kinara

A Technology White Paper

Measuring AI Accelerator Performance in Real Life

By Rehan Hameed

Co-Founder & CTO at Kinara

What is the Best Way to Measure Performance?

When comparing today's Edge AI accelerators, performance is a key metric for selecting the best accelerator for a given inference task. The best way to measure performance is to run a specific workload, typically ResNet-50, EfficientDet, a Transformer, or a custom model. However, this not always feasible. Thus, Trillions of Operations Per Second (TOPS) is often used as a single-unit performance metric.

While TOPS is easy to calculate for most systems, it falls short in assessing an accelerator's actual AI performance. In fact, real-world performance can be significantly lower than the TOPS value due to system-level inefficiencies and the accelerator's inability to maximize the workload's parallelism. This white paper explains how the real-world performance of Deep Vision's Ara-1 far exceeds its theoretical TOPS value.

What are TOPS?

TOPS measures the number of arithmetic operations AI accelerators can perform per second at 100% utilization. Traditional industry practice bases this calculation on the accelerator's maximum operating frequency and number of concurrent multiplications and additions hardware can perform, since these operations form the core of many processing algorithms such as matrix multiply, filtering, and convolutions.

Consider an accelerator with 2048 multipliers, 2048 adders, and a 1GHz peak operating frequency. If all multipliers and adders are active every cycle, this accelerator can perform (2048+2048) * 1 billion arithmetic operations per second, i.e., 4 TOPS. As the next section reveals, TOPS can be completely unreliable performance indicator on real-world applications.

TOPS can be an unreliable performance indicator on real-world applications.

TOPS as a Performance Measure

AI accelerators achieve a fraction of the theoretical TOPS ceiling due to factors such as idle compute units waiting for data from memory; synchronization overhead between different parts of the accelerator; and control overheads. Depending on the accelerator's architecture and workload characteristics, an accelerator might achieve only 5-10% of its theoretical TOPS value.

As an example, the Google TPU (Table 1) is a 92 TOPS machine but the LSTM1 workload utilizes only 2.8 TOPS - just 3% of the theoretical 92 TOPS! In fact, half the workloads utilize 10% or less of the peak TOPS. On the other hand, the CNN0 benchmark yields over 90% utilization, highlighting how the TOPS utilization is highly dependent on the AI algorithm.

Operations	Neural Network Applications					
	MLP0	MLP1	LSTM0	LSTM1	CNN0	CNN1
Active Compute Cycles (%)	12.5%	9.4%	8.2%	6.3%	78.2%	22.5%
Overhead Cycles (Compute Inactive)	12.5%	90.6%	91.8%	93.7%	21.8%	77.5%
Effective TOPS	**12.5%**	**9.7**	**3.7**	**2.8**	**86.0**	**14.1**

Table 1. Operation utilization of Google's TPU, derived from Google's paper [1] for various neural network workloads. The 'Active Compute Cycles' row gives the percentage of time the multiply-and-accumulate (MAC) units are active doing useful work. The inactive cycles are derived by adding weight stall cycles, weights shift cycles, non-matrix cycles, RAW stalls, input data stalls and unused MACs.
The effective TOPs show how much of the 92 peak TOPs are actually used.

An accelerator's architecture and software tools play an important role in determining how well the accelerator utilizes its computation resources while executing a specific workload. Two different AI accelerators with the same number of TOPs (e.g., 10 TOPS) can offer widely different performance for a neural network depending on the architecture's efficiency and how well the accelerator's compiler manages and schedules data movements through the system to minimize idle time and energy consumption. As a result, a well architected 5 TOPS accelerator can outperform an inadequately implemented 10 TOPS accelerator.

Moreover Table 1 shows that even for the same accelerator, TOPS utilization varies widely for different workloads with more than a 10x gap in utilization between the best and worst workloads. This is particularly the case because AI accelerators are often fixed function designs optimized for specific workload characteristics and perform poorly when the workloads differ from design assumptions.

Higher TOPS Equates to Higher Cost and Power

An argument can be made that even if TOPS is not an indicator of absolute performance, it is an indicator of relative performance, suggesting that an accelerator with a higher TOPS rating is better than an accelerator with a lower TOPS number. However, the reality is often the opposite.

Higher TOPS means a larger accelerator with more compute elements as well as more memory blocks to feed data to those compute units. This results in higher cost (i.e., die size) and power. An efficient accelerator offers high performance using a lower number of compute resources and thus has a lower TOPS rating compared to a less efficient design which requires more TOPS to deliver the required performance. Therefore, contrary to popular belief, a desirable AI accelerator is the one which provides high performance using low TOPS.

TOPS Doesn't Include All Computation Types

The TOPS metric only considers an accelerator's multipliers and adders. However, an accelerator can have other computation resources beyond multipliers and adders, and this is definitely the case with Deep Vision's architecture. For example, Deep Vision's architecture employs reduction trees instead of an adder array, resulting in significantly lower area and energy consumption - the TOPS metric fails to capture the reduction tree's computation capability. Similarly, other hardware structures such as transpose engines, lookup table memories, data shifters, and non-linearity computation blocks, all accelerate important parts of the AI computational workload but are not captured by the multiplies and additions of the TOPS metric. These points highlight that TOPS is an inadequate measure for assessing an AI accelerator's performance and we need better metrics to measure real-world performance. Standard neural networks such as ResNet50, MobileNet V1, and YOLO_v3, are useful for analysis when comparing different accelerators. These standard networks can also be used as a proxy for 'guesstimating' whether a given accelerator can meet the demands of a developer's own workloads.

Performance Metrics for Edge AI

The TOPS metric only considers an accelerator's multipliers and adders. However, an
An accelerator's efficiency can vary widely for different workload types. Therefore, we recommend using networks of different types, sizes, topologies, and input resolutions, to understand (Table 2). But an important point to note is that real-time processing dictates that inference latency is the most important performance metric (the accelerator's execution time to complete one inference of a specific AI model).

Model	Application	Size (Parameters)	Topology Complexity	Input Size
ResNet-50	Image Classification	26 million	Low	224x224
EfficientNet-B0	Image Classification	5.3 million	Medium	224x224
Yolo V3	Object Detection	62 million	Medium	416x416
FCN	Semantic Segmentation	11 million	Medium	1024x768
BERT Base	Natural Language Processing	110 million	High	Seq-length = 128
3D UNet	3D Image Segmentation		High	160x224x224

Table 2. A collection of networks that form the basis for a good benchmarking set.

Inference latency is the right metric to evaluate AI accelerator performance.

Inferences-Per-Second (IPS) indicates how many inferences of a model the accelerator completes per second. The problem is that this is a throughput measure and doesn't indicate the time for each inference. For example, if an AI accelerator runs ResNet50 at 200 IPS, you might be misled into thinking that each inference takes 5 msec to complete, however, the latency can be an order of magnitude higher than what you might expect. Therefore, inference latency is the right metric, as opposed to IPS.

Final Thoughts—Build Smart, Efficient Accelerators

While TOPS is an 'easy' metric to calculate, it is not a reliable performance indicator for real world workloads. Since TOPS only accounts for the number of multipliers and adders in an accelerator, the metric completely misses the mark for architectures such as Deep Vision's which include many other computational hardware structures for processing neural network models.

Benchmarking based on either publicly available models or developers' own applications offer a credible and more reliable alternative to TOPS.

An even greater downside of using TOPS is that it incentivizes building larger, power-hungry accelerators with more compute elements, rather than building smarter, more efficient accelerators. In turn, this leads to 'TOPS wars' in a manner similar to how CPU vendors engaged in GHz wars, before realizing more efficient, lower frequency CPUs were the better approach. Benchmarking based on either publicly available models or developers' own applications offer a credible and more reliable alternative to TOPS, and AI accelerators should provide ample tool support to facilitate these efficient evaluations. Contact Deep Vision today to get the details on how our accelerator leaves the TOPS contest winners behind and how we come out ahead on delivering real-world, real-time performance and high-quality tools. You can reach us at sales@deepvision.io.

01. https://arxiv.org/pdf/1704.04760.pdf

02. https://www.intel.com/content/www/us/en/gaming/resources/read-cpu-benchmarks.html

03. https://mlcommons.org/en/

KINARA | LEADING EDGE AI

4410 El Camino Real STE 110, Los Altos, CA 94022