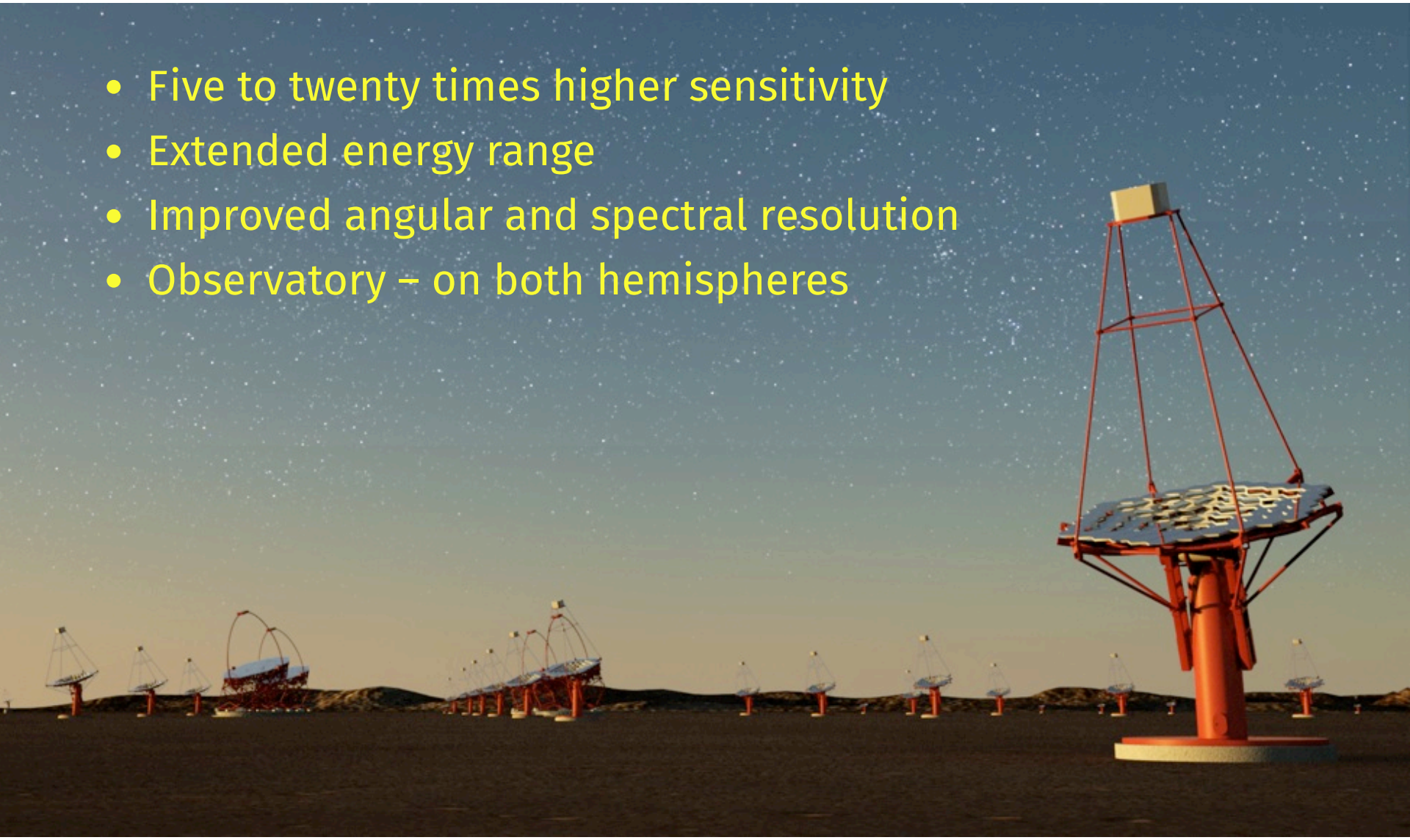
cherenkov telescope array

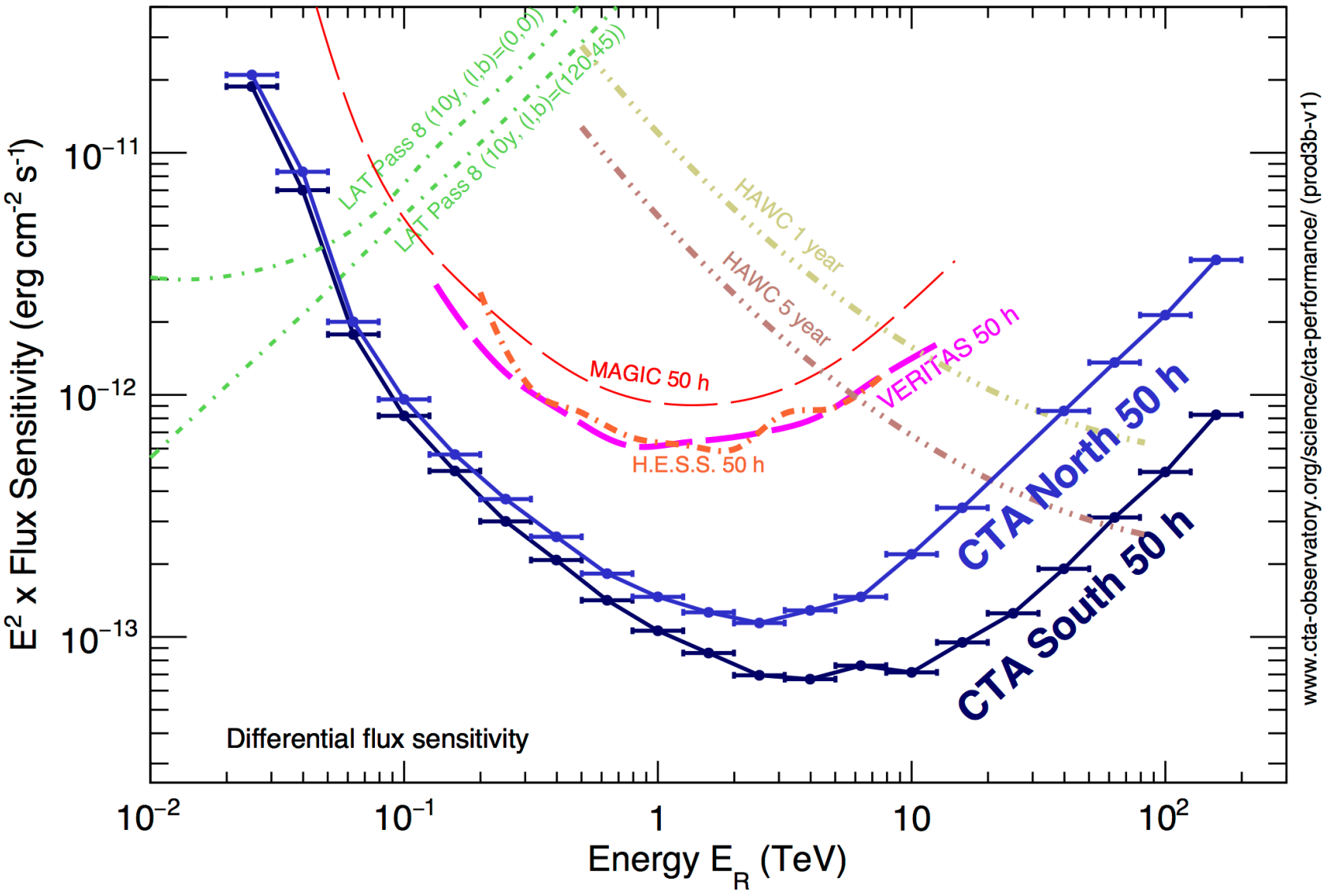
CTA: From Bulk Data Archive To Science Portal

Stefan Schlenstedt , CTAO
Computing Coordinator

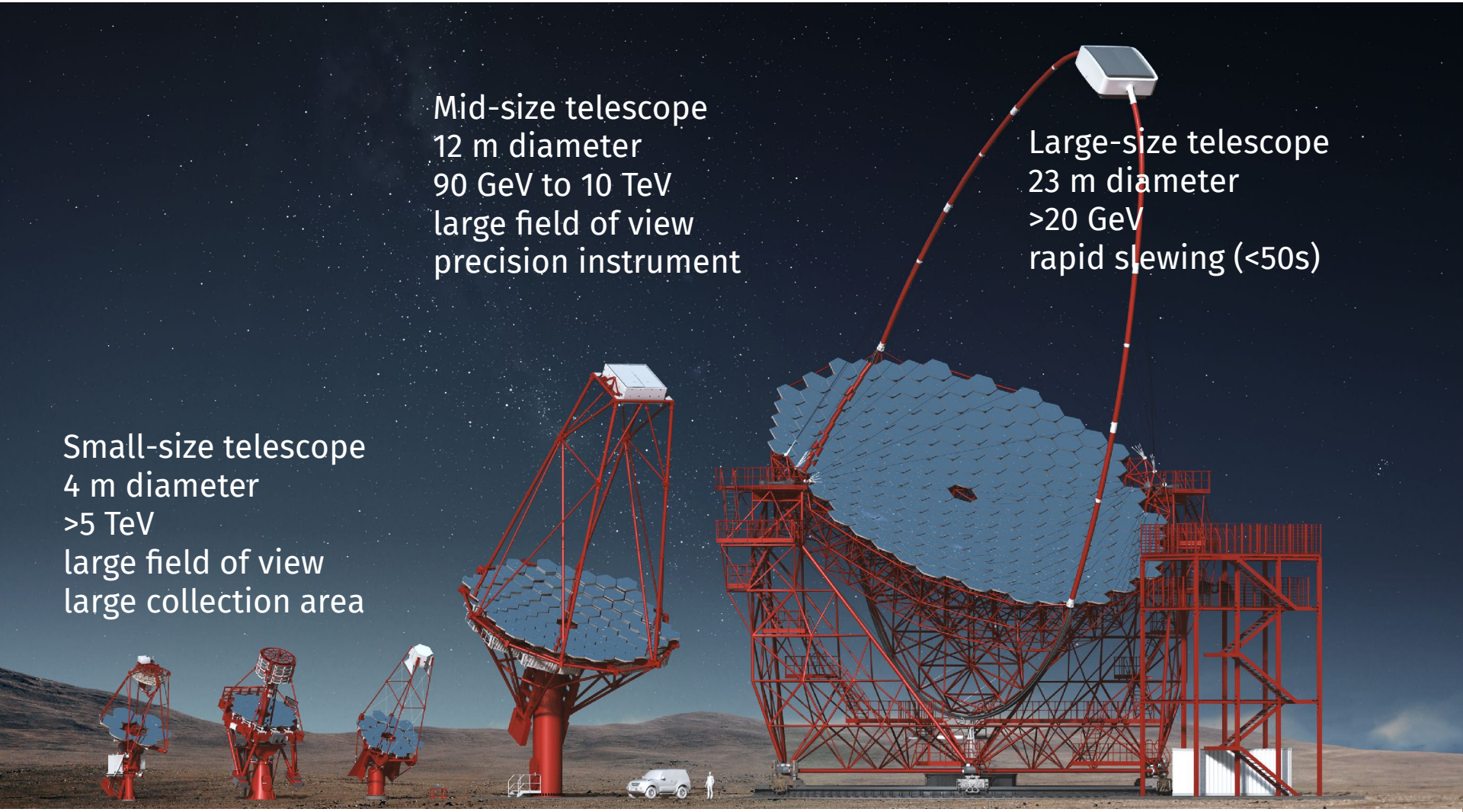
Cherenkov Telescope Array
CTA Observatory



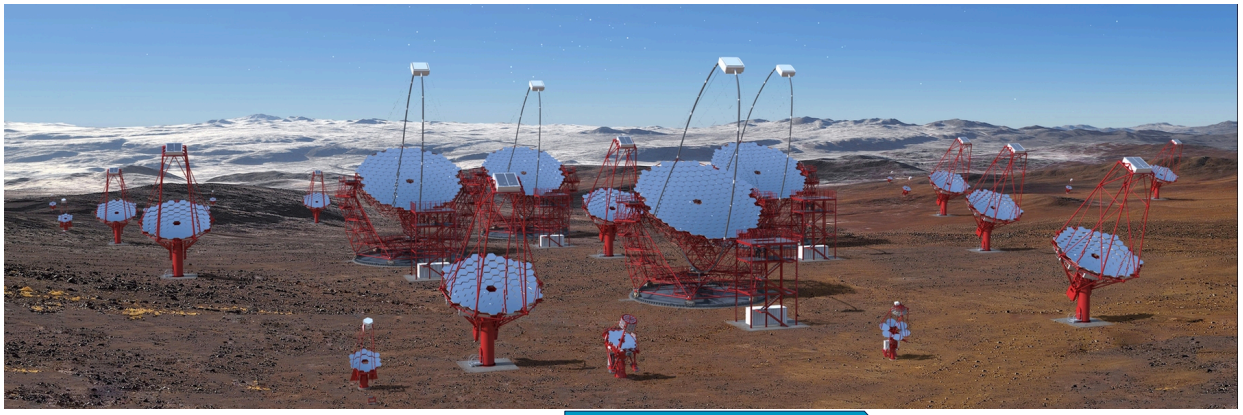
Sensitivity



Two Array of Telescopes
In Chile and Spain



Typically ~2000 pixel cameras with trigger rates: LST 15, MST 9, SST 0.6 kHz
Readout of roughly 60-100 ns with 0.25-1 GHz sampling



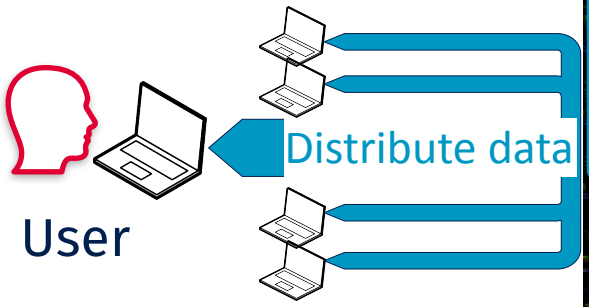
600 Gb/s
= 1000 PB/yr

On-site clusters

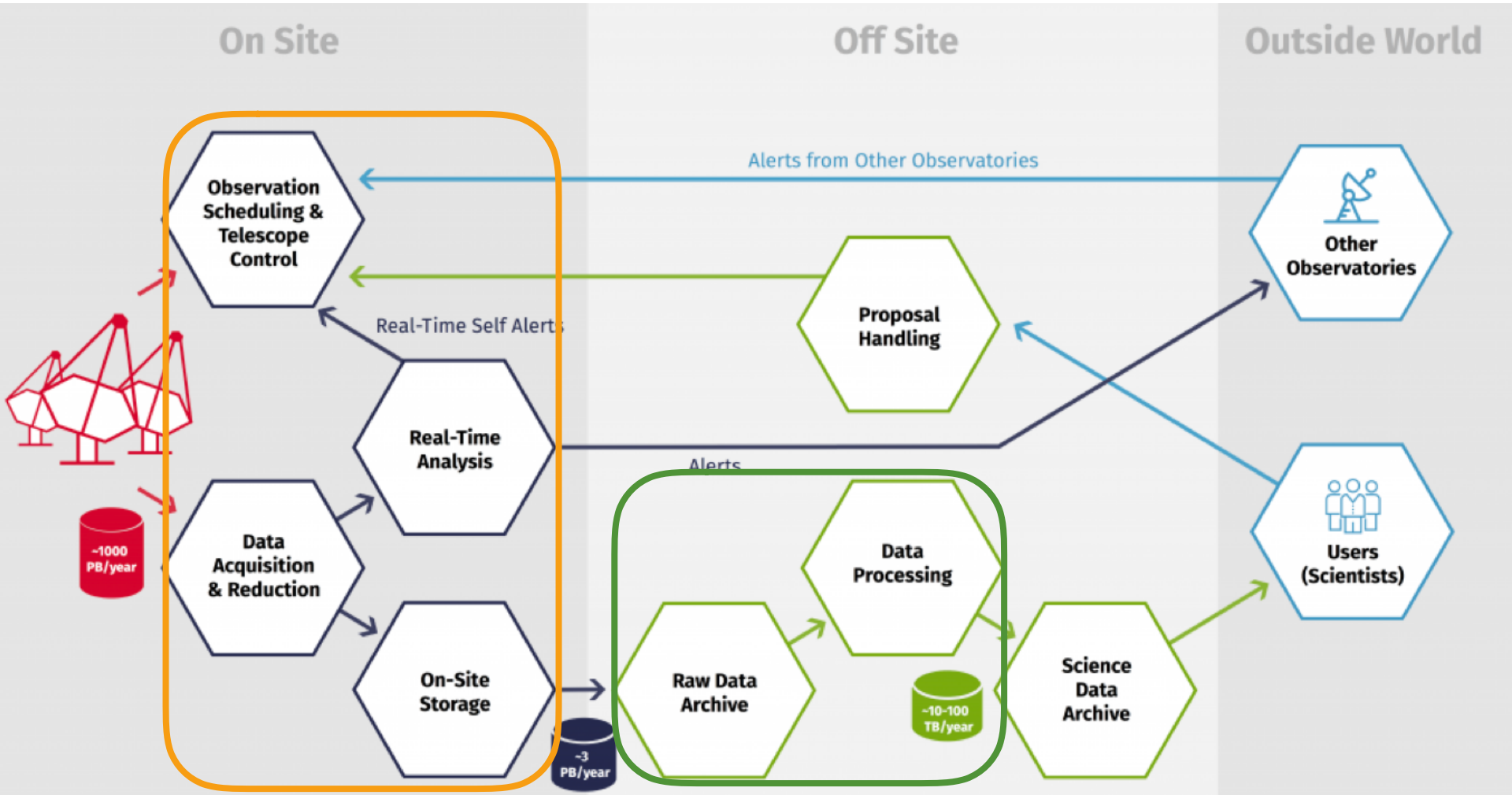
Process
Reduce

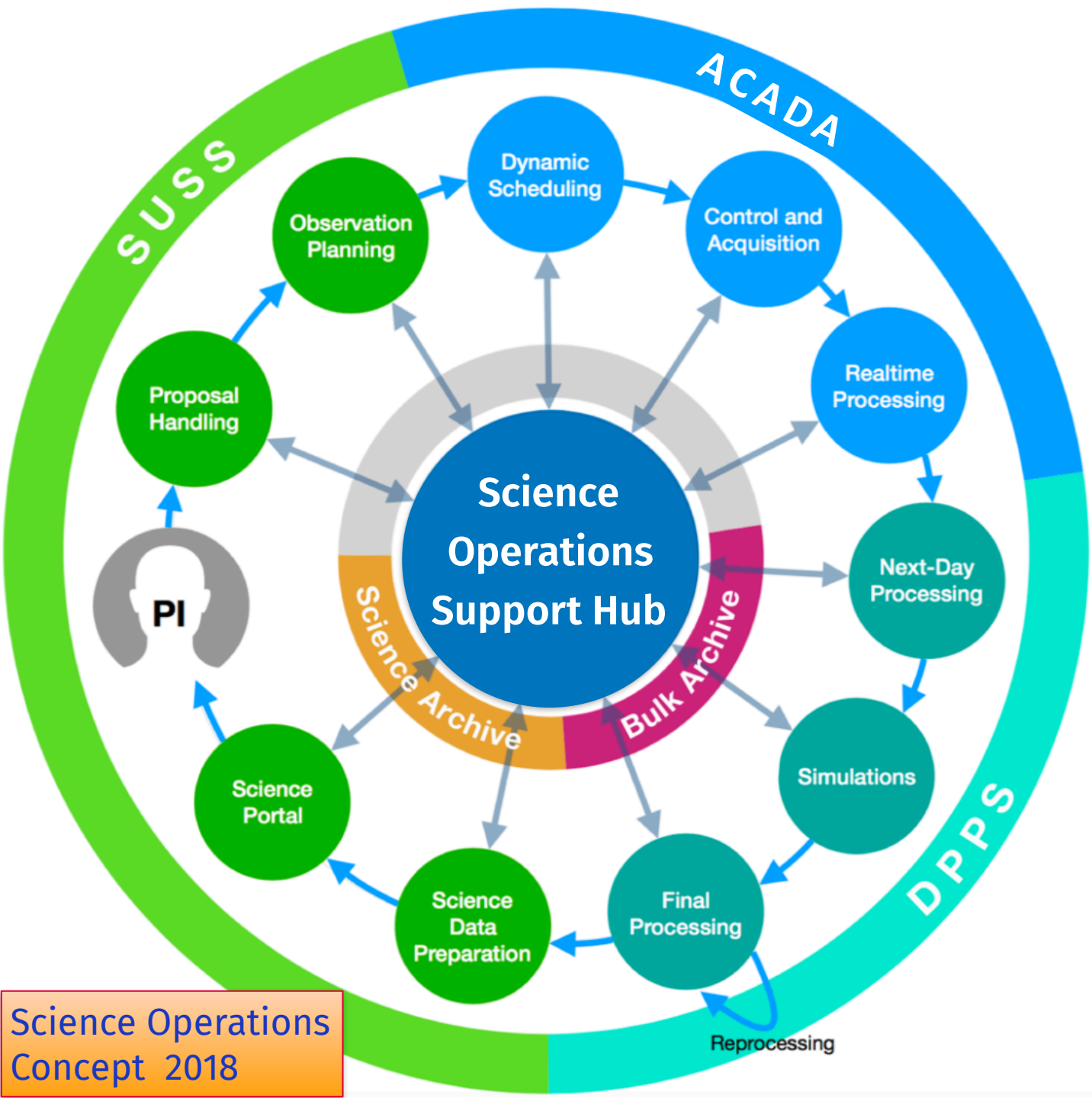
Limited bandwidth

Off-site computer centers



Process Reduce Store

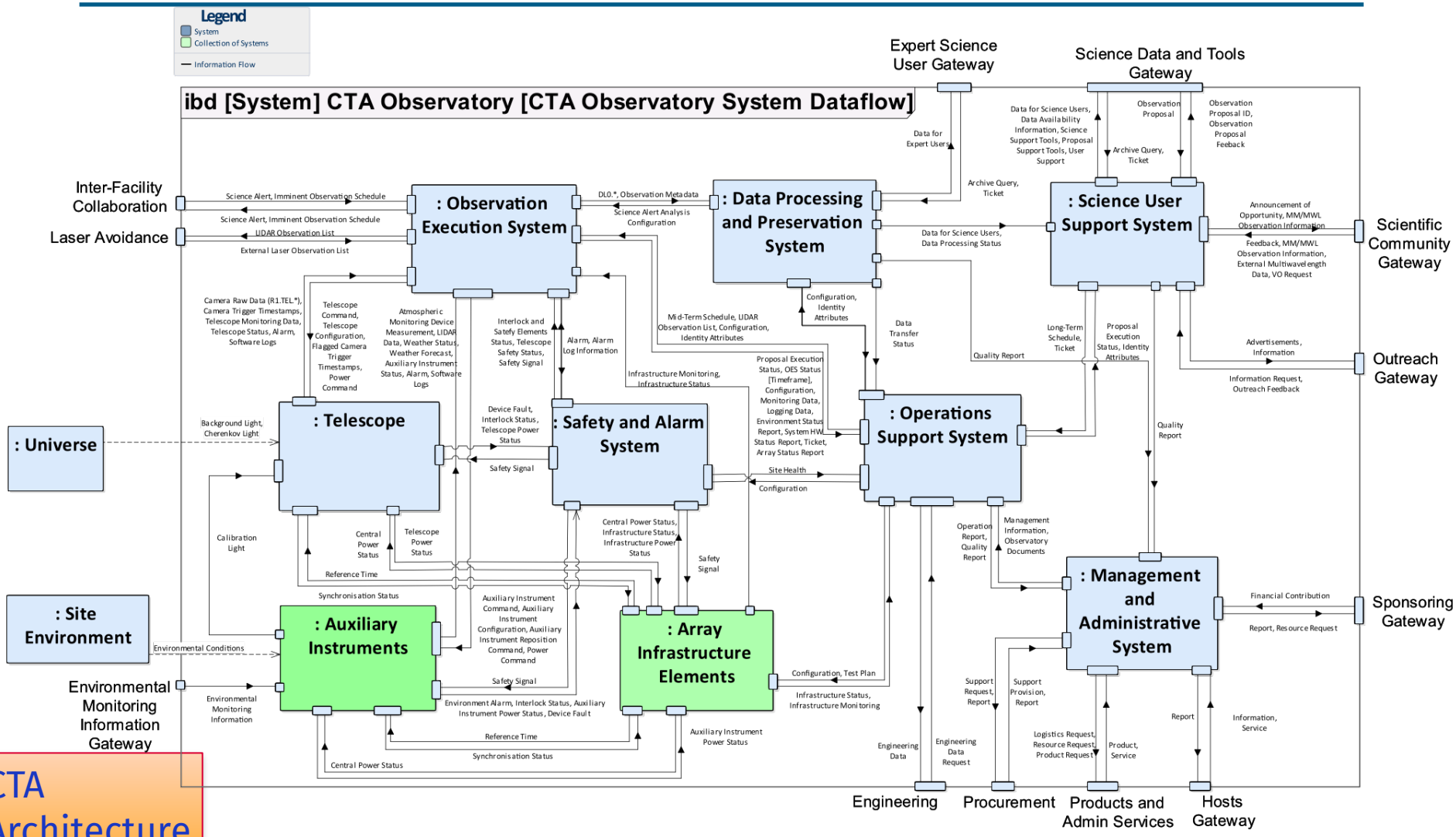

On Site
Off Site
Outside World
Observation Scheduling & Telescope Control
Alerts from Other Observatories
Other Observatories
Proposal Handling
Real-Time Self Alerts
Real-Time Analysis
Alerts
Data Acquisition & Reduction
~1000 PB/year
Data Processing
Users (Scientists)
On-Site Storage
Raw Data Archive
~10-100 TB/year
Science Data Archive
~3 PB/year
Array Control and Data Acquisition
Data Processing and Preservation
Science User Support System



Science Operations Concept 2018

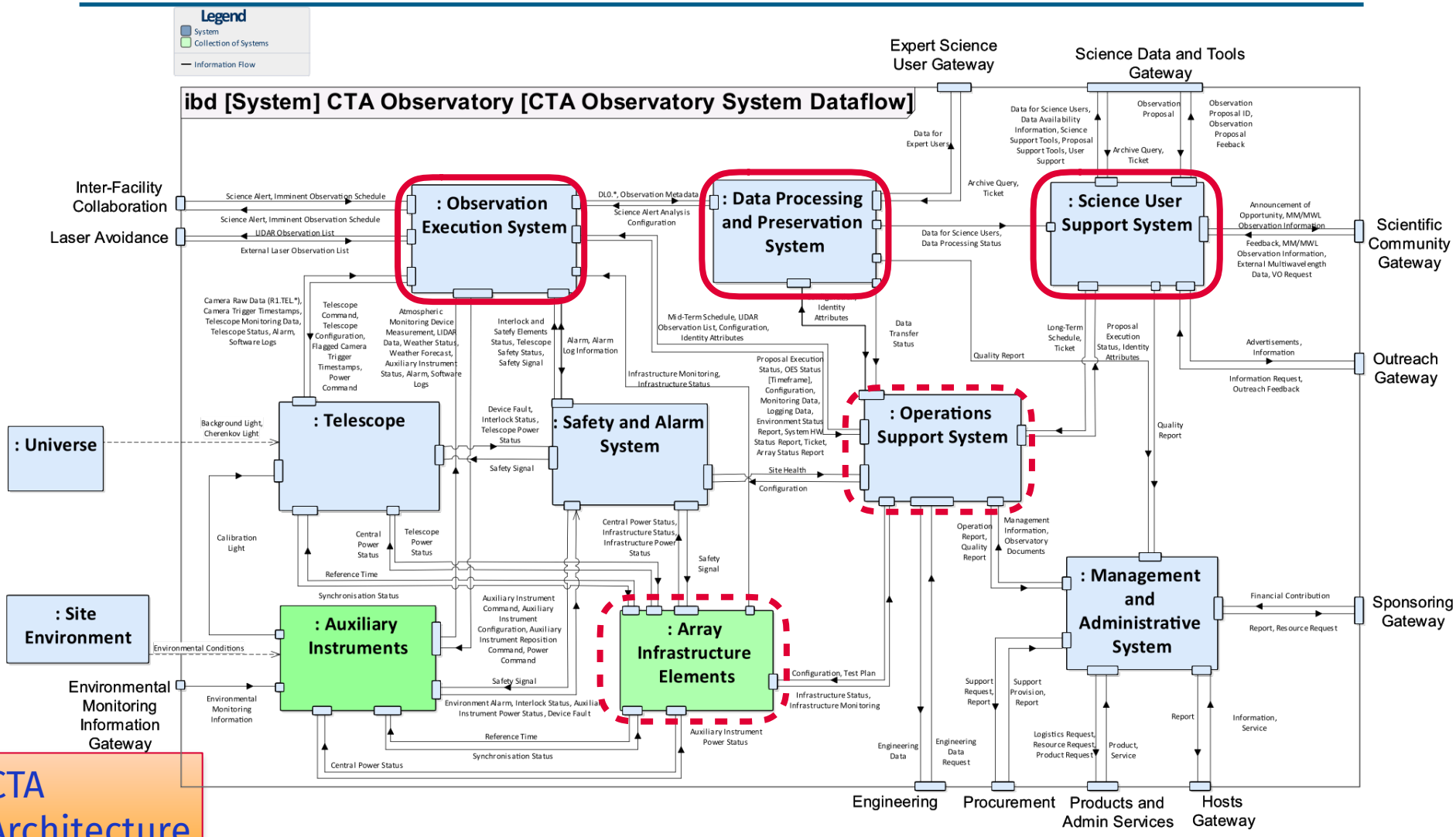
CTA Observatory science operations – primary processes

The CTA Systems Structure



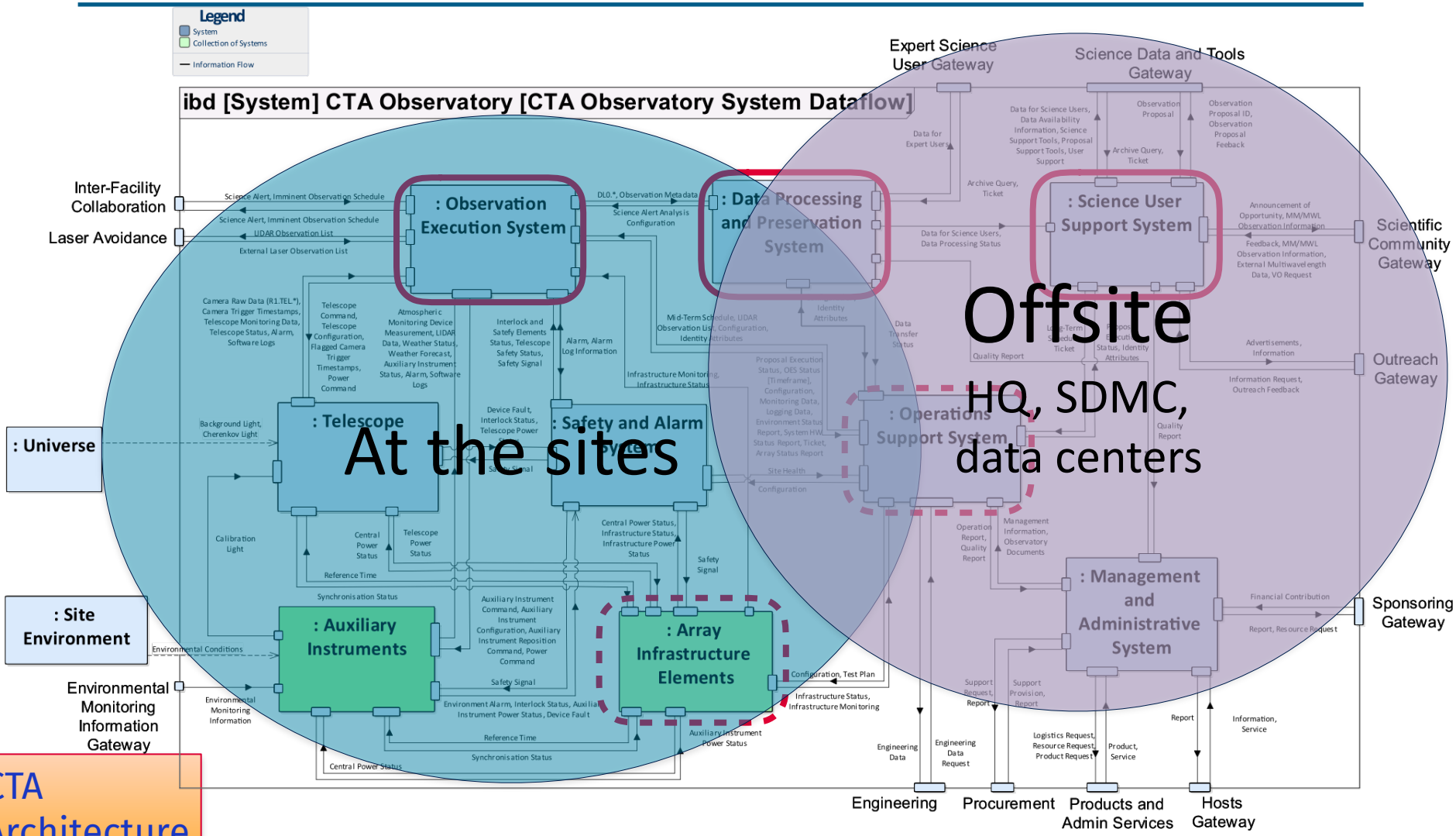
CTA Architecture 2018

The CTA Systems Structure



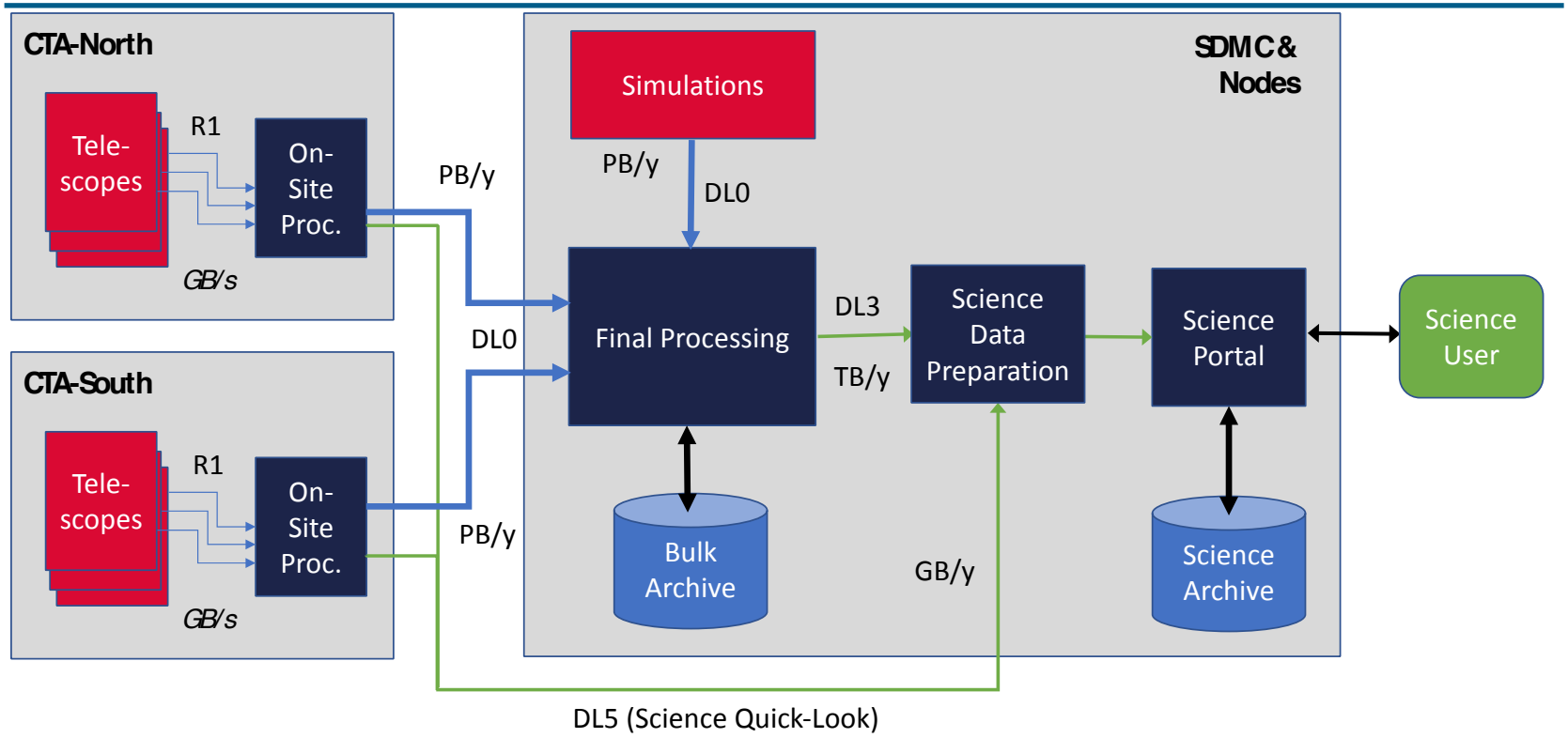
CTA Architecture 2018

The CTA Systems Structure



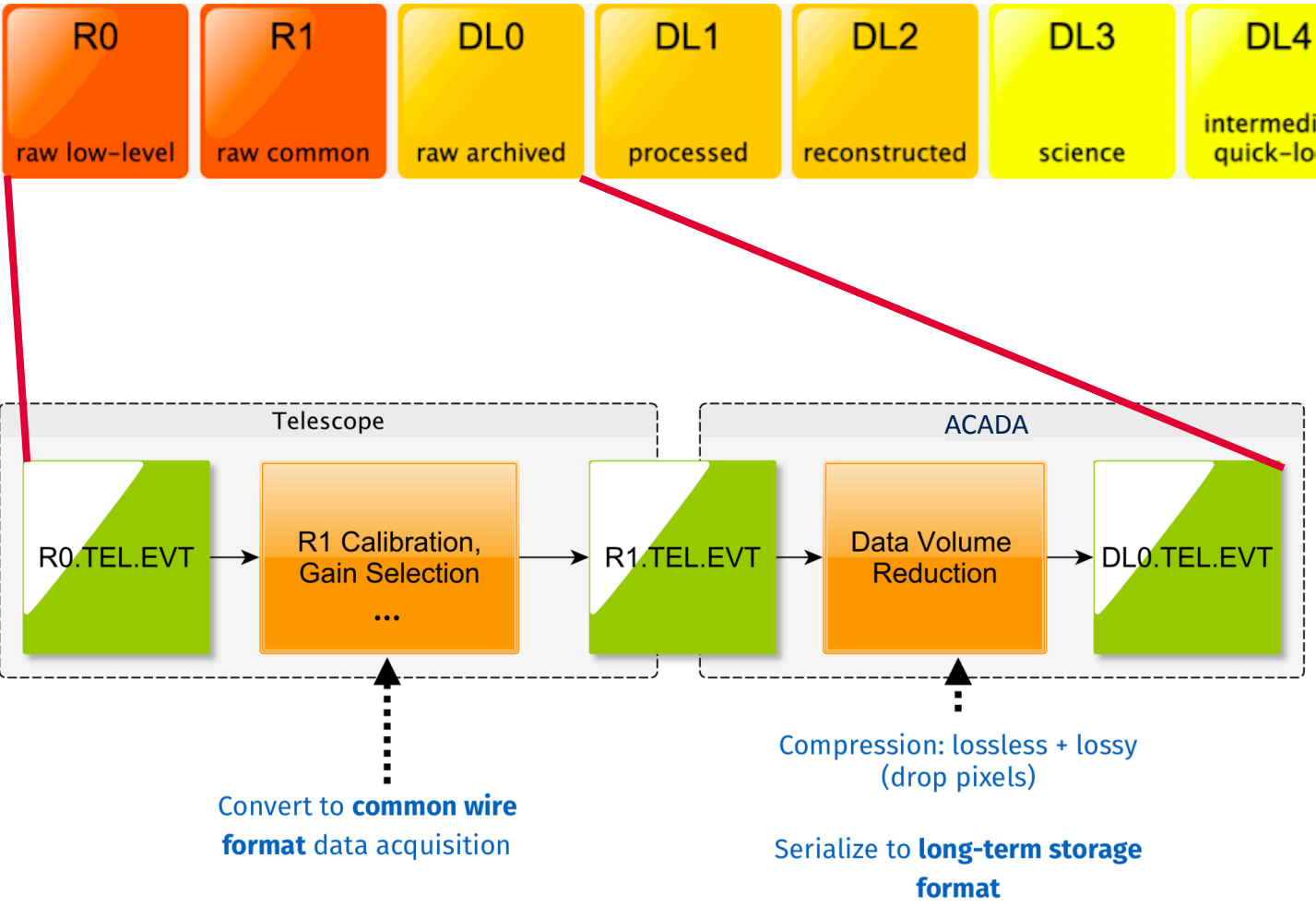
CTA Architecture 2018

Data Flow and Categories



- Strong data reduction along the processing steps
  - From PB/y (at raw data level) to GB/y (high-level science data)
- Open access through Science Portal
  - access to science archive, to science analysis tools
  - Exploration of quick-look data products

The Data



- Data model
- Data format
- Requirements
- Interfaces
- Prototypes
- Selection – Test procedure

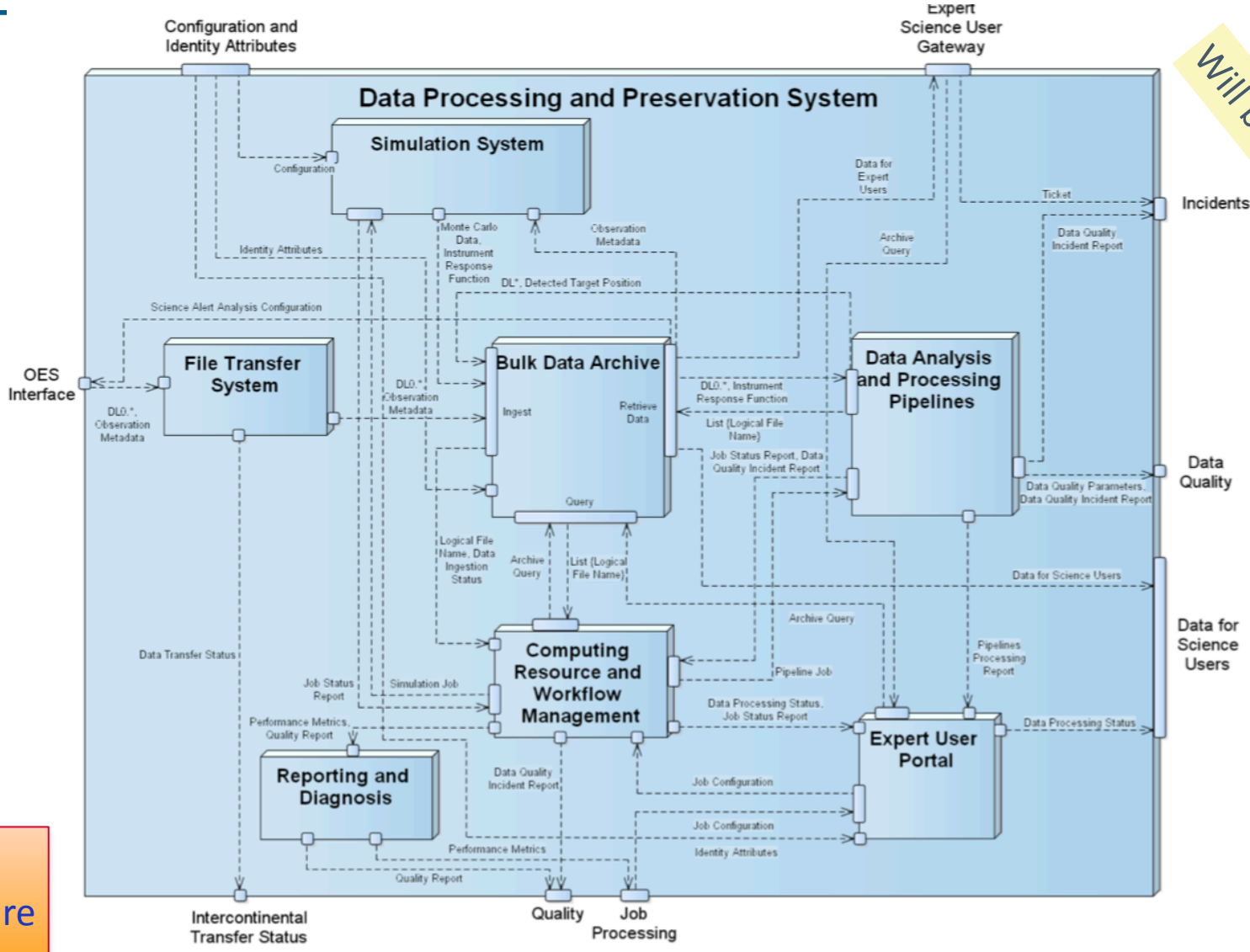
Distributed Data Processing and Storage

DPPS

- Manages the data
- Responsible for processing it down to a form usable by scientists
- Responsible for simulating the data needed to characterize the instrument
- Reprocessing

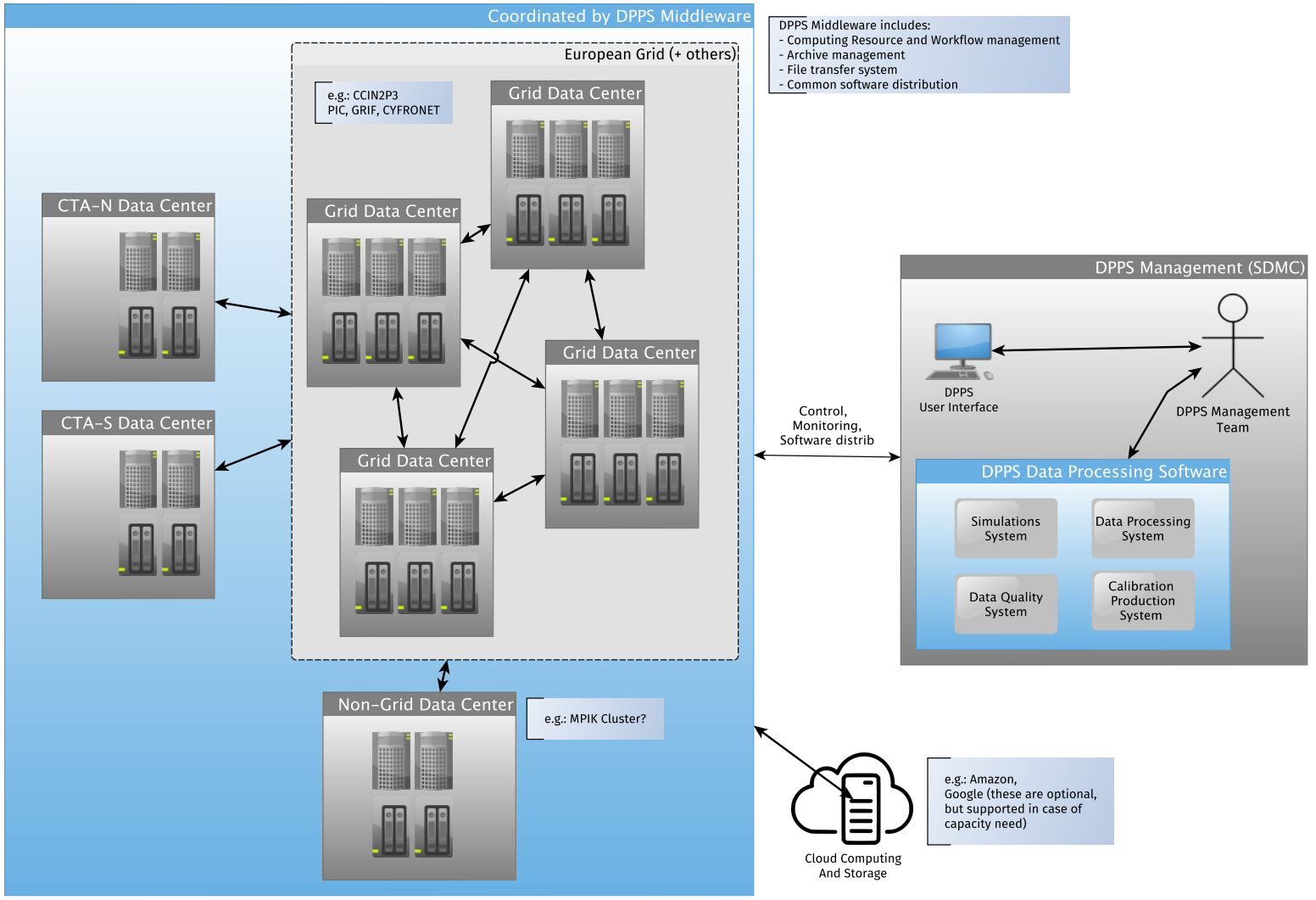
Requires a complex chain of algorithms running in parallel at multiple data centers

Distributed Data Processing and Storage DPPS



CTA Architecture 2018

DPPS Concept



DPPS Concept

- Distributed Data Processing and Storage is a “cloud”
  - Centrally managed at Science Data Management Centre (SDMC)
- Software is running at all Data Centers, including the sites
- The software running on a single data center is a Data processing and preservation node (DPPN)
  - What we deliver to the site is the software and data packages necessary to make the site’s Data Center a DPPN
  - ‘Just’ a sub-set of the packages needed to run the full DPPS, e.g.
    - No simulation pipelines
    - Data processing pipelines only need subset of functionality

DPPS Products

Full system managed by SDMC

- Documentation

Middleware

- Common Libraries and Frameworks
- Computing Resource and Workflow Management Middleware
- Bulk Archive Management Middleware
- File Transfer Middleware

Pipelines

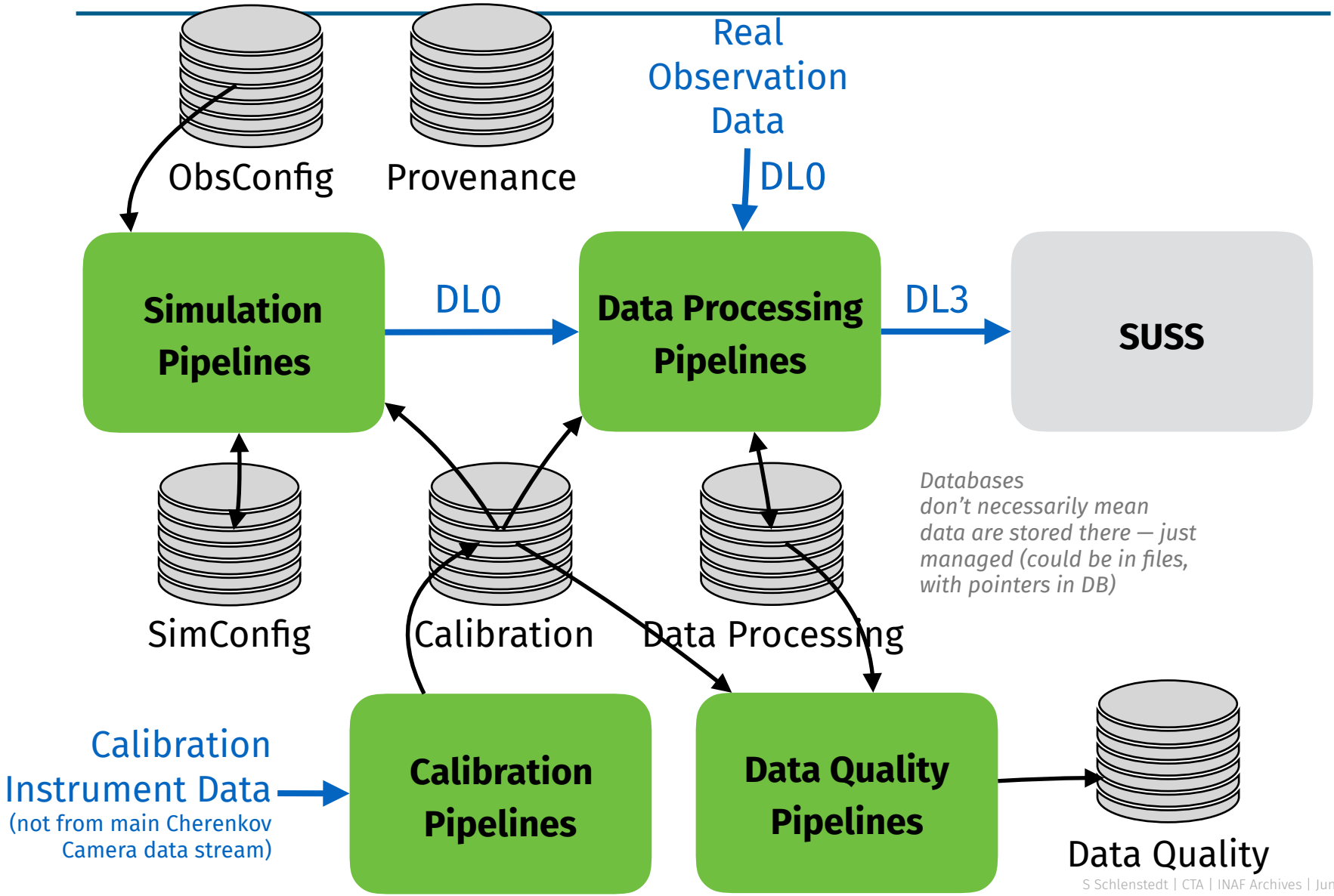
- Simulation Pipelines System
- Data Processing Pipelines System
- Calibration Pipelines System
- Data Quality Pipelines System

Pipeline systems each have similar substructure:

- Framework
- Tools
- Workflows
- Database

- Operations UI
- Data Quality UI and Reporting

DPPS Pipelines Concept



Archives

- Archives are in the centre of science operations from begin to end
- Archiving of data products and software at different levels
  - Bulk Archive for DL0-DL3
  - Science Archive for DL3, DL5, DL6
- Archiving of metadata linked to the different levels, including provenance information
- Archiving of additional information (the Central Hub / Science Operations Support System):
  - Proposals and Schedules
  - Status information
- Archiving of monitoring and engineering data

Approach to Archives in CTA

A lot of definition and prototyping work done in the past and documented in the DATA TDR (2016), e.g.

- Detailed data volume calculations
- Design studies for archive systems
- Use cases for archive access

Prototyping work in framework of European Projects

- European H2020 Projects where CTA participates as a major use case study (e.g. ASTERICS, XDC, ESCAPE)

Some key aspects:

- Established OAIS model as basis for the archive
- Based on assumptions on operational model of CTA data handling and access

Data Usage Scenarios

- CTA Data at DL0 and above (except intermediate data) needs to be preserved long-term
- CTAO internal users need access to DL0 and DL2 data for reprocessing and diagnostics / quality monitoring → bulk data archive (DPPS)
- End/science users need flexible query-based access to DL3-6 products → science archive (SUSS)
- End/science users and CTAO internal users need fast, flexible, query-based access to CTA meta-data in bulk archive and science archive depending on the planned data access
  - Access to lower-level data for small number of end/science users
  - Large number of workflows associated with operations
  - Concept in need of elaboration

CTA Data Archives

DPPS Requirements Workshop 2018

- Long-term Preservation of Bulk Data (in DPPS)
  - **Large** amounts of **data** [10s of PB/yr]
  - **Small** number of **users**
  - Need for coordinated computing and storage (likely distributed)
  - No need to be maintained after end of CTA operational phase
- Long-term Preservation of Science Data (in SUSS)
  - **Small** amount of **data** [10+ years of archive fits on an SDCard]
  - **Large** number of **users**
  - Outlives operational life of CTA by 10+ years
  - Can be centralized
- Clear difference in access needs
- Functional split – can use the same technology / be implemented as a single archive or as two separate (implementation detail)

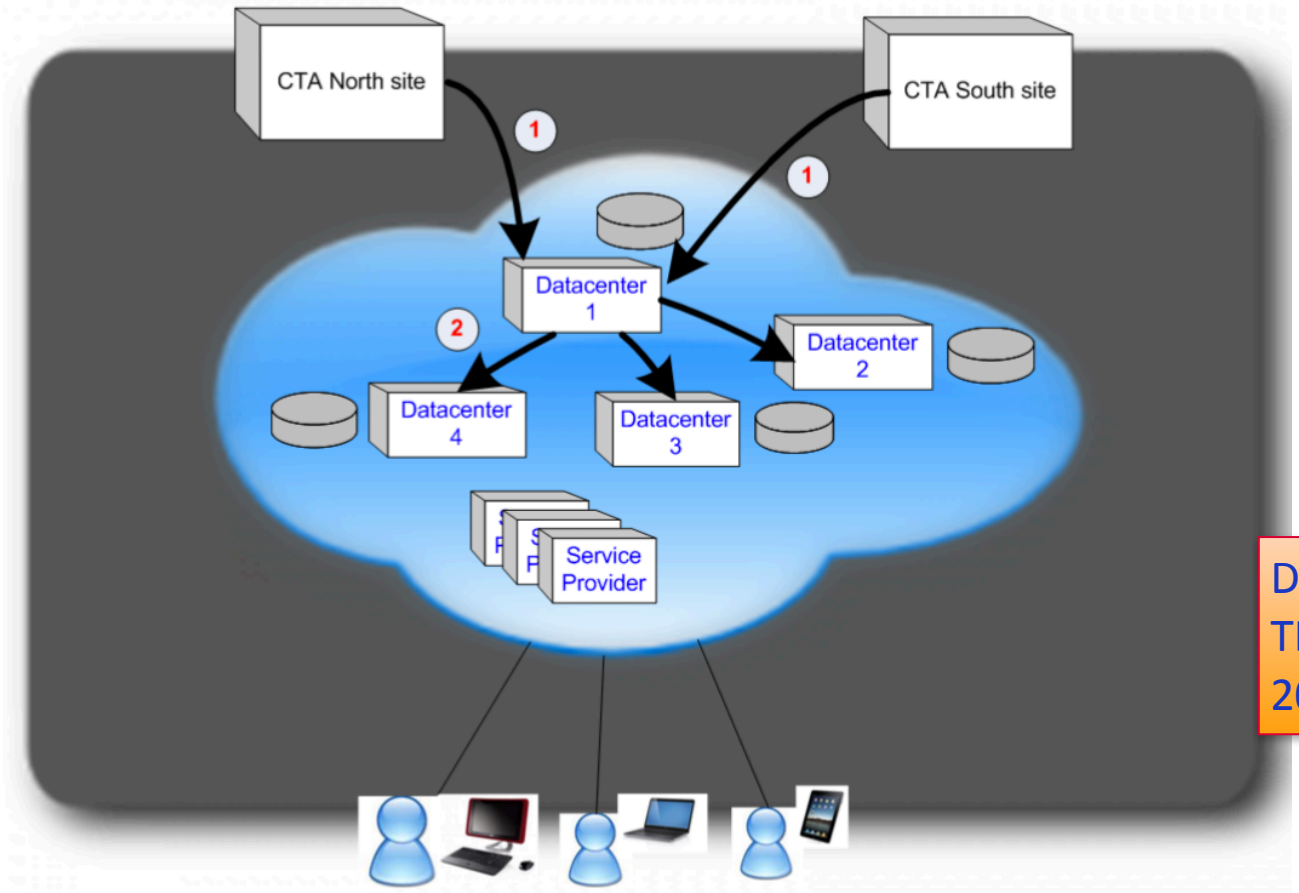
The Bulk Archive

- DPPS provides *Archive Management Middleware* (software)
  - The software that manages the preservation of CTA data, and ensures its availability for access for (re)processing.
  - Ensures “RAID-like” replication between *N* Data Centers (subject to computing model decision):
    - each file must be at at least two Data Centers
    - but not necessary that there are *only* two Data Centers
  - Provide interface like Open Archival Information System (ingest, retrieve, query)
    - implies also a file catalogue with search on file meta-data
- This software and the service agreements with *N* data centers, integrated and centrally managed by the SDMC is The Bulk Archive
  - Therefore an IKC to produce the DPPS Archive Management Software is not creating the full archive, rather

The Distributed Computing Model

- Under responsibility of SDMC

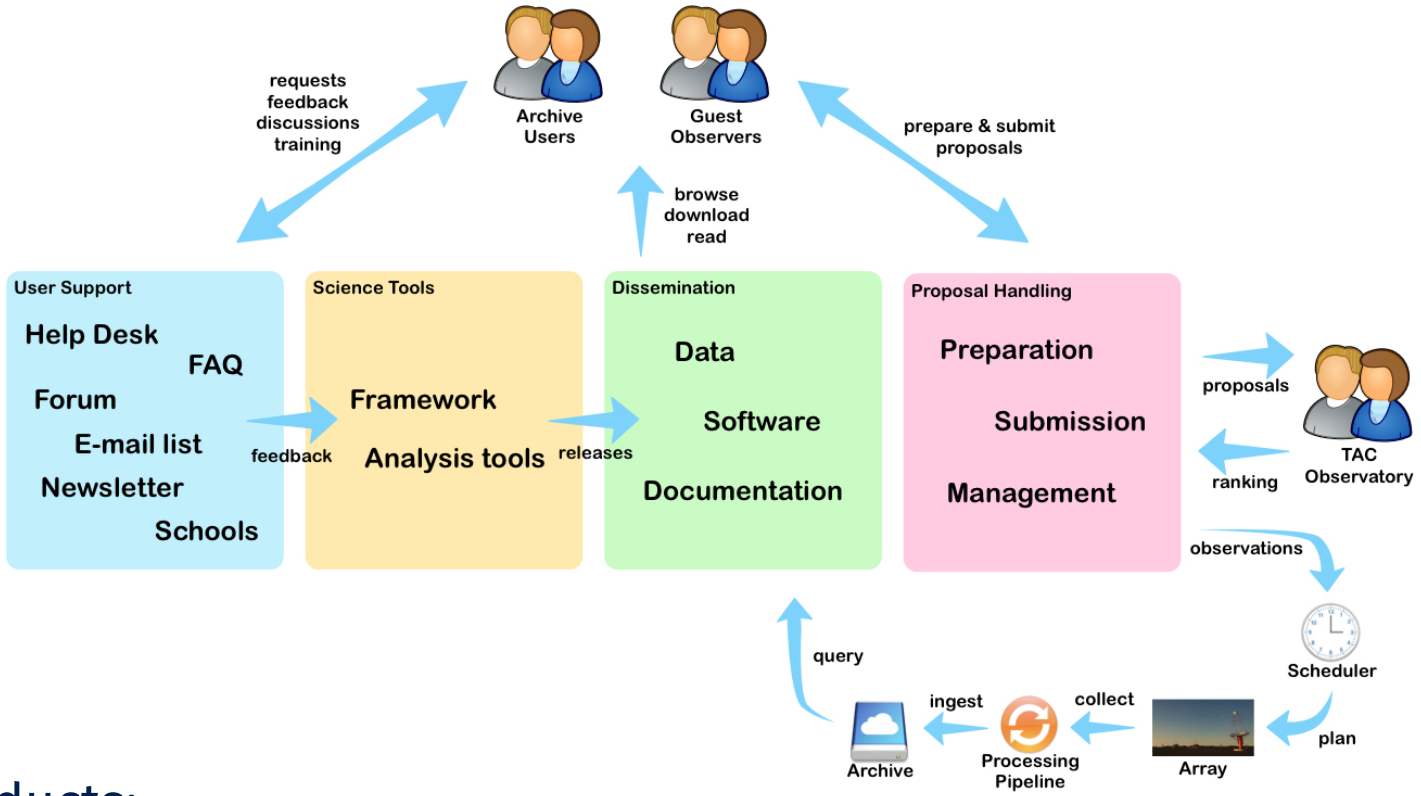
Under investigation



DATA
TDR v2.0
2016

- #data centres and data flow – updated in the Computing Model

Science User Perspective



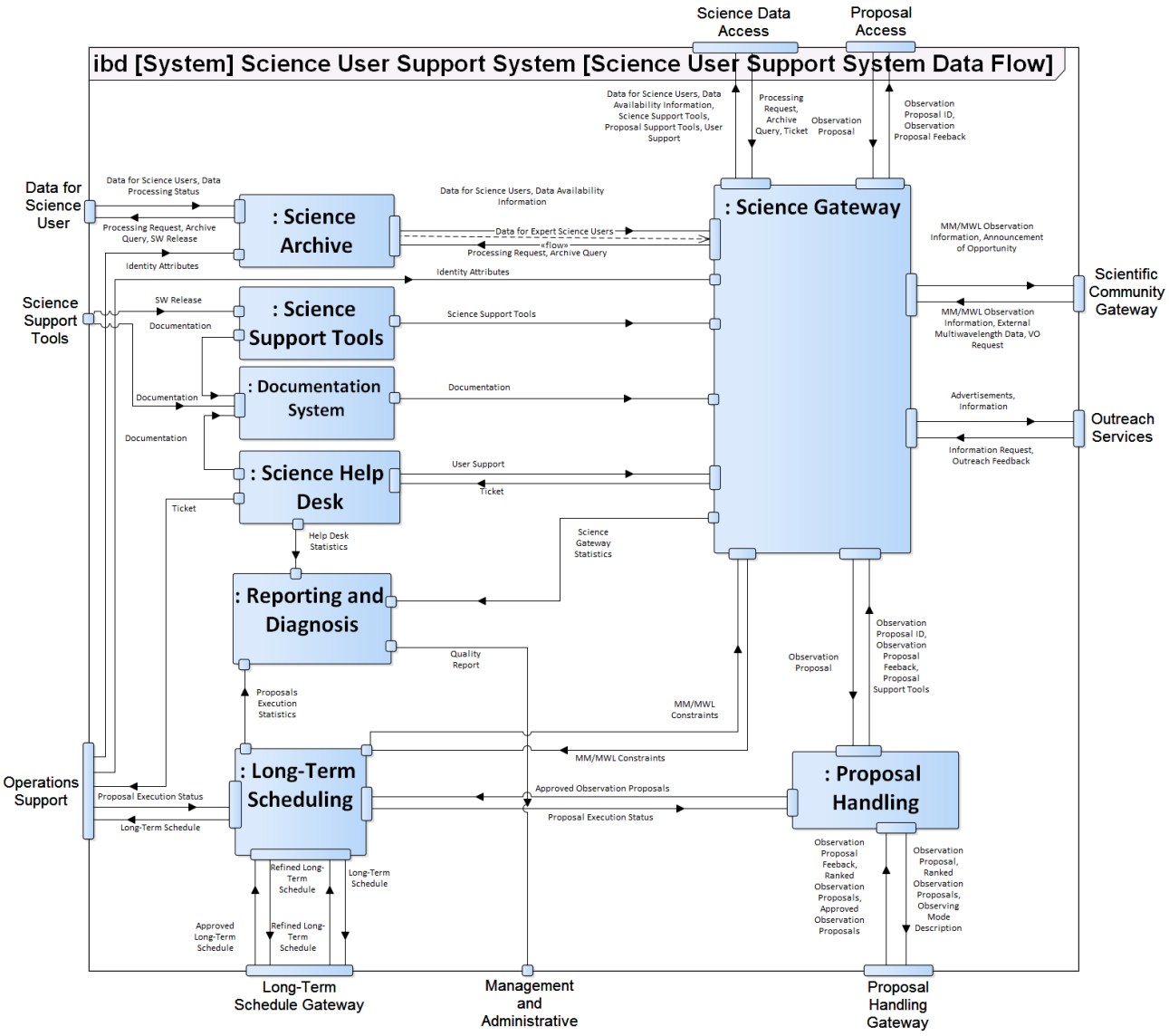
- Products:
  - Photon (candidate) event list data (FITS)
  - Instrument response functions, background model
  - Science analysis tool suite, supporting docs

Science User Support System
SUSS

The software systems responsible for science operations – gateway to the world

- Includes Software for :
  - Proposal Handling
  - Long-Term and Mid-Term Scheduling
  - Automatic Data Product Preparation (DL3→DL5) and Verification
  - Science Analysis Tools
  - Science Archive
  - Science Portal
  - Help Desk and User Support
  - Reporting/Diagnosis

Science User Support System

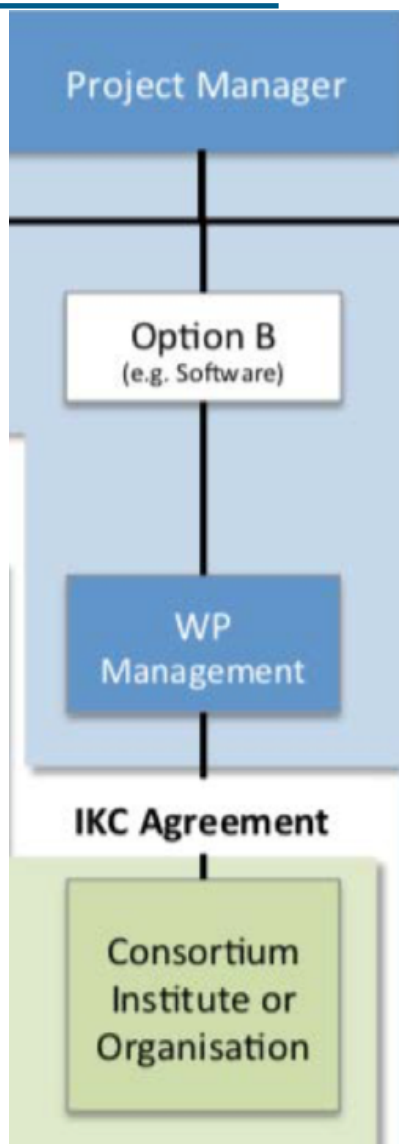


Will be updated

CTA Architecture 2018

CTA Software

- Responsibility for the **construction** of all software is with CTAO:
  - Management and Coordination
  - Architecture, Design, Specification
  - Definition of software Standards
  - Quality Assurance, Release, Testing, Integration, Acceptance
  - Coordination of implementation
    - via IKC, contracts with company, in-house implementation
- Strong involvement from contributors needed for implementation of software products
  - including detailed design, quality assurance, testing, integration



Bulk Archive Requirements 1

- No public access, only CTAO staff
- Preserve bulk raw data and higher-level data products (DL0-DL3) over the lifetime of CTA
- Preserve associated metadata and provenance information
- Preserve DL0 simulation data for at least 3 years after production
- Located at least at two sites with 300 km distance
- Handle increasing data volume of at least 6 PB/yr
- Allow fast (re-)processing of data (annual reprocessing of all data within 1 month)
- No data loss over the full lifetime of CTA
- Bulk Archive will follow OAIS standard (or reference model or reference architecture)
  - Note: Bulk Archive is not planned to be publicly open, so it follows OAIS as much as possible

Bulk Archive Requirements 2

- Storage-related requirements
  - Access rights management supporting specific roles for CTAO staff (Archive Manager, Data Processing Manager)
  - Unique identifiers for all data products that are independent of the storage location or number of copies
  - Versioning of data products
  - Placement, replacement, duplication, migration of data products and metadata
  - Bulk Archive validation and preservation of archive organization
- Ingest/Access-related requirements
  - Metadata extraction and browsing (separate DB for faster queries?)
  - Update and regeneration of metadata from data products
  - Confirmation of the availability of requested data products and estimation of the retrieval time < 1s (on average)

Science Archive Requirements

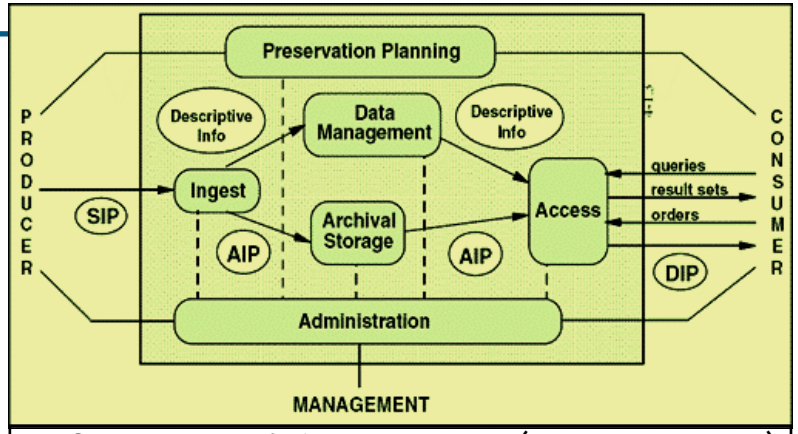
- Public access with data rights management (proprietary period)
- Preserve science data products > 10 yr beyond the lifetime of CTA
- Preserve associated metadata and provenance information
- Versioning of data products (unique identifiers) and software
- Flexible queries for users
- Fast access (product searches within 1 min)
- Support automatic data processing (DL3→DL5) – verification
- Support data browsing and interactive exploration
- Support standard interfaces (VO compliance)
- Highly available, high quality of service and products

Further considered:

- User-contributed part of the science archive for end-users
- Link of data products identifiers with publications and usage

Archives – Status

- Requirements and use-cases and operating model
- Compliance with Open Archival Information System (OAIS) – organization of people and systems



Reference Model *for* an OAIS (ISO 14721:2012)

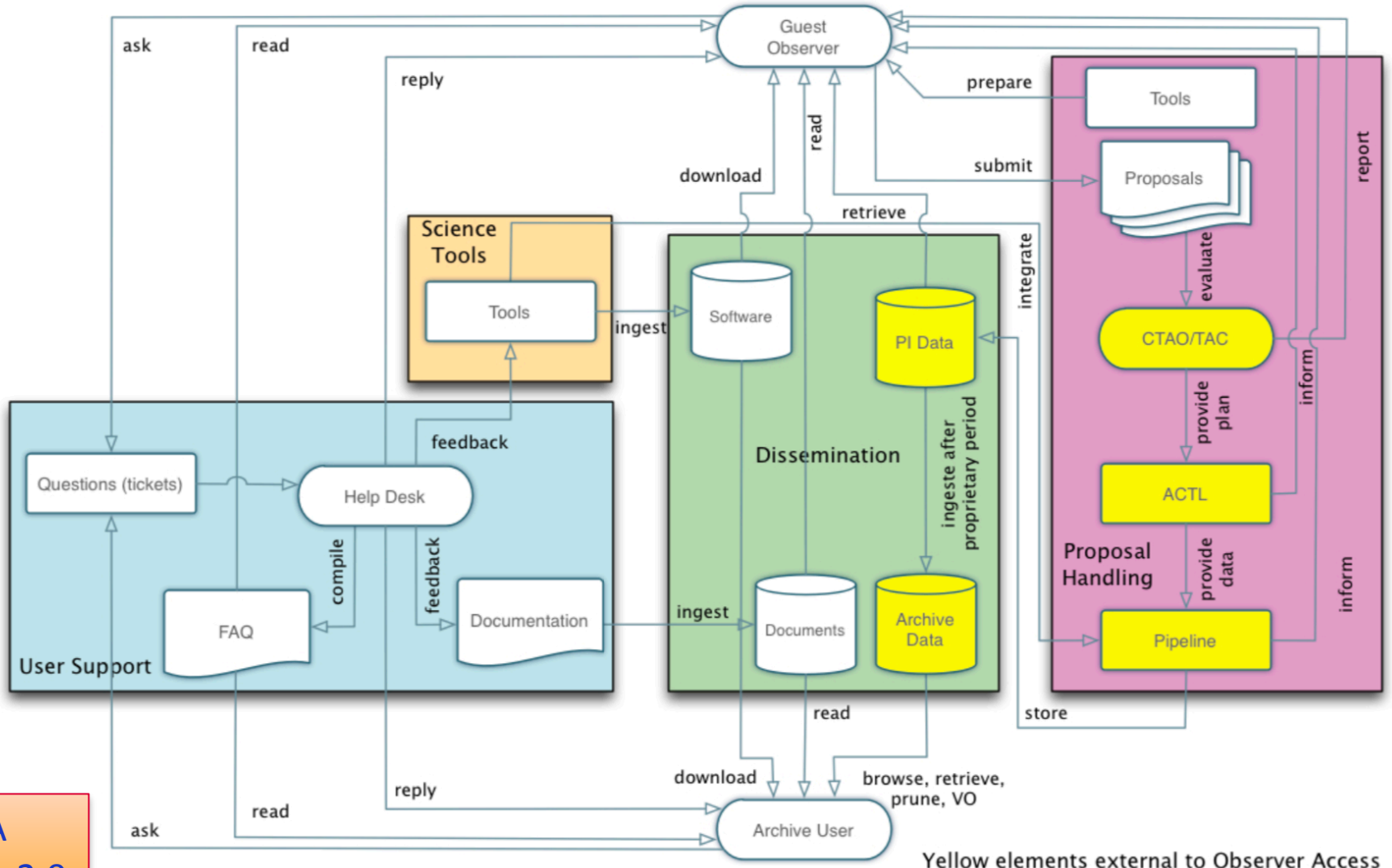
- CTA archive prototypes exist (some of them for both Bulk and Science Archive) – discussed at the CTAC meeting
  - INAF/ SSDC: INDIGO-DataCloud (H2020) based on OneData
  - LAPP: eXtreme DataCloud (H2020) based on OneData+dcache
  - UniGe/ETH: GAMAS - python-based development
  - LUPM: Data management using DIRAC

- CTAO Computing Department will organize Archive Workshop

Science User Access

- Major interface between CTAO and the scientific user community
  - Services and tools that are needed by Guest Observers and Archive Users to perform a successful scientific analysis
- User Support
- Dissemination of software to access high-level data (Science Data Access), documentation, and software
- Science Tools – the software to derive images, spectra and light curves (DL4) from processed CTA high-level event data (DL3)
- Proposal Handling

Science User Access



DATA
TDR v2.0
2016

Interoperability and Archives

In the time of CTA, SKA, LSST…

- Interoperability of the observatories is key to enable the astronomers to generate the best science
  - Interactions with science community, data products, observatory operations (e.g. scheduling, transients handling/science alerts)
  - Our archives will support interoperability wrt access to archives (and archived data of many observatories)
- CTA contributes to standards definition and develop prototype projects to improve interoperability
  - Member institutes actively involved in IVOA working groups
  - CTA fully supports VO standards and tools – data model and metadata for the data products
  - In the ASTERICS H2020 program: operations, data products, …
  - In the ESCAPE H2020 program: build a prototype for the European Open Science Cloud

Archives and Interoperability

- CTA Science Archive will support the VO standards and tools
  - CTA data will follow FAIR principles
    - Findable, Accessible, Interoperable, Re-usable
  - Configuration and Provenance data model in IVOA
  - Data findable via VO registry and VO tools
  - Expose data via (machine-readable) interfaces
- One step further:
  - Integration of Science Archive + Science Gateway + Interactive Data Exploration and Processing to a Science Platform
- Interests/ explorations beyond archives with community/ in H2020:
  - Sharing of observation schedules for coordinated campaigns
  - Common proposal handling tools
  - Setup of multi-observatory federation for A&A
  - Best practices for data processing and science workflows

The Computing Task

Major challenges ahead of us to get ready on two CTA-sites and SDMC

- Requirements, use-cases, architecture, plans, Computing Model, data formats…
- Transparent planning between the Computing Department of CTAO, in-kind contributors from the CTA consortium and industry
- From software prototypes to roll-out of quality software
- Milestones: first telescope acceptance in 2020 and Early Science in 2022

Get ICT and software ready for telescopes and instruments to deliver CTA data