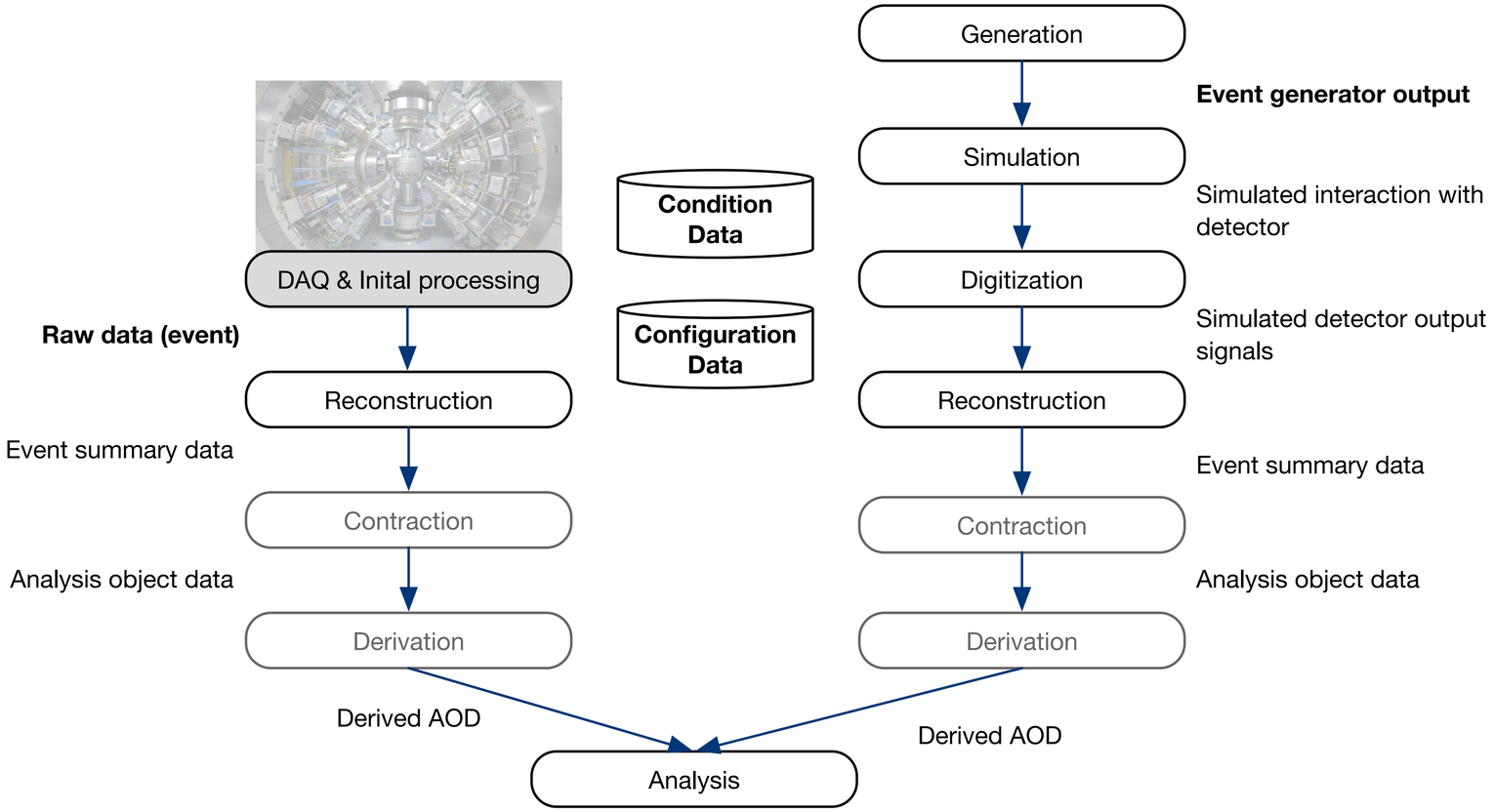
The SPD software and computing project

Danila Oleynik, Conference on High Energy Physics, 03.10.2025

Data processing in HEP

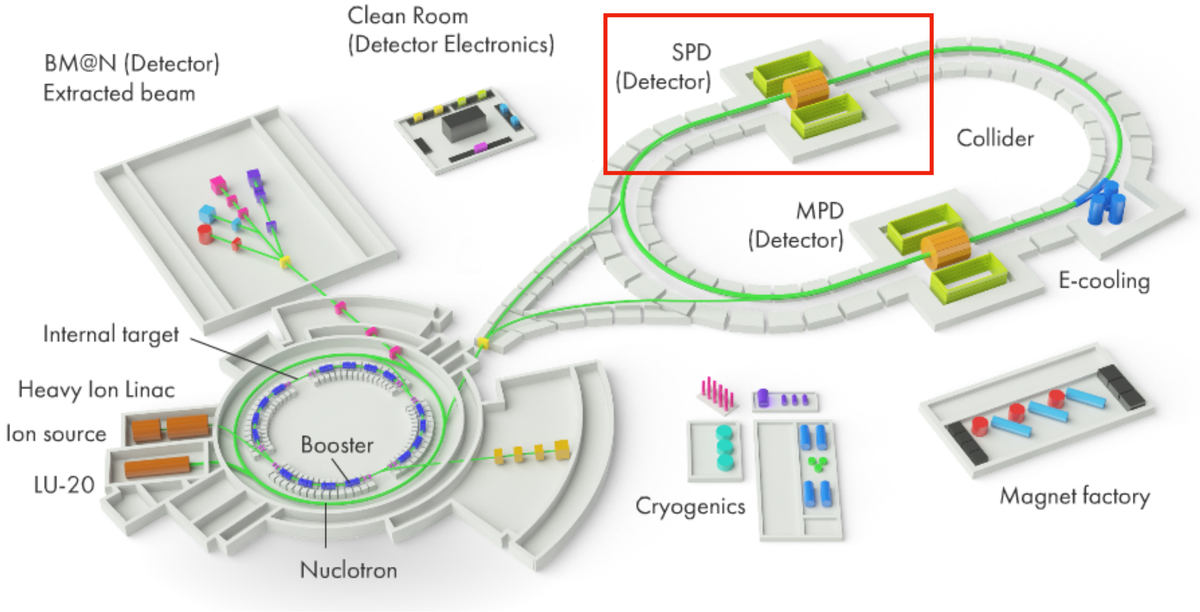
High-throughput computing (HTC) involves running many independent tasks that require a large amount of computing power.

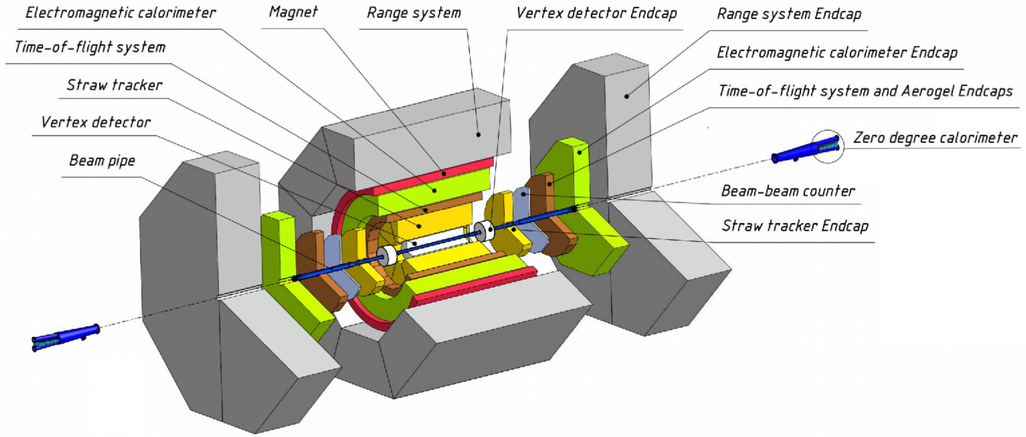
- Event is the least data unit in HEP.
  - Each event may be processed independently
- As reconstruction as simulation – are multistep workflows
  - Each step produces own data type, which correspond to different representation of events
  - So size of event will be different in different data type
- Why we need different types?
  - Some types of processing, like raw data, quite expensive or unique, producing of other types is resource consuming, another types good for long term storage but not optimal for final analysis because of redundancy



SPD Spin Physics Detector

Study of the nucleon spin structure and spin-related phenomena in polarized *p-p*, *d-d* and *p-d* collisions





- SPD - a universal facility for comprehensive study of gluon content in proton and deuteron

SPD Collaboration

Participants (MOUs signed)

Joint Institute for Nuclear Research (JINR)
Dubna, Russia — A. Guskov, V. Ladygin

National Research Nuclear University MEPhI
Moscow, Russia — P. Teterin

Budker Institute of Nuclear Physics of the Russian Academy of Sciences
Novosibirsk, Russia — A. Barnyakov

Institute for Nuclear Problems of BSU
Minsk, Belarus — A. Lobko

Tomsk State University
Tomsk, Russia — S.Filimonov, I. Shreyber

Skobeltsyn Institute of Nuclear Physics of the Moscow State University
Moscow, Russia — A. Berezhnoy

Samara National Research University
Samara, Russia — V. Saleev

Petersburg Nuclear Physics Institute (NRC KI – PNPI)
Gatchina, Russia — V. Kim

Peter the Great St. Petersburg Polytechnic University (SPbPU)
St. Petersburg, Russia — Ya. Berdnikov

National Science Laboratory
Yerevan, Armenia — N. Ivanov

University of Belgrade
Belgrade, Serbia — D. Maletic

Lebedev Physical Institute of the Russian Academy of Sciences
Moscow, Russia — V. Andreev

Institute of Nuclear Physics
Almaty, Kazakhstan — S. Sakhiyev

Belgorod National Research University
Belgorod, Russia — A. Kubankin

Institute for Nuclear Research RAS
Troitsk, Russia — E. Usenko

St. Petersburg State University
St. Petersburg, Russia — V. Vechernin

iThemba LABS
Somerset West, South Africa — N. Stodart

Participants

National Research Center Kurchatov Institute
Moscow, Russia — I. Alexeev

Higher Institute of Technologies and Applied Sciences (InSTEC)
Havana, Cuba — K. Shtejer

Cairo University
Cairo, Egypt — R. El-Kholy

Higher School of Econimics
Moscow, Russia — F. Ratnikov

Tsinghua University
Beijing, China — Y. Wang

Institute of applied physics of the NAS of Belarus
Minsk, Belarus — R. Shulyakovsky

CTEPP, UNAB
Santiago, Chile — S. Kuleshov

SAPHIR
Santiago, Chile — S. Kuleshov

China Institute of Atomic Energy
Beijing, China — X. Li

Francisk Skorina Gomel State University
Gomel, Belarus — V. Andreev

B.I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus
Minsk, Belarus — Yu. Kulchitsky

National University of Science and Technology
Moscow, Russia — M. Gorshenkov

Shandong University
Shandong, P.R.China — J. Zhang

Institute for High Energy Physics
Protvino, Russia — S. Golovnya

SPD as a data source

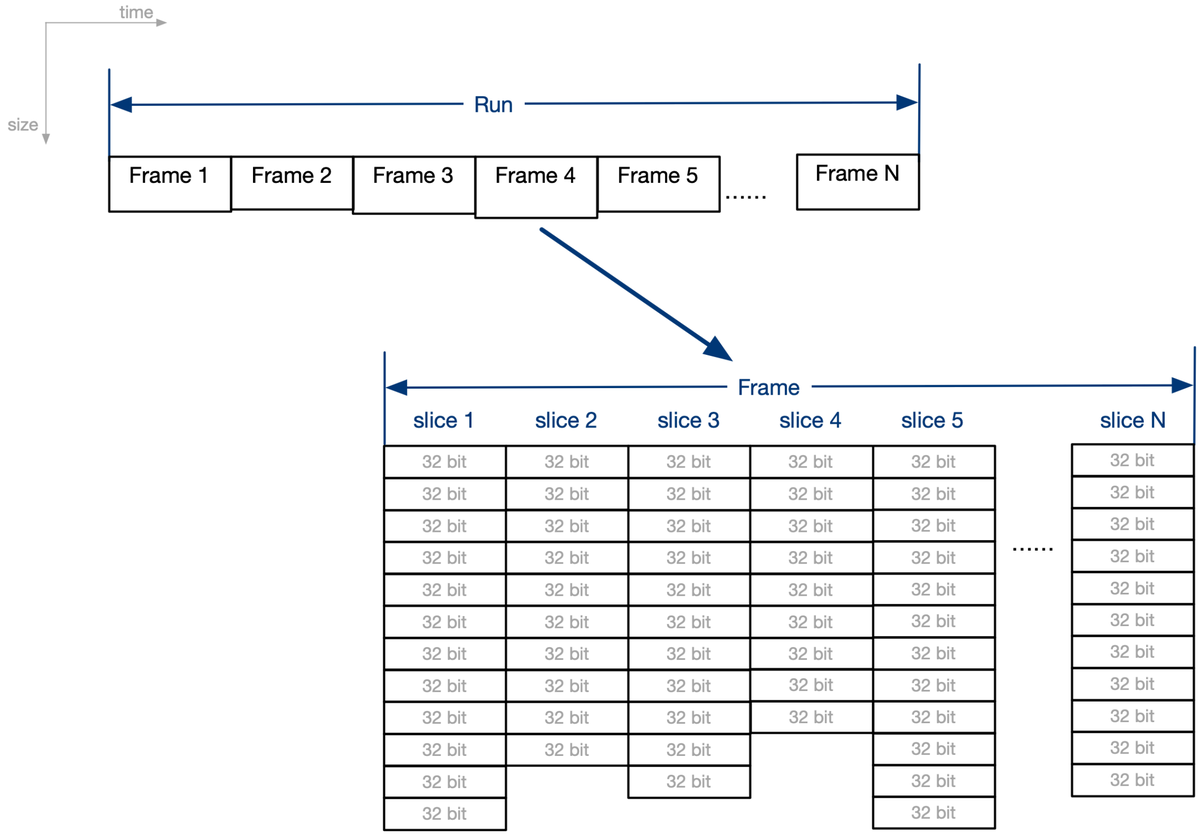
The SPD detector is a medium scale setup in size, but a large scale one in data rate!

- Bunch crossing every 76,3 ns = crossing rate 13 MHz
  - ~ 3 MHz event rate (at 10^{32} $cm^{-2}s^{-1}$ design luminosity)
- **20 GB/s** (or **200 PB/year** "raw" data, **~3*10^{13}** events/year)
  - “Only" **~1,5*10^{12}** events/year are interesting for detailed study (~10PB/year) of data
  - Selection of physics signal requires momentum and vertex reconstruction → no **simple trigger** is possible
- Comparable amount of simulated data

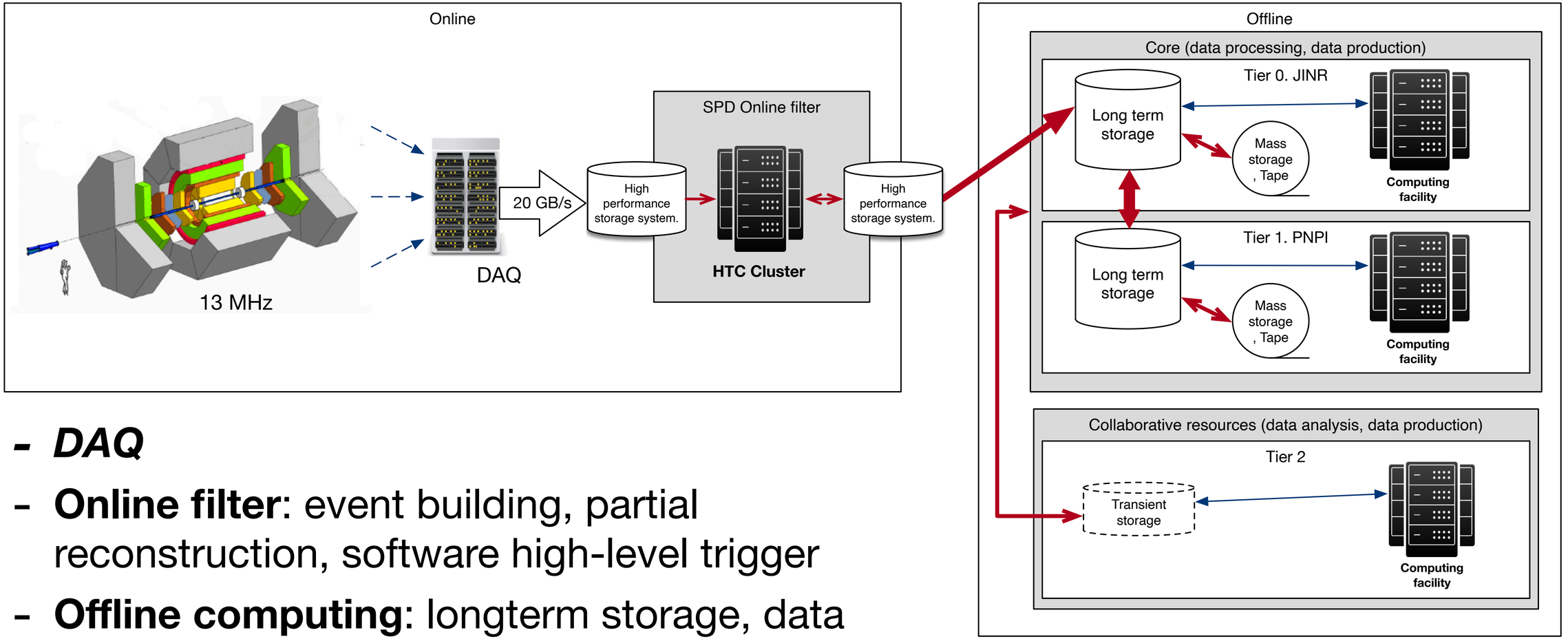
Free run DAQ and (raw) data format

No trigger = No "classical" events at the start

- Free run DAQ, means that the output of the system will not be a dataset of raw events, but a set of signals from detectors organized in time slices
- Primary data unit: time slice (1 μs — 8.3 ms)
  Time slices combined in time frames (up to 549 s, 16 GB max, < 160 MB to fullfil 20 GB/s limit)
- Intermediate units — time chunks of 0.1-0.2 s (2-4 GB or ~10^5- 10^6 events) are being discussed now
- Every time slices will contain signals from a few to many collisions (events)
- Event building have to unscramble events from a series of time slices



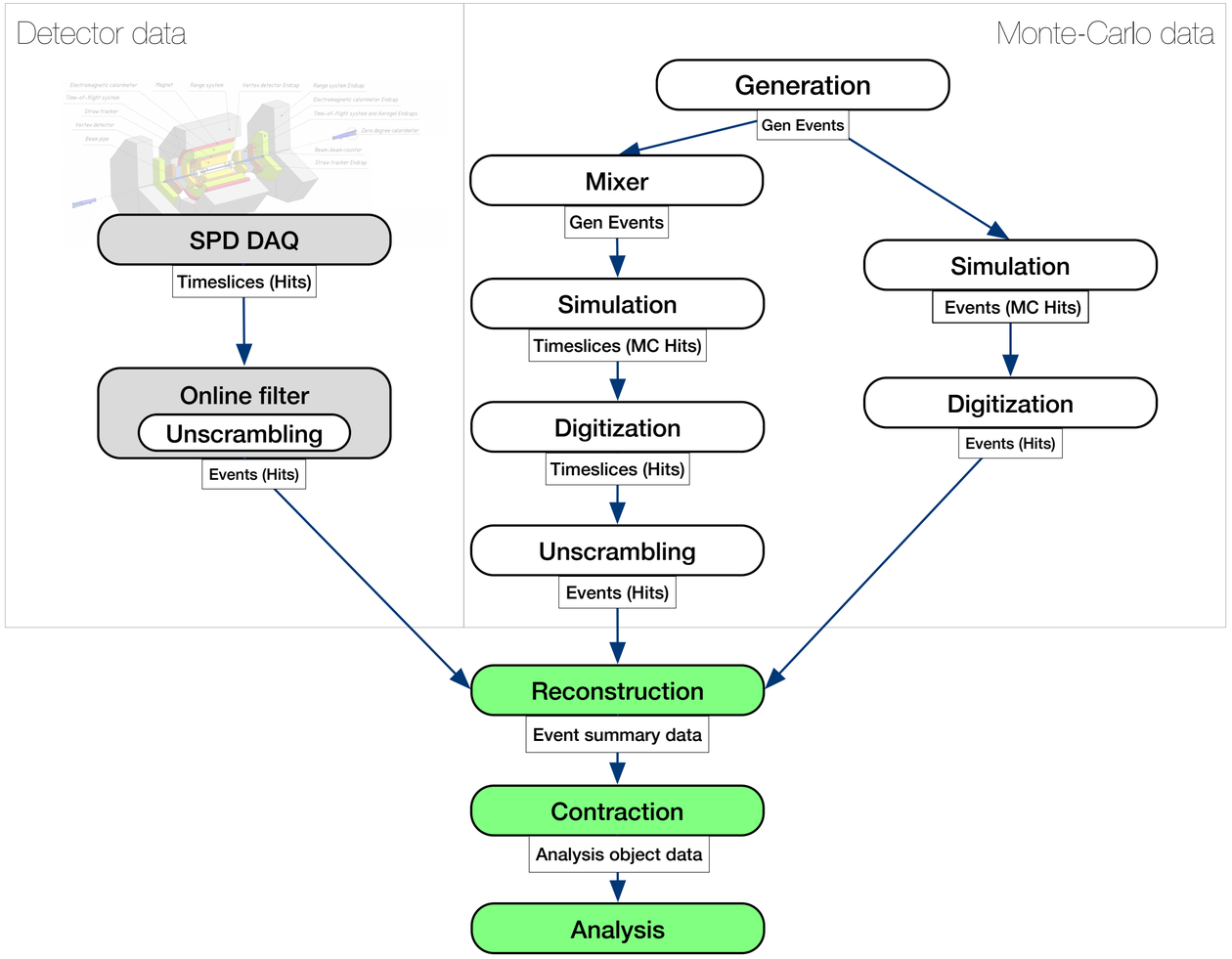
Data flow



- ***DAQ***
- **Online filter**: event building, partial reconstruction, software high-level trigger
- **Offline computing**: longterm storage, data production, data processing and analysis

SPD data processing

- Free run DAQ increased complexity of data production (Monte-Carlo)
- Raw data in timeslices should be simulated along with events simulation



“Mixer”

- We assume, that time slice contains information from a few collisions (events).
- For simulation of time slice we take an information from a few generated events
  - “Mixer" prepares a mix of generated events which may be produced by different generators
  - There is no significant changes expected for Geant4 itself, but Sampo framework should allow to read set of generated events
  - Obviously, output of simulation will represent new data type - ‘Timeslice’

Events unscrambling

- For each time slice
  - Reconstruct tracks and associate them with vertices
  - Determine bunch crossing time for each vertex
  - Associate ECAL and RS hits with each vertex (by timestamp)
  - Attach unassociated tracker hits in a selected time window according to bunch crossing time
  - Attach raw data from other subdetectors according to bunch crossing time
  - Call the block of information associated with each vertex an event
  - Store reconstructed events

SPD Online filter

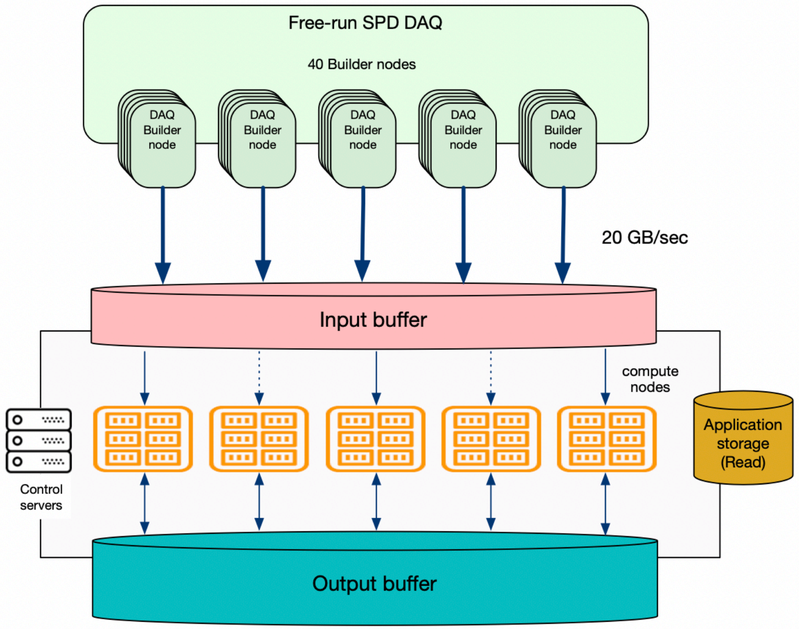
SPD Online Filter

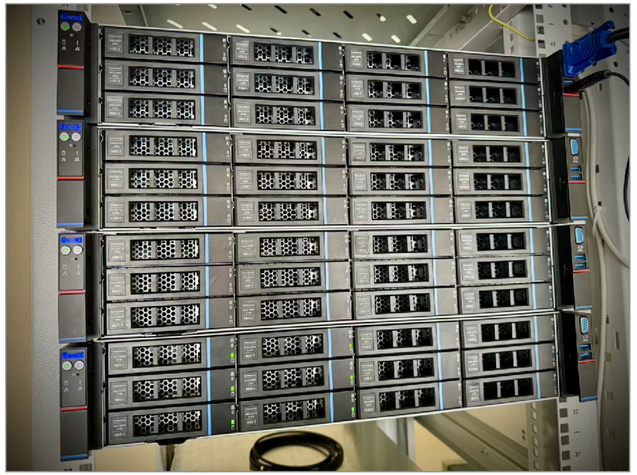
Online filter is the first stage in data processing chain for SPD Experiment (right after DAQ)

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.





Middleware functionality

Data management;

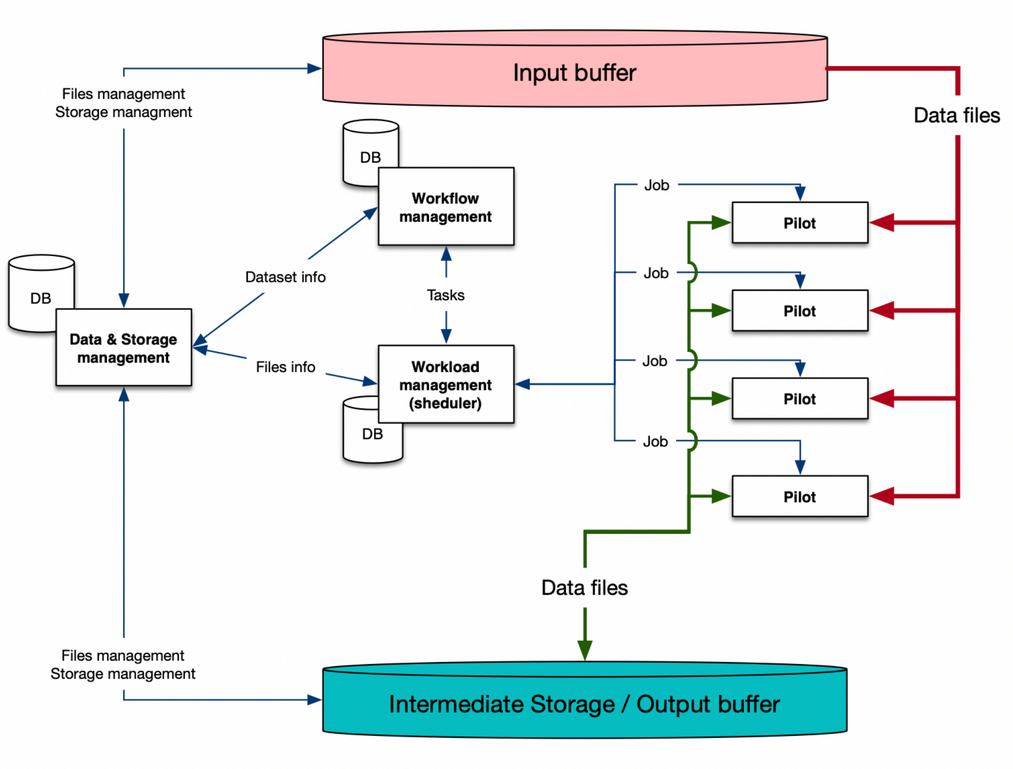
- *Support of data life-cycle and storage usage;*

Workflow management;

- *Definition of processing chains;*
- *Realisation of processing chains as set of computations tasks;*
- *Management of tasks execution;*

Workload management:

- *Generation of required number of processing jobs for performing of task;*
- *Control of jobs executions through pilots, which works on compute nodes;*



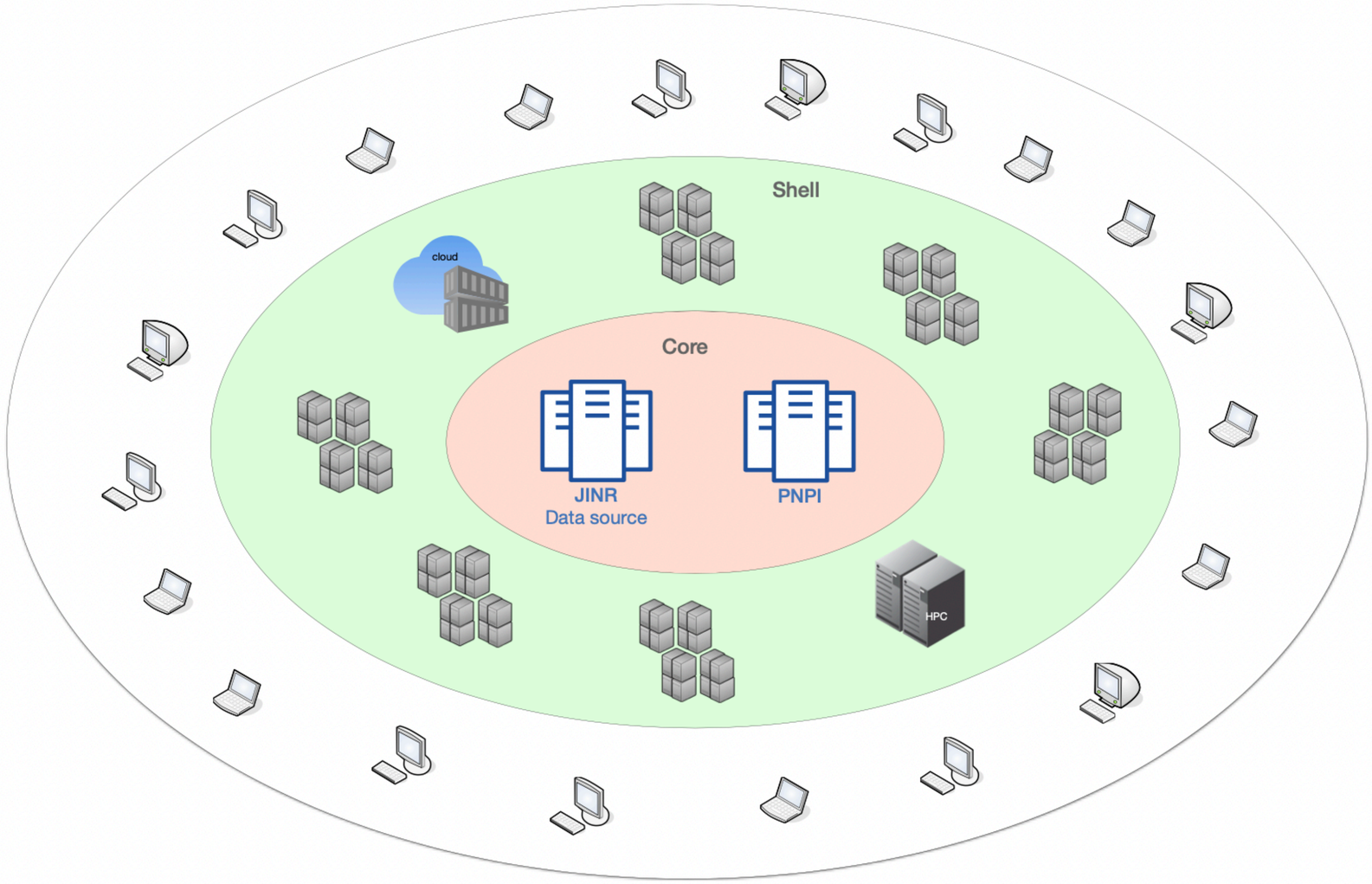
SPD Offline computing

Expected data volumes

Data volumes != required storage capacity ;-)

- **Preparation for the experiment.**
  - Monte Carlo simulation from 2024 to 2028 will provide 2 PB per year.
  - Total per stage: **10 PB**.
- **Stage I: running at low luminosity of the NICA collider.**
  - Monte Carlo simulation and real data taking from 2028 to 2030 will provide 4 PB per year. Reprocessing: 2 PB per year.
  - Total per stage: **18 PB.**
- **Upgrade of the setup for operation at high luminosity.**
  - Monte Carlo simulation from 2031 to 2032 will provide 2 PB per year. Reprocessing: 2 PB per year.
  - Total per stage: **8 PB**.
- **Stage II: running at maximum design luminosity of the NICA collider**.
  - Monte Carlo simulation and real data taking from 2033 to 2036 will provide 20 PB per year. Reprocessing: 10 PB per year.
  - Total per stage: **120 PB.**
- Total for all stages: **156 PB.**

SPD Offline computing system



- **Core sites (JINR, PNPI)** – data long term storage, main data processing and producing
- **Shell sites** – data analysis, data producing

Distributed data processing system

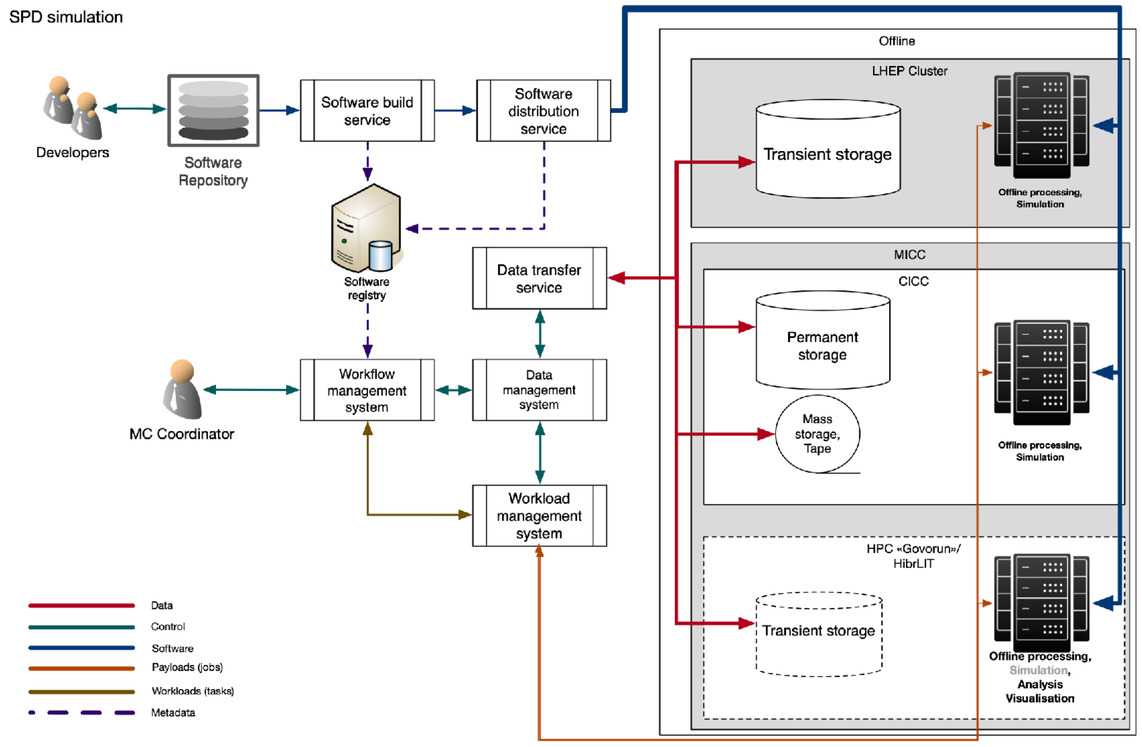
- Authentication and authorization
- Workflow and workload management
- Data organization and management
- Data transfers
- Software distribution
- Common catalog of computing and storage resources (information system)

Most of basic components are already available from LHC experiments:

- a lot of work required to adapt of components to work together for particular experiment

SPD Distributed computing middleware

- **INDIGO 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 WMS** —is a data-driven workload management system capable of operating at massive data processing scale, designed to have the flexibility to adapt to emerging computing technologies in processing, storage, networking and distributed computing middleware
- **Rucio DMS** — responsible for data management, including data catalog, data integrity and data lifetime management strategies
- **FTS DTS** — enables massive data transfers



SPD distributed computing in production

PanDA monitor Dash Tasks Jobs Errors Users Sites Harvester My BigPanDA Job by ID Enter... Help Danila

PanDA sites 6d5c69db8ef4 | 07:12:02,Refresh

Site attribute summary	
copytool (1)	xrdcp (5)
country (1)	Russian Federation (5)
gocname (4)	JINR (2) PNPI (1) SPbSU (1) SSAU (1)
harvester (1)	ST-221-126 (5)
status (2)	brokeroff (2) online (3)
system (1)	
tier (2)	T1 (2) T2 (3)
type (1)	production (5)

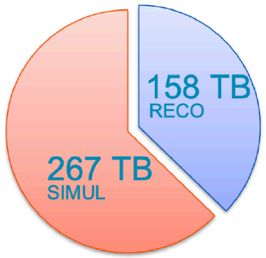
5 PanDA queues

Show All entries Search:

Cloud	GOC site name	Tier	Queue name	Status	Type	Workflow	System	Copytools	Associated Harvester	Max RSS [GB]	Max time [hours]	Max input size [GB]	Last comment
RU	JINR	T1	JINR_SPD_PROD	online	production	---	---	xrdcp	ST-221-126	1.9	72	---	no active blacklisting rules defined
RU	JINR	T1	JINR_SPD_TEST	brokeroff	production	---	---	xrdcp	ST-221-126	1.9	72	---	Only for testing
RU	PNPI	T2	PNPI_SPD_PROD	online	production	---	---	xrdcp	ST-221-126	1.9	---	---	no active blacklisting rules defined
RU	SPbSU	T2	SPbSU_SPD_PROD	brokeroff	production	---	---	xrdcp	ST-221-126	1.9	---	---	Setup & validation
RU	SSAU	T2	SSAU_SPD_PROD	online	production	---	---	xrdcp	ST-221-126	1.9	---	---	no active blacklisting rules defined

Showing 1 to 5 of 5 entries Previous 1 Next

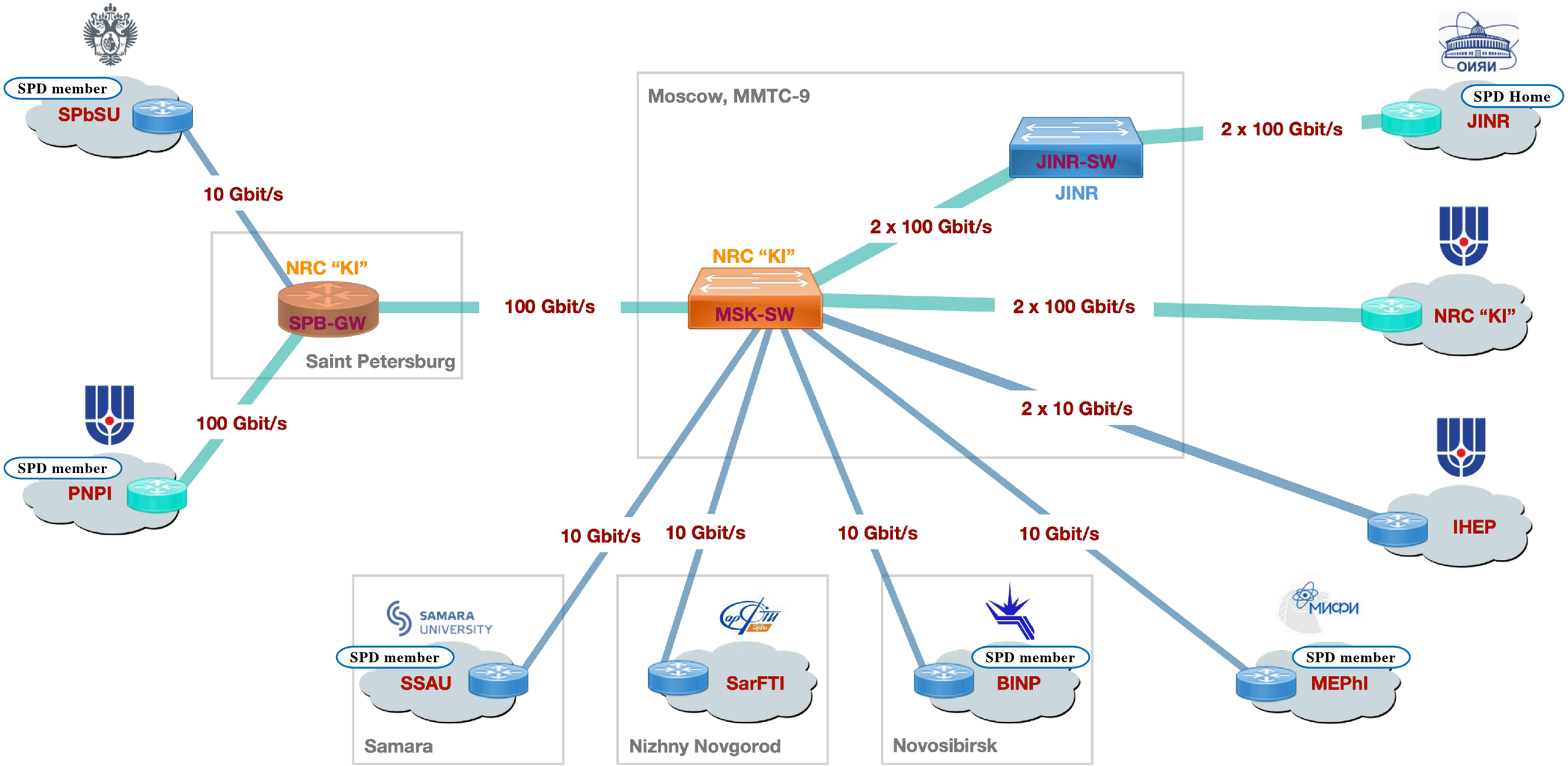
- Successfully processed about 300k jobs across 55 tasks



- Total output datasets volume – more than 425 TB

Distributed infrastructure

Russian scientific backbone



Required SPD computing resources

	CPU (cores)	Disk storage (PB)	Tape storage (PB)
SPD Online filter (stage 1)	3000	2	
Offline computing (stage 1)	20000	5	6 per year
SPD Online filter (stage 2)	6000	4	
Offline computing (stage 2)	60000	15	30 per year

- Tier-0 at JINR will provide about 25-30% of all computing resources
- Tier-1 at PNPI is going to contribute about 25%
- The rest should be distributed between the participating institutes

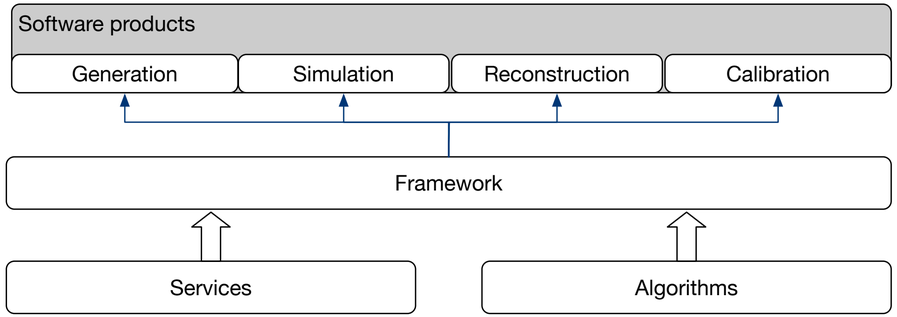
Information systems & databases

- Collaboration management data
- Detector hardware database and mapping (detector elements, cabling etc)
- Data production requests (including MC input configurations)
- Offline DB: Geometry versions, Calib&Align, Magnetic field
- Event index - is the set of special information systems which allows to store and navigate across all produced events
  - In simple words Event index allows identify dataset or even file where particular event is stored.
  - Quite important system as only you start to use hundreds of thousands files
- Condition database - stores data which is not related with event production itself, but status of environment during data tacking
- Configuration database - stores detector hardware setup and other hardware related information
- A PostgreSQL RDBMS is considered as a database platform

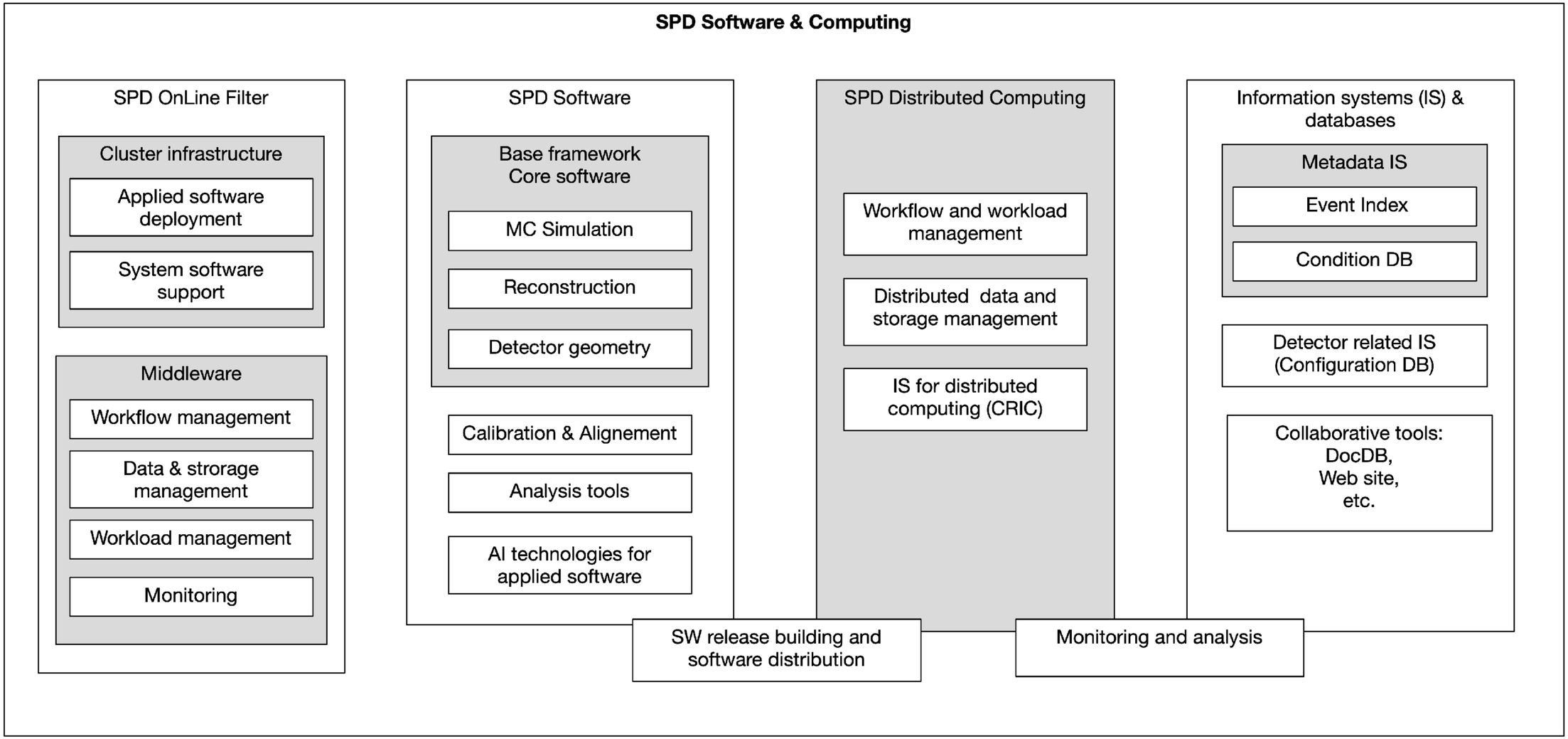
SPD Software

Information processing of physics data of SPD Experiment

- A Gaudi-based software framework is being developed:
  - Geometry description: GeoModel
  - Generators: Pythia8, FTF
    - In general any other generator can be used while it supports hepmc3 output format.
  - Simulation: Geant4
  - Reconstruction: MdcHough track finding, ACTS (Kalman filter) for track fitting, Kfparticle for vertex reconstruction, own algorithms for other subsystems
- Current simulation and performance studies are done by another framework SpdRoot, based on FairRoot software
  - Does not fit well for massive data processing



SPD Software and computing project



Summary

- The SPD Software and computing project is quite wide by the different IT aspects:
  - A set of existed services and frameworks allows significantly decrease requirements in manpower and decrease time gap to move systems to production
  - We face a lot of work for applied software framework and algorithms and with SPD Online Filter machinery
  - Computing part gradually grows, most of new development related with adoption for particular experiment requirements
- Laboratory of Information technology provides full support of the project not only from infrastructure part but also from methodology and expertise as in software so in computing

Thank you!


С САМОГО НАЧАЛА
У МЕНЯ БЫЛА КАКАЯ-ТО
ТАКТИКА
И Я ЕЁ ПРИДЕРЖИВАЛСЯ

Some basic definitions

reminder… from 1964

- **DATA** – a representation of facts or ideas in a formalized version, capable of being communicated or manipulated in some process.
- **INFORMATION –** in automatic data processing the meaning that a human assigns to data by means of the known conventions used in its representation.
- **DATA PROCESSING** – the execution of a systematic sequence of operations, performed with data, e.g. handling, merging, sorting, computing.
  - Note: Where data processing is performed in order to increase the value or significance (from a certain point of view) of the information conveyed by the data, it may be called **INFORMATION PROCESSING.**

THE TERMINOLOGY WORK OF IFIP (International Federation for Information Processing) AND ICC (International Computing Centre)
I. H. GOULD and G. C. TOOTILL

A few more definitions

- **SPD Software** - a set of activities related with development, support and evolution of applied software for **information processing of physics data of SPD Experiment**.
- **SPD Computing** - a set of activities devoted to setup and operation of distributed computing environment for **data processing of SPD Experiment**.
  - *Usually we call it - distributed data processing*
- **Infrastructure** - a set of computing and storage resources provided by collaboration members for shared usage in distributed data processing