SPD offline computing system: status report

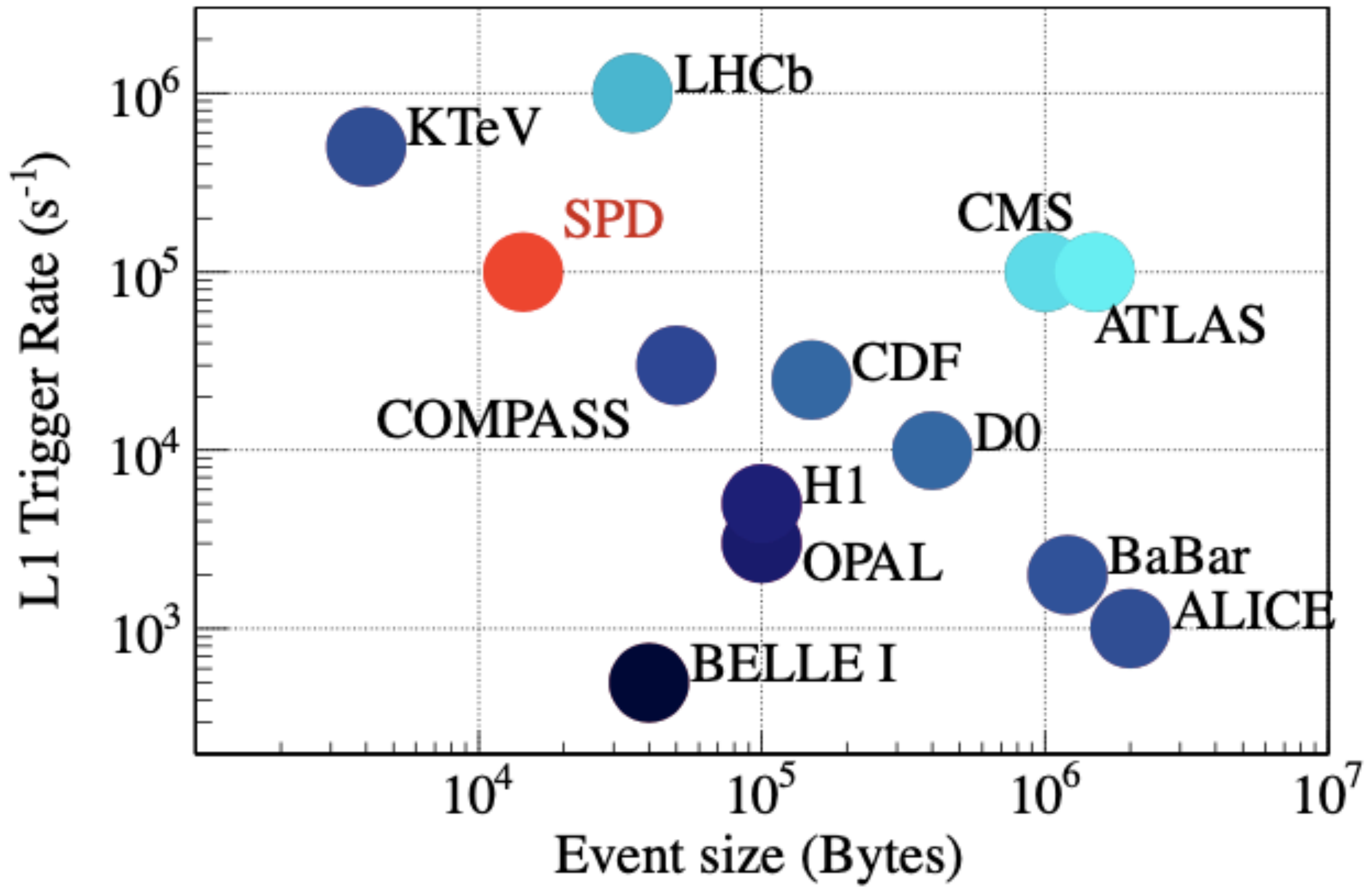
A. Petrosyan (MLIT JINR)

VI SPD Collaboration Meeting and Workshop on Information Technology in Natural Sciences
Samara University, Samara, October 26, 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

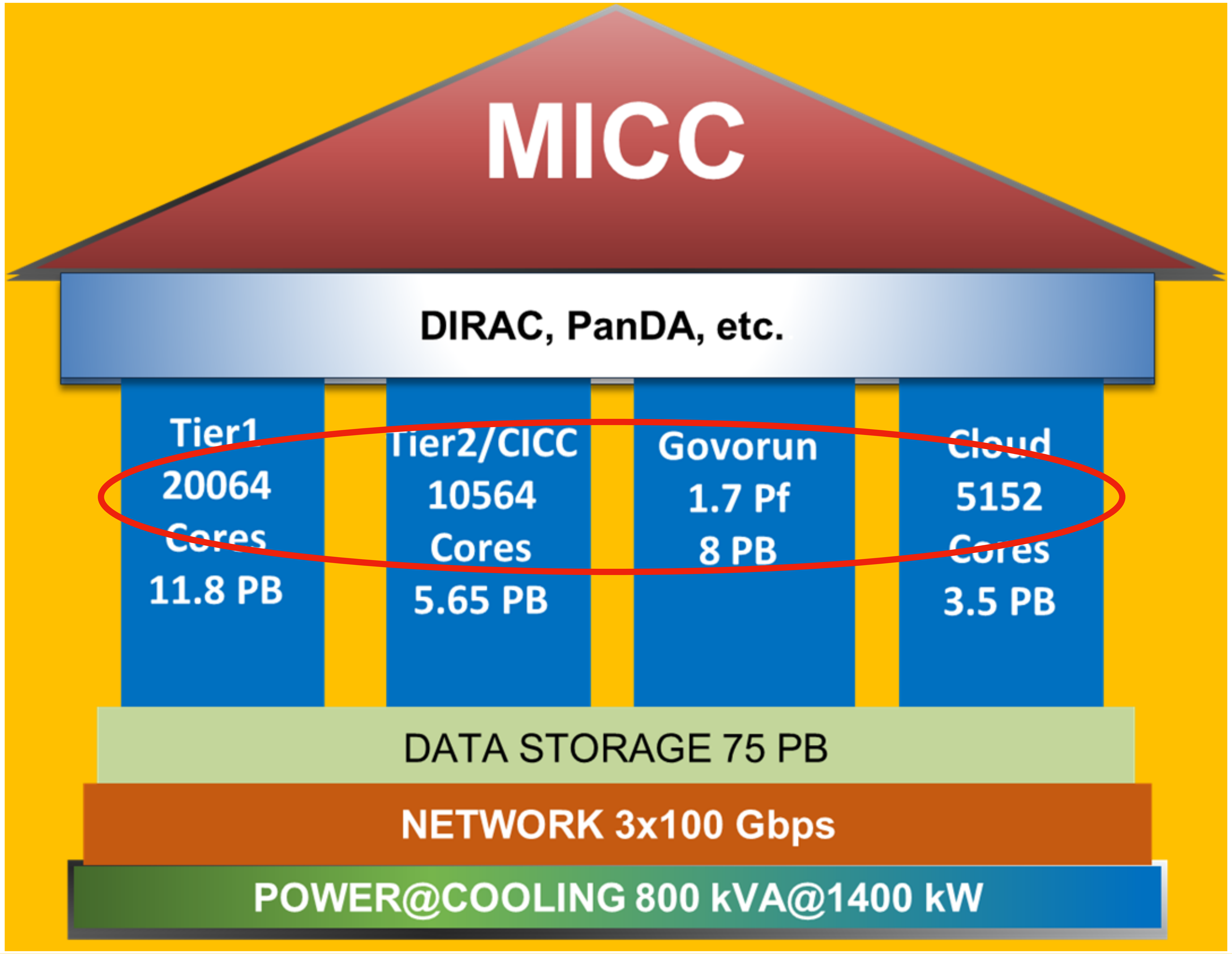


The goal of the online filter is at least to decrease the data rate by a factor of 20, so that the annual growth of data, including the simulated samples, stays within 10 PB. Then, data are transferred to the Tier-1 facility, where a full reconstruction takes place and the data is stored permanently. 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.

SPD as datasource

- We can expect ~10 PB of data each year, both MC and “real” data
- Having event size around 10-15 KB
- And 1 second to process each event
- We will need a computing system of around 60000 CPU
- One can say that during the first stage we’ll have 10 times less data, but at the moment our processing rate is far from 1 sec per event, it is close to 5-6 secs per event
- We need to control file sizes: too small ones will overload data catalogs, file systems and workload management system (1 file = 1 job), while too large will be too hard to transfer and store, will be taking too long to process one piece of work
- The optimal size is a file which can be processed within 6-8 hours, with size around 6-10 GB

Multifunctional Information and Computing Complex (MICC)



4 advanced software and hardware components

- Tier1 grid site
- Tier2/CICC site
- hyperconverged “Govorun” supercomputer
- cloud infrastructure

Distributed multi-layer data storage system

- Disks
- Robotized tape library

Network

- Wide Area Network
- Local Area Network

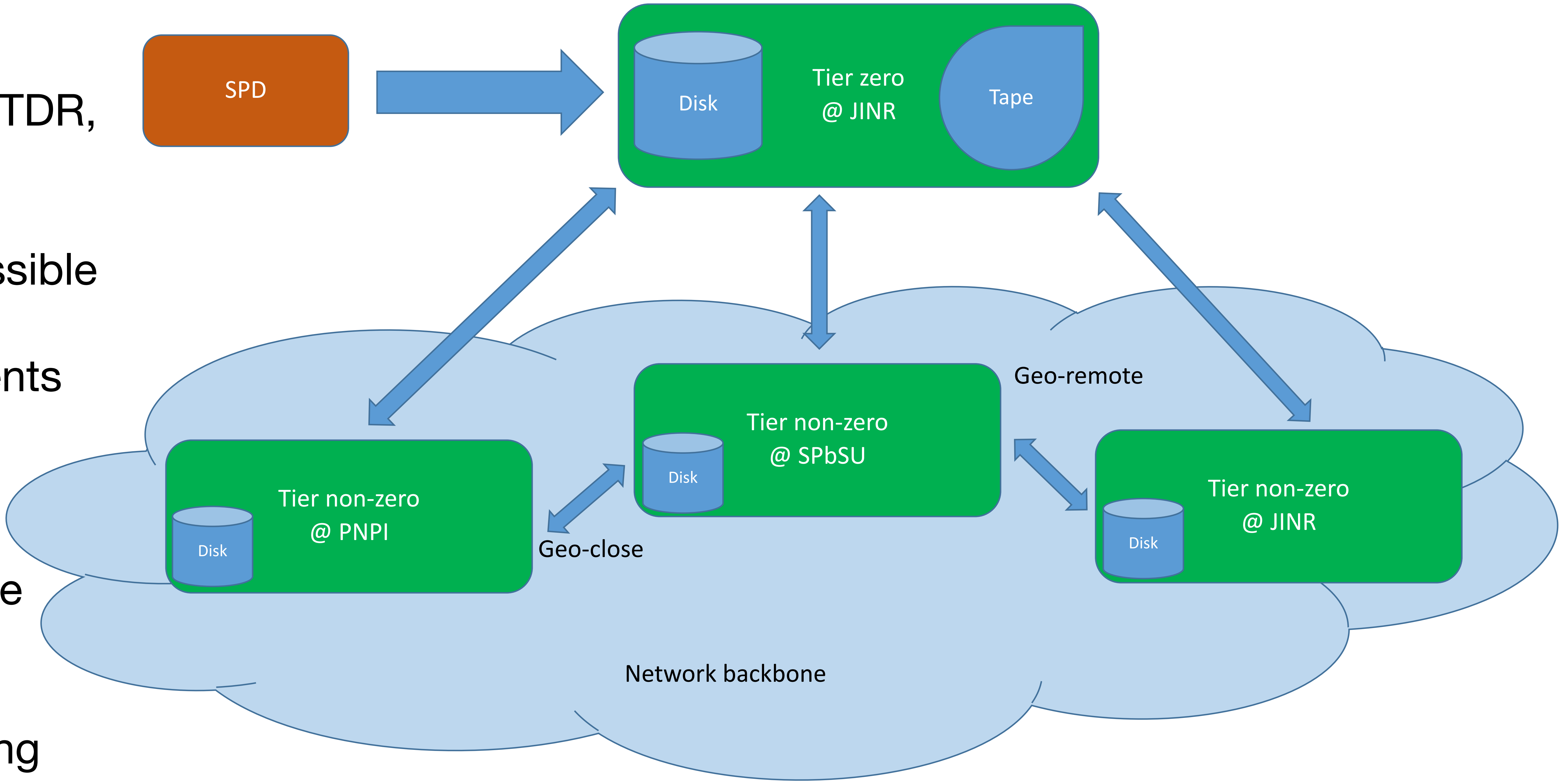
Engineering infrastructure

- Power
- Cooling

The main objective of the project is to ensure multifunctionality, scalability, high performance, reliability and availability in 24x7x365 mode for different user groups that carry out scientific studies within the JINR Topical Plan.

External participants

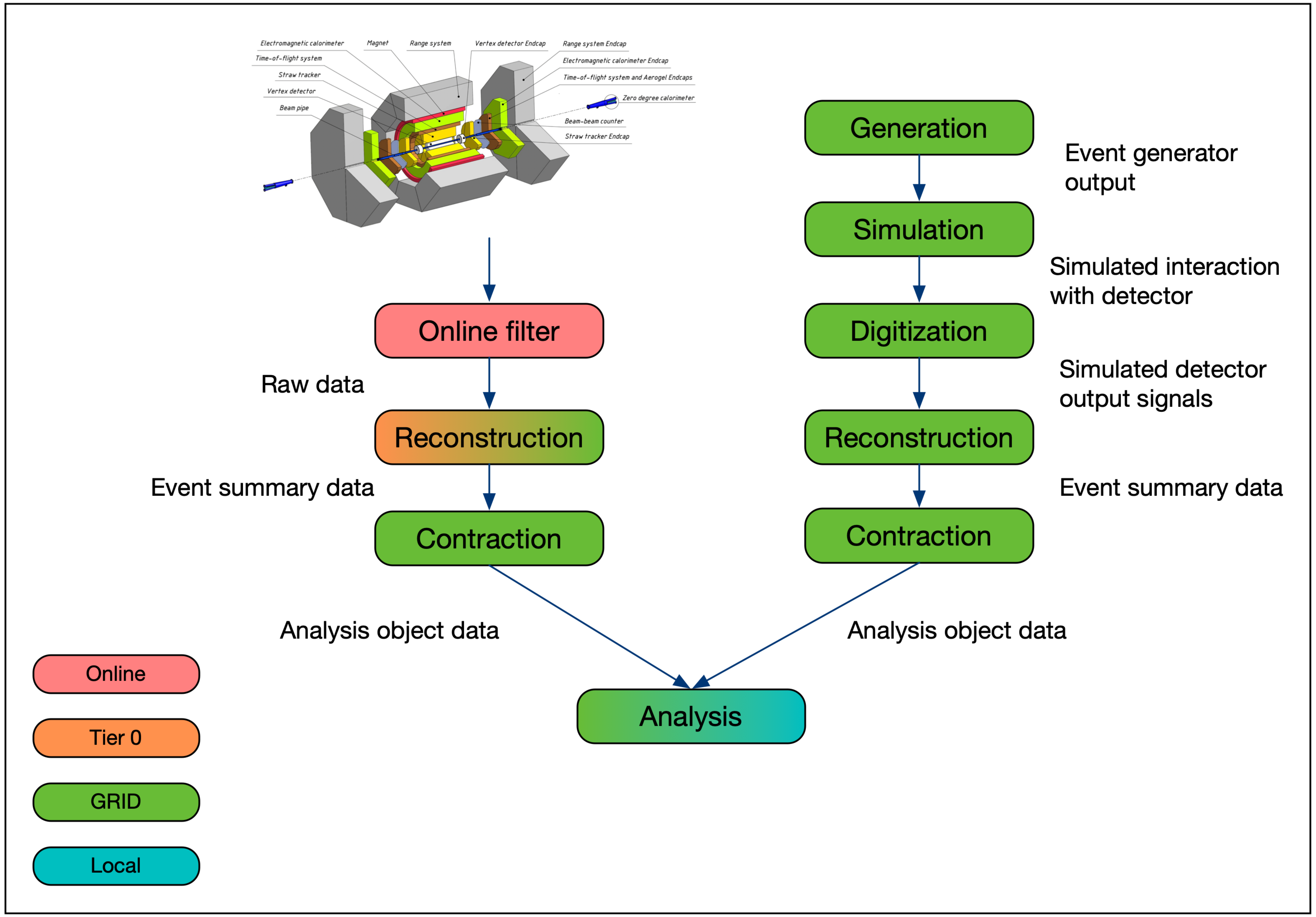
- Data volume mandates some baselines
  - >10 Gbps network per site (from TDR)
  - >500 TB storage capacity per site (not from TDR, but might be added to the next version)
- Try to use existing free software as much as possible
  - Experience comes from large LCG experiments
- Optimize management and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



Picture by Andrey Kiryanov from V SPD Collaboration Meeting

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



Online filter

Raw data

Reconstruction

Event summary data

Contraction

Analysis object data

Generation

Event generator output

Simulation

Simulated interaction with detector

Digitization

Simulated detector output signals

Reconstruction

Event summary data

Contraction

Analysis object data

Analysis

🟥 Online

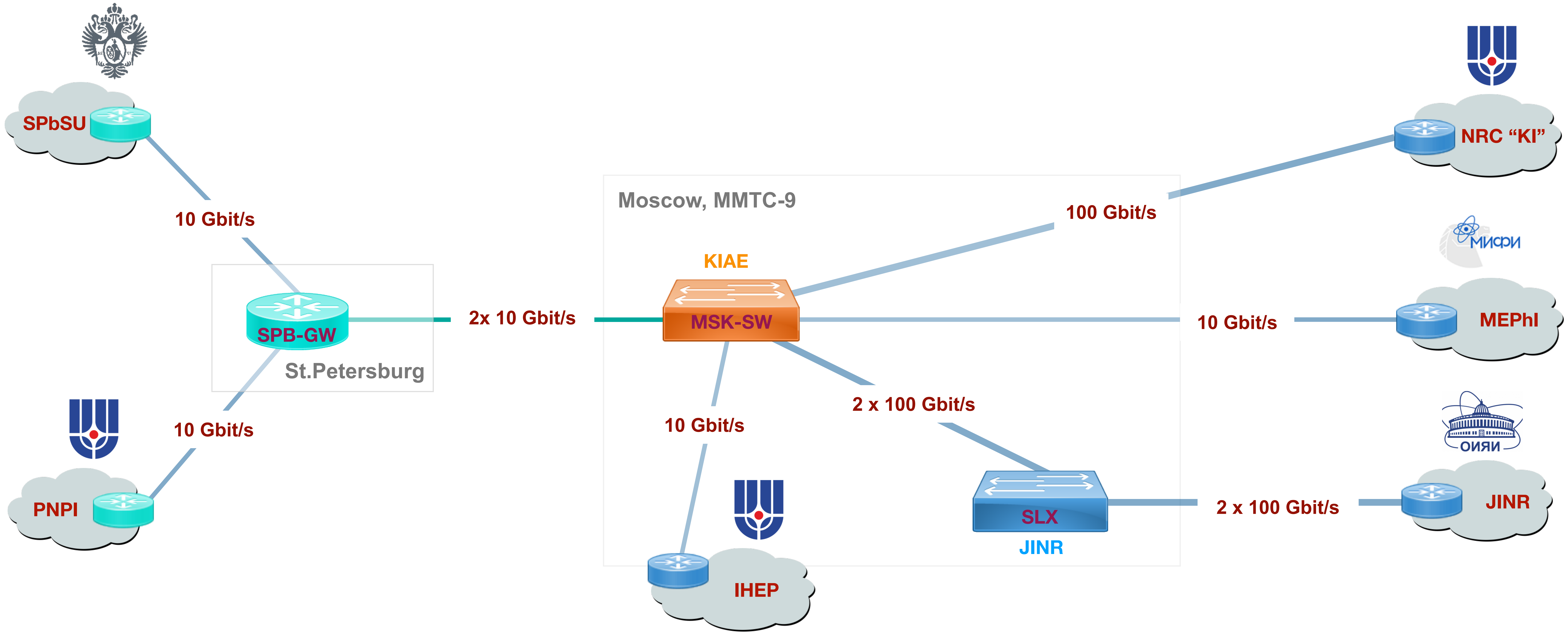
🟧 Tier 0

🟩 GRID

🟦 Local

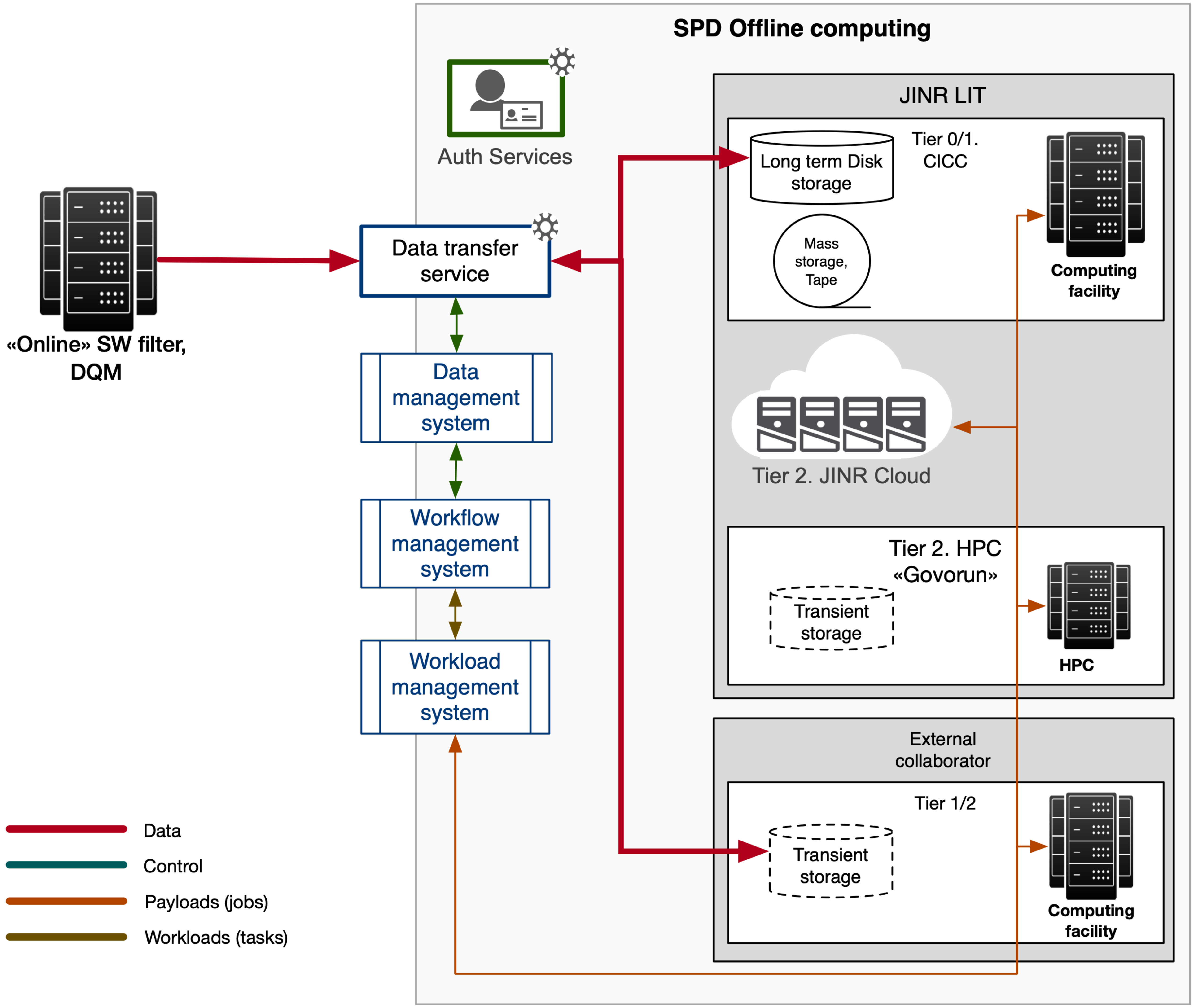
Russian WLCG network backbone

- 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 centers are the most likely candidates for this role



Computing system components

- CRIC information system — the main integration component of the 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



Computing system services status

CRIC

Computing Resource Information Catalog

- 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

- 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 SPD-driven extensions

- 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 used as data catalog for jobs and tasks data

FTS

- Not yet installed, we are going to install and support FTS as a central service for many our experiments, not only for the SPD

Production system

- In two words the production system is a database with a web UI which will be used to define SPD data processing tasks and campaigns, store the history and via API manages tasks, jobs, data and transfers in the computing system services
- What must be kept for each task?
  - Location of the applied software (SpdRoot) at CVMFS
  - Location of input parameters, like input files, datasets or parameters (number of events, flags and so on) to be passed to SpdRoot (or Gaudi) or other applied software
  - Location and metadata of the input and output data
- Once being prepared in the ProdSys, the task will be sent to the WFMS and, as soon as it’s done, the task is closed and can be used as an input data by for later processing
- The ProdSys database in combination with the frozen sandboxes at CVMFS allows to reproduce any processing at any time

Production setup

- CVMFS as an entry point to the “official” versions of SpdRoot:
  - /cvmfs/spd.jinr.ru/images/spdroot-version.sif
- Production setups on the CVMFS in form of frozen sandboxes
  - /cvmfs/spd.jinr.ru/production/processing type?/year?/campaign?/testOpenCharm/simu_VA_OC.C
- Each new production means new directory with all dependencies on CVMFS
- Each production on CVMFS corresponds to path with the same name on EOS
  - /eos/nica/spd/production/processing type?/year?/campaign?/testOpenCharm/
- Directory to store results of the production on EOS with strict access rights in order not to be deleted accidentally

MC task work flow

- Starts from request done by the MC production manager via the web UI of the ProdSys, all other steps to be done automatically
- Send jobs and processing on the computing nodes of the available computing resources
- Results are written to LIT EOS
- In case of merging data to be taken from EOS and sent to the nodes, results are written to EOS
- Once all processing steps are finished, migrate results to LIT tapes through CTA
- End

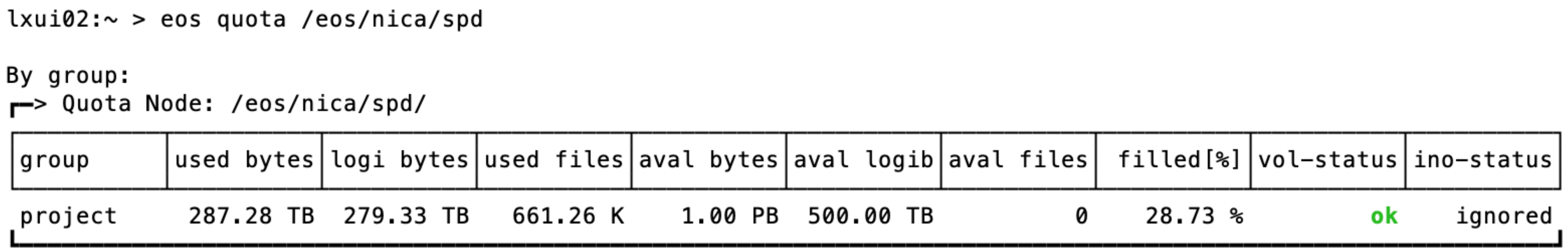
Storage

- There are many issues with JINR EOS (on both sites), there is ongoing discussion in LIT with what disc storage system we will proceed

- CTA is being prepared at LIT, we hope that soon we'll be able to store our sensitive data on tapes, EOS will be used only as disk buffer for a limited period of time

- We expect to have much larger quota at tapes than on disks

- Meanwhile, please look after your data at EOS carefully and do not forget that our quota at EOS is not infinite

- 

```
lxui02:~ > eos quota /eos/nica/spd

By group:
┏━> Quota Node: /eos/nica/spd/
┌──────────┬──────────┬──────────┬──────────┬──────────┬──────────┬──────────┬──────────┬──────────┬──────────┐
│group     │used bytes│logi bytes│used files│aval bytes│aval logib│aval files│ filled[%]│vol-status│ino-status│
└──────────┴──────────┴──────────┴──────────┴──────────┴──────────┴──────────┴──────────┴──────────┴──────────┘
 project      287.28 TB  279.33 TB   661.26 K    1.00 PB  500.00 TB          0    28.73 %         ok    ignored
└─────────────────────────────────────────────────────────────────────────────────────────────────────────────────┘
```

- We're working to organize a separated quota at EOS for users and productions

External storage

- There are already propositions to store some of our data at the external storages, for example, at Minsk
- In order to start doing this we must build a data catalog to know where and which our data are stored to avoid creating dark data
- The natural way to manage data on the external source is to do it through the data management service

Authentication

- At the moment we’re still using X.509 certificates, issued by RDIG (Russian Data Intensive Grid)
- To reduce the dependency on external services we have no control over, our own certification authority was prepared at LIT, with all necessary background machinery
- JINR CA issues user and host certificates immediately after being requested via web UI
- Work of enabling host certificates for many machines for power users (integration with IPDB) is ongoing
- User identification is done by JINR SSO, so at the end of the story anyone wishing to use computing resources as user of the SPD collaboration will have to get a user account at JINR
- Now in test operation
- We are preparing a transition from X.509 certificates to JSON Web Tokens, user identification will still be based on the JINR SSO

Manpower

- CRIC 0/0.25 FTE
- WMS/WFMS/Harvester/Pilot 0.5/2 FTE, Artem Petrosyan
- Rucio 1/1 FTE till July 2024, Alexey Konak
- FTS 0/0.25 FTE
- Production system 0.5/1 FTE, Artem Petrosyan
- Production manager 0/0.5 FTE
- Monitoring 0/1 FTE

Summary

- 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 minimum bias) is now ongoing
  - In September we finally managed to run a test task in containers with SpdRoot through the system
- We have several external participants willing to participate in the software development and data processing: PNPI, SPbSU and INP BSU, their CE are already connected to the distributed computing infrastructure of the experiment
- Our next steps will lay in the field of data and workflow management: continued Rucio integration, FTS deployment, Production system development
- We need to make efforts in the following directions: data and metadata catalog preparation, data types definition, data lifetime definition, data management service configuration
- We expect to have a naming convention for our data types to enable automation and jobs and data flows through the production system

Thank you!