LHAASO数据处理平台建设及进展

高能所计算中心 程耀东

2017年1月18日

云南-昆明

主要内容

- 计算需求
- 数据处理平台方案
- 建设进展
- 工作计划

计算资源需求

- 初设报告：计算资源需求，包括模拟、重建两个部分，共需要5776个CPU核
- 没有覆盖物理分析部分，经费来源？

探测器	数据重建（CPU核）	数据样本	模拟产生（CPU核）
KM2A	500	伽玛天文样本	800
WCDA	3610	多参数宇宙线样本	400
WFCTA	360	10^16-10^17能量区间的Chereknov模拟样本	60
		Sub-EeV荧光模拟样本	46
小计	4470	小计	1306
合计			5776

原始与重建数据存储需求

探测器	原始数据（TB/年）	重建数据（TB/年）	数据样本	存储量（TB）
KM2A	220	33	伽玛天文样本	1600
WCDA	1250	150	多参数宇宙线样本	600
WFCTA	200	30	10^16-10^17能量区间的Chereknov模拟样本	42
小计	1670	214	Sub-EeV荧光模拟样本	23
			小计	2265
合计		1884TB/年	合计（2年）	4530TB

原始数据：～2PB/年；模拟数据：　～4.5PB数据

建设期规模

- 存储系统
  - 建设期磁带预算5PB，只能保存2年左右的数据
    - 按照10年计算，需要40PB（20PB在线+20PB离线）磁带
  - 建设期4PB以上磁盘高性能并行文件系统
    - 如果CPU核增加，要满足I/O需求，实际需求会更大
- 计算系统：6000CPU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1884TB，至少需要500Mbps的带宽
  - 考虑带宽利用率，建议租用1Gbps的链路
- 国内数据量最大的科学装置之一

数据处理平台组成

- 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

在站小型数据中心

稻城海子山
观测基地

稻城县城
测控基地

远程运行控制 中心

租用带宽

北京大型数据中心

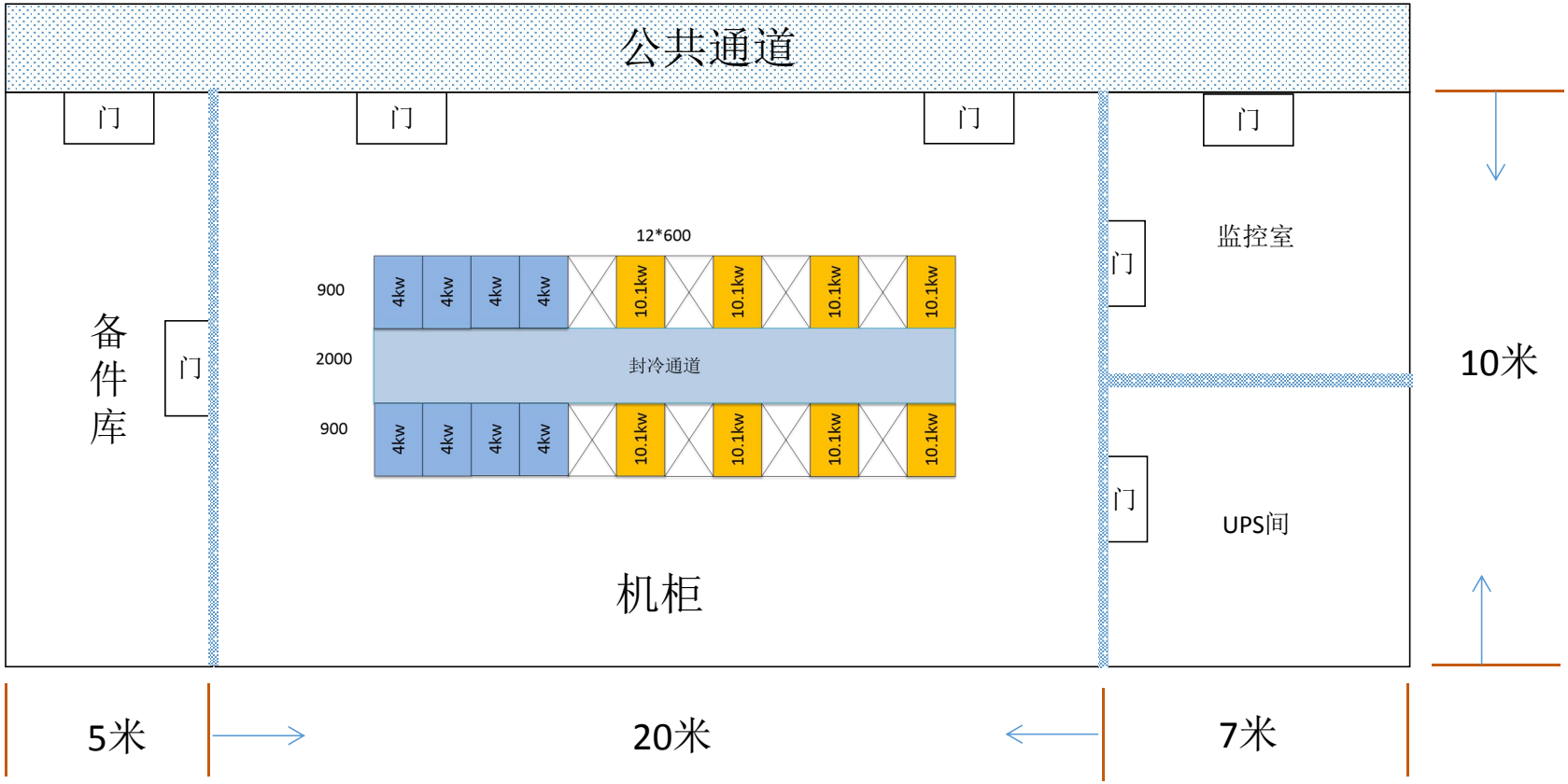
北京高能所
计算中心

分布式计算平台

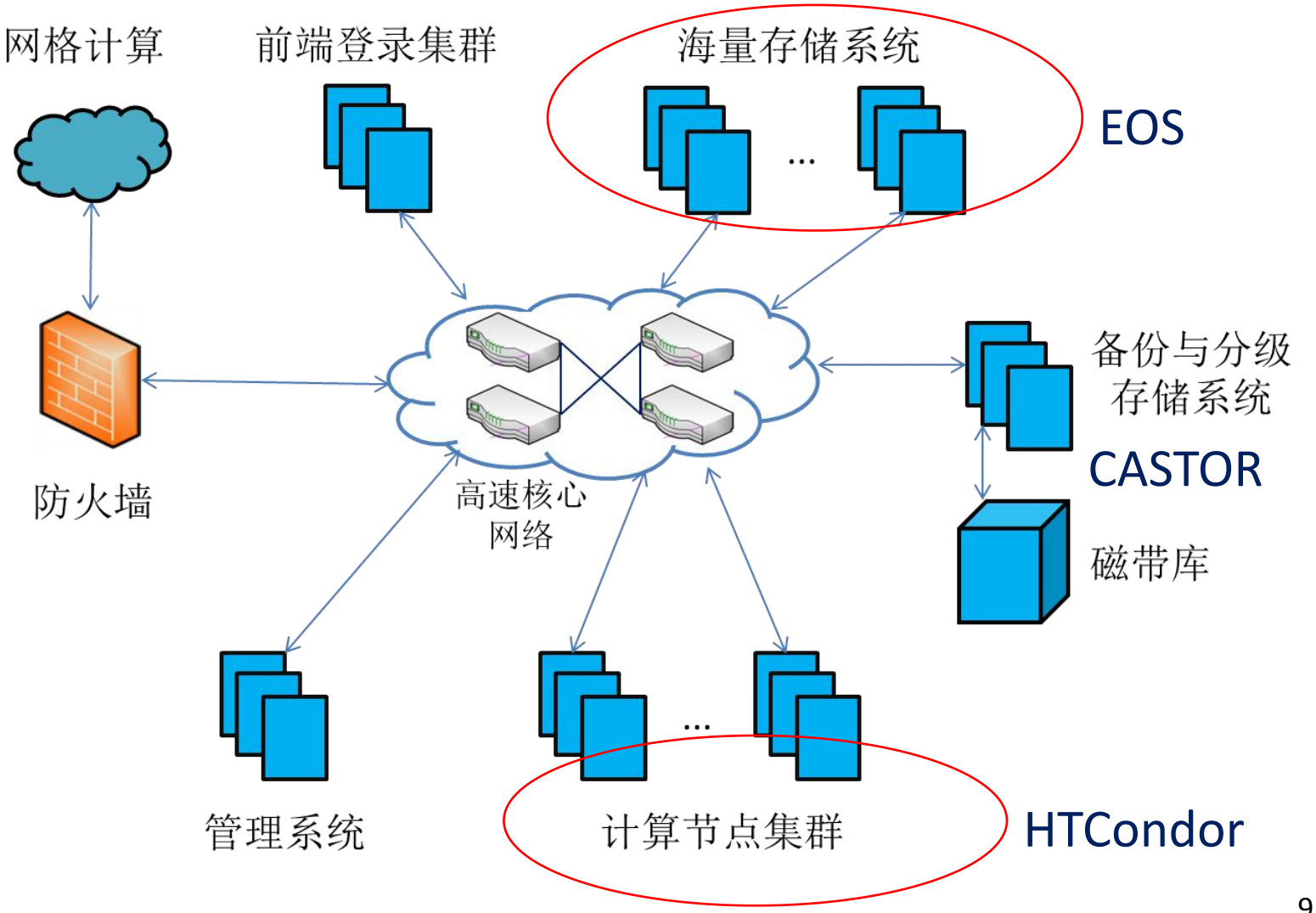
在站小型机房设计

- 满足DAQ以及快速重建等计算需求
- 规划面积：~300m²
- 16个机柜（8个高密，8个低密）
  - 必要时低密可以转化为高密
- 16箱刀片，>5000CPU核，300TB存储
  - **DAQ数据获取**：5箱IO刀片，共70~80个节点
  - **DAQ计算**： 5.5箱，2464~2816个CPU核，WCDA触发、直方图等。
  - **快速重建**： 5.5箱， 2464~2816个CPU核
- 总功率：~250KW

在站机房规划示意图


公共通道
门
门
门
门
备件库
门
12*600
900
2000
900
4kw
10.1kw
封冷通道
门
监控室
门
UPS间
机柜
10米
5米
20米
7米

离线数据处理平台架构


网格计算
前端登录集群
海量存储系统
EOS
防火墙
高速核心
网络
备份与分级
存储系统
CASTOR
磁带库
管理系统
计算节点集群
HTCondor

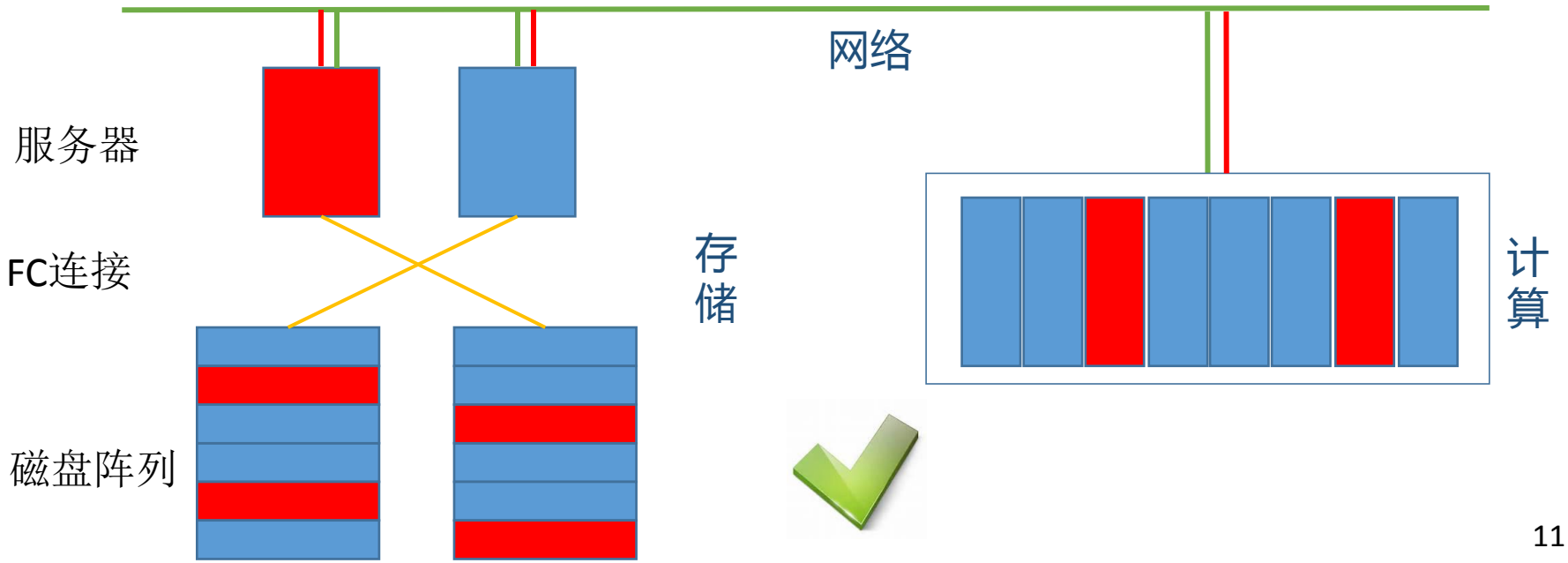
LHAASO数据处理平台的特点

- 异地取数，与BES等本地实验不同
- 数据量大，比大亚湾、羊八井等实验数据大10倍以上
- 异地机房建设、异地取数、异地处理、多数据中心、分布式计算
- 高海拔机房，环境苛刻
- 无人值守，运维要求高
- 高可靠，高可用

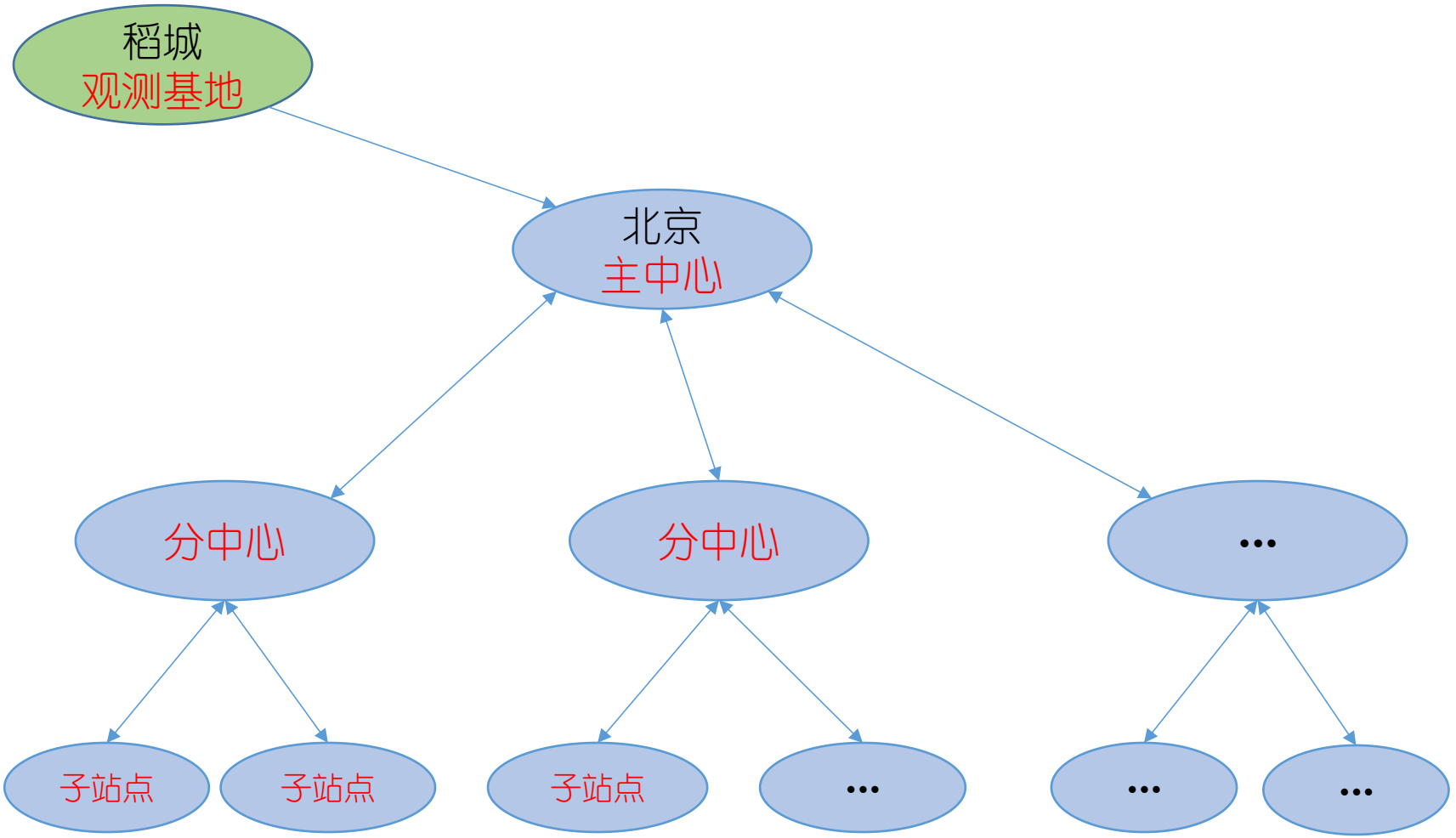
许多方案都是第一次采用，技术难度高，研发工作量大

系统可靠性设计

- 磁盘阵列：Raid6（10+2），12块中坏2块不影响
- 计算节点集群：任一计算节点损坏，作业重新调度，不影响使用
- 数据双副本：两个磁盘磁盘阵列中坏一个阵列不受影响
- 服务器双活：两台服务器中坏一台不受影响
- 网络冗余：单服务器，多网络连接，任一网卡损坏不影响网络连接
- 充足备件库，有问题及时更换；机柜摄像头，远程实时查看



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

主要进展

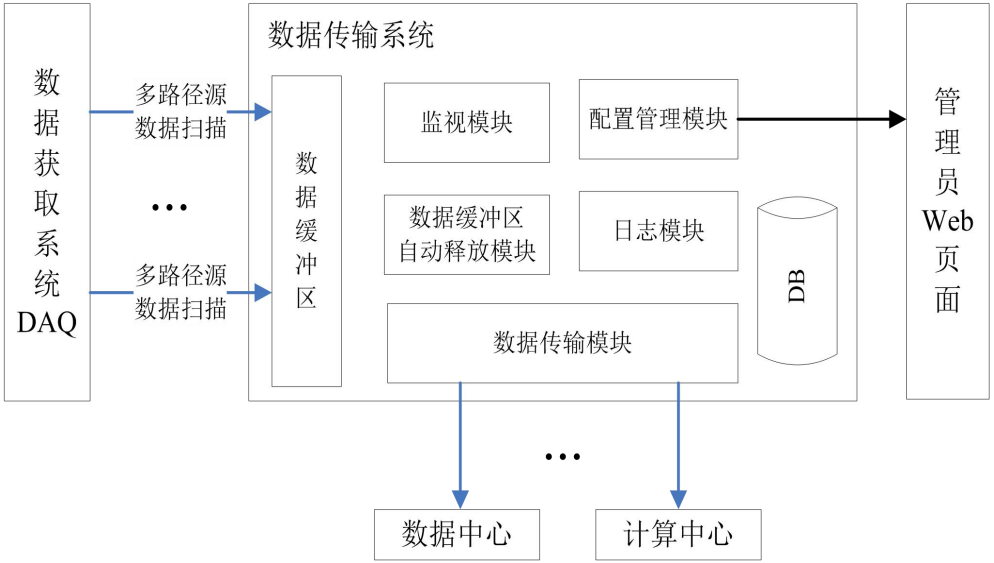
- 完成LHAASO数据数据处理平台部分的初设报告并提交
- 基本确定LHAASO观测基地机房建设的方案
  - 涉及多个部门，包括基建装修、通用供电、DAQ等
  - 计算中心任务：机柜、配电、消防、制冷、安防、新风排烟、监控等等
- 计算中心提供资源，建设LHAASO离线计算平台原型系统
  - 基于IHEPCloud云计算平台和HTCondor建设离线计算原型系统
  - 2016年提供了120万CPU小时的计算能力
  - 基于EOS建设了存储系统，提供253TB的空间，已经使用100TB
- 协助建设成都文献情报中心分站点
  - 提供120TB存储以及200CPU核
  - 使用高能所AFS认证，使用方式与所里一样
- 建设了120CPU核和150TB存储的Hadoop数据分析平台
  - 单独系统，存储不能共享

目前LHAASO计算环境

- 计算系统
  - 基于IHEPCloud云计算，全所实验共享
  - HTCondor: ~600 CPU core （共享）+ 768 CPU core（已到货）
- 存储系统
  - Gluster：347TB，剩余46TB
  - EOS：231TB，剩余160TB （共享）+ 300TB（已到货）
  - 未来以EOS为主，NFS、Gluster将逐步淘汰
- Hadoop平台
  - 84 CPU Core, 35TB；新增120 CPU Core，140TB（已使用50TB）
- 分布式计算系统
  - 成都情报文献中心，200CPUCore，120TB
- 2017年的需求？
  - 如果经费到位，计划购买1PB存储，1000CPU核

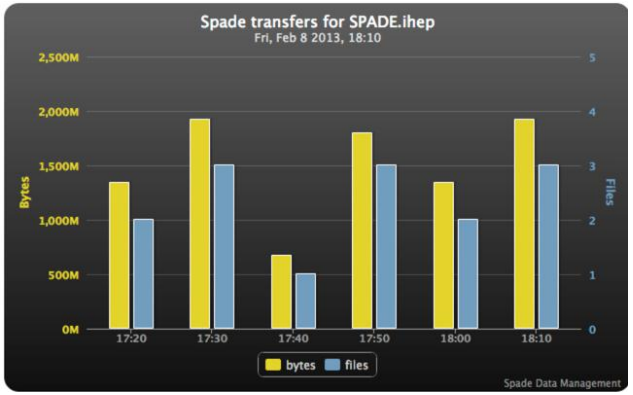
数据传输系统

- 负责将数据实时从观测基地传输到高能所以及向分站点进行数据分发
- 2017年争取在观测基地等站点部署，并与数据获取系统联调，与高能所进行数据传输测试
- 已经完成了数据传输系统的系统设计与开发
  - 支持一对多的数据传输，支持多种形式的数据传输（scp、gridftp等）
  - 更加通用的系统设置，提供接口与数据获取系统进行对接联调

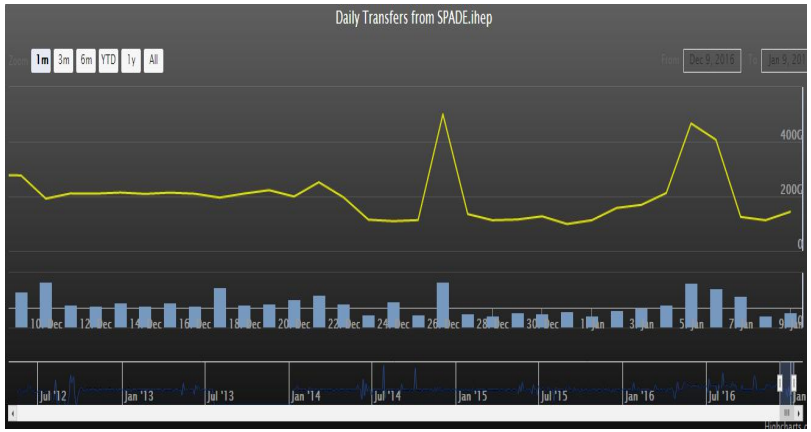


系统界面

- 传输文件大小和文件数量监视



- 历史信息监视
  - 统计信息
  - 单个文件传输过程历史信息



daq.Neutrino.0037445.Physics.EH1-Merged.SFO-1._0140

History for Ticket #2790555 (daq.Neutrino.0037445.Physics.EH1-Merged.SFO-1._0140)	
Ticketed	Fri, Feb 8 2013, 19:40
Compressed File	Fri, Feb 8 2013, 19:40
Wrapped File	Fri, Feb 8 2013, 19:40
Binary File	Fri, Feb 8 2013, 19:40
Metadata File	Fri, Feb 8 2013, 19:40
Placed in Warehouse	Fri, Feb 8 2013, 19:40
Packaged File	Fri, Feb 8 2013, 19:40
Dispatch Started	Fri, Feb 8 2013, 19:40
Dispatch Completed	Fri, Feb 8 2013, 19:41
Verification Started	Fri, Feb 8 2013, 19:41

未来工作安排

- 高海拔在站机房建设
  - 计划：2017.10-2018.5
- 现有系统运行与扩容
  - 计算、存储等，提供稳定服务
- 技术研究与应用
  - 实验数据传输系统 （2017）
  - 支持多级别数据冗余的高可靠磁盘存储系统（2017）
  - 新磁带类型LTO6/LTO7支持 （2017）
  - 基于广域网的远程文件研发
  - 分布式运维、资源整合与动态调度
  - 基于Hadoop的数据分析平台
  - GPU等高性能计算技术应用

人员安排

- 计算中心投入15人：程耀东作为联系人
- 机房建设：王新华
- 网络建设，包括广域网与两个园区网：齐法制
  - 齐孟尧 、 夏明山：观测基地与测控基地园区网
  - 孙智慧：广域网
  - 安德海：网络安全
- 数据传输：观测基地到高能所
  - 曾珊、齐孟尧
- 计算平台：石京燕
  - 黄秋兰，李海波，姚秋玲：存储系统，磁盘磁带等
  - 姜晓巍，胡庆宝，郑伟：计算系统与运维

小结

- 总体架构和方案基本确定，按照工程要求按序推进建设
- 已经建设了原型的数据处理平台
- 云计算等新技术的研究和应用正在展开
- 希望合作组的单位能够贡献IT方面的人力和资源，加入LHAASO分布式计算环境

谢谢

chyd@ihep.ac.cn