21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Applications of supercomputer Tianhe-II in BESIII

Jingkun Chen,[1] Biying Hu,[2] Qiumei Ma,[3] Jian Tang,[2] Ye Yuan,[3] Yao Zhang,[3] Wei Zheng,[3] and Xiaomei Zhang[3]
[1]National Supercomputer Center in Guangzhou, China
[2]Sun Yat-sen University, China
[3]Institute of High Energy Physics, China

Introduction

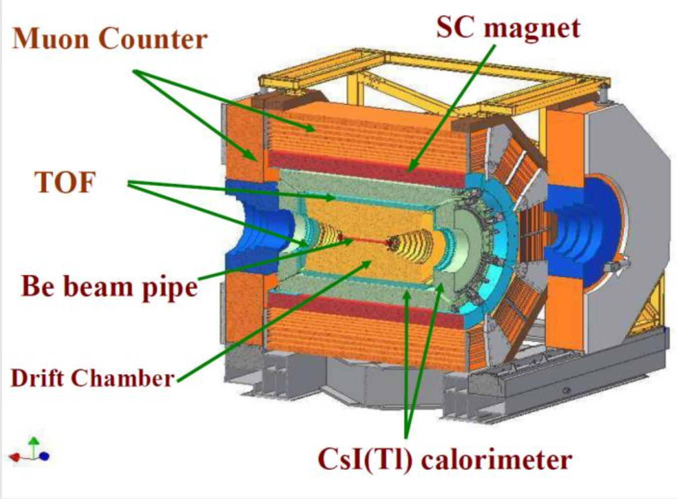
High energy physics experiments are pushing forward the precision measurements and searching for new physics beyond standard model. Taking the BESIII experiment as an illustration, we deploy the offline software BOSS into the top-tier supercomputer "Tianhe-II" with the help of Singularity. With very 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.

1. 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, 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 tau-charm physical energy zone in the world.

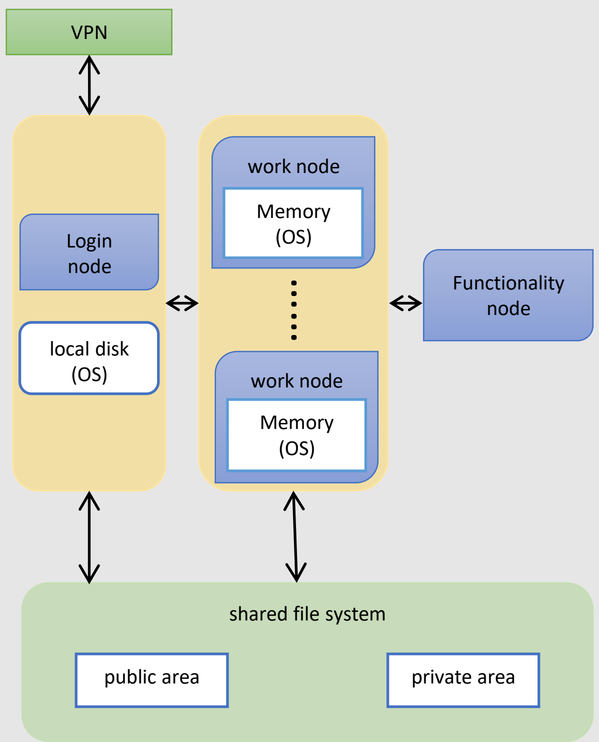
The rich physics program of the BESIII experiment includes:

- Tests of electroweak interactions with very high precision in both the quark and lepton sectors.
- High statistics studies of light hadron spectroscopy and decay properties.
- Studies of the production and decay properties of J/ψ 、ψ (2S) and ψ (3770) states.
- Studies of τ-physics.
- Studies of charm physics, including the decay properties of D and D_s and charmed baryons.



2. Supercomputer Tianhe-II

- Located in Sun Yat-sen University, National Supercomputing center in Guangzhou, China, http://en.nscc-gz.cn/
- 16,000 node, total 3,120,000 CPU core
- 30.65Pflops (world's fastest at 2013 ~ 2015)
- 88GB memory/node, CPU: 64G, MIC: 24G
- 12.4PB disk array
- Tested in work nodes with 0.4224Tflop, 64GB memory and 24-core CPUs.



3. Install CVMFS on Tianhe-II

The CernVM-File System (CernVM-FS) provides a scalable, reliable and low-maintenance software distribution service.

Classic approach:

1. **General install**:
   - ➢ Install a binary CVMFS client package in path "/CVMFS“.
   - ➢ Deploy a dedicate work node with connection from "/CVMFS" to the file system.

   Defects:
   - ➢ The dedicated node prevents us from using abundant resources of Tianhe-II.
   - ➢ The limitation of mount lock restricted the numbers of work node which was mounting the code-base.
2. **Using Cvmfsexec**:
   - ➢ Set the mount's destination on "/tmp" to get around the mount lock.
   - ➢ Requires the CVMFS libraries and the libraries at "/CVMFS/lib"

   Defects:
   - ➢ Requires a unique OS mirror and limits the computing resources.
   - ➢ CVMFSexec reports an unresolved warning: “unshare: unshare failed: Invalid argument”.
3. **Parrot-mount**:
   - ➢ Install CVMFS through virtual machine Parrot to avoid the path “/CVMFS”.

   Defects:
   - ➢ Virtual machine cannot be installed on Tianhe-II.

Our approach: **compile CVMFS from source code**

- Fundamentally solve the problem of “/CVMFS” path.
- Compile the new version fuse, uuid and libcap to deceive CVMFS version check.
- Load fuse module in worknode OS.
- Mount on path “/tmp/cvmfs” to avoid mounting lock.
- Deploy a Squid service in a functionality node.

4. Network topology

In addition of CVMFS, some supports are needed by running BOSS.

- A **fast network** for data transmission.
- A **MySQL slave server** in Tianhe-II.

Network status:

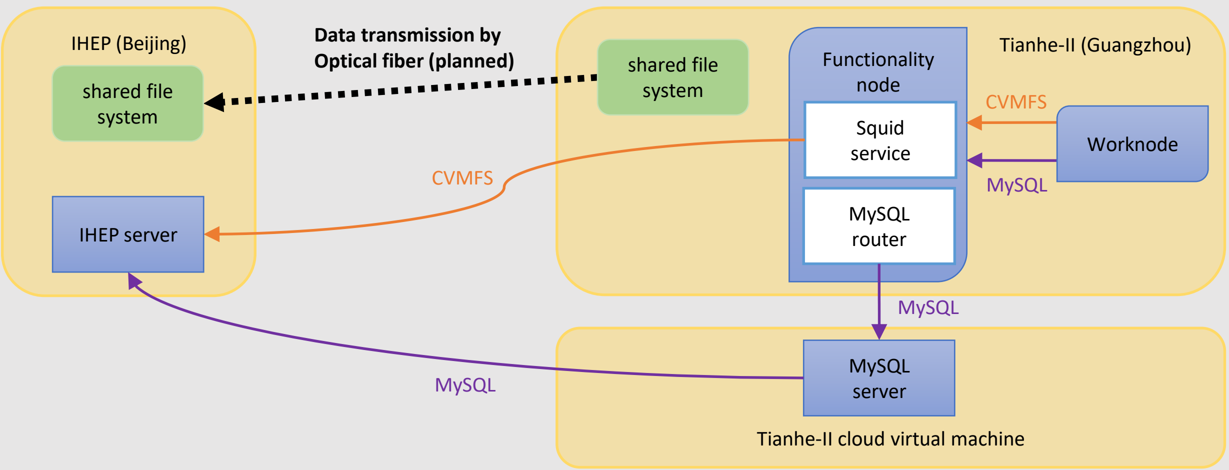
- All users (over 1400 research groups now) share 3Gb bandwidth.
- The speed test for IPV4 only reach to 20MB/s

Network solutions:

- Change to **IPV6**: The operators of IHEP and Tianhe-II are different.
- Set up an **Internet Leased Line**(ILL): cost too much.
- Set up a bare **optical fiber**: Security and stability is less than ILL but much cheaper.

MySQL server:

- Deploy a MySQL server in cloud virtual machine.
- Install a MySQL router in a functionality node as a bridge.

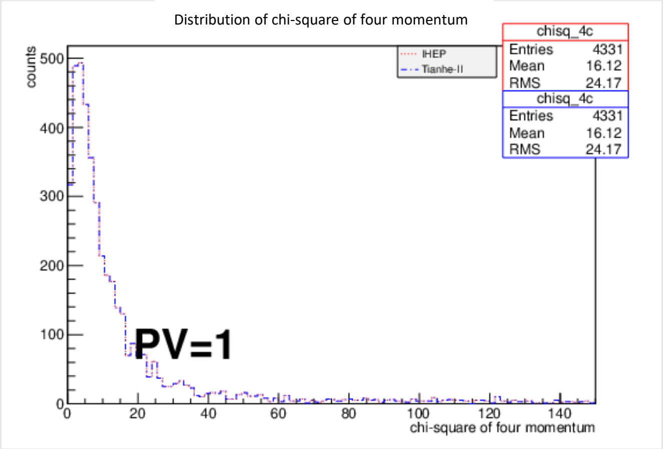


5. Validation of BOSS

BESIII **O**ffline **S**oftware **S**ystem (BOSS), 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 on the Scientific Linux CERN (**SLC**) operating system with **CMT** 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, we ran the same simulation script at Tianhe-II and IHEP and compared the results.

The results shows that data from two platforms are completely consistent.

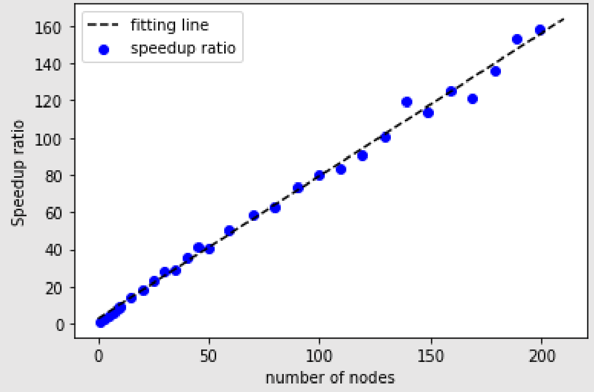


6. A Large-scale performance test

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,

- ➢ Develop a MPI python submitting script to pack the HTC jobs and control the I/O.
- ➢ Create initial files in memory instead of reading in file system.
- ➢ Save the output in memory first and move to file system in order.

The fitting line of the speedup ratio



Conclusion: 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.

中国科学院高能物理研究所
Institute of High Energy Physics Chinese Academy of Sciences
中山大學
SUN YAT-SEN UNIVERSITY
国家超级计算广州中心
NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU