Distributed computing for SPD offline data processing and massive MC simulation

A. Petrosyan, D. Oleynik, A. Zhemchugov

SPD Collaboration Meeting
June 9, 2021

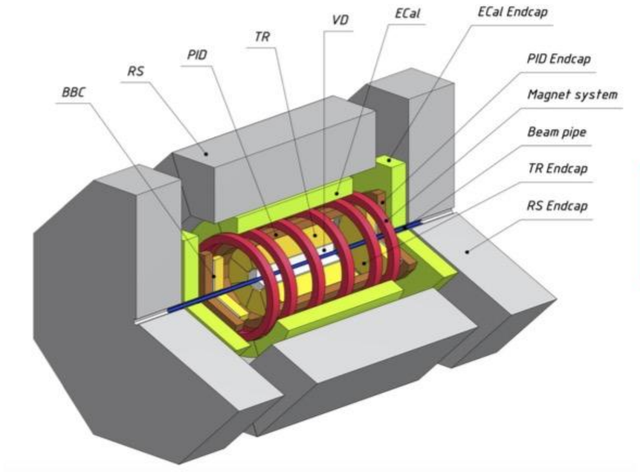
SPD INTERNATIONAL COLLABORATION

32 institutes from 14 states, ~300 members

The SPD international collaboration is forming actively

June, 4 - birthday of the SPD collaboration

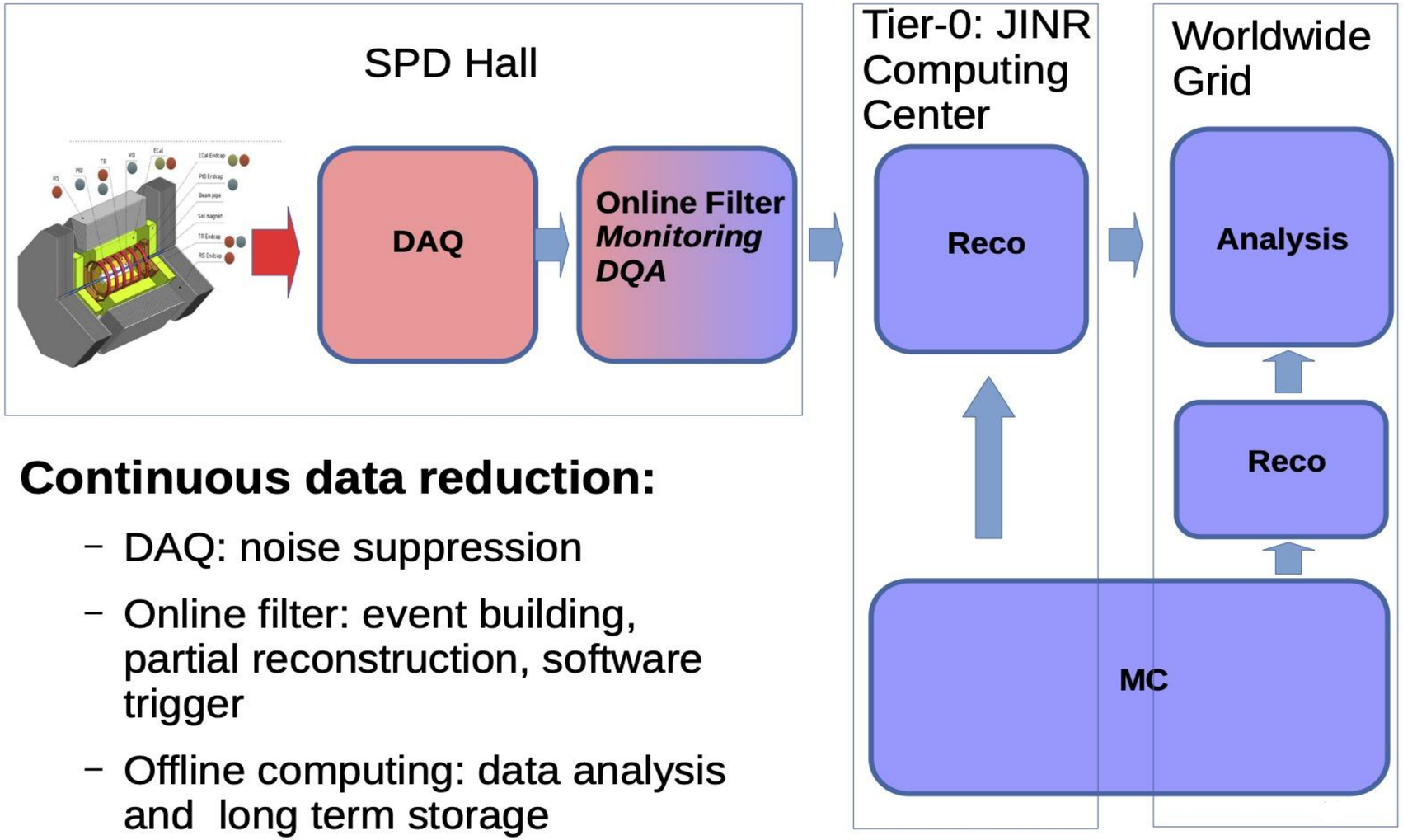
SPD as a data source



- Bunch crossing every 80 ns = crossing rate 12.5 MHz
- ~ 3 MHz event rate (at 10^{32} $cm^{-2}s^{-1}$ design luminosity) = pileups
- 20 GB/s (or 200 PB/year (raw data), $3*10^{13}$ events/year)
- Selection of physics signal requires momentum and vertex reconstruction → no simple trigger is possible

The SPD detector is a medium scale setup in size, but a large scale one in data rate!

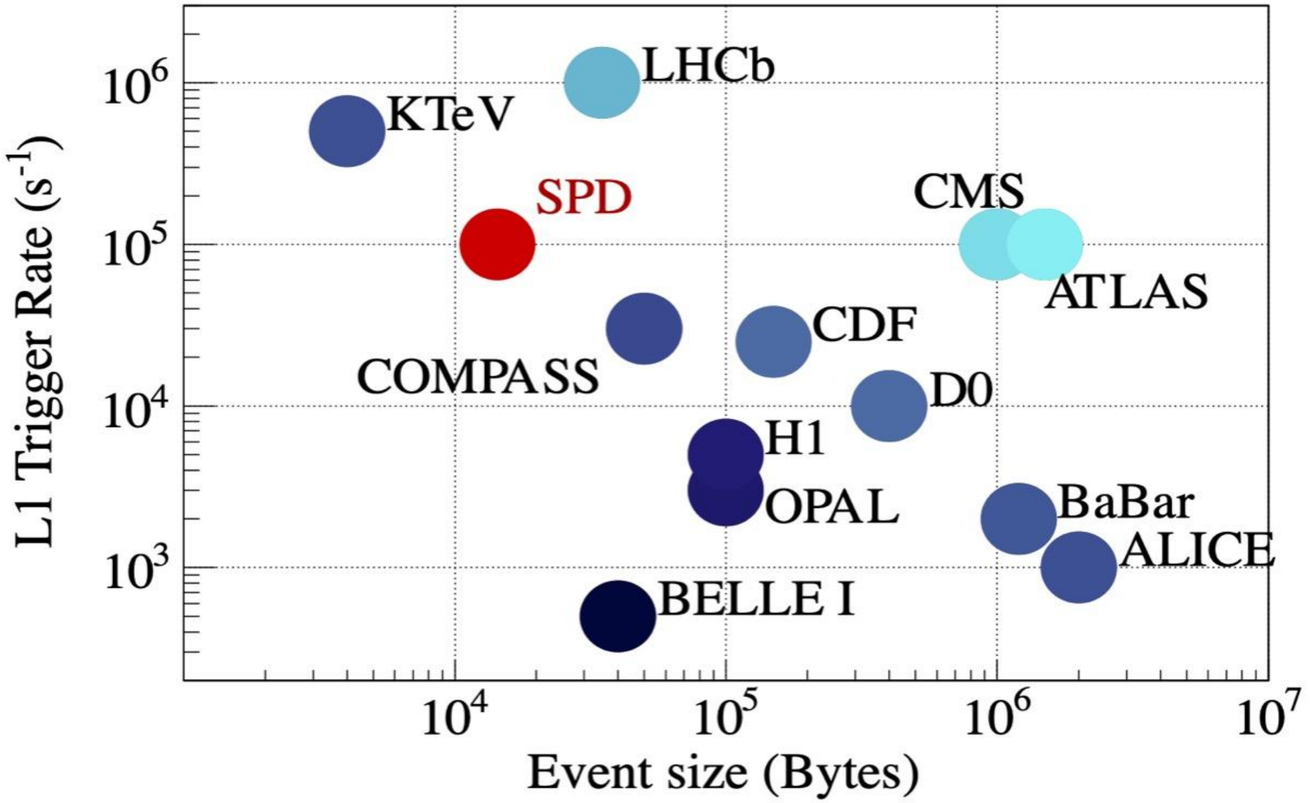
Data workflow



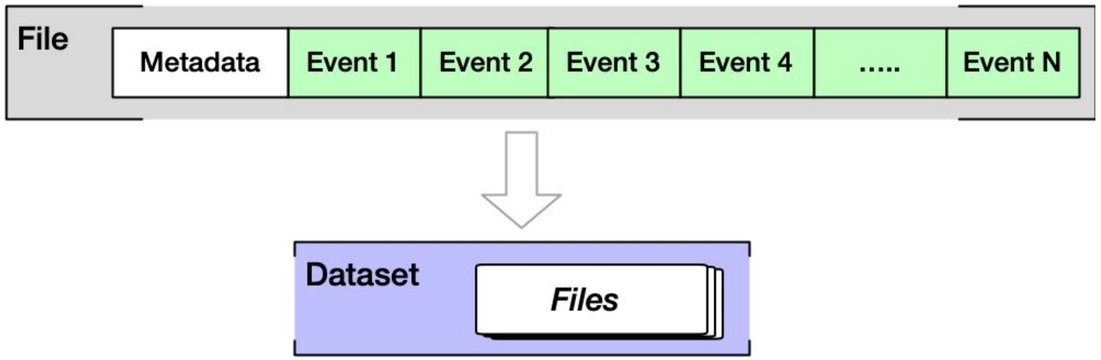
Continuous data reduction:

- DAQ: noise suppression
- Online filter: event building, partial reconstruction, software trigger
- Offline computing: data analysis and long term storage

After the online filter

- 

HEP Data & Processing speciality



Physics event is a unit of data

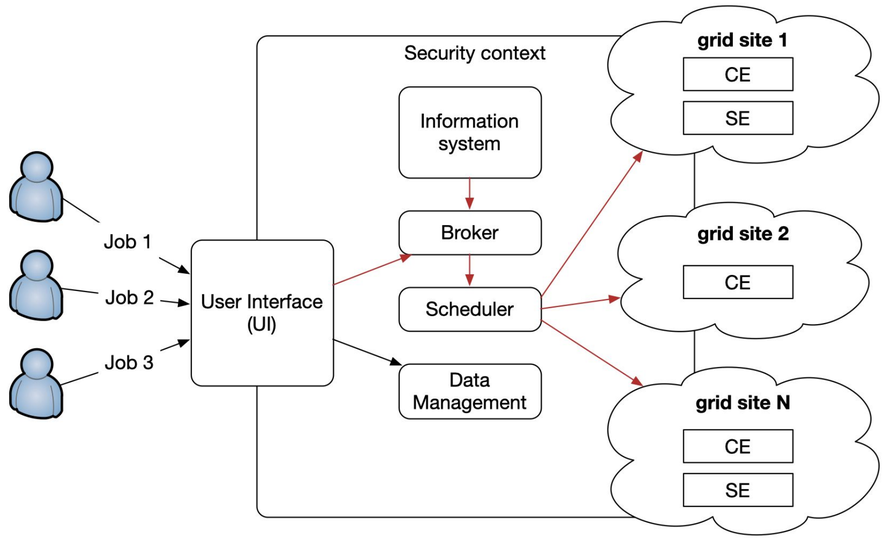
Each event can be processed independently, but processing of one event quite short, so one job should process a set of events to have reasonable number of jobs and reasonable amount of files.

HTC model works good - processing can be splitted to independent phases and distributed across a set of resources

No needs to have a huge single facility!

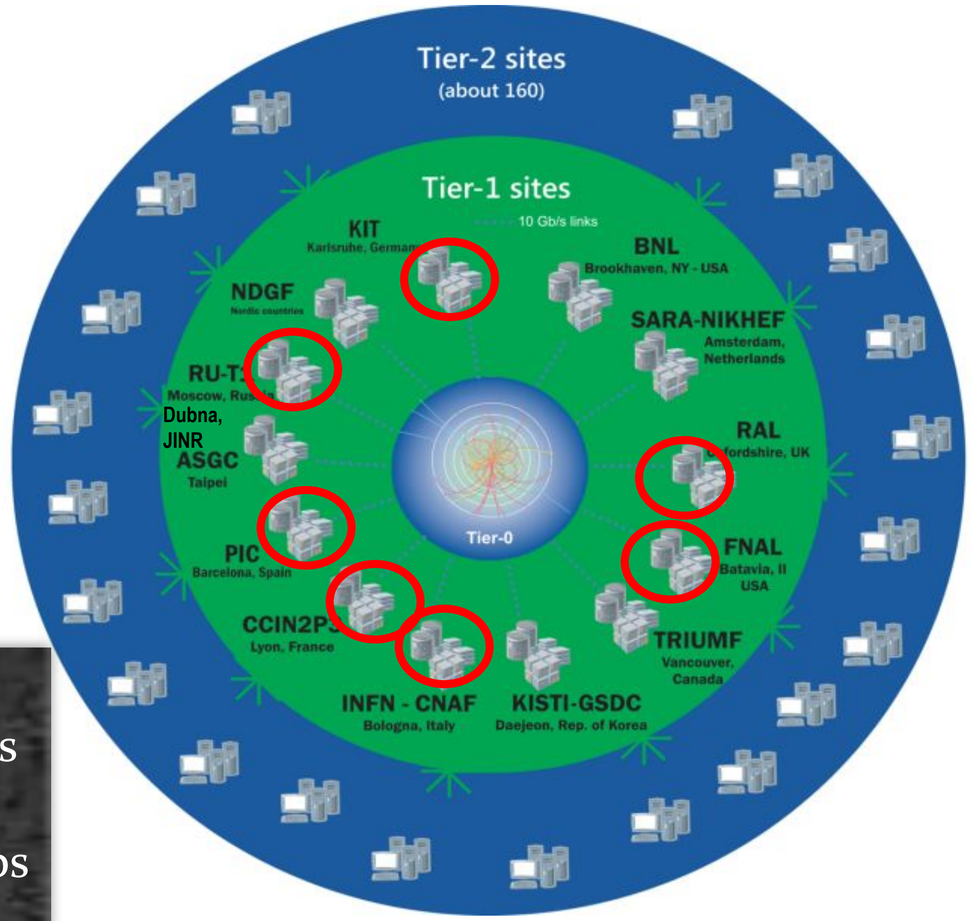
HTC: grid computing

Grid computing is the collection of computer resources from multiple locations to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. Grid computing is distinguished from conventional high performance computing systems such as cluster computing in that grid computers have each node set to perform a different task/application.



LHC computing model

The Worldwide LHC Computing Grid (WLCG): integrates computer centres globally to provide computing and storage resources into a single infrastructure accessible by all LHC physicists for data analysis



42 countries
~300,000 cores
173 PB storage
> 2 million jobs per day
>10 Gb links

Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11→14 centres):
- Permanent storage
- Re-processing
- Analysis
- Simulation

Tier-2 (>200 centres):
- Simulation
- End-user analysis

Paradigm shift in HEP computing

Batch processing

- Distributed resources are **independent entities**
- Groups of users utilize **specific resources** (whether locally or remotely)
- Fair shares, priorities and policies are managed **locally**, for each resource
- **Uneven user experience** at different sites, based on local support and experience
- Privileged users have access to **special resources**

Global distributed processing

- Distributed resources are seamlessly **integrated worldwide** through a single submission system
- **Hide middleware** while supporting diversity
- All users have access to **same resources**
- **Global** fair share, priorities and policies allow efficient management of resources
- **Automation, error handling, and other features** improve user experience
- Central support coordination
- All users have access to **same resources**

Data processing in a distributed computing infrastructure

- Distributed computing infrastructure is a computing system whose components are located on different remote computers and components interact with one another to achieve a common goal.
  - One well-known example: WLCG (Worldwide LHC Computing Grid)
  - JINR already have a set of facilities which can (should) be integrated into the distributed computing infrastructure
- Advantages of using distributed computing systems:
  - **High fault tolerance**: failure of a single computing facility is not a blocker for data processing chain
  - **Flexibility**: wide range of computing resources can be in integrated into a common infrastructure
  - **Balanced support expenses**: no needs to upgrade all computing facilities at the same time (etc.) Support expenses mostly on the facilities provider side

From Systems to Products

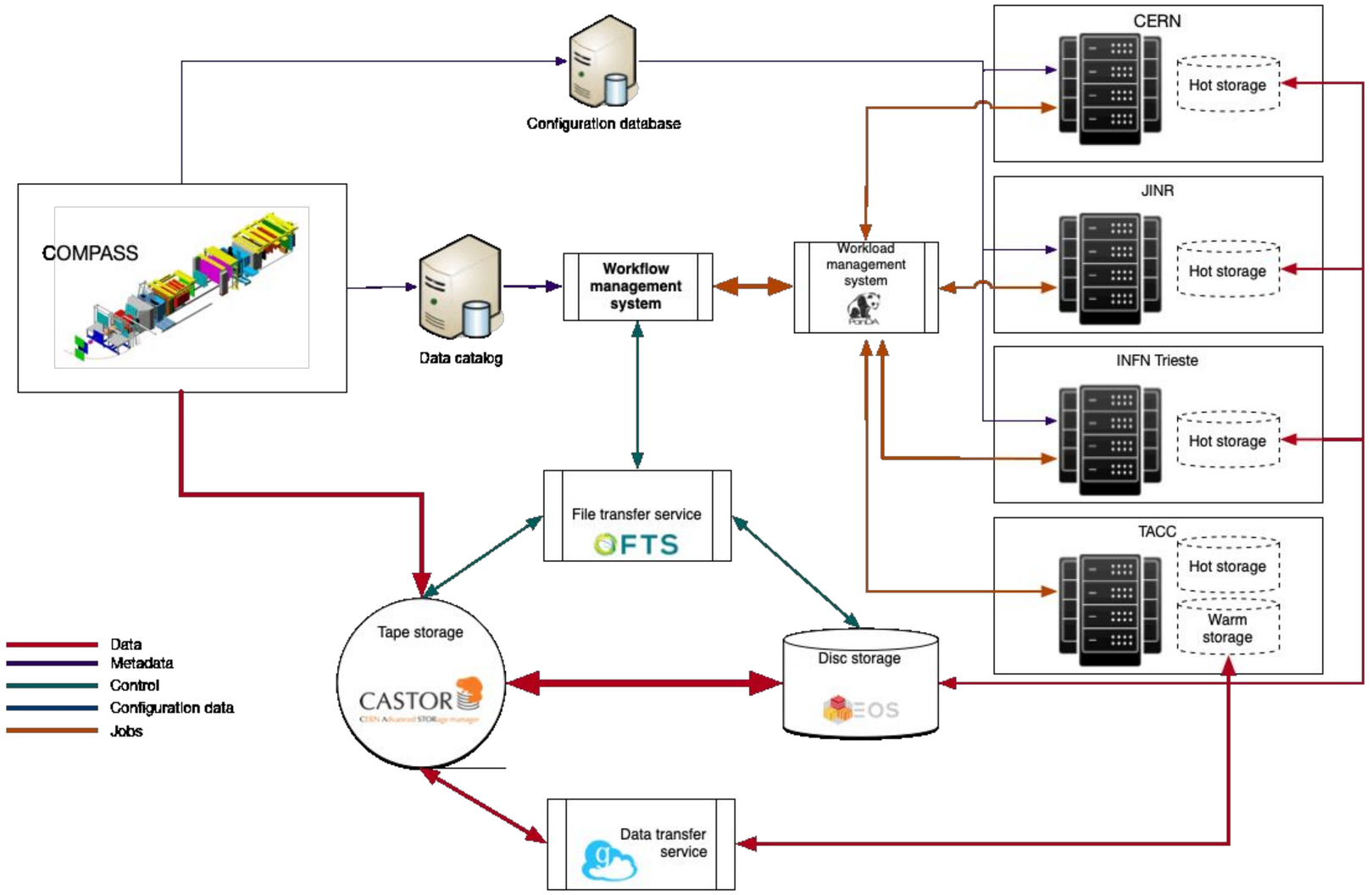
- Rucio: ATLAS -> AMS, CMS, BNL, Belle II, etc.
- AGIS/CRIC: ATLAS -> AMS, CMS, COMPASS, WLCG, etc.
- PanDA: ATLAS -> LSST, nEDM, SciDAC-4, CHARMM, AMS, IceCube, Blue Brain, COMPASS, etc.
- Dirac: LHCb -> Belle II, BES III, IHEP, ILC, Ibergrid, etc.

COMPASS mass data processing management

LIT is in charge of the development and support of the production management system for the COMPASS experiment since 2017

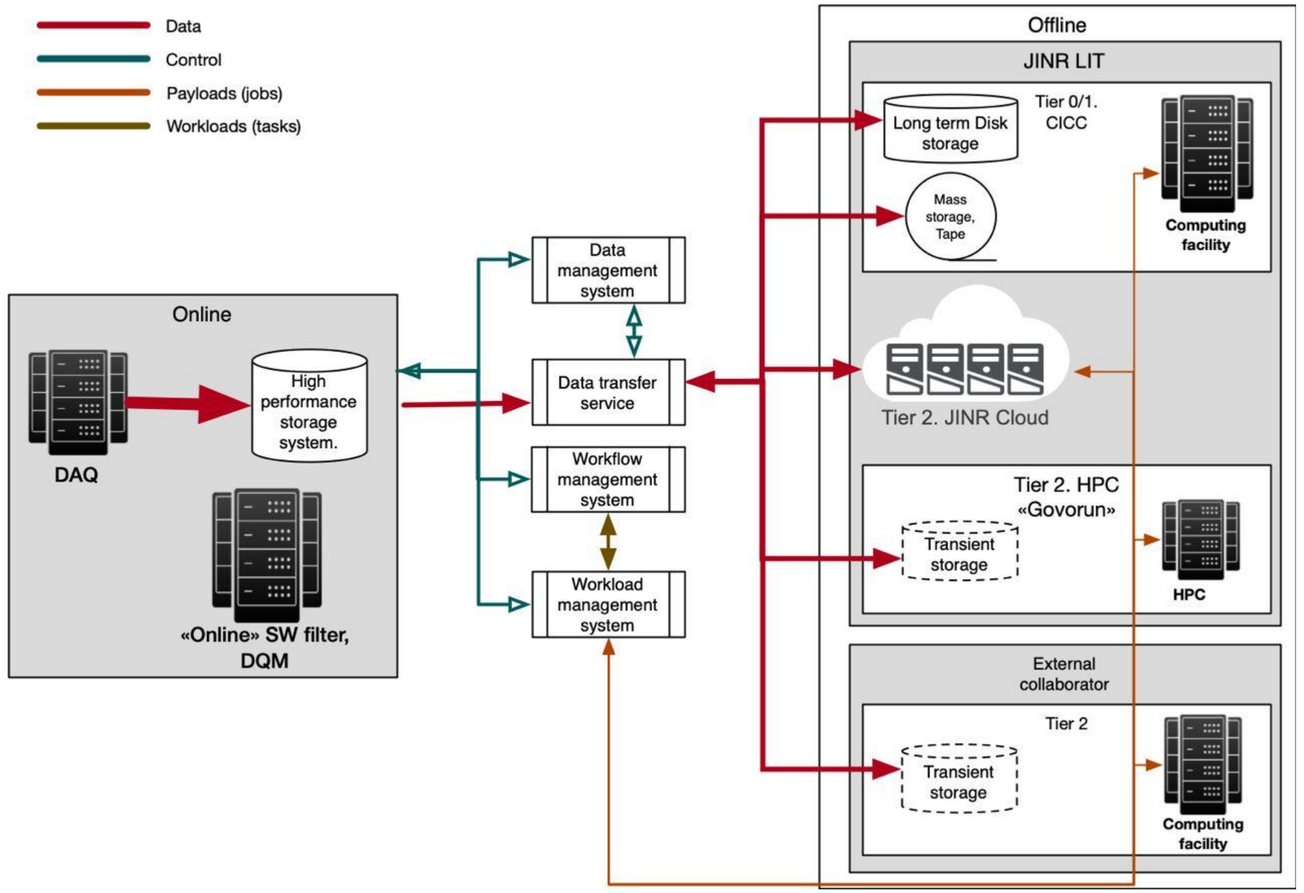
Workflow and workload management services, data processing monitoring deployed at LIT

During last three years ~600 tasks, 7 000 000 raw data files, 150 000 000 events, 12 000 000 jobs were processed by the system



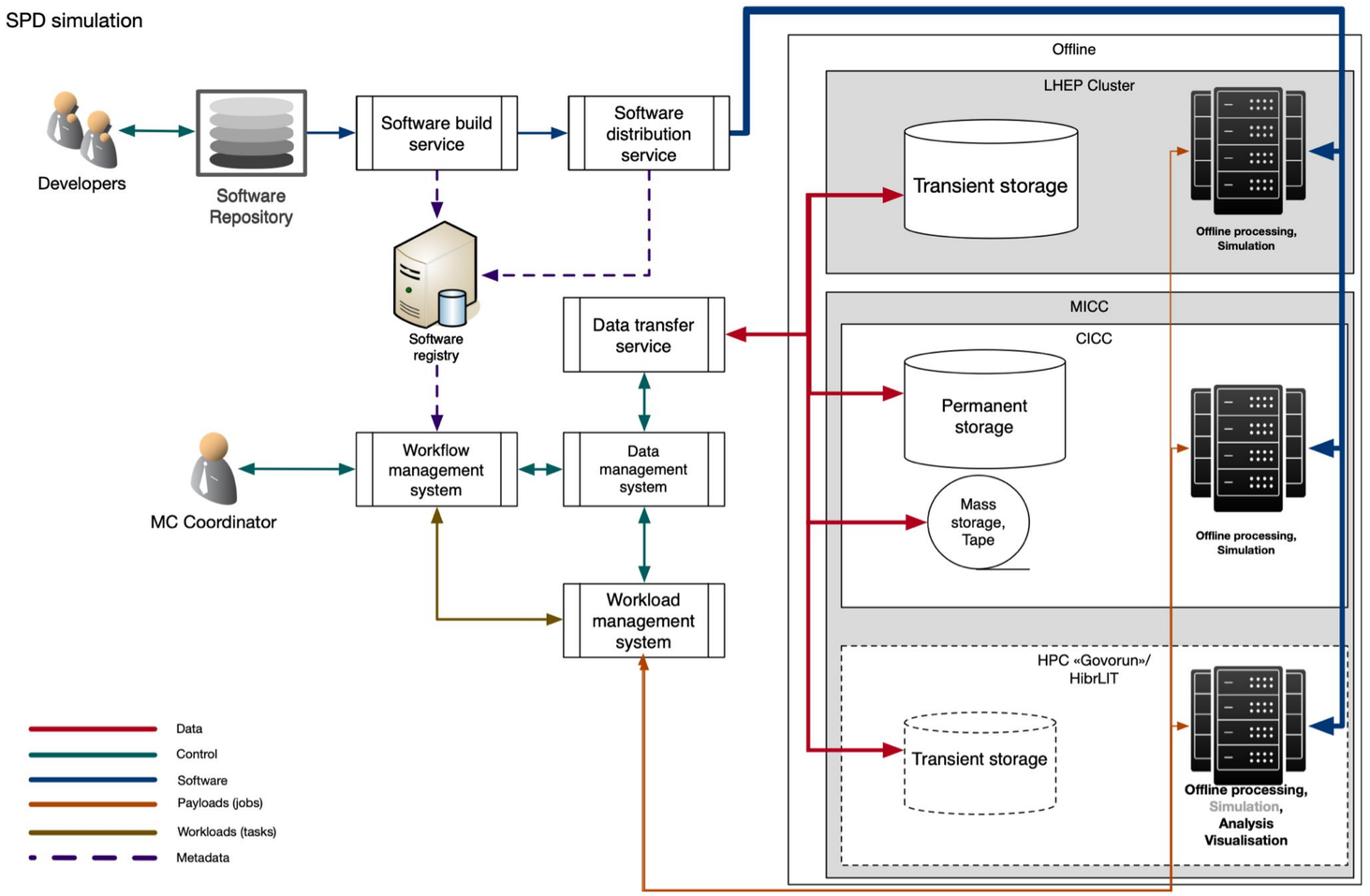
Use case: SPD mass data processing

- The main consumer of distributed computing resources
- It is very important to find the most optimal ratio of the size and number of files to be processed by the system for the most efficient use of the available computing infrastructure
- Key components:
  - WFMS
  - WMS
  - DMS & DTS
  - Software distribution service - service which allow automatic deployment of new versions of SW in heterogeneous environment

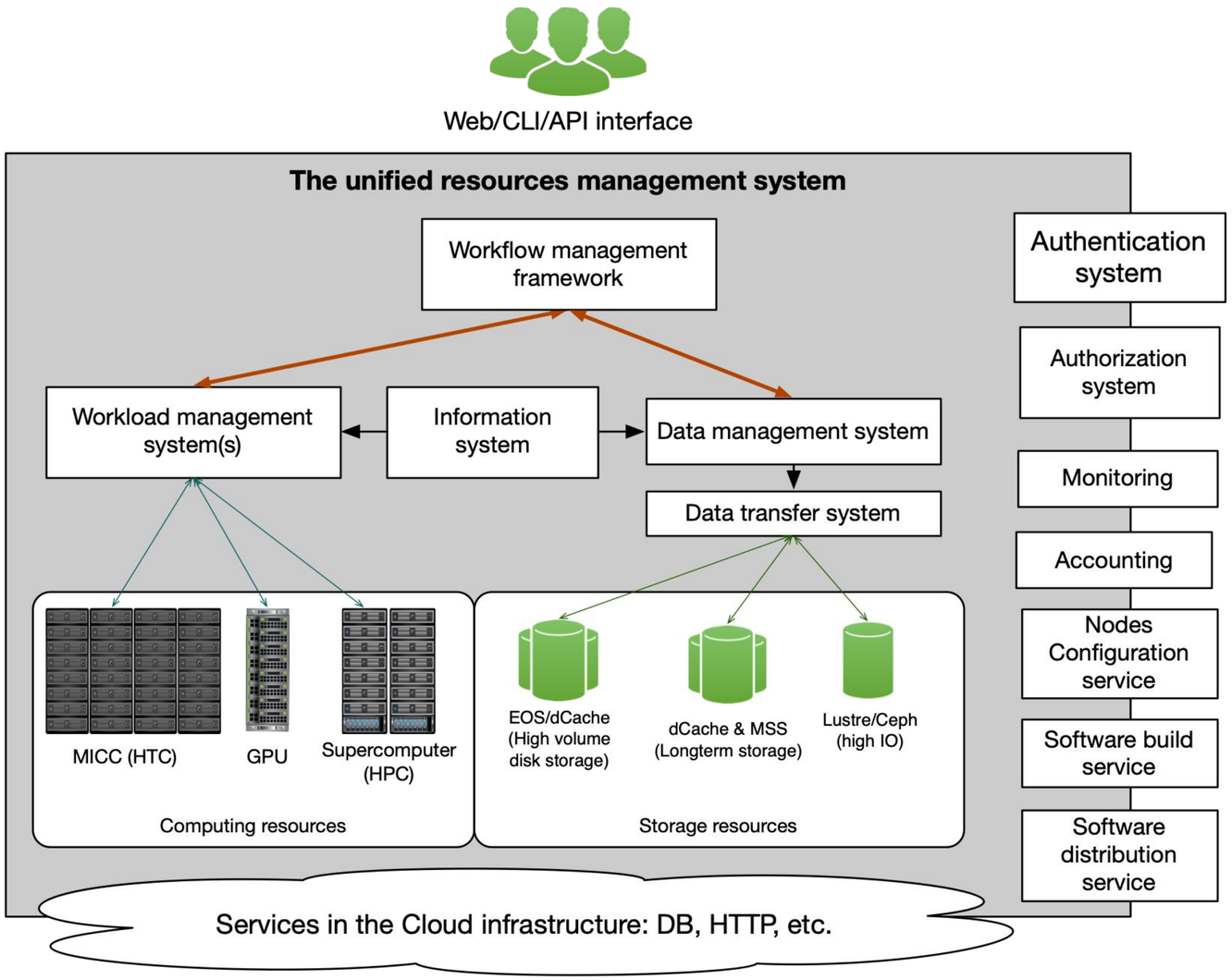


Use case: SPD simulation

- Simulation – another huge consumer of computing resources
- Can be (should be) started before facility will be ready to collect data
- Accompanied by intensive software development
- Key components:
  - WFMS
  - WMS
  - DMS & DTS
  - Software build service – required for automation of building of new releases of SW
  - Software distribution service - service which allow automatic deployment of new versions of SW in heterogeneous environment

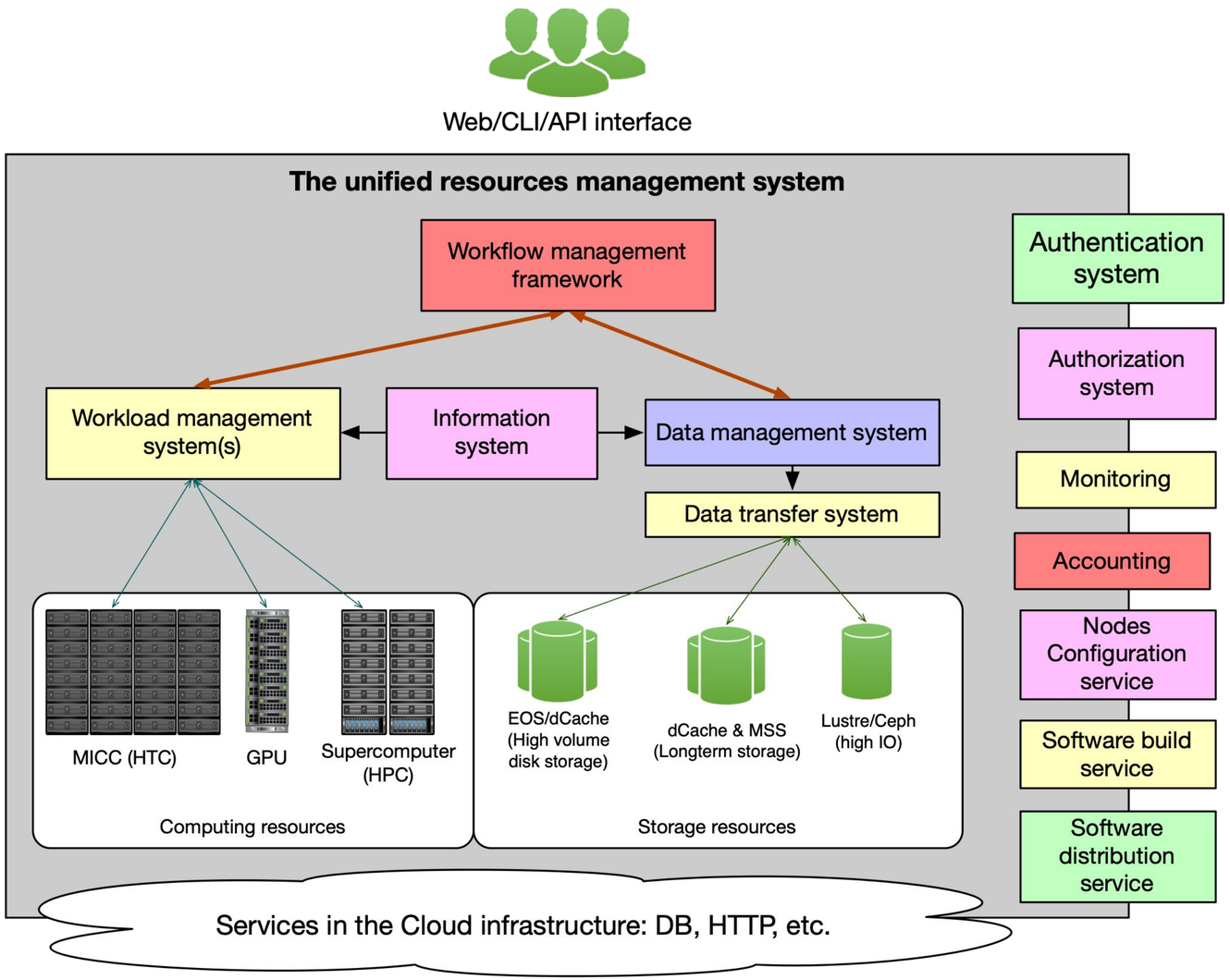


Unified Resource Management System



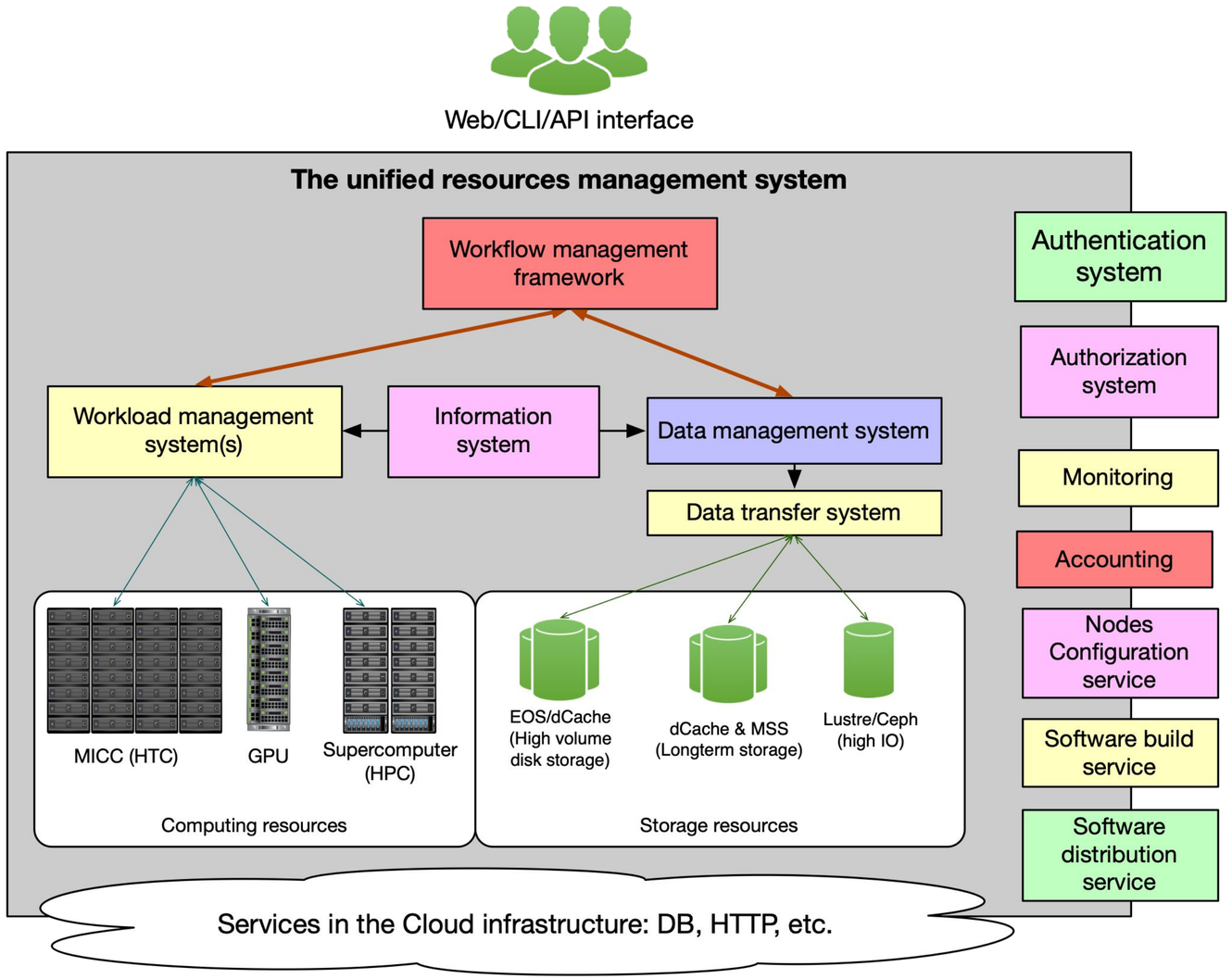
- The Unified Resource Management System is a IT ecosystem composed from the set of subsystem and services which should:
  - Unify of access to the data and compute resources in a heterogeneous distributed environment
  - Automate most of the operations related to massive data processing
  - Avoid duplication of basic functionality, through sharing of systems across different users (if it possible)
  - As a result - reduce operational cost, increase the efficiency of usage of resources,
  - Transparent accounting of usage of resources

URMS: first steps



- Some core subsystem already exist in JINR
  - Authentication system (Kerberos based, with SSO supporting for Web applications)
  - CVMFS as Software distribution service
- In progress:
  - deployment of FTS as the core of Data transfer system
  - We already have some infrastructure monitoring
  - A lot of research in WFMS and WMS fields, we may declare a list of requirements:
    - We should avoid limitations by scale as much as possible.
    - Advanced monitoring system
    - WMS with MultiVO support
    - Priority and share management
    - Task-based job management
  - Looks like that Rucio will be natural choice as cross experiment Data Management System
  - Software build service prototype already exists in the Cloud infrastructure

URMS: next steps



- Common Authorization System which will be used to manage user access to resources. The closest candidate is VOMS - but, we need to be coherent with Authentication System
- Accounting is required to understand system behaviour and analysing of bottlenecks.
- Nodes configuration - should be automated as much as possible
- Information system store and provide a description of computing and storage resources, including availability (shutdowns) of resources.

Status and plans

- We’re collecting info about requirements, data volumes and data flows
- A prototype, based on the components, which have proven their ability to handle expected data volumes, is being built
- Some services, like authorization/authentication, VOMS, CVMFS, FTS are already deployed and will be supported by LIT as part of the Unified Resource Management System
- We’re looking forward to integrate external computing and storage resources, which members of our collaboration will be ready to provide

Thank you!