

NVIDIA TESLA V100 GPU 加速器

前所未有最先進的資料中心 GPU

NVIDIA® Tesla® V100 是為加快 AI、HPC 及繪圖速度而建立，同時也是前所未有、全球最先進的資料中心 GPU。Tesla V100 搭載最新的 GPU 架構：NVIDIA Volta，在單一 GPU 中提供高達 100 倍 CPU 效能——讓數據科學家、研究人員與工程師能因應先前無法解決的挑戰。

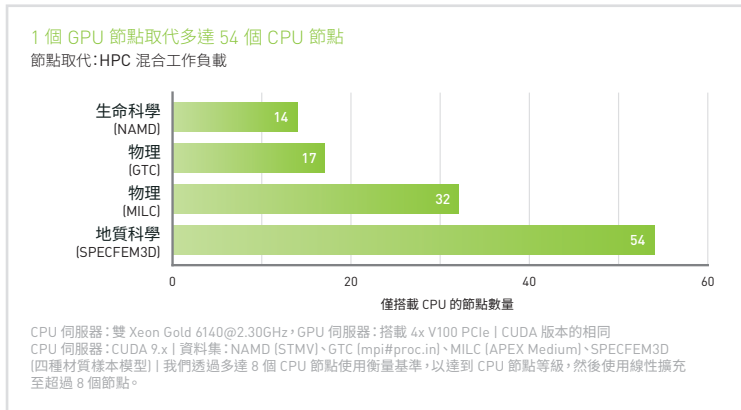
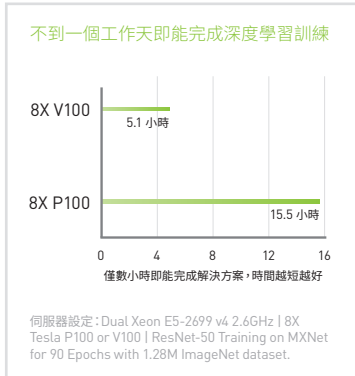
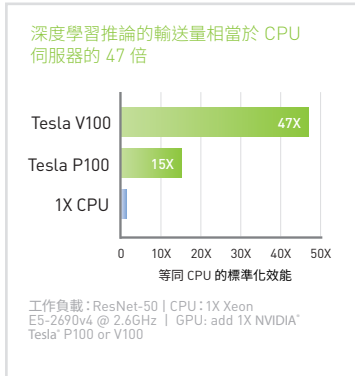
規格



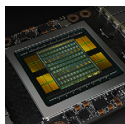
Tesla V100 PCIe

Tesla V100 SXM2

GPU 架構	NVIDIA Volta	
NVIDIA Tensor 核心	640	
NVIDIA CUDA® 核心	5,120	
雙精度浮點運算效能	7 TFLOPS	7.8 TFLOPS
單精度浮點運算效能	14 TFLOPS	15.7 TFLOPS
Tensor 效能	112 TFLOPS	125 TFLOPS
GPU 記憶體	32GB / 16GB HBM2	
記憶體頻寬	900GB/sec	
ECC	支援	
互連頻寬	32GB/sec	300GB/sec
系統介面	PCIe Gen3	NVIDIA NVLink
尺寸	全高/全長型 PCIe	SXM2
最大功耗	250 W	300 W
散熱解決方案	被動式	
運算 API	CUDA、DirectCompute、OpenCL™、OpenACC	

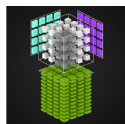


突破性創新



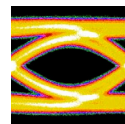
VOLTA 架構

配備 Tesla V100 GPU 的單一伺服器，可在整合架構內，將 CUDA 核心與 Tensor 核心進行配對，取代上百部供傳統 HPC 和深度學習使用的市售 CPU 伺服器。



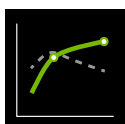
TENSOR 核心

Tesla V100 配備了 640 個 Tensor 核心，可展現 125 teraFLOPS 的深度學習效能。相較於 NVIDIA Pascal™ GPU，深度學習訓練可達 12 倍 Tensor FLOPS，而深度學習推論能力可達 6 倍 Tensor FLOPS。



新一代 NVLINK

相較於舊世代，Tesla V100 的 NVIDIA NVLink 能展現高達 2 倍的輸送量。具備高達 300GB/s 的速度，最多可與八部 Tesla V100 加速器互連，在單一伺服器上展現最高的應用效能。



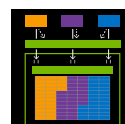
最大效率模式

資料中心可利用全新的最大效率模式，使現有的功率分配，達到更高的每一機架 40% 之運算容量。在此模式中，Tesla V100 是使用峰值處理效率運行，使功耗減半，提供高達 80% 的效能。



HBM2

Tesla V100 可在結合改良 900 GB/s 的原始頻寬和高達 95% 的 DRAM 使用效率後，如同於 STREAM 上測量般，展現相當於 1.5 倍的 Pascal GPU 記憶體頻寬。Tesla V100 現在可支援 32GB 組態，使標準 16GB 產品的記憶體加倍。



編程能力

Tesla V100 從最初即採用精簡編程能力的架構。全新的獨立執行緒排程是以分享小型工作的資源，達到細粒度同步及改善 GPU 使用率。

Tesla V100 是 Tesla 資料中心運算平台用於進行深度學習、HPC 與繪圖的旗艦產品。Tesla 平台能為超過 550 種 HPC 應用項目及各種主要深度學習架構加速，其使用範圍從桌上型電腦到伺服器，再擴展至雲端服務，能提供大幅提升效能與節省成本的機會。

各種深度學習架構

Caffe2 Microsoft Cognitive Toolkit
mxnet PYTORCH
TensorFlow theano

超過 550 種 GPU 加速應用

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

想要進一步瞭解 Tesla V100，請造訪 www.nvidia.com/v100