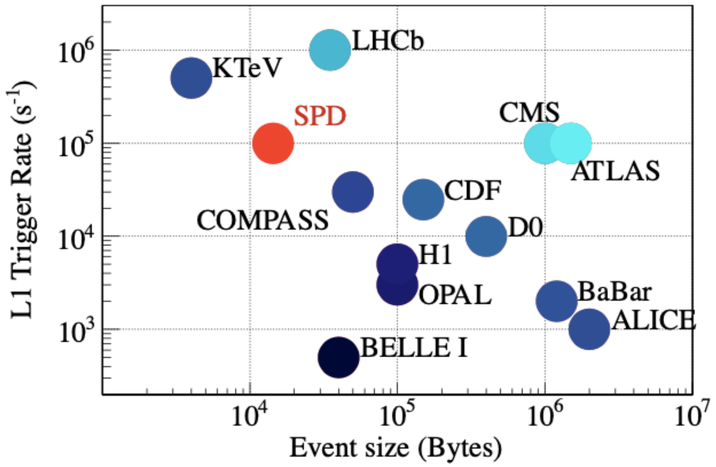
Offline data processing for the SPD experiment at NICA collider

A. Petrosyan
Conference on High Energy Physics, AANL, Yerevan, Armenia
September 14, 2023

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.

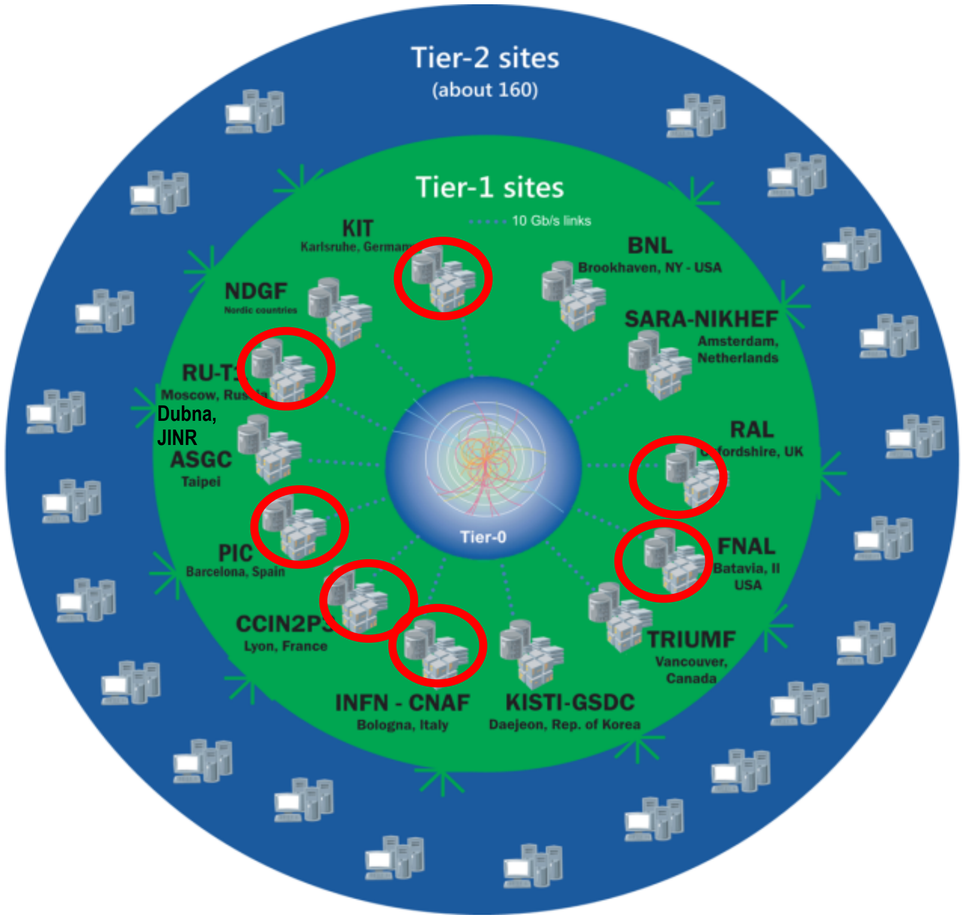
- SPD TDR



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.

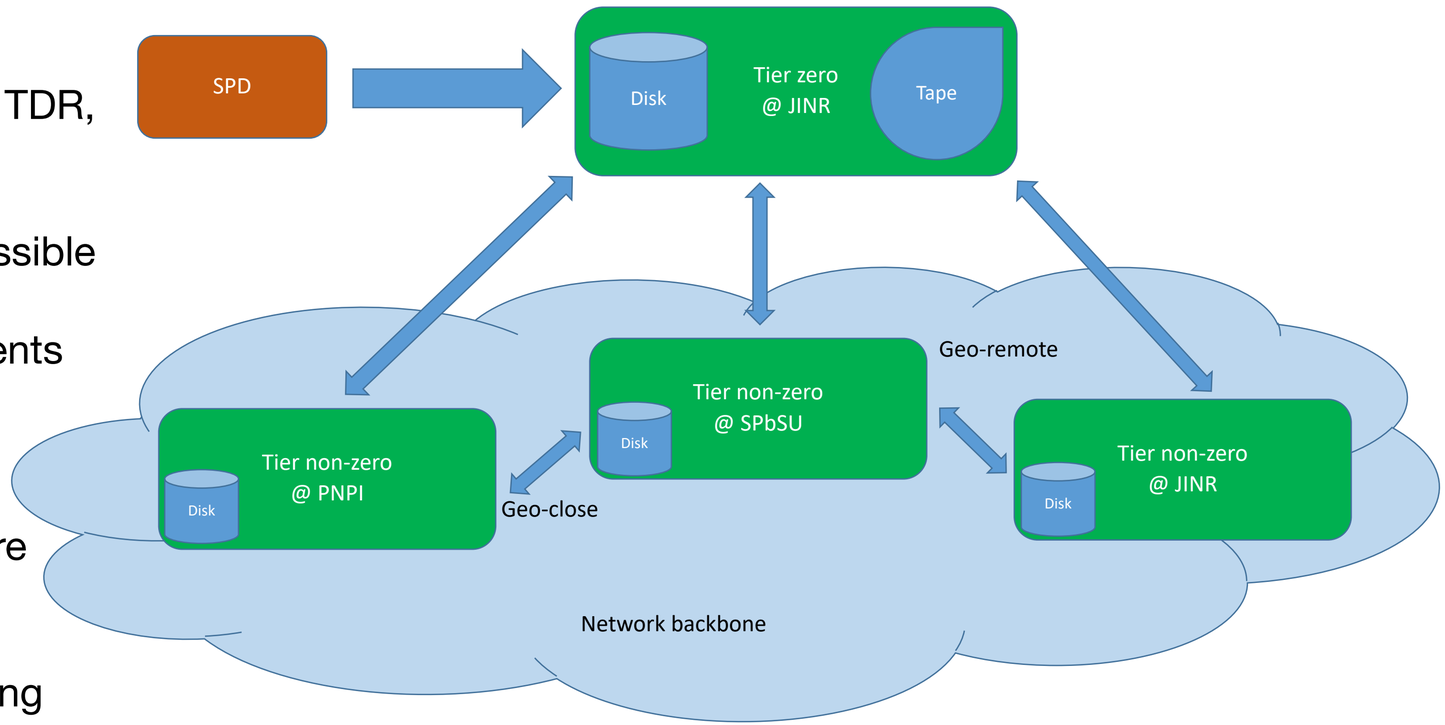
What about resources?

- In June 2023 it was announced that that JINR will not cover more than 25% of computing needs of the experiment
- Finding external collaborators for data processing is one of the highest priority tasks
- The picture at the right looks so familiar and it’s our natural way of development in order to build a geographically distributed computing environment
- We are not going to inherit a fixed tiers model, in our case there will be a data transfers mesh based on the several metrics: network, load, data availability, etc.



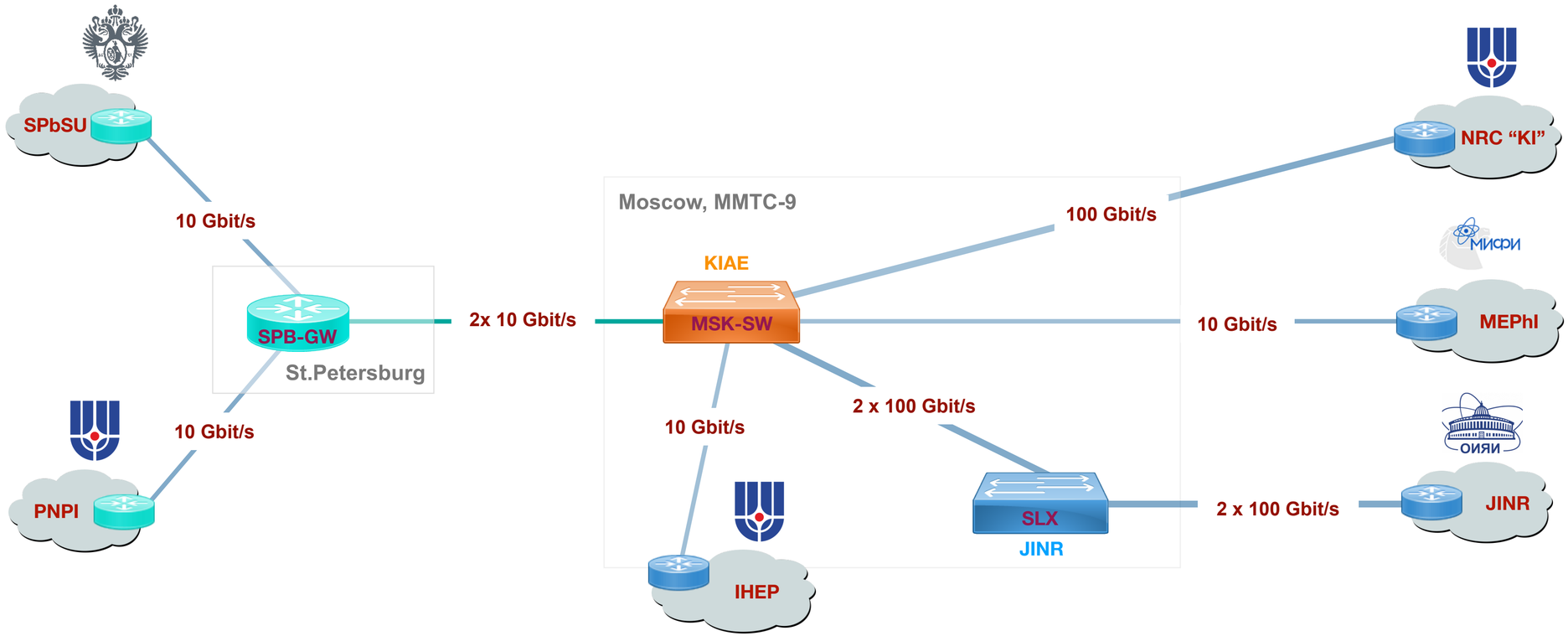
External participants

- 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 physical resources with central data management services



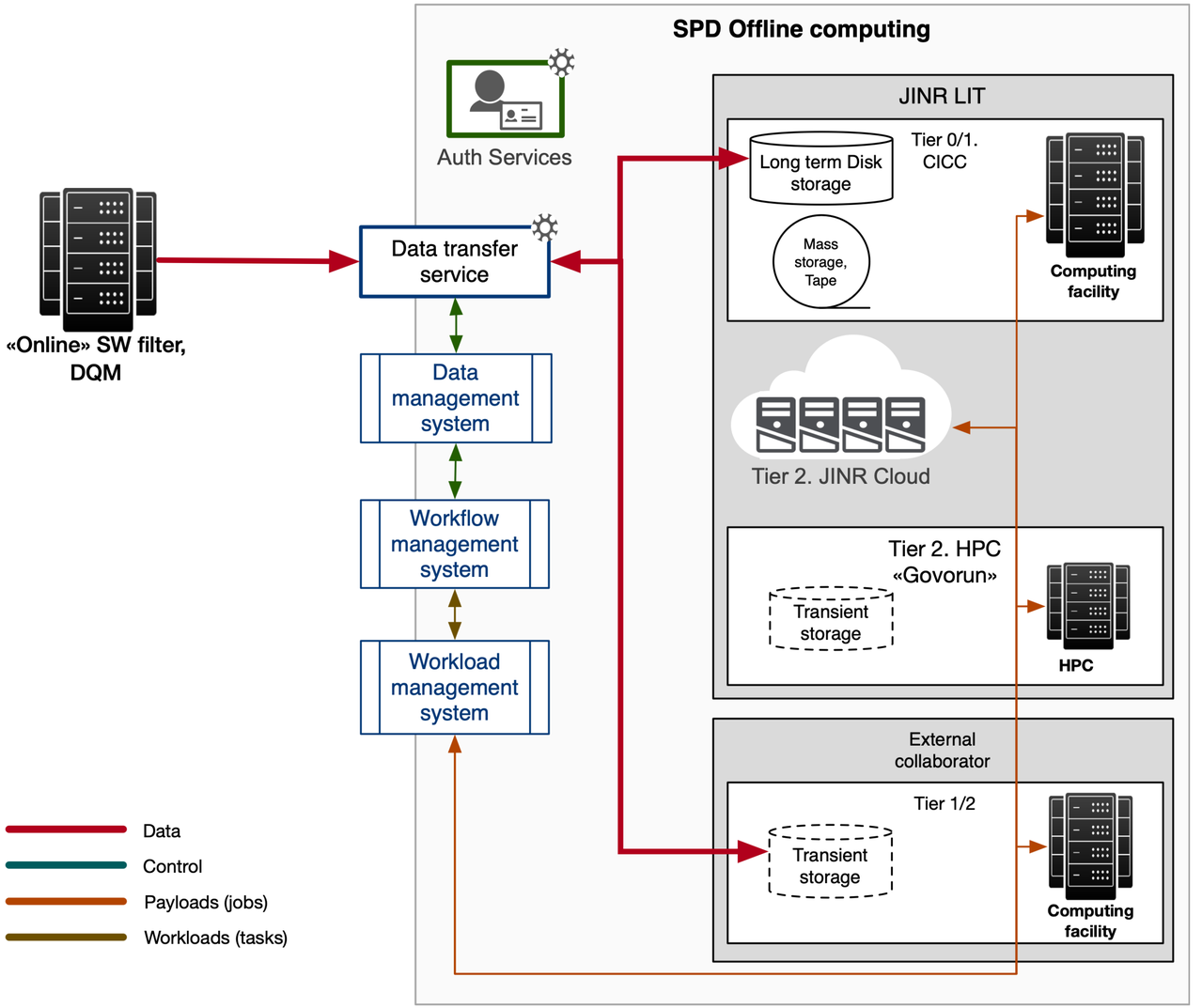
Russian WLCG network backbone

- Network bandwidth, amount of CPU and storage capacity is a combination of factors which allow to take part in SPD computing
- Russian “old school” WLCG computing centres are the most likely candidates for such role



Computing system components

- CRIC information system — the main integration component of the system: gathers 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



Computing Resource Information Catalog

- Initially developed for ATLAS collaboration at LHC, named AGIS at that time: ATLAS Grid Information System
- Now CRIC is an information system for WLCG (Worldwide LHC Computing Grid) and individual experiments, like ATLAS, CMS and COMPASS
- CRIC was deployed at LIT in 2020, tested with BM@N and now is the information system for the SPD experiment
- Our installation is not the same as ATLAS has, we support our own branch on the top of the CRIC core

PanDA workflow management system

- Initially developed for ATLAS collaboration at LHC
- Now PanDA running jobs for ATLAS at CERN, COMPASS at CERN, sPHENIX at BNL and Vera C. Rubin Observatory in Chile
- PanDA was deployed at LIT in 2015 in order to manage COMPASS data processing, another instance of PanDA was deployed in 2020, tested with BM@N jobs and now is the workflow/workload management system for the SPD experiment
- We are running the latest version of PanDA with experiment-driven extensions

Rucio scientific data management system

- Initially developed for ATLAS collaboration at LHC
- Rucio was deployed at LIT in 2022, at the moment we’re learning how to work with this system, but it is already integrated with PanDA and is being already used as data catalog

File Transfer Service

- Developed by the CERN IT in order to enable massive parallel transfers
- Allows to organize asynchronous data transfers, takes care of all checksum, retries, etc. machinery, enables transfers from and to discs and tapes
- A lot of experiments use FTS to transfer their data
- Not yet installed, we are going to install and support FTS as a central service for many our experiments, not only for the SPD

Resources and services provided by MLIT

- Certification Authority, VOMS/IAM, CVMFS
- MICC
  - Cloud (IaaS, dedicated VM, spinning disks and SSD)
  - CICC (Slurm batch, grid ARC6 CE)
  - Govorun HPC, HybriLIT
- Disk and tape storage (EOS, grid SE)
- It would be great to organize a DBOD (Database On Demand) service at MLIT with support of popular RDBMS systems, for example, PostgreSQL and MariaDB

GitLab
http://git.jinr.ru/

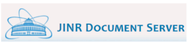
Use simply as a version control system or build a more complex development process using Git repositories, issue management and CI/CD.

Project managment
http://pm.jinr.ru/

Follow the progress of the project, create, execute tasks, coordinate the work of project participants.

JINR disk
http://disk.jinr.ru/

Sync your files between your devices. Share it with your colleagues, or just keep a reserve copy.



JDS
http://jds.jinr.ru/

An institutional Open Access (OA) repository of articles, preprints, books, audio and video lections, and other materials that reflect and promote research activities at JINR.

Helpdesk
http://helpdesk.jinr.ru/

Ask questions and get qualified answers.

Network service
http://noc.jinr.ru/

E-mail, remote access, digital telephony, mailing lists, access to scientific libraries.

Indico
http://indico.jinr.ru/

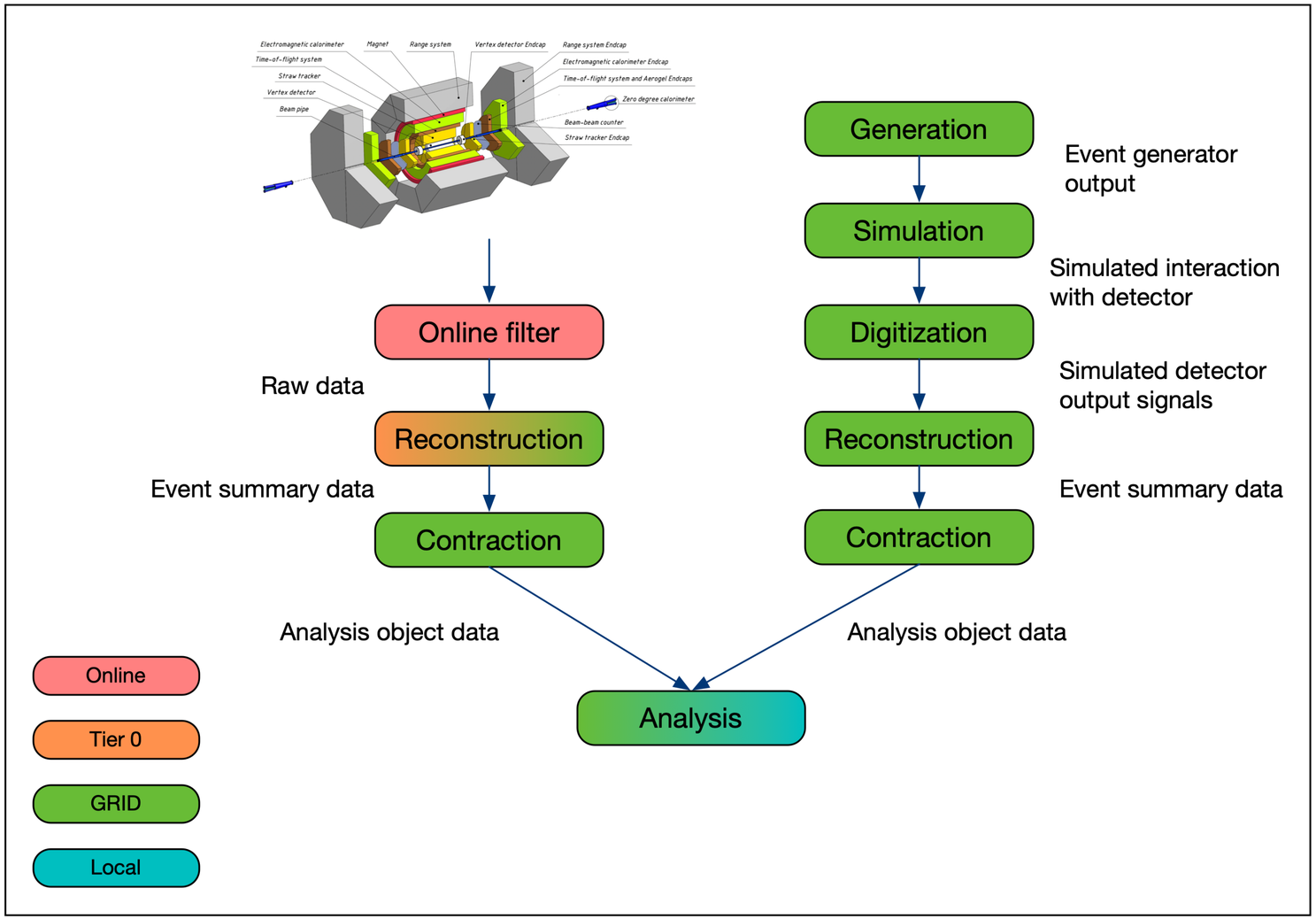
Manage conferences, seminars and meetings.

PIN
http://pin.jinr.ru/

Get detailed information about results of the research activities of the Institute staff.

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



How we recommend to use our computing resources

- Personal computers and laptops: development, initial testing
- JINR cloud service: development, testing in the environment, profiling, testing with containers
- Batch service: personal analysis, small tasks
- Production system over the distributed computing resources: large (thousands jobs) tasks over large datasets, chain of tasks and workflows, mass productions in the interest of the collaboration

Summary

- MLIT provides all necessary services for building data processing system of the experiment
- All needs of users wishing to organize personal calculations are covered within the JINR: they have access to cloud, batch, HPC and several storage systems
- Almost all components of the production system, which will be responsible for running massive calculations in the distributed environment have already deployed and configured
    - First test production (samples of D-meson decays and min. bias) is now ongoing
- External participants demonstrate interest and intention to participate in the software development and data processing: for example, an agreement of collaboration in computing with PNPI has just signed, grid queues of PNPI, SPbU and INP BSU were connected to the distributed computing infrastructure of the experiment in 2022
- There is an ongoing work on automatic image building (CI/CD) of SpdRoot in order to have the freshest “official” version of the framework always available at CVMFS
- Our next steps will lay in the field of data management and databases: Rucio integration, FTS deployment, geometry versions, calib&align, magnetic field, etc.

Thank you!

Backup

Principles of the computing system

- Jobs -> tasks -> trains of tasks -> workflows
- Files -> datasets
- Gather and keep metadata of each workflow/task/job, dataset/file and event
- Advanced strategies for managing data lifetime: some files to be deleted immediately, other will be kept till the end of the production, another will be stored for some period after being gathered, and the most important data must be kept forever
- Concentrating on the use of containers as an universal response to the variety of software versions on computing centres: we publish tagged version of SpdRoot at CVMFS so that anyone can use it
- We keep in mind multiprocessor, multithreading and non x86 architectures of the modern hardware — there is ongoing development of Gaudi based framework for SPD, which will replace SpdRoot, based on ROOT framework

Production setup

- CVMFS as an entry point to the “official” versions of SpdRoot
- Production setups on the CVMFS in form of frozen sandboxes
- Each new production means new directory with all dependencies on CVMFS
- Each production on CVMFS corresponds to path with the same name on EOS
- Production role in VOMS for users who run mass production
- Directory to store results of the production on EOS with strict access rights in order not to be deleted accidentally