AI Inference at the Rugged Edge: Meeting Performance with M.2 Accelerators

Balancing Power, Performance, Thermals, and Footprint Is the Next Hurdle in Data-Driven Applications
# Table of Contents

**Part 1**  
Data at the Rugged Edge: Enabling Capabilities, Accuracy and Speed Beyond Human Performance ......................................................... 4

**Part 2**  
Hitting an AI Wall: Compute Has Come a Long Way and AI Still Needs More ......................................................................................... 6

**Part 3**  
Domain-Specific Architectures (DSAs) With M.2 Accelerators: All Workloads Are Not Created Equal ............................................................. 9

**Part 4**  
Get To Know the M.2 Interface: A Compact, Versatile Next Generation Option ......................................................................................... 11

**Part 5**  
Throughput Matters: Understanding Benchmarks for Real-World AI Applications ......................................................................................... 15

**Part 6**  
Premio AI Edge Inference Products: Comparison Using Hailo M.2. Accelerators ......................................................................................... 17

**Part 7**  
Appendix A. Product Benchmarks ......................................................................................................................................................... 20

Appendix B. Hailo Modules Vs. Comparable GPUs ........................................................................................................................................... 25

Appendix C. Neural Network Comparisons (Total Power Vs. Frame Per Second) ......................................................................................... 26
The world of computing technology maintains a high bar for innovation, with systems continually becoming faster and more efficient. Think Moore’s Law and the steadfast expectation that CPU capabilities will naturally improve at a structured pace. And so they have…a reality amplified through smart design strategies developed to accommodate specific workloads. Developers and designers have found ways for systems to do more, pairing CPUs with floating point processors or adding GPUs to offload graphics-intensive processing applications like deep neural networks.

Today, groundbreaking applications like artificial intelligence (AI) and machine learning are bringing a new dose of reality to these design strategies. It’s not just the data-intensive nature of automation that is causing change – it’s where it is being implemented. As applications move out of the data center and into the world, more and more industrial and non-traditional computing settings are seeking greater competitive value from data in real-time. This can be defined as the rugged edge, and it is here that performance acceleration requires a new path.

Power and footprint are not the same steady values found in a data center, and neither are the workloads. In these scenarios, AI may keep drivers safe, factories online, supply chains moving, and so much more. Performance acceleration must balance power and costs with reliable and constant movement of data. These factors are contributing to M.2 technology’s increasing role as a performance accelerator optimized for the data-centric nature of AI and machine learning.

This paper will take a deeper look at performance acceleration at the rugged edge, and how modern hardware strategies can assure the reliable performance and smart data handling needed to fuel AI’s value further into modern life and enterprise business.

"AI and machine learning are paving the path for autonomous applications at the edge. Embedded systems help power these non-traditional compute systems."
Data drives business innovation, and most importantly, the ability for cognitive machine intelligence. On the factory floor, powering smart kiosks or advanced telematics, fueling surveillance and passenger services in infrastructure facilities like airports and train stations – data is everywhere and adds value when it can be revealed, captured, analyzed, and applied in real-time. These diverse industries are keen to capitalize on data to develop new services and improve business decisions. Yet for many of these applications performing in rigorous industrial environments, running small automated or AI tasks from a data center is simply too inefficient to add true value. In this traditional centralized compute structure, power and costs are too high. This is due to excessive, but necessary, use of compute, storage, and bandwidth resources. Performance trade-offs deepen the sacrifice, with factors such as high latency and insufficient data privacy.

Still, as the number of IoT (Internet of Things) and IIoT (industrial IoT) devices continues to increase, so does the volume and velocity of data they generate. This trend, combined with the realities of continually increasing numbers and types of connected devices, creates a wealth of new opportunities for purpose-built computing solutions. And one that also demands a different approach to hardware designs that enable optimized performance.

“Gartner predicts that by 2025, 75% of enterprise data will be processed outside a traditional centralized data center.”

Edge computing hardware is being deployed to cope with this increasing amount of data and to alleviate the ensuing burdens placed on the cloud and in data centers. Data-intensive workloads like AI inference and deep learning models are moving beyond the data center and into factories and other industrial computing environments. In turn, designers and developers are recognizing a current shift for performance acceleration closer to data sources such as IoT sensors. It’s a trend that is pushing edge hardware further, meeting the need to interact with AI workloads on-demand.
Acceleration Options: Flexibility and Workload Consolidation Is Pushing Edge Hardware Further

Performance acceleration may be dictated by a variety of hardware. There is a need for all these choices, however, it is key for designs to offer flexibility to support all options for workload consolidation. For example, depending on the specific software algorithm, an inference engine could lean on either CPU, GPU, or M.2 accelerators.

What are the fundamental differences between CPU, GPU, Google TPU (an M.2 format) and even new hardware accelerators? CPUs are processing units that function as the brains of a computer intended for general-purpose programming. A GPU is an alternative performance accelerator focused on enhancing computer graphics and AI workloads. CPU also are generally limited by their number of cores and are designed to process compute in a sequential fashion, whereas GPUs deploy many cores dedicated to parallel computing processes. TPUs are Google’s custom-developed processors that accelerate machine learning workloads using TensorFlow, a defined machine learning framework. Based on data growth coupled with the complexities of edge computing environments, today’s AI computing framework is moving from general CPU and GPU options to more specialized accelerators, such as smaller and more power-efficient acceleration modules in an M.2 standard.

<table>
<thead>
<tr>
<th>CPU</th>
<th>GPU</th>
<th>TPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Cores</td>
<td>Thousands of Cores</td>
<td>Matrix Processing Hardware</td>
</tr>
<tr>
<td>Low Latency</td>
<td>High Throughput</td>
<td>High Latency (Compared to CPU)</td>
</tr>
<tr>
<td>Serial Processing</td>
<td>Parallel Processing</td>
<td>Extreme Parallelism (Very High Throughput)</td>
</tr>
<tr>
<td>A handful of operations at once</td>
<td>Thousands of operations at once</td>
<td>Optimized for large batches</td>
</tr>
<tr>
<td>Large Memory Capacity</td>
<td>Low Memory Capacity</td>
<td>Convolutional Neural Network [CNN]</td>
</tr>
<tr>
<td>More Flexible Programs</td>
<td>Less Flexible Programs</td>
<td>AI Focused Programs</td>
</tr>
</tbody>
</table>

Source: Premio - What is M.2 Expansion Slot? The Future of Compact, Robust Technologies.
Accelerators provide the significant data processing necessary, filling the void caused by the slowing of Moore’s Law, a driving force of the electronics industry for decades. This long-proven principle states that the number of transistors on a chip will double every 18-24 months, yet when it comes to AI, industry experts point to Moore’s Law showing signs of wear. A less familiar principle, but one that also has impact today, is called Dennard scaling.

This observation by Robert Dennard indicated that as transistors get smaller, power density is constant. Power use remains in proportion with area, while voltage and current scale down with length. This theory ended about 20 years ago – yes, transistors kept shrinking but power current and voltage could not keep dropping and remain reliable.

At the same time, power budgets are not increasing, and mechanical and thermal performance still face limits. Additional cores per chip do not necessarily provide relief. The energy savings achieved by shifting from a single power-hungry processor to multiple, more efficient processors (i.e. cores per chip) has its limits as well. Amdahl’s Law states that any speedup in latency achieved by parallel computing is in turn limited by the sequential part of the task – this is particularly evident when the abundance of steady resources is unavailable in new mobile and industrial environments.

Clearly, at some point, limits prevail. All these factors mean that silicon evolution alone cannot support AI algorithms and the orders of magnitude greater processing performance they require. The necessary balance of performance, cost, and energy demands a new approach featuring more specialized domain-specific architectures.
Transferring more typically human tasks to computers and machines requires more data to support the process. Neural network processing demands become heavier in sync. This has created a surging demand for complex computing that is no longer satisfied by centralized processing in data centers. It’s a phenomenon propelling speedy evolution of AI inference at the edge, that is machine intelligence outside of the data center.

It’s useful to the conversation to gain an understanding of the elements that go into building a deep neural network, a process with two distinct phases. Deep learning training employs data sets to teach a deep neural network to complete an AI task, like image or voice recognition. Deep learning inference is the process of using the trained model, including feeding the same network with novel or new data, to predict what that data means based on its training.

**Training**: Deep learning training teaches the deep neural network to carry out AI tasks. This is accomplished by feeding the deep neural network data, allowing the network to make a prediction as to what the data represents.
Virtually all training is developed using floating point operations, a common benchmark for rating microprocessor speed. (Floating point operations per second, or FLOPs, involve fractional numbers; these take much longer to process than traditional integer operations.) GPUs, with their massively parallel architectures, have been used to date to accelerate FLOPS in bulk – a functional go-around for managing the data requirements of deep learning training applications, oftentimes in centralized data centers.

The inference process then typically obeys rigid response-time limits. This approach reduces the effectiveness of general-purpose computers offering speeds that vary with certain workloads – generally raising up the need for performance accelerators. For designers, meeting performance is likely to become a primary challenge, with AI inference models beginning to move to “productization,” or a fully developed and deployable application in the real-world.

Inference: Inference, on the other hand, exercises that training to predict the meaning of new and novel data. It is performed by feeding new data, such as new images, to the deep neural network, giving it the ability to classify the image.
Deep neural networks in the cloud have to date been powered by GPUs – proven as fundamental for building AI applications for deep learning. But developers must now consider that processor architectures are struggling to keep up with the growing demands of deep learning in an era when Moore’s Law is slowing. Today’s success is directly related to performance acceleration encompassing compute, storage, and connectivity for ideal workload consolidation in close proximity to where data is being generated.

“System architects now widely believe the only path left for major improvements in the performance-energy-cost equation is the domain-specific approach integrating M.2 performance acceleration.”

Domain-specific architectures handle only a few tasks, but they do so extremely well. This is an important shift for system developers and integrators, as the ultimate goal is to improve the cost vs performance ratio when comparing all accelerators, including CPUs, GPUs, and now M.2 options.

Domain-specific architectures are customized to execute a carefully defined workload. This is a core tenet for assuring performance that supports deep learning training and inference. Accelerating DSA processing can also dramatically increase the cost of building modern data centers, as data movement on all levels can be the costliest part of AI inference, i.e., between data centers or within a system.

“Using M.2 accelerators, DSAs operate deep neural networks from 15 to 30 times faster (with 30 to 80 times better energy efficiency) than a counterpart network relying on CPU and GPU technologies.”

While general-purpose GPUs are commonly used to provide enormous processing power for advanced AI algorithms, they are not optimized for edge deployments in remote and unstable environments. Drawbacks of size, power consumption, and heat management generate additional operating costs on top of the upfront cost of the GPU itself. Specialized accelerators such as TPU from Google and M.2 acceleration modules are new solutions that are compact, power efficient, and come purpose-built for driving machine learning algorithms at the edge with incredible performance.
Power Efficiency:
Total Cost of Ownership in Data Center Is Remarkably Different at the Edge

Data centers rely on high performance power, generally confident in the availability of a redundant, high quality power source. The total cost of ownership, or performance per watt equation, changes dramatically for edge environments. Edge computing requires a different approach, more often facing limited resources in power.

Considerations such as power source, thermal and energy limits, cooling, and space are more stringent – and potentially more challenging given the non-traditional environments where these systems must perform. Such factors can be optimized to help ensure reliable performance, for example, reduced power use means less cooling is required.

This may be ideal for more strenuous settings. M.2 takes this approach, given its ability to support better performance per watt in the same or smaller footprint in comparison to traditional processors such as CPUs or even GPUs.

There may be a number of performance constraints at the edge, for example, those related to particularly unique or rigorous environments or industries. However, power and heat dissipation are at the very heart of advanced deep learning at the edge and are directly related to its feasibility in non-traditional settings. An edge device’s limited resources call for maximum power efficiency in neural processing, which can be achieved using DSAs as dedicated processing architectures.

Original Source: NextCurve with modification to show power consumption
Source: Premio - Fog Computing vs Edge Computing
Intel® developed the M.2 [Next Generation Form Factor] interface for flexibility and powerful performance. M.2 supports multiple signal interfaces such as PCI Express (PCIe 3.0 and 4.0), Serial ATA (SATA 3.0), and USB 3.0. A range of bus interfaces enable M.2 expansion slots to be highly versatile for different storage protocols, performance accelerators, wireless connectivity, and I/O expansion modules. For example, M.2 expansion slots can be used to add wired and wireless capabilities or a range of M.2 SSDs with different sizes and specifications.

Further, M.2 offers both legacy and modern compatibility via its support for both SATA and NVMe storage protocols. Legacy standard SATA features an AHCI (Advanced Host Controller Interface), and is the storage protocol defined by Intel® to optimize data manipulation on spinning metal disks in an HDD (hard disk drive) type of storage. NVMe or Non-Volatile Memory Express offers an alternative, created to fully capitalize on flash chip storage (NAND chip) and the PCI Express Lane for super-fast SSD (solid-state drive) storage.

Besides connectivity and storage expansion modules, performance accelerators have quickly adopted the M.2 form factor to benefit from its compact and powerful interface. These performance accelerators include memory accelerators, AI accelerators, deep learning accelerators, inferencing accelerators, and more. These new specialized processors dedicated to AI workloads provide an improved power to performance ratio. This is demonstrated by domain-specific workloads handled by M.2 accelerators vs heterogeneous compute SoCs such as CPU and GPU resources.
Here are some of the top M.2 performance accelerators available today:

**M.2 Intel Optane Memory**
Intel's speed-boosting cache storage in an M.2 format, developed to accelerate cache for another drive to enable high-speed computing.

**M.2 VPU**
Intel's Movidius VPU (Vision Processing Unit), developed to enhance machine learning and inferencing for edge computer vision that requires robust and compact technologies.

**M.2 TPU**
Tensor Processing Unit, developed by Google to accelerate training on large and complex neural network models, a powerful and energy-efficient AI accelerator in a compact M.2 form factor.

**M.2 Hailo-8™**
AI Acceleration Module A best-in-class inference processor packaged in a module for AI applications; offers 26 tera operations per second and compatibility with NGFF M.2 form factor M, B+M, and A+E keys.
In contrast, here are some of the top heterogenous compute accelerators on today’s market:

**CPUs**

Central Processing Units

- **12th generation Intel® Core™ processors (Alder Lake)** – a family of six new desktop processors with turbo boost up to 5.2GHz and as many as 16 cores and 24 threads; new hybrid architecture enables scalable performance from 9W to 125W with a unique combination of performance and efficient cores (P-Core and E-Core)

- **3rd Gen Intel® Xeon® Scalable Processors (Ice Lake)** – this processor family offers from eight to 40 cores and a wide spectrum of frequency, feature, and power levels; their balanced, efficient architecture is optimized for diverse workloads (cloud, enterprise, network, security, and IoT); characterized as data center CPUs with unique built-in AI acceleration and end-to-end data science tools, these processors also feature increased performance of encryption-intensive workloads

- **3rd Gen AMD EPYC™** – based on AMD’s Zen 3 architecture, this processor family is designed to optimize latency-sensitive tasks like PC gaming and features integrated security that enables encryption of in-use data while it is being processed; upgrading from Zen 2 requires only a BIOS flash update, allowing server loads to access a 19% average IPC (instructions per clock/cycle) uplift.
Get To Know the M.2 Interface

Part 4

NVIDIA®'s T4 GPU – based on the new NVIDIA® Turing™ architecture and packaged in an energy-efficient 70-watt, small PCIe form factor; features multi-precision Turing Tensor Cores and new RT Cores

NVIDIA® GeForce RTX™ 4000 & A4000 Series GPUs – powered by Ampere, NVIDIA's 2nd gen RTX architecture, with new RT Cores, Tensor Cores; features streaming multiprocessors for realistic ray-traced graphics

NVIDIA® Jetson AGX Xavier – in a 100mm x 87mm footprint, this is NVIDIA®'s next-generation of performance for compute density, energy efficiency, and AI inferencing; paired with integrated AI tools and workflows, Jetson AGX Xavier modules offer up to 32 TOPS, or tera (trillions) operations per second, empowering developers to quickly train and deploy deep neural networks

Xilinx VCK5000 development card – this domain-specific FPGA is built on the Xilinx 7nm Versal® ACAP architecture and optimized for adaptable engine development in 5G, DC compute, AI, signal processing, radar, and other applications; its high-efficiency silicon (near 100% compute efficiency per watt in some benchmarks and NVIDIA GPU comparisons) is designed use in natural language processing, and convolutional and recurrent neural networks
Throughput Matters: Understanding Benchmarks for Real-World AI Applications

Even the metrics by which industry experts measure compute performance are changing to accommodate the data-rich nature of AI applications. TOPS, or tera (trillions) operations per second, is a measure of the maximum possible throughput rather than a measure of actual throughput. TOPS identifies the number of hardware-implemented computation elements times their clock speed.

While important, it is essentially a measure of what is possible if all the stars align in a given application, that is, steady data input, clean and consistent power sources, no memory limitations, and perfect synchronization between hardware and AI software algorithms. As a theoretical measurement, TOPS also does not offer any consideration for other tasks the hardware may need to perform. Engineers focused on silicon implementation may find specific value in TOPS data, but software and hardware systems engineers may find that it does not clearly indicate true, available performance for their real-world application.

"Throughput, not TOPS, is the more precise, real-world measurement."

Throughput references the amount of data that can be processed in a defined time period, for example FPS (frames per second) in vision processing terms or the number of inferences in deep learning edge application. Inferences or FPS per Watt, as related to a specific neural network task or application, is not only a more precise way of evaluating and comparing hardware but also a more clearly understood, real-world metric.

That said, it is very difficult to compare models for object detection. Accuracy and speed are critical measurements, but other real-world factors also affect performance. Which deep learning model is used, varying image input resolutions, variations in data augmentation or training sets, which feature maps layers are used for object detection – these and many other factors create impact. On this landscape, both ResNet-50 and YOLOv3 have emerged as leading options for AI performance evaluation, as well as use as a backbone in the development of new neural models.
ResNet-50, a pre-trained deep learning model

To solve computer vision challenges, machine learning developers working with convolutional neural networks (CNNs) add more stacking layers to their models. Problems are solved efficiently as each layer can be trained for specific tasks, which in turn help deliver accurate results. Stacked layers enrich the model’s features – to a point. Ultimately, a higher number of levels increases saturation such that it creates negative impact on performance of both the testing and training data models. ResNet-50 tackles this deterioration problem. Using residual blocks, or “skip connections” that simplify learning functions for each layer, ResNet-50 improves deep neural network efficiency while reducing errors. ResNet has a spectrum of variants – same concept, different number of layers – and ResNet-50 references the variant designed to handle 50 layers.

YOLOv3, a real-time object detection algorithm

As a CNN, YOLOv3 (You Only Look Once, Version 3) identifies specific objects in real-time for example in videos, live feeds, or images. YOLO’s classifier-based system interacts with input images as structured arrays of data – its goal is to recognize patterns between them and sort them into defined classes with similar characteristics. Only objects with similar characteristics are sorted; others are ignored unless system programming instructs attention. In YOLO, the size of its prediction map corresponds exactly to the size of its related feature map. These factors contribute to YOLO’s ability to operate much faster than other options, even as it maintains its correctness.

The algorithm allows the model to view the entire image during testing, ensuring its predictions are informed by a more global image context. YOLO assigns a score to regions based on how similar they are to other predefined classes, so a high scoring region is noted as a positive detection in its most closely identified class. A live traffic feed provides a good example – here, YOLO can detect various types of vehicles, examining high scoring regions and identifying similarities with certain predefined classes.
There is a clear differentiation between a general-purpose embedded computer and one that’s designed to handle inferencing algorithms. Premio has addressed this industry need with its AI Edge Inference Computers, combining next-generation processing and high speed storage technologies with the latest IoT connectivity features. The solution is designed from the ground up to deliver holistic inference analysis at the rugged edge. Systems also integrate M.2 performance accelerators to ensure a balance of performance, power, and cost. Integration for AI inference is seamless, resulting in hardware engineered for data-intensive performance directly through the PCIe I/O bus.

Premio’s AI Edge Inference Computers offer a range of features and options, incorporating advanced performance up to Intel® 9th and 10th Generation processors. Systems offer rich GPU support and scalable, hot-swappable NVMe capacity in versatile hardware designed to withstand deployment in challenging industrial environments. Because these systems are often exposed to dust, debris, shock, vibration, and extreme temperatures, Premio’s AI Edge Inference Computers are hardened to withstand exposure to these harsh environmental factors.
From external enclosure to internal components, every element of Premio’s AI Edge Inference Computer series is purpose-built through a combination of mechanical and thermal engineering to address environmental issues such as strong vibration, severe temperatures, and the presence of moisture or dirt. These industrial-grade computers are also validated to execute functions with extreme processing power and storage capacity, built to eliminate downtime and ensure stable 24/7 operation.

Deployed at the rugged edge, these specialized inference computers not only must endure temperature extremes, but also accommodate questionable power sources and mitigate kinetic factors as they process great volumes of data through a wide variety of I/O ports. Rich wireless connectivity permits uninterrupted communication, monitoring, data transfer, and automation to meet the incredible demands of rugged edge computing.

Premio’s AI Edge Inference Computers also feature EDGEBoost Node technology, developed by Premio to maximize system performance at the edge. The hardware nodes physically attach to the lower portion of the platform and provide hardware acceleration for edge-level workloads that require data acquisition for real-time insights. This two-piece modular design helps maintain the platform’s ruggedness while providing performance acceleration through NVMe solid-state disk drives in innovative cannister bricks and GPUs for parallel computing performance. Each EDGEBoost Node uses high-RPM active cooling to ensure reliability of those components.

These new nodes address current demands for performance acceleration needed to process and analyze large volumes of data for rugged edge computing. Specific computer vision and edge computing applications can benefit from real-time processing capabilities with the latest hardware acceleration technologies by incorporating Premio’s modular AI Edge Inference Computers.
Performance Accelerators for AI: Keeping Pace with Innovation

The massive influx of data – and applications that rely on it – is driving change in performance acceleration strategies. Specialization is required, creating systems that interact efficiently and on-demand with AI workloads. While the slowing of Moore’s Law has been managed using CPU/GPU-based designs, today these processor architectures are struggling to keep pace with the real-time data needs of automation and inference applications, particularly in more rigorous non-data center environments. Coupled with the increasing challenge of meeting price-performance-power demands, it’s more important than ever for performance acceleration to consider all factors necessary – compute, storage, and connectivity – for effective workload consolidation close to where data is generated, no matter how rugged the setting.

Here, M.2 is proving a powerful design option for system architects, offering domain-specific value that meets the precise needs of AI workloads. Contrasted to a counterpart system using CPU/GPU technologies, an M.2-based system operates deep neural networks 15 to 30 times faster with 30 to 80 times improved energy efficiency. These values are driving advanced system design optimized for the rugged edge. M.2 is helping system developers eliminate performance roadblocks and is playing a greater role as more and more systems are deployed in demanding, non-traditional environments where real-time computing makes a dramatic difference in services and operations. For system developers, the opportunity for purpose-built systems is immense – with smarter data handling poised to advance AI and inference applications even more broadly across global infrastructure industries.

Click <<here>> to set up a designer-to-designer discussion of your rugged computing needs, including real-time data processing and performance acceleration or <<here>> for more insight on Premio’s modular AI Edge Inference Computers and EDGEBoost Node technology.
# RCO-6000-CFL (w/ 1x Hailo8)

<table>
<thead>
<tr>
<th>Network</th>
<th>Resolution</th>
<th>FPS</th>
<th>Total Power [W]</th>
<th>FPS/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet_v1-50</td>
<td>224 x 224</td>
<td>1,276</td>
<td>83.1</td>
<td>15.4</td>
</tr>
<tr>
<td>RegNetx_800mf</td>
<td>224 x 224</td>
<td>2,404</td>
<td>81.9</td>
<td>29.4</td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>224 x 224</td>
<td>2,612</td>
<td>82.2</td>
<td>31.8</td>
</tr>
<tr>
<td>Efficientnet_M</td>
<td>240 x 240</td>
<td>891</td>
<td>83.2</td>
<td>10.7</td>
</tr>
<tr>
<td>MobileNet-V1 SSD</td>
<td>300 x 300</td>
<td>1,039</td>
<td>82.3</td>
<td>12.6</td>
</tr>
<tr>
<td>MobileNet_v34</td>
<td>224 x 224</td>
<td>3,489</td>
<td>82</td>
<td>42.5</td>
</tr>
<tr>
<td>YOLOv3s</td>
<td>608 x 608</td>
<td>60</td>
<td>84.3</td>
<td>0.7</td>
</tr>
<tr>
<td>YOLOv4s</td>
<td>512 x 512</td>
<td>68</td>
<td>83.1</td>
<td>0.8</td>
</tr>
<tr>
<td>YOLOv5m</td>
<td>640 x 640</td>
<td>218</td>
<td>84.4</td>
<td>2.6</td>
</tr>
<tr>
<td>NETWORK</td>
<td>RESOLUTION</td>
<td>FPS</td>
<td>TOTAL POWER [W]</td>
<td>FPS/W</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>------</td>
<td>-----------------</td>
<td>-------</td>
</tr>
<tr>
<td>ResNet_v1-50</td>
<td>224 x 224</td>
<td>5,104</td>
<td>92</td>
<td>55.5</td>
</tr>
<tr>
<td>RegNetx_800mf</td>
<td>224 x 224</td>
<td>9,616</td>
<td>88</td>
<td>109.3</td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>224 x 224</td>
<td>10,448</td>
<td>89</td>
<td>117.4</td>
</tr>
<tr>
<td>Efficientnet_M</td>
<td>240 x 240</td>
<td>3,564</td>
<td>93</td>
<td>38.3</td>
</tr>
<tr>
<td>MobileNet-V1 SSD</td>
<td>300 x 300</td>
<td>4,156</td>
<td>89</td>
<td>46.7</td>
</tr>
<tr>
<td>MobileNet_v34</td>
<td>224 x 224</td>
<td>13,956</td>
<td>88</td>
<td>158.6</td>
</tr>
<tr>
<td>YOLOv35</td>
<td>608 x 608</td>
<td>240</td>
<td>97</td>
<td>2.5</td>
</tr>
<tr>
<td>YOLOv45</td>
<td>512 x 512</td>
<td>272</td>
<td>92</td>
<td>3.0</td>
</tr>
<tr>
<td>YOLOv5m</td>
<td>640 x 640</td>
<td>872</td>
<td>98</td>
<td>8.9</td>
</tr>
</tbody>
</table>
## RCO-6000-CFL-4NS (w/ 8x Hailo8)

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>RESOLUTION</th>
<th>FPS</th>
<th>TOTAL POWER [W]</th>
<th>FPS/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet_v1-50</td>
<td>224 x 224</td>
<td>10,208</td>
<td>115</td>
<td>88.8</td>
</tr>
<tr>
<td>RegNetx_800mf</td>
<td>224 x 224</td>
<td>19,232</td>
<td>105</td>
<td>183.2</td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>224 x 224</td>
<td>20,896</td>
<td>108</td>
<td>193.5</td>
</tr>
<tr>
<td>Efficientnet_M</td>
<td>240 x 240</td>
<td>7,128</td>
<td>116</td>
<td>61.4</td>
</tr>
<tr>
<td>MobileNet-V1 SSD</td>
<td>300 x 300</td>
<td>8,312</td>
<td>108</td>
<td>77.0</td>
</tr>
<tr>
<td>MobileNet_v34</td>
<td>224 x 224</td>
<td>27,912</td>
<td>106</td>
<td>263.3</td>
</tr>
<tr>
<td>YOLOv3s</td>
<td>608 x 608</td>
<td>480</td>
<td>124</td>
<td>3.9</td>
</tr>
<tr>
<td>YOLOv4s</td>
<td>512 x 512</td>
<td>544</td>
<td>115</td>
<td>4.7</td>
</tr>
<tr>
<td>YOLOv5m</td>
<td>640 x 640</td>
<td>1,744</td>
<td>125</td>
<td>14.0</td>
</tr>
<tr>
<td>Network</td>
<td>Resolution</td>
<td>FPS</td>
<td>Total Power [W]</td>
<td>FPS/W</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------</td>
<td>-------</td>
<td>-----------------</td>
<td>-------</td>
</tr>
<tr>
<td>ResNet_v1-50</td>
<td>224 x 224</td>
<td>30,624</td>
<td>504</td>
<td>60.8</td>
</tr>
<tr>
<td>RegNetx_800mf</td>
<td>224 x 224</td>
<td>57,696</td>
<td>476</td>
<td>121.2</td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>224 x 224</td>
<td>62,688</td>
<td>483</td>
<td>129.8</td>
</tr>
<tr>
<td>Efficientnet_M</td>
<td>240 x 240</td>
<td>21,384</td>
<td>507</td>
<td>42.2</td>
</tr>
<tr>
<td>MobileNet-V1 SSD</td>
<td>300 x 300</td>
<td>24,936</td>
<td>485</td>
<td>51.4</td>
</tr>
<tr>
<td>MobileNet_v34</td>
<td>224 x 224</td>
<td>83,736</td>
<td>478</td>
<td>175.2</td>
</tr>
<tr>
<td>YOLOv3s</td>
<td>608 x 608</td>
<td>1,440</td>
<td>533</td>
<td>2.7</td>
</tr>
<tr>
<td>YOLOv4s</td>
<td>512 x 512</td>
<td>1,632</td>
<td>504</td>
<td>3.2</td>
</tr>
<tr>
<td>YOLOv5m</td>
<td>640 x 640</td>
<td>5,232</td>
<td>536</td>
<td>9.8</td>
</tr>
</tbody>
</table>
### FlacheSAN1N36M-UM (w/ 36x Hailo8)

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>RESOLUTION</th>
<th>FPS</th>
<th>TOTAL POWER [W]</th>
<th>FPS/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet_v1-50</td>
<td>224 x 224</td>
<td>45,936</td>
<td>562</td>
<td>81.7</td>
</tr>
<tr>
<td>RegNetx_800mf</td>
<td>224 x 224</td>
<td>86,544</td>
<td>518</td>
<td>167.1</td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>224 x 224</td>
<td>94,032</td>
<td>529</td>
<td>177.8</td>
</tr>
<tr>
<td>Efficientnet_M</td>
<td>240 x 240</td>
<td>32,076</td>
<td>565</td>
<td>56.8</td>
</tr>
<tr>
<td>MobileNet-V1 SSD</td>
<td>300 x 300</td>
<td>37,404</td>
<td>533</td>
<td>70.2</td>
</tr>
<tr>
<td>MobileNet_v34</td>
<td>224 x 224</td>
<td>125,604</td>
<td>522</td>
<td>240.6</td>
</tr>
<tr>
<td>YOLOv3s5</td>
<td>608 x 608</td>
<td>2,160</td>
<td>605</td>
<td>3.6</td>
</tr>
<tr>
<td>YOLOv4s5</td>
<td>512 x 512</td>
<td>2,448</td>
<td>562</td>
<td>4.4</td>
</tr>
<tr>
<td>YOLOv5s5m</td>
<td>640 x 640</td>
<td>7,848</td>
<td>608</td>
<td>12.9</td>
</tr>
</tbody>
</table>
## Appendix B: Hailo Modules Vs. Comparable GPUs

<table>
<thead>
<tr>
<th>PRODUCT</th>
<th>Max / Typical Power (W)</th>
<th>Max TOPS (INT8)</th>
<th>ResNet-50 Performance (FPS)</th>
<th>Form Factor</th>
<th>FPS/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCO-6000-CFL (w/ 1x Hailo-8)</td>
<td>6.25</td>
<td>26</td>
<td>1,329</td>
<td>Rugged Edge Computer</td>
<td>213</td>
</tr>
<tr>
<td>VCO-6000-CFL-4M2 (w/ 4x Hailo-8)</td>
<td>25</td>
<td>104</td>
<td>5,316</td>
<td>Rugged Edge Computer</td>
<td>213</td>
</tr>
<tr>
<td>RCO-6000-CFL-4NS (w/ 8x Hailo-8)</td>
<td>50</td>
<td>208</td>
<td>10,632</td>
<td>Rugged Edge Computer</td>
<td>213</td>
</tr>
<tr>
<td>FLACHESAN4N24M-UM (w/ 24x Hailo-8)</td>
<td>150</td>
<td>624</td>
<td>31,896</td>
<td>4U Tower Server</td>
<td>213</td>
</tr>
<tr>
<td>FlacheSAN1N36M-UM (w/ 36x Hailo-8)</td>
<td>225</td>
<td>936</td>
<td>47,844</td>
<td>1U Rackmount Server</td>
<td>213</td>
</tr>
<tr>
<td>RCO-6000-CFL (w/ Nvidia RTX 4000)</td>
<td>160</td>
<td>57</td>
<td>4,897</td>
<td>Rugged Edge Computer</td>
<td>31</td>
</tr>
<tr>
<td>RCO-6000-CFL (w/ Nvidia RTX A4000)</td>
<td>140</td>
<td>153</td>
<td>5,914</td>
<td>Rugged Edge Computer</td>
<td>42</td>
</tr>
</tbody>
</table>
Appendix C: Neural Network Comparisons
(Total Power Vs. Frame Per Second)