insight.tech POWERING IoT AND EMBEDDED INNOVATION

AI Benchmarks: A Barometer Before You Build

AI has spawned a new generation of chips designed to strike a balance between throughput, latency, and power consumption.

AI accelerators like GPUs, FPGAs, and vision processing units (VPUs) are optimized to compute neural network workloads. These processor architectures are empowering applications like computer vision (CV), speech recognition, and natural language processing. They are also enabling local AI inferencing on IoT edge devices.

But benchmarks show that these accelerators are not created equal. Choosing one can have serious implications on system throughput, latency, power consumption, and overall cost.

Understanding AI Inferencing

It's important to understand exactly what neural networks are and what's required to compute them. This will help clarify the benchmarks reviewed later.

Neural networking is a subset of AI technology modeled after the human brain. Rather than a single algorithm, neural networks are often a collection of multiple software algorithms organized into layers, like a cake.

Each layer analyzes an input data set and classifies it based on features learned during a training phase. After one layer classifies a specific feature, it passes the input on to a subsequent layer. In convolutional neural networks (CNNs), a linear mathematical operation (convolution) is performed one or more times to produce a cumulative expression of the layers.

In image classification, for example, the network would be fed a picture. One layer would classify a shape as a face. Another layer would classify four legs. A third could classify fur. After a convolution is applied, the neural network would eventually conclude that it's an image of a cat **(Figure 1)**. This process is known as inferencing.

A neural network processor must access input data from memory every time a new layer is computed. And this is where the tradeoffs begin.

The more layers and convolutions in a neural network, the more performance and high-bandwidth memory access required from an AI accelerator. But you can also sacrifice accuracy for speed, or speed for power consumption. It all depends on the needs of the application.

GPUs vs. FPGAs vs. VPUs

Throughput and latency benchmarks reveal how Intel® Arria® 10 FPGAs, Intel® Movidius™ Myriad™ X VPUs, and NVIDIA Tesla GPUs perform when running four compact image classification neural networks. The networks are GoogLeNetv1, ResNet-18, SqueezeNetv1.1, and ResNet-50.

Each processor is housed in an off-the-shelf acceleration card for real-world context:

Arria 10 FPGAs—This software-defined programmable logic device features up to 1.5 tera floating-point operations per second (TFLOPS) and integrated DSP blocks. It is represented in our benchmark by the IEI Integration Corp. Mustang F100-A10 AI Accelerator Card.

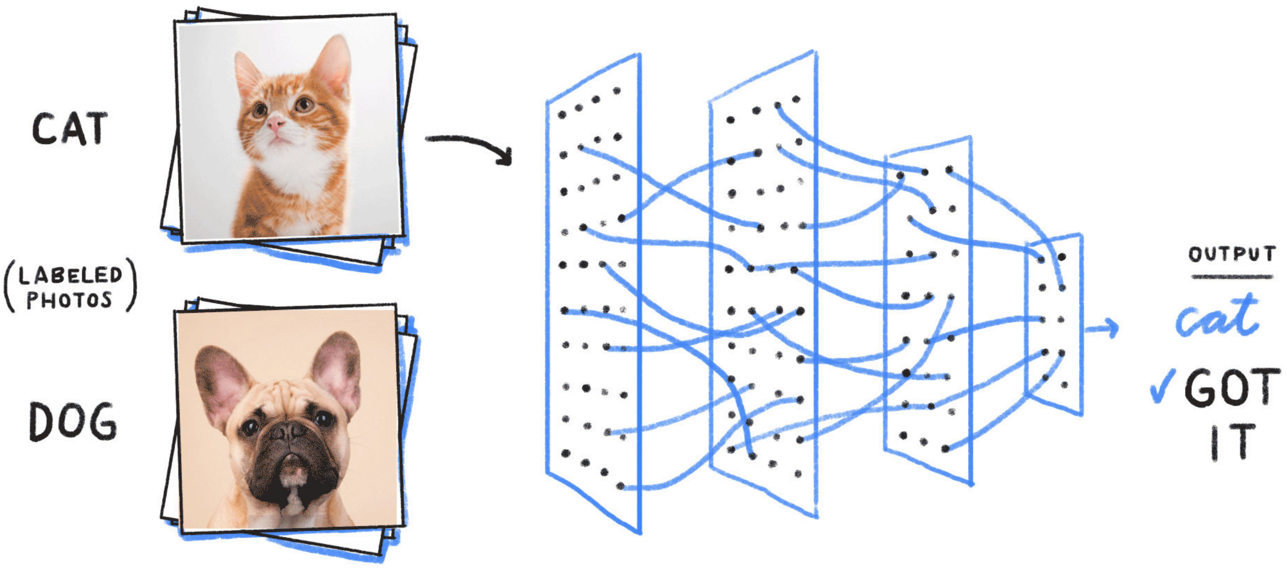


Figure 1. Each layer of a neural network analyzes input data to generate an output. (Source: Medium)

The Mustang F100-A10 includes 8 GB of 2400 MHz DDR4 memory and a PCIe Gen 3 x8 interface. These features support neural network inferencing on more than 20 simultaneous video channels **(Figure 2)**.

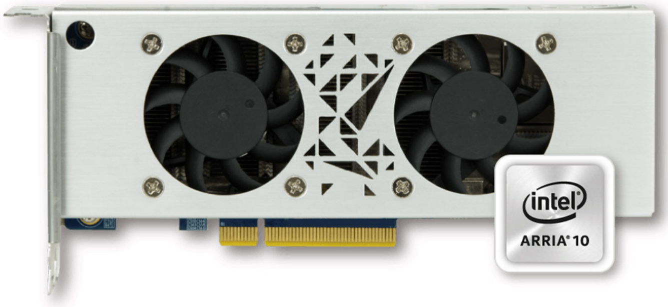


Figure 2. The IEI Mustang F100-A10 contains an Intel® Arria® 10 FPGA. (Source: IEI Integration Corp.)

Myriad X VPUs—These hardware accelerators integrate a Neural Compute Engine, 16 programmable SHAVE cores, an ultra-high throughput memory fabric, and 4K image signal processing (ISP) pipeline that supports up to eight HD camera sensors. They are included in the benchmark as part of the IEI Mustang-V100-MX8.

The Mustang-V100-MX8 integrates eight Movidius X VPUs, allowing it to execute neural network algorithms against multiple vision pipelines simultaneously **(Figure 3)**. Each of the VPUs consumes a mere 2.5 watts of power.

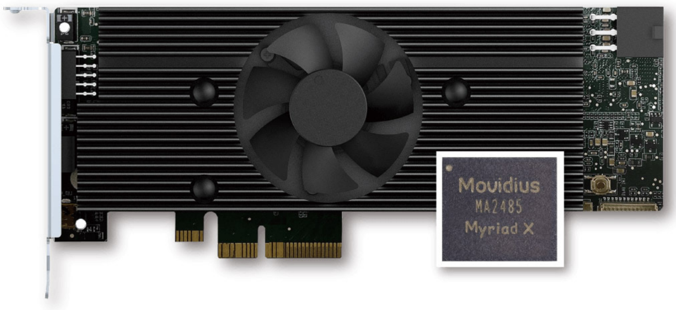


Figure 3. The IEI Mustang V100-MX8 contains eight Intel® Movidius™ Myriad™ X VPUs. (Source: IEI Integration Corp.)

NVIDIA Tesla GPUs—These inferencing accelerators are based on the NVIDIA Pascal architecture and provide 5.5 TFLOPS of performance at 15x less latency than CPUs. The NVIDIA P4 Hyperscale Inferencing Platform is the subject of the benchmark.

Figure 4 shows the benchmark results. Throughput is represented in terms of the number of images analyzed per second, while latency represents the time required to analyze each image (in milliseconds).

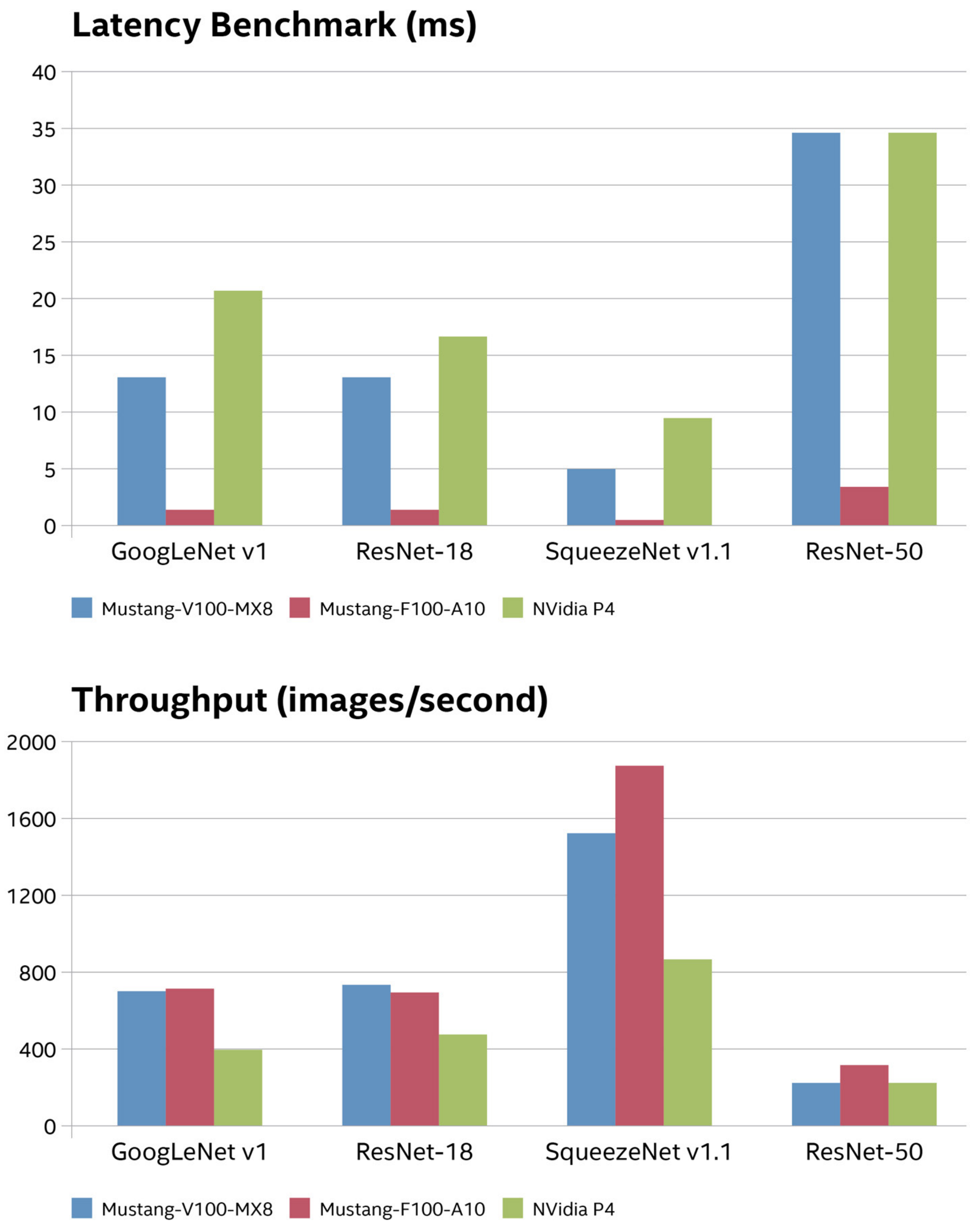


Figure 4. Latency (top) and throughput (bottom) benchmarks of the inferencing accelerators. (Source: IEI Integration Corp.)

The throughput and latency benchmarks reveal that FPGA and VPU accelerators perform considerably better than GPUs across the neural network workloads. And as shown in **Figure 5**, IEI Mustang products have much lower thermal design power (TDP) ratings and price points.

Behind the Benchmarks

The reason GPUs struggle in these small-batch processing tasks has a lot to do with architecture.

GPUs are typically organized into blocks of 32 cores, all of which execute the same instruction in parallel. This single-instruction, multiple-data (SIMD) architecture allows GPUs to burn through large, complex tasks more quickly than traditional processors.

But latency is associated with all of these cores accessing data from memory, which on the P4 is an external DDR5 SDRAM. In larger workloads, this latency is hidden by the fact that parallel processing can make up for it quickly by applying the performance of so many cores. In smaller workloads, the latency is more obvious.

In contrast, FPGAs and VPUs excel in smaller workloads because of their architectural flexibility.

Inside the Intel® Arria® 10 FPGA

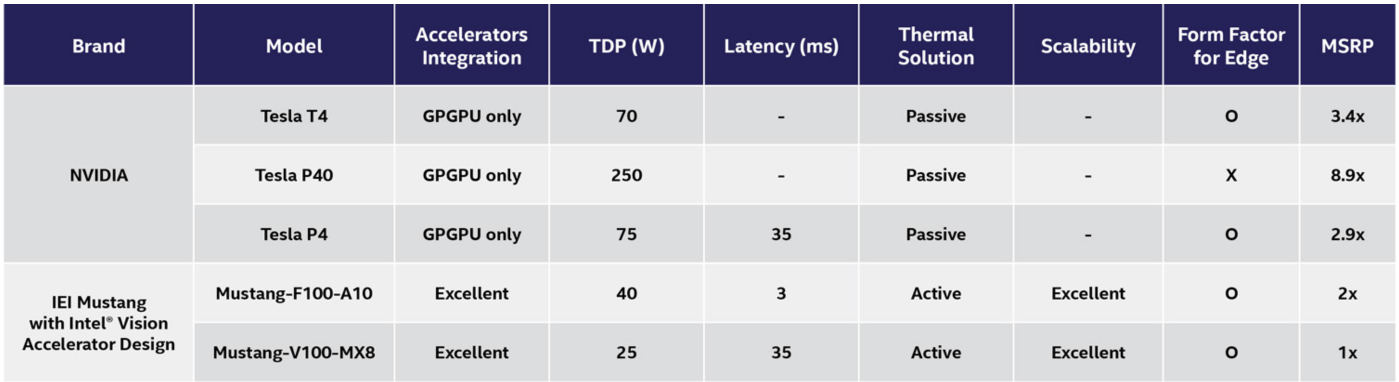
For instance, Arria 10 FPGA fabric can be reconfigured to support different logic, arithmetic, or register functions. And these functions can be organized into blocks of FPGA fabric that meet the exact requirements of specific neural network algorithms.

The devices also integrate variable-precision DSP blocks with floating-point capabilities **(Figure 6)**. This provides the parallelism of GPUs without the latency tradeoffs.

The ultra-low-latency is enabled by internal memory and high-bandwidth internal interconnects that provide logic blocks with direct data access. As a result, an Arria 10 device can retrieve and compute small-batch inferencing data much more quickly than GPUs, resulting in significantly higher throughput.

Flexible Compute and Memory with Intel® Movidius™ Myriad™ X VPUs

Meanwhile, the Myriad X VPU's Neural Compute Engine provides a dedicated on-chip AI accelerator. The Neural Compute Engine is a hardware block capable of processing neural networks at up to 1 tera operations per second (TOPS) with minimal power draw.



Brand	Model	Accelerators Integration	TDP (W)	Latency (ms)	Thermal Solution	Scalability	Form Factor for Edge	MSRP
NVIDIA	Tesla T4	GPGPU only	70	-	Passive	-	O	3.4x
	Tesla P40	GPGPU only	250	-	Passive	-	X	8.9x
	Tesla P4	GPGPU only	75	35	Passive	-	O	2.9x
IEI Mustang with Intel® Vision Accelerator Design	Mustang-F100-A10	Excellent	40	3	Active	Excellent	O	2x
	Mustang-V100-MX8	Excellent	25	35	Active	Excellent	O	1x

Figure 5. The IEI Mustang F-100-A10 and V100-MX8 consume considerably less power than alternatives. (Source: IEI Integration Corp.)

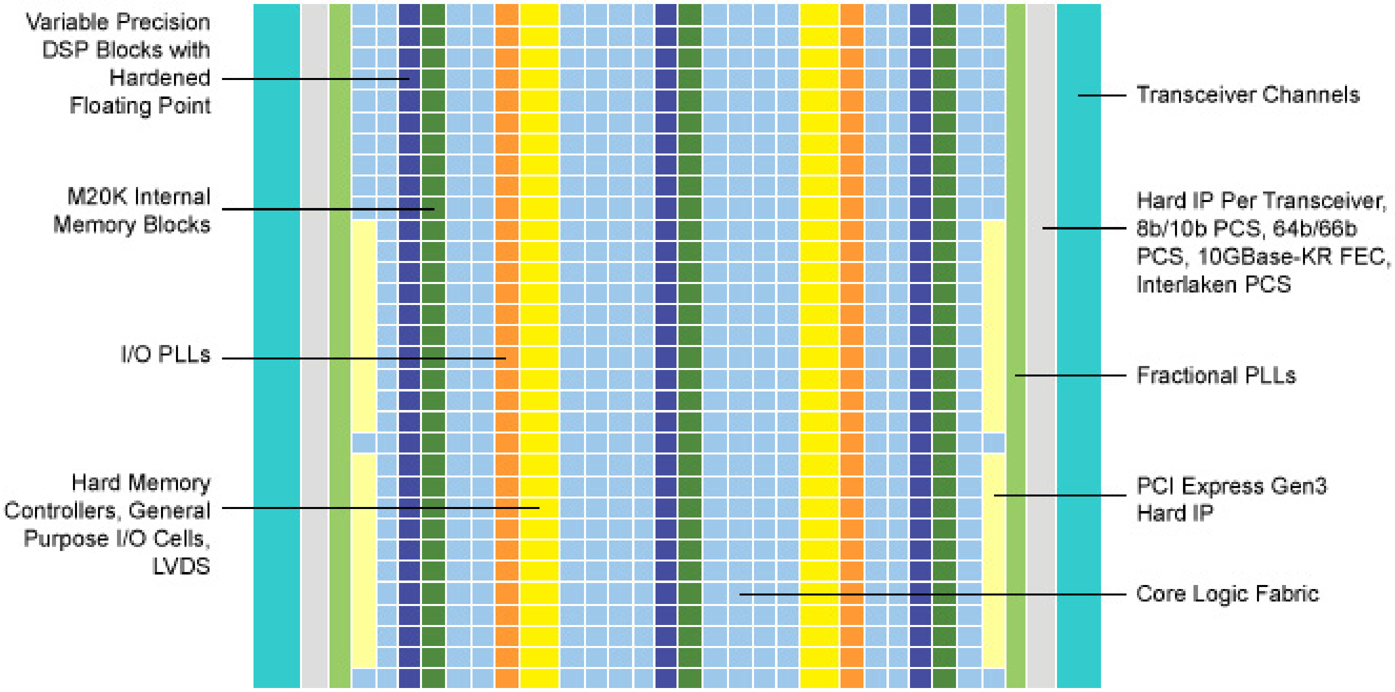

Variable Precision DSP Blocks with Hardened Floating Point
M20K Internal Memory Blocks
I/O PLLs
Hard Memory Controllers, General Purpose I/O Cells, LVDS
Transceiver Channels
Hard IP Per Transceiver, 8b/10b PCS, 64b/66b PCS, 10GBase-KR FEC, Interlaken PCS
Fractional PLLs
PCI Express Gen3 Hard IP
Core Logic Fabric

Figure 6. Intel® Arria® 10 FPGAs combine the parallelism of GPUs without the latency. (Source: Intel® Corp.)

The Neural Compute Engine is flanked by the 16 programmable SHAVE cores mentioned previously. These 128-bit vector processors combine with imaging accelerators and hardware encoders to create a high-throughput ISP pipeline. In fact, the SHAVE cores can collectively run multiple ISP pipelines at once.

Each component in the pipeline can access a common intelligent memory fabric **(Figure 7)**. So neural network workloads can terminate in the optimized Neural Compute Engine without expending the latency or power associated with multiple memory access.

Check the Benchmarks

This article has shown how innovation in chip architectures and hardware accelerators is enabling AI at the edge. While each architecture has its merits, it's critical to consider how these platforms impact the compute performance, power consumption, and latency of neural network operations and systems as a whole.

To that end, be sure to check the benchmarks before starting your next AI design.

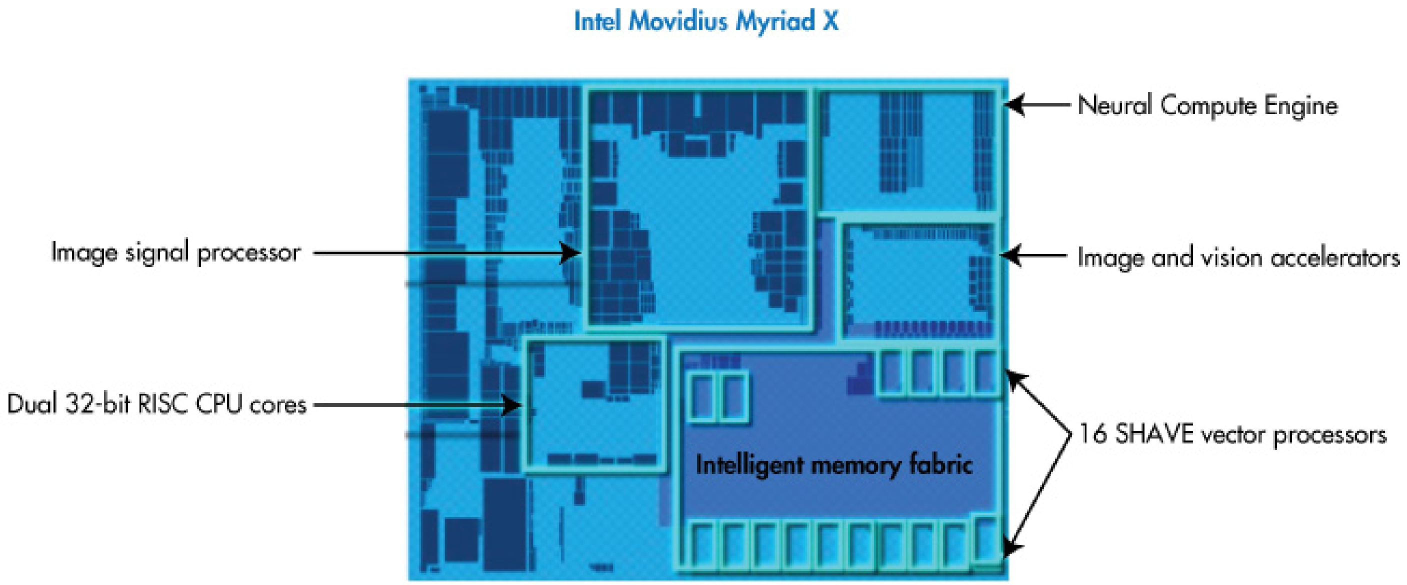

Intel Movidius Myriad X
Neural Compute Engine
Image signal processor
Image and vision accelerators
Dual 32-bit RISC CPU cores
16 SHAVE vector processors
Intelligent memory fabric

Figure 7. Myriad™ X VPUs contain a high-throughput image signal processing pipeline. (Source: MIPI Alliance)