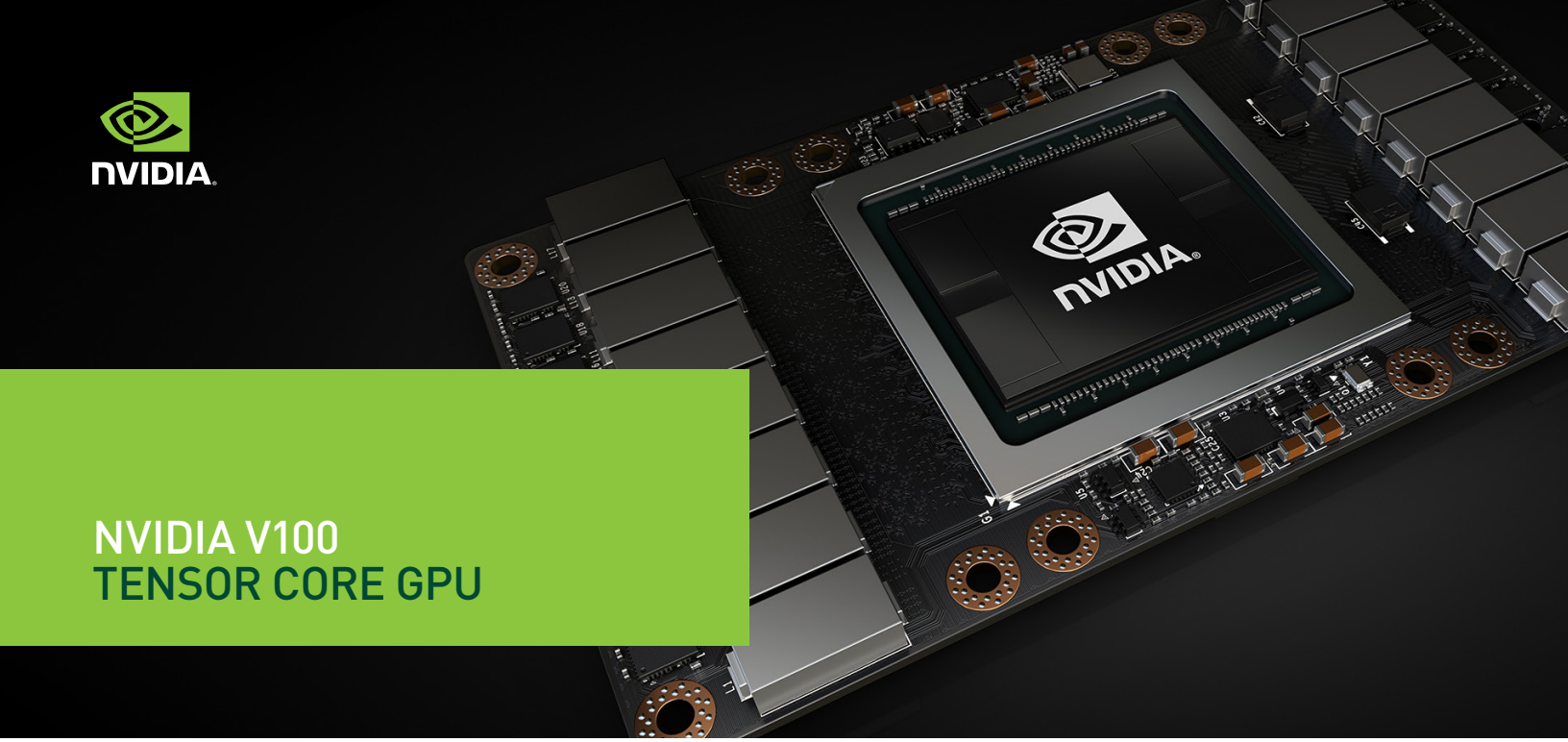
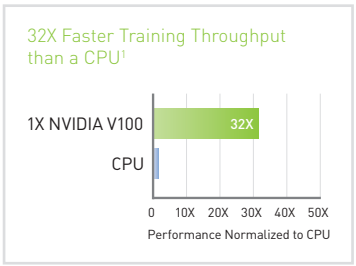
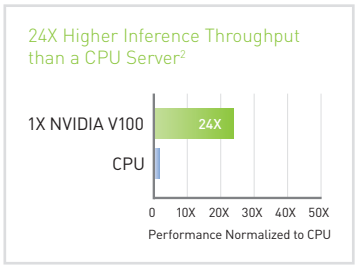


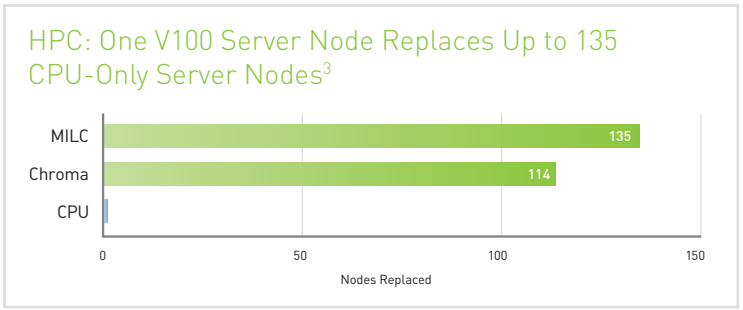
NVIDIA V100
TENSOR CORE GPU

The World's Most Powerful GPU

The NVIDIA® V100 Tensor Core GPU is the world's most powerful accelerator for deep learning, machine learning, high-performance computing (HPC), and graphics. Powered by NVIDIA Volta™, a single V100 Tensor Core GPU offers the performance of nearly 32 CPUs—enabling researchers to tackle challenges that were once unsolvable. The V100 won MLPerf, the first industry-wide AI benchmark, validating itself as the world's most powerful, scalable, and versatile computing platform.



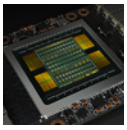




SPECIFICATIONS

	V100 PCIe	V100 SXM2	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	32 GB /16 GB HBM2		32 GB HBM2
Memory Bandwidth	900 GB/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 Comsumption	250 W	300 W	250 W
Thermal Solution	Passive		
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC®		

GROUNDBREAKING INNOVATIONS



VOLTA ARCHITECTURE

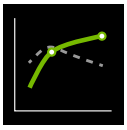
By pairing CUDA cores and Tensor Cores within a unified architecture, a single server with V100 GPUs can replace hundreds of commodity CPU servers for traditional HPC and deep learning.

TENSOR CORE

Equipped with 640 Tensor Cores, V100 delivers 130 teraFLOPS (TFLOPS) of deep learning performance. That's 12X Tensor FLOPS for deep learning training, and 6X Tensor FLOPS for deep learning inference when compared to NVIDIA Pascal™ GPUs.

NEXT-GENERATION NVLINK

NVIDIA NVLink in V100 delivers 2X higher throughput compared to the previous generation. Up to eight V100 accelerators can be interconnected at up to gigabytes per second (GB/sec) to unleash the highest application performance possible on a single server.

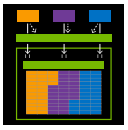


MAXIMUM EFFICIENCY MODE

The new maximum efficiency mode allows data centers to achieve up to 40% higher compute capacity per rack within the existing power budget. In this mode, V100 runs at peak processing efficiency, providing up to 80% of the performance at half the power consumption.

HBM2

With a combination of improved raw bandwidth of 900GB/s and higher DRAM utilization efficiency at 95%, V100 delivers 1.5X higher memory bandwidth over Pascal GPUs as measured on STREAM. V100 is now available in a 32GB configuration that doubles the memory of the standard 16GB offering.



PROGRAMMABILITY

V100 is architected from the ground up to simplify programmability. Its new independent thread scheduling enables finer-grain synchronization and improves GPU utilization by sharing resources among small jobs.

V100 is the flagship product of the NVIDIA data center platform for deep learning, HPC, and graphics. The platform accelerates over 600 HPC applications and every major deep learning framework. It's available everywhere, from desktops to servers to cloud services, delivering both dramatic performance gains and cost-savings opportunities.

EVERY DEEP LEARNING FRAMEWORK

Caffe2

Microsoft Cognitive Toolkit

mxnet

PYTORCH

TensorFlow

theano

600+ GPU-ACCELERATED APPLICATIONS

- AMBER
- ANSYS Fluent
- GAUSSIAN
- GROMACS
- LS-DYNA
- NAMD
- OpenFOAM
- Simulia Abaqus
- VASP
- WRF

To learn more about the NVIDIA V100 Tensor Core GPU, visit **www.nvidia.com/v100**

1 ResNet-50 training, dataset: ImageNet2012, BS=256 | NVIDIA V100 comparison: NVIDIA DGX-2™ server, 1x V100 SXM3-32GB, MXNet 1.5.1, container=19.11-py3, mixed precision, throughput: 1,525 images/sec | Intel comparison: Supermicro SYS-1029GQ-TRT, 1 socket Intel Gold 6240@2GHz/3.9Hz Turbo, Tensorflow 0.18, FP32 (only precision available), throughput: 48 images/sec

2 BERT Base fine-tuning inference, dataset: SQuADv1.1, BS=1, sequence length=128 | NVIDIA V100 comparison: Supermicro SYS-4029GP-TRT, 1x V100-PCIE-16GB, pre-release container, mixed precision, NVIDIA TensorRT™ 6.0, throughput: 557 sentences/sec | Intel comparison: 1 socket Intel Gold 6240@2.6GHz/3.9Hz Turbo, FP32 (only precision available), OpenVINO MKL-DNN v0.18, throughput: 23.5 sentences/sec

3 16x V100-SXM2-32GB in NVIDIA HGX-2™ | Application (dataset): MILC (APEX Medium) and Chroma (szscl21_24_128) | CPU server: dual-socket Intel Xeon Platinum 8280 (Cascade Lake)

NVIDIA