

CERN-LHCC-2005-039
CERN-LHCC-2001-040-Add1
4 November 2005

LHCb

Addendum to the LHCb Online System Technical Design Report

LHCb Collaboration

CERN
Geneva, 2005

Table of Contents

Chapter 1	Introduction.....	1
Chapter 2	System Architecture.....	2
	2.1 System Components	2
	2.1.1. Front-End Electronics	2
	2.1.2. Readout Network	2
	2.1.3. CPU Farm	2
	2.1.4. Event Distribution and Load Balancing.....	3
	2.2 Comparison with the old system	3
Chapter 3	System Design and Implementation	4
	3.1 Scale of the system	4
	3.1.1. Event Size	4
	3.1.2. Transport Overheads	6
	3.1.3. Number of Tell1 Boards and GbEthernet Links	7
	3.1.4. Number of Switch Output Ports.....	8
	3.1.5. Size of the CPU Farm	9
	3.2 Implementation	9
	3.2.1. Baseline Implementation	9
	3.2.2. Scalability	9
	3.2.3. Cost	10
Chapter 4	Potential Physics Benefits.....	11
Chapter 5	Summary	12
	5.1.1. Planning and Milestones	12
References	13

List of Figures

Figure 1 Architecture of the DAQ System as described in the Trigger TDR.....	1
Figure 2 Architecture of the proposed DAQ system.....	2
Figure 3 Architecture of a system designed for significantly larger event sizes.....	10

List of Tables

Table 1	Average event size per sub-detector from the current Monte-Carlo simulation	.5
Table 2	List of scale factors applied to the event size for estimating a conservative size of the system.	5
Table 3	Event Size per sub-detector as a result of the current Monte-Carlo simulations including the safety factors of Table 2	6
Table 4	Event and Fragment sizes and data rates per sub-detector with a packing factor of 13 and a Ethernet Maximum Transfer Unit of 1500 Bytes	7
Table 5	Event and Fragment sizes and data rates per sub-detector with a packing factor of 13 and a Ethernet Maximum Transfer Unit of 1500 Bytes	8
Table 6	Summary of the scale of the system	9
Table 7	Major milestones and planning check-points for the system development and installation	12

Chapter 1 Introduction

The LHCb online system was originally described in the Online System TDR submitted in 2001 [1]. A sub-system of the online system is the Data Acquisition (DAQ) system, which deals with the movement of the data from the detector front-end electronics, via the software triggers, to the storage. Originally, the DAQ system was only concerned with the High-Level Trigger (HLT) data flow and the Level-1 trigger was implemented in a dedicated system, external to the online system.

In the Trigger System TDR, submitted in 2003 [2], an extended DAQ system was described. The extension consisted in integrating the Level-1 trigger into the overall DAQ infrastructure as shown in Figure 1. The Level-1 trigger is integrated in the overall DAQ system's dataflow but is still a distinct entity. At Level-1, data of reduced precision from some of the sub-detectors are sent to the CPU farm at the rate of 1 MHz. The Level-1 algorithm running in the farm reduces this rate to 40 kHz. The full data of these selected events is then sent to the CPU farm where it is processed by the HLT algorithms.

Thus both the read out network and the farm must accommodate both Level 1 and HLT events.

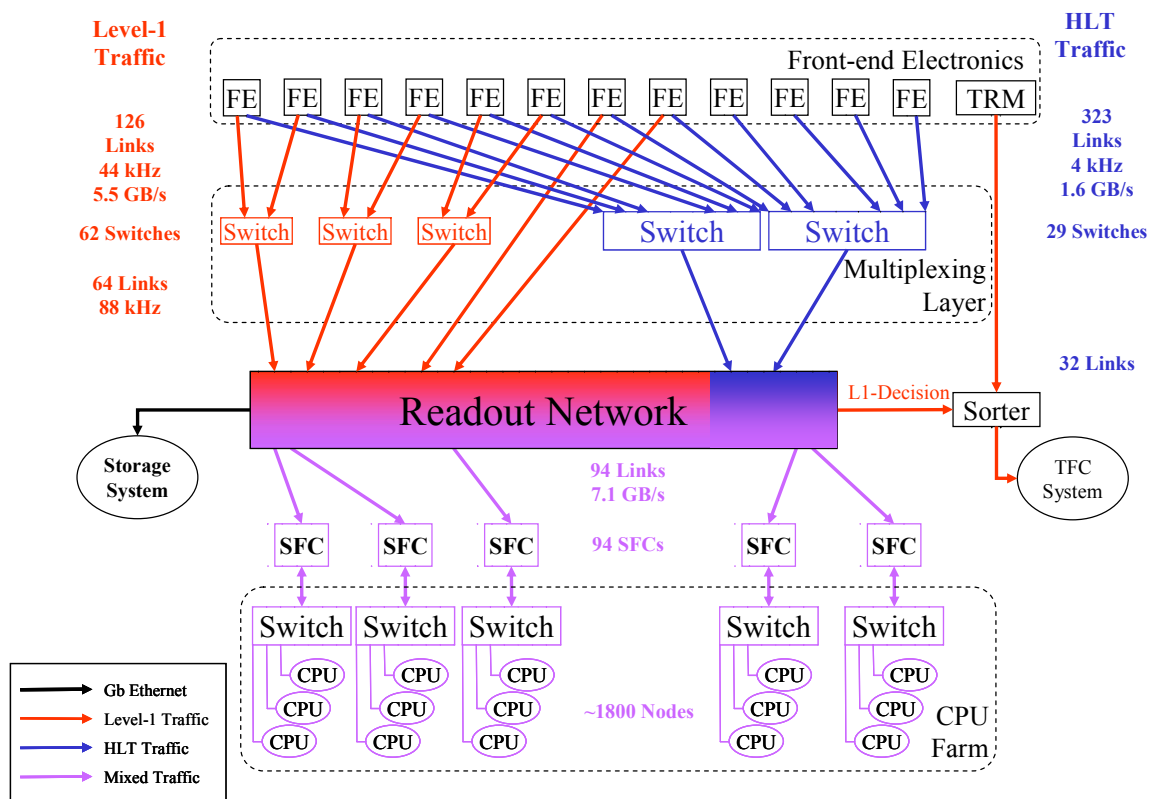


Figure 1 Architecture of the DAQ System as described in the Trigger TDR

The appearance of very high performance (>1 Tb/s switching capacity) routers/switches at affordable prices, with more than 1000 ports per chassis, allowed the evolution of the system described in this report. A system is now proposed in which the Level-1 trigger is eliminated as a distinct item and the entire detector is read out with full precision and the data sent to the CPU farm at the Level-0 accept rate of 1 MHz.

Chapter 2 System Architecture

In this chapter the new architecture of the system is described. It is depicted in Figure 2.

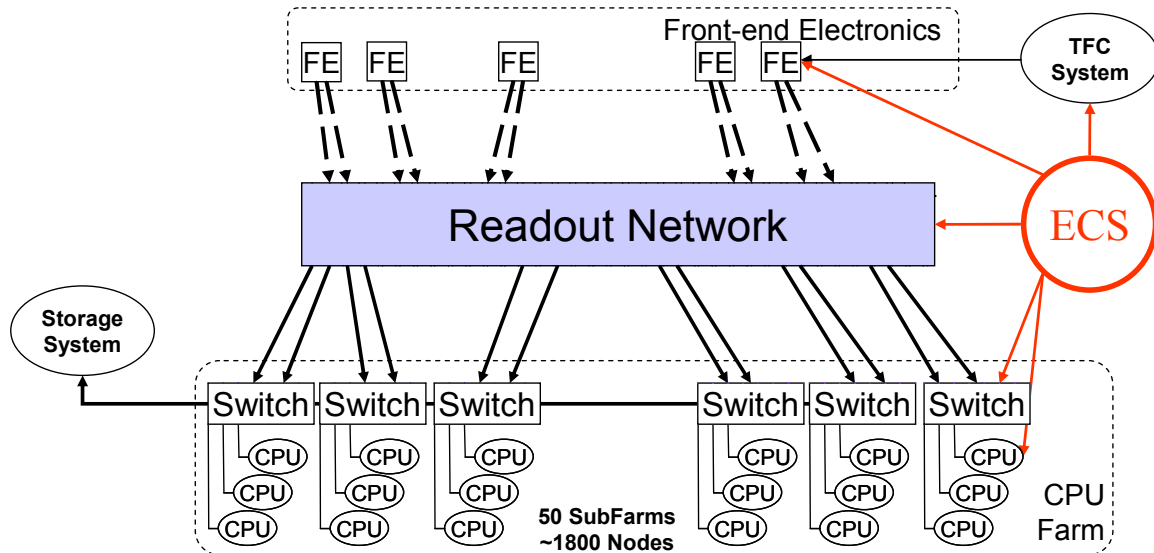


Figure 2 Architecture of the proposed DAQ system

2.1 System Components

2.1.1. Front-End Electronics

The front-end electronics interfaces the detector readout to the Data Acquisition (DAQ) system. It receives the data from the detector electronics, processes them (zero-suppression) and injects them through an LHCb-standard, 4-channel GbEthernet card into the DAQ system. The front-end electronics is, for most sub-detectors, implemented in a standard board, Tell1 [3]. The RICH detectors implement front-end electronics functionality in a similar board called UKL1¹.

2.1.2. Readout Network

The readout network provides the connectivity between the front-end electronics and the CPU farm. Each Tell1 board must be able to send data to any of the CPUs in the farm. The readout network must route Ethernet frames from the Tell1s to the farm nodes without packet loss, as long as the network is operated within the defined limits. For reasons of manageability and operational integrity, we prefer a single chassis or a small number of chassis, depending on availability (see Section 3.1.5).

2.1.3. CPU Farm

The CPU farm consists of two components, namely the distribution switches and the farm nodes themselves. The task of the distribution switches is to provide the connectivity between the output ports of the readout network and the individual nodes. This connectivity is not necessarily at wire

¹ In this text we will use the name Tell1 generically to include the UKL1 as well.

speed, since the number of output ports of the readout network is smaller than the number of nodes, thus leading to a de-multiplexing function of the switches.

2.1.4. Event Distribution and Load Balancing

In the old scheme, a two level load balancing scheme was foreseen: a static event distribution (round-robin) of events to the SFCs, and dynamic load balancing, based on a token scheme, of the farm nodes done by the SFC after the event building stage.

For the new system two scenarios have been studied and simulated

- Static load balancing: Events are distributed among the farm nodes according to their CPU power assigned statically. Statistical fluctuations of the processing time would be absorbed by large buffers inside the nodes. Since no Level 1 latency constraints apply, this scheme should work well.
- Dynamic load balancing: As an alternative, a token based scheme has been proposed. In this case, each node requests to the Readout Supervisor (RS) to receive events. The RS subsequently picks a node from the list of pending requests when a new Multi-Event Packet (MEP) is ready to be sent.

Our preference is the dynamic destination assignment and this is currently being implemented. It should be noted that neither of the above schemes implies hardware changes in the Readout Supervisor. There are sufficient resources available in the RS to handle any of the above schemes. For more details see [4].

2.2 Comparison with the old system

Comparing the old and new architecture, it is evident that the latter represents a major simplification, namely

- Reduction to only one uniform dataflow, i.e. there is only one kind of data flowing from the Front-end Electronics towards the CPU farm
- The TRM (Trigger Receiver Module) is not necessary anymore
- The decision sorter component is not necessary anymore
- The subfarm controllers (SFCs), a mandatory component in the old system, are eliminated
- The lack of buffering of events in the TELL1s during Level 1 processing removes latency constraints.

Other modifications/additions to the system are

- Zero suppression and data compression in some of the TELL1s
- Load balancing (static or dynamic) in the Readout Supervisor

The changes are in the data-flow system only. The other components, the Experiment Control System (ECS) and the Timing and Fast Control System (TFC) are not affected². The data transfer protocols, especially the concept of Multi-Event Packets (MEP), described in [2], are still valid. Of course, MEPs related to the Level-1 trigger have disappeared.

² The TFC System is marginally affected in that, of course, all items related to the Level-1 trigger are eliminated.

Chapter 3 System Design and Implementation

In this chapter we present the design and implementation of the system. We first discuss the expected scale of the system and subsequently present the physical implementation.

3.1 Scale of the system

The elimination of an explicit Level-1 trigger implies that all data have to be read out of the front-end electronics at the Level-0 rate (1 MHz)³. The following factors determine the scale of the system

- The average data size per trigger
- Overheads of the transport of the data
- The output bandwidth of the TELL1 boards which is limited to 4 GbEthernet ports.
- The maximum allowed load of each GbEthernet link

3.1.1. Event Size

The true event size will only be known once the LHC accelerator and the LHCb detector are operational. For the time being we have to rely to Monte-Carlo studies of the pp-collisions and apply plausible safety factors to cope with unforeseen effects.

Current Monte-Carlo studies result in an event size for the raw data of ~35 kB/event. In Table 1 the event size per sub-detector is listed from the current Monte-Carlo simulations.

³ The Level-0 rate is limited to an absolute maximum of 1.1 MHz by the fact that the readout of the front-end chips takes 900 ns. 1.1 MHz is thus the maximum instantaneous rate; 1 MHz is the maximum sustained rate and will be used throughout this document.

Table 1 Average event size per sub-detector from the current Monte-Carlo simulation

Subdetector	Event Size [Bytes]
Velo-r	5393
Velo-Phi	4424
RICH1	2448
RICH2	1873
TT	3827
IT	3449
OT	4761
PS/SPD	1167
Ecal	3350
Hcal	900
Muon M1	454
Muon M2	266
Muon M3	63
Muon M4+M5	58
L0 PU	275
L0 Calo	533
L0 DU	128
L0 Muon	634
PUS	506
Readout Supervisor	64
Total	34572

Where appropriate a safety factor of 1.2 was applied to the Pythia multiplicity. The Velo and ST have recently developed a new encoding scheme that reduces their data sizes by factors of 0.9 and 0.75 respectively. Data sizes of secondary interactions have been scaled up by factors depending on vulnerability of the various sub-detectors to these interactions. The outer tracker data size has been increased by factor of 1.5 to take into account reflections and a longer drift time necessary to cope with a slower gas. ECal is now using a data compression scheme rather than a zero suppression that results in only a 15% increase in data size for a factor of 3 increases in multiplicity.

Applying these conservative safety factors, listed in Table 2, results in a total event size of ~52 kB/event, decomposed by sub-detector as shown in Table 3.

Table 2 List of scale factors applied to the event size for estimating a conservative size of the system.

Subdetector	Pythia	Encoding	Secondary interactions	Noise and Spillover	Total
Velo	1.2	0.9			1.08
ST	1.2	0.75	1.2		1.08
OT			1.4	1.5	2.1
RICH	1.2				1.2
Muon1			2		
Muon2-5			3		
Ecal					3
SPD/PRS	1.2				1.2

Table 3 Event Size per sub-detector as a result of the current Monte-Carlo simulations including the safety factors of Table 2

Subdetector	Event Size [Bytes]
Velo-r	5825
Velo-Phi	4778
RICH1	2938
RICH2	2248
TT	4133
IT	3725
OT	13997
PS/SPD	1400
Ecal	6700
Hcal	1800
Muon M1	907
Muon M2	797
Muon M3	190
Muon M4+M5	174
L0 PU	275
L0 Calo	533
L0 DU	128
L0 Muon	634
PUS	506
Readout Supervisor	64
Total	51752

3.1.2. Transport Overheads

As described in [2] we pack the data originating from several triggers into Multi-Event Packets (MEPs) to reduce the overheads originating from the Ethernet and IP Protocol used. These overheads can be decomposed into three contributions, where r is the trigger rate and P is the packing factor ($r = 1$ MHz, $P = 13$ in the example)⁴

- Overhead per trigger $o_t = 4$ Bytes occurring at rate r , ~ 4 MB/s
- Overhead per Ethernet frame, including IP header $o_E = 58$ Bytes occurring at rate $\sim r/P$, ~ 4.5 MB/s
- Overhead per MEP fragment $o_f = 12$ Bytes occurring at rate r/P , ~ 0.9 MB/s

With these overheads for a packing factor of 13 and a Maximum Transfer Unit (MTU) on the network of 1500 Bytes, the data sizes of Table 3 are modified as show in Table 4.

⁴ These overheads have to be compared with the overall bandwidth of a GbEthernet link of $1.25 \cdot 10^8$ B/s, i.e. they represent $\sim < 10\%$ of the bandwidth at reasonable packing factors.

Table 4 Event and Fragment sizes and data rates per sub-detector with a packing factor of 13 and a Ethernet Maximum Transfer Unit of 1500 Bytes

Subdetector	Event Size [Bytes]	Average Event Size per Tell1	Average MEP Size per Tell1	Data Rate [MB/s]
Velo-r	5825	139	1867	6406
Velo-Phi	4778	114	1543	5306
RICH1	2938	226	3002	3176
RICH2	2248	321	4239	2376
TT	4133	86	1183	4650
IT	3725	89	1217	4173
OT	13997	292	3855	14876
PS/SPD	1400	175	2339	1511
Ecal	6700	258	3414	7176
Hcal	1800	225	2989	1946
Muon M1	907	227	3012	980
Muon M2	797	199	2656	853
Muon M3	190	95	1299	213
Muon M4+M5	174	87	1197	193
L0 PU	275	138	1852	303
L0 Calo	533	266	3527	569
L0 DU	128	128	1728	142
L0 Muon	634	127	1712	699
PUS	506	126	1708	561
Readout Supervisor	64	64	896	73
Total	51752			56183

Thus, according to Table 4, the total data rate originating in the front-end electronics is 56 GBytes/s. This is clearly substantial and would have been very difficult and expensive to implement only a few years ago.

3.1.3. Number of Tell1 Boards and GbEthernet Links

There are two factors that influence the number of Tell1 boards and GbEthernet links. Firstly, the geometrical or local decomposition of the readout of the sub-detectors, which gives a lower limit on the number of Tell1 boards. Secondly, the amount of data that a Tell1 board produces and the level at which the maximum 4 GbEthernet links are used. To preserve a safety margin we have chosen a maximum average link load on the output of the Tell1 boards of 80%.⁵ Taking this into account one obtains the scale of the system at the input of the readout network shown in Table 5.

⁵ The link load of 80% might seem high. It should be noted, though, that once a GbEthernet frame is sent out, the data is transferred at a rate of 1 bit/ns, independent of the link load.—a lower link load only reduces the number of frames per second. Under the assumption that the receiving end can handle the frame rate (rather than the bit rate) the link load is a rather arbitrary concept. What remains to be verified is the capability of the receiving ends to absorb the frame rate presented to them.

Table 5 Event and Fragment sizes and data rates per sub-detector with a packing factor of 13 and a Ethernet Maximum Transfer Unit of 1500 Bytes

Subdetector	Event Size [Bytes]	Number of Tell1	Average Event Size per board	Average MEP Size per board	Data Rate [MB/s]	Number of Links	Max. Link Load
Velo-r	5825	42	139	1867	6406	84	68%
Velo-Phi	4778	42	114	1543	5306	84	56%
RICH1	2938	13	226	3002	3176	39	65%
RICH2	2248	7	321	4239	2376	28	68%
TT	4133	48	86	1183	4650	68	80%
IT	3725	42	89	1217	4173	54	72%
OT	13997	48	292	3855	14876	192	62%
PS/SPD	1400	8	175	2339	1511	16	76%
Ecal	6700	26	258	3414	7176	78	74%
Hcal	1800	8	225	2989	1946	24	65%
Muon M1	907	4	227	3012	980	12	66%
Muon M2	797	4	199	2656	853	12	59%
Muon M3	190	2	95	1299	213	3	70%
Muon M4+M5	174	2	87	1197	193	3	71%
L0 PU	275	1	275	3639	293	3	78%
L0 Calo	533	2	266	3527	569	6	76%
L0 DU	128	1	128	1728	142	2	57%
L0 Muon	634	5	127	1712	699	9	68%
PUS	506	4	126	1708	561	8	58%
Readout Supervisor	64	0	64	896	73	1	59%
Total	51752	309			56173	726	

The calorimeter needed to change the data format because the original zero suppression schemes were not possible at a rate of 1 MHz. This resulted in an increase of the data size per event of ~factor 2 and consequently an increase in the number of Tell1 boards (+16 for ECal and +4 for HCal). The RICH detector was not included in the original Level-1 trigger. Its readout was thus designed including a very large multiplexing factor of detector channels onto one readout board (UKL1 in this case). At 1 MHz, the multiplexing factors have to be reduced, leading to an increase in the number of boards by 9 boards. One additional board is also necessary for the Level-0 Calorimeter readout. In total ~30 additional readout boards (Tell1 or UKL1) are necessary to cope with the additional data rates.

3.1.4. Number of Switch Output Ports

Downstream of the readout network the traffic is completely equilibrated and hence the number of output ports is determined simply by dividing the total input data rate (56.2 GByte/s) by the bandwidth of one link, moderated with a chosen link load. As maximum link load we chose 85 %. From this we obtain 530 as number of output links.

In Table 6 the summary of the scale of the system is presented. The system will have to support 726 input ports and 530 output ports. The average load of the input links is 62% and the average load on the output links is 85%, as expected.

On the receiving end of these connections, the sub-farm switches will perform ~1:3 demultiplexing⁶. Each farm node will thus be presented with an input link load of less than 35%.

⁶ assuming 50 farms of 36 nodes each

Table 6 Summary of the scale of the system

Packing Factor	13
Ethernet MTU	1500
Total Event Size [Bytes]	51752
Number of Tell1	309
Total Input Data Rate [MB/s]	56173
Average MEP Size [Bytes]	2234
Frame Rate [Mf/s]	49.9
Switch Buffering Needs [kB]	730.3
Average Link Load	61.9%
Readout Network Input Ports	726
Readout Network Output Ports	530
Number of ports	1256
Input	726
Output	530

3.1.5. Size of the CPU Farm

The size of the CPU farm will be determined mainly by the processing power needed to execute the trigger algorithms. It is not foreseen to perform what used to be the HLT algorithm at the Level-0 accept rate. There will still be phases in the trigger processing, such as determination of tracks with large impact parameters and determination of secondary vertices, which will allow decisions to be reached very quickly before more elaborated algorithms are being executed.

There is an increase in the number of nodes necessary for coping with the fact that the one of the functions of the SFC, i.e. the event building, is now performed in each Node of the farm. This is estimated to be of the order of 150 node equivalent. How much CPU power is needed to execute more elaborate trigger algorithms is difficult to assess before the algorithms are designed and implemented.

Nonetheless this system in principle allows a more flexible trigger scheme which could include data from all sub-detectors.

3.2 Implementation

3.2.1. Baseline Implementation

To implement the proposed scheme a readout network with more than 1256 GbEthernet ports is needed. Recently devices of this dimension appeared on the market at affordable prices. The device we are envisaging is the Terascale E1200 from Force 10 Networks [5]. This device features 14 line cards per chassis with 90 GbEthernet ports⁷ each, equalling to 1260 GbEthernet ports per chassis. Thus devices satisfying our requirements are commercially available.

3.2.2. Scalability

A total of 1260 available ports for 1256 needed might look marginal at first sight. It should be noted, though, that the required numbers are based on conservative assumptions and that the average link load is only 62%. Thus, there is already some safety margin built into the system.

Should the observed event size exceed the assumed figures of Table 5 an upgrade scenario could look as depicted in Figure 3. The idea is to decompose the system into two (almost) disjoint

⁷ These linecards are 90/44 over-committed, i.e. only the equivalent of 44 GbEthernet ports can be sustained run at full speed bi-directionally. Since our traffic is unidirectional (from the Front-end Electronics towards the farm nodes) this does not concern us, as long as we suitably mix input and output ports on the same linecard.

subsystems, consisting of two readout network switches and two CPU farms⁸. Each Tell1 board would then have to have at least one connection into each network switch. This is actually easily fulfilled since the standard GbEthernet mounted on each front-end board card features 4 outputs. All MEPs from all Tell1 boards belonging to the same range of triggers would thus be sent to the same switch. This system would allow up to 2520 ports in total or 1260 input and 1260 output ports. 1260 input ports are more than the installed number of Tell1s can provide (~1240).

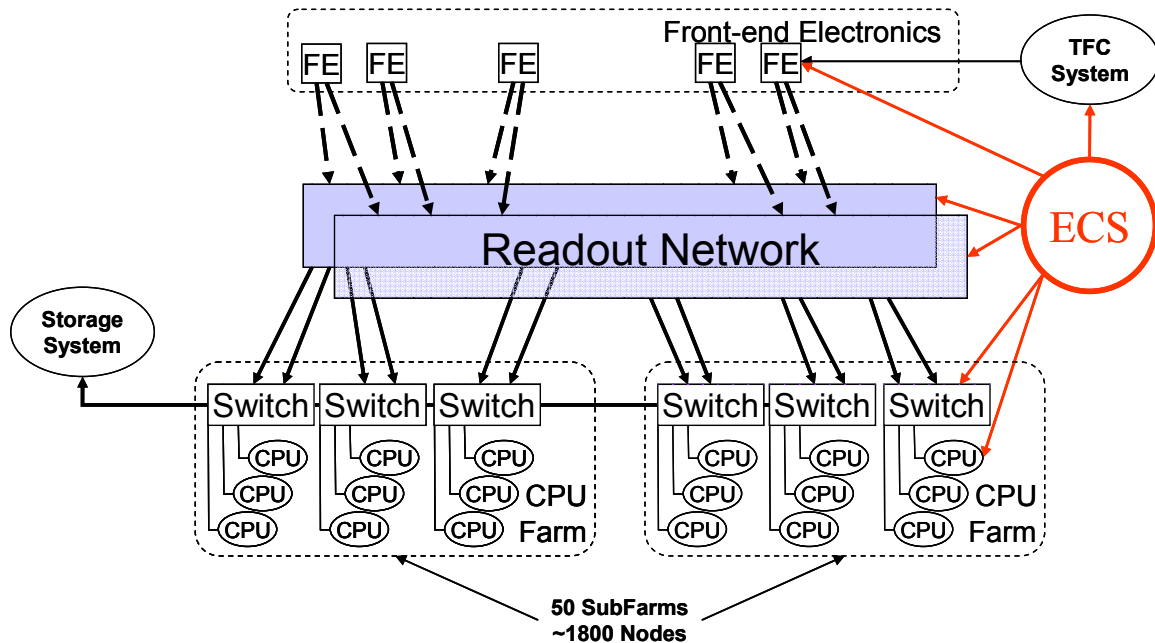


Figure 3 Architecture of a system designed for significantly larger event sizes.

3.2.3. Cost

Compared to the old system, the new system will need surely more switch ports and more link cables.

While the SFCs are eliminated their functionality will be taken collectively by the farm nodes. This leads to a saving, since the implementation of the SFCs needed high-end server PCs, whereas the farm nodes are of cheaper types.

As mentioned in section 3.1.3 some 30 additional Tell1 boards will be needed and those will be housed in 3 additional crates.

All in all, these components will lead to increased cost of ~300 kCHF compared to the system described in [2]. The resultant cost is, however, still below the cost estimates given in [1] and [2].

The cost increase of the upgraded system, such as shown in Figure 3 is estimated to be between 200 and 400 kCHF. The exact amount can only be estimated, once the true scale needed can be assessed, i.e. after the pilot run in 2007.

⁸ This is similar to the architectural concepts developed by CMS (see [6])

Chapter 4 Potential Physics Benefits

The physics performance should only gain from the adoption of the 1 MHz readout scheme, since the information previously used for the Level-1 decision would still be available, from VELO, TT, and the Level-0 decision unit, and exactly the same algorithms as before could be implemented. Indeed, this will be the baseline scenario, from which any improvements will be investigated. However, the information from all other sub-detectors will also be available for use in an extended Level-1 decision, rather than being strictly limited to use at the 40 kHz input rate of the HLT. When trying to use more information earlier in the trigger system, the overall CPU budget of the filter farm must be respected. However, if use of that extra information can lead to an earlier decision, one can gain. The major advantage of the increased simplicity of the online system should not be forgotten in terms of physics benefit, if it leads to a more robust system that would therefore increase the yields.

The first idea for extending the Level-1 decision, which was already discussed as an upgrade scenario in the Trigger TDR [2], is the inclusion of information from the tracking stations T1-T3 downstream of the spectrometer magnet. This information may be used to confirm the Level-0 candidates, by performing an “upstream” search for a track candidate in the T stations, that matches the calorimeter cluster or muon candidate that triggered at Level-0. It is estimated that about half of the hadron triggers from Level-0 have a true transverse momentum that is below the applied threshold, and the improvement in resolution that would be possible with the T station information should allow these candidates to be rejected.

Use of the information from muon stations M2-M5 in Level-1 has been studied, and an improvement in the $J/\psi \rightarrow \mu^+\mu^-$ selection has already been achieved by matching VELO track candidates to muon candidates reconstructed in the muon stations. The efficiency of selecting electron channels is significantly lower at Level-1 than for muons. There is therefore proportionately more room for gain in such channels, if calorimeter information could be used at Level-1 to validate the Level-0 candidate, or by using the calorimeter cluster to seed a T-station track search and comparing the track and calorimeter information.

The implementation of these ideas has not yet been worked out in detail. The main gain that is foreseen is an increased flexibility, and in particular one would be ready to improve the trigger if more funds (or cheaper CPUs) allow the filter farm to eventually be extended. The boundary between Level-1 and the HLT would no longer be fixed at 40 kHz, but could be adjusted according to the development of the algorithms in each. Indeed, more intermediate event selections can be introduced as required.

Chapter 5 Summary

The proposed system is a natural evolution of the LHCb DAQ system. Originally only designed to handle high-level trigger data and later extended to also handle the Level-1 trigger data, the system described in this report represents the ultimate step. It provides the final simplification, since only one data flow remains in the system and there is also only one type of processing software.

The benefits are

- Simplification of the dataflow and avoidance of interference between different data streams
- Reduction of the number of different kinds of modules and functions
- Only one single software process for triggering
- All event information available at all times in full precision. This should, eventually, allow improvements in the trigger efficiency

The system is affordable and the cost is covered within the limits set out in the relevant TDRs.

5.1.1. Planning and Milestones

The system proposed here is not fundamentally different from the one described in the Trigger system TDR [2]. It differs mainly in the simplification of the system (removal of components) and its larger scale. Thus there is no major change in the implementation procedures of the system.

The major milestones and planning check-points are shown in Table 7.

Table 7 Major milestones and planning check-points for the system development and installation

Milestone	Due Date	Description
Additional Cabling	01-Mar-06	Add additional cables in Pit to cope with additional bandwidth
Vertical Slice	01-Apr-06	Setup a vertical slice of the system consisting of 1 full Tell1 crate + TFC + (real) Switch+1 full subfarm
Install Readout Network	01-Sep-06	Install central switching network and farm switches
Farm	01-Oct-06	Install rudimentary farm (~300 Nodes) for commissioning
Farm Upgrade	01-Sep-07	Upgrade the CPU farm to full capacity, subject to funding

References

- [1] LHCb Collaboration, “LHCb Online System Technical Design Report”, CERN-LHCC-2001-040.
- [2] LHCb Collaboration, “LHCb Trigger System Technical Design Report”, CERN-LHCC-2003-031
- [3] A. Bay et al, “Common L1 read out board for LHCb specification”, LHCb 2003–007.
- [4] A. Barczyk et al. “1 MHz Readout”, LHCb 2005-062.
- [5] Force 10, <http://www.force10networks.com/>
- [6] CMS Collaboration, “Data Acquisition & High-Level Trigger Technical Design Report”, CERN-LHCC-2002-026