TECHNICAL OVERVIEW

NVIDIA AI INFERENCE PLATFORM

Giant Leaps in Performance and Efficiency for AI Services, from the Data Center to the Network's Edge

Introduction

The artificial intelligence revolution surges forward, igniting opportunities for businesses to reimagine how they solve customer challenges. We're racing toward a future where every customer interaction, every product, and every service offering will be touched and improved by AI. Realizing that future requires a computing platform that can accelerate the full diversity of modern AI, enabling businesses to create new customer experiences, reimagine how they meet—and exceed—customer demands, and cost-effectively scale their AI-based products and services.

While the machine learning field has been active for decades, deep learning (DL) has boomed over the last six years. In 2012, Alex Krizhevsky of the University of Toronto won the ImageNet image recognition competition using a deep neural network trained on NVIDIA GPUs—beating all human expert algorithms that had been honed for decades. That same year, recognizing that larger networks can learn more, Stanford's Andrew Ng and NVIDIA Research teamed up to develop a method for training networks using large-scale GPU computing systems. These seminal papers sparked the "big bang" of modern AI, setting off a string of "superhuman" achievements. In 2015, Google and Microsoft both beat the best human score in the ImageNet challenge. In 2016, DeepMind's AlphaGo recorded its historic win over Go champion Lee Sedol and Microsoft achieved human parity in speech recognition.

GPUs have proven to be incredibly effective at solving some of the most complex problems in deep learning, and while the NVIDIA deep learning platform is the standard industry solution for training, its inference capability is not as widely understood. Some of the world's leading enterprises from the data center to the edge have built their inference solution on NVIDIA GPUs.
Some examples include:

- **SAP's Brand Impact Service** achieved a 40X performance increase while reducing costs by 32X.
- **Bing Visual Search** improved latency by 60X and reduced costs by 10X.
- **Cisco's Spark Board and Spark Room Kit**, powered by NVIDIA® Jetson™ GPU, enable wireless 4K video sharing and use deep learning for voice and facial recognition.

The Deep Learning Workflow

The two major operations from which deep learning produces insights are training and inference. While similar, there are significant differences. Training feeds examples of objects to be detected or recognized, like animals, traffic signs, etc., into a neural network, allowing it to make predictions as to what these objects are. The training process reinforces correct predictions and corrects the wrong ones. Once trained, a production neural network can achieve upwards of 90 to 98 percent correct results. "Inference" is the deployment of a trained network to evaluate new objects and make predictions with similar predictive accuracy.

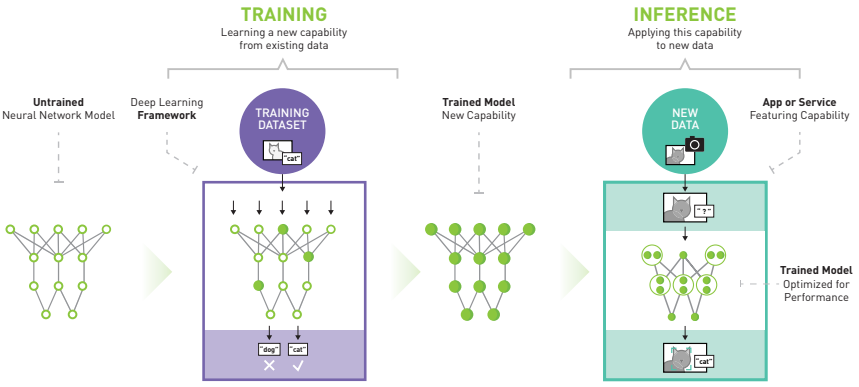

Figure 1

Both training and inference start with the forward propagation calculation, but training goes further. After forward propagation during training, the results are compared against the "ground truth" correct answer to compute an error value. The backward propagation phase then sends the error back through the network's layers and updates weights using the stochastic gradient descent to improve the network's performance on the task it's trying to learn. It's common to batch hundreds of training inputs (for example, images in an image classification network or spectrograms for speech recognition) and operate on them simultaneously during deep neural network (DNN) training to amortize loading weights from GPU memory across many inputs, greatly increasing computational efficiency.

Inference can also batch hundreds of samples to achieve optimal throughput on jobs run overnight in data centers to process substantial amounts of stored data. These jobs tend to emphasize throughput over latency. However, for real-time usages, high batch sizes also carry a latency penalty. For these usages, lower batch sizes (as low as a single sample) are used, trading off throughput for lowest latency. A hybrid approach, sometimes referred to as "auto-batching," sets a time

threshold—say, 10 milliseconds (ms)—and batches as many samples as possible within those 10ms before sending them on for inference. This approach achieves better throughput while maintaining a set latency amount.

TensorRT Hyperscale Inference Platform

The NVIDIA TensorRT™ Hyperscale Inference Platform is designed to make deep learning accessible to every developer and data scientist anywhere in the world. It all starts with the world's most advanced AI inference accelerator, the NVIDA Tesla® T4 GPU featuring NVIDIA Turing™ Tensor Cores. Based on NVIDIA's new Turing architecture, Tesla T4 accelerates all types of neural networks for images, speech, translation, and recommender systems, to name a few. Tesla T4 supports a wide variety of precisions and accelerates all major DL frameworks, including TensorFlow, PyTorch, MXNet, Chainer, and Caffe2.

Since great hardware needs great software, NVIDIA TensorRT, a high-performance deep learning inference optimizer and runtime, delivers low-latency, high-throughput inference for applications such as image classification, segmentation, object detection, machine language translation, speech, and recommendation engines. It can rapidly optimize, validate, and deploy trained neural network for inference to hyperscale data centers, embedded, or automotive GPU platforms. TensorRT 5 unlocks the power of Turing GPUs across a wide range of precisions, from FP32 all the way down to INT8. In addition, TensorRT has in-framework support for TensorFlow, MXNet, Caffe2 and MATLAB frameworks, and supports other frameworks via ONNX.

NVIDIA TensorRT Inference Server, available as a ready-to-run container at no charge from NVIDIA GPU Cloud, is a production-ready deep learning inference server for data center deployments. It reduces costs by maximizing utilization of GPU servers and saves time by integrating seamlessly into production architectures. NVIDIA TensorRT Inference Server simplifies workflows and streamlines the transition to a GPU-accelerated infrastructure for inference.

And for large-scale, multi-node deployments, Kubernetes on NVIDIA GPUs enables enterprises to scale up training and inference deployment to multi-cloud GPU clusters seamlessly. It lets software developers and DevOps engineers automate deployment, maintenance, scheduling, and operation of multiple GPU-accelerated application containers across clusters of nodes. With Kubernetes on NVIDIA GPUs, they can build and deploy GPU-accelerated deep learning training or inference applications to heterogeneous GPU clusters at scale seamlessly.

The Tesla T4 Tensor Core GPU, Based on NVIDIA Turing Architecture

The NVIDIA Tesla T4 GPU is the world's most advanced accelerator for all AI inference workloads. Powered by NVIDIA Turing™ Tensor Cores, T4 provides revolutionary multi-precision inference performance to accelerate the diverse applications of modern AI. T4 is a part of the NVIDIA AI Inference Platform that supports all AI frameworks and provides comprehensive tooling and integrations to drastically simplify the development and deployment of advanced AI.

Turing Tensor Cores are purpose-built to accelerate AI inference and Turing GPUs also inherit all of the enhancements introduced to the NVIDIA CUDA® platform with the NVIDIA Volta™ architecture, improving the capability, flexibility, productivity, and portability of compute applications. Features such as independent thread scheduling, hardware-accelerated Multi-Process Service (MPS) with address space isolation for multiple applications, unified memory with address translation services, and Cooperative Groups are all part of the Turing GPU architecture.

NVIDIA Turing Innovations

Figure 2: NVIDIA TURING TU102 GPU

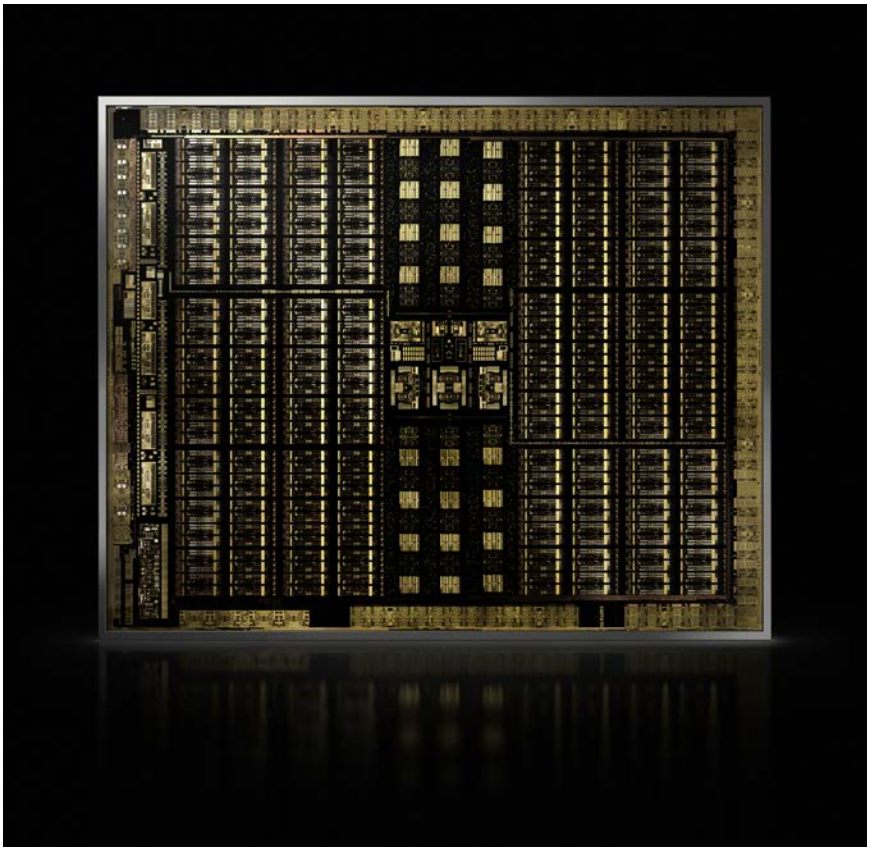

Figure 2

Turing Key Features

> **New Streaming Multiprocessor (SM) with Turing Tensor Cores**

The Turing SM builds on the major SM advancements of the Volta GV100 architecture and delivers major boosts in performance and energy efficiency compared to the previous-generation NVIDIA Pascal™ GPUs. Turing Tensor Cores not only provide FP16/FP32 mixed-precision matrix math like Volta Tensor Cores; they also add new INT8 and INT4 precision modes, massively accelerating a broad spectrum of deep learning inference applications.

Similar to Volta, the Turing SM provides independent floating-point and integer data paths, allowing a more efficient execution of common workloads with a mix of computation and address calculations. Warp-coherent computation can be accelerated with a uniform register file (URF) and data path (UDP). Also, independent thread scheduling enables finer-grain synchronization and cooperation among threads. Lastly, the combined shared memory and L1 cache improves performance significantly while simplifying programming.

> **Deep Learning Features for Inference**

Turing GPUs deliver exceptional inference performance, versatility, and efficiency. Turing Tensor Cores, along with continual improvements in TensorRT, CUDA, and CuDNN libraries, enable Turing GPUs to deliver outstanding performance for inference applications. Turing also includes experimental features such as support for INT4 and INT1 formats to further research and development in deep learning.

> **GDDR6 High-Performance Memory Subsystem**

Turing is the first GPU architecture to utilize GDDR6 memory, which represents the next big advance in high-bandwidth GDDR DRAM memory design that can deliver up to 320GB/sec. GDDR6 memory interface circuits in Turing GPUs have been completely redesigned for speed, power efficiency, and noise reduction. Turing's GDDR6 memory subsystem delivers a 40 percent speedup and a 20 percent power efficiency improvement over GDDR5X memory used in Pascal GPUs.

> **Twice the Video Decode Performance**

Video continues on its explosive growth trajectory, comprising over two-thirds of all internet traffic. Accurate video interpretation through AI is driving the most relevant content recommendations, finding the impact of brand placements in sports events, and delivering perception capabilities to autonomous vehicles, among other usages.

Tesla T4 delivers breakthrough performance for AI video applications, with dedicated hardware transcoding engines that bring twice the decoding performance of prior-generation GPUs. T4 can decode up to 38 full-HD video streams, making it easy to integrate scalable deep learning into the video pipeline to deliver innovative, smart video services. It features performance and efficiency modes to enable either fast encoding or the lowest bit-rate encoding without loss of video quality.

TensorRT 5 Features

The NVIDIA TensorRT Hyperscale Inference Platform is a complete inference solution that includes the cutting-edge Tesla T4 inference accelerator, the TensorRT 5 high-performance deep learning inference optimizer and runtime, and TensorRT Inference Server. This power trio delivers low latency and high throughput for deep learning inference applications and allows them to be quickly deployed. It can also leverage tools like Kubernetes, which can quickly scale containerized applications across multiple hosts. With TensorRT 5, neural network models can be optimized, calibrated for lower precision with high accuracy, and finally deployed to hyperscale data centers, embedded, or automotive product platforms. TensorRT-based applications on GPUs perform up to 50X faster than CPU during inference for models trained in all major frameworks.

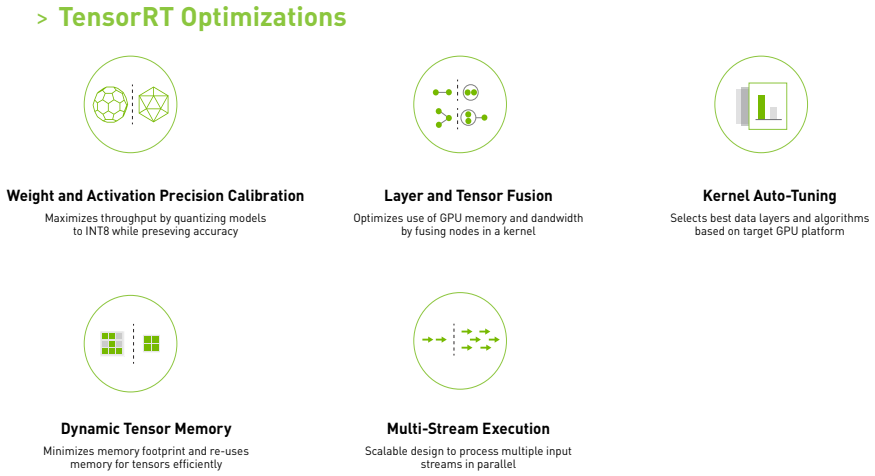

Figure 3

TensorRT provides INT8 and FP16 optimizations for production deployments of deep learning inference applications such as video streaming, speech recognition, recommendation, and natural language processing. Reduced-precision inference significantly lowers application latency while preserving model accuracy, which is a requirement for many real-time services as well as auto and embedded applications.

TensorRT and TensorFlow are now tightly integrated, giving developers the flexibility of TensorFlow with the powerful optimizations of TensorRT. And now, TensorFlow, MXNet, Caffe2 and MATLAB all have in-framework support to access TensorRT acceleration. In addition, the native ONNX parser in TensorRT provides an easy path to import models from frameworks such as Caffe2, Microsoft Cognitive Toolkit, and Chainer. MATLAB is integrated with TensorRT through GPU Coder so that engineers and scientists using MATLAB can automatically generate high-performance inference engines for Jetson, NVIDIA DRIVE™, and Tesla platforms.

TensorRT accelerates a wide diversity of usages, including images, video, speech recognition, neural machine translation, and recommender systems.

While it's possible to do inference operations within a deep learning framework, TensorRT easily optimizes networks to deliver far more performance and it includes new layers for multilayer perceptrons (MLP) and recurrent neural networks (RNNs). TensorRT also takes full advantage of the Turing architecture. Later in this paper, data will show how this combination delivers up to 45X more throughput than a CPU-only server.

Tesla GPUs, paired with the TensorRT inference optimizer, deliver massive performance gains, both on convolutional neural networks (CNNs), often used for image-based networks, as well as RNNs that are frequently used for speech and translation applications.

Inference Performance: Getting the Complete Picture

Measured performance in computing tends to fixate on speed of execution. But in deep learning inference performance, speed is one of seven critical factors that come into play. A simple acronym, PLASTER captures these seven factors. They are:

Programmability
Latency
Accuracy
Size of Model
Throughput
Energy Efficiency
Rate of Learning

Figure 4

Deep learning is a complex undertaking and so can be choosing the right deep learning platform. All seven of these elements should be included in any decision analysis, and many of these factors are interrelated. Consider these seven factors and the role each plays.

- **Programmability:** Machine learning is experiencing explosive growth, not only in the size and complexity of the models, but also in the burgeoning diversity of neural network architectures. NVIDIA addresses training and inference challenges with two key tools—CUDA and TensorRT, NVIDIA's programmable inference accelerator. In addition, NVIDIA's deep learning platform accelerates ALL deep learning frameworks, both for training and inference.

- **Latency:** Humans and machines need a response to an input to make decisions and take action. Latency is the time between requesting something and receiving a response. While AI continues to evolve rapidly the latency targets for real-time services remain a constant. For example, there is wide demand for digital assistants in both consumer and customer service applications. But when humans try to interface with digital assistants, a lag of even a few seconds starts to feel unnatural.

- **Accuracy:** While accuracy is important in every industry, healthcare needs especially high accuracy. Medical imaging has advanced significantly in the last couple of decades, increasing usage and requiring more analysis to identify medical issues. Medical imaging advancements and usage also mean that large volumes of data must be transmitted from medical machines to medical specialists for analysis. An advantage of deep learning is that it can be trained at high precision and implemented at lower precision.

- **Size of Network:** The size of a deep learning model and the capacity of the physical network between processors have impacts on performance, especially in the latency and throughput aspects of PLASTER. Deep learning network models are exploding in numbers. Their size and complexity are also increasing, enabling far more detailed analysis and driving the need for more powerful systems for training.

- **Throughput:** Developers are increasingly optimizing inference within a specified latency threshold. While the latency limit ensures good customer experience, maximizing throughput within that limit is critical to maximizing data center efficiency and revenue. There's been a tendency to use throughput as the only performance metric, as more computations per second generally lead to better performance across other areas. However, without the appropriate balance of throughput and latency, the result can be poor customer service, missing service-level agreements (SLAs), and potentially a failed service.

- **Energy Efficiency:** As DL accelerator performance improves, power consumption escalates. Providing a return on investment (ROI) for deep learning solutions involves more than looking at just

the inference performance of a system. Power consumption can quickly increase the cost of delivering a service, driving a need to focus on energy efficiency in both devices and systems. Therefore, the industry measures operational success in inferences per watt (higher is better). Hyperscale data centers seek to maximize energy efficiency for as many inferences as they can deliver within a fixed power budget.

- **Rate of Learning:** One of the two words in "AI" is "intelligence," and users want neural networks to learn and adapt within a reasonable time frame. For complex DL systems to gain traction in business, software tool developers must support the DevOps movement. DL models must be retrained periodically as inference services gather new data and as services grow and change. Therefore, IT organizations and software developers must increase the rate at which they can retrain models as new data arrives.

Throughput

Image-based networks are used for image and video search, video analytics, object classification and detection, and a host of other usages. Looking at an image-based dataset (ImageNet) run on three different networks, a single Tesla P4 GPU is 12X faster than a CPU-only server, while the Tesla V100 Tensor Core GPU is up to 45X faster than that same CPU-only server.

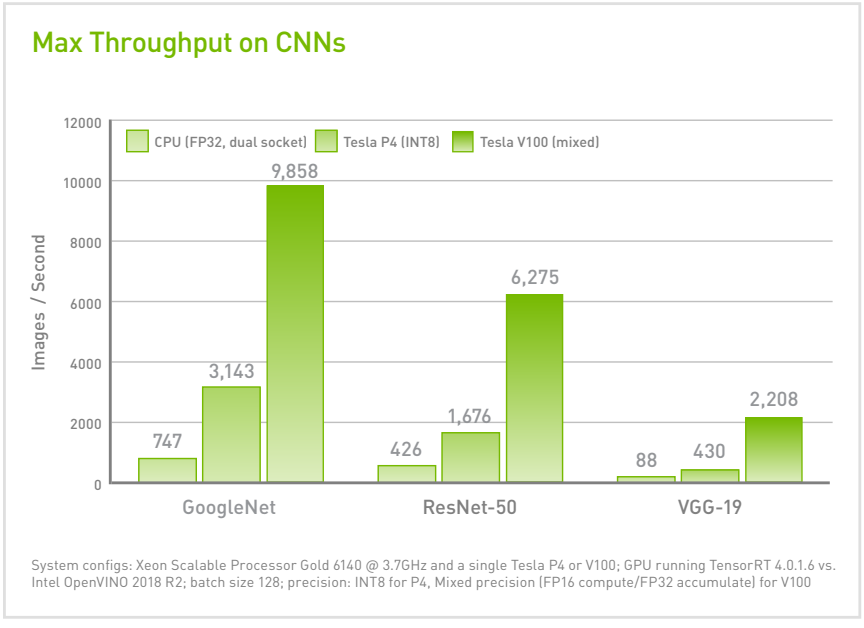

Chart 1

RNNs are used for time-series or sequential data and are often applied as solutions for translation, speech recognition, natural language

processing, and even speech synthesis. The data shown here are from the OpenNMT (neural machine translation) network, translating a dataset from German to English. The Tesla P4 is delivering 81X more throughput, while the Tesla V100 Tensor Core GPU is an even more impressive 352X faster than a CPU-only server.

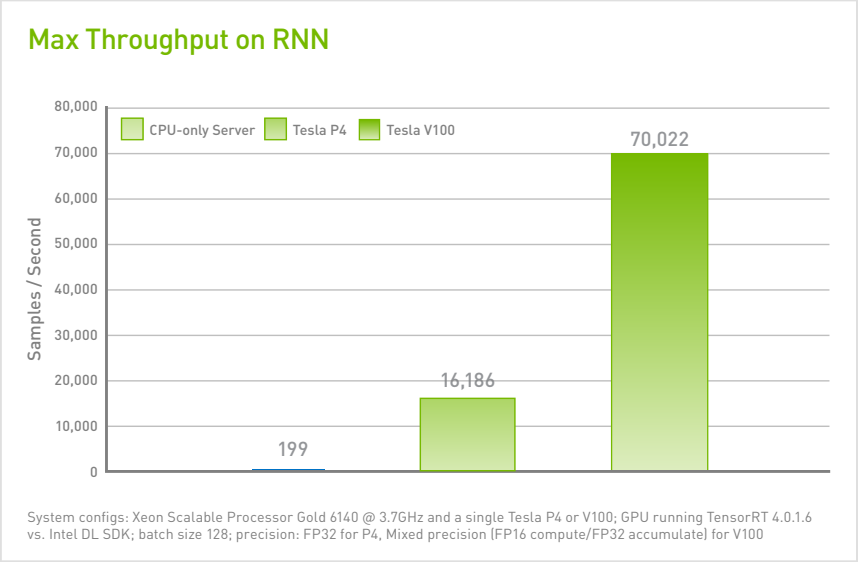

Chart 2

Low-Latency Throughput

While achieving high throughput is critical to inference performance, so is achieving latency. Many real-time use cases actually involve multiple inferences per query. For instance, a spoken question will traverse an automatic speech recognition (ASR), speech to text, natural language processing, a recommender system, text to speech and then speech synthesis. Each of these steps is a different inference operation. And while some pipelining is possible, the latency of each of those inference operations contributes to the latency of the overall experience. Shown here are low-latency throughputs for both CNNs as well as an RNN. Developers have generally approached low-latency inference two ways: 1) processing requests immediately as they come in with no batching (also called batch size 1) or 2) using an "auto-batching" technique where a latency limit is set (e.g., 7ms), samples are batched until either that limit is hit or a certain batch size is achieved (e.g., batch size 8), and then the work is sent through the network for inference processing. The former approach is easier to implement, while the latter approach has the advantage of delivering more inference throughput while preserving the prescribed latency limit. To that end, we present CNN results using the 7ms latency budget approach, while the RNN results are shown using a batch size of one.

Chart 3: With the emergence of real-time AI-based services, latency becomes an increasingly important facet of inference performance. Not only is high throughput critical, but so is delivering high throughput within a specified latency budget to optimize end-user experience. Google has stated1 that 7ms is an optimal latency target. Applying that latency target here, the above chart shows that Tesla V100 delivers up to 77X more performance than a CPU-only server within the 7ms latency budget. CPU latencies not meeting the 7ms requirement are shown in red. Meanwhile, the CPU server is unable to deliver its throughput within the specified latency budget for VGG-19.

1. ref: https://arxiv.org/ftp/arxiv/papers/1704/1704.04760.pdf

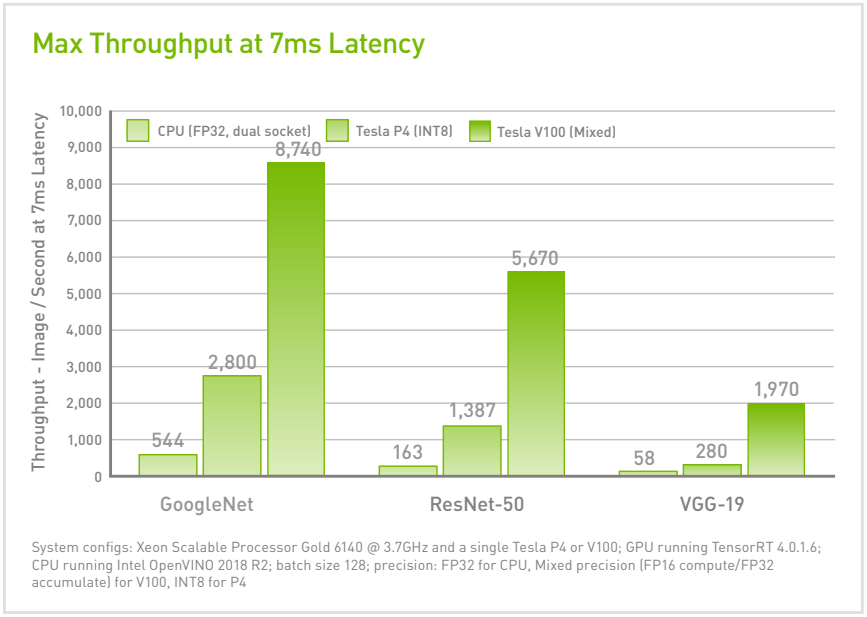

Chart 3

There have been some misconceptions that GPUs are unable to achieve very low latency at a batch size of one. However, as the chart below reflects, Tesla P4 and Tesla V100 are delivering 1.8 and 1.1ms, respectively, at a batch size of one, whereas a CPU server is at 6ms. In addition, the CPU server is delivering only 163 images per second, whereas Tesla P4 is at 562 images per second and the Tesla V100 is delivering 870 images per second.

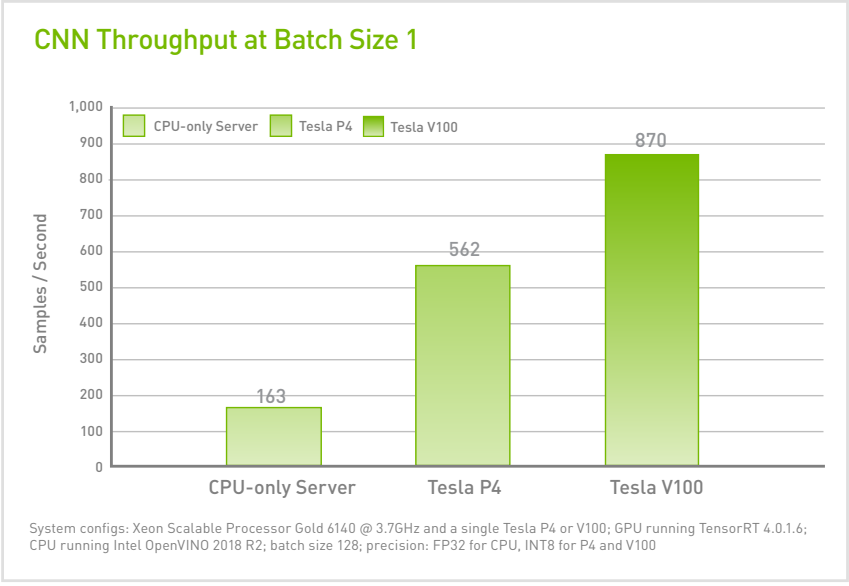

Chart 4

Performance Efficiency

We've covered maximum throughput already, and while very high throughput on deep learning workloads is a key consideration, so is how efficiently a platform can deliver that throughput.

Here we offer a first look at the Turing-based Tesla T4, whose efficiency far exceeds either the Tesla P4 or the Tesla V100. With its small form factor and 75-watt (W) footprint design, Tesla T4 is the world's most advanced universal inference accelerator. Powered by Turing Tensor Cores, T4 will bring revolutionary multi-precision inference performance to efficiently accelerate the diverse applications of modern AI. In addition, Tesla T4 is poised to more than double the efficiency of its predecessor, the Tesla P4.

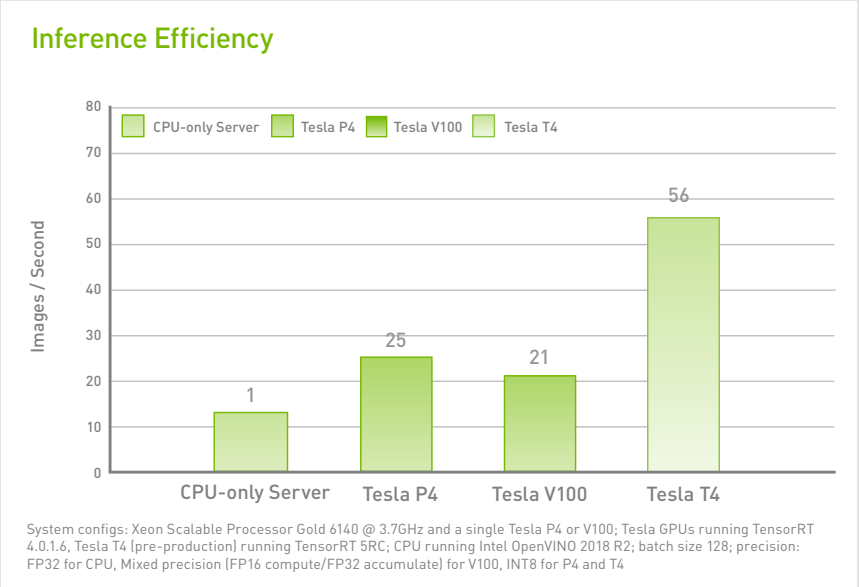

Chart 5

GPU Inference: Business Implications

Tesla V100 and P4 deliver massive performance boosts and power efficiency, but how does that benefit acquisition and operations budgets? Simply put, big performance translates into big savings.

The image below shows that a single server with 16 Tesla T4 GPUs running speech, NLP and video usages provides the same throughput performance as 200 CPU-based servers that take up four entire racks and require 60KW of power. The result? This single Tesla T4–equipped server will deliver a 30X reduction in power and a 200X reduction in server count.

Figure 5

Jetson: Inference at the Edge

NVIDIA Jetson TX2 is a credit card–sized open platform that delivers AI computing at the edge—opening the door to powerfully intelligent factory robots, commercial drones, and smart cameras for AI cities. Based on the NVIDIA Pascal architecture, Jetson TX2 offers twice the performance of its predecessor, or it can run at more than twice the power efficiency while drawing less than 7.5 watts (W) of power. This allows Jetson TX2 to run larger, deeper neural networks on edge devices. The result: smarter devices with higher accuracy and faster response times for tasks like image classification, navigation, and speech recognition. Deep learning developers can use the very same development tools for Jetson that they use on the Tesla platform, such as CUDA, cuDNN, and TensorRT.

Jetson TX2 was designed for peak processing efficiency at 7.5W of power. This level of performance, referred to as Max-Q, represents the maximum performance and maximum power efficiency range on the power/performance curve. Every component on the module, including the power supply, is optimized to provide the highest efficiency at this point. The Max-Q frequency for the GPU is 854MHz. For the ARM A57 CPUs, it's 1.2GHz. While dynamic voltage and frequency scaling (DVFS) permits the NVIDIA Tegra® "Parker" system on a chip (SoC), which Jetson TX2 is based on, to adjust clock speeds at run time according to user load and power consumption, the Max-Q configuration sets a cap on the clocks to ensure that the application operates in the most efficient range only.

Jetson enables real-time inference when connectivity to an AI data center is either not possible (e.g., in the case of remote sensing) or the end-to-end latency is too high for real time use (e.g., in the case of autonomous drones). Although most platforms with a limited power budget will benefit most from Max-Q behavior, others may prefer maximum clocks to attain peak throughput, albeit with higher power consumption and reduced efficiency. DVFS can be configured to run at a range of other

clock speeds, including underclocking and overclocking. Max-P, the other preset platform configuration, enables maximum system performance in less than 15W. The Max-P frequency is 1.12GHz for the GPU and 2GHz for the CPU when either the ARM A57 cluster is enabled or the Denver 2 cluster is enabled and 1.4GHz when both clusters are enabled.

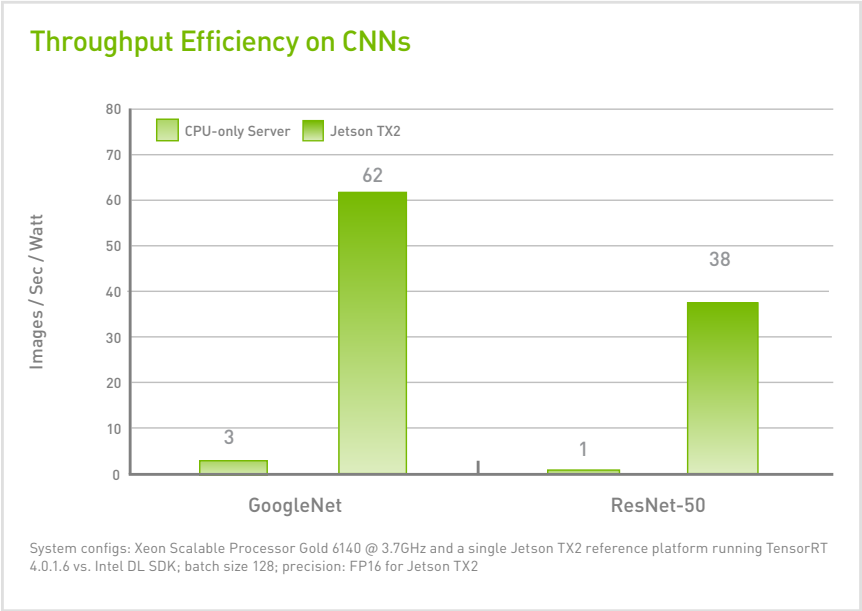

Chart 6

For many network-edge applications, low latency is a must-have. Executing inference on device is a far more optimal approach than trying to send this work over a wireless network and in and out of a CPU-based server in a remote data center. In addition to its on-device locality, Jetson TX2 also delivers outstanding low-latency on small-batch workloads, usually under 10ms. For comparison, a CPU-based server has a latency of around 23ms, and adding roundtrip network and data center travel time, that figure can be well over 100ms.

The Rise of Accelerated Computing

Google has announced its Cloud Tensor Processing Unit (TPU) and its applicability to deep learning training and inference. And while Google and NVIDIA chose different development paths, there are several themes common to both approaches. Specifically, AI requires accelerated computing. Accelerators provide the significant data processing necessary to keep up with the growing demands of deep learning in an era when Moore's law is slowing. Tensor processing—a major new workload that enterprises must consider when building modern data centers—is at the core of delivering performance for deep learning

training and inference. Accelerating tensor processing can dramatically reduce the cost of building modern data centers.

According to Google, the TPUv2 (also referred to as "TPU 2.0") has become available as a "Cloud TPU, which consists of four TPUv2 chips., But comparing chip to chip, a single TPU chip can deliver 45TFLOPS of computing horsepower per chip. NVIDIA's Tesla V100 can deliver 125TFLOPS of deep learning performance for both training and inference. An 8-GPU configuration such as NVIDIA DGX-1™ can now deliver a petaflop (PFLOP) of deep learning computing power.

NVIDIA's approach democratizes AI computing for every company, every industry, and every computing platform and accelerates every development framework—from the cloud, to the enterprise, to cars, and to the edge. Google and NVIDIA are the clear leaders, collaborating closely while taking different approaches to enable the world with AI.

Note on FPGAs

As the deep learning field continues to grow rapidly, other types of hardware have been proposed as potential solutions for inference, such as field-programmable gate arrays (FPGA). FPGAs are used for specific functions in network switches, 4G base stations, motor control in automotive, and test equipment in semiconductors, among other use cases. It's a sea of general-purpose programmable logic gates designed to simulate an application-specific integrated circuit (ASIC) for various usages, so long as the problem fits on the chip. But because these are programmable gates rather than a hard-wired ASIC, FPGAs are inherently less efficient.

At its recent Build conference, Microsoft claimed its FPGA-based Project BrainWave inference platform could deliver about 500 images per second on the ResNet-50 image network. However, to put this in perspective, a single Tesla P4 GPU can deliver more than 3X that throughput, or 1,676 images per second in a 75W solution. To compare further, shown here are projections made in a recent Intel whitepaper regarding their Altera and Stratix FPGAs. Note these results are run on the GoogleNet network.

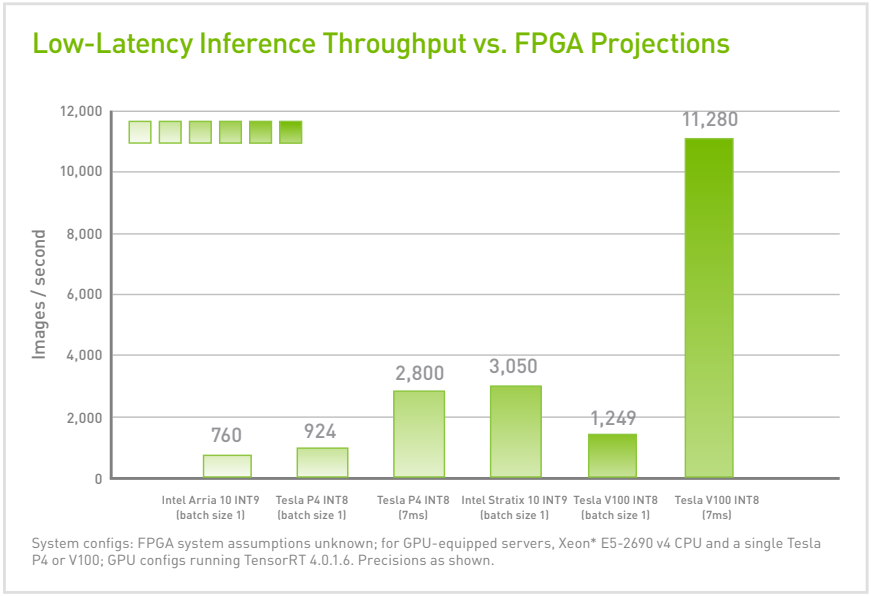

Chart 7

Programmability and Time to Solution Considerations

The speed of deep learning innovation drives the need for a programmable platform that enables developers to quickly try new network architectures, and iterate as new findings come to light. Recall that Programmability is the "P" in the PLASTER framework. There has been a Cambrian explosion of new network architectures that have emerged over the last several years, and this rate of innovation shows no signs of slowing.

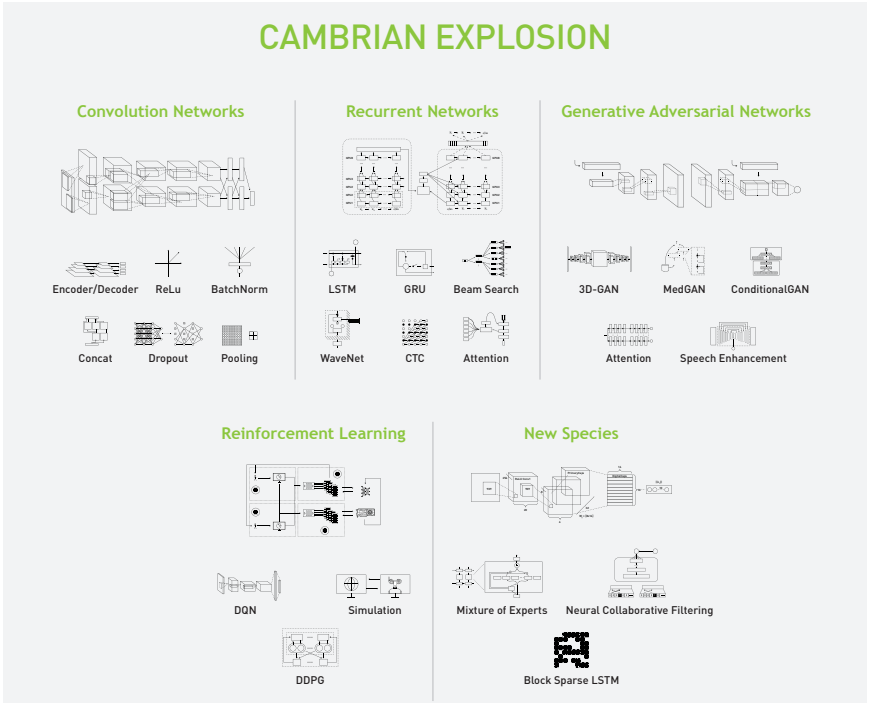

Figure 6

Another challenge posed by FPGAs is that, in addition to software development, they must be reconfigured at the hardware level to run each iteration of new neural network architectures. This complex hardware development slows time to solution and, hence, innovation by weeks and sometimes months. On the other hand, GPUs continue to be the most programmable platform of choice for quickly prototyping, testing, and iterating cutting-edge network designs, thanks to robust framework acceleration support, dedicated deep learning logic like Tesla V100's Tensor Cores, and TensorRT to optimize trained networks for deployed inference.

Conclusion

Deep learning is revolutionizing computing, impacting enterprises across multiple industrial sectors. The NVIDIA deep learning platform is the industry standard for training, and leading enterprises are already deploying GPUs for their inference workloads, leveraging its powerful benefits. Neural networks are rapidly becoming exponentially larger and more complex, driving massive computing demand and cost. In cases where AI services need to be responsive, modern networks are too compute-intensive for traditional CPUs.

Inference performance has seven aspects best remembered by PLASTER: programmability, latency, accuracy, size of network, throughput, efficiency, and rate of learning. All are critical to delivering both data center efficiency and great user experiences. This paper demonstrates how Tesla GPUs can deliver up to a 200X reduction in servers needed in the data center for "offline inference" use cases. In fact, the savings in energy costs alone more than pays for the Tesla-powered server. And at the network's edge, Jetson TX2 brings server-class inference performance in less than 10W of power and enables device-local inference to significantly cut inference latency times. These big improvements will enable state-of-the-art AI to be used end-to-end in real-time services that include speech recognition, speech-to-text, recommender systems, text-to-speech and speech synthesis.

An effective deep learning platform must have three distinct qualities: 1) It must have a processor custom-built for deep learning. 2) It must be software-programmable. 3) And industry frameworks must be optimized for it, powered by a developer ecosystem that is accessible and adopted around the world. The NVIDIA deep learning platform is designed around these three qualities and is the only end-to-end deep learning platform. From training to inference. From data center to the network's edge.

To learn more about NVIDIA's Tesla products, visit: **www.nvidia.com/tesla**

To learn more about Jetson TX2, visit: **www.nvidia.com/object/embedded-systems.html**

To learn more about TensorRT and other NVIDIA development tools, visit: **developer.nvidia.com/tensorrt**

To see the extensive list of applications that already take advantage of GPU acceleration, visit: **www.NVIDIA.com/GPU-applications**

Performance Data Tables

CNNs		TESLA P4 (INT8)			
NETWORK	BATCH SIZE	PERFORMANCE (IMAGES/ SEC)	TOTAL BOARD POWER	PERFORMANCE/ WATT	LATENCY (ms)
GoogLeNet	1	923	37	24.9	1.1
	2	1,215	40	30.4	1.6
	4	1,631	42	38.8	2.5
	8	2,197	46	47.8	3.7
	64	3,118	63	49.5	20
	128	3,191	64	49.1	40
ResNet-50	1	569	44	12.9	1.8
	2	736	44	16.7	2.7
	4	974	49	19.9	4.1
	8	1,291	57	22.6	6.2
	64	1,677	63	26.6	38
	128	1,676	62	27	76
VGG-19	1	206	55	3.7	4.9
	2	280	53	5.3	7.1
	4	346	60	5.8	12
	8	398	65	6.1	20
	64	429	63	6.8	149
	128	430	62	6.9	298

CNNs		TESLA V100 (MIXED PRECISION)			
NETWORK	BATCH SIZE	PERF (IMGS/ SEC)	TOTAL BOARD POWER	PERFORMANCE/ WATT	LATENCY (MS)
GoogLeNet	1	1,249	110	7.3	1.1
	2	1,686	114	10	1.6
	4	2,716	107	21.1	1.7
	8	4,571	127	30.3	2.1
	64	10,830	254	35.3	7.1
	128	11,280	284	34.7	13
ResNet-50	1	476	120	4	2.1
	2	880	109	8.1	2.3
	4	1,631	132	12.4	2.5
	8	2,685	153	17.5	3
	64	5,877	274	21.4	11
	128	6,275	285	22	20
VGG-19	1	497	151	3.3	2
	2	793	194	4.1	2.5
	4	1,194	220	5.4	3.4
	8	1,488	254	5.9	5.4
	64	2,161	290	7.5	30
	128	2,208	291	7.6	58

CNN		TESLA V100 (INT8)			
NETWORK	BATCH SIZE	PERF (TOKENS / SEC)	TOTAL BOARD POWER	PERFORMANCE / WATT	LATENCY (ms)
OpenNMT	1	894	103	8.7	1.1
	2	1,260	126	10	1.6
	4	1,746	129	13.5	2.3
	8	2,901	168	17.3	2.8
	64	5,903	289	20.4	11
	128	6,259	294	21.3	20

RNN		TESLA V100 (MIXED PRECISION)			
NETWORK	BATCH SIZE	PERF (TOKENS / SEC)	TOTAL BOARD POWER	PERFORMANCE / WATT	LATENCY (ms)
OpenNMT	1	3,457	96	36	15
	2	4,791	100	47.9	21
	4	8,076	105	76.9	25
	8	13,475	108	124.8	30
	64	50,758	74	685.9	64
	128	70,022	84	833.6	93

JETSON TX2 (MAX-Q MODE)

NETWORK	BATCH SIZE	PERF (IMGS/SEC)	AP+DRAM POWER UPSTREAM* (WATTS)	AP+DRAM PERFORMANCE / WATT	GPU POWER DOWNSTREAM* (WATTS)	GPU PERFORMANCE / WATT	LATENCY (MS)
AlexNet	1	119	6.6	18.0	2.3	52	8.4
	2	188	6.6	28.4	2.6	72	10.6
	4	264	6.7	39.3	2.9	91	15.2
	8	276	6.1	45.1	2.8	99	29.0
	64	400	6.4	62.6	3.2	125	160.0
	128	425	6.4	66.4	3.2	132.6	301.3
GoogLeNet	1	141	5.7	24.7	2.6	54.3	7.1
	2	156	5.9	26.2	2.7	57.6	12.8
	4	170	6.2	27.7	2.8	59.8	23.5
	8	180	6.4	28.2	3.0	60.6	44.5
	64	189	6.6	28.8	3.1	61.6	337.8
	128	191	6.6	28.9	3.1	61.6	671.8
ResNet-50	1	64	5.4	11.9	2.3	28.3	15.6
	2	77	5.3	14.4	2.3	33	26.2
	4	81	5.4	15.1	2.3	34.8	49.4
	8	83	5.4	15.4	2.4	35.4	95.9
	64	89	5.5	16.2	2.4	37	715.5
	128	90	5.5	16.2	2.4	37.7	1,424.3
VGG-19	1	19	7.2	2.6	3	7	53.1
	2	22	7.2	3.0	3.1	6.9	93.1
	4	23	7.3	3.1	3.1	7.2	176.8
	8	23	7.2	3.2	3.1	7.3	351.3
	64	23	7.2	3.2	3.2	7.1	2,792.4
	128	23	7.1	3.2	3.2	7.2	5,660.6

*Up = upstream power (above voltage regulators), and Down = downstream power (below the voltage regulators)

JETSON TX2 (MAX-P MODE)

NETWORK	BATCH SIZE	PERF (IMGS/SEC)	AP+DRAM POWER UPSTREAM* (WATTS)	AP+DRAM PERFORMANCE / WATT	GPU POWER DOWNSTREAM* (WATTS)	GPU PERFORMANCE / WATT	LATENCY (MS)
AlexNet	1	146	8.9	16.3	3.62	41	6.85
	2	231	9.2	25.2	4.00	57.7	8.66
	4	330	9.5	34.8	4.53	72.9	12.12
	8	349	8.8	39.8	4.42	79.0	22.90
	64	515	9.5	54.1	5.21	98.8	124.36
	128	546	9.6	56.9	5.28	103	234.32
GoogLeNet	1	179	8.2	21.8	4.14	43.2	5.6
	2	199	8.6	23.2	4.36	45.6	10.1
	4	218	9.0	24.2	4.61	47.2	18.4
	8	231	9.3	24.8	4.83	47.8	34.7
	64	243	9.7	25.1	5.03	49	263.6
	128	244	9.6	25.3	5.02	48.6	52

ResNet-50	1	82	7.4	11.1	3.49	23	12.2
	2	98	7.5	13.0	3.63	26.9	20.5
	4	104	7.6	13.6	3.71	27.9	38.6
	8	107	8.0	13.4	3.95	27.1	74.8
	64	115	7.9	14.6	3.81	30.1	558.9
	128	115	7.9	14.6	3.82	30.1	1,113.2
VGG-19	1	23.7	10	2.3	5	5.0	42.2
	2	26.8	10	2.6	4.93	5.4	74.7
	4	28.2	10	2.7	4.97	5.7	142.0
	8	28.3	10	2.8	4.96	5.7	282.7
	64	28.7	10	2.8	5.16	5.6	2,226.7
	128	28.4	10	2.8	5.09	5.6	4,514.0

*Up = upstream power (above voltage regulators), and Down = downstream power (below the voltage regulators)

JETSON TX1

NETWORK	BATCH SIZE	PERF (IMGS/ SEC)	AP+DRAM POWER UPSTREAM* (WATTS)	AP+DRAM PERFORMANCE / WATT	GPU POWER DOWNSTREAM* (WATTS)	GPU PERFORMANCE / WATT	LATENCY (MS)
AlexNet	1	95	9.2	10.3	5.1	18.6	10.5
	2	158	10.3	15.2	6.4	24.5	12.7
	4	244	11.3	21.7	7.6	32.0	16.4
	8	253	11.3	22.3	7.8	32.0	31.6
	64	418	12.5	33	9.4	44.0	153.2
	128	449	12.5	36	9.6	46.9	284.9
GoogLeNet	1	119	10.7	11.1	7.2	16.4	8.4
	2	133	11.2	12.0	7.7	17.4	15.0
	4	173	11.6	14.9	8.0	21.6	23.2
	8	185	12.3	15.1	9.0	20.6	43.2
	64	196	12.7	15.0	9.4	20.7	327.0
	128	196	12.7	15.0	9.5	20.7	651.7
ResNet-50	1	60.8	9.5	6.4	6.3	9.7	16.4
	2	67.8	9.8	6.9	6.5	10.0	29.5
	4	80.5	9.7	8.3	6.6	12.1	49.7
	8	84.2	10.2	8.3	7.0	12.0	95.0
	64	91.2	10.0	9.1	6.9	13.2	701.7
	128	91.5	10.4	8.8	7.3	12.6	1,399.3
VGG-19	1	13.3	11.3	1.2	7.6	1.7	75.0
	2	16.4	12.0	1.4	8.6	1.9	122.2
	4	19.2	12.2	1.6	8.9	2.2	207.8
	8	19.5	12.0	1.6	8.6	2.3	410.6
	64	20.3	12.2	1.7	9.1	2.2	3,149.6
	128	20.5	12.5	1.6	9.3	2.2	3,187.3

*Up = upstream power (above voltage regulators), and Down = downstream power (below the voltage regulators)

Test Methodology

For our performance analysis, we focus on four neural network architectures. AlexNet (2012 ImageNet winner) and the more recent GoogLeNet (2014 ImageNet winner), a much deeper and more complicated neural network compared to AlexNet, are two classical networks. VGG-19 and ResNet-50 are more recent ImageNet competition winners.

To cover a range of possible inference scenarios, we will consider two cases. The first case allows batching many input images together to model use cases like inference in the cloud where thousands of users submit images every second. Here, large batches are acceptable, as waiting for a batch to assemble doesn't add significant latency. The second case covers applications that are extremely latency-focused; in this case, some batching is usually still feasible, but for our testing, we consider the low-batch case of a batch size of two.

We compare five different devices: the NVIDIA Tegra X1 and X2 client-side processors, the NVIDIA Tesla P4 and V100, and the Intel Xeon data center processor. To run the neural networks on the GPU, we use TensorRT 2 EA, which will be released in a JetPack update slated for release in 2Q'17. For the Intel Xeon Scalable Processor Gold 6140, we run the Intel Deep Learning SDK v2016.1.0.861 Deployment Tool.

For all the GPU results, we run the "giexec" binary included in all builds of TensorRT. It takes the prototxt network descriptor and Caffe model files and populates the images with random image and weight data using a Gaussian distribution. For the CPU results, we run the "ModelOptimizer" binary with the prototxt network descriptor and Caffe model files to generate the .xml model file necessary to execute the "classification_sample" binary linked with MKL-DNN. We run the Intel Deep Learning SDK Inference Engine using images from imagenet12 rescaled and reformatted to RGB .bmp files. Both TensorRT and the Intel Deep Learning SDK Inference Engine use image sizes of 227 x 227 for AlexNet and 224 x 224 for GoogleNet, VGG-19, and ResNet-50. The Intel Deep Learning SDK Inference Engine threw the "bad_alloc" exception when running with a batch size of one for all networks we tested. Instead, we use Intel Caffe for a batch size of one linked with MKL 2017.1.132 where we start with the default_vgg_19 protocol buffer files and use Caffe's standard performance benchmarking mode "caffe time" with the same images as the Intel Deep Learning SDK.

We compare FP16 mixed-precision results on V100 and INT8 results on P4. All Tegra X1 and X2 results are using FP16. Intel Deep Learning SDK only supports FP32.

To compare power between different systems, it's important to measure power at a consistent point in the power distribution network. Power is distributed at a high voltage (pre-regulation), and then voltage regulators convert the high voltage to the correct level for the system on chip and DRAM (post-regulation). For our analysis, we are comparing pre-regulation power of the entire application processor (AP) and DRAM combined.

On the Intel Xeon Scalable Processor Gold 6140, the Intel OpenVINO libraries are running on a single-socket config for CNN testing (GoogleNet, ResNet-50 and VGG19). For RNN testing, we used Intel's Deep Learning SDK, since OpenVINO only supports CNNs. CPU socket and DRAM power are as reported by the pcm-power utility, which we believe are measured on the input side of the associated regulators. To measure pre-regulation (upstream) power for Tegra X1 and X2, we use production Jetson TX1 and TX2 modules, both powered by a 9V supply. TX1 has major supply rails instrumented at the input side of the regulators, and TX2 has onboard INA power monitors. On the Tesla P4 and V100, we report the total board power consumed by a production cards using the NVSMI utility. We don't include the system CPU's power in our Tesla measurements, as the entire computation is happening on the GPU; the CPU only submits the work to the GPU.

Legal Notices and Trademarks

conditions with regard to the purchase of the NVIDIA product referenced in this specification. NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk. NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to:(i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs. No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification Information published by NVIDIA regarding third party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.