The Design of Networking and Computing System for High Energy Photon Source

Fazhi QI, Qiulan Huang, Jingyan Shi, Lu Wang, Hao Hu, Haolai Tian, Hongmei Zhang, Yu Hu, Shan Zeng, Yanming Wang, Qingbao Hu, Haifeng Zhao

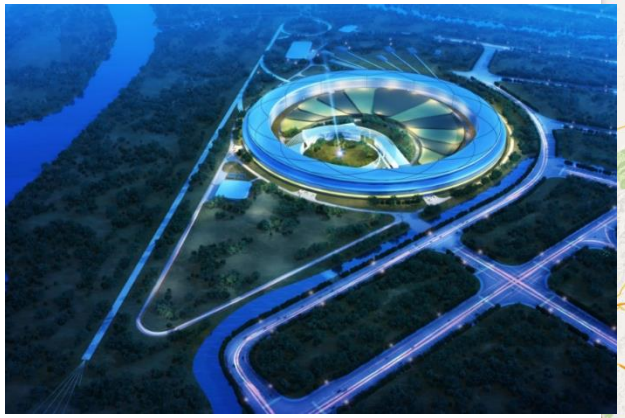
Institute of High Energy Physics, CAS

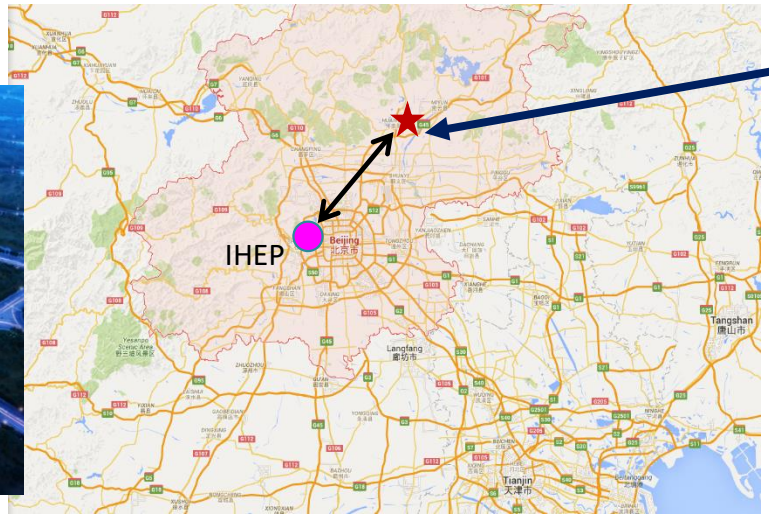
Outline

HEPS: High Energy Photon Source

- New light source in China — High energy, high brightness
- Located in Beijing - about 80KM from IHEP
- Officially approved in Dec. 2017, the construction was started at the end of 2018 and will be completed in 2024
- The whole project will be finished in mid-2025 after commissioning

A new photon science research center at the northern China



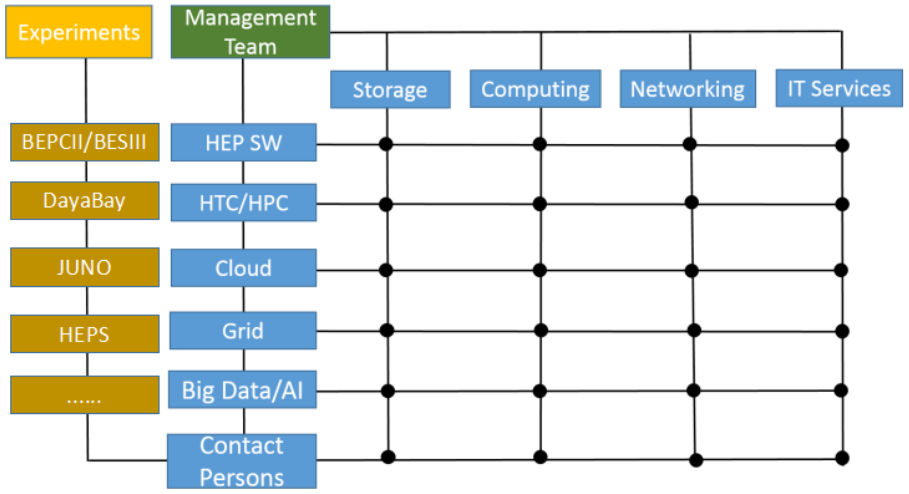


About 80 km from IHEP

Main parameters	Unit	Val
Beam energy	GeV	6
Circumference	m	1360
Emittance	pm·rad	< 60
Brightness	phs/s/mm²/mrad²/0.1%BW	>10²
Beam current	mA	200
Injection		Top-

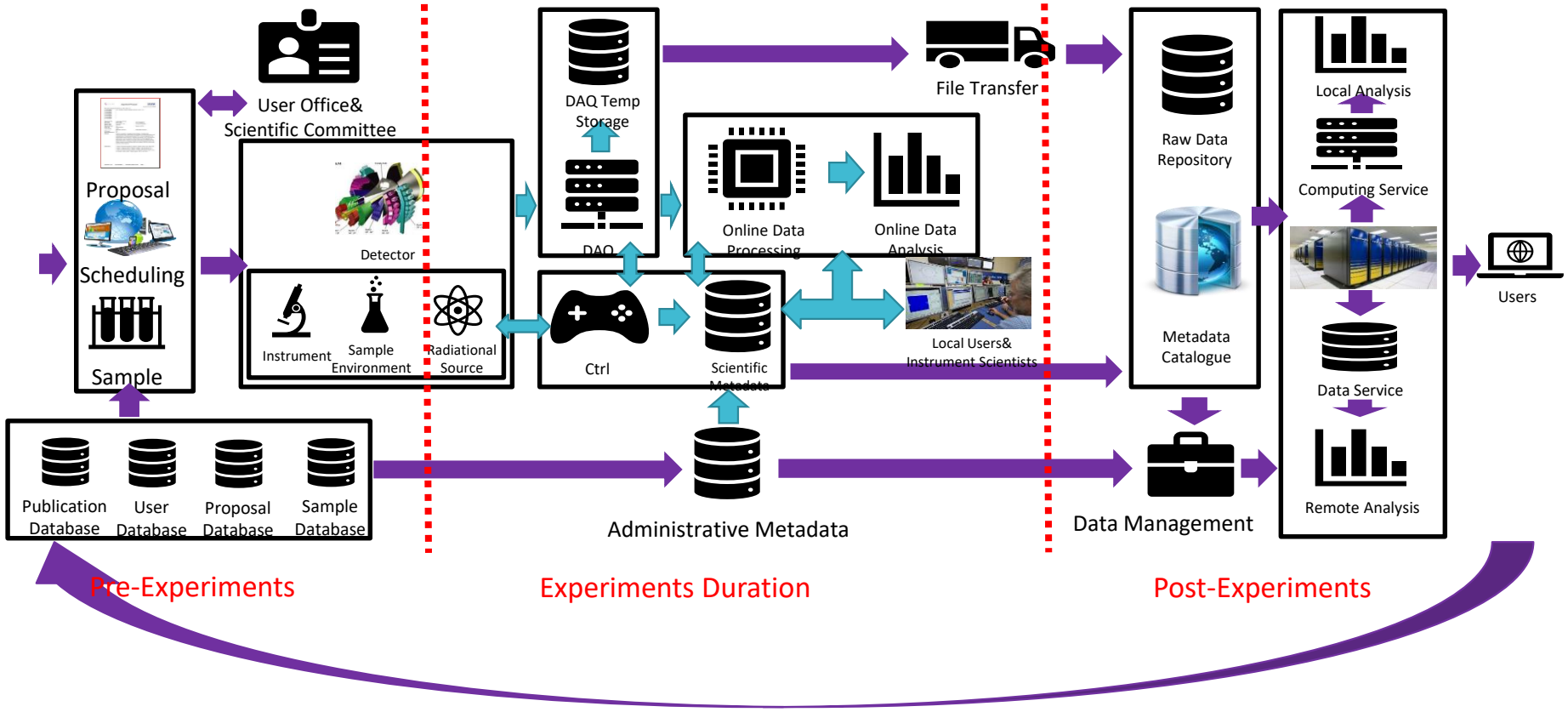
HEPS CC: the Computing & Communication System for HEPS

- 30+ members
  - But all the people are part time for HEPS
  - Most of the members are coming from Computing Center (IHEP CC)
  - 3 from CSNS/Computing and Software group
  - 1 from Beamline
- 7 work groups are set up according to the tasks
  - Infrastructure , Network, Computing & Storage, Scientific Software, Database & Public Service , Monitoring, Security.
- Matrix management
  - Across Group Boundaries and Experiments
  - Sharing talents and skills



Outline

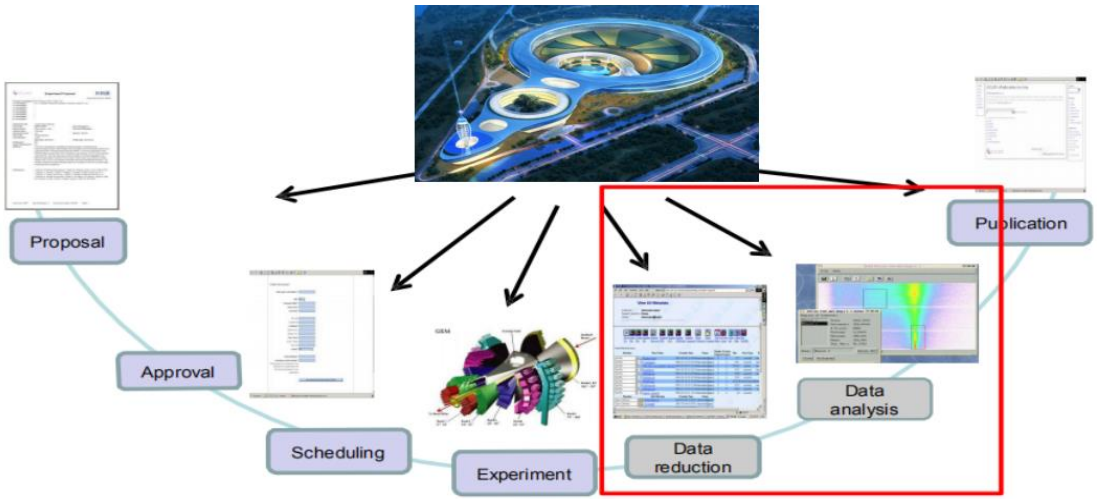
IT services & Beamline experiments



IT services are needed during the whole life-cycle of the Beamline experiments

HEPS CC Missions

- Provision of scientific data and user services for HEPS
  - Infrastructure
  - Network
  - Computing
  - Storage
  - Data Management
  - Scientific Software
  - Public Software and Services

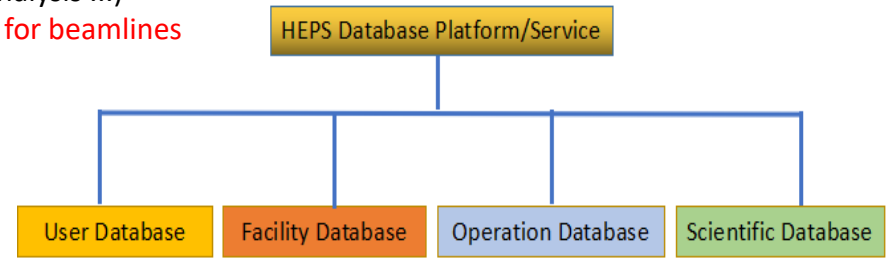


- Research on open IT technologies related to HEPS

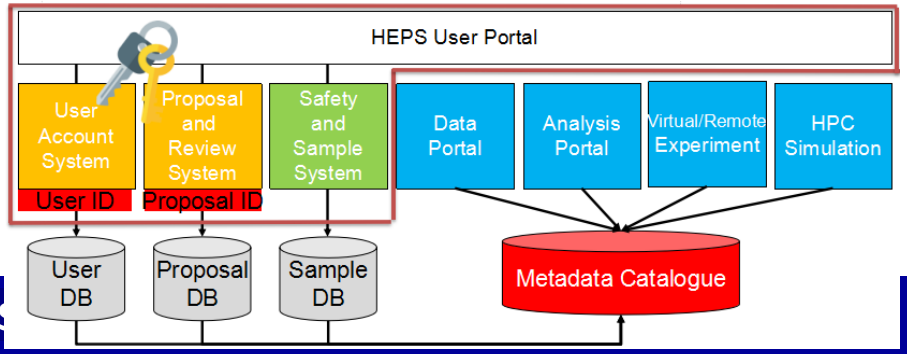
Challenges

- Large volume of Data
  - High I/O throughput/read(max: 6.94GB/s) & write speed(max:15GB/s)
  - Hierarchical storage management : Beamline Disk→ Central Disk→ Tape
  - High capability: Long-term data preservation
- Big Data Management
  - Metadata Catalogue
  - Scientific Data Management
- Fast Data Analysis
  - Simulation / For accelerator design
  - Online / For beamlines
  - Offline / For users and in-house scientists (Data Reconstruction, Analysis ...)
  - Capability / The more the better / At least meet the requirements for beamlines
- Scientific software
  - Scientific Software Framework
  - Layered and modularized software platform for data processing
  - Build or be involved in a self-sustaining software ecosystem
- Public & Data Service
  - Collaboration tools: Project , Sharepoint , Gitlab , Video conference
  - User management and authentication
  - Experiment process management / services
  - User interface for experiments, data access/sharing and analysis

Beam Line	Peak data Volume Per Day (TB)	Mean Data Volume Per Day (TB)	Capacity of Beam Line Storage (TB)	Write Performance of Beam Line Storage (GB/s)	Read Performance of Beam Line Storage (GB/s)	Read Performance of online Analysis (GB/s)	Capactiy of central storage (PB)
B1	600.00	200.00	200.00	5.56			17.58
B2	500.00	200.00	200.00	5.56	5.56		17.58
B3	8.00	3.00	3.00	0.80	0.08		0.26
B4	10.00	3.00	3.00	1.00	0.08	1.00	0.26
B5	10.00	1.00	1.00	0.10	0.03	0.10	0.09
B6	2.00	1.00	1.00	0.10	0.03		0.09
B7	1000.00	250.00	250.00	15.00	6.94		21.97
B8	80.00	10.00	10.00	1.00	0.28	0.10	0.88
B9	20.00	5.00	5.00	0.50	0.14		0.44
BA	35.00	10.00	10.00	1.00	0.28		0.88
BB	400.00	50.00	50.00	10.00	1.39		4.39
BC	1.00	0.20	0.20	0.10	0.01		0.02
BD	10.00	1.00	1.00	1.00	0.03		0.09
BE	25.00	11.20	11.20	2.00	0.31	2.00	0.98
BF	1000.00	60.00	60.00	5.00	1.67		
Total			805.4	48.72	14.84	3.1	65.51



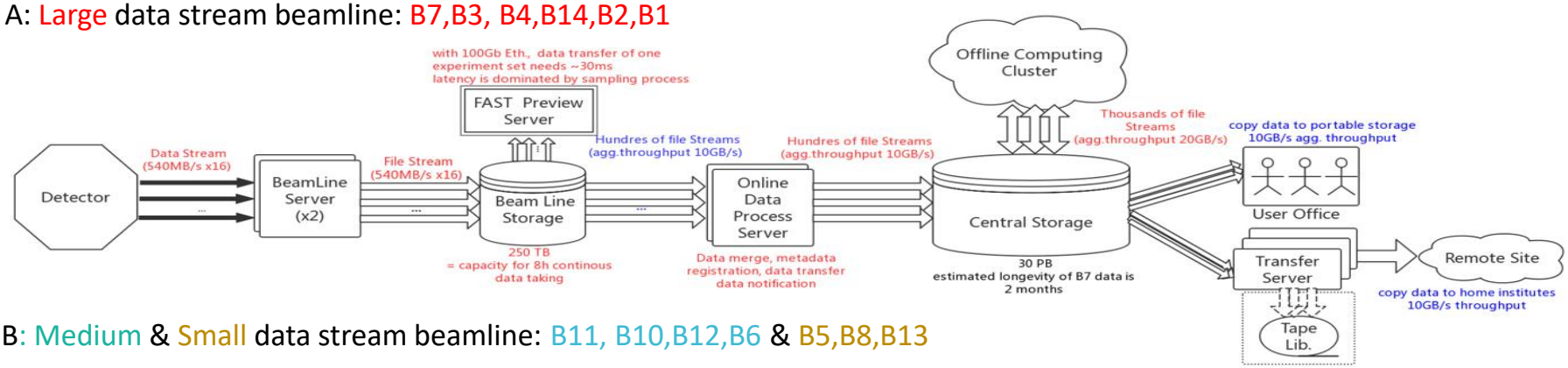
Beamline	Computational capacity demanded
B3	HPC, No parallel needed
B4	CPU >200 core, RAM >512GB, parallel
B7	Parallel >2000
BA	CPU >32 core, RAM >128GB, no parallel needed
BE	Parallel >240, RAM >768GB
合计	CPU >2472 core, RAM >1408GB, parallel



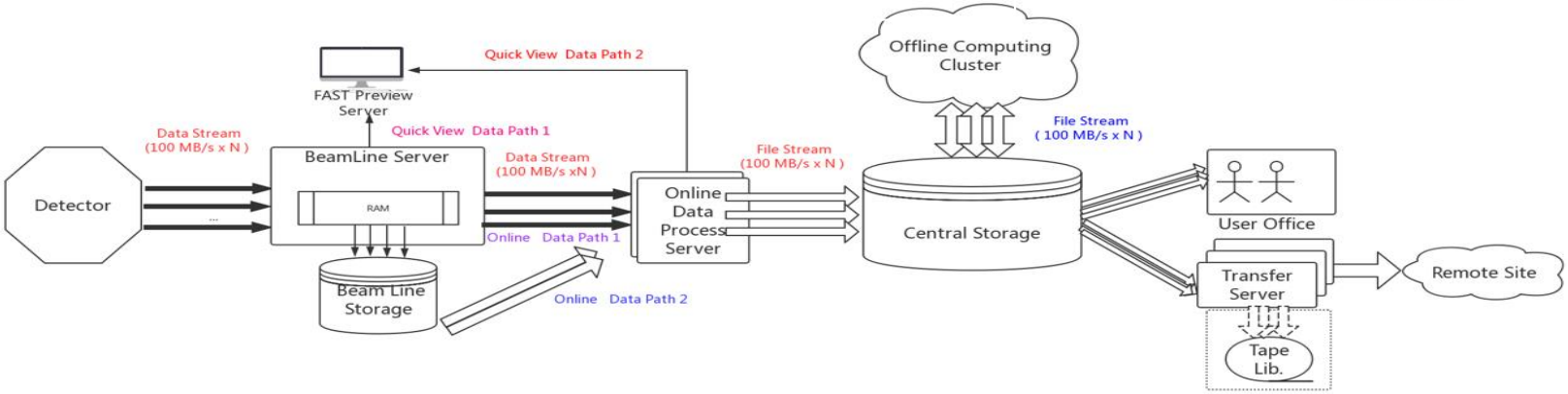
Outline

Different IT architectures for different beamline

Type A: Large data stream beamline: B7,B3, B4,B14,B2,B1



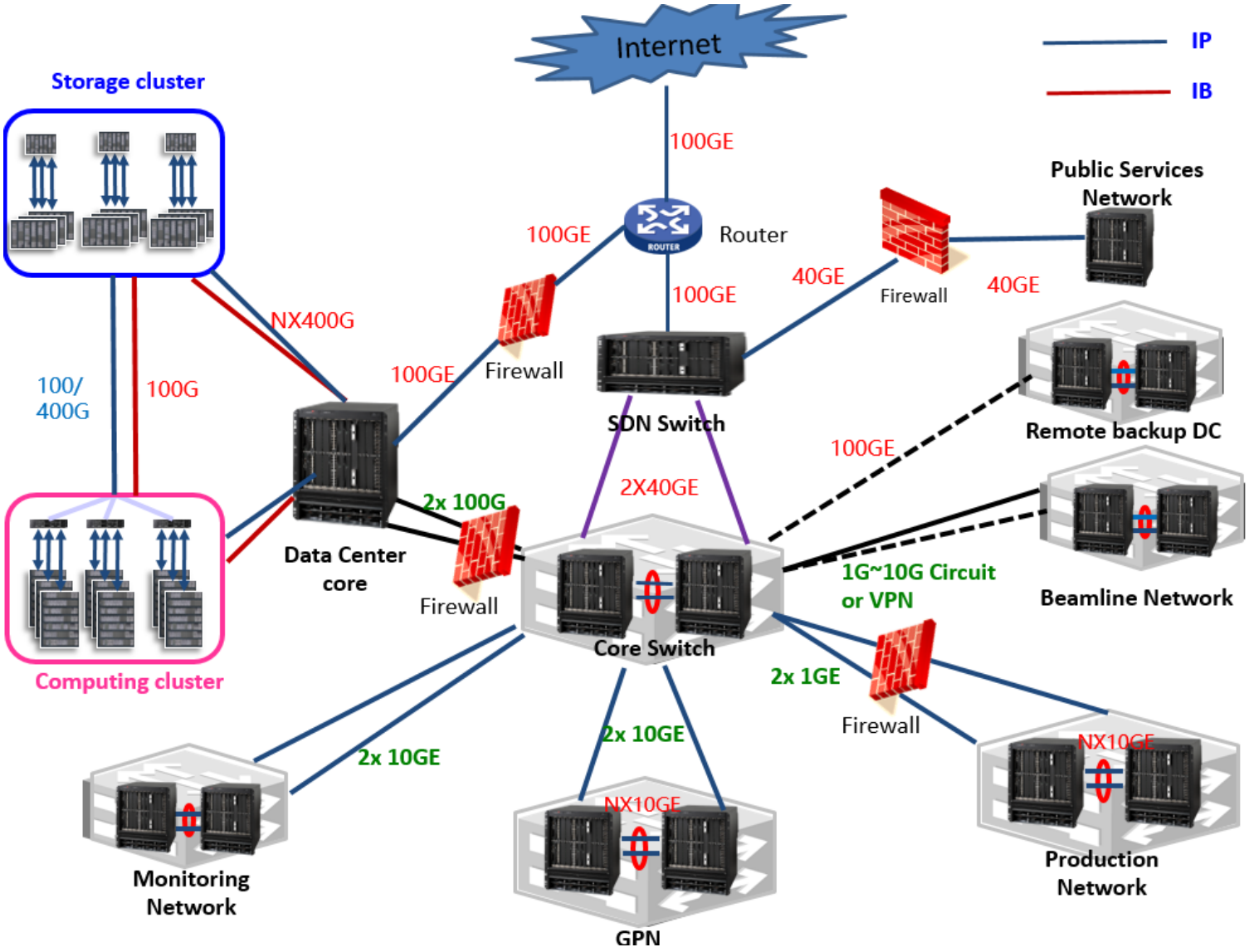
Type B: Medium & Small data stream beamline: B11, B10,B12,B6 & B5,B8,B13



System Design : Network

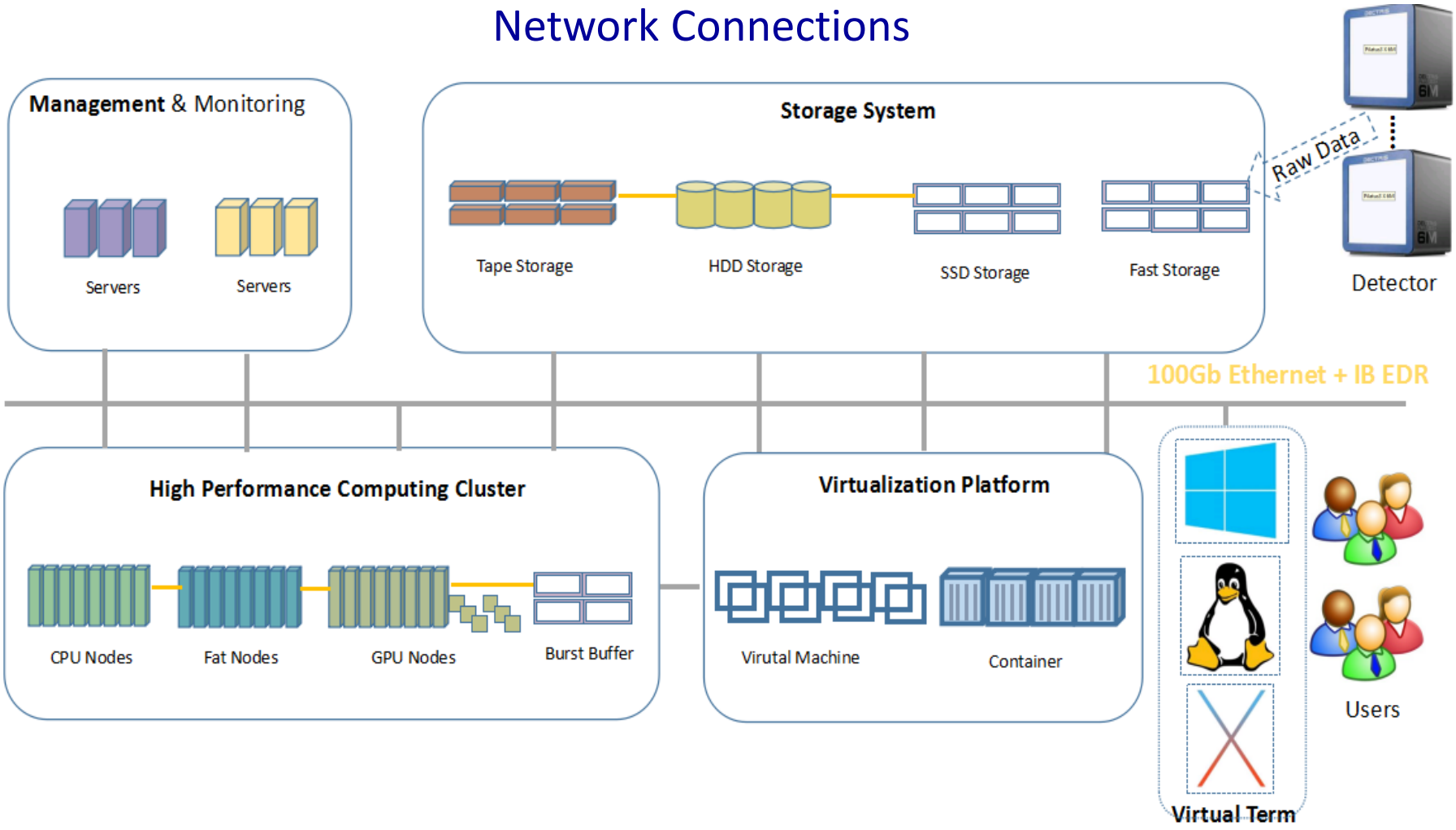
- General Purpose Network (GPN)
  - HEPS campus staff and visitors/Users
- Data Center Network (DCN)
  - Support both IB and IP connections between
    - Compute nodes, Storage nodes, Supporting nodes
  - Management network for DC servers and clusters
- Production Network
  - Network related to beamlines
    - Transfer the experiment data from beamlines to DCN
  - Accelerator/Beamlines control network (Deployed by Control System)
  - Remote Experiment
- WAN
  - Internet access for users
  - Dedicated virtual network to provide a high performance scientific data sharing with the collaborators

Network Overview



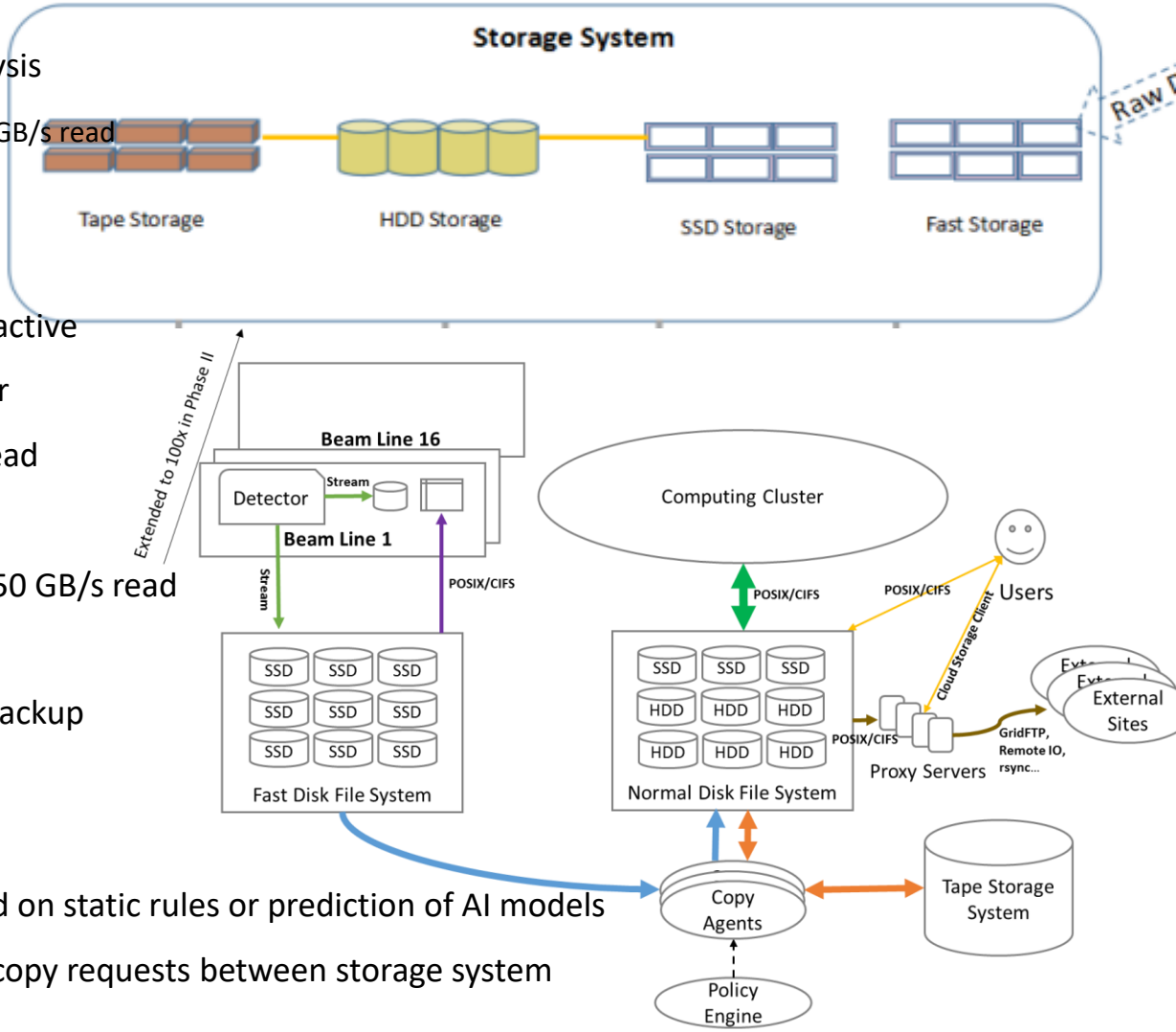
System Design : Computing & Storage Overview

Storage Area, Computing Area, User Interface / Area, Management Area Network Connections

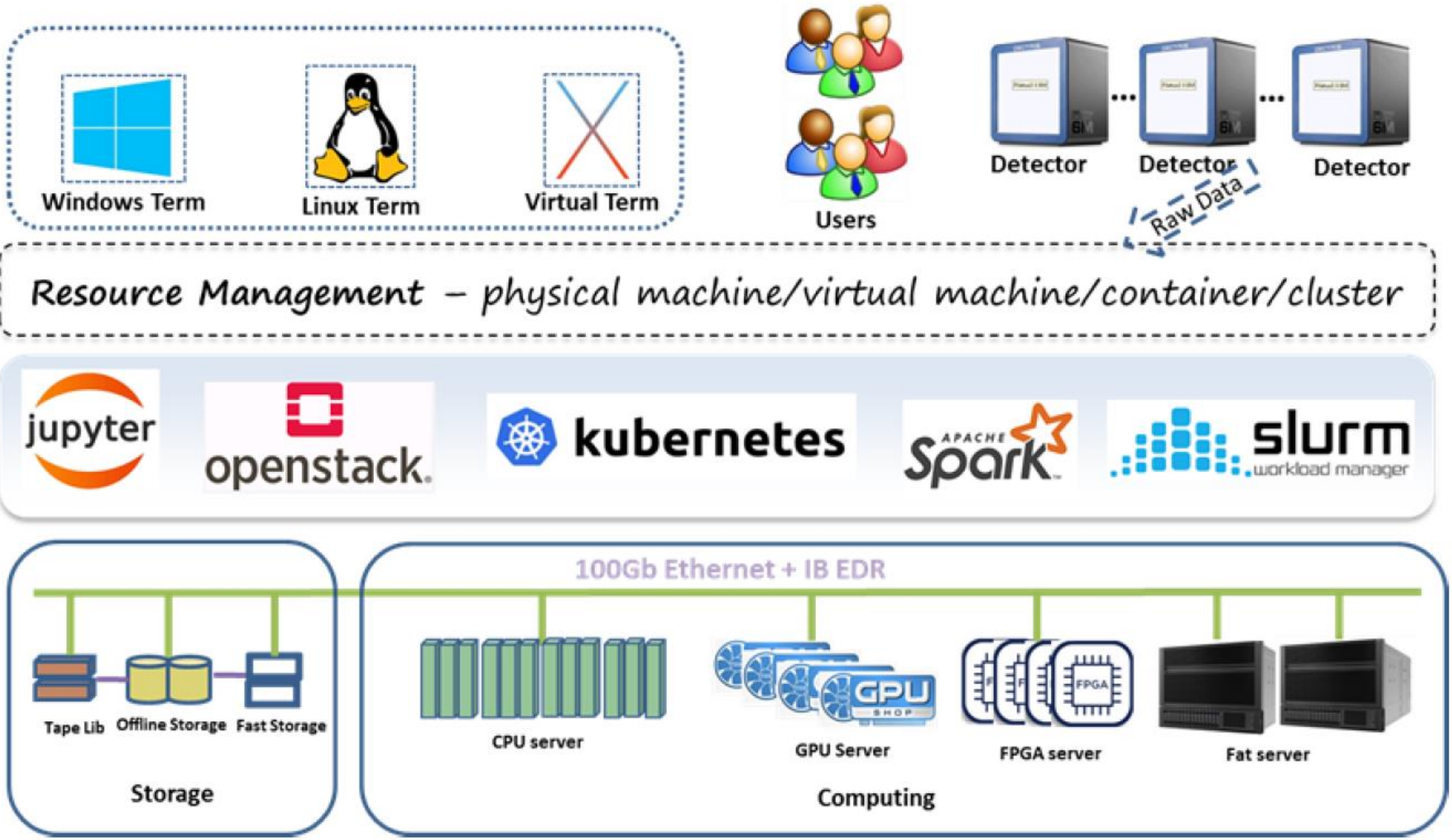


Storage Overview

- Beamline File System
  - Provides I/O for fast on-line analysis
  - High Throughput, 10GB/s write, 20GB/s read
  - Sequential read and write
- Central Storage
  - Provides I/O for batch jobs, interactive analysis, backup and data transfer
  - Mixed I/O patterns, sequential read and write, random read
  - High Throughput, 10GB/s write, 50 GB/s read
- Tape Storage
  - Used for data preservation and backup
  - Sequential Read and Write
- Policy engine
  - make decision of data copy based on static rules or prediction of AI models
  - scheduling and monitoring data copy requests between storage system

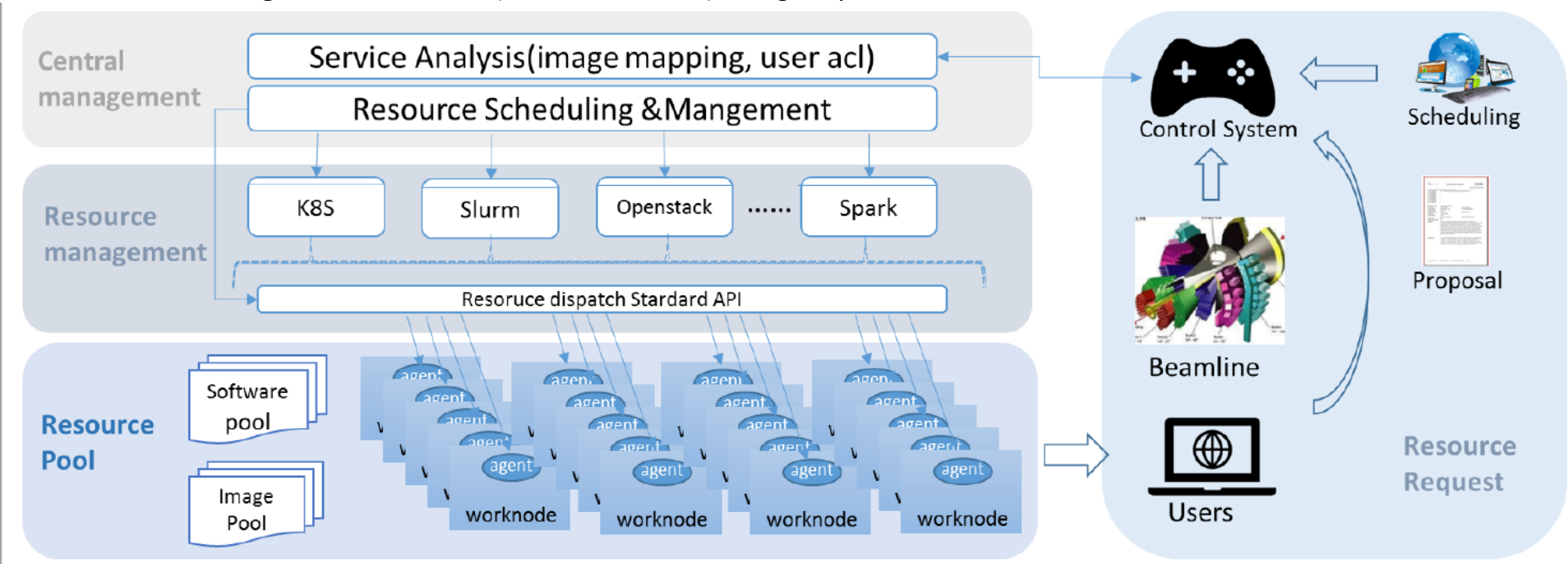


Computing Overview



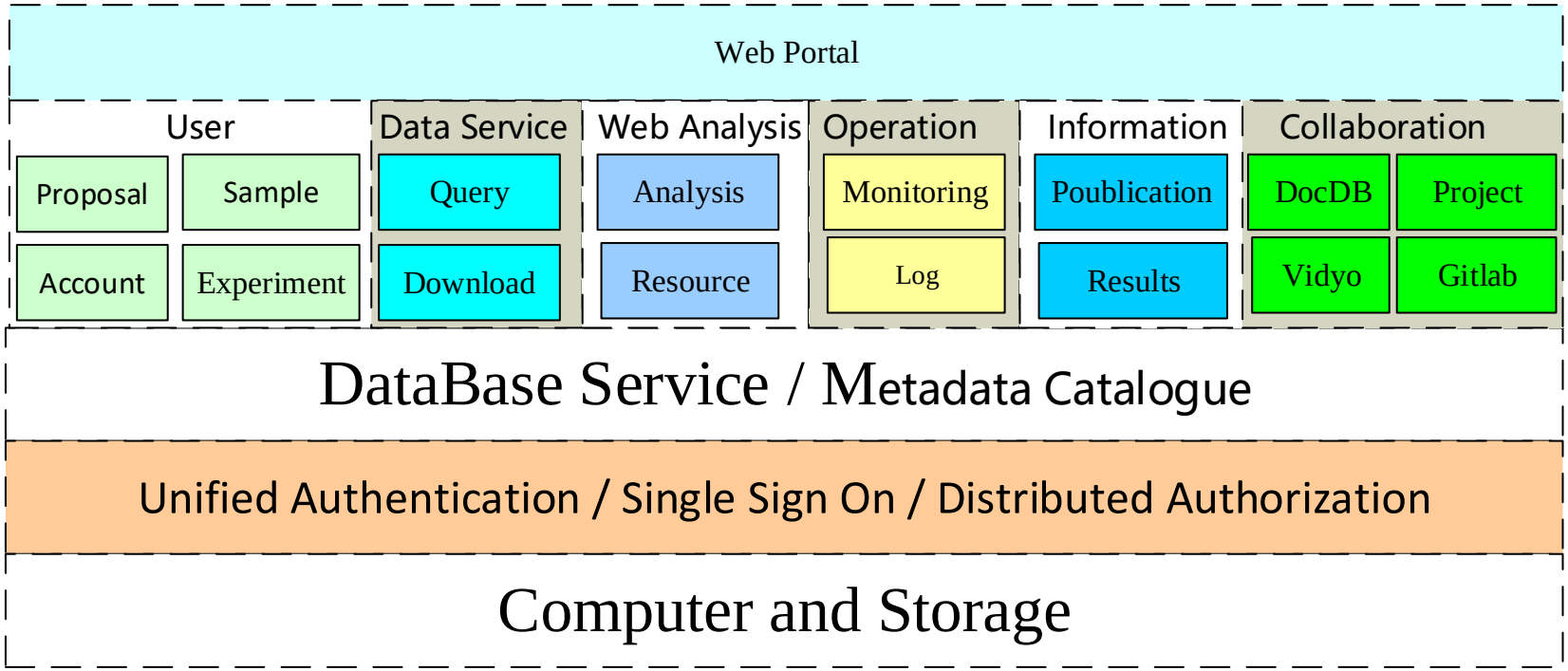
Computing Resources Schedule

- Unified resource management and scheduling(to provide Distributed/Central resources)
  - Implement the appropriate computing resources with necessary HEPS software and services to provide Support on-demand, scalable computing resources for users transparently and quickly (minutes-level)
  - HPC cluster, real-time data processing farm, cloud-based analysis and Web-based analysis
- Central management
  - Responsible for resource requests analysis and resource schedule like resource mapping and user access control
- Middleware
  - Resource management middleware like K8S, Slurm, Openstack…
- Resource Pool
  - Heterogeneous resources(CPU, GPU, FPGA) ,image repositories, software and so on



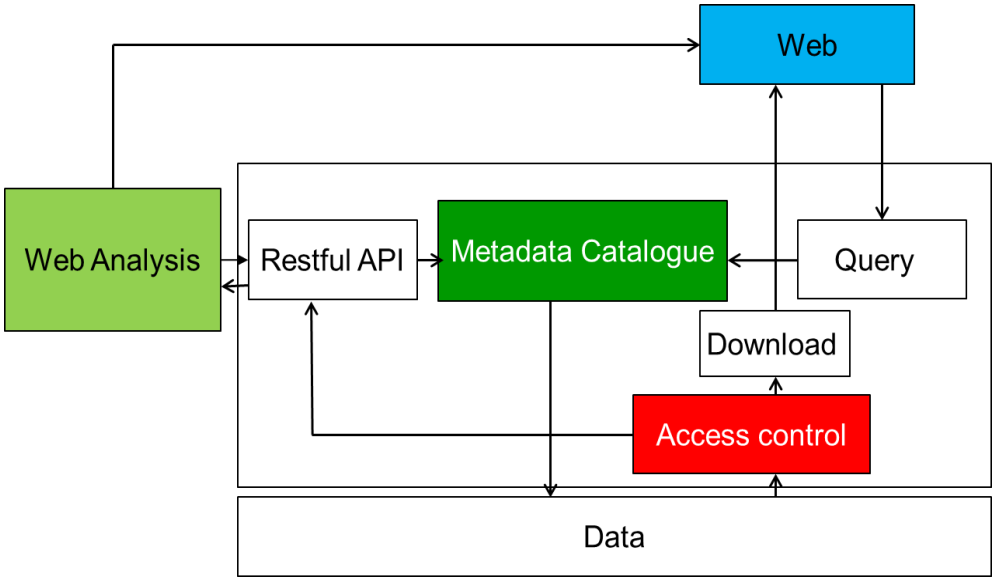
System Design : Data & Public Service

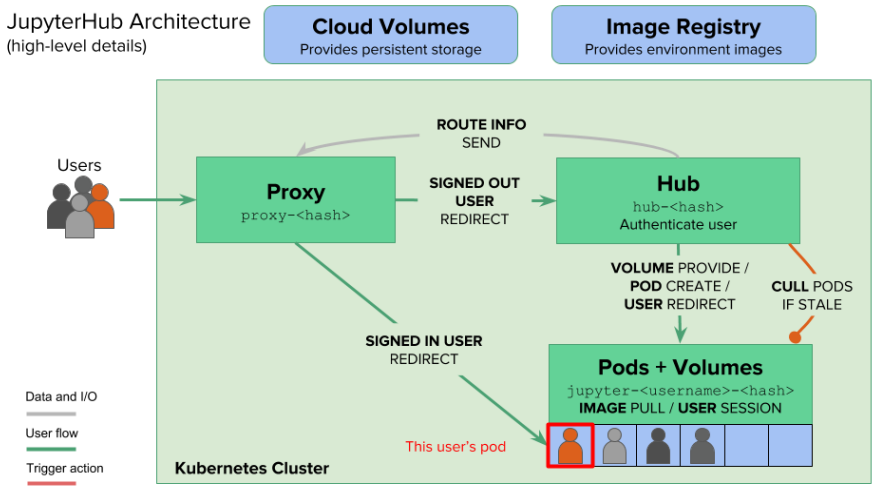
- Services
  - Proposal
  - Sample
  - Data Service
  - Web analysis
  - Results & information
  - Collaboration



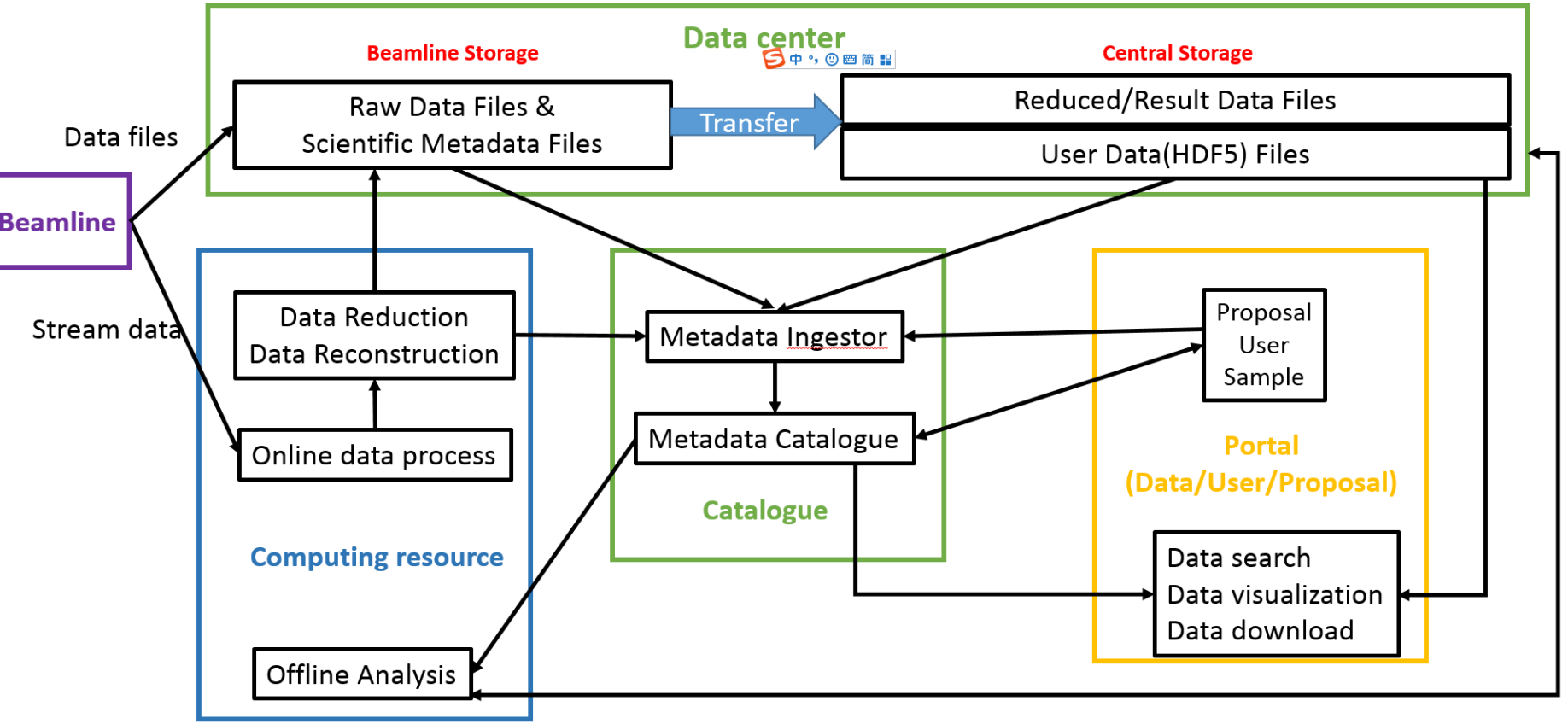
System Design : Data Service

- Standardized data format: NeXus / HDF5
- Data service
  - Search data based on the catalogue
  - Access share/download data based on data permissions
  - Data can be linked to proposals and samples
  - Data can be linked to publications
  - Tracking the data...
- Web based analysis
  - Frontend: Jupyter Notebook with Python support.
  - Middleware: HEPS specific analysis software integration.
  - Backend: opensource software JupyterHub + Kubernetes.
    - Support on-demand, scalable computing.
    - Computing resource is managed by Kubernetes.





Scientific Data Flow



System Design: Data Management

- Data policy
  - The acceptance of the Data policy must be a condition for the award of HEPS beam time
  - Provide disk storage for at least 3 months and permanent tape archive
  - Provide temporary storage for reduced data, processed data, results data, calibration data
  - Provide long-term storage for raw data, user data
  - Data service of HEPS will be restricted to registered users of data management system.
- Metadata catalogue
  - Support the management of the whole scientific data lifecycle
  - Catalogue and provide access to scientific metadata and raw experimental data
  - Cooperation with CSNS/SSRF : iCat / SciCat
- Data transfer system
  - Transfer all the data from beamline storage to central storage
  - Deploy data transfer instance （secure/easy to re-setup/light） in each beamline
- Data format
  - HDF5 is chosen as the standard data file format
  - Follows NeXus conventions (discussed with beamline scientists)

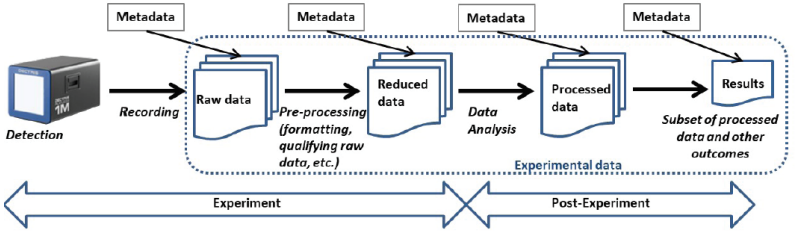
Data Storage Policies & Scheme

- Data storage policies

Control system/DAQ → **Raw data: long-term storage**

Data analysis→ **Reduced data/Processed data/Result data: temporary storage**

User data: standard format & long-term storage



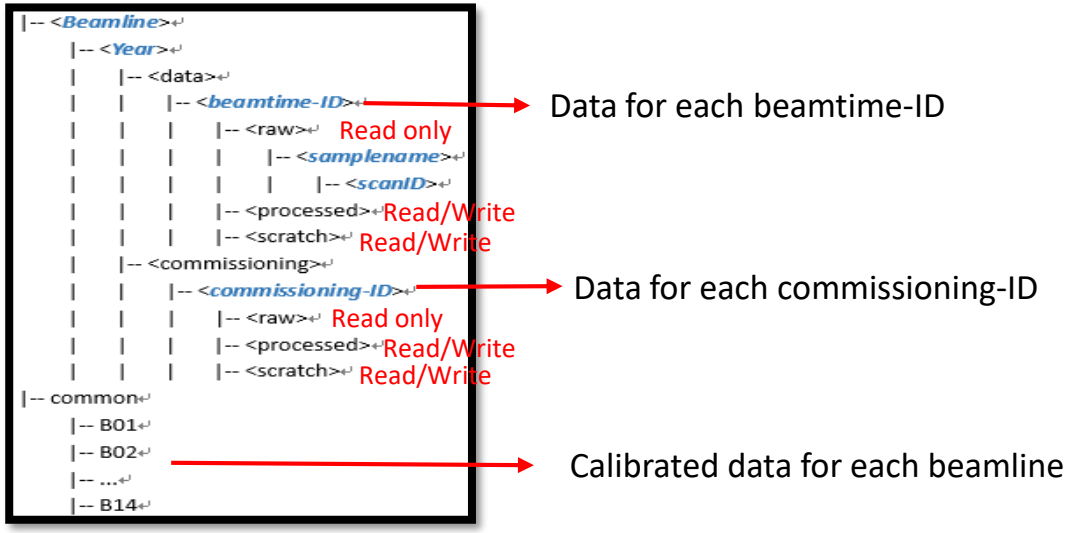
- Scheme

Storage directory structure

access rules

Beamline storage

Central storage

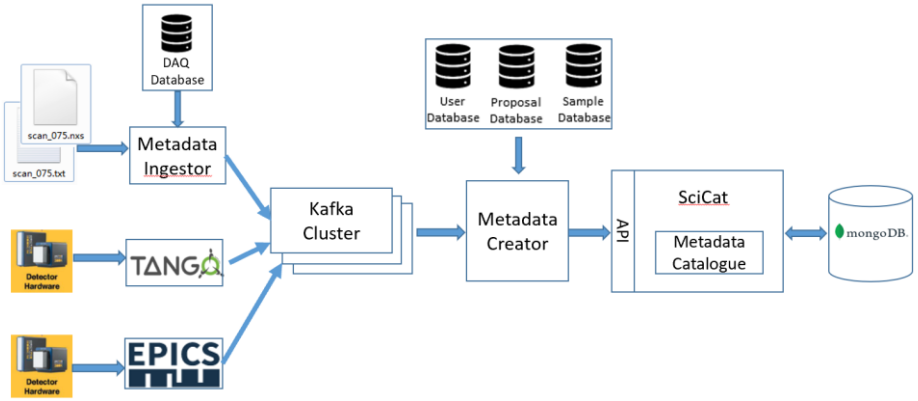


Metadata items to cataloging & Acquisition

Q: What kind of metadata are needed cataloging?

A: Those metadata necessary and significant for data searching and sharing!

Notice: Metadata with red mark should be provided by control system.	
Proposal Info	proposal ID, proposal name, principal investigator (PI), owner
Experiment Info	beamline, begin time, end time
Dataset Info	persistent ID, format, overall size, path
Data blocks	data files (file size, format, storage path, time, checksum)
Experiment metadata	Experimental technique specific Eg. detector Info (scan, x-ray exposure parameter…), Sample Info



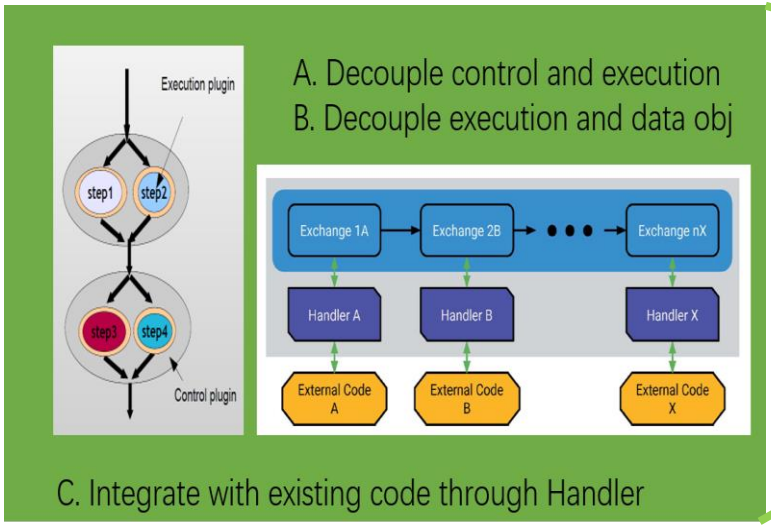
Depends on how metadata are provided

- Interfaces are provided for control system to write metadata to DMS(recommended)
- Metadata-ingestor plugins are designed to collect metadata from NeXus/txt/HDF5 files

System Design : Science Software Framework

Business Domain
- Algorithms
- Workflow

Running Time
- Workflow Engine
- Data Store



Scientific Analysis Application

Tomography | Spectroscopy | Diffraction | Other

GUI Widget
(Matplotlib & PyQt)
- 1D X vs Y plot
- 1D Histogram
- 2D Image view
- 3D surface
- 3D iso-surface and volume

CLI/Scripts
(ipython)
- Analysis Scripts
- Plotting Scripts
- Data processing workflow

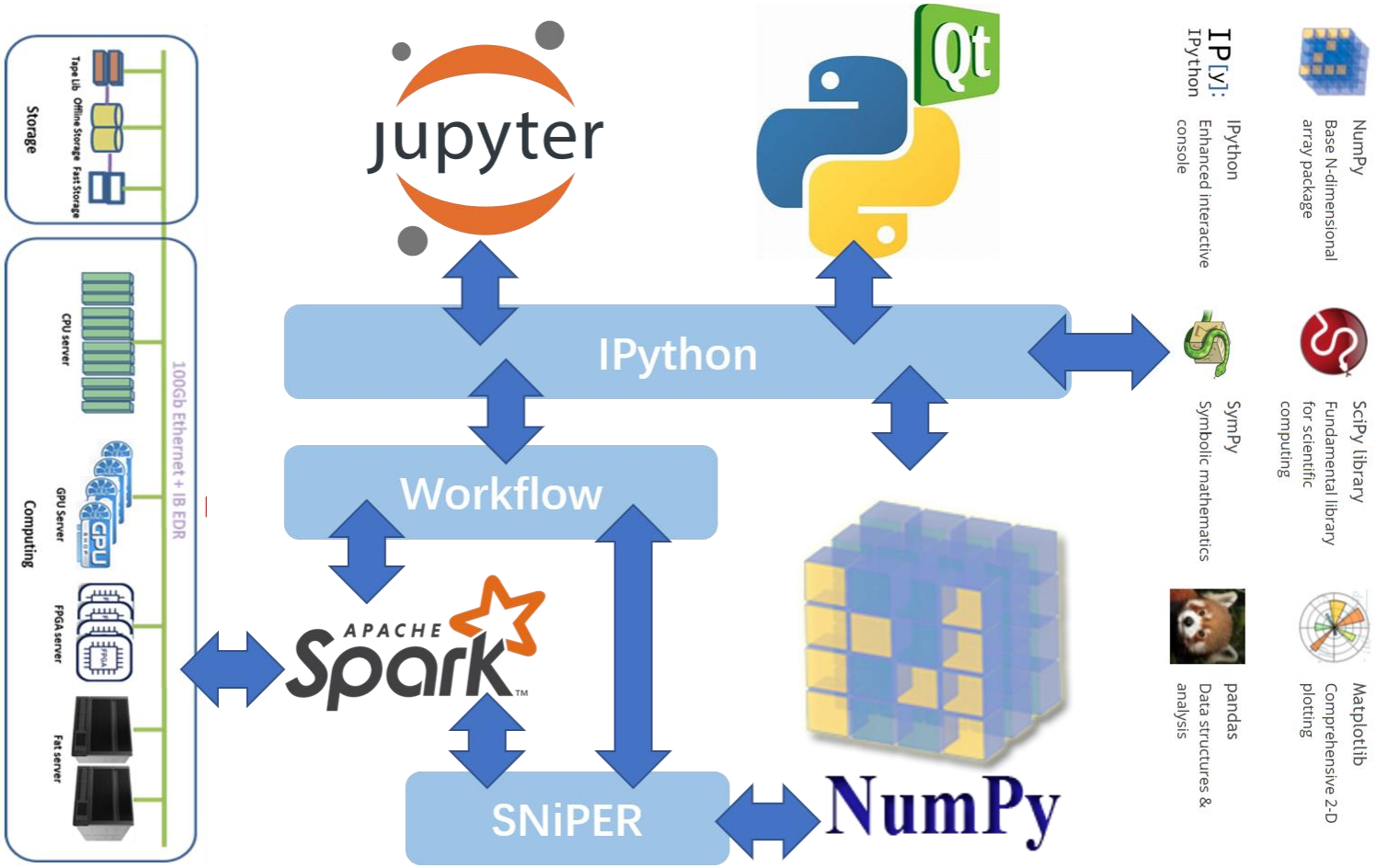
Middleware
- Workflow Engine
- Data Transfer Protocol

Framework
- Algorithm & Alg Mgr
- Data Object & DataObj Mgr

3rd Libs/Apps
- Python Ecosystem
- C/C++ Libraries
- Other Open Source Software

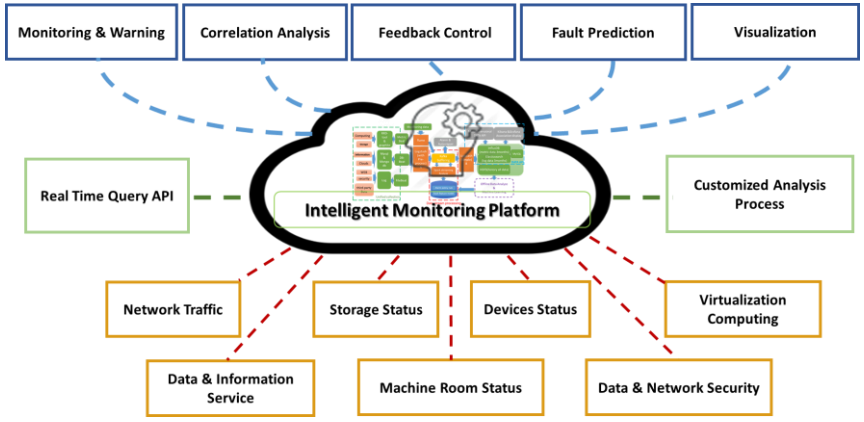
Computing Resource | **Network** | **Storage**

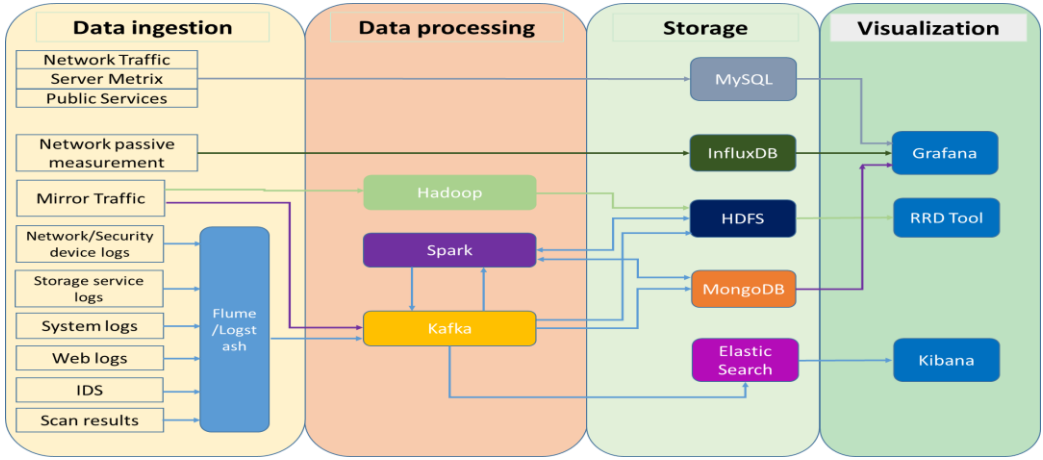
Executing Process of Workflow



System Design : Agile Intelligent Monitoring

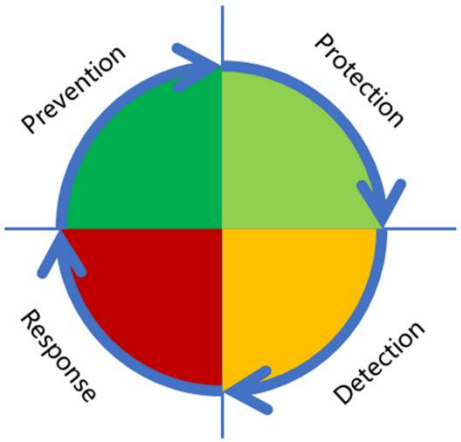
- Consists of data ingestion, data processing, storage and visualization
- Monitoring items : network / experimental data / Jobs / logs / security
- Help administrator to locate problem reasons and automatic fixing of system failures, improve the stability and reliability of system
- Provide rich, real-time, interactive visual panels



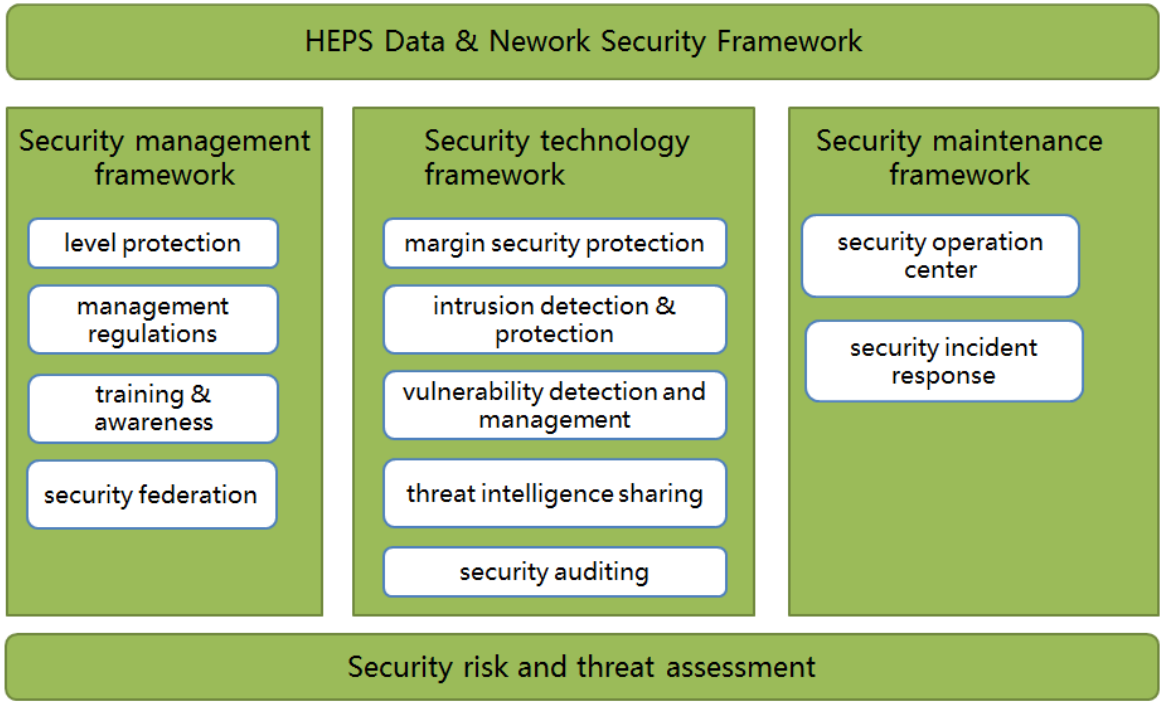


System Design : Security

- IT assets and vulnerabilities
- security rules and policies
- accounts and passwords/credentials
- data access control
- operations security
- incident response



Cybersecurity Risk Control Model



System Design : Capability

Items	Performance		
Machine Room	Space: 600 ㎡ Racks: 30(100)		14 beamlines phase I >90 beamlines phase II
GPN	1Gbps/10Gbps		
DCN	10Gbps/100Gbps		
Storage	30 PB Disk XXPB Tape		Tape storage will be provided according to the funding
Computing	CPU: 90 TFLOPS GPU:365 TFLOPS	2500 CPU Cores 48 GPU NVIDIA Tesla V100	

Big gap.....between the capability and the missions / requirements for the funding reason

But the system is scalable.....

Outline

Plan & Time table

- Plan & Time table
  - Network services should be ready from the beginning for civil construction
  - 2018-2019: Learn the requirements from beamlines, read documents and learn solutions from other light source facilities
  - 2020: Learn, understand the requirements, finish the detail design, start the test bed (we are here now!)
  - 2021-2023: Software development and testing.... More test beds
  - 2024: Production environment will be ready
  - 2025: Online for all the IT services
- Problems
  - Lack of the knowledge/experience for light source experiment data management & processing for HEPS CC team

More cooperation is needed

- With beamline in-house scientists
- With control system
- With other facilities: SSRF,SHINE,CSNS...
- With other computing center
- With computer science and scientists
- With software communities
- ……

Summary

- Many missions for HEPS CC
  - Facilities / Computing / Storage/ Network / Software / Services / Database / ...
- The system design has been finished
- Cooperation with other facilities and community is ongoing
  - data management, data format, software ……
- Should learn and understand more about the requirements from beamlines , scientists , and other similar facilities

Everything is ongoing……

References from SSRF, SHINE, ESRF, PSI and XFEL, we'd like to thank them.

Thanks for your attention

and

Welcome for the comments and suggestions