhere for recombination with the data processed on CPU-only resources. The DOE HPC machines are being designed to be very efficient at AI/ML with large numbers of GPUs and other accelerators. ML training requires large amounts of data transferred into the HPC's for use during the training. This will have an impact on the WAN network as data are transferred from WLCG data centers to HPCs in the US for ML training and inference.

To upgrade to HL-LHC will involve numerous changes to the accelerator. The LHC particle beam will need to be more intense and more focused than at present. New or upgraded components will need to be installed in various sections of the LHC's 27 kilometers ring. In addition to the accelerator upgrades, the ATLAS and CMS experiments also need to implement significant upgrades and detector replacements to handle the much more challenging environment of the HL-LHC.

6.12.2.4.1 ATLAS Computing Infrastructure

The HL-LHC refers to the upgraded LHC configuration that will be implemented during Long Shutdown 3 (LS3). Luminosity is a measure of the number of collisions per unit area for a given period of time. More luminosity means more particle collisions and more opportunities to measure new, interesting physics. A good way to measure discovery potential is by integrated luminosity which is measured in inverse femtobarns (fb^{-1}). An inverse femtobarn equates to 100 trillion collisions. The integrated luminosity for all LHC Runs 1 and 2 was about 150 inverse femtobarns of data and during the ongoing Run 3 also about 150 inverse femtobarns have been recorded so far. The HL-LHC will produce more than 250 inverse femtobarns of data per year and will be capable of collecting up to 3000 inverse femtobarns.

One of the significant challenges posed by HL-LHC is the increase in the average number of particle collisions each time particle bunches cross each other (which happens around 40 million times per second). The Run 3 average number of collisions per crossing is around 50 but HL-LHC will increase that to 150-200. It is important to note that the impact of this increase in processing power is worse than linear and a priori exponential, primarily because of the increased combinatorics involved in assigning particles to collisions.

In the ATLAS Software and Computing HL-LHC Roadmap, ATLAS has chosen to evaluate two scenarios based on how aggressively it is anticipated that the HL-LHC storage and computing R&D program will deliver improvements. They are useful straw-men to discuss model uncertainties and development costs.

The two scenarios, defined in terms of the R&D program activities described in the next section, are as follows:

Conservative R&D: "Conservative" R&D is assumed to be work that can be completed with the existing person-power in the ATLAS software and computing community, as long as the overall staffing levels are maintained and key experts remain in ATLAS. The research and development activities currently underway for Run 3

are assumed to be successful, including the data carousel, fast track reconstruction, and most of the detector simulation being performed with fast simulation. Almost universal adoption by the physics groups of DAOD_PHYSLITE would also fall into this category.

Aggressive R&D: The full program of "aggressive" higher-risk R&D can only be completed with new person-power, either in the form of new individuals, or existing contributors being permitted to spend a higher fraction of their time on software and computing development for ATLAS. There are a number of higher-risk projects where work is ongoing already with the available person-power. Since there is significant uncertainty in the impact on resources these projects might have, their potential impact is included in the "aggressive" R&D projection. ATLAS implements new developments that very significantly improve the speed and storage volumes of workflows that currently are heavy consumers of resources. For example, porting of high-precision generators to GPUs, sharing events with CMS, or speeding up the full simulation either by software efficiencies or porting parts of the code to GPUs. Developments of very high quality fast simulation that could replace full simulation in almost all cases, would also fall into this category. Some R&D activities in the aggressive category cannot yet be quantified in their impact and so are not yet included in the model. An example is ML models for fast simulation and reconstruction.

Figures 6.12.2 and 6.12.3 show the estimated CPU and storage needs under these different scenarios, to be compared with provisioned resource curves based on flat funding for CPU and storage, and capacity growth per unit cost of 10% and 20% per year. US ATLAS in its current best-estimate budget planning assumes 10% growth (the lower curve) for the years out to HL-LHC, based on an assessment of the technology landscape. Other estimates within the HEP community range from 5% to 20%. These studies demonstrate that ATLAS (and US ATLAS) must pursue vigorously aggressive R&D options to be able to meet the HL-LHC challenges in a sustained budget model.

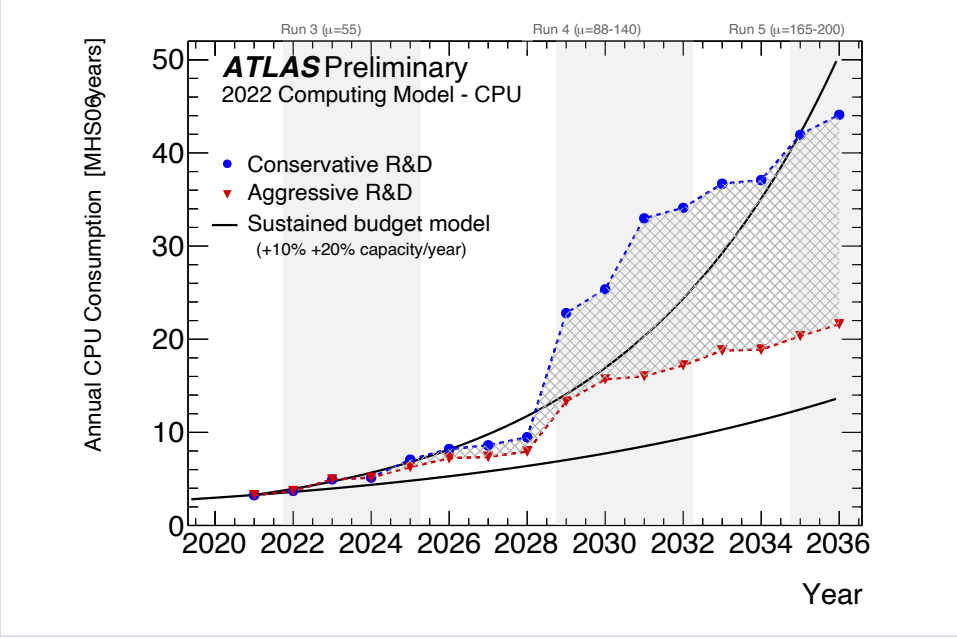


Figure 6.12.2: Estimated CPU resources needed for the years 2020 to 2036 under the different scenarios. The solid lines indicate annual improvements of 10% and 20% in the capacity of new hardware for a given cost, assuming a sustained level of annual investment. The US resources are typically 23% of the resources shown here. Note these will need to be updated to include the new start date for HL-LHC of May 2030.

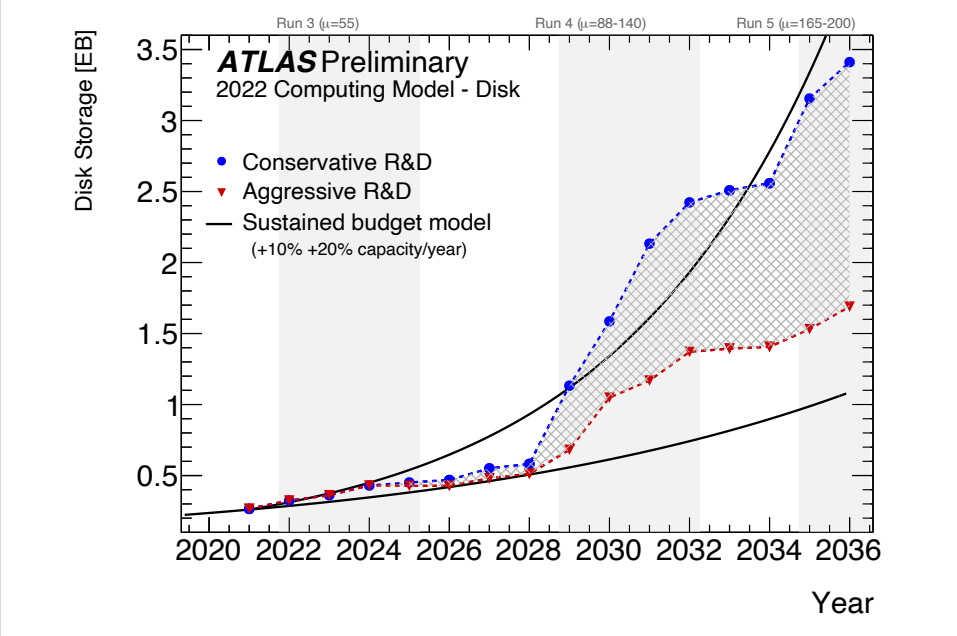


Figure 6.12.3: Estimated disk resources needed for the years 2020 to 2036 under the different scenarios. The solid lines indicate annual improvements of 10% and 20% in the capacity of new hardware for a given cost, assuming a sustained level of annual investment. The US resources are typically 23% of the resources shown here. Note these will need to be updated to include the new start date for HL-LHC of May 2030.

6.12.2.4.2 CMS Computing Infrastructure

In the HL-LHC era, it is expected that, in terms of the current computing model, the US will continue to provide one of the Tier-1 and several Tier-2 facilities, fully funded via DOE and NSF respectively, as resources for CMS. They are run as coordinated facilities via the US CMS Operations Program for central processing, generation of simulated datasets, and, at a smaller scale, data analysis. CERN will operate the Tier-0, a tape archive, and a Tier-2 for CERN-resident physicists. In addition, it is expected that US CMS will make use of any HPC centers funded by either DOE or NSF that will grant allocations. As discussed in section 6.12.2.9, commercial cloud providers are not expected to be a major component of the CMS computing infrastructure. The global infrastructure is currently relatively uniform, with approximately 70 compute and storage elements presenting an identical interface to CMS.

The data distribution model proposed by WLCG for the HL-LHC is commonly referred to as the "Data Lake Model." The data lake model envisions fewer regional entry points than the 70 storage elements currently available, allowing internal optimizations and potentially reducing the operational burden on international CMS. In a data lake, behind each entry point are generally multiple storage endpoints that are geographically distributed. These storage endpoints will be referred to as "data origins." An entry point may support multiple levels of QoS, depending on the cluster. QoS differences may include data access performance, latency, levels of protections against data loss, etc. The exact definitions of the QoS categories and their expected implementations are still a subject of R&D and discussions. Data transfer between two lakes is a top-down-controlled activity governed by Rucio and executed via FTS using third-party copy HTTPS transfer protocols with capability token authentication. Access to data from processing elements inside the lake is mediated either via caches or streaming directly from one of the data lake origins.

This vast quantity of data must be distributed around the globe for processing and physics analysis. The data distribution model for the HL-LHC is commonly referred to as the "Data Lakes Model." A lake is defined as a cluster of storage and nearby computing facilities that have a single entry point for interactions outside of the lake. Figure 6.12.4 shows this model as a graphic with two lakes, though many more lakes will be needed to meet CMS needs for processing and analysis worldwide. Further details on the data lake model are presented in Section 6.12.2.4.

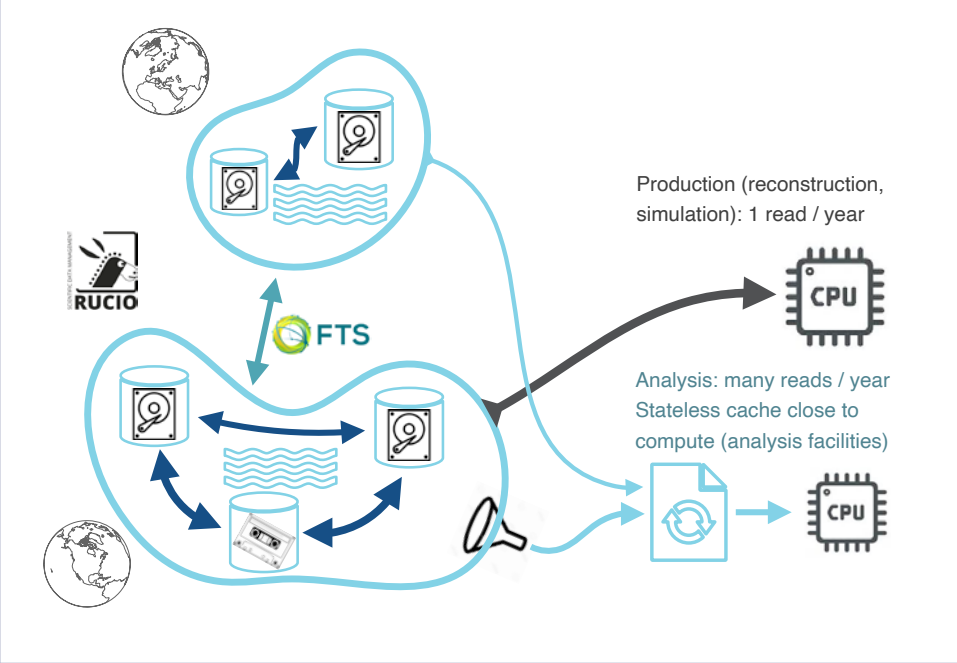


Figure 6.12.4: WLCG Data Lakes Model

6.12.2.5 Generalized Process of Science

The computing workflows and workloads described in Sections 6.9 and 6.10 for simulation, data processing, derivations and analysis will continue to be required, with some significant improvements noted below.

The HL-LHC, commencing with Run 4 in 2029, will deliver unprecedentedly complex events, with up to 200 interactions per proton-proton bunch crossing. These events will be collected at a prodigious rate: ATLAS and CMS expects to record data at 10kHz, approximately five to ten times more than during previous runs. By the end of Run 6 in 2041, the HL-LHC is expected to have delivered an integrated luminosity of up to 3000 fb-1, several times more than all previous runs combined. As well as the challenges involved in collecting, storing, reconstructing and analyzing such a colossal volume of data, MC simulation events will need to be produced in similar numbers. Taken together, the data and MC requirements of the HL-LHC physics program are formidable, and if the computing costs are to be kept within feasible levels, substantial improvements must be made in both compute and storage.

Networking has been fundamental to the success of ATLAS and CMS computing to date, enabling the exploitation of globally distributed resources for computationally limited science. This will remain the case to meet the budget-constrained computing challenges of HL-LHC. Strategies for HL-LHC computing are based on extensive use of powerful networks to reduce data replication by streaming over the network, and consolidating distributed resources into cohesive virtual federations such as data lakes. HL-LHC data and processing scales are large enough that optimizing for efficient bandwidth use will be essential.

As part of this optimization, the ability to mark network packets to identify sources of traffic, and to route traffic to control speed and cost may become vital. To this end, ATLAS initiated the HEPiX Network Function Virtualization WG in collaboration with other experiments and national entities including ESnet. US ATLAS will play central roles in the R&D in this area, drawing on the close relationship with ESnet, the long standing networking expertise within the facilities, and distributed computing leadership in PanDA and its Rucio integration; shaping network traffic through Rucio and PanDA will be crucial.

Economizing storage is an important goal for HL-LHC computing. Unlike CPU, the requirements for which will become approximately constant once the LHC reaches its design luminosity, storage needs will continue to increase during the lifetime of the HL-LHC. And while opportunistic computational resources exist, opportunistic storage does not. Innovative ideas for optimizing storage by breaking out of the disk/tape paradigm to a finer grained spectrum of storage cost-reliability-latency are being pursued. In particular, the US-driven ATLAS data carousel project is developing a mechanism to stage data (such as AODs) from tape to a sliding window disk buffer when they are required for processing, reducing by 50% or more the input sample volume resident on disk. The data carousel leverages the managed 'train production' approach already utilized by ATLAS to drive a tightly integrated orchestration of workflow and workload management that processes data with file-level granularity as soon as it appears from tape, keeping its disk residency to a minimum. With tape facilities at only ~10 sites, and processing resources distributed across over 100 sites, remote processing requiring remote data delivery is an essential element of the approach.

Remote data delivery reliant on powerful networks will in general be essential to minimize data replication and disk storage footprint while fully utilizing distributed processing resources. In order to most efficiently use bandwidth and minimize latencies, ATLAS (largely US ATLAS) is developing new services and workflows to deliver across the network only the data needed by the consuming workload, with fine granularity that avoids the inefficiencies and latencies of data pre-staging and enables the use of processing resources to be highly dynamic while remaining efficient. The intelligent data delivery service (iDDS) provides support for such fine-grained workflows, with its ServiceX component providing intelligent data transformations and filters to deliver just the needed data.

As a rough guide as to the network bandwidth that facilities in US CMS will require in the HL-LHC era for the purpose of reconstructing the raw data and producing corresponding simulation datasets, one can take the following estimate: from the 250 PB of raw data collected per year, about 40% (e.g., 100 PB) is expected to be reconstructed in the US. It is desired that these large-scale reconstruction campaigns take approximately 10 weeks, implying a data input rate to processing facilities of 10 PB/week, and an output rate of 4 PB/week. In parallel, simulation production campaigns will be executed continuously throughout the year, producing another 40% of the total 240 PB of output data (e.g., 96 PB) to be persisted on tape and/or disk storage within the US. In total, these workflows will require a sustained rate of up to 130 Gbps between processing and storage facilities. To account for burst traffic, and to support data analysis workloads at AFs discussed below, it is desired that all major processing facilities, namely the Fermilab Tier-1 and US CMS Tier-2 centers, are provisioned to 400 Gbps.

6.12.2.5.1 Analysis Facilities

A significant user-facing component of HL-LHC computing is distributed physics analysis, which is expected to be centered at a variety of AFs. These are dedicated pieces of infrastructure designed to provide access to large data sets and computational resources that enable rapid iterative analysis of physics data. Figure 3 shows how the Analysis Facility concept fits into the Data Lakes model. The AF is a relatively new concept discussed within IRIS-HEP, HSF, and WLCG. A motivation for this concept is the fact that columnar analysis of data formats like the NanoAOD promise to provide a very large increase in data analysis speed, as measured in number of events analyzed per second. Typical analyses of MiniAOD in Run 2 achieve processing rates O(10Hz) while columnar analysis of NanoAOD promises to achieve O(1–10kHz) or more, both per hyperthread. As such, these new processing paradigms allow an interactive feel. The AF enables physicist users to interact with their data in a

scalable way without sending several-hour-long batch jobs and waiting for their completion before digesting the results and preparing the next iteration. Several salient features are:

- One or more analysis software frameworks that take care of selecting the appropriate data access method, planning the query, and scaling out the resulting workflow to be executed. At present, users are largely responsible for hand-crafting each of these components.
    - Data access and query planning may also involve requesting data from the edge of a data lake nearby to the AF, potentially through the ServiceX platform. This pattern is much the same as predicate pushdown in local clusters, but targeting a reduction of WAN bandwidth.
- A modern user interface such as JupyterHub, while still retaining shell and scriptable access through well-designed APIs
- Facility hardware, provided by the US CMS operations program, including high-performance disk or NVMe caches and sufficient processing capabilities to provide a target "15-minute HL-LHC analysis turnaround time on NanoAOD"

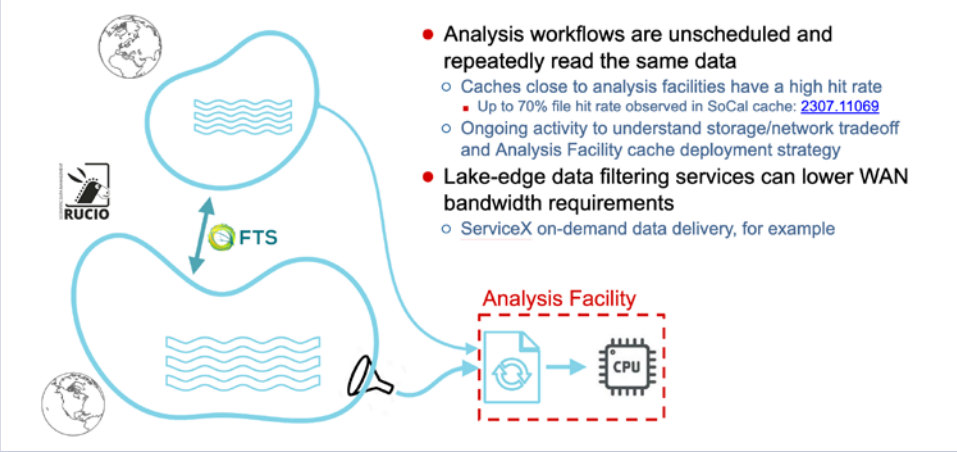


Figure 6.12.5: Depicts the AFs concept, how it fits into Data Lakes Model, and lists some of its features.

At present, it is assumed that the Tier-2 network requirements as specified above will be sufficient to support AFs at the Tier-2s. A user at an AF may expect to process a year's worth of MiniAOD data and simulation (27 PB) in a timescale of a week. If the data are not cached at the AF, that amounts to a sustained 400 Gbps network load from the data origin site. In practice, it is unlikely a user would read the entirety of the data catalog, though the exact fraction will vary significantly depending on the analysis goals. From Run2 experience, a typical analysis requires between 5–30% of a given data release version's data. As long as the use of a lake-edge query service such as ServiceX is structured such that it uses MiniAOD as an add-on to NanoAOD rather than as a replacement of NanoAOD, the IO requirements of ServiceX will probably be within 100 Gbps steady state and 400 Gbps burst capacity during the first three-year running period of HL-LHC. The reasoning is as follows: 10% of 27PB for an annual MINI version is 3PB. Reading this 10% in a week is a more modest 40 Gbps. If MiniAOD information was added only to a subset of filtered events, or only a few floats per event needed to be fetched via the WAN, then the bandwidth requirements could be very comfortably within the 400 Gbps burst capacity.

6.12.2.5.2 Inference As a Service

A novel network requirement CMS expects in the HL-LHC era is sourced from planned use of a disaggregated CPU-GPU architecture that allows to dynamically adjust the number of GPUs that facilities will require as the reconstruction software evolves to incorporate more algorithms amenable to GPU acceleration. These algorithms, presently in development, fall into two classes: ML algorithms for reconstruction trained on simulated data; or classical algorithms rewritten within performance portability frameworks such as Sycl or Alpaka. The SONIC

project has introduced a module that extends CMS reconstruction software to asynchronously send subsets of input data to a remote server with GPUs running the Nvidia Triton service for processing. It is an open question whether these coprocessors will only be located within the LAN or if CPU workflows running on a site can offload portions of the tasks to GPUs at another site over the WAN. In the latter case, it must be understood if the latency and bandwidth permits the CPU workflows to continue with good efficiency while waiting for the inference result to be returned.

6.12.2.6 Remote Science Activities

The HL-LHC computing infrastructure is inherently distributed. In other sections ATLAS has described the LHC computing model, with a central site at CERN (the Tier-0) and national scale centers around the globe (Tier-1s) as well as many regional centers (Tier-2s). This basic model is being evolved to meet the needs of HL-LHC, and it will certainly augment this infrastructure with additional resources that might be used, like HPC sites, commercial and research clouds and additional institutional resources. In practice you can consider all users as "remote" and almost all the resources that they use to pursue their science goals as "remote."

6.12.2.7 Software Infrastructure

6.12.2.7.1 ATLAS

The ATLAS Distributed Computing (ADC) system and organization today is a sophisticated ensemble of software systems, computing facilities and people that make it possible for ATLAS processing to run 24x7x365 on about 750k cores (in 2024) across over 100 worldwide sites and a wide range of resource types including grid clusters, HPCs, LCFs and clouds, processing in total about 2.8 exabytes (in 2024) of data per year.

The new HL-LHC directed capabilities and workflows described in previous sections are being developed on the proven and scalable foundation of this system, particularly the PanDA workload management system, Prodsys production management system and the Rucio distributed DMS (which were described in Section 6.9).

Beginning with Run 1 in the case of PanDA and Run 2 with ProdSys and Rucio, these systems have met the capability and scaling requirements of ATLAS, and have steadily been enhanced and hardened based on running experience. ATLAS has confidence that they have the scalability and robustness to continue to serve ATLAS through Run 4. Hence ATLAS has been able to focus on adding new HL-LHC directed capability to these systems.

PanDA, ProdSys and Rucio are being extended for the demands of HL-LHC in such projects as the data carousel and the intelligent data delivery service (iDDS). Rucio has been adopted as the LHC and WLCG standard. The iDDS service is in production use ATLAS and was developed to support fine-grained workflows like the ATLAS data carousel. ServiceX is being developed in a joint effort between ATLAS and IRIS-HEP to support fine-grained workflows with intelligent data transformations and filters to dynamically deliver just the needed analysis data across the network, where and when it is needed. While iDDS and ServiceX both work in close concert with ADC services like PanDA and Rucio, they are implemented as experiment-agnostic services.

6.12.2.7.2 CMS

At present, it is expected that a large portion of the software infrastructure described in Case Study 10 and 11 will continue to be used in the HL-LHC era. Notably, CMS and ATLAS now both use Rucio data management software, and have for some time both used the CERN FTS. A notable difference between the two collaborations is that CMS more heavily relies on streaming data from storage endpoints to remote processing centers via the XRootD protocol. CMS is currently re-evaluating its workflow management infrastructure solution, which may lead to changes in network usage patterns if a new system is adopted.

6.12.2.8 Network and Data Architecture

For the HL-LHC era, the predictions show a mismatch between the computing and storage resources the experiments can afford versus the resources needed to reach science goals. In response to this gap, the experiments are exploring alternatives in how to utilize storage, computing and network infrastructure. The network baseline currently being planned is composed of terabit-scale (1–2 Tbps) backbone networks with the largest resource sites connected at multiple 400G scale (800 Gbps+). Network use will be at least a factor of 10 larger than Run 3.

Global R&E network use will have different characteristics on the timescale of the HL-LHC, compared to the situation the WLCG experienced in Runs 1, 2, and 3. Specifically, the experiments foresee other science domains (astronomy, biology, engineering, and others) becoming global network users at scales equaling or exceeding the LHC users. In such an environment, it is not wise to assume the experiments can continue to treat the network as an infinite resource . The LHC community must explore options for effectively operating within a bandwidth constrained environment.

For HL-LHC four main requirements have been identified:

- Capacity: Run 3 is moving to multiple 100G links for large sites, while Run 4 (HL-LHC) is targeting Tbps links, perhaps based upon either 400G or 800G waves.
    - Capacity is fundamental for science at HL-LHC scales.
    - As noted, in a capacity constrained environment, it will be important to manage the capacity to do the most science possible.
- Capability: It is necessary to understand the impact of new features in networking (SDN/NFV) by testing, prototyping and evaluating impact. The experiments will need to evolve applications, facilities and computing models to meet the HL-LHC challenges; it will take time.
    - Traffic shaping activities underway in the Research Network Technical WG are a good capability example.
    - Network orchestration between sites is another good example and something that could help us more effectively exploit available resources.
- Visibility: As the ESnet Blueprinting meetings have shown, the ability to understand WAN network flows is limited. New methods to mark and monitor network use are needed.
    - Packet marking is viewed as a high priority so that everyone (experiments, sites and R&E networks) can understand the origin and intent of network flows at any point along their path.
    - The LHC community needs better mechanisms to coordinate the available monitoring resources from the sites, experiments and networks, to allow us to have a better understanding of how complex infrastructures are using the network.
- Testing: Developing, prototyping, and testing network features at suitable scale will be needed.
    - Networks of the future will likely have new features, capabilities and services that could be leveraged to do more with the resources that are available.
    - The LHC community challenge is identifying which features might be beneficial to try to integrate into operations, noting that such integration can require significant effort to realize.
    - Having at-scale network testbeds will be very important to understand the potential impact changes might have on operations.

These requirements should also motivate and guide network infrastructure upgrades and replacements and it will be critical to understand the bigger picture of R&E networking and its evolution, to ensure the experiments are able to take advantage of capacity and services available.

6.12.2.9 Emerging Technology Use

6.12.2.9.1 Cloud Services

As demonstrated in the ATLAS-Google project in 2022-23 and also in a smaller project with Amazon, PanDA+Rucio can use commercial cloud resources interchangeably with grid-based WLCG resources and also integrate new setups for interactive and parallel data analysis. Bulk MC production with bursts of up to 100k concurrently running jobs in Google Cloud have been demonstrated. The baseline plan for HL-LHC is to use grid and HPC resources. If some grid resources are set up as cloud resources, they will also be used. ATLAS will require a good network pipe between grid sites and commercial clouds when used. In this model, the network needs will be similar to university-based US Tier-2 sites.

CMS expects to have the ability to make use of Cloud resources, but at present, budgets to do so are not present, given the current business models of cloud providers. It is unclear whether or not such budgets will exist in the future. CMS has in the past evaluated the use of commercial cloud both for processing, and for TA transfer. US CMS finds that both are not cost-effective, at present. Experimentation has shown that making large-scale use of cloud resources if the cost structure were to change is possible. A key component of the cloud cost structure that makes this approach unattractive is the egress bandwidth charges. If ESnet were to be able to negotiate a favorable peering arrangement with cloud providers, the cost competitiveness of cloud resources would increase substantially.

6.12.2.9.2 AI and ML

From the ATLAS Run3 software and computing paper: ML is used in a wide range of applications in ATLAS, including physics analysis, simulation and physics object identification within the reconstruction and trigger system. Models trained for classification and regression, as well as generative models, are core components of the software and analysis chain.

During the HL-LHC era, it is expected that AI/ML tools and techniques will be used more frequently than is done during Run 3. The need for GPU resources will grow beyond the current levels provided by private local institutional resources . ATLAS is using GPU resources at NERSC HPC using a small but growing GPU allocation for AI/ML training.

As discussed in section 4, CMS plans to offload a portion of the event reconstruction algorithms to GPUs, both via ML and classical algorithms. The use of ML within CMS for reconstruction and data analysis is very well advanced, and underpins many of the major measurements and searches for new physics the collaboration has performed over the past years. A full review of these uses is outside the scope of this document.

Usage of ML within CMS workflow and data management, in contrast, is less well developed. R&D initiatives are underway to rectify this situation, with goals of reduced resource usage and personnel requirements through automation. ML anomaly detection approaches are actively being investigated for DQM in CMS as well, and some lessons learned from these activities may also be useful for monitoring computing infrastructure.

6.12.2.9.3 Wireless or Distributed Location Science Activities

No wireless capability studies or capability upgrades are planned at present since wireless use cases are only users accessing remote services that do the actual data operations.

6.12.2.10 Data-Related Resource Constraints

One interesting option to explore on the timescale of HL-LHC is the joint ATLAS and CMS use of Rucio for distributed data management. This could provide a mechanism to interact with ESnet (and other R&E networks), communicating near-term data-movement intents and perhaps negotiating for any required QoS or deadline requirements.

There have been significant efforts made to understand the computing and storage requirements for HL-LHC. The process began by extrapolating current methods forward, accounting for the data increases from HL-LHC and the additional complexity associated with high-luminosity operations of the collider. Further assumptions about the initially conservative estimates of computing and storage capacities would be increased for fixed cost and additionally assumed the notion of a "flat budget" for the foreseeable future.

To decrease the gap, significant efforts were made to better optimize how less resources could be utilized to do the same amount of science. This has helped decrease the difference between what is needed and what is affordable. While there are potentially some optimistic assumptions in this estimate, it is known that the planning process must change to ensure resource availability on the time scale of HL-LHC. At this point there is still a significant gap in the amount of storage believed to be needed versus what can be provisioned with current budget projections.

Another concern is the potential constraints in R&E network capacity on the timescale of the HL-LHC. One of the options that has been considered in addressing the gap in resources noted above, is taking more advantage of high-performance networks to help reduce storage requirements. However, this implicitly assumes the network continues to be unconstrained and free of cost to the HL-LHC collaborations. As noted elsewhere, network providers serving global high energy physics collaborations have been very successful in providing network capacities beyond the requirements of existing use cases, giving collaborations whatever bandwidths they have required. This is likely not going to be feasible for the HL-LHC era, not because of the significant growth in HL-LHC data, but, rather due to the expected rise in use from other global, data-intensive science domains.

The LHC collaborations have not faced such an environment to date and, if such a situation were to arise, the tools or methodologies to work effectively are not present. Capabilities like SENSE/Rucio, NOTED, and others, will help to provide potentially critical capabilities for a potential oversubscribed network environment.

ESnet and the LHC experiments (ATLAS and CMS) have started an R&D project that makes use of in-network data caches operated by ESnet using XRootD software. The pilot project focuses on caching of PHYSLITE data products to see if the time to reach physics results can be shorted, and the overall network traffic reduced for data analysis (only used parts of data files cached and reused).

6.12.2.11 Data Mobility Endpoints

On the timescale of HL-LHC, it is expected that the LHC experiments will evolve their computing models, hardware, software and architecture, guided by mini-challenges and the two remaining WLCG Network Data Challenges. So, even though "Data Mobility Endpoints" (e.g., Globus nodes or DTNs) are not currently a central component of the experiment's architecture, they could very well become a core component of WLCG. US ATLAS has already developed a proof of concept to use Rucio to manage transfers via Globus. US ATLAS, Rucio Core team and HEP-CCE (w/ HEP-CCE contributing a python code developer) are all collaborating testing such a setup at scale. This will be driven by architecture evolutions coupled with testing and at-scale validation. One side effect is that it may make WLCG infrastructure more compatible with other global science infrastructures.

6.12.2.12 Outstanding Issues

The individual LHC experiments are interested in discussing the set of features that must be present in future network designs. If it is possible to identify a minimum set of capabilities in switches and routers, to ensure the next generation of network components can meet the needs of the community, it would simplify the procurement and installation process. The sooner those capabilities are in place, the sooner prototyping and testing future capabilities will be possible.

Another important consideration is how to identify appropriate hardware and software configurations that can be used to optimize the ability to use the procured resources. In many cases, the information about such tunings can easily apply to many science domains and sharing the work to test and benchmark across different components would benefit all the LHC experiments and beyond.

6.12.2.13 Facility Profile Contributors

ATLAS Representation

Doug Benjamin, BNL, dbenjamin@bnl.gov

Johannes Elmsheuser, BNL, jelmsheuser@bnl.gov

Shawn Mckee, UM, smckee@umich.edu

CMS Representation

Nick Smith, Fermilab, ncsmith@fnal.gov

Syed Asif Shah, Fermilab, sshah@fnal.gov

David Lange, Princeton University, david.lange@princeton.edu

Andrew Melo, Vanderbilt University, andrew.m.melo@vanderbilt.edu

ESCC Representation

Vincent Bonafede, BNL, bonafede@bnl.gov

Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

Phil DeMar, Fermilab, demar@fnal.gov

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

6.13 Super Cryogenic Dark Matter Search (SuperCDMS SNOLAB)

Numerous astrophysical and cosmological observations indicate that nearly 85% of the mass content of the universe does not emit electromagnetic radiation at any wavelength, therefore being called dark matter. The composition of dark matter is unknown, and its properties imply that it cannot be made of known elementary particles. For this reason, dark matter is one of the strongest evidences of new fundamental physics. To date, experiments have set stringent constraints on the existence of dark matter particles with masses in the GeV regime, but the sub-GeV regime, that is also well-motivated from the theoretical side, has been much less explored.

SuperCDMS SNOLAB is a next-generation experiment aiming to detect sub-GeV dark matter particles. In the spring of 2015, SuperCDMS SNOLAB was selected by the DOE and the NSF as one of the three Generation 2 dark matter experiments. SuperCDMS SNOLAB is an experiment funded and operated under the DOE OS HEP program in the Cosmic Frontier, and co-funded by NSF. Non-US funding is provided through the CFI, and other smaller funding resources.

6.13.1 Discussion Summary

The LAN connectivity for SNOLAB is 2 x 10 Gbps, and the WAN connectivity consists of a 1 Gbps link that is also shared with 10 other experiments. This limitation implies that SuperCDMS must plan to work within a budget of no more than 10% of either of these connections during operation. This link allows SNOLAB to reach Laurentian University, which is then connected to ORION (the provincial R&E network of Ontario), and CANARIE (the national R&E network of Canada). CANARIE peers with ESnet, which facilitates the connectivity to reach SLAC. The network limitations were explored during a project review in 2022, at the time the cost was considered too high to upgrade the 1 Gbps WAN link, which will cause the experiments that share this resource to be more austere in their use. SuperCDMS will consider this in the design of their trigger mechanism, and computing and storage strategy. ESnet can discuss options with SuperCDMS regarding the use of other connectivity options.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- SuperCDMS SNOLAB is a next-generation experiment aiming to detect sub-GeV dark matter particles. SuperCDMS SNOLAB is an experiment funded and operated under the DOE OS HEP program in the Cosmic Frontier, and co-funded by NSF. Non-US funding is provided through the CFI, and other smaller funding resources. [Section 4.1]
- The SuperCDMS SNOLAB experiment is currently in the installation phase at SNOLAB (Canada), 2 km underground. Commissioning is expected to start in fall 2025, and science operations are expected to start at the beginning of 2026 and to last for at least three years, leaving open the possibility of upgrades (more detectors and/or new technology in order to leverage the cryogenic facility). [Section 4.1]
- The SuperCDMS collaboration is spread between many US, Canada, Germany, India, Spain, United Arab Emirates, and United Kingdom institutions. The approximate count is about 130 collaborators in 28 institutions. [Section 4.1]
- The estimated data volume of SuperCDMS SNOLAB is 350 TB/year, including calibration data and simulated data. [Section 4.2]
- SuperCDMS SNOLAB's raw data, in compressed MIDAS format, will be processed at SLAC, and the reconstructed output data, in ROOT format, can be directly used for data analyses. Collaborators will access and analyze processed data stored at SLAC using local computing resources. The data will also be transferred from SLAC to Fermilab, where it will be stored on tape for additional backup. Simulated data are generated at Texas A&M, transferred to SLAC, and then transferred to Fermilab for tape backup and to other sites as needed. [Section 4.3]

- SuperCDMS SNOLAB will not rely on SNOLAB for data analysis. Data analysis will happen primarily at SLAC, DRAC, the OSN, NSF ACCESS facilities, and at other smaller institutions. [Section 4.3]
- SNOLAB is minimally equipped with software, and each experiment has a dedicated subnetwork with limited bandwidth. Given the small amount of data to be moved from SNOLAB to SLAC, automated transfer tools based on rsync (single stream) with the option to move to bbcp (multiple streams, 4 by default), will be used. The limited outgoing bandwidth and the single destination make the use of such simple approaches ideal for this scenario. [Section 4.3]
- SNOLAB will have the following mobility patterns and tools for their data management: [Section 4.3]
    - Transfers between SLAC and Texas A&M happen via Globus.
    - Transfers between SLAC and OSN happen via S3 (other technologies can be available).
    - Transfers between SLAC and DRAC happen via rsync, Globus or https (small datasets ~10–100TB).
    - Transfers between SLAC and Fermilab have not been tested yet but we have a large choice of tools, from rsync to Globus.
- SuperCDMS SNOLAB data will be acquired underground at SNOLAB, transferred to the surface, and then transferred to the SLAC National Accelerator Laboratory where it will be processed. In order to ensure access to data to all members of the SuperCDMS collaboration (access to SLAC resources is restricted for some collaborators), processed data will be also copied on the OSN and the DRAC. [Section 4.3]
- The main network connections are going to be between: [Section 4.7]
    - SNOLAB underground laboratory to SNOLAB surface (raw data)
    - SNOLAB surface to SLAC, through the ORION network
    - SLAC to Fermilab (tape backup site for all data, raw, processed, simulated)
    - SLAC to OSN and DRAC (secondary sites for analysis, transfer of main processed datasets and other data on demand)
    - Texas A&M to SLAC (simulated data)
- The network link from underground to the surface is made of 2 x 10 Gbps aggregate links shared by about 10 experiments. Only ~0.2% is expected to be needed by SuperCDMS during peak usage. From the surface, the data are copied to SLAC through the SNOLAB network link to the ORION of 1 Gbps. From there, ESnet routes become available. [Section 4.7]
- The network link at SNOLAB from underground to surface (2 x 10 Gbps) is sufficient for the needs of the facility but it is subject to instabilities and interruptions, while the outgoing SNOLAB network is limited to a 1 Gbps link. While there are no plans yet, SNOLAB is looking into upgrading the outgoing bandwidth to 10 Gbps. [Section 4.7]
- SuperCDMS SNOLAB will make minimal or no use of cloud storage. Only a few services will be cloud-based, these include Slack and GitLab. [Section 4.8]
- SLAC is building support for the execution of AI/ML workflows with substantial investment in GPU. SuperCDMS SNOLAB reconstruction working group (WG) is looking into taking advantage of the GPUs for event reconstruction. TensorFlow is also used in some analyses. One use case is for datasets severely limited by noise coming from vibrations. With ML we can

remove the noise effectively while saving the low energy signal events, which are important for low-mass data management (DM) search. The same method can be used to predict the energy deposited in the crystals, which works well in low energies where other techniques do not perform well. [Section 4.8]

- SuperCDMS SNOLAB is not equipped, so far, to work in an IRI workflow. Its data structure, data taking, and data-processing patterns are simple enough that the need for advanced solutions has not presented itself, and would instead require additional personnel and resources that are not accessible with the current funding level. [Section 4.9]
- Limitations are imposed by the various collaborator sites depending on firewall rules, use of VPNs, and MFA requirements that often limit the amount of automation of the transfer tools. The lack of uniformity in the availability of tools (e.g., not all sites have Globus endpoints) and the different network protection stances also imply the need to develop data transfer tools tailored for each site. For a computing group of a small experiment like SuperCDMS, the lack of standardization creates additional overhead. [Section 5.3]
- In case of extensive, site-wide outages at SLAC, while we have on-site storage for a few weeks of data taking, it would also be important to have another data path from SNOLAB to an alternative computation site, namely DRAC and/or Texas A&M, in order to not interrupt the processing flow. The criteria needed to temporarily switch from SLAC to another site for storage and processing needs to be carefully evaluated and the process tested. [Section 5.6]
- The network connectivity between SNOLAB and SLAC is a critical component, given that all workflows and all scientific analysis happen off-site. And in particular, the bandwidth bottleneck between SNOLAB and the ORION network could present a challenge in the future if there are no upgrades to the bandwidth, given the tendency of data volumes from current and future experiments to grow rather than shrink. [Section 5.7]
- SNOLAB and SLAC could explore alternative network paths, e.g., those that do not rely on a terrestrial link, as an alternative path to address the critical need to keep networking active during experimentation. [Section 5.7]

6.13.2 SuperCDMS: Experiment Profile

SuperCDMS SNOLAB consists of an array of monocrystalline germanium and silicon detectors operating at cryogenic temperatures (~30 mK). Particle interactions occurring in SuperCDMS SNOLAB detectors produce recoiling nuclei or electrons, that in turn create prompt phonons (quanta of vibrations of the crystalline structure) and ionization charges. An external electric field is applied over the detector volume to drift the ionization charges that produce additional phonons through the Neganov-Trofimov-Luke (NTL) effect. The detectors are classified in two categories depending on their instrumentation: 1) iZIP detectors, that measure both phonon and ionization charge signals, enabling enhanced background rejection; and 2) high-voltage detectors, that measure the NTL phonons from charges drifting in an increased electric field, enabling lower energy thresholds.[118, 119]

6.13.2.1 Science Background

The SuperCDMS SNOLAB experiment is currently in the installation phase at SNOLAB (Canada), 2 km underground. Commissioning is expected to start in fall 2025, and science operations are expected to start at the beginning of 2026 and to last for at least three years, leaving open the possibility of upgrades (more detectors and/or new technology in order to leverage the cryogenic facility).

118 B. Neganov and V. Trofimov, Calorimetric method measuring ionizing radiation. Otkrytia i Izobret. 146, 215 1985.
119 P. N. Luke et al., Calorimetric ionization detector. Nuclear Instruments and Methods in Physics Research Section A 289, 406–409 1990

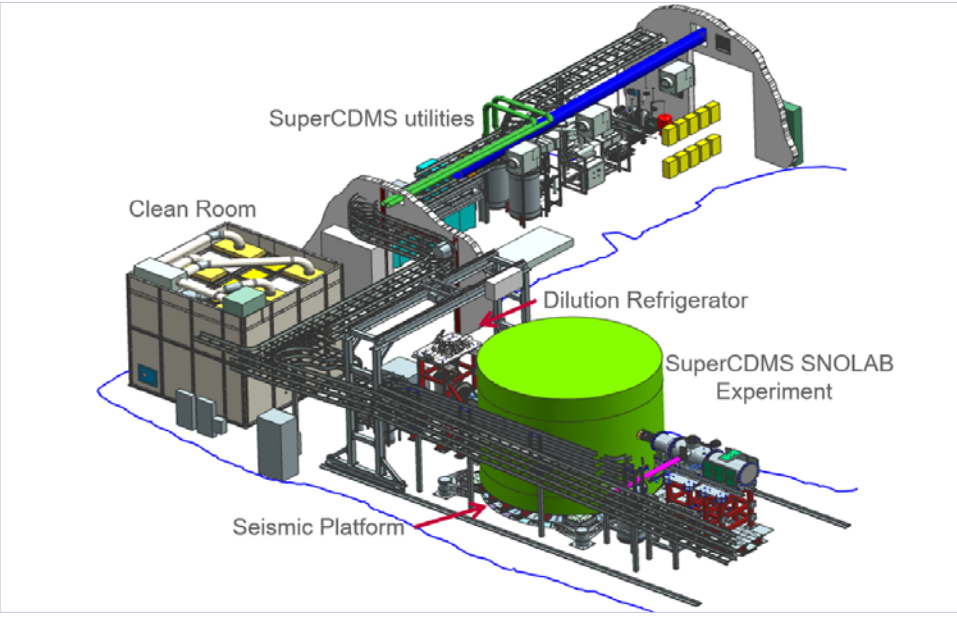


Figure 6.13.1: SuperCDMS SNOLAB experimental setup

SuperCDMS SNOLAB data will be acquired underground at SNOLAB, transferred to the surface, and then transferred to the SLAC National Accelerator Laboratory where it will be processed. In order to ensure access to data to all members of the SuperCDMS collaboration (access to SLAC resources is restricted for some collaborators), processed data will be also copied on the OSN and the DRAC.

The Collaboration will retain ownership of the data, except for small datasets released in support of publications.

6.13.2.2 Collaborators

The SuperCDMS collaboration is spread between many US, Canada, Germany, India, Spain, United Arab Emirates, and United Kingdom institutions. The approximate count is about 130 collaborators in 28 institutions. The full collection of SuperCDMS SNOLAB data (raw and processed data, as well as simulated data) is going to be hosted at SLAC (on disk and tape) and at Fermilab (on tape for recovery purposes only), while processed data sets will be also accessible at OSN and DRAC (due to the limited disk resources available). At SLAC, the SuperCDMS collaboration has ~100 active users from all SuperCDMS institutions. Most collaborators have accounts both at SLAC and on the Canadian DRAC.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
SNOLAB, UNDERGROUND	Temporary storage	Network transfer	200 TB/year	Daily	N	Unstable network
SNOLAB, SURFACE	Temporary storage	Network transfer	200 TB/year	Daily	N	Poor bandwidth
SLAC	Primary	Network transfer, data portal, disk, tape	350 TB/year	Daily	N	Tools and automation of data distribution
OSN	Secondary		150 TB/year	Ad hoc	N	Automation
DRAC	Secondary		150 TB/year	Ad hoc	N	Automation
TEXAS A&M	Temporary storage	Network transfer	60 TB/year		N	No
FERMILAB	Primary on tape only	Network transfer	~1–1.5 PB/ experiment lifetime	~Weekly	N	No

Table 6.13.1: SuperCDMS Data Mobility

These are estimates based on expected trigger rates and simulation needs and they may change with further studies closer to the start of the data taking.

6.13.2.3 IRI Readiness

SuperCDMS SNOLAB is not equipped, so far, to work in an IRI workflow. Its data structure, data taking, and data-processing patterns are simple enough that the need for advanced solutions has not presented itself, and would instead require additional personnel and resources that are not accessible with the current funding level.

6.13.2.4 Instruments and Facilities

The SuperCDMS SNOLAB DAQ and Trigger system is based on the MIDAS data acquisition (DAQ) framework, with off-the-shelf computers and custom trigger electronics. MIDAS is a flexible C++-based DAQ system, developed at Paul Scherrer Institute and TRIUMF. A custom software package manages all the underlying custom MIDAS front-ends and clients that control SuperCDMS DAQ hardware, data flow, and triggering system (which consists of two subsystems, the L1 firmware trigger and the L2 software trigger). These include the Online Data Base layout, device drivers, and low-level tools for data acquisition.

The DAQ and Trigger system is deployed on a cluster underground in a dedicated subnetwork. The gateway computer is connected to another dedicated SuperCDMS subnetwork on the surface to allow for data transfers from underground.

The lifetime goal is to be taking usable dark matter search data 80% of the time. To meet this high rate of experimental lifetime, the SNOLAB DQM system is designed to (1) promptly alert shifters as problems arise and (2) provide efficient access to data needed during debugging. In addition, data-quality information is used in the offline analysis to determine which data are reprocessed for science results and to inform analyzers of special conditions during particular runs. The DQM system is located on the underground cluster and it is mirrored on the surface and at SLAC.

The DAQ and DQM systems have redundant disk space both underground and on the surface. The surface server is used as data buffer, data transfer, underground database mirroring, and DQM.

Both the surface and underground computers have enough disk space to host about 3 weeks of data in case network problems occur. Additional surface resources include experimental control workstations for on-site shifters, a disk and CPU allocation on the shared SNOLAB Nearline Computing System, and lightweight virtual machines provided by SNOLAB IT.

On the surface, the data are automatically transferred to SLAC.

The raw data, in compressed MIDAS format, will be processed at SLAC, and the reconstructed output data, in ROOT format, can be directly used for data analyses. Collaborators will access and analyze processed data stored at SLAC using local computing resources. The data will also be transferred from SLAC to Fermilab, where it will be stored on tape for additional backup. Simulated data are generated at Texas A&M, transferred to SLAC, and then transferred to Fermilab for tape backup and to other sites as needed.

The estimated data volume is 350 TB/year, including calibration data and simulated data.

SLAC is the leading lab for the computing of the experiment, and it will host the full set of data (real and simulated), supports data processing, data analysis, user accounts, and web applications, including job management and bookkeeping. SLAC also hosts the public website, and collaborative tools. All SuperCDMS applications at SLAC are containerized and deployed through Kubernetes.

Software releases are deployed as Apptainer containers and are used at all participating SuperCDMS institutions for data analysis, software development, processing, and data simulation.

The major packages for simulation, reconstruction and calibration are written in C++ OO. Simulations are based on the Geant4 toolkit, and in particular SuperCDMS developed the open source Geant4 package G4CMP[120] to simulate the phonon and charge transport in cryogenic semiconductor crystals. SuperCDMS also has extensive software written in python to support workflows and analysis.

The full chain, from data acquisition, data transfers, data processing and analysis has been tested through data challenges and a test run with a quarter of the final detector payload at the Cryogenic Underground TEst Facility at SNOLAB (CUTE) facility[121]. All the stages, from data taking to making the processed data accessible to the analyzers are automated and functional.

6.13.2.5 Generalized Process of Science

SuperCDMS SNOLAB will not rely on SNOLAB for data analysis. Data analysis will happen primarily at SLAC, DRAC, the OSN, NSF ACCESS facilities, and at other smaller institutions.

Data are going to be processed at SLAC to extract quantities that can be directly used in physics analyses. During data processing, reconstruction software will extract various quantities from the raw pulses such as the pulse amplitude, area and start time, that provide information of the deposited energy and the interaction position within the detectors. These quantities from the processing are stored in Reduced Quantities files. After the data processing, the analyzers are going to apply a set of cuts to select only the relevant data for dark matter searches. A complete background model (including radiogenic, cosmic, and cavern sources) is being developed and it will be generated using DRAC resources while detector response simulation production based on Geant4 will happen at Texas A&M. After that, SuperCDMS SNOLAB analyzers will compare the processed data with the simulated background model, with a profile likelihood ratio statistic test.

120 Nuclear Instruments and Methods A 1055 (2023) 168473 (https://arxiv.org/abs/2302.05998)
121 "CUTE: A Cryogenic Underground TEst Facility at SNOLAB," https://arxiv.org/abs/2310.07930 , 2023

6.13.2.6 Remote Science Activities

The main network connections are going to be between:

- SNOLAB underground laboratory to SNOLAB surface (raw data)
- SNOLAB surface to SLAC, through the ORION network
- SLAC to Fermilab (tape backup site for all data, raw, processed, simulated)
- SLAC to OSN and DRAC (secondary sites for analysis, transfer of main processed datasets and other data on demand)
- Texas A&M to SLAC (simulated data)

In addition to that, SuperCDMS SNOLAB has some database replication between SNOLAB and SLAC that will require network usage. While the amount of replication data should be considered negligible, it may not actually be so for the limited bandwidth at SNOLAB toward the rest of the world.

6.13.2.7 Software Infrastructure

Data transfers between SuperCDMS sites happen with a variety of tools, depending on the availability of such tools at the source or destination site.

SNOLAB is minimally equipped with software (no Globus, XRootD, …) and each experiment has a dedicated subnetwork with limited bandwidth. Given the small amount of data that SuperCDMS will move from SNOLAB to SLAC, automated transfer tools based on rsync (single stream) with the option to move to bbcp (multiple streams, 4 by default) are currently in use, if needed. The limited outgoing bandwidth and the single destination make the use of such simple approaches ideal for this scenario.

SLAC is fully equipped with every possible transfer tool and SuperCDMS uses a variety of those. Transfers between SLAC and Texas A&M happen via Globus.

Transfers between SLAC and OSN happen via S3 (other technologies can be available).

Transfers between SLAC and DRAC happen via rsync, Globus or https (small datasets ~10–100TB).

Transfers between SLAC and Fermilab have not been tested yet but SuperCDMS has a large choice of tools, from rsync to Globus.

6.13.2.8 Network and Data Architecture

The network link from underground to the surface is made of 2 x 10 Gbps aggregate links shared by about 10 experiments. Only ~0.2% is expected to be needed by SuperCDMS during peak usage. From the surface, the data are copied to SLAC through the SNOLAB network link to the ORION of 1 Gbps. From there, ESnet routes become available and, given the amount of data moved, SuperCDMS does not generally experience any problems.

Once the data reaches SLAC, all subsequent main data transfers happen between SLAC and other laboratories with very large network capabilities (Fermilab, Texas A&M, OSN, DRAC).

6.13.2.9 Emerging Technology Use

6.13.2.9.1 Cloud Services

SuperCDMS SNOLAB will make minimal or no use of cloud storage. Only a few services will be cloud-based, these include Slack and GitLab.

6.13.2.9.2 AI and ML

SLAC is building support for the execution of AI/ML workflows with substantial investment in GPU. SuperCDMS SNOLAB reconstruction working group (WG) is looking into taking advantage of the GPUs for event reconstruction. TensorFlow is also used in some analyses. One use case is for datasets severely limited by noise coming from vibrations. With ML we can remove the noise effectively while saving the low energy signal events, which are important for low-mass data management (DM) search. The same method can be used to predict the energy deposited in the crystals, which works well in low energies where other techniques do not perform well.

6.13.2.9.3 Wireless or Distributed Location Science Activities

SuperCDMS SNOLAB does not require any connectivity to be provided by wireless or satellite connectivity, but could investigate this as an optional back-up strategy.

6.13.2.10 Data-Related Resource Constraints

The network link at SNOLAB from underground to surface (2 x 10 Gbps) is sufficient for the needs of the facility but it is subject to instabilities and interruptions, while the outgoing SNOLAB network is limited to a 1 Gbps link.

While there are no plans yet, SNOLAB is looking into upgrading the outgoing bandwidth to 10 Gbps.

Once the data reaches the US network the data distribution and accessibility problem becomes the dominant one.

Limitations are imposed by the various sites depending on firewall rules, use of VPNs, and MFA requirements that often limit the amount of automation of the transfer tools. The lack of uniformity in the availability of tools (e.g., not all sites have Globus endpoints) and the different network protection stances also imply the need to develop data transfer tools tailored for each site. For a computing group of a small experiment like SuperCDMS, the lack of standardization creates additional overhead.

6.13.2.11 Data Mobility Endpoints

SLAC endpoint that SuperCDMS currently uses:

- SLAC S3DF endpoint display name: slac#s3df_globus5

Texas A&M (TAMU) endpoint that SuperCDMS currently uses:

- TAMU Grace display name: TAMU grace-dtn

DRAC endpoint that SuperCDMS currently uses:

- Fir Globus endpoint display name: alliancecan#fir-globus

6.13.2.12 Outstanding Issues

The network connectivity between SNOLAB and SLAC is a critical component, given that all workflows and all scientific analysis happen off-site. And in particular, the bandwidth bottleneck between SNOLAB and the ORION network could present a challenge in the future if there are no upgrades to the bandwidth, given the tendency of data volumes from current and future experiments to grow rather than shrink.

In case of extensive, site-wide outages at SLAC, while SuperCDMS has on-site storage for a few weeks of data taking, it would also be important to have another data path from SNOLAB to an alternative computation site, namely DRAC and/or Texas A&M, in order to not interrupt the processing flow. The criteria needed to temporarily switch from SLAC to another site for storage and processing needs to be carefully evaluated and the process tested.

Finally, SuperCDMS need to also ensure the flow of processed data to secondary storage sites (OSN and DRAC).

6.13.2.13 Facility Profile Contributors

SuperCDMS Representation

Concetta Cartaro, SLAC, cartaro@slac.stanford.edu

Elias Lopez Asamar, U.A. Madrid, elias.lopez@uam.es

David Toback, Texas A&M University, toback@tamu.edu

Stefan Zatschler, SNOLAB, stefan.zatschler@snolab.ca

SLAC Representation

Mark Foster, SLAC, fosterm@slac.stanford.edu

6.14 Alpha Magnetic Spectrometer (AMS) Experiment

AMS is a precision particle physics detector installed on the ISS in May 2011, conducting a unique, long-duration mission of fundamental physics research in space. Figure 6.14.1 shows AMS on the ISS.

6.14.1 Discussion Summary

Due to limitations in the available bandwidth that TDRS can deliver, and the number of use cases that exist on the ISS, AMS is only able to access about 14Mbps of space-to-ground transmission. This limitation has not forced the adoption of high-speed data migration tools, as the data volumes will always remain small for this aspect of the workflow. NASA has been evaluating if the use of other satellite technologies (e.g., Starlink) could serve as an augmented communication mechanism, or a replacement for TDRS, but this would not be something AMS has any control over.

The AMS project creates a large number of simulations during their process of science, over 20PB have been produced since the start of the project. With the pending upgrade to AMS hardware in the coming years, it is anticipated that all produced MC will have to be reprocessed.

AMS occupies a unique space in the HEP portfolio, and is capturing a valuable set of research data that is not replicable on earth at the current time.

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- AMS is a precision particle physics detector installed on the ISS in May 2011, conducting a unique, long-duration mission of fundamental physics research in space. It is the largest instrument, and is scheduled to run until the ISS project ends, currently expected to be after 2030. Since 2011, AMS has collected more than 200 billion events. [Section 4.1]
- Construction, integration, testing and operation of the AMS is carried out by an international team, referred to as the AMS Collaboration, composed of about 600 physicists and 44 institutes and organizations from 14 countries. [Section 4.1]
- AMS produces ~40 TB of raw data per year, which are transmitted to CERN, and then reconstructed to 170 TB data per year available for analysis by collaboration members. The production is performed incrementally every six months. Reprocessing is only performed when there are changes to reconstruction software. [Section 4.2]
- Each version of full reproduction is called "pass." Currently, AMS acquired over 500 TB of raw data, from which a 2PB of pass 8 data were produced and copied to collaborators institutions for further analyses. [Section 4.2]
- AMS data analysis requires massive MC simulation, performed by computing centers on a regular time basis. Currently 23 PB of MC data were produced since the start of the project in 2011. [Section 4.2]
- AMS has more than 300K readout channels, and 1500 temperature sensors. The acquisition rate goes up to 2 kHz. There are more than 500 microprocessors onboard which reduce data rate from 7 Gbps to 10 Mb/s. Data taken on board of the ISS by AMS consist of raw events, and these are logically grouped in sequences called runs. Events inside one run are numbered in sequence. Every quarter of the orbit a scientific run starts. The duration of each run corresponds to 23 minutes, and every run contains on average 700k events. The average size of raw event is 2 KB, thus yearly raw data yield is 40 TB. [Section 4.2]
- The typical AMS raw file size is 1.5 GB, while the AMS data file in pass 8 production is 7 GB. At the moment, there are 250K pass 8 files were produced. [Section 4.2]

- By the end of 2025, AMS will undergo a significant upgrade on the ISS; it will involve the addition of a new tracker layer at the top of the instrument. These advancements are essential for progressing dark matter searches and improving precision measurements. The new configuration of AMS will require a complete overhaul of simulations and data analysis tools, which will be developed over the next three to five years. The total volume of data is expected to increase to 50 PB over the next five years. [Section 4.2]
- Currently, a new data analysis and computing center is being developed at the University of Hawaii at Manoa in the USA. This will be the first computing and analysis center on US soil dedicated exclusively to AMS data. About 20000 CPU cores and 10 PB of data storage will be employed for AMS data production, analysis, MC simulations, and archiving. [Section 4.3]
- Recorded data are packed to fixed length packets and transmitted via the TDRS satellites to ground at White Sands, NM, and then relayed to NASA MSFC, AL. At MSFC data are firstly transferred to the AMS relay computers using the UDP protocol and finally to the AMS SOC at CERN, Geneva, Switzerland with bbftp software. The sustained transferring speed from MSFC to SOC up to 50 Mbit/sec is achieved. [Section 4.3]
- Data arrived at SOC are de-framed and checked for consistency. Then events are indexed and reassembled into the AMS raw files, one file per run. Resulted files are moved to permanent storage and registered in the AMS database. Reconstructed data are available for the analysis in two hours after the raw data arrive in SOC. [Section 4.3]
- Each version of full reproduction, performed with updated software is called ”pass.” Second production uses the computing power of both CERN and remote computing centers. Data transfer between CERN and remote computing centers are performed with GridFTP and bbftp protocols. Calibration data are also available via cvmfs. [Section 4.3]
- The AMS data arrives at CERN, de-framed and processed at AMS SOC, making available detector calibrations. AMS SOC transfers raw data with calibrations to remote centers for distributed production of AMS Level 2 data. After production all data are copied back to CERN, stored at EOS, and archived on CTA. The full reprocessing of current AMS dataset takes about 1 month. Between rounds of AMS data production remote centers are employed for MC simulations of the various cosmic-ray species interactions with AMS instrument. The simulated events then undergo the same reconstruction procedure as used for events collected by AMS detector on ISS. These data are also copied to CERN. The collaborators are free to copy all available data for further analyses. [Section 4.3]
- AMS used to use Globus and GridFTP for cross-site data moving, but as CERN dropped Globus and GridFTP support, we now use XRDCP to transfer data across sites. [Section 4.3]
- AMS does not leverage any commercial cloud services for their computing or storage needs. [Section 4.8]
- AMS does not require any connectivity to be provided by wireless or satellite connectivity, beyond their core networking use case that leverages TDRS. [Section 4.8]

6.14.2 AMS: Experiment Profile

Since 2011, AMS has collected more than 200 billion events. The AMS collaboration has published 28 scientific articles, 13 of which are Editors' Suggestion, 6 are Featured in Physics and 2 received the Physics Viewpoint. The full list of the AMS publications can be found online[122]. The collection of the physics results from the first seven years with updated AMS data has been documented.[123]

AMS produces ~40 TB of raw data per year, which are transmitted to CERN, and then reconstructed to 170 TB data per year available for analysis by collaboration members. The production is performed incrementally every six months. However, in case there are significant changes in reconstruction software, a full reproduction is made.

6.14.2.1 Science Background

Each version of full reproduction, performed with updated software is called "pass." Currently, AMS acquired over 500 TB of raw data, from which a 2 PB of pass 8 data were produced and copied to collaborators institutions for further analyses.

AMS data analysis requires massive MC simulation, performed by computing centers on a regular time basis. Currently 23 PB of MC data were produced since the start of the project in 2011.

All AMS data reconstruction and simulation jobs are coordinated by AMS SOC at CERN.



Figure 6.14.1: AMS is a unique precision magnetic spectrometer on the ISS. AMS will operate on the ISS for the Station's lifetime.

6.14.2.2 Collaborators

Construction, integration, testing and operation of the AMS is carried out by an international team, referred to as the AMS Collaboration, composed of about 600 physicists and 44 institutes and organizations from 14 countries: North America (USA, Mexico), Asia (China, Korea, Taiwan), Europe, including nine European Space Agency (ESA) Member States (Denmark, Finland, France, Germany, Italy, The Netherlands, Portugal, Spain, Switzerland), and organized under an agreement between NASA and the US DOE. The geographical locations of the organizations are presented on Figure 6.14.2.

122 https://ams02.space/publications

123 M. Aguilar et al. (AMS collaboration). The Alpha Magnetic Spectrometer (AMS) on the International Space Station: Part II - Results from the first seven years. Physics Reports, 894:1–116, 2021. doi:10.1016/j.physrep.2020.09.003

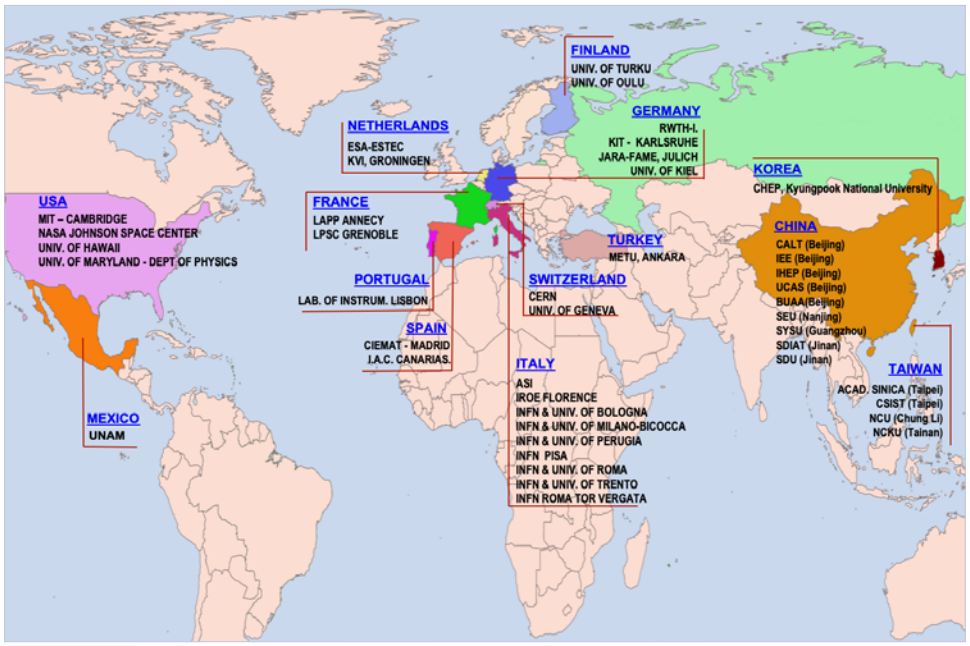


Figure 6.14.2: Overview of AMS-02: a US DOE led International Collaboration of 14 Countries, 44 Institutes and 600 Physicists.

AMS is not only the largest scientific instrument installed on the ISS, but it could also be considered the result of the largest international collaboration for a single experiment in space.

CERN, the European Organisation for Nuclear Research, is the main base of operations (at the Payload Operations Control Center, POCC) and research meetings. AMS also maintains strong cooperation with space agencies NASA and ESA. Computing centers, additional POCC facilities, and other crucial infrastructure are hosted by other AMS collaborating institutions.

A full list of AMS collaborators is as follows:

- China
  - BUAA, Beihang University, Beijing
  - IEE, Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing
  - IHEP, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing
  - University of Chinese Academy of Sciences, Beijing
  - SYSU, Sun Yat-set University, Guangzhou
  - ZJU, Zhejiang University, Hangzhou
  - SDIAT, Shandong Institute of Advanced Technology, Jinan
  - SDU, Shandong University, Jinan
  - SEU, Nanjing
- Finland
  - University of Oulu, Oulu
  - University of Turku, Turku

- France
  - LAPP Annecy, Laboratoire d'Annecy de Physique des Particules, Annecy
  - LPSC Grenoble, Laboratoire de Physique Subatomique & Cosmologie, Grenoble
- Germany
  - I. Physics Institute and JARA-FAME, RWTH Aachen University, Aachen
  - University of Kiel, Kiel
- Italy
  - INFN and University of Bologna, Bologna
  - IROE Florence, Florence
  - INFN and University of Milano-Bicocca, Milan
  - INFN and University of Perugia, Perugia
  - INFN and University of Trento, Povo, Trento
  - ASI, Italian Space Agency, Rome
  - INFN Roma Tor Vergata, Rome
  - INFN Rome and Sapienza University, Rome
- Korea
  - CHEP, Kyungpook National University, Daegu
- Mexico
  - UNAM, National Autonomous University of Mexico, Mexico City
- Netherlands
  - University of Groningen
- Portugal
  - LIP, Laboratory of Instrumentation and Experimental Particle Physics, Lisbon
- Spain
  - CIEMAT, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
  - IAC Canarias & University of La Laguna, Tenerife
- Switzerland
  - CERN, European Organization for Nuclear Research, Geneva
  - DPNC, Université de Genève
- Taiwan
  - NCSIST, National Chung-Shan Institute of Science and Technology, Longtan, Tao Yuan
  - Academia Sinica, Nankang, Taipei
  - Academia Sinica Grid Center, Nankang, Taipei
  - NCKU, National Cheng Kung University, Tainan
  - NCU, National Central University, Tao Yuan, Taiwan

- Turkey
  - METU, Middle East Technical University, Ankara
- USA
  - MIT, Cambridge, MA
  - University of Maryland (Institute for Physical Science and Technology, IPST), College Park, MD
  - University of Hawaii, Honolulu, HI
  - NASA, JSC, Johnson Space Center, Houston, TX
  - University of Maryland (EWCSS), Maryland

Since specific data analyses are performed at different locations, several international institutions provide computing support for AMS data storage, analysis, and MC data production. The amount of raw data, as well as reconstructed and simulated, can be well estimated.

A list of the analysis and computing centers dedicated to AMS is shown in the following table:

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g., 125 GB)	Frequency of data transfer or download? (e.g., ad hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
MSFC HUNTSVILLE, AL, USA	Data origination	UDP transfer	~40 TB/year	Continuous	No	N
CERN GENEVA, SWITZERLAND	Primary	XRootD, cvmfs	~210 TB/year (ISS) ~1 PB/year (MC)	Continuous	No	N
CNAF BOLOGNA, ITALY	Secondary	XRootD	~210 TB/year (ISS) ~100 TB/year (MC)	Continuous	Yes	N
RWTH AACHEN, GERMANY	Secondary	XRootD	~210 TB/year (ISS) ~100 TB/year (MC)	Continuous	Yes	N
SEU NANJING, CHINA	Secondary	XRootD	~210 TB/year (ISS) ~100 TB/year (MC)	Continuous	Yes	N
ACAD. SINICA TAIPEI, TAIWAN	Secondary	XRootD	~210 TB/year (ISS) ~100 TB/year (MC)	Continuous	Yes	N
SDIAT JINAN, CHINA	Secondary	XRootD	~210 TB/year (ISS) ~100 TB/year (MC)	Continuous	Yes	N

Table 6.14.1: AMS Data Mobility

Currently, a new data analysis and computing center is being developed at the University of Hawaii at Manoa in the USA. This will be the first computing and analysis center on US soil dedicated exclusively to AMS data. About 20000 CPU cores and 10 PB of data storage will be employed for AMS data production, analysis, MC simulations, and archiving.

6.14.2.3 IRI Readiness

AMS does not directly leverage DOE computing or networking resources, and is unable to address questions regarding IRI at this time. The project will monitor the progress of the IRI effort in the event that some of the advancements may apply to their process of science.

6.14.2.4 Instruments and Facilities

AMS has more than 300K readout channels and 1500 temperature sensors. The acquisition rate goes up to 2 kHz. There are more than 500 microprocessors onboard which reduce data rate from 7 Gbps to 10 Mb/s. Data taken on board of the ISS by AMS consist of raw events, where each event is the collection of the information read out from relevant AMS sub-detectors electronics after the latter are triggered by particle(s) crossing AMS sensitive volumes. Raw events are logically grouped in sequences called runs, each run having a unique 32 bit identifier. Events inside one run are numbered in sequence.

Every quarter of the orbit a scientific run starts. The duration of each run corresponds to 23 minutes, and every run contains on average 700k events. On orbits which are close to the Earth magnetic poles, the number of events in a run can exceed 2 million. The average size of raw event is 2 KB, thus yearly raw data yield is 40 TB.

On top of this data volume additional calibration information, like the AMS electronics parameters, as well as various temperatures, pressures and other health and status data are recorded regularly. This calibration related information corresponds to less than 1% of the total data volume.

Recorded data are packed to fixed length packets, or frames, and transmitted via the TDRS satellites to ground at White Sands, NM, and then relayed to NASA MSFC, AL. At MSFC data are firstly transferred to the AMS relay computers using the UDP protocol and finally to the AMS SOC at CERN, Geneva, Switzerland with bbftp[124] software, the latter using multiple Transmission Control Protocol (TCP) streams between client and server to speed up data transfer. The sustained transferring speed from MSFC to SOC up to 50 Mbit/sec is achieved.

Data arrived at SOC are de-framed and checked for consistency. Then events are indexed and reassembled into the AMS raw files, one file per run. Resulted files are moved to permanent storage and registered in the AMS database.

The AMS Data Production, i.e., massive data reconstruction is organized in two stages. Firstly, freshly arrived data are put into first production. It runs in a fully automated manner and produces the data summary files for the quick detector performance evaluation. The detector performance is monitored in 24/7 regime at AMS POCC at CERN. First production requires computing power of ~450 CPU cores (including 100% contingency) to cope with the data rate and to ensure nearly real-time processing. Usually, the reconstructed data are available for the analysis in two hours after the raw data arrive in SOC.

The data summary files from the first production are used subsequently to produce detector calibrations, which are required for second production. Second production uses all the available calibrations, alignments and ancillary data from the ISS as well as monitoring values (temperatures, pressures, and voltages) to produce physics analysis ready dataset. Second production usually runs every six months incrementally. However, in case there are significant changes in reconstruction software, a full reproduction may be needed.

Each version of full reproduction, performed with updated software is called "pass." Currently, pass 8 data are available for analysis. Second production uses the computing power of both CERN and remote computing centers. Data transfer between CERN and remote computing centers are performed with GridFTP and bbftp protocols. Calibration data are also available via cvmfs[125].

The typical AMS raw file size is 1.5 GB, while the AMS data file in pass 8 production is 7 GB. At the moment, there are 250K pass 8 files were produced.

124 Farrache G. bbftp. URL: http://ccweb.in2p3.fr/bbftp
125 cvmfs . URL: https://cvmfs.readthedocs.io/en/stable/

By the end of 2025, AMS will undergo a significant upgrade on the ISS, carried out by astronauts. This upgrade will involve the addition of a new tracker layer at the top of the instrument, as illustrated in Figure 3. The new layer will enhance the instrument's acceptance of high energy particles by 300% and increase the AMS sensitivity to high-Z nuclei. These advancements are essential for progressing dark matter searches and improving precision measurements of sub-Iron nuclei, which play a crucial role in refining cosmic-ray propagation models. Additionally, they enable the study of rare and exotic particles, such as strange quark matter, heavy antimatter nuclei, and supersymmetry (SUSY)T particles.

The new configuration of AMS will require a complete overhaul of simulations and data analysis tools, which will be developed over the next three to five years. Additionally, a new data analysis and computing center at UH Manoa, USA, will be established within the next two to three years. This center will handle extensive MC simulations, data production, and analyses. With both the current and upgraded geometries requiring simulated datasets, the total volume of data is expected to increase to 50 PB over the next five years.

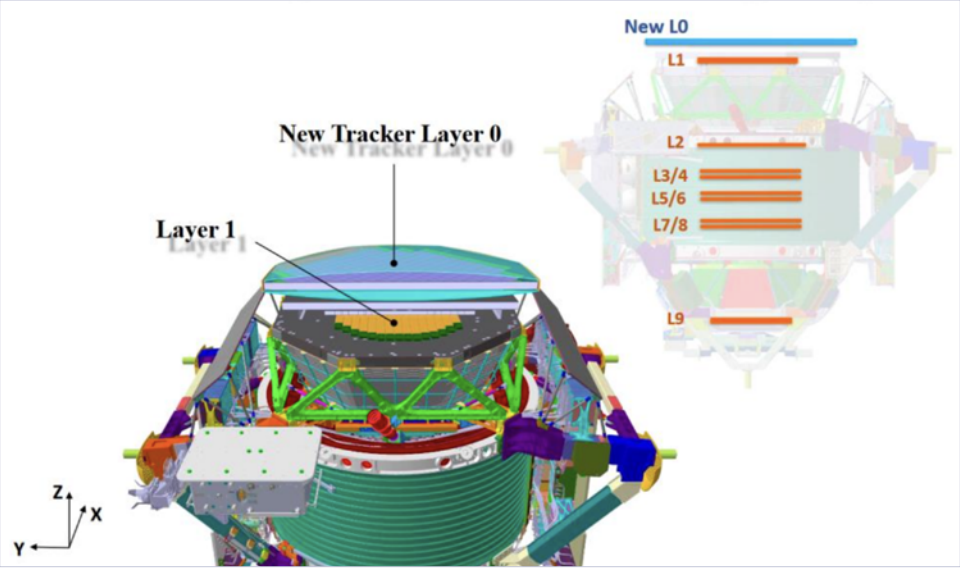


Figure 6.14.3: AMS instrument with new layer 0 added to the existing nine tracker layers

6.14.2.5 Generalized Process of Science

The general process of science is described in the previous section. Data are either reconstructed from AMS or simulated. Both are managed by SOC team at CERN. Raw data proceed through two stages of reconstruction, and analysis tuples after the second stage are available for transfer and offline analysis across the collaboration.

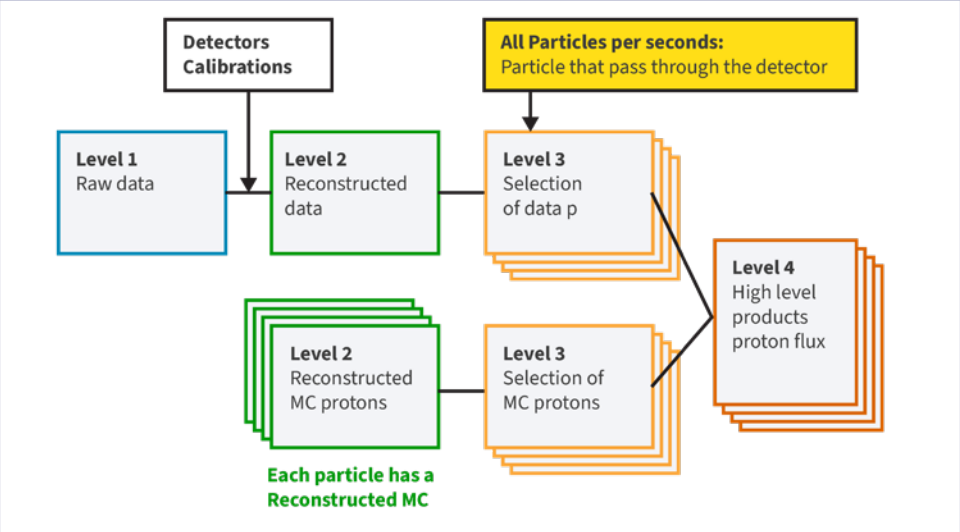


Figure 6.14.4: AMS data Levels. Level 1 data are raw data reconstructed from frame files. Level 2 data are produced with information from raw data and Detector Calibrations. Level 3 data, both for the actual detector and MC, include the number of particles per second (particle rates) of specific species with detector efficiency information to calculate the final flux. Level 4 data are the final product of the analysis, including flux and systematic errors, ready for publication.

The first stage of production is performed at AMS SOC and takes two or three hours after data frames are de-framed by checking raw events for consistency. Then, raw files are assembled, validated, and registered in the database as AMS Level 1 data, as shown in Figure 3. One raw file corresponds to one AMS run. The production jobs assemble data in ROOT files and Time-Dependent Variables (TDV) files. TDV files contain the environmental variables that are important for data analysis: for example, the temperature variations of the silicon Tracker sensors affect the alignment parameters and particle tracking efficiency. TDV files are stored on the Andrew File Service (AFS) and published to CVMFS weekly.

The AMS Level 1 data are processed every six months with a delay of 1 to 7 months since acquisition to create AMS Level 2 data. To produce Level 2 data, all the information in the TDV files, including detector calibrations, is necessary. MC simulations for all particle species are produced in parallel at Level 2. The AMS Level 3 data are generated by different groups, each of them selecting and analyzing different particle species at different time scales. Level 3 data includes the number of particles per second (particle rates) of specific species with detector efficiency information to calculate the final flux. Level 3 data are produced for both real detector data and MC data. Depending on the type of particle, Level 3 is produced with a time lag ranging anywhere from one to two years after acquisition. The AMS Level 4 data include full analysis of particle spectra after detector efficiency studies, MC corrections, and systematic error calculations. Extensive analyses and crosschecks between different groups are necessary to produce the Level 4 data that are finally ready for publication. Level 4 data are produced with a time lag of two to five years after acquisition. The software to generate Level 3 and Level 4 is not shared between the groups within the AMS collaboration; each group develops its own software independently.

6.14.2.6 Remote Science Activities

The AMS data arrives at CERN (Geneve), de-framed and processed at AMS SOC, making available detector calibrations. AMS SOC transfers raw data with calibrations to remote centers, listed in Table 1 for distributed production of AMS Level 2 data. After production all data are copied back to CERN, stored at EOS and archived on CTA. The full reprocessing of current AMS dataset takes about 1 month using 35,000-40,000 CPU cores. Between rounds of AMS data production remote centers are employed for MC simulations of the various cosmic-ray species interactions with AMS instrument. The simulated events then undergo the same reconstruction procedure as used for events collected by AMS detector on ISS. These data are also copied to CERN. The collaborators are free to copy all available data for further analyses.

6.14.2.7 Software Infrastructure

6.14.2.7.1 Data Resource Management

AMS SOC has developed its own software infrastructure to manage the primary data resources, i.e., the data at CERN. The data arrive at AMS POCC which is located at CERN in the format of "frame" files. The frame files are shared to AMS SOC by Network File System (NFS). The frame files are de-framed and re-packed into raw files, and the latter is put into EOS and CTA storage by XRDCP, at the same time, metadata are registered at Oracle database.

After raw files are registered, automatic job requesting software (also developed by AMS SOC) requests first production jobs through the Apache webserver. Jobs are distributed by the SOC production server to its producers, and in case local computing resources are not enough, jobs may be submitted to CERN LXBATCH computing farm using HTCondor. When first production jobs are done, SOC validating scripts check the produced ROOT files and copy them to EOS, and these data are backed up to CTA storage using the FTS platform.

Secondary "pass-N" production shares almost the same procedure except that more data moving may be involved which will be discussed in the next section.

6.14.2.7.2 Data mobility

Secondary "pass-N" production usually requires data transferring between CERN EOS and remote sites. AMS used to use Globus and GridFTP for cross-site data moving, but as CERN dropped Globus and GridFTP support, XRDCP is now used to transfer data across sites.

6.14.2.7.3 Data Process

GBATCH is a common software package, developed by AMS collaboration and used for data production from raw to AMS Level 2 data and MC simulations. It is linked against ROOT, Geant4, CERNLib, and some other HEP libraries and serves as a common framework for all collaboration data analyses.

The above-mentioned software tools may evolve over time, however, as AMS uses EOS and CTA as primary data storage, it is necessary to use the software tools supported by CERN IT.

6.14.2.8 Network and Data Architecture

6.14.2.8.1 The LAN Capabilities

AMS science data, including raw data and ROOT data ready for physics analysis, are stored at CERN EOS, so there is no LAN capability to be reported.

6.14.2.8.2 The Local Network Configuration

AMS science data, including raw data and ROOT data ready for physics analysis, are stored at CERN EOS, so there is no local network configuration to be reported.

6.14.2.8.3 Network Bandwidth Connection Speed

AMS science data, including raw data and ROOT data ready for physics analysis, are stored at CERN EOS, so the connection speed is determined by the bandwidth of CERN computing facilities.

6.14.2.9 Emerging Technology Use

6.14.2.9.1 Cloud Services

AMS does not leverage any commercial cloud services for their computing or storage needs.

6.14.2.9.2 AI and ML

AMS does not foresee any research activities into the use of AI or ML in the coming years.

6.14.2.9.3 Wireless or Distributed Location Science Activities

AMS does not require any connectivity to be provided by wireless or satellite connectivity, beyond their core networking use case that leverages TDRS.

6.14.2.10 Data-Related Resource Constraints

There are no data-related resource constraints to be reported.

6.14.2.11 Data Mobility Endpoints

AMS collaboration provides data to the science community with papers and tabulated datasets, accessible online. The AMS Level 4 data can be found in publications[126] and data archives:

- Grenoble cosmic-ray database (CRDB) database (France)[127]
- ASI CRDB database (Italy)[128]
- NASA HEASARC Archive (USA)[129]

Recently the work on providing AMS data to NASA HEASARC Archive started according to NASA/DOE and AMS Interagency Agreement regarding AMS Science Data Archiving. University of Hawai'i at Manoa, on behalf of AMS collaboration, will provide AMS Level 4 data as well as AMS Level 3 data related to all particle rates, lifetime, AMS location, and zenith angle at a given second.

6.14.2.12 Outstanding Issues

There are no outstanding issues to be reported.

6.14.2.13 Facility Profile Contributors

AMS Representation

Veronica Bindi, University of Hawaii at Manoa, bindi@hawaii.edu

Baosong Shan, BUAA, Beihang University, baosong.shan@cern.ch

Nikolay Nikonov, University of Hawaii at Manoa, nikonov@hawaii.edu

126 https://ams02.space/publications
127 https://lpsc.in2p3.fr/crdb/
128 https://tools.ssdc.asi.it/
129 https://heasarc.gsfc.nasa.gov

7 Technology Profiles

The technical profiles presented in this document are a written record of the current state of support offered by DOE SC–funded laboratories and user facilities in collaboration with HEP-funded activities. These case studies were discussed between July 2024 and December 2024. The profiled laboratories and user facilities include:

- NERSC
- ALCF
- OLCF
- Fermilab
- BNL
- SLAC National Accelerator Laboratory
- ANL

A summary of each will be presented prior to the case study document, along with a "Discussion Summary" that highlights key areas of conversation from authors and attendees. These brief write-ups are not meant to replace a full review of the case study, but will provide a snapshot of the discussion and focus during the in-person review.

7.1 NERSC

The mission of NERSC is to accelerate scientific discovery at the DOE OS through HPC and data analysis.

As the primary scientific computing facility for the US DOE SC, NERSC is the principal provider of HPC and data resources and services. Fundamental to the mission of NERSC is enabling computational science of scale, in which large, interdisciplinary teams of scientists attack fundamental problems in science and engineering that require massive calculations and have broad scientific and economic impacts.

7.1.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- NERSC supplies supercomputing, data storage, networking, auxiliary equipment and technical consulting to more than 10,000 users and 1,000 projects conducting research support by the DOE SC. NERSC supports approximately 100 HEP projects, both simulation and data focused. [Section 4.1]
- NERSC-10 is expected in the 2027 time frame, and is anticipated to be 10X the aggregate performance of Perlmutter. NERSC-10 is expected to offer more flexibility toward assembling and composing the required set of capabilities on a per-user basis, will be able to support urgent and interactive workflows from DOE SC's experimental and observational facilities, and provide seamless performance across converged networks, accelerators, and storage components. [Section 4.1]
- NERSC features 1 Tbps Aggregate Bandwidth to ESNET via 2 x 400 Gbps links and 2 x 100 Gbps links. NERSC will be expanding to a 4 x 400 Gbps architecture in future years, in addition to considering a migration to 800 Gbps technology. [Section 4.7]

- For NERSC-10 and beyond, NERSC will consider incorporating "cloud" technologies that add significant value to the NERSC user community and are affordable. This approach to provide cloud-like services includes NERSC Spin; a container-based platform where users can build their own websites, API endpoints, databases, and other network services that can integrate with center networks and data storage systems. In the next two to five years, NERSC is exploring deeper hybridization of Spin with the HPC system itself, which will open possibilities for performing computational tasks in a more cloud-like environment. [Section 4.8]

7.1.2 NERSC: Technology Profile

NERSC supplies supercomputing, data storage, networking, auxiliary equipment and technical consulting to more than 10,000 users and 1,000 projects conducting research support by the DOE SC.

7.1.2.1 Facilities Operated and Supported

7.1.2.1.1 Computing

Perlmutter is a HPE (Hewlett Packard Enterprise) Cray EX supercomputer, named in honor of Saul Perlmutter, an astrophysicist at Berkeley Lab who shared the 2011 Nobel Prize in Physics for his contributions to research showing that the expansion of the universe is accelerating.

Perlmutter, based on the HPE Cray Shasta platform, is a heterogeneous system comprising both CPU-only and GPU-accelerated nodes, with a performance of three to four times the previous Cori system.

Perlmutter consists of 1536 GPU-accelerated nodes with 1 AMD Milan processor and 4 Nvidia A100 GPUs, and 3072 CPU-only nodes with 2 AMD Milan processors.

- Perlmutter
  - 79 PF (Top 500 FP64 Linpack)
  - #12 in list of Top 500 most powerful supercomputers
  - 1,792 Nvidia A100 GPU-accelerated nodes (4 GPU/node)
  - 7,168 A100 GPUs total
  - 3,072 dual-socked AMD Milan CPU node
  - 35 PB Flash scratch file system

7.1.2.1.2 Storage

File System	Medium	Size
HOME	Disk	450 TB
GLOBAL COMMUNITY (CFS)	Disk	130 PB
PERLMUTTER SCRATCH	SSD	35 PB
ARCHIVE	Tape / Disk cache	300 PB

Table 7.1.1: NERSC Storage Resources

7.1.2.1.3 Networking

See Section 7.1.2.2.

7.1.2.1.4 Planned Computing

NERSC-10

- 2027 time frame
- 10X Perlmutter aggregate performance on applications
- The first configurable and programmable supercomputer that allows users to assemble and compose the set of capabilities that work best for their research project,
- The first supercomputing system designed to support urgent and interactive workflows from DOE SC's experimental and observational facilities, and
- The first HPC system to provide seamless performance across converged networks, accelerators, and storage components.

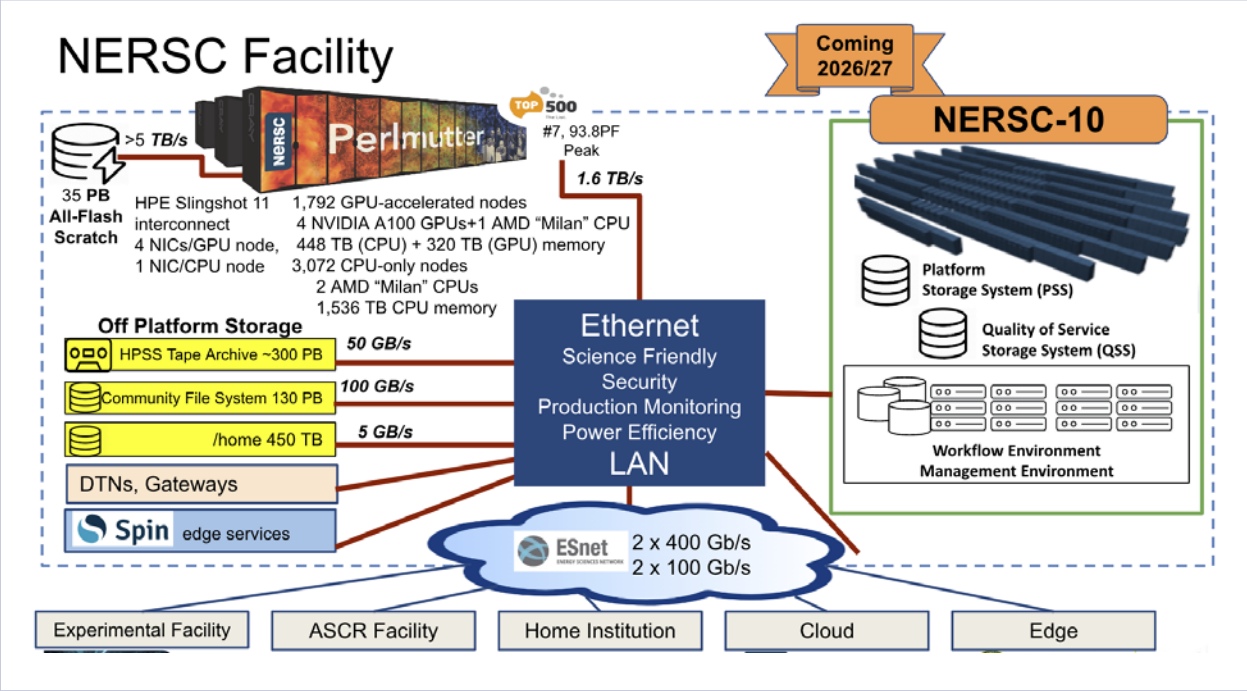


Figure 7.1.1: NERSC Facility

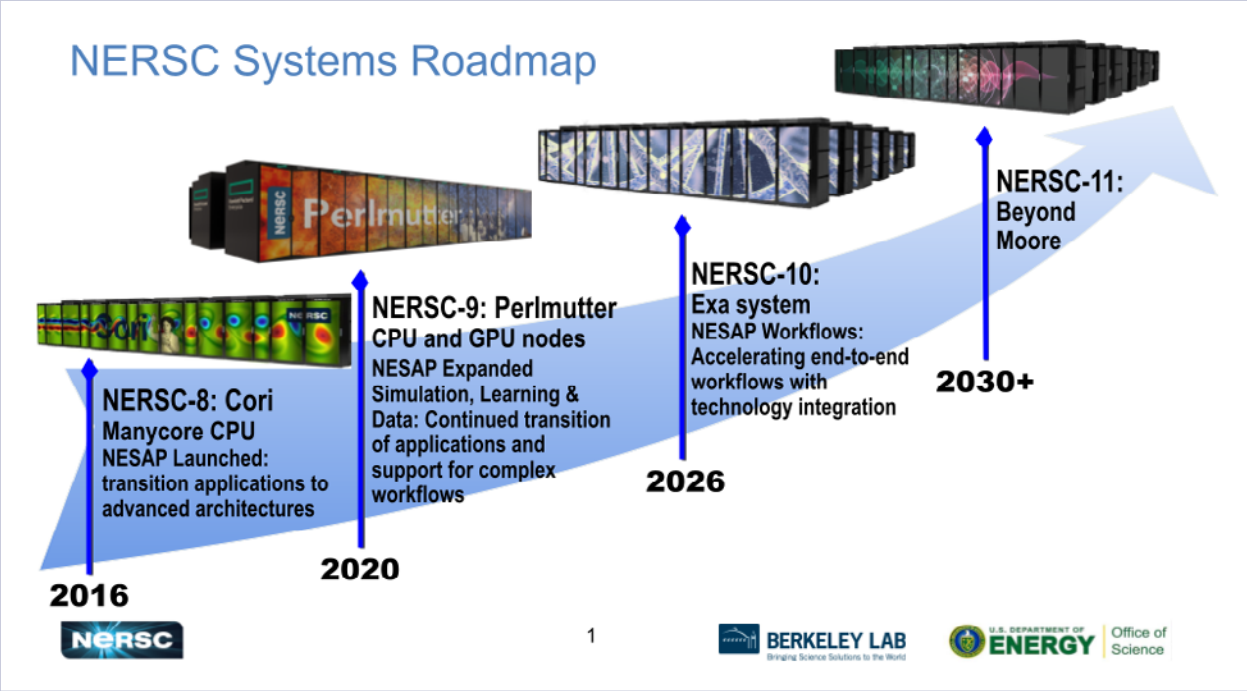


Figure 7.1.2: NERSC Systems Roadmap

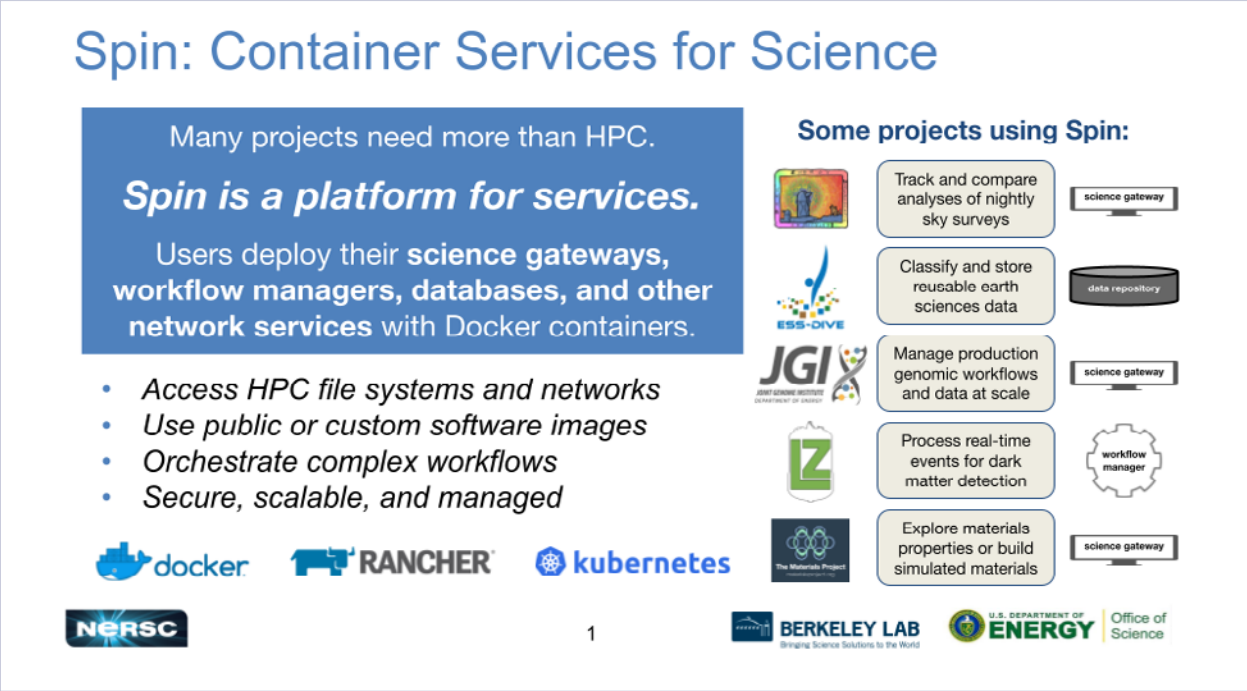


Figure 7.1.3: NERSC SPIN Service

7.1.2.1.5 HEP Projects Supported by NERSC

NERSC supports 100 HEP projects, both simulation and data focused.

7.1.2.1.5.1 Top 10 HEP Compute Projects

Total of 342 M CPU core hours used in 2024 to date

Total of 3.9 M Nvidia A100 GPU Hours used in 2024 to date

PI	Project Title	Organization	Science Category
GUTSCHE, OLIVER	Particle Physics with CMS at the LHC	Fermi National Accelerator Laboratory	Physics : High Energy Physics (Experimental)
CALAFIURA, PAOLO	Detector Simulation of the ATLAS Detector on NERSC HPCs	Lawrence Berkeley National Laboratory	Physics : High Energy Physics (Experimental)
GUPTA, RAJAN	Lattice QCD search for physics beyond the standard model	Los Alamos National Laboratory	Physics : High Energy Physics (Theory)
EL-KHADRA, AIDA	The anomalous magnetic moment of the muon	University of Illinois at Urbana-Champaign	Physics : High Energy Physics (Theory)
BAILEY, STEPHEN	DESI	Lawrence Berkeley National Laboratory	Physics : Cosmology
TSUNG, FRANK	Continuing studies of plasma based accelerators	University of California Los Angeles (UCLA)	Physics : Accelerator Science
ZUNTZ, JOE	Dark Energy Survey	University of Edinburgh	Physics : Cosmology
MONZANI, MARIA ELENA	LZ - LUX ZEPLIN	SLAC National Accelerator Laboratory	Physics : High Energy Physics (Experimental)
BARON, EDWARD	Synthetic Spectra of Astrophysical Objects	University of Oklahoma	Physics : Astrophysics
HABIB, SALMAN	Cosmological Hydrodynamic Simulations with CRK-HACC	ANL	Physics : Cosmology

Table 7.1.2: Top 10 HEP Compute Projects

7.1.2.1.5.1 Top 10 Data Projects (Data Stored)

Total of 66 PB of permanent data stored by NERSC by active HEP projects

PI	Project Title	Organization	Science Category	Permanent Data Storage at NERSC (TB)
BAILEY, STEPHEN	DESI	Lawrence Berkeley National Laboratory	Physics : Cosmology	13,701
MONZANI, MARIA ELENA	LZ - LUX ZEPLIN	SLAC National Accelerator Laboratory	Physics : High Energy Physics (Experimental)	12,204
DIGEL, SETH	LSST DESC Simulations and Analysis for Cosmology	SLAC National Accelerator Laboratory	Physics : Cosmology	6,131
GUPTA, RAJAN	Lattice QCD search for physics beyond the standard model	Los Alamos National Laboratory	Physics : High Energy Physics (Theory)	4,122
LUK, KAM-BIU	Daya Bay Reactor Neutrino Experiment	Lawrence Berkeley National Laboratory	Physics : High Energy Physics (Experimental)	3,797
KISNER, THEODORE	CMB Data Analysis Towards CMB-S4: Design Studies	Lawrence Berkeley National Laboratory	Physics : Cosmology	3,511
BAILEY, STEPHEN	Cosmology Data Repository	Lawrence Berkeley National Laboratory	Physics : Cosmology	3,151
NUGENT, PETER	DECaLS & the Zwicky Transient Facility	Lawrence Berkeley National Laboratory	Physics : Astrophysics	2,711
KISNER, THEODORE	CMB Data Analysis Towards CMB-S4	Lawrence Berkeley National Laboratory	Physics : Cosmology	2,667
WERTHIMER, JERROLD	Berkeley High Resolution Neutral Hydrogen Sky Survey	University of California Berkeley	Physics : Astrophysics	1,998

Table 7.1.3: Top 10 Data Projects (Data Stored)

7.1.2.2 Network and Data Architecture

7.1.2.2.1 Current NERSC Network

- 1 Tbps Aggregate Bandwidth to ESNET
  - 2 x 400 Gb links + 2 x 100 Gb links
- Non-Blocking Architecture for North-to-South data transfers, scalable model for providing bandwidth for East-West traffic
- Total 1.6 Tbps deployed bandwidth for Perlmutter to data center connectivity
- 3.2 Tbps planned deployment for CFS and DTN connectivity to Perlmutter after storage fabric deployment

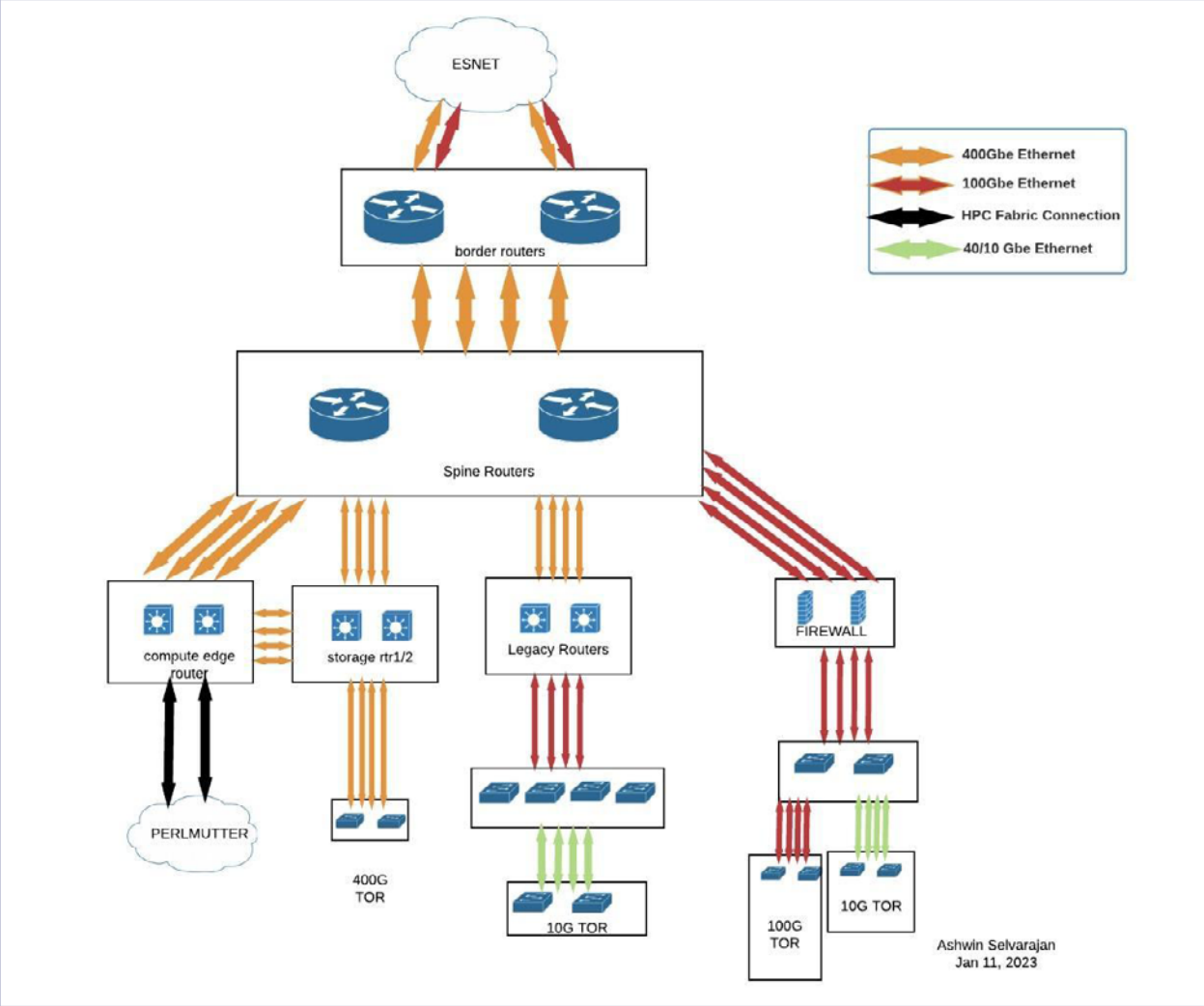


Figure 7.1.4: NERSC Network Diagram

7.1.2.2.2 Next (Two to Five Year) Network Design

- Looking to expand bandwidth to (4) x 400 Gb connections to ESnet, and include MP-BGP, EVPN and dynamic routing to the North-South traffic.
- Around the time of NERSC-10 there may be a transition to 800 Gb technology in support of Empirical Orthogonal Function workflows.

7.1.2.3 Computational and Storage Infrastructure

NERSC offers a robust infrastructure for efficient data handling, including high-performance DTNs, squid servers, a dedicated CVMFS cache for each compute node, and Frontier cache[130], along with a suite of data transfer tools such as Globus, XRootD TPC over HTTPS, scp, bbcp, and GridFTP.

The DTNs are specialized servers with dual 100-gigabit Ethernet links for fast external transfers over the ESnet WAN and 10-gigabit links for internal NERSC transfers. These nodes provide seamless access to storage resources like the CFS, Perlmutter Scratch and HPSS, ensuring high-speed data movement between NERSC and external sites.

Globus, as a widely used data transfer tool at NERSC, is configured with specific endpoints to support different workflows and storage requirements:

- NERSC DTN: A multi-node endpoint that enables high-speed access to all major NERSC file systems, including CFS, Global Homes, Global Common, and Perlmutter Scratch. This endpoint is ideal for most transfers into and out of NERSC.
- NERSC Perlmutter: Another multi-node endpoint similar to NERSC DTN, serving as a backup for Perlmutter-related transfers when the primary DTN endpoint is unavailable.
- NERSC HPSS: A single-node endpoint optimized for direct access to the HPSS tape archive, useful for archival data transfers though potentially slower for large volumes.
- NERSC SHARE: A read-only, single-node endpoint offering access to specific shared file systems for data distribution.
- NERSC S3: Designed for transferring data directly to and from Amazon S3 buckets, facilitating cloud integration and hybrid workflows.

Over the next two to five years, NERSC plans to maintain its current suite of data management tools while leveraging the expected evolution of these tools. As tools evolve, it is anticipated that improved capabilities for facilitating complex, multifacility workflows, real-time processing, and network QoS adjustments to prioritize critical data flows will be available.

7.1.2.4 Cloud Services

For NERSC 10 and beyond, NERSC will consider incorporating "cloud" technologies that add significant value to the NERSC user community and are affordable. The NERSC 10 system itself will be located at Berkeley Lab in the machine room of Shyh Wang Hall (Building 59).

NERSC Spin, a container-based platform where users can build their own websites, API endpoints, databases, and other network services that can integrate with center networks and data storage systems. Over 80 NERSC projects are using Spin to complement their use of batch HPC e.g., data dissemination to their research communities or to automate complex research workflows. In the next two to five years, NERSC is exploring deeper hybridization of Spin with the HPC system itself, which will open possibilities for performing computational tasks in a more cloud-like environment by a) making HPC resources accessible to containerized workloads and b) providing a platform for NERSC to deploy novel user-facing capabilities such as Function-as-a-Service

130 https://docs.nersc.gov/services/frontiercache/

7.1.2.5 AI and ML

NERSC has been involved in modern deep learning and AI for the last 10 years. This includes activities related to deployment of system hardware and software optimized for AI; deep application engagement with science teams; and broad empowerment of the community, through for example training and schools.

Current examples in this area include deployment of Perlmutter a system with over 7000 A100 GPUs (which was the optimum AI accelerator at time of deployment) and work with Nvidia and HPE to ensure performant large-scale deep learning training using NVIDIA Collective Communication Library (NCCL) with HPE slingshot interconnect. Application engagement has been through for example the NERSC Exascale Scientific Applications Program (NESAP) that has resulted in multiple publications in domain science and ML venues. Training events have included over 20 deep learning at-scale tutorials with thousands of participants, including at the SC conference (and made publicly available) since 2018.

In the next two to five years NERSC-10 will be deployed. This will provide significant improvements in performance for AI workloads as well as the integration of AI with traditional simulation and data pipelines. This will be complemented with software and services to facilitate the reuse of AI models, and continued work on applications, focused on large-scale training and the integration of fine-tuning and inference into science codes.

Starting in this period and moving beyond five years, NERSC's strategy is towards pervasive AI. This means enabling AI to accelerate science and HPC operations wherever and whenever it is able to do so. This is expected to include NERSC being available for the training, hosting and fine-tuning of mission-science 'foundational' AI models as well as AI-ready datasets, but also the operation of highly instrumented "self-driving" HPC systems, and providing an AI ecosystem for science including AI-assisted interfaces.

7.1.2.6 Wireless Capabilities

This section does not apply to the NERSC case study.

7.1.2.7 Data-Related Resource Constraints

No constraints identified.

7.1.2.8 Data Mobility Endpoints

The DTNs are NERSC servers dedicated to performing transfers between NERSC data storage resources such as HPSS, the Community File System, Perlmutter Scratch, and storage resources at other sites. These nodes are being managed (and monitored for performance) as part of a collaborative effort between ESnet and NERSC to enable high-performance data movement over the high-bandwidth 100 Gb ESnet WAN. The most up-to-date documentation can be found online.[131]

7.1.2.8.1 Configuration

All DTNs have :

- Two 100-gigabit ethernet links for transfers to internet
- Two 10-gigabit ethernet links to transfer to NERSC internal resources (HPSS)
- Two FDR IB connections to the global file system

Similar to other NERSC systems, users will maintain their own shell configuration files ("dot files").

7.1.2.8.2 Access

There are four DTNs available for interactive use (e.g., via ssh) at dtn0[1-4].nersc.gov, and a pool of nodes used for automated transfer services like Globus.

131 https://docs.nersc.gov/systems/dtn/

All NERSC users are automatically given access to the DTNs. The nodes support both interactive use via Su–Schrieffer–Heeger (SSH) (direct login) or data transfer using GridFTP services; however, it is recommended to use Globus for large data transfers.

7.1.2.8.3 Available File Systems

The NERSC DTNs provide access to Global Homes, Global Common, the CFS, Perlmutter Scratch, and HPSS. On the DTNs Perlmutter Scratch is mounted at /global/pscratch and can be accessed using the $SCRATCH environment variable.

7.1.2.8.4 File Transfer Software

For most cases, it is recommended to use Globus because it provides the best transfer rates. It makes transfers trivial so users do not have to learn command line options for manual performance tuning. Globus also does automatic performance tuning and has been shown to perform comparable to -- or even better (in some cases) than -- expert-tuned GridFTP. Users can also use GridFTP to transfer large files by hand.

For smaller files you can use Secure Copy (scp) or Secure file transfer protocol (sftp) or rsync to transfer files between two hosts.

7.1.2.8.5 Restrictions

In order to keep the DTNs performing optimally for data transfers, NERSC requests that users restrict interactive use of these systems to tasks that are related to preparing data for transfer or are directly related to data transfer of some form or fashion. Examples of intended usage would be running python scripts to download data from a remote source, running client software to load data from a file system into a remote database, or compressing (gzip) or bundling (tar) files in preparation for data transfer. Examples of what should not be done include running a database on the server to do data processing, or running tests that saturate the nodes resources. The goal is to maintain data transfer systems that have adequate memory and CPU available for interactive user data transfers.

7.1.2.9 Outstanding Issues

No issues identified.

7.1.2.10 Technology Profile Contributors

NERSC Representation

Debbie Bard, NERSC, djbard@lbl.gov

Wahid Bhimji, NERSC, wbhimji@lbl.gov

Shane Canon, NERSC, scanon@lbl.gov

Jack Deslippe, NERSC, jrdeslippe@lbl.gov

Pengfei Ding, NERSC, pding@lbl.gov

Richard Gerber, NERSC, ragerber@lbl.gov

Cory Snavely, NERSC, csnavely@lbl.gov

7.2 ALCF

The ALCF, a US DOE SC user facility located at ANL, enables breakthroughs in science and engineering by providing supercomputing resources and expertise to the research community.

7.2.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The ALCF operates multiple supercomputing systems and enclaves that support all communities, including HEP: [Section 4.1]
  - Polaris: a 560 node HPE Apollo 6500 Gen 10+ based system.
  - Sophia: a 24-node Nvidia DGX A100-based system
  - Aurora: a 10,624 node HPE Cray Ex based system
  - Crux: a 1PF CPU-only system that is currently being repurposed
  - The Joint Lab for Systems Evaluation (JLSE)
  - The ALCF AI Testbed
- The ALCF is planning for the following new systems in the two- to five-year time range: [Section 4.1]
  - GraceHopper Inference system
  - SambaNova Inference system
  - Education and Security Enclave
  - The ALCF-4 system is in early planning stages.

7.2.2 ALCF: Technology Profile

ALCF computing resources—available to researchers from academia, industry, and government agencies—support large-scale, computationally intensive projects aimed at solving some of the world's most complex and challenging scientific problems. Through awards of supercomputing time and support services, the ALCF enables its users to accelerate the pace of discovery and innovation across disciplines.

The ALCF's team of computational scientists, data scientists, performance engineers, system administrators, software developers, visualization experts, and support staff help ensure users get the most out of the facility's HPC systems. The ALCF also provides training and expertise to prepare researchers for the next generation of leadership computing resources.

7.2.2.1 Facilities Operated and Supported

The ALCF operates multiple supercomputing systems:

- Polaris is a 560 node HPE Apollo 6500 Gen 10+ based system. Each node has a single 2.8 GHz AMD EPYC Milan 7543P 32 core CPU with 512 GB of DDR4 RAM and four Nvidia A100 GPUs connected via NVLink, a pair of local 1.6TB of SSDs in RAID0 for the users' use, and a pair of Slingshot 11 network adapters. There are two nodes per chassis, seven chassis per rack, and 40 racks. More details about Aurora can be found on the ALCF Polaris page[132].

132 https://docs.alcf.anl.gov/polaris/hardware-overview/machine-overview

- Sophia is a 24-node Nvidia DGX A100-based system. The DGX A100 comprises eight Nvidia A100 GPUs that provide a total of 320 gigabytes of memory for training AI datasets, as well as high-speed Nvidia Mellanox ConnectX-6 network interfaces. More details about Sophia can be found on the ALCF Sophia page.[133]

Systems that are in flight:

- Aurora is a 10,624 node HPE Cray Ex based system. It has 166 racks with 21,248 CPUs and 63,744 GPUs. Each node consists of 2 Intel Xeon CPU Max Series (codename Sapphire Rapids or SPR) with on-package high-bandwidth memory (HBM) and six Intel Data Center GPU Max Series (codename Ponte Vecchio or PVC). Each Xeon has 52 physical cores supporting 2 hardware threads per core and 64 GB of HBM. Each CPU has 512 GB of DDR5. The GPUs are connected all-to-all with Intel XeLink interfaces. Each node has 8 HPE Slingshot-11 NICs, and the system is connected in a dragonfly topology. More details about Aurora can be found on the ALCF Aurora page.[134]
- Crux - No documentation is available currently. Crux is an approximately 1PF CPU-only system that is being repurposed.

Planned in the next two years (no more details available at this time):

- GraceHopper Inference system
- SambaNova Inference system
- Education and Security Enclave

Planned approximately five years out:

- The ALCF-4 system is in early planning stages.

Two other enclaves are also available:

- The JLSE aims to investigate alternative approaches to current and future deployments (both hardware and software) within ALCF. More details can be found online.[135]
- The ALCF AI Testbed houses some of the most advanced AI accelerators for scientific research. The AI accelerators complement the ALCF's current and next-generation supercomputers to provide a state-of-the-art computing environment that supports pioneering research at the intersection of AI, big data, and high-performance computing (HPC). More details can be found on the ALCF AI Testbed page.[136]

133 https://docs.alcf.anl.gov/sophia/hardware-overview/machine-overview
134 https://docs.alcf.anl.gov/aurora/hardware-overview/machine-overview
135 https://www.jlse.anl.gov
136 https://www.alcf.anl.gov/alcf-ai-testbed

7.2.2.2 Network and Data Architecture

The current ALCF Science DMZ is pictured in Figure 7.2.1.

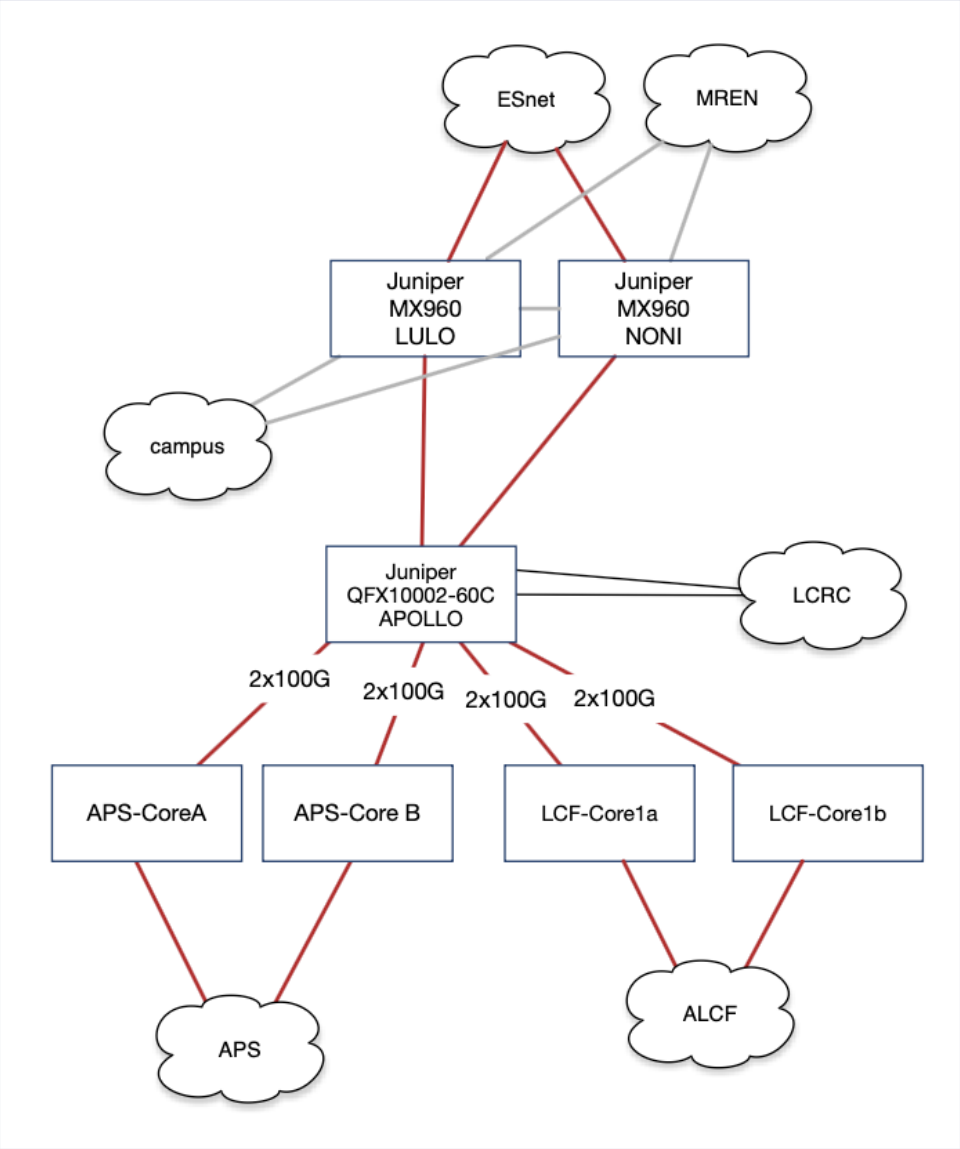


Figure 7.2.1: ALCF Science DMZ

ALCF has acquired hardware to upgrade the science DMZ and are in the process of coordinating the upgrade as shown in Figure 7.2.2.

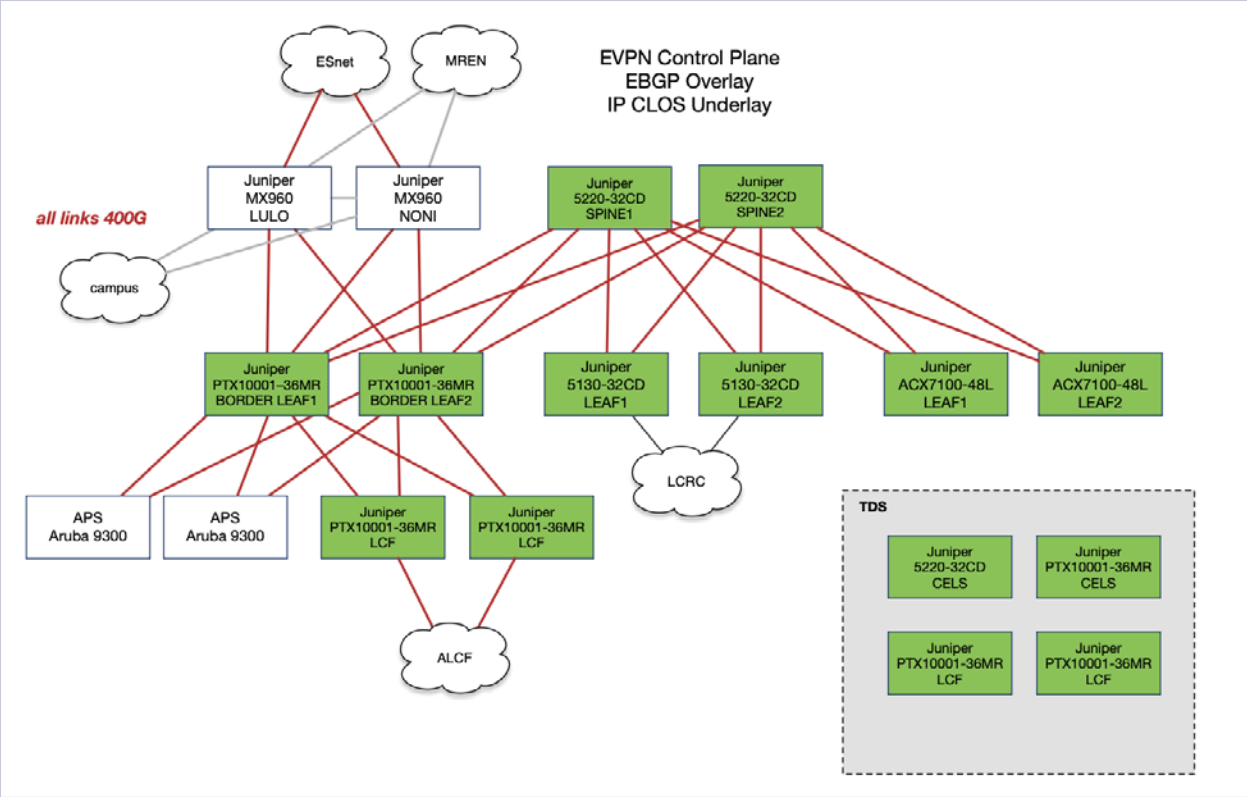


Figure 7.2.2: ALCF Network Upgrade

7.2.2.3 Computational and Storage Infrastructure

Besides the major systems listed in section 7.2.2.1, ALCF currently uses:

- Lustre File systems[137]
- PBS scheduler
- Globus tools (transfer, share, compute, flows, etc.)
- Workflow tools to support large, complex job campaigns
- JupyterHub support for running user notebooks

In the next two years ALCF is planning a Global File System upgrade. The RFI process has been completed, and the ALCF is now meeting with vendors before submitting an RFP.

7.2.2.4 Cloud Services

It is unlikely ALCF will significantly use an external cloud resources. The facility is working on building up an on-prem cloud / containerization / edge services infrastructure which should be online in the next two years.

137 https://docs.alcf.anl.gov/data-management/filesystem-and-storage/file-systems

7.2.2.5 AI and ML

To meet the growing demand for AI and ML components in user jobs, ALCF provides environments that include key AI software packages, including PyTorch, TensorFlow, Horovod, scikit-learn, and others. The facility maintains active relationships with vendors to support high performance with AI libraries on their hardware, including significant work during the Aurora procurement to ensure high performance on Aurora.

AI and ML have also become prominent in ALCF training offerings, including the Argonne Training Program on Extreme-Scale Computing (ATPESC) program, training throughout the year that blends AI/ML and simulation, and dedicated AI/ML workshops. ALCF also offers programs to engage graduate and undergraduate students.

ALCF support of AI pervades the local systems, including the AI testbed systems.

Present-2 years (current budget horizon)

- Aurora supercomputer
- Polaris supercomputer
- Sophia
- ALCF AI testbed
  - Cerebras
  - SambaNova
  - Groq
  - Graphcore
  - Habana
  - Tenstorrent
- Inference Clusters
  - Nvidia GH200 based
  - SambaNova SN40L

Next two to five years (current technology horizon)

- Aurora supercomputer
- Existing ALCF AI testbeds and Inference Clusters
- Continued investments in new AI testbed systems
- Continued investments in new Inference and Fine-tuning clusters

Beyond five years (strategic planning)

- ALCF 4
  - 100EF+ for reduced precision
- Continued investment in AI testbeds
- Continued investment in Inference and Finetuning clusters

7.2.2.6 Wireless Capabilities

There are no plans for the use of wireless.

7.2.2.7 Data-Related Resource Constraints

The ALCF monitors capacity and performance of storage systems and upgrade them as needed. As noted above, an upgrade is targeted for FY25.

7.2.2.8 Data Mobility Endpoints

ALCF operates multiple Globus endpoints for data movement in general. To enable sharing of data more publicly, ALCF has also deployed Globus Sharing on the Eagle filesystem. This functionality allows project owners to share subsets of their data directly from the filesystem, with named users, users in defined Globus groups, all Globus users, or publicly via https.

Current Globus transfer endpoints deployed at ALCF include the following:

- alcf#dtn_home for accessing /home (i.e., home directories on swift-home filesystem)
- alcf#dtn_grand for accessing /lus/grand/projects or /grand (.i.e., project directories on Grand filesystem)
- alcf#dtn_eagle for accessing /lus/eagle/projects or /eagle (.i.e., project directories on Eagle filesystem)
- alcf#dtn_hpss for HPSS

7.2.2.9 Outstanding Issues

ALCF has deployed a variety of services to support the IRI initiative, under the Argonne Nexus effort:

- Service accounts to support long-running experimental campaigns that integrate ALCF computing at experiment time. For example, in the same way that a user can obtain an account at ALCF, a light source beamline can have a dedicated account.
- On-demand queue for immediate access to compute nodes at ALCF. This queue is intended for near real-time computing, frequently required by experimental facilities but also useful in other time-sensitive scenarios.
- Globus Sharing allows ALCF users to share data directly to named collaborators with an ALCF account, or publicly via Globus to users without an ALCF account.

7.2.2.10 Technology Profile Contributors

ALCF Representation

William Allcock, ALCF, allcock@anl.gov

Tom Uram, ALCF, turam@anl.gov

Venkat Vishwanath, ALCF, venkat@anl.gov

7.3 OLCF

The OLCF is charged with helping researchers solve some of the world's most challenging scientific problems with a combination of world-class HPC resources and world-class expertise in scientific computing.

As a US DOE SC User Facility, the OLCF offers leadership-class computing resources to researchers from government, academia, and industry who have many of the largest computing problems in science. Every year, researchers use OLCF resources to achieve breakthroughs ranging from understanding the smallest building blocks of matter to how supernovae spread elements across the universe. The OLCF supports research in biology, chemistry, seismology, engineering, energy, and many other fields. OLCF simulations have improved the safety and performance of nuclear power plants, turbomachinery, and aircraft; accelerated development of new drugs and advanced materials; and informed design of an international fusion reactor. The simulations have explored hurricanes, biofuels, neurodegenerative diseases, and clean combustion for power and propulsion. Research challenges remain, and OLCF resources can help.

7.3.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- OLCF Frontier has been in operation since its debut as the first exaflop supercomputer on May 2022. Its predecessor, Summit, was decommissioned in November 2024, as OLCF prepares for the commissioning of the next HPC system, OLCF-6 and titled "Discovery." The next system is expected in the 2027-2028 time frame. [Section 4.1]
- To support Frontier, OLCF-6, and DOE programs and initiatives such as IRI and FASST, the OLCF has developed application, computing technology, and infrastructure pilots under the Advanced Computing Ecosystem (ACE) strategic initiative. ACE is a strategic computing framework to support domain science efforts by providing infrastructure to enhance research capabilities. [Section 4.1]
- ORNL has two 400 Gbit connections to ESnet and OLCF will have two 400 Gbps connections to the ITSD of ORNL. Additionally, there is a 100 Gbps dedicated WAN connection between Defiant, an HPC cluster on the ACE testbed, and ESnet testbed for IRI activities. [Section 4.7]
- OLCF anticipates more integration with outside clouds for certain applications. OLCF has been testing integration with things such as weights and biases; this is an ML/DL analytics dashboard. The heavy computations happen on Frontier, but the analytics dashboard may be onsite or off-site and integrated through API mechanisms. OLCF is surveying cloud resource performance for HPC workloads, particularly for large-scale leadership-class systems.[138] Current constraints around data in the cloud play a large role in limiting its use, particularly cost and performance of tightly-coupled parallel calculations. Having said that, with IRI, HPDF and FASST, data sources and destinations are expected to become more distributed. [Section 4.8]
- Strategic planning (five+ years) for OLCF shows LAN traffic will be increased to 800 GbE for Science DMZ and data-movement networks with individual 1.6 Tbit links for network interconnects. This will also imply that DMZ traffic (e.g., related to data-movement) will scale to 6–12Tbit/sec. All other LAN traffic will have multiple 800 Gbps connections. Firewalls in this time frame will be 4–6 Tbit/second in throughput. WAN capabilities at ORNL could have multiple 800 Gbps connections to ESnet, and OLCF could have multiple 800 Gbps connections to ITSD. The connections between the ESnet testbed and OLCF resources could be further expanded to multiple 400 Gbps interfaces. [Section 5.7]

138 Evaluating the Cloud for Capability Class Leadership Workloads" https://www.osti.gov/biblio/2000306

7.3.2 OLCF: Technology Profile

The OLCF is a multi-enclave facility operated by ORNL that supports researchers solving the most challenging scientific problems. OLCF has a long history in delivering, operating, and conducting research on world-leading supercomputers. The most recent being Frontier, the world's first exascale system[139]. The multi-enclave nature of the OLCF provides various levels of security, policies, and capabilities to researchers with diverse computing needs. These enclaves include open data, useful for domain-specific data portal development, supercomputing (or medium enclave) which provides the leadership-class computing environment, protected data enclave for ensuring a secure and protected computing infrastructure, and the testbed enclave, that supports exploring and developing future capabilities, workflows, interfaces, and technology and fits into the US DOE IRI program landscape.

7.3.2.1 Facilities Operated and Supported

The foremost system operated at OLCF, Frontier, debuted in 2022 and is based on HPE Cray's EX architecture and slingshot interconnect with optimized 3rd Gen AMD EPYC CPUs and AMD Instinct 250X GPUs. Frontier consists of ~9,856 nodes containing a hybrid node architecture design of 1 CPU to 4 GPU achieving 1.2 exaflop/sec performance. Further description of the architecture and specifications of Frontier are available[140]. Frontier has been accelerating innovation in research and technology and is expected to be operational for the next four to five years.

The predecessor to Frontier, Summit, was decommissioned in November 2024, as OLCF prepares for the commissioning of the next HPC system, OLCF-6 and titled "Discovery." The next system is expected in the 2027-2028 time frame. OLCF-6 is expected to further extend the limit of exascale leadership-class computing and support DOE through increased computational and data science capabilities, interoperate with leadership scale IRI, and be at the forefront in supporting domain scientists and application developers as they explore and integrate transformational AI technologies.

To support Frontier, OLCF-6, and DOE programs and initiatives such as IRI and FASST, the OLCF has developed application, computing technology, and infrastructure pilots under the Advanced Computing Ecosystem (ACE) strategic initiative. ACE is a strategic computing framework to support domain science efforts by providing infrastructure to enhance research capabilities[141]. ACE is the integration of Frontier, Odo (a small-scale HPC training system), and the ACE testbed. ACE supports advancing scientific research through four primary facets, 1) developing and deploying science projects and workflows that leverage IRI, 2) enabling testbeds for a controlled environment for testing and validating new technologies, 3) advancing technologies that support future computing systems, and 4) evaluating new computing strategies and technologies.

Access to the OLCF resources is limited to approved projects through the INCITE program, ALCC, or Director's Discretionary projects. Access to the ACE testbed is provided through a testbed project application and collaboration with OLCF staff. Many resources are made available to OLCF users across the government, academia, and industrial sectors to help facilitate the project's goals and maximize the utilization of computational resources.

Lastly, the OLCF supports quantum computing research efforts, including the Quantum Computing User Program (QCUP). This program is managed by the OLCF and connects researchers with quantum computing resources through ORNL's vendor partners. The OLCF is positioned at the forefront of high-performance computing to support the diverse and evolving scientific landscape at the international stage.

139 S. Abraham et al., "2023 Operational Assessment Oak Ridge Leadership Computing Facility," United States, 2024-04-01 2024. [Online]. Available: https://www.osti.gov/biblio/2378093

140 S. Atchley et al., "Frontier: Exploring Exascale," presented at the Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, CO, USA, 2023, 52.

141 Oral, Sarp, et al. "OLCF's Advanced Computing Ecosystem (ACE): FY24 Efforts for the DOE Integrated Research Infrastructure (IRI) Program." , Nov. 2024. https://doi.org/10.2172/2477506

7.3.2.2 Network and Data Architecture

The OLCF network has redundancy to remove single points of failure and contains various technologies to take advantage of all network paths. These technologies include link aggregation control protocol (LACP), equal cost multipath (ECMP), multi-chassis link aggregation (MLAG), and first hop redundancy protocol. Therefore, losing part of the network only results in loss of bandwidth. Currently, ESnet is working on bringing dark fiber to ORNL. This capability would allow multiplexing technology such that the fiber can be turned into many pairs of fibers dynamically and enables greater flexibility as capacity can be increased quickly when needed without waiting for new fiber to run underground. The networking capabilities at the OLCF are outlined with time frames following current (present to two years), forecasted technology horizon (two to five years), and strategic planning (five+ years). It is challenging to forecast some of the details, as the continued development of the future OLCF-6 system will dictate the network and data architecture.

Current OLCF LAN contains multiple 400 GbE connections for the Science DMZ and data movement. A 1.6 Tbit path is available for moving data between scratch storage (Orion/Wolf) and archive (Kronos/Themis) storage using moderate and open security DTNs and can be scaled up to two to three times that. All other network traffic is managed with multiple 100 GbitE connections. OLCF's firewalls are capable of 1.1 Tbit of aggregate traffic and can carry 100 Gbps streams. In terms of WAN, ORNL has two 400 Gbps connections to ESnet and OLCF will have two 400 Gbps connections to the ITSD of ORNL. Additionally, there is a 100 Gbit dedicated WAN connection between Defiant, an HPC cluster on the ACE testbed, and ESnet testbed for IRI activities.

In the forecasted technology horizon (two to five years), OLCF expects LAN traffic will be 400/800 GbE for Science DMZ and data-movement networks. Data streaming nodes (DSN) could exist on the LAN in this time frame with possibly 200 or 400 Gbit connections. For all other traffic, there will be the availability of multiple 400 Gbps or 800 Gbps connections. It is expected that the firewalls will be in the 3 TB range of aggregate traffic and connected via multiple 400 GbE or possibly 800 GbE interfaces. WAN capabilities at ORNL could be increased to have four or more 400 Gbit connections to ESnet, and OLCF could have four or more 400 Gbit connections to ITSD. The connections between the ESnet testbed and OLCF resources could be expanded to multiple 100 Gbit Interfaces in the next few years to support the IRI program.

Strategic planning (five+ years) for OLCF is expecting LAN traffic will be increased to 800 GbE for Science DMZ and data movement networks with individual 1.6 Tbit links for network interconnects. This will DMZ data-movement traffic to scale 6–12 Tbit/sec. All other LAN traffic will have multiple 800 Gbit connections. Firewalls in this time frame will be 4–6 Tbit/second in throughput. WAN capabilities at ORNL could have multiple 800 Gbit connections to ESnet and OLCF could have multiple 800 Gbit connections to ITSD. The connections between the ES.net testbed and OLCF resources could be further expanded to multiple 400 Gbit interfaces.

7.3.2.3 Computational and Storage Infrastructure

OLCF maintains several computing clusters with a variety of resources available to OLCF users and projects. The existing computing clusters are outlined in Table 1. The storage infrastructure at OLCF contains two components, a node-local storage and a center-wide parallel file system using Lustre (named Orion[142]). The node-local storage is user-managed and primarily used for storing reads and writes for various compute job types. The Orion system has a capacity of ~700 PB and allows a tiered archival system (disk + tape) with a distribution of flash performance, hard disk-based capacity and metadata servers.

142 https://docs.olcf.ornl.gov/data/index.html#orion-lustre-hpe-clusterstor-filesystem

Cluster	Brief Description
FRONTIER	Exascale leadership-class compute system with AMD EPYC CPUs and AMD Instinct GPUs.
ANDES	704-node commodity Linux cluster where each node contains two 16-core AMD EPYC 7302 processors with 256 GB memory.
DEFIANT	20 CPU-only nodes each containing two 32-core Intel Xeon Platinum CPUs and 2 GPU nodes each with 8 NVIDIA H200 GPUs. Future planned upgrades will include the addition of AMD GPU nodes.
WOMBAT	Heterogeneous AArch64 cluster featuring CPUs like Fujitsu A64FX, Ampere Computing Altra, Nvidia Grace, and Nvidia GPUs with InfiniBand networking.
HOLLY	Single Supermicro server equipped with 8 Nvidia H100 GPUs and 8 HDR IB network connections (1.6 tbps aggregate bandwidth).
QUOKKA	16 general-purpose nodes with Intel Xeon CPUs connected via NDR200 InfiniBand.
SLATE	OpenShift Kubernetes platform cluster for long-running or internet-facing services.

Table 7.3.1: Overview of the existing computational infrastructure available at OLCF.

Other existing software, workflow, and interfaces available at OLCF are listed below.

- CITADEL workflow system for enabling protected compute environments on Frontier (enabling PHI/ITAR data processing).
- Globus Framework (Transfer, Flows, and Compute).
- Secure Scientific Service Mesh (S3M) that enables a Facility API in the ACE testbed enclave. This is a forward-looking capability for IRI-related workflow integration.
- Slurm resource manager.
- OpenShift k8s-based environment for hosting databases and certain subcomponents of a workflow system that ties to the clusters.
- The OLCF in partnership with the QCUP facilitates access to an array of vendor hosted quantum computing platforms. This has enabled a pipeline for user onboarding, training, and workflow integration.

In summary, OLCF provides diverse computational and storage infrastructure across enclaves. In the Moderate enclave, the supercomputing platform (Frontier), a companion commodity Linux cluster for post/pre-processing (Andes), an additional k8s-based companion system, a parallel filesystem for supercomputing I/O, and tiered storage for archival and other non-supercomputing I/O are all made available. In the open enclave, OLCF primarily focuses on data-centric capabilities to enable science portals for the scientific domains to openly share data with their community. Additionally, OLCF hosts a testbed system for future looking capabilities that face IRI and the HPDF. This will help understand the larger operational ecosystem that must be developed in the two to five years and beyond.

For the future, during the next two to five years, OLCF is evaluating how this supercomputing platform will evolve. For instance, the next system, may see more emphasis on HBM on the computing platform. Additionally, there is interest in accelerated AI and analytics capabilities integrated into the platform. This could lead to a new AI-focused software stack and hardware.

Going beyond five years, the heterogeneity of the system will potentially continue to increase. Strategically speaking there is potential for tiered computational resources to facilitate diverse workload requirements as the leadership-class computing environment evolves. Especially, as the future of IRI and HPDF are evaluated. This will require new operational software capabilities to securely and optimally integrate distributed science systems.

Furthermore, OLCF could see quantum integration becoming more coupled with traditional HPC. These could be quantum resources hosted at the facility, within the larger leadership ecosystem at OLCF, with specialized software gluing capabilities together. [143]

7.3.2.4 Cloud Services

In the present- to two-year time frame OLCF provides access to an on-premise cloud resource. This is the OpenShift k8s-based resource called "Slate." OLCF projects are able to develop and deploy cloud-native applications on this resource. OLCF-staff also develop and deploy services for users on this system. Currently, OLCF provides JupyterHub as a service from this resource to enable a python portal for workflows. It has been possible to integrate the Dask workflow system from this JupyterHub, which has clients on supercomputing resources, in order to drive workflows across HPC systems.

OLCF provides cloud-based access to external quantum resources in the current time frame. This is through the QCUP. OLCF is the centralized onboarding point, which then enables access to an array of back-end vendor-provided quantum resources hosted outside of OLCF.

Going beyond the two-year time frame, more integration with the outside clouds for certain applications is possible. OLCF has been testing integration with things such as weights and biases for instance. This is an ML/DL analytics dashboard. The heavy computations happen on Frontier, but the analytics dashboard is off-site and integrated through API mechanisms.

Going beyond five years, it is possible things get even more distributed. Especially with AI. OLCF is surveying cloud resource performance for HPC workloads, particularly for large-scale leadership-class systems. Current constraints around data in the cloud play a large role in limiting its use, particularly cost and performance of tightly-coupled parallel calculations. Having said that, with IRI, HPDF and FASST, data sources and destinations are expected to become more distributed.

7.3.2.5 AI and ML

There are several options for various kinds of ML, deep learning, and data analytics tasks on OLCF systems.

- PyTorch on Frontier: PyTorch is a library for Python programs that pairs well with HPC resources and facilitates building DL projects. PyTorch emphasizes the flexibility and human-readableness of Python and allows deep learning models to be expressed in a similar manner. Compared to other frameworks and libraries, it is one of the more "beginner friendly" ML/DL packages due to its dynamic and familiar "Pythonic" nature. PyTorch is also useful when GPUs are involved because of its strong GPU acceleration ability. On Frontier, PyTorch is able to take advantage of the many AMD GPUs available on the system.
- JAX: a Python library for accelerator-oriented array computation and program transformation, designed for high-performance numerical computing and large-scale ML. JAX provides a familiar NumPy-style API that can execute on either CPUs, GPUs, or TPUs.

7.3.2.6 Wireless Capabilities

No known wireless capabilities currently or planned for OLCF resources.

143 T. Beck et al., "Integrating Quantum Computing Resources into Scientific HPC Ecosystems," Future Generation Computer Systems, 161, 11–25. https://doi.org/10.1016/j.future.2024.06.058

7.3.2.7 Data-Related Resource Constraints

Most constraints revolve around open and integrated access. Science projects funded by DOE science programs (BER, BES, HEP, etc..) in the DOE seem to be moving towards web-based applications, distributed across user facilities and cloud. How the data and analysis pipelines integrate with ASCR HPC facilities is certainly a challenge, from a security and automation perspective. Policies are a constraint limiting integration between web-based applications and ASCR HPC facilities on the back end.

In the present to two years at OLCF, the openShift k8s-based resource called "Slate" will be operational. Currently, most ASCR user facilities provide a peripheral resource like this for either workflows or facility-hosted web applications. This infrastructure is very useful for enabling data sharing capabilities. This will be used in the Open enclave to enable domain-specific data portals for projects that live at OLCF. Challenges still exist when projects want to host data at OLCF and share that data publicly.

In the next two to five years, OLCF anticipates more distributed and seamless access to data across the DOE landscape. This revolves around IRI. Additionally, there needs to be more integrated and automated access to leadership computing resources from within data portal applications, so users can search data and launch computational tasks seamlessly. The challenges of building trust for these integrations and workflows between the cloud and HPC user facilities will hopefully be more well established in this time frame. This will require supplemental interfaces and policies to enable this integrated future.

7.3.2.8 Data Mobility Endpoints

DTNs are hosts specifically designed to provide optimized data transfer between OLCF systems and systems outside of the OLCF network. These nodes perform well on local-area transfers as well as the wide-area data transfers for which they are tuned. The OLCF recommends that users use these nodes to improve transfer speed and reduce load on computational systems' login and service nodes. OLCF provides two sets of DTNs: one for systems in the moderate enclave and a second for systems in the open enclave. More information about DTNs can be found online.[144]

7.3.2.8.1 Interactive Access

DTN access is automatically granted to all enabled OLCF users. For interactive access to DTNs (ssh/scp/sftp), connect to dtn.ccs.ornl.gov (for the moderate enclave) or opendtn.ccs.ornl.gov (for the open enclave).

7.3.2.8.2 Access From Globus Online

DTNs are also accessible via the "OLCF DTN" (for moderate) and "NCCS Open DTN" (for open) Globus endpoints.

7.3.2.9 Outstanding Issues

No issues identified.

7.3.2.10 Technology Profile Contributors

OLCF Representation

Travis Croxdale, OLCF, croxdaletk@ornl.gov

Brian Etz, OLCF, etzbd@ornl.gov

Rafael Ferreira Da Silva, OLCF, silvarf@ornl.gov

Daniel Pelfrey, OLCF, dspelfrey@ornl.gov

Ryan Prout, OLCF, proutrc@ornl.gov

144 https://docs.olcf.ornl.gov/systems/dtn_user_guide.html

7.4 Fermilab

Fermilab is America's particle physics and accelerator laboratory. We bring the world together to solve the mysteries of matter, energy, space and time. Since 1967, Fermilab has worked to answer these and other fundamental questions and enhance our understanding of everything we see around us. As the United States' premier particle physics laboratory, we do science that matters. We work on the world's most advanced particle accelerators and dig down to the smallest building blocks of matter. We also probe the farthest reaches of the universe, seeking out the nature of dark matter and dark energy.

7.4.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- Fermilab is the DOE's particle physics and accelerator laboratory. It is a single purpose laboratory, with a mission to support DoE-sponsored HEP experiments and activities. Fermilab is the US Tier-1 center for the LHC/CMS experiment, hosting roughly 40% of CMS experiment data. The Laboratory is also the host institution for the DUNE experiment and supports the Tier-0 center for the experiment. Other active Fermilab-based experiments include SB/ND, NOvA, Mu2e, Muon g-2, ICARUS, and MicroBoone. [Section 4.1]
- The WLCG recommendation for large Tier-1s, such as Fermilab, at HL-LHC startup is a minimum of 1Tb/s. Fermilab is currently provisioned at 1.6 Tbps for science traffic and 800 Gbps for general-purpose traffic. The latter does include the fraction of the Laboratory's science that is not carried via the science networks. It is expected that this configuration should be adequate to satisfy the currently known requirements for 2024- 2028. If needed, additional 400G connections can be added to the existing LAG interfaces. It is expected that switching to 800G Ethernet technology might become justifiable in 2028–2030. [Section 4.7]
- In terms of aggregate WAN capacity out of the site, Fermilab currently has provisioned 2.4Tb/s of link capacity to ESnet, via geographically diverse connections. A LAG of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports the Laboratory's general internet traffic, which includes some portion of science traffic, for example, CMS traffic to/from sites not connected to LHCONE. [Section 4.7]
- The DUNE experiment extends across two facilities ~1300km apart. One site is Fermilab in Batavia, IL, and the second site, the FD location, is SURF, in Lead, South Dakota. The Fermilab-managed perimeter network will provide connectivity for the DUNE FD site to Fermilab. Sufficient capacity to move ~30PB of science data annually is required. The perimeter network equipment will reside in two geographically diverse buildings at SURF — the Ross Dry Building and the Yates Administrivia Building. Each building has a shaft which provides access to the underground facilities at the 4850ft level (4850L). WAN connectivity is to be provided and managed by ESnet. [Section 4.7]
- ESnet and Fermilab network engineering teams have drafted a plan for connectivity back to Fermilab to support DUNE. The plan calls for redundant, geographically diverse, vendor-diverse, 100 Gbps circuits from SURF to ESnet PoPs in Chicago and Denver, respectively. ESnet backbone services (OSCARS) would then extend those L2 paths back to Fermilab. Deployment of both 100 Gbps circuits should occur by the 2026–2027 time frame. Both pathways will be connected to ESnet in a production capacity. Currently, a dedicated 10 Gbps circuit between SURF and ESnet's Kansas City PoP provides DUNE FD connectivity back to the Fermilab campus. This circuit, provided by South Dakota's REED, will revert to a tertiary back-up role once the two 100 Gbps circuits are procured and deployed. [Section 5.7]

- Fermilab is concerned about not having the capability to adequately stress test the WAN infrastructure as network technology step-increases push bandwidth capacities to terabit scale. The ability to generate traffic at 400GE, 800GE, & 1.6TE is challenging in its own right. Fermilab desires to be able to routinely stress test the WAN infrastructure to ensure it is performing as expected. The complexity of having multiple types of virtual network services sharing the same physical infrastructure adds another level of difficulty to this. It is desirable to consider this as a capability to concurrently generate high traffic loads across various virtual network services (LHCOPN, LHCONE, default routed IP path, etc.). This capability is a critical tool in Fermilab's ability to troubleshoot aggregate data bottlenecks where multiple virtual network services are involved. Fermilab proposes that this should be developed as a service that ESnet provides to all its sites. [Section 5.7]
- Fermilab has a strong interest in making use of ESnet High Touch Services to generate data for AI/ML-based analysis of network traffic patterns. Fermilab is currently working cooperatively with ESnet on distinguishing between streaming-based CMS data movement and file-transfer-based CMS data movement. It is desirable not just to continue with this particular effort, but also to look toward additional opportunities to expand use of ESnet High Touch Service capabilities in analyzing WAN traffic. [Section 5.8]

7.4.2 Fermilab: Technology Profile

Fermilab is the DOE's particle physics and accelerator laboratory. It is a single purpose laboratory, with a mission to support DoE-sponsored HEP experiments and activities. Fermilab is the US Tier-1 center for the LHC/CMS experiment, hosting roughly 40% of CMS experiment data. The Laboratory is also the host institution for the DUNE experiment and supports the Tier-0 center for the experiment. Other active Fermilab-based experiments include SB/ND, NOvA, Mu2e, Muon g-2, ICARUS, and MicroBoone. In addition, Fermilab hosts one of the elements of the DOE's national Lattice QCD computational infrastructure. However, the Lattice QCD cluster is standalone infrastructure with no significant Fermilab WAN requirements and is not factored into this case study.

7.4.2.1 Facilities Operated and Supported

The network requirements for Fermilab experiments with significant WAN data movement can be found within the specific review case study applicable to each experiment (see table below). The experiments are listed in rough order of their expected experiment data volume.

Experiment	2024 ESnet HEP Requirements Case Study (Studies)
LHC/CMS (US TIER-1)	Section 6.10: LHC: CMS Experiment Section 6.11: LHC: LHC Operations Section 6.12: LHC: HL-LHC & Network Research
DUNE/FD (TIER-0)	Section 6.8: Neutrino Experiments at Fermilab: SBN and DUNE
SB/ND	Section 6.8: Neutrino Experiments at Fermilab: SBN and DUNE
NOVA	Section 7.4.2.2.5
MUON G-2	Section 6.6: Muon Experimentation at Fermilab: Muon g-2 and Mu2e
MU2E	Section 6.6: Muon Experimentation at Fermilab: Muon g-2 and Mu2e
MICROBOONE	<local experiment w/ no significant WAN requirements>
ICARUS	<local experiment w/ no significant WAN requirements>

Table 7.4.1: Fermilab HEP Support

The current schedule for active Fermilab experiments is shown in Figure 7.4.1.

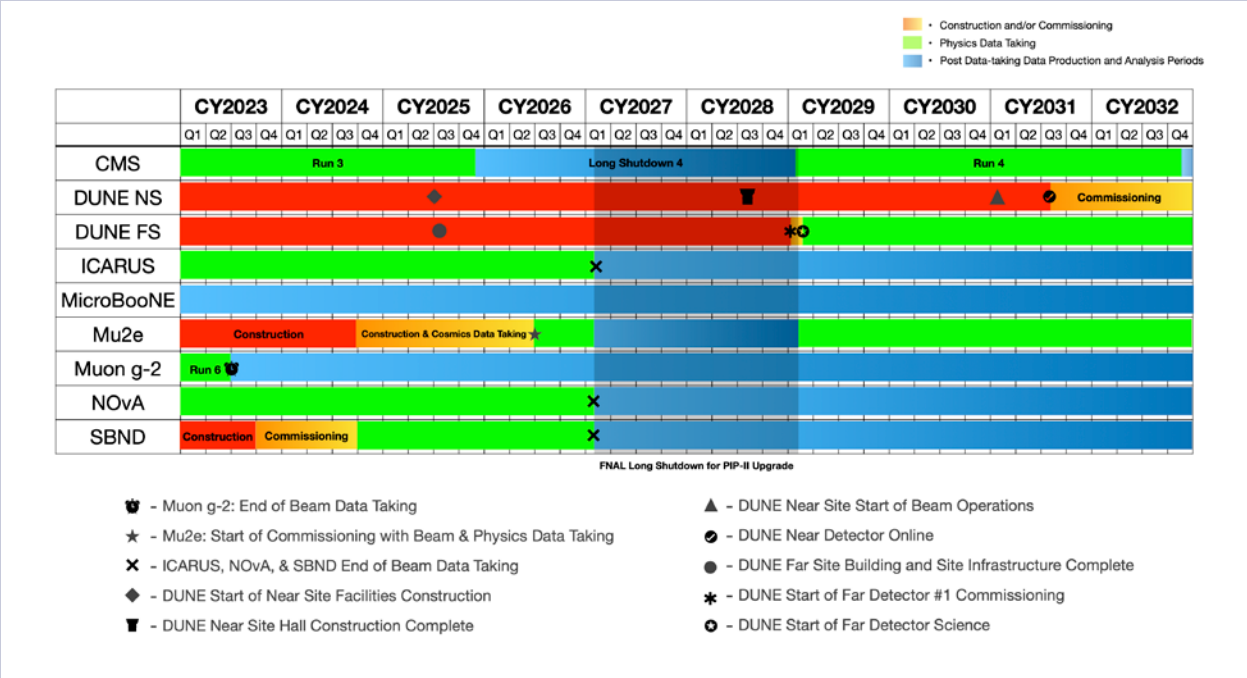


Figure 7.4.1: Fermilab Experiments Schedule 2022-2032

Computing resources to support these experiments are divided into two distinct entities, the US CMS Tier-1 facility for CMS computation, and a shared computation infrastructure for all the other experiments. The significantly higher computing, storage, and networking requirements for CMS computation are more easily and economically met with separate infrastructure than attempting to scale all Fermilab experiment computing infrastructure up to meet CMS's needs. As a result, CMS traffic into and out of Fermilab is relatively easy to identify, based on network addresses used. The rest of Fermilab's experiments share computing, storage, and network resources. Identifying specifically what WAN traffic belongs to which of those experiments presents a challenge.

Computing resources, whether part of the US CMS Tier-1 facility or the shared use infrastructure, are distributed across two data centers, the Feynman Computing Center (FCC) and the Grid Computing Center (GCC). Each computer center supports multiple computer rooms. The two data centers are approximately 2km apart, and heavily interconnected (>1000pr) with geographically diverse optical fiber. FCC has both uninterruptible power supply (UPS) and back-up generator support, while GCC has only limited UPS deployment and generator support. Storage nodes and other sensitive servers are normally located in FCC, while CPU worker nodes are normally located in GCC. This holds true for both the US CMS Tier-1 and the shared computing infrastructure for the rest of the Laboratory's experiments. The end result is virtual work group LANs that straddle both FCC and GCC data centers for both the US CMS Tier-1 facility and the shared computing infrastructure used by the rest of the Fermilab-supported experiments.

Future Computing Facility Upgrades/Expansion

While there has been discussion and very preliminary planning for a next-generation data center at Fermilab to meet its future local computation needs, there is nothing currently in place with a funding or construction roadmap. The existing FCC and GCC data centers are expected to accommodate planned expansion of local computing resources to meet HL-LHC and DUNE/FD startup requirements up to 2029-2030.

7.4.2.2 Network and Data Architecture

The Laboratory's network architecture is organized into a modular model. Each module consists of network resources dedicated to specific tasks and can be represented by logical groupings of subnets. These modules are typically structured based on organizational divisions, experiment affiliations, or, in some cases, geographical locations. Examples of network modules include:

- Fermilab's network perimeter
- Data center network spanning multiple buildings
- US CMS-Tier-1 network
- DUNE network architecture
- NOvA far detector network

All network modules connect to a central component known as the Site Interconnect, which facilitates connectivity between modules and enforces inter-module security policies. Modules can also have direct connections to one another, enabling certain traffic to bypass the Site Interconnect, provided that adequate security controls are in place. For instance, the US CMS-Tier-1 module has a direct link to the Perimeter module for high-impact data transfers, bypassing the Site Interconnect. However, general-purpose traffic from the US CMS-Tier-1 module, such as email and web traffic, still must pass through the Site Interconnect.

Network modules may be protected by dedicated firewalls at the module boundary to address specific security needs. The Network Perimeter module includes border firewalls that inspect general-purpose traffic before the traffic reaches the Site Interconnect. High-impact scientific data traffic may bypass these border firewalls to reach its destination module directly. Additionally, computing resources in commercial clouds can be integrated into Fermilab's network, functioning as an extension of the network modules connected to either the Site Interconnect or the Perimeter module. The conceptual network architecture is depicted in the figure below.

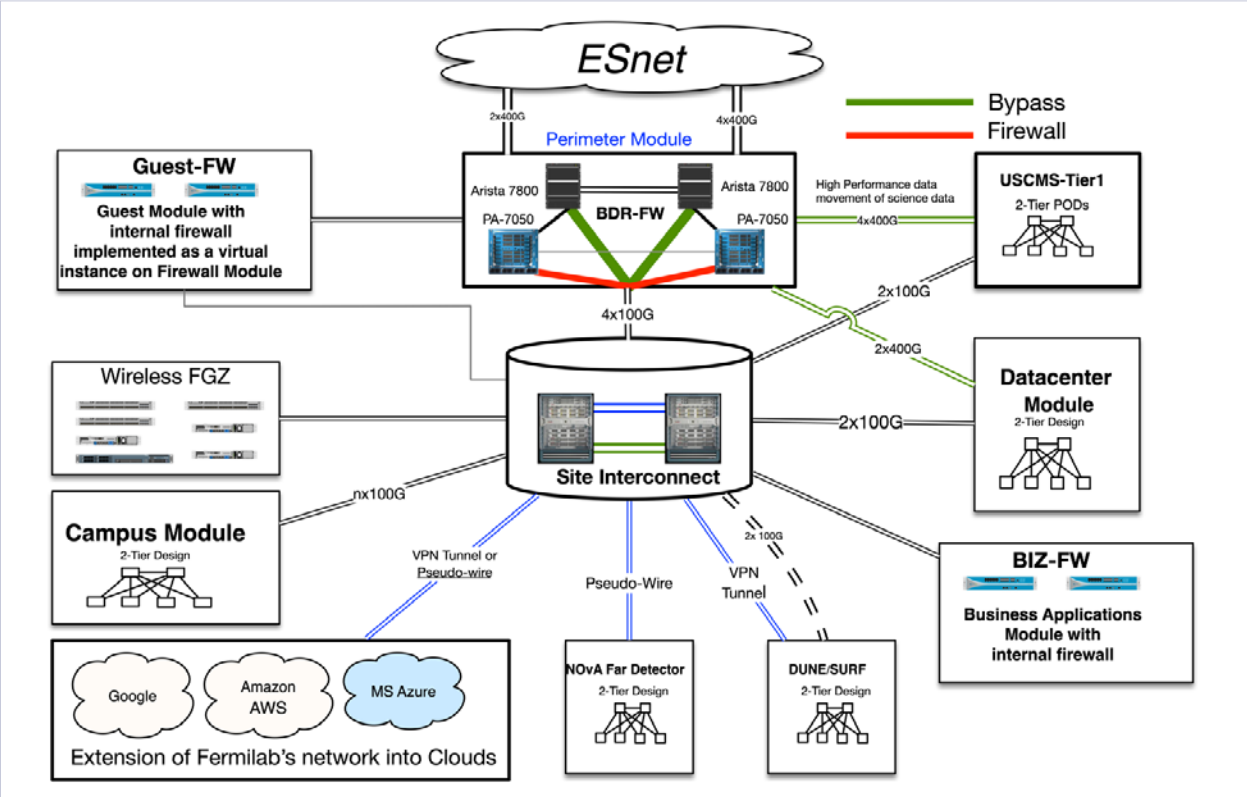


Figure 7.4.2: Fermilab Network Diagram

7.4.2.2.1 Network Perimeter

Currently, the Network Perimeter consists of:

- Two (2) Arista 7800 routers, each consisting of:
  - 2x 48-port 100G LC, 2x 36-port 400G LC
  - Deep buffers, 800G enabled
  - 2x 400G and 4x 400G off-site connections to ESnet at the geographically diverse GCC and FCC data centers respectively
- Two Palo Alto PA-7050 with 100G line cards, at the geographically diverse GCC and FCC data centers respectively
- DMZ Arista switches (not shown in diagram) with 10/100/400G ports

Fermilab off-site traffic is asymmetric by design, with science networks (LHCOPN, LHCONE, E2E circuits, etc.) traffic using one ESnet connection (FCC), and general routed internet traffic using the other ESnet connection (GCC). This inherently creates a situation for one (FCC) ESnet connection to require significantly larger bandwidth capacity than the other. If one connection fails, traffic will automatically switch to the other connection. This leads to a potential bottleneck if all scientific traffic shifts to the general IP path where lower bandwidth is provisioned. However, based on the previous experience, Fermilab can tolerate temporary congestion on one link. Additionally, with 800G of provisioned bandwidth on the lower capacity WAN link (GCC), there is currently ample capacity to manage this type of situation.

The WLCG recommendation for large Tier-1s at HL-LHC startup is 1Tb/s, with the largest Tier-1s (Fermilab specifically noted) needing significantly more than 1Tb/s. Fermilab is currently provisioned at 1.6 Tbps for science traffic and 800 Gbps for general-purpose traffic. The latter does include the fraction of the Laboratory's science that is not carried via the science networks. It is expected that this configuration should be adequate to satisfy the currently known requirements for 2024- 2028. If needed, additional 400G connections can be added to the existing LAG interfaces. It is expected that switching to 800G Ethernet technology might become justifiable in 2028–2030.

7.4.2.2.2 General Datacenter for Multiple Experiments

General data center serves multiple experiments (SBN, NOvA, G-2, Mu2E and others) with modest bandwidth requirements, and which utilize common computing and storage facilities such as FermiGrid, public dCache and public EnStore. The networks used in these experiments are typically directly connected to the general data center to streamline the path for collecting, commuting, and storing experimental data. This setup is illustrated in the diagram below for the SBN-FD/ND network. The general data center is built as the following:

- Three distribution switches in computer rooms in different buildings ~2km apart:
  - 400GE technology deployed, including nx400GE LAGs between buildings
- ToR access switches dual-homed to two distribution switches:
  - Large-scale deployment of 100GE host connections
  - Abundant of fiber utilized in dual-homing between computer rooms
  - Default 10GbaseT connection for all worker servers' (1350 connections)
  - 200 storage servers with 100G connections
  - 44x 1G connections
  - 92 / 20 mix of 100/400G uplinks respectively

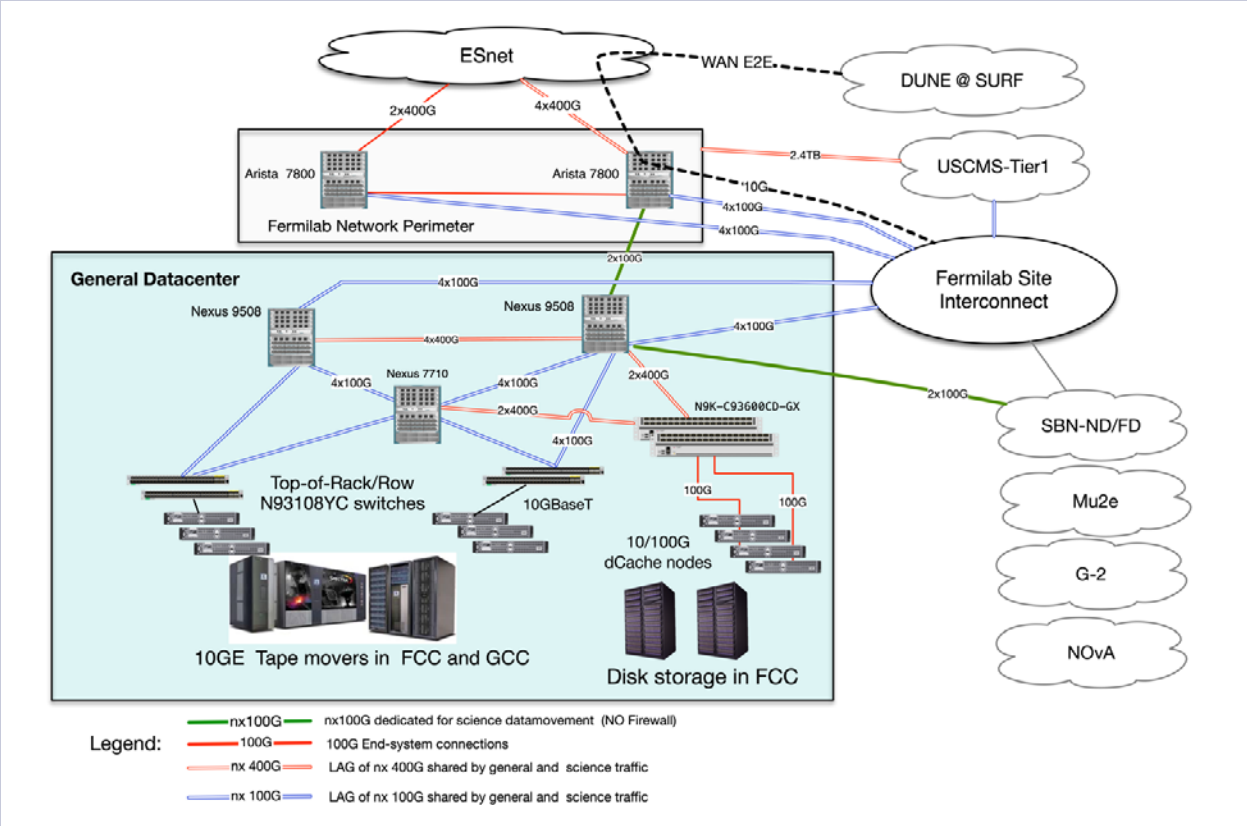


Figure 7.4.3: Fermilab Datacenter Architecture

7.4.2.2.3 The US CMS-Tier-1 Network

The US CMS Tier-1 Center at Fermilab is the largest of CMS Tier-1 facilities, currently providing 40% of the total CMS Tier-1 capacity. The Fermilab Tier-1 consists of dedicated computing infrastructure. For the details on computing and storage requirements, please see Section 6.11.

The US CMS Tier-1's local network infrastructure consists of fully redundant, high-performance switching fabric distributed across the Laboratory's two data centers. That switching fabric has the following characteristics:

- Currently based on 400GE network technology for inter-switch connectivity, with 10GE and 100GE connectivity available for host system connectivity.
- Extensive use of link aggregation (LAG) to scale bandwidth for inter-switch connectivity. At the present time, the US CMS Tier-1 LAN provides approximately 10Tb/s of network capacity in total.
- PBR techniques to route CMS high volume traffic over special purpose networks (such as LHCOPN, LHCONE) utilized for WAN data movement.

Fermilab's WAN architecture is based on separating its high-impact science data traffic from its general internet traffic. Conceptually, this design is analogous to a Science DMZ architecture. Most traffic into and out of the US CMS T1 is via the science data path(s). For CMS, those science data paths mean the LHCOPN, LHCONE, and end-to-end OSCARS circuits. Fermilab's LHCOPN connectivity supports movement of raw data from the T0 (CERN), as well as production data movement with some of the other CMS T1s. Fermilab's LHCOPN configuration consists of three OSCARS circuits (primary, secondary, and tertiary) to CERN, which provide levels of redundancy with differing bandwidth guarantees for that traffic. LHCONE supports production data movement between the Fermilab Tier-1 and most CMS Tier-2s, as well as CMS T1s that do not use LHCOPN for

T1-T1 data movement. Like the LHCOPN, connectivity to LHCONE is via geographically redundant (primary/failover) paths. US CMS Tier-1 WAN traffic that does not utilize either the LHCOPN or LHCONE paths traverses the Laboratory's general internet path.

In terms of aggregate WAN capacity out of the site, Fermilab currently has provisioned 2.4Tb/s of link capacity to ESnet, via geographically diverse connections. A LAG of four 400 Gbps links is used to support the science data network paths, including LHCOPN and LHCONE. A second 2 x 400 Gbps LAG supports the Laboratory's general internet traffic, which includes some portion of science traffic, for example, CMS traffic to/from sites not connected to LHCONE.

Future network enhancements:

- Upgrade of the US CMS Tier-1 LAN infrastructure to 800GE technology is expected for 2028–2030 time frame;
- Upgrade of the Laboratory network perimeter infrastructure to 800GE technology is expected also in 2028 - 2030 time frame, likely to be aligned with availability of 800GE services from ESnet;
- In 2026–2028 time frame, additional 400GE links deployed to WAN capacity. The 2 x 400GE GCC connection is expected to be upgraded to 4 x 400GE.

7.4.2.2.3.1 US CMS-Tier-2 Locations

Tier-1 - Tier-2 production computing is conducted between the US CMS Tier-1 at Fermilab and the following US CMS sites (includes the US CMS Heavy Ion T1 at Vanderbilt Univ. and the CMS Tier-2 sites in Brazil):

US CMS Site	WAN Bandwidth	Regional Provider	LHCONE Peering Pt
CALTECH (T2)	100 Gbps	CENIC	Sunnyvale
FLORIDA (T2)	200 Gbps	FLR	Atlanta (SoX)
MIT (T2)	100 Gbps	NoX	NY (MAN-LAN)
NEBRASKA (T2)	200 Gbps	GPN	Kansas City
PURDUE (T2)	100 Gbps	I-Light	Chicago
UCSD (T2)	100 Gbps	CENIC	Sunnyvale
WISCONSIN (T2)	100 Gbps	BRAA	Chicago
SPRACE (BR T2)	100 Gbps	RedNesp	AmLight (Miami)
UERJ (BR T2)	10 Gbps	RNP	AmLight (Miami)
VANDERBILT (HI T1)	100 Gbps	- - -	Atlanta (SoX)

Table 7.4.2: US CMS T2 Locations

7.4.2.2.4 DUNE Network Architecture

The DUNE experiment extends across two facilities ~1300km apart. One site is Fermilab in Batavia, IL, and the second site, the FD location, is SURF, in Lead, South Dakota. The local network at SURF to support the DUNE FD is made up of several components:

- Perimeter network — Fermi-managed infrastructure to provide connectivity for the DUNE FD site to Fermilab. Sufficient capacity to move ~30PB of science data annually is required. The perimeter network equipment will reside in two geographically diverse buildings at SURF — the Ross Dry Building and the Yates Administrative Building. Each building has a shaft which provides access to the underground facilities at the 4850ft level (4850L). WAN connectivity is to be provided and managed by ESnet.

- Experiment (DAQ/DCS) network - provides network services for the DAQ and the detector control systems. The experiment network moves data from the detectors down at the 4850L in the Ross cavern to the computing room in the Ross Dry main control room on the surface. The Experiment network is designed to have a high level of resilience to meet the DAQ requirements, even in the event of extended failure scenarios such as a complete fiber cut in one shaft.
- Technical network — provides network services for cryogenic control devices, building and computer room controls, and monitoring systems. The technical network extends from the surface to the cavern.
- General-purpose network - provides essential network services such as DNS, DHCP, KDC, and other IT services required at the FD location, both in the cavern and on the surface.
- Network management network - a separate network infrastructure to manage all DUNE networks.

The DUNE FD network at SURF is logically an extension of Fermilab site network. It is designed as a distinct network module within the Fermilab network architecture. The DUNE FD network module connects to the Site Interconnect module at the Fermilab Batavia site via the site-to-site WAN circuits. The high-level network architecture for DUNE is depicted in figure below:

- Logically within the network perimeter of Fermilab.
- Fully redundant, high-performance switching fabric, both down in the cavern and up on the surface.
- Based on 100GE/400GE network technologies for inter-switch connectivity, with redundant 10GE and 100GE connectivity available for host system connectivity.
- Fully redundant and geographically separated fiber connectivity between the cavern and the computing facilities up on the surface in Ross.
- Redundant essential network services (DNS. NTP, DHCP)
- A redundant network perimeter at SURF, with the border devices in both the Ross and Yates buildings.

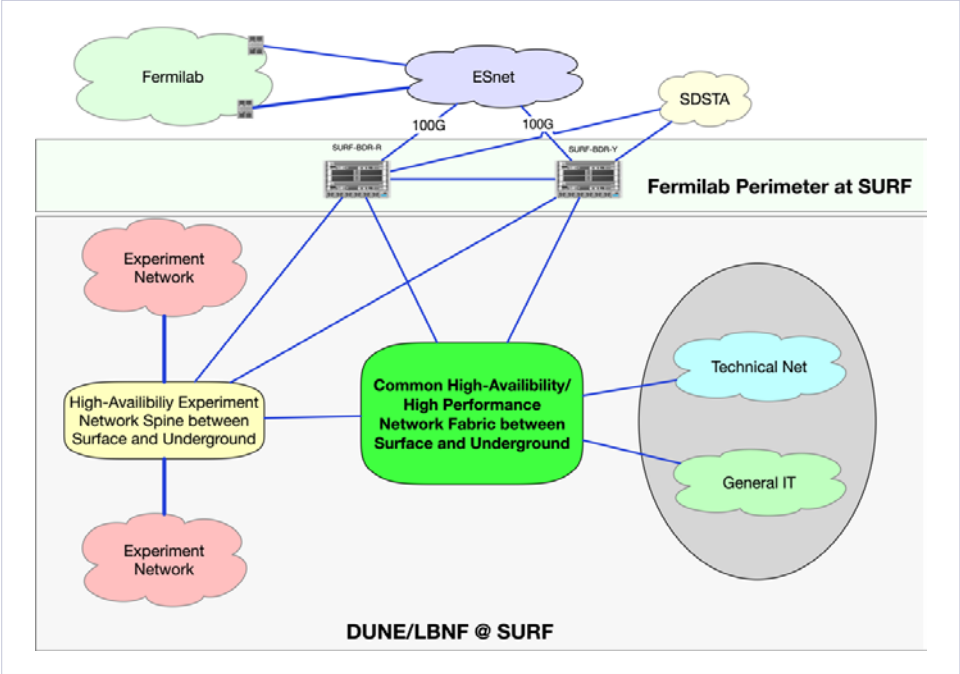


Figure 7.4.4: DUNE Network Architecture

7.4.2.2.4.1 DUNE WAN

Fermilab DUNE WAN connectivity has two components:

- Movement of detector-generated data from SURF to Fermilab for Tier-0 archiving;
- Distribution of experiment-generated data to other DUNE computation sites for production computing.

7.4.2.2.4.2 SURF -> Fermilab Data Movement

ESnet and Fermilab network engineering teams have drafted a plan for connectivity back to Fermilab for the DUNE FD complex. The plan calls for redundant, geographically diverse, vendor-diverse, 100 Gbps circuits from SURF to ESnet PoPs in Chicago and Denver, respectively. ESnet backbone services (OSCARS) would then extend those L2 paths back to Fermilab.

The revised DUNE FD WAN connectivity schedule is to now deploy both 100 Gbps circuits in the 2026–2027 time frame. Both pathways will be connected to ESnet in a production capacity. Currently, a dedicated 10 Gbps circuit between SURF and ESnet's Kansas City PoP provides DUNE FD connectivity back to the Fermilab campus. This circuit, provided by South Dakota's state R&E network (REED), will revert to a tertiary back-up role once the two 100 Gbps circuits are procured and deployed. See the Figure 7.4.5.

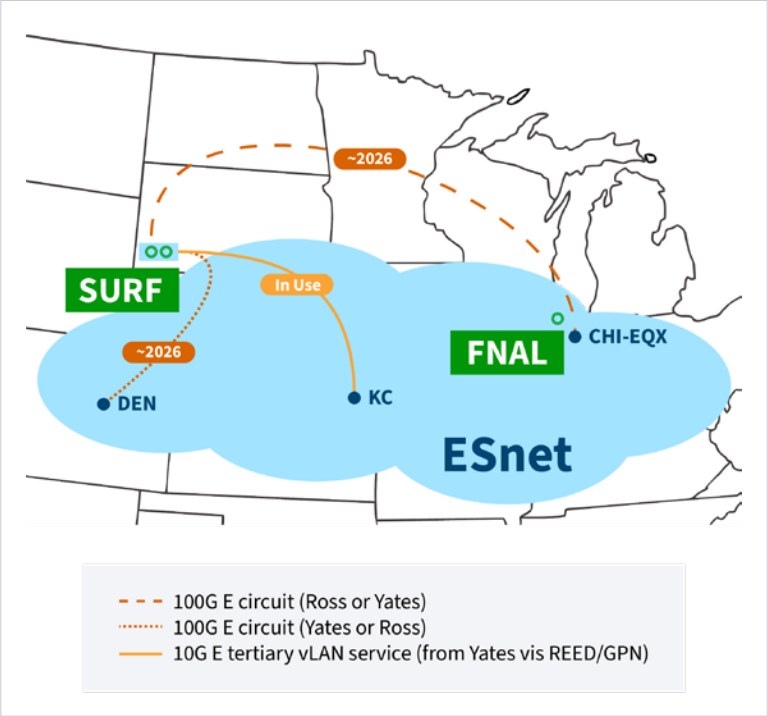


Figure 7.4.5: DUNE WAN Architecture

7.4.2.2.4.3 Fermilab Distribution of DUNE Data To Collaboration Sites

Like the LHC experiments, DUNE will utilize a distributed model for its production computing. Approximately 50% of the production computing will be performed at other DUNE computing facilities, mostly in Europe. In 2023, the DUNE experiment was approved for use of LHCONE for its offline computing. The expectation is that the majority of DUNE offline WAN traffic will traverse LHCONE. The table below lists the RSE locations, where the majority of the remote DUNE production computing is expected to be performed:

Rucio Storage Element	Expected WAN Path	WAN Service Provider
CERN	LHCONE	<many>
PIC	LHCONE	RED-IRIS
IN2P3	LHCONE	RENETAR
FZU	LHCONE	CSNET
NIKHEF	LHCONE	SURF
SURFSARA	LHCONE	SURF
LANCASTER (*)	R&E Paths	JANET
MANCHESTER (*)	R&E Paths	JANET
QMUL (*)	R&E Paths	JANET
RAL_ECHO (*)	LHCONE	JANET
RAL_PP (*)	LHCONE	JANET
BNL	LHCONE	ESNET
NERSC	R&E Paths	ESNET
TIFR	<not commissioned yet>	<not commissioned yet>
JINR	<not commissioned yet>	<not commissioned yet>

Table 7.4.3: DUNE Data Distribution

The DUNE experiment projects to generate approximately 30PB of data per year, roughly the equivalent of CMS's Run-2 data generation.

7.4.2.2.5 NOvA (NuMI Off-axis νe Appearance)

The NOvA experiment utilizes two detectors: a 300-metric-ton near detector at Fermilab and a much larger 14-metric-kiloton far detector at Ash River, Minnesota, just south of the US-Canada border. Although NOvA has surpassed its initial planned schedule, it remains an active experiment, and a decommissioning date has not yet been set.

Currently, ESnet and the University of Minnesota jointly provide two geographically separate 1G OSCARS circuits (Vlan 201 & 203) between Batavia and the NOvA Far Detector, ensuring redundancy across ESnet and partially through University of Minnesota (UMN). From the UMN router in Duluth, these circuits are merged and provisioned through a single physical pipe to Ash River provided by a local Internet service provider (ISP). These connections need to be maintained and kept operational during the lifetime of the experiment. Additionally, the Fermilab public dCache prefix is advertised on LHCONE for the NOvA experiment. This service is primarily used for production NOvA computing performed at FZU (Prague, Institute for Physics of the Czech Academy of Sciences).

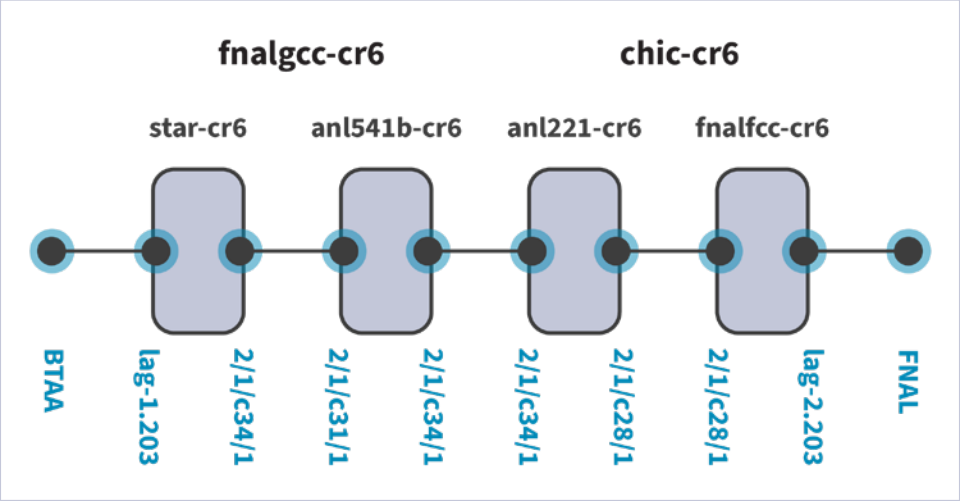


Figure 7.4.6: NOvANetworking

7.4.2.3 Computational and Storage Infrastructure

Current computation and storage needs by individual Fermilab experiments, as well as projected future needs, should be detailed in the use case study which covers that experiment. Consult the table in section 1 for the applicable use case study. An aggregate view of current hardware computing infrastructure for all Fermilab is described in Sections 7.4.2.3.1 and 7.4.2.3.2.

7.4.2.3.1 Storage

The US-CMS Tier-1 has dedicated storage, both disk and tape. The dCache distributed storage system is used to provide disk storage for production CMS computing. CMS dCache servers are (nearly...) all connected to the network with 100GE NICs. There is also a much smaller EOS storage system for derived data objects used in analysis by physicists. The EnStore tape system currently provides tape storage for archiving and back-end storage of production CMS computing data. The rest of the experiments supported at Fermilab use shared storage facilities. Disk storage for them is provided on the public dCache storage facility. The EnStore tape system is used for tape storage by those experiments, although the storage elements are separate from CMS. Table 7.4.4 depicts the approximate current capacity/capabilities of Fermilab storage systems used to support its experiments.

Experiment	Type	# of Servers	Storage Space
CMS	Disk (dCache)	220	70PB
CMS	Disk (EOS)	160	12PB
NON-CMS EXPERIMENTS	Disk (dCache)	160	46PB
CMS	Tape (EnStore)	50	150PB
NON-CMS EXPERIMENTS	Tape (EnStore)	70	200PB

Table 7.4.4: Fermilab Storage Breakout

7.4.2.3.2 Compute

The US-CMS Tier-1 facility has dedicated compute nodes as well. These worker nodes are partitioned into use for production CMS computing and for analysis use by CMS physicists. The latter is referred to as the LHC Physics Center (LPC). The partitioning is not dynamic, but can be changed as production computing demands fluctuate. The rest of the Fermilab-supported experiments use a shared compute infrastructure, named the FermiGrid Compute Facility. In addition, there is a smaller, general use compute facility (referred to as Wilson) for more casual use and as a development environment. Wilson does include a modest level of GPU capabilities. The current capacity/capabilities of the Fermilab compute facilities is shown in Table 7.4.5.

Experiment	# of Worker Nodes	# of Cores	# of GPUs
CMS (INCLUDING LPC)	500	40k	0
NON-CMS EXPERIMENTS	530	36k	0
DEVELOPMENT	100	1.9k	25

Table 7.4.5: Fermilab Compute Breakout

7.4.2.3.3 Future Storage/Compute Resource Upgrades/Expansion

Modest incremental growth in storage and compute is expected over the short-term time frame (~2 yr). Starting in roughly the 2028-2030 time frame, the US-CMS Tier-1 storage and compute facilities will need to expand significantly to meet the demands of High-Luminosity LHC. In addition, DUNE/FD will be ramping up to take beam data. This expansion of compute and storage resources will continue into the early 2030s.

7.4.2.4 Cloud Services

See the individual experiment use case studies identified in Table 7.4.1 (Section 7.4.2.1) for discussion of cloud services being utilized or under consideration by individual Fermilab experiments.

7.4.2.5 Artificial Intelligence and ML

See the individual experiment use case studies identified in Table 7.4.1 (Section 7.4.2.1) or discussion of artificial intelligence and ML techniques being utilized or under consideration by individual Fermilab experiments.

In addition, Fermilab networking has an interest in applying AI/ML techniques to its WAN network traffic data. Potential applications for AI/ML-based analysis include:

- Active network flow performance analysis, in terms of identifying sub-optimal performance within individual data transfer flows, and predicting whether said performance problems are likely to be network-related or end system-related;

- Network path asymmetry detection, and predicting whether this is or is not significantly affecting data-movement performance;
- Jumbo frame use/non-use, and predicting whether jumbo frames are/will significantly impact performance, either positively or negatively.

Fermilab has a strong interest in making use of ESnet High-Touch Services to generate data for AI/ML-based analysis of network traffic patterns. Fermilab is currently working cooperatively with ESnet on distinguishing between streaming-based CMS data movement and file-transfer-based CMS data movement. It is desirable not just to continue with this particular effort, but also to look toward additional opportunities to expand use of High-Touch Service capabilities in analyzing WAN traffic.

7.4.2.6 Wireless Capabilities

No programmatic uses of wireless technologies beyond normal user connectivity are currently planned. There have been several R&D investigations with use of 5G technologies for both beamline operations activities and advanced detector design. However, none of those efforts have gone beyond the early R&D stage. There are no roadmaps for deployment of either at this time.

7.4.2.7 Data-Related Resource Constraints

See the individual experiment use case studies identified in Table 7.4.1 (Section 7.4.2.1) for data-related resource constraint concerns on the part of individual Fermilab experiments.

7.4.2.8 Data Mobility Endpoints

Globus is not deployed at Fermilab as a general use service. See the individual experiment use case studies identified in Table 1 (Section 1) for discussion of Globus and other data transfer software middleware being utilized by individual Fermilab experiments.

7.4.2.9 Outstanding Issues

Fermilab is concerned about not having the capability to adequately stress test the WAN infrastructure as network technology step-increases push bandwidth capacities to terabit scale. The ability to generate traffic at 400GE, 800GE, & 1.6TE is challenging in its own right. Fermilab desires to be able to routinely stress test the WAN infrastructure to ensure it is performing as expected. The complexity of having multiple types of virtual network services sharing the same physical infrastructure adds another level of difficulty to this. It is desirable to consider this as a capability to concurrently generate high traffic loads across various virtual network services (LHCOPN, LHCONE, default routed IP path, etc.). This capability can be seen as being a critical tool in Fermilab's ability to troubleshoot aggregate data bottlenecks where multiple virtual network services are involved. Fermilab proposes that this should be developed as a service that ESnet provides to all its sites.

7.4.2.10 Technology Profile Contributors

Fermilab Representation

Andrey Bobyshev, Fermilab, bobyshev@fnal.gov

Phil Demar, Fermilab, demar@fnal.gov

Ryan Heath, Fermilab, rheath@fnal.gov

Ray Pasetes, Fermilab, rayp@fnal.gov

7.5 BNL

BNL delivers discovery science and transformative technology to power and secure the nation's future. Primarily supported by the US DOE SC, BNL is a multidisciplinary laboratory with seven Nobel Prize-winning discoveries, 37 R&D 100 Awards, and more than 70 years of pioneering research.

7.5.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- BNL delivers discovery science and transformative technology to power and secure the nation's future. Primarily supported by the US DOE Office of Science, Brookhaven Lab is a multidisciplinary laboratory. BNL supports a number of HEP experiments, including: [Section 4.1]
  - USATLAS
  - Belle II
  - DUNE
- The SDCC at Brookhaven Lab began in 1997 when the RHIC and ATLAS Computing Facility was established. Then called the RHIC & ATLAS Computing Facility (RACF), the full-service scientific computing facility has since supported some of the most notable physics experiments, including Broad RAnge Hadron Magnetic Spectrometers (BRAHMS), Pioneering High Energy Nuclear Interaction eXperiment (PHENIX), PHOBOS Collaboration, and Solenoidal Tracker at RHIC (STAR), by providing dedicated data processing, storage, and analysis resources for these diverse, expansive experiments with general computing capabilities and support for users. Today, this history of providing useful, resilient, and large-scale computational science, data management, and analysis infrastructure has grown, and the SDCC has evolved to support additional facilities and experiments. These include the National Synchrotron Light Source II (NSLS-II) and CFN at Brookhaven Lab, as well as other DOE SC User Facilities; the ATLAS experiment at CERN's LHC in Europe; and Belle-II at KEK in Japan and DESY in Germany. SDCC also is planning for its role in future experiments with the sPHENIX, DUNE, and Electron-Ion Collider. [Section 4.1]
- The BNL Network Perimeter consists of two redundant Nokia Routers which support interface speeds of 10 GbE, 40 GbE, 100 GbE and 400 GbE. These routers are split between BNL's two data centers, one in building 515 and the other in building 725. Each BNL router connects to a local ESnet router at 800 Gbps (2 x 400 GbE) totaling 1.6 Tbps of Internet/WAN connectivity. Each 800 Gbps circuit is utilized by all scientific and administrative communities at BNL, all data traffic to and from BNL flows through either of these circuits. [Section 4.7]
- Each WAN/Internet circuit provided from ESnet is redundantly connected to each BNL Network Perimeter router for resiliency. Only one circuit from each service is active at any given time. To support the large data flows to and from US ATLAS resources BNL offsets the LHCOPN and LHCONE primary circuits so each service in theory could transfer 800 Gbps of data. [Section 4.7]
- To adhere to the LHCONE and LHCOPN AUP, various complex routing policies and features must be implemented on BNL Network Perimeter. This AUP states only specific HEP resources at BNL (e.g., US ATLAS, Belle II, DUNE, etc.) should have access to the LHCONE and LHCOPN services, which leaves all other resources at BNL able only to traverse the general Internet services. To provide this exclusivity for LHCONE and LHCOPN, separate campus Virtual Routing and Forwarding (VRF) instances have been created on both perimeter routers. Policies are then applied to only import default, black hole, and local BNL routes into

the campus VRFs. Then policy-based routing (PBR) policies are configured on all perimeter downstream ingress interfaces. In short, these PBR policies state that all private IPv4 and IPv6 and HEP subnets forward to the default routing table, which has full visibility to all WAN circuits, while all other BNL subnets should forward to the campus VRF instance, which can only see the general Internet services. [Section 5.7]

7.5.2 BNL: Technology Profile

BNL delivers discovery science and transformative technology to power and secure the nation's future. Primarily supported by the US DOE SC, Brookhaven Lab is a multidisciplinary laboratory.

The Laboratory's staff members lead and support diverse research teams that address the DOE mission to ensure the nation's security and prosperity by addressing its energy, environmental, and nuclear challenges through transformative science and technology solutions. Among Brookhaven Lab's current initiatives are nuclear science, energy science, data science, particle physics, accelerator science and technology, quantitative plant science, and quantum information science.

7.5.2.1 Facilities Operated and Supported

Brookhaven Lab staff are entrusted by DOE on the nation's behalf to safely build and operate facilities that house unique tools used by interdisciplinary teams comprising researchers from Brookhaven as well as those from other national labs, academia, and industry.

BNL supports a number of HEP experiments, including:

- US ATLAS
- Belle II
- DUNE

Additionally, BNL has provided input into the ESnet requirements review process for other facilities and experiments:

- CFN
- PHENIX & sPHENIX
- STAR
- US Lattice Quantum Chromodynamics (USQCD)
- Electron-Ion Collider (EIC)

7.5.2.2 Network and Data Architecture

7.5.2.2.1 BNL Network Architectures

The BNL network infrastructure's purview is to provide data communications that are highly available and robust in order to facilitate and support operations and research at BNL. It provides the foundation underlying the critical services within the enterprise and scientific computing such as, HTC, phone service, security and safety and monitoring.

The success of the BNL network infrastructure in supporting its critical services can be attributed to the following fundamental principles, reliability and resiliency, modularity and hierarchy. These fundamental principles are architected and implemented throughout all the various network infrastructures at BNL.

1. Reliability and Resiliency
    - A highly available and reliable network that can survive any component failure without human intervention.
2. Modularity
    - A network divided into multiple functional modules, each serving a specific role in the network. Modules should be easy to replicate, redesign and expand. Modularity helps promote, ease of network manageability, adaptability, scalability and creates fault isolation domains.
3. Hierarchy
    - Provides the ability to have more choke points within the network to aggregate traffic, reachability and topology information.

As of today, the BNL network infrastructure is separated into three main areas, the Network Perimeter, Campus Network and the HTSN. Each of these areas support different critical BNL services:

1. Network Perimeter
    - A highly available, Terabit per second (Tbps) network infrastructure to support the timely transfer of large amounts of scientific data from various systems or experiments to remote locations via the Internet or WAN.
2. Campus Network
    - Foundation underlying the critical services within the enterprise such as workstations, phone service, security, safety and monitoring and enterprise computing.
3. HTSN
    - A centralized, highly available, fault-tolerant, scalable, Terabit per second network infrastructure to support all collaborations compute and storage requirements.

7.5.2.2.2 BNL Network Perimeter

The BNL Network Perimeter consists of two redundant Nokia Routers which support interface speeds of 10 GbE, 40 GbE, 100 GbE and 400 GbE. These routers are split between BNL's two data centers, one in building 515 and the other in building 725. Each BNL router connects to a local ESnet router at 800 Gbps (2 x 400 GbE) totaling 1.6 Tbps of Internet/WAN connectivity. Each 800 Gbps circuit is utilized by all scientific and administrative communities at BNL, all data traffic to and from BNL flows through either of these circuits.

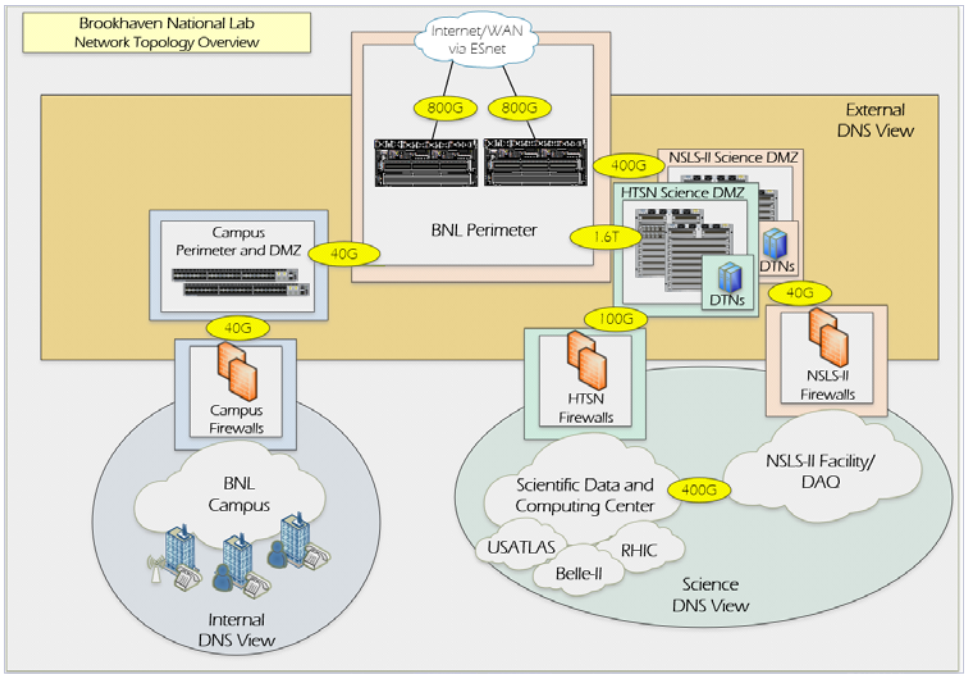


Figure 7.5.1: High-Level Overview of the BNL Network Perimeter and DNS Architecture (FY24)

7.5.2.2.2.1 BNL WAN and Internet Circuits

As of today, the BNL Network Perimeter supports various WAN circuits along with its Internet circuits. These WAN/Internet circuits all share the same 1.6 Tbps of total connectivity to ESnet. There is no QoS, traffic shaping or prioritization on any data sources over the BNL WAN/Internet circuits. The following are BNL's current WAN/Internet circuits:

1. General IPv4 Internet
2. General IPv6 Internet
3. LHCONE Service
4. LHCOPN Service

Each WAN/Internet circuit provided from ESnet is redundantly connected to each BNL Network Perimeter router for resiliency. Only one circuit from each service is active at any given time. To support the large data flows to and from US ATLAS resources BNL offsets the LHCOPN and LHCONE primary circuits so each service in theory could transfer 800 Gbps of data.

All WAN/Internet circuits terminate on the BNL Network Perimeter and utilize EBGP and Bidirectional Forwarding Detection (BFD). BNL does not allow any Layer-2/VPLS (ESnet OSCARS) circuits to traverse the Network Perimeter since Layer-3 routing is strictly enforced.

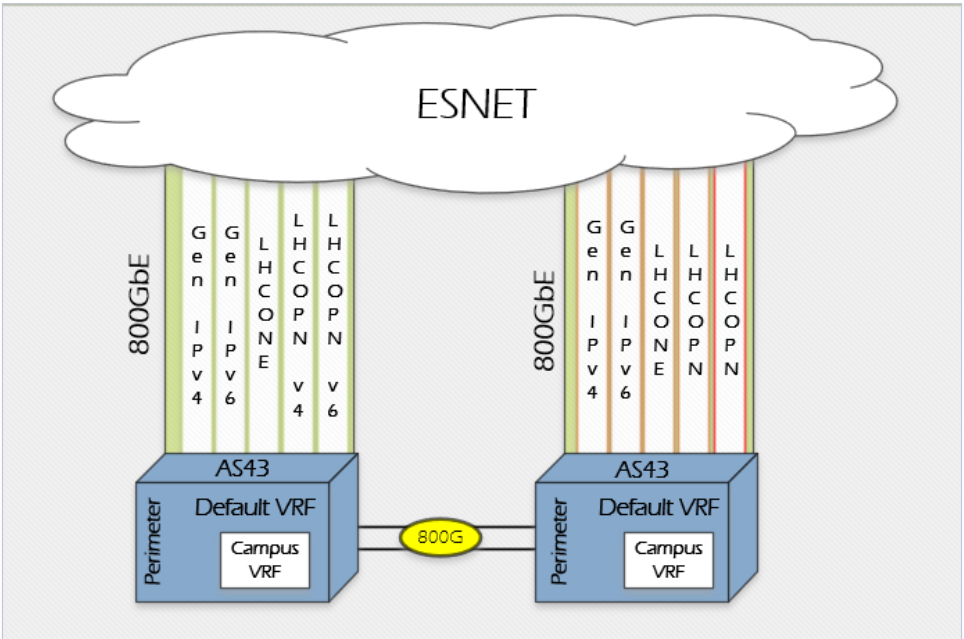


Figure 7.5.2: High-Level Overview of the BNL's WAN and Internet Circuits provided from ESnet (FY24)

7.5.2.2.2.2 Connectivity to Science DMZ(s) and Campus Network

The BNL Network Perimeter strictly utilizes Layer-3 routing protocols such as OSPF/OSFPv3 and BGP. Between both Perimeter routers Open Shortest Path First (OSPF0 / OSPFv3 and iBGP) are utilized but when connecting to other downstream devices (e.g., Science DMZ(s), Campus Network) EBGP and BFD are used. All downstream devices utilizing EBGP are configured with a private ASN, since these ASNs are private and should not be advised externally they are stripped when their routes are advertised to ESnet or any other WAN peering's.

The utilization of EBGP between the Network Perimeter and downstream devices allows for better traffic engineering and provides BNL with the ability to steer large data flows to the correct Perimeter router that has the active Internet/WAN peering. Utilizing EBGP also eliminates the need to redistribute BGP routes into and Interior Gateway Protocol such as OSPF.

As depicted above in "High-Level Overview of the BNL Network Perimeter and DNS Architecture" the following connectivity is provided to downstream devices:

1. BNL Campus
    a. 40 Gbps in aggregate (4 x 10 GbE)
2. HTSN Science DMZ
    a. 1.6 Tbps in aggregate (4 x 400 GbE)
3. NSLS-II Science DMZ
    a. 400 Gbps in aggregate (4 x 100 GbE)

7.5.2.2.2.3 BNL's Support for LHCOPN and LHCONE Services

To adhere to the LHCONE and LHCOPN AUP, various complex routing policies and features must be implemented on BNL Network Perimeter. This AUP states only specific HEP resources at BNL (e.g., US ATLAS, Belle II, DUNE, etc.) should have access to the LHCONE and LHCOPN services, which leaves all other resources at BNL able only to traverse the general Internet services.

To provide this exclusivity for LHCONE and LHCOPN, separate campus Virtual Routing and Forwarding (VRF) instances have been created on both perimeter routers. Policies are then applied to only import default, black hole, and local BNL routes into the campus VRFs. Then policy-based routing (PBR) policies are configured on all perimeter downstream ingress interfaces. In short, these PBR policies state that all private IPv4 and IPv6 and HEP subnets forward to the default routing table, which has full visibility to all WAN circuits, while all other BNL subnets should forward to the campus VRF instance, which can only see the general Internet services.

Below is a diagram depicting how and where the specific VRFs and policies are placed on the Perimeter routers.

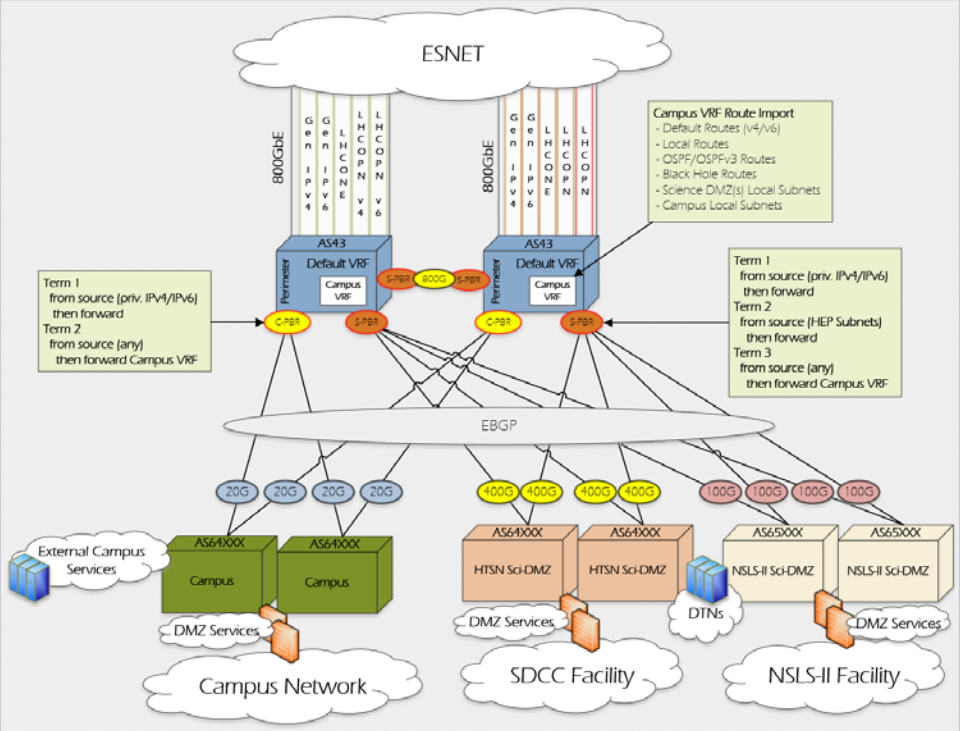


Figure 7.5.3: High-Level Overview of the BNL's PBR and VRFs to Support LHCONE/LHCOPN (FY24)

7.5.2.2.2.4 Network Perimeter Security Features

The BNL Network Perimeter implements Access Control Lists (ACLs) on all Internet peering's as well as utilizing Black Hole Routing using BGP.

7.5.2.2.3 BNL Campus Network

The BNL Campus Network supports critical services within the enterprise such as workstations, phone service, security, safety and monitoring and enterprise computing. The BNL campus is over 5,000 acres and contains over 300 buildings with over 100 of these buildings connected to the BNL Campus Network through various types of Layer-1 technologies such as, DSL, wireless point-to-point bridges or through the large campus fiber optic cabling infrastructure. The BNL Campus Network is comprised of over 800+ switches, 800+ wireless access points and numerous pairs of firewalls and load balancers from various vendors.

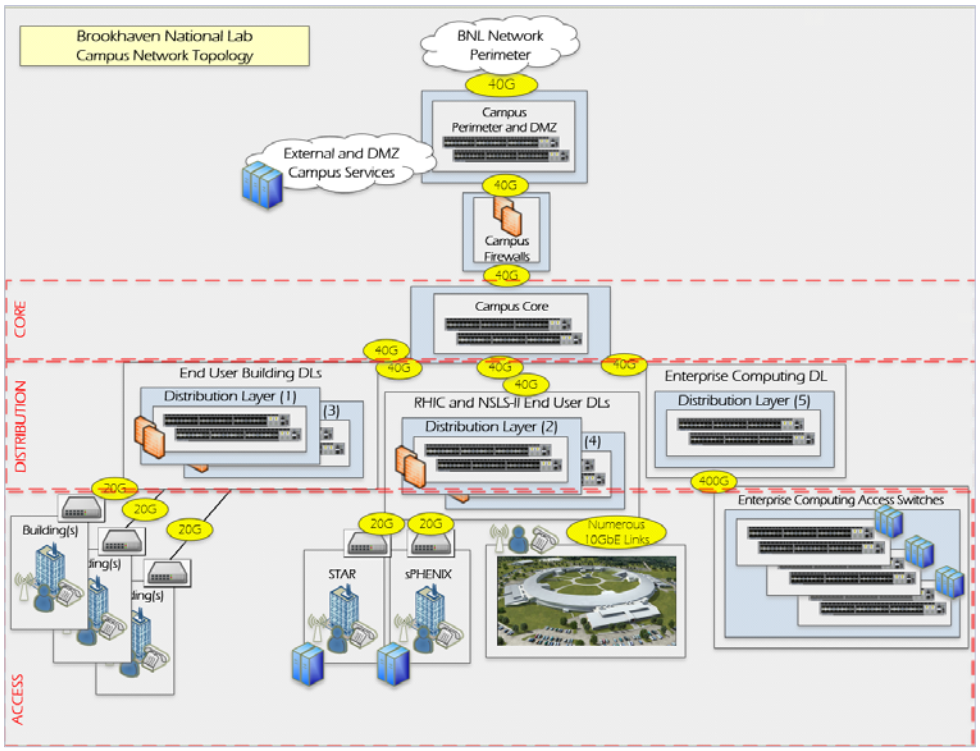


Figure 7.5.4: High-Level Overview of the BNL's Campus Network Topology (FY24)

7.5.2.2.3.1 BNL Campus Architecture

The BNL Campus Network architecture follows the same fundamental principles as discussed in the beginning of the section: modularity, resiliency and hierarchy. The Campus architecture follows a well-defined and traditional three-tiered model comprising the Core, Distribution, and Access layers.

1. Core Layer
    a. Is the least complex layer yet the most critical part of the Campus Network.
    b. It provides a limited set of services and is designed to be highly resilient.
    c. The responsibility of the Core is fast packet forwarding and flexibility, not policy insertion or server termination.
    d. Serves as Area 0 in the OSPF and OSPFv3 topologies.
2. Distribution Layer (DL)
    a. Connects directly to the Campus Core via OSPF/OSPFv3 via Area 0.
    b. Defines a summarization boundary for OSPF and OSFPv3 since each DL is a separate OSPF Area.
    c. Each distribution switch pair is configured as a MLAG cluster, which creates one logical Spanning-Tree and LACP switch from the access layer switches perspective. This allows access layer switches to redundantly connect to both DL switches via LACP in an active/active fashion.
    d. Acts as an active/active default gateway for all Non-Firewalled Switch Virtual Interfaces (SVIs) and Spanning-Tree root bridge for all virtual LANs (VLANs) that are defined within the DL.

e. Is an aggregation point for all the access switches.

f. Services and control boundary between the access layer and Campus Core.

g. Each DL usually has a pair of firewalls for various services that require access control from other internal BNL subnets.

3. Access Layer

a. Is where devices (e.g., workstations, access points, phones, printers, etc.) connect to the network.

b. Various interfaces speeds and duplex's must be supported (e.g., 10 Mbps, 100 Mbps, 1 Gbps, half duplex, full duplex).

c. Must support Power over Ethernet (PoE) for phones, access points and other devices requiring power over the network.

7.5.2.2.4 HTSN

BNL's HTSN is a vendor agnostic, resilient, scalable and modular Terabit Per Second (Tbps) network infrastructure which serves as the primary network transport for all data-intensive collaborations at BNL. It provides high-throughput connectivity to all High-Performance Compute (HPC) and HTC collaborations and supports the timely transfer of large amounts of scientific data via the Internet.

The HTSN has five key components:

1. Network Perimeter

a. Discussed in section 2.1

2. Science DMZ

a. Supports open, high-speed WAN and Internet access for all scientific collaborations throughout the BNL campus.

b. Is configured as an MLAG cluster which allows all DTNs to redundantly connect to each chassis in an active/active fashion via LACP.

c. Supports 10, 25, 40, 50, 100 and 400 GbE interface speeds.

3. Science Core

a. A Tbps Science and Data Center Interconnect for data-intensive collaborations at BNL. This Science Interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, sPHENIX, CAD, CFN, NSLS-II, HPC Clusters and the SDCC.

b. Intelligence and routing policies are applied within the Science Core to restrict or grant access to specific resources within the SDCC.

c. Supports 10,25,40 and 100 GbE interface speeds.

4. Spine

a. A Tbps network Spine that interconnects all Leaf switches. Leaf switches consist of ToR or chassis-based switches that connect compute, storage or general infrastructure service servers.

b. The responsibility of the Spine is fast packet forwarding and flexibility, not policy insertion or server termination.

c. EBGP is utilized throughout the HTSN. EBGP was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each Spine group

shares the same ASN but does not have IBGP peering between them. Each Leaf or pairs of Leaves will require their own ASN.

d. Supports 10,25,40,50,100 and 400 GbE interface speeds.

5. Storage Core

a. A redundant Tbps switching block that aggregates high-performance storage services.

b. Is configured as an MLAG cluster which allows all DTNs to redundantly connect to each chassis in an active/active fashion via LACP.

c. Supports 10,25,40,50,100 and 400 GbE interface speeds.

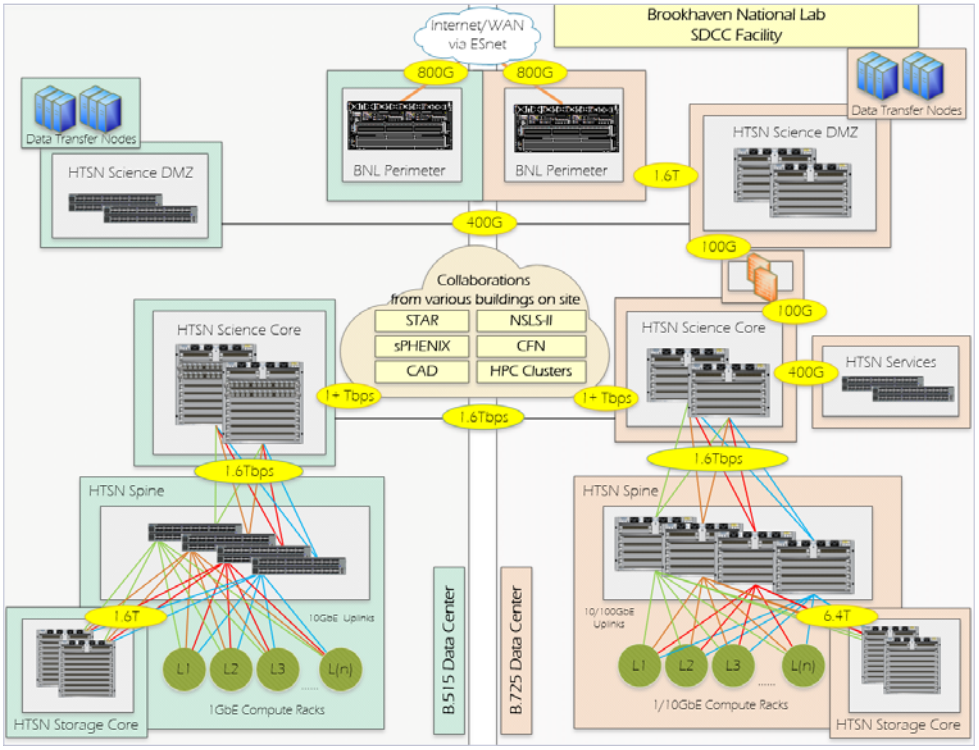


Figure 7.5.5: High-Level Overview of the BNL HTSN in both BNL data centers (FY24)

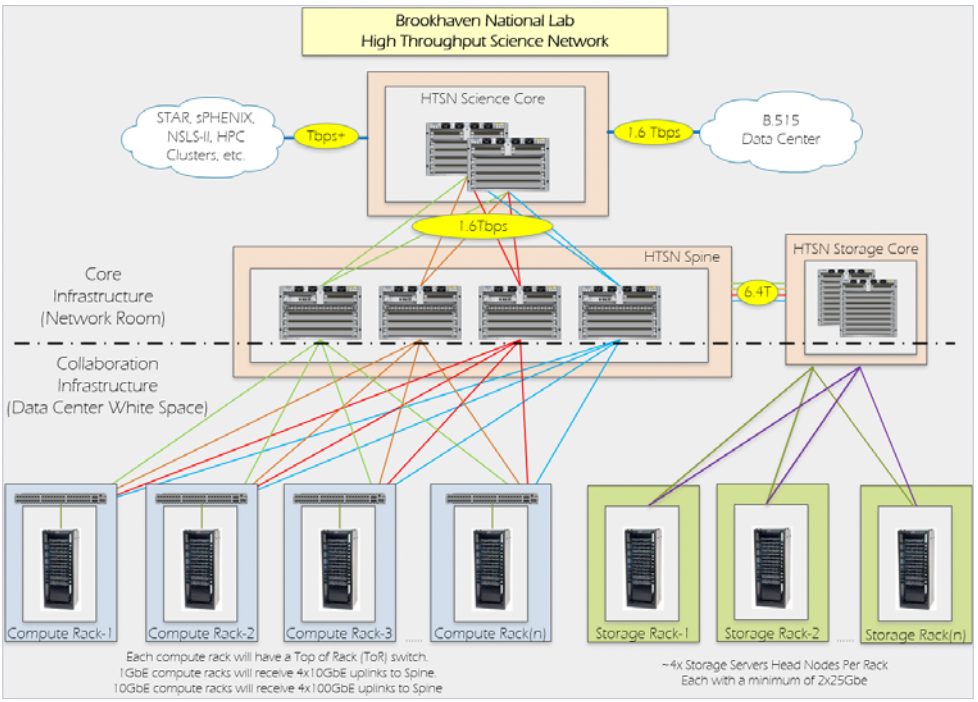


Figure 7.5.6: High-Level Overview of the BNL HTSN Demarcation Points Between Core and Collaboration Infrastructures (FY24)

7.5.2.2.5 NSLS2 Network Infrastructure

The NSL2 network infrastructure comprises of three functional areas:

1. NSLS2 Science Network (aka N2SN)
    a. A highly available Terabit per second (Tbps) network infrastructure to support the operation of science experiments performed at the individual Beamlines.
2. NSLS2 Accelerator Network (aka Controls)
    a. A highly available Gigabit per second (Gbps) network infrastructure to support the operations and monitoring of the medium-energy electron storage ring.
3. NSLS2 Network Management
    a. A highly available Gigabit per second (Gbps) network infrastructure to support the operations and monitoring of the network switches and firewalls which comprise of the NSLS2 Infrastructure.

7.5.2.2.5.1 NSLS2 Science Network (N2SN) Components

1. N2SN Core
    a. Supports the connectivity between the N2SN components, the NSLS2 Accelerator network, NSLS2 Network Management as well as the HTSN Science Core.
2. N2SN Beamline block (1 per Beamline)
    a. Supports services for the operation of individual Beamline experiments.
3. N2SN Data Services block
    a. Supports data services such as DNS, DHCP, etc., as well as Storage.

4. N2SN DMZ block
    a. Supports connectivity to the Perimeter routers for Internet and BNL Campus traffic from both the N2SN and Controls functional areas.

7.5.2.2.5.2 NSLS2 Accelerator/Controls Network Components

1. High-Availability pair of firewalls
    a. Used to separate and enforce policy for traffic to/from the Accelerator/Controls network.
2. Accelerator Aggregation block
    a. Supports the connectivity between the Accelerator components, as well as Storage.
3. Accelerator Data Services block
    a. Supports data services such as DNS, DHCP, VM IOC servers, etc.
4. Accelerator Mezzanine cell block (30)
    a. Supports the individual cells comprising of the accelerator ring.
5. Accelerator RF Cavity block
    a. Supports the function of the accelerator ring.
6. Accelerator Injection Service Building block
    a. Supports the function of the accelerator ring.
7. Accelerator Control Room
    a. Supports the operations and monitoring of the accelerator ring.

7.5.2.2.5.3 N2SN Core

Supports the connectivity between the N2SN components, the NSLS2 Accelerator network, NSLS2 Network Management as well as the HTSN Science Core.

1. Consists of redundant Tbps chassis configured in an MLAG configuration capable of supporting 10 GbE, 25 GbE, 40 GbE and 100 GbE interface speeds.
2. Connectivity between the N2SN Core and the Beamlines is done via a 20 Gbps (2 x 20 GbE) or 200 Gbps (2 x 100 GbE) connection.
3. Connectivity between the N2SN Core and the N2SN Data Services block is done via a 400 Gbps (4 x 100 GbE) connection.
4. Connectivity between the N2SN Core and the N2SN DMZ is separated by a redundant pair of firewalls at 80 Gbps (2 x 40 GbE).
    a. Only required traffic is permitted between these areas.
    b. Static routes are used between N2SN Core and the N2SN DMZ firewall.
5. Connectivity between the N2SN Core and the HTSN Science Core is done via redundant 400 Gbps connections.
    a. 4 x 100 GbE to B515 Data Center
    b. 4 x 100 GbE to B725 Data Center
    c. Only required traffic is permitted between these areas which are controlled by ACLs.
6. EBGP routing protocol with BFD is utilized to support connectivity to the N2SN Beamline blocks, N2SN Data Services block, the NSLS2 Network Management block, as well as the HTSN Science Core.

7. Connectivity between the N2SN Core and the NSLS2 Accelerator network is separated by redundant pair of firewalls at 80 Gbps (2 x 40 GbE).
    a. Only required traffic is permitted between these areas.
    b. Static routes are used between N2SN Core and the NSLS2 Accelerator firewall.
8. Connectivity between the N2SN Core and the NSLS2 Network Management block is done via a 40 Gbps (4 x 10 GbE) connection.

7.5.2.2.5.4 N2SN Beamline Network

Supports services for the operation of individual Beamline experiments. Each beamline has its own network infrastructure. There are currently 30+ beamline network infrastructures.

Each of the 30+ individual beamline infrastructures consist of the following:

1. A High-speed aggregation switch capable of supporting 10 GbE, 25 GbE, 40 GbE and 100 GbE interface speeds. Redundant High-speed aggregation switches is available should the beamline require it.
2. Multiple low-speed switches capable of supporting 10 MbE, 100 MbE and 1 GbE interface speeds with PoE.
3. Connectivity between the N2SN Core and a Beamline is done via redundant 20 Gbps (2 x 10 GbE) or 200 Gbps (2 x 100 GbE) connection.
4. EBGP routing protocol with BFD is utilized to support connectivity to the N2SN Core.
5. BGP EVPN VxLAN is utilized to transport L2 EPICS, Camera, and Instrumentation traffic between the N2SN Beamlines and ESXi VM IOC servers residing on the N2SN Data Services block.
6. Local ACLs are used to limit traffic between individual Beamlines.

7.5.2.2.5.5 NSLS2 Data Services Block

Supports data services such as DNS, DHCP, etc., as well as Storage.

1. Consists of redundant Tbps switches in an MLAG configuration capable of supporting 10 GbE, 25 GbE, 40 GbE and 100 GbE interface speeds
2. Connectivity between the N2SN Core and the N2SN Data Services block is done via a 400 Gbps (4 x 100 GbE) connection.
3. Connects directly to the following:
    a. Storage nodes via 100 Gbps (4 x 25 GbE) connection
    b. Redundant High-speed switches also in an MLAG configuration capable of supporting 10 GbE, 25 GbE, 40 GbE and 100 GbE interface speeds and are connected via 400 Gbps (4 x 100 GbE)
4. EBGP routing protocol with BFD is utilized to support connectivity to the N2SN Core
5. BGP EVPN VxLAN is utilized to transport L2 EPICS, Camera, and Instrumentation traffic between the N2SN Beamlines and ESXi VM IOC servers residing on the N2SN Data Services block.

7.5.2.2.5.6 NSLS2 DMZ

Supports connectivity to the Perimeter routers for Internet and BNL Campus traffic from both the N2SN and NSLS2 Accelerator/Controls functional areas.

1. Consist of the following:
    a. Redundant Tbps switches in an MLAG configuration capable of supporting 10 GbE, 25 GbE, 40 GbE and 100 GbE interface speeds.
    b. High-Availability pair of firewalls
2. Connectivity between the N2SN Core and the N2SN DMZ is separated by this redundant pair of firewalls at 80 Gbps (2 x 40 GbE).
    a. Only required traffic is permitted between these areas.
    b. Static routes are used between N2SN Core and the N2SN DMZ
3. Connectivity between the N2SN DMZ and the Perimeter routers is done via redundant 200 Gbps connections.
    a. 2 x 100 GbE to B515 Data Center
    b. 2 x 100 GbE to B725 Data Center
4. EBGP routing protocol with BFD is utilized to support connectivity to the Perimeter routers.

7.5.2.2.5.7 NSLS2 Accelerator/Controls Infrastructure

Supports critical services for the operation and monitoring of the medium-energy electron storage ring. This infrastructure consists of an Aggregation block, a Data Services block, 30 Mezzanine cell blocks, an RF Cavity block, an Injection Service Building block, and a Controls room block. This infrastructure is comprised of 220+ Layer 2 switches interconnected using fiber optic cabling.

1. Connectivity between the N2SN Core and the NSLS2 Accelerator network is separated by redundant pair of firewalls at 80 Gbps (2 x 40 GbE).
    a. Only required traffic is permitted between these areas.
    b. Static routes are used between N2SN Core and the NSLS2 Accelerator firewall.
1. The RF Cavity block, the Injection Service Building block and Controls room all consist of:
    a. A pair of High-speed switches in an MLAG configuration to support redundant connectivity both downstream and upstream.
    b. Multiple low-speed switches capable of supporting 10MbE, 100MbE and 1 GbE interface speeds with PoE.
2. This infrastructure consists of a large Layer 2 network with Layer 3 SVIs terminating on the main aggregation switch or the firewall depending on security requirements.

7.5.2.2.5.8 NSLS2 Network Management

A highly available Gigabit per second (Gbps) network infrastructure to support the operations and monitoring of the network switches and firewalls which comprise of the NSLS2 Infrastructure.

1. Consist of the following:
    a. Redundant Gbps switches configured in an MLAG configuration capable of supporting 10 GbE and 25 GbE interface speeds.
    b. Low-speed switch capable of supporting 10 MbE, 100 MbE and 1 GbE interface speeds used for Out-of-band management of Data Center switches only.

2. Connectivity between the N2SN Core and the NSLS2 Network Management block is done via a 40 Gbps (4 x 10 GbE) connection for in-band switch management.
3. EBGP routing protocol with BFD is utilized to support connectivity to the N2SN Core.

7.5.2.3 Computational and Storage Infrastructure

The SDCC at Brookhaven Lab began in 1997 when the RHIC and ATLAS Computing Facility was established. Then called RACF, the full-service scientific computing facility has since supported some of the most notable physics experiments, including BRAHMS, Pioneering High Energy Nuclear Interaction eXperiment (PHENIX), PHOBOS Collaboration, and Solenoidal Tracker at RHIC (STAR), by providing dedicated data processing, storage, and analysis resources for these diverse, expansive experiments with general computing capabilities and support for users.

Today, this history of providing useful, resilient, and large-scale computational science, data management, and analysis infrastructure has grown, and the SDCC has evolved to support additional facilities and experiments. These include the National Synchrotron Light Source II (NSLS-II) and CFN at Brookhaven Lab, as well as other DOE SC user facilities; the ATLAS experiment at CERN's LHC in Europe; and Belle II at KEK in Japan and DESY in Germany. SDCC also is planning for its role in future experiments with the sPHENIX, DUNE, and Electron-Ion Collider.

For users, SDCC offers a single-point facility with high-throughput, high-performance, and data-intensive computing along with data management, storage, analysis, and preservation capabilities that support scientific programs in domains spanning physics, biology, climate, and energy. SDCC also serves as the core for Brookhaven Lab's institutional computing needs, connecting scientists and researchers at home and abroad.

SDCC maintains a 1,050,000 HEPSPEC 2006 (HS06) unit processing farm (currently with 110,000 processing cores), a distributed and centralized disk storage farm (more than 80 PB of online disk storage), robotic tape storage silos (250 PB of storage), a 400 terabit-per-second internal network with up to 2 x 400 gigabits per second of WAN connectivity, and data grid and cloud computing software infrastructure. The hardware is a combination of commodity-based processing servers, enterprise-class Unix servers, and highly specialized mass storage systems, all connected by high-speed network infrastructure.

7.5.2.4 Cloud Services

See Section 6.7, 6.9, 6.11, and 6.12 for discussion of cloud services being utilized at BNL to support different experiments.

7.5.2.5 Artificial Intelligence and ML

See Section 6.7, 6.9, 6.11, and 6.12 for discussion of AI and ML being utilized at BNL to support different experiments.

7.5.2.6 Wireless Capabilities

BNL utilizes wireless services to provide network access for enterprise users, enable science, monitor remote pump stations using LoRaWan, and enable internet connectivity in the apartments.

7.5.2.6.1 Enterprise Wireless

1. Provides wireless network access to users and staff
2. Open service set identifier (SSID)
3. Authenticated via captive portal
    a. Active Directory for staff and users
    b. Guests can register for a temporary account using their phone number
4. Limited access to BNL internal networks.

7.5.2.6.2 LoRaWAN Remote Monitoring

The BNL Energy and Utilities division has several wells located across the site. Many of these locations have no network connectivity available. The need to monitor the level of these wells has been needed for some time as operators currently have to do a visual inspection of the sites. It would be far too costly to install traditional wired network services to these locations, so another solution was needed. LoRaWAN was chosen due to its long range, ease of integration and relatively low cost to fill this need.

LoRa gateways are currently being deployed in three locations throughout the site to cover the needed areas. These gateways will receive data from various Lora sensors which will be sent back to an on-premises server. The server has native integration with the Modbus protocol, so the BNL energy and utilities division can poll these data points using their building management software.

The vendor for this project is Activity. They provide the application server, ThingPark Enterprise and the Long Range Relay (LRR) software for the gateways. The gateways are Kerlink Wirnet iStations.

The most challenging aspect of this project has been sourcing the necessary components. Originally, BNL was looking to Cisco for this before they decided to exit the LoRa business. Cisco directed BNL towards Actility for the software components. Special arrangements were made with vendors to purchase the necessary hardware, which was difficult since the technology was not being resold at the time.

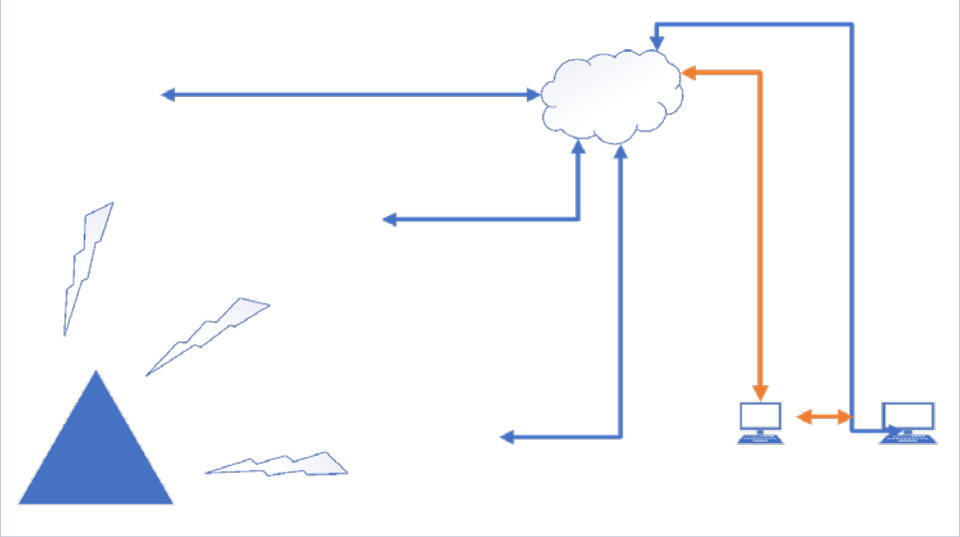


Figure 7.5.7: High-Level LoraWan Architectures
This project is currently in production, so the final architecture is subject to change.

7.5.2.6.3 Site Cellular Network Connectivity

Various locations throughout the BNL campus have network connectivity through legacy DSL uplinks. Due to growing bandwidth consumption in these areas, the need for modern network connectivity has been long overdue. After being unable to procure the necessary funds to have fiber connectivity brought to these locations the idea of using a form of wireless connectivity for these areas was thought to be the best solution. Ultimately cellular capable network routers were used to provide the needed connectivity to these buildings.

There are basically two use cases for these routers. A use case where connectivity is required back to the BNL campus network and another where just internet connectivity is required. Where connectivity is required back to the BNL campus network, the routers build a VPN tunnel to BNL's VPN concentrator via the cellular network. Using this method, BNL can deploy wireless access points, VoIP phones, and additional drop-down switches to provide network connectivity. Other buildings, such as the apartments simply require connectivity to the internet, so for these use cases a cellular router and a wireless access point is deployed for user connectivity to the internet.

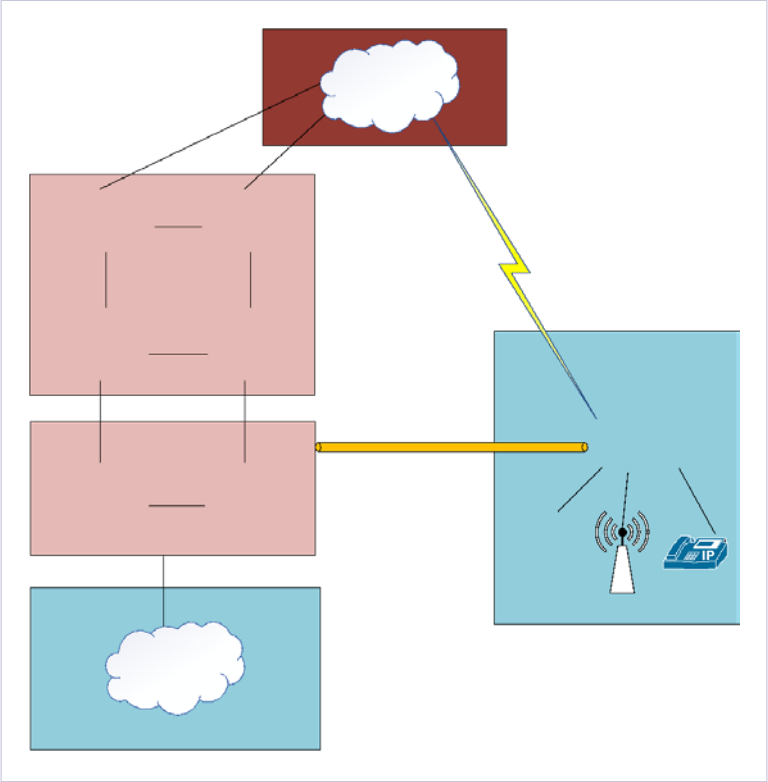


Figure 7.5.8: Cellular Router Connectivity to BNL Campus Network via VPN

7.5.2.6.4 NSLS2 Wireless Network

As described previously, the enterprise wireless network is an open SSID authenticated via a captive portal with limited access to the BNL networks. The need for additional access to the BNL network via a wireless has arisen over time, specifically in the National Synchrotron Light Source II (NSLS2) facility. Many of the devices needing this type of connectivity are headless devices and unable to authenticate through the captive portal. The answer to this was to provision a new hidden SSID using a pre-shared key for authentication with specific access to parts of the BNL network at NSLS2 and the internet. The SSID is hidden to prevent normal users from attempting to connect to it and the passkey is only shared with authorized users.

7.5.2.7 Data-Related Resource Constraints

No constraints were identified.

7.5.2.8 Data Mobility Endpoints

The BNL Science DMZ supports open, high-speed WAN and Internet access for all scientific collaborations throughout the BNL campus. To this end, a number of DTNs are available on a per-experiment or facilitate basis, and are redundantly connected to the network and storage infrastructure.

7.5.2.9 Outstanding Issues

No issues were identified.

7.5.2.10 Technology Profile Contributors

BNL Representation

Vincent Bonafede, BNL, bonafede@bnl.gov

Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

7.6 SLAC National Accelerator Laboratory

SLAC National Accelerator Laboratory, originally named the Stanford Linear Accelerator Center, is located in Menlo Park, California. Founded in 1962, the laboratory is sponsored by the US DOE and administrated by Stanford University.

7.6.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- The primary scientific computation and storage facilities at SLAC are provisioned in the S3DF (SLAC Shared Scientific Data Facility). The S3DF is a compute, storage and network architecture designed to support massive-scale analytics required by all SLAC experimental facilities and programs, including LCLS/LCLS-II, UED, cryo-EM, the accelerator, and the Rubin observatory. The S3DF infrastructure is optimized for data analytics and is characterized by large, massive throughput, high-concurrency storage systems. [Section 4.1]
- The SLAC networking core currently operates at multiple 100 GbE for connected enclaves and groups. The networking core of S3DF operates at 100 GbE, with a pair of 2 x 100 GbE links to the SLAC core. SLAC connects to ESnet via two pairs of 100 GbE links; these link pairs typically operate as primary/secondary. There is not general load balancing or dual use of the links. One pair of the SLAC/ESnet connections is dedicated for use by the Rubin Observatory Virtual Routed Network (Rubin Long Haul Network - LHN) over ESnet. [Section 4.7]
- Within the next two years, SLAC expects to augment the SLAC/ESnet connections with a pair of 400 GbE links. These two links are expected to replace two of the 100 GbE connections, while the two 100 GbE connections for Rubin LHN remain in place specifically for the Rubin traffic. Thus, SLAC/ESnet will be a pair of 400 GbE links and a pair of 100 GbE links. The SLAC core network will also be upgraded to N x 400 GbE to accommodate higher speed connectivity among the enclaves and groups, including S3DF. [Section 4.7]
- Within the next five years, SLAC expects to upgrade the 2 x 400 GbE links to 2x or 4x 800 GbE links. LCLS-II has a requirement for 1–2 Tbps capacity between SLAC and NERSC. This requirement is expected to be met with the N x 800 GbE upgrades. The SLAC core network will also be upgraded to N x 800 GbE. [Section 5.7]

7.6.2 SLAC: Technology Profile

The primary scientific computation and storage facilities at SLAC are provisioned in the S3DF (SLAC Shared Scientific Data Facility). The S3DF is a compute, storage and network architecture designed to support massive-scale analytics required by all SLAC experimental facilities and programs, including LCLS/LCLS-II, UED, cryo-EM, the accelerator, and the Rubin observatory. The S3DF infrastructure is optimized for data analytics and is characterized by large, massive throughput, high-concurrency storage systems. More information is available online.[145]

7.6.2.1 Facilities Operated and Supported

SLAC Shared Science Data Facility (S3DF) offers:

- Modern hardware for massive-scale analytics in a modern data center designed ground-up for research computing
- High-speed networking storage solutions for data-intensive pipelines/applications
    - Scale up to Tb/s links with hundreds of PBs online

145 http://s3df.slac.stanford.edu

- Support for the entire data pipeline from detector input through to results publication
- Funding plans that incorporate hardware lifecycle

TID Scientific Computing Systems division at SLAC is responsible for delivering S3DF. Funding is provided by SLAC for both core staff and core infrastructure, and all science directorates will be supported. The S3DF has the strategic goal of effectively replacing legacy siloed infrastructure. To this end, goals are to create high-throughput pipelines by leveraging existing technology, but also centralizing services that can be shared by the entire SLAC ecosystem.

7.6.2.2 Network and Data Architecture

The SLAC networking core currently operates at multiple 100 GbE for connected enclaves and groups. The networking core of S3DF operates at 100 GbE, with a pair of 2 x 100 GbE links to the SLAC core. SLAC connects to ESnet via two pairs of 100 GbE links; these link pairs typically operate as primary/secondary. There is not general load balancing or dual use of the links. One pair of the SLAC/ESnet connections is dedicated for use by the Rubin Observatory Virtual Routed Network (Rubin Long Haul Network - LHN) over ESnet.

Within the next two years, SLAC expects to augment the SLAC/ESnet connections with a pair of 400 GbE links. These two links are expected to replace two of the 100 GbE connections, while the two 100 GbE connections for Rubin LHN remain in place specifically for the Rubin traffic. Thus, SLAC/ESnet will be a pair of 400 GbE links and a pair of 100 GbE links. The SLAC core network will also be upgraded to N x 400 GbE to accommodate higher speed connectivity among the enclaves and groups, including S3DF.

Within the next five years, SLAC expects to upgrade the 2 x 400 GbE links to 2x or 4x 800 GbE links. LCLS-II has a requirement for 1–2 Tbps capacity between SLAC and NERSC. This requirement is expected to be met with the N x 800 GbE upgrades. The SLAC core network will also be upgraded to N x 800 GbE.

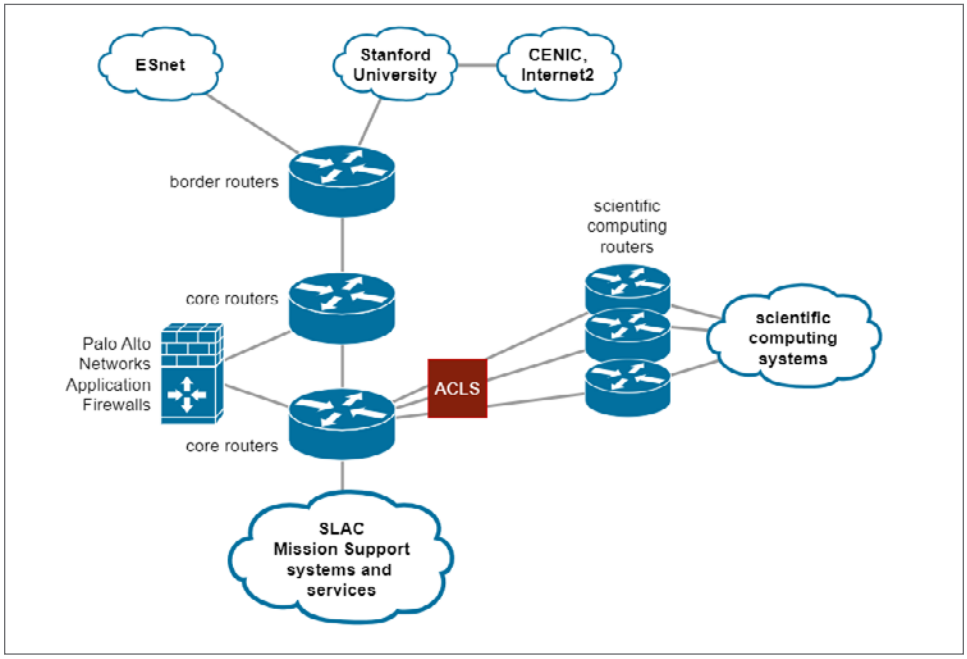


Figure 7.6.1: SLAC Network Architecture

7.6.2.3 Computational and Storage Infrastructure

There are three different ways of utilizing S3DF:

- **Interactive:** Commands that you issue are executed immediately. This is the most common approach for activities like building and debugging code, running simple analysis which require limited resources and/or require interaction with plots and logs. There are two modes for doing interactive work in S3DF: through a terminal, by opening a shell on one of the interactive pools, or through a browser, via OnDemand.
- **Batch:** Jobs are submitted to a queue and are executed as soon as resources become available. This is the most common approach for running large jobs or for executing many jobs. Note that this is the most efficient mechanism form a facility perspective because it provides the best use of the available resources. Also, note that, by far, the largest fraction of S3DF computing cycles is in the batch system. S3DF uses SLURM as workload manager for batch jobs.
- **Service:** This approach is for running long-lived jobs that run in the background waiting for data to analyze. This method is for service activities and require specific resources. These resources may be dedicated hardware acquired by your organization or may be dynamically allocated from a larger pool. For the latter approach, S3DF uses Kubernetes, an open-source framework for automating deployment, scaling, and management of containerized applications.

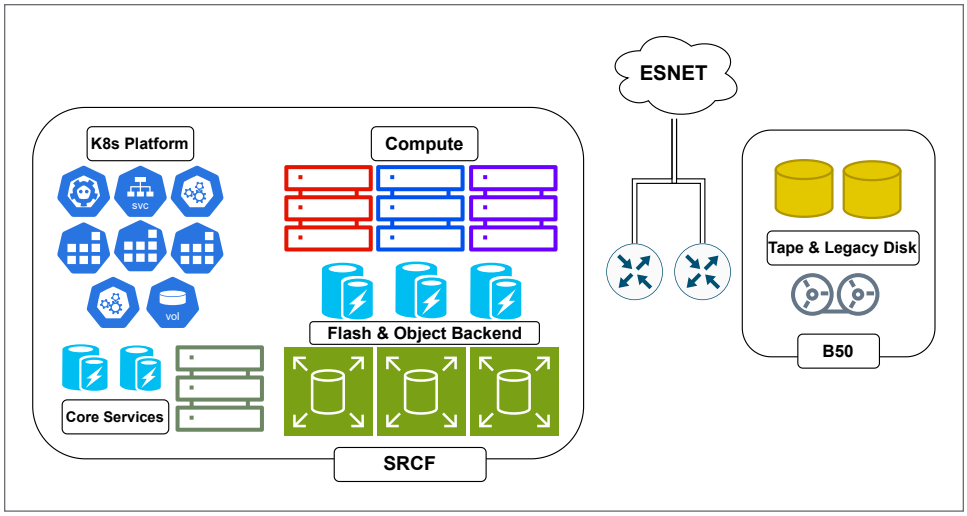


Figure 7.6.2: S3DF High-Level Architecture

7.6.2.3.1 Interactive Computing Resources

The following table represents resource allocations within S3DF for different classes of interactive computing.

Pool name	Facility	Resources
IANA	For all S3DF users	4 servers, 40 HT cores and 384 GB per server
RUBIN-DEVL	Rubin	4 servers, 128 cores and 512 GB per server
PSANA	LCLS	4 servers, 40 hyper-threading (HT) cores and 384 GB per server
FERMI-DEVL	Fermi	1 server, 64 HT cores and 512 GB per server
FADERS	FADERS	1 server, 128 HT cores and 512 GB per server
LDMX	LDMX	1 server, 128 HT cores and 512 GB per server
AD	AD	3 servers, 128 HT cores and 512 GB per server
EPPTHEORY	EPPTheory	2 servers, 128 HT cores and 512 GB per server
CDMS	SuperCDMS	(points to iana)
SUNCAT	SUNCAT	(points to iana)
NEUTRINO	Neutrino	(points to iana)
MLI	MLI (ML Initiative)	(points to iana)

Table 7.6.1: SLAC Interactive Computing Resources

7.6.2.3.2 Batch Computing Resources

The following table represents resource allocations within S3DF for different classes of batch computing.

Partition name	CPU model	Usable cores per node	Usable memory per node	GPU model	GPUs per node	Local scratch	Number of nodes
ROMA	Rome 7702	120	480 GB	-	-	300 GB	129
MILANO	Milan 7713	120	480 GB	-	-	300 GB	136
AMPERE	Rome 7542	112 (hyper-threaded)	952 GB	Tesla A100 (40GB)	4	14 TB	42
TURING	Intel Xeon Gold 5118	40 (hyper-threaded)	160 GB	Nvidia GeForce 2080Ti	10	300 GB	27
ADA	AMD EPYC 9454	72 (hyper-threaded)	702 GB	Nvidia L40S	10	21 TB	6

Table 7.6.2: SLAC Batch Computing Resources

7.6.2.3.3 Service Computing Resources

S3DF is currently supporting websites and web applications for several groups, programs, and experiments:

Site/Application	URL
USDF RSP	https://usdf-rsp.slac.stanford.edu
LCLS	https://pswww.slac.stanford.edu
SUPERCDMS	https://supercdms-dev.slac.stanford.edu
CRYO-EM E-LOGBOOK	https://cryoem-logbook.slac.stanford.edu

Table 7.6.3: SLAC Service Computing Resources

7.6.2.4 Cloud Services

SLAC uses the following cloud services for mission support (infrastructure and services that are not directly part of science/HEP programs or projects).;

- Azure
- AWS

In addition, there are is considerable use for collaboration, documentation, and related efforts among both mission support and science projects:

- Google Cloud

In 2024, the bandwidth requirements to/from these services is fairly modest (a few Gbps for connections between SLAC and the provider). Some of these services are implemented with VPN tunnels, some use public addresses, via ESnet.

During the next two years, SLAC expects to move more support services into Azure and AWS, and anticipates bandwidth increase of two to four times during that time period, perhaps sporadically reaching peaks of 10 Gbps.

7.6.2.5 AI and ML

SLAC is working toward use of Yurts.AI for knowledge base and searching/cataloging of a broad variety of online information resources, including scientific papers, technical notes, SLAC provided web content, and other sources.

Individual project teams are exploring use of ML and AI technologies directed at their specific needs. Much of that effort is coalesced within the resources provided in S3DF.

7.6.2.6 Wireless Capabilities

SLAC implements WiFi throughout the facility using Cisco Access Points and Wireless LAN Controllers. Some areas of SLAC do not have WiFi coverage due to types of building construction or distance/terrain. Outdoor coverage is limited. Most use of WIFI is in 2.4G and 5G bands, although 2.4G is being phased out in some areas due to interference and noise from non-WiFi equipment. A WiFi lifecycle process implements ongoing upgrades, including WiFi-6 support and future technologies as they become available.

The primary linear beamline tunnel has WiFi coverage via remote antennas. This facilitates service operations during planned maintenance activities, when the beamline is not in operation.

Commercial cellular coverage is available in numerous areas around SLAC, but that coverage is more limited than WiFi, due to terrain and building construction. Three carriers maintain a macro site on the roof of the primary SLAC communications building (B050).

7.6.2.7 Data-Related Resource Constraints

No additional constraints were identified.

7.6.2.8 Data Mobility Endpoints

s3dfdtn.slac.stanford.edu is a load-balanced DNS name which points to a pool of dedicated DTNs. It is open to everyone with an S3DF account. Common tools like scp/sftp/rsync are available for casual data transfers. For serious large-volume data transfer, you may consider bbcp and globus.

7.6.2.8.1 bbcp

This is a high-performance multi-stream data transfer tool developed at SLAC. In its simplest form, the bbcp command line is similar to that of scp.

7.6.2.8.2 Globus

S3DF has a Globus 5 testing endpoint slac#s3df_globus5. This service is available to everyone with an S3DF account.

7.6.2.9 Outstanding Issues

No issues were identified.

7.6.2.10 Technology Profile Contributors

SLAC Representation

Mark Foster, SLAC, fosterm@slac.stanford.edu

7.7 ANL

Argonne is a multidisciplinary science and engineering research center, where talented scientists and engineers work together to answer the biggest questions facing humanity, from how to obtain affordable clean energy to protecting ourselves and our environment. Ever since we were born out of the University of Chicago's work on the Manhattan Project in the 1940s, our goal has been to make an impact — from the atomic to the human to the global scale.

The laboratory works in concert with universities, industry, and other national laboratories on questions and experiments too large for any one institution to do by itself. Through collaborations here and around the world, we strive to discover new ways to develop energy innovations through science, create novel materials molecule-by-molecule, and gain a deeper understanding of our planet, our climate, and the cosmos.

7.7.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion found in Section 8. Several of these findings and actions also appear in Sections 1, 4, and 5.

- External network connectivity to ANL is provided by ESnet, MREN, Internet2, and Internet2 Peering Exchange (I2PX). ESnet has optical transport and routers deployed on the Argonne campus in a redundant configuration supporting up to 400 gigabits per second (Gbps) connections. The connections to MREN are 2 x 100 Gbps. The border routers are connected to each other with 4 x 100 Gbps connections. The connections to the campus core network are 2 x 40 Gbps and 40 Gbps to the perimeter firewall. By the end of FY25, the perimeter firewall will be replaced with a Palo Alto PA-5450 with 100 Gbps connectivity. [Section 4.7]
- Several data-related resource constraints were identified for ANL's support of scientific efforts: [Section 5.7]
    - Within the CELS directorate, the recent upgrades to 400G WAN & Science DMZ paths which very seldom exceed 25% sustained utilization has resulted in very few concerns on data-related constraints
    - Outside of the Science DMZ, the replacement of 40G firewall paths with 100G should provide room for future growth of non-CELS scientific divisional network traffic
    - However, tooling such as perfSONAR has not been upgraded. In some cases, it was decommissioned without immediate replacement due to the CentOS7 end of life
    - Rebuilding the fleet of perfSONAR nodes throughout campus is necessary for us to proactively identify data-related constraints, however this has been a challenge to prioritize due to other initiatives
- ANL has interest in learning how they can leverage the MetrANOVA platform, similar to how perfSONAR has been adopted. [Section 5.8]

7.7.2 ANL: Technology Profile

ANL is a multidisciplinary science and engineering research center, where scientists and engineers work together to answer the biggest questions facing humanity, from how to obtain affordable clean energy to protecting ourselves and the environment.

The laboratory works in concert with universities, industry, and other national laboratories on questions and experiments too large for any one institution to do by itself. Through collaborations here and around the world, ANL strives to discover new ways to develop energy innovations through science, create novel materials molecule-by-molecule, and gain a deeper understanding of the planet, climate, and the cosmos.

7.7.2.1 Facilities Operated and Supported

ANL capabilities span 18 of the 24 institutional strengths that the DOE defines for national laboratories:

- Propel discovery through national user facilities
- Produce new understanding of the world through basic research
- Drive advances in computation and analysis to solve the toughest problems
- Shape and secure the energy infrastructure future by applying science and engineering

Partnerships and teamwork, within and outside Argonne, are integral to success in achieving these outcomes. Active multidisciplinary collaborations across the laboratory, with the University of Chicago and other leading research institutions, are a cornerstone of ANL's approach to science and engineering.

ANL supports scientific use of networking in a number of core ways:

- Argonne's CELS directorate co-manages a pair of border routers with the Business Information Services Directorate (BIS).
- CELS HPC resources (including ALCF) are behind a Juniper QFX10002-60C switch configured in a manner similar to that of the Science DMZ "Supercomputer Center Network" architecture.
- Scientific directorate HPC resources outside of CELS (Advanced Energy Technologies, Nuclear Technologies and National Security, Photon Sciences, Physical Sciences & Engineering) along with all laboratory operations IT assets are behind a high-availability pair of Cisco Firepower 9300 firewalls running Adaptive Security Appliance (ASA) software. The firewalls and campus core switches are managed by BIS, and HPC resources themselves are managed by scientific directorate IT administrators.
- Argonne's HEP division is a member of the Physical Sciences & Engineering directorate.
- Both CELS HPC resources and Physical Sciences & Engineering directorate IT resources are used for HEP work at Argonne.
- The non-CELS scientific directorate HPC clusters are typically less than six <25kVA cabinets, running a wide variety of operating systems to support user needs. Scientific directorate IT administrators typically manage all cluster + storage interconnects using a wide variety of infrastructure. The BIS networking team manages all uplinks from these small HPC deployments into the campus network switches.
- Approximately 100 buildings and 12 datacenters are interconnected in the Argonne campus via a new, robust, single-mode fiber cable plant with capacity and ability to expand.
- Most CELS directorate infrastructure is located in Argonne Building 240, with a purpose-built datacenter for HPC. The focus in this building is supporting ultra-high-density, non-redundant cabinets of scientific computing equipment.
- Most non-CELS scientific directorate infrastructure is located at Argonne Building 386, a purpose-built datacenter for mission-critical data services. The focus in this building is low to medium density, redundant cabinets of enterprise computing equipment.
- All ESnet equipment, border routers and campus core switches at the Argonne campus are co-located with the fiber plant terminating in buildings 221 and 541b.
- A small number of legacy datacenters outside of Buildings 221, 240, 541b and 386 are being used throughout campus.

Research in the High Energy Physics division is driven by the goal of understanding the fundamental constituents of matter and energy, and illuminating the ultimate nature of space and time.

In their search for scientific discovery, high energy physicists have built some of the world's largest, most complex and sophisticated devices, including powerful accelerators, massive particle detectors and sensitive cosmological survey instruments. Theorists in the field have contributed to the deep level of understanding of the natural world. Cutting-edge computing has long been associated with high energy physics, from extreme-scale number-crunching with supercomputers to handling enormous amounts of data produced by the experiments. Because of the size and complexity of many projects, which can involve many institutions, the national laboratories have a special role to play in high energy physics.

Argonne maintains two state-of-the-art facilities for high energy physics. The Argonne Wakefield Accelerator Facility[146] aims to achieve high accelerating gradients for next-generation particle accelerators. The 4 Tesla Magnet Facility reuses hospital MRI magnets for prove and detector testing under high magnetic fields. It provides benchmarking for muon experiments that will be performed at Fermilab. Additionally, the ATLAS Support Center provides computing and training facilities for members of the ATLAS collaboration.

Additionally, ANL supports a number of research groups (some of which are profiled as case studies in this review):

- The Accelerator Physics Group carries out research to enable the next generation of particle accelerators.
- The ATLAS at the LHC experiment will enable us to learn about basic universal forces and allow us to further probe possible extra dimensions of space, the unification of fundamental forces, and evidence for dark matter candidates in the universe.
- The Cosmological Physics and Advanced Computing group studies some of the most exciting fundamental science problems: dark energy, dark matter, and the nature of primordial fluctuations—uniting theorists, observers, computer and computational scientists, applied mathematicians, and statisticians.
- The Experimental Cosmology Group focuses on studies of the CMB, the relic radiation left over from the early universe. Group members work on cutting-edge sensor development and related technologies.
- The Instrumentation Research Group is involved in the development of new instrumentation, particularly detector technology, provides a common thread that links together a wide range of high energy physics projects and programs.
- Researchers at the Intensity Frontier investigate some of the rarest processes in nature, including unusual interactions of fundamental particles and subtle effects that require large data sets to observe and measure.
- The Theoretical High Energy Physics Research Group concentrates on the interplay between theory and experiment. The Argonne Theory Group is active in exploring the particle physics SM and in searching for avenues to go beyond it.

7.7.2.2 Network and Data Architecture

7.7.2.2.1 Overview

External network connectivity is provided by ESnet, MREN, Internet2, and Internet2 Peering Exchange (I2PX). ESnet recently migrated Argonne to the new ESnet6 network and has optical transport and routers deployed on the Argonne campus in a redundant configuration supporting up to 400 gigabits per second (Gbps) connections.

146 https://www.anl.gov/awa

There are two primary networking node locations distributed between the north (Building 221) and south (Building 541B) sides of the campus to provide network redundancy. Generally, all buildings and data centers have connections to each of these networking nodes. The fiber used to connect to off-site providers is diverse and exits on the east and west sides of the campus.

Argonne networks support many connection speeds. The WAN currently consists of 100 Gbps and 400 Gbps paths. The campus core network supports speeds of 1 Gbps, 10 Gbps, 40 Gbps, and 100 Gbps. Building connections are generally multiple 10-Gbps links with some 40-Gbps links, and data centers are 10 Gbps, 40 Gbps, and 100 Gbps. LAN connections within buildings are generally 1 Gbps and 10 Gbps.

7.7.2.2.2 WAN

The Argonne WAN network connectivity is provided by ESnet and MREN (Internet2) gateway, Peer Exchange (I2PX)). ESnet has optical and routing equipment on the Argonne campus. Argonne has two Juniper MX960 border routers that are split between Building 221 and Building 541B for redundancy and diversity. These will most likely be replaced in FY26 due to a significant amount of the MX960 cards going end of life in June 2026.

The Argonne connections to ESnet are 2 x 400 Gbps. The connections to MREN are 2 x 100 Gbps. The border routers are connected to each other with 4 x 100 Gbps connections. The connections to the campus core network are 2 x 40 Gbps and 40 Gbps to the perimeter firewall. By the end of FY25, the perimeter firewall will be replaced with a Palo Alto PA-5450 with 100 Gbps connectivity.

7.7.2.2.3 Science DMZ

The Argonne Science DMZ is composed of Juniper QFX series equipment in Building 240 that is connected directly to the Argonne border routers with redundant 4 x 100 Gbps links. The Science DMZ provides high-speed connectivity between scientific organizations for the exchange of data without taking the traditional path through a firewall.

Connections into the Science DMZ are a minimum of 100 Gbps with most facilities connecting at 4 x 100 Gbps. This allows collaborators to exchange data in a high-speed environment that does not affect commodity network connectivity.

7.7.2.2.4 Core

The Argonne core network is provided by two Cisco Nexus 7710s spread across Buildings 221 and Buildings 541B for redundancy. These switches offer network speeds at 1 Gbps, 10 Gbps, 40 Gbps, and 100 Gbps. The core provides connections to buildings and many of the data centers on the Argonne campus.

7.7.2.2.5 LAN

The Argonne LAN consists of the building access and aggregation switches. The Laboratory has standardized on Cisco and Aruba switches in the LAN. There is a combination of 1 Gbps, 10 Gbps, and 40 Gbps network speeds in the LAN. Desktops are connected at 1 Gbps.

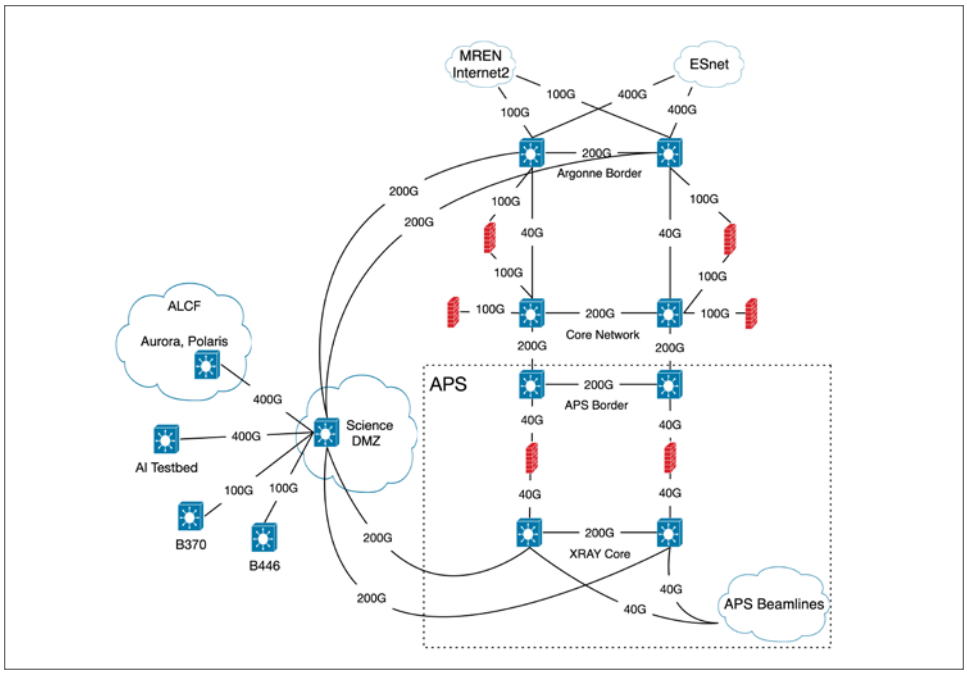


Figure 7.7.1: ANL Network Diagram

In addition to the Building 240 Science DMZ, a new redundant path was provisioned in spring 2024 to allow Advanced Photon Source (APS) DTNs direct internet access into the Argonne border routers via the redundant BIS campus core switches. This was mainly done to eliminate concerns about Building 240 datacenter losing power or cooling while APS is in the middle of a run. However, now that this work is completed, it could easily be adapted to support DTNs in the non-CELS scientific directorates throughout the ANL campus.

7.7.2.3 Computational and Storage Infrastructure

The CELS directorate integrates Argonne's research in the life sciences with the environmental sciences and the computing sciences and has established a leadership computing facility essential for advanced simulation. The mission of Argonne's CELS directorate is to enable groundbreaking scientific and technical accomplishments in areas of critical importance to the 21st century.

CELS supports two National User Facilities:

- The ALCF provides supercomputing capabilities to the scientific and engineering community to advance fundamental discovery and understanding in a range of disciplines.
- Atmospheric Radiation Measurement User Facility (ARM) is a multi-platform scientific user facility with instruments at fixed and varying locations worldwide for obtaining continuous field measurements of atmospheric data. Argonne manages the Southern Great Plains & Bankhead National Forest observatories.

7.7.2.4 Cloud Services

See the individual experiment use case studies for discussion of cloud computing resources being utilized or under consideration by individual ANL experiments and facilities. ANL provides assistance to groups that need to maintain network paths to cloud providers when applicable.

7.7.2.5 AI and ML

See the individual experiment use case studies for discussion of artificial intelligence research and resources being utilized or under consideration by individual ANL experiments and facilities.

7.7.2.6 Wireless Capabilities

Many experiments utilizing both WiFi and 5G CBRS are occurring within the Argonne campus.

Within the CELS directorate, the Waggle sensor project[147] has deployed sensors both on-site and off-campus for initiates[148]. In some cases, sensors are pre-configured at Argonne on the eduroam wireless infrastructure, and automatically join the same eduroam SSID at their final location (such as Northwestern University) seamlessly. A number of 5G projects have been launched at Argonne, including an end-to-end open-source 5G testbed.

The APS operates their own Aruba WiFi deployment for their experiment hall and directly attached Lab/Office Modules. All other WiFi is managed by the BIS directorate, consisting of Cisco Wireless LAN Controllers.

New Cisco 9800 series controllers were installed by BIS in 2024, paving the way towards WiFi 6E and 7 becoming available for general use. There is no estimated time of arrival for when 6ghz wireless will become widely available on-site due to the large number of end-of-life WiFi 5 (and earlier) access points which must be replaced first.

7.7.2.7 Data-Related Resource Constraints

Several data-related resource constraints were identified for ANL's support of scientific efforts:

- Within the CELS directorate, the recent upgrades to 400G WAN & Science DMZ paths which very seldom exceed 25% sustained utilization has resulted in very few concerns on data-related constraints
- Outside of the Science DMZ, the replacement of 40G firewall paths with 100G should provide room for future growth of non-CELS scientific divisional network traffic
- The perfSONAR infrastructure, used for network monitoring and testing, has not been upgraded in a number of year and in some cases, it was decommissioned without immediate replacement due to operating system support factors.
- Rebuilding the fleet of perfSONAR nodes throughout campus is necessary for us to proactively identify data-related constraints, however this has been a challenge to prioritize due to other initiatives

7.7.2.8 Data Mobility Endpoints

While there are a wide variety of other DTNs throughout Argonne, the lion's share of research data sharing occurs at ALCF and APS, both of which are profiled under different ESnet requirements review efforts.

7.7.2.9 Outstanding Issues

Once MetrANOVA is shipping, it would be great to see a nudge towards that platform included in these questionnaires, similar to the way perfSONAR is mentioned in today's document.

7.7.2.10 Technology Profile Contributors

ANL Representation

Tom Costello, ANL, tcostello@anl.gov

Brandon Siegel, ANL, bsiegel@anl.gov

Linda Winkler, ANL, winkler@anl.gov

147 https://github.com/waggle-sensor
148 https://crocus-urban.org/locations/

8 Group Discussion

In September 2024, an in-person review was held that invited participation from the HEP community, HEP activity host institutions, DOE SC staff spanning ASCR and HEP, and ESnet staff. The purpose of this review was to go over the case studies presented in [Section 6] and the technical profiles discussed in [Section 7], and discuss common findings and potential actions for the future support of the HEP community by ESnet.

8.1 Topic Summary

Discussion centered on four broad topic areas, heavily based on the early finding and actions identified through case study review, and the interests of the representatives present for the in-person review:

- IRI and the HEP community.
- HEP R&D efforts.
- Data challenges.
- LHCONE support and international networking.

The notes that follow are a summary of the discussions, many of which led to the drafting of additional findings and recommendations for ESnet, ASCR, and HEP.

8.1.1 IRI and the HEP Community

A common theme across many of the HEP case studies was the surrounding challenges of data access as collaboration space grows larger. This is not purely a technical problem; the LHC collaboration numbers in the thousands and has a structured ecosystem that facilitates access to groups and individuals through several tools that manage identity and permissions across a federated grid infrastructure. Other collaborations that have less structure may struggle with the available tools to manage collaboration size long term. DESI, for example, has users from around the world attempting to use resources at NERSC, and each will have to apply and be granted access to NERSC systems; access may not be allowed due to political factors such as citizenship and country of origin. A potential way to address this later concern is brokering through neutral sites like cloud providers; the RSP approach[149] offers a layer of policy and identity between the users and the collaboration systems directly managing the data. The approach of using a cloud-based front-end may see more consideration in future collaborations that span international boundaries, and may be present for future work within the IRI effort as they define standards and approaches to address federated identity and permissions that span facility boundaries.

The DOE IRI effort has identified "cybersecurity and federated access practice" as one of the six core practice areas for future work. This practice area will focus on creating novel solutions that enable seamless scientific collaboration within a secure and trusted IRI ecosystem. The discussions within this effort are nascent, but a core understanding will be recognizing that the security posture and policies of participating facilities may differ, and there must be a way to determine and agree on appropriate mechanisms to implement identity and access control.

The discussions within the HEP community touched on factors that should be considered by the IRI effort. A first step will be to establish a hierarchy of needs: an example being that the ability to access a set of resources or tools that spans administrative boundaries is very different than the ability to use different tools and resources within the same location. A similar policy language and control structure could solve both if there are ways to differentiate the needs and articulate the required control structures.

149 https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13101/131012B/Rubins-hybrid-on-premises-cloud-data-access-center/10.1117/12.3018005.short#_=_

Additionally, it must be acknowledged that authentication, authorization, and accounting approaches have limitations. Typically, it is possible to specify who can do what and keep a record of what has happened. Within an organizational structure this is not too much of a challenge, but across organizations with different policies this can be complicated as different implementations and validation methods are employed.

Opportunities exist to develop potentially new, and disruptive, models for managing access across reassures. Currently one HPC facility cannot create users, permissions, and allocation of resources for someone to access a different HPC facility. There is a natural expectation that different accounts are needed with different expectations, tools, and culture at each. Creating a policy layer that spans facilities is one matter, but ways to bridge the underlying permissions, APIs, and access methods are completely different. Compounding this is the technical debt that may be required for a facility, collaboration, or individual PI to commit to in attempting to provide or use resources.

Suggested ways forward are to define problems and categorize the relative ease toward a solution. For example, limited access to data, storage, and systems via web interfaces is much easier to manage than lower-level methods, such as system account, which allows for direct access. Existing efforts to migrate permissions across boundaries are widely accepted for some simple use cases (https://eduroam.org/), and can be built upon.

8.1.2 HEP R&D Efforts

ESnet has a long history of participation in R&D efforts with the HEP community and several ongoing efforts with the LHC. These currently include work on SENSE and Rucio, in-network caching of experimental data, network capacity testing, supporting cloud pilots, identification of science traffic, and the concept of data lakes to support scientific analysis. All these R&D efforts tie together networking capabilities to the workflow of LHC science in some meaningful way. Some are approaching a level of maturity that will see them deployed as production services in the coming years; others are pilots providing valuable information for the development of future services.

Integration of SENSE and Rucio, as described in the LHC case studies, is an effort to impart network control abilities into the Rucio workflow manager. When deployed on participating networks and sites, it allows scheduling of data-movement tasks across traffic-engineered paths, which allows for greater performance capabilities, and less impact on other traffic flows. CMS and ATLAS have been moving this capability into production at several of their Tier-2 sites.

In-network caching has gone through several successful trials, and is under active development. As discussed in the LHC case studies, questions remain as to the effectiveness of the approach, and whether it will translate into a useful production service. Early findings indicate several major differences in the operation of a cache depending on the number of users and facilities it is designed to support. For example, an "in-facility" cache that serves a small to medium number of users, many of which may be operating on similar subsets of data, can see reduced WAN download benefits. As the number of users or facilities grows, e.g., a cache deployed to service a region, the data sets analyzed rarely overlap. To see any cache benefit, the size of the cache must increase, and the policy for replacement must be altered. Caching has many benefits, but how it is deployed, and how it interacts with similar efforts such as the data lake project, are being worked out.

During DC24, ESnet directly assisted sites in "full capacity" testing of their network links. This was done using multiple perfSONAR servers to generate traffic and changing routing policy on both the network and the site end. This testing helped to reveal configuration bottlenecks and was invaluable to gain an understanding of the maximum expected performance. Future DC activities can take a similar approach, but there are also capabilities that ESnet's High Touch servers (that feature FPGAs) could assist with. High Touch servers could conceivably be programmed to create network traffic and direct it at resources located within a lab or facility perimeter. ESnet will follow up with interested sites on this prototype concept.

A number of HEP case studies outlined their production, proto-duction, and experimental use of commercial cloud resources. ESnet, as a network carrier, is identifying ways to assist with the execution of the scientific mission, and will continue to assist with connectivity and peering when applicable. During some HEP evaluations of cloud solutions for portions of analysis workflows, the network was identified to be a bottleneck in certain situations: mainly when traffic remained within a cloud versus being allowed to egress and use more predicable R&E routing and transit. In situations like this, ESnet could serve as a valuable partner between connected end sites and cloud resources, to try to force routing choices that take advantage of R&E WAN paths versus the internal networking a cloud may use.

SciTags is an initiative promoting identification of the science domains and their high-level activities at the network level. Simply stated, these are network flow and packet marking approaches that can be applied by scientific users and used by networks to better understand usage patterns. The approach is based on a network-tagging mechanism that marks network packets and/or network flows using the science domain and activity fields. The current state of use for SciTags is purely voluntary; end users can adopt this approach for their own traffic flows, and participating networks may choose to add the information on data they observe and are able to identify. Some networks may choose to act on this information for either monitoring or traffic prioritization, but are not required to do so. The effort is active, and ESnet is participating with the LHC community on several demonstrations and pilot efforts (including demonstrations at the SC conference).

The concept of a data lake, a method to access experimental data that abstracts away the locations and technologies that will serve the data, has been prototyped within the LHC community. The concept is tied to the notion of an AF, which offers capabilities to perform user analysis of data in different locations around the world. Tools such as Rucio are used to manage the overall workflow, but the data lake abstraction facilitates retrieval of the underlying dataset transparently for the user, and imparts some level of caching such that the data set could be reused if it remains popular for other users of the same AF. To date these have been deployed within LHC resources, and not on networks like ESnet. The R&D cycle for this effort is high, and many questions remain on implementation: how the data lake interfaces with storage, what metrics are used to optimize retrieval, and how the network itself plays into requests to move data. ESnet has no immediate follow-up on this R&D, but will be available for future collaboration.

8.1.3 Data Challenges

The LHC community and ESnet are collecting an enormous amount of data on the performance of data mobility tools, the networks that are a part of data transfer, and the individual hardware components (routers, switches, servers, storage). These data were incredibly useful during the DC24 activity, and helped to identify transfer bottlenecks and other gaps to delivering the requirements for HL-LHC operation. Despite having visibility, reviewing and acting on all the collected telemetry data is not possible. Cultivating or hiring a data analyst to review the data and suggest ways they can be better incorporated into future tools would benefit both efforts.

DC24 provided critical insight into the behaviors of all layers of the technology stack for the LHC communities: data management and ability software, computation and storage, as well as local and wide area networking. The previously planned DC26 challenges may slip due to the changing schedules of the LHC, but it is expected that this activity will also be a crucial preparation step for the HL-LHC operation. Beyond the major milestones of a full DC, the mini-DC activities[150] in the run-up to a full DC will be useful ways to validate specific portions of technology support structure. ESnet plans to engage with the LHC community on ways to participate in all these.

ESnet requested additional clarifications on the organized, and unorganized, nature of DC activities within the HEP community. Full data challenges will also be coordinated, and could involve more than just the core LHC experiments: affiliated efforts like DUNE and Belle II may choose to participate as well given the overlap in operational technology they share with the LHC experiments. Mini-DC activities are also organized, but to a lesser extent. Any site can choose to run DC activities (alone, or with partners) with the intention of making small

150 https://indico.cern.ch/event/1374962/contributions/5778763/attachments/2921661/5128230/Mini-Challenges_%20Building%20from%20DC24.pdf

incremental changes to hardware, software, or network configuration and analyzing the resulting performance. For example, ATLAS will be performing US-based load testing at the end of 2024 between BNL and the Tier-2 sites. This will be done on an individual basis first, and then to all simultaneously. ATLAS within Europe will do the same shortly after. The ISIS HEP effort will conduct a series of tests in early 2025. Each test has a set of deliverables and milestones that lead into the preparation for the larger DC efforts.

Uncertainly remains regarding the schedule of the next DC. Tentatively this was set for the end of 2026, but the LHC operational schedule has shifted and the LHC DOMA group will be meeting at the end of 2024 to adjust schedules accordingly. Once the dates are decided, an effort will be required within the greater US R&E community to prepare as we did in the run-up to DC 24. ESnet will request some visibility of this topic in the upcoming community meetings (The Quilt, Internet2, etc.).

Lastly, the changes to hardware at Tier-1 and Tier-2 sites will impact some of the performance aspects of future DC activities, but the schedule of hardware upgrades is hard to predict. Some sites plan on upgrading hardware yearly, so their upgrades appear more "rolling"; when a DC happens they will use what is available. Others may be able to schedule larger hardware purchases and plan to install and test prior to a large DC activity. Many sites will not get funding until closer to the DC (e.g., end of 2026, start of 2027); thus they may not be able to upgrade anything before the next major challenge.

8.1.4 LHCONE Support and International Networking

A number of facilities are not connected to LHCONE currently, despite being heavily involved in LHC-affiliated computation and storage activities. These include major computational sites (e.g., NERSC, and TACC) used for simulation workflows, SLAC (which has operated AFs in the past), and when applicable, commercial clouds. No formal agreement with any of these facilities exists now to be part of the LHC collaboration in a structured way, but investigating their ability to connect to LHCONE to support a number of important use cases is recommended.

LHCONE runs a hybrid approach to supporting IPv6 currently. The upcoming government mandate (e.g., M-21-07) will imply that Fermilab and BNL will move toward IPv6 capabilities, and subsequently the Tier-2s will have to move toward dual stack support at a minimum. One critical issue that cannot be solved by the US community will be CERN's ability to deliver IPv6 capabilities from its service infrastructure; some capabilities may be IPv4 only in the coming years. ESnet will continue to work with US collaborators to ease the transition to IPv6 operation.

Several international networking partners factored heavily into the discussions this year; connectivity to Europe, South America, and the Asia-Pacific region is accomplished through relationships with a number of R&E partners. ESnet maintains peering relations with most of these partners, and their connectivity carries significant amounts of HEP traffic daily (e.g., LHC and ProtoDUNE traffic from Europe, Rubin Observatory and CMB-S4 traffic from South America, Belle II traffic from Japan, etc.). ESnet will include current network capability diagrams into this report to document the current state of international networking.

8.2 Outcomes

The following discussion points, many of which are core findings and actions of this report and denoted in [Section 4] and [Section 5], were created from the discussions in [Section 8.1]. These observations will help to drive future collaboration.

8.2.1 Facility Management and Readiness

- A common problem across several HEP efforts is the management of credentials to access computation and storage across facilities (e.g., DOE HPC centers), and how these may differ from accounts used to manage other aspects of the scientific workflow (e.g., home institutional accounts).

- Several facilities are not connected to LHCONE currently, despite being heavily involved in LHC-affiliated computation and storage activities. These include major computational sites (e.g., NERSC, and TACC) used for simulation workflows, SLAC (which has operated AFs in the past), and when applicable, commercial clouds. No formal agreement with any of these facilities exists now to be part of the LHC collaboration in a structured way, but investigating their ability to connect to LHCONE to support several important use cases is recommended.

8.2.2 Scientific Data Management

- A common theme across many of the HEP case studies was the surrounding challenges of data access as collaboration space grows larger. The LHC collaboration has a highly structured ecosystem that facilitates access to groups, and individuals, through a number of tools that manage identity and permissions across a federated grid infrastructure. Other collaborations that have less structure may struggle with the available tools to manage collaboration size long term and are attempting other solutions (e.g., neutral site hosting through cloud providers, or adopting the policies of the strictest member of a collaboration). [Section 5.2]
- The DC24 exercise provided critical insight into the behaviors of all layers of the technology stack for the LHC communities: data management and ability software, computation and storage, as well as local and wide area networking. The DC27 is expected to be a crucial preparation step for the HL-LHC operation. Beyond the major milestones of a full DC, the mini-DC activities in the run-up to a full DC will be useful ways to validate specific portions of technology support structure. ESnet plans to engage with the LHC community on ways to participate in all these. [Section 5.2]

8.2.3 Scientific Workflow

- Several HEP experiments and facilities have adopted a core set of tools for data and workflow management, including Rucio, FTS, and XRootD. Adoption of a common set of tools will greatly increase the ability to make the analysis work portable across available resources, particularly at DOE HPC centers. [Section 5.3]

8.2.4 Computational and Storage Requirements

- The barrier to using DOE HPC resources within the LHC community remains integrating the systems as resources into the software components, and normalizing the way the resources could be accounted for (expected versus ad hoc or opportunistic). More specifically, there is a desire to create agreements or other arrangements, similar to an SLA, with DOE HPC facilities such that resources can be considered pledged. This helps with planning for both the HEP community and the ASCR resources. Some of this may require discussion at the programmatic level. [Section 5.4]
- The LHC community experiences friction with supporting a streaming-based workflow at large HPC facilities. Workarounds have been established to support this: either by allowing the worker nodes stream directly, or by leveraging caches and data lakes that are deployed close topologically. All these design choices come with trade-offs, which implies a set of R&D activities that will extend in the coming years. [Section 5.4]

8.2.5 Remote Collaboration and Operational Requirements

- Removing the need to worry about data locality allows researchers focused on developing cosmological simulation to invest more time on the science, versus spending energy in activities to stage data and allocate technology resources. It also prevents the half-measures being used today to create mechanisms to perform analysis on partial data sets, or adopt streaming approaches that are not always successful on larger computational resources not designed to support that mode of operation. [Section 8], [Section 4.5]

- Several HEP experiments rely on interactive computing methods, namely Jupyter Notebooks. It is desirable to be able to leverage computing options that may not be tied to a single facility, such as a cloud platform (like those that the Rubin Observatory has adopted). [Section 4.5]

8.2.6 Multifacility Computational Workflows

- The HEP community is interested in moving a model that would help to reduce the number of times and locations that need to be involved to archive long-term data products, and is looking for assistance in scoping and realizing this design goal. Establishing permanent storage facilities would allow the community to establish a more regular pattern of collaboration; a smaller set of data portals that can be used to catalog and retrieve long-term campaigns of products along with a simplified way to support analysis activities, potentially leveraging interactive computing approaches, at a well-known number of sites.

8.2.7 Networking for Local and Wide-area Data Mobility

- During DC24, ESnet directly assisted several sites in "full capacity" testing of their network links. This was done using multiple perfSONAR servers to generate traffic, and changing routing policy on both the network and the site end. This testing helped to reveal configuration bottlenecks, and was invaluable in gaining an understanding of the maximum expected performance. Future DC activities can take a similar approach, but there are also capabilities that ESnet's High Touch servers (that feature FPGAs) could assist with. Several High Touch servers could conceivably be programmed to create network traffic and direct it at resources located within a lab or facility perimeter. ESnet will follow up with interested sites on this prototype concept. [Section 5.7]
- The LHC community and ESnet are collecting an enormous amount of data on the performance of data mobility tools, the networks that are a part of data transfer, and the individual hardware components (routers, switches, servers, storage). These data were incredibly useful during the DC24 activity, and helped to identify transfer bottlenecks and other gaps to delivering the requirements for HL-LHC operation. Despite having visibility, reviewing and acting on all the collected telemetry data is not possible. Cultivating or hiring a data analyst to review the data and suggest ways they can be better incorporated into future tools would benefit both efforts. [Section 5.7]
- LHCONE runs a hybrid approach to supporting IPv6 currently. The upcoming government mandate (e.g., M-21-07) will imply that Fermilab and BNL will move toward IPv6 capabilities, and subsequently the Tier-2s will have to move toward dual stack support at a minimum. One critical issue that cannot be solved by the US community is CERN's ability to deliver IPv6 capabilities from its service infrastructure; some capabilities may still be IPv4 only in the coming years. ESnet will continue to work with US collaborators on this to ease the transition to IPv6 operation. [Section 5.7]
- Several international networking partners factored heavily into the discussions this year; connectivity to Europe, South America, and the Asia-Pacific region is accomplished through relationships with several R&E partners. ESnet maintains peering relations with most of these partners, and their connectivity carries significant amounts of HEP traffic daily (e.g., LHC and ProtoDUNE traffic from Europe, Rubin Observatory and CMB-S4 traffic from South America, Belle II traffic from Japan, etc.). ESnet will include current network capability diagrams into this report to document the current state of international networking. [Section 5.7]

8.2.8 Emerging Needs

- Commercial cloud resources are an intriguing addition to simulation workflows. While the cost of using these resources remains high, there may become a way to augment existing resources during periods of heavy use and justify the increased expenditure. One could imagine a situation where the use of AI/ML frameworks becomes an instantaneous source of large-scale data; if difficulty exists in allocating local computational and storage resources to meet the demand, these workflows may need to be integrated with commercial cloud infrastructure that can expand as needed to fit the needs.
- The use of in-network caching has had some impact on network use for the LHC community. Much of the R&D efforts in the recent years have been determining where caches may be beneficial for placement, how large they need to be to be effective, and what strategy should be used to store and replace data sets. Questions remain about deployment strategies to work out for both CMS and ATLAS. Regional caches were found to not be as effective (since the overlap of data use was not high enough), but caches within a facility, and ensuring that the workflow management tools know to check the caches, produced the most reliable results. [Section 4.8]
- ESnet will continue to work with the LHC community on co-design efforts like SENSE and Rucio, SciTags, and other potential opportunities. The experiments need to be specific in what they are hoping to achieve, however. The aim of joint projects could be to co-design an effective system that makes best use of the available network resources, coordinated with the site resources, and manage joint site and network operations cooperatively. This needs to be spelled out in more detail, and discussed with ESnet leadership. [Section 5.8]
- GPU use to support some aspects of AI and ML is expected to continue; what is unknown currently is how this may impact WAN usage. Experiments have shown that bursting to a commercial cloud to utilize GPUs is possible, but leveraging DOE HPC resource would allow better control over the WAN components. [Section 5.8]
- The LHC work on data lakes will continue, and several things have been learned from the years of R&D and use during activities like DC24. There is now an effort to reduce the number of "entry points" into storage resources. The XRootD tool now is not bound by factors such as network capacity or distance/latency, and may find nonoptimal answers to requests for data from clients. The overall goal of a data lake should be to simplify the way data are accessed, along with providing the most performant ways to access and retrieve them. Data lakes can also leverage some of the work on caching, and allow the caches (which often do have more intelligence) to assist in making optimal choices about retrieval. [Section 5.8]
- Authentication, authorization, and accounting approaches have limitations. Typically, it is possible to specify who can do what and keep a record of what has happened. Within an organizational structure this is not too much of a challenge, but across organizations with different policies this can be complicated, as different implementations and validation methods are employed. [Section 5.8]
- Several HEP case studies outlined their production, proto-duction, and experimental use of commercial cloud resources. ESnet, as a network carrier, is identifying ways to assist with the execution of the scientific mission, and will continue to help with connectivity and peering when applicable. During some HEP evaluations of cloud solutions for portions of analysis workflows, the network was identified to be a bottleneck in certain situations: mainly when traffic remained within a cloud versus being allowed to egress and use more predicable R&E routing and transit. In situations like this, ESnet could serve as a valuable partner between connected end sites and cloud resources to try to force routing choices that take advantage of R&E WAN paths versus the internal networking a cloud may use. [Section 5.8]

8.2.9 IRI Responsiveness

- Several HEP experiments and facilitates could take advantage of some aspects of the IRI design approaches: namely the use of DOE HPC facilities for aspects of their simulation and analysis workflows. GPU resources are expected to be incredibly valuable as the role of AI and ML approaches increases. Additionally, the announcement to create an HPDF may serve as a focal point for a long-term location for the storage of public data archives.
- The DOE IRI effort has identified cybersecurity and federated access practice as one of the six core practice areas for future work. This practice area will focus on creating novel solutions that enable seamless scientific collaboration within a secure and trusted IRI ecosystem. The discussions within this effort are nascent, but a core understanding will be to recognize that the security posture and policies of participating facilities may differ, and that there must be a way to determine and agree on appropriate mechanisms to implement identity and access control. [Section 5.9]
- The discussions within the HEP community touched on several factors that should be considered by the IRI effort. A first step will be to establish a hierarchy of needs. For example, it is very different to access a set of resources or tools that spans administrative boundaries versus having the ability to use different tools and resources within the same location. A similar policy language and control structure could solve both, if there are ways to differentiate the needs and articulate the required control structures. [Section 5.9]

List of Abbreviations

AAA	Any data, anytime, anywhere
ACE	Advanced Computing Ecosystem
ACL	Access Control Lists
ADC	ATLAS Distributed Computing
ADQL	Astronomical Data Query Language
AEON	Alert Event Observatory Network
AF	Analysis Facilities
AFS	Andrew File Service
ALCC	ASCR Leadership Computing Challenge
ALCF	Argonne Leadership Computing Facility
ALMA	Atacama Large Millimeter Array
AMD	Advanced Micro Devices
AMD	Advanced Micro Devices
AMGA	ARDA Metadata Grid Application
AMPATH	Americas Pathway
AMS	Alpha Magnetic Spectrometer
ANL	Argonne National Laboratory
ANTARES	Arizona–NOIRLab Temporal Analysis and Response to Events System
AOD	Analysis Object Data
APAN	Asia Pacific Advanced Network
APDB	Alert Production Database
APS	Advanced Photon Source
ARC	Advanced Resource Connector
ARM	Acorn RISC Machine
ASA	Adaptive Security Appliance
ASCR	Advanced Scientific Computing Research
ASN	Autonomous System Number
ATLAS	A Toroidal LHC Apparatus
ATLAS	Atlantic Research Platform
ATPESC	Argonne Training Program on Extreme-Scale Computing
AUP	Analysis Object Data
AURA	Asia Pacific Advanced Network
AVS	Alert Production Database
AWS	Advanced Photon Source
BFD	Advanced Resource Connector
BGP	Acorn RISC Machine
BHR	Adaptive Security Appliance
BIS	Advanced Scientific Computing Research
BNB	Autonomous System Number

BNL	Brookhaven National Lab
BRAHMS	Broad RAnge Hadron Magnetic Spectrometers
BSM	Beyond the Standard Model
CANARIE	Canadian Network for the Advancement of Research, Industry and Education
CCD	Charge-coupled device
CDB	Conditions database
CDMS	Cryogenic Dark Matter Search
CDN	Content Delivery Network
CDR	Conceptual design report
CELS	Computing and Life Sciences
CVMFS	CernVM Files System
CFI	Canadian Foundation for Innovation
CFN	Center for Functional Nanomaterials
CFS	Community File System
CMB	Cosmic Microwave Background
CMS	Compact Muon Solenoid
CNAF	National center of INFN
CoCB	Configuration Control Board
CONICET	National Scientific and Technical Research Council, Argentina
CPPM	Centre de Physique des Particules de Marseille
CPU	Central processing unit
CRDB	Cosmic-ray database
CRDC	Crescent Road Data Center
CRIC	Computing Resource Information Catalog
CSAID	Computational Science and Artificial Intelligence Directorate
CSCS	Swiss National Supercomputing Centre
CSDC	Community Science and Data Center
CSI	Caesium iodide
CSRG	Compact Stars Research Group
CTA	Cloud Tiering Appliance
CTIO	Cerro Tololo Inter-American Observatory
CVMFS	CERN Virtual Machine File System
CXL	Compute Express Link
DAC	Data access center
DAOD	Derived AOD
DAQ	Data acquisition
DCI	Data Center Interconnect
DD	Director's Discretionary
DES	Dark Energy Survey
DESC	Dark Energy Science Collaboration
DESI	Dark Energy Spectroscopic Instrument

DESY	Deutsches Elektronen-Synchrotron
DIRAC	Distributed Infrastructure with Remote Agent Control
DL	Distribution Layer
DM	Data Management
DMM	Data Movement Manager
DMS	Data Management System
DNS	Domain name service
DOE	Department of Energy
DPDD	Data Products Definition Document
DQM	Data Quality Monitoring
DRAC	Digital Research Alliance of Canada
DSL	Dynamic light scattering
DSN	Data streaming nodes
DTN	Data Transfer Note
DUNE	Deep Underground Neutrino Experiment
DWDM	Dense wavelength division multiplexing
EBGP	External Border Gateway Protocol
EDC	Education and Public Outreach Data Center
EFD	Engineering & Facility Database
EGI	European Grid Infrastructure
EOS	Exabyte Object Storage
EPO	Education & Public Outreach
ESA	European Space Agency
ESCC	ESnet Site Coordinators Committee
ESNET	Energy Sciences Network
EWCSS	Entanglement wedge cross sections
FaaS	Function-as-a-Service
FABRIC	FABRIC is Adaptive ProgrammaBle Research Infrastructure for Computer Science and Science Applications
FASST	Frontiers in Artificial Intelligence for Science, Security and Technology
FCC	Feynman Computing Center
FD	DUNE far detector
FHRP	First hop redundancy protocol
FIFE	FabrIc for Frontier Experiments
FITS	Flexible Image Transport System
FPGA	Field Programmable Gate Arrays
FTP	File transfer protocol
FTS	File Transfer Service
FUSE	Filesystem in Userspace
FZU	Institute of Physics of the Czech Academy of Sciences
GCC	Grid Computing Center

GCP	Google Cloud Platform
Geant4	GEometry ANd Tracking
GeV	Giga-electronvolt
GPU	Graphics processing unit
GSFC	Goddard Space Flight Center
HBM	High Bandwidth Memory
HEP	High Energy Physics
HLT	High-Level Trigger
HPC	High-performance computing
HPDF	High-Performance Data Facility
HPE	Hewlett Packard Enterprise
HPSS	High-performance storage system
HT	Hyper-threading
HTAR	HPSS Tape Archiver
HTAR	HPSS Tape Archiver
HTC	High Throughput Computing
HTSN	High Throughput Science Network
HV	High voltage
IBGP	Internal Border Gateway Protocol
ICARUS	Imaging Cosmic And Rare Underground Signals
ICTP	Abdus Salam International Centre for Theoretical Physics
IDAC	Independent Data Access Centers
IDF	Interim Data Facility
IFAE	Institut de Física d'Altes Energies
IFIBA	Instituto de Física de Buenos Aires
IFIC	Instituto de Física Corpuscular
INCITE	Innovative and Novel Computing Theory and Experiment
INFN	Italian Institute for Nuclear Physics
IO	Input/output
IP	Internet Protocol
IPST	Institute for Physical Science and Technology
IRI	Integrated Research Infrastructure
IRIS	Institute for Research and Innovation in Software for High Energy Physics
ISP	Internet service provider
ISR	Instrument signature removal
ISS	International Space Station
IT	Information technology
ITSD	Information Technology Services Directorate
JEDI	Job Execution and Definition Interface
JGN	Japan Gigabit Network
JINR	Joint Institute for Nuclear Research

JLSE	Joint Lab for Systems Evaluation
KEK	Japanese High-Energy Accelerator Research Organization
KNL	"Knights Landing" architecture for Intel CPUs
KNPO	Kitt Peak National Observatory
LACP	Link aggregation control protocol
LAG	Link aggregation group
LAMBDA	Legacy Archive for Microwave Background Data Analysis
LAN	Local area network
LArTPC	Liquid Argon Time Projection Chamber
LBNF	Long-Baseline Neutrino Facility
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership computing facilities
LCG	LHC Computing Grid
LEO	Low Earth orbit
LHC	Large Hadron Collider
LHCC	LHC Coordinating Committee
LHCONE	Large Hadron Collider Open Network Environment
LHCOPN	Large Hadron Collider Optical Private Network
LHN	Long-haul network
LINCC	LSST Interdisciplinary Network for Collaboration and Computing
LPC	LHC Physics Center
LRR	Long Range Relay
LSST	Legacy Survey of Space and Time
LSSTC	LSST Corporation
MB	Megabyte
MC	Monte Carlo
MCR	Main control room
MFA	Multi-factor authentication
MGHPCC	Massachusetts Green High Performance Computing Center
MIDAS	Maximum Integrated Data Acquisition System
MINOS	Main injector neutrino oscillation search
MIT	Massachusetts Institute of Technology
ML	Machine learning
MLAG	Multi-chassis link aggregation
MOA	Memorandum of agreement
MOE	Ministry of Education
MONIT	Monitoring for CERN IT Data Centre and WLCG Infrastructure
MOU	Memorandum of understanding
MPC	Minor Planet Center
MPLS	Multiprotocol Label Switching
MREN	Metropolitan Research & Education Network

MSFC	Marshall Space Flight Center
MSU	Michigan State University
NCCL	NVIDIA Collective Communication Library
NCCS	National Center for Computational Sciences
NCSA	National Center for Supercomputing Applications at University of Illinois at Urbana-Champaign
NeIC	Nordic e-Infrastructure Collaboration
NERSC	National Energy Research Scientific Computing Center
NESAP	NERSC Exascale Scientific Applications Program
NN	Neural networks
NOAO	National Optical Astronomy Observatory
NOvA	NuMI Off-axis ν_e Appearance
NSF	National Science Foundation
NTL	Neganov-Trofimov-Luke
OCS	Observatory Control System
ODB	Online Data Base
OLCF	Oak Ridge Leadership Computing Facility
OODS	Observatory Operations Data Service
ORION	Ontario Research and Innovation Optical Network
ORNL	Oak Ridge National Laboratory
OSCARS	On-demand Secure Circuits and Reservation System
OSG	Open Science Grid
OSN	Open Storage Network
OSPF	Open Shortest Path First
PanDA	PanDA Production and Distributed Analysis System
PB	Petabyte
PBR	Policy-based routing
PI	Principal investigators
PIREN	Pacific Islands Research
POCC	Payload Operations Control Center
PoE	Power over Ethernet
POI	Point of interest
PoP	Points-of-presence
POSIX	Portable Operating System Interface
POT	Protons-on-target
PSF	Point spread function
PVI	Processed visit image
QA	Quality assurance
QCD	Quantum chromodynamics
QCUP	Quantum Computing User Program
QoS	Quality of service

R&E	Research and education
RACF	RHIC & ATLAS Computing Facility
REED	Research Education and Economic Development Network
REUNA	Red Universitaria Nacional
RF	Radio frequency
RFI	Request for information
RHIC	Relativistic Heavy Ion Collider
RM	Resource Managers
ROADM	Reconfigurable optical add-drop multiplexer
RQ	Reduced Quantities
RSE	Rucio Storage Element
RSP	Rubin Science Platform
SAC	Science Advisory Committee
SAM	Security Account Manager
SBN	Short-Baseline Neutrino
SBND	Short-Baseline Near Detector
SC	Office of Science
SCOC	Survey Cadence Optimization Committee
SDCC	Scientific Data and Computer Center
SDN	Software-Defined Networking
SDSC	San Diego Supercomputing Center
SDSS	Sloan Digital Sky Survey
SE	Storage Element
SENSE	Software-Defined Network for End-to-end Networked Science at the Exascale
SINET	Science Information NETwork
SLA	Service Level Agreement
SLURM	Simple Linux Utility for Resource Management
SM	Standard Model
SOC	Science Operation Center
SONIC	Services for Optimized Network Inference on Coprocessors
SPT	South Pole Telescope
SSD	Solid-state drive
SSH	Su–Schrieffer–Heeger
SSID	Service set identifier
STFC	Science and Technology Facilities Council
SURF	Sanford Underground Research Facility
SUSY	Supersymmetry
SVI	Switch Virtual Interfaces
TA	Transatlantic
TACC	Texas Advanced Computing Center
TB	Terabyte

TCP	Transmission Control Protocol
TDAQ	Triggering and data acquisition
TDRS	Tracking and Data Relay Satellite
TDV	Time-Dependent Variables
TOPCAT	Tool for OPerations on Catalogues And Tables
ToR	Top of rack
TPC	Third-Party Copy
TPU	Tensor Processing Units
TriDAS	LHC's trigger and data-acquisition system
UC	University of Chicago
UCSD	University of California San Diego
UDP	User Datagram Protocol
UFJF	Universidade Federal de Juiz de Fora
UID	Unique Identifier
UKDF	UK Data Facility
UM	University of Michigan
UMN	University of Minnesota
UNL	University of Nebraska-Lincoln
US ATLAS	A Toroidal LHC ApparatuS
USDF	US Data Facility
VLAN	Virtual LANs
VO	Virtual Organization
VOMS	Virtual Organization Membership Service
VPN	Virtual private network
VRF	Virtual Routing and Forwarding
WAN	Wide-area network
WF	Workflow
WG	Working group
WLCG	Worldwide LHC Computing Grid
ZTF	Zwicky Transient Facility