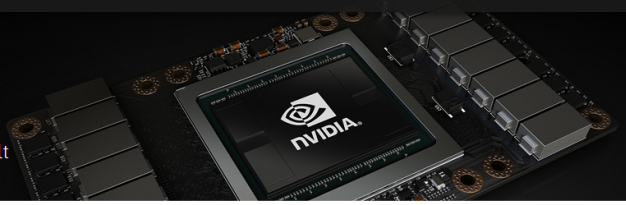
NVIDIA TESLA V100 TENSOR CORE GPU

The Most Advanced Data Center GPU Ever Built



WELCOME TO THE ERA OF AI.

Finding the insights hidden in oceans of data can transform entire industries, from personalized cancer therapy to helping virtual personal assistants converse naturally and predicting the next big hurricane.

NVIDIA® Tesla® V100 Tensor Core is the most advanced data center GPU ever built to accelerate AI, High Performance Computing (HPC), and graphics. It's powered by NVIDIA Volta architecture, comes in 16 and 32GB configurations, and offers the performance of up to 100 CPUs in a single GPU. Data scientists, researchers, and engineers can now spend less time optimizing memory usage and more time designing the next AI breakthrough.

Download V100 Datasheet

3 Reasons Why

V100 Performance Guide

Server Config: Dual Xeon E5-2699 v4 2.6 GHz | 8X NVIDIA® Tesla® P100 or V100 | ResNet-50 Training on MXNet for 90 Epochs with 1.28M ImageNet Dataset.

AI TRAINING

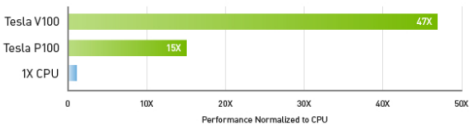
From recognizing speech to training virtual personal assistants and teaching autonomous cars to drive, data scientists are taking on increasingly complex challenges with AI. Solving these kinds of problems requires training deep learning models that are exponentially growing in complexity, in a practical amount of time.

With 640 Tensor Cores, Tesla V100 is the world's first GPU to break the 100 teraFLOPS (TFLOPS) barrier of deep learning performance. The next generation of NVIDIA NVLink™ connects multiple V100 GPUs at up to 300 GB/s to create the world's most powerful computing servers. AI models that would consume weeks of computing resources on previous systems can now be trained in a few days. With this dramatic reduction in training time, a whole new world of problems will now be solvable with AI.

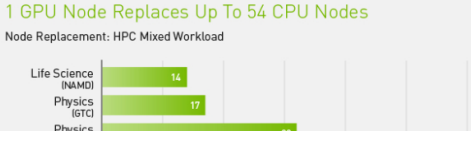
AI INFERENCE

To connect us with the most relevant information, services, and products, hyperscale companies have started to tap into AI. However, keeping up with user demand is a daunting challenge. For example, the world's largest hyperscale company recently estimated that they would need to double their data center capacity if every user spent just three minutes a day using their speech recognition service.

Tesla V100 is engineered to provide maximum performance in existing hyperscale server racks. With AI at its core, Tesla V100 GPU delivers 47X higher inference performance than a CPU server. This giant leap in throughput and efficiency will make the scale-out of AI services practical.

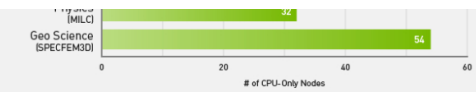


Workload: ResNet-50 | CPU: 1X Xeon E5-2690v4 @ 2.6 GHz | GPU: Add 1X Tesla P100 or V100



HIGH PERFORMANCE COMPUTING (HPC)

HPC is a fundamental pillar of modern science. From predicting weather to discovering drugs to finding new energy sources, researchers use large computing systems to simulate and predict our world. AI extends traditional HPC by allowing



CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ 4x V100 PCIe | CUDA Version: CUDA 9.x| Dataset: NAMD (STMV), GTC (mpi#proc.in), MILC (APEX Medium), SPECFEM3D (four_material_simple_model) | To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

researchers to analyze large volumes of data for rapid insights where simulation alone cannot fully predict the real world.

Tesla V100 is engineered for the convergence of AI and HPC. It offers a platform for HPC systems to excel at both computational science for scientific simulation and data science for finding insights in data. By pairing NVIDIA CUDA® cores and Tensor Cores within a unified architecture, a single server with Tesla V100 GPUs can replace hundreds of commodity CPU-only servers for both traditional HPC and AI workloads. Every researcher and engineer can now afford an AI supercomputer to tackle their most challenging work.

DATA CENTER GPUs

NVIDIA TESLA V100 FOR NVLINK

Ultimate performance for deep learning.

NVIDIA TESLA V100 FOR PCIe

Highest versatility for all workloads.

NVIDIA TESLA V100 SPECIFICATIONS

	Tesla V100 for NVLink	Tesla V100 for PCIe
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 teraFLOPS SINGLE-PRECISION 15.7 teraFLOPS DEEP LEARNING 125 teraFLOPS	DOUBLE-PRECISION 7 teraFLOPS SINGLE-PRECISION 14 teraFLOPS DEEP LEARNING 112 teraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 32/16 GB HBM2 BANDWIDTH 900 GB/s	
POWER Max Consumption	300 WATTS	250 WATTS

TAKE A FREE TEST DRIVE

The World's Fastest GPU Accelerators for HPC and Deep Learning.

GPU TEST DRIVE

WHERE TO BUY

Find an NVIDIA Accelerated Computing Partner through our NVIDIA Partner Network (NPN).

FIND A PARTNER

NVIDIA Tesla P100
NVIDIA Tesla P4/P40
NVIDIA DGX Systems
NVIDIA DGX Station
NVIDIA DGX-1
NVIDIA DGX-2
NVIDIA HGX
NVIDIA GPU Cloud

NVLink/NVSwitch
Tensor Cores
IndeX ParaView Plugin

GPU Apps Catalog
Tesla Product Literature
GPU Test Drive
Where to Buy - DGX
Where to Buy - Tesla
Qualified Server Catalog

SIGN UP FOR DATA CENTER NEWS SUBSCRIBE

Follow Data Center

USA - United States