EPJ Web of Conferences **245**, 03034 (2020) https://doi.org/10.1051/epjconf/202024503034
CHEP 2019

Distributed resources of Czech WLCG Tier-2 center

Martin Adam[1,*], *Dagmar* Adamová[1,**], *Jiří* Chudoba[2,***], *Alexandr* Mikula[2,****], *Michal* Svatoš[2,†], *Jana* Uhlířová[2,‡], and *Petr* Vokáč[3,§]

[1]Nuclear Physics Institute of the Czech Academy of Sciences, Řež 130, 25068, Řež, Czech Republic

[2]Institute of Physics of the CAS, Na Slovance 1999/2, Prague, 18221, Czech Republic

[3]Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Břehová 7, Prague, 115 19, Czech Republic

Abstract. The Computing Center of the Institute of Physics (CC IoP) of the Czech Academy of Sciences provides compute and storage capacity to several physics experiments. Most resources are used by two LHC experiments, ALICE and ATLAS. In the WLCG, which coordinates computing activities for the LHC experiments, the computing center is Tier-2. The rest of computing resources is used by astroparticle experiments like the Pierre Auger Observatory (PAO) and the Cherenkov Telescope Array (CTA) or particle experiments like NOvA and DUNE. Storage capacity is distributed to several locations. DPM servers used by the ATLAS and the PAO are all in the same server room. ALICE uses several xrootd servers located at the Nuclear Physics Institute in Rez, about 10 km away. The storage capacity for the ATLAS and the PAO is extended by resources of the CESNET (the Czech National Grid Initiative representative) located in Ostrava, more than 100 km away from the CC IoP. Storage is managed by dCache instance, which is published in the CC IoP BDII. ATLAS users can use these resources using the standard ATLAS tools in the same way as the local storage without noticing this geographical distribution. The computing center provides about 8k CPU cores which are used by the experiments based on fair-share. The CPUs are distributed amongst server rooms in the Institute of Physics, in the Faculty of Mathematics and Physics of the Charles University, and in CESNET. For the ATLAS experiment, the resources are extended by opportunistic usage of the Salomon HPC provided by the Czech national HPC center IT4Innovations in Ostrava. The HPC provides 24-core nodes. The maximum number of allowed single-node jobs in the batch system is 200. The contribution of the HPC to the CPU consumption by the ATLAS experiment is about 15% on average.

*e-mail: madam@fzu.cz
**e-mail: adamova@ujf.cas.cz
***e-mail: Jiri.Chudoba@cern.ch
****e-mail: mikula@fzu.cz
†e-mail: Michal.Svatos@cern.ch
‡e-mail: uhlirova@fzu.cz
§e-mail: petr.vokac@cern.ch

1 Introduction

The Czech WLCG Tier-2 center (praguelcg2) combines CPU and storage resources located in several different cities. The main compute and storage resources are located in Prague. There is additional storage for the ALICE experiment located in Řež. In Ostrava, there is storage capacity provided the CESNET (the Czech National Grid Initiative representative) and HPC resources provided by the Czech national HPC center IT4Innovations. Figure 1 shows their location on the map of the Czech Republic.

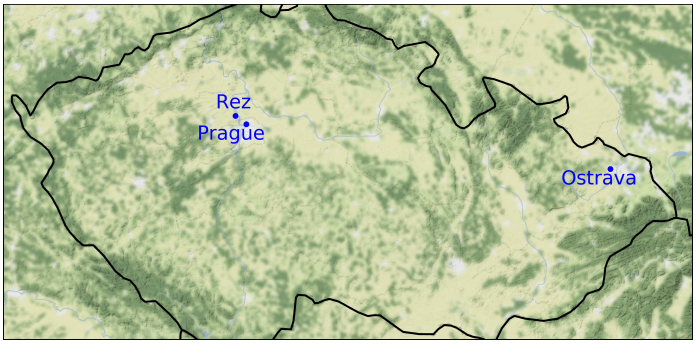

Figure 1. Location of computing centers federated under praguelcg2

2 The site

For core services, the praguelcg2 uses high-availability enterprise storage mounted to several virtualization machines with the ability of live migration from one virtualization machine to another. Worker nodes use standard server hardware. There is detailed monitoring in place which can automatically remove problematic worker node from batch system. The storage is based on RAID6. Similarly to worker nodes, there is detailed monitoring in place. In case any problems are observed in the monitoring, we try to get it fixed as soon as possible. Almost all machines are controlled by puppet [1]. This system works very well but for HL-HLC era the infrastructure will need to evolve. What will the result of the evolution is quite unclear now as it depends on many factors (funding available during HL-LHC era, maturity of various storage system at the time, etc.).
The authentication and authorization to resources is aligned with computing models of supported VOs. It is x509 now and tokens in the future. Jobs are running in containers and therefore their payloads are isolated. The cgroups help with control of resource usage.
The site provides resources to experiments based on MoU for T2 center.

2.1 Storage

The Computing Center of the Institute of Physics (CC IoP) of the Czech Academy of Sciences in Prague provides ~ 3 PB for ATLAS, ~ 200 TB for Auger, ~ 300 TB for Dune in its DPM storage system [2]. It also provides ~ 1 PB for ALICE in XRootD system [3] and small fast StashCache (XCache) [4] to improve Fermilab jobs efficiency by caching big static files from the CVMFS. The Nuclear Physics Institute in Řež provides additional ~ 1 PB for ALICE in XRootD system. There is also a tape archiving for local users available in dCache [5] in Ostrava.

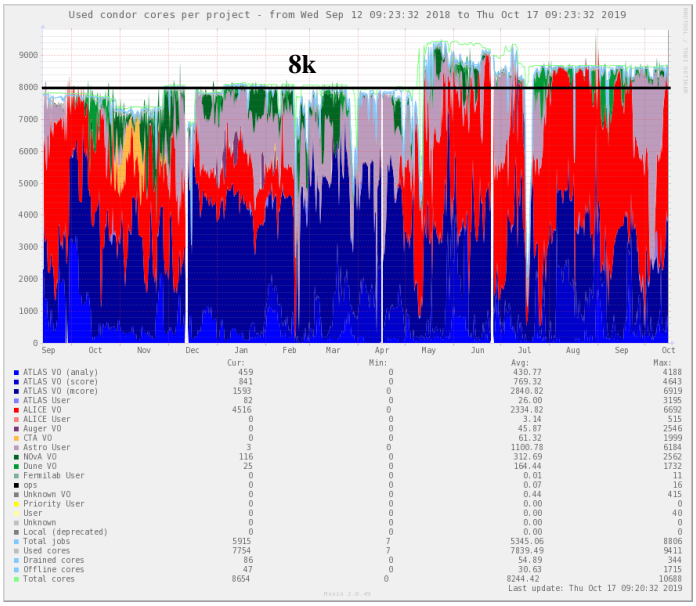

Figure 2. Batch slots usage in the HTCondor between September 2018 and October 2019. Colours represent different VOs but also different activities of a VO.

The DPM headnode is a virtual machine on high-availability system, i.e. in case it breaks down it can be quickly re-created. The whole DPM storage is split into ten dpmpools. As the files are spread over them, the load is more less evenly spread.

2.2 CPU resources

The HTCondor batch system [6] allows usage of off-site machines. Figure 2 shows usage of ~ 8k CPU cores, consisting of Intel and AMD (EPYC) processors, distributed amongst the Institute of Physics (the Computing Center), the Faculty of Mathematics and Physics of the Charles University, Faculty of Nuclear Sciences and Physical Engineering of the Czech Technical University and CESNET. Usage by the experiments is based on fair-share.

2.3 Network

The network configuration of the Computing Center is shown on Figure 3. The Computing Center is connected to several external locations/providers, namely, by 100 Gbps to LHCONE, 40 Gbps to Pasnet/Internet, and 10 Gbps to the Charles University and to Řež. The internal connectivity differs for storage and worker nodes. For storage nodes, it is typically 40 Gbps (or at least 10 Gbps for older servers). For worker nodes, it is 10 Gbps for all new machines with many cores.
All machines providing core and grid services are configured in dual-stack mode with IPv6 protocol preference. Majority of Prague site worker nodes use private IPv4 addresses and direct external access is thereby limited by 20 Gbps NAT. Dual-stack is configured also on all our worker nodes and for IPv6 external data transfers are limited only by our upstream connectivity, currently 140 Gbps.
The monitoring of suspicious behaviour is done by CESNET. As NREN, it monitors its whole infrastructure.

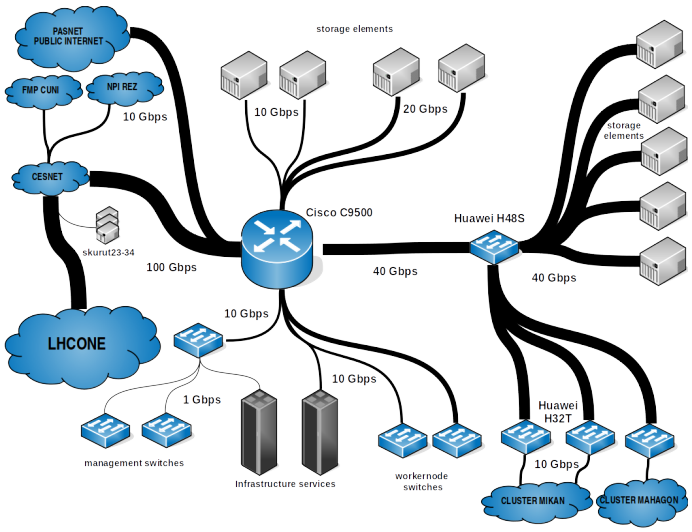

Figure 3. Schematic of praguelcg2 networking

3 Usage

The praguelcg2 provides resources to several Virtual Organizations (VOs):

3.1 ATLAS

The ATLAS experiment [7] is using computing resources of the Computing Center as well as off-site machines (ATLAS VO bands on Figure 2) under its HTCondor batch system. In addition to that, it uses opportunistically the Salomon HPC cluster in Ostrava (usage on Figure 4) via ARC-CE machines installed at the Computing Center (setup details in [8]). Minor contribution to the computing resources also comes from BOINC which runs ATLAS@Home [9] on an unused desktop.

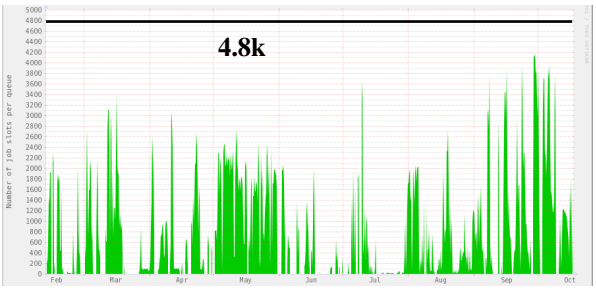

Figure 4. Number of batch slots provided opportunistically by the Salomon HPC cluster

3.2 ALICE

The ALICE experiment has set mandatory delivery of CPU and storage capacities specific for each participating institution. The Czech ALICE group has been in sync with storage and disk space requests but during last 2 or 3 years the delivery of CPU resources was not

completely matching the ALICE requirements. The cause of this situation is in insufficient hardware capacities dedicated to ALICE.
The balancing factor is the ALICE system perfect ability to use free opportunistic resources immediately. This way in 2019 most of the year was ALICE delivering according to the central requirement, over-fulfilling it occasionally, cf. EGI Accounting Portal [10]. ALICE's requirement was 12,744,000 HS06 hours.
The average efficiency of ALICE jobs over 4 months was above 80% (left plot on Figure 5). Its XRootD storage is spread over two locations. ALICE jobs can require enough data to saturate their 10 Gbps connection (right plot on Figure 5).

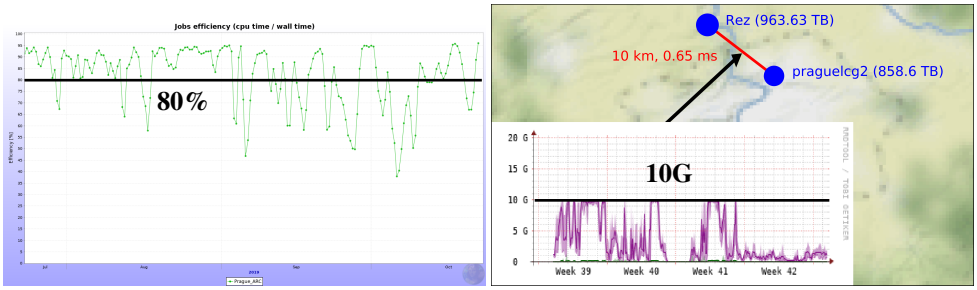

Figure 5. ALICE CPU efficiency and networking usage

3.3 Astrophysics

Supported astrophysics project Pierre Auger Observatory (PAO) and Cherenkov Telescope Array (CTA) use grid resources for simulations of cosmic ray showers induced by various primary particles. These bulk simulations are done in campaigns which can last from several weeks to several months. Here we see the added value of shared resources. These projects can get more computing cores than their contributed share during these campaigns and thus shorten the total time needed to finish simulations. And their resources are not idle when these projects do not use them.
The sharing of resources does not work for data storage. Volumes of tens or hundreds of terabytes of data files cannot be migrated fast enough to enable flexible sharing of storage. CTA uses only short term storage at praguelcg2 site for output of simulations done on local resources and later moves these files for a long term storage on a few selected grid sites. VO auger uses DPM at praguelcg2 for long term storage. A second copy of output data files is

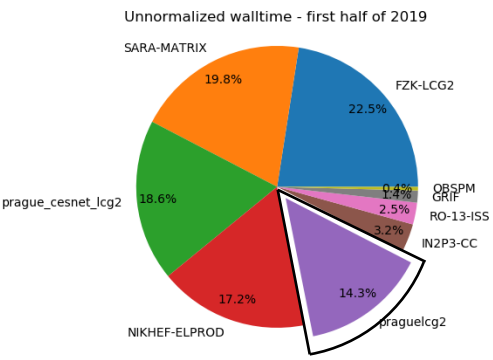

Figure 6. VO auger uses several sites connected to the EGI grid infrastructure for bulk simulations of cosmic ray showers. Usage of praguelcg2 resources by local user jobs is not included in this plot.

stored in CC IN2P3 in iRODS system to protect against possible data loss due to a hardware issues. The total volume of the VO auger data stored at praguelcg2 exceeded 200 TB in 2019.

3.4 Fermilab neutrino experiments

Czech physicists are involved in two Fermilab neutrino experiments - NuMI Off-axis ν_e Appearance (NOvA) and Deep Underground Neutrino Experiment (DUNE). Our distributed computing center provides CPU and storage resources fully integrated in the OSG infrastructure used by Fermilab grid jobs. We provide resources available at our computing center since DZero experiment and to make our integration simple we use OSG middle-ware including HTCondor-CE as a gateway for grid jobs.
CPU resources for Fermilab experiments comes from CC IoP neutrino fair-share and worker nodes located at Charles University. Neither NOvA nor DUNE have continuous flow of jobs, because they run bigger production campaign only several times per year and only small fraction of user analysis jobs use grid resources. When neutrino CPU resources are not fully filled they can be used opportunistically by other experiments. CPU resources, named FZU in OSG, provided second most significant contribution in processing neutrino offsite jobs (Figure 7).
Close location with reasonable low latency on network between two CPU sites and CESNET dedicated network connection allow us to have all storage resources located just centrally at CC IoP. To provide better job CPU efficiency neutrino jobs benefits from StashCache (XCache) caching service which improve access time and file read throughput for big static physics data distributed via CVMFS distributed filesystem. DUNE data model rely on distributed storage managed by Rucio and fraction of DUNE data are replicated to our DPM storage. Local storage can further improve job efficiency, because relatively slow data transfers from distant storage can't sufficiently utilize CPU resources. Average job CPU efficiency is much better for DUNE jobs compared to the older NOvA experiment.

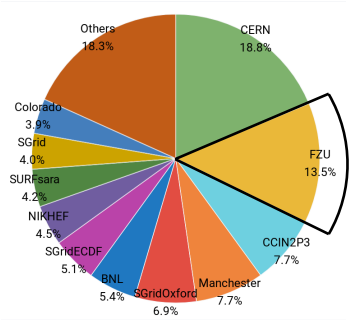

Figure 7. Computing sites CPU time contribution for Fermilab neutrino experiments (NOvA, DUNE) jobs eligible for off-site processing including our Prague (FZU) site

4 Summary and Conclusion

The praguelcg2 successfully provides computing resources to several experiments. Resources located in several cities are presented as one site to them.
In the future, several changes and upgrades are planned (token based authentication, migration to version 6 of the ARC-CE, etc.). They are related to computing models of VOs. Further expansion of the infrastructure depends on future funding.

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.

References

[1] *Puppet - software configuration management tool*, https://puppet.com/

[2] F. Furano, O. Keeble, A. Manzi, G. Bitzes, EPJ Web Conf. **214**, 04018 (2019)

[3] W. Yang, A.B. Hanushevsky, J. Phys. Conf. Ser. **898**, 062046 (2017)

[4] D. Weitzel, M. Zvada, I. Vukotic, R. Gardner, B. Bockelman, M. Rynge, E.F. Hernandez, B. Lin, M. Selmeci, *StashCache: A Distributed Caching Federation for the Open Science Grid*, in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)* (ACM, New York, NY, USA, 2019), PEARC '19, pp. 58:1–58:7, ISBN 978-1-4503-7227-5, http://doi.acm.org/10.1145/3332186.3332212

[5] T. Mkrtchyan, O. Adeyemi, P. Fuhrmann, V. Garonne, D. Litvintsev, P. Millar, A. Rossi, M. Sahakyan, J. Starek, S. Yasar, EPJ Web Conf. **214**, 04042 (2019)

[6] D. Thain, T. Tannenbaum, M. Livny, Concurrency - Practice and Experience **17**, 323 (2005)

[7] ATLAS Collaboration, JINST **3**, S08003 (2008)

[8] M. Svatos, J. Chudoba, P. Vokac, EPJ Web Conf. **214**, 03005 (2019)

[9] D. Cameron, W. Wu, A. Bogdanchikov, R. Bianchi, EPJ Web Conf. **214**, 03011 (2019)

[10] *EGI accounting portal*, https://accounting.egi.eu/wlcg/tier2/site/praguelcg2/normelap_processors/VO/DATE/2019/1/2019/12/lhc/onlyinfrajobs/