



# BESIII软件在天河2号的部署和测试应用

陈璟锟<sup>1</sup>, 胡碧莹<sup>1</sup>, 季晓斌<sup>2</sup>, 马秋梅<sup>2</sup>, 唐健<sup>1</sup>, 袁野<sup>2</sup>, 张晓梅<sup>2</sup>,  
张瑶<sup>2</sup>, 赵问问<sup>1</sup>, 郑伟<sup>2</sup>

1. 中山大学
2. 高能所

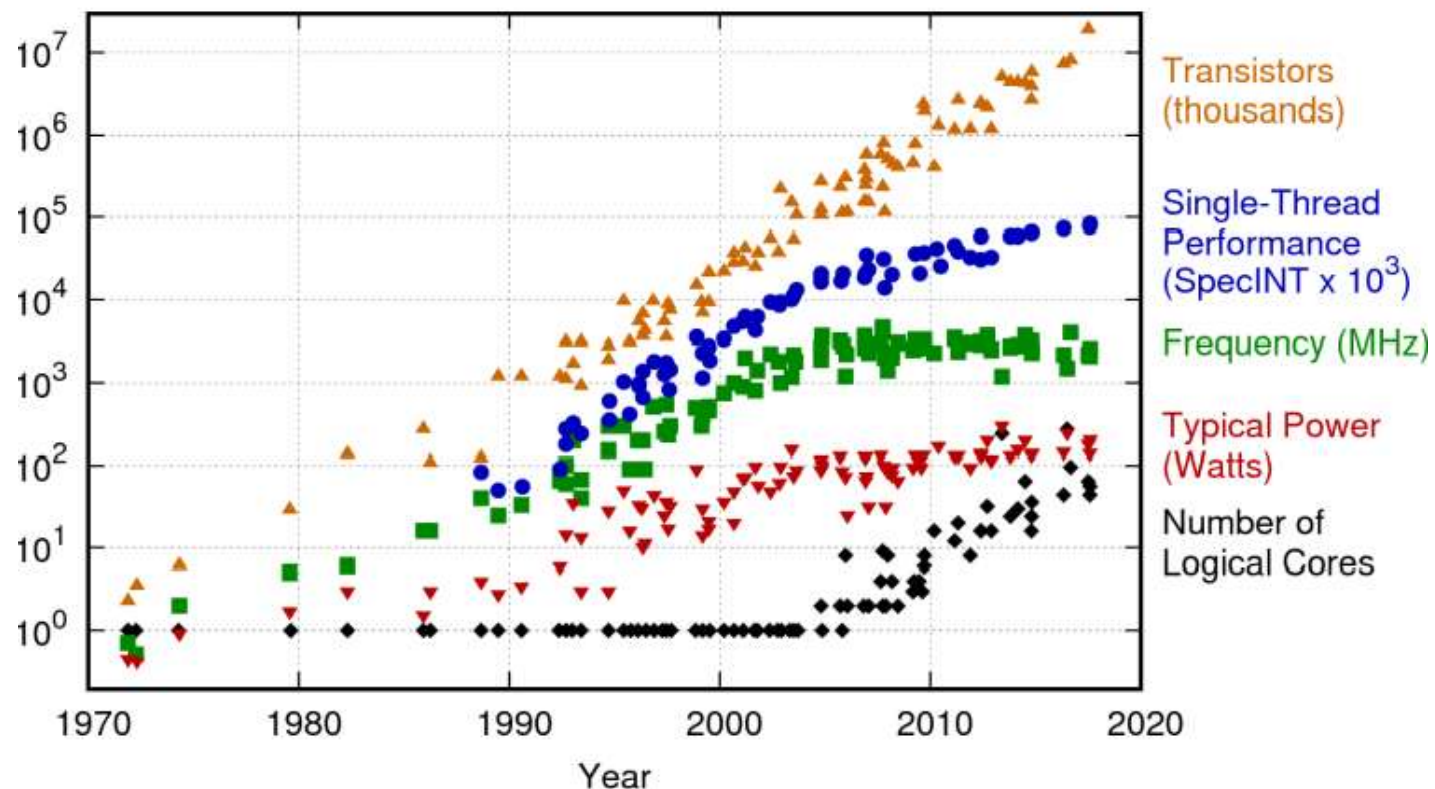
OUTLINE

目录

- 1 介绍
- 2 在天河二号上适配BESIII
- 3 远程提交 workflow
- 4 大规模性能测试
- 5 总结

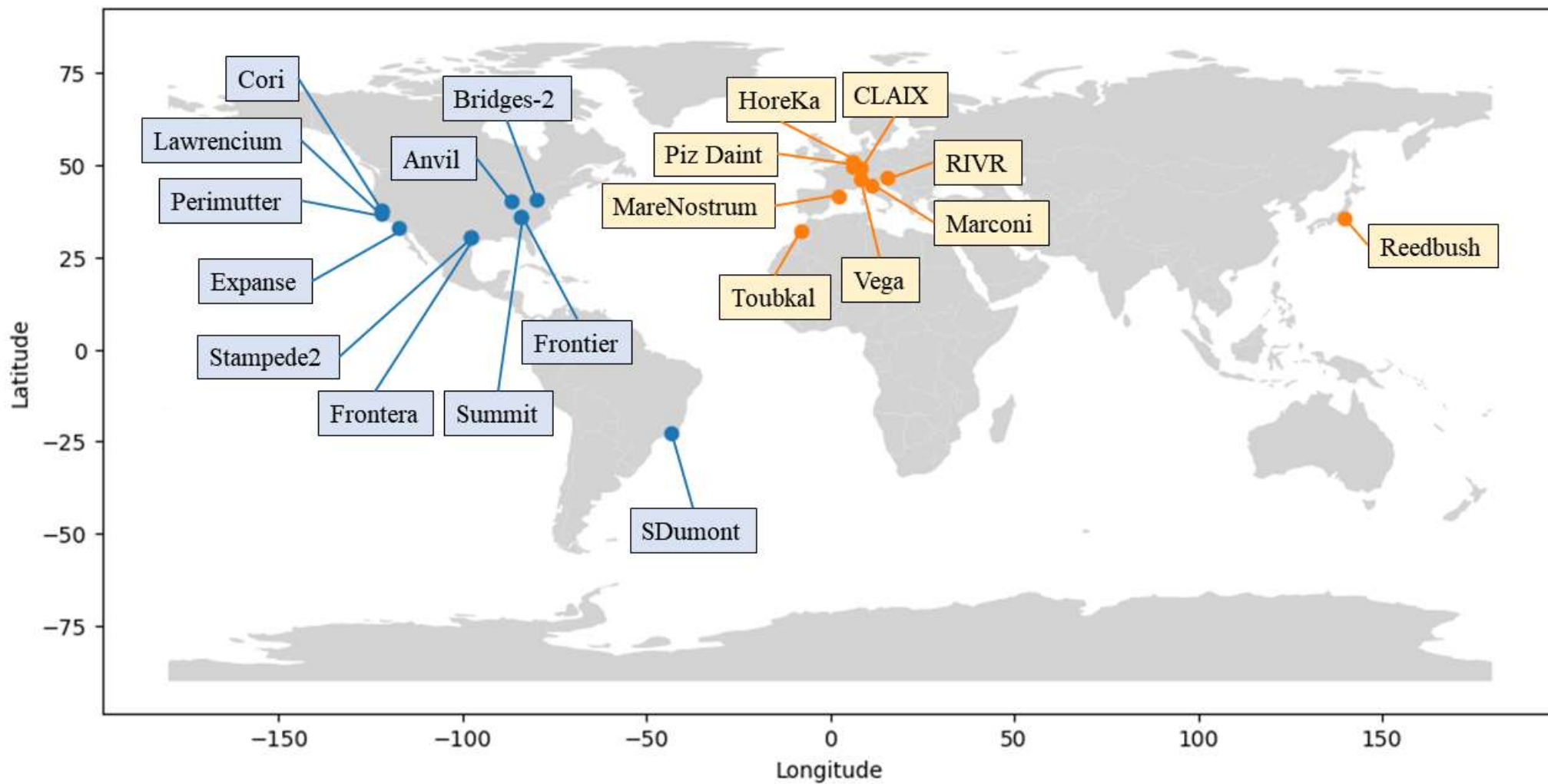
# 数据产生和逻辑

42 Years of Microprocessor Trend Data

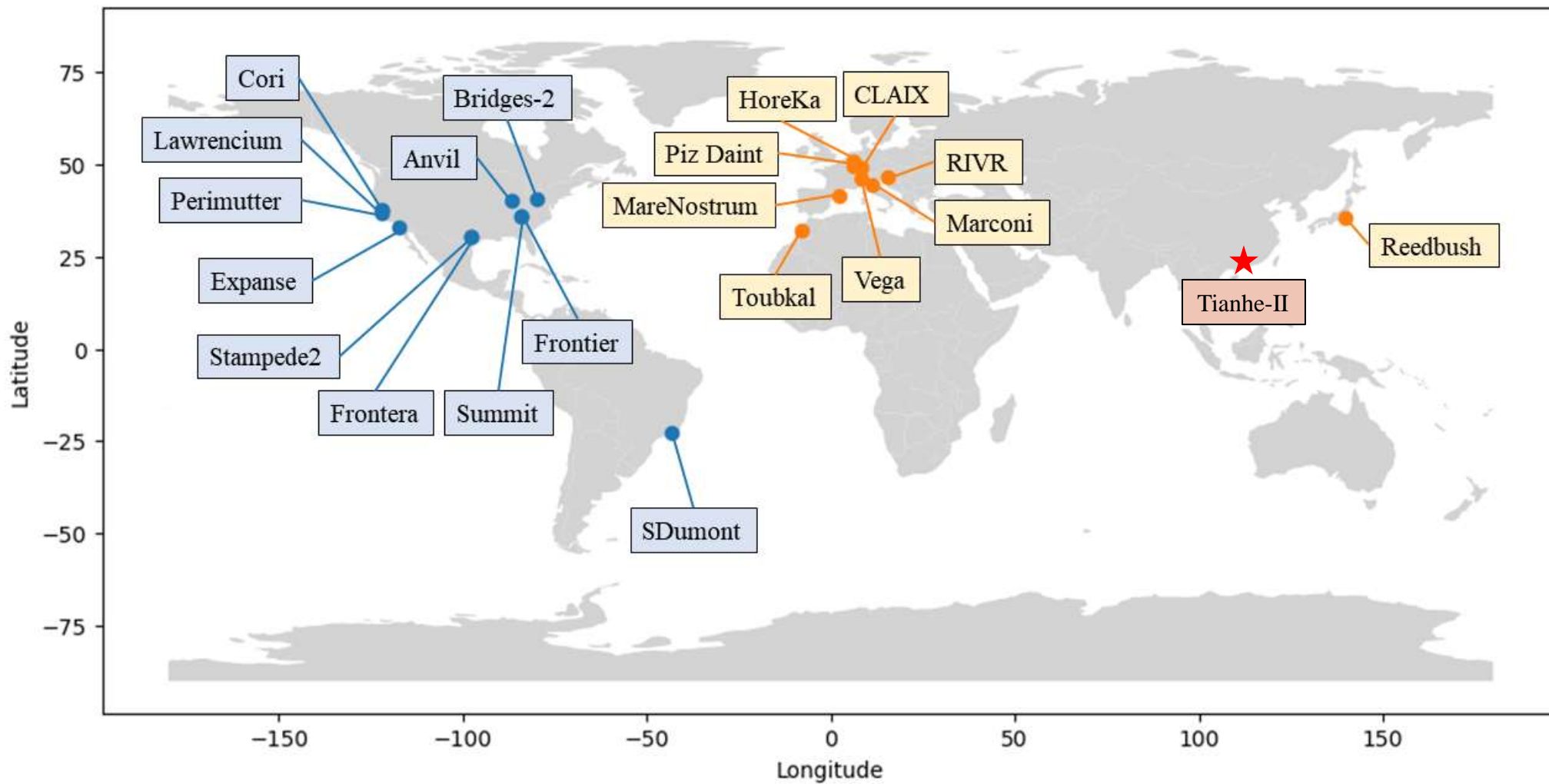


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten  
New plot and data collected for 2010-2017 by K. Rupp

# 承担HPE数据处理任务的超算中心

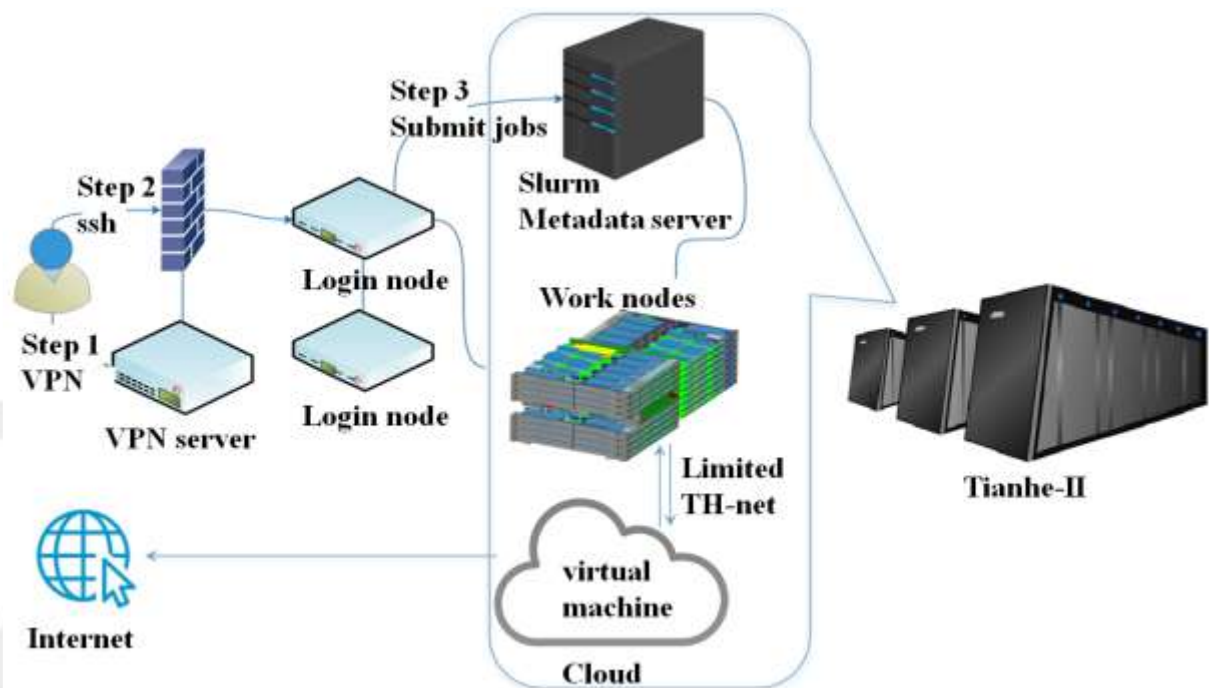


# 承担HPE数据处理任务的超算中心



# 天河2号

- 坐落于中山大学，国家超级计算广州中心
- 总计312万CPU核
- 持续性能30.65 Pflops ( 2013 ~ 2015 超算6连冠)
- 共享式15PB的Slurm文件系统搭配若干元数据服务器
- 每节点24CPU核，64GB内存，没有本地缓存
- 新机器即将登场



# 高通量计算机(HTC) & 高性能计算机 (HPC)

## HTC

- 长时间
- 数据密集型
- 更适合离散数据处理任务, 例如HEP

### 目标:

- 适配BESIII软件
- 异地远程提交技术
- 大规模性能测试

## HPC

- 短时间
- 计算密集型
- 对高通量作业支持不足

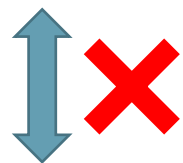
### 挑战:

- 没有root权限
- 受限的对外网络连接行为
- 瞬时的海量读写容易引发Slurm文件系统的崩溃

**所有的HPC在运行HEP作业时都要面对上述的问题, 又因为管理细则的差异导致, 导致不存在一个通解。**

# 虚拟镜像技术

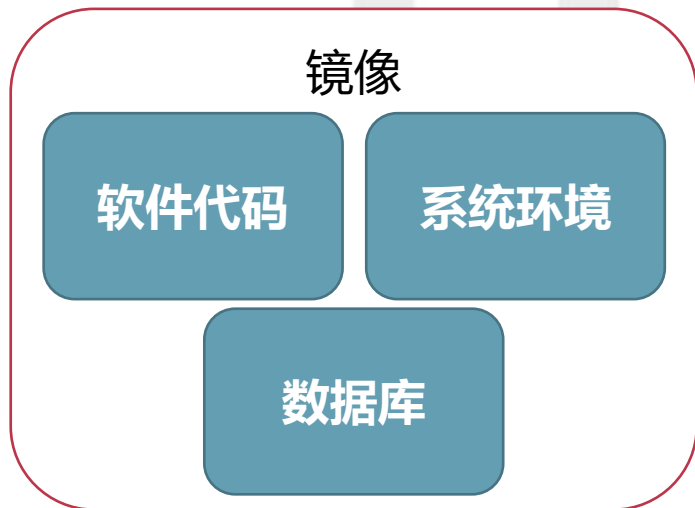
普通用户，没有  
管理员权限



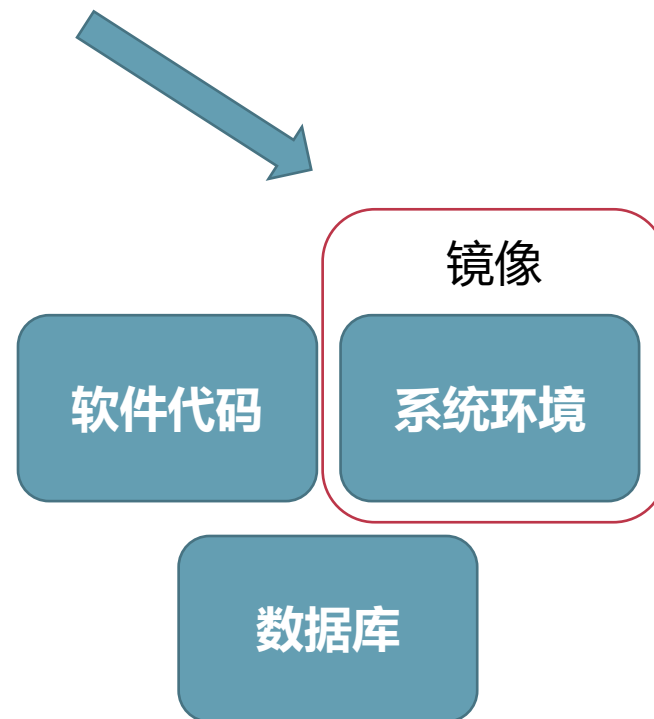
BOSS工作路径  
位于管理员目录



容器技术  
Singularity  
规避管理员权限



胖容器







轻容器



	胖容器	轻容器
特点		
优点	不需要考虑网络，不需要考虑系统版本兼容性，部署速度快	和传统使用方式几乎保持一致，实时更新BOSS和数据库资料，镜像小（3GB）
困难	难以应对频繁更新的BOSS和数据库资料	需要解决CVMFS的安装适配
	需要占用大量的存储空间（100GB）	需要解决计算节点的对外网络连接
	在版本维护上需要耗费大量的人力	数据库的部署和与计算节点的连接问题
	改变使用习惯	
技术路线	保底方案	优先选择

# CVMFS安装

CVMFS的安装也涉及root权限和系统版本的限制

- 传统的apt-get install——涉及root权限 
- 使用已经编译好的Cvmfsexec程序——系统版本冲突 
- Parrot-mount安装虚拟机——涉及root权限 
- 我们的解决方法: 从源代码编译CVMFS 

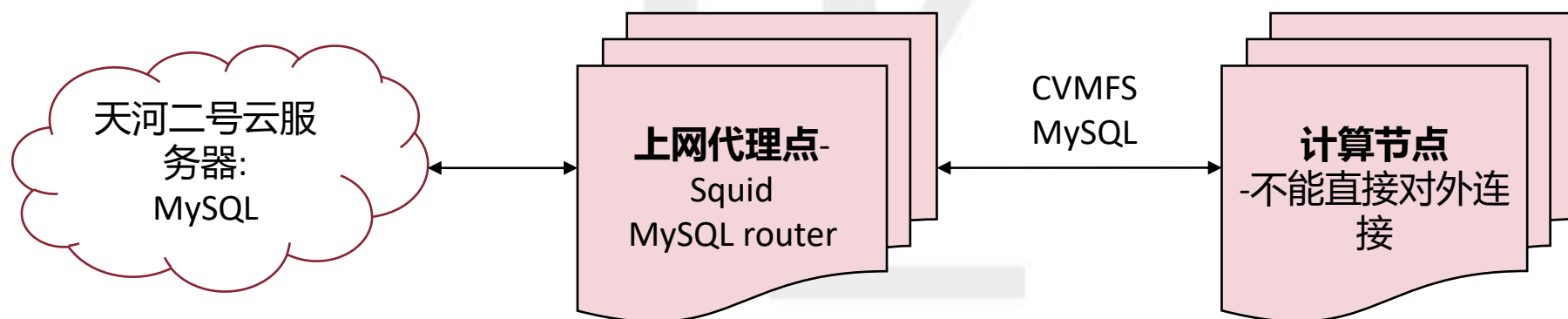
# 连接数据库

## Squid-cvmfs

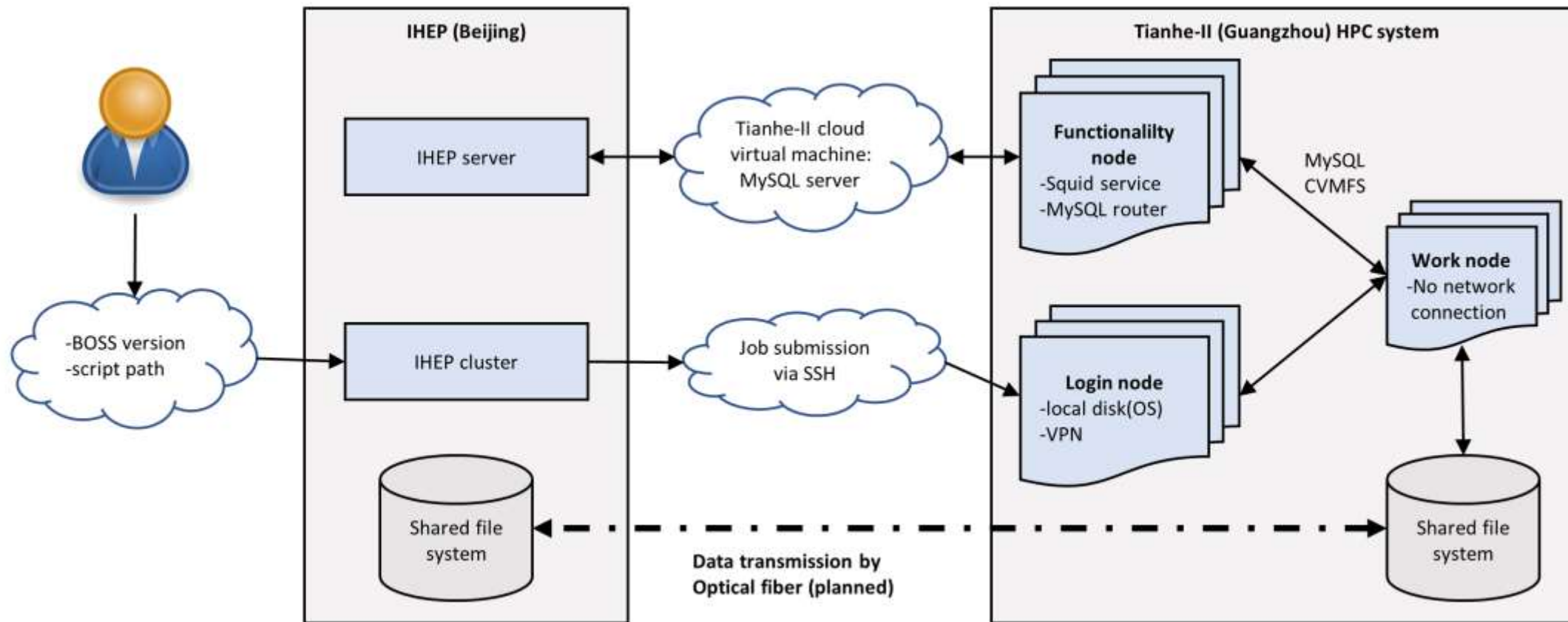
- 常见的HTTP缓存代理工具
- 天河2号通过上网代理点能有限的访问外网——IHEP的代码库
- 搭配更大的内存满足CVMFS性能

## MySQL

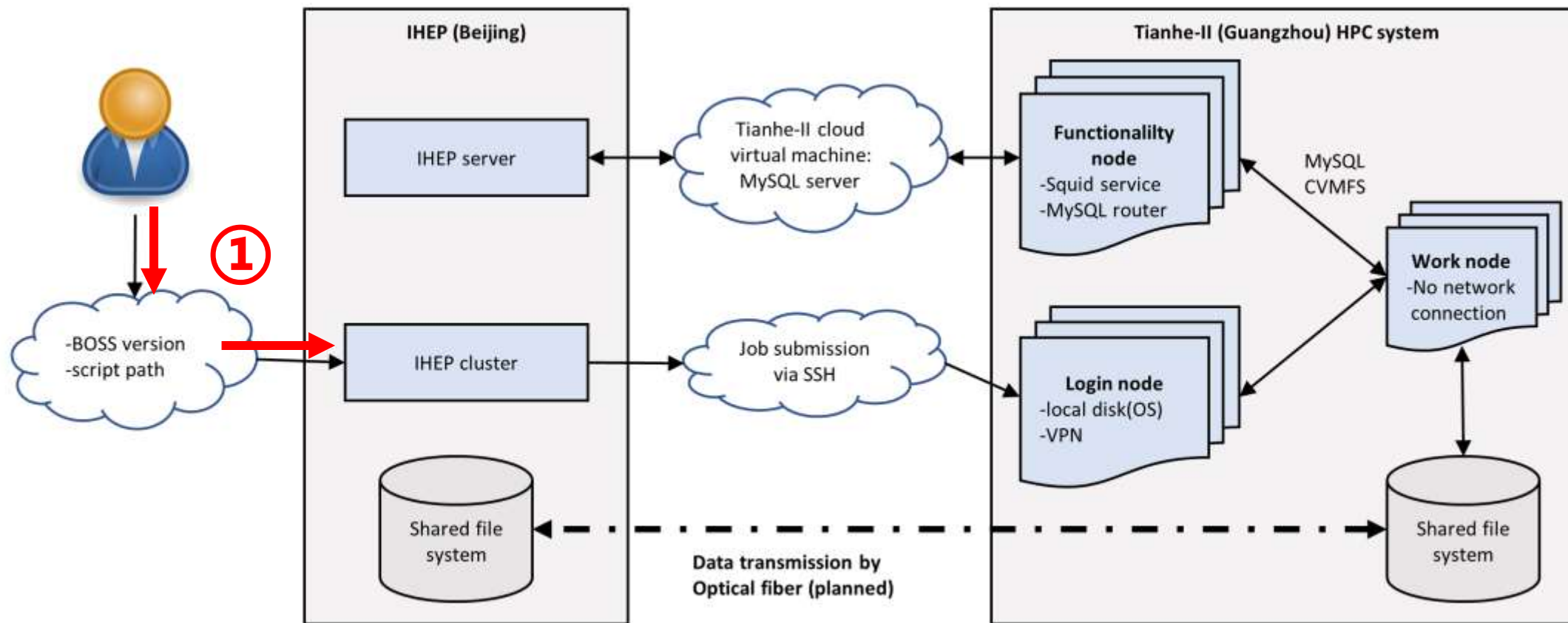
- 部署mysql在云平台.
- 云平台和高性能计算节点之间连接受限
- 在上网代理点部署MySQL-router实现计算节点和数据库之间的连接



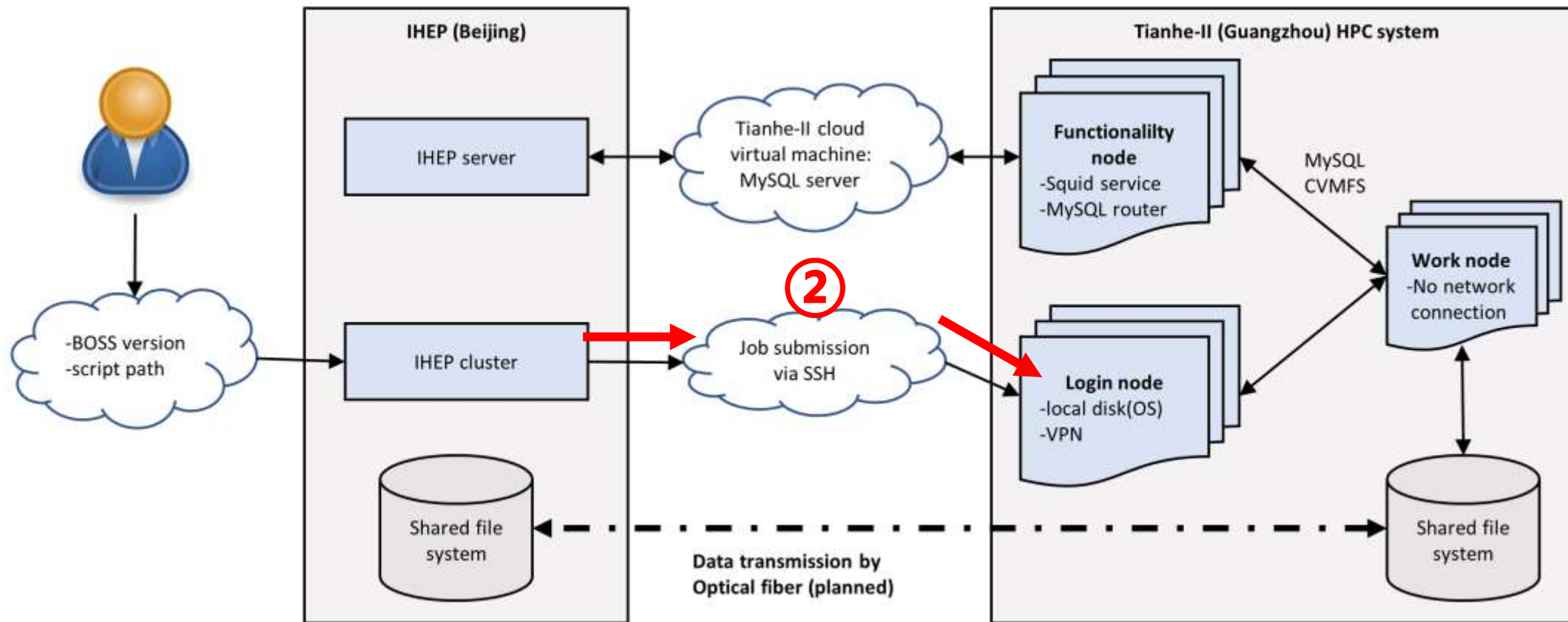
# workflow



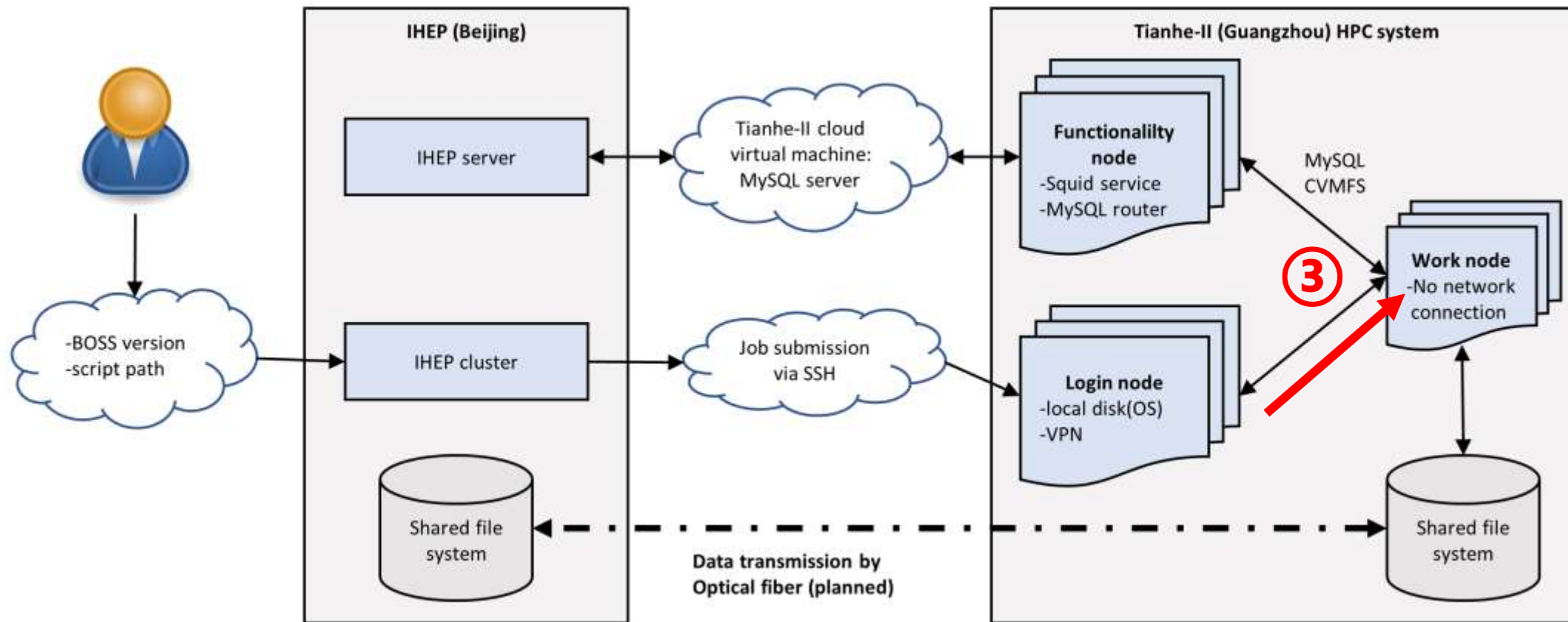
# workflow



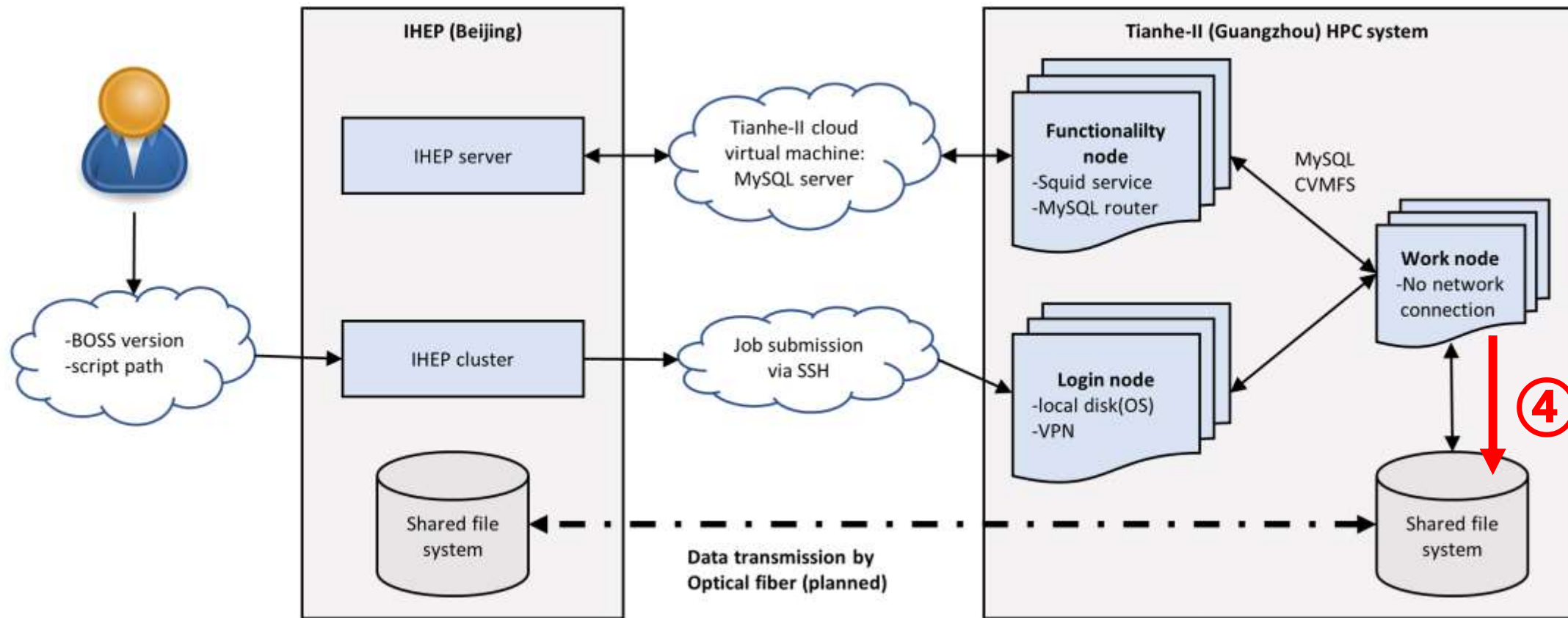
# workflow



# workflow

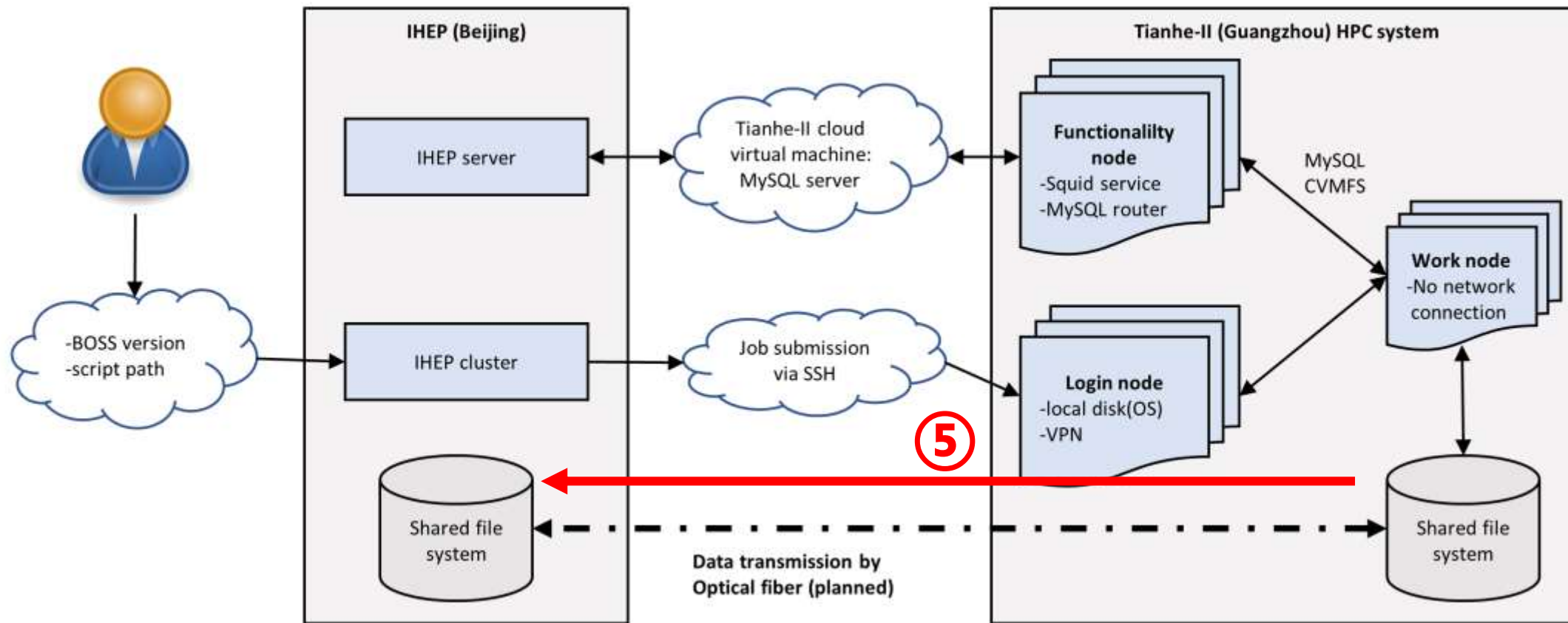


# workflow



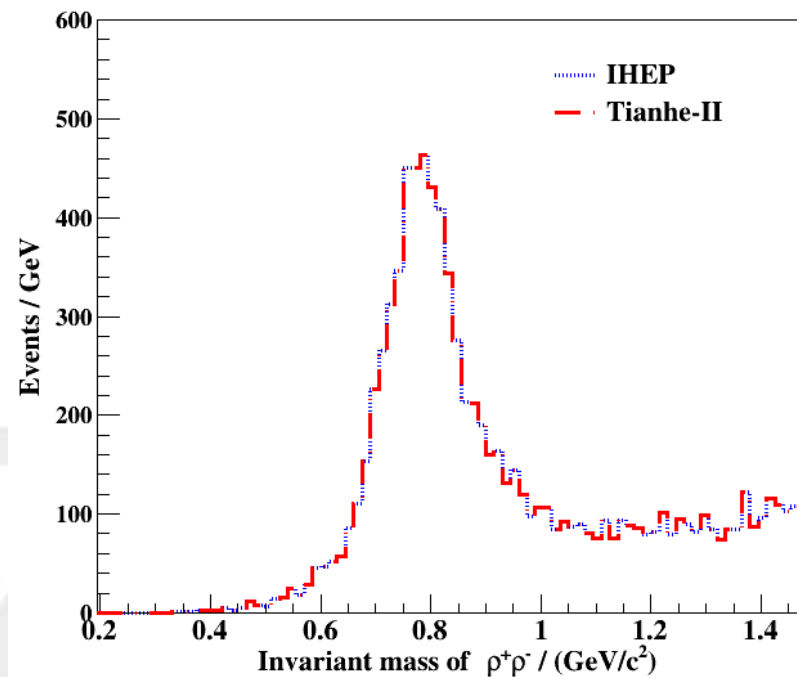
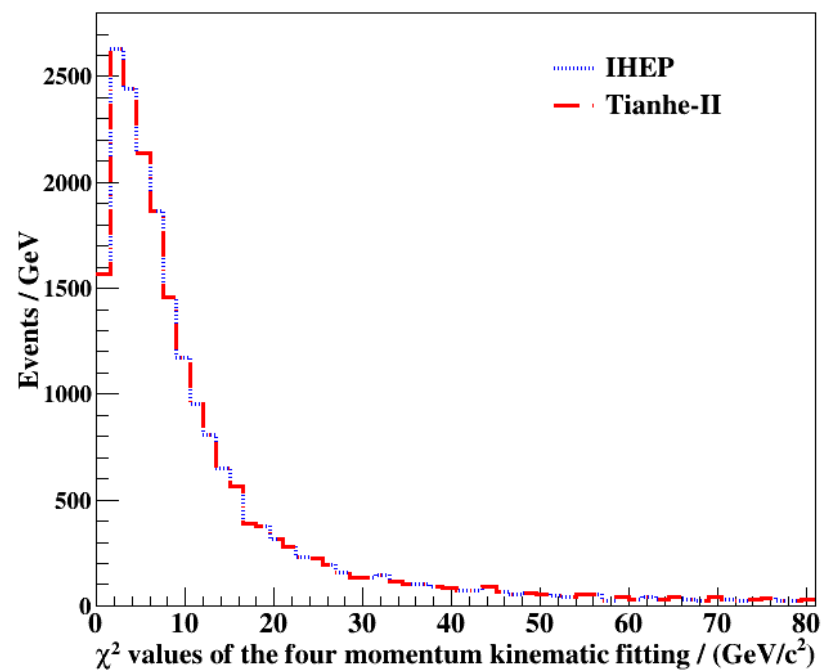


# workflow



# 验证

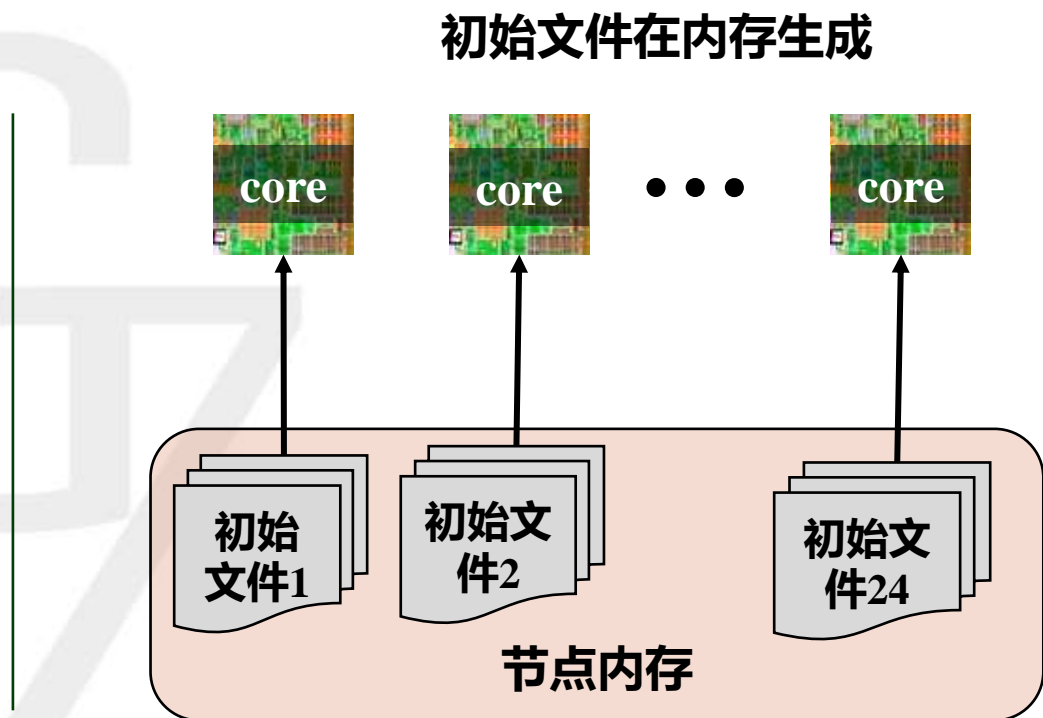
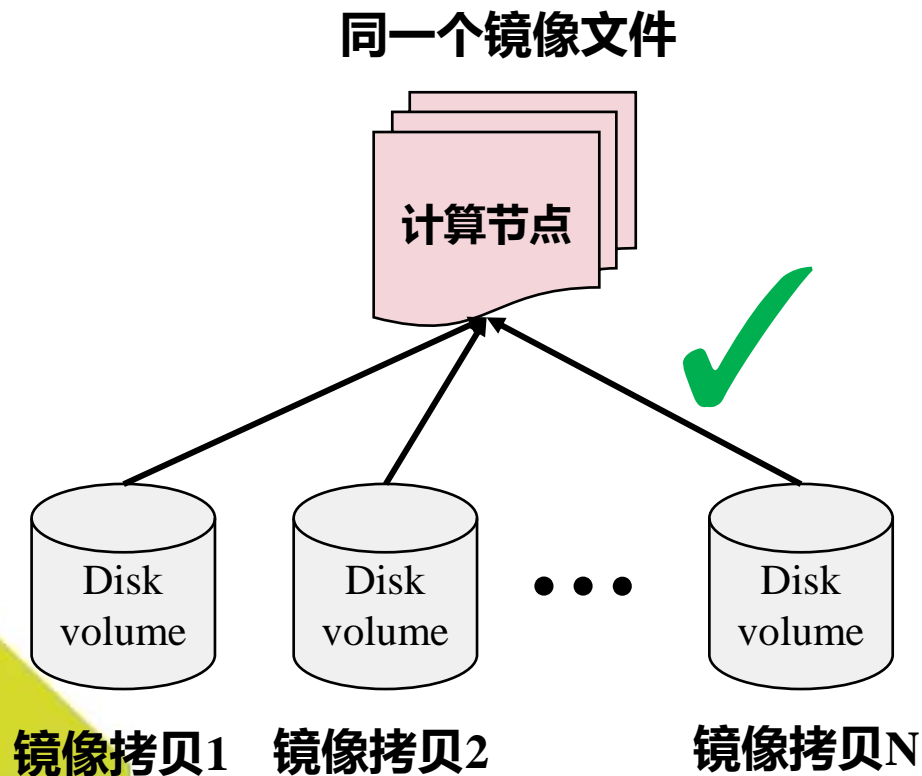
$$e^+ e^- \rightarrow J/\psi \rightarrow \rho\pi$$



IHEP和天河2号分别进行模式，物理结果一致。

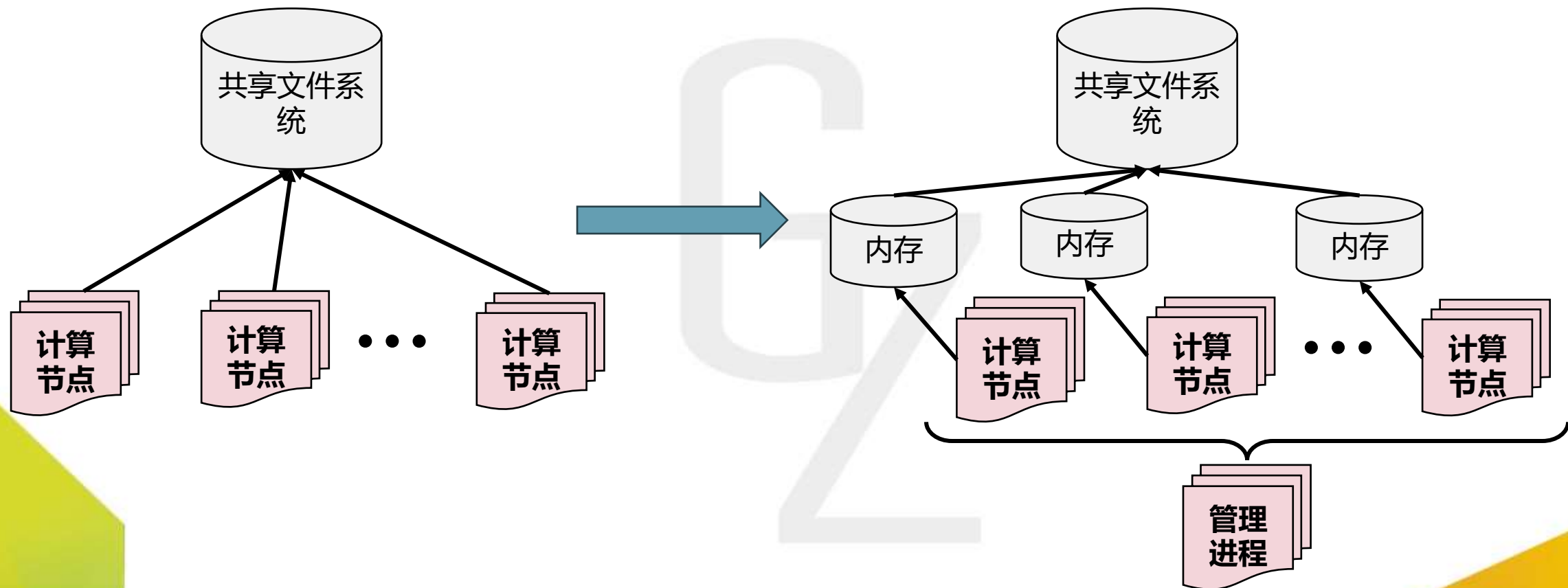
# 输入问题

- 多个计算进程从共享文件系统中同时读取**同一个文件**或同时读入**太多文件**，都容易引发IO堵塞。



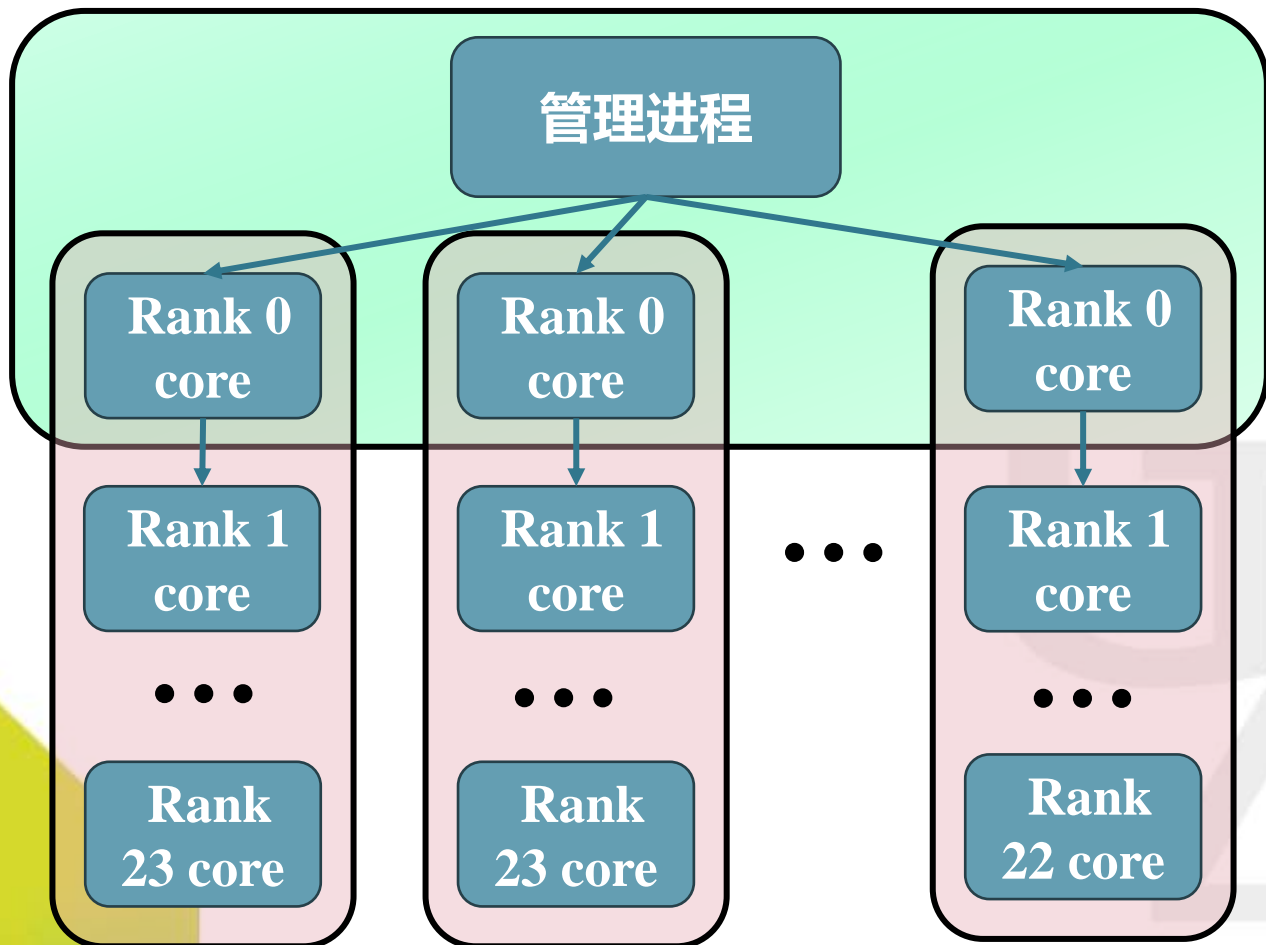
# 输出问题

- Slurm系统面对高通量作业短时间内输出太多小文件会引发系统崩溃。
- 采用先保存在结点的内存里，然后再排队输出的策略。



# 封装提交脚本

- MPI+python封装提交脚本
- 最后一个进程作为管理进程



```
def main():
    if rank 0:
        生成初始文件, 设定随机种子和输出地址

    if work_task:
        select (全局序号%硬盘的卷数):
            case 0:
                加载镜像0, 开始计算
            case 1:
                加载镜像1, 开始计算

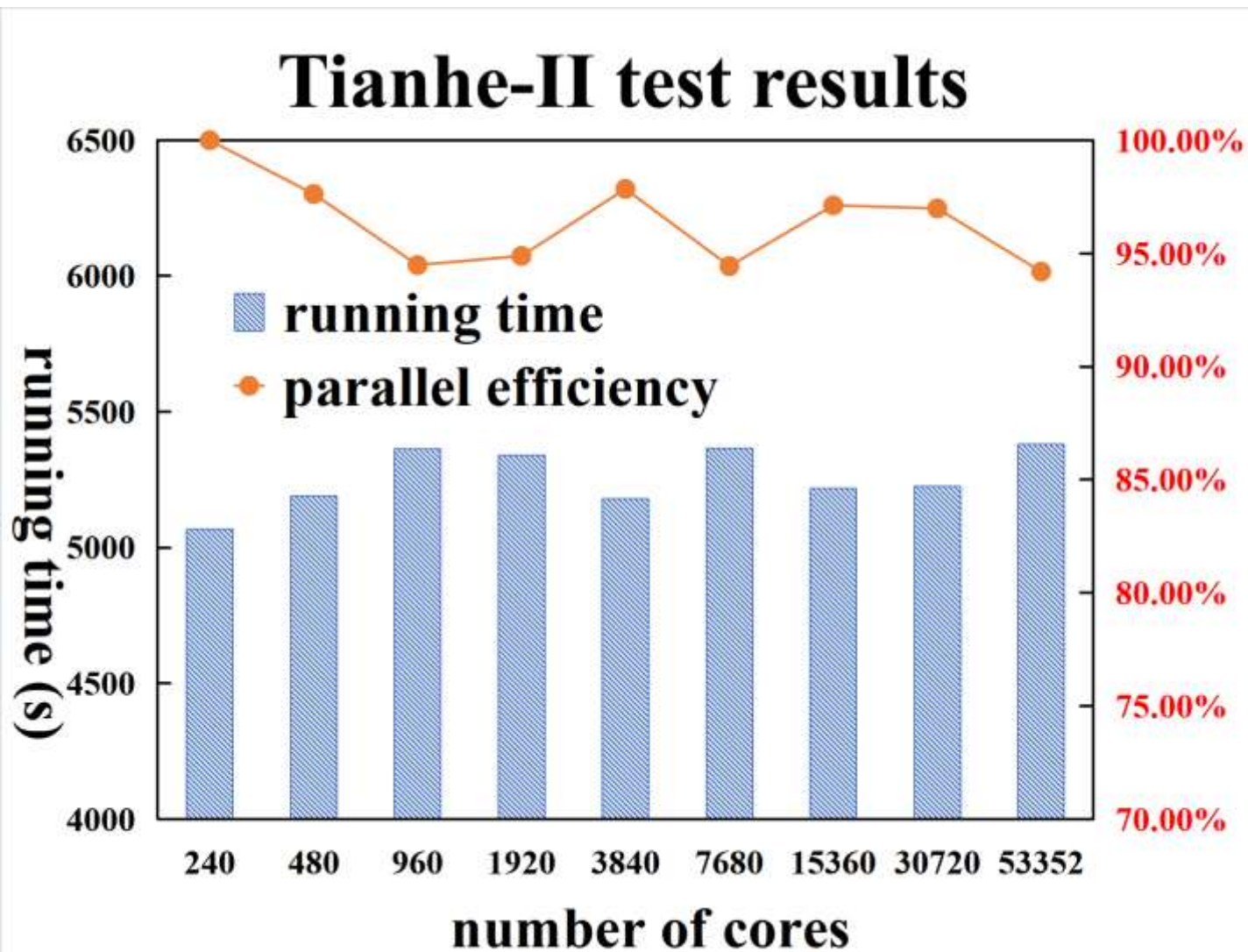
    if rank0:
        if 本节点其余进程都运算完毕
            加入管理进程的队列
    if 管理进程:
        管理队列, 前10名允许输出
```

初始化

运算

输出

# 大规模性能测试



使用**并行效率**来评价大规模性能表现

并行效率定义是:

$$E(N) = \frac{t_1}{N * t_N}$$

$t_1$  单节点耗时

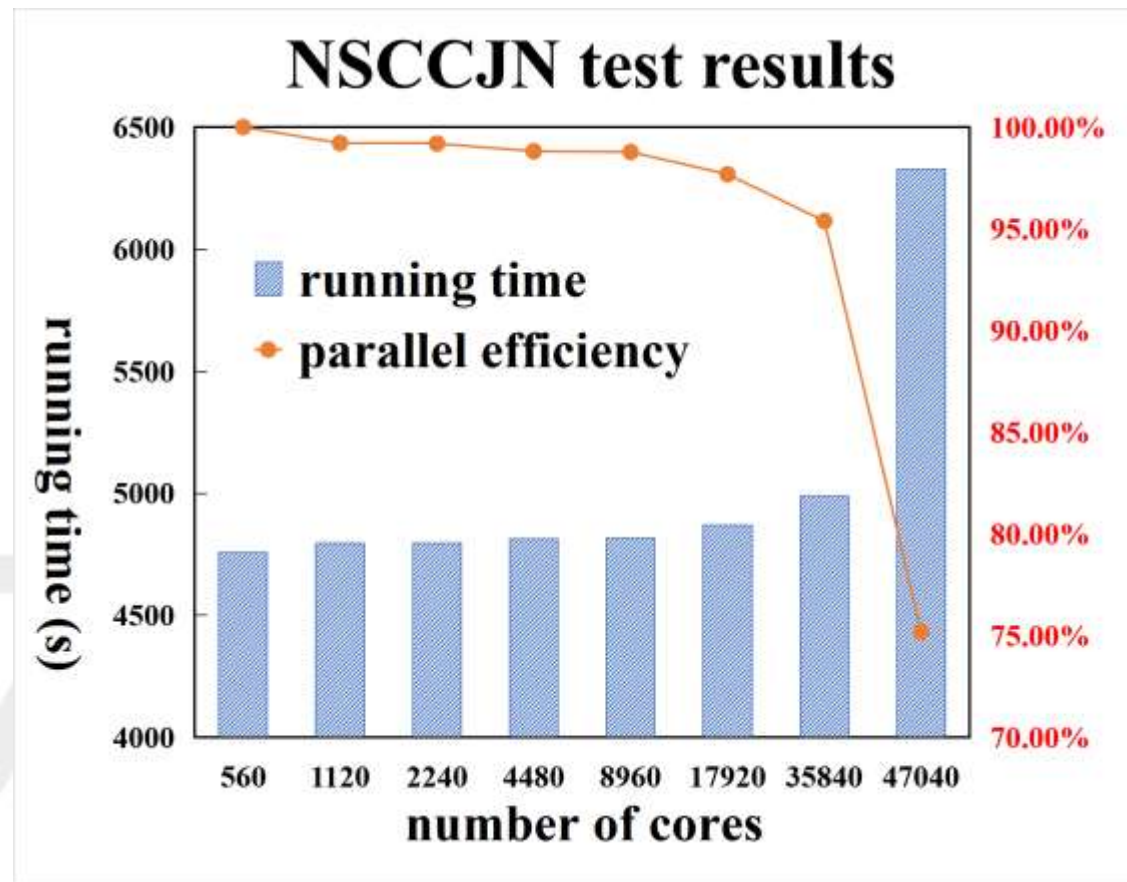
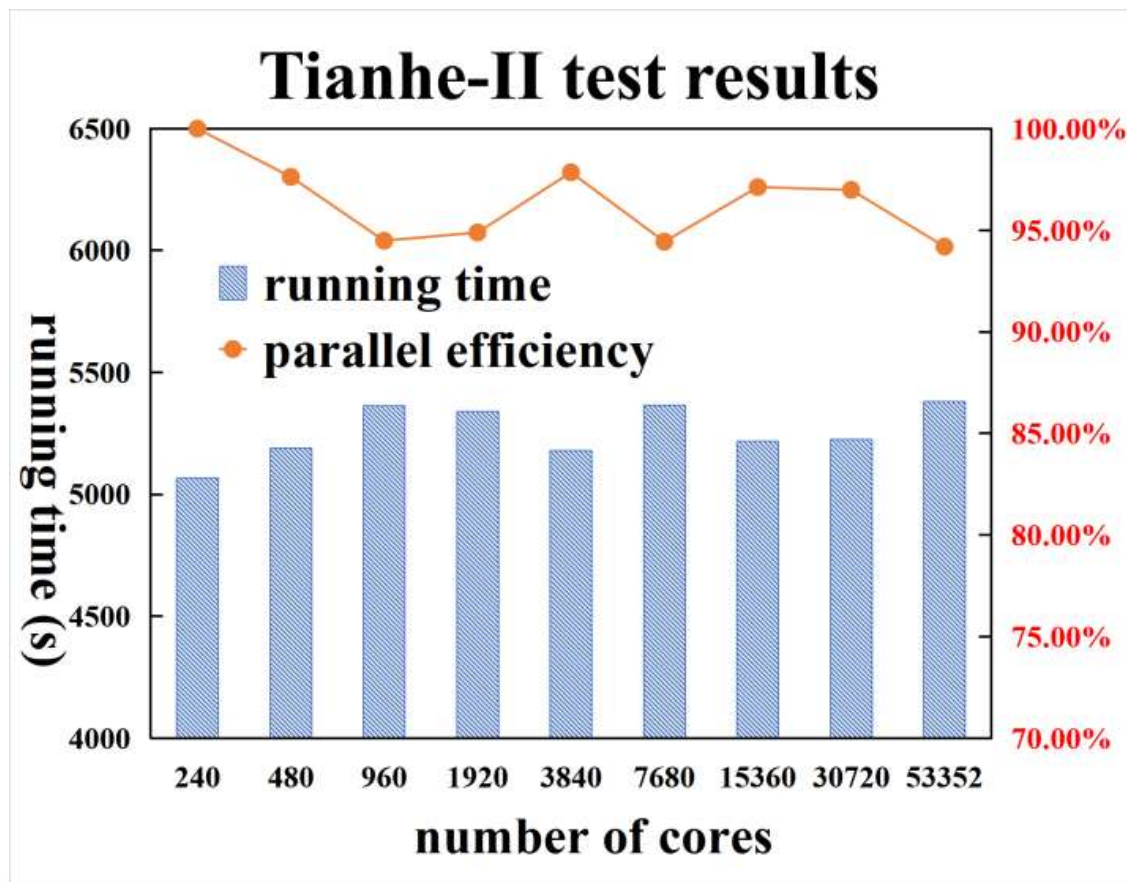
$t_N$  N个节点时的耗时

70%以上是可接受

80%以上是好

90%以上是非常好

# 大规模性能测试



在天河2号采用轻容器而在济南超算上使用胖容器同时运行大规模测试.

济南超算则在接近IO瓶颈时, 并行效率发生了快速下降。



# 总结和展望

## • 总结

- 在天河2号上实现了实时更新BOSS版本，保证核心部分使用方式不变。
- 从高能所远程提交作业到天河2号的功能正在开发中。
- 完成了大规模性能测试。

## • 展望

- 希望有真实使用的用户和我们合作。
- 优化数据传输能力，更好满足实际需求。
- 从BESIII扩展到更多。
- 更适应HPC的版本。





多谢

