

# TESLA V100 PERFORMANCE GUIDE

Deep Learning and HPC Applications

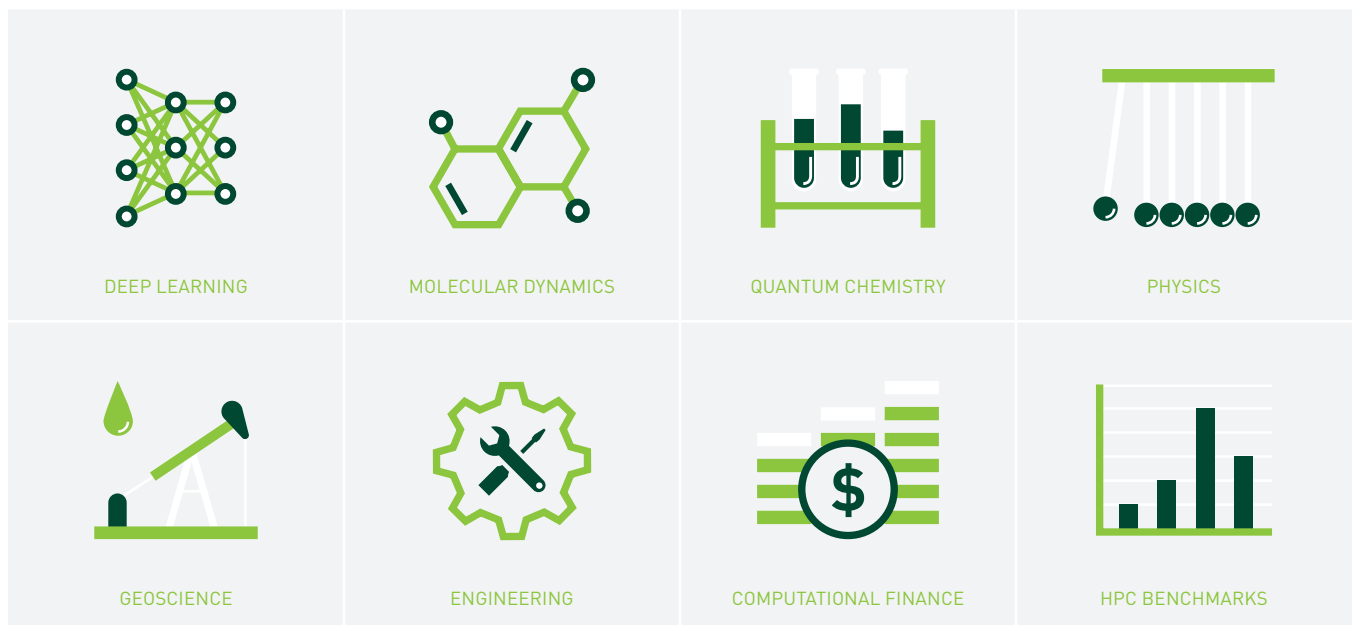


# TESLA V100 PERFORMANCE GUIDE

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, less power consumption, and reduced networking overhead, resulting in total cost savings of 5X-10X. Each GPU-accelerated server provides the performance of dozens of commodity CPU servers, delivering a dramatic boost in application throughput. 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 550 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 550 HPC applications and all deep learning frameworks are GPU-accelerated.

> To get the latest catalog of GPU-accelerated applications visit:

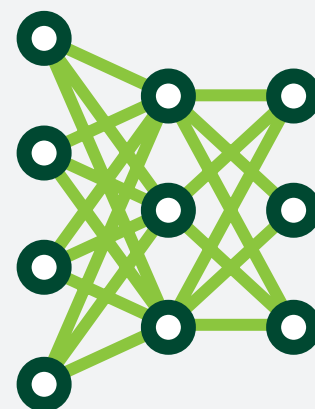
[www.nvidia.com/teslaapps](http://www.nvidia.com/teslaapps)

> 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](http://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 over 90% in server and infrastructure acquisition costs.

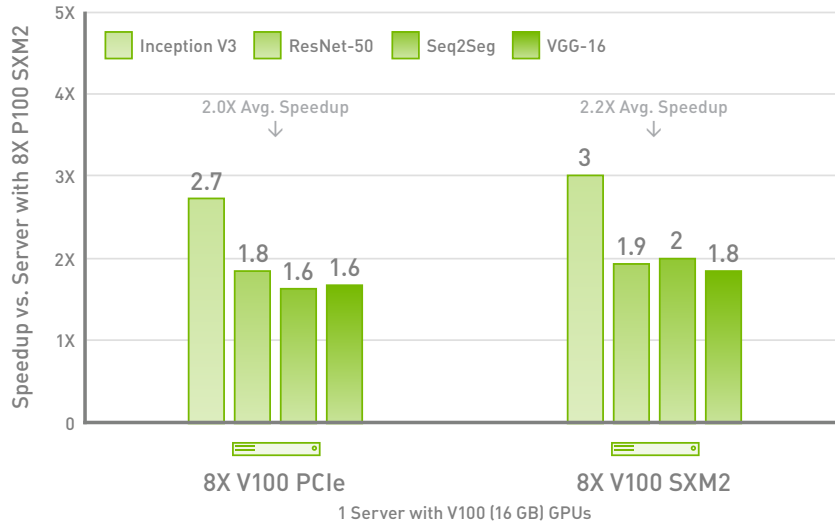
### KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR DEEP LEARNING TRAINING

- > 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 per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s memory bandwidth per GPU

View all related applications at:  
[www.nvidia.com/deep-learning-apps](http://www.nvidia.com/deep-learning-apps)

## Caffe2 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 | NVIDIA CUDA® Version: 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | Driver 384.66 | Data set: ImageNet | Batch sizes: Inception V3 and ResNet-50 64 for P100 SXM2 for P100 SXM2 and 128 for Tesla V100, Seq2Seq 192, VGG16 96

### CAFFE2

A popular, GPU-accelerated Deep Learning framework developed at UC Berkeley and Facebook

### VERSION

1.0

### ACCELERATED FEATURES

Full framework accelerated

### SCALABILITY

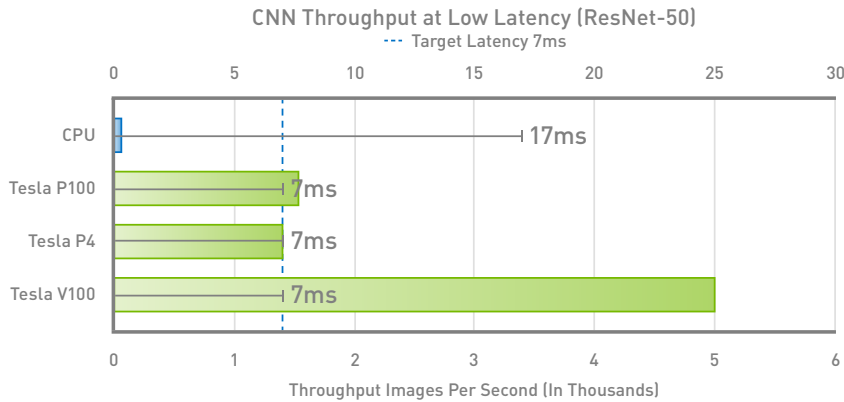
Multi-GPU

### MORE INFORMATION

[research.fb.com/downloads/caffe2](https://research.fb.com/downloads/caffe2)

## NVIDIA TENSORRT 3

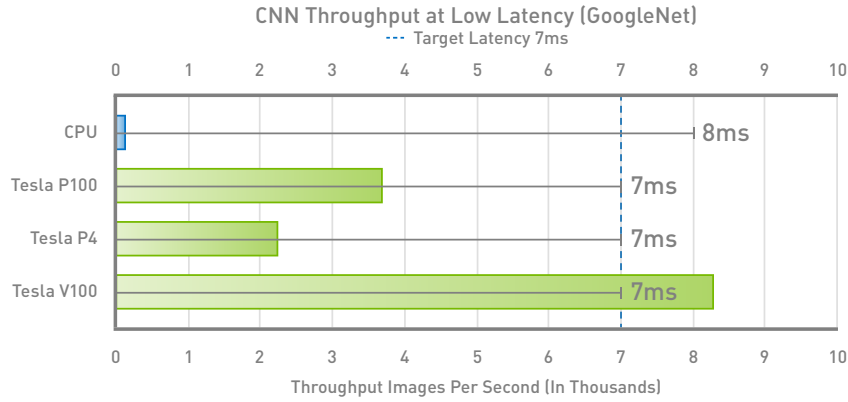
Massive Throughput and Amazing Efficiency at Low Latency



CPU throughput based on measured-inference-throughput performance on Broadwell-based Xeon E2690 v4 CPU and doubled to reflect Intel's stated claim that Xeon Scalable Processor will deliver 2X the performance of Broadwell-based Xeon CPUs on deep learning inference.

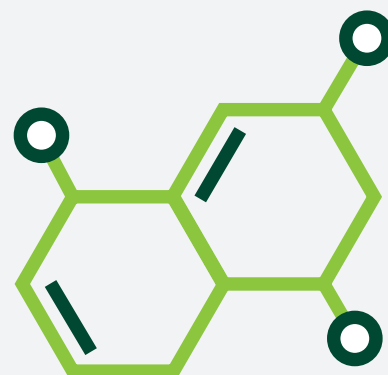
# NVIDIA TENSORRT 3

## Massive Throughput and Amazing Efficiency at Low Latency



CPU throughput based on measured-inference-throughput performance on Broadwell-based Xeon E2690 v4 CPU and doubled to reflect Intel's stated claim that Xeon Scalable Processor will deliver 2X the performance of Broadwell-based Xeon CPUs on deep learning inference.

# 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 over 90% in server and infrastructure acquisition costs.

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

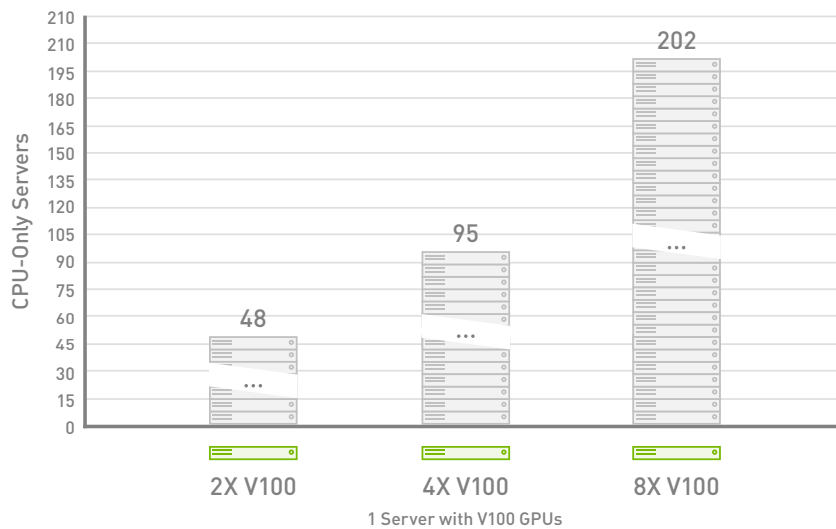
- > Servers with V100 replace over 202 CPU servers for applications such as Amber and HOOMD-blue
- > 100% of the top MD applications are GPU-accelerated
- > Key math libraries like FFT and BLAS
- > Up to 15.7 TFLOPS of single precision performance per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s of memory bandwidth per GPU

View all related applications at:

[www.nvidia.com/molecular-dynamics-apps](http://www.nvidia.com/molecular-dynamics-apps)

## AMBER Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU servers: same CPU server with NVIDIA® Tesla® V100 PCIe or V100 SXM2 on 8X V100 config | NVIDIA CUDA® version: CUDA 9.0.176 | Dataset: PME-Cellulose\_NVE | To arrive at CPU node equivalency, we use a 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

16.12

### ACCELERATED FEATURES

PMEMD explicit solvent and GB; explicit and implicit solvent, REMD, aMD

### SCALABILITY

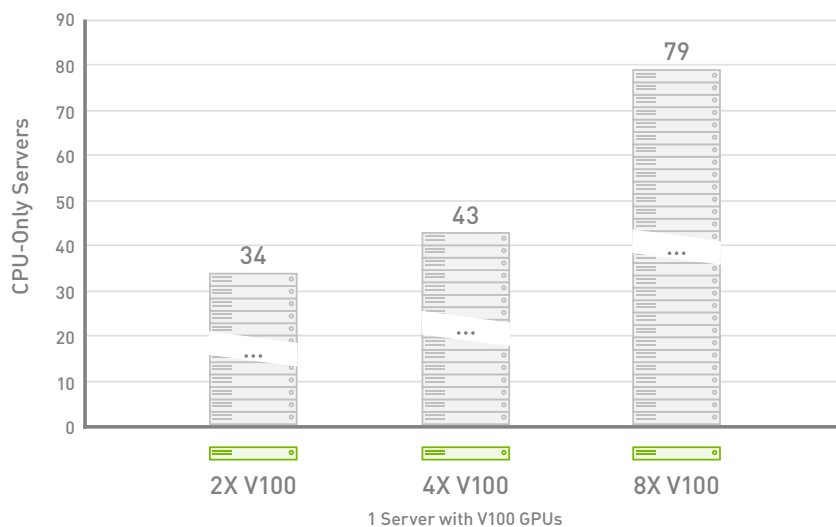
Multi-GPU and Single-Node

### MORE INFORMATION

<http://ambermd.org/gpus>

## HOOMD-blue Performance Equivalency

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe or V100 SXM2 on 8X V100 config | NVIDIA CUDA® Version: 9.0.145 | Dataset: Microsphere | To arrive at CPU node equivalency, 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 written grounds up for GPUs

### VERSION

2.2.2

### ACCELERATED FEATURES

CPU and GPU versions available

### SCALABILITY

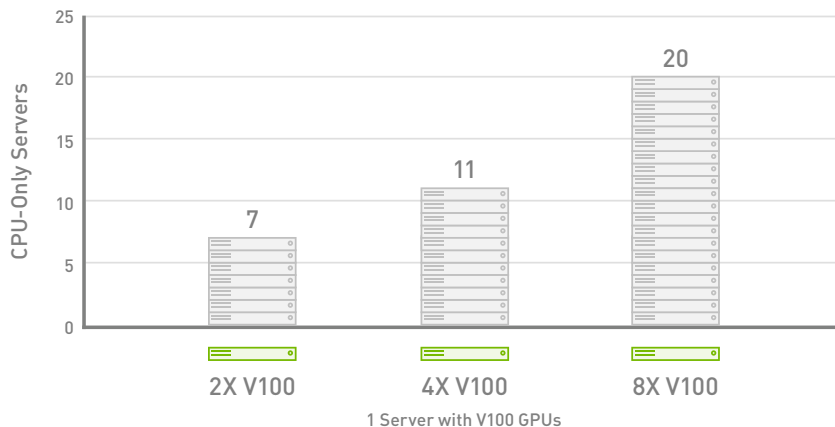
Multi-GPU and Multi-Node

### MORE INFORMATION

<http://codeblue.umich.edu/hoomd-blue/index.html>

## LAMMPS Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Gold 6140 @ 2.30GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe or V100 SXM2 on 8X V100 config | NVIDIA CUDA® Version: 9.0.176 | Dataset: atomic-fluid Lennard-Jones 2.5 cutoff | To arrive at CPU node equivalence, we use a measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

### LAMMPS

Classical molecular dynamics package

### VERSION

2018

### ACCELERATED FEATURES

Lennard-Jones, Gay-Berne, Tersoff, many more potentials

### SCALABILITY

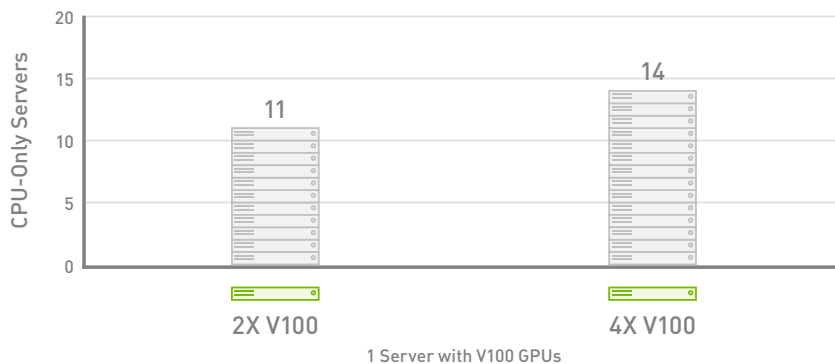
Multi-GPU and Multi-Node

### MORE INFORMATION

<http://lammps.sandia.gov/index.html>

## NAMD Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.176 | Dataset: STMV | To arrive at CPU node equivalence, we use a measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

### NAMD

Designed for high-performance simulation of large molecular systems

### VERSION

2.13

### ACCELERATED FEATURES

Full electrostatics with PME and most simulation features

### SCALABILITY

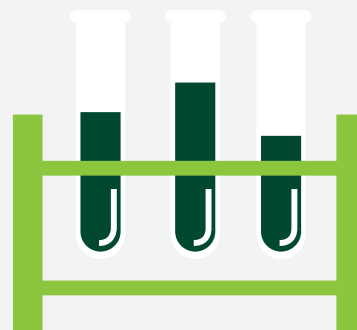
Up to 100M atom capable, multi-GPU, scales to 2x Tesla P100

### MORE INFORMATION

<http://www.ks.uiuc.edu/Research/namd>



# 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 60% in server and infrastructure acquisition costs.

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

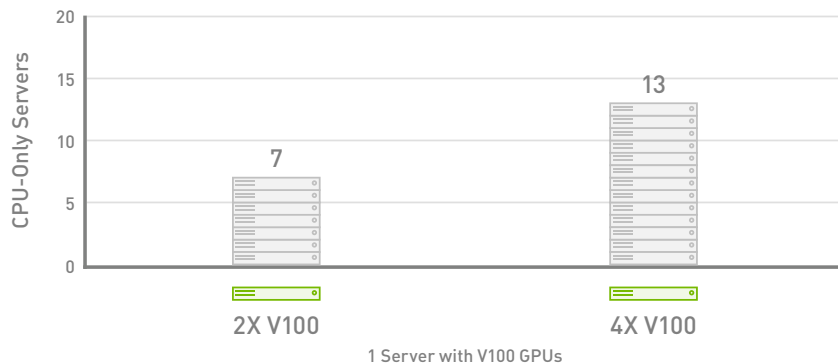
- > Servers with V100 replace up to 13 CPU servers for applications such as Quantum Espresso
- > 60% of the top QC applications are GPU-accelerated
- > Key math libraries like FFT and BLAS
- > Up to 7.8 TFLOPS per second of double precision performance per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s of memory bandwidth per GPU

View all related applications at:

[www.nvidia.com/quantum-chemistry-apps](http://www.nvidia.com/quantum-chemistry-apps)

## Quantum Espresso Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.176 | Dataset: AUSURF112-jR | To arrive at CPU node equivalence, we use a measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

### QUANTUM ESPRESSO

An open-source suite of computer codes for electronic structure calculations and materials modeling at the nanoscale

### VERSION

6.2.1

### ACCELERATED FEATURES

Linear algebra (matrix multiply), explicit computational kernels, 3D FFTs

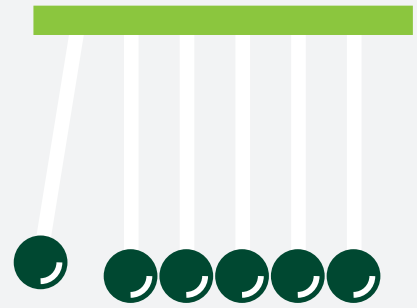
### SCALABILITY

Multi-GPU and Multi-Node

### MORE INFORMATION

<http://www.quantum-espresso.org>

# PHYSICS



From fusion energy to high energy particles, physics simulations span a wide range of applications in the HPC data center. All of the top physics applications are GPU-accelerated, enabling insights previously not possible. A data center with Tesla V100 GPUs can save over 85% in server acquisition cost when running GPU-accelerated physics applications.

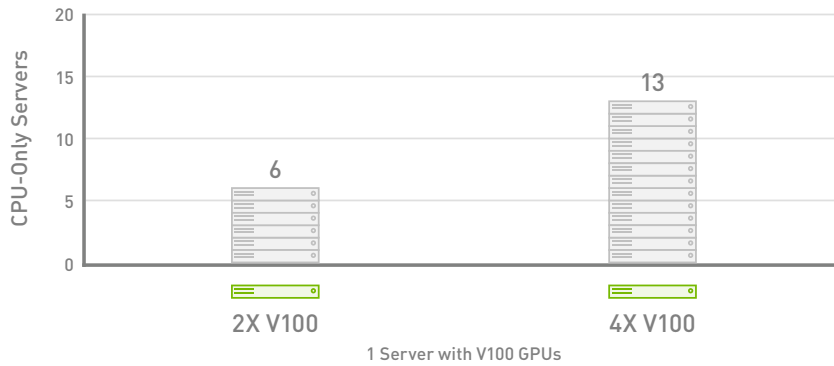
## KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR PHYSICS

- > Servers with V100 replace up to 89 CPU servers for applications such as Chroma, GTC, MILC, and QUDA
- > Most of the top physics applications are GPU-accelerated
- > Up to 7.8 TFLOPS of double precision floating point performance per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s memory bandwidth per GPU

View all related applications at:  
[www.nvidia.com/physics-apps](http://www.nvidia.com/physics-apps)

## Chroma Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.176 | Dataset: szscl21\_24\_128 | To arrive at CPU node equivalence, we use a measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

### CHROMA

Lattice quantum chromodynamics (LQCD)

### VERSION

2018

### ACCELERATED FEATURES

Wilson-clover fermions, Krylov solvers, Domain-decomposition

### SCALABILITY

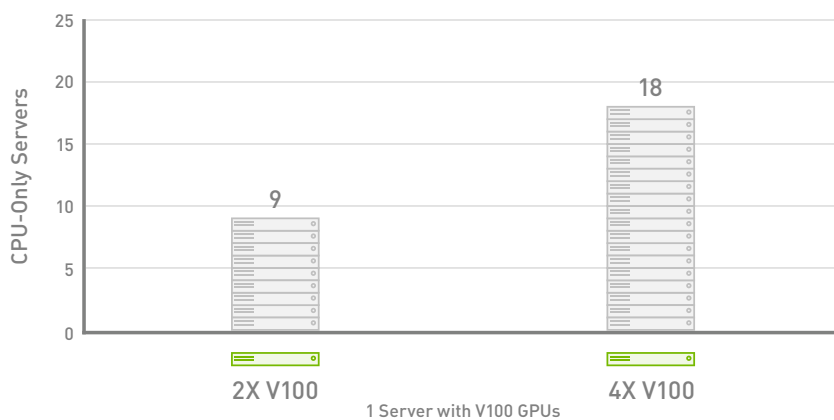
Multi-GPU

### MORE INFORMATION

<http://jeffersonlab.github.io/chroma>

## GTC Performance Equivalence

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.176 | Dataset: mpi#proc.in | To arrive at CPU node equivalence, we use a measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

### GTC

Used for gyrokinetic particle simulation of turbulent transport in burning plasmas

### VERSION

2017

### ACCELERATED FEATURES

Push, shift, collision

### SCALABILITY

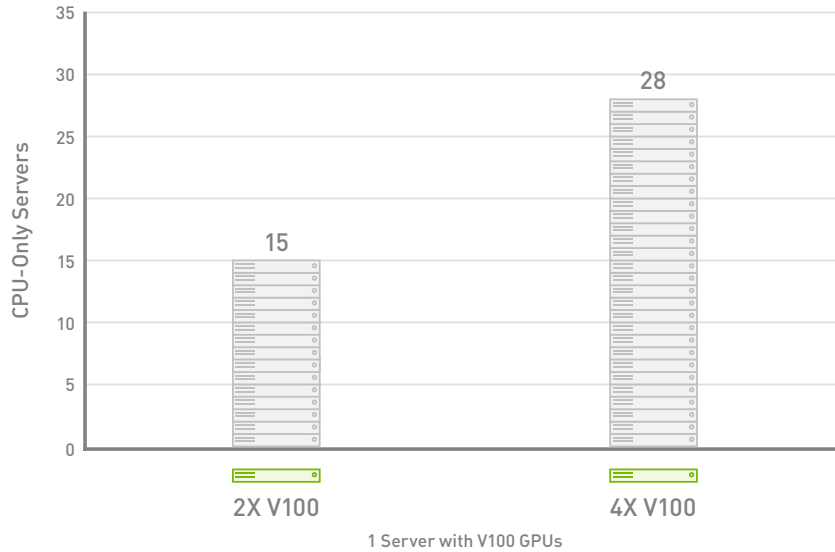
Multi-GPU

### MORE INFORMATION

[www.nvidia.com/gtc-p](http://www.nvidia.com/gtc-p)

## MILC Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30GHz, GPU Servers: Dual Xeon Gold 6140 @ 2.30GHz with NVIDIA® Tesla® V100 PCIe (16 GB) | NVIDIA CUDA® Version: 9.0.176 | Dataset: MILC APEX Medium | To arrive at CPU node equivalence, we use a 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**  
2018

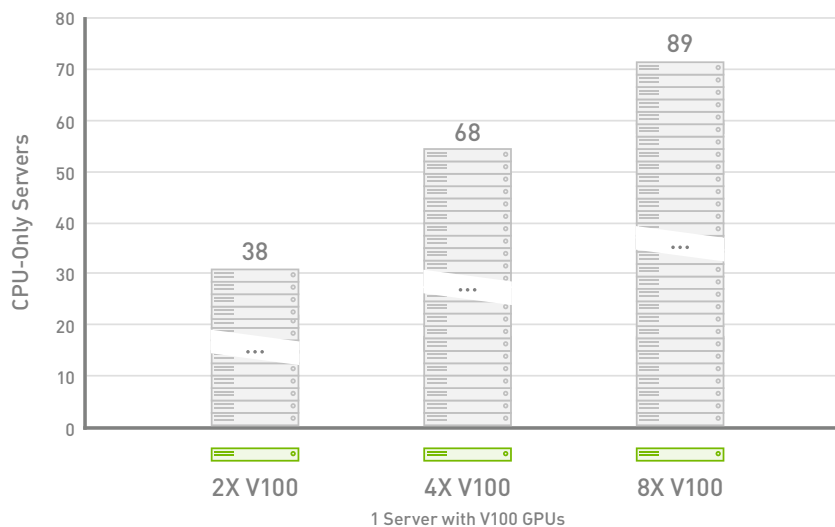
**ACCELERATED FEATURES**  
Staggered fermions, Krylov solvers, gauge-link fattening

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

**MORE INFORMATION**  
[www.nvidia.com/milc](http://www.nvidia.com/milc)

## QUDA Performance Equivalence

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe or V100 SXM2 on 8X V100 config | 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 chromodynamics on GPUs

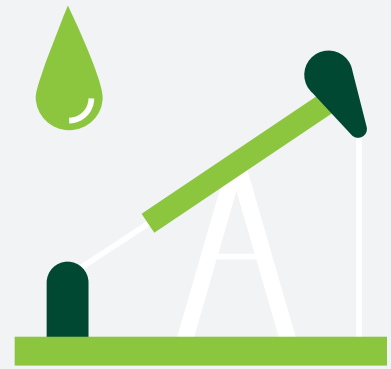
**VERSION**  
2017

**ACCELERATED FEATURES**  
All

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

**MORE INFORMATION**  
[www.nvidia.com/quda](http://www.nvidia.com/quda)

# 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 90% in server and infrastructure acquisition costs.

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

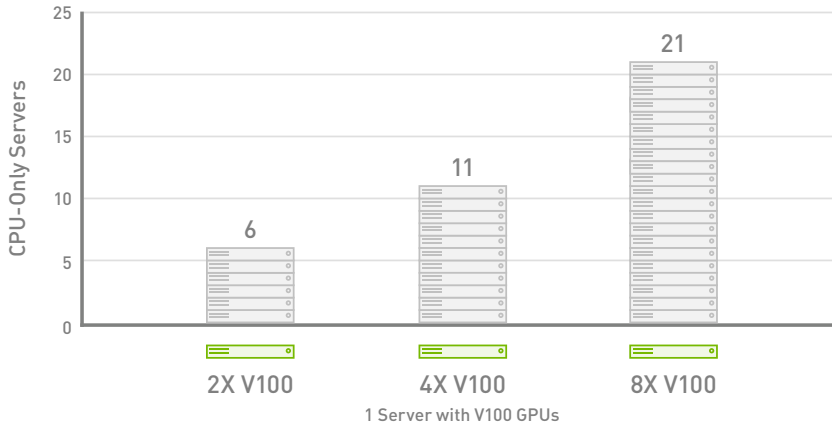
- > Servers with V100 replace up to 54 CPU servers for applications such as RTM and SPECFEM3D
- > Top Geoscience applications are GPU-accelerated
- > Up to 15.7 TFLOPS of single precision floating point performance
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s memory bandwidth per GPU

View all related applications at:

[www.nvidia.com/oil-and-gas-apps](http://www.nvidia.com/oil-and-gas-apps)

## RTM Performance Equivalence

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe or V100 SXM2 on 8X V100 config | 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 node.

### RTM

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

### VERSION

2017

### ACCELERATED FEATURES

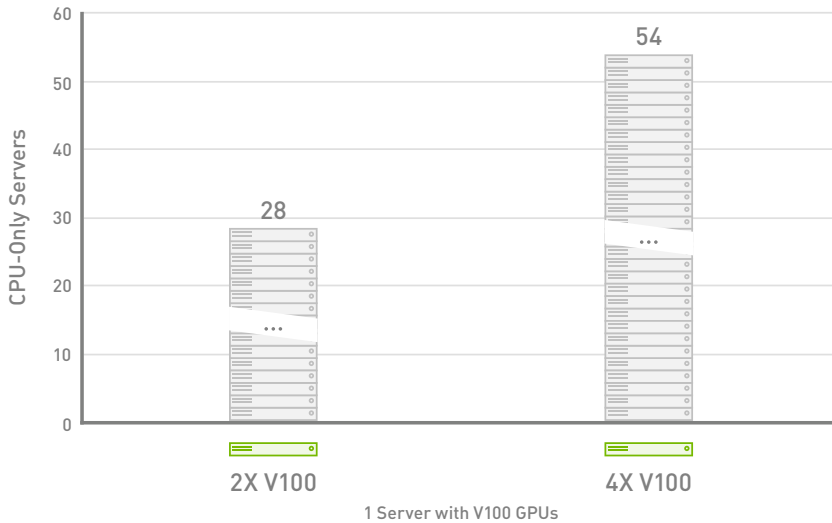
Batch algorithm

### SCALABILITY

Multi-GPU and Multi-Node

## SPECFEM3D Performance Equivalence

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.176 | Dataset: four\_material\_simple\_model | To arrive at CPU node equivalence, we use linear scaling to scale beyond 1 node.

### SPECFEM3D

Simulates Seismic wave propagation

### VERSION

2.0.2

### SCALABILITY

Multi-GPU and Single-Node

### MORE INFORMATION

[https://geodynamics.org/cig/software/specfem3d\\_globe](https://geodynamics.org/cig/software/specfem3d_globe)

# 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 70% in software licensing costs and 50% in server and infrastructure acquisition costs.

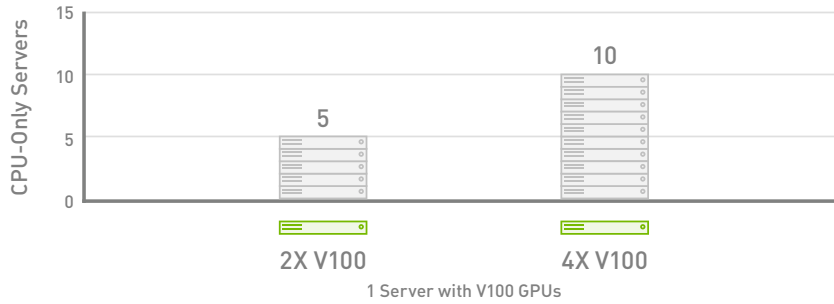
## KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR ENGINEERING

- > Servers with Tesla V100 replace up to 10 CPU servers for applications such as SIMULIA Abaqus and ANSYS FLUENT
- > The top engineering applications are GPU-accelerated
- > Up to 7.8 TFLOPS of double precision floating point performance per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s memory bandwidth per GPU



## SIMULIA Abaqus Performance Equivalency

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 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 eigensolver, steady-state dynamics solver

#### SCALABILITY

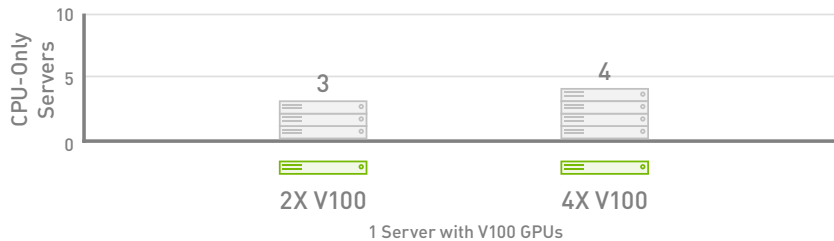
Multi-GPU and Multi-Node

#### MORE INFORMATION

<http://www.nvidia.com/simulia-abaqus>

## ANSYS Fluent Performance Equivalency

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 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](http://www.nvidia.com/ansys-fluent)

# COMPUTATIONAL FINANCE



Computational finance applications are essential to the success of global financial service firms when performing market and counterparty risk analytics, asset pricing, and portfolio risk management analysis. This analysis requires numerical methods that are computationally intensive. And because time is money in financial analysis, several of the leading computational finance applications are GPU-accelerated. Computational finance applications using Tesla V100 GPUs can improve performance by over 50X and save up to 80% in server and infrastructure acquisition costs.

## KEY FEATURES OF THE TESLA PLATFORM AND V100 FOR COMPUTATIONAL FINANCE

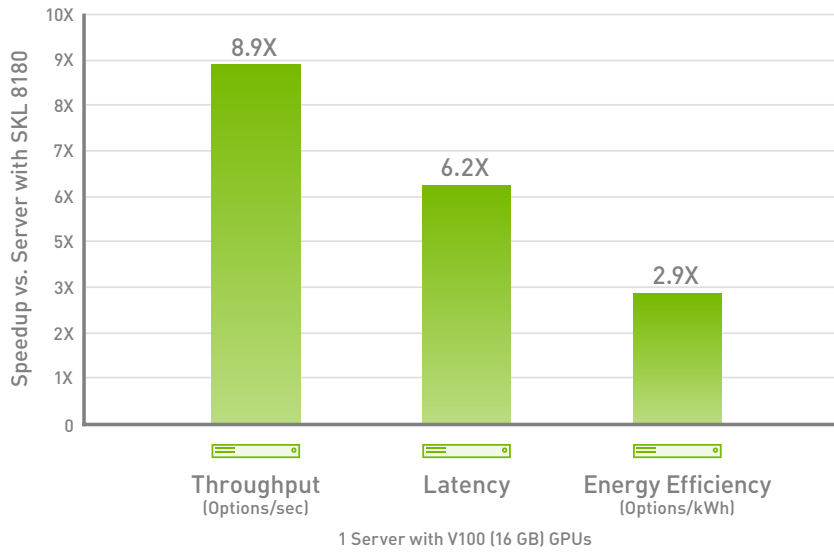
- > Servers with V100 outperform CPU servers by nearly 9X based on STAC-A2 benchmark results
- > Top Computational Finance applications are GPU-accelerated
- > Up to 7.8 TFLOPS per second of double precision performance per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s memory bandwidth per GPU

View all related applications at:

[www.nvidia.com/en-us/data-center/gpu-accelerated-applications/catalog](http://www.nvidia.com/en-us/data-center/gpu-accelerated-applications/catalog)

## STAC-A2 Benchmark Performance Results

8X V100 GPU Server vs Dual Skylake Platinum 8180 Sever



System Configuration: GPU Server SUT ID: NVDA171020 | STAC-A2 Pack for CUDA (Rev D) | GPU Server: 8X NVIDIA® Tesla® V100 (Volta) GPU, 2X Intel® Xeon E5-2680 v4 @ 2.4 GHz | CUDA Version 9.0 | Compared to: CPU Server SUT ID: INTC170920, | STAC-A2 Pack for Intel Composer XE (Rev K) | 2X 28-Core Intel Xeon Platinum 8180 @ 2.5GHz "Throughput" is STAC-A2.B2.HPORTFOLIO.SPEED, "Latency" is STAC-A2.B2.GREEKS.WARM, and "Energy Efficiency" is STAC-A2.B2.ENERG\_EFF.I "STAC" and all STAC names are trademarks or registered trademarks of the Securities Technology Analysis Center, LLC.

### STAC-A2

Financial risk management benchmark created by leading global banks working with the Securities Technology Analysis Center (STAC) used to assess financial compute solutions

### VERSION

STAC-A2 (Beta 2)

### ACCELERATED FEATURES

Full framework accelerated

### SCALABILITY

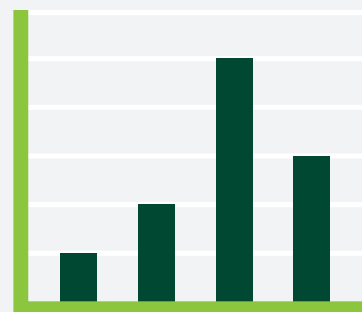
Multi-GPU

### MORE INFORMATION

[www.STACresearch.com/nvidia](http://www.STACresearch.com/nvidia)

# TESLA V100 PERFORMANCE GUIDE

## HPC BENCHMARKS



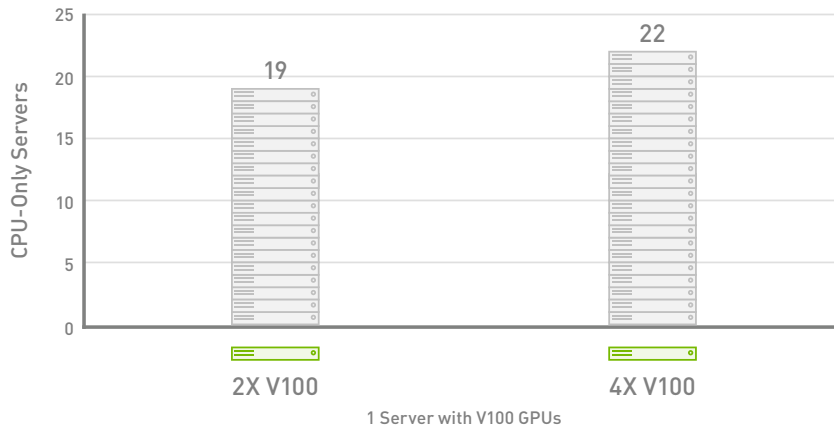
Benchmarks provide an approximation of how a system will perform at production-scale and help to assess the relative performance of different systems. The top benchmarks have GPU-accelerated versions and can help you understand the benefits of running GPUs in your data center.

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

- > Servers with Tesla V100 replace up to 23 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 performance per GPU
- > Up to 32 GB of memory capacity per GPU
- > Up to 900 GB/s memory bandwidth per GPU

## Cloverleaf Performance Equivalency

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 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

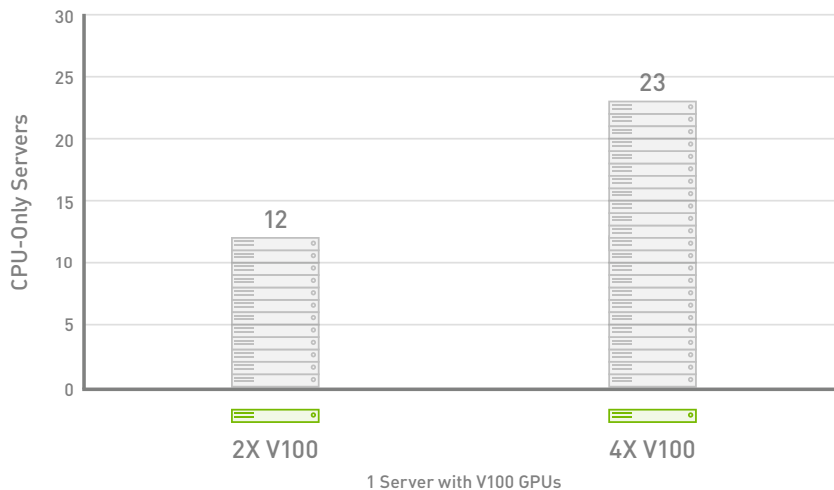
Multi-Node (MPI)

### MORE INFORMATION

<http://uk-mac.github.io/CloverLeaf>

## HPCG Performance Equivalency

Single GPU Server vs Multiple Broadwell CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 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

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

### VERSION

3

### ACCELERATED FEATURES

All

### SCALABILITY

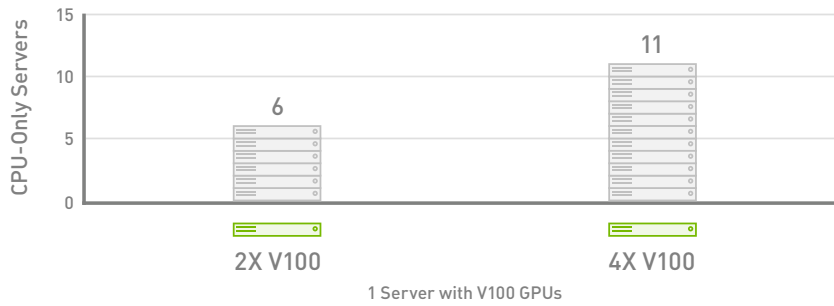
Multi-GPU and Multi-Node

### MORE INFORMATION

<http://www.hpcg-benchmark.org/index.html>

## Linpack Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: HPL.dat | To arrive at CPU node equivalence, we use a 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

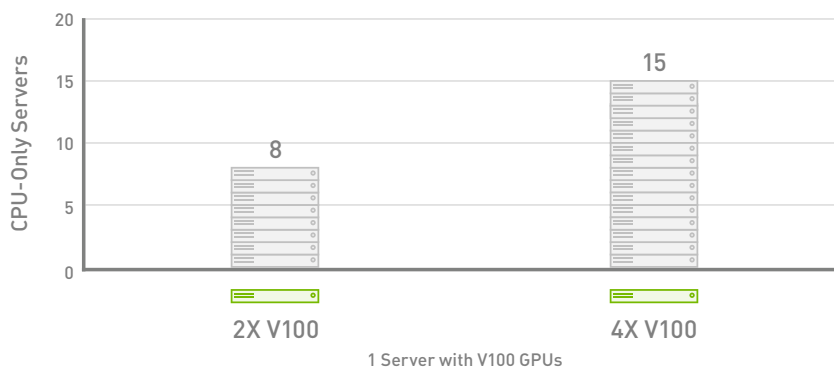
Multi-Node and Multi-Node

### MORE INFORMATION

[www.top500.org/project/linpack](http://www.top500.org/project/linpack)

## MiniFE Performance Equivalency

Single GPU Server vs Multiple Skylake CPU-Only Servers



CPU Server: Dual Xeon Gold 6140 @ 2.30 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® V100 PCIe | NVIDIA CUDA® Version: 9.0.103 | Dataset: 350x350x350 | To arrive at CPU node equivalence, we use a 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

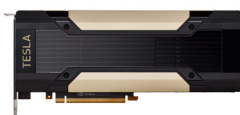
### SCALABILITY

Multi-GPU

### MORE INFORMATION

<https://mantevo.org/about/applications>

## 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	32 GB / 16 GB	32 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.