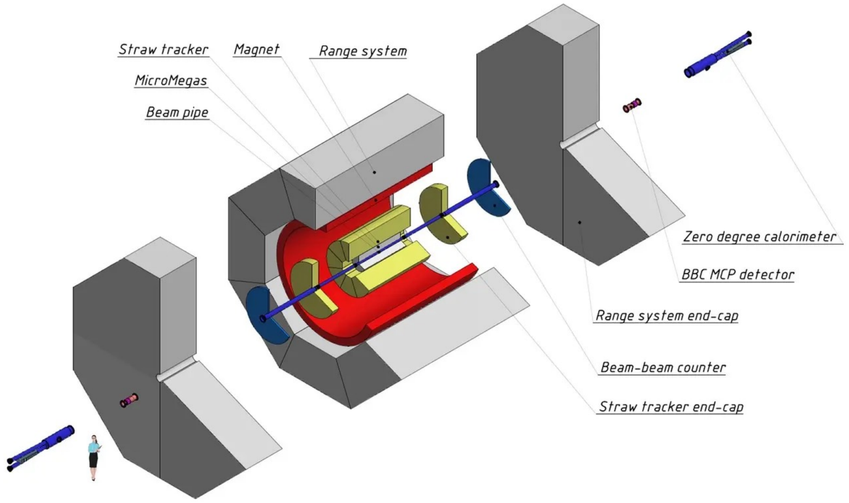
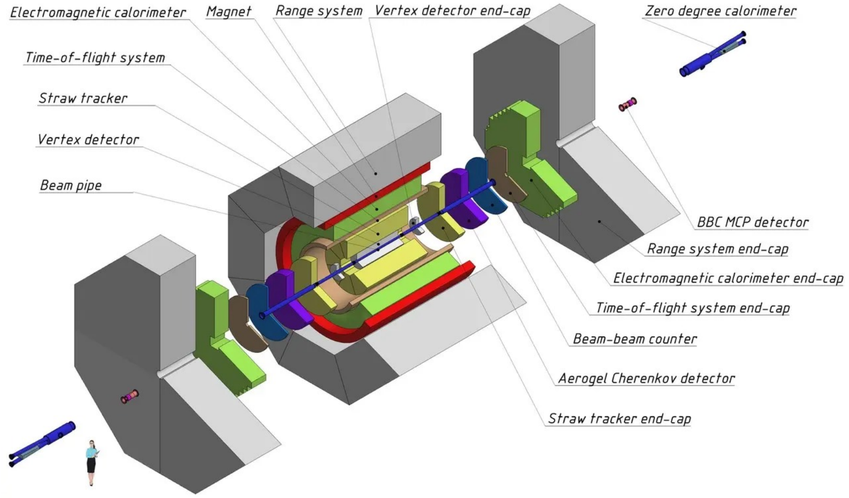
Production System of the SPD Experiment



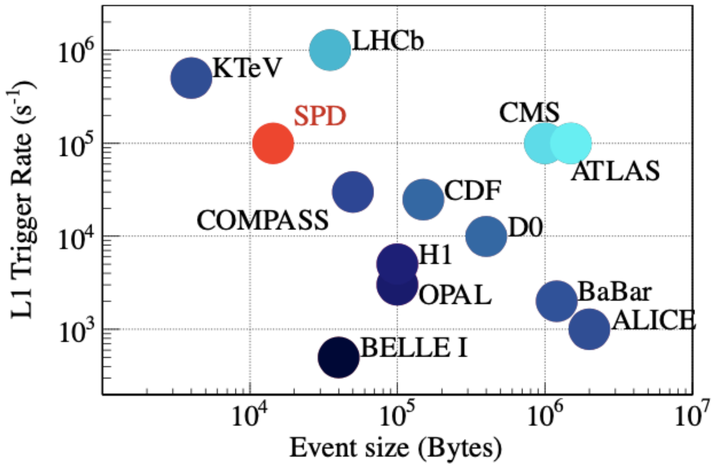


Artem Petrosyan, MLIT, JINR
MMCP 2024, Yerevan, Armenia
October 22, 2024

Introduction

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.

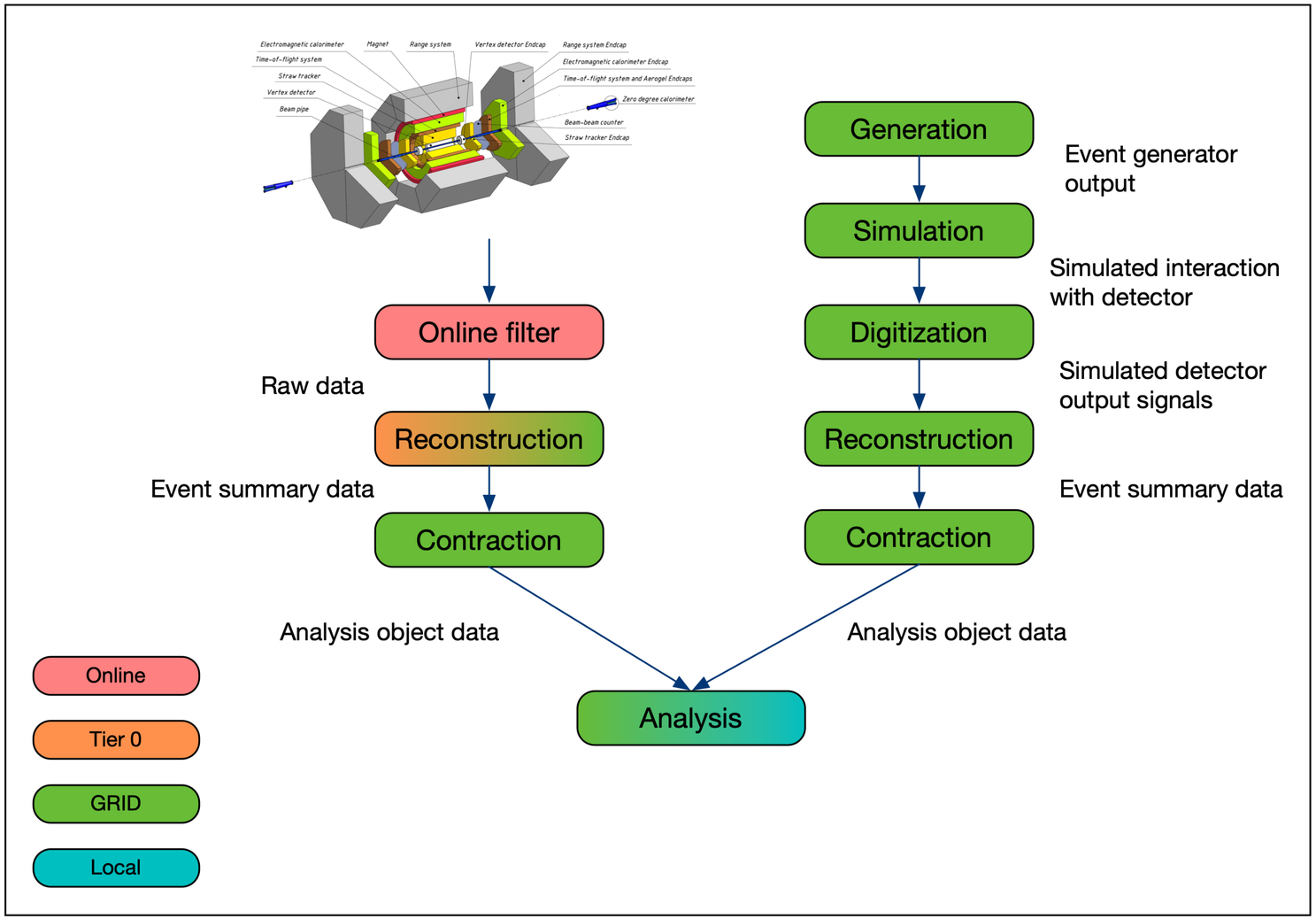
- SPD TDR



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.

Processing steps distribution over computing resource types

- Execution of events reconstruction and reprocessing jobs is accompanied by intensive I/O operations and will be done mostly on the dedicated farms on JINR site as Tier 0 component of the distributed computing system
- The use of Tier 0 is dictated by huge amount of initial data, gathered by the physics facility — data must be reduced as much as possible in order to be ready for distribution
- Less I/O intensive steps, especially Monte-Carlo production, can be performed on the remote computing centres
- User analysis can be run on every close to user resource

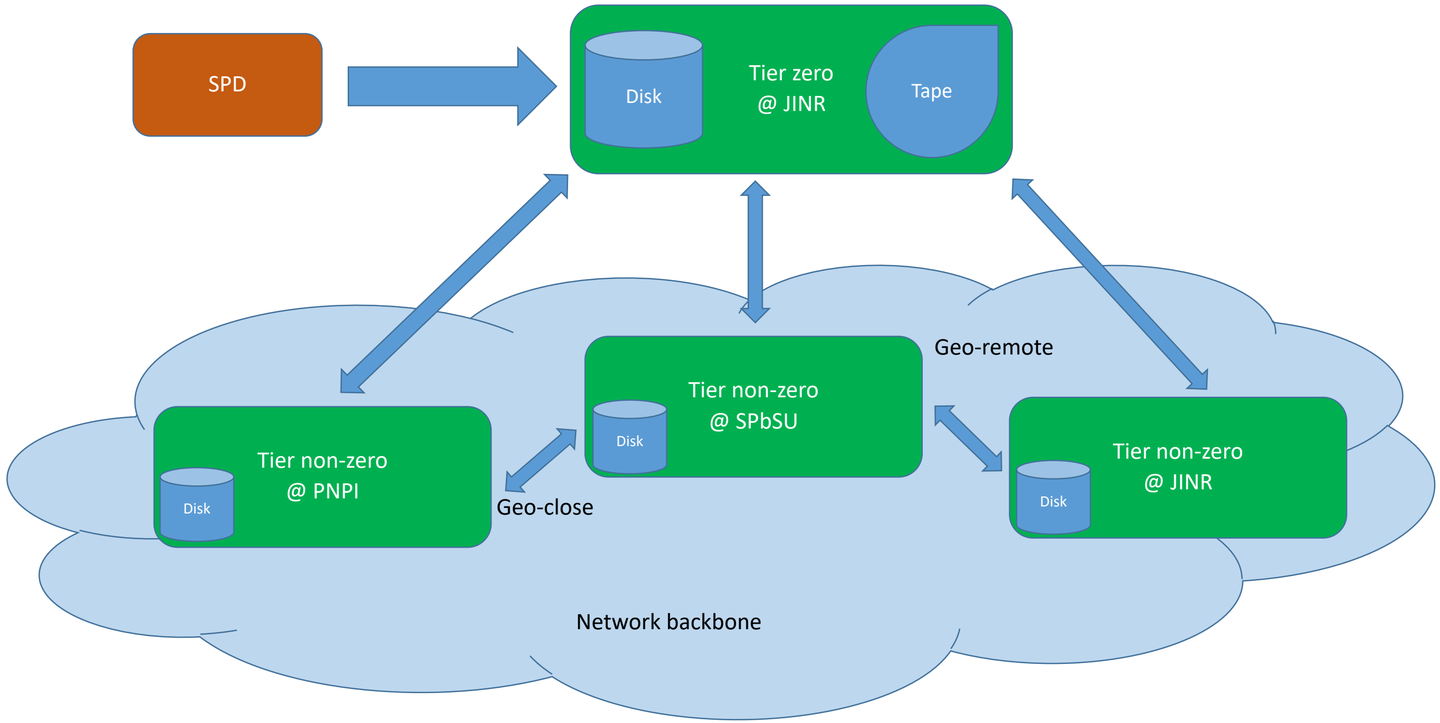


Why do we need a Production System?

- Processing requests for simulation and reconstruction algorithms tuning require significant computing resources, and the results of data processing occupy a fairly large disk space
  - For example, a standard request for the generation of 20 million events involves two calculation steps (simulation and reconstruction) and the generation of 10 terabytes of data and 15 thousand output files, not taking into account service files with processing logs
  - We plan to use all available computing resources, both located at the institute and outside it
- We plan to store several copies of the most important data: one at JINR and one somewhere else
- In order to quickly generate the required datasets, it is necessary to create a system that would automate the generation of jobs, distribute them across available computing resources and manage the output data
- In order to organize managing of these requests in the most efficient way, a dedicated system must be built
- Production system is a set of systems and services under the control of a high-level system, the so-called orchestrator, which manages the underlying services and allows you to organize highly automated mass processing of data and their movement in accordance with the computing and storage model of the experiment

SPD Computing Cloud

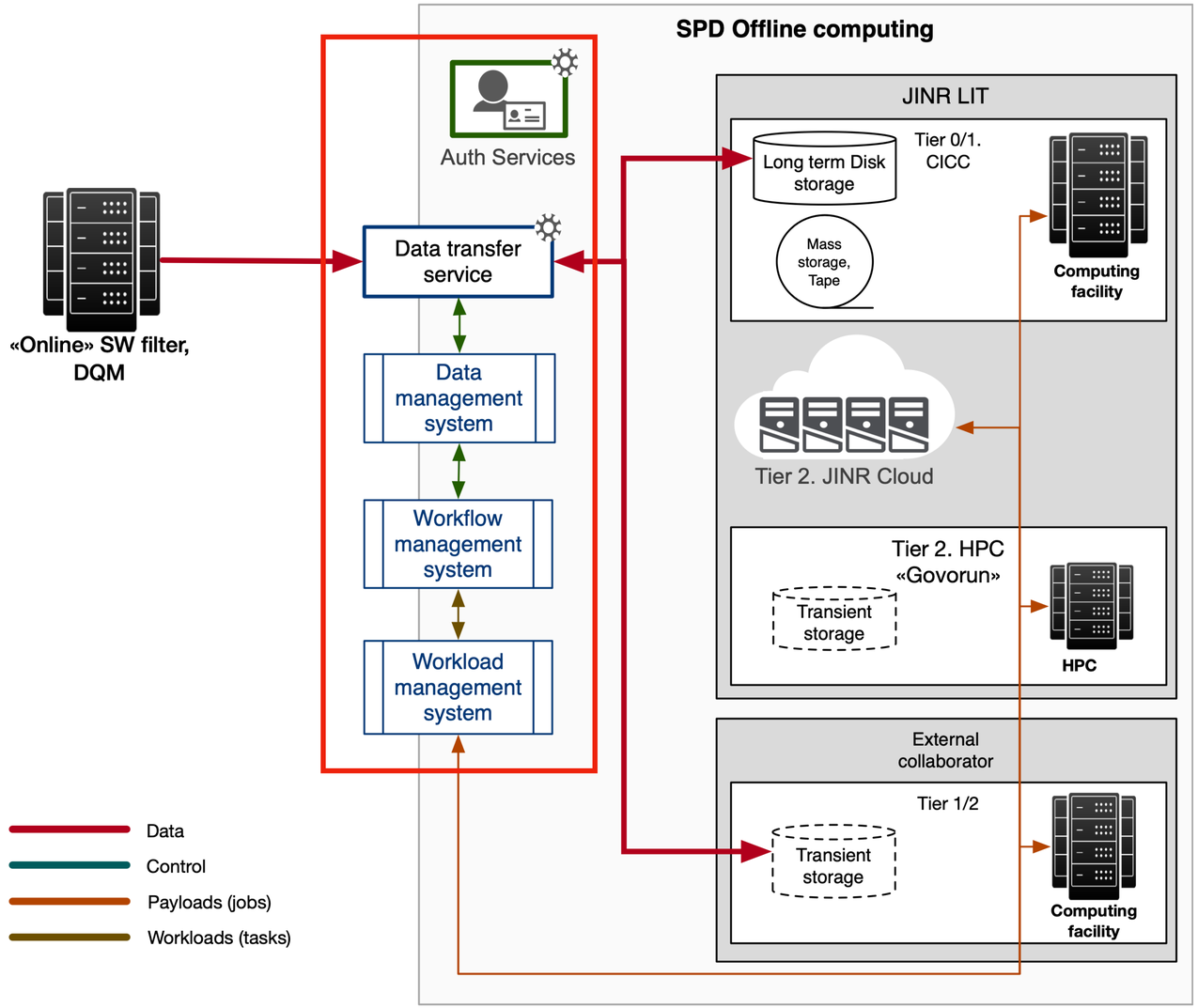
- Data volume mandates some baselines
  - >10 Gbps network per site (from TDR)
  - >500 TB storage capacity per site (not from TDR, but might be added to the next version)
- Try to use existing free software as much as possible
  - Experience comes from large LCG experiments
- Optimize management and operation effort
  - Do not deploy home-grown solutions that are different from site to site
  - Provide a reasonable guidelines for interfacing computing resources with central data management services



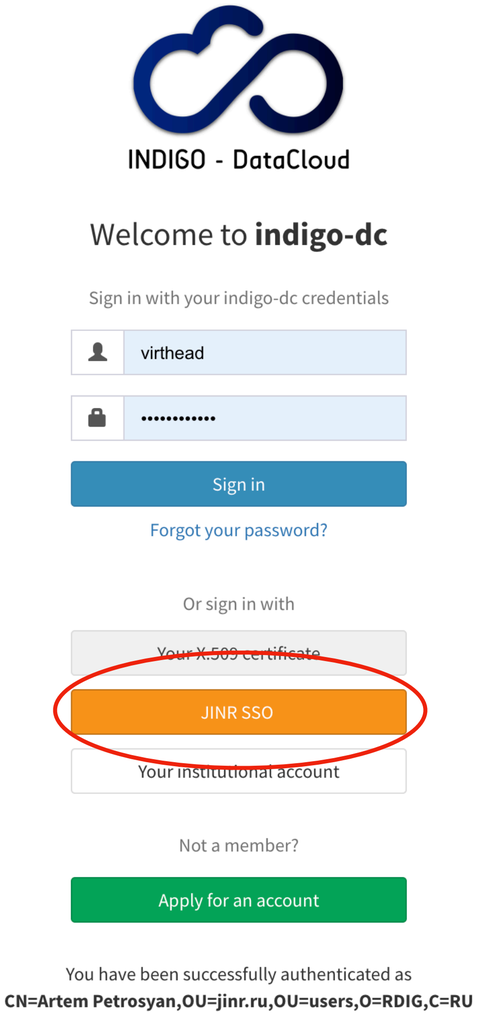
Components of the Production System

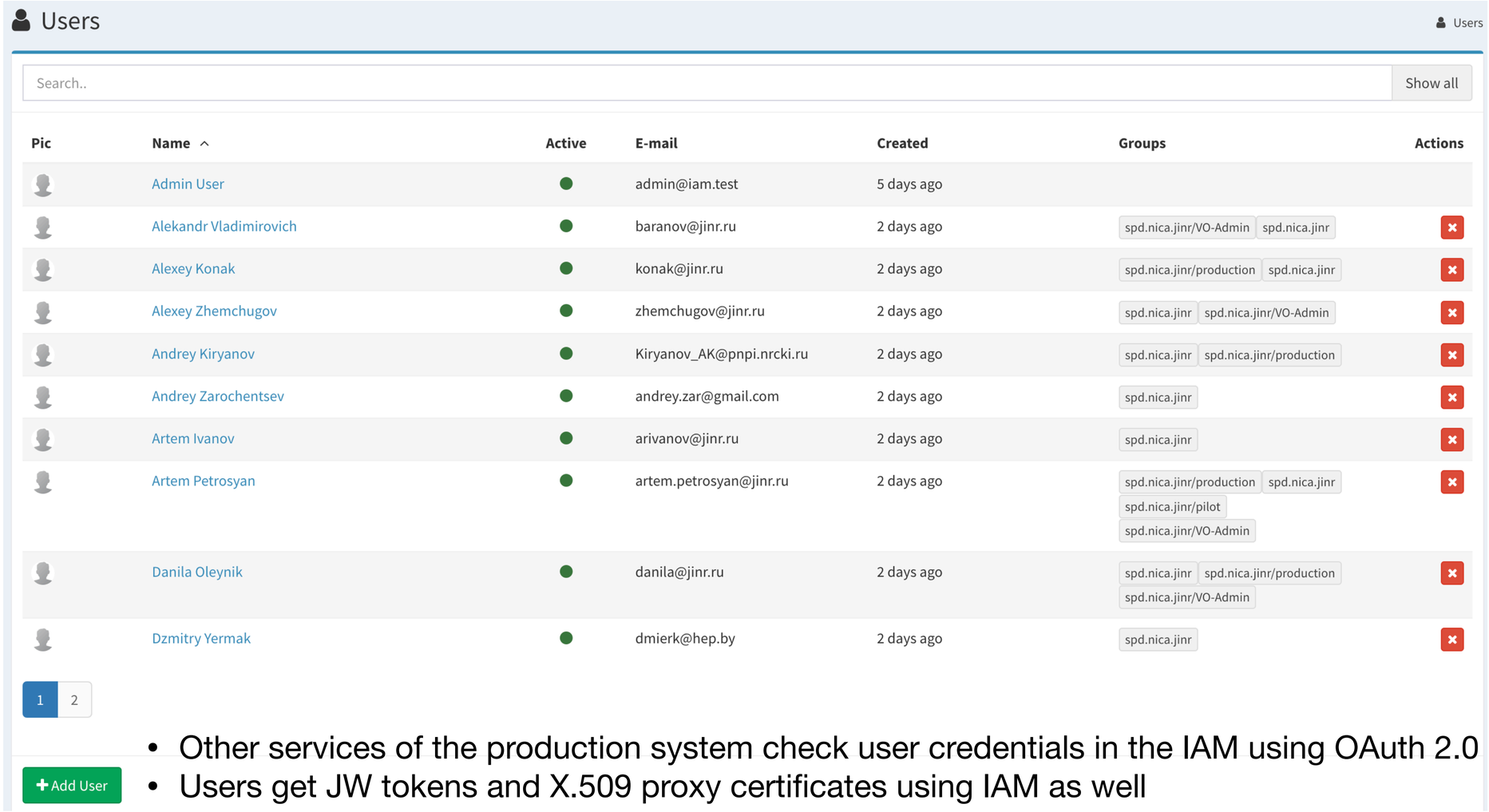
- IAM — an entry point to all members of the computing services of the collaboration: stores user profiles, their roles and rights to perform certain actions
- CRIC information system — the main integration component of the computing system: contains info about all computing and storage resources, access protocols, entry points, and many other things in one place and distributes this info via API to all other components mentioned below
- PanDA WFMS/WMS — manages data processing at the highest level of chains of tasks and datasets or periods and campaigns, finds the best computing resource for task to be executed on, manages individual jobs (usually 1 job means 1 input file) processing
- Rucio DMS — responsible for data management, including data catalog, data integrity and data lifetime management strategies
- FTS DTS — enables massive data transfers

- All services deployed in the JINR Cloud Service



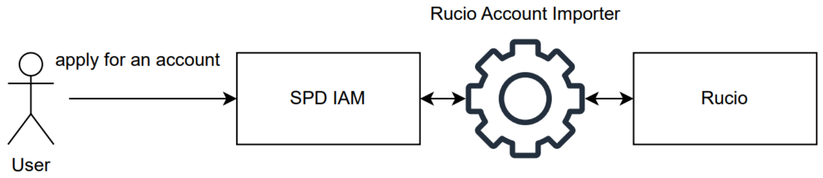
Identity and Access Management Service





- Other services of the production system check user credentials in the IAM using OAuth 2.0
- Users get JW tokens and X.509 proxy certificates using IAM as well

Data Management Service

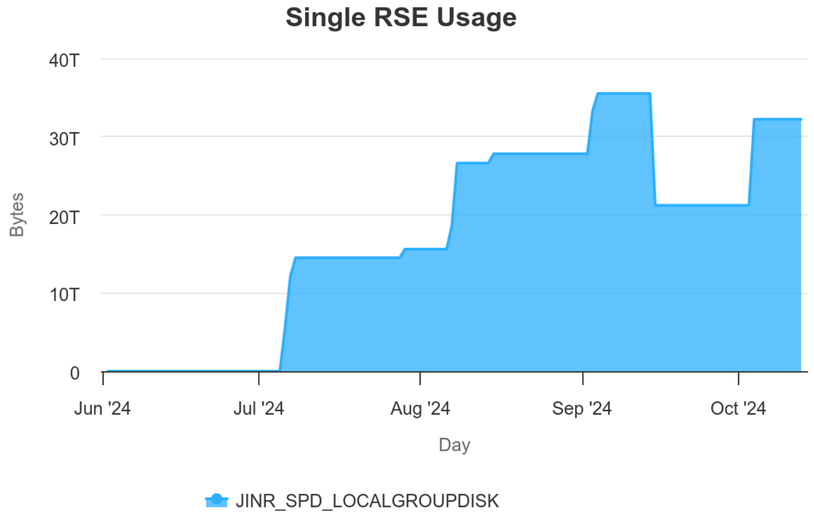


The Rucio Account Importer was implemented to import accounts and their user identification information from SPD IAM to Rucio

Data Production

- The scope jeditest was used to test the interaction of PanDA and Rucio

- Scopes 2024 and archive are significant data that appeared after the production



Name	Account	RSE Expression	Creation Date	Remaining Lifetime	State	Locks OK	Locks Replicating	Locks Stuck
2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.log	panda	JINR_SPD_LOCALGROUPDISK	2024-08-07T07:55:11.000Z	-	OK	5006	0	0
2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.P	panda	JINR_SPD_LOCALGROUPDISK	2024-08-07T07:55:12.000Z	-	OK	5000	0	0
2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.S	panda	JINR_SPD_LOCALGROUPDISK	2024-08-07T07:55:12.000Z	-	OK	5000	0	0
2024:2024.MC.27GeV.test-minbias.00001.RECO.6f25043e-689f-40f7-951f-ba6e0c9f4d14.1.R	panda	JINR_SPD_LOCALGROUPDISK	2024-08-15T09:07:38.000Z	-	OK	1426	0	0
2024:2024.MC.27GeV.test-minbias.00001.RECO.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0	panda	JINR_SPD_LOCALGROUPDISK	2024-09-26T11:43:57.000Z	-	OK	23	0	0
2024:2024.MC.27GeV.test-minbias.00001.RECO.6f25043e-689f-40f7-951f-ba6e0c9f4d14.1.log	panda	JINR_SPD_LOCALGROUPDISK	2024-09-26T11:44:52.000Z	-	OK	1451	0	0
2024:2024.MC.27GeV.test-minbias.00001.RECO.2.log	panda	JINR_SPD_LOCALGROUPDISK	2024-10-02T12:36:46.000Z	-	OK	4785	0	0
2024:2024.MC.27GeV.test-minbias.00001.RECO.2.R	panda	JINR_SPD_LOCALGROUPDISK	2024-10-02T12:36:46.000Z	-	OK	4753	0	0

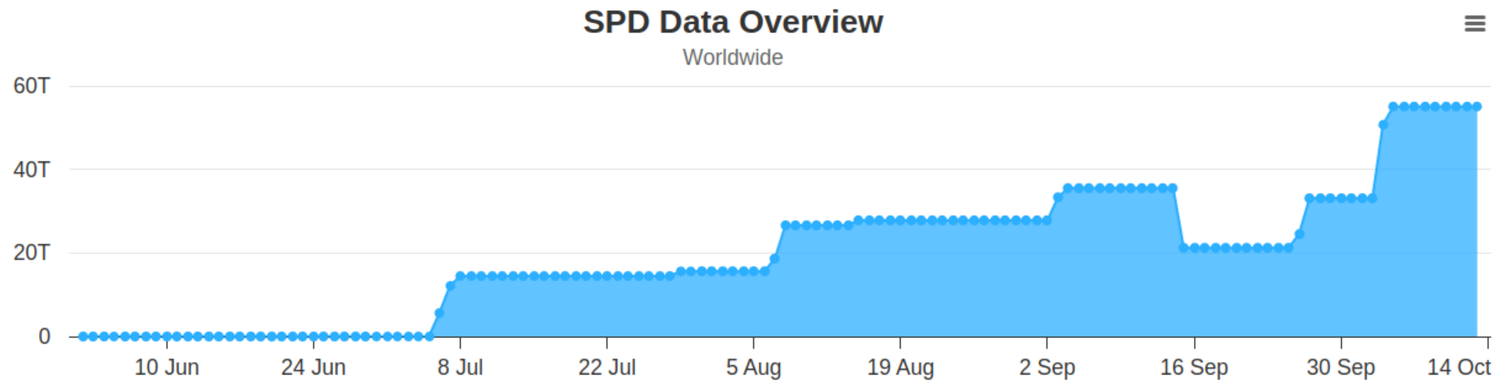
Data Replication 1/2

- It is planned to have at least two copies of important data – origin data at JINR and replicas somewhere else, at the moment only at PNPI

Name	Account	RSE Expression	Creation Date	Remaining Lifetime	State	Locks OK	Locks Replicating	Locks Stuck
2024:2024.MC.27GeV.test-minbias.00001.RECO.2.R	panda	PNPI_PROD_DATADISK	2024-10-02T12:39:47.000Z	-	STUCK	4741	0	12
2024:2024.MC.27GeV.test-minbias.00001.RECO.2.log	panda	PNPI_PROD_DATADISK	2024-10-02T12:39:46.000Z	-	STUCK	4769	0	16
2024:2024.MC.27GeV.test-minbias.00001.RECO.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0	panda	PNPI_PROD_DATADISK	2024-09-26T12:41:18.000Z	-	SUSPENDED	8	0	15
2024:2024.MC.27GeV.test-minbias.00001.RECO.6f25043e-689f-40f7-951f-ba6e0c9f4d14.1.log	panda	PNPI_PROD_DATADISK	2024-09-26T12:40:58.000Z	-	SUSPENDED	1429	0	22
2024:2024.MC.27GeV.test-minbias.00001.RECO.6f25043e-689f-40f7-951f-ba6e0c9f4d14.1.R	panda	PNPI_PROD_DATADISK	2024-09-26T12:40:38.000Z	-	SUSPENDED	537	0	889
2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.S	panda	PNPI_PROD_DATADISK	2024-09-26T12:39:55.000Z	-	SUSPENDED	4982	0	18
2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.P	panda	PNPI_PROD_DATADISK	2024-09-26T12:39:10.000Z	-	SUSPENDED	4977	0	23
2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.log	panda	PNPI_PROD_DATADISK	2024-09-26T12:36:57.000Z	-	SUSPENDED	4987	0	19

- 20 TB in the form of replicas are stored on the remote storage via FTS service

Data Replication 2/2



- Original files are stored in JINR storage, replicas — in PNPI storage
- After the expiration of the lifetime in the specified lifetime model this data were deleted automatically
- The file deletion rate in the current configuration is approximately 16 files per second

Example of the MC task definition 1/2

- Step 1: Simulation
- User defines an output dataset name
- Desired total number of events and events per job
- The system divides the total number of events by the number events per job and generates the required number of jobs
- User can specify either a specific computing queue or a cloud; in the second case, the jobs will be distributed among the queues of the specified cloud
- Jobs execution is performed in the container

```
TaskName = '2024.27GeV.test-MB.2st.DSSD.simu'
DatasetName = '2024.MC.27GeV.test-minbias.00001.SIMUL.0'
#DatasetName = 'jeditest.000023.simu'

taskParamMap = {}

taskParamMap['nEventsPerJob'] = 4000
taskParamMap['nEvents'] = 20000000
taskParamMap['noInput'] = True
taskParamMap['skipScout'] = True
taskParamMap['taskName'] = TaskName
taskParamMap['userName'] = 'Artem Petrosyan'
taskParamMap['vo'] = 'spd.nica.jinr'
taskParamMap['taskPriority'] = 900
taskParamMap['architecture'] = 'x86_64'
taskParamMap['transUses'] = 'A'
taskParamMap['transHome'] = None
taskParamMap['transPath'] = 'https://159.93.221.125:8080/spd_simu_VA_transform.sh'
taskParamMap['processingType'] = 'step1'
taskParamMap['prodSourceLabel'] = 'managed'
taskParamMap['taskType'] = 'test'
taskParamMap['workingGroup'] = 'spd.nica.jinr'
taskParamMap['cloud'] = 'JINR'
taskParamMap['ramCount'] = 1900

outDatasetNameLog = '{0}.log'.format(DatasetName)
outDatasetNameS = '{0}.S'.format(DatasetName)
outDatasetNameP = '{0}.P'.format(DatasetName)

taskParamMap['log'] = {'dataset': outDatasetNameLog,
                       'type':'template',
                       'param_type':'log',
                       'token':'DATADISK',
                       'value':'{0}.${{SN}}.log.tgz'.format(DatasetName)}

taskParamMap['jobParameters'] = [
    {'type': 'constant',
     'value': '''singularity run --bind /cvmfs/spd.jinr.ru/production/MC/2024.27GeV.test-MB.2st.DSSD:/prod -H
     ./:/WORKDIR
    /cvmfs/spd.jinr.ru/images/spdroot-4.1.6.sif spdroot.py -b -q \'/prod/simu.C({0}, '''.format(taskParamMap['nEventsPerJob']),
     },
```

Example of the MC task definition 2/2

- Step 2: Reconstruction
- User defines a name of the input dataset, in this example there are two input datasets of the same size (have the same number of files)
- Sets a name of the output dataset
- Set how many jobs needs to be created per each file in the dataset
- At the job generation stage, the workload management system communicates with the data management service, reads the size (number of files) of the dataset and generates the appropriate number of jobs
- The input files will be staged-in from the storage closest to the computing node

```
scope = '2024'
inDatasetName = '2024.MC.27GeV.test-minbias.00001.SIMUL.0'
outDatasetName = '2024.MC.27GeV.test-minbias.00001.RECO.2'

inDatasetNameS = '{0}.S'.format(inDatasetName)
inDatasetNameP = '{0}.P'.format(inDatasetName)
outDatasetNameR = '{0}.R'.format(outDatasetName)
outDatasetNameLog = '{0}.log'.format(outDatasetName)

taskParamMap = {}

taskParamMap['nFilesPerJob'] = 1
taskParamMap['nEventsPerJob'] = 4000
taskParamMap['noInput'] = False
taskParamMap['taskName'] = TaskName
taskParamMap['userName'] = 'Artem Petrosyan'
taskParamMap['vo'] = 'spd.nica.jinr'
taskParamMap['taskPriority'] = 900
taskParamMap['architecture'] = 'x86_64'
taskParamMap['transUses'] = 'A'
taskParamMap['transHome'] = None
taskParamMap['transPath'] = 'https://159.93.221.125:8080/spd_simu_VA_transform.sh'
taskParamMap['processingType'] = 'step2'
taskParamMap['prodSourceLabel'] = 'managed'
taskParamMap['taskType'] = 'test'
taskParamMap['workingGroup'] = 'spd.nica.jinr'
taskParamMap['cloud'] = 'JINR'
taskParamMap['ramCount'] = 1900

taskParamMap['log'] = {'dataset': outDatasetNameLog,
                       'type': 'template',
                       'param_type': 'log',
                       'token': 'DATADISK',
                       'value': '{0}.${{SN}}.log.tgz'.format(outDatasetName)}

taskParamMap['jobParameters'] = [
    {'type': 'constant',
     'value': '''singularity run --bind /cvmfs/spd.jinr.ru/production/MC/2024.27GeV.test-MB.2st.DSSD:/prod -H
     ./:/WORKDIR /cvmfs/spd.jinr.ru/images/spdroot-4.1.6.1.sif spdroot.py -b -q \'/prod/reco.C({0}, '''
     .format(taskParamMap['nEventsPerJob'])
     },
```

Production Manager Control Panel 1/3

Welcome, monakov Create task Tasks Log out

Select field | Enter value | Filter

Task ID	Task name	Parent ID	Creator	Status	Done jobs	Default/Current priority	Total events	Submit time	Start time
211	2024.27GeV.test-MB.2st.DSSD.reco	211	Artem Petrosyan	finished	4753	900/900	0	12:36, 02 Oct 2024	12:36, 02 Oct 2024
210	2024.27GeV.test-MB.2st.DSSD.reco	210	Artem Petrosyan	aborted	532	900/900	0	09:07, 15 Aug 2024	18:20, 15 Aug 2024
209	2024.27GeV.test-MB.2st.DSSD.reco	209	Artem Petrosyan	aborted	0	900/900	0	13:07, 14 Aug 2024	00:22, 15 Aug 2024
208	2024.27GeV.test-MB.2st.DSSD.reco	208	Artem Petrosyan	aborted	0	1000/1000	0	12:43, 14 Aug 2024	12:44, 14 Aug 2024
207	2024.27GeV.test-MB.2st.DSSD.reco	207	Artem Petrosyan	aborted	0	900/900	0	12:06, 14 Aug 2024	12:07, 14 Aug 2024
206	2024.27GeV.test-MB.2st.DSSD.reco	206	Artem Petrosyan	broken	0	900/900	0	12:44, 13 Aug 2024	15:19, 13 Aug 2024
205	2024.27GeV.test-MB.2st.DSSD.reco	205	Artem Petrosyan	broken	0	900/900	0	12:42, 13 Aug 2024	None
204	2024.27GeV.test-MB.2st.DSSD.simu	204	Monakov Nikita Glebovich	submitting	0	1000/1000	0	15:49, 12 Aug 2024	15:50, 12 Aug 2024
203	2024.27GeV.test-MB.2st.DSSD.simu	203	Monakov Nikita Glebovich	submitting	3	1000/1000	200	15:29, 12 Aug 2024	10:53, 15 Aug 2024
202	2024.27GeV.test-MB.2st.DSSD.simu	202	Monakov Nikita Glebovich	submitting	1	1000/1000	100	15:29, 12 Aug 2024	12:34, 02 Oct 2024

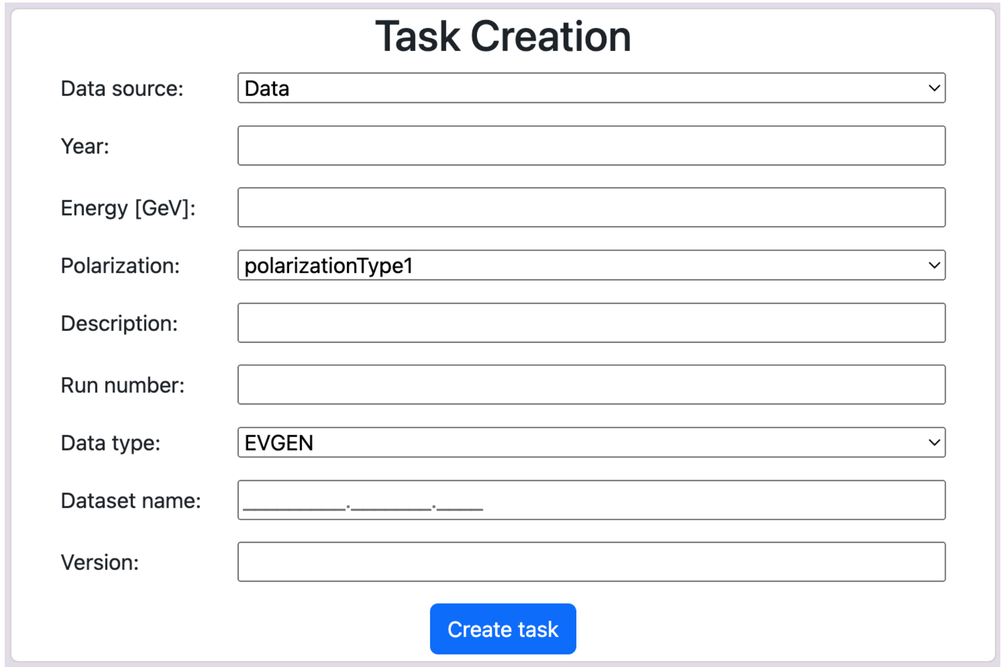
Page: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

- A control panel for pipelines submissions is under development

Production Manager Control Panel 2/3

Task ID	211
Parent task ID	211
Chain task ID	None
Name	2024.27GeV.test-MB.2st.DSSD.reco
Type	managed
VO	spd.nica.jinr
Creator	Artem Petrosyan
State status	finished
Total events	0
Total jobs done	4753
Total requested jobs	8365
Priority	900
Current priority	900
Submit time	12:36 October 2, 2024
Start Time	12:36 October 2, 2024
Time stamp	2:36 October 9, 2024
JEDI parameters	{"nFilesPerJob": 1, "nEventsPerJob": 4000, "noInput": false, "taskName": "2024.27GeV.test-MB.2st.DSSD.reco", "userName": "Artem Petrosyan", "vo": "spd.nica.jinr", "taskPriority": 900, "architecture": "x86_64", "transUses": "A", "transHome": null, "transPath": "http://159.93.221.125:8080/spd_simu_VA_transform.sh", "processingType": "step1", "prodSourceLabel": "managed", "taskType": "test", "workingGroup": "spd.nica.jinr", "cloud": "JINR", "ramCount": 1900, "log": {"dataset": "2024.MC.27GeV.test-minbias.00001.RECO.2.log", "type": "template", "param_type": "log", "token": "DATADISK", "value": "2024.MC.27GeV.test-minbias.00001.RECO.2.${SN}.log.tgz"}, "jobParameters": [{"type": "constant", "value": "singularity run --bind /cvmfs/spd.jinr.ru/production/MC/2024.27GeV.test-MB.2st.DSSD:/prod -H ./:/WORKDIR /cvmfs/spd.jinr.ru/images/spdroot-4.1.6.1.sif spdroot.py -b -q '/prod/reco.C(4000, "}, {"type": "constant", "value": "\\\""}, {"type": "template", "param_type": "input", "value": "${INS}", "dataset": "2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.S"}, {"type": "constant", "value": "\\\","}, {"type": "constant", "value": "\\\""}, {"type": "template", "param_type": "input", "value": "${INP}", "dataset": "2024:2024.MC.27GeV.test-minbias.00001.SIMUL.6f25043e-689f-40f7-951f-ba6e0c9f4d14.0.P"}, {"type": "constant", "value": "\\\","}, {"type": "constant", "value": "\\\""}, {"type": "template", "param_type": "output", "token": "DATADISK", "value": "r.2024.MC.27GeV.test-minbias.00001.RECO.2.${SN/P}.root", "dataset": "2024.MC.27GeV.test-minbias.00001.RECO.2.R"}, {"type": "constant", "value": "\\\","}, {"type": "template", "value": "${RNDMSEED}", "param_type": "number"}, {"type": "constant", "value": ")'"}]}

Production Manager Control Panel 3/3



Grouping tier	Field	Description	Example
0	[YEAR]	Main Scope - the year of data production	2050
1	[MC | DATA]	Real data or simulated data	DATA
2	[energy][polarization]		250LT
3	[desc]	Short name of physics aim	minbias
4	[RunNumber]	Run number for DATA, ID for MC	27189
5	[data type]	EVGEN, SIMUL, RECO….	RAW
6	[DatasetUID]	unique ID of the dataset	636763fd78df7d
7	[Version]	for reprocessing	0

- In order to ease metadata catalog navigation, data filtration, identification, etc., a datasets naming convention was proposed in November 2023
- Dataset name example: 2025.MC.250LT.minbias.27189.RAW.636763fd78df7d.0
- Control panel checks input parameters and ensures that dataset is created in accordance with the naming convention

Summary

- The SPD detector will generate data streams that cannot be guaranteed to be stored and processed within a single data center
- We plan to use all possible computing resources for processing, and try to use them in the most optimal way, distributing jobs to the most suitable centers
- To control the data processing processes of the SPD experiment, we are building a highly automated system that takes into account a variety of parameters: file sizes, their location, CPUs and memory suitable for each stage of processing, the state of network connections, and so on.
- Such complex distributed systems require special attention to security issues, in which, among other things, a multi-level monitoring system plays an important role
- Our nearest activities lie in the field of automation, production manager control panel enhancement and processing monitoring development

Thank you for attention!