HPC2N site round

NDGF All Hands 2019-1, Ljubljana

Power outage 2018-11-12

- Almost city-wide power outage 14:05 - 14:31
- Never got a good explanation from Umeå Energi

Power outage 2018-11-26

- Campus-wide outage 09:41 - 13:10 (HPC2N UPS)
  - Approx 1 hour down according to official sources
- Excavator meets high-voltage campus feed cable
  - Reportedly the excavator didn’t hit the cable, but the frozen ground lifted included the cable...
- UmU campus do have dual feeds, but it’s a manual switchover arrangement
- Took quite a while to isolate/reconnect to enable power from secondary feed
  - Mariehem station
- Universitetet station is being replaced with a new one
  - Needed to cope with the upgraded 145 kV regional network
  - All cables from the old station is being dug up and rerouted
  - Main feed for UmU campus

Cooling outage 2019-01-30

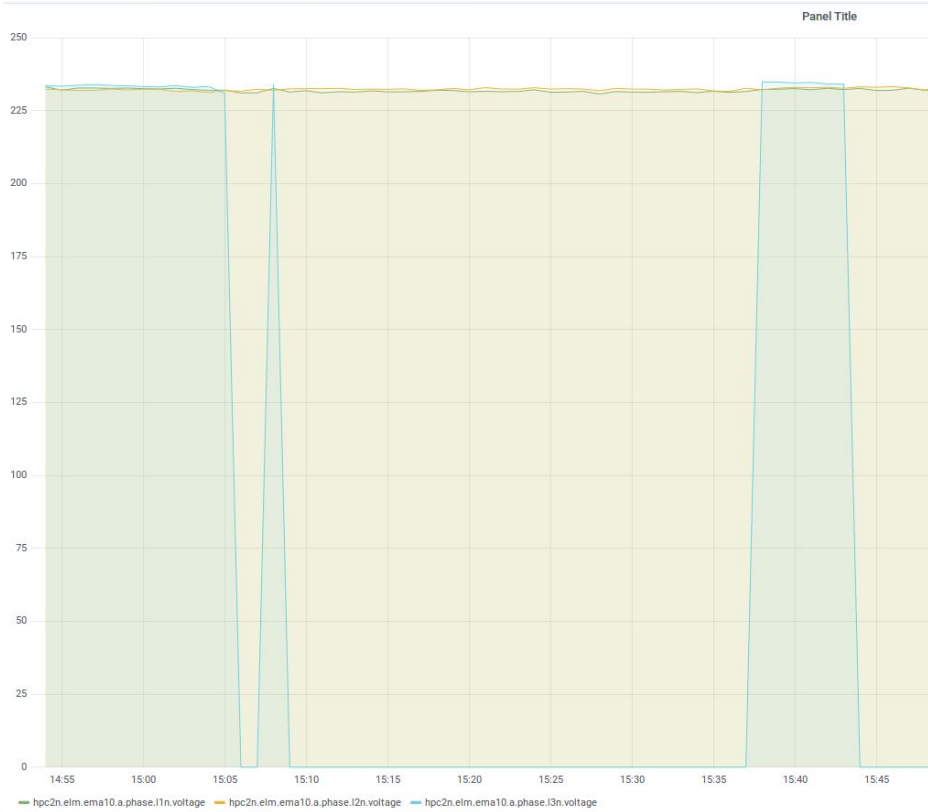
- Lost cooling at 19:20
- Cooling restored 1-2 hours later
- Abisko back in production by 22:40
- Kebnekaise faced IB/Lustre issues, back by 17:40 next day
- Same cause as last time: We got warm water on the cold water feed
- Root cause: On-roof chillers activates to compensate for less heating
  - Control systems behave different after hw upgrade
    - Same program, go Siemens!
    - Valves and circulation pumps start, but not the fans...
- Workaround: Disable on-roof chillers, only enable with supervision during office hours
- A different firm has been hired to sort out the PLC programming/issues

Power outage 2019-03-13

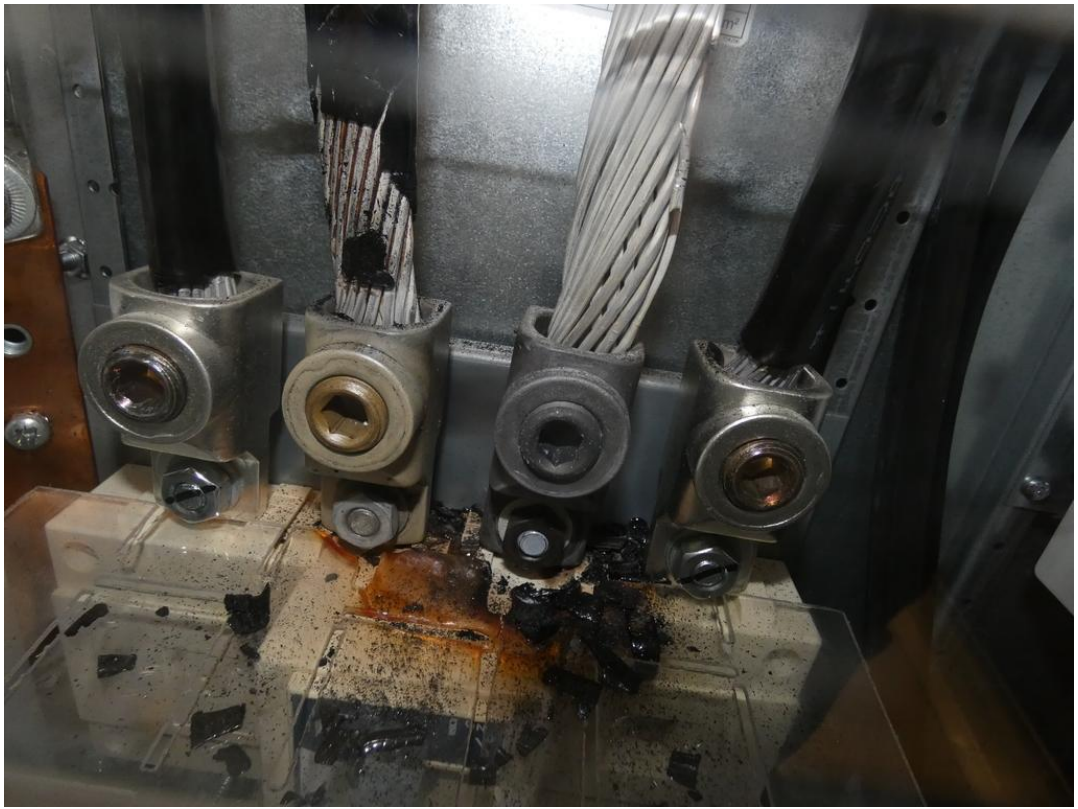
- Campus-wide outage 09:28 - 09:49
- Caused by excavator finding cable in Anumark, about 6 km from Campus
  - Connected to Mariehem station
  - Umeå Energi trusted map/GPS instead of measuring the true cable location...
- Campus was supplied from secondary feed due to work on campus and primary feed, due to previous unplanned outage due to excavator
  - Ie, powered from Mariehem station
- Tripped ground fault protection inside campus high voltage network
  - Could likely have been prevented by direction aware protection units
  - Campus power network was initially built for a single feed
- Now we’re running on the primary feed from Universitet station again
  - Awaiting excavator?

Power distribution board failure 2019-03-15 (1)

- Our Kebnekaise main guy noticed multiple nodes with power issues
- Voltage graph for A2A telling
  - One of six power dist boards
- Main facility guy gone for the (fri)day
  - On-call guy still on campus, decides that it’s not normal
- Bravida electrician called on site
  - Judges the distribution board not fit for duty and cuts power



Power distribution board failure 2019-03-15 (2)



Power distribution board failure 2019-03-15 (3)

Power distribution board failure 2019-03-15 (4)

- Main switch needed replacement
  - In stock
    - Though not at Garo in Sweden, but at their supplier in France…
    - But of course the delivery time they stated was to Garo warehouse, not to us in Umeå.
- Supply cable needed replacement
  - Insulation melted
  - Easier to run a new cable the few meters needed than fit a junction
- Internal cables from main switch L3 also needed replacement
- Power restored on 2019-03-28 09:58
- All nodes had normal power routing restored the day after
- Thermography done of all electrical distribution cabinets
  - Did not reveal anything urgent
  - Not without comments though, can wait until next full stop

Power distribution board failure 2019-03-15 (5)

- No cluster equipment damaged
  - In general not an issue during normal power outages
- Abisko switches do not have redundant power
  - Main IB switch rack was of course connected to the affected distribution board
  - All running jobs on Abisko were lost
- Kebnekaise does have redundant power supplies on everything
  - While good for networking, proved to be cumbersome for nodes

On compute nodes with redundant power

- Abisko nodes have a single PSU
  - Failure mode: If some power is gone, some nodes die
  - Only affects jobs running on the affected node
  - New jobs won’t start
- Kebnekaise has chassis with multiple PSUs and redundancy
  - Failure mode: If some power is gone, a lot of nodes throttle
  - Affects a whole lot of jobs in all affected chassis
  - When old job hits the walltime limit, new jobs start
- A time consuming task during this event was to figure out which Kebnekaise nodes to turn off in order for the rest not to throttle

Linux kernel update breaks MOFED/NVidia

The upgrade of Ubuntu Xenial kernel from 4.4.0-142 to 4.4.0-143 included a change to get_user_pages() that made both Nvidia and MOFED (3.3) kernel drivers configure fail to detect the correct number of arguments. (MOFED 3.3 didn’t even check for different number of arguments).

A quick analysis gave at hand that upgrading to the HWE kernel (4.15) and MOFED 4.4 would solve the problem. And since we already had this combination deployed on the Grid nodes of Abisko it was an easy decision.

However, this required us to recompile all versions of OpenMPI and UCX on all architectures and make a migration with the system still in production…

2.5 weeks later we’re finally done.

HPC2N /pfs fs Lustre breakage

- On 2019-04-04 Lustre OSS (servers) started rebooting due to LBUGs.
- Caused by recent upgrade of lustre client versions.
- Triggered checks in lustre servers that previously hadn’t run.
- Workaround found and put in place.
- Sort of resolved on 2019-04-06 (cluster queues resumed).
    - LNET routers back-leveled to fix another problem that showed up at the same time.
- Some (in-flight?) end-user files were corrupted.

Tarpooling

- In progress for disk pools
  - brytbas test pool
  - n-x02 production pool reinstalled successfully without wiping data
  - The rest to be handled during one/multiple downtime(s), likely April 16th or 17th
- Tape pools?
  - Need to define who does what
  - Nikke/Maswan uses the HPC2N tape pools for early production ENDIT testing
  - If NDGF takes on ENDIT/tape responsibility the tape system monitoring skills needs to improve to detect performance/behavior issues