ATLAS SCALE-UP TEST ON PIZ DAINT

Gianfranco Sciacca

AEC - Laboratory for High Energy Physics, University of Bern, Switzerland

LHConCray WG - 7 November 2017

PREPARATION

- **Panda queue created: CSCS-LCG2-HPC_MCORE_TEST**
  - **pointing to arc04.lcg.cscs.ch, queue atltest, corecount=16**
  - **reservation with 11 nodes set up by Miguel**
- **Sent some HC jobs, helped Miguel tuning the ARC conf**
- **Simulation task created by ATLAS: https://bigpanda.cern.ch/task/12491843/**
  - **4M events, 40k input files, up to 148MB/file (mostly 115MB)**
  - **jobs tuned to ~1h duration (maxEvents=100)**
  - **ramCount=900 MBPerCore**
  - **Output expected: ~70MB/job**
- **Started submitting jobs, 2 Nov at 4PM**
  - **jobs accounted to be using up to 32GB of mem and got killed**
  - **load spike on GPFS**
  - **removed memory limits, jobs started running**

PREPARATION

- **Identified an issue with the ARC infosys:**
  - **jobs of the wlcg partition were published correctly, jobs of the atltest partition were not.**
  - **This would break submission from the aCT, causing the system to drain every few hours**
- **After many attempts to fix it, it was decided late on Friday to switch to arc05 and have only the arcds1 do the staging.**
- **New CE host hardcoded in aCT so we did not need to wait for the ATLAS infosys to propagate the change**
  - **jobs started to run, and ran stable over the weekend, filling the 11-node allocation**
- **On account of the low memory usage, ATLAS proposed on Sunday to try out 18-core jobs in order to fill the nodes**
  - **Miguel switched to allow 72-cores per node un Sunday evening**
  - **ATLAS overrode the 16-core setting for the task directly on the aCT**
  - **18-core jobs ran stable overnight**

SCALING UP

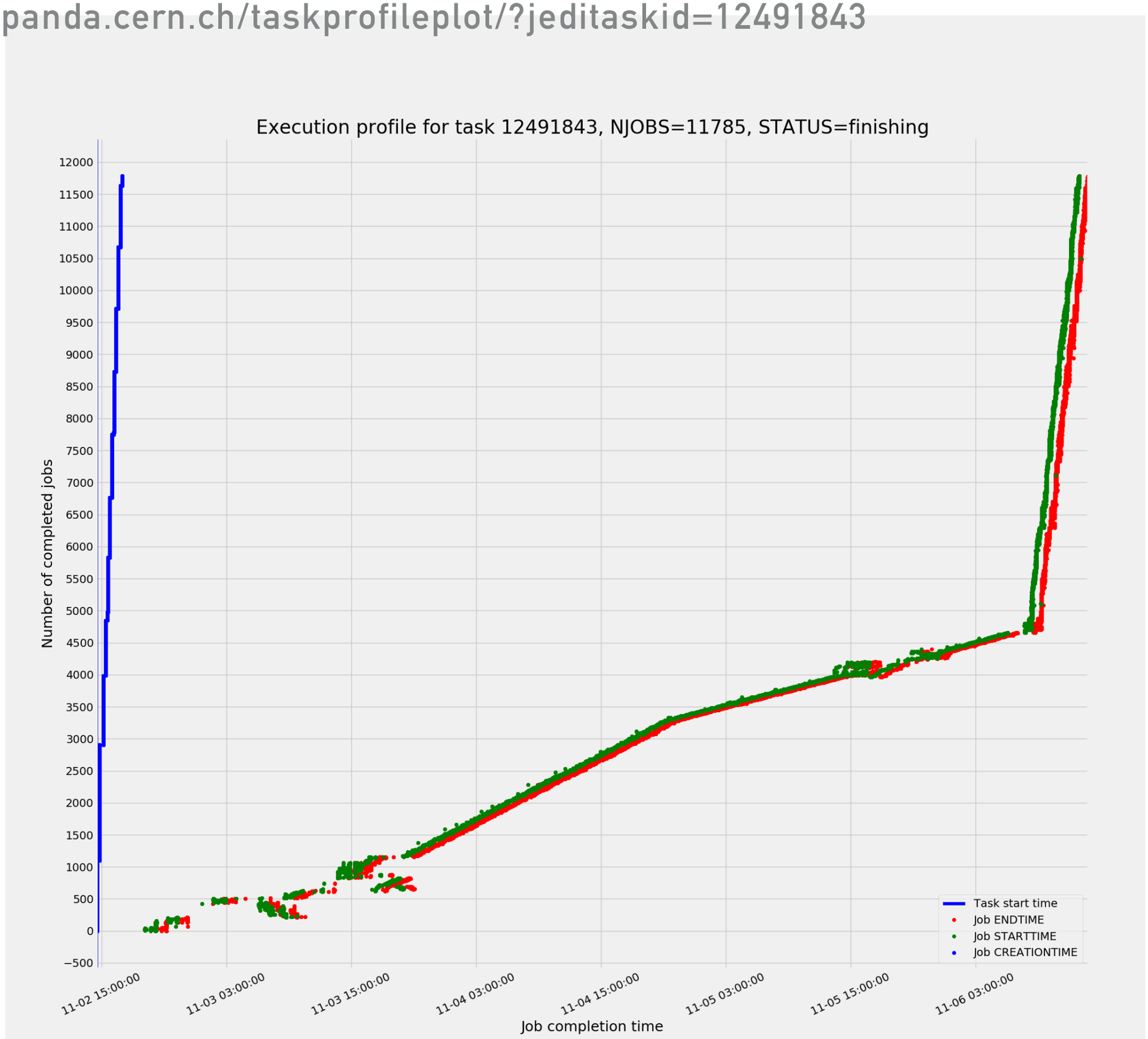
Started 06 Nov 8 AM

- **Decided initially to ramp up in 5 stages in order to avoid nasties**
- **Eventually went for all-at-once**

- **Reached 1420 jobs (25560 cores) in ~1h**
  - **aside from a glitch with ARC that added about ~20 min delay**
  - **fairly linear otherwise, 27 jobs/min**
  - **seemingly dominated by slurm**
- **ARC unstable, a-rex getting stuck repeatedly, needs to be restarted by hand**
  - **Realised we did not have the latest bugfix version**
  - **Debated whether upgrade on the fly vs babysit**
  - **Went for the latter, many restarts needed**
  - **Incresed the maxqueued on the aCT to have a large enough buffer and avoid draining between restarts**
- **Stable running for 3h from 11 AM**
- **Stopped submission at 2 PM**
- **Killed all running from the aCT at 14:45**
- **System clean at 3 PM**

TEST SUMMARY

https://bigpanda.cern.ch/taskprofileplot/?jeditaskid=12491843



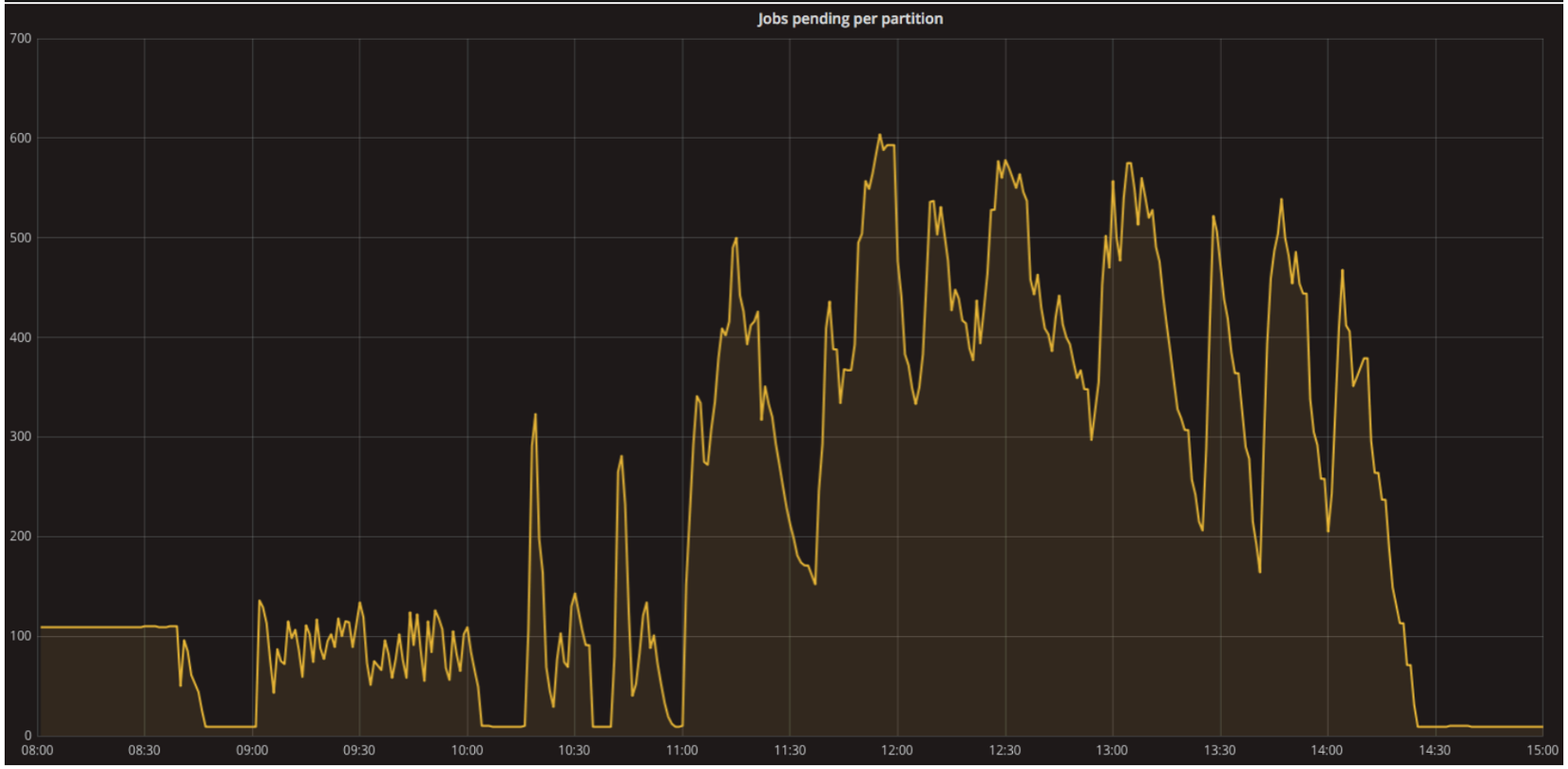
TEST SUMMARY

boosted nr. queued

a-rex still dying

a-rex dying

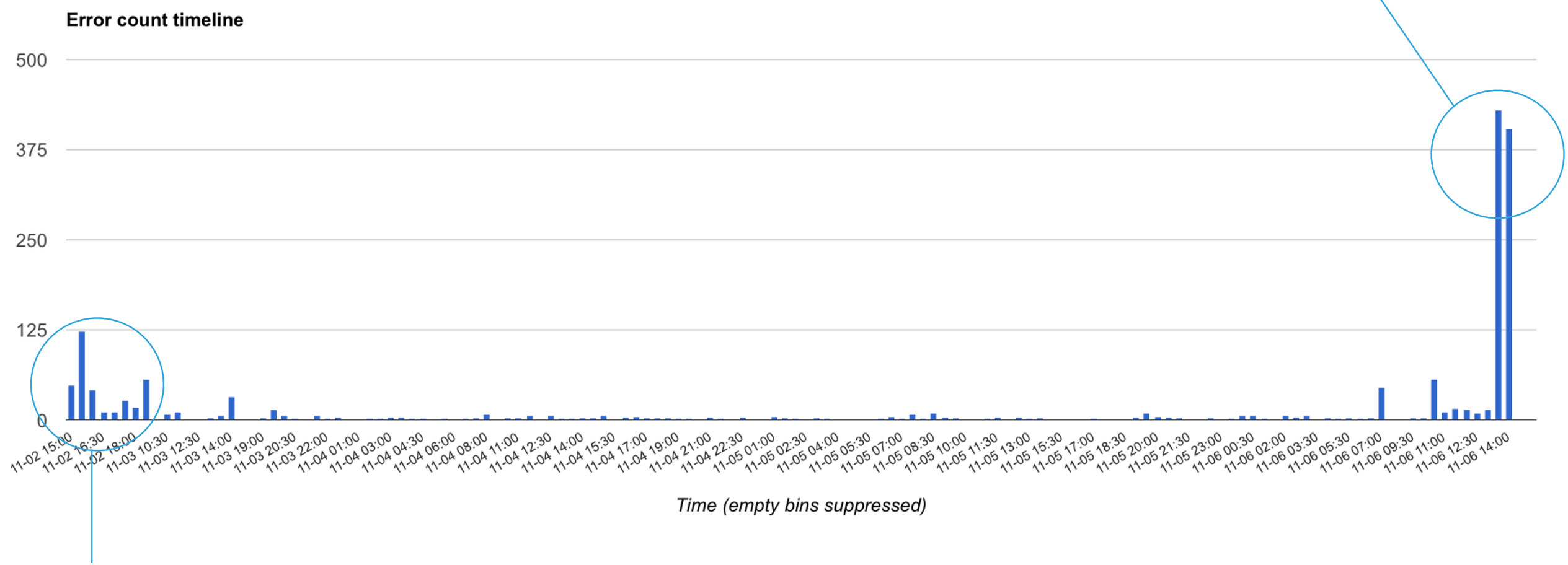
a-rex died



TEST SUMMARY

https://bigpanda.cern.ch/errors/?jeditaskid=12491843

jobs killed at task end

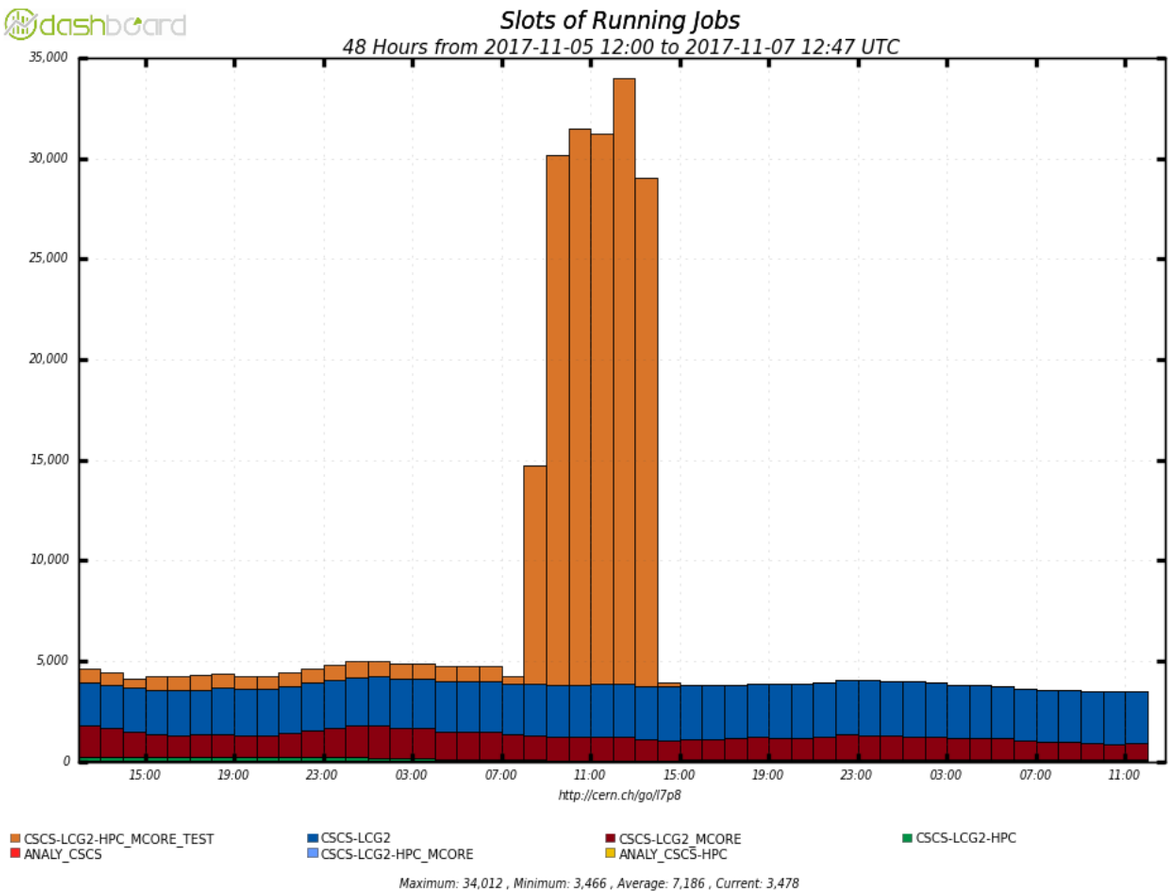


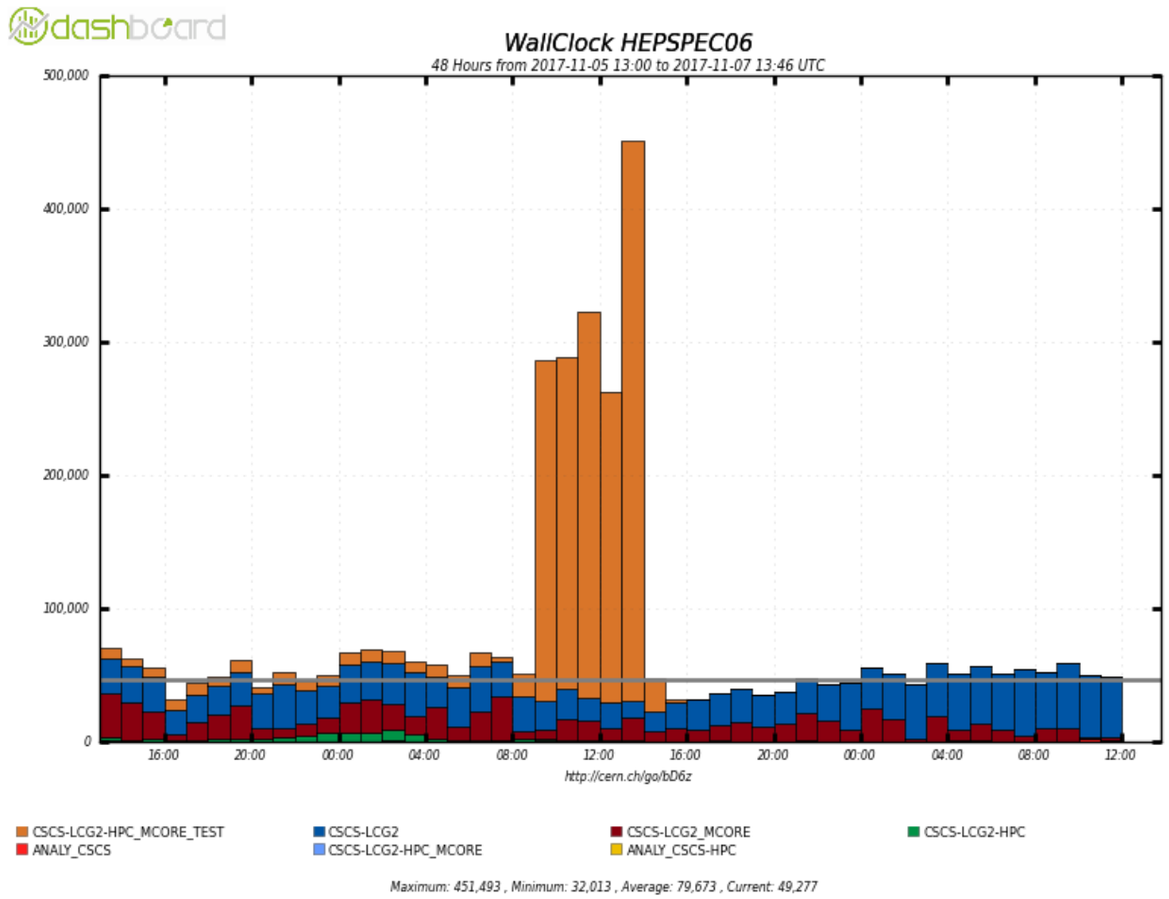
slurm memory management issues

TEST SUMMARY

http://dashb-atlas-job.cern.ch/dashboard/request.py/dailysummary#button=resourceutil&sites%5B%5D=CSCS-LCG2&sitesCat%5B%5D=CH-CHIPP-CSCS&resourcetype=All&sitesSort=2&sitesCatSort=2&start=null&end=null&timerange=last48&granularity=Hourly&generic=0&sortby=16&series=30&activities%5B%5D=all

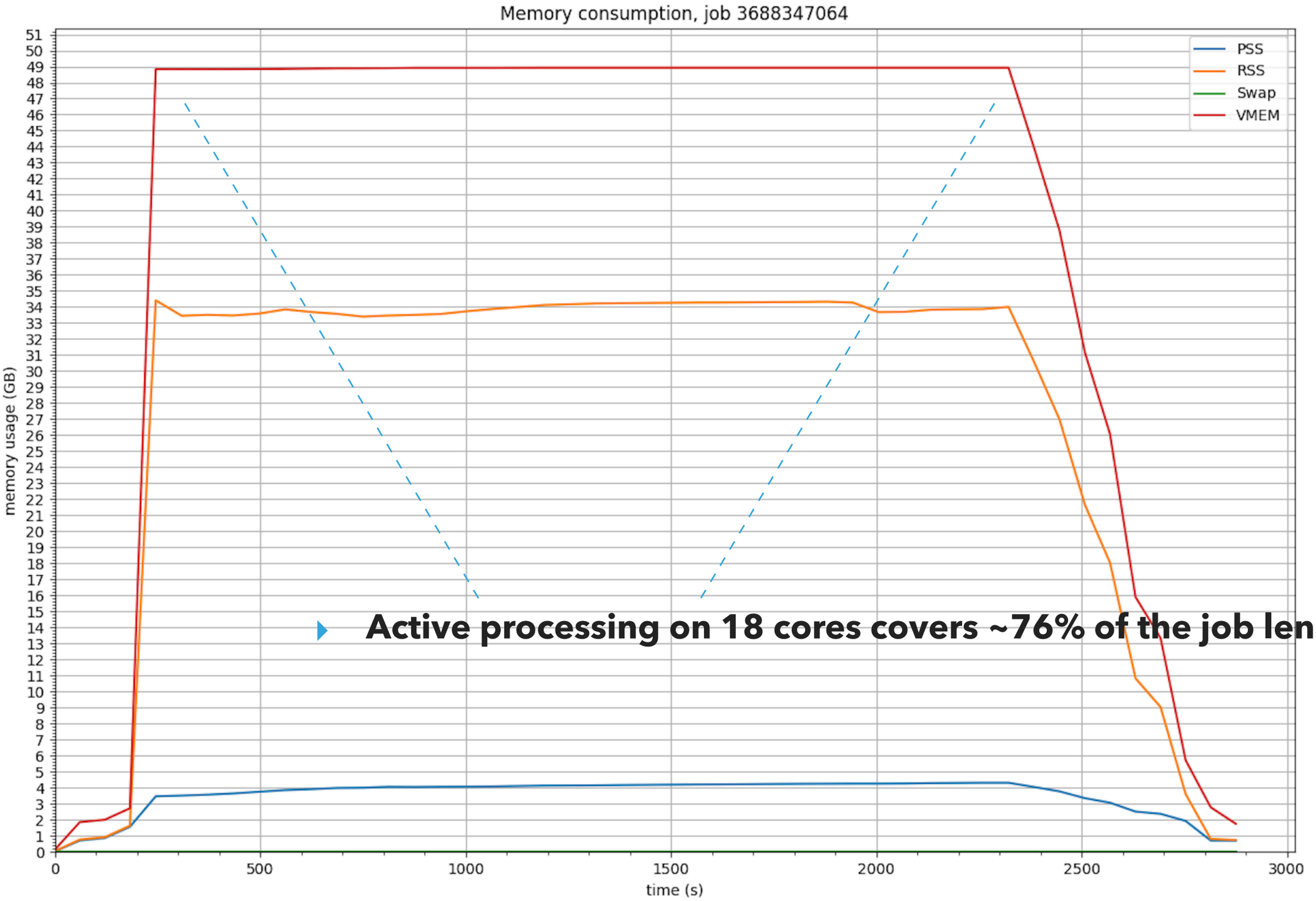
- **1M events processed (25% of total): 10162 jobs (out of 11785)**
- **Total input size: 1TB (no ARC caching), output size: 0.7TB (to the Nucleus in Spain)**
- **Max running jobs reached 1432 (25774/27648 cores - 93.22%)**
- **Failure rate <1% (but all retried), CPU/WC eff 0.76 (due to artificial job length)**





JOB PROFILE

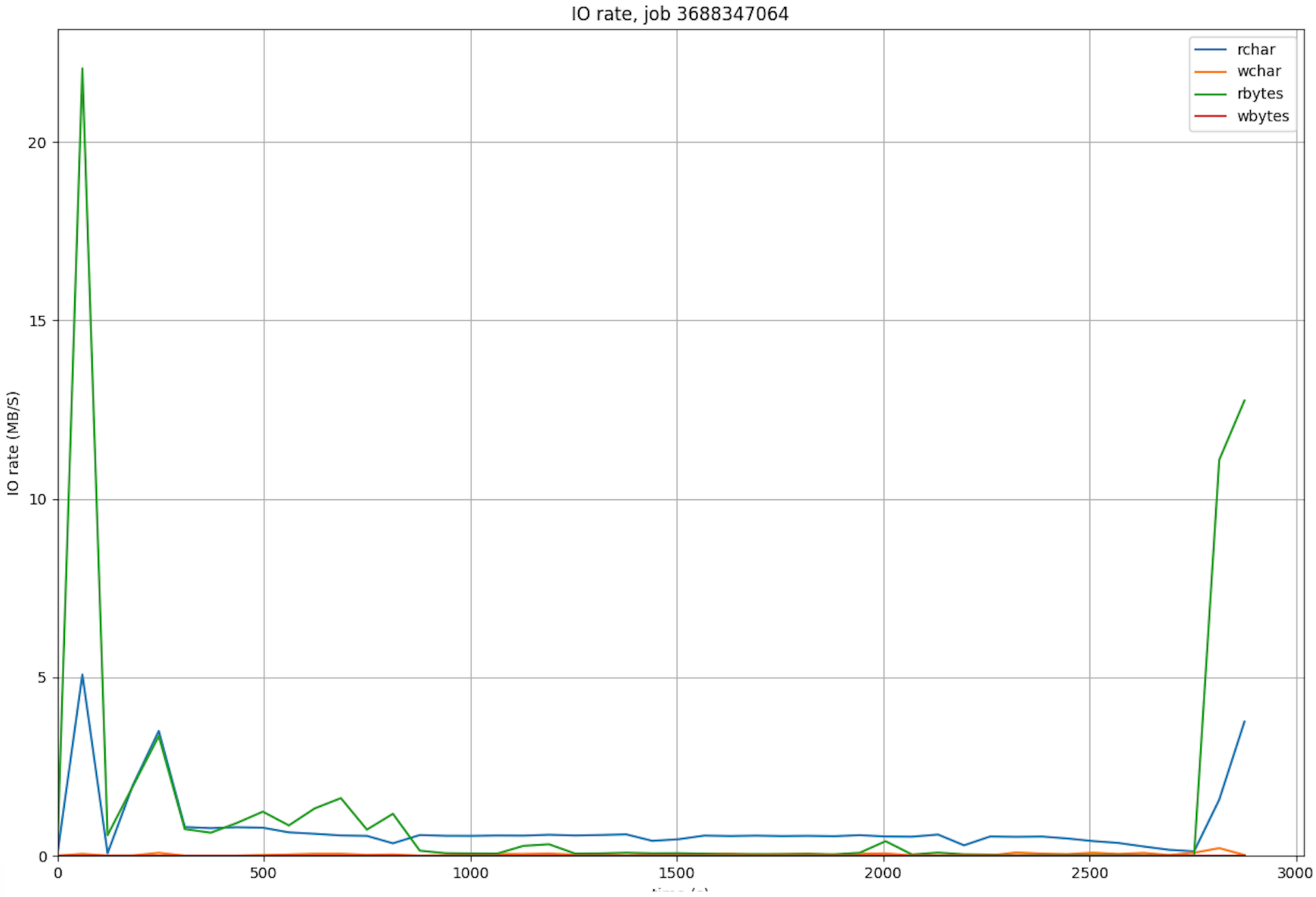
https://bigpanda.cern.ch/memoryplot/?pandaid=3688347064



- **Active processing on 18 cores covers ~76% of the job length**

JOB PROFILE

https://bigpanda.cern.ch/memoryplot/?pandaid=3688347064



CONCLUSIONS AND OUTLOOK

1/2

- **The exercise has been very successful**
- **This is the first large scale (>10k) exercise pursued by ATLAS, that uses a "standard" Tier-2 setup**

- **The weak link at runtime turned out to be the middleware**
  - **we realised too late that we did not have the optimal version**

- **Scale up time not limited by the middleware and/or the shared file system**
  - **I/O is very modest for these jobs**

- **Job failure rate negligible**

- **CPU/WC constrained by the artificially short job duration**
  - **job start up and wrap up are executed on a single core**
  - **the job profile plots show that the effective CPU/WC eff is close to 100%**
  - **jobs of this kind would last several hours in normal production conditions**

CONCLUSIONS AND OUTLOOK

2/2

- **This test can be considered a milestone for our efforts of porting the WLCG data processing to Piz Daint**

- **It has been crucial to demonstrate that simulation runs and scales up as expected before even thinking to port other workloads, such as data processing**

Immediate plans:

- **I have agreed with ATLAS to run the same simulation task on Phoenix**
  - **This should give us some benchmarking information beyond the HS06**

- **The same task is also running on Titan at Oak Ridge (Cray XK-7, #4 in the top500)**
  - **We will carry out a direct comparison with that system, since we know the HS06 rating of both**