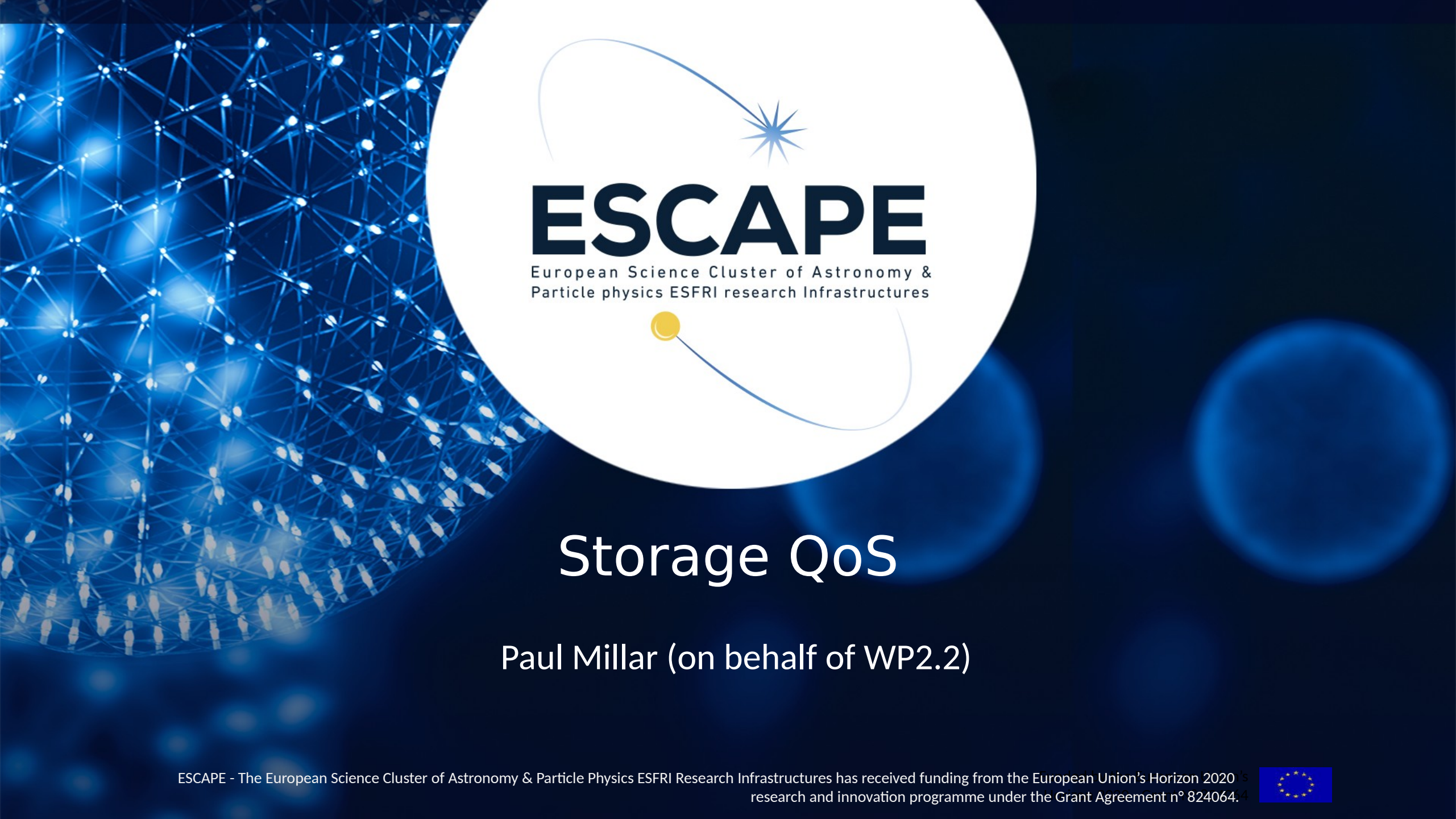


Storage QoS

Paul Millar (on behalf of WP2.2)

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

How expensive is recreating lost data?

- Data from detectors: **very expensive**.
  - Researcher would need to reapply for beam time.
  - (Likely) need to prepare fresh samples.
  - (Likely) travel back to the facility.
  - Overall, considerable expense and loss of research time.
- Output from simple file processing: **almost no cost**.
  - No need to book time at the facility, no need to prepare new samples, no travelling.
  - Recovery would be very quick, no significant loss of research time.
- Data-loss can **very different costs**, depending on the data.
- Making storage robust against data-loss is expensive

  Why do it for data where durability doesn't matter so much?

Funded by the European Union's Horizon 2020 - Grant N° 824064

How “fast” does data access need to be?

- Data being processed by HPC/GPGPU farm: **very fast**.
  - HPC / GPGPU resources are expensive – limited resources.
  - Want to maximise use – Do not want HPC/GPU waiting for data.
- Long-term data storage: **not fast at all**.
  - Data where all expected processing has taken place.
  - Data stored to fulfil FAIR requirements.
  - Can afford for it to take time to read data.
- Data has a range of **different performance requirements**.
- Making storage high performance is **expensive**.

  Why store data on high-performance capacity when the data doesn’t need it?

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

How to build different storage options?

- “Disk” → “Disk” + TAPE
- “Disk” → Different options from …

  Media layer: PMR / NVMe-SSD / SMR / [HAMR, MAMR, SSD-low-endurance, …] / TAPE

  Media compositing layer: JBOD/RAID-0, RAID-5, RAID-6, RAID-Z, RAID-Z1 …

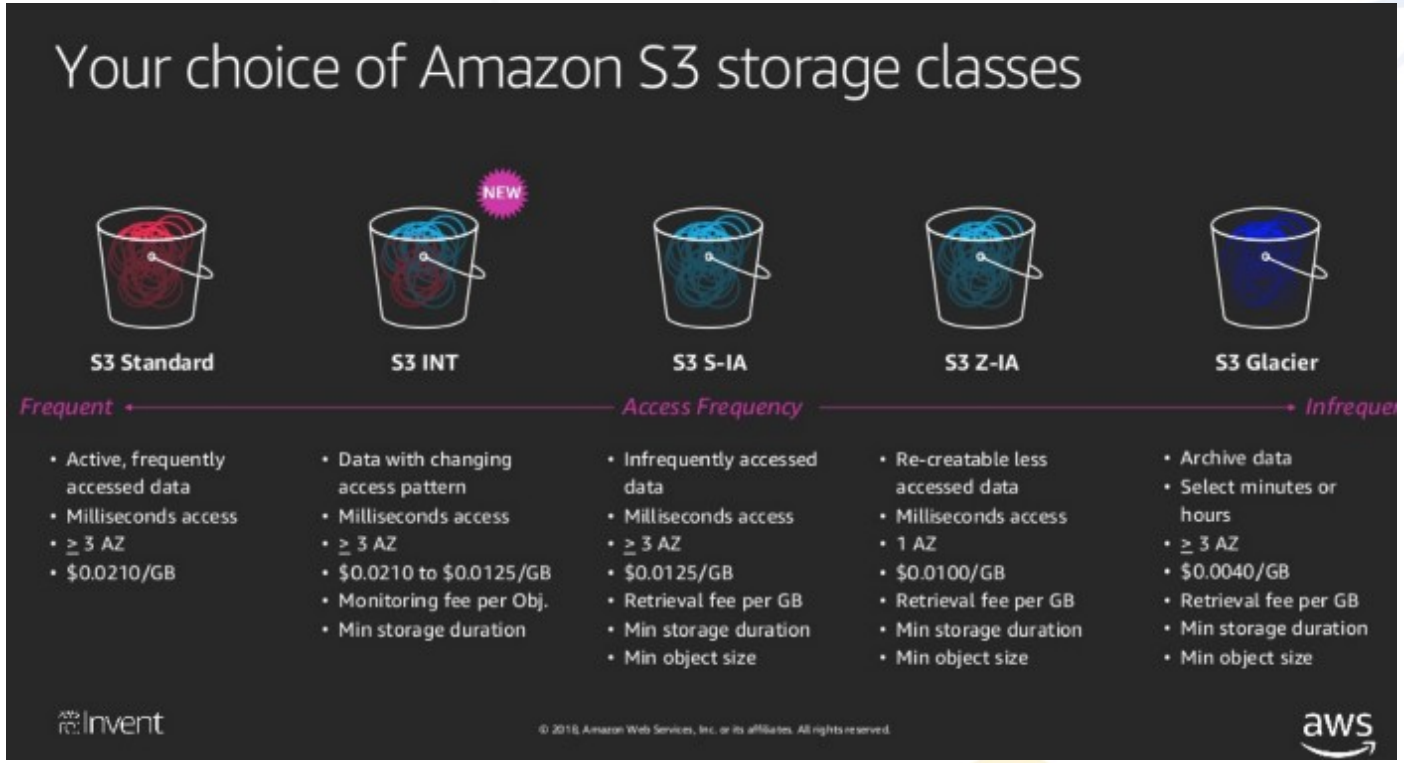
  Intra-cluster compositing layer: Single-copy, Replication (n), Erasure code (n,m)

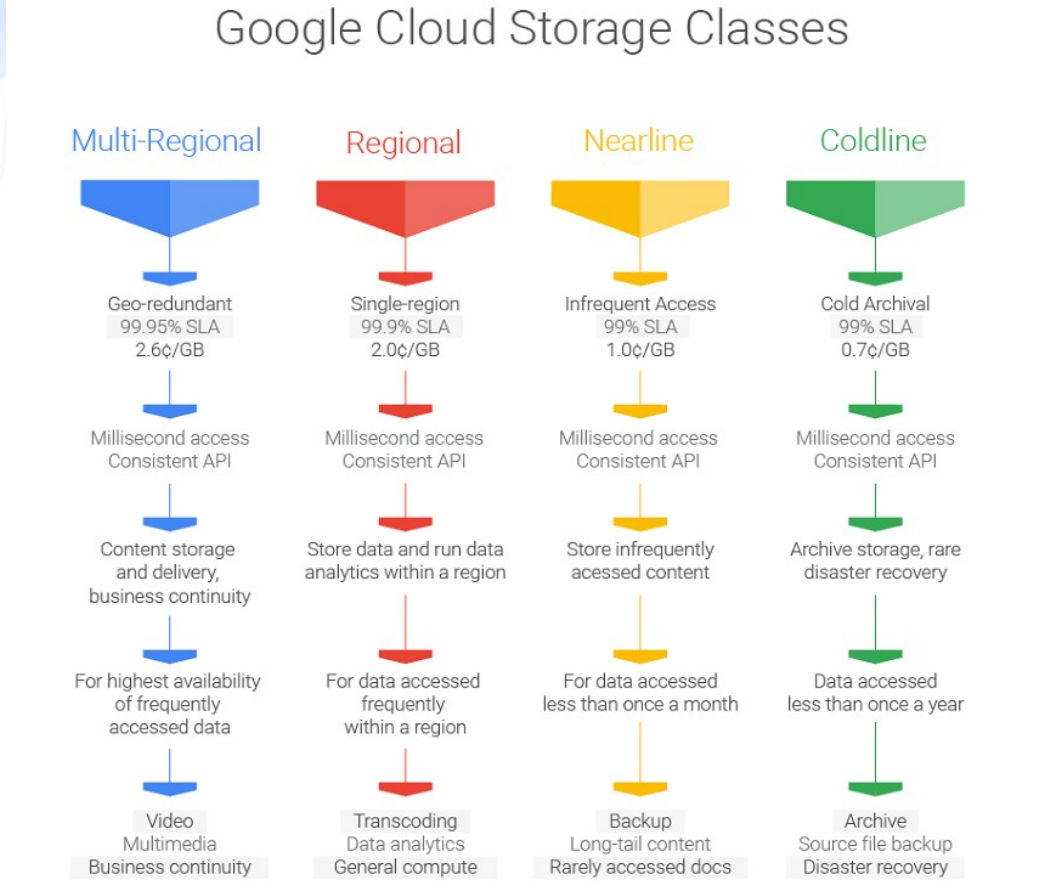
  Inter-storage compositing layer: How many replicas? … which which characteristics?

- As a researcher you **don’t want to see** this complexity!
- Abstraction is a common solution: **hide the details**
  - Instead, you say what kind of storage you want in general terms, and the system uses the most reasonable available option.
- This abstration is called **(storage) QoS**.

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

Storage QoS is what industry is doing





Funded by the European Union's Horizon 2020 - Grant N° 824064

Why use QoS?

- **Save money** – limited budget.
  - At some point “just buy more hardware” doesn’t work.
  - Get more done with the available budget.
- Provide tailored storage behaviour.
  - **Faster storage**, when processing data.
  - **More durable storage**, where storing important data.
- Support possible **future storage technology**.
  - You don’t know what changes tomorrow will bring. How will you handle changes needed when new storage types become available?

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

How to make use of QoS

- Data is **not homogenous**.
  - Some data has low(er)-value; for example, some data may be recreated if lost.
  - Some data is not used in time-critical analysis (e.g., log files).
- Data use **changes over time**.
  - Data fresh from camera/detector is of interest; month-old or year-old data is less interesting.
  - Only use expensive storage for embargoed data; public data is stored on cheaper (and slower) storage.
  - Pre-emptively move data to hot storage for thematic analysis runs.
- Different users/work-flows have **different needs**.
  - Identify “power users” (or work-flows) and store their data on more capable hardware.

… potentially more options! …

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

QoS in ESCAPE

- Building up **knowledge** within scientific communities.
  - Storage QoS is a new concept; new ideas are hard. It requires new ways of thinking about storage.
  - In ESCAPE, we are working on an iterative approach to building up case studies on how scientific communties can benefit from QoS.
- Working with **technology providers**.
  - Require some new concepts at the data management layer.
  - Storage technologies can be improved to make QoS transitions “faster” or “better”.
  - In ESCAPE, we are working closely with the Rucio development team to deploy and test new features, and with the storage providers to allow testing of these concepts.
- Proving **this really works**, in practice.
  - Use the ESCAPE testbed to gain (as close to) real-world experience as possible

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

Closing remarks

- This probably all sounds quite complicated, but it isn't really.
  - The job is to **hide complexity**, to make managing storage managable.
- QoS is **baked-in** to the DataLake concept.
  - You don't need to add anything to start using it.
- QoS is an **optional feature**.
  - You don't need to use QoS if you want to use the Data Lake concept.

Funded by the European Union's
Horizon 2020 - Grant N° 824064

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

… and now for something different

Funded by the European Union's Horizon 2020 - Grant N° 824064

Managed transfers without grid storage

- Data Lake concept (to a large extent) come from **WLCG and "The Grid"**
  - Grid storage software (dCache, EOS, StoRM, …) support Data Lakes out-of-the-box.
- **Grid storage** software tends to provide "closed world" solutions.
  - All data access is through the software, no by-pass access.
  - It's just like how IBM Spectrum Scale storage is always accessed through Spectrum Scale servers.
- KIT and DESY are developing an **"open world" alternative**.
  - Funded by German government to support better data management within Germany.
  - Simple, open-source solution, based on Apache, you can deploy to allow access to your existing storage solution.
- Currently being tested within **ESCAPE testbed**.
  - Deployed within the ESCAPE testbed to gain (as close to) real-world experience.

Funded by the European Union's Horizon 2020 - Grant N° 824064

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Thanks for listening!

Funded by the European Union's Horizon 2020 - Grant N° 824064

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Backup slides

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Why are you using storage?

- To store my data (… well … yes, obviously)
- Some specific examples:

  “Scratch” – to store analysis output, which can be recreated if needed

  “Reliable” – to store important data where performance isn’t an overriding goal

  “High Performance” – for analysis in HPC /GPGPU cluster

  “Archive” – long-term storage for unknown/possible future analysis

  “Public” – non embargoed data that external users can access

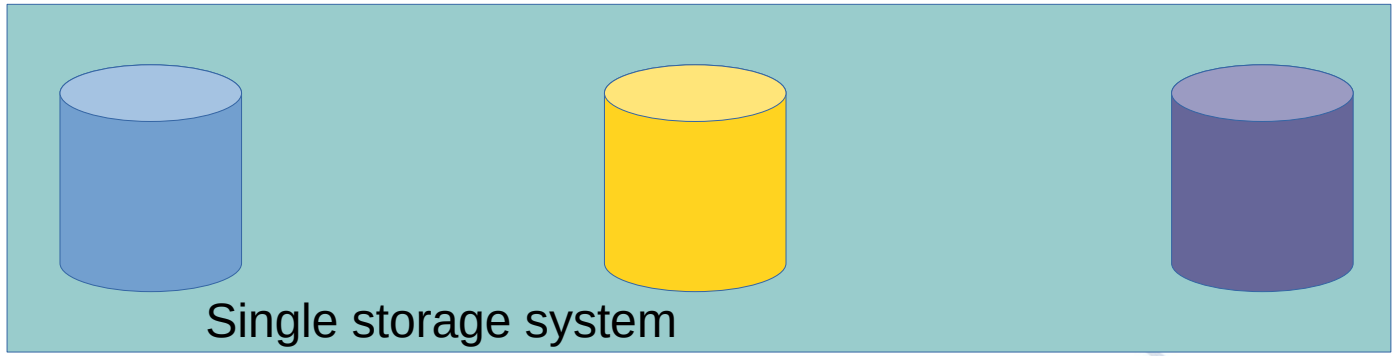
  “Ingest” – storage that is tuned to for high performance write operations

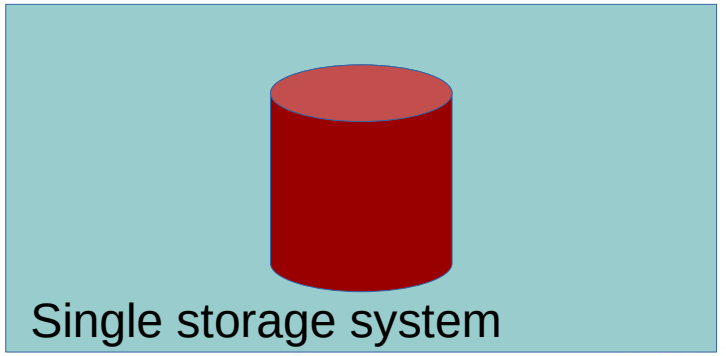
  … etc …

- A single storage that supports all use-cases will be expensive.
- Storage QoS (supporting specialised storage) allows you to cut costs.

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

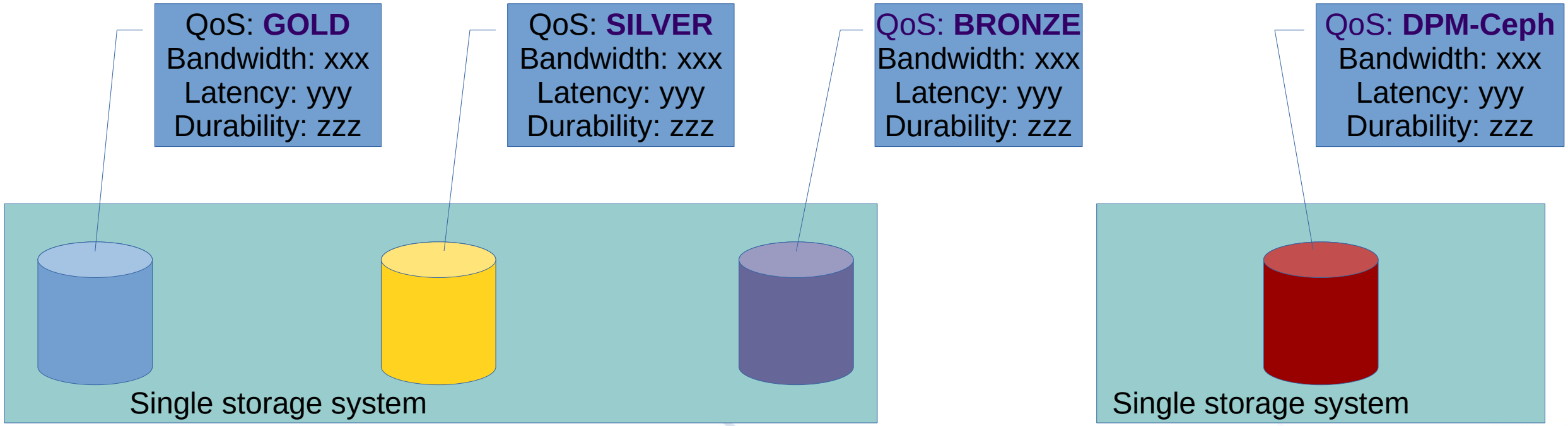
Concepts: Storage QoS classes





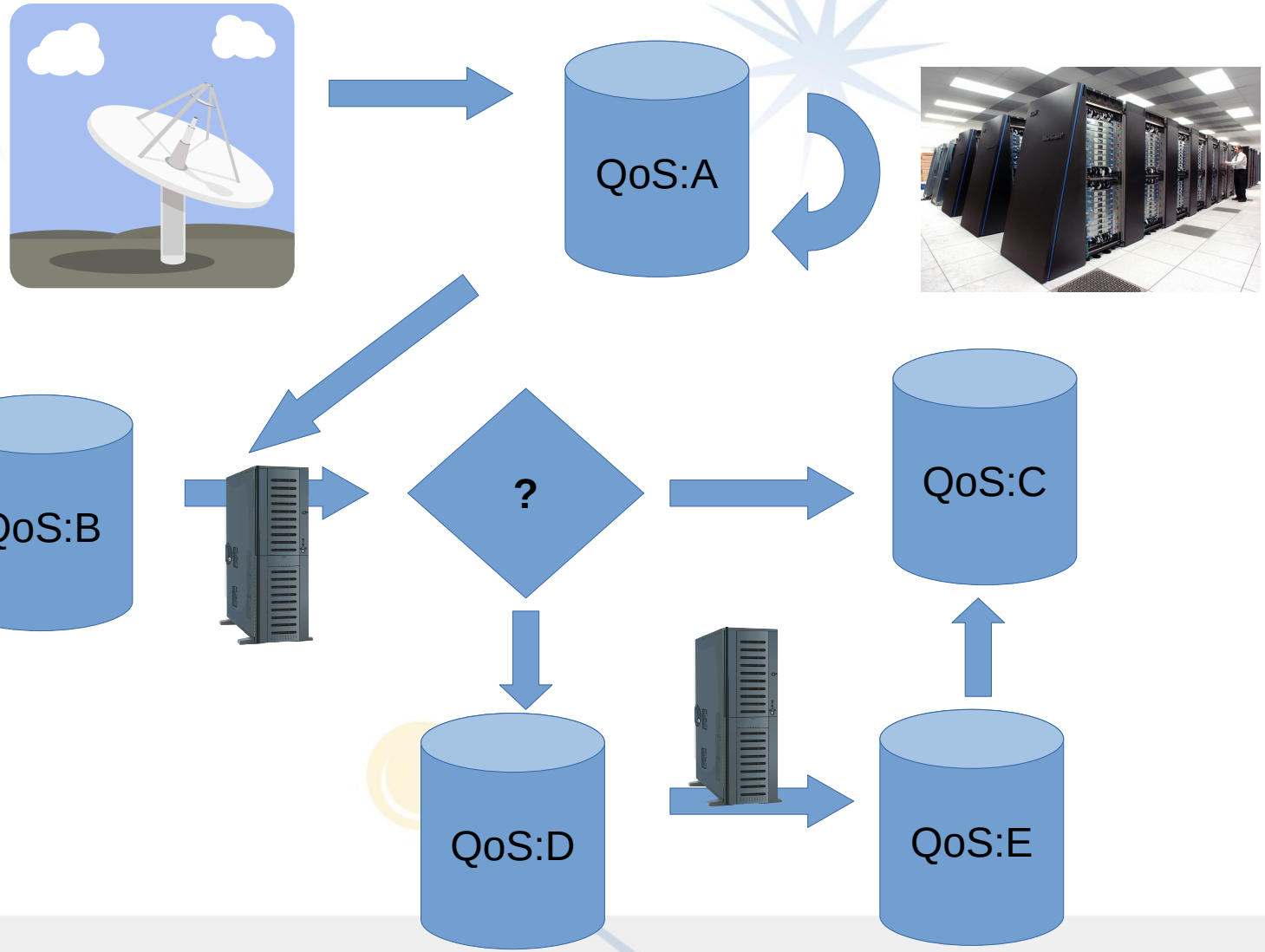
Funded by the European Union's
Horizon 2020 - Grant N° 824064

Concepts: Storage QoS classes



Funded by the European Union's Horizon 2020 - Grant N° 824064

Concepts: work-flow / data lifecycle



Funded by the European Union's Horizon 2020 - Grant N° 824064

Concepts: VO-QoS-Policies

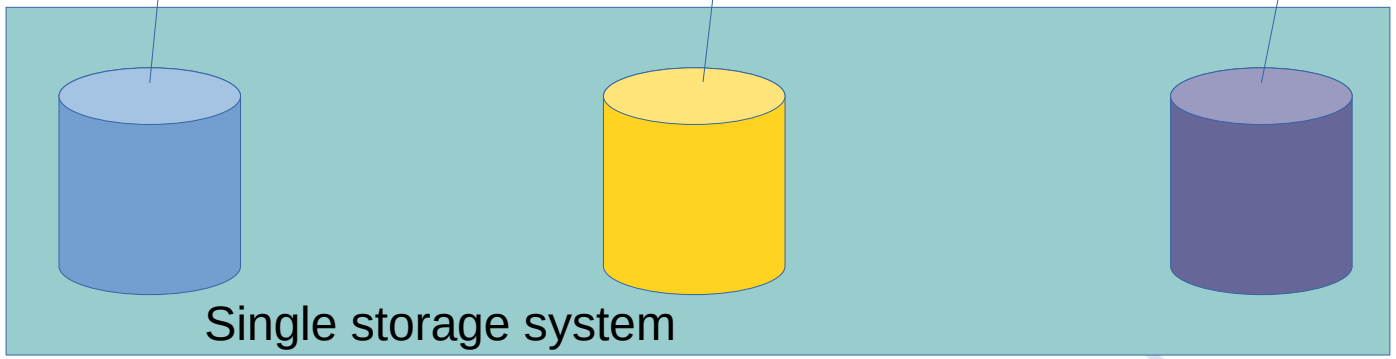
QoS:A | QoS:B | QoS:C | QoS:D | QoS:E

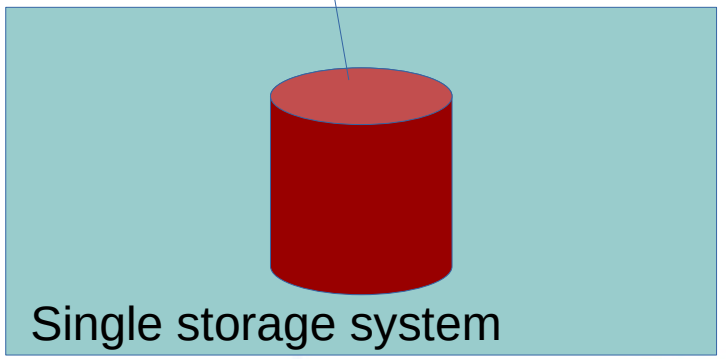
QoS: **GOLD**
Bandwidth: xxx
Latency: yyy
Durability: zzz

QoS: **SILVER**
Bandwidth: xxx
Latency: yyy
Durability: zzz

QoS: **BRONZE**
Bandwidth: xxx
Latency: yyy
Durability: zzz

QoS: **DPM-Ceph**
Bandwidth: xxx
Latency: yyy
Durability: zzz





Funded by the European Union's Horizon 2020 - Grant N° 824064

Concepts: VO-QoS-Policies

QoS:A | QoS:B | QoS:C | QoS:D | QoS:E

QoS: **GOLD**
Bandwidth: xxx
Latency: yyy
Durability: zzz

QoS: **SILVER**
Bandwidth: xxx
Latency: yyy
Durability: zzz

QoS: **BRONZE**
Bandwidth: xxx
Latency: yyy
Durability: zzz

QoS: **DPM-Ceph**
Bandwidth: xxx
Latency: yyy
Durability: zzz

Single storage system

Single storage system

Funded by the European Union's Horizon 2020 - Grant N° 824064

General approach in Task 2.2

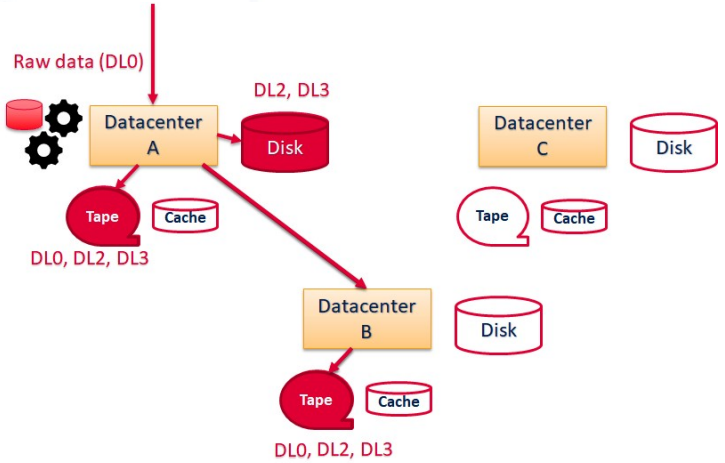
- **Capture** each ESFRI communities QoS requires.
  - Initial interview, to go through the document template, describing what are the different elements and what they mean.
  - The ESFRI representatives discuss the document with other scientists in their community.
    - This might trigger further, ad-hoc meetings to discuss details.
  - Once document is complete, it is presented at the T2.2 regular meetings.
- **Build** a coherent document that takes input from these documents
  - More than just chapters for each community.
  - An initial version, that will contain some open questions
- **Update** the document, based on further discussion.

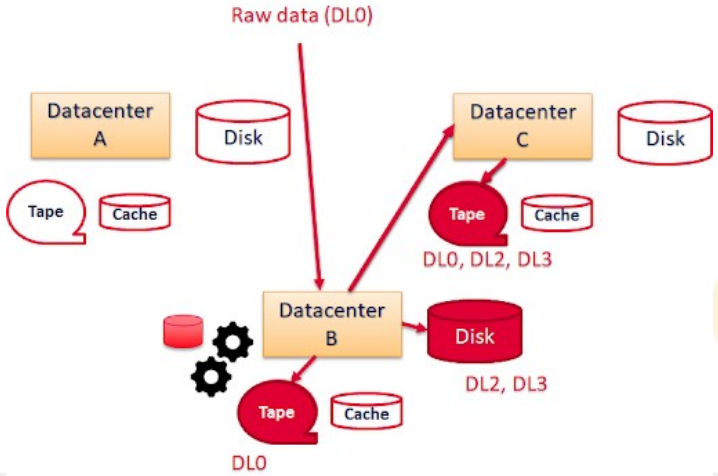
Funded by the European Union's Horizon 2020 - Grant N° 824064

Document format

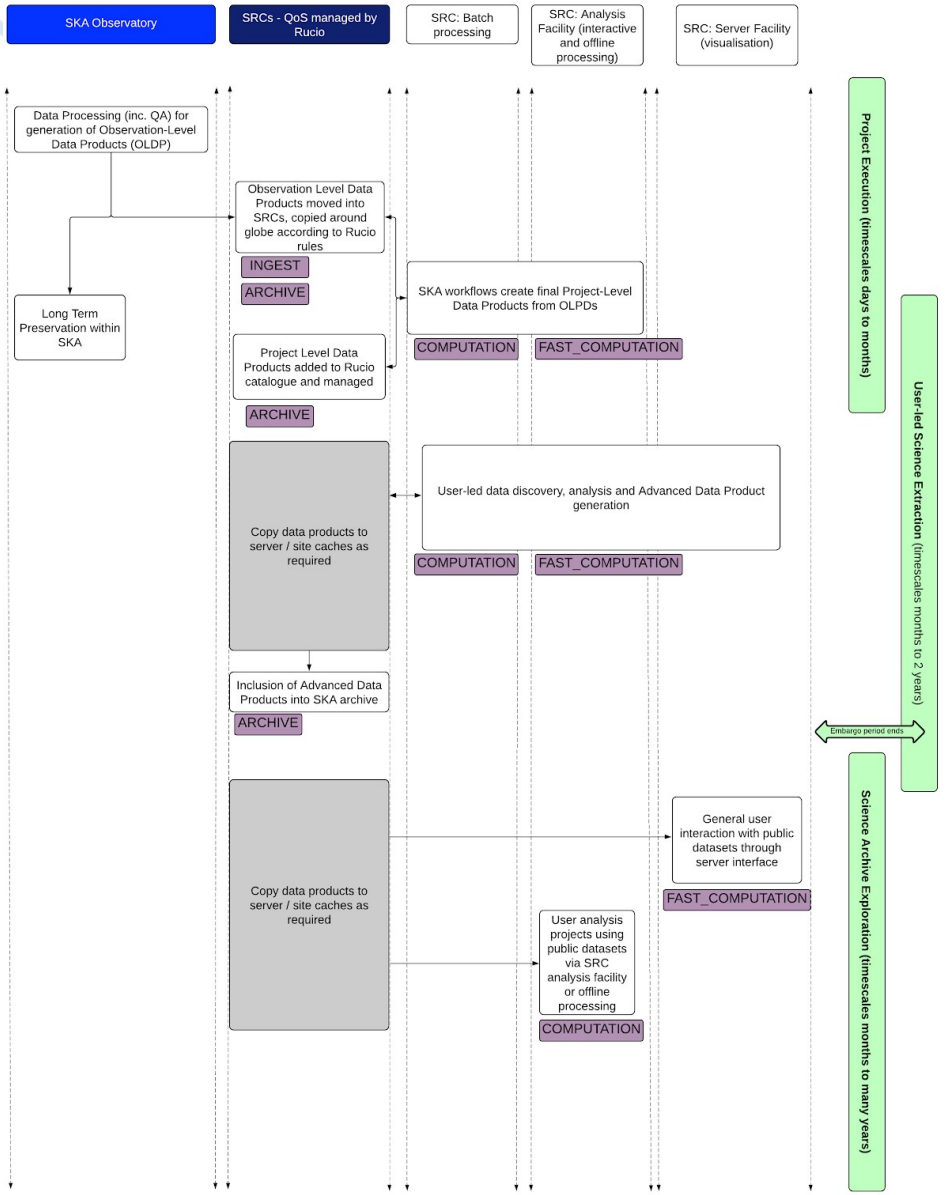
- The ESFRI-specific QoS document is split into four sections.
- The **QoS policies section** is a table, containing information about how the experiment sees QoS.
  - Name, Where it is used, Important characteristics, Example media.
- The **data life-cycle/work-flows section** is more free-form.
  - Describes the best guess at how data will be handled.
- The **interactions with Rucio** section describes how the experiment framework should achieve the desired QoS
  - List of operations to satisfy life-cycles
- The **use-cases section** provide terse description of user interactions.
  - Bring previous three items together.

Funded by the European Union's
Horizon 2020 - Grant N° 824064





CTA workflow (cred. Nadine Neyroud)



SKA data lifecycle (cred. Rohini Joshi)

Funded by the European Union's Horizon 2020 - Grant N° 824064

Current status

- **Received documents** from ATLAS, CTA and SKA
- Document from FAIR is **in preparation**.
  - Killian and Marek for FAIR,
- Next step is to build the **combined document**, providing a single reference for:
  - Driving testbed deployment.
  - Driving software development.
  - Testing QoS support in Rucio.
  - Stimulate further discussion within ESFRI communities.
- This document is (necessarily) an **initial version**:
  - Anticipate an updated version, based on feedback from ESFRI communities

Funded by the European Union's
Horizon 2020 - Grant N° 824064

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Storage endpoint deployment

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Testbed: the storage QoS classes

- Currently, we have **storage endpoints** from (in no particular order):
  - CERN, SurfSARA, CNAF, Napoli, Roma, IN2P3, PIC, LAPP, GSI and DESY.
- There is a **wiki page** that documented what storage technology is being used.
- From this, devised five **storage QoS classes**:
  - JBOD, RAID, EC1, EC2, TAPE.
- Five is a **balance** between
  - too few → difficult to make specific requirements; each class is too general.
  - too many → difficult to work with; each class is not general enough.

Funded by the European Union's
Horizon 2020 - Grant N° 824064

The “Aleem” QoS demo

- Answering the **question**: what can we do with what we have now?
- How it **works**:
  - In Rucio, label RSE endpoint, using the attribute **`QoS=<storage-class>`**
  - When creating rules, use this attribute to describe desired storage QoS
    - E.g., **`(QoS==JBOD || QoS== RAID) && Site=IN2P3`**
  - Rucio takes care to copy data to specific locations.
  - QoS transitions involve adding or removing rules.
- **Positive**: we can do this right now
- **Negative**: we lose the VO-QoS policy level abstraction
  - This makes the system fragile; changes become difficult.
- Demoed with LOFAR data, ESFRI representatives are replicating this.

Funded by the European Union’s
Horizon 2020 - Grant N° 824064

Next steps

- FAIR complete their QoS document.
- Complete 1st version of the combined ESCAPE QoS document.
- New version of **Rucio** is released:
  - Once deployed on the ESCAPE testbed, we can start experimenting with the new VO-QoS-policy support.
  - Need to create a mapping between VO QoS Policy and Storage QoS Classes (e.g., in CRIC)
  - Rucio needs to obtain this mapping
- Start **integrating** QoS work with computation workflows:
  - Somehow tag workflows with desired QoS
  - Rules are added as a precursor to running analysis jobs.

Funded by the European Union's Horizon 2020 - Grant N° 824064

Thanks due to …

- Thanks to everyone for their help so far.
- In particular the ESFRI representatives in their work building the QoS documentation:

  Mario, Frederic, Nadine, Rosie, Rohini, Kilian and Marek

- … and the sites for providing storage and QoS information
- … and Aris in helping establish the QoS attributes.

Funded by the European Union's
Horizon 2020 - Grant N° 824064