ATLAS Sim@P1 Upgrades During Long Shutdown Two

F Berghaus[1], F Brasolin[2], A Di Girolamo[3], M Ebert[1], C Leavett-Brown[1], C Lee[4], P Love[5], E Pozo Astigarraga[3], DA Scannicchio[6], J Schovancova[3], R Seuster[1], R Sobie[1]

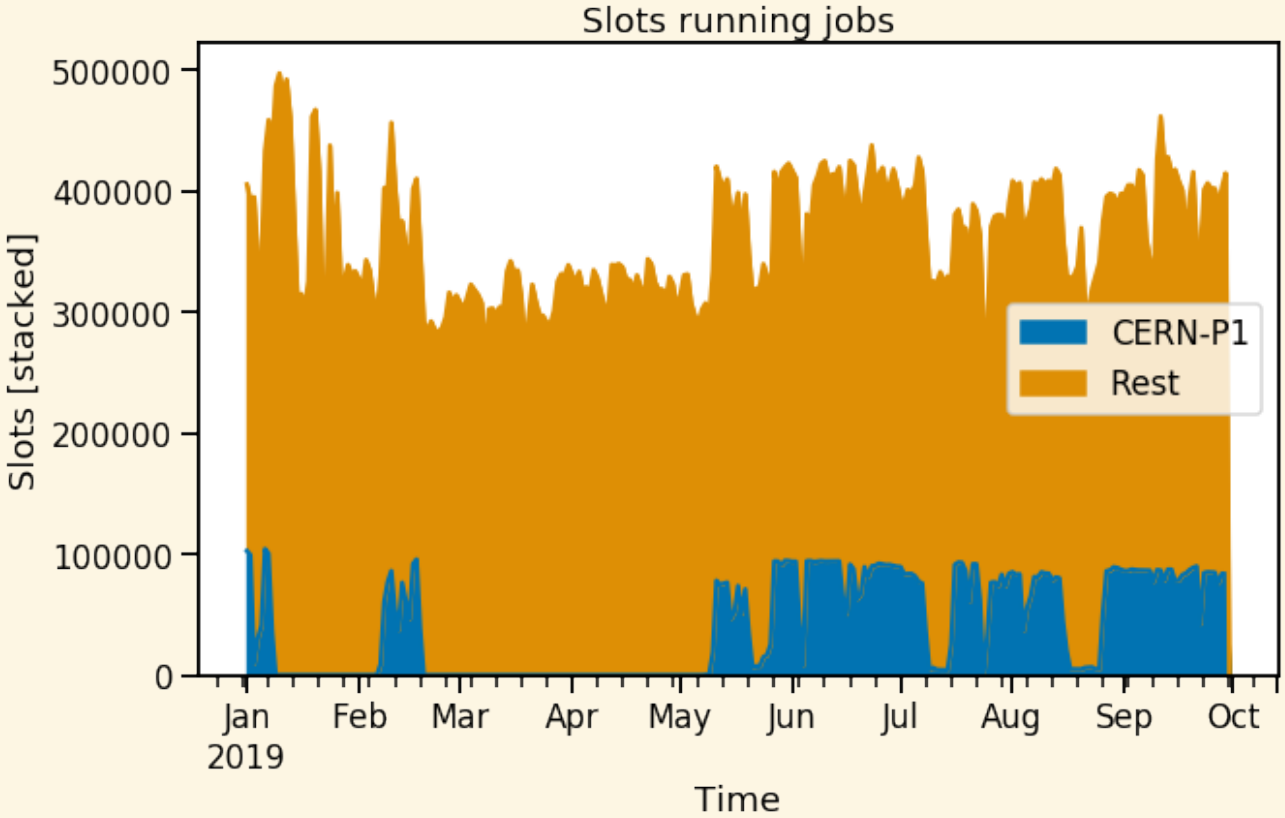
1. University of Victoria [CA]
2. INFN Bologna [IT]
3. CERN
4. University of Cape Town [ZA]
5. Lancaster University [GB]
6. University of California Irvine [US]

Simulation at point 1 [Sim@P1]

- Opportunistic usage of the **ATLAS** Trigger and Data Acquisition [TDAQ] High Level Trigger [HLT] for offline processing

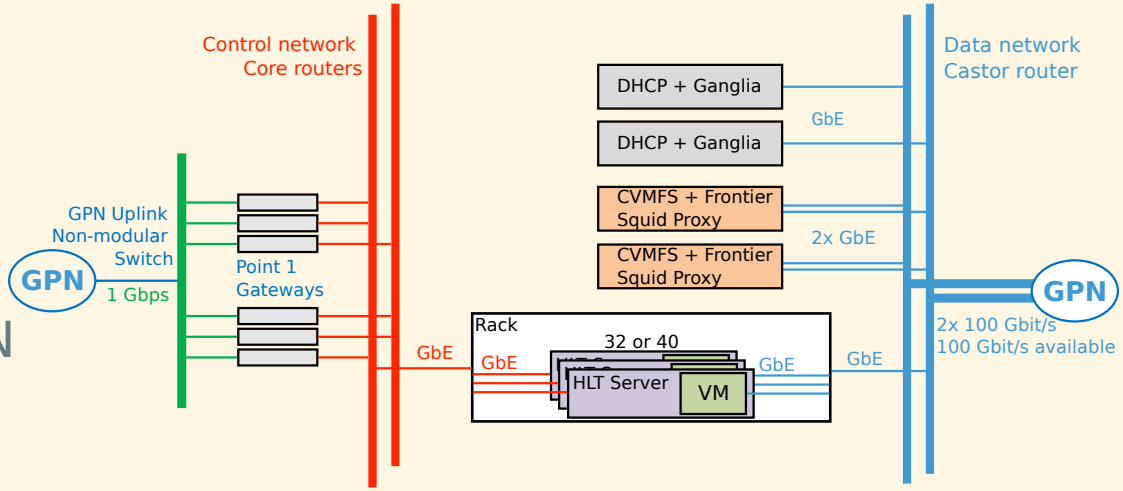
- When the experiment is not taking data and there are no other TDAQ activities, for example:
  - long shutdowns
  - technical stops

- 2.5k servers with 95k cores



Switch between HLT and Sim@P1

- Isolate offline environment from detector control:
  - VLAN:
    - On data network
    - Limited access to CERN GPN
  - Virtual machines
- Machine reconfiguration
  - QEMU creates ephemeral disk:
    - 20GB per core (max 90% disk)
  - Libvirt boots virtual machines with
    - CernVM image
    - Config ISO
    - Ephemeral disk



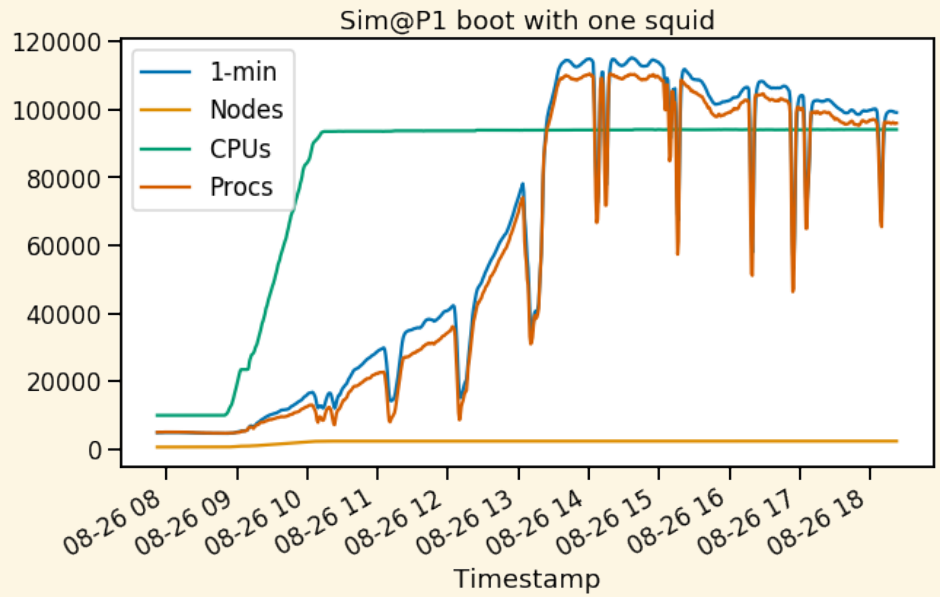

```
<disk type='file' device='disk'>
  <source file='/dsk1/sp1/ephemeral/disk.local'/>
  <target dev='hda' bus='ide'/>
  ...
</disk>
<disk type='file' device='disk'>
  <source file='/dsk1/sp1/permanent/cernvm.hdd'/>
  <target dev='hdb' bus='ide'/>
  ...
</disk>
<disk type='file' device='cdrom'>
  <source file='/dsk1/sp1/permanent/config.iso'/>
  <target dev='hdc' bus='ide'/>
  ...
</disk>
```

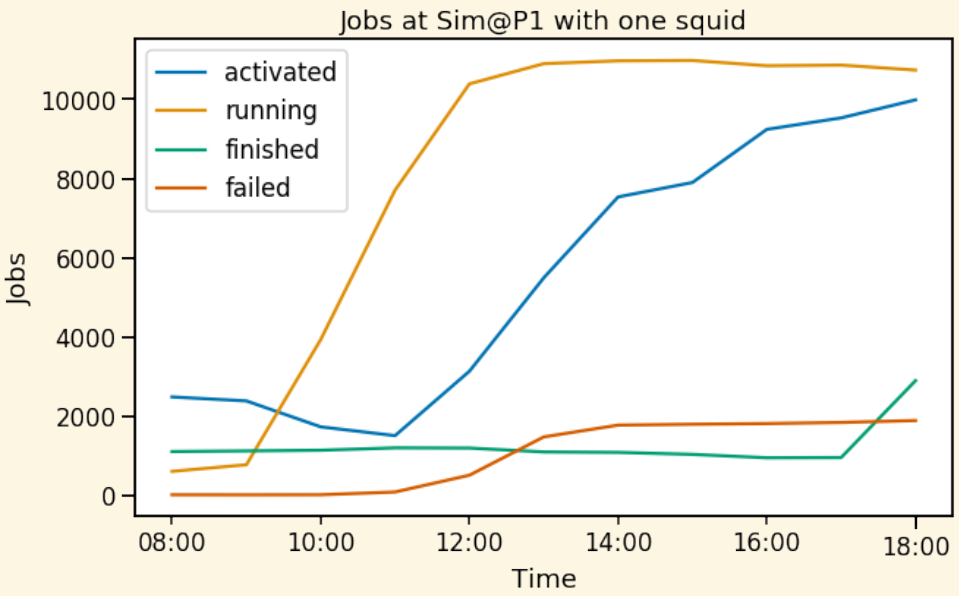

CernVM
Micro


Config
Disk


Disk

Switch between HLT and Sim@P1





1) Shifter switches racks
   - Old racks always in Sim@P1 mode
2) Reconfiguration runs within one hour
   - Takes 15 minutes
   - Sim@P1 -> HLT puppet is run immediately
3) Virtual machines advertise to HTCondor
4) Resources receive jobs
5) CVMFS caches in needed software
6) Payload begins running

Dedicated services (managing 100k cores)

- @P1:
  - DHCP : `pc-sp1-ganglia-01`
  - Monitoring

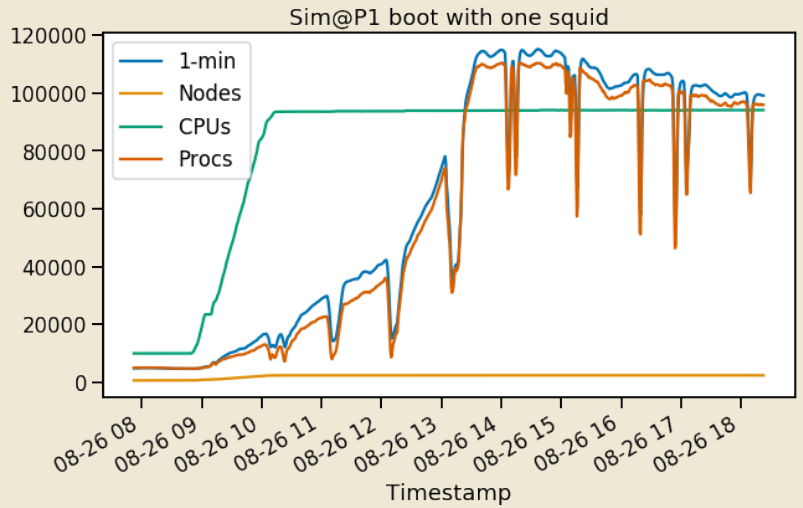
    `pc-sp1-ganglia-01` `pc-sp1-ganglia-02 [off]`
  - Squid cache for CVMFS and Frontier

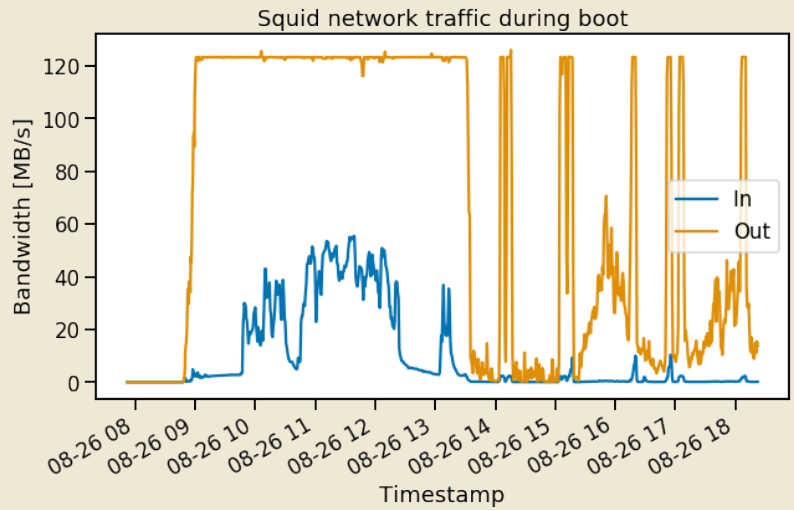
    `pc-sp1-front-01` `pc-sp1-front-02`
- @CERN
  - HTCondor

    `simatp1-cm` `simatp1-sched0[1-4]`
  - Harvester (CERN_central_0)

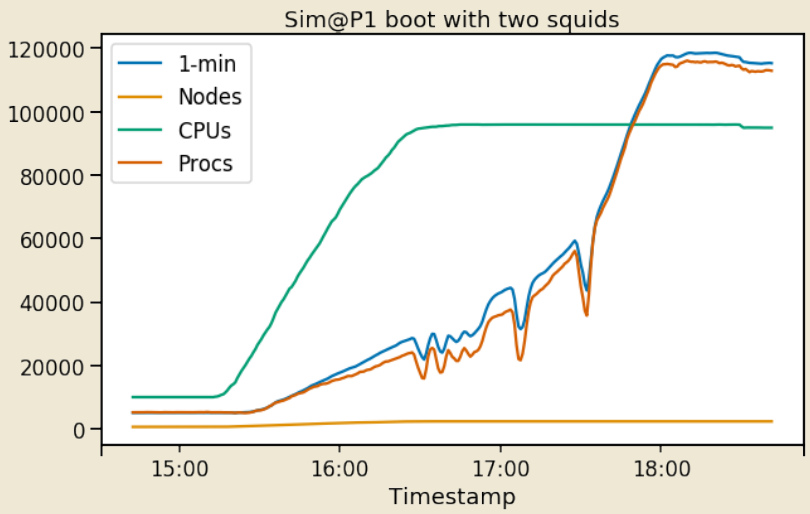
    `aipanda175`

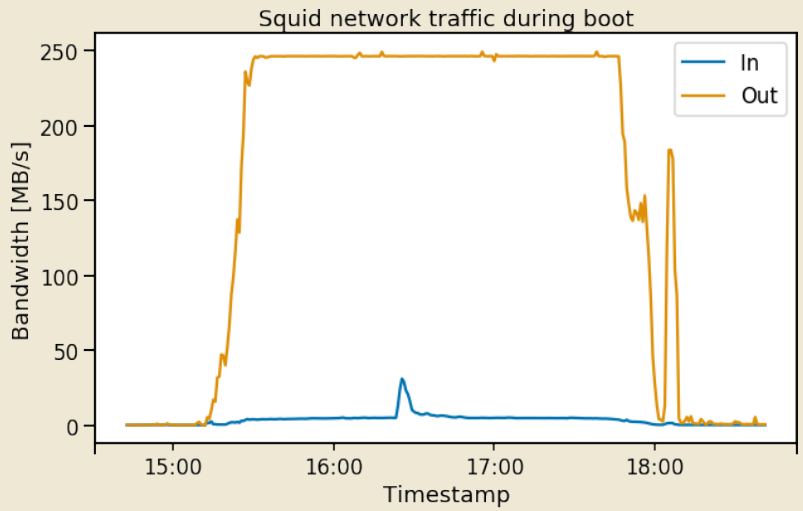
Effect of squid performance





One Squid: ~5h





Two Squids: ~3h

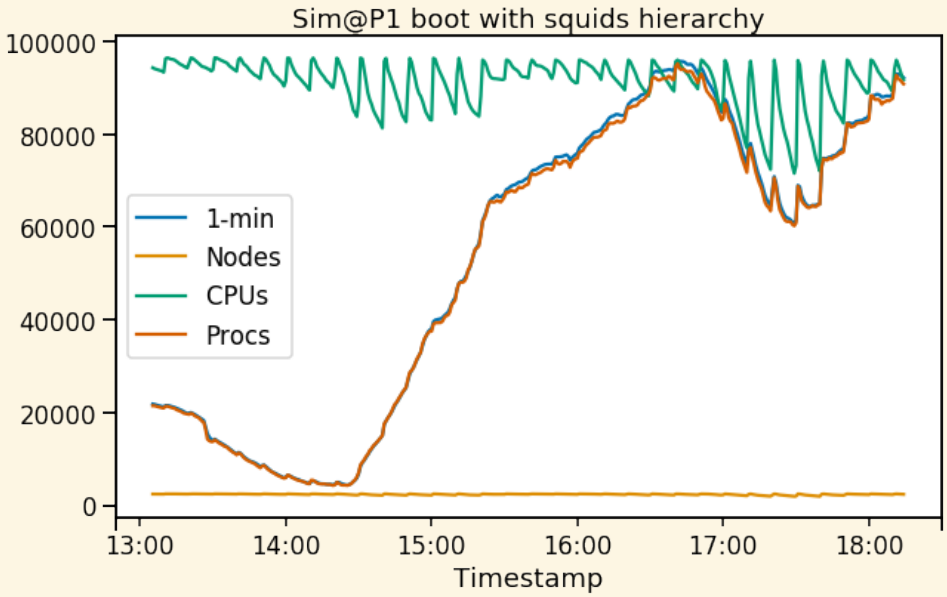
P1 in 2019

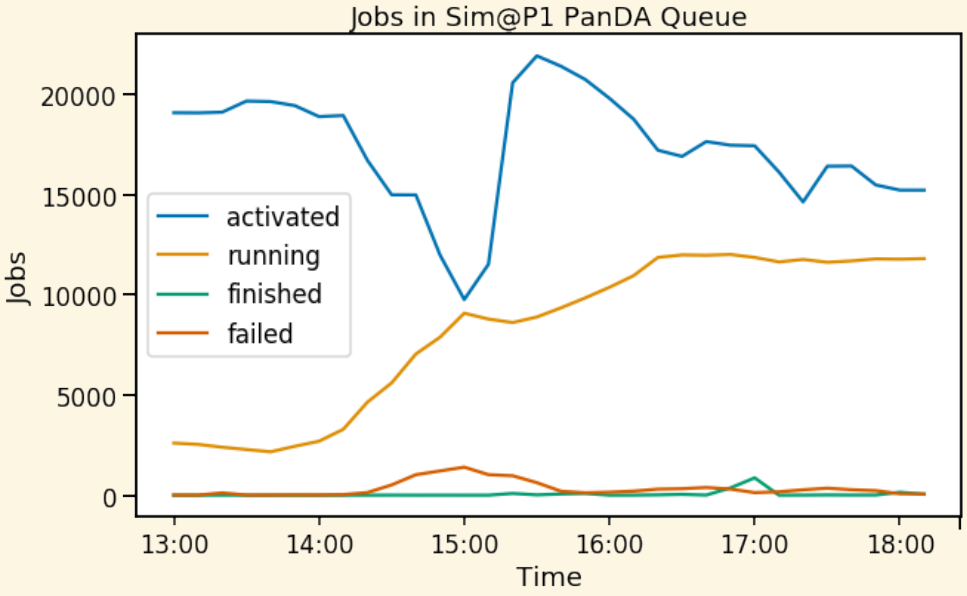
Use Sim@P1 on in long breaks between LHC fills:

- Improve turn-on performance (future goal <30 minutes)
  - Persistent CVMFS cache *requires*: xxGB / server
  - Squid hierarchy *requires*: 2 CPUs and 4Gbyte memory per squid
  - Replace old squid hardware requires*:* money 
- Short running jobs: event service

Squid hierarchy

- Use two server in each rack as squid
  - From different chassis for resilience
- Caveats:
  - Use Web Proxy Auto Discovery to find squids
  - VMs wait to boot until at least one squid is up
  - Squid servers use central P1 squid
  - Squids in rack are siblings
  - Central squids are parents



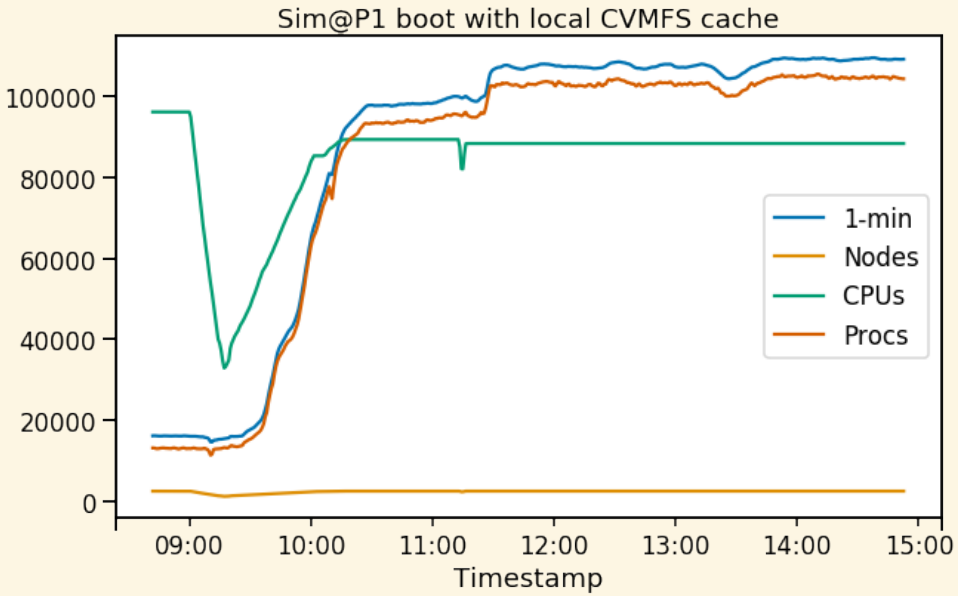


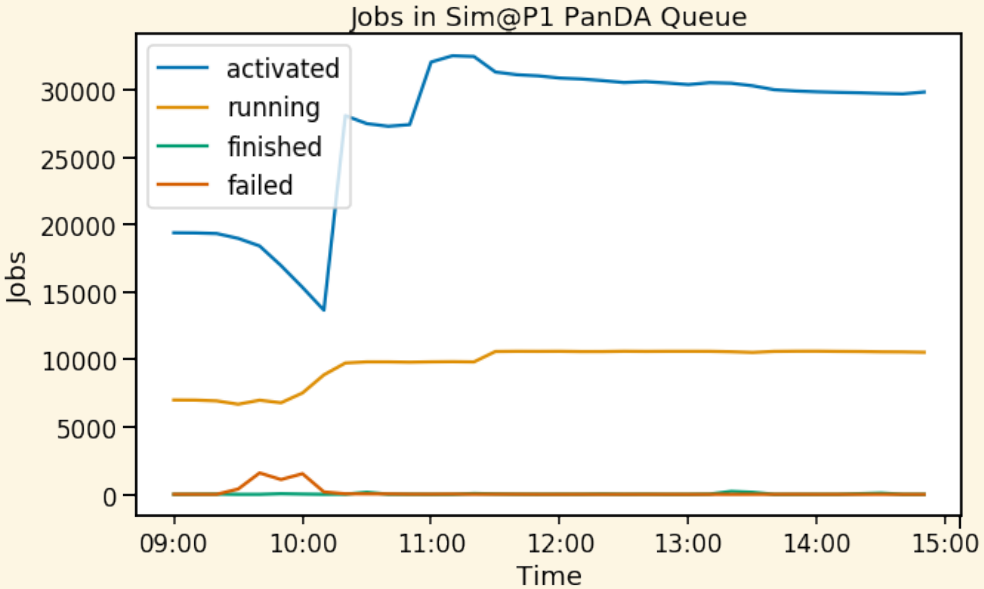
Persistent CVMFS cache

Method:

- If not there create a persistent virtual disk
- Format disk with label cache
- Mount partition by label CVMFS cache

- Many system files and ATLAS software releases already present on boot





Sim@P1 in 2019 so far

TDAQ HLT updated to CC7

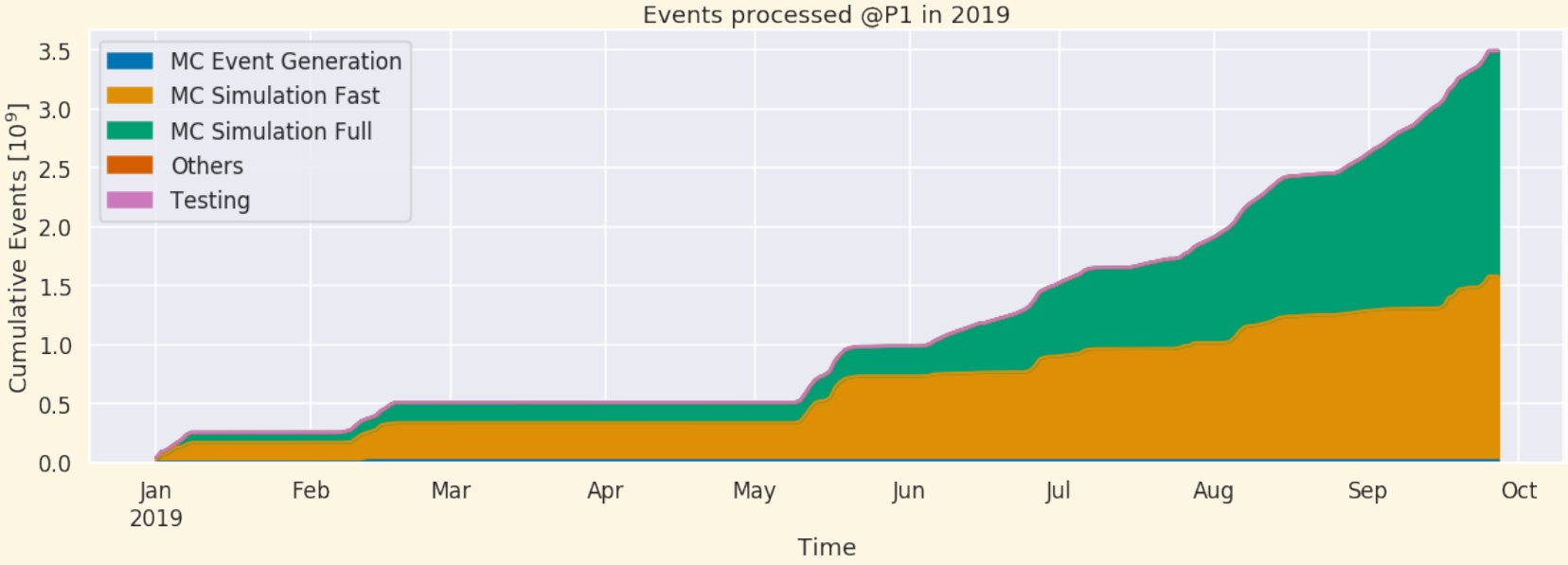
Switching from OpenStack to qemu and `libvirt`

- Faster farm switch over
- Enhanced reliability

Can run 90k cores with a small team

Improve turn-on efficiency

Study more I/O production intensive work flows @P1



Credits

- Solarized colour scheme by Ethan Schoonover
- Original compact disk image by Sakurambo