Preparation of the Multi-Site Data Processing at the Vera C. Rubin Observatory

FERMILAB-CONF-25-0809-CSAID

Preparation of the Multi-Site Data Processing at the Vera C. Rubin Observatory

Wei Yang[2*],*Jennifer* Adelman-McCarthy[1], *Greg* Daues[4], *Richard* Dubois[2], *Wen* Guan[3], *Yuyi* Guo[1], *Michelle* Gower[4], *Fabio* Hernandez[5], *Tim* Jenness[6], *Edward* Karavakis[3], *Kian-Tat* Lim[2], *Peter* Love[7], *Timothy* Noble[8], *Stephen* Pietrowicz[4], *Brandon* White[1], *Zhaoyu* Yang[3], *Brian* Yanny[1]

[1]Fermi National Accelerator Laboratory, Kirk Road and Pine Street, Batavia, IL 60510, USA
[2]SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
[3]Brookhaven National Laboratory, Upton, NY 11973, USA
[4]NCSA, University of Illinois at Urbana-Champaign, 1205 W. Clark St. Urbana, IL 61801, USA
[5]CNRS, CC-IN2P3, 21 avenue Pierre de Coubertin, CS70202, F-69627, Villeurbanne cedex, France
[6]Vera C. Rubin Observatory Project Office, 950 N. Cherry Ave., Tucson, AZ 85719, USA
[7]Lancaster University, Lancaster, UK
[8]Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell, UK

Abstract. The Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST) Camera is scheduled to start taking data in the summer of 2025. The Data Release Production will run the LSST Science Pipe software at data facilities in the US, France and the UK. The LSST Science Pipeline consists of complex directed acyclic graphs (DAGs) of tasks. Rubin will use the Production and Distributed Analysis (PanDA) workflow and workload management system to orchestrate this complex workflow and the distribution of workloads to the data facilities. When run end-to-end by a team of data production staff, this processing (the Science Pipelines, distributed by the workflow and workload management system) is referred to as a 'campaign'. This paper describes the central services and data facility specific services that support this multi-site data process model, including the service deployment infrastructure, the workload and workflow system, the Campaign Management tools, and connection to Rubin Data Management. This paper will also mention the experience of processing the Rubin Commissioning Camera data. All these are part of the effort to scale up the processing capabilities for the expected very large data volume from the LSST Camera.

1 Introduction

The NSF-DOE Vera C. Rubin Observatory, funded by the U.S. National Science Foundation and the U.S. Department of Energy's Office of Science, including the *Legacy Survey of Space and Time* (LSST) camera which is located at a site in Cerro Pachón, Chile, is scheduled to start taking sky images in the summer of 2025. The camera is about 1.65m by 3m in size and has a resolution of 3.2-gigapixel. Its large-aperture, wide-field optical imager is designed to provide a 3.5-degree field of view, and is sensitive to light with

[*] Corresponding author e-mail: yangw@slac.stanford.edu

wavelength ranging from 0.3µm to 1µm. The observatory will take a full picture of the visible southern sky every few nights. The survey operation will continue for a period of 10 years [1,2].

The LSST camera records science image data using 189 16-megapixel silicon detectors. After each 15 to 30 second exposure, these 189 image files will be transferred within 7 seconds from Cerro Pachón, Chile, to Rubin's US Data Facility (USDF) within the SLAC Shared Science Data Facility (S3DF) at the SLAC National Accelerator Laboratory in California, USA. From there two types of processing of the camera data will happen: near real time Prompt Processing and Data Release Production (DRP).

The prompt processing system will detect minute changes in groups of exposures on the same area of sky, known as *visits*. With each visit, prompt processing will compare to previous images and a database of known sources and distribute alerts on detected transients in under a few minutes. The conceptual picture and current architecture are detailed in DMTN-219 [3] and DMTN-260 [4].

For the Data Release Production, the images will be processed in greater detail in order to make high-precision measurements on astronomical objects. These measurements will then be made available to science users for analysis. DRP will use the LSST Science Pipelines software [5] and will be processed at Rubin's data facilities in the USDF, in the French CC-IN2P3 data facility (FrDF) and in the UK data facilities (UKDF) at the University of Lancaster and Rutherford Appleton Laboratory. The DRP will be organized in campaigns. Each year, Rubin will process all previous image data with the most current version of the Science Pipelines software release in a Data Release Processing campaign. The preparation work to this complex, distributed processing at the three data facilities will be covered in later sections of this paper.

2 Service Infrastructure Supporting the DRP

Rubin plans to process 35% of the raw data at the USDF, 40% at the FrDF and 25% at the UKDF. Both USDF and FrDF will keep a full copy of the raw image data, and the UKDF plans to keep the raw image data they processed. USDF will also keep a copy of all data products. Each site has already deployed batch systems and disk storage to support the DRP. In the case of USDF and FrDF, tape storage will also be used.

2.1 DRP Service Components

Over the last decade, Rubin developed the Batch Production System (BPS) [5] and Data Butler [6] to support large-scale data processing. BPS manages batch job submission based on job dependency. Data Butler refers to both a data store (including a database to keep track of data file locations, type, grouping, etc.) and an API to access those data.

To support data processing at multiple sites, Rubin DRP adopted a distributed computing model developed by the LHC experiments. This model includes the CVMFS [7] infrastructure to manage the distribution of software releases and small amount of static data such as configuration files, software from the Rucio [8] ecosystem (including Rucio, FTS3, XRootD and dCache) for distributed data management, Grid Computing Element (CE) [9] as a gateway for remote job submissions, and the Production and Distributed Analysis (PanDA) [10] system for workflow/workload management.

Fig. 1 shows the deployed cyber infrastructure from the service component's point of view. Some of the services/components in Fig. 1 need to be deployed at all data facilities, while others are considered central services that will only need to be deployed at one facility. The PanDA system, Rucio system (including FTS3 [11]), VOMS (Virtual Organization Management System) and analytic platform OpenSearch are central services deployed at the USDF. The CVMFS Stratum 0 (and stratum-1) is a central service deployed at the FrDF, and the backup VOMS [12] and FTS3 are central services deployed at the UKDF.

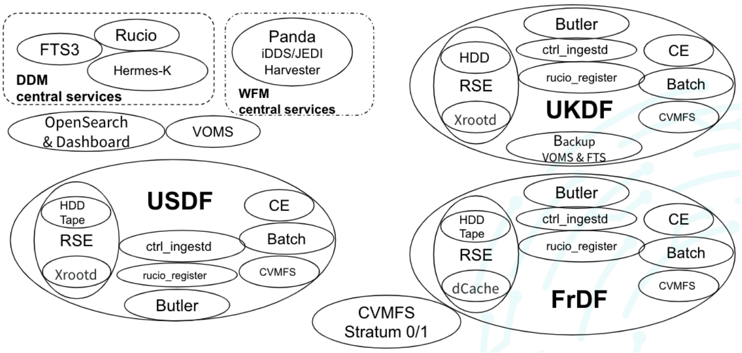


Figure 1. Components of Rubin distributed computing infrastructure

The LSST Science Pipeline features a large number of processing tasks in acyclic graphs, long and short in processing time, and somewhat hard-to-predict data-dependent memory usage. Rubin has selected the PanDA system to orchestrate this complex workflow and to manage the distributed workload. The plugin architecture of the BPS software [13] allows the PanDA system to be integrated with it as one of its workflow management service classes [14], along with other classes such as HTCondor and Parsl.

Note that both the Data Butler and Rucio are disparate data management systems, each with different properties and capabilities. Their functions and integration of them at Rubin will be presented in another paper [15]. The Grid computing model and its associated X509 infrastructure are well known technologies [9].

2.2 Kubernetes Based Service Deployment Infrastructure at the USDF

Most central services consist of several components/daemons. Many of them were initially designed to run on bare metal machines. However, at the USDF most of them are deployed in SLAC's Kubernetes infrastructure. In this environment, cloud native forms of software components are often used. For example, instead of containerizing and deploying a Postgres database on its own, the CloudNativePG (CNPG) Kubernetes operator is used to deploy databases for PanDA, Rucio and Butler. The use of Kubernetes and the associated CNPG operator provide both benefits and challenges: On one hand this cloud native deployment enables rapid database deployment, high availability, easy scaling, rolling updates and high (hardware) resource utilization. But on the other hand, it increases deployment complexity, and it makes it difficult to identify and debug issues because of the added orchestration layer between containerized applications and bare metal infrastructure.

3 Multi-site Data Release Production

After an embargo period, and when permitted by policy, raw image data at the USDF will be copied from the embargoed storage to the main USDF storage, under a space associated with the Butler repository for the raw data. The Butler's database will also be populated with the appropriate metadata. This point marks the beginning of the first step of the multi-site DRP for the raw data.

3.1 The Processing Flow of the Multi-site Data Release Production

Fig. 2 shows the flowchart of the multi-site processing from the USDF to one of the data facilities in Europe - either the FrDF or UKDF, designated below as EUDF. Similar processing flow also exists (step 2 to 7) in between the two EU data facilities. This multi-site processing is built on top of, and uses the components described in section 2.1.

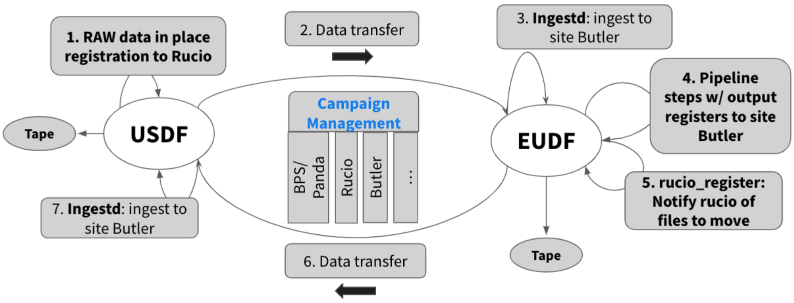

1. RAW data in place registration to Rucio
2. Data transfer
3. Ingestd: ingest to site Butler
4. Pipeline steps w/ output registers to site Butler
5. rucio_register: Notify rucio of files to move
6. Data transfer
7. Ingestd: ingest to site Butler
Tape
USDF
EUDF
Tape
Campaign Management
BPS/ Panda
Rucio
Butler
...

Figure 2. Multi-site DRP processing flowchart. The EUDF represents either the FrDF or the UKDF. The Campaign Management Tools will coordinate the processing flow.

While the movement of raw data (the step 1, 2 and 3 above) will start once the raw data leaves the USDF embargo storage, the actual processing (step 4 and later) of the data will be organized by campaigns.

3.2 Execution of the LSST Science Pipeline

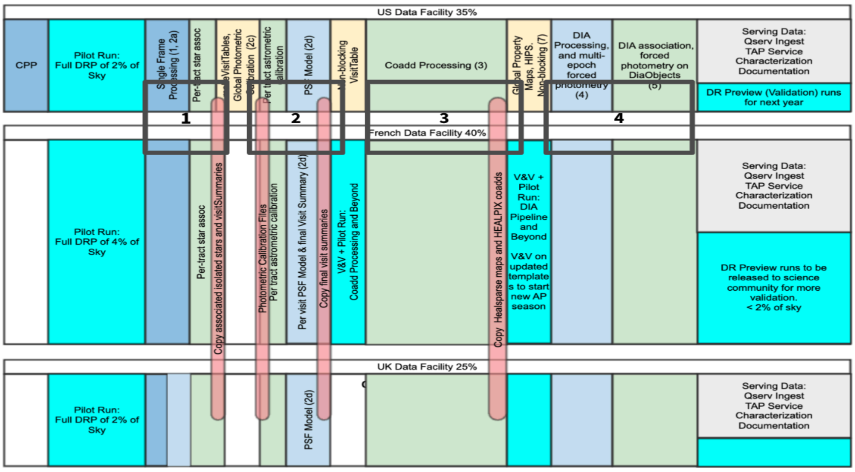

US Data Facility 35%
CPP
Pilot Run: Full DRP of 2% of Sky
Single Frame Processing (1, 2a)
Per-tract star assoc
VisitTables, Global Photometric Calibration (2c)
Per tract astrometric calibration
PSF Model (2d)
Non-blocking VisitTable
Coadd Processing (3)
Global Property Maps, HIPS, Non-blocking (7)
DIA Processing, and multi-epoch forced photometry (4)
DIA association, forced photometry on DiaObjects (5)
Serving Data: Qserv Ingest TAP Service Characterization Documentation
DR Preview (Validation) runs for next year
1
2
3
4
French Data Facility 40%
Pilot Run: Full DRP of 4% of Sky
Per-tract star assoc
Copy associated isolated stars and visitSummaries
Photometric Calibration Files
Per tract astrometric calibration
Per visit PSF Model & final Visit Summary (2d)
Copy final visit summaries
V&V + Pilot Run: Coadd Processing and Beyond
Copy Healsparse maps and HEALPIX coadds
V&V + Pilot Run: DIA Pipeline and Beyond
V&V on updated templates to start new AP season
Serving Data: Qserv Ingest TAP Service Characterization Documentation
DR Preview runs to be released to science community for more validation. < 2% of sky
UK Data Facility 25%
Pilot Run: Full DRP of 2% of Sky
PSF Model (2d)
Serving Data: Qserv Ingest TAP Service Characterization Documentation

Figure 3. LSST Science Pipeline Stages (four boxes): 1) initial, 2) recalibration, 3) coadds, 4) revisit.

Step 4 uses the LSST Science Pipelines software to process the data. It runs at all DFs, including the USDF. Step 5 and later move data products back to the USDF or to other Data Facilities (DFs).

3.2.1 Recapitulation of the LSST Science Pipeline

The pipeline consists of a long chain of tasks whose output will be used as the next task's input. These tasks are further grouped into subsets, steps, and stages. Fig. 3 shows the pipeline tasks in four high level stages

Pipeline execution is not a single linear process. It will fork and form a directed acyclic graph (DAG). In LSST, we represent these custom DAGs known as Quantum Graphs. The Quantum Graphs are generated at each data facility, based on the data to be processed there and the pipeline tasks that will be applied to the data.

3.2.2 Orchestration of the Science Pipeline Execution

Through the BPS submission interface, PanDA will orchestrate execution of pipeline tasks. As the backbone of the multi-site processing, PanDA is capable of following complex workflows and managing the execution of a high volume of distributed workload. The main components and their logical relations within the PanDA system are shown in Fig. 4. The following paragraphs give short descriptions of the components relevant to this paper.

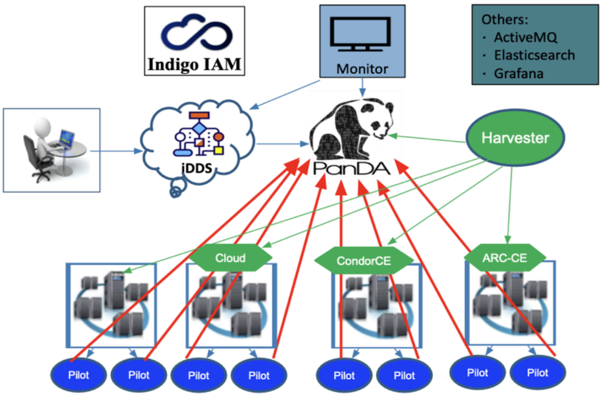


Figure 4. The components of the PanDA workflow and workload management system. The Campaign Management Tools (the user icon) submit a workflow to iDDS, which generates the corresponding tasks. Those tasks are released to the Panda server (the Panda icon) when their dependencies are met. Harvesters will submit pilot jobs to the data facilities via their ARC-CE or Condor CE, etc.). Those pilot jobs will ask the Panda Server for tasks to execute.

The main function of the iDDS (intelligent data delivery service) is to orchestrate workflows. It takes a Quantum Graph and maps it to a managed workflow of tasks. It will release a task to the PanDA Server when the task's dependency is met.

PanDA Server (the "PanDA" in Fig 4.) is a workload management system. It dispatches and manages the execution of pipeline tasks, and manages retries if failure happens.

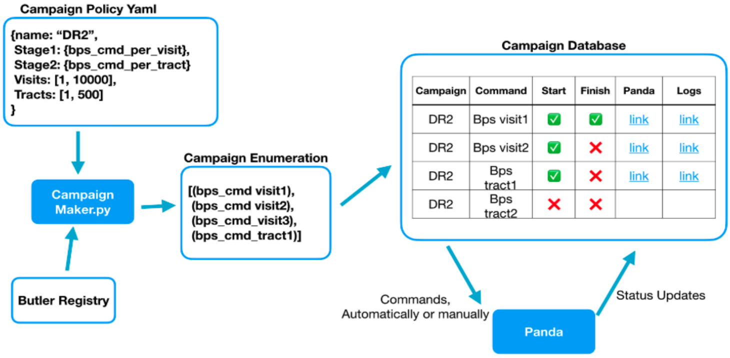
Pilots is a wrapper script around the actual payload. It is what gets run by the batch system. It fetches real payload (Science Pipeline tasks) from the Panda Server. It also handles the staging of the input and output, and reports back to PanDA the status and heartbeat of its execution.

The Harvester functions as a pilot factory. It submits pilots to data facilities' batch system, via the Grid CE or directly to the batch system (when possible).
PanDA Monitoring allows monitoring the progress of a group of identical pipeline tasks (aka, PanDA tasks), and allows diving into the logs of pilot or payload. Some of this information is also fed to an OpenSearch analytic cluster at the USDF.

3.3 Orchestrating the Data Release Production Campaign

Functioning as a command centre, Rubin Campaign Management team will plan, prepare, launch and monitor the execution of the DRP campaigns at data facilities. The Campaign Management team essentially needs to coordinate all the steps in the multi-site processing flowchart (Fig. 2) – not just pipeline processing, but also data (pre)placement and data movement.

When it comes to the pipeline processing, the Campaign Management team has its own tools for automation, as shown in Fig. 5.



```
{name: “DR2”,
 Stage1: {bps_cmd_per_visit},
 Stage2: {bps_cmd_per_tract}
 Visits: [1, 10000],
 Tracts: [1, 500]
}
```

```
[(bps_cmd visit1),
 (bps_cmd visit2),
 (bps_cmd_visit3),
 (bps_cmd_tract1)]
```

Campaign	Command	Start	Finish	Panda	Logs
DR2	Bps visit1	✓	✓	link	link
DR2	Bps visit2	✓	✗	link	link
DR2	Bps tract1	✓	✗	link	link
DR2	Bps tract2	✗	✗		

Figure 5. Rubin Campaign Management database and tools to conduct campaigns

4 Multi-site DRP Campaign on Rubin Commissioning Camera Data

Since the beginning of the Science Pipelines development for DRP, we have used Hyper Suprime-Cam [16] and simulated LSSTCam data from the DESC Collaboration for testing purposes. We also use them to test our multi-site processing deployment and capabilities. From 24-Oct-2024 to 12-Dec-2024, the Vera C. Rubin Observatory took a survey with the full mirrors and telescope in place but used a Commissioning Camera (ComCam) [17]. The ComCam is similar to the LSST Camera but has only 9 16-megapixel silicon detectors (in a 3x3 matrix). Those CCD images were processed using the PanDA based multi-site infrastructure at the USDF. The processing was smooth with few glitches. Fig. 6 shows the usage of computer cores and estimated timeline of the LSST science pipeline steps at the USDF during one day of such processing.

Given the limited data volume, the DRP processing of these data was run at the USDF. The LSST Science Pipelines software, the PanDA system, the Campaign Management tooling, BPS and Data Butler were used, but data movement between multiple sites was not. Nonetheless, this exercise demonstrated the successful integration of the involved components. The processing turnaround time was about one day, and that made it easy to test science code changes.

This processing exercise also provided a valuable view into the intrinsic characteristics of the Rubin processing: step 1, 4 and 5 have millions of short duration jobs at around 2 minutes but step 2 and 3 have hundreds of long duration jobs (~12 hours). Rubin is looking into grouping those short jobs into clusters to improve efficiency.

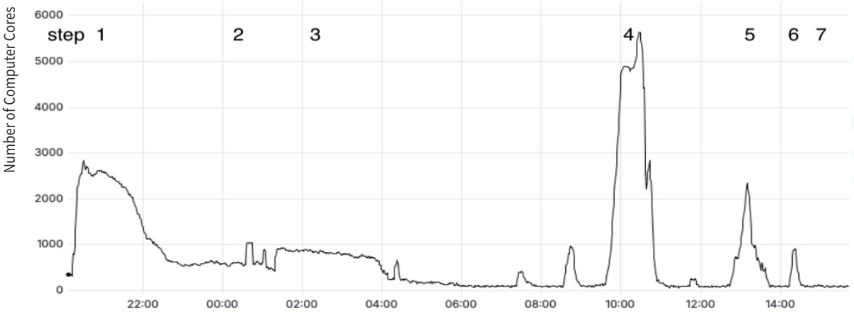


Figure 6. Computer core used at the USDF for one day processing of the ComCam data, with a rough estimation of processing timeline for each of the high-level Science Pipeline steps. Due to the rapid evolution of the LSST Science Pipeline software, these steps do not match the newer stages and steps described in Fig. 3.

5 Conclusions

The successful processing of the ComCam data at the USDF, along with the progress made on data movement are good indications that Rubin's choice of distributed DRP architecture fits its needs, and the multi-site processing is starting to all work together.

Compared to the Commissioning Camera, the LSST Camera has 21 times more CCD image detectors. The LSST Camera will take images every night when weather permits, and the DRP will only have a 200-day per year window to (re)process all previous raw data using new Science Pipelines software. Therefore continuous improvement of the science code, PanDA, Rucio, Campaign Management tools, and reliable and high-performance data facilities are vital to the success of DRP. Rubin is planning on several phases of large-scale multi-site processing campaigns. The goal is to sharpen software components, hone operational procedures, and be ready when the real data comes.

6 Acknowledgements

This material is based upon work supported in part by the US National Science Foundation through Cooperative Agreement AST-1258333 and Cooperative Support Agreement AST-1202910 managed by the Association of Universities for Research in Astronomy (AURA), and the US Department of Energy under Contract No. DE-AC02-76SF00515 with the SLAC National Accelerator Laboratory managed by Stanford University. Additional Rubin Observatory funding comes from private donations, grants to universities, and in-kind support from LSSTC Institutional Members.

This work has been supported by the UK Science and Technology Facilities Council (STFC) funding for UK participation in LSST, through grants ST/X001334/1 and 204 ST/Y003004/1.

This work has been supported by the French National Institute of Nuclear and Particle Physics (IN2P3) through dedicated funding provided by the National Center for Scientific Research (CNRS).

References

1. Z. Ivezic *et al.*, LSST: From Science Drivers To Reference Design And Anticipated Data Products , Astrophys. J., 873, 111, 2019, https://doi.org/10.3847/1538-4357/ab042c
2. https://www6.slac.stanford.edu/lsst
3. K. Lim, Proposal and Prototype for Prompt Processing, Vera C. Rubin Observatory Data Management Technical Note DMTN-219, 2022. https://dmtn-219.lsst.io
4. K. Findeisen, *et al.*, Failure Modes and Error Handling for Prompt Processing, Vera C. Rubin Observatory Data Management Technical Note, DMTN-260, 2004 https://dmtn-260.lsst.io
5. J. Bosch, Y. AlSayyad, *et al.*, An Overview of the LSST Image Processing Pipelines, ASP Conf Ser 523, 521, 2019 https://doi.org/10.48550/arXiv.1812.03248
6. T. Jenness, J.F. Bosch, A. Salnikov, N.B. Lust, N.M. Pease, M.Gower, M. Kowalik, G.P. Dubois-Felsmann, F. Mueller and P. Schellart "The Vera C. Rubin Observatory Data Butler and pipeline execution system", Proc. SPIE 12189, Software and Cyberinfrastructure for Astronomy VII, 1218911 (29 August 2022); https://doi.org/10.1117/12.2629569
7. C. Aguado Sanchez, J. Bloomer, P. Buncic, L. Franco, S. Klemer and P. Mato, CVMFS-A File System for the CERNVM Virtual Appliance, Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research vol 1 p 52, 2008, https://pos.sissa.it/070/012/pdf
8. Barisits, M., Beermann, T., Berghaus, F. *et al.*, Rucio – Scientific data management Comput Softw Big Sci (2019) 3: 11, https://doi.org/10.1007/s41781-019-0026-3
9. Foster, I., Kesselman, C., High Performance Computing: From Grids and Clouds to Exascale, IOS Press, pages 3-30, 2011, https://doi.org/10.3233/978-1-60750-803-8-3
10. Maeno, T., Alekseev, A., Barreiro Megino, F.H. *et al.,* PanDA: Production and Distributed Analysis System. *Comput Softw Big Sci* **8**, 4 (2024). https://doi.org/10.1007/s41781-024-00114-3
11. A A Ayllon *et al*., FTS3 New Data Movement Service for WLCG, 2014 J. Phys.: Conf. Ser. 513 032081, https://doi.org/10.1088/1742-6596/513/3/032081
12. Alfieri, R. *et al.* (2004). VOMS, an Authorization System for Virtual Organizations. In: Fernández Rivera, F., Bubak, M., Gómez Tato, A., Doallo, R. (eds) Grid Computing. AxGrids 2003. Lecture Notes in Computer Science, vol 2970. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24689-3_5
13. M. Gower, M. Kowalik, N.B. Lust, J.F. Bosch, T. Jenness, Adding Workflow Management Flexibility to LSST Pipeline Execution, Astronomical Data Analysis Software and Systems XXXII, 2022, https://doi.org/10.48550/arXiv.2211.15795
14. E. Karavakis, W. Guan, *et al*, Integrating the PanDA Workload Management System with the Vera C. Rubin Observatory, EPJ Web of Conferences **295**, 04026 (2024) https://doi.org/10.1051/epjconf/202429504026
15. F. Hernandez, *et al.,* Data Movement Model for the Vera C. Rubin Observatory, Submitted to the proceeding for CHEP 2024
16. H. Aihara, N. Arimoto, R. Armstrong, and et al., “The Hyper SuprimeCam SSP Survey: Overview and survey design,” Publications of the Astronomical Society of Japan, vol. 70, p. S4, Jan. 2018, https://doi.org/10.1093/pasj/psx066
17. B. Stalder, *et. al.* "Rubin Observatory Commissioning Camera: summit integration", Proc. SPIE 12184, Ground-based and Airborne Instrumentation for Astronomy IX, 121840J (29 August 2022); https://doi.org/10.1117/12.2630184