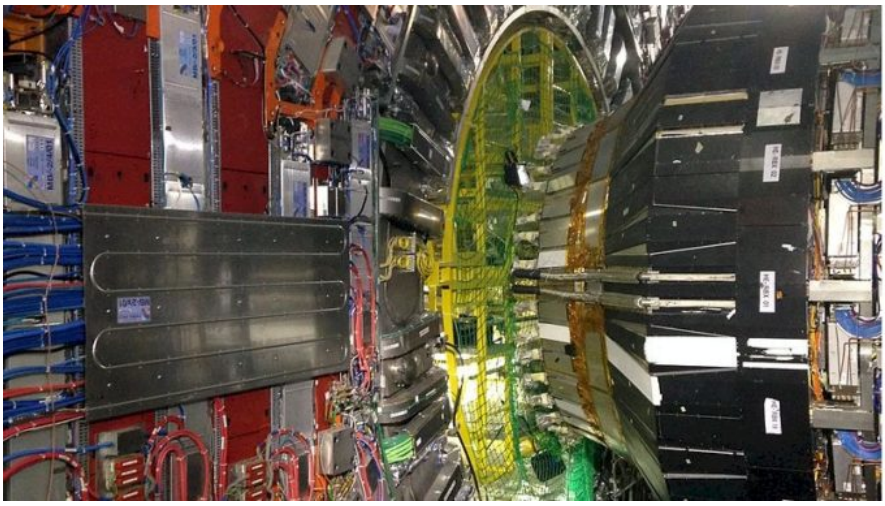
CERN USES DLBOOST, ONEAPI TO JUICE INFERENCE WITHOUT ACCURACY LOSS

February 1, 2021 James Reinders



Investigations, conducted together with scientists at CERN, show promising results – with breakthrough performance – in their pursuit of faster Monte Carlo based simulations, which are an important part of many scientific, engineering, and financial applications. In order to help address the future needs of CERN's LHC (Large Hadron Collider), which is the world's largest particle accelerator), researchers at CERN, SURFsara, and Intel have been investigating approaches for supplying extraordinary new levels of simulation.

Early results from a promising approach that relied on trained Generative Adversarial Networks (GANs) were first revealed at the International Supercomputing Conference in 2018, where it was awarded **best poster** in the category "programming models and systems software."

Now, they have demonstrated success in accelerating inferencing nearly two-fold by using reduced precision without compromising accuracy at all. Their technique highlights a general approach to inferencing acceleration – that is supported by Intel Xeon SP processors today, and by GPUs coming to market. A paper outlining this work will be presented this week at the 10th **International Conference on Pattern Recognition Applications and Methods**.

WHY FASTER SIMULATIONS MATTER TO CERN

Simulations, critical to data analytics, consume vast amounts of computing for Monte Carlo computations. Dr. Sofia Vallecorsa, a CERN physicist specializing in AI and Quantum research, has sought to ease the vast amount of compute currently consumed for Monte Carlo simulations. She observes, for instance, that more than half of the computing in the Worldwide LHC Computing Grid (WLCG – a global collaboration of more than 170 computing centers in 42 countries) is used for simulation.

Future plans to upgrade CERN's LHC will dramatically increase particle collision rates. This makes it is important to investigate new frameworks like this and assess their potential in helping to ensure computing requirements remain manageable.

A team of researchers at CERN, SURFsara, and Intel, are investigating the use of deep learning engines for fast simulation. 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.

> *If you want your inferencing to run faster, or your Monte Carlo operations to run faster, then results emerging from CERN are worth examining very closely.*

The project team has found a way to encapsulate expensive computations into a neural net via training, and extract –through inferencing – a result much faster than the standard algorithmic approaches of a classical Monte Carlo simulation. The team successfully simulated a calorimeter for a potential future accelerator – using a conditional generative adversarial network (GAN) – using only a fraction of the compute resources previously needed.

The key test for their resulting work is that their machine-learning-generated distribution is indistinguishable from other high-fidelity methods in physics-based simulations. Having achieved this, the team at CERN has shown a way to achieve orders of magnitude faster performance.

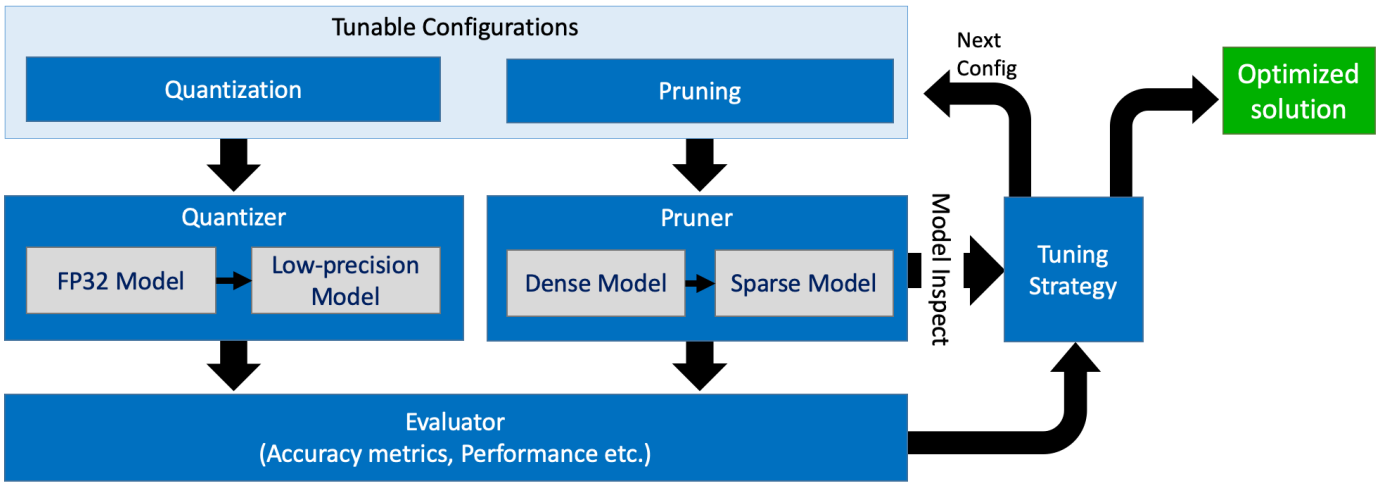
INFERENCE WITH AS LITTLE PRECISION AS POSSIBLE, BUT NO LESS

In 1950, the *New York Times* reported that Albert Einstein observed that "Everything should be made as simple as possible, but no simpler."

Such is the challenge when you use reduced precision data types within a neural network. Many papers have explored training networks using less and less precision. The motivation is performance, the challenge is loss of accuracy.

One can envision two approaches to using INT8 data type: one is to train a network using INT8, and another is to train using FP32 data type and then quantize to INT8. The latter has the advantage that the quantization can be selective in quantizing only parts of the network that do not adversely affect accuracy. The quantization is achieved by iterative trials using reducing precision, and measuring the resulting change in accuracy. This feedback loop is illustrated below and is guided by a tuning strategy we can control.

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 speed up deployment of low-precision inferencing solutions on popular DL frameworks including TensorFlow, PyTorch, MXNet, and so forth. In addition to the GitHub site, it is included in **Intel AI Analytics Toolkit** along with Intel optimized versions of TensorFlow, PyTorch, and pre-trained models to accelerate deep learning workflows.



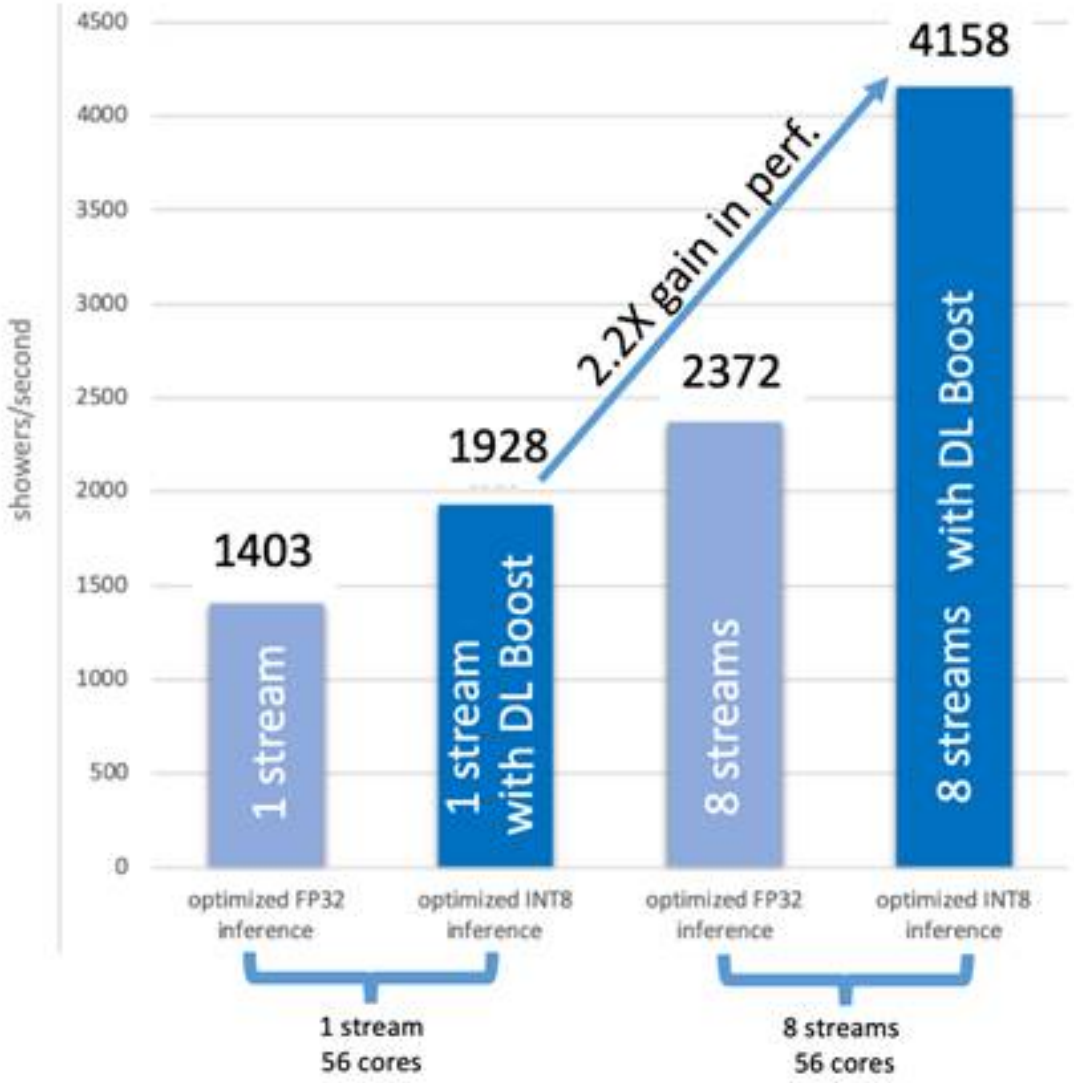
Quantization is achieved with full control over accuracy trade-offs, resulting in significant performance gains for the inferencing

In their work, they found that about half of the computations in the GAN could switch from FP32 to INT8 numerical precision, as supported by Intel DL Boost, without loss of accuracy. They saw nearly a doubling in performance as a result. That matches our intuition since we expect a complete conversion from FP32 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 a little under a 2X gain was achieved when 4X was the theoretical maximum for a complete conversion.

It is important to note that this significant gain comes without sacrificing accuracy. A complete conversion to INT8 would give better performance, but with a loss of accuracy. Quantization is an important technique made relatively easy thanks to tools supporting automatic accuracy-driven tuning. This allows us to achieve performance boosts while managing accuracy to whatever level we wish. The CERN team wished to maintain accuracy and, as illustrated below, they still saw 1.8X in gains from quantization alone for their complex GAN model inferencing. It shows better accuracy as well (lower is better: INT8 accuracy of 0.05324 versus FP32 accuracy of 0.061227).

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. (See the 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.)

Both FP32 and INT8 inferencing were previously 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 compared to single-stream quantized inference, using the same Intel Xeon Platinum 8280 system.” This is illustrated below:



Multistreaming the inferencing boosted performance 2.2X on an Intel Xeon SP-8280 Platinum processor. The version using Intel DL Boost with eight streams was significantly (1.8X) superior to other versions.

QUANTIZATION, INTEL DL BOOST, AND ONEAPI

Quantization is proving to be an effective way to accelerate inferencing, and Intel Xeon Scalable processors 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 inferencing while maintaining accuracy. The maintaining of accuracy is possible thanks to the open-source quantization tool used to manage accuracy to the needs of the developer. There are many more details of the work in their paper referenced above.

INT8 has broad support thanks to Intel Xeon SP processors, and it is also supported in Intel X^e GPUs. FPGAs can certainly support INT8 and other reduced precision formats. Quantization methods offer effective ways to use powerful hardware support in many forms.

> *The secret sauce underlying this work and making it even better: oneAPI makes Intel DL Boost and other acceleration easily available without locking in applications to a single vendor or device*

It is worth mentioning how oneAPI adds value to this type of work. Key parts of the 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 done at CERN described in this article 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.

SUMMARY

Boosting performance for inferencing has wide applicability, and is within reach thanks to Intel Xeon Scalable processors with Intel DL BoostAdditionally, 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.

The work at CERN to accelerate data analytics for probabilistic inference by quantizing and then using Intel Xeon Scalable processors is remarkably effective. It highlights why quantization has been gaining interest and will definitely become the norm for performance tuning inferencing workloads. Intel Xeon Scalable processors with Intel DL Boost were designed by engineers who foresaw this trend and Intel already has strong support in the processors today for this important inferencing acceleration capability.

LEARN MORE LINKS. . . .

- 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 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. **http://www.icpram.org/**
- Intel Low Precision Optimization Tool (for quantization): **https://github.com/intel/lp-opt-tool/**

James Reinders likes fast computers and the software tools to make them speedy. With over 30 years in High Performance Computing (HPC) and Parallel Computing, including 27 years at Intel, he is also the author of ten books in the HPC field as well as numerous papers and blogs.