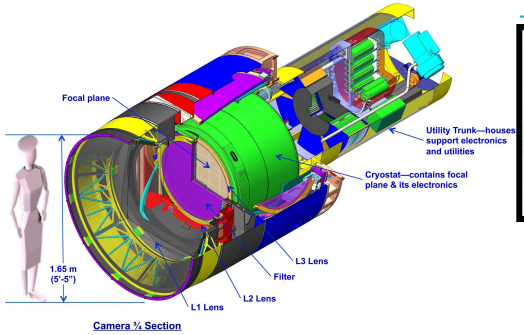
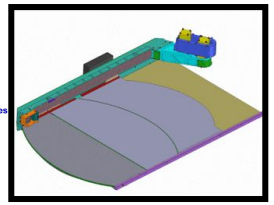
Preparation of Multi-Site Data Processing at the Vera C. Rubin Observatory

Offline Data Release Production

B.Yanny E.Karavakis F.Hernandez J.Adelman-McCarthy K.-T.Lim M.Gower P.Love R.Dubois S.Pietrowicz T.Jenness T.Noble W.Guan W.Yang Z.Yang on behalf of the Vera C. Rubin Observatory

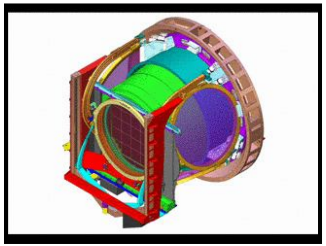
The Camera and the Observatory





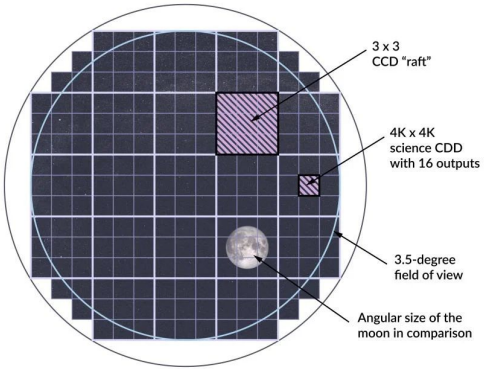
Shutter

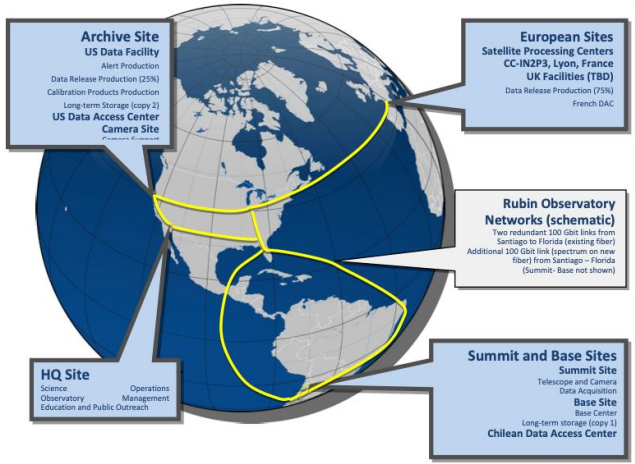
EM-ware filters (bands)



Telescope : 8.4m diameter

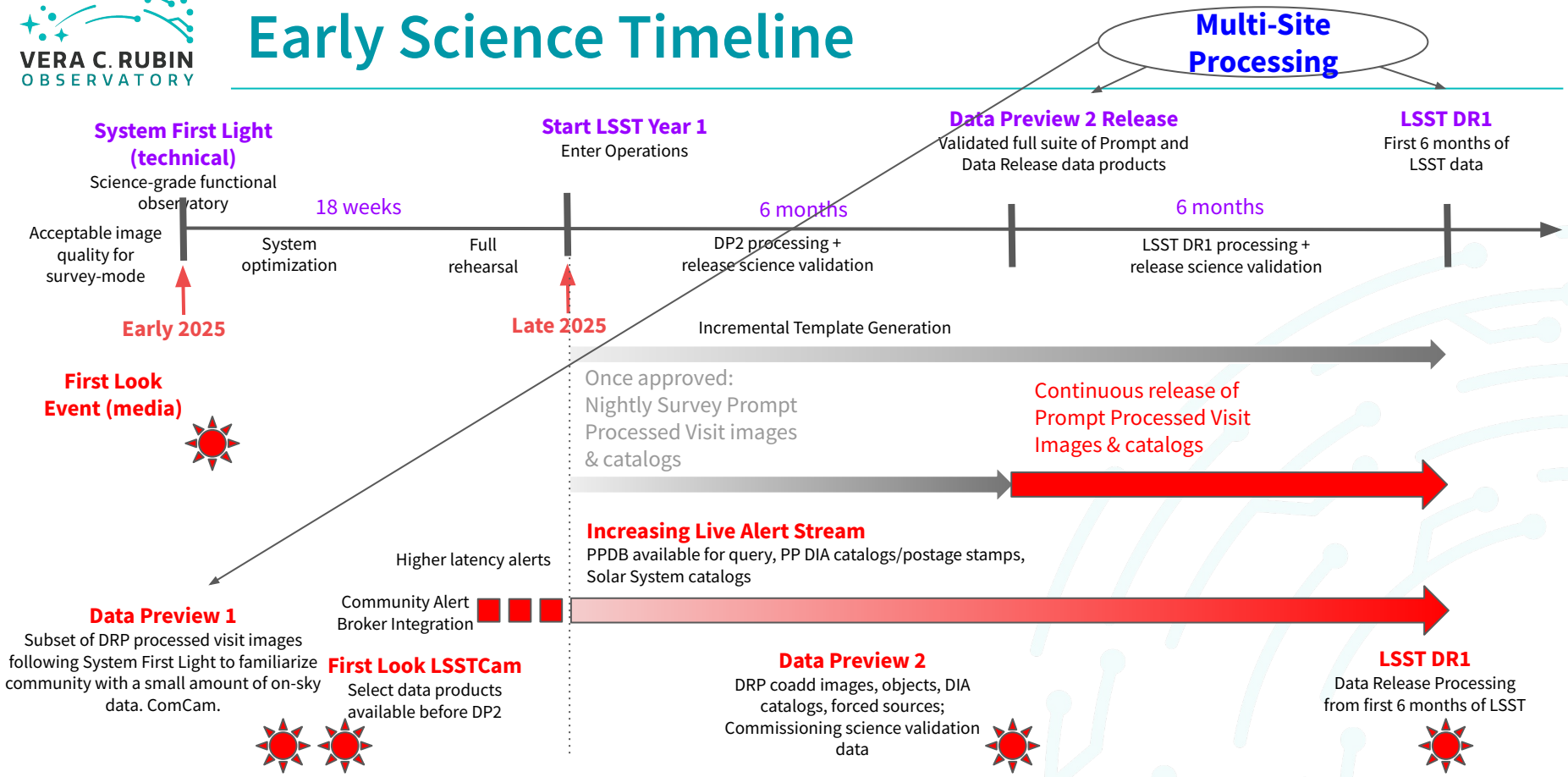
Camera: 3.2G pixel





- Legacy Survey of Space and Time (LSST) Camera @ Cerro Pachon, Chile
- All visible southern sky: 18000 deg^2 in 6 EM bands.
- 10 year survey starting at 2025
- **Offline data processing will be at US, France and UK**

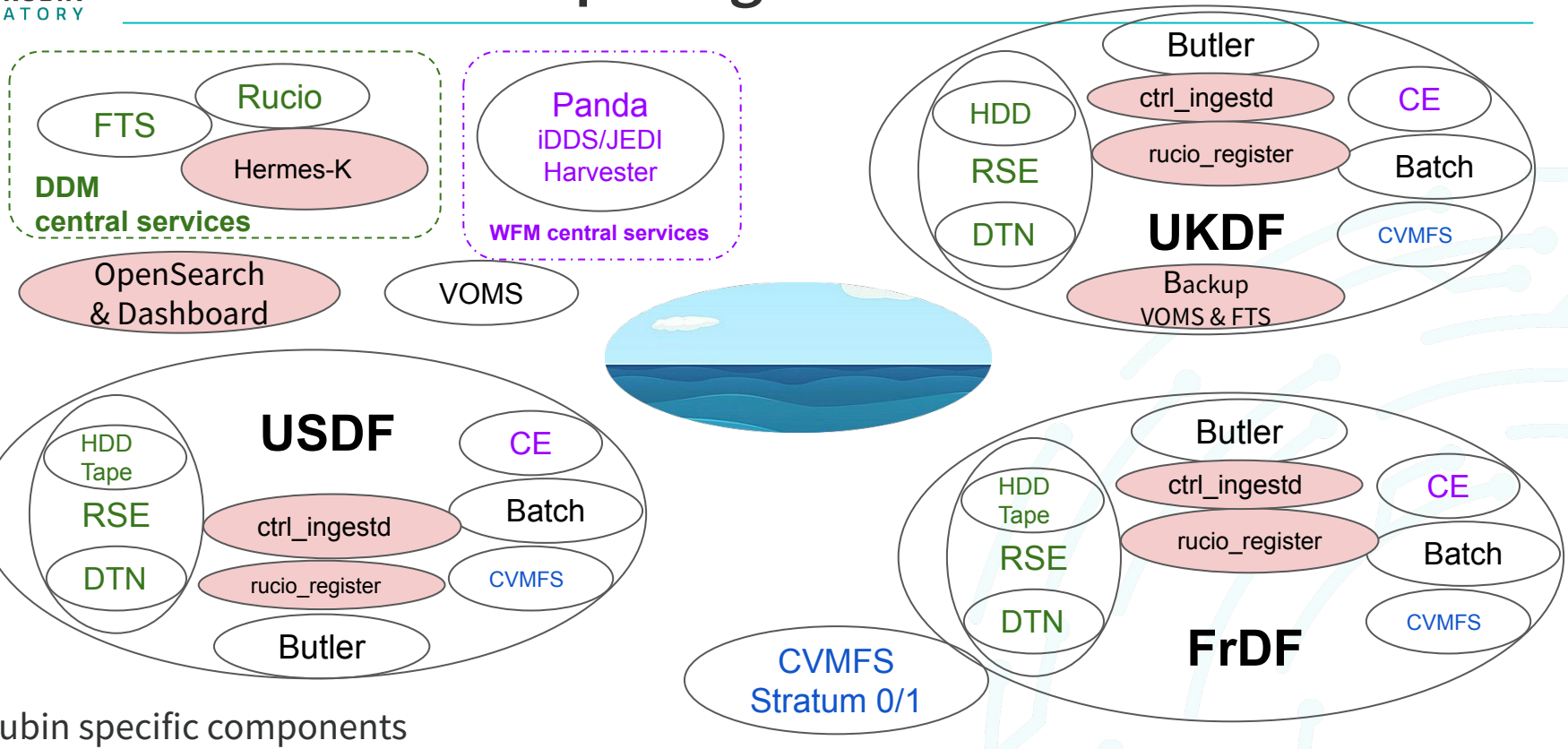
Early Science Timeline



Details: [Operations Timeline Data Releases]

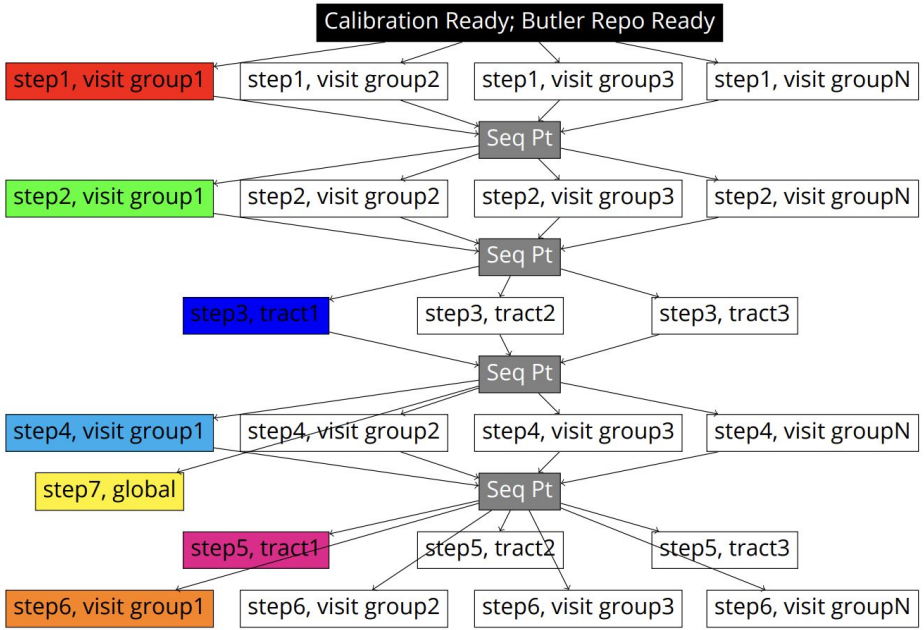
Acronyms & Glossary

Distributed Computing Infrastructure



Campaign Management for Multi-Site Processing

- Campaign Management (CM) is a Python interface to a backend database which organizes, launches and monitors data processing campaigns for Rubin.
- Campaigns have a hierarchical structure: campaign → steps → groups → jobs
- CM has single campaign database to design and launch campaigns, is able to direct jobs to any Data Facility for execution based on algorithm, can monitor all DFs and combine outputs from several DFs when needed for special 'Seq point'.

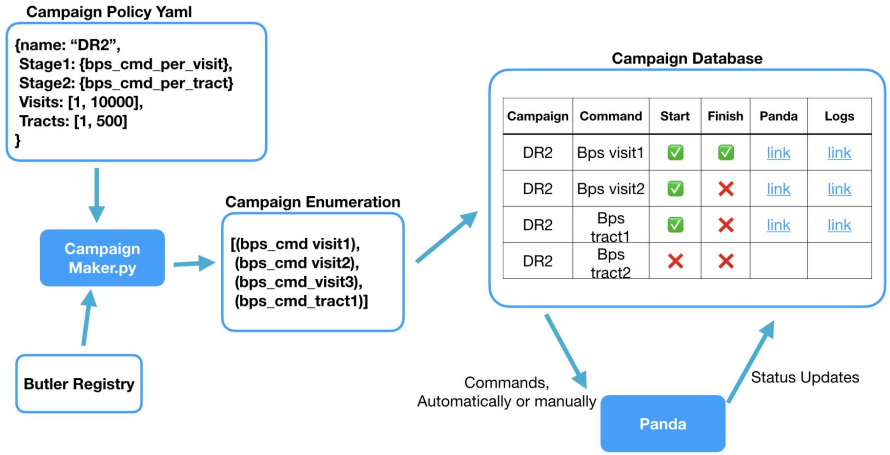


Multi-site Full Chain Plan

1. Perform monthly small (3 sq. degree) tests at individual DFs with latest software release.
2. Process 2 tract (5 sq. degree) sky area at each of three Data Facilities (DF): FrDF, UKDF, USDF
3. Move star catalogs to USDF for global calibration fit. Distribute solution to: FRDF, UKDF
4. Complete processing at 3 DFs. Return retained outputs to USDF for analysis, tape archival
5. Repeat steps 2-4 for larger 70 tract (200 sq. degree) sky area (3PB data volume) test.
6. Upgrade CPU (10K CPU cores) and storage (30PB). Network is in place.
7. Roll out multi-site production to handle data from LSSTCam (3.2 Gpixel camera on telescope) starting in late-2025.

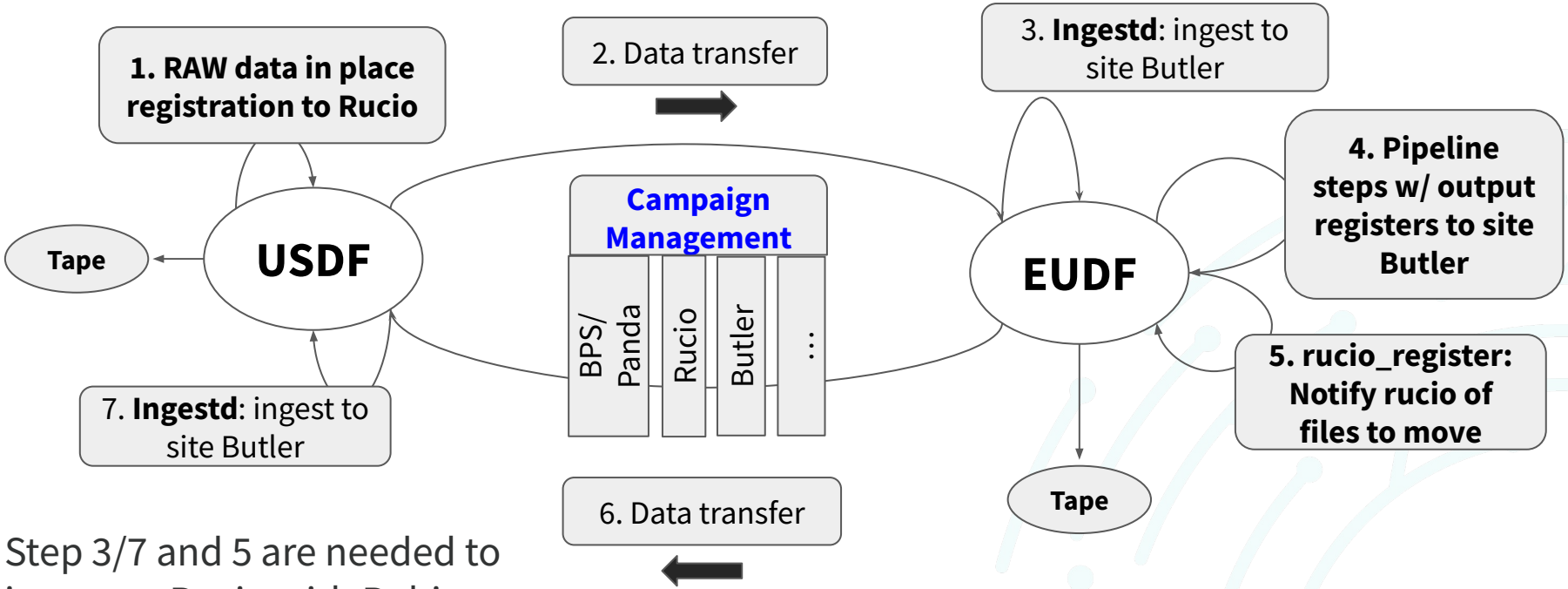
Campaign Management for Multi-site: status

- Campaign management for single facility is in regular production
- For multi-site, two features are needed:
  - Mechanism for running short, generic ‘connection’ housekeeping scripts at remote facility
  - Gluing together Rucio DM steps to single site processing system.
- Also need to enhance cross-facility job monitoring



Aiming for end of CY2024 to have fully functioning multi-site campaign management system.

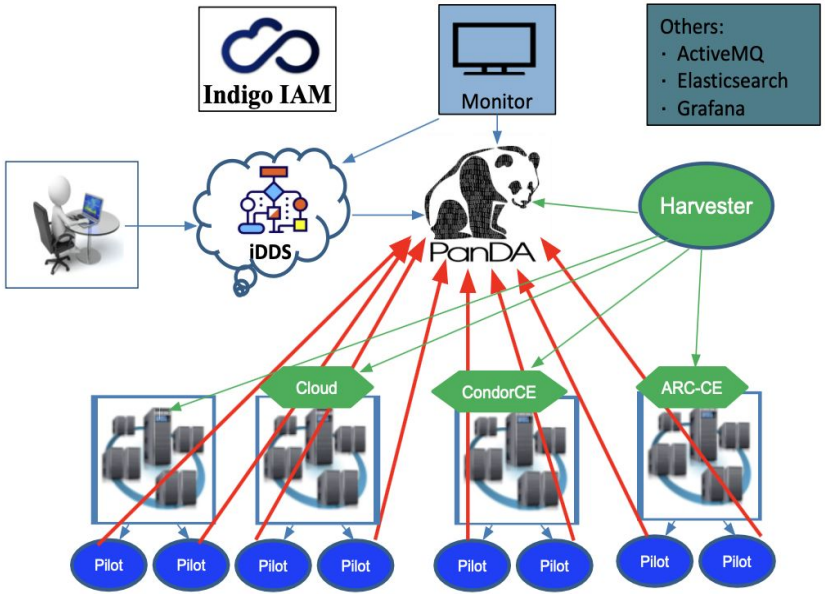
Multi-Site Processing Workflow



Step 3/7 and 5 are needed to integrate Rucio with Rubin native data access layer.

Multi-site processing includes data movement. See CHEP2024 presentation [Data Movement Model for the Vera C. Rubin Observatory]

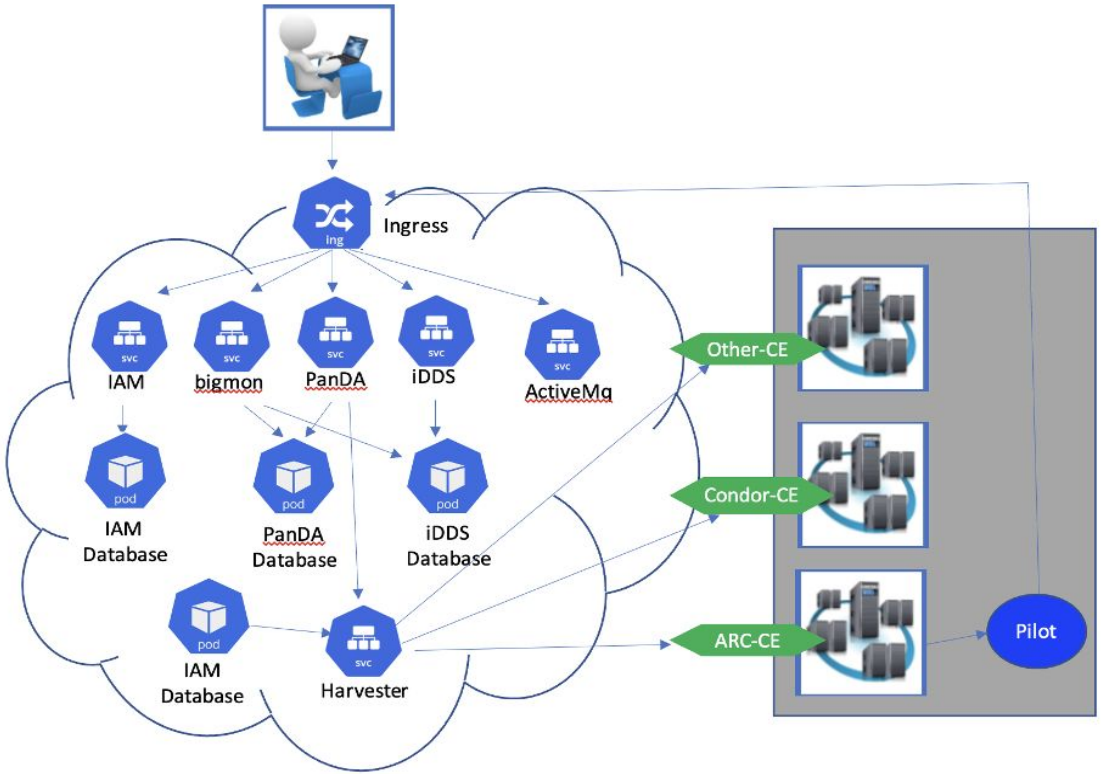
PanDA: the backbone of multi-site processing



- **PanDA**. PanDA is the workload manager. It manages/schedules tasks and jobs. It includes panda-server (job management), panda-JEDI (task management) and panda-database (PostgreSQL for Rubin).
- **iDDS**. iDDS is the workflow manager. It manages the dependencies of tasks and jobs. It includes the Restful service, the daemon agents and the database (PostgreSQL for Rubin).
- **Harvester**. It's the resource facing service to submit jobs to Grid/Cloud. It submittes jobs to CEs, such as Cloud, CondorCE, ARC-CE and others. It includes the Harvester service and a Mariadb.
- **Pilot**. The pilot runs as an agent at remote worker nodes to manage the user payload execution.
- **Monitors**. The main monitor is a PanDA monitor. It can also be integrated with Grafana, ElasticSearch.
- **Messaging**. The system employs a messaging service, for example ActiveMQ, to communicate between each other.
- **Indigo IAM**. The Indigo IAM is employed to manage OIDC based authentication and authorization operations.

PanDA: the backbone of multi-site processing

- PanDA system deployed on kubernetes at USDF
- The PanDA system employes Nordugrid's ARC-CE to send and monitor jobs at remote sites
- USDF, FrDF and UKDF are integrated for data processing



Logging and analytics platform

- Processing logs:
  - PanDA pilot (sophisticated jobs wrapper) log.
    - Exported through Harvester http service for latest logs, old logs are cleaned automatically
  - Pilot payload logs
    - Realtime logs from pipeline jobs: send to Google log service while the payload is running and rotation cleaning, looking to send to USDF Loki
    - Log files: send to Google Object store at the end of each payload, linked on PanDA monitor and kept for long term
  - PanDA system logs
    - Info from Panda system logs parsed by logstash, and sent to OpenSearch.
    - ATLAS developed extensive high level analytic dashboards. Rubin is interested.
- Data movement logs:
  - Rucio/FTS logs
    - Run in Kubernetes: All rucio daemons send logs to stdout. K8s captures them and sends to Loki
    - Looking into sending analytic info to OpenSearch,
  - Xrootd/dCache storage logs
    - Site level logs, USDF using Xrootd monitoring streams to build dashboards in Grafana
  - Rucio/Butler nexus logs
    - Under construction

Monitoring & Dashboard

- Site-level Monitoring and Dashboard
  - Mostly for site internal monitoring and investigation. Site-specific implementation.
- Multi-site subsystem level
  - We need to be able to drill down to
    - Data processing: Task progress, job status and logs
      - These are mostly available in Panda Bigmon
    - Data transfer: datasets transfer progress, site-level and file-level transfer status and logs
      - We are building these.
  - Currently investigating how to monitor Butler and Rucio interaction
- Campaign-level monitoring and dashboard
  - PanDA has years of experience to provide campaign-level dashboard for the ATLAS experiment. Ongoing work to leverage that for Rubin-specific needs
  - Will need dashboards for other areas. Under discussion.

Summary

- All distributed processing components are deployed and being regularly tested
  - We are ready to enter early production phase.
- We are using full chain data processing at multi-site to test them and drive forward
  - Will look into reliability/stability and scalability issues, and fix them
  - Will look into areas that we can optimize: already identified a few areas
- Still needed:
  - Building monitoring/dashboard for better understanding of processing campaigns.
  - Meet the challenge of storing enormous number of small files to tape
    - See: CHEP2024 presentation A Tape RSE for Extremely Large Data Collection Backups