Storage cost optimisation on COSMA

Alastair Basden, Peter Draper, Mark Lovell, Fawada Qaiser, Richard Regan, Paul Walker

Alastair Basden
DiRAC / Durham University

Durham University
Institute for Computational Cosmology

DiRAC
High Performance Computing Facility

ExCALIBUR 10

Storage on HPC

- Traditionally: home, data and scratch
  - data and scratch are often the same
  - Balancing these helps to reduce cost
    - Select the right technology for the right place
- Thousands of nodes will write at once
  - Requires a parallel file system
  - Typical write patterns:
    - One file per node/rank
    - One file per job

Alastair Basden
DiRAC / Durham University

Storage on COSMA

- POSIX file systems
  - /cosma/home - ~100TB
  - /cosma/apps - ~100TB
  - /cosma/local - ~100TB
  - /cosma[5,6,7,8] - up to 20PB
  - /snap[7,8] - up to 1.2PB
- Object
  - StorJ
  - Ceph
- Tape - ~30PB

"home": snapshotted, backed up

"data": redundancy

"scratch": no redundancy

Alastair Basden
DiRAC / Durham University

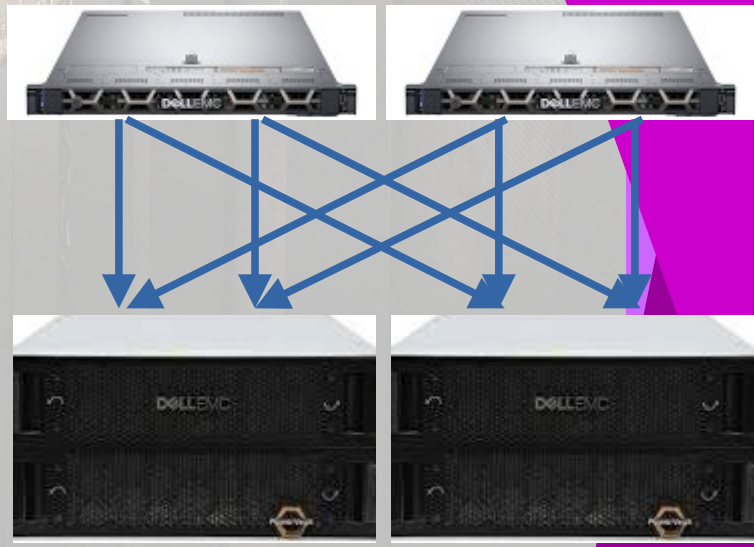
Bulk “data” storage

- Always in demand
- Aim to make it cheap: £50k/PB (inc)
- Needs to be performant
  - Parallel file system
- Ideally expandable

Alastair Basden
DiRAC / Durham University

Standardised design

- Lustre with ZFS
- Pairs of servers and 84-drive JBODS
  - 12-16TB drives
  - 7 Zpools per server: Z2 (10+2)
  - Manual failover
  - Redundant controllers, HBAs and cables
- Overheads: x0.67 (RAID, TB to TiB, reserved)
- Expandable - ~20GB/s/pair
- Self-installed: No software costs



Alastair Basden
DiRAC / Durham University

Issues

- Rebuild times - up to 4 days
  - DRAID should reduce this
- Performance bottlenecks
  - Badly behaved codes
    - User education
  - Or hardware faults

Scratch tier

- Lustre on NVMe
- High performance: ~400GB/s
- Non-redundant
  - Component failure leads to data loss
    - <1/year
- Regular clear out

Alastair Basden
DiRAC / Durham University

Standardised design

- Servers hosting 8 NVMe drives
  - One NIC per socket (InfiniBand)
  - Each drive becomes a storage target
    - No RAID

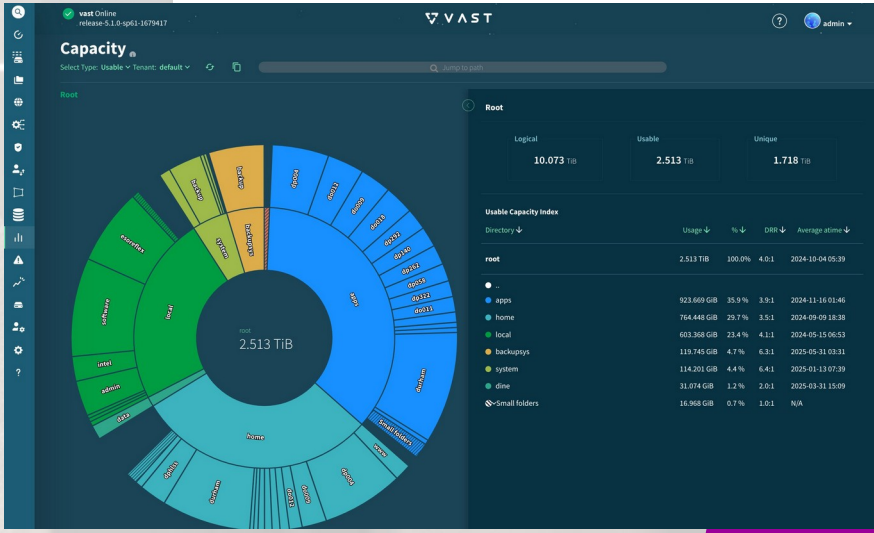
Alastair Basden
DiRAC / Durham University

Hardware reuse

- Storage warranties 5-7 years
  - We operate beyond this:
    - Repurposed spares
    - System shrinkage
    - Reinstallation and repurposing
      - JBODS are good for this
    - User awareness
    - Replacement disks (cheap)

Home space

- VAST
  - Good compression (4:1)
    - Home space is well suited
  - Not cheap
    - 2.5x considering compression
    - 10x otherwise
  - Reliable, performant
- Replaced a Ext4/NFS solution
  - Far better tooling, visualisation, snapshots, etc
- Backed up to tape daily (60d), snapshotted hourly (7d)
- Limited to 10GB/user



Lustre tends to struggle if used as homespace

Alastair Basden
DiRAC / Durham University

Application storage

- /cosma/local: System installed modules
- /cosma/apps: User space for software, venvs, etc
  - Snapshotted but not backed up
  - 100GB
- Using the VAST system

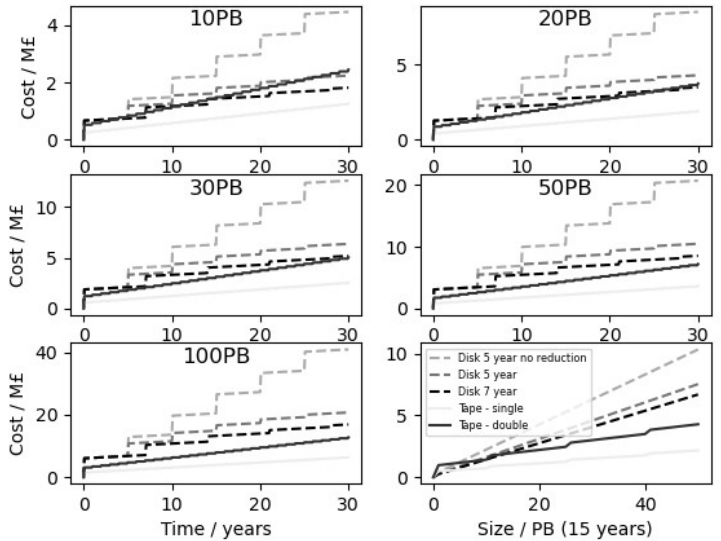
Object storage

- StorJ object storage:
  - Distributed cloud storage on-premise
    - Hosted at 3 DiRAC sites
    - Provides S3 access
    - Reusing old hardware
- Ceph object storage:
  - Required for VM/Container storage etc
  - Not as cost-effective or performant as Lustre
  - Old hardware can be used
  - Current IRIS funding for Ceph storage attached to Azimuth cloud
    - 3 Storage nodes, 24 drives each

Alastair Basden
DiRAC / Durham University

Tape

- Cost effective for volumes >30PB
  - Based on our historic purchases of disk and tape
    - And various assumptions on lifetime etc
    - Would be lower if our storage was more expensive
  - High initial costs:
    - Libraries, slot licencing, software licences
- 2 libraries - separate locations
  - Data replicated
- Soon to be 3 libraries with erasure coding



Alastair Basden
DiRAC / Durham University

Conclusions

- Storage does not need to be expensive
  - Assuming you can DIY
    - Expertise needed
- Important to differentiate use cases
  - User awareness / education helps
- 20PB for £1m

Alastair Basden
DiRAC / Durham University