System Design and Prototyping for the CMS Level-1 Trigger at the High-Luminosity LHC

A. Tapper on behalf of the CMS Collaboration

***Abstract*—For the High-Luminosity Large Hadron Collider era, the trigger and data acquisition system of the Compact Muon Solenoid experiment will be entirely replaced. Novel design choices have been explored, including ATCA prototyping platforms with SoC controllers and newly available interconnect technologies with serial optical links with data rates up to 28 Gb/s. Trigger data analysis will be performed through sophisticated algorithms, including widespread use of Machine Learning, in large FPGAs, such as the Xilinx Ultrascale family. The system will process over 60 Tb/s of detector data with an event rate of 750 kHz. The system design and prototyping are described and examples of trigger algorithms reviewed.**

***Index Terms*—Field programmable gate arrays, Machine learning, Physics**

I. Introduction

The Large Hadron Collider (LHC) started operation with Run 1 in 2011. Since then the collision energy and luminosty have increased, resulting in proton-proton collisions at a centre-of-mass energy of 13 TeV and an instantaneous luminosity of over 2x10^{34}cm^{-2}s^{-1} during Run 2. In order the extend the LHC physics programme to enable increased discovery potential for searches and high-precision measurements, the LHC will be upgraded to deliver higher luminosity, resulting in larger data samples [1]. The High Luminosity LHC (HL-LHC) upgrade aims to reach instantaneous luminosities of 5-7.5x10^{34}cm^{-2}s^{-1}, a factor of 5 to 7.5 beyond the original design specification of 1x10^{34}cm^{-2}s^{-1}, which was alread exceeded in Run 2. With instantaneous luminosities at these levels the HL-LHC aims to deliver datasets of 3000-4000 fb^{-1} of integrated luminosity to the LHC experiments, allowing a significant extension to their physics reach. Fig. 1 illustrates the planned evolution of the LHC.

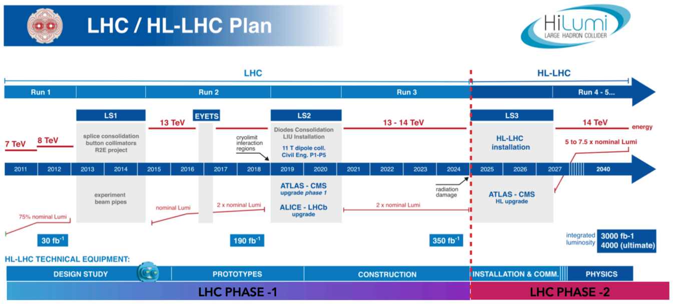

Fig. 1. Large Hadron Collider project schedule, showing the evolving performance plans for the accelerator, culminating in the high-luminosity era, denoted as Phase-2.

Imperial College London (e-mail: a.tapper@imperial.ac.uk).

II. Detector challenges

The higher instantaneous luminosity from the LHC, required to deliver larger datasets, results in increasingly challenging detector conditions.

The number of simultaneous proton-proton interactions (pileup) increases strongly from the design specification of an average of around 20 interactions per LHC bunch crossing to 140-200 interactions per bunch crossing, for the high-luminosity scenarios. The higher pileup results in degraded performance in many cases, for example higher occupancy in detectors often leads to failures in pattern recognition algorithms. Fig. 2 illustrates the expected charged particle multiplicity in HL-LHC running conditions.

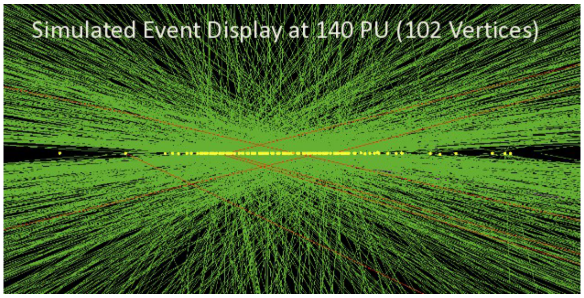

Fig. 2. Simulated high-luminosity collision with 140 pileup interactions, illustrating the charged particle multiplicity. The event has 102 reconstructed interaction vertices spread over a distance of around 10 cm along the beamline.

The increased particle flux also leads to a high radiation dose [2]. This is also a source of detector performance degradation, since radiation damage increases noise in detectors and leads to a lower response. Fig. 3 shows a simulation of the radiation dose expected for different areas of the Compact Muon Solenoid (CMS) detector, showing high doses for detectors close to the beamline.

Finally, trigger rates increase with instantaneous luminosity and trigger performance also degrades with increased pileup, since quantities such as isolation used to control trigger rates are less effective in high pileup conditions.

III. Detector upgrades

The consequences for detector performance described in the previous section motivate a large-scale upgrade to the CMS detector [2], [3], which will be undertaken in preparation for HL-LHC running.

Radiation damage and high detector occupancy require a complete replacement of the CMS tracker system [4]. The new

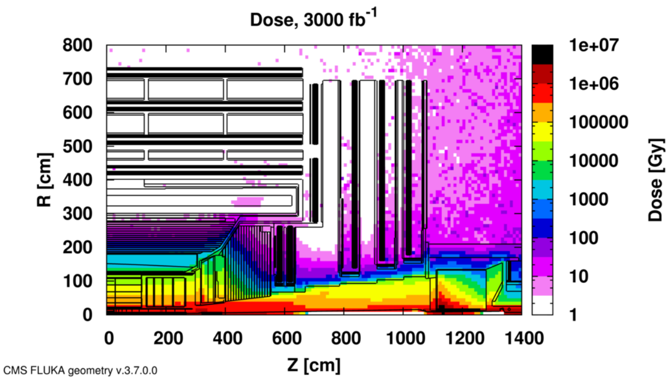

Fig. 3. Simulated radiation dose in the CMS detector after HL-LHC running of 3000 fb^{-1}, as a function detector radius and distance along the beamline from the nominal interaction position.

detector will again be constructed fully from silicon, using pixels for inner layers and strips in outer layers. Over 200 m^2 of silicon will be required, and the detector will have 10^9 channels.

The outer strip tracker is comprised of six detector layers in the barrel and four disks in the endcaps, with a strip pitch of 100 μm. For the first time the tracker will contribute to the CMS Level-1 trigger, delivering full tracks, above a p_T threshold of around 2 GeV, for pseudorapidities up to 2.4, for every event.

To achieve this the outer tracker is instrumented with p_T modules, doublet sensors which have common electronics to correlate signals and form stubs to be used in reconstructing tracks. Through careful choice of the spacing between sensors and correlation conditions a lower p_T threshold is achieved [5], reducing the data rate from the tracker by an order of magnitude, to a rate which is feasible for FPGA-based track reconstruction with a latency of around 4 μs [6], [7].

Fig. 4 shows schematically the concept of stacked, doublet tracking and illustrates the forming of p_T thresholds, which lead to data reduction.

The calorimeter endcap detectors must also be replaced since they were not designed to withstand the radiation environment expected at HL-LHC. A novel, high-granularity calorimeter supporting four dimentional (space-time) shower measurement will replace the current electromagnetic and hadronic calorimeters [8].

The new calorimeter will be a sampling calorimeter with layers of silcon sensors (600 m^2) and absorbers. The design is optimised for the HL-LHC high-pileup conditions with around 10^6 channels, leading to granularity of around 1 cm^2 and precison timing, with resolution of better than 50 ps.

The high-granularity calorimeter will also provide information to the Level-1 trigger, at a reduced granularity (4 cm^2) compared to the offline readout. Three dimensional clusters will be forwarded to the trigger with a latency of around 4 μs. The bandwidth and latency are insufficient to provide timing information.

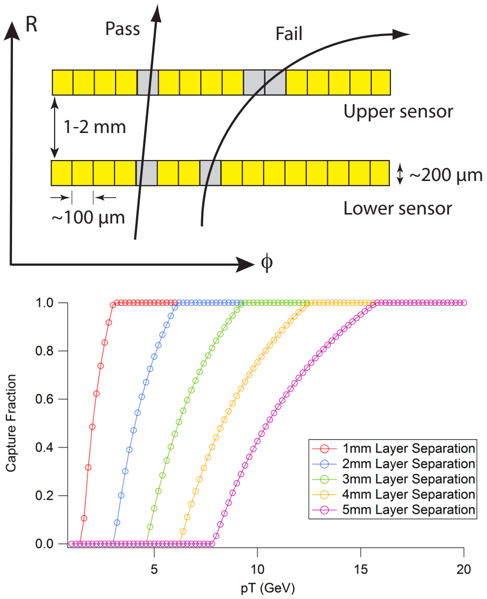

Fig. 4. Schematic of the doublet sensor design (top) and illustration of p_T threshold formation to reduce data volume (bottom).

The CMS muon system will also be upgraded, adding additional muon chambers in the forward regions of the detector, to improve performance in HL-LHC conditions [9]. This system also provides information to the Level-1 trigger.

IV. Technology R&D examples

In addition to the replacement of key detector components within CMS the corresponding electronics systems must also be replaced to handle the increased data volumes from higher granularity detectors with high channel counts and higher particle flux. A programme of electronics R&D has been undertaken to develop the required prototype electronics [10].

Generic high I/O processing boards, based on the Advanced Telecommunications standard (ATCA) and carrying state-of-the-art FPGAs have been developed. The I/O is based on high-speed serial optical links, to allow the transport and processing of large data rates to be achieved.

Several development projects have proceeded in parallel. Here two are described.

The APx consortium, comprising of several institutes, have developed prototype ATCA cards based on the Xilinx Virtex Ultrascale+ (VU9P) FPGA, with optical links running at up to 28 Gb/s. Board control is achieved via a Xilinx Zync SoC, with a dual core ARM processor. There is an optional 128 GB memory card which may be mounted on a rear transition module, to allow for large look-up-tables to be implemented if required.

The Serenity collaboration, again comprising of several institutes, have designed and produced prototype carrier boards, with with two sites for daughter cards. High density, low profile interposers are used to mount daughter cards with a variety of FPGAs. The boards also support optical links running up to 28 Gb/s. A commercial COM express card with an x86 CPU is used for board control.

Fig. 5 shows prototype APx and Serenity ATCA cards, used for technology tests.

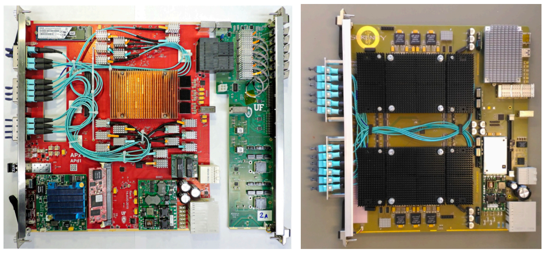

Fig. 5. Prototype APx (left) and Serenity (right) generic ATCA data-processing electronics.

The prototypes produced have been subject to a wide range of testing to ensure the technologies are suitable for widespread use in the upgrade to the CMS detector, and robust enough to function reliably. Examples of the thorough testing regime are shown in Fig. 6, which presents extensive, simultaneous link tests, for multiple channels at 28 Gb/s and Fig. 7 which shows the results of thermal cycle testing and corresponding thermal simulations.

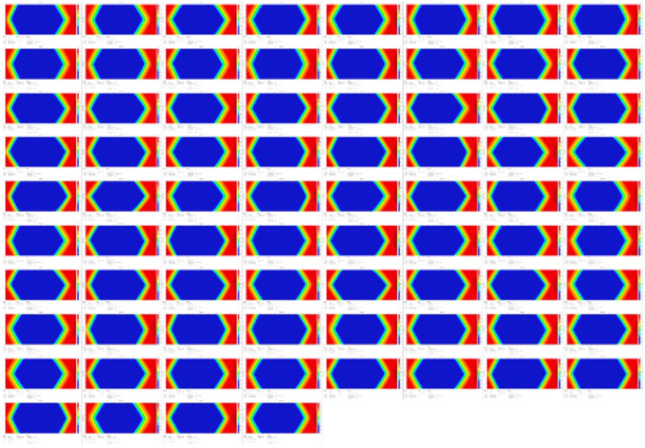

Fig. 6. Examples of electronics testing: results of a test of optical link signal integrity for multiple channels running simultaneously at 28 Gb/s.

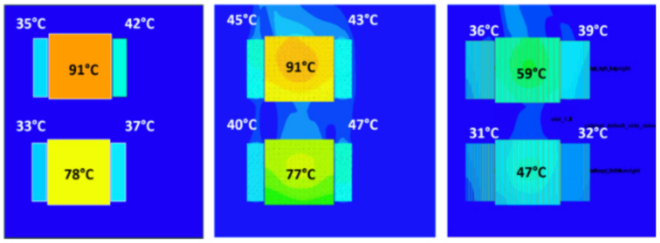

Fig. 7. Examples of electronics testing: results of thermal cycle tests, including comparison to simulation results.

V. Trigger System Design

The principal challenge of the Level-1 trigger design is to efficiently process the huge amount of input data and maintain robustness when detectors are functioning in a sub-optimal state and flexibility to evolve with conditions and an evolving physics programme.

The design illustrated in Fig. 8 is the result of these considerations and experienced gained with demonstrators systems, using the technology R&D described in the previous section. The design provides robust independent triggers for calorimeter, muon and tracking systems separately, and a particle flow trigger, which combines detector information, all feeding into a global trigger where the final trigger decision is made.

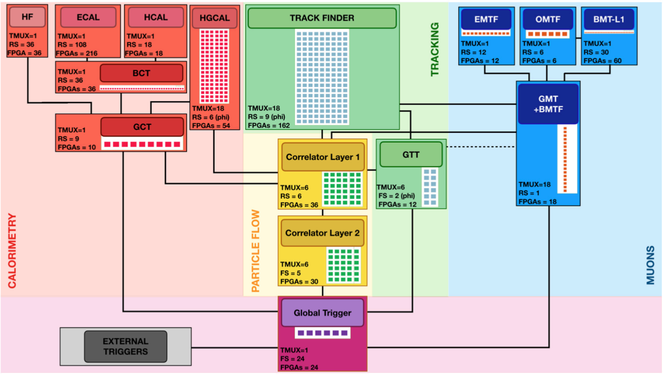

Fig. 8. Schematic of the Level-1 trigger design. The design provides independent trigger paths for calorimeter, muon and tracking systems separately, and a particle flow trigger, which combines detector information, all feeding into a global trigger. The small panels within each sub-component represent the number of FPGAs in each system, segmented in time (y axis) and detector region (x axis).

In order to meet the requirements the maximum trigger rate was increased from the current 100 kHz to 750 kHz. Similarly the latency was increased from 3.8 μs to 12.5 μs. The total amount of input data which is required to be processed is over 60 Tb/s. Data is transported via high-speed serial optical links, running at speeds of 16 and 25 Gb/s as appropriate for the interface requirements. Large FPGA parts (e.g. Xilinx Virtex Ultrascale+ VU9P/VU13P) are specified in areas where the trigger is processing bound and smaller parts (e.g. Xilinx Kintex Ultrascale) where processing is less critical. Overall over 200 FPGAs are required, with processing partitioned regionally and in time as appropriate.

In addition, the design includes a dedicated scouting system streaming data from key parts of the trigger at 40 MHz, via FPGAs into HPC resources. The scouting system provides unprecedented flexibility for parasitic debugging and commissioning of new ideas and is also being investigated for physics channels which are impossible with traditional triggering methods.

VI. Algorithm Examples

A complete set of baseline trigger algorithms has been developed and tested to ensure that the system designed will meet the requirements of the CMS HL-LHC physics programme. Two representative examples are discussed further below.

A. Particle Flow

Particle flow algorithms [11] aim to reconstruct and identify all particles in an event using all sub-detector information and have been shown to provide excellent performance in CMS offline and at the high level trigger. With the addition of efficient reconstruction of charged particles in the tracker, down to a threshold of 2 GeV and fine granularity calorimetry, to resolve the contributions from neighbouring particles, particle flow algorithms may be feasible in the upgraded CMS Level-1 trigger designed for HL-LHC.

A simplified particle flow algorithm was implemented in firmware, bringing together data from all sub-detectors available in the Level-1 trigger. Fig. 9 shows a schematic of how the information is used to reconstruct and identify particles, which then go on to be used to reconstruct trigger objects, such as hadronic jets and missing transverse momentum.

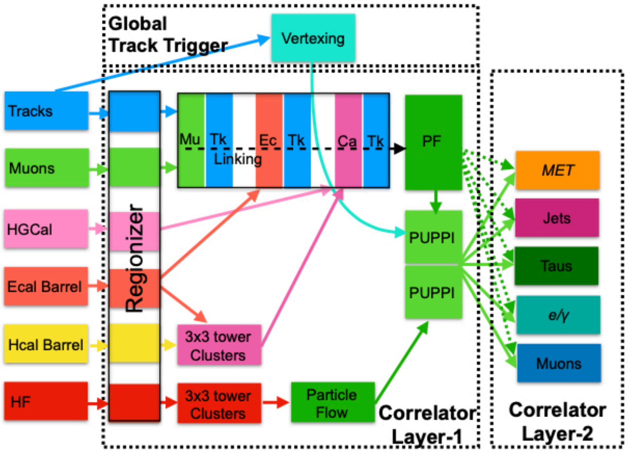

Fig. 9. Schematic of the particle flow trigger, illustrating how data from all sub-detectors is used to provide optimised reconstruction.

The interaction vertex, reconstructed using tracks, is used to implement the PUPPI algorithm [12], which filters particles based on a measure of their probability of coming from pileup.

This ambitious prototype algorithm was implemented in firmware and build for a Xilinx Virtex Ultrascale+ (VU9P) FPGA. The algorithm uses less than 50% of the FPGA resources (see Fig. 10) and runs with a latency of 0.7 μs, meeting the requirements for the HL-LHC trigger.

B. Vector Boson Fusion Higgs

The production of the Higgs boson in the vector boson fusion channel is a key process in the study of the properties of the Higgs and also to search for physics beyond the Standard Model.

The process is characterised by a pair of high-energy hadronic jets, with a large separation in rapidity from the production process with the decay products of the Higgs lying between the jets (see Fig. 11).

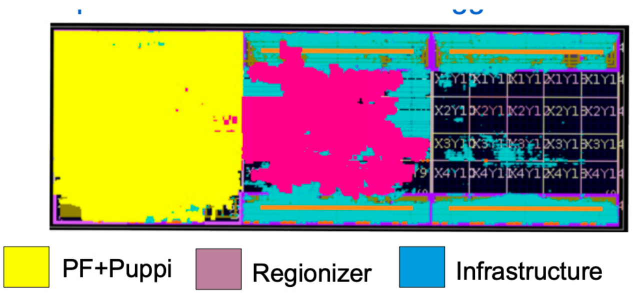

Fig. 10. The FPGA floorplan for the prototype particle flow algorithm, showing how the elements of the algorithm have been placed in the FPGA fabric.

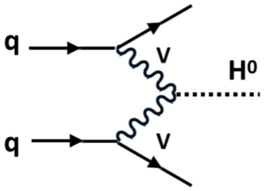

Fig. 11. Feynman diagram representing the production of the Higgs boson via the vector boson fusion process.

The topology of the final state, including jet pair, is exploited by trigger algorithms to achieve high efficiency at an acceptable rate. In the current CMS trigger, described in Ref. [13], it is possible to apply requirements on correlations between multiple objects, calulating invariant masses and differences in rapidity or azimuth for example. A natual continuation of such a strategy is to use modern Machine Learning tools to build powerful multivariate discriminators, instead of simple one dimensional criteria. Software tools now exist to port Machine Learning algorithms to FPGA firmware [14] and modern FPGAs are sufficient to host such algorithms with acceptable resource use and latency.

A study was performed as a proof of principle, for such a trigger in the HL-LHC environment. A Deep Neural Network, with three hidden layers with 72 nodes each, was designed with input variables based on jets and missing energy kinematics. It was trained to distinguish vector boson Higgs processes from multijet backgrounds. Efficiencies and rates were calculated and are shown in Fig. 12. In addition the model was built in firmware for a Xilinx Virtex Ultrascale+ (VU9P) FPGA, yielding 4300 multiplications per inference, a latency of 0.5 μs and DSP usage of 40% of the FPGA. The study proves the feasibility of such algorithms for HL-LHC and shows an improvement over the current trigger logic, which may be further improved in future studies.

VII. Conclusion

The Level-1 trigger system of the CMS experiment will be entirely replaced for the HL-LHC era. The new system will be required to process over 60 Tb/s of detector data with an event rate of 750 kHz. A novel design, allowing for flexibility while ensuring robust operation has been developed, based on

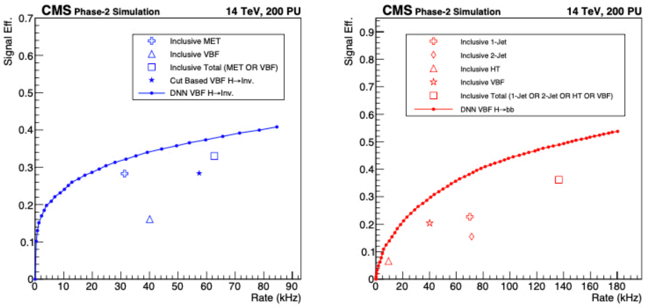

Fig. 12. Results from the feasibility study for Neural Network based vector boson fusion Higgs triggers. Efficiency and rates are shown for invisible decays of the Higgs (left) and decays to b-quarks (right). In both cases the Neural Network based trigger (solid line) outperforms existing triggers (shown as markers).

a programme of technology R&D and studies of the detector conditions. Examples of sophisticated algorithms, of the level necessary for HL-LHC running, have been implemented successfully, including algorithms based on Machine Learning.

References

[1] G. Apollinari et al., “High-Luminosity Large Hadron Collider (HL-LHC)”, CERN Yellow Rep. Monogr. 4 (2017) 1.

[2] CMS Collaboration, “Technical Proposal for the Phase-II Upgrade of the CMS Detector”, Technical Report CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, CERN, Geneva, Jun, 2015.

[3] CMS Collaboration, JINST 3 S08004 (2008).

[4] CMS Collaboration, “The Phase-2 Upgrade of the CMS tracker”, Technical Report CERN-LHCC-2017-009. CMS-TDR-014, CERN, Geneva, Jun, 2017.

[5] J. Jones, G. Hall, C. Foudas and A. Rose, A pixel detector for Level-1 triggering at SLHC, in 11th Workshop on Electronics for LHC Experiments, Heidelberg Germany, September 2005 CERN-2005-011.

[6] R. Aggleton et al., “An FPGA based track finder for the L1 trigger of the CMS experiment at the High Luminosity LHC”, JINST 12 P12019 (2017).

[7] E. Bartz et al., “FPGA-based tracking for the CMS Level-1 trigger using the tracklet algorithm”, JINST 15 P06024 (2020).

[8] CMS Collaboration, “The Phase-2 Upgrade of the CMS Endcap Calorimeter”, Technical Report CERN-LHCC-2017-023. CMS-TDR-019, CERN, Geneva, 2017.

[9] CMS Collaboration, “The Phase-2 Upgrade of the CMS Muon Detectors”, Technical Report CERN-LHCC-2012-023. CMS-TDR-016, CERN, Geneva, 2017.

[10] CMS Collaboration, ”The Phase-2 Upgrade of the CMS Level-1 Trigger”, Technical Report CERN-LHCC-2020-004. CMS-TDR-021, CERN, Geneva, 2020.

[11] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector”, JINST 12 P10003 (2017) .

[12] D. Bertolini, P. Harris, M. Low, and N. Tran, “Pileup Per Particle Identification”, JHEP 10 (2014) 059.

[13] CMS Collaboration, “Performance of the CMS Level-1 Trigger in proton-proton collisions at $\sqrt{s}$ = 13 TeV”, JINST 15 P10017 (2020).

[14] e.g. J. Duarte et al., “Fast inference of deep neural networks in FPGAs for particle physics”, JINST 13 P07027 (2018).