Journal of Physics: Conference Series **664** (2015) 082023 doi:10.1088/1742-6596/664/8/082023

The ALICE High Level Trigger: status and plans

Mikolaj Krzewicki, David Rohr, Sergey Gorbunov, Timo Breitner, Johannes Lehrbach, Volker Lindenstruth and Dario Berzano for the ALICE Collaboration

Frankfurt Institute for Advanced Studies, Ruth-Moufang-Str. 1, 60438 Frankfurt, Germany

E-mail: mikolaj.krzewicki@cern.ch

Abstract. The ALICE High Level Trigger (HLT) is an online reconstruction, triggering and data compression system used in the ALICE experiment at CERN. Unique among the LHC experiments, it extensively uses modern coprocessor technologies like general purpose graphic processing units (GPGPU) and field programmable gate arrays (FPGA) in the data flow. Real-time data compression is performed using a cluster finder algorithm implemented on FPGA boards. These data, instead of raw clusters, are used in the subsequent processing and storage, resulting in a compression factor of around 4. Track finding is performed using a cellular automaton and a Kalman filter algorithm on GPGPU hardware, where both CUDA and OpenCL technologies can be used interchangeably. The ALICE upgrade requires further development of online concepts to include detector calibration and stronger data compression. The current HLT farm will be used as a test bed for online calibration and both synchronous and asynchronous processing frameworks already before the upgrade, during Run 2. For opportunistic use as a Grid computing site during periods of inactivity of the experiment a virtualisation based setup is deployed.

1. Introduction

The ALICE HLT has been successfully used throughout the LHC Run 1 as an online data reconstruction and compression facility. New concepts and technologies like the use of hardware acceleration for cluster finding using FPGAs and tracking on GPUs have been implemented and proven to be stable in production [1]. For the LHC Run 2 the ALICE TPC data readout has been upgraded [2], [3] to take advantage of the expected increase of luminosity. To cope with the higher data throughput related to this upgrade the HLT farm has been equipped with new hardware and a new networking layout has been implemented. Parts of the old production system comprise the development cluster used for software development, validation and testing. The main functions of the HLT in Run 2 will be:

- Data compression.
- Online event reconstruction.
- Online calibration.

Additionally, during periods of inactivity, the CPU resources will be opportunistically used for offline computing.

2. The computing farm setup

The new production system consists of 180 worker nodes featuring:

- 2x Intel Xeon E5-2697 CPUs running at 2.7 GHz, 12 physical cores each, with support for Hyper-threading,
- 128 GB RAM,
- 2x 240 GB SSD (RAID1),
- 1x 500 GB SSD,
- 1 AMD FirePro S9000 GPU,

and 8 infrastructure nodes equipped with

- 2x Intel Xeon E5-2690 (3.0 GHz 10 Cores each + Hyper Threading),
- 128 GB RAM,
- 2x 240 GB SSD (RAID 1).

Common ReadOut Receiver Cards (C-RORC)[4] serving as entry points for detector data links are installed in 74 worker nodes. Additionally a number of older machines serve various different roles like managing the OpenStack infrastructure for offline use.

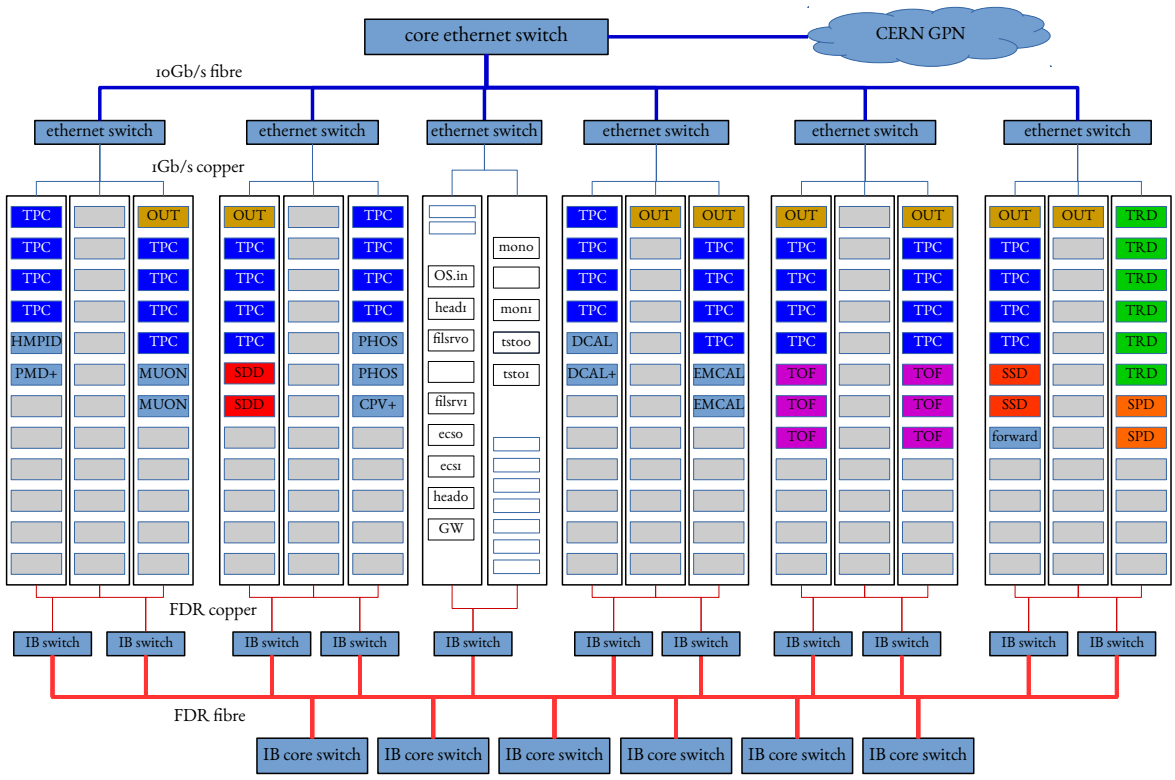

Figure 1. The layout of the ALICE HLT. The filled boxes represent the worker nodes, grouped in building blocks of 3 racks. Machines equipped with C-RORC cards are highlighted and tagged with the name of connected detector system. The white boxes represent the infrastructure machines.

The entire production system occupies 1 row of 17 racks instead of 3 rows as was the case in the previous setup, lowering the overall maintenance cost. Thanks to the size reduction the entire system is connected to an existing uninterruptible power supply (UPS).

The machines are grouped in building blocks of 36 machines (3 racks). Each machine features two network connections: gigabit ethernet (1 GigE) used for management and monitoring and 56Gbit/s FDR InfiniBand (IB) for data transport. The machines in each building block connect to a single 48 port 1 GigE switch and to two 36 port IB switches (see figure 1). On the ethernet side each building block switch connects to a 24 ports 10 GigE core switch using optic fibre connects. In each of the IB switches 18 ports are used to connect the nodes and 18 are used as uplinks to the 6 backbone switches using optic fibre interconnects, supporting full bisection bandwidth interconnection between all nodes in the system. This layout leaves 36 spare ports in the backbone network, a number of which are in use by the infrastructure nodes leaving some room for adding additional nodes (1 building block), should the need arise.

3. Infrastructure

Provisioning, software deployment and system configuration is fully automatised using Foreman and Puppet. Foreman manages all the basic networking infrastructure like DHCP, DNS, TFTP and centralises the Puppet management, which handles the setup of each worker node. All nodes in the system currently use Fedora 20 as operating system; CentOS 7 is also being considered. The machines are centrally monitored using Zabbix and Ganglia with all the benefits of those: automatic alarms, a web interface, trending, etc.

4. Core functionality

4.1. Online data processing

The online processing chain on the HLT starts with the detector data entering the C-RORC card over the DDL links [3] from the data acquisition system (DAQ).

The C-RORC is an 8 lane PCIe Gen2 card developed by the ALICE collaboration and currently used by both the ALICE and ATLAS detector readout. It features up to 12 bidirectional optical links with a maximum data transfer rate of 6.6 Gbps per channel. It hosts the Xilinx Virtex-6 FPGA, essential for the TPC cluster finding performed on-chip and the subsequent data compression scheme.

Cluster finding on the FPGA reduces the data volume by about 20% and enables further compression of the TPC data by data format optimisation and a Huffman encoding scheme. On average, the compression factor achieved is about a factor 4.2, illustrated in figure 2. The raw TPC clusters comprise most of the volume of ALICE data. Replacing them with the clusters compressed in the HLT significantly improves the data taking capability of the entire system, as was proven in previous years [1].

The reconstructed clusters serve as input to the GPU based cellular automaton (CA) track finding algorithm [6].

During Run 1, the HLT employed NVIDIA Fermi cards using the CUDA framework. The usage of CUDA imposes a vendor lock, and in order to become vendor independent, the tracker has been ported to the OpenCL framework during the shutdown period after Run 1. All code is written in a generic way: A common source file contains the entire algorithm while small wrappers for different APIs contain the API specific part. The generic algorithm file contains about 90% of the code, such that the majority of the sources is shared between the different implementations: for the CUDA framework, for OpenCL, and for the CPU reference implementation. The new HLT farm is equipped with AMD FirePro S9000 GPUs, yielding a performance boost in track reconstruction compared to the older model used in Run 1 [7].

Compared to the CPU version, the GPU tracker is about a factor 10 faster, see figure 3. As the final track merging and Kalman filter based refit is performed on CPU, the GPU needs to be assisted by 3 CPU cores to perform the entire tracking procedure. The performance of 1 GPU assisted by 3 cores is roughly equivalent to 27 CPU cores.

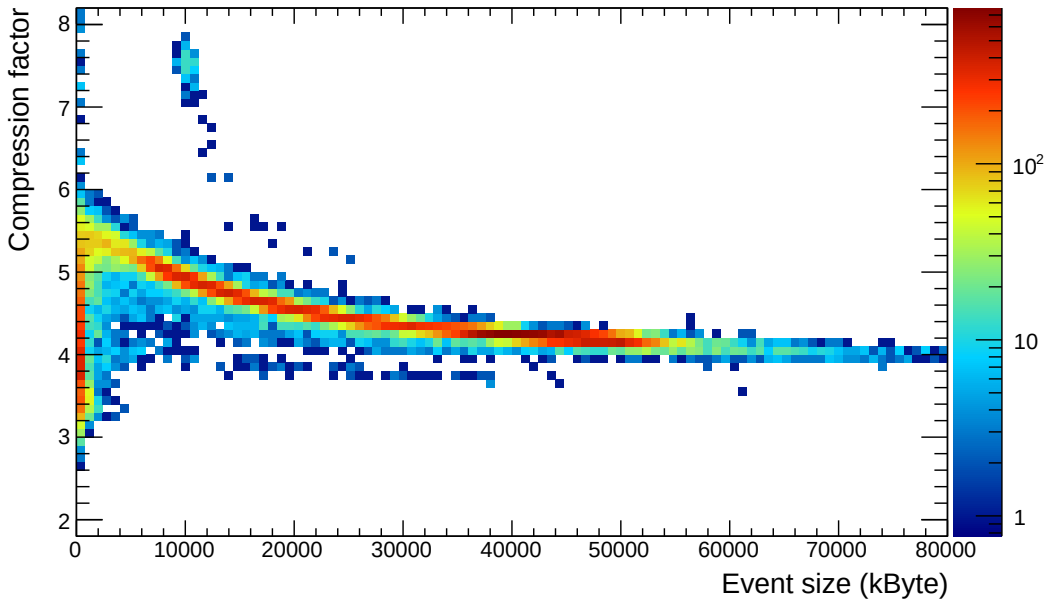

Figure 2. TPC cluster compression factor as function of the event size in the 2011 Pb-Pb data taking.

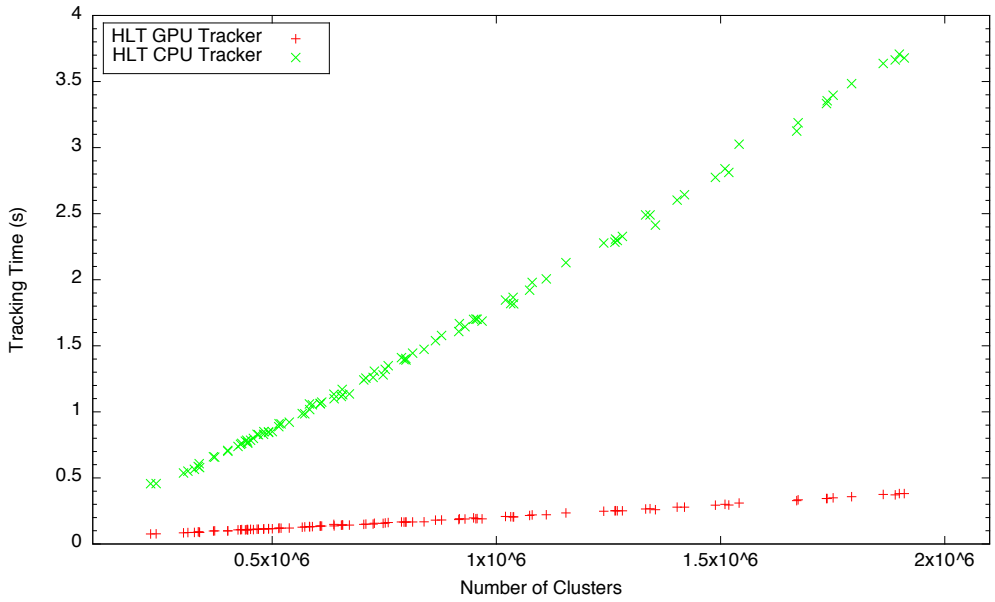

Figure 3. Tracking time of the HLT TPC CA tracker on Nehalem CPU (6 Cores) and NVIDIA Fermi GPU.

4.2. Online calibration

Calibration of the ALICE detector is currently performed offline in 2 passes: first the data is reconstructed and the TPC tracking calibration is performed. Then, in the second pass other detectors use the calibrated TPC tracking information to perform their calibration procedures.

The plan for Run 2 is to make use of the HLT reconstruction capabilities to perform the first

step of the TPC tracking calibration online. Other calibration procedures are also considered, depending on available resources.

Existing offline calibration code has originally been designed with processing complex offline data structures (Event Summary Data or ESD) in mind. The internal data representation in the HLT has been optimised for online performance and is largely incompatible with the offline calibration software. The main difference from an online processing perspective is that offine ESDs are non-contiguous in memory and need to be serialised/de-serialised, which introduces a non negligible overhead in some cases. A memory-contiguous (flat) ESD representation together with a common interface for flat and non-flat ESDs was developed to alleviate this overhead and allow the same code to run in both offline and online environments. Also a wrapper of the offline processing framework was implemented at the HLT component framework level to improve maintainability and consistency.

An important constraint is that the core HLT data transport framework needs to remain untouched in order not to introduce problems into the proven production system. The entire online calibration procedure is designed to run on top of the base software as a (asynchroneous) component within the HLT component framework.

Another issue is the performance of the TPC calibration procedure. Gaseous detectors are inherently difficult to calibrate and (in most cases) only limited performance gains could be achieved in a realistic time frame. In order to allow longer running procedures to be executed without blocking the normal HLT processing chain a procedure to run the calibration in an asynchroneous thread was developed and described in [7].

The described procedure effectively runs in parallel to the existing HLT processing chain without disturbing the data flow. Looking forward to the future, the next planned step is to apply the calibration in real time to online reconstruction and deliver already calibrated data, eliminating even further steps from offline processing. The ALICE post Run 2 upgrade will require full online calibration. Developing parts of this scheme already in Run 2 will be an important exercise. This closed loop approach requires the introduction of new data transport concepts. The current HLT framework does not support feedback loops, this is being developed in the component framework using ZeroMQ as transport, see [7].

4.3. Opportunistic use as a Grid processing site

During periods of inactivity in the data taking the computing resources of the HLT can be utilised to perform computing tasks normally associated with "offline" computing, i.e. Monte Carlo simulations, data analysis, etc. Given the fact that the HLT can host around 7000 job slots (using hyper-threading) it becomes a non-negligible resource.

An OpenStack based cloud running the ALICE Grid framework has been developed. It runs AliEn services and uses HTCondor as the underlying batch system. The number of running virtual machines is regulated based on the batch system load by a custom system (elastiq [8]). Virtual machines are only allowed to run on machines explicitly donated to the OpenStack pool by the HLT operator at any given time. The strategy for running offline jobs on the HLT is that jobs are expendable - the HLT can reclaim the resources at any time and the killed jobs will be automatically rescheduled to run elsewhere.

For security reasons a full separation between the internal HLT resources and the virtualised OpenStack setup is achieved using VLANs, see figure 4.

The system deployment and configuration is fully automatised using Puppet.

5. Conclusion

The ALICE HLT has been commissioned to participate in the upcoming LHC Run 2 data taking period. Consolidated with the new data readout requirements, featuring new, more powerful hardware and a new networking layout it will continue to perform it's vital role in the ALICE

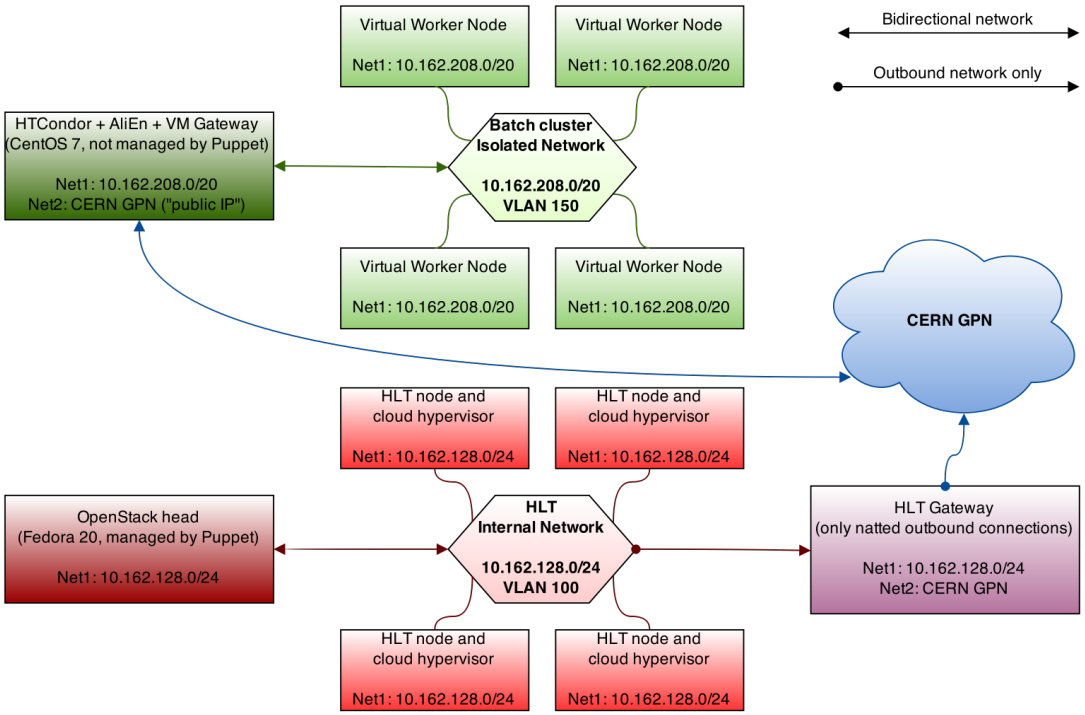

Figure 4. A schema of the network separation between the online processing resources and the HLT Grid cloud.

experiment as an online data reconstruction and compression farm, enabling the processing and storage of the envisioned extensive data volume.

New use scenarios are being implemented to more efficiently utilise the computing hardware available to the experiment. Online calibration is expected to remove the need for an entire offline reconstruction pass over the data and be a playing ground for the future upgrades of the experiment where online calibration will be a necessity.

The computing resources, when not needed for online processing will be used for offline processing by means of an OpenStack cloud participating in the ALICE Computing Grid as a sizeable grid site, even further optimising the overall computing resource utilisation of the ALICE experiment.

References

[1] Kollegger, T. 'The ALICE High Level Trigger: The 2011 Run Experience'. 2012 18th IEEE-NPSS Real Time Conference (2012): n. pag. Web. 14 May 2015.
[2] Alme, J. et al. 'RCU2 - The ALICE TPC Readout Electronics Consolidation For Run2'. J. Inst. 8.12 (2013): C12032-C12032. Web. 14 May 2015.
[3] Carena, F. et al. 'DDL, The ALICE Data Transmission Protocol And Its Evolution From 2 To 6 Gb/S'. J. Inst. 10.04 (2015): C04008-C04008. Web. 14 May 2015.
[4] Borga, A. et al. 'The C-RORC Pcie Card And Its Application In The ALICE And ATLAS Experiments'. J. Inst. 10.02 (2015): C02022-C02022. Web. 14 May 2015.
[5] Alt, T. 'High- speed algorithms for event reconstruction in FPGAs', presented at 17th Real Time Conference, Lisbon, 2009.
[6] Rohr, D. et al. 'ALICE HLT TPC Tracking Of Pb-Pb Events On GPUs'. J. Phys.: Conf. Ser. 396.1 (2012): 012044. Web. 14 May 2015.
[7] Rohr, D. 'Fast TPC Online Tracking on GPUs and Asynchronous Data Processing in the ALICE HLT to facilitate Online Calibration', these conference proceedings.
[8] https://github.com/dberzano/elastiq