IHEP Site Report

4th Asia Tier Forum, Bangkok

Shi, Jingyan
shijy@ihep.ac.cn

Computing Center, IHEP

Outline

1. Introduction to IHEP-CC
2. Computing, Storage and Network Resources
3. Architecture and Services
4. Beijing LCG Tier 2 Site
5. Summary

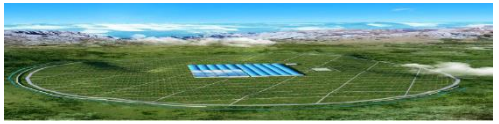
Experiments We Support

BESIII (Beijing Spectrometer III at BEPCII)

DYB (Daya Bay Reactor Neutrino Experiment)

JUNO (Jiangmen Underground Neutrino Observatory)

YBJ (Tibet-ASgamma ARGO-YBJ Experiments)



LHAASO (Large High Altitude Air Shower Observatory)

HXMT(Hard X-Ray Moderate Telescope)

CMS

ATLAS EXPERIMENT

CEPC

HEPS (High Energy Photon Source)

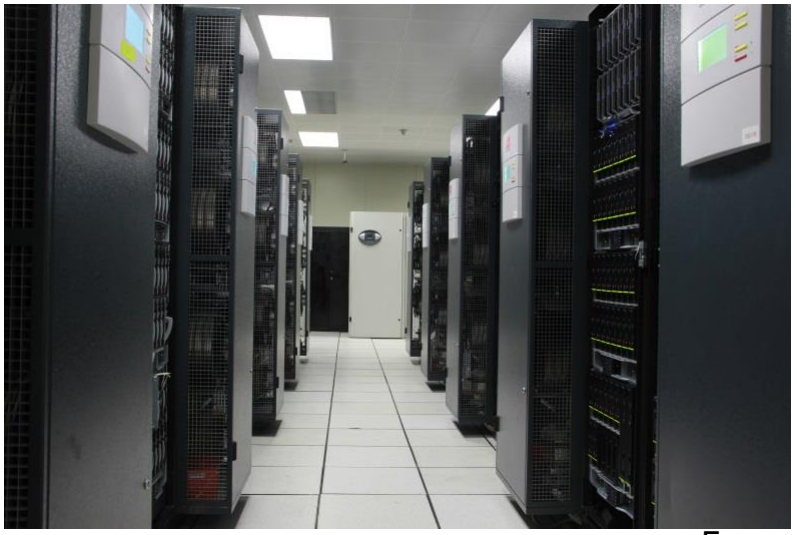
New supported

Outline

1. Introduction to IHEP-CC
2. Computing, Storage and Network Resources
3. Architecture and Services
4. Beijing LCG Tier 2 Site
5. Summary

Computing Resources

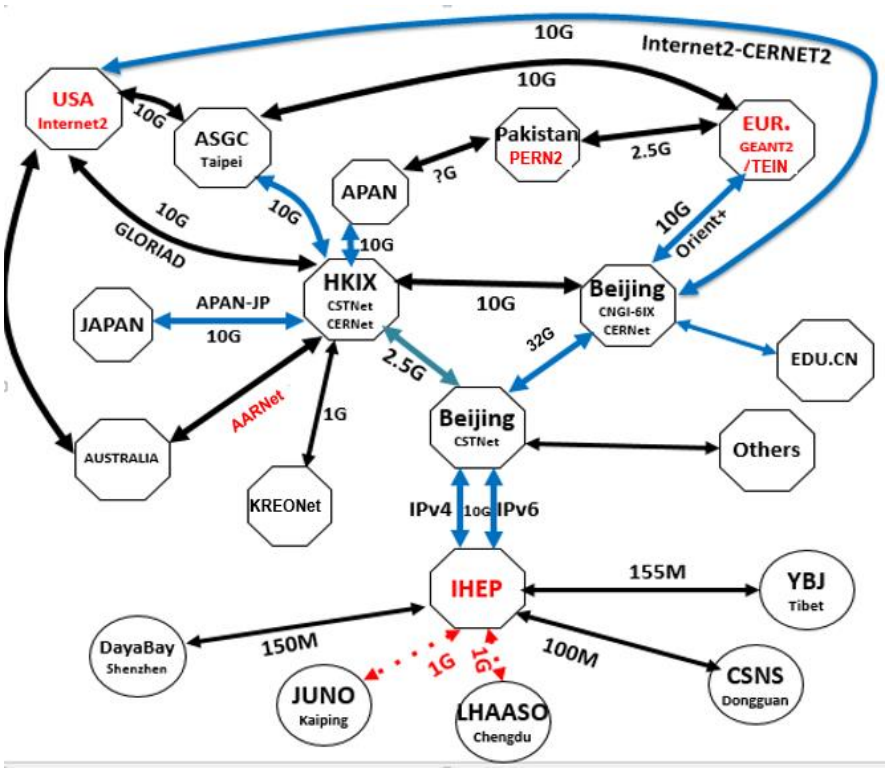
- HTCondor cluster
  - HTC jobs: series jobs
  - ~13,400 CPU cores
- Slurm clusters
  - Cpu nodes + GPU nodes
  - HPC jobs: parallel jobs + GPU jobs
  - 3,384 CPU cores + 8 GPU cards
  - 80 GPU cards will be added
- Grid site (WLCG)
  - Tier 2 site
  - 1,200 CPU Cores
  - 1000 CPU cores for LHCb will be added
- The BESIII DIRAC-based distributed computing system
  - ~ 3,500 CPU cores
- IHEPCloud based on Openstack
  - ~ 2,000 CPU cores

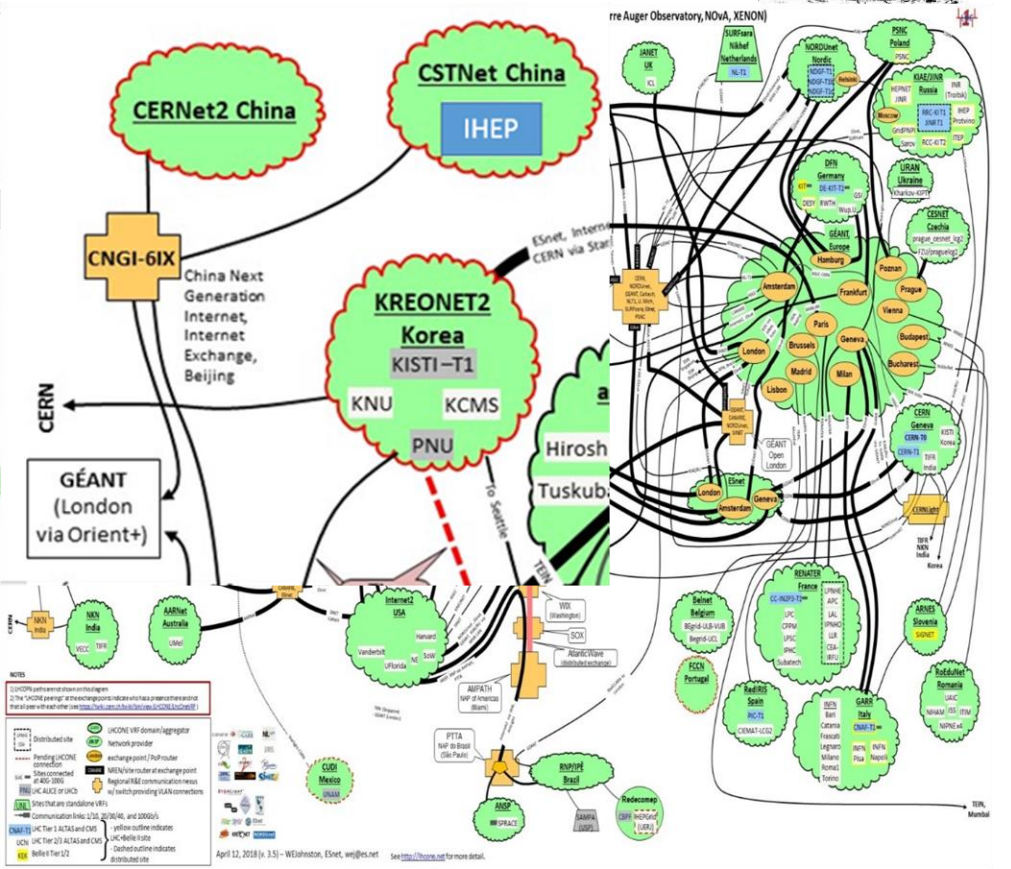


Storage Resoruce

- Local cluster
  - 15 PB+ disk storage
    - Lustre: 11 PB
    - EOS: 2.3PB
    - Other: 1.5PB
  - 5 PB tape storage: Castor
- Grid site
  - DPM: 400TB
  - dCache: 540TB

International Network

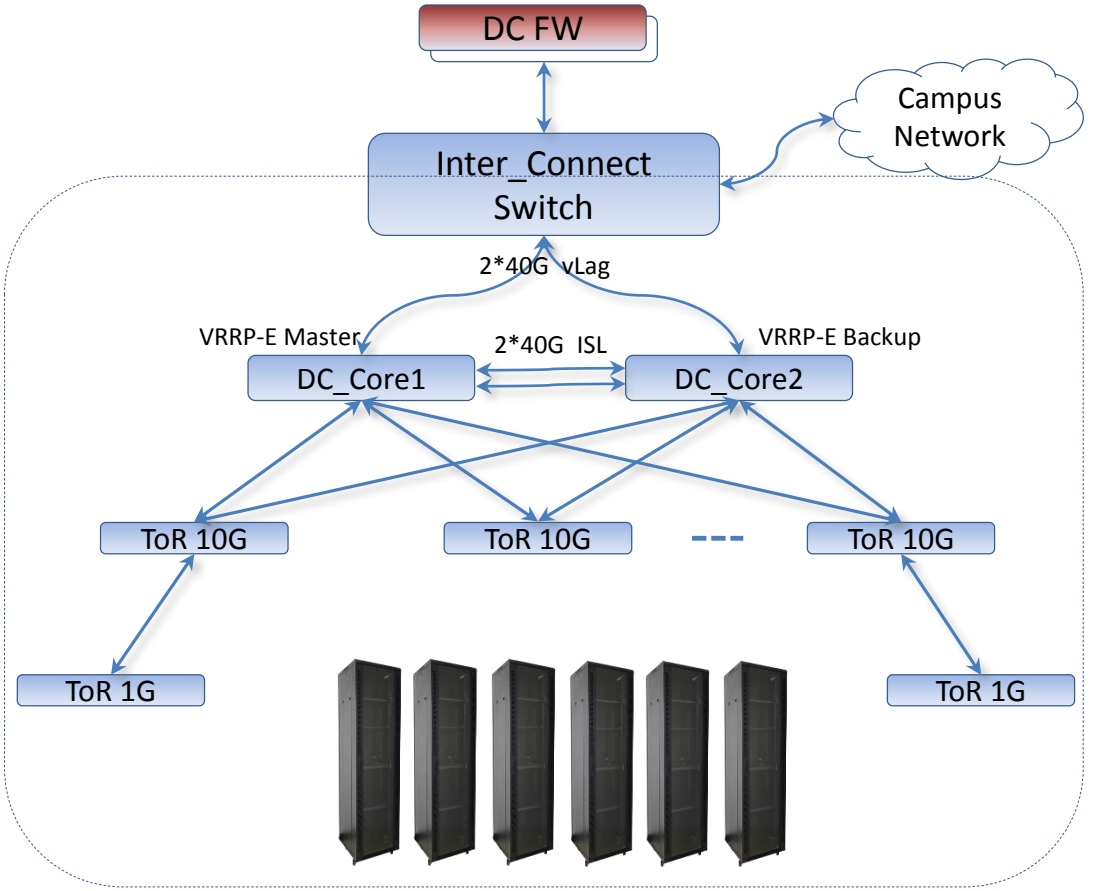




- IHEP joined LHCONE (LHC Open Private Network) in March, 2018.

DC Network Topology

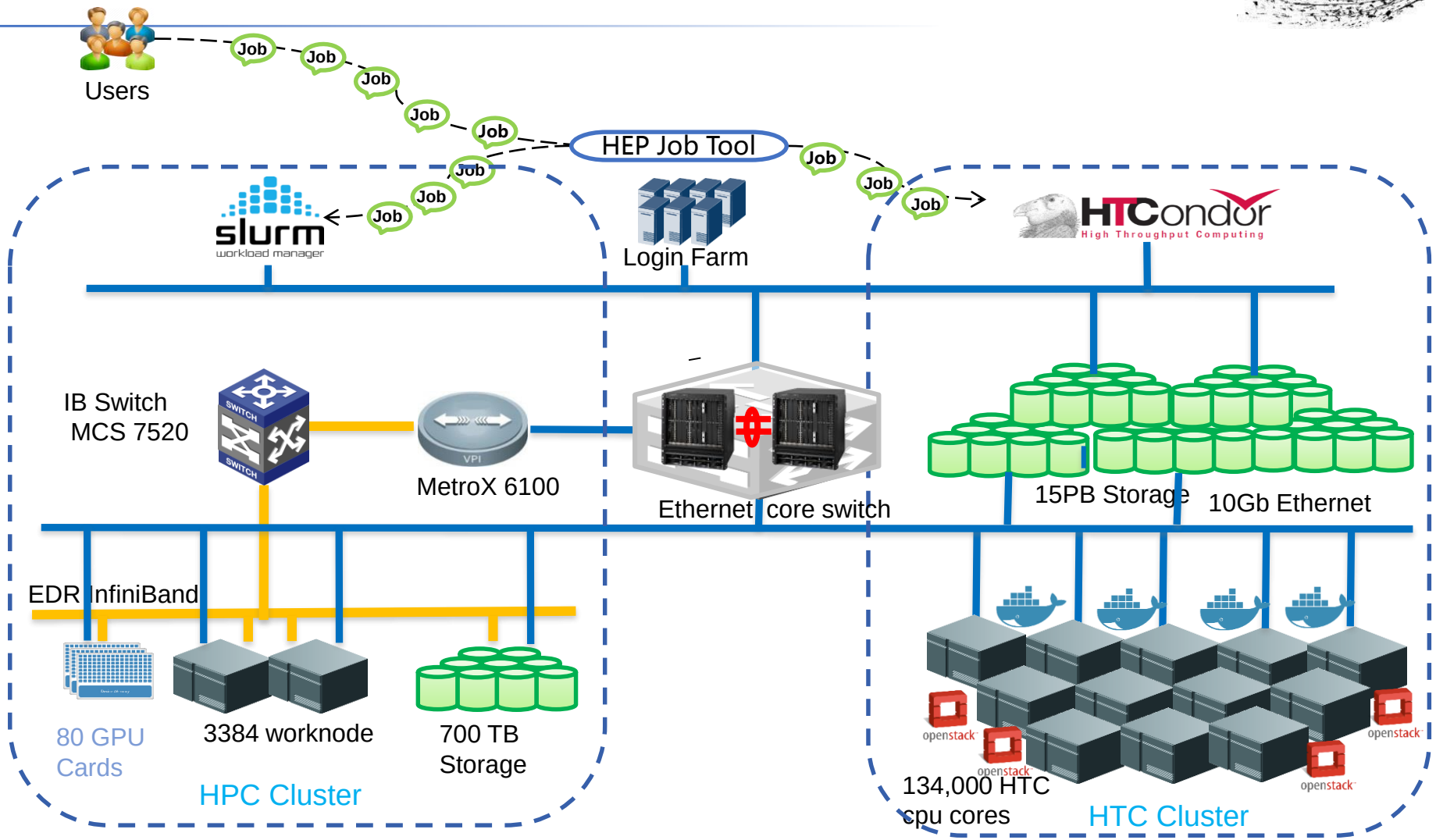
- From single Core switch to double Core Switches with VRRP-E (Brocade VDX 8770)
  - VRRP extended (VRRP-E) is an extended version of the VRRP protocol. Brocade developed VRRP-E as a proprietary protocol to address some limitations in standards-based VRRP
  - ISL(inter switch link) 80G
  - Uplink 80G
  - Backbone bandwidth 160G
  - Storage server 20G
- Network equipments
  - 2 Core switch VDX 8770
  - 20 10G ToR Switch
  - 25 1G ToR Switch



Outline

1. Introduction to IHEP-CC
2. Computing, Storage and Network Resources
3. Architecture and Services
4. Beijing LCG Tier 2 Site
5. Summary

HPC+HTC Architecture

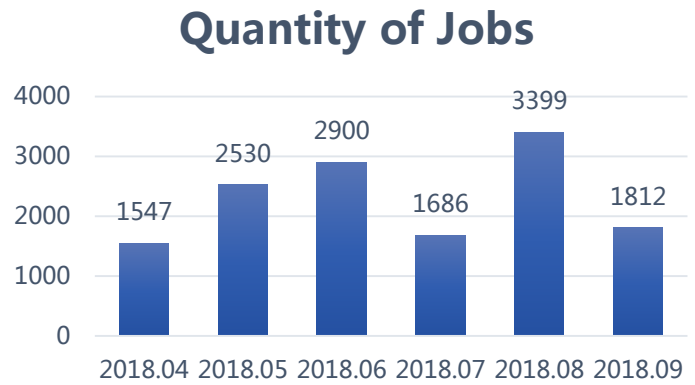


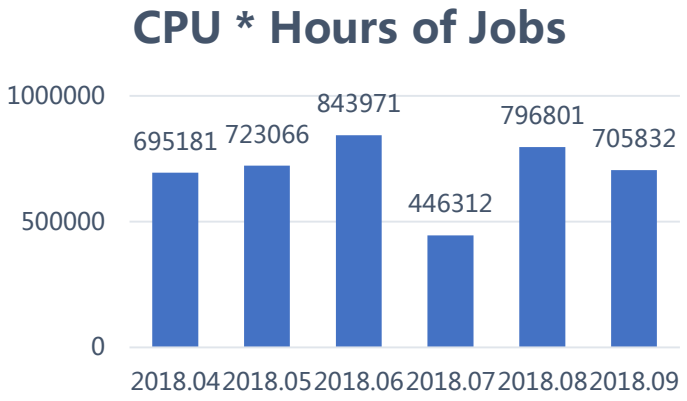
HTC Shared Scheduling Policy

- New scheduling policy for HTCondor cluster
  - Resource contributed by all experiments
  - A shared pool includes the most job slots accept jobs from all experiments
  - Linux group quota guarantee the fairness among experiments
  - Surplus scheduling policy to promote the job slots utilization
  - HTCondor provides fairness policy among the same experiment users
- High job slots utilization → **~95%**

Slurm Cluster

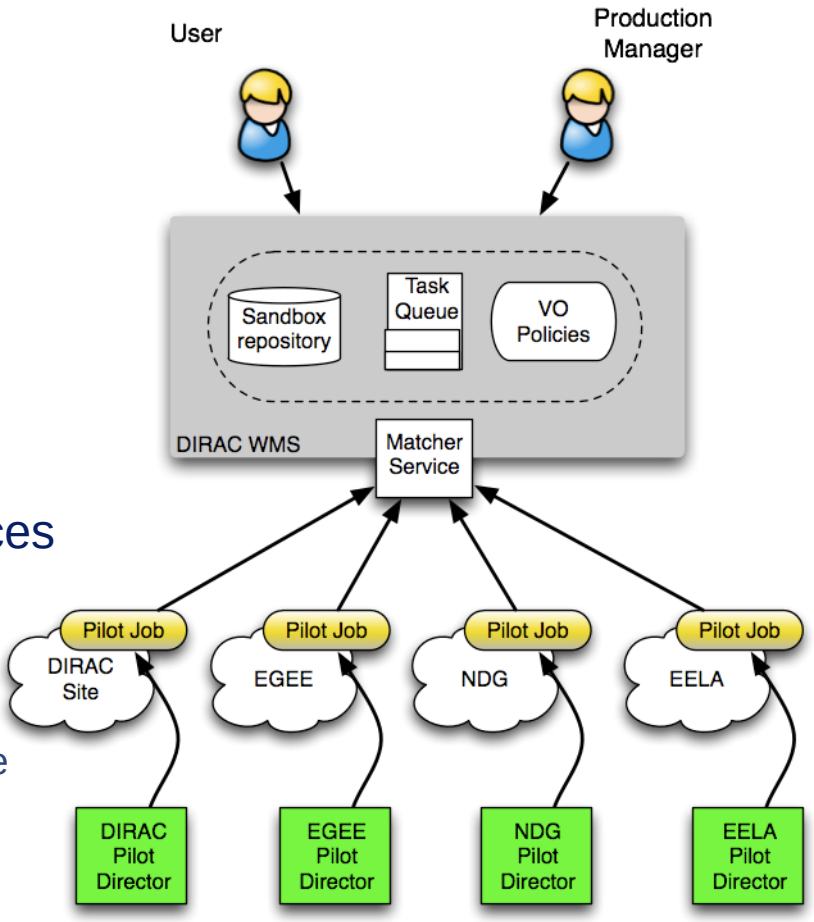
- Resources
  - 1 master node, 1 accounting & monitoring node
  - 16 login nodes
  - 141 work nodes: 3,384 CPU cores + 8 GPU cards
- Jobs (2018.04~2018.09)
  - Jobs number : ~14K
  - CPU hours : ~4.2 million
- GPU servers procurement
  - 80 GPU cards : NVIDIA Tesla V100 nvlink 32GB
  - 1 PFLOPs (single precision)
  - INSPUR won the bid





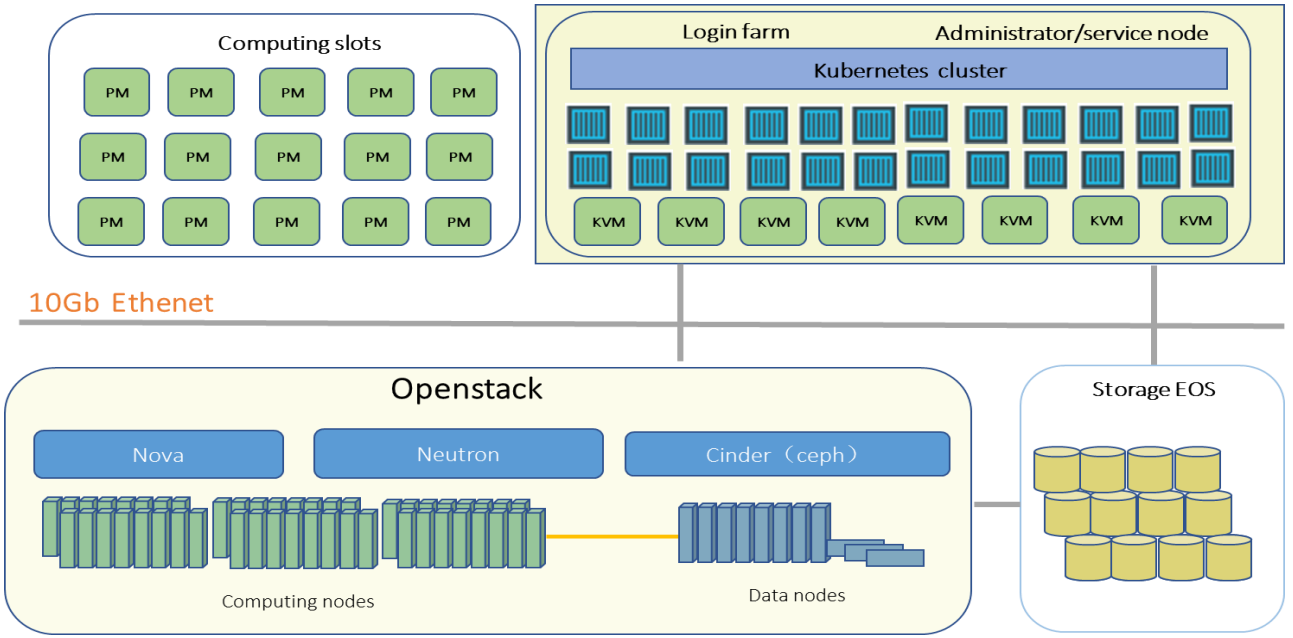
Distributed Computing

- Distributed Infrastructure with Remote Agent Control (DIRAC)
  - A general purpose open source distributed computing framework
- Pilot based Workload Management provides abstraction of Computing Resources
  - Allows to combine heterogeneous resources in a transparent way
- IHEP distributed computing built on DIRAC implemented integration of computing resources among collaborations
  - About 14 sites from USA, Italy, Russia, China universities
  - About 3,500 CPU cores and 500 TB disk storage have been integrated
  - Support Grid, Cluster, Cloud and Volunteer computing



Computing Service for LHAASO

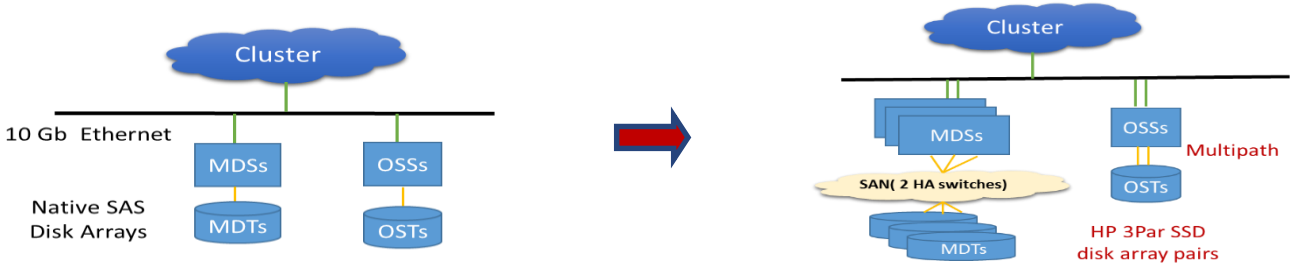
- A new remote site for LHAASO, located in Dao Cheng (at the altitude of 4,410m), Sichuan Province, China
- Cloud-based service to reduce the operation and maintenance cost
  - Login and administration nodes are managed by openstack + kubernetes
  - Jobs are scheduled by HTCondor



Lustre Storage

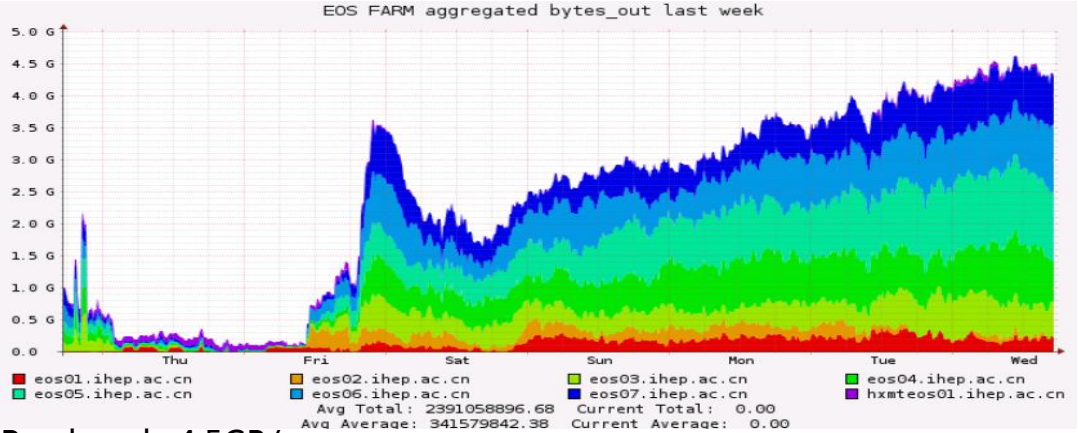
- Lustre has been adapted for 10 years with rich experiences
- Currently 10 Lustre instances, totally 11 PB
- Throughput
  - Read: 15 GB/s Peak, 10 GB/s Average
  - Write:1.1 GB/s Peak, 0.2 GB/s Average

- New robust architecture without single failure point
- Evaluation of new features in Lustre 2.11+
  - File Replication, Data on MDT, Progressive File Layout …

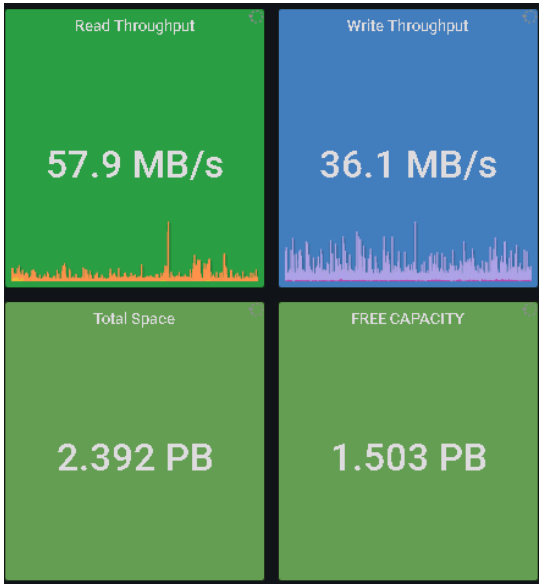


New Storage at IHEP

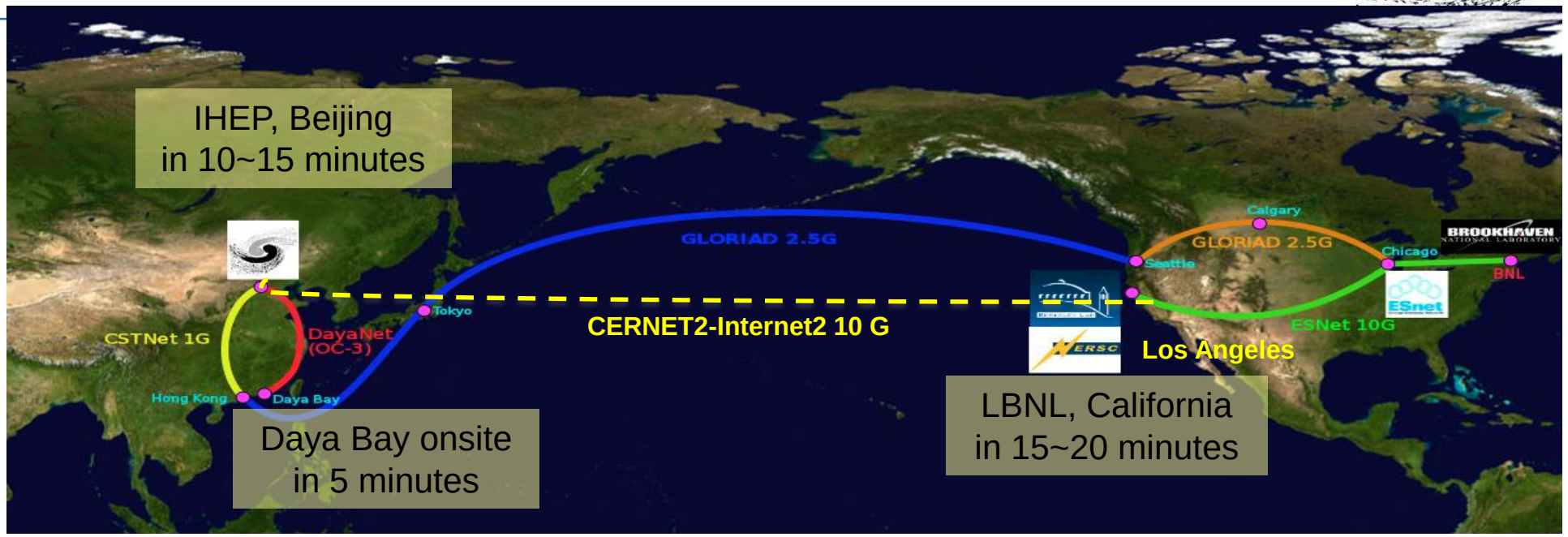
- EOS
  - EB-level storage software developed by CERN
  - To be the future solution for HEP storage
- Deployed at IHEP in 2016
  - 2 instances for physics data storage
    - LHAASO: 2.3PB
    - HXMT: 330TB
  - 1 instance for users' own data (IHEPBox based on owncloud)

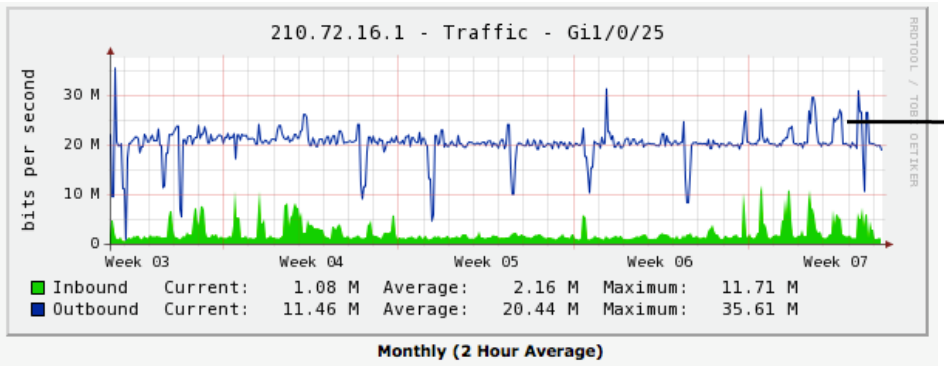


Read peak: 4.5GB/s
Peak values allowed by the environment (mainly 5 FST each has 10Gb Ethernet)

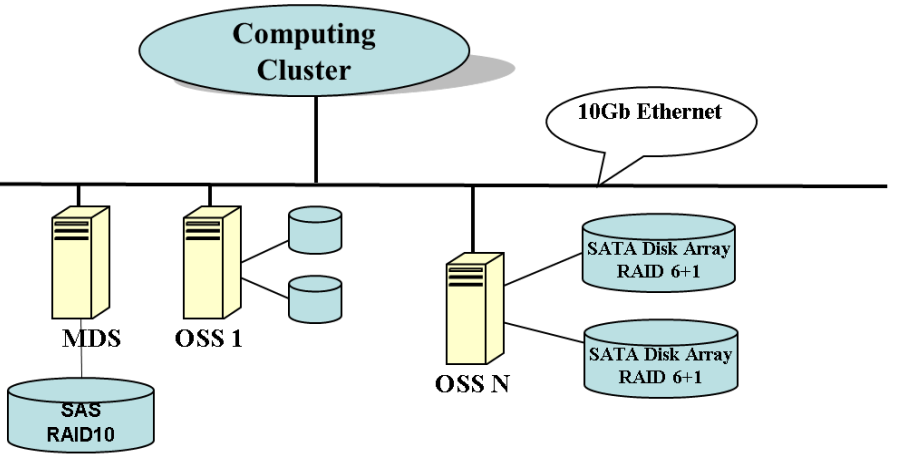


Data Transfer





Daya Bay onsite network monitoring

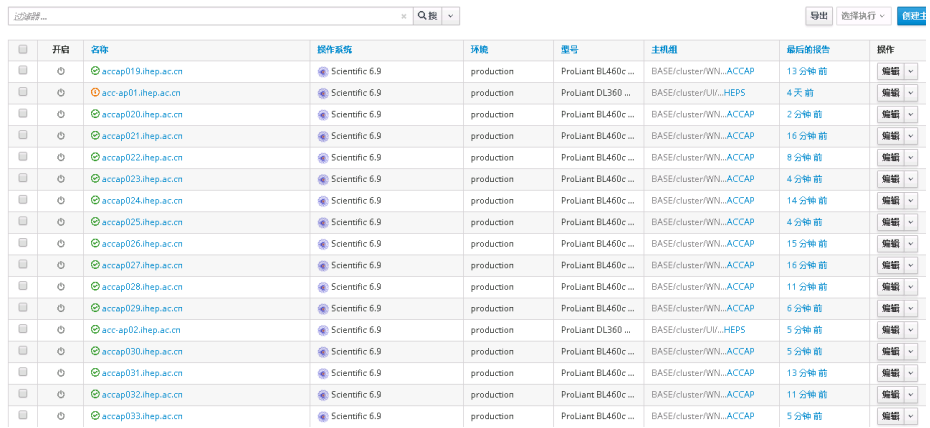


Infrastructure of data storage

Deployment & Monitor

- Software automatic deployment
  - Quattor → Puppet: More flexible and better scalability
  - More than 2,000 machines: OS installation & software upgrade automatically
- Monitoring
  - NMS: 20,000+ service metrics from all machines are under real time monitoring
    - Mini-error: fixed automatically
    - Serious-error: warning
      - Message, email, wechat

主机

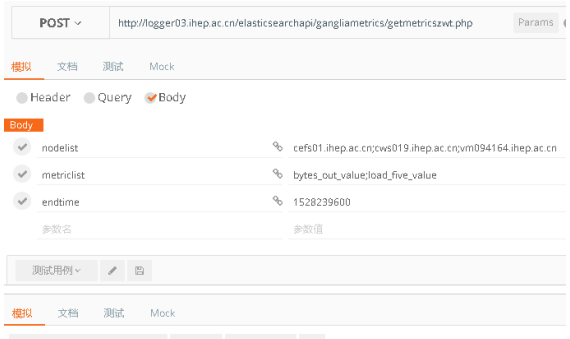


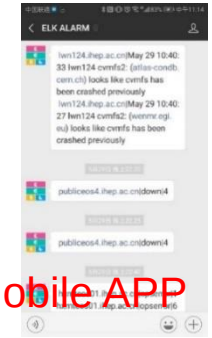
开启	名称	操作系统	环境	型号	主机组	最后的报告	操作
	accap019.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	13 分钟 前	编辑
	acc-ap01.ihep.ac.cn	Scientific 6.9	production	ProLiant DL360 ...	BASE/cluster/UI/...HEPS	4 天 前	编辑
	accap020.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	2 分钟 前	编辑
	accap021.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	16 分钟 前	编辑
	accap022.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	8 分钟 前	编辑
	accap023.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	4 分钟 前	编辑
	accap024.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	14 分钟 前	编辑
	accap025.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	4 分钟 前	编辑
	accap026.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	15 分钟 前	编辑
	accap027.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	16 分钟 前	编辑
	accap028.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	11 分钟 前	编辑
	accap029.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	6 分钟 前	编辑
	acc-ap02.ihep.ac.cn	Scientific 6.9	production	ProLiant DL360 ...	BASE/cluster/UI/...HEPS	5 分钟 前	编辑
	accap030.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	5 分钟 前	编辑
	accap031.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	13 分钟 前	编辑
	accap032.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	11 分钟 前	编辑
	accap033.ihep.ac.cn	Scientific 6.9	production	ProLiant BL460c ...	BASE/cluster/WN...ACCAP	5 分钟 前	编辑

Service Status Totals

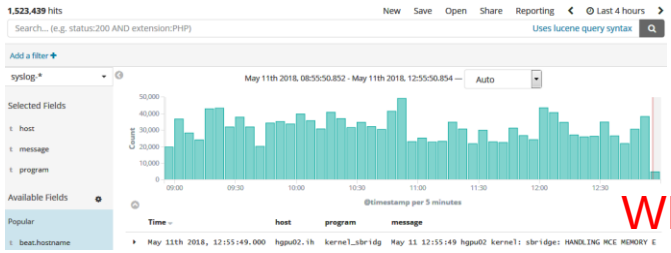
Ok	Warning	Unknown	Critical	Pending
19285	22	16	35	0
3999	1	4	307	0
630	7	0	1	0
184	4	0	1	0
5	6	0	4	0

New Monitoring System - ELK





Mobile APP



WEB

To: "Jiangxw@mail.ihep.ac.cn" <Jiangxw@mail.ihep.ac.cn>, "YAN Tian" <yant@ihep.ac.cn>
Cc:
Subject: 作业信息获取API

作业统计接口 API1： ES直接返回结果
curl -XPOST "http://logger01.ihep.ac.cn:9200/programinfo-*/_search" -H 'Content-Type: appl
{
"_source": ["cmdabpath","leaf","command","cpu","pid","user","host","stat","runtime","slotid"
"query": {
"bool": {
"must": [
{
"query_string": {
"query": "leaf:1 AND cpu:[88 TO *] AND NOT cmdabpath: (\"/afs/ihep.ac.cn/soft/\" C
"analyze_wildcard": true,
"default_field": "*"
}
},
{
"range": {

Mail

Outline

1. Introduction to IHEP-CC
2. Computing, Storage and Network Resources
3. Architecture and Services
4. Beijing LCG Tier 2 Site
5. Summary

BEIJING LCG Site

- CPU: 888 cores
  - Intel E2680V3: 696 Cores
  - Intel X5650 192 Cores
- Batch: Torque

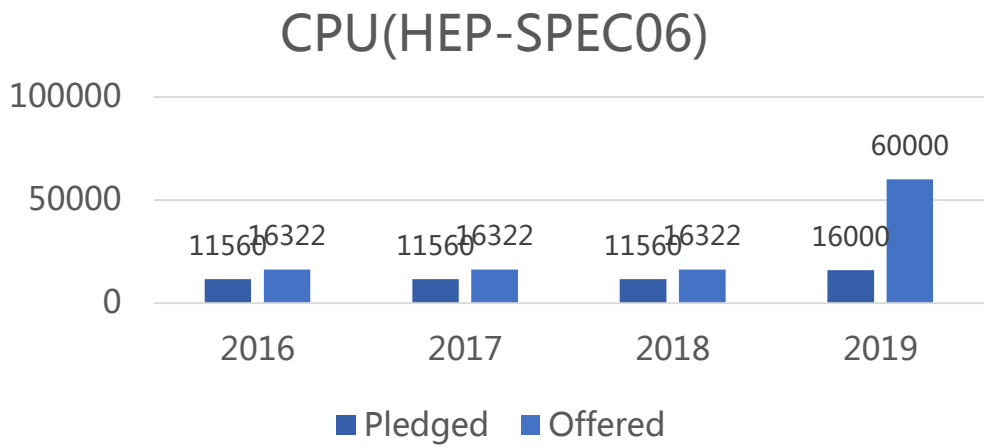
- DPM: 400TB
  - 4TB * 24slots with Raid 6, 5 Array boxes
- dCache: 540TB
  - 4TB * 24slots with Raid 6, 8 Array boxes
  - 3TB * 24slots with Raid 6, 1 Array box

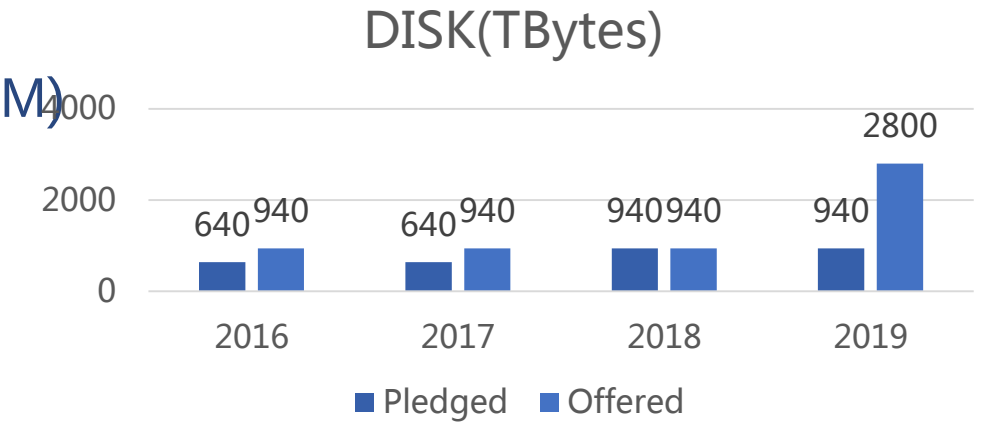
The Site keeps a good reliability at most of the time

	04/18	05/18	06/18	07/18	08/18	09/18
ATLAS	100.00%	97.00%	100.00%	100.00%	100.00%	97.00%
CMS	99.00%	99.00%	100.00%	99.00%	100.00%	100.00%

The Plan of Grid Tier-2 in 2019

- Grid Tier-2 for LHCb
  - 1080CPU cores
  - 360TB
- Resources replacement
  - CPU: Intel Golden 6140, 3456 cores
  - HEPSPEC06: 60,000
  - Storage: 2,800TB (dCache + DPM)
  - ATLAS:CMS:LHCb ~1:1:1
- CE: HTcondor-CE





Summary

- IHEP site has been providing computing service to HEP experiments
- IHEP site scale will be expanded
- Trying to keep up with the new technology trends
- Challenge are always in front of us

Thank you !
Question?