The New LHCb Trigger and DAQ Strategy: A System Architecture Based on Gigabit-Ethernet

Artur Barczyk, Jean-Pierre Dufey, Clara Gaspar, Philippe Gavillet, Richard Jacobsson, Beat Jost, Niko Neufeld, and Philippe Vannerem

Abstract—The LHCb software trigger has two levels: a high-speed trigger running at 1 MHz with strictly limited latency and a second level running below 40 kHz without latency limitations. The trigger strategy requires full flexibility in the distribution of the installed CPU power to the two software trigger levels because of the unknown background levels and event topology distribution at the time the LHC accelerator will start its operation. This requirement suggests using a common CPU farm for both trigger levels fed by a common data acquisition (DAQ) infrastructure. The limited latency budget of the first level of software trigger has an impact on the organization of the CPU farm performing the trigger function for optimal usage of the installed CPU power.

We will present the architecture and the design of the hardware infrastructure for the entire LHCb software triggering system based on Ethernet as link technology that fulfills these requirements. The performance of the event-building of the combined traffic of both software trigger levels, as well as the expected scale of the system will be presented.

Index Terms—Data acquisition (DAQ), networking, trigger.

I. INTRODUCTION

THE LHCb experiment planned at the Large Hadron Collider (LHC) at CERN, uses a three level trigger system to reduce the primary interaction rate of ~ 10 MHz to a chosen output rate of 200 Hz. A hardware trigger, Level-0, based on calorimetry and muon detector information reduces the rate of accepted events to 1 MHz.

The Level-1 trigger algorithm is designed to operate on a general purpose CPU. The input data are coming from the front-end electronics (FEE) of the detectors included in this system which are the Vertex Detector and a dedicated Trigger Tracker, both semiconductor detectors. The Level-1 algorithm attempts to reconstruct secondary decay vertices, to enhance in the selected sample the population of beauty events with respect to minimum-bias events.

The task of the Level-1 trigger system consists in collecting the data-fragments from over 100 sources and assembling them into a single event in a CPU on a large computer farm. The event data are buffered in the FEE boards until a decision has been taken. The time for taking a decision (maximum latency) is, therefore, determined by the limited size of the buffers in the FEE.

The algorithms will be tuned such that the accept rate of the Level-1 trigger and hence the input rate of the final trigger stage, the Higher Level Trigger (HLT), is 40 kHz. For the HLT the full detector is read out, the maximum time available is only limited by the amount of installed CPUs. The algorithms, which are more complex and slower than their Level-1 counter-parts, will be tuned to a final rate of events to permanent storage of 200 Hz.

The technological challenge in the system consists in handling the high rate of data using commercial and (to a large extent) commodity equipment, while transporting and assembling the data as quickly as possible.

From a data acquisition (DAQ) point of view Level-1 and the HLT are quite similar. The HLT has ~ 350 data sources sending an aggregated event of 38.2 kB, while Level-1 has ~ 130 sending an aggregated event of 4.8 kB. However the *aggregated* data traffic is significantly lower for the HLT. In the baseline scenario there are 5.5 GB/s for Level-1 and 1.6 GB/s for HLT. The foreseeable upgrade of the Level-1 system by including more sub-detectors will increase the Level-1 to 11.2 GB/s, while leaving the the HLT rate unchanged.

A system for doing the data acquisition of the HLT alone has been described in [1]. The system described here is an evolution of the architecture described there, which does the data acquisition and event assembly and trigger processing for both trigger levels using the same infrastructure. The key characteristics of this system are as follows.

- Gigabit Ethernet as a link technology. The connectivity between the sources and the destinations is provided by a large switching network.
- Rate reduction achieved by packing of several events into one Ethernet frame.
- Data being pushed through, every source sending when it is ready to do so, flow control being implemented centrally by disabling the (preceding) trigger via the timing and fast control (TFC) [2] system.
- HLT and Level-1 data share the infrastructure and the HLT and Level-1 algorithms run concurrently on the CPUs.

II. ARCHITECTURE

The architecture is most easily explained by following the data flow from the sources, the FEE boards, to the ultimate destinations, the CPU nodes, as shown in the top, respectively, bottom of Fig. 1.

Manuscript received June 5, 2003; revised January 16, 2004.

A. Barczyk, J.-P. Dufey, C. Gaspar, P. Gavillet, R. Jacobsson, B. Jost, and P. Vannerem are with CERN, CH-1211 Geneva, Switzerland.

N. Neufeld is with CERN, CH-1211 Geneva, Switzerland, and also with the Laboratoire de Physique des Hautes Energies, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (e-mail: niko.neufeld@cern.ch).

Digital Object Identifier 10.1109/TNS.2004.828600

Fig. 1. The architecture of the Level-1 trigger system. Not shown here for clarity are the ECS and its separate network.

A. FEE

The FEE are required to be able to store 58 254 events [3], resulting in a absolute maximum latency for the Level-1 decision to arrive within ~ 58 ms. The decisions are distributed via the TFC system, described in [1], [2]. All front-end boards use the same standard Gigabit Ethernet plug-in card to send the data [4]. This card has dedicated output links, for HLT data and for Level-1 data, if applicable. The card packs the data from the board into the appropriate transport format. Since an Ethernet frame cannot contain more than 1500 bytes of payload [5], the card must be capable of splitting up the data block in several frames if necessary. This can be done either using a custom format defined directly on top of Ethernet or using the standard Internet Protocol (IPv4 [6], [7]). The card also assigns a destination, a subfarm controller (introduced below). The algorithm for assigning a destination is running centrally in the Readout Supervisor, the central element of the TFC system. Upon an accepted event an index into a local address table is transmitted to all FEE boards.

For efficient link usage, but primarily to reduce the packet rate resulting from the high trigger rate of 1 MHz, event fragments are packed into multievent packets (MEPs). The number of events in one packet is an adjustable parameter. The expected data sizes from a full detector simulation indicate that packing factors of 25 for the Level-1 and 10 for the HLT are good working points. Higher packing factors are possible,

however packet reduction can only be achieved as long as all event fragments fit into a single Ethernet frame (1500 bytes).

B. Network

In order to improve the link occupancy at the entrance to the main event building switch a first stage of multiplexing using comparatively cheap and small Ethernet switches is employed. This is the multiplexing layer in Fig. 1. The 349 links for the HLT data are thus reduced to only 33 links into the readout network (RN)¹. The data are then pushed through a large high-performance Ethernet switch or Layer 3 (IP) router to one of ~ 90 subfarm controllers.

C. CPU Farm and Subfarm Controllers (SFC)

The SFC sit at the downstream end of the event-builder switch. They perform the event-building, where individual event fragments from the multievent packets are assembled in correct order into events. They distribute the event to the compute nodes connected to them via another Gigabit Ethernet switch. There are some 16 CPUs in a subfarm. The SFC exercises dynamic load balancing among the nodes. Events which cannot be distributed are buffered in the SFC. Each CPU is

¹The effective packet rate on these links becomes then rather high: ($O(400)$ kHz). Assuming an event building network with as many inputs as outputs, this high packet rate can be a problem for the subfarm controller. For this reason, some packing of events is foreseen also for the HLT.

Fig. 2. The TFC system distributes the 40-MHz clock and the trigger decisions to the FEE. At 1 MHz, L0-triggers are accepted and part of the detector is read out through the RN (Level-1 trigger). Level-1 decisions are taken in the CPUs (bottom) and sent to the Level-1 decision sorter (shown on the left). The TFC system distributes the Level-1 decisions to the FEE.

allowed to process at most one Level-1 event at any given time. This means that Level-1 events will have to queue in the SFC, when there is no CPU free to process them. Simulation shows that the additional latency due to this effect is of the order of one mean processing time only. The total processing time of an event is checked in the individual CPUs, and strict time-outs are enforced. The SFC checks for timeouts in processing time as well.

Since the mean time for reaching a Level-1 decision is much shorter than the allowed maximum, the nodes will not always be busy. Optimal usage of the total available CPU power is achieved by running the HLT as a background task, which is interrupted whenever a Level-1 event needs to be processed.

D. Trigger Decisions

The trigger decisions are sent back to the SFC. For a Level-1 event the decision contains only a short summary block, which is forwarded to the Level-1 decision sorter described in the next paragraph. In the case of the HLT data, accepted events will be undergo a full reprocessing, using all available detector data, including the ones which were not used for reaching a trigger decision. These “reconstructed” data will be sent together with the raw data back through the event building network to permanent storage. The anticipated rate of events to storage is 200 Hz. One gigabit Ethernet port is sufficient to accommodate the small amount of traffic to permanent storage.

E. Trigger Distribution

The synchronous part of the readout and the distribution of the trigger decisions are handled by the TFC system described

in more detail in [2]. Only a few important facts are mentioned here. The system and its connection with the DAQ are schematically shown in Fig. 2.

Level-0 decisions are broadcast to the FEE by the TFC system, after they have been received from the L0-trigger hardware.

The Level-1 decisions, produced in the CPU farm, are contained in small Ethernet packets, which are sent to the *Level-1 decision sorter*. The sorter has been informed in advance, when the event entered the system².

The sorter can thus react to timeouts and force a default decision to avoid buffer overflows in the front end. It also sends out the decisions sorted by event number, which is required by the FEE. Because the decisions must go out in order, the front-end buffers will always be almost full. The arrival time of the decision is determined by the most computing intensive event in the system. Simulation shows that this is not a problem for the system, if some safety margins are implemented. The sorter sends its decisions to the Readout Supervisor, the central component of the TFC system, which ultimately decides, whether an event will be accepted or not.

It should be mentioned here that all the software and hardware components described earlier will be configured, controlled, and monitored using the experiment control system (ECS), described in [1] and [8]. The ECS interfaces to the equipment via a separate network. Separation of data and control paths is strictly observed throughout the system.

²Depending on the technology chosen for the decision sorter this might require a simple converter module to create an Ethernet frame from the TFC signal to be sent to the sorter, this is shown in Fig. 1 as trigger receiver module (TRM). If the sorter functionality was implemented directly in the readout supervisor, then the TRM could be removed from the system.

TABLE I

BASE TARGET PARAMETERS FOR THE L1/HLT IMPLEMENTATION. THE UPPER PART LISTS CONSTANT, UNCHANGEABLE PARAMETERS, THE LOWER PART LISTS CONSTANT BUT CHOSEN PARAMETERS

L0 accept rate	1 MHz
L1 accept rate	40 kHz
L1 transport overhead	56 Bytes
HLT transport overhead	64 Bytes
L1 packing factor	25
HLT packing factor	10
Input link rate	100 MB/s
Output link rate	100 MB/s
Number of CPUs	1400
Frame rate at output	80 kHz
Port-number in aggregation switch	16

III. IMPLEMENTATION

The common Ethernet plug-in card is the only custom component in the system. Its design is currently under way, in many aspects it corresponds to a network interface controller. This ensures that all the various custom electronics boards, provided by the different subdetector teams, use the same interface to the common DAQ system. The card complies with the PMC specifications [4] in terms of dimensions and power consumption, but it uses a custom connector to accommodate the data traffic for up to four Gigabit streams.

All the other components are commercially available. The aggregation and subfarm switches are relatively cheap gigabit Ethernet switches, typically found in high performance LAN installations. Full connectivity at maximum speed is not required, because most of the links are never fully loaded. Since we have a protocol which does not guarantee delivery of frames (it operates on a “best effort” policy), the RN must be dimensioned in such a way as to avoid congestion. Simulation studies show that this can be realized, in our case, with “reasonable” amounts of internal buffering. The event-builder switch must provide this amount of buffering to cope with the traffic pattern. Such devices are typically found in the backbone of large campus networks. Key parameters for this switch can be extracted from simulation. If devices with a sufficient number of ports cannot be found, or are very expensive, then the switch can be build from smaller components. Different interconnection topologies are possible, detailed studies can be found in [9].

The SFC is simply a high-performance PC. The emphasis on performance is mainly on the I/O capabilities, because it is required to handle at least two gigabit/s of data. Such PCs are already available. To achieve maximum throughput, care must be taken in selecting high-performance network interfaces, which support advanced DMA features and buffering. Custom software may need to be developed to tightly couple the event-building with the data movement, which gives maximum performance.

The farm-nodes will be chosen according to the best obtainable price/CPU-performance ratio. They will be operated diskless and require apart from CPU power, a lot of memory and two network interfaces in order to maintain the separation of data and control paths.

TABLE II

KEY PERFORMANCE FIGURES FOR THE LHCb TRIGGER SYSTEM. THE UPPER PART SUMMARIZES THE RATES OBSERVED IN THE RN. THE LOWER PART GIVES EVENT RATES FOR THE SUBFARM

Event Building	
Total Frame Rate at RN input [kHz]	7248
RN Input Links	97
RN Output Links	91
RN Output Link Rate [MB/s]	48.9
Frame Rate (L1) per link [kHz]	61.9
Frame Rate (HLT) per link [kHz]	17.8
Total Frame Rate at RN output [kHz]	79.6
Multi-Event Packet Rate (L1) per Link [kHz]	0.48
Multi-Event Packet Rate (HLT) per Link [kHz]	0.04
Total Multi-Event Packet Rate	0.53
Trigger Farms	
Sub-farms	91
Event Rate/Sub-farm (L1) [kHz]	12.1
Event Rate/Sub-farm (HLT) [kHz]	0.4
Total Event Rate/Sub-farm [kHz]	12.5
Processors/Sub-farm	16
Event Rate per Processor (L1) [kHz]	0.76
Event Rate per Processor (HLT) [kHz]	0.03
Total Event Rate/Processor [kHz]	0.78

The implementation of the Level-1 decision sorter and the TRM is currently under study. It may be possible to combine the functionalities, or to integrate them in the readout supervisor.

The implementation of the remaining infrastructure and hardware is as described in [1].

In order to get the size of the system the number front-end board electronics is one basic input. This determines the number of links to be read out. Assuming a packing factor of 25 (i.e., 25 events are sent in one packet) for the Level-1 and 10 for the HLT, multiplexing factors can be calculated to get to an average link load of less than 80%. Another input parameter to this is the total number of CPU nodes available at the start-up of the experiment. Multiplexing is done with relatively cheap aggregation switches, to save on the cost of the main event builder switch.

The target parameters are summarized in Table I. They either represent available resources like CPU nodes and switch ports, or load factors like the link rates and frame rates. The latter have been chosen, either based on experience or on results from simulation.

The other input into the system design are the expected average data sizes per fragment per front-end board. These numbers are taken from the latest full detector simulation and include all overheads due to data formatting, noise, crosstalk and the like. At the level of the system design, only the overheads due to the data transport are added. A relatively straightforward minimization exercise yields than the final required numbers of switch ports in the event-building switch. Extra ports need to be added to connect the storage-system and the L1-decision sorter. From a system management point of view, it is desirable to have only one large core switch. It might be cheaper however to build the switch from several smaller devices.

Table II summarizes the key parameters of the system. It can be seen that a fairly large gigabit Ethernet switch with approximately 200 ports is needed. These devices exist today, or they can be build from smaller ones. The number of links out of the network and, hence, the number of subfarms is dominated by

limiting the frame rate at the output to 80 kHz. We are confident that using custom software (“zero copy sockets”) and high-end network interface controllers (NICs) in the SFC, it will be possible to accept higher rates and thus reduce the scale of the system.

IV. CONCLUSION

LHCb will build the software trigger levels operating at rates of 1 MHz and 40 kHz, respectively, mostly from commercial components. Taking advantage of the intelligence available in modern FPGAs, the data sources, mounted on the FEE boards of the experiment, can be interfaced directly to a Ethernet switching network. This network is a gigabit Ethernet LAN as is common today nowadays on large campuses. The trigger algorithms will run on a large PC farm, which is connected to the same network. This system is scalable, robust, and affordable. Building it mostly from standard commercial components allows us to leverage technological progress in the networking and PC industry, while at the same time profiting from the very competitive price structure in these markets. Still, a lot of challenging software has to be written for PCs and FPGAs, but these efforts will profit from the vast experience gained with standard PCs, programming-languages, tools, and operating systems.

ACKNOWLEDGMENT

The authors would like to thank S. Cittolin of the PH-Department, CERN, who during a review of the system presented earlier, came up with the idea of reducing the packet rate by packing several events together. This led to a significant simplification compared to the original architecture.

REFERENCES

- [1] “LHCb Online System Technical Design Report,” LHCb Collaboration, CERN, Geneva, Switzerland, 2000.
- [2] Z. Guzik, R. Jacobsson, and B. Jost, “Driving the LHCb front-end readout,” *Trans. Nucl. Sci.*, vol. 53, pp. 508–512, June 2003.
- [3] J. Christiansen, “Requirements to the LI Front-End Electronics,” CERN, Geneva, Switzerland, Tech. Rep. LHCb 2003-078, Sept. 2003.
- [4] H. Muller, F. Bal, and A. Guirao, “GiGabit Ethernet Mezzanines for DAQ and Trigger Links of LHCb,” CERN, Geneva, Switzerland, Tech. Rep. LHCb DAQ 2003-021 Rev.2, Apr. 2003.
- [5] *Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, 2000.
- [6] R. Braden, “Requirements for Internet Hosts—Communication Layers,” RFC 1122, Oct. 1989.
- [7] J. Postel, “Internet Protocol DARPA Internet Program Protocol Specification,” RFC 791, Sept. 1981.
- [8] C. Caspar, B. Franek, R. Jacobsson, B. Jost, S. Morlini, N. Neufeld, and P. Vannerem, “An integrated experiment control system, architecture and benefits: The LHCb approach,” *Trans. Nucl. Sci.*, vol. 53, pp. 513–520, June 2003.
- [9] J.-P. Dufey, B. Jost, and N. Neufeld, “The LHCb event building strategy,” presented at the IEEE-NPSS Real Time Conf. 2001, 2001.