SPD Online Filter

High-Throughput Processing Middleware

Nikita Greben

Meshcheryakov Laboratory of Information Technologies

11th International Conference
"Distributed Computing and Grid Technologies in Science and Education" (GRID'2025)

July 7-11, 2025

SPD experiment at NICA collider

Detector

- Polarized proton and deuteron beams.
- Collision energy up to 27 GeV.
- Luminosity up to 10^{32} cm^{-2} s^{-1}.
- Bunch crossing every 80 ns = crossing rate 12.5 MHz.

Key Challenges

- Number of registration channels in SPD $\approx$ 500000.
- Physics signal selection requires momentum and vertex reconstruction $\rightarrow$ no simple **trigger** is possible.
- The goal of the online filter is to reduce the data stream so that the annual increase in data, including modeled samples, does not exceed 10 PB.

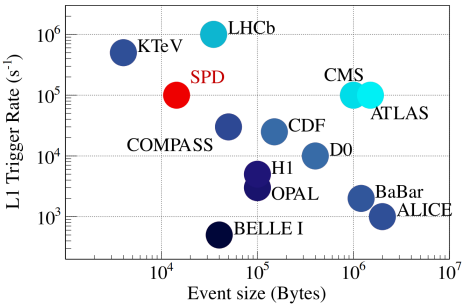


Figure 1: Expected event size and event rate of the SPD setup after the online filter, compared with some other experiments.

SPD DAQ

Data Acquisition System

The **DAQ** system takes raw data from detector sensors and ensures that only the most interesting events (collisions) are recorded for later analysis, while discarding unimportant ones due to limitations in storage and processing.

Triggerless DAQ

Triggerless DAQ means that the output of the system is not a set of raw events, but a set of signals from sub-detectors organized into time slices.

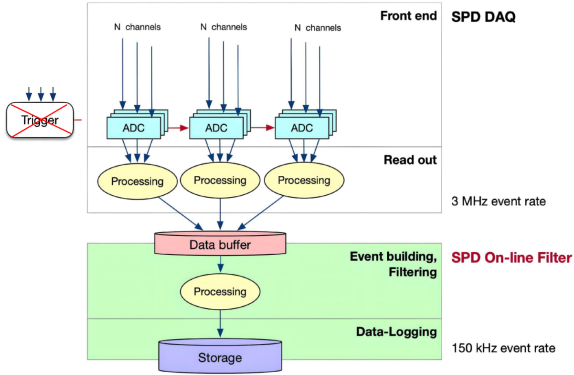


Figure 2: Triggerless dataflow in SPD

SPD Online Filter

SPD Online Filter is a **primary** data processing facility designed for the high-throughput, multi-step processing of data from the SPD detector.

Hardware component

Compute cluster with two storage systems and set of working nodes: multi-CPU and hybrid multi CPU + Neural Network Accelerators (GPU, FPGA etc.)

Middleware component

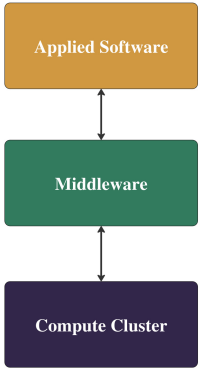
Software complex for management of multistep data processing and efficient loading (usage) of computing facility.

Applied software

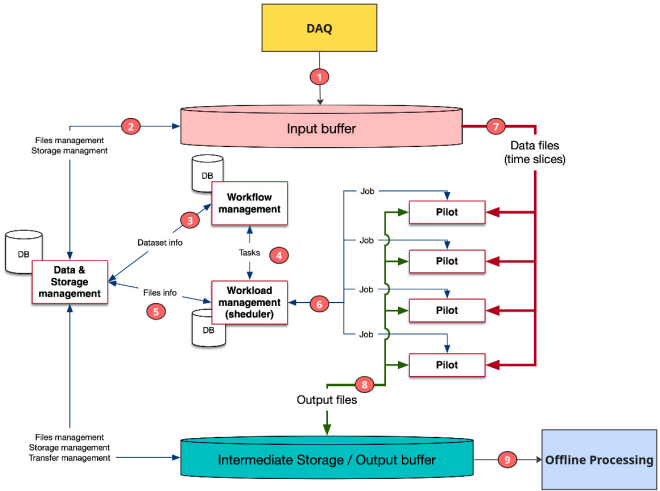
Performs informational processing of data.

SPD Online Filter Middleware

- **Data Management System**
  - Data lifecycle support (data catalog, consistency check, cleanup, storage);
- **Workflow Management System**
  - Define and execute data processing chains by generating the required number of computational tasks;
- **Workload management System**
  - Create the required number of processing jobs to perform the task;
  - Control job execution through pilots working on compute nodes;
  - Handles efficient use of resources.



High-level architecture of SPD Online Filter



High-throughput computing

Definition

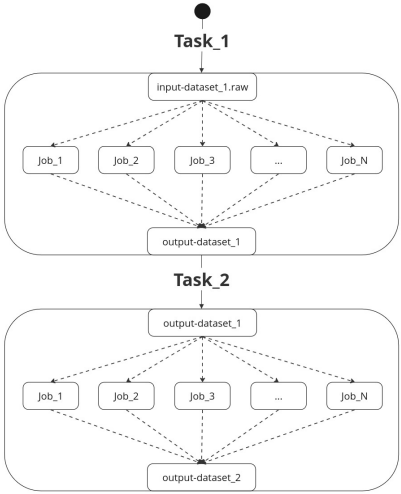
The European Grid Infrastructure defines **HTC** as "a computing paradigm that focuses on the efficient execution of a large number of loosely-coupled tasks".

Focus

Maximizing the number of tasks processed per unit of time.

Reliability

HTC systems are mostly designed to provide high reliability and make sure that all tasks run efficiently even if any one of the individual components fails.



Tasks and Jobs relationship

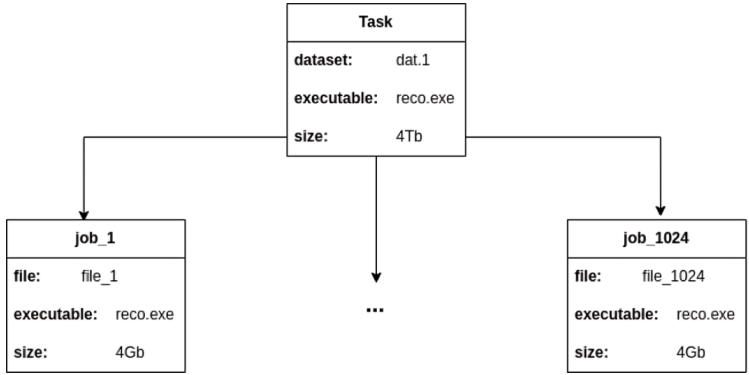


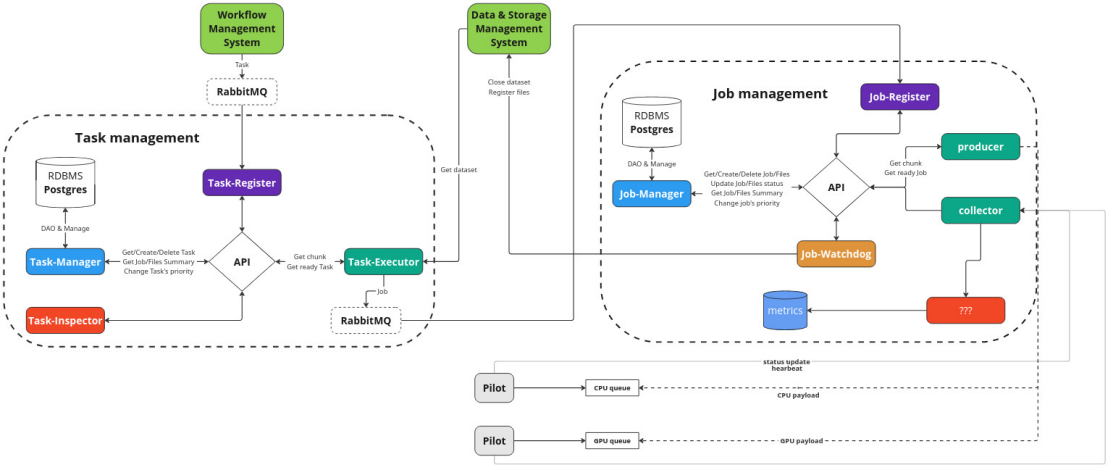
Figure 3: The goal of the entire task is to completely process the input dataset. Each job is assigned a file for further processing. The sum of the resulting output data from all jobs is the output of the entire task

Workload Management System

Responsibilities

The **Workload Management System** is responsible for partitioning each task into individual processing jobs, dispatching those jobs to **Pilot** agents on compute nodes, and monitoring their execution. It ensures efficient resource utilization by generating the appropriate number of jobs, tracking status, handling retries or failures, and aggregating output files into new datasets.

Workload Management System Architecture



Workload Management System as Main Scheduling Mechanism

Role of Scheduling in WMS

In the Workload Management System (WMS), scheduling fulfills two primary functions:

- Partitioning each task (dataset) into quanta for job generation based on dataset priority – IWRR based scheduler.
- Distributing ready jobs to compute nodes (Pilots) according to job priority – rank-based scheduler.

Interleaved Weighted Round-Robin (IWRR) Based Scheduler

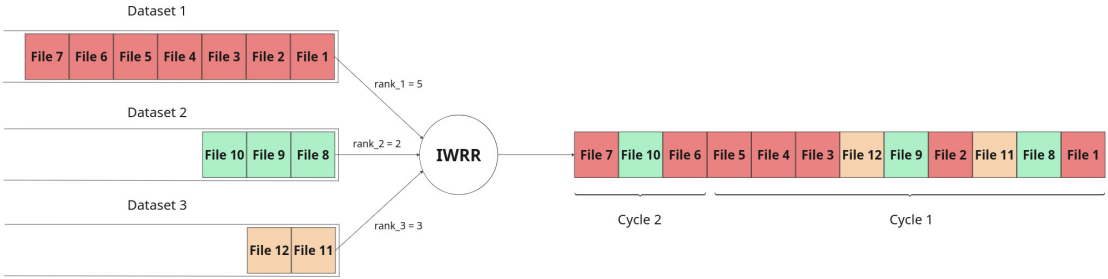


Figure 4: An example of the job generation procedure across several datasets being processed concurrently.

Rank-Based Job Distribution Scheduler

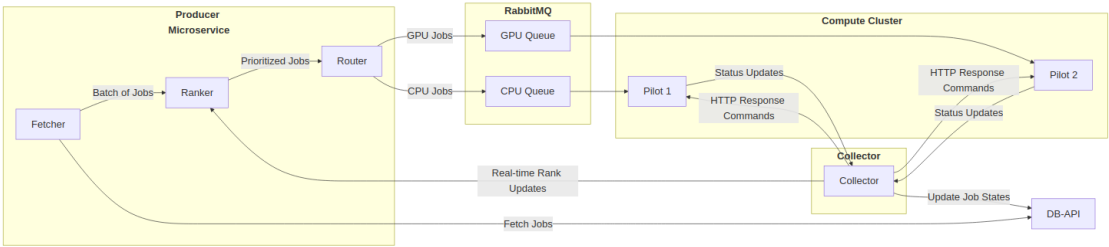


Figure 5: Simplified diagram of the Job Distribution Scheduler's working process.

Analogy with an Operating System

OS Component	SPD Online Filter	Explanation
Kernel	Middleware (WMS, WfMS, DSM)	The middleware manages resources, schedules tasks/jobs, and handles data flow, similar to an OS kernel coordinating system operations.
Scheduler	Task Executor in WMS	The task executor distributes jobs to pilots using IWRR based on task ranks, similar to an OS scheduler's role in allocating CPU time to processes/threads.
File System	Data & Storage Management (DSM)	DSM manages data lifecycle (cataloging, consistency, storage), similar to an OS file system handling file storage and access.
Inter-Process Communication (IPC)	RabbitMQ/Message Broker	RabbitMQ facilitates communication between components (e.g., task executor to job register, producer to pilots).
System Calls	REST APIs (to WfMS, WMS, DSM)	The middleware uses REST APIs for communication between its microservices. This is similar to how an OS uses system calls or kernel APIs to coordinate between kernel components or provide services to user processes.

First “load testing”

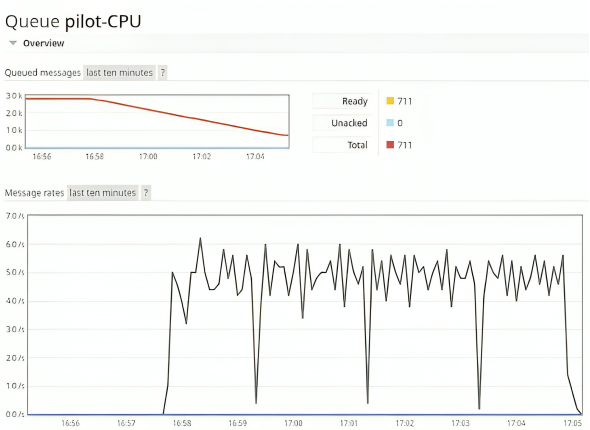


Figure 6: 100 concurrent Pilots processed ≈2,100 jobs in 7 minutes (≈15 s/job including stage-in/out) on standard JINR Cloud VMs using a simplified synthetic payload.

First “load testing”

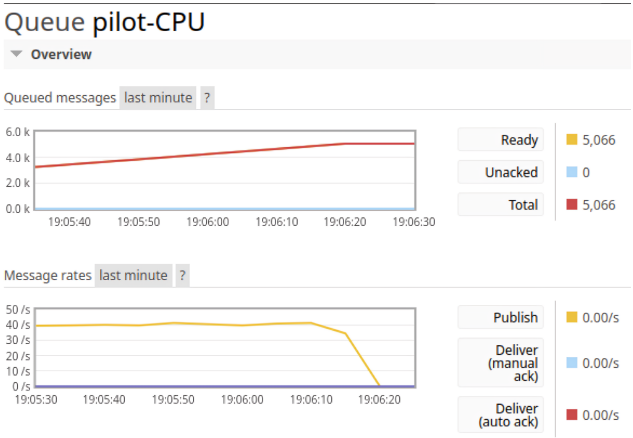


Figure 7: **Workload Management System** generates $\approx$ 5000 jobs in less than a minute.

Summary

- **Codebase and deployment**
  - Around ≈25 000 lines of code for the entire SPD Online Filter Middleware;
  - Full deployment requires ≈16 Docker containers: one container per microservice;
  - Integration infrastructure deployed on JINR Cloud resources on 9 VM's.
- **Task and workflow processing has been achieved**
  - Execution of the entire workflow set up on the level of Workflow Management System;
  - The major cycle of refactoring and test coverage is required.

Next steps and milestones

- **Middleware and applied software integration**
  - Requires prototyped applied software and simulated data;
  - Non-functional requirements for applied software;
  - Move to the execution of the jobs on the pilot with a "real" payload.
- **Middleware deployment and release management**
  - Focus on shipping SPD Online Filter as standalone software;
  - Work on the deployment on the **upcoming testbed** (256 CPU Cores, 1TB RAM, 120TB HDD);
  - Select the appropriate release management strategy.

Thank you for your attention!

Backup slides

SPD Online Filter

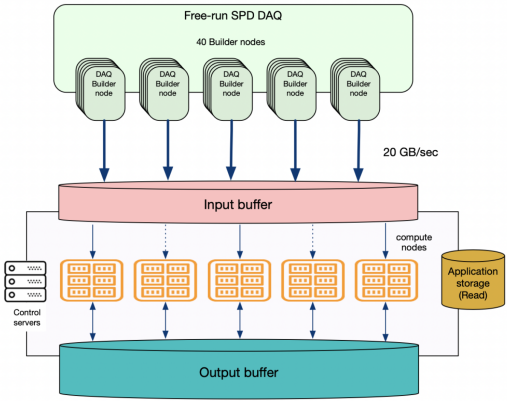


Figure 8: SPD Online Filter Facility

Data Processing in SPD Online Filter

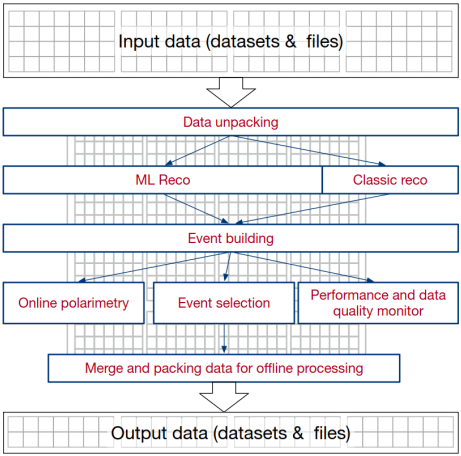


Figure 9: Example of multi-step processing scheme

Workflows execution

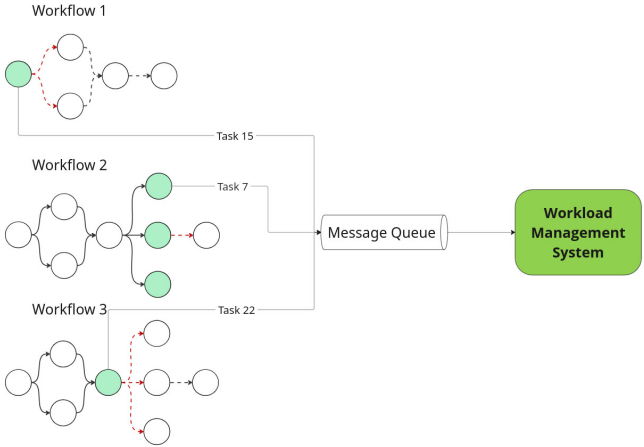
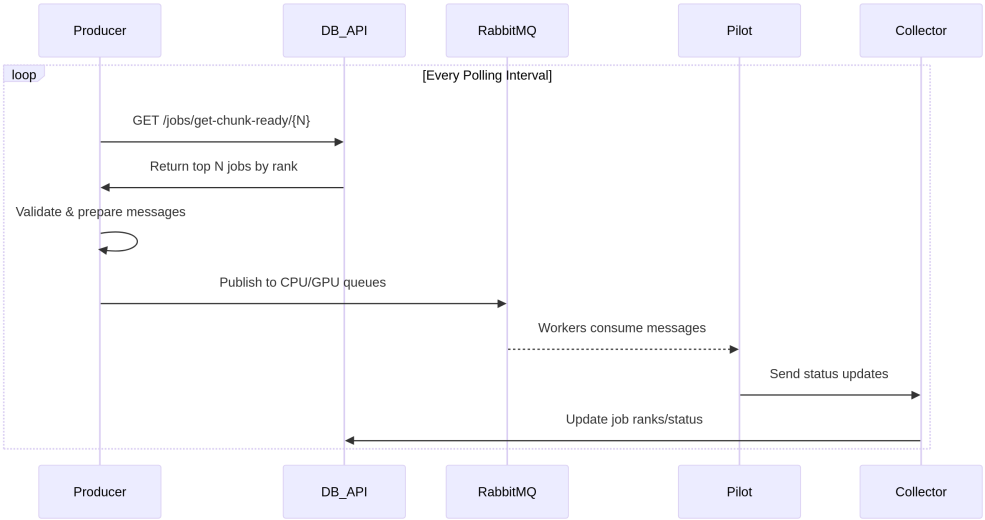


Figure 10: An example of the concurrent execution of several workflows.

Rank-Based Job Distribution Scheduler



Next Life Plan - Control Theory's Dynamic Adaptability Scheduler

- Each dataset has a rank (priority) that determines its processing order;
- Tasks are processed in priority order, with dynamic updates to maintain system responsiveness;
- **Priority-based job scheduling mechanism** is expected, with rank update scheme involving **Control Theory** (option to be explored later);
- Not applicable at this stage of the development process.

$$r_{i+1} = \underbrace{\alpha \ln(x_i + 1)}_{\text{Aging}} - \underbrace{\beta 2^{y_i}}_{\text{Retry Penalty}} + \underbrace{\gamma r_i}_{\text{History}} + \underbrace{\delta(1 - L)}_{\text{Load}}$$

$$\mathbf{r}_{i+1} = \Gamma \mathbf{r}_i + \alpha \ln(\mathbf{x}_i + \mathbf{1}) - \beta \cdot 2^{\mathbf{y}_i} + \delta(1 - L)\mathbf{1}$$

$\Gamma = \text{diag}(\gamma_1, ..., \gamma_N)$ (job-specific history weights)
$\mathbf{x}_i = [x_i^{(1)}, ..., x_i^{(N)}]^\top$ (job ages)
$\mathbf{y}_i = [y_i^{(1)}, ..., y_i^{(N)}]^\top$ (retry counts)