The data acquisition and reduction challenge at the Large Hadron Collider

BY SERGIO CITTOLIN*

CERN/PH div., CH-1211, Geneva 23, Switzerland

The Large Hadron Collider detectors are technological marvels—which resemble, in functionality, three-dimensional digital cameras with 100 Mpixels—capable of observing proton–proton (pp) collisions at the crossing rate of 40 MHz. Data handling limitations at the recording end imply the selection of only one pp event out of each $10^5$. The readout and processing of this huge amount of information, along with the selection of the best approximately 200 events every second, is carried out by a trigger and data acquisition system, supplemented by a sophisticated control and monitor system. This paper presents an overview of the challenges that the development of these systems has presented over the past 15 years. It concludes with a short historical perspective, some lessons learnt and a few thoughts on the future.

Keywords: trigger; data acquisition; high-level trigger; Large Hadron Collider; Compact Muon Solenoid; A Large Ion Collider Experiment

1. Introduction

From the earliest moments of one’s involvement in the design and construction of the Large Hadron Collider (LHC) experiments, it was clear that the development of every single element was going to be a challenge that would require one or more major breakthroughs with respect to the then-known experimental techniques and technology. One such challenge was the absorption of the huge amount of information that was to be generated by the experiments as they registered proton–proton (pp) collisions at the LHC.

In the early 1990s, when the first seeds of the experiments were being put in place, all were in awe of the requirements. The beams would cross each other 40 million times per second, leaving a mere 25 ns between individual crossings; the intensity of those beams would be more than two orders of magnitude higher than what had been achieved by then, and this, in turn, would translate into not just one but roughly 20 pp interactions with every crossing; the detectors themselves were growing, at least in their active elements, and talk of 100 million active channels was common. Even after breaking down the problem into a

*sergio.cittolin@cern.ch

One contribution of 15 to a Discussion Meeting Issue ‘Physics at the high-energy frontier: the Large Hadron Collider project’.
number of modular subsystems, using multiple triggering stages (the ‘trigger levels’ introduced in the following section), the resulting systems far exceeded those of the previous generation of experiments at a large collider—e.g. at the Fermi National Accelerator Laboratory Tevatron. Figure 1 is a graph of the maximum trigger rate and event size for the experiments of the past three decades. The outermost points in this graph are clearly given by the four LHC experiments, with the general-purpose experiments (Compact Muon Solenoid (CMS) and A Toroidal LHC Apparatus (ATLAS)) pushing both the rate and size frontier, and the two specialized LHC experiments pushing the rate (Large Hadron Collider beauty experiment; LHCb) and event size (A Large Ion Collider Experiment; ALICE) frontier, respectively.

All this had to be done subject to some new unusual constraints: the experiments (and thus the systems that would read them out, select the best interactions and store them for later processing) would have to operate more than a decade later (back in the 1990s, the target was the early 2000s)—rendering discussions of what technology would do by then closer to the art of reading a crystal ball than to the discipline of system design. Another interesting constraint came from budget limitations. While at the time, the problem could simply not be solved with any of the available technology, the cost was fairly fixed. What made planning quite difficult was that there were two extrapolations into the
future: one involved technology (everything from the size of memories to the speed of microprocessors was extrapolated using the usual exponential power laws); budgets, on the other hand, used a linear scale, and changes in the estimated cost of more than 20–30% were seen as ‘wild fluctuations’.

These constraints set the scene for what was going to be another long, painstaking, but also equally rewarding, development in the context of experimentation at the LHC, namely the readout of the detector information and the selection, in real time, of a very small fraction of the pp collisions for storage and offline processing. This paper is split into two sections. First, a broad overview of the solutions found to the problem of online selection and data handling will be given. A necessarily short summary of the highlights of the various ‘subsystems’ follows. Some first performance results from the actual data-taking runs in 2009 and 2010 complete the presentation of the work carried out. The second part of the paper will attempt a review of the basic parameters and decisions that led to the development of these complex systems over the course of the past 20 years.

2. The concept and the overall architecture

Figure 2 is a schematic of the function of the full trigger and data acquisition (DAQ) system and its key parameters. In brief, at the highest design luminosities of the LHC, and a pp crossing interval of 25 ns, there is a GHz of pp interactions occurring inside the detectors. Storing and processing the large amount of data from these ‘events’ is simply impossible, thus necessitating a drastic rate reduction of a factor approximately $10^6$ in the number of events selected for later processing. The selection is carried out by the trigger system, which performs the selection
in a set of successive approximations, usually referred to as ‘trigger levels’, with the uninspired terms of ‘Level-1’, ‘Level-2’ and so on used for the first decision, second decision and so on.

The Level-1 trigger has to operate under the tightest time constraint, with only about a few microseconds available for its decision-making processing. It is implemented using custom-designed electronics. The subsequent filtering decisions, collectively referred to as ‘high-level trigger’ (HLT), are software systems running in a large farm of processors. The key parameters that determine the functionality of this system are the latency and the fraction of events selected by each processing step. Once these two parameters are defined for each step in the selection, the corresponding buffering requirement for that step, for the system to operate in an essentially dead-timeless way, is fixed.

Even if a slightly modified version of the offline reconstruction is applied, leading to a shorter overall processing time per event, the latency per event is roughly 1s. This relatively long latency associated with the final selection step implies that if the full output of the Level-1 step is fed directly to the final selection, a very significant amount of buffering on input, a huge bandwidth connecting the detector readout with the online farm and a significant amount of the processing time must be made available online. The numbers are sobering: at a Level-1 output of 100kHz, and with a mean event size of approximately 1 MB, feeding all the data into the HLT requires a network capable of providing a total sustained throughput of approximately 1 Tbs$^{-1}$ ($10^{12}$ bit s$^{-1}$), and adequate buffering for about $10^5$ events. At the time of the conceptual designs, the corresponding processing power for the online HLT processor farm was estimated at roughly 50 000 CPUs. An alternative to developing a system that would provide this unprecedented performance was the introduction of an intermediate selection step, akin to the Level-1 trigger, the ‘Level-2’ trigger, whose task would be to reduce the rate of events accepted by the Level-1 system by some factor of approximately 100. This factor would result in significant savings in the amount of networking, processing and buffering needed for the final selection step (which—in this scheme—is called the ‘Level-3’ trigger). The savings, of course, do not come for free: the Level-2 trigger is a complex system: it has to act on intermediate time scales (those of a millisecond), providing some of the functionality of the final Level-3 step, but based, again, on selected information from the various sub-detectors. Furthermore, to provide a small selection rate, it must use information from multiple detectors, thus necessitating the introduction of another network with significant high-connectivity requirements, as well as a complex data router that forwards different parts of the detector information based on the type of trigger provided by the Level-1 system for each event.

These two alternatives, shown in figure 2, constitute the two opposite ends of the possible architectures of a system that can cope with the demands of a general-purpose experiment at the LHC and represent different philosophies or ‘extent of belief’ in the evolution of networking and computing technologies. In the first case, i.e. that of a single HLT farm that is linked to the detector readout buffers, the aggressive expectation at the time of the conceptual design was the firm belief that those technologies would continue their exponential growth though the following two decades. In the second case, the extrapolations into what the commercial world would supply were less aggressive, albeit at the cost
of an increased in-house development. Fast-forwarding into the present time, both architectures have been designed in detail, implemented and commissioned with success [1,2].

Once the trigger architecture is defined, what remains is the design of a system that supplies the connectivity between the data sources, which in this case are the detector readout buffers, and the ultimate destinations, i.e. the processors that provide the final event selection. The process is displayed graphically in figure 2. Briefly, upon the reception of a Level-1 trigger accept signal, a number of buffers, each connected to several detector front-end channels, stores the data belonging to the bunch crossing in question. The next level in the selection, whether it is an intermediate Level-2 trigger or the full final Level-3, needs to collect data from all these independent buffers, in order to ‘build’ the data from these geographically and electronically distinct sources into a single ‘event’ whose final destiny will be determined by the HLT. In all cases, the processors running the final selection and the data sources must be connected via a high-speed network with a few thousand ports and capable of sustaining up to Tbs$^{-1}$ throughputs. The design and implementation of such a large networking system will be described briefly in the following section. Once the events are ‘built’, a special version of the offline reconstruction (which is usually based on the full offline algorithms appropriately steered; so only the parts necessary for deciding whether an event is to be kept for offline storage are active) is run. The result is an output of a few hundred events per second that are forwarded to an intermediate online store and eventually to the ‘Tier-0’ centre at CERN where the full offline reconstruction is applied.

3. A brief description of the main subsystems

Figure 3, along with the description of the core functionality of the DAQ system, reveals a fundamental property of the system that makes its design tractable: the different stages in the life of an event in the online system can be factorized into five essentially independent stages. The near-independence of these stages allows the design of a modular system that can be developed, tested and commissioned in parallel. This factorization is made possible via the deployment of buffering of adequate depth inbetween each of these stages. The primary purpose of these buffers is to balance the very different operating rates of the processing at each stage.

In the following, we concentrate on the DAQ system of the CMS experiment [3]. This is done for two reasons: first, a concrete implementation allows the description to concentrate on the major characteristics and development challenges, instead of comparing small differences across the four experiments; second, the CMS system is based on a two-level trigger system, i.e. a Level-1 and a HLT system. The requirements on the DAQ were, therefore, the most demanding, resulting in the most interesting ‘case’ to be reviewed.

(a) Stage 1: data pipelines and Level-1 trigger decision

This is the very first stage in the life of a pp crossing as recorded by the detector. At an input of one crossing per 25 ns and a desired maximum accept rate of one event per 10 ms, there are two major challenges. The data from the faster and easier-to-reconstruct detectors (calorimeters and muons) must be digitized and
transferred to a common place where they can be inspected. Just the transport of the data to this ‘intelligence’ and the propagation of the decision to accept or reject the crossing in question back to the detector front-end electronics take more than 1 μs. Allocating a minimum of 1–2 μs for the actual decision to be formed results in a total latency for the Level-1 trigger decision of approximately 3 μs, a time period during which the data must be stored in the detector front-end electronics. This is the first buffering stage, in the form of sequential storage or ‘pipelines’ capable of holding up to 132 crossings.

The corresponding development of the Level-1 trigger elements has been a long, painstaking process that has occupied multiple physicists and engineers from tens of institutions around the world. In the context of this overview, only a very short summary can be provided, and the full description for ATLAS, CMS, ALICE and LHCb experiments trigger systems is explained in the CMS [3], ALICE [4], ATLAS [5] and LHCb [6] Collaborations.

The CMS Level-1 trigger uses three different types of information, namely local (typically referring to a few detector channels); regional (usually referring to multiple local regions adjacent in space); and finally, global information combining...
identified objects from multiple detector elements. Local triggers are based on energy deposits in coarse regions of the calorimeter towers and hit patterns, possibly reconstructed as track segments, in the muon chambers. Regional triggers combine the use of look-up tables and pre-defined patterns to link local trigger information into trigger objects, e.g. electron or muon candidates. The information from the regional trigger is processed by the global calorimeter and global muon triggers, which determine the ‘best’ calorimeter and muon objects in the event. The definition of ‘best’ is typically a combination of thresholds on the momentum of the objects found and the quality of the reconstructed object. The ultimate decision to keep or reject an event is made by the global trigger, which takes into account not only the objects found by the processors, but also the readiness of the associated sub-detectors and the DAQ system.

A very important component of the system is the communication of the Level-1 trigger decision to the sub-detectors through the timing, trigger and control (TTC) system [7], which, in itself, was a major development that involved very precise laser systems and optoelectronics. While some of the Level-1 trigger components are located on the detector, the processors are hosted in the underground control room some 90 m from the experimental cavern. Taking into account the propagation delays for the signals to get to the processors and for the decision signal to come back through the TTC system, the total Level-1 trigger latency is 3.2 $\mu$s, i.e. some 128 bunch crossings.

(b) Stage 2: front-end readout and pushing data on approximately 500 parallel links

The next stage in event handling consists of the readout of the data in the detector front-end electronics and their intermediate storage in the ‘front-end buffers’ (referred to as FEDs), which are detector specific.

A key concept that facilitated the development was the introduction of a common protocol for all detector FEDs. Irrespective of the specific requirement and timings of each detector element, at their output, the detector front-ends were required to communicate via a standard protocol that would be the same across all detectors. This was achieved via the deployment of a common link, the FRL, which could combine data from two FEDs.

Since the occupancy of the sub-detectors in an LHC experiment varies across the detector, the resulting data per buffer can have significantly different sizes. A merging of FEDs, e.g. of two of them into a single link, is thus employed in order to ensure that each readout link, which represents a data source to the next stage in the system, delivers an average amount of approximately 2 kB. The FEDs and FRLs are located in the underground electronics room.

The introduction of the FRL made the development of this stage of the readout straightforward, leaving the complexity to the problem of interconnecting all these parallel buffers to the multitude of processors in the HLT farm. This is the subject of the next stage.

(c) Stage 3: event building stage

Here, all the data corresponding to a single event are collected from the FRLs via a large switch network. The specifications for this system were, at the time, beyond anything that was available in a single switching fabric. For this reason,
Figure 4. Implementation of the CMS event builder as a series of multi-stage switching fabrics. The system is modular, in that the second stage can use one or more (up to eight) switches, each providing an equal bandwidth for the event building process. At the output of each switch is a processor farm that executes the algorithms of the HLT. GPN, general purpose network; EPS, express processing; TPS, throttle trigger system; TTC, timing, trigger and control; GTP, global trigger processor; FED, front-end driver; FRL, front-end readout link; EVM, event manager; RU, readout unit; BU, builder unit; EVB, event builder; DAQ, data acquisition; DQM, data quality monitoring; FMM, fast monitor module. (Online version in colour.)
along with the desire to implement a system that would be as modular as possible, the event building network was designed as a two-stage switching fabric that could potentially use different technologies for each stage.

The actual implementation is shown in figure 4. In brief, the first stage of the event builder provided the interconnection to bring the data from multiple FEDs (via the FRLs) into a single ‘readout unit’ (RU). The first stage thus consists of a number of ‘FED builders’ implemented via Myrinet crossbar switches [8]. The second stage combines data from multiple RUs, via an ‘RU builder’ to build the full event into the memory of one of the processors in the HLT farm. The outputs from the first stage are cross connected to the inputs of the second stage in a manner that allows the operation of the entire system with only one RU. The DAQ is thus split into eight ‘slices’, each slice consisting of an RU builder appropriately connected to the existing FED builders. Each slice that gets installed brings a linear increase in the bandwidth of the system, while the interconnection topology implies an equally linear increase in the throughput that the overall system could sustain.

(d) Stage 4: the high-level (trigger) selection stage

Here the HLT algorithms are executed with the goal of providing a selection of 1 event in 1000. This is a small selection fraction, and therefore a significant amount of computing, corresponding to \(O(10^3)\) CPU nodes, is required. It is estimated that at a luminosity of \(2 \times 10^{33} \text{ cm}^{-2} \text{s}^{-1}\), the HLT algorithms will demand a mean of 100 ms of processing time on a 2 GHz Xeon CPU core. This implies for the 50 kHz DAQ system that an equivalent of about 1250 nodes with two dual-core processors must be deployed for the HLT. Extrapolating to the LHC design luminosity, where a total of eight DAQ slices will be operating at a Level-1 rate of 100 kHz, and assuming some gains from further optimization in the HLT algorithms, the required computing power is estimated to be roughly twice this figure.

(e) Stage 5: analysis/storage stage

Here, the events selected by the HLT are sent to a number of services, either for storage or for further analysis. This further processing can consist of a number of online monitoring tasks, e.g. for detector and trigger monitoring or for calibration purposes.

4. The system at work

After a long two decades of designs, redesigns, prototyping, testing and eventually going into ‘production mode’, procuring large numbers of processors and switches, the DAQ system was installed in time for the first LHC runs. The system commissioned for the first high-energy run at 7 TeV in 2010 was as follows.

— The full detector read-out, consisting of a total of 633 FEDs, was installed. The FRLs, i.e. the modules that provide the common interface between the subdetector-specific FEDs and the central DAQ, were also in place.
— Eight DAQ ‘slices’ with a full event-building capacity of 100 GB s$^{-1}$—corresponding to a nominal 2 kB per FRL at a Level-1 trigger rate of 100 kHz.

— An event filter farm comprising 720 PCs, each with two quad-core 2.6 GHz CPUs, which was allocated to running the HLT algorithms. The farm provided about 80 ms per event when running at an input rate of 100 kHz. The farm was extended in 2011 with the addition of 288 PCs with two six-core 2.6 GHz CPUs for a total of 8215 cores (see layout in figure 5).

— A 16 node storage manager system that allowed a writing rate that exceeded 1 GB s$^{-1}$, with concurrent transfers to the Tier-0 CERN computing centre at the same rate, and a total storage capacity of 250 TB. It also forwarded events to the online data quality monitoring (DQM).

By the end of the proton physics run in 2010, the LHC was operating with approximately 350 colliding bunches, reaching peak luminosities of approximately $2 \times 10^{32}$ cm$^{-2}$ s$^{-1}$. The typical operation of the DAQ was with an approximately 80 kHz Level-1 trigger rate, a raw event size of approximately 400 kB and an approximate 200 Hz recording of the physics stream [9,10].
Figure 6. The evolution of key technologies with time. Some important dates in the development of the trigger and DAQ systems are indicated on the time axis. CMS, Compact Muon Solenoid; TCP/IP, transmission control protocol/Internet protocol; LoI, letter of intent; DAQ TDR, data acquisition technical design report; LHC, Large Hadron Collider; TP, technical proposal. (Online version in colour.)

5. Historical perspective

Looking back at this period—with all the wisdom that hindsight allows—reveals a few major axes (or ‘bets’ on the future) along which the design and development were based:

— that the basic digital elements—namely microprocessors, memories and clock speeds—would (continue to) evolve at an amazing exponential rate;
— that there would be a true revolution in networking technology;
— that home-grown solutions should be limited to the electronics systems which are very close to the detectors. A corollary here was that industry would provide pieces of the solution that would be less expensive and more reliable than any home-grown solution; and
— that the then-infant Web would form not only a new way of communicating, but also a new way of controlling and monitoring complex instrumentation from afar.

The four hunches proved to be correct and over the past 15 years, the evolution of technology has provided the necessary tools and elements that were needed to put together the final trigger and DAQ systems of the LHC experiments. Figure 6 displays a summary of the evolution of key technologies over the past
40 years. Particularly noteworthy is the change in the slope of the Internet traffic line in the early 1980s. Closer to the early 1990s, at around the time of the technical proposal [11] that presented the conceptual design of the system, a major question was whether the exponential growth of networking needs (and technology) would continue at the same rate since the 1980s or a slowdown was imminent. Luckily, the original growth rate was resumed and sustained throughout the following decade: by the year 2000, the total traffic for data had exceeded that of voice (for global communications). The traffic for data has since continued its exponential growth.

Another very important element was the advent of industry standards and a high-reliability set of products by industry. The combination of the two really rendered the launch of any home-grown development unnecessary. Home-grown hardware became the solution of ‘last resort’, confined to those cases where the requirements were so specific or customized that the industry did not care to supply a ready-made solution. An interesting side effect was the parallel phasing out of ‘hardware designers’ and their substitution by software designers over the past 20 decades—at least within the community that was building the LHC experiments.

Perhaps, the best example to illustrate the above points is the development of the event builder. It is instructive to consider this example in some detail because it illustrates several of the issues that had to be addressed at the time.

The starting point here was the magnitude of the task: as mentioned earlier, designing, let alone constructing and commissioning, such a system within the specification and within the financial restrictions, was a very difficult task—at least in the early 2000s when these systems were being put in place [2,5]. While it was clear by then that networking technology would continue in its stride, and for this reason, the choice of actual networking technology should be delayed for as long as possible; it was also necessary to decide and freeze at least the layout of the system that would transport the data to the surface control room, where the full event builder and its output would be housed. At the same time, it was clear that technology could not provide, at least in time, a single fabric with 1000 ports and the bandwidth required for the full transport of all the data from the events passing the Level-1 trigger system.

A number of other requirements were also at work: there was, in the early 2000s, a significant uncertainty on the exact size of the system that should be installed in time for the first physics run, as well as on early performance of the accelerator and the experiment. The evolution of this performance was also an open question. With the rapid technological progress and the associated huge drop in cost per transfer bandwidth, it was clearly to the advantage of the experiment to deploy a design that could involve a phased installation of the full system. While the bandwidth of the event builder would be a function of time, the number of inputs was fixed: the number of readout buffers was dictated by the mean event size expected at the design luminosity. Furthermore, the prospect of additional installations or interventions in the readout part of the system, once the DAQ was deployed for the first run, were prohibitively complex. A final consideration was the significant uncertainty over the resources that would actually be made available for the completion of the DAQ, at around the middle of the decade, once all the detector subsystems had been constructed.
The result of these considerations was the decision to implement the event builder in the most modular possible way, to allow for a phased installation, and to also benefit from the expected improvements in technology in the period between physics runs. Modularity, along with a fixed number of inputs, implied designing the event building network as a two-stage switching fabric consisting of potentially different technologies. Only the first layer of switching, which would be connected directly to the data sources (the readout buffers), needed to be defined at the time, leaving the second-stage switches to be determined, procured and installed later. As mentioned earlier, this is the path that was followed, and the result has been the deployment of a high-performance, high-reliability system with a natural modularity that allowed the procurement and installation of the entire system in phases.

As for the Web, and its usage in today’s experiments, no words can do justice to the revolution it has introduced in the online systems of the LHC experiments. As much as the Web is currently inextricably linked to our daily experience, even daily life, it has become unthinkable to contemplate operating the LHC experiments in the absence of Web technologies. A few examples may shed light on this strong statement: the entire monitor and control systems are based on web tools and web services [12]. Remote shifts, data monitoring and even controls are available through Web interfaces; the full set of activities related to group work for the development, construction and operation of the system is based on the Web as well; lastly, all the system documentation is based on, and located on, the Web.

One can contemplate whether the field of particle physics has offered a good set of requirements and challenges that have inspired some of these developments. Leaving this question to historians, what did happen was that beginning with the early 2000s, the feasibility of acquisition of the individual components was no longer in doubt. The big question was whether the large numbers of electronics modules, computers, networks and controls and monitors that would be necessary to actually observe 40 MHz of crossings and select only about a few hundred per second, could actually be put together.

As mentioned earlier, the key to making this system a reality was the design of the two-stage event builder, which decoupled the development of the early stage of the DAQ system from that of the ‘back-end’, which consists of the filter farm. Idealizing the case study, one can regard the current DAQ systems as a collection of intelligent units, which are all interconnected via a network, as shown in figure 7.

Interestingly, the figure does not look different from that of the Internet (or a part of it) linking everything from servers to clients, control processes and data generators. What we have been observing in the past decade has been the gradual fusion of the networking and processing elements into single entities, which are capable of sending, receiving and processing data.

Extrapolating the current rate of progress in this area, it is difficult to believe that the next generation of trigger and DAQ systems in high-energy physics will be using dedicated fabrics for handling the data from the detectors. It is not unthinkable to imagine that the data sources may be directly connected to the World Wide Web of the time, even using some of its contextual extensions for routing the data, processing it and routing it to appropriate clients at collaborating institutions. If history is to serve as guide, there should also
be one or two unexpected developments along the way, some of which will introduce capabilities that are currently unimaginable. Let us close with one such unimaginable possibility: could it be that one day, the future experiments will be able to look closely at every single crossing of the beams?

Entertaining these extrapolations into the future, of course, is a luxury that is made possible by the successful operation of the current system. Figure 8 shows an instantaneous display from the online monitor of the DAQ from a data-taking run in June 2011. The density of information on the display conveys some of the complexity of the underlying systems, while the operating parameters that are visible are testaments to the smooth, high-efficiency operation. The event shown at the upper left in the mini event display was selected by the system, evidently

Phil. Trans. R. Soc. A (2012)
because it was characterized as very special: it was the best event out of about 1000 others that occurred in the same 10 ms. Perhaps the event contains a Higgs boson, or the long decay chain of two supersymmetric particles. Whatever it may turn out to be, it is by now, along with millions other events, safely stored for further detailed analysis.

The baton of the challenge is now with the analysis of these events.

References

10 Mommsen, R. K. et al. 2010 The data-acquisition system of the CMS experiment at the LHC. In Int. Conf. on Computing in High Energy and Nuclear Physics (CHEP10), Taipei, 18–22 October 2010.