The LHCb DAQ System

Beat Jost[1] on behalf of the LHCb Collaboration
[1]Cern EP-Division, 1211 Geneva 23, Switzerland

Abstract

The LHCb experiment is the most recently approved of the 4 experiments under construction at CERN's LHC accelerator. It is a special purpose experiment designed to precisely measure the CP violation parameters in the B-B system.

Triggering poses special problems since the interesting events containing B-mesons are immersed in a large background of inelastic p-p reactions. We therefore decided to implement a 4 level triggering scheme.

The LHCb Data Acquisition (DAQ) system will have to cope with an average trigger rate of ~40 kHz, after two levels of hardware triggers, and an average event size of ~150 kB. Thus an event-building network which can sustain an average bandwidth of 6 GB/s is required. A powerful software trigger farm will have to be installed to reduce the rate from the 40 kHz to ~200 Hz of events written to permanent storage.

In this paper we will concentrate on the networking aspects of the LHCb data acquisition and the controls system.

I Introduction

LHCb [1] is an experiment being constructed at CERN's LHC accelerator for the purpose of studying precisely the CP violation parameters in B-meson decays by detecting many final states. The LHCb detector is a forward single-dipole spectrometer, consisting of a microvertex detector, a tracking system, aerogel and gas RICH detectors, electromagnetic and hadron calorimeter system, and a muon detector. The total number of detector channels is approximately 1 million. The layout of the experiment is shown in Figure 1.

The expected b-quark production cross-section of 500 μbarn, at the LHCb working luminosity of $1.5 \cdot 10^{32} cm^{-2} s^{-1}$, leads to a rate of about 75 kHz of B-meson events.

This is embedded in a total inelastic interaction rate of some 15 MHz. Typical branching ratios for the interesting final states of B-meson events lie between 10^{-5} and 10^{-4} leading to a rate of interesting events of ~5 Hz. For rare decay modes the branching ratios are as low as 10^{-9}.

Thus triggering encounters special problems, since the B-meson events of interest are a small fraction of all the events containing B-mesons. Minimum bias events also are a source of severe background.

Table 1 summarizes the most important parameters of the LHCb trigger and data-acquisition system

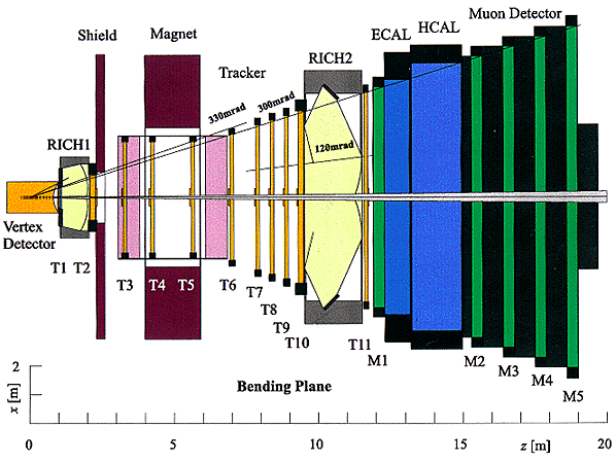

Figure 1 The LHCb detector.

Table 1
Summary of the LHCb Trigger and DAQ Parameters

Number of Channels	~1.1 M
Bunch crossing rate	40 MHz
Level-0 accept rate	1 MHz
Level-1 accept rate	40 kHz
Readout rate	40 kHz
Event size	~150 kB
Event-building bandwidth	~6 GB/s
Level-2 accept rate	~5 kHz
Level-3 accept rate	~200 Hz
Level-2/3 CPU power	~100 kSI95
Data rate to storage	~50 MB/s

The role of the DAQ system is to collect the data, zero-suppressed in the front-end electronics, and assemble complete events in CPUs for further data-reduction by the Level-2 and Level-3 triggers

Figure 2 shows schematically the overall architecture of the LHCb trigger and DAQ system. The main functional components are:

- Timing and Fast Control [2] to distribute a common clock synchronous to the accelerator and the Level-0 and Level-1 decisions to all components needing this information, such as Front-end electronics, Trigger, etc.
- Two levels of 'hardware' triggers: Level-0 and Level-1
- The Front-end electronics where data are buffered during the latencies of the hardware triggers and subsequently processed (zero-suppression, formatting, etc.) being passed to the DAQ system.

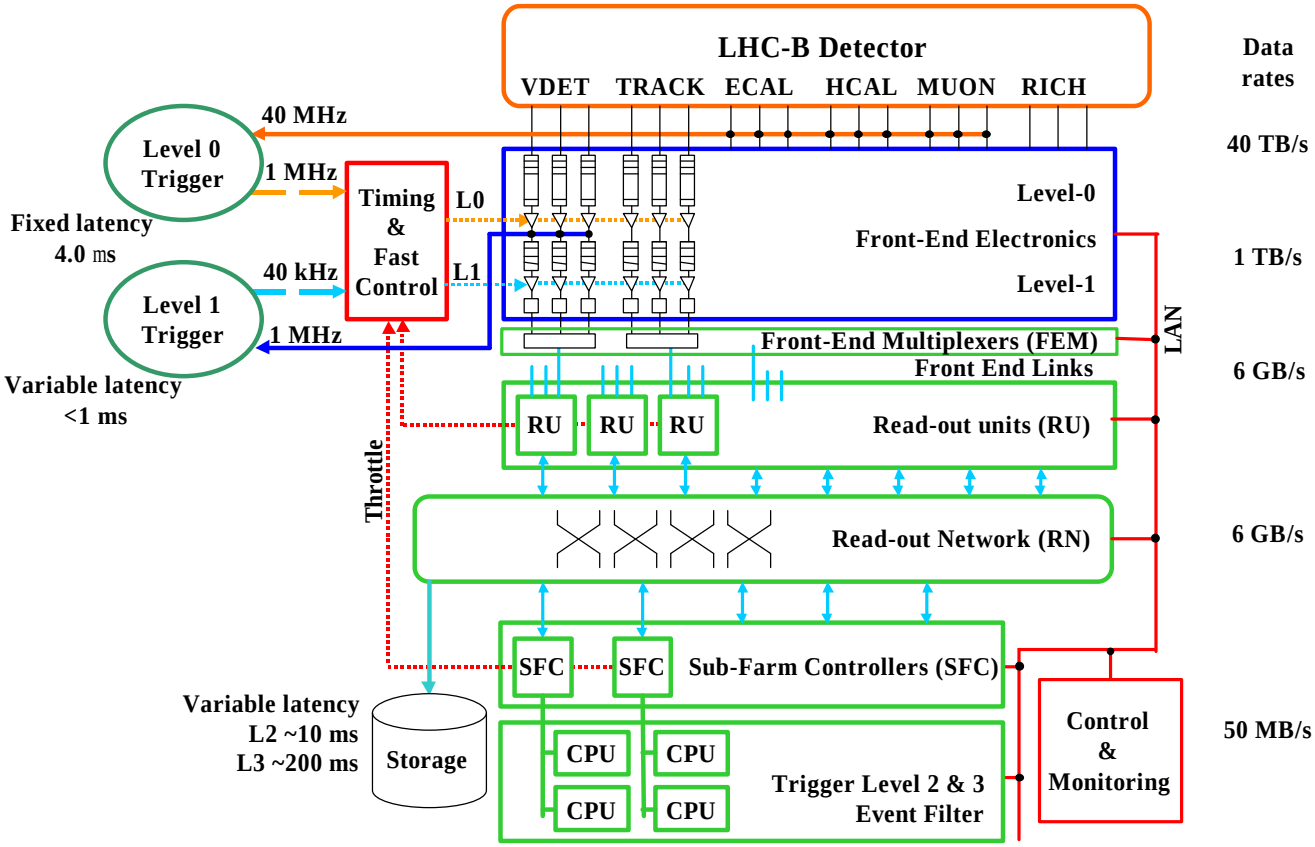

Figure 2 Schematic diagram of the general Trigger and DAQ architecture for the LHCb experiment.

- The DAQ system with as its main components
  - The Readout Units (RU) [3] acting as a multiplexer of Front-end links and as a interface to the Readout Network (RN). The same basic module is used as interface to the Readout Network and also as a Front-End Multiplexer.
  - The Readout Network (RN) which provides support for event-building, i.e. assembling all event fragments buffered in the RUs in one place
  - Sub-Farm Controllers (SFC) which act as an interface between the RN and the processor farm, which will run the higher-level triggers (Level-2 and Level-3)
  - CPU farm to execute the higher level trigger algorithms (Leve-2 and Level-3)
- The whole experiment will be controlled by an integrated experiment control system (ECS) which is in charge of setting the operational states of the detector (traditional slow control) and setting-up and controlling the state of the DAQ system (traditional run control).

In this paper we will focus on the use of networking technologies within the LHCb Trigger and DAQ system, specifically in the following areas:

- Event-Builder (Readout Network)
- Sub-Farms (Event Filter)
- Controls system

II THE LHCB EVENT-BUILDER

A *Requirements and Scale of the System*

The role of the DAQ system is to collect the event fragments originating at the Level-1 electronics, buffer them in the Readout Units and to assemble those belonging to the same bunch crossing in the memory of one of the processors in the CPU farm. The latter process is called event-building. This process should obviously be error-free or at least if errors occur they should be detected and the events flagged as being erroneous. The required performance figures are compiled in Table 2.

Comparing the numbers in Table 2 with those of the large LHC experiments, Atlas and CMS, one can notice that the readout rate is comparable. However the estimated average event size is roughly a factor of 10 smaller. This is also reflected in the expected scale of the system summarized in

Table 3. However the CPU power required in LHCb to execute the high-level triggering algorithms is within a factor of 2 the same.

Table 2
Performance Requirements on the DAQ system

Level-1 Rate	40 kHz
Average Event Size	150 kB
Sustained Bandwidth through Readout Network	6 GB/s
CPU Power in Farm	~10^5 SI95

Table 3
Summary of the approximate scale of the LHCb DAQ system

Number of Front-end Links	~160
Number of Readout Units(RU)	~120
Number of Links in Readout Network	~120
Number of Outputs of Readout Network	~120
Number of Subfarm Controllers	~120
Number of CPUs in Farm (100 SI95/CPU)	~1000

B Readout Protocol

One of the main design criteria of the LHCb DAQ system is simplicity, both in hardware and in the readout protocol. Hence we are favoring a pure push-through protocol, where each source of the RN (in our case the RU) would push its data to a destination of the RN (SFC) as soon as they are available. The algorithm governing the destination selection is based on the event number and is identical in all RUs belonging to the same partition. This scheme has several nice features:

- No central control to communicate with sources and destinations on an event-by-event basis is needed. This in principle leads to perfect scalability.
- The functionality of the RU is very simple in that it only has to multiplex the input links onto an output link[1] using basically a FIFO to isolate the input from the output. In this sense the RU acts as a gateway between the front-end links and the RN
- Simple functionality of the SFC: assemble event fragments arriving from RUs and send complete events to one of the CPUs. Probably some load-balancing algorithm will be implemented in the SFC to level the load among the CPUs connected to one SFC.
- Since all data of one trigger is always available there are no constraints imposed on the Level-2 and Level-3 algorithms.

Obviously there is also a price to pay with this simple protocol, such as

- An elevated sustained bandwidth across the readout network is required (~6 GB/s at nominal rates)
- No direct feedback between sources and destinations of the Readout network. If anywhere in the system a buffer gets too occupied, a general throttle 'signal' is issued to the trigger to disable the flow of events
- An elevated sensitivity to transmission errors or lost fragments in that there is no possibility of retransmission.

The last point must obviously be a subject of extensive simulation on the behavior of the system. In these simulations clearly the network technology enters, but also the detailed behavior of the switches with respect to potential frame or fragment losses.

Partitioning, i.e. concurrent and independent DAQ activities in the system, is supported easily by this protocol, in that the destination assignment algorithm will be configured in such a ways that the RUs belonging to one partition will only send their event fragments to SFCs that are associated to that partition. The partitioning granularity in the event filter farm, i.e. the smallest entity that can be independently used, is thus one sub-farm. Since the number of sub-farms is of the order 100, this granularity should be sufficient.

We have studied alternatives to this protocol ([4] and [5]), namely a phased readout, in which in a first stage only the data needed for the Level-2 algorithm are transferred from the appropriate RUs to the SFCs. Only after a positive Level-2 decision would the rest of the data be transferred. The reduction of the needed bandwidth through the readout network obviously depends on two parameters, namely on the fraction of the data needed for the Level-2 algorithm and the fraction of the Level-2 "Yes" decisions. In our studies we assumed a rate reduction in Level-2 of a factor of 8. This would be achieved by reading ~60% of the data [6]. With these figures one still needs roughly 65% of the bandwidth required for the full readout protocol. Hence the gain is marginal.

We believe therefore that the simplicity in the protocol and the hardware and the additional flexibility for the trigger-software outweighs the disadvantages mentioned. We are convinced that the network technologies and the trend in industry will allow us to find an affordable solution to our bandwidth problem at the time we have to decide (2002).

Clearly the simple push-through protocol relies on a low rate of erroneous frames and frames lost in the readout network, since there is no possibility to re-transmit a damaged frame. While obviously the estimation of the error rate is very difficult, the rate of frames lost in the readout network will have to be simulated in depth.[2]

C Readout Network Technologies

Several candidate technologies have been investigated to date for implementing the readout network. Using the knowledge gathered by the RD31 collaboration on ATM we have studied the use of this technology for the readout network as part of the feasibility studies for the LHCb technical proposal [5]. The result of these studies was that ATM could indeed be used for this purpose with the technology available at the time (1998). A detailed design of the event-builder showed that the cost of an ATM-based solution was almost prohibitively high. Due to this and since no decision had to be taken immediately we started looking on the market for alternative technologies.

[1] Actually the RU does some event building in the sense that it re-formats the packets it receives on the input links into one larger packet.

[2] We expect the bit error rate to be small, since the network wil consist only of short connections and will be confined to a well-controlled environment.

One technology that looked interesting was Myrinet[3]. We setup a simulation framework to investigate the performance of Myrinet under the traffic patterns related to event building. It turned out that the wormhole routing of Myrinet makes large networks, specially aggregated with several layers of interconnected switches not very scalable, unless between each layer of switches FIFO buffers are introduced [9].

Our current activities are centered on the Gigabit Ethernet technology. We are studying "Smart NICs" that should offload the host processor from the interrupt handling [11] and plan to reuse the simulation framework setup earlier to study Myrinet by implementing Gigabit Ethernet as a transport media. Figure 3 shows a sketch of a possible implementation of the readout network using almost today's technology.[4] The network depicted in Figure 3 would have (on paper) a bandwidth of 15 GB/s, which would be largely enough to satisfy the LHCb requirements. Of course this would be one of the configurations that need to be simulated.

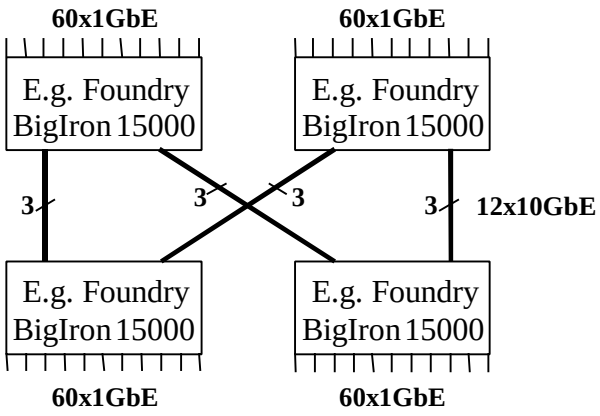

Figure 3 Possible implementation of the LHCb readout network using "today's" technology, in this case Gigabit Ethernet

III THE LHCb EVENT FILTER FARM

The main purpose of the event filter farm is to provide the CPU power and the general infrastructure for running the high-level trigger algorithms. The number of CPUs expected to constitute the farm is expected to be of the order of 1000. The event filter farm will be organized as a number of (identical) sub-farms. The number of sub-farms will be equal to the number of destinations (SFCs) of the readout network.

In Figure 4 the general architecture of a sub-farm of the LHCb event filter farm is shown.

While the technology of the readout network is not yet determined it is quite clear, at least from today's perspective, that the network technology within the sub-farm will be of Ethernet based. The reasons for this are primarily cost. The prices for the data switches and the network interfaces of the CPUs will surely be lowest for Ethernet compared to any other technology[5].

Figure 4 also illustrates the implementation of a general design principle in the LHCb trigger and DAQ system, namely the rigorous and consequent separation of the controls path from the data path. We believe this to be very important to ensure proper and reliable operation of the experiment since any "blockage" of the data path would immediately have also a blockage of the controls path as a consequence. Hence any kind of investigation and debugging of the problem would be rendered difficult if not impossible.

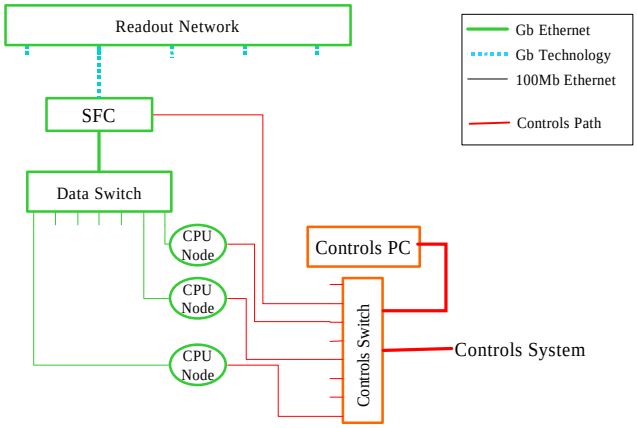

Figure 4 Architecture of a sub-farm of the LHCb event filter farm

The exact number of sub-farms that will be controlled by one Controls-PC will be determined by the amount of data that has to be downloaded to the CPU nodes of the sub-farm and the requirements on the time this download is allowed to take. Most likely there will be one or two sub-farms per Controls-PC.

IV CONTROLS NETWORK

As can be seen from the previous chapter and also from Figure 2 there will be a very substantial controls network installed in LHCb. For the controls of the filter farm alone more than 1000 links are envisaged and more than 1000 additional links will be needed for controlling all the front-end electronics chips and boards.

Figure 5 shows the general architecture of the controls network. The network technology used for the controls network will most likely be Ethernet with all its varieties of speed, downto the level of the individual electronics boards[6].

---

[3] Myrinet is a 1.28 Gb/s parallel technology with an Xon/Xoff protocol for flow control. Myrinet switches are ideal non-blocking crossbar switches with wormhole routing. Paths through the network are defined at the source (source routing). More information can be found in [8]

[4] For the specifications of Foundry's BigIron 15000 switch see e.g. [10]

[5] The price for a NIC in the CPUs of the sub-farm plays a certain role since there are ten times more CPUs in the sub-farm than SFCs.

[6] We currently favor PCs with a form-factor of a credit card and an Ethernet interface plugged on each electronics board as the interface to the Experiment Controls System (ECS). This would act as a replacement of using e.g. VME as a controls bus for electronics boards.

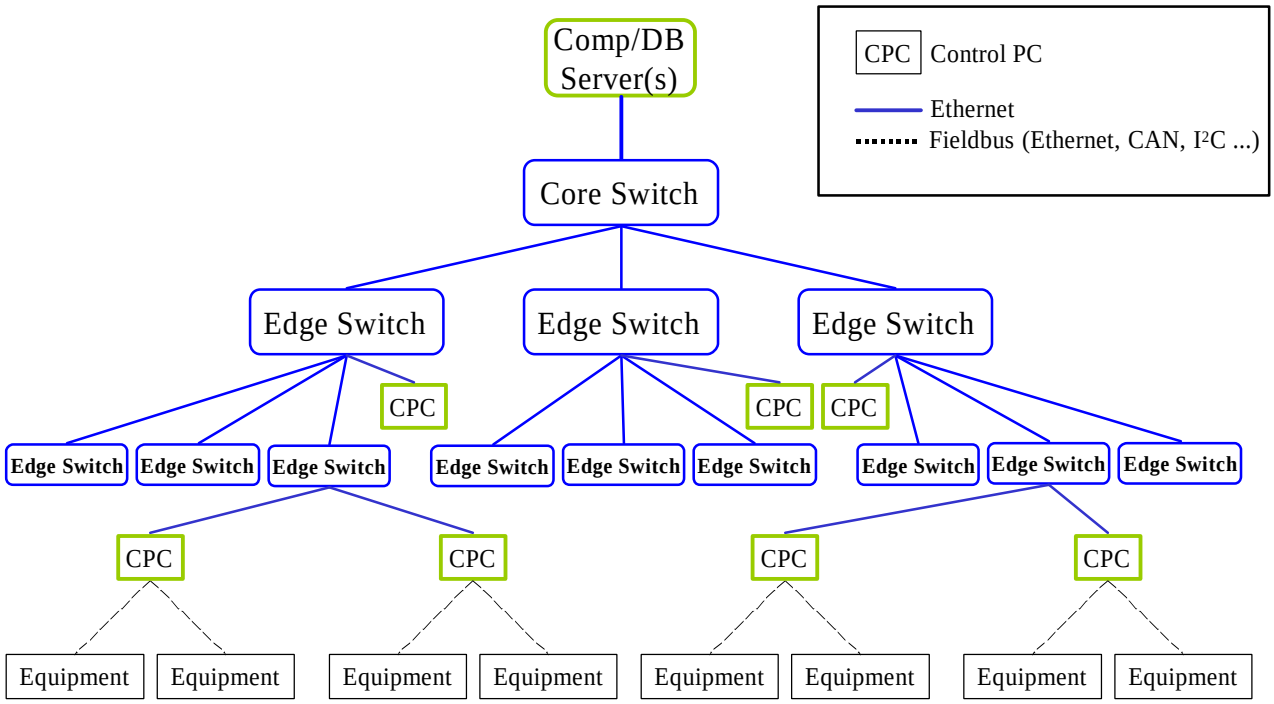

Figure 5 Architecture of the LHCb Controls network.

The topology of the network will be tree-like, which represents the nature of the application, i.e. exercising control over the experiment equipment from an operator console.

The number of Control PCs will be determined by the requirements on the performance and scalability of the controls system. Currently we expect that several GB of data need to be

downloaded into the electronics (Front-End and Readout) of the experiment at the startup of data taking. Also several GB of calibration and configuration data will be needed in the CPUs of the event filter farm. If this operation should take of the order of 100 seconds, a network bandwidth of many tens of MB/s will be needed. This gives an idea of the scale of the controls network

V Conclusion

LHCb will need a quite substantial network infrastructure for the data acquisition and the controls of the experiment. The requirements for the DAQ system (Event-Building) are of the order of 6 GB/s sustained throughput. The readout protocol envisaged is a pure push-through protocol with a throttle mechanism to the trigger distribution system for flow control. We are convinced that the network technology available by the time we have to decide (~2003) will enable us to acquire the necessary equipment from industry. We are currently focussing our activities on the Ethernet family, specifically Gigabit Ethernet, where there already products on the market, which would satisfy our requirements. Detailed simulations of the readout network in terms of performance but also issues like transmission errors and fragment losses still have to be done, in order to validate the approach.

LHCb will implement a controls network which will be completely separated from the DAQ network, for ease of operation and debugging. The scale of this network, in terms of number of nodes, will at least of the same magnitude compared to the Readout Network, while clearly the required bandwidth will be lower. From all that can be said today, the technology for this network will be Ethernet.

VI Acknowledgements

This paper is the result of many discussions with a lot of people within the LHCb computing group, which are hereby gratefully acknowledged.

VII References

[1] LHCb Collaboration, "LHCb Technical Proposal", CERN/LHCC-98-4.

[2] B. Jost, "Timing and Fast Control", LHCb internal Note LHCb 99-001 (unpublished).

[3] H. Mueller et al., Draft document available under http://hmuller.home.cern.ch/hmuller/lhcb_projects.htm

[4] B. Jost et al., "DAQ Architecture and Read-Out Ptotocols", LHCb internal Note LHCb 98-028 (unpublished).

[5] J.-P. Dufey, "DAQ Implementation Studies", LHCb internal Note LHCb 98-029 (unpublished)

[6] M. Frank and F. Harris, "LHCB Dataflow Requirements", LHCb internal Note LHCb 98-027 (unpublished)

[7] Ptolemy Website, http://ptolemy.eecs.berkeley.edu/

[8] Myricom Website, http://www.myri.com/

[9] J.-P. Dufey et al, "The LHCb trigger and data acquisition system", IEEE Trans. Nucl. Sci. : 47 (2000) no.2, pt.1, pp.86-90

[10] Foundry Networks Website, www.foundrynetworks.com/

[11] J.-P. Dufey et al, "Event Building in an Intelligetnt Network Interface Card for the LHCb Readout Network", these NSS-MIC proceedings.