GPU/MIC Acceleration of the LHC High Level Trigger to Extend the Physics Reach at the LHC:

Principal Investigator:

Valerie Halyo
(908) 239 5351
vhalyo@gmail.com

Application type:

USCMS Phase II USCMS R&D

The quest for rare new physics phenomena leads the PI [3] to propose evaluation of coprocessors based on Graphics Processing Units (GPUs) and the Intel Many Integrated Core (MIC) architecture for integration into the trigger system at LHC. This will require development of a new massively parallel implementation of the well known Combinatorial Track Finder which uses the Kalman Filter to accelerate processing of data from the silicon pixel and microstrip detectors and reconstruct the trajectory of all charged particles down to momentums of 100 MeV. It is expected to run at least one order of magnitude faster than an equivalent algorithm on a quad core CPU for extreme pileup scenarios of 100 interactions per bunch crossing. The new tracking algorithms will be developed and optimized separately on the GPU and Intel MIC and then evaluated against each other for performance and power efficiency. The results will be used to project the cost of the proposed hardware architectures for the HLT server farm, taking into account the long term projections of the main vendors in the market (AMD, Intel, and NVIDIA) over the next 10 years. Extensive experience and familiarity of the PI with the LHC tracker and trigger requirements led to the development of a complementary tracking algorithm that is described in [arxiv: 1305.4855], [arxiv: 1309.6275] and preliminary results accepted to JINST.

The recently published work by the PI used a machine vision and pattern recognition algorithm, which are typically a good fit for the GPU architecture. However, the primary motivation was the similarity of the computational problem with image analysis and feature detection using Hough Transform that has been successfully used in other image processing applications. Many lessons and technical challenges have been learned in the process of developing the complementary tracking algorithms using massively parallel computing on the GPU/MIC. A few of the challenges include: the technical expertise of the personnel in algorithms, hardware, parallel computing, and physics; the complexity and time required to develop and optimize massively parallel algorithms; the tools necessary to test and debug the massively parallel implementations; and the test data that is required in order to compare the new code to the old code. As an example, one of the necessary tools needed to compare the new tracking algorithm to the original CMSSW tracking algorithm for the purposes of debugging, testing, and performance characterization is a stand-alone version of the CMS HLT tracking algorithm that can run using ROOT on any developers laptop. Completely isolating the code from CMSSW and making sure it yields the same results took a few months. Last but not least is the preparation of various input data and the design of input data formats that are coupled to the design of the algorithm. The input data format is needed to test the parallel tracking algorithm in the most realistic way possible.

Enhancement of the HLT might allow for track reconstruction of the full event up to design luminosity instead of regional tracking at the Level-1 event rate, alleviating the problems associated with high pile up. The proposed HLT tracking would have a far-reaching impact on the physics program set by the HLT where the ultimate goal is to select the events of interest for CMS. Beside reconstruction of the tracks down to 100MeV, the new tracking algorithm would also introduce new possible trigger paths that are currently not possible due to the extensive processing time that would be required using CPUs alone. The robust parallel processing of the tracking on the hybrid CPU and coprocessor system would allow reconstruction of not only charged prompt tracks from the interaction point but also reconstruction in real time of tracks, displaced vertices, and displaced jets in the tracker far beyond what is possible in both CMS and ATLAS. Therefore, the new parallel tracking algorithm on the coprocessor will enrich the physics program and allow for searches of models with new topological signatures that were not possible or suppressed before [4]. A few of the new topological models that could be selected in real time in the HLT include Hidden valley models, high jet multiplicity events [2] possible due to strongly interacting particles, models with boosted jets, long lived neutral particles decaying at macroscopic distances from the primary vertex [4], displaced jets and black holes [2], lower mass Higgs that decay to long lived particles [4], and last but not least more efficient selection of inclusive and exclusive b-decay channels. (Details and references can be found in [1],[2],[4])

Plan and Time Scale

Phase I (4 months) The PI plans to design the data parallel implementation of the full track reconstruction in the silicon tracker using the Kalman Filter that includes primary vertex finding. Hardware cost, energy consumption, and performance improvements will be compared with multicore CPUs. In addition, the use of RDMA in the form of GPUDirect and InfiniBand for L1 tracking will be evaluated.

Phase II (14 months) The design for the massively parallel implementation of the Kalman Filter will be implemented, The algorithm includes seed track finding in the pixel detector along with the algorithm to extrapolate the particle trajectories through the rest of the silicon tracker. Performance testing of the algorithms with MC using ideal conditions will initially be done. In parallel, input data from LHC and MC at high pileup conditions would be prepared for future comprehensive testing.

Phase III (6 months) The PI would carry out a comprehensive evaluation to compare time performance and tracking efficiency/purity of the GPU implementation against the Intel MIC based implementation and the original CPU algorithm. The test will include the setup of a small server farm to process real data as an end to end test.

Financial scale of the project

Hardware $30K for 4 systems, 2 with GPU cards and 2 with Intel MIC cards

Personnel 2 FTE at a cost of $130-150K for 2 years

In addition, 50% FTE of 4 to 6 graduate students from any institutions who wish to participate in the project would be welcomed.

To conclude, the new tracking algorithm will accelerate data processing to reconstruct in real time all of the charged tracks even in extreme pile up conditions, and also allow selection of events with new topological signatures for physics beyond the Standard Model. Such signatures could have evaded previous detection so this will provide the opportunity to extend the reach for new physics at the LHC. The full track reconstruction in the silicon tracker using Kalman Filter can potentially spin off other upgrades for each of the HLT subsystems (muon, calo), new parallel algorithms for CMS analysis tools, for LHC computing GRID, and simulation. All applications require implementations that are high performance and energy efficient, and most importantly can extend the physics reach of CMS at the LHC.

References

[1] V. Halyo, A. Hunt, P. Jindal, P. LeGresley, P. Lujan, *GPU Enhancement of the Trigger to Extend Physics Reach at the LHC*, [arXiv:1305.4855 [physics.ins-det] Accepted to JINST].

[2] V. Halyo, P. LeGresley and P. Lujan, *"Massively Parallel Computing and the Search for Jets and Black Holes at the LHC",* [arXiv:1309.6275 [physics.comp-ph] submitted to JINST.]

[3] Valerie Halyo interview in Nvidia Spotlight Newsletter
http://www.nvidia.com/content/cuda/spotlights/valerie-halyo-princeton.html

[4] V. Halyo, H. K. Lou, P. Lujan and W. Zhu, *"Data Driven Search in the Displaced $b\bar{b}$ Pair Channel for a Higgs Boson Decaying to Long-Lived Neutral Particles",* arXiv:1308.6213 [hep-ph] submitted to JHEP.