# QoS Session
# WLCG Workshop

eXtreme DataCloud

Data Management for extreme scale computing

Oliver Keeble on behalf of the working group

European Commission

**eXtreme DataCloud is co-funded by the Horizon2020 Framework Program – Grant Agreement 777367**

# WLCG Workshop : QoS Session

- Brief Intro
- QoS WG activities
  - Survey
  - White Paper
- Experiment input
- Storage providers
- Discussion

# Introduction

- “Quality of Service”
  - A quantitative measure of service performance characteristics
    - Intended to be associated with a cost and a workflow
  - “Unreliable and cheap”, “Fast and expensive”
- QoS is asking questions such as:
  - Are there places in experiment work-flows where it makes sense to trade performance/reliability for increased storage capacity?
  - Are there places in experiment work-flows where a small amount of higher performance storage would yield significant benefits?
- QoS our umbrella term for finding the cheapest possible solution to a given problem (workflow)
  - Concentrating on storage

# Introduction

- Is this new?
  - Have you always tried to meet your pledge at the lowest possible cost?
    -
  - Do you wonder how you could deliver your services more cheaply? Or if your users could manage with something different?
    -
  - Do you think this is going to get any easier?
    -
- QoS
  - Is not new
  - Is a new label to group existing efforts
- Now is the time to
  - Give it some more emphasis
  - Coordinate efforts

# Introduction

- Is this new?
  - Have you always tried to meet your pledge at the lowest possible cost?
    - You have always cared about QoS
  - Do you wonder how you could deliver your services more cheaply? Or if your users could manage with something different?
  - Do you think this is going to get any easier?
- QoS
  - Is not new
  - Is a new label to group existing efforts
- Now is the time to
  - Give it some more emphasis
  - Coordinate efforts

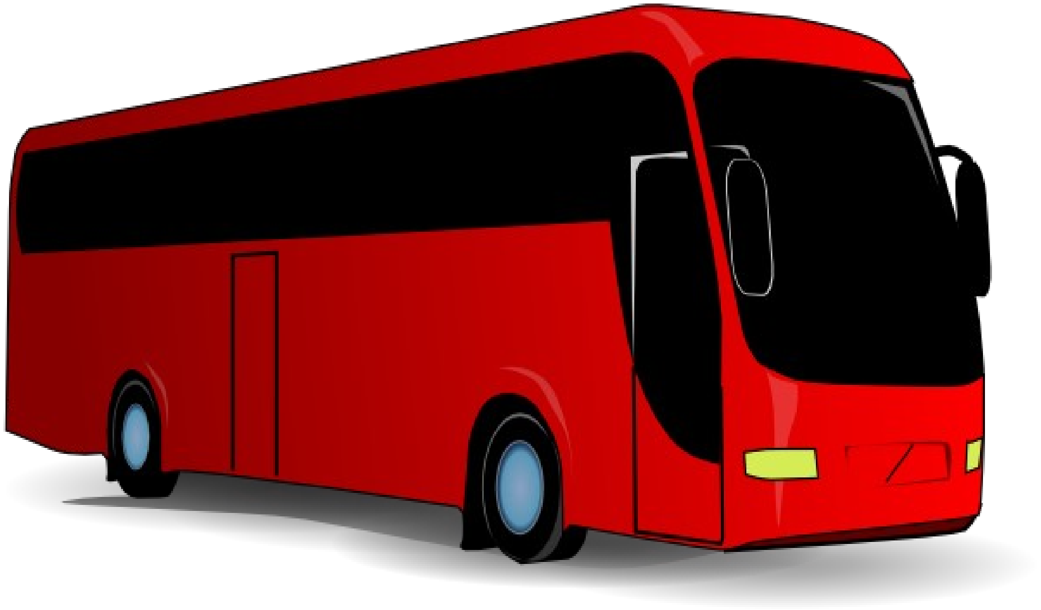
# Introduction

- Is this new?
  - Have you always tried to meet your pledge at the lowest possible cost?
    - You have always cared about QoS
  - Do you wonder how you could deliver your services more cheaply? Or if your users could manage with something different?
    - You care about QoS
  - Do you think this is going to get any easier?
    -
- QoS
  - Is not new
  - Is a new label to group existing efforts
- Now is the time to
  - Give it some more emphasis
  - Coordinate efforts

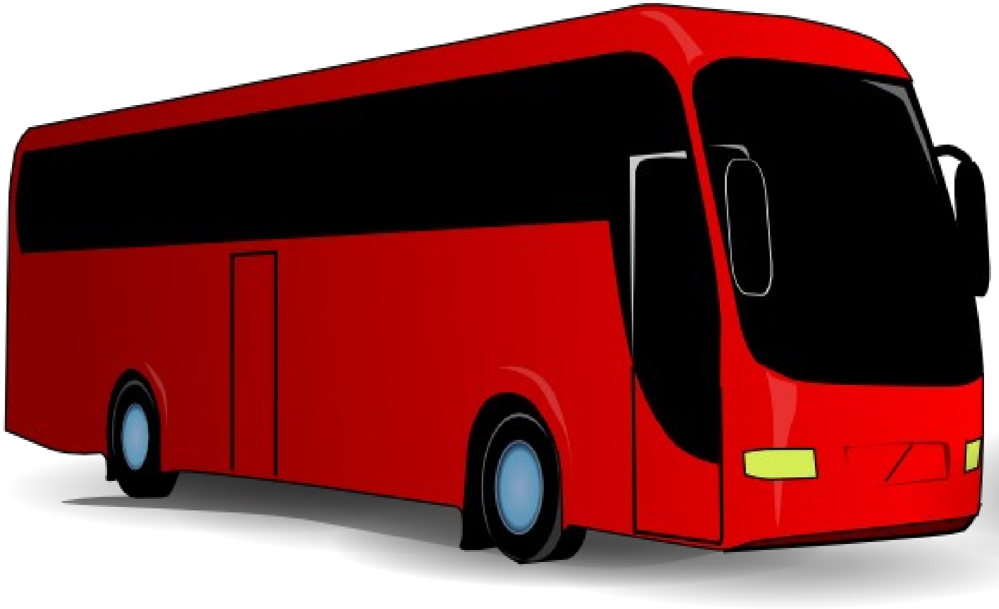
# Introduction

- Is this new?
  - Have you always tried to meet your pledge at the lowest possible cost?
    - You have always cared about QoS
  - Do you wonder how you could deliver your services more cheaply? Or if your users could manage with something different?
    - You care about QoS
  - Do you think this is going to get any easier?
    - You will only care more about QoS
- QoS
  - Is not new
  - Is a new label to group existing efforts
- Now is the time to
  - Give it some more emphasis
  - Coordinate efforts

# Introduction – an analogy

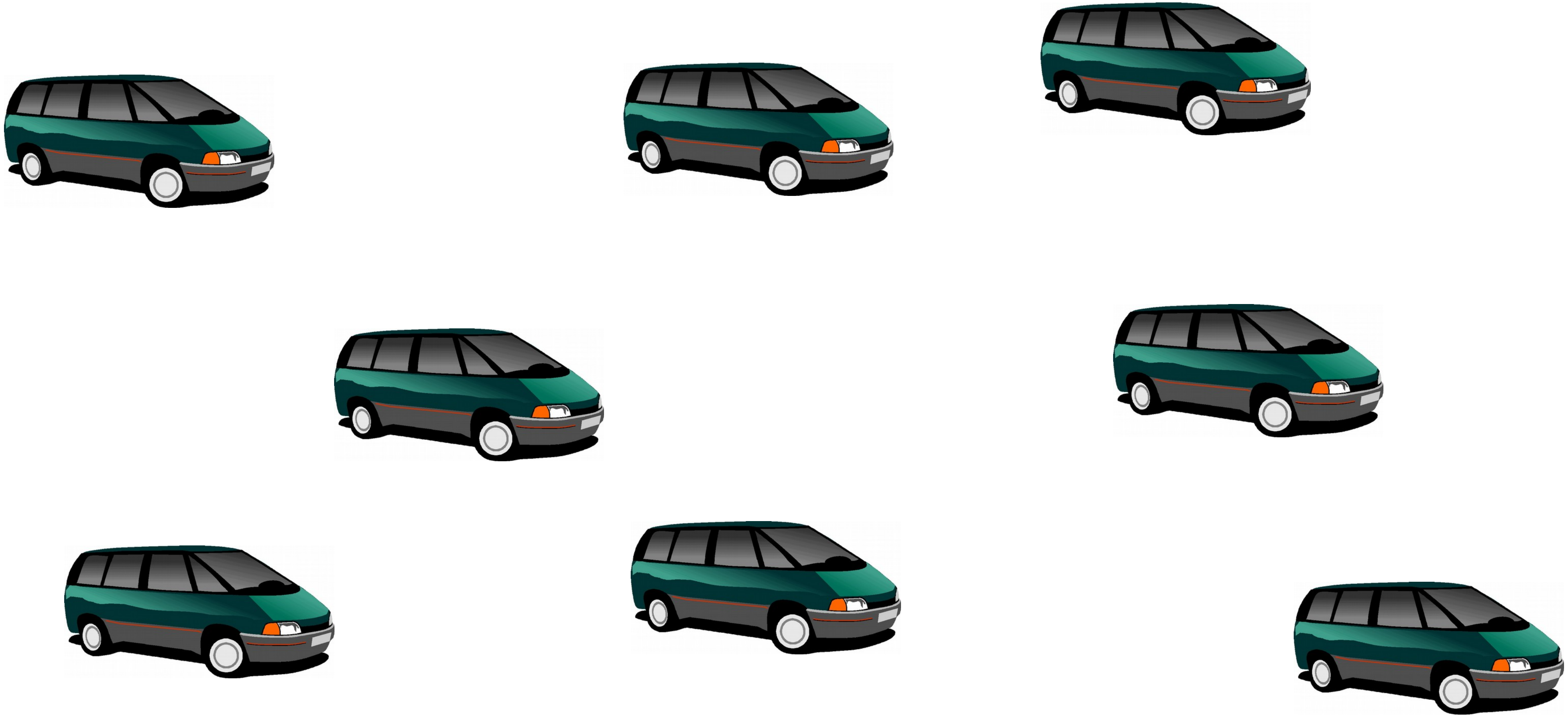


# Introduction

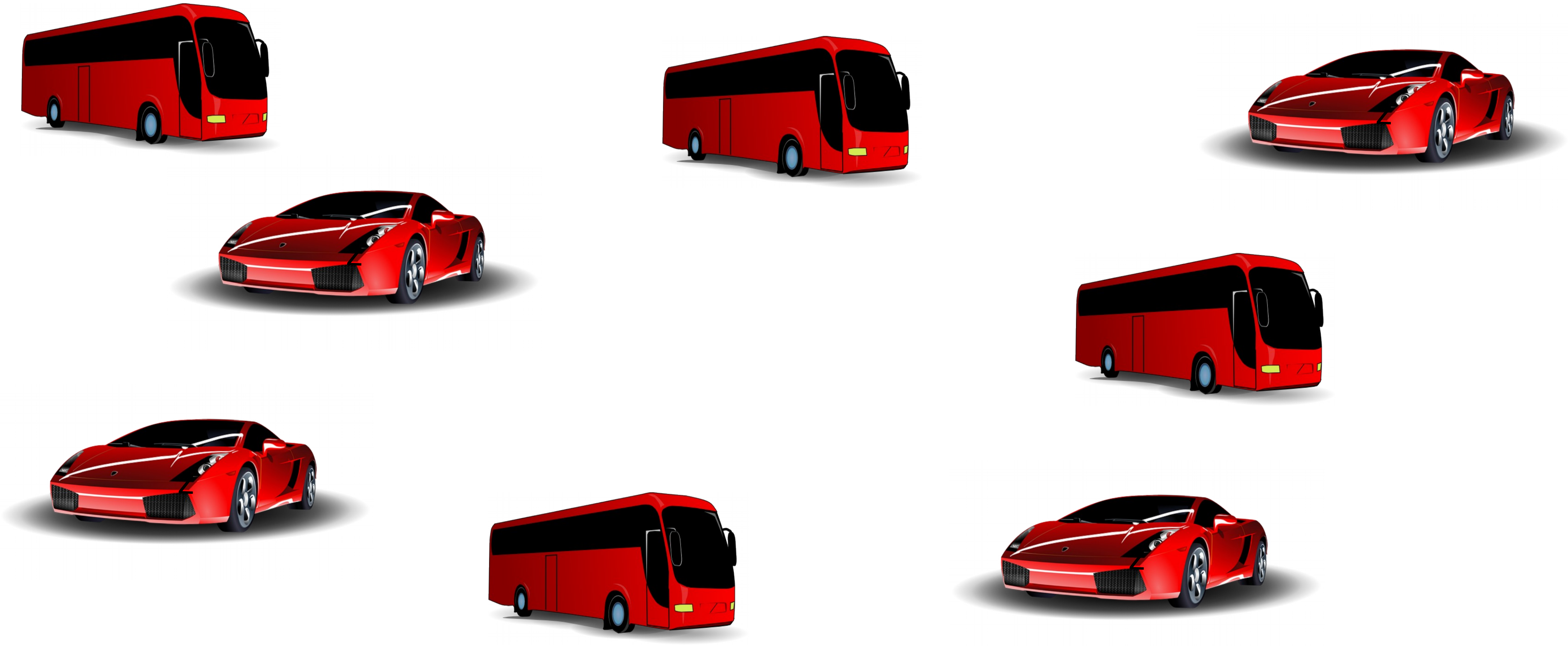


# Introduction

# Introduction



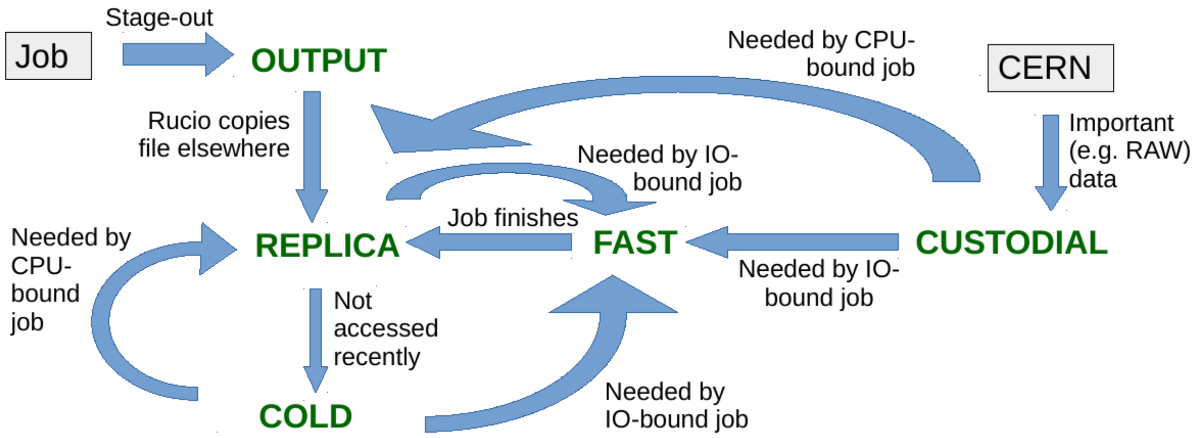
# Introduction



# Familiar QoS concepts

- Disk
  - Huge QoS variations possible under this category
  - All relevant workflows mapped onto this
    - For a particular workflow, can be overspecified in some ways (e.g. reliability) and underspecified in others (e.g. concurrent clients)
- Tape
  - Covers both durability and low-cost
- → "Disk", "Tape"

- Example additional storage QoS possibilities:
  - Enterprise HDD as RAID: OUTPUT, REPLICA, COLD
  - Consumer HDD as JBOD: REPLICA
  - (public) cloud storage: COLD
  - SSD as JBOD: FAST
  - Internal replicas existing on multiple server nodes: FAST



# The DOMA Working Group

# WG Activities

- Site Survey
  - Understand the current and potential QoS landscape
- Experiment Contact
  - Map workflows onto QoS (i.e. onto different systems, reconfigured systems...)
- White Paper
  - A short reference on status and opportunities for cost savings through QoS in WLCG
- Gathering storage provider input
- Contact with other activities : Access WG, Storage Resource Reporting, Cost Modelling ...
- Get involved: sites, experiments and storage providers are very welcome!
  - https://twiki.cern.ch/twiki/bin/view/LCG/QoS
- Egroup: WLCG-DOMA-QoS
  - https://e-groups.cern.ch/e-groups/EgroupsSubscription.do?egroupName=wlcg-doma-qos

# Site Survey

eXtreme DataCloud

- Describe your current system
- Describe your users and use cases
- R&D involvement, future directions
- Will be sent out with example responses filled in by CERN and DESY
- CERN
  - EOS erasure encoding, Server hardware configuration, Tactical deployment of SSDs, Tape backends …
- DESY
  - …

# Experiment Input

# ALICE - two QoS types in the future (same as today)

- **Disk** - primary holder of analysis objects
  - No use case for complicated disk structures
  - Current implementation is OK - the size of the site (CPU) and nearby SE I/O performance are usually matching well
  - In very special cases (Analysis Facility) - direct negotiation with the site providing the AF
- **Custodial** (@present=Tape)
  - Single instance of RAW data and replica of the reco/MC output
  - Strictly controlled recall/access
  - SSD caches as tape buffer are very interesting concept
- **The trend** - software configurable storage, inexpensive hardware (JBODs, no hardware RAID, no special FS)
  - ALICE is fully on board with this
  - Sites manage the infrastructure, combined storage (aka ‘data lakes’) for close and well connected sites is working and we support it through the ALICE DM system

# Rucio & QoS Short Summary

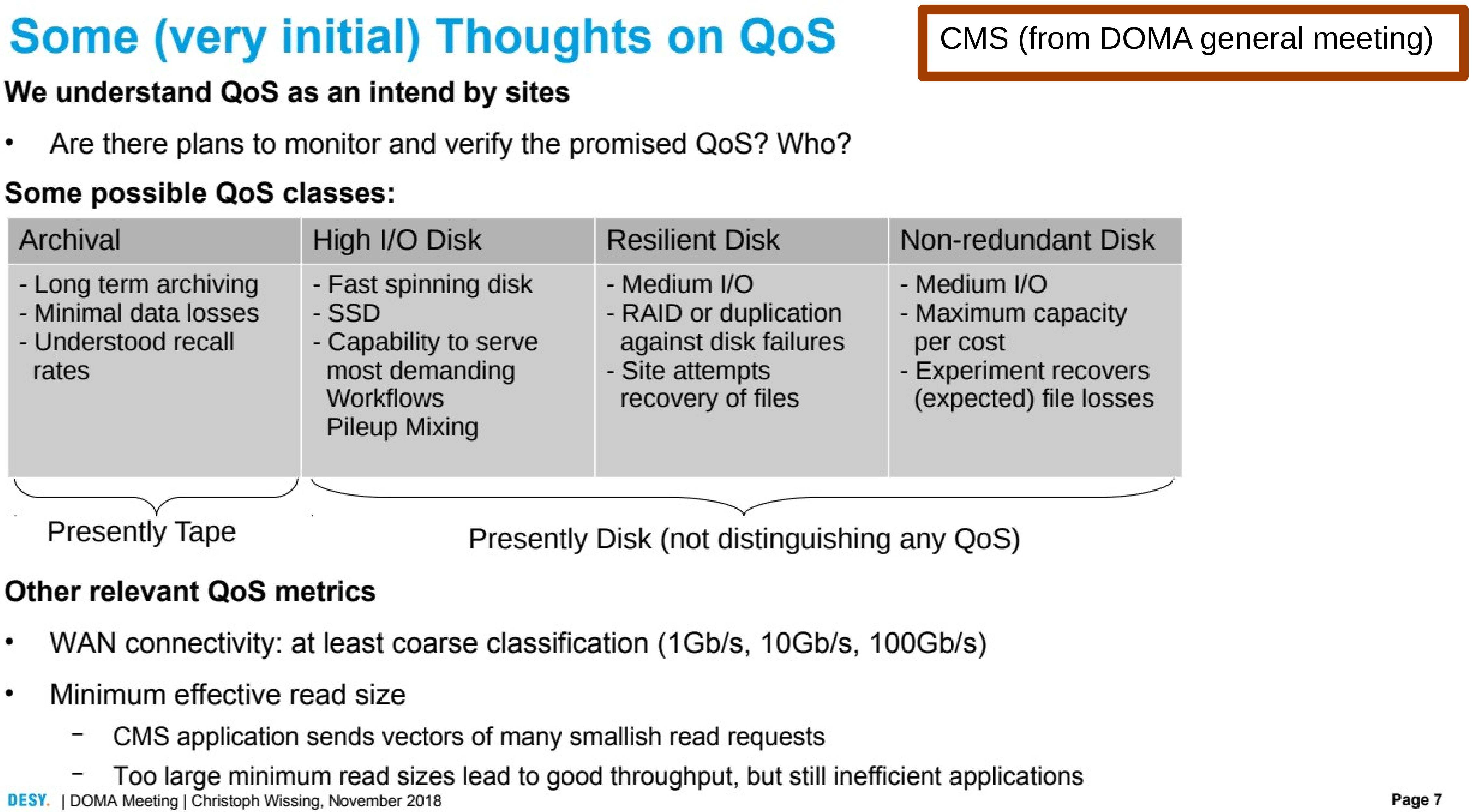
```
rucio add-rule my.dataset copies=2 country=UK lifetime=1w qos=cheap-med-latency
rucio add-rule my.dataset copies=2 country=UK lifetime=1w qos={latency:<100,
                                                               cost:<1000}
```

class or property?

What is important for us?

- Common language for the definition of QoS classes and QoS properties
- Common API + data structure to ask for for QoS transition
- QoS capabilities and zones from each storage need to be published and kept up to date
  - Rucio needs to know in which QoS zone the data is for internal scheduling
- Storage can automatically transition between "lower" QoS properties, but must never exceed constraint
  - e.g., move between cheaper zones without affecting combined cost and latency constraint
  - Must notify Rucio when such a transition happens
- Rucio would continuously check all QoS constraints at the rule level
  - Request transitions as necessary to keep rules satisfied

CMS (from DOMA general meeting)



# Some (very initial) Thoughts on QoS

**We understand QoS as an intend by sites**

- Are there plans to monitor and verify the promised QoS? Who?

**Some possible QoS classes:**

| Archival | High I/O Disk | Resilient Disk | Non-redundant Disk |
|---|---|---|---|
| - Long term archiving<br>- Minimal data losses<br>- Understood recall rates | - Fast spinning disk<br>- SSD<br>- Capability to serve most demanding Workflows<br>Pileup Mixing | - Medium I/O<br>- RAID or duplication against disk failures<br>- Site attempts recovery of files | - Medium I/O<br>- Maximum capacity per cost<br>- Experiment recovers (expected) file losses |

Presently Tape

Presently Disk (not distinguishing any QoS)

**Other relevant QoS metrics**

- WAN connectivity: at least coarse classification (1Gb/s, 10Gb/s, 100Gb/s)
- Minimum effective read size
  - CMS application sends vectors of many smallish read requests
  - Too large minimum read sizes lead to good throughput, but still inefficient applications

DESY. | DOMA Meeting | Christoph Wissing, November 2018

Page 7

# Experiment input : LHCb

- QoS appears through the “Service class”.
  - In LHCbDirac: configuration linked with operational requests. No software definition
- T1D0: used for archive, this very precious data.
  - Operationally 2 replicas for RAW data but only 1 for other (derived) datasets
  - Heavy task to reproduce derived dataset in case of loss => high reliability required
- T0D1: used for 3 purposes
  - Datasets for physics: usually >1 disk replica + 1 archive => loss is not a disaster, can be recovered
  - Temporary datasets (before further processing/merging): a single replica with life time of a few days => loss created operational complications, although re-creation is possible but painful
  - User private data )e.g. nTuples): usually 1 disk replica, can be re-created with operational complications (users are less experienced). Also used for input sandboxes, this availability is usually a problem (jobs cannot run if SB is unavailable)
- T0D2 : EOS @ T0
- Possible improvements
  - New class with very high QoS for temporary data (also for user data?)
  - Important:  New classes should be available through separate endpoints or explicit prefixes

# Storage Systems

# Storage Systems

- Storage Systems' QoS support generally already exceeds what we currently use in WLCG
  - All support pools with different media types
    - Most distinguishable by prefix
  - All support multi replica either natively or through the backend system
  - Almost all support multi-site operation
  - Most have hierarchical support with potentially automated QoS transitions
    - Including tape backends
  - Some have volatile or caching modes of operation
  - Some support CDMI, an interface extendable with support for QoS operations
- What's missing?
  - Production-grade QoS Management interface (but do we need it?)

# Discussion

# Discussion points

- Is a new “contract” desirable/possible between sites and experiments?
  - What characteristics do we care about? (i/o, durability, ...)
  - Does the pledge system need a review?
  - How would new QoS classes be validated?
- What are sites interested in trying?
- What technology should we be reviewing?
- How can the experiments adapt their workflows to exploit QoS savings?
- What QoS transitions on a single system are desirable?
  - Is a community discussion required for a post-SRM tape interface?
- What other QoS initiatives are there (Escape)?
- What should the WG be concentrating on?