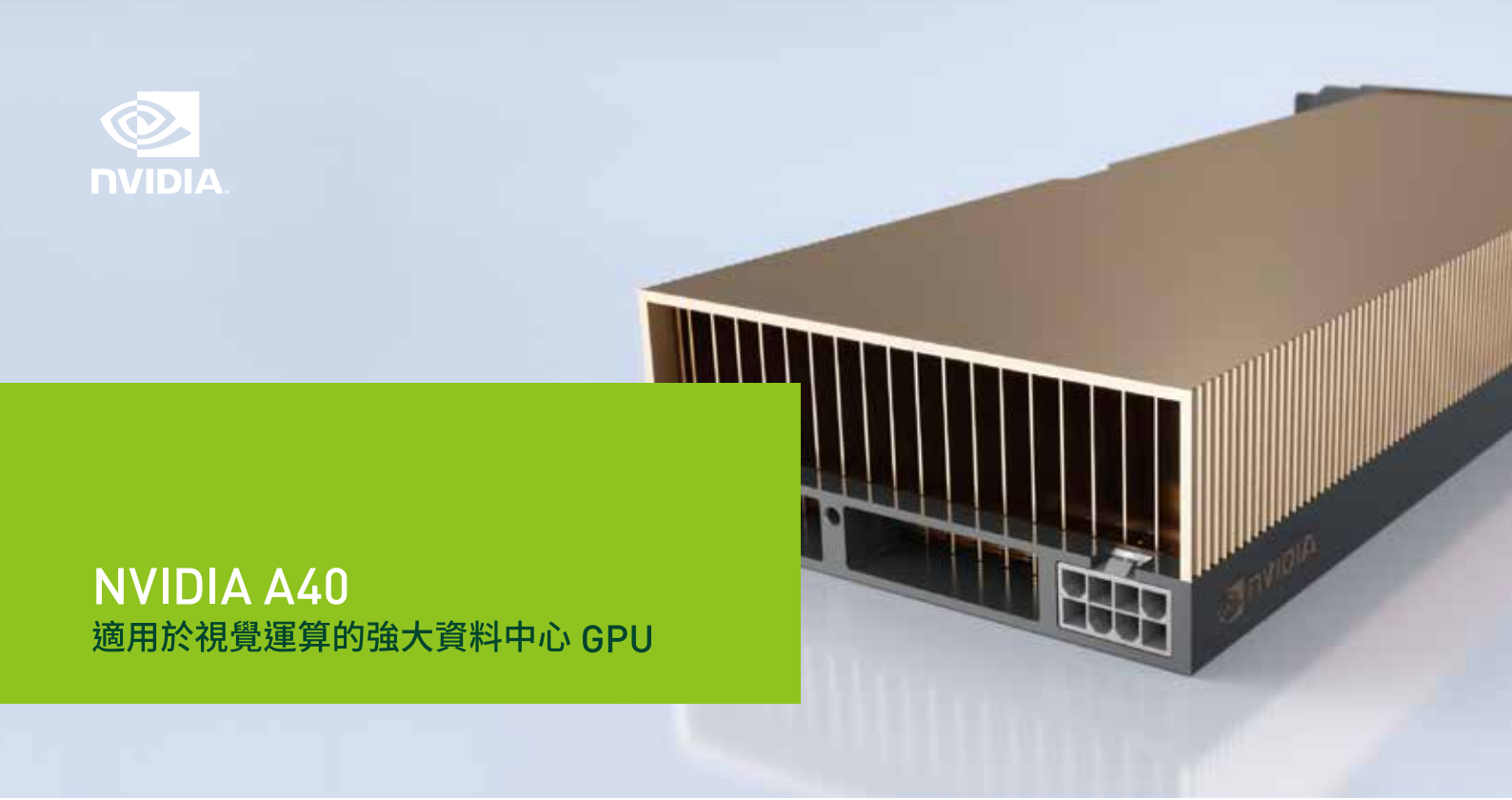
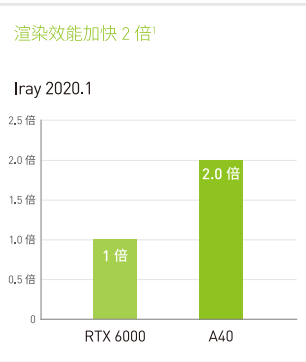


NVIDIA A40

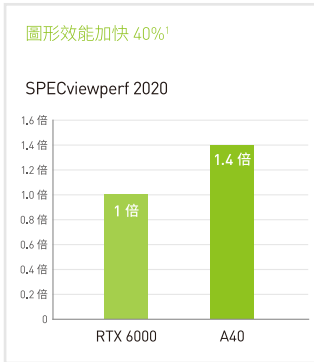
適用於視覺運算的強大資料中心 GPU

NVIDIA A40 將最新 NVIDIA Ampere 架構 RT 核心、Tensor 核心、CUDA® 核心與 48 GB 的圖形記憶體結合，以加快資料中心的高需求視覺運算工作負載。從隨處存取強大的虛擬工作站，到專用的渲染節點，NVIDIA A40 將次世代 NVIDIA RTX™ 技術帶入資料中心，以支援最先進的專業視覺化工作負載。

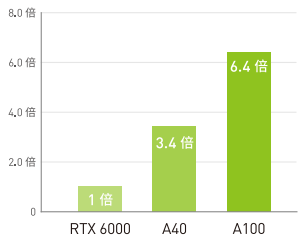
渲染效能加快 2 倍[1]



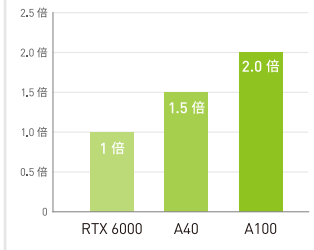
圖形效能加快 40%[1]



人工智慧訓練效能加快 3 倍[2]



單精度 (FP32) 高效能運算效能加快 50%[2]



規格

GPU 架構	**NVIDIA Ampere 架構**
GPU 記憶體	**48 GB GDDR6 含 ECC**
記憶體頻寬	**696 GB/s**
互連介面	**NVIDIA® NVLink® 112.5 GB/s（雙向）[3] PCIe Gen4 31.5 GB/s（雙向）**
以 NVIDIA Ampere 架構為基礎的 CUDA 核心	**10,752**
NVIDIA 第二代 RT 核心	**84**
NVIDIA 第三代 Tensor 核心	**336**
峰值 FP32 TFLOPS（非 Tensor）	**37.4**
FP16 累計的峰值 FP16 Tensor TFLOPS	**149.7 | 299.4***
峰值 TF32 Tensor TFLOPS	**74.8 | 149.6***
RT 核心效能 TFLOPS	**73.1**
FP32 累計的峰值 BF16 Tensor TFLOPS	**149.7 | 299.4***
峰值 INT8 Tensor TOPS 峰值 INT 4 Tensor TOPS	**299.3 | 598.6*** **598.7 | 1,197.4***
尺寸	**4.4"（H）x 10.5"（L）雙槽**
顯示連接埠	**3x DisplayPort 1.4**、支援 NVIDIA Mosaic 和 Quadro® Sync[4]**
最大功耗	**300 W**
電源接頭	**8-pin CPU**
散熱解決方案	**被動式**
虛擬化 GPU（vGPU）軟體支援	**NVIDIA 虛擬 PC / 虛擬應用程式、NVIDIA RTX 虛擬工作站、NVIDIA 虛擬運算伺服器**
支援的 vGPU 設定檔	**請參閱虛擬化 GPU 授權指南**
NVENC | NVDEC	**1x | 2x（包含 AV1 解碼）**
透過硬體信任根進行安全與測量開機	**有**
NEBS 就緒	**Level 3**
運算 API	**CUDA、DirectCompute、OpenCL™、OpenACC®**
圖形 API	**DirectX 12.07[5]、Shader Model 5.17[5]、OpenGL 4.68[4]、Vulkan 1.18[4]**
MIG 支援	**無**

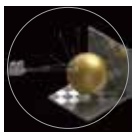
* 啟用結構稀疏

** 預設將 A40 配置用於虛擬化，並停用實體顯示器接頭。可以透過管理軟體工具，啟用顯示器輸出。

探索 NVIDIA Ampere 架構

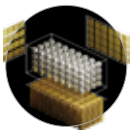
NVIDIA AMPERE 架構 CUDA 核心

以雙倍速度處理單精度浮點（FP32）運算，並改善功率效率，大幅提升圖形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 核心的傳輸量為上一代的 2 倍，且能同時執行光線追蹤、著色與去雜訊功能，以大幅加快電影內容的真實感渲染、架構設計評估、製作產品設計虛擬原型等工作負載。此技術同時加快了光線追蹤動態模糊的渲染速度，進而能以更高的視覺準確性，更快獲得結果。



第三代 TENSOR 核心

Tensor Float 32（TF32）精度提供之訓練傳輸量為上一代的 5 倍，可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。對結構稀疏的硬體支援，提供 2 倍的推論傳輸量。Tensor 核心同時將人工智慧帶入圖形，以提供深度學習超級取樣（**DLSS**）、人工智慧去雜訊、強化編輯特定應用程式等功能。

搭載 NVLINK 的 48 GB GDDR6 記憶體

超快速的 GDDR6 記憶體，可以透過 **NVLink**[3] 擴充至 96 GB，為資料科學家、工程師和創作者提供處理龐大資料集和工作負載（例如資料科學和模擬）所需要的大型記憶體。

第四代 PCIE EXPRESS

第四代 PCIE Express 將第三代 PCIe 的頻寬加倍，提高了 CPU 記憶體的資料傳輸速度，以滿足人工智慧、資料科學、3D 設計等資料密集型的任務。更快的 PCIe 效能可同時加快 GPU 直接記憶體存取（DMA）傳輸，在 GPU 與搭載 **GPUDirect® for Video** 的裝置之間提供更快的視訊資料輸入 / 輸出通訊，造就強大的直播解決方案。A40 向下相容於第三代 PCI Express，確保了部署的靈活性。

資料中心效率和安全性

NVIDIA A40 採用雙插槽的高效率設計，其功率效率為上一代的 2 倍，並與全球 OEM 的各種伺服器相容。NVIDIA A40 透過硬體信任根技術進行安全與測量開機，確保韌體不會遭到竄改或毀損。

NVIDIA A40 GPU 提供最先進的視覺運算功能，包括即時光線追蹤、人工智慧加速和多工作負載靈活性，以加快深度學習、資料科學以及運算工作負載。搭載 NVIDIA A40、NVIDIA RTX 虛擬工作站（vWS）和 NVIDIA 虛擬運算伺服器軟體的虛擬工作站，受益於各種產業應用程式和專業軟體的廣泛測試，可確保最佳效能和穩定性。

各種深度學習架構

mxnet
PYTORCH
Apache Spark
TensorFlow

RTX 可適用於專業應用程式

Adobe Premiere Pro
SOLIDWORKS
PLM Software SIEMENS NX
AUTODESK ARNOLD
REDSHIFT
AUTODESK VRED
KeyShot
UNREAL ENGINE
blender
octanerender
V-RAY

欲深入瞭解 NVIDIA A40 GPU，請造訪 **www.nvidia.com/a40**

1 渲染和圖形測試是在 2x Xeon Gold 6126 2.6GHz（3.7GHz Turbo）上執行。256GB 系統記憶體。NVIDIA 驅動程式 461.09。渲染測試：Iray 2020.1，NVIDIA Endeavor 場景的渲染時間。圖形測試：SPECviewperf 2020 Subtest、4K medical-03 Composite | 2 AI 和 HPC 測試，是在 AMD EPYC 7742@2.25GHz（3.4GHz Turbo）上執行。512GB 系統記憶體。NVIDIA 驅動程式 460.14。AI 訓練：BERT 預先訓練傳輸量。PyTorch（2/3）第 1 階段和（1/3）第 2 階段。精度 FP32 適用於 RTX 6000，TF32 適用於 A40 和 A100。第 1 階段的序列長度 = 128。第 2 階段 = 512。單精度 HPC：NAMD 版本 3.0a7，stmv_nve_cuda、精度 =FP32、ns/ 天，CUDA 版本：11.1.74 | 3 連接兩張 NVIDIA A40 卡與 NVLink，可以將效能和記憶體容量擴充至 96 GB，若應用程式可支援 NVLink 技術。請聯繫應用程式供應商，以確認是否支援 NVLink。| 4 Quadro Sync II 卡另售。在 Windows 10 和 Linux 上支援 Mosaic。| 5 GPU 支援 DX 12.0 API，硬體功能層級 12 + 1。| 6 產品以公布的 Khronos 規格為準，預計在上市時通過 Khronos 符合性測試流程。在 www.krhonos.org/conformance 上可以找到目前的符合性狀態

ZERONE 零壹科技

NVIDIA