NVIDIA TESLA
ONE PLATFORM. UNLIMITED DATA CENTER ACCELERATION.

The Exponential Growth of Computing

We are at the dawn of a new age of intelligence, where AI and high performance computing (HPC) are transforming our world. From autonomous vehicles to global climate simulations, new challenges are emerging that demand enormous computing resources to solve. Similarly, life-like graphics and IT remote infrastructure applications require massive processing capability. With the slowing down of Moore's law, a new platform is needed that can provide unprecedented computing capability for scientists and engineers to run these diverse applications.

NVIDIA GPUs have evolved to become the essential tool at the intersection of these applications, powerful enough to support even the most demanding of tasks. NVIDIA® Tesla® is a world-leading GPU platform, adopted specifically for the data center. Deployed by the largest supercomputing centers and enterprises, it enables breakthrough performance with fewer, more powerful servers, resulting in faster scientific discoveries and productivity while saving money.

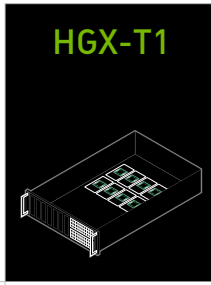
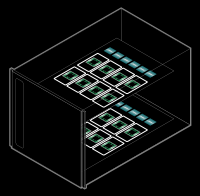
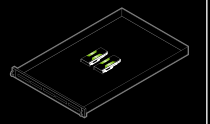
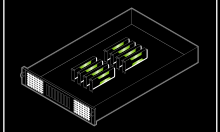
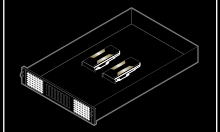
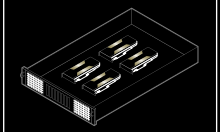
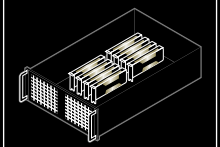
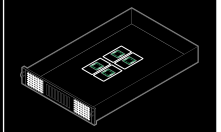
Family of NVIDIA GPU-Accelerated Server Platforms

A broad array of accelerated workloads, from AI training and inference to supercomputing and virtual desktop infrastructure (VDI) applications, demand diverse classes of servers for optimal performance. NVIDIA GPU-Accelerated Server Platforms defines these server classes by recommending the optimal mix of GPUs, CPUs, and interconnects for diverse Training (HGX-T), Inference (HGX-I), and Supercomputing (SCX) applications.

A customer can choose the class that most closely matches their workload to identify the ideal server for that workload. For example, a customer with a life sciences application would look for servers from the SCX-E2 class. A researcher running deep learning training would acquire servers from the HGX-T2 class. The table on the following page discloses the full list of server classes and their corresponding optimal workloads.

To find an NVIDIA partner that provides servers in these classes, please visit www.nvidia.com/qualified-server-catalog

GPU-Accelerated Server Platforms

	HGX-T1	HGX-T2	HGX-I1	HGX-I2	SCX-E1	SCX-E2	SCX-E3	SCX-E4
								
GPU	8x V100 NVIDIA NVLINK™	16x V100 NVLINK + NVIDIA NVSwitch™	2x P4 PCIE	8x P4 PCIE	2x V100 PCIE	4x V100 PCIE	8x V100 PCIE	4x V100 NVLINK
CPU	Dual Xeon							
SYSTEM POWER	3,200 W	10,000 W	600 W	1,400 W	1,200 W	1,800 W	3,000 W	2,000 W
APPLICATIONS								
AI Training	✓	✓						
AI Inference	✓	✓	✓	✓				
HPC	✓	✓			✓	✓	✓	✓
IVA			✓	✓				
VDI/RWS				✓	✓	✓		
Rendering					✓	✓		
KEY BENEFIT	> Optimal for deep learning training and batch inference	> Ultimate deep learning training performance for the largest AI models	> Scale-out deployment with minimal space > Excellent for low-latency speech and language inference	> Excellent for low-latency video and image inference	> Entry-level HPC	> Optimal for life sciences: VASP, AMBER, and GROMACS	> Optimal for seismic, computational fluid dynamics, defense, and data analytics	> Most versatile for supercomputing sites running HPC and AI

HGX-T: AI Training
HGX-I: AI Inference
SCX: Supercomputing and HPC

IVA: Intelligent Video Analytics
VDI: Virtual Desktop Infrastructure
RWS: Remote Workstation

NVIDIA