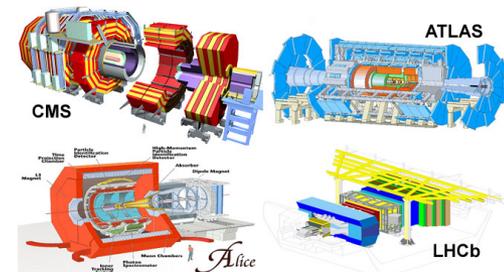


高能物理计算平台的 建设与运行

中科院高能物理所 计算中心
石京燕

高能物理计算

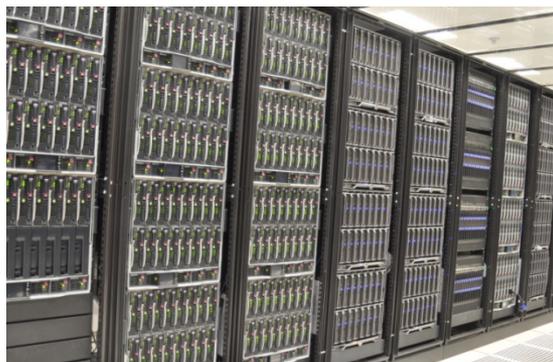
- BEPCII & BESIII
 - 5年累积数据5PB以上的数据
- 西藏YBJ宇宙线实验
 - 每年200TB的原始数据
- 大亚湾中微子实验
 - 累积数据：4PB数据
- 大型强子对撞机LHC
 - 每年15PB的数据
- 将来的项目：
 - 云南 Lhaasso: 预计1.2PB*10年
 - 江门中微子: 预计1PB*10年



高能物理计算特点与需求

- 特点：
 - 串行作业
 - 数据密集性计算
- 对数据访问与存储的基本需求
 - 数据快速访问：I/O性能
 - 数据长期保存：海量数据保存
 - 国际合作共享数据：数据传输

高能所计算环境结构



~10000 CPU cores



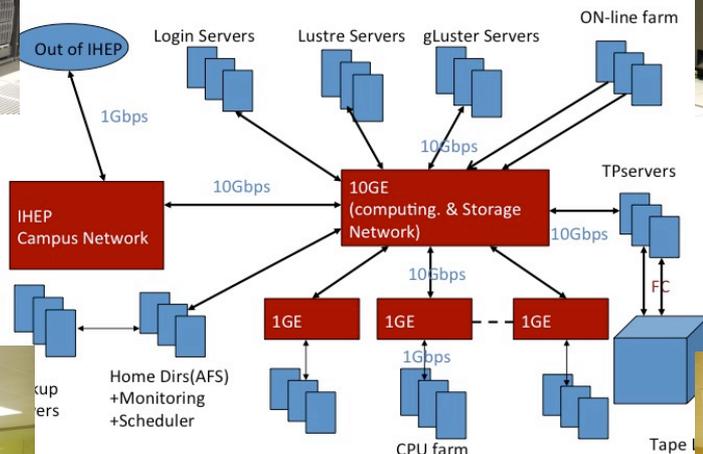
5PB tape library



>3PB disk space



Power supply, cooling

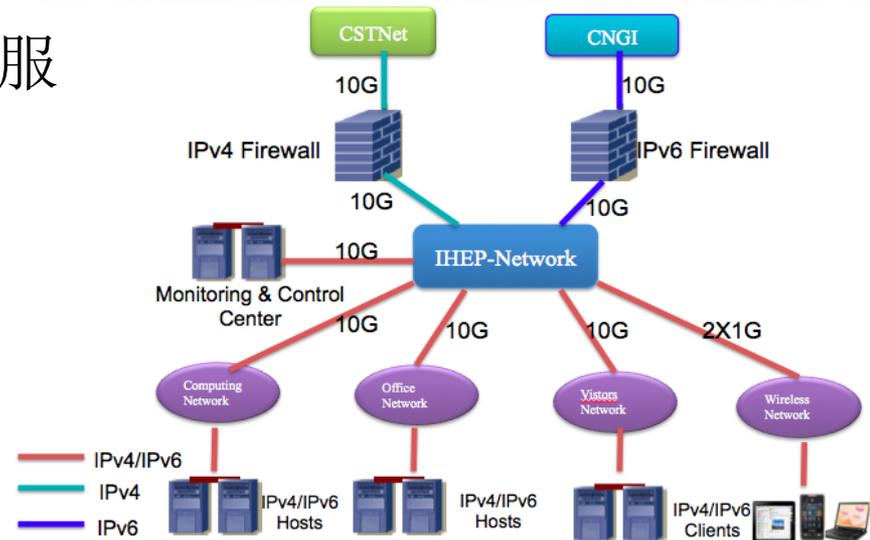
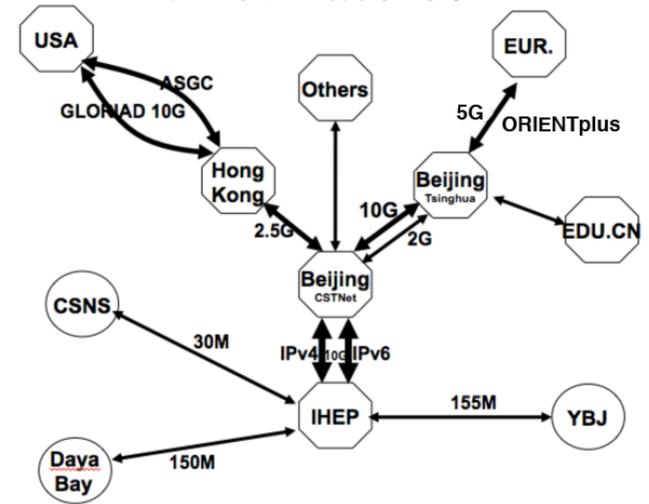


高性能计算环境

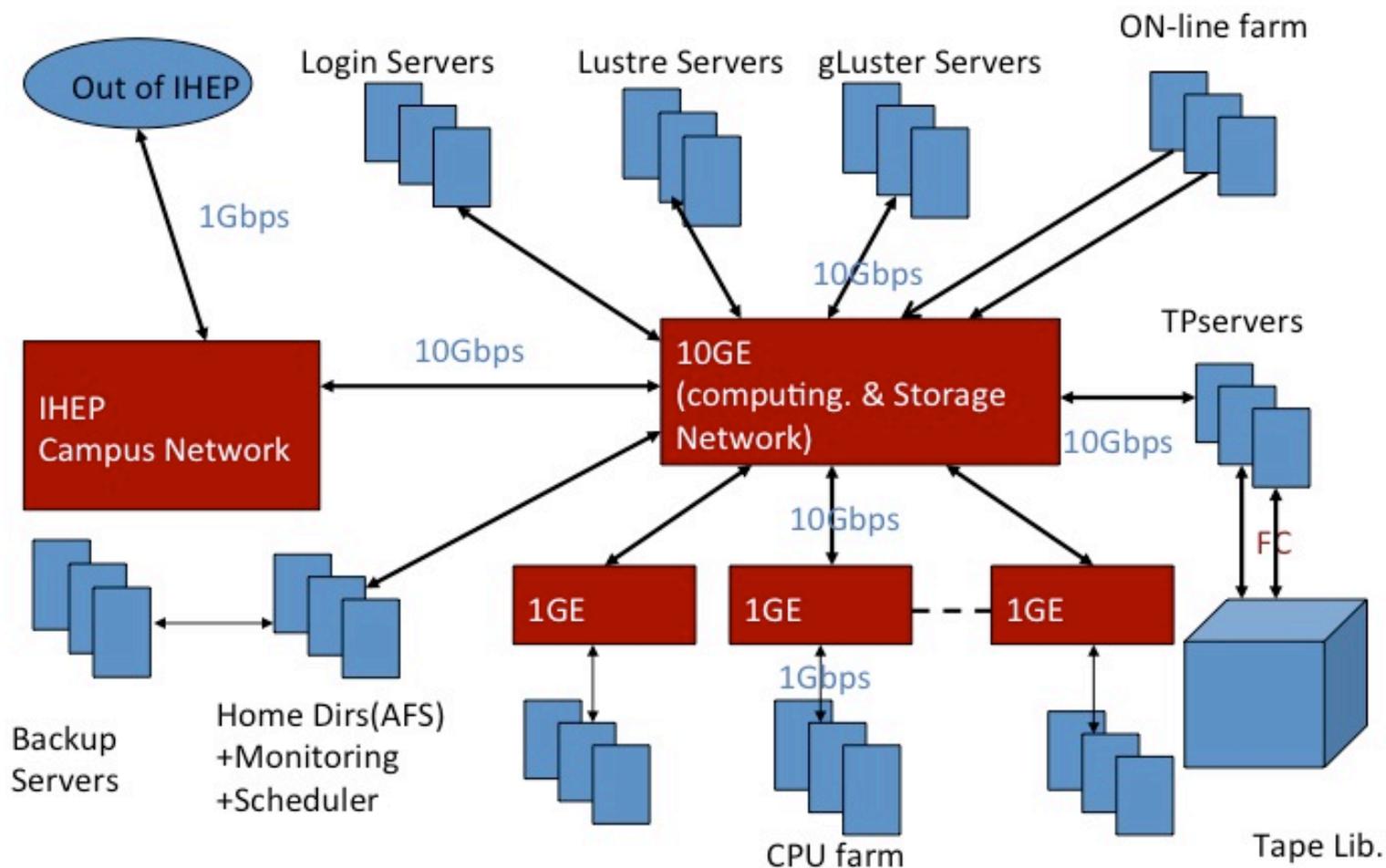
- 完全自主建设与集成（软件，硬件）
 - 高性价比的高密度刀片组成的集群结构，构成大规模批处理以及MPI计算系统，支持高能物理，生物科学等多个学科的高性能计算
 - 服务器+阵列的磁盘存储
 - 大型带库
- 二部分组成：
 - 大型本地集群环境
 - 网格站点
- 开发部署一系列管理软件
 - 操作系统和应用软件的自动安装与部署
 - 计算环境运行状态细粒度监控
 - 计算资源使用情况的统计与记账
 - 用户信息，设备信息管理
 - 存储访问模式分析
 - 运行维护日志记录
 - 。 。 。 。

高速网络

- 科学院最大校园网，最高出口带宽
 - 10G backbone
 - IPv4/IPv6 双栈
 - 校园网无线全覆盖
- 约500个网络服务器及设备
- 提供邮件，网络，协同工作等服务
- >3000 用户
- 每年约4 PB的数据传输

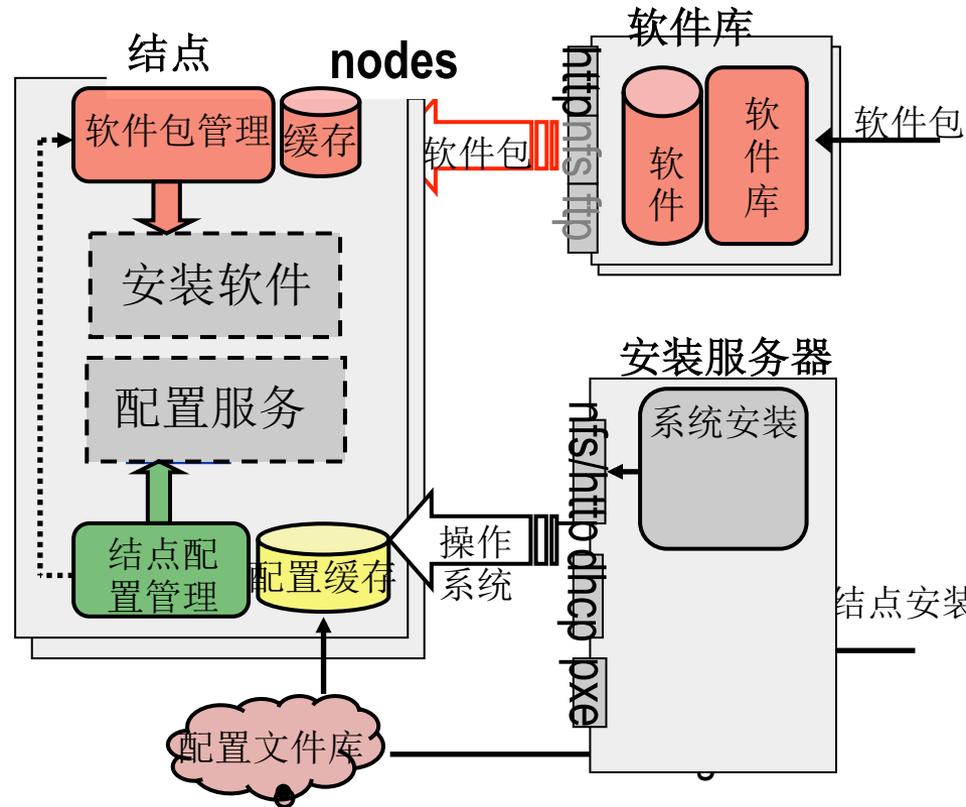


体系结构



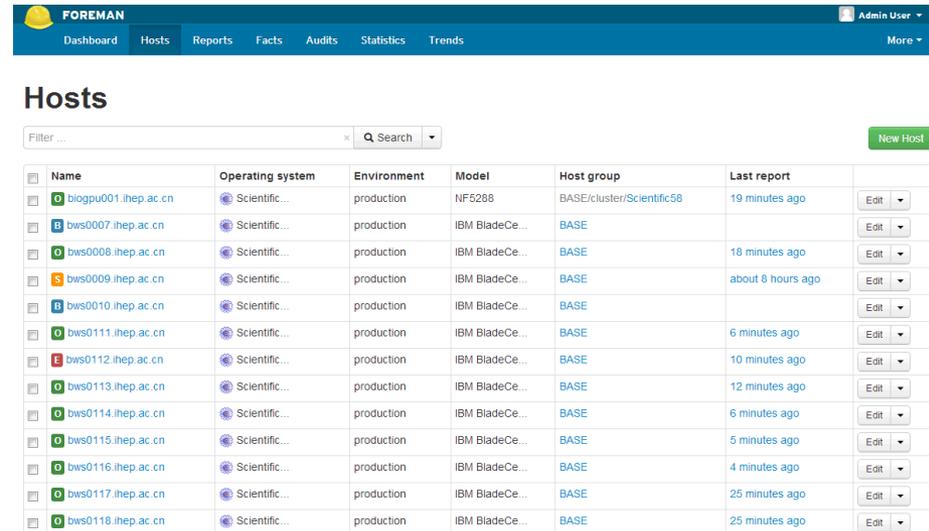
自动安装与部署-- quattor

- 自动为结点服务器 安装/卸载/升级 操作系统、应用软件、网络中间件等软件包
 - 自动配置结点的服务
 - 结点配置文件
 - 配置信息用XML文件进行描述
 - 每一个结点对应一个配置文件
- 配置文件的内容: 机群信息、系统、软件、硬件
- 一小时完成几百台机器的系统及应用软件安装与配置



Quattor → Puppet

- 使用模式的不同
 - Quattor: 发布
 - Puppet: agent自动更新
- 使用更加简单:
 - Dashboard: 便于操作
- 虚拟化管理的需求
- 开源的活跃社区
- 性能的需求



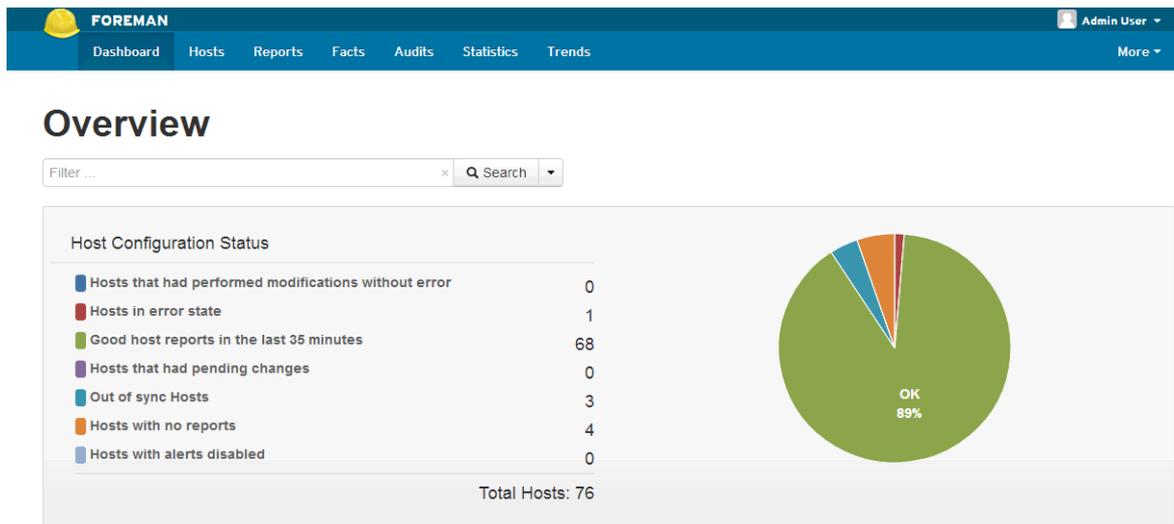
FOREMAN Admin User

Dashboard Hosts Reports Facts Audits Statistics Trends

Hosts

Filter ... Search

Name	Operating system	Environment	Model	Host group	Last report	
bioqpu001.ihep.ac.cn	Scientific...	production	NF5288	BASE/cluster/Scientific58	19 minutes ago	Edit
bws0007.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE		Edit
bws0008.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	18 minutes ago	Edit
bws0009.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	about 8 hours ago	Edit
bws0010.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE		Edit
bws0111.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	6 minutes ago	Edit
bws0112.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	10 minutes ago	Edit
bws0113.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	12 minutes ago	Edit
bws0114.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	6 minutes ago	Edit
bws0115.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	5 minutes ago	Edit
bws0116.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	4 minutes ago	Edit
bws0117.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	25 minutes ago	Edit
bws0118.ihep.ac.cn	Scientific...	production	IBM BladeCe...	BASE	25 minutes ago	Edit



FOREMAN Admin User

Dashboard Hosts Reports Facts Audits Statistics Trends

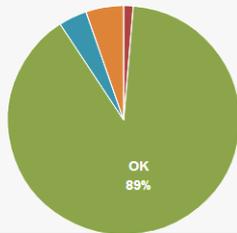
Overview

Filter ... Search

Host Configuration Status

Hosts that had performed modifications without error	0
Hosts in error state	1
Good host reports in the last 35 minutes	68
Hosts that had pending changes	0
Out of sync Hosts	3
Hosts with no reports	4
Hosts with alerts disabled	0

Total Hosts: 76



OK 89%

监视管理

- Nagios → Icinga

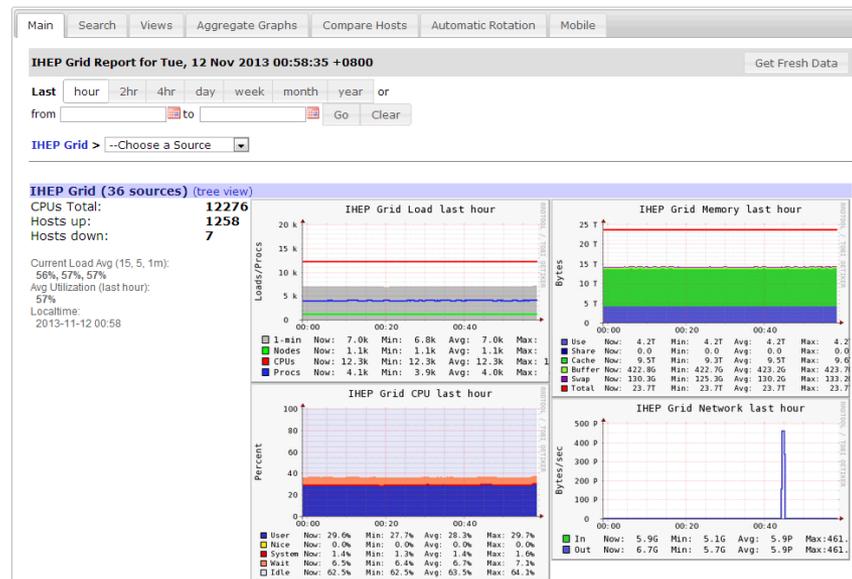
- 性能优化与二次开发

- 机器分类，服务分组
 - 推拉结合的监视程序
 - 根据实际需求的监视脚本的编写
 - 与批作业的结合
 - 与日志系统的结合
 - 访问权限设定与统一认证

- Ganglia

- 计算环境当前及历史状态的记录

DYB计算节点负责人, 系统组值班人员 (DYB-Servers)	41 运行	492 正常
数据服务器负责人, 杜国红, 杨毅 (Data-Servers)	12 运行	12 正常
GPU负责人, 文硕, 频6067 (GPU-Servers)	29 运行	319 正常
	1宕机 1已确认	11严重 11主机上的故障
存储: 姚秋玲, 郑伟 (GRASS-Servers)	5 运行	10 正常
负责人, 系统组值班人员 (Gluster-Servers)	4 运行	24 正常
Gridtb计算节点负责人, 系统组值班人员 (Gridtb-Servers)	8 运行	72 正常
HTTPServers (HTTP-Servers)	1 运行	8 正常
登录节点负责人, 杜国红, 杨毅 (Login-Servers)	20 运行	194 正常
	1宕机 1已确认	9严重 9主机上的故障
Lustre-server: 汪璐 (Lustre-servers)	48 运行	144 正常
MBH计算节点负责人, 系统组值班人员 (Mbh-Servers)	32 运行	384 正常
SNS节点负责人, 闫晓飞 (Sns-Servers)	14 运行	70 正常
YBJ计算节点负责人, 系统组值班人员 (Ybj-Servers)	75 运行	900 正常



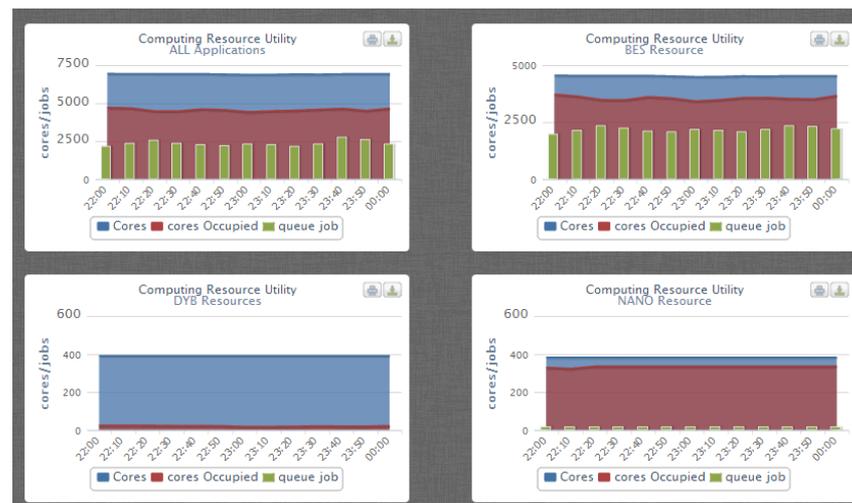
大型本地集群

- 提供串行作业，MPI作业，GPU作业计算服务
- 当前
 - 7000+ 批处理作业槽
 - 支持串行作业与mpi作业
 - 250+ gpu作业槽
 - 1000+用户，活跃用户300个
- 今后：将增加到10,000 至20,000作业槽

作业管理

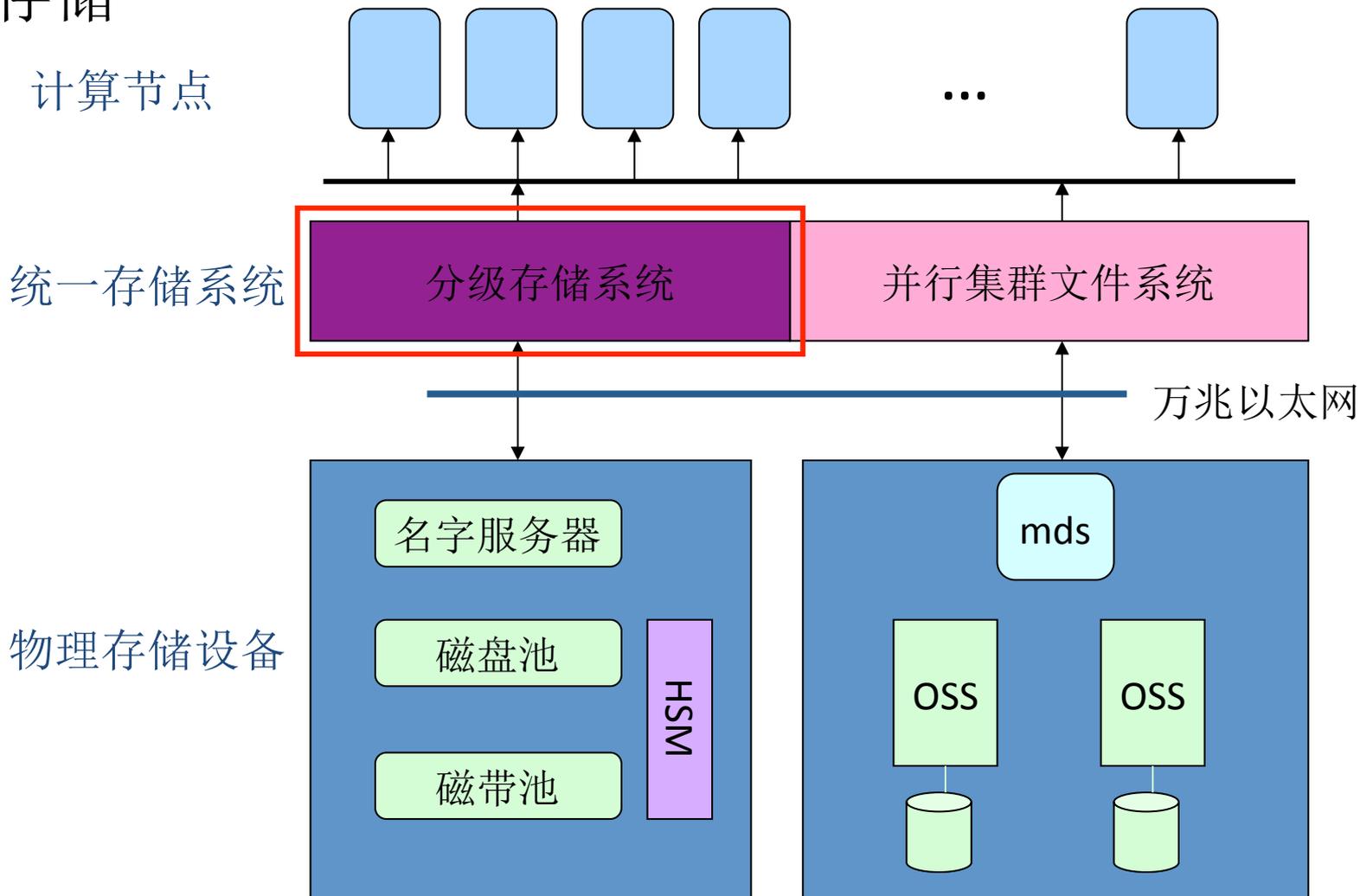
- torque + Maui
 - 开源软件
 - 50+队列
 - 性能监视
 - 记账功能
 - Gpu作业调度
 - 调度性能调优
 - 增加短队列
 - 动态调整运行参数

- 性能瓶颈
 - 作业量增加导致健壮性，作业调度性能下降
 - 新的调度规划



高性能存储

- 采用高性能的并行集群文件系统，磁盘，带库分级存储



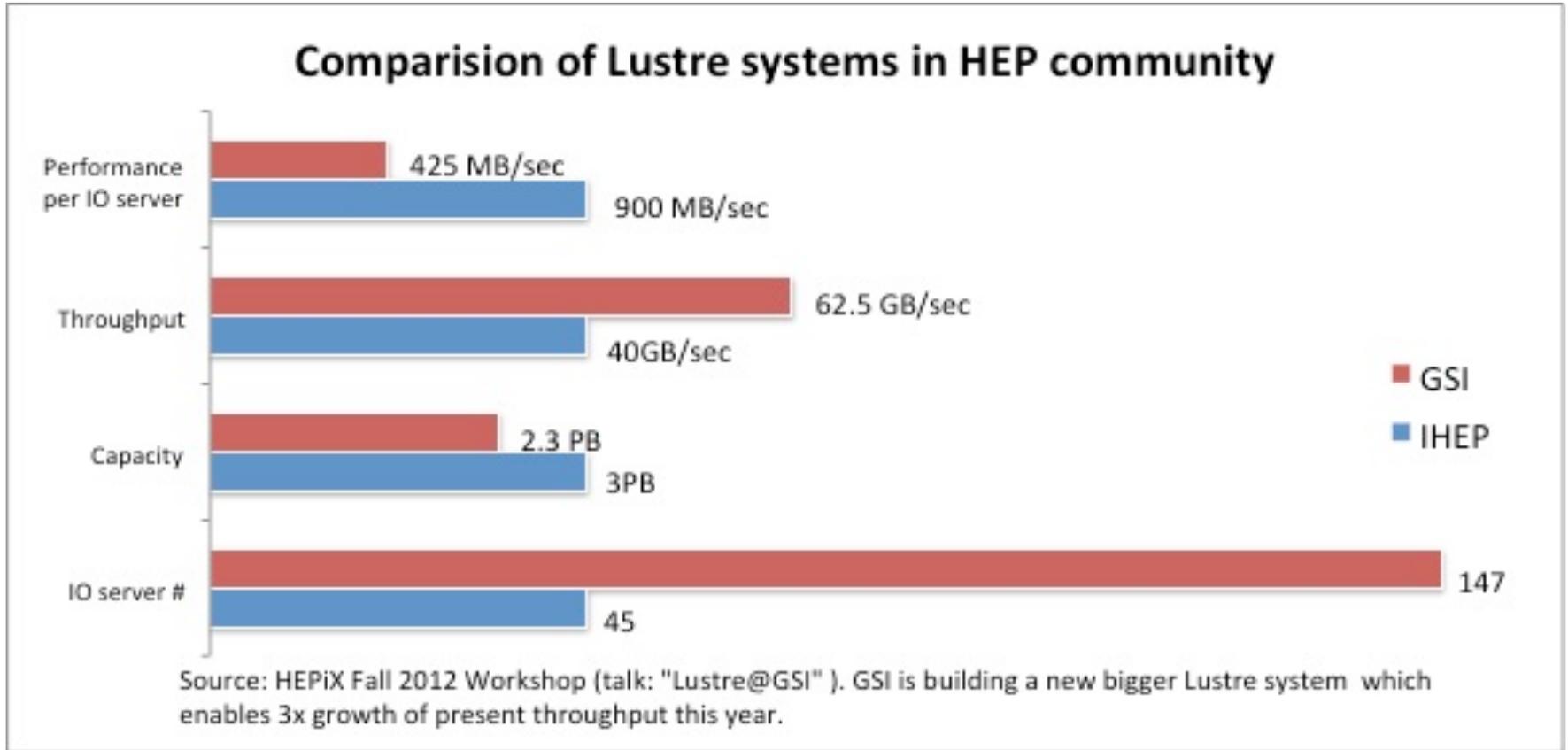
分级存储

- 根据高能物理计算的需求，自主开发
- 自动管理磁盘、磁带等不同的存储介质，多个磁带被虚拟成磁带池，多个磁盘被虚拟成磁盘池
- 最后，通过名字服务器，将这些存储设备虚拟成统一的文件命名空间。用户不用关心数据存放在什么存储设备，只需要按照统一的文件名来访问即可
- 根据文件访问频度等因素，自动在磁盘与磁带之间迁移数据，在保证海量数据存储的同时，也实现较高的数据访问性能，即分级存储管理HSM
- 目前状况
 - 26个LTO4磁带驱动器
 - ~1500个槽位，可以扩充到5PB的磁带空间
 - 10个磁带服务器，10个磁盘服务器，120TB磁盘缓存

分布式存储

- Lustre, gLuster
 - 针对不同的使用需求
 - 大规模分布式存储系统
 - 性能调优与二次开发
 - 划分多个不同用途的存储区域
- 3PB+的磁盘存储
- 实际存储：上亿个数据文件
- 聚合带宽达到41GB/sec

与国外同类存储比较



用户及设备管理

- 自主开发
 - 用户管理系统
 - 设备及备件管理
 - 运行日志管理



设备

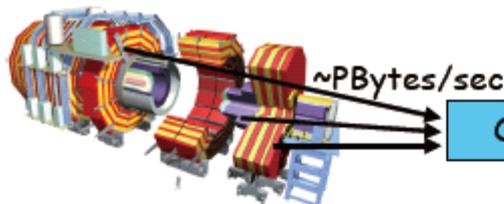
设备名 应用

序号	设备名	IP	应用	类别	排 列	高 度	负责人	状态
1	sched02.ihep.ac.cn	192.168.61.254	TEST	test服务器	1 2 37	2	阎博文	1
2	sched01.ihep.ac.cn	192.168.61.253	TEST	test服务器	1 2 35	2	阎博文	1
3	xas04.ihep.ac.cn	192.168.61.98	OTHER	集群计算节点	4 1 27	2	闫小飞/赵海峰	1
4	xas03.ihep.ac.cn	192.168.61.97	OTHER	集群计算节点	4 1 27	2	闫小飞/赵海峰	1
5	xas02.ihep.ac.cn	192.168.61.96	OTHER	集群计算节点	4 1 27	2	闫小飞/赵海峰	1
6	hzmt062.ihep.ac.cn	192.168.61.222	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
7	hzmt061.ihep.ac.cn	192.168.61.221	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
8	hzmt060.ihep.ac.cn	192.168.61.220	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
9	hzmt059.ihep.ac.cn	192.168.61.219	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
10	hzmt058.ihep.ac.cn	192.168.61.218	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
11	hzmt057.ihep.ac.cn	192.168.61.217	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
12	hzmt056.ihep.ac.cn	192.168.61.216	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1
13	hzmt055.ihep.ac.cn	192.168.61.215	HZMT	集群计算节点	3 4 3 10		yanxf/慕建胤	1

序号	设备名	操作日期	出错硬件	现象	操作人	基本操作
1	cmspn02-1-2	2013-11-11	电源	左侧电源坏	杜国红	编辑
2	cmspn03-2-1	2013-11-11	硬盘	7# 1TB硬盘坏	杜国红	编辑
3	bws0503.ihep.ac.cn	2013-11-11	其它	Server Blade Enclosure Inadequate Power To Power On: Not Enough Power (Enclosure Address 0000, Slot 16)	杨毅	编辑
4	gpul11.ihep.ac.cn	2013-11-11	电源	加不上电	杨毅	编辑
5	cmspn03-2-2	2013-11-11	硬盘	5# 1TB硬盘坏	杜国红	编辑
6	cmspn03-1-1	2013-11-11	硬盘	2# 1TB硬盘坏	杜国红	编辑
7	com05-1-2	2013-11-11	硬盘	10# 1TB硬盘坏	杜国红	编辑
8	com05-1-1	2013-11-11	硬盘	8# 1TB 硬盘坏	杜国红	编辑
9	bws0576.ihep.ac.cn	2013-11-07	其它	cpu i报错, 报修	杜国红	编辑
10	sns024.ihep.ac.cn	2013-11-06	其它	磁盘报错, 只读, 重启	杜国红	编辑
11	dybfs02-1	2013-11-06	硬盘	21# 3TB 磁盘坏, 24#rebuilding 12:00	杜国红	编辑
12	boss39-2	2013-11-06	硬盘	16# 3TB磁盘坏14#rebuilding 11-05 15:41	杜国红	编辑

网格计算

- 网格计算 —— Cern:
 - 1999年：开始研究分布式网格计算，用于分析LHC的实验数据。
 - 2001年： DataGrid Project，开发网格中间件，用于连接应用软件和硬件资源，提供高能物理，生物信息等学科计算资源
 - 2004年： Enabling Grid for E-science (EGEE) 提供世界范围内的科研人员不间断计算服务。每天完成2,000,000个计算作业。
 - 2010年： European Grid Infrastructure (EGI) 支持欧洲网格服务和基础设施的长期发展计划，扩展与其它网格项目的合作。



Online System ~100 MBytes/sec

1 TIPS = 25,000 SpecInt95
 PC (today) = 10-15 SpecInt95

Bunch crossing per 25 nsecs.
 100 triggers per second
 Event is ~1 MByte in size

Offline Farm ~20 TIPS

~100 MBytes/sec

~622 Mbits/sec **Tier 0**
 or Air Freight

CERN Computer Center



Tier 1

France Regional Center

Germany Regional Center

Italy Regional Center

Fermilab ~4 TIPS



~2.4 Gbits/sec

Tier 2

Tier2 Center ~1 TIPS

Tier2 Center ~1 TIPS

Center TIPS

Center TIPS

Center TIPS

Tier 3

~622 Mbits/sec

Institute ~0.25TIPS

Institute

Institute

Institute

Physics data cache

100 - 1000 Mbits/sec

Workstations

Tier 4

Physicists work on analysis "channels".
 Each institute has ~10 physicists working on one or more channels
 Data for these channels should be cached by the institute server

北京网格站点

- 支持ATLAS, CMS二个高能物理实验



- 1000+作业槽, 400TB存储
- 5Gbps网络连接欧洲, 2.5Gbps网络连接美国
- 每年提供10,000,000 cpu小时计算时间, 完成5,000,000网格计算作业
- 高可靠性与有效性

Tier-2 Availability and Reliability Report
OPS
June 2013

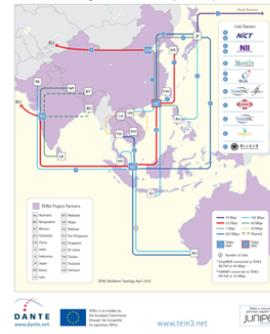
Federation Summary - Sorted by Reliability

Colour coding: N/A <30% <60% <90% >=90%

Availability Algorithm: (OSG-CE + CREAM-CE + KRC-CE)* (USG-SRM v2 + USG-Beaman/Krood + SRM + SRM v2)* Site-BDI

Federation	Availability	Reliability	Federation	Availability	Reliability
CN-IHEP	100%	100%	JP-Tokyo-ATLAS-T2	97%	97%
DE-DESY-LHCB	100%	100%	AT-HEPHY-VIENNA-UIBK	97%	97%
FHHP-T2	100%	100%	TR-Tier2-federation	97%	97%
FR-IN2P3-SUBATECH	100%	100%	US-AGLT2	97%	97%
HU-HGCC-T2	100%	100%	KR-KNU-T2	97%	97%
SE-SNIC-T2	100%	100%	PT-LIP-LCG-Tier2	97%	98%
T2_US_Nebraska	100%	100%	DE-DESY-GOE-ATLAS-T2	97%	98%
TW-FIT-T2	100%	100%	IL-HEPTier-2	97%	97%
T2_US_UCSD	100%	100%	PK-CMS-T2	97%	97%
DE-DESY-ATLAS-T2	100%	100%	EE-NICPB	96%	97%
FR-IN2P3-LPC	100%	100%	UK-LondonTier2	96%	98%
FR-IN2P3-LAPP	100%	100%	FR-IN2P3-CPPM	96%	97%
CH-CHIPP-CSCS	100%	100%	CA-WEST-T2	96%	100%
NO-NORGRID-T2	100%	100%	ES-ATLAS-T2	96%	97%
SI-SIGNET	100%	100%	T2_US_Purdue	95%	95%
US-SWT2	100%	100%	UK-SouthGrid	95%	95%
T2_US_MIT	100%	100%	SK-Tier2-Federation	94%	98%
FR-IN2P3-IPHC	100%	100%	FR-IN2P3-CC-T2	92%	100%
US-MWT2	100%	100%	FR-GRIF	92%	100%
T2_US_Florida	100%	100%	ES-LHCb-T2	92%	92%
CZ-Prague-T2	100%	100%	DE-MCAT	90%	90%
KR-KISTI-T2	99%	99%	US-LBNL-ALICE	90%	90%
T2_US_Wisconsin	99%	99%	RU-RDIG	90%	98%
BR-SP-SPRACE	99%	99%	BE-TIER2	88%	88%
UK-NorthGrid	99%	99%	RO-LCG	87%	88%
AJ-ATLAS	99%	100%	GR-Ioannina-HEP	86%	92%
DE-DESY-RWTH-CMS-T2	99%	99%	T2_US_Gattech	85%	85%
DE-FREIBURGWUPPERTAL	98%	98%	UA-Tier2-Federation	82%	85%
UK-ScotGrid	98%	98%	US-NET2	79%	79%
IT-INFN-T2	98%	98%	IN-DAE-KOLKATA-TIER2	77%	77%
ES-CMS-T2	98%	98%	IN-INDIACMS-TIFR	68%	68%
2-WLGG	98%	98%	CA-EAST-T2	61%	77%
3-LPSC	98%	98%	US-LBNL-ALICE	4%	100%
			DE-GSI	N/A	N/A

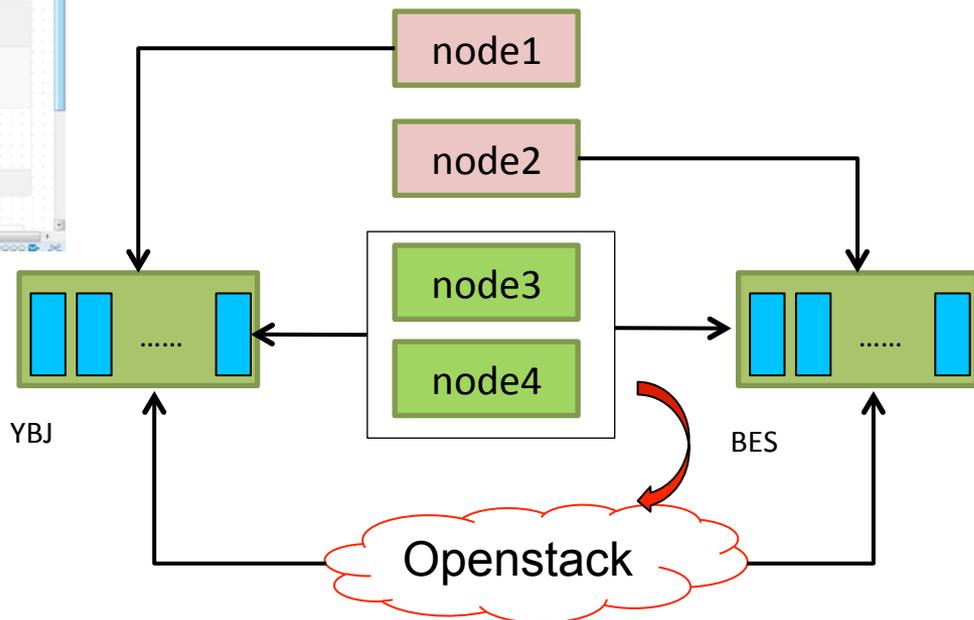
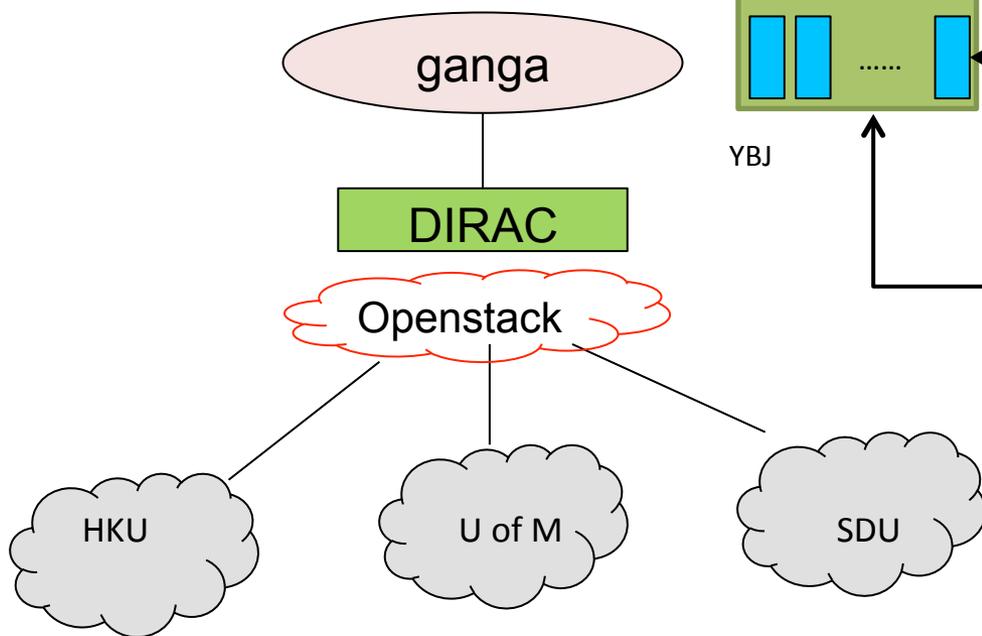
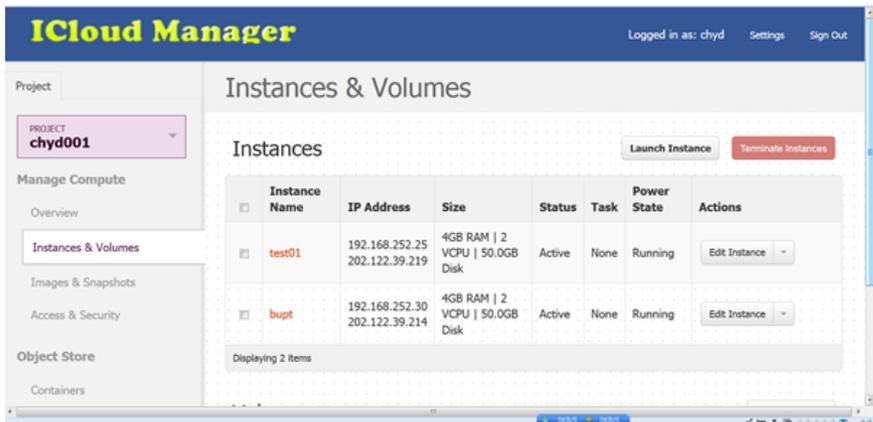
TEIN3
The research and education network for Asia-Pacific
Linking Asia-Pacific to Europe and beyond



云计算

- 云计算：借助虚拟化技术、分布式海量存储技术等动态创建高度虚拟化的IT资源池，通过网络提供服务
- 高能物理对云计算有强烈的需求
 - 将OS与APP打包发布，提供轻量级的虚拟应用程序
 - 将计算节点虚拟化，提高资源利用率
- 开源云中间件：OpenStack, OpenNebula
 - 技术发展趋势
 - 活跃社区，广泛支持
- 研究进行中
 - 节点虚拟化管理、应用虚拟化封装等，即可提供大规模的云计算服务，主要包括虚拟机（IaaS）、虚拟集群/网格（PaaS）以及云存储（SaaS）服务

IHEP Openstack



面临的困难

- 硬件老化
 - 2 / 3的硬件已经超过保修期，仍在服役
 - 硬件维护任务艰巨，工作量增大
- 应用不断提出新的需求
 - 存储性能
 - 作业调度性能
- 开源软件的使用健壮性难以保证
 - 需求牵引，探索前进

谢谢！

问题？