NVIDIA GPU & Dell EMC Server Recommendations by Workload

Tech Note by

Matt Ogle
Ramesh Radhakrishnan

Summary

The NVIDIA product portfolio includes GPU models to address different use cases and applications. Deciding which GPU model and Dell EMC server to purchase based on intended workloads can become very complex for customers looking to leverage GPU acceleration.

Workload categories that leverage GPUs to improve application performance and achieve better TCO include compute intensive use cases like AI training and inference, High-Performance Computing (HPC) and Database Analytics. VDI, rendering and ray tracing are use cases that leverage the graphical computing capability of GPUs.

This DfD will educate Dell EMC customers on four popular NVIDIA GPU models and how-to best pair them to PowerEdge servers based on the intended workload.

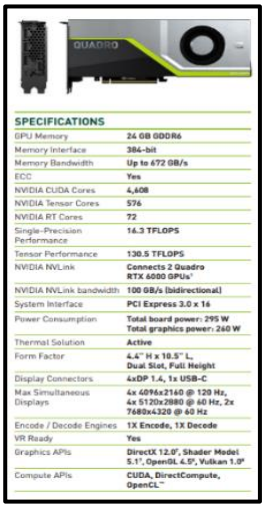
PowerEdge servers support various NVIDIA GPU models. Each model is designed to accelerate demanding applications by acting as a powerful assistant to the CPU. It is vital to understand which GPUs and PowerEdge products work best together to enable breakthrough performance for an intended workload. This paper will compare four popular NVIDIA GPUs on the market today, as shown in Figure 1, as well as educate Dell EMC customers on PowerEdge servers and specific workloads best suited for each GPU.

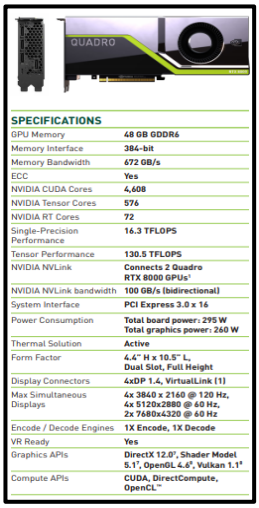
GPU Model	CUDA Cores	Single Precision (FP32)	Mixed Precision (FP16/FP32)	Double Precision (FP64)	Memory Size / Bus	Memory Bandwidth	Power Consumption
RTX6000	4608	15 TFLOPS	120 TFLOPS	N/A	24GB GDDR6	624 GB/s	250W
RTX8000	4608	15 TFLOPS	120 TFLOPS	N/A	48GB GDDR6	624 GB/s	250W
T4	2560	8.1 TFLOPS	65 TFLOPS	N/A	16GB GDDR6	300 GB/s	70W
V100 *(PCIe)*	5120	14 TFLOPS	112 TFLOPS	7 TFLOPS	32GB HBM2	900 GB/s	250W
V100 *(SXM2)*	5120	15.7 TFLOPS	125 TFLOPS	7.8 TFLOPS	32GB HBM2	900 GB/s	300W
V100S	5120	16.4 TFLOPS	130 TFLOPS	8.2 TFLOPS	32GB HBM2	1134 GB/s	250W
M10	2560	5 TFLOPS	N/A	N/A	32GB GDDR5	332 GB/s	225W

Figure 1 – Table comparing popular NVIDIA GPU specifications

1. Quadro RTX 6000 & 8000

The latest additions to the NVIDIA datacenter roadmap are the RTX 6000 and 8000. The Quadro RTX 6000/8000 will best accelerate performance graphics, render farms and Edge computing. In addition to having high CUDA core counts, memory speeds and floating-point performances, these GPUs have unique features that make them ideal for graphics, such ray tracing cores and NVLINK capability for supporting large memory capacities.





Figures 2 & 3 – RTX 6000 (left) and RTX 8000 (right) specifications

It is important to remember that the workload dictates which server to choose for best results. The RTX 6000/8000 supports high-performance graphics workloads and optimizing this type of workload will require sourcing as many GPUs as possible into datacenter racks. For this reason, we recommend the DSS8440 as a first option, as it can support up to 10 GPUs, with the R740 and R7525 as second options, which are commonly used compute nodes in render farms.

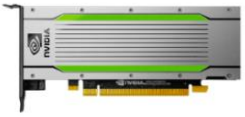
Supported Workloads: *Graphics, Render Farms, Edge Computing Training, AI Inference, IVA, VDI, Video Analytics*
Recommended Workloads: *Graphics, Render Farms, Edge Computing*
Recommended PowerEdge Servers: *DSS8440, R740, R7525*

2. Tesla T4

The Tesla T4 will best accelerate AI inference, training, general-purpose compute applications and graphics. The T4 introduced the Turing Tensor Core technology with multi-precision computing ranging from FP32/FP16 for floating point arithmetic to INT8/INT4 integer precision capability to handle diverse workloads. With low power consumption, modest pricing and a low-profile, single-width form factor, the T4 is both versatile in functionality and easy to integrate into most PowerEdge servers, making it ideal for accelerating general purpose workloads. It is an optimized solution for workloads that don't need high precision (FP64) capabilities.

The servers that we recommend populating with T4s are the R640, R740, R740xd and DSS8440. Users can add 1-2 T4 GPUs for inference on R640, 1-6 T4 GPUs on the R740(xd) for more demanding applications and up to 16 T4 GPUs on the DSS8440 for applications requiring highly dense GPU compute capability.

Supported Workloads: *AI Training, AI Inference, IVA, VDI, Video Analytics, General Purpose Computing*
Recommended Workloads: *AI Inference, General Purpose Computing*
Recommended PowerEdge Servers: *R640, R740, R740xd, DSS8440*



SPECIFICATIONS

GPU Architecture	NVIDIA Turing
NVIDIA Turing Tensor Cores	320
NVIDIA CUDA® Cores	2,560
Single-Precision	8.1 TFLOPS
Mixed-Precision (FP16/FP32)	65 TFLOPS
INT8	130 TOPS
INT4	260 TOPS
GPU Memory	16 GB GDDR6 300 GB/sec
ECC	Yes
Interconnect Bandwidth	32 GB/sec
System Interface	x16 PCIe Gen3
Form Factor	Low-Profile PCIe
Thermal Solution	Passive
Compute APIs	CUDA, NVIDIA TensorRT™, ONNX

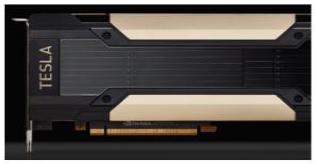
Figure 4 – T4 specifications

3. Tesla V100

The V100 will best accelerate high performance computing (HPC) and dedicated AI training workloads. The V100 is equipped with the double-precision performance required by various HPC applications such as engineering simulation, weather prediction and molecular dynamics . The V100 is also equipped with 32GB of memory that can run at 900GB/s to support the memory bandwidth requirements of HPC workloads. The V100S is the latest addition to the V100 family and can speed up HPC applications with its increased memory bandwidth capability of 1134 GB/s. AI training workloads leverage the processing capability of multi-GPUs using scale-out distributed training techniques to improve performance. Using the V100 SXM2 GPU with the NVLink capabilities enables direct communication between GPUs with bandwidth of up to 300GB/s; further increasing performance of AI training workloads.

The Tesla V100 powered by NVIDIA Volta architecture is the most widely used accelerator for scientific computing and artificial intelligence. HPC and scientific computing workloads are recommended to use the V100/V100S PCIe in R740 (1-3GPUs), R7425(1-3GPUs) and PowerEdge C4140 (4 GPUs). Deep Learning training workloads can leverage NVLink capability of the V100 SXM2 GPUs on the C4140 with NVLink capabilities or DSS8440 that support up to 10 V100 PCIe GPUs. The R840 and R940xa combine larger server memory capacities and GPU acceleration for accelerating Analytics workloads

Supported Workloads: *HPC, AI Training, AI Inference, VDI, Video Analytics*
Recommended Workloads: *HPC, Dedicated AI Training*
Recommended PowerEdge Servers: *C4140, R7425, R840, R940xa, DSS8440*



SPECIFICATIONS

	Tesla V100 PCIe	Tesla V100 SXM2	Tesla V100S PCIe
GPU Architecture	NVIDIA Volta		
NVIDIA Tensor Cores	640		
NVIDIA CUDA® Cores	5,120		
Double-Precision Performance	7 TFLOPS	7.8 TFLOPS	8.2 TFLOPS
Single-Precision Performance	14 TFLOPS	15.7 TFLOPS	16.4 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS	130 TFLOPS
GPU Memory	32GB /16GB HBM2		32GB HBM2
Memory Bandwidth	900GB/sec		1134 GB/sec
ECC	Yes		
Interconnect Bandwidth	32 GB/sec	300 GB/sec	32 GB/sec
System Interface	PCIe Gen3	NVIDIA NVLink	PCIe Gen3
Form Factor	PCIe Full Height/Length	SXM2	PCIe Full Height/Length
Max Power Consumption	250 W	300 W	250 W
Thermal Solution	Passive		
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC		

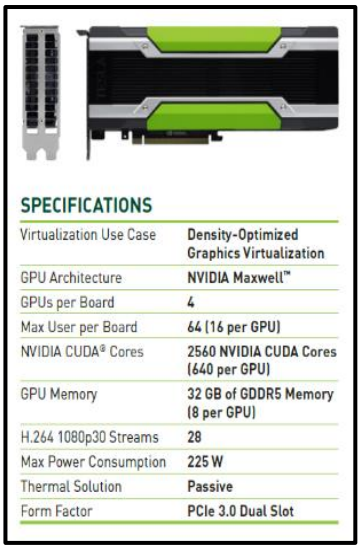
Figure 5 – V100 specifications

4. Tesla M10

The M10 will best accelerate Virtual Desktop Infrastructure (VDI) and mainstream graphics applications. This legacy GPU has maintained popularity with customers due to its large 32GB memory capacity and affordable price point, giving it a compelling TCO for VDI and mainstream graphics.

For VDI applications, we suggest running the M10 on a PowerEdge server that has enough CPU cores, memory and PCIe cores to support a large number of virtual desktop users, such as the R7425. For mainstream graphics we suggest a 2U PowerEdge server that has a high frequency CPU and adequate PCIe slots for population, such as the R740 or R740xd.

Supported Workloads: *VDI, Mainstream Graphics, IVA, AI Training, AI Inference, General Purpose Computing*
Recommended Workloads: *VDI, Mainstream Graphics*
Recommended PowerEdge Servers: *R740, R740xd, R7425*



SPECIFICATIONS

Virtualization Use Case	Density-Optimized Graphics Virtualization
GPU Architecture	NVIDIA Maxwell™
GPUs per Board	4
Max User per Board	64 (16 per GPU)
NVIDIA CUDA® Cores	2560 NVIDIA CUDA Cores (640 per GPU)
GPU Memory	32 GB of GDDR5 Memory (8 per GPU)
H.264 1080p30 Streams	28
Max Power Consumption	225 W
Thermal Solution	Passive
Form Factor	PCIe 3.0 Dual Slot

Figure 6 – M10 specifications

Conclusion

The NVIDIA GPU catalog offers a wide variety of GPU models that were designed to accelerate diverse workloads. A properly configured server will enable the workloads to utilize the capabilities of a GPU working in concert with other system components to yield the best performance. In this DfD we have discussed the value proposition of four popular NVIDIA GPU models, as well as what Dell EMC servers and workloads would work best for each.

Learn More

DellEMC GPU eBook
Demystifying Deep Learning Infrastructure Choices using MLPerf Benchmark Suite
HPC at Dell

PowerEdge DfD Repository
For more technical learning

Contact Us
For feedback and requests

Follow Us
For PowerEdge news