The LHCb Trigger and Data Acquisition System

J.-P. Dufey[1], M. Frank[1], F. Harris[2], J. Harvey[1], B. Jost[1], P.Mato[1], H. Mueller[1]
[1] CERN, 1211 Geneva, Switzerland
[2] Physics Department, Oxford University, 1, Keble Road, Oxford OX1 3NP, U.K.

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-$\bar{B}$ 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 ~100 kB. Thus an event-building network which can sustain an average bandwidth of 4 GB/s is required. A powerful software trigger farm will have to be installed to reduce the rate from the 40 kHz to ~100 Hz of events written to permanent storage

In this paper we will outline the general architecture of the Trigger and DAQ system and the readout protocols we plan to implement. First results of simulations of the behavior of the event-building network implementations under the expected traffic patterns will be presented.

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 calorimeters, and a muon detector. The layout of the experiment is shown in Figure 1.

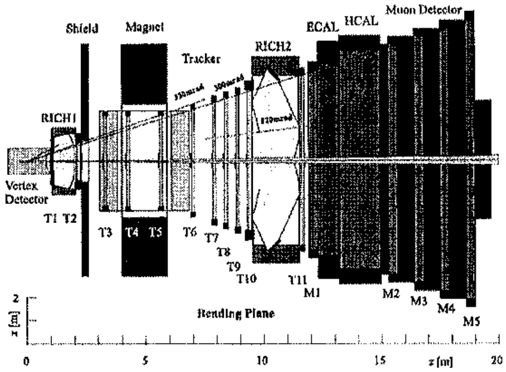

Figure 1 The LHCb detector.

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 offer a severe background.

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.

II. THE LHCb TRIGGER AND DAQ SYSTEM

A. General Architecture

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.) and multiplexed before being passed to the DAQ system.
- 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 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 which is in charge of setting the operational states of the

0018-9499/00$10.00 © 2000 IEEE

detector (traditional slow control) and setting-up and controlling the state of the DAQ system (traditional run control).

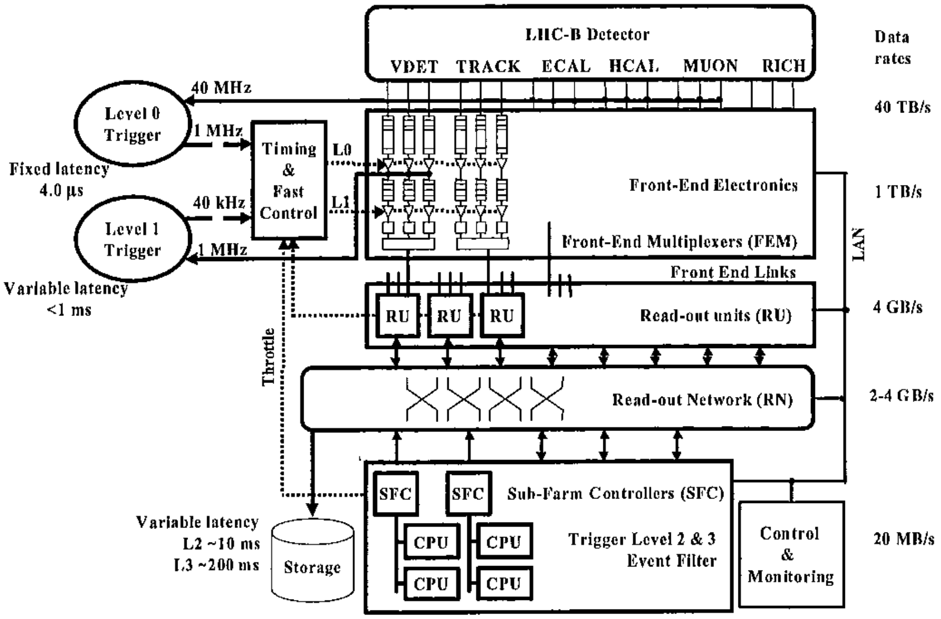

Figure 2 Schematic diagram of the general Trigger and DAQ architecture for the LHCb experiment.

In the following sections the low-level triggers (Level-0 and Level-1) and components of the LHCb DAQ system are described in more detail.

B. The LHCb Trigger System

The LHCb Trigger system is responsible for selecting reliably and efficiently B-meson events out of all the p-p inelastic interactions. Given the high bunch crossing rate of 40 MHz, and the difficulty to distinguish events containing B-mesons from background from inelastic p-p collisions, we have adopted a 4-level triggering scheme[1] (Table 1).

Table 1
Characteristics of LHCb Trigger system

Level	Input Rate	Output Rate	Latency
0	40 MHz	1 MHz	Fixed 4.0 μs
1	1 MHz	40 kHz	Var <1 ms
2	40 kHz	5 kHz	Var <10 ms
3	5 kHz	~100 Hz	Var <200 ms

The first 2 levels (Level-0 and Level-1) are acting only on data from specific detectors whereas the subsequent levels are pure software triggers deciding on the basis of all data from the detector at their full granularity after event building.

1) Level-0 Trigger

Level-0 is primarily based on calorimeter information plus data from the muon identification system, and data from special silicon detectors to reject multiple interactions per bunch crossing. It is designed to select preferentially events with large transverse electromagnetic or hadronic energy, or events which have a muon carrying large transverse momentum. To reject events with multiple interactions in one bunch crossing, a pile-up veto logic is part of the Level-0 trigger. During the fixed latency of 4.0 μs the data of all channels of the detector are stored in pipelines in the front-end electronics.

2) Level-1 Trigger

The Level-1 trigger is searching for events that have a displaced secondary vertex from decaying long-lived particles, which is the case for B-mesons. The average decay length of B-mesons produced at the LHC energies is of the order of 7 mm. To perform this decision based on the event topology the Level-1 trigger uses the data from the vertex detector, whose geometry has been specially chosen to support the Level-1 trigger algorithm. The algorithm is quite sophisticated, doing first a 2-dimensional track reconstruction in the r-z-projection and an impact parameter analysis with respect to the primary vertex, followed by a 3-dimensional reconstruction of the tracks that have a large 2-dimensional impact parameter[2]. The algorithm will run on a farm of CPUs connected to the data-sources via a switching network (Figure 3). The expected data rate is

[1] The distinction in different trigger levels is basically done either on the basis of where the detector data is stored during the decision time of the appropriate level (Level-0 and Level-1) or based on an algorithmic criterion (Level-2 and Level-3).

[2] For a detailed discussion of the algorithm and its performance see [4].

3 GB/s and the input rate of the data is 1 MHz. It is obvious that these requirements are very challenging, where the challenge lays primarily in the fact that the sources and the destinations of the switching network have to handle a fragment rate of 1 MHz.

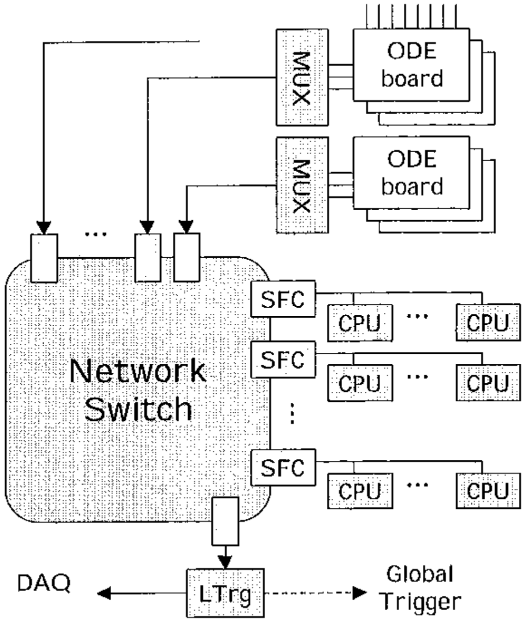

Figure 3 Schematic diagram of the architecture of the Level-1 trigger.

During the latency of the Level-1 trigger, which is expected to be smaller than 1 ms, the detector data that have been transferred out of the front-end electronics will be buffered in the so-called 'Off Detector Electronics' (ODE). Upon a positive Level-1 decision data will be retrieved from the intermediate buffer, and any algorithms necessary to zero-suppress and process the data will be applied before they are forwarded to the DAQ system.

3) Level-2 and Level-3 Triggers

The higher-level trigger algorithms (Level-2 and Level-3) will be applied after events have gone through the DAQ system and will be run on a large processor farm. The strategies for the algorithms are not well defined yet. Current thinking is that Level-2 would harden the Level-1 trigger by taking into account momentum information from the tracking system, and so remove false triggers stemming from multiple scattering in the silicon detector mimicking secondary vertices. This algorithm is expected to reduce the rate by a factor of ~8. After Level-2 we expect that the events contain mostly B-events and charm events. The Level-3 trigger is supposed to distinguish B-decays interesting for CP-violation studies from the total sample. This task will need full final state reconstruction. This aims to reduce the rate to 100-200 Hz.

C. The LHCb DAQ System

1) Requirements and Scale of the System

The role of the DAQ system is to collect the event fragments originating from the ODE and to assemble those belonging to the same bunch crossing in the memory of one of the processors in the CPU farm. 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.

Table 2
Performance Requirements on the DAQ system

Level-1 Rate	40 kHz
Average Event Size	100 kB
Sustained Bandwidth through Readout Network	4 GB/s
CPU Power in Farm	1.4 10^6 MIPS

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

Number of Front-end Links	~160
Number of Readout Units(RU)	~100
Number of Links in Readout Network	~100
Number of Outputs of Readout Network	~100
Number of Subfarm Controllers	~100
Number of CPUs in Farm (1000 MIPS/CPU)	~200 0

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.

2) 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 favouring 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. 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[3] using basically a FIFO to isolate the input from

[3] 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.

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 (4 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

We have studied alternatives to this protocol [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).

3) Simulation Studies of Readout Network

We have built a simulation framework for the event building network for testing different technologies and different readout protocols. For this we use the PTOLEMY discrete event simulation framework [7].

In Figure 4 the model implemented in the simulation software is depicted. Figure 5 shows a blow-up of the composite switching network of Figure 4for the case of a 64x64 network.

The technology currently simulated is Myrinet[4], which is based on non-blocking cross-bar switches, without buffering. The basic problem with any technology of this kind is the question of scalability, i.e. the question whether one can build bigger and bigger switching networks out of small switching elements and what would be the effectively usable bandwidth of the combined switching fabric. One can easily convince oneself that if a large switching fabric is built out of small switching elements scalability is destroyed. We have however found a way to restore the scalability by introducing FIFO buffers between each level of switching elements.

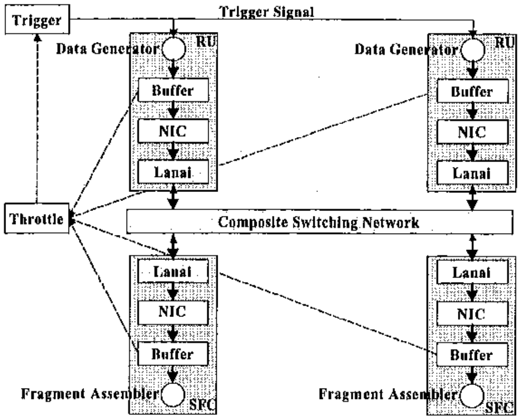

Figure 4 Simulation Model. The shaded areas represent parts of functional components of the LHCb DAQ architecture, whereas the dotted boxes are sub-components specific to the simulated technology, in our example Myrinet.

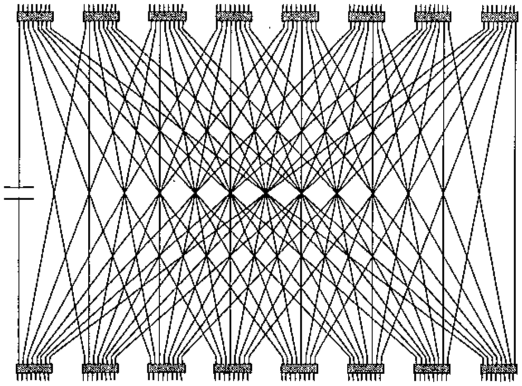

Figure 5 Blow-up of the composite switching network of Figure 4 for the case of a 64x64 switch. The basic switching elements are 8x8 crossbar switches. To simplify the figure the FIFO buffers are only drawn on one connection.

Table 4 shows this effect for different sizes of composite switching fabrics. The drop in efficiency between a single switch configuration (8x8) and a three-level configuration (128x128) is insignificant if FIFOs are introduced, however without FIFOs the loss is almost prohibitive. We believe that the loss of efficiency is due to the fact that a transfer between a source and a destination can be blocked within the composite switching network by a transfer between another source and another destination just because the two transfers happen to compete for some internal path. In this case the introduction of FIFO buffers will de-couple the switching layers and thus each layer will work independently and subsequently will reach the performance of a basic switching element.

[4] 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]

These results show that even with technology existing today, the LHCb readout network could be implemented at reasonable cost. We plan to enlarge the scope of the simulation to other technologies, such as Gigabit Ethernet, and to simulate the complete DAQ system. In this way we will prepare the ground for deciding eventually on a technology to adopt, and also will be able to study the behavior of the system as a whole (virtual prototype).

Table 4 Efficiencies for different sizes of composite switches. All configurations are made out of 8x8 switches. The efficiency is relative to the bisection bandwidth of the switching fabric.

Switch Size	Fifo Size KB	Switching Levels	Efficiency
8x8	NA	1	52.5%
32x32	0	2	37.3%
32x32	256	2	51.8%
64x64	0	2	38.5%
64x64	256	2	51.4%
96x96	0	3	27.6%
96x96	256	3	50.7%
128x128	0	3	27.5%
128x128	256	3	51.5%

III. SUMMARY

We have outlined the architecture of the LHCb trigger and DAQ system and described in some detail the low-level triggers (Level-0 and Level-1) and the main components of the DAQ system. The design of the DAQ system is governed by simplicity, which in turn leads to stronger requirements on the event-building network. However our first simulations show that already today readout networks could be built that satisfy our requirements.

IV. ACKNOWLEDGMENTS

For the simulation part we are indebted to B. Rensch for making available to us his implementation of the Myrinet protocol within the Ptolemy framework. Also the work of M. Tujula getting the simulation of B. Rensch running at CERN is very much appreciated.

V. 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://nicewww.cern.ch/~hmuller/lhcb.htm

[4] H. Dijkstra and T. Ruf, "The L1 vertex trigger algorithm and its performance", LHCb internal Note LHCb 98-006 (unpublished).

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