FUTUREDAQ FOR CBM: ON-LINE EVENT SELECTION

H.G. Essel
CBM collaboration

Abstract

At the upcoming new Facility for Antiproton and Ion Research FAIR at GSI the Compressed Baryonic Matter experiment CBM requires a new architecture of front-end electronics, data acquisition, and event processing. The detector systems of CBM are a Silicon Tracker System, RICH detectors, a TRD, RPCs, and an electromagnetic calorimeter. The envisioned interaction rate of 10 MHz produces a data rate of up to 1 TByte/s. Because of the complexity and variability of trigger decisions no common trigger will be applied. Instead, the front-end electronics of all detectors will be self-triggered and marked by time stamps. The full data rate must be switched through a high speed network fabric into a computational network with configurable processing resources for event building and filtering. The decision for selecting candidate events requires tracking, primary vertex reconstruction, and secondary vertex finding in the STS at the full interaction rate. The essential performance factor is now computational throughput rather than decision latency, which results in a much better utilization of the processing resources especially in the case of heavy ion collisions with strongly varying multiplicities. The development of key components is supported by the FutureDAQ project of the European Union (FP6 I3HP JRA1).

THE NEW FACILITIES

The GSI future project FAIR will provide unprecedented accelerator facilities to investigate physics cases in the fields of nuclear structure physics and nuclear astrophysics, hadron physics, physics of nuclear matter, plasma physics, atomic physics, and applied physics. The FAIR accelerators will provide heavy ion beams up to Uranium at beam energies ranging from 2 – 45 AGeV (for Z/A=0.5) and up to 35 AGeV for Z/A=0.4. The maximum proton beam energy is 90 GeV.

THE CBM EXPERIMENT

The Physics Case

The nucleus-nucleus collisions research program of CBM [1] will focus on the search for

- in-medium modifications of hadrons in super-dense matter as signal for the onset of chiral symmetry restoration,
- a deconfined phase at high baryon densities, and
- the critical endpoint of the deconfinement phase transition.

Many of the signatures pursued with the CBM experiment are based on rare processes. To achieve an adequate sensitivity, the detector systems are designed to operate at interaction rates of up to 10 MHz for A-A collisions and up to several 100 MHz for p-p and p-A collisions.

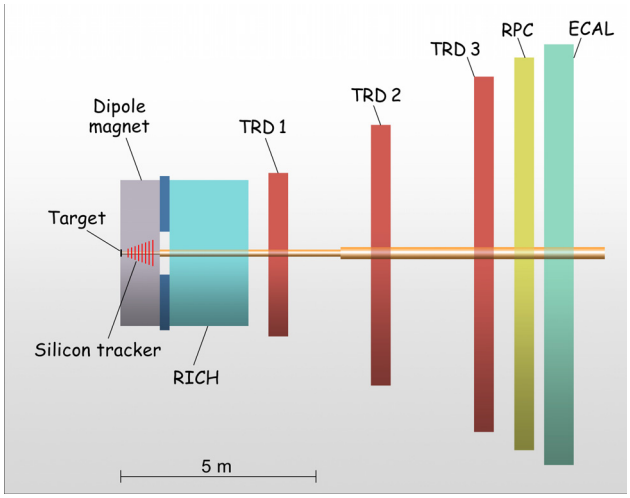

Figure 1: CBM detector setup.

Fig. 1 shows the detector schemes. The setup consists of a superconducting dipole magnet with a Silicon Tracker System inside, a Rich Imaging Cherenkov detector (RICH) for electron identification, three Transition Radiation Detectors (TRD), a Time-of-Flight (TOF) wall built as Resistive Plate Chamber (RPC), and an electromagnetic calorimeter.

It is the task of the data acquisition and event selection system to identify the candidate events for the physics signals under study and send them to archival storage. The most challenging aspect is here the measurement of open and hidden charm production in heavy ion collisions down to very low cross sections. The D mesons will be identified via the displaced vertices of their hadronic decays, the decision for selecting candidate events thus requires tracking, primary vertex reconstruction, and secondary vertex finding in the STS. In addition, the system has to be configurable to handle a wide range of physics signals, ranging from D and J/Ψ in A-A collisions over low-mass dileptons in p-A collisions to Y in p-p collisions.

Triggered DAQ

The conventional system design with triggered front-end electronics allows to keep the event information for a

limited time, usually a few µsec, in the front-end electronics while a fast first level trigger decision is determined from a subset of the data. Upon a positive trigger decision, the data acquisition system transports the selected event to higher level trigger processing or archival storage. A system with such a fixed trigger latency constraint is not well matched to the complex algorithms needed for a D trigger, especially in the case of heavy ion interactions, where the multiplicities and thus the numerical effort needed for a decision varies strongly from event to event.

CBM DAQ

The concept adopted for CBM will use self-triggered front-end electronics, where each particle hit is autonomously detected and the measured hit parameters are stored with precise timestamps in large buffer pools. The event building, done by evaluating the time correlation of hits, and the selection of interesting events is then performed by processing resources accessing these buffers via a high speed network fabric. The large size of the buffer pool ensures that the essential performance factor is the total computational throughput rather than decision latency. Since we avoid dedicated trigger data-paths, all detectors can contribute to event selection decisions at all levels, yielding the required flexibility to cope with different operation modes.

In this approach we have no physical trigger signal, which prompts a data acquisition system to read a selected event and transport it to further processing or storage. We thus avoid the term 'trigger' in this paper. The role of the data acquisition system is to transport data from the front-end to processing resources and finally to archival storage. The event selection is done in several layers of processing resources, reminiscent of the trigger level hierarchy in conventional systems.

One consequence of using self-triggered front-end electronics is a much higher data flow coming from the front-ends on the detector. For CBM a data rate of about 1 TByte/sec is expected. However, communication cost is currently improving faster over time than processing cost, an observation sometimes termed Gilder's law, making such a concept not only feasible but also cost effective.

DAQ ARCHITECTURE

The communication and processing needed between the front-end electronics, generating digitized detector information, and the archival storage, where the complete context of selected candidate events is recorded, can be structured and organized in several ways. The solution described here is guided by two principles: processing is done after event building and it is done in a structured processor farm. It is well adapted to the type of processing needed in the CBM experiment and leads to a straightforward and modular architecture.

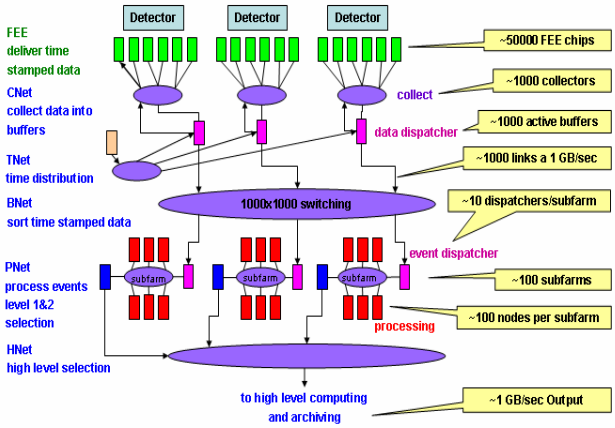

Figure 2: Overall DAQ architecture

A logical data flow diagram is shown in Fig. 2, indicating the data sources and processing elements as boxes and every form of interconnection networks as ovals. The main components are:

Front-end electronics (FEE)

The front-end detects autonomously every particle hit and sends the hit parameters together with a precise timestamp and channel address information over the concentrator network (CNet) to a buffer pool. A rough estimate for the data volume generated by a detector channel can be deduced from typical CBM operation and detector parameters: 10 MHz interaction rate, 10% occupancy for central collisions, a ratio of 1/4 for minimum bias to central multiplicity, and a typical cluster size of 3 fired electronics channels per physical particle hit gives a channel count rate of about 750 kHz. Assuming 8 byte per hit yields a data flow of about 6 MB/sec and channel. For a typical FEE unit with 16 channels this results in a data rate of 100 MB/sec which can be transported over a single GBit serial link.

Clock and time distribution (TNet)

The timestamps of each hit are used to associate hits with events but also in drift and flight time measurements. Thus not only a time scale, in practice a frequency, but also information about the absolute time has to be communicated to all front-end units. The most stringent requirements come from the START and RPC detectors, where the contribution from the clock jitter should be below 25 ps sigma.

The most straightforward approach is to distribute a common clock frequency and to provide a mechanism for broadcasting information with clock cycle precise latency to all units. The minimal required functionality is a global clock reset at the begin of the measurement, or alternatively, distribution of tick marks every second as provided by the planned campus-wide frequency and time normal.

The TNet is thus a dedicated broadcast network, connecting a central controller logically with all front-end

units. The last hop to the front-end units may be implemented with the part of the CNet infrastructure, as indicated by the connection of TNet to CNet in Fig. 2.

Concentrator Network (CNet)

The role of the concentrator network is to collect the data from the individual front-end units and aggregate the traffic on a set of high speed links which connect the detector with the area where the data buffers and the data processing is located. A rough estimate for the total data rate is 1 TB/sec which could be finally transported off the detector with about 1000 links with 10 Gbps each.

In addition other communication tasks like control traffic or time distribution can be handled by the CNet infrastructure.

Active Buffers

The next stage in the data flow is a large buffer pool. The units are indicated as data dispatchers in Fig. 2. They are dubbed 'active buffers' because the data is not only stored but potentially also reformatted and reorganized. They are also hand-over points between CNet, BNet, and PNet. On the output side of the BNet the active buffers are indicated as event dispatchers. Actually, the two dispatchers will be on one board using bi-directional links in/out of BNet.

Build Network (BNet)

For the event selection processing, the data of one event distributed over all data dispatchers must assembled into one event dispatcher as entry into a farm node. This is done by the build network and the active buffers. The FEE sends a stream of time-stamped hits, and it is one of the tasks of the data processing, to first identify at what times interactions occur, and in a second step, to associate the hits with those events.

This *event tagging* processing can be done before or after data traverses the BNet. In the first case, event tagging is handled in the active buffers, and the entities being assembled in the BNet transfers are indeed (buffers of) events. In the second case, only the timestamp information is available, and all the data of a time interval, i.e. defined by epoch markers, is assembled. One possible method for the event definition is to analyze the multiplicity over time of the silicon trackers. The multiplicities of ca. 50 STS concentrator modules must be summed up per time interval (~2ns) in a multiplicity histogram. In this histogram the event times can be recognized.

A reasonable choice for a dispatching unit is an event interval of about 100 events or equivalently a time interval in the order of 10 µsec, which also matches well with the anticipated frame times of a MAPS detector system.

The simplest solution is a time interval based build logic with a traffic shaping and scheduling setup similar to the one developed for the LHCb first level trigger [2]. The more involved solution based on event interval switching may suppress event incoherent backgrounds. The active buffers are responsible for organization of the data flow, in particular for traffic shaping and appropriate scheduling of transfers.

Plausible candidates for the transport are Ethernet, InfiniBand, or ASI.

Figure 3 shows a schematic view of a factorized switch. With n = 32, 1024 bi-directional channels can be implemented by 496 interconnections. The triangles are the active buffers where the data dispatchers (DD) send the data from CNet through Bnet via the event dispatchers (ED) into the PNet. In addition some of the free inner ports can be used by the BNet controller. Some active buffers also function as histogrammers and histogram collectors. The current approach is to use the BNet itself for all necessary communication of scheduling and histogramming (event tagging).

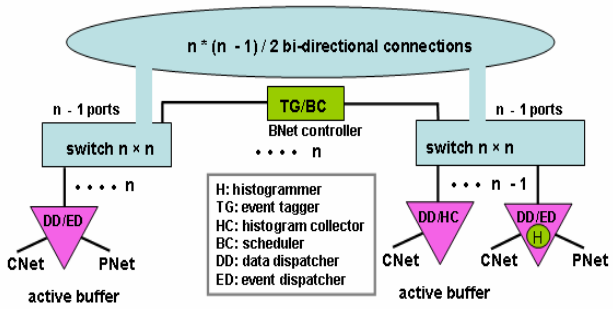

Figure 3: Generic BNet topology.

A SystemC framework has been set up to simulate such a configuration for n=10. Fig. 4 shows the simulated occupation of BNet by the different data types. Over 75% of the nominal bandwidth are used by data tranfers, only a few % by meta data.

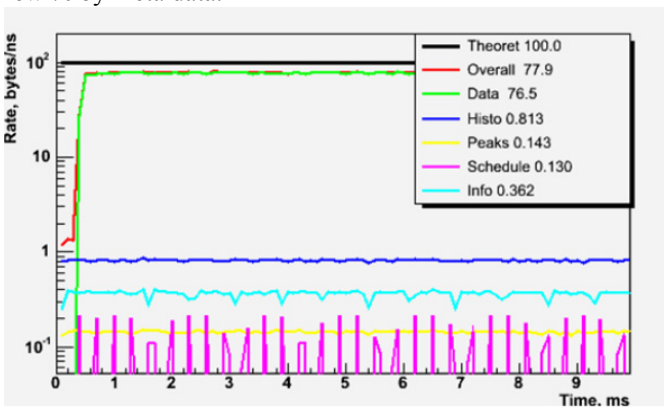

Figure 4: Network utilization

Processing Resources

The first level of event selection processing has to handle the full event rate on a substantial fraction of the total data volume. Processing a data flow on the scale of a

TByte/sec is likely to require a computational bandwidth on the scale of 10^{15} operations/sec.

With todays technology, the most promising approach is a hybrid system using a combination of hardware processors, implemented with programmable logic components like FPGA's, and software processors, implemented with commodity PC's.

New trends already apparent in recent developments like the compute ASIC for IBM's Blue Gene system, the STI cell processor, or the Stretch S5000 series CPUs [3] may provide new solutions and are watched carefully.

The overall architecture is easily adaptable to more integrated forms of configurable computing.

Processing Network (PNet)

The processing resources are grouped in farm nodes. Each farm node is organized around a local processing network which provides the communication between the associated processing resources and active buffers, which act as central data repository and as gateway to the BNet.

The PNet is thus structured into many local networks. The number of hardware and software processors aggregated to one farm node is determined by the amount of resources needed to efficiently handle all the algorithms needed for an event selection decision.

It is expected that the resources of a farm are concentrated in a crate or are at least in close proximity. The PNet can therefore use technologies designed for short distance interconnects, plausible candidates are from todays perspective PCIexpress or ASI.

High-level Network (HNet)

A further reduction will be needed to reduce the data volume to a level suitable for archival storage. This will be handled by a more conventional processing farm. The HNet provides the connection to this high-level computing.

FUTUREDAQ

We acknowledge the support of the European Community-Research Infrastructure Activity under the FP6 "Structuring the European Research Area" programme (HadronPhysics, contract number RII3-CT-2004-506078). Besides GSI, universities of Heidelberg, Mannheim, Munich, Katowice, Krakow, Warsow, Giessen, Budapest, and Turino are participating.

STATUS

A small scale demonstrator as shown in fig.5 for the hierarchy chain will be implemented the next two years. The main goal is to investigate the critical technology, the triggerless data flow, some real detector tests, and finally replacement of existing DAQ systems. This system could scale up to 10k channels and 160 CPUs.

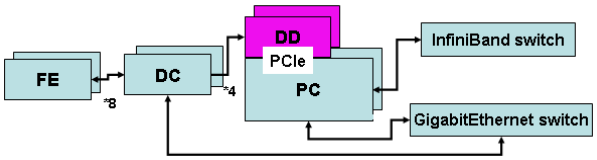

Figure 5: Demonstrator

Up to eight front-end boards (FE) are connected with bi-directional 2.5 Gbps links to a data combiner board (DC). Four of these are connected by 2.5 Gbps links to a PCIexpress board (DD). The hosting PCs act as data dispatchers into a bi-directional InfiniBand network and as event filters. The time clock is distributed by the DCs to the FEs. A first prototype of a DC/DD board utilizing a Virtex 4 FPGA has been assembled. It allows for investigating several kinds of links, access of external memory and network by the imbedded PPC and driving the PCIe interface.

A little 10 Gb InfiniBand cluster with four dual opteron PCs has been installed. Mellanox IB Gold distribution v1.8 [4] was installed. Communication libraries such as MPI (Message Passing Interface) [5] and uDAPL (user Direct Access Programming Library) [6] have been tested (msg, RDMA) (900KB/s with 20KB buffers).

For the DAQ framework the xdaq [7] system developed for CMS is installed and under investigation as well as EPICS for controls.

REFERENCES

[1] P.Senger, J.Phys.G: Nucl.Part.Phys.28(2002)1869
[2] D.Atanasov et.al.,*Ascalable 1 MHz Trigger Farm Prototype with Event-Coherent DMA Input*, Proceedings of the 13th Real Time Conference 2003, Valencia Spain
[3] Stretch S5000 Technical Overview, http://www.stretchinc.com/products_overview.php
[4] http://www.mellanox.com/
[5] http://www.mpi-forum.org/
[6] http://www.datcollaborative.org/
[7] http://xdaqwiki.cern.ch/index.php/Main_Page