XDC-QoS / DOMA-QoS

eXtreme DataCloud

Data Management for extreme scale computing

Paul Millar

paul.millar@desy.de

dCache workshop

Tuesday 21th May 2019

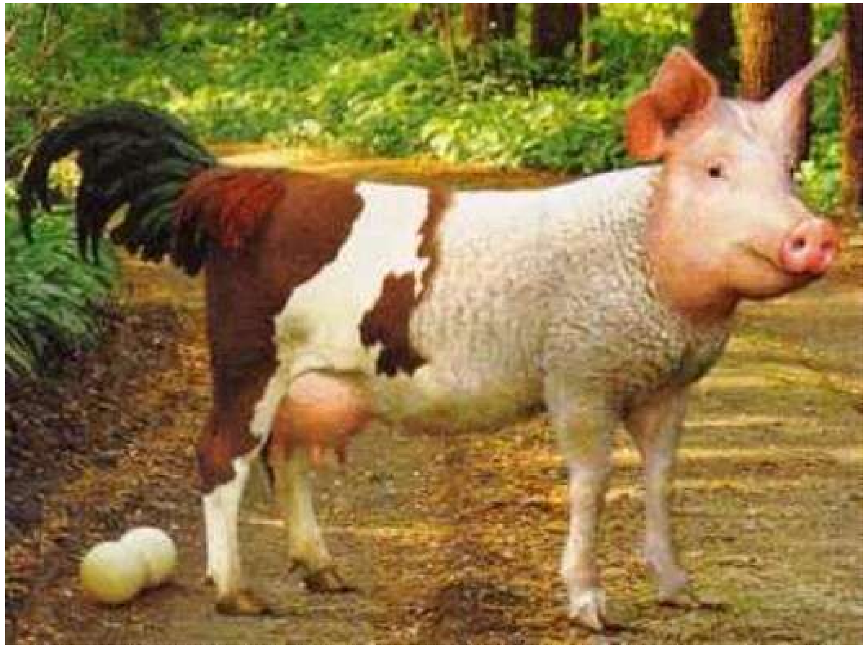
eXtreme DataCloud is co-funded by the Horizon2020 Framework Program – Grant Agreement 777367

Why storage-QoS?

Have cheapest possible storage

Get the “most science” from a finite budget

Why storage-QoS?



Eierlegende wollmilchsau

Building hybrid solutions, as no ***single*** storage technology can match desired behaviour.

Example: cheap storage that is both robust (“tape”-like), and fast (SSD-like).

What is storage-QoS: an analogy

Idea stolen from **Oliver Keeble** (thanks!)

What is storage-QoS: an analogy

Idea stolen from **Oliver Keeble** (thanks!)

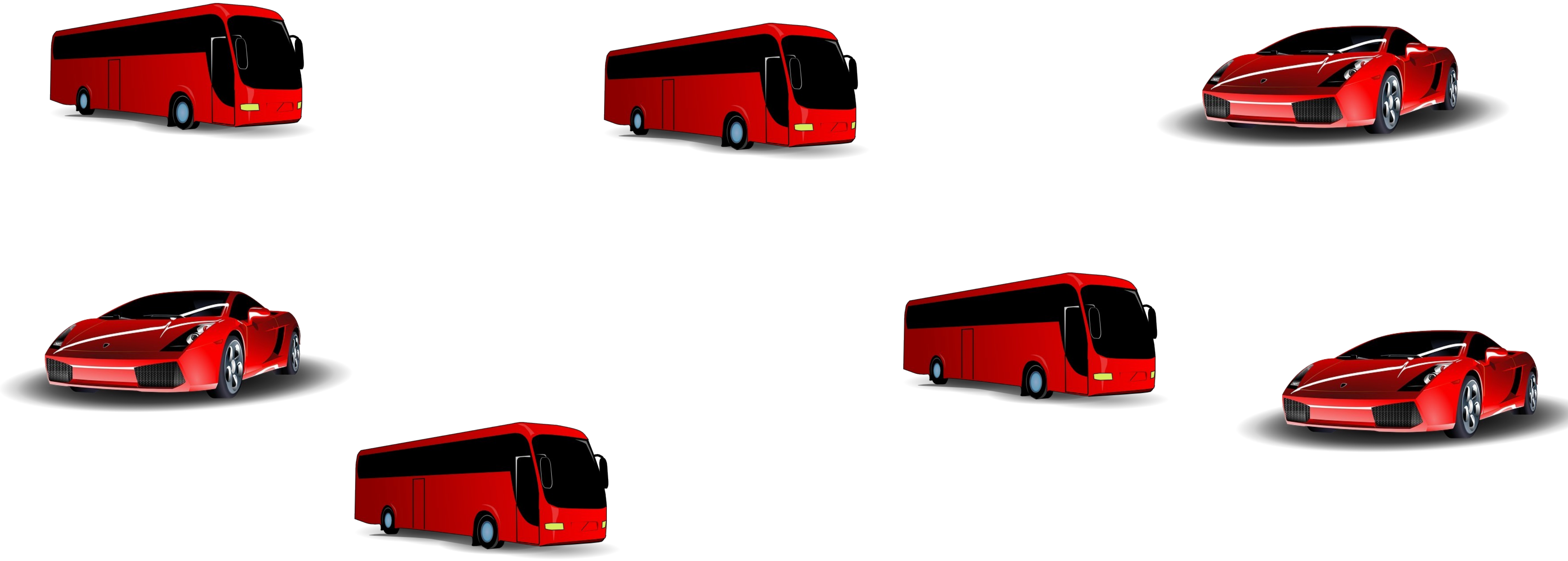
What is storage-QoS: an analogy

Idea stolen from **Oliver Keeble** (thanks!)

What is storage-QoS: an analogy

Idea stolen from **Oliver Keeble** (thanks!)

What is storage-QoS: an analogy



Idea stolen from **Oliver Keeble** (thanks!)

Different behaviour, different costs

- Different media options have different characteristics
  - Tape, “cheap” disks, “enterprise” disks, SSD, …
  - Different combinations of media: RAID, RAIN, JBOD, Erasure coding
- These also have different costs
  - Cost in terms of raw capacity used to store a 1 GiB file (JBOD vs RAID vs Erasure coding vs multiple-copies)
  - Cost in terms of money/budget-usage
- This is all very complicated – too complicated to deal with
- Better to describe **expectations**, rather than dictate how storage operates.

QoS as an agreement

Users

Storage behaves
how I expect

Storage providers

Promises on how storage
behaves, not on technology

QoS as an agreement

- Experiements decide what they **really** need
  - How bad is data loss, how much can you handle?
- Sites aim to provide what is desired – at a minimum cost.
- This works fine, provided everyone is honest
- It also allows for innovation:
  - new storage technology can be integrated if it matches minimum requirements
  - We have a framework for discussing new technologies.

QoS as a *qualified* agreement

latency: …
bandwidth: …
durability: …
cost-model: …

QoS as a *qualified* agreement

Users

Storage behaves how I expect

Storage providers

Free to innovate on how this is implemented.

latency: ....
bandwidth: ...
durability: ....
cost-model: ...

Available QoS at a site level

- A site provides finite choices, not arbitrary selection
  - You can chose from these options: QoS-A, QoS-B or QoS-C.
  - These choices may be influenced by discussion with experiements, but that happens on a longer time-scale.
- QoS options at a site:
  - A site may provide a single QoS.
  - A site could provide multiple storage system, each with a single QoS.
  - A site could provide storage systems with multiple QoS.

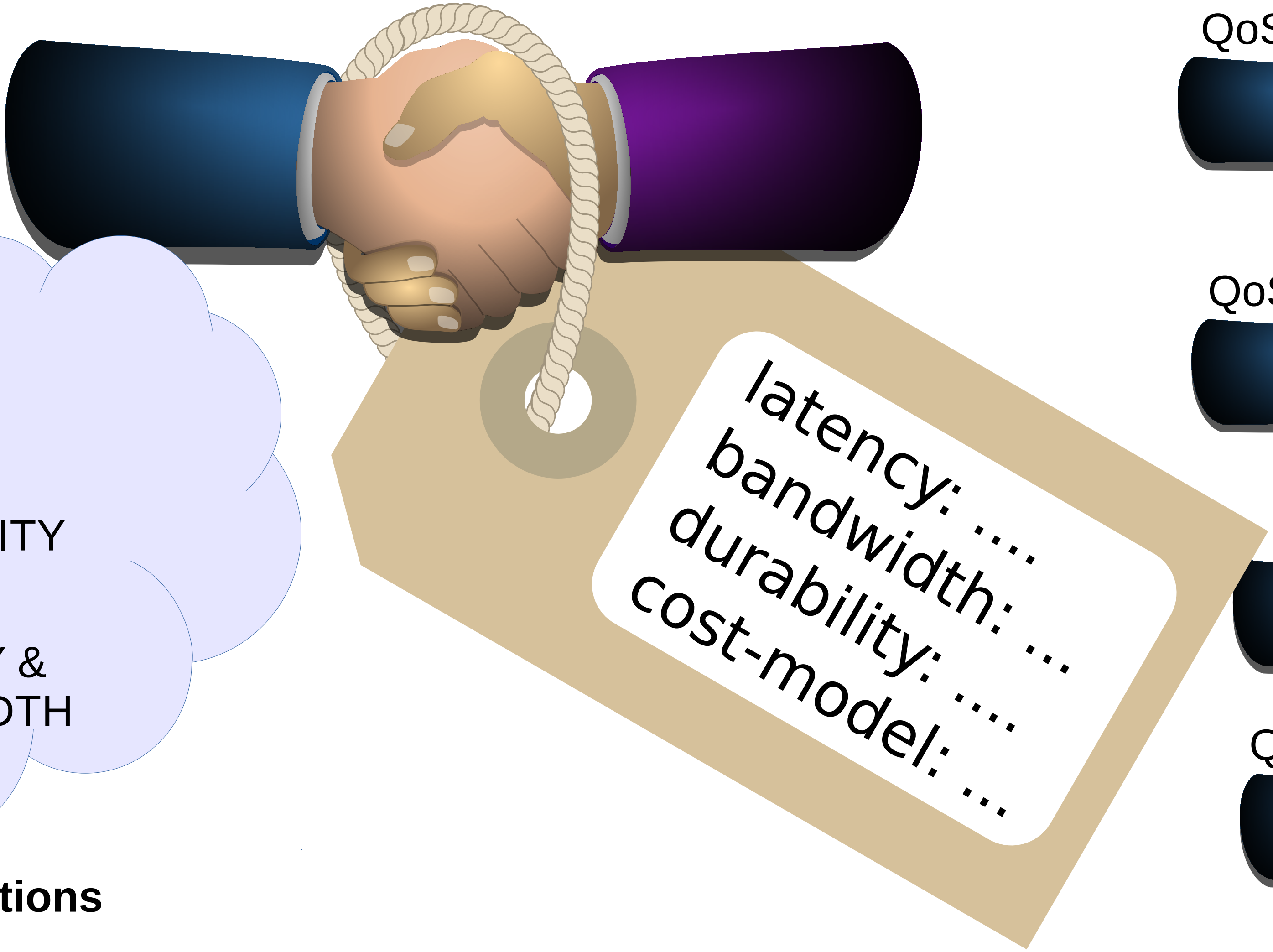
QoS as an agreement on behaviour

“**SCRATCH**”
(latency)

“**ARCHIVAL**”
DURABILITY

“**FAST**”
LATENCY &
BANDWIDTH

User expectations



QoS #1: **SCRATCH**

QoS #2: **SCRATCH, FAST**

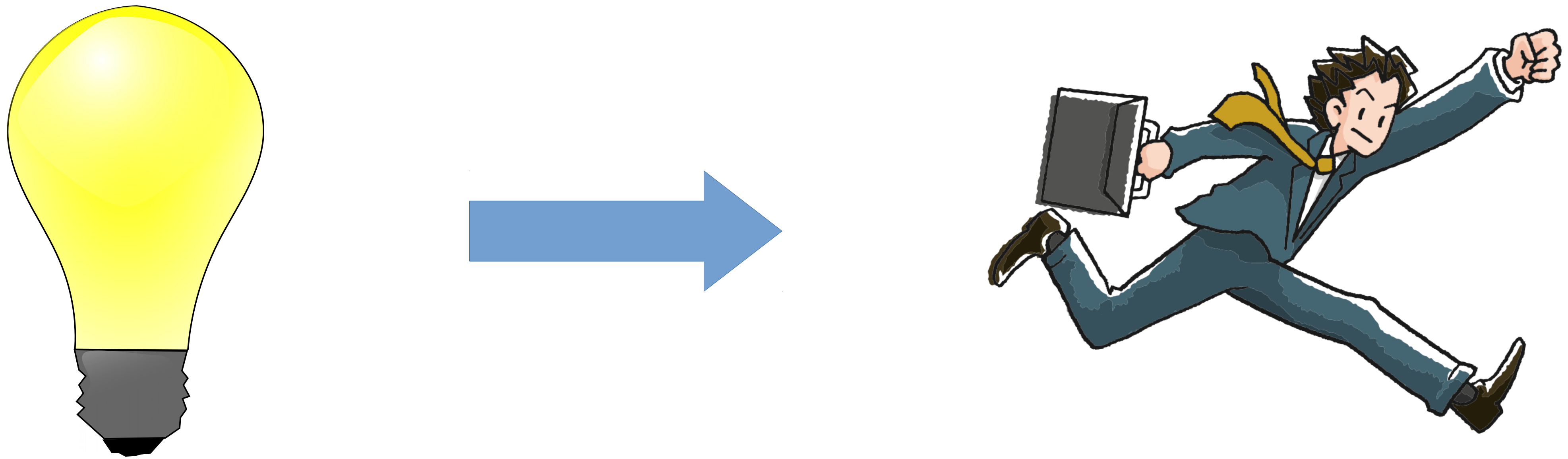
QoS #3: **ARCHIVAL**

QoS #4:

Case study: WLCG with DISK and TAPE

- WLCG has a long tradition of working with QoS
  - It just wasn’t called QoS.
- Different storage media was used:
  - Data was stored on TAPE because it is cheap.
  - Data was sometimes stored on DISK because it was just produced, or needs to be processed / analylised.
- Data is stored: on TAPE only, on DISK only, on TAPE and DISK
  - Different QoS: different characteristics for durability (likelihood of data-loss) and access latency (time to deliver first byte).
- Moving data from different QoS is automated, based on experiment polices.

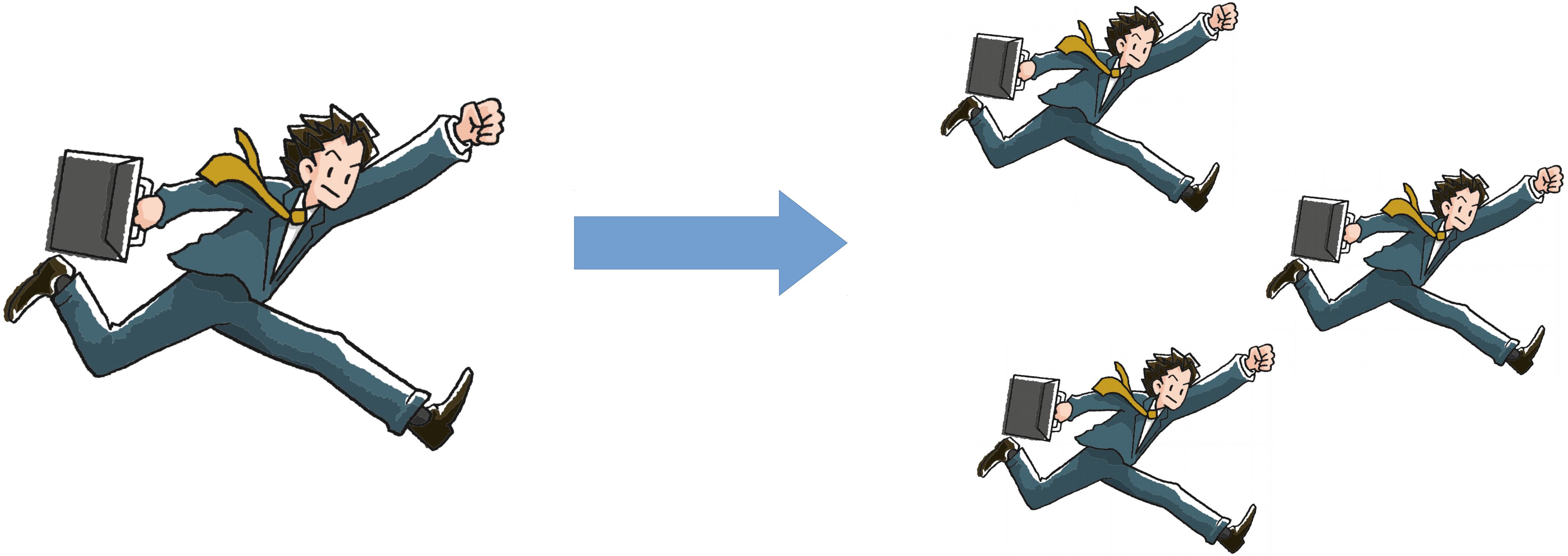
WLCG: Data Lake → DOMA



Data Lake
An idea

Data Organisation Management Access (DOMA)
A WLCG working group

WLCG: DOMA and DOMA activites



Data Organisation Management Access (DOMA)
A WLCG working group

DOMA activities
Each activity is a group with specific focus, all under a common DOMA umbrella

DOMA-QoS: two rhetorical questions

- QoS is asking two questions:
  - 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?

    (Note that these questions are strongly experiment focused: this effort will only be successful with strong input from experiments.)
- Assuming the answer to these questions is “yes” then how do we achieve these trade-offs?

DOMA-QoS: our motivation

“Given the expected **flat budget** for High-Lumi / RUN 4, create a mechanism to allow a **diversity** where **sites** can offer specific QoS options through innovative solutions that **save cost**. Through this **competition**, drive down the total cost of storage, while allowing **experiments** to optimise their **storage usage**.”

from DOMA-QoS Mandate

DOMA-QoS: our motivation

eXtreme DataCloud

> “Given the expected **flat budget** for High-Lumi / RUN 4, create a mechanism to allow a **diversity** where **sites** can offer specific QoS options through innovative solutions that **save cost**. Through this **competition**, drive down the total cost of storage, while allowing **experiments** to optimise their **storage usage**.”

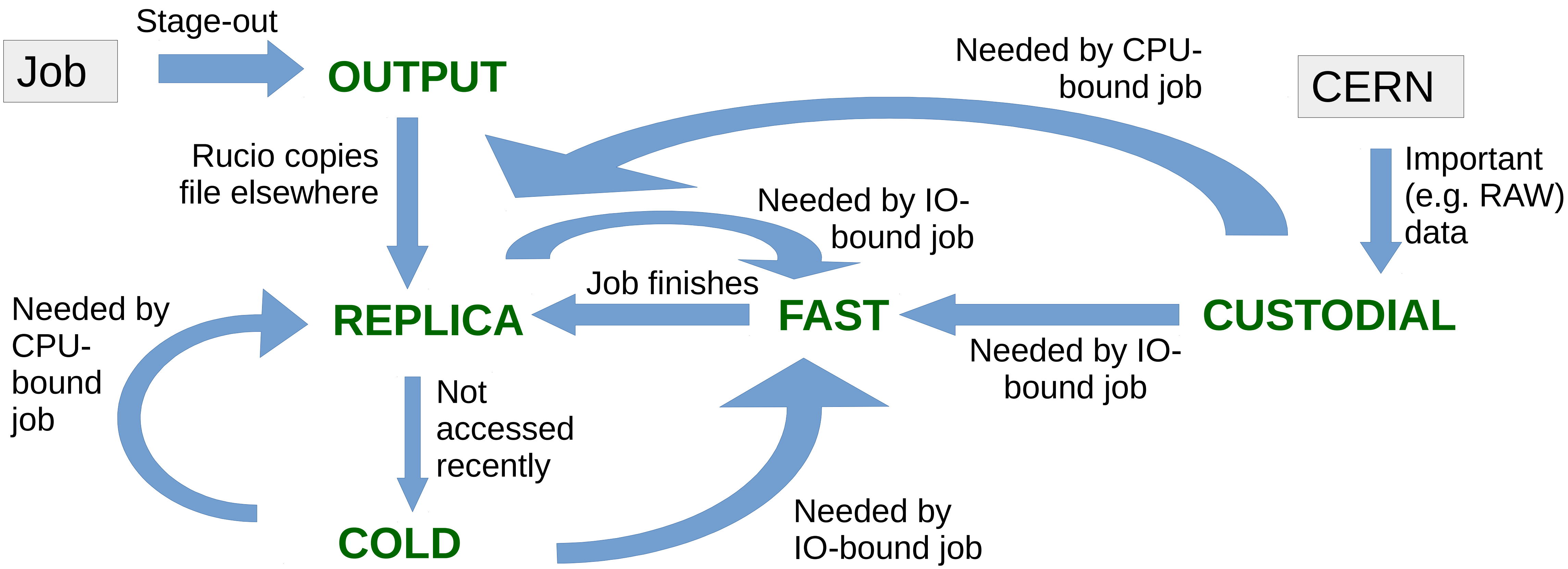
SAVE MONEY

from DOMA-QoS Mandate

DOMA-QoS: strawman model

- DISK → OUTPUT, REPLICA
  - **OUTPUT** storing only existing copy of data
  - **REPLICA** data also exists elsewhere (data loss more acceptable)
- TAPE → CUSTODIAL, COLD
  - **CUSTODIAL** storing data that must not be lost.
  - **COLD** data that is only used in bursts, and currently not being used.
- DISK → {OUTPUT/REPLICA}, FAST
  - **OUTPUT/REPLICA** input data for non-IO bound (analysis) jobs
  - **FAST** input data for IO bound jobs.

DOMA-QoS: strawman model



DOMA-QoS: strawman examples

- Example storage QoS:
  - 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**
- Same site could have multiple QoS that have required QoS label
  - For example, enterprise RAID and consumer JBOD both have **REPLICA** label.
  - Use “cost” to drive decision: cheaper to store data on JBOD than RAID.
- Different sites could implement QoS using different technologies
  - As above, would like “cost” to drive decision.

DOMA-QoS: current activity

- Engage with **experiments** to explore adapting workflows to include QoS concepts: **white paper**,
- Engage with **sites** to learn what technologies are currently available, and from their experiences of technologies that are currently not available to experiments: **site survey**,
- **Coordinate** our activities within the wider community: other DOMA activities, WLCG workgroups, and (potentially) further afield.

eXtreme DataCloud XDC

XDC: Developing QoS

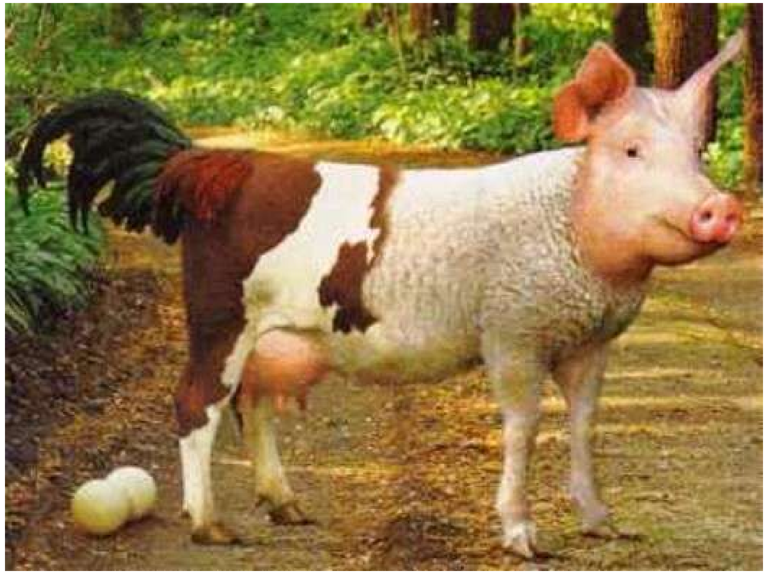
- EU-H2020 project, user-community driven development.
  - WLCG is one of these user-communities
- WP4 is a development activity, with which task 1 (→XDC-4.1) is working on QoS development.
- QoS activity continues the QoS work started in the INDIGO-DataCloud project.
- Focus has mainly been on adding OIDC and QoS support in FTS: using FTS to manage QoS transitions.
- Currently also supporting DOMA-QoS.

dCache developments

- New concept: data-placement policy
  - Says where data should be located, how many copies on disk or on tape, etc.
  - Different from (pool-manager) links, which is client driven
- A typical dCache has a handful of data-placement policies
  - A DPP corresponds to a QoS class.
  - Can assign metadata to policies, which become QoS attributes
- Each file is assigned one of these data-placement policies.
- If a file's replicas do not match the file's data-placement policy, dCache fixes the problem.

DataLake QoS orchestration

DataLake QoS orchestration



Providing aggregate of site QoS

- Select “appropriate” storage:

  E.g., only select sites that have agreed to support a research community.

- QoS aware data placement:
  - Move data to storage that meets requirements, as requirements change.
  - Data is now no longer embargoed, should be on “public appropriate” storage
  - Data is now cited in paper, should be on long-term storage.
- QoS to drive down cost
  - e.g., Cheaper to store data on JBOD than replicated-storage.
- Different sites could implement QoS using different technologies
  - As above, would like “cost” to drive decision.

Take-away messages

- QoS is motivated by:
  - Saving money
  - Building something “better” than any one site can provide.
- QoS is an abstraction of storage.
- QoS is an experiment driven activity:
  - It only makes sense if integrated into experiment work-flows
  - this is HARD.

eXtreme DataCloud

Thanks for listening!

Backup slides