Integration of the Chinese HPC Grid in ATLAS Distributed Computing

A Filipčič[1] for the ATLAS Collaboration

[1] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

E-mail: andrej.filipcic@ijs.si

Abstract.

Fifteen Chinese High-Performance Computing sites, many of them on the TOP500 list of most powerful supercomputers, are integrated into a common infrastructure providing coherent access to a user through an interface based on a RESTful interface called SCEAPI. These resources have been integrated into the ATLAS Grid production system using a bridge between ATLAS and SCEAPI which translates the authorization and job submission protocols between the two environments. The ARC Computing Element (ARC-CE) forms the bridge using an extended batch system interface to allow job submission to SCEAPI. The ARC-CE was setup at the Institute for High Energy Physics, Beijing, in order to be as close as possible to the SCEAPI front-end interface at the Computing Network Information Center, also in Beijing. This paper describes the technical details of the integration between ARC-CE and SCEAPI and presents results so far with two supercomputer centers, Tianhe-IA and ERA. These two centers have been the pilots for ATLAS Monte Carlo Simulation in SCEAPI and have been providing CPU power since fall 2015.

ATL-SOFT-PROC-2017-041
03 February 2017

1. Chinese HPC CNGrid

The Chinese HPC CNGrid [1] is a network of 15 High-Performance Computing centers which participate in a transparently accessible infrastructure, including the top-class largest sites such as MilkyWay 1 and 2 supercomputers. Some of those supercomputer centers are interested to provide opportunistic CPU resources to ATLAS [2], such as CNIC ERA [3] and Tianhe-1A in Tianjin [4].

The SCEAPI [5] is a RESTful middleware interconnecting the HPC centers, that provides secure https access, while authentication and authorization is based on username and password with a token stored in a json file (Figure 1). The registered applications are pre-installed on the selected HPC centers where a user has active allocation. The targeted HPC center is chosen at the job submission level in the job description file. The file interface is used to upload the input files to the HPC center and download the produced output files to the client machine. The jobs can be monitored and controlled through the job status interface and the job and file management interface.

SCEAPI architecture resembles the Compute Element as used in the grid, with the difference that it is highly optimised for submission of massively parallel jobs. The supercomputers do not provide an outbound connectivity, so all the data management and transfers need to be done externally. Only approved applications are able to run, so the ATLAS job wrappers are fixed and pre-installed on the supercomputers, while the payload itself can be submitted as an input

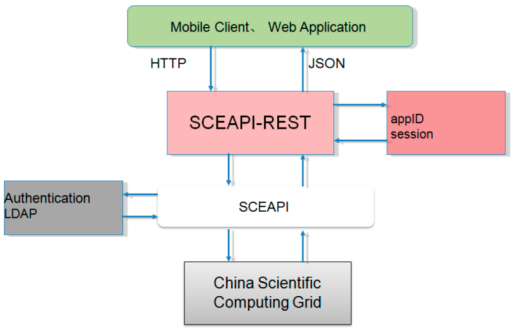

Figure 1. SCEAPI Interface.

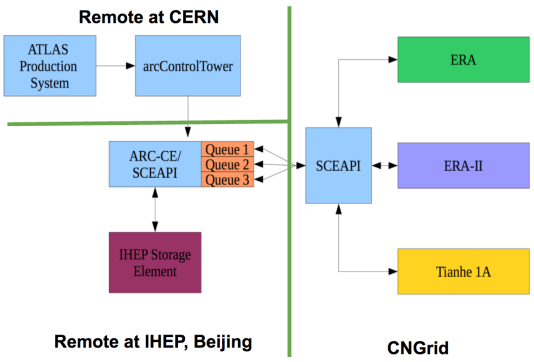

Figure 2. Submission of ATLAS jobs through arcControlTower and SCEAPI.

file. The ATLAS software needs to be installed locally on the HPC center shared filesystem, and the deployment is done by the supercomputer administrators.

2. Connecting CNGrid to ATLAS Production System

arcControlTower [6] and ARC Compute Element (ARC-CE) [7] have been used since 2007 to submit ATLAS jobs to sites architected around Nordugrid middleware [8] with a distributed dCache [9] NDGF Tier-1 storage [10]. The data transfers are controlled by ARC-CE using native input file caching, which is suitable for sites with a capable shared filesystem.

The ARC-CE backend was extended to submit to SCEAPI (Figure 2), where the latter was treated just as another batch system flavour. The ARC-CE queue which usually submits to a specific batch queue, is mapped to a targeted HPC center. A single ARC-CE instance is therefore used to submit ATLAS production jobs to all available CNGrid HPC centers.

Supercomputers are usually designed for CPU-intensive payloads, therefore only ATLAS Monte-Carlo GEANT-4 [11] simulation was enabled for job execution. In addition, a typical Monte-Carlo campaign spans a period of several months thus limiting a software installation to a single ATLAS software release. A part of ATLAS CernVM-FS [12] tree was packed into tarballs, including the targeted ATLAS software release, supporting common software such as compilers, wrappers and third-party libraries, and the geometry and ATLAS detector conditions packed in a DBRelease sqlite distribution. A deployment script was provided to the supercomputer system administrators to install the software and fix for the path relocation on the shared filesystem. A new version of software is installed by CNGrid upon the request by ATLAS.

3. Job Submission to CNGrid

Each HPC center has its own PanDA [13] queue, for example BEIJING-ERAII_MCORE and BEIJING-TIANJIN-TH-1A_MCORE. arcControlTower submits activated jobs to ARC-CE at IHEP Tier-2 in Beijing. ARC-CE then transfers the input data from the IHEP Storage Element to a selected CNGrid HPC center and subsequently submits the payload through SCEAPI. After job completion, the outputs are delivered to the IHEP Storage Element.

ATLAS was using 3 HPC centers in 2016 (Figure 3), although the usage was rather sporadic during that period. The Monte-Carlo simulation campaign was not active all the time, there were longer maintenance periods in both IHEP and CNGrid. The interface is not yet bug free and requires frequent manual interventions to fix unexpected issues.

Nevertheless, CNGrid has simulated about 1% of ATLAS Monte-Carlo events contributing 3.5M CPU hours on processors which are twice faster than the ones on a typical grid site.

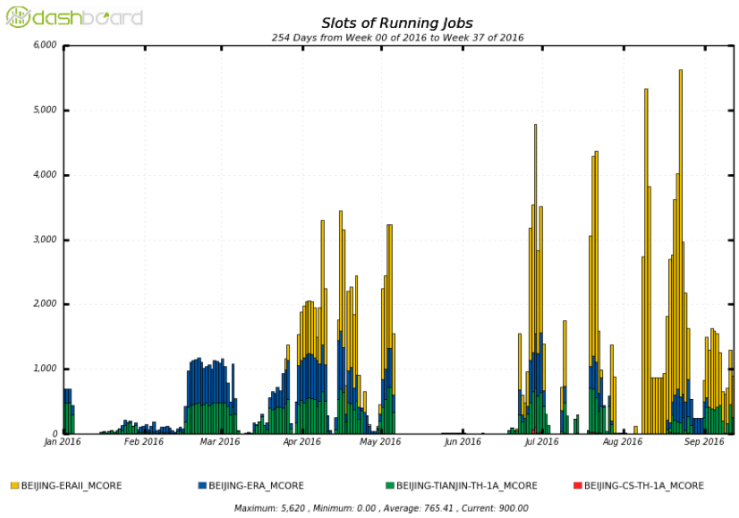

Figure 3. ATLAS Production on Chinese HPC centers.

4. Conclusions

ATLAS is very efficient in including non-standard sites into its production system as was already demonstrated by ATLAS@Home [14] and submission to computing clouds. ARC-CE architecture was the most appropriate to extend the submission mechanism to the SCEAPI RESTful interface by customising the batch system backend and by exploiting the ARC-CE data management and caching support. ATLAS used limited resources at CNGrid for production up to now, but there are possibilities for a significant cputime allocation on several of the worlds largest machines in the near future.

References

[1] CNGrid web Page URL http://www.cngrid.org/
[2] ATLAS Collaboration 2008 *JINST* **3** S08005
[3] ERA HPC web Page URL http://english.cnic.cas.cn/ns/es/201407/t20140707_123836.html
[4] Tianhe-1A HPC web Page URL http://nscc-tj.gov.cn/en/index.asp
[5] Rongqiang C et al 2016 SCEAPI: A Unified Restful Web APIs for High-Performance Computing Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.
[6] Filipcic A 2011 *J. Phys.: Conf. Ser.* **331** 072013
[7] Ellert M, Grønager M, Konstantinov A *et al.* 2007 *Future Gener. Comput. Syst.* **23** 219–240 ISSN 0167-739X
[8] The NorduGrid Collaboration web site URL http://www.nordugrid.org
[9] de Riese M *et al. dCache Book* URL http://www.dcache.org/manuals/Book/Book-a4.pdf
[10] NeIC web site URL https://neic.nordforsk.org/activities/nt1/
[11] Geant 4 web site URL http://geant4.cern.ch
[12] CernVM File System web site URL http://cernvm.cern.ch/portal/filesystem
[13] Maeno T et al, on behalf of the ATLAS Collaboration 2011 *J. Phys.: Conf. Ser.* **331** 072024
[14] Cameron D et al, on behalf of the ATLAS Collaboration 2016 Volunteer computing experience with ATLAS@Home. Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.