PROCEEDINGS OF SCIENCE

Tracking Triggers for the High Luminosity LHC

Fabrizio Palla[*†]
INFN- Pisa and CERN
E-mail: Fabrizio.Palla@cern.ch

The planned High Luminosity Phase of the LHC (HL-LHC) will increase the collision rate in the ATLAS and CMS detectors by nearly an order of magnitude beyond the maximum luminosity for which the detectors have been designed. In that scenario, the number of proton-proton interactions per bunch crossing is expected to be about 140, on average. This very high pileup environment represents a major challenge for the L1 trigger of the experiments. The inclusion of the high granularity information coming from the Silicon Tracking detectors increases the performance of traditional triggers, based on Muon and Calorimeter information only. This poses new challenges in the design and integration of the novel inner tracking detectors in both ATLAS and CMS. On one hand, this is accomplished by modules capable of transverse momentum (p_T) discrimination, to only readout hits from relatively high p_T particles. A second stage performs pattern recognition and tracking at the first level trigger in a few μs, then combined with the rest of the Muon and Calorimeter information. This presentation will discuss the track trigger strategies, the expected performances of L1 tracking, and use of the L1 tracks in the trigger.

22nd International Workshop on Vertex Detectors,
15-20 September 2013
Lake Starnberg, Germany

[*]Speaker.
[†]On behalf of the ATLAS, CMS and LHCb Collaborations

PoS(Vertex 2013)034

1. Introduction

The LHC will undergo two major luminosity increases in the next two decades [1]. The first one, starting around 2019, will double the nominal luminosity of the LHC (thus reaching $2\times10^{34}cm^{-2}s^{-1}$), delivering nearly more than100 fb^{-1} in the following three years in ATLAS and CMS. After a long shutdown of 2 years, the luminosity will reach $5\times10^{34}cm^{-2}s^{-1}$, with nearly 140 proton-proton interactions per bunch crossing, and a yearly integrated luminosity of more than 250 fb^{-1}, for ATLAS and CMS. The LHCb will run at lower intensity, but will continue to integrate up to nearly 70 fb^{-1} by the end of 2028. Also the detectors will undergo an important rebuild of the inner trackers, and the main electronics. These two upgrades will allow the study of processes with very small cross sections, among the many, those useful to study the properties of the Higgs boson. The increased instantaneous luminosity and detector granularity will enhance the input trigger rate by a factor 100 compared to Run1. In order to keep the same level of selectivity with 10 times higher background rates, a higher background rejection power is mandatory , even in the case of a possible increase of the Level 1 (L1) rate. Using the tracker information, as currently done at the high level trigger (HLT) [2], is the only viable solution. However, the HLT algorithms run on commercial processors with times of a few tens ms, too long with respect to the available latency at L1 (a few μs). Even applying More's law to the commercial processors performance in 10 years from now it is impossible to match the goal, hence ad-hoc processors need to be developed, at least for ATLAS and CMS. In the case of LHCb, increasing the collision rate by running at higher luminosity requires that the thresholds of the Level 0 trigger should be raised which in turn implies, in particular for the hadronic channels, a drop in efficiency and to an even lower total event yield, thus requiring a revision of the selection strategy.

2. Trigger architectures

We will review the strategies employed by ATLAS, CMS and LHCb to use the information of the Tracker in the early stages of event selection. Two possible approaches have been developed, with different implications in the Tracker and Trigger electronics. In the first one - the "push" path - the trigger data from the Tracker are combined with the trigger data coming from other subdetectors (muon or calorimeters) and a "global-trigger" combines them as to make physics objects with refined information, like muons, electrons, jets etc. In this case the high granular information of the tracker is readout at 40 MHz, either after an on-detector data reduction like in CMS, or using the full information like in LHCb. In another approach, the tracker information is readout at a rate smaller than 40 MHz, following a "Level-0" signal from the muon or calorimeters in some region of interests, to then form tracker primitives to be combined at Level-1 with the calorimeter and muons ones in a Global trigger. This second approach, exploited by ATLAS, is less demanding in terms of data volume to be readout from the Tracker.

3. The Fast Track Trigger in ATLAS

The Fast Track project (FTK)[3] represents the state of the art technology which uses hardware processors to reconstruct tracks in real time: it is currently under development and is expected to be deployed in 2016 inside the experiment.

FTK is a custom electronics system that rapidly finds and fits tracks in the ATLAS inner track detectors for every event that passes the Level-1 accept. A conceptual design is close to complete. It uses all 11 silicon layers, 3 of the pixel detector (PXD), 8 of the SemiConductor Tracker (SCT), over the full rapidity range covered by the barrel and the disks. It receives a parallel copy of the PXD and SCT data at the full speed of the Level-1 accept from the detector specific ReadOut Drivers (ROD) to the ReadOut Systems (ROS). After processing the hits FTK sends out the helix parameters of all tracks with transverse momentum p_T above a minimum value, typically 1 GeV/c. The FTK system is a scalable system and can be easily expanded to operate at higher luminosities. The core algorithm consists of two sequential steps. In step 1, pattern recognition is carried out by a dedicated Associative Memory [4] (AM) device, which finds track candidates in coarse resolution patterns, i.e., roads. This step uses massive parallelism to carry out what is usually the most CPU intensive aspect of tracking by processing hundreds of millions of roads nearly simultaneously as the silicon data pass through. When a road has hits in all silicon layers or all but one, step 2 is to perform the fit using the hits with the full resolution in the road to determine the track helix parameters and a goodness of fit. Only those tracks that pass a χ^2 cut are kept. The road width must be optimized to balance the workload between the two steps. Too narrow roads would require large AM size and therefore too high cost, while too wide roads would increase the track fitting time drastically. The system is divided into eight core 9U VME crates, each covering 45^o in azimuth plus 10^o overlap. Each core crate is further divided into four regions of η and two regions in ϕ . This yields 8 $\eta - \phi$ towers per crate and 64 trigger towers altogether. A core crate could hold up to 16 AM boards, with 128 AM chips on each. With such a detector segmentation, the data can be distributed on 8 parallel buses at the full 100 kHz rate for the detector occupancy expected at the LHC design luminosity. A sketch is given in figure 1.

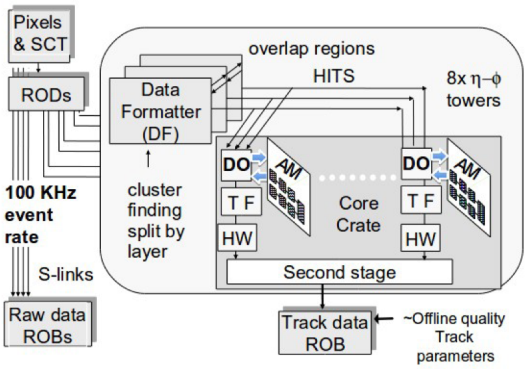

Figure 1: ATLAS FTK functional sketch. See text for explanations.

The PIX and SCT data are transmitted from the RODs and received by the Data Formatters (DF) which perform cluster finding. The cluster centroids in each logical layer are sent to the Data Organizers (DO). The DFs are not partitioned into regions but organize the detector data into the FTK $\eta - \phi$ tower structure and deliver them to the core crates, including the data in the overlap range. The DO boards store hits with full resolution and also convert hits to coarser resolution superstrips (SS) appropriate for pattern recognition in the AM. The DOs hold the smart databases

where full resolution hits are stored in a format that allows rapid access based on the pattern recognition road ID and then the hits are retrieved when the AM finds roads with the requisite number of hits. The AM boards hold AM chips which contain a very large number of preloaded patterns, corresponding to the possible combinations for real tracks passing through a SS in each silicon layer. These are determined in advance from a full ATLAS simulation of single tracks using detector alignment extracted from real data. The AM is a massively parallel system in that each hit is compared with all patterns almost simultaneously. The AM chip is currently under design in 65 nm technology. It features 8 serial input buses, each per layer, running at approximately 2 Gb/s and a clock frequency of up to 100 MHz. Every chip contains 128k patterns. When a road is found, the AM sends the road back to the DOs. A DO immediately fetches the associated full resolution hits and sends them and the road to the Track Fitter (TF). Because each road is quite narrow, the TF can obtain helix parameters with high resolution via a fit with the local coordinates in each layer. Such a fit is extremely fast and a modern FPGA can fit approximately 10^9 track candidates per second. The overall latency of the system is being evaluated to be of the order of a few 100 μs.

4. Level 1 Track Triggers R&D

We will now move to illustrate the current R&D on the Level-1 Track triggers.

4.1 LHCb

At the present time the LHCb trigger system has two levels: L0 is a hardware trigger to reduce a maximal event rate of 1MHz, followed by a HLT, run on a CPU farm. The hardware trigger uses information from objects of high transverse energy in the calorimeters and the muon chambers.

The upgrade of LHCb [5] foresees to increase the readout rate up to 40MHz and implement a flexible event filtering in software on a larger CPU-farm. The baseline strategy for LHCb is to develop a low level hardware trigger (LLT) with similar functionality to L0, using the calorimeters and muon chambers informations. The LLT will enrich the selected sample with interesting events, not to just pre-scale to a rate acceptable by the DAQ and the CPU-farm. Like the L0, LLT will protect against occupancy fluctuations that prevent full readout, and will have a tunable output rate to cope with whatever CPU power will be available. The high level algorithms running in the CPU-farm will have the whole event information available. It is anticipated that a factor up to 2 in efficiency can be reached for the hadronic modes. In order to reduce the "pressure" on the HLT and make a better usage of the CPU resources, the LHCb experiment is studying a Low Level Track Trigger (LLTT) solution, to be run at 40 MHz and provide tracks already at L0. The current R&D is concentrating in using the pixel hit information coming from 6 upgraded VELO planes and 2 Track Trigger planes, capable to deliver good quality tracks. Data from these planes are merged via a switching network and split into several boards equipped with FPGAs, running a real time track finding algorithm based on an "artificial retina" [6] in less than 1 μs latency.

4.2 CMS

The CMS current trigger architecture is based on a L1 trigger based on calorimeters and muon detectors informations, with an L1-accept rate of 100 kHz and a latency of 6.4 μs, followed by a HLT run on a CPU farm. The trigger upgrade strategy of CMS [7] foresees to increase the

L1 latency up to 10 μs and a L1-accept rate up to 1 MHz. The Track Trigger makes use of the information coming from the detectors located at radii above 30 cm. The aim is to provide tracks with $p_T > 2$ GeV/c in the full range of pseudo-rapidity covered by the Tracker ($|\eta| < 2.4$), thus also reducing the data rate to be used by the Track Trigger by roughly a factor 20. This is achieved by using ad-hoc developed modules (p_T-modules), exploiting the correlation, due to the high CMS magnetic field, between hits on sensors separated by a few mm [8]. The latest CMS Tracker layout is shown in Fig. 2 There are two types of (p_T-modules): those below 60 cm in radius are composed

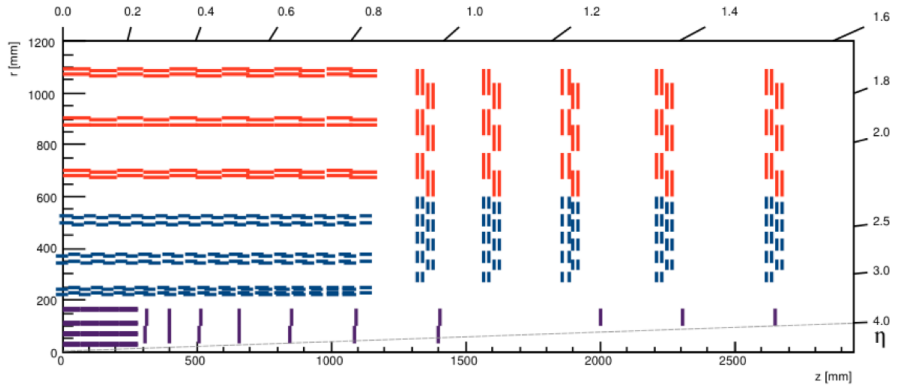

Figure 2: Sketch of a quadrant of the Tracker Layout. Outer Tracker: blue lines correspond to PS modules, red lines to 2S modules (see text for explanations of the module types). The Inner Pixel detector, with forward extension, is shown in purple.

by pairing a Silicon pixelated sensor with a Silicon strip sensor on top, and are called PS-modules. The pitch of the pixel and of the strips is 100μm. A special ASIC, still under design [9], performs cluster finding and form a pair of clusters (stub) compatible to being produced by tracks exceeding a transverse momentum of 2 GeV/c. The modules between 60 and 100 cm from the beam line are composed of pairs of Silicon strip sensors, and are called 2S modules. The strip pitch is 90 μm and are readout by a CBC chip [10] which performs digitization, cluster finding, filters clusters coming from low p_T tracks and finds finally stubs compatible to come from tracks with $p_T > 2$ GeV/c. In both 2S and PS modules, the ASIC has a tunable window to select the desired p_T threshold, which depends also from the sensor spacing which is instead fixed and optimized for accepting tracks above 2 GeV/c in p_T. The PS modules, being closer to the beam and made of one strip and one pixel sensor provide also a fair pointing precision, that allows to achieve a resolution of about one mm in the z impact parameter of the L1-trigger tracks, slightly larger than the average spacing of the vertices of the pileup events. The stub data sent to the trigger processors consist of the barycenter of the clusters in the two adjacent sensors and the direction of the vector joining them, whose slope is indirectly proportional to the p_T of the track candidate [11]. The stub data are sent to the trigger processors from each individual module, via GBT links [12]. The p_T modules select very efficiently the hits coming from tracks with p_T >2 GeV/c with a sharp turn on-curve, in both the barrel as well as the end-cap detectors, as shown in Fig. 3. However, a large fraction of stubs are produced by secondary tracks, mainly generated by interactions in the Tracker material.

Two different approaches are currently under study for L1-track reconstruction. In either case,

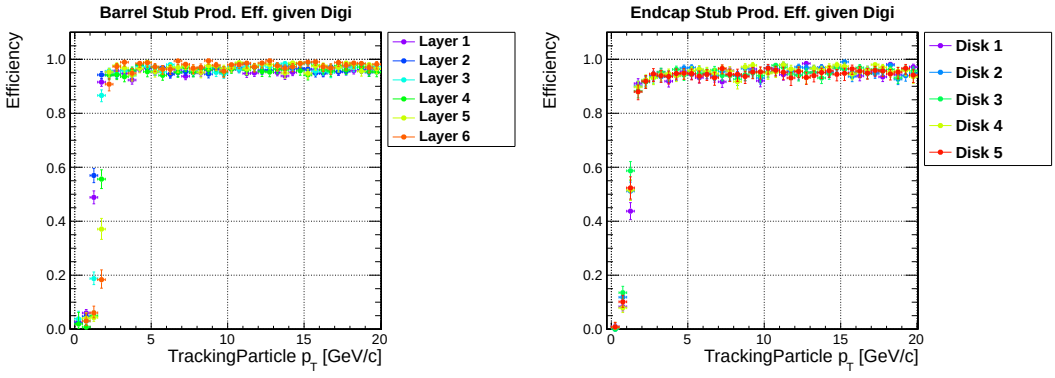

Figure 3: Stub efficiency finding for single pions in the barrel (left) and in disks (right) as a function of track p_T.

the trigger processors work in parallel using the data coming from several $\eta - \phi$ Tracker sectors. Due to the fact that low p_T tracks might originate in a given sector and bend in ϕ the adjacent one, or due to the large luminous region (roughly Gaussian with 5 cm width) along the beam line, there is the necessity to send a fraction of data from adjacent sectors in a given trigger processor.

One approach is based on using FPGA only. It first divides the Tracker into 16 ϕ and 2 η sectors. Tracks are searched by using an iterative algorithm which combines pairs of stubs in two subsequent detector layers: the seed (tracklet) formed by a pair of stubs is extrapolated to neighboring layers and another stub is added if it is compatible with the extrapolated track; the search is stopped if there are no compatible stubs.

The other approach makes use of a two stage procedure, similar to the one of FTK. Here the detector is subdivided into $8(\phi) \times 6(\eta)$ trigger sectors. In a first stage, data are formatted with coarser resolution than the detector pitch, and the pattern recognition is performed by matching compatible sequences of low resolution stubs in 6 detector layers with pre-computed pattern banks, residing in the AM chips. The high resolution stub data belonging to the matched pattern are then retrieved and a second step of track reconstruction is performed on a FPGA, thus dealing with a smaller combinatorial problem. Every single sector receives on average 200 stubs, divided into 6 detector layers. This number is about a factor of 5 to 10 smaller than the one from FTK, although they run at a lower LHC luminosity, because of the excellent data reduction provided by the p_T modules. As in the FTK approach the number of patterns required in each sector depends on the stub resolution and the minimum p_T of the track trigger tracks: coarser stub resolution as well as higher p_T threshold lower the number of needed patterns, but complicates the subsequent step of track finding in the FPGA, which in turn affects the latency of the trigger algorithm and the purity of the tracks. Viceversa, too precise stub information increases the number of patterns required, thus resulting in an unpractical number of AM chips. These parameters are now under a global optimization, and preliminary indications result in about 1-2 Million patterns for a p_T threshold of 2 GeV/c. To give an example, a sector will require about a dozen of FTK chips.

One last open question regards the system latency. CMS has now started a vigorous R&D with

the aim to build a Vertical Slice Demonstration System. This system will comprise a full tracking trigger path and will be used with simulated high-luminosity data to measure trigger latency and efficiency, to study overall system performance, and to identify potential bottlenecks and appropriate solutions; the demonstration system will investigate both approaches, those using full FPGA and the other one using AM chips and FPGA. The system is using state-of-the-art technologies, and is just used for demonstration purposes, since the final system will be needed at later stages, and would profit of additional R&D that will be performed in the future. The architecture under study is based on ATCA with full-mesh backplane as shown in Fig. 4. The large inter-board communication bandwidth provided by the full-mesh backplane is used to time multiplex the high volume (about 50 Tbps) of incoming data in such a way that the I/O bandwidth demands are manageable at the board and chip level, making it possible for an early technical demonstration with existing technology. The resulting architecture is scalable, flexible and open. For example, it allows different pattern recognition architectures and algorithms to be explored and compared within the same platform.

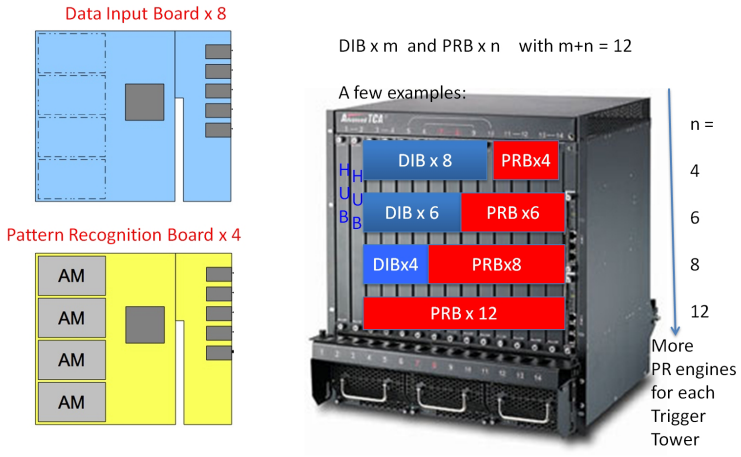

Figure 4: Possible demonstrator system for the CMS Level 1 Track Trigger. On the left, two types of boards: on top a "data input board" which receives the stub data, and on bottom a "pattern recognition board" that performs pattern recognition and track reconstruction using Associative Memory chips. On the right a ATCA crate equipped with several combinations of these boards.

CMS is also starting to investigate a possible L1 trigger using inner pixel detectors, located below 20 cm in radius from the beam line. Several studies are currently ongoing, like providing better precision primary vertices, or early rejections of photon conversions in the E-gamma triggers by identifying electrons tracks in the pixel, or tau lepton reconstruction. Given the high data rate expected, most likely a pull architecture (see Sec. 2) it is needed with additional latency up to a total of 20 μs to be included in the Level-1 trigger. These studies are yet preliminary and would deserve more time before being presented.

4.3 ATLAS

The current ATLAS Trigger is composed of three levels. The first level (L1) is implemented in custom-built electronics with a L1-accept rate of 100 kHz, followed by a the two-stage High

Level Trigger (HLT) is implemented in software executed on large computing farms. The ATLAS experiment trigger upgrade [13] is studying two different approaches. One approach is similar to the one just discussed for CMS, though the ATLAS Tracker uses double sensor modules only in some of the outer layers, given its lower magnetic field, as compared to the CMS one. The other solution uses an architecture sketched in Fig. 5, where the coarse calorimeter and muon detectors data provide a Level-0 (L0) accept signal at a rate of 500 kHz with a latency of 6.4 μs in several Regions of Interest (ROI). The data in the tracker front-end are kept in the pipeline and readout after a L0 accept only in some ROI. The L1 trigger is issued after combining the tracker data and a refined information from the calorimeters and muon detectors, at a rate of 200 kHz and a latency of 20 μs. This requires in turn that the data from the Tracker front-end chips should arrive to the L1 with a maximum latency of about 5 μs. This is done in the front-end chip by exploiting a dual buffer scheme for the the readout. The data coming from a stave of silicon strip sensors are buffered in a first FIFO; upon a L0-accept the data corresponding to that particular event are moved in a second buffer in the front-end chip, that upon a Regional Readout Request (R3) are read to the L1 Track system with a prioritization scheme, for that given ROI. In this way only about 10% of the readout data are effectively used for the trigger, since the L0-accept rate is 500 kHz. Simulations shows that using such a scheme, the data are available to the L1 Track finding logic in about 5 μs at the expenses of an increase of the total bandwidth of 50% with respect to the readout the track trigger.

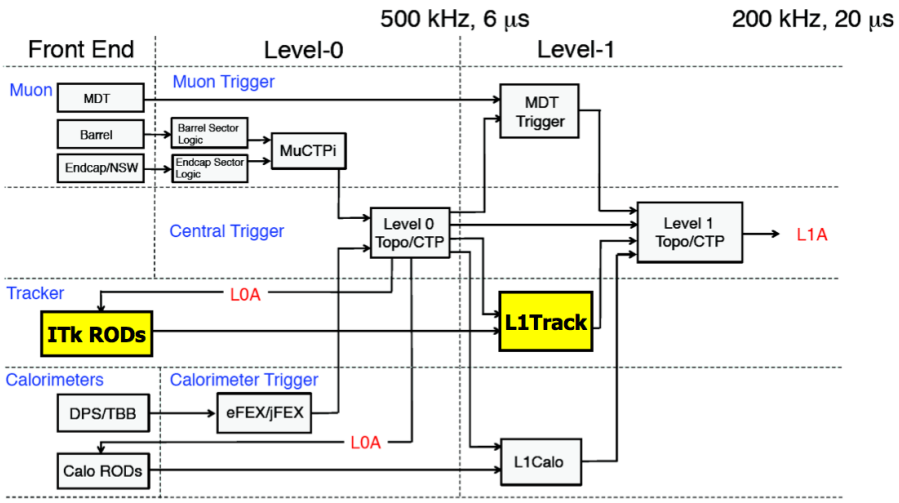

Figure 5: ATLAS L0 and L1 triggers readouts.

4.4 Tracking Triggers performances

Both ATLAS and CMS have started to evaluate the impact of including reconstructed tracks at Level-1, using simulated events. The main studies have concentrated on physics objects, like muons, electrons, taus or jets. For instance, ATLAS showed a rejection factor from 3 to 5 for electrons, while keeping 95% efficiency on electrons from W decays, with respect to the ones

without the inclusion of track-trigger, by matching the transverse energy in the calorimeter with the p_T of the electron candidate. Table 1 shows the preliminary studies performed by CMS [7].

Table 1: Overview of the projected improvement factors (as compared to the algorithm without Track-trigger) for key trigger objects in CMS.

Trigger object	Single Muon	Single Electron	Single Tau	Single Photon	Multijet
Main Track Trigger input	Improved p_T, matching and isolation	Matching with cluster, isolation	Isolation and track-calo matching	Track Isolation	Match jet vertex
Improvement factor	≈6	≈10	≈5	≈2	≈4

References

[1] http://hilumilhc.web.cern.ch/HiLumiLHC/index.html

[2] See for instance Eur. Phys. J. C 46, 605-667 (2006).

[3] CERN-LHCC-2013-007.

[4] A. Annovi et al. IEEE Nucl. Sci. Symp. Conf. Rec. 2011 (2011) 141-146.

[5] LHCb Collaboration, "Framework TDR for the LHCb Upgrade: Technical Design Report", CERN/LHCC-2012-007, LHCb-TDR-12 - 2012.

[6] L. Ristori, Nucl. Instrum. Meth. A 453:425-429, 2000

[7] CMS Collaboration, CMS Phase 2 Upgrade: Preliminary Plan and Cost Estimate, CERN-RRB-2013-124.

[8] D. Eckstein, these proceedings.

[9] D. Abbaneo and A. Marchioro, JINST 7 C09001.

[10] W. Ferguson et al 2012 JINST 7 C08006.

[11] F. Palla and G. Parrini, Pos VERTEX2007:034, 2007

[12] P. Moreira et al., "The GBT Project", in proceedings of the Topical Workshop on Electronics for Particle Physics TWEPP 2009, CERN-2009-006.

[13] ATLAS Collaboration, Letter of Intent for the Phase-II Upgrade of the ATLAS Experiment, CERN-LHCC-2012-022