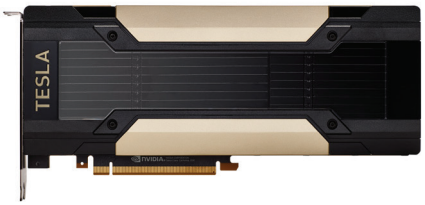
BOSTON
Servers | Storage | Solutions

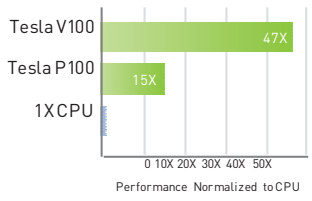
NVIDIA TESLA V100 GPU ACCELERATOR

The Most Advanced Data Center GPU Ever Built.

NVIDIA® Tesla® V100 is the world's most advanced data center GPU ever built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta™, the latest GPU architecture, Tesla V100 offers the performance of up to 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once thought impossible.

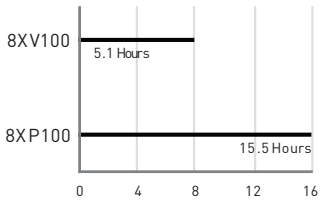


47X Higher Throughput than CPU Server on Deep Learning Inference



Workload: ResNet-50 | CPU: 1X Xeon E5-2690v4 @ 2.6GHz | GPU: add 1X NVIDIA® Tesla® P100 or V100

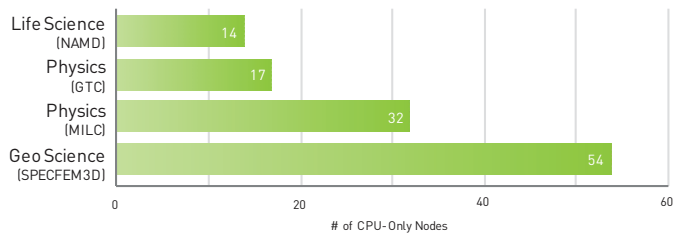
Deep Learning Training in Less Than a Workday



Server Confg: Dual Xeon E5-2699 v4 2.6GHz | 8X Tesla P100 or V100 | ResNet-50 Training on MXNet for 90 Epochs with 1.28M ImageNet dataset.

1 GPU Node Replaces Up To 54 CPU Nodes

Node Replacement: HPC Mixed Workload



CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ 4x V100 PCIe | CUDA Version: CUDA 9.x| Dataset: NAMD (STMV), GTC (mpi#proc.in), MILC (APEX Medium), SPECFEM3D (four_material_simple_model) | 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.

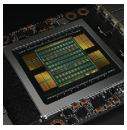
SPECIFICATIONS

Part Number	TCSV100MPCIE-PB TCSV100M-32GB-PB TCSV100SM-32GB-PB	
	TESLA V100 PCIe	**TESLA V100S PCIe**
GPU Architecture	**NVIDIA Volta**	
NVIDIA Tensor Cores	**640**	
NVIDIA CUDA® Cores	**5,120**	
Double-Precision Performance	**7 TFLOPS**	**8,2 TFLOPS**
Single-Precision Performance	**14 TFLOPS**	**16,4 Tflops**
Tensor Performance	**112 TFLOPS**	**130 Tflops**
GPU Memory	**16/32 GB HBM2**	**32 GB HBM2**
Memory Bandwidth	**900 GB/sec**	**1134 GB/s**
ECC	**Yes**	
System Interface	**PCIe Gen3**	
Form Factor	**PCIe Full Height**	
Max Power Comsumption	**250 W**	
Thermal Solution	**Passive**	
Compute APIs	**CUDA, DirectCompute, OpenCL™, OpenACC**	

BOSTON
Servers | Storage | Solutions

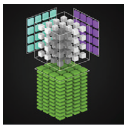
NVIDIA

GROUNDBREAKING INNOVATIONS



VOLTA ARCHITECTURE

By pairing CUDA Cores and Tensor Cores within a unified architecture, a single server with Tesla V100 GPUs can replace hundreds of commodity CPU servers for traditional HPC and Deep Learning.

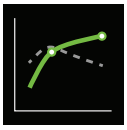


TENSOR CORE

Equipped with 640 Tensor Cores, Tesla V100 delivers 120 TeraFLOPS of deep learning performance. That's 12X Tensor FLOPS for DL Training, and 6X Tensor FLOPS for DL Inference when compared to NVIDIA Pascal™ GPUs.

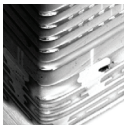
NEXT GENERATION NVLINK

NVIDIA NVLink in Tesla V100 delivers 2X higher throughput compared to the previous generation. Up to eight Tesla V100 accelerators can be interconnected at up to 300 GB/s to unleash the highest application performance possible on a single server.



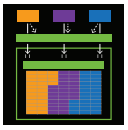
MAXIMUM EFFICIENCY MODE

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, Tesla 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 900 GB/s and higher DRAM utilization efficiency at 95%, Tesla V100 delivers 1.5X higher memory bandwidth over Pascal GPUs as measured on STREAM.



PROGRAMMABILITY

Tesla 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.

Tesla V100 is the flagship product of Tesla data center computing platform for deep learning, HPC, and graphics. The Tesla platform accelerates over 550 HPC applications and every major deep learning framework. It is available everywhere from desktops to servers to cloud services, delivering both dramatic performance gains and cost savings opportunities.

EVERY DEEP LEARNING FRAMEWORK

Caffe2

Microsoft Cognitive Toolkit

mxnet

PYTORCH

TensorFlow

theano

600+ GPU-ACCELERATED APPLICATIONS

- AMBER
- ANSYS Fluent
- GAUSSIAN
- GROMACS
- LS-DYNA
- NAMD
- OpenFOAM
- Simulia Abaqus
- VASP
- WRF

SPEAK TO BOSTON TODAY FOR MORE INFORMATION

WEB: WWW.BOSTON.CO.UK
EMAIL: SALES@BOSTON.CO.UK
PHONE: +44 (0) 1727 876 100

© 2019 NVIDIA Corporation and PNY. All rights reserved. NVIDIA, the NVIDIA logo, Quadro, nView, CUDA, NVIDIA Pascal, and 3D Vision are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. The PNY logotype is a registered trademark of PNY Technologies. 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.

BOSTON
Servers I Storage I Solutions

NVIDIA