Status of IHEP Site

Yaodong CHENG
Computing Center, IHEP, CAS
2016 Fall HEPiX Workshop

Outline

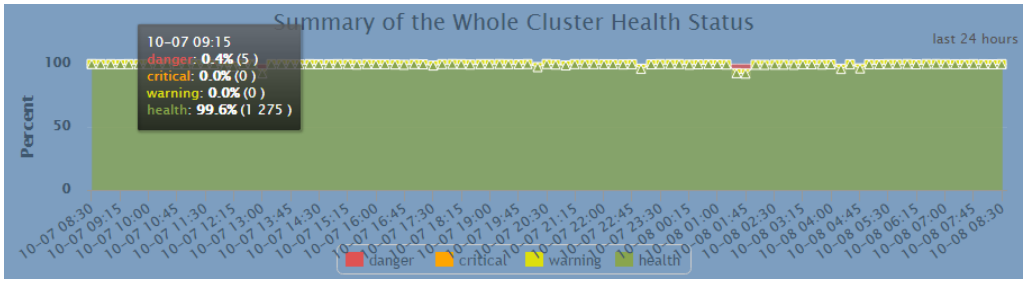
- Local computing cluster
- Grid Tier2 site
- IHEP Cloud
- Internet and domestic Network
- Next Plan

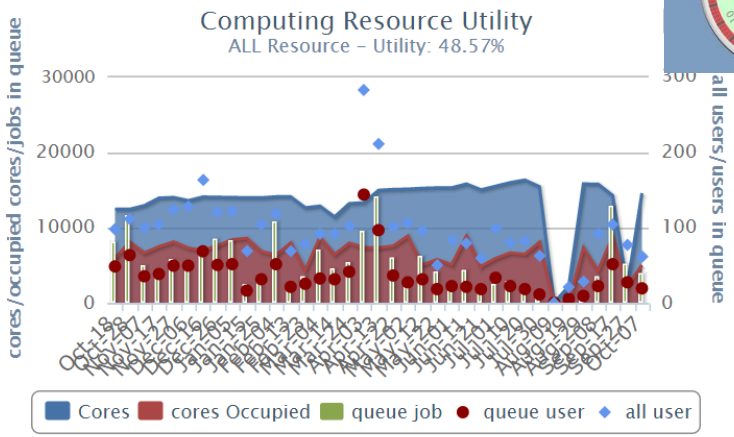
Local computing cluster

- Support BESIII, Da-ya Bay, JUNO, astrophysics experiments ……
- Computing
  - ~13,500 CPU cores, 300 GPU cards
  - Mainly managed by Torque/Maui
  - 1/6 has been migrated to HTCondor
  - HTCondor will replace Torque/Maui this year
- Storage
  - 5PB LTO4 tapes managed by CASTOR 1
  - 5.7 PB of Lustre. Another 2.5 PB will be added this year
  - 734 TB of gLuster with replica feature
  - 400TB of EOS
  - 1.2 PB of other disk spaces

Local computing cluster (2)

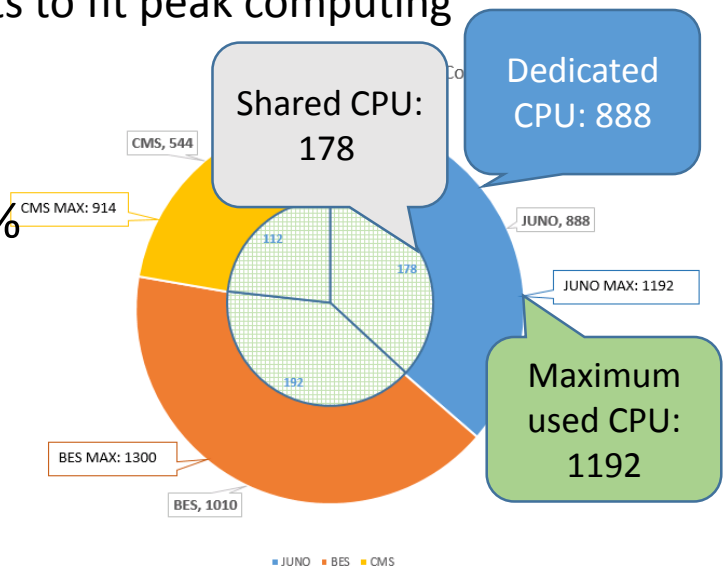
- Automatic deployment: Puppet + Foreman
- Monitor:
  - Icinga (Nagios)
  - Ganglia
- Flume+ElasticSearch
- Annual resource utilization rate: ~48%





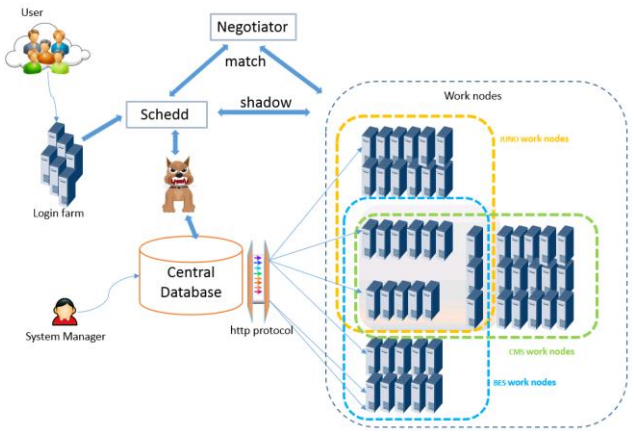
HTCondor Cluster (1/2)

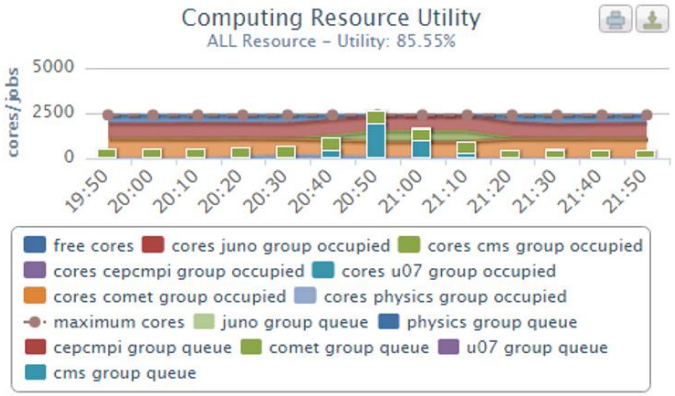
- 153 blade servers with 2500 CPU cores
- New share policy to support 3 experiments: BES, JUNO, CMS
  - CPU cores are contributed separately by different experiments
  - Each experiment has its dedicated CPU cores and contributes some of CPU cores to one shared pool
  - Dedicated CPU cores run jobs belonging to owner's experiment
  - Shared CPU cores run jobs from all experiments to fit peak computing requirement
- Resource utility has been promoted a lot: ~80%
- Scale to 6000 CPU cores in October



HTCondor Cluster (2/2)

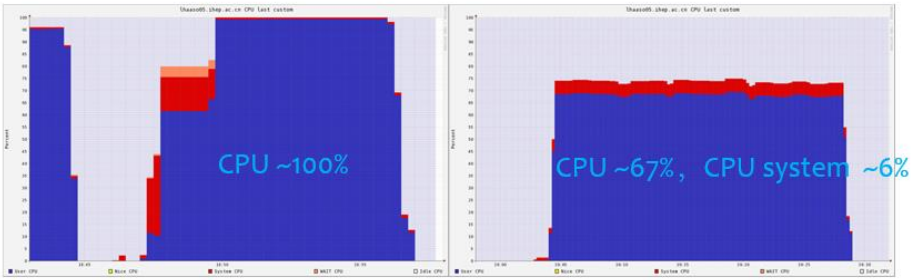
- Central Database
  - Store device information
  - the only place where scheduler, monitoring , puppet get their configuration info
  - Work nodes info includes supported experiments, user priority, current status etc.
- Worker node info published via the Apache server
  - Work nodes info is published in real time (<5m)
  - HTCondor machines update their configuration via crond
  - Cluster will still work properly even if both database and Apache crash
- Failure nodes detected by the monitor system would be excluded automatically



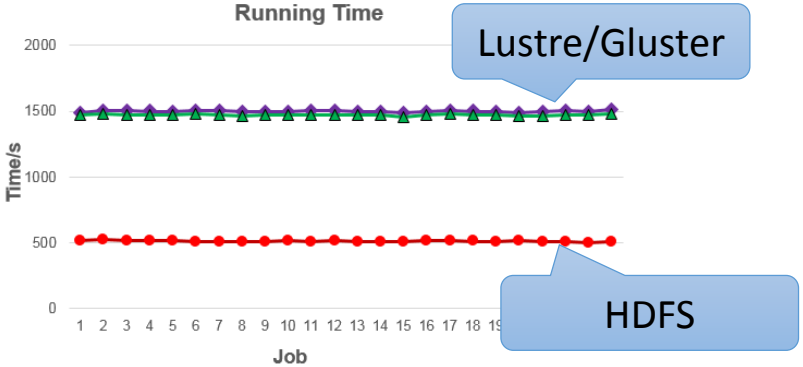


Hadoop data analysis platform

- Built for more than one year
  - Old system: 84 CPU Cores, 35TB disk
  - New system: 120 CPU Cores, 150TB disk
  - Support BESIII, cosmic ray experiments
- Test results
  - Cosmic ray experiment application: medea++ 3.10.02
  - CPU utilization on Gluster much lower than HDFS
  - HDFS job running time is about one-third of Gluster/Lustre FS

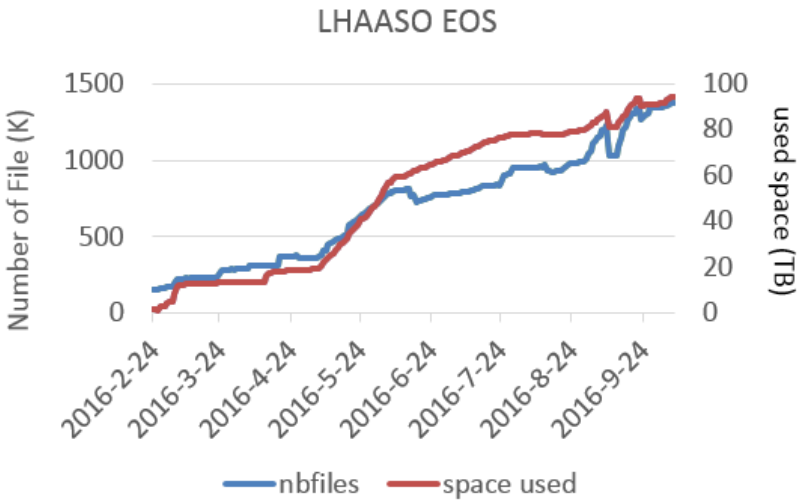


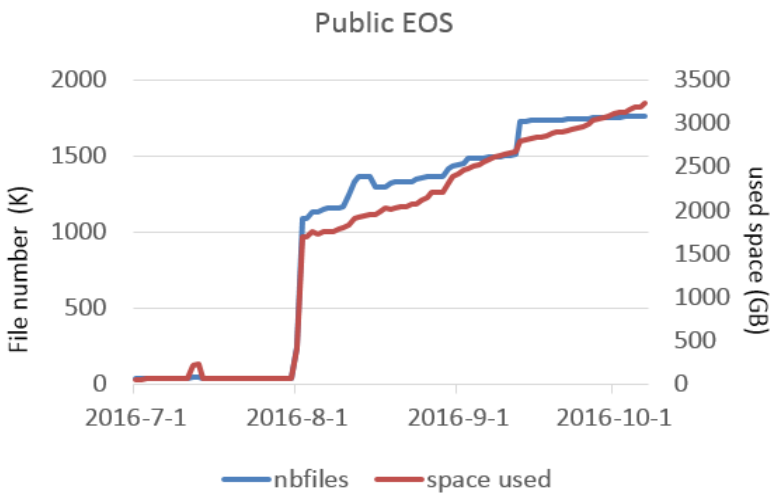
HDFS Gluster



EOS deployment

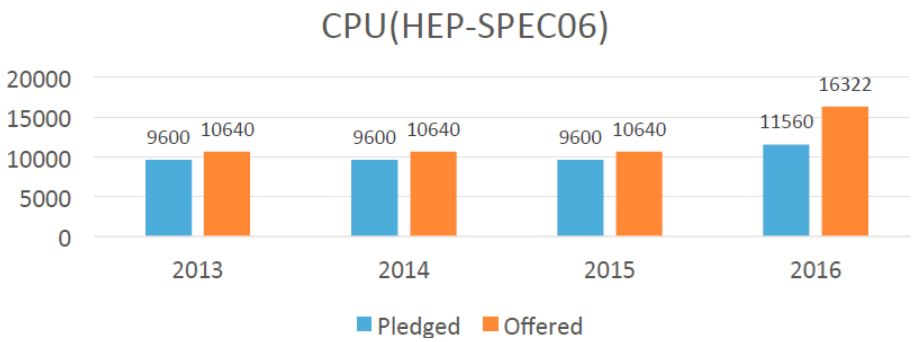
- Two instances
- One for batch computing (LHAASO experiment)
  - 3 servers with 86 X 4T SATA disks
  - 3 dell disk array box (raid6)
  - Each server has 10Gb network link
- One for user and public services
  - IHEPBox based on Owncloud: user file synchronization
  - Web server and other public services
  - 4 servers
  - Each server with 12 disks and 1Gb link
  - Replication mode
- Plan to support more applications

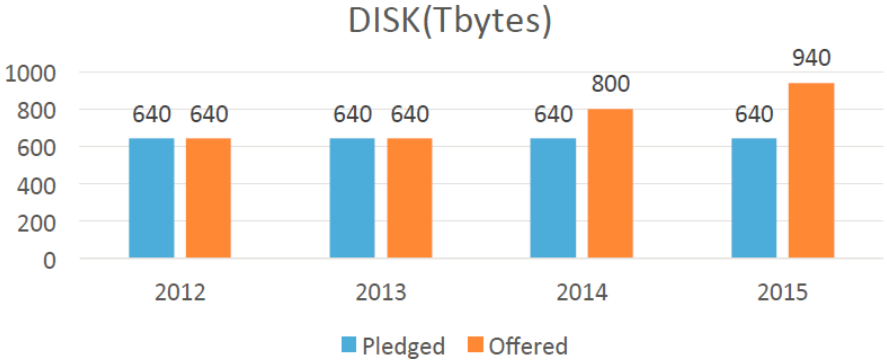




Grid Tier 2 Site (1)

- CMS and Atlas Tier2 Site
- CPU: 888 Cores
  - Intel E2680V3: 696 Cores
  - Intel X5650 192 Cores
  - Batch: Torque/Maui
- Storage: 940 TB
  - DPM: 400TB
    - 4TB X 24slots With Raid 6
    - 5 Array boxes
  - dCache: 540TB
    - 4TB X 24slots With Raid 6. 8 Array box.
    - 3TB X 24slots With Raid 6. 1 Array box
- Network
  - 5 Gbps link to Europe visa Orient-plus and 10 Gbps to US



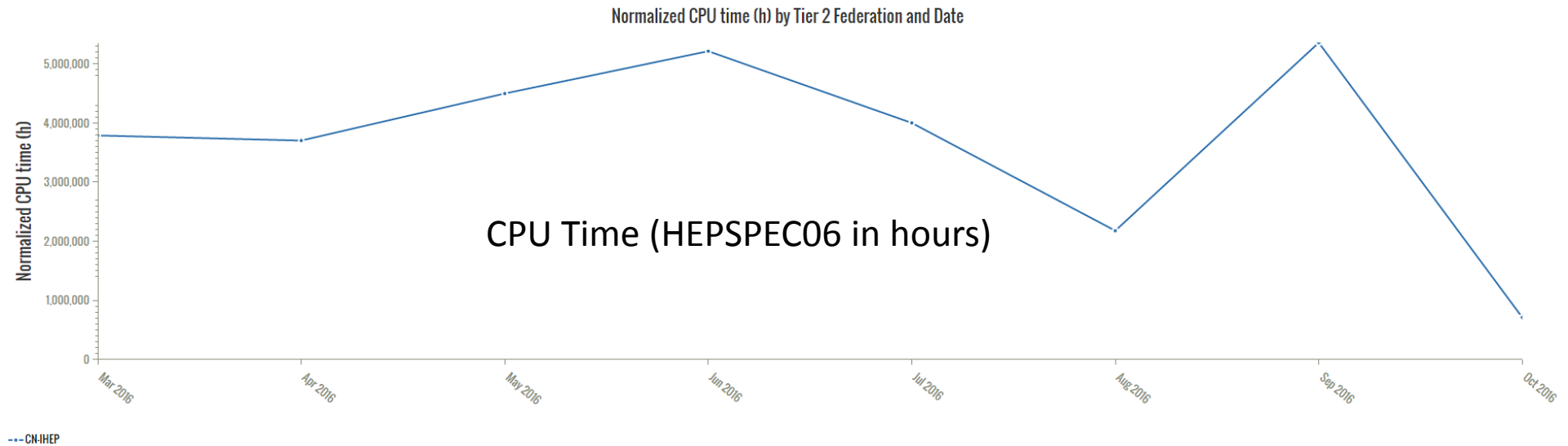


Grid Tier 2 Site (2)

- Site Statistics last half year

29,445,996

VO	Mar 2016	Apr 2016	May 2016	Jun 2016	Jul 2016	Aug 2016	Sep 2016	Oct 2016		Percent
atlas	2,239,566	2,689,636	3,725,879	2,756,796	1,981,214	1,851,600	3,875,671	453,156	19,573,518	66.47%
cms	1,548,616	1,012,453	771,408	2,456,745	2,018,930	325,453	1,481,687	257,186	9,872,478	33.53%
Total	3,788,182	3,702,089	4,497,287	5,213,542	4,000,144	2,177,053	5,357,358	710,343	29,445,996	
Percent	12.86%	12.57%	15.27%	17.71%	13.58%	7.39%	18.19%	2.41%		



CPU Time (HEPSPEC06 in hours)

BESIII Distributed Computing

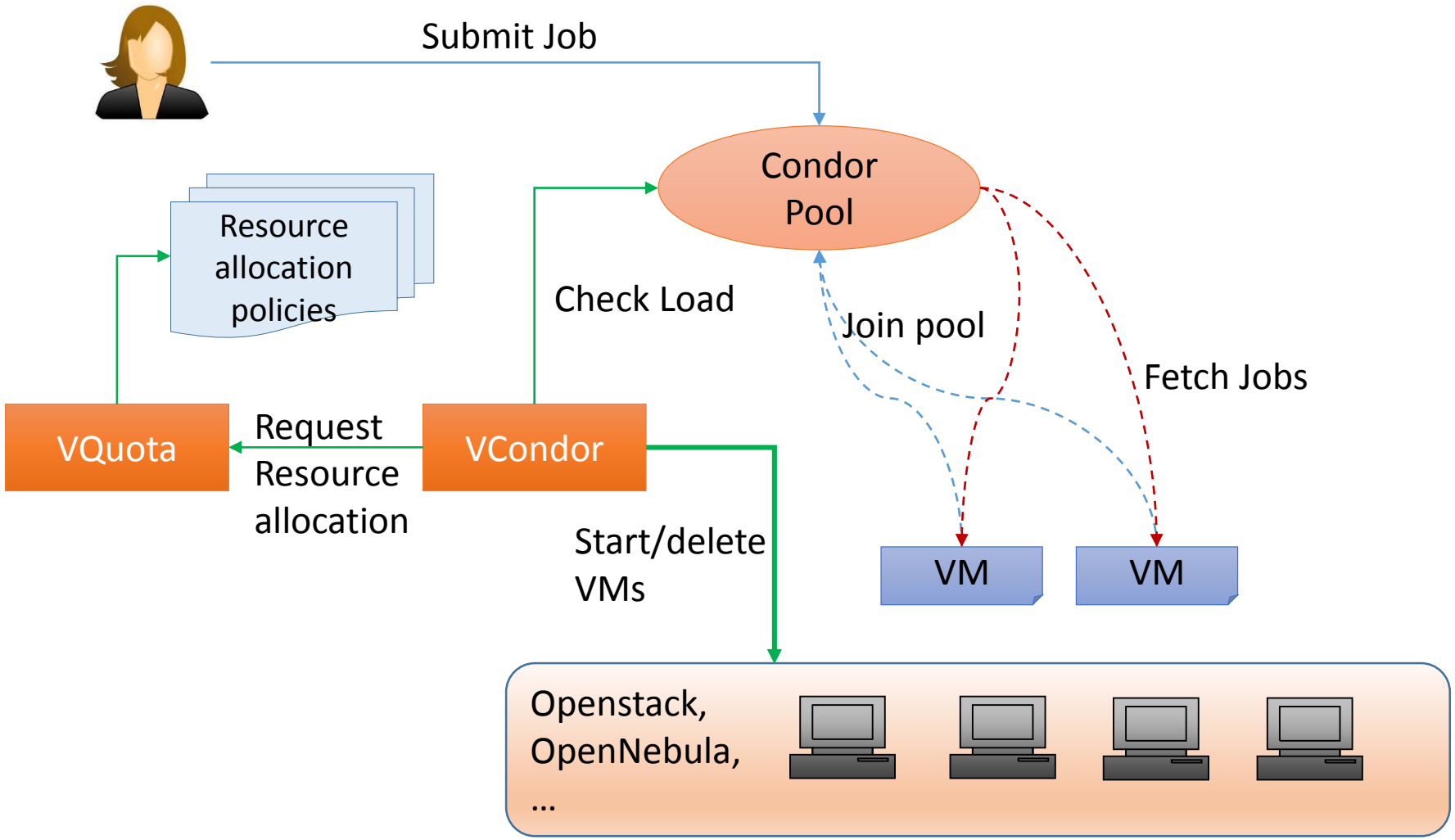
- During August summer maintenance, DIRAC server has been successfully upgraded from v6r13 to v6r15
  - Prepare to support multi-core jobs in the near future
  - VMDirac has been upgraded to 2.0, which greatly simplifies the procedure to adopt new cloud sites
- New monitoring system has been put into production, which gives a clear view of real-time site status
- Total resources: 14 sites, ~3000 CPU cores, ~500TB storage
- Support CEPC and other experiments

#	Site Name	CPU Cores	Storage	#	Site Name	CPU Cores	Storage
1	CLOUD.IHEP.cn	210	214 TB	9	GRID.JINR.ru	100 ~ 200	30 TB
2	CLUSTER.UCAS.cn	152		10	GRID.INFN-Torino.it	200	60 TB
3	CLUSTER.USTC.cn	200 ~ 600	24 TB	11	CLOUD.TORINO.it	128	
4	CLUSTER.PKU.cn	100		12	CLUSTER.SDU.cn	100	
5	CLUSTER.WHU.cn	120 ~ 300	39 TB	13	CLUSTER.BUAA.cn	100	
6	CLUSTER.UMN.us	768	50 TB	14	GRID.INFN-ReCas.it	50	30 TB
7	CLUSTER.SJTU.cn	100		15	CLOUD.CNIC.cn	50	50 TB
8	CLUSTER.IPAS.tw	300		16	CLUSTER.NEU.tr	50	

IHEP Cloud

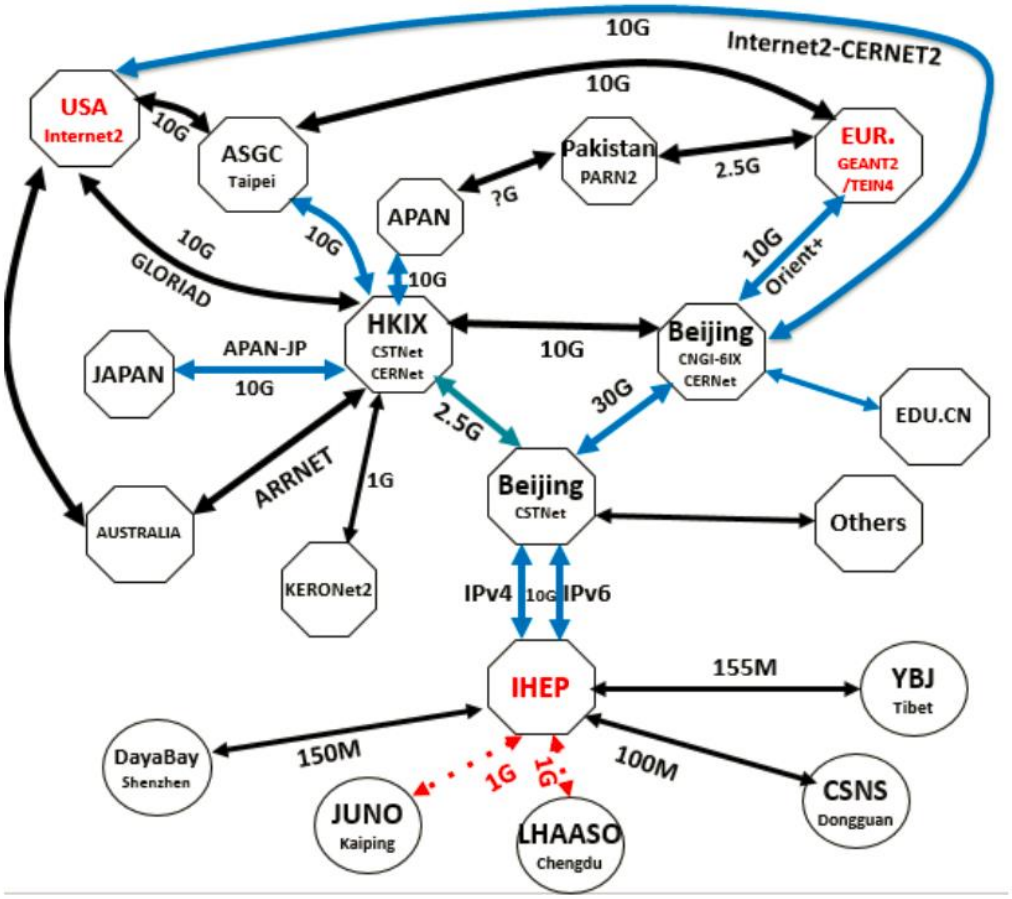
- IHEP private cloud
  - Provide cloud services for IHEP users and computing demand
- Based on Openstack
  - Launched in May 2014
  - Upgraded from Icehouse to Kilo in 2016
  - Single sign on with IHEP UMT (oauth2.0)
- Resources
  - User self-service Virtual Machines service
    - 21 computing nodes, 336 CPU cores
  - Dynamic Virtual computing cluster
    - 28 computing nodes, 672 CPU cores
    - Support BESIII, JUNO, LHAASO, …
  - Will add 768 CPU cores this year
- More detail in Cui Tao's talk

VCondor



Github: https://github.com/hep-gnu/VCondor

Internet and Domestic Network



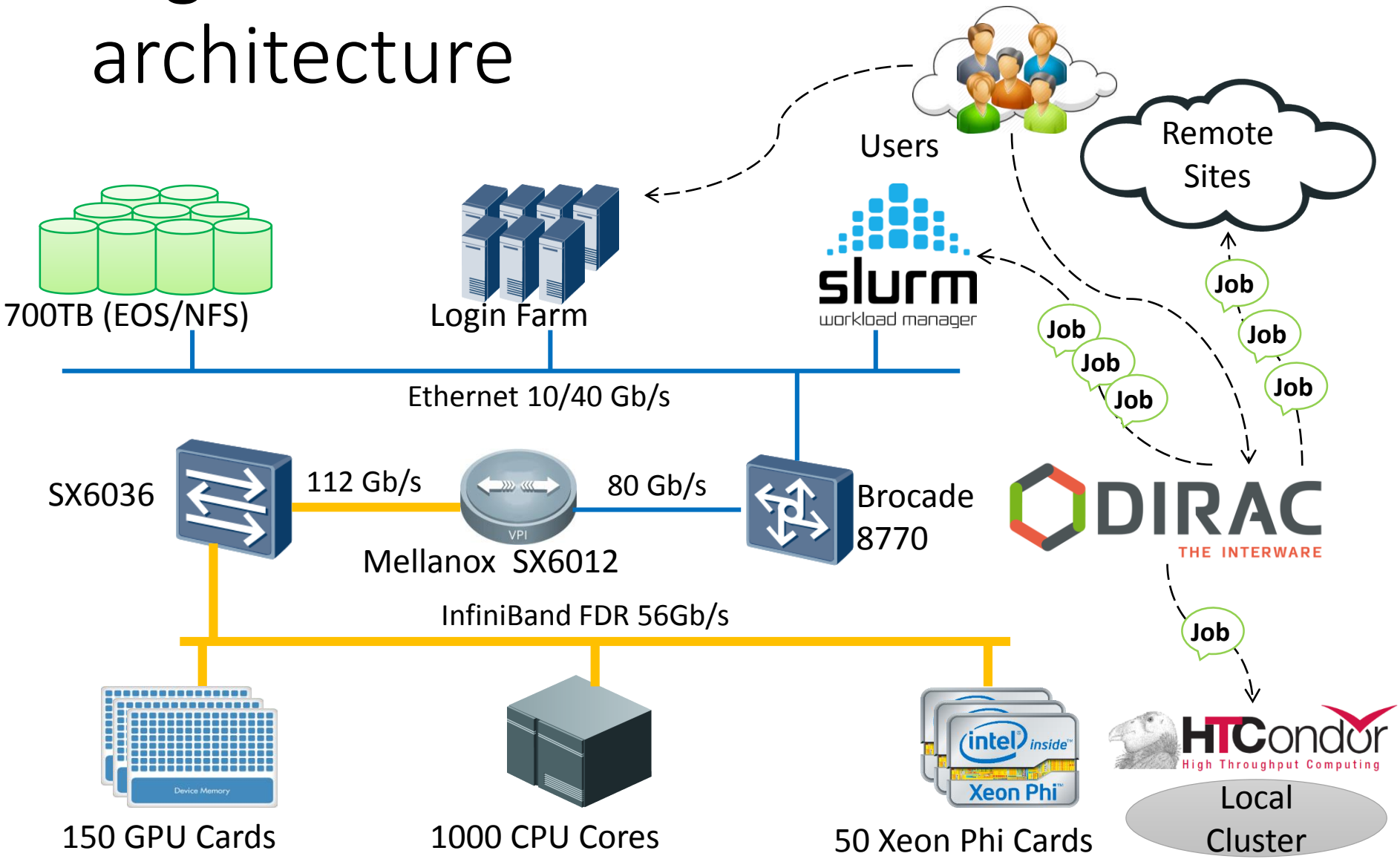
- IHEP-EUR: 10Gbps
- IHEP-USA: 10Gbps
- IHEP-Asia: 2.5Gbps
- IHEP-Univ: 10Gbps

SDN@IHEP: Zeng Shan's talk

High Performance Cluster

- A new heterogeneous hardware platform : CPU, Intel Xeon Phi, GPU
- Parallel programming supports : MPI, OpenMP, CUDA, OpenCL …
- Potential usage cases : simulation, partial wave analysis …
- SLURM as the scheduler
  - Test bed is ready: version 16.05
    - Virtual machines: 1 control node, 4 computing nodes
    - Physical servers: 1 control node, 26 computing nodes(2 GPU servers included)
  - Undergoing scheduler evaluation
    - Two scheduler algorithms evaluated: sched/backfill, sched/builtin.
    - Undergoing integration with DIRAC
- Network architecture & technologies
    - InfiniBand network for HPC test bed was already built

High Performance Cluster architecture



Next step

- ~2.5PB (available storage) will be added
- Migration from PBS to HTCondor will be completed by the end of this year
- IHEP will provide routine maintenance service to more remote sites
- Modify CASTOR 1 to integrate new LTO6 tape
- HPC cluster will be provided next year

Thank you!

chyd@ihep.ac.cn