VERA C. RUBIN
OBSERVATORY

Rubin Observatory Multi-Site Testing

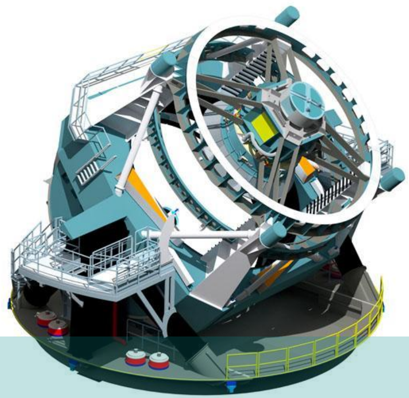
Richard Dubois (USDF, SLAC)

1 May 2024

Data Management System Vision

Raw Data: 20TB/night

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



LSST
SCIENCE
PIPELINES

Prompt Data Products

Alerts: up to 10 million per night

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

Solar System Objects: ~ 6 million

Data Release Data Products

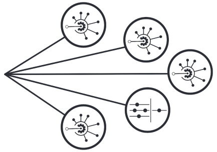
Final 10yr Data Release:

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

60s via nightly alert streams

24h via Prompt Products Database

via Data Releases



Community Brokers

Alert Filtering Service

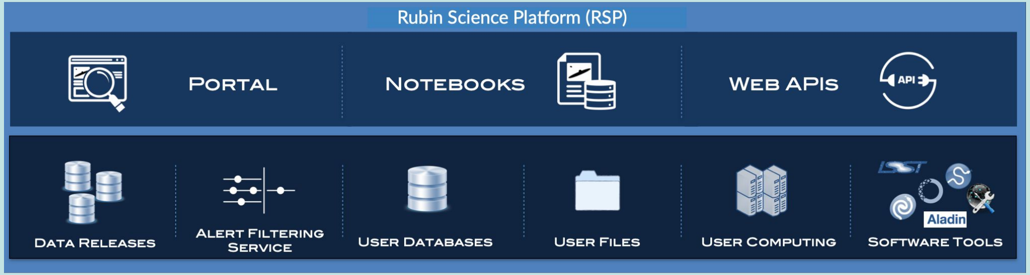
Rubin DACs (DFs & Chile)

Independent DACs (iDACs)

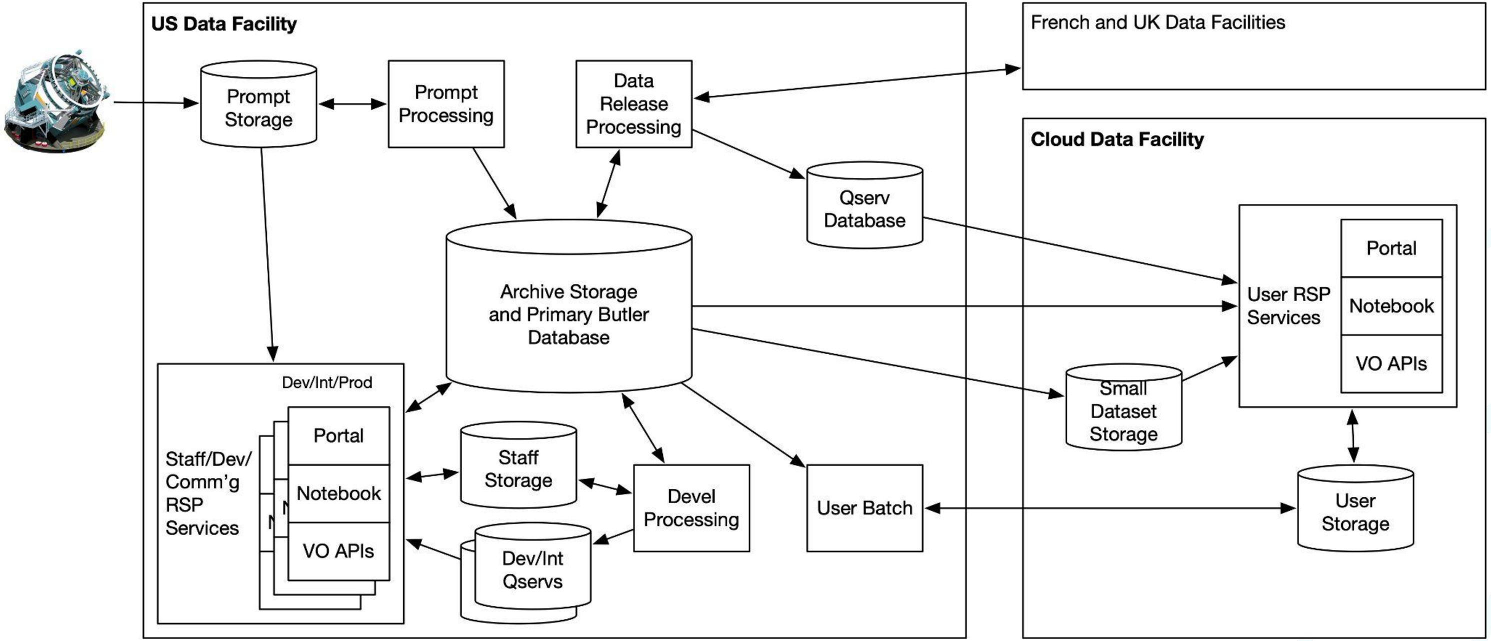
Rubin Science Platform

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

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



USDF: A Mix of On-prem and Cloud



Hybrid model: Data at SLAC but users on the Cloud.

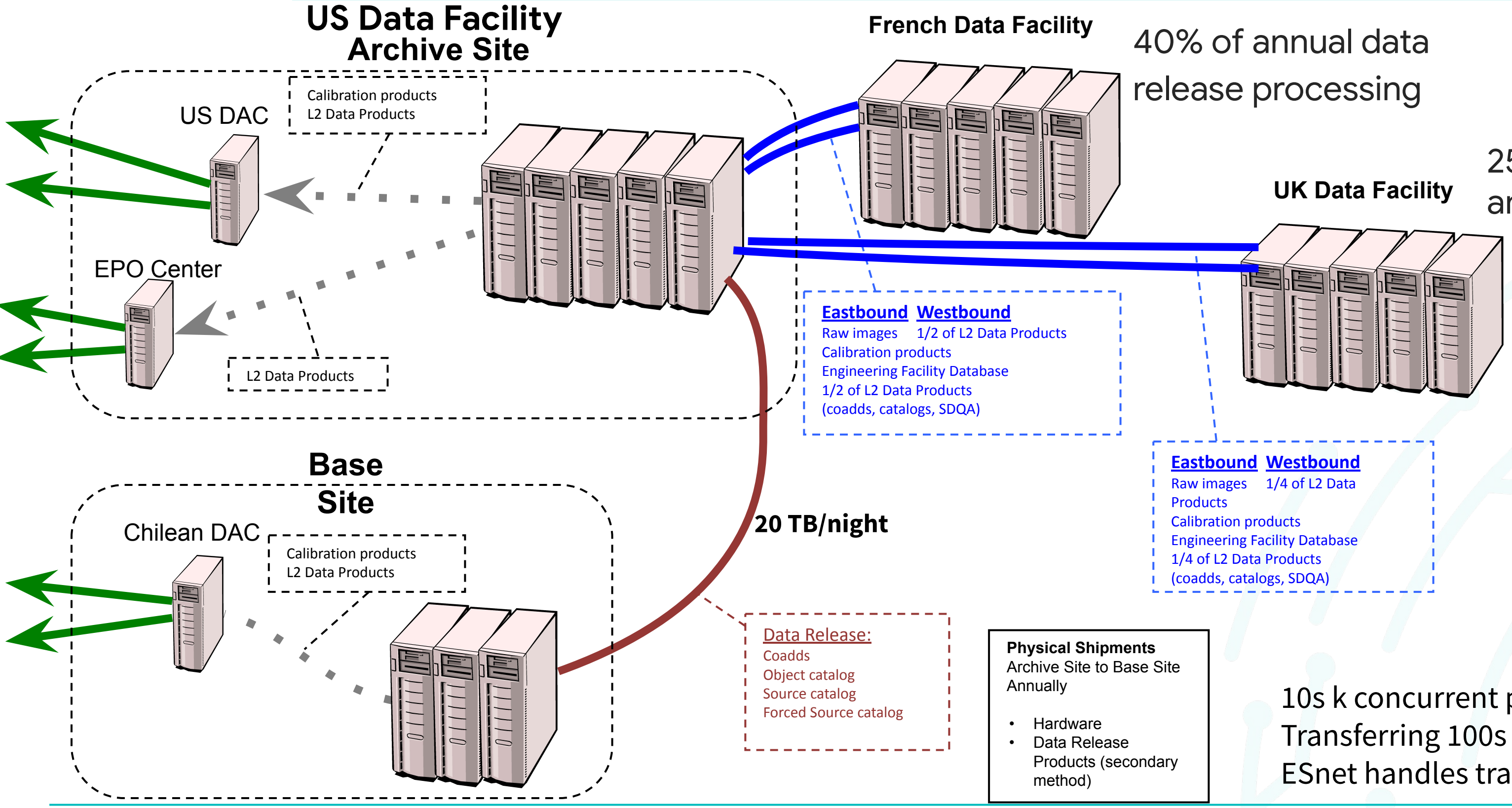
Allows:

- Separation of security concerns
- Burst response
- Reduced risk

(see DMTN-209)

RSP = Rubin Science Platform

Data Flows: Prompt & Data Release Processing



US Data Facility Archive Site

US DAC

Calibration products
L2 Data Products

EPO Center

L2 Data Products

French Data Facility

40% of annual data release processing

Eastbound **Westbound**

Raw images 1/2 of L2 Data Products
Calibration products
Engineering Facility Database
1/2 of L2 Data Products
(coadds, catalogs, SDQA)

UK Data Facility

25% of annual DRP

Eastbound **Westbound**

Raw images 1/4 of L2 Data Products
Calibration products
Engineering Facility Database
1/4 of L2 Data Products
(coadds, catalogs, SDQA)

20 TB/night

Base Site

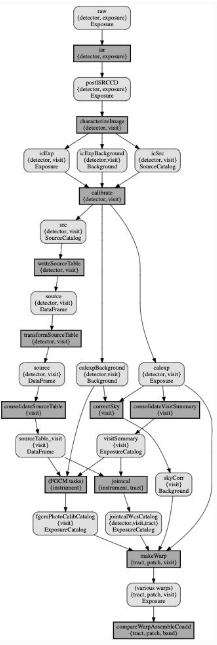
Chilean DAC

Calibration products
L2 Data Products

Data Release:
Coadds
Object catalog
Source catalog
Forced Source catalog

Physical Shipments
Archive Site to Base Site
Annually

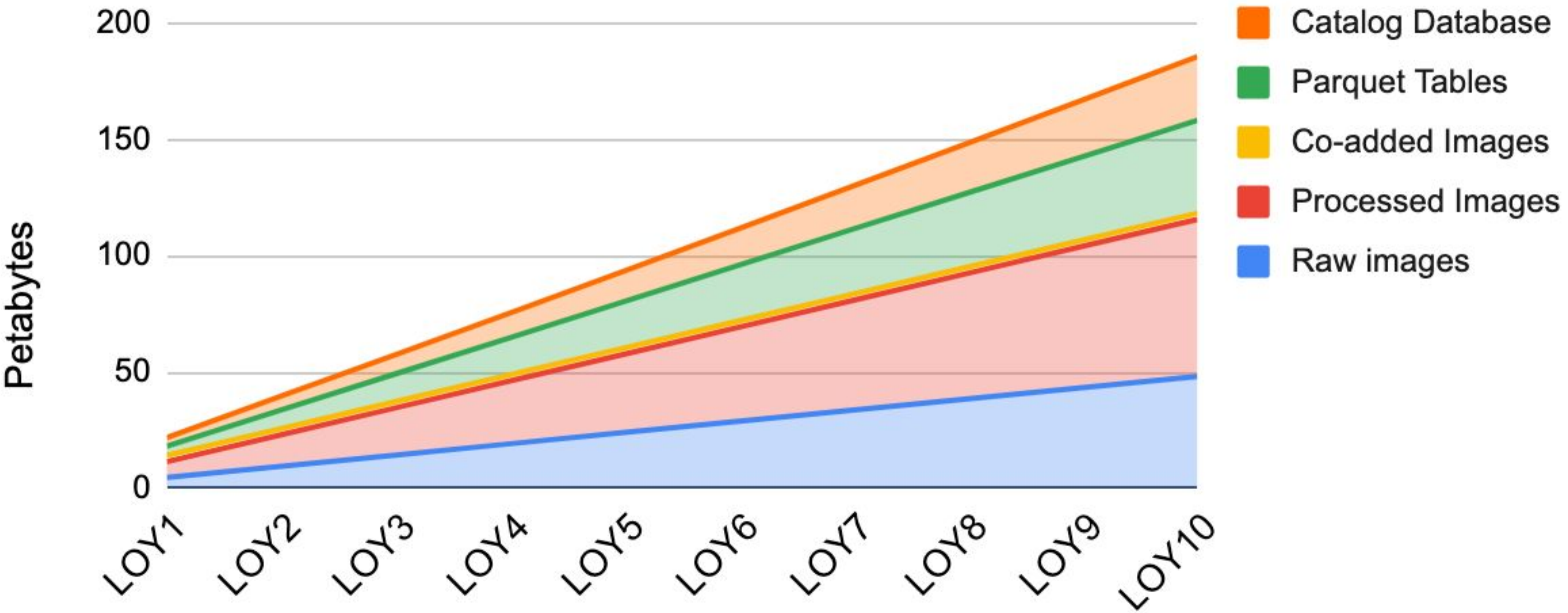
- Hardware
- Data Release Products (secondary method)



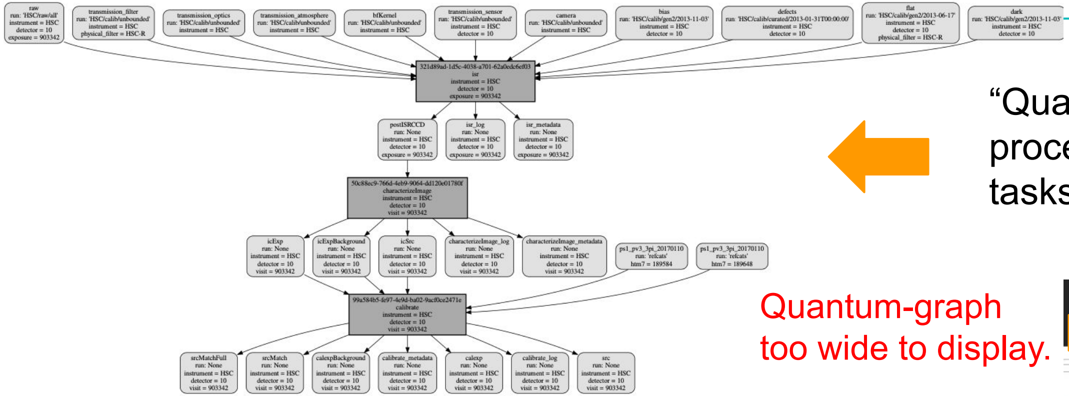
raw {detector, exposure} Exposure
isr {detector, exposure}
postISRCCD {detector, exposure} Exposure
characterizeImage {detector, visit}
icExp {detector, visit} Exposure
icExpBackground {detector,visit} Background
icSrc {detector, visit} SourceCatalog
calibrate {detector, visit}
src {detector, visit} SourceCatalog
writeSourceTable {detector, visit}
source {detector, visit} DataFrame
transformSourceTable {detector, visit}
source {detector, visit} DataFrame
calexpBackground {detector,visit} Background
calexp {detector, visit} Exposure
consolidateSourceTable {visit}
correctSky {visit}
consolidateVisitSummary {visit}
sourceTable_visit {visit} DataFrame
visitSummary {visit} ExposureCatalog
(FGCM tasks) {instrument}
jointcal {instrument, tract}
skyCorr {visit} Background
fgcmPhotoCalibCatalog {visit} ExposureCatalog
jointcalWcsCatalog {detector,visit,tract} ExposureCatalog
makeWarp {tract, patch, visit}
(various warps) {tract, patch, visit} Exposure
compareWarpAssembleCoadd {tract, patch, band}

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

Pretty Big Data

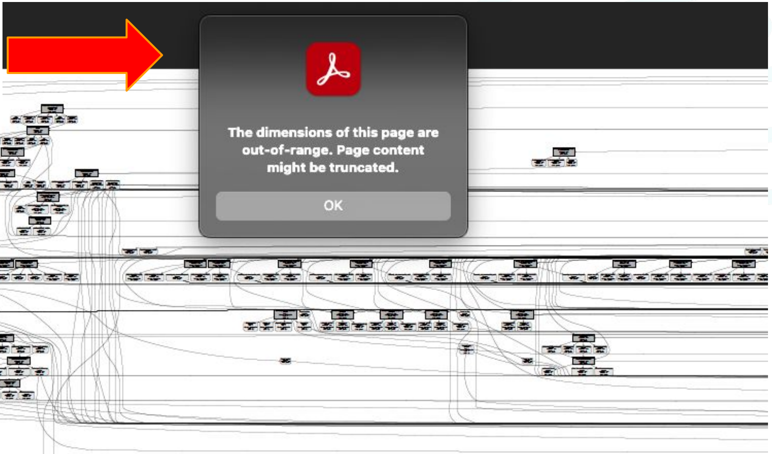


Complexity of Rubin Data Processing Pipeline



"Quantum Graph" of Rubin Science Pipeline to process a single LSST CCD image, showing 3 tasks (top to bottom) and input/outputs

Quantum-graph too wide to display.



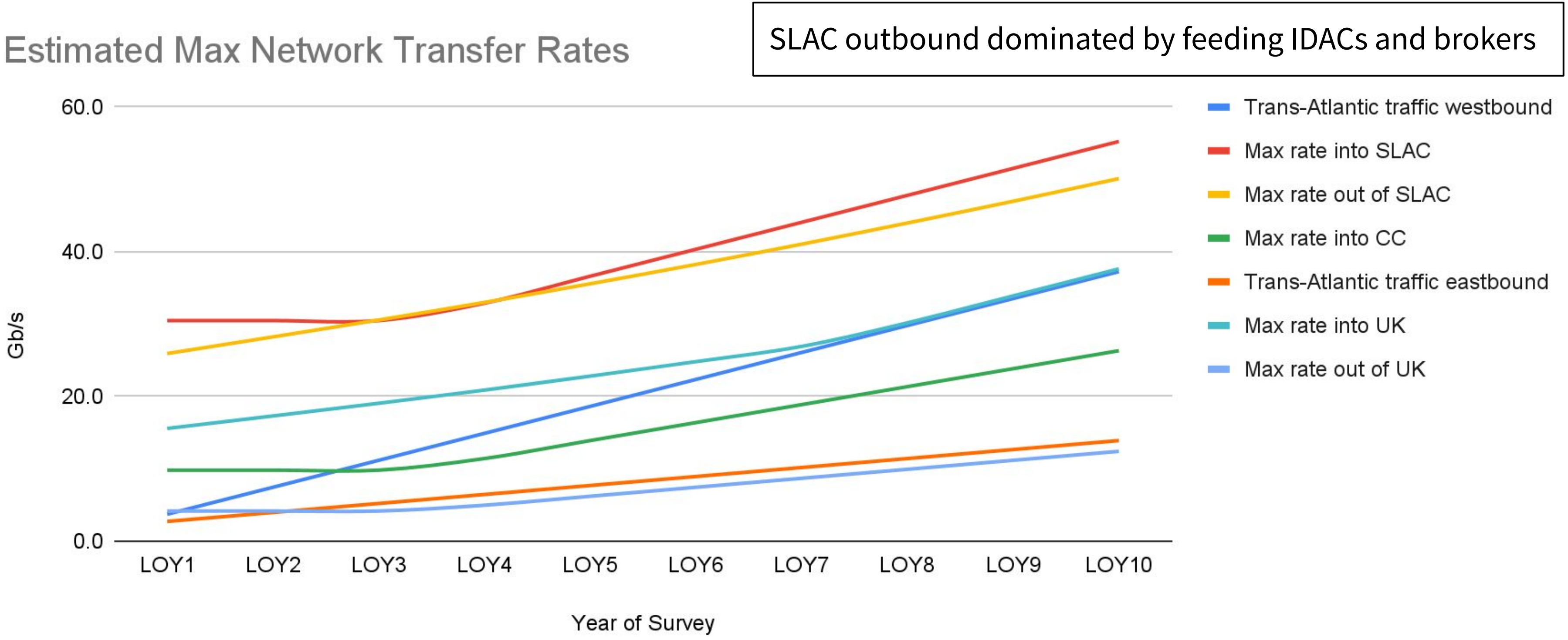
The dimensions of this page are out-of-range. Page content might be truncated.

Rubin Science Pipeline to perform single frame and coadd processing based on HSC engineering test data

Actual Rubin DRP pipeline will be a lot more complicated than the above, **They are grouped into 7 steps**

Projected Network Transfer Rates

SLAC outbound dominated by feeding IDACs and brokers



Assumes DRP transfers can proceed in parallel with processing

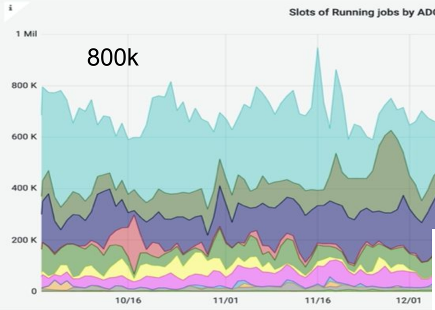
Technologies Adopted for Multi-Site

- PanDA - Workflow management
  - Used for DP0.2 in the Google Cloud Interim Data Facility
  - Exercised for routine CI reprocessing and HSC-PDR2
  - Multi-site testing underway
- Rucio - Data management & movement (FTS)
  - Data replication demonstrated to all Facility sites
  - Wrapping up interface to Butler
- 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

Workflow & Workload Management System

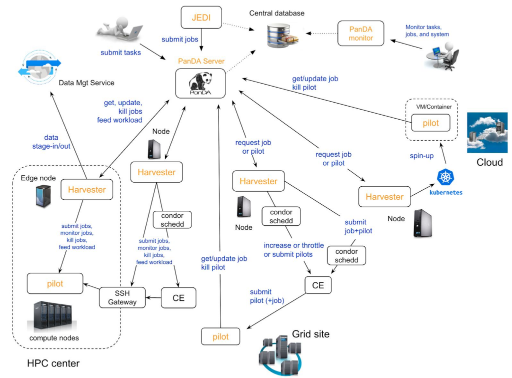
Rubin Batch Production Service (BPS) will use

- Panda/JEDI to manage workload
  - Manage concurrent jobs at multi-DFs
  - This is a proven technology used by LHC ATLAS for 1+ decade
- iDDS to manage workflow
  - Handle complex dependencies in workflow
  - Rubin DAG will likely drive iDDS usage toward wide and deep.
- ARC-CE to interface between Panda and local batch
- CVMFS to distribute software environment
  - and small amount of static data

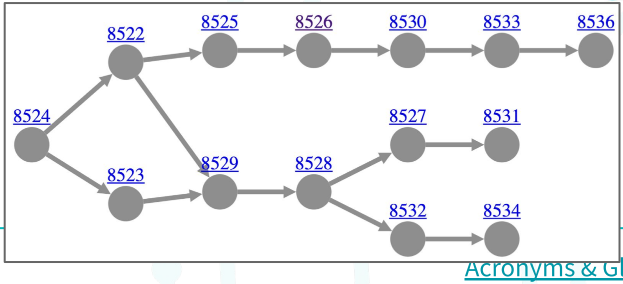


Panda workload ⬆️

Panda system diagram ➡



Rubin DAG by iDDS ⬇️

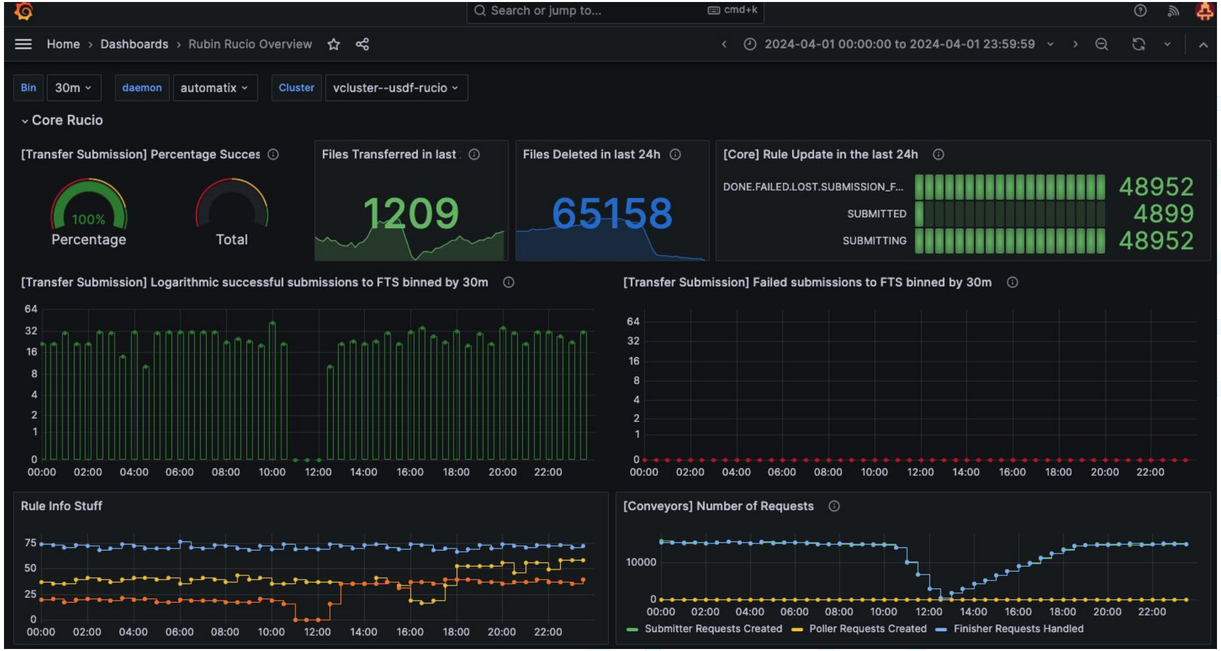


Distributed Data Management System

- **Rucio** ecosystem:
  - Rucio: also developed by LHC ATLAS and used for 1+ decade
    - data classification, keeping track of data location, drive data movement
    - Rubin will have several times more file/object in Rucio than the current LHC ones
      - A big challenge for the backend database. Rubin will drive this forward
  - FTS: also 1+ decade history
    - Think of it as a batch system dedicated to data transfer jobs.
    - Again, efficiently transfer large numbers of small files is a challenge
  - Xrootd: has been around for 2+ decades
    - Mostly used as data transfer agent, to replace GridFTP
    - Rubin prefers object stores, and is driving xrootd based data transfer to/from OS/Cloud
- Butler
  - The original Rubin data management system
  - DB of metadata and pointers to data
  - software layer to access Rubin data
  - Must coordinate with Rucio

Automatix - transfer “CI”

- Send modest number of files among the Data Facilities routinely (every ~30 mins) to test end-to-end functionality



Home › Dashboards › Rubin Rucio Overview

2024-04-01 00:00:00 to 2024-04-01 23:59:59

Bin 30m · daemon automatix · Cluster vcluster--usdf-rucio

Core Rucio

[Transfer Submission] Percentage Succes: 100% Percentage · Total

Files Transferred in last: 1209

Files Deleted in last 24h: 65158

[Core] Rule Update in the last 24h:

DONE.FAILED.LOST.SUBMISSION_F... 48952
SUBMITTED 4899
SUBMITTING 48952

[Transfer Submission] Logarithmic successful submissions to FTS binned by 30m

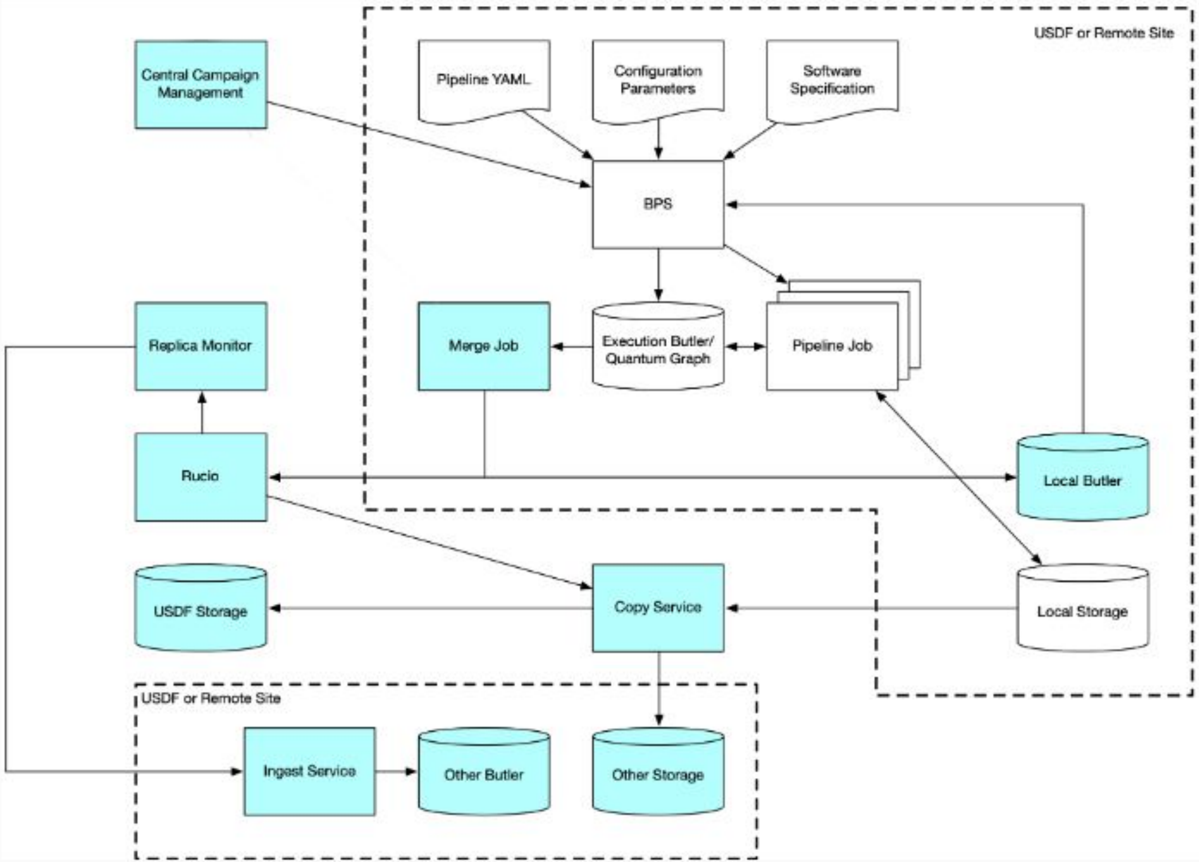
[Transfer Submission] Failed submissions to FTS binned by 30m

Rule Info Stuff

[Conveyors] Number of Requests

Pulling the Pieces Together (DMTN-213)

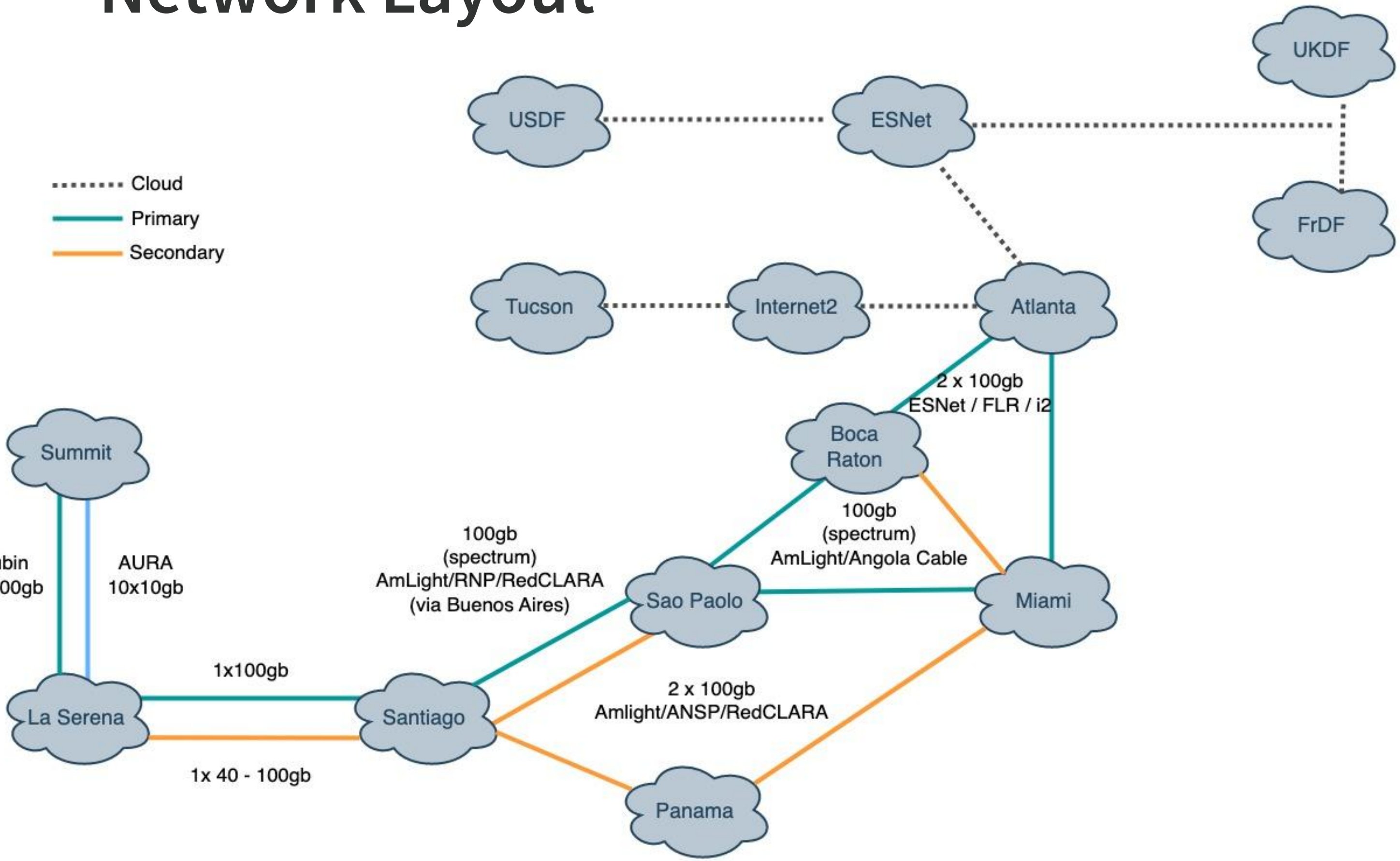
- Processing submitted to 3 sites via PanDA, where local Butlers and storage are used
- Last step in pipelines is Merge Job: register datasets to Rucio and Local Butler
- Rucio transfers the files by FTS around DFs as needed
- Replica Monitor/Ingest Service registers files in the DF Butlers triggered by a kafka message stream from the Rucio server at SLAC (DMTN-198)



Status

- Tooling
    - developed a first version of the tools required to extract from a Butler repo the files we need to transfer to another facility and to configure Rucio to drive the replication.
    - exercising these tools and noticed some issues with Rucio server that we currently trying to understand
- Throughput
    - Memory to memory tests show we can fill the available bandwidth
    - Modest rates (220MB/sec) to transfer 24 MB compressed files using parallel streams
        - We did demonstrate 5 GB/s from Fermilab during our NCSA transition, but throttled to 3.5 GB/s to not stress HPSS - for 100 GB tar files.
    - We’re looking into zipping files prior to transfer across the Atlantic
        - We plan to save the zip files to tape, rather than individual files

Network Layout



Multi-site Status

- Access to 3k cores each at FrDF and UKDF
    - Demonstrated ability to submit and run jobs there to capacity (not yet at the same time)
- Rucio installed and configured:
    - Server at SLAC; Rucio Storage Elements at each site
    - Can routinely exchange data amongst sites
        - eg: transfer 7700 files, 3.5 TB - peak rate to CC-IN2P3 of 1.4 GB/s via FTS
- HSC PDR2 reprocessing
    - First two steps complete; working on transferring output products back to USDF for global calibrations step
    - Shakedown of “Campaign Management” tools
- Automated transfers of Full Camera data from SLAC to FrDF demonstrated
    - Transfers performed by FTS3 based on Rucio rules
- Finishing up connectors between Rucio and Butler before ramping up full multi-site capability:
    - Both Rucio and Butler act as repositories of dataset information - need to keep them in sync