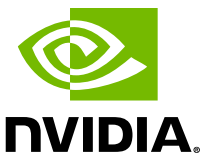
TECHNICAL OVERVIEW

NVIDIA® TESLA® P100:
INFINITE COMPUTE POWER FOR THE MODERN DATA CENTER



Nearly a decade ago, NVIDIA pioneered the use of GPUs to accelerate parallel computing with the introduction of the G80 GPU and the NVIDIA® CUDA® platform.

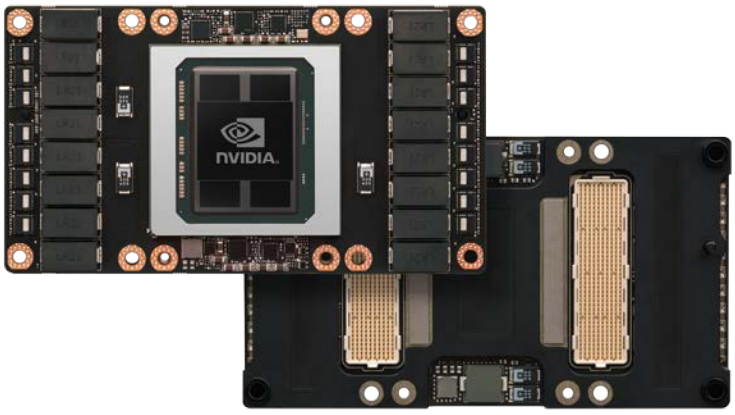
From the desk-side to the data center, supercomputing capabilities were made accessible to thousands of researchers worldwide aspiring to accelerate their most important work.

SUMMARY

The NVIDIA Tesla P100 is the most advanced data center GPU ever created, built on the new NVIDIA Pascal™ architecture. From silicon to software, Tesla P100 is engineered with four key technological breakthroughs to deliver the highest absolute performance. This technical brief describes these breakthroughs in more detail.

- Pascal Architecture
- CoWoS with HBM2
- NVIDIA NVLink™
- Page Migration Engine and Unified Memory

Today, accelerated computing is revolutionizing the data center. The Tesla platform powers some of the world's fastest supercomputers in HPC, enabling groundbreaking Artificial Intelligence (AI) and deep learning systems.



NVIDIA Tesla P100 with the new Pascal architecture

Pascal Architecture: A Quantum Leap for Data Center Applications

The Tesla P100 delivers unprecedented performance for hyperscale and HPC applications. It offers 5.3 TeraFLOPS of peak double-precision performance—3X faster than the previous-generation Tesla K40 GPU. Double-precision (FP64) arithmetic is at the heart of many HPC applications, such as linear algebra, numerical simulation, and quantum chemistry. It also delivers 10.6 TeraFLOPS of peak single-precision performance to accelerate applications in energy exploration and molecular dynamics.

Exponential HPC and hyperscale performance

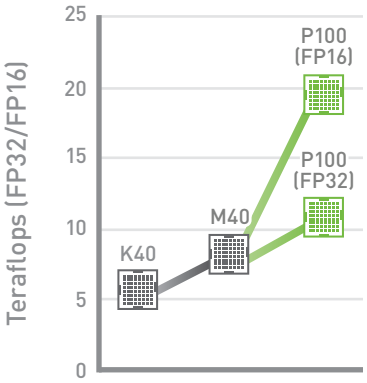


Figure 1: The Tesla P100 significantly exceeds the compute performance of past GPU generations

3X memory boost

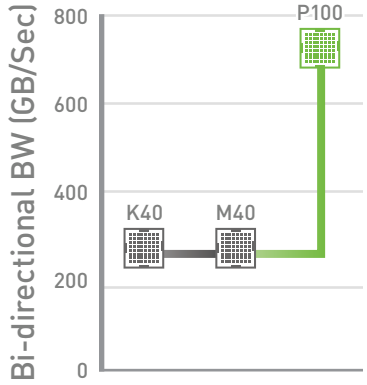


Figure 2: The Tesla P100 with HBM2 significantly exceeds memory bandwidth of past GPU generations

The NVIDIA GPU has been the engine powering the big bang of deep learning. The world's largest hyperscale companies—such as Baidu, Facebook, Google, and Microsoft—are now delivering services with superhuman performance using deep learning. Applications include recognizing objects in images, recognizing speech, and optimizing search results. Using NVIDIA GPUs, deep neural networks can reduce the training time from weeks to days.

Deep learning training workloads typically operate on 32-bit floating point data today. But leading techniques have demonstrated lower-precision FP16 operations that provide higher performance with similar accuracy. The Tesla P100 is the world's first accelerator built for deep learning, and has native hardware ISA support for FP16 arithmetic, delivering over 21 TeraFLOPS of FP16 processing power.

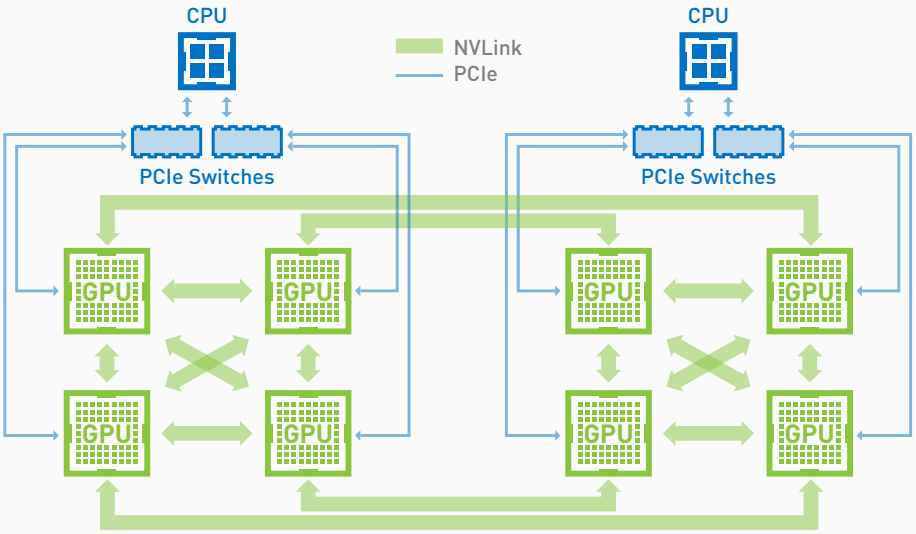
CoWoS with HBM2 Stacked Memory: Unifying Compute and Data into a Single Package for Ultra-Efficient Computing

The biggest inefficiency in computing is data movement. In fact, applications spend more time moving data from memory than processing it. To solve this problem, the Tesla P100 tightly integrates compute and data on the same package by adding Chip on Wafer on Substrate (CoWoS) with HBM2 technology. Using a 4096 bit-wide interface with HBM2, the Tesla P100 delivers 720 GB/s, which is 3X the memory bandwidth of Tesla K40 and M40 GPUs[1].

HBM2 memory has native support for error correcting code (ECC) functionality, while GDDR5 does not. GDDR5 lacks support for internal ECC protection of memory content and is limited to error detection of the GDDR5 bus only. Therefore, Tesla K40 and K80 offered ECC protection by allocating 6.25% of the overall GDDR5 memory capacity for ECC bits. In addition, ECC reduces memory bandwidth. The Tesla P100 with HBM2 has no ECC overhead, both in memory capacity and bandwidth.

Another key benefit of HBM2 memory is its small footprint. Even with 16 GB of memory capacity, the Tesla P100 board is approximately 1/3 the size of Tesla K40 because memory stacks are co-located with the GPU in a single package. Smaller module design enables a new class of highly dense server designs.

[1] Comparison with ECC turned on



NVIDIA NVLink Hybrid Cube Mesh

This figure shows an 8-GPU server design based on NVLink and the Hybrid Cube Mesh topology. Four GPUs are directly connected and four cross-links connect the two quads to each other.

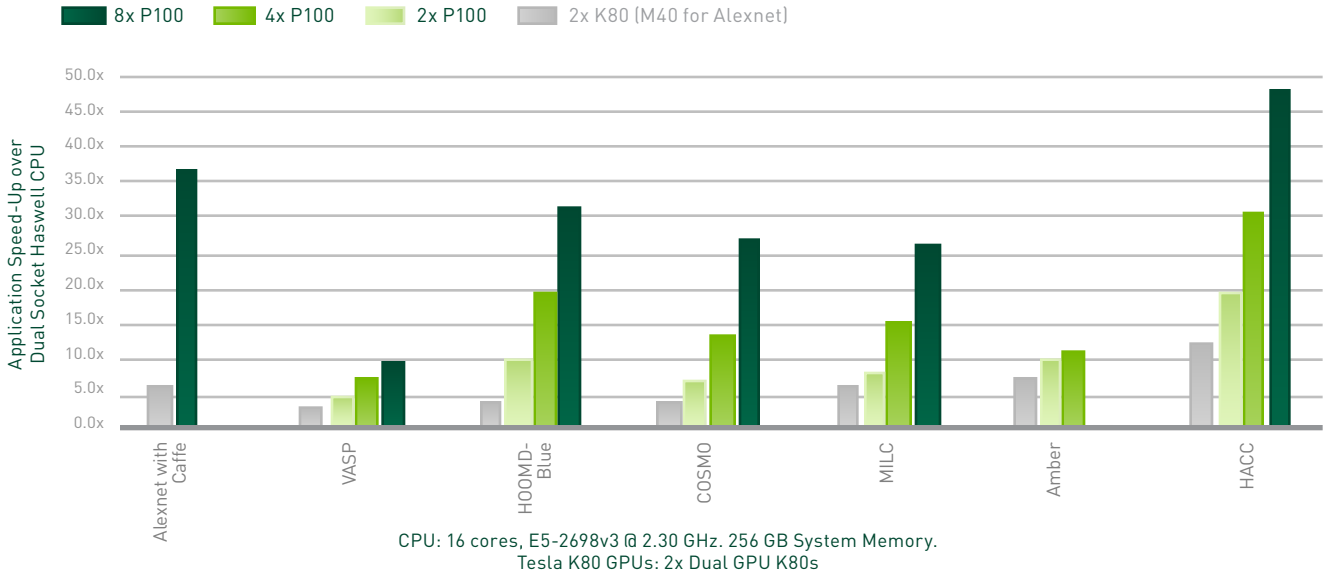
NVIDIA® NVLink™: The World's First GPU-to-GPU and GPU-to-CPU High-Speed Interconnect

As accelerated computing becomes the de-facto standard in the data center, increasing numbers of highly dense GPU server nodes are being deploy. While 4-GPU and 8-GPU system configurations are commonplace to solve bigger problems, interconnect bandwidth between GPUs often becomes a significant bottleneck to application performance. That's because, GPUs in a node communicate through a PCIe switch. Bandwidth is also shared with other devices, such as Ethernet and InfiniBand NICs.

NVLink is the world's first high-speed interconnect for NVIDIA GPUs and solves the interconnect problem. With four NVLink connections per GPU — each delivering with 40 GB/sec bi-directional interconnect bandwidth, Tesla P100 delivers 160 GB/s bidirectional bandwidth in total. This is over 5X the bandwidth of PCI Express Gen3. The PCIe interface is still available for communication with x86 CPU or NIC interfaces.

NVIDIA TESLA P100 PERFORMANCE

The following chart shows the performance for various workloads demonstrating the performance scalability a server can achieve with eight Tesla P100 GPUs connected via NVLink. Applications can scale almost linearly to deliver the highest absolute performance in a node.



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

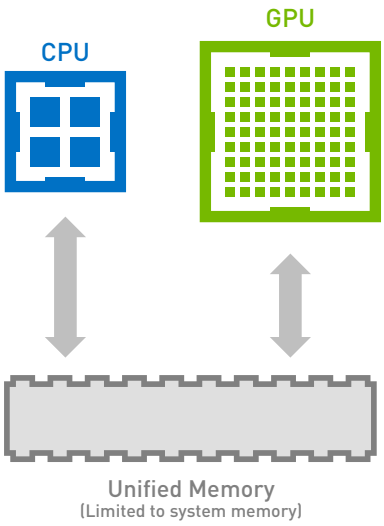


Figure 3: Page Migration Engine feature of the NVIDIA Pascal GPU architecture

Page Migration Engine: Simplified Parallel Programming with Unified Memory

NVIDIA Pascal™ is the first GPU architecture to incorporate virtual memory paging and page faulting support in hardware — called Page Migration Engine. This allows applications with massive datasets to scale beyond the physical memory size of a system. With prior-generation GPUs, developers were limited to executing on datasets that fit into the physical limits of GPU memory size.

Using Page Migration Engine in the Pascal architecture, datasets move seamlessly in the background and on-demand across the physical boundaries of the CPU and GPU memory based on the demand of the application. Applications are permitted to oversubscribe the memory system: they can allocate, access, and share arrays larger than the total physical capacity of the system, enabling out-of-core processing of very large datasets.

Unified Memory, now accelerated by the Page Migration Engine, reduces the GPU computing learning curve. Explicit device memory management becomes a performance optimization, rather than a requirement. Programmers can focus on developing parallel code without getting bogged down in the details of allocating and copying device memory. This makes it easier to learn to program GPUs and bring new workloads into the domain of accelerated computing.

The new NVIDIA Tesla P100 accelerator, built on the Pascal architecture, combines breakthrough technologies to enable science and deep learning workloads that demand unlimited computing resources. Incorporating innovations in architectural efficiency, memory bandwidth, capacity, connectivity, and power efficiency, the NVIDIA Tesla P100 delivers the highest absolute performance for next-generation HPC and AI systems.

To learn more about NVIDIA Tesla visit **www.nvidia.com/tesla**

JOIN US ONLINE

blogs.nvidia.com

@GPUComputing

linkedin.com/company/nvidia

Google.com/+NVIDIA

NVIDIA