Data processing in HEP experiments

Oleynik Danila. Senior researcher, MLIT JINR.

A long times ago….



Tycho Brahe

1546 - 1601

Danish astronomer who designed and constructed greatly improved astronomical instruments. This increased the accuracy of measurements.

18 years of astronomy observations, precision measurement and systematisation of collected data by Tycho Brahe were complexly analysed by Johannes Kepler, which finally allowed to formulate "Kepler's laws of planetary motion" and eventually paving the way for Isaac Newton theory of universal gravitation.



Johannes Kepler

1570 - 1630

Mathematician who used Tycho's observations of the heavens to validate the Copernican model.

Data Acquisition

Data Analysis

Not so long time ago: 1964

Basic definitions from another science

- **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

- **Software** - a set of activities related with development, support and evolution of ***applied*** software for **information processing of an experiment data**.
- **Computing** - a set of activities devoted to setup and operation of computational environment for **data processing of an experiment**.
  - *Usually we call it - distributed data processing*
- **Infrastructure** - a set of compute, storage and network resources required for data processing

Simple, Complicated, Complex

Simple Systems

A simple system is simple – it has a minimal set of subsystems and is understood easily.

Complicated Systems

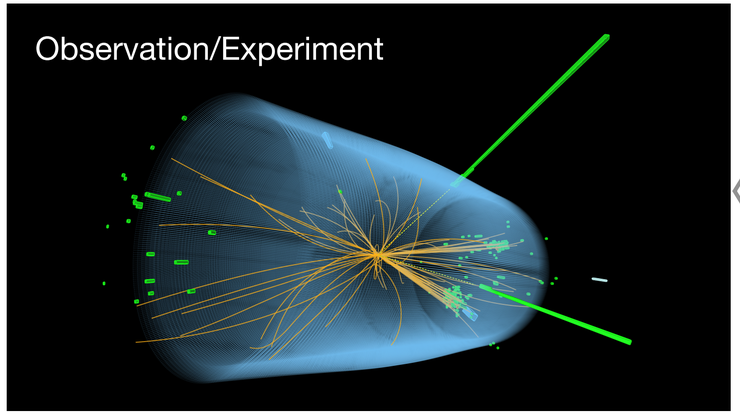
These consist of subsystems, usually tightly coupled, interacting in a definite pattern. Despite the surface similarity to complex systems; complicated systems exhibit predictable behaviours.

The challenge lies in understanding the intricate links between the multiple subsystems. Examples of complicated processes and systems include timepieces, food recipes and *algorithms*.

Complex Systems

Complex systems are also made up of multiple interacting subsystems. However, they differ because it is impossible to fully understand the system by studying its components. No algorithm or recipe can tell you how to solve a complex problem – it is the realm of unknown unknowns.

Where is data processing in physics research?



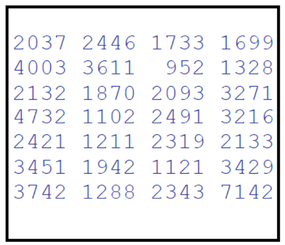
Computation / Analysis

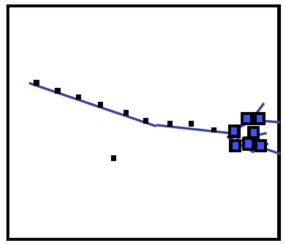
Theory

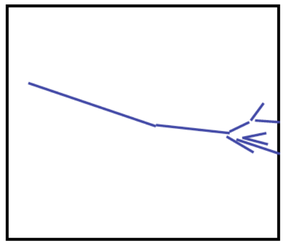
$$\mathcal{S}[\varphi_i] = \int \mathcal{L}[\varphi_i(s)]\, d^n s,$$

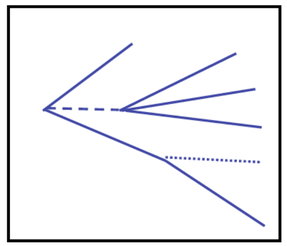
- Simulation of the experiment & Construction of experimental setup
- Data tacking
- Preparation of physics data
- Analysis of physics data & obtaining of physics results

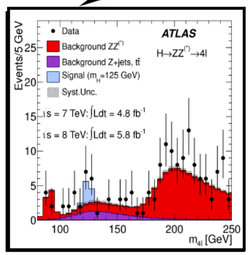
Base streams of data processing











RAW DATA

CONVERT TO PHYSICS QUANTITIES

DETECTOR RESPONSE APPLY CALIBRATION, ALIGNMENT,

INTERACTION WITH DETECTOR MATERIAL PATTERN, RECOGNITION, PARTICLE IDENTIFICATION

FRAGMENTATION, DECAY PHYSICS ANALYSIS

BASIC PHYSICS

RESULTS

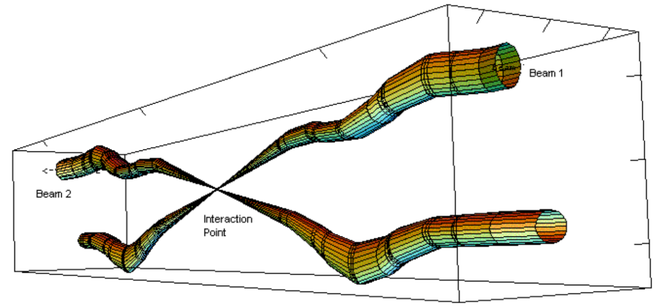
RECONSTRUCTION

ANALYSIS

SIMULATION (MONTE-CARLO)

Data granularity

What is “Event”?



Relative beam sizes around IP1 (Atlas) in collision

In physics, and in particular relativity, an event is the instantaneous physical situation or occurrence associated with a point in spacetime (wikipedia).

For HEP:

- **Interaction:** Two particles interact and somehow produce a change, be it in energy, trajectory or identity.
- **Collision:** Two particles are made to approach one another and actually undergo an interaction.
- **Beam crossing:** Two beam bunches pass through one another in the center of a detector.
- **Event:** During a beam crossing, one pair or multiple pairs of particles undergo a collision. In an event, often one collision dominates the signature in the detector.

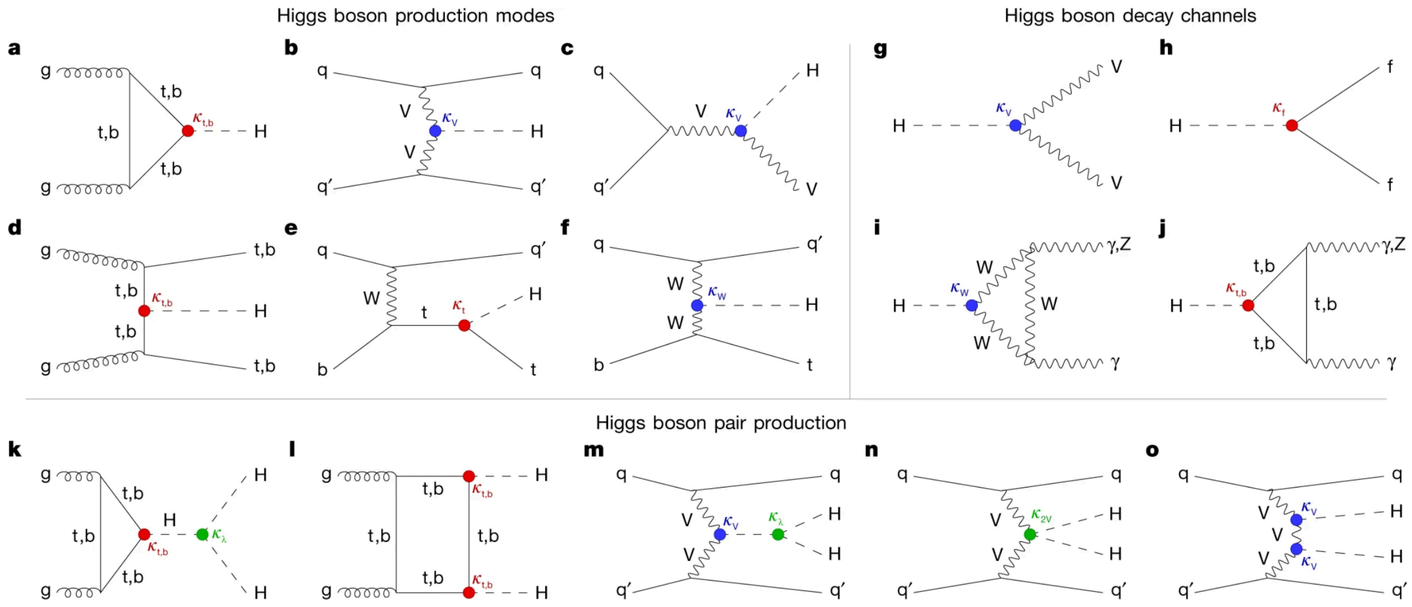
Each event may be processed independently, so **for HEP Data processing - event is the least data unit**.

By knowing the size of event, complicity of event processing and expected number of event for processing, you can estimate requirements for computing system.

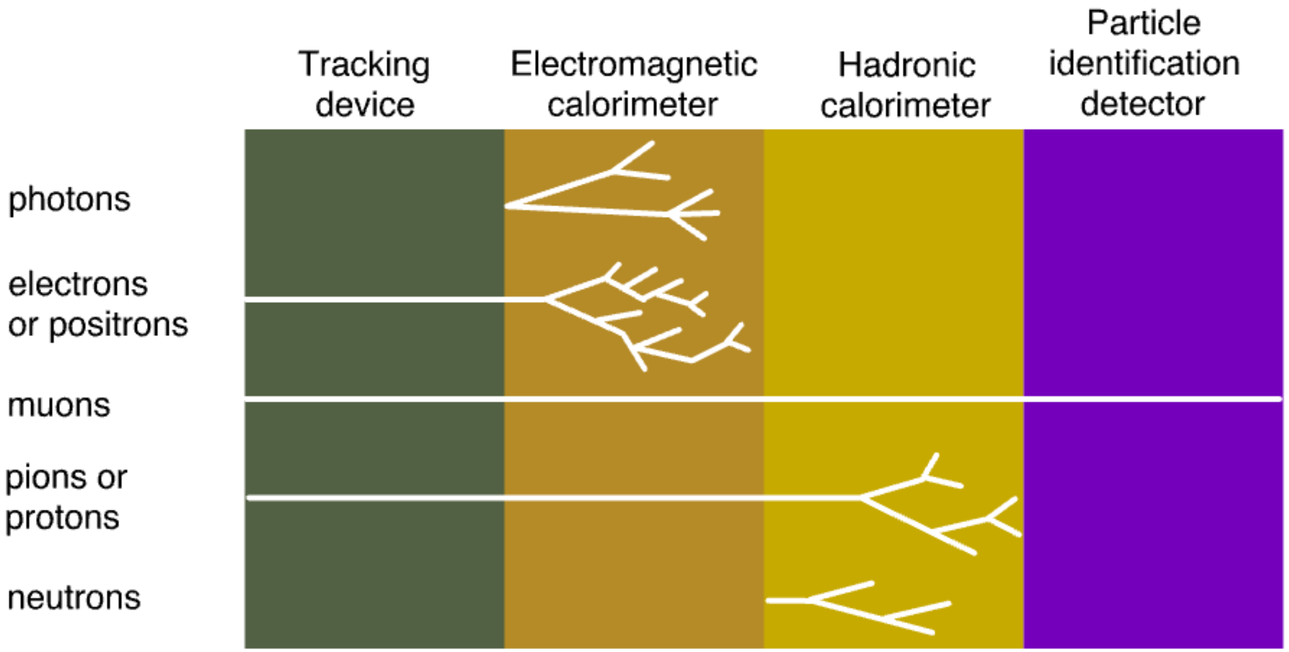
Why we care about “Event rate”

Higgs boson example

- Frequency of producing a Higgs boson that has decayed to two Z bosons each of which has decayed to an electron-positron pair is extremely rare: once in 10^{13} = 10 000 000 000 000

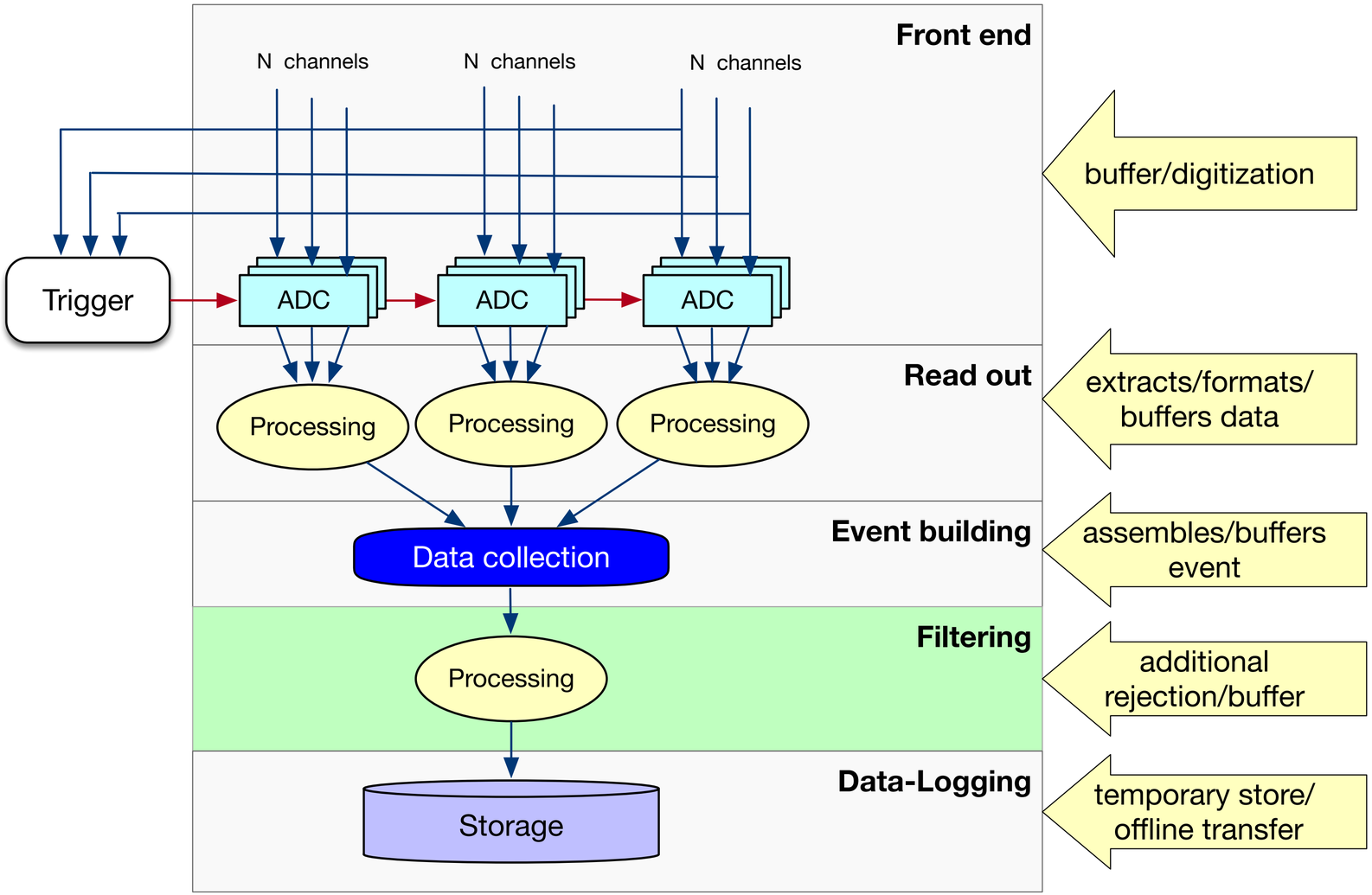


Generic HEP detector



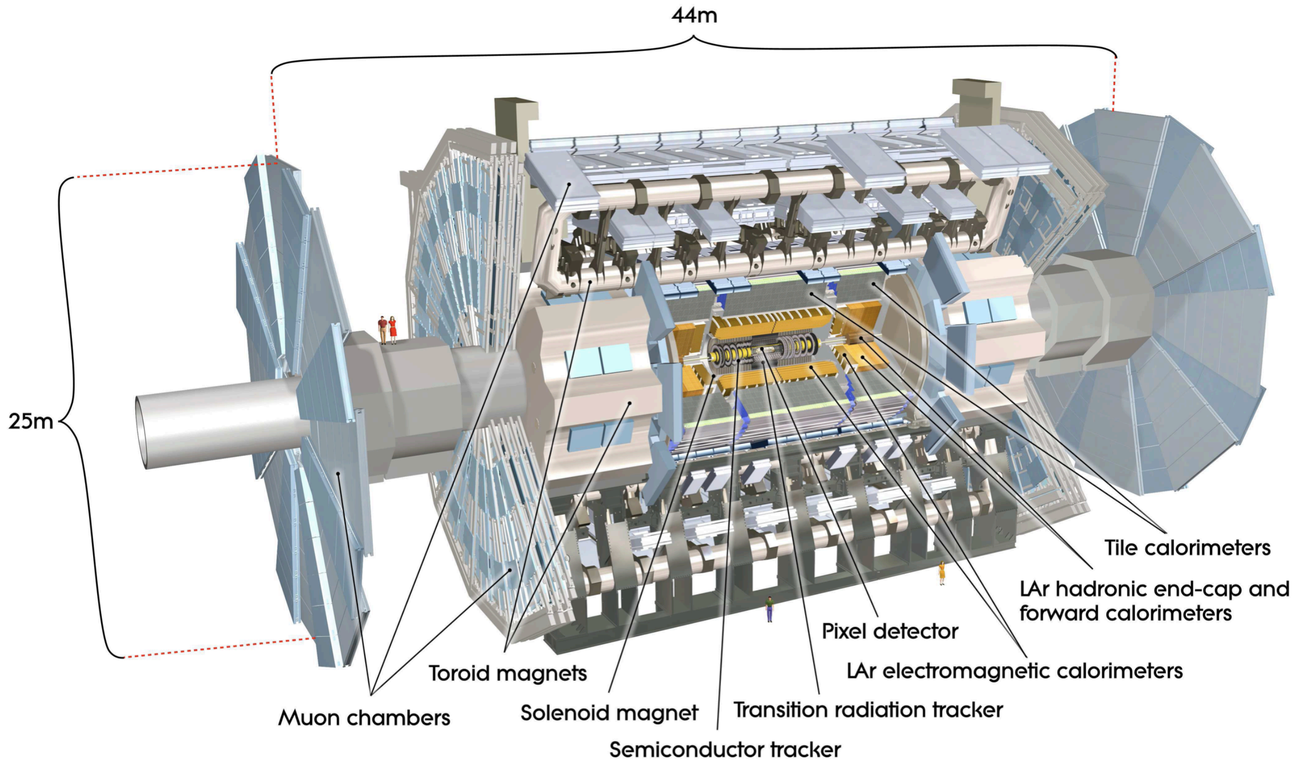
- **Tracking device** – detects and reveals the path of a particle
- **Calorimeter** – stops, absorbs and measures the energy of a particle
- **Particle identification detector** – identifies the type of particle using various techniques

“Classical" DAQ



An example: ATLAS Detector at LHC

Largest volume particle detector ever constructed



LHC (Large Hadron Collider)

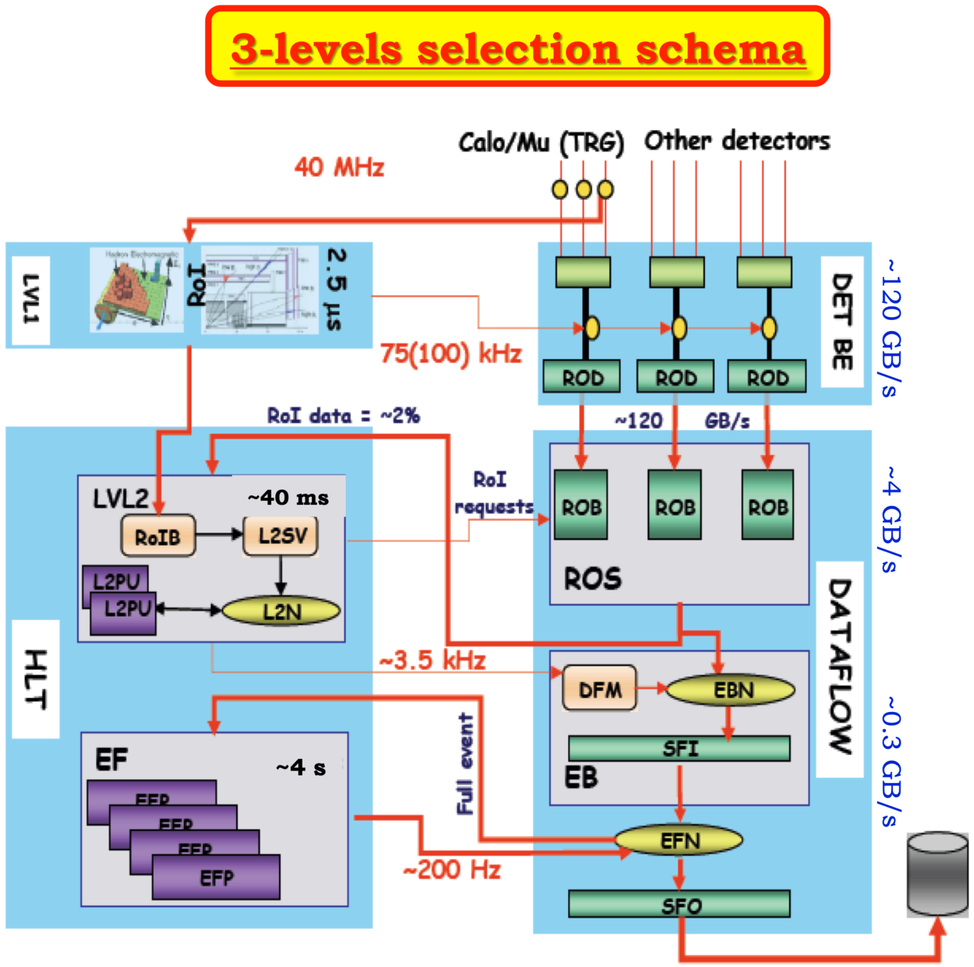
- proton-proton collisions at 14 TeV (design)
- 25 ns bunch-crossing interval, **40 MHz bunch-crossing rate**
- **Design luminosity 10^{34} cm^{-2} s^{-1}** (~23 interactions per crossing)

ATLAS Detector

- Size: 25m x 44m
- Barrel toroid length: 26 m
- Overall weight: 7 000 tonnes
- **~ 100 million electronic channels**
- ~ 3 000 km of cables

An example: ATLAS Detector at LHC

ATLAS TDAQ



Level 1 (LVL1)

- Custom electronics
- Reduced granularity (calorimeter, muon)
- Latency 2.5 μs
- Select rate 75 kHz

Level 2 (LVL2)

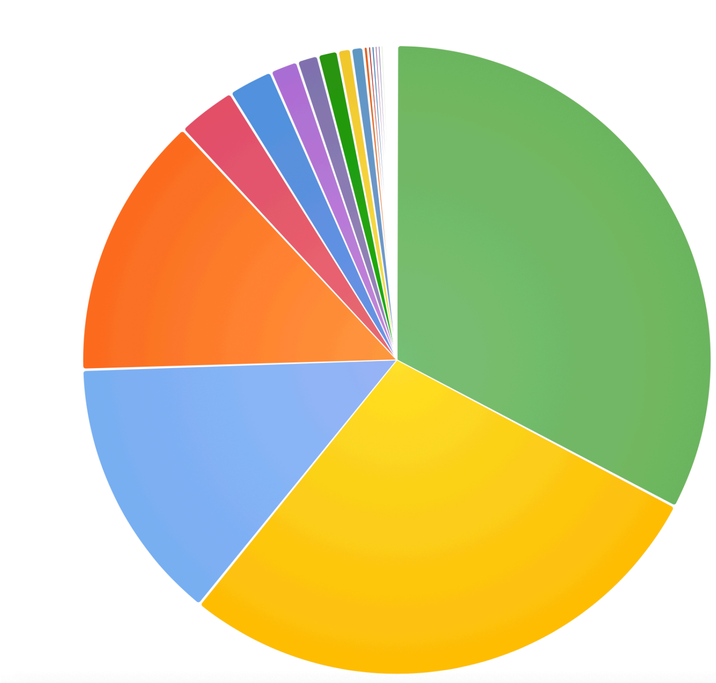
- Software based, specialized algorithms
- Full granularity
- Region Of Interest (2% of data transferred)
- Event processing time: ~40 ms
- Event Filter (EF)
  - Software based, offline algorithms
  - Full event available
  - Seeded by LVL2
  - Event processing time: ~4s

An example: ATLAS Detector at LHC

Data volume

Recorded	per event	per year
raw data	1.6 Mbytes	3 200 Tbytes
reconstructed data	1 Mbytes	2 000 Tbytes
physics data	0.1 Mbytes	200 Tbytes

Logical Size	Disk Size	Primary Size	Disk Files	Primary Files
785 PB	426 PB	323 PB	987871720	933224511
Number of Eve...	**Tape Size**	**Secondary Size**	**Tape Files**	**Secondary Files**
6213703842233	814 PB	103 PB	416870561	54647209
	Total Size		**Total Files**	
	1.24 EB		1404742281	



	Value	Percent
RAW	413 PB	33%
AOD	354 PB	29%
HITS	172 PB	14%
DAOD	171 PB	14%
EVNT	37 PB	3%
RDO	28 PB	2%
DRAW	17 PB	1%
DESD	13 PB	1%
no_name	12 PB	1%
NTUP	8 PB	1%
ESD	7 PB	1%
log	2 PB	0%
TXT	1 PB	0%
group	1 PB	0%
HIST	950 TB	0%
user	920 TB	0%
HEPMC	870 TB	0%
HITS_FILT	288 TB	0%
D2AOD	170 TB	0%
DPD	43 TB	0%
LOGARC	26 TB	0%
TAG	17 TB	0%
ESD_FILTERED	9 TB	0%
CBNT	7 TB	0%
cond	5 TB	0%

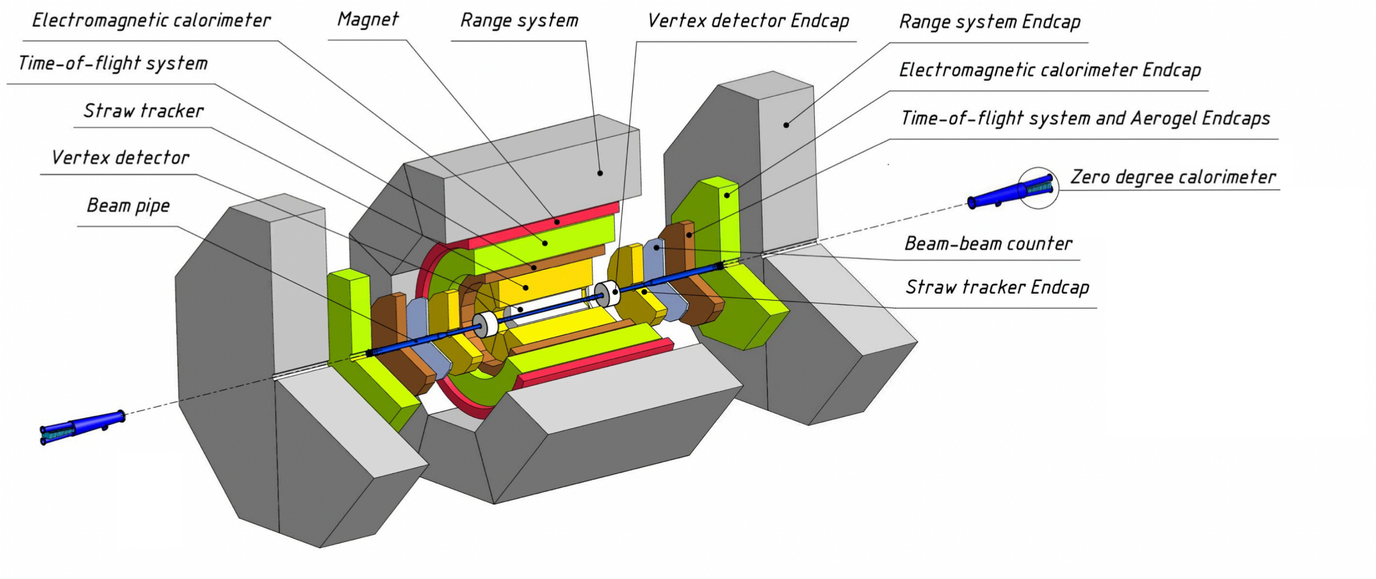
ATLAS Experiment

- 38 countries / 181 Research centres
- 5500 members (3000 researches)

Physics Analysis Groups	Performance Groups, Simulation and Fora
Standard Model	Tracking CP
B physics and light states	Electron and Photon
Top physics	Muon CP
Higgs physics	Tau
Higgs and Diboson searches	Jet and Etmiss
Supersymmetry searches	Flavour Tagging
Exotics searches	Simulation
Heavy ions physics	Machine Learning Forum
Upgrade physics studies	
Physics Modelling (PMG)	

SPD Spin Physics Detector

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



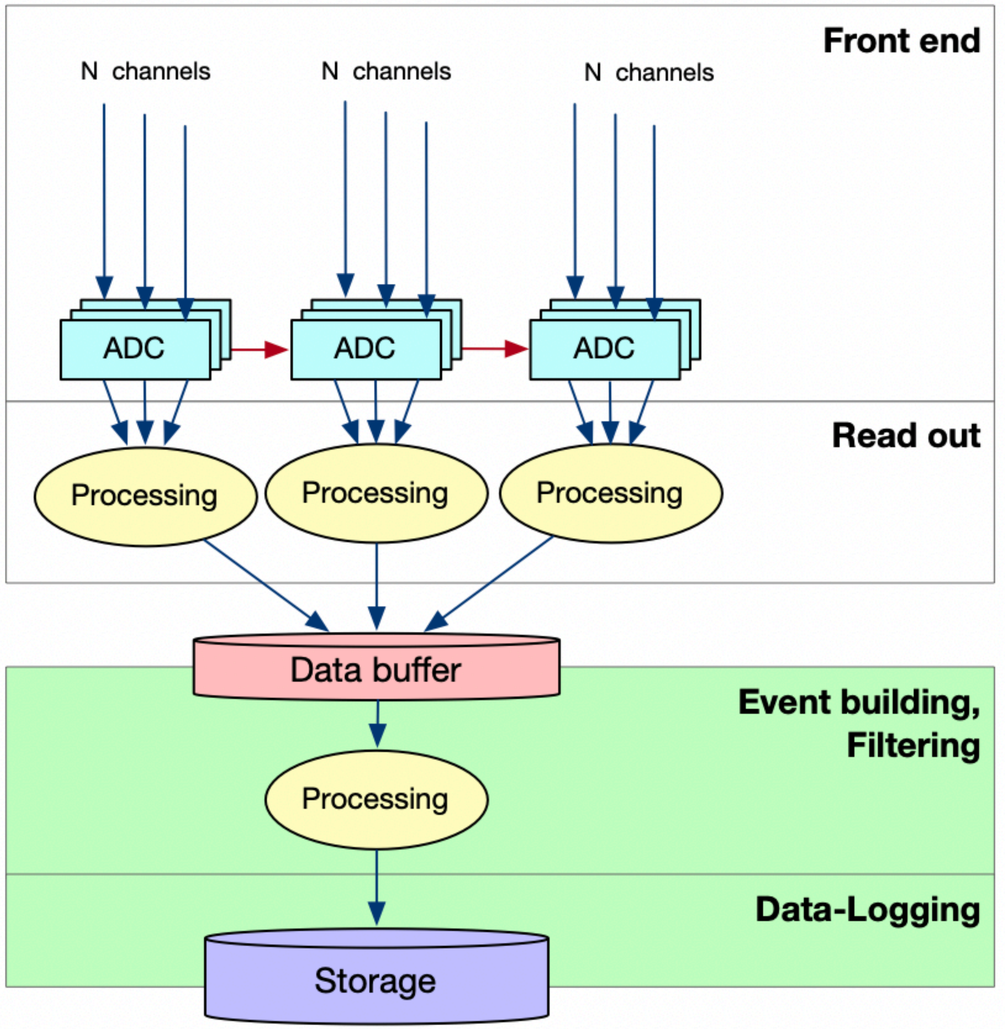
SPD - a universal facility for comprehensive study of gluon content in proton and deuteron

SPD as data source

- Bunch crossing every 80 ns = crossing rate 12.5 MHz
  - ~ 3 MHz event rate (at 10^{32} $cm^{-2}s^{-1}$ design luminosity)
  - 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
- SPD is the triggerless detector setup
  - **20 GB/s** (or **200 PB/year** "raw" data, **~3*10^{13}** events/year)
  - Comparable amount of simulated data

Trigerless DAQ

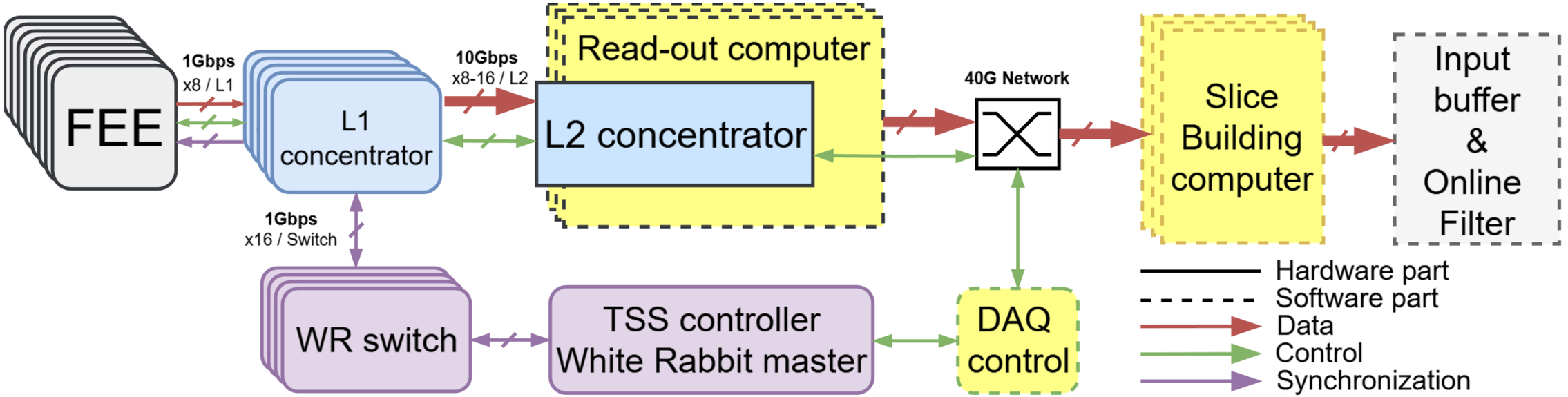
- Triggerless DAQ, means that the output of the system will not be a dataset of raw events, but a set of signals from sub-detectors organized in time slices
- To get data in proper format for future processing (reconstruction) and filtering of ‘boring’ events special computing facility named "Online Filter" in progress



SPD DAQ Structure

- L1 concentrator
- L2 concentrator
- TSS – synchronization system
- Computer Read-out and Slice Building system.

SPD Stage	FEE Outputs	L1 (8 ports)	L2 (16 ports)	Read-out PC	Slice Builders
First	~1292	>165	>12	>6-12	>5
Second	~5800	>730	>50	>25-50	>40

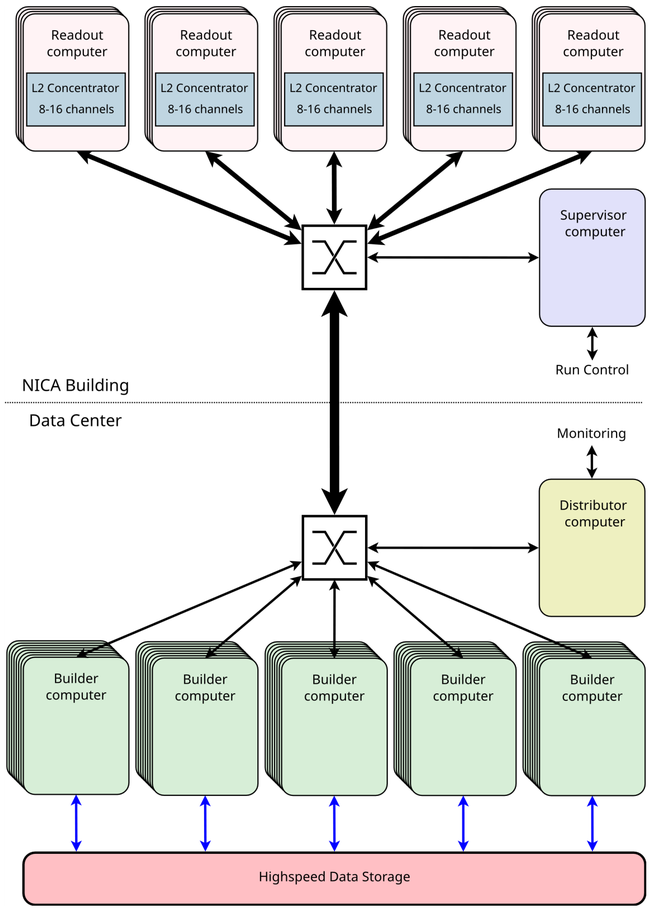


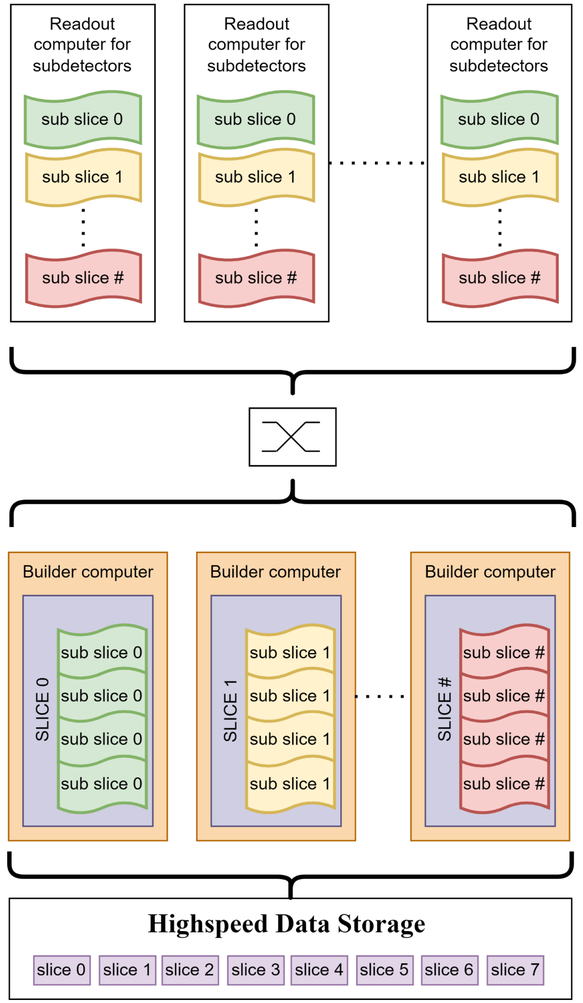
SPD initial data organisation

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.

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 a monitoring data

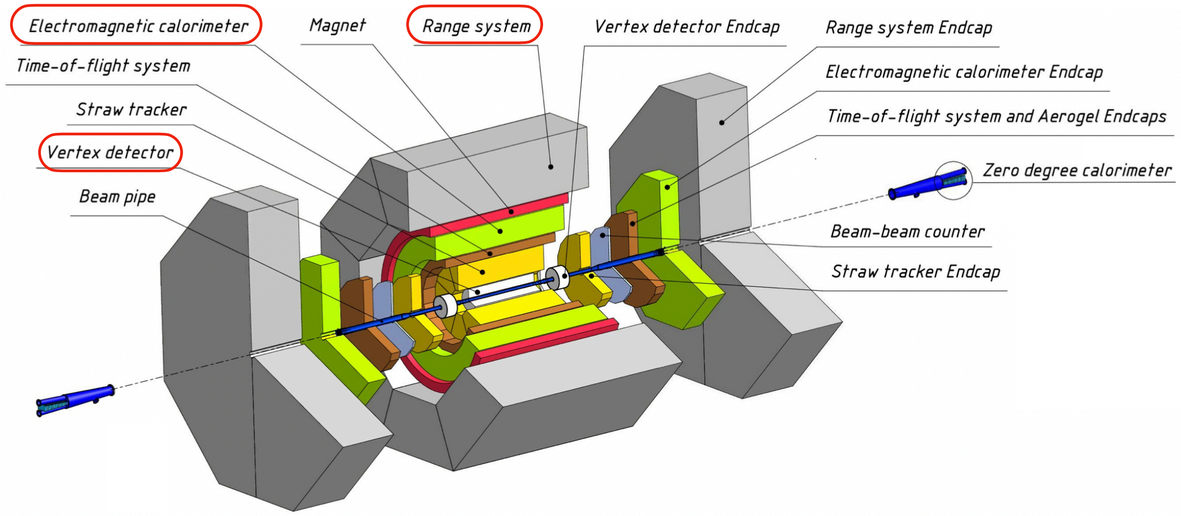




Event unscrambling

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



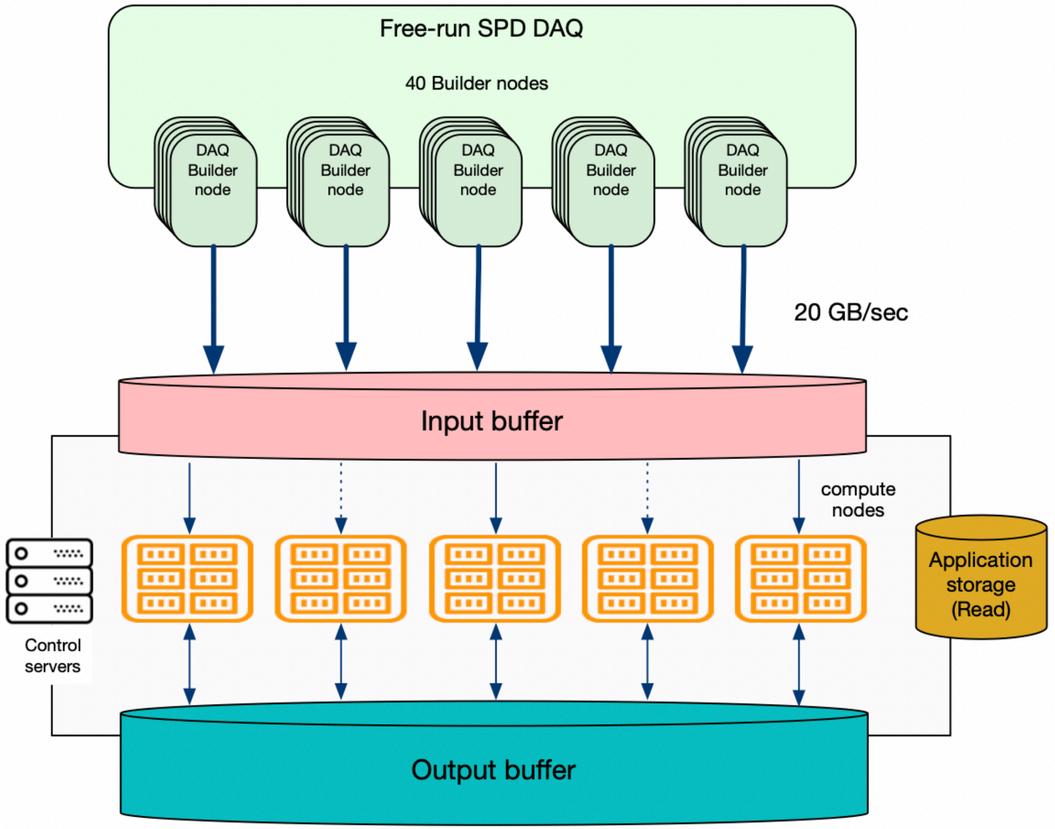
High Throughput Computing

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

- HPC tasks are characterized as needing large amounts of computing power for short periods of time, whereas HTC tasks also require large amounts of computing, but for much longer times (months and years, rather than hours and days). HPC environments are often measured in terms of FLOPS.
- The HTC community, however, is not concerned about operations per second, but rather operations per month or per year. Therefore, the HTC field is more interested in how many jobs can be completed over a long period of time instead of how fast an individual job can complete.
- As an alternative definition, the European Grid Infrastructure defines HTC as “a computing paradigm that focuses on the efficient execution of a large number of loosely-coupled tasks”, while HPC systems tend to focus on tightly coupled parallel jobs, and as such they must execute within a particular site with low-latency interconnects. Conversely, HTC systems are independent, sequential jobs that can be individually scheduled on many different computing resources across multiple administrative boundaries. HTC systems achieve this using various grid computing technologies and techniques.

Online filter infrastructure

- SPD Online Filter is a high performance computing system for high throughput processing
  - High speed (parallel) storage system for input data written by DAQ.
  - Compute cluster with two types of units: multi-CPU and hybrid multi CPU + Neural network accelerators (GPU, FPGA etc.) because we are going to use AI…
  - A set of dedicated servers for middleware which will manage processing workflow, monitoring and other service needs.
  - Buffer for intermediate output and for data prepared for transfer to long-term storage and future processing.



Middleware basic functionality

Data management;

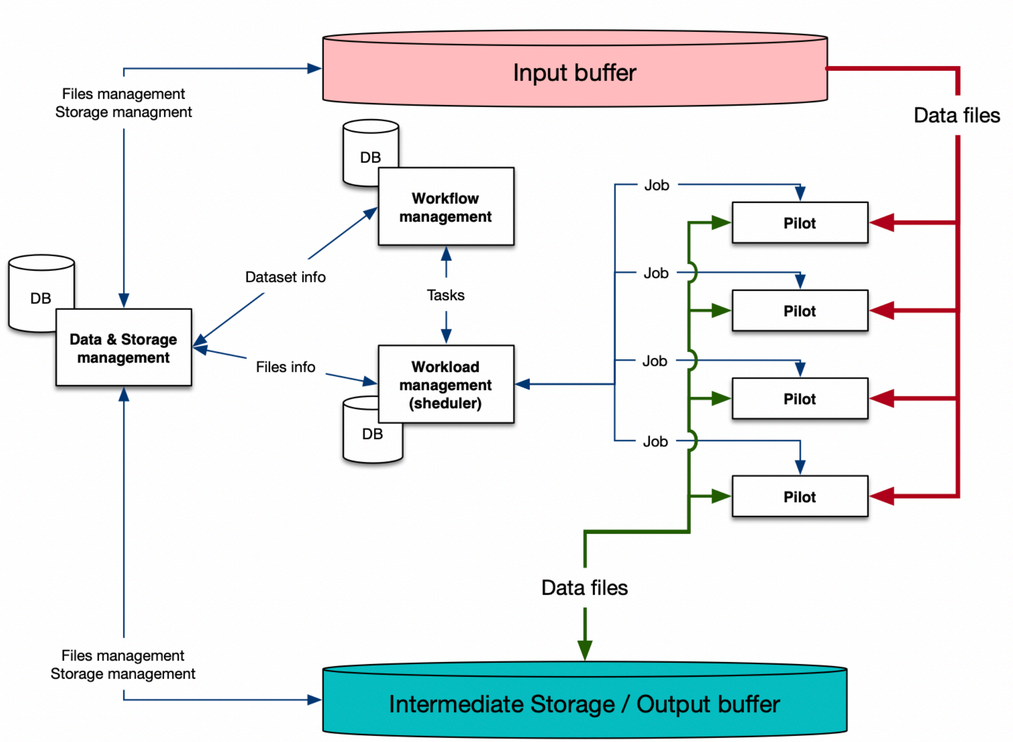
- *Support of data life-cycle and storage usage;*

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 Expected data volumes

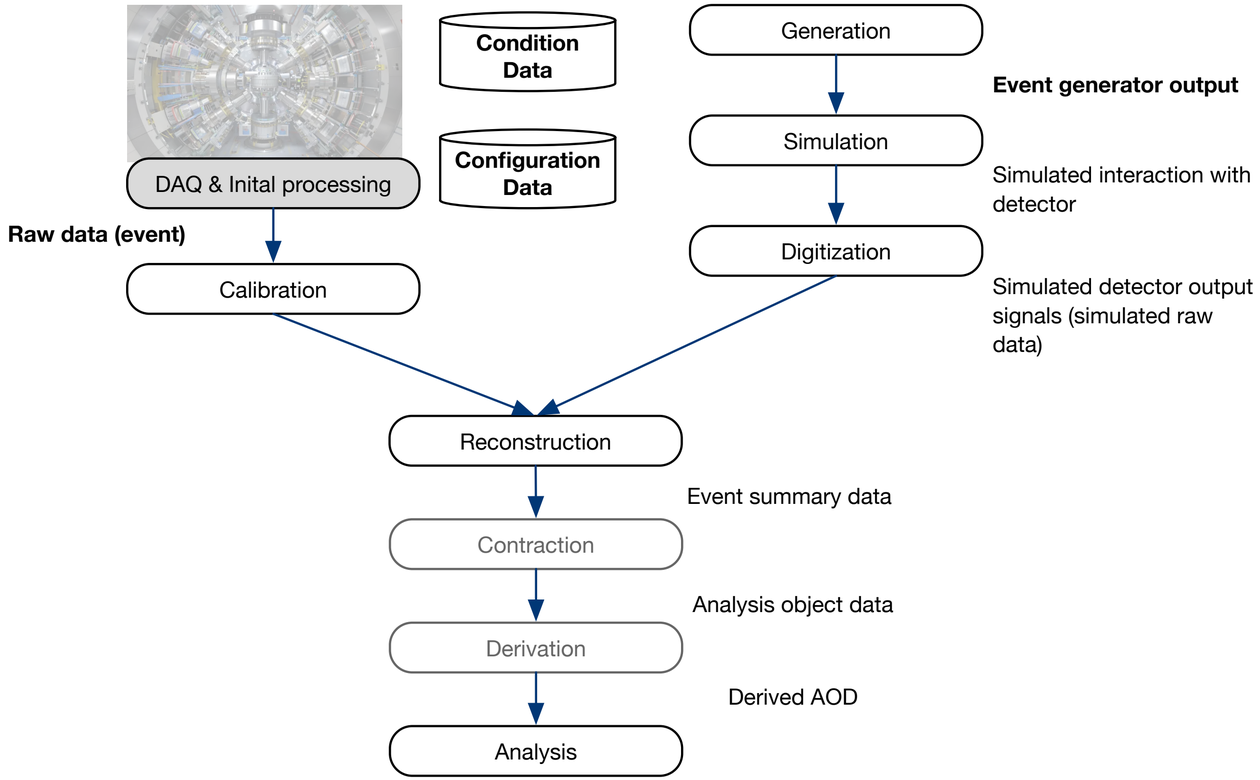
Data volumes != required storage capacity ;-)

- **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.**

Offline processing

Reconstruction, Simulation

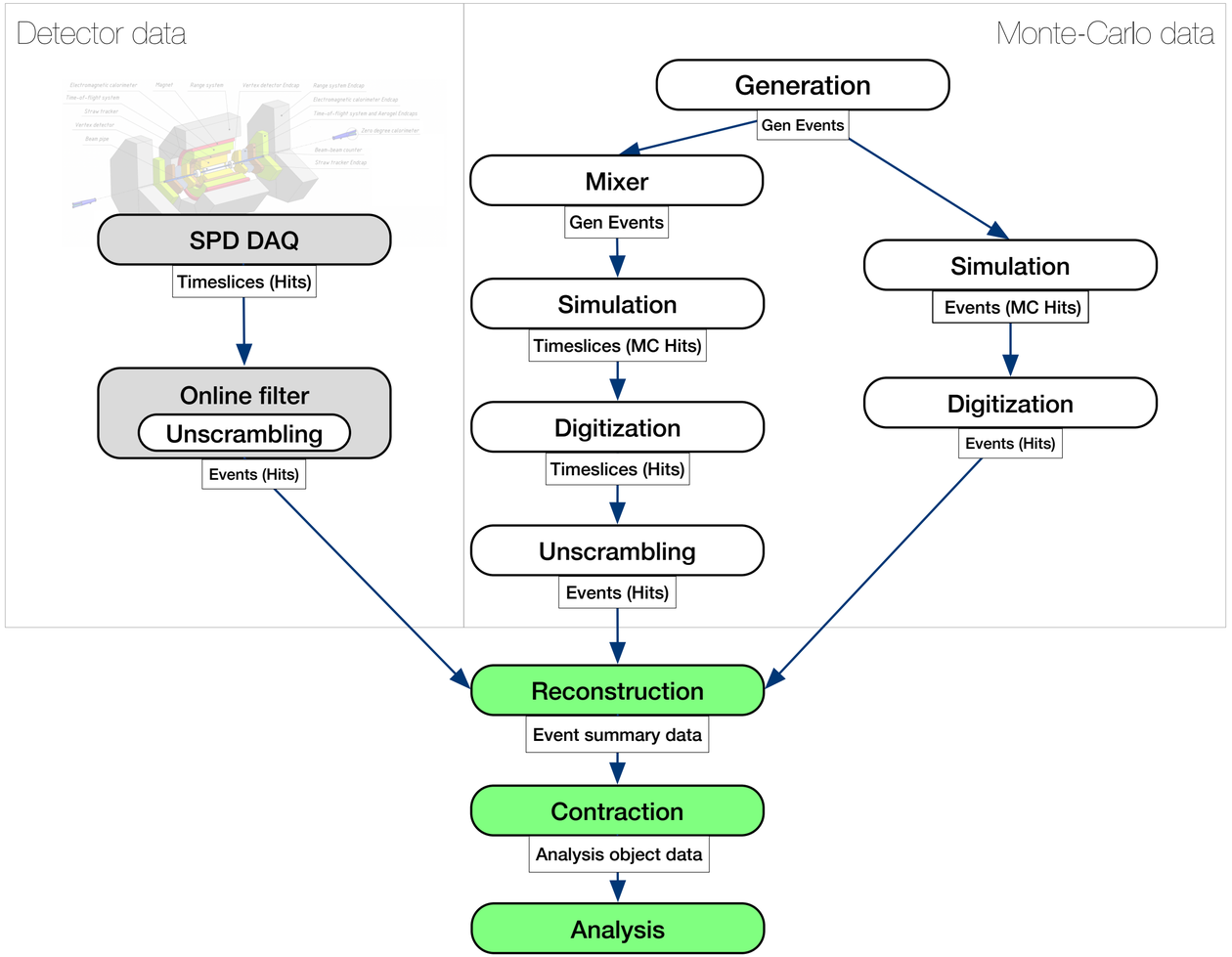
- Collected data is not ready for analysis yet
  - Events contain raw or partially reconstructed data
  - Calibration and alignment is not applied yet
- Simulation - pure computation processing (will start mach early than apparatus will be ready)
  - Will require significant amount of computing resources
- Even after processing amount of data are still huge, so event selection became a special processing step.



- Another type of computing facility required for routine offline processing – distributed data processing system (aka grid)

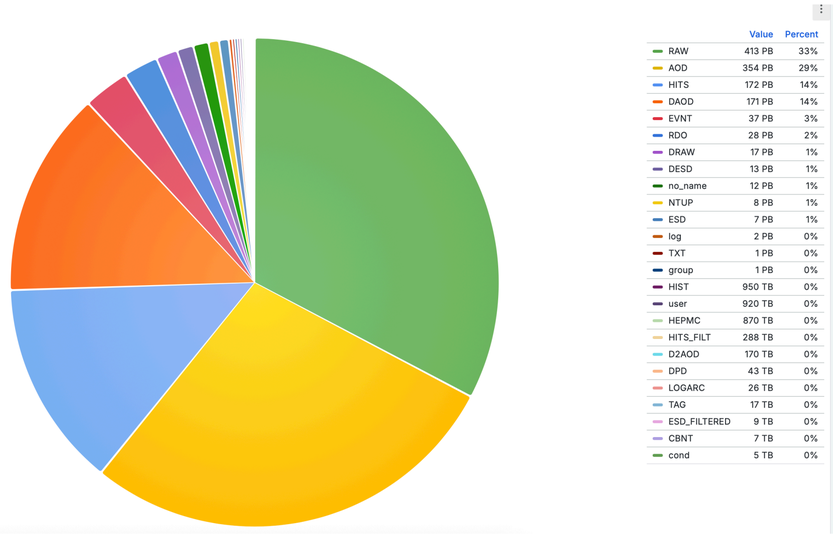
SPD data processing

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



Processing steps and data types

- 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



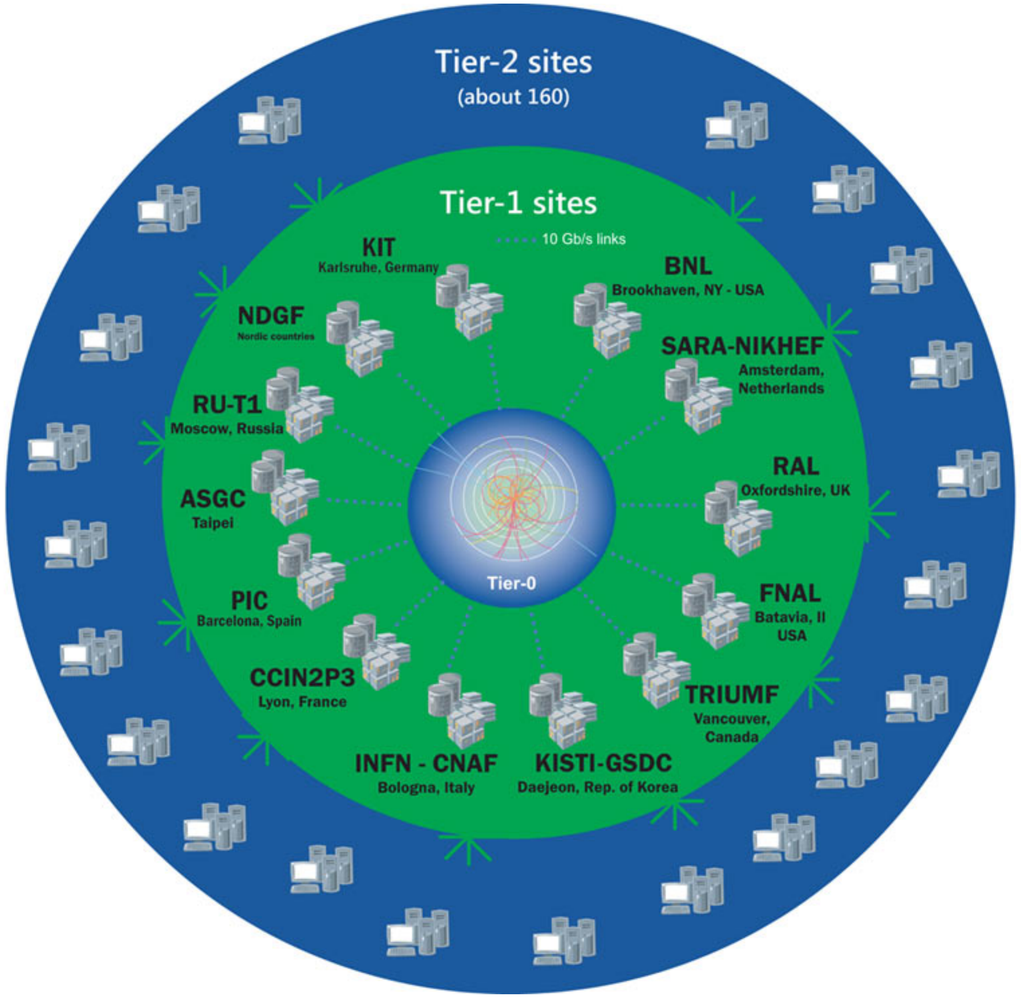
grid computing

Grid computing is the collection of computer resources from multiple locations to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. Grid computing is distinguished from conventional high performance computing systems such as cluster computing in that grid computers have each node set to perform a different task/application.

Basis of grid infrastructure

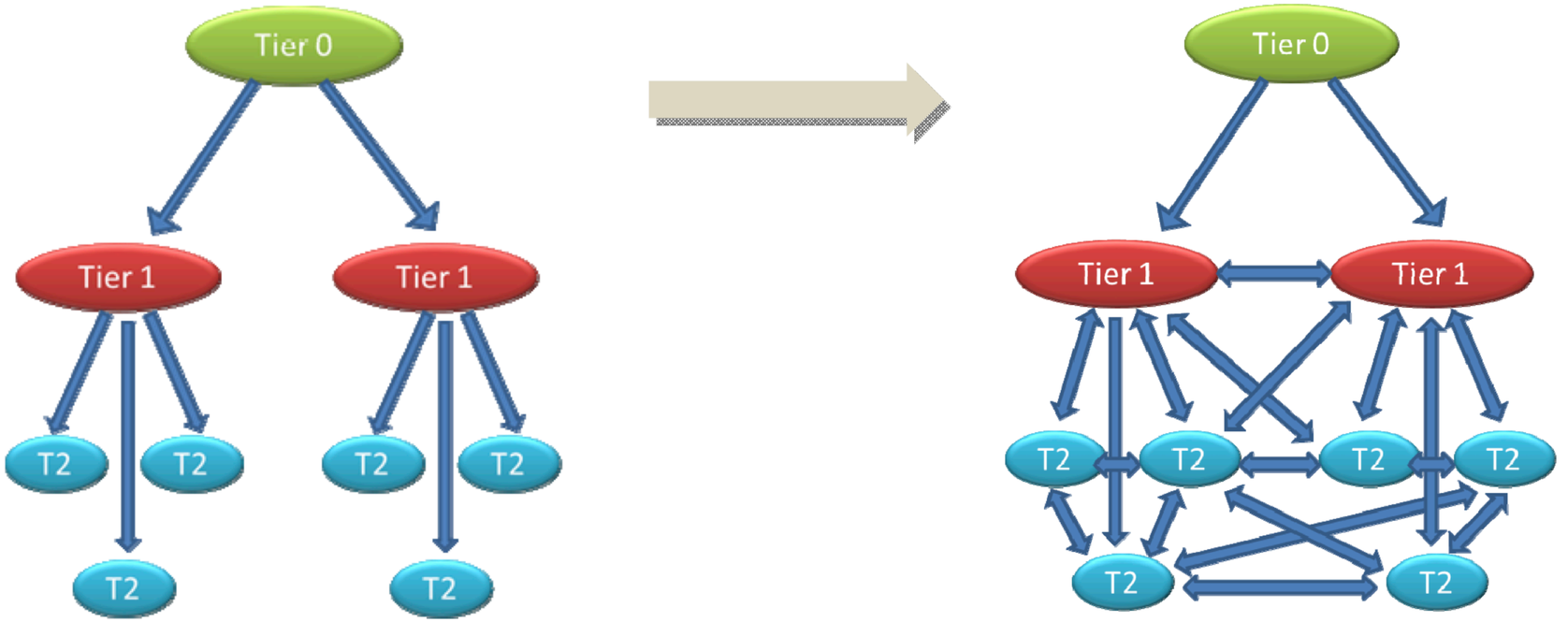
- Agreed rules of usage, shares of provided resources and level of participation
- Common authentication and authorization across infrastructure
- Common set of protocols and instruments for access to compute resources and data
- Information system with all required information about infrastructure
- A service which takes care of proper data catalog and data distribution
- A service which manage jobs execution

WLCG



- The mission of the Worldwide LHC Computing Grid (WLCG) is to provide global computing resources for the storage, distribution and analysis of the data generated by the LHC.
- WLCG combines about 1.4 million computer cores and 1.5 exabytes of storage from over 170 sites in 42 countries. This massive distributed computing infrastructure provides more than 12 000 physicists around the world with near real-time access to LHC data, and the power to process it.
- It runs over 2 million tasks per day and, global transfer rates exceeded 260 GB/s.
- CERN provides about 20% of the resources of WLCG.

From MONARCH to mesh



General requirements for distributed computing 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)

A lot of All basic components are already available from LHC experiments: CRIC (information system), FTS, CVMFS, Rucio, PanDA etc.

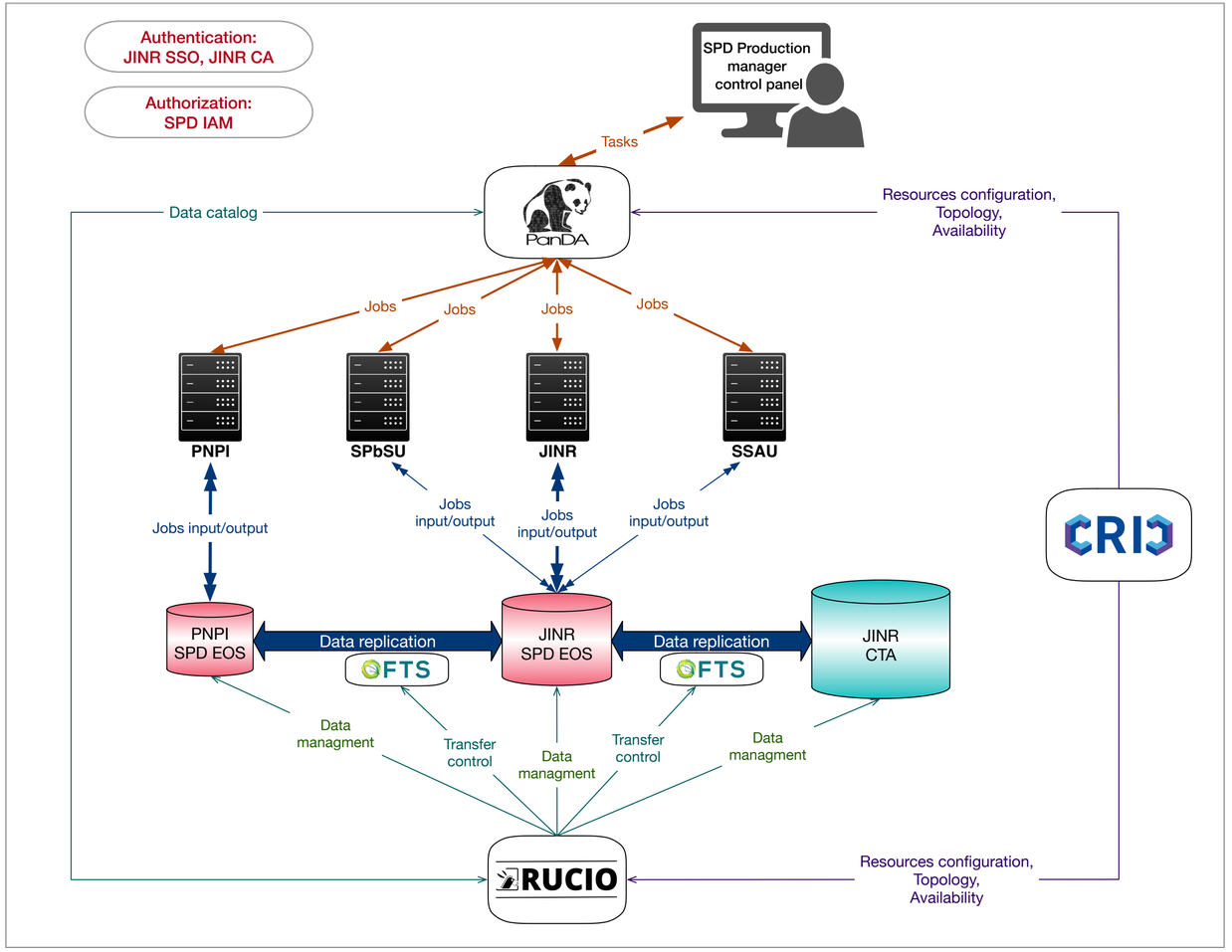
- a lot of work required to adapt of components to work together for particular experiment

Managing of data processing
in heterogenous distributed data processing system

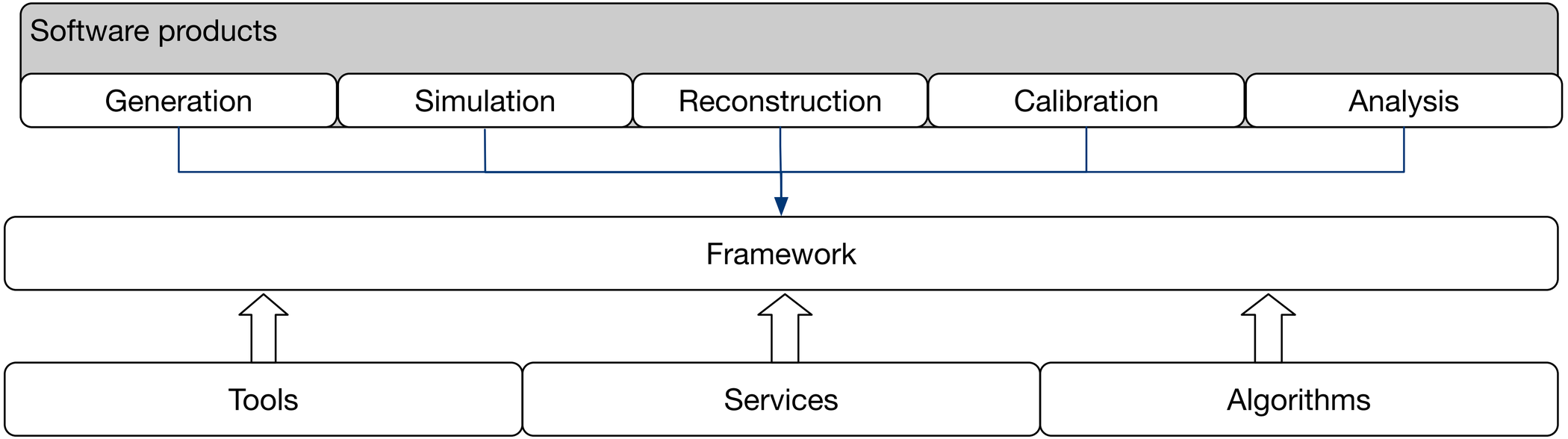
- Key middleware components required for efficient processing in grid:
  - **Workflow management system** - control the process of processing of data on each step of processing. Produce tasks, which required for processing of certain amount of data, manages of tasks execution.
  - **Workload management system** - processes tasks execution by the splitting of the task to the small jobs, where each job process a small amount of data. Manage the distribution of jobs across the set of computing resources. Takes care about generation of a proper number of jobs till task will not be completed (or failed)
  - **Data management system** - responsible for distribution of all data across computing facilities, managing of data (storing, replicating, deleting etc.)
  - **Data transfer service**: takes care about major data transfers. Allow asynchronous bulk data transfers.

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



Applied software framework



- Athena
- AliROOT
- MPDRoot
- FairROOT

SPD Software

Information processing of physics data of SPD Experiment

- A Gaudi-based software framework is being developed:
  - Geometry description: GeoModel
  - Generators: Pythia8, FTF, UrQMD + capability to add more generators
  - 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

Application software deployment

How to deliver software to particular site

- Software repositories acceptable, in case:
  - Software framework is not huge
  - No needs to have multiple version of software
  - No needs support of thousands simultaneous jobs
- Shared file system
  - Works good for single cluster
- CVMFS - wide shared read-oriented filesystem
- Containers to cover OS and environment differences

Data for users analysis.

Final step

- Huge dataset of collected and simulated data are not well fit for usage of particular used due to its size
- Usually, only a set of particular properties from a subset of events is needed
  - Derivation - data reduction process which achieved by slimming, skimming and thinning procedures.
    - Slimming - subset of files form datasets
    - Skimming - subset of events from files
    - Thinning - subset of only required properties of events

Event index & problem oriented databases

- 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

Scientific digital services

- Collaborative tools:
  - E-log - digital shifter logbook
  - Project management and issue tracking system
  - Digital libraries for managing of documentation and publication
  - Group calendars, meetings support etc…

Conclusions

- HEP Data processing requires significant efforts from IT side
- Amount and structure of HEP data may be easy qualified as BigData
  - Efficient algorithms and methods required for data processing
  - Sophisticated information systems is needed for managing of processing in distributed and high performance computing systems
- Collaborative scientific activities requires at supporting by different digital services