NVIDIA® TESLA®.
ONE PLATFORM. UNLIMITED DATA CENTER ACCELERATION.

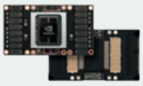
The Exponential Growth of Computing

Accelerating scientific discovery, visualizing big data for insights, and providing smart services to consumers are everyday challenges for researchers and engineers. Solving these challenges takes increasingly complex and precise simulations, the processing of tremendous amounts of data, or training sophisticated deep learning networks. These workloads also require accelerating data centers to meet the growing demand for exponential computing.

NVIDIA Tesla is the world's leading platform for accelerated data centers, deployed by some of the world's largest supercomputing centers and enterprises. It combines GPU accelerators, accelerated computing systems, interconnect technologies, development tools, and applications to enable faster scientific discoveries and big data insights.

At the heart of the NVIDIA Tesla platform are the massively parallel GPU accelerators that provide dramatically higher throughput for compute-intensive workloads—without increasing the power budget and physical footprint of data centers.

Choose the Right NVIDIA® Tesla® Solution for You

PRODUCT	DESIGNED FOR	BENEFITS	KEY FEATURES	RECOMMENDED SERVER CONFIGURATIONS
Tesla P100 PCIe	HPC and Deep Learning	Replace 32 CPU servers with a single P100 server for HPC and deep learning	> 4.7 TeraFLOPS of double-precision performance > 9.3 TeraFLOPS of single-precision performance > 720 GB/s memory bandwidth (540 GB/s option available) > 16 GB of HBM2 memory (12 GB option available)	2-4 GPUs per node
Tesla P100 with NVLink™ 	Deep Learning Training	10X faster deep learning training vs. last-gen GPUs	> 21 TeraFLOPS of half-precision performance > 11 TeraFLOPS of single-precision performance > 160 GB/s NVIDIA NVLink™ > Interconnect > 720 GB/s memory bandwidth > 16 GB of HBM2 memory	4-8 GPUs per node
Tesla P40	Deep Learning Training and Inference	40X faster deep learning inference than a CPU server	> 47 TeraOPS of INT8 inference performance > 12 TeraFLOPS of single-precision performance > 24 GB of GDDR5 Memory > 1 decode and 2 encode video engines	Up to 8 GPUs per node
Tesla P4	Deep Learning Inference and Video Trancoding	40X higher energy efficiency than a CPU for inference	> 22 TeraOPS 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	1-2 GPUs per node