The online event selection architecture of the CBM experiment

J. de Cuveland*, D. Hutter*, H. Hartmann*, V. Lindenstruth*

* Frankfurt Institute for Advanced Studies, Johann-Wolfgang-Goethe-Universität Frankfurt am Main, Germany

Motivation

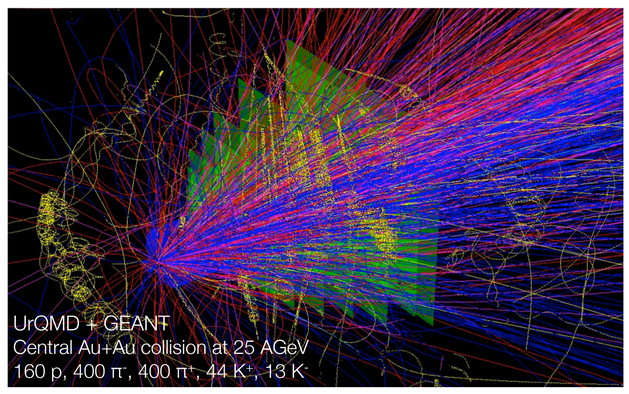
CBM

The goal of the Compressed Baryonic Matter (CBM) experiment at FAIR is to explore the QCD phase diagram in the region of high baryon densities. The experimental task is to identify hadrons and leptons in collisions with up to 1000 charged particles at event rates of up to 10 MHz. These measurements require fast self-triggered readout electronics, a high-speed data acquisition architecture, and an appropriate high-level event selection concept.

The First-level Event Selector (FLES) is the central event selection system of the CBM experiment. Designed as a high-performance computing cluster, its task is an online analysis of the physics data at a total data rate of > 1 TByte/s.

The FLES has to combine the data from approximate 1000 input links to self-contained, overlapping processing intervals (timeslices), on which full 4-D event reconstruction is performed in real-time. Specialized event selection algorithms identify events and select those relevant for storage.

Challenge

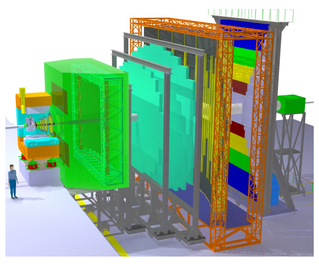


- Extreme reaction rates up to **10 MHz**
- Up to **1000 charged tracks** in aperture
- Hit densities up to 1/mm^2
- High-precision vertex reconstruction
- Identification of leptons and hadrons
- No conventional trigger architecture possible
- Self-triggering readout electronics
- Event selection exclusively done in a high-performance computing cluster
- **Full online event reconstruction needed**

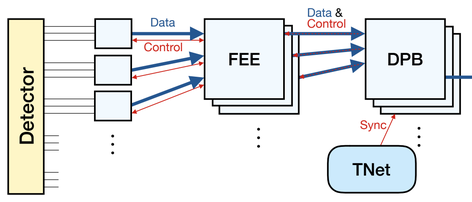
CBM First-level Event Selector: FLES

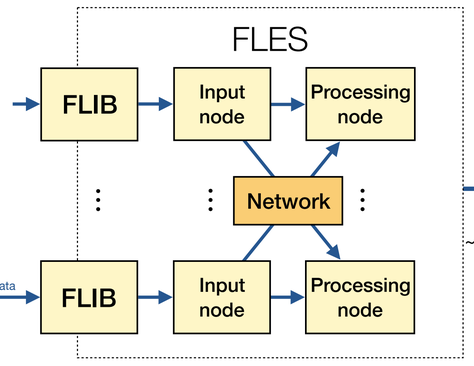
- Central physics selection system for CBM
- HPC processor farm with FPGAs, GPUs and fast interconnect
- > 1 TByte/s input data stream
- High-throughput event building
- Online analysis using fast, vectorized many-core track reconstruction algorithms

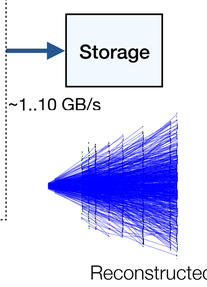
10^7 events/s
- Self-triggering front-end
- Data push readout
- Complex global triggers



CBM







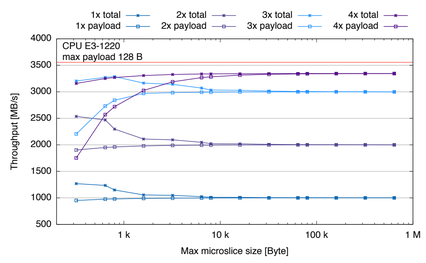
Reconstructed event in STS

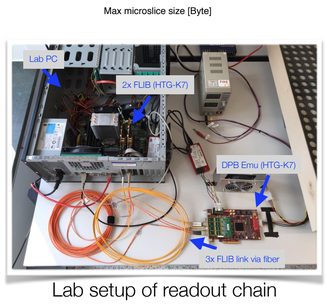
High-throughput event building
- **>1 TByte/s** input data rate
- ~ 1000 input nodes
- ~ 60.000 cores

Online Event Selection
- 4-D tracking
- Identification of leptons and hadrons
- High-precision vertex reconstruction

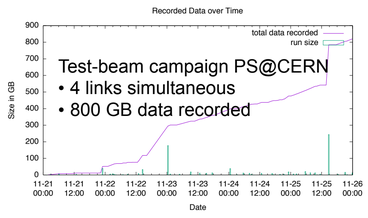
Input Interface

- FPGA-based PCIe board: FLIB
  - Consumes microslices received from DPBs
  - Preparation and indexing of microslices for timeslice building
  - Transfers microslices and index data to PC memory
- High input data rate of > 10 Gbit/s sustained
- Custom PCIe DMA interface
- Optimized data scheme for zero-copy timeslice building

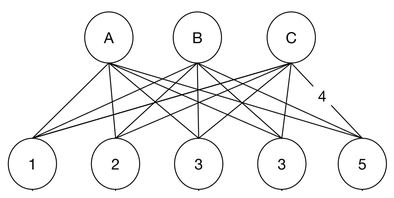




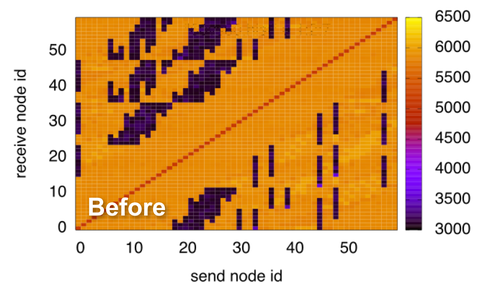
Lab setup of readout chain



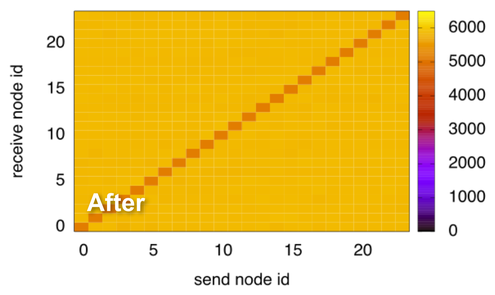
Network Performance

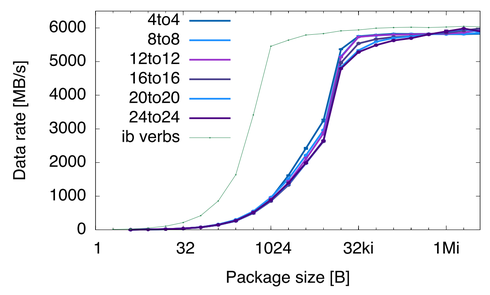


L-CSC arranged in a 5-3 fat-tree network
12 nodes per leaf switch, 4 uplinks



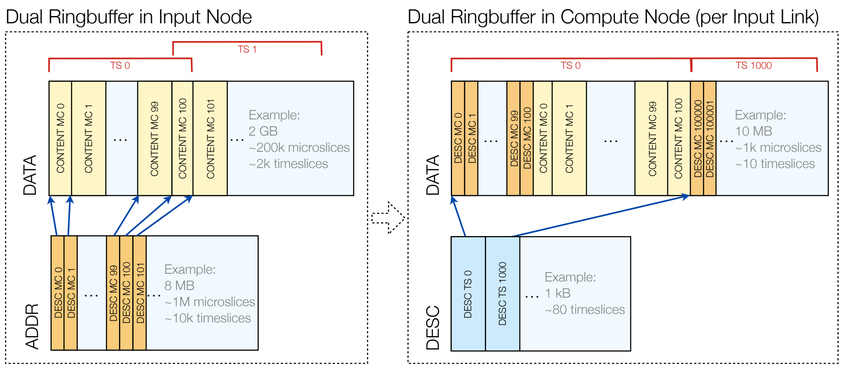
- Standard routing pattern suboptimal for continuous all-to-all communication
- **Optimized routing scheme leads to excellent performance (>5 GB/s per node)** (tested on for 24 nodes using InfiniBand verbs and custom MPI benchmark)





Timeslice Building & Data Management Framework

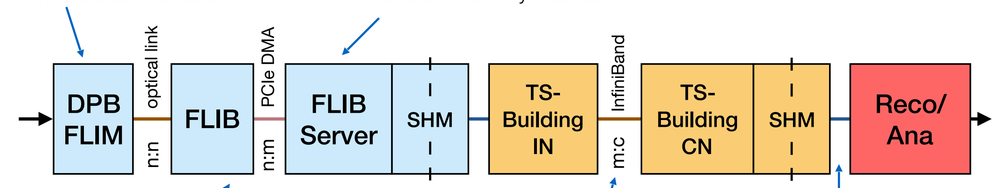
- Data input to FLES is distributed across approx. 1000 input nodes (IN)
- In the process of timeslice building, matching time intervals from all input links are combined to one "timeslice" (processing interval) on one particular compute node (CN)
- Timeslice substructure: microslices (constant in experiment time), allow overlapping timeslices



- Timeslice building and readout software: FLESnet
- Delivers fully built timeslice to reconstruction code
- Based on indexed ring buffers and RDMA
- Paradigms:
  - Maximize throughput
  - Do not copy data in memory

Data Flow

- 10 GBit/s custom optical link
- Microslice interface
- Direct DMA to InfiniBand send buffers
- Shared memory interface



- Data preprocessing for timeslice building
- Full-offload PCIe DMA engine
- Timeslice building
- InfiniBand, true zero-copy
- Indexed access to timeslice data

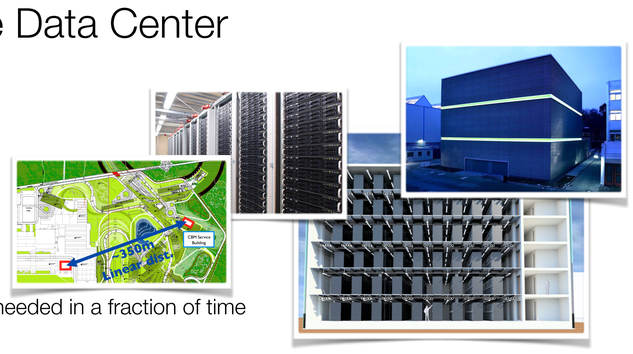
Online Event Selection

- Full online event reconstruction prior to selection
- High-throughput, up to 10^7 events/s
- No event separation by previous trigger
- Reconstruction in 4-D (including time)
- Extensive use of vectorization (SIMD) and many-core architectures (e.g., GPU)

Parallel speed-up of CBM reconstruction [2]

Location: FAIR Green-Cube Data Center

- Outline
  - 780 water-cooled racks in 3-D architecture
  - Max cooling power: 12 MW
  - Fully redundant (N+1), target PUE: 1.05
- Ideal environment for FLES
  - Cost-efficient infrastructure sharing
  - Maximum FLES online computing power only needed in a fraction of time → combine and share computing resources
- Fiber lengths to experiment site exceeding 500 m



- In operation since Jan 2016 at GSI

References

[1] J de Cuveland et al, 2011 J. Phys.: Conf. Ser. 331 022006

[2] I Kisel and V Akishina, RT-2014 *Online 4-Dimensional Event Building in the CBM Experiment*

FIAS Frankfurt Institute for Advanced Studies

Bundesministerium für Bildung und Forschung