
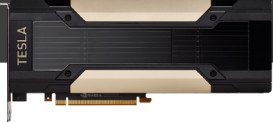




Choose the Right NVIDIA Data Center Product for You.

NVIDIA Tesla V100 with NVIDIA NVLink	NVIDIA Tesla V100 PCIe	NVIDIA Tesla P4	NVIDIA Tesla P40	NVIDIA Tesla P6
				
DESIGNED FOR Deep Learning	DESIGNED FOR HPC and Deep Learning	DESIGNED FOR Deep Learning Inference and Video Transcoding	DESIGNED FOR GPU Virtualization - Graphics and Compute	DESIGNED FOR GPU Virtualization - Graphics and Compute
Up to 3X faster time-to-solution over P100	Up to 5X lower Total Cost of Ownership (TCO) than CPUs for mixed workloads	40X higher energy efficiency than CPUs for inference	Up to 24 virtual GPUs per board	Up to 16 virtual GPUs per board
Ultimate deep learning training performance	Most versatility for mixed HPC workloads 32 GB memory configuration for memory-intensive HPC applications	Low power, low profile optimized for scale out deep learning inference deployment	Industry's highest graphics performance for virtualized environments Run multiple virtualized graphics and compute workloads	Maximum performance for any virtualized workload in a blade-optimized form factor Double the frame buffer of previous generation NVIDIA Maxwell™
KEY FEATURES <ul style="list-style-type: none"> > 125 TeraFLOPS of tensor operations for deep learning > 15.7 TeraFLOPS of single-precision performance > 7.8 TeraFLOPS of double-precision performance > 300 GB/s NVIDIA NVLink™ Interconnect > 900 GB/s memory bandwidth > 32 GB / 16 GB HBM2 memory 	KEY FEATURES <ul style="list-style-type: none"> > 112 TeraFLOPS of tensor operations for deep learning > 14 TeraFLOPS of single-precision performance > 7 TeraFLOPS of double-precision performance > 900 GB/s memory bandwidth > 32 GB / 16 GB HBM2 memory 	KEY FEATURES <ul style="list-style-type: none"> > 22 TeraFLOPS of INT8 inference performance > 5.5 TeraFLOPS of single-precision performance > 1 decode and 2 encode video engines > 50 W/75 W power > Low profile form factor 	KEY FEATURES <ul style="list-style-type: none"> > 24 GB memory > 24 H.264 1080p30 streams > Up to 24 vGPU instances > PCIe 3.0 dual slot form factor > 250 W power 	KEY FEATURES <ul style="list-style-type: none"> > 16 GB memory > 24 H.264 1080p30 streams > Up to 16 vGPU instances > MXM form factor > 90 W (70 W opt) power
RECOMMENDED SERVER CONFIGURATIONS 8-way NVIDIA NVLink hybrid cube mesh (HGX)	RECOMMENDED SERVER CONFIGURATIONS 2-4 GPUs per node	RECOMMENDED SERVER CONFIGURATIONS 1-2 GPUs per node	RECOMMENDED SERVER CONFIGURATIONS 2-4 GPUs per node	RECOMMENDED SERVER CONFIGURATIONS GPUs per node dependent on the blade server