VERA C. RUBIN OBSERVATORY

Rubin Observatory Multi-Site Testing

Richard Dubois (USDF, SLAC)

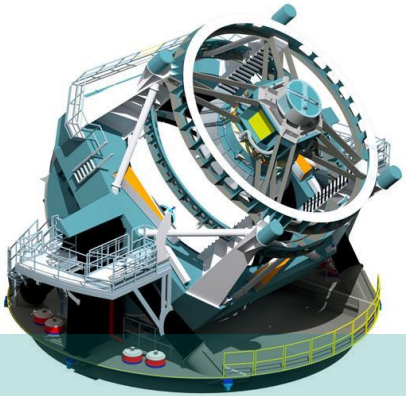
20 April 2022

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

Data Production System Vision

Raw Data: 20TB/night

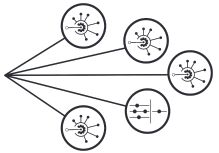
Sequential 30s images covering the entire visible sky every few days



Prompt Data Products

Alerts: up to 10 million per night

60s via nightly alert streams



Community Brokers

Alert Filtering Service

Results of Difference Image Analysis (DIA): transient and variable sources

Solar System Objects: ~ 6 million

24h via Prompt Products Database

Rubin DACs (DFs & Chile)

Independent DACs (iDACs)

Data Release Data Products

Final 10yr Data Release:

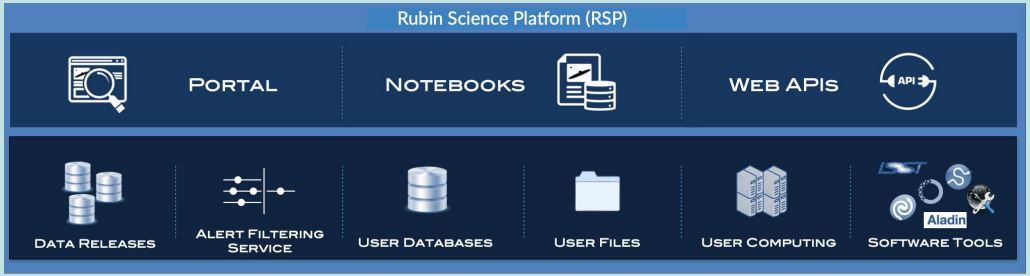
- Images: 5.5 million x 3.2 Gpx
- Catalog: 15PB, 37 billion objects

via Data Releases

Access to proprietary data and the Science Platform require Rubin data rights

Rubin Science Platform

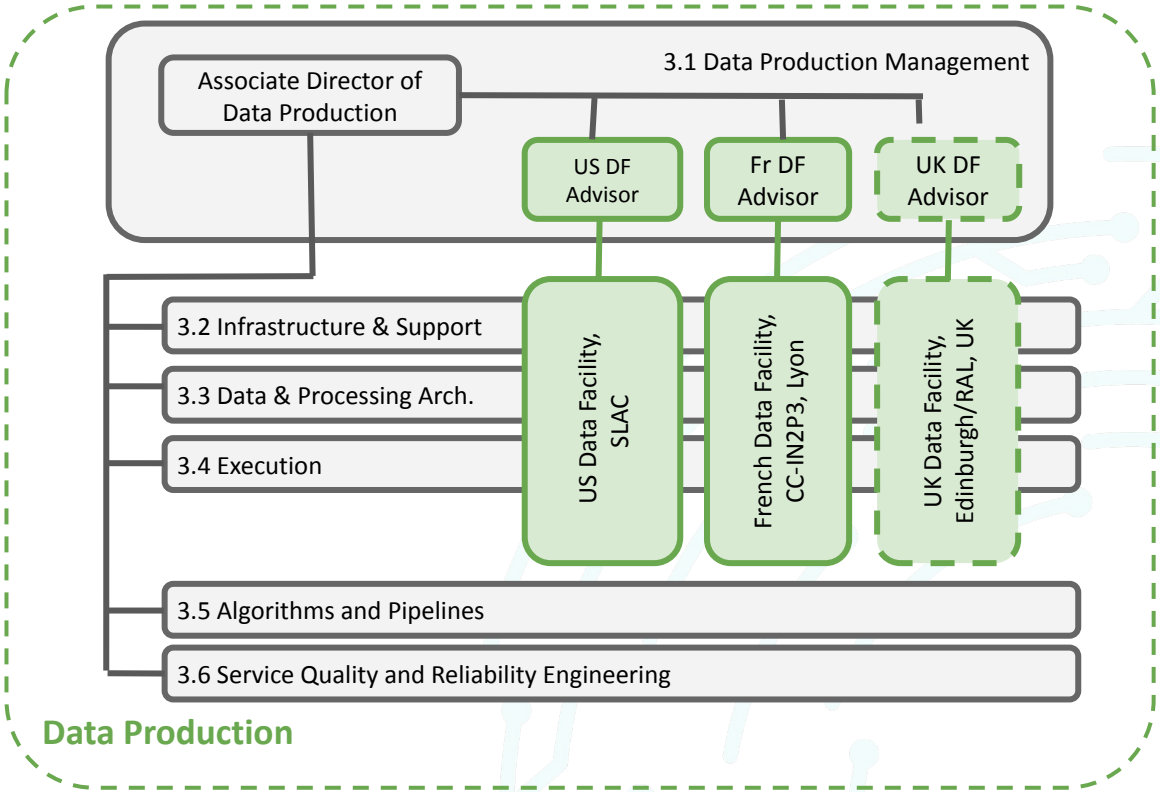
Provides access to Rubin Data Products and services for all science users and project staff



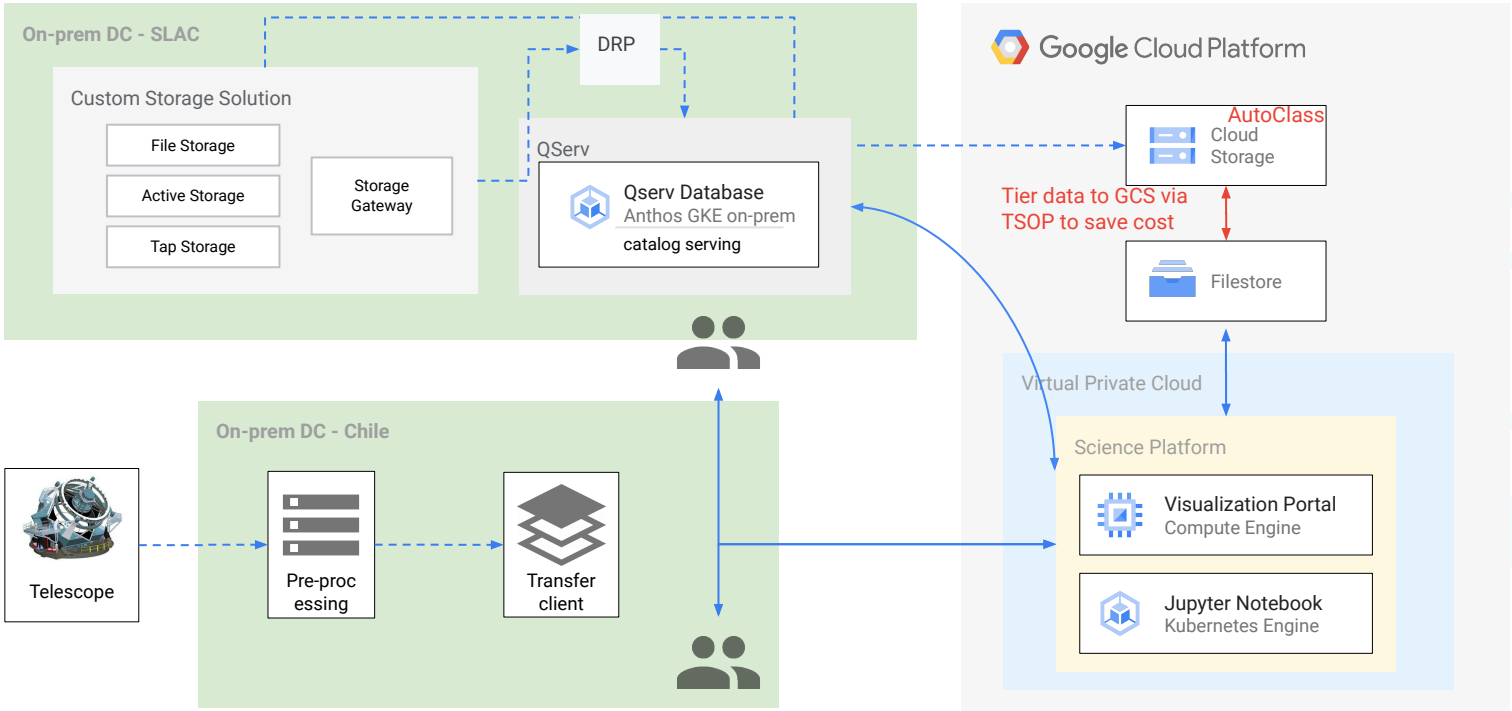
Data Production as a Matrixed, International, Multi-center Department

Multiple data facilities

- **United States Data Facility (USDF)**
  - ~10% of compute available to users
  - 25% of data release processing
- French Data Facility at CC-IN2P3
  - 50% of data release processing
- UK Data Facility
  - 25% of data release processing
- Independent Data Access Centers
  - May serve only a subset of data
- Clouds



Hybrid model: Cloud/On-prem



Sizing Model: Compute

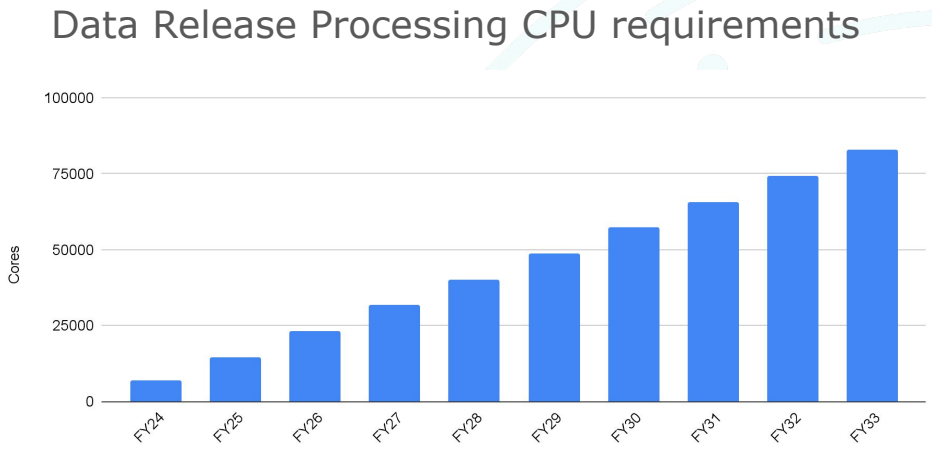
Size estimates based on:

- Benchmarking construction-era codes;
- Comparison with precursors (HSC, DES).

Rubin data products are ambitious; still some uncertainty in compute needs.

- 25% of total DRP processing power at SLAC
- ramp up from ~1700 cores in Yr 1 to ~20,000 cores in Yr 10 for ~200 day annual turnaround
- Current plan: AMD Romes 128 cores/512 GB RAM

Prompt Processing (nightly alert generation, daily solar system processing) is sized to consume ~1200 cores continuously when observing.

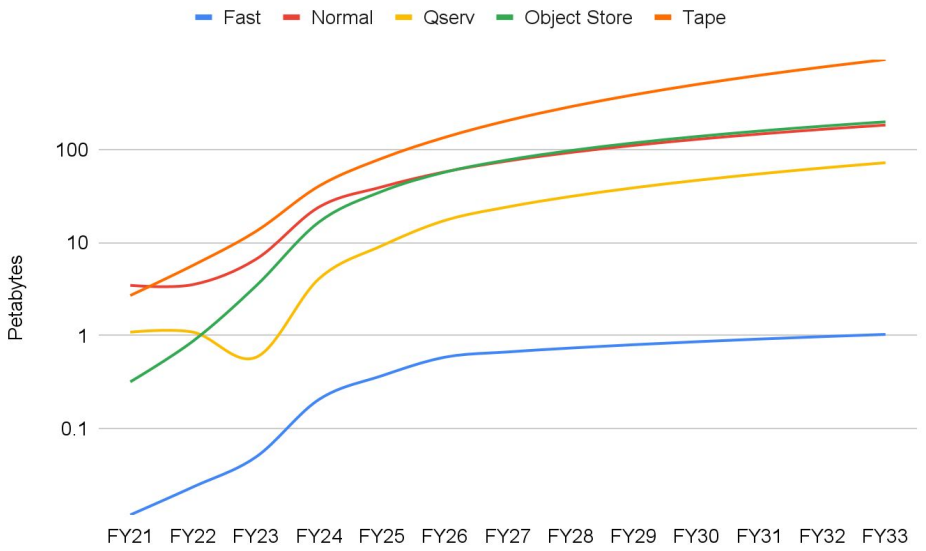


Sizing Model: Storage

Storage needs are relentless: data will flow and we have to process and store.

However the raw data rate is steady: once the system is bootstrapped it is easy to anticipate.

The first years will be the most challenging

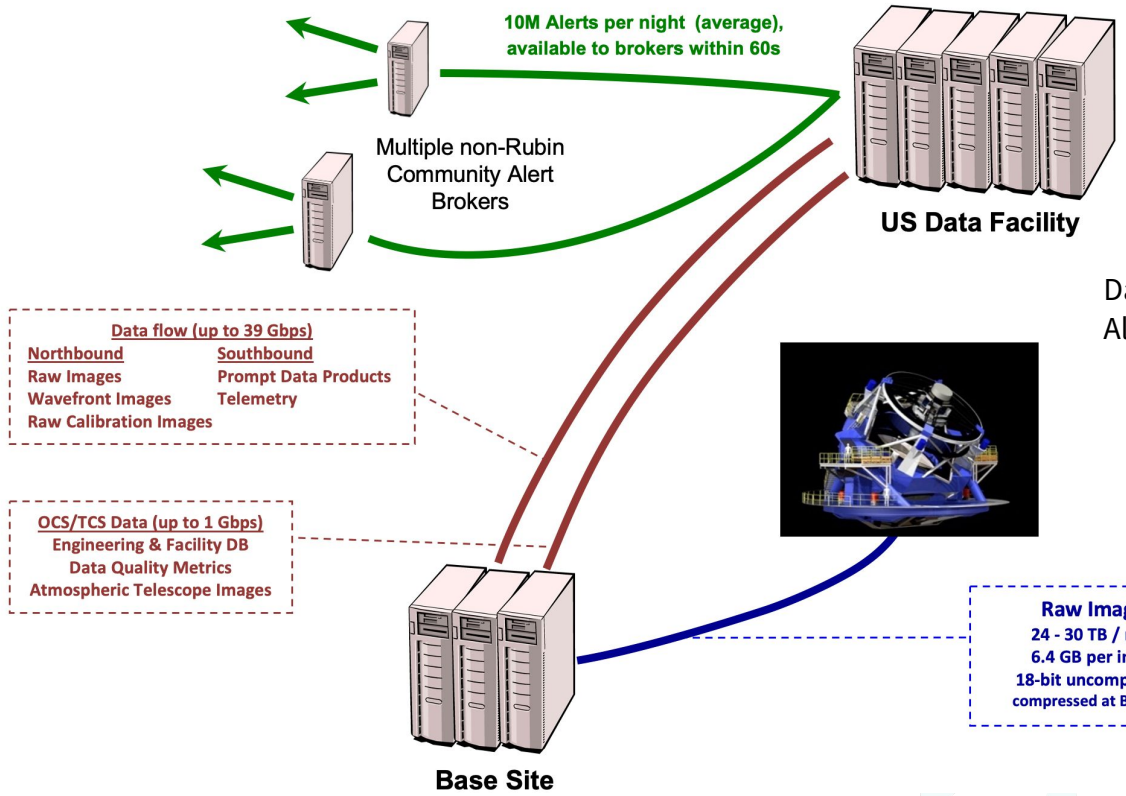


Legend:
Fast - SSD
Normal - regular POSIX R/W disk - used during DRP
Qserv - local disk per node
Object store - R/O by url
Tape - well, tape.

Now thinking that all image data (rw, ro) will be object store

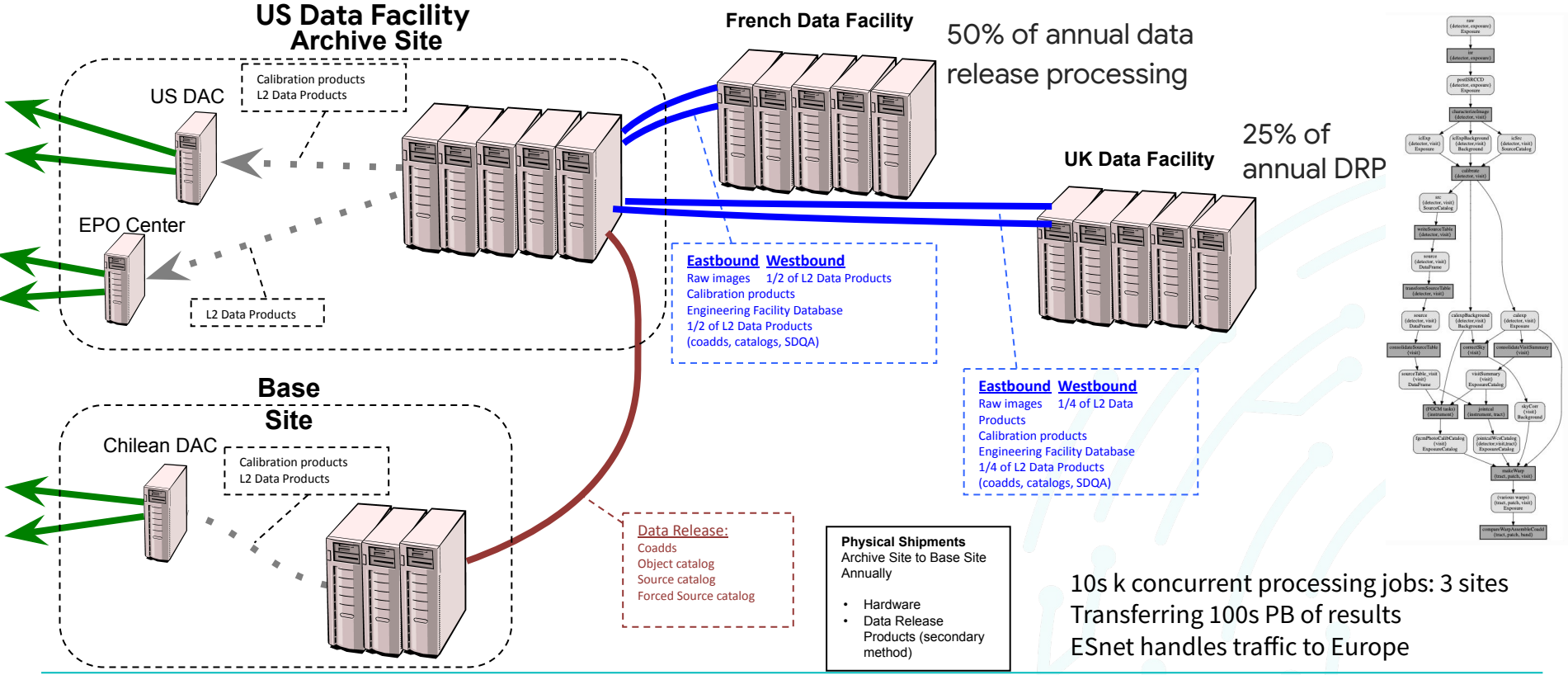
Pretty Big Data

Data Flows: Prompt Processing



Data to SLAC in 7 secs
Alerts generated in 1 min

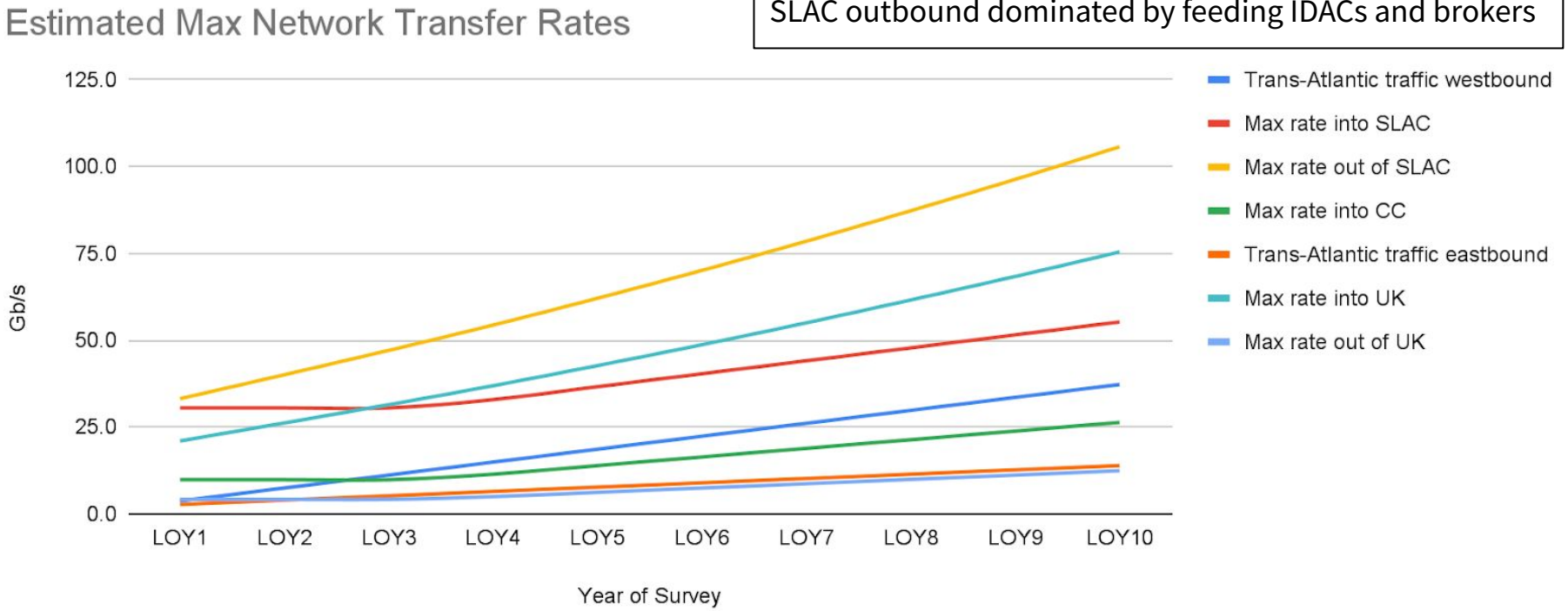
Data Flows: Data Release Processing



10s k concurrent processing jobs: 3 sites
Transferring 100s PB of results
ESnet handles traffic to Europe

Projected Network Transfer Rates

SLAC outbound dominated by feeding IDACs and brokers



Assumes DRP transfers can proceed in parallel with processing

Technologies Adopted for Multi-Site

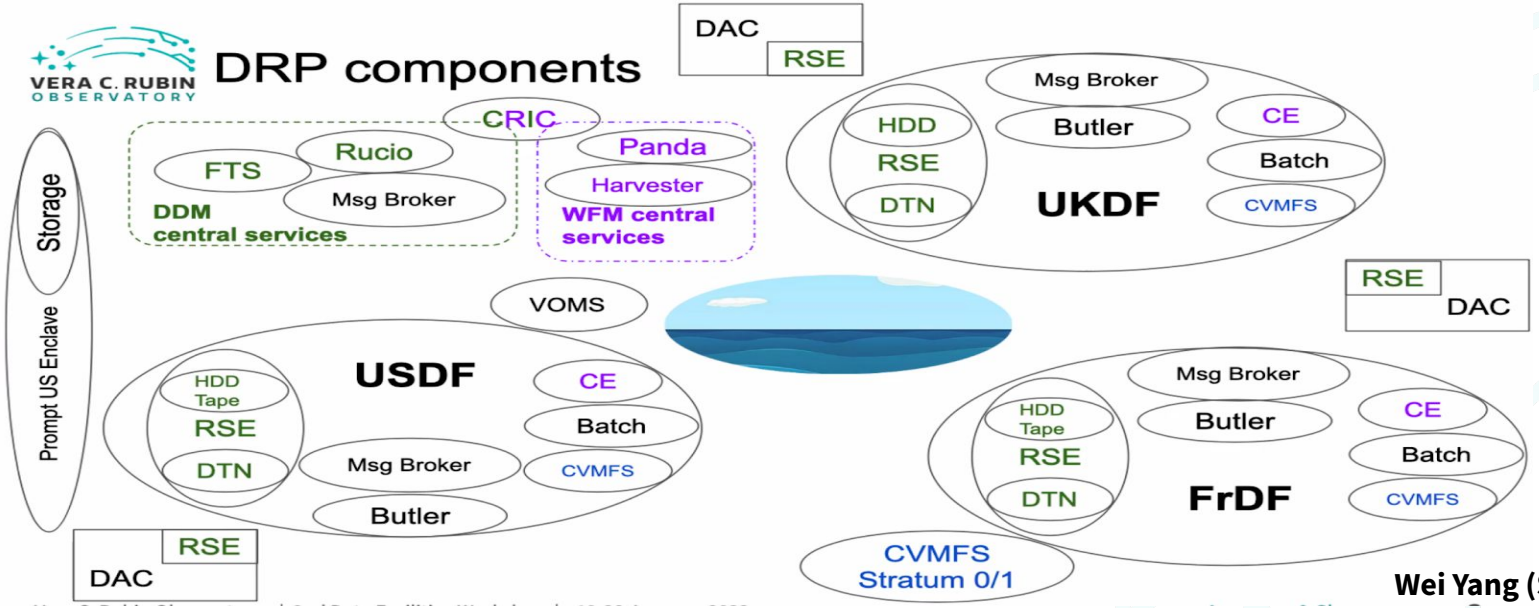
- PanDA - Workflow management
  - In use for DP0.2 in the Google Cloud Interim Data Facility
  - Multi-site testing just getting underway
- Rucio - Data movement
  - Multi-site testing underway, including to Chile
  - Envisaged as the basis of the Data Backbone
  - No alternatives in mind
- cvmfs - code distribution
  - Stratum 0 hosted by CC-IN2P3 and in use for Rubin code in a variety of places
  - There are other options, but this appears to work

Technologies Adopted

CE: PanDA Compute Element
RSE: Rucio Storage Element
DTN: Data transfer node
DAC: Data Access Center

- Workflow: PanDA
- Data movement: Rucio/FTS
- Code distributions: cvmfs

Facing the issue of bolting them onto butler



Rucio: Well-supported by HEP/LHC

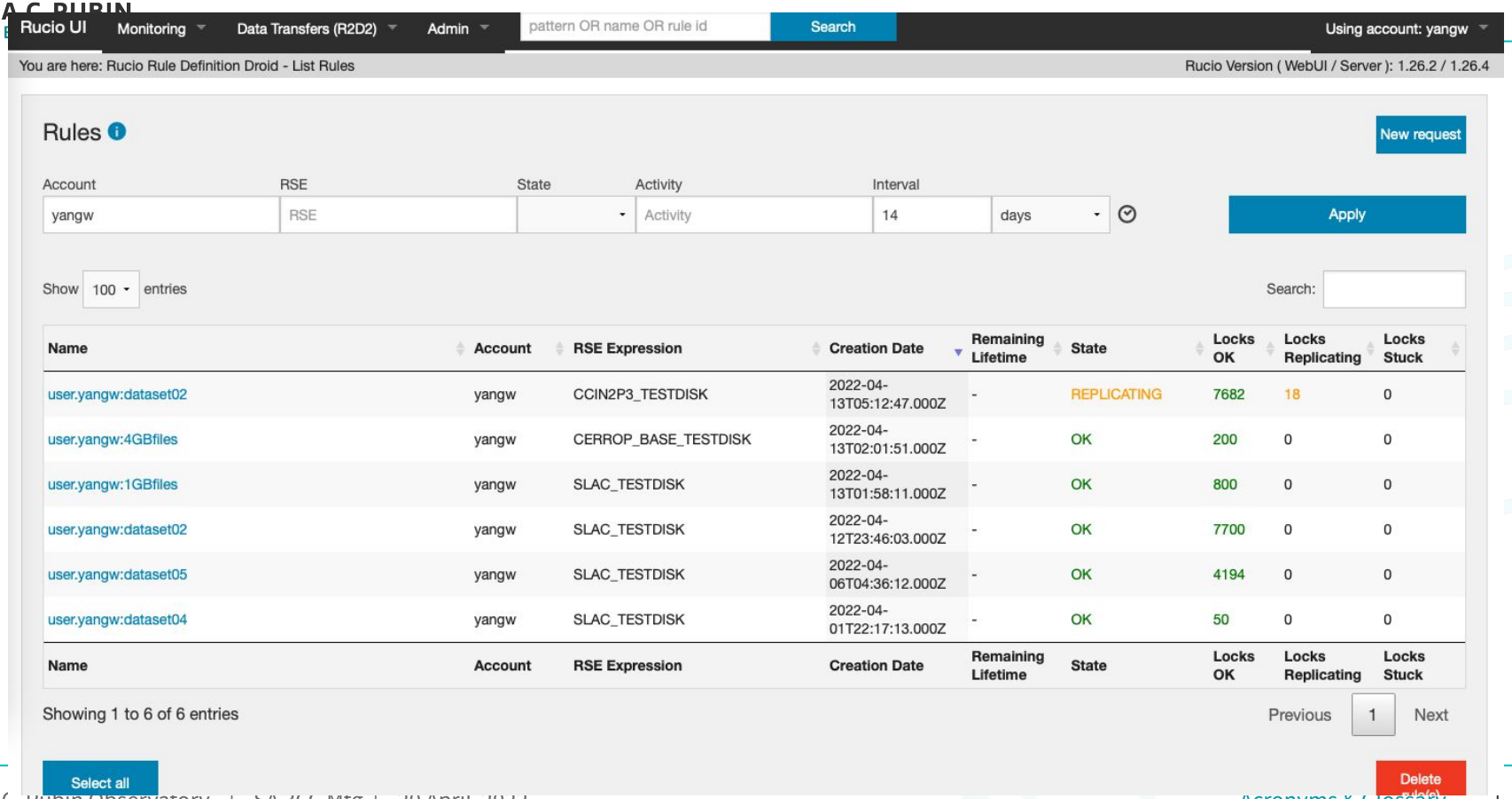
Overview

Rucio in a nutshell

Presented by Rucio developers

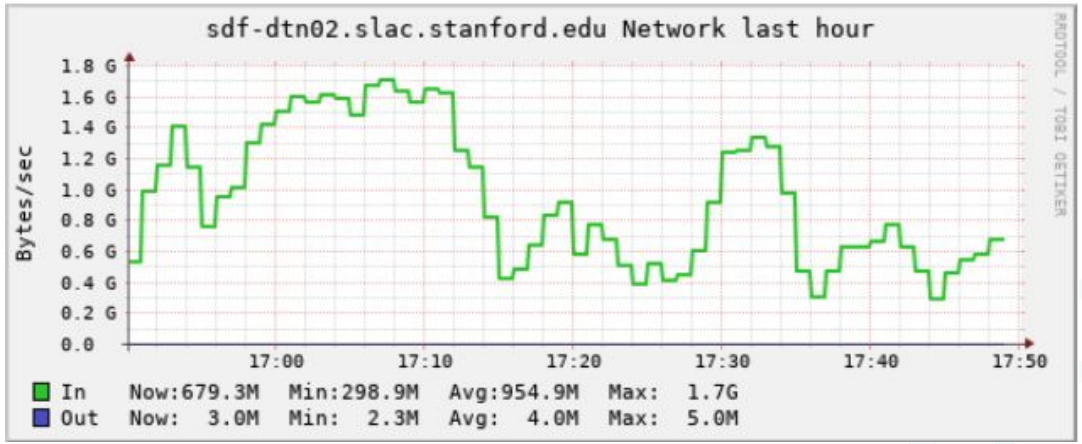
- Initially developed by the ATLAS experiment
- Provides services and libraries for scientific collaborations/experiments/communities
  - Designed with more than 10 years of operational experience in data management
  - Full, complete and generic data management service
  - The number of data intensive instruments generating unprecedented data volume is growing
- Store, manage, and process data in a heterogeneous distributed environment
  - Data can be scientific observations, measurements, objects, events, images saved in files
  - Manage transfers, deletions, and storage
  - Connects with workflow management systems
  - Supports both low-level and high-level policies and enforces them
  - A rich set of advanced features and use cases supported
  - Facilities can be distributed at various locations belonging to different administrative domain

Rucio Web UI


Rucio UI
Monitoring
Data Transfers (R2D2)
Admin
pattern OR name OR rule id
Search
Using account: yangw
You are here: Rucio Rule Definition Droid - List Rules
Rucio Version (WebUI / Server): 1.26.2 / 1.26.4
Rules
New request
Account
RSE
State
Activity
Interval
yangw
RSE
Activity
14
days
Apply
Show 100 entries
Search:
Name
Account
RSE Expression
Creation Date
Remaining Lifetime
State
Locks OK
Locks Replicating
Locks Stuck
user.yangw:dataset02
yangw
CCIN2P3_TESTDISK
2022-04-13T05:12:47.000Z
-
REPLICATING
7682
18
0
user.yangw:4GBfiles
yangw
CERROP_BASE_TESTDISK
2022-04-13T02:01:51.000Z
-
OK
200
0
0
user.yangw:1GBfiles
yangw
SLAC_TESTDISK
2022-04-13T01:58:11.000Z
-
OK
800
0
0
user.yangw:dataset02
yangw
SLAC_TESTDISK
2022-04-12T23:46:03.000Z
-
OK
7700
0
0
user.yangw:dataset05
yangw
SLAC_TESTDISK
2022-04-06T04:36:12.000Z
-
OK
4194
0
0
user.yangw:dataset04
yangw
SLAC_TESTDISK
2022-04-01T22:17:13.000Z
-
OK
50
0
0
Showing 1 to 6 of 6 entries
Previous
1
Next
Select all
Delete

CC-IN2P3 to SLAC

Mixture of 3506 large and medium size files (100MB, 1GB, 4GB), ~3.5TB
No attempts at network tuning yet



Source	Destination	VO	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	Rate (last 1h)	Thr.
+ davs://ccdavlsst.in2p3.f	davs://sdf-dtn02.slac.stanford.edu	lsst	221	150	-	-	-	3134	-	-	100.00 %	1156.24 MB/s
			221	150	0	0	0	3134	0	0	100.00 %	-

Status and Issues

- We are kicking the tires now for multi-site testing - making sure all the connections work and testing network throughput
- Issues
  - Rucio is not appropriate for low latency transfers - we'll need a different solution for the summit → SLAC transfers
  - We have mostly small files (10s of MB) and a LOT of them
  - In fact, we're expecting more than 10^9 per year - more than current LHC expts
  - We'll be moving 100s of PB by year 10 - we depend on being able to transfer files across the Atlantic as we make them. We don't know yet if there are any global gather steps in our pipeline graphs.
  - We are bolting Rucio to an existing metadata handling tool written internally called Butler. There is overlap in functionality that we to have diagonalise.
    - In particular we need bulletproof connections between the two for registering datasets in both tools
    - Minimize trans-Atlantic traffic to the central Butler/Rucio servers at SLAC

Operations timeline

DP1 & DP2 - multi-site practice
DR1 in late 2024 - real thing

Data Preview/Release		Date	Milestone Date
DP0.1	DC2 Simulated Sky Survey	June 2021	2021-09-30
DP0.2	Reprocessed DC2 Survey	June 2022	2022-06-30
DP1	ComCam On-Sky Data	Jan 2024 - Apr 2024	2024-01-01
DP2	LSSTCam On-Sky Data	Jul 2024 - Sep 2024	2024-07-01
DR1	LSST First 6 Months Data	Apr 2025 - Jul 2025	2025-04-01
DR2	LSST Year 1 Data	Sep 2025 - Dec 2025	2025-09-30
DR3	LSST Year 2 Data	Oct 2026 - Dec 2026	2026-10-01

US DF Development Activity		Start Date	Milestone Date
Start-up	Hardware Installation	October 2021	2021-10-01
Start-up	Test Processing	March 2022	2022-03-02
DP1	Hardware Installation	January 2023	2023-01-01
DP1	Processing	June 2023	2023-06-02
DP1	Data Release	January 2024	2024-01-01
DP2	Hardware Installation	July 2023	2023-07-02
DP2	Processing	December 2023	2023-12-01
DP2	Data Release	July 2024	2024-07-01
DR1	Hardware Installation	April 2024	2024-04-01
DR1	Processing	August 2024	2024-08-31
DR1	Data Release	April 2025	2025-04-01
DR2	Hardware Installation	September 2024	2024-09-30
DR2	Processing	March 2025	2025-03-01
DR2	Data Release	September 2025	2025-09-30
DR3	Hardware Installation	October 2025	2025-10-01
DR3	Processing	March 2026	2026-03-02
DR3	Data Release	October 2026	2026-10-01
DR4	Hardware Installation	October 2026	2026-10-01
DR4	Processing	March 2027	2027-03-02
DR4	Data Release	October 2027	2027-10-01