Survey Astronomy in the Big Data Era with LSST

Timothy Noble, STFC

GridPP / Swift-HEP

Thursday 30th March 2023

Acknowledgement: Rubin Obs/NSF/AURA

Vera C. Rubin Observatory

- New observatory being constructed in northern Chile (Cerro Pachón)

- High altitude and very dry climate
  - best site for (ground-based) optical astronomy

- Hosts 8.4m Simonyi optical telescope
  - 3-mirror configuration gives extremely wide field of view
 - Survey whole (Southern) sky in 3 nights

Acknowledgement: Rubin Obs/NSF/AURA

Legacy Survey of Space and Time (LSST)

- 10-year survey, starting in 2024 (*)

- Imaging 30,000 sq. deg. of sky (in three nights)

- Observes each patch 800+ times during survey

- Stacked images help identify very faint objects
  - 24Bn galaxies, 14Bn stars plus solar-system objects

- Frequent, regular visits allow time-series for dynamic objects
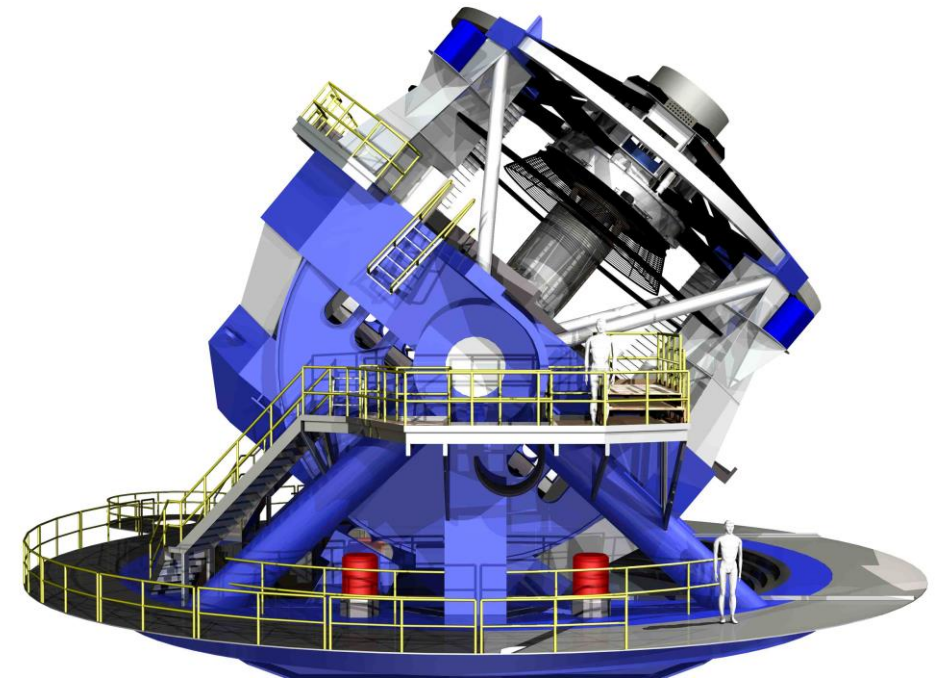 - Near-earth Objects
  - Supernovae
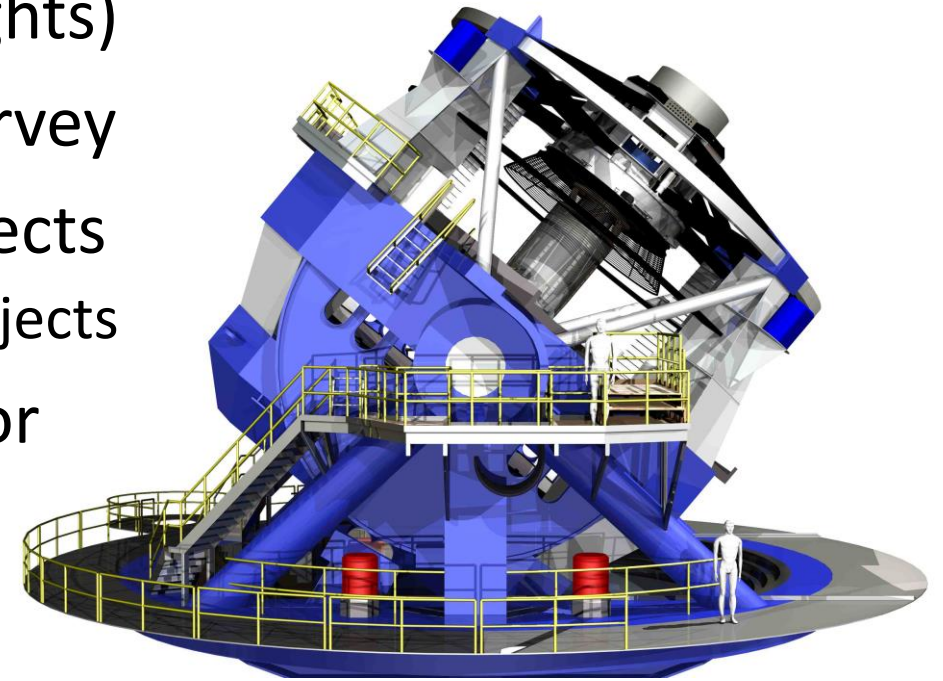
(*) Construction delayed by Covid-19

Primary Science Drivers

- Understanding Dark Energy and Dark Matter

- Creating inventory of solar system

- Cataloguing transient sky

- Mapping Milky Way

However, LSST data expected to have huge impact on almost all areas of astronomy

Data Products

- ## Annual LSST Data Releases (*)
  - Processed visit images, co-adds, and object catalogues
  - Published annually (two releases in first year)
  - Served to community through Data Access Centres

- ## Nightly Alert Stream/ Prompt Products
  - Near-real-time issue of detected transient activity
  - Expect ~10M alerts per night (though not all interesting)
  - Streamed to small number of Community Broker sites

- ## User-generated Products (*)
  - Downstream enhancements from key communities/ groups

(*) Data is proprietary for two years, only for designated data rights holders

LSST:UK Consortium

- Representation from every (that is, 36) UK astronomy groups

LSST:UK Consortium

- Remit
  - Responsible for securing and managing funding for UK Programme – called *Science Centre Programme*
  - Delegated authority for administration of Rubin data-rights in UK
- Roles
  - Project Leader – Bob Mann (Edinburgh)
  - Project Scientist – Stephen Smartt (QUB)
  - Commissioning Coordinator – Graham Smith (Birmingham)
  - Project Managers – George Beckett and Terry Sloan (Edinburgh)
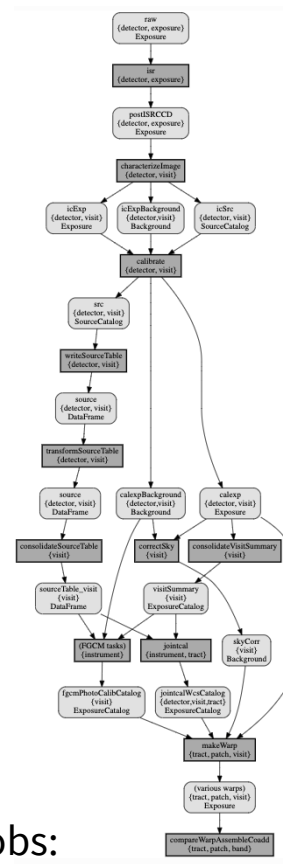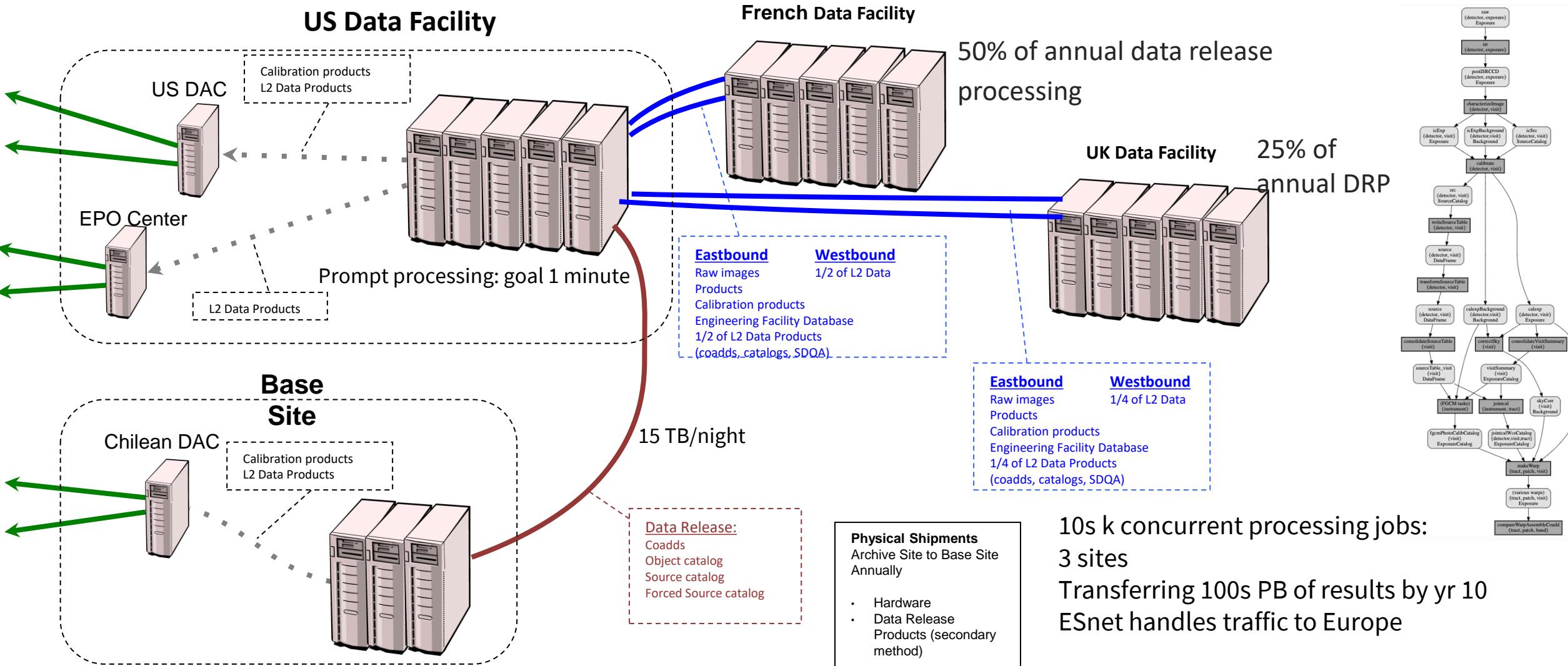 - Consortium Board Chair – Mike Watson (Leicester)

UK In-kind Contribution Make-up

- Activity groups
  - DRP – Contribute to Data Release Processing
    - to process 25% of survey data (alongside CC-IN2P3, France and SLAC, USA)
  - DAC – Design, deploy, and operate (full) Independent Data Access Centre
    - Serving 20% of international Rubin community (1,500 data-rights holders)
  - DEV – Develop software and services for UK priority science areas
    - Creation of User-generated Products
    - (Operation of Community Broker to receive and process nightly alert stream)
 - Plus contributions to Commissioning, International Coordination and EPO

Data Flows

US Data Facility

US DAC

Calibration products
L2 Data Products

EPO Center

Prompt processing: goal 1 minute

L2 Data Products

French Data Facility

50% of annual data release processing

UK Data Facility

25% of annual DRP

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

Westbound
1/2 of L2 Data

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

Westbound
1/4 of L2 Data

Base Site

Chilean DAC

Calibration products
L2 Data Products

15 TB/night

Data Release:
Coadds
Object catalog
Source catalog
Forced Source catalog

Physical Shipments
Archive Site to Base Site Annually

- Hardware
- Data Release Products (secondary method)

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

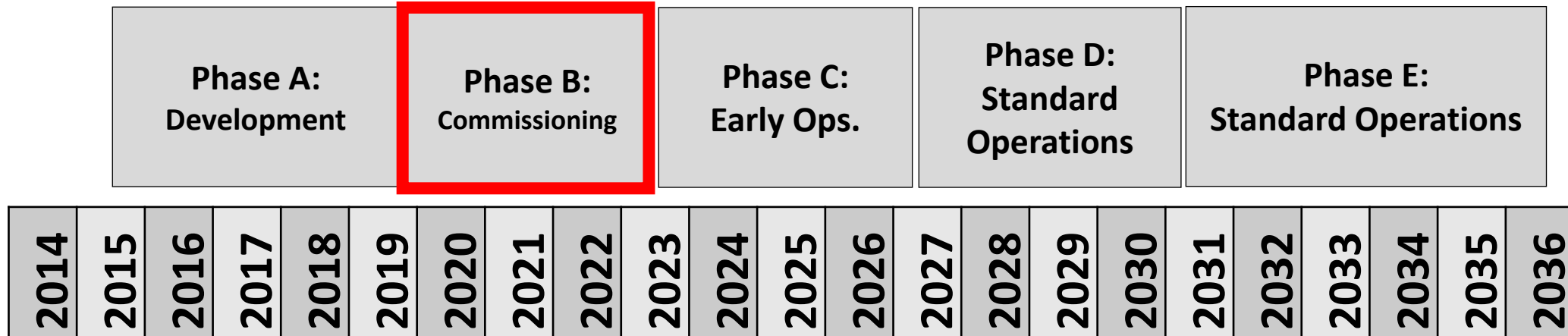# Data Release Processing (DRP)

- Work shared across three facilities:
  - US Data Facility – SLAC
  - French Data Facility – CC-IN2P3
  - UK Data Facility – IRIS
- Looking to HEP technologies for distributed processing
  - Rucio – distributed data management
  - PanDA – workload distribution/ management
  - FTS – data transfer
- Integrated with LSST Stack technologies
  - LSST Science Pipeline – processing
  - Data Butler – data/ metadata curation
  - Qserv – catalogue database
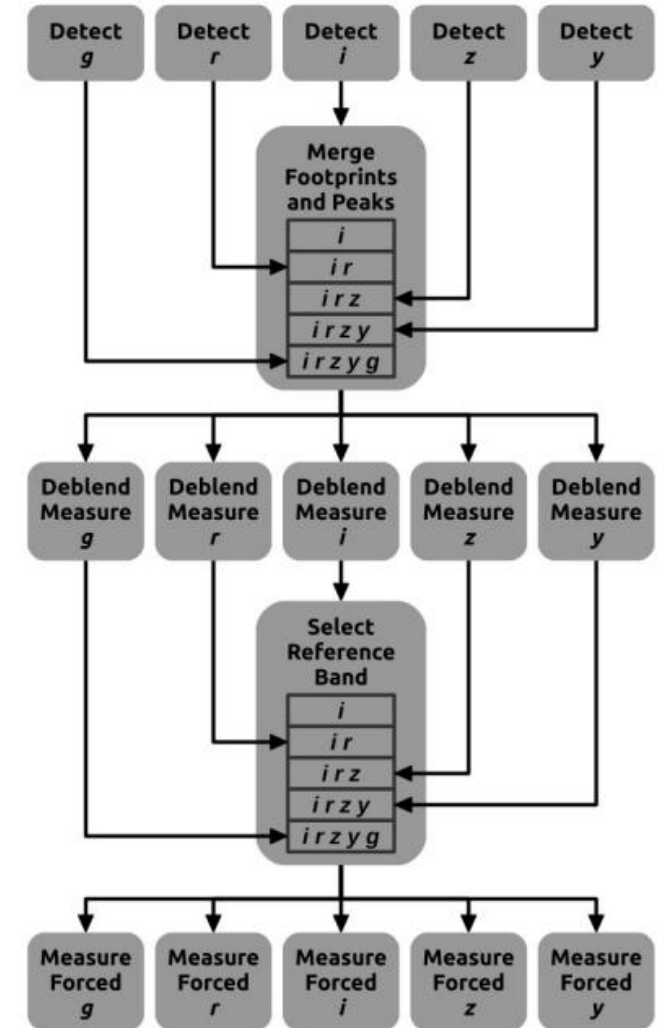
Preparing for DRP (SLAC, France, UK)

- DP0.2 done, using PanDA
  - Redone at CC-IN2P3, though with different workflow tool (Parsl)
- Routine HSC reprocessing in progress at the USDF for a couple of months now
  - Engagement of Campaign Mgmt, Pipelines and Infrastructure groups
  - Pipelines group proposing a full HSC PDR3 multi-site reprocessing by July
- Rucio and PanDA servers installed at SLAC; in final testing
- Multi-site testing/scaling work is underway, via increasingly complex stages
  - Start with job submission to central PanDA server from each site, to be executed at those sites (done)
  - Expand to central submission
  - Use Rucio/FTS to move input/output files among sites
  - Test scaling up numbers of processes and volume of data

LSST:UK Science Centre Programme

Phase A: Development	Phase B: Commissioning	Phase C: Early Ops.	Phase D: Standard Operations	Phase E: Standard Operations

2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036

- Five-phase programme
  - Phase A (Jul'15—Jun'19)   –   Development
  - Phase B (Jul'19—Mar'23)   –   Commissioning
  - Phase C (Apr'23—Mar'27)   –   Early Operations
  - Phase D (Apr'27—Mar'31*)   –   Standard Operations
  - Phase E (Apr'31--~2036*)   –   Standard Operations (cont'd)
- Forecast budget of £56M (including capital for DAC and DRP)
  - Infrastructure element funded through IRIS Programme
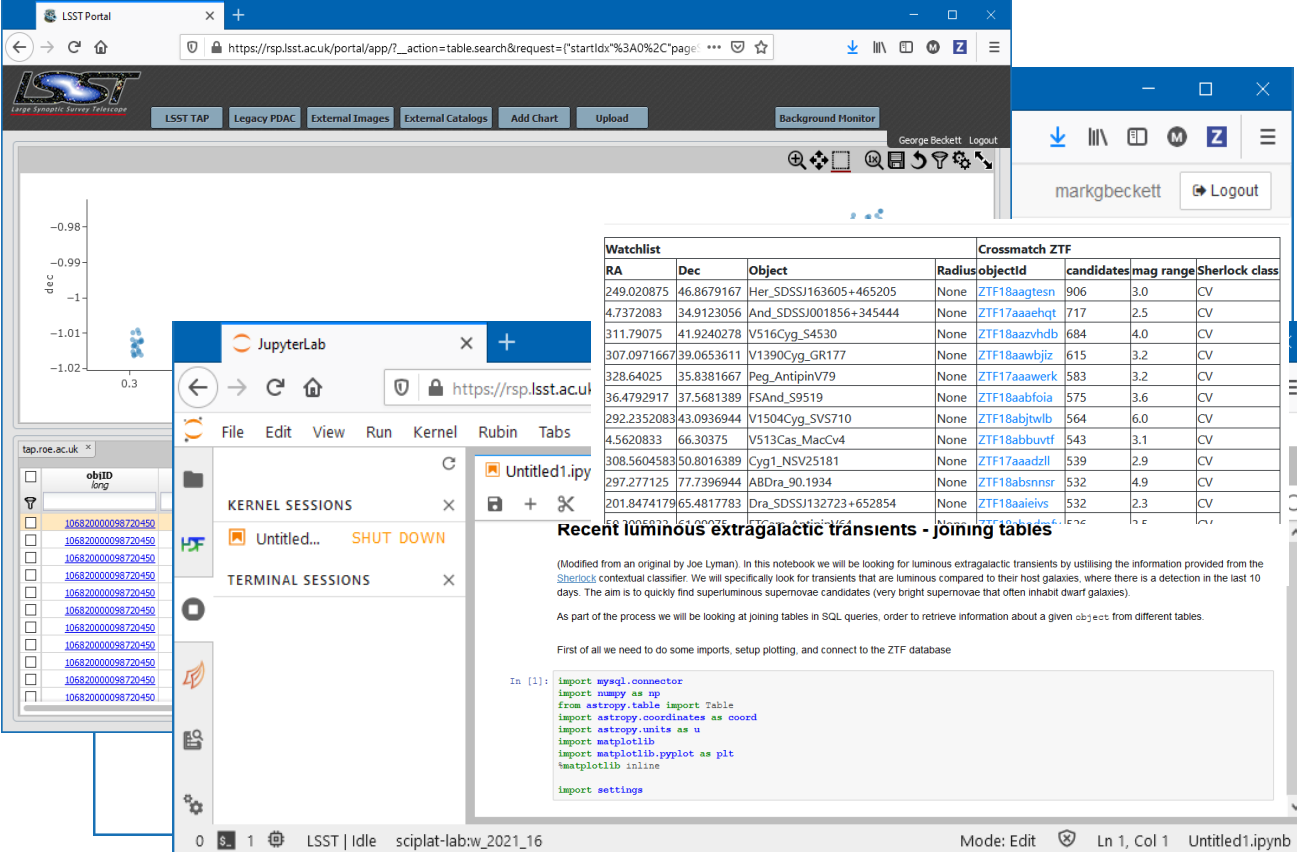
STFC IRIS Programme

- Infrastructure provision for large-scale UK PPAN facilities
  - LSST, LHC, Dune, SKA, Euclid, Diamond Light Source, …
  - LSST:UK founding partner of IRIS Programme
- Distributed provision of cloud, grid, HPC, storage (+ supporting infra)
- LSST:UK developing long-term sizing model (2022—2036)
  - Cloud, HPC and storage for DAC services
  - HPC and storage for DEV
  - Grid and storage for DRP
  - (Networking bandwidth for all aspects of operation)

UK Data Access Centre

- Proto-DAC
  - Based on Rubin Science Platform
- Analysis Platform for LSST Data
  - Bring compute to data
  - Catalogue and image services
  - Portal, notebooks (and batch) UI
  - Containerised (Kubernetes)
- Running on IRIS cloud
  - Hosting early datasets (see later)
  - Deployed on Scientific OpenStack
  - Running at Edinburgh (and RAL)

Rubin Science Platform/ Lasair:ZTF

Running on IRIS Cloud in Edinburgh
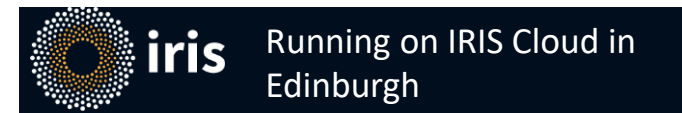
Community Broker for astronomical transients

- UK Community Broker confirmed for operations (one of seven)
- Rubin Alerts with enhancements
  - Sherlock (sky context), known supernovae (TNS), user-made watchlists/ areas
- SQL access to alerts, both on-demand and real-time streaming
- Users can "annotate" the data objects with classifications
  - Zooniverse, Fastfinder, Fink, Alerce
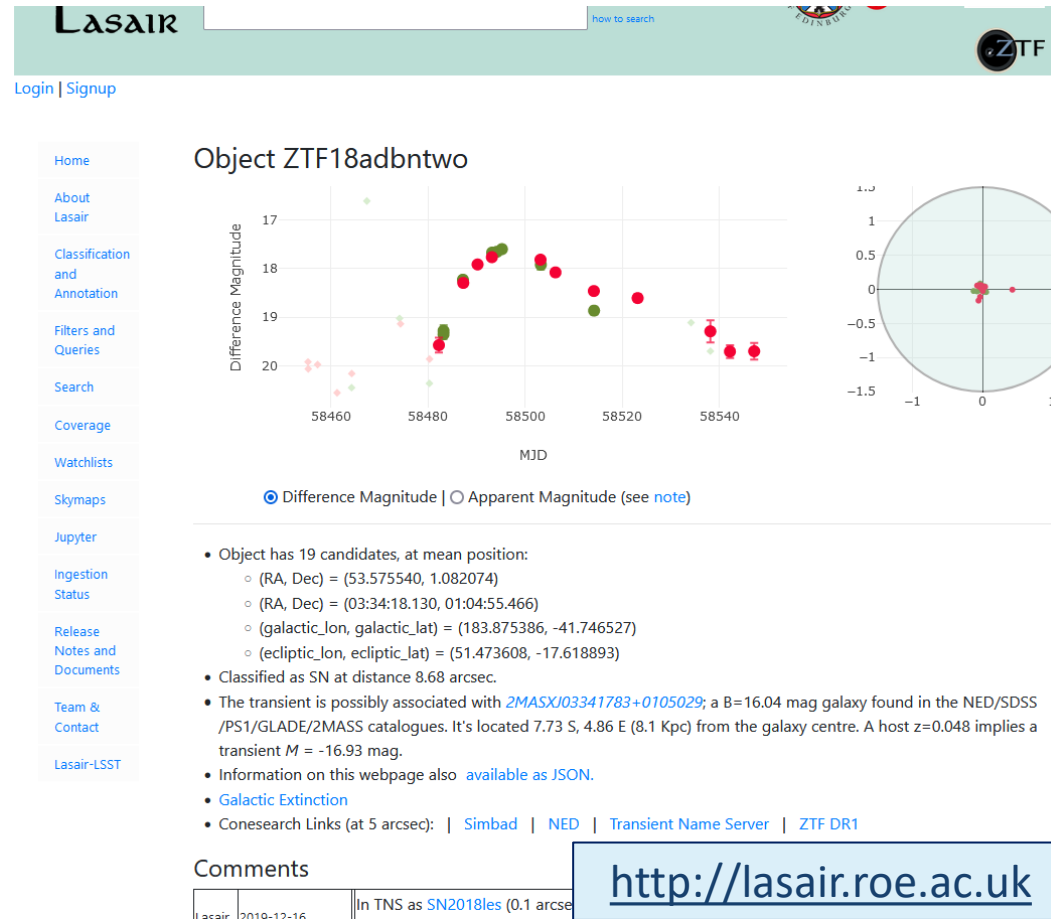- Machine-to-machine filtered output streams via public Kafka service

Running on IRIS Cloud in RAL and Edinburgh

University of Edinburgh and Queen's University Belfast

Lasair:ZTF—Precursor for Rubin

http://lasair.roe.ac.uk

- Running on IRIS cloud (Edinburgh and RAL)
- Ingest nightly alerts from ZTF
  - 400,000 alerts per night (20kb each)
  - Filter, annotate, and follow-up
- Publish via web interface
  - Standard queries
  - Cross-match to other surveys (5TB catalogues)
  - Personalised watch lists/ alerts
- … and notebook service
  - Flexible scripted analysis
  - Astronomy/ Rubin software and Lasair API
- Collaborations with
  - Zooniverse, Fink & Ampel, 4MOST/Tides

UK Contributions

UK DAC Data movement

- Pre-LSST data generated and processed at Cambridge (CSD3)
  - Need to move data from CSD3 to RAL for long-term storage and access
- Challenge – Move data from Cambridge to RAL and register metadata in Data Butler
  - Rucio to be used to manage the data movement
  - Then register the files and their metadata to Data Butler for user lookup
  - Data should then be accessible for Users in Edinburgh and highly used data moved to Edinburgh to reduce network traffic
- To be carried out by Mathew Sims at RAL

Data Transfer Monitoring

- Begun setting up monitoring pages to help all data facilities understand data flows and troubleshoot problems

- More infrastructure is needed for further information, desired monitoring like Edinburgh

- Use of RAL production FTS has highlighted monitoring hole, and to be worked on
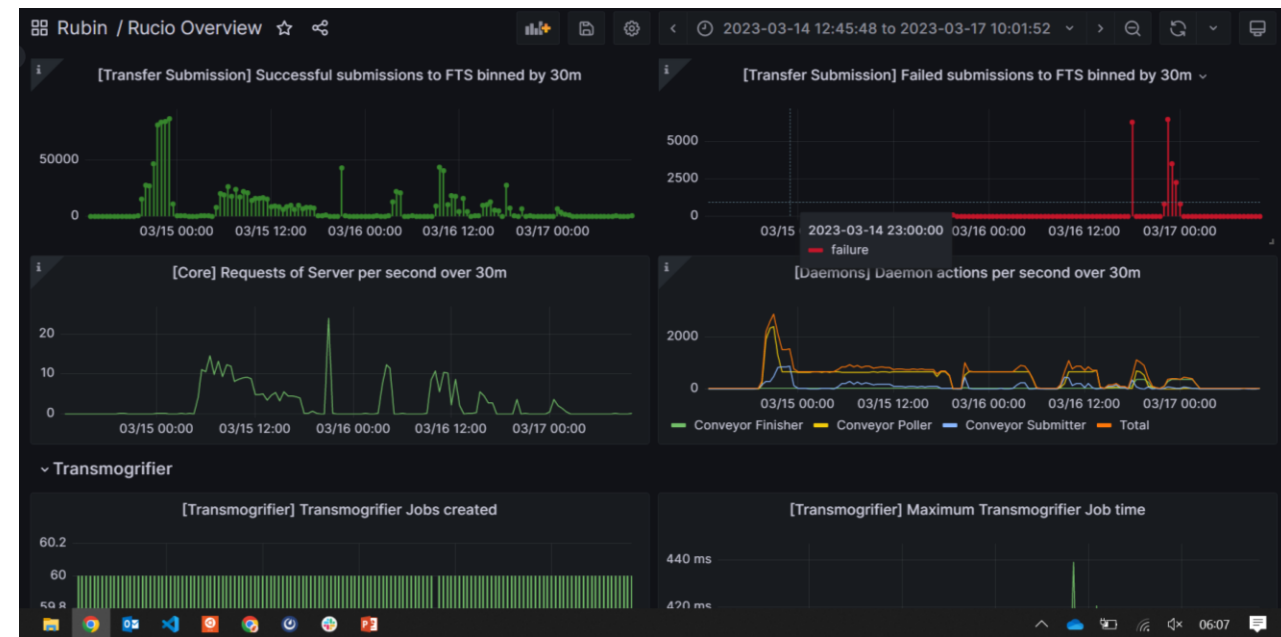
LSST Requirements for future

Sizing Model (High-level Summary)

DRP

	Preops	Survey Operations (based on commencement of survey in FY24)										Post-ops	
Capability	FY23	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10	FY34	FY35
CPU (M core hrs)	11	11	21	30	40	50	63	73	83	93	103	100	50
Normal/ Object (PB)	8.0	9.0	16.0	23.0	30.0	37.0	44.0	51.0	58.0	65.0	72.0	61.4	0

DAC

	Preops	Survey Operations (based on commencement of survey in FY24)										Post-ops	
Capability	FY23	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10	FY34	FY35
CPU (M core hrs)	0.53	0.88	2.10	2.45	2.63	3.94	5.26	5.26	6.57	7.88	7.88	7.88	7.9
Normal/ Object (PB)	2.2	21.9	50.0	72.1	94.9	117.6	140.2	162.9	185.6	208.3	231.0	231.0	231.0

Sizing Model (cont.)

Lasair

	Preops	Survey Operations (based on commencement of survey in FY24)										Post-ops	
Capability	FY23	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10	FY34	FY35
CPU (M core hrs)	0	4	4	4	4	4	4	4	4	4	4	4	4
Normal/ Object (PB)	0.0	0.3	0.5	0.8	1.1	1.3	1.6	1.8	2.1	2.4	2.6	2.6	2.6
Fast/ NVMe (TB)	115	136	156	177	197	218	238	259	279	300	300	300	136

- User compute/ storage is part of RSP

DEV

- ??? – though less than DRP

Summary

- UK astronomers and sites working together to maximise opportunity of Vera C. Rubin Observatory

  - In-kind programme means more involvement/ influence in Operations

- Ambitious plans

  - Operate UK-based Data Access Centre (and Community Broker)
  - Extend LSST science potential through user-generated products and integration with ancillary surveys
  - Undertake contribution to Data Release Processing

- Significant requirements for diverse computing infrastructure over next 10—15 years

Thank you