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**

<table>
<thead>
<tr>
<th></th>
<th>V100 PCIe</th>
<th>V100 SXM2</th>
<th>V100S PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Architecture</td>
<td>NVIDIA Volta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVIDIA Tensor Cores</td>
<td>640</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVIDIA CUDA® Cores</td>
<td>5,120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double-Precision Performance</td>
<td>7 TFLOPS</td>
<td>7.8 TFLOPS</td>
<td>8.2 TFLOPS</td>
</tr>
<tr>
<td>Single-Precision Performance</td>
<td>14 TFLOPS</td>
<td>15.7 TFLOPS</td>
<td>16.4 TFLOPS</td>
</tr>
<tr>
<td>Tensor Performance</td>
<td>112 TFLOPS</td>
<td>125 TFLOPS</td>
<td>130 TFLOPS</td>
</tr>
<tr>
<td>GPU Memory</td>
<td>32 GB /16 GB HBM2</td>
<td>32 GB HBM2</td>
<td></td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>900 GB/sec</td>
<td>1134 GB/sec</td>
<td></td>
</tr>
<tr>
<td>ECC</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interconnect Bandwidth</td>
<td>32 GB/sec</td>
<td>300 GB/sec</td>
<td>32 GB/sec</td>
</tr>
<tr>
<td>System Interface</td>
<td>PCIe Gen3</td>
<td>NVIDIA NVLink™</td>
<td>PCIe Gen3</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe Full Height/Length</td>
<td>SXM2</td>
<td>PCIe Full Height/Length</td>
</tr>
<tr>
<td>Max Power Consumption</td>
<td>250 W</td>
<td>300 W</td>
<td>250 W</td>
</tr>
<tr>
<td>Thermal Solution</td>
<td>Passive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compute APIs</td>
<td>CUDA, DirectCompute, OpenCL™, OpenACC®</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HPC: One V100 Server Node Replaces Up to 135 CPU-Only Server Nodes³
**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
- 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](http://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 (szsc212_24_128) | GPU server: dual-socket Intel Xeon Platinum 8280 (Cascade Lake)

© 2019 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Volta, CUDA, NVLink, OpenACC, TensorRT, DGX, HGX, and Pascal are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc. All other trademarks and copyrights are the property of their respective owners. DEC19