Site Report HPC2N

NDGF AHM 2022
2022-06-21 — 2022-06-22

Center storage, Lustre

- Versions
  - SFA400NVXE plus 4 enclosures BIOS DDN2.02.07
  - Exascaler 5.2.3
  - Insight 4.0.1
  - CentOS 7.9
- A couple of new LBUGs triggered
- The weird problem with shared libs (and other mmaped files) caused by hotpool still lurking. Hotpool Heatfn turned off to "solve" the problem
- Two occasions of servers either crashing or getting force-rebooted, probably related to low-level fsck problems and/or striped directories.
- Multiple occasions of unmotivated slowness, when rebooting servers one by one it's obvious that it's caused by a single server. DDN working on the issue.
- Upgrade of Exascaler planned for late June or early August

Kebnekaise/Cluster

- Ubuntu 20.04.4 LTS
- Slurm 21.08.7
  - having the support was really useful for advance warning and patches for the CVE's
- HW works really good
  - Some nodes have needed new CPU paste due to reoccuring warnings
  - Quite a few enclosures have had some of their fans replaced
- Normally we have a 100% percent usage of available nodes for users
  - The exception is the KNL nodes which have fewer projects using them, but currently there are 3 projects that really takes advantage and uses them fully.
- A few extra maintenance windows due to modifications of data center
  - mainly the ventilation and power supply structure, see later slide

Abisko dismantled, central IT borrows space

- University IT (ITS, formerly UMDAC) will renovate their data center
- ITS now temporarily housed in the old Abisko racks
  - Only approx 60-80kW, so not noticeable wrt power/cooling
- Temporary kludges for UPS power of ITS equipment
  - Including emergency cooling, which also needs to cater for HPC2N equipment if we get cooling loss but still have power.

Power and cooling kludges - quirks

- When testing that the data center works on UPS, it would have been smart to disable the controlled-shutdown-on-power-outage function in the ventilation control system.
- When using fresh/cold water for cooling it's at a higher pressure than the regular cooling circuit. The data center gets cold if the valve motors are too weak to stop the flow.
- When adding an additional power source to a data center, remembering to cater for emergency power cutoff is a clever thing to do.

Power and cooling kludges – quirks (2)

- Have a test plan, have all parties attend the meeting discussing this
- Have margins for delivery times in your problem-fixing schedule, you can't expect issues to be solved on short notice if parts have multi-week delivery times...

Tape/backup

- Hardware unchanged
  - IBM TS4500 library (2000 slot capacity)
  - 6x TS1155 tape drives
  - Dell R740xd with approx 200T spinning disk, 4x16G FC, 2x25GigE
- Added more SNIC tapes to handle increased copy pool volume
- Most hardware issues related to tapes being broken
  - Not that many, but more than you expect of tapes with "lifetime" warranty
- TSM server software on version 8.1.14.100
  - Support TOTP MFA for admin accounts, seems to work

Tape/backup (2)

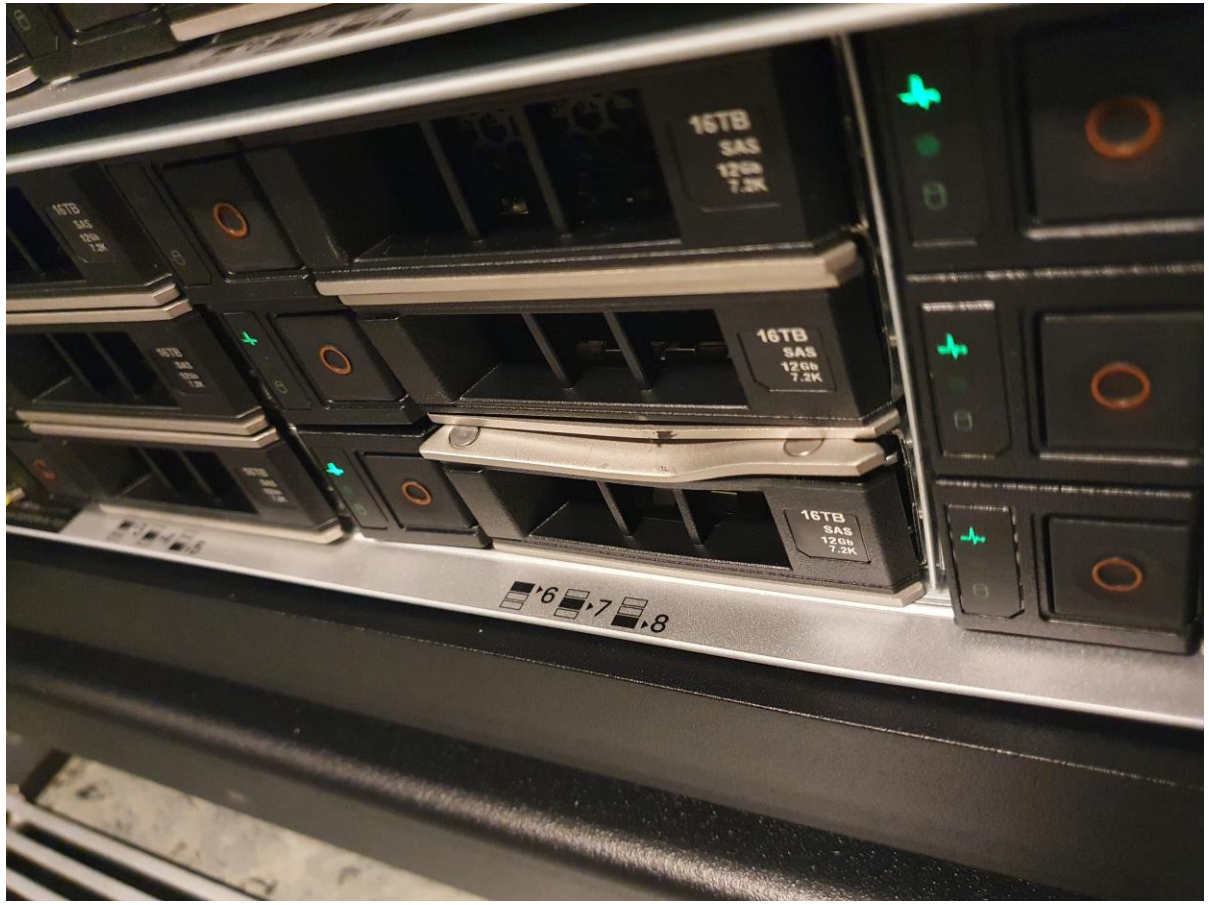
- WLCG specific tape status
  - No ALICE tape data (migrated to UIB)
  - ATLAS part mostly full
    - Some space being freed up by reclamation, needed to pause that for a while since it was so full there wasn't enough tapes to do reclamation in full speed.

NDGF dCache

- Purchased more WLCG/NDGF dCache storage
  - Replaced old storage (8x Dell R730xd)
  - 7x Dell R740xd2, 26x 18T HDDs
  - Purchased with some parts like NICs separate due to uncertain delivery times.
  - In the end, everything was delivered in time to be installed before service on old storage expired.

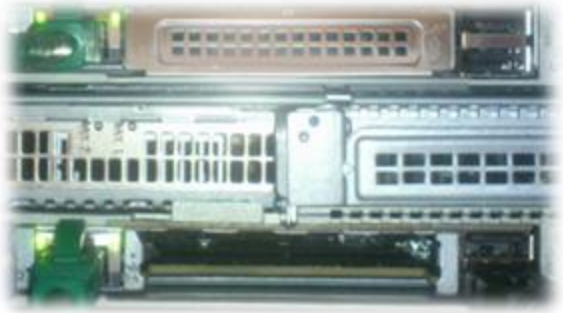
Swestore dCache

- Recommendation:
  - Do not drop servers
  - And if you drop them, don't drop them onto human body parts or other servers
- No data loss
  - Top drive lost contact for ~20 minutes, then returned and rebuilt
  - Likely due to SAS backplane flexed back
- Out-of-warranty repair
  - Replaced SAS backplane + 2x drive carriers
  - Realigned bent parts in drive cage



The continuing saga of HPE support (aka q-h36)

- Part of 4 node setup for ARC cache. Delayed delivery last week of last July
- 1 node broken @ delivery
- Took HPE well over 6 month to fix it
  - HPE tried replacing pretty much everything. Both HW and FW. Most of the time it just got worse.
  - When they finally fixed the original problem there were issues with the root-drives and a missing PCIe-cover plate.
  - They wanted fix root-drives issue with more new parts. We suggested a reseat first. That solved that problem.
  - It then took them 3 weeks to get the correct cover plate and then finally (March 16th) the machine was ready for production.



But wait there's more …

- Since it took HPE so long to fix the problem we asked for compensation
  - Proposed an elongation of the service contract … and to avoid having one odd machine spread it over all 4 machines (~ 2 month extra)
  - Initial response was positive but … because of reasons HPE seems to be unable to this directly.
  - Instead, they will compensate us fully with credits for the server. We are in the process of using that credit to get prolonged service. And maybe more if there is money over. 😀
- On Saturday May 7th q-h36 lost contact with it root drives
  - The raid-controller had hiccupped. Mon: figuring out how to remove machine from production (top-tip: avoid having the person that knows it, be sick and not document it. Puppet helped as secondary documentation), figure out what happened, reseat raid-cards. Tues: issue reported to HPE. Thurs: raid-controller replaced. Fri: firmware updates. Machine tested Mon-Wed. In production on Thursday.