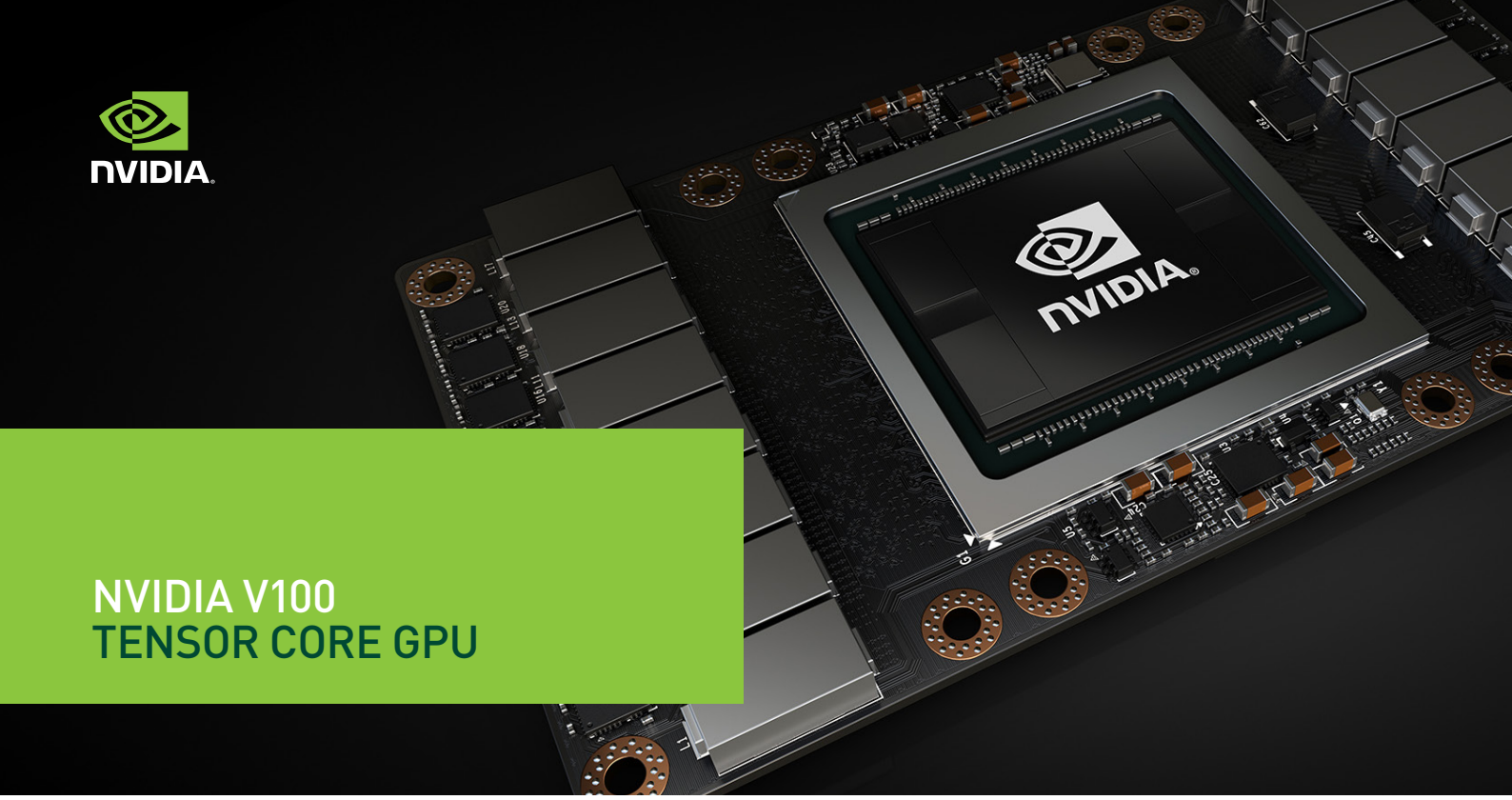
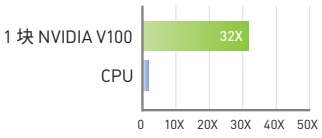


NVIDIA V100
TENSOR CORE GPU

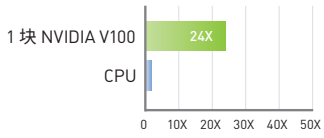
世界上强大的 GPU

NVIDIA® V100 Tensor Core GPU 是深度学习、机器学习、高性能计算 (HPC) 和图形计算的强力加速器。V100 Tensor Core GPU 采用 NVIDIA Volta™ 架构，可在单个 GPU 中提供近 32 个 CPU 的性能，助力研究人员攻克以前无法应对的挑战。V100 已在业界首个 AI 基准测试 MLPerf 中拔得头筹，以出色的成绩证明了其是具有巨大可扩展性和通用性的当今世界上强大的计算平台。

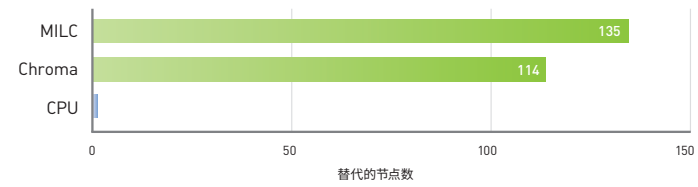
对比 CPU，相同吞吐量训练速度快 32 倍[1]



相比使用 CPU 服务器，推理吞吐量高达 24 倍[2]



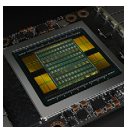
高性能计算 (HPC)：1 个 V100 服务器节点最多可以替代 135 个 CPU 服务器节点[3]



规格

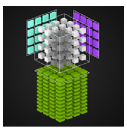
	V100 PCIe	V100 SXM2	V100S PCIe
GPU 架构	NVIDIA Volta		
NVIDIA Tensor 核心数量	640		
NVIDIA CUDA® 核心数量	5120		
双精度浮点运算性能	7 TFLOPS	7.8 TFLOPS	8.2 TFLOPS
单精度浮点运算性能	14 TFLOPS	15.7 TFLOPS	16.4 TFLOPS
Tensor 性能	112 TFLOPS	125 TFLOPS	130 TFLOPS
GPU 显存	32 GB 或 16 GB HBM2		32 GB HBM2
显存带宽	900 GB/s		1134 GB/s
纠错码	支持		
互联带宽	32 GB/s	300 GB/s	32 GB/s
系统接口	PCIe 3.0	NVIDIA NVLink™	PCIe 3.0
外形尺寸	PCIe 全高 / 全长	SXM2	PCIe 全高 / 全长
最大功耗	250 瓦	300 瓦	250 瓦
散热解决方案	被动式		
计算 API	CUDA、DirectCompute、OpenCL™、OpenACC®		

突破性的创新



VOLTA 架构

通过在一个统一架构内搭配使用 CUDA Core 和 Tensor Core，配备 V100 GPU 的单台服务器可以取代数百台通用 CPU 服务器，以处理传统的 HPC 和深度学习工作。

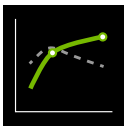


TENSOR CORE 技术

V100 配有 640 个 Tensor Core，可提供 130 teraFLOPS (TFLOPS) 的深度学习性能。与 NVIDIA Pascal™ GPU 相比，可为深度学习训练提供 12 倍张量浮点运算性能，为深度学习推理提供 6 倍张量浮点运算性能。

新一代 NVLINK

V100 中采用的 NVIDIA NVLink 可提供两倍于上一代的吞吐量。八块 V100 加速器能以每秒高达千兆字节（GB/s）的速度互联，从而发挥出单台服务器所能提供的极高应用性能。

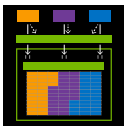


超强节能模式

全新的最大节能模式可允许数据中心在现有的功耗预算内，使每个机架最高提升 40% 的计算能力。在此模式下，V100 以最大处理效率运行时，只需一半的功耗，即可提供高达 80% 的性能。

HBM2 显存

V100 将 900 GB/s 的改良版原始带宽与高达 95% 的 DRAM 利用效率相结合，在 STREAM 上测量时可提供相较 Pascal GPU 高达 1.5 倍的显存带宽。V100 现提供 32 GB 显存配置，比标准的 16 GB 版增加一倍显存空间。



可编程性

V100 架构的设计初衷即是为了简化可编程性。其全新的独立线程调度能力可实现细粒度同步，并能通过在琐碎的工作之间共享资源来提升 GPU 的利用率。

V100 是 NVIDIA 数据中心平台在深度学习、HPC 和图形领域的强大产品。该平台可为 600 余款 HPC 应用和各大深度学习框架提供加速。此平台适用于桌面、服务器以及云服务，不仅能显著提升性能，还能节省成本。

各个深度学习框架

Caffe2

Microsoft Cognitive Toolkit

mxnet

PYTORCH

TensorFlow

theano

600 余款 GPU 加速应用

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

如需详细了解 NVIDIA V100 Tensor Core GPU，
请访问 **https://www.nvidia.cn/data-center/v100/**

1 ResNet-50 训练，数据集：ImageNet2012，批量大小 = 256 | NVIDIA V100 比较数据：NVIDIA DGX-2™ 服务器，1 块 V100 SXM3-32GB 显卡，MXNet 1.5.1，容器 = 19.11-py3，混合精度，吞吐量：1525 张图像 /s | 英特尔比较数据：Supermicro SYS-1029GQ-TRT，单路英特尔至强 Gold 6240 处理器（2GHz/3.9GHz Turbo 频率），Tensorflow 0.18，FP32（唯一可用的精度），吞吐量：48 张图像 /s

2 BERT Base 微调推理，数据集：SQuADv1.1，批量大小 = 1，序列长度 = 128 | NVIDIA V100 比较数据：Supermicro SYS-4029GP-TRT，1 块 V100-PCIe-16GB 显卡，预发布容器，混合精度，NVIDIA TensorRT™ 6.0，吞吐量：557 句 /s | 英特尔比较数据：单路英特尔至强 Gold 6240 处理器（2.6GHz/3.9GHz Turbo 频率），FP32（唯一可用的精度），OpenVINO MKL-DNN v0.18，吞吐量：23.5 句 /s

3 基于 NVIDIA HGX-2™ 的 16 块 V100-SXM2-32GB 显卡 | 应用（数据集）：MILC (APEX Medium) 和 Chroma (szscl21_24_128) | CPU 服务器：双路英特尔至强 Platinum 8280 (Cascade Lake)

NVIDIA