Results from the CBM mini-FLES Online Computing Cluster Demonstrator

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

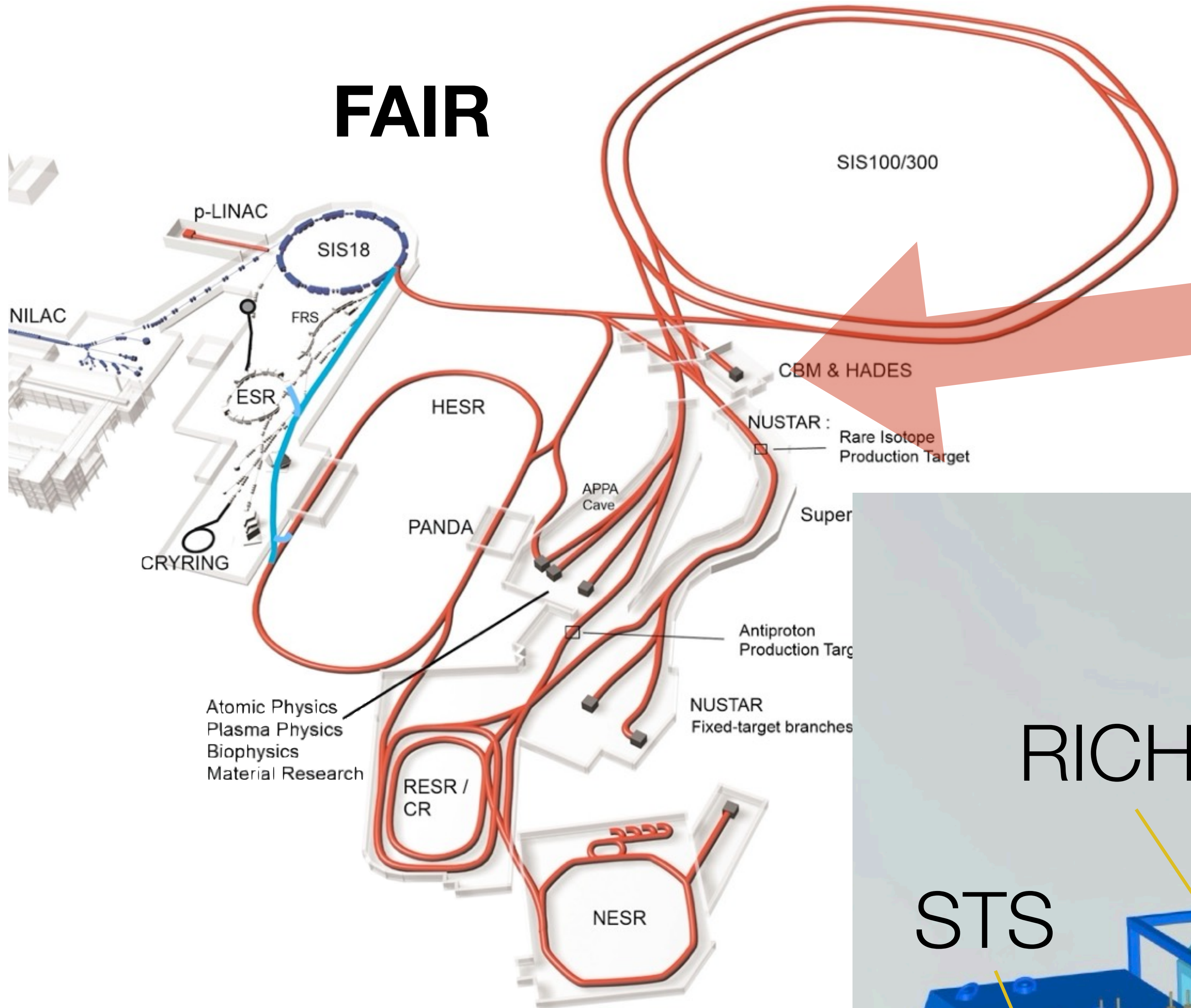
FIAS Frankfurt Institute for Advanced Studies
Goethe-Universität Frankfurt am Main, Germany

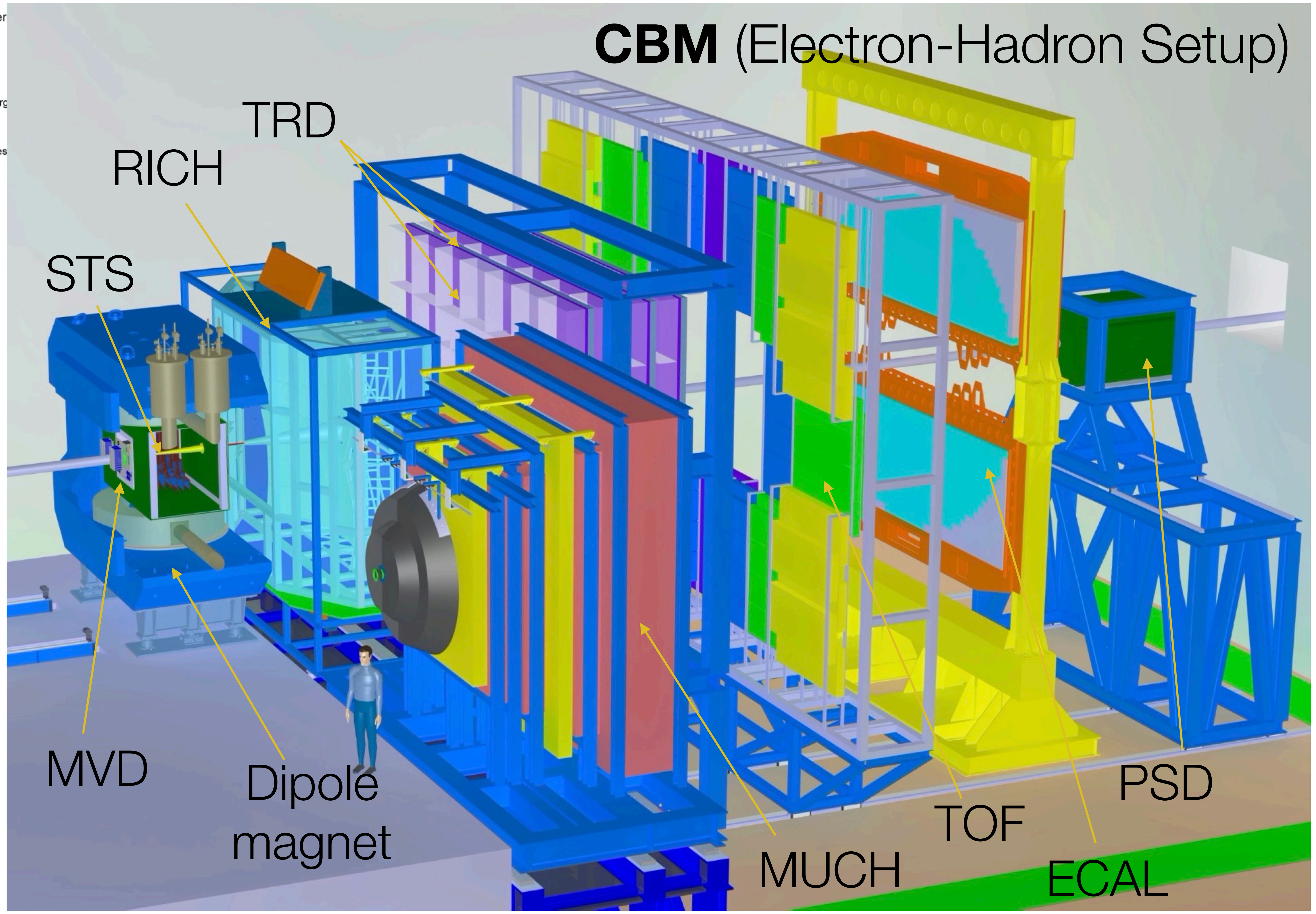
SPONSORED BY THE
Federal Ministry
of Education
and Research

CBM

CHEP 2019 Conference
2019-11-05 in Adelaide, Australia

The CBM Experiment at FAIR





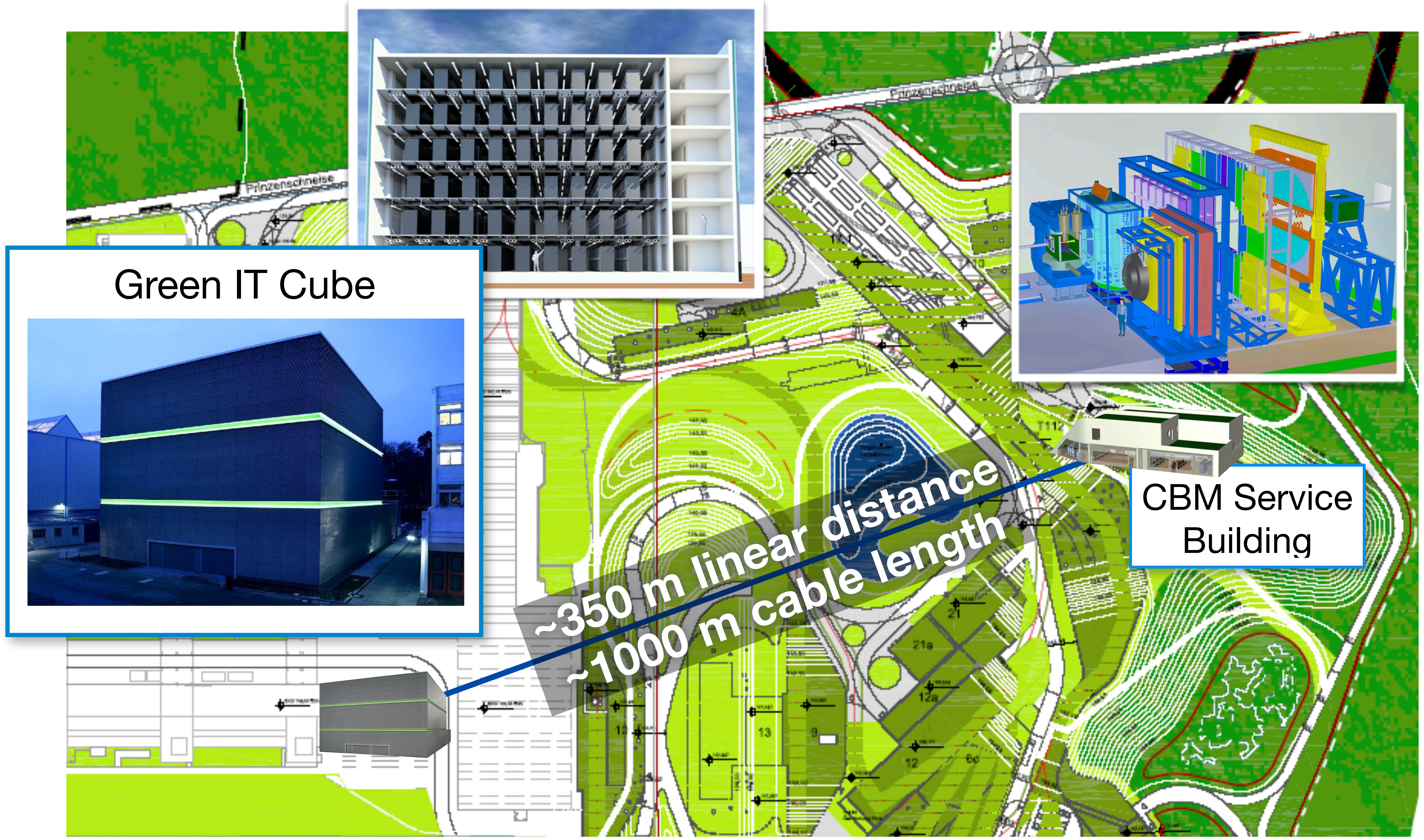
- Fixed target heavy ion experiment at FAIR
- Physics goal: exploration of the QCD phase diagram
- Complex (topological) trigger signatures
- Extreme reaction rates of up to **10 MHz** and track densities up to 1000 tracks in aperture
- Full **online event reconstruction** needed

➡ Self-triggering free-streaming readout electronics

➡ Event selection exclusively done in FLES HPC cluster

First-level Event Selector (FLES)

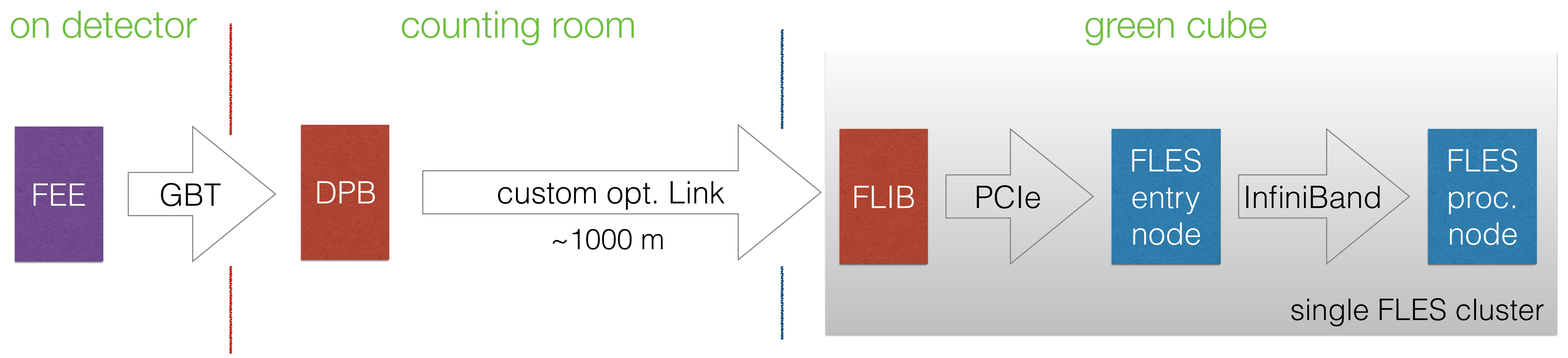
- FLES is designed as an **HPC cluster**
  - Commodity PC hardware
  - FPGA-based custom PCIe input interface
  - Total input data rate > 1 TByte/s
- Located in the Green IT Cube data center
  - Cost-efficient infrastructure sharing
  - Maximum CBM online computing power only needed in a fraction of time
→ combine and share computing resources



Consequences

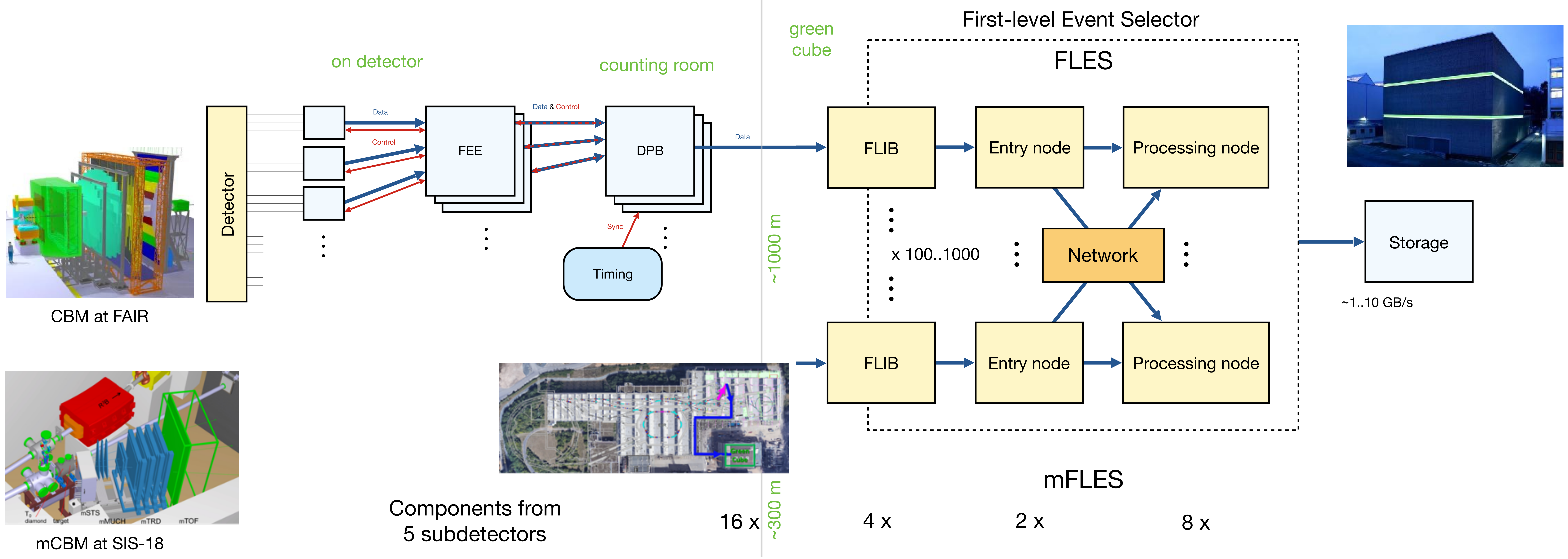
- Transmit 1 TByte/s over 1000 m distance
- Boundary condition for online computing architecture

CBM DAQ/FLES Architecture



- Initial DAQ/FLES architecture → basis for mini-CBM setup
  - **Single flat cluster** design
  - Two **FPGA**-based stages: Data Processing Board (DPB) and FLES Interface Board (FLIB)
  - Long-range connection to Green Cube via **custom optical links**
- Side note: test results with standard network components will allow a revised architecture
  - Long-range connection to Green Cube via **standard network** equipment (e.g., long-range InfiniBand)
  - Split computing into 2 **dedicated clusters**: entry cluster and compute cluster
  - **Combine DPB and FLIB** to single FPGA board (similar to ATLAS, LHCb and ALICE)

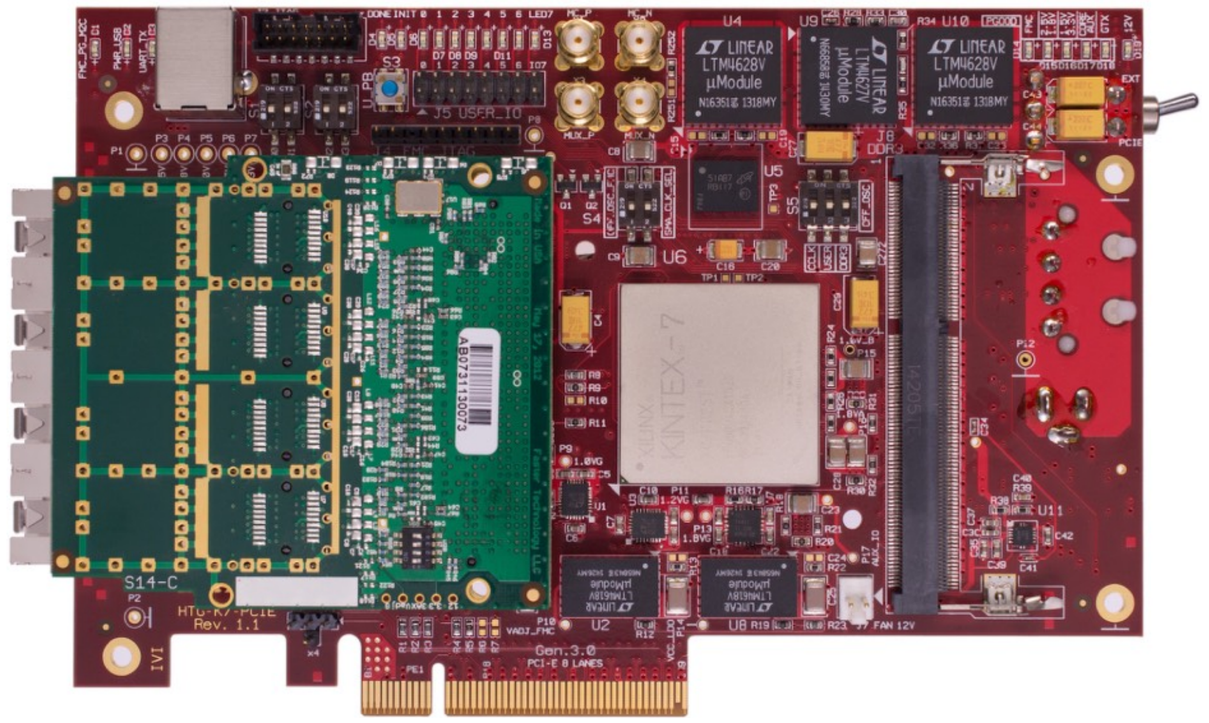
A slice of CBM: mini-CBM (mCBM)



- mCBM:
  - A **complete slice** of the full CBM system (hardware and software)
  - Study **integration** (and identify missing pieces)
  - Eventually, apply online analysis to live physics data
- mFLES:
  - **Online system demonstrator** with all data path components
  - Study integration and verify concepts
  - Extensive online monitoring (online reconstruction still WIP)
  - Hardware currently approx. 2 % of foreseen FLES system

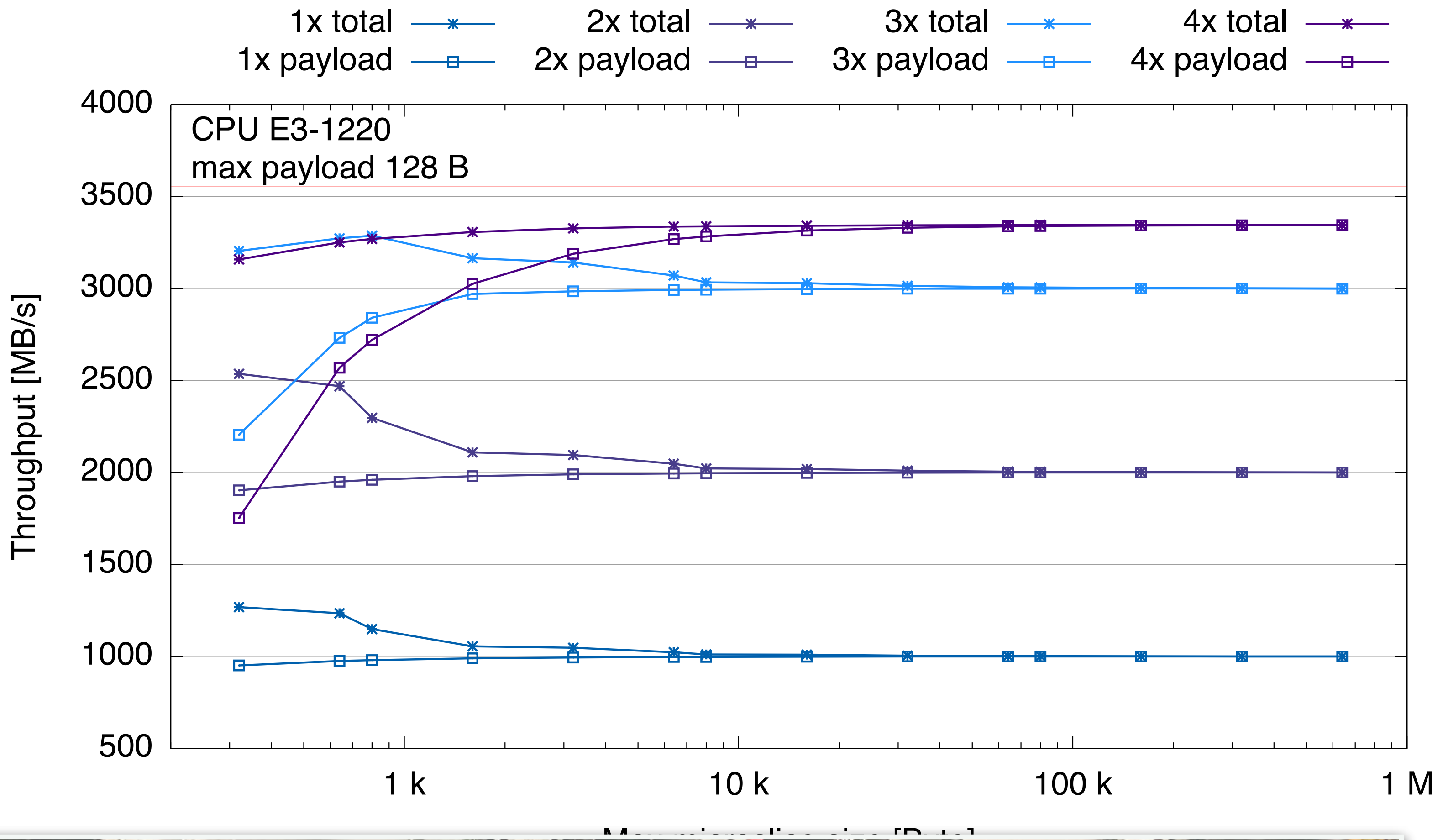
FLES Input Interface

- FPGA-based PCIe board: FLIB
  - Prepares and indexes data for timeslice building
  - Custom PCIe DMA interface, full offload engine



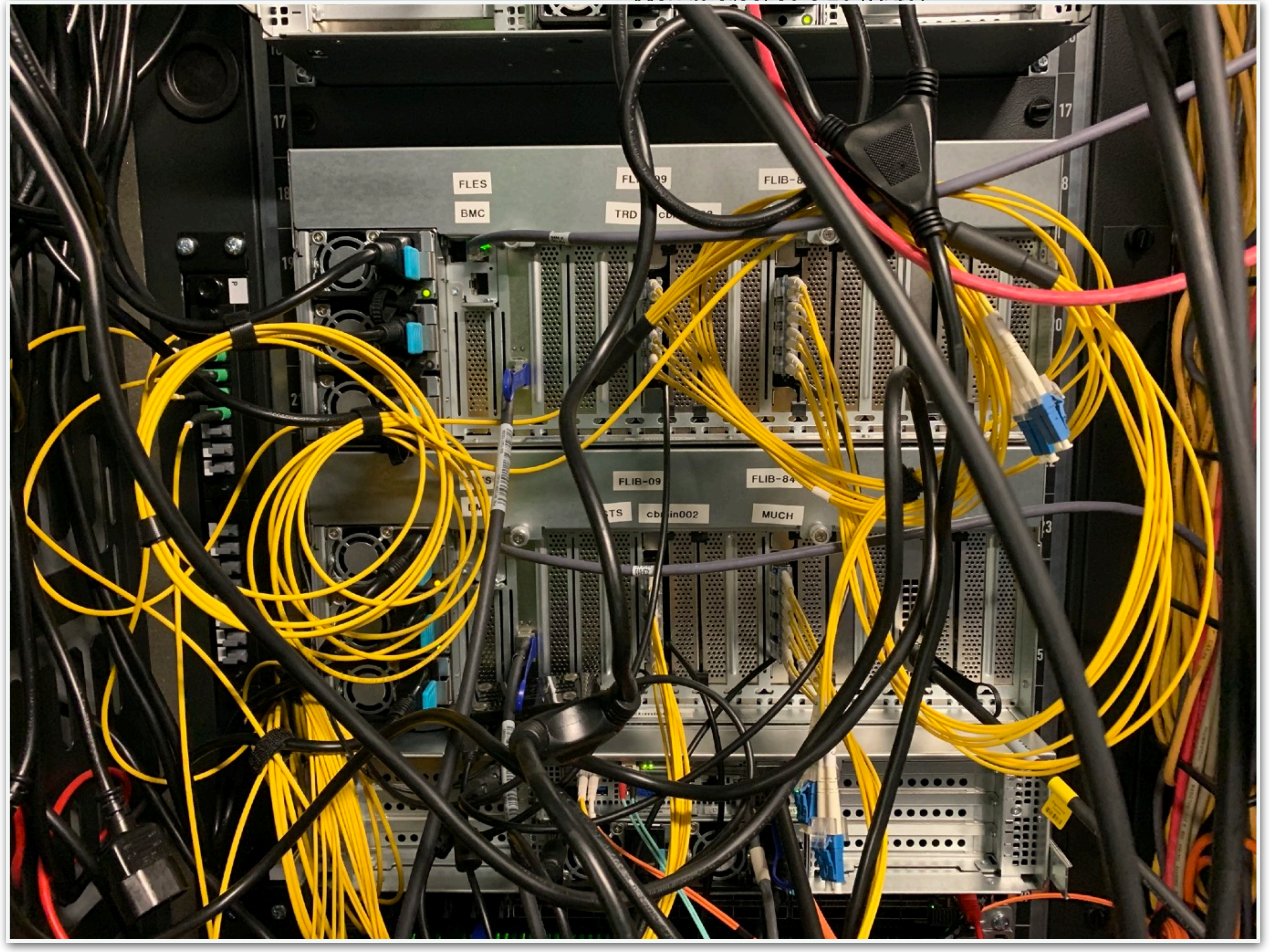
- Optimized data scheme for zero-copy timeslice building
  - Transmit microslices via PCIe/DMA directly to userspace buffers
  - Buffer placed in Posix shared memory, can be registered in parallel for InfiniBand RDMA

Measured FLIB PCIe throughput

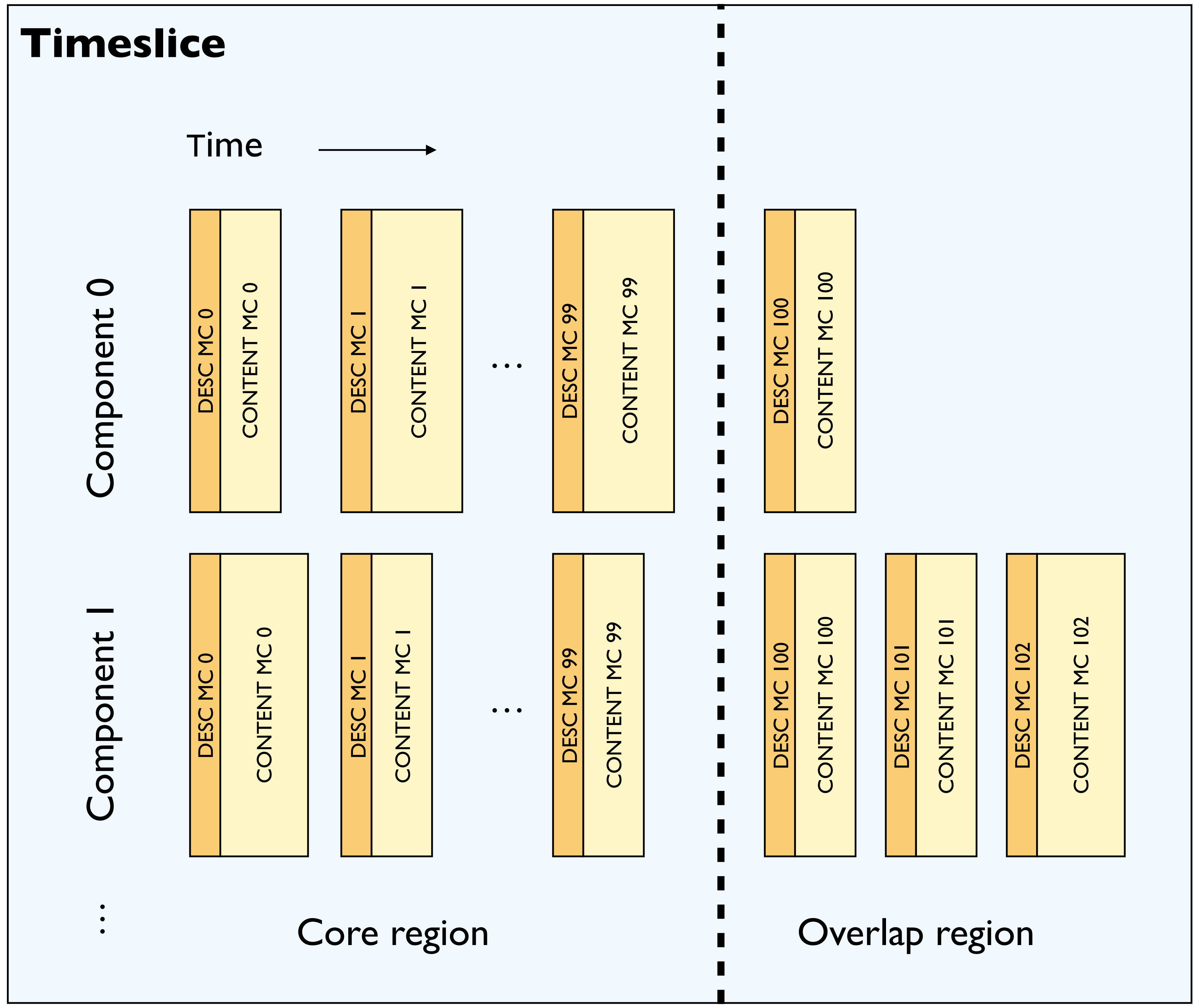


- mCBM: 4 FLIBs in 2 nodes
  - 12 input links connected to detectors
  - Implemented on HTG-K7 development boards

- Front-end interface employed at mCBM
  - Custom link, FLIM module
  - Input link commissioning with BER < 4.6e-16 (808 TB, 0 errors)



Timeslice Building and Online Analysis Interface



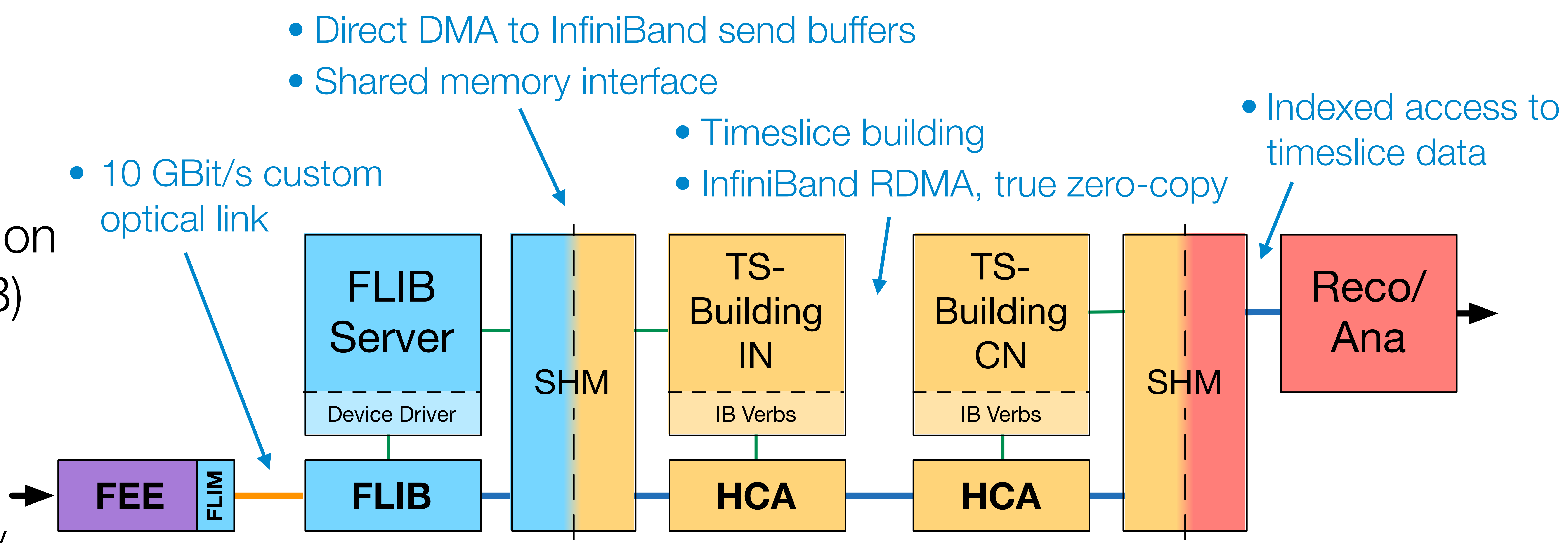
- Timeslice building: **combine matching time intervals** from all input links to one "timeslice" (processing interval)
- Distribute different timeslices to different processing nodes
- Timeslice data management concept
  - Timeslice is self-contained
  - Calibration and configuration data distributed to all nodes
  - **No network communication** required during reconstruction and analysis

Timeslice
- Two-dimensional indexed access to microslices
- Overlap allows limited timing calibration in front-end
- Interface to online reconstruction software

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

FLES Data Management Framework

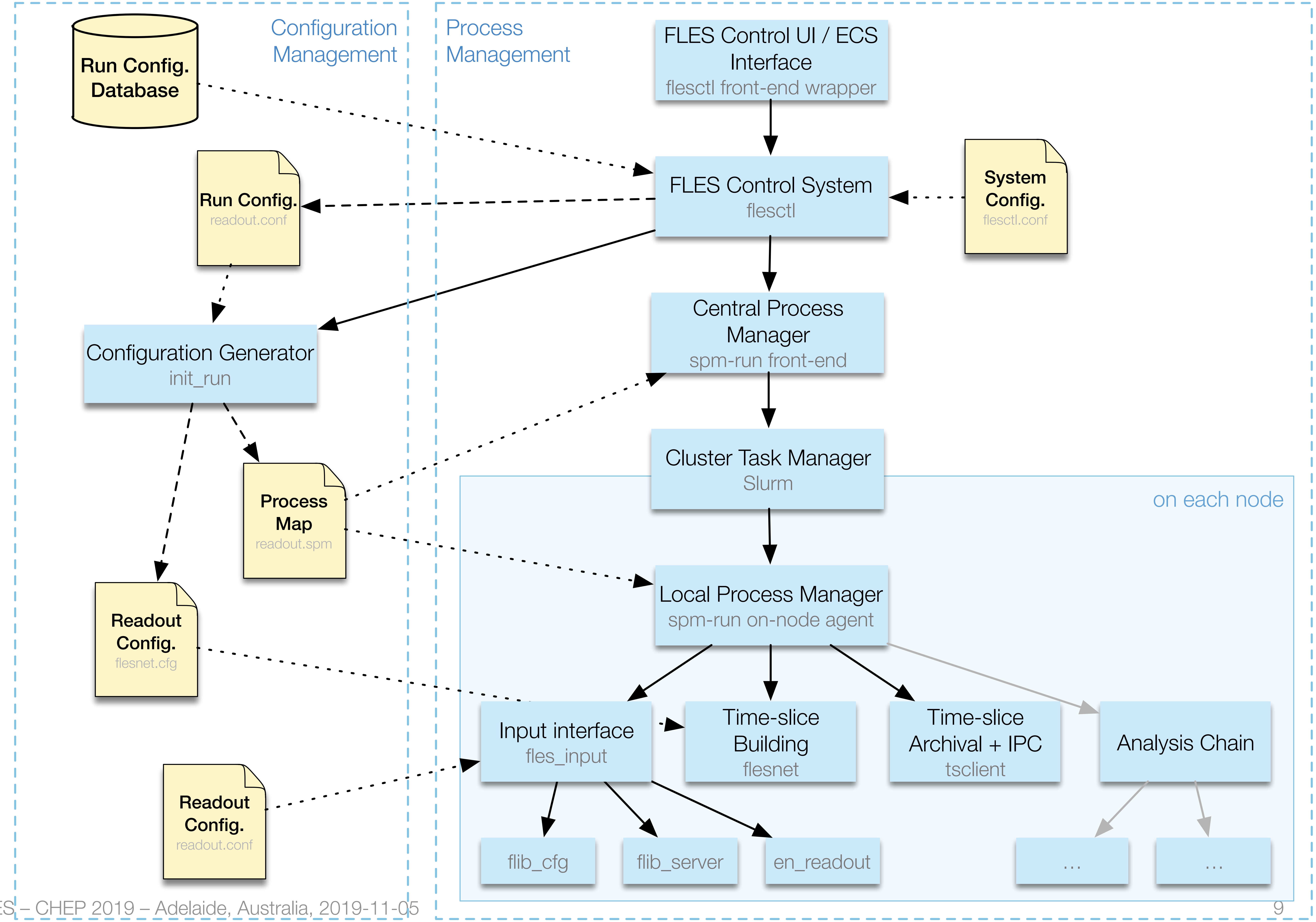
- RDMA-based timeslice building *(flesnet)*
- Works in close conjunction with input interface (FLIB) hardware design
- Paradigms:
  - Do not copy data in memory
  - Maximize throughput
- Based on microslices, configurable overlap
- Delivers fully built timeslice to reconstruction code



- Initial implementation of all components available
  - C++, Boost, IB verbs
  - Critical network performance optimized for > 1 TB/s
- Full data chain software employed at mCBM

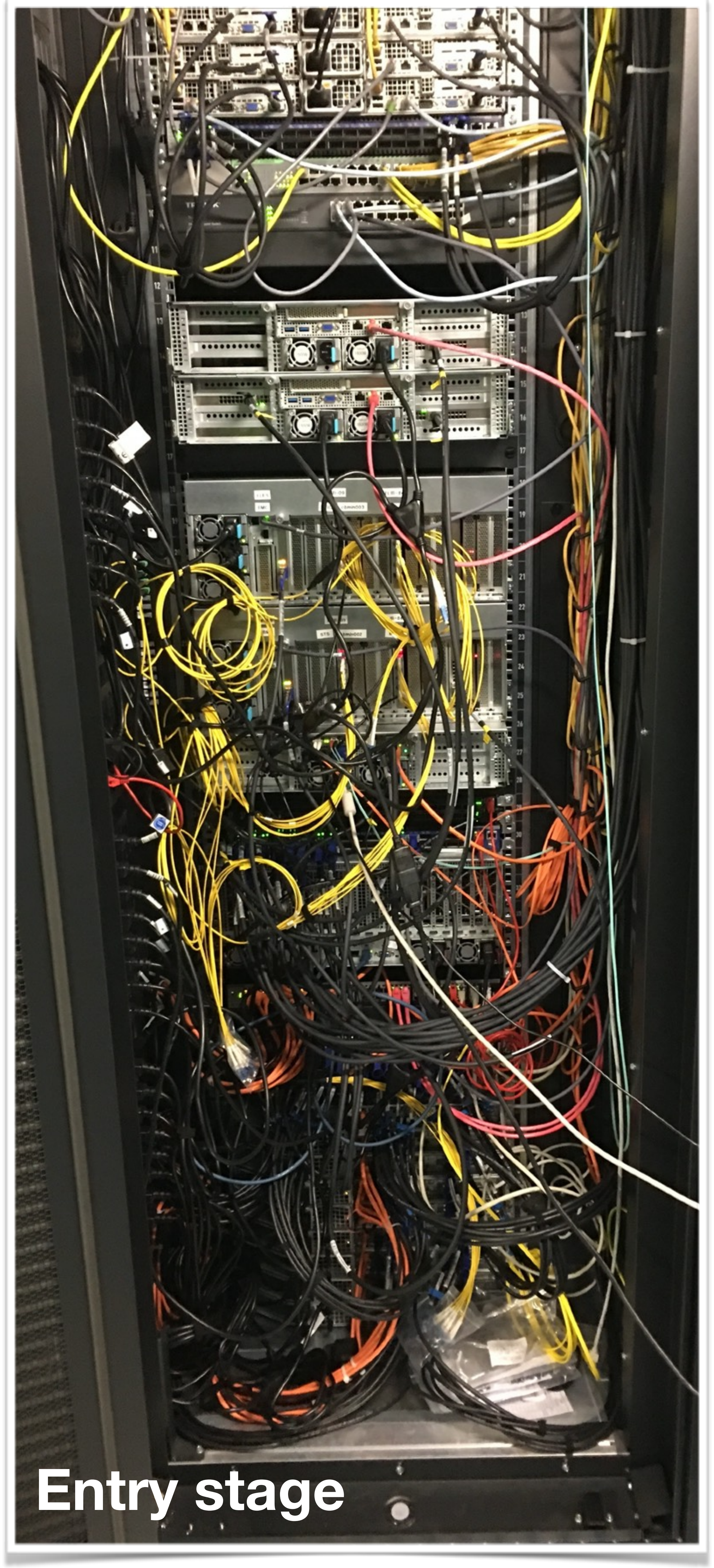
FLES Control

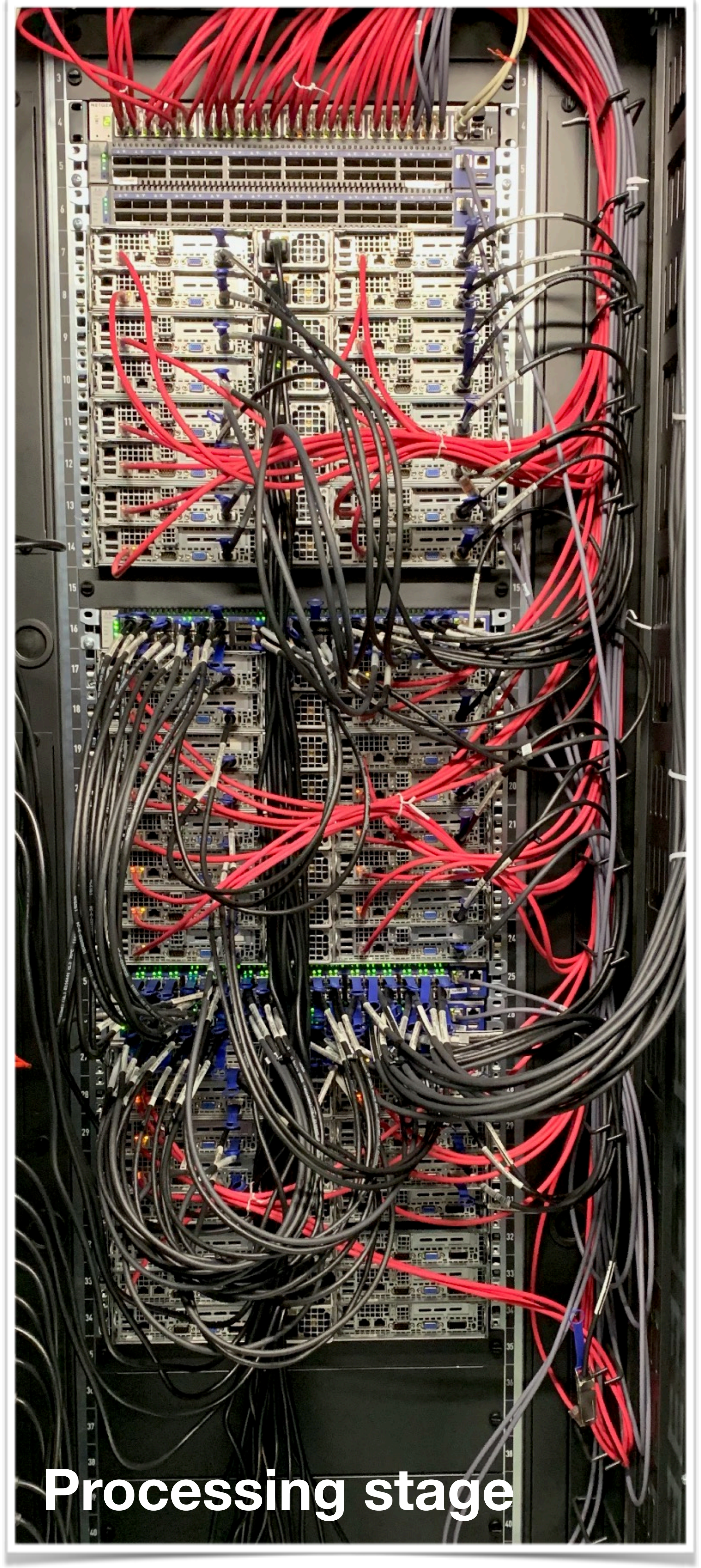
- Prototype **configuration** and **process management** on mFLES cluster
- Reproducible data taking on **multiple nodes**, timeslice building from EN to PN
- Successfully employed in all global mCBM runs



mFLES Setup and Functionalities Summary

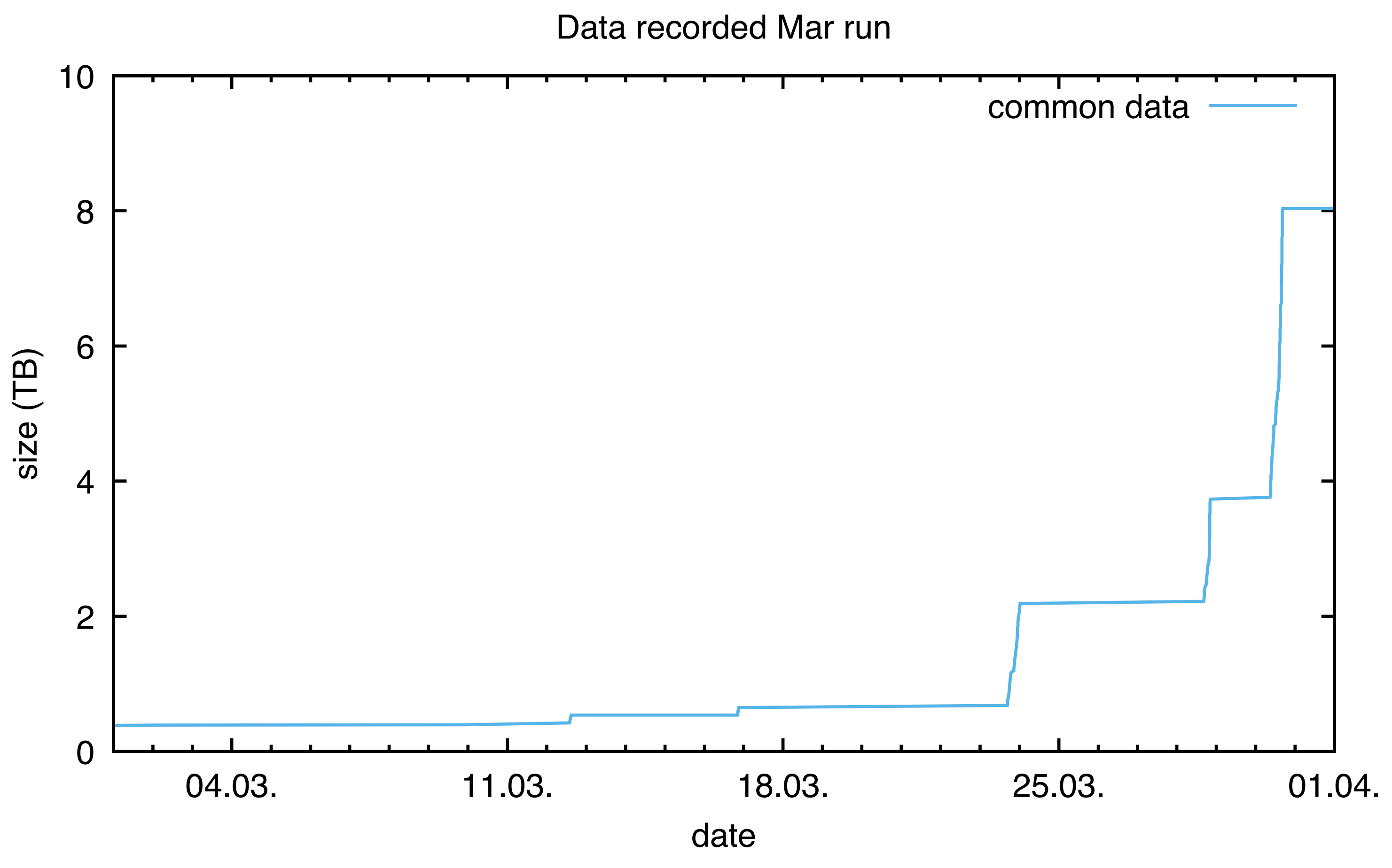
- mFLES setup
  - 4 FLIB input cards, **12 FLIM links** (of 16/32), 2 entry nodes, **3–10 processing nodes** (of 36), InfiniBand network
  - Single-mode links from mCBM to Green Cube
- mFLES software
  - **Distributed data taking**
  - Full flesnet chain with **timeslice building**
  - Automated run control with prototype configuration and process management

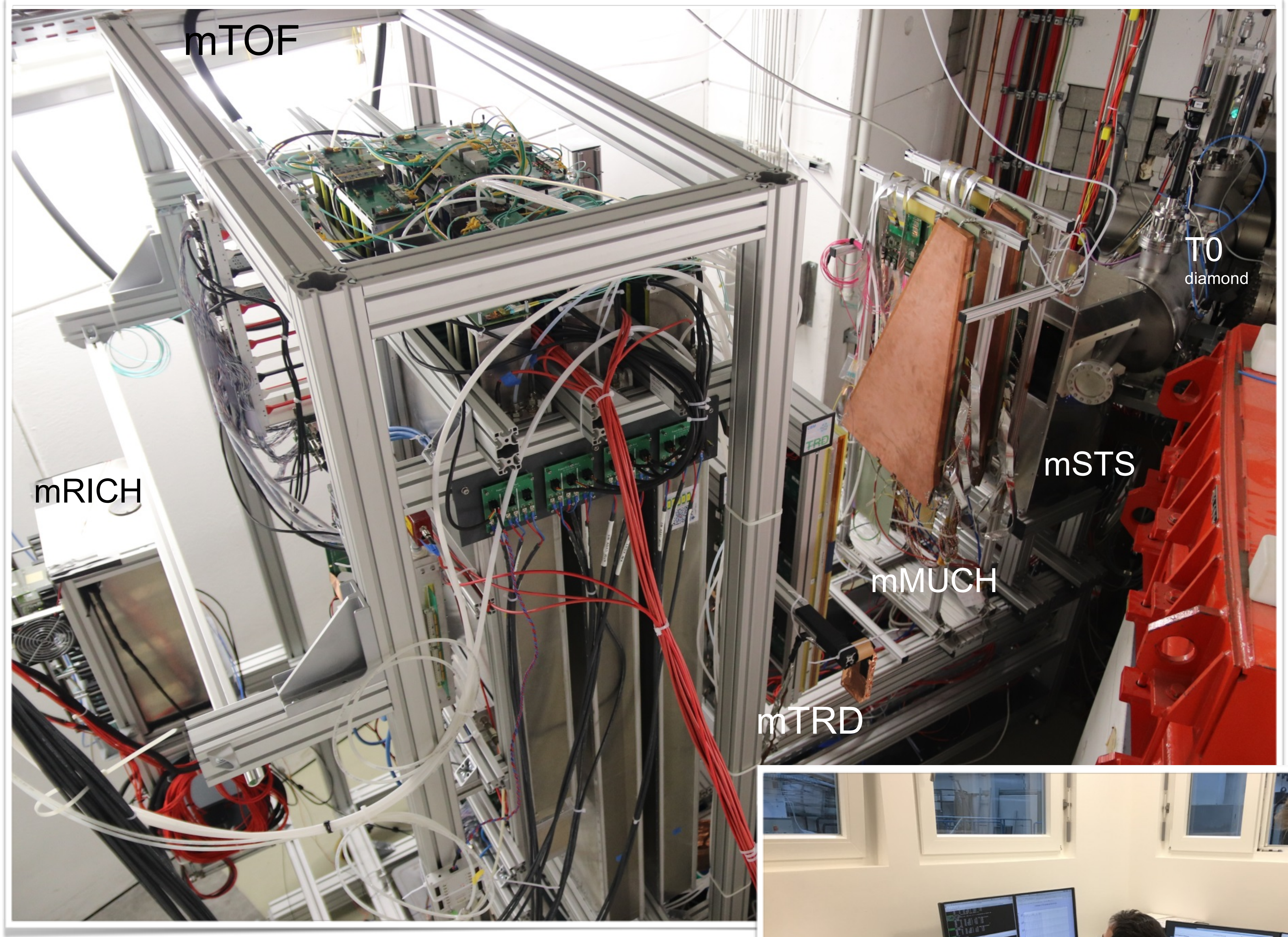




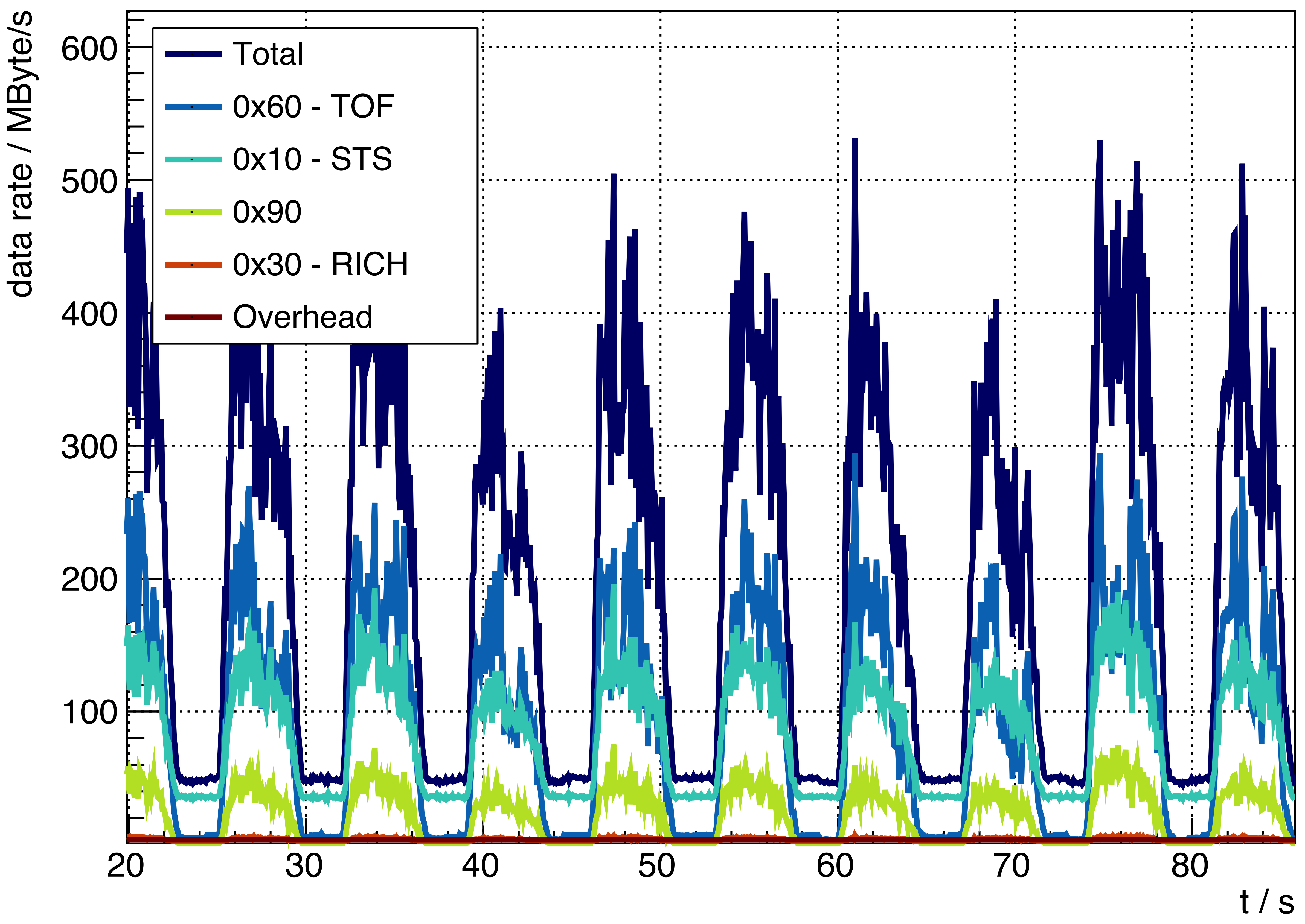
mCBM Campaigns

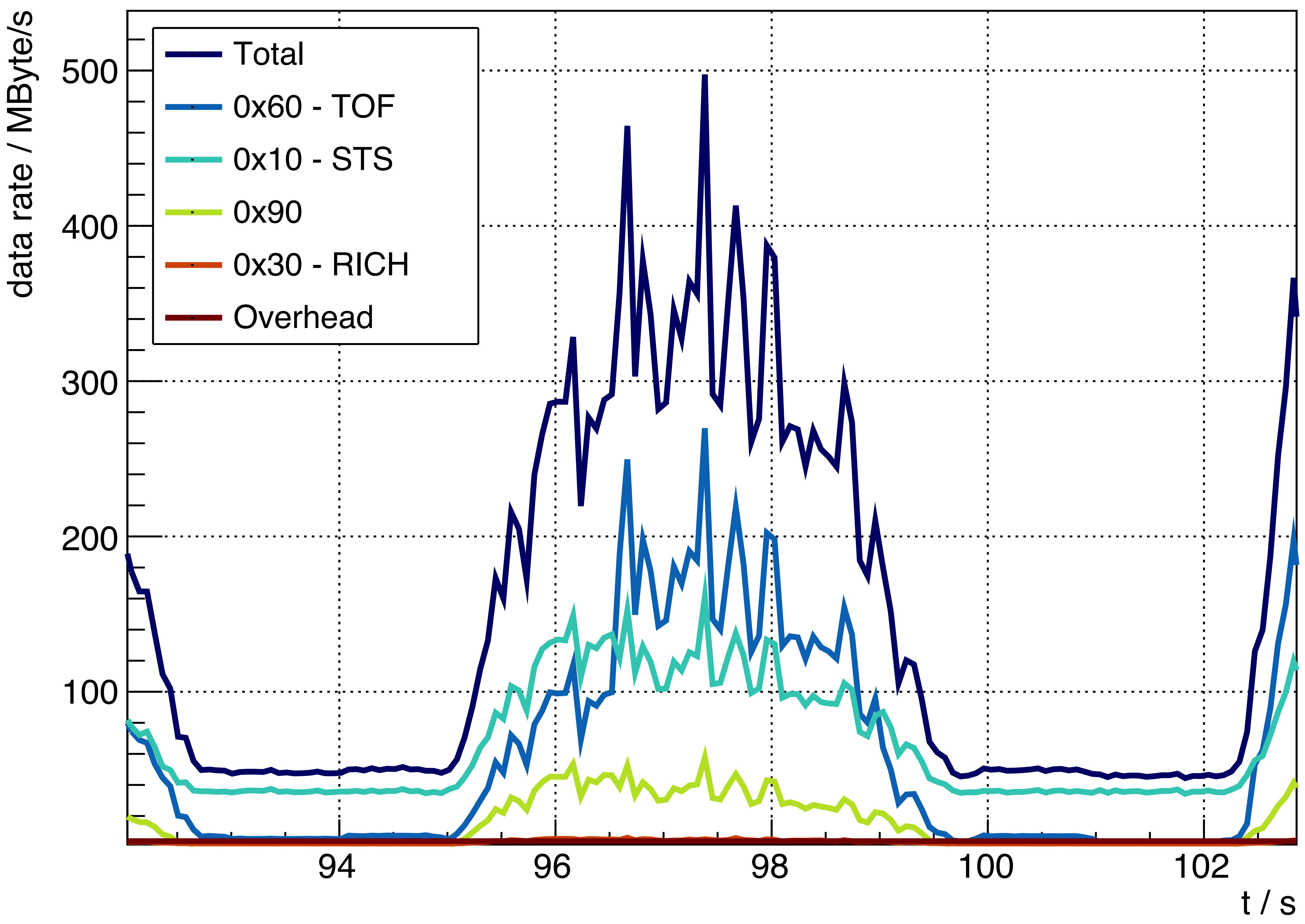
- Two mCBM campaigns with beam: Dec 2018 and Mar 2019 (next: Nov 2019)
- **mini-FLES**: central readout element
- **8 TB** recorded, high-rate runs on last two days
  - Physics data at SIS-18, Ag-Ag
  - With detector systems: T0, STS, TOF, RICH, MUCH





mCBM Stability and Observed Total Data Rate

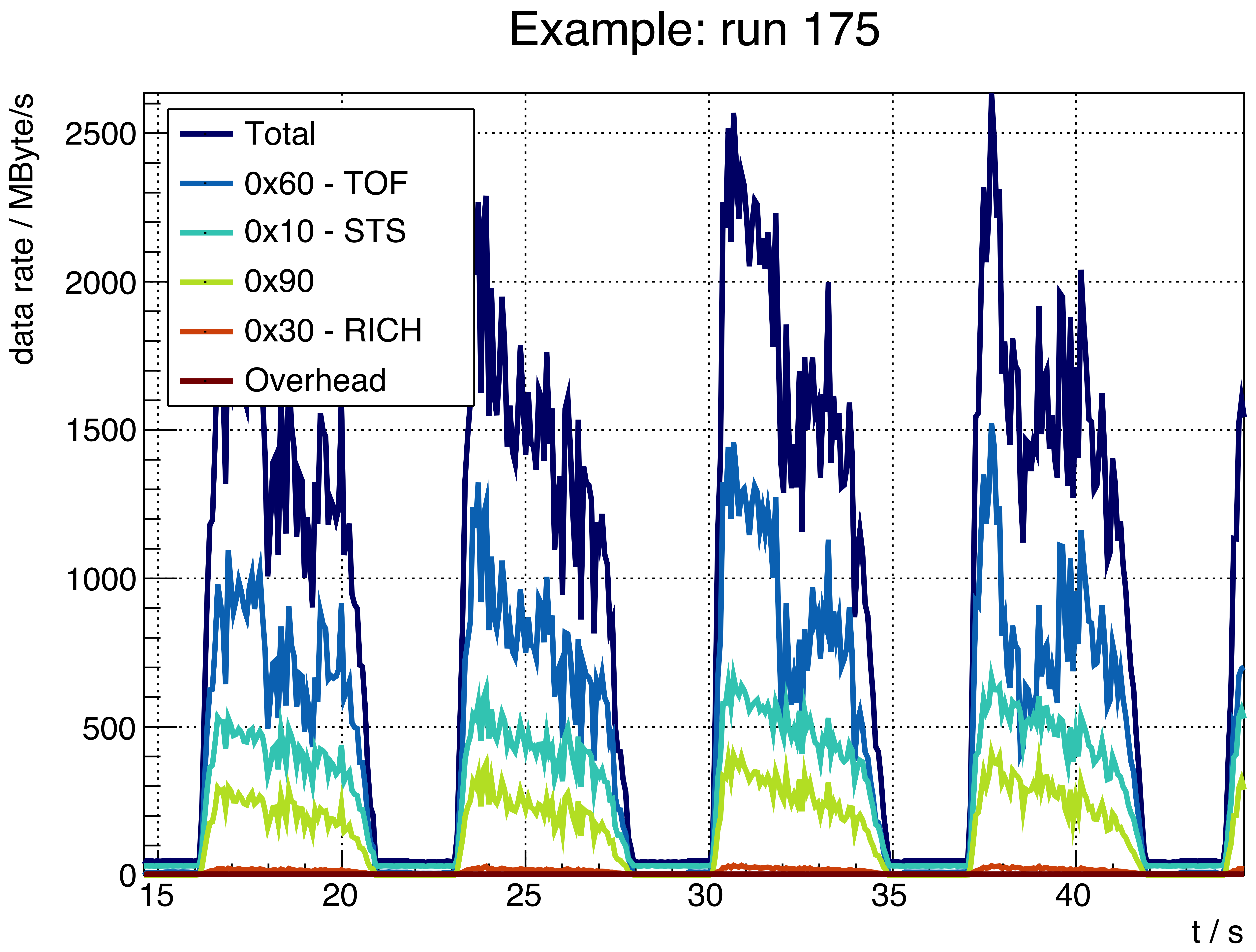




- Typical example: run 155
  - Configuration tag: `sts2_much4_tof5_rich_7pn_rec`
  - 2.5 mm gold target; ~ 3×10^6 ions/s
- No major issues related to FLES components seen

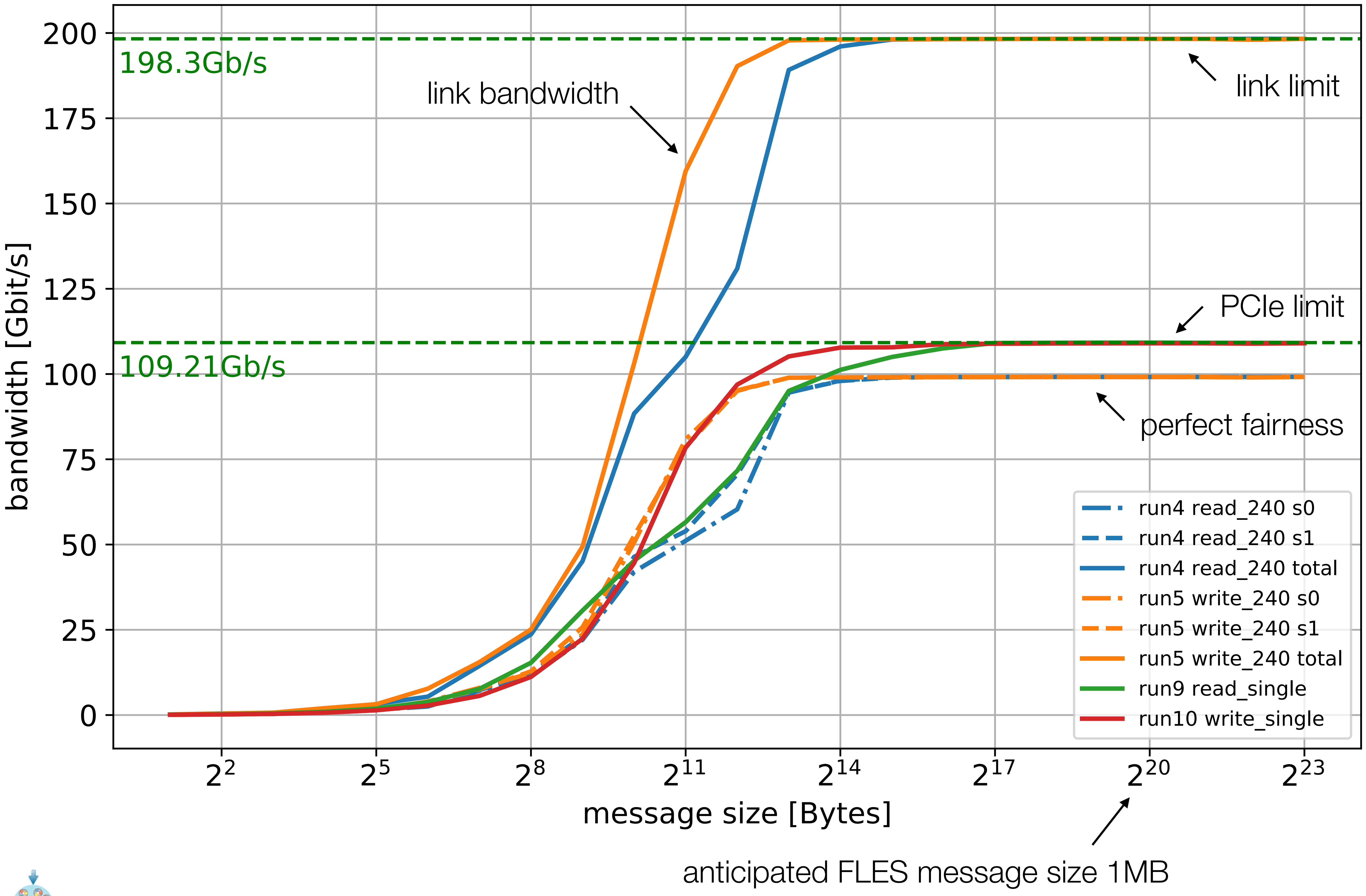
- FLES data path worked without problems
  - Internal pattern generators and automated data integrity checks proved useful for commissioning with detectors
- Timeslice building successfully scaled to several nodes

Observed Total Data Rate (Highest Intensity)



- Peak data rate **> 2.5 GByte/s**
  - Highest intensity (T0 in saturation)
  - Configuration tag: `sts2_much4_tof5_rich_7pn_rec`
- Recorded to 7 PNs, employed Flesnet buffers to average the data rates
  - Derandomization working perfectly, no data loss
  - mFLES well below performance limit

Perspective: InfiniBand HDR Network



- CBM was one of the very first customers in EMEA running an InfiniBand HDR setup
  - Installed in mFLES cluster
- HCA implements 2 PCIe devices
  - Simultaneous streams on both PCIe devices
- Measured maximum link bandwidth: **198.3 Gbit/s**

Summary

- **Compressed Baryonic Matter (CBM)** experiment at FAIR
  - High event rates (10^7 Hz), complex (topological) trigger signatures
  - Self-triggered detector front-ends, data push readout architecture
- **Central CBM physics selection system: First-Level Event Selector (FLES)**
  - **HPC processor farm** including FPGAs (at entry stage) and many-core architectures (e.g., GPUs)
  - >1 TByte/s input data stream, timeslice building in RDMA-enabled network
- **FLES demonstrator: mini-FLES**
  - Slice of the foreseen full FLES system, in live operation as part of mini-CBM
  - All data path components including **interface hardware** and **timeslice building**
  - Long-term developments fully demonstrated for the first time
  - FLES data path worked without problems, well below performance limit
  - Overall **successful operation**, further extending scope for next campaigns

SPONSORED BY THE
Federal Ministry of Education and Research

CBM

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