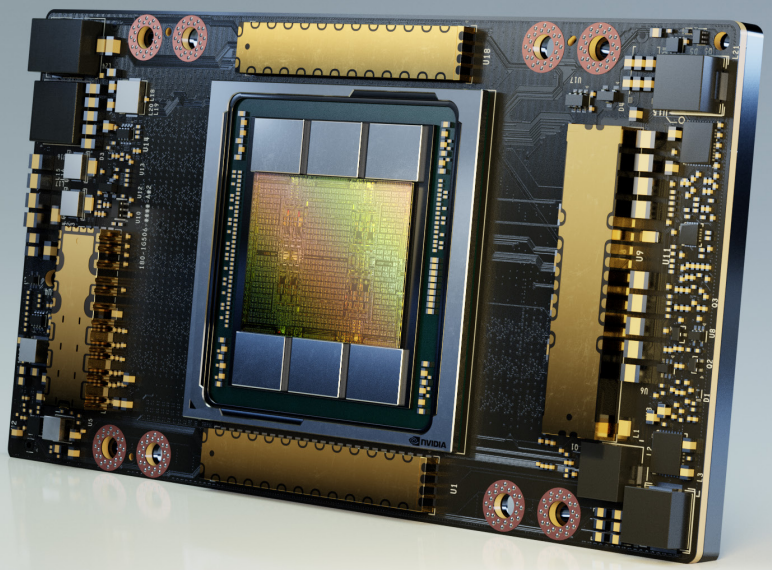
NVIDIA A100 TENSOR 核心 GPU

稱霸各級距的空前規模



適合各種工作負載的最強大運算平台

NVIDIA® A100 Tensor 核心 GPU 在各級距提供無與倫比的加速能力，以針對人工智慧、資料分析與高效運算（HPC）應用程式，強化全球最高效能的彈性資料中心。由於具備 NVIDIA 資料中心平台引擎，A100 可以提供比舊世代 NVIDIA Volta™ 高出 20 倍的高效能。A100 可以藉由提供統一平台的多執行階段 GPU（MIG），有效擴充或分割為 7 個分隔的 GPU 執行階段，以便能調整彈性的資料中心動態，配合不斷變化的工作負載要求。

NVIDIA A100 Tensor 核心技術可以支援廣泛的數學精密度，並為所有的工作負載提供單一加速器。最新一代的 A100 80GB 可以將 GPU 記憶體增加 1 倍，並首度推出每秒 2TB/s 且全球最快的記憶體頻寬，以加速最大模型與最大量資料集合的解決方案。

A100 是完整 NVIDIA 資料中心解決方案的一部分，包含了硬體、網路連線、軟體、程式庫，以及來自 NGC™ 的最佳化人工智慧模型與應用程式的建構基礎。其代表適用於資料中心之最強有力的端對端人工智慧與高效能運算平台，並允許研究人員傳遞即時結果以及在產品中大規模部署解決方案。

系統規格

	NVIDIA A100 適用於 NVLink		NVIDIA A100 適用於 PCIe
Peak FP64	9.7 TF		9.7 TF
Peak FP64 Tensor Core	19.5 TF		19.5 TF
Peak FP32	19.5 TF		19.5 TF
Tensor Float 32（TF32）	156 TF | 312 TF*		156 TF | 312 TF*
Peak BFLOAT16 Tensor Core	312 TF | 624 TF*		312 TF | 624 TF*
Peak FP16 Tensor Core	312 TF | 624 TF*		312 TF | 624 TF*
Peak INT8 Tensor Core	624 TOPS | 1,248 TOPS*		624 TOPS | 1,248 TOPS*
Peak INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*		1,248 TOPS | 2,496 TOPS*
GPU 記憶體	40GB	80GB	40GB
GPU 記憶體頻寬	1,555 GB/s	2,039 GB/s	1,555 GB/s
互連	NVIDIA NVLink 600 GB/s** PCIe Gen4 64 GB/s		NVIDIA NVLink 600 GB/s** PCIe Gen4 64 GB/s
多重執行個體 GPU（MIG）	各執行個體大小最高達 7 MIGs @ 10 GB		各執行個體大小最高達 7 MIGs @ 5 GB
外型規格	NVIDIA HGX™ A100 上的 4/8 SXM		PCIe
最大 TDP 功率	400 W	400 W	250 W

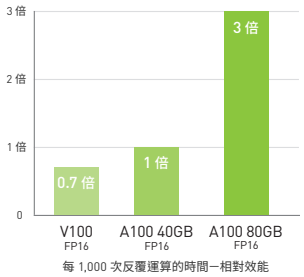
* 有稀疏性

** 經 HGX A100 伺服器板的 SXM GPU、經 NVLink 橋接器的 PCIe GPU 高達 2 GPU

工作負載中驚人的效能表現

在最大模型上的人工智慧訓練
高達 3 倍

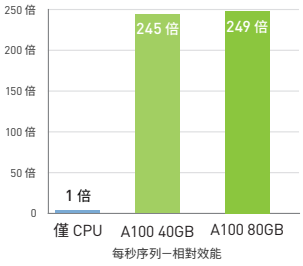
DLRM 訓練



HugeCTR 架構上的 DLRM，精確度 = FP16 | NVIDIA A100 80GB 批次大小 = 48 | NVIDIA A100 40GB 批次大小 = 32 | NVIDIA V100 32GB 批次大小 = 32。

CPU 上的人工智慧推論效能
高達 249 倍

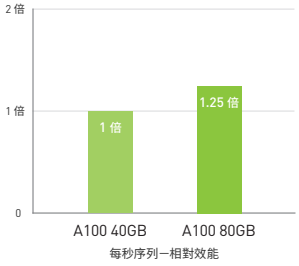
BERT－大量推論



BERT－大量推論 | 僅 CPU：雙 Xeon Gold 6240 @2.60 GHz，精確度 = FP32，批次大小 = 128 | V100：NVIDIA Tensor- RT™（TRT） 7.2，精確度 = INT8，批次大小 = 256 | A100 40GB 和 80GB，批次大小 = 256，精確度 = INT8 有稀疏性。

A100 40GB 上的人工智慧推論效能
高達 1.25 倍

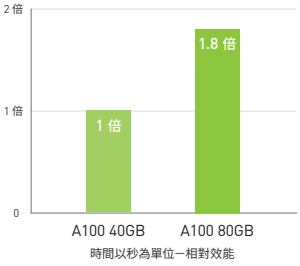
RNN-T 推論：單一串流



MLPerf 0.7 RNN-T 測得含（1/7）MIG 配量。架構：TensorRT 7.2，資料集合 = LibriSpeech，精準度 = FP16。

高效能運算應用程式效能
高達 1.8 倍

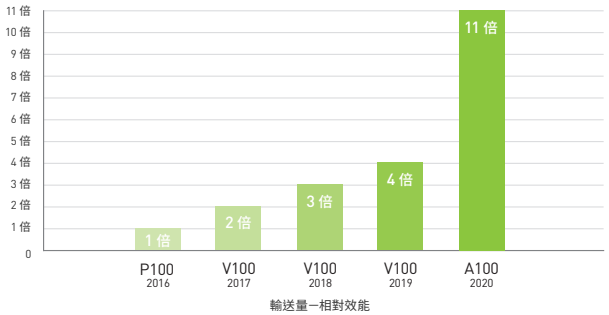
Quantum Espresso



使用 CNT10POR8 資料集合測量Quantum Espresso，精確度 = FP64。

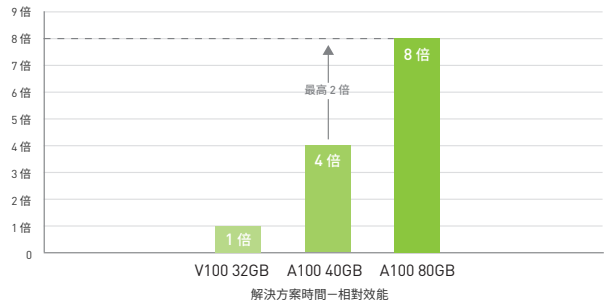
4 年內高效能運算效能成長 11 倍

頂級高效能運算應用程式傳輸量



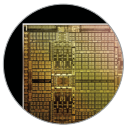
應用程式加速比的幾何平均數對比P100：基準測試應用程式：Amber [PME-Cellulose_NVE]、Chroma [szscl21_24_128]、GROMACS [ADH Dodec]、MILC [Apex Medium]、NAMD [stmv_nve_cuda]、PyTorch (BERT-Large Fine Tuner]、Quantum Espresso [AUSURF112-jR]；Random Forest FP32 [make_blobs (160000 x 64：10)]、TensorFlow [ResNet-50]、VASP 6 [Si Huge] | GPU 節點含雙插槽 CPU 及 4x NVIDIA P100、V100 或 A100 GPU。

比巨量資料分析基準測試的 A100 40GB 快 2 倍



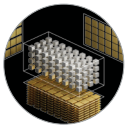
巨量資料分析基準測試 | GPU-BDB 是源自於 TPCx-BB 基準測試，且使用於內部效能試驗。GPU-BDB 的結果無法與 TPCx-BB 比較 | 30 分析零售查詢，10TB 資料集合上的 ETL、ML、NLP | V100 32GB，RAPIDS™/Dask | A100 40GB 與 A100 80GB，RAPIDS/Dask/BlazingSQL

突破性創新



NVIDIA AMPERE 架構

無論是使用 MIG，將 A100 GPU 分割為更小的執行個體，或使用 NVIDIA NVLink®，將多部 GPU 連接至快速大規模工作負載，A100 都能迅速掌握從最小工作到最大多節點工作負載之不同大小的加速需求。A100 的全面性，代表 IT 經理可以全天候最大化資料中心每一部 GPU 的效用。

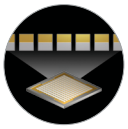


第三代 TENSOR 核心

NVIDIA A100 提供 312 teraFLOPS（TFLOPS）的深度學習效能。相較於 NVIDIA Volta GPU，Tensor FLOPS 深度學習訓練高出 20 倍，而 Tensor TOPS 深度學習推論也高出 20 倍。

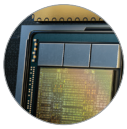
新一代 NVLINK

相較於舊世代，A100 中的 NVIDIA NVLink 可以提供高 2 倍的輸送量。當與 NVIDIA NVSwitch™ 結合時，可以於每秒高達 600GB，連接達 16 部 A100 GPU，釋放單一伺服器上的最高應用程式效能。NVLink 可以經由 HGX A100 伺服器板，使用於 A100 SXM GPU 中，而經由 NVLink 橋接器，可以在 PCIe GPU 中使用多達 2 個 GPU。



多重執行個體 GPU（MIG）

一個 A100 GPU 可以分割為多達 7 個 GPU 執行個體，在含高頻寬記憶體、快取和運算核心的硬體層級下完全分隔。多重執行個體 GPU（MIG）可以為開發人員提供存取適合所有應用程式的突破性加速能力，而 IT 管理員可以為所有工作提供適當大小的 GPU 加速，以最佳化使用性及擴大存取所有的使用者與應用程式。



HBM2E

在搭配高達 80GB 的高頻寬記憶體（HBM2e）之後，A100 可以提供每秒超過 2TB 全球第一的 GPU 記憶體頻寬，以及高達 95% 的動態隨機存取記憶體（DRAM）使用效率。A100 提供的記憶體頻寬比舊版本高出 1.7 倍。

結構稀疏性

人工智慧網路擁有數百萬到數十億組參數。在精準預測時不需要所有的這些參數，其中的部分參數可以轉換為零，以確保模型稀疏性，且不會損及精度。A100 中的 Tensor 核心可以為稀疏性模型提供高達 2 倍的效能。儘管稀疏性特色可以很快使人工智慧推論受益，但是也可以改善模型訓練的效能。

NVIDIA A100 Tensor 核心 GPU 是 NVIDIA 資料中心平台
用於深度學習、高效能運算（HPC）和資料分析的旗艦產品。
此平台可以為超過 1,800 個應用程式加速，包括所有主要
深度學習的架構。A100 可以隨處使用，從桌上型電腦到
伺服器，再到雲端服務，不僅能提升動態效能，且能增加
節省成本的機會。

所有深度學習架構

1800+ GPU 加速應用程式

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF

想要進一步了解 NVIDIA A100 Tensor 核心 GPU，請造訪 **www.nvidia.com/a100**

ZERONE 零壹科技 NVIDIA