EXPERIENCE WITH ITEP-FRRC HPC FACILITY

I.E. Korolko [a], M.S. Prokudin [b] and V.A. Kolosov [c]

Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of National Research Centre "Kurchatov Institute"

E-mail: [a] Ivan.Korolko@cern.ch, [b] Mikhail.Prokudin@cern.ch, [c] Victor.Kolosov@itep.ru

ITEP-FRRC HPC facility was built in 2011-2014 as a common project of State Atomic Energy Corporation ROSATOM and Helmholtz Association of German Research Canters. It utilizes the concept of "green computing" which was invented by GSI/FAIR scientists. Facility is used for FAIR related studies by various groups from ITEP and other Russian physics centers. After 7 years of successful HPC facility operation we want to summarize the experience we got running the hardware and supporting the requested software.

Keywords: high performance computing, FAIR project

1. Introduction

The success of the FAIR experiments depends to a large extent on the availability of an adequate computing infrastructure to support their physics program. While each of the four FAIR pillars APPA, CBM, NUSTAR and PANDA has its own challenging requirements, from the computing perspective mainly CBM and PANDA are setting the scale. Both will have a data rate of up to 1 TeraByte/s coming from the detector systems into the computing center at FAIR. This enormous amount of data needs to be processed, the volume reduced without affecting the physics quality and finally stored. Each of the experiments will store several PetaBytes per year on the permanent mass storage system. The raw data has to be further processed for event reconstruction and physics analysis. In addition detailed detector simulations for the various physics signals are required to achieve the final physics results. These tasks are traditionally distributed between the computing centers of the participating countries. For the FAIR experiments a similar approach is envisioned. While the exact requirements are currently under study, depending on the exact detector technology chosen, first estimates are available. For the core functionality at the FAIR Tier 0 computing center, 300.000 cores and 25 PetaBytes of both disk and permanent mass storage per year are required. For a timely physics analysis, this has to be augmented with at least similar resources provided by external partners.

In 2011, group of enthusiasts from GSI and ITEP came forward with a proposal to strengthen the IT infrastructure available for FAIR-Russia. Rosatom and Helmholtz approved the proposal which aims at building a super-computer center serving as the TIER-1 FAIR center in Russia at FRRC and ITEP with an excellent IT network between the Russian institutes. Funded by Helmholtz, a mobile container with high density water cooled racks [1] coupled to an exterior water cooling tower has been designed for the climate of Moscow and has been transported to the FRRC/ITEP. As of September 19, 2014, this High Power Computing Center (HPC) has been successfully commissioned and is now fully operational with 10240 cores, interconnected with Ethernet and Infiniband networks. For theory groups we have 5 special servers equipped with 8 (each) GPU accelerators, with 40 TFlops peak performance. The HPC center is also equipped with a 1080 TB mass storage system. Supercomputer container is shown in Figure 1.

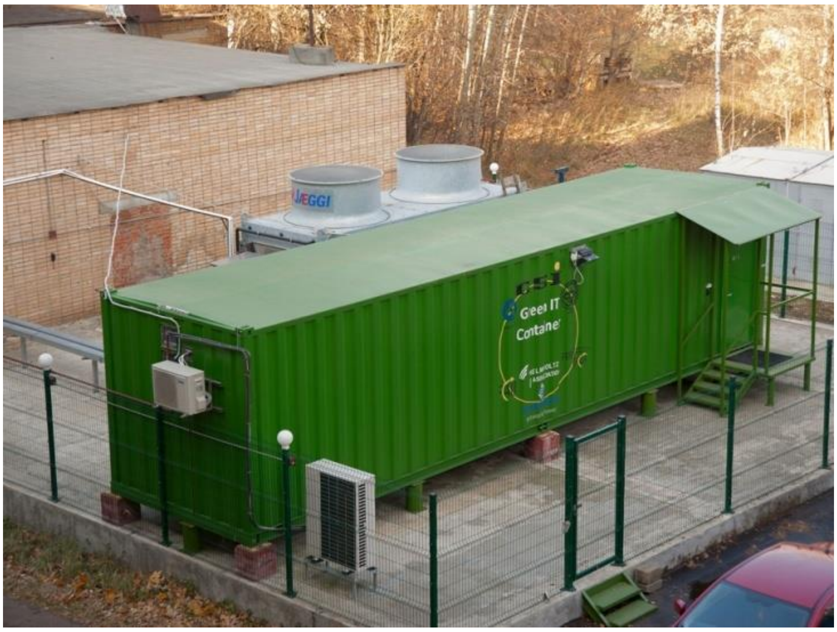

Figure 1. ITEP-FRRC HPC facility

2. ITEP-FRRC HPC facility

The hardware for supercomputer was chosen to meet the following requirements:

- compact enough to provide more computational power per unit in standard rack;
- acceptable power consumption and as a consequence heat emission;
- provide enough possibilities for remote administration (IPMI, KVM etc);
- Ethernet for access to some common local and external resources (1GbE is considered to be enough for these purpose) and Infiniband for data exchange (memory-memory) in the parallel calculations and for access to the shared parallel file system.

We choose 7U blade system with 10 twin modules (20 2xCPU nodes per blade) as a solution which fits the requirements in a best way. Each node has:

- 2xCPU AMD 6272 (16 cores, 2.1 GHz);
- 64 GB RAM (2 GB per core);
- 1xGbE and 1x 4xQDR links (fault tolerance is not an issue for such systems).

The supercomputer as a whole has 10240 cores and power consumption up to 120 kW (10-16 kW per rack). Cooling system consumes up to 15 kW and provide PUE (Power Usage Effectiveness) of 1.13 or better (depending on season). For blade servers interconnections we used fat tree topology of Infiniband network. Lustre high performance distributed parallel file system has capacity of 1080 TB. Linpack tests give the productivity of the supercomputer approximately 58 Tflops. In addition to pure CPU based system we have hybrid segment with GPU Nvidia Kepler K20X (20 nodes with 2 GPU in each). Each node has 64 GB RAM, 1xGbE and 1x 4xQDR links. GPU segment gives us approximately 40 Tflops of computing power.

The following software is installed on every node:

- CentOS 6.x as operating system;
- Torque batch system with maui scheduler;
- OpenMPI with Torque integration (local build);
- OpenMP (for tasks which do not “like” OpenMPI);
- BLAS, lapack including ATLAS versions, ACML;
- CUDA 6.5.14 (for GPU nodes).

We used Redhat kickstart based infrastructure for software installation with networked PXE boot for automation of the system deployment process. Initially we used the custom script based system for package and configuration management but recently we started the migration to Salt open source configuration management software and remote execution engine. This engine provides several models of the system management: centralized server-client with cronbased execution of the batches (“states” in terms of Salt), event processing etc, clients with execution “states” directly on the client, centralized using ssh as a transport like in Ansible.

The monitoring system is an essential part of the supercomputer software. This is our eyes to watch for every part of the supercomputer infrastructure: hardware, operating system and the software. We choose Zabbix as a powerful solution for monitoring with great possibilities for customization. The key features of Zabbix which were important for us:

- Possibility to use the agent in active mode which is much more efficient for work in large systems;
- SNMP monitoring including possibility of low level discovery of configurations and processing of SNMP traps;
- IPMI monitoring out of box though this part were replaced custom scripts which were more efficient for large system;
- Triggers and events for group of hosts including the usage of the aggregate functions;
- Powerful and flexible tools for triggers and actions configuration;
- Presentation of the data in many ways including usage of Zabbix API.

3. Software installation for FAIR users

Many FAIR experiments use FairSoft and FairROOT software packages for simulation, reconstruction and physical analysis. Total size of program libraries and tables included in these packages is as large as 1400 megabytes for FairSoft and about 200 megabytes for FairROOT. Allocation of FairROOT and FairSoft at network disks is undesirable because it will produce an addition network and storage load and will increase run time of each job in case of massive data production/reconstruction. Alternative solution is installation of these packages at each computing node of supercomputer. So program libraries should be prepared for the installation according to supercomputer system administrator requirements:

1. All program libraries, physical tables and other data, to be installed locally at the computing nodes, should be packed into RPM-files compatible with CentOS 6 64-bit, which is used as operation system at supercomputer.

2. Provided RPM-files should be installed with command of the form:

```
rpm -i --relocate /=/opt/dirname softwarename-XX.YY.x86_64.rpm
```

, where softwarename is the name of the package to be installed (FairSoft in our case), XX.YY is its version (year and month of the release in case of FairSoft) and /opt/dirname is an installation destination. Using such bulky installation command ensures integrity of operation system files and data in case of malformed RPM-package.

4. Gained experience

After 7 years of running ITEP-FRRC HPC facility we have learned the following:

1. One should never believe in advertising promises. It is much more efficient to test the performance of different hardware solutions for the typical applications in your field;
2. The amount of RAM is an issue, especially for multi core systems. Seven years ago 2 GB per core looked enough, while today 8Gb is required;
3. Three-five years warranty is required for the purchased hardware, most of the failures happens during first year of operation. Special attention is needed for server power supplies where failure of one dollar fan could require the replacement of 100 dollar block;
4. Simulation of experimental setups and theoretical tasks perfectly fit in modern HPC architecture, while analysis of big experimental (and Monte Carlo) data samples still require optimization;
5. Green cooling system has demonstrated adequate performance in Moscow weather conditions, as illustrated in Figure 2. For two hot summer months additional active cooling could be useful. Small computing facilities should be protected from direct sun.

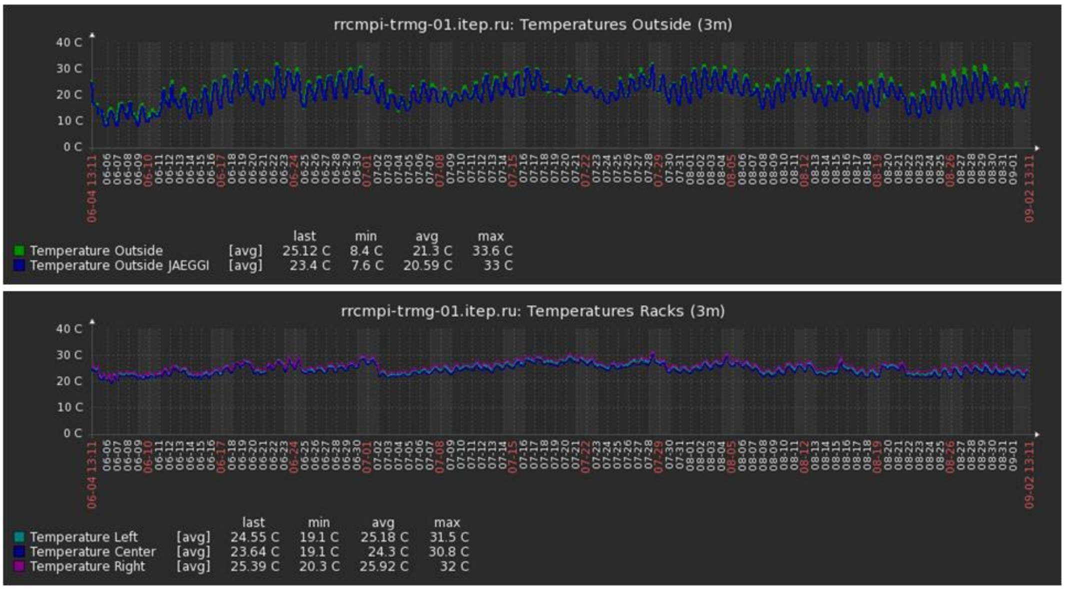

Figure 2. Air temperature outside (top plot) and inside (bottom plot) container during three summer month in 2018

5. Conclusion

In 2016 scientists from GSI and ITEP have prepared a proposal for a Russian FAIR Tier 1 center, containing about 100 000 cores, 400 GPUs, 20 PetaByte disk space, and a power consumption of about 3 MW, based on ‘green computing’ technology. This Russian center should also be linked to a high speed network to all Russian FAIR Institutes and to FAIR at Darmstadt.

References

[1] M. Bach, M. Kretz, V. Lindenstruth, D. Rohr, Optimized HPL for AMD GPU and multi-core CPU usage, Comput. Sci. 26 (3-4) (2011) 153-164. doi:10.1007/s00450-011-0161-5.