SAND2017-8906PE

ATDM Data Warehouse:
Data Management Services for Exascale Computing

Craig Ulmer (PI)

Ron Oldfield (PM)

Todd Kordenbrock, Scott Levy, Jay Lofstead,
Shyamali Mukherjee, Gary Templet, Patrick Widener

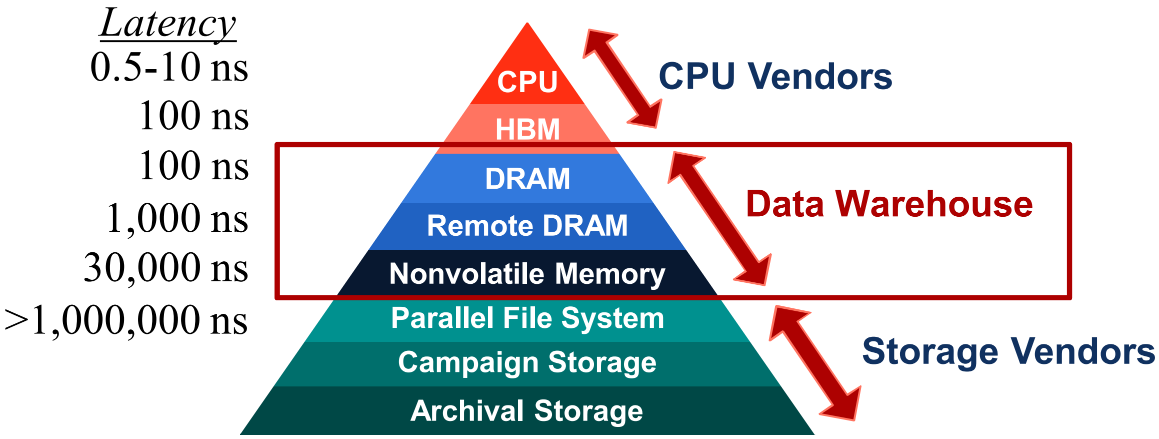
U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Overview

- Sandia relies on HPC to answer NNSA’s questions about stockpile
  - **Concern**: Current codes may not scale on Exascale computing platforms
- Advanced Technology Development and Mitigation (ATDM) (2015-2020)
  - Develop ***performance portability*** layers to insulate users from hardware
  - CS Efforts: Kokkos, DARMA, On-node Runtime, **Data Warehouse**
- *How do we migrate datasets through distributed memory resources?*
  - Develop better ***data management services***
  - ATDM and Exascale Computing Project (ECP), also relevant to HPC and Big Data

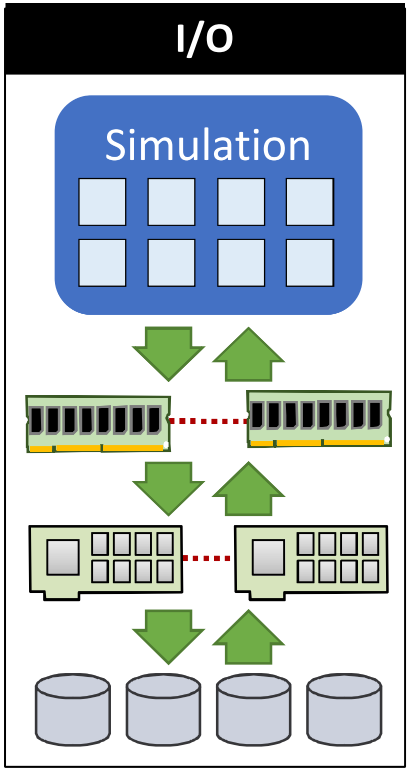


Deep Memory/Storage Hierarchy

Memory Hierarchy Critical for Production HPC

Storage is **slow**.

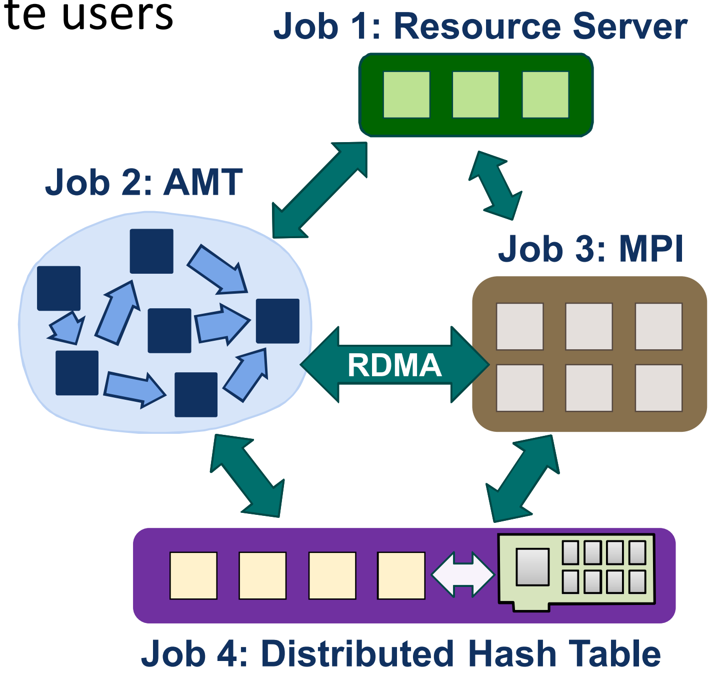
Use ***memory*** and ***nonvolatile memory*** to speed up data handoffs



Need a portable ***Data Management Layer*** to make migrations easy

Designing ***Better*** Data Management Services

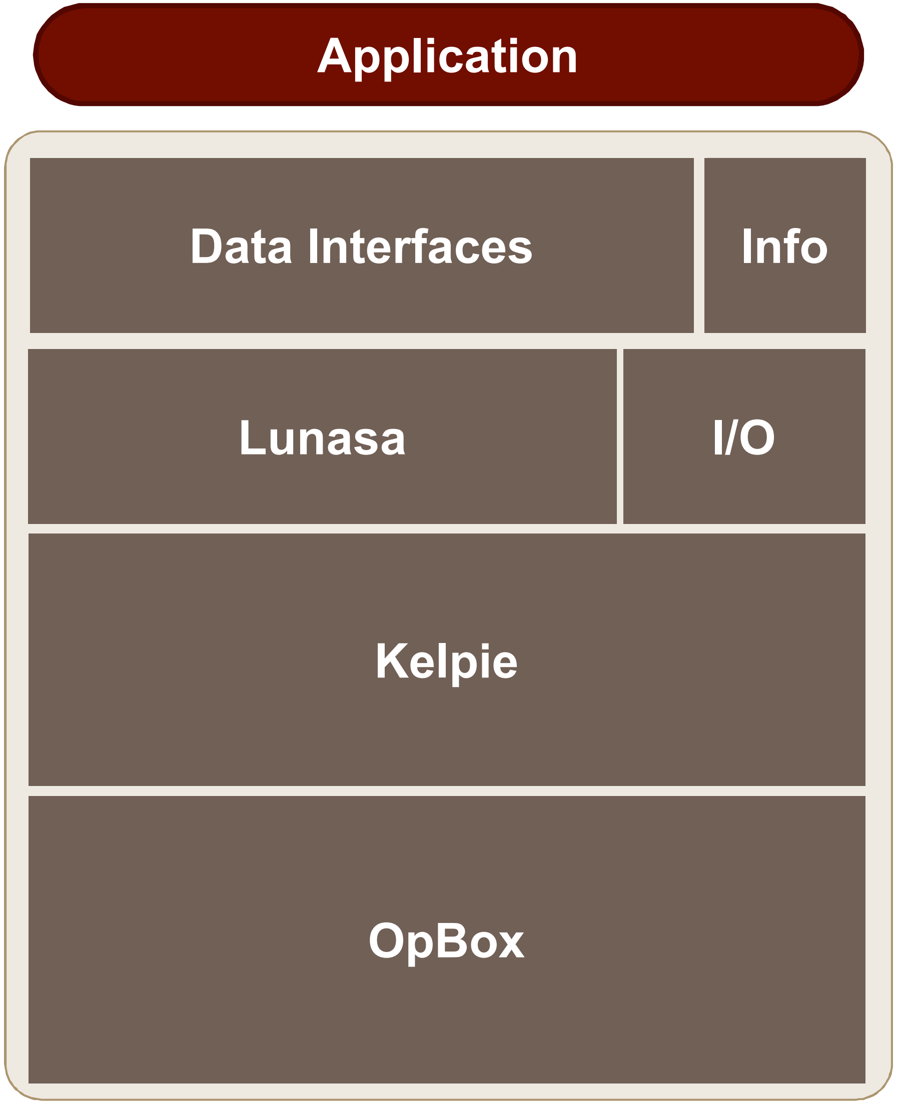
- Goals
  - Portable, Performant Data Management Services
  - Abstractions that hide complexity and insulate users
- Requirements
  - Job-to-job Communication
  - Asynchronous and Event-Driven
  - Modern C++ primitives (Lambdas, Futures)
  - Support our production environments
- ATDM Data Warehouse
- Why develop at Sandia?
  - Currently, no single solution for these problems
  - Optimize for Sandia technologies (e.g., Kokkos)



Related Work

Domain	Examples	Issues
AMT Runtimes	DARMA, Charm++, Legion, Uintah	Lack job-to-job communication Lack storage support Runtime lock-in
RDMA Libraries	GASnet, Mercury, Nessie, libfabric, UCX, Converse…	Too low-level, lack storage support Only target Client/Server Many lack job-to-job
Key/Value Stores	MDHIM, HERD, Pilaf, FaRM, RAMCloud, Accumulo, Memcached, LevelDB, …	Most are early in technical readiness (TRL-5) Some lack job-to-job or HPC support Philosophy mismatches
Code Coupling	DataSpaces, GLEAN, Catalyst	Target specific use cases (e.g., Viz) Lack I/O support Opportunities for improvement
I/O	ADIOS, HDF5/DE	Focused on making persistent I/O faster Files instead of memory abstractions

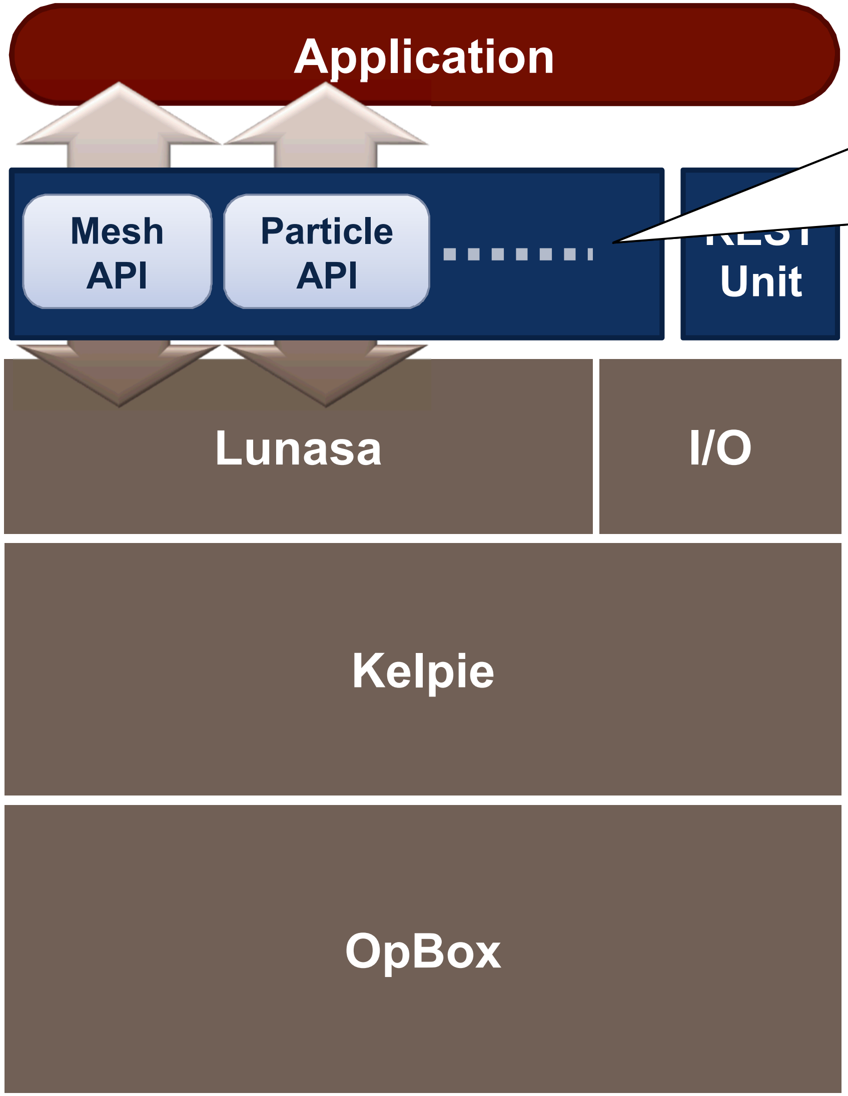
How does the Data Warehouse work?



ATDM Data Warehouse

Collection of libraries for developing data management services.

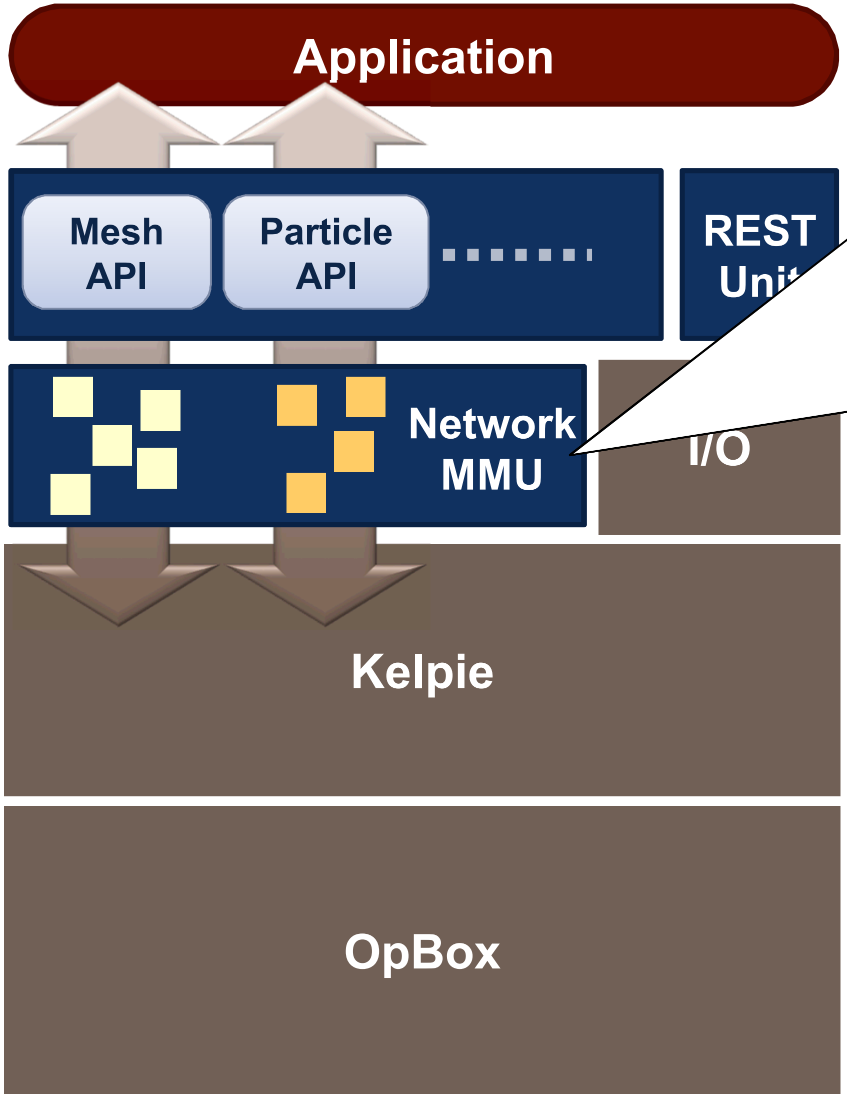
How does the Data Warehouse work?



Data Interface Modules (DIMs)

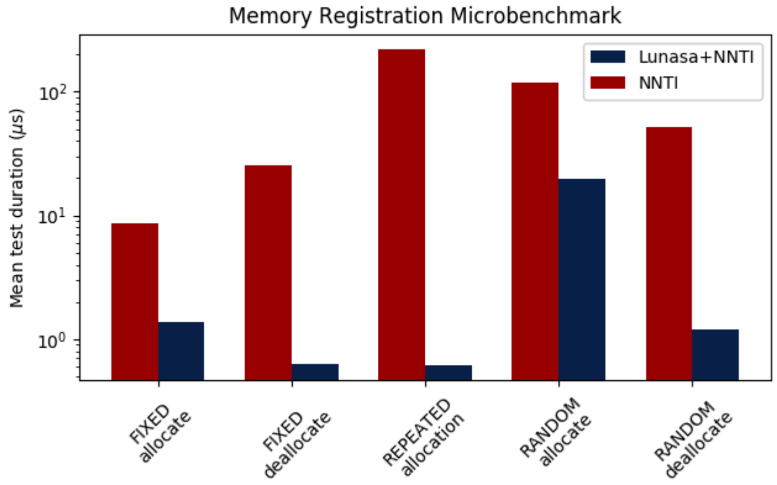
- No single API for all datasets
- Develop new modules for each dataset
- Top: Implement familiar user API
- Bottom: Data warehouse calls

How does the Data Warehouse work?

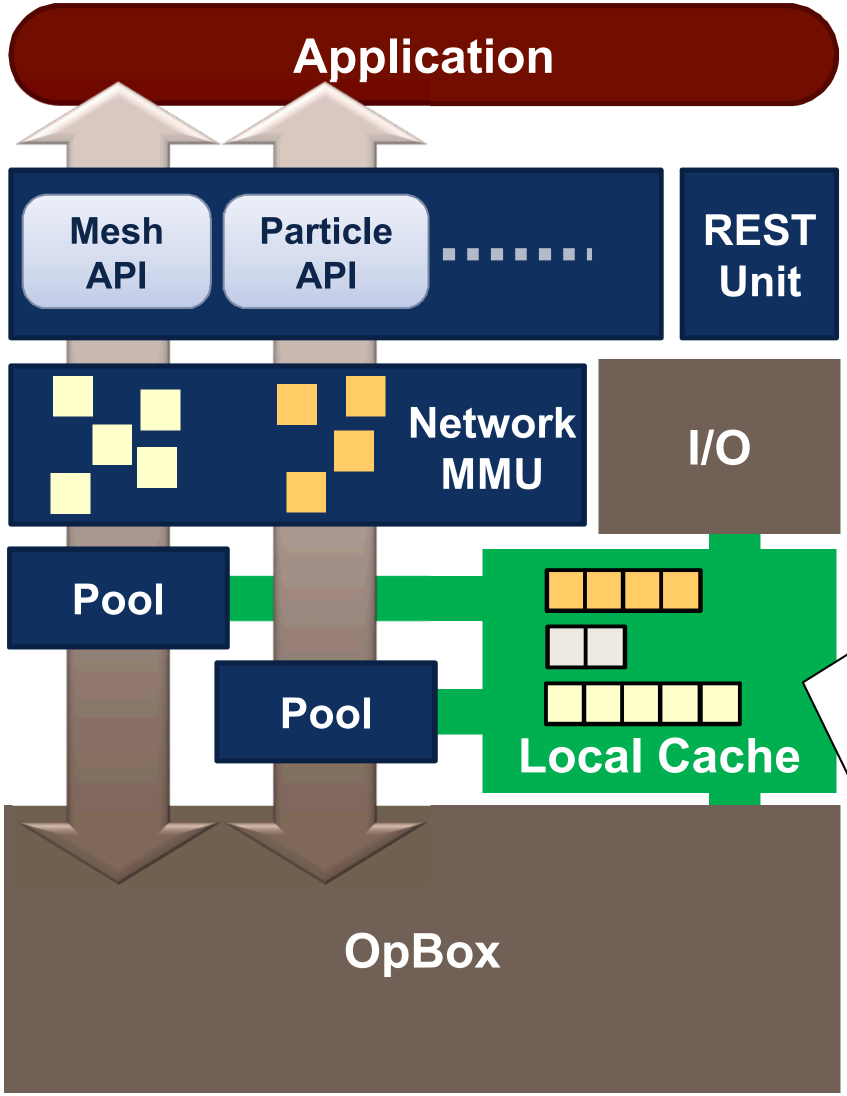


Lunasa: Network Memory Management

- Network memory requires *registration*
- Registration can be expensive
- Suballocate memory with tcmalloc

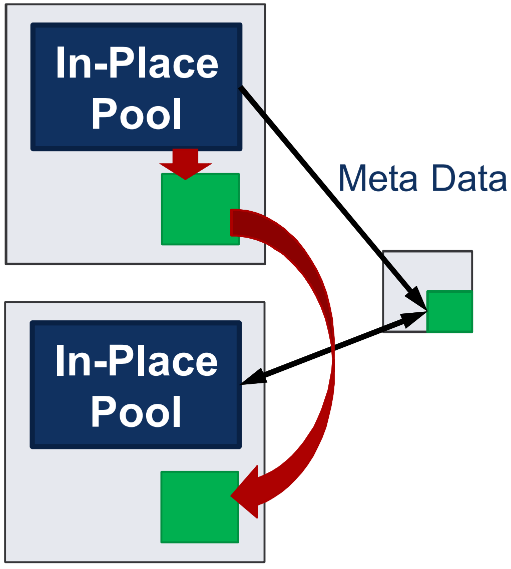


How does the Data Warehouse work?



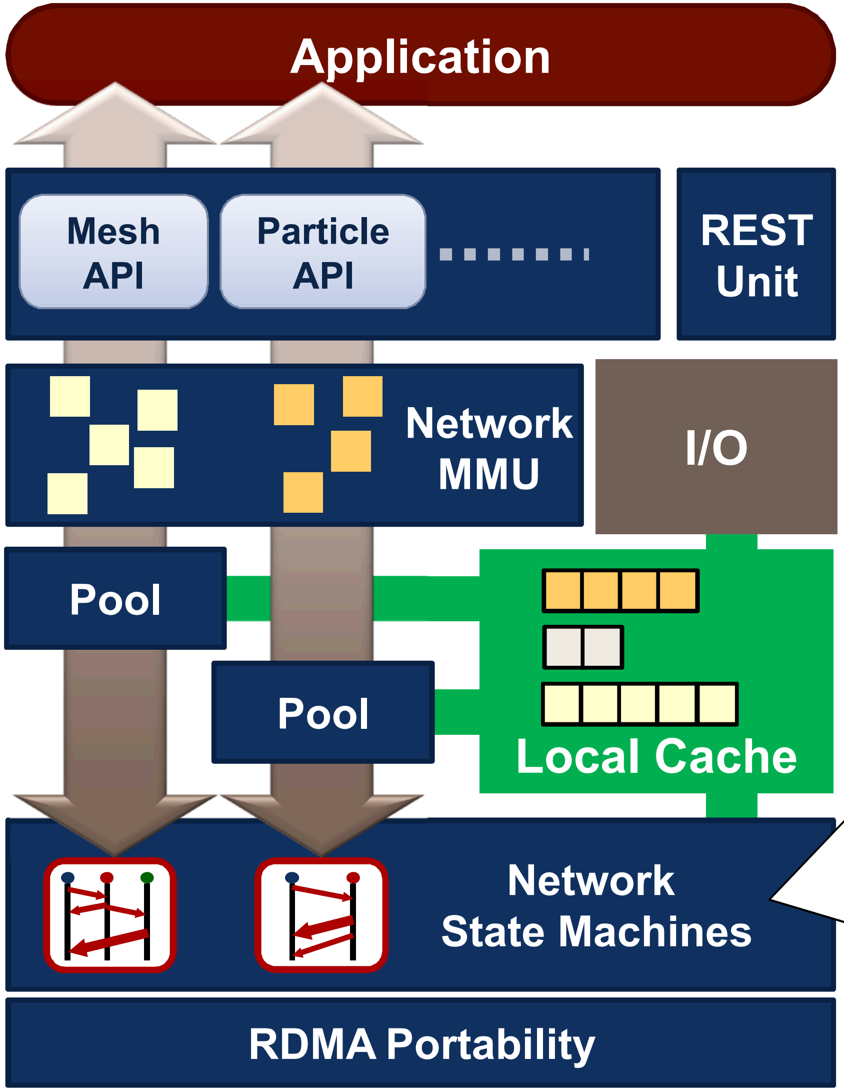
Kelpie: Distributed Key/Blob Service

- User-controlled ***Local Cache***
- Leave callbacks for objects
- "Pool" controls object distribution



In-Place Publish

How does the Data Warehouse work?



Application

Mesh API

Particle API

REST Unit

Network MMU

I/O

Pool

Pool

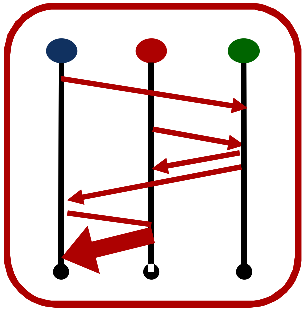
Local Cache

Network State Machines

RDMA Portability

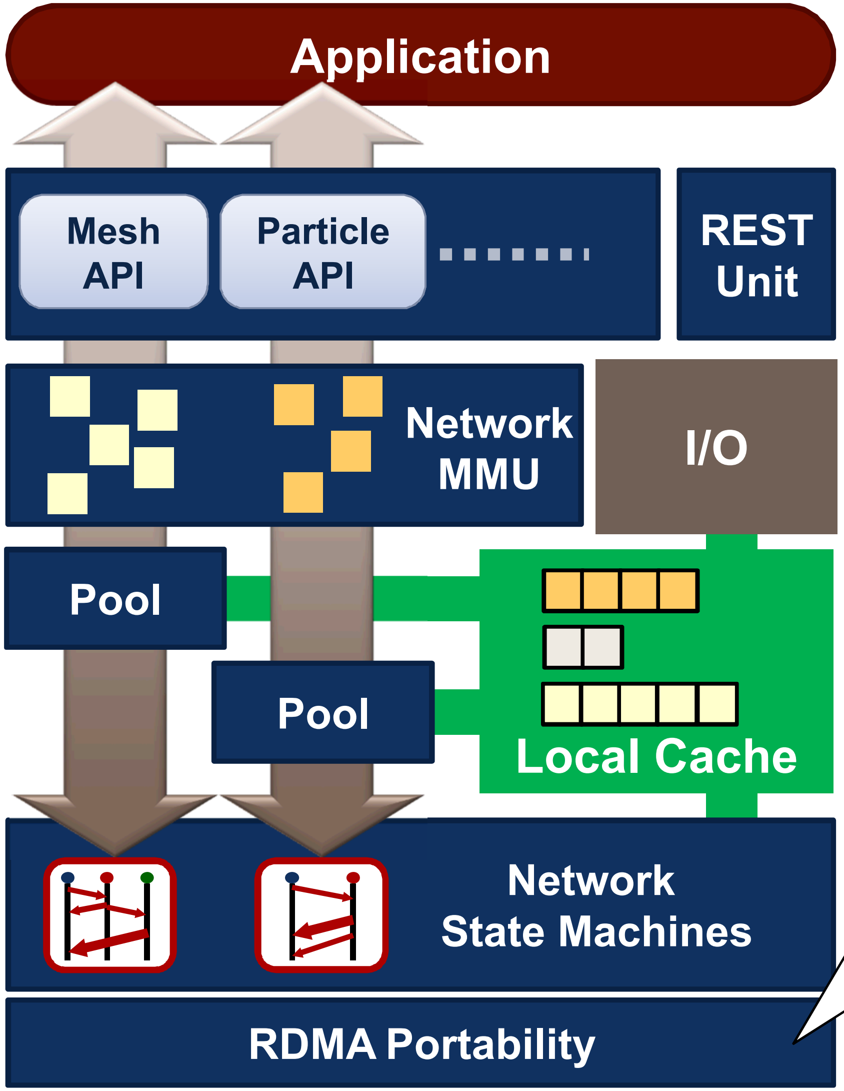
OpBox: Network State Machines

- Remote Procedure Calls insufficient
- Implement transfers in *state machines*
- More clarity, better error handling
- OpBox can make progress on Ops



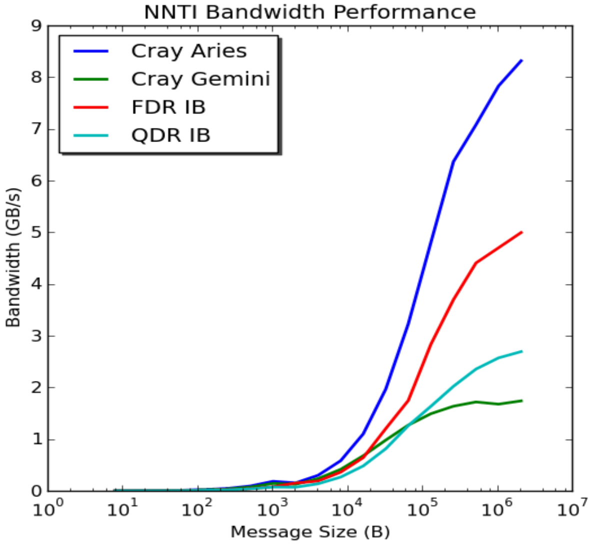
= Op

How does the Data Warehouse work?

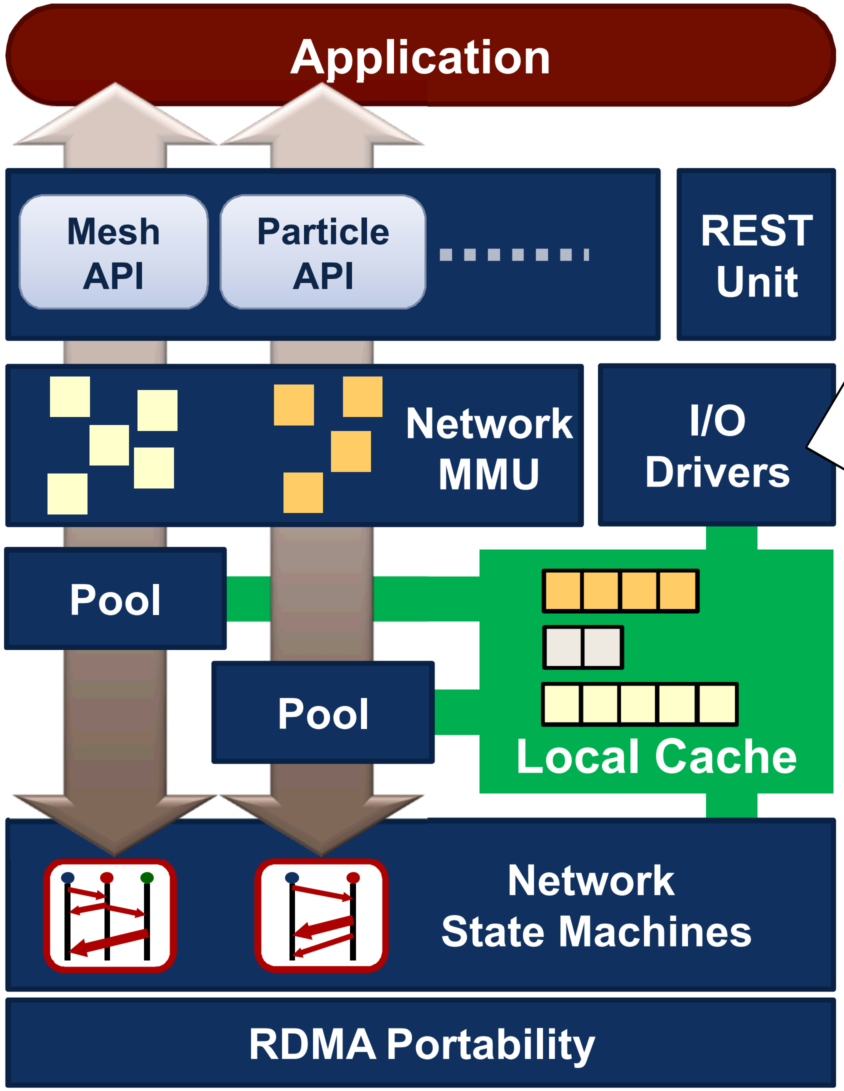


RDMA Portability

- Low-level network transfers
- Support **NNTI** (SNL)
- or libfabric (OpenFabrics)



How does the Data Warehouse work?



Application

Mesh API

Particle API

REST Unit

Network MMU

I/O Drivers

Pool

Pool

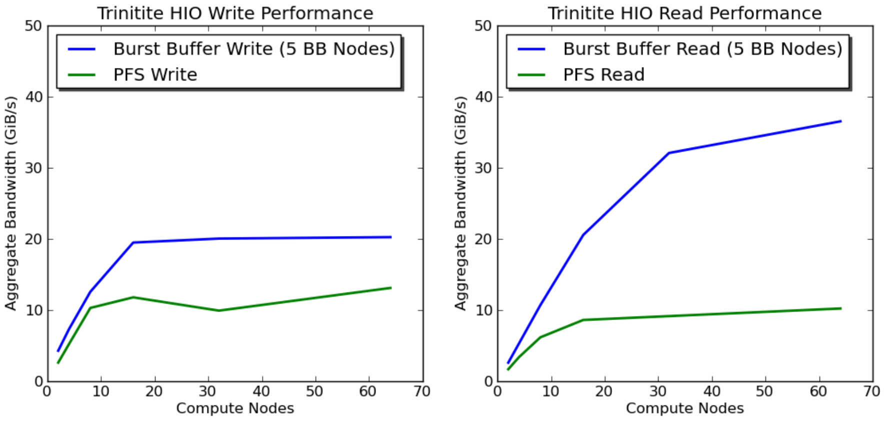
Local Cache

Network State Machines

RDMA Portability

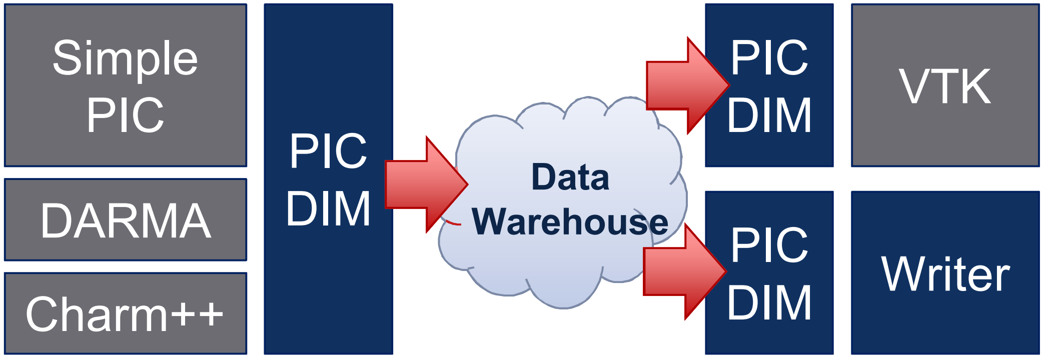
I/O Drivers

- Interface to Burst Buffers, NVMe, PFS
- Currently use **HIO** from **LANL**
- Support for XC40 DataWarp and PFS



Examples: FY17 ATDM Prototypes

EMPIRE: Particle-in-Cell (PIC)



Summary

- Data management is a critical aspect of production HPC work
  - Workflows, AMT, checkpointing, coupling, I/O
- ATDM Building data management services with new capabilities
  - Data Warehouse provides flexible libraries for new services
  - Philosophy for Impact:
    - Build custom services for high-value customers
    - Broad use by plugging into existing data libraries (IOSS)
- Status: Prototype, transitioning to Production (FY18-FY19)
  - Building support for ATDM's applications
  - Drivers for LLNL Sierra Platform's NVMe devices
  - Integration with SAW for workflows

Deliverables

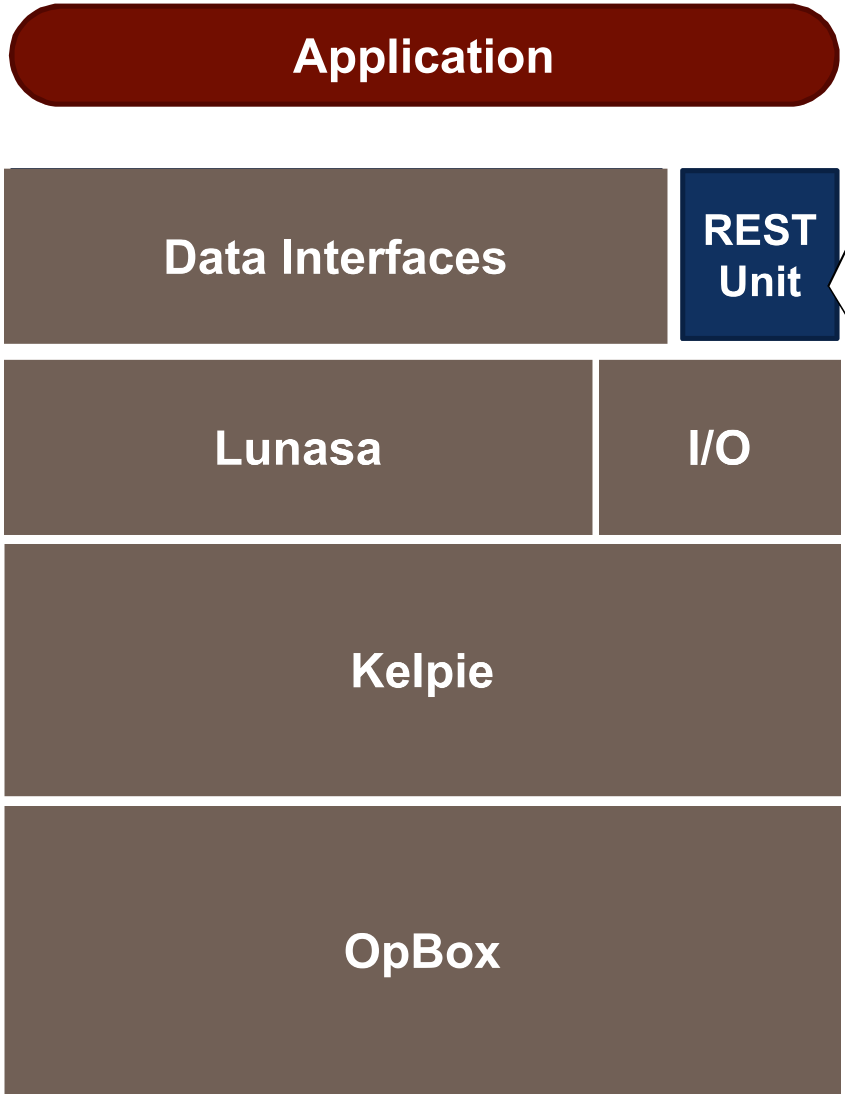
- Code: data-warehouse v1707 (Amigo)
  - Internal: https://gitlab.sandia.gov/nessie-dev/data-warehouse-release
  - External release plans: After publication and copyrights
- Design Docs
  - Data Warehouse Design: Support for ATDM Datasets
  - Data Warehouse Design: Support for AMT Requirements
- PoC: Craig Ulmer / cdulmer@sandia.gov
- Community Interactions

Date	Venue	Type
Jan 2017	ECP All Hands	Poster
Nov 2016	ECP PI Meeting	Briefing
Feb 2016	ECP/ATDM Leaders	Briefing
Feb 2016	JOWOG/LANL	Talk
Nov 2015	ASC Trilab Review/LLNL	Talk
Apr 2013	Salishan/SNL	Talk

ECP:	Exascale Computing Project
JOWOG:	US/UK Joint Working Group

ADDITIONAL INFORMATION

How does the Data Warehouse work?



Application

Data Interfaces

REST Unit

Lunasa

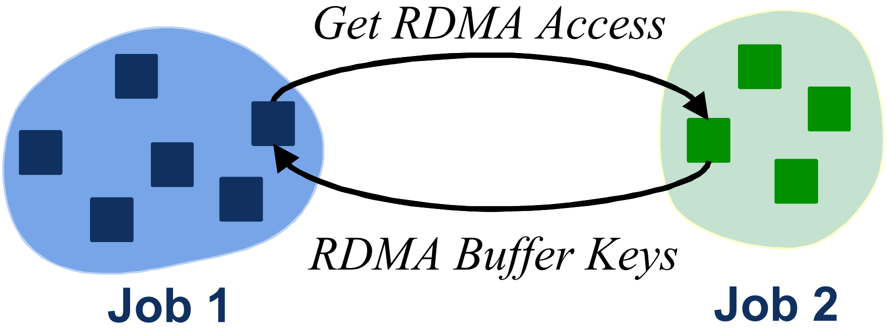
I/O

Kelpie

OpBox

REST Unit: Low-bandwidth Info

- Multipurpose http server
- Add callbacks via C++ lambdas
- Negotiate ***job-to-job*** communication
- Also, Excellent debugging interface

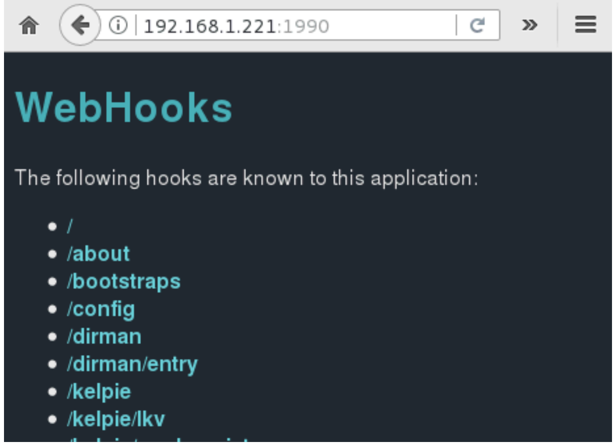


Get RDMA Access

RDMA Buffer Keys

Job 1

Job 2



WebHooks

The following hooks are known to this application:

- /
- /about
- /bootstraps
- /config
- /dirman
- /dirman/entry
- /kelpie
- /kelpie/lkv