TruePoint™ Technology.

Stop Compromising Accuracy for Performance.

Legal Statements

TruePoint™ Technology.

Groq Tech Doc

Overview

Groq TruePoint™ technology is what powers numerical representation in Groq's core matrix processing functional unit, which performs the bulk of GroqChip™ accelerator calculations common in artificial intelligence (AI), machine learning (ML), deep learning (DL), high performance computing (HPC), and linear algebra computing and applications. Supporting a variety of standard reduced bit formats today (and an expanding format set in the future), TruePoint utilizes less power and memory while giving customers the performance boost they seek without compromising accuracy, a typically forced tradeoff.

Uniquely, the GroqChip accelerator contains both floating point (FP) and integer numeric "primitive" formats, making it possible to both train and deploy inference with the same device. Most solutions train on one device and infer on another, using distinct hardware types with dissimilar numerics. GroqChip architecture uses TruePoint to harmonize the development flow between training and inference, achieving the benefits of mixed numerics without compatibility issues. Truepoint allows conversion from 32bit models to lower precision (16bit models for instance), without undermining accuracy, thereby impacting training and convergence time - all on a single device, providing a recipe for significant reduction in Total Cost of Ownership (TCO).

The need for more precision has limited options to migrate away from FP32 (single-precision floating point format) and FP64 (double-precision floating point format) to available lower precision formats for large problems. While these numbers are more precise and, if well-managed, can lead to greater accuracy, handling single- and double-precision numbers requires a huge amount of processing power. Moving from, say FP16 to FP32, typically requires a 4x increase in power.

Because floating point and fixed point multiplication are cornerstones of compute, TruePoint technology focuses on these operations, minimizing error accumulation where it makes the biggest impact, while preserving precision. TruePoint delivers the benefit of data accuracy typical with higher precision numerics, along with the performance and power savings of standard, reduced precision inputs.

Three Points That Matter

1

TruePoint fuses, multiplies, and rounds one time over large numbers of operations

2

TruePoint achieves unprecedented high accuracy with standard, reduced precision numerics

3

TruePoint reduces time to solution with a significantly better TCO

Benefits

Velocity

Unlike traditional architectures that train and infer in different environments, due to differences in deeply-embedded numerics, the GroqChip both trains and deploys inference with the same device. The result is greater time efficiency and development speed.

Accuracy

TruePoint enables better accuracy at lower input bit sizes which results in better performance, lower power, and less memory usage for your workloads. The benefit for large complex iterative algorithms is less accumulated error; for example, a training algorithm can see faster convergence.

TCO

TruePoint delivers more efficient cost of ownership: leveraging simpler numerics is faster, while delivering high performance and low power, plus reduced development and processing time by eliminating the accumulation of avoidable error in the algorithm result.

Use Cases

HPC Convergence to ML Methods

Today, High Performance Computing is embracing AI workloads. Instead of trying to simulate giant datasets one after another, they take smaller datasets and develop models that are trained against data, then run millions of inferences to discover new properties. These smaller models employ machine learning, typically with FP32 rather than FP64. Developers are reluctant to reduce precision further for fear of compromising accuracy on these highly sensitive models. TruePoint was designed to enable the use of reduced precision for accuracy sensitive applications that are highly iterative and tend to accumulate residual error. Large convolution layers or dense matrix multiply are common examples where error can accumulate. TruePoint technology handles these specific types of computations without loss of precision seen on other devices like a GPU or CPU.

Computational Science

High Performance Computing has become the go-to platform for genomics, computational chemistry, and seismic imaging. Years ago, HPC was primarily used in research, but today it's gaining wider adoption with availability in public clouds. A scientist employing HPC is often decomposing an algorithm using large 64bit, highly precise data. One challenge is the ability to scale performance using these large bit size numerics - which are in short supply for most CPU and GPU based cloud systems. But where they can, they would like to selectively reduce precision where possible, without compromising the system level results. Often this requires costly iterative refinement processing to recoup accuracy losses. For these highly iterative algorithms TruePoint can reduce or even completely eliminate the need for iterative refinement steps. Furthermore, Groq has analyzed HPC algorithms and demonstrated the ability to increase performance (via lower precision) without compromises to accuracy (See Figures 2 & 3). This ability to achieve high accuracy and high performance has broad application across all sciences that require extremely high precision, from oil and gas exploration, to weapon simulation, to biotech.

By the Numbers

FP32-level Accuracy at FP16 Precision

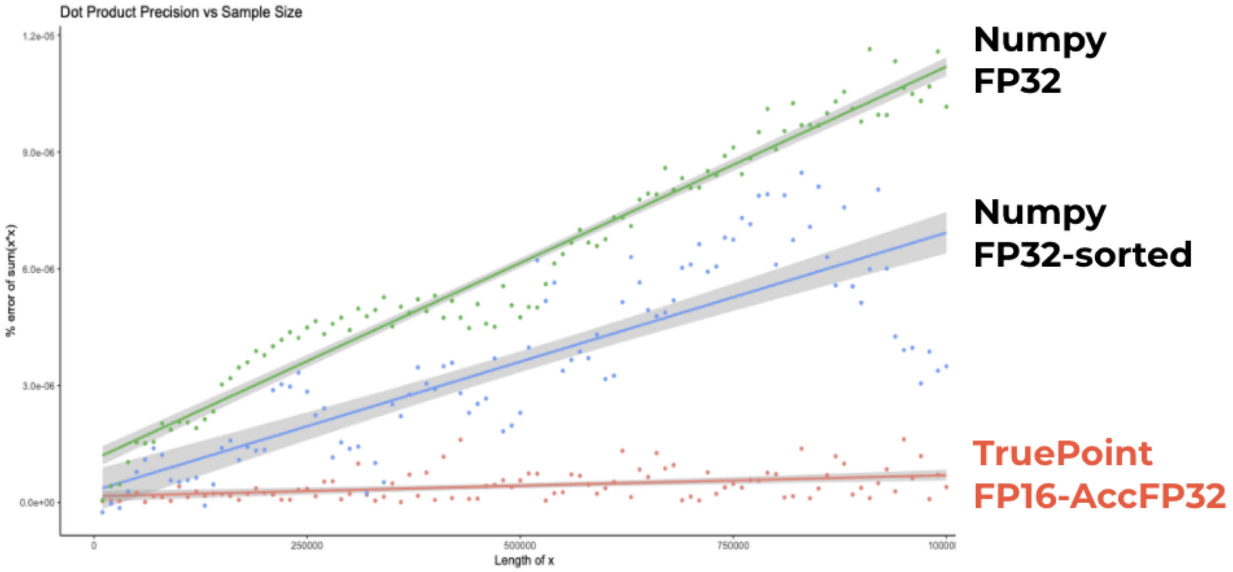
After quantization, losses can continue to accumulate through a series of discrete computations. TruePoint takes advantage of mixed-precision in a 320-element fused dot product with a single rounding step, each dot product then accumulated in FP32. For FP16, GroqChip is able to effectively retain higher precision than standard FP16 or FP16 scaled data with a FP32 data type.

GroqChip architecture is deterministic, which allows one to predict the order of operations and create a testbed that more accurately expresses the actual numerical evaluation of the program. The result is FP32 accuracy from FP16 multipliers. With GroqChip architecture, TruePoint fuses, multiplies, and rounds one time, delivering premium accuracy.

The graph in Figure 1 shows how the rate of accumulated error is significantly lower for TruePoint than that of FP32 on the host, indicating a robust improvement in precision. In fact, TruePoint outperforms standard IEEE FP32 over long compute lengths (for data pre-scaled within FP16 ranges).

Residual MSE vs Dot Product Length

ML workloads can take advantage of lower-precision numerics like FP16 or INT8 for quantized models



Linear fits with 95% confidence intervals shown (robust improvement in precision).
Compares against inputs in FP32 but within the range of FP16 values (remove quantization error effects).
Sorted line shows best-case FP32 MSE assuming deterministic compute, like GroqChip™.
Compared against FP64 oracle.

Figure 1. Groq TruePoint technology outperforms standard IEEE FP32 over long compute lengths.

Drug Discovery in Minutes at Argonne National Laboratory

The graph in Figure 2 and table in Figure 3 show how Groq TruePoint™ technology was the key to maintaining superior accuracy while exploiting the efficiency of lower precision, further enhancing performance of Argonne National Laboratory's SARS-CoV-2 drug discovery workload. At FP32, Groq was able to achieve 185x throughput versus the baseline (Figure 2). When reducing precision to FP16, Groq not only boosted relative throughput from 185x to 333x, but also maintained almost the same accuracy, with an 0.02% average error of the highest prediction and 0.05% max average error versus the FP32 baseline (Figure 2 & Figure 3). Learn more about Groq's work with Argonne National Laboratory here.

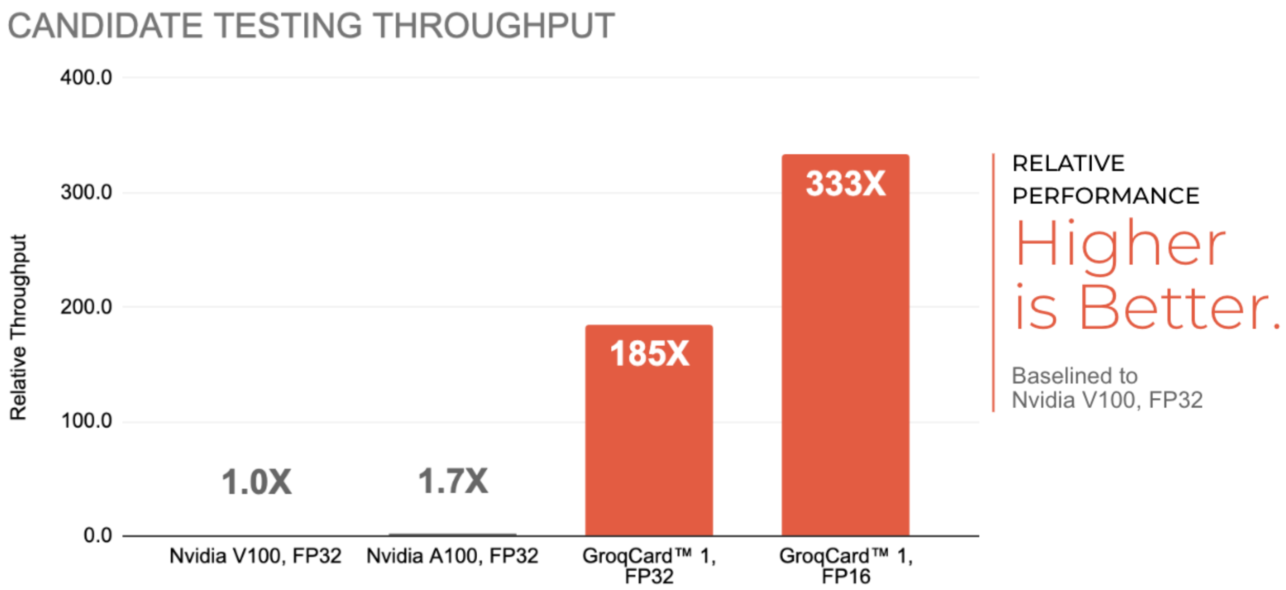


Figure 2. GroqCard 1 FP16 delivers up to 333X relative throughput versus the baseline as measured by Argonne National Laboratory at their Argonne Leadership Computing Facility AI Testbed.

Groq TruePoint™ FP16 vs FP32 Base
Relative Error Across 6.5M Compounds and 32 Models. Lower is Better.

Average Error of Highest Prediction	Max Average Error
0.02%	0.05%

Figure 3. GroqCard TruePoint™ FP16 delivers an 0.02% average error of the highest prediction and 0.05% max average error (versus FP32 baseline).

Interested in Learning More?

For more information on Groq technology and products, contact us, follow us on YouTube and Twitter, and connect with us on LinkedIn.

Additional Related Documents

- Latency Tech Doc
- Predictability Tech Doc
- Scalability Tech Doc
- Velocity Tech Doc