Computing in CBM

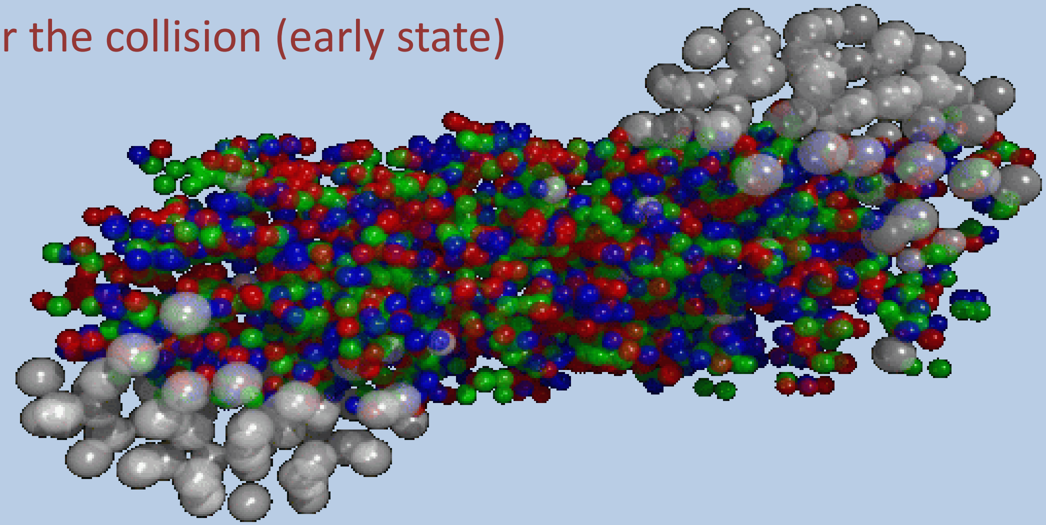
Volker Friese
GSI Darmstadt

International Conference on Matter under High Densities
21 – 23 June 2016
Sikkim Manipal Institute of Technology, Rangpo

Complex Tasks

The Hope:

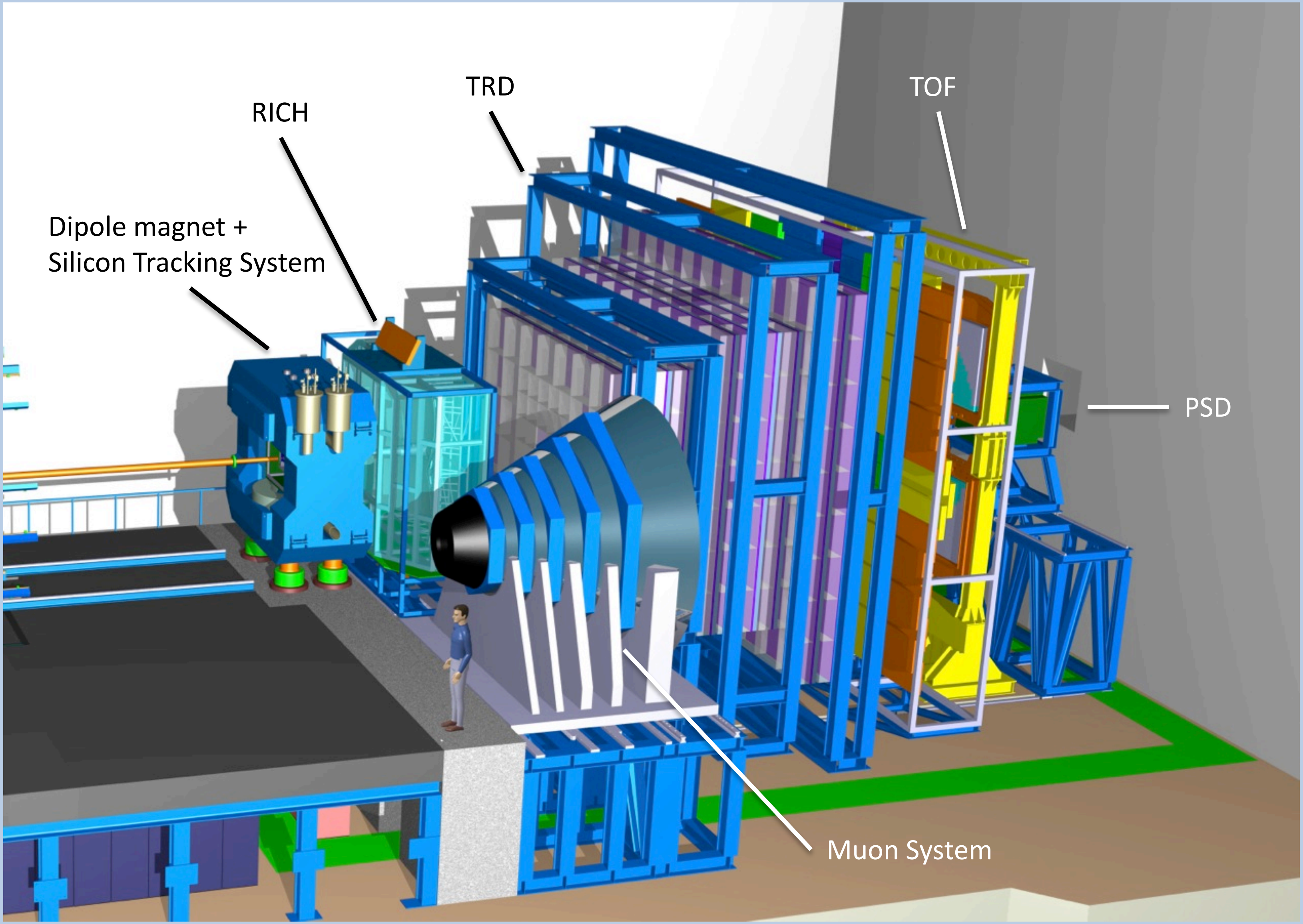
Learn from the multitude of emitted particles (final state) about the state of the matter immediately after the collision (early state)



The Task:

Detect the final-state particles as completely as possible and characterise them w.r.t. momentum and identity (π, K. p, …)

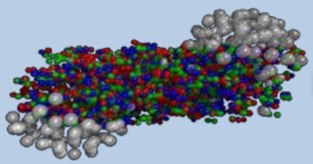
Complex Devices



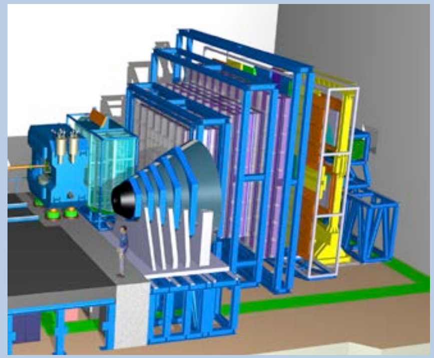
Tasks for reconstruction

„Reconstruction“

Particles



Detector



Raw Data

00110
10110
10101
00011
0001
11100
10011
0

Hi, I am channel 31846002. If you have a DAQ map, you will know that this is strip 348 of module l4u03 of the fourth layer of the main tracker. I just saw a charge of 11 ADC units. My clock reads 723948532 ns.

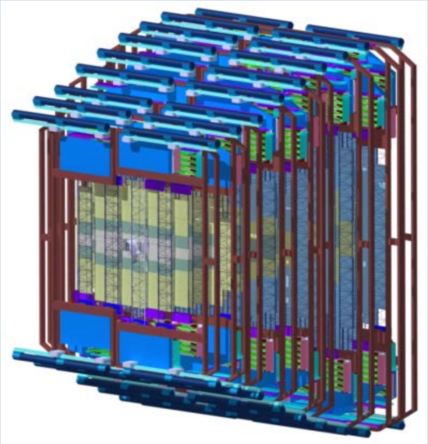
Reconstruction delivers a list of particles (trajectories) with momenta (and ID), which is the input for the higher-level physics analysis.

Reconstruction Tasks

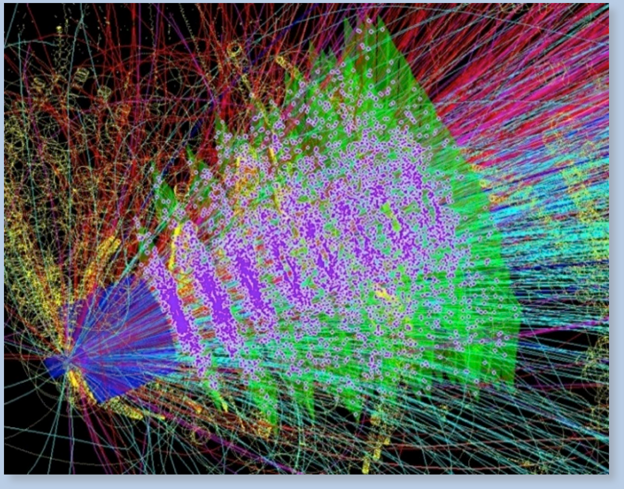
- are typically pattern recognition issues
- act both local…
  - cluster / hit finding
  - local track finding
- … or global
  - connect track information from different detector systems
- Algorithms are usually developed by experts and executed in a central production.
- Results are provided to the physics users in some suitable event format.
  - user normally does not see raw data

Example: Track Finding in the Main Tracker

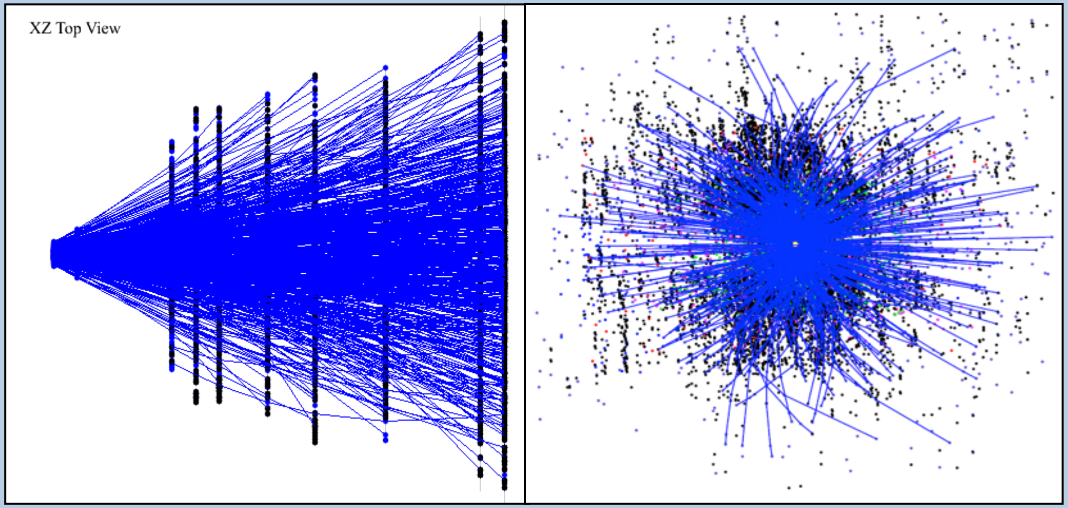
Silicon Tracking System



Event Display



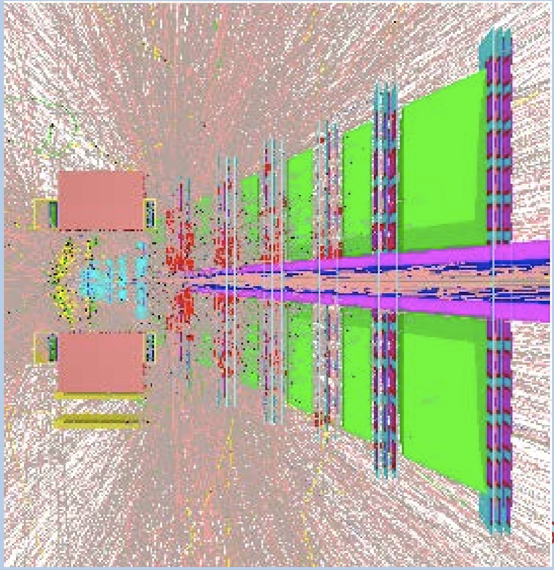
Reconstructed Tracks

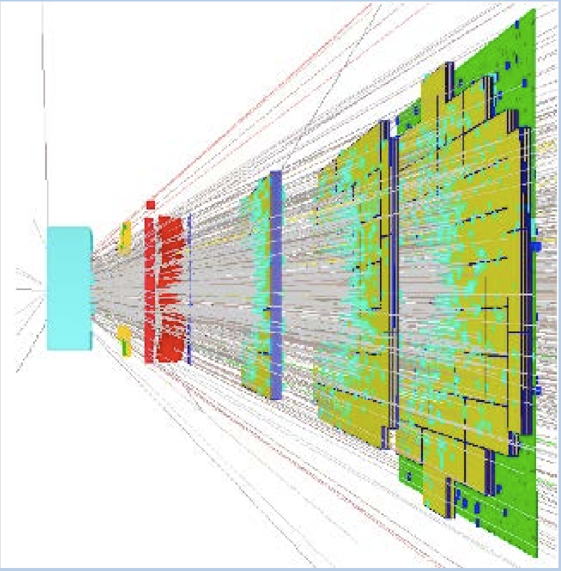


Track finding: define a set of measurements (hits) which belong to one and the same trajectory (particle)

Difficulty: Large number of hits; fake hits exceed true hits by factors.

Similar: Track Finding in the MUCH and TRD

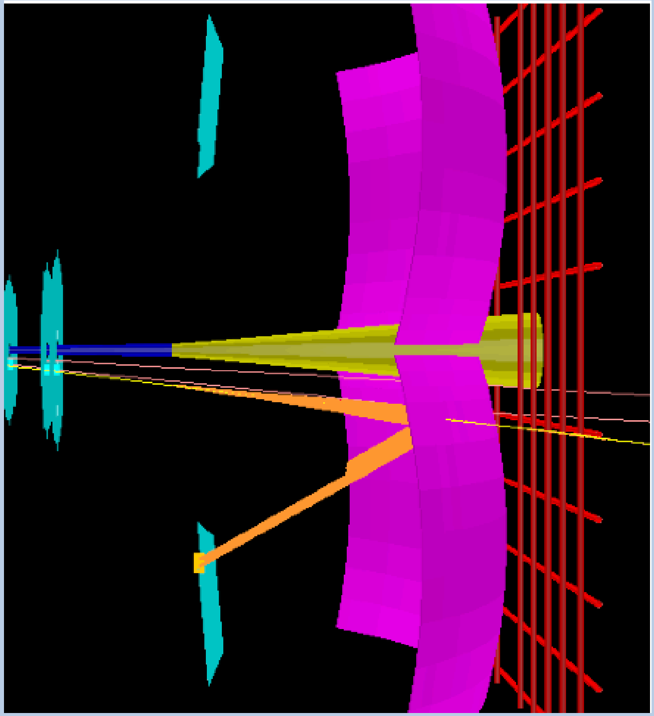




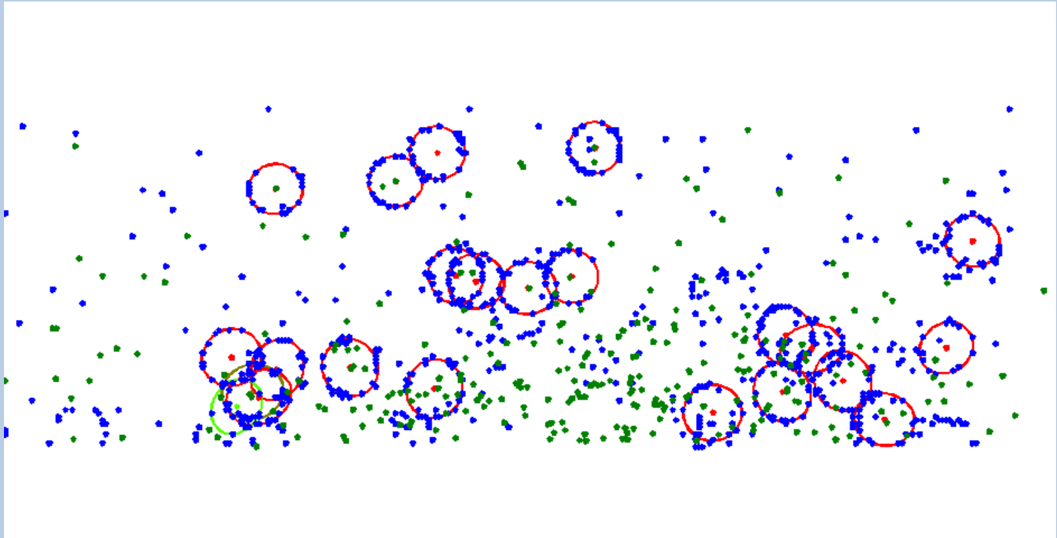
Muon System: Outside magnetic field (straight tracks), but absorber layers complicate tracking

TRD: Outside magnetic field (straight tracks); coordinate resolution worse than STS

Another Example: Ring Finding in the RICH



Event Display



Cherenkov light emitted by electrons in the radiator is mirrored and focused into rings onto the photodetector plane.

Problems:

- High hit / ring density
- Overlapping rings
- Ring distortions

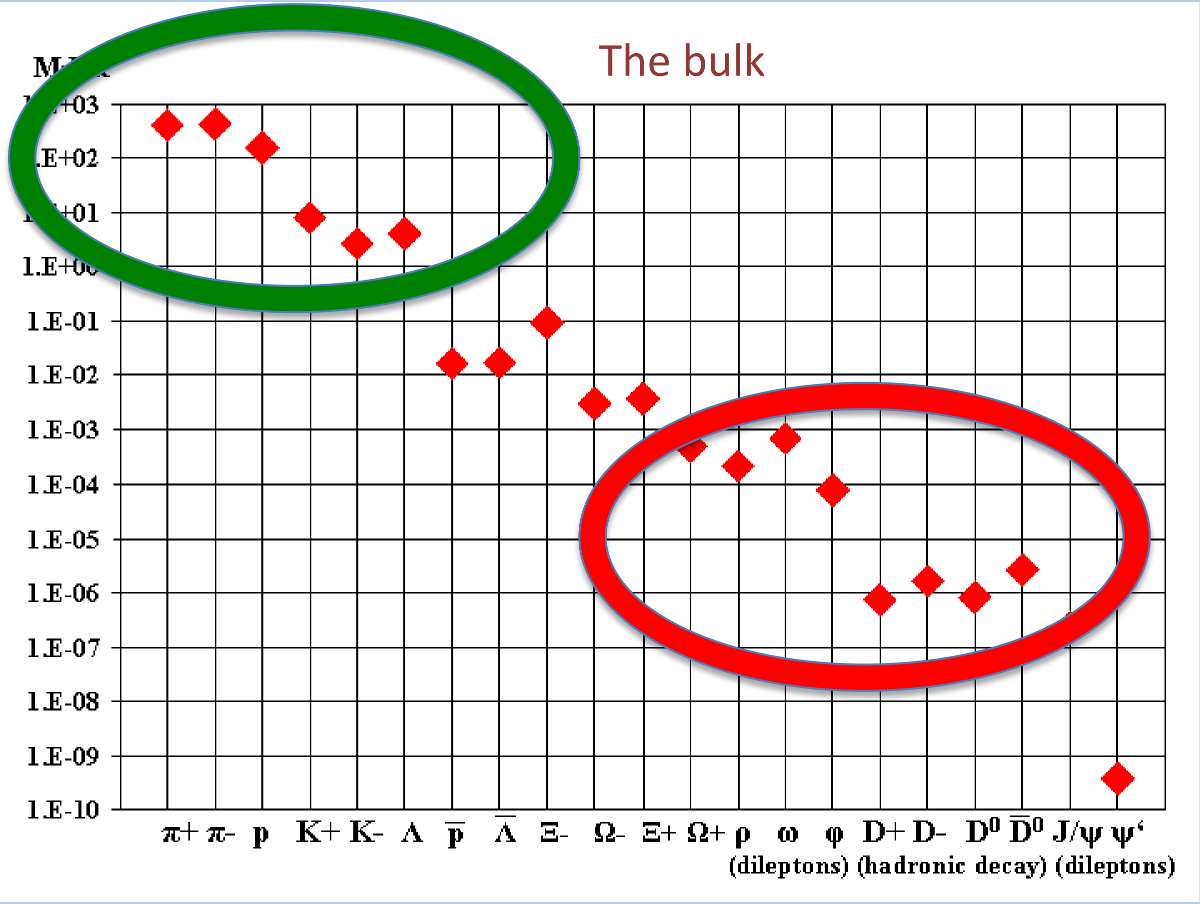
Event Reconstruction: A First Upshot

- The pattern recognition problems in event reconstruction from raw data are common to HEP experiments.
- A bunch of methods were applied in previous or current experiments.
- There is, however, nothing “off the shelf”; methods and algorithms have to be adapted / tuned / developed for each experiment.
- Reconstruction in heavy-ion experiments is more demanding than in particle physics because of the large track multiplicity -> high complexity of the event topology.
- There are solutions developed for CBM which fulfil our demands in terms of efficiency and precision.

High Rates and the Data Challenge

The Shopping List

Model predictions of particle multiplicities (Au+Au, 25A GeV)

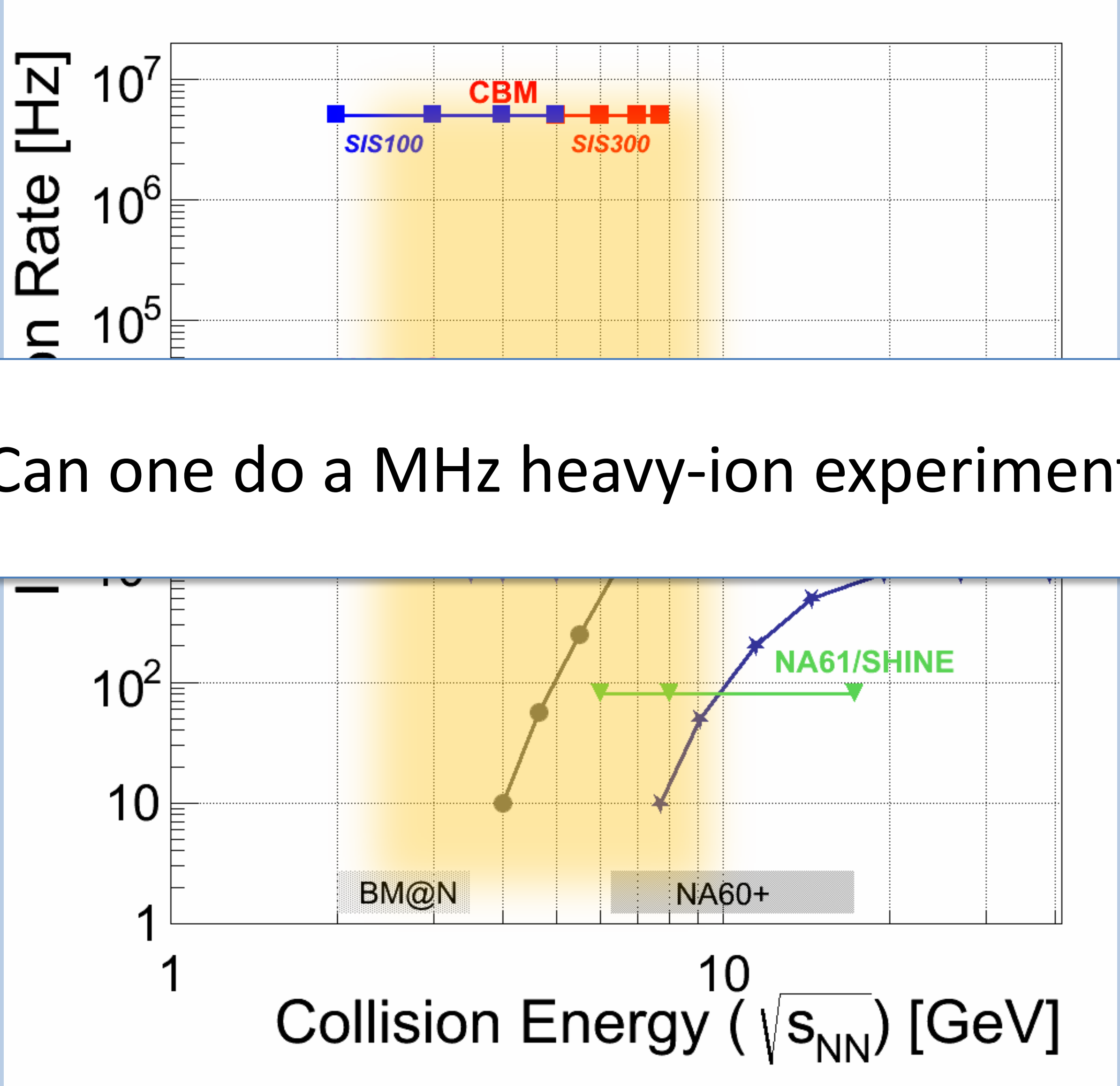


The needles in the haystack

Measurement of very rare probes: requires extreme interaction rates

Which rate is possible in heavy-ion reactions?

The Rate Landscape



Can one do a MHz heavy-ion experiment?

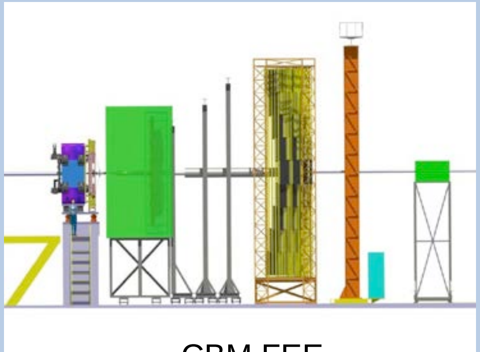
Limiting Factors

- At an interaction rate of 10 MHz (Au+Au), the average time between subsequent events is 100 ns.
  - but there are many with much less time spacing.
- Detectors have to be fast
  - possible (i.e., solid-state detectors instead of gas drift chambers)
- Read-out electronics has to be fast
  - not trivial, but possible
- The limiting factor is the data rate to be shipped from the detectors to the permanent storage.

The Data Rate Problem

- At an interaction rate of 10 MHz (Au+Au), the raw data rate from the detector is about 1 TB/s.
- What can we store?
  - to tape: several GB/s
  - to disk: 100 GB/s no problem (e.g., GSI Lustre FS)
- What do we want to store?
  - at 1 GB/s archival rate, the amount of data after 2 months of beam time is 5 PB.
- The storage bandwidth is rather limited by the cost of storage media than by technological constraints.
- For a given rare observable, 99% of the raw data are physically uninteresting.
  - we would like to trigger on such observables.

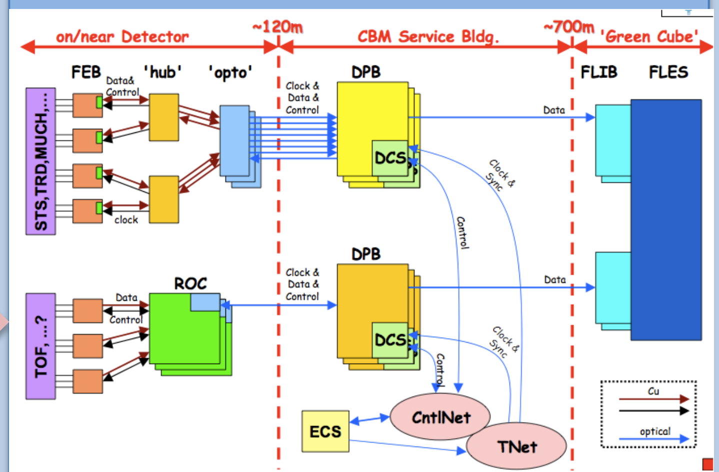
The Online Task



CBM FEE

1 TB/s

at max. interaction rate



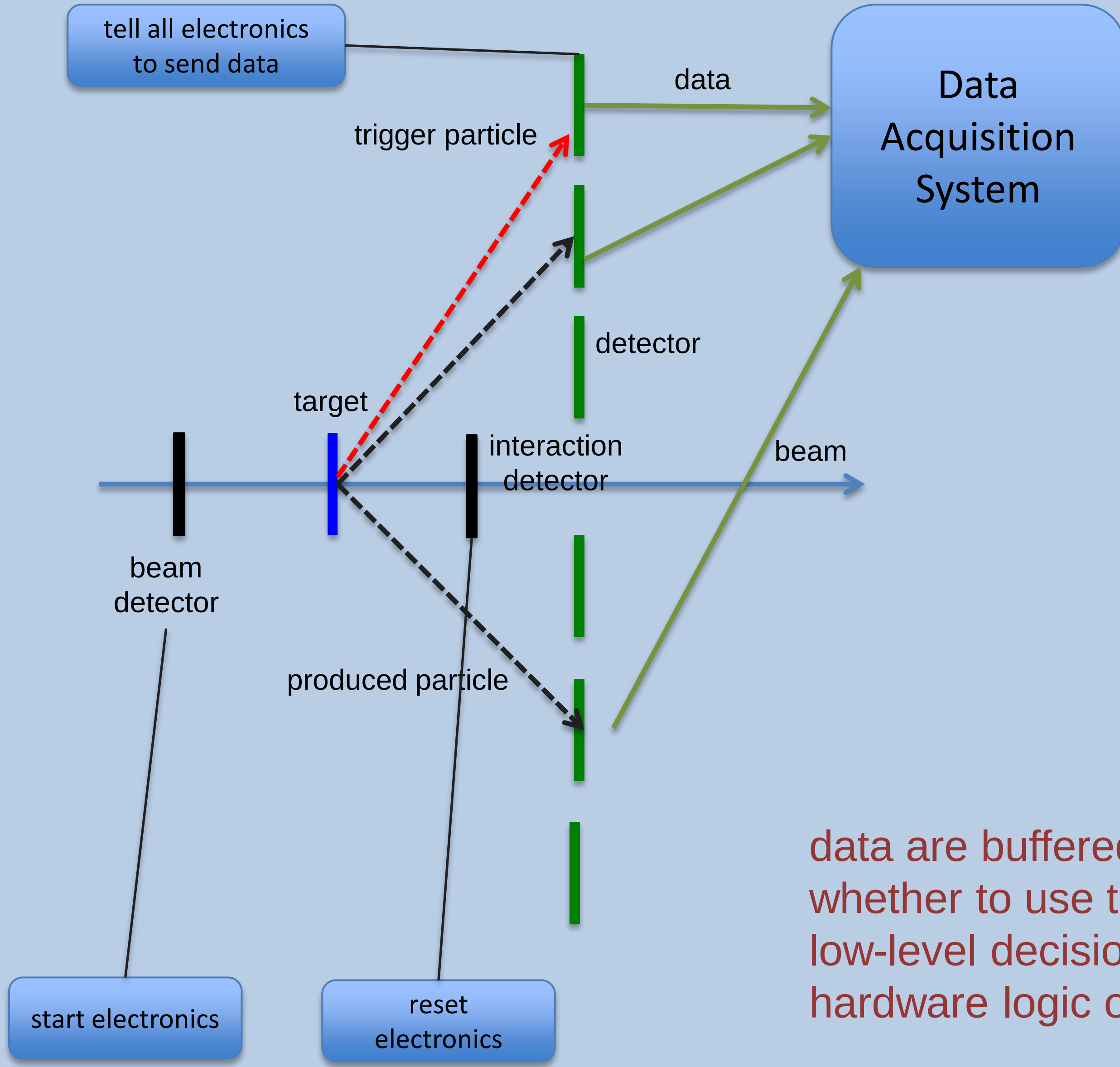
~ GB/s

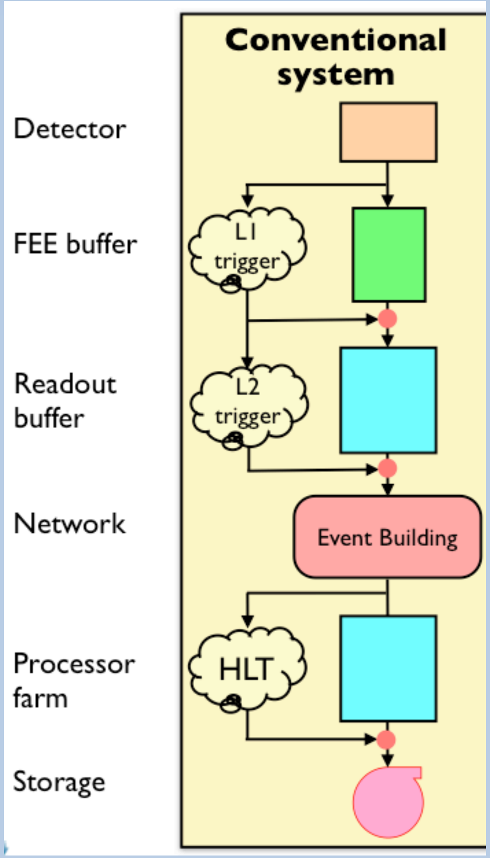
Mass Storage

Trigger:

- tells the readout electronics when to read out the detector
- tells the electronics whether to send the collected data further or discard then

A Trigger Example (Schematic)



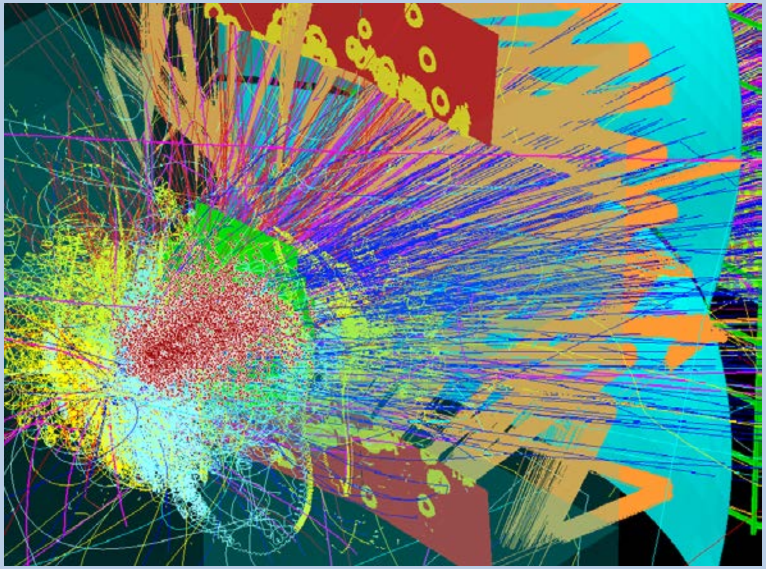


data are buffered on-chip until decision whether to use them or not

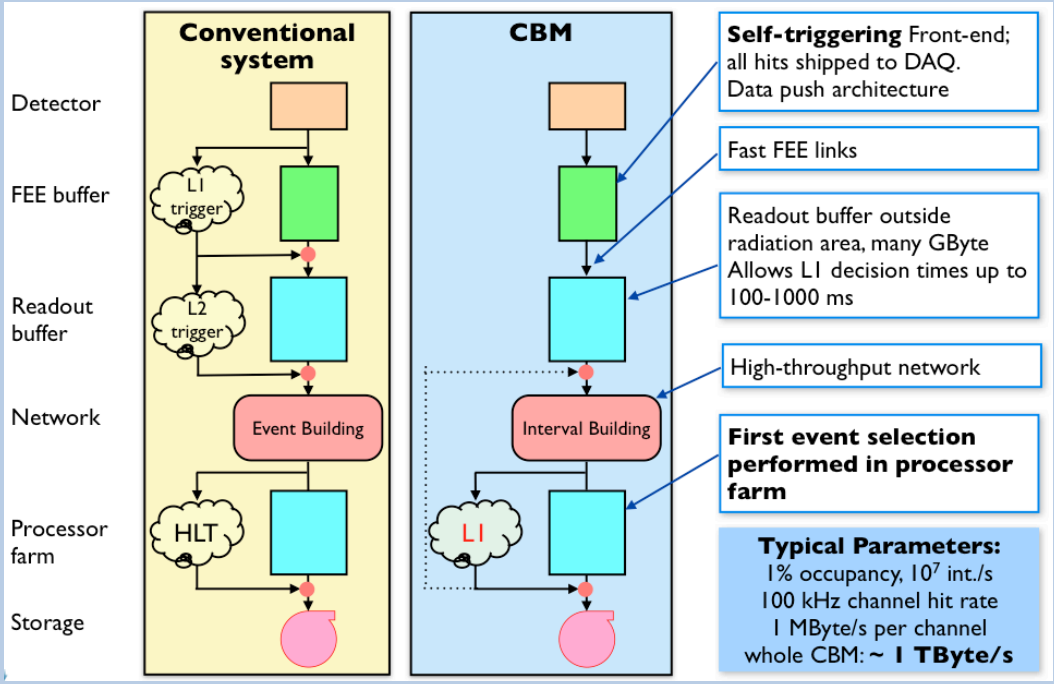
low-level decision is evaluated in hardware logic or on FPGAs

Why a Triggered Readout Does Not Work for CBM

- The trigger signatures (e.g. Ω^+ cascade decay) are complex and involve several detector systems at a time (e.g., STS + TOF)
  - not possible to implement in hardware
- 1 MHz or above does not allow for high trigger latency
  - FEE buffer is difficult and expensive, in particular if radiation hard
  - not possible to wait for trigger decision in software



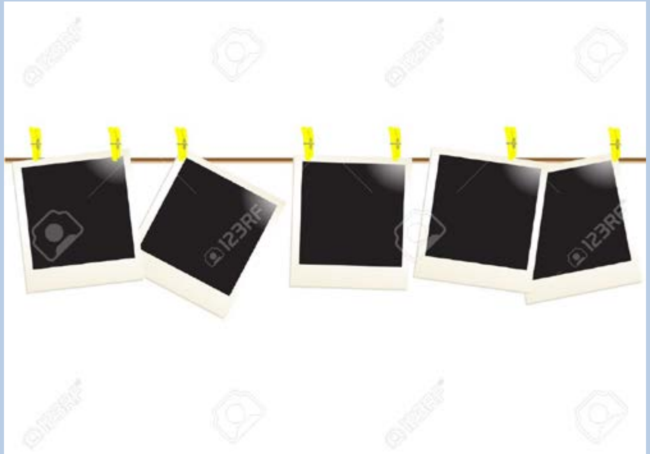
Free Streaming Read-out and Data Aqcuisition



- Continuous readout by FEE
- FEE sends data message on each signal above threshold ("self-triggered")
- Hit message come with a time stamp
- DAQ aggregates messages based on their time stamp into "time slices"
- Time slices are delivered to the online computing farm (FLES)
- Decision on data selection is done in the FLES (in software)

Triggered and Free-Running Readout

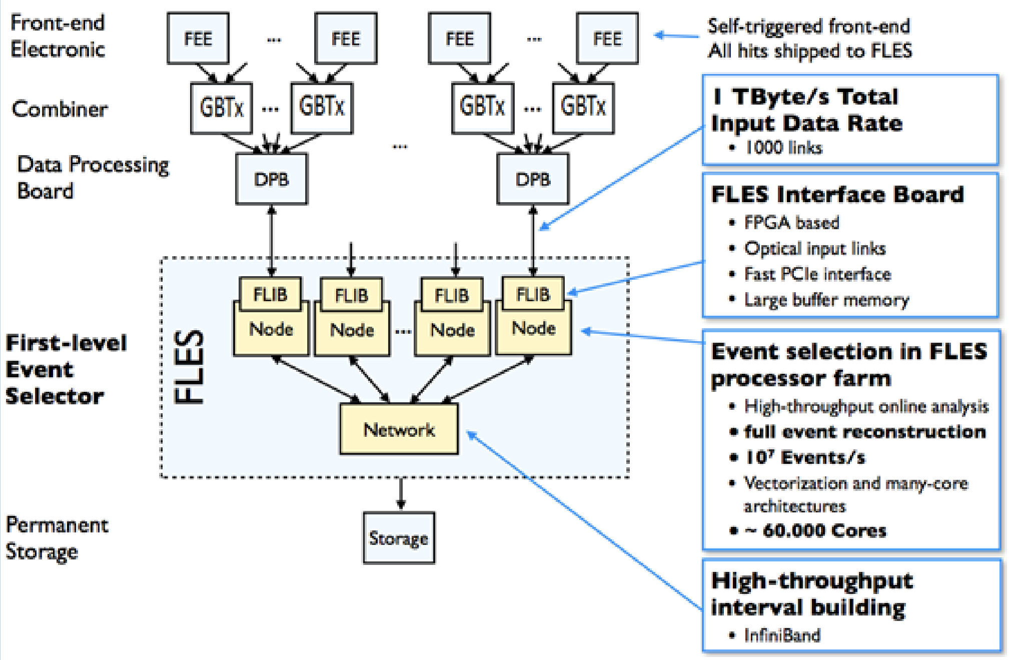
Trigger: snapshots of the detector



Trigger-less: a movie of the detector

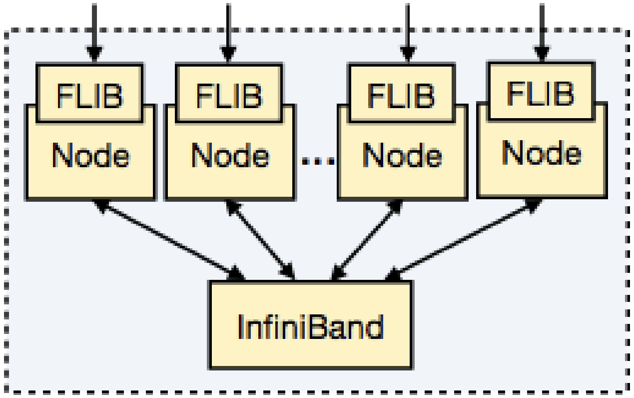
N.b.: Too large to be stored! Will be cut into pieces in the photo lab (= FLES).

DAQ and First-Level Event Selector



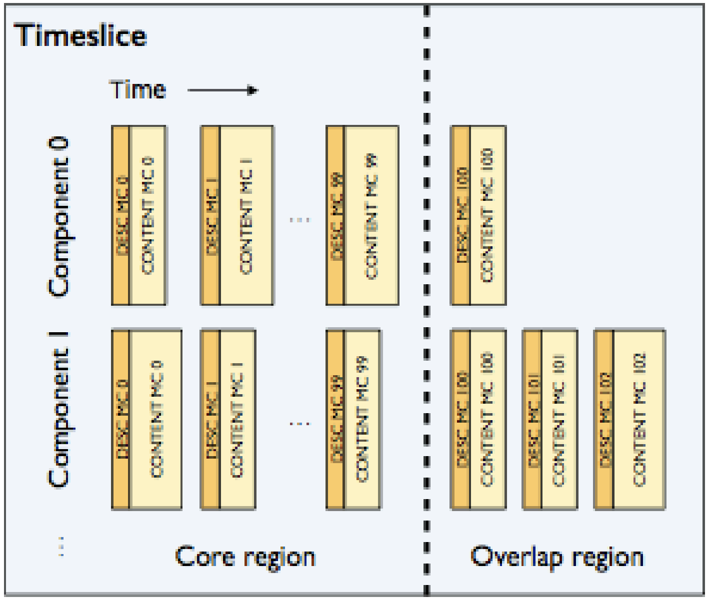
FLES prototype: Loewe CSC Frankfurt

FLES Architecture



- FLES is designed as an HPC cluster
  - Commodity PC hardware
  - GPGPU accelerators
  - Custom input interface
- Total input data rate ~1 TB/s
- InfiniBand network for interval building
  - High throughput, low latency switched fabric communications
  - Provides RDMA data transfer, very convenient for interval building
  - Most-used system interconnect in latest TOP500 (224 systems)*
- Flat structure w/o dedicated input nodes
  Inputs are distributed over the cluster
  - Makes use of full-duplex bidirectional InfiniBand bandwith
  - Input data is concise, no need for processing before interval building
- Decision on actual commodity hardware components as late as possible
  - First phase: full input connectivity, but limited processing and networking

Time Slice: Interface to Online Reconstruction



Timeslice
- Two-dimensional indexed access to microslices
- Overlap according to detector time precision
- Interface to online reconstruction software

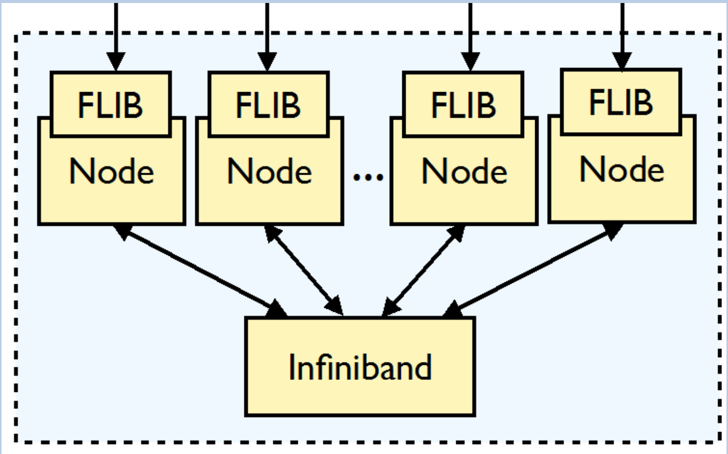
- Basic idea: For each timeslice, an instance of the reconstruction code...
  - is given direct indexed access to all corresponding data
  - uses detector-specific code to understand the contents of the MCs
  - applies adjustments (fine calibration) to detector timestamps if necessary
  - finds, reconstructs and analyzes the contained events

Consequences for Online Computing

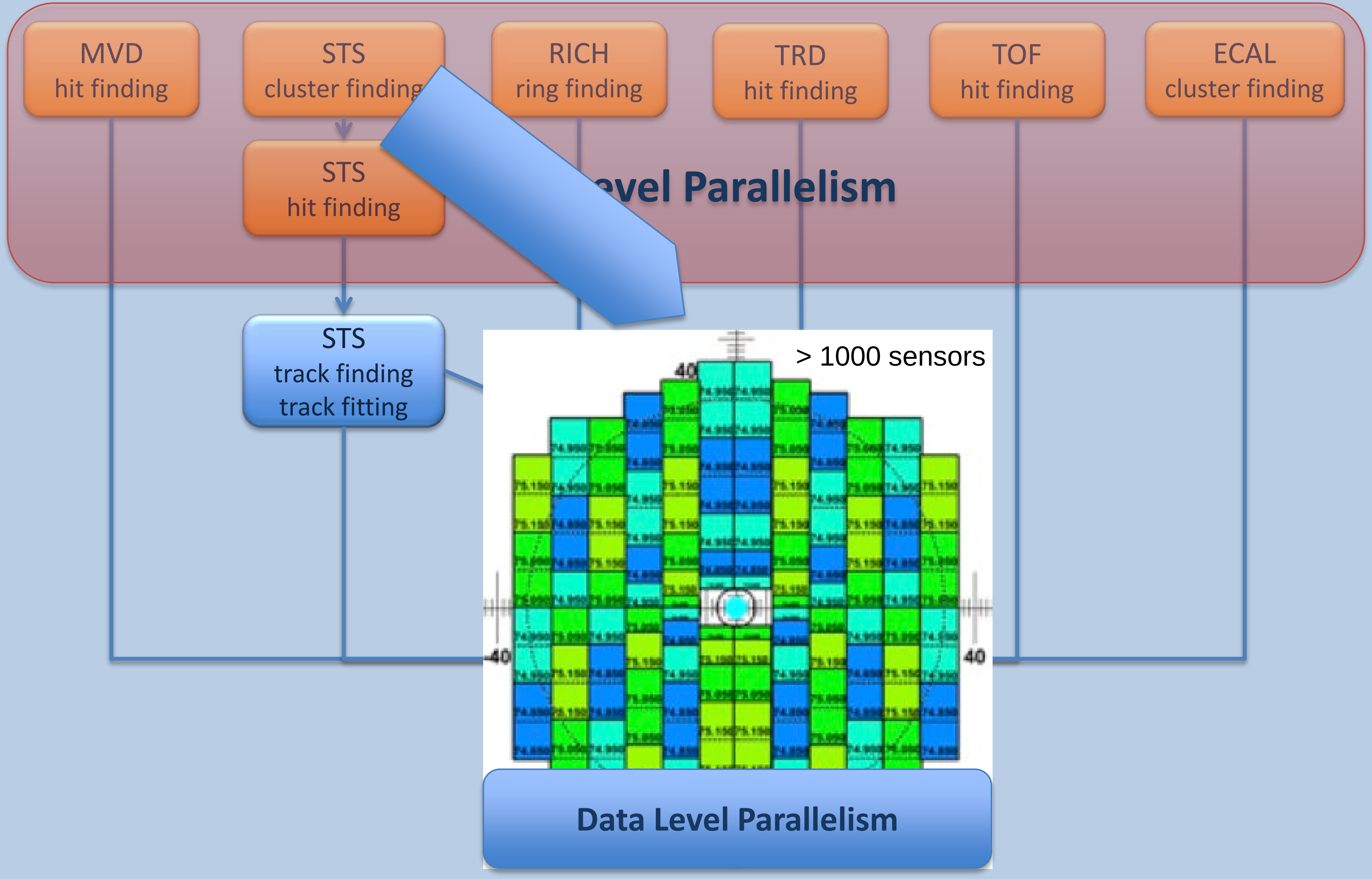
- CBM tries to achieve high interaction rate capability with a novel readout paradigm.
- Data will be continuously read out; the full data volume will be shipped to CPU.
- Online data selection will happen on the online compute farm ($O(10^5)$ cores), where data reconstruction will be performed in real-time.
- The quality criteria for reconstruction algorithms are not only efficiency and precision, but also, and mostly, execution speed.
- To achieve the required performance, the computing parallelism offered my modern computer architectures must be exploited.
- High-performance online software is a necessary pre-requisite for the successful operation of CBM.

Where to parallelize?

- FLES is designed as HPC cluster
  - commodity hardware
  - GPGPU accelerators
- Chunks of data („time slice“) distributed to independently operating computing nodes.
- Obvious data parallelism on event / time-slice level
- But: each computing node will have a large number of cores
- Need in addition parallelism within event / time slice



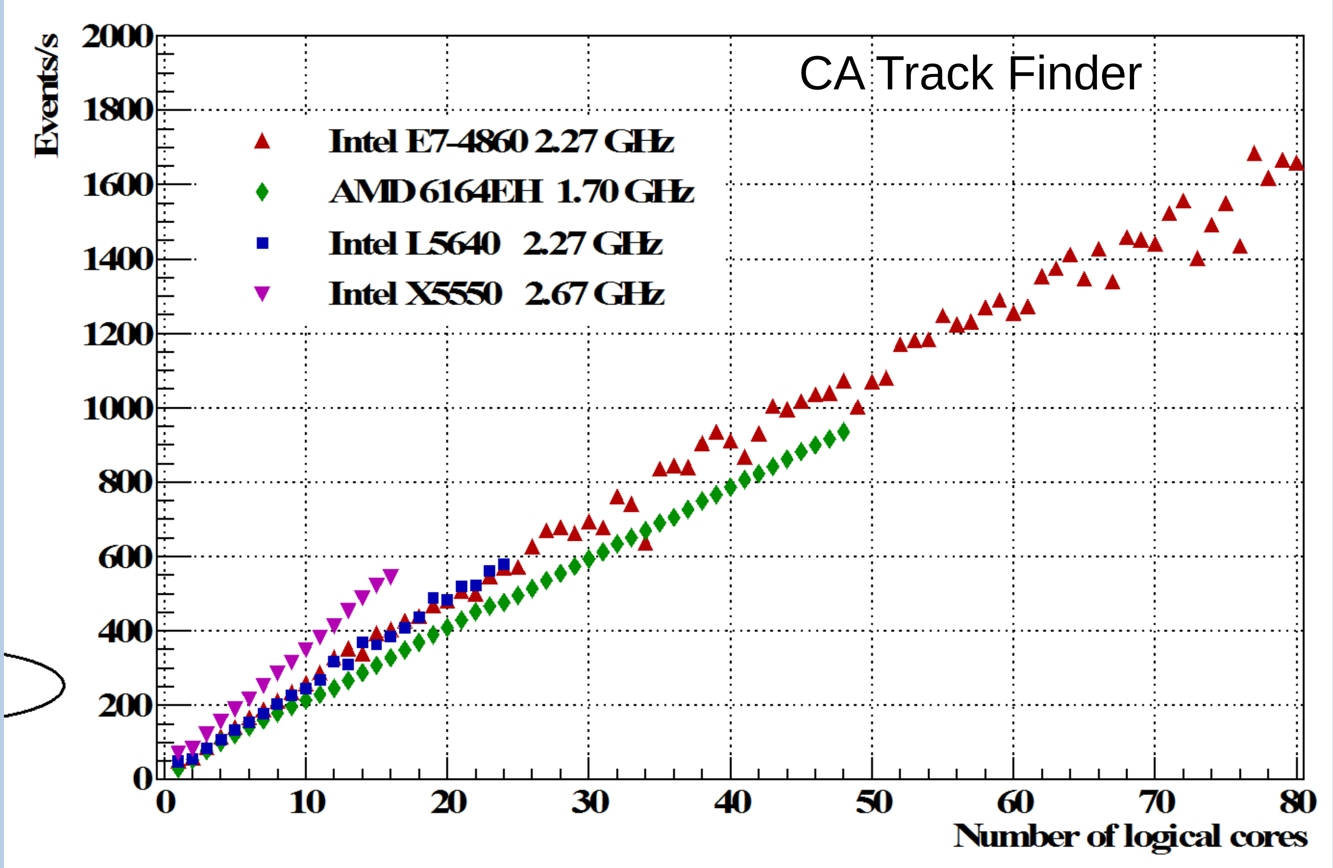
Parallelisation within event



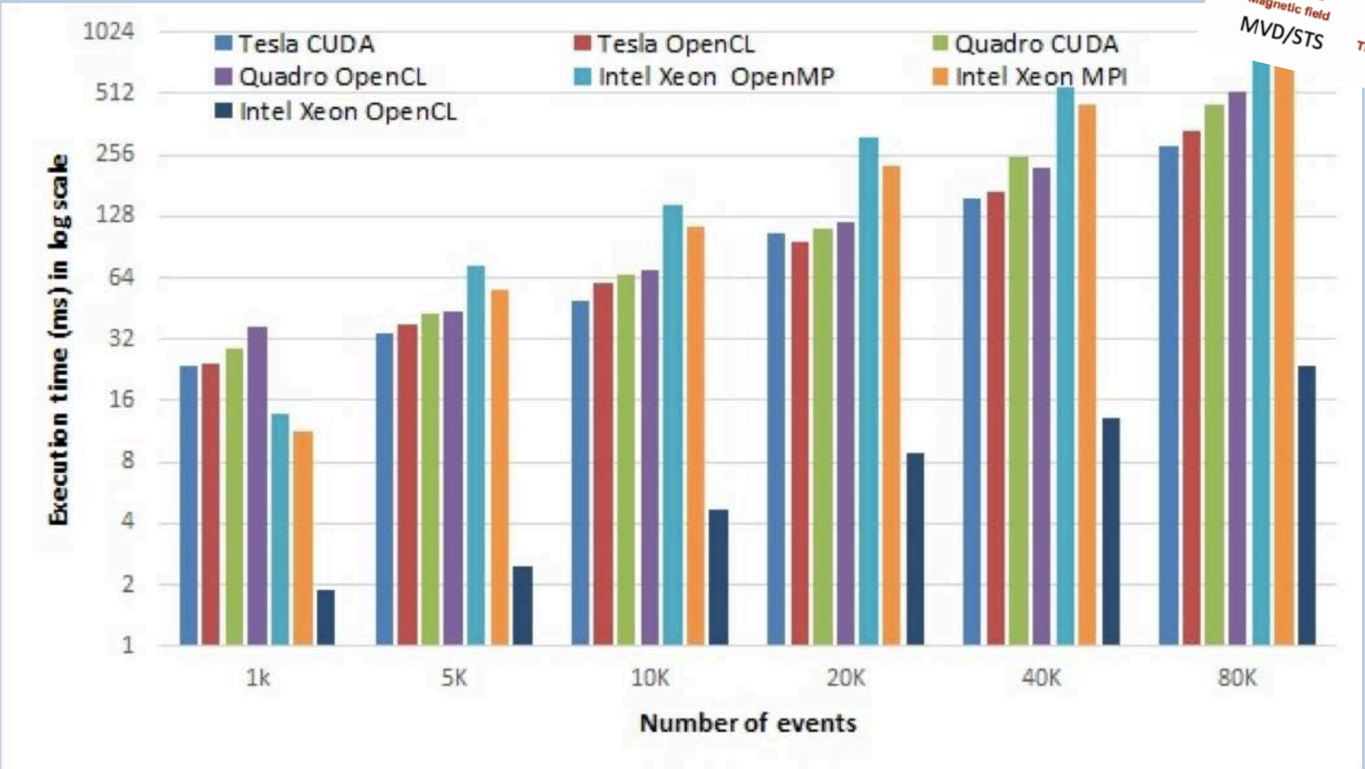
Ways to parallelize: multi-threading

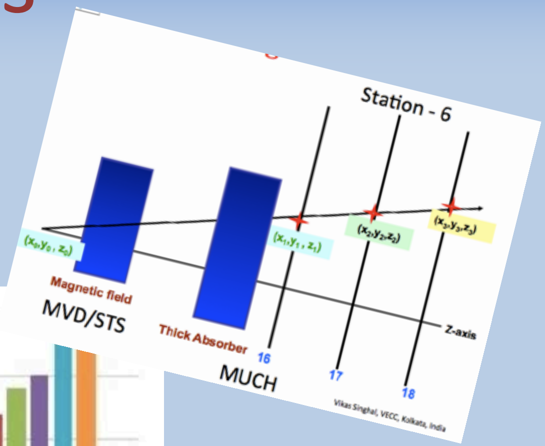
"Trivial" event-level parallelism: reconstruction with independent processes.

Exploit many-core systems with multi-threading: 1 thread per logical core, 1000 events per core. Gives good scalability.



Muon trigger studies





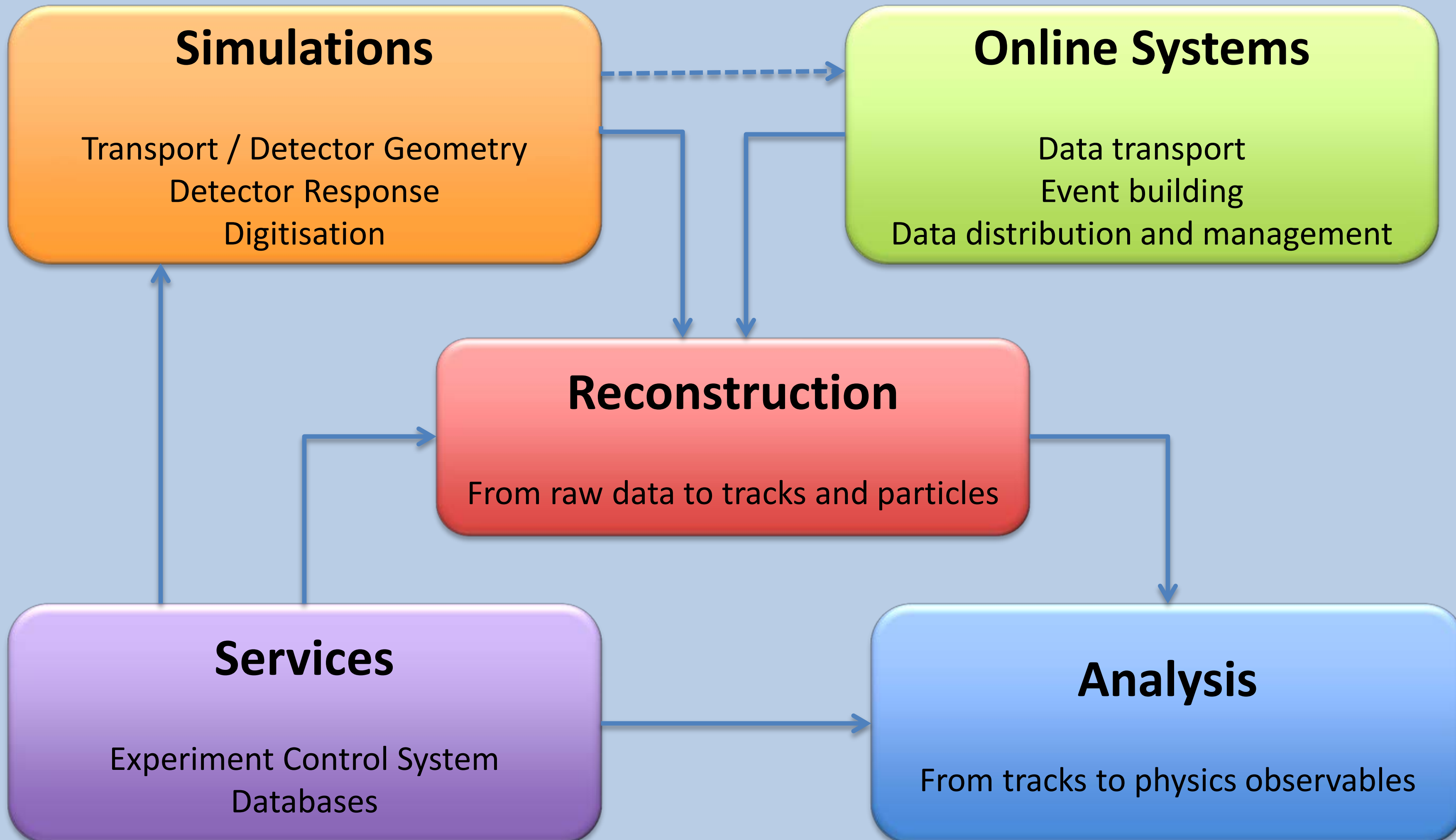
Investigation on several CPU / GPU architectures. Strong differences between different computing paradigms on same architecture.

Offline Computing

Assuming we manage to solve the online computing issues, we will have some 5 PB per year on permanent storage.

How to efficiently get physics results out of them?

Computing Tasks



Distributed Computing

- 10 years ago, LHC experiments were confronted with a similar problem. The amount of data was too large to be handled on a single site.
- The answer was GRID computing: data and data processing are distributed on many sites worldwide.
- To be considered are three factors:
  - Computing power
  - Storage capacity
  - Data transfer between sites
- Experience shows that the limiting factors are
  - Network bandwidth
  - Site administration

New Paradigms for CBM / FAIR

- Today, the entire LHC grid compute power could easily be concentrated in one large computing centre.
- The CBM Online Compute Cluster (~10^5 cores) is commodity hardware; it can be used for offline computing between runtimes (¾ of the year).
- Since reconstruction must be fast (performable in real-time!), there is no need to store reconstructed data. Analysis can always start from raw data.
- Strategy tends towards a small number of large centres connected by high-speed links - details still to be worked out.

Thanks for the attention!