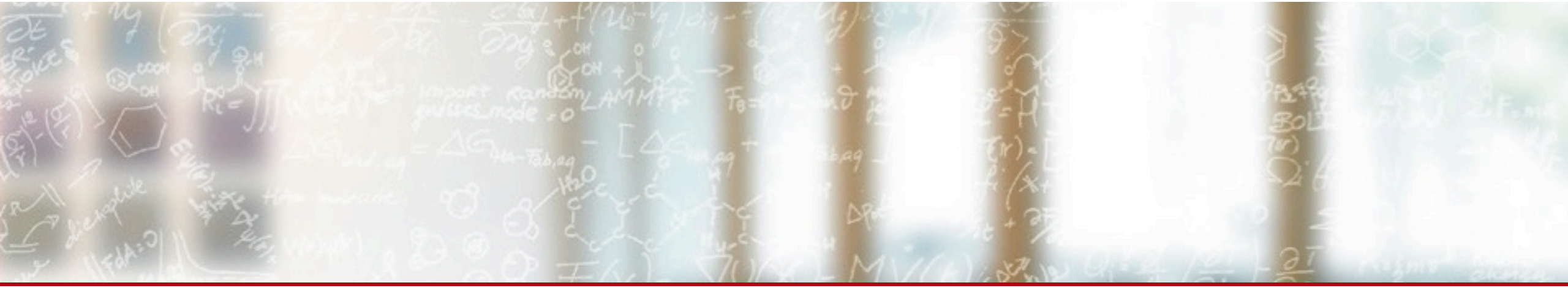
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



LHConCRAY

Acceptance Tests 2017 – Run6 System Report (Sep 01 2017 – Oct 01 2017)

Miguel Gila, CSCS

October 02, 2017

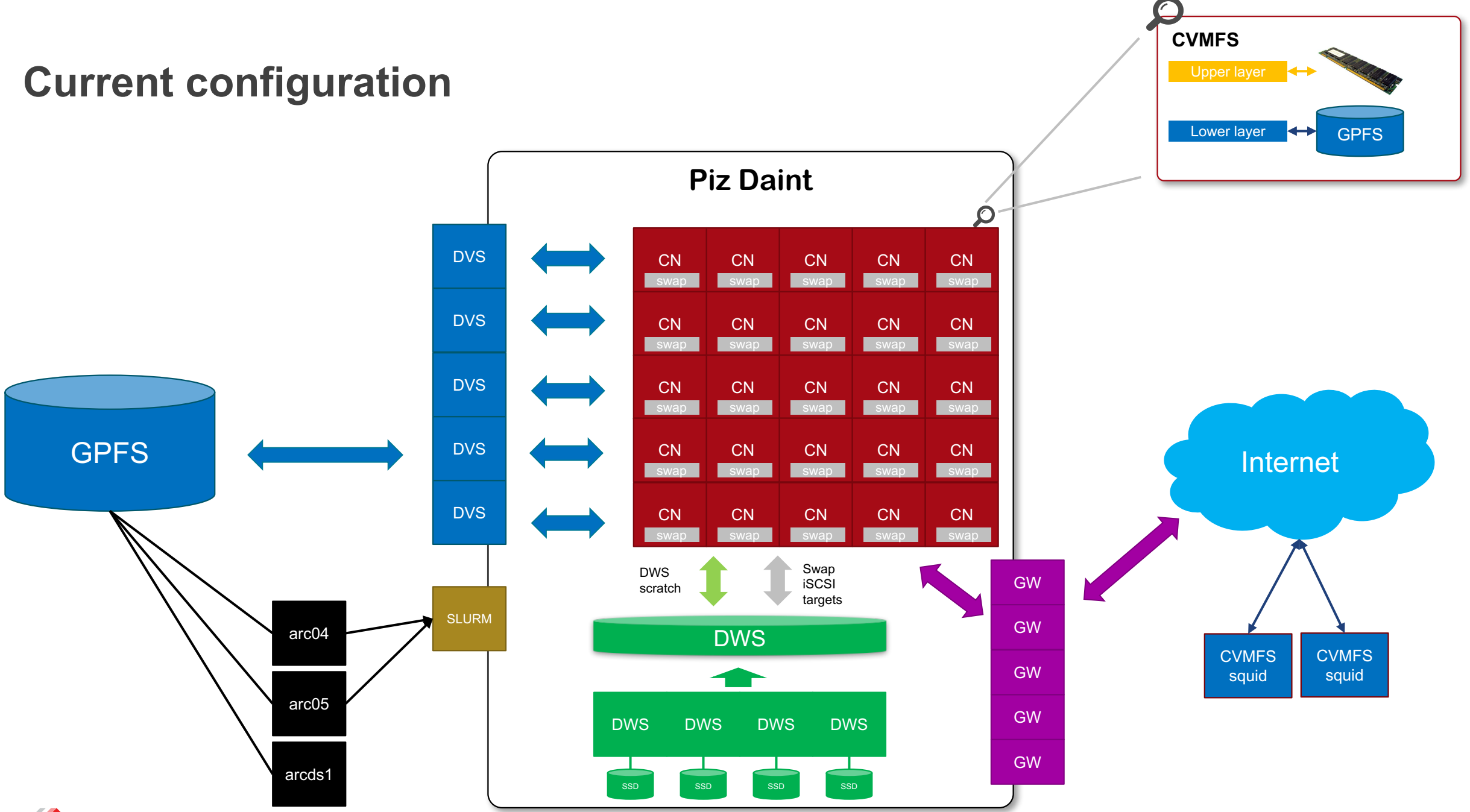
Current configuration

Current configuration

Allocating 68-core per node instead of 64!

- 25 compute nodes (72core, 128GB RAM, diskless)
- 1 production ARC + 1 data staging ARC + 1 test ARC (internal)
- Dedicated GPFS filesystem shared with Phoenix
- 5 DVS nodes exposing GPFS to CNs
- CVMFS running natively on CNs using tiered cache and workspaces
  - Upper layer: 6 GB in-RAM shared to all experiments
  - Lower layer: preloaded cache on GPFS, mount on CNs RO with caching enabled
- Memory limits NOT really enforced
  - Hard limit of 6000MB/core to catch rogue jobs
- Swap on DataWarp enabled (64GB per node)
- ARC caching not enabled (ATLAS)
  - Each job has a copy of the files, even if they're the same on multiple jobs.
- Arc delegation database converted to sqlite for better performance (22.08.2017)
- New blackhole detector that will drain a node if more than 5 jobs have failed in the last 5min, but only if more than 10 jobs finished in the period (28.08.2017)
- Improved node auto-drain mechanism to be more permissive and drain less often (28.08.2017)
- CLE 6.0 updated to .UP04 release on (27.09.2017)
- DataWarp tested as /scratch

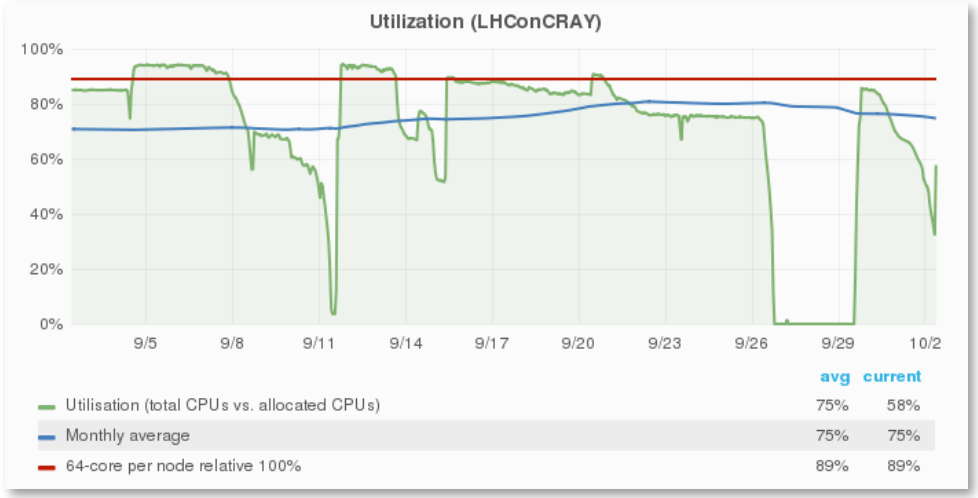
Current configuration

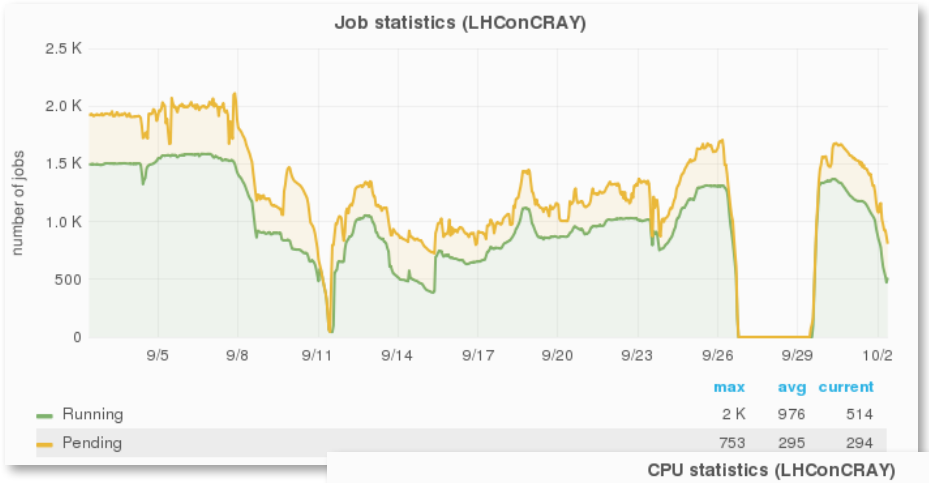


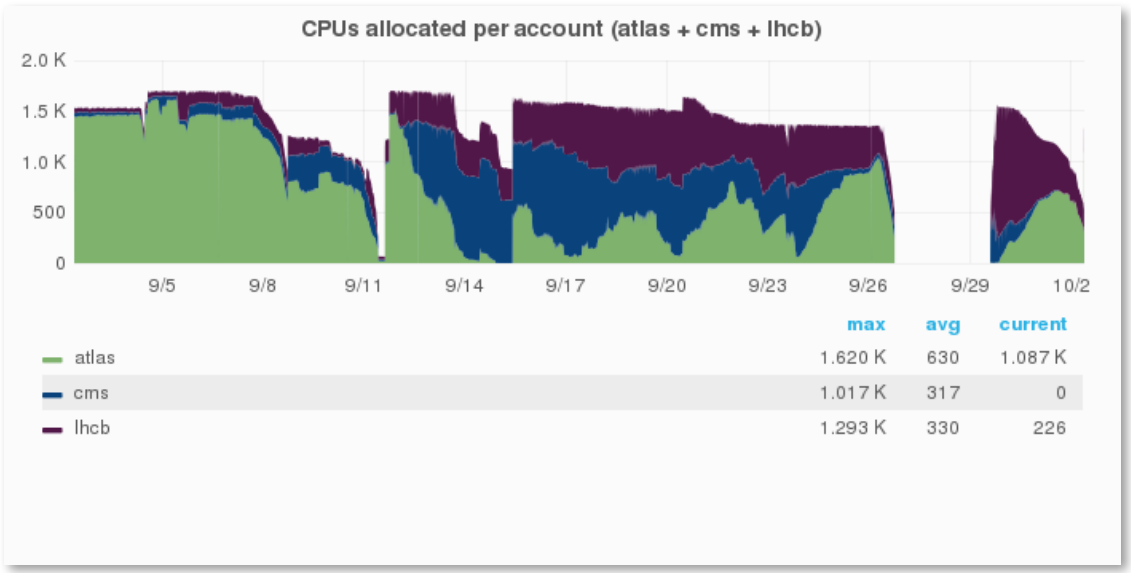
System statistics & report

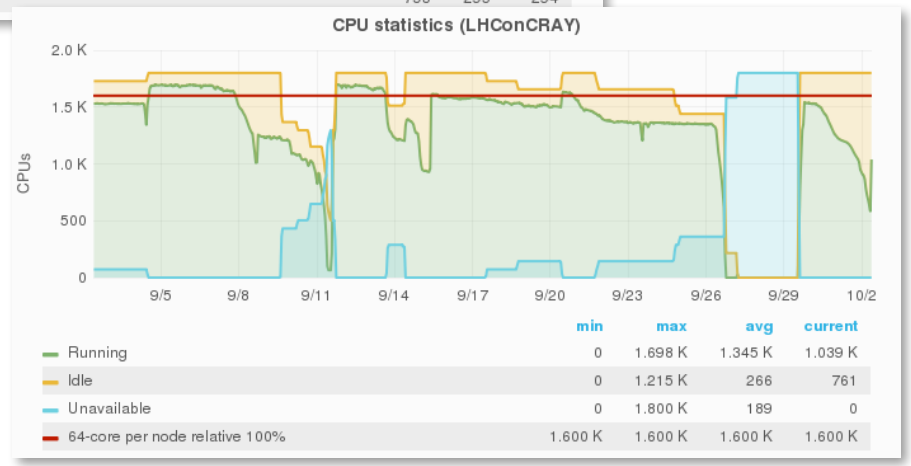
System utilization and issues

- High utilization during the period
- VO distribution seems to start stabilizing
- Intermittent issues with some nodes due to timeouts in the health checks









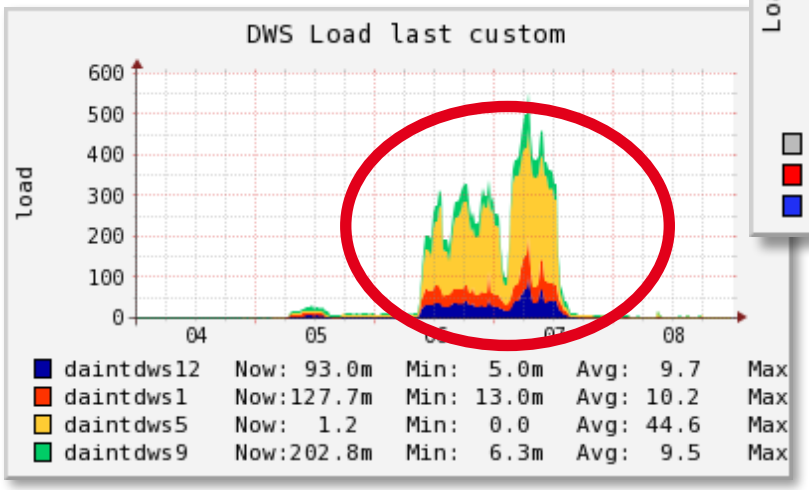
Node statistics

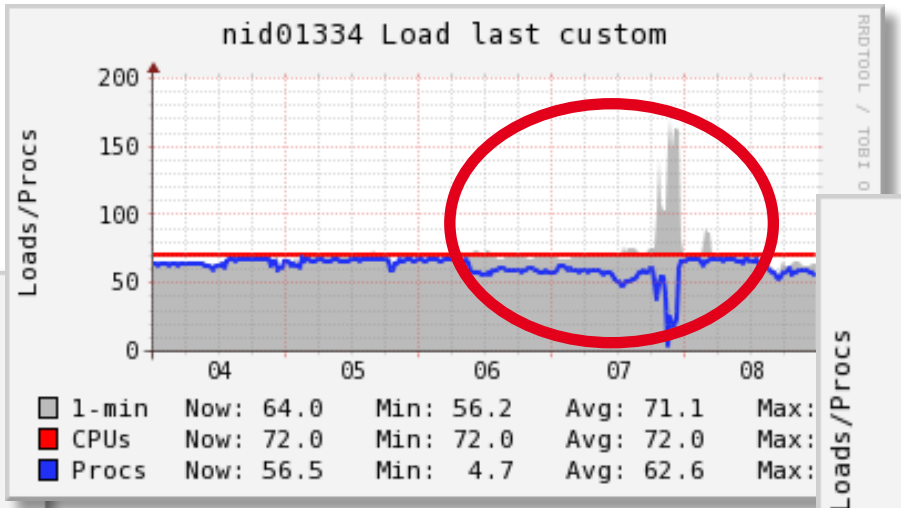
- Load
  - Number of procs in line with load
  - Some load peaks due to IO
- CPU utilization
  - Very variable, likely due to storage (dCache) or VO-Boxes
- Memory utilization
  - SWAP never used
- Network
  - No significant activity

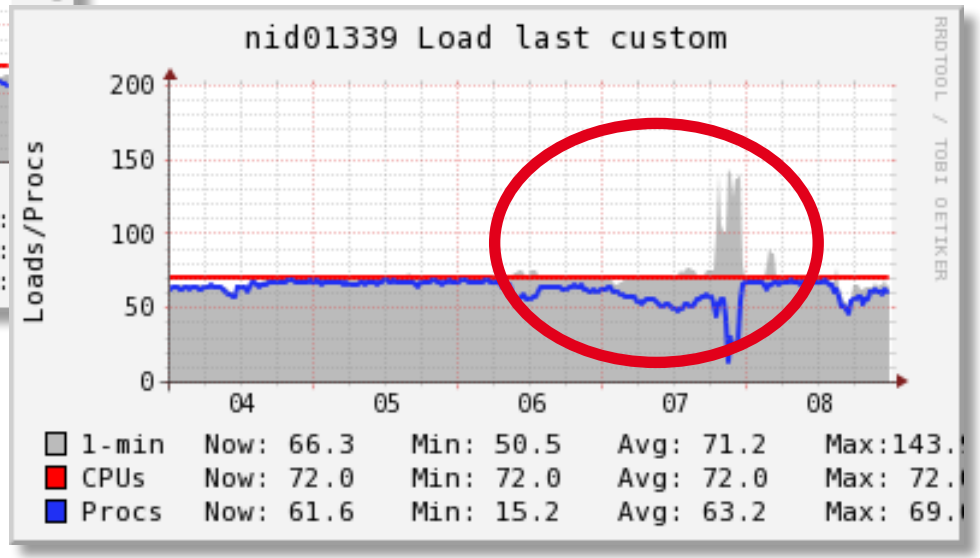
DataWarp Report

However, other configurations are possible and we are actively looking into all of them!

- DataWarp does not seem to be a viable candidate in its current configuration for LHConCRAY jobs:
  - *Using static allocations leads to single-server metadata IO problems*
  - Using multiple allocations (dividing nodes in blocks) is very complex to implement
  - Using Datawarp and Slurm burst buffer directives seems to have a bad effect in the scheduling of jobs



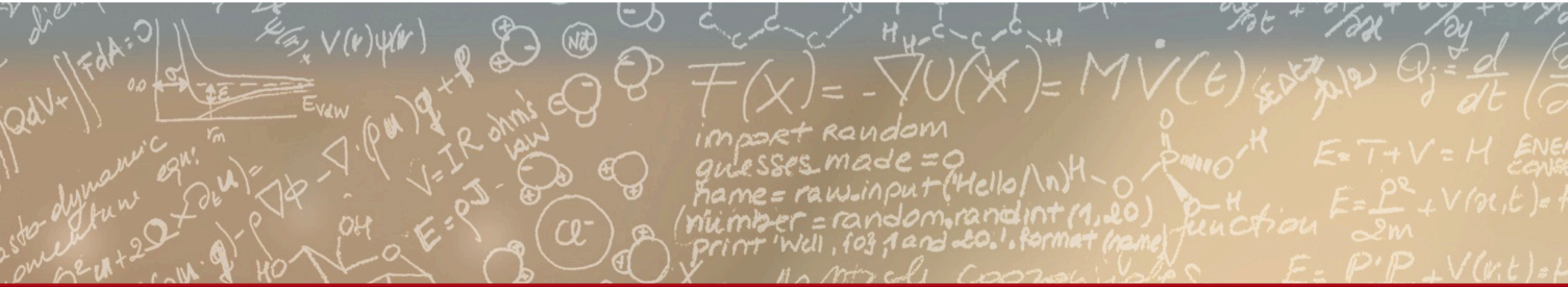




Report

- Overall utilization reaching relative maximum
- Swap not really used so far
- There seems to be room to allocate more cores/node
- CVMFS in RAM seems to work quite well, not a single issue in the period
- DVS and node load high at times due to IO
- ATLAS and CMS have picked up CPU hours to LHCb
- Issues affecting Phoenix (dCache, CVMFS, VO-Boxes) also affect Piz Daint. This reflects in CPUs allocated, but processes waiting and not really using the CPU

Piz Daint	ATLAS	492‘955	51.35%
Piz Daint	CMS	238’614	24.85%
Piz Daint	LHCb	228’466	23.80%
Piz Daint	TOTAL	960’035	



Thank you for your attention.