Computing Status of JUNO

2018-05-11 @ Wuhan University

IHEP – Computing Center

Zou Jiaheng

Outline

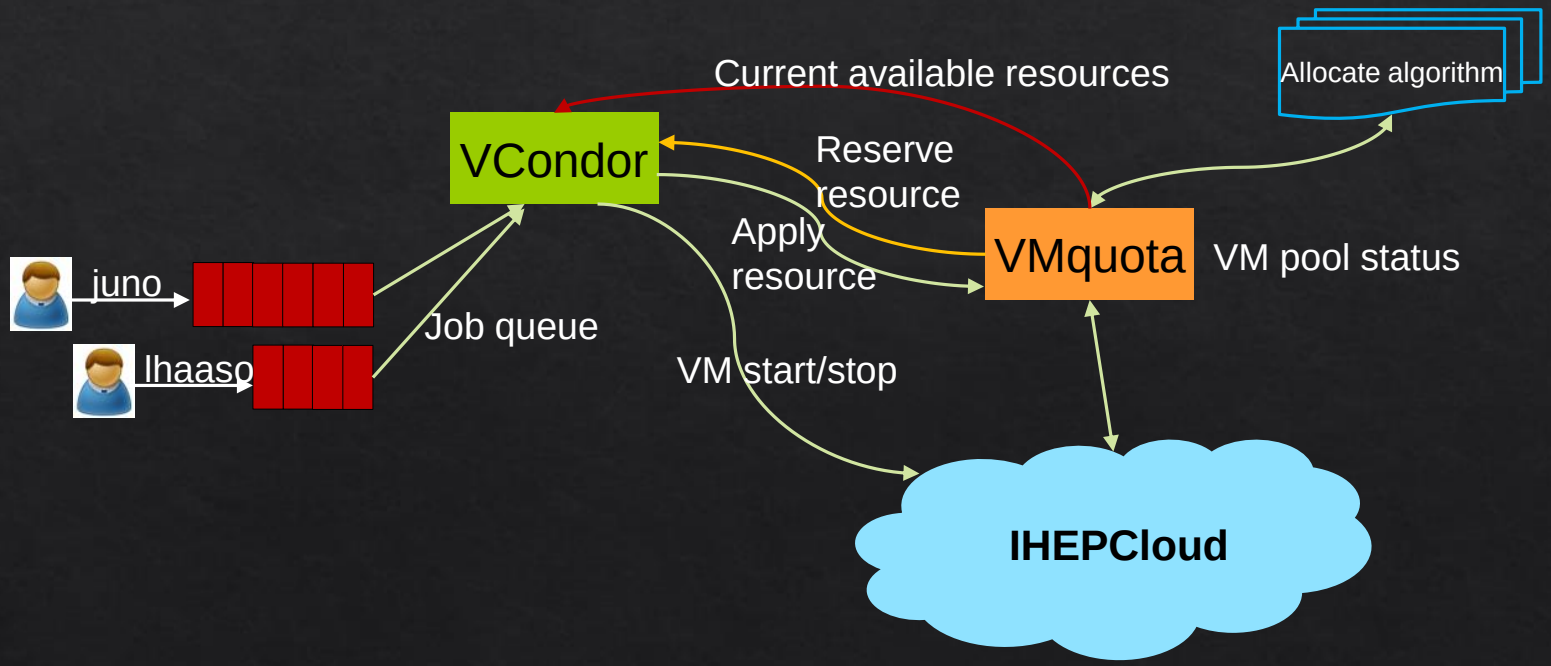
- Computing resources
  - HTC
  - HPC
  - Cloud
- Storage
- Network and Data transfer
- Summary

HTC Cluster

- HTC - High Throughput Computing
- Totally 10,000+ CPU cores
  - JUNO: 888 CPU cores
- HTCondor as job scheduler
  - Optimized for HTC job scheduling policies
  - Capacity for large scale computing clusters
- Resources are shared between all experiments
  - There are always some busy experiments and some free experiments
  - Busy experiments/groups can take benefits from free experiments/groups
  - There can be more than 3000 JUNO jobs running at the same time
  - The overall resource utility is greatly improved

Virtual Computing Cluster

- Based on OpenStack Kilo
- Resource provision dynamically to meet peak requirements
- Improve resource sharing between different experiments
- Transparent to users



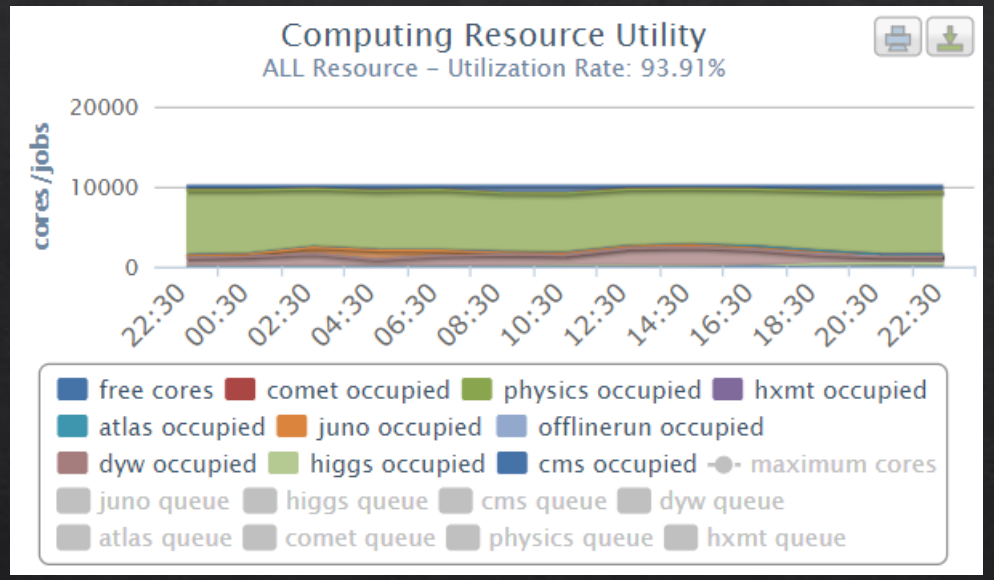
Statistics of JUNO HTC Jobs (2018.1-2018.5)

- Total
  - Job count ~ 2,365,547
  - CPU Walltime ~ 2,187,801 hours
  - Activity users ~ 61
- Physical resources usage
  - Job count ~ 2,283,210
  - CPU Walltime ~ 2,185,250 hours
- Virtual resources usage
  - Job count ~ 82337
  - CPU Walltime ~ 2,551 hours

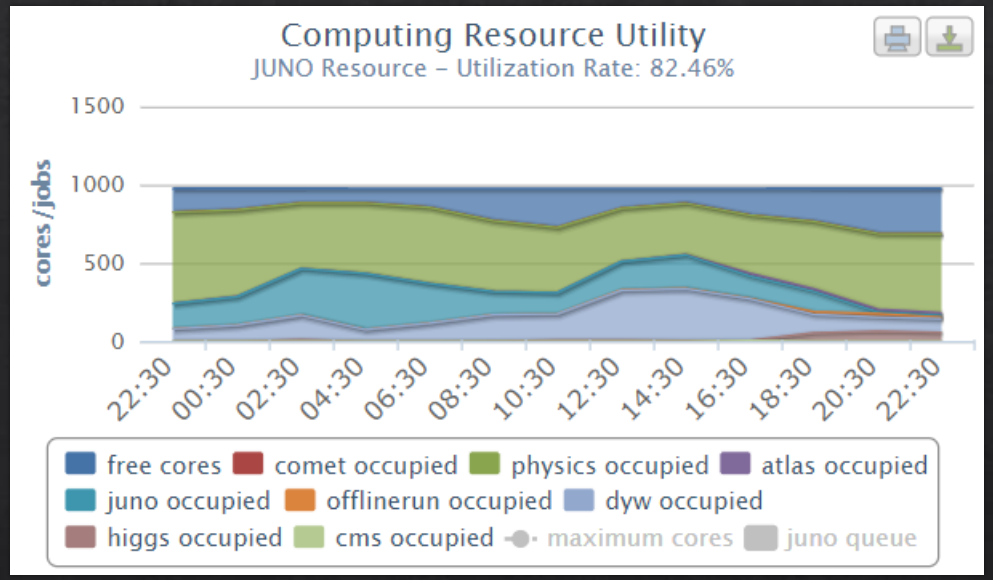
username	count	walltime(h)
zhangfy	948938	765559
huangyb	205308	261046
guoyh	355633	163499
wenjie	70817	144030
gorchakov	24726	139446
zhangqm	34932	137015
liangjs	16689	113098
huanggh	81161	85678
weilh	122512	73892
xujl	25323	63251
rengl	81003	56293
yyzhang	82605	44626
lihl	97650	16300
zhaobq	2076	14431
zhangyp	7657	9558
xuyu	7126	9217
lizy	12163	8746
waseem	1088	8644
yury	3020	8091
huangqh	4956	7467
dingxf	8734	6836

Resource Utility

- The overall resource utility keeps more than 90% on workdays



The overall resource utility

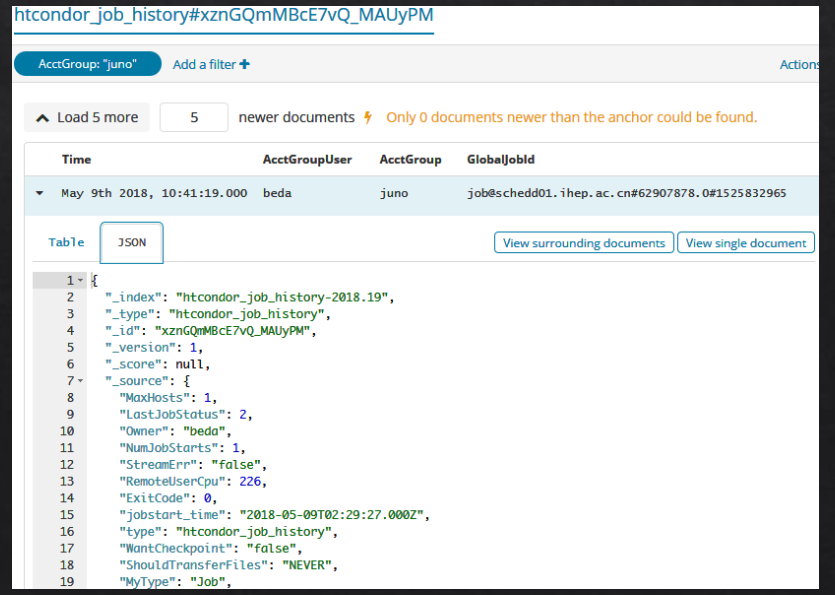


The JUNO resource utility

- JUNO resources can be occupied by others when it is free
- JUNO jobs can be scheduled to others' free resource when it is busy

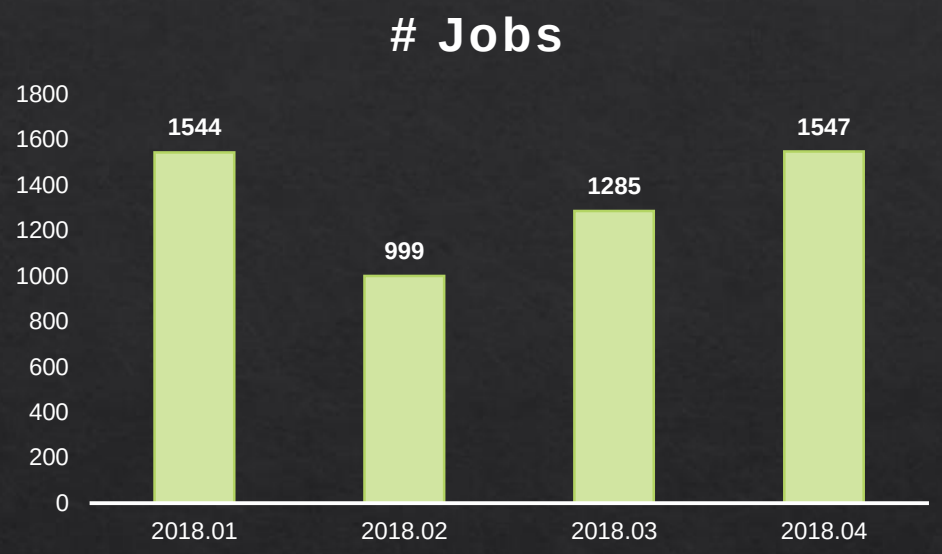
Integrated Monitoring

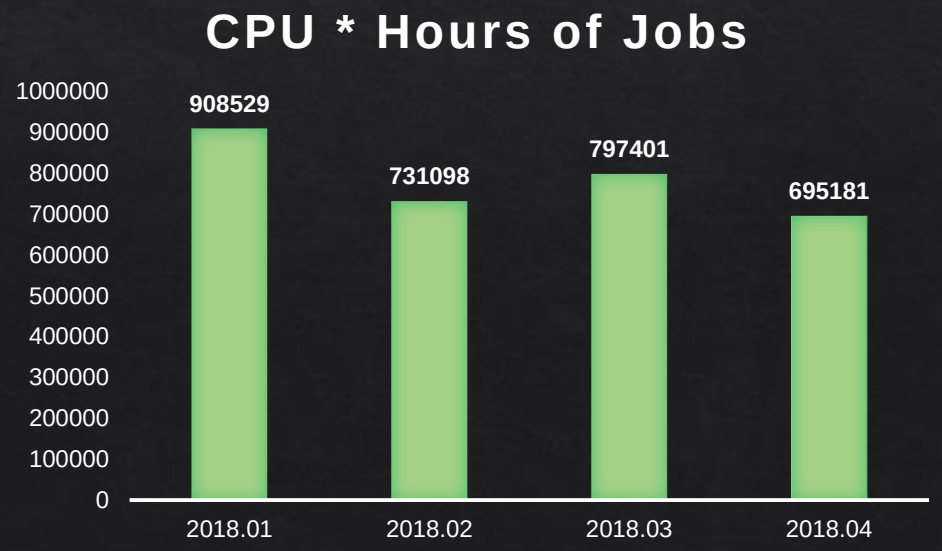
- HTCondor job logs are collected into Elasticsearch
- History Job statistics are displayed via Grafana
- Multi-core jobs will be detected and terminated in time, through the data flow processing module of integrated monitoring



HPC Cluster

- HPC - High Performance Computing
- Resources
  - 1 master node
  - 1 accounting & monitoring node
  - 16 login nodes
  - 125 work nodes: 2,808 CPU cores + 8 GPU cards
- SLURM as job scheduler
- Jobs (2018.1~2018.4)
  - # Jobs : ~5,300
  - CPU hours : ~3 million
- Plan of GPU servers procurement
  - 72 GPU cards: NVIDIA Tesla V100
  - Expected to be done by the end of 2018



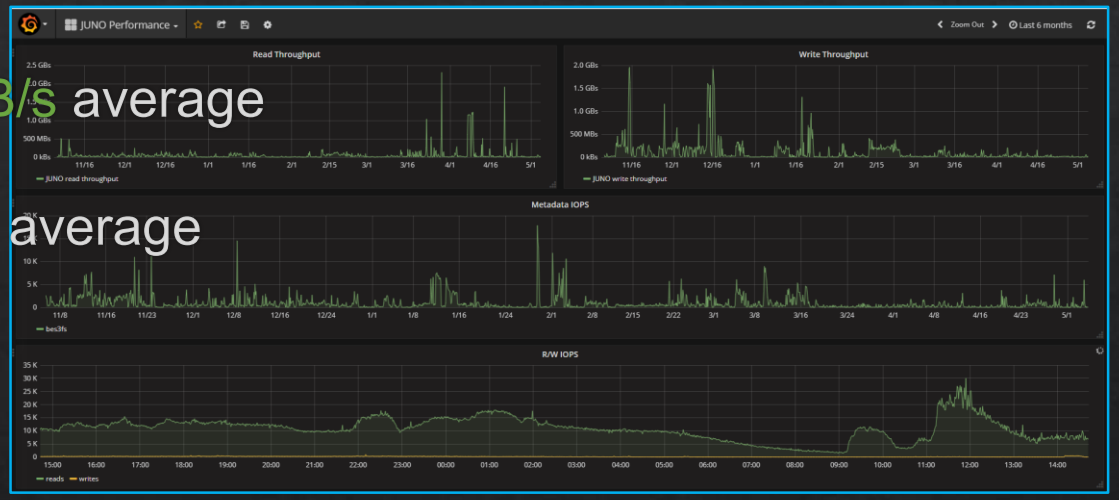


IHEP Cloud Service

- Based on Openstack Kilo
- Infrastructure As A Service
  - 14 compute nodes – **352** virtual cores
    - **329** cores are being used
    - **222** virtual machines are running
  - User Oriented Self Service
- Authentication: IHEP SSO account

/Junofs File System

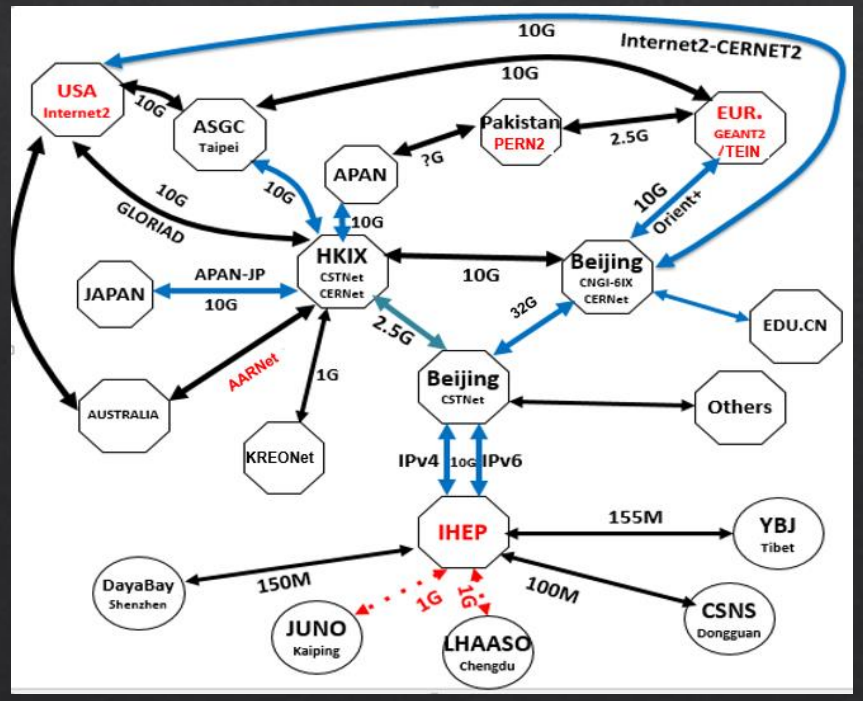
- Totally 500 TB, Used 240TB, 5 disk servers, 1 metadata server
- Performance since last meeting
  - Read throughput: 2.4 GB/s peak, 200 MB/s average
  - Write throughput: 2 GB/s peak, 200 MB/s average
  - Metadata OPS: 18K peak, 5K average
  - RW OPS: 60K peak, 20K average



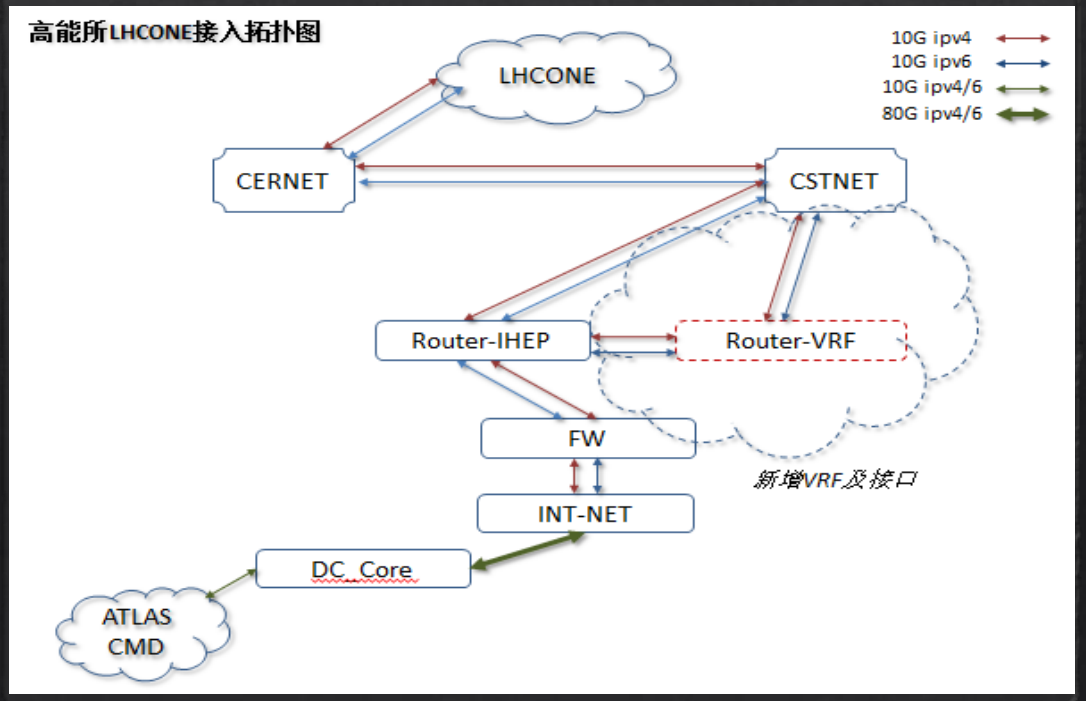
- Plan to replace metadata storage to SAN connected SSD disk array pair during summer maintenance, for better performance and reliability

Internet Connections

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

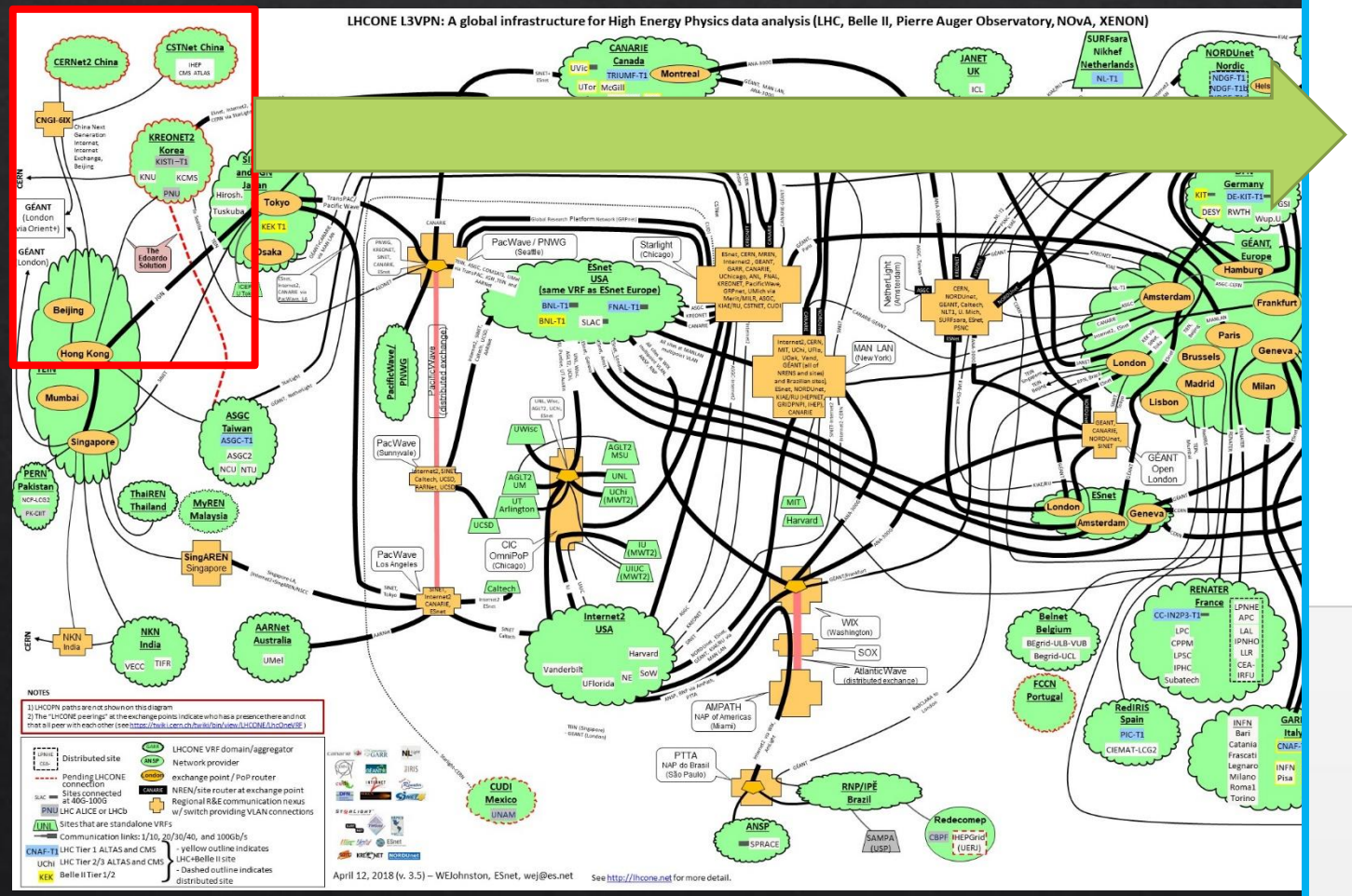


- LHCONE:
  - IPv4(10Gbps)+IPv6(10Gbps)



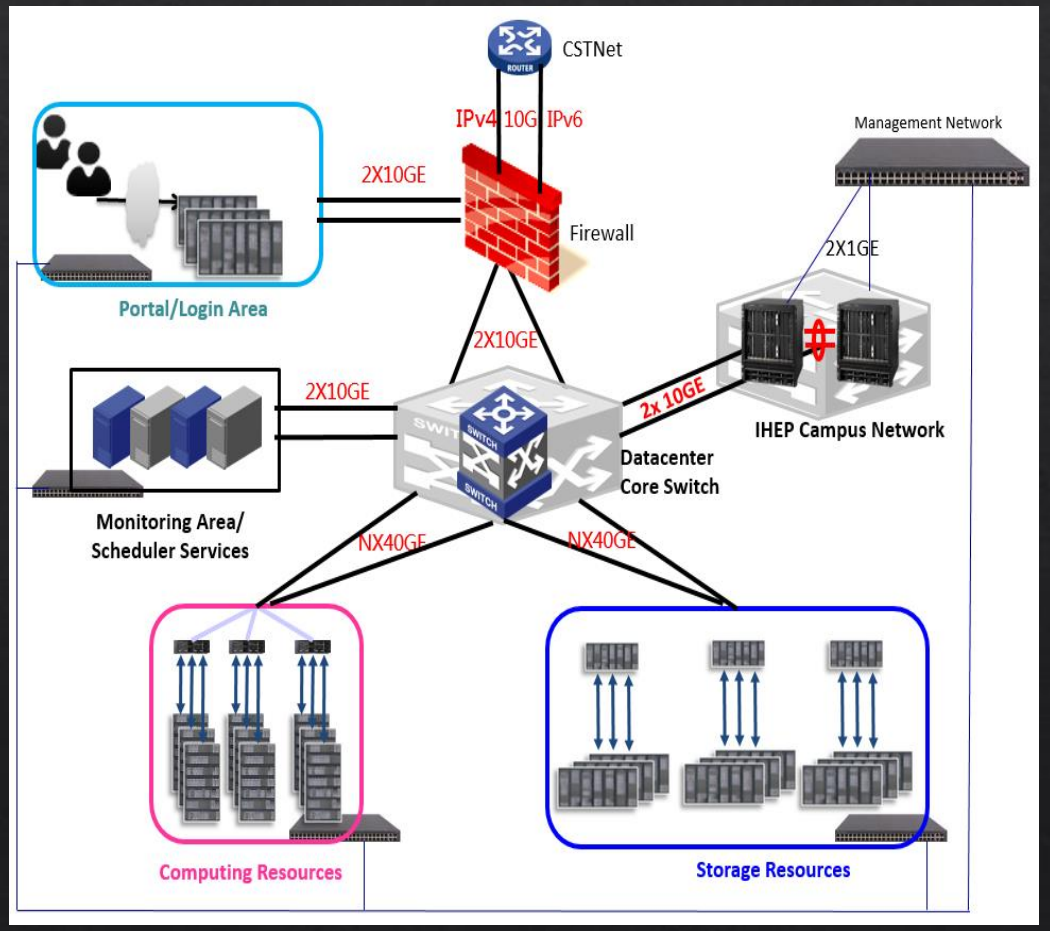
IHEP became the member of LHCONE(LHC Open Private Network)

since Mar. 2018



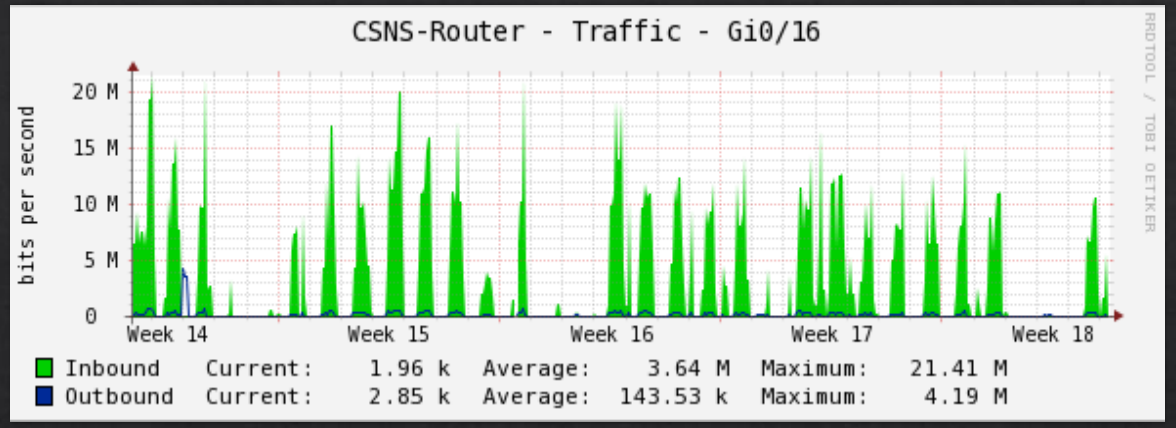
Data Center Network

- Function Area
  - Internal: Computing/Storage/AFS/DNS/Monitoring
  - DMZ: Public Servers/Login Nodes/…
  - Internet: Performance Measurement nodes
- Data Exchange Network
  - 160Gbps Backbone
  - IPv6 enabled
- Storage Network/FC
  - Now : File Server → Disk Array
  - Plan: File Server →SAN Switch →Disk Array
  - SAN Switch: Brocade 6520 with 48-port/ 16G SFP
- Management Network
  - Remote management/control
  - 1Gbps



Data transfer status for PMT testing

- Data transfer system works well
  - Data volume: 12.7TB
  - We strongly suggest an unified style of source data's file name
- A monitoring dashboard has been deployed
  - Real-time transfer status
  - Historical transfer status
    - File count
    - File size



Summary

- HTCondor works well on HTC cluster
- Job scheduler is able to cooperate with the monitoring system
- High performance computing cluster will be scaled
- New Storage architecture is undergoing
- IHEP became a member of LHCONE
- Data transfer system works well for PMT testing data

Thanks!