A RoCE-based network framework for science workloads in HEPS data center

Shan Zeng，Tao Cui，Fazhi Qi

on behalf of HEPS-CC

Funded by NSFC (No. 12175258)

zengshan@ihep.ac.cn

2024/10/23

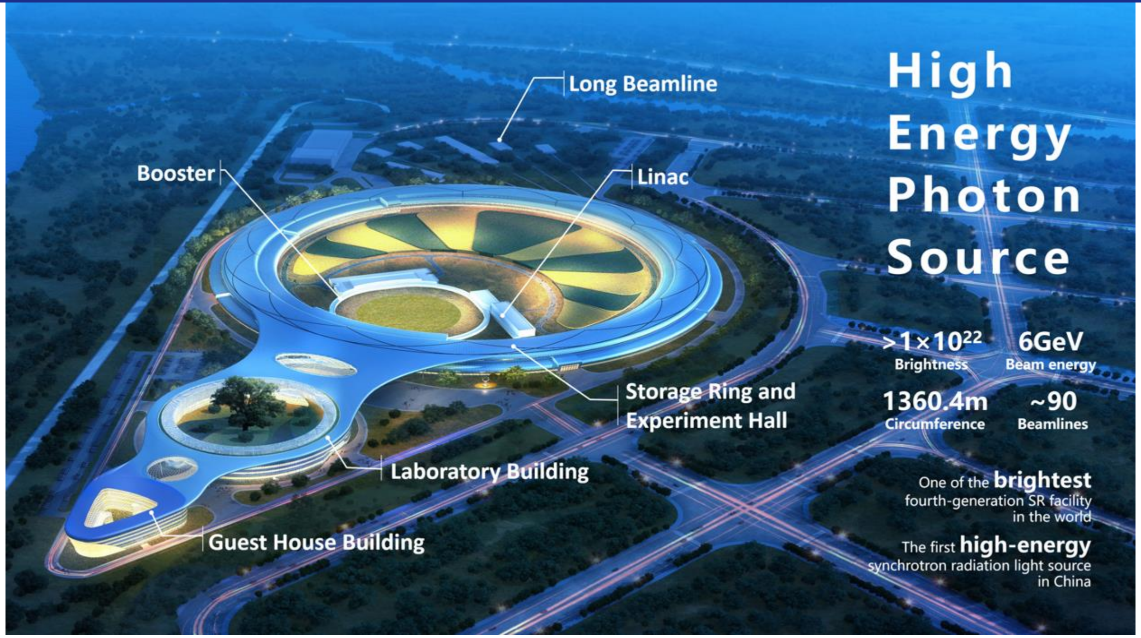
Outline

- Introduction
- Network architecture design
- Running status
- Future plan
- Summary

Overview of HEPS

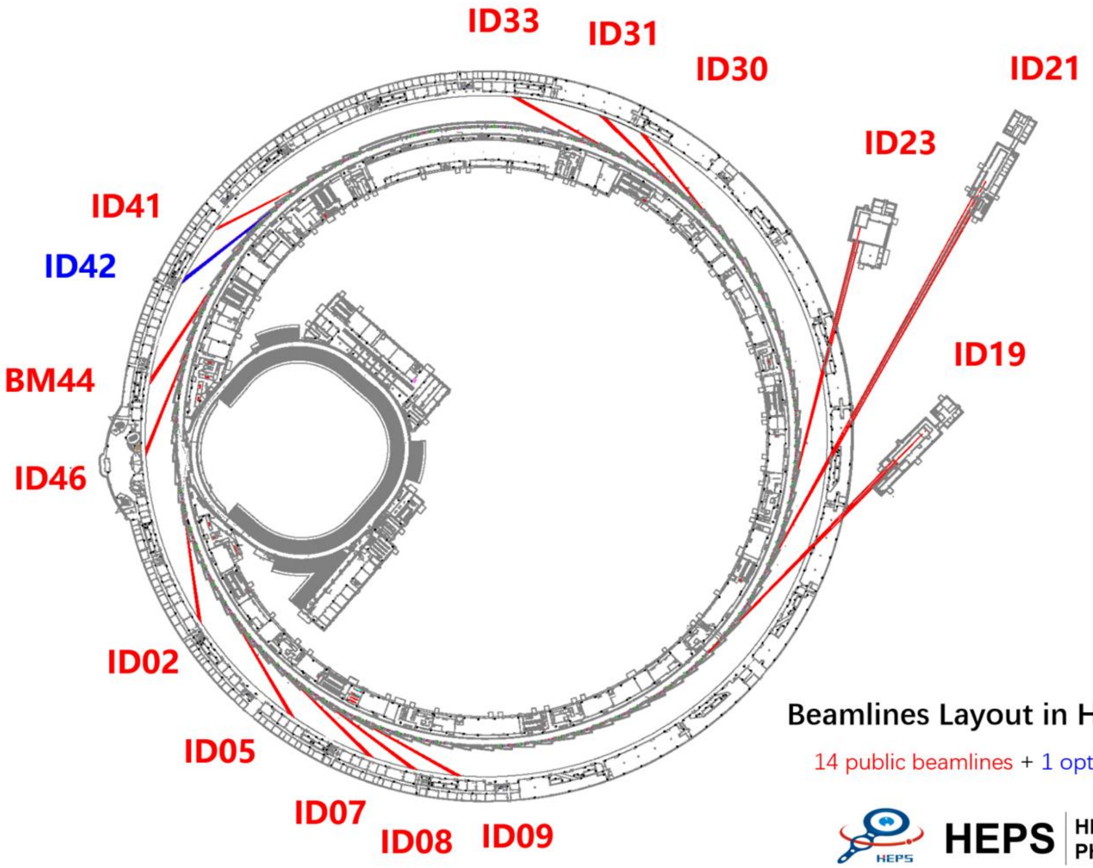
High Energy Photon Source (HEPS)

- The first 4th generation synchrotron in Asia: High energy, High brightness
- Located in Beijing - about 80KM from IHEP
- Civil construction completed in 2022
- Expected to put into use in 2025


Long Beamline
Booster
Linac
High Energy Photon Source
>1×10²² Brightness
6GeV Beam energy
1360.4m Circumference
~90 Beamlines
Storage Ring and Experiment Hall
Laboratory Building
Guest House Building
One of the brightest fourth-generation SR facility in the world
The first high-energy synchrotron radiation light source in China

Main parameters	Unit	Value
Beam energy	GeV	6
Circumference	m	1360.4
Emittance	pm·rad	< 60
Brightness	phs/s/mm^2/mrad^2/0.1%BW	>1x10^{22}
Beam current	mA	200

HEPS Beamlines in Phase I



Beamlines Layout in HEPS phase I

14 public beamlines + 1 optics test beamline

HEPS HIGH ENERGY PHOTON SOURCE

Microfocusing X-Ray Protein Crystallography-ID02 Beamline

Low-Dimensional Structure Probe Beamline-ID05

Engineering Materials Beamline-ID07

Hard X-Ray Coherent Scattering Beamline-ID09

Pink Beam SAXS Beamline-ID08

Hard X-Ray Nanoprobe Multimodal Imaging-ID19 Beamline

Hard X-Ray Imaging Beamline-ID21

Structural Dynamics Beamline-ID23

ID30-Transmission X-Ray Microscopic Beamline

ID31-High Pressure Beamline

ID33-Hard X-Ray High Resolution Spectroscopy Beamline

BM44-Tender X-Ray Beamline

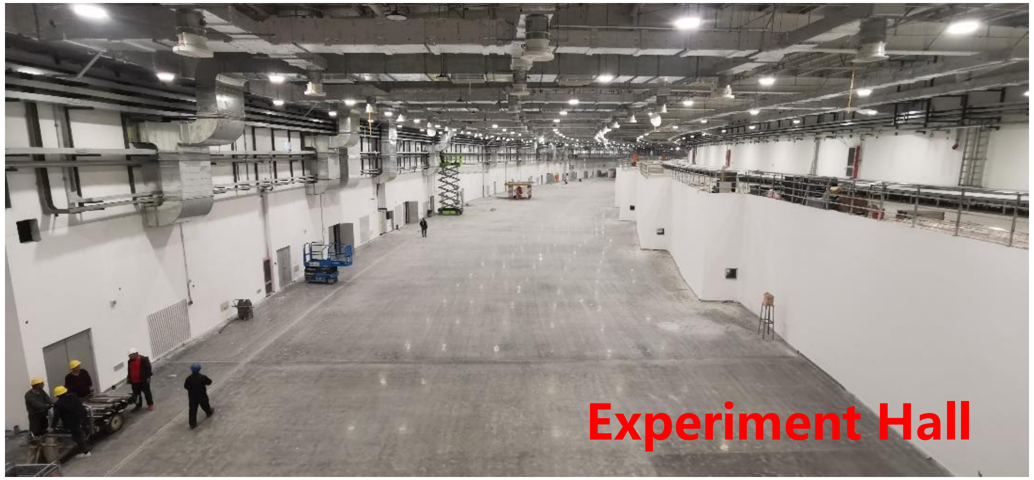
ID41-High Resolution Nanoscale Electronic Structure Spectroscopy Beamline

ID42-Optics Test Beamline

ID46-X-Ray Absorption Spectroscopy Beamline

14 public beamlines + 1 optics test beamline in Phase I

Can accommodate over 90 beamlines in total



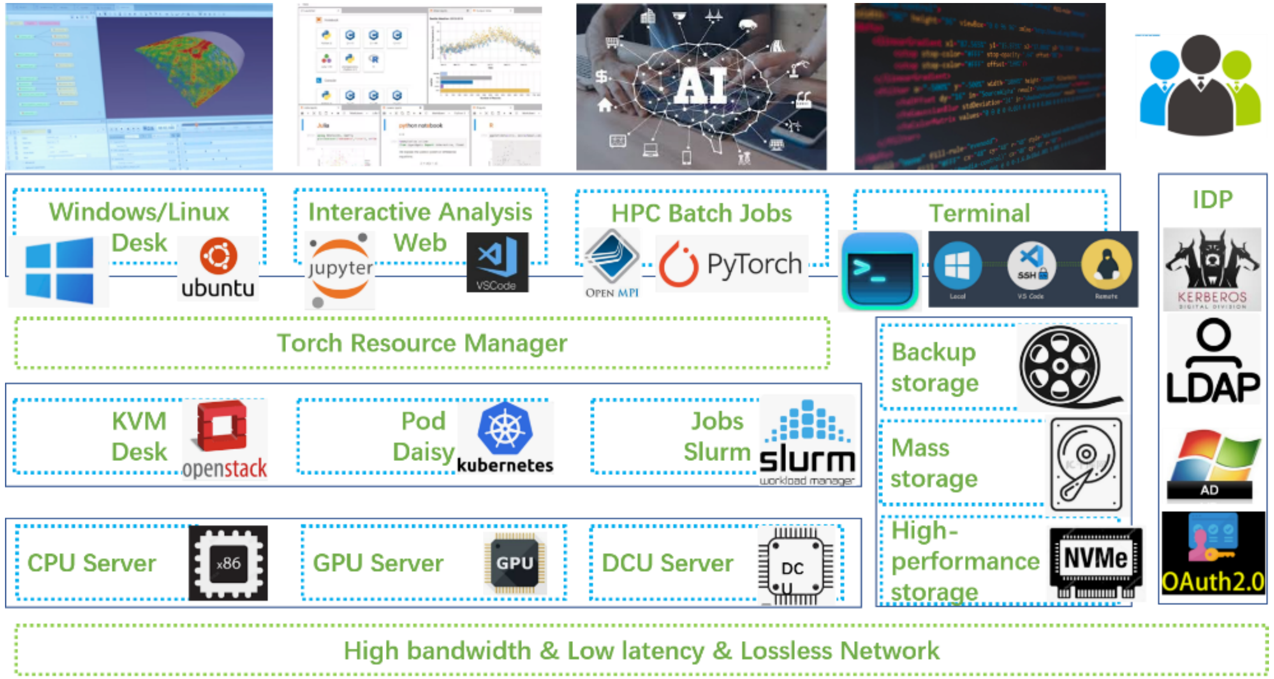
HEPS Data Center

HEPS DC computer room space

- The total area is approximately 900㎡. The main computer room is 530㎡, the UPS room is 165㎡, and the tape library is 155㎡.
- The maximum planned layout is approximately 130 cabinet positions

HEPS computing platform was designed to satisfy the complex data analysis requirements in the field of synchrotron radiation

- Openstack
  - provide users with remote desktop access services
- Kubernetes
  - manages container clusters, and starts container images with multiple methodological software according to user analysis requirements
- Slurm
  - provide HPC computing services and meet offline data analysis need



Windows/Linux Desk

Interactive Analysis Web

HPC Batch Jobs

Terminal

IDP

Torch Resource Manager

KVM Desk

Pod Daisy

Jobs Slurm

Backup storage

Mass storage

High-performance storage

CPU Server

GPU Server

DCU Server

High bandwidth & Low latency & Lossless Network

See Hu Qingbao's talk

HEPS Data Center Network Challenges

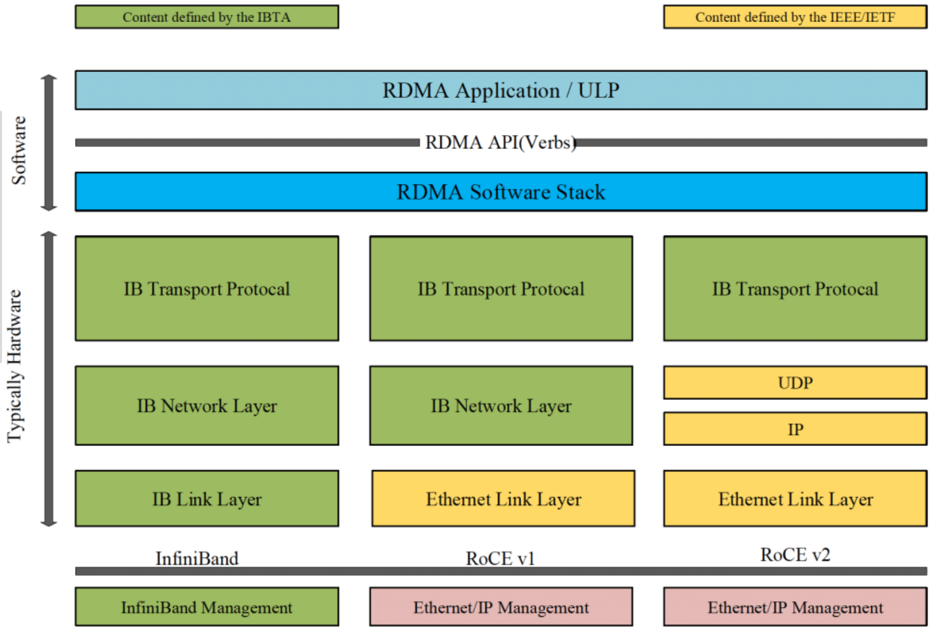
- **During phase I, HEPS will produce more than 300PB/year of raw data, requiring high performance in network to assure data moving and analysis**
  - High bandwidth
  - Lower latency
- **AI applications will be deployed in some beamlines, requiring lossless feature in network**
- **Different construction phases have different numbers of beamlines, requiring the network to provide expandable capabilities**
- **As a remote site of IHEP, to reduce labor costs, intelligent operation and maintenance is also a key issue that needs to be considered.**

RDMA technologies: RoCE vs IB

- **RDMA is a technology that allows servers in a network to exchange data in main memory without involving the processor, cache or operating system of either server, which can provide high bandwidth and low latency**

IB stands for InfiniBand. It is a high-performance computer networking technology used in data centers and high-performance computing environments. It offers low latency and high bandwidth for applications that require fast data transfer and communication between servers and storage systems

RoCE is a network protocol defined in the InfiniBand Trade Association (IBTA) standard, allowing RDMA over converged Ethernet network. Shortly, it can be regarded as the application of RDMA technology in hyper-converged data centers, cloud, storage, and virtualized environments.

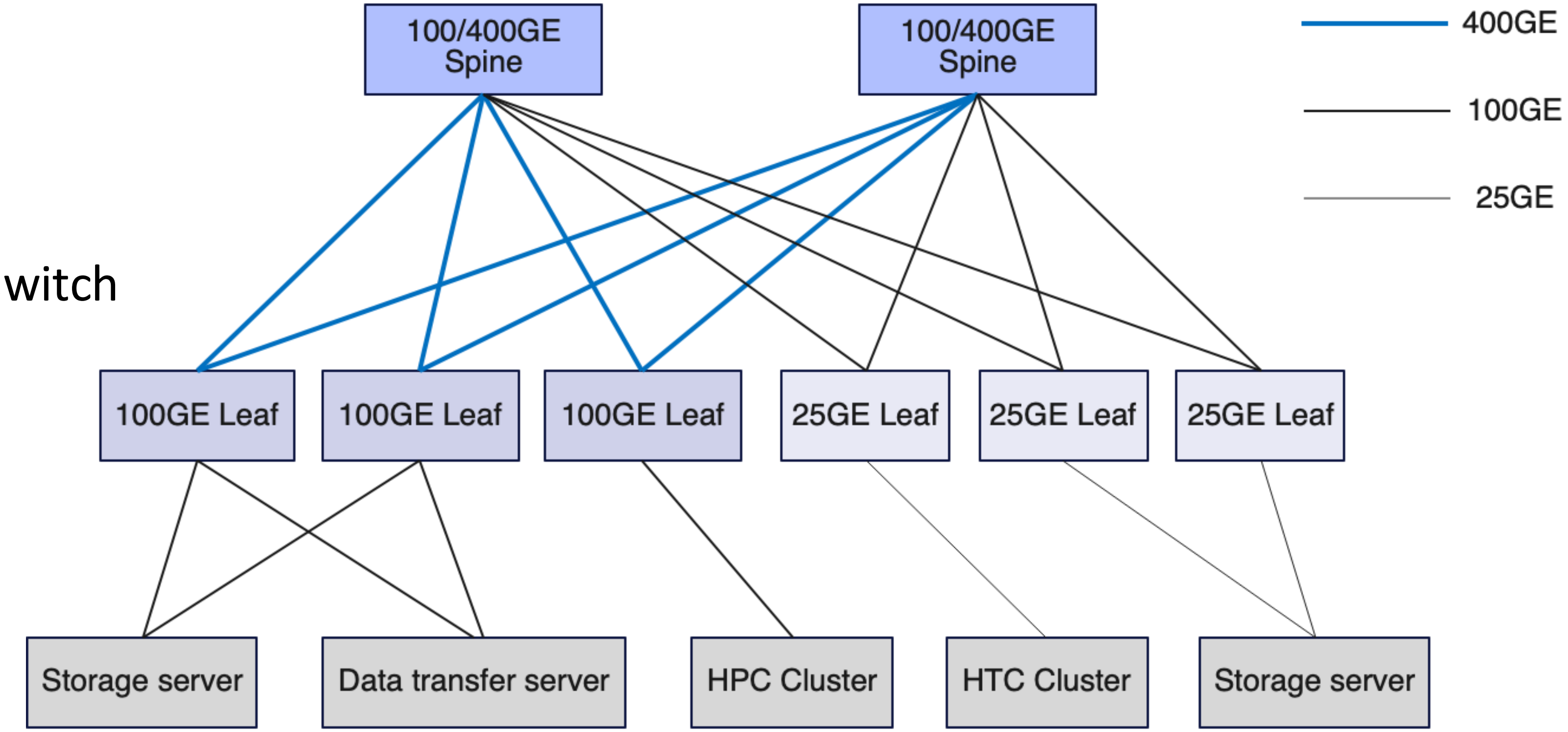


Underlying ISO Stacks of the Flavors of RDMA

	Performance	Cost	Scalability	Compatibility
RoCE	Higher latency especially in large scales	Lower	Performance may be affected in large scales	can be integrated with existing Ethernet networks, easier to deploy
IB	lower latency	Higher	can support thousands of nodes	requires a dedicated network

Network Architecture Design

- **Concerning about the scale, cost and compatibility, we designed a RoCE-based DC network**
- **Support the mixed running of RoCE and traditional TCP**
- **Spine-Leaf architecture**
  - Easy to scale out
  - Convergence ratio is 1:1
  - Gateway for each server is on Leaf switch



- **Performance test**
  - Bandwidth test is perfect
  - Latency is acceptable
  - details refer to our paper in CHEP2021

https://doi.org/10.1051/epjconf/202125102018

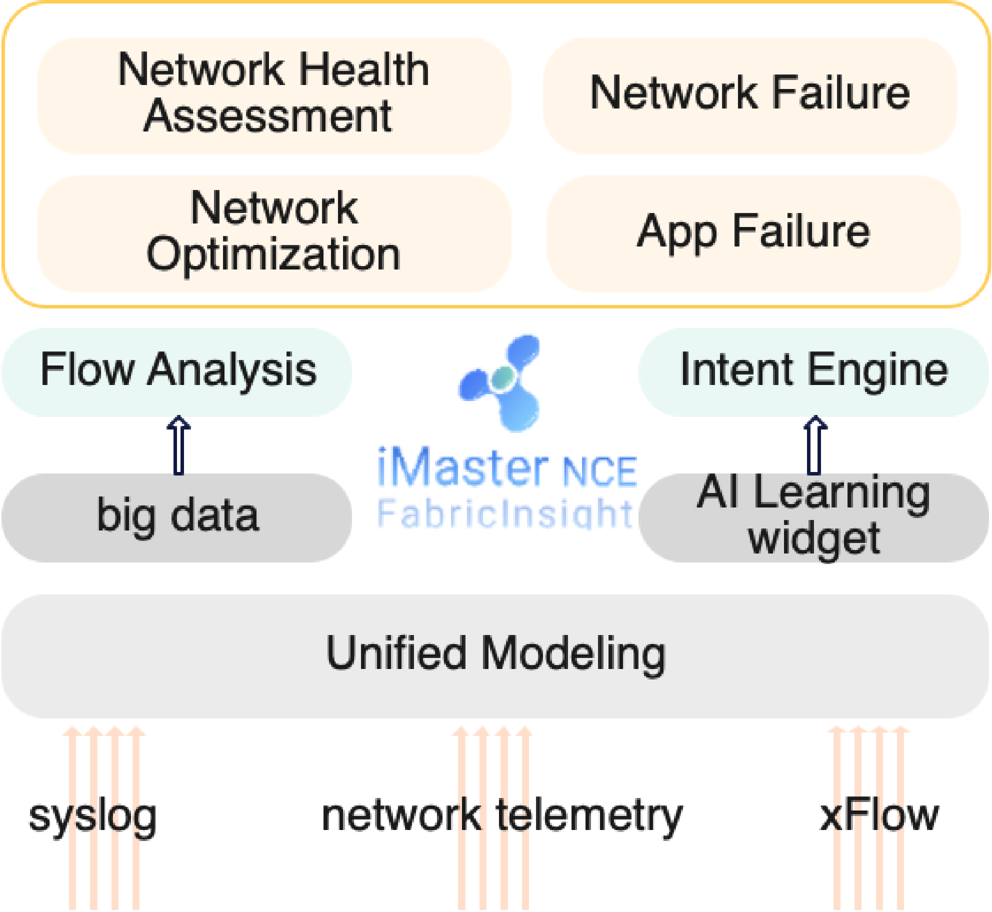
Network Monitoring

What we concerned

- When failures happened?
- What kinds of failure they are?
- How we can handle/optimize them?

Monitoring technologies

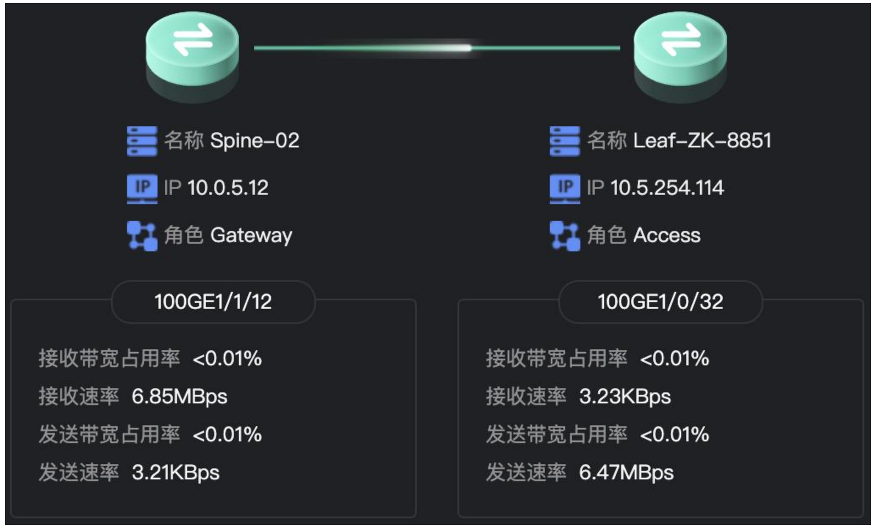
- Network data capture technology
  - syslog
  - network telemetry
  - xFlow
- Create an intelligent brain to analyze the big data
  - Flow analysis
  - AI learning to produce an intent engine
- Provide a network monitoring service
  - Network health assessment
  - Network failures report
  - Application failures report
  - Network optimization suggestions



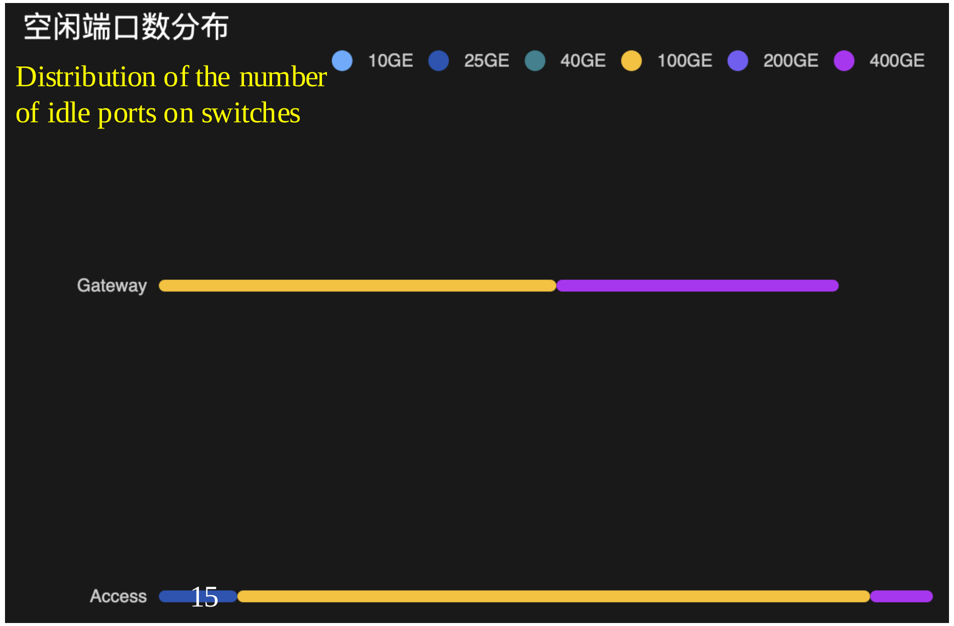
Running Status

- The HEPS Data Center Network was put into use in October 2023, and has been running stably

- Online devices
  - 8 switches, 697 ports, 339 optical modules
  - Provide 10G/25G/100G/400G access abilities

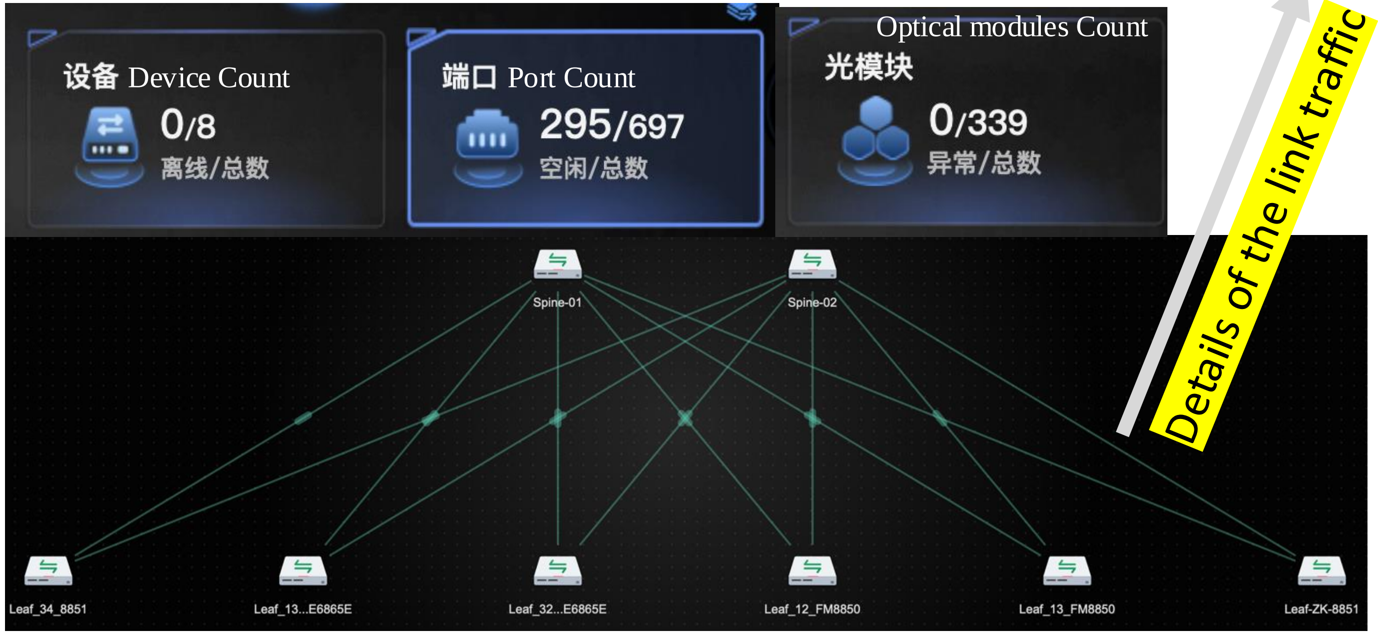


名称 Spine-02	名称 Leaf-ZK-8851
IP 10.0.5.12	IP 10.5.254.114
角色 Gateway	角色 Access
100GE1/1/12	100GE1/0/32
接收带宽占用率 <0.01%	接收带宽占用率 <0.01%
接收速率 6.85MBps	接收速率 3.23KBps
发送带宽占用率 <0.01%	发送带宽占用率 <0.01%
发送速率 3.21KBps	发送速率 6.47MBps



空闲端口数分布

Distribution of the number of idle ports on switches



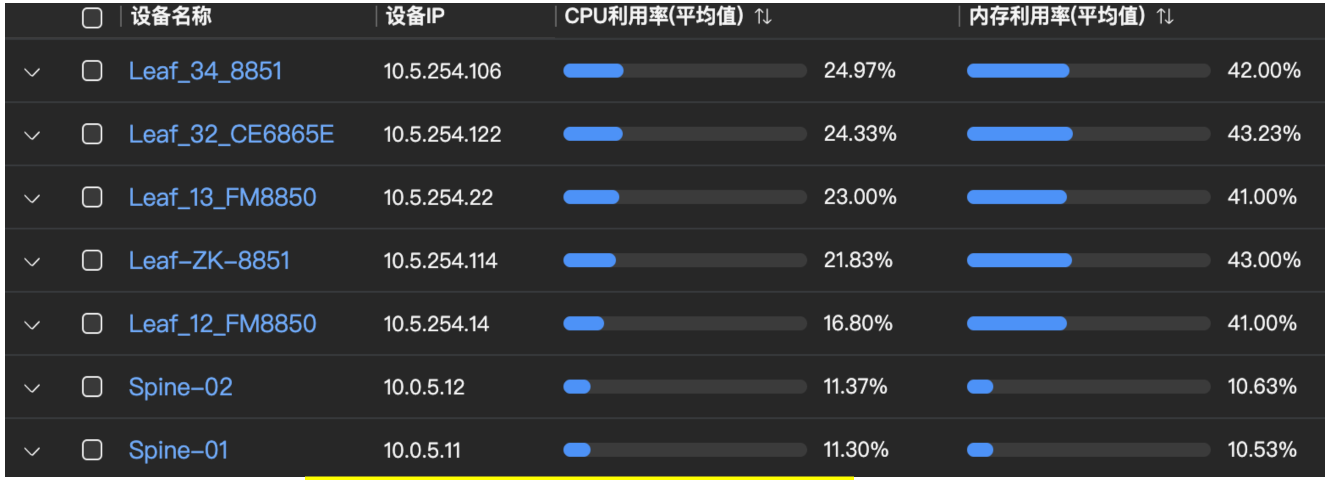
设备 Device Count
0/8
离线/总数

端口 Port Count
295/697
空闲/总数

Optical modules Count
光模块
0/339
异常/总数

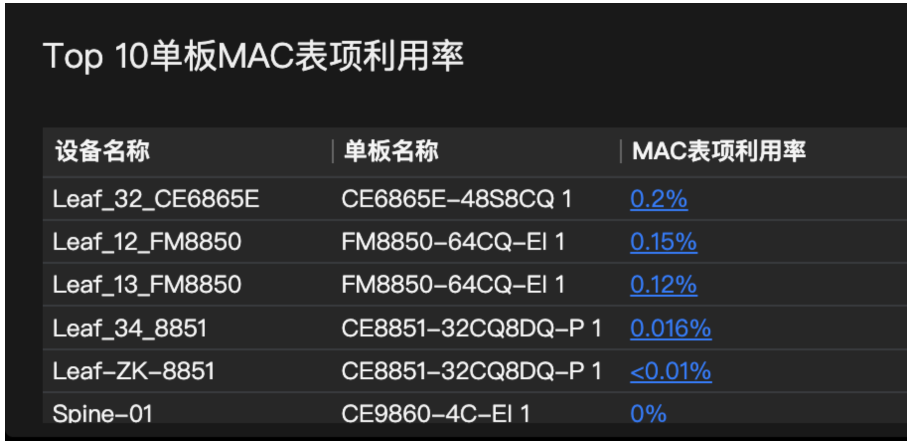
Details of the link traffic

Monitoring Statistics



	□	设备名称	设备IP	CPU利用率(平均值) ⇅	内存利用率(平均值) ⇅
⌄	□	Leaf_34_8851	10.5.254.106	24.97%	42.00%
⌄	□	Leaf_32_CE6865E	10.5.254.122	24.33%	43.23%
⌄	□	Leaf_13_FM8850	10.5.254.22	23.00%	41.00%
⌄	□	Leaf-ZK-8851	10.5.254.114	21.83%	43.00%
⌄	□	Leaf_12_FM8850	10.5.254.14	16.80%	41.00%
⌄	□	Spine-02	10.0.5.12	11.37%	10.63%
⌄	□	Spine-01	10.0.5.11	11.30%	10.53%

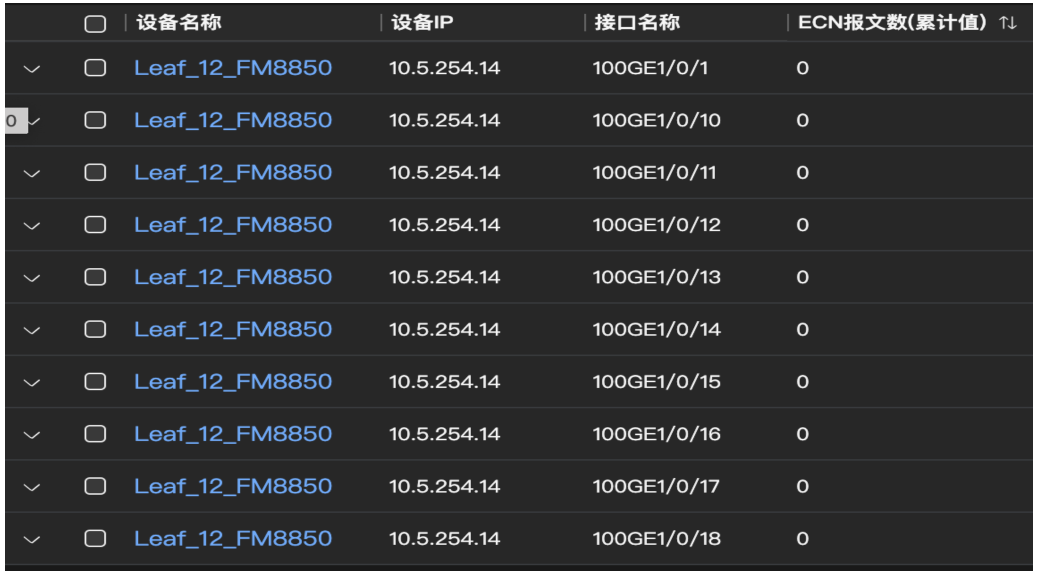
CPU/Memory usage of switches



Top 10单板MAC表项利用率

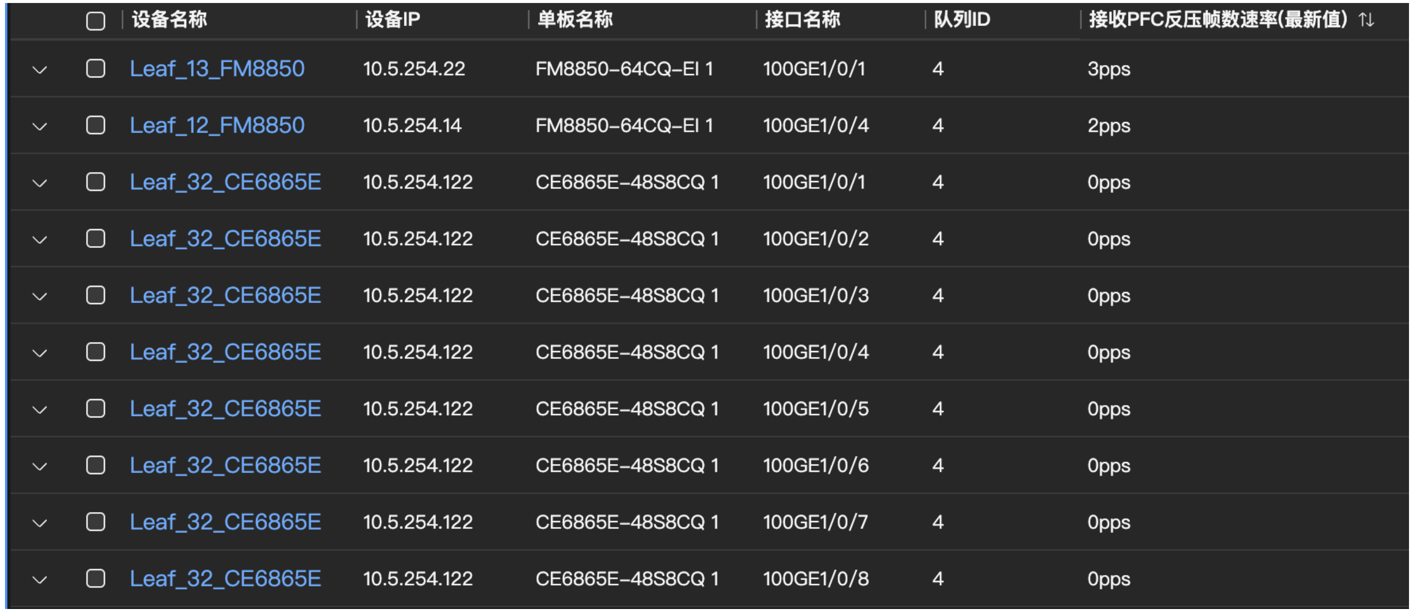
设备名称	单板名称	MAC表项利用率
Leaf_32_CE6865E	CE6865E-48S8CQ 1	0.2%
Leaf_12_FM8850	FM8850-64CQ-EI 1	0.15%
Leaf_13_FM8850	FM8850-64CQ-EI 1	0.12%
Leaf_34_8851	CE8851-32CQ8DQ-P 1	0.016%
Leaf-ZK-8851	CE8851-32CQ8DQ-P 1	<0.01%
Spine-01	CE9860-4C-EI 1	0%

Top10 MAC table usage



	□	设备名称	设备IP	接口名称	ECN报文数(累计值) ⇅
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/1	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/10	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/11	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/12	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/13	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/14	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/15	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/16	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/17	0
⌄	□	Leaf_12_FM8850	10.5.254.14	100GE1/0/18	0

ECN count of each Port



	□	设备名称	设备IP	单板名称	接口名称	队列ID	接收PFC反压帧数速率(最新值) ⇅
⌄	□	Leaf_13_FM8850	10.5.254.22	FM8850-64CQ-EI 1	100GE1/0/1	4	3pps
⌄	□	Leaf_12_FM8850	10.5.254.14	FM8850-64CQ-EI 1	100GE1/0/4	4	2pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/1	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/2	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/3	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/4	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/5	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/6	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/7	4	0pps
⌄	□	Leaf_32_CE6865E	10.5.254.122	CE6865E-48S8CQ 1	100GE1/0/8	4	0pps

PFC count of each RoCE Queue

Future Plan

- More switches will be added for providing the access ability of 25GE/100GE nodes
- Automatically alarm of data center network problems will be developed
- Monitoring data will be considered to be called by 3rd party applications through RESTful API to develop more fancy monitoring dashboard

Summary

- HEPS data center network is designed based on RoCE
- It works fine since it launched in October 2023
- More services will be in production, and we will keep a close eye on the network performance and monitoring metrics
- Any suggestions and cooperation are welcomed

Thanks for your attentions

Questions, Comments, Suggestions?