NVIDIA 虛擬化運算伺服器
以虛擬化 GPU 驅動運算最密集的工作負載

改變虛擬化運算

隨著資料中心的 GPU 伺服器數量增加，IT 管理員希望從 VMware、Red Hat、Nutanix、Citrix 等標準伺服器虛擬化平台管理伺服器。根據 Gartner 研究顯示，以虛擬管理器為基礎的伺服器虛擬化已相當成熟，許多中大型企業將其 80-90% 的伺服器工作負載在虛擬機器（VM）上執行。[1] 然而，這類傳統的利用虛擬管理器虛擬化的資料中心基礎架構僅限於純 CPU 伺服器，但 VDI 除外。因此，執行 AI、深度學習和高效能運算（HPC）工作負載的 GPU 加速的伺服器通常與資料中心隔離，限制利用率、靈活性和可管理性。

NVIDIA® vComputeServer 能夠將以虛擬機管理器為基礎之伺服器虛擬化的效益帶給搭載 GPU 加速的伺服器。資料中心管理員現在能夠執行任何在虛擬機器（VM）中需要 GPU 的運算密集型工作負載。

vComputeServer 軟體將 NVIDIA GPU 虛擬化以加快運算密集型工作負載，包括超過 600 個用於 AI、深度學習、資料科學和 HPC 的 GPU 加速應用程式。透過 GPU 共用，單一 GPU 可驅動多個 VM，將利用率和可負擔性最大化，或者可由多個虛擬 GPU 驅動單一 VM，使運算最密集的工作負載成為可能。由於支援所有主要的虛擬機管理器虛擬化平台，資料中心管理員可以使用相同的管理工具管理 GPU 加速的伺服器。

針對運算的授權

不像 NVIDIA® GRID® vPC/vApp 和 Quadro® vDWS 是以同時使用者（CCU）為單位授權的用戶端運算產品，vComputeServer 是以 GPU 為單位授權的 1 年期訂閱，包括 NVIDIA 企業支援。這讓多個 VM 中的多個運算工作負載可在單一 GPU 上執行，將資源利用率和 ROI 最大化。

透過 NGC 軟體針對容器而最佳化

vComputeServer 支援用於深度學習、機器學習和 HPC 的 NVIDIA NGC GPU 最佳化軟體。NGC 軟體包括經過 NVIDIA 調整、測試及最佳化的主要 AI 和資料科學軟體容器，以及經過全面測試的 HPC 應用和資料分析容器。

NGC 也為針對 NVIDIA Tensor 核心 GPU 最佳化的各種常見 AI 任務提供經過預先訓練的模型，並包括逐步說明和指令碼，建立具有樣本效能和準確度指標的深度學習模型。NGC 讓資料科學家、開發人員和研究人員在虛擬化環境中降低部署時間和專案複雜性，以便專注於建構解決方案、收集洞見和創造業務價值。

特色

- **GPU 效能** – 在虛擬化環境中存取最強大的 GPU。
- **管理和監控** – 利用以虛擬機管理器為基礎的工具簡化資料中心可管理性。
- **即時移轉** – 即時移轉 GPU 加速 VM 而不中斷，簡化維護和升級。
- **將利用率最大化** – 透過 GPU 共用和多 GPU 聚合提高利用率和生產力。
- **安全性** – 擴展伺服器虛擬化的優勢至 GPU 工作負載。
- **多租戶** – 隔離工作負載並安全地支援多個使用者。
- **快速部署** – 快速部署用於 AI、資料科學和 HPC 的 GPU 最佳化 NGC 容器。
- **可靠性** – 錯誤修正代碼（ECC）和動態頁面退役可防止資料毀損。
- **企業軟體支援** – NVIDIA 企業和 NVIDIA NGC 支援服務提供全面支援。

NVIDIA VCOMPUTESERVER 特色清單

配置和部署	
GPU 共用（分化）	✓
GPU 聚合（多 vGPU）	✓
點對點 NVLink	✓
ECC 和動態頁面退役	✓
Linux 作業系統支援	✓
Windows 作業系統支援	X
NVIDIA 運算驅動程式	✓
NVIDIA 顯示卡驅動程式	X
NVIDIA Quadro 驅動程式	X
服務品質排程	✓

資料中心管理	
主機、客機和應用程式層級監控	✓
即時移轉	✓
支援	
NVIDIA 直接企業級技術支援	✓
維護版本、瑕疵解決和安全性修補最長可達三年 [2]	✓
NGC 支援服務 [3]	✓

vCOMPUTESERVER 規格

支援的最大畫格緩衝區	48GB
支援的最小畫格緩衝區	4GB
最大多租戶	8:1
可用規格	4C、6C、8C、12C、16C、24C[4]、32C[5]、48C[6]

VCOMPUTESERVER 建議使用的 GPU

	NVIDIA T4	NVIDIA V100（SXM2）
RT 核心	48	-
Tensor 核心	320	640
CUDA® 核心	2,560	5,120
記憶體	16 GB GDDR6	32 GB HBM2
FP 16/FP 32（混合精度）	64 TFLOPS	125 TFLOPS
FP 32（單精度）	8.1 TFLOPS	15.7 TFLOPS
FP 64（雙精度）	-	7.8 TFLOPS
NVLink：每個 VM 的 GPU 數量	-	最多 8 個
ECC 和頁面退役	✓	✓
每個 VM 的多 GPU	最多 16 個	最多 16 個

其他支援的 GPU

NVIDIA® Quadro RTX™ 6000、RTX 8000、NVIDIA P40、P100 和 P6 刀鋒型規格。

[1] 資料來自 Gartner 的「Market Guide for Server Virtualization」，2019 年 4 月，ID G00350674。
[2] 適用於有效的支援、更新與維護（SUM）合約。
[3] 不包含在 vComputeServer 授權中，但透過 NVIDIA NGC 支援服務合作夥伴另外提供。
[4] Quadro RTX 6000 和 RTX 8000 提供 24C 規格。
[5] NVIDIA V100 提供 32C 規格。
[6] Quadro RTX 8000 支援 48C 規格。