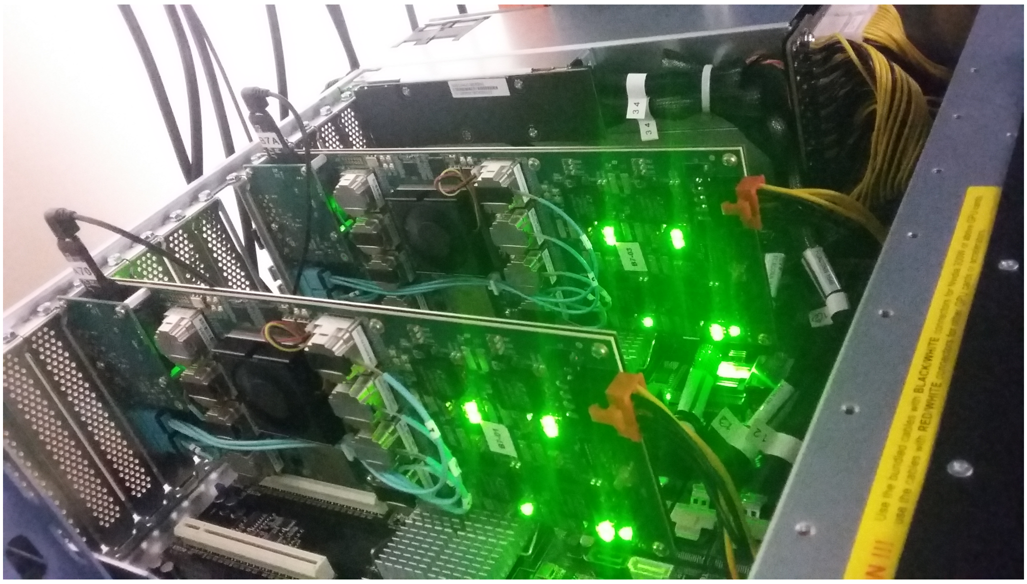
Compressed Baryonic Matter Experiment

Technical Design Report for the CBM

Online Systems (Part I)

The CBM Collaboration



January 2022

Contents

Chapter 1

The Compressed Baryonic Matter Experiment

1.1 Exploring the phase diagram of nuclear matter

Substantial experimental and theoretical efforts worldwide are devoted to the exploration of the phase diagram of nuclear matter. Figure 1.1 illustrates the possible phases of nuclear matter and their boundaries in a diagram of temperature versus the baryon chemical potential. Cold nuclear matter - as found in normal nuclei with a net-baryon density equal to one - consists of protons and neutrons (i.e. nucleons) only. At moderate temperatures and densities, nucleons are excited to short-lived states (baryonic resonances) which decay by the emission of mesons. At higher temperatures, also baryon-antibaryon pairs are created. This mixture of baryons, anti-baryons and mesons, all strongly interacting particles, is generally called hadronic matter, or baryonic matter if baryons prevail. At very high temperatures or densities the hadrons melt, and the constituents, the quarks and gluons, form a new phase: the Quark-Gluon-Plasma (QGP). For very low net-baryon densities where the numbers of particles and anti-particles are approximately equal, Quantum Chromo-Dynamics (QCD) on the lattice predicts that hadrons dissolve into quarks and gluons above a temperature of about 160 MeV Bazavov et al., "The chiral and deconfinement aspects of the QCD transition"; Borsanyi et al., "Is there still any T_c mystery in lattice QCD? Results with physical masses in the continuum limit III." The inverse process happened in the universe during the first few microseconds after the big bang: the quarks and gluons were confined into hadrons. In this region of the phase diagram the transition is expected to be a smooth crossover from partonic to hadronic matter Aoki et al., "The order of the quantum chromodynamics transition predicted by the standard model of particle physics." Calculations suggest a critical endpoint at relatively large values of the baryon chemical potential Fodor and Katz, "Critical point of QCD at finite T and μ_B, lattice results for physical quark masses." Beyond this critical endpoint, for larger values of net-baryon densities (and for lower temperatures), one expects a phase transition from hadronic to partonic matter with a phase coexistence region in between. A new phase of so called quarkyonic matter has been proposed to exist beyond the first order phase transition at large baryon chemical potentials and moderate temperatures Andronic et al., "Hadron production in ultra-relativistic nuclear collisions: quarkyonic matter and a triple point in the phase diagram of QCD." High-density but cold nuclear matter is expected to exist in the core of neutron stars, and at very high densities correlated quark-quark pairs are predicted to form a color superconductor.

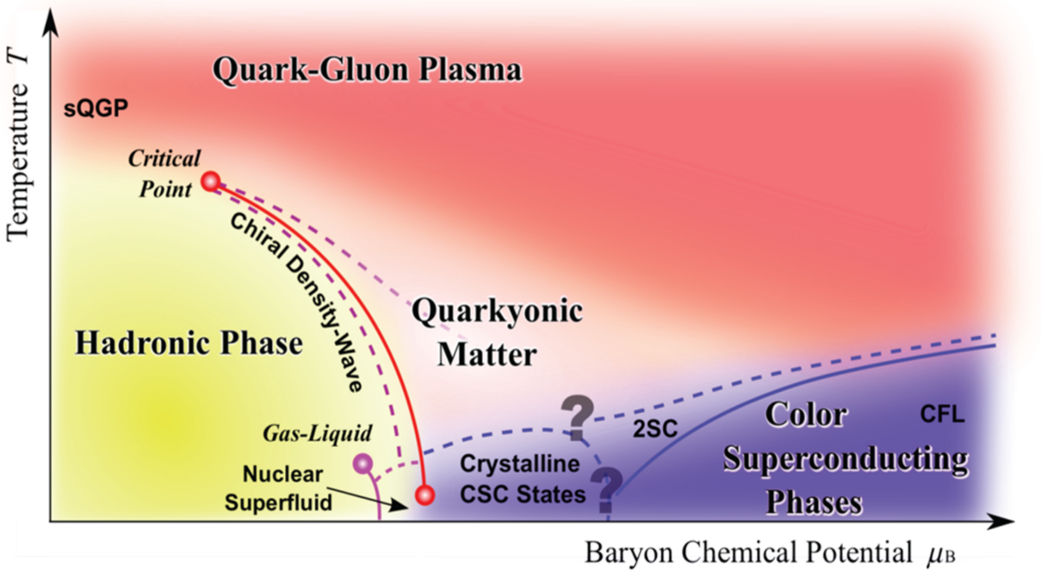


Figure 1.1: Sketch of the phase diagram for strongly-interacting matter (taken from Fukushima and Hatsuda, “The phase diagram of dense QCD”).

As illustrated in Fig. 1.1, it is expected that the QCD phase diagram exhibits a rich structure at finite values of baryon chemical potentials, such as the critical point, the predicted first order phase transition between hadronic and partonic or quarkyonic matter, and the chiral phase transition. The experimental discovery of these prominent landmarks of the QCD phase diagram would be a major breakthrough in our understanding of the properties of nuclear matter. Equally important is quantitative experimental information on the properties of hadrons in dense matter which may shed light on chiral symmetry restoration and the origin of hadron masses.

In the laboratory hot and dense nuclear matter is generated in a wide range of temperatures and densities by colliding atomic nuclei at high energies. The goal of the experiments at RHIC and LHC is to investigate the properties of deconfined QCD matter at very high temperatures and almost zero net-baryon densities. Several experimental programs are devoted to the exploration of the QCD phase diagram at high net-baryon densities. The STAR collaboration at RHIC scanned the beam energies in order to search for the QCD critical endpoint Schmah et al., “Highlights of the beam energy scan from STAR.” For the same reason, measurements are performed at the CERN-SPS with the upgraded NA49 detector (NA61) using light and medium size ion beams Aduszkiewicz et al., “NA61/SHINE at the CERN SPS: plans, status and first results.” At the Joint Institute for Nuclear Research (JINR) in Dubna, a heavy-ion collider project (NICA) is planned with the goal to search for the coexistence phase of nuclear matter Blaschke et al., “Topical issue on exploring strongly interacting matter at high densities - NICA white paper." However, due to luminosity or detector limitations these experiments are constrained to the investigation of particles which are abundantly produced. In contrast, the Compressed Baryonic Matter (CBM) experiment at the Facility for Antiproton and Ion Research (FAIR) in Darmstadt is designed for precision measurements of multidimensional observables including particles with very low production cross sections using the high-intensity heavy-ion beams provided by the FAIR accelerators.

The SIS100/300 accelerators at FAIR are very well suited to create high net-baryon densities. This is illustrated in Fig. 1.2 which depicts results of transport code calculations for central Au + Au collisions. According to these calculations, densities of up to seven times saturation density can be produced already at beam energies of 10 *A*GeV. Under these conditions the nucleons overlap, and theory predicts a transition to a mixed phase of baryons and quarks.

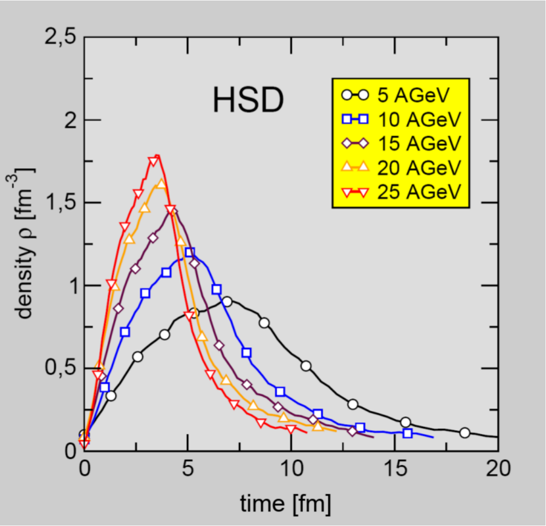


Figure 1.2: Baryon density as function of elapsed time for central Au + Au collisions at different energies as calculated with the HSD transport code Ehehalt and Cassing, "Relativistic transport approach for nucleus nucleus collisions from SIS to SPS energies."

1.2 Diagnostic probes of the high-density fireball

Figure 1.3 depicts three snapshots of the evolution of a heavy-ion collision at FAIR energies as calculated with the UrQMD transport code Bass et al., "Microscopic models for ultrarelativistic heavy ion collisions," and illustrates the time of production and eventual emission of various particle species. Particles containing charm quarks are expected to be created in the very first stage of the reaction. Then, D mesons and J/ψ mesons may serve as probes for the dense fireball and its degrees of freedom. Vector mesons like ω, ρ and ϕ

mesons are produced continuously via $\pi\pi$ annihilation during the course of the reaction, and decay either again into mesons or into a pair of leptons. However, as leptons are not affected by final-state interactions, the dileptonic decay offers the possibility to look into the fireball. In particular, the short-lived ρ meson is a promising diagnostic probe of hot and dense nuclear matter. Due to their small hadronic cross sections, also multi-strange hyperons and ϕ mesons carry information on the dense phase of the collision, in particular via their collective flow. Finally, the bulk of the particles freezes out at densities below saturation density. Up to date, essentially only these bulk particles have been measured in heavy-ion collisions at beam energies between 2 and 40 AGeV (on stationary target). Diagnostic probes of the dense stage of the fireball such as multi-strange baryons, dilepton pairs and charmed particles will be measured for the first time by the CBM experiment in this beam energy range. Therefore, the CBM experiment has a unique discovery potential both at SIS100 and SIS300 energies.

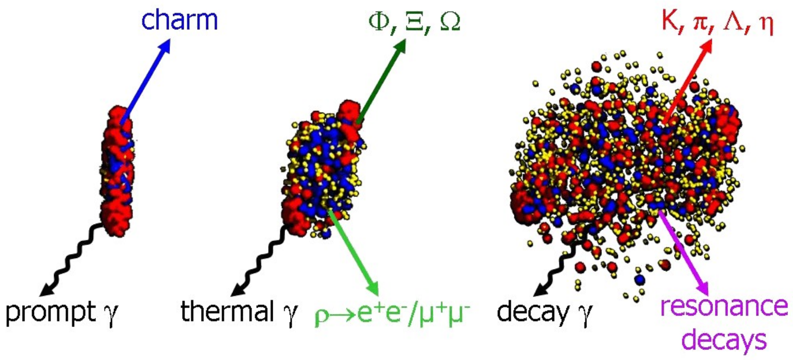


Figure 1.3: Three stages of a U + U collision at a laboratory beam energy of 23 AGeV as calculated with the UrQMD model Bass et al., “Microscopic models for ultrarelativistic heavy ion collisions”: The initial stage where the two Lorentz-contracted nuclei overlap (left), the high density phase (middle), and the final stage (“freeze-out”) when all hadrons have been formed (right). Different particles are created in different stages of the collisions or escape from the interaction region at different times (see text). Almost 1000 charged particles are created in such a collision, most of them are pions.

The experimental challenge is to measure multi-differential observables and particles with very low production cross sections such as multi-strange (anti-)hyperons, particles with charm and lepton pairs with unprecedented precision. The situation is illustrated in the left panel of Fig. 1.4 which depicts the multiplicities for various particle species produced in central Au + Au collisions at 4 AGeV. The data points are calculated using the thermal model based on the corresponding temperature and baryon-chemical potential Andronic et al., “Hadron production in central nucleus-nucleus collisions at chemical freeze-out.” The dilepton decay of vector mesons, here illustrated for the ϕ meson, is suppressed by the square of the electromagnetic coupling constant $(1/137)^2$, resulting in a dilepton yield which is about six orders of magnitude below the pion yield, similar to the multiplicity of multi-strange anti-hyperons.

In order to produce high-statistics data even for the particles with the lowest production cross sections, the CBM experiment is designed to run at reaction rates of 100 kHz up to 1 MHz. For charmonium measurements - where a trigger on high-energy lepton pairs can be generated - reaction rates up to 10 MHz are envisaged. This exceeds the rate capabilities of other existing and planned heavy-ion experiments by orders of magnitude, as illustrated in the right panel of Fig. 1.4.

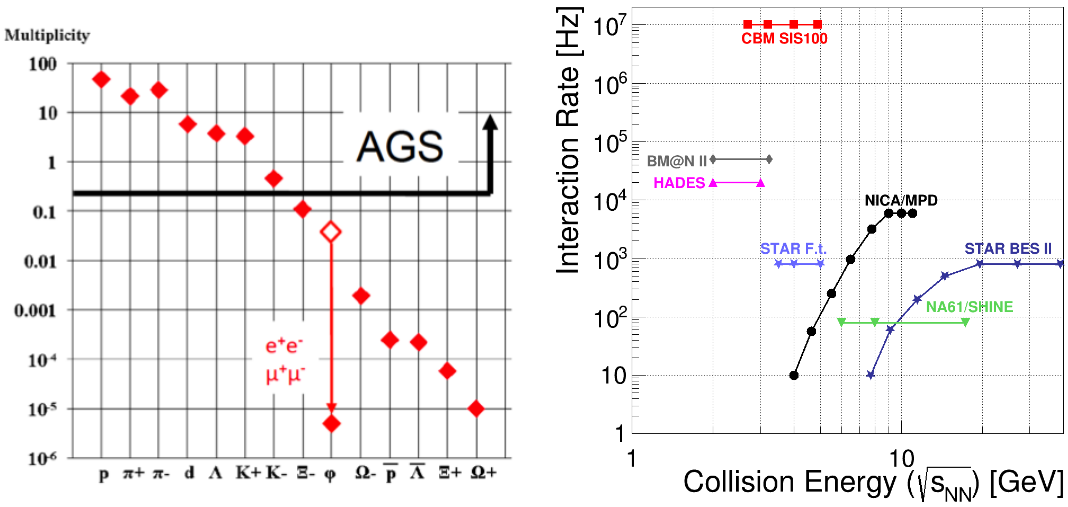


Figure 1.4: Left: Particle multiplicities for central Au + Au collisions at 4 *A*GeV as calculated with a statistical model Andronic et al., “Hadron production in central nucleus-nucleus collisions at chemical freeze-out.” For the ϕ meson also the branching fraction for the decay into lepton pairs is included (open symbol). The black line roughly indicates the multiplicities that were available to the AGS heavy-ion program at BNL at this energy. Right: Interaction rates achieved by existing anf planned heavy-ion experiments as a function of center-of-mass energy Ablyazimov et al., “Challenges in QCD matter physics – The scientific programme of the Compressed Baryonic Matter experiment at FAIR.” “STAR F.t.” denotes the fixed-target operation of STAR.

1.3 CBM physics cases and observables

The CBM research program is focused on the following physics cases:

The equation-of-state of baryonic matter at neutron star densities.
The relevant measurements are:

- The excitation function of the collective flow of hadrons which is driven by the pressure created in the early fireball (SIS100).

- The excitation functions of multi-strange hyperon yields in Au + Au and C + C collisions at energies from 2 to 11 AGeV (SIS100). At sub-threshold energies, Ξ and Ω hyperons are produced in sequential collisions involving kaons and Λ, and are therefore sensitive to the density in the fireball.

In-medium properties of hadrons.
The restoration of chiral symmetry in dense baryonic matter will modify the properties of hadrons. The relevant measurements are:

- The in-medium mass distribution of vector mesons decaying in lepton pairs in heavy-ion collisions at different energies ($2-45$ AGeV), and for different collision systems. Leptons are penetrating probes carrying the information out of the dense fireball (SIS100/300).
- Yields and transverse mass distributions of charmed mesons in heavy-ion collision as a function of collision energy (SIS100/300).

Phase transitions from hadronic matter to quarkyonic or partonic matter at high net-baryon densities.
Already at SIS100 energies densities of up to seven times of the normal nuclear density are reached in central collisions between heavy-ions. A discontinuity or sudden variation in the excitation functions of sensitive observables would be indicative of a transition. The relevant measurements are:

- The excitation function of yields, spectra and collective flow of strange particles in heavy-ion collisions from $6-45$ AGeV (SIS100/300).
- The excitation function of yields, spectra and collective flow of charmed particles in heavy-ion collisions from $6-45$ AGeV (SIS100/300).
- The excitation function of yields and spectra of lepton pairs in the intermediate mass region in heavy-ion collisions from $6-45$ AGeV (SIS100/300).
- Event-by-event fluctuations of conserved quantities like baryons, strangeness, net-charge etc. in heavy-ion collisions with high precision as function of beam energy from $6-45$ AGeV (SIS100/300).

Hypernuclei, strange dibaryons and massive strange objects.
Theoretical models predict that single and double hypernuclei, strange dibaryons and heavy multi-strange short-lived objects are produced via coalescence in heavy-ion collisions with the maximum yield in the region of SIS100 energies. The planned measurements include:

- The decay chains of single and double hypernuclei in heavy-ion collisions at SIS100 energies.
- Search for strange matter in the form of strange dibaryons and heavy multi-strange short-lived objects. If these multi-strange particles decay into charged hadrons including hyperons they can be identified via their decay products.

Charm production mechanisms, charm propagation and in-medium properties of charmed particles in (dense) nuclear matter.
The relevant measurements are:

- Cross sections and momentum spectra of open charm (D-mesons) in proton-nucleus collisions at SIS100/300 energies. In-medium properties of D-mesons can be derived from the transparency ratio $T_A = (\sigma_{pA} \to DX)/(A \times \sigma_{pN} \to DX)$ measured for different size target nuclei.
- Cross sections, momentum spectra and collective flow of open charm (D-mesons) in nucleus-nucleus collisions at SIS300 energies.
- Cross sections, momentum spectra and collective flow of charmonium (J/ψ) in proton-nucleus and nucleus-nucleus collisions at SIS100/300 energies.

As discussed above, a substantial part of the CBM physics cases can be addressed already with beams from the SIS100 synchrotron Ablyazimov et al., "Challenges in QCD matter physics – The scientific programme of the Compressed Baryonic Matter experiment at FAIR." The intended measurements at SIS100 including the results of simulations and count rate estimates are described in Senger, V. Friese, et al., *Nuclear matter physics at SIS-100*. A general review of the physics of compressed baryonic matter, the theoretical concepts, the available experimental results and predictions for relevant observables in future heavy-ion collision experiments can be found in the CBM Physics Book Friman et al., "The CBM physics book: Compressed baryonic matter in laboratory experiments."

1.4 The Facility for Antiproton and Ion Research (FAIR)

The international Facility for Antiproton and Ion Research (FAIR) in Darmstadt will provide unique research opportunities in the fields of nuclear, hadron, atomic and plasma physics Gutbrod et al., *FAIR baseline technical report*. The research program devoted to the exploration of compressed baryonic matter will start with primary beams from the SIS100 synchrotron (protons up to 29 GeV, Au up to 11 AGeV, nuclei with $Z/A = 0.5$ up to 14 AGeV), and will be continued with beams from the SIS300 synchrotron (protons up to 90 GeV, Au up to 35 AGeV, nuclei with $Z/A = 0.5$ up to 45 AGeV). The layout of FAIR is presented in Fig. 1.5. The beam extracted to the CBM cave reaches intensities up to 10^9 Au ions per second.

1.5 The Compressed Baryonic Matter (CBM) experiment

The CBM experimental strategy is to perform systematic both integral and differential measurements of almost all the particles produced in nuclear collisions (i.e. yields, phase-space distributions, correlations and fluctuations) with unprecedented precision and statistics. These measurements will be performed in nucleus-nucleus, proton-nucleus, and - for

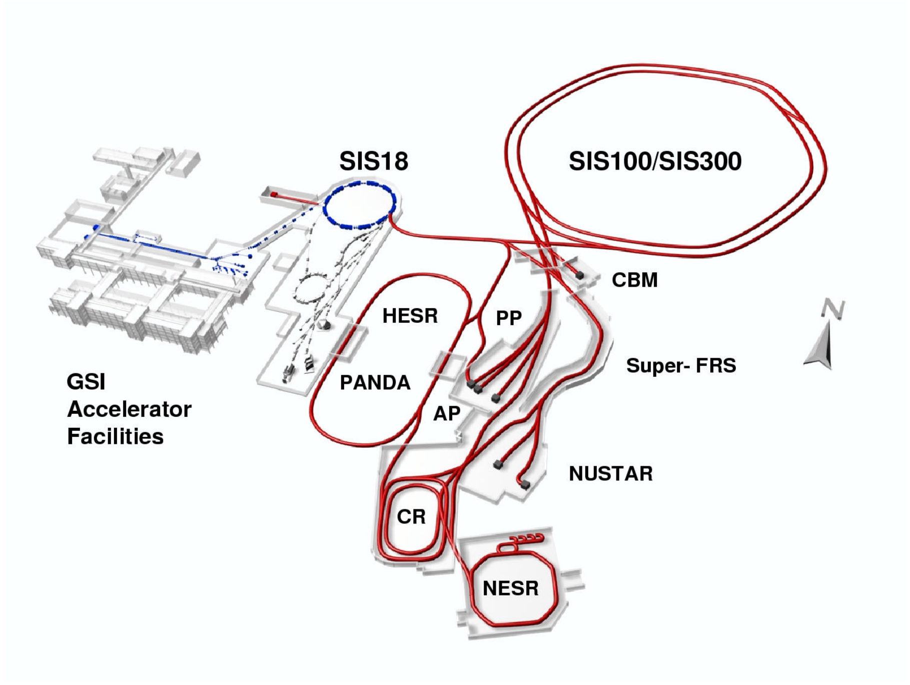


Figure 1.5: Layout of the Facility for Antiproton and Ion Research (FAIR) Gutbrod et al., *FAIR baseline technical report*.

baseline determination - proton-proton collisions at different beam energies. The identification of multi-strange hyperons, hypernuclei, particles with charm quarks and vector mesons decaying into lepton pairs requires efficient background suppression and very high interaction rates. In order to select events containing those rare observables, the tracks of each collision have to be reconstructed and filtered online with respect to physical signatures. This concept represents a paradigm shift for data taking in high-energy physics experiments: CBM will run without hierarchical trigger system. Self-triggered readout electronics, a high-speed data processing and acquisition system, fast algorithms, and, last but not least, radiation hard detectors are indispensable prerequisites for a successful operation of the experiment. Figure 1.6 and depict the CBM experimental setup for SIS100. The CBM experiment comprises the following components:

Dipole magnet
The dipole magnet will be superconducting in order to reduce the operation costs. It has a large aperture of ±25° polar angle, and provides a magnetic field integral of 1 Tm.

Micro-Vertex Detector (MVD)
The MVD will provide excellent position resolution and low material budget as required for the identification of open charm particles by the measurement of their displaced decay vertex. It consist of four layers of Monolithic Active Pixel Sensor (MAPS) detectors located

from 5 cm to 20 cm downstream of the target in vacuum. The detector arrangement provides a resolution of secondary vertices of about $50 - 100\mu$m along the beam axis.

Silicon Tracking System (STS)
The task of the STS is to provide track reconstruction and momentum determination of charged particles. The system consists of eight tracking layers of silicon strip detectors, located downstream of the target at distances between 30 cm and 100 cm inside the magnetic dipole field, and provides a momentum resolution of about $\Delta p/p = 1.5\%$.

Ring Imaging Cherenkov Detector (RICH)
The RICH detector will provide the identification of electrons via the measurement of their Cherenkov radiation. This will be achieved using a gaseous RICH detector build in a standard projective geometry with focusing mirror elements and a photon detector. The detector will be positioned behind the dipole magnet about 1.6 m downstream of the target. It will consist of a 1.7 m long gas radiator (overall length approximately 2 m) and two arrays of mirrors and photon detector planes. The design of the photon detector plane is based on MAPMTs in order to provide high granularity, high geometrical acceptance, high detection efficiency of photons also in the near UV region and a reliable operation.

Muon Chamber System (MUCH)
The concept of the muon detection system is to track the particles through a hadron absorber and thus perform a momentum dependent muon identification. The absorber/detector system is placed downstream of the STS, which determines the particle momentum. In order to reduce meson decays into muons the absorber/detector system is designed as compact as possible. It consists of six hadron absorber layers made of iron plates and 18 gaseous tracking chambers located in triplets behind each iron slab (SIS300 setup). The trigger concept is based on the measurement of short track segments in the last tracking station triplet, and extrapolation of these tracks to the target. For J/ψ measurements at SIS100 a MUCH start version with three chamber triplets is sufficient.

Transition Radiation Detector (TRD)
The Transition Radiation Detector, consisting of four detector layers grouped into one station in the SIS100 configuration (ten layers in three stations for SIS300), will serve for particle tracking and for the identification of electrons and positrons with $p > 1.0$ GeV/c ($\gamma \geq 1000$). The detector layers are located at approximately 4.1 m to 5.9 m downstream of the target, the total active detector area amounts to about 114 m^2 (SIS100). The TRD readout will be realized in rectangular pads giving a resolution of $\sim 300\mu$m across and $3 - 30$ mm along the pad. Every second TRD layer is rotated by 90°.

Time-Of-Flight System (TOF)
An array of Multi-gap Resistive Plate Chambers (MRPC) will be used for hadron identification via TOF measurements. The TOF wall covers an active area of about 120 m^2

and is located about 6 m downstream of the target for measurements at SIS100, and at 10 m at SIS300. The required time resolution is on the order of 80 ps. At small deflection angles the pad size is about 5 cm^2 corresponding to an occupancy of below 5% for central Au + Au collisions at 25 *A*GeV.

Electromagnetic Calorimeter (ECAL)

A "shashlik" type calorimeter as installed in the HERA-B, PHENIX and LHCb experiments will be used to measure direct photons and neutral mesons (π^0, η) decaying into photons. The ECAL will be composed of modules which consist of 140 layers of lead and scintillator sheets. The shashlik modules can be arranged either as a wall or in a tower geometry with variable distance from the target.

Projectile Spectator Detector (PSD)

The PSD will be used to determine the collision centrality and the orientation of the reaction plane. The detector is designed to measure the number of non-interacting nucleons from a projectile nucleus in nucleus-nucleus collisions. The PSD is a fully compensating modular lead-scintillator calorimeter which provides very good and uniform energy resolution. The calorimeter comprises 44 individual modules, each consisting of 60 lead/scintillator layers.

Online event selection and data acquisition

High-statistics measurements of particles with very small production cross sections require high reaction rates. The CBM detectors, the online event selection systems and the data acquisition will be designed for event rates of 10 MHz, corresponding to a beam intensity of 10^9 ions/s and a 1 % interaction target, for example. Assuming an archiving rate of 1 GByte/s and an event volume of about 10 kByte for minimum bias Au + Au collisions, an event rate of 100 kHz can be accepted by the data acquisition. Therefore, measurements with event rates of 10 MHz require online event selection algorithms (and hardware) which reject the background events (which contain no signal) by a factor of 100 or more. The event selection system will be based on a fast online event reconstruction running on a high-performance computer farm equipped with many-core CPUs and graphics cards (GSI GreenIT cube). Track reconstruction, which is the most time consuming combinatorial stage of the event reconstruction, will be based on parallel track finding and fitting algorithms, implementing the Cellular Automaton and Kalman Filter methods. For open charm production the trigger will be based on an online search for secondary vertices, which requires high speed tracking and event reconstruction in the STS and MVD. The highest suppression factor has to be achieved for J/ψ mesons where a high-energetic pair of electrons or muons is required in the TRD or in the MUCH. For low-mass electron pairs no online selection is possible due to the large number of rings/event in the RICH caused by the material budget of the STS. In the case of low-mass muon pairs some background rejection might be feasible.

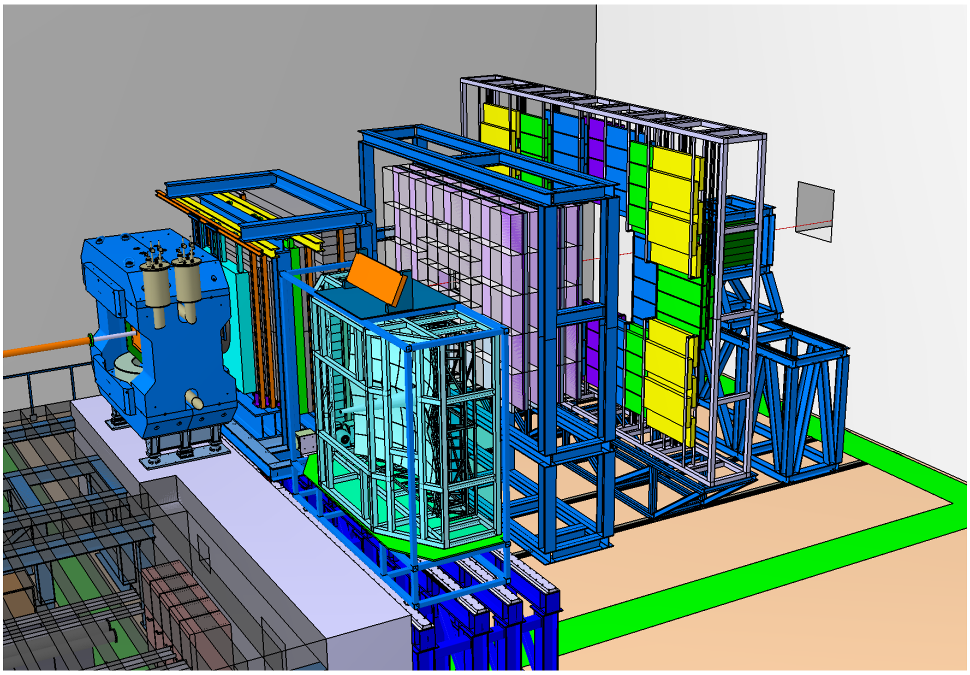


Figure 1.6: Drawing of the experimental setup of CBM for the SIS100.

Chapter 2

Task and Concept

(Editorial Board)

2.1 Data reduction and signatures

The goal of the Compressed Baryonic Matter (CBM) experiment at the FAIR SIS100/300 accelerator system is to explore the QCD phase diagram in the region of very high baryon densities. In particular, the experiments will focus on the search for the phase transition between hadronic and quark-gluon matter, the QCD critical endpoint, new forms of strange matter, in-medium modifications of hadrons and the onset of chiral symmetry restoration. A detailed discussion of the physics of compressed baryonic matter can be found in.

The SIS-100 accelerator will deliver beams of heavy ions (Au) up to 11A GeV ($\sqrt{s_{\mathrm{NN}}} =$ 4.7 GeV), light ions (e.g. Ca) up to 14A GeV ($\sqrt{s_{\mathrm{NN}}} = 5.3$ GeV) and protons up to 29 GeV ($\sqrt{s_{\mathrm{NN}}} = 7.5$GeV). The research programme, which can be conducted with a basis version of the CBM setup, will address the equation of state of nuclear matter and the properties of hadrons at large baryonic densities, as well as strangeness and charm production near thresholds, where the production rates depend on the conditions inside the early fireball.

The energy regime of the SIS-300 accelerator, reaching beam energies up to 40A GeV, will open the phase-space to a more abundant production of rare penetrating probes, which might result in more stringent signatures of the transient deconfined plasma. In particular, the production of newly created quark flavors, like charm, with energies well above the charm production threshold, offers the possibilities to study in-medium effects on D-mesons and charmonia, which are promising observables to probes the chiral dynamics at high baryon densities.

The most promising probes, which are sensitive to high density effects and phase transitions, are short-lived particles containing charm quarks, or multi-strange hyperons (Ξ, Ω), or quarkonia. One of the early preduction for a signature of a deconfined plasma of quarks and gluons, socalled Quark Gluon Plasma (QGP) was the increased production of strangeness resulting in an enhanced yield of strange (and even more multi-strange) particles after hadronization. Also, the anomalous suppression of charmonium due to screening effects was predicted to be an experimental signature of the QGP. While the role of these probes as a signature for a deconfinement phase transition is still under discussion, and various phenomena have emerged from the experimental data leading to potentially opposite effects in the measured rates, strangeness and charm still play a central role in the studies of nuclear matter.

The production rate of such probes however is low and the beam energies at FAIR are close to the kinematic threshold. Therefor this challenging research programme requires high-performance subsystems such as highrate detectors, free-streaming read-out electronics, high-speed data acquisition, and an ultra-fast online event selection performed on a computer farm consisting of many-core CPUs accelerated with graphics cards.

As a general-purpose experiment, though optimized for heavy-ion reactions, CBM has to be able to track and identify particles from very low (... MeV/c) up to fairly high (... GeV/c) transverse momentum, to reconstruct short-lived particles such as hyperons and charmed mesons, and to perform these tasks in an environment of large particle density (up to 1000 charged particles) at extreme interaction rates (10 MHz). This is with detectors that measure and identify hadrons, electrons, photons, and muons produced in proton-proton, proton-nucleus and nucleus-nucleus interactions up to 40A GeV.

Hadrons, electrons and photons are detected and identified by a complex system of detectors placed in a moderate magnetic dipole field (... T). Tracking relies on a set of Silicon pixel (and strip) detectors (STS) close to the target which will measure the trajectories of the produced particles, determine their momenta and reconstruct hyperons by their decay topology. A micro-vertex detector (MVD) can be employed for the high precision measurement of the decay vertices. An intermediate matching point for the tracking can be provided by the start version of the Transition Radioation Detector (TRD). Particle identification is performed by a large-area time-of-flight detector (TOF) consisting of multi-gap resistive plate chambers for the identication of pions, kaons and protons, and by alternating a Ring Imaging Cherenkov (RICH) detector for electron identification and a muon detection system (MUCH) for muon identification.

The experimental challengs for the online event selection consist in developing complex reconstruction algorithms that require information from several detector sub-system. For example, in case of strangeness and open charm, the event selection will be based on an online search for secondary vertices, which requires high-speed tracking and event reconstruction of STS (and MVD, in case of charm) data. The highest suppression factor has to be achieved for J/ψ mesons, where a pair of high-energy electrons or muons is required in the TRD or MUCH detectors.

2.2 Overall architecture

Overall architecture (W. Müller)

2.2.1 PCA - DCA - SCA

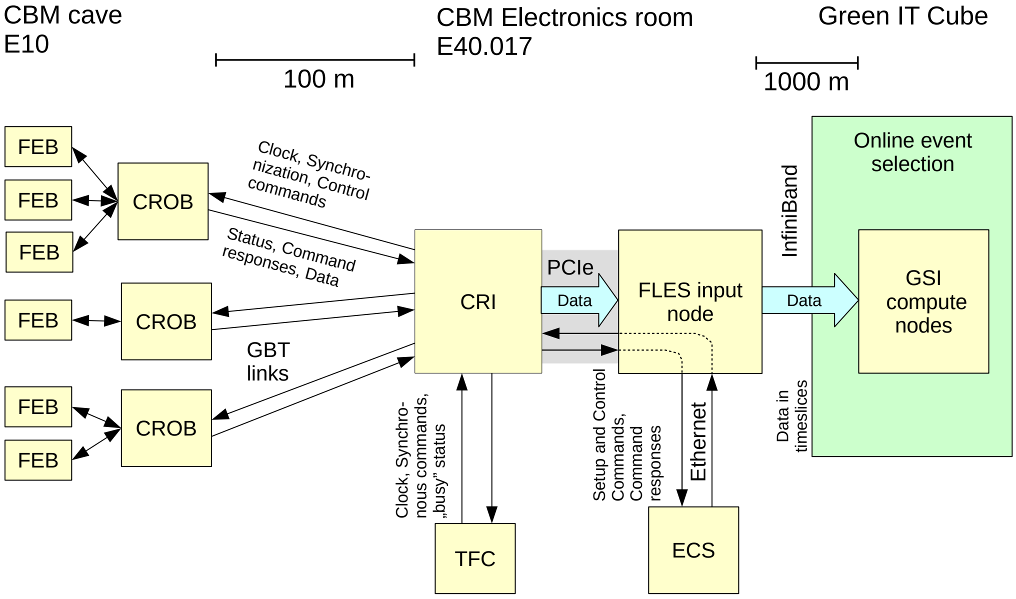


Figure 2.1: The CRI-based CBM readout chain.

Chapter 3

Requirements and Constraints

(Editorial Board)

3.1 Requirements

3.1.1 Data Rates

A 2018 prediction of the detector data rates can be found in the CBM computing note 18001: `https://indico.gsi.de/event/7449/contributions/33311/attachments/24092/30155/cbm-cn-18001.pdf`

Questions to the detector subsystems:

1. Are the numbers given in Tables 1 to 3 of the CBM computing note 18001 for your subsystem up-to-date and in line with the most recent geometry?
2. Is there any data overhead *per message*, i.e. is there additional data being transmitted to the CRI depending on the amount of messages? This might include, e.g., an average estimate for deduplicated epoch markers.
3. How large is the data overhead *per time*? This might include, e.g., periodic epoch markers and status messages.
4. What is the expected dark rate in messages/s?

We will use these numbers to compute aggregate data rates for the different setups in common tables.

In addition, we will need some text:

- A short text per detector subsystem explaining any non-obvious numbers w.r.t. questions 1–3.
- For question 4 (noise), please explain the estimate (e.g., threshold-to-noise ratio).
- Comparison to mCBM: How do these numbers (especially dark rate and overheads) relate to mCBM measurements?

3.1.2 Connectivity

Questions to the detector subsystems:

1. How many GBT links (or other technology) does your subsystem operate?
2. Data distribution over links: Please provide a histogram of the average number of messages per event per link.
3. Channel distribution over links: Please provide a histogram of the number of channels per link.

Again, we will use these numbers to assemble aggregate tables. The histograms are needed to define the requirements on the individual data paths.

3.2 Constraints

Chapter 4

Data Sources

(D. Emschermann)

This chapter describes for each subsystem the tree of readout electronics interfacing a single CRI. The focus is on the digital part of the readout system, not the analog front-end. Subsystems are supposed to decribe which part of their detector electronics is to be controlled and readout by one CRI. For large subsystems the number of CRI boards will be scaled to allow for the readout of the entire subsystem.

4.1 GBTx-based Readout-chains

(WUT)

4.1.1 STS and MUCH

(Jörg Lehnert)

4.1.1.1 STS/MUCH-XYTER

(Jörg Lehnert)

4.1.1.2 HCTSP-Protocol

(Michal Kruszewski)

4.1.1.3 STS

(Jörg Lehnert)

Sample text for the STS, to be completed by the subsystem:

The front-end electonics of the STS subsystem consists of front-end boards (FEB8) populated with 8 STS-XYTER ASICs (SMX). There are three variants of the FEB8 board differing in the amount of e-links used by each ASIC:

- FEB8-1 with 1 e-link connection per SMX ASIC - 8 e-links in total
- FEB8-2 with 2 e-link connections per SMX ASIC - 16 e-links in total
- FEB8-5 with 5 e-link connections per SMX ASIC - 40 e-links in total

These FEB8-x are interfaced to GBTx-Readout Boards (ROB3), populated with 1 GBTx-Master and 2 GBTx-Slave ASICs. Each of these ROB3 can interface up to $(3 * 14 =)$ 42 e-links. The following mapping of FEB8-x to ROB3 or any combination of these will be realised:

1. 5x FEB8-1 $(5 * 8 =)$ 40-e-links interfacing 1x ROB3
2. 5x FEB8-2 $(2 * 16 =)$ 80-e-links interfacing 2x ROB3
3. 1x FEB8-5 $(1 * 40 =)$ 40-e-links interfacing 1x ROB3 or combinations
4. 3x FEB8-1 + 1x FEB8-2 $(3 * 8 + 1 * 16 =)$ 40-e-links interfacing 1x ROB3
5. 1x FEB8-1 + 2x FEB8-2 $(1 * 8 + 2 * 16 =)$ 40-e-links interfacing 1x ROB3

The ROB3 are connected to the optical backbone with LC-MTP24 fan-ins. Each ROB3 provides 1x Rx- and 3x Tx-links. The MTP24 connector carries 12x Rx and 12x Tx fibers, allowing to interface 4x ROB3 boards on 1x MTP24 connector.

The CRI2 will be designed with 3x MTP24 connectors. In terms of optics 1x CRI2 can be interfaced up to:

- 12x ROB3 - $(3 * 4)$
- 60x FEB8-1 - $(3 * 4 * 5)$
- 480x SMX ASICs - $(3 * 4 * 5 * 8)$

The readout tree of a CRI2 can consist of a mix of FEB8-1, FEB8-2 and FEB8-5 boards.

The distribtuion of clock downlinks from the CRI to the FEE is foreseen as follows ...

The total amount of electronics in the STS system amount to xxx FEB8 and yyy ROB3.

4.1.1.4 MUCH-GEM

(Jogender Saini)

4.1.1.5 MUCH-RPC

(Jogender Saini)

4.1.2 TRD

(Phillip Kähler)

4.1.2.1 SPADIC

(Peter Fischer)

4.1.2.2 FASP

(Alexandru Bercuci)

4.1.3 TOF

(Jochen Fruehauf)

4.1.3.1 GET4

(Holger Flemming)

4.2 FPGA-based Readout-chains

(Jan Michel)

4.2.1 RICH

(Adrian Weber)

4.2.2 BMON

(Jerzy Pietraszko)

4.2.3 PSD

(Dmitry Finogeev)

4.3 Other Readout-chains

4.3.1 MVD

(Christian Muentz - Michael Deveaux - Jan Michel)

Chapter 5

Common Readout Interface

The Data Processing Board (DPB) layer is an intermediate layer located between the Readout Boards connected to the Front End Electronics (FEE) and the Data Acquisition First-level Event Selector (FLES).

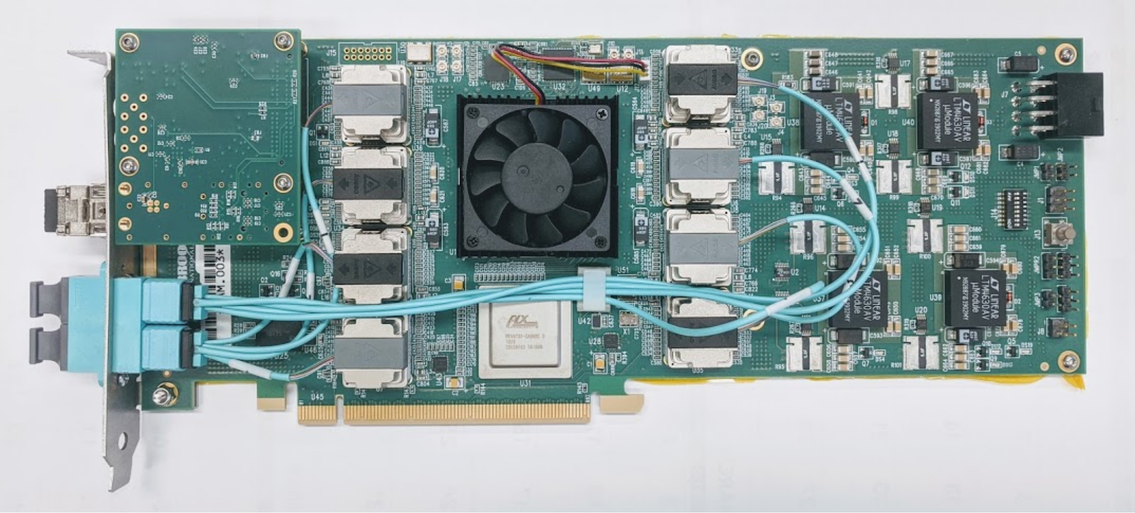


Figure 5.1: Photo of the CRI board (BNL-712).

The DPB boards will be located in the E40 area, near to detectors, but outside the irradiated area. Therefore it will be possible to base them on standard electronic components.

The main reason to introduce the DPB layer is to reduce the cost of the links needed to transmit the data from the Front End Electronics to the First-level Event Selector. The straight line distance between the FEE and FLES will be equal to 350 m, but due to additional constraints affecting placement of fibers, the length of the link will be significantly higher. The optical links transmitting the data from FEE will work at data rate not higher than 5 Gb/s. Use of higher speed optical links between the DPB and FLES will therefere allow reduction of links number (e.g. for 10 Gb/s links between the DPB and FLES the number of links will be reduced by factor of 2). Additional reduction of necessary links may be achieved by aggregation and preprocessing of data.

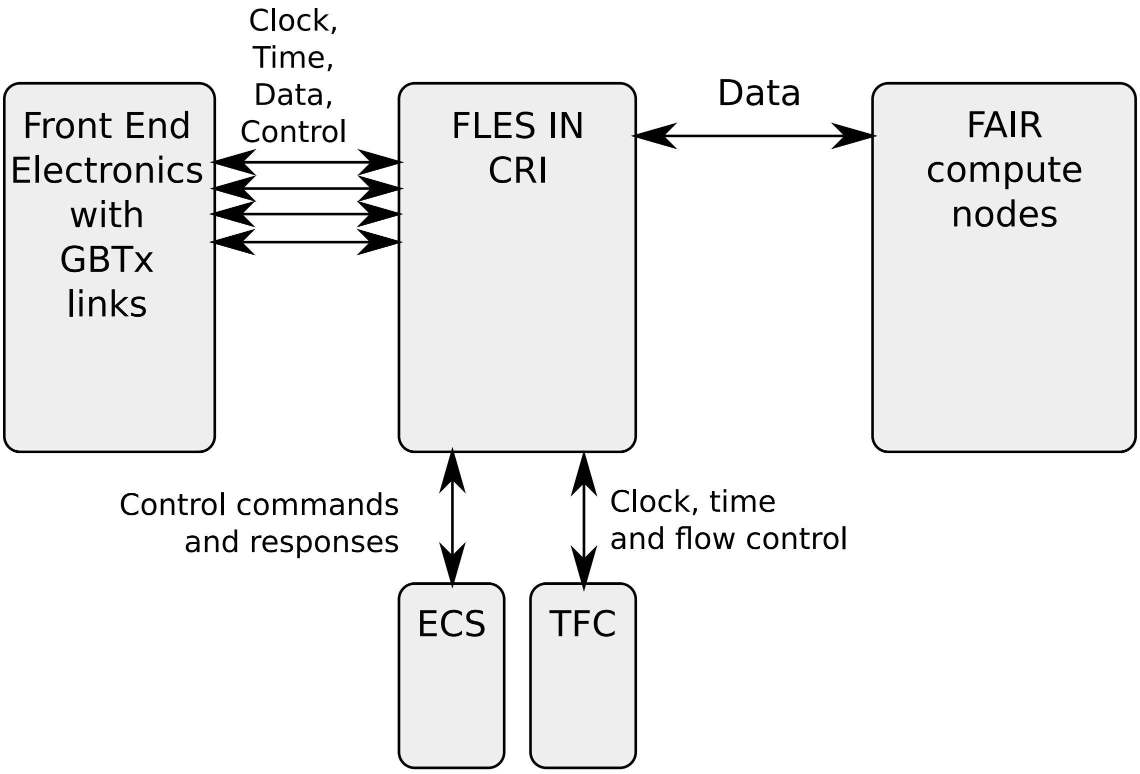

Clock,
Time,
Data,
Control
Front End Electronics with GBTx links
FLES IN CRI
Data
FAIR compute nodes
Control commands and responses
Clock, time and flow control
ECS
TFC

Figure 5.2: Location of the CRI boards in the CBM experiment

All functionalities of the DPB layer are shown in Figure 5.2. The DPB boards will communicate with four subsystems of the experiment:

- **The Front End Electronics (FEE)** - receiving the acquired data, providing FEE with the master clock and with synchronization commands, sending the control commands and receiving responses to them, receiving informations about the status of FEE.

- **The Experiment Control System (ECS)** - receiving the control commands and sending responses to them

- **The Timing and Flow Control system (TFC)** - receiving the master clock and synchronization commands, receiving the flow control commands and sending the flow control status.

The detailed tasks associated with the above main tasks will be described in the next sections.

5.1 Technology options

Location of the DPB boards in the area with low radiation level allows use of standard electronic components, which simplifies design. To achieve design flexibility and allow upgrades or corrections, the DPB layer should be based on programmable FPGA chips. Choice of technology used to implement the DPB layer is constrained by the requirements associated with all tasks listed in the previous section. The important function of DPB boards is communication with FEE and FLES via high speed optical links. Fortunately modern FPGA chips are offered also in versions equipped with multiple gigabit transceivers, which may be used for that purpose. Good examples may be the Xilinx Series 7☢**To-Do: Add reference here: url-xlx-ser7-ovrv**☢ or Altera Stratix V or Arria V. FPGA chips allow to work with data rates 10 Gb/s and higher, but also rates below 5 Gb/s are available. Implementation of both: links communicating with FEE and links communicating with FLES in the same chip, allows to fully utilize high bandwith of internal FPGA connections. Therefore it is desirable, that most functionalities of the DPB board should be implemented in a single FPGA. Next subsections describe the required functionalities in more detail.

5.1.1 Communication with FEE

Communication with the FEE subsystem will be provided by the optical links with speed up to 5Gb/s. The protocol used in these links should allow combining multiple data streams in a bidirectional link:

- transmission of master clock in the downlink (to FEE) direction
- sending of the acquired data with high througput in the uplink (from FEE) direction
- sending the synchronization messages with deterministic latency in downlink direction
- bidirectional transfer of control commands and responses

As a viable options two protocols were considered: the CBMnet Lemke and Bruening, “A hierarchical synchronized data acquisition network for CBM”; Lemke, Slogsnat, et al., “A unified DAQ interconnection network with precise time synchronization” and the GBTBaron et al., “Implementing the GBT data transmission protocol in FPGAs.” The IP cores for both protocols are available for modern FPGA chips from most significant manufacturers like Altera and Xilinx. Both protocols have also similar requirements regarding the hardware design of the DPB board.

The important part of the technology choice is the selection of appropriate optical transceivers for FEE links. It is important, that necessary bandwidth in both directions significantly differs, therefore the number of the uplink connections may be significantly higher than number od the downlink connections. Additionally it is important to consider the price factor and size of transceivers. From currently available solutions, the Avago MicroPOD or MiniPOD transceivers *http://www.avagotech.com/pages/minipod_micropod* may be a reasonable choice, however it is necessary to consider the cost of additional components (e.g. patch-pannels).

5.1.2 Communication with FLES

Connection to the FLES subsystem requires optical links of significant length. Therefore to minimize the cost, it is important to use cheaper - single mode fibers, and to reduce number of necessary links. Considering current state of technology, use of 10 Gb/s links seems to be reasonable. The necessary 10 Gb/s capable gigabit transceivers are available in most contemporary advanced FPGA chips. Another important decision is the selection of the appropriate protocol, which may affect necessary hardware resources. Possible solutions are summarized below:

- Use of private protocol, using directly the 10 gigabit transceivers
- Use of 10 gigabit transceivers as Ethernet 10GBASE-R phy, with private layer 3 protocol
- Use of 10 gigabit transceivers as Ethernet 10GBASE-R phy with standard TCP protocol

The first approach allows use higher number of uplink fibers, than number of downlink fibers. Even if we decide to use the acknowledge/retransmit mechanism, a single donwlink fiber may transmit acknowledgments for a few uplink fibers. For example the Xilinx Aurora 64/66B *LogiCORE IP Aurora 64B/66B v9.2 Product Guide* protocol may be used in such configuration.

The Ethernet based solutions require bidirectional, two-fiber links, which may increase the total cost, but allow to use the standard Ethernet adapters at the FLES side.

Use of private layer 3 protocol with optimized kernel driver in the FLES node allows to minimize acknowledge latency and decrease requirements for data buffering memory in the DPB. The IP core needed to implement such a solution is not significantly more complicated, that the one needed in the first solution.

Use of standard TCP protocol requires implementation of the TCP/IP stack in the FPGA, which consumes significant resources. The lightweight implementation was developped in CERN (the FEROLBauer et al., “10 Gbps TCP/IP streams from the FPGA for the CMS DAQ eventbuilder network” core), but another drawback is necessity to use significant amount of memory for buffering of not acknowledged packets.

It can be stated that only choice between the first and next solutions must be done early, as it affects number of links needed. Selection of the second or third solution may be done later, assuming that the prototype DPB is equipped with sufficiently big FPGA and external memory with sufficient throughput.

5.1.3 Communication with ECS and TFC

For communication with the Experiment Control System, the Ethernet interface should be used, and 1Gb/s throughput should be sufficient for that purpose.

Communication link between the DPB layer and the TFC must transmit the master clock, and the synchronization signal (e.g. the Pulse Per Second - PPS signal). Transmission of this signals may be done using the White Rabbit [*The White Rabbit Project*] technology, which ensures transmission of the 125 MHz master clock and the PPS signal. Synchronous commands can be in this approach transmitted in advance using the White Rabbit "critical data" channel, and then scheduled for execution at the particular time.

The second possibility is use of the GBT technology for TFC connection, however most of the bandwidth provided by this solution will be not used, and therefore its high cost seems to be not justified from the economical point of view.

The third possibility may be use of a technology similar to the TTC system [Taylor, "TTC Distribution for LHC Detectors"] used in CERN, with the PPS signal and synchronous commands transmitted using the 160 MHz clock or 80 MHz clock and biphase encoding. Possible problem in this approach is the fact that encoded data increase jitter of the tranmitted clock, and therefore a PLL based jitter cleaner will be needed to produce clock of required quality.

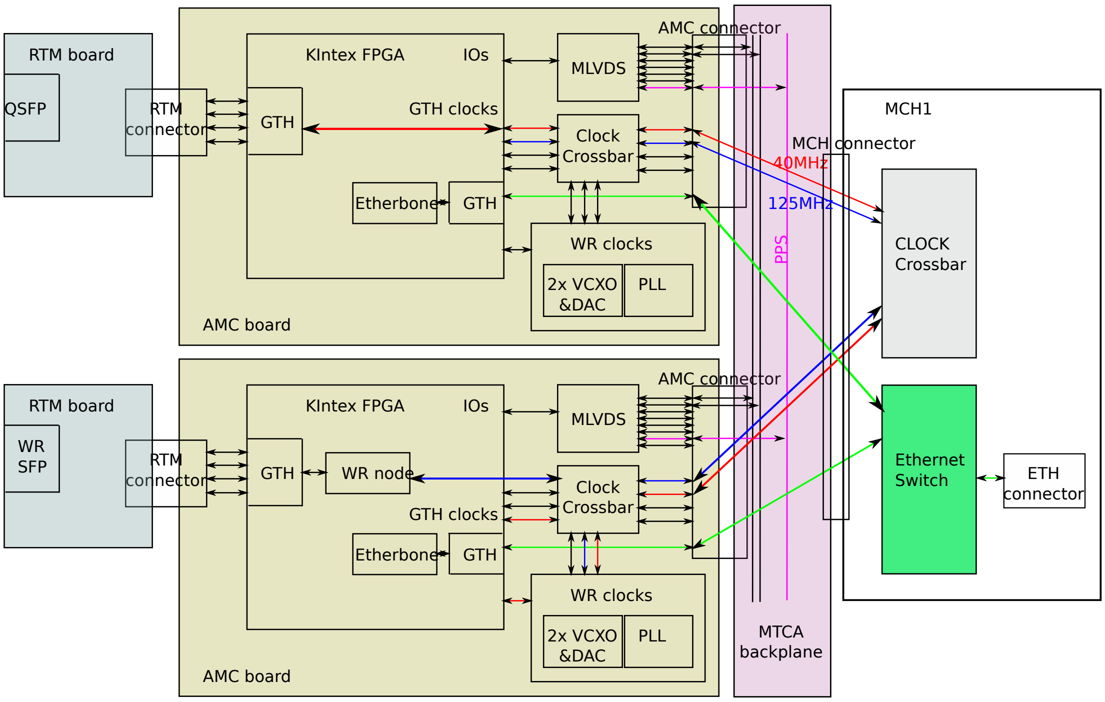
The DPB layer must also provide the TFC system with information about the "Busy" status of connected FEE systems and of DPB board itself. This information may be transmitted via single differential copper link, using the biphase or Manchester encoding to allow AC-coupling.

5.1.4 PCIe technology

5.2 Clock and time distribution

When DPB boards will be located in crates, the connection with the TFC system may be implemented only in one of them. This board should implement the TFC receiver, as described in Section 5.1.3. The received clock should be then converted to 120 MHz and submitted to jitter cleaning, to produce the reference clock for GBT-FPGA blocks used to communicate with FEE. The reference clock and the PPS signals will be transmitted to other boards located in the same crate using the clock crossbar resources on the backplane (e.g. the TCLKA for the Master Clock and TCLKB for the PPS signal).

Figure 5.3 presents a concept with White Rabbit based distribution of clock and synchronization.



]

Figure 5.3: White Rabbit based distribution of clock and synchronization
One of the AMC boards is configured as WR receiver. It receives WR signal using SFP transceiver installed on the RTM board. It has embedded PLL circuit that recovers WR clock and creates 120 MHz clock which is distributed using clock crossbar to the AMC connector. Together with the 120 MHz clock the PPS signal is distributed.

5.3 Flow control signals distribution

Each DPB board will either receive the information about buffer occupancy of connected FEE systems, or should actively monitor them. The received information will be used to elaborate the “Busy” status of the underlying FEE systems and of the DPB board itself. The “Busy status” of the whole crate may be found as the logical “or” of “Busy status” of all boards in the crate. This function may be easily implemented in the MTCA.4 (for Physics) crate, using “wired or” on one of 8 M-LVDS lines in the backplane, connected to the type 2 MLVDS receiver in the board connected to the TFC system. Second M-LVDS line may be used to transmit the “Stop DAQ” signal from the TFC system, requesting all DPB boards to stop data acquisition in connected FEE systems and to flush data buffers. The data acquisition will be restored later on using the synchronous command.

5.4 Synchronous commands distribution

Synchronous commands transmitted by the TFC system may be distributed from the board implementing the TFC receiver to other boards in the DPB crate using the remaining 6 MLVDS lines. If more than 6 commands are needed, it will be possible to encode commands. The backplane MLVDS lines are relatively slow, but if we assume that the time granularity for the synchronous commands may be only 12.5 ns or 25 ns, it should be possible to reliably transfer synchronous commands via these lines.

5.5 Control interface

Proposed solution allows two control scenarios:

- Usage of Ethernet switch (MCH) and Ethernet to Wishbone bridge to control FPGA and AFE resources. It is simple, low cost but requires additional FPGA resources needed to implement MAC and bridge.
- Usage of embedded AMC CPU, MCH PCIe switch and PCI Express to Wishbone bridge in FPGA to control resources. It saves FPGA resources since PCIe core is already hardcoded but requires additional CPU. It also provides faster configuration since PCIe accesses can be pipelined efficiently.

Both scenarios use standard Ethernet interface visible from outside. Figure 5.3 also presents the first scenario of control interface data flow. The Wishbone interface is used for creating the control interface. It is general purpose control bus foreseen for System On Chip solutions. From the outside world, through MCH switch, a standard Ethernet protocol together with IP library (Etherbone) can be used together with software library to directly control Wishbone over Ethernet. Etherbone transfers can be pipelined to increase the throughput of the system. The same interface will be used for AMC hardware diagnostics, FEE configuration and control of data processing algorithms.

The second approach resembles a standard PC architecture.

The Fat Pipe 1 (FP1) MTCA interface is used to connect all the FPGAs through the PCIe switch to the AMC CPU. Since PCIe node does not consume much resources, this approach gives a memory-mapped, easy to use interface for slow control, diagnostic data storage and reading.

5.6 Data path

The DPB layer should allow to minimize cost of the link between the FEE layer and the FLES layer by reduction of number of necessary links. This should be achieved both by transmitting the data via higher speed links (10 Gb/s instead of 5 Gb/s or less), and by aggregation of data received from multiple channels of the FEE layer. Depending on the detector, the aggregation may include local data processing and feature extraction. Before the aggregation, the data should be time odered. This implies implementation of additional buffer memory needed to compensate different latency of data received from different input channels. Moreover if the particular detector does not ensure delivery of time ordered data, an additional sorting stage is needed (however in this case, if combined data will be sorted, amount of memory needed to buffer incoming data may be reduced). The time sorted and aggregated data should be encapsulated in the microslice containersHutter, Cuveland, and V., *CBM Readout and Online Processing Overview and Recent Developments* and transferred to the block responsible for transmission to the FLES.

Degree of data reduction achievable for data from different detectors is detector specific.

There should follow subsections describing possibility of local data processing and aggregation for different detectors

5.7 Prototype, Evaluation

5.8 Next steps towards a CRI2

The CRI2 will probably be based on the Xilinx KU15P. This FPGA does not have SLRs, therfore the following approach seems prudent:

- use 36 GTH for 3 * 12 GBTx links
- use 16 GTY for a single Gen3 x16 interface -> no PEX chip anymore

The board will be designed with 3x MTP24 connectors, offering optical connectivity for up to 12x ROB3.

5.9 FPGA Design Overview

(Ingo Froehlich)

As all firmware projects are based on the same hardware, namely the CRI, the designs contain common and detector- (or ASIC-) specific modules. Here, one important goal is that all projects are based on the same common modules in order to allow for an easy and maintainable integration into the ECS system. In addition, the design flow must ensure, that all projects are up-to-date with the most recent versions of the common modules.

5.9.1 Common structure and workflow

5.9.2 Common modules

5.9.2.1 Wishbone interface

5.9.2.2 Zeropage

5.9.2.3 Alarm handler

5.10 Subsystem specific data processing

The subsystems are supposed to explain in the following subsections how the data streams originating from the data sources described in Chapter 4 are processed in the FGPA of the CRI. Please answer the following questions:

- How does the output data rate of the processing logic relate to the input data rate?
- How many input links are aggregated into one data stream?

5.10.1 BMON

(Adrian Rost)

Section 4.2.2

5.10.2 MVD

(Christian Muentz)

Section 4.3.1

5.10.3 STS

(Wojciech Zabolotny)

Sample text for the STS, to be completed by the subsystem:

The electronics subtree of the STS connecting to a single CRI was introduced in section 4.1.1.3. (There might be synergies between STS and TRD-SPADIC.) In the FGPA the data generated by the front-end electronics is processed as follows:

5.10.3.1 HCTSP-Bridge

(Michal Kruszewski)

5.10.4 MUCH-GEM

(Jogender Saini)

Section 4.1.1.4

5.10.5 MUCH-RPC

(Jogender Saini)

Section 4.1.1.5

5.10.6 RICH

(Adrian Weber)

Section 4.2.1

5.10.7 TRD-1D

(David Schledt)

Section 4.1.2.1

Sample text for the TRD, to be completed by the subsystem:

The electronics subtree of the TRD connecting to a single CRI was introduced in section 4.1.2. (There might be synergies between STS and TRD-SPADIC.) In the FGPA the data generated by the front-end electronics is processed as follows:

5.10.8 TRD-2D

(Claudiu Schiaua)

Section 4.1.2.2

5.10.9 TOF

(Esteban Rubio)

Section 4.1.3

Describe would you handle the data streams originating from the GET4. And how the data submitted to the FLIM interface is structured.

5.10.10 PSD

(Dmitry Finogeev)

Section 4.2.3

5.11 DCA - detector control agent

(Walter Mueller)

Chapter 6

Timing and Fast Control System

6.1 Overall concept

(Vladimo Sidorenko)

6.2 Timing and Fast Control Implementation

☢To-Do: Add text about TFC System here.☢

6.3 Data Throttling

The Compressed Baryonic Matter experiment (CBM) operating at high heavy-ion interaction rates of up to 10 MHz excludes conventional, latency-limited trigger architectures. Instead, CBM opted for a free-running readout. To be able to operate the experiment at highest interaction rates, despite beam intensity fluctuations introduced by the slow extraction of the SIS 100 synchrotron, a time-based throttling mechanism is investigated. To achieve a high number of complete readout events, different throttling strategies are compared and the throttling algorithms and parameters are optimized.

6.3.1 Introduction

The CBM experiment will study rare probes in heavy-ion collisions at the FAIR facility. The observation of detached vertices requires a topological trigger, which is too complex to be defined in hardware, and fully realized in software. CBM uses fast and radiation hard detectors with self-triggered front-end electronics and a free-streaming readout system. The data messages, time-stamped on activation of the respective detector channel are passed via a high-throughput data acquisition network to a large computer farm (First-Level Event Selector, FLES), where online event building and event selection are performed Toia and Selyuzhenkov, *CBM Progress Report 2017*Volker Friese, “Simulation and reconstruction of free-streaming data in CBM.”

At the maximum interaction rate, a raw data flow of about 1 TB/s from the front-end electronics through the Data Acquisition (DAQ) system is expected. However, fluctuations of the beam intensity are significant because of the slow extraction of the synchrotron Singh et al., “Slow Extraction Spill Characterization From Micro to Milli-Second Scale.” Consequently, on short time scales the data rates can exceed the maximum readout bandwidth and buffer capacity of the front-end electronics in the detectors. Such data overflow directly leads to incomplete events. A preliminary research has proven that the probability of complete or almost complete events can be significantly improved through a throttling mechanism.

Based on the available functionality of the detector front-end electronics, different strategies are compared in various experimental conditions and the throttling algorithms and parameters are optimized. The simulation structure is presented in Section 6.3.2. The throttling system and strategies for the STS-XYTER front-end ASIC of the CBM silicon tracking system (STS) are described in detail in Section 6.3.3. Section 6.3.4 compares the throttling strategies with various parameter settings according to the simulation results.

6.3.2 Simulation structure

A closed-loop simulation model comprising a hit generator, a data flow model and a result evaluation (see figure 6.1) has been set up. As the core, the data flow model is implemented in the Questa framework using the System Verilog language *ModelSim User’s Manual*. It calls Linux shells to invoke the front and back stages realized in C++/ROOT “IEEE Standard for SystemVerilog–Unified Hardware Design, Specification, and Verification Language.”

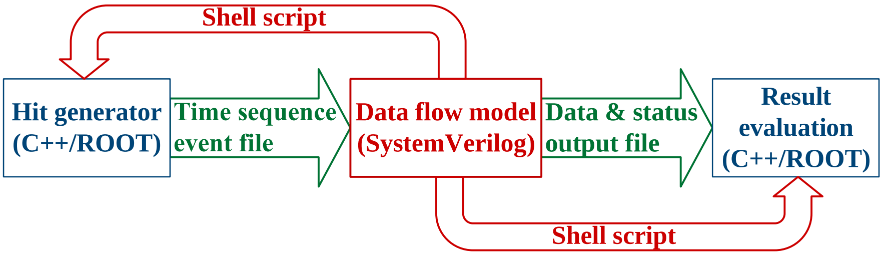

Shell script
Hit generator (C++/ROOT)
Time sequence event file
Data flow model (SystemVerilog)
Data & status output file
Result evaluation (C++/ROOT)
Shell script

Figure 6.1: Closed-loop simulation structure

A simplified hit generator is used for the initial simulations and shown in figure 6.2. Here, one event is one collision. The event size is random with a uniform distribution, which means the average number of hits per event is constant. For each event, hits have a uniform distribution on all detector channels. This can later be extended to more realistic distributions in a more complex model.

The hit rate R_{hit} is calculated with equation (6.1).

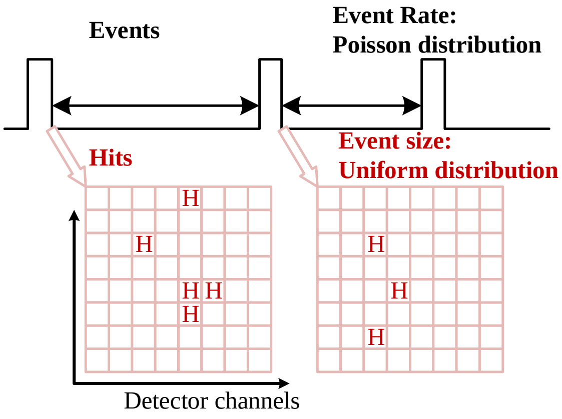


Figure 6.2: Hit generator with Poisson distribution

$$R_{hit} = S_{event} * R_{event} \tag{6.1}$$

In equation (6.1), S_{event} represents the event size, and R_{event} is the event rate. The event rate R_{event} is subject to Poissonian fluctuations and proportional to the beam intensity. For a more intuitive understanding of throttling results, the normalized hit rate R_{hit_N} is defined as the ratio of hit rate to readout bandwidth, as shown in equation (6.2).

$$R_{hit\_N} = R_{hit}/BW \tag{6.2}$$

In equation (6.2), BW represents the readout bandwidth of each ASIC, which depends on the configurable number of serial electrical readout links ("elinks") in the front-end ASIC.

As the input of the data flow model, the hits with the information of channel address and time stamp (i.e. the time when the hit is generated in the ASIC) are organized in a time sequence. The time sequence event file only includes valid hits after pileup correction. When two or more hits fire on the same detector channel, the hit pulses extended by the fast or slow shaper are piled up. So these pulses cannot be distinguished and are recognised as one hit by the front-end electronics of the STS-XYTER. To mimic the dead time of the ASIC in the simulation, if the interval of the adjacent hits on the same channel is less than 300 ns, only the first hit is saved and the others are removed.

The hits acquired at the end of the data flow model will be saved in output files for result evaluation. A "good" event which can be restored in the physics analysis is defined as an event where at least 95% of hits are saved at the end of the acquisition. The absolute number of good events in the same simulation time is considered as the criterion of the comparison of different strategies and without throttling. The comparison covers a range of the normalized hit rate R_{hit_N} which extends from a small fraction of the readout bandwidth up to twice the bandwidth limit.

6.3.3 Throttling system

6.3.3.1 The data flow model

This study is based on the CBM Silicon Tracking System (STS), which is the main tracking system closest to the target with a readout bandwidth of up to 50 MHits/s per front-end ASIC. The readout electronics (see figure 6.3) comprises 14,400 STS-XYTER ASICs, populating 1,800 Front-End Boards (FEB-8), interfacing to about 600 GBTx Readout Boards (ROB-3), connecting to about 80 Common Readout Interface cards (CRI) connected to a central Timing and Fast Control system (TFC). All of the components are under the supervision of the Experiment Control System (ECS)Toia and Selyuzhenkov, *CBM Progress Report 2017*Kasinski, Szczygiel, Zabolotny, et al., "A protocol for hit and control synchronous transfer for the front-end electronics at the CBM experiment."

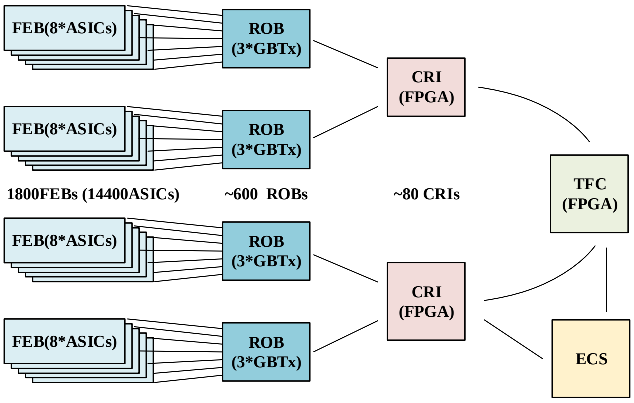


Figure 6.3: Hierarchy of the readout tree of the STS subsystem

Figure 6.4 is the throttling hardware functional diagram. Various parts of the throttling logic are implemented in the front-end ASIC and in the firmware for the CRI and TFC. In this simplified simulation, the model implements 32 STS-XYTERs. Each STS-XYTER comprises 128 readout channels. Each channel has a buffer FIFO of 8 words depth. However, the first word is actually the data latch stage that is stored in the front-end digital part, awaiting hits writing to the FIFO.

The throttling functionality is implemented as shown in Figure 6.5. A FIFO-almost-full flag is asserted once the FIFO contains 7 words. The ASIC counts the number of channels with almost-full FIFOs. The threshold for the number of channels in almost-full condition before triggering an alert is programmable Kasinski, Szczygiel, and Zabolotny, "Back-end and interface implementation of the STS-XYTER2 prototype ASIC for the CBM experiment." If exceeded, the ASIC reports an almost full condition to the CRI by immediately sending an alert frame in the next uplink frame slot. In the CRIs, the Alert Unit (AU) receives the alert frame, and transfers this alert signal to the FIFO Full Indicator (FFI). Meanwhile, the AU resets the ASIC status bits immediately and thus

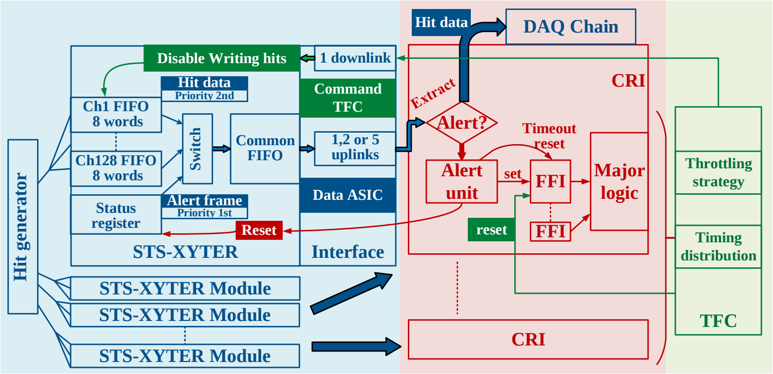
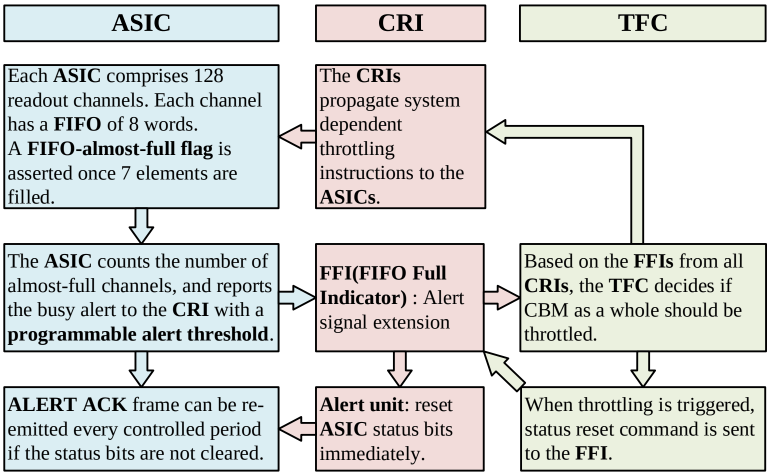


Figure 6.4: Hardware functional diagram

guarantees proper arbitration of incoming alerts and the resulting action. The alert signal is locked in the FFI and released by a timeout or throttling reset. Through FFI and AU, busy ASIC information is temporarily saved in the CRIs. After collecting the busy information from all CRIs at 50MHz clock, the TFC decides if CBM as a whole should be throttled. The CRIs receive the throttling decision from the TFC and propagate system dependent throttling instructions to the ASICs. According to current test work of the CBM DAQ system, the round trip time between ASICs, TFC and back is approximately 6 μs.

In the initial simulation, 5 elinks/ASIC in the uplink direction are active, providing approximately 50 MHits/ASIC/s as bandwidth limit. In equation (6.1), the S_{event} per ASIC averages 5 hits. When the average of normalized hit rate R_{hit_N} equals to 1, in other words, when the average hit rate R_{hit} equals the readout bandwidth BW, the averaged event rate R_{event} should be 10MHz, which is the goal of CBM experiment.



Each **ASIC** comprises 128 readout channels. Each channel has a **FIFO** of 8 words. A **FIFO-almost-full flag** is asserted once 7 elements are filled.

The **CRIs** propagate system dependent throttling instructions to the **ASICs**.

The **ASIC** counts the number of almost-full channels, and reports the busy alert to the **CRI** with a **programmable alert threshold**.

FFI(FIFO Full Indicator) : Alert signal extension

Based on the **FFIs** from all **CRIs**, the **TFC** decides if CBM as a whole should be throttled.

ALERT ACK frame can be re-emitted every controlled period if the status bits are not cleared.

Alert unit: reset **ASIC** status bits immediately.

When throttling is triggered, status reset command is sent to the **FFI**.

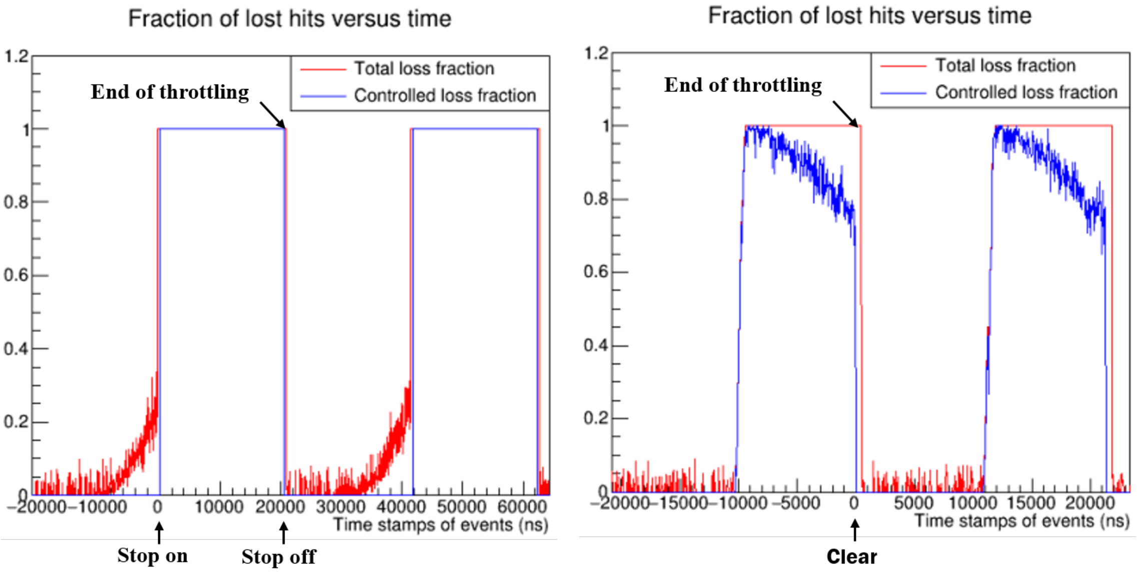
Figure 6.5: Communication structure

6.3.3.2 Two throttling strategies

In the present paper, two throttling strategies are investigated. They are abbreviated as “Stop” and “Clear” strategy. Upon a throttling decision, the “Stop” strategy is to stop accepting new hits in the channel FIFOs for a time long enough to fully read out these FIFOs and restart accepting hits afterwards, while the “Clear” strategy is to clear the buffers by resetting the channel FIFOs, then quickly re-enable data taking by releasing the reset. At the end of the throttling cycle all buffers are empty. A throttling cycle is defined from one end of throttling to the next end of throttling as shown in Figure 6.6. In the simulation, approximate 20 μs is expected to drain out all 128 channel FIFOs for each ASIC due to 5 active uplinks. According to the simulation, one throttling cycle is around 2 to 10 times the drain time. The total simulated time of 10ms is enough to cover in the order of hundreds of throttling cycles.

The throttling performance can be analyzed through the distributions of the fraction of lost hits shown in Figure 6.6. The total hit losses should be divided into two stages of data acquisition and data processing. The losses in the data acquisition stage can be classified into controlled and uncontrolled losses. If the hit rate exceeds the readout bandwidth for a short time, the channel FIFO becomes full and starts to miss next hit data. So these uncontrolled losses come from the channel FIFO overflow, where no detailed information about lost hits is available. Controlled losses represent hits lost due to the throttling interventions, i.e. masked at the FIFO inputs in the “Stop” strategy or cleared from the channel FIFOs in the “Clear” strategy. The controlled losses are the blue lines included in total losses of the red lines. The uncontrolled losses have adverse effect on the event reconstruction. To reduce uncontrolled losses and improve the fraction of the controlled losses in DAQ, two programmable throttling parameters are available. The first parameter is the alert threshold per ASIC, which is the number of channels reporting FIFO-Almost-Full (AF) conditions before generating an alert frame. The second parameter is the fraction of ASICs reporting alerts in the STS.

Assuming identical thresholds, the effect of uncontrolled losses in one throttling operation at different hit rates are schematically shown in Figure 6.7 in order to discuss the basic behavior. The four shaded areas represent the hits affected by the throttling. They are the saved hits with “Stop” strategy, while they show the discarded hits with “Clear” strategy. As defined in Section 6.3.2, one event which has more than 5% hit losses is considered as the bad event. From high to low hit rates, three possible consequences on the affected events can be distinguished. For the highest hit rate in purple, all shaded areas are filled with diagonal grids, representing all hits in bad events. The middle structure in red has hits in both of good and bad events. For the lowest hit rate in green, all shaded areas are filled with square grids and hits in good events. It’s obvious that more bad events are discarded by “Clear” strategy at higher hit rate, while more good events are saved by “Stop” strategy at lower hit rate. This is why they have different hit loss distributions in Figure 6.6 at a high hit rate ($R_{hit\_N} = 1.9$). The uncontrolled losses due to the round trip delay of the throttling decision can be clearly observed as the red lines before “stop on” in the “Stop” strategy. Depending on the analysis, the two strategies have respective



(a) (b)

Figure 6.6: The distributions of hit loss fraction versus time for 2 throttling cycles in the simulations ($R_{hit\_N} = 1.9$)

advantages in the opposite hit rate ranges. The same analysis and trend can also be extended to different alert thresholds.

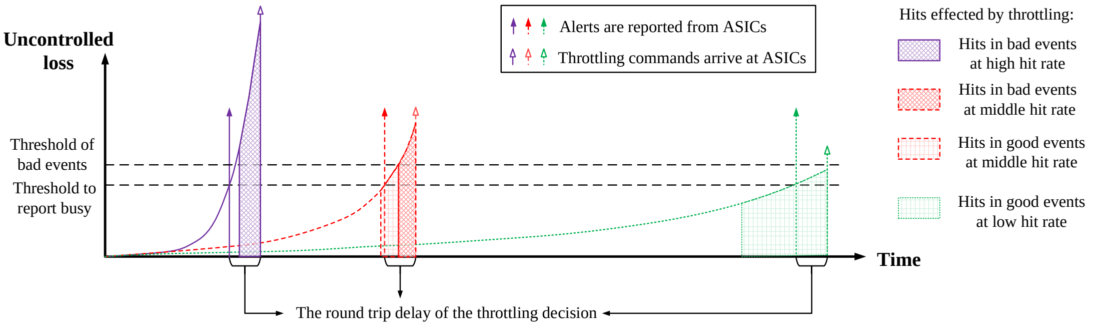

Uncontrolled loss
Threshold of bad events
Threshold to report busy
Time
Alerts are reported from ASICs
Throttling commands arrive at ASICs
Hits effected by throttling:
Hits in bad events at high hit rate
Hits in bad events at middle hit rate
Hits in good events at middle hit rate
Hits in good events at low hit rate
The round trip delay of the throttling decision

Figure 6.7: The discrimination of bad events and good events depending on the hit rate and the round trip delay of the throttling decision

The losses in the data processing stage are due to the command slot granularity of the CRI. In other words, the throttling operations among ASICs in the detector have jitters of 375ns. So the data processing loss is defined as the fiducial time interval of 375ns before and after each throttling operation.

6.3.4 Simulation results

6.3.4.1 Comparison of strategies

The “Stop” and “Clear” strategies are simulated with two temporal beam profiles of stable beam intensity and realistic beam intensity fluctuation.

Assuming a stable beam intensity, the event rate is only subject to Poissonian fluctuations. Shown in Figure 6.8 is, as a function of the normalized hit rates, how the number of good event changes. Green lines represent “all events", which is the event count of the hit generator.

Without throttling, the good events go quickly down to 0 when exceeding the bandwidth limitation. Using throttling, the number of good events improves significantly. When the hit rate exceeds the bandwidth limit, a large fraction of the available bandwidth is still used for almost complete, good events. With “Clear” strategy, the good events stay on plateau. However, with “Stop” strategy the good events have a slightly decreasing slope.

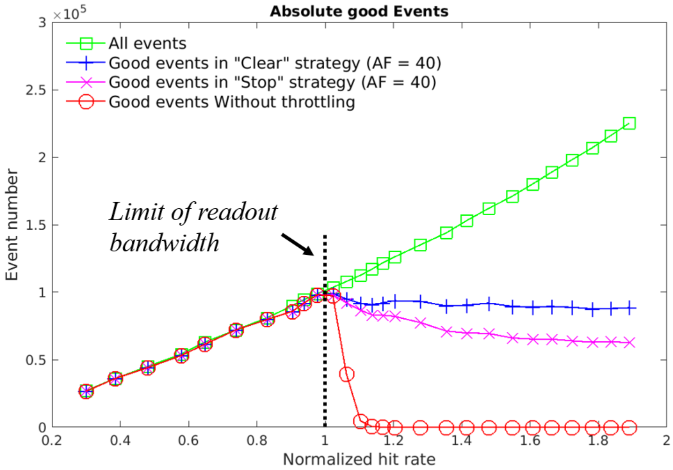
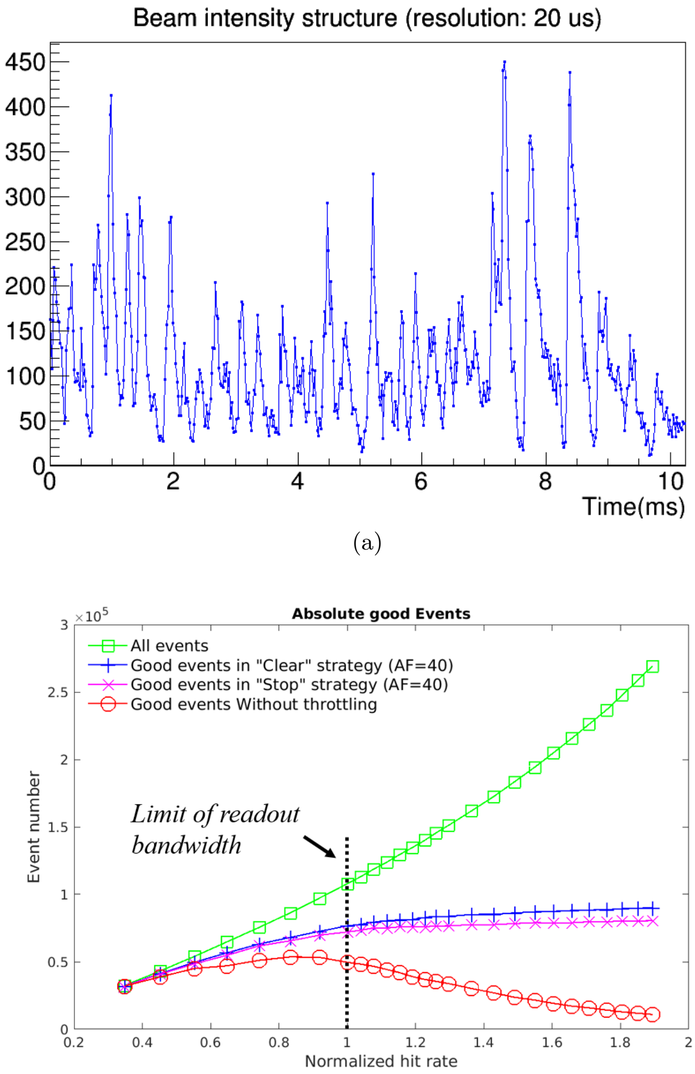


Figure 6.8: Simulation with stable beam intensity

Figure 6.9 (a) is the realistic beam intensity structure measured at the SIS 18 accelerator in a beam time in March 2019 Rost et al., “Performance of the CVD Diamond Based Beam Quality Monitoring System in the HADES Experiment at GSI*.” The measurement time resolution is 20 μs. In the simulation, the event rate is varied and scaled proportionally to the measured beam intensity, and obeys a Poisson process within each 20 μs time slot. Similar improvement of data quality with throttling can be observed in Figure 6.9 (b). Without throttling, the high intensity fluctuations lead to event losses already at rates below the bandwidth limit. The throttling significantly improves the number of good events over all hit rates, and provides some advantage already at rates below the bandwidth limit. When the hit rate exceeds the bandwidth, the good events exhibit a plateau with



(a)

(b)

Figure 6.9: Simulation with realistic beam intensity fluctuation

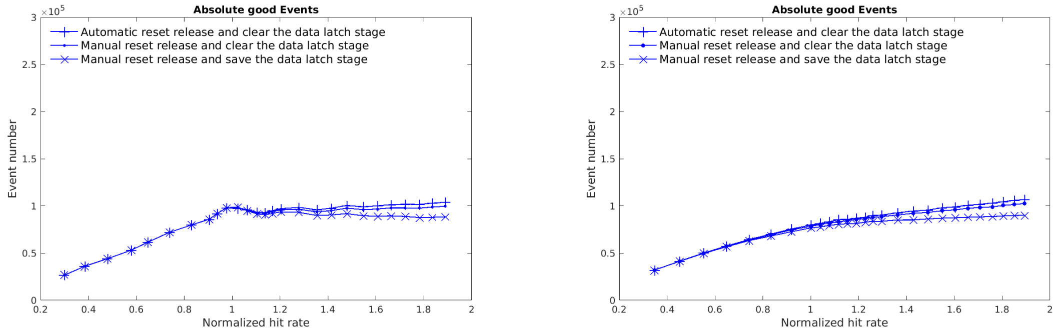
both strategies. However, the good events with “Clear” strategy are a little more than “Stop” strategy.

For both strategies, the alert thresholds per ASIC are 40 almost full FIFOs. But in the two simulations with different beam structures, “Clear” strategy has a better performance when the beam intensity increases. The difference can be explained with Figure 6.7 in Section 6.3.3.2. With higher hit rates, more bad events generated during the transport latency are saved by “Stop” strategy, which are discarded by “Clear” strategy. Meanwhile, more good events are masked in the drain out time of “Stop” strategy. In Section 6.3.4.3, the effect of different thresholds on the performance is investigated.

6.3.4.2 Optimization of “Clear” strategy based on hardware

Our model is based on the STS-XYTER ASIC. In the current version, after reset channel FIFOs, the ASIC is waiting for a command to release the FIFO reset. To reduce the latency of the throttling decision, the reset should be released automatically instead of manually. Because at least 2 downlink frames are required to release the reset by firmware, and the duration of a single downlink frame is 375ns, this modification can save 650ns.

In the introduction in Section 6.3.3.1, the data latch stage is the first word of the channel FIFO. However, the operation of the data latch stage is independent of the other words of the channel FIFO. In the simulation in Section 6.3.4.1, the data latch stage stays invariable when throttling is executed. As shown in Figure 6.10, to clear the data latch stage at the same time improves performance of “Clear” strategy obviously. With automatic FIFO reset release and synchronized clear of the data latch stage, the performance of the “Clear” strategy is improved as shown by the comparison in Figure 6.10.



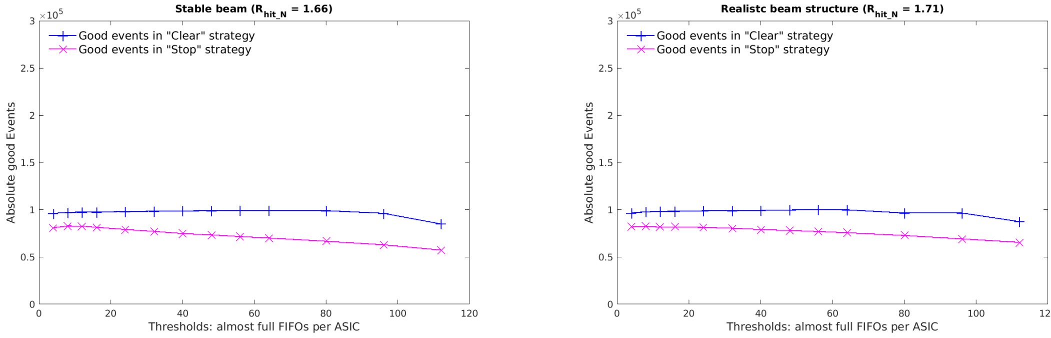
(a)

(b)

Figure 6.10: Simulation with the optimization of “Clear” strategy

6.3.4.3 Optimization of thresholds

After the optimization of the "Clear" strategy, the selection of thresholds is discussed in this section. In Section 6.3.4.1, both strategies use the identical thresholds. According to Section 6.3.3.2, properties of the two strategies mandate an earlier reporting of a "busy" condition in case of the "Stop" strategy than the "Clear" strategy. Though lower thresholds to trigger throttling improve the ratio of controlled loss, frequent throttling also increases the fiducial loss. So we investigate their performance changes as a function of the alert thresholds per ASIC. In any case, throttling is triggered when half of the ASICs report busy alerts.



(a)

(b)

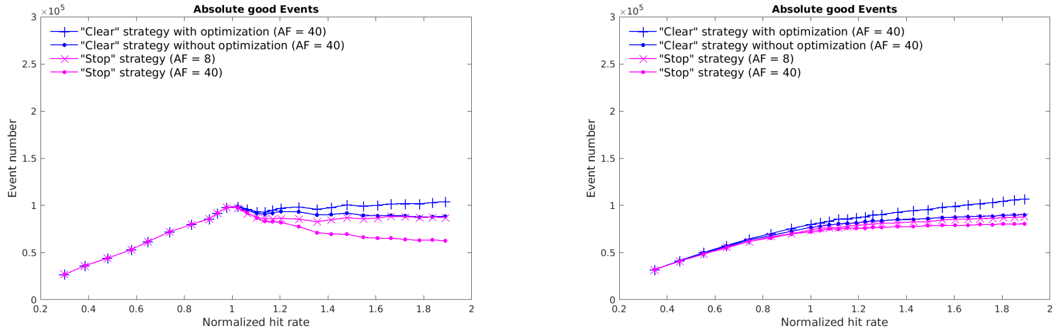
Figure 6.11: Scan of alert thresholds

In Figure 6.11, the almost full FIFO threshold is scanned in the stable beam and realistic beam intensity fluctuation. Their normalized hit rates are 1.66 and 1.71 respectively. The two situations show similar characteristics. With the identical threshold, the "Clear" strategy is better than the "Stop" strategy. In general, the "Clear" strategy is less sensitive to the threshold than the "Stop" strategy. The "Stop" strategy reached its best performance at a threshold of 8 almost full FIFOs per ASIC.

After the investigation of the alert parameters, the optimized performances of the two strategies are compared with the initial simulation of Section 6.3.4.1 in Figure 6.12. In both beam structures, the advantage of the optimization is more obvious with higher beam intensity. The optimized "Clear" strategy consistently shows better performance than the "Stop" strategy.

6.3.5 Conclusion

This time-based throttling is designed to adapt data acquisition to the high interaction rate of the CBM experiment with beam intensity fluctuation. Ideally, events saved into the large computer farm are almost complete.



(a)

(b)

Figure 6.12: Simulation with optimized parameters

In this paper, the “Stop” strategy and “Clear” strategy are compared. The performance of “Clear” strategy can be improved with optimization based on the electronics of the STS. And the “Stop” strategy is more sensitive to the throttling threshold. After the investigation of the basic throttling behavior in the present study, different event sizes distributed on different detector ASICs will be simulated in a more complex model based on real physics simulations. In this case, the “Clear” strategy is more flexible, while the “Stop” strategy needs longer drain out time because the active elink numbers of different ASICs have a distribution according to the detector geometry, leading to a lower readout bandwidth. According to the above simulation and analysis, the optimized “Clear” strategy gives better overall performance and more robust system behaviour.

Chapter 7

FLES Entry Stage

The First-level Event Selector (FLES) is the endmost part in the CBM read-out chain. Its task is to select data for storage by performing an online analysis of the physics data delivered by the detectors. As CBM is based on free-running, self-triggered detectors delivering timestamped data streams, there is no inherent event separation. Thus, classical approaches for global event building and event selection are not applicable. Instead of event building, the FLES has to perform *timeslice building*. Goal of the timeslice building is to combine the data from all given input links to time intervals and to distribute them to compute nodes. Event selection algorithms subsequently analyze these processing intervals and identify data for storage. For efficient event selection it is crucial that the algorithms can run on single nodes without much internode communication. Depending on the chosen CBM subsystem setup and selection scenario, two-staged interval building and event selection is conceivable. In this case, only partial interval building (based on a subset of all input links) is performed at the first stage, and the event selection algorithm can request subintervals for further analysis in a second step.

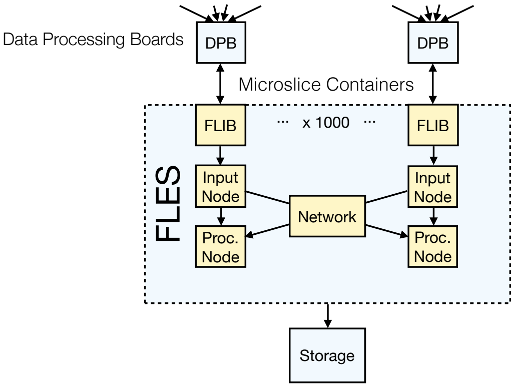


Figure 7.1: Architecture of the First Level Event Selector (FLES).

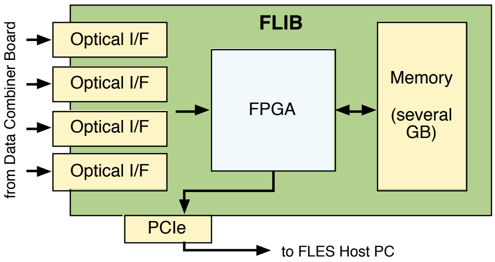


Figure 7.2: The FLES Interface Board (FLIB).

7.1 FLES Architecture

The FLES architecture is specially laid out for achieving the required throughput of the incoming data exceeding 1 TByte/s and to provide the necessary computation efficiency for flexible data processing in real-time. The conceptual design of the FLES is shown schematically in Fig. 7.1.
Special FPGA-based input cards collect the data coming from the detectors via 1000 10-GB/s optical fibers, pre-process and forward them to the HPC Cluster. This cluster is implemented as a scalable super computer consisting out of 1000 custom ware nodes, thus providing a flexible, maintainable and cost efficient system. Some of the nodes are, additionally to their data processing task, assigned as input nodes where the data segments arrive and are then distributed throughout all processing nodes forming complete sets of data. This task requires a fast event building network and furthermore special designed software that makes best use of the computer architecture.

The FLES system will be situated at the new FAIR data center ("Green IT Cube"). The building yields an ideal infrastructure and allows for an economic and energy efficient use of the PC's. This local separation from the CBM read-out chain is possible, because the FLES collects the experimental data untriggered and does not require fast access to the front-end electronics.

7.1.1 Input Interface

The FLES input interface defines the junction between FLES and the DAQ system. It has to handle the incoming detector data stream, buffer and prepare data for subsequent timeslice building. To enable timeslice building in the FLES three mayor requirements have to be fulfilled:

- FLES has to be able to *partition subsystem data into time intervals* that can be combined to processing intervals, each containing data from the same interval of hit time for all contributing subsystems.
- A certain overlap between subsequent intervals is required to avoid losing physical events at the borders of processing intervals without requiring expensive internode communication. Thus, the input interface has to facilitate the *selection of overlap regions.*
- For partial timeslice building, a means of *selecting subintervals* for the second stage of event selection is needed.

As described in Chapter 4 the CBM read-out tree is based on streaming protocols using continuous data streams that contain messages in subsystem-specific formats. In general, these messages are stateful, i. e., the information depends on the stream history. Epoch markers, which carry the uppermost bits of the time information, are an example of this. In addition not all data sources guarantee strict time sorting of the individual messages inside a stream. This leads to several challenges when trying to fulfill the given requirements.

In this section, a concept as well as a prototype implementation for the FLES input interface is presented. Its aim is to allow for efficient timeslice building in the FLES without the need to convert all detector data messages into a common global format.

7.1.2 The Microslice Concept

The fundamental idea of the microslice concept is to partition the message streams from the detectors into specific fragments, *microslices.* Prior to processing in the FLES, all data messages are encapsulated in *microslice containers (MCs)* of a globally defined structure. Microslices have two major properties: they cover a constant timeframe of hit time, which is the same for all subsystems, and the data content of an MC is self-contained.

Each MC consists of a descriptor in a global format and a block of subsystem-specific data. As MCs are constant in time, the size of the data block may vary. The descriptor provides information needed for timeslice building – especially the start time of the microslice. Thus, timeslice building can rely on combining subsequent MCs from all subsystems to one processing interval. As all needed information for timeslice building is available in the MC descriptor, there is no need for a global detector data format. Consequently subsystems are free to implement data formats solely optimized for online analysis without side effects on FLES timeslice building. Self-containment of MCs ensures that no information is lost in the timeslice building step by cutting the streams. Overlap between processing intervals can be achieved by duplicating MCs at the borders of the processing interval. For partial timeslice building, one MC is the smallest amount of data which can be requested for the second stage.

The length in time of a microslice has to be comparably small against the length of a processing interval to allow for efficient handling of interval overlap and partial timeslice building. On the other hand, smaller microslices lead to bigger overhead due to MC descriptors. Assuming a microslice length of $T = 1\,\mu s$ and an average input link rate of 1 GByte/s, the average data size per MC is 1 kByte. In this case, a 32-byte MC descriptor leads to an acceptable overhead of less than 4 %. Assuming a length of a processing interval of 100 µs, the timeslice building system can select subintervals at a granularity of 1 %.

The creation of MCs has to take place before any further processing steps in the FLES input interface. For each consecutive, fixed-length time interval, exactly one MC has to be generated by a subsystem-specific design. Therefore, the design has to keep track of the timing information in the data stream and partition it accordingly. During event analysis in the FLES, there is no guarantee which and how many MCs are processed together on the same node. Therefore, each microslice block of data has to be stateless and self-contained, i. e., all information needed by the event reconstruction algorithm of the corresponding subsystem has to be available in each MC.

The microslice generation process implies that all detector front-ends have to be synchronized, and possible FEE timing offsets have to be calibrated at least at the level of creation of MCs to enable the required conversion from FEE local time to the global experiment time. To limit the FEE calibration effort at this stage, a defined small uncertainty in time is acceptable when assigning messages to microslices.

As the DPB is the last data processing component in the CBM readout chain that by design has to contain subsystem-specific firmware, it is feasible to implement the generation of MCs in the DPB layer. Nevertheless this is no consequence of the microslice concept but is considered to be the most practical way. If it turns out superior the generation of MCs can also be implemented on the FLES Interface Board.

7.1.3 Microslice Containers

In order to implement the input interface into the software framework of the FLES, a shared communication ground is established, meaning a mutual data format that fulfills the requirements of the host PC to build complete datasets. The solution to these demand yields the concept of Microslice containers.

Microslice containers introduce a special global container format holding the data of the detector in a globally defined and self-contained container format, which generates an overhead of less than two percent. An important feature of the Microslice container is, that they hold all necessary information about the data (i.e. start time, origin, etc. ...) and thus all Microslices packed in containers can be distributed independently of each other to the input nodes.

Furthermore, a Microslice covers a constant timeframe of hit time 1 µs, which is the same for all subsystems. As the time interval is constant the size of the data block may vary because data production rates fluctuates. For each microslice a container is generated. The content may be empty if no data is available for the corresponding time interval. All in all, the container format facilitates a timestamp based, detector independent access of the experimental data coming from all detector segments.

7.1.4 Timeslices

The data from the input interface is forwarded as Microslices to the input nodes. At this point all Microslice containers from one detector arrive at the same input node. However, the goal here is to concentrate data that has a similar time stamp from all detector segment links in a single processing node and to combine them to a Timeslice. A Timeslice is the fundamental data structure that contains an interval of the continuous detector run-time of all detector links and provides access to the Microslices. This is necessary because the experimental data is streamed unfiltered (i.e., CBM detects all particles untriggered) and Timeslices allow to access this data for subsequent unbiased event reconstruction. This poses a great benefit over typical collider experiments, which are triggered and therefore only a small amount of information are saved in already categorized events.
The Timeslices are arbitrarily cut at a point in time disregarding the actual data contents. Hence, an event might start on one Timeslice and end on a different one, which would require both Timeslices for later reconstruction. However, this causes additional traffic on the network, that is best avoided. Furthermore differences in detector readout timing performance need to be accommodated. The solution to these problems is to create an *overlap region*. The overlap region includes a small number of Microslices, which are duplicated at the border of a Timeslice and then contained in two consecutive Timeslices at the cutting edge. This procedure ensures that every Timeslice can be processed independently.

- **☢To-Do: über länge der overlap region philosophieren☢**

Efficient Timeslice building requires a fast network, wich allows to distribute the great amount of incoming data, i. e. Microslices, from the input nodes to the processing nodes. This is best resolved using an Infiniband FDR network and a high quality and special developed software, cf. 7.3.

☢To-Do:

- **partial Timeslice building**
- **different trigger scenarios**

☢

7.1.5 Event Reconstruction

For each Timeslice the associated processing node performs the required feature extraction and online analysis described in Chapter **☢To-Do: Add reference here: cha:FLES Computing☢**. The challenge here is to deal with the high event rate of 10MHz in combination with non-trivial trigger criteria requiring partial of full event reconstruction. To reach the necessary computing power, all modern computing architectures are utilized, including many-core CPUs, graphics cards (GPUs), as week as novel languages such as OpenCl to allow for parallel programming and processing. Furthermore the FLES building network is used as an interface to the mass storage system, making perfect use of the given resources.

7.1.6 FLES Start-up Version

☢To-Do:

- **initial computing requerements**
- **full connectivity v.s. processing power**

☢

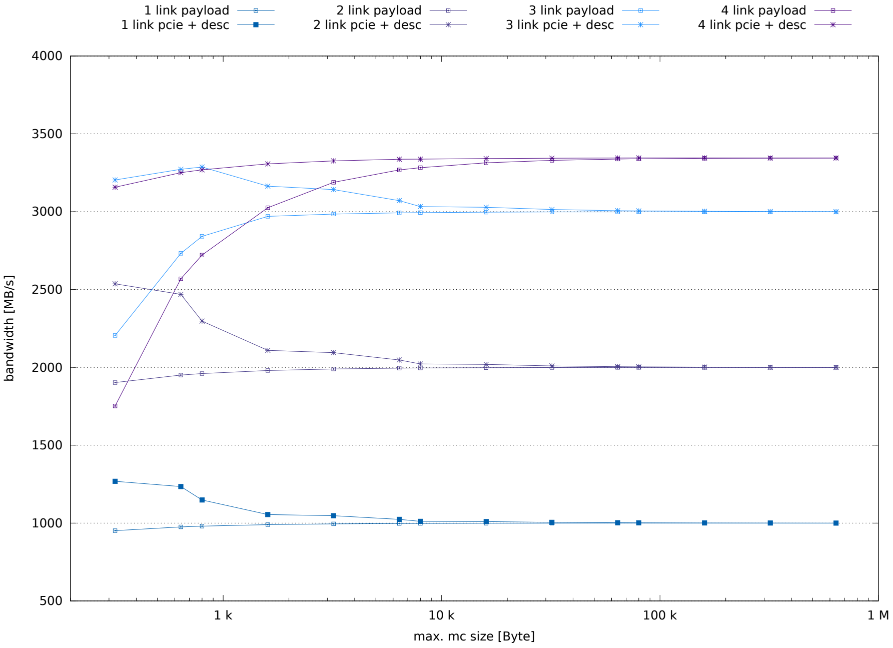


Figure 7.3: PCIe bandwidth messurement

7.2 Input Interface

7.2.0.1 Messurements

7.3 Timeslice Building

7.3.1 Time slice building on microslices

☢To-Do: Contents: Flesnet (buffer management, real world, real data) [Jan]☢

In the CBM experiment, detector raw data consists of a continuous stream of detector messages with no immediate association to a collision event. To efficiently manage this raw data, it is first packaged into Microslices (see Section 0), each representing data from a specific short amount of time and a specific part of a detector. The Microslices for a certain time interval are then combined over all detector parts in a single node. This *Timeslice building* replaces the traditional step of *event building* in a triggered experiment.

As several important event selection scenarios in CBM require an analysis of the full event, this Timeslice building has to be performed at the full data rate generated by the detectors. To implement overlapping timeslices (see Section 0), it should in addition be possible to duplicate a configurable number of microslices at the timeslice borders, thus generating Timeslices that can be analyzed independently. At an incoming data rate exceeding 1 TByte/s, Timeslice building is a challenge even for today's computer architectures. The implementation should be especially optimized with respect to memory bandwidth usage and network load.

In this section, the results of a prototype software implementation (*flesnet*), which addresses these challenges, are presentend.

7.3.1.1 Efficient memory management

Limited memory bandwidth is one of the most important bottlenecks in today's computers, and memory interface speed continues to increase slower than CPU core performance or density. The general method of mitigating this situation is the use of caches. However, in the case of Timeslice building, the required data buffer sizes are too large to fit in a cache. To archieve a good use of the available resources, it is therefore crucial to minimize memory access in the respective software.

InfiniBand InfiniBand is currently the leading networking technology at the world's fastest HPC clusters. In the TOP500 list of November 2013, of the top 50 clusters with non-custom interconnect, 40 were using InfiniBand.

The InfiniBand networking technology has a favorable architecture that allows for efficient data transfer using remote DMA (RDMA) semantics. Data is directly transfered from and into buffers managed by the user application. Using an optimized buffer structure, no memory bandwith has to be wasted for expendable copy operations.

The buffer structures used for handling detector data in microslices are shown in Figure 0. Per input link, the FLIB (see Section 0) writes to two large ring buffers, a data buffer and a descriptor buffer. **☢To-Do: continue here!☢**

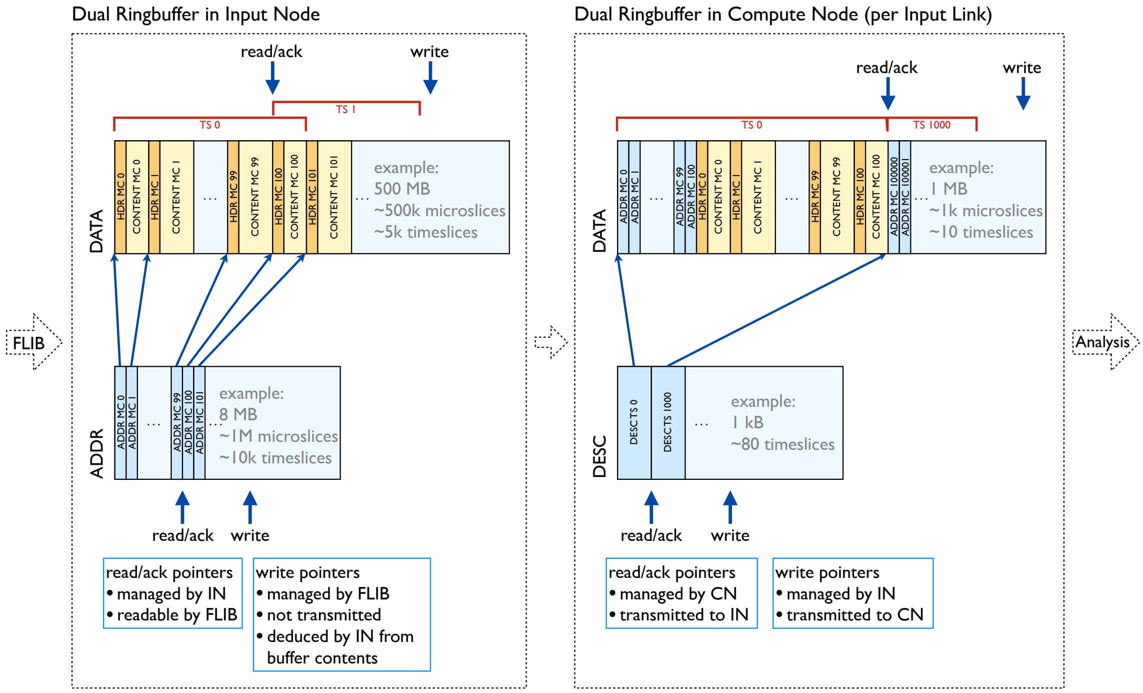


Figure 7.4: Overview of FLESnet buffer management. Both sender and receiver maintain sets of dual ring buffers to store data and descriptors.

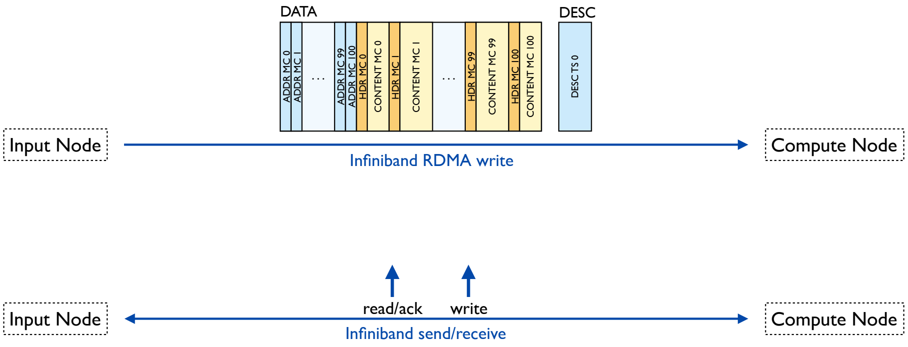


Figure 7.5: InfiniBand Transactions during Timeslice building. Data and descriptors are transmittend via RDMA write, asyncronous read/write pointer updates independently via send/receive.

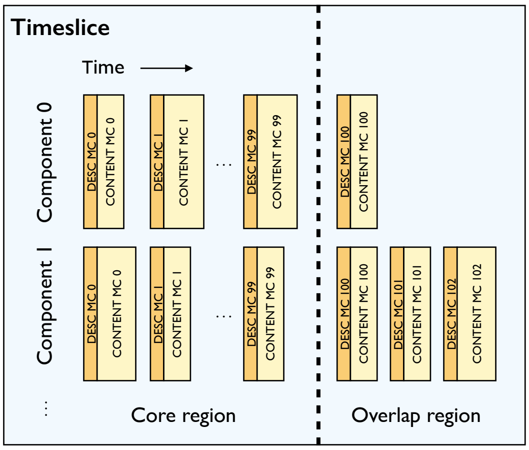


Figure 7.6: The Timeslice data structure is the primary interface to the online reconstruction code.

7.3.1.2 Software interface to the reconstruction code

7.3.1.3 Timeslice building performance

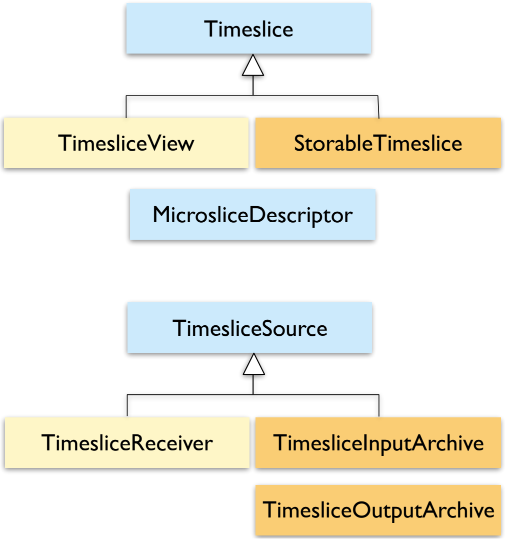


Figure 7.7: The FLES Timeslice API provides identical access to online and stored Timeslices (here: main access classes).

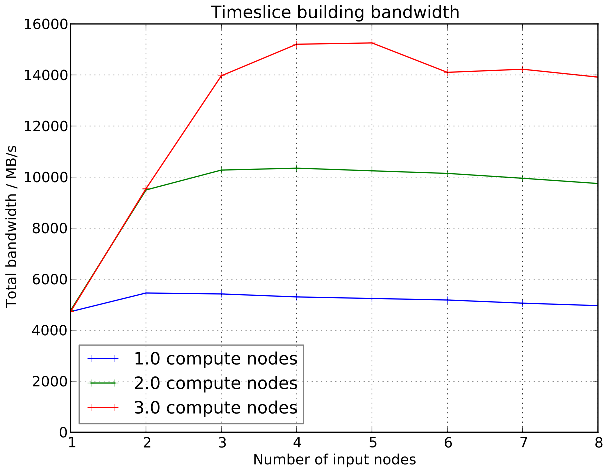


Figure 7.8: Timeslice building bandwidth. TODO: Plot of better results!

Figure 7.9: Timeslice building test setup (MicroFLES) with InfiniBand FDR switch.

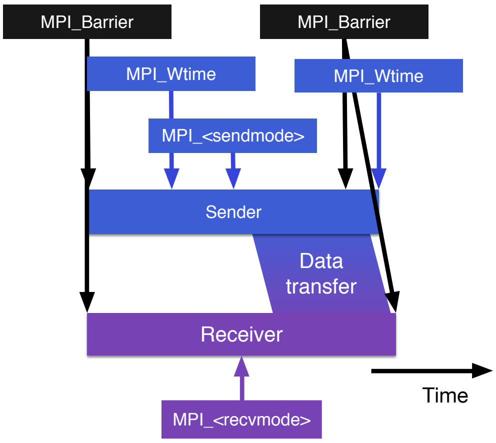


Figure 7.10: Layout of the MPI micro benchmark to test the bandwidth for Timeslice building via MPI.

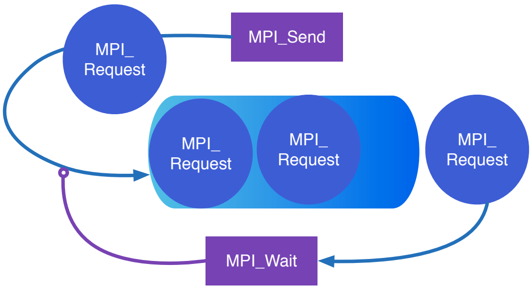


Figure 7.11: Pipeline scheme employed to improve the bandwidth reached using MPI for Timeslice building.

7.3.2 Timeslice building performance

☢To-Do: Contents: MPI (small scale, large scale) [Helvi]☢

The Message Passing Interface (MPI) is a high-level API that is able to handle multiple network systems. Therefore, it is a promising technology to use for the Timeslice building rather than depending on software specially designed for Infiniband. The micro benchmark, described in the following sections, tests whether or not MPI is capable of reaching the necessary bandwidth required for Timeslice building.

7.3.2.1 Microbenchmark

The benchmark is schematically drawn in Fig. 7.10 and run using OpenMPI 1.6.5 on the MicroFles test system at GSI (cf. 7.4.2).
Two processes, sender and receiver, are set up and a certain amount of bytes is transferred in packages between them. Directly before and after the data transfer the time since an arbitrary time in the past is recorded via the function MPI_Wtime and used later for bandwidth calculations. It is important to post a MPI_Barrier before the time measurement. Prior to the data transfer it assures that both processes are at the same point in the program and no additional waiting time is recorded. At the end, the time measurement stops only if sender and receiver returned to secure that the whole message has passed the network and arrived at the receiver.

The MPI send/receive call can be chosen from various options posed by MPI. There are three major communication procedures: blocking, non-blocking and persistent. Each of the send procedures can be used in one of the following modes: synchronous, ready, buffered or default. The MPI massage passing matches a push communication mechanism, where data transfer is effected by the sender. Hence, there is only one receive mode provided for all three receive procedures. The benchmark revealed that the nonblocking procedure with a synchronous sending mode is the best method for reaching the highest bandwidths when implemented with a pipeline.

This result is owed to the fact that the non-blocking procedure separates communication from the main program instructions. The sender returns immediately continuing with the program and the message is sent in an additional thread in the background. The process can continue with other tasks until an MPI_Wait is posted. The wait assures that the message has been transferred successfully and it is safe to reuse the send buffer. The input to MPI_Wait is an MPI request object handle, i. e. MPI_Request. This handle is returned by the non-blocking communication calls and enables access to the request object. This object is created when an MPI_Issend is posted and it contains all the information in the arguments to the message passing call, plus the communication mode and most important its status.

The pipeline depicted in Fig. 7.11 functions as follows. When the send call is posted, an MPI_Request object handle is created and stored in a C++ std::queue. This queue can hold a defined amount of requests. When it is full, the first request is taken out and checked for the status of the message. If the message was successfully delivered the next send can be posted and the request stored in the queue. This procedure allows to make the best use of the MPI capacity: while a message is transferred the next sends can be posted and checked later for correct transmission.

The results are shown in Fig. 7.12. The first plot demonstrates the pure performance of

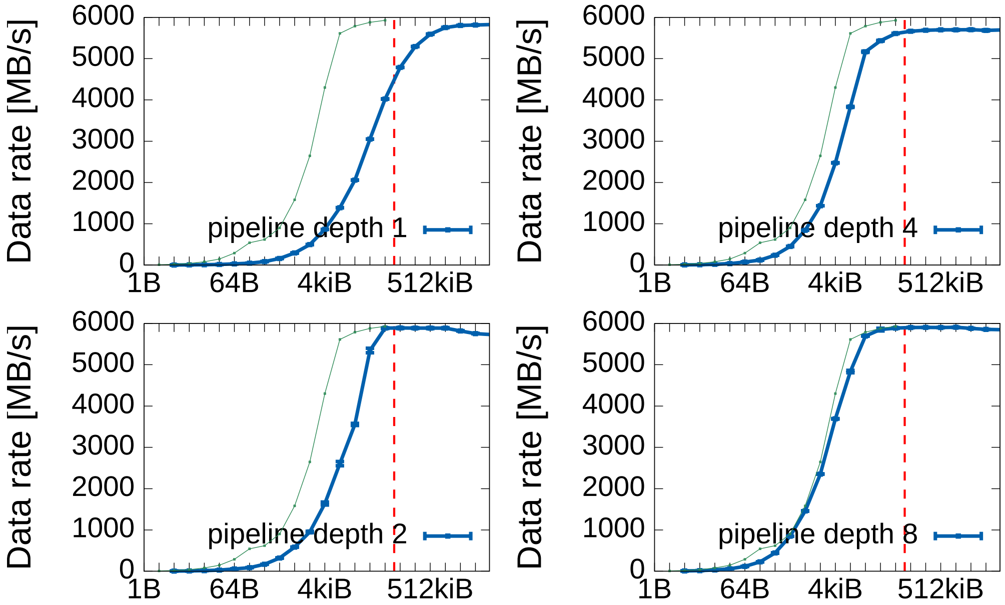


Figure 7.12: Data rate plots for MPI_Issend (blue curve) compared to direct data transfer using infiniband verbs (green curve) and typical Microslice container size of 100 kByte (red line)

sending messages via MPI. The pipeline depth of one implicates a blocking send, because the sender waits for correct data transmission directly after the call is posted. For an expected size of 100 kBytes for the Microslice containers, that are exchanged throughout the network, it does not yield the optimal data rate as compared to Infiniband Verbs. This situation changes significantly when the pipeline is deployed. With a pipeline depth of two, the peak bandwidth of 6 GByte/s can already be reached. For or a pipeline depth of 8, all packages can be sent at the highest possible rate provided by Infiniband.
This is a promising result. For a simple test set-up, the bandwidth for relevant package sizes of the high-level API MPI is as high performant as with the low-level Infiniband Verbs. On this basis, MPI can be considered as a possible technology for Timeslice building, opening the opportunity to develop a software framework more independently of the underlying network technology.

7.3.3 Large scale Benchmark

☢To-Do: still to come☢

7.4 Prototypes

7.4.1 mCBM CRI

7.4.2 Micro-FLES prototype system

☢To-Do: [Jan]☢

7.4.3 LoeweCSC HPC cluster

Chapter 8

Physical Connections and Layout

8.1 Optical network in the CBM building

8.2 Entry Node cluster in Room E40.017

8.3 Data backbone to the Green IT Cube

The data produced by the detector front-end electronics is in most cases aggregated by GBTx data concentrator ASICs and sent via optical links to the ntry nodes in the DAQ room located in E40.017. The links run at 4.8 Gbps, use OM4 multi-mode fibers, and are terminated by CRI PCIe cards in the entry nodes. After suitable preprocessing the data is sent from the entry nodes in the CBM building to the worker nodes in the GSI Green IT Cube. This connection will most likely be based on InfiniBand over single-mode fibers. The overall concept is visualized in Fig. 8.1.

The optical links from cave to DAQ room are sub-divided into two segments

- a set of high density trunk cables from the DAQ room to fixed patch panels in the cave, which offer MTP-12 interconnections. These panels are located left and right of the magnet for MVD and STS, and on the right cave wall for all other detectors.
- cables from these patch panels to the location of the on-detector sites where the data sources are located. These cables have lower density, will be flexed during operation when detectors are moved, and are therefore easily replaceable.

The E10 to E40 trunk cables as well as the CBM to Green IT Cube cables are considered common infrastructure, mainly to ensure a uniform system which is centrally planned and installed. The optical connection from E10 patch panel to the detectors is part of the respective detector projects.

In the DAQ room the MTP-12 fiber connections need to be re-grouped into MTP-48 bundles in order to match the interface of the CRI cards. The cost estimate shown in Table 8.1 is based on 144 fiber cables which have been successfully used in the mini-CBM setup.

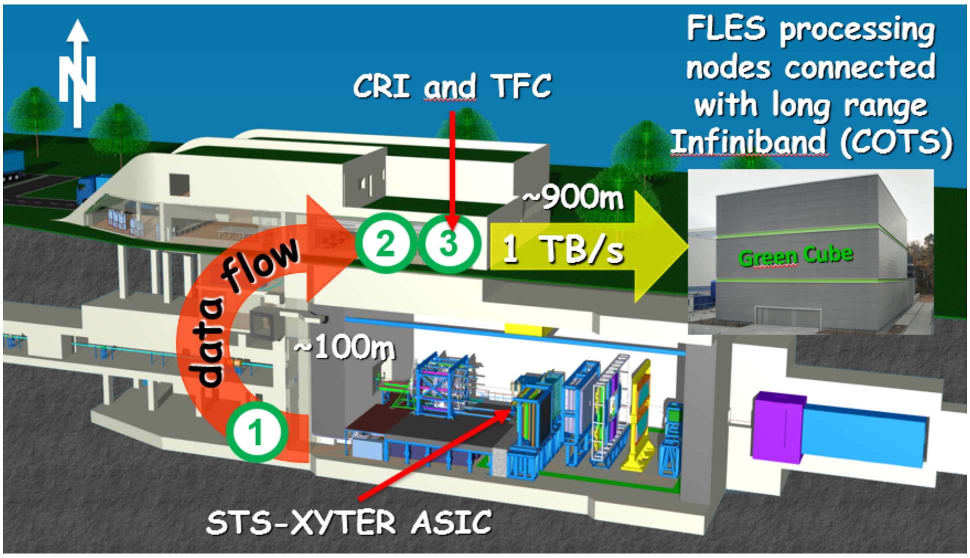

N
CRI and TFC
FLES processing nodes connected with long range Infiniband (COTS)
~900m
1 TB/s
Green Cube
data flow
~100m
1
2
3
STS-XYTER ASIC

Figure 8.1: Sketch of the data flow of the CBM experiment. The data backbone links the subsystems in location (1) to the DAQ room in location (3). The experiment is operated from the Control Room in location (2).

The cost for the CBM to Green IT Cube optical fibers is also based on 144 fiber cables, see Table 8.2.

Table 8.1: Cost of data backbone (E10 to E40) (in kEuro).

item	units	price	sum	comment
MPO/MTP-trunk-cable, 100m, OM4	56	6.200	347.2	
patch panels (24x MTP each)	56	0.100	5.6	
installation cost per 100 m	56	0.220	12.3	
installation cost per hour	140	0.033	4.6	
4x MTP-12 to 1x MTP-48 adapter	168	0.315	52.9	
sum overall			422.6	

Table 8.2: Cost of data backbone (E40 to Green IT Cube) (in kEuro).

item	units	price	sum	comment
MPO/MTP-trunk-cable, 1000m, SM	10	9.500	95.0	
patch panels (24x MTP each)	10	0.100	1.0	
installation cost per 100 m	100	0.220	22.0	
installation cost per hour	100	0.033	3.3	
sum overall			120.3	

Chapter 9

Project Organization

Project organization (V. Lindenstruth et al)

9.1 Schedule and Milestones

9.2 Installation and Commissioning

9.3 Costs

9.4 Responsibilities

Bibliography

CBM Progress Report 2017. Tech. rep. CBM Progress Report 2017. Darmstadt, 2018, V, 214 Seiten : Illustrationen, grafische Darstellungen.

Ablyazimov, T. et al. "Challenges in QCD matter physics – The scientific programme of the Compressed Baryonic Matter experiment at FAIR." In: *Eur. Phys. J.* A53.3 (2017), p. 60. DOI: 10.1140/epja/i2017-12248-y. arXiv: 1607.01487 [nucl-ex].

Aduszkiewicz, A. et al. "NA61/SHINE at the CERN SPS: plans, status and first results." In: *Acta Phys. Polon.* B43 (2012), p. 635. DOI: 10.5506/APhysPolB.43.635. arXiv: 1201.5879 [nucl-ex].

Andronic, A. et al. "Hadron production in central nucleus-nucleus collisions at chemical freeze-out." In: *Nucl. Phys.* A772 (2006), pp. 167–199. DOI: 10.1016/j.nuclphysa.2006.03.012. arXiv: nucl-th/0511071 [nucl-th].

– "Hadron production in ultra-relativistic nuclear collisions: quarkyonic matter and a triple point in the phase diagram of QCD." In: *Nucl. Phys.* A837 (2010), pp. 65–86. DOI: 10.1016/j.nuclphysa.2010.02.005. arXiv: 0911.4806 [hep-ph].

Aoki, Y. et al. "The order of the quantum chromodynamics transition predicted by the standard model of particle physics." In: *Nature* 443 (2006), pp. 675–678. DOI: 10.1038/nature05120. arXiv: hep-lat/0611014 [hep-lat].

Baron, S. et al. "Implementing the GBT data transmission protocol in FPGAs." In: 2009, pp. 631–635. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84884194925&partnerID=40&md5=7581aad60877ae322fa0267e896266f9.

Bass, S. A. et al. "Microscopic models for ultrarelativistic heavy ion collisions." In: *Prog. Part. Nucl. Phys.* 41 (1998), pp. 255–369. DOI: 10.1016/S0146-6410(98)00058-1. arXiv: nucl-th/9803035 [nucl-th].

Bauer, G. et al. "10 Gbps TCP/IP streams from the FPGA for the CMS DAQ event-builder network." In: *Journal of Instrumentation* 8.12 (2013). URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84892896325&partnerID=40&md5=a968cb3e91459d3b8348d826541cd9dd.

Bazavov, A. et al. "The chiral and deconfinement aspects of the QCD transition." In: *Phys. Rev.* D85 (2012), p. 054503. DOI: 10.1103/PhysRevD.85.054503. arXiv: 1111.1710 [hep-lat].

Blaschke, D. et al. "Topical issue on exploring strongly interacting matter at high densities - NICA white paper." In: *Eur. Phys. J. A* 52.8 (2016), p. 1. ISSN: 1434-601X. DOI: 10.1140/epja/i2016-16267-x.

Borsanyi, S. et al. "Is there still any T_c mystery in lattice QCD? Results with physical masses in the continuum limit III." In: *JHEP* 09 (2010), p. 073. DOI: 10.1007/JHEP09(2010)073. arXiv: 1005.3508 [hep-lat].

Ehehalt, W. and W. Cassing. "Relativistic transport approach for nucleus nucleus collisions from SIS to SPS energies." In: *Nucl. Phys.* A602 (1996), pp. 449–486. DOI: 10.1016/0375-9474(96)00097-8.

Fodor, Z. and S. D. Katz. "Critical point of QCD at finite T and μ_B, lattice results for physical quark masses." In: *JHEP* 04 (2004), p. 050. DOI: 10.1088/1126-6708/2004/04/050. arXiv: hep-lat/0402006 [hep-lat].

Friese, Volker. "Simulation and reconstruction of free-streaming data in CBM." In: *Journal of Physics: Conference Series* 331.3 (Dec. 2011), p. 032008.

Friman, B. et al. "The CBM physics book: Compressed baryonic matter in laboratory experiments." In: *Lect. Notes Phys.* 814 (2011), pp.1–980. DOI: 10.1007/978-3-642-13293-3.

Fukushima, K. and T. Hatsuda. "The phase diagram of dense QCD." In: *Rept. Prog. Phys.* 74 (2011), p. 014001. DOI: 10.1088/0034-4885/74/1/014001. arXiv: 1005.4814 [hep-ph].

Gutbrod, H. et al. *FAIR baseline technical report.* ISBN 3-9811298-0-6 and ISBN 978-3-9811298-0-9. Darmstadt, 2006.

http://www.avagotech.com/pages/minipod_micropod. http://www.avagotech.com/pages/minipod_micropod. [Online; accessed 12-May-2014].

Hutter, D., J. de Cuveland, and Lindenstruth V. *CBM Readout and Online Processing Overview and Recent Developments.* https://www-alt.gsi.de/documents/DOC-2013-Apr-27-1.pdf. [Online; accessed 17-June-2014]. Mar. 2013.

"IEEE Standard for SystemVerilog–Unified Hardware Design, Specification, and Verification Language." In: *IEEE Std 1800-2017 (Revision of IEEE Std 1800-2012)* (Feb. 2018), pp. 1–1315. ISSN: null.

Kasinski, K., R. Szczygiel, and W. Zabolotny. "Back-end and interface implementation of the STS-XYTER2 prototype ASIC for the CBM experiment." In: *Journal of Instrumentation* 11.11 (Nov. 2016), pp. C11018–C11018.

Kasinski, K., R. Szczygiel, W. Zabolotny, et al. "A protocol for hit and control synchronous transfer for the front-end electronics at the CBM experiment." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 835 (2016), pp. 66–73. ISSN: 0168-9002.

Lemke, F. and U. Bruening. "A hierarchical synchronized data acquisition network for CBM." In: *IEEE Transactions on Nuclear Science* 60.5 (2013), pp. 3654–3660. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84885950108&partnerID=40&md5=66592497478ce1bcd852e89eadcdf704.

Lemke, F., D. Slogsnat, et al. "A unified DAQ interconnection network with precise time synchronization." In: *IEEE Transactions on Nuclear Science* 57.2 PART 1 (2010), pp. 412–418. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-77951166866&partnerID=40&md5=d8951f1ad61222ad6b2d899370707d8a.

LogiCORE IP Aurora 64B/66B v9.2 Product Guide. http://www.xilinx.com/support/documentation/ip_documentation/aurora_64b66b/v9_2/pg074-aurora-64b66b.pdf. [Online; accessed 15-July-2014]. June 2014.

ModelSim User's Manual. Version Software Version 10.5c. Mentor Graphics Corporation. 808 pp. 1991-2016.

Rost, Adrian et al. "Performance of the CVD Diamond Based Beam Quality Monitoring System in the HADES Experiment at GSI*." In: *Proceedings, 10th International Particle Accelerator Conference (IPAC2019): Melbourne, Australia, May 19-24, 2019*. 2019, WEPGW019.

Schmah, A. et al. "Highlights of the beam energy scan from STAR." In: *Central Eur. J. Phys.* 10 (2012), pp. 1238–1241. DOI: `10.2478/s11534-012-0149-1`. arXiv: `1202.2389 [nucl-ex]`.

Senger, P., V. Friese, et al. *Nuclear matter physics at SIS-100*. CBM Report 2012-01. 2011.

Singh, R et al. "Slow Extraction Spill Characterization From Micro to Milli-Second Scale." In: *Journal of Physics: Conference Series* 1067 (Sept. 2018), p. 072002.

Taylor, E.G. "TTC Distribution for LHC Detectors." In: *IEEE Transactions on Nuclear Science* 45.3 PART 1 (1998), pp. 821–828. URL: `http://www.scopus.com/inward/record.url?eid=2-s2.0-0032098094&partnerID=40&md5=25c6985995d50fe559a660c8749f1dc4`.

The White Rabbit Project. `http://www.ohwr.org/attachments/2528/IBIC2013_WR.pdf`. [Online; accessed 2013]. 2013.