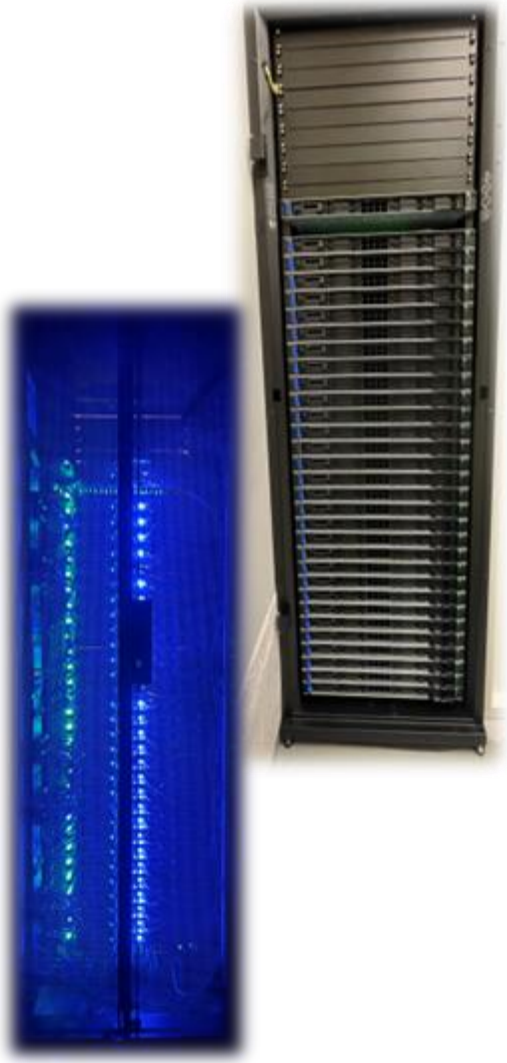
Site Report HPC2N

NDGF All Hands Copenhagen

2024-04-23

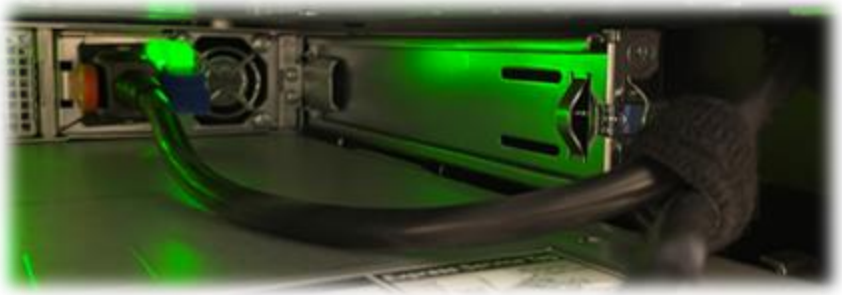
New WLCG compute – Aigert (aka g-ce01)

- 31 Dell PowerEdge R6625 nodes
- 256 cores/node (2 x AMD EPYC 9754)
- 3 GB/core, 3.2TB/node (12GB/core + 50GB), 25Gb/s
- Peak performance 393.6 GFlops, Est. HPL 217 GFlops
  - 25.4 HS23/core (~6510/node)
- 1 ethernet switch (100Gb/s for uplink and ARC-cache)
- 1 management switch
- ~~2~~ 3 PDUs



Aigert - issues

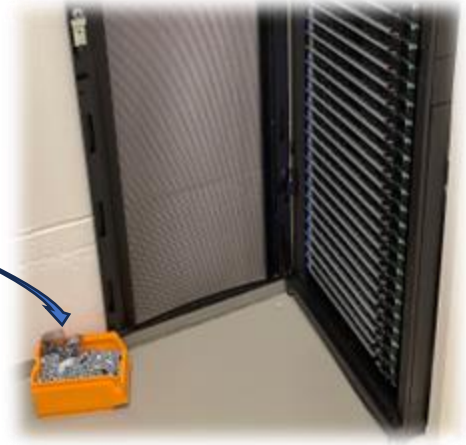
- Network cables were a bit short
  - Original plan was to have switches top and bottom.
  - Put ethernet switch closer to nodes (below top-node to avoid cooling leakage)
  - Problem! How to get cables to the switch? Nodes block access
  - Solution: Use some old brackets bought for old cloud



- 1 node slower when benchmarking HS23
  - 22% slower but 10 times longer execution time!?
  - Dell swapped CPUs between nodes
    - Without applying new CPU-paste?!
  - And then both nodes worked fine!
- Have an issue with it being unable to PXE-boot
  - Could be that PXE-boot is stuck in using HTTP instead of TFTP
  - Believed to be a minor problem, not even reported because of it not being that important!

Aigert – issues (cont.)

- Some nodes had hotter in-let temperature
  - 26-27°C vs 21-22°C
  - Temp. (no pun intended) kludge was to steal an ITS solution
  - Real solution – remove door! 😉



- 2 PDUs not sufficient
  - Supplier had calculated with power usage at around 1200W, we were hitting 1400 with HS23
  - HPL even hit 1500! and managed to trip a PDU fuse
  - In fact, could use more amps than 16-amp outlet
  - Solution – one more PDU

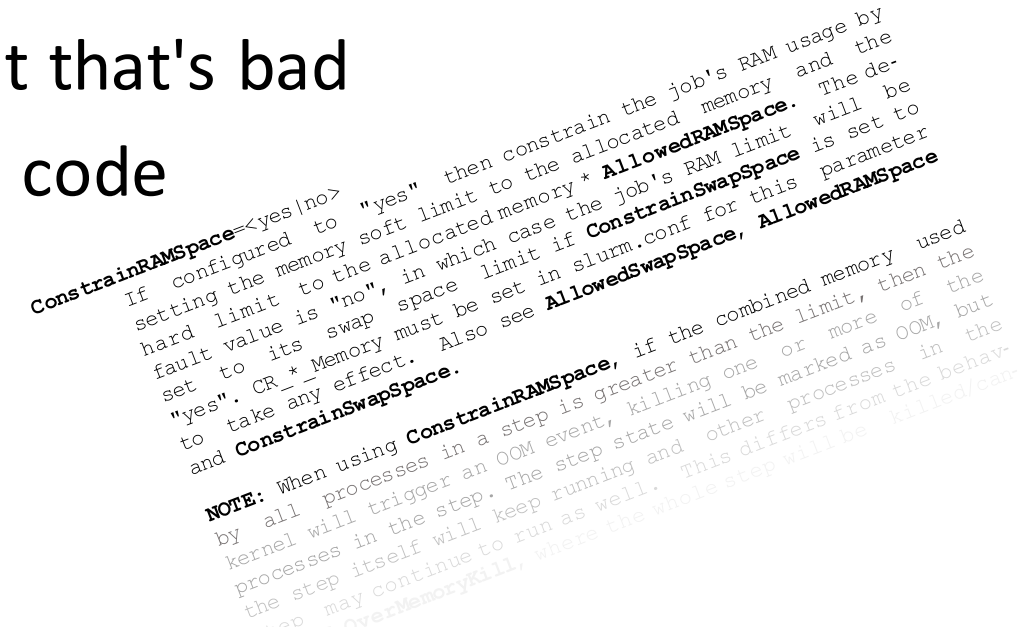
- Non-locking power cables
  - Particularly bad towards PDU. Cables would disconnect when working near-by cables
  - Solution – new locking cables will be delivered

Aigert and ARC cache

- Use the same ARC cache as before, but with 100Gb/s optics
- But old kebnekaise-ce still in production, so must do staged testing
- Picked q-h36(!) as first deployment victim
  - Weird network issues
  - Also weird host behavior
  - Turned out we had two separate issues
    - q-h36 (still?) broken
    - 100G optics didn't work with the NIC
- Using another cache and different 100G optics from a later purchase worked
  - Purchased from Direktronik via ATEA
    - Needed to argue for quite some time with ATEA for them to even contact Direktronik to handle return.
- As for q-h36; riser-card replaced with no difference, now waiting for new network card. The story continues...

Slurm for Aigert

- OOMs on some test jobs! Should be no OOMs on the test jobs!
- Change slurm-config to not be so strict with memory limits
- Should be easy, eh! `ConstrainRamSpace=no`
  - No! Still memory limit on cgroup
- Completely turn off cgroup worked, but that's bad
- Read syslog, find nice text, grep source code
  ⇒ `ConstrainSwapSpace=no`
  - Success!



WLCG compute current status

- Because of two ARC cache machines moved to Aigert, drained half of old kebnekaise-ce nodes
- g-ce01 seems to be running test jobs fine. No local error visible
- Only 2 nodes are active. 1 ARC cache machine is not enough!
- Todo!
  - Install the extra PDU, move all power cables
  - Decommission kebnekaise-ce from production
  - Move the rest of the ARC cache nodes to Aigert
  - Move g-ce01 to production ...

How do one get a CE in production?

- What is the procedure to get a CE into NDGF nagios and NDGF dashboard?
- What is the **actual** procedure to get a cluster into PANDA?
  - Felt like a lot of handwaving
  - Difficult to have two
- Having access to CRIC seems to be critical
  - That fact is undocumented
  - How do one get access to CRIC?

UMU compute - Kebnekaise

- Still Kebnekaise, but split into 5 partitions
  - Largemem (broadwell, 3TB)
  - GPU (skylake)
  - Normal batch (skylake)
  - 2 AMD partitions, one for pure CPU, one for GPUs
- New cluster HW ordered, Lenovo
  - 10 dual L40s nodes, 2x24cores AMD, 384G memory, HDR IB
  - 8 2x128core AMD, 768G memory, HDR IB
  - 2 4xH100 HGX, 2x48core AMD, 768G memory, HDR IB

Network

- New SUNET gear in place
- UMU (finally) moving towards a 100G based core network
- HPC2N (finally) will get its 100G core switch
  - With 100G uplink
- LHCOPN 100G
  - In production since autumn (late September 2023)
  - Now that new SUNET gear is in place we're planning on moving to a 2x100G trunk to have less panic if there's fibre/optics issues.

Tape/backup

- Not same as last time
- IBM TS4500 library (2550 slot capacity)
- 6x TS1155 tape drives (JD tapes, 15T, 360 MB/s)
- New: 6x TS1170 tape drives (JF tapes, 50T, 400 MB/s)
- New: Dell R750
  - 2x100G Ethernet
  - 4x32G FC
  - Approx 30T NVMe for DB and incoming stgpool
  - A few TB of SAS SSD for log mirrors etc
  - Approx 250T spinning disk for on-disk backup storage
  - 256 G RAM, 2xIntel Gold 6334 (total 16 cores @ 3.6 GHz)

Tape/backup (2)

- Tape technology upgrade order, try 1
  - Purchased via server/storage agreement
  - 12xIBM TS1170 tape drives
  - 500 tapes that fit
  - FC/SAN switch
  - Result: Got 12 new tapedrives and 500 tapes that fit our old drives
  - Mistake at config time by distributor, apparently JF tapes are way more expensive than JD tapes to that's why we got everything cheap.
  - Quirk: WLCG money would expire by end of year so couldn't argue for supplier to own their mistake indefinitely.
  - Ended up cancelling this order and retry-ish.

Tape/backup (3)

- Tape technology upgrade ordered, try 1,5
  - 6xIBM TS1170 tape drives
  - The 500 JD tapes that we already got
  - 200 JF tapes with different labels
  - FC switch
  - Result: 200 JF tapes with same labels as the JD tapes we got
    - Quirk: For 3592/Jaguar the default is 6 char labels ie without the format letters.
  - After some time they actually sent new labels and an IBM technician to relabel all tapes...
  - In production since March
    - New generation tape used only for WLCG

Incidents - Squid

- On 2024-04-11 a security update of squid 4.10 was auto installed.
- This made our CVMFS squids to start failing on startup.
  - No software for local Kebnekaise users, no CERN-stuff for the WLCG compute.
- After an hour or so of investigating we found out that one of the CVE fixes had broken things severely.
- We solved it in two ways
  - One squid instance got upgrade to old 5.2 version, 3 days later they had fixed it and now back to the 4.10 series on Ubuntu 20.04 Focal.
  - Other squid instance, got a distro upgrade from focal to jammy

Incidents – SSH gssapi-keyex patch mismerge

- We noticed that sshd was spewing error messages in syslog about gssapi-keyex logins from root but still succeeding with gssapi-with-mic.
- Gssapi-keyex is not part of upstream openssh, but an addition that Debian/Ubuntu is providing. During the change from openssh 8.2 on focal to 8.9 on jammy they had not done their job properly.
  - The struct for handling these extensions, in the upstream openssh, had changed from 3 to 4 elements, with the new element in position 2 (i.e. not last). However, the patch for gssapi-keyex still used 3 elements...
  - Long story short, this has now been fixed and new tests added to Debian/Ubuntu builds of openssh to verify gssapi-kexec functionality.

Misc

- Internal wiki migrated to mkdocs
  - Magnus-ninja-converter-script plus manual work
- Center storage – Lustre
  - ExaScaler and Insight upgrade being planned, probably after summer
  - Stability has been good, only one two minor interruptions since last SONC

Staff

- Niklas Edmundsson – Tape storage
- Mattias Wadenstein – NEIC NT1
- Erik Andersson – Cluster/Storage/SSC
- ~~Magnus Jonsson – SUPR/SAMS~~
- **My Karlsson** – SUPR/SAMS
- Roger Oscarsson – ARC/WLCG
- **Paul Dulaud** – DDLS developer
- **Abdullah Aziz** – DDLS developer
- Lars Viklund – App Expert
- Åke Sandgren – App Expert
- Birgitte Brydsö - App Expert
- Pedro Ojeda May – App Expert
- Björn Torkelsson – Tech director
- Lena Hellman – Administrative asst.
- Mikael Rännar - WLCG
- Jerry Eriksson – Advanced Consultant
- Paolo Bientinesi - Director