A 32 Terabit/s Data Acquisition from Mostly COTS Components

Rainer Schwemmer and Niko Neufeld

Abstract—The Large Hadron Collider beauty (LHCb) data acquisition after 2019 will need to perform event-building at an aggregated band-width of 32 Tbit/s. Apart from the technological challenges described in various papers also at this conference, the key challenge is to come up with an architecture which minimises the cost, while providing a system which can be maintained by a small team for a long time and which scales well. In this paper we present the analyses we have been doing to minimise the cost, the R&D topics we derived from that and how we combined all this into a coherent proposal which allows us to come up with a system which not only today fits the budgetary constraints of LHCb, but also will allow profiting from any main-stream technological development. We achieve this by aligning our system needs as much as possible to data-centre mass-market commercial of the shelf (COTS) products; by minimising the number of optical interconnects and by optimising the physical layout of the system. This system requires only one piece of custom-made hardware, and even this could, for a smaller setup be replaced by a commercially available item. We believe that the reasoning behind this design can be beneficial to any large, high-rate data acquisition system.

Index Terms—Data acquisition, multiprocessor interconnection networks, network topology, next generation networking, real-time systems.

I. INTRODUCTION

THE LHCb [1] collaboration has decided to remove the hardware based first level trigger completely from its data acquisition (DAQ) for the third run of the LHC, starting in 2019/2020. The current trigger is using only the information from a small fraction of the detector. While performant, given the circumstances, this trigger is very inefficient due to its limited input data. A good example here are hadronic events which can be selected more efficiently when particle track parameters are available. Currently LHCb uses only calorimeter and muon signals for the trigger decision, which misses most of these decays [2]. To overcome these inefficiencies it is necessary to look at the detector in its completeness, which means the triggering has to happen after the data has passed through the event-building process. In order to cope with the expected data rates we have done extensive studies in network architectures, upcoming network technologies, computer I/O capabilities and

pricing to come up with a solution which is scalable to O(Tbit/s) read-out bandwidths and at the same time cost effective [3].

Over the past decade the I/O capability of modern computers has reached Tbit/s speeds. Applications that used to be in the realm of FPGAs and ASICs are nowadays within the grasp of common CPUs. A good example here is the mass storage market. Ten years ago the state of the art, high performance storage systems were mostly based on FPGAs and custom ASICs. Today these architectures have all but died out. This is true for most cases where complex algorithms are needed and FPGAs were used solely because of their latency or I/O capability advantage. Most of these systems use common servers with CPUs and Unix based operating systems under the hood today.

With the advent of PCIe Gen3 and the Intel Ivy Bridge architecture, a single, dual-socket server with 80 PCIe Gen3 lanes can reach a total I/O bandwidth of 640 Gbit/s in- and 640 Gbit/s output for a total bidirectional bandwidth of approximately 1300 Gbit/s. As a comparison, the current DAQ of the LHCb experiment has a total bandwidth of approximately 400 Gbit/s [5] and could be streamed through a single server.

Another cornerstone of large scale DAQ systems, the read-out network has recently also seen a substantial boost in performance. 100 Gbit/s capable network cards are available today on InfiniBand (IB), and 100 Gbit/s Ethernet is being worked on.

This explosion of throughput capability has made it possible, in fact necessary, to abandon the classical, expensive, crate based read-out board solutions and bring the data into a computer one step earlier in the read-out chain.

II. THE HIDDEN COST OF LARGE SCALE CRATE BASED READ-OUT

On the scale of several million input channels it is almost impossible to use only local information for the filtering of interesting data sets. At some point the detector state has to be brought together in one computer and treated as either sufficiently large regions of interest or as a whole. This filtering computer will typically be busy for some time, processing a single detector time slice and therefore many computers are necessary to keep up with the high interaction rates in modern accelerators. These filtering units (FU) are normally limited by CPU resources instead of I/O capability and can usually be fed via low speed I/O interfaces.

Since the entire detector read-out does not fit into a single crate with a cross connecting back-plane, some sort of network has to be deployed to connect the individual source units to a set of event building units (BU) which then feed the filtering

Manuscript received June 16, 2014; revised January 23, 2015; accepted February 26, 2015. Date of publication July 09, 2015; date of current version August 14, 2015.

The authors are with the PH Department, CERN, 1211 Geneva 23, Switzerland (e-mail: rainer.schwemmer@cern.ch).

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.2015.2435902

units. These networks can be built with FPGAs or self made ASICs and if part of the filtering or event building can be done on these dedicated network processors this might be economic. However, due to economies of scale it is very hard today to compete with commercial network equipment and usually some local area network (LAN) technology is deployed for interconnecting the sources with the filtering units.

The cost of a crate based read-out is not only the price of the hardware and infrastructure. A crate based read-out attached to a local area network either uses expensive, single-board computers which connect to the back plane of the crate or FPGA-based boards, which transforms the detector data into an industry standard network format. Simple event-building protocols are preferred to reduce overall complexity and FPGA resource usage. As a consequence data is often put onto the network without ensuring its arrival at the destination.

This simple push mode usually necessitates large, very expensive buffers in the read-out network, which is one of the main cost factors at this scale. These costs can be mitigated by adding buffer memory to the FPGA boards and adding flow control and traffic shaping algorithms to the firmware. However, this increases the complexity of the firmware with more complicated protocols and memory interface code which would use large portions of the FPGA's resources [6].

Another aspect is the network interface itself. This can either be solved by instantiating a network core on the FPGA or by adding an interface card to the board. Today this means Ethernet only, which is not necessarily the most cost efficient solution anymore at very high bandwidths. Also one might want to use precious logic cells for something more important than network code.

Yet another disadvantage is actually the high density of these solutions. While small, optical links are coming down in prices, copper cable-assemblies are still cheaper if power consumption is not a concern and distances can be kept short. These cable assemblies require significant front-panel space and sometimes additional ASICs. In custom boards this drives up the cost of the solution, since one cannot easily exploit the economy of scale, which drives the COTS market.

III. A PCIe BASED READ-OUT

To overcome these disadvantages, LHCb, similar to other LHC experiments [7] [8], is currently developing an FPGA based PCIe card which connects directly to the detector.

Additionally the read-out, controls and fast control are currently all performed by different, custom made hardware. The slow control of the detector is performed by the copper based Serial Protocol for the Experiment Control System (SPECS) [9]. The fast control of the experiment is done by the optical Timing and Fast Control (TFC) [10] system. For the upgrade we will use the Versatile Link with the GigaBit Transceiver (GBT) protocol developed by CERN [11]. Since all three tasks will use the same physical link we can use the same board for them. This is a first step in reducing the complexity and removing a lot of custom hardware in the system.

The card uses high density optical connectors on the front plate to connect the detector. It uses a built-in PCIe hard core, which comes with most modern FPGAs for negligible cost and

Fig. 1. Schematic of the PCIe card currently in development for the upgraded LHCb DAQ system. The card utilises modular, pluggable optical receivers and transceivers. Depending on the use case of DAQ, Slow- or Fast-control, the card can be equipped with a variable number of receivers and transmitters. Since FPGAs are currently only offering two 8 lane PCIe cores instead of one 16 lane core, we are planning to use a bridge chip to merge the two 8 lane interfaces into one.

with very low foot print in terms of logical cell usage. The card will be Gen3 and 16 lanes wide. It will be capable of sustained 100 Gbit/s throughput for DAQ and controls purposes.

Since this card plugs into a server, the expensive buffer task can be moved from the network to the server, where memory is extremely cheap and plentiful. The CPU also allows the implementation of more complex event-building algorithms, which can make the read-out more robust and are easier to develop and maintain¹.

Once the data is inside the server, the choice of network technology is only limited by what is available. It allows the choice to be made at a much later time, which means more bandwidth per price unit and a safer, more future proof choice. This is one of the major advantages of this solution. Typically the R&D phase of a large scale experiment is of the order of five years.

The read-out board has to be there typically three years before the experiment goes online to allow proper testing of the board and prototype detector front-ends. The network market is moving at a much faster pace though and one is almost guaranteed to end up with obsolete hardware from the very start. PCI has a long history of backward compatibility and is a much more stable interface.

IV. CURRENT R&D PROJECTS

Our current R&D is focusing on two major subjects. The first major topic is the development of the card and the drivers for it. Writing a high speed I/O driver is not a simple task and surprisingly little knowledge and resources are currently available within the high energy physics and the open source software community. We are currently using Altera Stratix V based PCIe development cards for developing the firmware and drivers. We have completed a first version of an 8 lane interface which works stable and can sustain throughput at more than 50 Gbit/s. Our next goal is to enhance this interface to a 16 lane version. This version uses two 8 lane interfaces as a base and a PLX PCIe bridge to merge them into a single 16 lane stream. A schematic of the planned interface card can be seen in Fig. 1.

The second major topic is the change of the classic read-out network topology and event-building protocols. Since the server

¹C/C++ Code instead of a Hardware Description Language

Fig. 2. Schematic of the old baseline network topology. Data flows in a unidirectional way through a core router based network.

the card plugs into has plenty of computing capability, we are trying to move the usually distributed event-building task into the servers that house the read-out board. This compaction allows us to do the event-building within a very small, dedicated, high bandwidth network, illustrated in Fig. 3. This in turn allows the use of short cables which can be copper instead of optical.

We have already shown the feasibility of doing a 100 Gbit/s event building within a single machine [12]. The studies were performed on an InfiniBand network using two servers. One server was running the event building software under test, while the other server provided simulated data traffic from all the other nodes in the event building network. A GPU was used to simulate the data coming from a local read-out board on server one. Our current focus of R&D in this area is a test of a larger network. We are currently scaling up the software used in the two machine test to run on bigger clusters. The goal will be to test at least one full switch with typically 48 ports as a first step. We are also negotiating for access to super computing sites to do a test with more than 100 network nodes.

Another side effect of this change in network architecture is the separation of the event-building and the event filter network. Since the filtering of events will be CPU bound, high speed I/O will not be necessary in the filter network. This opens the door for a cheaper/slower solution, where only a limited number of high speed links are necessary as uplink to the event-building network. An additional chance for cost reduction is the fan-out stage to the farm. To have some margin for fluctuations and safety factors for yet undetermined detector occupancies, most detector links will actually run at far less than the maximum bandwidth of 100 Gbit/s. This means we can save on output links from the event building servers to the farm by doing the building only on a subset of the servers. This subset will run close to 100 Gbit/s while the other building servers will only relay their data to them via the building network. We can then connect only these active servers to the farm and roll out high

Fig. 3. Network architecture based on PCIe cards. The read-out network has been divided into event-building and filter network. The read-out board sits inside the combined Read-out/Builder unit. Events are built via the top network and then passed to the Filter Units in the bottom network.

speed uplinks according to the actual bandwidth needs of the detector.

V. COST CONSIDERATIONS

We would like to show some price comparisons of the different architectures we investigated. Since many of the prices we use here are based on confidential quotes from the hardware manufacturers we have to obfuscate the actual prices with arbitrary currency units and can not show a bill of materials.

We are showing two example architectures on the extreme ends of the price scale. The first is a scale-up of the current, 1 Gbit/s Ethernet read-out network to a 40 Gbit/s Ethernet based network using ATCA based read-out boards. In the past this used to be the baseline architecture for the upgrade in 2019/20.

The second example is a network based on 56 Gbit/s InfiniBand technology using the separated network topology described earlier and PCIe based read-out boards. It is the current baseline for the upgrade.

A. Ethernet/ATCA

Fig. 2 illustrates the topology of this network. The read-out units are AMC based, custom built boards which connect via 40GBASE-SR4 optical links to the core network.

The boards have no buffering and are sending their data as soon as it is available and with high synchronicity to the designated combined event building and filtering node on the other side of the core network. Since the boards have no buffers they are sending a large amount of small messages on the order of kilobytes to the network. To reduce FPGA resource consumption there is typically also no space for flow control mechanisms so the packets are sent in the hopes that there is always enough buffer space available on the next hop in the network.

This kind of data flow is the prototype example of how to cause network congestion and packet loss: the synchronised sources are sending packets to the same target within a short

Fig. 4. Overall event building network costs for crate based read-out vs. PCIe based read-out.

time window. This kind of traffic can only be handled by expensive, large scale telecom level routers.

Moreover, each link is only used in one direction, wasting half of the potential bandwidth. These are expensive, optical links which have to run both ways to be compatible with Ethernet. This adds a significant contribution to the overall price even though they are not used at all.

B. InfiniBand/PCIe

InfiniBand is a high bandwidth, low latency network which is mostly used in High Performance Computing. The low latency is achieved by using cut-through instead of store-and-forward switching and a highly sophisticated offload architecture, which runs almost the entire protocol stack in hardware on the network interface cards. This technique needs almost no buffering in the network switches.

At first glance this kind of network is the complete opposite of what is needed in large scale DAQ. The network has almost no buffers and the protocol stack is too sophisticated and complicated to be run on an FPGA. It has however excellent flow control mechanisms based on a credit system which prevents unreliable transmission like in Ethernet. On top of that, InfiniBand is significantly cheaper per port per Gbit/s than its Ethernet counterparts and there are already interface cards capable of 100 Gbit/s I/O [13]. Due to the complexity of the protocol, it is currently not economical to use an FPGA for talking to the InfiniBand interface ASICs.

The topology of the network is depicted in Fig. 1. The expensive optical links in case of the AMC board have been replaced by the PCIe copper back-plane link on the server mainboard. The read-out board now uses the hard implementation of PCIe available on current high-end FPGAs which frees logical units on the FPGAs for other uses.

The event building is taking place on the read-out units themselves, which are interconnected with a 100 Gbit/s InfiniBand based event building network. This combination of read-out and

Fig. 5. Breakdown of the network cost item in Fig. 4.

event building also makes sure that the high speed links are used in both directions.

Since the read-out units are now also CPU servers, there is no shortage of buffer memory. This buffering allows for a more sophisticated event building which can avoid congestion in the network in conjunction with the networks flow control mechanisms. The builder unit will typically accumulate $O(10.000)$ event fragments, which span several Megabytes. A distributed program running in parallel on the builder units assigns this event range block to a particular builder node and all other builders send their data block to that node. Since the messages are large the flow control of the network has time to regulate these message streams which prevents congestion. Once events are fully assembled, they are sent from the building/read-out servers to the filter farm via a second, one-directional link. This second link connects to a fast uplink port in the fan-out network which then distributes the data to the filter units.

This architecture is of course not limited to InfiniBand, as the AMC solution was to Ethernet. It can use any interconnect which works on industry standard computers.

C. Comparison

Fig. 4 shows a cost comparison of the two network topologies, including the price for the additional CPU servers for the combined read-out, builder units. Replacing the telecom level core routers with IB switches has reduced the network cost dramatically. Even though we now need to buy additional CPUs to host the PCIe cards, the second topology is only half as expensive.

Also the move from AMC/ATCA to PCIe has brought down the price for the read-out boards significantly. This is partly due to the lower speed grade of the necessary FPGAs and less optical transmitters on the board. Another price reducing factor is the removal of an intermediate board between the AMC boards and

the crate back plane. This intermediate board would have been responsible for distributing the clock and fast control signals to the AMC read-out boards. Instead clock and fast control are now distributed via optical links to the single SFP+ connector on the PCIe board and driven by four additional PCIe cards.

Fig. 5 shows a breakdown of only the network costs. As pointed out earlier, the major cost factor here are the network switches. However we are also saving a lot of money on not using optical links. Since we need network cards now on the read-out/building units, the cost there increases a bit, but the overall cost is still less than half.

VI. CONCLUSION

We have analysed major cost factors in the current LHCb DAQ system and what they mean for the future. By replacing classical, crate based read-out solutions with PCIe we keep our options open for future network technologies and prevent an early lock-in to essentially Ethernet.

By utilising the combination of Versatile Link and GBT we are able to unify controls and read-out into a common link which allows us to use the same hardware and reduce complexity and cost.

Furthermore, the utilisation of PCIe allows us to move data into computers much earlier in the DAQ process. We can now use the virtually limitless memory and the CPU power of a modern day server to implement more sophisticated event building protocols that allow the usage of cheap, non-buffering switches. The computer based read-out also allows us to chose the physical transport (copper or optical) closer to the deployment stage of the system.

We have shown that a 100 Gbit/s event building and read-out are already possible today and are well on our way of having a larger scale demonstration system soon.

REFERENCES

- [1] The LHCb Collaboration et al., The LHCb Detector at the LHC, 2008, DOI: 10.1088/1748-0221/3/08/S08005.
- [2] LHCb Trigger and Online Upgrade Technical Design Report, CERN-LHCC-2014-016; LHCb-TDR-016 [Online]. Available: <https://cds.cern.ch/record/1701361>
- [3] The LHCb Collaboration et al., Framework TDR for the LHCb Upgrade: Technical Design Report, CERN-LHCC-2012-007 LHCb-TDR-12, 2012 [Online]. Available: <http://cds.cern.ch/record/1443882/files/LHCb-TDR-012.pdf>
- [4] Intel Corporation, Intel Xeon Processor E5 v2 Family: Datasheet 2014, vol. 1 [Online]. Available: <http://http://www.intel.com/content/www/us/en/processors/xeon/xeon-e5-v2-datasheet-vol-1.html>
- [5] N. Neufeld et al., "Improvements in the LHCb DAQ," in *Proc. Real Time Conf.*, 2014.
- [6] M. Bellato et al., "A PCIe Gen3 based readout for the LHCb upgrade," *J. Phys., Conf. Ser.*, vol. 513, p. 012023, 2014, DOI: 10.1088/1742-6596/513/1/012023.
- [7] H. Engel and U. Kebschull, "Common read-out receiver card for ALICE Run2," *J. Instrum.*, vol. 8, C12016, DOI: 10.1088/1748-0221/8/12/C12016.
- [8] R. Cranfield et al., "The ATLAS ROBIN," *J. Instrum.*, vol. 3, T01002, 2008, DOI: 10.1088/1748-0221/3/01/T01002.
- [9] D. Breton and D. Charlet, SPECS: The Serial Protocol for the Experiment Control System of LHCb, CERN-LHCb 2003-004, 2003.
- [10] R. Jacobsson et al., The LHCb Timing and Fast Control System, CERN-LHCb 2001-16, 2000.
- [11] P. Moreira et al., "The GBT project," in *Proc. Topical Workshop Electronics for Particle Physics*, Paris, France, Sep. 21–25, 2009, pp. 342–346, DOI: 10.5170/CERN-2009-006.342.
- [12] D. H. Campora Perez, R. Schwemmer, and N. Neufeld, "Protocol-Independent Event Building Evaluator for the LHCb DAQ System," in *Proc. Real Time Conf.*, 2014.
- [13] Connect-IB: Architecture for Scalable High Performance Computing, Mellanox Technologies [Online]. Available: http://www.mellanox.com/related-docs/applications/SB_Connect-IB.pdf