Adapting the ATLAS job submission systems to changes in HPC computing

M. Svatoš, J. Chudoba, P. Vokáč

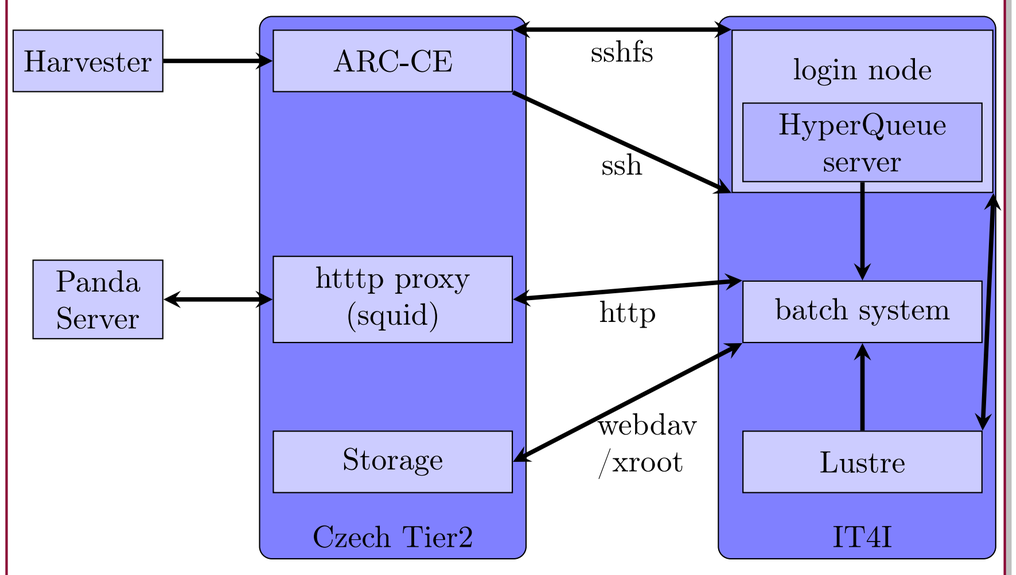
7th Users Conference of IT4Innovations

30.-31.10.2023

Introduction

For several years, the distributed computing of the ATLAS experiment at the LHC (ADC) has been granted access to computing resources of the Czech national HPC centre, IT4Innovations. Currently, it means running jobs on Karolina and Barbora clusters. The submission system is being improved and adapted to the evolving HPC environment.

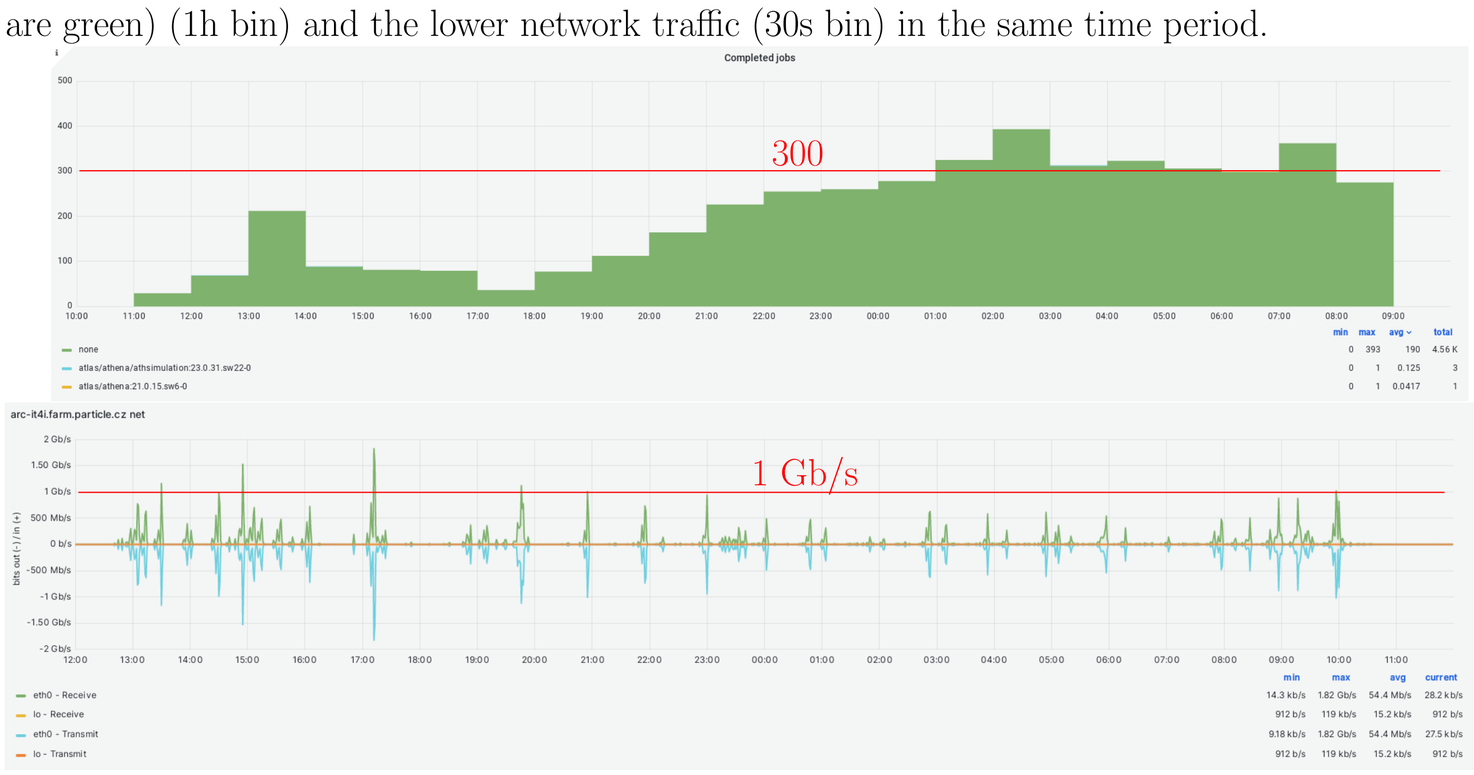
Submission system



- the ARC-CE shares storage with the Lustre via sshfs connection through a login node and communicates with the batch system via ssh connection (through the same login node)
- when the ARC-CE receives a job, it translates the job description into a script that can be run in the batch system, puts necessary files into a folder within sshfs shared area and submits the job via ssh connection to the HyperQueue server running on a login node
- the HyperQueue server buffers the jobs and when there are enough of them, it submits jobs into the batch system
- when the batch job starts, HyperQueue jobs start in it (in sufficient numbers to fill the worker node - if available)
- in each HyperQueue job, pilot wrapper starts, launching the pilot
- pilot contacts panda server through http proxy (Czech Tier2 squid) to receive a payload (as there are only few open ports)
- when it receives the payload, it gets input file from the Czech Tier2 storage via xroot or webdav
- then it starts the calculation
- when the payload finishes, it sends outputs to the Czech Tier2 storage via xroot or webdav
- when this is finished, pilot will request another payload (if it can expect that the batch queue setting would allow it to finish)

cvmfsexec

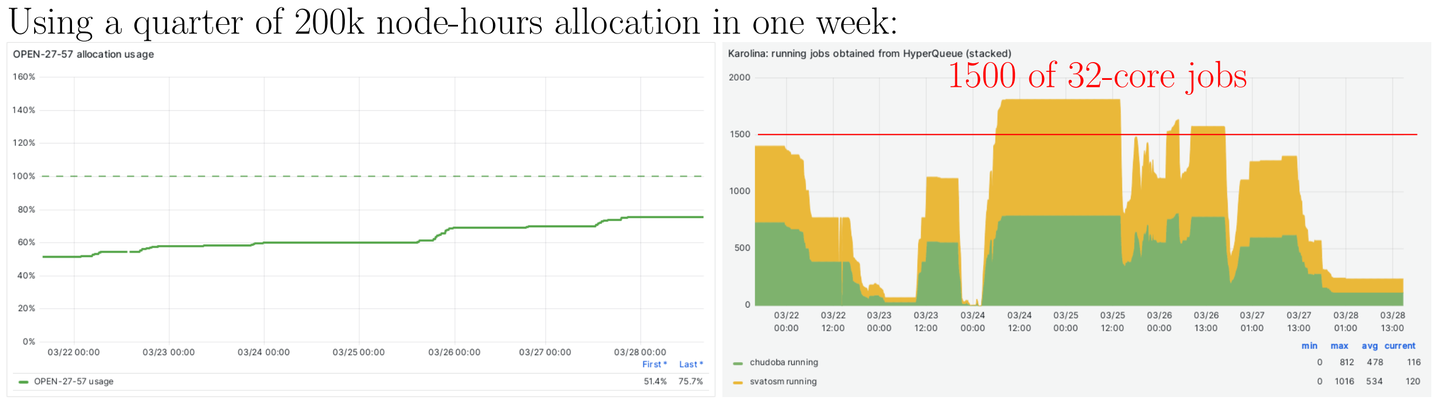
On clusters, where CVMFS software mount is not available, the software can be obtained using cvmfsexec. The downside is that the necessary files are downloaded for every job.
The upper plot shows successfully completed jobs on Karolina (jobs using software via cvmfsexec are green) (1h bin) and the lower network traffic (30s bin) in the same time period.



Allocation and opportunistic usage

Allocation

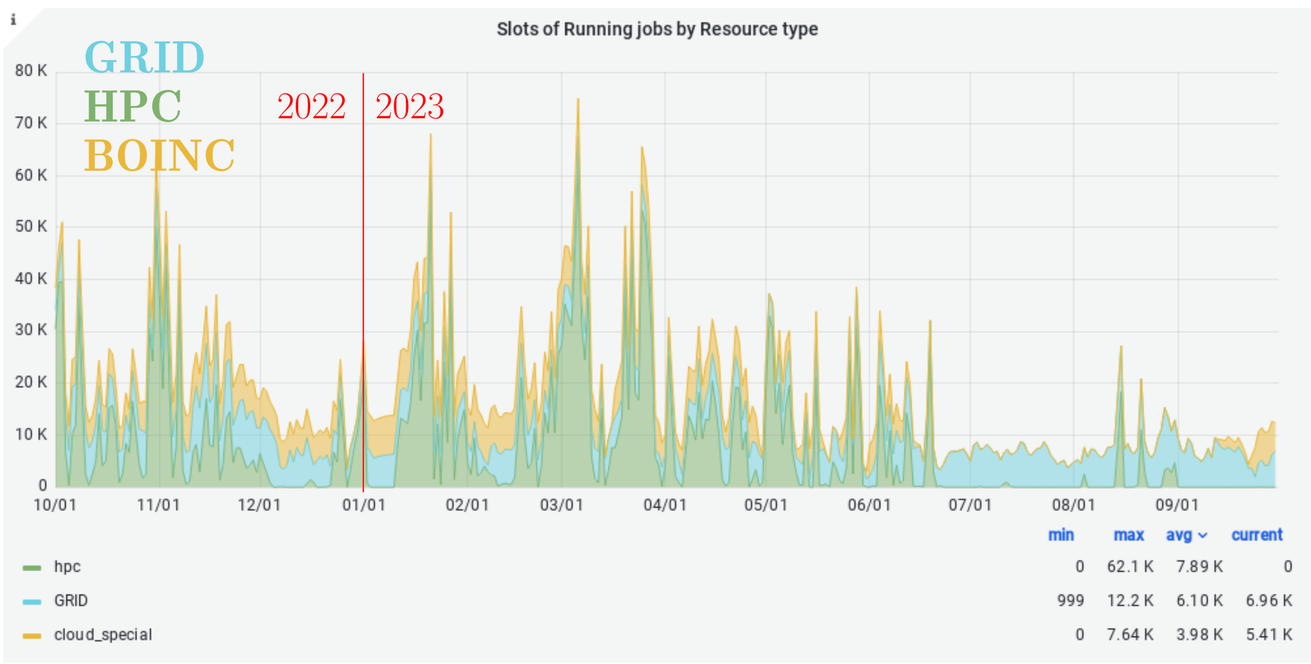
Using a quarter of 200k node-hours allocation in one week:

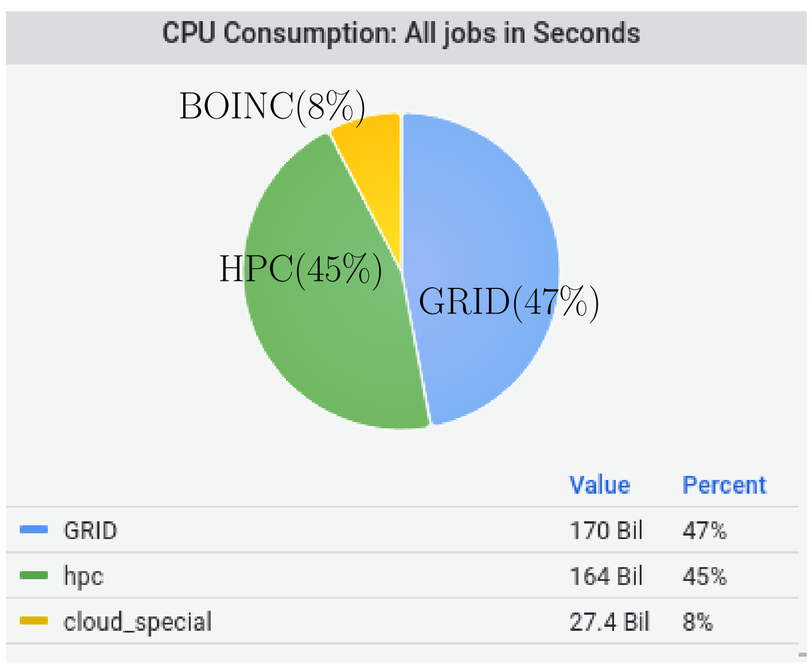


Opportunistic usage

To decrease number of jobs killed on pre-emptive batch queue, the submission system can adapt speed of job submission to situation in the batch system.

Performance





- the HPCs of IT4Innovations (Karolina and Barbora) provide significant resources to Czech Tier2
- the period of more stable running during winter and spring of 2023 corresponds to usage of allocation

Acknowledgement

This work and computing resources were supported by the Ministry of Education, Youth and Sports (MEYS) of the Czech Republic through the e-INFRA CZ (ID:90254) and CERN-CZ (LM2018104 and LM2023040). The computing resources at FZU were co-financed by projects CERN-C (CZ.02.1.01/0.0/0.0/16013/0001404) and CERN-CD (CZ.02.1.01/0.0/0.0/18 046/0016013) from EU funds and MEYS.