# TESLA V100 PERFORMANCE GUIDE

Life Sciences Applications

NVIDIA

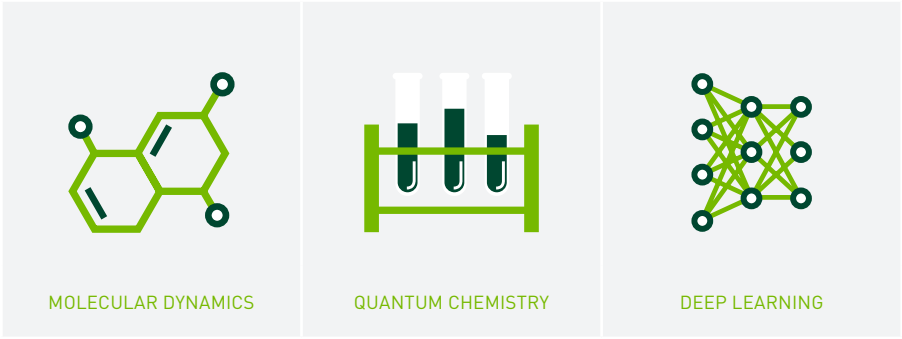
NOVEMBER 2017

# TESLA V100 PERFORMANCE GUIDE

Modern high performance computing (HPC) data centers are key to solving some of the world's most important scientific and engineering challenges. NVIDIA® Tesla® accelerated computing platform powers these modern data centers with the industry-leading applications to accelerate HPC and AI workloads. The intersection of AI and HPC is extending the reach of science and accelerating the pace of scientific innovation like never before. The Tesla V100 GPU is the engine of the modern data center, delivering breakthrough performance with fewer servers resulting in faster insights and dramatically lower costs. Improved performance and time-to-solution can also have significant favorable impacts on revenue and productivity.

Every HPC data center can benefit from the Tesla platform. Over 500 HPC applications in a broad range of domains are optimized for GPUs, including all 15 of the top 15 HPC applications and every major deep learning framework.

## RESEARCH DOMAINS WITH GPU-ACCELERATED APPLICATIONS INCLUDE:

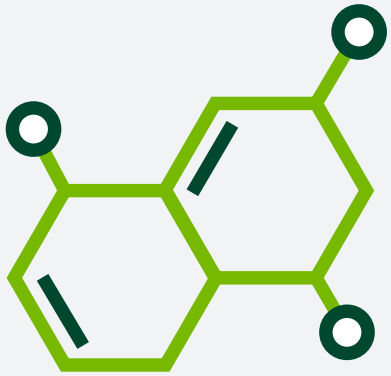



Over 500 HPC applications and all deep learning frameworks are GPU-accelerated.

- To get the latest catalog of GPU-accelerated applications visit: **www.nvidia.com/teslaapps**
- To get up and running fast on GPUs with a simple set of instructions for a wide range of accelerated applications visit: **www.nvidia.com/gpu-ready-apps**

TESLA V100 PERFORMANCE GUIDE

# MOLECULAR DYNAMICS



Molecular Dynamics (MD) represents a large share of the workload in an HPC data center. 100% of the top MD applications are GPU-accelerated, enabling scientists to run simulations they couldn't perform before with traditional CPU-only versions of these applications. When running MD applications, a data center with Tesla V100 GPUs can save up to 80% in server and infrastructure acquisition costs.

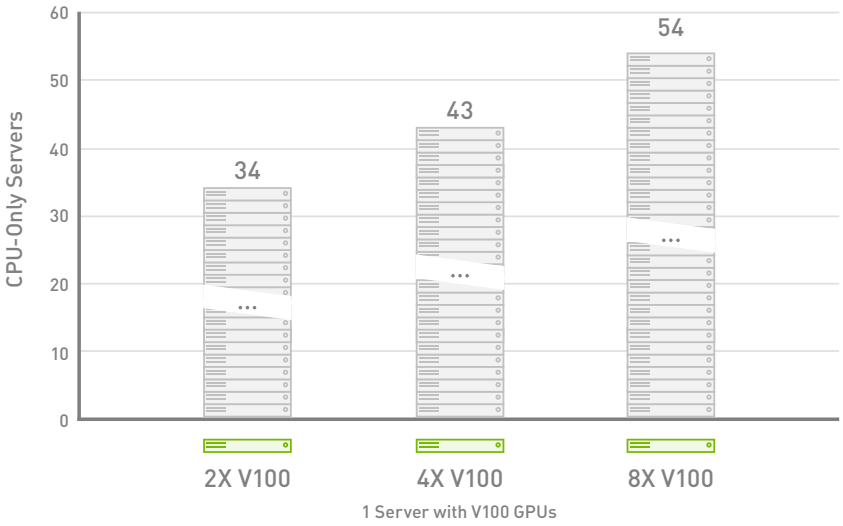
## KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR MD

- Servers with V100 replace up to 54 CPU servers for applications such as HOOMD-Blue and Amber
- 100% of the top MD applications are GPU-accelerated
- Key math libraries like FFT and BLAS
- Up to 15.7 TFLOPS per second of single precision performance per GPU
- Up to 900 GB per second of memory bandwidth per GPU

View all related applications at:
**www.nvidia.com/molecular-dynamics-apps**

## HOOMD-Blue Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers




CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 9.0.145 | Dataset: Microsphere | To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

**HOOMD-BLUE**
Particle dynamics package is written from the ground up for GPUs

**VERSION**
2.1.6

**ACCELERATED FEATURES**
CPU & GPU versions available

**SCALABILITY**
Multi-GPU and Multi-Node

**MORE INFORMATION**
http://codeblue.umich.edu/hoomd-blue/index.html

## AMBER Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers


CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: PME-Cellulose_NVE | To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

**AMBER**
Suite of programs to simulate molecular dynamics on biomolecule

**VERSION**
16.8

**ACCELERATED FEATURES**
PMEMD Explicit Solvent & GB; Explicit
& Implicit Solvent, REMD, aMD

**SCALABILITY**
Multi-GPU and Single-Node

**MORE INFORMATION**
http://ambermd.org/gpus

TESLA V100 PERFORMANCE GUIDE

# QUANTUM CHEMISTRY

Quantum chemistry (QC) simulations are key to the discovery of new drugs and materials and consume a large part of the HPC data center's workload. 60% of the top QC applications are accelerated with GPUs today. When running QC applications, a data center's workload with Tesla V100 GPUs can save over 30% in server and infrastructure acquisition costs.

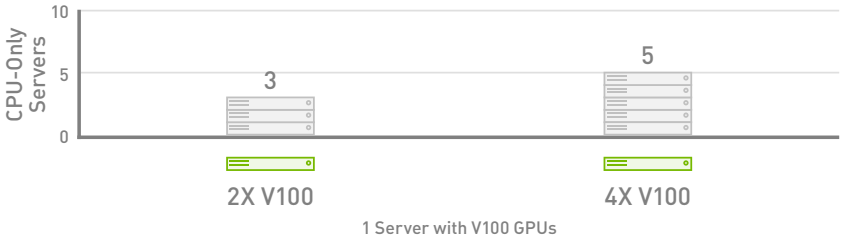
## KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR QC

- Servers with V100 replace up to 5 CPU servers for applications such as VASP
- 60% of the top QC applications are GPU-accelerated
- Key math libraries like FFT and BLAS
- Up to 7.8 TFLOPS per second of double precision performance per GPU
- Up to 16 GB of memory capacity for large datasets

View all related applications at:
**www.nvidia.com/quantum-chemistry-apps**

## VASP Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers




1 Server with V100 GPUs

CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: Si-Huge | To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

**VASP**
Package for performing ab-initio quantum-mechanical molecular dynamics (MD) simulations

**VERSION**
5.4.4

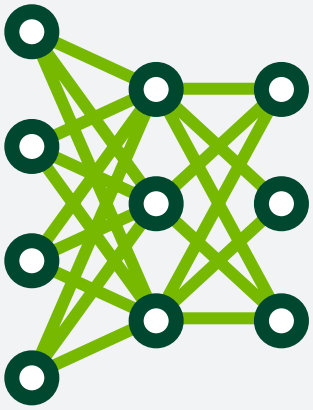
**ACCELERATED FEATURES**
RMM-DIIS, Blocked Davidson, K-points, and exact-exchange

**SCALABILITY**
Multi-GPU and Multi-Node

**MORE INFORMATION**
www.nvidia.com/vasp

TESLA V100 PERFORMANCE GUIDE

# DEEP LEARNING



Deep Learning is solving important scientific, enterprise, and consumer problems that seemed beyond our reach just a few years back. Every major deep learning framework is optimized for NVIDIA GPUs, enabling data scientists and researchers to leverage artificial intelligence for their work. When running deep learning training and inference frameworks, a data center with Tesla V100 GPUs can save up to 85% in server and infrastructure acquisition costs.

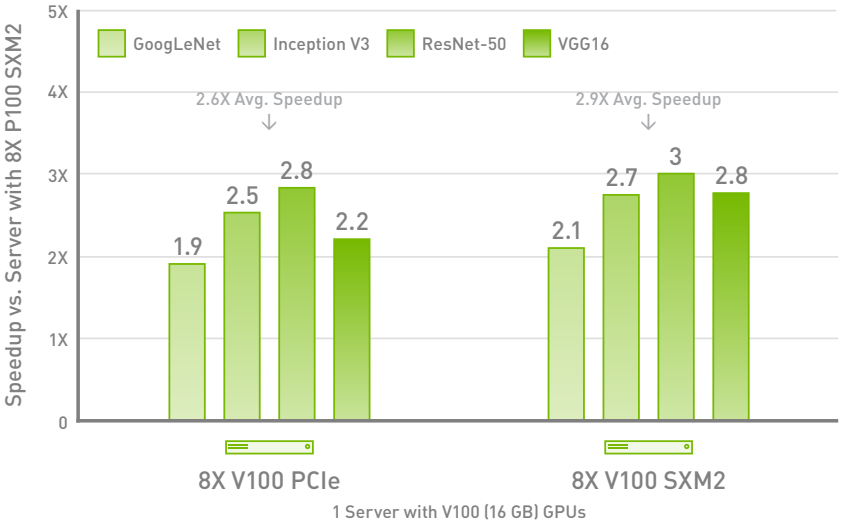
## KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR DEEP LEARNING TRAINING

- Caffe, TensorFlow, and CNTK are up to 3x faster with Tesla V100 compared to P100
- 100% of the top deep learning frameworks are GPU-accelerated
- Up to 125 TFLOPS of TensorFlow operations
- Up to 16 GB of memory capacity with up to 900 GB/s memory bandwidth

View all related applications at:
**www.nvidia.com/deep-learning-apps**

## Caffe Deep Learning Framework

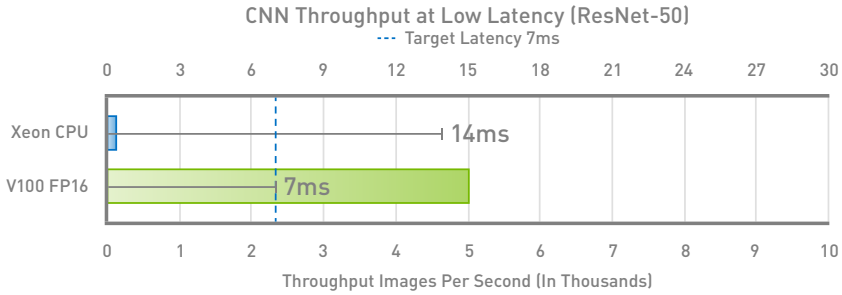
Training on 8X V100 GPU Server vs 8X P100 GPU Server




CPU Server: Dual Xeon E5-2698 v4 @ 3.6GHz, GPU servers as shown | Ubuntu 14.04.5 | CUDA Version: CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | Driver 384.66 | Data set: ImageNet | Batch sizes: GoogleNet 192, Inception V3 96, ResNet-50 64 for P100 SXM2 and 128 for Tesla P100, VGG16 96

## LOW-LATENCY CNN INFERENCE PERFORMANCE

Massive Throughput and Amazing Efficiency at Low Latency




System configs: Single-socket Xeon E2690 v4 @ 3.5GHz, and a single NVIDIA® Tesla® V100, GPU running TensorRT 3 RC vs. Intel DL SDK beta 2 | Ubuntu 14.04.5 | CUDA Version: 7.0.1.13 | CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | Driver 384.66 | Precision: CPU FP32, NVIDIA Tesla V100 FP16

**CAFFE**
A popular, GPU-accelerated Deep Learning framework developed at UC Berkeley

**VERSION**
1.0

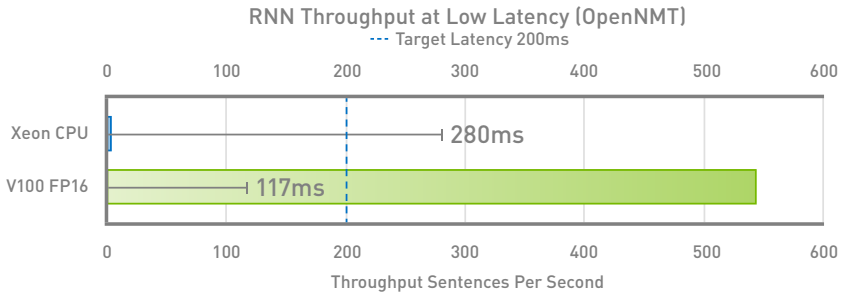
**ACCELERATED FEATURES**
Full framework accelerated

**SCALABILITY**
Multi-GPU

**MORE INFORMATION**
caffe.berkeleyvision.org

## LOW-LATENCY RNN INFERENCE PERFORMANCE

Massive Throughput and Amazing Efficiency at Low Latency




System configs: Single-socket Xeon E2690 v4 @ 3.5GHz, and a single NVIDIA® Tesla® V100, GPU running TensorRT 3 RC vs. Intel DL SDK beta 2 | Ubuntu 14.04.5 | CUDA Version: 7.0.1.13 | CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | Driver 384.66 | Precision: CPU FP32, NVIDIA Tesla V100 FP16

# TESLA V100 PRODUCT SPECIFICATIONS


| | NVIDIA Tesla V100 for PCIe-Based Servers | NVIDIA Tesla V100 for NVLink-Optimized Servers |
|---|---|---|
| Double-Precision Performance | up to 7 TFLOPS | up to 7.8 TFLOPS |
| Single-Precision Performance | up to 14 TFLOPS | up to 15.7 TFLOPS |
| Deep Learning | up to 112 TFLOPS | up to 125 TFLOPS |
| NVIDIA NVLink™ Interconnect Bandwidth | - | 300 GB/s |
| PCIe x 16 Interconnect Bandwidth | 32 GB/s | 32 GB/s |
| CoWoS HBM2 Stacked Memory Capacity | 16 GB | 16 GB |
| CoWoS HBM2 Stacked Memory Bandwidth | 900 GB/s | 900 GB/s |


ABC Product (Model) Name

## ABC PRODUCT (MODEL) NAME

Partner product description paragraph. One hundred words maximum. Xeris exeria nobis exerferis dolupt.

- Spec 1: Some Data
- Spec 2: Some Data
- Spec 3: Some Data
- Spec 4: Some Data

ABC COMPANY, INC.

## COMPANY NAME

Optional company brief description paragraph. No more than fifty words. Explia consequam il ilis escipiducium remd. Xeris exeria nobis exerferis dolupt, qui quo volores dolori blab iliquate il il excerum excesequi dolori manaianisi mintes.

www.abccompany.com | +1 (123) 555-678 | jdoe@abccompany.com

Assumptions and Disclaimers

The percentage of top applications that are GPU-accelerated is from top 50 app list in the i360 report: HPC Support for GPU Computing.
Calculation of throughput and cost savings assumes a workload profile where applications benchmarked in the domain take equal compute cycles:
**http://www.intersect360.com/industry/reports.php?id=131**
The number of CPU nodes required to match single GPU node is calculated using lab performance results of the GPU node application speed-up and the Multi-CPU node scaling performance. For example, the Molecular Dynamics application HOOMD-Blue has a GPU Node application speed-up of 37.9X. When scaling CPU nodes to an 8 node cluster, the total system output is 7.1X. So the scaling factor is 8 divided by 7.1 (or 1.13). To calculate the number of CPU nodes required to match the performance of a single GPU node, you multiply 37.9 (GPU Node application speed-up) by 1.13 (CPU node scaling factor) which gives you 43 nodes.



NVIDIA