Journal of Physics: Conference Series **119** (2008) 022003 doi:10.1088/1742-6596/119/2/022003

LHCb Online event processing and filtering

F. Alessio[1], C. Barandela[1], L. Brarda[1], M. Frank[1], B. Franek[2], D. Galli[3], C. Gaspar[1], E.v. Herwijnen[1], R. Jacobsson[1], B. Jost[1], S. Köstner[1], G. Moine[1], N. Neufeld[1], P. Somogyi[1], R. Stoica[1], S. Suman[1]

[1]CERN, Geneva, Switzwerland
[2]RAL, Oxford, United Kingdom
[3]Universita di Bologna, Bologna, Italy

E-mail: niko.neufeld@cern.ch

Abstract.
The first level trigger of LHCb accepts one million events per second. After preprocessing in custom FPGA-based boards these events are distributed to a large farm of PC-servers using a high-speed Gigabit Ethernet network. Synchronisation and event management is achieved by the Timing and Trigger system of LHCb. Due to the complex nature of the selection of B-events, which are the main interest of LHCb, a full event-readout is required. Event processing on the servers is parallelised on an event basis. The reduction factor is typically 1/500. The remaining events are forwarded to a formatting layer, where the raw data files are formed and temporarily stored. A small part of the events is also forwarded to a dedicated farm for calibration and monitoring. The files are subsequently shipped to the CERN Tier0 facility for permanent storage and from there to the various Tier1 sites for reconstruction. In parallel files are used by various monitoring and calibration processes running within the LHCb Online system. The entire data-flow is controlled and configured by means of a SCADA system and several databases. After an overview of the LHCb data acquisition and its design principles this paper will emphasize the LHCb event filter system, which is now implemented using the final hardware and will be ready for data-taking for the LHC startup. Control, configuration and security aspects will also be discussed.

1. Introduction

LHCb [1] is an experiment dedicated to the study of the decays of B-hadrons produced at the Large Hadron Collider (LHC). Visible[1] events are produced in the experiment at $\sim$ 10 MHz. The transverse momentum information measured in the calorimeter and muon sub-systems of LHCb is used to reduce the rate of event to 1 MHz. At this rate the entire detector is read out, since the reconstruction of secondary vertices, which is the key discriminant for LHCb physics, requires the full event data. Simulation studies indicate that at the nominal conditions expected at the LHC, the total event-size should $\sim$ 35 kB.

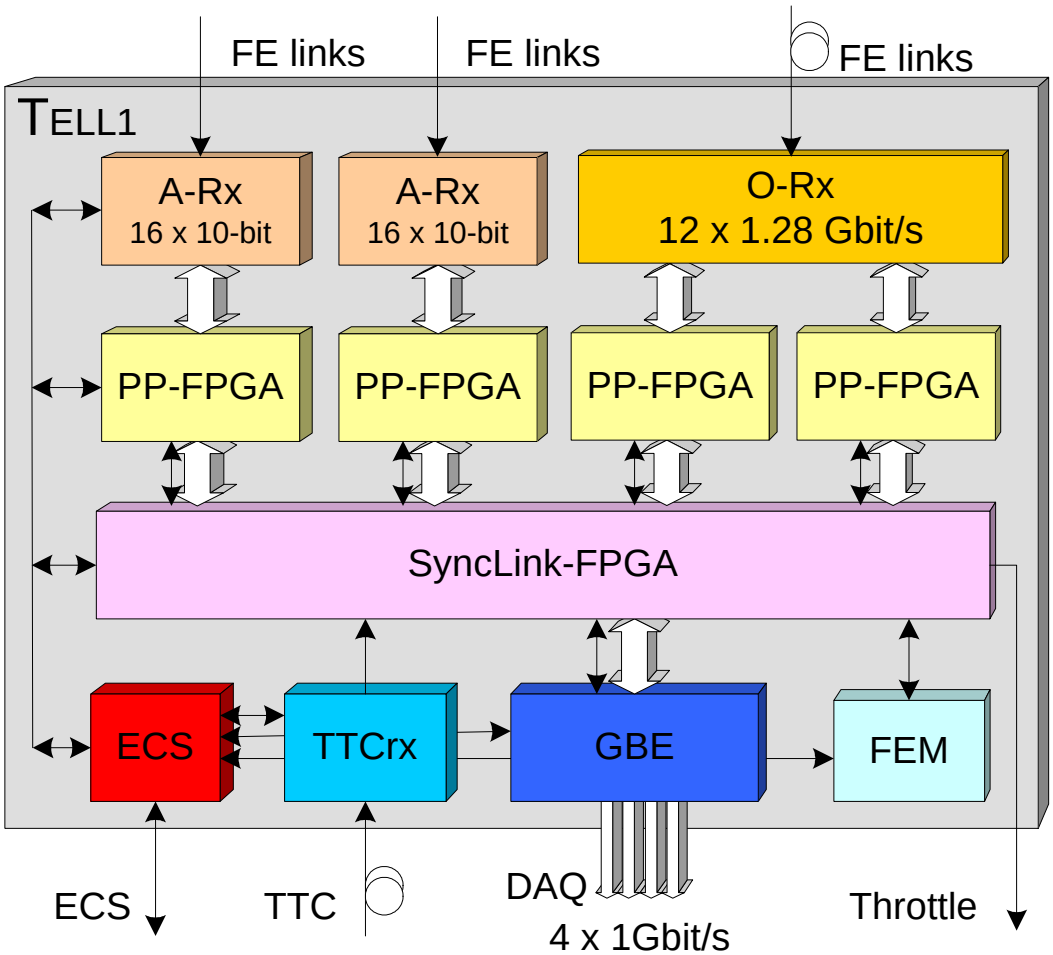
Vertex reconstruction is the first stage of the *High Level Trigger*, which is entirely implemented in software running on a large farm of commodity servers (the *Event Filter Farm, EFF*).

[1] An event is considered visible when it has a minimum energy deposit in the calorimeters and a minimum number of charged tracks.

In the following sections the event-builder, the farm-architecture, the event-processing after selection and finally on the control and configuration of the system will be described.

2. Data acquisition and event-building

The event data are read-out from the detector via approximately 5000 optical and analogue links, each of which has a maximum raw band-width of 160 MB/s. These links are fed into $\sim$ 330 *Readout Boards*, which do analogue and digital pre-processing of the data. Zero-suppression and high-level feature extraction in FPGAs allow reducing the data rate by a factor 4 to 8, depending on the attached sub-detector. There are two functionally equivalent variants of the readout-board called TELL1 [2] and UKL1 [3] respectively. Both are connected to the data acquisition via a plug-in card, which provides 4 1000 BaseT (copper) Gigabit Ethernet ports. A block-diagram of the TELL1 board is shown in Figure 1.



!

Figure 1. Functional block-diagram of the TELL1 readout-board. Both options for the input mezzanine cards are shown for illustration. In reality the board can only be either optical or analogue.

2.1. Event-building

Event-building requires that the fragments belonging to the same original collision, which are spread out over all Readout Boards, are collected in the same place. The CPU requirements of

the sub-sequent High Level trigger are such that a farm of approximately 8000 cores[2] will be needed. The resulting connectivity problem is solved by means of a large switching network. Gigabit Ethernet has been chosen as the network technology, for its excellent price/performance, the huge number of vendors and its firm establishment as a long-lasting industry standard.

Initially all Readout Boards are connected to either of two aggregation switches[3] which are connected to a single large router, through which they send IP packets to one of the servers in the farm. Each board is connected with up to four 1000 BaseT links to the router. Should the bandwidth requirements exceed the initial expectations another core router will be added. The router in turn is sending the data to edge routers, which attach to the individual servers. A few technical details deserve mentioning in this context,

- The network protocol used between Readout Boards and farm-nodes is a light-weight, unreliable datagram protocol on top of IP, very similar to UDP. This was chosen to keep the implementation in the FPGA based Readout Boards simple. LHCb does not need an intermediate layer of computers to do a protocol adaptation.
- The event size in LHCb is quite small, on average every Readout Board will have only ~ 100 Bytes per trigger. In order to reduce the overhead from protocol headers and mitigate the packet rate at the receiving node, several triggers are packed into a *multi-event packet (MEP)*.
- The destination address and the number of triggers to pack into one MEP are assigned by the Timing and Fast Control (TFC) system, which is used for all synchronous information transfer in the experiment.
- The TFC system is also used to implement a simple load-balancing. The farm-nodes announce their availability. The TFC system will disable (*throttle*) the trigger, when there are no destinations left.

Overall the readout architecture is a single-stage, full readout based on push-protocol, with a centralized load-balancing, shown in Figure 2. Excepting the custom build readout-boards [3], [2], the system is entirely built out of commercial of the shelf (COTS) components. The network equipment however is truly high-end, in order to cope with the high amount of traffic and the very unusual[4] traffic patterns.

2.2. Network

The network consists of an aggregation layer at Layer-2 (Ethernet), where two HP5400 [4] routers are used to aggregate underused links. In the commissioning phase and the first year of data-talking, the nominal rate to read out the detector is expected to stay below 20 % of the nominal rate. In order to profit from the latest developments, the full deployment of the core network, capable of absorbing the full 1 MHz, has been deferred. The aggregation layer is used to connect one link from each board, while deliberately overcommiting its up-links into the core network.

The core network consists of a single E1200 router by Force10 networks [5]. This is the highest density Gigabit Ethernet router available[5]. While LHCb uses only Layer 2 switching in the aggregation layer, in the core proper routing is used, i.e. switching based on Layer 3 (IP).

[2] equivalent of a 3 GHz Intel WoodCrest core.

[3] This aggregation layer is only present in the startup phase until end of 2008. Then all boards will be connected directly to the main router. This staging allows buying the very expensive linecards of the router as late as possible.

[4] The traffic is unusual compared to standard random traffic on a usual office/campus LAN, in that it is very bursty in nature, with strong instantaneous overcommitment.

[5] It can accommodate 14 line-cards of 90 Gigabit ports each. Currently these are 1 to 2 overcommited, but this limitation will be removed in the near future.

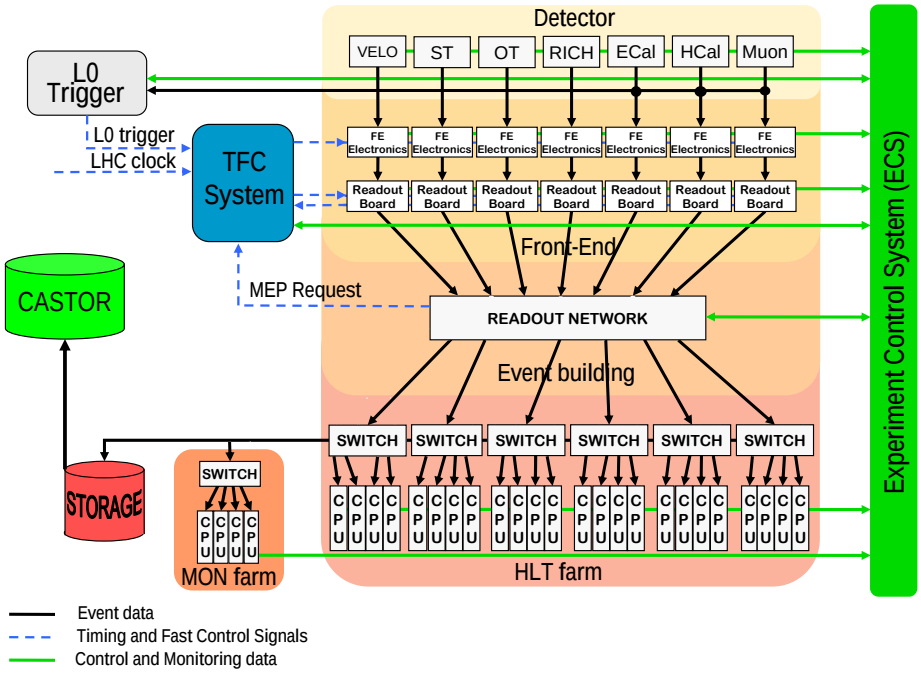


Figure 2. The LHCb readout-system

The CPU farm is physically installed in 50 racks with space for up to 44 servers. In each rack a single HP3500 [4] is connected to the core with up to 12 uplinks of which currently four are used. This edge-switch is also run as a router and connects to the farm-nodes.

2.3. Synchronisation and event management
Event-building requires that all readout boards send the MEP to same destination for the same trigger. The trigger and timing is transmitted by the *Timing and Fast Control (TFC) system* [6], the LHCb implementation of the standard CERN TTC system. In LHCb the second channel ("B-channel") of the TTC system is used to broadcast the IP address of the destination node to all Readout Boards . Conversely the event filter farm nodes themselves announce their availability to the central module of the TFC system, the *Readout Supervisor*. This scheme leads to simple, dynamic load-balancing.

2.4. Throttling
As has been described the LHCb data acquisition uses a push-protocol. There is no back-pressure implemented on the various links. In order to avoid overflows and random truncation of event fragments, a central *throttle* mechanism is used. This the second important task of the TFC system from the DAQ point of view. Any entity in the system that runs low in resources in a way jeopardizing its capability to accept more events can send a throttle signal to the Readout

Supervisor. The Readout Supervisor will then disable the trigger and thus allow the system to settle down. Buffering is nevertheless required at all stages, because it must be possible that all data which are still stored in the pipeline above the component can send their data. The suitable combination of water-marks and buffer size avoids buffer-overflows. This does not bias the trigger, because "problematic" events will always go through at the cost of other *potential* events being rejected at the source.

3. High Level Trigger

Each server in the event filter farm runs one event-builder process, which distributes the assembled events to trigger processes. There will be as many trigger-processes as there are CPU cores. Each server also runs one instance of the data-writer, which sends accepted events to the streaming and formatting layers (below).

Data are not actually moved, but put into a shared memory area. Descriptors to the events are passed between processes by means of a shared buffer library. All processes are implemented using the standard LHCb software frame-work, GAUDI, where some of the services have been specifically re-implemented for the Online environment. These include the logging service, event-reading and writing layers. These are described in detail in [7].

The selection algorithms in the event filter farm reduce the incoming data rate of 1 MHz to an output rate of 2 to 5 kHz, following the luminosity of the accelerator.

4. From raw events to files

Accepted events are sent off the event filter farm servers via TCP/IP. They are sent to a layer of servers, which consolidate the many TCP connections into a few streams. Events are put into streams according to the classification they got by the trigger process and their *partition*, described in more detail in the next section. From the streaming layer events are sent to the formatting layer, where events are sent to multiple destinations using a publish subscribe scheme. Monitoring processes can subscribe to event streams on a best-effort basis. File writer processes are running on the *storage layer* . They make sure that all streams are written into files. A stream can be written into multiple files in parallel, if the rate requires this. The storage layer implements full fail-over and fail-back. The servers are connected to storage area network (SAN), where they create the files.

4.1. Files

Raw data files are normally 2 GB in size. They are registered in a run-database and are transferred to the CERN mass-storage facility, CASTOR, as soon as they are closed. For this transfer the standard LHCb production system, DIRAC is used. DIRAC does the registration in the Grid file-catalogue. A special daemon process copies the data to CASTOR in collaboration with the DIRAC agent processes and after a successful transfer, registers the file in the LHCb book-keeping database. The details of this transfer are described in [8].

5. Configuration, Control and Monitoring

Configuration, control and monitoring of the data acquisition is done by the Experiment Control System [9]. In LHCb run-control and detector control are on equal footing. They use the same tools, namely the PVSS SCADA system and the software framework developed on top of PVSS by the joint controls project (JCOP).

DAQ processes are modelled as "devices". They are integrated into PVSS using the DIM protocol. A dedicated software package has been developed for this purpose: *Farm Management and Control, FMC* [10].

Higher level control is based on a system of finite state machines implemented using the SMI++ engine. In particular this is used to implement *partitioning*.

5.1. Partitioning
Partitioning involves the independent, parallel running of several parts of the detector for debugging purposes. Everything which is *partition-able* must be able to run in multiple instances in parallel. In LHCb the TFC system, the event filter farm and the storage system are all partition-able. Some sub-detectors can also be split into independent partitions.

5.2. Databases
The ECS relies on several databases. The *Configuration database* stores the configuration of all devices and their connectivity. The *Conditions Database* stores all information required for the reconstruction and trigger processes, which is not acquired as raw data. Examples are temperatures and other environmental data, acquired independently from the collider status, calibration constants derived from offline processing and similar constants. Details about the Conditions database can be found in [11]. Information about runs and associated files is stored in the run database, explained in detail in [12]. All databases are running on Oracle10g infrastructure.

5.3. Online environment
The ECS runs on a private, dedicated network. This network is not directly accessible from the Internet or the CERN network and is also strictly separated from the data acquisition LAN. Access to the ECS network is via dedicated, dual-homed gateway machines.

Single-sign on (SSO) is used, meaning that each user has only one unique account that can be used to access all resources. A global shared file-system is available on all nodes for both Windows and Linux operating systems. NFS is used on Linux and Samba is used for Windows clients. Currently user information such as group membership, privileges and id-numbers are currently managed via the Network Information System (NIS). This will be migrated to LDAP, which will allow integrating the authentication at the level of PVSS. On Windows systems we use pGina to defer the authentication to a Linux Kerberos server. All authentication is based on Kerberos v5.

6. Status of installation and deployment
Network equipment and computers are constantly getting cheaper. In order to take advantage of this, the deployment of the core network and the event filter farm is staged in accordance with the expected performance of the LHC. In 2008 a system which can accept approximately 250 kHz will be installed. That means between 20 and 25% of the network and farm.

The base network is almost completely deployed: 300 readout boards with 4 connections each, 1 of them connected to the aggregation layer, the core connected to 10 fan-out switches. In total currently there are almost 1700 Gigabit patch-connections.

The rest will be installed right in time for the 2009 run, when LHCb anticipates full luminosity. The infrastructure for the control system, the databases and the storage is already deployed completely, because it is needed for commissioning and early data-taking.

All software components exist and have been tested. Final integration for global running is ongoing.

Conclusion
LHCb's data acquisition system is based on a simple, robust and scalable architecture. Industry standard, commercial of the shelf components are used wherever possible. All hardware necessary for a full readout at a reduced rate of 100 to 200 kHz has been installed and tested. For 2009, network and event filter farm will be upgraded to the capacity required for a full detector readout at 1 MHz. This will be done by simply adding hardware: line-cards in the core

network and servers in the event filter farm. Synchronous information and load-balancing is provided by the TFC system, which is in full operation. The entire DAQ (including the TFC) is controlled, configured and monitored by one Experiment Control System. This system is based on a commercial SCADA software product, customised by a CERN-wide common framework. The same software framework is also used for controlling and monitoring the detector hardware. This architecture allows leveraging synergy between the different areas and providing a coherent view of and access to *all* components of LHCb. The ECS is completed by a set of databases, all running on an industry strength Oracle10g infrastructure. The infrastructure software in the event filter farm and the rest of the Online system is mostly based on LHCb's standard offline software framework, GAUDI, with some online-specific adaptations. The transfer of data to permanent storage is using the same Grid enabled software framework, DIRAC, which is also driving LHCb's offline production and data replication to the Tier-1 centres. DAQ, ECS and Grid data-transfer have been successfully tested and are ready for the start of data-taking.

Acknowledgments

The implementation of this system would not be possible without the excellent collaboration and dedication of all the sub-detector teams in LHCb. The authors would also like to thank the CERN IT department, in particular the teams of IT/FIO (Castor operations), IT/CS (campus networking) and IT/CO (controls) for their tight collaboration and effective support.

This work has been supported by the Marie Curie Early Stage Research Training program of the European Community's Sixth Framework Program under contract numbers MEST-CT-2004-007307-MITELCO and MEST-CT-2005-020216-ELACCO.

References

[1] LHCb Collaboration 2003 *LHCb Reoptimized Detector: Design and Performance* CERN-LHCC-2003-030 (CERN)
[2] Bay A, Gong A, Gong H, Haefeli G, Neufeld M M and Schneider O 2006 *Nuclear Instruments and Methods* **A560** 494–502
[3] Barham C, Katvars S and Wotton S 2006 RICH L1 Technical manual Edms note University of Cambrdige
[4] Hewlett Packard http://www.hp.com
[5] Force10 Networks http://www.force10networks.com
[6] TFC homepage http://cern.ch/lhcb-online/TFC
[7] Cherukuwada S S, Frank M, Gaspar C, van Herwijnen E, Jost B, Neufeld N, Somogyi P and Stoica R 2007 *Proc. International Conference on Computing in High Energy and Nuclear Physics* (Victoria, BC)
[8] Frank M, Jost B, Neufeld N, Smith A C, Stoica R and Suman S 2007 *Proc. International Conference on Computing in High Energy and Nuclear Physics* (Victoria, BC)
[9] ECS homepage http://cern.ch/lhcb-online/ecs
[10] Galli D *et al.* 2007 *Proc. 15th IEEE NPSS Real Time Conference* (Fermilab, Batavia, Ill)
[11] del Carmen Barandela-Pazos M 2007 *Proc. International Conference on Computing in High Energy and Nuclear Physics*
[12] Frank M, Neufeld N, Smith A C and Stoica R 2007 *Proc. 15th IEEE NPSS Real Time Conference* (Fermilab, Batavia, Il)