Operations Boot Camp

Rubin Observatory

Tuesday 13 OCT 2020, 08:00 PDT

Introduction Part 2: Rubin Data Production and System Performance Activities

Session Webpage: Confluence

Slack Channel: #rubinops-bootcamp

In this talk

Rubin Observatory

PPP and DRP high level overview

System Performance in Operations

Rubin Observatory

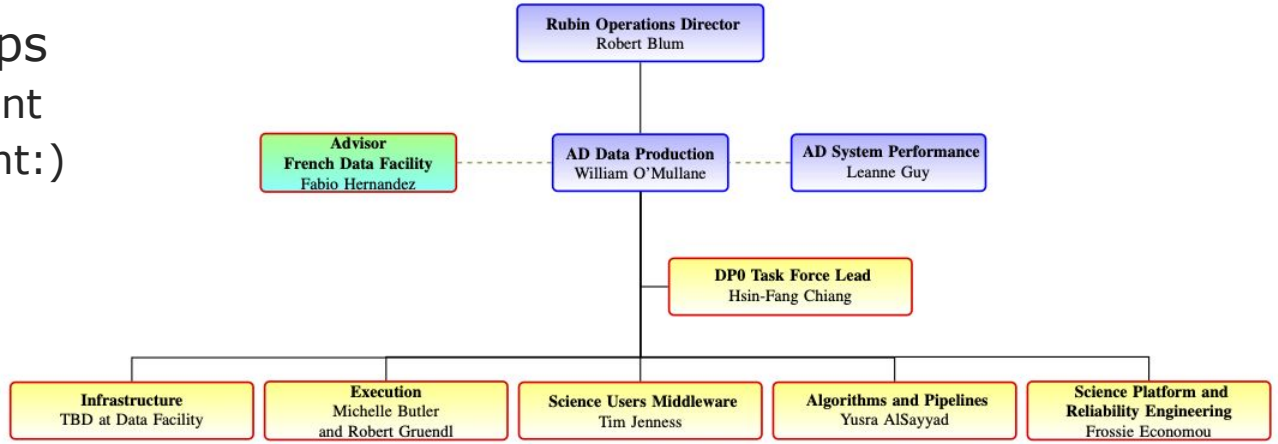
PPP and DRP high level overview

William O'Mullane and Yusra AlSayyad

Operations Organisation

Rubin Observatory

- 5 teams for ops
  - 6 if you count management:)
- Advisors from facilities
- Strong link to System performance



- This is not set in stone its the first attempt at an ops structure which is not just DM construction
- For the initial preops DP0 is treated as a special task force

Pipelines

Rubin Observatory

Previous slide reminds us of the teams in Data Production. This talk will concentrate more on pipelines.

- The Algorithms and Pipelines Team is responsible for developing and maintaining the scientific logic which generates data products.
- The team will evolve algorithms to take advantage of new hardware capabilities (in conjunction with the Infrastructure Team) and community needs (in conjunction with System Performance).
- Identifies and resolves errors in conjunction with the V&V team.
- DP will continue to follow the Dev Guide

Acronym Definitions Available Here

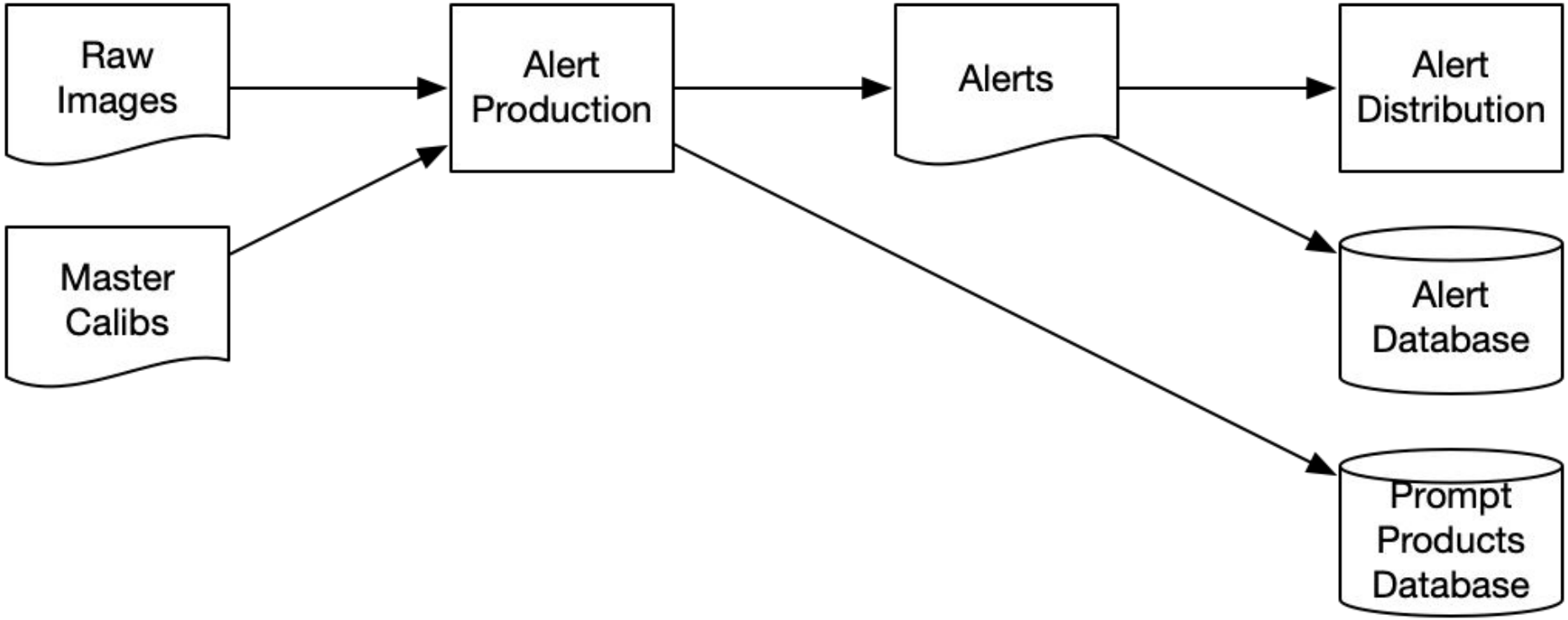
Transition to maintenance

Rubin Observatory

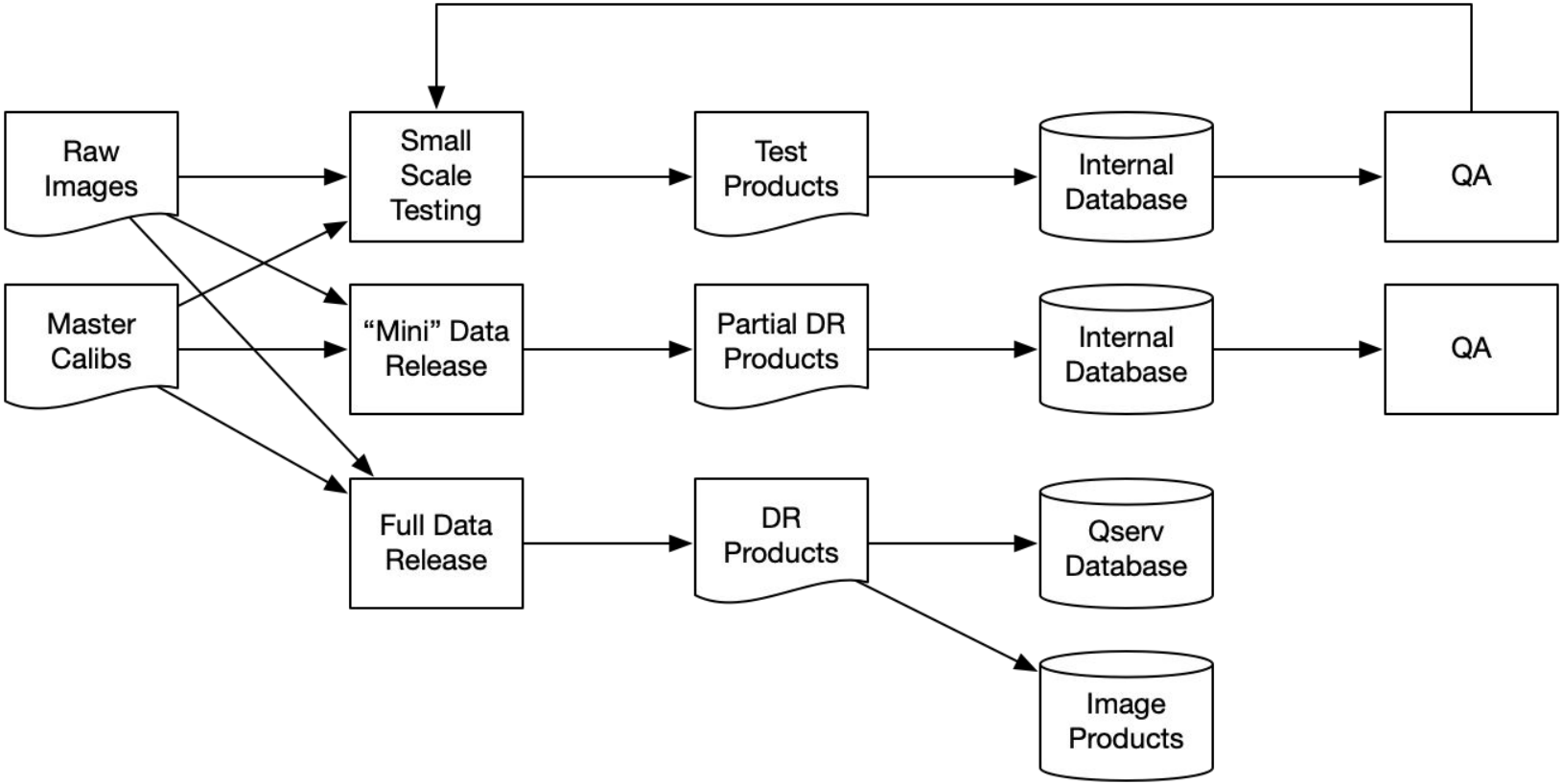
- Construction and Operations will develop on the same codebase
- When milestones are completed, the software component is then "in maintenance" - updates/fixes/improvements are operation activities
- DM software will start to be used in PreOps activities, e.g., Interim Data Facility (IDF) and Data Preview 0 (DP0)
- Work on any of these activities can be considered PreOps work.
  - Features needed specifically for DP0
  - Work done on the IDF
  - Getting ops team members up to speed (yourself)
- There is a PREOPS Jira project
  - For pipelines development, the separation is at the level of **Jira Epics** (i.e., DM stories on PreOps activities can go on PREOPS epics)
  - We may be audited

Data Product Lifecycle (Prompt)

Rubin Observatory



Data Product Lifecycle (Data Release)



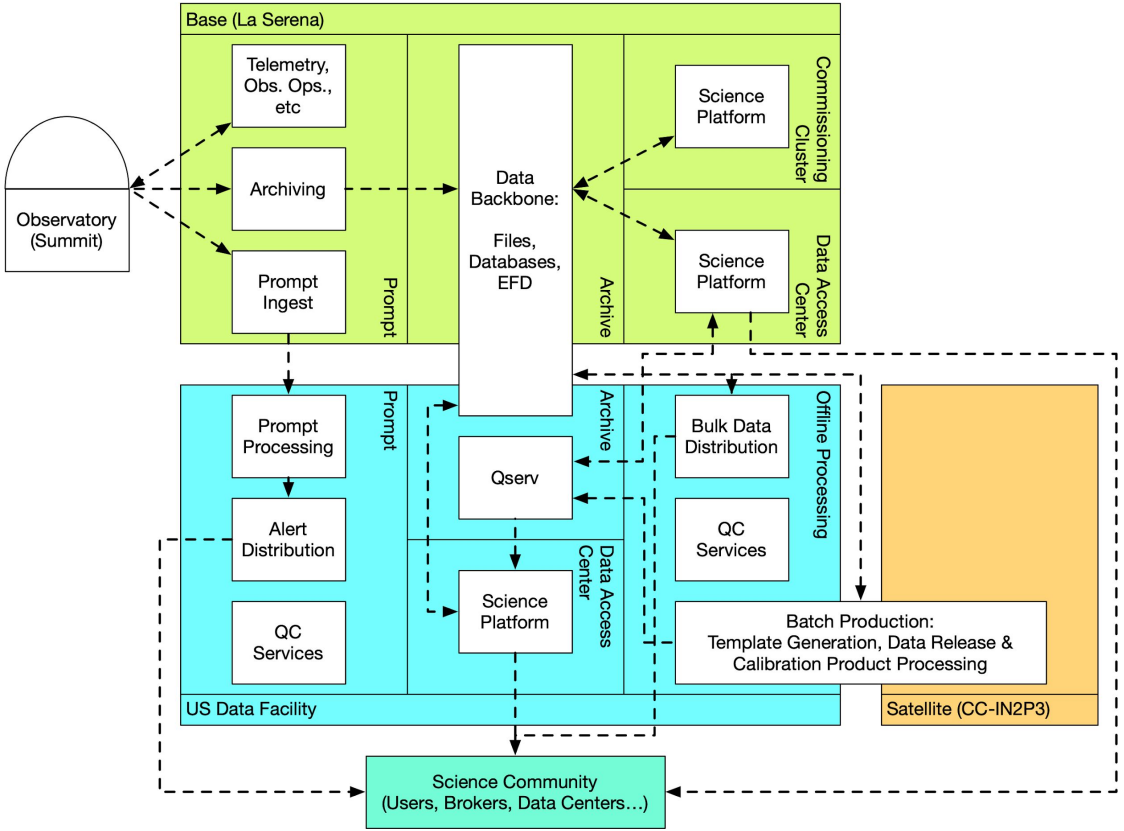
Data Product Lifecycle

- Data progresses from raw to released products and then to the permanent (offline) archive. There may be iterations in the generation of products as problems are discovered (e.g. via performance metric tracking, QA tools); as inputs, code, and configurations are refined during the transition to Operations; and during testing and even during annual Data Release production. A mini cycle is foreseen for DRP.
- We currently execute similar processes by performing monthly reprocessings of HSC RC2 and DESC DC2 data. Communication between the science and data production teams for these is via Jira tickets, but we are looking to improve these processes as we transition into Operations. Any Rubin staff member can look at and analyze the reprocessing results.
- In pre-operations there are planned activities to make sure we have captured and understood procedures. We plan to practice the releases and processing cycles. The releases are called Data Previews and we intend to open them to the science collaborations. The initial details of planning Data Preview 0 is given in RTN-001 which includes serving data to the community and a DRP cycle on simulated data in 2021.

Data Product Provenance

- The Prompt Products and Data Release databases will contain metadata about data products, particularly Processed Visit Images (PVIs), including QA metrics, which can be combined with information from the Engineering and Facility Database (EFD). The information for these tables comes from our single-frame processing pipelines. We will restart loading these databases as we finish integration with the Science Data Model. This information will be used to help determine what reprocessing may be necessary and to select (or de-select) inputs to reprocessings.
- Pipeline inputs, code versions, and configuration are captured by the "Gen3" middleware, enabling us to know how any given data product was generated.
- All released data products and their provenance will be ingested into the Data Backbone as part of the archive of the survey. The provenance will be available through the Gen3 Butler Registry as a database with a Python API.

Architectural Overview from Construction

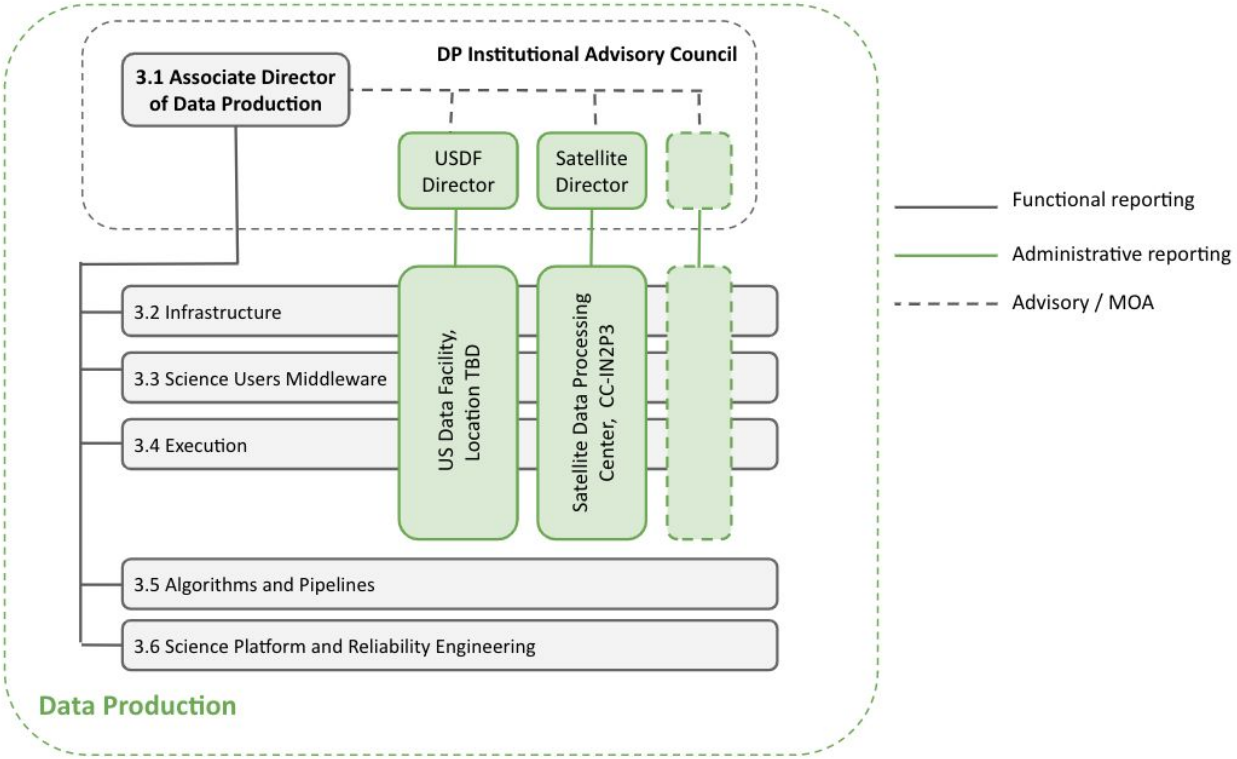


Satellite Data Facilities

Rubin Observatory

Multiple data facilities, including the USDF, will form constituent parts of one integrated Data Production Department.

- Satellite Data Processing Center at CC-IN2P3 will perform 50% of annual Data Release Processing.
- We are anticipating a UK DF taking on up to 25% of DRP



CC-IN2P3 French Data Facility

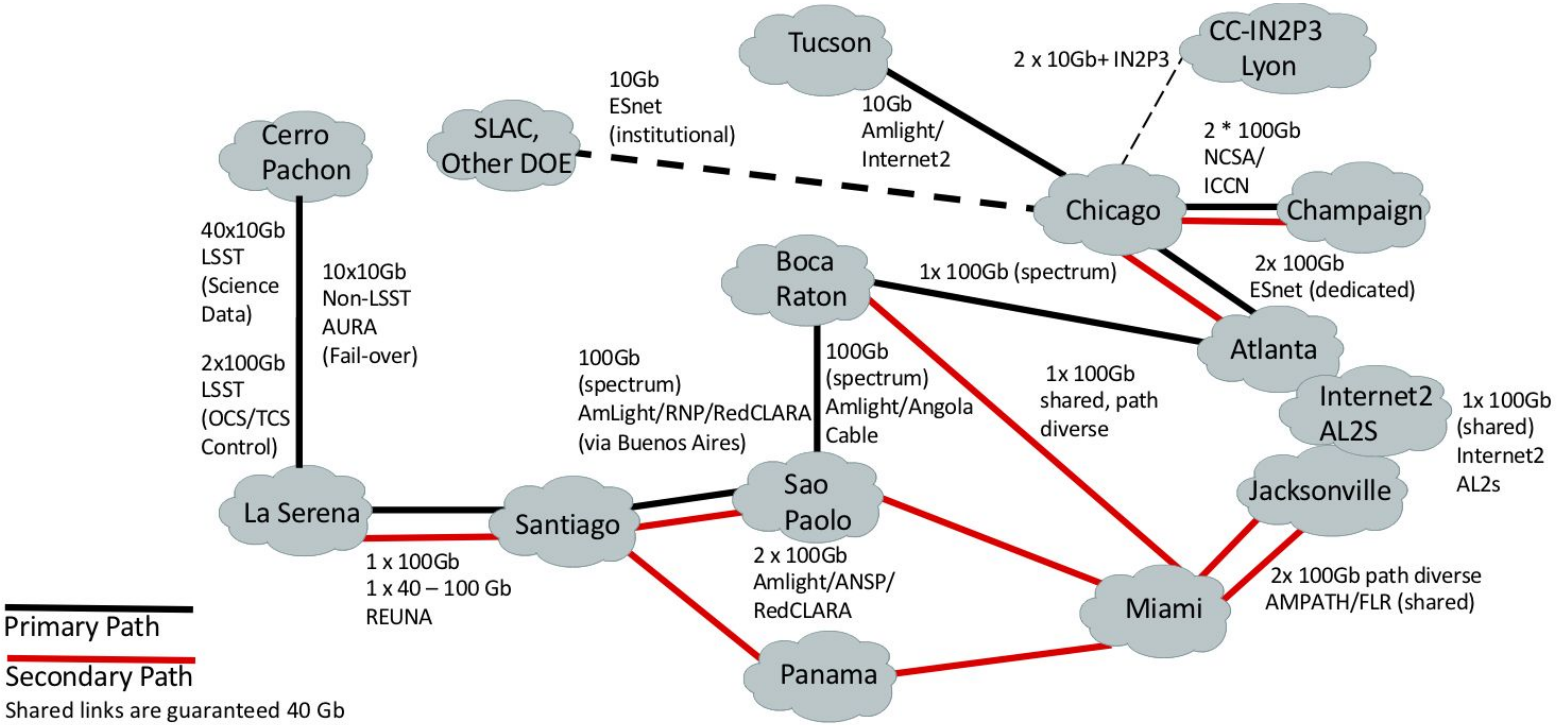
Rubin Observatory

Activities:

- Import copy of raw data (images, calibration, etc.) from USDF and store it in permanent storage (tape);
- Annually, locally process 50% of the raw data collected since the beginning of the survey, to produce science-ready images and data for populating the catalog(DRP)
- Export locally-produced data products to USDF and import theirs;
- Store a copy of released data products on permanent storage.

CC-IN2P3 is a shared data processing facility contributing to several international projects, e.g. LHC, VIRGO-LIGO, EUCLID, CTA, AUGER, HESS, etc.

Rubin Obs Long Haul Network Links (FY22)

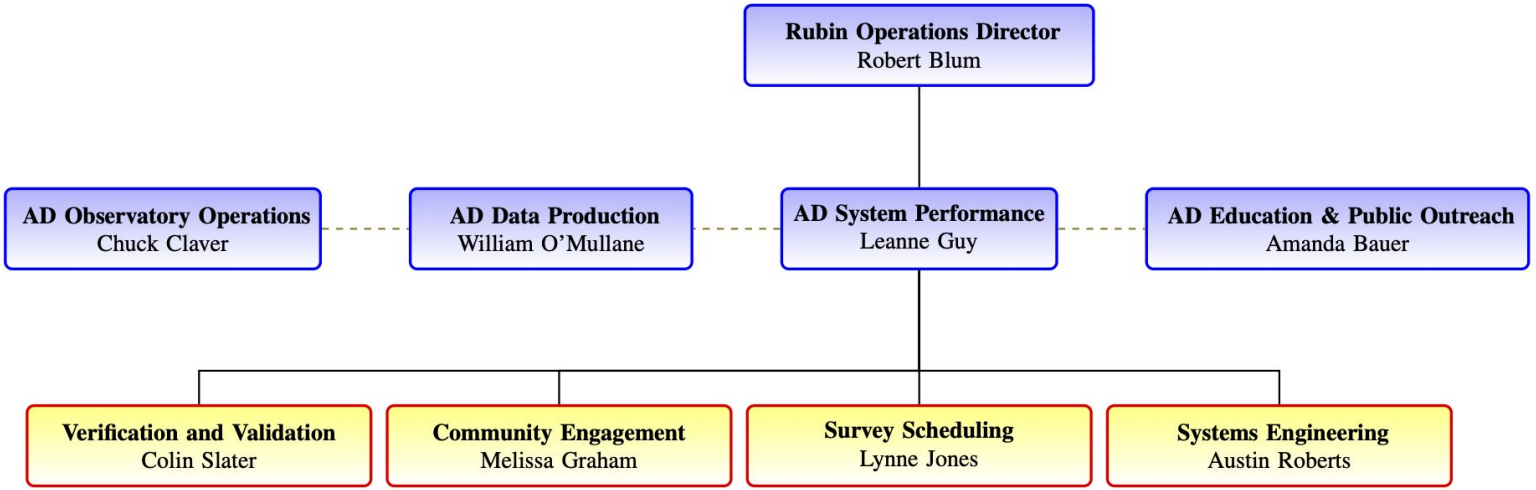


Rubin Observatory

System Performance Activities

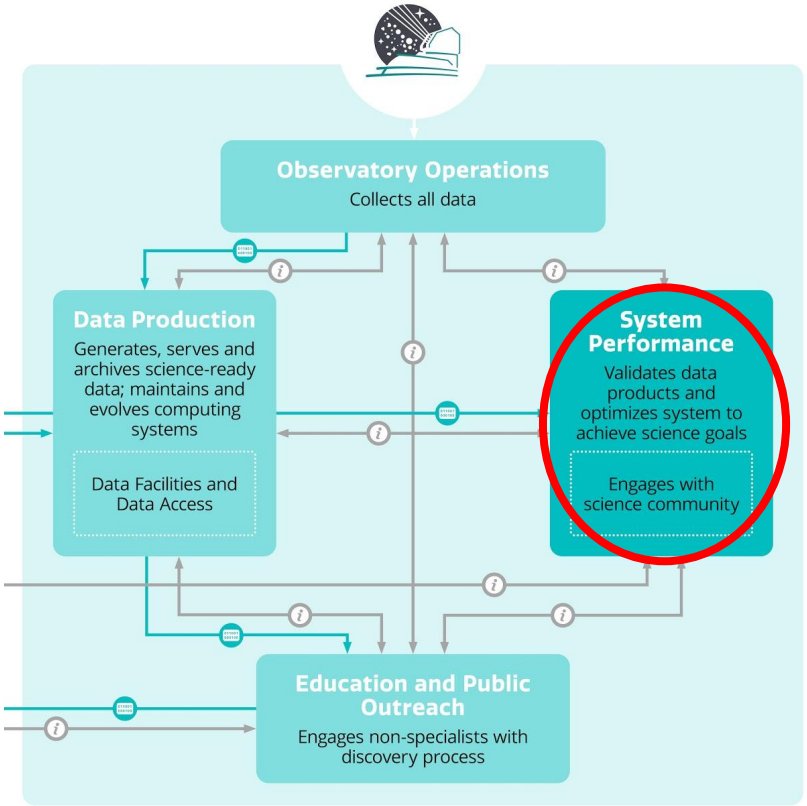
Melissa Graham, Colin Slater, Leanne Guy

System Performance in Operations



- DP0:
  - Verification & Validation and Community Engagement teams only
  - Systems Engineering will be following the processes
  - Working very closely with Data Production department

Verification and Validation Team



- Goal of System Performance: Validate data products and optimize the system to achieve science goals.
- V&V works towards this goal by developing a thorough understanding and characterization of the data.

Verification and Validation Team

- Continually assess the scientific performance of the observatory
  - In comparison to formal requirements, model expectations, and past experience
- Discover, characterize, and diagnose undesirable data quality issues
  - Could have hardware, software, or operational process origins; our job is to track it down
  - Emphasis on systematic effects that might not be apparent in aggregate performance metrics
- Focus is on both raw images and processed data products

Verification and Validation Team

- Scope is broad, team must be interdisciplinary
  - Early in the diagnosis it's hard to tell where an effect is coming from
  - Draw on experience from different construction subsystems and from outside of construction.
  - Use and expand on construction-era tooling for performance tracking and data quality assessment.
- Coordinate with and learn from SITCOM during commissioning and early operations.

V&V Interfaces

- Primary deliverable is feedback to the other departments on what we see, and how we can improve the observatory
  - Propose process changes to Systems Engineering, for implementation by Observatory Operations
  - Algorithmic improvements to Data Production
  - Knowledge of known effects to Community Engagement, to better handle inquiries.

V&V Collaborations

- V&V serves as a collaborative resource
  - During data releases -- key time for V&V collaboration with Data Production
  - Understanding planned process changes with Observatory Operations
  - Work with Community Engagement to learn from our users, both for uncovering new data quality issues and to draw on community expertise in testing and solving problems.
  - Collaborate with Survey Scheduling and Systems Engineering in tracking observatory performance and survey progress.

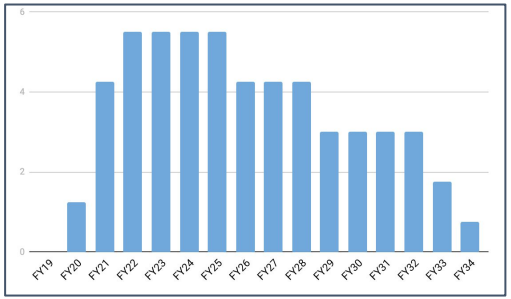
Community Engagement Team (CET)

Rubin Observatory

Maximize scientific results from Rubin by engaging the community.

In Operations, the CET will:

- promote inclusive and equitable engagement in LSST science
  - diversity in representation in user committees
  - documentation and tutorials that target all experience levels
- facilitate access to and analysis of data products and services
  - enable the community to crowd-source solutions
  - curate resources (documentation, tutorials, online forums)
  - interact with scientists (workshops, Community.lsst.org)
- coordinate expertise within the community and the project
  - help to guide issue resolution
  - inform RSP developments
- support diverse research methodology such as citizen science
  - prepare datasets for the EPO Data Centre



The plan for CET staffing is still in development.

At least 7 FTE with a quick ramp-up and a slower ramp-down over 10 years.

All staff will be scientists with expertise across the four Rubin science pillars.

CET's Transition to Operations

Six CET members drawn from construction-era staff at NOIR and DOE labs have begun preparatory and transitional work with fractional FTE assignments.

The CET's current activities include:

- developing a comprehensive model for community engagement
- building CE use-cases for model-based systems engineering analysis
- running an discussion series for community input on engagement initiatives

- studying documentation and tutorials from Rubin and other facilities
- investigating options for online courses, discussion forum enhancements

- assisting with, e.g., the Stack Club, the Project & Community Workshop
- participating in user support via Community.lsst.org
- helping to stand up the Users Committee

- **preparing to engage the community in the Data Previews**

Community Engagement in DP0

Rubin Observatory

Use terminology to set expectations that will deliver success on DP0's goals.

The goals are: (1) to test the DMS/RSP and inform further development, and
(2) to prepare the community to use data products and services.

These goals rely on community engagement in DP0.

The ~300 community participants, "**DP0 Delegates**", will have the important role of *representing* the community as *learners* and as *contributors*. This role comes with expectations and responsibilities as well as the benefit of early RSP access.

"**Onboarding resources**" to the IDF and the RSP must be created to ensure that delegates learn a sufficient amount to enable their contributions.

"**Delegate contributions**" might include creating and sharing materials with the community, or testing aspects and providing feedback requested by Rubin staff.

On Thursday at 10:30am the CET will solicit your ideas for **onboarding resources** and **delegate contributions** that would help you achieve your DP0 goals.

An RTN of "Guidelines for Community Participation in DP0" is in prep.

CET Interfaces to Other Groups

Rubin Observatory

During Operations, the CET will interface with many Rubin groups.

Much discussion between all groups will be held openly on Community.lsst.org.

Here are a few specific groups and how else the CET might interface with them.

- **SP - SE, SS, SV:** regular SP leadership team meetings, collaborative work
- **OO, DP, SQuaRE:** topical interactions to solve problems; co-work on Jira tickets
- **RSP Devs/Users:** facilitate Users Committee (twice-yearly reports)
- **EPO & Comms:** regular collaboration re. citizen science, website, workshops etc.
- **Science Collaborations:** formalize Rubin liaison program; build relationship with chairs
- **In-Kind Contributors:** assist with the ingestion of deliverables into the Rubin system
- **NOIRLab:** maintain contact; share resources for common engagement goals
- **LSSTC:** maintain contact; shared goal of enhancing scientists' ability to secure funding
- **science community:** [*the many interfaces as previously discussed*]