Production experience and performance for ATLAS data processing on a Cray XC-50 at CSCS

F G Sciacca and M Weber on behalf of the ATLAS Collaboration

Albert Einstein Center for fundamental Physics, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

E-mail: gianfranco.sciacca@lhep.unibe.ch

Abstract. Prediction for requirements for the LHC computing for Run 3 and for Run 4 (HL-LHC) over the course of the next 10 years, show a considerable gap between required and available resources, assuming budgets will globally remain flat at best. This will require some radical changes to the computing models for the data processing of the LHC experiments. The use of large scale computational resources at HPC centres worldwide is expected to increase substantially the cost-efficiency of the processing. In order to pave the path towards the HL-LHC data processing, the Swiss Institute of Particle Physics (CHIPP) has taken the strategic decision to migrate the processing of all the Tier-2 workloads for ATLAS and other LHC experiments from a dedicated x86_64 cluster that has been in continuous operation and evolution since 2007, to Piz Daint, the current European flagship HPC, which ranks third in the TOP500 at the time of writing. We report on the technical challenges and solutions adopted to migrate to Piz Daint, and on the experience and measured performance for ATLAS in over one year of running in production.

1. Introduction

The High Energy and Nuclear Physics computing community will face several challenges with respect to the computing requirements for the next decade and beyond. For the LHC community, computing requirements for the High-Luminosity LHC runs (2025-2034) are expected to be a factor of 50 higher compared to the resources available today. Since available budgets are expected to remain at the current level at best, and technology improvements will cover only a fraction of the increase in the needs, a considerable gap still remains between resources needed and available, estimated to be somewhere between a factor 4 and 12 [1]. This will require optimisation of several aspects of experiment software, resource provisioning and usage, system performance and efficiency, data processing infrastructure.

HPC machines, such as *Piz Daint* [2] at CSCS, are increasingly powerful, they feature high-end hardware and are operated cost-effectively and competently in large specialised centres. They also profit from economy of scales when it comes to hardware procurements. This should translate to a cost benefit for the end customer, with the potential to deliver more computing for the same budget for the LHC computing use case.

Since HPCs are typically self-contained systems, subject to strict access and connectivity policies, one of the biggest challenges, however, is the transparent integration with the complex experiment data processing frameworks.

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

2. Piz Daint integration challenges

Piz Daint is a Cray XC50, CPU/GPU hybrid supercomputer, featuring 5320 hybrid and 1431 multicore (CPU only) compute nodes, with a total core count of 361,760 and a total peak performance of 25.326 Petaflops for the hybrid partition and 1731 Petaflops for the multicore partition. Featuring *NVIDIA Tesla P100* and *Xeon E5-2695v4 2.1 GHz*, over 0.5 PB of memory in total, Cray *Aries* routing and communications ASIC, and *Dragonfly* network topology, two Lustre *Sonexion* file systems for a total capacity of 8.7 PB and a peak performance of 138 GB/s. In addition, a GPFS file system [3] is interfaced to *Piz Daint*, dedicated exclusively to the LHC experiments data processing.

Also featured are two Cray proprietary technologies: *Data Virtualisation Service (DVS)* [4], allowing the interfacing of external systems to the internal Cray network, and *Data Warp Service* [5], which can provide posix file systems on demand on a *Burst Buffer* SSD-based storage layer.

As for most supercomputers however, some of the features are not suited to the typical High Energy Physics (HEP) workflows. Nodes feature no local disk, which has a negative impact on a lot of standard Linux workflows. The operating system is not what the HEP workflows expect: the *Cray Linux Environment* is a stripped down version of *SUSE*, designed to accelerate parallel software and not undergoing frequent upgrades. The memory available is limited: most of the nodes in the multicore partition feature 1 GB per core, with a limited amount of them featuring 2 GB per core. Network connectivity from/to the outside is not guaranteed, it must be negotiated with the centre. Data exchange implementations need gateways and interfacing external services is not straightforward, e.g. mounting a directory.

In order to address such challenges, the The Swiss HEP computing community and CSCS have started working on the HPC integration with the LHC experiment Tier-2 facilities in 2014.

3. ATLAS HPC integration at CSCS.

3.1. Non invasive approach

The first efforts back in 2014 came from the ATLAS collaboration [6] and resulted in succeeding to run the ATLAS Geant4 [7] simulation in production on a Cray XK7 at CSCS for over six months [8]. For this effort, we obtained access to a Cray integration system with one user account and a home directory for the user on the scratch area of the system. The integration scheme made use of a modified ARC Compute Element [9] submitting via *ssh* from outside the CSCS perimeter and managing the job lifecycle via a persistent *sshfs* mount of the job directories from the Cray scratch to the remote ARC CE.

3.2. The LHConCRAY project

Following this successful exercise, CSCS and the Swiss HEP community (ATLAS, CMS, LHCb) launched the *LHConCRAY* project. This aimed at integrating *Piz Daint* with the three LHC experiment frameworks and targeted all their workflows, including user analysis. The project ran for about two years in 2016-17, and went into Tier-2 level production in April 2017 with 1.6k cores, in order to measure performance in a production environment. For this project, CSCS relaxed some policies, allowing outbound IP connectivity from the supercomputer nodes, and allowing ARC CE services and a GPFS file system which is part of the classic Tier-2 centre implementation to be interfaced to the internal supercomputer network.

Late in 2017, a scale-up test was performed, which saw ATLAS running on up to 27k *Piz Daint* cores with Geant4 simulation jobs, and sustaining the pressure for the agreed test period of 6 hours. By the end of 2017, the decision was taken to migrate the CSCS WLCG Tier-2 facilities from *PHOENIX*, a dedicated linux cluster that has been operated for over 10 years, to *Piz Daint*. Over 4k cores are running in production since April 2017, expected to raise to over 10k cores by April 2019, when the transition will be completed and *PHOENIX* will be decommissioned.

Within the scope of the project, was also to agree on a pricing for the resources for up to the expected lifetime of *Pix Daint* (2021). The pricing is slightly favourable, allowing our pledged CPU resources to grow several % faster within this period, compared to the expected growth of the dedicated linux cluster. We are not allowed opportunistic usage of additional resources.

3.3. Further integration R&D

Over the past few months, some R&D has been initiated, in order to further integrate *Piz Daint*. These studies have involved commissioning the execution of ATLAS and CMS Tier-0 workloads, and are currently in progress.

4. System architecture

The current architecture, with *PHOENIX* and *Piz Daint* sharing the execution of the CSCS Tier-2 workflows is shown in Fig. 1 and a more detailed view of the supercomputer systems and services is in Fig. 2.

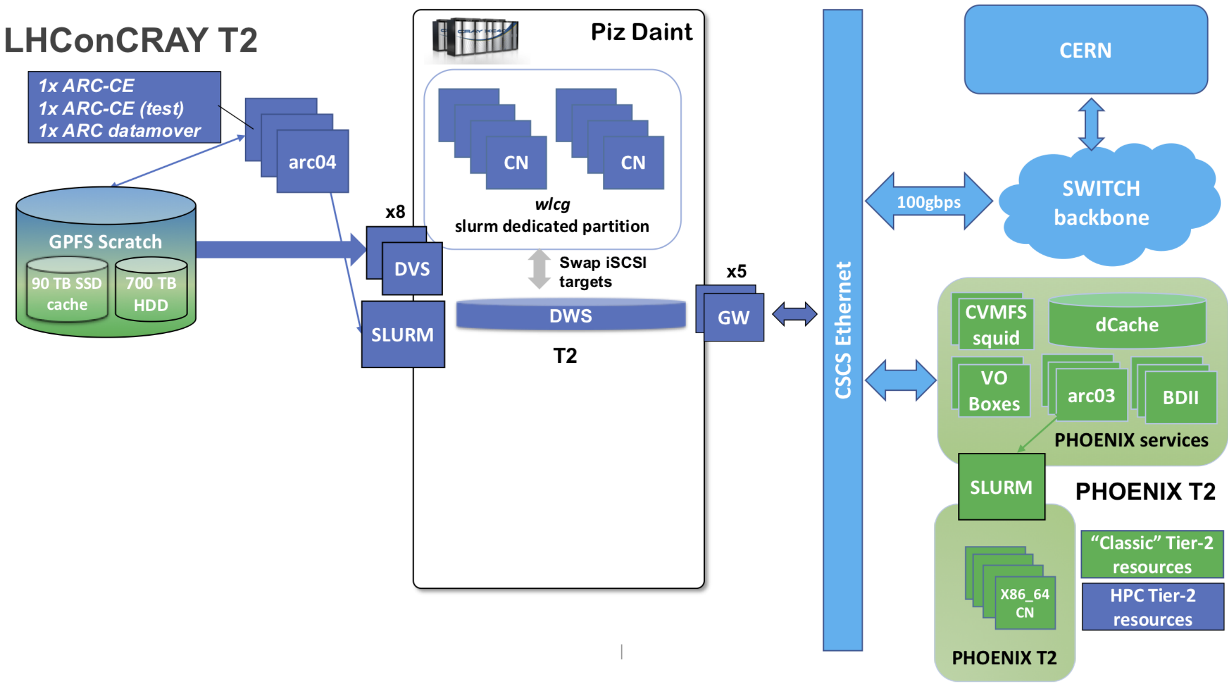

Figure 1. Shared CSCS Tier-2 architecture with PHOENIX and Piz Daint

4.1. HPC system architecture highlights

4.1.1. OS and memory requirements Jobs run on the supercomputer nodes in Docker containers using Shifter. The images are full WLCG compliant worker nodes and are sourced from one of the Cray own Lustre file systems. Each node features 128 GB of memory and 68 out of the 72 logical cores are used. Jobs are allowed to use an arbitrary number of cores and are not restricted to be node-exclusive as for all other HPC users. Memory is not made consumable in *slurm*, so its allocation is not tracked on the nodes and limits are not enforced, exception made for a 6 GB/core limit in order to catch rouge jobs. This allows to run workloads whose memory usage may peak well above 2 GB/core without compromising the stability of the system, since memory usage of a variety of workloads ends up balancing out.

4.1.2. Software provisioning Thanks to the HPC centre policy relaxation, we can *FUSE* mount *CVMFS* [10] on the nodes dedicated to the WLCG data processing. The cache makes use of a novel two-tier scheme developed specifically for this application [11]: the upper tier is in-RAM, for which 6 GB per node are allocated for the three VOs. The lower tier is preloaded on the GPFS shared file system. The performance and stability has been noted to be exceptionally good.

4.1.3. Network The supercomputer internal network is interfaced to the public CSCS network by means of gateways. The nodes have public IP addresses with standard Linux IP packet forwarding. Services that are external to the Cray systems, like GPFS and the ARC compute elements are interfaced by means of the Cray *Data Virtualisation Service*, providing the needed mount points for the shared areas, e.g., the job session directories for ARC, the *CVMFS* preloaded cache and the job scratch area.

4.1.4. Additional services The Cray *Data Warp* technology has ben investigated, mainly in order to provide on demand posix file systems. The main application is provisioning swap on demand. Also the provisioning of */tmp* areas on the compute nodes is under investigation.

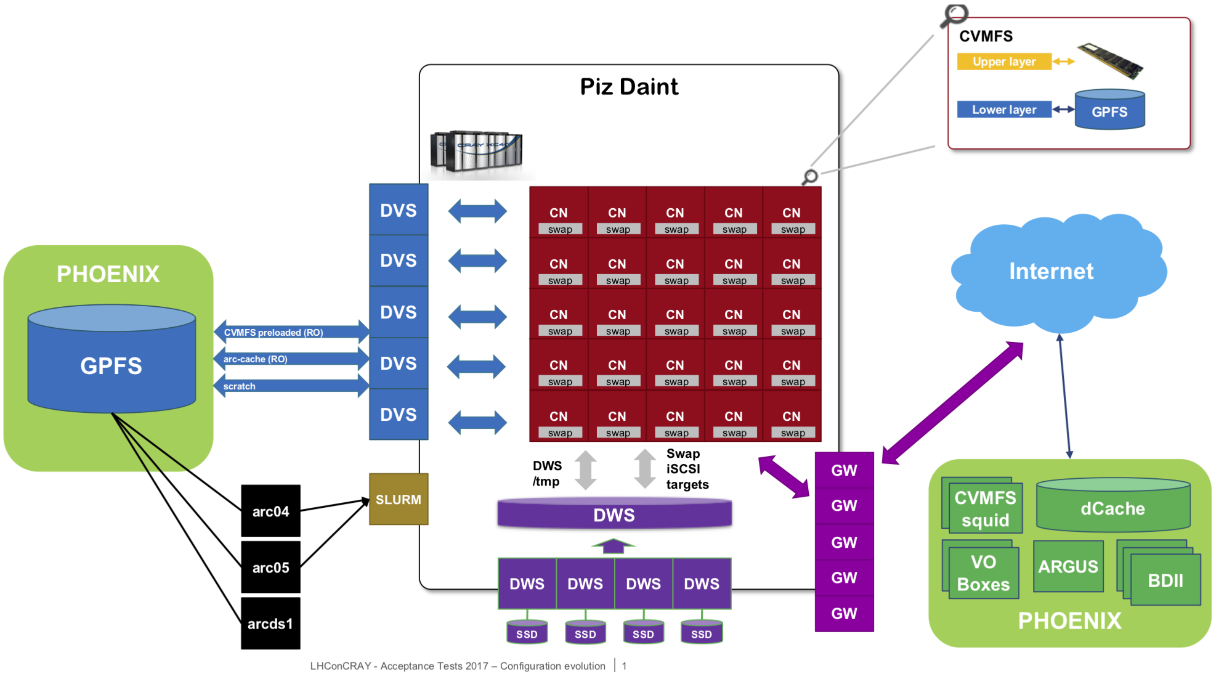

Figure 2. HPC system architecture highlights

5. Scale-up test and Tier-2 production performance

Shortly before the conclusion of the *LHConCRAY* project, a scale-up test was performed by ATLAS, in order to exercise the full framework chain at a scale. This consisted of running ATLAS Geant4 simulation jobs on up to 27k *Piz Daint* cores. The ramp-up time was approximately one hour and the job pressure was sustained for the agreed test period of six hours (Fig. 3).

Following the December 2017 decision to migrate the Tier-2 facility to the HPC services, the number of Tier-2 production cores has been increased from 1.6k to over 4k for the pledge period

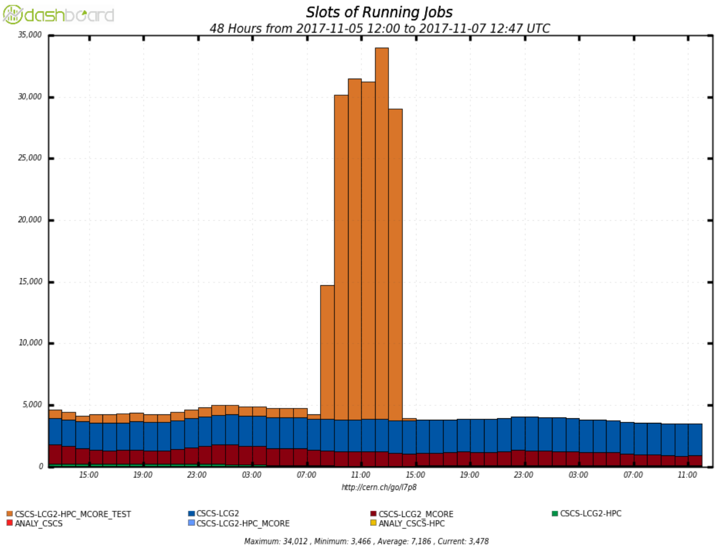

Figure 3. Number of running slots for ATLAS at CSCS during the scale-up test. In orange the slots used by the test, on top of the pedestal of the slots used by the PHOENIX Tier-2

starting April 2018. Since then, about half of the ATLAS Tier-2 workload are processed on the HPC, as shown in Fig. 4 (right). The ATLAS nominal share of the total resources is 40%.

In order to measure the production performance of the systems, we look at:

(i) Wall-clock efficiency based on successful over all jobs

(ii) CPU/Wall-clock efficiency for the successful jobs

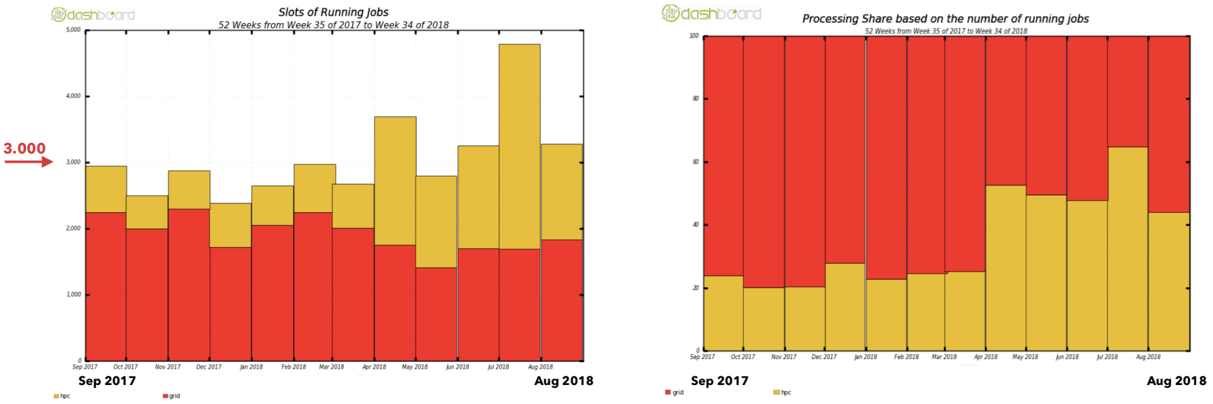

Figure 4. 1-year history of the tier-2 slots used by ATLAS (left) and relative processing share (right) for PHONENIX (red) and Piz Daint (yellow)

The two quantities are shown for the period of one year in Fig. 5, respectively left and right, and can be seen to be equivalent within fluctuations for the two systems. The slight dip observed in July 2018 for the HCP systems is due to some R&D activities beyond the Tier-2 that are detailed in the next section. When looking at these metrics only, there is neither advantage nor disadvantage in running on either system. This is largely due to the fact that both systems share the GPFS scratch file system and the site *dCache* [12] storage element. These two systems are understood to be the dominating factor with regard to the measured efficiencies.

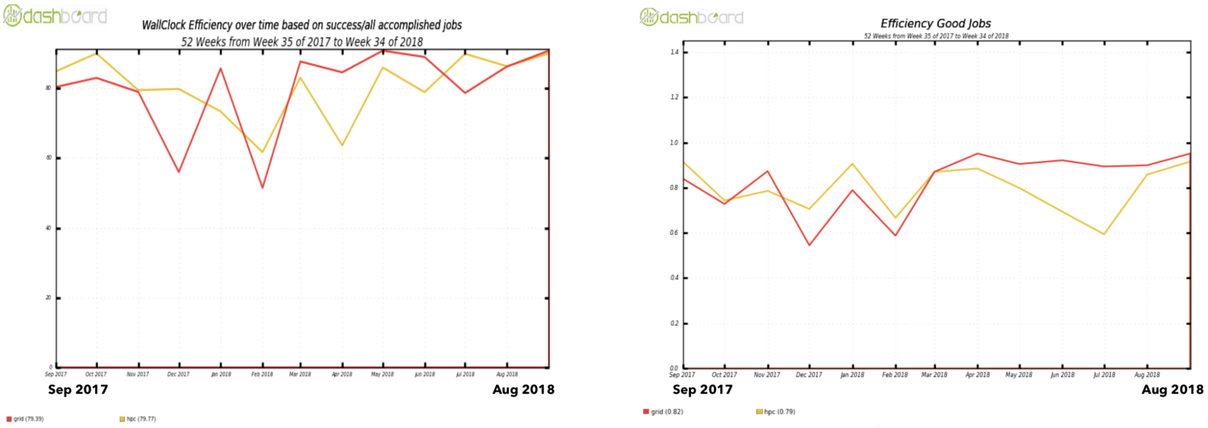

Figure 5. 1-year history of efficiencies for ATLAS on PHONENIX (red) and Piz Daint (yellow). The average values are: PHOENIX 79%, HPC 80% for the left plot; PHOENIX 82%, HPC 79% for the right plot

6. Additional integration R&D

In the wake of the successful HPC integration at CSCS, the ATLAS computing community has arranged to pursue further the venture by launching an R&D project aimed at integrating the prompt reconstruction of the experiment RAW data, an activity performed exclusively at the Tier-0 at CERN. Such activities could not be carried out if we were to be restricted to a dedicated grid cluster such as *PHOENIX*, since that would not allow the provisioning of additional large scale resources on demand in order to accommodate computational peaks. A powerful HPC machine by contrast offers such opportunities.

The integration scheme for the Tier-0 workloads is based on the proven Tier-2 architecture. The challenge however, goes a few steps further, since such workloads are more demanding in terms of resources: higher memory consumption and heavier input/output rates and patterns. Further to that, the integration foresees to run such jobs on the nodes that are shared with other users. The Tier-2 nodes, by contrast, are dedicated to the WLCG jobs, which is crucial, since in the architecture we have described, these nodes are configured differently compared to those used by the general HPC users.

This activity has made very good progress, and ATLAS have reached an integration level that could allow to go to production. Also CMS have decided to carry out a similar exercise and their work is in progress. Detailed reports will follow in the future.

7. Conclusions

We have successfully integrated the HPC systems of *Piz Daint* at the CSCS with the ATLAS and other LHC experiment computing frameworks, and have started migrating the main Swiss Tier-2 facility from a dedicated linux cluster to the shared HPC systems. The integration effort has taken place over the course of over two years and has partially overlapped with the production phase. We operate fully in production on the HPC systems since April 2018 and expect to complete the transition by April 2019. Both systems perform comparably in terms of CPU efficiencies, but we can profit from slightly more favourable pricing conditions for the life time of *Piz Daint*.

The integration has opened the doors to further activities targeted at on demand resource provisioning to accommodate computational peaks, novel computing models and software optimisation, which would not have been possible if restricted to the dedicated linux cluster. In addition, we hope that the requirements of the HEP community will be taken into account during the specification process of the next generation HPC at the centre.

Acknowledgments

We acknowledge the support of the Swiss National Science Foundation and thank CSCS for the provision of the integration systems.

Copyright

[2018] CERN for the benefit of the [ATLAS Collaboration]. CC-BY-4.0 license.

References

[1] Schmidt B *The High-Luminosity upgrade of the LHC: Physics and Technology Challenges for the Accelerator and the Experiments* JPCS **706** 2016 022002

[2] https://www.cscs.ch/computers/piz-daint/

[3] Schmuck F and HAskin R *GPFS: A Shared-Disk File System for Large Computing Clusters* Proceedings of the FAST'02 Conference on File and Storage Technologies, USENIX, 2002 231-244

[4] Sugiyama S and Wallace D *Cray DVS: Data Virtualization Service* in Cray User Group Conference (CUG) 2008

[5] Liu N, Cope J, Carns P, Carothers C, Ross R, Grider G, Crume A, and Maltzahn C, *On the role of burst buffers in leadership-class storage systems* in Mass Storage Systems and Technologies (MSST), 2012 IEEE 28th Symposium on. IEEE 2012 1-11

[6] ATLAS Collaboration *The ATLAS Experiment at the CERN LHC* JINST **3** 2008 S08003

[7] Agostinelli S et al *Geant4, a simulation toolkit* Nucl. Instrum. Meth. A **506** 2003 250-303

[8] Hostettler M *Enabling the ATLAS Experiment at the LHC for High Performance Computing*, Masterarbeit an der philosophisch-naturwissenschaftlichen Fakultaet der Universitaet Bern 2015

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

[10] Blomer J and Fuhrmann T *A Fully Decentralized File System Cache for the CernVM-FS* Computer Communications and Networks (ICCCN) 2010 1-6

[11] Blomer J et al *New directions in the CernVM file system* J. Phys.: Conf. Ser. 2017 **898** 062031

[12] The dCache Project, http://www.dcache.org