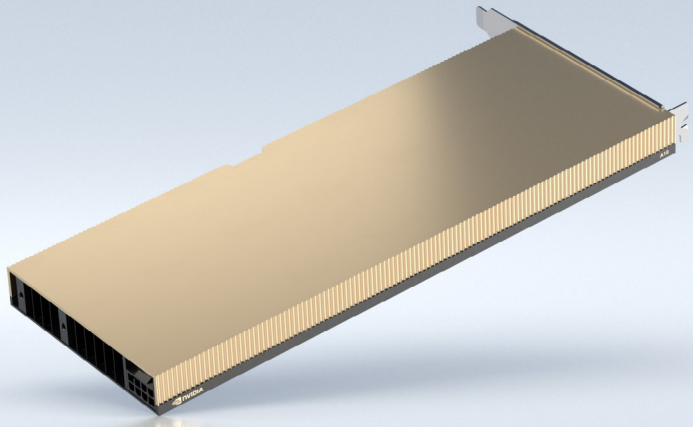
NVIDIA A10

以適用於主流企業伺服器的人工智慧提供加速繪圖和視訊

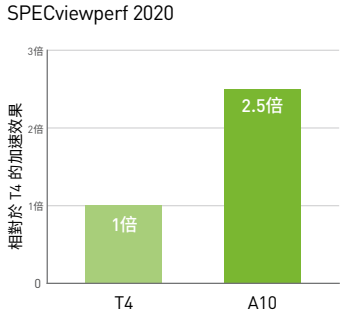


利用強大的人工智慧豐富繪圖和視訊應用程式

NVIDIA A10 Tensor 核心 GPU 與 NVIDIA RTX 虛擬工作站 (vWS) 軟體結合，將具有人工智慧服務的主流繪圖和視訊帶入主流企業伺服器，為設計師、工程師、創作者和科學家提供克服現今挑戰需要的解決方案。A10 以最新 NVIDIA Ampere 架構為基礎，將第二代 RT 核心、第三代 Tensor 核心以及具有 24 GB GDDR6 記憶體的新型串流微處理器，全部結合在 150W 功率包絡中，以提供多用途繪圖、渲染、人工智慧以及運算效能。從可隨處存取的虛擬工作站，到將節點渲染至執行各種工作負載的資料中心，A10 皆以單寬、全高、全長 PCIe 外型規格提供最佳效能。

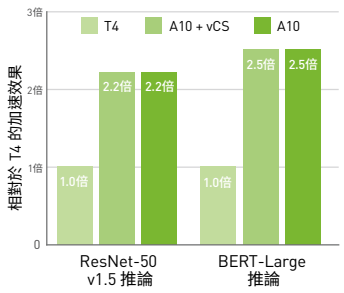
NVIDIA A10 是 NVIDIA-Certified Systems™ 的一部分，在內部部署的資料中心、雲端以及邊緣皆可支援。NVIDIA A10 是以來自 NVIDIA NGC™ 目錄的人工智慧框架、CUDA-X™ 函式庫、超過 230 位開發人員，以及 1,800 多款 GPU 最佳化應用程式的豐富生態系統為基礎，可以協助企業解決最關鍵的業務挑戰。

A10 提供比 T4 快 2.5 倍的虛擬工作站效能[1]



A10 提供的推論效能比 T4 高 2.5 倍[2]

NVIDIA 虛擬運算伺服器 (vCS) 可提供與裸機相當的效能



規格

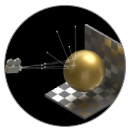
FP32	**31.2 TF**
TF32 Tensor 核心	**62.5 TF | 125 TF***
BFLOAT16 Tensor 核心	**125 TF | 250 TF***
FP16 Tensor 核心	**125 TF | 250 TF***
INT8 Tensor 核心	**250 TOPS | 500 TOPS***
INT4 Tensor 核心	**500 TOPS | 1000 TOPS***
RT 核心	**72**
編碼/解碼	**1 個編碼器 1 個解碼器 (+AV1 解碼)**
GPU 記憶體	**24 GB GDDR6**
GPU 記憶體頻寬	**600 GB/s**
互連	**PCIe Gen4: 64 GB/s**
外型規格	**單插槽 FHFL**
最大 TDP 功率	**150W**
虛擬化 GPU 軟體支援	**NVIDIA 虛擬 PC (vPC)/ 虛擬應用程式 (vApp)、NVIDIA RTX™ 虛擬工作站 (vWS)、NVIDIA 虛擬運算伺服器 (vCS)**
透過硬體信任根進行安全和測量開機	**有**
NEBS 就緒	**Level 3**
電源接頭	**PEX 8**

*有稀疏性

NVIDIA Ampere 架構概覽

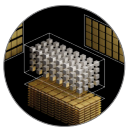
NVIDIA AMPERE 架構 CUDA 核心

單精度浮點 (FP32) 運算的雙速處理以及更高的功率效率，可以明顯提升繪圖和運算工作流程的效能，例如複雜的 3D 電腦輔助設計 (CAD) 和電腦輔助工程 (CAE)。



第二代 RT 核心

第二代 RT 核心的傳輸量比上一代高 2 倍，且能同時執行光線追蹤與著色或去雜訊功能，使工作負載大幅加速，例如渲染電影內容的真實感、評估建築設計，以及製作產品設計的虛擬原型。此技術同時加快了光線追蹤動態模糊的渲染速度，能以更高的視覺準確性，更快速地獲得結果。



第三代 TENSOR 核心

Tensor Float 32 (TF32) 精度提供比上一代高 5 倍的訓練傳輸量，以加快人工智慧和資料科學模型訓練，而無須變更任何程式碼。對於結構稀疏性的硬體支援，可以將推論傳輸量提高至兩倍。Tensor 核心也具備深度學習超級取樣 (DLSS)、人工智慧去雜訊、針對特定應用程式的強化編輯等功能，可以將人工智慧帶入繪圖中。

24 GB GDDR6

超快速 GDDR6 記憶體，提供 600 GB/s 的頻寬，適用於渲染、資料科學、工程模擬，以及其他 GPU 記憶體密集型工作負載。

第四代 PCIE EXPRESS

第四代 PCI Express 的頻寬是第三代 PCIe 的 2 倍，提高了 CPU 記憶體的資料傳輸速度，適用於人工智慧、資料科學、3D 設計等資料密集型任務。更快的 PCIe 效能，同時加快了 GPU 直接記憶體存取 (DMA) 傳輸速度，使 GPU 與支援 **NVIDIA GPUDirect® for video**之裝置間的視訊資料能更快速地輸入/輸出通訊，為直播提供強大的解決方案。A10 也回溯與第三代 PCI Express 相容，以確保部署的靈活性。.

資料中心效率和安全性

NVIDIA A10 採用單插槽全高、全長高功率效率設計，能與來自全球 OEM 的各種伺服器相容。NVIDIA A10 可以透過硬體信任根技術，進行安全和測量開機，確保韌體不會遭到竄改或毀損。

NVIDIA A10 Tensor 核心 GPU 是結合人工智慧之主流繪圖與視訊的理想選擇。第二代 RT 核心和第三代 Tensor 核心以強大的人工智慧，採用適合主流伺服器的 150W TDP，豐富繪圖和視訊應用程式。

NVIDIA A10 同時與 NVIDIA 虛擬化 GPU (vGPU) 軟體結合，透過易於管理、安全、靈活及可擴充，以適應資源需求的基礎架構，加快眾多資料中心的工作負載，從繪圖豐富的虛擬桌面基礎架構 (VDI) 到高效能虛擬工作站 (vWS)，再到人工智慧。

所有的深度學習框架

mxnet

PYTORCH

Apache Spark

TensorFlow

RTX 適用於專業應用程式

AUTODESK® REVIT®

CATIA

SOLIDWORKS

creo®

Rhinoceros® design, model, present, analyze, realize...

SIEMENS

欲深入瞭解 NVIDIA A10 Tensor 核心 GPU，請造訪 **www.nvidia.com/a10**

1 在搭載 2x Xeon Gold 6154 3.0GHz (3.7GHz Turbo)、NVIDIA RTX vWS 軟體 VMware ESXi 7 U2、主機/客體驅動程式 461.33 的伺服器上執行測試。| SPECviewperf 2020 Subtest，以及 HD 3dsmax-07 複合。

2 BERT Large 推論 NVIDIA TensorRT7.2，序列長度 =128，批次大小 =128，NGC 容器：21.02-py3 | ResNet-50 v1.5：NVIDIA TensorRT7.2，INT8 精度批次大小 = 128 NGC 容器：20.12-py3 | NVIDIA A10 搭載虛擬運算伺服器軟體、VMware ESXi 7 U2 主機/客體驅動程式 461.33