

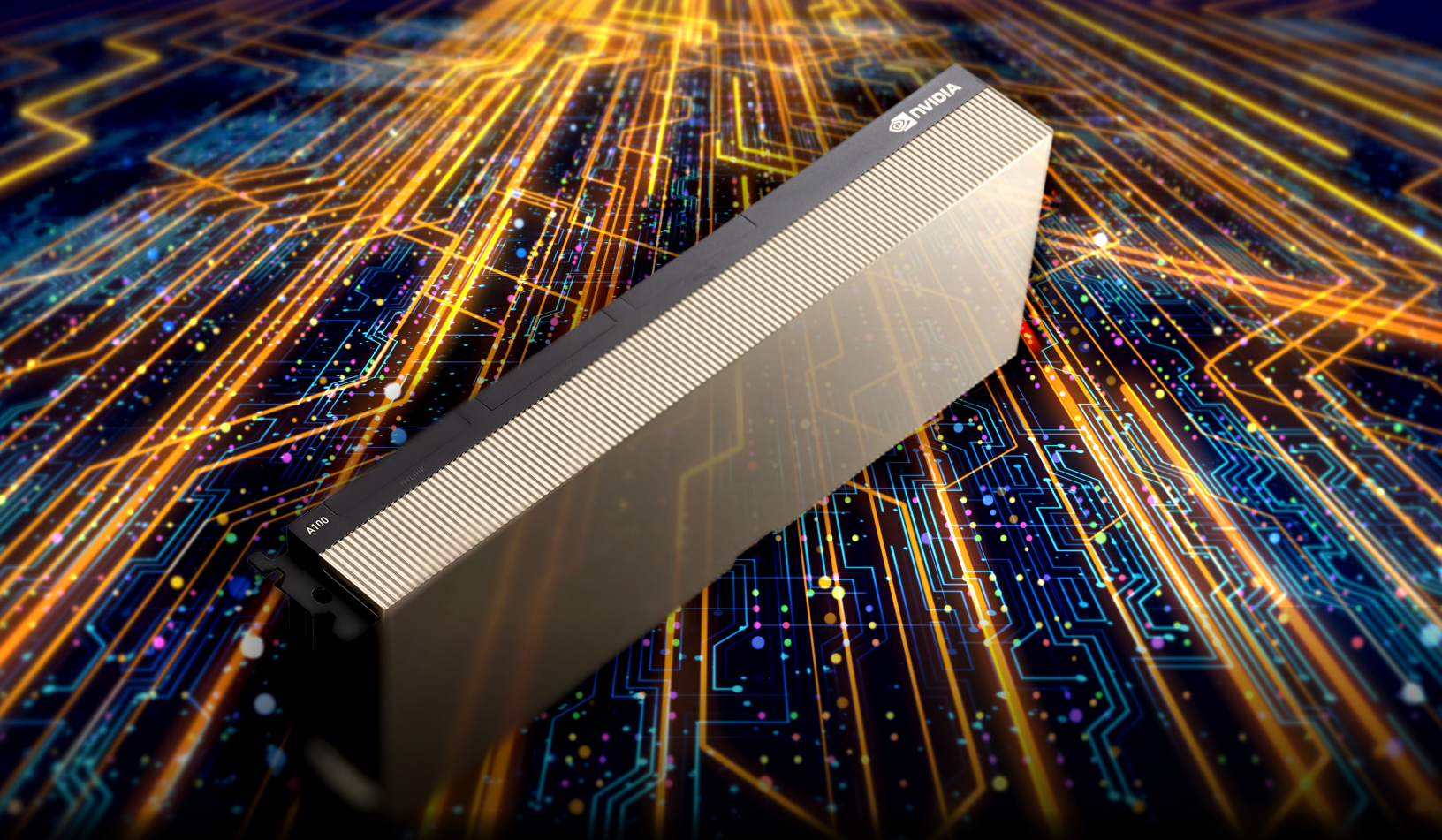


PNY



EXXACT

NVIDIA RTX & DATA CENTER GPU
Benchmarks for Deep Learning



CPUs AND GPUS ARE SIMILAR IN THE SENSE THAT BOTH ARE SILICON-BASED MICROPROCESSORS PROVIDING CRITICAL COMPUTING ENGINES CAPABLE OF HANDLING DATA.

With so many workstation configuration options for deep learning and life sciences, how do you know which will provide optimal results or significant performance increases? This white paper will compare results for image recognition by adding various NVIDIA® RTX™ workstation GPUs (graphics processing units), as well as AMBER 22 benchmarks using NVIDIA Ampere architecture-based data center GPUs.

The two most important components in deep learning and life sciences workstations are the CPU (central processing unit) and GPU. A CPU is commonly referred to as the brain of a workstation and is essential to all computing systems. It is like the taskmaster of the overall system and coordinates a wide range of general-purpose computing tasks, along with others such as input/output (I/O) operations, basic arithmetic functions, and logic operations. GPUs were originally designed to create images for computer graphics and video games, but as they became more technically sophisticated proved able to accelerate calculations involving massive amounts of data by running parallel algorithms far more

efficiently than any CPU. The GPU complements the CPU by allowing essential calculations within an application to be run on the massively parallel GPU, while the main program continues to run on the CPU, resulting in increased performance and data throughput.

CPUs and GPUs are similar in the sense that both are silicon-based microprocessors providing critical computing engines capable of handling data. They have different architectures and are built for different purposes. A CPU is designed with far fewer cores than a GPU, but the individual cores are faster (measured by clock speeds) and smarter (measured by available instruction sets), making it able to handle a wide range of tasks quickly, but limited in the number of tasks that can run concurrently. A GPU is designed with thousands of processor cores that run simultaneously, which enables massive parallelism where each core is focused on making efficient calculations. GPUs can complete more work in the same amount of time as a CPU. This makes GPUs ideally suited for repetitive and highly parallel computing tasks, such as deep learning and molecular dynamics

simulations. GPUs have been able to dramatically boost deep learning performance and accelerate neural network training of real-world use cases; all while considerably reducing the time required and lowering hardware acquisition costs, and TCO (Total Cost of Ownership) when factors like performance per watt and the number of servers required is taken into consideration.

Because the GPU does not work in isolation and complements the CPU, it is important to make sure the two implement a balanced architecture. To maximize a GPU's potential, it requires an equally powerful CPU. The CPU controls the management and assignment of work to the GPUs, while the GPUs do the heavy lifting of transforming, loading, and analyzing data. If the CPU is unable to perform its function fast enough, it causes a CPU bottleneck that affects the amount of data the GPU gets to process. With molecular dynamics programs utilizing GPU acceleration, it not only greatly impacts which GPU to use, but also the CPU to choose that allows maximum GPU performance.

RESNET-50

ResNet-50 is an AI benchmark for image classification and is a popular standard for measuring performance of machine learning accelerators. It is a convolutional neural network that is 50 layers deep where users can load a pre-trained version of the network trained on more than ten million images from the ImageNet database. The pretrained network can classify images into 1,000 object categories, such as different animals and vehicles. Users benefit from these abilities to make predictions using a trained neural network for deep learning on either a CPU or GPU.





For this particular ResNet-50 testing environment, Exxact Corporation's TensorEX TS4-173535991 EMLI 4U system was used. EMLI, or Exxact Machine Learning Images, is a customizable production ready, open source machine learning environment for accelerating AI research prototyping and production deployment. Deep learning systems with EMLI ship with the latest AI development tools installed in a way that best suits developers' needs, whether they prefer containerized environments or natively installed frameworks.

The testing environment technical specifications are listed below:

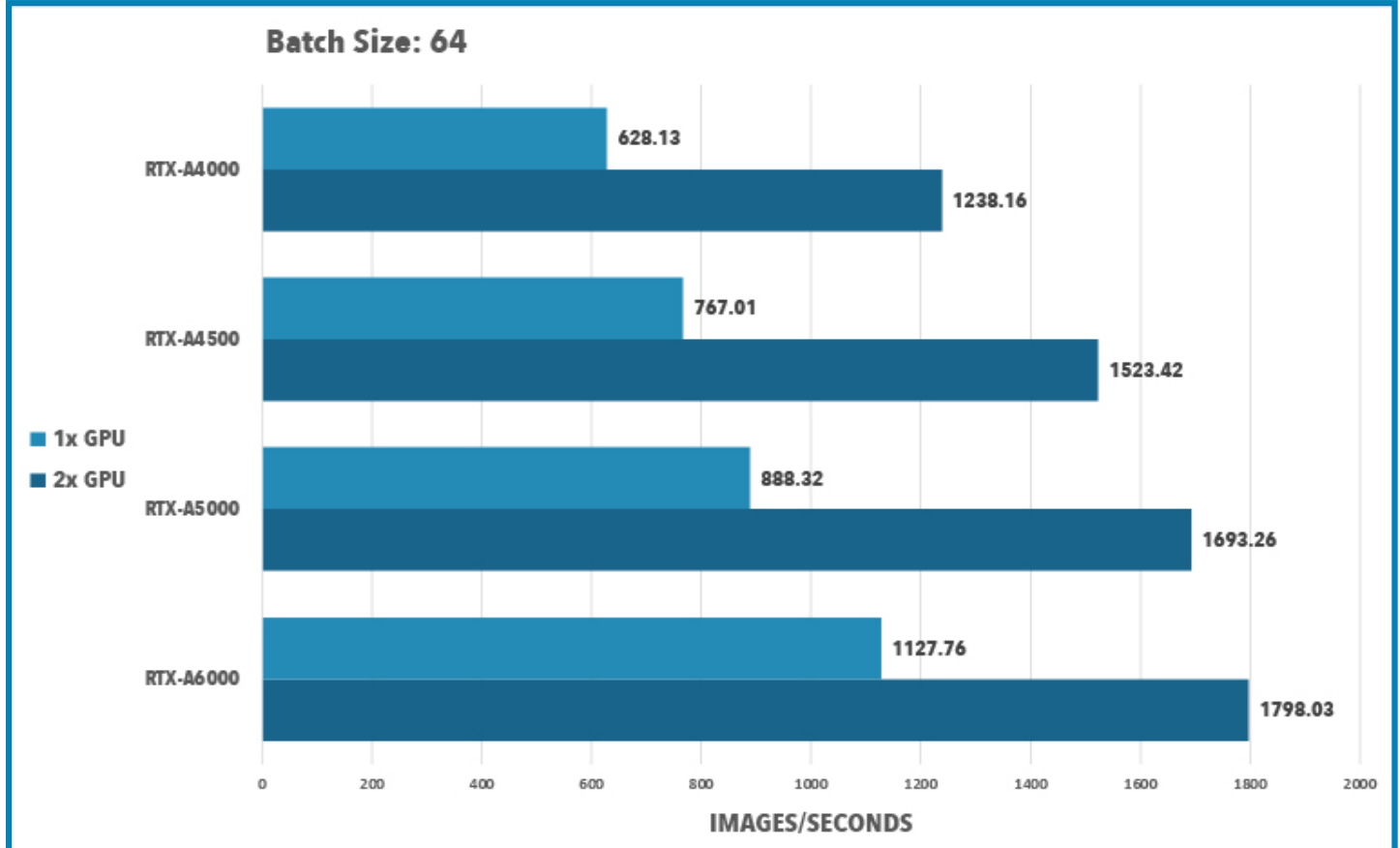
- CPU: 2x AMD EPYC™ 7352 24-Core
- Operating System: Ubuntu 20.04
- Graphics Driver: NVIDIA Driver 465.27
- Memory: 512 GB DDR4 ECC
- CUDA Toolkit: 11.2
- CuDNN Version: 7 (NVIDIA CUDA® Deep Neural Network library)
- TensorFlow Version: 1.15

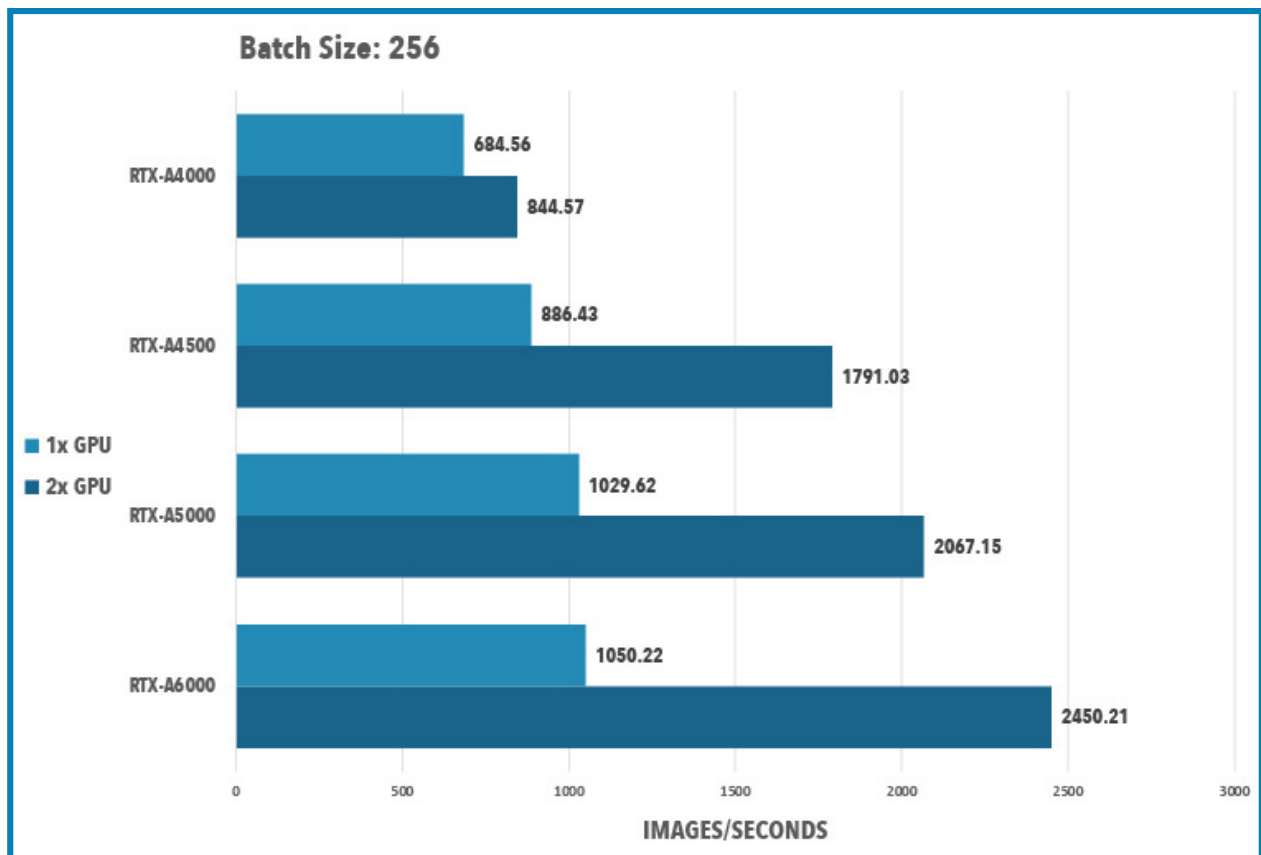
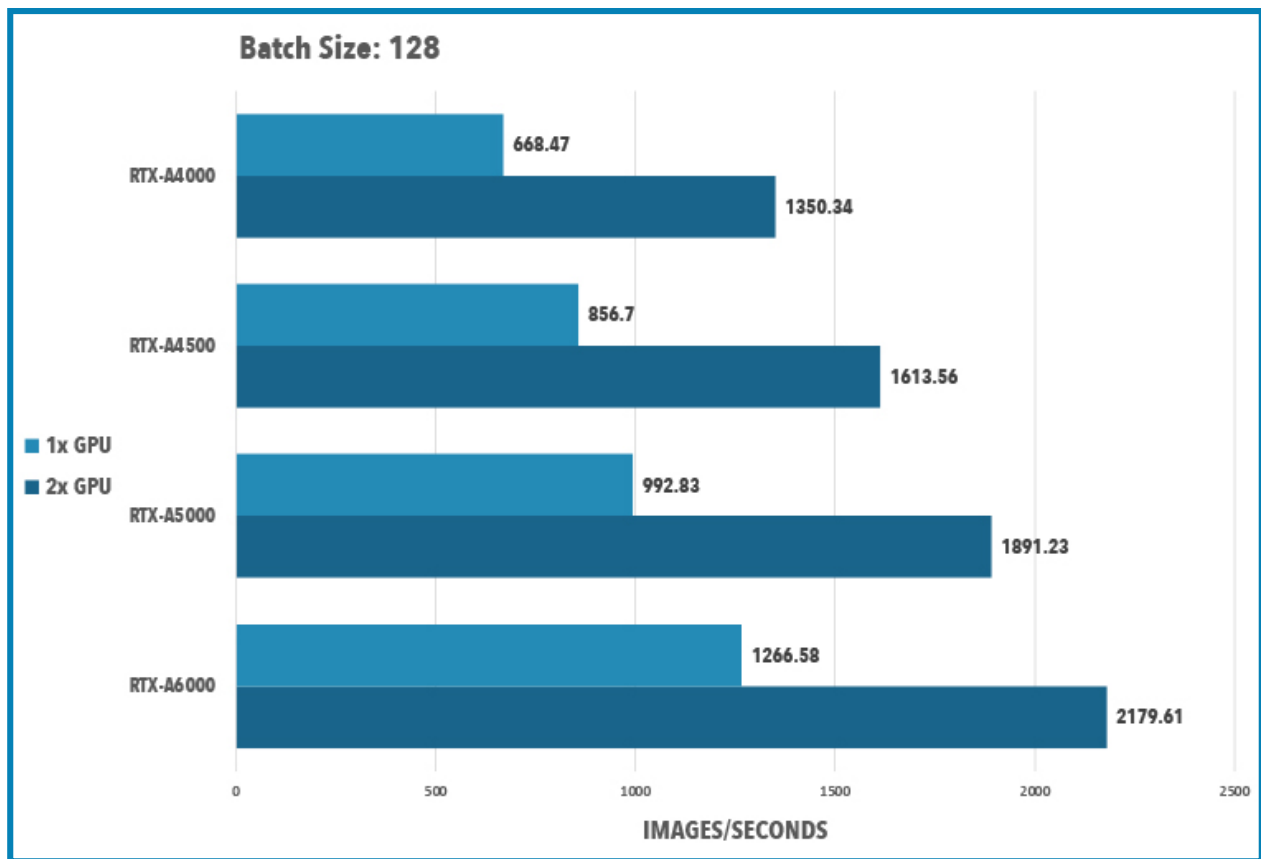
AMD's EPYC CPU sets superior standards for performance, security, and scalability for the most demanding workloads. This dual AMD EPYC CPU powered system was designed to run one or more high performance GPUs, and deliver optimized performance per Watt, large L3 cache for low latency access to data, and an industry leading 8 channels of DDR4-3200 memory with up to 128 lanes of PCIe® 4.0 connectivity to reduce bottlenecks. With an optimal cooling system and PCIe 4.0 slots directly connected to the CPU to support each GPU, any of the NVIDIA Ampere-based GPUs installed benefit by doubling the data transfer rates between the CPUs and GPUs.

NVIDIA RTX PROFESSIONAL WORKSTATION GRAPHICS BOARDS

	NVIDIA RTX A4000	NVIDIA RTX A4500	NVIDIA RTX A5000	NVIDIA RTX A6000
PRODUCT				
SUITABLE FOR	AI Training/Inference, Life Sciences, Visualization	Pro-viz, AI Training/ Inference, Life Sciences, Visualization	Pro-viz, AI Training/ Inference, Life Sciences, Visualization	Pro-viz, AI Training/ Inference, Life Sciences, Visualization
MEMORY	16 GB GDDR6	20 GB GDDR6	24 GB GDDR6	48 GB GDDR6
FORM FACTOR	Full Height, Half Width	Full Height, Full Width	Full Height, Full Width	Full Height, Full Width







For the benchmarks below, batch size specifies how many propagations of the network are done in parallel, the results of each propagation are averaged among the batch, and then the result is applied to adjust the weights of the network. Using an optimal batch size is one way to optimize the workload for each GPU. As the batch sizes expand in size, parallelism increases and GPU core utilization improves. It is important that the batch size does not exceed available GPU memory. Doing so – in most cases – will reduce performance or possibly even cause the application to crash.





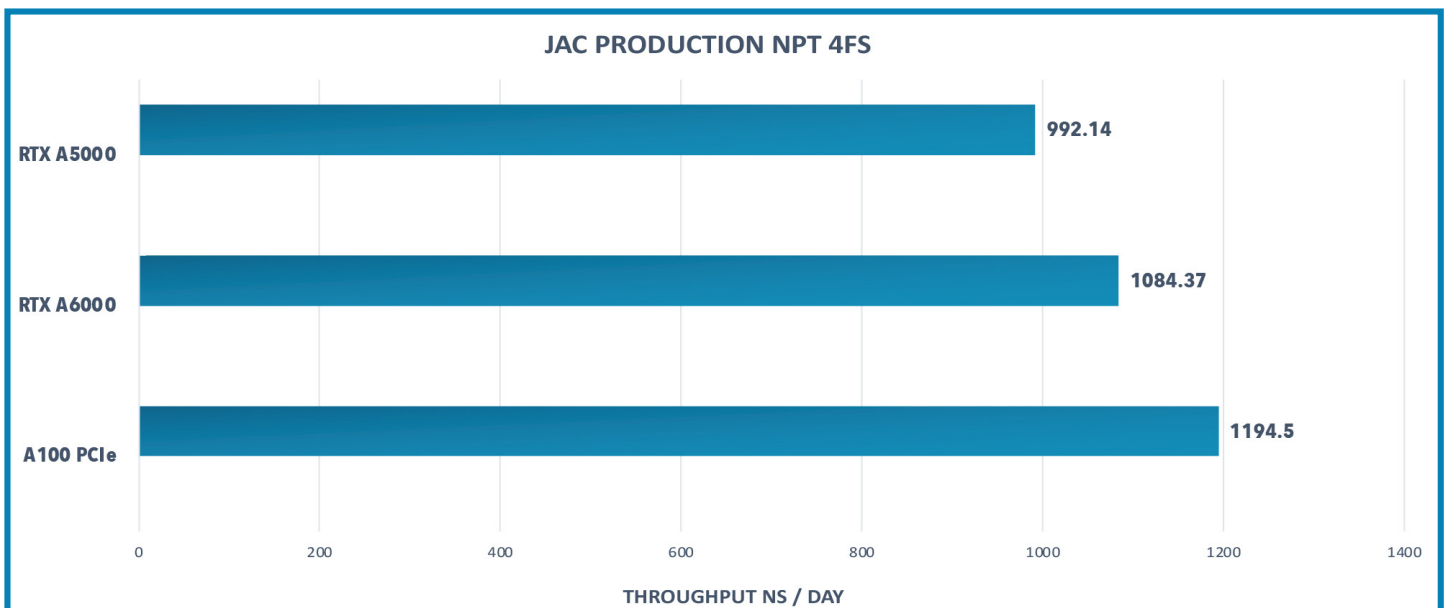
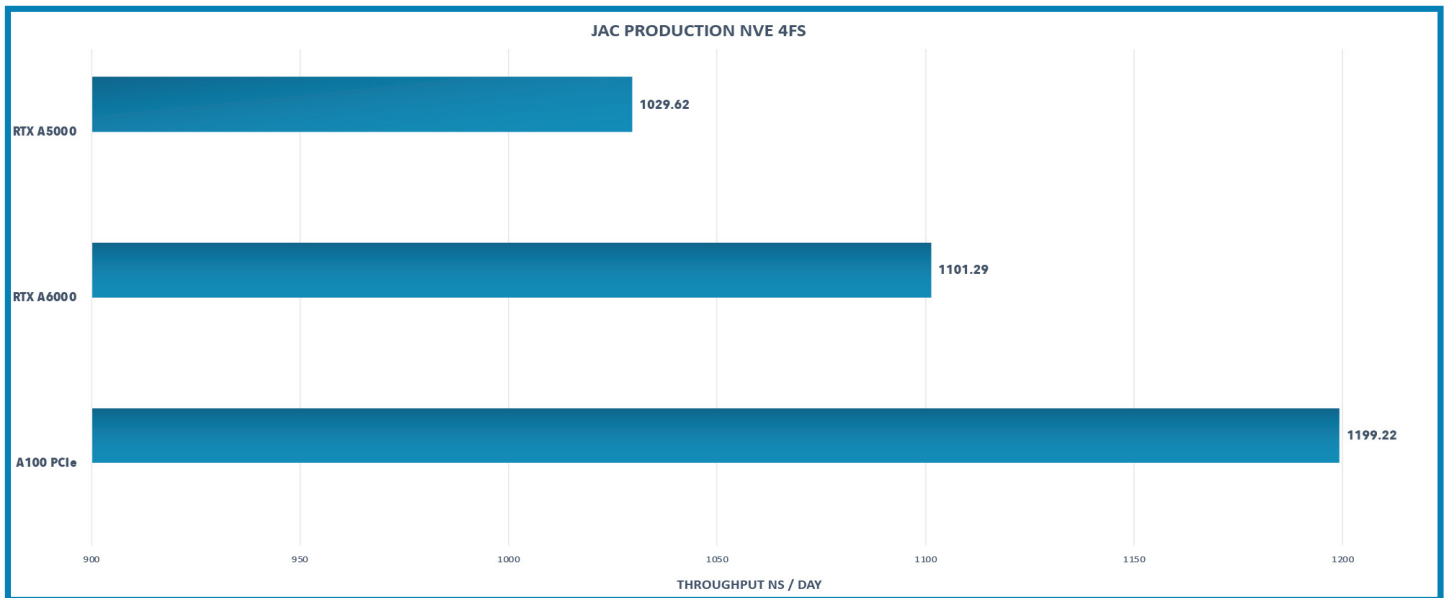
NVIDIA DATA CENTER GPUs

In addition to deep learning, GPUs have a firm foothold in data science and can be used to accelerate the processing and visualization of large dataframes. NVIDIA products like the A100 80GB PCIe and A30 provide the artificial intelligence (AI) and high-performance computing (HPC) capabilities required to solve the world's most pressing scientific, industrial, and business challenges. The NVIDIA A40 renders innovative product or architectural concepts and advances M&E digital storytelling. For easily and flexibly virtualizing remote workforces the NVIDIA A16 is unmatched. A10 brings accelerated graphics and video with AI for mainstream enterprise servers. Finally, the NVIDIA A2 brings entry-level GPU acceleration to almost any industry standard CPU-based server – in the data center or at the edge.

NVIDIA A2	NVIDIA A10	NVIDIA A16	NVIDIA A30	NVIDIA A40	NVIDIA A100
					
BENEFITS					
Ideal for entry-level high-density server or at the edge deployments and is particularly suited for IVA (Intelligent Video Analytics) and GPU accelerated AI inference in any server. Low profile, low power for thermally constrained systems	Power high-performance design and engineering virtual workstations to AI in an easily managed, secure and flexible infrastructure that scales to accommodate any need.	Take remote work to the next level with GPU-accelerated virtual desktops powered by NVIDIA virtual GPU software.	The most versatile mainstream compute GPU for diverse workloads. Ideal for inference at scale, it delivers maximum value for mainstream enterprises.	Ideal for AEC, Manufacturing, or Media & Entertainment contexts where real-time photorealistic ray tracing is required, along with AI and powerful compute capabilities.	Unprecedented AI and HPC acceleration at every scale for mainstream enterprise servers.
MEMORY					
16 GB GDDR6	24 GB GDDR6	64 GB GDDR6 w/ ECC	24 GB HBM2 w/ ECC	48 GB GDDR6 w/ECC	80 GB HBM2e w/ ECC
FORM FACTOR					
Full Height, Half Width	Full Height, Full Width	Full Height, Full Width	Full Height, Full Width	Full Height, Full Width	Full Height, Full Width

The AMBER 22 benchmarks below compare the NVIDIA A100 80GB PCIe data center GPU with a two top-tier Ampere architecture-based NVIDIA RTX workstation GPUs, specifically the NVIDIA RTX A6000 and RTX A5000. The benchmarks were divided by NVE, which is ideal for dynamical properties, and NPT, which is preferred when calculating thermo-physical properties like density. The NVIDIA A100 80 GB PCIe (as expected) outperformed even the most powerful RTX workstation GPU.

AMBER 22 GPU Benchmark: JAC Production NVE 4fs



Final Thoughts

For the ResNet-50 inference benchmarks in this white paper, swapping an NVIDIA RTX professional workstation GPU for the next level up may not yield a significant performance increase. Installing an additional NVIDIA RTX GPUs typically resulted in double the images classified per second. When debating whether to upgrade or add another GPU, adding a second is the recommended upgrade path. NVLink can be used to bridge two RTX GPUs at greater than PCIe Gen 4 bandwidth, but application support is required to realize any performance benefit.

For AMBER 22 benchmarks, the NVIDIA A100 80GB data center GPU outperformed even the most powerful RTX workstation GPU – the NVIDIA RTX A6000. The NVIDIA A100 80GB Tensor Core GPU delivers unprecedented acceleration at every scale to power the world's highest-performing elastic data centers for AI, data analytics, and HPC.

However, it should be noted that benchmarks may not accurately represent real-world use cases. For example, using a ResNet-50 trained model for a security system that has many video streams as the input might require a larger GPU to ingest, process, and validate large numbers of video streams. A benchmark may not take factors like this into account.