VERA C. RUBIN OBSERVATORY

Rucio at LSST/Rubin

7th Rucio Community Workshop

Yuyi Guo / Fermilab

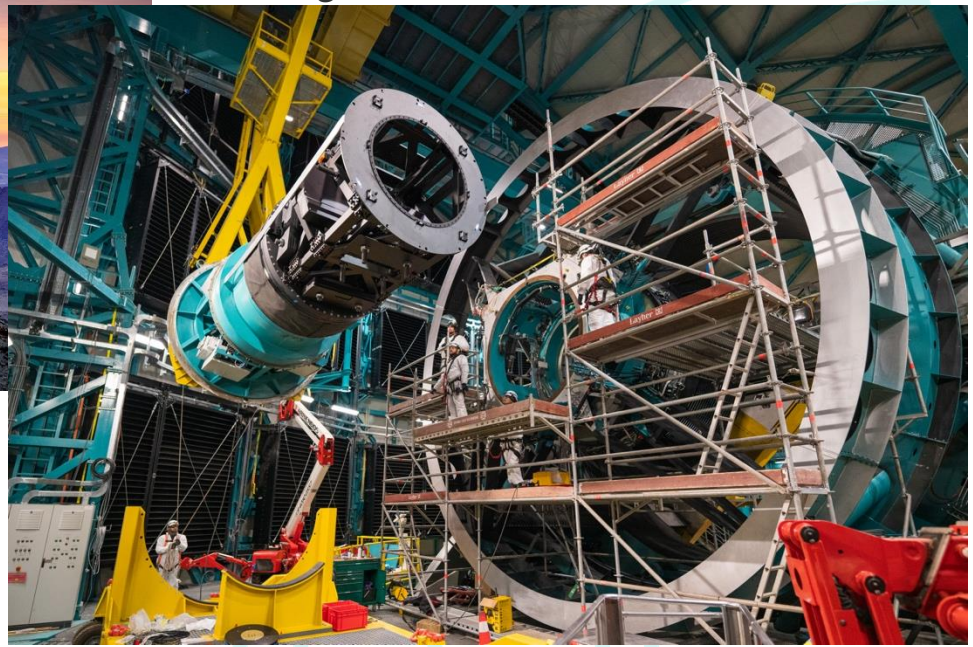
NSF AURA U.S. DEPARTMENT OF ENERGY SLAC CHARLES AND LISA SIMONYI FUND FOR ARTS AND SCIENCES LSST CORPORATION

Outline

- Rubin Intro – How Rubin will use Rucio
- Rubin data model
- Multi-site processing
- Rucio performance and scalability evaluations
- Tape RSE project
- Suggestions

Cerro Pachón Chile – 3.2 Gpixel CCD Camera

8.4m diameter primary mirror
2000 exposures/night 10-year survey starting ~2025



Hundreds of repeat scans of 25,000 square degrees of sky – search for moving (Near Earth Objects) and transients (Supernovae, Gamma Ray Bursts, Gravitational Wave counterparts)

Introduction to Rubin – How Rucio will be used 1/3

Rubin Data Volume:

- 5 PB of raw data annually, with cumulative reprocessing over 10 years projected to reach ~1 EB by 2035, including processed data.
- 2000 exposures (20 TB) collected nightly from the telescope in Chile and transmitted to SLAC for processing and distribution.

Introduction to Rubin – How Rucio will be used 2/3

Rucio's Role in Data Movement:

1. **Annual Data Release Processing (DRP):**
   - Cumulative processing, with one DRP each year generating coadded images, object catalogs, and object light curves (detection timelines of individual objects).
   - Requires the transfer of **millions of small image files daily**, necessitating packaging methods such as ZIP.
   - Four processing centers (FrDF, UKDFs, USDF) equipped with 3K-20K cores, 4 GB RAM/core, and over 5 PB storage each.
   - Daily data transfer: 20 TB moved to processing centers, with a subset of 2 TB/day of compressed, selected outputs returned to the central DF (USDF@SLAC).

2. Tape robot RSE at SLAC for raw, key output **backup storage (up to 30TB/day).**

Introduction to Rubin – How Rucio will be used 3/3

3. **Distribution of Curated DRP Data**
    - Data subsets will be shared with community scientists through Independent Data Access Centers (IDACs) and other analysis facilities, reaching 30 remote universities and institutions. This distribution can provide up to several hundred GB per day per institution.
4. **Calibration Data Transfer**
    - Certified calibration data will be transferred from the USDF to the Summit for use in observing systems.
5. **Supporting Data Distribution**
    - Possible distribution of supporting materials, including postage stamp image cutouts, catalogs, and background templates, associated with alerts for potential objects of interest (e.g., supernovae, Near-Earth Asteroids).

Rubin Data Model 1/2

- **Rucio** serves as the authoritative source for file locations.
- **Integration**: Rucio and the Rubin Data Butler (catalog) work together, each tracking files.
- **Identity Mapping**: Use the mapping from Logical File Name (LFN) to Physical File Name (PFN) to streamline integration with the Butler.
- **Data Transfers**: Rucio handles the transfer of raw, Quantum Graph (QG), and published data products among Data Facilities (DFs).
- **Dataset Management**: Utilize Rucio Datasets to group files by data type and sky location, facilitating targeted transfers through Rucio Subscriptions.

Rubin Data Model 2/2

- **File Packaging**: Rucio Datasets also enable the consolidation of small files into larger ones.

- **Future Considerations**: While Rucio Containers are not currently a priority, they may become valuable in the future.

- **Data Storage**
  - US Data Facility (USDF): Will hosts 100% of raw images and all published data products.
  - French Data Facility (FrDF): Will host a 100% of raw images and a fraction of published data.
  - UK Data Facility (UKDF): Will host 25% of raw images and a fraction of the published data.

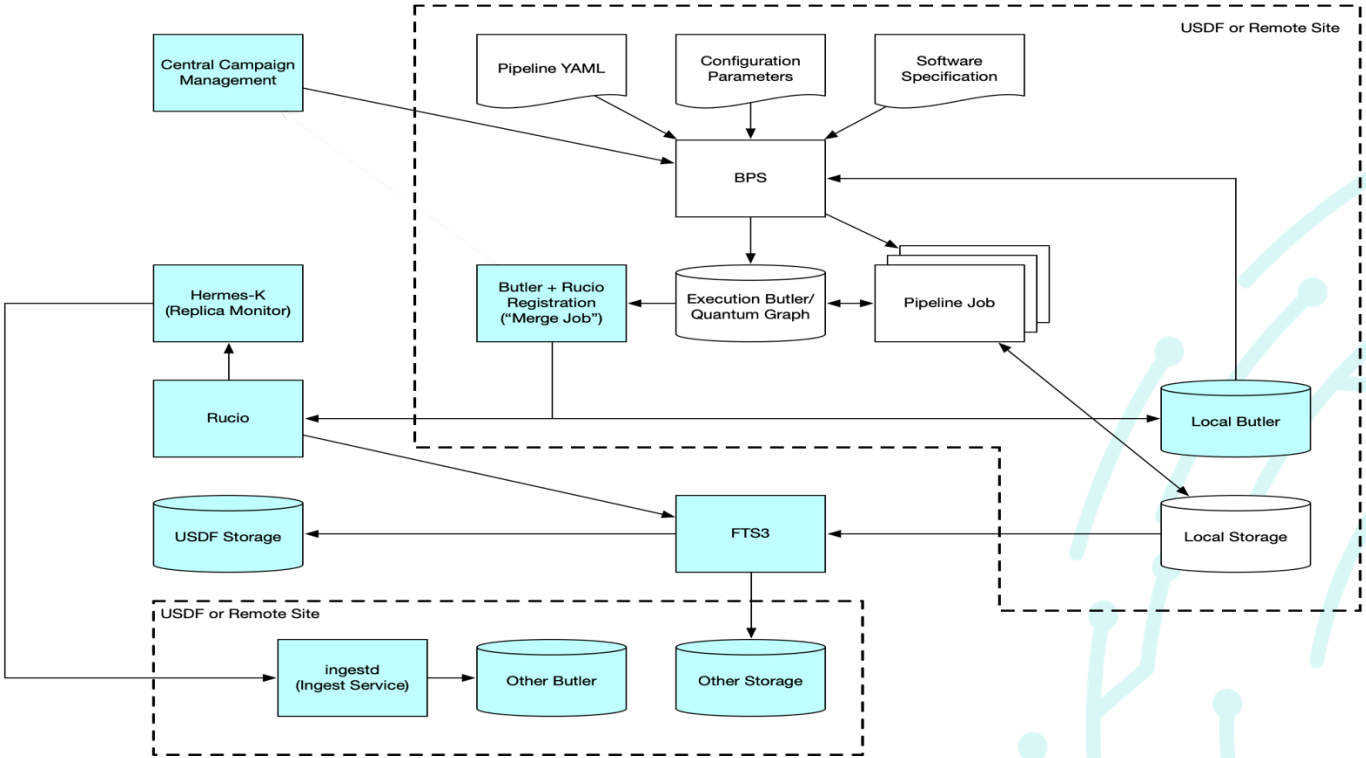
Multi-Site Processing 1/3

- **Rubin Data Release Production (DRP)** will operate across multiple sites.
- **Central Campaign Management** oversees the submission and sequencing of workflows to Data Facilities (DFs).
- **Workflow Execution**: Employ single-site execution with Rucio for result replication, as temporary files are large and numerous.
- **Batch Production Service (BPS)**: Works alongside the central PanDA to define and control processing workflows at each DF.
- **Job Submission**: PanDA jobs are submitted remotely to execute BPS on localhost, generating the Quantum Graph (QG).
- **Merging Process**: The merge job integrates the QG into the local Butler at each site, registers files with Rucio, and replicates them to the USDF, which then submits additional PanDA jobs based on the QG.

Multi-Site Processing 2/3

- **Final File Management**: Rucio ingests final files in place and automates replication to other sites.
- **Replication Automation**: Achieved through Rucio subscriptions, utilizing metadata provided at registration.
- **Data Availability**: The ingestd ingests results into the local Butler, making them accessible for subsequent workflows and jobs.
- **Testing Overview**: Initial testing involved one-way transfers from FrDF and UKDF to USDF, with automated ingestion from FrDF.
- **Plans**: Additional testing is ongoing.

Multi-Site Processing 3/3



Replica Monitor & Ingest Service toolkit

- **Rucio-register**: registers replicas (datasets of files) from Rubin's source end 'Butler' (metadata database) with Rucio.
- **Rucio rules**: Declare replicas to be made at source and destination ends
- **FTS3**: moves replicas from source to destination RSEs based on rules
- **HermesK**: message system – one message per file/dataset works with Rucio+FTS3
- **Ingestd**: listens for messages from HermesK, registers just-arrived files in RSE with 'Butler' on receiving end (reverse of rucio-register).

HermesK

- HermesK monitors Rucio to determine when a replica of a file/dataset has been fully transferred. It then sends out messages on a Kafka topic.
- HermesK has a plugin architecture which is used to filter messages. We use it to filter messages, looking for **transfer-done**. Once found, we send a Kafka message on topics matching the RSE name.
- When files are registered with Rucio, **Rubin metadata** is added to information kept by Rucio. When HermesK sends out messages, this metadata is added as part of the message.
- HermesK is under internal review. Once we review it, we'll send this to the Rucio team for review.

Ingestd

- Ingestd is local to a data facility. Kafka services only receive messages on Kafka topics which match RSEs at that data facility.
- Ingestd uses PFN to identify the file and metadata in the messages to reconstitute the Butler registry entry for the file.
- Ingestd automatically ingests that dataset into a local Butler.

Rucio performance and scalability evaluations 1/4

- Rucio is for orchestrating the data movement among Rubin's data facilities for DRP, distribution of curated DRP data subsets to universities and institutes, and possible distribution of supporting data.
- Evaluations performed in terms of
  - number of files per unit of time
  - file sizes
- The goal is to understand Rucio in the context of Rubin's data needs and the underlying infrastructure.

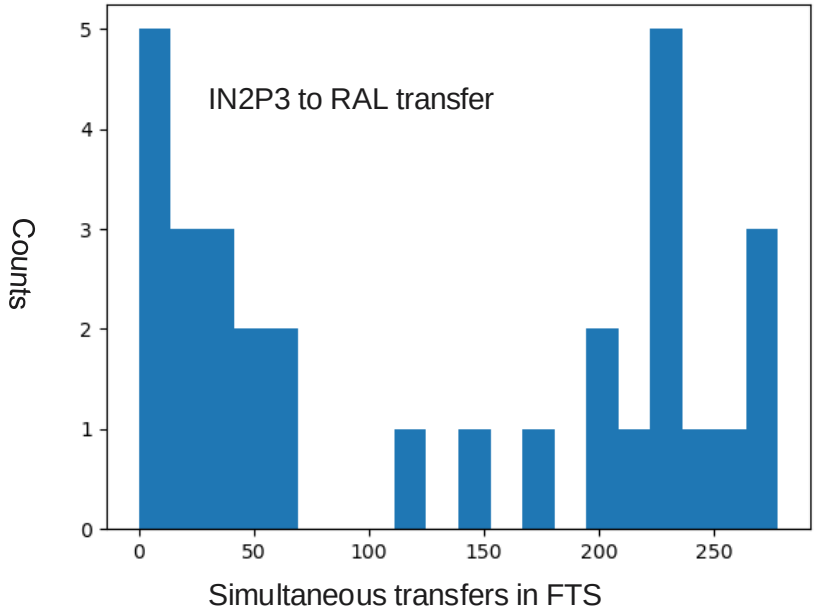
Rucio performance and scalability evaluations 2/4

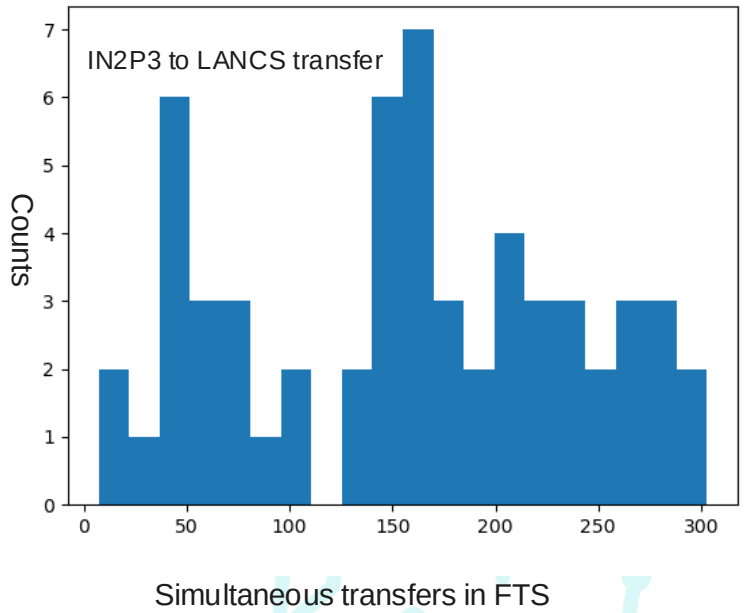
- Rucio replication between different Rubin DFs
  - a set of ~35,000 calexp files in size about 55MB/file.

Average Throughputs	
source=>Destination	Throughput (MB/s)
SLAC=>IN2P3	400 - 500
SLAC=>LANCS	300 - 400
SLAC=>RAL	300 - 500
IN2P3=>LANCS	400 - 700
IN2P3=>RAL	800 - 1400

Rucio performance and scalability evaluations 3/4

103,855 Calexp Background files with size in 50KB/file tested for simultaneous transfers. We are looking for a way to increase the concurrent transfers.

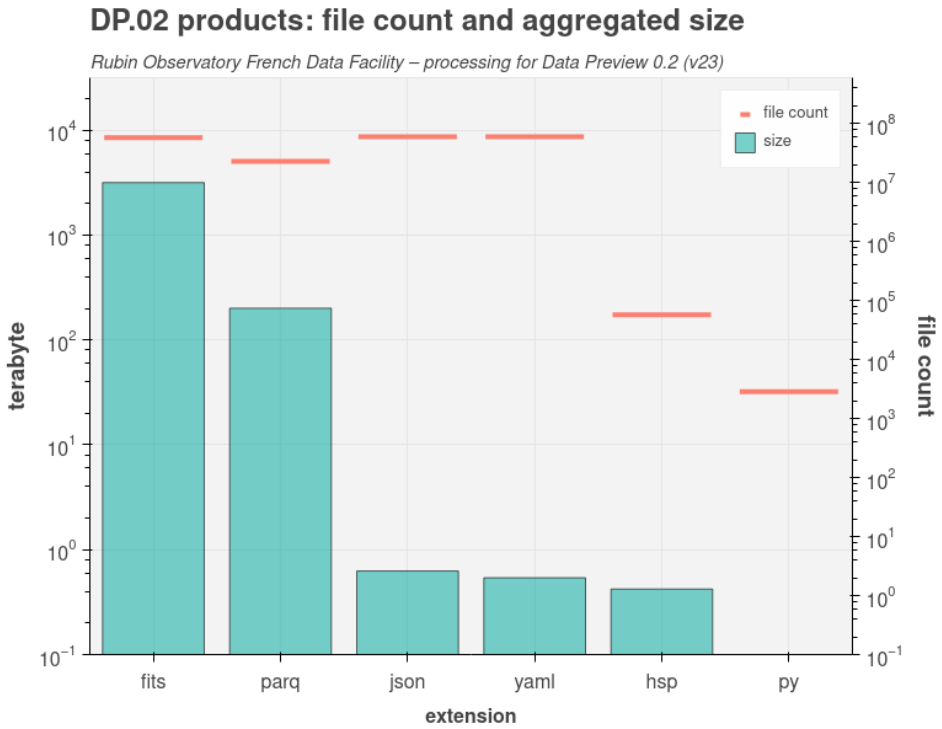




Rucio performance and scalability evaluations 4/4

- Summary of transfer throughputs between Rubin Data Facilities in limited tests
  - 55MB Calexp files ~500 MB/s
  - 12MB raw files ~250 MB/s
  - 50K files ~ 850 KB/s
  - 2K files ~100 KB/s
- Rubin needs a mechanism to pack the files into larger sizes.

Tape RSE Project



Characteristics of Rubin data files

Click here to interact with this plot

Tape RSE Project

The factors

- Rubin files are mostly small between O(10KB) and O(10MB).
- Deal with more files (no in bytes) than the LHC.

The goals

- Reduce the number of files.
- Decrease the pressure to storage system.
- Increase the efficiency of data transfer system.
- Improve the performance for Rucio database.

The Tools and requirements

- Python zip archive and zipinfo that allows Posix-like IO on members.
- Butler needs to know individual member files.
- Rucio, FTS and storage only know the zip archives.

The 3rd Phase of Tape RSE Project

- We have an alpha implementation to be tested. We did below:
  - Used Rucio metadata to tag datasets that need to be zipped.
  - Made a daemon to find the datasets to zip and zip them.
  - Wrote the zipped datasets to tape.
  - Ingested in place the datasets to Rucio.
- Currently we only did this at USDF.
- The plan is to zip the files at where they generated, and all the zipped datasets will be transferred back to USDF.

Suggestions

- Rucio documentation about API needs improvement. We had to look at the Rucio code for clarification sometimes.
- Enhanced QoS support
  - Improve the preparation time in Rucio for a multi-source transfer
  - In Rucio client, a way to list rules for DIDs matching a pattern. For example, **rucio list-rules raw:Dataset/LSSTCam/202408'*'**
- Completion of VO-specific Policy Package testing
- Provide documentation/guidance for Independent Data Access Centers (IDACs) to deploy a Rucio RSE.

Acknowledgements

I would like to extend my heartfelt thanks to my colleagues Brian Yanny, Kiant-Tat Lim, Wei Yang, Brandon White, Steve Pietrowicz, Greg Daues, Fabio Hermandez, Jhonatan Amado and others for their invaluable contributions to the work presented in this talk, as well as for providing essential information and slides.

Thank You !