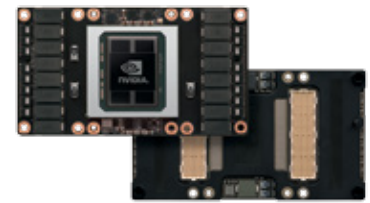# NVIDIA® TESLA® P100
GPU ACCELERATOR

Infinite compute power for the modern data center

Artificial intelligence for self-driving cars. Predicting our climate's future. A new drug to treat cancer. The world's most important challenges require tremendous amounts of computing to become reality. But today's data centers rely on many interconnected commodity compute nodes, limiting the performance needed to drive important HPC and hyperscale workloads.

The NVIDIA Tesla P100 is the most advanced data center accelerator ever built, leveraging the groundbreaking NVIDIA Pascal™ GPU architecture to deliver the world's fastest compute node. It's powered by four innovative technologies with huge jumps in performance for HPC and deep learning workloads.

The Tesla P100 also features NVIDIA NVLink™ technology that enables superior strong-scaling performance for HPC and hyperscale applications. Up to eight Tesla P100 GPUs interconnected in a single node can deliver the performance of racks of commodity CPU servers.

TESLA P100 AND NVLINK DELIVERS UP TO 50X PERFORMANCE BOOST FOR DATA CENTER APPLICATIONS
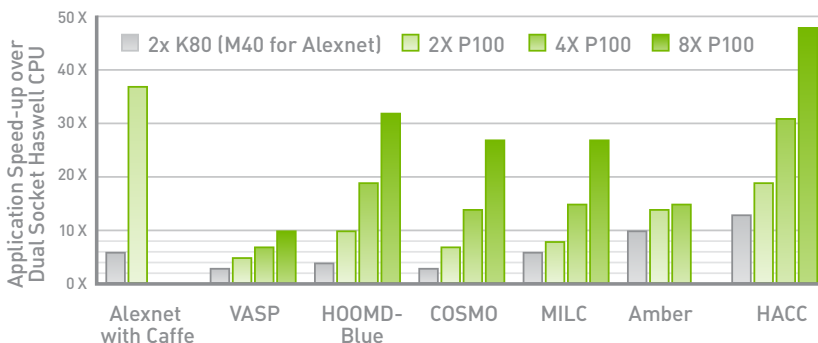

NVIDIA Tesla P100 Performance

Legend: ■ 2x K80 (M40 for Alexnet) ■ 2X P100 ■ 4X P100 ■ 8X P100

Y-axis: Application Speed-up over Dual Socket Haswell CPU (0 X – 50 X)

X-axis: Alexnet with Caffe, VASP, HOOMD-Blue, COSMO, MILC, Amber, HACC

CPU: 16 cores, E5-2698v3 @ 2.30GHz. 256GB System Memory. Tesla K80 GPUs: 2x Dual GPU K80s, Pre-Production Tesla P100

SPECIFICATIONS

GPU Architecture	**NVIDIA Pascal**
NVIDIA CUDA® Cores	**3584**
Double-Precision Performance	**5.3 TeraFLOPS**
Single-Precision Performance	**10.6 TeraFLOPS**
Half-Precision Performance	**21.2 TeraFLOPS**
GPU Memory	**16 GB CoWoS HBM2**
Memory Bandwidth	**732 GB/s**
Interconnect	**NVIDIA NVLink**
Max Power Consumption	**300 W**
ECC	**Native support with no capacity or performance overhead**
Thermal Solution	**Passive**
Form Factor	**SXM2**
Compute APIs	**NVIDIA CUDA, DirectCompute, OpenCL™, OpenACC**

TeraFLOPS measurements with NVIDIA GPU Boost™ technology

EXPERIENCE A GIANT LEAP IN EVERYTHING.

The Tesla P100 is reimagined from silicon to software, crafted with innovation at every level. Each groundbreaking technology delivers a dramatic jump in performance to inspire the creation of the world's fastest compute node.
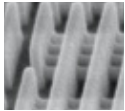
PASCAL ARCHITECTURE
More than 21 TeraFLOPS of FP16, 10 TeraFLOPS of FP32, and 5 TeraFLOPS of FP64 performance powers new possibilities in deep learning and HPC workloads.

Exponential HPC and hyperscale performance
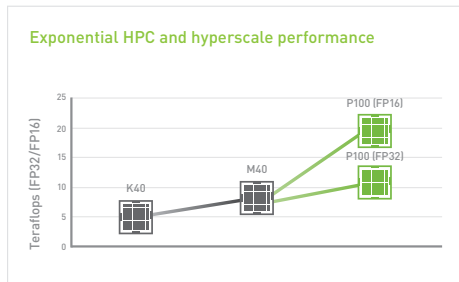
COWOS HBM2
Compute and data are integrated on the same package using Chip-on-Wafer-on-Substrate with HBM2 technology for 3X memory performance over the previous-generation architecture.

3X memory boost
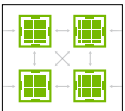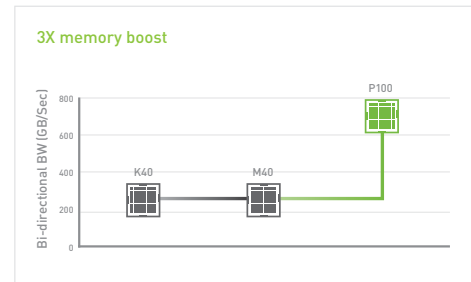
NVLINK INTERCONNECT
This high-speed bidirectional interconnect scales applications across multiple GPUs for 5X higher performance than current best-in-class technology.

5X improvement in interconnect performance
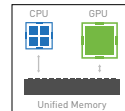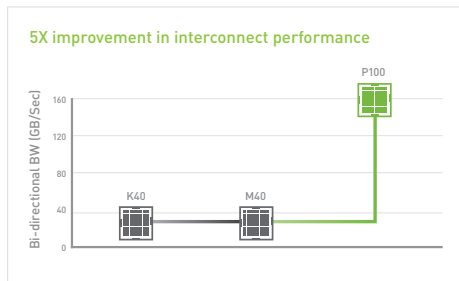
PAGE MIGRATION ENGINE
Simpler programming and computing performance tuning means that applications can now scale beyond the GPU's physical memory size to virtually limitless levels.

Virtually limitless memory scalability
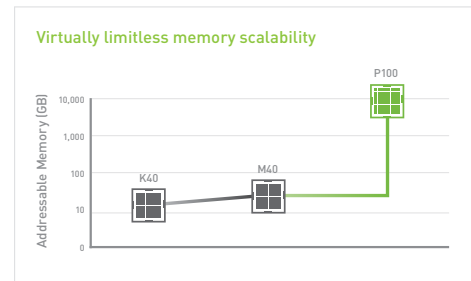
To learn more about the NVIDIA Tesla P100 visit **www.nvidia.com/tesla**