Available online at www.sciencedirect.com

ScienceDirect

Nuclear Physics B Proceedings Supplement 00 (2014) 1–7

Nuclear Physics B Proceedings Supplement

The ATLAS Trigger System: Past, Present and Future

F. Pastore

Royal Holloway, University of London (UK)

on behalf of the ATLAS Collaboration

Abstract

The ATLAS trigger has been used very successfully for the online event selection during the first run of the LHC between 2009-2013 at a centre-of-mass energy between 900 GeV and 8 TeV. The trigger system consists of a hardware Level-1 (L1) and a software based high-level trigger (HLT) that reduces the event rate from the design bunch-crossing rate of 40 MHz to an average recording rate of a few hundred Hz.

We will briefly review the performance of the ATLAS trigger system during the past data-taking period and point out the challenges for the trigger system during the next LHC run in early 2015 with a smaller bunch spacing, almost twice the centre-of-mass energy and higher peak luminosity. We will show the ongoing improvements and upgrades to the existing system that will ensure an even better performing trigger system despite the harsher machine conditions. This includes changes to the L1 calorimeter trigger, the introduction of a new L1 topological trigger module, improvements in the L1 muon system and the merging of the previously two-level HLT system into a single event filter farm. In addition, we will give an overview of the algorithmic improvements in the various HLT algorithms used to identify leptons, hadrons and global event quantities like missing transverse energy.

Keywords: Trigger, ATLAS, evolution

ATLAS [1] is a general purpose experiment collecting data from p-p collisions at the Large Hadron Collider (LHC) [2]. In this paper we summarize the roadmap of the evolution of the ATLAS trigger system, starting from Run-1 performance, describing some of the hardware and software developments happening during the current Long Shutdown 1 of the LHC (LS1, 2013-2014), and showing the strategy followed for the forthcoming runs.

1. LHC Upgrades and new conditions for Run-2

The Large Hadron Collider (LHC) has been designed to provide proton-proton collisions at a centre-of-mass energy of 14 TeV with $\sim 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ peak luminosity. During the so-called Run-1 (2009-2013) a first step was reached, with the energy set to 7/8 GeV and a maximum peak luminosity of $0.7{\times}10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Run-2 should come much closer to meet or even exceed the design parameters. LHC is expected to operate at 13 TeV and reach around $1.6{\times}10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ peak luminosity. Further in the future, in order to extend the physics potential of the experiments and respond to radiation damage of the components, the accelerator will undergo a series of upgrades, that will take it to up to five times the designed luminosity in 2025, with the High Luminosity LHC (HL-LHC). To record the results of such collisions, the experiments will also be upgraded according to a similar schedule.

Starting by the end of 2014, Run-2 will conclude the first long shutdown (LS1), a 2-year long period planned to enhance the performance of both machine and experiments in preparation for the new conditions. On one side, the higher centre-of-mass energy will increase by factor 2 to 4 the production cross-sections of electroweak processes (including W, Z, H). On the other, the higher peak luminosity raises the number of interactions per proton bunch collision (~43 are expected), meaning higher density of interactions in space and time, and higher detector occupancy. Moreover the bunch-spacing will reduce from 50 to 25 ns (the design value), with a consequent smaller in-time pile-up (reduced number of interactions within the same bunch crossing), but increased out-of-time pile-up (due to superposition of interactions from close bunches).

2. The ATLAS experiment and its trigger system

The ATLAS experiment has been mainly designed for understanding the Brout-Englert-Higgs mechanism and for searching for evidence of physics Beyond the Standard Model (BSM). During Run-2, the experiment is then devoted to both investigate the electroweak energy scale and remain open to the discovery of new scenarios with very inclusive signatures. In particular, for a Higgs boson mass of 125 GeV, SM decays like $H \to \gamma\gamma, H \to ZZ, H \to WW, H \to bb$ and $H \to \tau\tau$ become relevant and a good measurement of the signals strengths in these channels enriches the list of ingredients needed for the accurate study of the Higgs parameters, like mass, spin, parity and coupling constants. So while Higgs studies are mainly being carried out with light leptons (and photons) at relatively low momentum, with higher luminosity the hadronic final states, like those containing forward jets or taus, will be given increasing importance. Beyond the Higgs, the study of the SUSY parameter space can be extended through final states with soft leptons, missing transverse energy (MET) or b-quarks, while many exotic scenarios can be identified with either very energetic signatures or multi-body decays.

Based on this physics program, the ATLAS trigger system has been designed to select a few hundred interesting events per second for permanent storage and subsequent analysis, discarding online most of the abundant low-momentum-transfer interactions. During Run-1, the rate of selected events was reduced from the initial 20 MHz LHC bunch-crossing frequency to average 700 Hz (reaching peaks of 1 kHz), with a rejection factor of the order of 30,000. To achieve this with negligible deadtime, the trigger system is split into three levels. The first-level trigger (L1) is a synchronous system with 40 MHz clock frequency and fixed latency of 2.5 μs. The L1 fast electronics process coarse resolution information from dedicated detectors in the muon spectrometer and/or in the calorimeter to identify high energy activity in the so-called Region-of-Interest (RoI). At this level, due to the limited latency, no tracking information from the high-resolution inner detectors can be read out. The high-level trigger (HLT) comprises two trigger levels, which both handle more complexity using fast software running on commercial CPUs. Having longer latency available, they can make use of the full resolution of all the detectors and apply quasi-offline selections. While the second-level trigger (L2) can access partial event data mainly from the L1 RoIs requiring few tens of ms, the Event Filter (EF) is designed to access the full event and process it with high detail, within a few seconds. The final trigger selection is based on combinations of candidate physics objects such as electrons, photons, muons, taus, jets, jets with b-flavor tagging (b-jets) as well as global event properties such as missing transverse energy and summed transverse energy (ΣE_T).

3. The trigger transition from Run-1 to Run-2

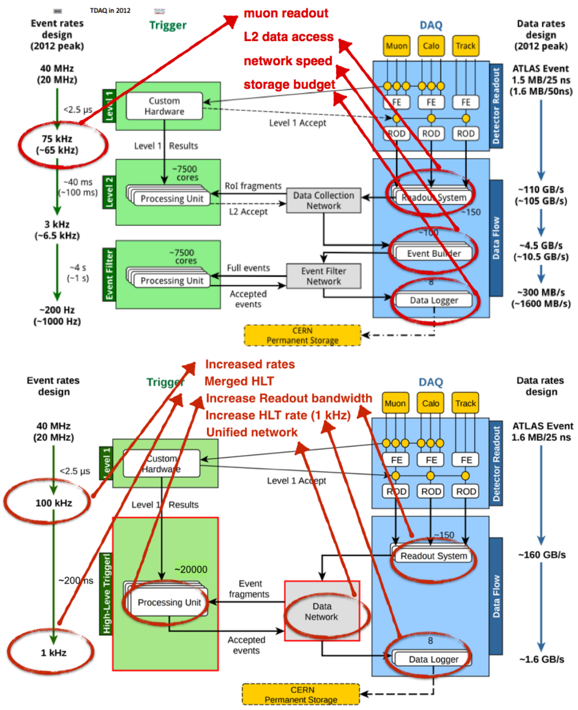

Figure 1: Overview of the ATLAS TDAQ architecture during Run-1 (top) and Run-2 (bottom), showing both the design values of latencies and sustained rates and the maximum reached during Run-1. The bottlenecks and the expected values for Run-2 are also pointed out.

During Run-1, the trigger system worked successfully. The trigger/DAQ architecture has been able to work within the design values, sometimes going beyond the expectations. As shown in Fig.1, there were several places where the system reached its limits, so it would not scale up to the new Run-2 specifications and needs to be upgraded. Examples of bottlenecks are: maximum L1 readout rate at 75 kHz (due to detector readout limitations), high data request rate from the L2 system to the Readout System (ROS), request of high network traffic for event-building and limited storage buffer (called "data logger"), which bounded the final HLT rate to maximum 1 kHz.

3.1. A new T/DAQ architecture

For Run-2, the plan is to apply limited changes to move to a more scalable, resilient and maintainable system, also exploiting new available technologies. The first remedy for higher rates is to increase the maximum L1 accept rate from 75 to 100 kHz, which requires radical changes in the Front-End of some muon detectors (Cathode Strip Chambers) and force others to operate in less redundant conditions (like reducing the readout samplings of the Liquid Argon Calorimeter). In general the actual limit for many systems depends on occupancy, so testing during commissioning and with early data will be critical.

A new DAQ architecture has been conceived, for better traffic sharing and more scalable design: in particular a faster Readout System (ROS) and a new network design merging the two existing networks into one. The details of the new Run-2 DAQ architecture and the status of its implementation are described in [3]. The upgraded ROS supports much higher data readout rates, up to 50 kHz. This enables a significant shift in strategy for Run-2: calorimeter-based high-level triggers that are sensitive to pileup such as multi-jet and missing transverse energy can potentially benefit from full detector readout. A new HLT architecture, matching the network evolution, is being prepared, in which L2 and EF levels merge and run, together with the event-building, within the same processing unit (L2/EF merging), as shown in Fig. 2. This has several advantages. The unlimited memory transfer between the two levels allows tighter coupling of the algorithms and reduces CPU and network usage. Moreover, a flexible event-building point will be possible, optimized on the trigger strategy requirements and the network traffic. It also gives the opportunity for re-optimizing the algorithms, to make them faster. In Run-1, multiple HLT processes were run on each computer in order to achieve high utilisation of the multi-core CPUs. In Run-2, the memory required increases, due to the switch from 32 to 64-bit executables and the merger of L2/EF into a single process, and becomes a limiting factor. Inspired by

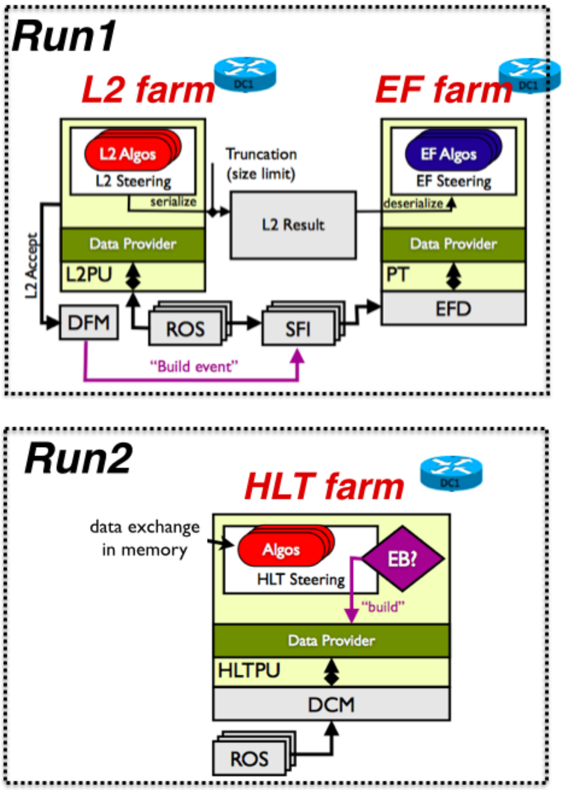

Figure 2: Schematic view of the structure of the HLT code, as used for Run-1 (top) and for Run-2 (bottom). The simplified network architecture, with one HLT farm, is reflected into a merged L2/EF infrastructure, more simple and compact. The Steering is the module scheduling the algorithms, while the Data Provider accesses the data from the ROSes.

earlier attempts to implement a multi-threaded HLT [4] and a multi-process development of the Athena framework [5], a multi-process HLT worker node has been designed. Memory is shared between worker processes by using copy-on-write Linux kernel feature when the processes are spawned, thus reducing the total memory footprint significantly [6].

Lastly, the speedup of the offline reconstruction software and a new analysis model allowed a further rationalization of the ATLAS offline computing and storage resources, so that the average output rate of the HLT can be raised to 1 kHz. The expected new design is shown in the bottom schema in Fig.1.

3.2. Towards a trigger strategy for Run-2

During Run-1, the available bandwidth was shared between the main groups of signatures: muons/B-physics, jet/taus/MET, electrons/photons, plus supporting triggers, calibrations and monitoring. Every selection was optimized for maximum efficiency and the threshold values chosen to keep rates within the allowed limits. The trigger strategy for Run-2 is still under study, but the final plan will have to ensure that the trigger works efficiently and with full physics coverage, up to L=2×10^{34}cm^{-2}s^{-1}. Following the main ATLAS physics goals, the trigger selection needs to strike a balance between maintaining sensitivity to the electroweak scale with inclusive single leptons, and keeping space for exclusive and/or multi-object triggers, single tau and hadronic (MET) selections.

The main limitation comes from the first-level trigger, since the higher energy and luminosity will lead to a factor 5 increase of most of the rates compared to Run-1, with larger increase for jets, due to the rise of parton luminosities functions at higher momentum, while lower contribution is expected for muons, dominated by fakes. So for L1, simply scaling Run-1 strategy to the new conditions, slightly rising the thresholds, would abundantly exceed the total rate over the 100 kHz limit.

Moreover, some trigger inefficiency is expected at higher pile-up conditions for all the selections. Higher radiation will indeed increase the fake rate, mainly in the forward regions, and will worsen the electron identification, highly contaminated by the increasing jet production: the isolation becomes the most powerful tool for distinction from this background. In addition, the worst resolution in the calorimeters will reduce the rejection power of the algorithms and leads to a less effective isolation and difficult pattern recognition. Similar challenges were faced already in Run-1, and actions were taken to mitigate the pile-up dependency for all the most sensitive signatures: some examples are the optimization of the calorimeter noise cuts, for the L1 missing-energy calculation (Fig.3) at different pile-up conditions, and the optimization of the HLT tracking algorithms, to maintain high tracking efficiency without incurring too great an increase in processing time.

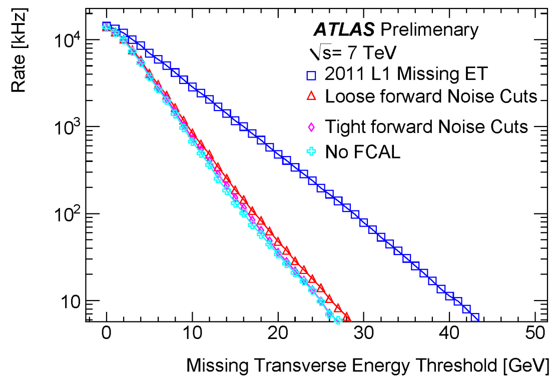

Figure 3: Level-1 MET rates as a function of threshold, for several pile-up noise cut scenarios. The 2011 configuration corresponds to noise cuts of ~1 GeV. The loose forward noise cut applies cuts in the first Forward Calorimeter (FCAL) layer and in the second layer at $|\eta|$ >3.5. The tighter noise cuts raise these by maximum 1 GeV, and the final case removes FCAL entirely from calculation. At Run-1 luminosity, the bulk reduction is achieved with the loose cuts [7].

4. Improvements in the first-level trigger system

4.1. L1 Muon Upgrades

The Run-2 upgrades for the L1 muon system will ensure higher rejection in the forward endcap regions ($|\eta| > 1.05$) and will increase the acceptance in the central barrel region.

Likely, the main single L1 muon trigger will require p_T > 20 GeV/c, with a target rate around 30 kHz. Raising the p_T threshold above this value does not reduce the rate significantly, due to the limited resolution of the selection (affected by multiple scattering and the extended beam-spot size). The dominant background comes from the forward regions, where low energy protons from the beam-pipe produce late fake coincidences (out-of-time pile-up). New requirements will be added to increase the robustness of the system in this region: one additional coincidence with (TGC) chambers upstream of the toroids and an additional coincidence with the outermost layer of the Tile Calorimeter, in a reduced η region (the transition region between the barrel and endcaps toroids). The rate reduction expected applying these requirements is around 40%. If further reduction is required, excluding the low B-field regions (with < 5% acceptance loss) or the higher η region ($|\eta| > 2$) would be also possible. Future upgrades for HL-LHC [8] will add new trigger detectors in the inner forward part, providing even higher rejection factors.

In the central barrel region, the trigger coverage is limited to 70% due to the inactive regions left for the toroid structure. In Run-2, the acceptance will be increased by factor ~ 4%, by completing the installation of the RPC chambers, as planned in the initial design.

4.2. L1 Calorimeter Upgrades

Under Run-2 conditions, the target is to apply soft isolation requirements to maintain the main L1 single electro-magnetic selection with p_T > 30 GeV/c. Having more flexible isolation algorithms is one of the requirements for changing the L1 calorimeter trigger electronics. The main challenge though is to ensure good efficiency for missing-energy and multi-jet selections in this high pile-up environment. Studies demonstrate that a rejection factor of more than 5 can be achieved on these signatures, by improving the energy scale and the resolution with better noise and pile-up suppression. Two main features are being addressed in the fast electronics: a first-order pedestal shifts based on the global cell occupancy and an energy correction based on the bunch position in the LHC bunch-trains. Indeed it was observed that an unbalanced overlaying of bipolar calorimeter signal shapes in bunches at the beginning of the train distorts the missing-energy calculation, raising the rate to uncontrollable values. An example of the impact on the MET rates is visible in Fig.4, where the new simulated filter and noise cuts show a clear advantage over those used in Run-1. To ensure these capabilities,

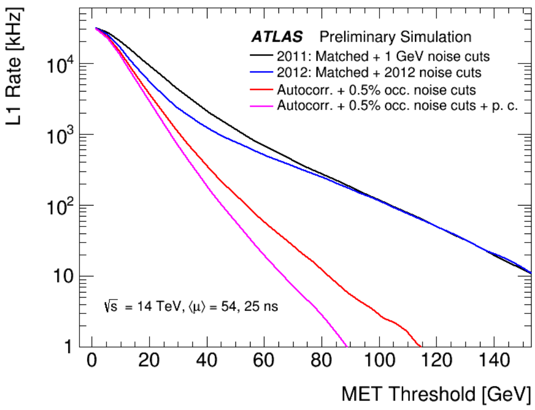

Figure 4: The estimated L1 MET trigger rate as a function of threshold, from 14 TeV minimum bias Monte Carlo. Shown are the operation scenarios with 2011 and 2012 noise cuts using matched FIR filters and two options for Run-2 with noise cuts optimized for 0.5% occupancy, using auto-correlation FIR filters with and without pedestal correction (p.c.) which are possible with the upgraded L1 calorimeter trigger system [9].

limited changes are applied to the L1 calorimeter electronics: new FADCs for faster digitization (80 MHz) with low noise, change from ASIC to FPGA technology to increase flexibility, implementation of digital filters and BCID algorithms, exploitation of full bandwidth to increase output details (fast optical links and faster crate communication). More details on these upgrades can be found in [10].

Future upgrades for Run-3 and Run-4 (HL-LHC) foresee the full replacement of the trigger electronics, with fast digitization already on the Front-End and increased cell resolution.

4.3. The new L1-Topological processor

The L1-Topological processor is a new component able to make up to 128 combinations of L1 signatures (like electron/photon, muon, tau, jets and MET) and quickly calculate correlation with simple algorithms based on event topology, like measuring angles, invariant masses, etc... The algorithms are implemented in FPGAs and allow 100 kHz output results with a latency of 100 ns. The large bandwidth, 1 Tbps, is managed by ATCA technology. This module ensures the possibility of studying exclusive signatures which would otherwise suffer from reduced rate at L1, applying exclusive requirements made anyway in the physics analyses. There are a number of benchmark physics cases, among which: the reduction of dominant background from di-jet with angular cuts, the combination of muon + jet to identify punch-through jets, the improvement of b-tagging or b-physics selections with soft muons. For example, by selecting the angular distance between a muon and a jet, a factor 5 rate reduction is expected for the identification of $b(b)H/A \rightarrow b(b)\bar{b}b$ with 33% efficiency loss, while gaining a factor 2 in the significance. A very clear example of signal/background separation using topological variable is visible in Fig. 5, for ZH→ $\nu\bar{\nu}$bb, using the ϕ angle between MET and the central jets in the event. In order to use the topological

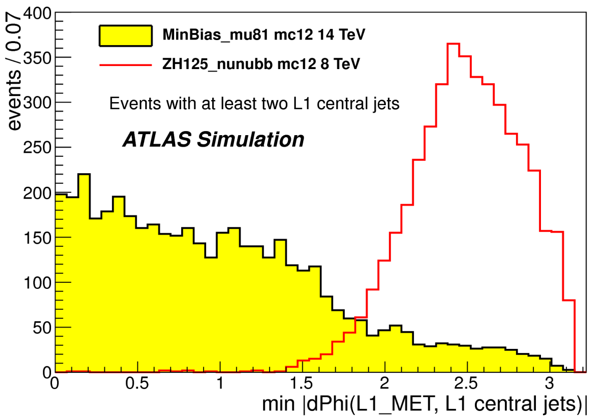

Figure 5: Minimum azimuthal angular distance between L1 MET and central jets with p_T >20 GeV and $|\eta| < 2.5$ for minimum bias and ZH events with at least two central jets [8].

processor, a change in the type of L1 information is needed. For example, the new L1 calorimeter modules need to provide the information of the full trigger object (energy, position, type of identified object), instead of the only trigger threshold, as was in Run-1. This is the main reason for increasing the data throughput in these modules, as seen in the previous section.

Lastly, the L1 Central Trigger Processor (CTP) will be also upgraded, in order to increase three times the number of inputs, making room for more trigger sources (like the topological processor), and to double the number of outputs, for more exclusive combinations. The additional requirements are fulfilled with extended capabilities, in terms of memory and latency, of the completely new hardware.

5. The HLT Upgrades for Run-2

The new DAQ architecture for Run-2 will achieve a global factor 100 rejection for the High Level Trigger system (from 100 kHz to roughly 1 kHz), similar with Run-1, but relaxes the data access requirements and allows more complex algorithms to be run, which can be much closer to the offline reconstruction algorithms. In this way, unnecessary inefficiencies introduced by different online-offline criteria are avoided, and closer thresholds can be applied. With offline-like algorithms in the HLT, better resolution can be achieved, and hence thresholds can be raised without adding inefficiency. It also allows for event-level quantities to be exploited, for example for pile-up suppression, and complex algorithms, based on likelihood or multivariate-analysis techniques, to be used for object selections. Examples of that are the identification of electrons and taus, as well as the b-tagging of jets.

The limited latency available for the HLT along with the large combinatorics at high pileup presents a challenge to develop software that will be fast enough. A lot of work has been done during LS1 to speed up the processing, by optimising the algorithms and devising more efficient strategies. For example, the "full scan" muon-finding algorithm, which searches the whole of the muon spectrometer and inner detectors for tracks and then tried to match them together, is very time consuming. By using the results of the muon spectrometer scan to guide the search into specific regions of the inner detector, a speed up of almost a factor of 4 was a achieved. Similarly, the algorithm forming the calorimeter cells (and their summations) is expected to be 6 (30) times faster, saving considerable time for the application of subsequent complex clusterizations, which are more robust against pile-up. Further changes concern the calorimeter clustering and the tracking algorithms, which are described in details in the sections below.

5.1. The Calorimeter clustering upgrades

Calorimeter clustering performance is highly sensitive to high levels of pile-up. As in the first-level trigger, also HLT software algorithms need to improve resolution and restore the calorimeter response. This is particularly important for jet and MET algorithms, which are

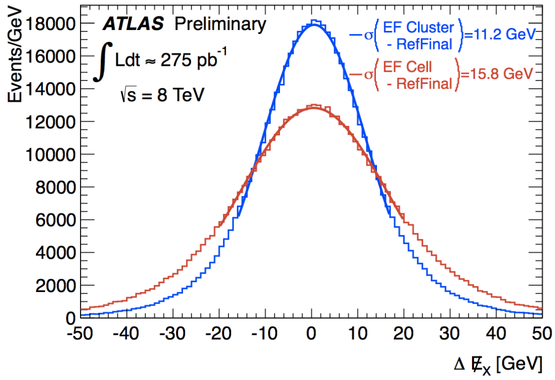

Figure 6: The resolution of the MET x component for 2012 data for EF, calculated as residuals with the offline measurement, for two different algorithms: in blue the EF topological cluster calculation and in red the EF cell based calculation [11].

based primarily on the cluster energy calculation. The most powerful way is to measure pile-up activity event-by-event and jet-by-jet. This is usually done by calculating the median p_T density of the event, excluding the hard physics components, but requires the readout of the full calorimeter for each hadronically-triggered event (at more than 20 kHz). This is one of the main motivation for the upgrade of the readout-system (ROS).

The topological clustering algorithm [1], used extensively offline, is considered highly efficient at suppressing noise in large clusters. In Run-1, improved MET resolution using the topological cluster algorithm in the EF was measured, as shown in Fig. 6. However it was too slow to use without a more simple pre-selection, which resulted in some inefficiency with respect to offline reconstruction. To overcome this, a 40% speed up has been achieved solely through code profiling and optimisation. So for Run-2, it will be available at the first selection step. Studies are ongoing to quantify the benefits for various triggers (mainly jet, MET and tau) and to check the behaviour with respect to pile-up. Some details on the reconstruction and identification of jets can be found in [12].

5.2. The tracking evolution

Many selections profit from high tracking performance, and the key point for rate control is to enable tracking algorithms as soon as possible in the trigger sequences, even if they can become expensive in terms

[1] A topological cluster is a group of calorimeter cells topologically connected, in contrast with the fixed-size cluster already used for the sliding-window technique adopted for electrons and photons.

of processing time due to high combinatorics as luminosity increases. For Run-2, the HLT tracking algorithms profit from the merged L2/EF levels by reusing the tracklets identified by a fast tracking stage, which then directly seeds the precision tracking, rather than running the offline pattern recognition. To mitigate the effects of higher pile-up, the RoI handling has been enhanced to avoid duplicated processing of overlapping regions. This is crucial for example to allow the b-jet identification at high rate, which is based on the measurement of the primary and secondary vertex positions. All of this motivates the total re-design of the trigger tracking system, with great benefits in terms of processing time. The new strategy has been demonstrated to be three times faster, with no impact on efficiency or resolution. Independent work to speed-up offline tracking code, used also for trigger tracks, can provide an additional factor two in performance.

Studies are on going to evaluate the potential of GPUs to accelerate tracking algorithms for next trigger upgrades. Having processing performance almost completely independent from the combinatorics, GPUs are also good candidates for other purposes, such as data decoding and jet reconstruction.

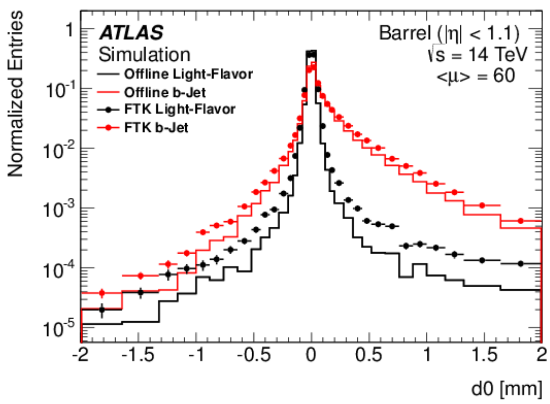

Figure 7: The transverse impact parameter for tracks associated to light-flavor (black) and heavy-flavor (red) jets. The solid lines show the distribution for the offline tracks, whereas the points show the FTK tracks [13].

An evolution of the tracking system is expected during Run-2, with the inclusion of the Fast TracKer (FTK) by the end of 2015. This is a "Level-1.5" hardware processor dedicated to extrapolate tracks from the full silicon detectors at L1 output rate (100 kHz). FTK is a massively parallel processor, based on Associative Memory technology, able to reach quasi-offline performance with a latency less than 100 μs, up to 70 pile-up events. Reducing the timing required for tracking, important features at event-level can be calculated in the first steps of the filter, like for example the primary vertex position and the multiplicity of tracks and vertices. As is shown in Fig. 7, this will be crucial for example for tagging the b-quarks [14] and improve the identification efficiency of taus [15].

For the high-luminosity scenario of HL-LHC, the use of tracking trigger is foreseen at limited latency in the Level-1 trigger [16].

6. Summary and future plans

The ATLAS trigger system is being developed to improve on the successful performance of Run-1 and meet the challenges of Run-2. It will be enriched with more sophisticated and flexible algorithms, in hardware and software. New ingredients will be added, like L1 topology, pile-up corrections and robustness and a faster tracking. Complex trigger menu and optimized algorithms will take advantage of the new homogeneous architecture (one HLT farm and network) for better timing performance. This is the result of efforts of hundreds of people, involved in prototyping and developing new tools during the 2-year LS1 shutdown period. Work still remains to integrate and carefully validate all the components, re-commissioning the whole trigger system before the LHC restart. The luminosity ramp-up could be very fast, so the system needs to be ready from the beginning, to avoid loss of physics data.

The upgrades developed so far are also compatible with the road-map for future LHC expansions. Preparations for the Run-3 upgrade have already started, mainly for L1 systems, such as the New Small Wheel detectors for muons and a brand new electronics for almost all the calorimeter.

References

[1] ATLAS Collaboration, 2008 JINST 3 S08003
[2] L. Evans and P. Bryant (editors) 2008 JINST 3 S08001
[3] See W. Panduro Vazquez's contribution at this Conference
[4] S. Armstrong et al, IEEE Trans. Nucl. Sci. 3, 915-920
[5] Crooks, D. et al. (2012), ICHEP Procs 2012
[6] ATLAS Collaboration, ATL-DAQ-PROC-2014-002
[7] ATLAS Collaboration, ATL-COM-DAQ-2011-152
[8] ATLAS Collaboration, https://cds.cern.ch/record/1602235
[9] ATLAS Collaboration, ATL-COM-DAQ-2013-150
[10] See A. Ezhilov's contribution at this Conference
[11] https://twiki.cern.ch/twiki/bin/view/AtlasPublic/MissingEtTriggerPublicResults
[12] See S. Cheatham's contribution at this Conference
[13] ATLAS Collaboration, CERN-LHCC-2013-007, updated version with IBL
[14] See D. Madaffari's contribution at this Conference
[15] See T. Mitani's contribution at this Conference
[16] See A. De Santo's contribution at this Conference