The LHCb Readout System and Real-Time Event Management

F. Alessio, C. Barandela, L. Brarda, O. Callot, M. Frank, J.-C. Garnier, D. Galli, C. Gaspar, Z. Guzik, E. van Herwijnen, R. Jacobsson, *Member, IEEE*, B. Jost, A. Mazurov, G. Moine, N. Neufeld, M. Pepe-Altarelli, A. Sambade Varela, R. Schwemmer, P. Somogyi, D. Sonnick, and R. Stoica

***Abstract*—The LHCb Experiment is a hadronic precision experiment at the LHC accelerator aimed at mainly studying b-physics by profiting from the large b-anti-b-production at LHC. The challenge of high trigger efficiency has driven the choice of a readout architecture allowing the main event filtering to be performed by a software trigger with access to all detector information on a processing farm based on commercial multi-core PCs. The readout architecture therefore features only a relatively relaxed hardware trigger with a fixed and short latency accepting events at 1 MHz out of a nominal proton collision rate of 30 MHz, and high bandwidth with event fragment assembly over Gigabit Ethernet. A fast central system performs the entire synchronization, event labelling and control of the readout, as well as event management including destination control, dynamic load balancing of the readout network and the farm, and handling of special events for calibrations and luminosity measurements. The event filter farm processes the events in parallel and reduces the physics event rate to about 2 kHz which are formatted and written to disk before transfer to the offline processing. A spy mechanism allows processing and reconstructing a fraction of the events for online quality checking. In addition a 5 Hz subset of the events are sent as express stream to offline for checking calibrations and software before launching the full offline processing on the main event stream.**

In this paper, we will give an overview of the readout system, and describe the real-time event management and the experience with the system during the commissioning phase with cosmic rays and first LHC beams.

***Index Terms*—LHCb, readout control, readout system, real time event management.**

I. Introduction

THE LHCb experiment [1] is located at CERN and is one of the four experiments at the Large Hadron Collider (Fig. 1). Its goal is the search for new physics through the study of CP-violation in B-decays, and the search for rare decays. As such it is a precision experiment at a hadron collider with the implication that it has to cope with high particle multiplicities, very high background, and small branching ratio of interesting B-meson decays. With less beam focusing than the General Purpose Detectors at LHC to avoid interaction pile-up, the LHCb design luminosity is $2 * 10^{32}$ cm^{-2}s^{-1}. LHCb thus expects a 10 MHz rate of visible interactions out of the 30 MHz of proton collisions with the nominal filling scheme, and 100 kHz of b-anti-b-pair production. Based on these facts, the emphasis of the experiment is a good decay time resolution, and efficient particle identification, a good mass resolution, and in particular an efficient trigger for many B decay topologies.

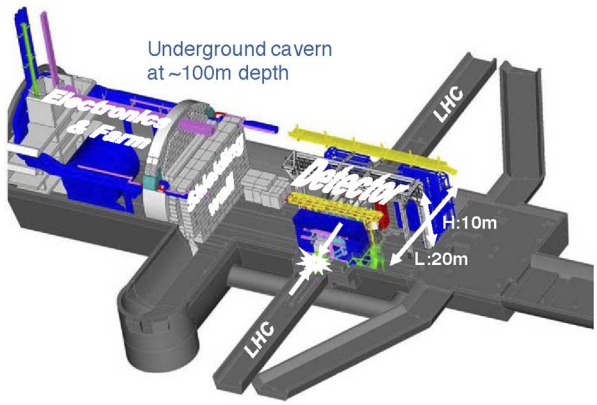

Fig. 1. The LHCb cavern and detector. The backend electronics and online processing farm is protected by a shielding wall from the radiation in the detector area.

In the initial conception [2], the LHCb experiment included three levels of triggers: a hardware low latency high-rate trigger, and two levels of software triggers performed on a CPU farm, the first of which accessed partial detector data and required storing the full data in back-end electronics during the processing. The full data was only sent to the High-Level Trigger farm following the decision of the partial software trigger. This partial software trigger has been eliminated thanks to an increased readout bandwidth in order to greatly simplify the system and allow the High-Level Trigger to have full access to the events accepted by the hardware trigger.

II. Trigger Architecture

The current LHCb trigger architecture features only two levels of triggers (Fig. 2): a high-rate hardware trigger (Level-0 or L0 for short) and a software High- Level Trigger (HLT).

The Level-0 trigger is a low-latency trigger implemented in FPGAs and selects events containing muon, electron, photon or hadron candidates with relatively high transverse momenta. The processing is synchronous with the LHC 40 MHz bunch

Manuscript received May 23, 2009; revised August 14, 2009. Current version published April 14, 2010.

F. Alessio, C. Barandela, L. Brarda, M. Frank, J.-C. Garnier, C. Gaspar, E. van Herwijnen, R. Jacobsson, B. Jost, A. Mazurov, G. Moine, N. Neufeld, M. Pepe-Altarelli, A. Sambade Varela, R. Schwemmer, P. Somogyi, D. Sonnick, and R. Stoica are with CERN, CH-1211 Geneva 23, Switzerland (e-mail: Richard.Jacobsson@cern.ch).

O. Callot is with LAL, Centre d'Orsay, F-91898 Orsay, France.

D. Galli is with the University of Bologna, 40126 Bologna, Italy.

Z. Guzik is with Soltan Institute for Nuclear Studies, PL-00681 Warsaw, Poland.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNS.2010.2042069

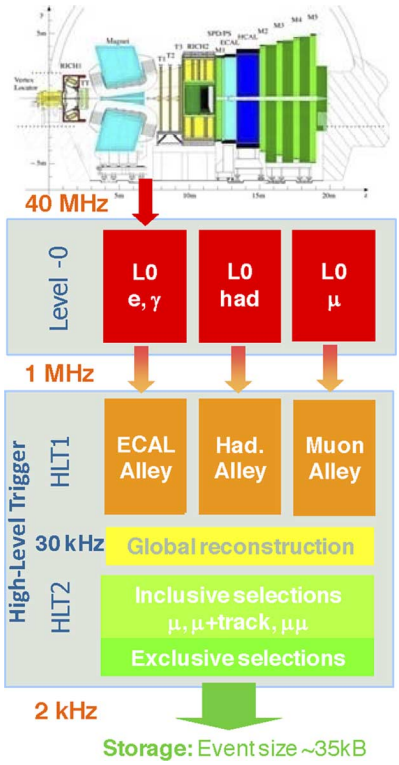

Fig. 2. LHCb Trigger Architecture.

crossing clock and requires 4 μs including data transfer and trigger distribution. The latency implies an intermediate pipelining of 160 events in the detector Front-End electronics. In the nominal LHC filling scheme about $3/4$ of the proton bunch buckets are filled, producing a collision rate of 30 MHz. With the luminosity tuned to avoid pileup and optimize for single proton interactions, the rate of visible collisions in LHCb is 10 MHz. The L0 trigger reduces the rate of 10 MHz visible interactions to a full readout of 1 MHz of events.

A unique feature among the LHC experiments is that LHCb is designed to allow accepting consecutive 25 ns events (max 16). In addition to reduced dead-time ($<0.5\%$ at nominal luminosity), the feature allows easy time alignment of the sub-detectors and optimization of the signal to spill-over ratio.

The High-Level Trigger (HLT) runs on a processing farm with of the order of 1000 quad-core processing nodes. In the first stage the HLT tries to confirm the Level-0 candidate with more complete information. It also applies impact parameter and lifetime cuts, thereby reducing the rate from 1 MHz to about 30 kHz. The second stage of the HLT consists of a global event reconstruction and subsequent selections reducing the rate to an event storage rate of 2 kHz. The processing time available to the HLT is several milliseconds.

A particularity of this setup is that the HLT requires knowledge about the settings of the L0 trigger, which may need optimization during the LHC fill to follow the luminosity. In order to avoid the lengthy procedure of interrupting the data taking and reconfiguring the 1000 nodes, a Trigger Configuration Key, corresponding to a set of preloaded settings, is distributed in the event data from the central readout control system. In this manner the update of the parameters used by the HLT is immediate.

III. Readout Architecture

Fig. 3 shows the LHCb readout architecture. It consists of a set of detector-specific custom-made front-end electronic boards connected to a set of 320 common Readout Boards [3] via approximately 5000 optical digital links and copper analog links with a total data thru-put of the order of 4 Tb/s. The Readout Boards perform zero-suppression and interface the custom electronics to the Readout Network based on Gigabit Ethernet. A router with the highest Gigabit Ethernet port density available assures full connectivity with the processing farm and a bandwidth of 35 GB/s. Each Readout Board holds a fragment of an event. The event building and HLT processing is performed by sending all fragments to the same node in the processing farm. With an average event fragment size of 120 Bytes and an IP/Ethernet overhead of 58 Bytes, a packing of event fragments into Multi-Event Packets (MEP) is done to improve the network utilization. The MEP packing factor is typically around ten. The MEP protocol used between the Readout Boards and the processing farm is a simple datagram protocol on top of IP/Ethernet. The TCP acknowledgement and retransmission features were considered unnecessary and would have complicated significantly the FPGA implementation in the Readout Boards. A packet loss would be detected by the Event Builders and would be equivalent to detector inefficiency for the event to which the lost data fragment belongs. Up to now no packet loss due to the network transmission has been observed under normal running conditions.

The processing farm and the Monitoring Farm, which spies on the stream of accepted events, are based on commodity servers located in one of the counting houses of the LHCb experiments. From the processing farm the data is transmitted via TCP/IP to a storage cluster with a capacity of 50 TB located in the surface building. With an HLT accept rate of 2 kHz and an event size of 35 kB, LHCb has an output rate of 70 MB/s.

The entire readout from the front-end electronics up to the processing farm is driven and controlled by the so called Timing and Fast Control (TFC) system.

IV. Centralized Real-Time Event Management

The Timing and Fast Control system [4] controls the synchronous readout between the Front-End electronics and the Readout Boards, and the asynchronous readout between the Readout Board and the processing farm.

The asynchronous readout control has two aspects of real-time event management. On the one hand it consists of managing the pure data transfer from the Readout Boards, that is, the packing of the events into Multi-Event Packets in the Readout Boards and the assignment of the destination in the processing farm for the next set of events. The destination assignment includes a load-balancing scheme of the readout network and the processing farm, and a rate control. The readout control also allows partitioning, that is, it allows operating any ensembles of sub-detectors autonomously and concurrently for commissioning, calibration and debugging.

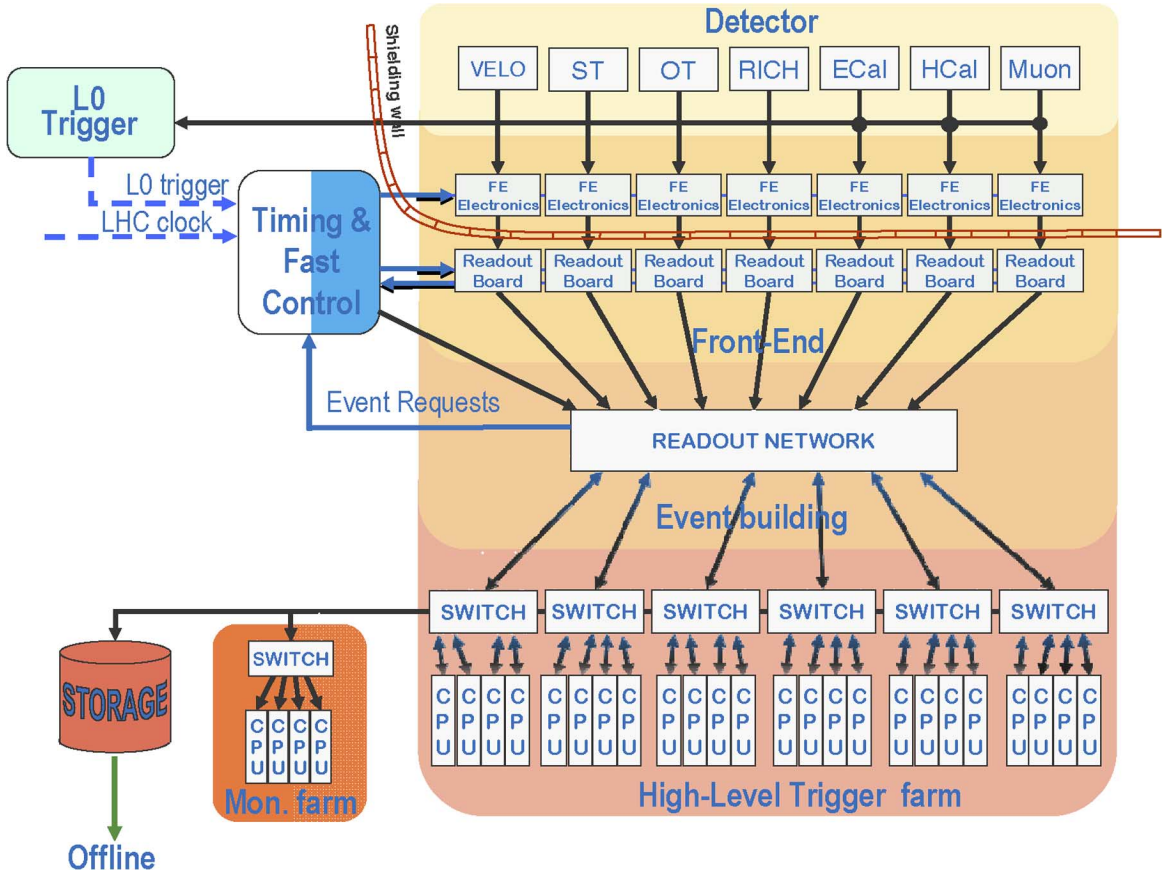

Fig. 3. The LHCb Readout Architecture.

On the other hand, the readout control also consists of managing the different events types and their associated types of farm destination and processing. For instance, it allows differentiating real-time between physics triggered events, calibration events, special events for luminosity measurements, events containing non-zero suppressed data, and special events based on consecutive triggers for timing alignment purposes.

The readout control is implemented in the master of the Timing and Fast Control System, the Readout Supervisor. The Readout Supervisor is entirely based on FPGAs and performs the control by distributing beam-synchronous clocks, the Level-0 triggers, synchronous resets, and fast commands to the readout electronics via a dedicated optical network. It also receives back-pressure from the Readout Boards via a dedicated optical network.

As shown in Fig. 4, the Readout Supervisor is interfaced with most of the systems to perform the control. In particular, it is connected to the Readout Network in order to transmit an event data bank which is appended to each event during the event building, and in order to receive so called Multi-Event Requests from the processing farm nodes.

The Readout Supervisor appears to represent a single point of failure in the system. However, apart from being designed with emphasis on reliability, the Timing and Fast Control system includes ten completely equivalent Readout Supervisors, nine of which are normally invoked via the Experiment Control System for concurrent stand-alone runs with the sub-detectors,

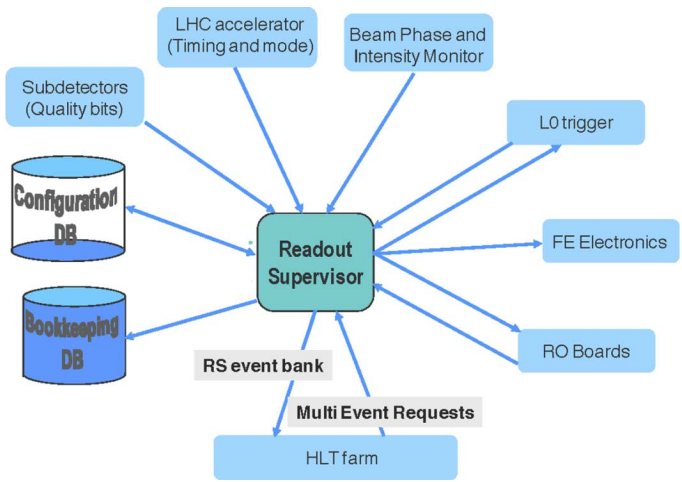

Fig. 4. Logical picture of the TFC information flow. It shows all the systems to which the Readout Supervisor is connected in order to perform the control of the readout and the management of events.

but which may also be immediately invoked for the global data taking without any re-cabling.

A. *Data Transfer Control*

The Readout Supervisor controls the MEP packing by broadcasting a Trigger Type command to the Readout Boards for each trigger, interleaved at a configurable nominal interval by

a Destination command. The Trigger Type command is used by the Readout Boards to start adding each event to the current Multi-Event Packet. The command also contains an event identifier which allows checking the synchronicity of the system and a trigger type which may be used to control the processing in the Readout Boards according to the source of the trigger. The Destination command on the other hand acts as an instruction to close the current MEPs which are being filled in the Readout Boards, and carries the IP address to where the MEPs should be sent in the processing farm. In practice this means that the Readout Supervisor controls the packing factor dynamically and that it may vary the packing factor, for instance, to avoid MEPs with mixed trigger types. It also allows closing MEPs prematurely to flush the system due to for instance End-of-Run or just before a synchronous reset sequence.

The farm destinations are retrieved from currently two different destination tables, one of which is for the HLT farm and the other which is for a special calibration farm. The trigger type determines which destination type should be used.

The destination is chosen based on a credit scheme. The credit associated with each destination function as a counting semaphore. The farm nodes declare themselves as ready to receive at the beginning of a run and at the end of the processing of each MEP by transmitting 'MEP Requests' to the Readout Supervisor. The Readout Supervisor increments the credit of a destination based on the MEP Request and decrements the credit whenever a destination is used for the next MEP. If the destination has a zero or negative credit, the destination is skipped but the credit is still decremented. This means that the LHCb readout system between the Readout Boards and the processing farm effectively works in a push mode with a passive pull mechanism. Thus the scheme allows a static load balancing of the Readout Network according to the organization of the destinations in the tables and a dynamic load-balancing of the processing farm. This also means that event loss is minimized in case of a failing, blocked or overloaded links or nodes. Failing nodes or links may be identified rapidly since the credit of a destination becomes increasingly negative if the node does not transmit a MEP Request and since the Experiment Control System monitors the credit table.

In addition the Readout Supervisor monitors the total credit as a global counting semaphore which is used ultimately to regulate the trigger rate.

B. Event Management

As mentioned above, the Readout Supervisor controls the type of destination to which an event is sent according to the source of the trigger. In addition the Readout Supervisor transmits an event data bank in the same way as any Readout Board which is appended to the event during the event building. The data bank contains information about the identity of an event (Run Number, Event Number, LHC Orbit Identifier and Bunch Identifier, and Universal Time) which is used in the online and offline bookkeeping. It also contains a few detector quality bits from each sub-detector which are derived from the status of the detector hardware (HV, LV, positioning etc) and which may be used by the farm processing to judge the usefulness of an event. In addition it contains a number of words indicating the type of event (trigger source, calibration type, bunch crossing type, and the number of consecutive triggers taken for timing alignment purposes) which determines the processing in the online farm. The Trigger Configuration Key discussed in Section II is also part of the Readout Supervisor data bank.

V. Data Monitoring

To ensure data integrity and to detect malfunctioning components early, both the hardware and the software components involved in the data taking process are monitored [5]. In addition, a dedicated Monitoring Farm consisting of a few tens of general purpose processors spies at best effort on the streams of events accepted by the HLT. The monitoring of the HLT itself may thus produce information mostly on the rejected events.

The Monitoring Farm produces statistics based on both an analysis of the raw data, and on an analysis of the output from a few reconstruction tasks. The reconstruction allows checking the data using high level physics objects which thus allows verifying the calibration and space alignment, as well as the overall performance of the detector online.

The output of the monitoring tasks comprises summary information and statistical data in the form of histograms and scalars, such as counters. The distributed processing on the many nodes of the HLT and of the Monitoring Farm, require that the information from the different sources is summed. This is achieved by a tree-like structure of "Adder" tasks. All the monitoring information is subscribed to and saved to disk regularly and finalized at the end of a run by special "Saver" tasks. The online monitoring information may at any time be inspected and analyzed using an interactive Histogram Presenter. The Histogram Presenter may also store and retrieve monitoring views from a Histogram Database.

At the end of each run, an automatic histogram analysis is performed which compares the collected monitoring information with reference data.

A more rigorous data quality check is made offline after the full reconstruction of the events. However, before launching the processing of the files of the full 2 kHz data stream, a special fast quality check is made on a duplicate 5 Hz sub-sample. This so called Express Stream allows checking regularly the integrity of the offline software and validating the final calibration and alignment parameters.

VI. Global System Commissioning

The commissioning of LHCb has consisted of two years of intense work between 2006 and 2008 in which the main aim has been to operate the detector as a unit with common procedures and tools, and to understand and calibrate the sub-detectors. The aim has also been to operate the experiment with only two shifters and reach an operational efficiency which allows bringing the entire experiment to running from a cold-start in less than 10 minutes and restarting data-taking in less than a minute. While these goals have been reached technically, more shifter friendly diagnostics tools are in development to reach full operation efficiency with non-expert shifters.

The exercises have been performed using calibration pulses and radioactive sources in addition to various auto-triggers, but

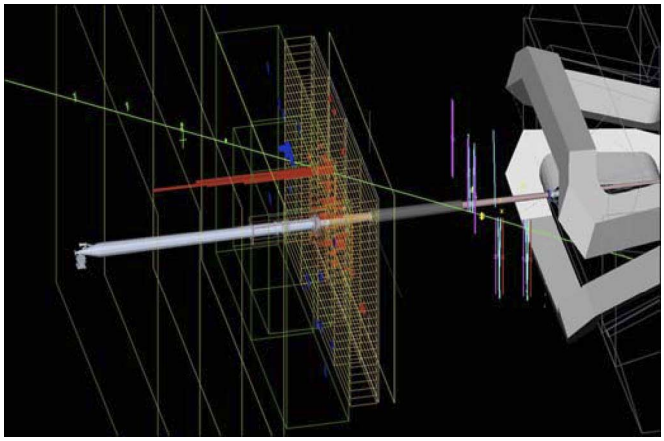

Fig. 5. Cosmic ray event recorded for calibration purposes. A muon is travelling from left to right in the picture and is detected with the correct timing in the muon chambers, the calorimeter and in the straw-tube based Outer Tracker on the right.

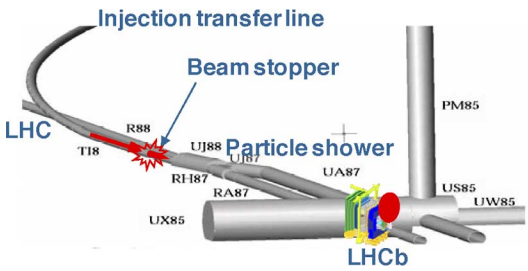

Fig. 6. LHCb was able to use the LHC transfer line tests for space alignment by recording the induced particle showers produced when the pilot bunch was dumped on the beam stopper at the end of the transfer line.

more importantly also using cosmic rays and the first LHC injections and circulating beams. The LHCb readout system has been in use successfully during the entire installation and commissioning phase.

A crucial tool in the adjustment of the readout timing of the sub-detectors is the possibility of reading out consecutive 25 ns-crossings around an activity trigger as detected by a subsystem, and processing each set of consecutive crossings as special Timing Alignment Events. It also allows measuring the signal leakage into the preceding and subsequent clock cycles and, consequently, optimizing the signal-to-spillover ratio.

Although the LHCb geometry is not well-suited for cosmic rays, more than one and a half million events were recorded with the large sub-detectors in the summer of 2008. Fig. 5 shows a cosmic ray traversing the Muon chambers, the Calorimeters and the Outer Tracker. LHCb also benefitted from the first LHC transfer line tests in which a pilot bunch was continuously dumped on a beam stopper at the end of transfer line (Fig. 6). The beam stopper is almost perfectly in line with the LHCb experiment and thus gave rise to particle showers through LHCb which were very useful for space alignment of the small silicon trackers.

Fig. 7 shows an event from the first days with circulating beam in the LHC on September 10, 2008. The splash comes from a pilot bunch being dumped on a collimator upstream of LHCb.

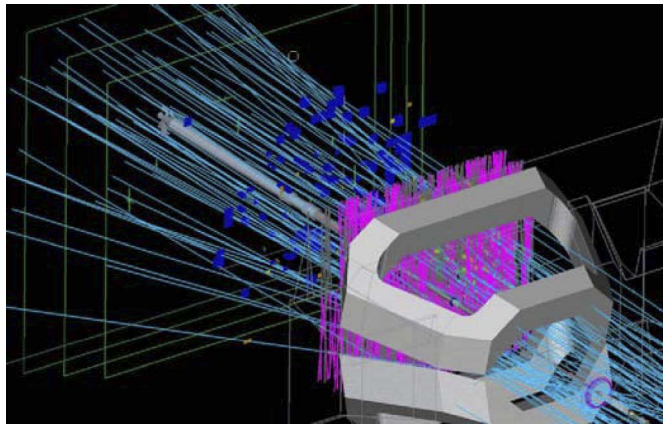

Fig. 7. Event recorded while the LHC pilot bunch was dumped on the LHC collimator upstream of LHCb on September 10, 2008. A large shower of particles produced in the interaction with the collimator travels through the LHCb detector.

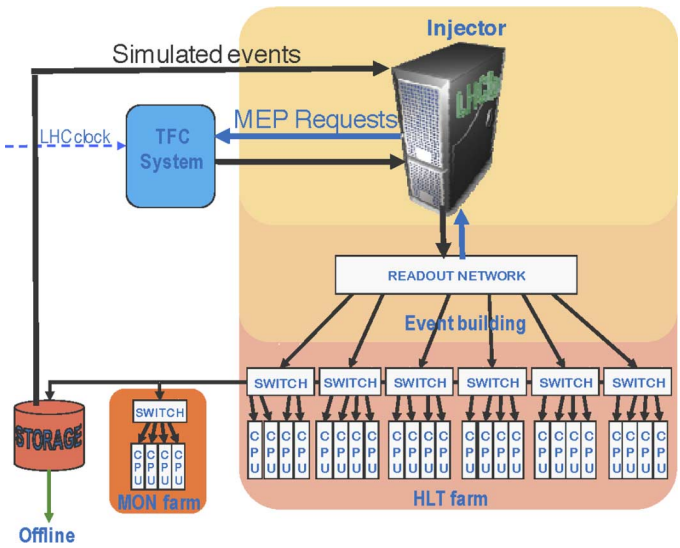

Fig. 8. A real-time system test which includes the entire offline processing chain as well has been conceived. Simulated data is injected in the online processing farm at the HLT rate by a high-end server.

A. *Real-Time Full System Test*

In the absence of beam during the shutdown 2008–2009, a complete real-time system test [6] has been conceived which allows validating the High Level Trigger, the data flow and the offline processing. Said differently, it is aimed at demonstrating the readiness of the entire chain to record, transfer, process, and analyze the about 7 million events which LHCb will get per hour starting from the first day with collisions.

The system test includes the online processing farm, the storage system, the transfer to the offline and the entire sequence of offline processing, including the databases and the bookkeeping. As Fig. 8 shows, the test consists of injecting simulated Level-0 accepted events using a software "Injector" running on a high-end server in the Readout Network in order to feed the HLT processing farm with a random rate of 2 kHz. The Timing and Fast Control System is part of the test to generate the random rate and to provide the Readout Supervisor event data bank.

VII. Upgrade Path

LHCb has recently submitted an Expression of Interest for an LHCb Upgrade [7] which would allow operating LHCb at ten times the current design luminosity and allow improving the trigger efficiencies in order to collect more than ten times the statistics foreseen in the first phase of LHCb. Improving the trigger efficiencies requires in practice eliminating the hardware trigger entirely and install a trigger-free 40 MHz complete event readout in order to perform the event selection on a processing farm with only a high-level software trigger with access to all detector information.

The R&D activities are already well underway.

VIII. Conclusion

LHCb has become an operational experiment "waiting" for beam. The readout system is based on a simple, robust and scalable architecture with advanced readout control and real-time event management implemented in a central FPGA-based Readout Supervisor. The single point-of-failure aspect is solved by having several equivalent optional Readout Supervisors which are normally used for stand-alone runs of the various sub-detectors. The system is very mature and has been extensively tested. The experience shows that a good compromise has been reached between the use of Commercial-Off-The-Shelf equipment and custom electronics.

The global commissioning of the readout system together with all the sub-detectors has produced a proof-of-concept although it has also showed that improved non-expert diagnostics tools for the shifters are needed to reach full operational efficiency.

Apart from the use of cosmic rays to calibrate and understand the sub-detectors, the LHC injection tests at the end of August 2008 gave LHCb the possibility to record the first ever LHC-induced tracks.

LHCb has already started an activity aimed at studying the possibilities of upgrading LHCb. This includes a full 40 MHz trigger-free readout in which the event-selection is only performed on a processing farm. This is a challenging concept which is receiving increasing attention also at future accelerators.

References

[1] LHCb Reoptimized Detector Design and Performance: Tech. Design Rep., CERN-LHCC-2003-030, LHCb Collaboration.

[2] LHCb Online System TDR, Dec. 19, 2001, CERN/LHCC 2001-040, The LHCb Collaboration.

[3] A. Bay, A. Gong, H. Gong, G. Haefeli, N. Neufeld, and O. Schneider, "Control and operation of the LHCb readout boards using embedded microcontrollers and the PVSS II SCADA system," *Nucl. Instrum. Methods Phys. Res. A*, vol. A560, pp. 494–502.

[4] Z. Guzik, R. Jacobsson, and B. Jost, "Driving the LHCb front-end readout," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 3, pp. 508–512, Jun. 2004.

[5] O. Callot *et al.*, "Online data monitoring in the LHCb experiment," in *J. Phys.: Conf. Series*, 2008, vol. 119, p. 021015.

[6] M. Frank, J. C. Garnier, C. Gaspar, and N. Neufeld, "Online test-bench for LHCb high level trigger validation," *J. Phys.: Conf. Series*, CHEP2009, submitted for publication.

[7] Expression of Interest for an LHCb Upgrade, Apr. 22, 2008, CERN/LHCC/2008-007, LHCb Collaboration.