
CERN

Qos Site Survey

Results, Conclusions, Actions

Oliver Keeble
On behalf of the QoS WG

Site Survey

- Around 80 sites responded
- Results
  - https://twiki.cern.ch/twiki/bin/view/LCG/QoSSurveyAnswers
- Analysis
  - https://twiki.cern.ch/twiki/bin/view/LCG/QoSSurveyConclusions
  - This contains much more detail than presented here
- QoS
  - Understand how embracing diversity in storage can save money
    - e.g. use cheaper, less reliable systems where possible
- Survey purpose
  - Capture relevant ongoing efforts and directions relating to cost-saving and QoS

Overview of answers

Q1 - Underlying media

- Majority of sites report using enterprise media
  - Some on consumer drives
- CERN's “consumer drive” experience is inconclusive
  - Consumer doesn't necessarily work out cheaper when you factor in bulk availability and prices
- Isolated experimentation with alternative media types
  - No clear direction here
  - But where can sites share their experiences with this?

Q2 - Media combinations

- RAID6 with 12-16 disks represents over 2/3 of sites
  - This does not give much margin for further cost savings
- Rest do JBOD
  - Redundancy comes from Ceph, EOS, HDFS and GPFS
  - Not much info on the replication level here :(
    - CERN is deploying EOS + EC for targeted uses and has a variety of Ceph deployments
- Where do we go from here?
- Abandoning redundancy would give ~15%
  - Under what circumstances would it be worth it?

Q3 - Storage systems

- Few surprises in grid storage systems
- Underlying resource more diverse
  - Local mounted fs
  - Intermediate layer
    - Shared fs (Ceph, Lustre)
    - Block (Ceph)
    - Deeper integration (HDFS and Ceph)
- Almost always a shared POSIX fs too
  - with a variety of uses
- Cost savings here would come from consolidation and reduction in the number of different solutions required at each site

Q4 – Effort

- Inconclusive
  - No clear mapping of “efficiency” onto system
- Difficult to quantify effort
  - But then how will we know if we save?
- T1s report ~2.5 FTE on storage
- T2s report ~0.6 FTE on storage

Q5 - "Storageless sites"

- The vast majority of T2s (who responded) are neither planning nor wanting to move to storageless setups
  - Local storage is needed
  - This can cut off independent lines of funding
  - Mixed/diverging opinions regarding caching solutions
    - we are interested / we are not interested / depends on which kind of cache you're talking about
- T1s indicate uncertainty about the impact of storageless sites on their systems

Q6 - Non-WLCG communities

- Practically all T1s are already sharing their resources across WLCG and other communities
  - No major problems identified
  - LHC experiments typically on separate systems
- T2s and T3s: approximately half of the sites are shared

Q7 - Future directions

- Ceph is cited in numerous contexts
  - Redundancy layer over JBOD, provision of S3, provision of posix fs
- No strong signal on site directions regarding cost saving
  - e.g. novel media (shingled disks), server densification, volatile storage …
- Many sites have no exploratory activities
  - Others are concerned about breaching MoU

Q8 - Experiment workflows

- Concern about granularity of matching workflow to resources → QoS classes
- Concern about cost and disruption of WAN access to custodial sites
- Desire for better tools provided by the experiments to the sites

What now?

Site level activities

- Procurement, densification and media
  - Purchasing strategy, server density and overheads, SMR, SSDs
  - Networking & implications of the data lake model on origin storage
  - This is trying to provide the current QoS, unchanged, at lower price
- Software defined storage ("SDS")
  - Configurable storage characteristics
    - Replication, erasure coding, media transitions
    - Future of RAID-6 and higher capacity disks
    - Pure JBOD operation
  - Stack consolidation - serving as many different use cases as possible with the same system
  - Introducing "SDS" into the stack
    - Use of Ceph, HDFS etc behind existing grid storage systems
    - Direct use of cluster FS tech by experiments (e.g. CephFS).
  - Identification of which configurations can be mapped onto which WLCG workflows
    - -> QoS classes

Site level actions

- Invite sites to begin a classification of their current offerings
  - Do this after the first set of recommended classes has been produced
  - Sites could also add any they provide (or are interested in providing) but do not figure in the list
- Invite sites to report on current relevant directions
  - Media diversity
  - Redundancy layer
    - In particular, introducing this into existing systems
- Attempt the "JBOD experiment" in conjunction with an experiment.
  - Remove all redundancy and work on handling data loss gracefully
  - As a cache? Is this being done already in the Access WG?
- Understand where this should be progressed
  - Many points are under the experiment radar and more aligned with forums like HEPiX

Grid level activities

- WLCG QoS classes
  - Definition of the most useful set of QoS classes allowing WLCG to progress beyond "disk" and "tape"
  - Tagging and brokering
  - MoU
- Client-driven QoS
  - Interfaces, clients, orchestration
    - inc bring-online
  - Data Lifecycle

Grid level actions

- Organise a dedicated consultation meeting with each of Alice, CMS and LHCb.
  - Identification of QoS classes useful to the experiment
  - Some reference use cases for these classes
  - Plan for how a new QoS class can be introduced, integrated, tested and exploited
    - Start with one (potentially already available)
- The WG intends to publish a white paper
  - Informed by the results of the experiment consultation, along with the survey conclusions
  - A statement about the potential of QoS and what we would need to do to achieve identified objectives
- Sites should be solicited to provide new experimental storage areas with novel QoS features
  - Trigger exploration of what adaptations are needed over the stack
  - Experiment/Site combinations could be encouraged to report.