Case Study

High Performance Computing (HPC)
Intel® Xeon® Scalable Processors
Intel® Deep Learning Boost
oneAPI

Researchers from Intel, CERN and SURFsara Demonstrate Significant Deep Learning Inference Speedup

No loss of accuracy[1] using Intel Xeon Scalable processors with Intel Deep Learning Boost and oneAPI; Performance gains will be critical for handling future dramatic increase of data from Large Hadron Collider experiments

Training GANs and using Intel DL Boost to accelerate via quantization without sacrificing accuracy opens up exciting new possibilities for all applications that use Monte Carlo simulations

Executive Summary

In seeking to accelerate simulation workloads, researchers made good use of techniques that are likely to have wide applicability for accelerating Monte Carlo based simulations and deep learning inferencing in general. Their results helped highlight the effectiveness of the AI acceleration capabilities present in Intel Xeon Scalable processors.

Challenge

Physicists and researchers at CERN, the European Organization for Nuclear Research, utilize a unique range of particle accelerator facilities to study the most basic constituents of matter—fundamental particles. The Worldwide LHC Computing Grid, a global collaboration of more than 170 computing centers in 42 countries, is used for simulation and also to analyze and store the vast amount of data generated by this research.

Engineers watch as the small wheel Muon chamber is lowered into the cavern in preparation for recent large-scale general-purpose ATLAS experiments. The chamber detects the collisions which provide the data for analysis.
Photograph used with permission © CERN

In order to help address future needs for CERN's LHC (Large Hadron Collider—the world's largest particle accelerator), researchers at CERN, SURFsara, and Intel have been rethinking approaches for supplying extraordinary new levels of Monte Carlo based simulations. Future upgrades to the LHC will result in dramatically increased particle collision rates. Following collisions at the LHC experiments, calorimeters measure the energy a particle loses as it passes through the detector. Interpreting data from calorimeters is done through Monte Carlo based simulations which effectively reconstruct the collisions.

The researcher team wanted to accelerate a deep learning inferencing workload that held the promise of yielding results much faster than Monte Carlo based simulations. This work is being carried out as part of Intel's long-standing collaboration with CERN through CERN openlab. CERN openlab is a public-private partnership, founded in 2001, which works to help accelerate innovation in Information and Communications Technology (ICT). Today, Intel and CERN are working together on a broad range of investigations, from hardware evaluation to HPC and AI.

Solution

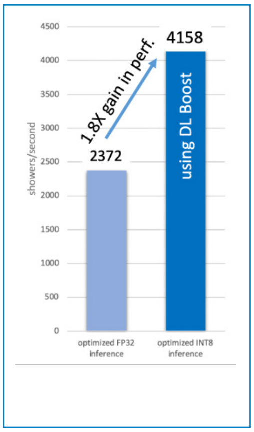
The researchers used the Intel AI Analytics Toolkit to obtain higher performance for inferencing in Intel Xeon Scalable processors with Intel Deep Learning Boost (Intel DL Boost). Intel DL Boost extends the AVX-512 instruction set to deliver significantly more efficient inference acceleration for deep learning workloads.

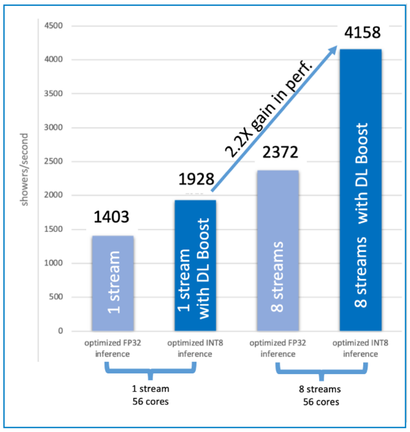
Furthermore, these investigations provide insights on how to accelerate models dependent on Monte Carlo simulations that could be useful in many other fields.

The built-in AI acceleration provided by Intel DL Boost was central to the project's performance gains. Intel DL Boost was shown to accelerate inferencing without sacrificing accuracy.

Results

Researchers demonstrated performance gains by simulating a calorimeter for a potential future particle accelerator—using a conditional generative adversarial network (GAN)—with only a fraction of the compute resources previously needed. Their approach of training GANs, and using Intel DL Boost to accelerate via quantization without sacrificing accuracy, opens up exciting new possibilities for all applications that use Monte Carlo simulations.





(Figure 2, left) Quantization led to a 1.8X[1] speed up by utilizing Intel DL Boost (specifically INT8 computations) on an Intel Xeon Platinum 8280 processor, and it shows slightly improved accuracy as well.

(Figure 3, right) Multistreaming the inferencing boosted performance 2.2X[1] on an Intel Xeon Platinum 8280 processor with Intel DL Boost.

This work has widespread implications. As Dr. Sofia Vallecorsa, a physicist specializing in AI and Quantum research at CERN observes, more than half the computing in the Worldwide LHC Computing Grid is used for simulation. Performance, cost, and accuracy are all critically important in the deployment of their trained model.

As illustrated in Figure 2, the team saw 1.8X gains for their complex GAN model inferencing. It shows slightly improved accuracy as well (lower is better: INT8 accuracy of 0.05324 vs. FP32 accuracy of 0.061227).[1]

Quantization led to a 1.8X speed up by utilizing Intel DL Boost (specifically INT8 computations) on an Intel Xeon Platinum 8280 processor, and it shows slightly improved accuracy as well.[1]

Solution Summary

In order to adopt their model to use Intel DL Boost without any loss of accuracy, the researchers at CERN used the Intel Low Precision Optimization Tool, which is a new open-source Python library that supports automatic accuracy-driven tuning strategies. The tool helps to accelerate deployment of low-precision inferencing solutions on popular DL frameworks including TensorFlow, PyTorch, MXNet, etc. The tool is available on the GitHub site and is included in Intel AI Analytics Toolkit along with Intel-optimized versions of TensorFlow, PyTorch, and pre-trained models to accelerate deep learning workflows. Figure 4 shows the flow used during the automated quantization auto-tuning.

CERN researchers found that about half of the computations in their network could switch from float32 to INT8 numerical precision, as supported by Intel DL Boost, without loss of accuracy. They saw nearly a doubling in performance[1] as a result. That matches the expectation that a complete conversion from float32 to INT8 could yield up to a theoretical maximum 4X gain in performance because of additional computational performance and reduction in memory bandwidth. With half the network converted, it makes sense that slightly less than a 2X gain was achieved when 4X was the theoretical maximum for a complete conversion.

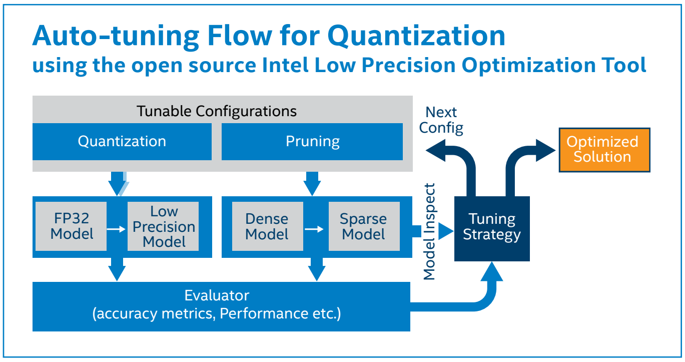
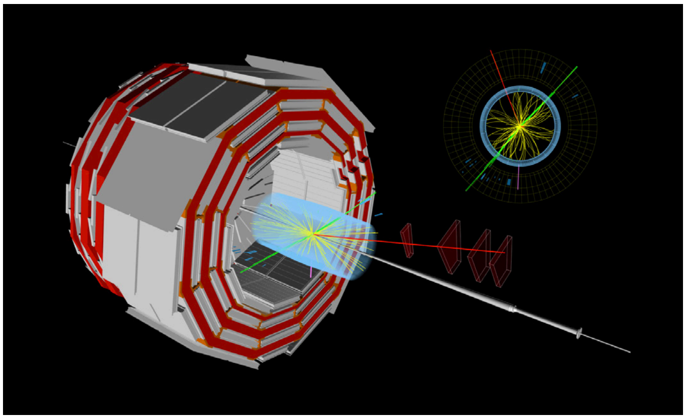


Figure 4. Quantization is achieved with full control over accuracy trade-offs, resulting in significant performance gains for inferencing



This display, similar to one of the two large general-purpose experiments that are famous for discovery of the Higgs, showcases the Compact Muon Solenoid (CMS) detector as well as the candidate event in which three W bosons are produced. Image used with permission © CERN.

> *The Intel DL Boost support found in Intel Xeon Scalable processors was shown to be a good fit for accelerating inferencing, without sacrificing accuracy.*

It is important to note that this significant gain came without sacrificing accuracy. A complete conversion to INT8 would give better performance, but with a loss of accuracy that this team did not want for their application. Quantization is an important technique made relatively easy thanks to tools supporting automatic accuracy-driven tuning. This allows users to achieve performance boosts while managing accuracy to whatever level is desired.

Quantization is proving to be an effective way to accelerate inferencing, and Intel Xeon Scalable processors with built-in support for AI acceleration (Intel DL Boost) with INT8 shows just how powerful this can be. Performance was nearly doubled compared with the prior 32-bit. Accuracy was maintained thanks to the open-source quantization tool.

FP32 and INT8 inferencing were both optimized for multicore. Valeriu Codreanu, Head of High-Performance Computing and Visualization at SURF, explains this performance optimization: “Since inferencing is less computationally expensive than training (as only the generator part of the GAN is being used), the hardware efficiency when using multiple cores in this process is not optimal. To overcome this, we have used multistream quantized inference, achieving a speed-up of 2.2X[1] compared to single-stream quantized inference, using the same Intel Xeon Platinum 8280 system.” This is illustrated in Figure 3.

Multistreaming the inferencing boosted performance 2.2X[1] on an Intel Xeon Platinum 8280 processor with Intel DL Boost.

Key parts of tools used, including the acceleration tucked inside Tensorflow and Python, utilize libraries with oneAPI support. That means they are openly ready for heterogeneous systems instead of being specific to only one vendor or one product (e.g. GPU).

oneAPI is a cross-industry, open, standards-based unified programming model that delivers a common developer experience across accelerator architectures. Intel helped create oneAPI and supports it with a range of open source compilers, libraries, and other tools.

By programming to use INT8 via oneAPI, the kind of work discussed in this case study could be carried out using Intel X^e GPUs, FPGAs, or any other device supporting INT8 or other numerical formats for which they may quantize.

Solution Ingredients

Intel High Performance Computing

Intel Xeon Scalable processors

Intel DL Boost

Intel Low Precision Optimization tool

Video presentation “Increasing AI Inference with Low-Precision Optimization Tool with Intel Deep Learning Boost–A High Energy Physics Use Case” by Haihao Shen (Intel) and Dr. Sofia Vallecorsa (CERN openlab).

CERN paper, “Reduced Precision Strategies for Deep Learning: A High Energy Physics Generative Adversarial Network Use Case”, to be presented at the 10th International Conference on Pattern Recognition Applications and Methods in February.

CERN GAN work

[1] CERN paper “Reduced Precision Strategies for Deep Learning: A High Energy Physics Generative Adversarial Network Use Case”, to be presented at the 10th International Conference on Pattern Recognition Applications and Methods in February. http://www.icpram.org/

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.