


LHAASO数据处理平台建设及进展

程耀东

高能所计算中心

2016年8月15日

主要内容

- 计算需求
- 数据处理平台方案
- 建设进展
- 工作安排计划

离线数据处理平台

- 实验数据经过DAQ获取之后，进入离线计算平台
- 提供数据存储、传输、共享、分析处理的支撑服务

小型在站数据中心

稻城海子山
观测基地

当地
电信网络

稻城县城
测控基地

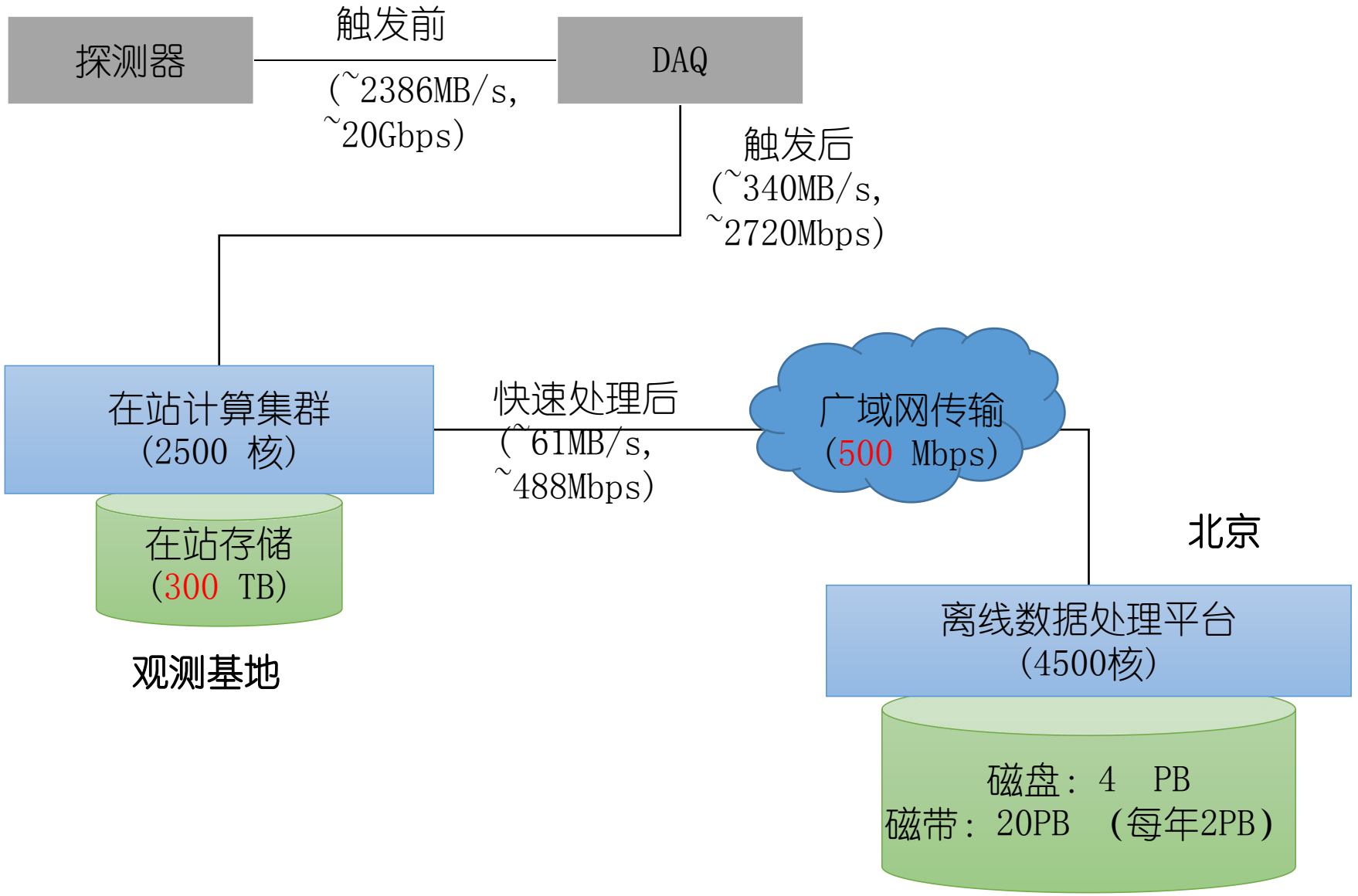
租用带宽

离线数据处理中心

北京高能所
计算中心

分布式计算平台

数据处理过程

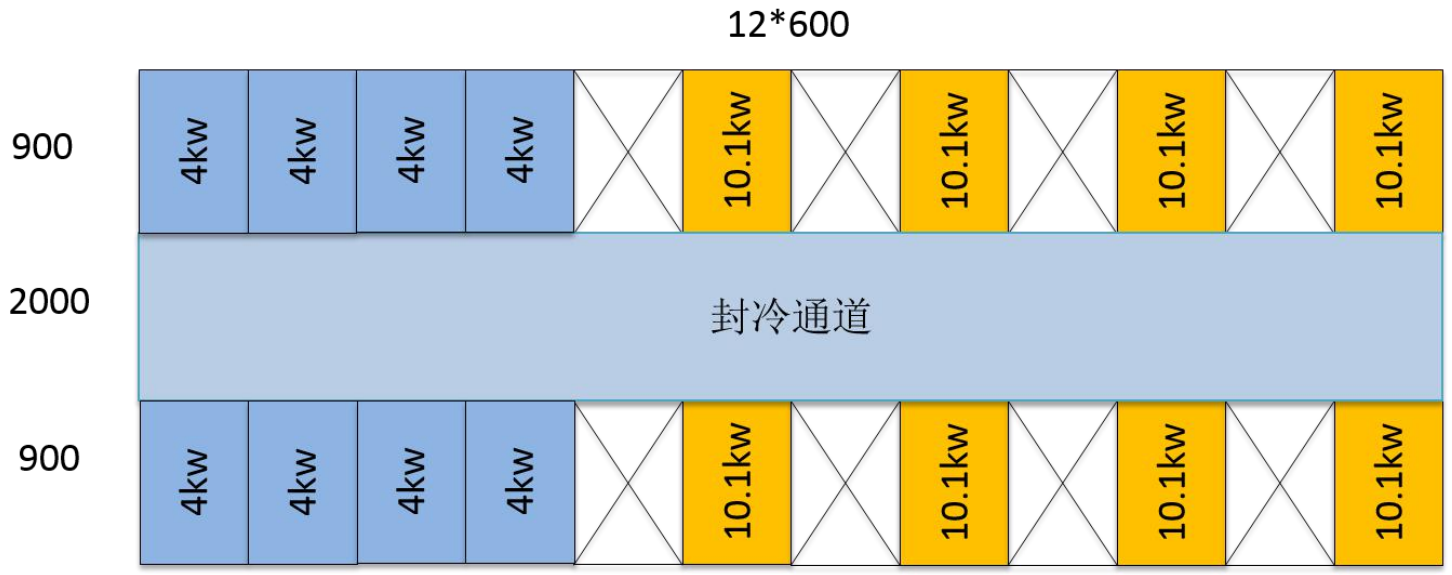

探测器
触发前
(~2386MB/s,
~20Gbps)
DAQ
触发后
(~340MB/s,
~2720Mbps)
在站计算集群
(2500 核)
快速处理后
(~61MB/s,
~488Mbps)
广域网传输
(500 Mbps)
在站存储
(300 TB)
观测基地
北京
离线数据处理平台
(4500核)
磁盘：4 PB
磁带：20PB (每年2PB)

计算资源需求

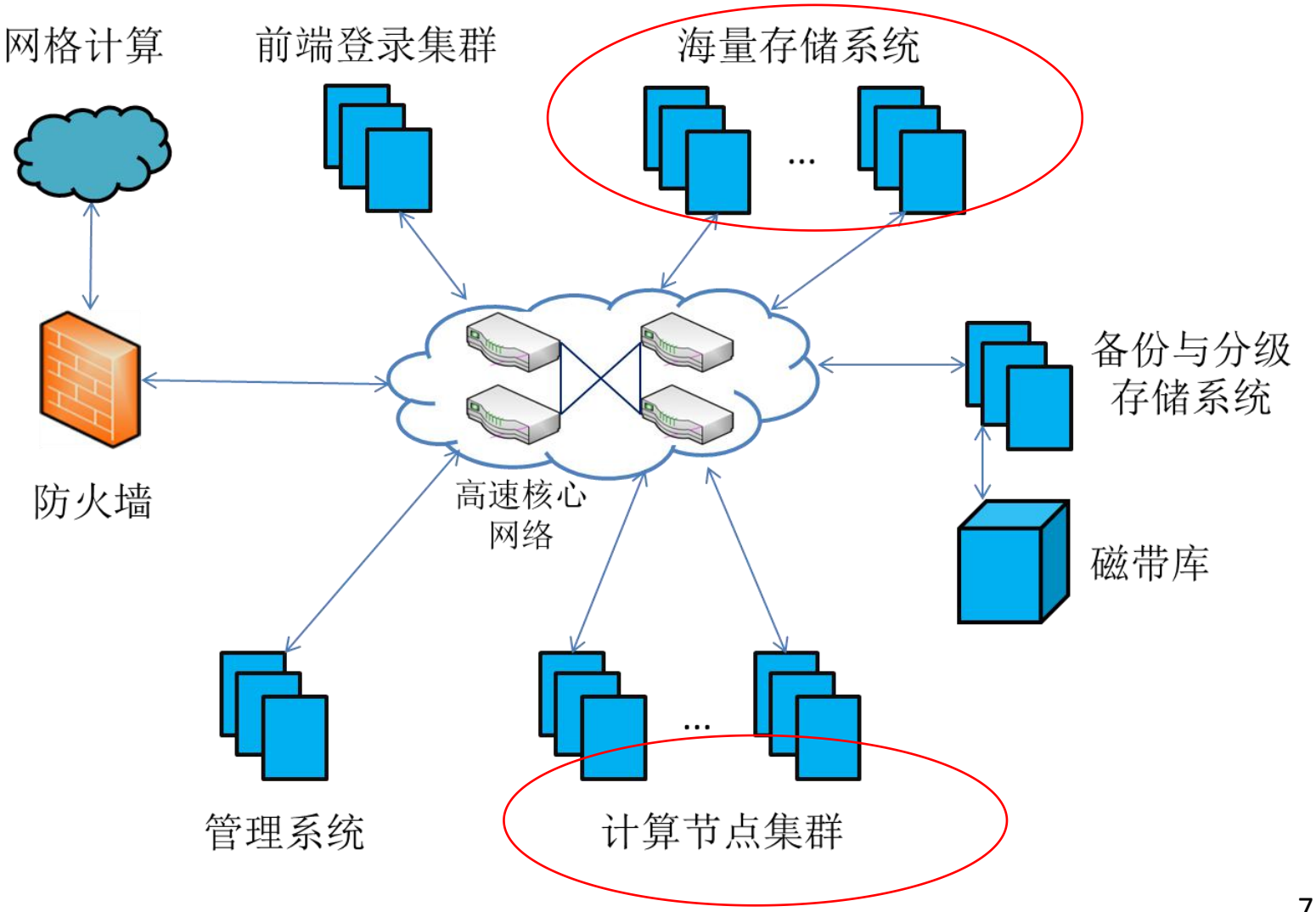
- 存储需求
  - 原始与重建数据存入磁带，每年1834TB
  - 共2265TB的模拟数据
  - 每年产生~2PB数据，共需要20PB以上的磁带库存储
  - 另外需要20PB的离线备份磁带柜以及4PB以上的高性能并行文件系统
- 计算系统
  - 7000CPU核的高性能计算系统
  - 4500核用于重建；1500核用于模拟；1000核用于分析
- 网络系统
  - 每年需要从实验站传输1834TB，需要500Mbps的带宽
  - 考虑带宽利用率，建议租用1Gbps的链路
- 目前国内产生数据量最大的科学实验装置之一

在站小型机房设计

- 满足DAQ以及快速重建等计算需求
- 2015年已经确定初步方案，并完成内部评审
- 规划面积：~190m²
- 16个机柜（8个高密，8个低密）
- 16箱刀片，>4000CPU核，300TB存储
- 总功率~100KW

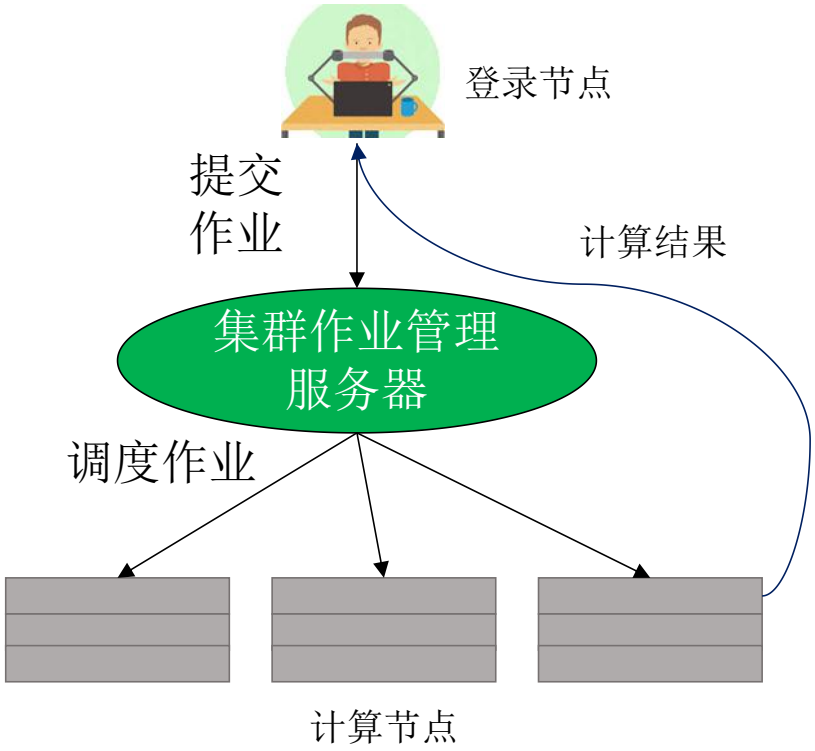


离线数据处理平台架构



本地计算集群

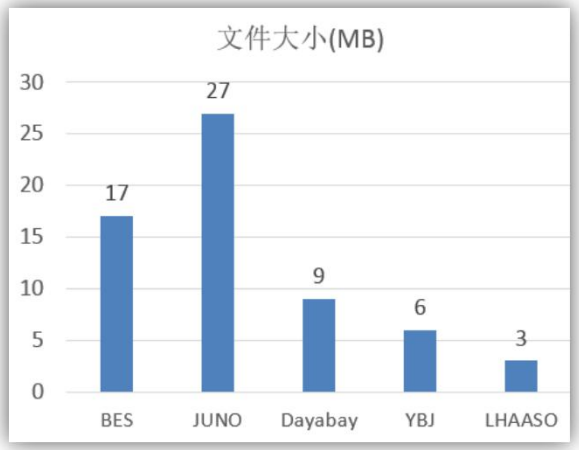
- 管理计算节点，调度作业
- 提供用户提交作业接口
- PBS
  - 历史悠久
  - OpenPBS, PBS Pro, Torque
  - IHEPCC当前主要调度系统
- HTCondor
  - 更好的性能，更多的功能
  - 调度算法更为公平
  - 已经应用在JUNO等实验
  - IHEPCC计划全部迁移到HTCondor
- SLURM: 高性能计算调度
- LSF：商业调度软件



LSF

存储系统

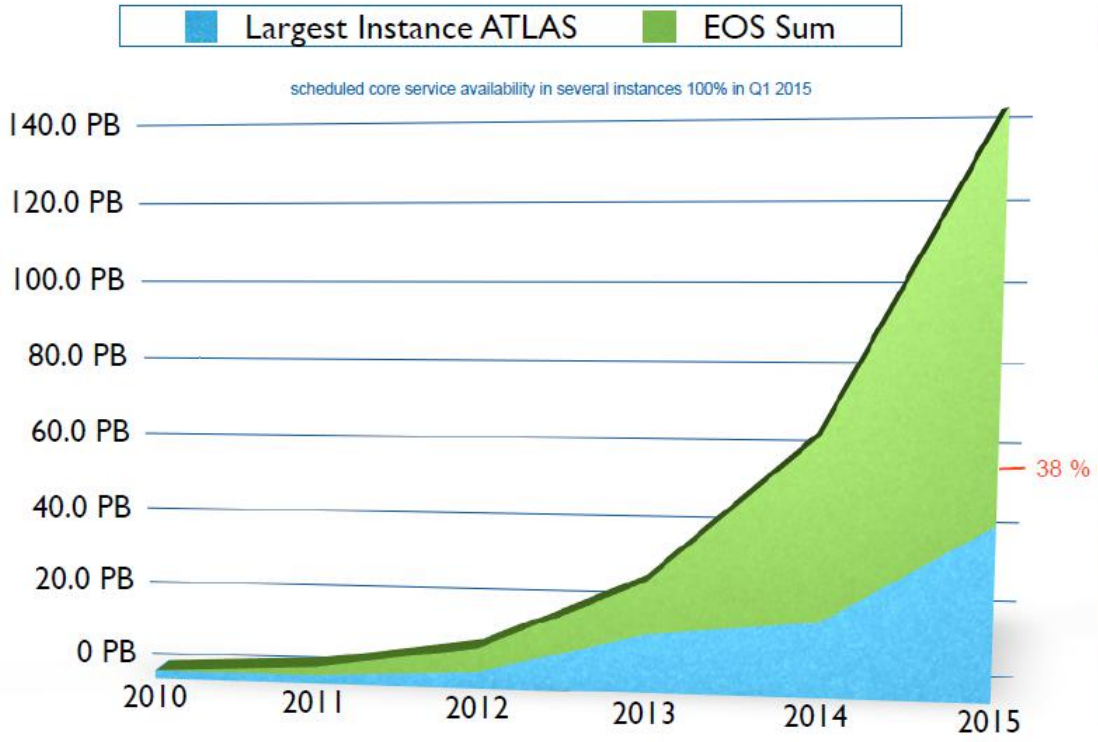
- 磁带存储系统
  - 将顺序设备映射成类似于存储系统的树形目录
  - CERN CASTOR, enstore等开源软件
  - TSM等商业软件
- 磁盘存储
  - Lustre
  - Gluster
  - EOS
- 用户目录
  - AFS
- 软件库共享系统
  - AFS, CVMFS



常见文件系统的特点对比

	EOS	CephFS	GlusterFS	Lustre
存储硬件	JBOD	JBOD	专业磁盘阵列/JBOD	专业磁盘阵列
元数据服务器	双MDS互相备份，内存存放元数据，不存在单点故障，存在瓶颈	有多个MDS，不存在单点故障和瓶颈	无MDS，动态算法替代，不存在单点故障	双MDS互相备份，不可扩展，存在瓶颈
冗余保护/副本	支持Replica、Archive、N+M等多副本	N+M	镜像	无
数据可靠性	多副本提供可靠性	多副本提供可靠性	镜像提供可靠性	由存储节点上的RAID1或RAID5/6提供可靠性
故障恢复	支持主备模式，数据多副本，自动恢复	节点失效自动迁移数据、重新复制副本	系统自动处理故障	无
扩展性	元数据服务器不可扩展、存储节点可扩展	元数据服务器、存储节点可扩展	存储节点可扩展	存储节点可扩展，元数据服务器不可扩展
典型应用场景	海量数据分析 CERN大规模使用	云计算环境中的块存储	多媒体应用、互联网	超级计算系统

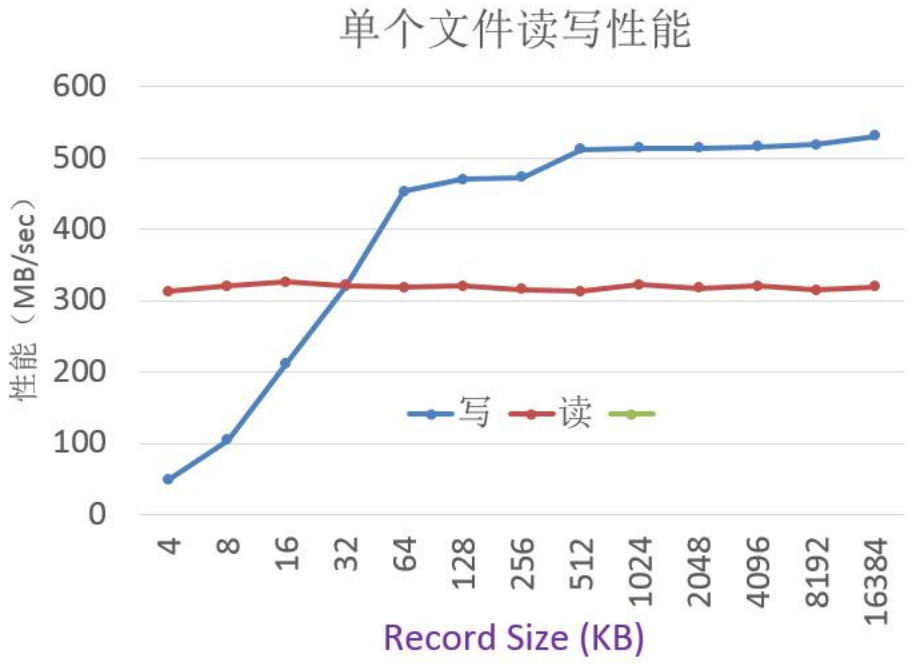
CERN EOS应用情况

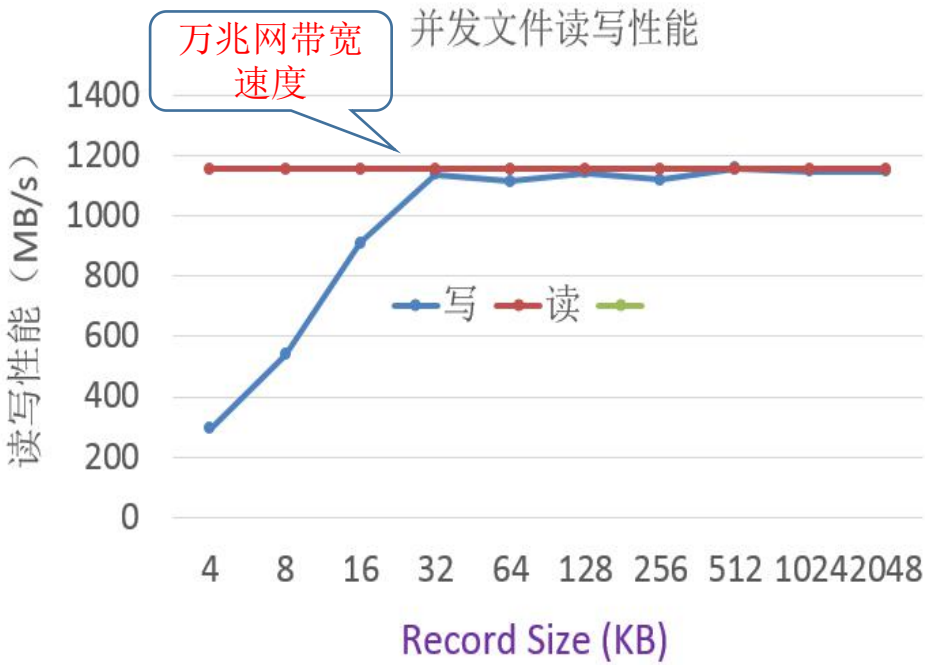


April 2015	
Capacity	140 PB
Server	1.400
Hard Disks	44k
Files	271 M
Directories	26 M
Replicas	0.5 B
Connectivity [theor.]	13 Tbit
random IOPS	2.2 M
Disk BW [theor.]	3.3 TB/s
Internal Messaging	150 kHz
State Machine	3M kv pairs
Users storing data	~3k
Quota rules	9.600

美国、俄罗斯、英国等高能物理实验室部署数十PB

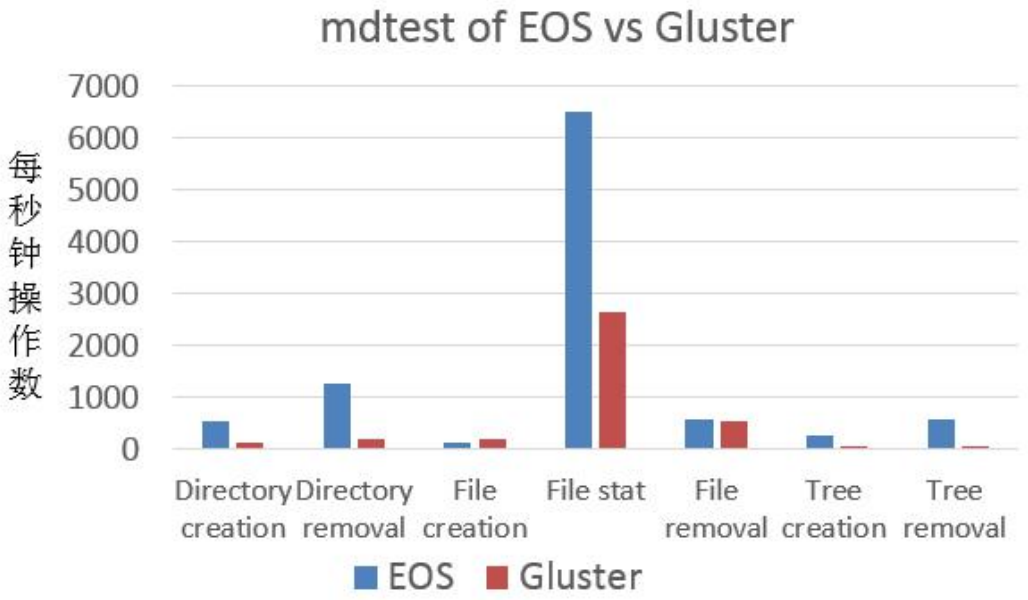
EOS文件读写性能测试

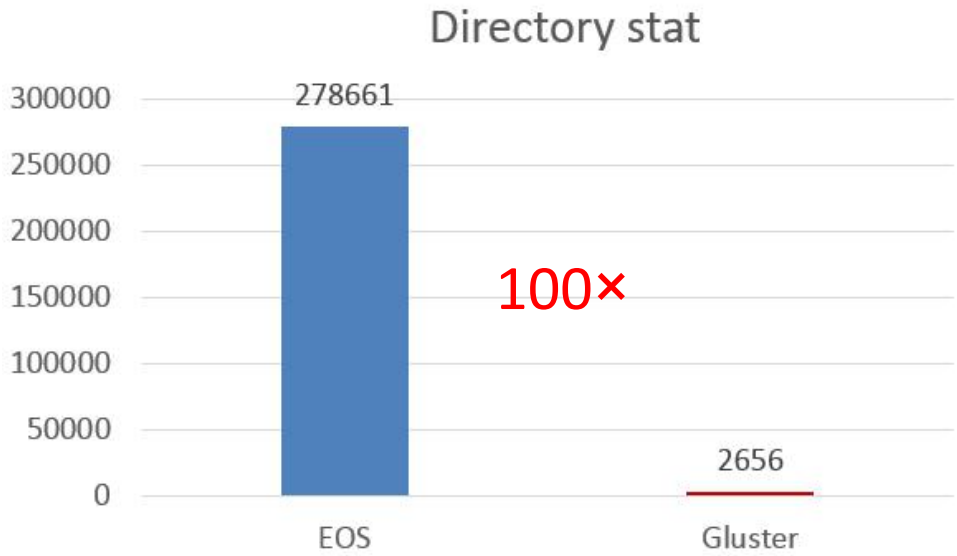




1. 用户调用write函数写入文件时，尽量采用128KB以上的缓冲区大小
2. 万兆网环境下单文件的读写性能超过300MB/sec以上
3. 单客户端万兆网环境下，并发文件的读写性能可以达到1000MB/sec，跑满网络带宽
4. 高并发情况下服务器带宽全部可以利用

EOS元数据操作性能



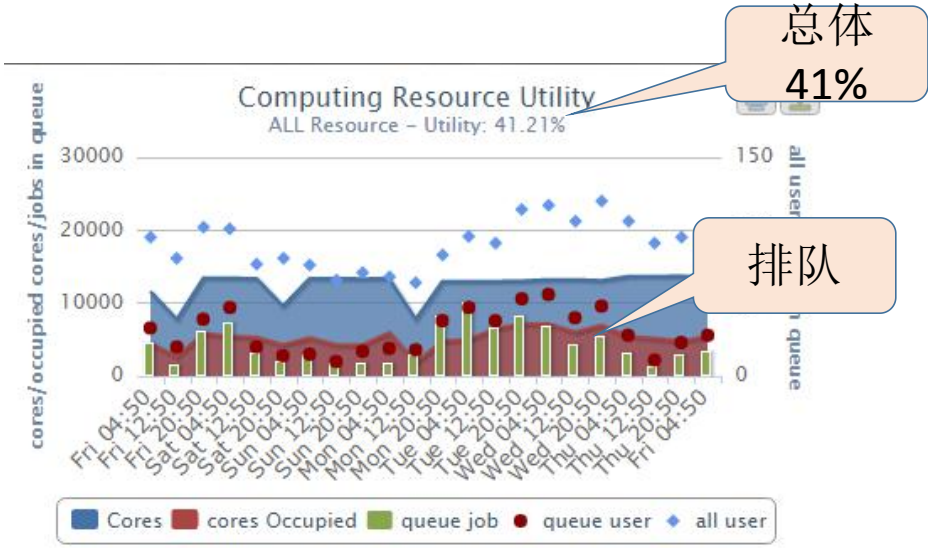


CERN的系统监视显示

- 多线程服务器可达到每秒百万次的stat请求

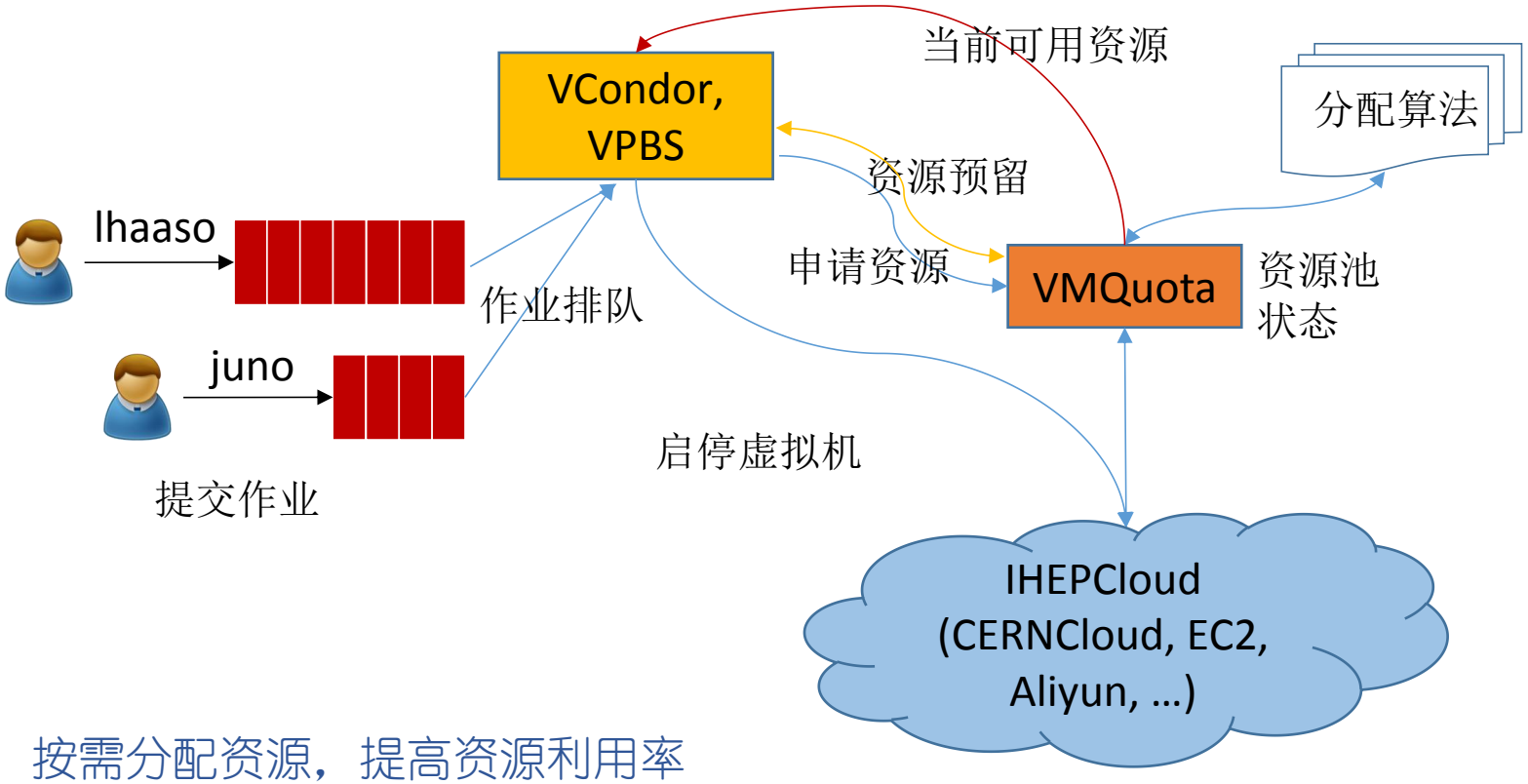
传统集群存在问题

- 队列使用控制
  - 不允许其它人或其它组使用
- 操作系统等运行环境不兼容
  - 不同实验组之间不能互相运行作业
- 不支持抢占，资源回收慢
- 调度不灵活，运维成本高



引入虚拟化和云计算

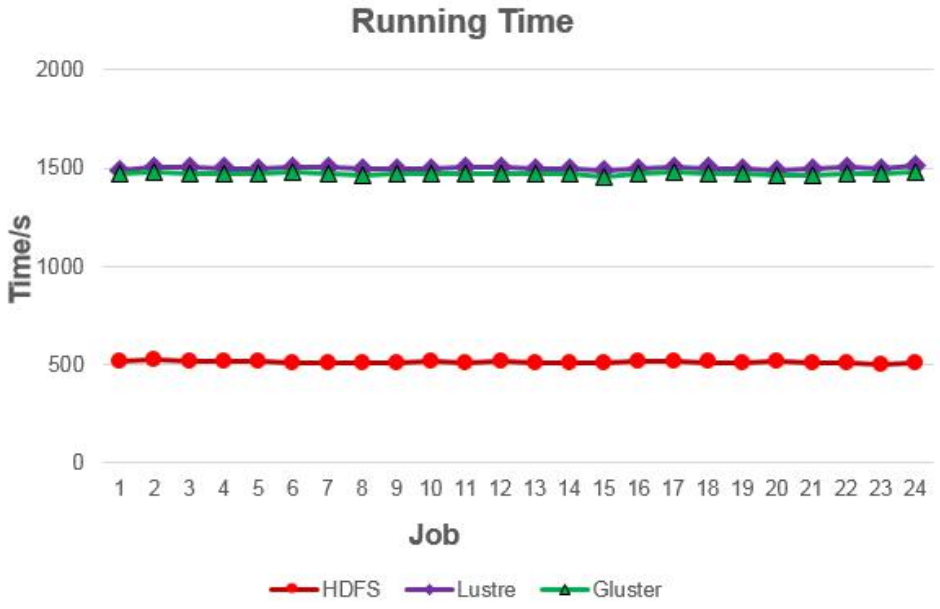
基于云计算的虚拟集群



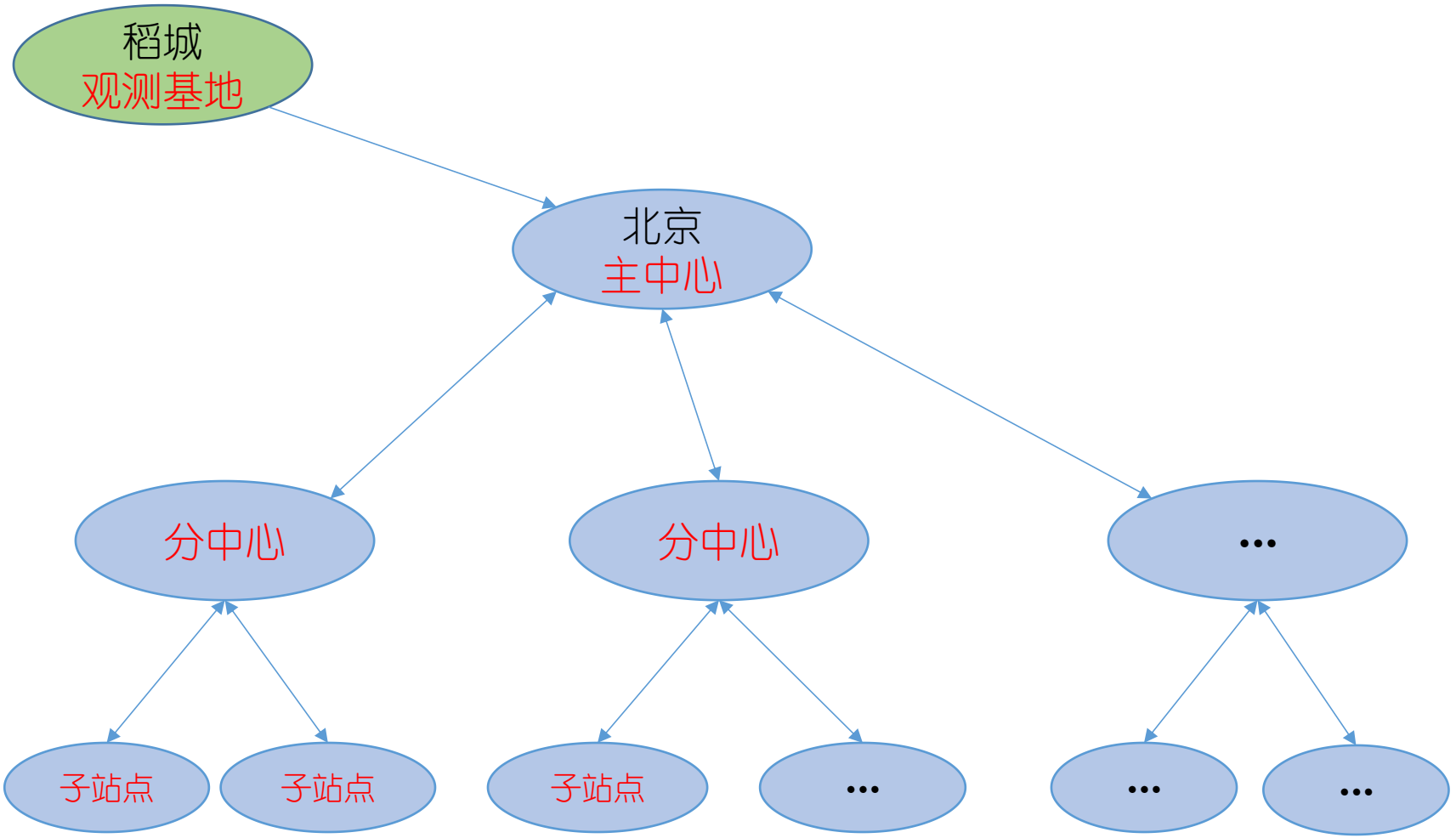
- 按需分配资源，提高资源利用率
- 实现资源整合，共享不同实验/组织的计算资源
- 满足峰值需求

基于Hadoop的数据分析平台

- 区别于传统的“存储-计算”分离的架构，采用“存储-计算”统一节点的方式，试图解决I/O瓶颈问题
- 初步测试结果显示，本地数据访问具有更好的数据分析性能，运行时间只有网络文件系统的1/3
- 需要进一步细化工作
  - 调度策略
  - 资源共享
  - 数据迁移
  - 用户易用性



LHAASO分布式计算方案

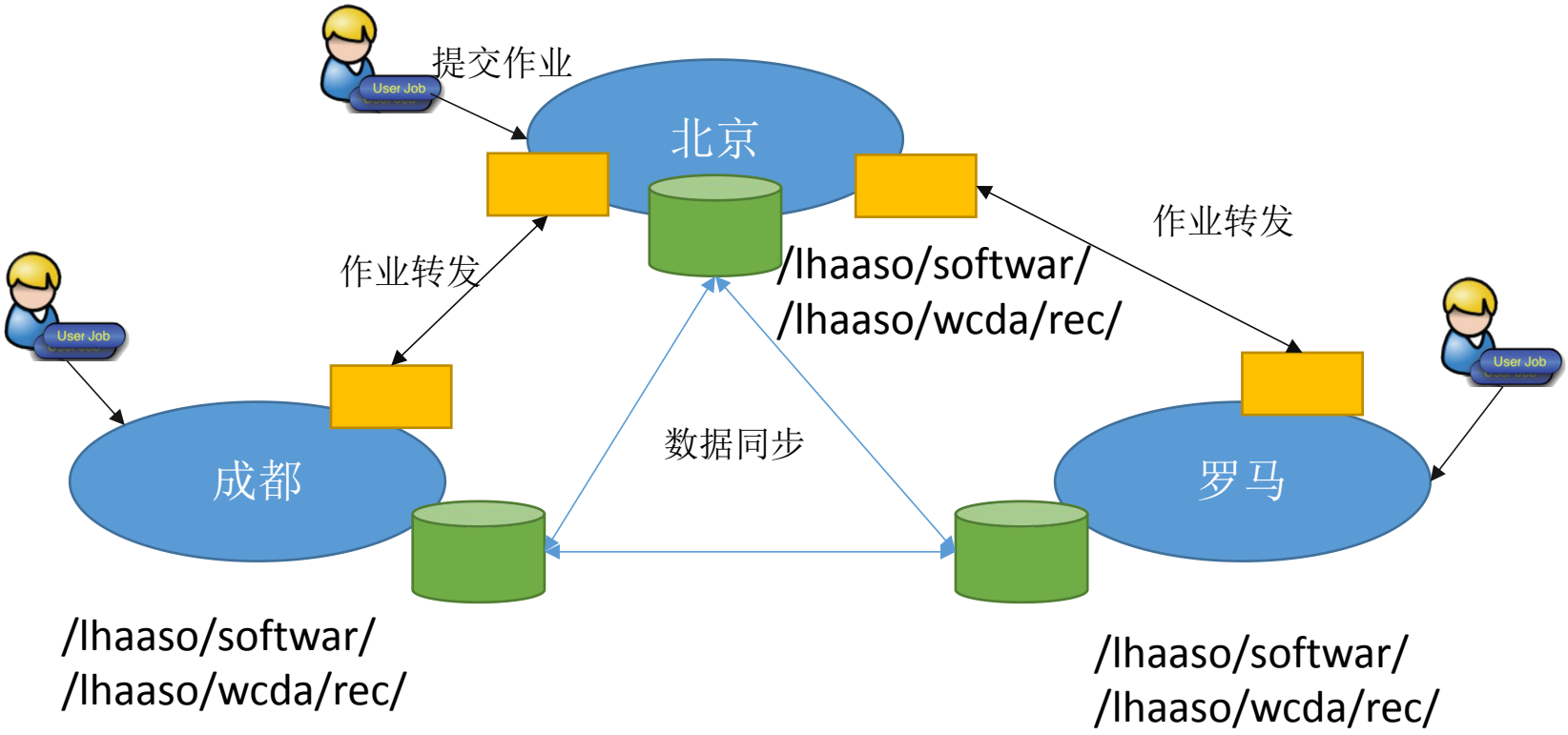


各中心功能

- 稻城观测基地
  - DAQ、数据过滤、快速重建、压缩等
  - 将原始数据和快速重建数据传输到主中心
- 主中心
  - 所有数据（原始、重建、模拟、分析等）安全存储
  - 全部数据重建计算
  - 将重建数据分发到各个分中心
  - 接收来自分中心的模拟和分析数据
  - 负责LHAASO分布式计算系统（包括分中心、子站点）建设
  - 负责LHAASO分布式计算系统（包括分中心、子站点）技术支持
- 分中心、子站点
  - 模拟、分析

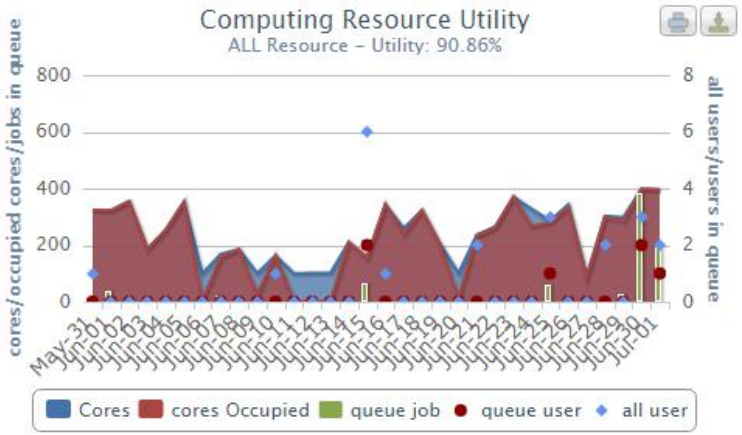
基本技术方案

- 基于云计算技术，多中心统一管理，远程运维
- 多中心统一系统视图，存储和计算环境一体化，对用户完全透明
- 用户作业优先在本中心运行，资源不够时自动转发到其它中心



目前LHAASO计算环境

- 集群计算系统
  - Torque：~500 CPUcore，将迁移到HTCondor
  - HTCondor: ~600 CPU core （共享）
- 存储系统
  - Gluster：347TB，剩余46TB
  - EOS：231TB，剩余160TB （共享）
- 云计算，
  - IHEPCloud，支持HTCondor/EOS
  - 已经为LHAASO提供80万CPU hour的服务
- Hadoop平台
  - 84 CPU Core, 35TB；新增120 CPU Core，150TB
- 分布式计算系统
  - 成都情报文献中心，200CPUCore，145TB



工作安排

- 2016年：方案细化
  - 在站机房方案细化，充分考虑高原环境与运维问题
    - 系统可靠性方案、设备可靠性测试
  - 离线数据处理平台系统的方案验证
  - 现有系统运行，提供稳定服务
- 2017年~2018年：系统建设，提供服务
  - 在站机房的建设
  - 计算、存储、网络以及管理系统的建设
- 2018年以后，系统扩容，运行维护
- 新技术研究与应用
  - 基于磁盘服务器等新型存储系统研究
  - GPU等高性能计算技术应用
  - 分布式计算、远程文件系统等
  - 基于Hadoop的数据分析平台

小结

- 总体架构基本确定，需要进一步细化方案和测试
- 已经建设了小型的离线数据处理平台
- 云计算等新技术的研究和应用正在开展
- 希望更多的人才加入到IT技术的研究
- 合作单位贡献资源，加入LHAASO分布式计算环境

谢谢

chyd@ihep.ac.cn

- backup

计算资源需求

- 计算资源需求，包括模拟、重建两个部分，共需要5776个CPU核;物理分析需要1000CPU核

分总体	重建CPU需求（核）	备注
KM2A	500	
WCDA	3560	其中一部分放在在站机房
WFCTA	360	
总计	4420	

	模拟CPU需求（核）
伽玛天文样本	800
多参数宇宙线样本	300
1016-1017能量区间的Chereknov模拟样本	30
Sub-EeV荧光模拟样本	26
总计	1156

原始与重建数据存储需求

分总体	原始数据(TB/年)	重建数据量(TB/年)	标定数据(TB/年)	备注
KM2A	220	33	忽略	磁盘暂存半年原始数据，所有数据带库永久保存
WCDA	1200	150	忽略	原始数据是经过在线重建以后的数据；重建数据是去掉hit信息以后的数据。磁盘保存所有重建数据，原始与重建数据写入磁带
WFCTA	200	30	1	磁盘暂存半年原始数据，所有数据带库永久保存
总计	1620	213	1	共1834TB/年

LHASSO模拟计算产生2265TB数据