Centre de Calcul
de l'Institut National de Physique Nucléaire
et de Physique des Particules

Updates on CC-IN2P3 Computing Infrastructure for LSST

fabio hernandez, quentin le boulc'h, gabriele mainetti

Journées Rubin LSST-France, Lyon, December 14th, 2023

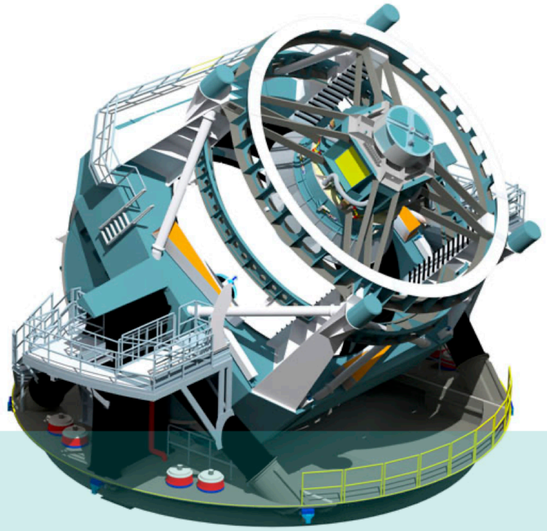
doc.lsst.eu

Big Picture

LSST DATA

Raw Data: 20TB/night

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



Prompt Data Products

- Alerts incl. science, template and difference image cutouts
- Catalogs of detections incl. difference images, transient, variable & solar system sources
- Raw & processed visit images (PVIs), difference images

Data Release Data Products

Final 10yr Data Release:

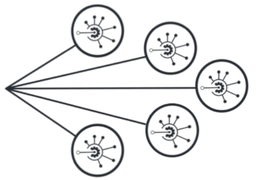
- Images: 5.5 million x 3.2 Gpixels
- Catalog: 15PB, 37 billion objects

via Alert Streams

via Prompt Products

via Image Services

via Data Releases



Community Brokers

Rubin Data Access Centres (DACs)

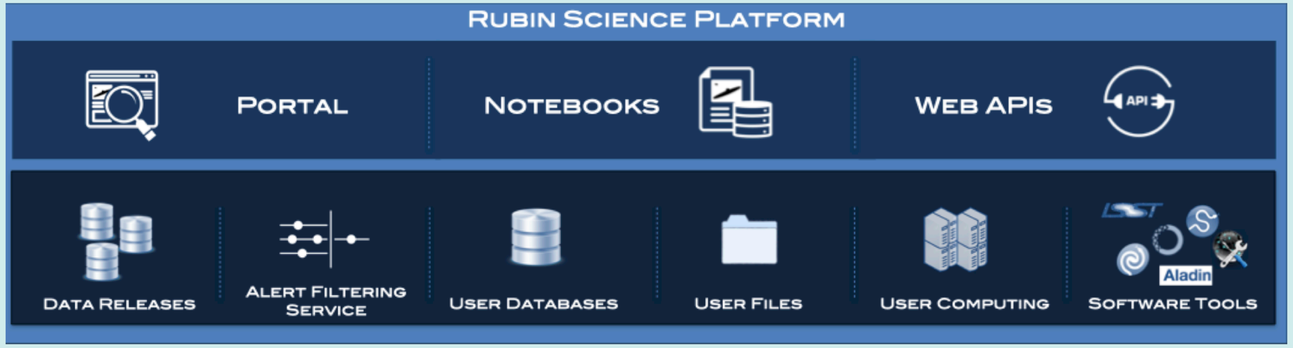
USA (USDF)
Chile (CLDF)
France (FRDF)
United Kingdom (UKDF)

Independent Data Access Centers (IDACs)

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

Rubin Science Platform

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

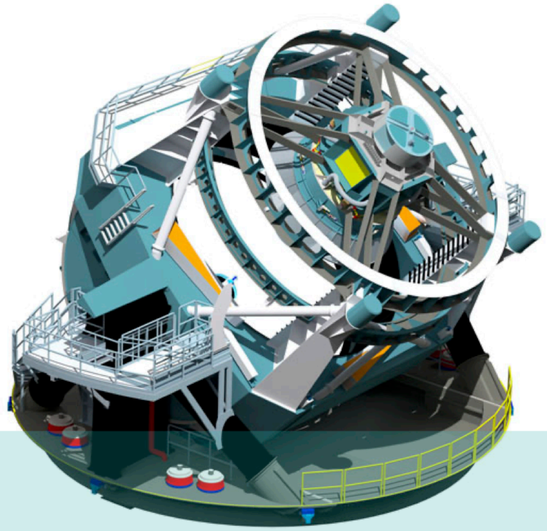


Credit: Leanne Guy

LSST DATA

Raw Data: 20TB/night

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



Prompt Data Products

- Alerts incl. science, template and difference image cutouts
- Catalogs of detections incl. difference images, transient, variable & solar system sources
- Raw & processed visit images (PVIs), difference images

60s via Alert Streams

24h via Prompt Products

80h via Image Services

Data Release Data Products

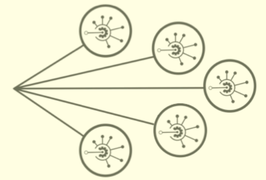
DRP

Final 10yr Data Release:

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

via Data Releases

Fink



Community Brokers

Rubin Data Access Centres (DACs)

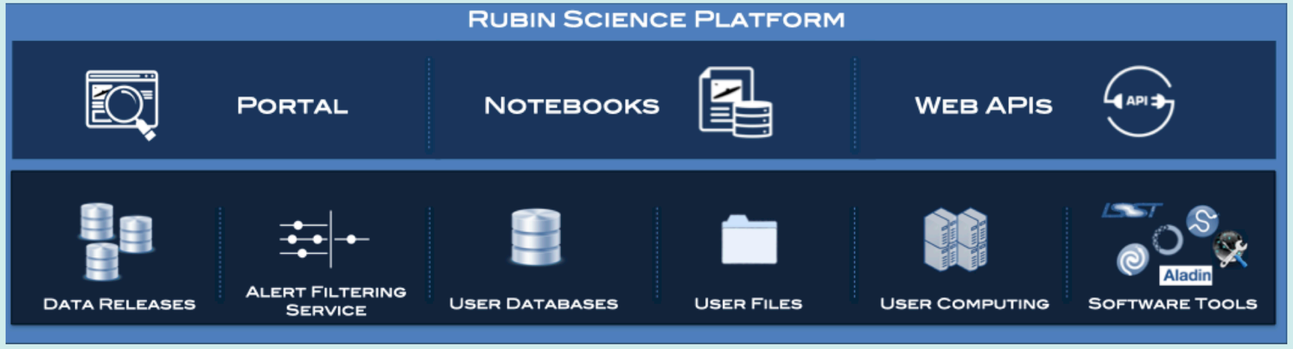
USA (USDF)
Chile (CLDF)
France (FRDF)
United Kingdom (UKDF)

Independent Data Access Centers (IDACs)

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

Rubin Science Platform

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



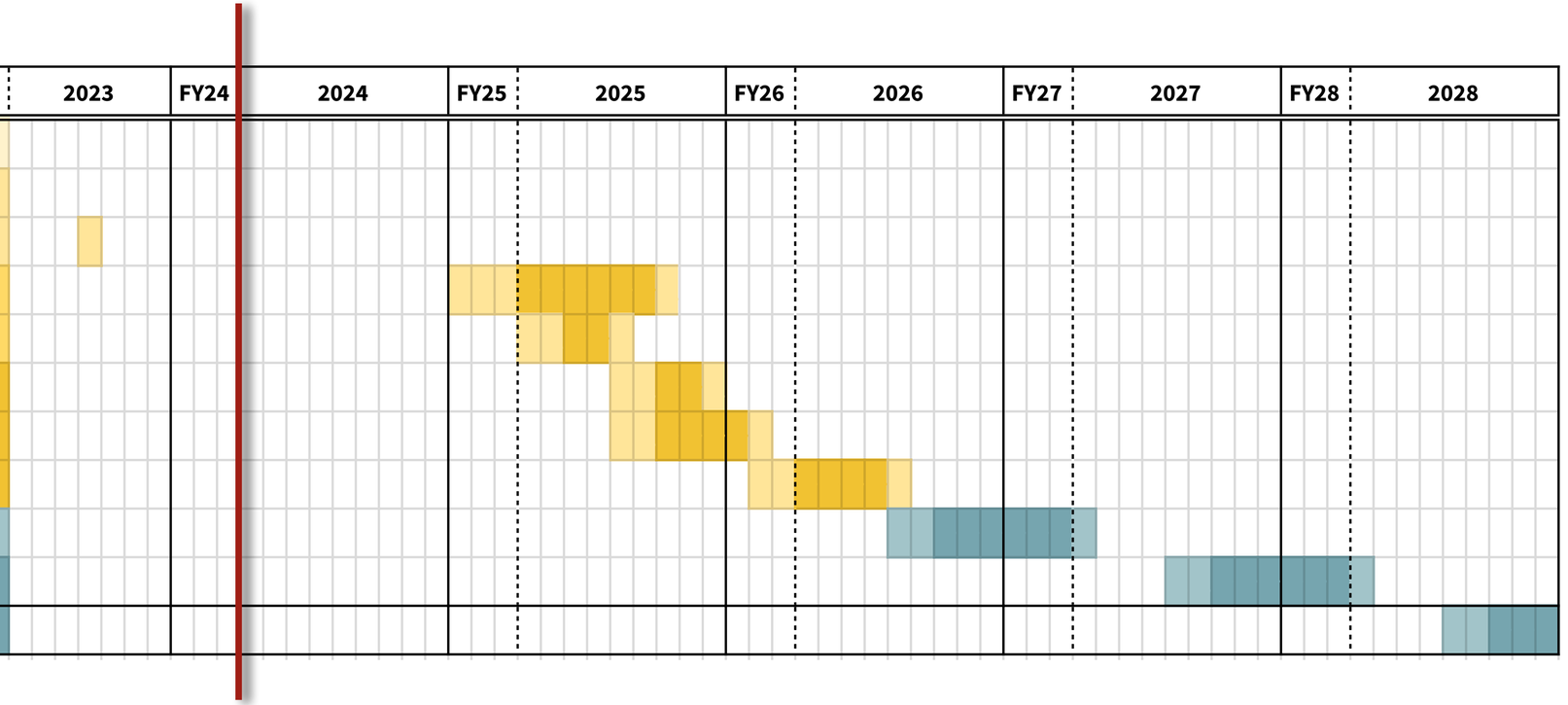
Credit: Leanne Guy

Data Release Timeline

Rubin Operations Survey and Data Release Timeline

Nominal LSST Survey Start Date: August 2025

Event		Date Range
DP0.1	DC2 Simulated Sky Survey	June 2021
DP0.2	Reprocessed DC2 Survey	June 2022
DP0.3	Solar System PPDB Simulation	June 2023
DP1	ComCam/LSSTCam Data	Oct 2024 - Jul 2025
FL	System First Light	Jan 2025 - May 2025
OPS	Start of Operations	May 2025 - Sep 2025
SVY	Start of Survey	May 2025 - Nov 2025
DP2	LSSTCam Science Validation Data	Nov 2025 - May 2026
DR1	LSST First 6 Months Data	May 2026 - Jan 2027
DR2	LSST Year 1 Data	May 2027 - Jan 2028
DR3	LSST Year 2 Data	May 2028 - Nov 2028



DATA RELEASE PROCESSING (DRP)

- Reprocessing of the full raw image dataset to produce the annual data release to be jointly performed at 3 data facilities*

Organisation	*Country*	*Share*
CC-IN2P3	*France*	*40 %*
SLAC	*US*	*35 %*
IRIS Network	*UK*	*25 %*

- SLAC to host the archive center for Rubin and the US data access center

  Univ. Edinburgh to host the UK data access center

** for details see arXiv:2311.13981*

Cloud

EPO Data Center

Dedicated Long Haul Networks

Two redundant 100 Gb/s links from Santiago to Florida (existing fiber)
Additional 100 Gb/s link (spectrum on new fiber) from Santiago-Florida (Chile and US national links not shown)

UK Data Facility
IRIS Network, UK

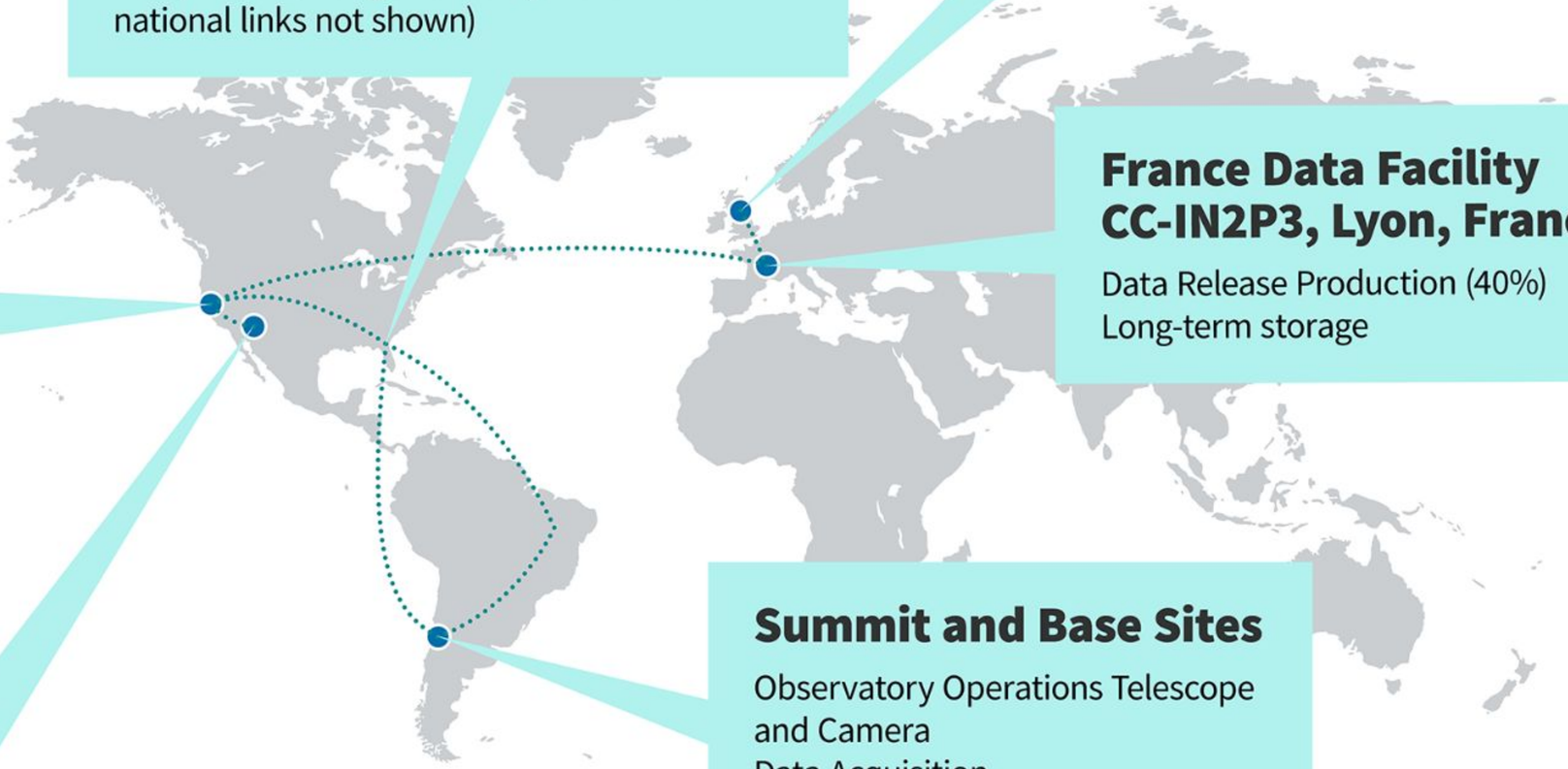
Data Release Production (25%)

US Data Facility
SLAC, California, USA

Archive Center
Alert Production
Data Release Production (35%)
Calibration Products Production
Long-term storage
Data Access Center
Data Access and User Services

France Data Facility
CC-IN2P3, Lyon, France

Data Release Production (40%)
Long-term storage



HQ Site
AURA, Tucson, USA

Observatory Management
Data Production
System Performance
Education and Public Outreach

Summit and Base Sites

Observatory Operations Telescope and Camera
Data Acquisition
Long-term storage
Chilean Data Access Center

VERA C. RUBIN OBSERVATORY

LSST AT CC-IN2P3

- We are preparing to get ready to

  *continuously **import** from SLAC raw image data and locally store a full copy: 20 TB per observing night, 5 PB of new data every year*

  ***process** 40% of the raw data set accumulated since the beginning of the survey, once every year*

  ***export** to SLAC the final products of the reprocessing to prepare the publication of the data release*

  ***repeat** the process every year, for 10 years*

LSST AT CC-IN2P3 (CONT.)

- Several ingredients are required

  capacity to import and export data to SLAC at the required rates

  capacity to locally store raw image data, intermediate and final products of the annual reprocessing

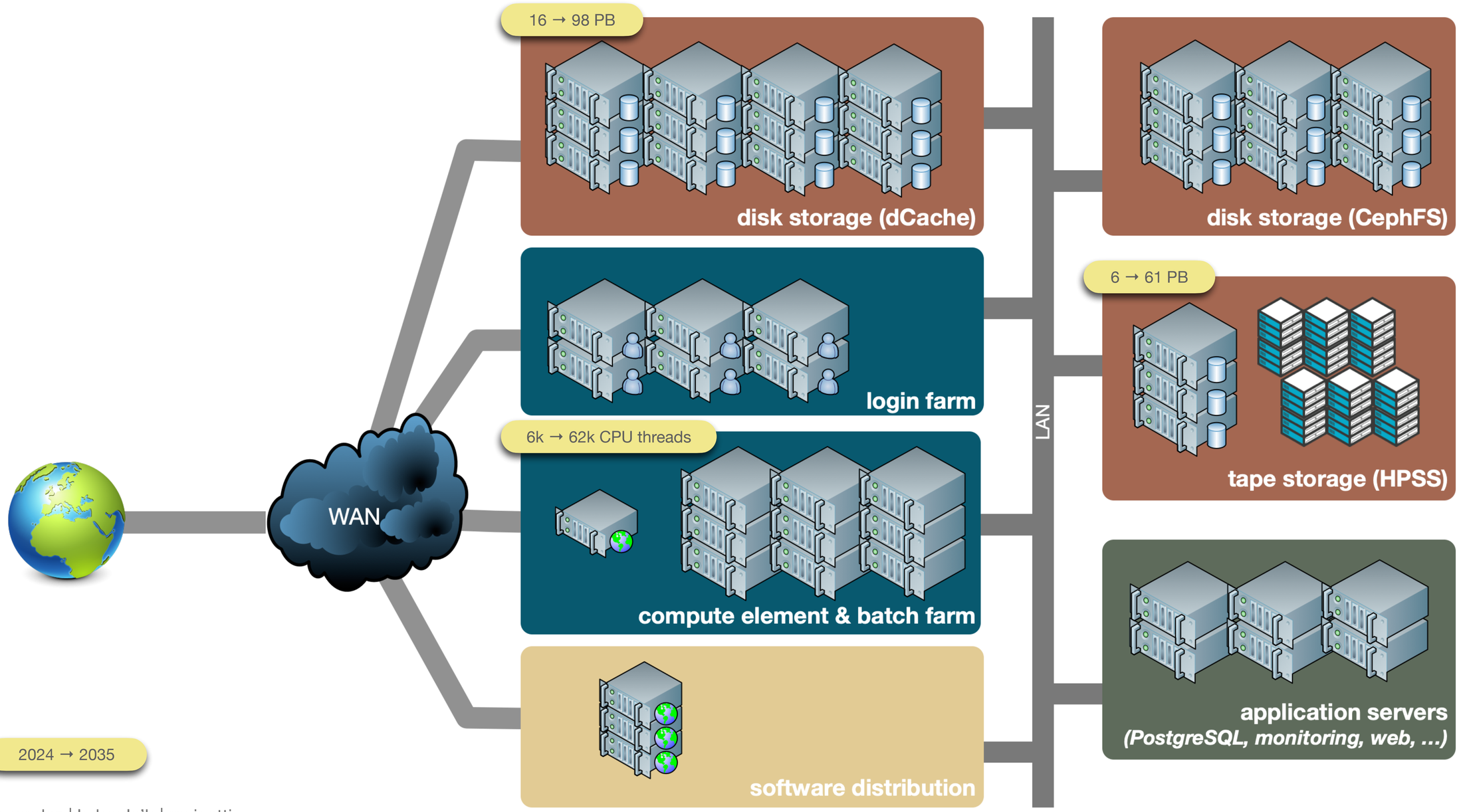
  capacity to execute the LSST Science Pipelines to efficiently process our share of images using the local batch farm

  all of this within a rather tight time budget: 200 days/year

- All this requires work by all CC-IN2P3 teams

  systems, networking, compute farm, storage, applications, support

ARCHITECTURE OVERVIEW



Recent Improvements

Improvements: Compute

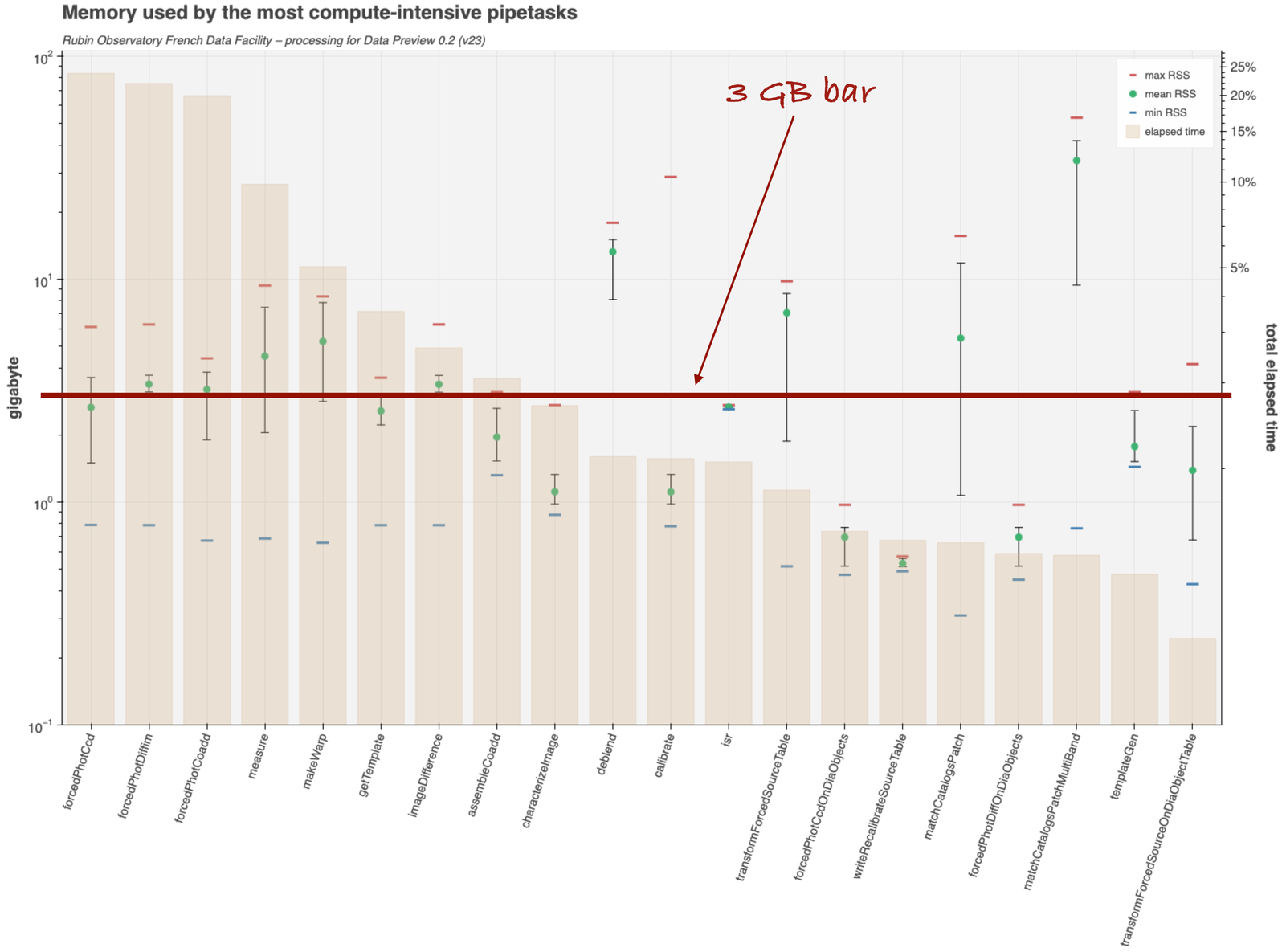
- **Software**

  Slurm replaced GridEngine as workload manager of the batch farm in Apr.'22

- **Hardware**

  the typical ratio memory per CPU thread is 3 GB / CPU thread

  our experience processing DESC simulated data for Rubin's Data Preview 0.2 showed that more RAM is needed



NOTE: the pipetasks shown consume in aggregate 98% of the total elapsed time of the DP0.2 campaign. Whiskers show 5th to 95th RSS percentiles.

Improvements: Compute (cont.)

- Hardware (cont.)

  1300+ additional CPU threads with 9 GB each were put in production in Oct.'23

  currently configured in a separate Slurm partition devoted to LSST users

  we are experimenting with that configuration which should allow us to use more effectively the available CPU capacity

  drawback: purchase cost is 42% higher than the typical compute node

Improvements: Compute (cont.)

- ARM architecture

  started experimenting with compute nodes equipped with ARM processors, which performance per watt is reportedly better than Intel and AMD

  identified some issues building the LSST Science Pipelines on RockyLinux 9, which are being sorted out

- Uncertainty regarding the Linux distribution to use

  in principle, Rubin settled on using AlmaLinux but given the announcements made by RedHat it is unclear if this is a sustainable choice

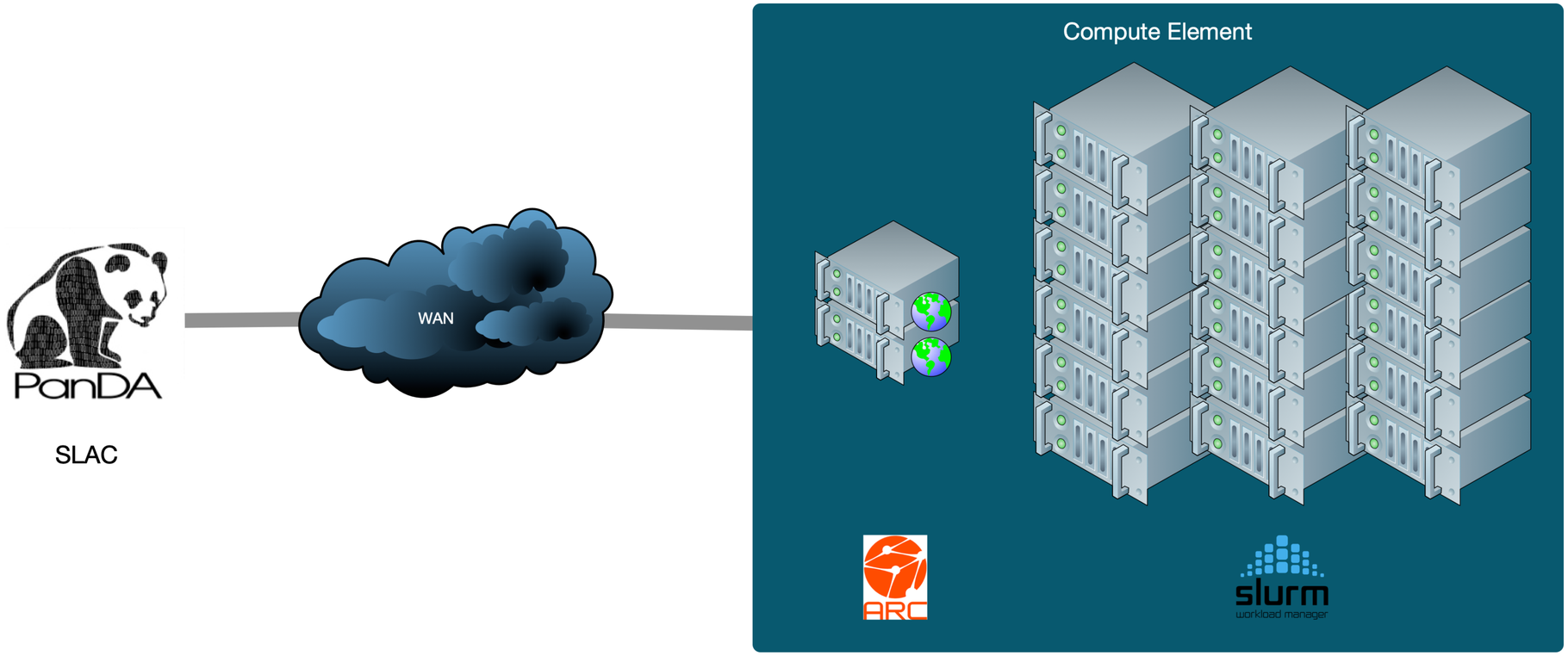
  following closely the discussions in the LHC community

  only relevant for compute nodes

Improvements: Compute (cont.)

- Compute element

  deployment of an ARC compute element to securely expose CC-IN2P3's Slurm farm to Rubin job orchestrator (PanDA) located at SLAC



IMPROVEMENTS: STORAGE

- **Storage**

  for storing LSST data we are using 2 storage systems: CephFS (6 PB) and dCache (11 PB)

  CephFS: POSIX interface, `/sps/lsst`

  dCache: webDAV interface, devoted to DRP for both local processing and inter-facility data exchange

- **CephFS**

  used intensively for producing the Data Preview 0.2

  if you use `/sps/lsst` *you are using this system*

IMPROVEMENTS: STORAGE (CONT.)

- dCache

  over 2023, we conducted more than 15 intensive, realistic test campaigns to certify dCache as butler datastore, to serve image data to 3k to 5k simultaneous Slurm jobs

  we have a better understanding of the adequate hardware and software configuration to sustain the load

  the load on dCache induced by LSST processing is between 5 to 30 times higher, in terms of requests per unit of time, compared to the load induced by LHC experiments

  LSST handles a sheer number of small files relative to the typical size of LHC individual data files

Data Exchange

- Since last summer, we started experimenting Rubin inter-facility data replication using focal plane data: SLAC → CC-IN2P3
  - *realistic in terms of file size, file format and file contents*
  - *8.4M files, 52k exposures, 130 TB*
- Data replication orchestrated by Rucio and executed by FTS3
  - *at CC-IN2P3, those files land in dCache and we immediately copy them under* `/sps/lsst` *to make processing more convenient by Thibault G. et al. (details in the Camera and commissioning parallel session earlier today)*
- Overall a good exercise and we demonstrated the system works
  - *we still need to improve several components to make replication more reliable and to reach and sustain the target rates*
  - *testing of the automatic butler ingestion-upon-arrival with the final tool to be done in the next few weeks*
- Rubin data will be transported across the Atlantic by ESnet and within Europe by GEANT and RENATER
  - *no network links dedicated to transport Rubin data between US and Europe nor within Europe*

Catalog Database

- Experimentations with a local instance of the Rubin catalog database (Qserv) continued

  we loaded into CC-IN2P3's Qserv the catalog data products resulting from processing DESC simulated images for Data Preview 0.2

  catalog products generated at the Rubin Interim Data Facility (Google Cloud) and those independently produced at CC-IN2P3 were both ingested: a comparison of those products performed by Gabriele M. showed good agreement

- ⚠ Important: the budget for storing Rubin catalog database at CC-IN2P3 is not secured yet

  still unclear if you will be able to use LSST catalog database at CC-IN2P3

  more on this on next talk

Rubin Science Platform

- We continue experimenting with a local instance of the Rubin Science Platform: data-dev.lsst.eu

  well integrated with CC-IN2P3 environment, in particular in terms of authentication and access to the file systems you have access to

  includes a catalog and tables viewer, an image viewer and analyzer, an advanced programmatic analysis through the LSST Python software stack using Jupyter notebooks

- ⚠ Important: the budget for operating an instance of the RSP at CC-IN2P3 is not secured yet

  still unclear which tools we will need / have for LSST data analysis at CC-IN2P3

  more on this on next talk

PYTHON NOTEBOOKS

- CC-IN2P3 Jupyter Python notebooks service continues to improve: notebook.cc.in2p3.fr

  although not dedicated to LSST, many of its recent developments were motivated by ZTF and LSST use cases and now benefit many other projects

- You can now also use Dask

  from the comfort of your Python notebook you can create and drive your own ephemeral cluster using compute nodes of the Slurm batch farm

  well suited for interactive analysis, in particular, if your analysis can exploit data parallelism

- Details in the documentation

  example use cases (by Mickael R. and Dominique B.) and the internals of the service available in this recent talk by B. Chambon

Distribution of the LSST Science Pipelines

- CC-IN2P3 operates a central repository of stable and weekly releases of the LSST Science Pipelines

  globally distributed via the CernVM file system

  details at sw.lsst.eu

- This distribution is used by the 3 Rubin data facilities to get the image analysis software to produce the data releases

  all the facilities will use a bit-by-bit identical distribution of the pipelines: good for reproducibility

MONITORING

- Ongoing work to consolidate our monitoring tools

  mon.lsst.eu

  mostly based on ElasticSearch + Grafana

- Our goal is to be able to observe all the activity induced by LSST from a single entry point

  compute, storage, butler databases, catalog database, data exchange, etc.

PROFILING

- Ongoing work to profile the behavior of the pipelines

  first target is to understand the memory usage patterns

- See Johan's talk earlier today

  further details in this talk by Quentin Le Boulc'h

PERSPECTIVES

What is next

- Jointly processing of HSC public data release 2 by all the Rubin data facilities

  started last week, now ramping up

  Rubin Campaign Management team is conducting this exercise

- Iterate over the inter-site data exchange activity

  increase reliability, improve monitoring

- Exercise the mechanism to store raw data on tape
- Test, test, test

SUMMARY

SUMMARY

- We have made significant progress preparing CC-IN2P3's infrastructure for the Rubin challenge

  compute, storage, data exchange, catalog database, butler, connectivity, etc.

  we also improved our understanding of how the system is expected to work

- Production of LSST data releases will be a real challenge for CC-IN2P3

  we are in a better position now that one year ago, thanks to the work done in several areas by many people

  we feel we are building on top of solid foundations but a lot of work is ahead of us