QoS:
CNAF experience

Vladimir Sapunenko
Head of Storage and Data management group
Outline

- Redundancy
  - Experience with RAID and its sustainability

- Media
  - Incorporation of novel media types
  - Consumer vs enterprise drives
  - SMR disks
  - Fast cache

- Purchasing strategy
  - Server densification, reducing overheads

- Stack consolidation
  - Cost management through converging multiple user communities on the same system

- use of post-warranty hardware
Redundancy on the storage level

- Everything is under HW RAID protection
- Older systems are protected by traditional RAID6 (8+2)
  - Reconstruction for a 8TB NL-SAS hdd under normal load takes \(\sim 50\) Hours
- More recent systems are protected by Distributed RAID (similar to EC)
  - Bigger Storage Pools (70-180 HDDs)
  - Single stripe similar to RAID6 (8+2) cycling over entire Storage Pool
  - Reserved Capacity (like Hot Swap disks) - equal to the capacity of 2 or 3 disks
  - Reconstruction (recovery of missing blocks) starts automatically using Reserved Capacity
  - Since many disks participating in this reconstruction the whole process is much faster (\(\sim 4-6\) hours for 8 TB disks) and produces lower load on participating nodes
- Overall efficiency of disk space usage 75% (Usable/Raw)
Media

- **Capacity drives - Enterprise class NL-SAS drives**
  - Capacity 4 - 8 TB (14TB hdds are being installed)
  - Sector size 512B or 4KB
    - 4KB sectors are more performant from throughput point of view
- **Fast drives (SAS and SSD)**
  - Used for metadata or for some special FS (like DB or users home ~3.5TB)
  - Mid range SSD with >3 rewrites per day (DWPD)
- **No SMR drive in use nor expected to be used**
  - Needs specific driver and support on application level
  - Still expensive to be considered as tape replacement
- **HDD Failure rate 1 hdd/week (over +4400 disks deployed)**
Fast cache (NVMe)

With the price drop NVMe may become interesting for use as

- **Local cache on clients (on every WN)**
  - Volatile - OK
  - Can be used as local SCRATCH or as local CACHE for GPFS

- **Storage for tape buffer**
  - Persistent (RAID-protected) is better
  - Tape dives becomes faster and faster (easily archivable 400MB/s per drive)
  - One server can provide optimal (full) performance for up to 10 drives
    - 100GbE and/or IB for tapeserver->diskserver
    - 2xFC16 to/from tape

- **Dedicated storage for metadata**
  - Only persistent (protected by RAID)
Purchasing strategy

● Bigger integrated storage systems
  ○ Easy to manage
  ○ Faster in recovery
● Small number of servers
  ○ single server can manage different services providing up to 10 GB/s I/O rate
● 2x100 GbE for server to LAN connectivity
  ○ Redundancy on the LAN level
● FDR or EDR on IB for the server to disks connectivity
● 800-3000 TB /server (in prod now)
  ○ Guaranteed at least 3MB/s/TB in data access
● SW costs (licenses) is at ~3% to overall HW price
## Storage HW overview (disk)

<table>
<thead>
<tr>
<th>Manuf</th>
<th>Model</th>
<th>N. units</th>
<th>N. Of cont/unit</th>
<th>N. Encl./unit</th>
<th>RAID configuratio</th>
<th>SSD, N*Capacity (GB)</th>
<th>SAS, N*Capacity (GB)</th>
<th>NL-SAS, N*Capacity (TB)</th>
<th>Year in service</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELL</td>
<td>MD3860f</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>DDP 180(8+2)</td>
<td>0</td>
<td>0</td>
<td>180 * 4</td>
<td>2014</td>
</tr>
<tr>
<td>DDN</td>
<td>SFA12K</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>6(8+2)</td>
<td>0</td>
<td>0</td>
<td>36 * 300</td>
<td>2015</td>
</tr>
<tr>
<td>Huawei</td>
<td>OceanStor 6800v5</td>
<td>1</td>
<td>6</td>
<td>12</td>
<td>2.0 95(6+2)</td>
<td>25 * 600</td>
<td>0</td>
<td>855 * 6</td>
<td>2017</td>
</tr>
<tr>
<td>DELL</td>
<td>MD200f</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1(1+1)</td>
<td>24 * 300</td>
<td>0</td>
<td>0</td>
<td>2015</td>
</tr>
<tr>
<td>Huawei</td>
<td>OceanStor 18000v5</td>
<td>5</td>
<td>4</td>
<td>34</td>
<td>2.0 67(8+2)</td>
<td>12 * 1800</td>
<td>0</td>
<td>408*6000</td>
<td>2018</td>
</tr>
<tr>
<td>DDN</td>
<td>SFA 7990</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>DCR 56(8+2)</td>
<td>0</td>
<td>0</td>
<td>232x14</td>
<td>2020</td>
</tr>
<tr>
<td>DDN</td>
<td>SFA 220NV</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>DCR 7(4+2)</td>
<td>7 * 3200**</td>
<td>0</td>
<td>0</td>
<td>2020</td>
</tr>
</tbody>
</table>

- Disk space efficiency usage (usable/raw) – 75%
- * - 4KN sectors
- ** NVMe disks
LHCb example

**Servers** (dedicated):
- 4 as GridFTP, XrootD and NSD
- 2 as metadata servers
- 1 (VM) as StoRM FE/BE
- 1 as HSM

**Storage** *(shared with other exps)*
- 2 Dell MD3820f as metadata storage
- 3 Huawei OS18Kv5 (main storage)
- 1 SL8500 Tape library

So, only 4 I/O servers and 4 service nodes for 6PB of data!
MSS @CNAF

- 1 TSM commont server for all experiments
- 1 HSM server (tape server) for each LHC experiment
  - 2xFC16 to tape drives
  - 2xFDR IB or 2x10GbE to disks
- 2 tape libraries
  - only one actively used, the second one just entered in service
- 16 T10kD + 19 ts1160 tape drives

ATLAS example during tape carousel last week:
- Up to 8 tape drives (when available)
- Up to **84 TB/day** recall rate

NB: doing so with just 1 tape server
Stack consolidation

- One filesystem for more user groups
- Fileset in GPFS
- Use of fileset quotas makes possible to overbook filesystem
- “df” on user dir will show quota as available disk space
- Example
  - 4PB filesystem for 44 experiments
  - Sum of all quotas = 7.5PB
  - Space usage 85%
Use of post-warranty HW

Something that we are doing regularly, but

- It costs a lot in sense of power ether electrical and human
  - No HW can be run unattended, To keep it running support is needed
    - Disks are failing
      - “disks exist in two states: failed or about to fail” (© A.Maslennikov)
  - Always additional costs:
    - Using local man-power
      - Manageable until it is ease accessible - no way to keep it on remote site
    - Using lower level of external support
- Not sure if it makes sense in production environment
That’s All Folks