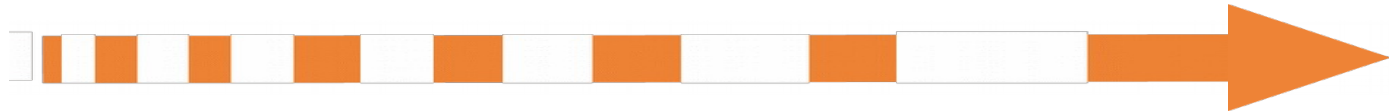


QoS Overview and White Paper



Data Management for extreme scale computing



Paul Millar
on behalf of the DOMA-QoS WG



What is DOMA?

- ✘ An acronym for “Data Organisation, Management and Access”
- ✘ A WLCG R&D project to develop new approaches to data, to be delivered in time for the **High-Luminosity LHC**.
 - currently scheduled for 2027
- ✘ Three main (“fat”) working groups: Access, Third-Party Copy, and (storage) Quality of Service.
 - **DOMA-Access** involves the “last mile”, delivering data to the application,
 - **DOMA-TPC** involves moving data between sites,
 - **DOMA-QoS** involves data “at rest”, as it is stored at sites.

DOMA-QoS motivation

✘ The **funding** problem:

- ☛→ High Luminosity LHC is expected to produce vastly more data than previous runs (including RUN 3).
- ☛→ Funding is “flat”: no additional money will be available.
- ☛→ Advances in storage (e.g., improved areal density) will help, but not enough.

✘ New **technologies**:

- ☛→ Faster storage technology is available, but is expensive.
- ☛→ Can WLCG experiments use a small amount of “high performance” storage?

The challenge

- ✘ To bootstrap a new way of working with storage, we need:
 - ☛→ Something that **works for sites**: the storage people are deploying and the software they are using
 - ☛→ A **common framework** that works for a diverse set of use-cases from the experiments
 - ☛→ No “big bang” changes: must **co-exist** with the existing deployment
 - ☛→ Support an **evolving set** of use-cases.
- ✘ The important point is this only works with the cooperation of the sites and the experiments.

How to achieve this?

- ✘ Work with the sites to understand what storage options are available
 - ☛→ Conduct a site survey
 - ☛→ Continue the dialogue with a **QoS workshop** and dedicated DOMA-QoS meetings

- ✘ Work with experiments to understand where (in their work-flows) different QoS makes sense.
 - ☛→ Create a white paper that describes what QoS means
 - ☛→ Invite feedback: in a **QoS workshop** and dedicated DOMA-QoS meetings

Purpose of the workshop

✘ An opportunity...

- ...→ for **experiments** to exchange their QoS ideas.
- ...→ to learn from **sites**' QoS experiences.

✘ Start receiving **feedback** about the QoS white paper

Dialogue will continue in DOMA-QoS regular meetings and direct interviews.

✘ The major objectives are:

- ...→ First step in establishing a **consensus** on what should be pursued on the experiment side, and where WLCG-level coordination is desirable.
- ...→ Identification of **common themes** at the infrastructure level which could attract interested sites and be exposed to the experiments.

DOMA-QoS Output

Site survey results

- ✘ Detailed information available at DOMA-QoS twiki.

https://twiki.cern.ch/twiki/bin/view/LCG/QoS_SurveyAnswers

- ✘ Oliver presented an excellent summary at 2019-10-09 GDB:

https://indico.cern.ch/event/739883/contributions/3577297/attachments/1922942/3181621/QoS_Site_Survey.pdf

- ✘ Summary of the summary of the survey:

- ☛→ Some sites are investigating possibilities: **CEPH** is a favourite technology.
- ☛→ **No strong direction** for saving cost.
- ☛→ **Concerns** about how novel storage can be reconciled with WLCG pledges.

White paper: the process

- ✘ The white paper is a tool to support a **dialogue** with the experiments.
Defines specific terms, to speed up communication and avoid confusion.
- ✘ Is an **iterative** approach
 - ☛→ Currently at v1.0
 - ☛→ We are expecting feedback from experiments, related projects and sites.
 - ☛→ Use this feedback to update the white paper (v1.1, v1.2, ...).
- ✘ Once a **consensus** is reached, the final version of white paper is cut.
- ✘ More details are available from our wiki:
<https://twiki.cern.ch/twiki/bin/view/LCG/QoSWhitePaper>

White paper: v1.0

- ✘ An 18 page document that provides a description of QoS
 - Deliberately technology and implementation agnostic
- ✘ Suggests a model that allows:
 - Sites to adopt new storage technology
 - Experiments to include novel storage in their work-flows
 - No sudden, disruptive deployments
 - Allows storage to evolve, with changing technologies and software over time.
- ✘ Identifies some **open questions** and proposes **possible solutions**.

White paper ideas: files and replicas

- ✘ **Files** are a sequence of bytes with some metadata; a VO concept.
- ✘ **Replicas** are the bytes of a file stored at a specific site.
 - A file must have at least one replica, otherwise the data is lost.
- ✘ Each replica will have some **well-defined QoS**
- ✘ Files may have **multiple replicas** (potentially with different QoS).

White paper ideas: Storage QoS Class



✘ Different kinds of storage are abstracted as **Storage QoS Classes**.

⋯→ Homogeneous storage will have a single Storage QoS Class.

⋯→ Heterogeneous storage may have multiple Storage QoS Classes.

✘ DISK and TAPE are **examples** of Storage QoS Classes.

✘ Sites may provision **novel storage**, with a new Storage QoS Class.


The new storage QoS Class prevents accidental use.

✘ **Open question:** should Storage QoS Classes be composable.

Replica has Class:FOO and Class:BAR or only a list of supported classes, which could include Class:FOO_AND_BAR?

Example of Storage QoS Classes

Your choice of Amazon S3 storage classes

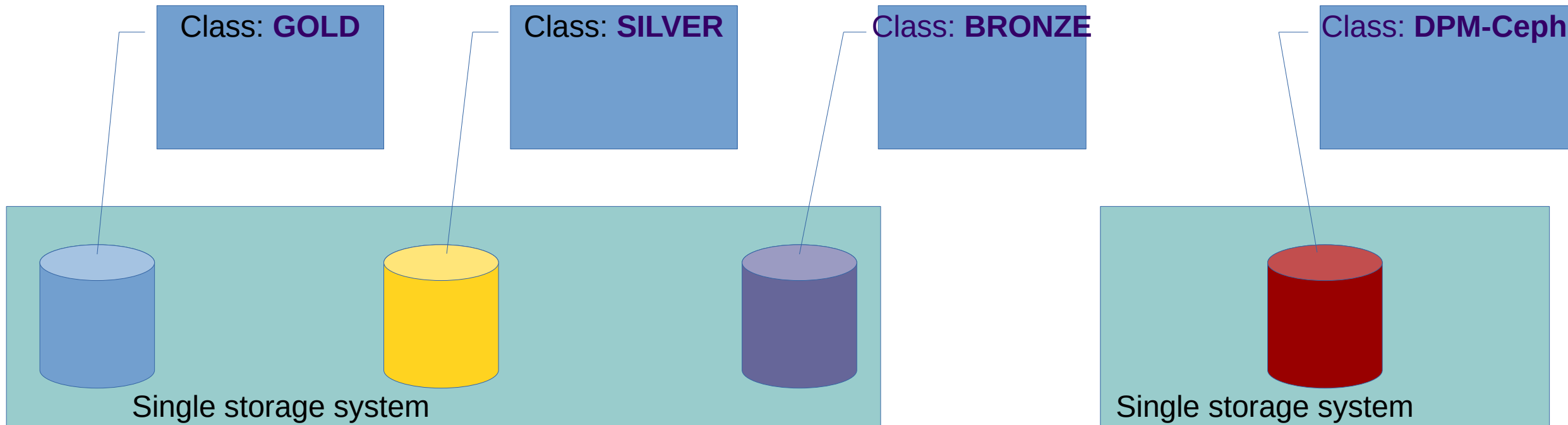


Frequent ← *Access Frequency* → *Infrequent*

S3 Standard	S3 INT	S3 S-IA	S3 Z-IA	S3 Glacier
<ul style="list-style-type: none">• Active, frequently accessed data• Milliseconds access• ≥ 3 AZ• \$0.0210/GB	<ul style="list-style-type: none">• Data with changing access pattern• Milliseconds access• ≥ 3 AZ• \$0.0210 to \$0.0125/GB• Monitoring fee per Obj.• Min storage duration	<ul style="list-style-type: none">• Infrequently accessed data• Milliseconds access• ≥ 3 AZ• \$0.0125/GB• Retrieval fee per GB• Min storage duration• Min object size	<ul style="list-style-type: none">• Re-creatable less accessed data• Milliseconds access• 1 AZ• \$0.0100/GB• Retrieval fee per GB• Min storage duration• Min object size	<ul style="list-style-type: none">• Archive data• Select minutes or hours• ≥ 3 AZ• \$0.0040/GB• Retrieval fee per GB• Min storage duration• Min object size

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

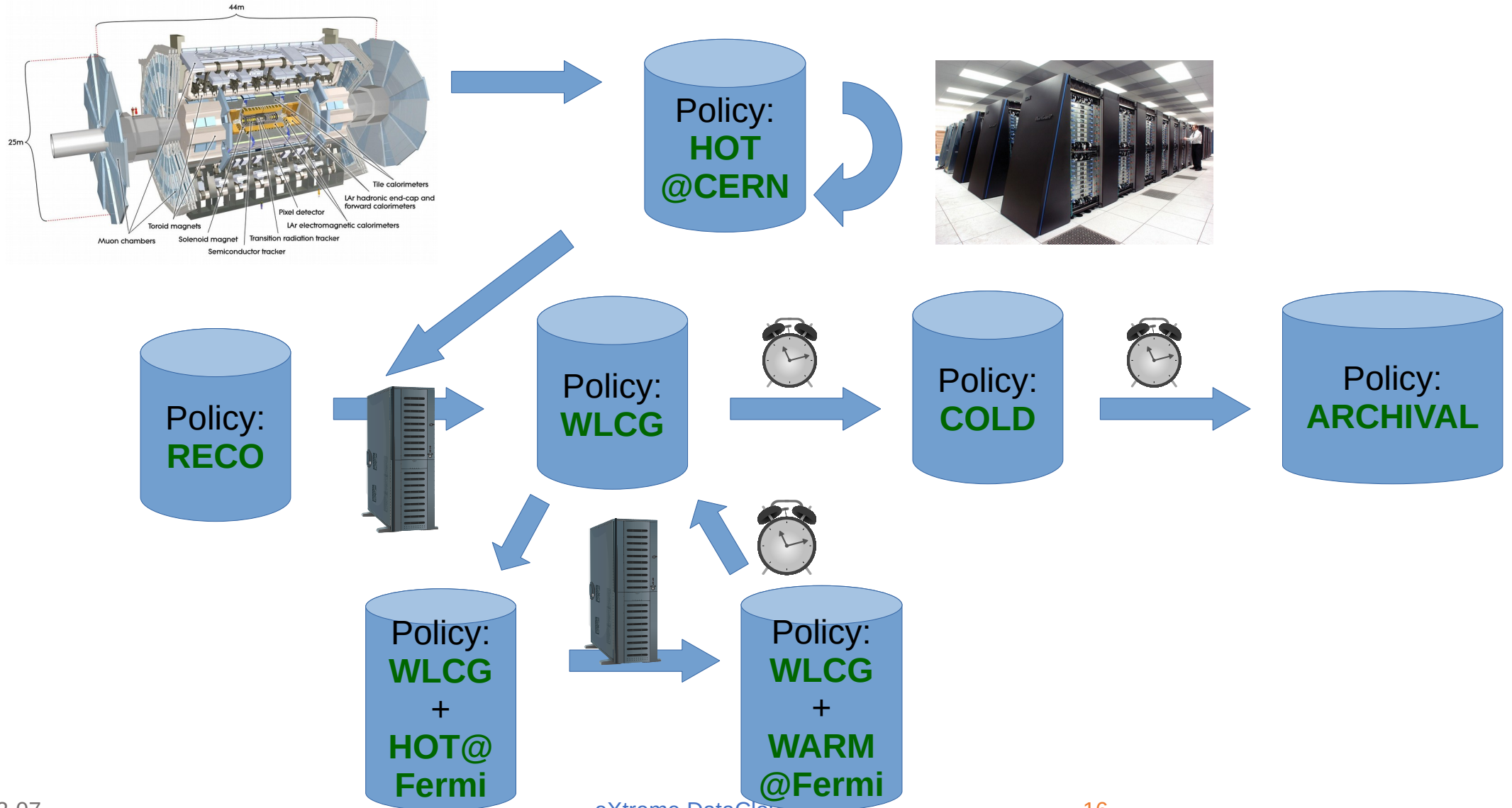
Storage QoS Classes



White paper ideas: VO QoS Policies

- ✘ At the VO level, data orchestration is achieved with **VO QoS Policies**.
- ✘ A VO QoS Policy describes which kind of replicas must exist:
 - For example, that a file is stored with two replicas, in different countries, on with Storage QoS Classes where the data is very unlikely to be lost, but performance may be slow.
- ✘ VO QoS Policies are **composable**: a file may have multiple policies, all of which must be satisfied.
- ✘ VO QoS Policies are (very likely) linked with **experiment work-flows**.
- ✘ The VO QoS Policies of a file will **change over time**, depending on what is happening with that file.

VO QoS Policy change over time



White paper ideas: policies → classes

✘ **Three models** considered for mapping VO QoS Policies to Storage QoS Classes:

- ☛ Storage QoS Classes **are well-define**, with well-defined names; VO QoS Policies use these names.
- ☛ Storage QoS Classes **metadata describes the policies** with which they are compatible.
- ☛ Storage QoS Classes **metadata contains attributes**; the VO QoS Policy is defined in terms of minimum (or maximum) allowed values.

VO QoS Policies: well defined classes



Policy: **WLCG**
1x**ROBUST**

Policy: **COLD**
2x**VOLATILE**

Policy: **HOT@CERN**
FAST@CERN

Policy: **HOT@BNL**
FAST@BNL

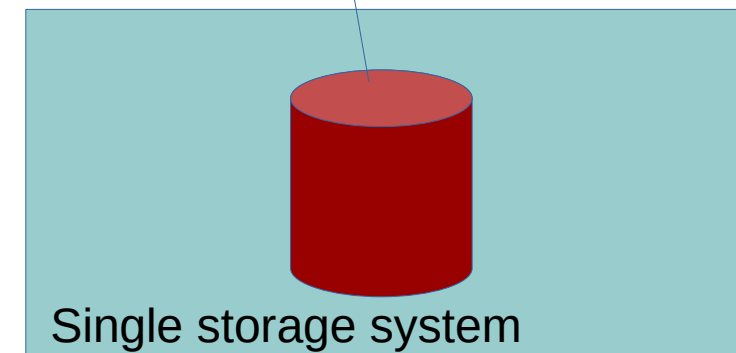
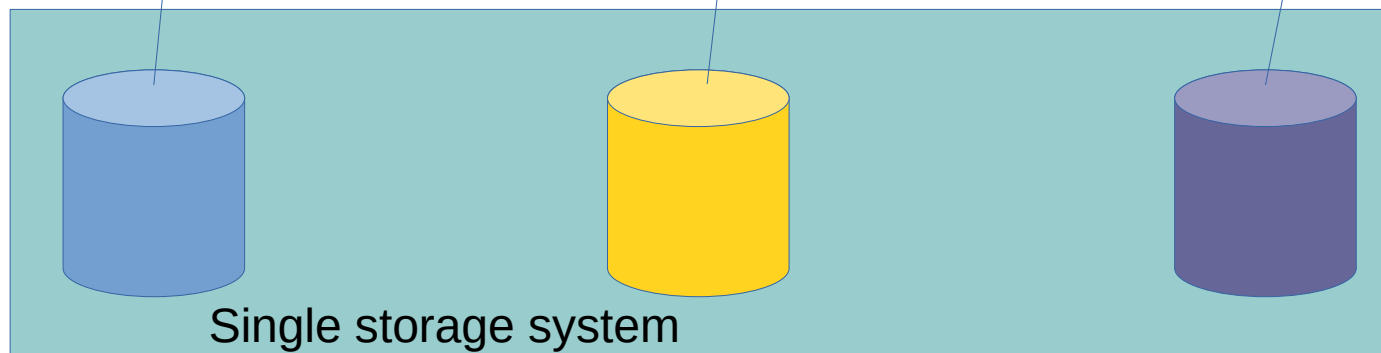
Policy: **HOT@DESY**
FAST@DESY

Class: **FAST**

Class: **ROBUST**

Class: **VOLATILE**

Class: **ROBUST**



VO QoS Policies: suggested usage



Policy: **WLCG**

Policy: **COLD**

Policy:
HOT@CERN

Policy:
HOT@BNL

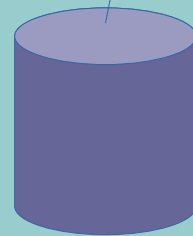
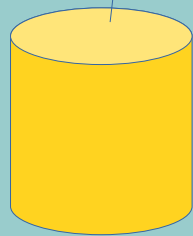
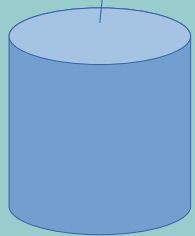
Policy:
HOT@DESY

Class: **GOLD**
Good for: **HOT**

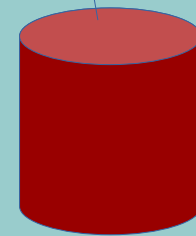
Class: **SILVER**
Good for: **WLCG**

Class: **BRONZE**
Good for: **COLD**

Class: **DPM-Ceph**
Good for: **WLCG**



Single storage system



Single storage system

VO QoS Policies: attributes



Policy: **WLCG**

Durability > Z1

Policy: **COLD**

Durability > Z2

Policy:
HOT@CERN

Bandwidth > N

Site = CERN

Policy:
HOT@BNL

Bandwidth > N

Site = BNL

Policy:
HOT@DESY

Bandwidth > N

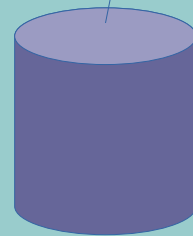
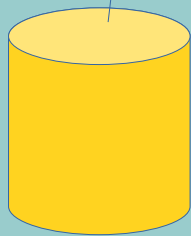
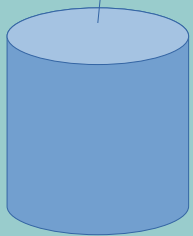
Site = DESY

Class: **GOLD**
Bandwidth: nnn
Latency: yyy
Durability: zzz

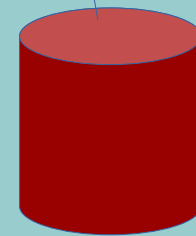
Class: **SILVER**
Bandwidth: nnn
Latency: yyy
Durability: zzz

Class: **BRONZE**
Bandwidth: nnn
Latency: yyy
Durability: zzz

Class: **DPM-Ceph**
Bandwidth: nnn
Latency: yyy
Durability: zzz



Single storage system



Single storage system

Summary

- ✘ First version of white paper exists:
 - ☛→ Defines a fairly complete model for how QoS could work.
 - ☛→ Identifies some open questions.
- ✘ We now need feedback from sites, related projects and experiments, to build a consensus.
 - We will update the white paper, based on this feedback.
- ✘ This workshop is an important step to build this consensus and identify common themes.

... just one more thing

- ✘ DOMA-QoS are contributing to the High Luminosity LHC Computing Review, by writing a chapter in the DOMA supplementary document.
 - One section in that chapter is on “related activity.”
 - We would like this to be as complete as possible.
- ✘ We are inviting all QoS projects to get in touch with us
 - Either join the DOMA-QoS group or just drop Oliver and myself an email.
- ✘ The document is currently being prepared, so please get in touch quickly.