Chapter 8

Online System

The task of the Online system is to ensure the transfer of data from the front-end electronics to permanent storage under known and controlled conditions. This includes not only the movement of the data themselves, but also the configuration of all operational parameters and the monitoring of these, as well as environmental parameters, such as temperatures or pressures. The online system also must ensure that all detector channels are properly synchronized with the LHC clock. The LHCb Online system is described in detail in [214, 217, 218]

8.1 System decomposition and architecture

The LHCb Online system consists of three components:

- the Data Acquisition (DAQ) system,
- the Timing and Fast Control (TFC) system,
- the Experiment Control System (ECS).

The general architecture of the LHCb online system is shown in figure 8.1.

8.2 Data Acquisition System

The purpose of the Data Acquisition (DAQ) system is the transport of the data belonging to a given bunch crossing,and identified by the trigger, from the detector front-end electronics to permanent storage. The design principles for the DAQ architecture (figure 8.1) are:

- Simplicity: simple protocols and a small number of components with simple functionalities
- Scalability: ability to react to changing system parameters, such as event sizes, trigger rates or the CPU needs of trigger algorithms.
- Only point-to-point links: components are connected through point-to-point links only. No buses are used (outside monolithical boards). This leads to a more robust system.

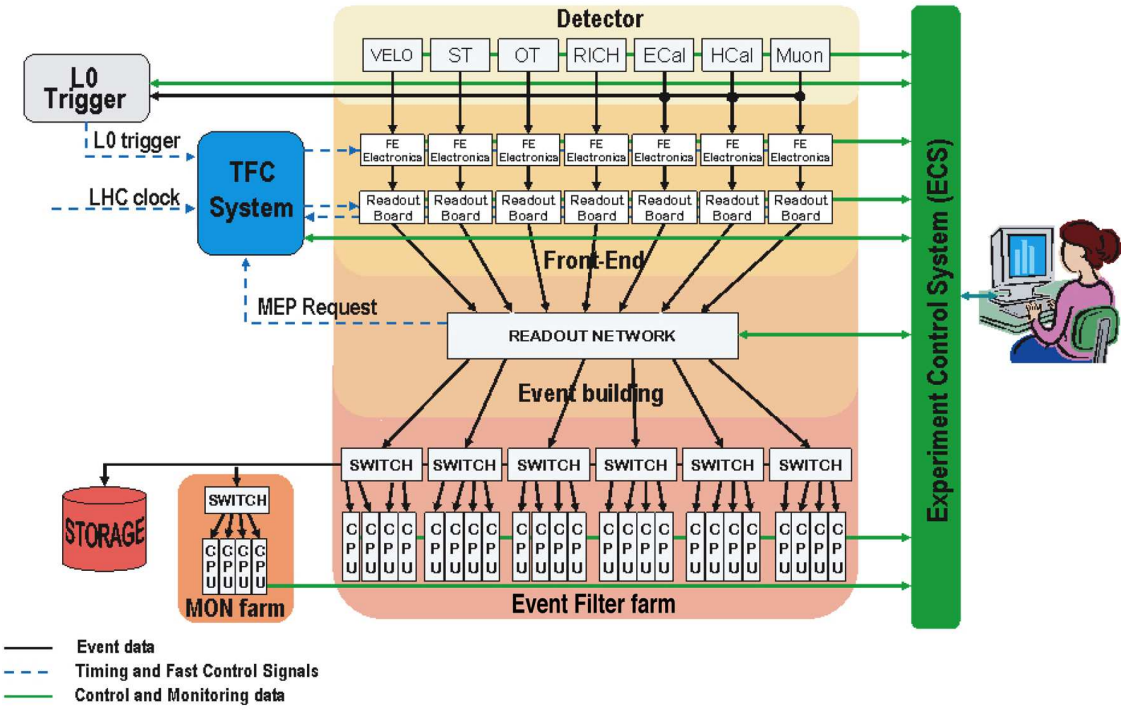

Figure 8.1: General architecture of the LHCb Online system with its three major components: Timing and Fast Controls, Data Acquisition and Experiment Control System. A blow-up of the the TFC box is shown in figure 8.3.

- Use of Commercial off-the-shelf (COTS) products and, wherever possible, commodity components and protocols.

These principles allowed to construct a reliable and robust system with enough flexibility to cope with possible new requirements, motivated by experience with real data.

Data from the on/near-detector electronics (front-end electronics) are collected in LHCb-wide standardized readout boards (TELL1).[1] Figure 8.2 shows a simplified block diagram of the TELL1 board. A detailed description can be found in [10].

Data are received from the detector electronics either by optical or analogue receiver cards and processed in four pre-processing FPGAs,[2] where common-mode processing, zero-suppression or data compression is performed depending on the needs of individual detectors. The resulting data fragments are collected by a fifth FPGA (SyncLink) and formatted into a raw IP-packet that is subsequently sent to the DAQ system via the 4-channel GbEthernet mezzanine card. The board interfaces to the Experiment Control System (ECS) by means of a credit-card sized PC mounted on the board. Clock and synchronization signals (e.g. triggers) are transmitted through the on-board Trigger, Timing and Control (TTC) interface [219]. Flow control to the TFC system is performed through the throttle signal driven by the SyncLink FPGA.

[1] The RICH detectors use a specific board (c.f. 6.1) which is, from the data readout point of view, functionally identical to the TELL1.

[2] Altera Stratix 1S25.

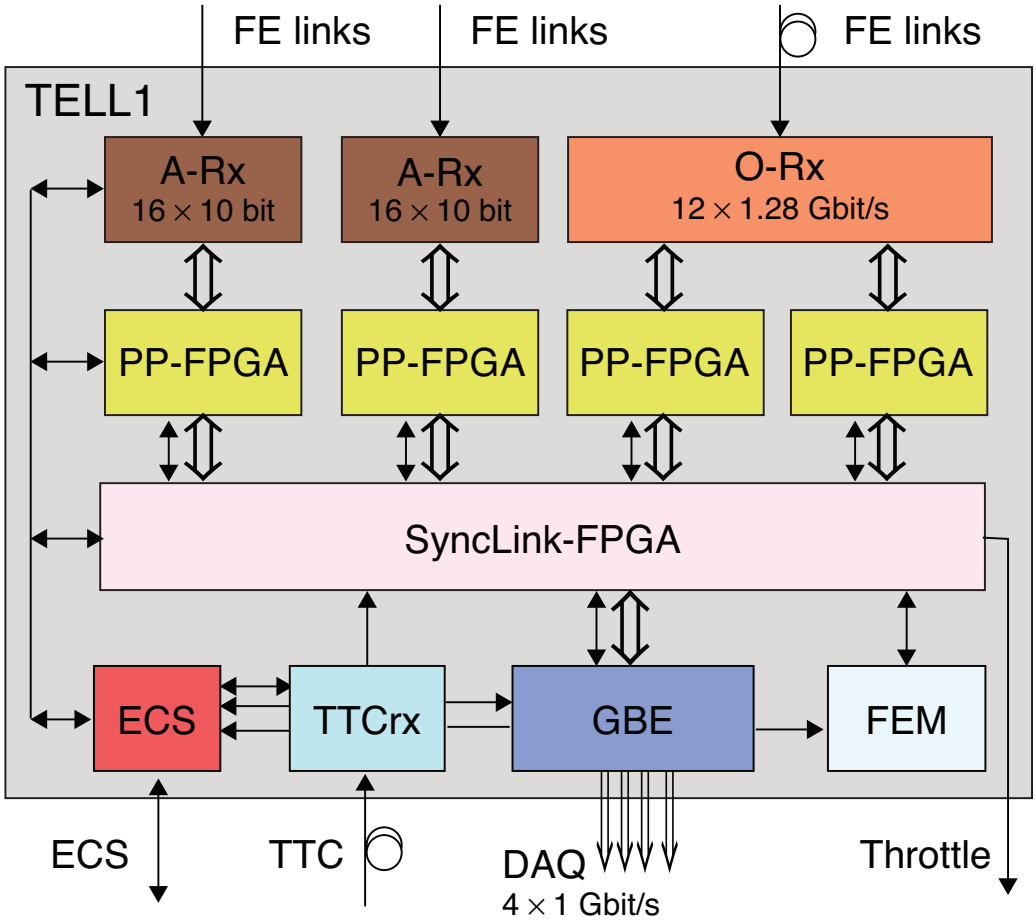

Figure 8.2: Simplified block diagram of the common readout board TELL1. The FEM (Front-End Multiplexer) allows to merge event fragments from several input links to form one output fragment.

In the CPU farm, the HLT algorithm selects interesting interactions; upon a positive decision, the data are subsequently sent to permanent storage. The HLT (see 7.2) is expected to reduce the overall rate from the original trigger rate of 1 MHz to ~2 kHz, hence by a factor of 500. The storage system is expected to have a capacity of ~40 TB, which should offer sufficient buffer space to cope with possible interruptions of the transfer to permanent storage at CERN. Gigabit-Ethernet was chosen as link technology, mainly because of its wide, almost monopoly-like, acceptance in the LAN market and its low price. The very wide range of speed from 10 Mb/s to 10 Gb/s, and the availability of very big switches (>1200 ports per chassis) are also important assets.

The TELL1 board offers 4 Gb Ethernet ports as output stages. Some of these are fed into a large switching network providing the connectivity between the TELL1 boards and the individual Farm nodes. To overcome the significant overhead per frame of Ethernet, the concept of Multi-Event Packets has been devised, in which the data of several triggers (~10) are collected in one IP packet and transferred subsequently through the network. The size of the CPU farm running the HLT trigger algorithms is determined by the average execution time of the HLT algorithm per event but also possibly by the maximum bandwidth into an individual processing node: if the execution

time were to be very low, the input bandwidth might constitute the limiting factor and the number of *boxes* would have to be increased. The HLT algorithms are executed on a sizeable farm of CPUs. It's is expected to consist of 1000–2000 1U servers containing CPUs with multi-core technologies. The starting size of the farm will be about 200 servers. The maximum available space is 2200 U. The large number of CPUs is organized into 50 sub-farms of 20–40 CPUs each. The scalability is then guaranteed since one sub-farm is a functional unit and there is no cross-communication between sub-farms.

The quality of the acquired data is checked in a separate monitoring farm that will receive events accepted by the HLT and will house user-defined algorithms to determine e.g. the efficiencies of detector channels or the mass resolution of the detector. Also, some rate of L0 accepted and random triggers will be used to monitor the trigger itself.

8.3 Timing and Fast Control

The TFC system drives all stages of the data readout of the LHCb detector between the front-end electronics and the online processing farm by distributing the beam-synchronous clock, the L0 trigger, synchronous resets and fast control commands. The system is a combination of electronic components common to all LHC experiments and LHCb custom electronics. The TFC architecture shown in figure 8.3 can be described in terms of three main ingredients, the TFC distribution network, the trigger throttle network, and the TFC master (Readout Supervisor).

The TFC optical distribution network with transmitters and receivers is based on the LHC-wide TTC system developed at CERN [219]. In addition to transmitting the beam synchronous clock, the protocol features a low-latency trigger channel and a second channel with framed user data used to encode the control commands. A switch has been developed and introduced into the distribution network to allow a dynamic partitioning of the LHCb detector to support independent and concurrent sub-detector activities such as commissioning, calibration and testing.

The optical throttle network is used to transmit back-pressure, that is a trigger inhibit, from the asynchronous parts of the readout system to the Readout Supervisor in case of congestion of the data path. The network incorporates a Throttle Switch to support the requirement that the readout system is partitionable, and to allow modules to perform an *OR* of the throttle signals of each sub-system locally.

The heart of the system, the Readout Supervisor, implements the interface between the LHCb trigger system and the readout chain. The Readout Supervisor synchronizes trigger decisions and beam-synchronous commands to the LHC clock and orbit signal provided by the LHC. It is also capable of producing a variety of auto-triggers for sub-detector calibration and tests, and performs the trigger control as a function of the load on the readout system. In order to perform dynamic load balancing among the nodes in the online processing farm, the Readout Supervisor also selects and broadcasts the destination for the next set of events to the Readout Boards based on a credit scheme in which the farm nodes send data requests directly to the Readout Supervisor.

For each trigger the Readout Supervisor transmits a data bank over the readout network which is appended to the event data and which contains the identifier of the event, the time and the source of the trigger.

2008 JINST 3 S08005

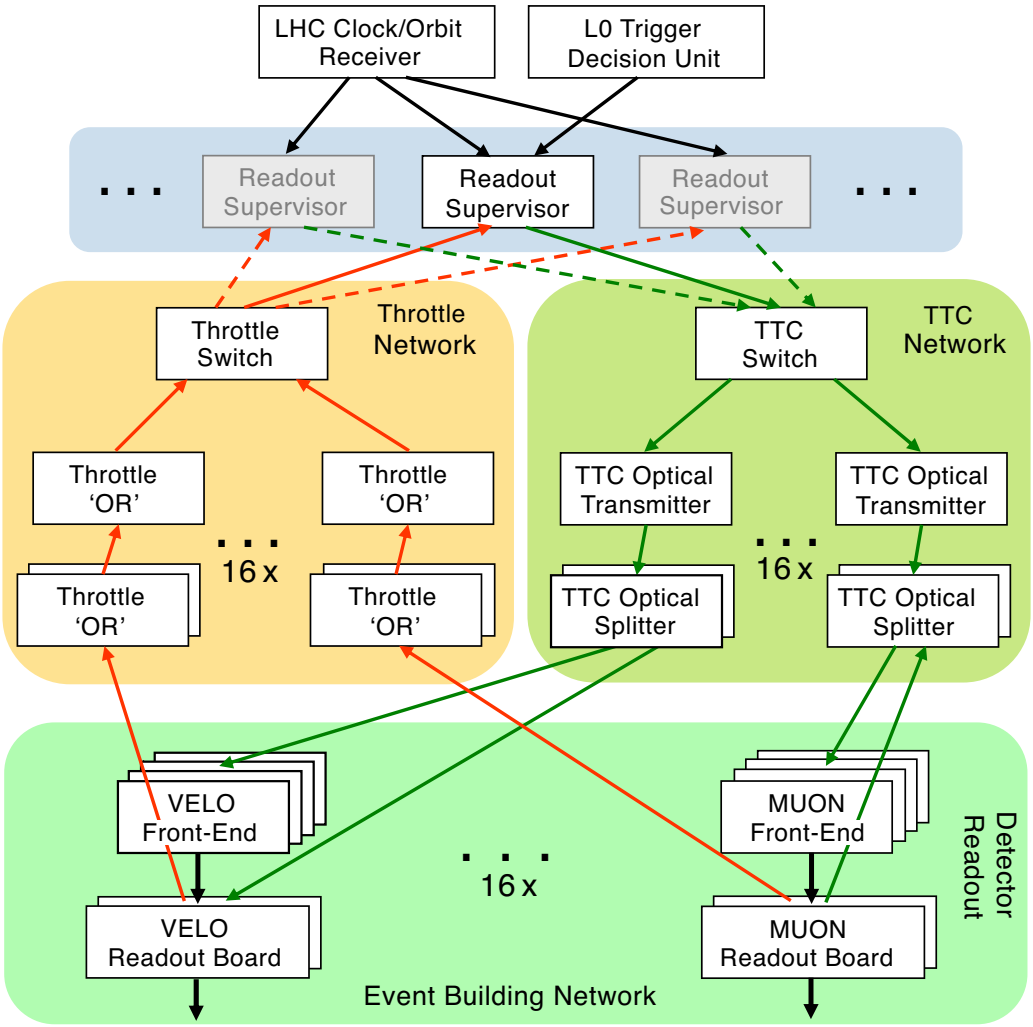

Figure 8.3: Schematic diagram of the TFC architecture. For a given partition there is only one RS, but several partitions can operate at the same time.

8.4 Experiment Control System

The Experiment Control System (ECS) ensures the control and monitoring of the operational state of the entire LHCb detector. This encompasses not only the traditional detector control domains, such as high and low voltages, temperatures, gas flows, or pressures, but also the control and monitoring of the Trigger, TFC, and DAQ systems. The hardware components of the ECS are somewhat diverse, mainly as a consequence of the variety of the equipment to be controlled, ranging from standard crates and power supplies to individual electronics boards. In LHCb, a large effort was made to minimize the number of different types of interfaces and connecting busses. The field busses have been restricted to:

- SPECS, Serial Protocol for ECS, a serial bus providing high-speed, 10Mb/s, control access to front-end electronics [41],

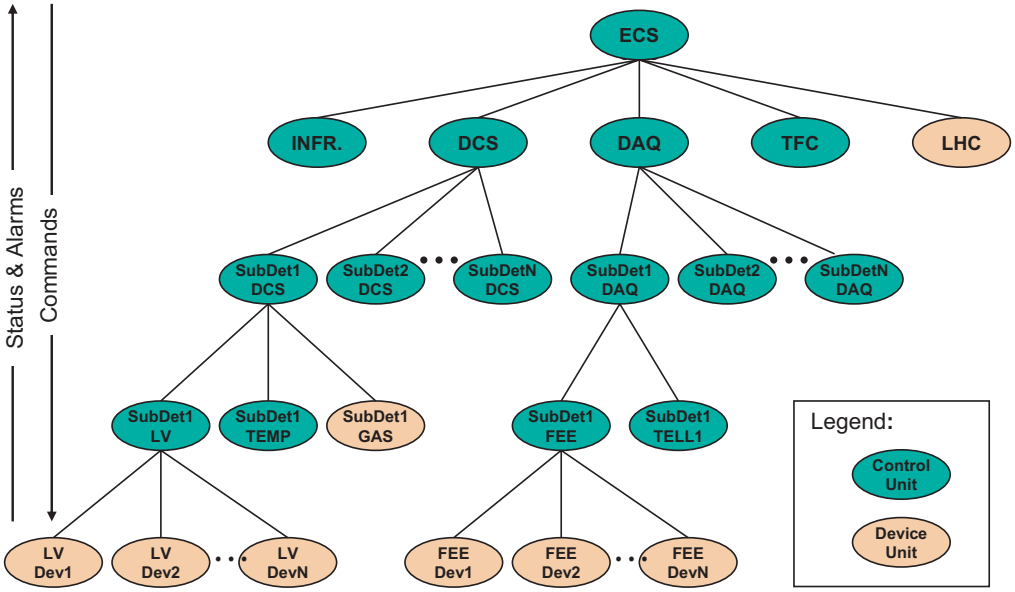

Figure 8.4: ECS architecture.

- CAN (Controller Area Network[3]),
- (fast)Ethernet.

The first two, SPECS and CAN, are mainly used for equipment residing in the high radiation area close to the detector. The associated interfaces tolerate modest levels of radioactivity but are not radiation hard. Ethernet is only used in the radiation free areas, such as the electronics barracks or on the surface. Ethernet is used to control individual PCs, as well as the individual electronics boards usedfor the readout through Credit-Card sized PCs mounted directly on each board. This choice allows the use of normal PCs over their standard Ethernet interfaces for controlling the readout electronics.

The ECS software is based on PVSS II, a commercial SCADA (Supervisory Control And Data Acquisition) system. This toolkit provides the infrastructure needed for building the ECS system, such as a configuration database and communication between distributed components, graphical libraries to build operations panels, and an alarm system as well as components, such as OPC clients. Based on PVSS, a hierarchical and distributed system was designed as depicted in figure 8.4.

Device Units, in figure 8.4, denote low-level access components which model the physical device and typically communicate directly with the hardware. In general they only implement a very simple state machine which is exclusively driven by the controlling Control Unit. Examples of Device Units are power supplies, and software processes, such as the HLT processes.

Control Units implement high-level states and transitions and also local logic to support recovery from errors of subordinate Device Units. Typical examples of Control Units are a HV subsystem, or the component that controls the ensemble of crates of a sub-detector or an entire

[3] *ISO Standard 11898*, see e.g. www.iso.org.

sub-farm of the Event Filter Farm. Control Units can be controlled by other Control Units, to allow the building of a hierarchy of arbitrary depth. State sequencing in the ECS system is achieved using a Finite State Machine package, based on SMI++ that allows creating complex logic needed, for example, for implementing elaborate sequencing or automatic error recovery.

The distributed components of the ECS system are connected with a large Ethernet network consisting of several hundred Gigabit and Fast Ethernet links.