ATLAS utilisation of the Czech national HPC center

M. Svatoš, J. Chudoba, P. Vokáč

On behalf of the ATLAS Collaboration

CHEP 2018, 9-13.7.2018

Introduction

Czech National Supercomputer Center IT4Innovations in Ostrava operates Salomon HPC system which is the most powerful computer in the Czech Republic listed in Top500 and currently ranked 87th in the world. Salomon was built in 2015 providing 2 PFLOPs in peak. It consists of 1008 computational nodes with 24 cores of Intel Xeon E5 CPUs and 128 GB of RAM per node interconnected with Infiniband (56 Gbps). 432 nodes also contain 61 core Intel Xeon Phi accelerator. So, in total there are 24192 cores on the computational nodes with additional 52704 accelerator cores. Salomon runs CentOS 6 based nodes with PBSpro as a batch system. Nodes have very limited external connectivity through http/https proxy.

The ATLAS Experiment at CERN uses the Salomon cluster in opportunistic fashion via the Czech Tier2 site (praguelcg2) [1]. Only non-accelerated nodes are available for opportunistic usage and unlike other opportunistic HPC resources, there is no job pre-emption. HPC batch scheduler selects jobs of projects without dedicated computing time allocation to increase supercomputer utilization efficiency by filling empty batch slots with opportunistic jobs. Therefore, lack of available resources will manifest as closed jobs (jobs that never started running), not as jobs killed by the batch system.

First successful ATLAS job submitted to the Salomon via ARC-CE finished in December 2017 and the ARC-CE is submitting jobs since.

ARC-CE

Traditional pilot based grid jobs do not always deal well with restrictive HPC environment. ATLAS developed ARC Control Tower (aCT) to be more flexible with utilizing supercomputers with specific configurations. ATLAS jobs coming from aCT are transformed by ARC-CE [2] computing element into a script runnable in HPC's PBSPro batch system. Computing element also downloads input files which are put into the session directory together with run script and later uploads output files. Several modifications of ARC-CE scripts were necessary to be able to submit job to a login node via ssh. Also, proxy setting was added to allow jobs access to condition data on squids.

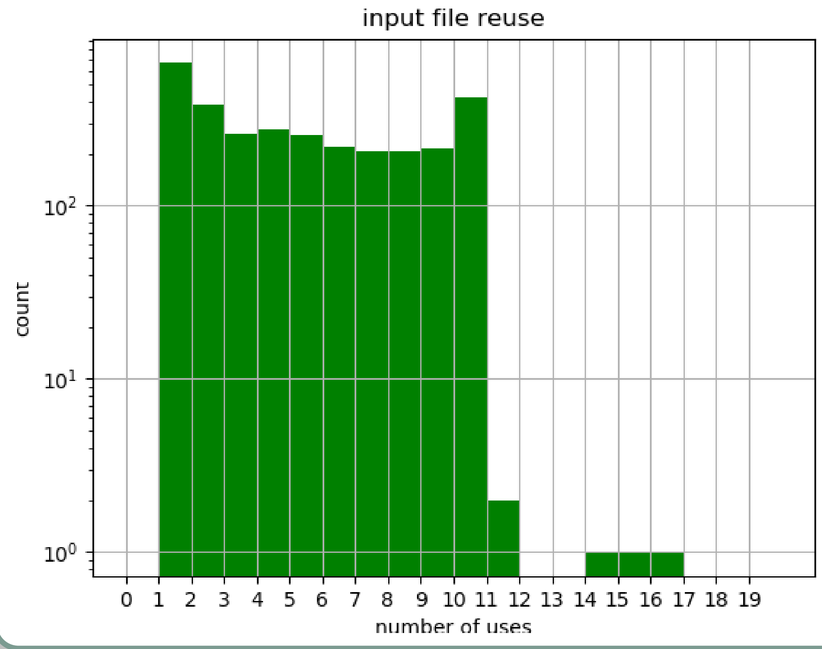
Shared folder

Session, runtime, and cache directories are mounted via sshfs on dedicated scratch space on Salomon's Lustre filesystem.

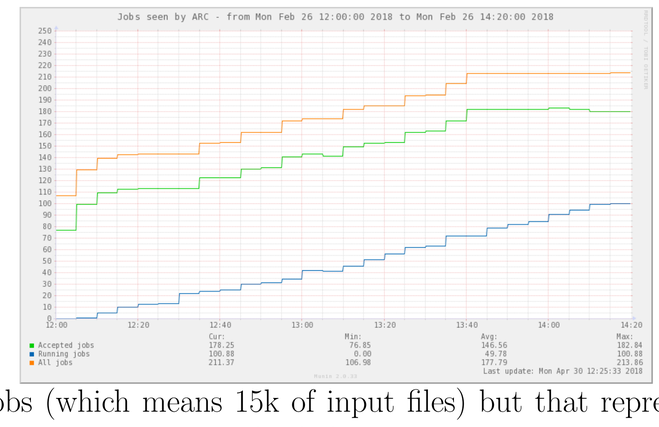
Software

CVMFS [3] is mounted to the ARC-CE and necessary software is copied into scratch area of the HPC via `rsync` once a day. Current size of the software is slightly over half TB in ∼10 million files.

Cache

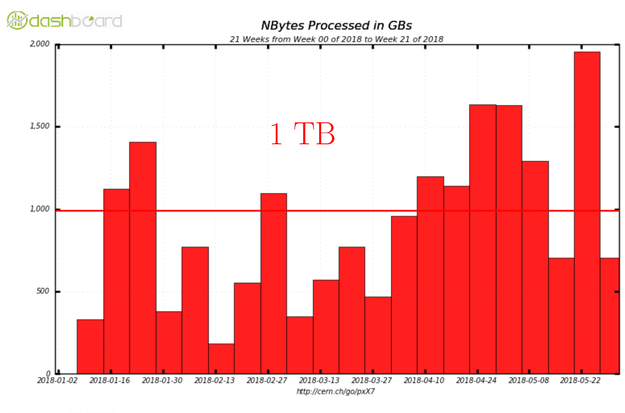


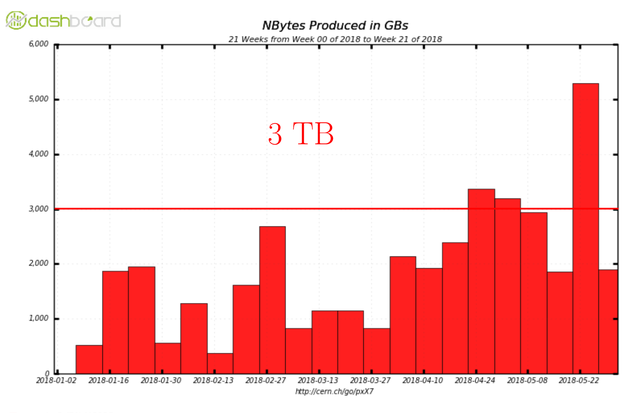
The analysis of input file reuse show that many input files are used more than once. The analysis was performed on successfully finished jobs from end of December 2017 until the beginning of March 2018. During this time period, there were 15k of successful jobs (which means 15k of input files) but that represents only 3k of unique files (Figure on the left). This was sufficient motivation to deploy an ARC-CE cache located on the shared file system.

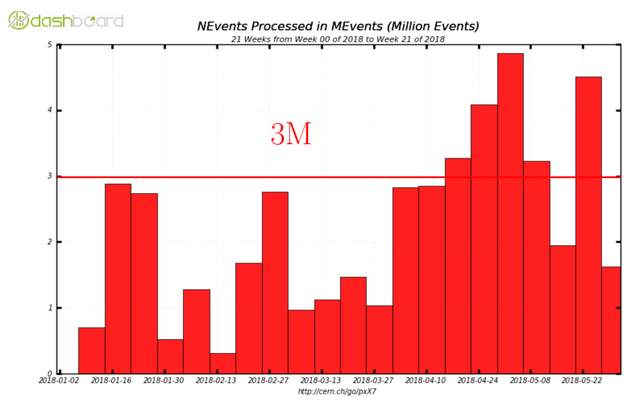


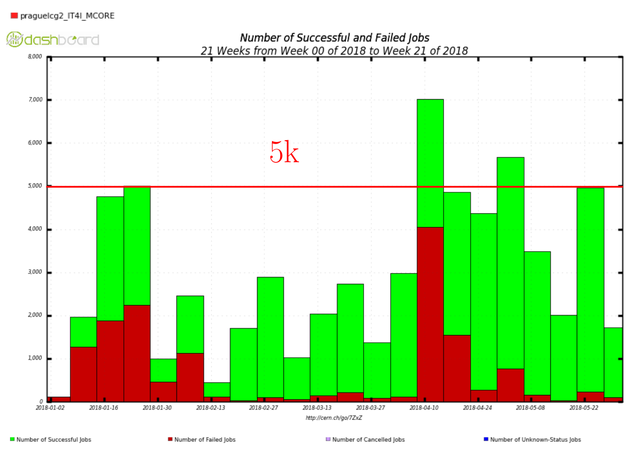
The effect on speed of job start-up is significant. Before the cache was deployed, it took $\mathcal{O}(1)$ h (middle Figure, each step is 5 min) to create jobs for all 100 slots. With the cache, it takes $\mathcal{O}(1)$ min (right Figure). Cache size is $\mathcal{O}(1)$ TB because only files which were not used in the last 60 days are deleted.

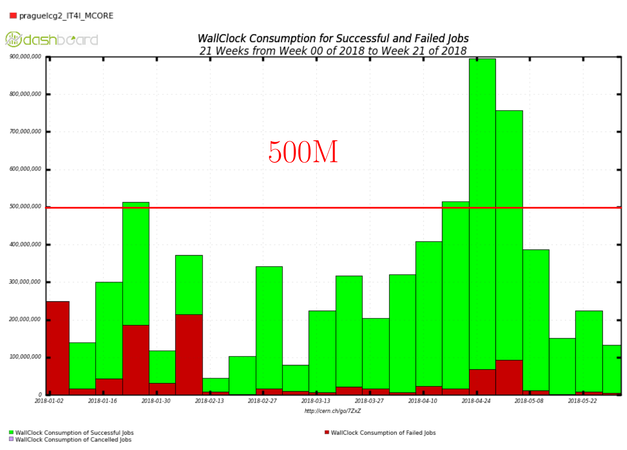
Job statistics (January-May 2018)

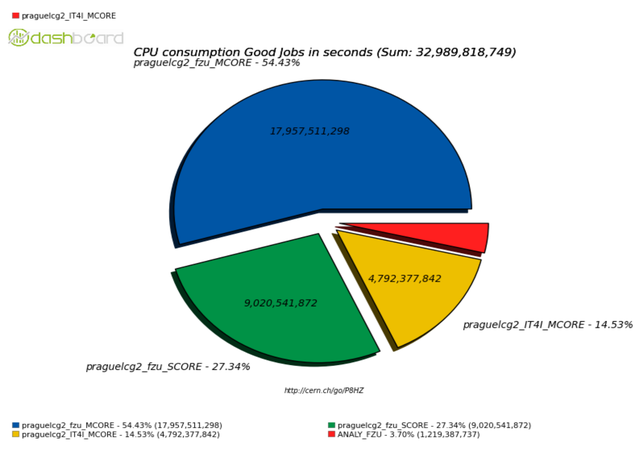












Weekly numbers:

- amount of input (top left Figure) is usually below or around TB
- amount of output (top middle Figure) is usually few TB
- few millions events are processed (top right Figure)
- hundreds to thounsands jobs finishes every week (bottom left Figure) - the major sources of failed jobs on are caused by PBS filehandle leak (until first half of February) and installation issue (April)
- wallclock (bottom middle Figure) shows the installation issue did not waste significant amount of resources
- comparison of CPU consumption of good jobs (bottom right Figure) of all active praguelcg2 queues shows that the HPC provides significant amount of resources (about 15%).

References

References

[1] J. Chudoba *et al.*, A multipurpose computing center with distributed resources, J. Phys. Conf. Ser. **898** (2017) no.8, 082034. doi:10.1088/1742-6596/898/8/082034

[2] M.Ellert et al., Advanced Resource Connector middleware for lightweight computational Grids, Future Generation Computer Systems 23 (2007) 219-240.

[3] `http://cernvm.cern.ch/`

Acknowledgement

Computing resources as co-financed by projects Research infrastructure CERN (CERN-CZ) and OP RDE CERN Computing (CZ.02.1.01/0.0/0.0/16013/0001404) from EU funds and MŠMT.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS