Applications of supercomputer Tianhe-II in BESIII

Jing-Kun Chen[1], Bi-Ying Hu[1], Qiu-Mei Ma[2], Jian Tang[1], Ye Yuan[2], Xiao-Mei Zhang[2], Yao Zhang[2], Wen-Wen Zhao[1]

[1]School of Physics, Sun Yat-sen University, Guangzhou, 510275, China
[2]Institute of High Energy Physics, Beijing, 100049, China

E-mail: maqm@ihep.ac.cn, tangjian5@mail.sysu.edu.cn

Abstract. High energy physics experiments are pushing forward the precision measurements and searching for new physics beyond standard model. It is urgent to simulate and generate mass data to meet requirements from physics. It is one of the most popular areas to make good use of existing power of supercomputers for high energy physics computing. Taking the BESIII experiment as an illustration, we deploy the offline software BOSS into the top-tier supercomputer "Tianhe-II" with the help of Singularity. With very limited internet connection bandwidth and without root privilege, we synchronize and maintain the simulation software up to date through CVMFS successfully, and an acceleration rate in a comparison of HPC and HTC is realized for the same large-scale task. There are two creative ideas to be shared in the community: on one hand, common users constantly meet problems in the real-time internet connection and the conflict of loading locker. We solve these two problems by deployment a squid server and using fuse in memory in each computing node. On the other hand, we provide a MPI python interface for high throughput parallel computation in TianheII. Meanwhile, the program to deal with data output is also specially aligned so that there is no queue issue in the I/O task. The acceleration rate in simulation reaches 80% so far, as we have done the simulation tests up to 15 K processes in parallel.

1. Introduction

Precision measurements and new physics searches require massive computation in high energy physics experiments. Taking the BESIII experiment as an illustration, we deploy the offline software BOSS into the top-tier supercomputer "Tianhe-II" with the help of Singularity. With very limited internet connection bandwidth and without root privilege, we synchronize and maintain the simulation software up to date through CernVM-FS successfully, and an acceleration rate in a comparison of HPC (High Performance Computing) and HTC (High Throughout Computing) is realized for the same large-scale task. We deploy a squid server and use fuse in memory in each computing node to update docker. We provide a MPI python interface for high throughput parallel computation in Tianhe-II. Meanwhile, the program to deal with data output is also specially aligned so that there is no queue issue in the I/O task.

2. BESIII

The Beijing Electron Positron Collider (BEPC), designed to operate τ-charm energy region, and its detectors, the Beijing Spectrometer (BES) and the upgraded BESIII[1], were operated at the Institute of High Energy Physics Chinese Academy of Science (IHEP) in Beijing. BEPC focuses on investigating τ-charm physics and Hadron physics, with the collision energies in the range from 2 to 5 GeV and the peak luminosity of $\sim 1 \times 10^{33} cm^{-2} s^{-1}$, which is the highest luminosity in τ-charm physical energy zone in the world.

3. Supercomputer Tianhe-II

Tianhe-2 is one of the TOP500 supercomputers based at Sun Yet-sen University's National Supercomputer Center in Guangzhou[2]. It has a total of 16000 worker nodes, each with 24 CPU cores, a 15PB shared file system, a self-developed internal high-speed network similar to Infinite Band, and 54.9PFlops in Linpack testing. Tianhe-2 network topology is shown in Fig.1. Users connect to login nodes via VPN, a worker node can only connect to login nodes, another worker node and functional nodes can connect to the internet via a high-speed internal network, and all other nodes except the cloud can connect to the internet for security reasons.

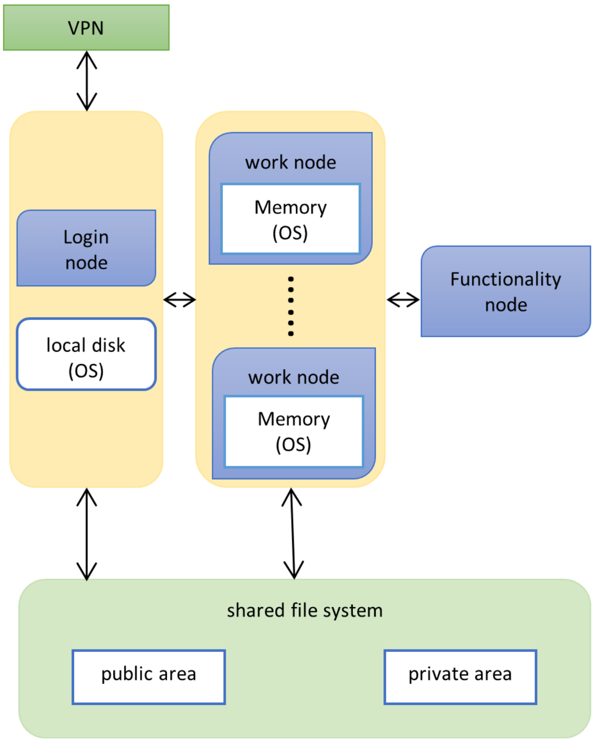

Figure 1. The network topology of Tianhe-II

4. Install CVMFS on Tianhe-II

The CernVM-File System (CVMFS) provides a scalable, reliable and low-maintenance software distribution service, which is a very important and widely used tool for HEP groups, including BESIII[3]. There is no universal method for installing CVMFS on all HPCs because not all HPCs can support it. After consulting with some fellow workers, including the developer of CVMFS, we went through the following approaches.

- General install: A traditional method is installing a binary CVMFS client package in path "/CVMFS" by yum or apt-get command. However, as a normal user, we cannot use yum, which is prohibited in Tianhe II, to install a package in a root-only path. We install a binary CVMFS client package in path "/CVMFS" and deploy a dedicated worker node with connection from "/CVMFS" to the file system. However, the dedicated worker node prevents us from using abundant resources of Tianhe-II. Also, the limitation of mount lock restricted the numbers of worker node which was mounting the code-base.

- Using Cvmfsexec: Using Cvmfsexec, we set the mount destination on ”/tmp” to get around the mount lock and root privileges. However, we meet two new problems. The first is that CVMFSexec requests the CVMFS libraries in the root belong path ”/CVMFS/lib”, which means it needs a unique OS and limits the computing resources. The second is that CVMFSexec reports an unresolved warning: “unshare: unshare failed: Invalid argument”.
- Parrot-mount: We attempt to install CVMFS through virtual machine Parrot to avoid the root-only path “/CVMFS”. Nevertheless, virtual machine cannot be installed on Tianhe-II, which means this method has also failed.

We discover that the fundamental challenge is a binary package of CVMFS or CVMFSexec that requires a path that belongs to root after making the aforementioned attempts. As a result, compiling CVMFS from source code is one solution to this problem. CVMFS is successfully installed in a public location after compilation, and anyone can use it in any worker node without root privileges. With this approach, we deploy the CVMFS system in the HPC and enabled frequent software updates, as is customary in an HTC.

5. Network topology

Like the majority of supercomputers worldwide, Tianhe-II lacks a direct internet connection for security reasons. Thus, it is necessary to establish the network connections with the computing center in IHEP and ensure the data synchronization during the submission of large-scale simulation tasks. The whole network topology between IHEP and Tianhe-II is shown in Fig.2.

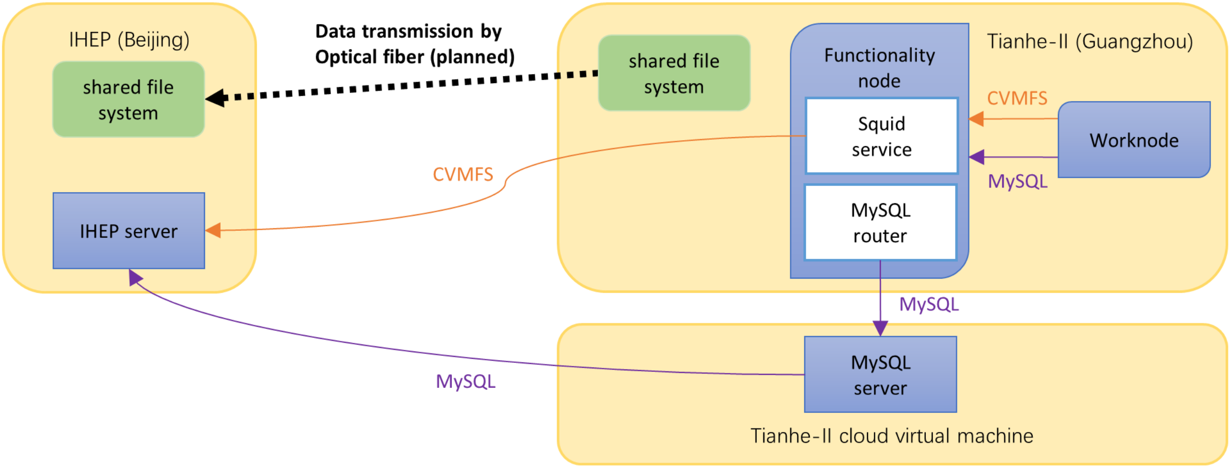

Figure 2. Network topology between IHEP and Tianhe-II

- Data transfer: Data transfer is another issue in a whole use flow, especially the amount of input and output data of HEP experiment is up to TB magnitude. We made a speed test in a realistic environment and find the transfer speed is just up to 20MB/s.
  A speed of 20MB/s is enough to work but not comfortable. As a result of that, we still try to find a faster scheme. The first plan is to use IPV6 instead of IPV4, however IHEP using CASnet and Tianhe-II using EDUnet, a message from CASnet IPV6 address to

EDUnet IPV6 address will be blocked by EDUnet's router. The second plan is to build a Internet Leased Line(ILL) between Guangzhou Nation Super Computer Center(NSCC) and Dongguan China Spallation Neutron Source(CSNS). IHEP has a 10GB and 2GB ILL between Beijing database and CSNS. The price of ILL between NSCC and CSNS is depends on different schemes.

- MySQL: Another support service needed by running BOSS is a slaver MySQL. We install a MySQL in Cloud platform virtual machine and follow the manual to set it be a IHEP's MySQL slaver. A problems is Cloud platform and HPC can't connect directly for security reason. So, we also deploy MySQL Router in functional nodes. The application of MySQL Router[4] is similar to squid but MySQL Router is a proxy especially for MySQL. Cloud platform can easily expand the number of virtual machine.

6. Validation of BOSS

BESIII Offline Software System (BOSS)[5] is an object-oriented data processing software system, mainly used for the simulation, calibration, reconstruction and analysis of the data collected by BESIII. BOSS utilizes the C++ language and GAUDI framework[6] on the Scientific Linux CERN (SLC) operating system with CMT[7] as the configuration management tool and MySQL as the data server.

We are able to operate any version of BOSS at Tianhe-II thanks to CVMFS. To ensure the accuracy of the data, which may be influenced by differences in operator systems, we ran the same simulation script at Tianhe-II and IHEP and compared the results.

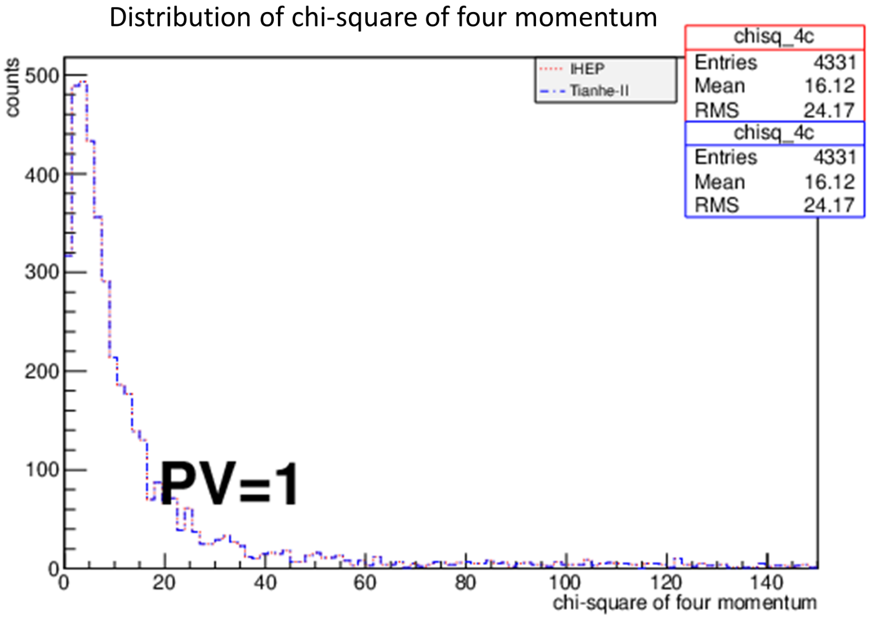

Figure 3. chi-square values of the four momentum

Part of the essential results such as χ^2 values of the four momentum is given as a demonstration of validation in Fig. 3. The results shows that data from two platforms are completely consistent.

7. A Large-scale performance test

The Tianhe-II SLURM system could become stuck or crash if numerous large-scale HTC jobs are submitted one by one in a short time. Also, it is not a good idea to input/output too many files from/to the share file system. Thus, we develop a MPI python submitting script to pack the HTC jobs and control the I/O. Also, we create initial files in memory instead of reading in file system, and save the output in memory first and move to file system in order.

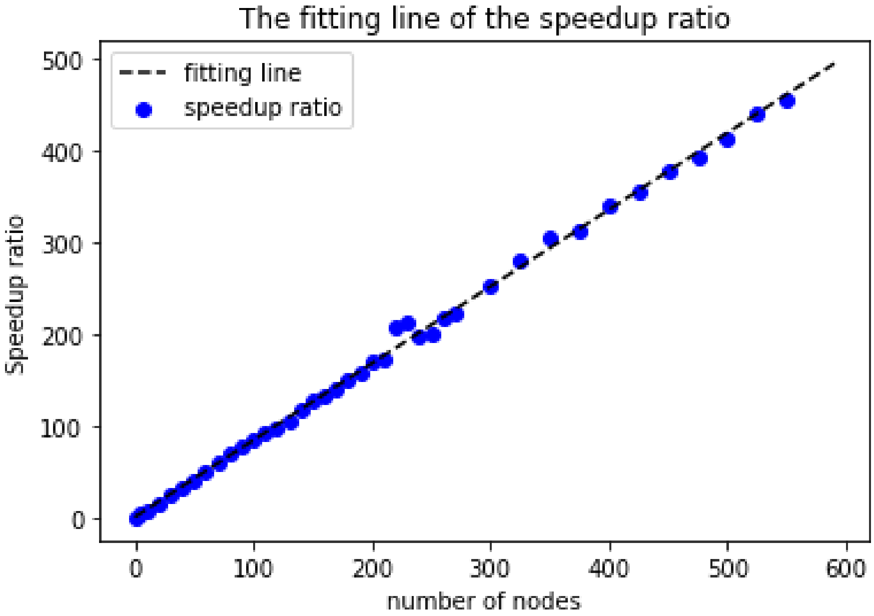

Figure 4. the speedup ratio with 500 nodes

The speedup ratio results are shown in Fig.4. The slope of the speedup ratio fitting line is 0.768, which is close to 1, and the acceleration rate in simulation reaches 80% so far, as we have done the simulation tests up to 15 K processes in parallel.

8. Summary and conclusion

In this paper, we present our work on resolution in a supercomputer application for the HEP experiment, using BOSS at Tianhe-II as an example. First, in order to support regular users who may access the executable file via HPC worker nodes, we have installed a Squid and constructed a CVMFS by compiling the source code at Tianhe-II. Subsequently, the MySQL server and MySQL-router were set up to facilitate the real usage. Our network speed testing have also shown that the IPV4 port is sufficient for the demonstration work. However, further support, such as IPV6 or ILL, will be required going forward. Following that, we have finished a large-scale simulation test to show that Tianhe-II can perform HTC jobs and provide helpful recommendations along the way. Lastly, we have demonstrated the ability to send a large-scale job to Tianhe-II from the IHEP server, however, additional work has to be done to optimize it in a more user-friendly manner.

Acknowledgments

We acknowledge the National Key R&D Program of China under Contract No. 2020YFA0406304. We thank the support of the National Natural Science Foundation of China (NSFC) under Grants U1832204, U1932105 and 12175259.

References

[1] Medina Ablikim, ZH An, JZ Bai, Niklaus Berger, JM Bian, X Cai, GF Cao, XX Cao, JF Chang, C Chen, et al. Design and construction of the besiii detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 614(3):345–399, 2010.

[2] Official website of tianhe-2. `http://www.nscc-gz.cn/`.

[3] Jakob Blomer, Carlos Aguado-Sánchez, Predrag Buncic, and Artem Harutyunyan. Distributing lhc application software and conditions databases using the cernvm file system. In *Journal of Physics: Conference Series*, volume 331, page 042003. IOP Publishing, 2011.

[4] Official website of mysql. `https://www.mysql.com/products/enterprise/router.html`.

[5] Wei-Dong Li, Huai-Min Liu, Ziyan Deng, Kanglin He, Miao He, Xiaobin Ji, Linli Jiang, Haibo Li, Chunxiu Liu, Qiumei Ma, et al. The offline software for the besiii experiment. In *Proceeding of CHEP*, volume 27, 2006.

[6] GAUDI& Barrand, I Belyaev, P Binko, M Cattaneo, R Chytracek, G Corti, M Frank, G Gracia, J Harvey, Eric Van Herwijnen, et al. Gaudi—a software architecture and framework for building hep data processing applications. *Computer Physics Communications*, 140(1-2):45–55, 2001.

[7] Christian Arnault. Cmt: A software configuration management tool. In *Prepared for International Conference on Computing in High-Energy Physics and Nuclear Physics (CHEP 2000), Padova, Italy*, pages 7–11, 2000.