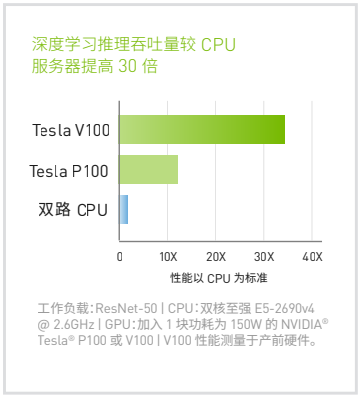
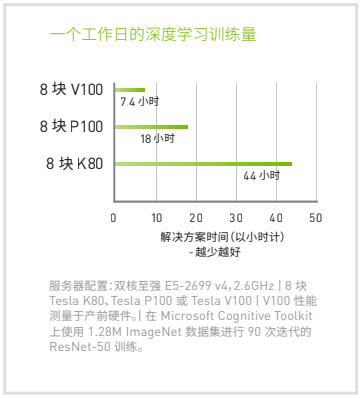
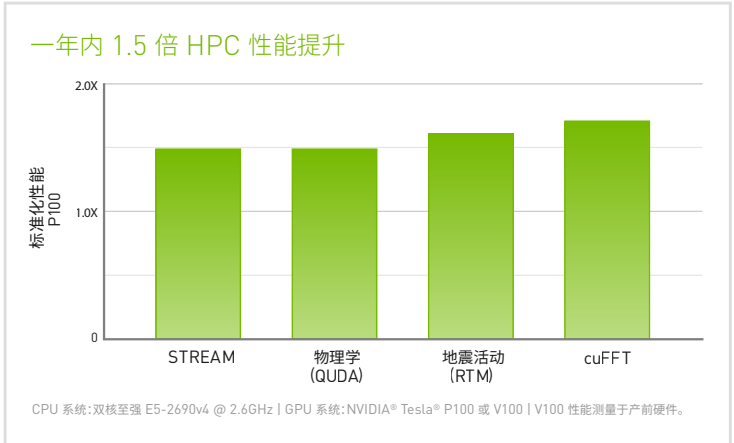
NVIDIA TESLA V100
GPU 加速器

当今市场上数据中心 GPU 中的精尖之作。

NVIDIA® Tesla® V100 是当今市场上为加速人工智能、高性能计算和图形的数据中心 GPU 中的精尖之作。Tesla V100 采用全新一代 NVIDIA Volta™ 架构，可在单个 GPU 中提供高达 100 个 CPU 的性能，助力数据科学家、研究人员和工程师解决以前无法应对的难题。



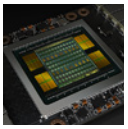




规格

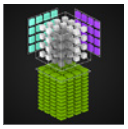

	Tesla V100 PCIe	Tesla V100 SXM2
GPU 架构	NVIDIA Volta	
NVIDIA Tensor 核心数量	640	
NVIDIA CUDA® 核心数量	5,120	
双精度浮点运算能力	7 TFLOPS	7.5 TFLOPS
单精度浮点运算能力	14 TFLOPS	15 TFLOPS
Tensor 性能	112 TFLOPS	120 TFLOPS
GPU 内存	16 GB HBM2	
显存带宽	900 GB/秒	
ECC	是	
互联带宽*	32 GB/秒	300 GB/秒
系统接口	PCIe Gen3	NVIDIA NVLink
外形尺寸	PCIe 全高/全长	SXM2
最大功耗	250 W	300 W
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 内核和 Tensor 内核，配备 Tesla V100 GPU 的单台服务器可以取代数百台通用 CPU 服务器来处理传统的 HPC 和深度学习。

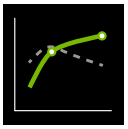


TENSOR 核心

Tesla V100 配有 640 个Tensor 核心，可提供 120 万亿次级的深度学习性能。与 NVIDIA Pascal™ GPU 相比，可为深度学习训练提供 12 倍张量浮点运算能力；为深度学习推理提供 6 倍张量浮点运算能力。

新一代 NVLINK

Tesla V100 中采用的 NVIDIA NVLink 可提供 2 倍于上一代的吞吐量。8 块 Tesla V100 加速器能以高达 300 GB/s 的速度互联，从而发挥出单个服务器所能提供的最高应用性能。

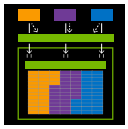


最大节能模式

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，Tesla V100 以最大处理效率运行时，可提供高达 80% 的性能而只需一半的功耗。

HBM2

Tesla 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 上测量时可提供高于 Pascal GPU 1.5 倍的显存带宽。



可编程性

Tesla V100 的架构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进而提升 GPU 的利用率。

Tesla V100 是 Tesla 数据中心计算平台在深度学习、HPC 和图形领域的旗舰产品。Tesla 平台可为 450 余项 HPC 应用程序和各大深度学习框架提供加速。从桌面、服务器到云服务，均可使用此平台，不仅能带来巨额性能收益，还能创造众多成本节约机会。

各深度学习框架

Caffe2

Microsoft Cognitive Toolkit

mxnet

PYTORCH

TensorFlow

theano

450+ GPU 加速应用程序

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

如需详细了解 Tesla V100，请访问 www.nvidia.cn/v100

NVIDIA