PROCEEDINGS OF SCIENCE

HLT for HL-LHC: technology and architecture for next decade TDAQ

Silvia Amerio[*]
University of Padova and INFN
E-mail: silvia.amerio@pd.infn.it

In view of LHC phase 1 and phase 2 upgrades, all LHC experiments are facing significant challenges in their real-time selection systems, due to higher production rate and greater complexity of the events. In this paper we review the main requirements for the High Level Triggers (HLT) of the experiments to sustain the extreme data taking conditions of the future LHC, and the solutions being developed to meet them.

INFN Workshop on Future Detectors for HL-LHC,
March 11-13, 2014
Trento, Italy

[*] Speaker.

1. Introduction

The goal of High Energy Physics (HEP) experiments, like LHC ones, is to answer the questions still open on the fundamental nature of matter and energy, space and time. As far as HEP experiments are concerned, such answers can be searched for in the production and detection of new particles, not yet discovered (as in ATLAS and CMS), or in subtle effects on properties of well known particles (LHCb), or in the study of the creation of ordinary matter out of the extreme conditions just after the big bang (ALICE). To produce new particles, ATLAS and CMS need higher collision energy and higher luminosity, which results in very complex events: more than 100 multiple interactions are foreseen in the LHC high luminosity phase after 2022. Such complex events will be produced at a rate of 40 MHz, while data write to disk for analysis will not exceed a few kHz. Similar problems will be faced by LHCb, which will greatly increase the size of the collected samples reading the full detector at 40 MHz, and by ALICE, which will need to sample the full 50 kHz interaction rate to collect the statistics necessary for its physics program. Detectors and search methods may be different, but the problems the experiments have to face in terms of event reconstruction and selection are very similar. All experiments need to develop an efficient and fast online selection system, in order to reduce the amount of information to be stored permanently to tape for future analysis. The selection system (trigger) is usually organized in successive levels, each capable of performing a finer selection on more complex physics objects describing the event. Trigger systems usually comprise a first level based on custom hardware, followed by one or two levels usually based on farms of general purpose processors (High Level Trigger, HLT). HLT systems are based on farms of CPUs. This solution has several advantages: general purpose processor farms are easier to maintain and upgrade, the HLT code can be run offline in simulation programs as it is and, time constraints permitting, it can be similar to the offline code. Finally, HLT farms are additional computing power for offline processing when the experiments is not taking data.

HLT systems are facing many challenges in view of LHC upgrade: given the higher event rate and pile-up, much more computing power is needed to reach a reasonable reduction rate, with increased costs for the experiments. Moreover we are facing a significant technological challenge due to the widespread use of many-core devices and parallel computing, which may reduce the cost of the farms at the price of re-thinking the experiment software.

In this paper we review the challenges being faced by the experiments in the construction of their HLT systems in view of the future LHC upgrades.

2. A continuous read-out for ALICE phase 1 upgrade

The physics objective of ALICE phase 1 upgrade (Run 3, 2019) is to precisely determine the Quark Gluon Plasma (QGP) properties, which will be accessible through measurement of heavy-flavor transport parameters, quarkonia down to zero transverse momentum and low mass di-leptons states. Since these processes do not exhibit signatures that can be selected by hardware triggers, they can only be collected by a zero bias (minimum bias) trigger. Moreover, to reach the necessary sensitivity, a sample of at least 10 fb^{-1}, 100 times the current one, is necessary. This can only be obtained sampling the full 50 Hz Pb-Pb interaction rate, where each collision is shipped to the online system, either upon a minimum bias trigger or in a self-triggered or continuous fashion. So the upgraded online system must be capable of handling an input rate of 9.2 TB/s, and perform a partial reconstruction and compression [1]. The upgrade architecture is shown in Fig. 1

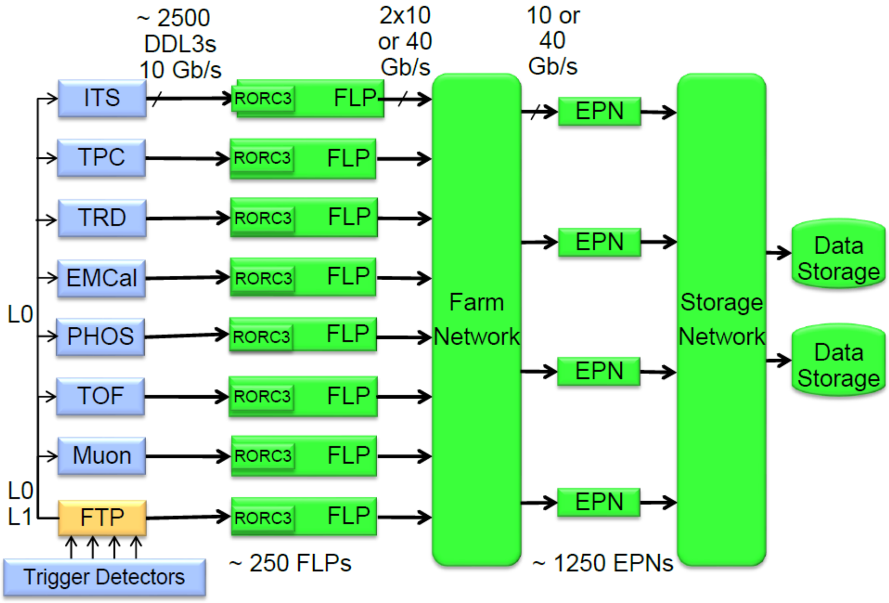

Figure 1: Architecture of ALICE trigger upgrade.

The Time Projection Chamber (TPC) and the Inner Tracking System (ITS) will have a continuous readout, while triggered readout will be used for the other subdetectors via the Fast Trigger Processor (FTP). Data will be transmitted with optical GBT links to a PC farm (First Level Processors, FLP) via a dedicated PCIe Gen3 card (RORC3) whose FPGA will perform a first compression of the data. The FLP farm will take care of the first processing of the data (clustering and local track reconstruction) and it will consist of about 250 nodes. Data will then enter a second farm (Event Building and Processing Nodes, EPN) made of about 1,250 nodes, for complete event building and processing, included global track reconstruction. As far as the network technology is concerned, current 10-40 Gb Ethernet or IB 56 GB will be sufficient[1]. The EPN nodes will be equipped with GPUs which have already been used during Run 1 with great success for track reconstruction in the TPC. A typical Pb-Pb collision event contains more than 20,000 tracks with several millions of hits in the chamber which makes online track reconstruction the most time consuming task performed in the HLT. A track reconstruction algorithm implemented partially on GPUs allowed to reduce the total processing time from 500 to 170 ms (see Figure 2 and [2]).

3. LHCb towards a full software trigger

LHCb goal for phase 1 upgrade is to collect a data sample up to 50 fb^{-1}, exploiting the greater luminosity offered by LHC. One of the main limitations of the current detector is that the collision

[1] Ethernet exists today in 10 Gbit/s and 40 Gbit/s versions (10G and 40G) and FDR InfiniBand offers effectively about 50Gbit/s. In both cases a variant with 100 Gbit/s speed will be available at the time of the upgrade which will be cheaper and reduce the number of necessary links.

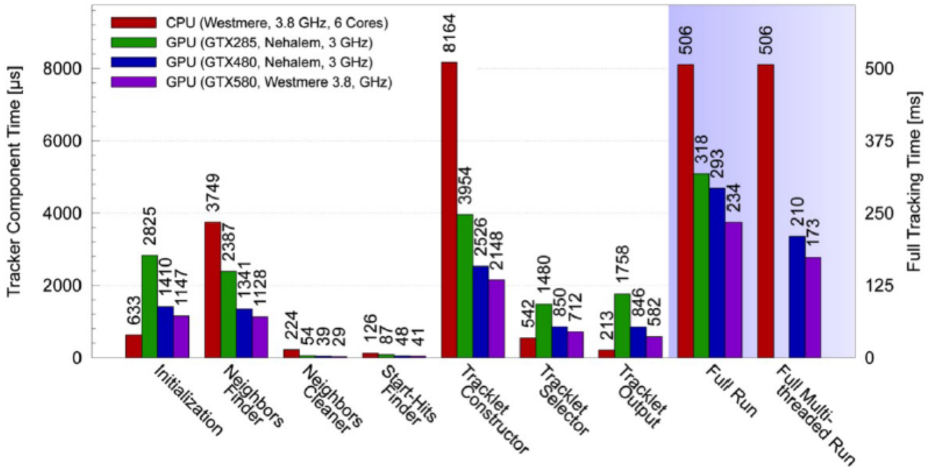

Figure 2: Performance of Alice online track reconstruction algorithm on different devices.

rate must be reduced to the readout rate of 1 MHz within a fixed latency. This reduction is achieved using the basic signatures available to the Level-0 hardware trigger. The largest inefficiencies in the entire trigger chain, especially for purely hadronic decays, occur at the Level-0 decision. Therefore, one of the main objectives of the LHCb upgrade is to remove this bottleneck by implementing a trigger-less readout system able to process the full inelastic collision rate of 30 MHz.

The future DAQ and trigger architecture will have to sustain an input rate of 32 Tb/s. The data will be processed by two farms based on commodity processors, the first for the event building (Event Building Farm, EBF), the second running the HLT selections (Event Filter Farm, EFF) as shown in Fig. 3. A LAN based on Ethernet (IEEE 802.3) or InfiniBand will bring the data from all readout boards to a single CPU node in the event building farm [3].

On each EB node a dedicated PCIe board (PCIe40) will receive input data through 24 4.8 Gb/s GBT links and write it via DMA on the node memory (Fig. 4). The EB node performs stable at 100 Gb/s with a total throughput of 400 Gb/s. Moreover, as the CPU load is rather modest during the event building, with about 80% of the CPU resources free, other applications can be run on the node. As an example, up to 18 instances of LHCb HLT application run in parallel without negatively influencing the event-building. This means that very conservatively at least 80% of the event-building server will be available for opportunistic use by the high-level trigger or a software version of the low-level trigger.

As shown in Fig. 5 a low level trigger, evolution of the current Level-0, is still foreseen as a backup, but the core of the selection will be in the EFF where the HLT algorithms are run. After the upgrade about 25% of the events are expected to contain a *b* or *c* hadron, thus the goal of the HLT algorithms has to change from background rejection to signal categorization. Full track reconstruction with the full input rate and offline-line particle identification will be two of the main ingredients of the upgraded HLT. The usage of many-core technology (GPUs, Intel-Xeon Phi) is also under study [4] to optimize the cost/performance ratio for the EFF.

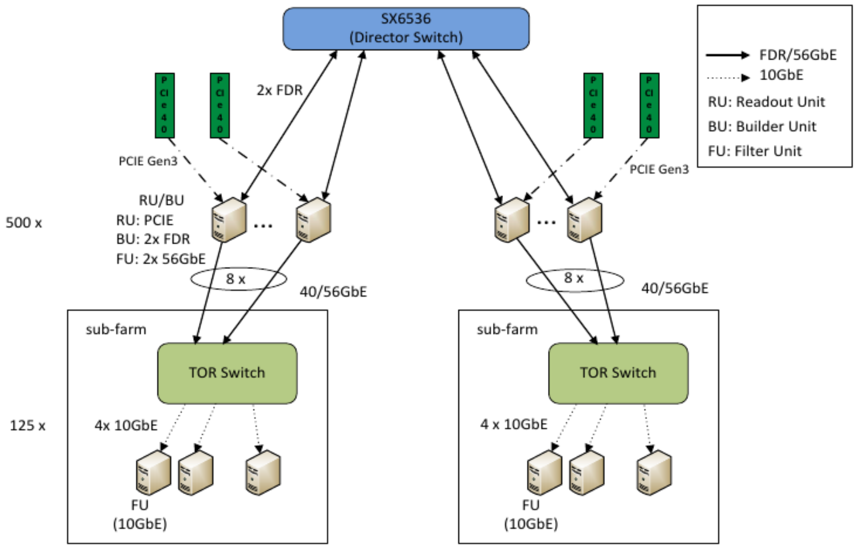

Figure 3: LHCb full-software trigger farms and links for phase 1 upgrade.

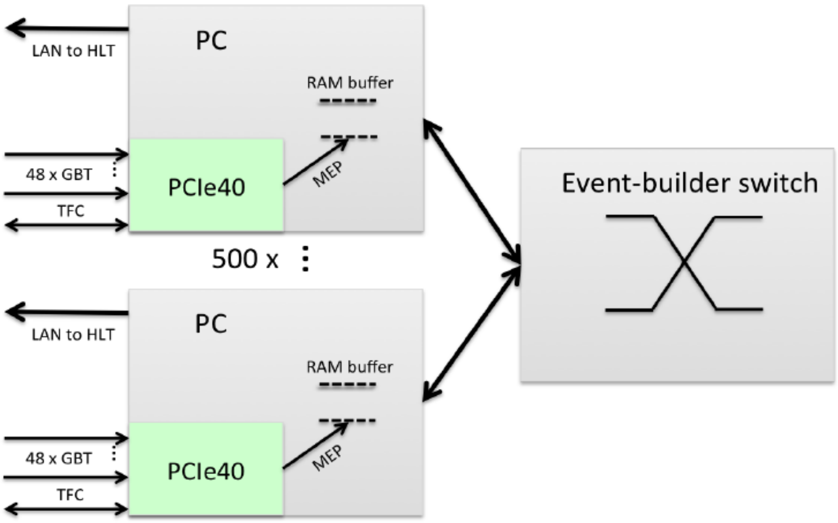

Figure 4: PCIe based readout system in LHCb.

4. Parallelism at all levels for CMS and ATLAS HLT upgrades

While ALICE and LHCb need to significantly upgrade their trigger systems already for phase 1, CMS and ATLAS trigger upgrades are foreseen for phase 2 (Run 4, 2025), when LHC will provide an instantaneous luminosity of $5\times10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ and more than 100 multiple interactions per event are expected [5] [6]. CMS and ATLAS HLT farms will have to reduce the event rate from 200 to 5-10 kHz and from 500 to 10 kHz respectively. This will require about a factor 50 more

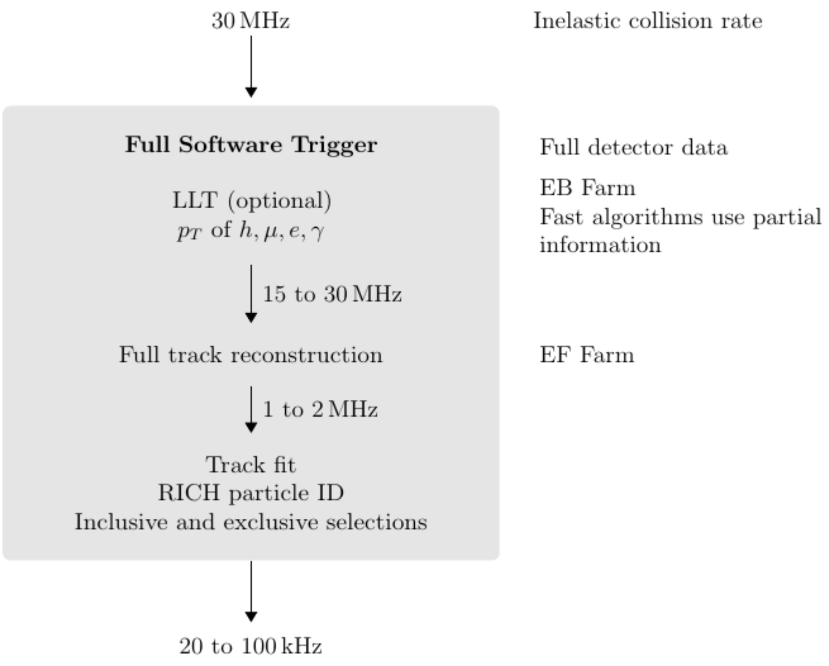

Figure 5: Schematic view of LHCb full software trigger.

computing power. Moore's law can lower this value by a factor ten, so improvements at different levels (software and hardware) are necessary to recover the missing reduction factor. A possibility on which all experiments are actively working is the implementation of parallelism at different levels to fully exploit multi-core CPUs: at event level, thanks to Copy-on-Write (CoW) and forking techniques to reduce memory needs; at algorithm level, thanks to multi-threading, running in parallel the different algorithms of a single data processing job; and inside the algorithm itself. This requires a lot of effort in the design of thread-safe algorithms and data processing frameworks which can handle all level of parallelism. AthenaMP in ATLAS [7] is an example of event-level multi-processing framework, with AthenaMT its evolution for multi-threading. CMSSW framework is also being upgraded to handle multi-processing and multi-threading [8]. The usage of many-core technology is also being explored in both experiments. As an example, ATLAS developed a GPU-based algorithm for the Inner Detector track reconstruction at HLT level. As shown in Fig 6, a significant gain with respect to the CPU-based version of the algorihtm was obtained both in the data preparation and in the track reconstruction phases [9].

5. Conclusions

All LHC experiments need to develop and implement new solutions for managing the large and complex datasets which will be produced by the future LHC. Many challenges are raised like acquisition, capture, storage and analysis which cannot be handled by the traditional acquisition and computing models of the experiments. The high production rate and event complexity will

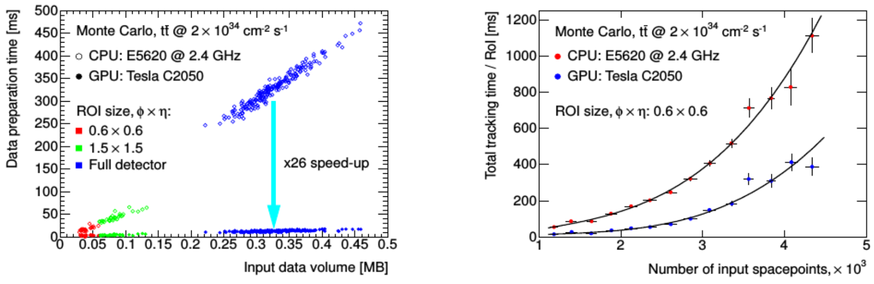

Figure 6: Performance improvement for data preparation (left) and tracking (right) steps with ATLAS GPU-based tracking algorithm.

make the extraction of relevant information much more challenging than in current experimental conditions.

ALICE and LHCb will need to reconstruct and classify the events in real-time with offline-like quality and will move towards trigger-less or full software trigger systems, implemented on commercial network and computing devices. This solution has the advantage of reducing the gap between online and offline processing and is easier to maintain and upgrade. The usage of PCIe-based readout boards allows to postpone the final decision on the technology and thus adopt the most advanced devices at the time of the upgrade.

ATLAS and CMS will still have low level hardware triggers, but are investigating many-core technology and parallel computing for their HLT systems. The introduction of this new technology is very challenging as it requires to re-design the experiments analysis frameworks and code to implement parallelism at different levels and to support different devices, e.g. from standard CPUs to GPUs and Intel-Xeon Phi.

All experiments are indeed facing a profound change not only in their real-time selection systems, but in their computing model in general.

References

[1] ALICE trigger TDR, CERN-LHCC-2013-019

[2] S.Gorbunov et al., ALICE HLT High Speed Tracking on GPU, IEEE TNS, Vol.58, Issue 4, 2011

[3] LHCb trigger TDR, CERN-LHCC-2014-016

[4] A. Badalov et al., GPGPU opportunities at the LHCb trigger, LHCb-PUB-2014-034. CERN-LHCb-PUB-2014-034. May, 2014.

[5] ATLAS trigger TDR, CERN-LHCC-2013-018

[6] CMS L1 trigger TDR, CERN-LHCC-2013-011

[7] Multi-core job submission and grid resource scheduling for ATLAS AthenaMP, D Crooks et al 2012 J. Phys.: Conf. Ser. 396 032115

[8] Multi-core processing and scheduling performance in CMS, J M HernÃąndez et al 2012 J. Phys.: Conf. Ser. 396 032055

[9] GPU-Based Tracking Algorithms for the ATLAS High-Level Trigger, D Emeliyanov and J Howard 2012 J. Phys.: Conf. Ser. 396 012018