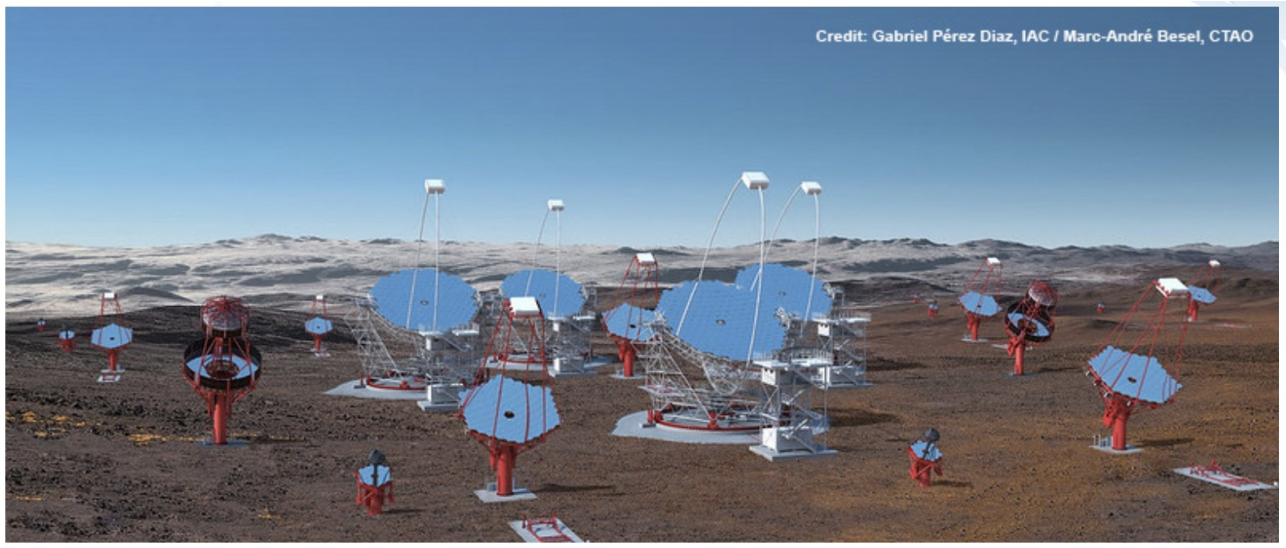
Rucio framework in the Bulk Data Management System for the CTA Archive

Update status on design and prototype(s)

5th Rucio Community Workshop, November 10-11, 2022
Lancaster University, UK

Georgios Zacharis, Stefano Gallozzi, Fabrizio Lucarelli
INAF – OAR

The Cherenkov Telescope Array Project



- Two sites (*CTA North* and *CTA South*) taking data with four data centers in Europe
- O(10TB) of raw reduced data required to be transferred ‘daily’
- O(50PB)/year from both sites to be archived
- Data must to be duplicated at the European DCs and removed at origin
- Processed up to science data
- Stored long-term on tape and periodically reprocessed

BDMS TEAM

- **INAF**: Stefano Gallozzi (Team Leader), Fabrizio Lucarelli (Deputy), Georgios Zacharis (Developer)

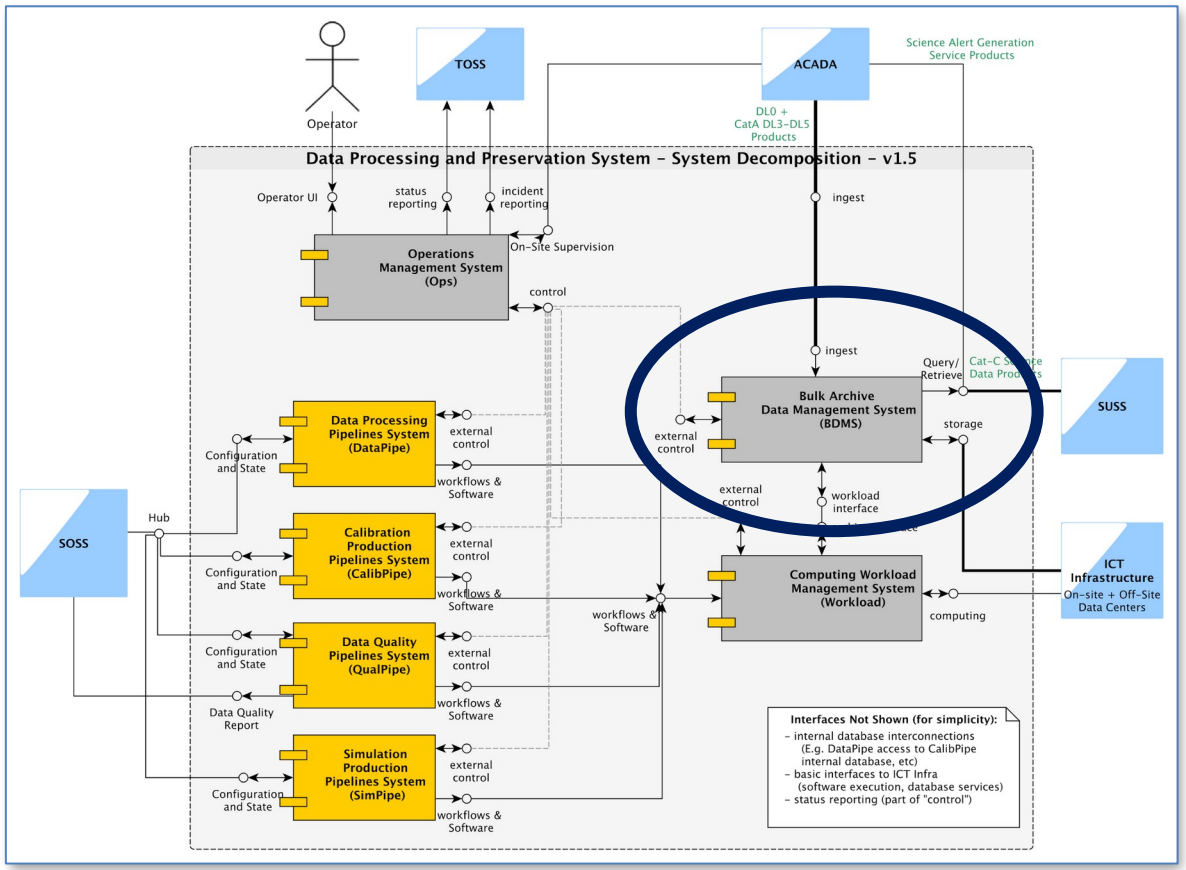
- **Swiss Contribution**:

  Roland Walter, Etienne Lyard (University of Geneva, ISDC)

  Syed Hasan (ETH Zurich)

  Volodymyr Savchenko (EPFL, SwissDC)

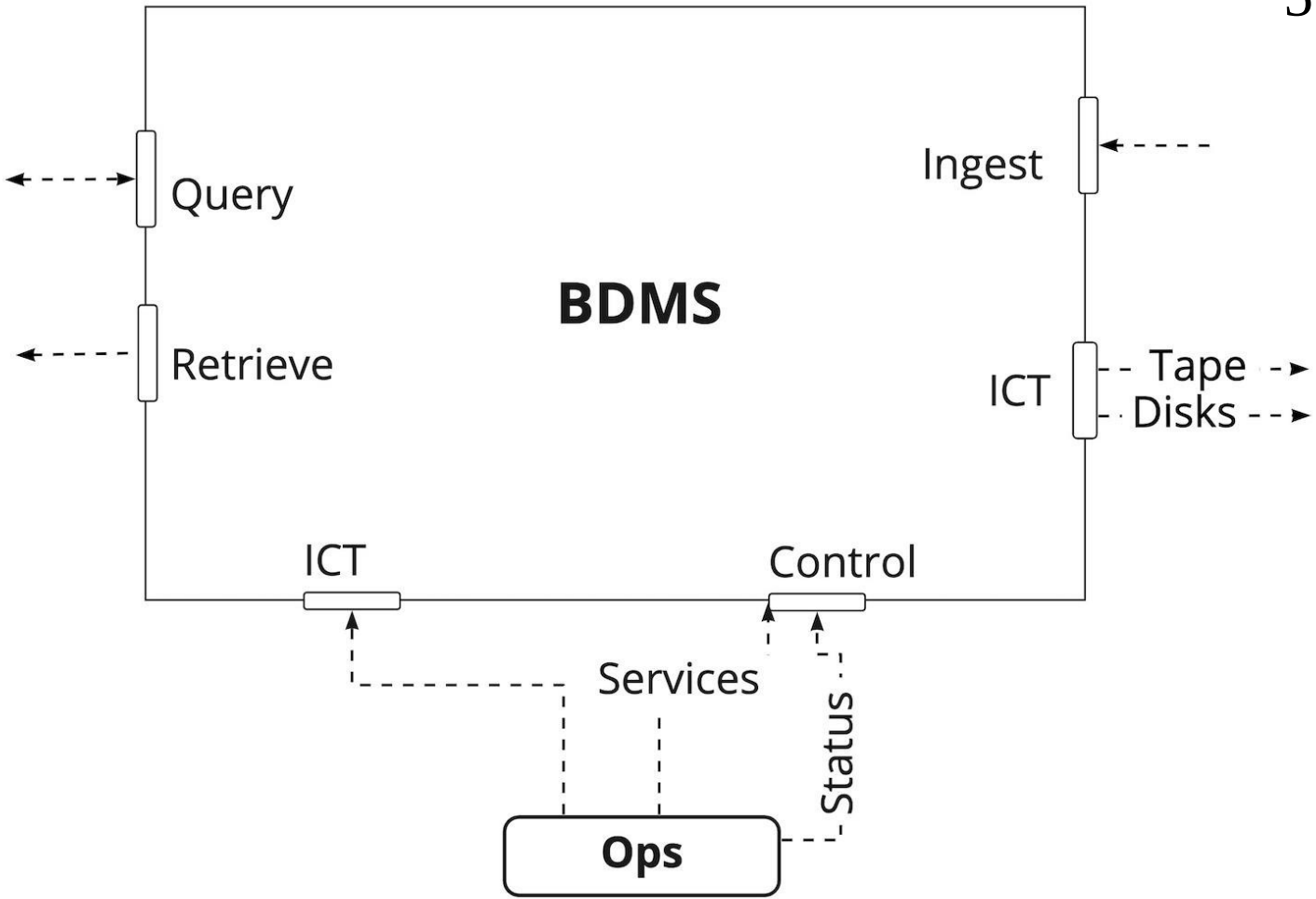
DPPS – System Decomposition (v2.5)



Bulk Data Management System (BDMS) (from DPPS MGT plan v2.0):

Provides a software system that manages the movement, replication, and thus preservation of data at a distributed set of storage elements located at DPPNs, both on- and off-site, and ensures the availability of data products for (re)processing by the ***Workload*** *system…*

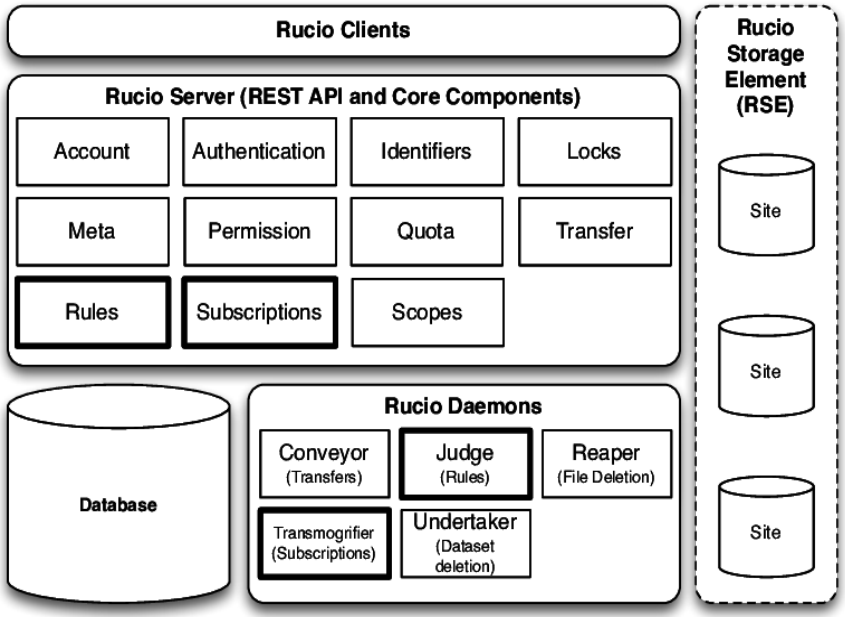
BULK DATA MANAGEMENT SYSTEM



Rucio for Bulk archive data management

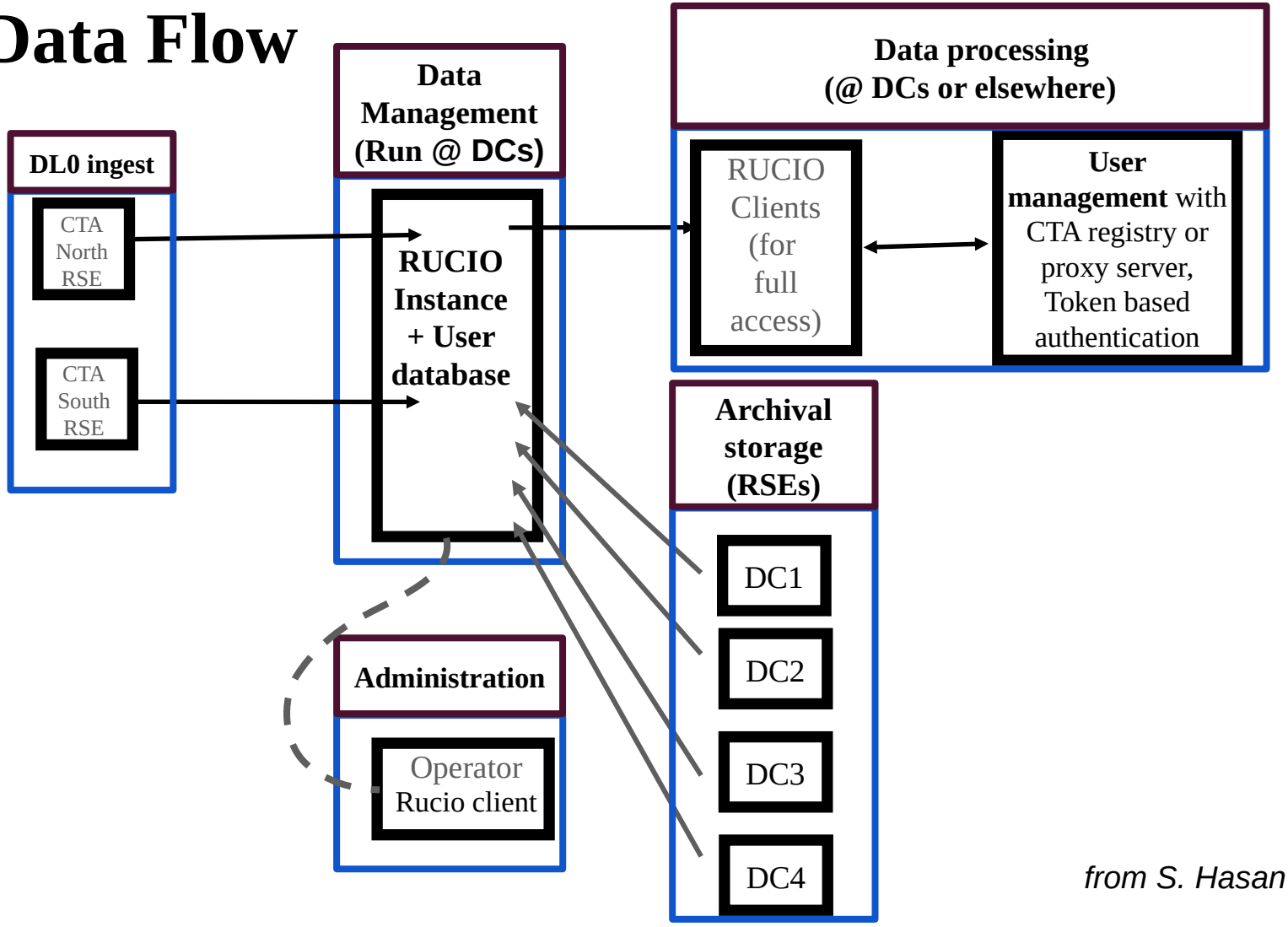
Rucio provides declarative engine for distributed archive management

- General and replicated catalog of **global files names – data identifiers (DID)** associated to **file copies on specific storages** (dCache, XrootD, EOS, etc)
  - grouping of files in datasets
  - domain (“physics”) metadata
  - specific datacenter file locations (unless location is computed
  - from the global name)
- User accounts and read/write rights (easy interface to retrieve and access data)
- Collection of file **transfer rules** describing which kind of storages should contain which files and with how many replicas



Rucio has been identified as the best tool to federate storage and manage sites in the CTA Bulk-Archive; can be used and adapted for all its functionalities?

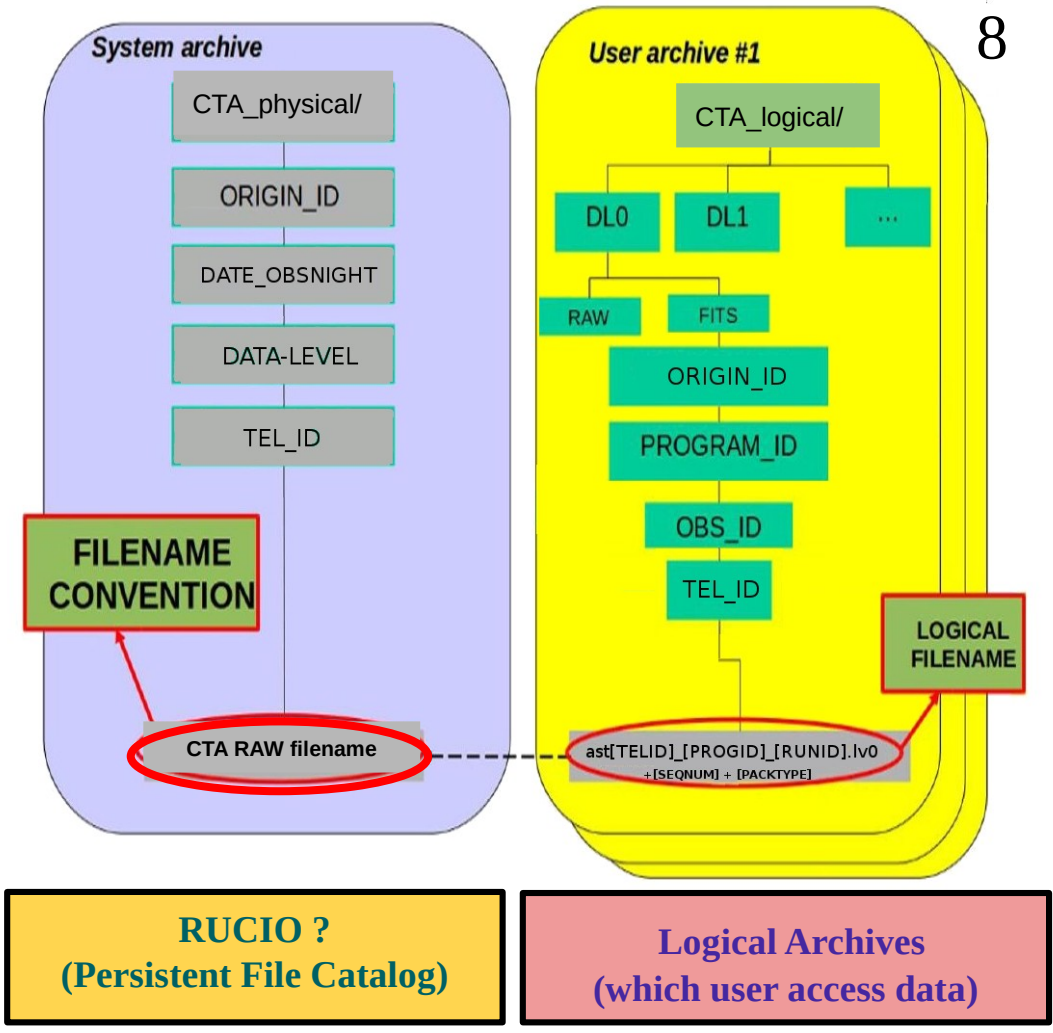
CTA Bulk Data Flow



from S. Hasan

BDMS (preliminary) DESIGN

- One Physical Archive
(structured as CTA datamodel(s))
+ many Logical Archives
i.e. depending on which data
the archive user needs to
access (use-case & user-stories)



→ User Archives identifies Logical Archives:
A single *physical* system archive and a few logical *user* archives

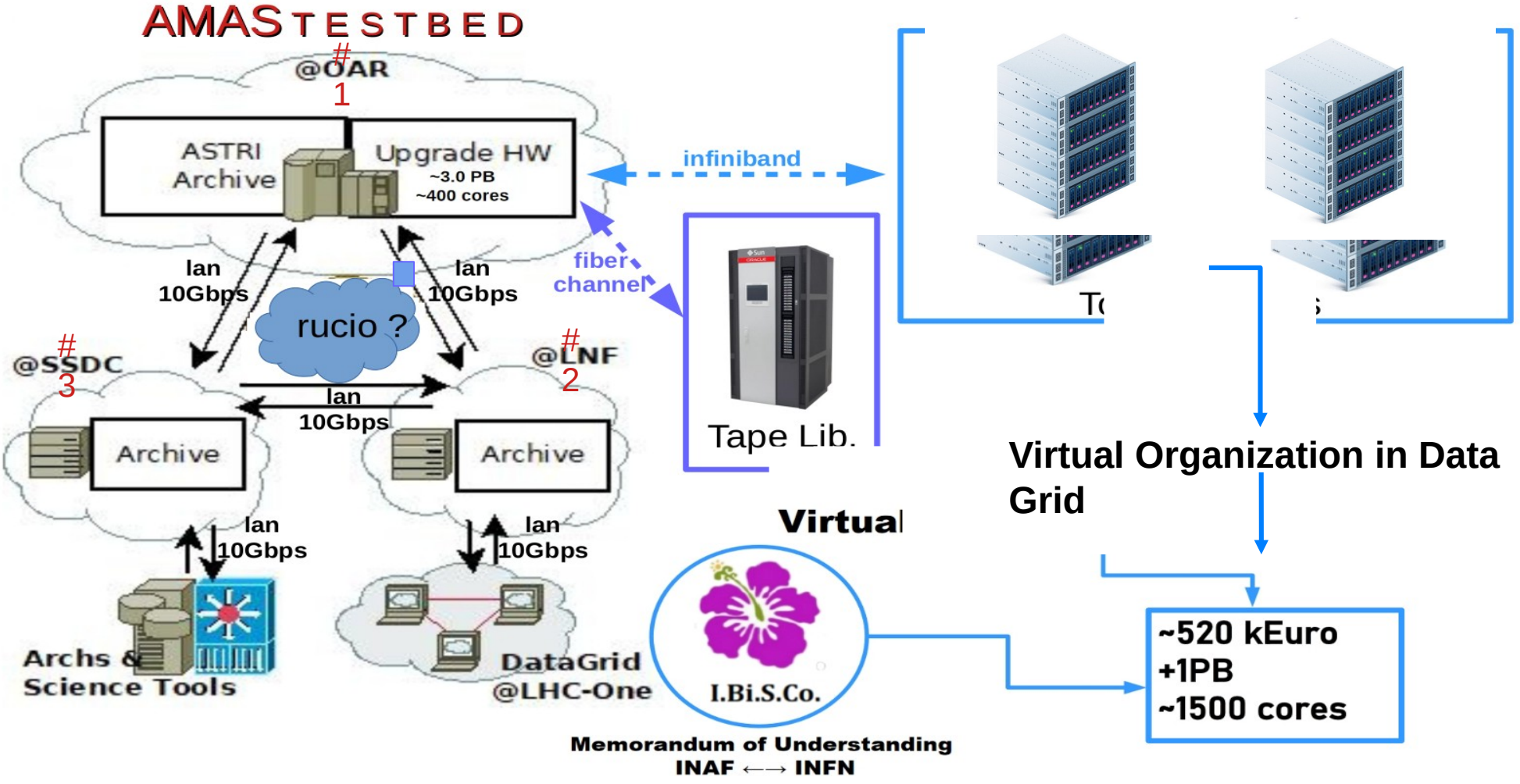
PROTOTYPING ACTIVITIES

- Testbed @Rome for ASTRI Mini-Array Data Archive System (SST data)
- Testbed @CSCS in Lugano for Large Size Telescope (LST data)

BDMS for CTA

BDMS prototype implementation – Test set-up

- Kubernetes K8s instance for installing RUCIO (and its services)
- dedicated FTS
- dCache


AMAS TESTBED
1
@OAR
ASTRI Archive
Upgrade HW
~3.0 PB
~400 cores
lan 10Gbps
lan 10Gbps
rucio ?
infiniband
fiber channel
Tape Lib.
3
@SSDC
2
@LNF
lan 10Gbps
Archive
Archive
lan 10Gbps
lan 10Gbps
Archs & Science Tools
DataGrid @LHC-One
Virtual
I.Bi.S.Co.
Memorandum of Understanding
INAF ⟷ INFN
Virtual Organization in Data Grid
~520 kEuro
+1PB
~1500 cores

dCache installation at CSCS and its set-up as Rucio RSE (*from S. Hasan*)

dedicated FTS deployed in CSCS k8s
https://github.com/cta-epfl/helm-charts/tree/master/charts/fts

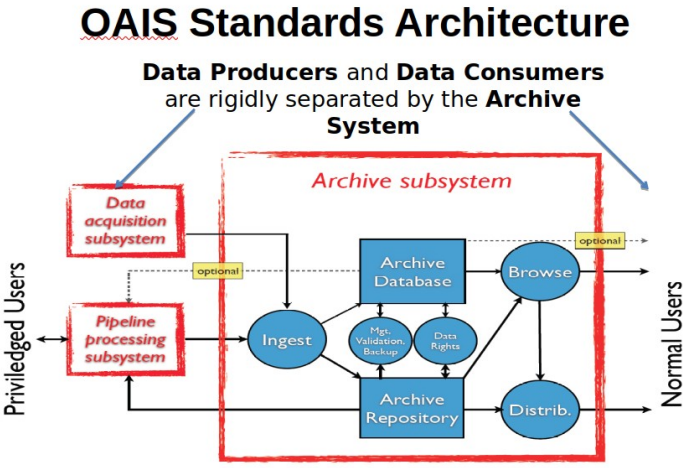
- **CTA dcache** service with **0.5Pb, primarily used for CTA Monte Carlo simulations**
  - **RSE name:** CTA-DC-CSCS
  - webdav (port 2880), Xrootd (port 1094)
  - deterministic: True; disk
  - Attributes - fts: https://fts:8446/
  - Protocols: https
    - domains: LAN, WAN, **TPC** (read, write, delete)
    - hostname: **dcache.cta.cscs.ch**
    - impl: rucio.rse.protocols.webdav.Default
    - root path or prefix: **/pnfs/cta.cscs.ch/dteam/bulk-archive/dc-cscs**
    - rse_id: 2c39b68d0c6747708217dabc11eecf89

```
rucio-admin rse add CTA-DC-CSCS || true
rucio-admin -v rse \
        add-protocol \
        CTA-DC-CSCS \
        --hostname dcache.cta.cscs.ch \
        --port 2880 \
        --scheme https \
        --prefix "/pnfs/cta.cscs.ch/dteam/bulk-archive/dc-cscs"  \
        --impl rucio.rse.protocols.webdav.Default \
        --domain-json '{"wan": {"read": 1, "write": 1, "delete": 1, "third_party_copy": 1}, "lan": {"read": 1, "write": 1, "delete": 1}}'

rucio-admin account set-limits root CTA-DC-CSCS 1073741824

rucio-admin rse update-distance --distance 10 --ranking 1 CTA-SITE CTA-DC-CSCS
rucio-admin rse update-distance --distance 10 --ranking 1 CTA-DC-CSCS CTA-SITE

rucio-admin rse update-distance --distance 1 --ranking 1 CTA-ECOGIA CTA-DC-CSCS
rucio-admin rse update-distance --distance 1 --ranking 1 CTA-DC-CSCS CTA-ECOGIA
```

- An example showing how the set-up and access is done

```
$ gfal-ls -l https://dcache.cta.cscs.ch:2880/pnfs/cta.cscs.ch/dteam/bulk-archive/site/ctaarc/05/78/cta1Mb-00003
-rwxrwxrwx   0 0     0       1048576 Oct  3 12:52 https://dcache.cta.cscs.ch:2880/pnfs/cta.cscs.ch/dteam/bulk-archive/site/ctaarc/05/78/cta1Mb-00003
```

BDMS Prototypes: Next Steps



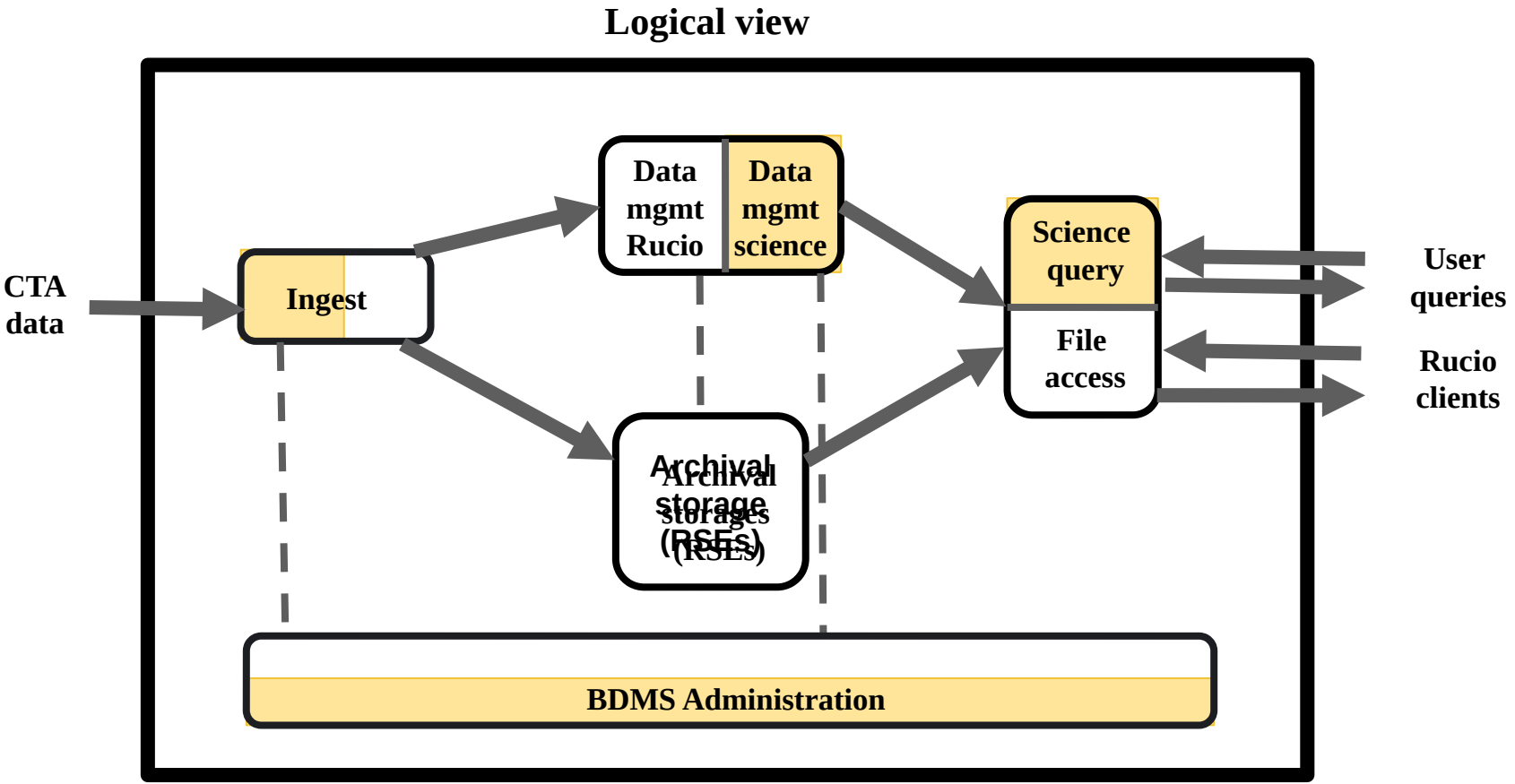
- Implementing and testing the replication (and file deletion) through RUCIO between two or more RSEs (data distribution from OAIS paradigm)
- Commit RUCIO catalog as physical storage for federated archive in BDMS: ingestion module from OAIS paradigm to be adapted to the data model and related metadata extracted and ingested in DBs (to be queried as logical data)
- Link logical archives view to physical RUCIO organization to BDMS (browsing module from OAIS paradigm as archive user interface)
- Working on interfaces implementation to/from BDMS to find dataset needed by “any” archive user (based on user reqs)
- Develop customized CTA policy package for CTA user access (based on A&A and to data rights & ownership)
- Create a FULL BDMS virtual environment for the whole system using K8s (or any other) (as deliverable for mini-DPPS release?)

BACKUP SLIDES

User authentication and transfers with RUCIO

- Support X.509 certificate based authentication with Virtual organization (VO) management proxy
- Rucio daemons need X.509 certificate support to enable it to access with Rucio storage elements (RSEs) and make transfers
- Rucio supports Token-based support for user authentication and works well with the RSEs that supports tokens
- Rucio also works with Identity providers like IAM (which is also based on Tokens) - ESCAPE project demonstrated RUCIO capabilities with IAM
- For the CTA project for Bulk archive access, X.509 certificates could be used. Tokens based support with Token issuers or even CTA's own Token service could be used
- There is also possibility to make RUCIO as the token issuing authority with OIDC based token workflows (according to RUCIO team)

BDMS architecture with RUCIO



Experimenting Rucio set-up with RSEs at different geographies (IP networks)

- RSE at UNIGE:
  - RSE name: CTA-ECOGIA RSE
  - https://www.isdc.unige.ch/~savchenk/cta/
  - Protocols: **ssh** and **http(s)**; two protocols used because https here is read-only (apache server) and to check that the same physical file location on an RSE can be accessed with different RSE protocols

- Rucio upload

```
$ rucio upload --scope ctaarc --rse CTA-ECOGIA cta1Mb-00005 --protocol ssh
2022-10-05 22:51:25,319 INFO    Preparing upload for file cta1Mb-00005
2022-10-05 22:51:25,508 INFO    Successfully added replica in Rucio catalogue at CTA-ECOGIA
2022-10-05 22:51:25,724 INFO    Successfully added replication rule at CTA-ECOGIA
2022-10-05 22:51:37,081 INFO    Trying upload with ssh to CTA-ECOGIA
2022-10-05 22:51:49,740 INFO    Successful upload of temporary file. ssh://login02.astro.unige.ch:22/www/people/savchenk/public_html/cta/ctaarc/70/42/cta1Mb-00005.rucio.upload
2022-10-05 22:52:16,063 INFO    Successfully uploaded file cta1Mb-00005
$ rucio list-file-replicas ctaarc:cta1Mb-00005  --protocol ssh,http
+---------+--------------+------------+-----------+-----------------------------------------------------------------------------------------------------------------+
| SCOPE   | NAME         | FILESIZE   | ADLER32   | RSE: REPLICA                                                                                                    |
|---------+--------------+------------+-----------+-----------------------------------------------------------------------------------------------------------------|
| ctaarc  | cta1Mb-00005 | 1.049 MB   | 4fea810a  | CTA-ECOGIA: http://www.isdc.unige.ch:80/~savchenk/cta/ctaarc/70/42/cta1Mb-00005                                 |
| ctaarc  | cta1Mb-00005 | 1.049 MB   | 4fea810a  | CTA-ECOGIA: ssh://login02.astro.unige.ch:22/www/people/savchenk/public_html/cta/ctaarc/70/42/cta1Mb-00005       |
+---------+--------------+------------+-----------+-----------------------------------------------------------------------------------------------------------------+
```

ASTRI prototype

- detects Cherenkov radiation using compact, sensitive and
- extremely fast silicon (SiPM) sensors
- located at INAF observative site OACT in Serra La Nave
- (Etna – Sicily)
- Mini-Array (located in Tenerife) project is based on ASTRI
- prototype
- AMAS: ASTRI Mini-Array Archiving System

AMAS Expected Data

Considering:
Packet dim. **13.052kB** & **9 telescopes**

In the **Worst Case**: **1.0 kHz** trigger rate, **11hr** acquisition/dd

In the **Average Case**: **150kHz** trigger rate, **8hr** acquisition/dd

- Optimal **HD Space** →
- (hot+MC [yearly]) ~ **0.75 PB**

- Optimal **Tape Space** →
- (cold + hot+MC [yearly])
- ~ **1.15 PB**
- **[+1.2PB per SI3]**

Archive/DB Units	GB/day MAX	GB/day AVG	TOT MAX (AVG) [TB/yr]
Bulk Archive (only RAW)	5117[2]	558[3]	604 (91)[4]
Bulk Archive (DL0 FITS + pipe products)[5]	15680	1710	1853 (278)
Science Archive	250	200	~35 (~25)
Swap-tmp Loc.Repo[6]			200
Simulation Archive (MC)[7]			100
Quality Archive	33	24	3.9 (3.9)
Log / Monitor / Alarm Archive	54	27	~20 (~10)
System Configuration DB	5	4	~0.6 (~0.5)
CALDB			0.2-0.4 (TBD)
Performance DB			0.5-1.0 (TBD)
Interferometry Instrument (SI3)			**1200 (??)**
hot-storage TOTALS:	**5405**	**786**	**~640 (~120)**
cold-storage Backup	**~16000**	**1737**	**~1873 (~290)**