VERA C. RUBIN OBSERVATORY

Image processing infrastructure to produce the Legacy Survey of Space and Time (LSST)

G. Beckett, P. Clark, M. Doidge, F. Hernandez, T. Jenness, E. Karavakis, Q. Le Boulc'h, P. Love, G. Mainetti, T. Noble, B. White, W. Yang

CHEP 2023, May 8-12, 2023

NSF NOIR Lab AURA SLAC U.S. DEPARTMENT OF ENERGY

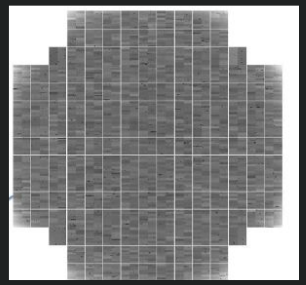
CCIN2P3 UKRI Science and Technology Facilities Council iris GridPP UK Computing for Particle Physics Lancaster University THE UNIVERSITY of EDINBURGH

Contents

- Overview of the Legacy Survey of Space and Time (LSST)
  - https://rubinobservatory.org
- Distributed image processing
- Ongoing work

Overview of LSST

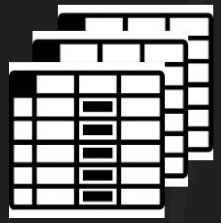
Legacy Survey of Space and Time



raw images

alerts

science-ready images



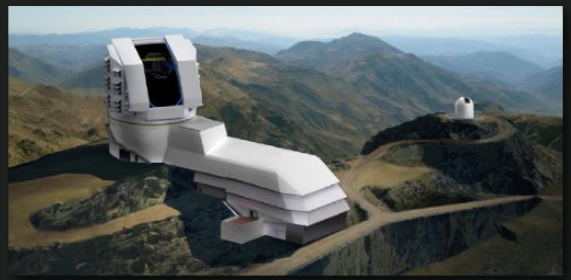
astronomical catalog

science collaborations

*LSST aims to deliver a catalog of **20 billion galaxies** and **17 billion stars** with their associated physical properties*

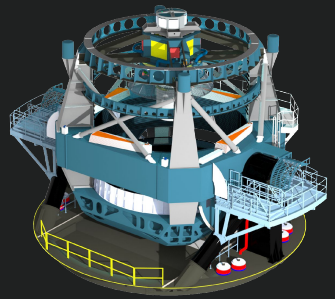
Legacy Survey of Space and Time (cont.)

OBSERVATORY



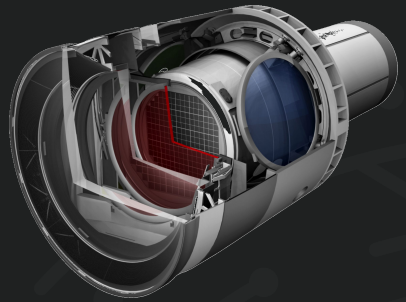
southern hemisphere | 2647m a.s.l. | stable air | clear sky | dark nights | good infrastructure

TELESCOPE



main mirror ∅ 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 | ∅ 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

Source: LSST: from Science Drivers to Reference Design and Anticipated Data Products

Legacy Survey of Space and Time (cont.)

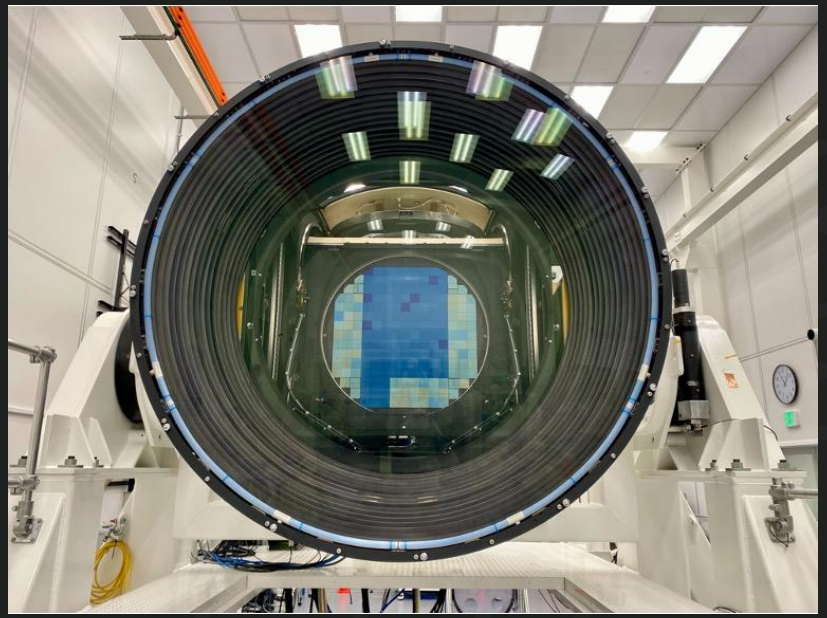
Raw data

6.4 GB per exposure (compressed)
2000 science + 500 calibration images per night
20 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 early 2025

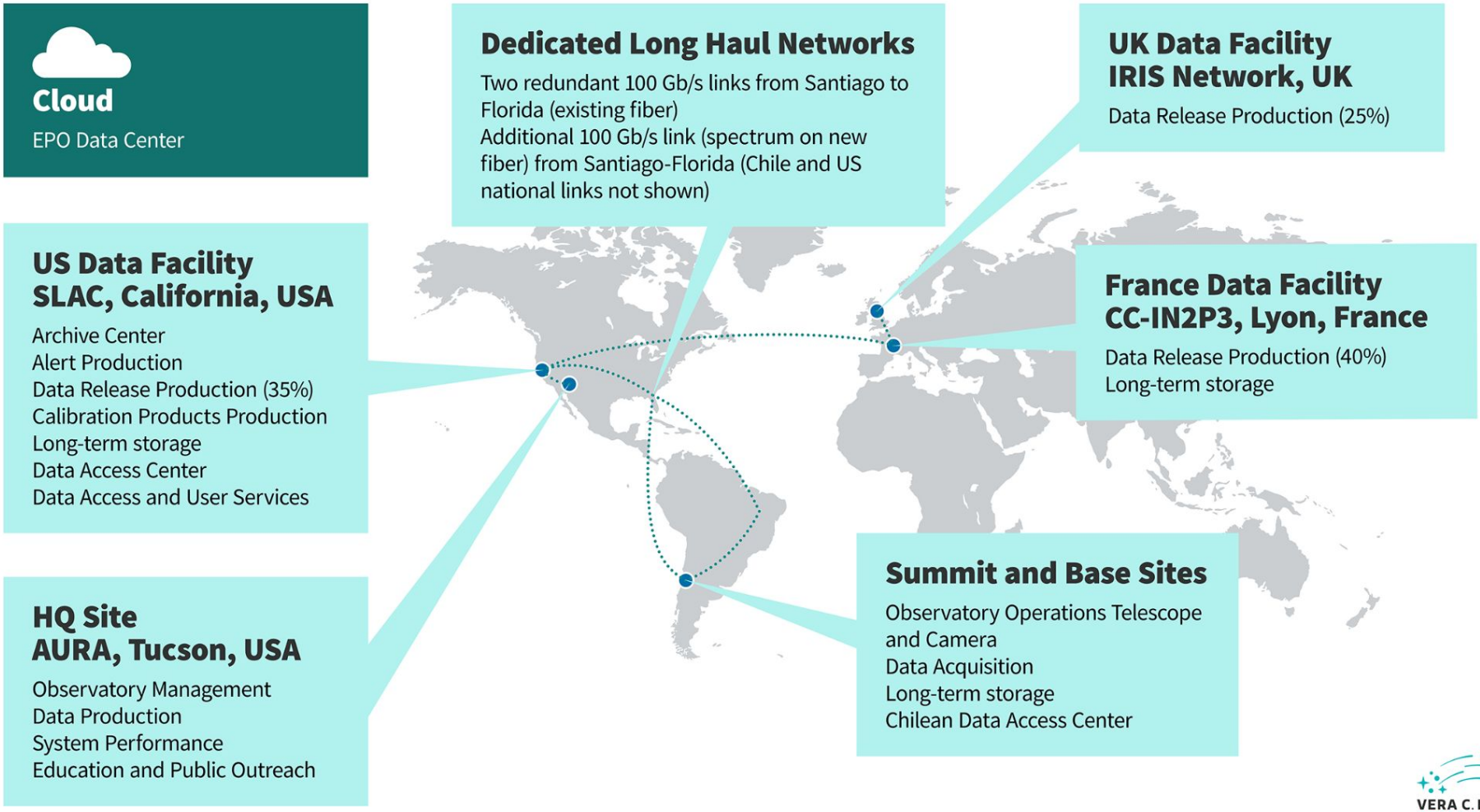


Source: Rubin Observatory System & LSST Survey Key Numbers

Distributed processing

Rubin Data Facilities

- Image processing for producing the annual data release to be performed at 3 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 data facility to store an integral copy of raw and published data products
  - *implies replication of the entire dataset across the Atlantic*
- Connectivity among those facilities provided by ESnet (transatlantic segment from/to SLAC), GEANT (within Europe), JANET (UK) and RENATER (FR)
  - *facilities specifically configured not to use LHCONE*



VERA C. RUBIN OBSERVATORY

LSST Science Pipelines

Major processing steps

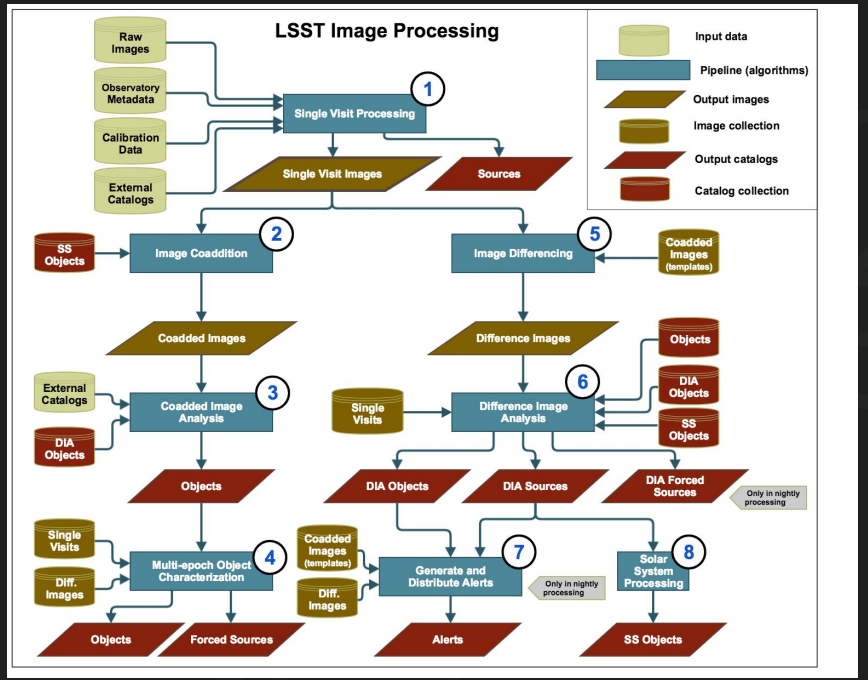
- Single-frame processing
- Calibration
- Image coaddition
- Coadd processing
- Catalog production

Lower layer written in C++ for performance (150 KLOC), upper layer in Python for expressivity and convenience (350 KLOC)

Expose CLI and Python APIs

Open source development: github.com/lsst

Documentation: pipelines.lsst.io



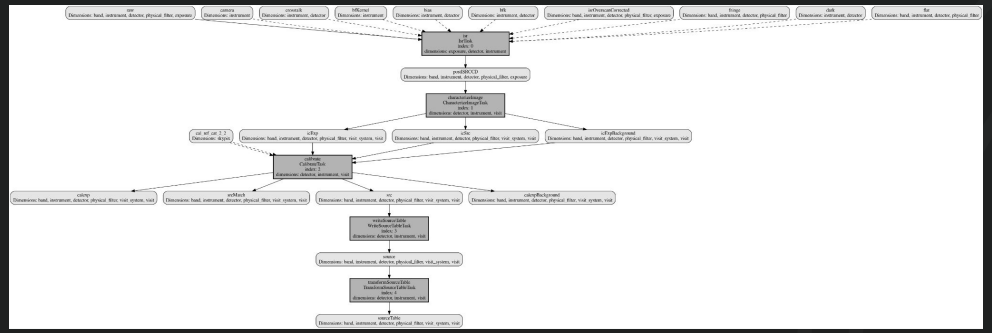
More information: [An Overview of the LSST Image Processing Pipelines](), [Rubin Observatory Data Products Definition Document]()

LSST Science Pipelines (cont.)

- Packaged and distributed via several mechanisms
  - *Conda- and container-based (Docker, Apptainer)*
  - *Intended for installation at both individual scientists' personal computers and at data facilities*
  - *Linux (CentOS 7 and others) and macOS*
- Batch farms in the 3 data facilities mount a single CernVM-FS repository
  - *Image processing jobs can use the conda-based distribution or Apptainer container images to execute the pipelines*
  - *Details: https://sw.lsst.eu*

Middleware

- Image 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*
  - *It uses a database to track locations of items in a data repository and relationships between them.*
- Batch Production Services (BPS) executes workflows composed of `PipelineTasks`, managing sequential dataflow and distributed data-parallel execution
  - *Uses plugins to interface with workflow management systems (PanDA, HTCondor, Parsl, Pegasus)*



Additional information: The Vera C. Rubin Observatory Data Butler and Pipeline Execution System

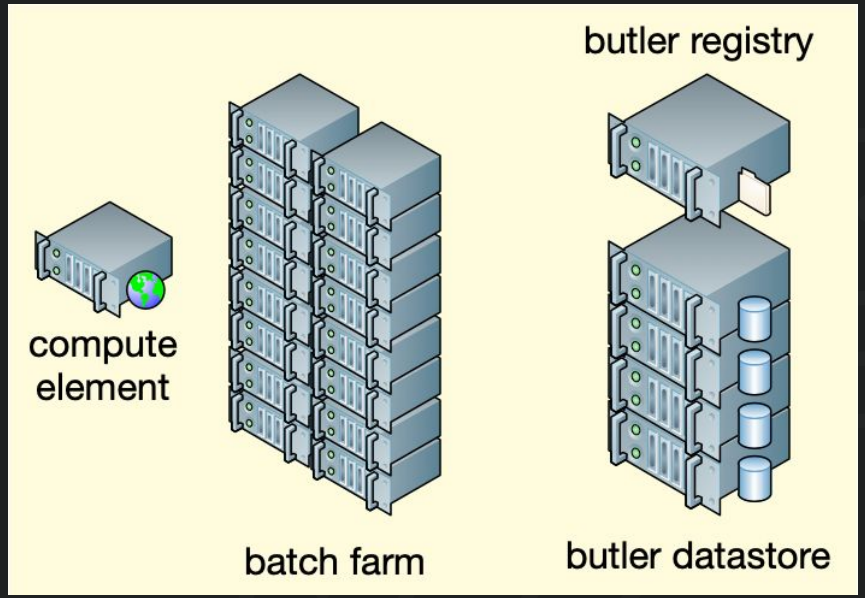
Typical Rubin data facility

Compute element

- Exposes the site's batch farm to the workflow executor
- Typically composed of ARC CE and Slurm

Butler repository

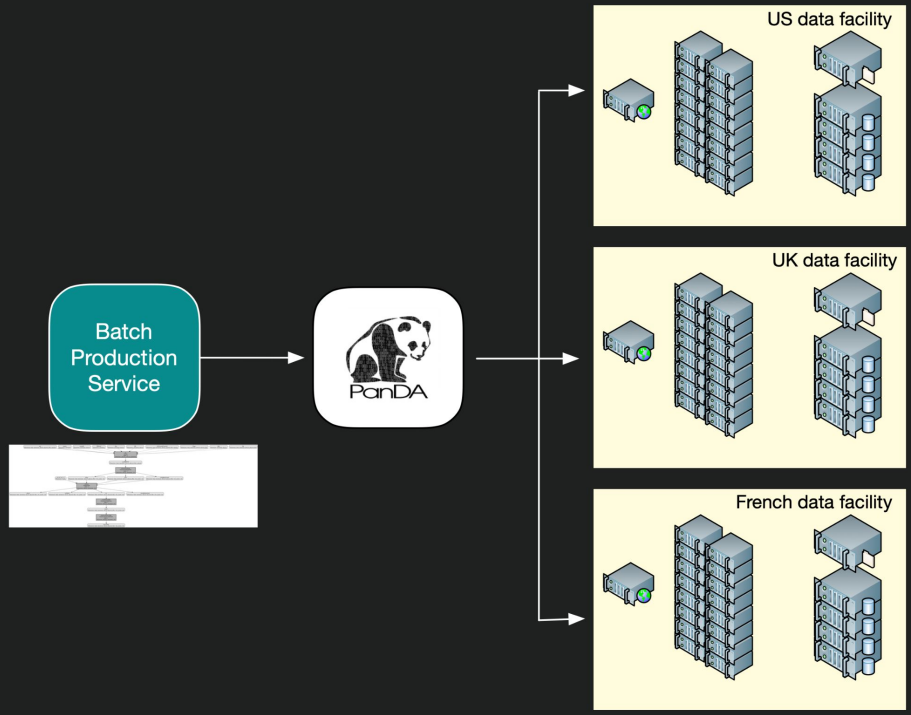
- *Registry*: database which contains the location of the data and their relationships (PostgreSQL)
- *Datastore*: storage system where the data files are located. Weka (S3), Google Storage, dCache (webDAV), XRootD (webDAV), CephFS, Lustre



Additional information: The Vera C. Rubin Observatory Data Butler and Pipeline Execution System

Distributed image processing

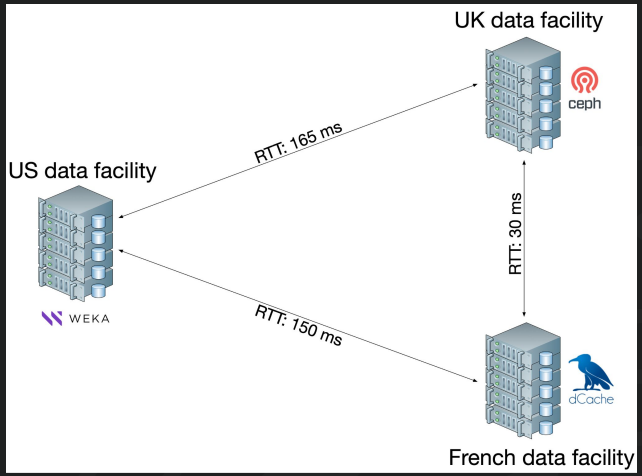
- Batch Production Service (BPS)
  - *Generates the workflow to be executed at each facility: a directed acyclic graph of independent units of work*
  - *Takes into account data dependencies and data location*
- PanDA
  - *Creates pilot jobs and coordinates the execution of the workflow*
  - *Each job executes one or several science algorithms over a set of input data, stores output data in the butler repository local to the facility*



For details see: Integrating the PanDA Workload Management System with the Vera C. Rubin Observatory, track 4, today 2pm.

Inter-site data replication

- Data replication will be achieved with open-source software: Rucio and FTS3
  - *Proven to work at scale by the ATLAS and CMS collaborations, among others*
- Rucio
  - *Replica catalog: Where does my data live?*
  - *Data policy enforcement: How many copies of the data, and where?*
  - *Transfer scheduling: Arranges to satisfy your policies with external services!*
- FTS3
  - *Executes transfers scheduled externally on behalf of Rucio*
  - *Highly configurable for tuning handling of many transfers to many sites*
- Rubin-specific tools
  - *To identify data which needs replication among the facilities (e.g. exclude intermediates) and ask Rucio to replicate it*
  - *To trigger actions at each facility to timely ingest replicated data into the local data butler repository*



Data replication over high-latency network links

Ongoing work

- Regularly performing image processing exercises of increasing complexity at each facility
  - *Using data sets of simulated images or images from other telescopes*
  - *Modest scale (a few thousands CPU cores) so far relative to the required scale*
  - *Orchestrated processing using the 3 facilities to be demonstrated: depends on Rucio and Butler integration*
- Performing regular Rucio-driven data replication exercises across the Atlantic
  - *Significant amount of small files (by HEP) standards could become an issue*
  - *Routine replication of relevant scale among the 3 facilities to be demonstrated*

Backup slides

Details of each Rubin facility

- US data facility
  - *Serves as the archive site of the observatory*
  - *ARC CE, Slurm, Weka (S3, datastore), PostgreSQL (registry database)*
  - *Hosts central services: PanDA, Rucio, FTS, logging facility*
- UK data facility
  - *ARC CE, CephFS/XRootD (webDAV, datastore), PostgreSQL (registry database), Kafka messaging*
  - *Approximately 3 FTEs available over 6 persons*
- French data facility
  - *ARC CE, Slurm, dCache (webDAV, datastore), PostgreSQL (registry database)*
  - *Hosts the stratum 0 of the CernVM-FS repository*
  - *Hosts an instance of the astronomical catalog database and of the analysis platform*