Applications of Supercomputer Tianhe-II in BESIII

Jingkun Chen[1], **Biying Hu[1]**, Xiaobin Ji[2], Qiumei Ma[2], Jian Tang[1], Ye Yuan[2] , Xiaomei Zhang[2], Yao Zhang[2], Wenwen Zhao[1], Wei Zheng[2]

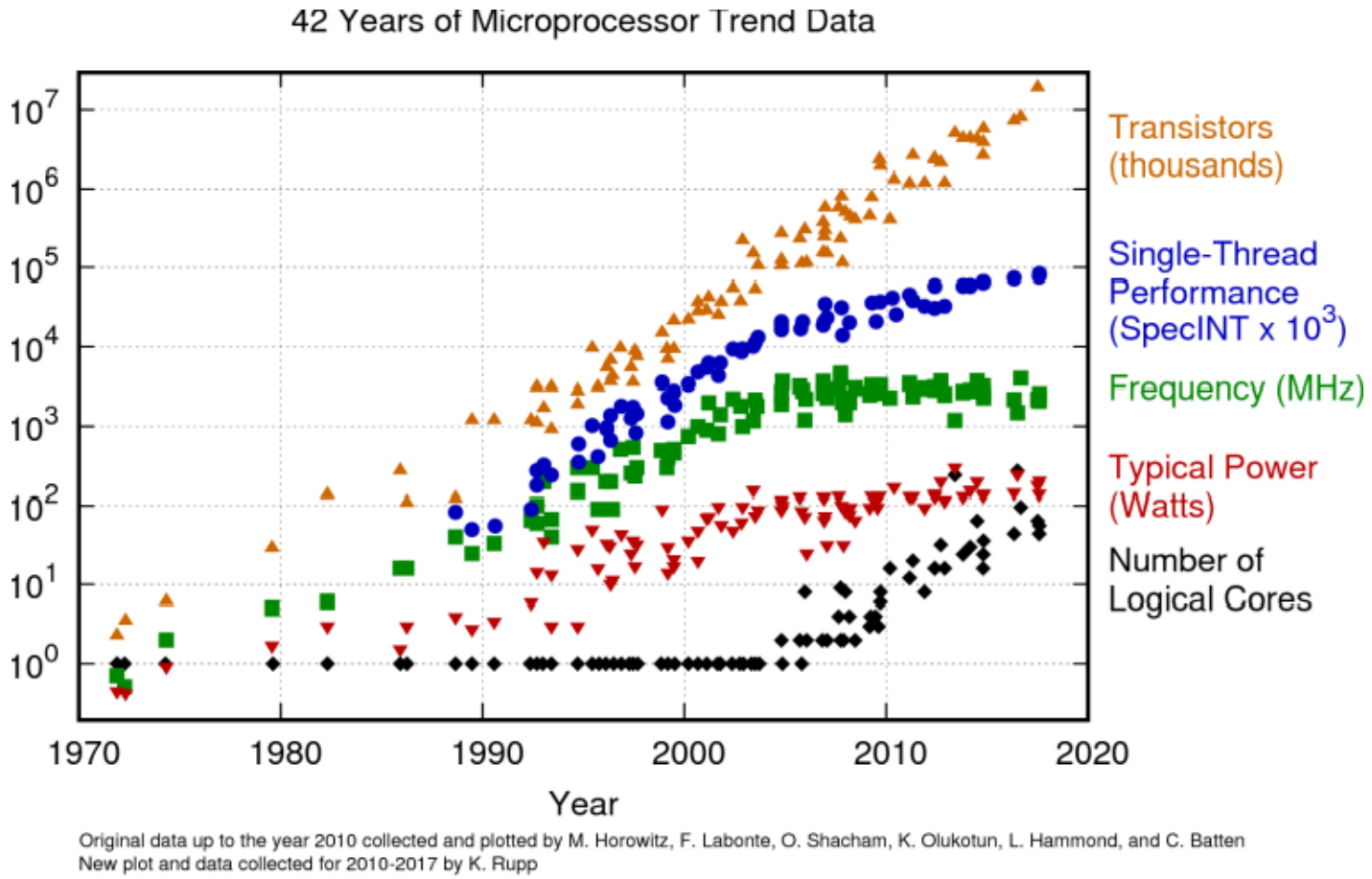
1. Sun Yat-Sen University
2. Institute of High Energy Physics, China

CHEP at Norfolk, Virginia, USA, May 11th, 2023

Outline

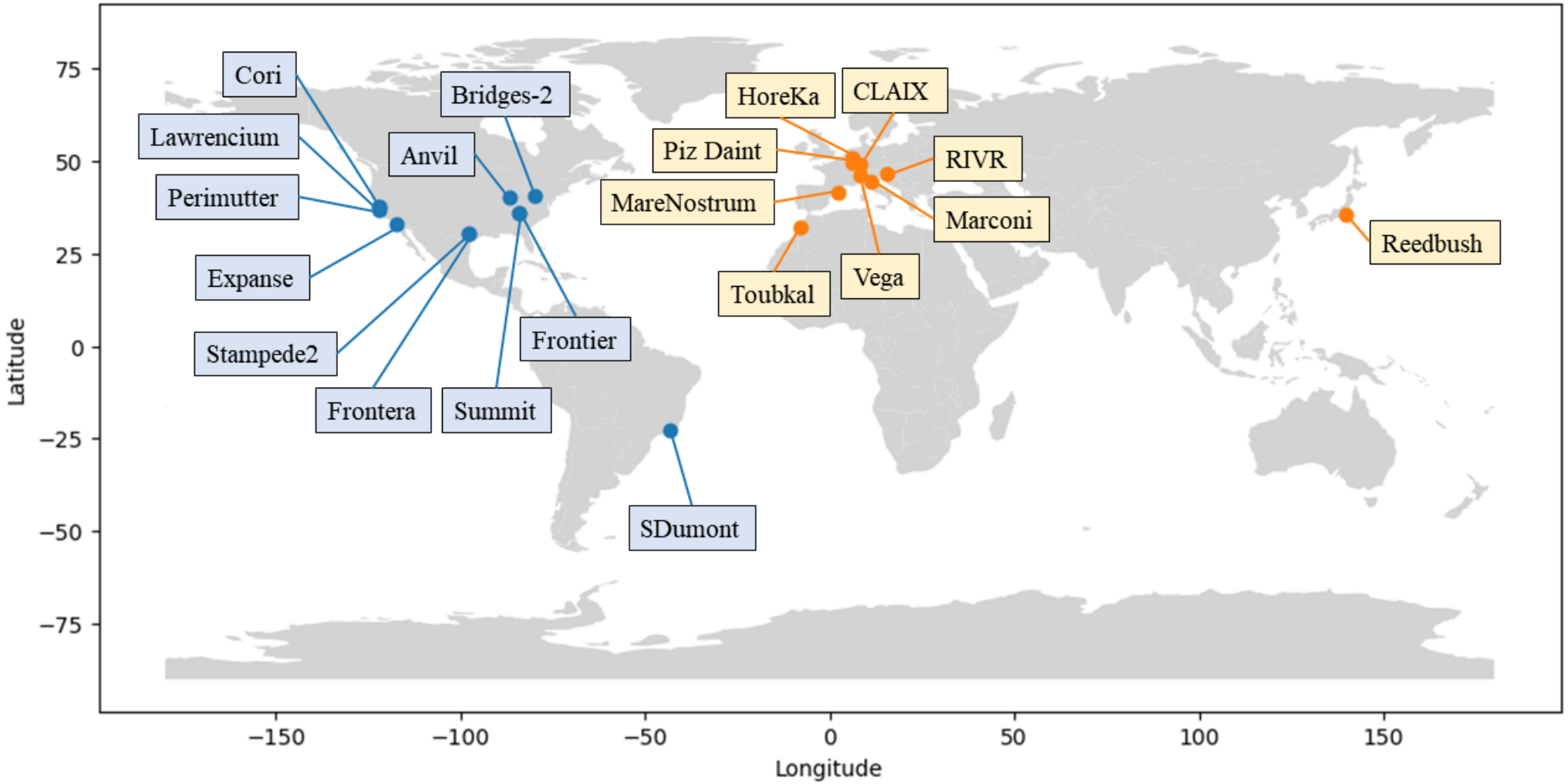
1 Introduction

Worldwide trend

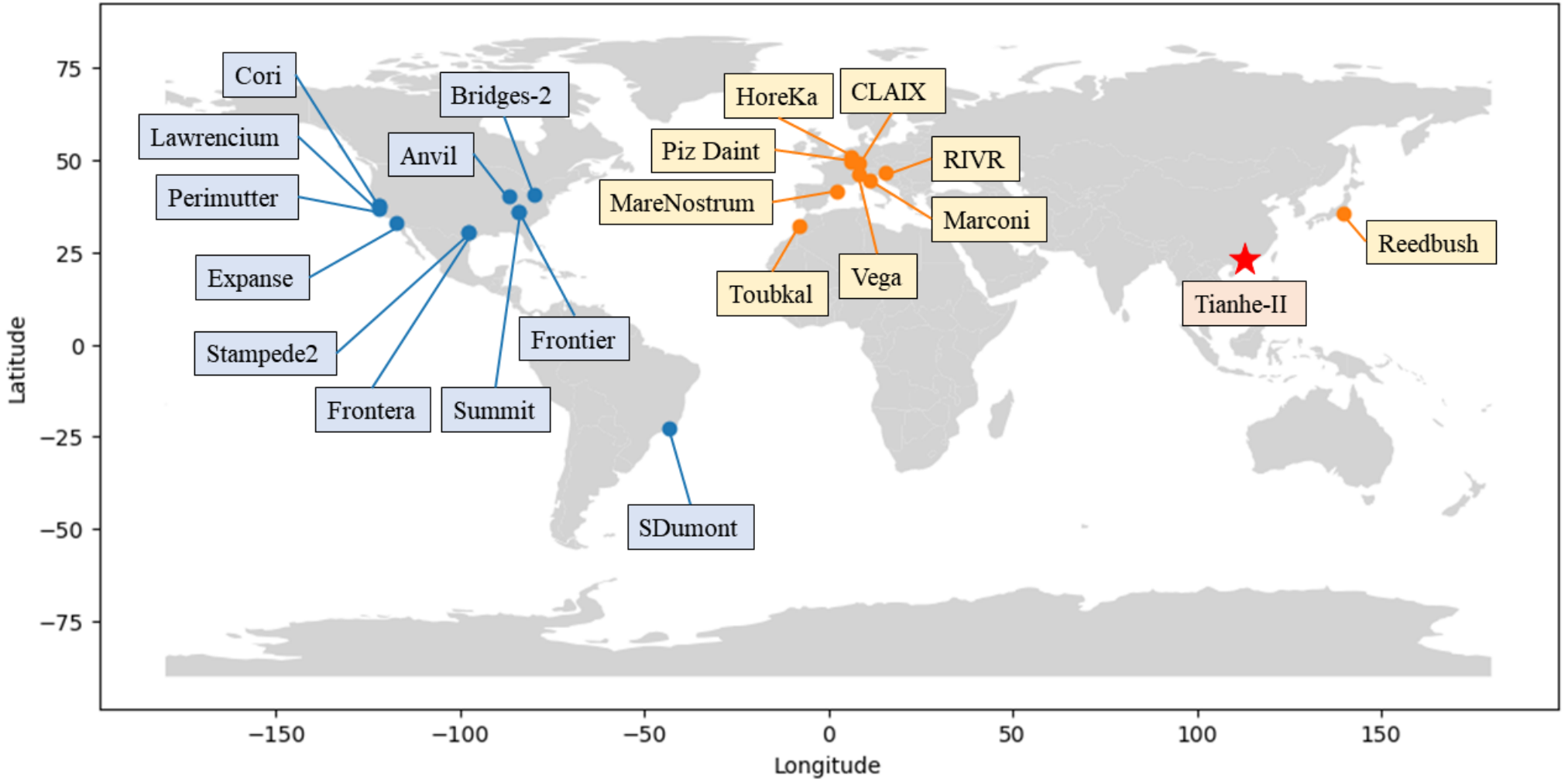


Strategic worldwide trend: massive investments into supercomputers.

Worldwide HPC used for HEP experiment

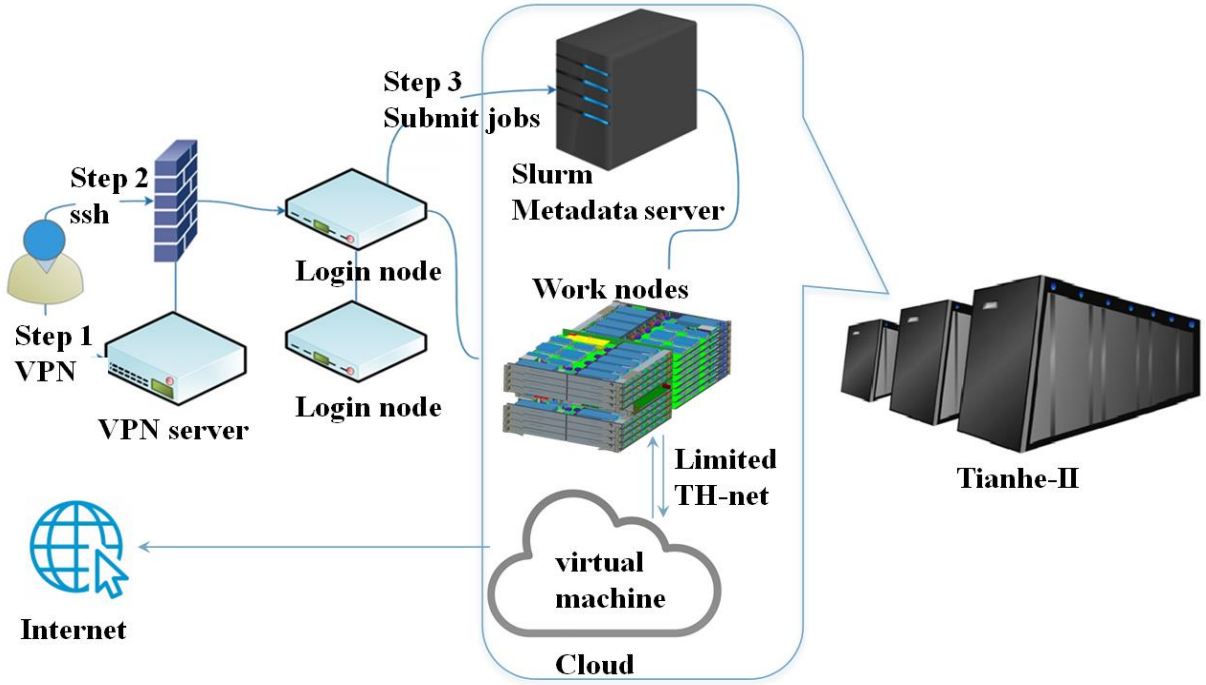


Worldwide HPC used for HEP experiment



Tianhe-II

- Located in Sun Yat-sen University, National Supercomputing center in Guangzhou, China
- Total **3,120,000 CPU cores**
- 30.65 Pflops (**world's fastest** at 2013 ~ 2015)
- Adapts serially SLURM system and metadata server
- 24 CPU cores per node and **15 PB** shared file system Lustre
- A new data center to support applications.



HTC & HPC

HTC (High-throughput Computing)

- Long timescale
- Data intensive
- Designed for HEP need

HPC (High-performance Computing)

- Short timescale
- Computing intensive
- Not designed for HEP

Challenges:

- Strict security policy.
- Limit of network.
- I/O crash caused by large input and output data.

All supercomputers will encounter these challenges,

but no general solution for HPC to run HTC jobs.

Need to achieve

- BESIII Offline Software System (BOSS) deployment
- Remote submission
- Large-scale performance test

2 BOSS deployment

Virtualization container

- **No root authority** on Tianhe-II
- Singularity is designed to run the container without a root authority.

- Two types of container:
  - Fat container with all the required software and relevant database.
  - Light container only includes the BOSS environment that applies to all BOSS versions.

Deployment Solution 1 (done)

- Solution 1
  - Use container virtualization technology and a **fat container**

- Fat container disadvantages
  - Follow-up update of BOSS is very troublesome.
  - Requires **large storage space** at Tianhe-II
  - Requires **great human effort** to maintain many containers

Deployment Solution 2

- Solution 2
  - Use Cern-VM File System (**CVMFS**) and a **light container**
- Two problems:
  - 1. Install the **cvmfs client** on Tianhe-2;
  - 2. Realize real-time access to the **database**

Install CVMFS

Without root authority, we have failed in many approaches.

- General installation ✗
- Using a pre-installed executable CVMFS Cvmfsexec ✗
- Parrot-mount CVMFS ✗

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

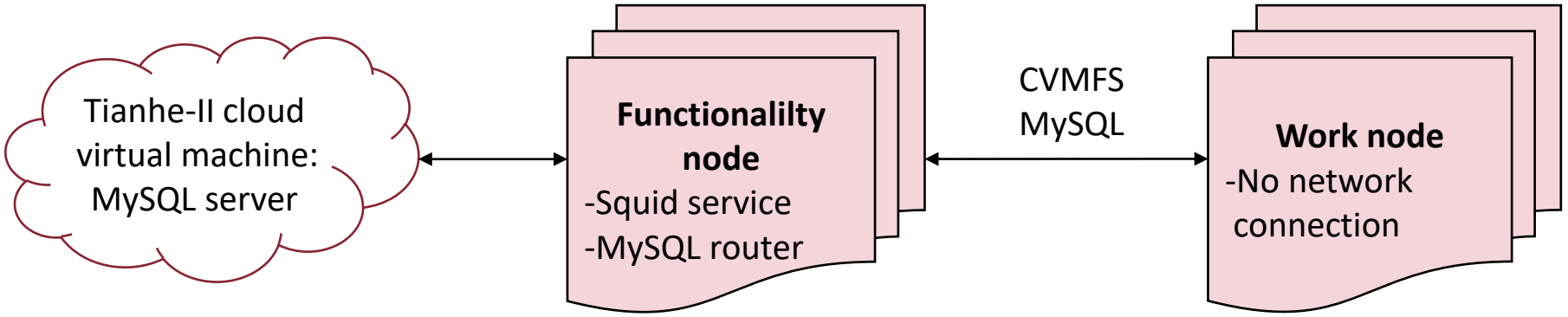
Database connection

Squid

- Squid is a caching and forwarding HTTP web proxy.
- Install Squid on functionality nodes.
- Allow CVMFS connect to the IHEP code database.

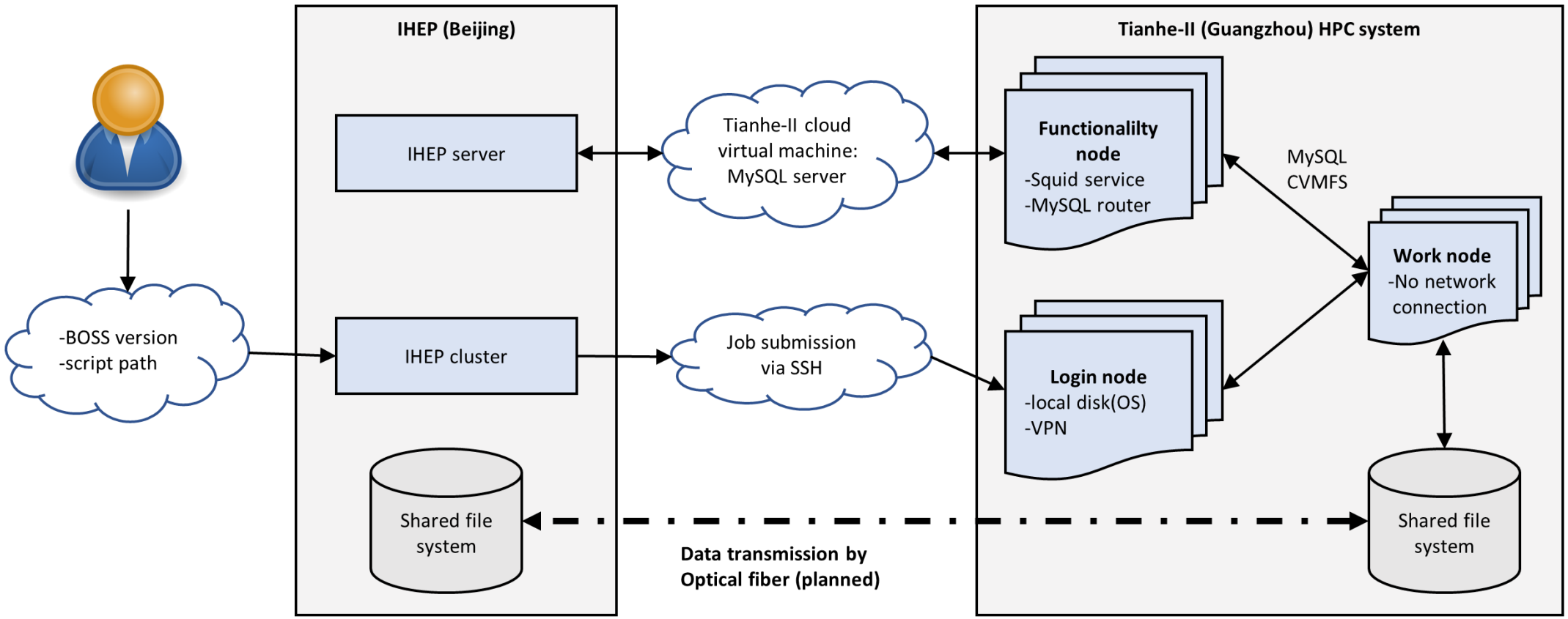
MySQL

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



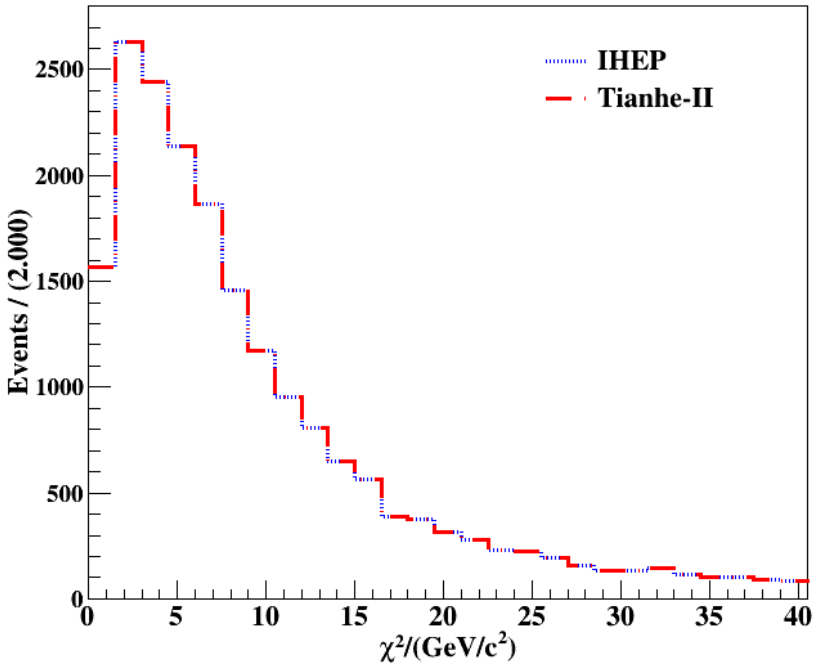
3 Submission flow

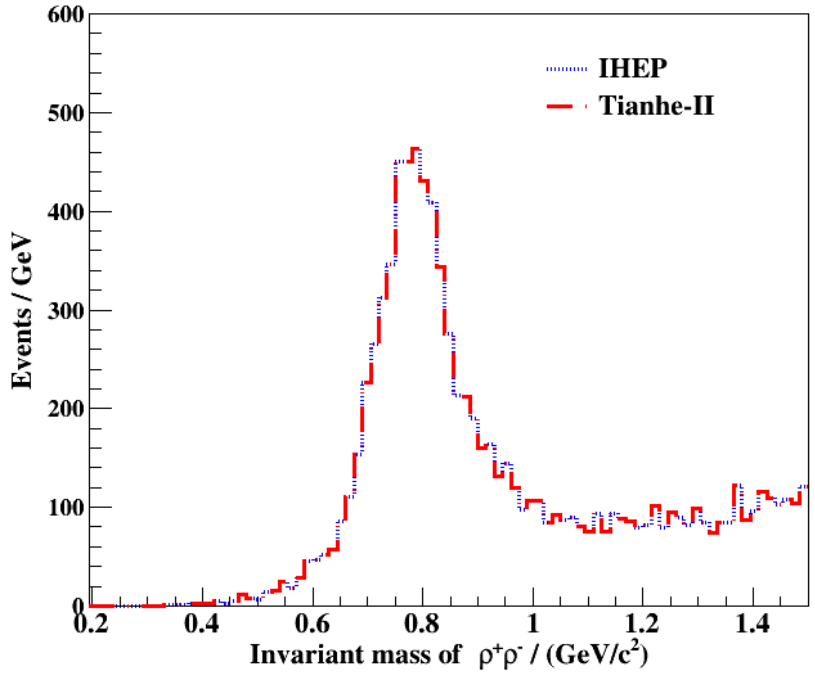
Workflow


IHEP (Beijing)
Tianhe-II (Guangzhou) HPC system
IHEP server
Tianhe-II cloud virtual machine: MySQL server
Functionalilty node
-Squid service
-MySQL router
MySQL
CVMFS
Work node
-No network connection
-BOSS version
-script path
IHEP cluster
Job submission via SSH
Login node
-local disk(OS)
-VPN
Shared file system
Shared file system
Data transmission by Optical fiber (planned)

Validation

$e^+ e^- \rightarrow J/\psi \rightarrow \rho\pi$





Invariant mass spectrums of ρ meson and chi-square values of the four momentum from IHEP and Tianhe-II simulation results are completely the same.

What can we do on Tianhe-II?

Network status:

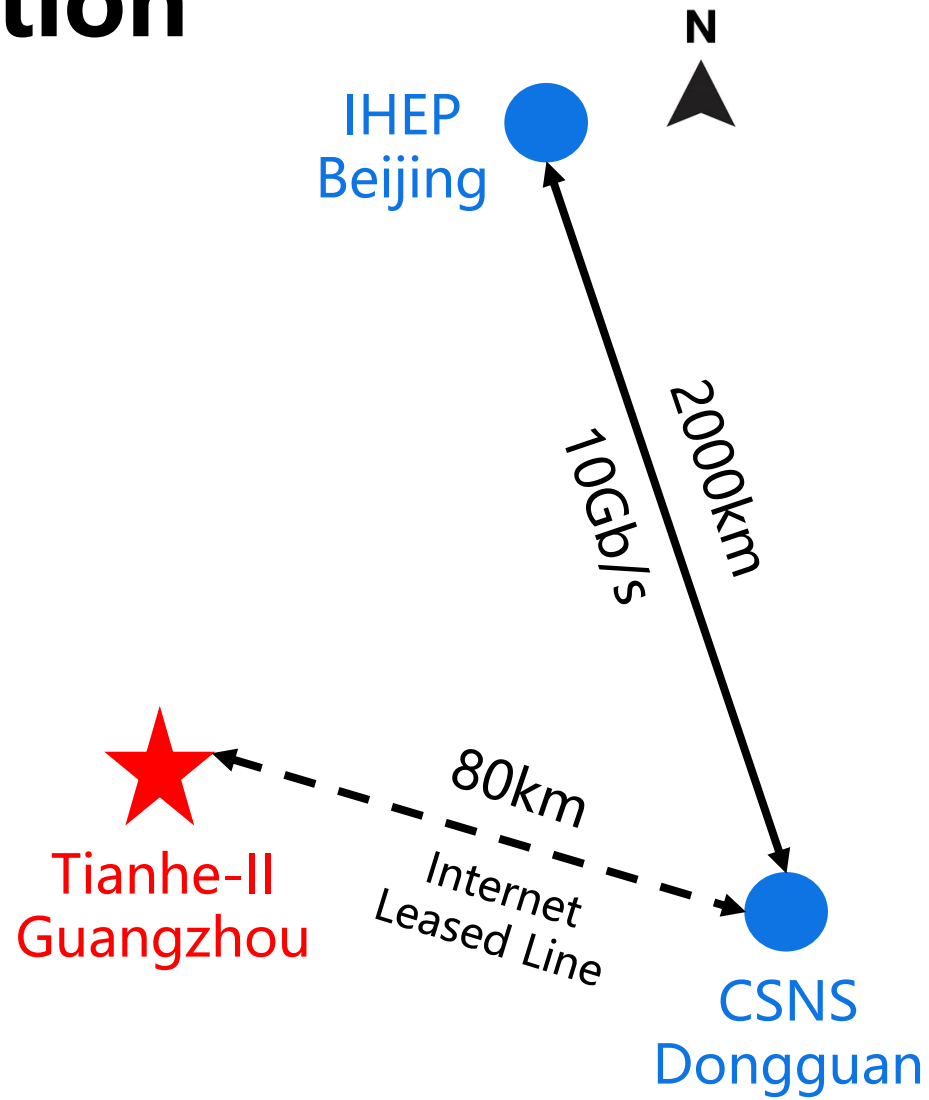
- The speed test for IPV4 only reach to **20MB/s**.
- Adequate for software deployment and database update, but **insufficient** for mass data transmission.
- The **operators** of IHEP and Tianhe-II are **different**, and IPV6 has only 50Mb bandwidth.

How to use:

- Base on **simulation**, which requires very few input files.
- **Reconstruct** and **analyze** the mass simulation data on Tianhe-II instead of sending back IHEP.

Network Solution

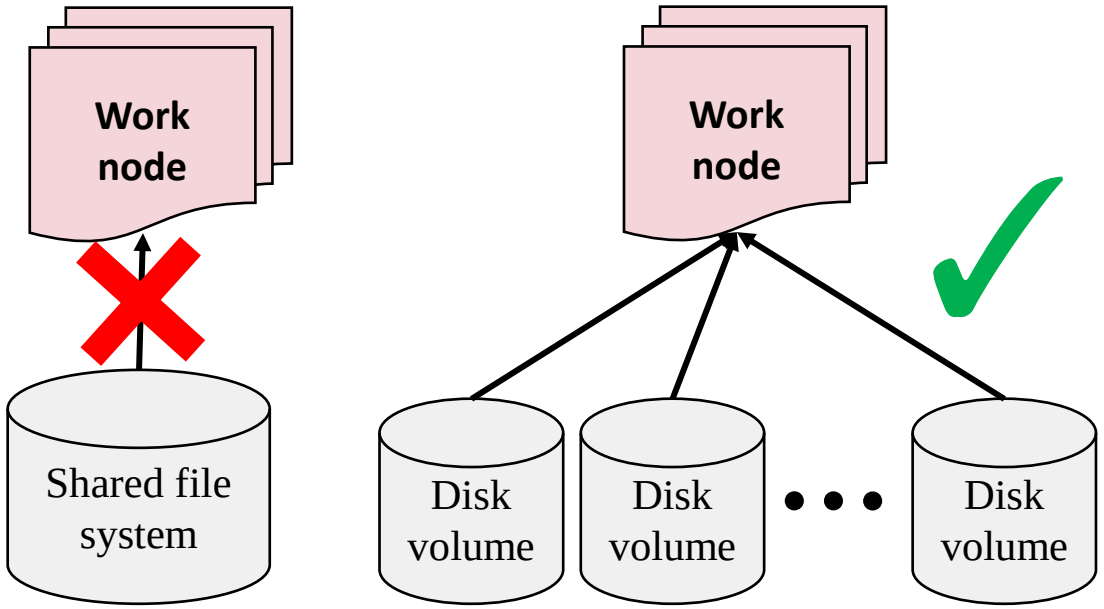
The **China Spallation Neutron Source (CSNS)** is the pulsed neutron source facility located at **Dongguan City**, which is next to **Guangzhou** City. To meet the need for data services, the National HEP Science Data Center (NHEPSDC) in **Beijing** set up its Branch Center at the CSNS. The Branch Center has more than **10 PB** of storage space, an international network link of **10,000 megabits per second**, as well as a complete information support system.

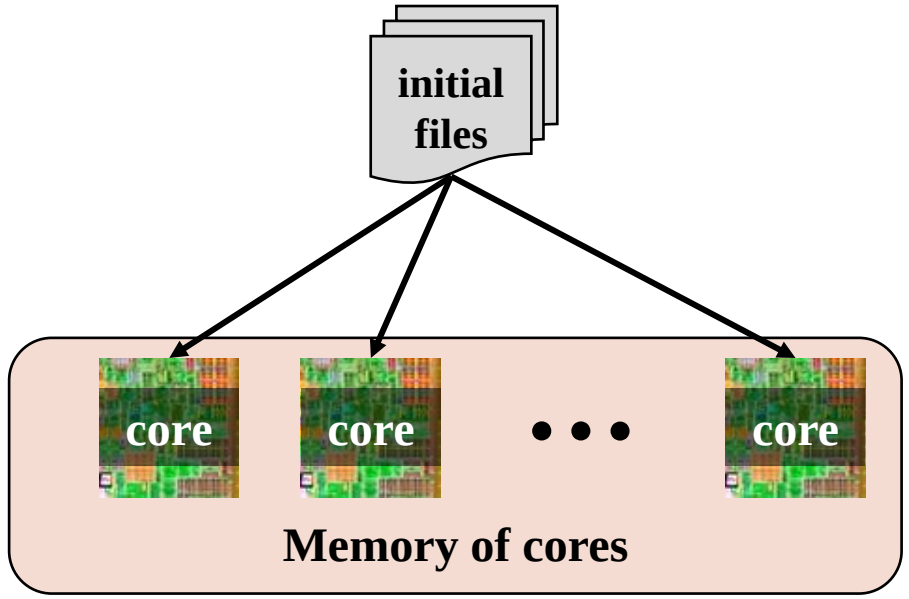


4 Large-scale performance

Input issue

- Input too many files from the share file system may cause IO congestion.
  - ➢ Storage mirror files in **each disk volume**
  - ➢ Create **initial files in memory** instead of reading in file system.



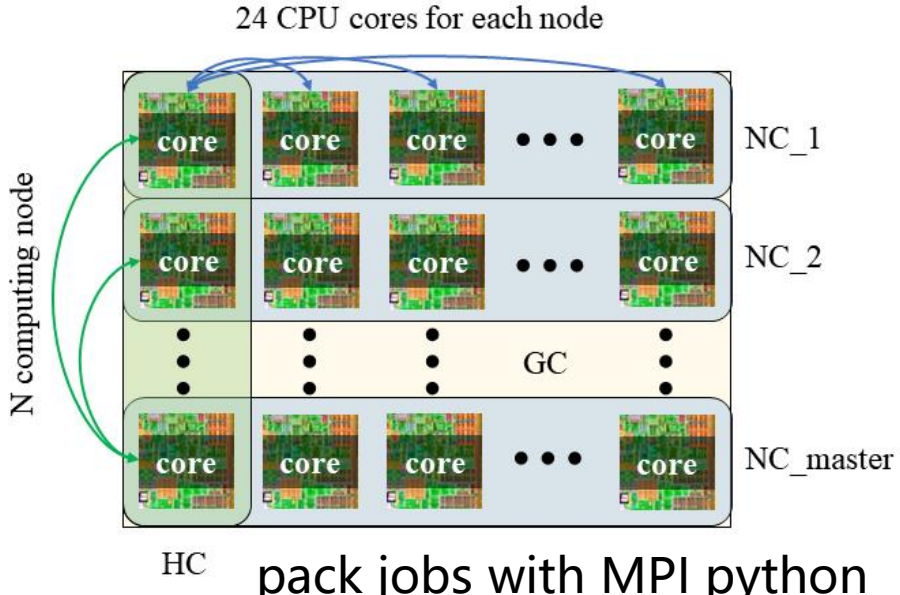


Output issue

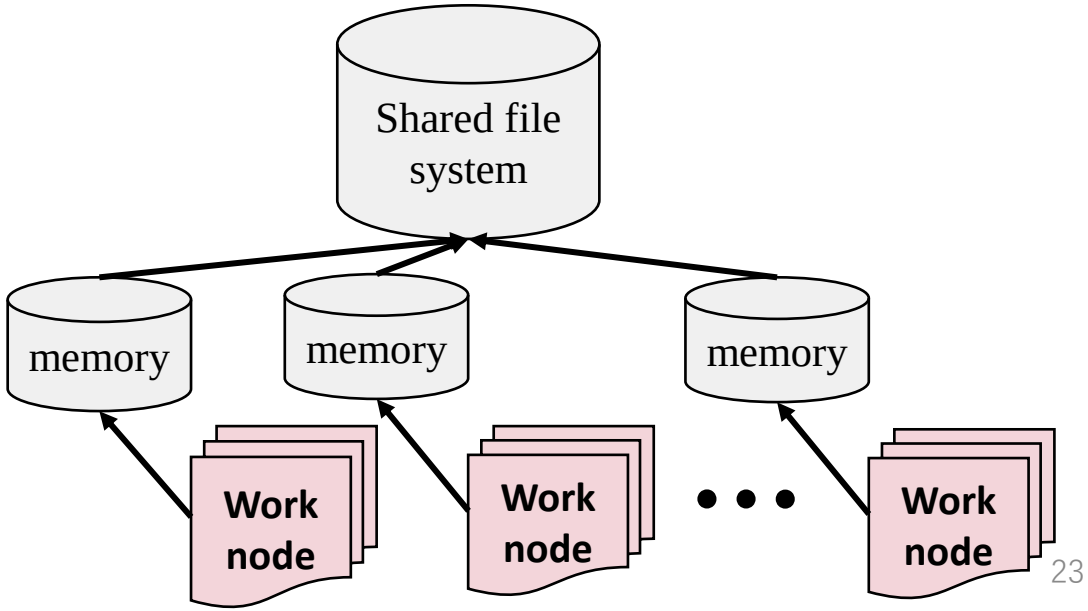
- The Tianhe-II SLURM system could become **crash** if numerous HTC jobs are submitted in a short time.

➢ Develop a **MPI python submitting script** to pack the HTC jobs and control the I/O.

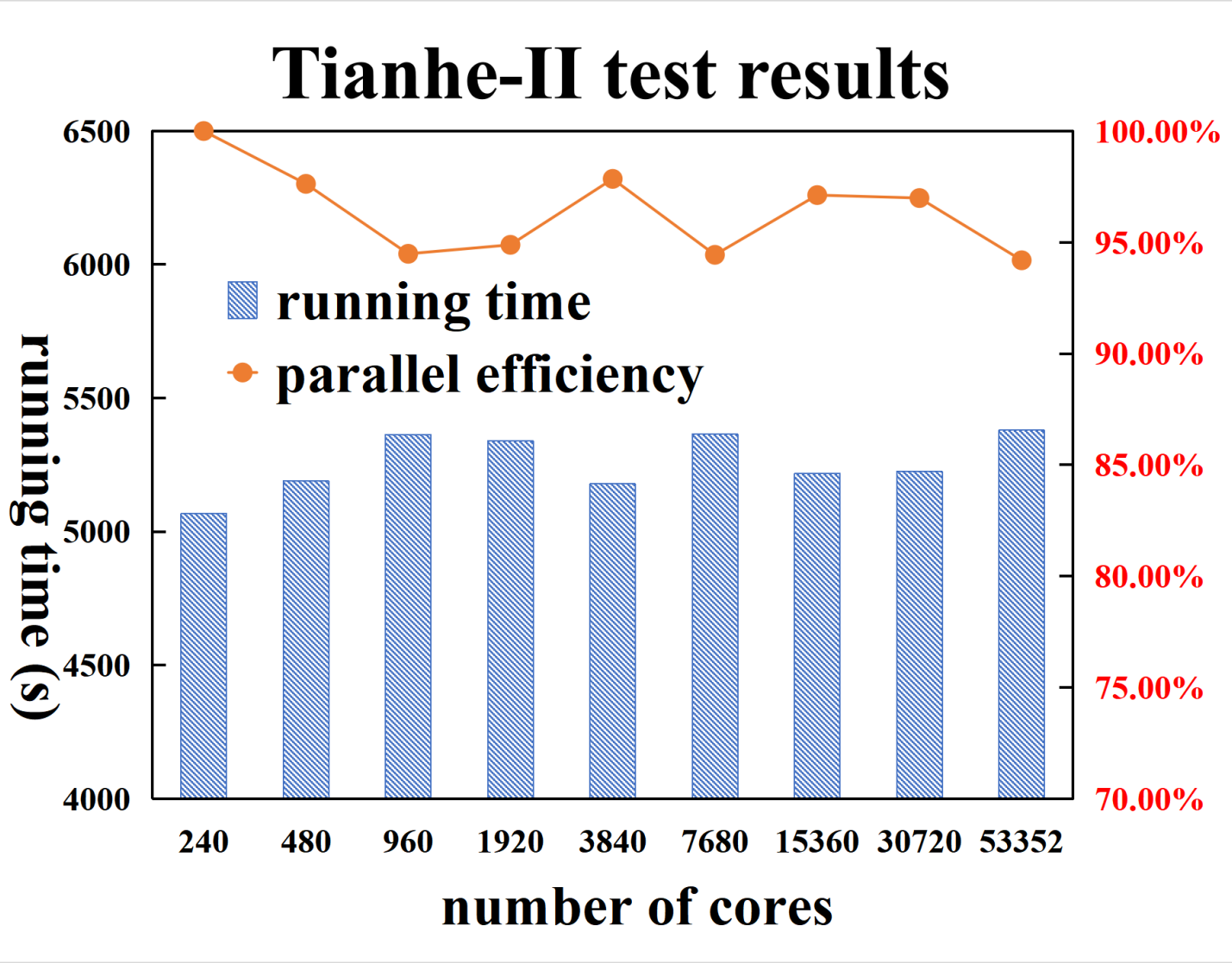
➢ Save the **output in memory** first and move to file system **in order**.



pack jobs with MPI python



Large-scale performance result



Parallel efficiency is used to measure the **performance of parallel systems**.

Parallel efficiency define as:

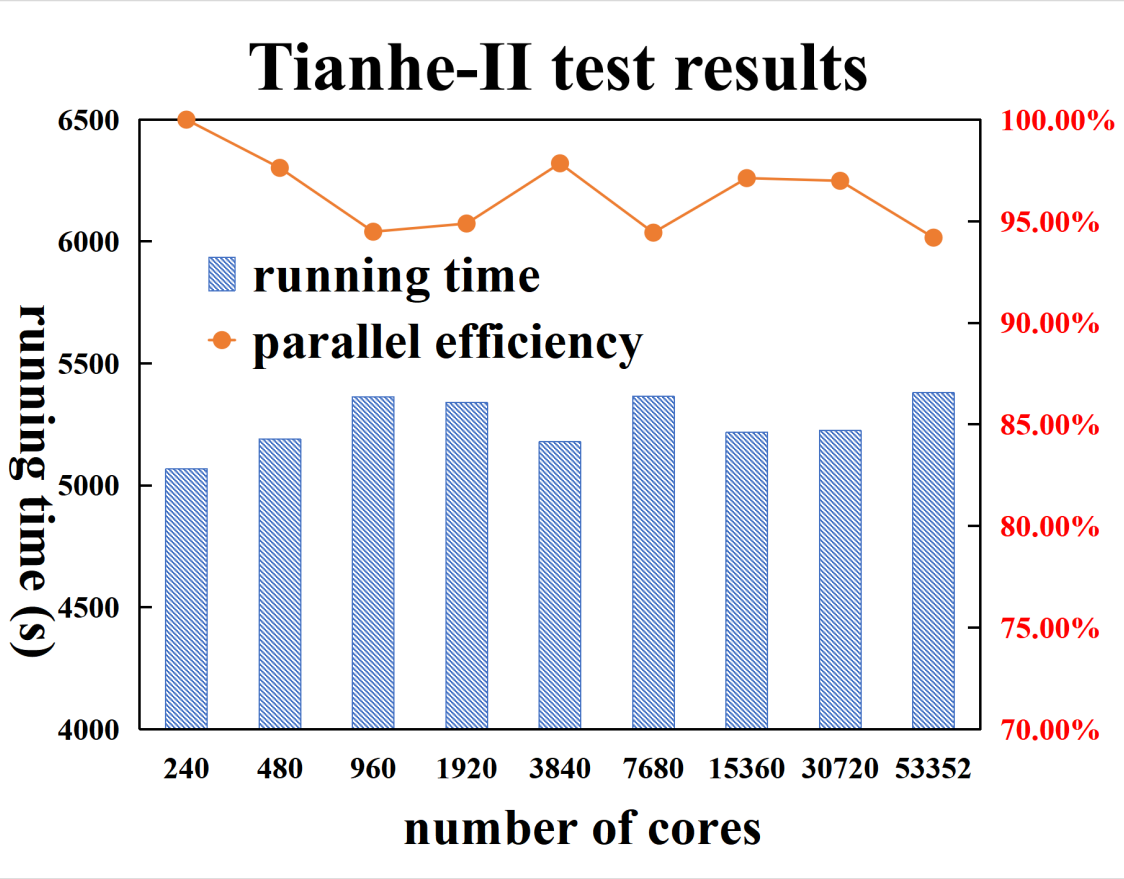
$$E(P,N) = \frac{t_1}{t_P}$$

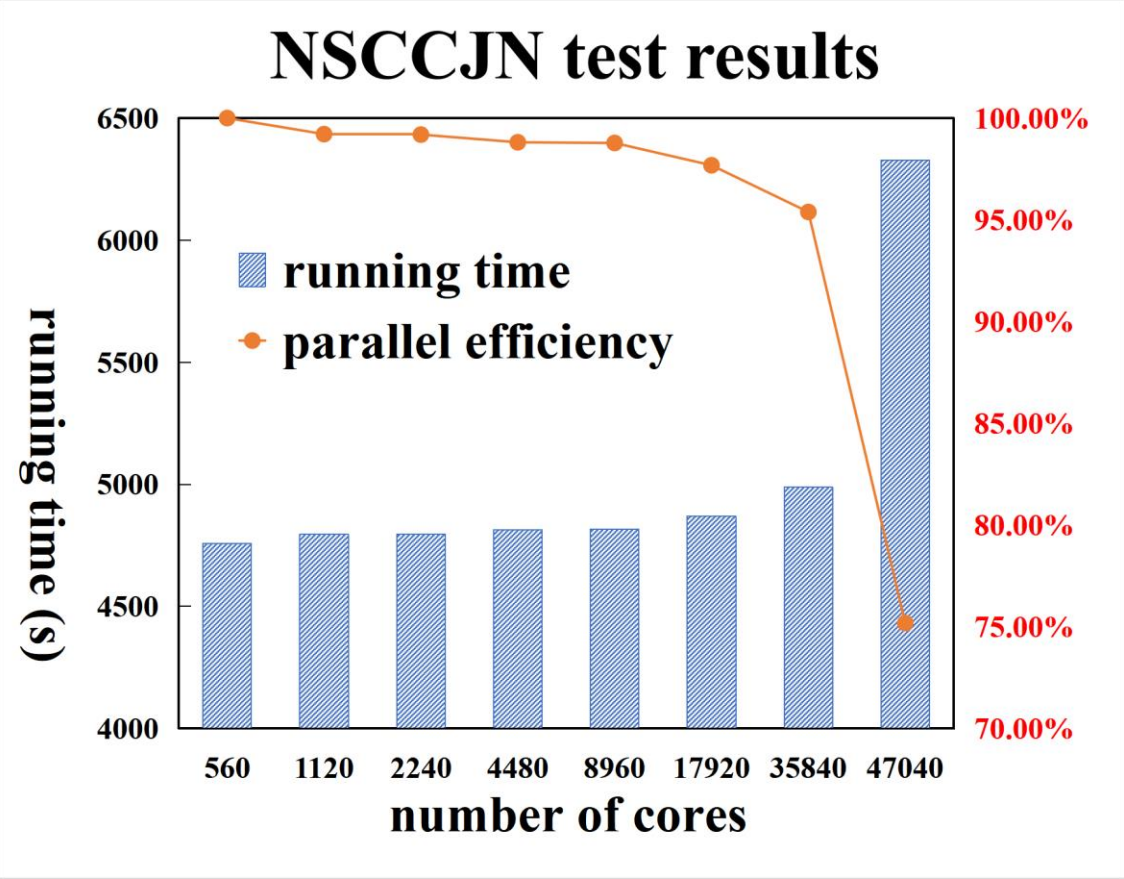
t_1 is the running time on single nodes,

t_P is the running time on P nodes.

The ideal parallel efficiency is 100%, and **80%** is considered as good performance.

Large-scale performance result





We ran BOSS with a light container on Tianhe-II and a fat container on NSCCJN.

The parallel efficiency of NSCCJN drops to 75% due to the **IO crash**.

Summary & Prospects

- Summary
  - Real-time updates of BOSS on Tianhe-II is realized.
  - Remote job submission to Tianhe-II from IHEP farm is scheduled.
  - Large-scale performance running is done.
- Prospects
  - Look forward to cooperating with the physics groups.
  - Improve the data transfer speed, for example, set up a bare optical fiber between Tianhe-II and CSNS.
  - More sites and resources to join in the future.

Ref: JINST18(2023), T03003

Thank you!