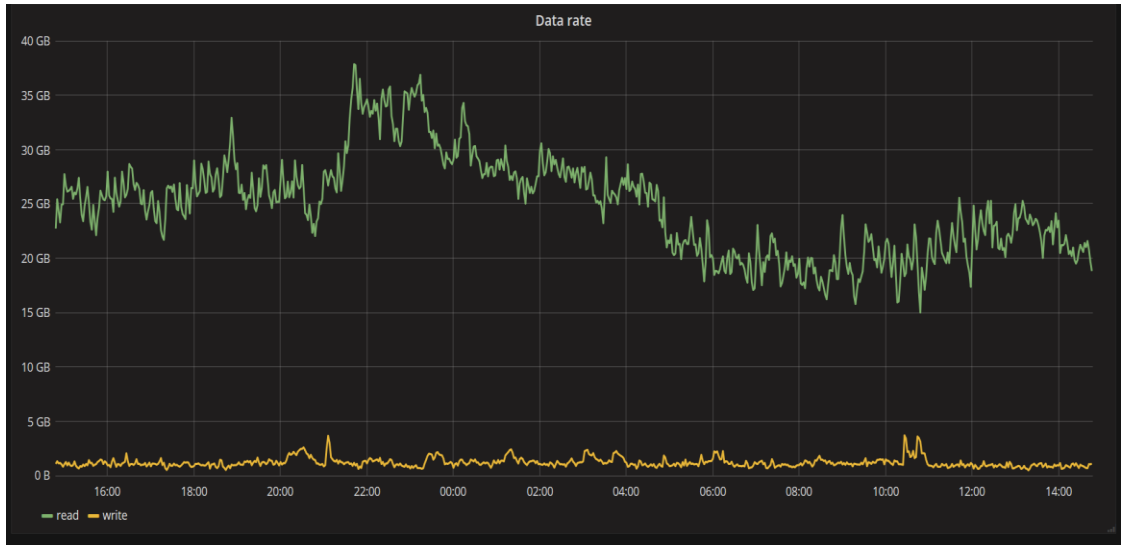
Disk Storage

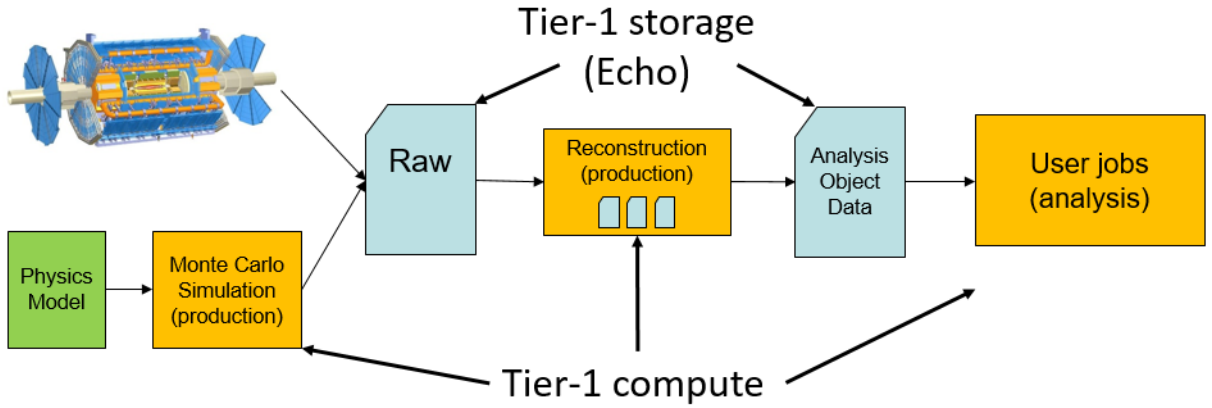
Tom Byrne

Introduction

- The LHC experiments require large amounts of local storage to support their compute at each Tier-1 site.
  - Datasets written in from the LHC, and files/events read by jobs running on local analysis machines
  - High throughput computing – many independent transfers
- I am the technical expert and storage architect for the Tier-1 disk storage known as "**Echo**".
  - ~60PB of Ceph object storage that stores data from the LHC experiments.
  - ~200 storage servers, ~6000 drives
- We also run other (mainly Ceph) storage for a variety of different use cases



Typical Echo read load from batch farm jobs



Current Hardware (Echo storage)

Large scale object storage for the LHC

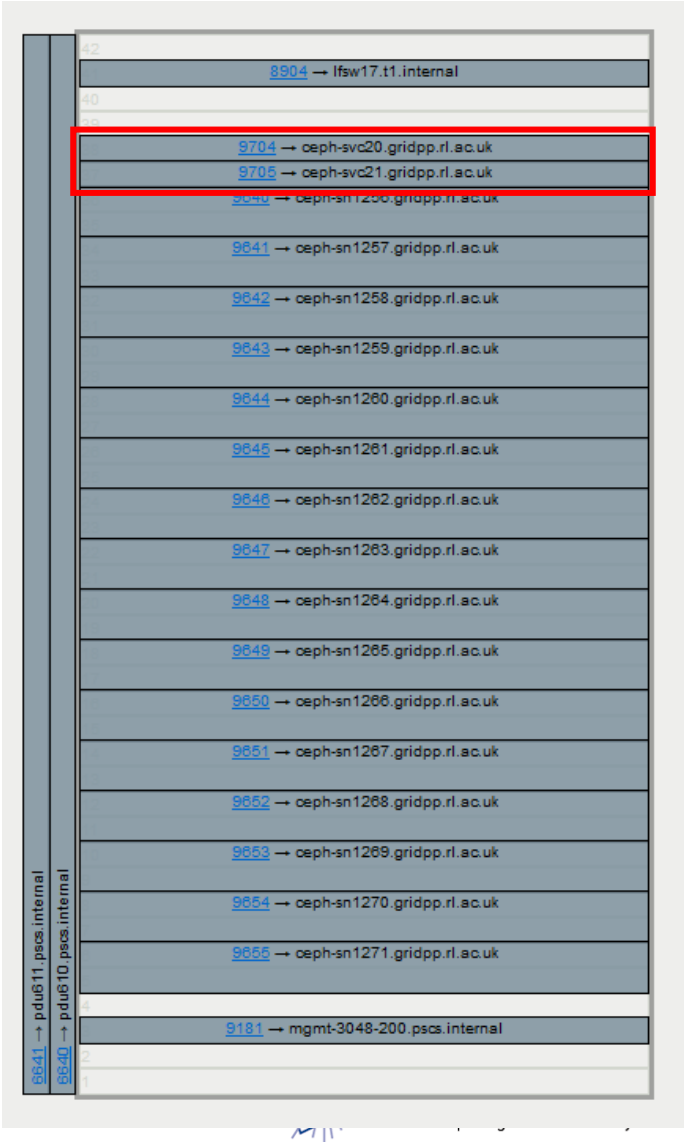
	Specification
Storage node size	24 x 20TB conventional HDD 1 x 4TB SSD
CPU	24C/48T (~1 core per disk)
Memory	192GB (~6GB per disk)
Network	1 x 25Gb Mellanox (~1Gb per disk)

- In each of the last few years we have purchased ~30PB of Echo storage
- The 2U, 24 drive form factor with ‘the biggest HDDs we can get’ has been working for us…
  - But we’re open to other possibilities if they have benefits.
  - No specific node size/capacity limits as long as the above relationships hold
- The main requirement here is the bulk storage at low cost, but we are starting to add faster storage tiers to this cluster, and would like to continue this in future years

Current Hardware (Echo service node)

- Echo service nodes fulfil the ‘non-storage’ functions of the service:
  - Ceph cluster monitoring and management
  - Gateway nodes for external access
- Our design has two 1U service nodes per rack of storage
- Generic ‘compute’ spec to allow for purchasing to be decoupled from requirement at time of use
  - Higher frequency CPU preferred compared to Echo storage
- 25Gb Network is starting to become a bottleneck for ‘gateway’ use-case

Current service node spec:	
CPU	32C/64T (2.4GHz base)
Disk	~1TB SSD
Memory	256GB
Network	1 x 25Gb Mellanox



Current Hardware (Other clusters)

Supporting on-site and UK science

Cluster	Deneb	Arided/Antares	Sirius
Purpose:	Large scale file storage	Fast file storage/ Buffer for tape archive	Block device storage
Size:	7PB / 60 nodes / 800 disks	3PB / 30 nodes / 800 disks *	1PB / 26 nodes / 208 disks
Storage node size:	16 x 8TB HDD 2 x 1.6TB SSD (~200GB per HDD)	24x4TB SSD	8 x 4TB NVMe
CPU:	32C/64T (2 core per HDD)	48C/96T (2 core per SSD)	32C/64T (4 core per NVMe)
Memory:	128GB (~8GB per HDD)	256GB (~10GB per SSD)	128GB (~16GB per NVMe)
Network:	1 x 25Gb Mellanox	1 x 25Gb Mellanox	1 x 25Gb Mellanox
Notes	• Hybrid storage needed for capacity per cost currently • An all flash (similar to the Arided/Antares) spec could be interesting to consider	• A few different iterations of SSD size and count have been bought recently, but would be good to standardise in the future	• In general – the faster the underlying storage the supporting more ‘resources’ needed • Additionally – the faster the storage the more preferable high frequency CPUs are.

*split over two different storage technologies and use cases

Science and
Technology
Facilities Council

Questions?