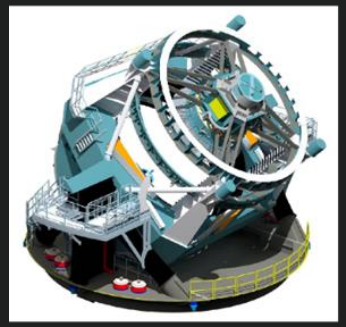
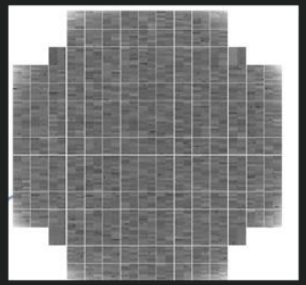
Rucio at the Rubin Observatory

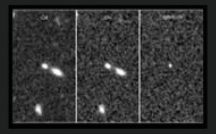
Rucio Workshop : October 2023

Legacy Survey of Space and Time





raw images



alerts

science-ready images

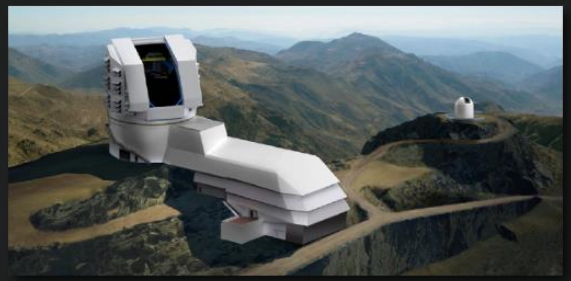
astronomical catalog

science collaborations

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

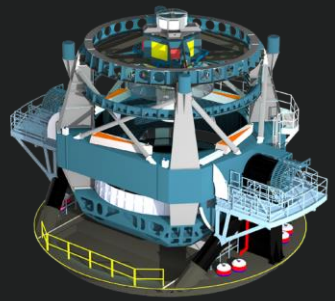
Legacy Survey of Space and Time (cont.)

OBSERVATORY



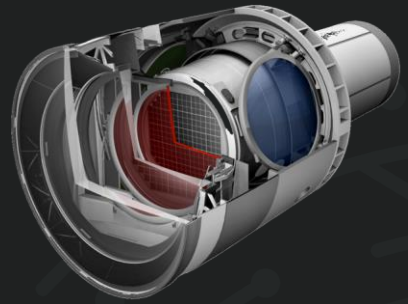
southern hemisphere | 2647m a.s.l. | stable air | clear sky | dark nights | good infrastructure

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.)

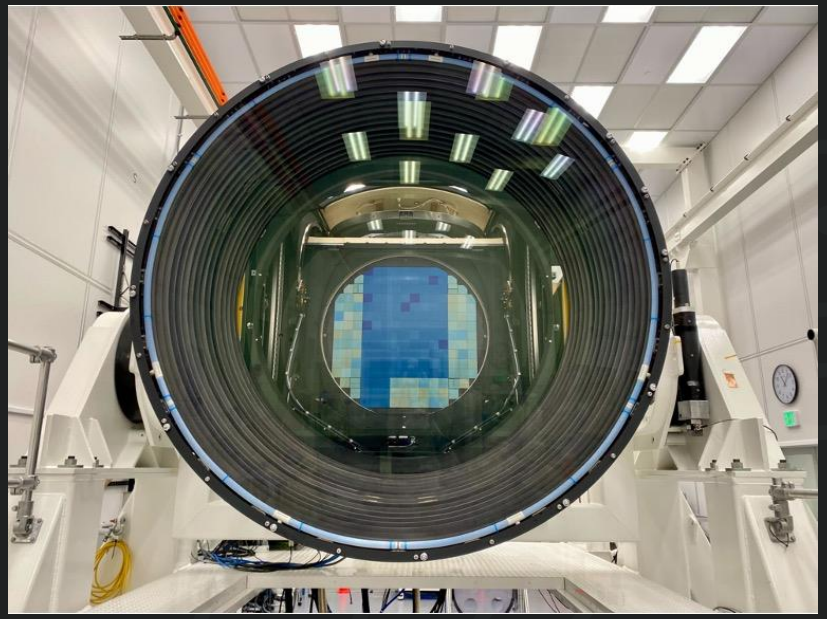
Raw data

6.4 GB per exposure (compressed)
2000 science + 500 calibration images per night
20 TB per night, ~5 PB raw data per year

Aggregated data over 10 year science mission

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

First light 2024
Operations to start early 2025



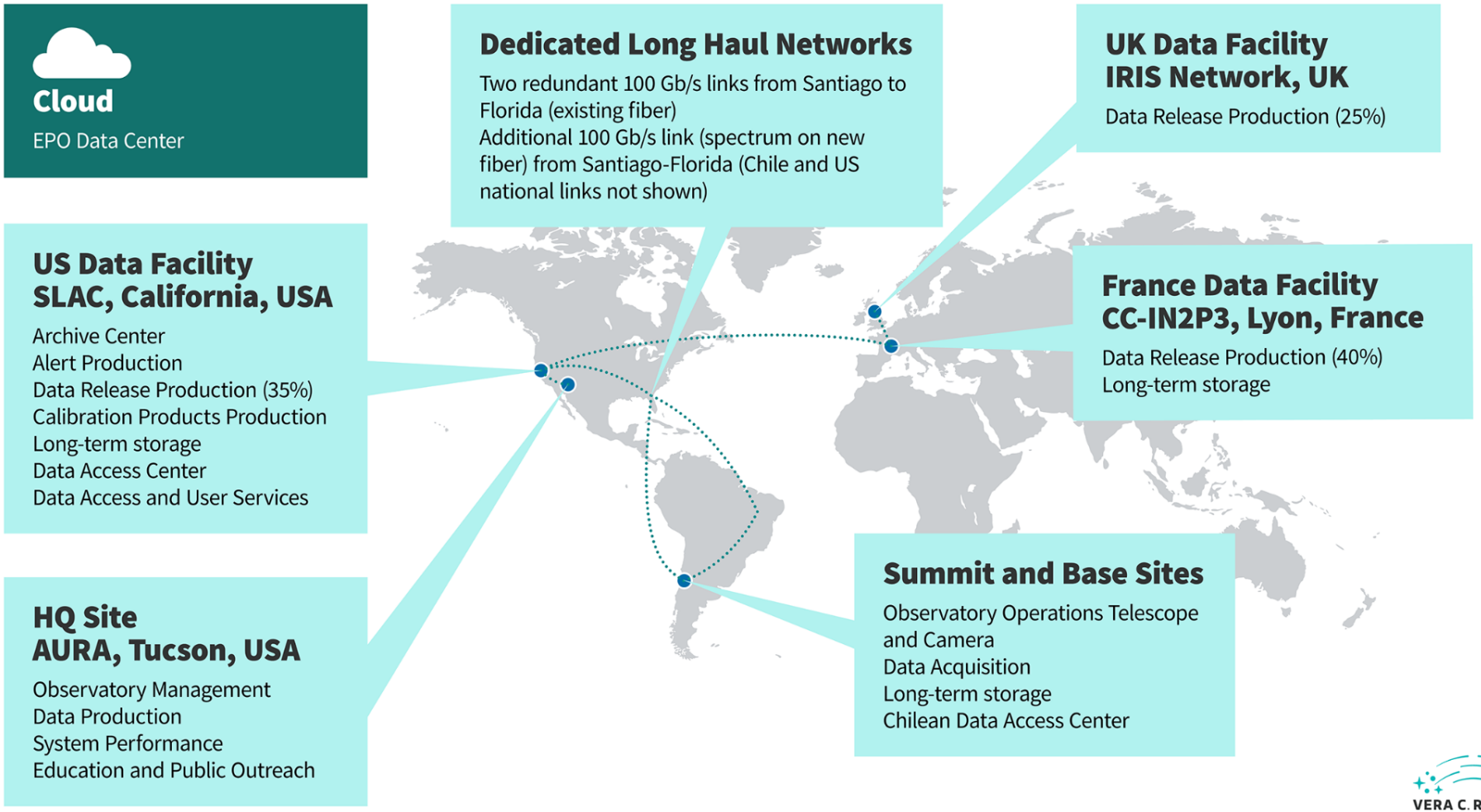
Source: Rubin Observatory System & LSST Survey Key Numbers

Multisite Data Processing

- 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*



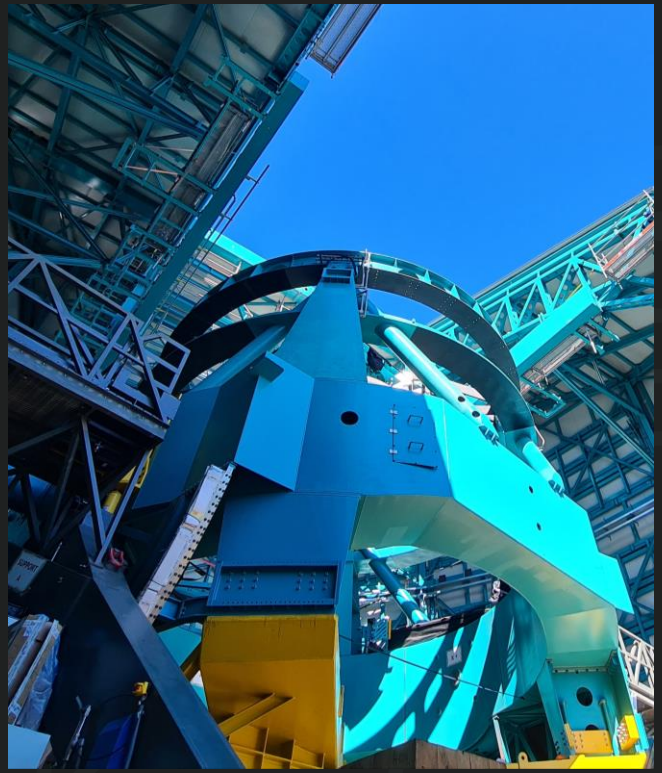
VERA C. RUBIN OBSERVATORY

US Data Facility

- US Data Facility is hosted at SLAC National Laboratory
- USDF will provide
  - Storage
  - Compute
  - Hosting of core services
- Rucio will form the backbone of the Rubin data archive
  - Tape archive RSE w/custom compression and retrieval
  - Movement to the Data Facilities
  - Movement to Independent Data Access Centers
- Rucio will form the backbone of the Rubin data archive

Rucio Deployment

- Physical Kubernetes cluster at SLAC partitioned into virtual clusters for multitenancy support
  - Virtual clusters provisioned with vCluster software from Loft
  - Virtual clusters are nearly indistinguishable from a physical Kubernetes cluster (compared to other platforms aka Openshift)
- Infrastructure definitions located alongside Rucio's
  - CloudNativePG PostgreSQL
  - Strimzi Kafka
  - ActiveMQ – Coming Soon!
  - Storage provisioned from SLAC's Shared Scientific Data Facility Weka filesystem via K8 StorageClass
- Currently running 1.29LTS



Data Movement

- Large % of raw data products UK and FRDF
  - 40% of all raw to FrDF and 25% to UKDF
- Complicated by the number of files
  - Need to determine performance impact of file volume
- All derived data products replicated back to USDF
- Connectivity among DFs provided by ESNet (transatlantic from/to SLAC). GEANT (EU), JANET (UK), RENATER (FR)\
- Annual Data Releases (at minimum) delivered to Independent Data Access Centers

Tape RSE Implementation

- SLAC maintains a HPSS deployment on a Spectra TFinity ExaScale Tape Library
  - Loaded with LTO-9 media (18TB uncompressed)
- Order of magnitude more files going to tape than HEP
  - ~40 million/week
  - Decision was made to compress Rubin data files into zip format
  - Zip archives chosen due to native XRootD support for file operations
  - DIDs will be matched with PFNs with a special form

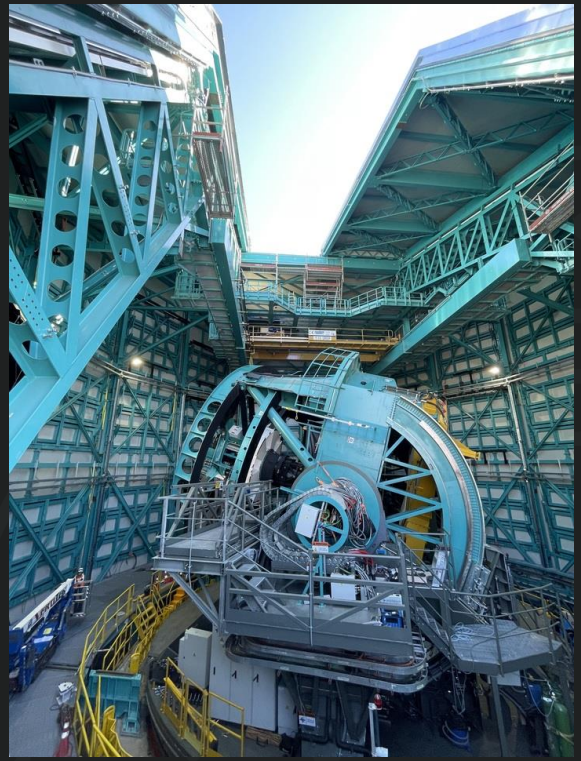
    datasetname.rubinzip?xrdcl.unzip=<pfn_from_deterministic_RSE>

Transfer Testing Results

- Transfers SLAC->LANCS w/real data have poor throughput
  - FTS computation of SRC/DEST checksum is current suspect
  - 7MB HSC datafiles – 230MB/s aggregate transfer rate
  - iperf3 tests have demonstrated that the 10Gb/s link can be saturated
- Another test, SLAC -> IN2P3
  - 7000 * ~1GB files @ 1.4GB/s
- Official Prometheus monitoring would help standardize transfer analysis
  - Incorporate the work of Tim (RAL) for Rubin?

HermesKafka Daemon

- Custom version of the 1.29 Hermes daemon
  - Designed by Steve Pietrowicz at the National Center for Supercomputing Applications
- Supports delivery of messages to listed Kafka topics
- Topics replicated to sites with MirrorMaker2 for resilience
- Plugin architecture
  - Default plugin allows transmission to multiple Kafka topics
  - Rubin-specific plugin transmits based on destination RSE while adding custom metadata
- Transfer messages delivered to topic at storage site
- Daemon at site reads message and ingests newly arrived file into Data Butler



Future Concerns

- Client distribution
  - Gfal-python compiled independently for each Python version
- Transfer speed testing and optimization
- Establish a tape storage element
- Rucio/Data Butler integration