Intelligence is Your Edge

Powered by Intel

congatec

Decoding AI Inferencing Hardware Performance: Usable Metrics Beyond TOPS

Written by Maximilian "Max" Gerstl, Product Line Manager at congatec

Benchmarking the AI inference performance of different pieces of AI hardware can be like comparing apples to oranges. How are you supposed to properly compare them? Let's clear up the confusion.

History repeating itself.

Remember the days when buying a new processor was all about which could crank up the GHz the most? It was a simpler time when you could just look at the clock speed to determine performance and make your decision. But then multi-core processors came along and changed the game. Suddenly, different processor architectures were not as easily comparable as before. It was not just about raw frequency anymore - it was about having the right processor technology for the specific tasks. Can this processor run my game smoothly? Can my Digital Audio Workstation (DAW) handle all the tracks I want to record and process in real time? And how fast will it apply filters to my photos when I am editing them? These were the questions that drove my buying decisions. And these questions could usually not be answered by pure GHz. Do I need the highest frequency CPUs, or should I invest in massive multi-core technology instead?

Many of our customers find themselves in the same situation today when it comes to AI use cases and required performance. How much AI inference performance do I need to fluently run tasks like computer vision, natural language processing, or autonomous navigation? There is a lot of confusion out there and there is not that one value that should drive your decision. So, let's take a detailed look of what should be considered to make a better and well informed buying decision for AI hardware.

TOPS – The answer to all AI-questions?

The biggest culprit of confusion around AI inference performance is undoubtedly the metric that all hardware manufacturers list - including congatec. Tera Operations Per Second or TOPS for short. TOPS is the theoretical computational peak performance of a processor. It is defined as a value, of how many trillions of operations silicon can perform in a second on an AI workload. As such, TOPS are a very simplifying metric and can be misleading, when you want to judge the true AI performance.

First of all, TOPS do not differentiate between the quality or type of operation that the hardware performs. Do you work with convolutional algorithms or does your application rely on transformer-based models? And what about the datatype of the operation – Do you need just INT8 precision, or do you need to process INT16 or even FP32? Also, memory bandwidth can have a massive affect in real-life applications limiting the real performance. Don't get me wrong: The metric has its use, in my opinion. Due to its simplicity, it makes a great ballpark figure to understand general differences between hardware if the underlying architecture is similar, but that's it. It most likely doesn't help you to determine if a specific silicon suits your application.

Complex workloads need specific answers

Let me work it out by the latest Intel Core Ultra Series 1 and Series 2 processor technology. These processors offer a unique combination of heterogeneous compute engines, including Central Processing Unit (CPU), your "conventional" processor, a Graphics Processing Unit (GPU) you can also use as a General Purpose GPU (GPGPU) for AI tasks such as image classification and the Neural Processing Unit (NPU). All three different compute architectures contribute to an overall AI performance that elevate the new Intel Core Ultra Processors Series 2 to the 100 TOPS class. The Neural Processing Unit (NPU) called Intel® AI Boost adds advanced neural processing capabilities to the overall computational architecture. The integrated NPU enables highly efficient integration of advanced artificial intelligence workloads at lower system complexity, power, and costs than discrete accelerators. In detail, it adds 13 TOPs. Next is the integrated Intel Xe-LPG+ graphics processor. It provides not only immersive graphics but can also be used to process parallel, high-throughput AI workloads, delivering up to 77 TOPS. Last but not least Intel® deep learning boost by Intel AVX512 Vector Neural Network Instructions (VNNI) accelerates inner convolutional neural network loops with up to 9 TOPS. These operations are run on the x86 P-Cores and E-Cores of the Intel Core Ultra processors.

Detailed benchmarks provide the answers.

As you can see: AI is undergoing the same development that the CPU benchmarks went through. Nowadays, you test performance with many metrics. For different use cases, different metrics are more or less useful. In the AI world, we look at the accuracy of a model and the inferencing time or throughput. The AI benchmarks that actually help to cut through all this noise, have to consider specific workloads like object detection or text classification.

Since more and more AI professionals are looking for hardware solutions that are more cost-effective alternatives to NVIDIA's offerings, understanding the true performance metrics becomes even more crucial. And the current possibilities of integrated AI accelerators for computer-on-modules are becoming more and more relevant.

No more guess work

In the rapidly evolving world of AI hardware, relying solely on metrics like TOPS can be misleading. To make informed decisions, it's crucial to benchmark AI hardware using specific workloads and models relevant to your applications. This approach ensures that the hardware you choose will meet your performance and accuracy requirements.

At congatec, we provide actual AI benchmarks based on Geekbench AI to provide a meaningful score and Inferences per second (IPS) values classified for the common datatypes of INT8, FP16 and FP32.This gives you a way sharper picture of the actual performance your AI inferencing application can reach. Sounds promising? Don't hesitate to reach out to us. We're here to help you find the best hardware for your AI needs.

Partner Name

congatec GmbH

Booth Info

Hall 3
Booth #241

embeddedworld
Exhibition&Conference