Online computing architecture for the CBM experiment at FAIR

Jan de Cuveland
cuveland@compeng.uni-frankfurt.de

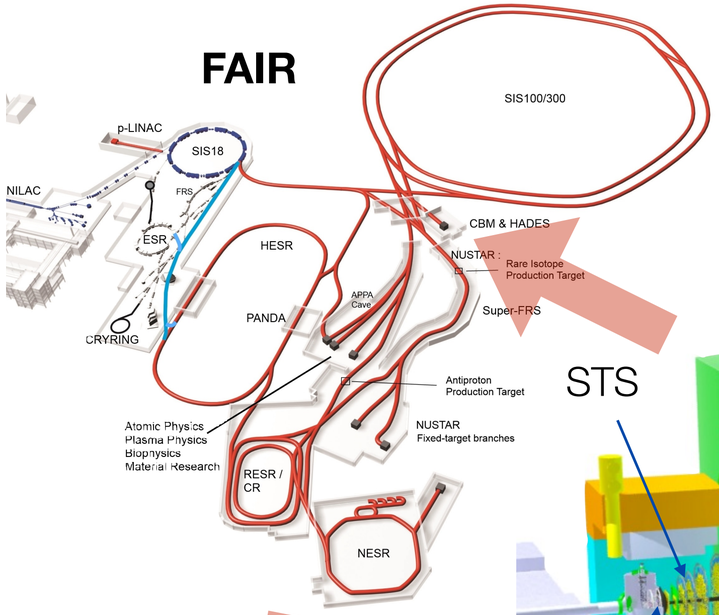
Prof. Dr. Volker Lindenstruth
FIAS Frankfurt Institute for Advanced Studies
Goethe-Universität Frankfurt am Main, Germany

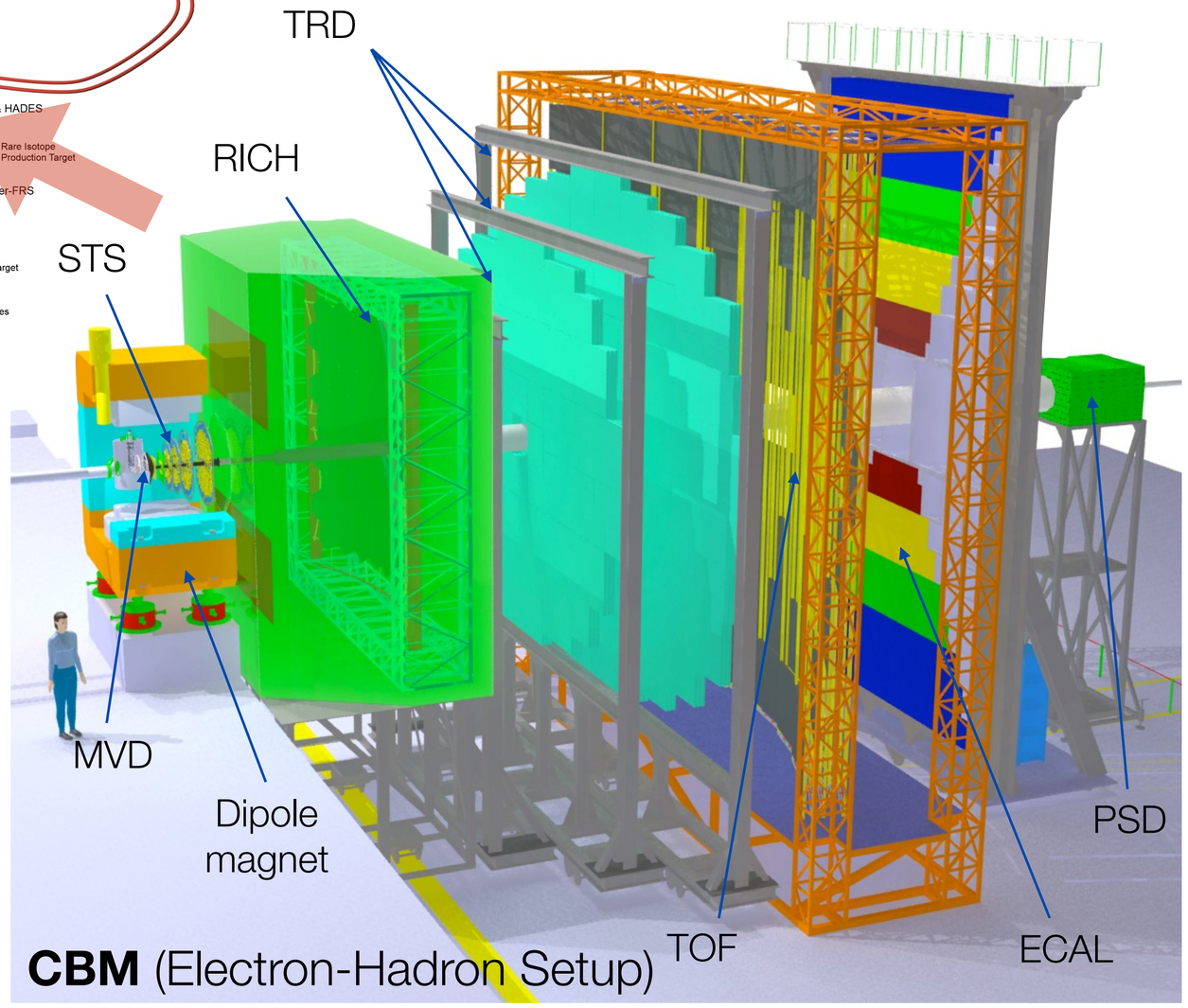
SPONSORED BY THE
Federal Ministry of Education and Research

CHEP 2016 Conference

2016-10-11 in San Francisco, USA

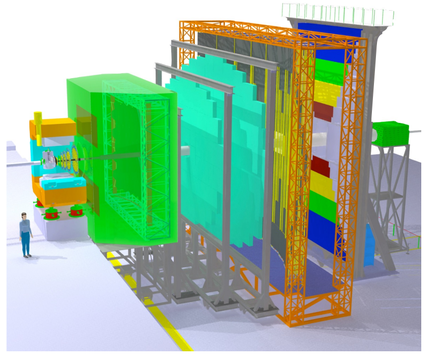
The CBM Experiment at FAIR



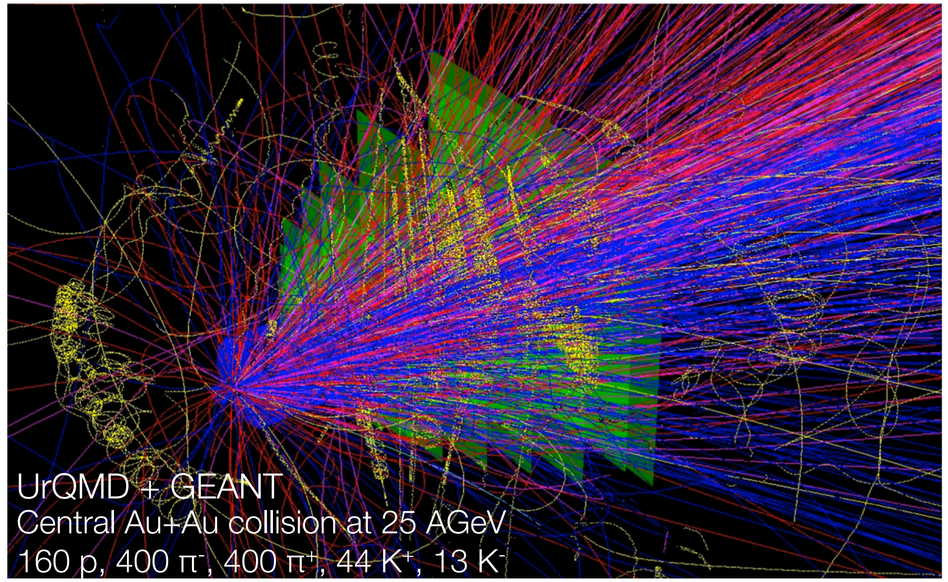


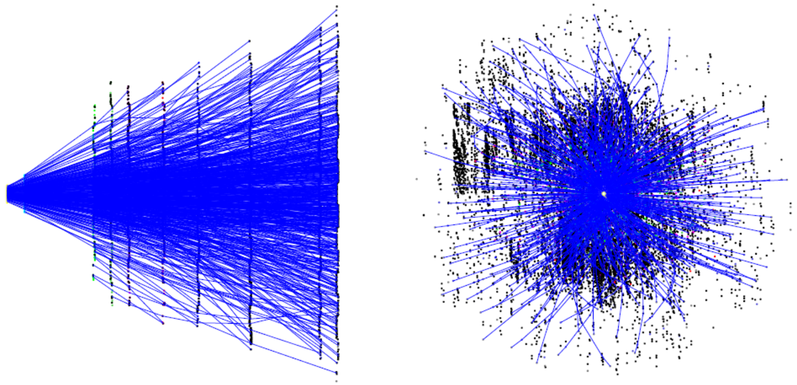
- Fixed target heavy ion experiment
- Under construction (time frame similar to LHC Run3)
- Physics goal: exploration of the QCD phase diagram at highest baryon densities and moderate temperatures
- E_{kin} = 2.0 – 35 A GeV
  $\sqrt{s_{NN}}$ = 2.7 – 8.3 GeV
- 10^5 – 10^7 Hz interaction rates
- Modular detector setup

CBM Challenges



CBM Setup





Reconstructed event in STS

- 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

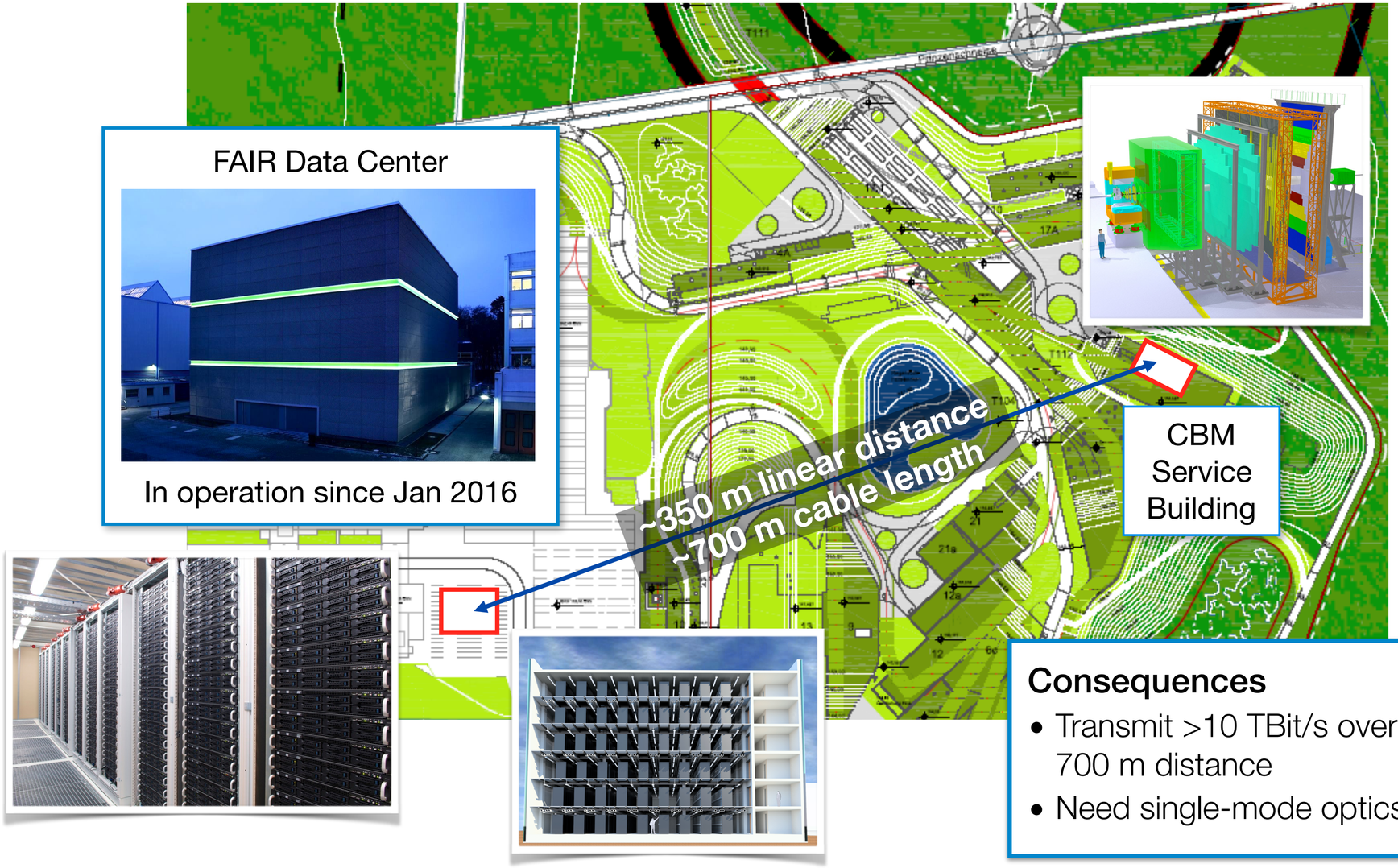
  → Self-triggering readout electronics

- **Full online event reconstruction needed**

  → Event selection exclusively done in a high-performance computing cluster

FAIR Data Center "Green-IT Cube"

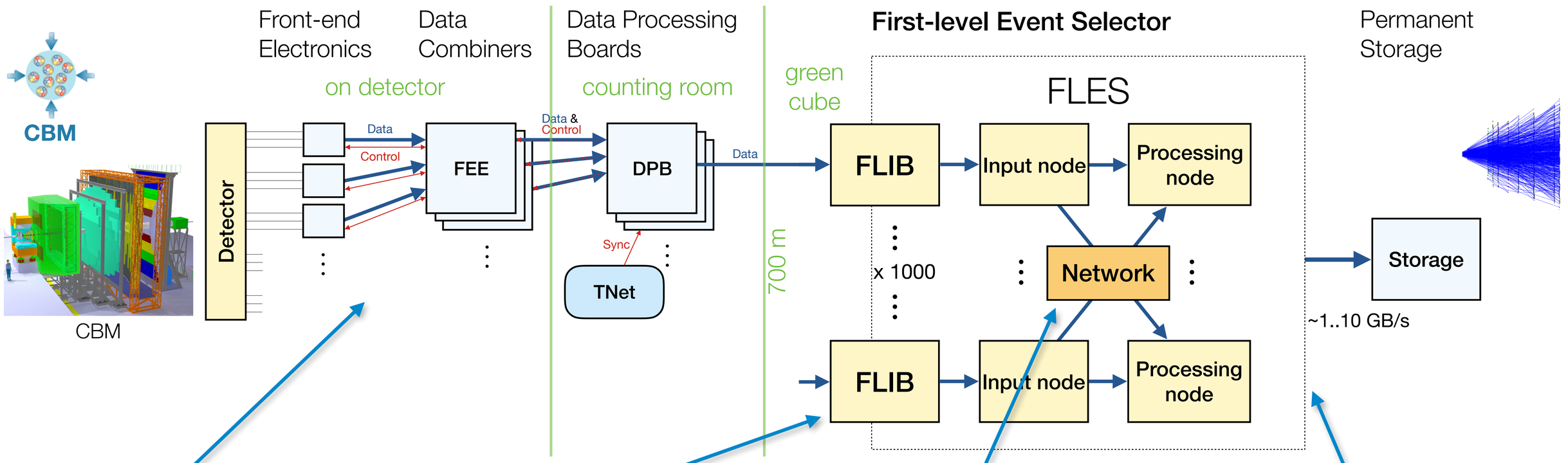
- Outline
  - 780 water-cooled racks in 3-D architecture
  - Max cooling power: 12 MW
  - Fully redundant (N+1), target PUE: 1.05
- Location of CBM online computing
  - Cost-efficient infrastructure sharing
  - Maximum CBM online computing power only needed in a fraction of time → combine and share computing resources
- Fiber lengths to experiment site approx. 700 m



Consequences

- Transmit >10 TBit/s over 700 m distance
- Need single-mode optics

CBM Readout Structure



Detector Front-ends

- Self-triggering front-end
- **10^7 events/s**
- Data push architecture
- All hits shipped to FLES

FLES Input Interface

- **FPGA**-based PCIe board
- Long-distance links to front-end
- Preprocessing and indexing for timeslice building

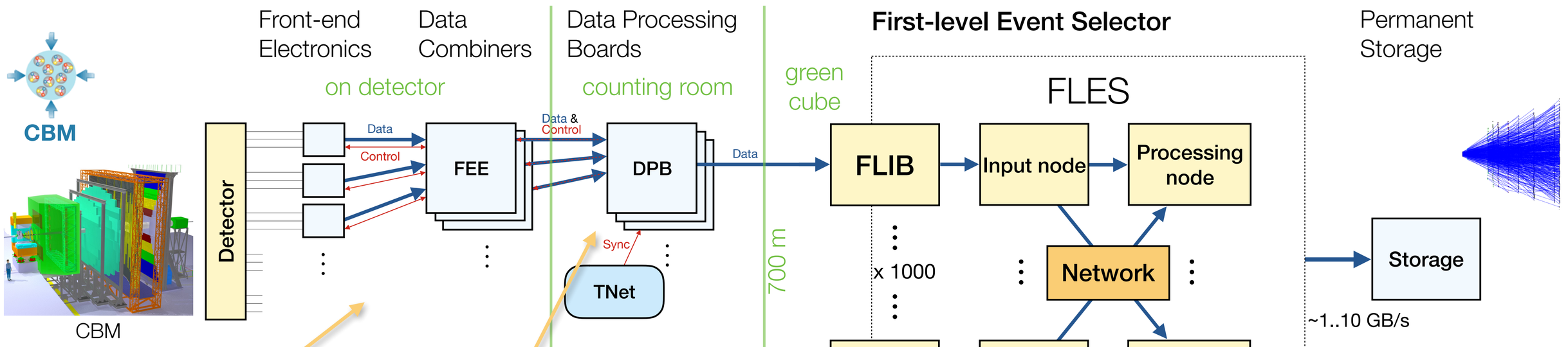
High-throughput event building

- **>1 TByte/s** input data rate
- ~ 1000 input streams
- RDMA-enabled network
- Deliver **global timeslices** to reconstruction code

Online Event Selection

- HPC processor farm with FPGAs, GPUs and fast interconnect
- ~ **60.000 cores**
- Fast, vectorized many-core track reconstruction algorithms
- Full event reconstruction

CBM Online Computing



Detector Front-ends

- Autonomous hit detection and **zero-suppression**
- Associate **time stamp** with each hit, aggregate data

Data Processing Board (DPB)

- Local data **preprocessing**: Feature extraction, time sort messages, data reformatting, merging input streams
- Convert to **global time**

FLES Interface Board (FLIB)

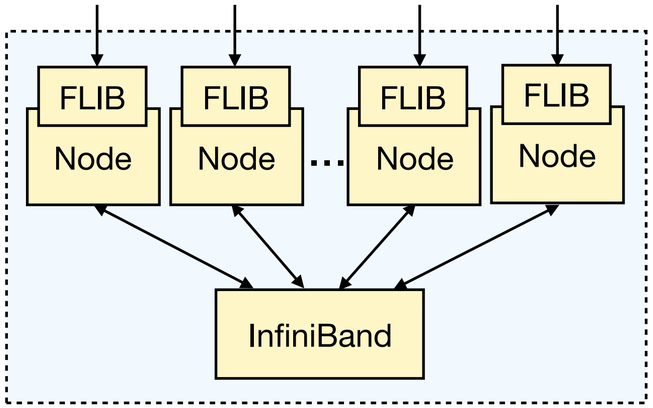
- **Time indexing** and buffering of microslices

FLES Nodes

- Calibration and global feature extraction
- **Tracking in 4 dimensions** (including time)
- Full reconstruction, associate hits with events
- Identification of leptons and hadrons
- High-precision vertex reconstruction
- Event selection

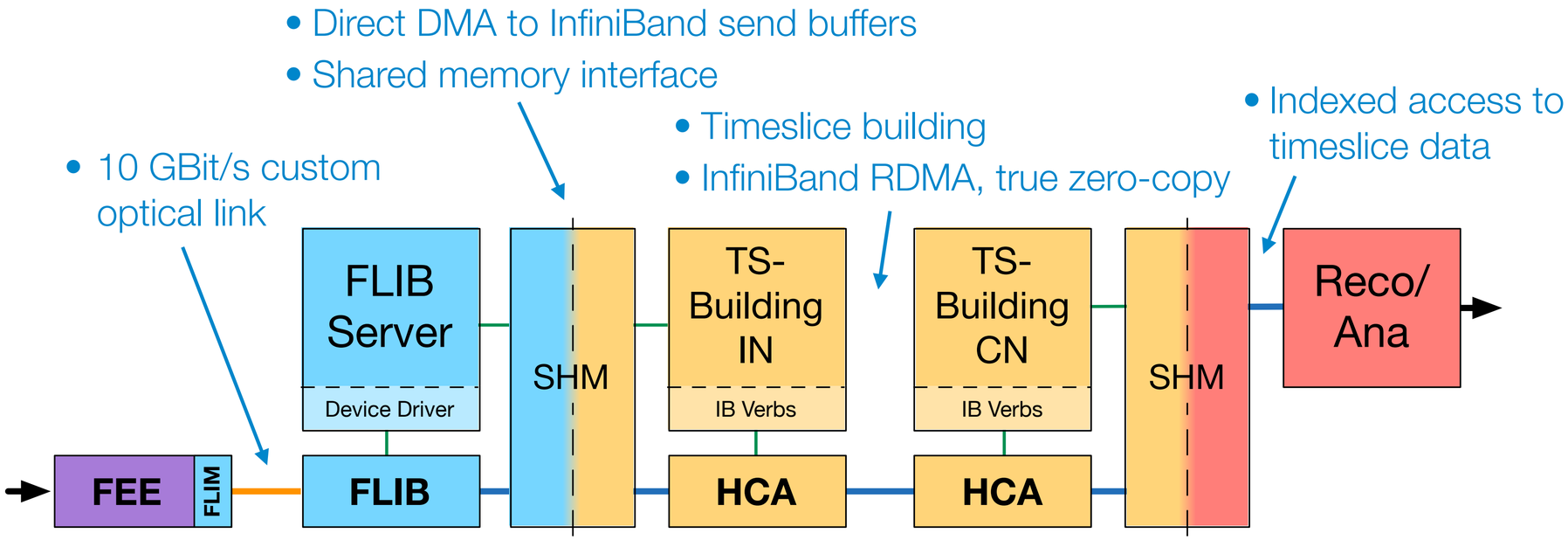
First-level Event Selector (FLES) Architecture

- 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 timeslice building
  - RDMA data transfer, very convenient for timeslice building
- Flat structure w/o dedicated input nodes
  Inputs are distributed over the cluster
  - Makes use of full-duplex bidirectional InfiniBand bandwidth
  - Input data is concise, no need for processing before timeslice building
- Decision on actual commodity hardware components as late as possible
  - First phase: full input connectivity, but limited processing and networking



FLES Data Management Framework

- RDMA-based timeslice building *(flesnet)*
- Works in close conjunction with FLIB hardware design
- Paradigms:
  - Do not copy data in memory
  - Maximize throughput
- Based on microslices, configurable overlap
- Delivers fully built timeslice to reconstruction code

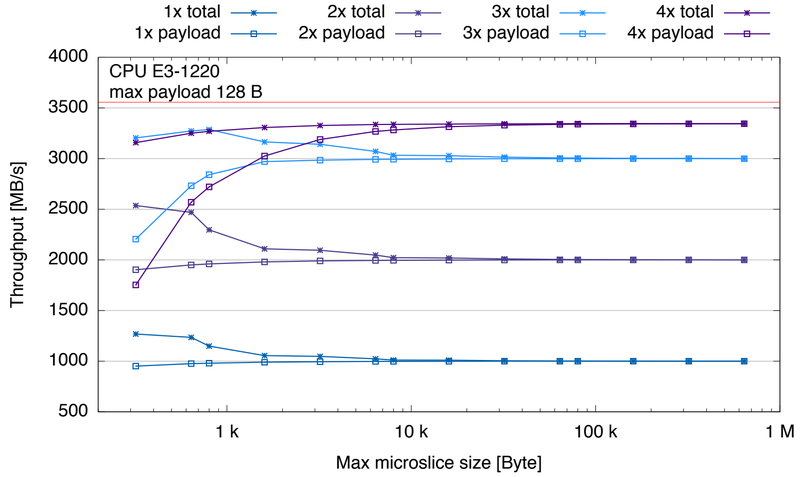


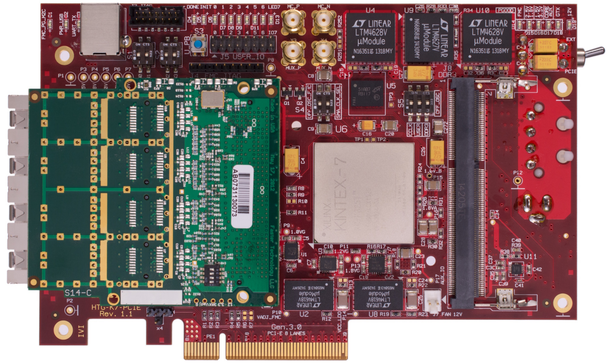
- Prototype implementation available
  - C++, Boost, IB verbs
- Measured flesnet timeslice building (8+8 nodes, including ring buffer synchronization, overlapping timeslices):
  - ~5 GByte/s throughput per node
- **Prototype software successfully used in several CBM beam tests**

FLES Input Interface

- FPGA-based PCIe board: FLIB
  - Consumes microslices received from DPBs
  - Prepares and indexes microslices for timeslice building
  - Transfers microslices and index data to PC memory
- Custom PCIe DMA interface
- Optimized data scheme for zero-copy timeslice building
- Common HDL interface module in front-end
- Status
  - Complete **design available**, implemented on HTG-K7 development board
  - Combined FLIB and DPB functionalities for beam test usage available
  - Successfully used in numerous setups

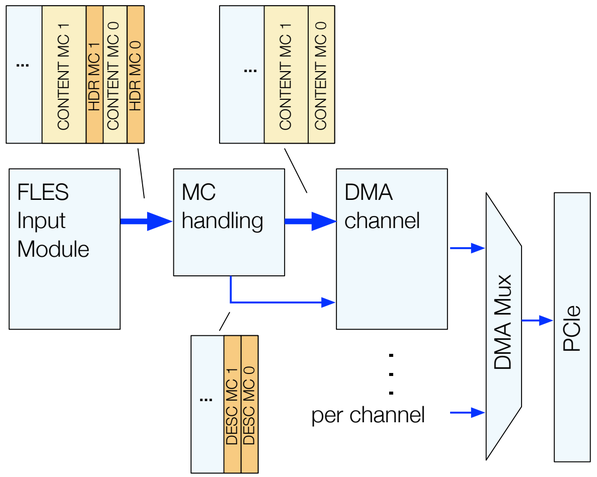
Measured FLIB PCIe throughput

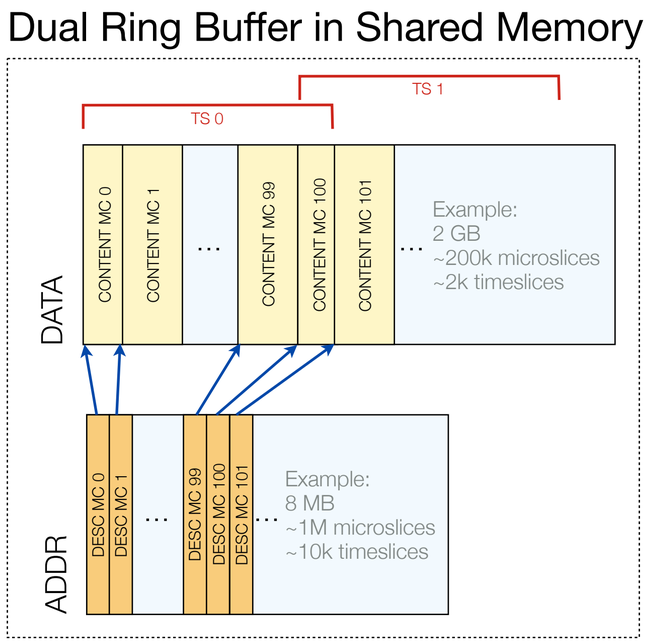




Poster presentation today at 15:30 by **D. Hutter**

FLES Input Data Path

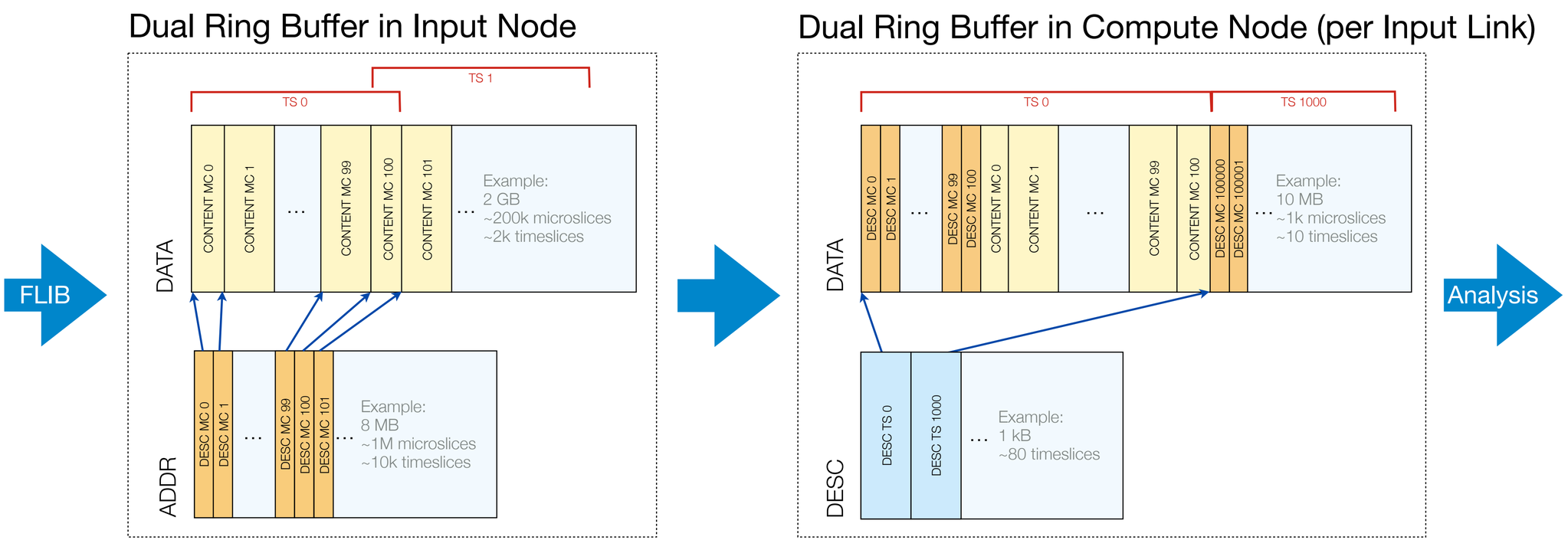




Microslice
- Timeslice substructure
- Constant in experiment time
- Allow overlapping timeslices

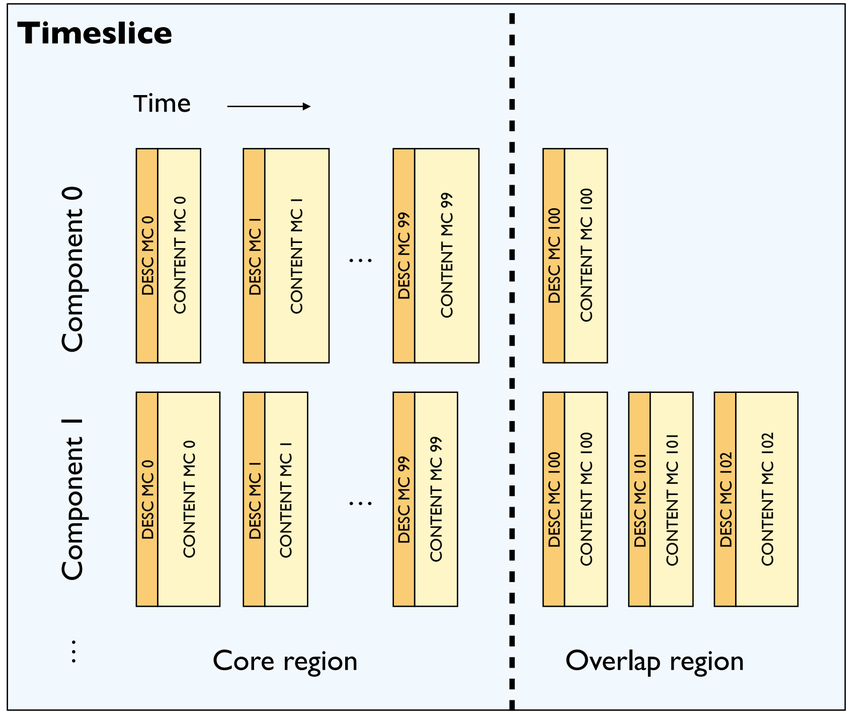
- Full offload DMA engine
- Transmit microslices via PCIe/DMA directly to userspace buffers
  - Buffer placed in Posix shared memory, can be registered in parallel for InfiniBand RDMA
- Pair of ring buffers for each link
  - Data buffer for microslice data content
  - Descriptor buffer for index table and microslice meta data

RDMA Timeslice Building



- Two pairs of ring buffers for each input link
  - Second buffer: index table to variable-sized data in first buffer
- Copy contiguous block of microslices via RDMA (exception: borders)
- Lazy update of buffer status between nodes, reduce transaction rate

Interface to Online Reconstruction Code

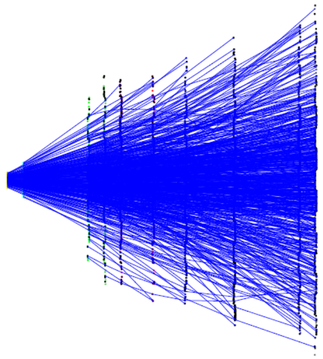


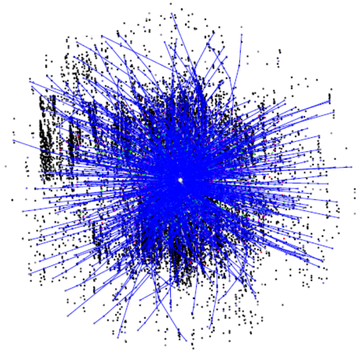
Timeslice
- Two-dimensional indexed access to microslices
- Overlap according to detector time precision
- Interface to online reconstruction software

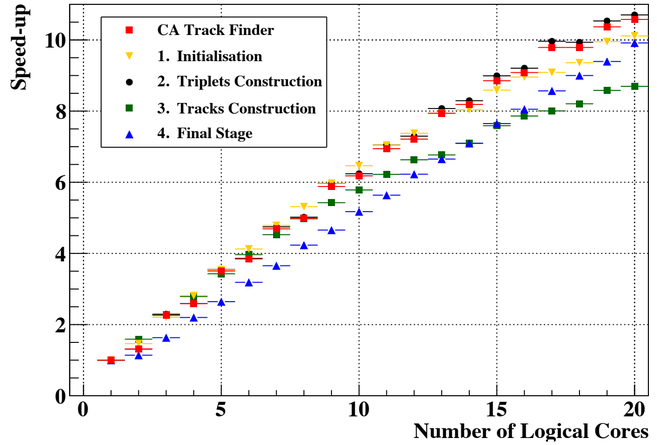
- Basic idea: For each timeslice, an instance of the reconstruction code...
  - ...is given direct **indexed access** to all corresponding data
  - ...uses **detector-specific** code to understand the **contents** of the microslices
  - ...applies **adjustments** (fine calibration) to detector time stamps if necessary
  - ...finds, **reconstructs and analyzes** the contained events
- Timeslice data management concept
  - Timeslice is self-contained
  - Calibration and configuration data distributed to all nodes
  - **No network communication** required during reconstruction and analysis

Online Event Selection

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







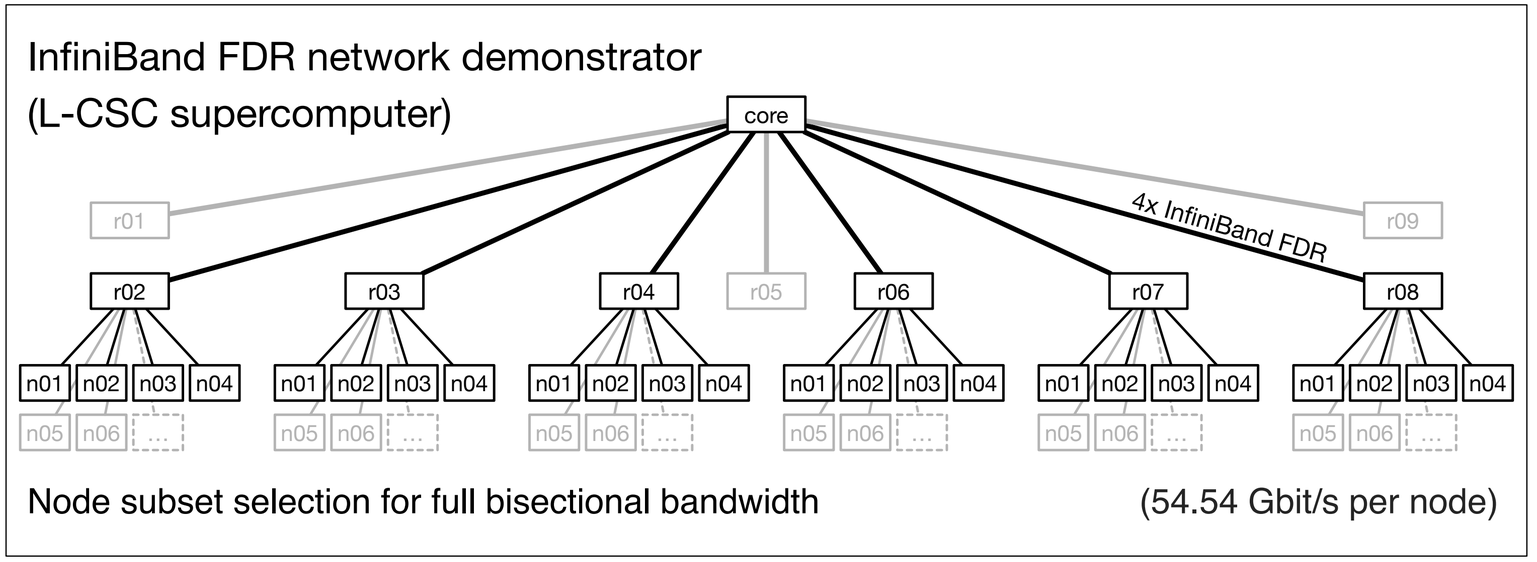
Parallel speed-up of CBM reconstruction [I. Kisel]

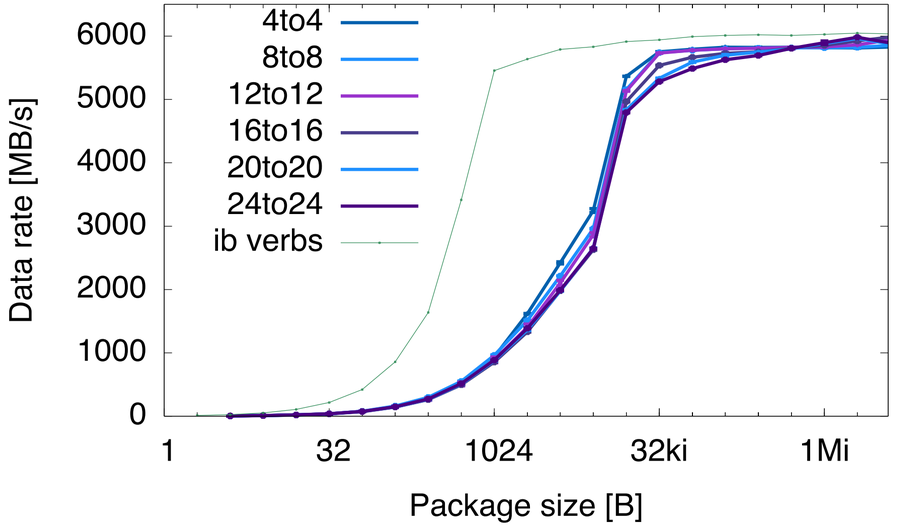
Massive parallelization

- Many independent processing nodes
- Multiple timeslices simultaneously per node
- Multi-threaded, vectorized reconstruction code

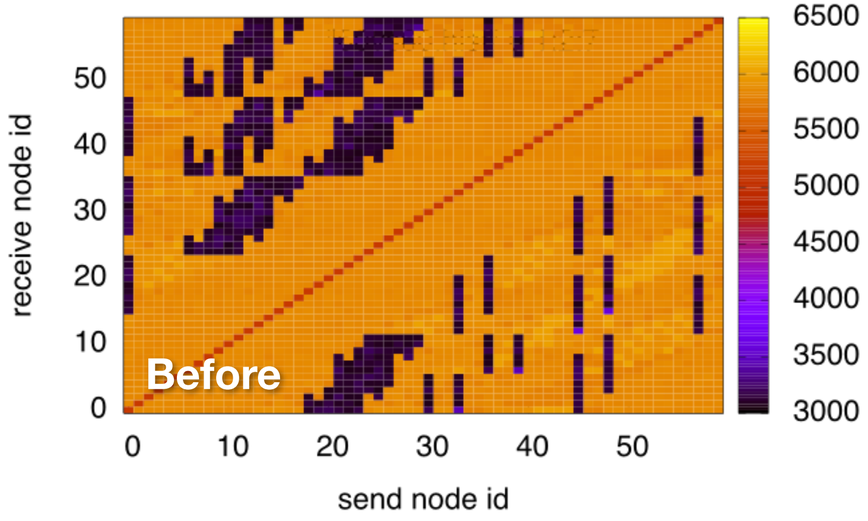
Provided by framework

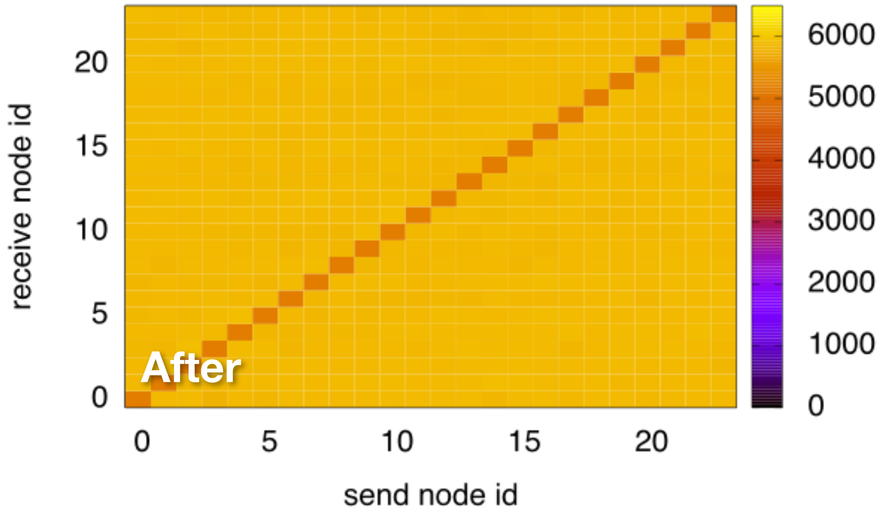
FLES Network Performance Study



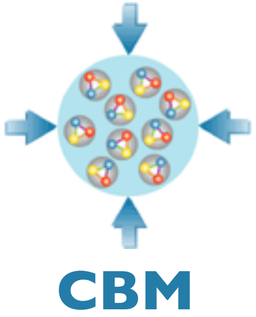


- Standard routing pattern suboptimal for continuous all-to-all communication
- **Optimized routing scheme** leads to excellent performance (>5 GB/s per node) (tested on for 24 nodes using InfiniBand verbs and custom MPI benchmark)



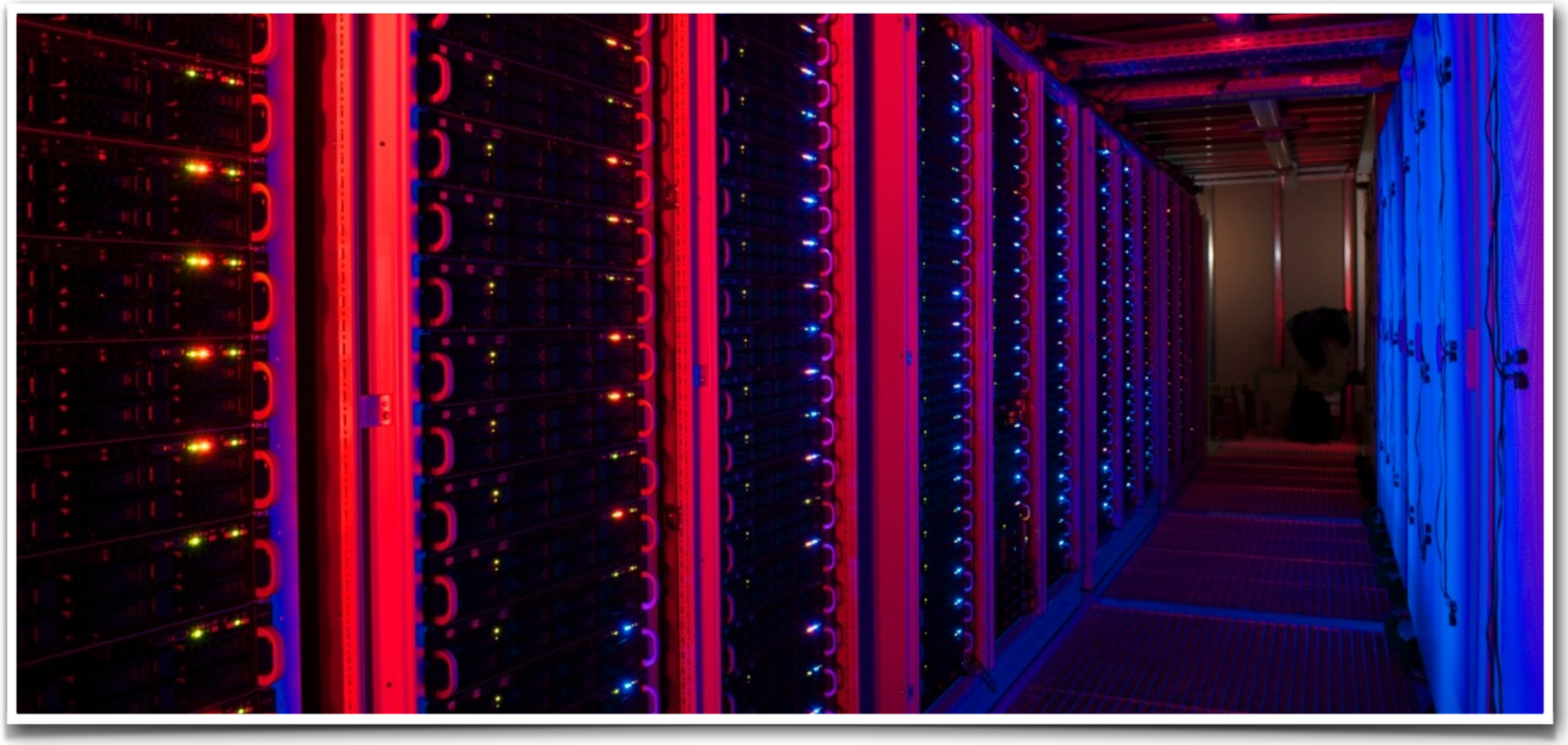


Summary



- Compressed Baryonic Matter (CBM) experiment at FAIR
  - High event rates (10^7 Hz), complex global triggers
  - Self-triggered detector front-ends
  - Data push readout architecture
- Central physics selection system: First-Level Event Selector (FLES)
  - HPC processor farm including FPGAs (at input stages) and heterogenous many-core architectures (e.g., GPUs)
  - >1 TByte/s input data stream
  - Timeslice building in RDMA-enabled network
  - 4-D event reconstruction using fast, vectorized track reconstruction algorithms
- Online computing architecture – status
  - Demonstrator implementations available, data chain field-tested in beam tests
  - Architecture still being refined towards final system
  - Aim for first phase: full input connectivity, but limited processing and networking

Thanks for your attention



SPONSORED BY THE

Federal Ministry
of Education
and Research

Jan de Cuveland
cuveland@compeng.uni-frankfurt.de

CBM