XILINX

CASE STUDY

Artificial Intelligence Accelerates Dark Matter Search

Integrating Inference Acceleration with Sensor Pre-processing in Xilinx FPGAs Delivers Performance Unachievable by GPUs and CPUs

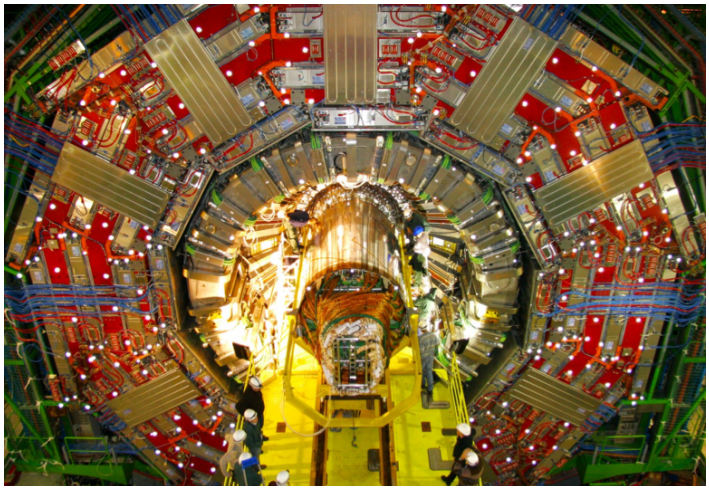
AT A GLANCE:

Customer: High energy physics researchers from leading international institutions conducting experiments at CERN (Conseil Européen pour la Recherche Nucléaire)

Industry: Scientific Research

Employees: CERN has more than 2500 operations staff with over 17,500 global scientific collaborators representing 110 nationalities from institutes in 70 countries.

https://home.cern/



Project: https://hls-fpga-machine-learning.github.io/hls4ml/

CHALLENGE:

Processing massive quantities of high energy particle physics data at extremely fast rates to find clues to the origins of the universe. Filtering sensor data in real-time, known as the trigger, to identify novel particle substructures that could contain evidence of the existence of dark matter and other physical phenomena.

SOLUTION:

Xilinx acceleration of AI inference together with performance-critical sensor pre-processing. Open-source hls4ml software tool with Xilinx Vivado® High Level Synthesis (HLS) that accelerates machine learning neural network algorithm development and deployment on Xilinx Virtex Ultrascale+™ FPGAs installed in the Large Hadron Collider (LHC) at CERN.

RESULTS:

Extremely low-latency neural-network inference on the order of 100 nanoseconds. Significant improvement in trigger identification capabilities through machine learning, along with a dramatic reduction in the time required to create and test specialized trigger algorithms.

CHALLENGE:

Optimize Trigger Filter Algorithm Development
CERN is one of the premier scientific research institutions where a global community conducts studies in fundamental physics to better understand what comprises the universe and how it works. The LHC at CERN is the place where sub-atomic particles are accelerated and "smashed" into their elemental parts revealed by an array of detectors and sensor systems. High energy particle physics experiments at CERN, like the recent observation of the Higgs boson, are the key to advancing the frontiers of human knowledge about the universe.

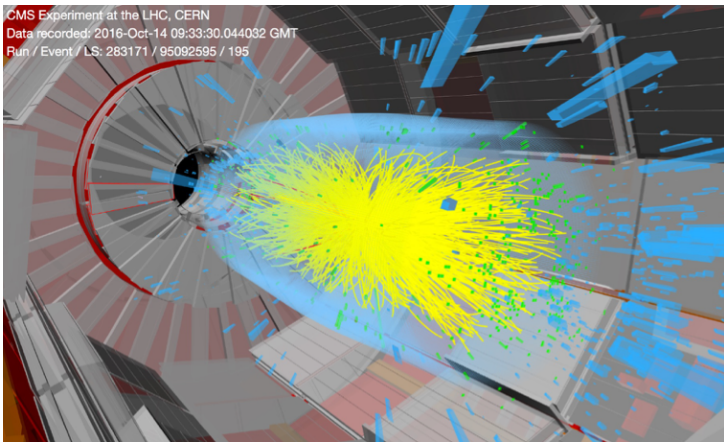


Figure 1: CMS event data from collisions happening every 25 nanoseconds.

A growing team of physicists and engineers from CERN, Fermilab, MIT, UIC, and UF led by Philip Harris, MIT and Nhan Tran, Fermilab, wanted to have a flexible way to optimize custom event filters in the Compact Muon Solenoid (CMS) detector they are working with at CERN. The very high data rates (150 Terabytes/second) in the CMS detector requires event processing in real-time, but trigger filter algorithm development hindered the team's ability to make progress. (see figures 1 and 2) Custom event triggers typically required many months of development.
Harris explained the project's genesis, "We were inspired after talking to a few people who had been working on machine learning with FPGAs from the Microsoft brainwave team and seeing on Github some very simple machine learning inference code written by EJ Kreinar using Xilinx' Vivado HLS tool. The combination of those two got us very excited because we could actually see the potential to do this hls4ml project to enable fast ML-based event triggers."

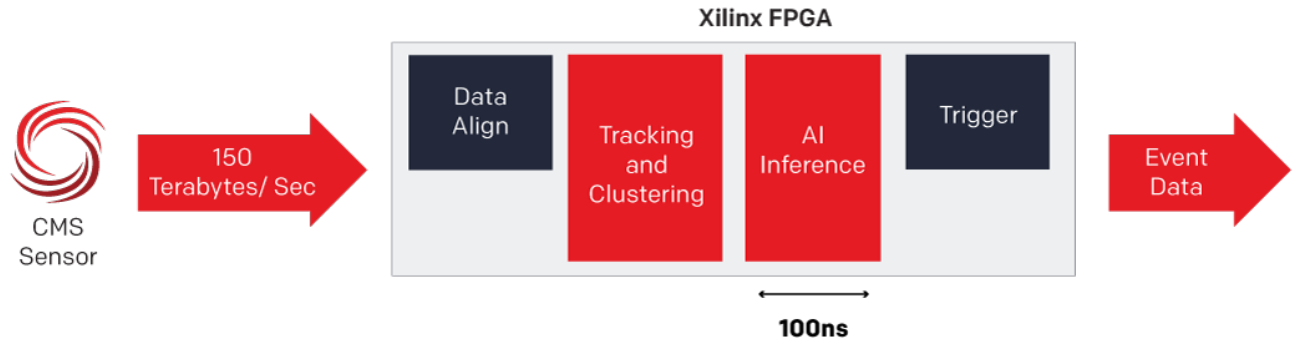


Figure 2: Compared to alternative devices such as GPUs and ASICs, FPGAs are the only viable choice for the event trigger processing because they provide extremely low latency. While the large numerical processing capability of GPUs is attractive, these technologies are optimized for high throughput, not low latency.

SOLUTION:

Developed a Customized hls4ml Flow Leveraging Vivado HLS

The team set out to develop and benchmark a tool flow (see figure 3), based around Xilinx Vivado HLS, that would shorten the "time-to-physics" for creating machine learning algorithms for the CMS level one trigger. The hls4ml tool has a number of configurable parameters that enable users to customize the space of latency, initiation interval, and resource usage tradeoffs for their application. Because every application is different, the goal of hls4ml is to empower users to perform this optimization through automated neural network translation and FPGA design iteration.

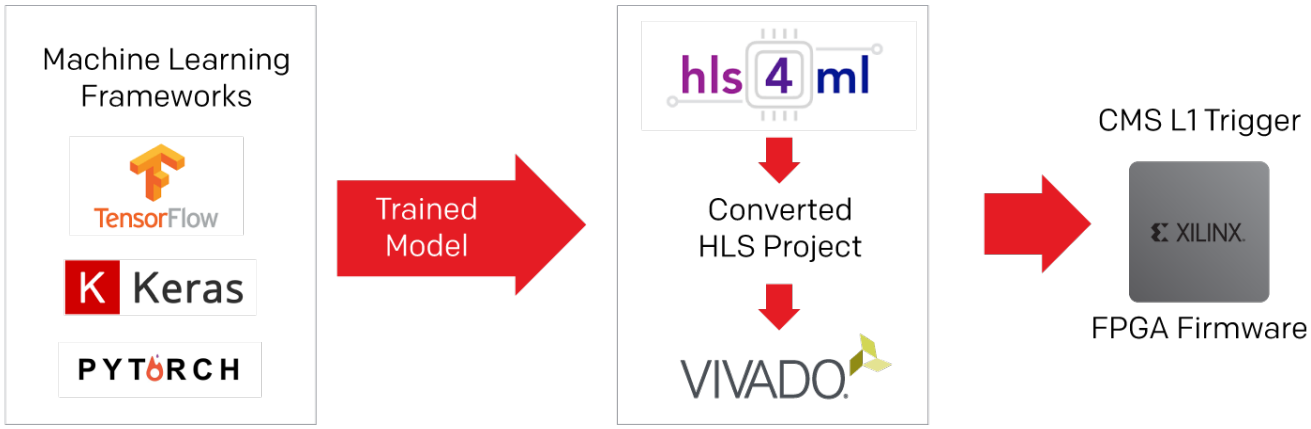


Figure 3: hls4ml tool flow

Prior to the team's work to create hls4ml, physicists would have to manually create simple trigger algorithms and engineers would then program the FPGAs in Verilog or VHDL. This was a very time-consuming process that could take several man-months of work by expert physicists and engineers.

Tran said, "We envisioned at a very high level putting neural networks into the level one trigger. Nobody had really considered the possibility of generically putting neural networks of all different types there. Once you give that capability to the community, then it can be everywhere. We're seeing it in muon identification, tau leptons, photons, electrons-- all the particles that we see--we can improve the performance using these more sophisticated techniques."

Raising the level of abstraction with hls4ml allows the physicists to perform model optimization with big data industry standard open source frameworks such as Keras, TensorFlow or PyTorch. The output of these frameworks is used by hls4ml to generate the FPGA acceleration firmware. This automation was a big time saver as Tran stated, "Electrical engineers are a scarce resource in physics and they're quite expensive. The more we can use physicists to develop the algorithms and electrical engineers to design the systems, the better off we are. Making machine learning algorithms more accessible to the physicist helps a lot. That's the beauty of why we started with HLS and not the Verilog or VHDL level. Now, we can do the whole chain from training to testing on an FPGA in a day."

How do the physicists search for dark matter using machine learning algorithms when they don't know what it actually looks like in order to train the neural networks?

Harris answered, "We make a hypothesis for what it will look like and write down a list of all the signatures we would expect for dark matter." According to Harris and Tran, there is a rich history of doing this in high energy physics. Tran added, "We're training on a very generic class of signatures. For example, dark matter by nature will be missing energy in the detector because it will go right through it. If we can use machine learning techniques to optimize the performance to understand missing energy that improves our sensitivity to dark matter as well."

RESULTS:

Achieving 100ns Inference Latency on 150 Terabytes/Second Data Rates

The data rate of the CMS detector is staggering and what makes the trigger filtering problem such a unique challenge. To overcome these challenges, extremely low-latency inference times are produced by the team's machine learning algorithms running on the Xilinx FPGAs. The data rates coming into the CMS are measured in hundreds of terabytes/second. The FPGAs receive and align sensor data, perform tracking and clustering, machine learning object identification, and trigger functions, before formatting and delivery of event data.

"Whether it's low-level aggregation of hits in some calorimeter all the way up to taking the full event and optimizing for a particular topology. It allows the spread and adoption of machine learning more quickly across the experiment."

The team uses multi-layer perceptron neural networks with a limited number of layers to meet the 100 nanosecond, real-time performance requirements of triggers. In addition to AI inference, the FPGAs provide the sensor communications, data formatting, and pre-filtering compute required for the incoming raw sensor data prior to the inference driven trigger, thereby accelerating the whole detector application.

Tran summarized hls4ml project benefits, "In our day-to-day work, it really allows us to access machine learning at every level across the experiment with the trigger. Before, you would have to think about a very specific application and work really hard on developing the model and the firmware for either VHDL or Verilog. Now, you can think more broadly about how we can improve the physics, whether it's low-level aggregation of hits in some calorimeter all the way up to taking the full event and optimizing for a particular topology. It allows the spread and adoption of machine learning more quickly across the experiment."

Asked if he expected his team's adoption of machine learning techniques might bring more practitioners into the high energy physics field, Harris concluded, "This has traditionally been the case. Even with the Higgs discovery there was a lot of machine learning involved. People like to come to the field because you can do cutting edge machine learning and deal with very large-scale computing problems. There is a lot of interest and people are excited about employing some of these techniques."

Additional Resources:

EJ Kreinar's Vivado HLS Presentation

hls4ml Jet Substructure Benchmark Paper