CEPC Computing Platform Design and Vision

Xiaowei Jiang
On behalf of CEPC computing team and IHEP-CC

October 26, 2024
The 2024 CEPC International Workshop

Contents

- Status of CEPC Distributed Computing
- CEPC Distributed Computing Infrastructure (CEPC DCI)
  - Distributed Computing System
  - Distributed Storage System
  - Network and Data Transfer
  - User Authentication and Authorization
  - Other systems
- Summary

CEPC Distributed Computing Status

- DIRAC is chosen as distributed computing framework
  - Originally from LHCb, now used for many new experiments: BELLEII, ILC, CTA, SKA…..
- CVMFS for software distribution
  - stratum0 operated @IHEP : /cvmfs/cepc.ihep.ac.cn/, stratum1 @IHEP and @RAL
- VOMS for managing CEPC users
  - VOMS hosted @IHEP : https://voms.ihep.ac.cn:8443/voms/cepc/
- CEPC users can access resources everywhere with web or client
  - Web sites: https://dirac.ihep.ac.cn
  - IHEPDIRAC Client in cvmfs: /cvmfs/dcomputing.ihep.ac.cn/dirac/IHEPDIRAC/

Resources and Sites

- About 4600 cores in the system
  - IHEP has dedicated resources
  - CPU: 2000 cores (640 cores shared with ILC in grid)
    - Several thousands of CPU cores will be added next year
  - Storage: 3.7 PB
    - Several PBs would be added next year
- Five joint sites from UK and other China universities
  - ~2600 CPU cores
  - Shared with other experiments
- Network
  - A shared network link with 100 Gbps bandwidth between China and Europe

CEPC Distributed Computing Infrastructure

- CEPC Distributed Computing Infrastructure (**CEPC DCI**) is responsible for CEPC data processing
  - **Data processing** -> Distributed computing system
  - **Data access** -> Distributed storage system
  - **Data distribution** -> Network and data transfer system
  - **Data privilege management** -> Authentication and authorization system
  - Receive data from detector and also engineering data, support data processing, scientific research and international collaboration of grid computing etc.

CEPC Data Processing Requirements (1)

- CEPC experiment is an international collaborative experiment, the data processing needs across different regions and multiple data centers
  - **Unified data storage across data centers**
    - All types of data are stored in IHEP
    - Data replicas in the regional center
    - Coordinated and shared storage usage among data centers
  - **Data transfer between data centers**
    - IHEP: RAW and reconstruction
    - Regional center and Chinese center: RAW or reconstruction
    - Among normal sites: mainly simulation data
  - **Computing resources sharing and collaborative scheduling of computing tasks across multiple data centers**
    - Computing resources (CPU and GPU) are managed by a unified computing platform and allocated based on the characteristics of CEPC data processing task
    - Computing tasks are submitted from a unified entrance with the standard methods

CEPC Data Processing Requirements (2)

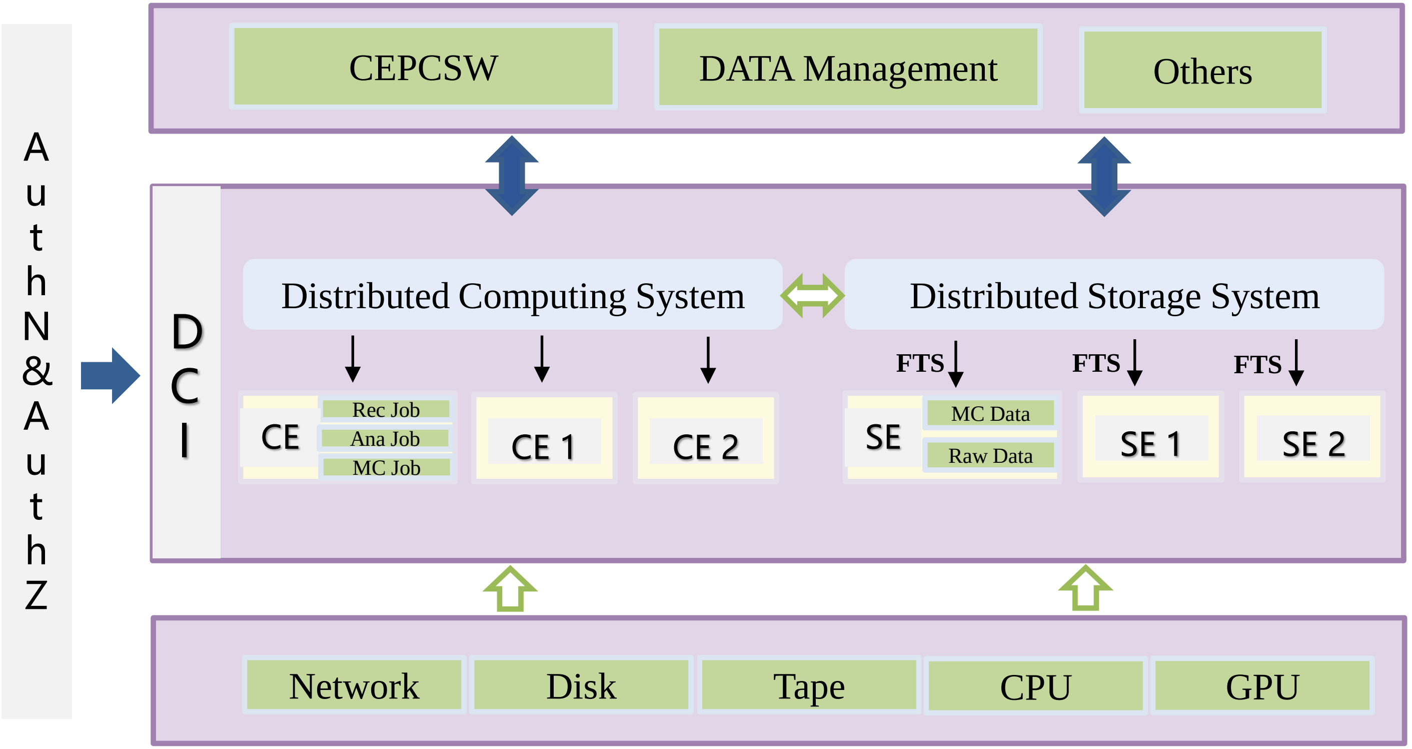
- CEPC experiment is an international collaborative experiment, the data processing needs across different regions and multiple data centers
  - **User authentication and authorization**
    - Uniformly manage the identities of users in the CEPC collaboration, providing a standard method for joining CEPC collaboration
    - Unified management of permissions for information service, computing, storage, data systems, etc.
  - **Information services**
    - Including documentation, meetings, code repositories, websites and visualization
- A set of distributed computing software suites should be developed and deployed in CEPC DCI

Tier Model of CEPC DCI

- Inspired by WLCG Tier Model
  - T0 -> T1 -> T2 -> T3
- Tier-0 sites: Central site
  - IHEP: All types data storage and data distribution source
- Tier-1 site: Regional center site
  - SIM and REC data storage, computing resources
- Tier-2 site: SIM data processing
  - SIM and ANA
- Tier-3 site
  - Basically local sites

Structure of CEPC DCI

- Software in CEPC DCI
  - Distributed Computing System
  - Distributed Storage System
  - Network and Data Transfer
  - AuthN & AuthZ
  - Other Systems
    - Software publish/deployment
    - Unified DCI software distribution

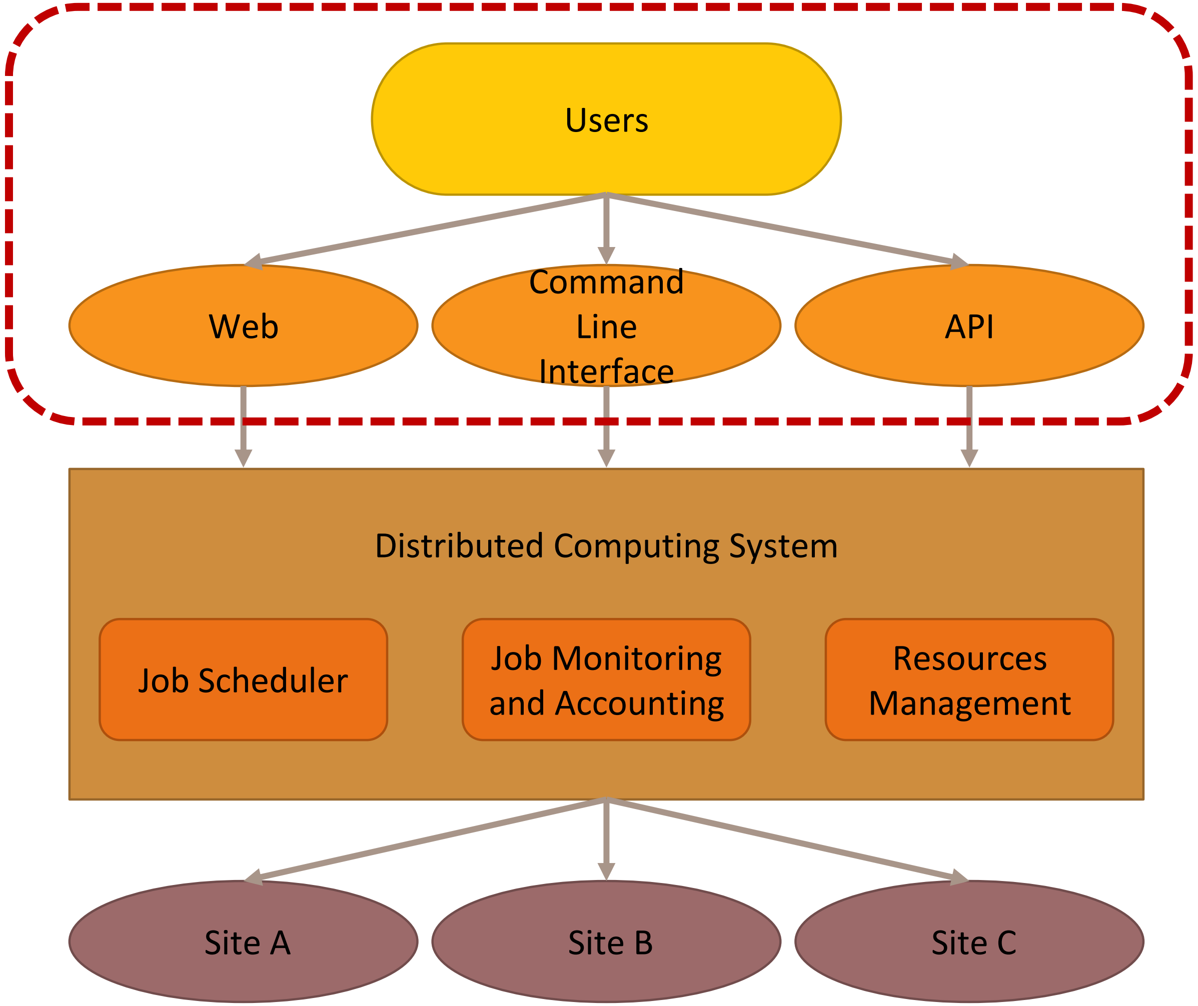


Distributed Computing System

- CEPC computing system serves:
  - **Official data processing**
    - SIM and REC data production
  - **User analysis data processing**
  - Special computing tasks
    - tasks on supercomputing sites, GPU sites, etc.
- CEPC computing system manages:
  - Distributed computing sites around the world, by distributed computing system
    - For official data processing and special tasks
  - Computing resources from sites, by site computing service
    - For user analysis

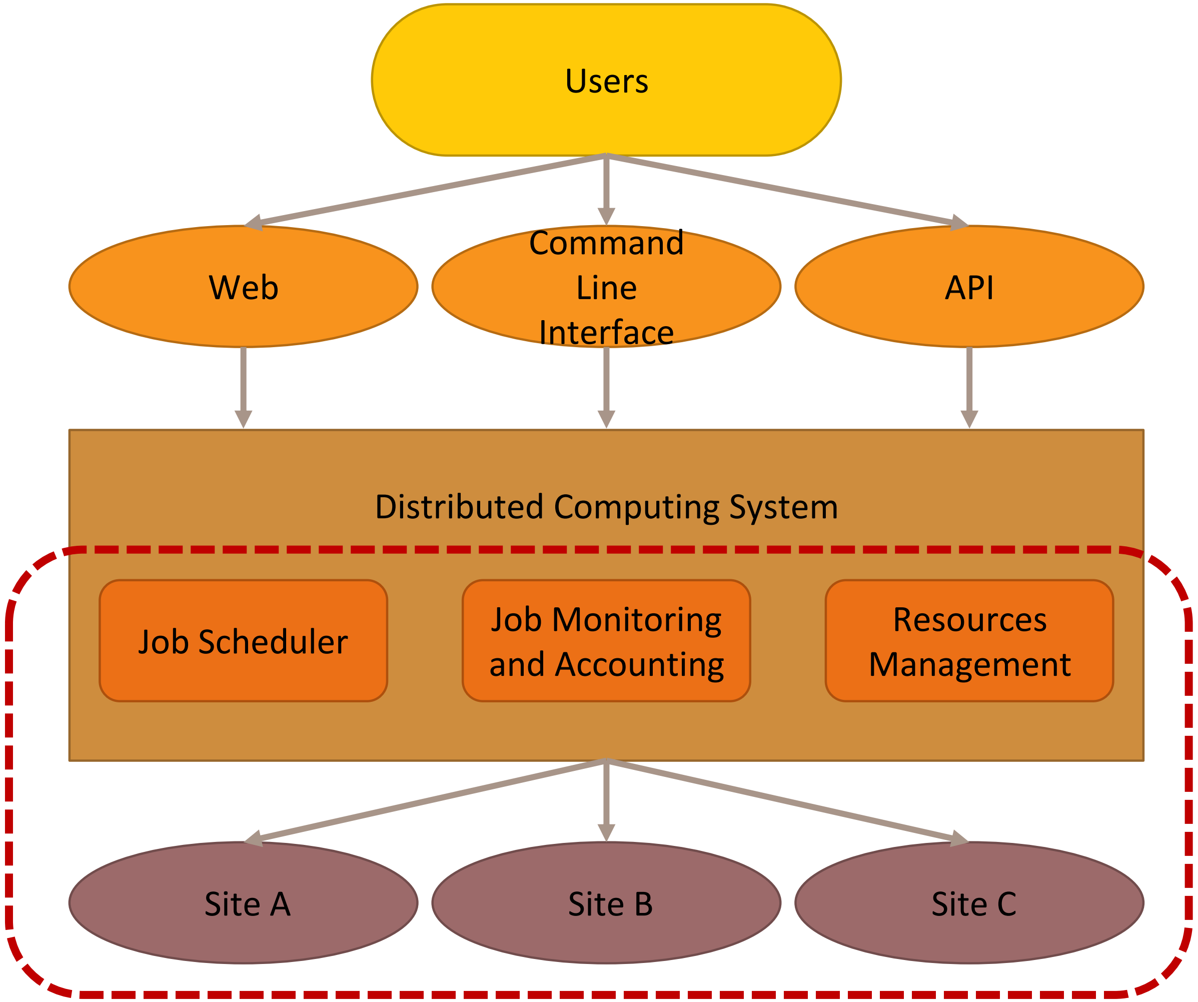
CEPC Distributed Computing System

- Distributed Computing System
  - To manage the distributed computing resources from the world
  - Mainly for official data processing
- For Users:
  - To unify computing sites with heterogeneous computing systems
    - HTCondor, Slurm, Cloud computing, supercomputing, local cluster, etc.
  - To supply unified job management interface
    - For users and production system
    - By Web, Command line interface and APIs



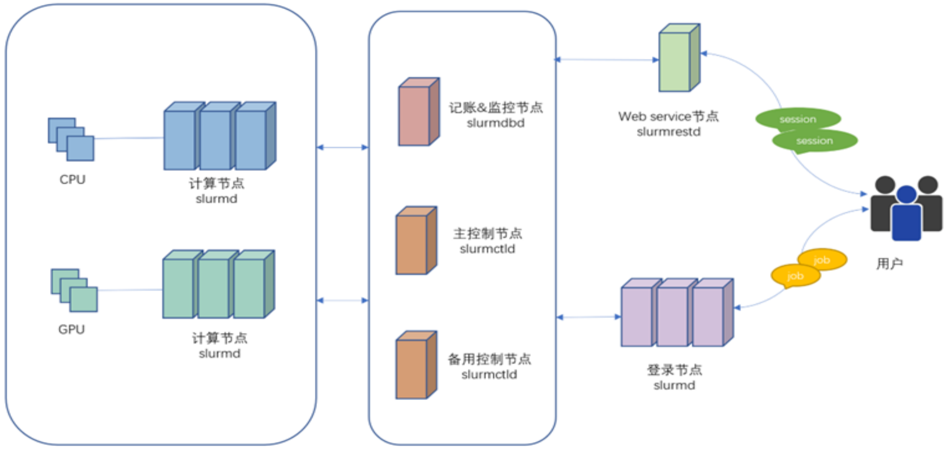
CEPC Distributed Computing System

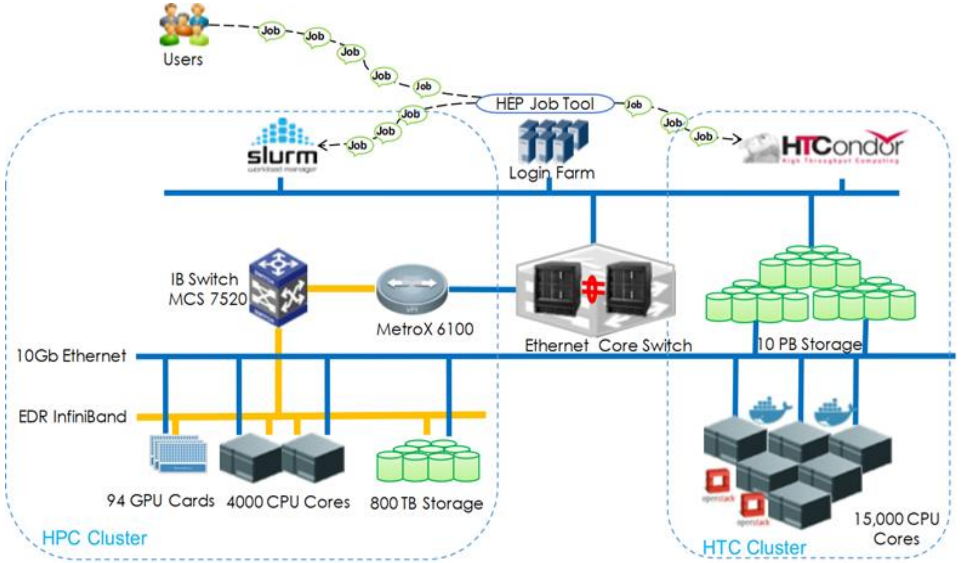
- For Sites:
  - To schedule jobs to computing resources
  - Optimize jobs distribution among sites
  - Monitoring computing resources status
  - Generate site reports and accounting sites usage



Site Computing Service

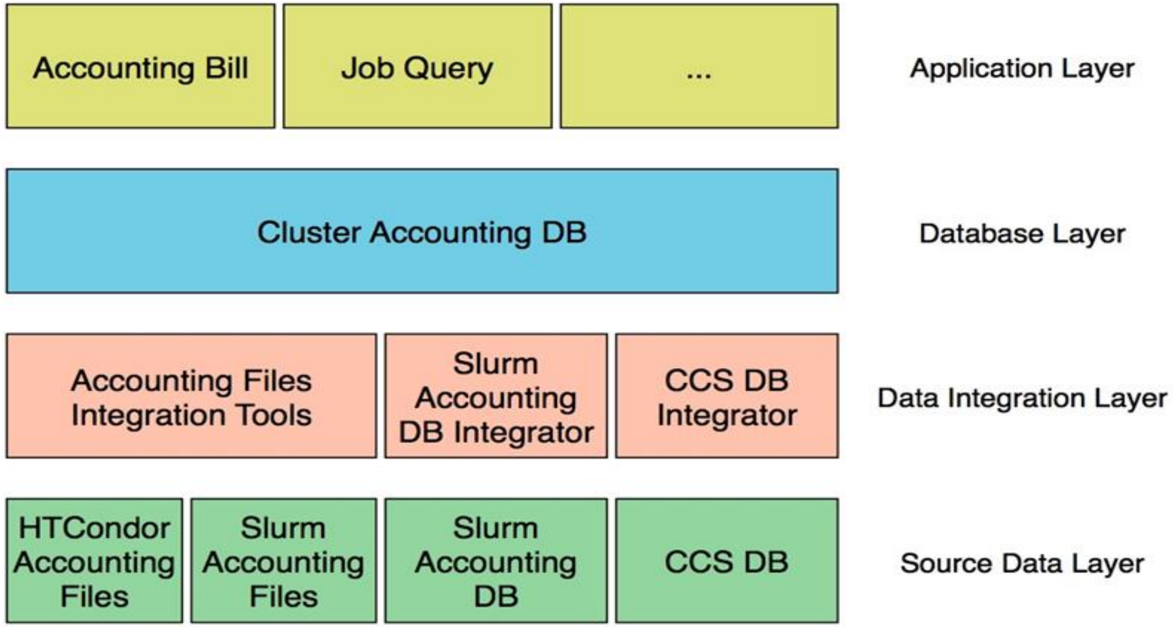
- Type of CEPC computing jobs
  - Single-core job or multi-core job within one node: simulation, reconstruction, analysis
  - Multi-core job on multi nodes or GPU job: part of reconstruction, AI training
- CEPC site computing service is based on HTCondor/Slurm
  - HTC service for single-core job or multi-core job within one node
    - Support 1,000,000 jobs queuing and 100,000 jobs running
  - HPC service for big multi-core job or GPU job
    - Support big-scale parallel job and GPU
- Service components
  - Resource management and allocation
  - Computing task management





Job Accounting and Monitoring

- Architecture
  - data source layer, data integration layer, database layer and application layer
- Metrics
  - users, groups, and experiments
  - CPU, memory, walltime,…
- Support for multiple sites and multiple computing services



Accounting Bill | Job Query | … — Application Layer

Cluster Accounting DB — Database Layer

Accounting Files Integration Tools | Slurm Accounting DB Integrator | CCS DB Integrator — Data Integration Layer

HTCondor Accounting Files | Slurm Accounting Files | Slurm Accounting DB | CCS DB — Source Data Layer

CEPC Distributed Storage Management

- CEPC distributed storage management
  - To produce and distribute data from distributed computing and storage sites
  - To manage distributed data access requests from other data systems or users
  - Based on Rucio system, a popular grid data management system in HEP
- Storage management services manages data production
  - RAW data distribution, IHEP Tier0 site
  - SIM and REC data distribution, replicate among Tier1 and Tier2 sites
  - Official data adding, deleting, modifying, querying in distributed storage sites
- For normal users
  - Supply data access in developed CEPC Storage APIs and client

Distributed Storage Workflow

- When a data distribution subscripted

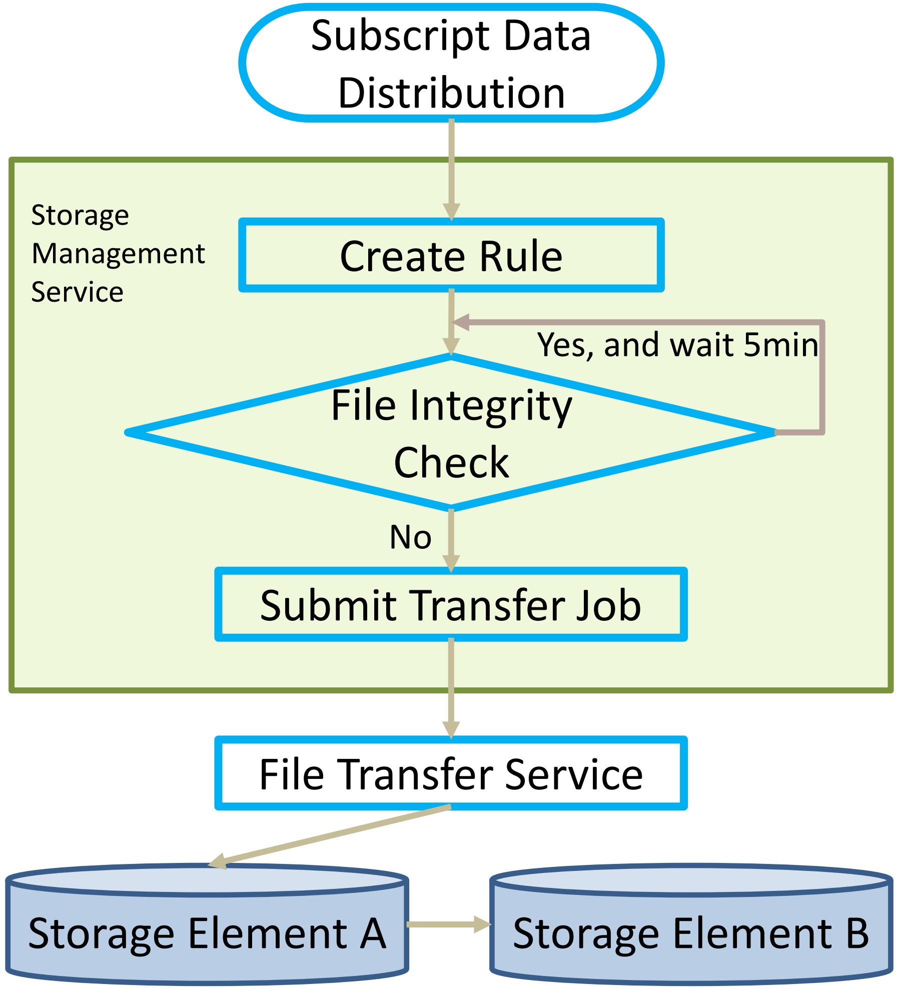
  ① A replication rule created in database

  ② File Integrity check daemon scans database and find incomplete files

  ③ Submit new transfer jobs to file transfer service for incomplete rule

  ④ If all file completed, waits for 5 min and restarts scan then

- Database and Daemon design
  - Multi-threads to exceeds rule processing
  - Avoid status loss in message system

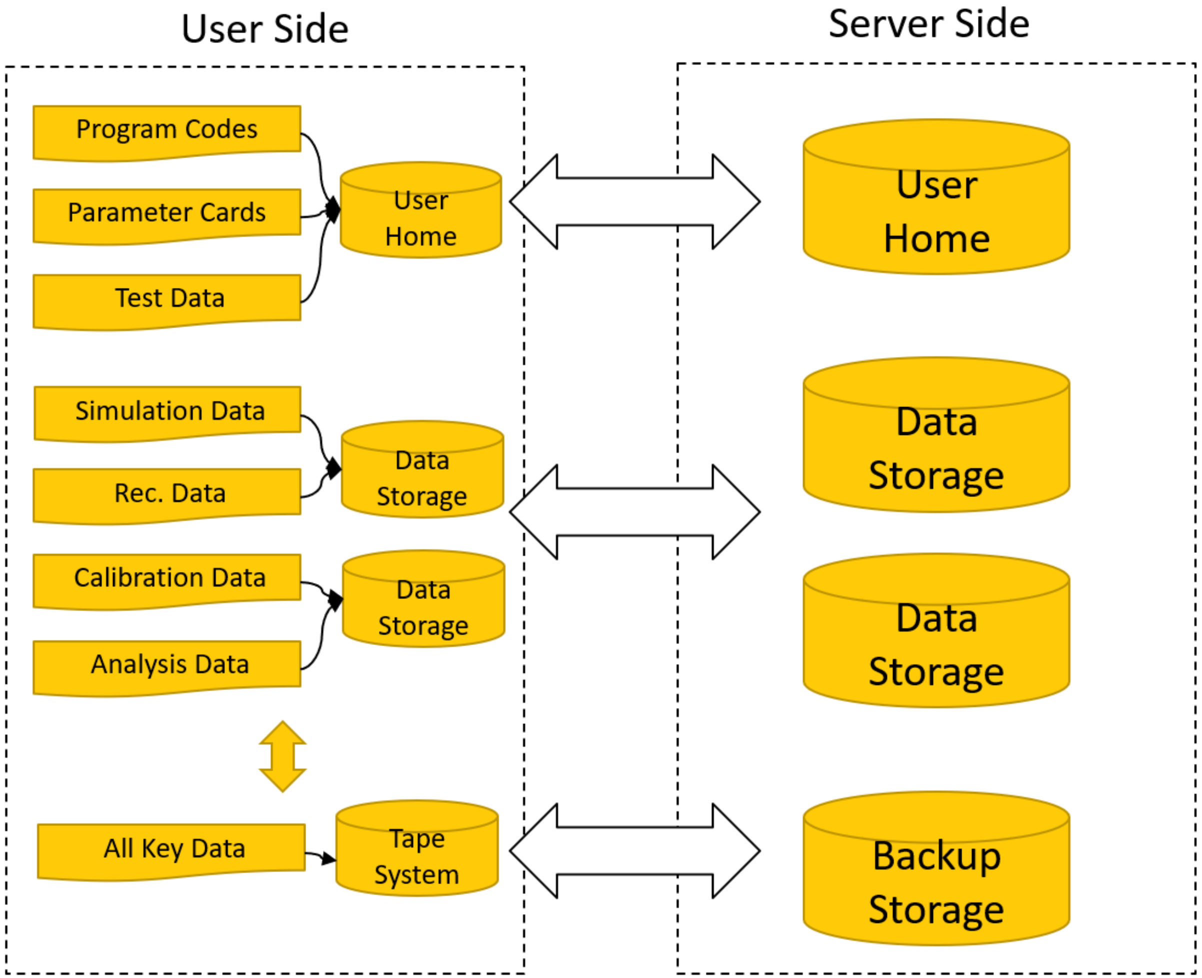


Storage Access Permission Management

- User permission is authorized by AuthZ service
  - Only production group user could add, delete, modify data in CEPC
  - Fine-grained permission is managed by CEPC permission policy
    - Could be managed by user group and user name
    - Could manage every single file execution command and system command

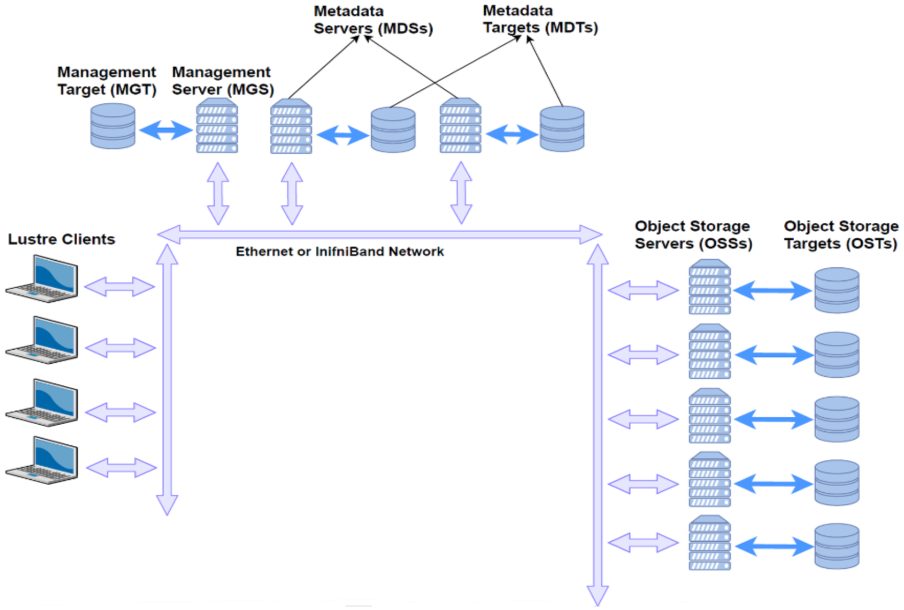
Site Storage Services

- All the CEPC data physically stores in the storage system
- CEPC experiment data
  - sim/rec/cali/ana/…
  - Large size (GB per file, PB in total)
  - Huge amount of files (hundreds of millions)
- User Personal data
  - codes/parameter/test data
  - Small size but big number of files
- Key data needs backup
  - Part of experiment data for permanent backup
  - Personal data backup periodically



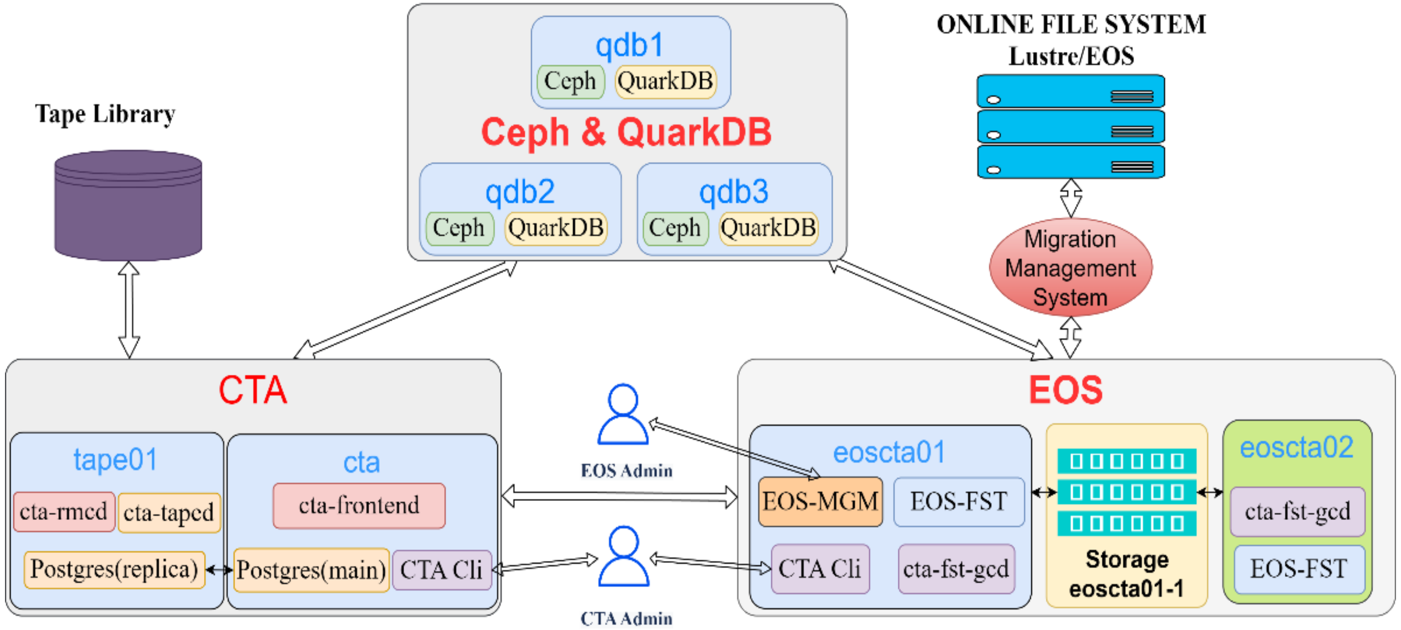
Site Storage – Disk Service

- User home storage is developed based on Lustre
  - Lustre is open-source and support for massive data storage
- Development and deployment
  - Service development over Lustre
    - Lustre client/management service/accounting/monitoring/…
  - Deploy an independent user home storage instance for CEPC
    - MGT/MGS/MDS/MDT/…
- Data storage is developed based on EOS
  - EOS is open-source and popular data storage in high energy physics
- Large-scale of disk storage: hundreds of storage servers
  - Dedicated storage pool for CEPC
  - Service development
  - External APIs for CEPC: Production system/Job system/…
  - Support for xrootd protocol and http protocol



Site Storage – Tape Service

- CEPC key data archives in tape system for long-term data storage (backup)
  - Tape is cheaper than disk and good for long-term storage
- CEPC backup system is developed based on EOS-CTA
- Tape system components
  - Tape buffer/Tape server/Tape library
- Integrate external software
  - ceph/quarkdb/postgre/eos/xrootd/…
- API developments
  - Throughput monitoring and optimization
  - API for external systems
    - DMS/Production system/…



CEPC Data Transfer System (1)

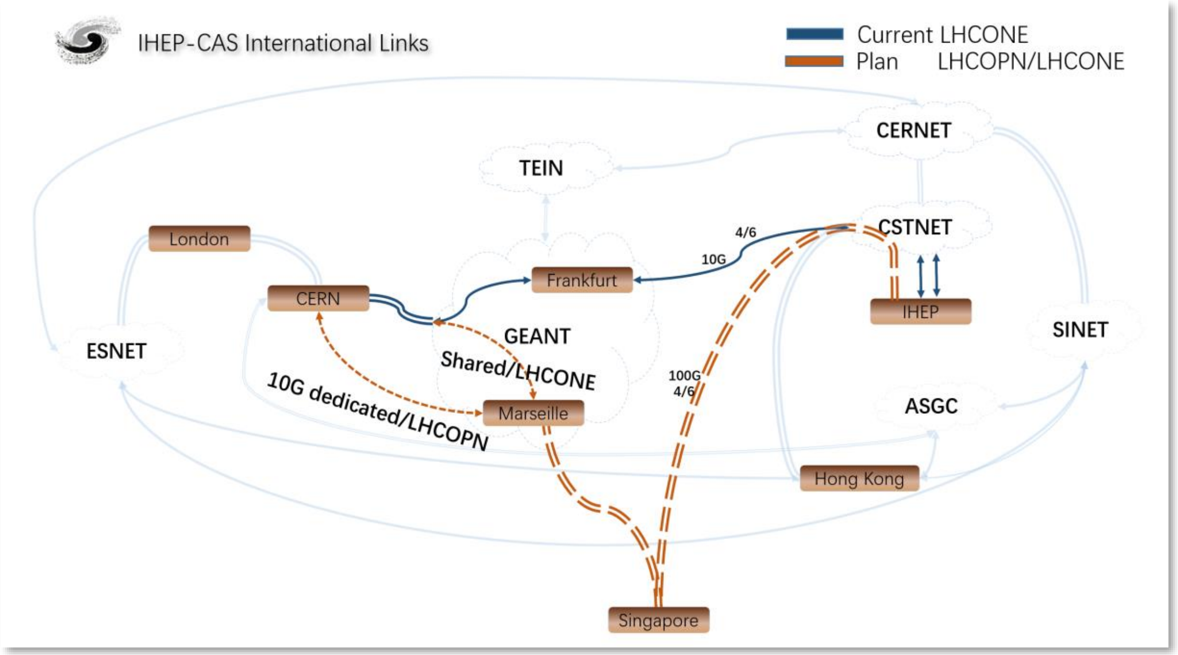
- CEPC needs to transfer official RAW, MC, REC data among sites
  - MC data flow: Basically T2 -> T1
  - REC data flow: Basically T1 -> T2, T2->T1
- Data Transfer needs:
  - Support Grid transfer among IHEP and other T1, T2 and T3. Transfer job submitted by CEPC storage management system
  - Support Token-based TPC transfer
- Data Transfer is an infrastructure service, cannot be used by normal user

CEPC Data Transfer System (2)

- Token-based protocols
  - Root (xrootd): Origin from ROOT framework and good support for ROOT file
  - HTTP (WebDav): Common protocol in Internet and support for more systems and services
- TPC transfers
  - Directly from site A to site B, no Client as temporary middle storage
  - Root and HTTP has already supported TPC copy
- Transfer Tools
  - FTS3 and Gfal2, Grid transfer standard tools
  - Monitoring and accounting will be set at IHEP for transfer

Network System

- CEPC data transfer and information interaction depends on network system
  - Especially data transfer need a stable network link with enough bandwidth
- Network Topology should be established between CEPC data centers
  - Establish international export links between IHEP and other sites

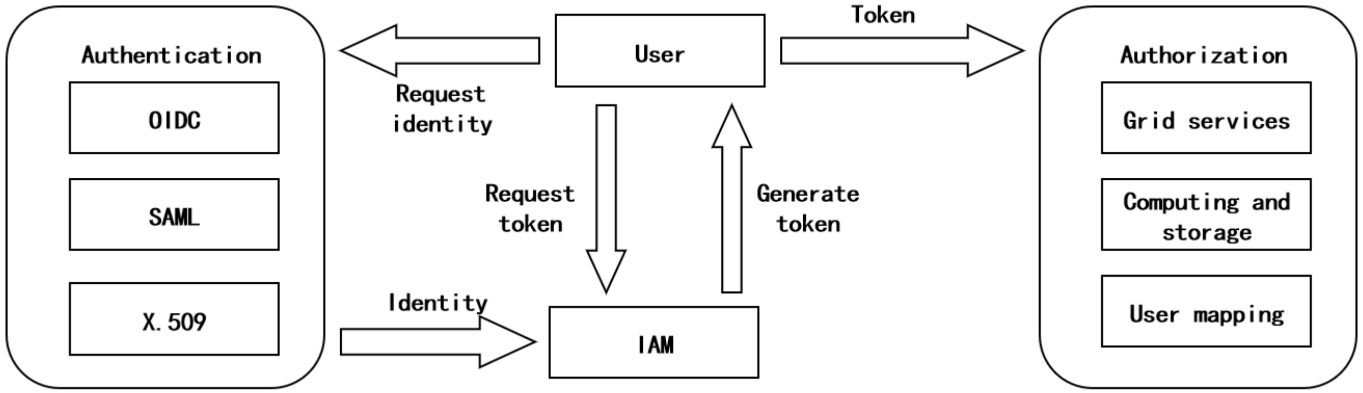


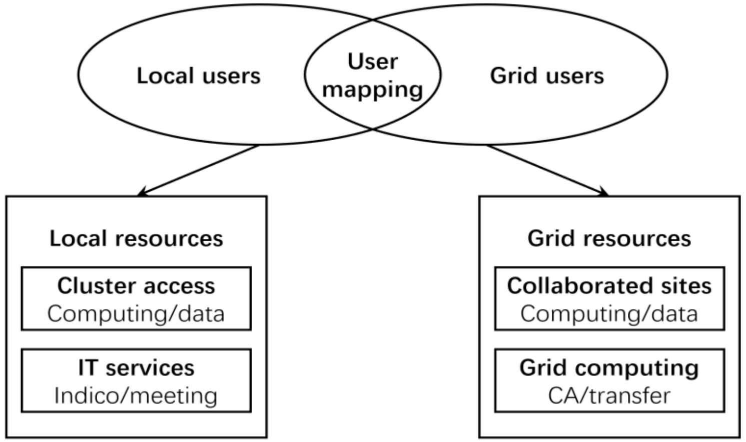
User AuthN&AuthZ System

- CEPC member users should have an identity in CEPC DCI
  - A certain identity to safely access the multiple systems of CEPC
  - The identity is obtained from a unified identity authentication system
- Different roles of CEPC user should have different permissions of accessing different system or service
  - Data permissions (who can read/modify/delete what data)
  - Resource permissions (who can submit jobs to request resources)
  - Other service permissions: indico/gitlab/docs/…
- CEPC user authn&authz system covers two types of users
  - Grid users and local users

User AuthN&AuthZ System – Grid User

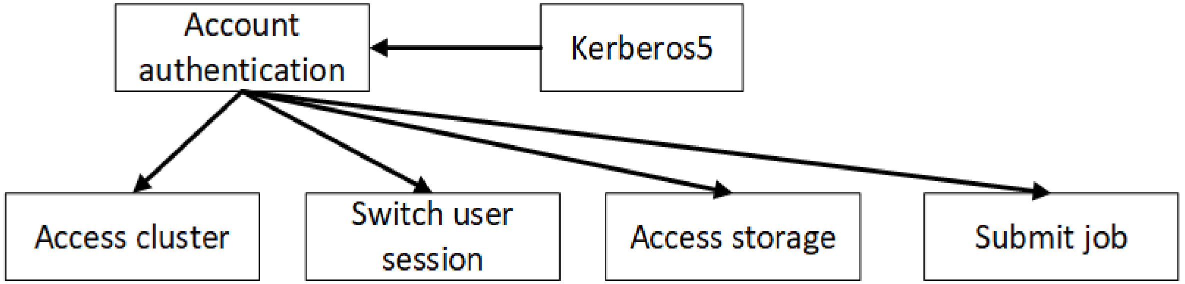
- Certificate and Token are the main authz&authn methods in grid computing
  - Many HEP sites have supported certificates and WLCG SCIToken in their site services
  - IHEP grid sites also support certificate and token to do authentication and authorization
- CEPC user authz&authn system is built and developed based on IAM
  - It is the suggested grid user management system by WLCG
  - Support user management, Access control, Authentication, Auditing and monitoring
  - Will be highly integrated to data processing with grid resources
  - Already support user authentication by INFN and IHEP SSO with eduGain

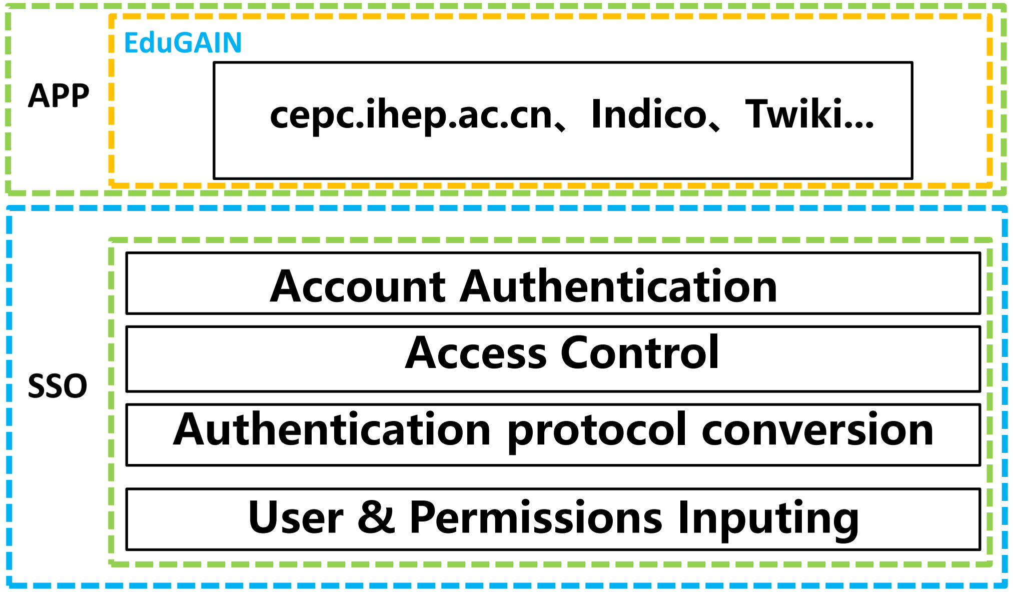




User AuthN&AuthZ System – Local User

- CEPC local user has two types of identity: computing and SSO accounts
- Computing account for local computing
  - Application, approval, creation, locking, password change, permission change
  - Ticket management based on Kerberos5 Token
- SSO account for public services
  - Implement by integrated in IHEP SSO
  - CEPC public services should be put behind IHEP SSO
  - Support CEPC web application, twiki, indico, etc.



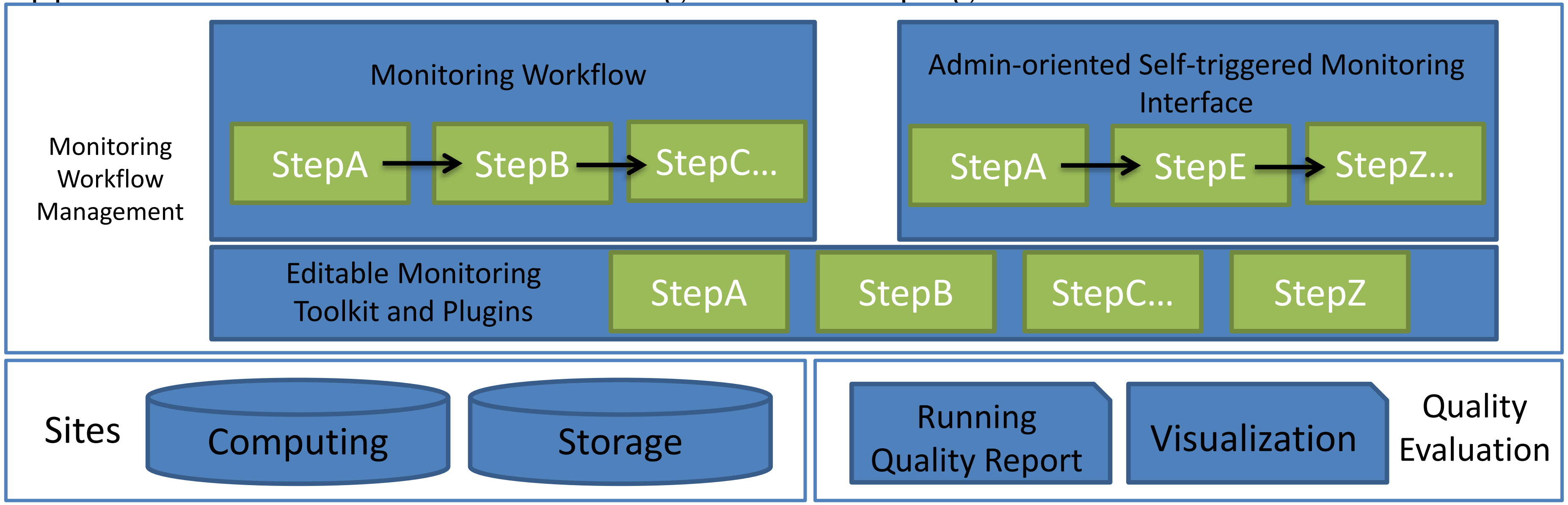

APP
EduGAIN
cepc.ihep.ac.cn、Indico、Twiki…
SSO
Account Authentication
Access Control
Authentication protocol conversion
User & Permissions Inputing

Other Systems

- Site Monitoring System
- Site Middleware and Service

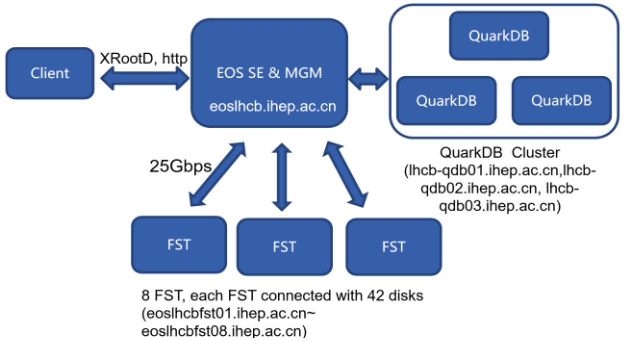
Site and Service Monitoring

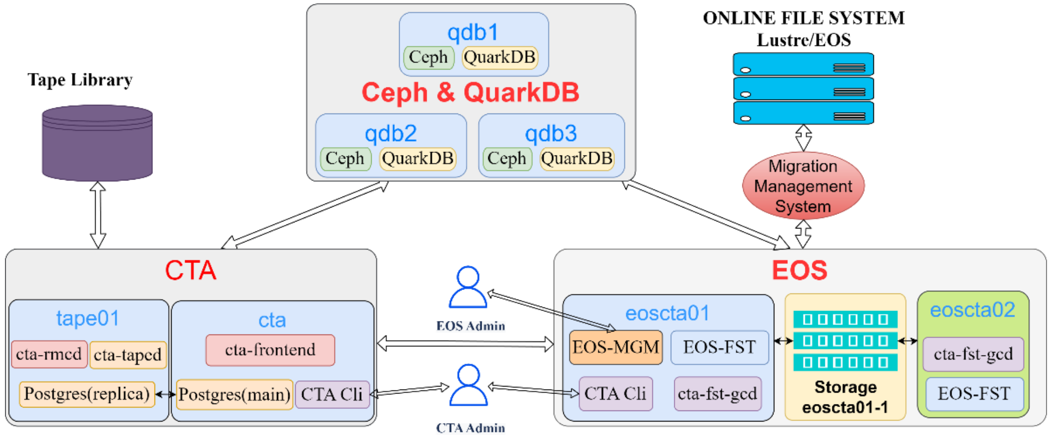
- To monitor each site status and service availability
  - Develop a monitoring platform, provide sites running status collection and metrics visualization.
  - Based on workflow system with developed site monitoring probers.
  - Provide a running quality evaluation system for each sites.
  - Support site admin-oriented monitoring toolkit and plugins interface.

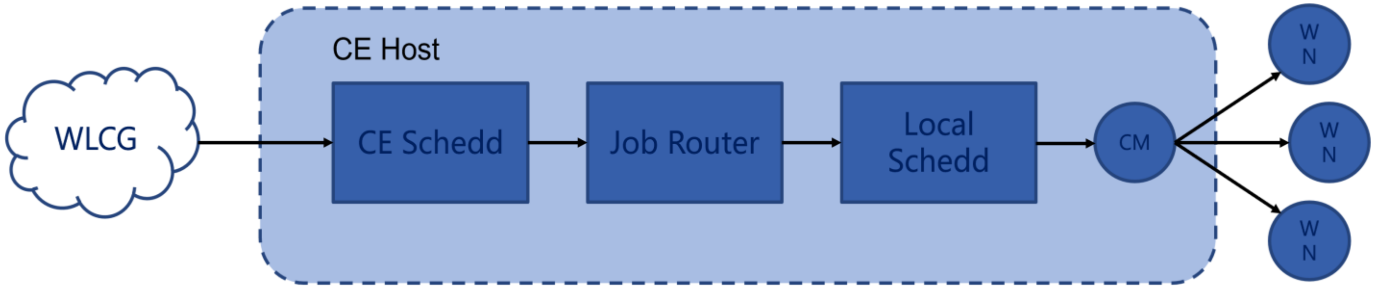


Site Middleware and Service

- A CEPC site need to equip with a set of middleware
  - Build a site middleware repository required by building a CEPC site
  - Including middleware, like CE/SE/Authentication/tape/…
- Disk storage: EOS
  - services: QuarkDB, MGM, FST
  - protocol: xrootd and http
- Tape storage: EOS & EOS-CTA
  - Protocols: xrootd and http
- CE: HTCondor-CE & HTCondor
  - Support for SCIToken and GSI
- Other middle software
  - Argus, BDII, APEL


Client
XRootD, http
EOS SE & MGM
eoslhcb.ihep.ac.cn
QuarkDB
QuarkDB
QuarkDB
QuarkDB Cluster
(lhcb-qdb01.ihep.ac.cn,lhcb-qdb02.ihep.ac.cn, lhcb-qdb03.ihep.ac.cn)
25Gbps
FST
FST
FST
8 FST, each FST connected with 42 disks
(eoslhcbfst01.ihep.ac.cn~
eoslhcbfst08.ihep.ac.cn)





Summary

- CEPC computing platform is developed based on WLCG standard, with specific development meeting CEPC requirements
- Distributed computing system
  - Manage the CEPC sites all over the world
  - Manage and dispatch the CEPC jobs to the worker nodes from multiple sites
- Distributed storage system
  - Manage the CEPC storage from the multiple sites, including disk and tape
  - Provide the policies of data distribution and data placement
- Network and data transfer system
  - Provide the functions to transfer data from site to site and support the popular protocal
  - Manage the network and monitor the status
- User authentication and authorization
  - IAM for Grid user management and IHEP-SSO for local user management
- Site/service monitoring and accounting

Thanks!

Q&A