1 MHz Readout

LHCb Technical Note

Issue: Final
Revision: 1.0

Reference: LHCb 2005—62
Created: 9 March, 2005
Last modified: 7 September 2005

Prepared By: Artur Barczyk, Guido Haefeli, Richard Jacobsson, Beat Jost, and Niko Neufeld

Abstract

We present the design and possible implementation of a single-stage readout of the LHCb detector. This means the readout of the full detector at Level-0 accept rate of 1 MHz.
Conclusion is that technically the proposed scheme can be implemented.

Document Status Sheet

Table 1 Document Status Sheet

1. Document Title: 1 MHz Readout			
2. Document Reference Number: LHCb 2005—62			
3. Issue	**4. Revision**	**5. Date**	**6. Reason for change**
Final	1.0	7 September 2005	Initial release

Table of Contents

List of Figures

List of Tables

1 Introduction

In this note, we explore the possibility of upgrading the LHCb read-out to a system reading out the full detector at the maximum possible rate of 1 MHz. This has always been the natural, ultimate upgrade from a data-acquisition point-of-view. Originally, something anticipated for 2009 or 2010, current technology and price developments in the networking industry bring a date for such a system even 2007 within reach.
It is evident that such a system is not for free, but as will be seen from this document, the simplification brought about by a single-level read-out also leads to some cost-savings, so that the overall cost increase should not be unreasonably high.
A detailed costing depends of course on the required bandwidth, the dates of purchase and the selected equipment and is not included here, but can be in short time, once the boundary conditions are defined.
In the following we discuss the various constrains in the current baseline and discuss several scenarios for a 1 MHz readout. The important question of zero-suppression is also addressed. We finish the document by a review of the technical and operational advantages.

Disclaimer:
This note does not try to explore or evaluate the potential gains in physics performance of the experiment, which of course must be the ultimate judge of such an endeavour, resources permitting.

2 Constraints and Boundary Conditions

There are several constraints, which we will assume for the remainder of this paper. They will be listed in the following and we will quickly discuss how "hard" they are and what it would ensue should they be softened or lifted.

2.1 Output bandwidth of a TELL1[1]

Each TELL1 is equipped with a 4-channel Gigabit Ethernet card (GE-card). The maximum data throughput out of such a card is theoretically 4 Gigabit/s or roughly 480 MB/s (net). This is less than the SPI3 bus used between the SyncLink FPGA and the GE-card can handle, when it is operated in single-direction mode. Since the data are pushed in one direction only and we have no intention to change the protocol, it should possible to use a 10 Gigabit Ethernet card. Obviously such a card would have to be designed and tested first. Moreover, also currently available 10 Gigabit technology is based exclusively on fibre-optics. It is not yet clear when a copper 10 Gigabit physical layer will be available, which would allow to reuse the existing cabling. It should be noted that obviously not all TELL1s might need to be upgraded in this way, just the ones where the data-rate really goes beyond the 4 Gigabit.

2.2 Number of TELL1 boards

Currently there are 276 TELL1 boards in the experiment. Each is equipped with a Gigabit Ethernet Quad-port card. If processing power on the TELL1s is not sufficient to do the full zero-suppression, one can increase the number of TELL1 boards. Apart from the cost, the availability of components at a later time will be a big concern. A smaller inconvenience is that the cabling has been done only with the 276 currently foreseen TELL1s in mind, for full bandwidth, some additional cabling is necessary.

2.3 Cabled bandwidth into the event-filter-farm

Clearly a minor issue, it should not be forgotten that currently "only" 350 Gigabit links are available into the racks in D1. Clearly, with the very likely event of 10-Gigabit Ethernet over copper this is no problem[2]. Alternatively laying a few hundred more cables should not pose a significant problem.

---

[1] Throughout this note wherever a TELL1 board is mentioned implicitly also the UKL1 board is meant

[2] For the current status of 10GBase-T, i.e. 10 Gb Ethernet over UTP see e.g. http://www.npiconnect.com/techUpdate.htm.

2.4 Computing Power

In the first instance, one might be misled to the conclusion that reading out the whole detector at 1 MHz would imply an enormous increase in required computing power, because the current HL T algorithm would be applied to 25-times more events. This is however not what would be done in practice, where a, still-to-be defined, algorithm will reduce the incoming rate in the fast and efficient way the current Level-1 trigger and the primary stage of the HLT trigger do. Clearly, it is impossible to execute the HLT algorithm on the full 1 MHz event rate. There is still a sequence of algorithms necessary reducing the output rate to the originally foreseen rates. Still there will be a sort of algorithm selecting events with detached secondary vertices, like in the original Level-1 trigger. The main advantage of the scheme elaborated here lies in the fact that **ALL** information is available at any time with **FULL** precision and granularity. This should allow considerably more efficient signal/background separation. There is no major increase in CPU power necessary in the new scheme.
Some additional CPU will be needed to cope with additional amounts of data that have to be copied around and also, in case only partial zero-suppression is possible, to finalize the zero suppression in software. We believe that, with the maximum available space for 2200 1U boxes enough CPU power can be provided for all practical purposes.

3 System Design

3.1 System Architecture and Size

The architecture of the LHCb Online system for the 1 MHz readout, emphasizing the Dataflow subsystem, is shown in Figure 1.

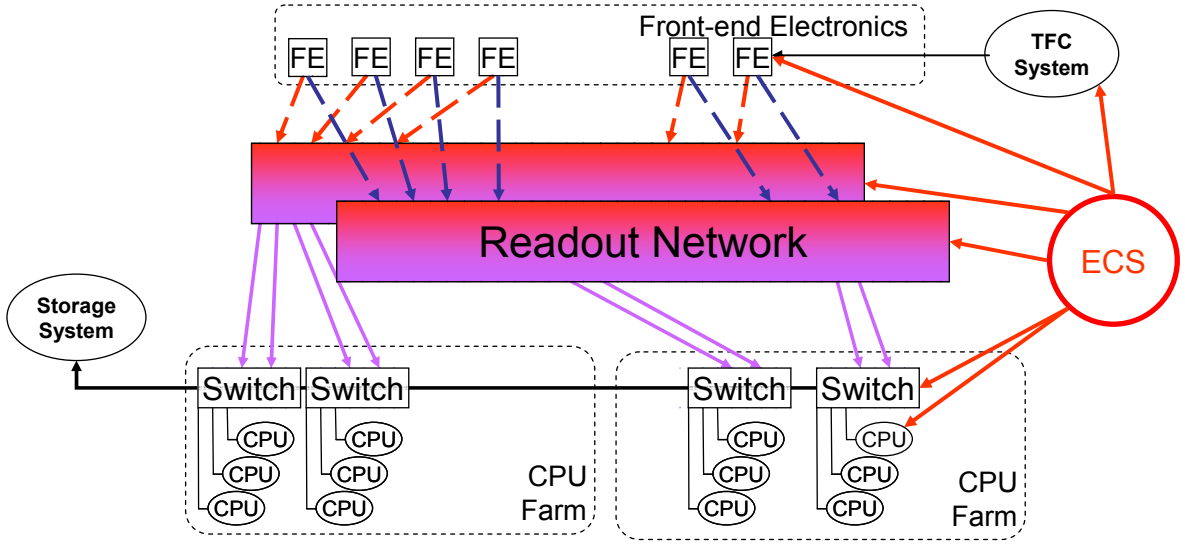

Figure 1 Resulting architecture of the LHCb online system after elimination of an explicit Level-1 trigger.

The system is noticeably simplified due to the elimination of the Level-1 trigger. As a consequence the decision sorter and the TRM module have disappeared, but also, and more important, there is only one data flow in the readout system. To cope with potentially very large bandwidth requirements through the system (beyond the capacity of one switch) we apply here the concept of decomposing the system in 'independent' subsystems as developed in CMS[3]. Every subsystem is independent and isolated from the other. The only point of contact is at the level of the TELL1 front-end electronics boards and at the Storage sub-system. All Tell1 boards send their MEP to the same CPU located on the same subsystem. Like this the MEPs are assembled on a CPU and subsequently processed. This scheme is possible since there is no need for two CPUs to communicate with eachother and hence the connectivity between CPUs is not necessary. Thus the overall bandwidth through the system is doubled and scaling perfectly.

The sizing of the system crucially depends on the data volume transferred. This is, of course, given

[3] See e.g. CMS "Data Acquisition & High-Level Trigger" Technical Design Report CERN/LHCC 2002-26, Chapter 3.

by the trigger rate and the data size per trigger. The former is fixed (per requirement) to 1 MHz. The latter depends very much on the zero-suppression capabilities of the front-end electronics and the occupancy of the detector, which is given by the physics processes at the LHC collision energy and the electronics noise of the detectors.
Currently the situation concerning zero suppression is as follows:

- The electromagnetic calorimeter cannot perform the "2D" zero suppression, originally foreseen at 40 kHz, at the rate of 1 MHz. The group studied the problem and came up with a solution applying (loss-less) data compression algorithms instead of zero suppression. This increases the average event size of the ECal by about 50%, but still represents a significant reduction with respect to a full (non-zero suppressed) readout.
- The other subdetectors do not see a problem executing the 'full' zero suppression at a rate of 1 MHz. There can be local problems with the output capabilities of the Tell1 board, which can easily be overcome by adding more boards.

Table 2 shows a summary of the size of the system

Table 2 Size of the system reading at 1 MHz

Subdetector	Event Size [Bytes]	Number of Tell1	Average MEP Size per board	Data Rate [MB/s]	Number of Links	Max. Link Load
Velo-r	5825	42	1867	6406	84	68%
Velo-Phi	4778	42	1543	5306	84	56%
RICH1	2938	13	3002	3176	39	65%
RICH2	2248	7	4239	2376	28	68%
TT	4133	48	1183	4650	68	80%
IT	3725	42	1217	4173	54	72%
OT	13997	48	3855	14876	192	62%
PS/SPD	1400	8	2339	1511	16	76%
Ecal	6700	26	3414	7176	78	74%
Hcal	1800	8	2989	1946	24	65%
Muon M1	907	4	3012	980	12	66%
Muon M2	797	4	2656	853	12	59%
Muon M3	190	2	1299	213	3	70%
Muon M4+M5	174	2	1197	193	3	71%
L0 PU	275	2	1852	303	4	61%
L0 Calo	533	2	3527	569	6	76%
L0 DU	128	1	1728	142	2	57%
L0 Muon	634	5	1712	699	9	68%
PUS	506	4	1708	561	8	58%
Readout Supervisor	64	1	896	73	1	59%

Table 2 was generated using estimates for the event sizes with the following assumptions

- Luminosity $5 \cdot 10^{32}/cm^2/s$
- 30% increase at generator level (Pythia)
- Increase of the event size by a factor of 2 for Muon station M1 and 3 for Muon stations M2-M5
- ECal contains still the ECal trigger information although this information can be injectively reconstructed from the compressed digitization

- HCal also uses data compression (like ECal), leading to an increase of the size by factor ~2.
- Maximum link load at input and output of the readout network limited to 80% and 85% respectively (including all overheads)

The safety factors applied for the data sizes are summarized in Table 3

Table 3 Summary of Safety Factors applied to determine the average event sizes. All data have been generated at luminosity of $5 \cdot 10^{32}/cm^2/s$

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				2.25	2.25
RICH	1.2				1.2
Muon1-2			2		
Muon3-5			3		
ECal					3
SPD/PRS	1.2				1.2

Ecal non-zero suppressed depends only weakly on multiplicity, safety factor 3 only causes <15% increase in size.

In Table 4 the basic parameters describing the size of the system are summarized. The average event size is now 52 kB. A total of ~1260 GbEthernet ports are needed to implement the system.

Table 4 Summary of the scale of the resulting system

Totals	
Number of Switches	1
Packing Factor	13
Ethernet MTU	1500
Total Event Size [Bytes]	51752
Number of Tell1	311
Total Input Data Rate [MB/s]	56183
Average MEP Size [Bytes}	2227
Frame Rate [Mf/s]	50.0
Switch Buffering Needs [kB]	730.4
Average Link Load	61.8%
Readout Network Input Ports	727
Readout Network Output Ports	530
Ports per Switch	1257
Input	727
Output	530
Total Number of Ports	1257

Higher port counts can be accommodated using a second switch and so implement the two-subsystem configuration as depicted in Figure 1. Beyond the real event size observed at LHCb, clearly, the availability of concrete hardware will dictate whether a second switch is needed. In Appendix B. we give the scale of the current baseline system for comparison.

3.2 Hardware Availability, To-Date.

At this level, we try to make give a plausible argumentation that this scenario is feasible. As an example of a supplier of equipment, we take Force10 and their Terascale technology.
The basic characteristics are

- 16 Slot Chassis, with 14 slots available for line cards
- 48 GbEthernet port line cards (wire speed)
- 90-GbEthernet port line cards (90/48 over-committed, today)
- 1.6 Tb/s (aggregated) switching fabric; backplane capacity ~5 Tb/s

If we mix inputs and outputs e.g. on the 90-port blade, we can take advantage of the bi-directional nature of the connection to the backplane and in this sense the 90-port blade becomes full-speed for our traffic pattern. In this mode, the available bandwidth today is 672 Gb/s or ~84 GB/s.

Comparing the availability of suitable switches and the number of ports needed from Table 2 and Table 4 it can be seen that one switch is sufficient for the size foreseen to be needed. Should the event size encountered at real collisions in the accelerator be significantly higher than foreseen, i.e. significantly higher than the safety margins as stated in Table 3, more Tell1 boards would be needed, at least in certain places that represent hot spots. This would lead to more input (and output) links and would eventually exceed the number of ports on the foreseen switch. Hence additional chassis will be needed to accommodate the total number of ports, especially since 90 ports per blade seems to represent some sort of practical maximum concerning the front-panel space for the I/O ports. Using two switches we can accommodate up-to 2500 ports, which we consider sufficient for all practical purposes.

3.3 Redesign of the event-filter farm

In the current baseline design the functionality of a subfarm controller is mandatory due to the limited latency of the Level-1 trigger. It would be impossible to maintain the limited latency foreseen if e.g. 25 triggers were sent to one CPU and queue there for processing. Currently an SFC can perform event-building at 2 GBit/s. Limiting their number to 100 means that even in the minimal scenario a SFC must handle 3.6 GBit/s in and out sustained. This seems formidable, though not out of reach with high-end servers. However it is not to be expected that the even higher rates needed for the other scenarios can be handled anytime soon by reasonably priced hardware. It is thus more scalable to eliminate the SFC altogether and let a lightweight event-builder task run on each node. This event-building is of course considerably simpler than today, because there is only one stream and no latency restriction. The CPU cycles needed by the event-builder task must be bought of course in equivalent farm nodes. First measurements on prototype software show that a maximum of 20% of one CPU per box is needed to perform the event-building function. This, assuming 900 boxes (1800 CPUs), necessitates the acquisition of 150 additional boxes. The cost for these would be compensated by the elimination of the SFCs.

3.4 Additional Cabling Needs

As mentioned in section 2.3 there are 350 Cat6 cables installed to the CPU farm. According to Table 4 there are a minimum of ~550 links needed to implement the system. Hence at least 200 more cables are needed (to cope with a small safety factor, prob. 250 more cables will be installed). Unfortunately the 350 existing cables were installed assuming the existence of SFCs and thus originate not at the readout network, but rather at the (foreseen) location of the SFCs. Hence the 350 existing links will need to be rerouted.

3.5 Implementation Strategy

The uncertainties connected to the unknown multiplicities and event sizes at the LHC energy demands a flexible implementation strategy and a scalable architecture. The scalability is demonstrated in previous sections, however the associated cost to scale up is significant and hence there has to be a balance between 'being on the safe side' and the cost of the acquisition of the material. We propose the following scenario

- Upgrade the cabling in the barracks to cope above outlined scenario
- Installation of a system of the size as outlined in Table 4
- Commissioning of that system in the pilot run of the LHC and determining the real working point (event-size, etc.). Throttling the trigger will ensure that the rate will be stabilized to a level supported by the installed capacity.
- Upgrade of the system (if necessary) to match the real event size in the shut-down after the pilot run. Since the (long distance) cabling is already installed it's only necessary to acquire the necessary (switching) hardware. This can be achieved within ~2 months and assuming one month for cabling and commissioning the entire upgrade can be achieved within 3 months.

4 Benefits

4.1 Technical

The elimination of the Level-1 trigger clearly signifies a simplification of the DAQ system. It therefore has a beneficial effect on the overall system reliability. Obviously none of the hardware components specific to the Level-1 trigger, such as decision sorter and TRM module, are necessary anymore. This reduces the maintenance effort for the system.
The elimination of the SFCs leads to a harmonization of the hardware base and also leads to a slight cost reduction.
Currently it seems that industry's tendency in terms of processor development does not go in the direction of increasing clock speeds but rather keeping the clock speed of the processor cores constant and increase the overall performance by the introduction of multi-core architectures. This tendency does not favor the current LHCb trigger strategy, since the Level-1 trigger imposes a limited latency. Hence a 'slow' processor cannot be compensated by more processors because more processors will not reduce the latency. Multi-core architectures, however, are much more suited for high event rates without latency limitations. Therefore, the 1MHz-Readout actually seems to be more adequate to the future processor architectures.

4.2 Physics

We have not studied the benefits of the proposed scheme in term of physics performance. It is however plausible that more information available at an earlier stage can only have a positive effect on the physics performance. In addition, the complete freedom of algorithm design will lead gradually to better and better performance, especially with the advent of more and more powerful processors in multi-core technologies.

5 Cost

Table 5 summarizes the cost of the system.

Table 5 Cost for the 1 MHz system, its upgrade and the current baseline system for comparison. All numbers in kCHF

	1 MHz					Baseline		
	2-Switch Startup			2-Switch Upgrade	Comments			
	Quantity	Unit Cost	Cost			Quantity	Unit Cost	Cost
Router (Chassis + Linecards)	1	387.6	**387.6**	206.7		1.0	206.7	206.7
Add'l Cabling			**49.7**					
Add Tell1 Cables	88	0.090	7.9					
Add'l Patch Panels for Tell1	6	0.360	2.2					
Addt'l cables D2-D1 (250)	250	0.085	21.3					
Addt'l patch panels	12	0.360	4.3					
Recabling D2	350	0.040	14.0					
Add'l Tell1	31	4.550	141.1					
Add'l Crates	3	7.500	22.5					
SF Switches	50	3.000	150.0		Limit of CPU farm to 1800 Boxes	50.0	3.0	**150.0**
"SFC"	150	1.800	270.0			100.0	3.5	**350.0**
Total			**1020.8**	**206.7**				**706.7**

It is assumed that the initial 1 MHz system will consist of a one-switch solution and (maybe) be upgraded to two switches. It should be noted that the additional cost for the DAQ system proper is 'only' ~150 kCHF and the additional ~150 kCHF originate in additional Tell1 boards and crates. The figure for the upgrade presented in Table 5 only includes the cost for a minimal upgrade of the system to two switches, not including an increase effective bandwidth (additional linecards).

Appendix A. Destination Assignment Strategy in the TFC system

The destination assignment[4] in the proposed 1 MHz scheme, as was foreseen in the baseline system, will be done by the TFC system. One consequence of the transition to the 1 MHz readout and the elimination of the SFCs is that the addresses of the individual CPUs are exhibited to the TFC system. This is not a problem a priori, since there is enough space available in the Readout Supervisor to handle tables of that size[5].

In the baseline design, however, it was foreseen to implement static loads balancing between subfarms and dynamic load balancing inside a subfarm. The latter is particularly necessary for Level-1 data, because of the finite latency.

In the new scheme we foresee actually that the CPUs send tokens to the Readout Supervisor via the Readout network to request a new Multi-Event. This would implement a perfect dynamic load balancing, since if a CPU is busy it will simply not issue a token. The request tokens are, at least at the network level, the equivalent of the Multi-decision packets that were sent to the Level-1 decision sorter and hence don't represent an additional network load.

It is also foreseen that the Readout Supervisor issues a TTC broadcast for each Level-0 accept to the Front-End electronics in order to inform the logic on the Tell1 of the nature of the trigger. This allows the Tell1 to execute different algorithms for different types of triggers, specifically calibration triggers. In fact, it's foreseen that calibration triggers would be routed (directed) by the Readout Supervisor to a different set of CPUs. This would decouple the software handling the calibration triggers from physics triggers and thus eases the software management.

[4] Destination assignment means telling the Tell1 boards where the next MEP should be addressed to.

[5] The maximum number of CPUs in the farm is 2200 1U boxes.

Appendix B. Scale of the LHCb Baseline Readout System

For completeness we give in this appendix the size of the current baseline system. Table 6 shows the decomposition of the system by subdetector for both dataflows (Level-1 and HLT).

Table 6 Scale of the baseline system. Both dataflows are shown.

Subdetector	Event Size [Bytes]	Number of Tell1	Average MEP Size per board	Data Rate [MB/s]	Number of Links	Max. Link Load
Level-1 Trigger						
Velo-r L1	1431	42	964	1717	42	35%
Velo-Phi L1	1342	42	911	1627	42	33%
TT L1	1125	48	698	1452	48	41%
L0 Trigger L1	90	3	861	110	3	38%
TRM	24	1	712	31	1	25%
HLT						
Velo-r	5825	42	1439	257	42	6%
Velo-Phi	4778	42	1190	210	42	4%
RICH	5187	14	3757	220	14	13%
TT	4133	48	913	189	48	6%
IT	3725	42	939	170	42	6%
OT	13997	48	2968	603	48	10%
PS/SPD	1400	8	1802	61	8	6%
Ecal	6700	10	6752	282	10	23%
Hcal	1800	4	4552	77	4	15%
Muon M1	907	2	4587	39	2	15%
Muon M2	797	2	4039	34	2	13%
Muon M3	190	2	1002	8	2	3%
Muon M4+M5	174	2	923	8	2	3%
PUS	506	4	1316	22	4	5%
L0 PU	275	1	2802	12	1	9%
L0 Calo	533	1	5380	22	1	18%
L0 DU	128	1	1332	6	1	4%
L0 Muon	634	5	1320	28	5	5%
Readout Supervisor	64	1	692	3	1	2%

In the following table, we summarize the system's parameters.

Table 7 Summary of the basic parameters of the system

Totals			
	General	**HLT**	**Level-1**
Total Event Size [Bytes]		51752	4012
Number of Tell1	280		
Total Input Data Rate [MB/s]	7187	2250	4937
Readout Network Input Ports	415	279	136
Readout Network Output Ports	100		
Total Number of Ports	515	279	136
Packing Factor		10	25
Ethernet MTU	1500		

Appendix C. Glossary of Terms

SPI3 — a parallel bus at 133 MHz with a width of 64 bits. This bus is used to access the GbEthernet MAC on the standard GbEthernet card

GE-card — a 4-channel Gigabit Ethernet 1000 BaseT (copper) interface used by the TELL1 boards to send data to the DAQ network

TELL1 — The common LHCb Front-End electronics board which receives the data from the on-detector electronics and provides the buffering and formatting capabilities for the LHCb data flow.

UKL1 — the analogon to the TELL1 for the RICH detector

kB, GB, TB — kilobyte (1024 bytes), Gigabyte (~ 1 x 10^9 bytes), Terabyte (1 x 10x^{12} bytes)

Gb, Tb — Gigabit (1 x 10^9 bits), Terabit (1 x 10^{12} bits). These are pure powers of ten contrary to kB, GB, TB which are defined as powers of 2.

SFC — Subfarm Controller. A CPU that performs in the original LHCb DAQ system the functionality of event-building and distributing the events to the individual nodes of the CPU farm.