TESLA V100 PERFORMANCE GUIDE

Deep Learning and HPC 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 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:

DEEP LEARNING

MOLECULAR DYNAMICS

QUANTUM CHEMISTRY

PHYSICS

GEOSCIENCE

ENGINEERING

HPC BENCHMARKS

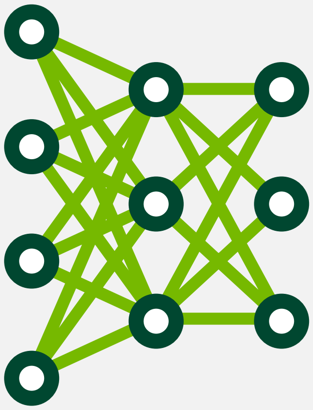
Over 500 HPC applications and all deep learning frameworks are GPU-accelerated.

> 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

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

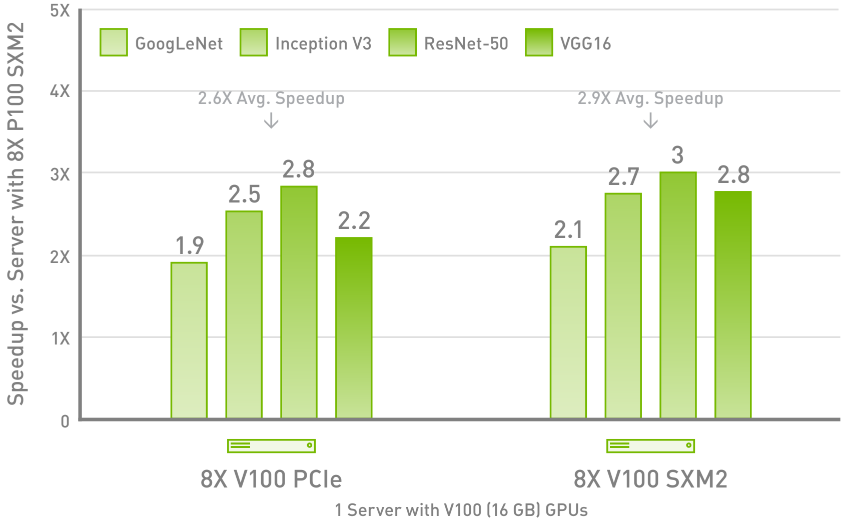
> Up to 125 TFLOPS of TensorFlow operations

> 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

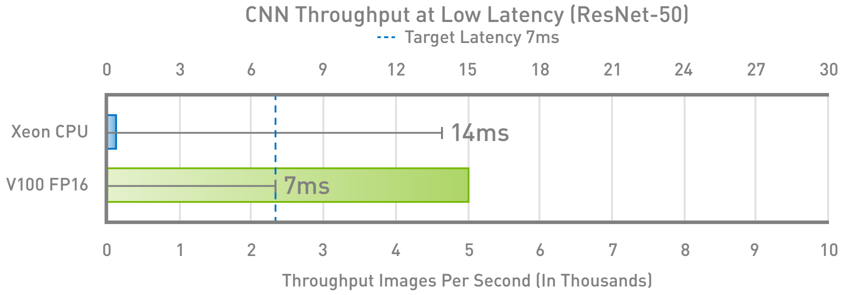
Training on 8X V100 GPU Server vs 8X P100 GPU Server



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

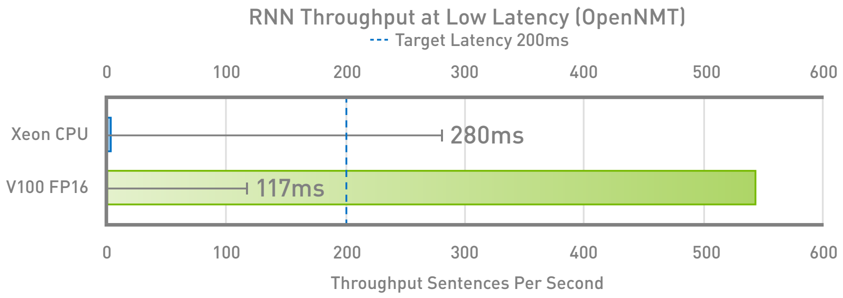
ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU

MORE INFORMATION
caffe.berkeleyvision.org

LOW-LATENCY RNN INFERENCE PERFORMANCE

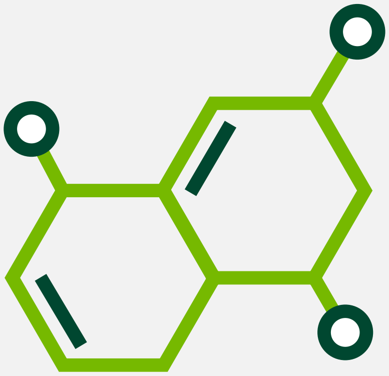
Massive Throughput and Amazing Efficiency at Low Latency



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 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.

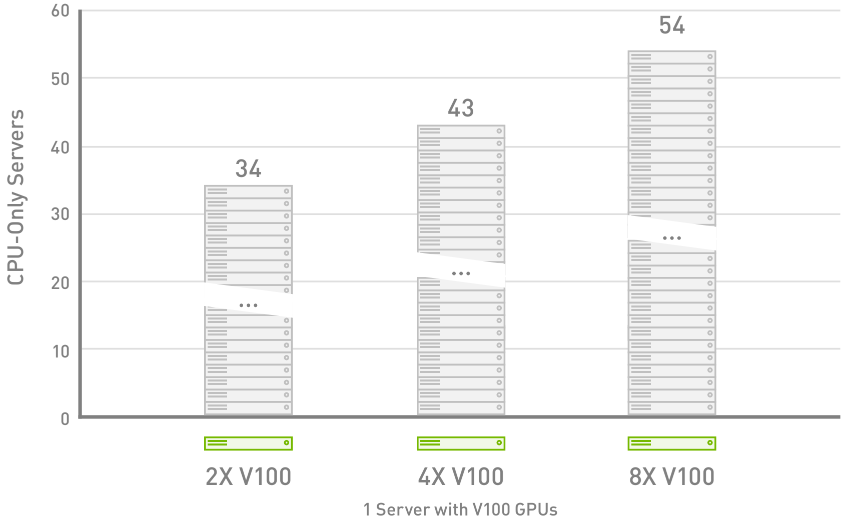
KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR MD

- 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

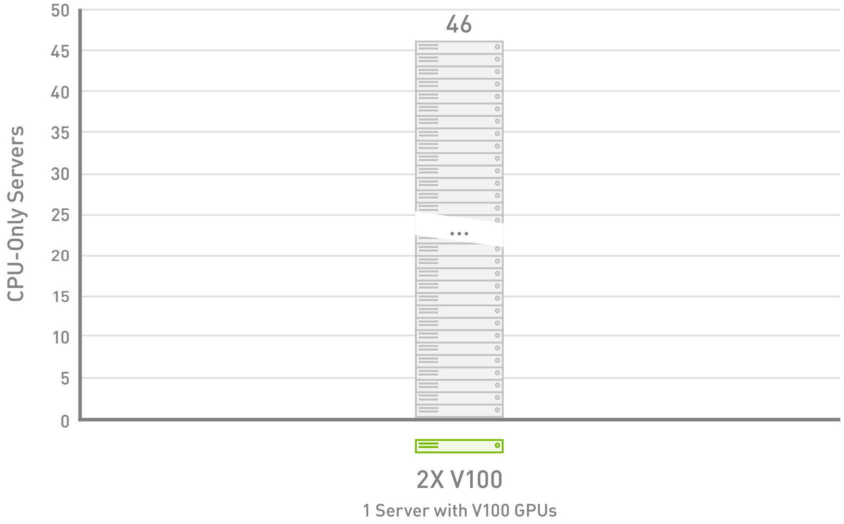
ACCELERATED FEATURES
CPU & GPU versions available

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.

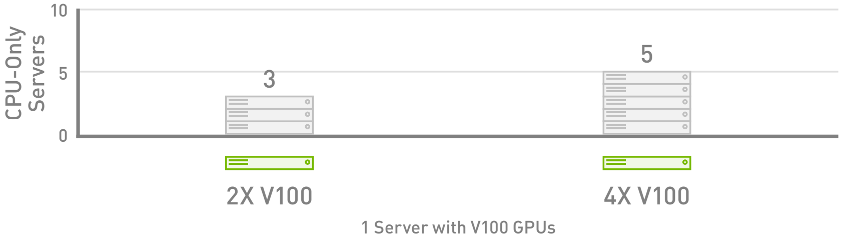
KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR QC

- 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



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

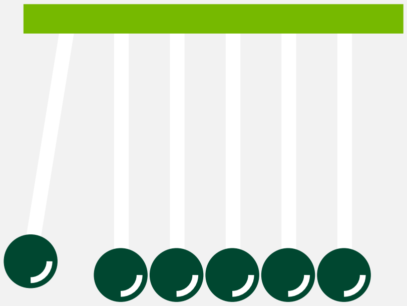
ACCELERATED FEATURES
RMM-DIIS, Blocked Davidson, K-points, and exact-exchange

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/vasp

TESLA V100 PERFORMANCE GUIDE

PHYSICS



From fusion energy to high energy particles, physics simulations span a wide range of applications in the HPC data center. Many of the top physics applications are GPU-accelerated, enabling insights previously not possible. A data center with Tesla V100 GPUs can save up to 75% in server acquisition cost when running GPU-accelerated physics applications.

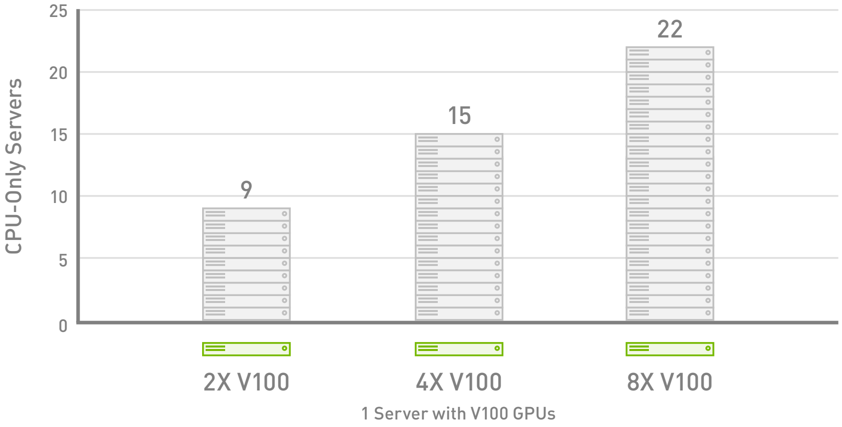
KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR PHYSICS

- Servers with V100 replace up to 75 CPU servers for applications such as GTC-P, QUDA, and MILC
- Most of the top physics applications are GPU-accelerated
- Up to 7.8 TFLOPS of double precision floating point performance
- Up to 16 GB of memory capacity with up to 900 GB/s memory bandwidth

View all related applications at:
www.nvidia.com/physics-apps

GTC-P 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: A.txt | 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.

GTC-P
A development code for optimization of plasma physics

VERSION
2017

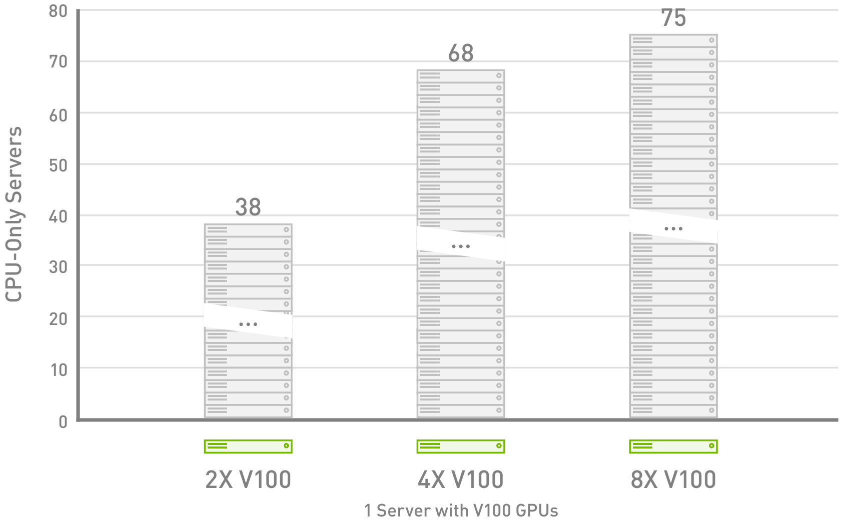
ACCELERATED FEATURES
Push, shift, and collision

SCALABILITY
Multi-GPU

MORE INFORMATION
www.nvidia.com/gtc-p

QUDA 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: Dslash Wilson-Clove; Precision: Single; Gauge Compression/Recon: 12; Problem Size 32x32x32x64 | 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.

QUDA
A library for Lattice Quantum Chromo Dynamics on GPUs

VERSION
2017

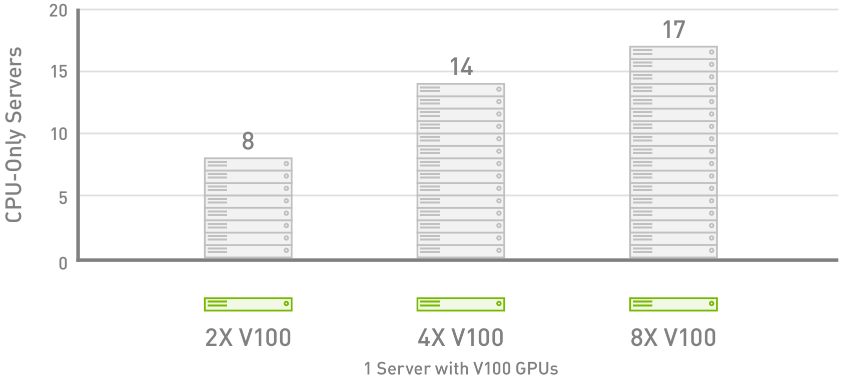
ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/quda

MILC 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: Precision=FP64 | 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.

MILC
Lattice Quantum Chromodynamics (LQCD) codes simulate how elemental particles are formed and bound by the "strong force" to create larger particles like protons and neutrons

VERSION
2017

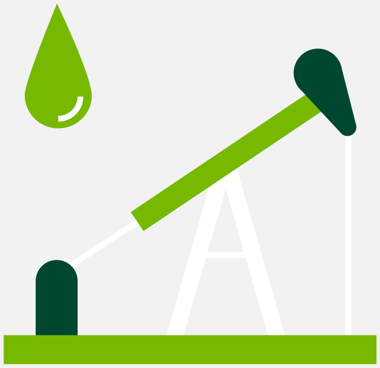
ACCELERATED FEATURES
Staggered fermions, Krylov solvers, Gauge-link fattening

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/milc

TESLA V100 PERFORMANCE GUIDE

GEOSCIENCE



Geoscience simulations are key to the discovery of oil and gas and performing geological modeling. Many of the top geoscience applications are accelerated with GPUs today. When running Geoscience applications, a data center with Tesla V100 GPUs can save up to 70% in server and infrastructure acquisition costs.

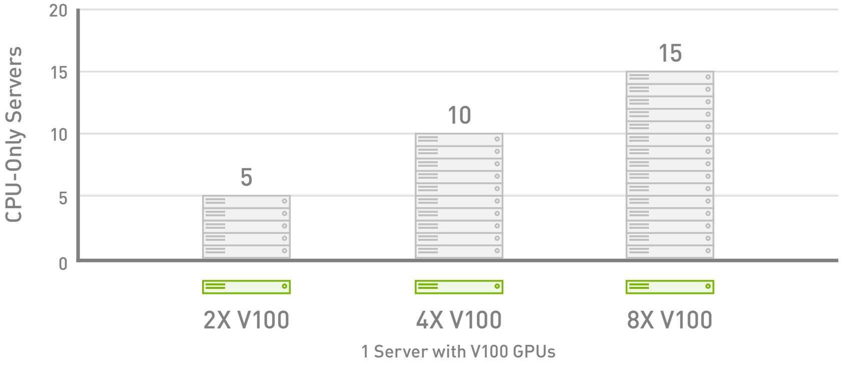
KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR GEOSCIENCE

- Servers with V100 replace up to 82 CPU servers for applications such as RTM and SPECFEM 3D
- Top Oil and Gas applications are GPU-accelerated
- Up to 15.7 TFLOPS of single precision floating point performance
- Up to 16 GB of memory capacity with up to 900 GB/s memory bandwidth

View all related applications at:
www.nvidia.com/oil-and-gas-apps

RTM 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: TTI RX 2pass mgpu | To arrive at CPU node equivalence, we use linear scaling to scale beyond 1 nodes.

RTM

Reverse time migration (RTM) modeling is a critical component in the seismic processing workflow of oil and gas exploration

VERSION

2017

ACCELERATED FEATURES

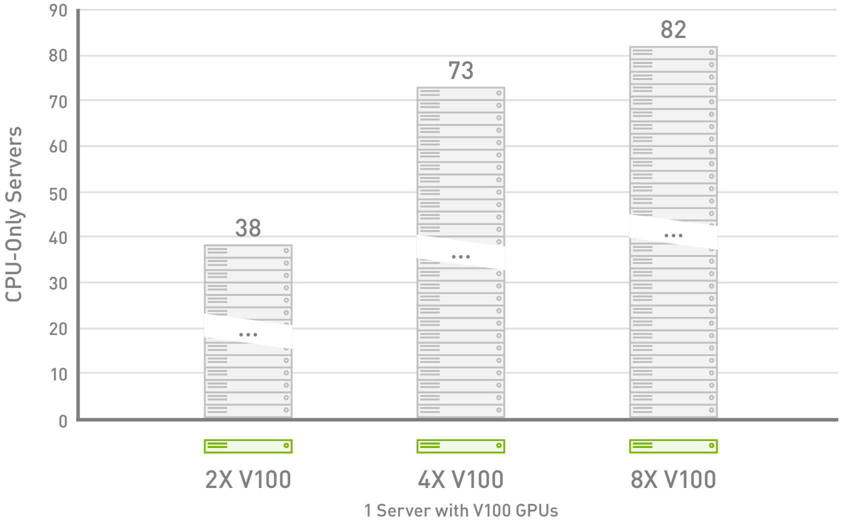
Batch algorithm

SCALABILITY

Multi-GPU and Multi-Node

SPECFEM 3D 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: 288x64, 100 mins | To arrive at CPU node equivalence, we use linear scaling to scale beyond 1 nodes.

SPECFEM 3D

Simulates Seismic wave propagation

VERSION

7.0.0

SCALABILITY

Multi-GPU and Multi-Node

MORE INFORMATION

https://geodynamics.org/cig/software/specfem3d_globe

TESLA V100 PERFORMANCE GUIDE

ENGINEERING

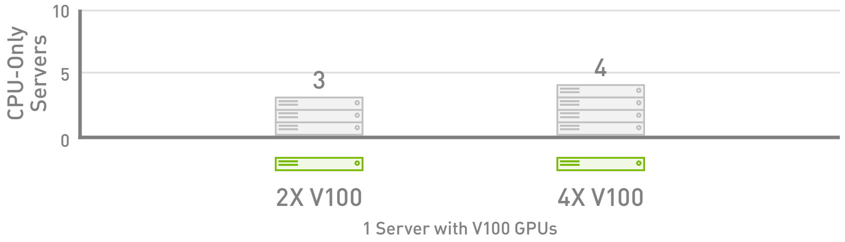
Engineering simulations are key to developing new products across industries by modeling flows, heat transfers, finite element analysis and more. Many of the top Engineering applications are accelerated with GPUs today. When running Engineering applications, a data center with NVIDIA® Tesla® V100 GPUs can save over 20% in server and infrastructure acquisition costs and over 50% in software licensing costs.

KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR ENGINEERING

- Servers with Tesla V100 replace up to 4 CPU servers for applications such as SIMULIA Abaqus and ANSYS FLUENT
- The top engineering applications are GPU-accelerated
- Up to 16 GB of memory capacity
- Up to 900 GB/s memory bandwidth
- Up to 7.8 TFLOPS of double precision floating point

SIMULIA Abaqus Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 7.5 | Dataset: LS-EPP-Combined-WC-Mkl (RR) | 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.

SIMULIA ABAQUS
Simulation tool for analysis of structures

VERSION
2017

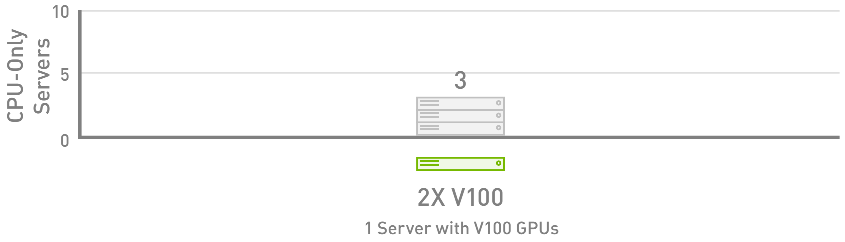
ACCELERATED FEATURES
Direct Sparse Solver
AMS Eigen Solver
Steady-state Dynamics Solver

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/simulia-abaqus

ANSYS Fluent Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 6.0 | Dataset: Water Jacket | 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.

ANSYS FLUENT
General purpose software for the simulation of fluid dynamics

VERSION
18

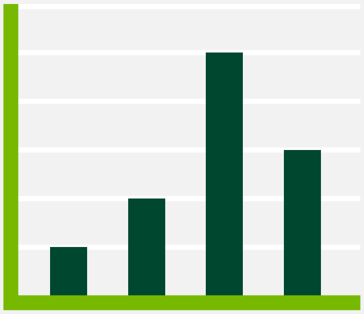
ACCELERATED FEATURES
Pressure-based Coupled Solver and Radiation Heat Transfer

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/ansys-fluent

TESLA V100 PERFORMANCE GUIDE

HPC BENCHMARKS



Benchmarks provide an approximation of how a system will perform at production-scale and help to assess the relative performance of different systems. The top benchmarks have GPU-accelerated versions and can help you understand the benefits of running GPUs in your data center.

KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR BENCHMARKING

> Servers with Tesla V100 replace up to 67 CPU servers for benchmarks such as Cloverleaf, MiniFE, Linpack, and HPCG

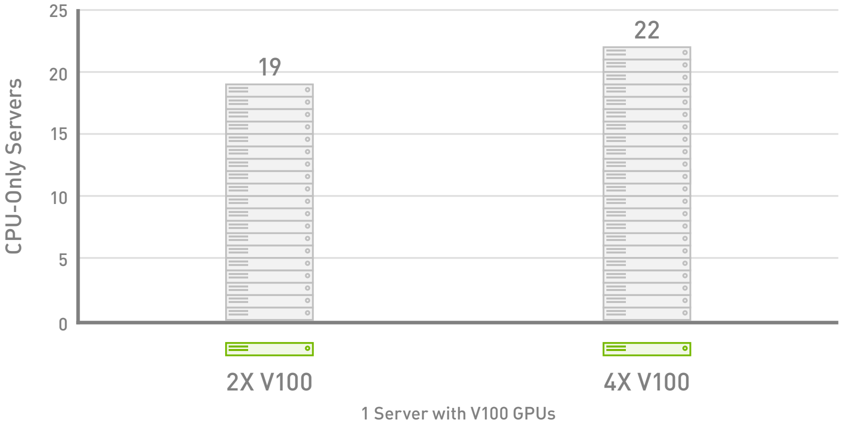
> The top benchmarks are GPU-accelerated

> Up to 7.8 TFLOPS of double precision floating point up to 16 GB of memory capacity

> Up to 900 GB/s memory bandwidth

Cloverleaf Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: bm32 | 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.

CLOVERLEAF
Benchmark – Mini-App
Hydrodynamics

VERSION
1.3

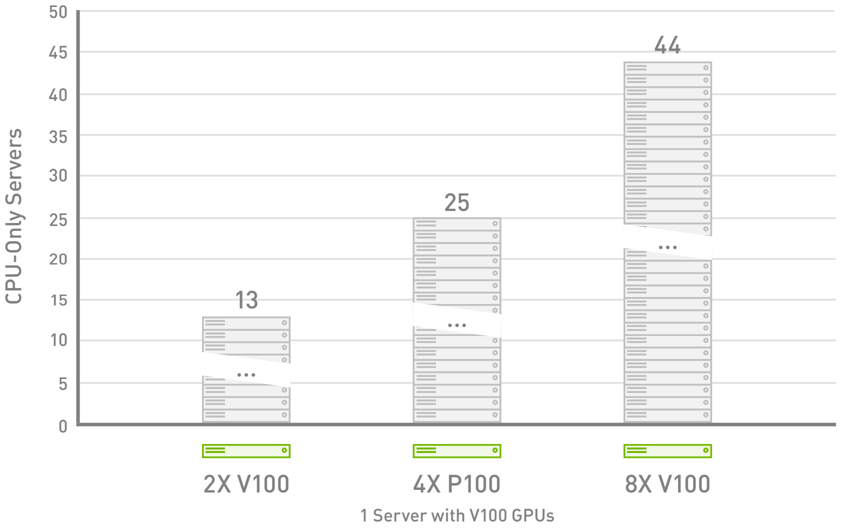
ACCELERATED FEATURES
Lagrangian-Eulerian
explicit hydrodynamics mini-application

SCALABILITY
Multi-Node (MPI)

MORE INFORMATION
http://uk-mac.github.io/CloverLeaf

MiniFE Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Single Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: 350x350x350 | 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.

MINIFE
Benchmark – Mini-App
Finite Element Analysis

VERSION
0.3

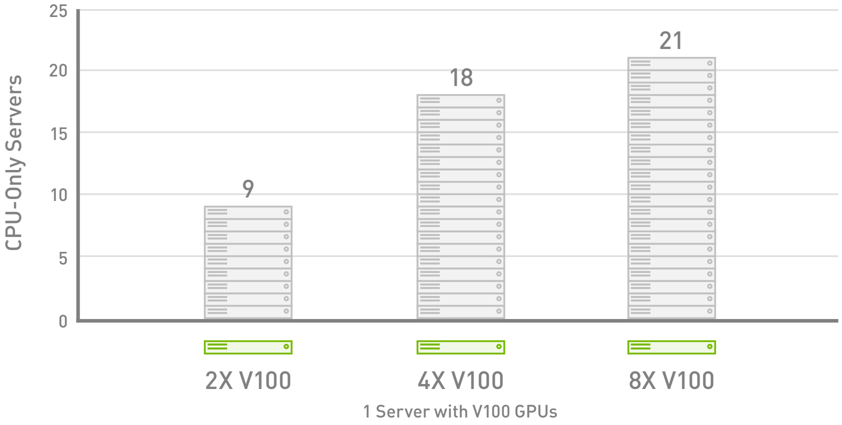
ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU

MORE INFORMATION
https://mantevo.org/about/applications

Linpack Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 9.0.103 | Dataset: HPL.dat | 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.

LINPACK
Benchmark – Measures floating point computing power

VERSION
2.1

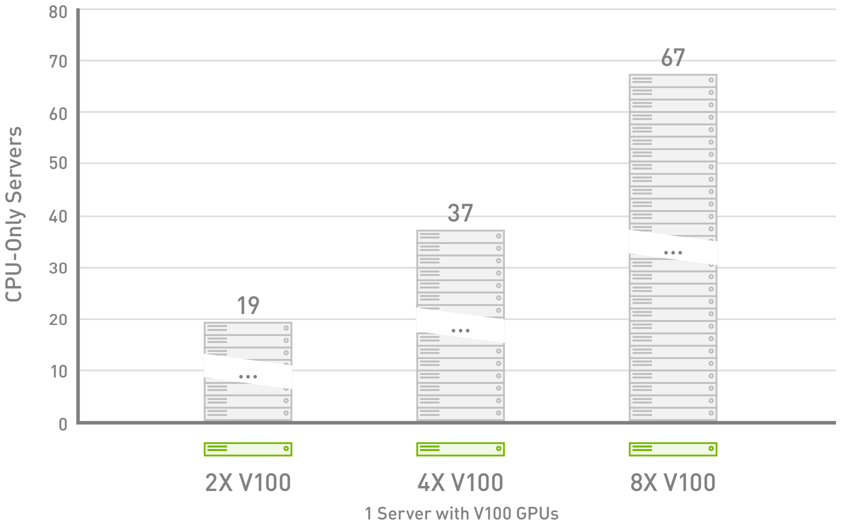
ACCELERATED FEATURES
All

SCALABILITY
Multi-Node and Multi-Node

MORE INFORMATION
www.top500.org/project/linpack

HPCG Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 for PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: 256x256x256 local size | 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.

HPCG
Benchmark – Exercises computational and data access patterns that closely match a broad set of important HPC applications

VERSION
3

ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.hpcg-benchmark.org/index.html

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

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.