TESLA P100 PERFORMANCE GUIDE

Deep Learning and HPC Applications

SEPTEMBER 2017

TESLA P100 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 P100 GPU is the engine of the modern data center, delivering breakthrough performance with fewer servers resulting in faster insights and dramatically lower costs.

Every HPC data center can benefit from the Tesla platform. Over 450 HPC applications in a broad range of domains are optimized for GPUs, including all 10 of the top 10 HPC applications and every major deep learning framework.

RESEARCH DOMAINS WITH GPU-ACCELERATED APPLICATIONS INCLUDE:

DEEP LEARNING

MOLECULAR DYNAMICS

QUANTUM CHEMISTRY

PHYSICS

OIL AND GAS

ENGINEERING

HPC BENCHMARKS

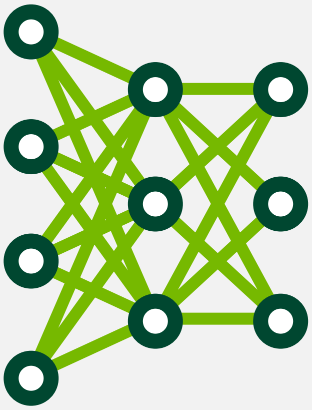
Over 450 HPC applications and all deep learning frameworks are GPU-accelerated.

> 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 P100 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 frameworks, a data center with Tesla P100 GPUs can save up to 70% in server acquisition cost.

KEY FEATURES OF THE TESLA PLATFORM AND P100 FOR DEEP LEARNING TRAINING

> Caffe, TensorFlow, and CNTK are up to 3x faster with Tesla P100 compared to K80

> 100% of the top deep learning frameworks are GPU-accelerated

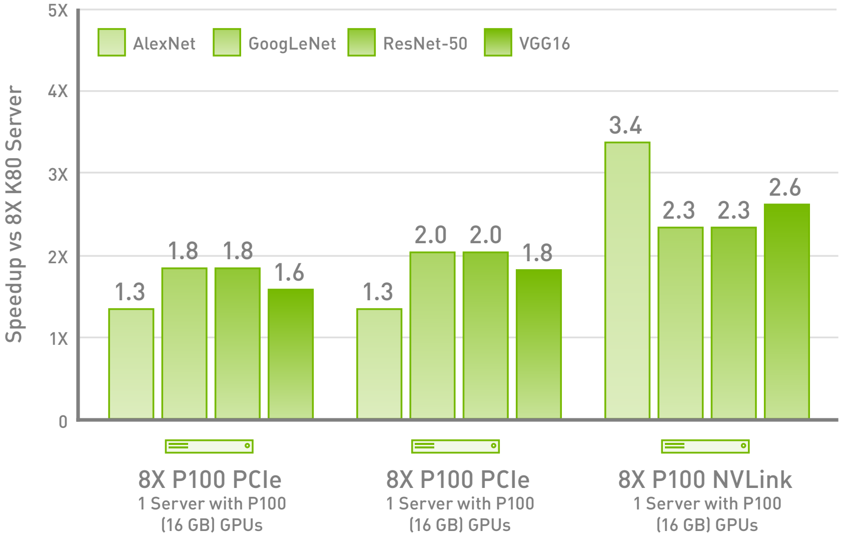
> Up to 21.2 TFLOPS of native half precision floating point

> Up to 16 GB of memory capacity with up to 732 GB/s memory bandwidth

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

Caffe Deep Learning Relative Performance

Training on P100 Servers vs K80 Server



CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | Ubuntu: 14.04.5 | NVIDIA CUDA® Version: 8.0.54 | cuDNN: 6.0.5 Dataset: ImageNet | Batch Sizes: AlexNet (128), GoogleNet (256), ResNet-50 (64) VGG-16 (32) | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

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

VERSION
0.16

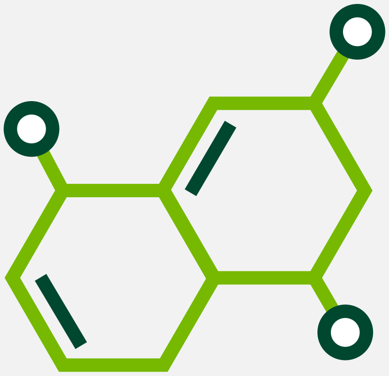
ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU

MORE INFORMATION
www.nvidia.com/caffe

TESLA P100 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 P100 GPUs can save up to 60% in server acquisition cost.

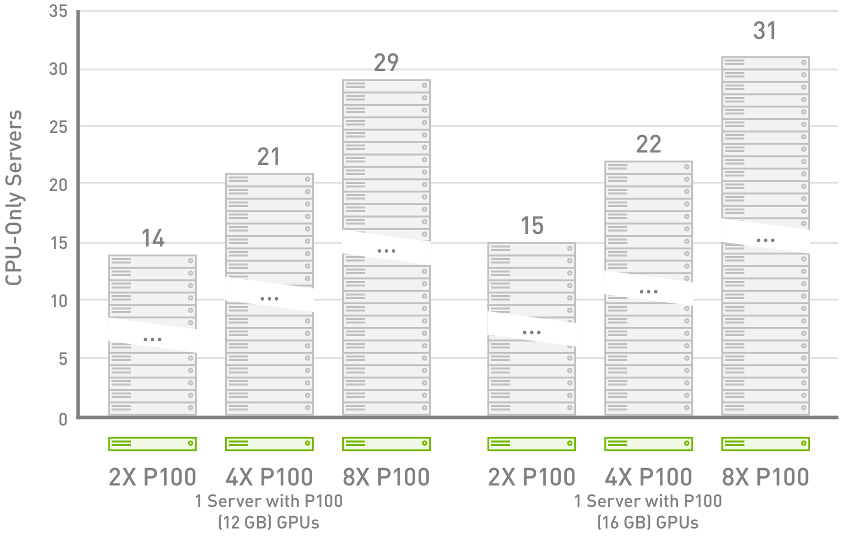
KEY FEATURES OF THE TESLA PLATFORM AND P100 FOR MD

- Servers with P100 replace up to 40 CPU servers for applications such as HOOMD-Blue, LAMMPS, AMBER, GROMACS, and NAMD
- 100% of the top MD applications are GPU-accelerated
- Key math libraries like FFT and BLAS
- Up to 11 TFLOPS per second of single precision performance per GPU
- Up to 732 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: Microsphere | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

HOOMD-BLUE

Particle dynamics package is written from the ground up for GPUs

VERSION

1.3.3

ACCELERATED FEATURES

CPU & GPU versions available

SCALABILITY

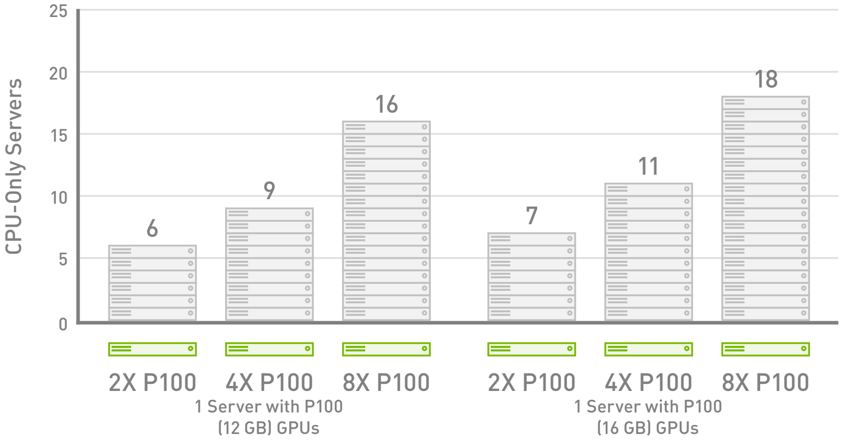
Multi-GPU and Multi-Node

MORE INFORMATION

www.nvidia.com/hoomd-blue

LAMMPS 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: EAM | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

LAMMPS

Classical molecular dynamics package

VERSION

2016

ACCELERATED FEATURES

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

SCALABILITY

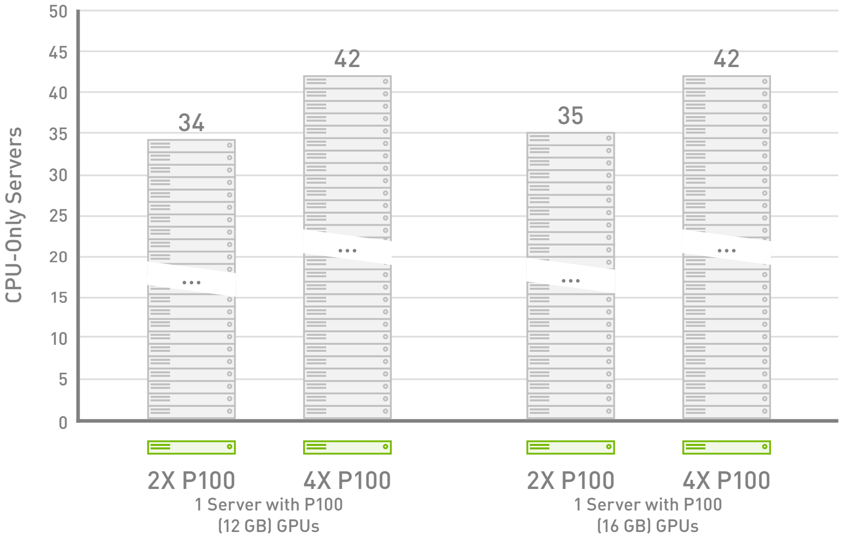
Multi-GPU and Multi-Node

MORE INFORMATION

www.nvidia.com/lammps

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: GB-Myoglobin | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

AMBER
Suite of programs to simulate molecular dynamics on biomolecule

VERSION
16.3

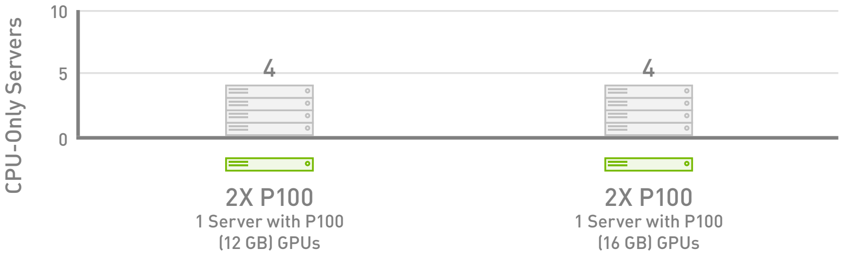
ACCELERATED FEATURES
PMEMD Explicit Solvent & GB; Explicit & Implicit Solvent, REMD, aMD

SCALABILITY
Multi-GPU and Single-Node

MORE INFORMATION
www.nvidia.com/amber

GROMACS 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: Water 3M | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes.

GROMACS
Simulation of biochemical molecules with complicated bond interactions

VERSION
5.1.2

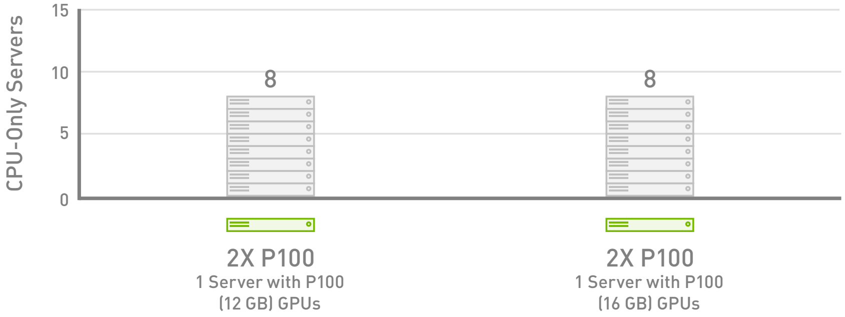
ACCELERATED FEATURES
PME, Explicit, and Implicit Solvent

SCALABILITY
Multi-GPU and Multi-Node
Scales to 4xP100

MORE INFORMATION
www.nvidia.com/gromacs

NAMD 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: STMV | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

NAMD
Designed for high-performance simulation of large molecular systems

VERSION
2.11

ACCELERATED FEATURES
Full electrostatics with PME and many simulation features

SCALABILITY
Up to 100M atom capable, Multi-GPU, Scales to 2xP100

MORE INFORMATION
www.nvidia.com/namd

TESLA P100 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 P100 GPUs can save up to 40% in server acquisition cost.

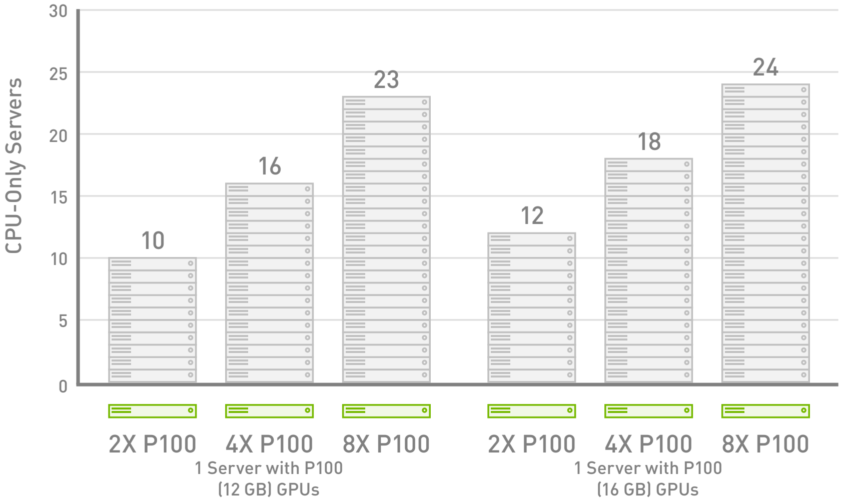
KEY FEATURES OF THE TESLA PLATFORM AND P100 FOR QC

- Servers with P100 replace up to 36 CPU servers for applications such as VASP and LSMS
- 60% of the top QC applications are GPU-accelerated
- Key math libraries like FFT and BLAS
- Up to 5.3 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



PU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: B_hR105| To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

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

VERSION
5.4.1

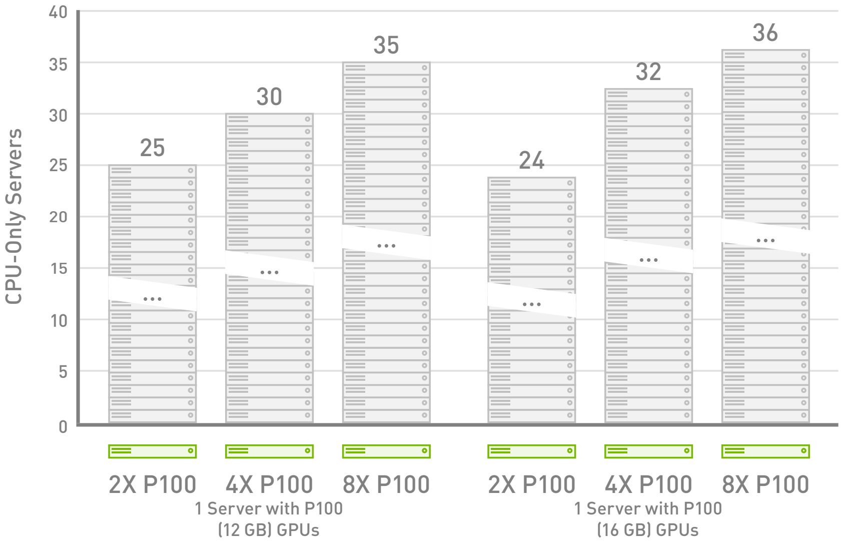
ACCELERATED FEATURES
RMM-DIIS, Blocked Davidson, K-points, and exact-exchange

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/vasp

LSMS 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: Fe16_new | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

LSMS
Materials code for investigating the effects of temperature on magnetism

VERSION
3

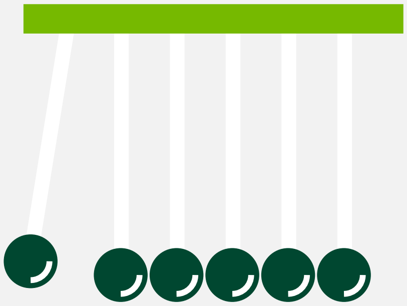
ACCELERATED FEATURES
Generalized Wang-Landau method

SCALABILITY
Multi-GPU

MORE INFORMATION
www.nvidia.com/lsms

TESLA P100 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 P100 GPUs can save up to 70% in server acquisition cost when running GPU-accelerated physics applications.

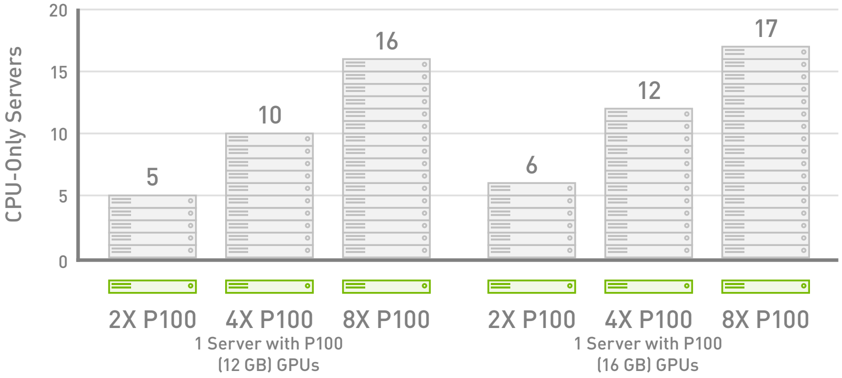
KEY FEATURES OF THE TESLA PLATFORM AND P100 FOR PHYSICS

- Servers with P100 replace up to 50 CPU servers for applications such as GTC-P, QUDA, MILC and Chroma
- Most of the top physics applications are GPU-accelerated
- Up to 5.3 TFLOPS of double precision floating point performance
- Up to 16 GB of memory capacity with up to 732 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: A.txt | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

GTC-P
A development code for optimization of plasma physics

VERSION
2016

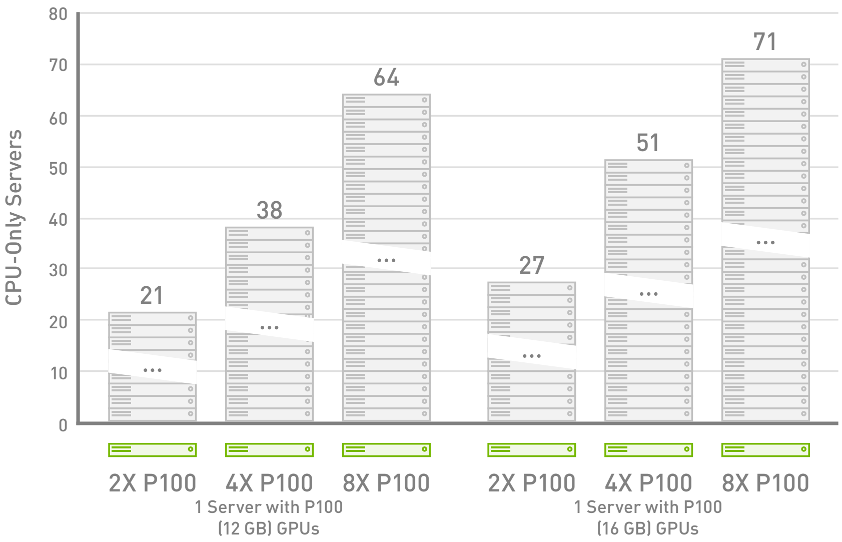
ACCELERATED FEATURES
Push, shift, and collision

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: Glove Precision Single, Gauge Compression/Recon: 12; Problem Size 32x32x32x64 | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

QUDA
A library for Lattice Quantum Chromo Dynamics on GPUs

VERSION
2017

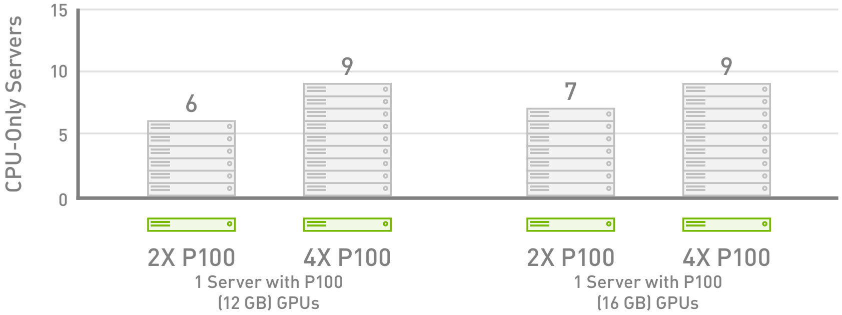
ACCELERATED FEATURES
All

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: Double Precision | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling 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
7.8.0

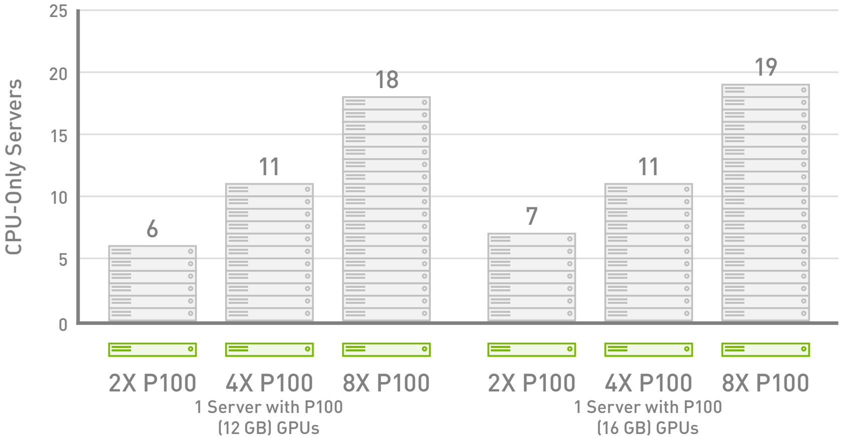
ACCELERATED FEATURES
Staggered fermions, Krylov solvers, and Gauge-link fattening
Scales to 4xP100

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/milc

Chroma 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: szscl21_24_128 (total time sec) | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

CHROMA
Lattice Quantum Chromodynamics (LQCD)

VERSION
2016

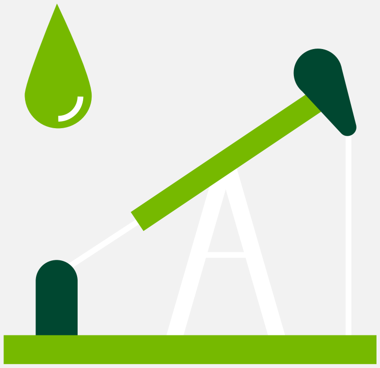
ACCELERATED FEATURES
Wilson-clover fermions, Krylov solvers, and Domain-decomposition

SCALABILITY
Multi-GPU

MORE INFORMATION
www.nvidia.com/chroma

TESLA P100 PERFORMANCE GUIDE

OIL AND GAS



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 P100 GPUs can save up to 65% in server acquisition cost.

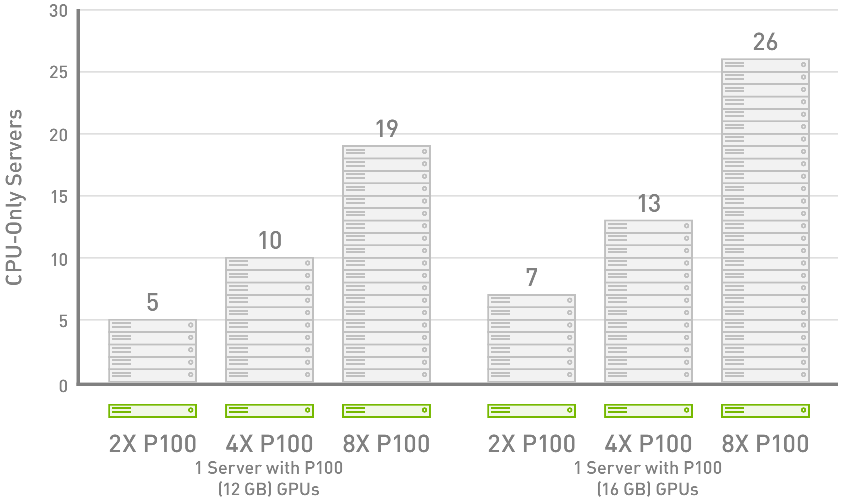
KEY FEATURES OF THE TESLA PLATFORM AND P100 FOR GEOSCIENCE

- Servers with P100 replace up to 50 CPU servers for applications such as RTM and SPECFEM 3D
- Top Oil and Gas applications are GPU-accelerated
- Up to 10.6 TFLOPS of single precision floating point performance
- Up to 16 GB of memory capacity with up to 732 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: TTI R8 3 pass | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

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

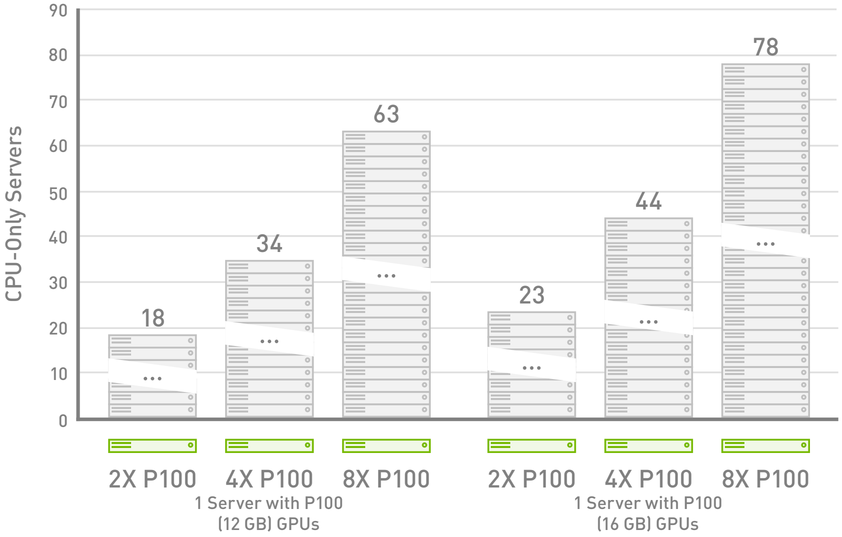
VERSION
2016

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: Globe 112x64, 100 mins | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

SPECFEM 3D
Simulates Seismic wave propagation

VERSION
7.0.0

ACCELERATED FEATURES
Wilson-clover fermions, Krylov solvers, and Domain-decomposition

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/specfem3d-globe

TESLA P100 PERFORMANCE GUIDE

ENGINEERING

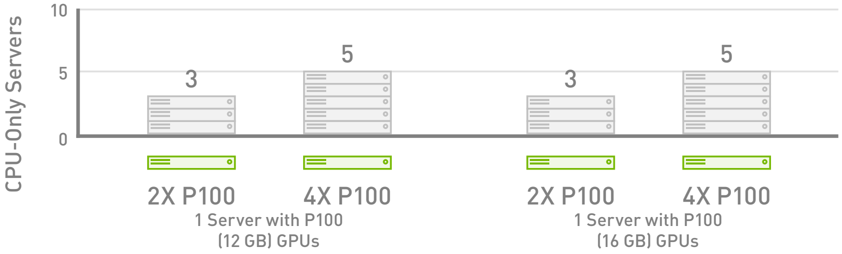
Engineering simulations are key to developing new products across industries by modeling flows, heat transfers, finite element analysis and more. Many of the top Engineering applications are accelerated with GPUs today. When running Engineering applications, a data center with NVIDIA® Tesla® P100 GPUs can save over 25% in server acquisition costs and over 50% in software licensing costs.

KEY FEATURES OF THE TESLA PLATFORM AND P100 FOR ENGINEERING

- Servers with Tesla P100 replace up to 5 CPU-only 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 732 GB/s memory bandwidth
- Up to 5.3 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 7.5 | 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

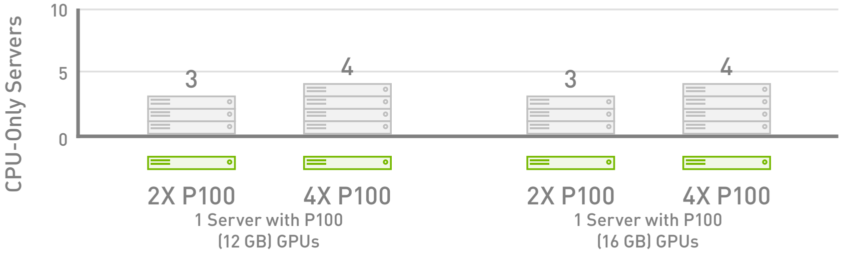
ACCELERATED FEATURES
Direct Sparse Solver
AMS Eigen Solver
Steady-state Dynamics Solver

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® P100 for PCIe (12 GB or 16 GB) | 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

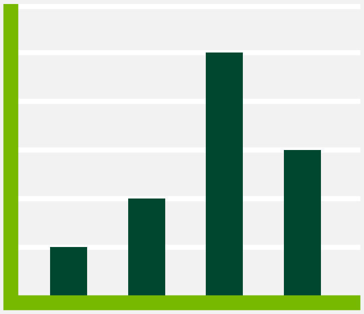
ACCELERATED FEATURES
Pressure-based Coupled Solver and Radiation Heat Transfer

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
www.nvidia.com/ansys-fluent

TESLA P100 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 P100 FOR BENCHMARKING

> Servers with Tesla P100 replace up to 39 CPU servers for benchmarks such as Cloverleaf, MiniFE, Linpack, and HPCG

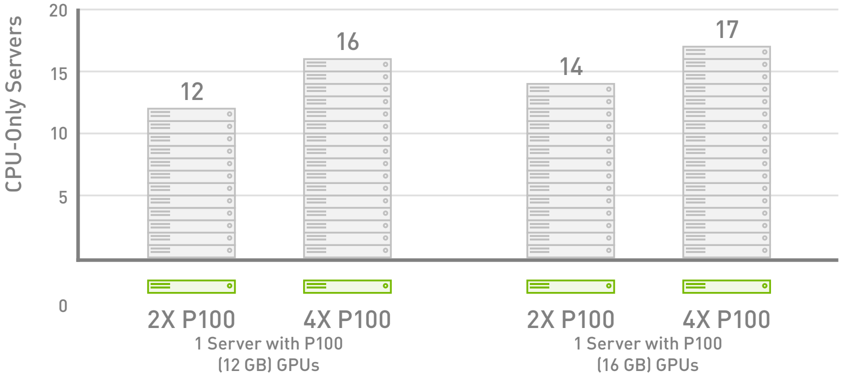
> The top benchmarks are GPU-accelerated

> Up to 5.3 TFLOPS of double precision floating point up to 16 GB of memory capacity

> Up to 732 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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | 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

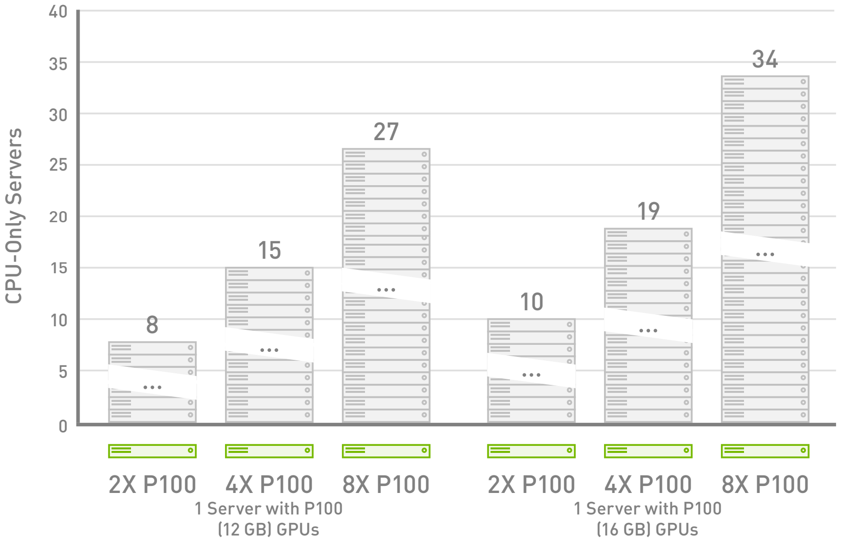
ACCELERATED FEATURES
Lagrangian-Eulerian
explicit hydrodynamics mini-application

SCALABILITY
Multi-Node (MPI)
Scales to 4X P100

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | 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

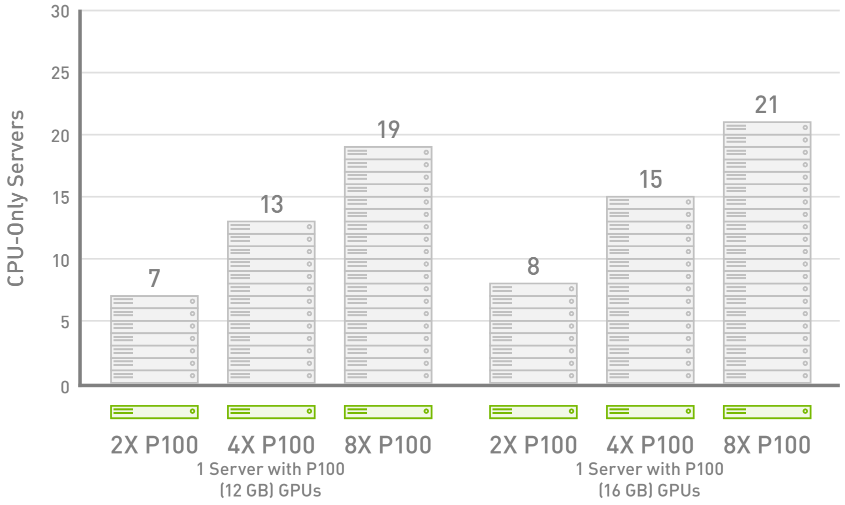
ACCELERATED FEATURES
All

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | 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

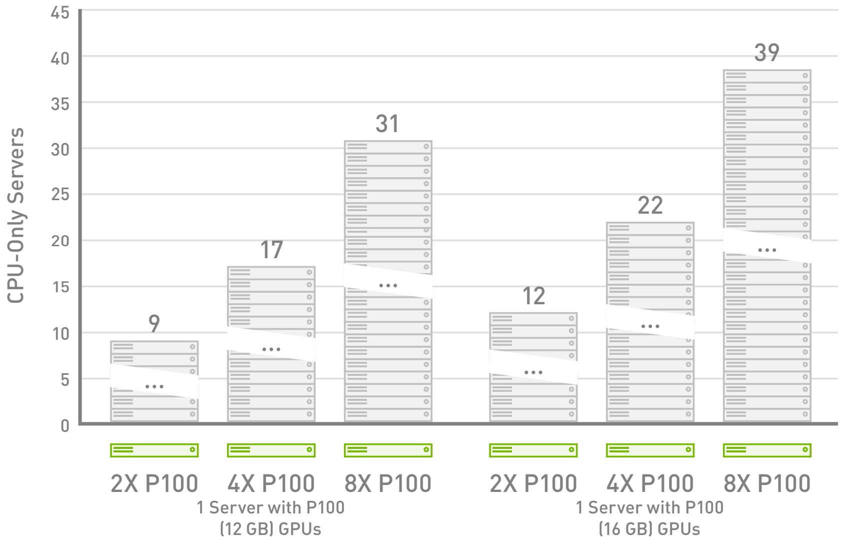
ACCELERATED FEATURES
All

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® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | 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

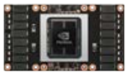
VERSION
3

ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU and Multi-Node

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

TESLA P100 PRODUCT SPECIFICATIONS



	NVIDIA Tesla P100 for PCIe-Based Servers	NVIDIA Tesla P100 for NVLink-Optimized Servers
Double-Precision Performance	up to 4.7 TFLOPS	up to 5.3 TFLOPS
Single-Precision Performance	up to 9.3 TFLOPS	up to 10.6 TFLOPS
Half-Precision Performance	up to 18.7 TFLOPS	up to 21.2 TFLOPS
NVIDIA NVLink™ Interconnect Bandwidth	-	160 GB/s
PCIe x 16 Interconnect Bandwidth	32 GB/s	32 GB/s
CoWoS HBM2 Stacked Memory Capacity	16 GB or 12 GB	16 GB
CoWoS HBM2 Stacked Memory Bandwidth	732 GB/s or 549 GB/s	732 GB/s

Assumptions and Disclaimers
The percentage of top applications that are GPU-accelerated is from top 50 app list in the i360 report: H PC Application Support for GPU Computing. Calculation of throughput and cost savings assumes a workload profile where applications benchmarked in the domain take equal compute cycles.