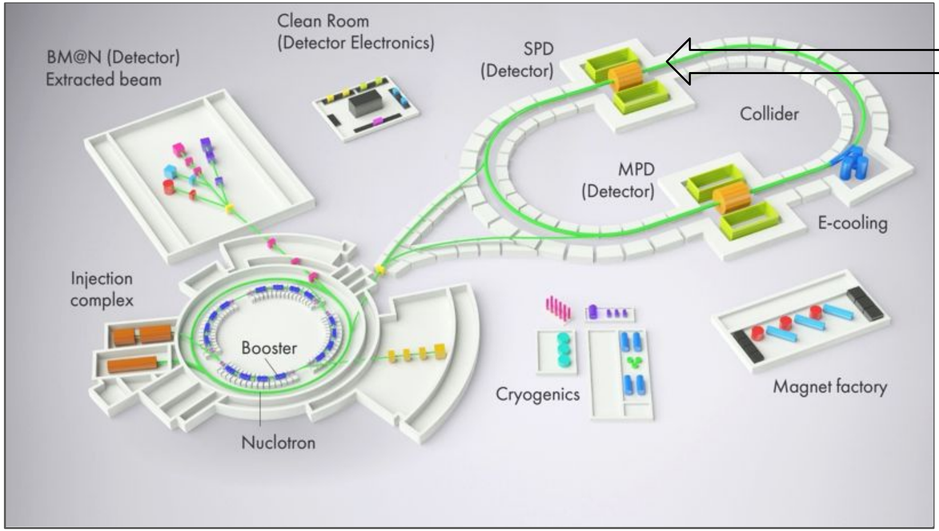
Rucio data management system for the SPD experiment

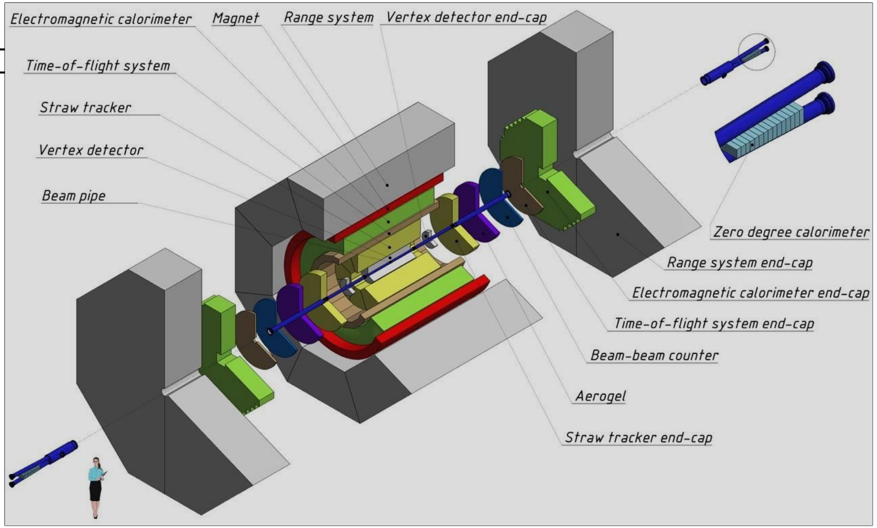
Alexey Konak, Danila Oleynik, Artem Petrosyan
JINR MLIT
konak@jinr.ru

Eighth Rucio Community Workshop
Square Kilometre Array Observatory Global Headquarter
05.11.2025

Spin Physics Detector (SPD)

The spin structure of the **nucleon** is one of the fundamental properties of matter. The spin of a nucleon is distributed between its components — **quarks** and **gluons**, and their mutual movement.
The EMC, HERMES, and COMPASS experiments have made it possible to study in detail the contribution of **quarks** to spin. However, the role of **gluons** remains poorly understood and requires further research.





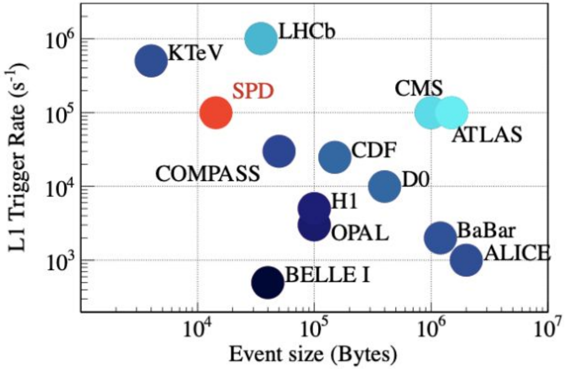
The SPD facility is being developed for a more accurate study of the contribution of **gluons** to the spin of the **nucleon**.

V. Abazov et al. (SPD Collaboration). "Technical Design Report of the Spin Physics Detector at NICA" Natural Sci. Rev. 1 1 (2024) https://doi.org/10.54546/NaturalSciRev.100101

SPD as data source

The expected event rate of the SPD experiment is about 3 MHz (pp collisions at $\sqrt{s}$ = 27 GeV and 10^{32} $cm^{-2}s^{-1}$ design luminosity). This is equivalent to a raw data rate of 20 GB/s or 200 PB/year, assuming a detector duty cycle is 0.3, while the signal-to-background ratio is expected to be on the order of 10^{-5}. Taking into account the bunch-crossing rate of 12.5 MHz, one may conclude that pile-up probability cannot be neglected.

- SPD TDR



The goal of the online filter is at least to decrease the data rate by a factor of 20, so that the annual growth of data, including the simulated samples, stays within 10 PB. Then, data are transferred to the Tier-1 facility, where a full reconstruction takes place and the data is stored permanently. The data analysis and Monte-Carlo simulation will likely run at the remote computing centres (Tier-2s). Given the large data volume, a thorough optimization of the event model and performance of the reconstruction and simulation algorithms are necessary.

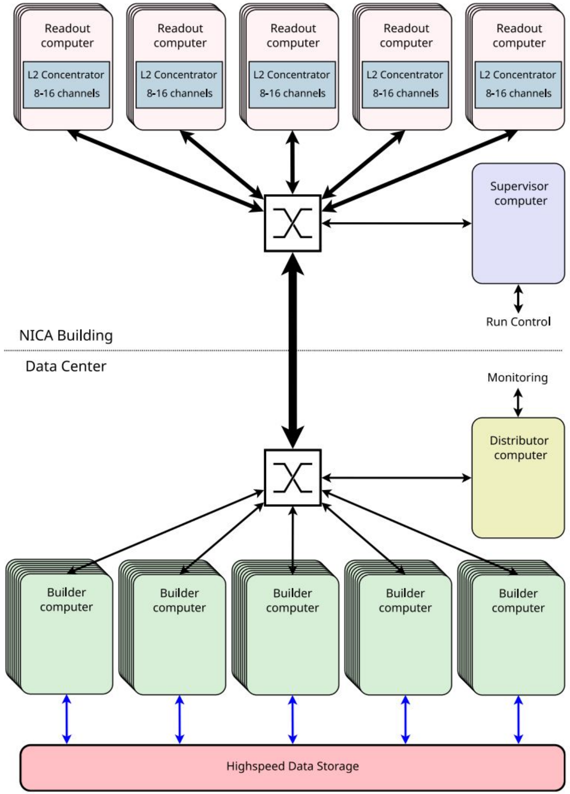
SPD DAQ

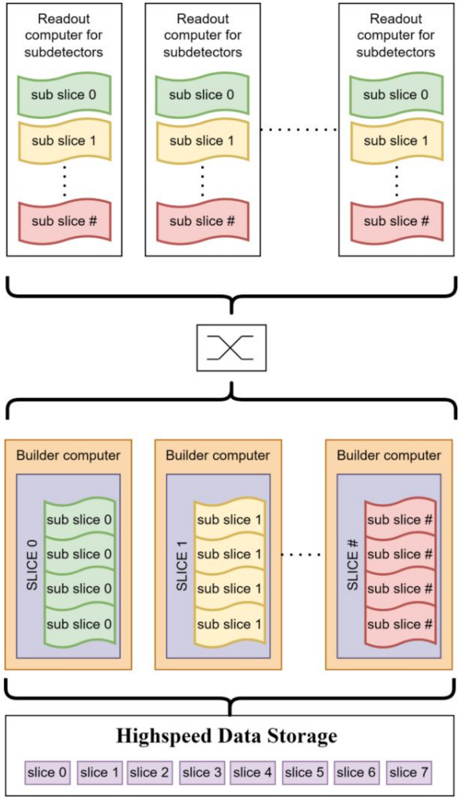
Why triggerless DAQ:

- No fast detectors to form a “classic” trigger signal.
- No 4π detector with 100% efficiency to detect the collisions.
- The wide SPD physical program eliminates a possibility of rejecting events at a hardware level.

Free run DAQ Speciality

- DAQ provide data organized in time slices which placed in files with reasonable size (a few GB).
- Each of these file may be processed independently as a part of top-level workflow chain.





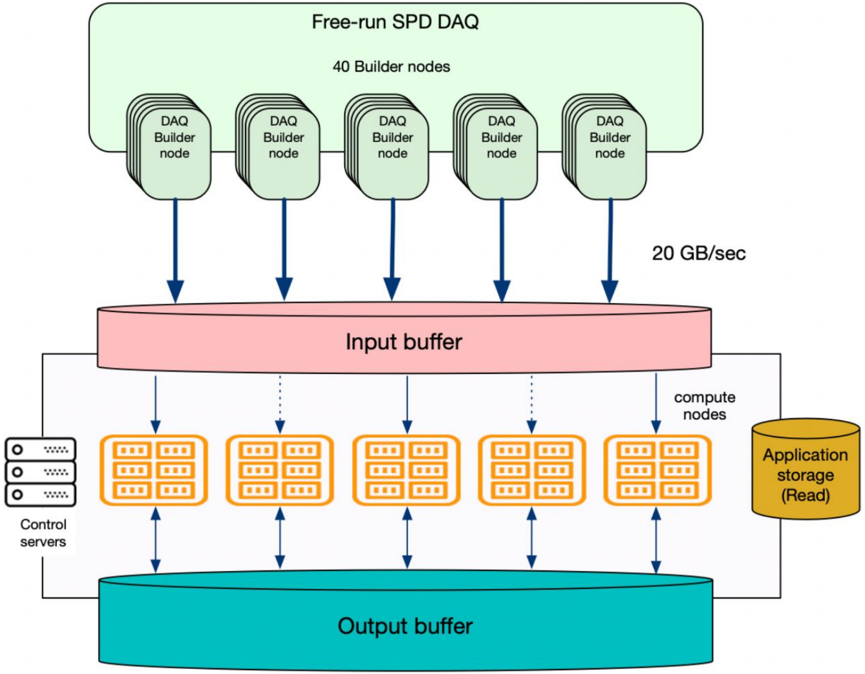
SPD Online Filter

Online filter is the first stage in data processing chain for SPD Experiment (right after DAQ)
Main goals:

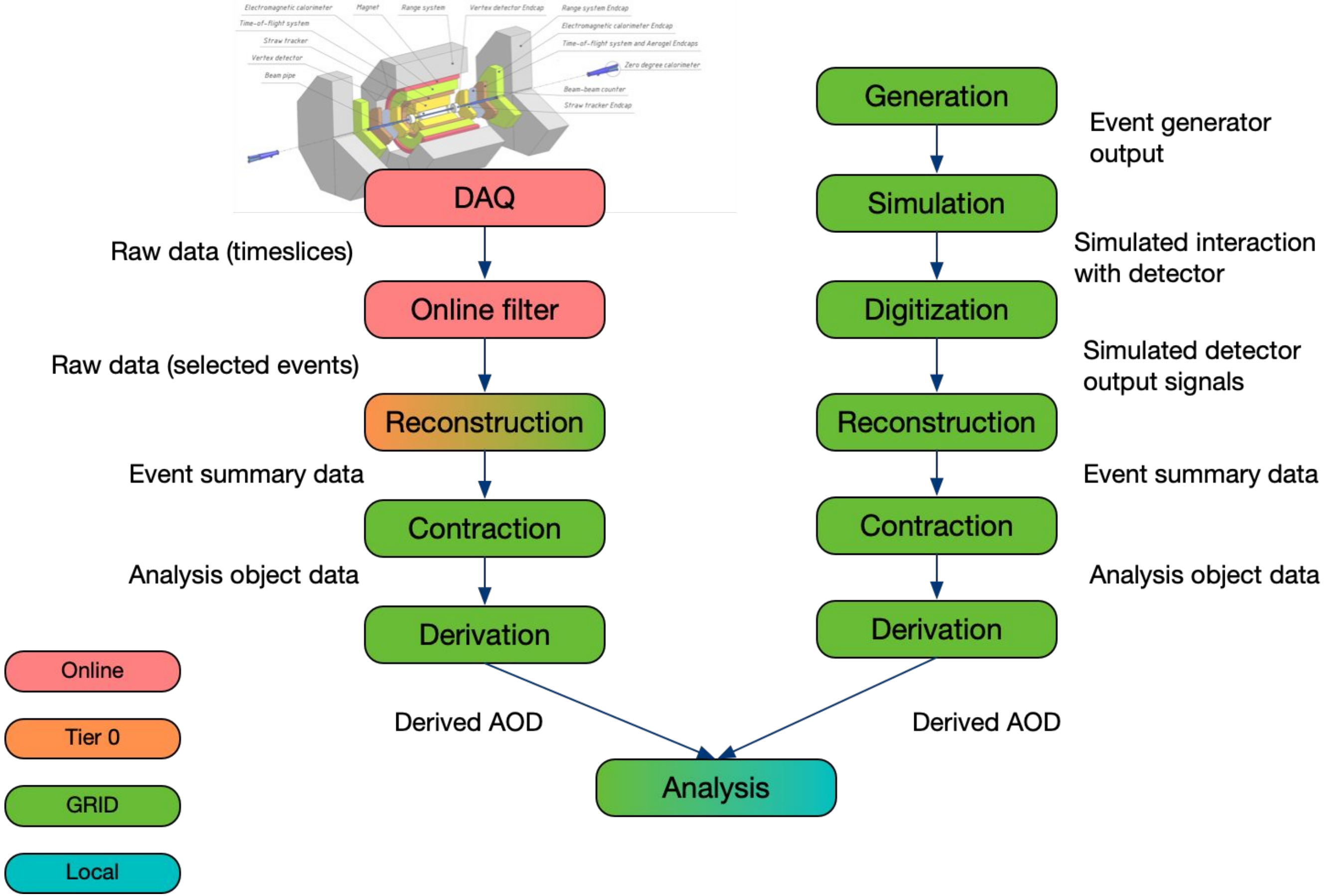
- Events unscrambling through partial reconstruction.
- Software trigger, which essentially is event filter.

SPD Online Filter is a high performance computing system for high throughput processing.

- 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.)
- Applied software: performs informational processing of data. Had to use same framework as 'offline' applied software.
- Middleware component: software complex for management of multistep data processing and efficient loading (usage) of computing facility.



SPD Distributed data processing



DAQ

Raw data (timeslices)

Online filter

Raw data (selected events)

Reconstruction

Event summary data

Contraction

Analysis object data

Derivation

Derived AOD

Generation

Event generator output

Simulation

Simulated interaction with detector

Digitization

Simulated detector output signals

Reconstruction

Event summary data

Contraction

Analysis object data

Derivation

Derived AOD

Analysis

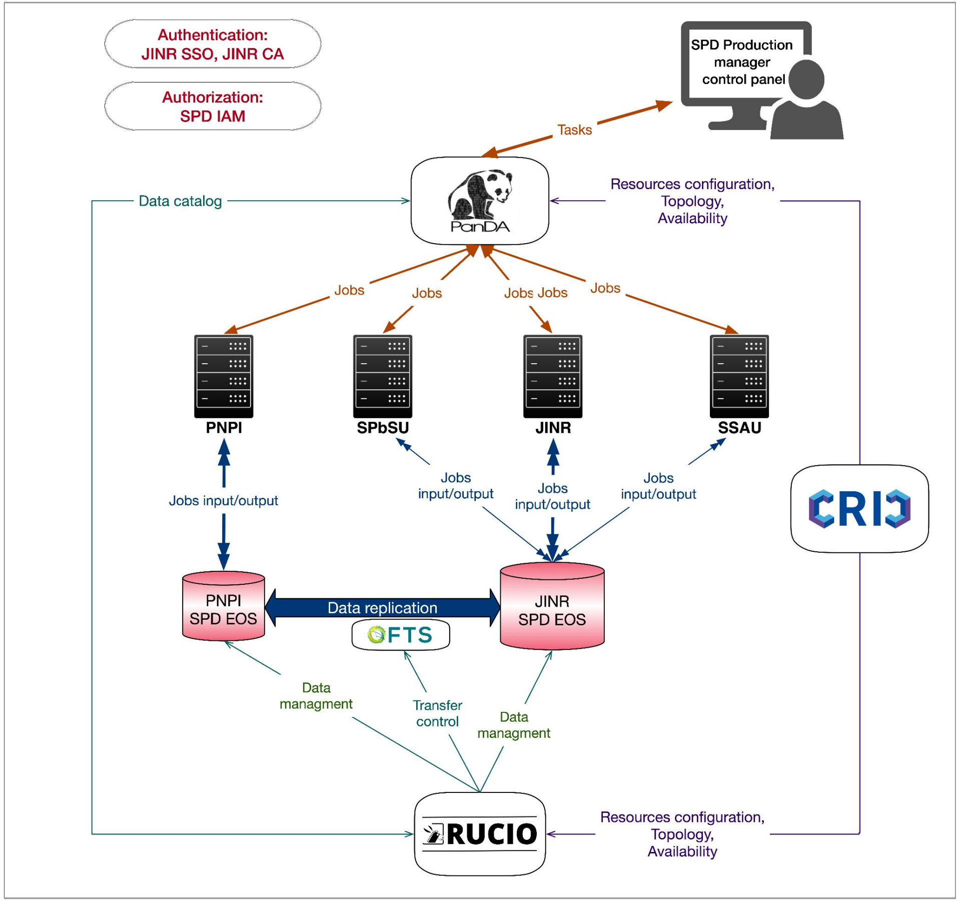
Online

Tier 0

GRID

Local

SPD distributed data processing system

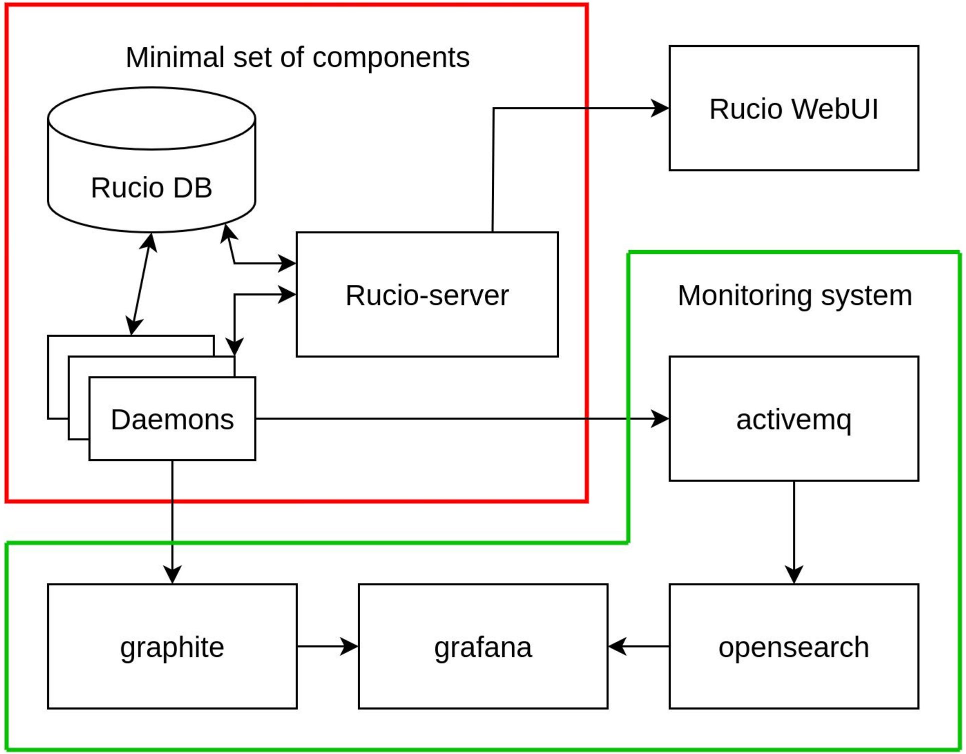


Offline data processing system includes the following components:

- Authentication system: JINR SSO
- Authorization system: IAM
- Information system: CRIC
- Software distribution service: CVMFS
- Data management system: **Rucio**
- Data transfer service: **FTS**
- Workflow management system: SPD own solution & PanDA
- Workload management system: PanDA

SPD Rucio deployment

At the moment, the required set of system components of three Rucio-servers are deployed in Docker containers based on JINR cloud computing infrastructure

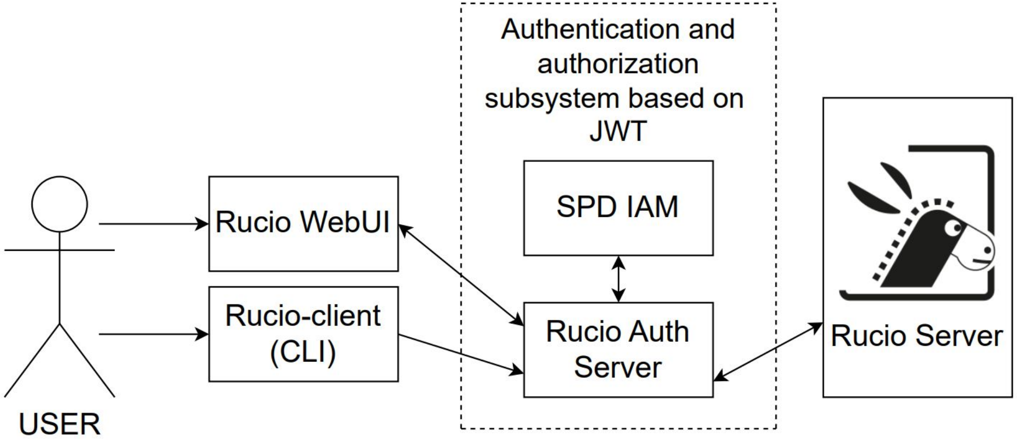


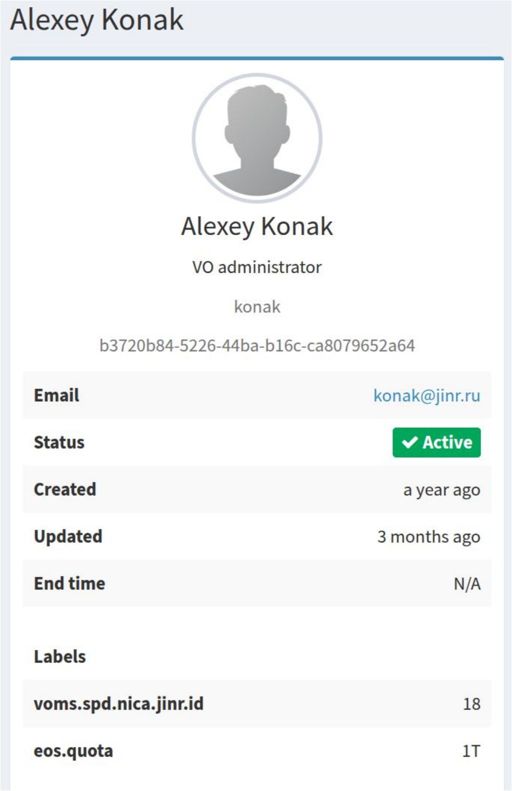
- **Production Infrastructure** with main Rucio-server which works stable and used for the needs of the SPD collaboration.
- **Development environment.** This infrastructure is used for development, testing and debugging.
- **Integration environment.** All updates and innovations are checked and tested on this installation before being put on the Prod infrastructure.

- PostgreSQL as the RDBMS backend

IAM integration

- IAM is a single source of info about users and their rights in the distributed computing environment of the SPD experiment.
- IAM provides authorization to all services and systems (including Rucio) with an access token and an ID token obtained during authorization in SPD IAM.
- Information about users imports from IAM to Rucio.
- User quota is defined in the IAM it is taken into account in Rucio and in EOS.




Alexey Konak

Alexey Konak

VO administrator

konak

b3720b84-5226-44ba-b16c-ca8079652a64

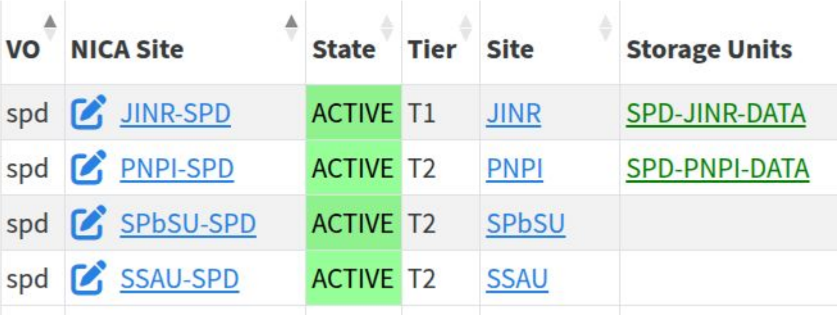
Email	konak@jinr.ru
Status	✓ Active
Created	a year ago
Updated	3 months ago
End time	N/A

Labels

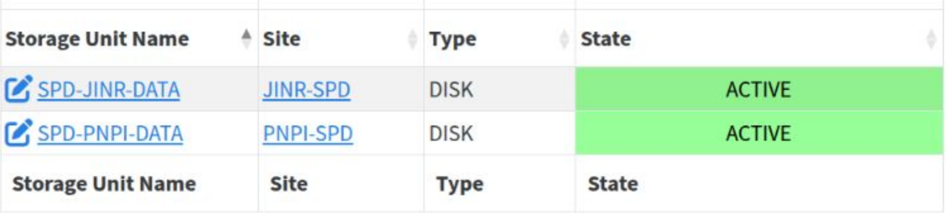
voms.spd.nica.jinr.id	18
eos.quota	1T

CRIC integration

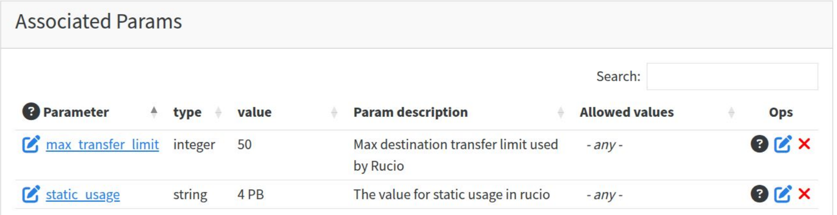
- Importing configuration information about storage systems from CRIC to Rucio.
- Management of storage system configuration from a single location.
- Creation of new RSE is possible with CRIC Controller.



VO	NICA Site	State	Tier	Site	Storage Units
spd	JINR-SPD	ACTIVE	T1	JINR	SPD-JINR-DATA
spd	PNPI-SPD	ACTIVE	T2	PNPI	SPD-PNPI-DATA
spd	SPbSU-SPD	ACTIVE	T2	SPbSU	
spd	SSAU-SPD	ACTIVE	T2	SSAU	



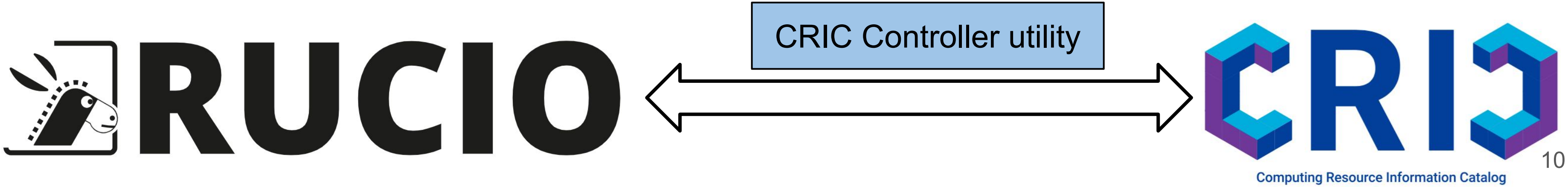
Storage Unit Name	Site	Type	State
SPD-JINR-DATA	JINR-SPD	DISK	ACTIVE
SPD-PNPI-DATA	PNPI-SPD	DISK	ACTIVE
Storage Unit Name	Site	Type	State



Associated Params

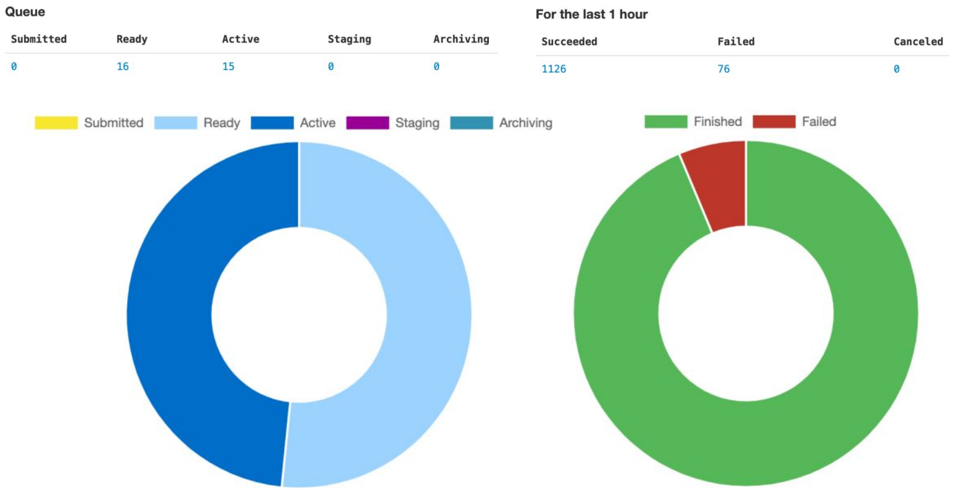
Search:

Parameter	type	value	Param description	Allowed values	Ops
max_transfer_limit	integer	50	Max destination transfer limit used by Rucio	- any -	
static_usage	string	4 PB	The value for static usage in rucio	- any -	



FTS3 integration

- One instance of data transfer service.
- Transfers and delegation with x509_proxy credentials
  (oidc not implemented yet).
- Custom docker image with conveyor-daemons.
- No limitation of number of transfers (in progress).
- Keeping two replicas of productions on two sites.

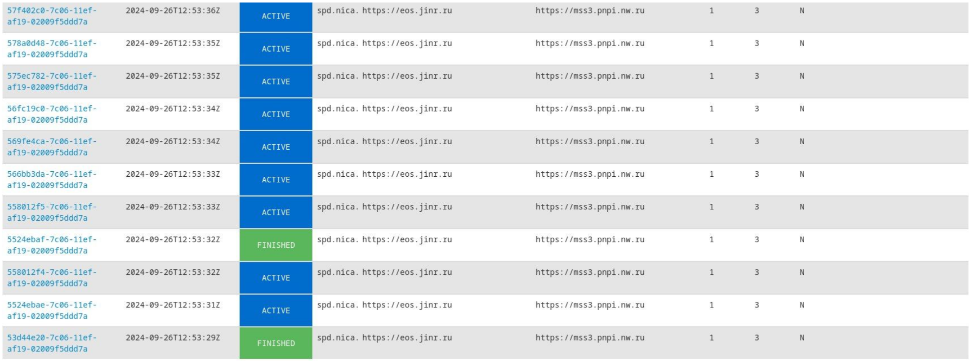


Queue

Submitted	Ready	Active	Staging	Archiving
0	16	15	0	0

For the last 1 hour

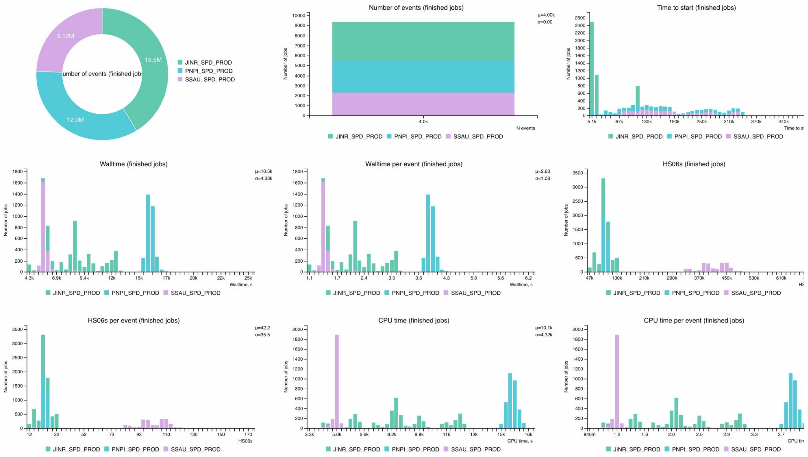
Succeeded	Failed	Canceled
1126	76	0

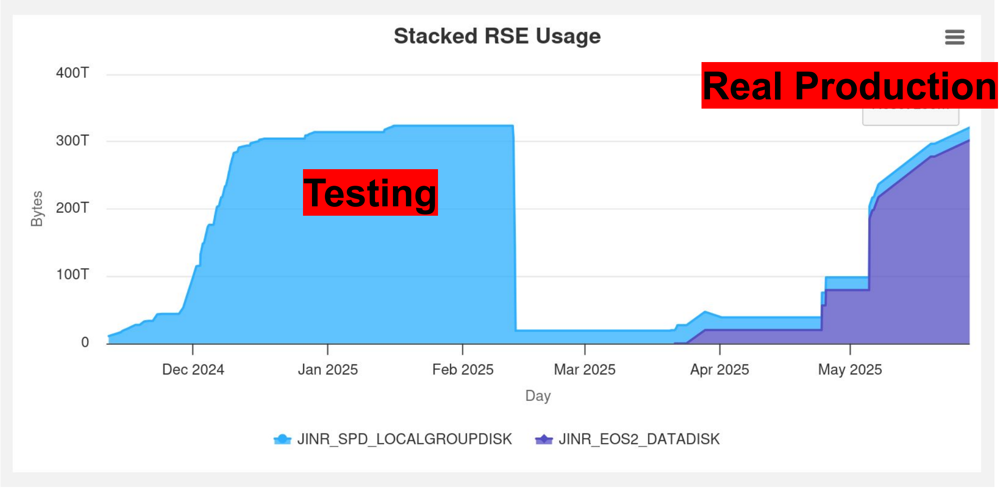


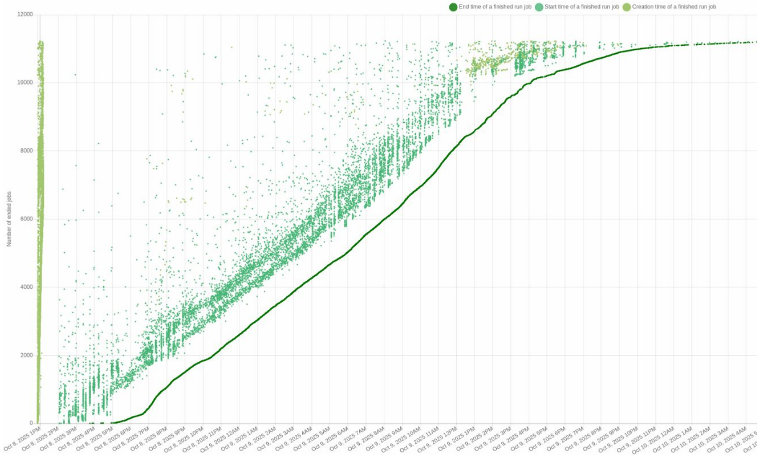
57f402c0-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:36Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
578a0d48-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:35Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
575ec782-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:35Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
56fc19c0-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:34Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
569fe4ca-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:34Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
566bb3da-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:33Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
558012f5-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:33Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
5524ebaf-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:32Z	FINISHED	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
558012f4-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:32Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
5524ebae-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:31Z	ACTIVE	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N
53d44e20-7c06-11ef-af19-02009f5ddd7a	2024-09-26T12:53:29Z	FINISHED	spd.nica.	https://eos.jinr.ru	https://mss3.pnpi.nw.ru	1	3	N

PanDA integration

- PanDA and Rucio is used for mass production of SPD.
- Interaction of PanDA and Rucio was tested in various forms as well as different data organisation.
- Rucio informs PanDA about storage space left.



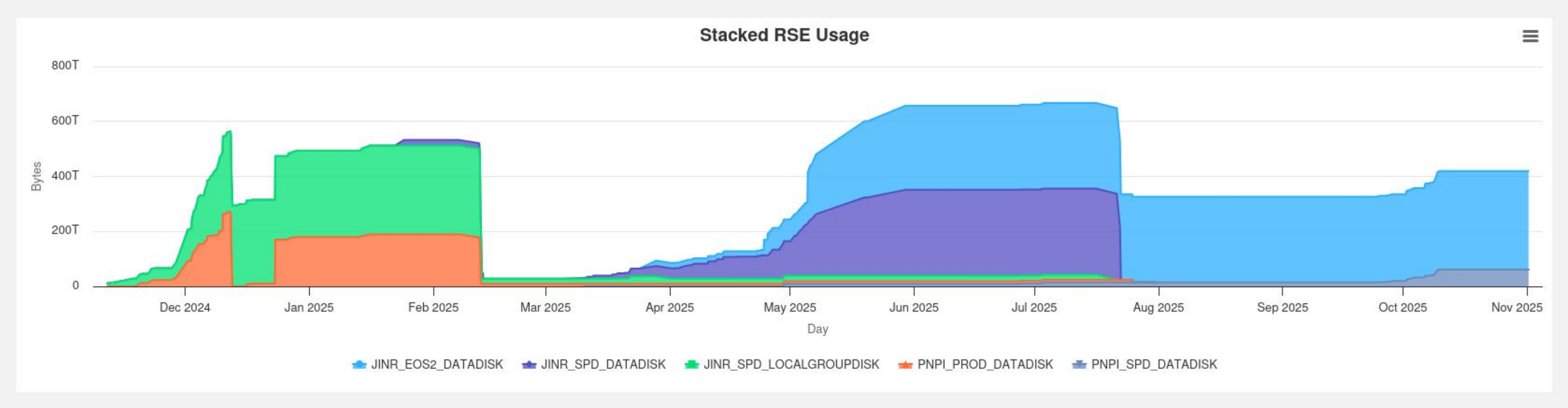




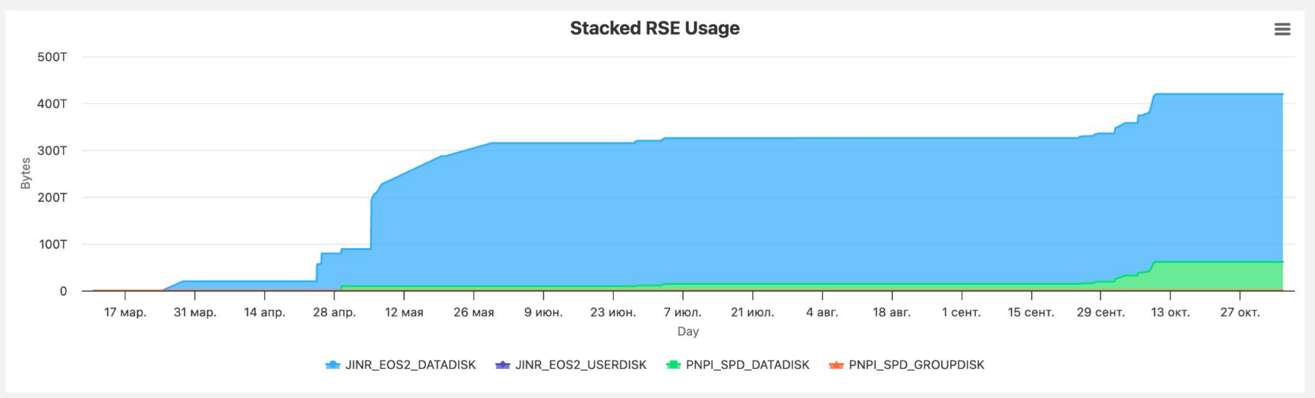
Data Challenge

We have a few RSE in JINR and PNPI:

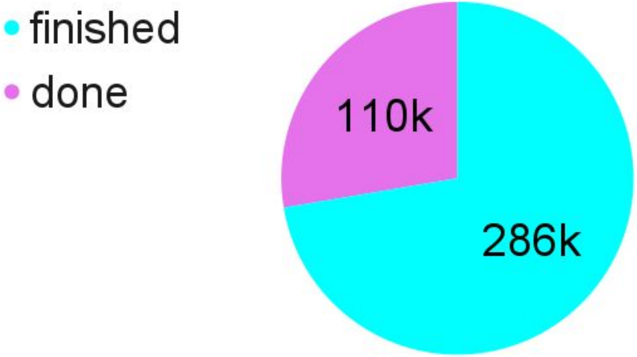
- SPD_JINR_LOCALGROUPDISK and PNPI_PROD_DATADISK were used for testing interactions between Rucio-PanDA-FTS3 on scale.
- JINR_SPD_DATADISK – SPD production data storage on shared JINR EOS facility ;
- JINR_EOS2_DATADISK – SPD production data on the SPD own storage facility at JINR;
- PNPI_SPD_DATADISK – SPD production data on the SPD storage at PNPI.



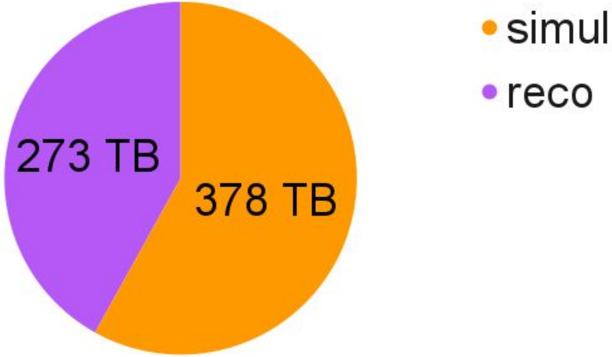
Productions results



Successfully processed jobs



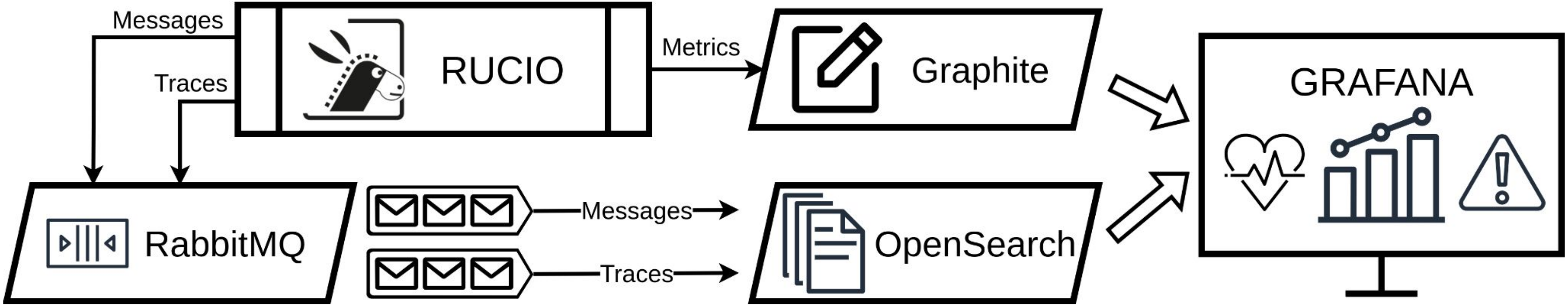
Total output datasets volume



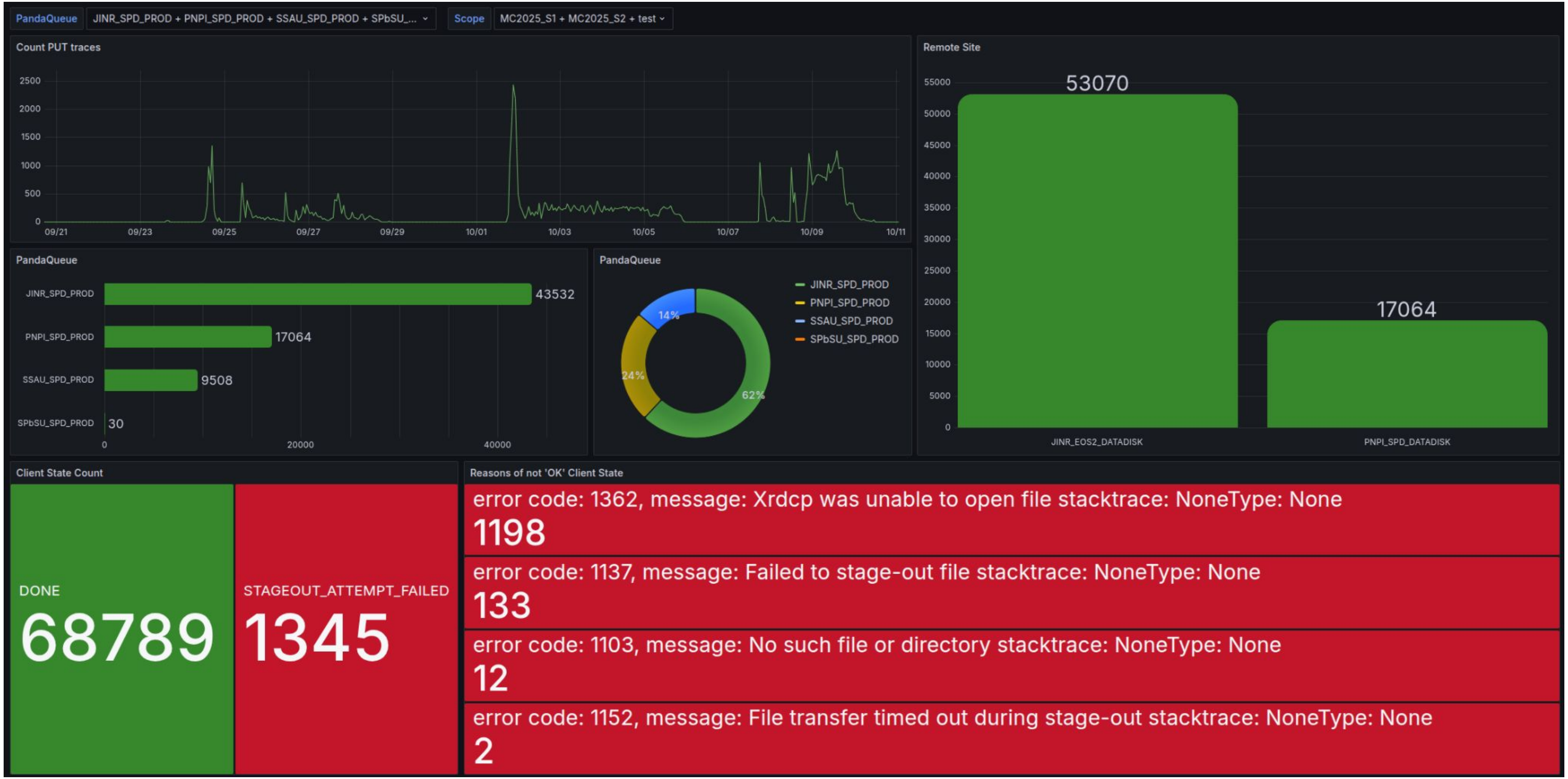
Monitoring System

Components of Monitoring System:

- **Graphite**: collects metrics from Rucio server.
- **RabbitMQ**: queues with messages and traces.
- **OpenSearch**: documents with messages and traces (long term storage).
- **Grafana**: visualizations.



Monitoring System [1/2]



PandaQueue: JINR_SPD_PROD + PNPI_SPD_PROD + SSAU_SPD_PROD + SPbSU_...

Scope: MC2025_S1 + MC2025_S2 + test

Client State Count

DONE 68789

STAGEOUT_ATTEMPT_FAILED 1345

Reasons of not 'OK' Client State

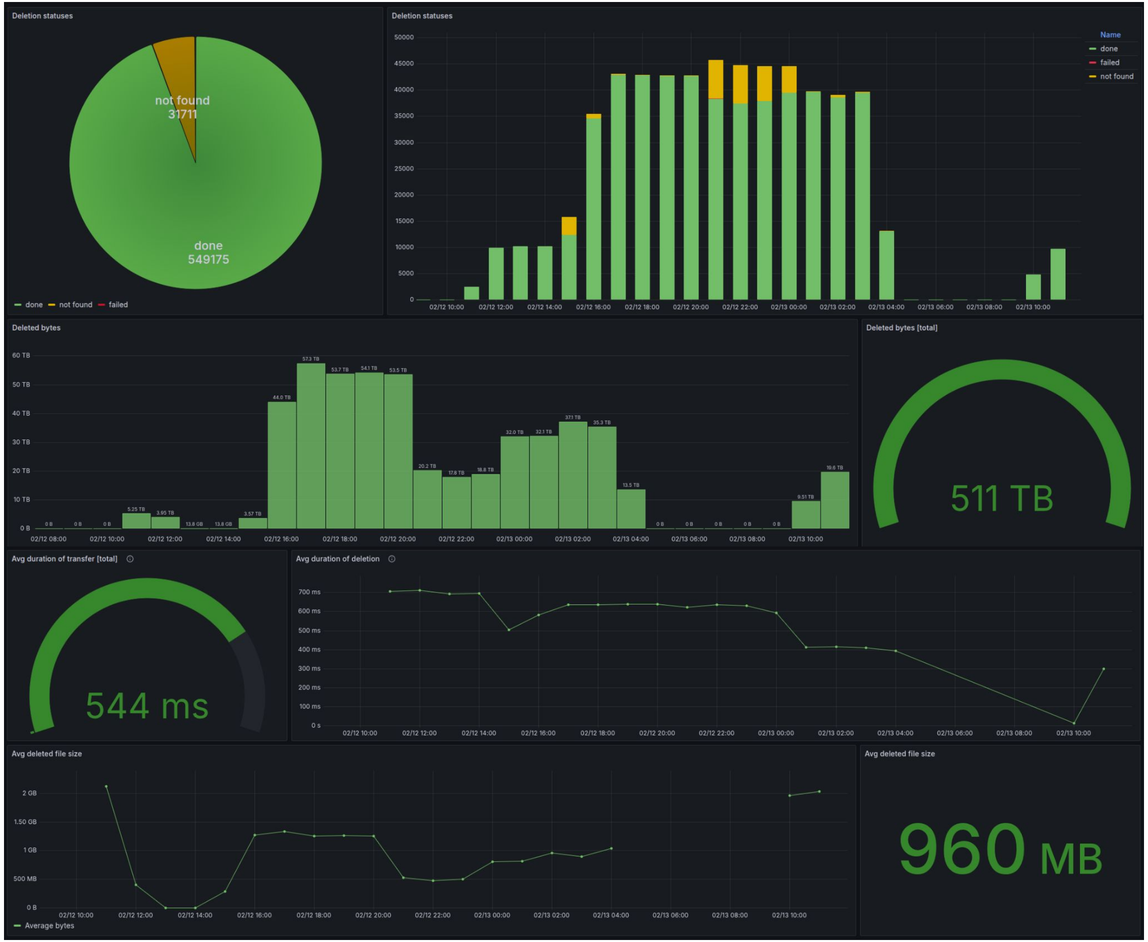
error code: 1362, message: Xrdcp was unable to open file stacktrace: NoneType: None
1198

error code: 1137, message: Failed to stage-out file stacktrace: NoneType: None
133

error code: 1103, message: No such file or directory stacktrace: NoneType: None
12

error code: 1152, message: File transfer timed out during stage-out stacktrace: NoneType: None
2

Monitoring System [2/2]



Summary

- Rucio allows the SPD experiment to solve data management tasks.
- Easy installation of the demo version. Problems arise during the deployment and configuration of the production environment.
    - Installing and configuring the system requires reasonable time and some skill, but it is generally a technical task.
    - There is a lack of documentation on configuring daemons and deploying monitoring.
    - Some solutions involve using the CERN infrastructure which is also not reflected in the documentation, and therefore we have to add our own implementations.

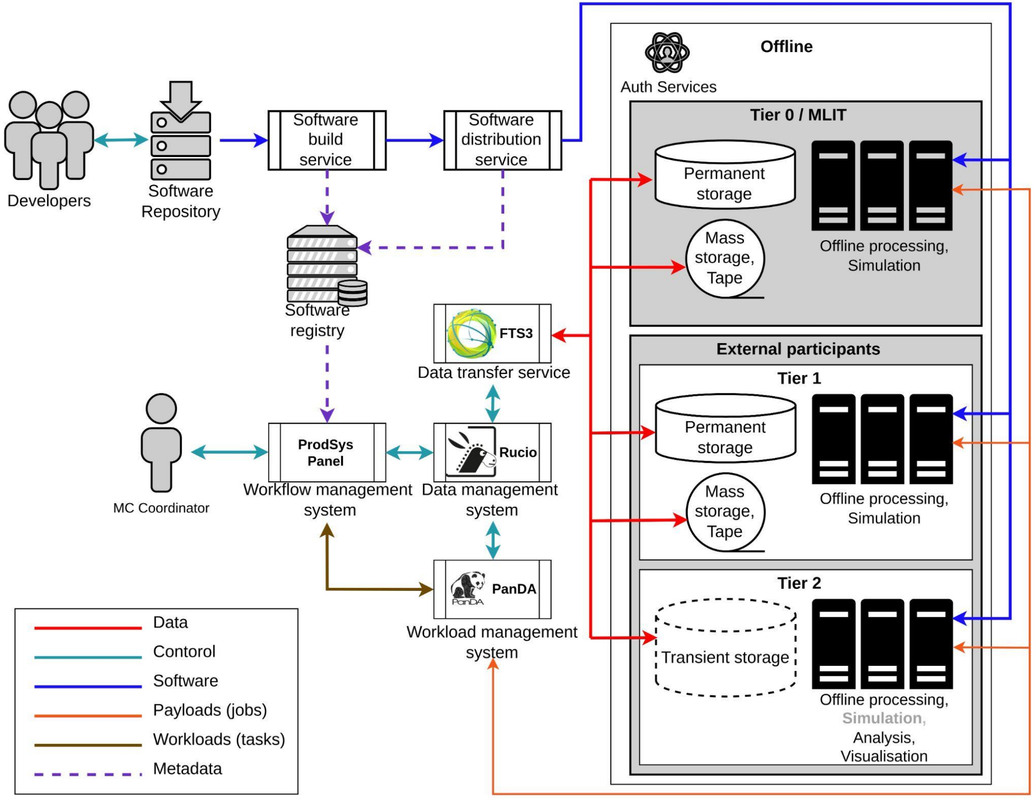
Future plans

- The development of monitoring – already begun.
- Consistency-check – need to integrate standart Rucio solution and developing own tools.
- User policy – dividing users into groups and reviewing the allowed actions for these groups.
- User support – development documentation for users.
- Start using TAPE storages.
- SPD Production system in progress with fulfilling of metadata in Rucio.

Thank you for your attention!

Backup slides

SPD distributed data processing system



Offline data processing system includes the following components:

- workload management system (WMS) – PanDA,
- workflow management system (WFMS) – ProdSys Panel,
- data management system (DMS) – Rucio,
- data transfer service (DTS) – File Transfer Service 3 (FTS3),
- information system (IS) – Computing Resource Information Catalog (CRIC)

SPD Identity and Access Management (SPD IAM) provides authentication to all services and systems.

MC Production data organisation

PanDA	Rucio	Status
Campaings MC_2025_S1, MC_2025_S2	Scopes MC2025_S1, MC2025_S2	Ready
Requests PROD2025-020	Containers MC2025_S2:PROD2025-020	Being developed
Tasks PROD2025-020.RECO.2	Datasets MC2025_S2:PROD2025-020:minbias-P8-spdroot4174-dev.27GeV-UU.PROD2025-020.RECO.2	Ready
Jobs PanDA ID	Files MC2025_S2:r.MC2025_S2.minbias-P8-spdroot4173-dev.27GeV-UU.PROD2025-018.RECO.1.001225.root.1	Ready

Rucio Account Importer [1/3]

Rucio Account Importer is designed to import accounts from SPD IAM to Rucio and also to adding identities to rucio accounts.

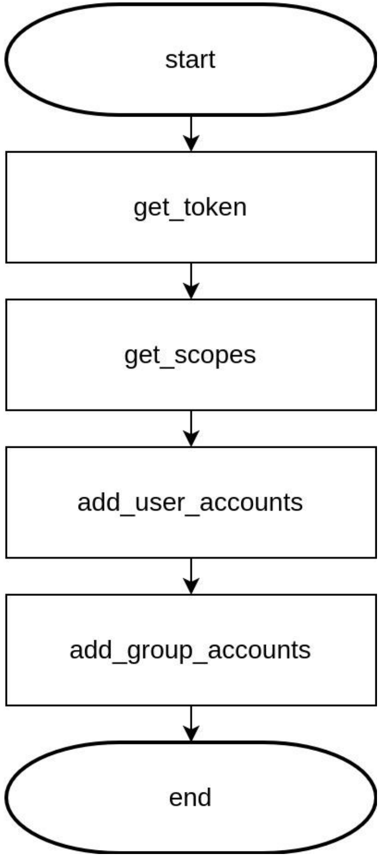
Implemented the addition of the OIDC identities for authentication in Rucio via SPD IAM and addition of the user certificates subject DN for authentication in rucio with usercert and proxy certificates. There is a functionality for adding a subject DN string in legacy format. Also, for each user will be added their standard user scope and a global limit.

Rucio Account Importer [2/3]

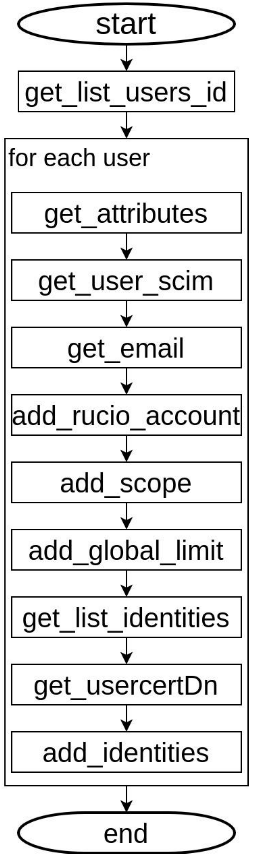
The developed utility interacts with SPD IAM using an access, issued to the client who performs actions to import accounts and identification information. This client is registered in SPD IAM with the scope iam.admin:read, scim:read.

The access token contains only these two scopes, which allow the client (in this case, the developed utility) to obtain information about users, their identification information, groups, etc. from the SPD IAM. The lifetime of the token is five minutes. During this time, all operations are performed to obtain information about SPD IAM users and add new information to Rucio. The token grant flow is client_credentials (client_id + client_secret).

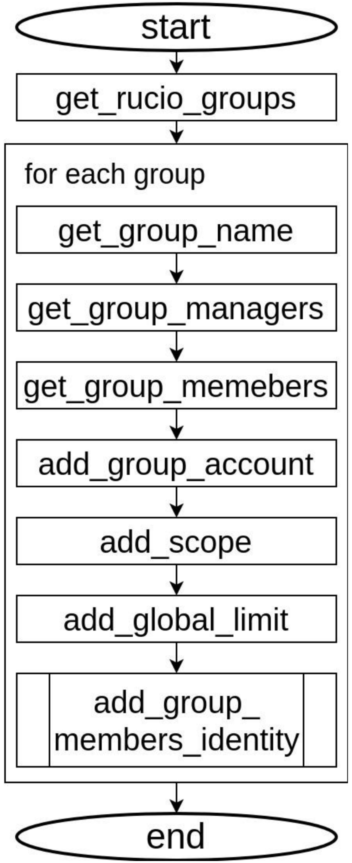
Rucio Account Importer [3/3]



(1) General algorithm

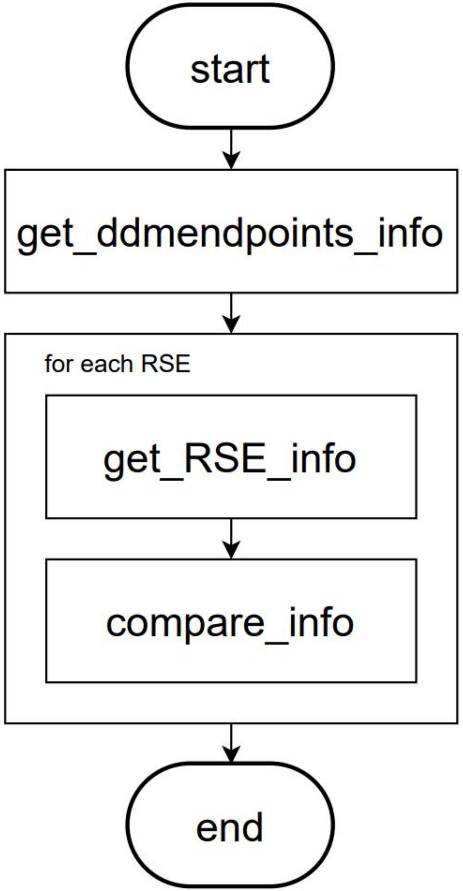


(2) add_user_accounts algorithm



(3) add_group_accounts algorithm

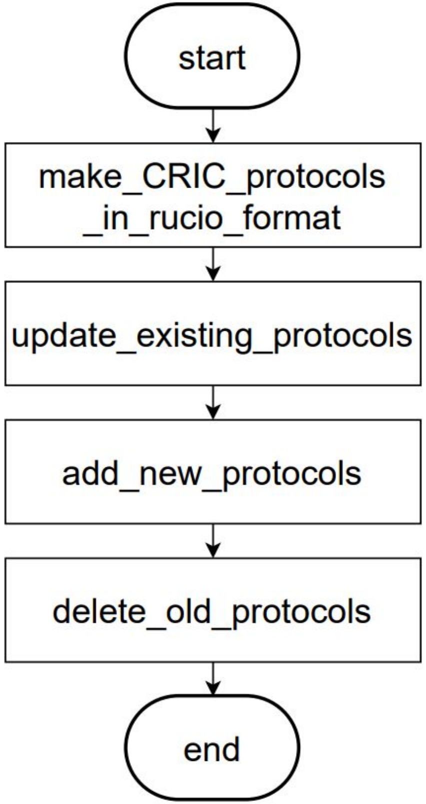
CRIC integration [1/2]



1) Takes information from CRIC about all storage systems registered in it.
2) Requests protocols and attributes of RSE from Rucio.
3) Compares info (changes it if necessary):

- checks attributes;
- checks FTS;
- checks protocols.

CRIC integration [2/2]


start
make_CRIC_protocols _in_rucio_format
update_existing_protocols
add_new_protocols
delete_old_protocols
end

Rucio protocol description

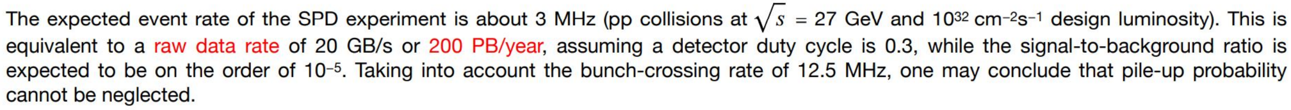
```
{
 "domains": {
   "lan": {
     "delete": 0,
     "read": 0,
     "write": 0
   },
   "wan": {
     "delete": 0,
     "read": 0,
     "third_party_copy_read": 1,
     "third_party_copy_write": 1,
     "write": 0
   }
 },
 "extended_attributes": null,
 "hostname": "somehostname.jinr.ru",
 "impl": "rucio.rse.protocols.webdav.Default",
 "port": 8000,
 "prefix": "/eos/rucio/spd",
 "scheme": "https"
}
```

Hostname, **port** and **scheme** are key attributes of protocol in Rucio.

If any other attribute has been changed -> update_existing_protocol

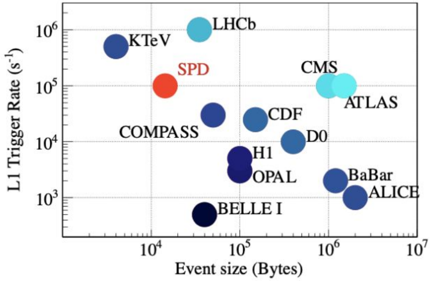
If any key attribute has been changed -> add_new_protocol and delete_old_protocol

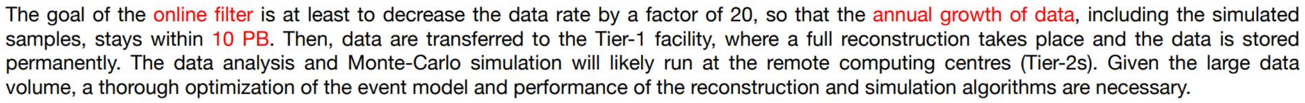
SPD as data source



The expected event rate of the SPD experiment is about 3 MHz (pp collisions at $\sqrt{s}$ = 27 GeV and 10^{32} cm^{-2}s^{-1} design luminosity). This is equivalent to a raw data rate of 20 GB/s or 200 PB/year, assuming a detector duty cycle is 0.3, while the signal-to-background ratio is expected to be on the order of 10^{-5}. Taking into account the bunch-crossing rate of 12.5 MHz, one may conclude that pile-up probability cannot be neglected.

- SPD TDR





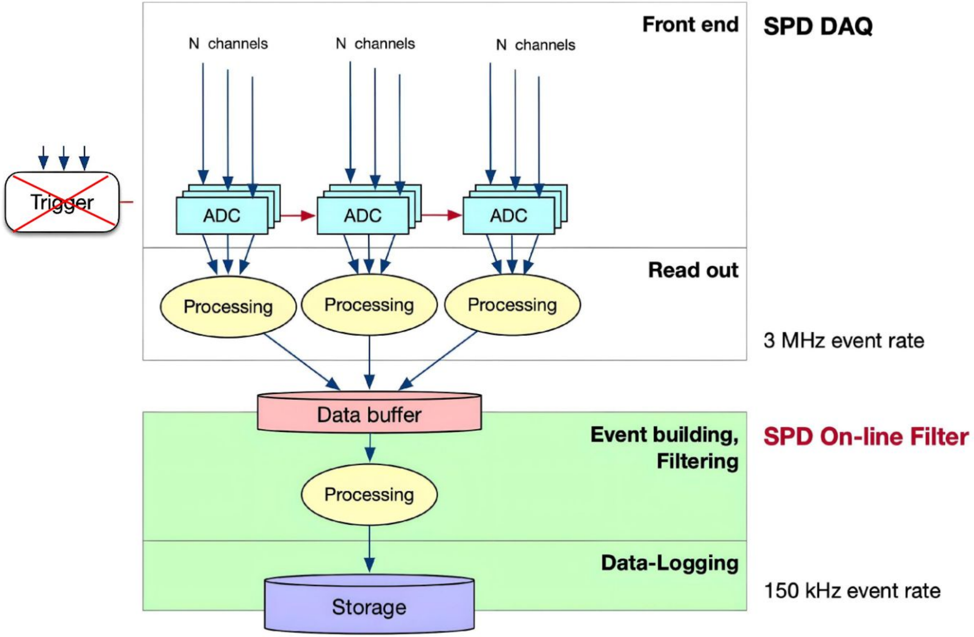
The goal of the online filter is at least to decrease the data rate by a factor of 20, so that the annual growth of data, including the simulated samples, stays within 10 PB. Then, data are transferred to the Tier-1 facility, where a full reconstruction takes place and the data is stored permanently. The data analysis and Monte-Carlo simulation will likely run at the remote computing centres (Tier-2s). Given the large data volume, a thorough optimization of the event model and performance of the reconstruction and simulation algorithms are necessary.

- Data from the detector – 20 GB/s (or 200 PB/year "raw" data, ~3*10^13 events/year)
- Simulation results – ??? (the exact volume is unknown, but there will be no less of them than the data from the detector.)
- Data of various intermediate formats along the way from “raw” to ready for analysis by physical groups – ??? (there will be a lot of them...)

SPD 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.

- DAQ provide data organized in time frames which placed in files with reasonable size (a few GB).
- Each of these file may be processed independently as a part of top-level workflow chain.
- No needs to exchange of any information during handling of each initial file, but results of may be used as input for next step of processing.



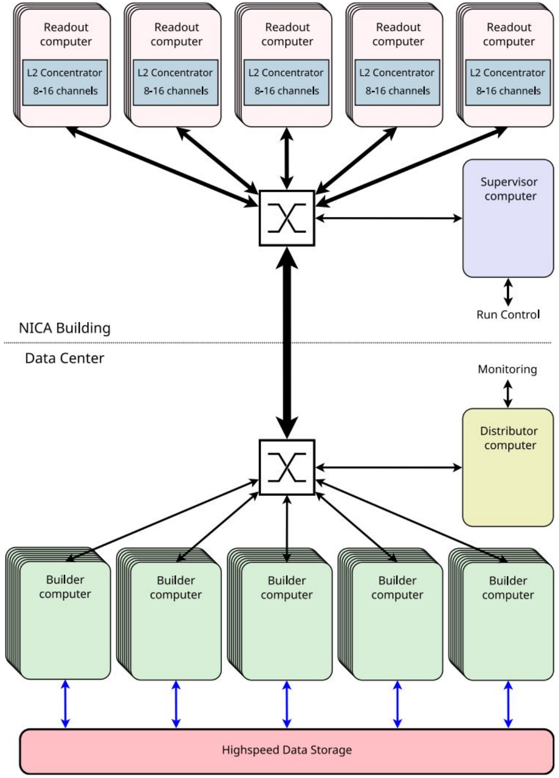
Computer Read-out and Slice Building system

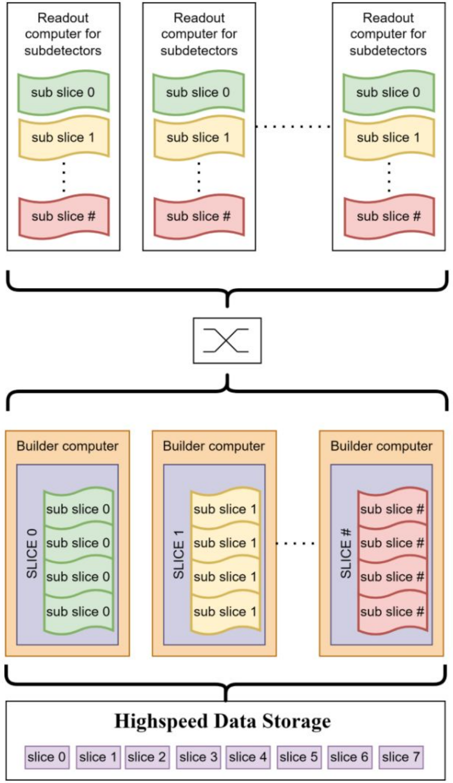
The purpose of slice building systems is to collect data (sub-slices) from all readout computers, build a complete slice and write it to the fast intermediate storage.

Additionally, the system will provide access to raw data (sub-slice) for detector groups in order to monitor the collected data before processing in the online filter.

Components:

- Read-out – getting a sub-slice from each detector system
- Builder – combining sub-slices and forming a single data structure
- Supervisor – arbitration of the slice construction system
- Distributor – provides raw data for monitoring





To be continued...