



Data Movement Model for the Vera C. Rubin Observatory

G. Beckett, A. Hanushevsky, F. Hernandez, T. Jenness,
K-T. Lim, P. Love, T. Noble, S. Pietrowicz, W. Yang

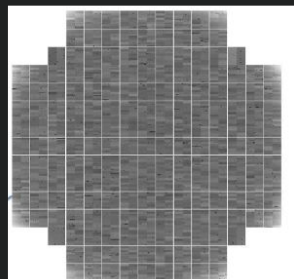
CHEP 2024

Oct. 23rd, 2024



Rubin Observatory Overview

Legacy Survey of Space and Time



raw images



alerts



science-ready images



astronomical catalog

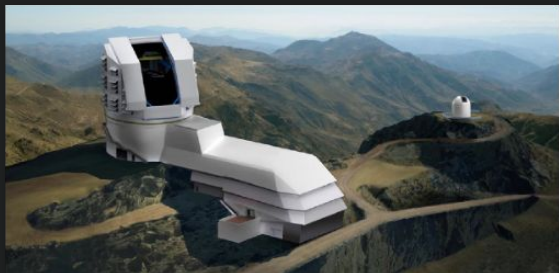


science
collaborations

Rubin Observatory's mission is to survey the sky to build and deliver LSST, a catalog of 20 billion galaxies and 17 billion stars with their associated physical properties

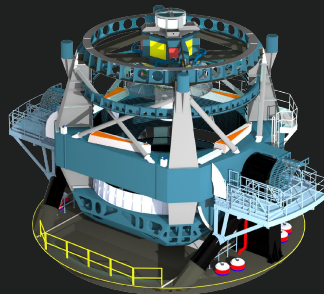
Legacy Survey of Space and Time (cont.)

OBSERVATORY



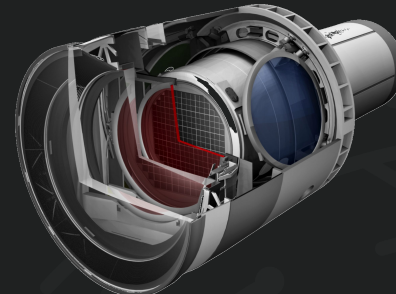
southern hemisphere | 2647m a.s.l.
| stable air | clear sky | dark nights |
good infrastructure

TELESCOPE



main mirror \varnothing 8.4 m (effective 6.4
m) | large aperture: f/1.234 | wide
field of view | 350 ton | compact |
to be repositioned about 3M
times over 10 years of operations

CAMERA



3.2 G pixels | \varnothing 1.65 m | 3.7
m long | 3 ton | 3 lenses | 3.5°
field of view | 9.6 deg^2 | 6 filters
ugrizy | 320-1050 nm | focal
plane and electronics in cryostat
at 173K

Source: [LSST: from Science Drivers to Reference Design and Anticipated Data Products](#)

Legacy Survey of Space and Time (cont.)

Raw data

6.4 GB per exposure (compressed)
2000 science + 500 calibration images per night
300 observing nights per year
16 TB per night, ~5 PB per year

Aggregated data over 10 years of operations

image collection: ~6 million exposures
derived data set: ~0.5 EB
final astronomical catalog database: 15 PB

Operations to start late 2025

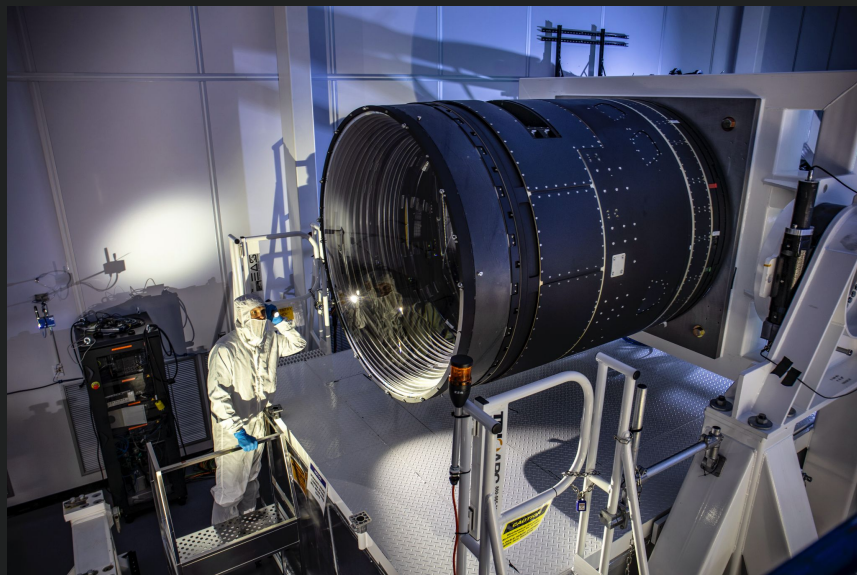


Image credit: Jacqueline Ramseyer Orrell / SLAC National Accelerator Laboratory

Source: [LSST Key Numbers](#)



Cloud

EPO Data Center

US Data Facility SLAC, California, USA

Archive Center
Alert Production
Data Release Production (35%)
Calibration Products Production
Long-term storage
Data Access Center
Data Access and User Services

HQ Site AURA, Tucson, USA

Observatory Management
Data Production
System Performance
Education and Public Outreach

Dedicated Long Haul Networks

Two redundant 100 Gb/s links from Santiago to Florida (existing fiber)
Additional 100 Gb/s link (spectrum on new fiber) from Santiago-Florida (Chile and US national links not shown)

UK Data Facility IRIS Network, UK

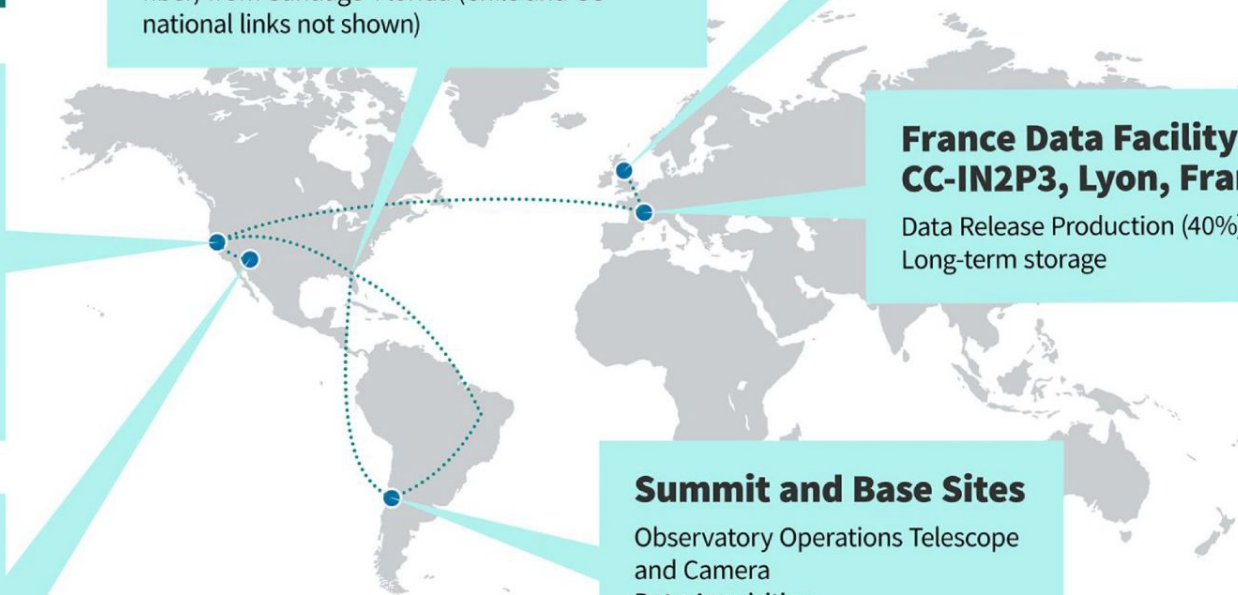
Data Release Production (25%)

France Data Facility CC-IN2P3, Lyon, France

Data Release Production (40%)
Long-term storage

Summit and Base Sites

Observatory Operations Telescope and Camera
Data Acquisition
Long-term storage
Chilean Data Access Center



- Data movement use cases and tools
 - *summit site* → *archive center*
 - *archive center* ↔ *data facilities*
 - *archive center* → *data access centers*
- Status
- Summary

Data movement: summit → archive center

Summit → Archive Center

- Images are transferred using direct writes via the S3 API to an object store at the US Data Facility archive at SLAC
 - *specialized connection pooling/keep-alive code and TCP tuning is used to make this efficient over the very high bandwidth-delay product international network*
 - *target end-to-end latency, including data compression and other overheads: 7 seconds for more than 4 GB per exposure*
- Telemetry and a few databases are replicated using native protocols for timeliness
- A relatively small number of certified calibration files are transferred infrequently from the Archive back to the Summit (and other locations)
 - *currently done using ad hoc rsync, but we are working to convert it to using Rucio*

Data movement: archive center ↔ data facilities

Rubin Data Facilities

- Annual reprocessing of the entire image dataset recorded since the beginning of the survey for producing a data release to be performed at three data facilities
 - *US data facility ([SLAC National Accelerator Laboratory, CA, USA](#)) – 35%*
 - *UK data facility ([IRIS](#) and [GridPP, UK](#)) – 25%*
 - *French data facility ([CC-IN2P3, Lyon, FR](#)) – 40%*
- US and French data facilities to store each a full copy of the raw image dataset
- US to store the gold copy of published data products and to operate the observatory's archive center
- Raw image data to be transferred eastward and data products westward across the Atlantic
 - *up to 165 ms RTT*
- Connectivity between processing facilities provided by ESnet (transatlantic segment from/to SLAC), GEANT (within Europe), JANET (UK) and RENATER (FR)

LSST Science Pipelines Middleware

- Processing is organized into `PipelineTasks` that execute scientific algorithms on data
- The *Data Butler* is the sole client library used to retrieve and persist data items specified using scientifically-relevant identifiers (not pathnames) to and from in-memory Python objects
 - *PipelineTasks never read files directly, they rely on the Butler to retrieve the input data they need to process and to persist the data they generate*
 - *Butler uses a database to track locations of items in a data repository and relationships between them*
- Batch Processing Service (BPS) executes *workflows* composed of `PipelineTasks`, managing sequential dataflow and distributed data-parallel execution using plugins to interface with workflow management systems (PanDA, HTCondor, Parsl, Pegasus)

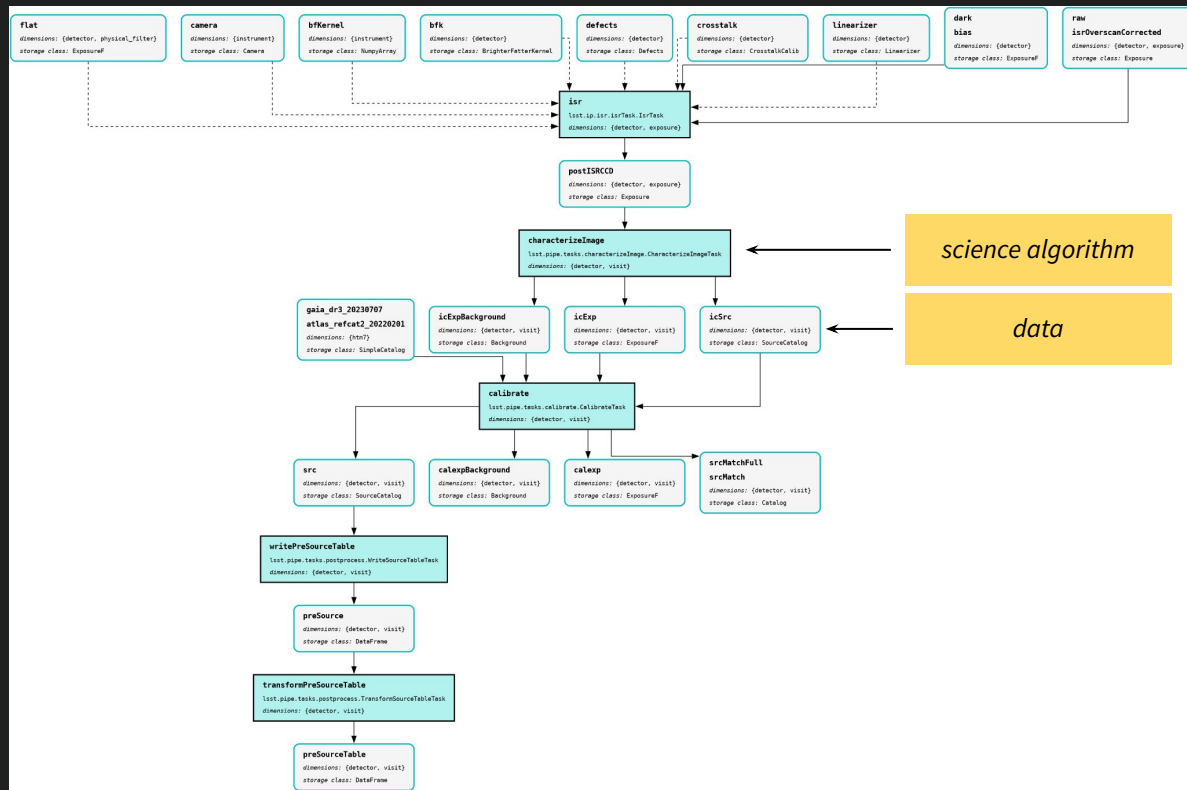
For further details see [“The Vera C. Rubin Observatory Data Butler and Pipeline Execution System”](#)

LSST Science Pipelines Middleware (cont.)

Example of a relatively simple workflow generated by Batch Processing Service (BPS).

BPS queries Butler to build the workflow according to input specifications.

BPS uses PanDA to orchestrate execution of workflows at data facilities. PanDA submits jobs to each facility's batch farm. Those jobs use data known to the Butler local to the execution facility.



Data movement model

- Processing divided naturally into spatial regions with little to no overlap of needed intermediate data between them
 - *each data facility gets assigned a subset of those regions to process*
 - *only final and selected intermediate products need to be transferred to the archive site*
 - *intermediate products can be deleted when the campaign ends*
- Inter-site data transfer use cases for annual data release processing:
 - *input data (e.g. raw data, calibrations): archive center → processing facilities*
 - *all final products and selected intermediate: processing facilities → archive center*
- The Data Butler knows only about the files locally present at a single facility
 - *it controls the location of those files under the repository root location*
 - *to be usable, replicated files must be ingested into the receiving facility's local Butler*

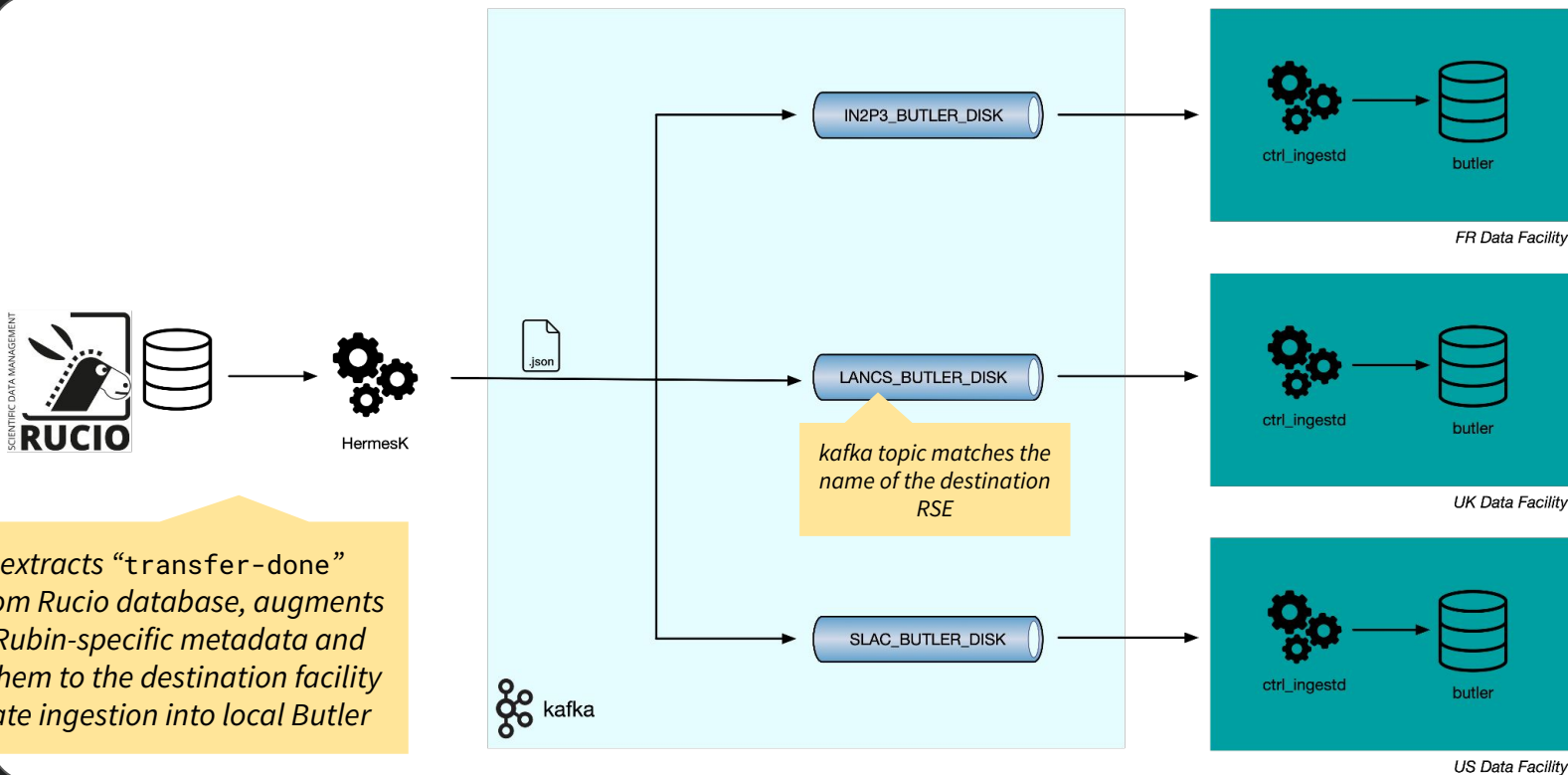
Data movement tools

- We use Rucio and FTS to orchestrate data movement among the processing facilities
- Each facility exposes a few Rucio storage elements (RSE)
 - *one devoted to store raw data, with more restrictive write access permissions to reduce unintended modification or deletion*
 - *one devoted to store the facility's Butler repositories which contain data products*
- RSEs are configured with **identity** LFN-to-PFN algorithm and labeled with attributes we conveniently use to express Rucio replication rules
 - *Preserving the file pathname, relative to the local Butler repository's root location, is a must*
- New raw and calibration data is ingested daily into the archive center's Butler repository and registered in place into Rucio for replication to the European facilities
 - *reception at the destination facility triggers immediate ingestion into the local Butler repository*
- Final data products get registered into Rucio and transferred to the archive center as they are produced
 - *similarly, reception at archive center triggers ingestion into the local Butler repository for consolidation*
 - *time budget for producing the annual data release: 200 days*

Data movement tools (cont.)

- **rucio_register** registers files in place into Rucio catalog, associates Rubin-specific metadata to them and creates Rucio Datasets according to some conventions
 - *metadata is a JSON dictionary with details required for ingestion into destination Butler (~ 500 bytes)*
 - *typical size of a Rucio Dataset with raw data: ~20k files (100 exposures)*
 - *raw data ingestion frequency: ~10 Hz*
- **ctrl_rucio_ingest** (a.k.a. HermesK) retrieves file replication messages from Rucio database, augments them and emits them via a multi-site Kafka cluster
 - *the payload of those messages include Rubin-specific metadata about the file, the scope, name and the destination Rucio storage element (RSE)*
 - *the Kafka topic each message is emitted through is named after the destination RSE*
- **ctrl_ingestd** runs as a daemon at each processing facility, listens for file transfer messages in the relevant topic and ingests replicated files into the local Butler
 - *once ingested, those files become known to the local Butler and thus available for processing*
 - *files are replicated by Rucio at the location expected by Butler*

Data movement tools (cont.)



Data movement: archive center → data access centers

Archive center → Data access centers

- Annually released data products to be transferred to several data access centers in the Americas, Europe and Asia-Pacific for scientific analysis
- Foreseen technical solution uses Rucio to distribute annual data products to data access centers
 - *15 to 20 sites of varying needs and size located around the world*
- Advantages
 - *distribution can be centrally managed, using the same tools used for interfacility data exchange, according to flexible data replication rules*
 - *Rucio & FTS ensure the integrity of data when transported to the analysis centers*
 - *distribution can be orchestrated not only from the archive center to analysis centers but also between analysis centers, sharing the load and reducing the network bandwidth out of the archive site*
- Data access centers would need to configure and operate a Rucio storage element
 - *supported authentication mechanisms (tokens, X.509 certificates, etc.) may pose problems to some sites*

For further details see [“Bulk Data Transfer Policies and Procedures”](#)

Status

- Development of Rubin-specific tools mostly complete
 - *ongoing work to demonstrate reliable end-to-end functionality: registration in place + replication + ingestion at reception*
 - *some improvements opportunities identified but not yet implemented*
- Scalability tests to be performed next
 - *Rubin image processing results in a large number of files of relatively small size: significant fraction in the KB to 100 MB range*
 - *technical solution for aggregating small files into zip files identified but not yet fully implemented; zip allows byte range extraction of individual components unlike tar/cpio/pax*
- Ongoing effort to build monitoring tools for observing inter-facility file replication activity

Summary

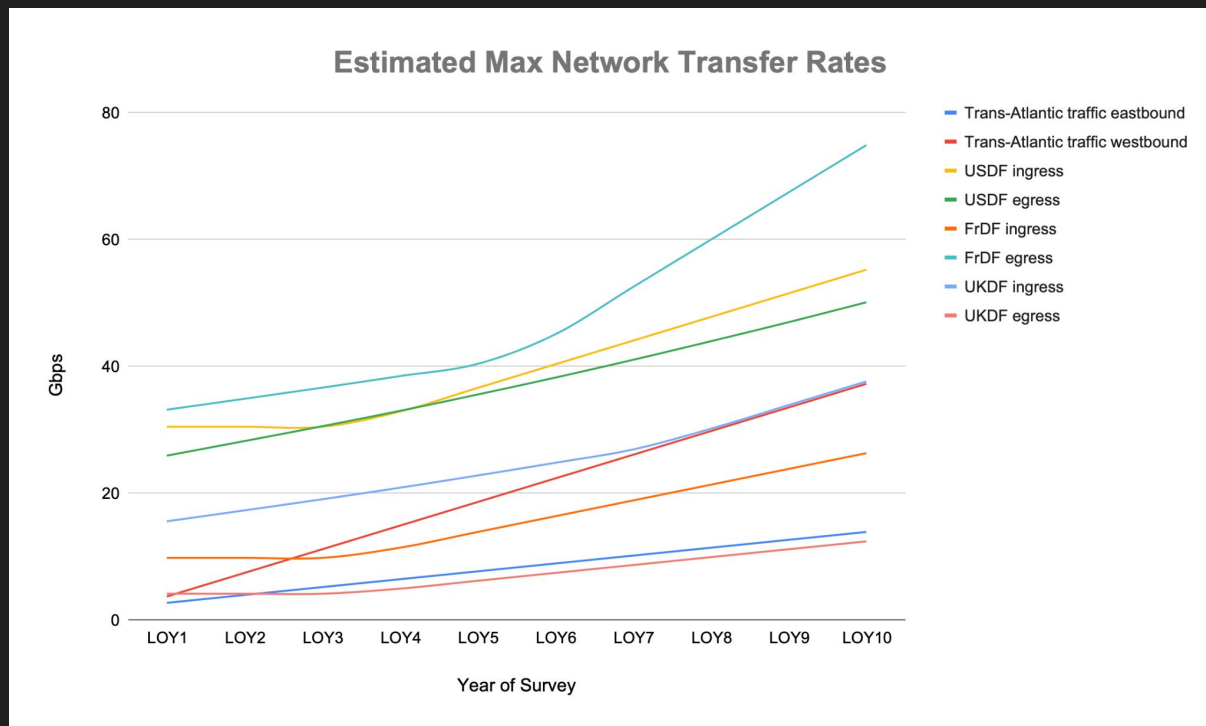
Summary

- Rubin uses PanDA for driving execution of payloads of LSST Science Pipelines at the processing facilities
 - *further details in talk [“Preparation for Multi-Site Processing at the Vera C. Rubin Observatory”](#) in track 4*
- Rubin is basing its inter-site replication machinery on Rucio & FTS
 - *extended Rucio to use Kafka as a replication control plane between the processing facilities and the archive center*
 - *developed tools to integrate with Rubin middleware to perform actions on successful replication events*
- The end-to-end system is currently being tested in reasonably realistic conditions
 - *scalability tests to be performed next*
- Distribution of released data to analysis centers still to be prototyped

Reference material

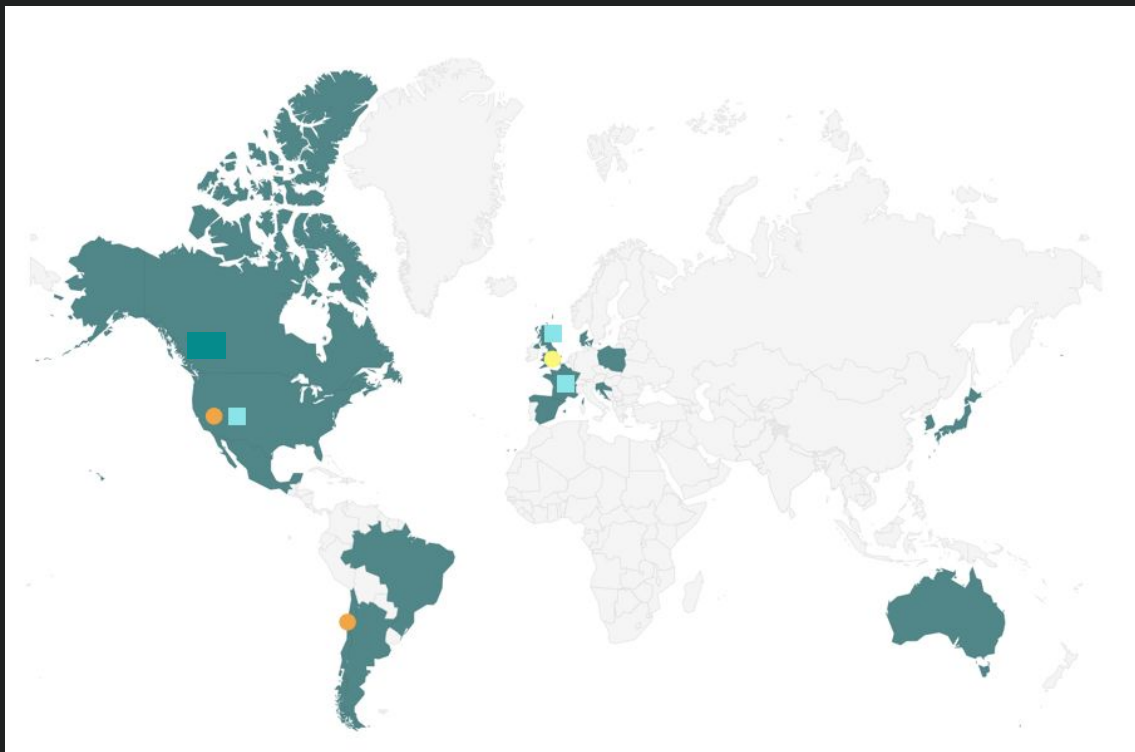
Projected data transfer rates




These estimations make some assumptions that we may need to revisit as we learn how data reprocessing will proceed in real-life conditions



Adapted from Richard Dubois, Rubin Observatory

Rubin Data Access Centers

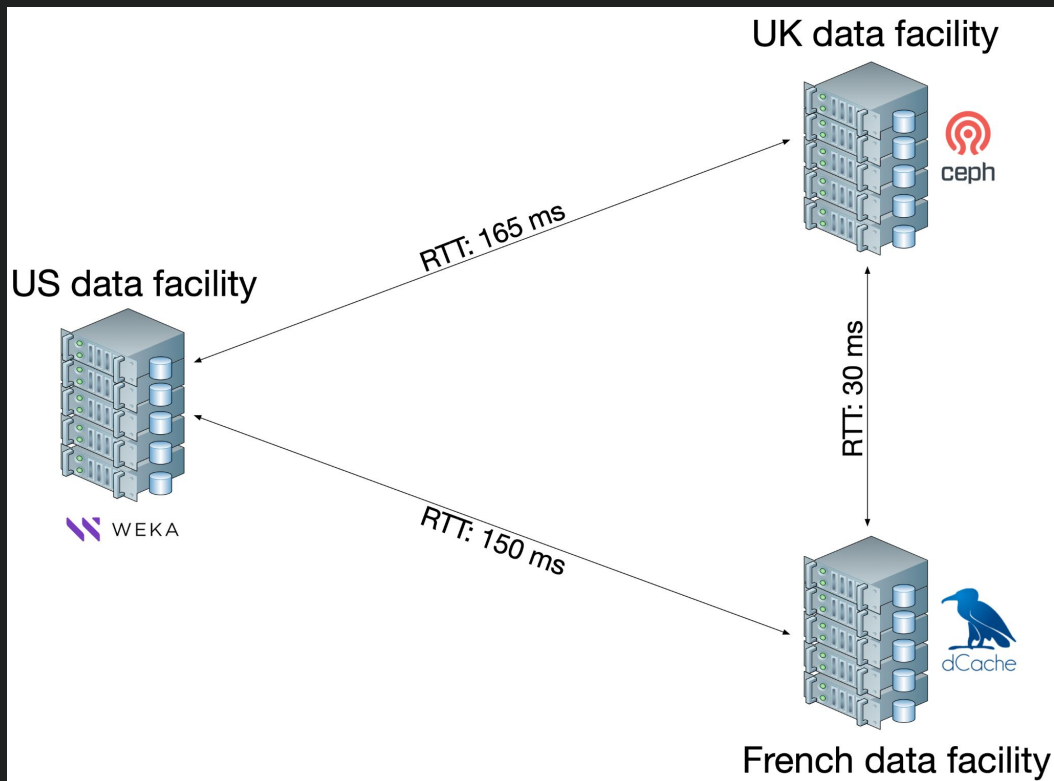


Data Access Center		Chile, US
Full Independent DAC		UK
Lite Independent DAC		Argentina, Australia, Brazil, Canada, Denmark, Japan (x2), Mexico, Poland, Slovenia, South Korea, Spain
Scientific Processing Center		Croatia
Data Facilities		France, UK, US

Two Data Access Centers managed by the Rubin project to serve released data to scientific communities.

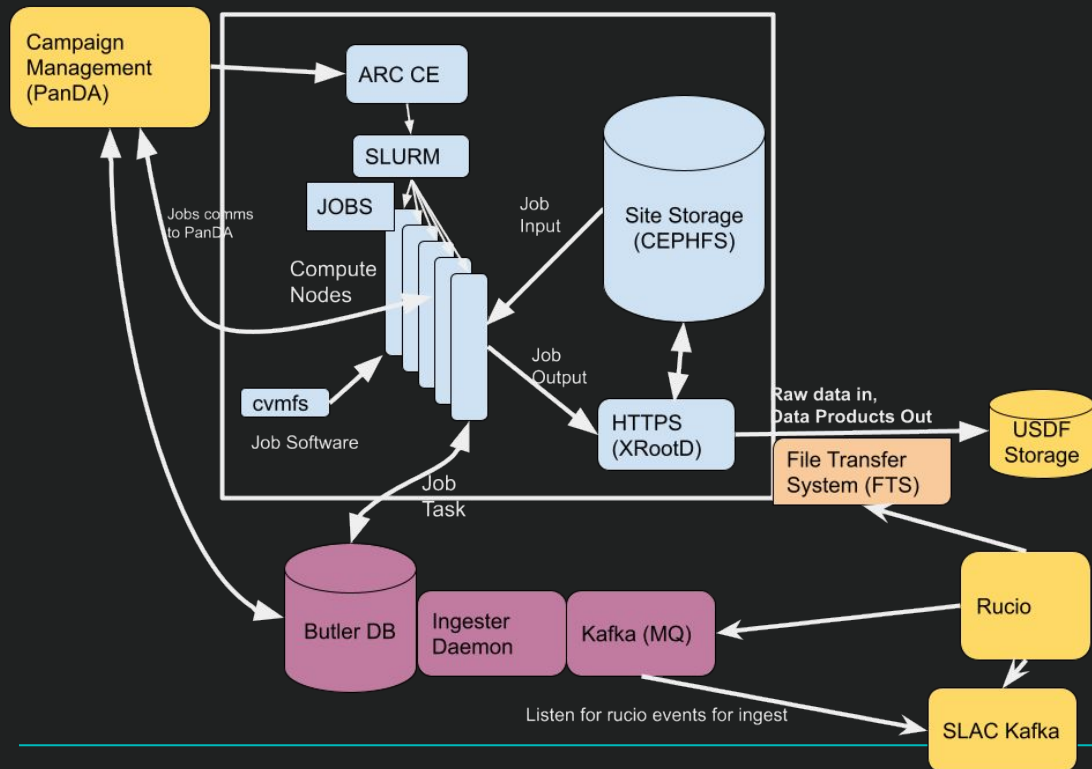
Lite Independent Data Access centers are in-kind contributions. They intend to store and serve a subset of data.

Rubin Data Facilities Network Latency



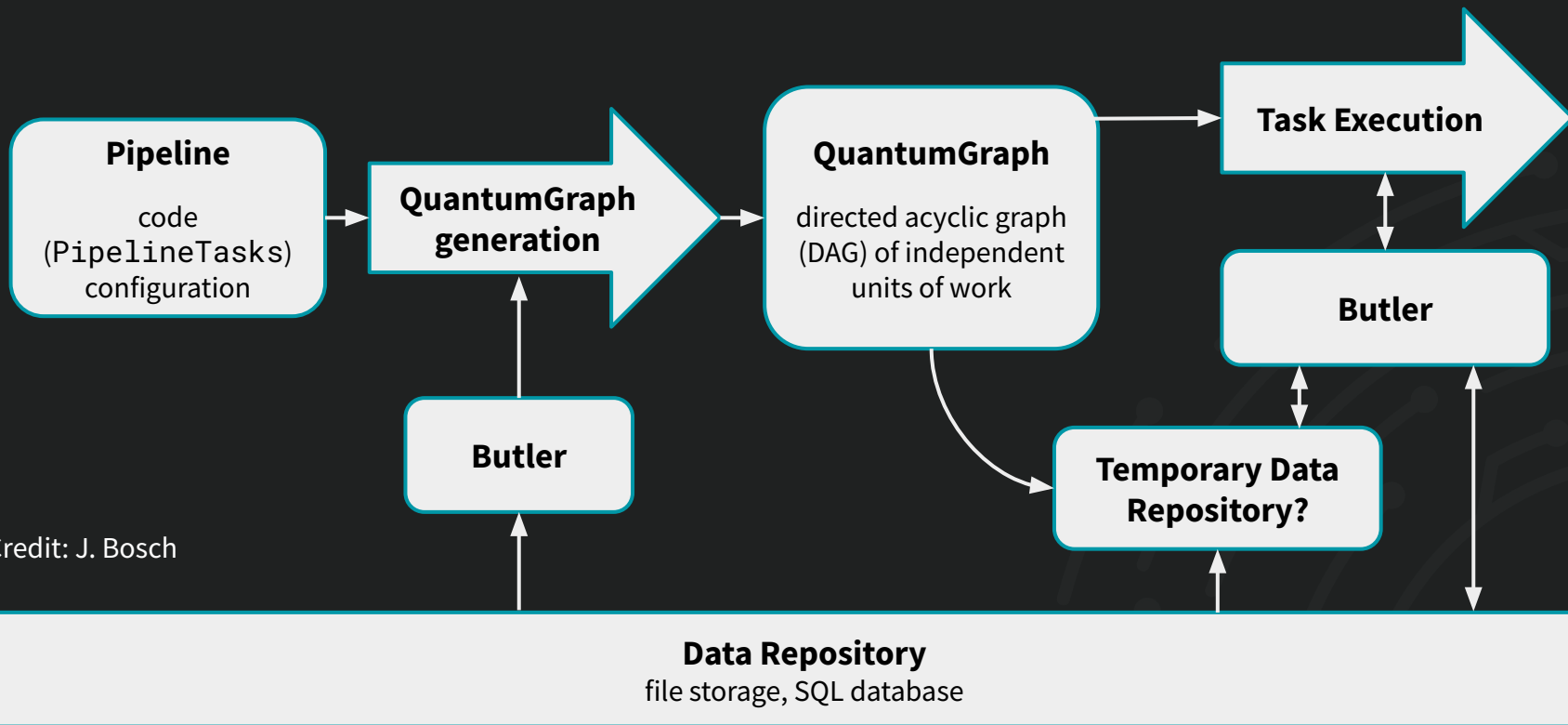
Data replication over high-latency network links

DRP Site Anatomy (Lancaster)



- Pilot Jobs are submitted from PanDA to the ARC CE (grid “gatekeeper”) at the site.
- This in turn submits batch jobs to the SLURM batch system.
- Jobs communicate with PanDA and Butler to get workloads.
- Job software is served over CVMFS.
- Input data is read directly from the read-only CEPHFS mount.
- Output (intermediate or final) is uploaded via HTTPS for traceability.
- Data transfers are handled using the FTS, and orchestrated via Rucio.
- Rucio events are communicated via Kafka.
- This triggers ingestion into the Butler for use by jobs.

Middleware before and during execution



Credit: J. Bosch