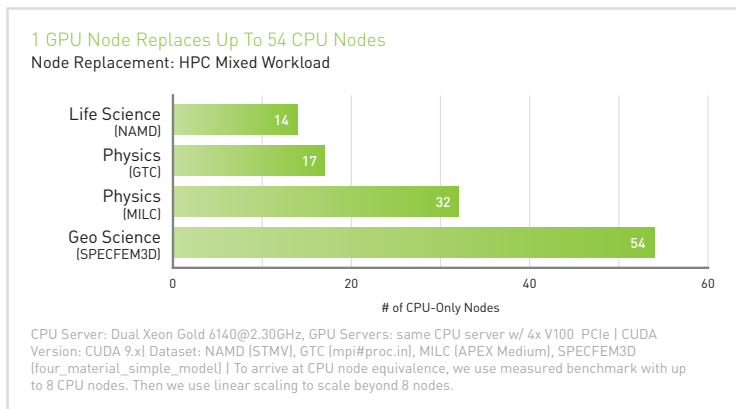
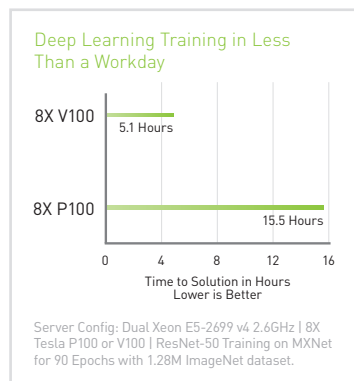
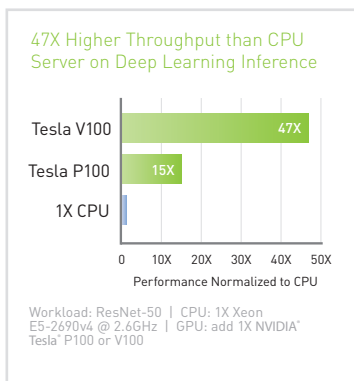


NVIDIA TESLA V100 GPU ACCELERATOR

The Most Advanced Data Center GPU Ever Built.

NVIDIA® Tesla® V100 is the world's most advanced data center GPU ever built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta, the latest GPU architecture, Tesla V100 offers the performance of up to 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once thought impossible.

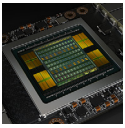


SPECIFICATIONS



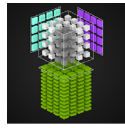
| | Tesla V100 PCIe | Tesla V100 SXM2 |
|------------------------------|---------------------------------------|-----------------|
| GPU Architecture | NVIDIA Volta | |
| NVIDIA Tensor Cores | 640 | |
| NVIDIA CUDA® Cores | 5,120 | |
| Double-Precision Performance | 7 TFLOPS | 7.8 TFLOPS |
| Single-Precision Performance | 14 TFLOPS | 15.7 TFLOPS |
| Tensor Performance | 112 TFLOPS | 125 TFLOPS |
| GPU Memory | 32GB /16GB HBM2 | |
| Memory Bandwidth | 900GB/sec | |
| ECC | Yes | |
| Interconnect Bandwidth | 32GB/sec | 300GB/sec |
| System Interface | PCIe Gen3 | NVIDIA NVLink |
| Form Factor | PCIe Full Height/Length | SXM2 |
| Max Power Consumption | 250 W | 300 W |
| Thermal Solution | Passive | |
| Compute APIs | CUDA, DirectCompute, OpenCL™, OpenACC | |

GROUNDBREAKING INNOVATIONS



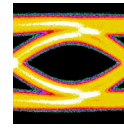
VOLTA ARCHITECTURE

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



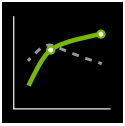
TENSOR CORE

Equipped with 640 Tensor Cores, Tesla V100 delivers 125 teraFLOPS of deep learning performance. That's 12X Tensor FLOPS for DL Training, and 6X Tensor FLOPS for DL Inference when compared to NVIDIA Pascal™ GPUs.



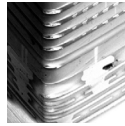
NEXT GENERATION NVLINK

NVIDIA NVLink in Tesla V100 delivers 2X higher throughput compared to the previous generation. Up to eight Tesla V100 accelerators can be interconnected at up to 300GB/s 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, Tesla 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%, Tesla V100 delivers 1.5X higher memory bandwidth over Pascal GPUs as measured on STREAM. Tesla V100 is now available in a 32GB configuration that doubles the memory of the standard 16GB offering.



PROGRAMMABILITY

Tesla 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.

Tesla V100 is the flagship product of Tesla data center computing platform for deep learning, HPC, and graphics. The Tesla platform accelerates over 550 HPC applications and every major deep learning framework. It is available everywhere from desktops to servers to cloud services, delivering both dramatic performance gains and cost savings opportunities.

EVERY DEEP LEARNING FRAMEWORK

550+ GPU-ACCELERATED APPLICATIONS

To learn more about the Tesla V100 visit www.nvidia.com/v100

