SemiEngineering Blog – 18
December 2018

Lies, Damn Lies, and TOPS/Watt

There are almost a dozen vendors promoting inferencing IP but none of them gives even a ResNet-50 benchmark.

The only information they state typically is TOPS (Tera-Operations/Second) and TOPS/Watt.

Let's discuss why these two indicators of performance and power efficiency are almost useless by themselves.

What does X TOPS really tell you about performance for your application?

When a vendor says our ABC-inferencing-engine does X TOPS, you would assume that in one second it will perform X Trillion Operations. Let's show you how a 130 TOPS engine actually delivers 27 TOPS of useable throughput.

First, what is an operation? Some vendors count a multiply (typically INT 8 times INT 8) as one operation and an accumulation (addition, typically INT 32) as one operation. So a single Multiply-Accumulate = 2 Operations. But be careful, some vendors include other types of operations in their TOPS specification.

Second, did the vendor specify operating conditions for their TOPS spec? If not, it's likely they are basing the number on typical conditions: room temperature, nominal voltage and typical process. Usually they will mention which process node they are referring to, say 16nm, but operating speeds differ between different vendors. And most processes are offered with 2, 3 or more nominal voltage. Since performance is a function of frequency, and frequency is a function of voltage, you can get more than twice the performance at 0.9V than at 0.6V.

Let's look at how frequency varies depending on the conditions/assumptions: MACS that run at 2 GHz typical (typical process, 0.8V, 25C) would run at about 2.3GHz at 0.9V or 2.5GHz at 0.99V (0.9V + 10%). But the synthesis worst case number (using Slow-Slow process corner, 0.72V, and +125C junction temperature), the value given before place-and-route, would be about 1.5GHz. But after place-and-route, when actual metal routing and loads are factored in, the actual worst case performance would be about 1GHz (Slow-Slow, 0.72V and +125C junction).

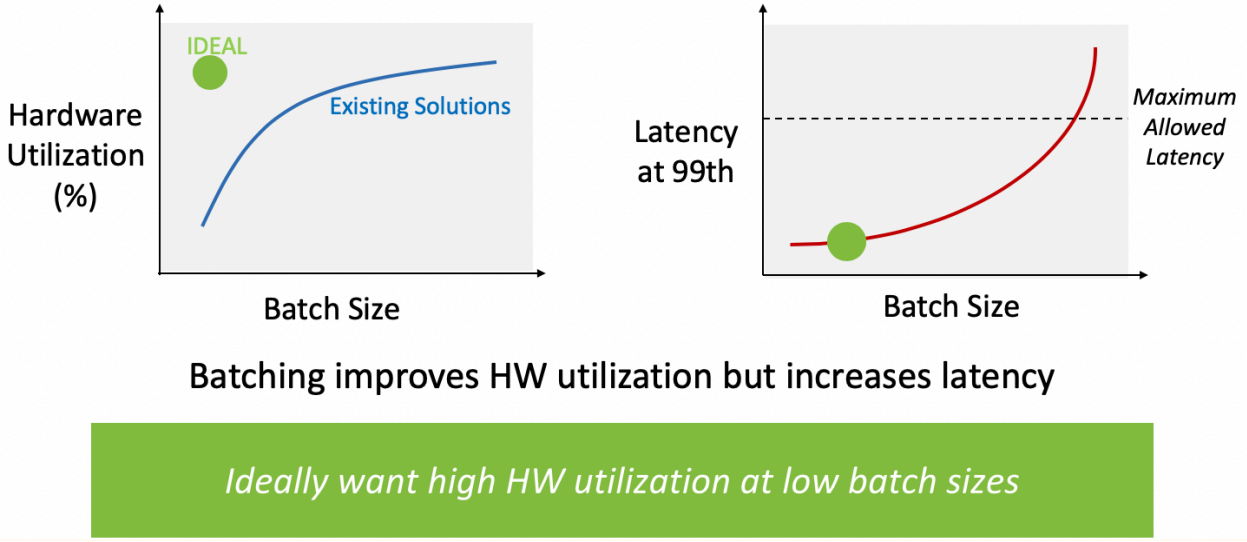
So if the ABC-inference-engine has 1000 MACs, they could say they have 2.5 TOPS, 2.3 TOPS, 2 TOPS, 1.5 TOPS or 1 TOPS depending which of the frequencies they are using. If they don't give specific conditions, assume they are giving you optimistic numbers.

Third, let's assume the vendor has given you worst-case TOPS: the worst-case frequency times the number of available MACs. Do all of those operations actually contribute to computing your neural network model? No. No inferencing engine has 100% utilization of all of the MACs all of the time. In fact, the actual utilization can be very low.

Consider batch size. Batching is where weights are loaded for a given layer and process multiple data sets at the same time: the reason to do this is to improve throughput, but the give-up is longer latency. ResNet-50 has over 20 Million weights; YOLOv3 has over 60 Million weights; every weight must be fetched and loaded into the MAC structure for every image. There are too many weights to keep them all resident in the MAC structure.

As Microsoft said at their HotChips talk, the ideal inferencing engine would have high hardware utilization at Batch = 1: few do.

For example, Habana's Goya does 15,012 images/second for ResNet-50 at batch = 10. But at Batch =1 their performance drops more than 50% to 7,107 images/second.



So to understand the actual performance of any neural inference engine, you must know what batch size they used in giving you performance information.

Fourth, Not all neural networks behave the same. What is the actual MAC utilization for the neural inference engine for the neural network model you want to deploy, at the batch size you require.

Almost no one gives even a ResNet-50 benchmark and if they do they often don't give batch size (if they don't, you can assume it will be a large batch size that maximizes their hardware utilization %). Only a few suppliers give ResNet-50 benchmarks with batch size information.

Nvidia Tesla T4's web page lists their inferencing capacity as 130 TOPS. On ResNet-50 it benchmarks at batch size = 28 as processing 3,920 images/second (image size = 224x224 pixels). We know that ResNet-50 requires 3.5 Billion MACs/image = 7 Billion Operations. Telas T4 actually performs 3920 images/second x 7 Billion Operations/image = 27,440 Billion Operations/second = 27.4 Trillion Operations/Second = 27.4 TOPS. As a result, 130 TOPS is

actually 27.4 TOPS of real throughput = <25% hardware utilization. What about batch =1? They don't say, but it's likely much less.

And ResNet-50 isn't really a very helpful benchmark, some people call it a "toy benchmark". YOLOv3 for example requires 100 times more operations to process a 2 Megapixel image. Hardware utilization will be even more challenged on "real world" models.

We haven't even talked about the effect of

- Winograd Transform: if they use it, they may state the number of TOPS they would have done if they hadn't used it, inflating their MACs by 2x or more
- Model pruning: if they use it, they may state the number of TOPS based on the model before pruning
- Larger multiplies: if they do say a 16x8 multiply in hardware, they could count this as two 8x8 multiplies in their TOPS
- Layers: some layers have higher utilization than others, they may be stating TOPS for the best layer
- Sparsity: they may skip MACs for highly-sparse arrays and report TOPS as if they hadn't

Conclusion: if a vendor says they deliver X TOPS without stating the conditions, you don't know the process corner, the voltage, the temperature, the batch size or the model. So TOPS is pretty useless.

What you should ask your vendor is: how many images/second can you process for your model, say YOLOv3, at batch size = A & at XYZ PVT conditions. Only if you specify all of the conditions and assumptions do you get useful data. And only if you get the same data from all your vendors can you really do intelligent comparisons.

What does TOPS/Watt really say about power efficiency?

We've already seen that TOPS can be computed many different ways so TOPS/W is already suspect because of the numerator, TOPS.

But let's look at the divisor, Watts.

Again, do they state the assumptions and conditions? If not, likely they are giving you optimistic data.

They may be stating TOPS based on conditions that maximize throughput while computing Watts based on conditions that minimize power – of course, this is meaningless because you want to know TOPS and Watts for the same conditions.

They are likely stating TOPS based on 100% hardware utilization, but if they have done actual power calculations it will be based on their actual much lower hardware utilization, which will be lower power because they are doing less calculations.

For example, Nvidia Tesla T4 has a TDP (thermal design power) spec of 75W, this is a worst case number or close to worst case number.

Their 130 TOPS claim leads to 130 TOPS divided by 75W = 1.7 TOPS/Watt.

But we saw above that actual throughput in TOPS is 27.4 TOPS; this divided by 75W = 0.36 TOPS/Watt, quite a difference.

The Tesla T4 is an inferencing board with 8 DRAMs, so the power number includes the DRAM power which is a significant part of the total power.

Your vendor's TOPS/Watt number almost certainly does NOT include the power of DRAM, and most inferencing solutions require 100's of GB of DRAM bandwidth: Habana Goya and Nvidia Tesla T4 are both 8 DRAMs, 256-bit DRAM data bus width and about 320GB/second of peak DRAM bandwidth.

Excluding DRAM power from TOPS/W understates what matters to you: you care about the power of your total inferencing subsystem.

Also, watch out for weight compression: they may give you a very optimistic benchmark based on a very sparse weight matrix, but if you run a model with less sparsity your DRAM bandwidth and power in reality will jump.

Conclusion: Just like you should ask our vendor for images/second throughput for your model at your desired batch size and at specified PVT conditions, insist for the same for power and have them include DRAM power as well as inferencing-silicon power.

Geoff Tate, CEO, Flex Logix