HPE Compute for AI

Leading your enterprise AI into the future

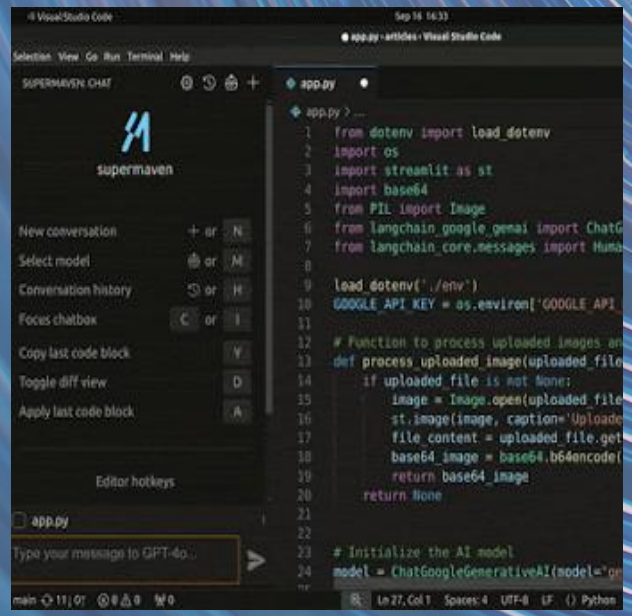
黃光平 **Compute Specialist**

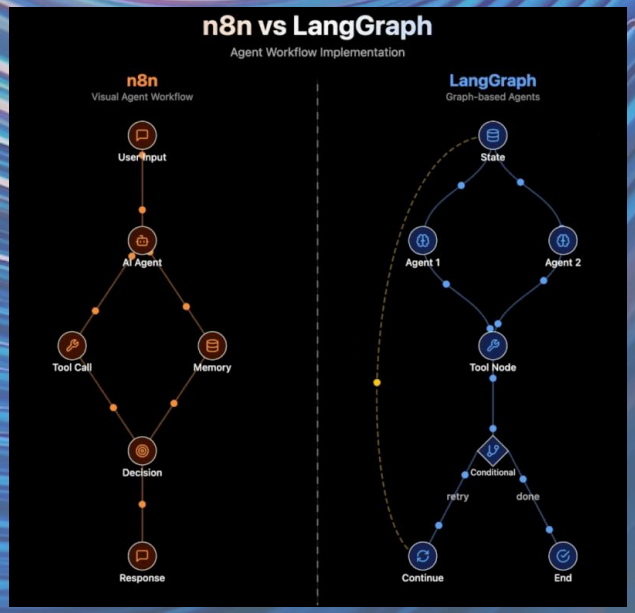
Jan. 2026

“人工智慧將成為改變世界的奇異點”

有可能以我們一生中任何科技都無法比擬的規模改變我們的生活和工作方式

AI驅動轉型的新時代





Generative AI

Agentic AI

Physical AI

人工智慧使用案例遍佈各行各業

金融

詐欺偵測
個人化銀行服務
投資洞察

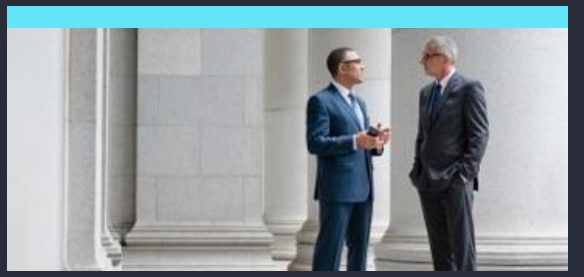
製造

工廠模擬
產品設計
預測性維護

醫療保健

護理交班
分子模擬
藥物研發

政府



文件摘要
審計合規
AI虛擬助手

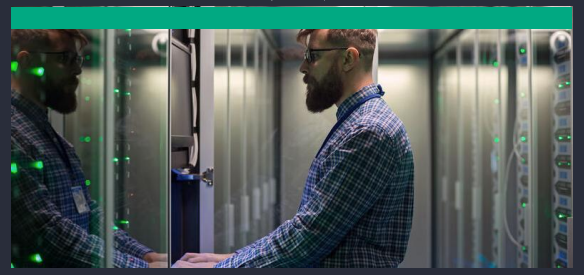
零售

防損管理
自動化盤點
客戶體驗分析

娛樂媒體

角色塑造
影片編輯與影像創作
數位資產管理

IT維運



網路安全
基礎設施優化
AIOps

能源

預測性維護
知識庫問答
供應鏈分析

企業成功導入人工智慧的關鍵要素

資料品質與可存取性
確保資料準確、完整且易於取得，是推動 AI 成效的基礎

明確的應用場景願景
需具備清晰的策略方向，以辨識並優先推動具影響力的 AI 應用

可擴展且安全的基礎架構
建構具彈性與高安全性的技術環境，以支援 AI 解決方案的長期發展

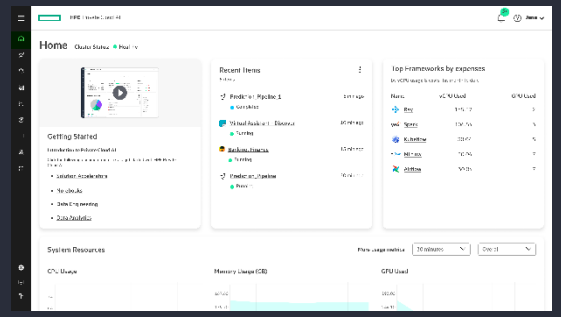
專業人才與跨部門協作
結合技術專才與業務團隊，促進知識整合與創新落地

NVIDIA AI Computing by HPE

Enterprise Grade AI

AI Optimized
Compute and Storage

Engineered Systems



Turnkey Private Cloud AI Solutions

Model Building : Research : Service Providers

HPC and Super Computing

AI Services | AI Software Platform | Data Management | Hybrid Cloud Platform

企業發展 AI 應用場景現況

	AI 專案 1	AI 專案 2	AI 專案 3	AI 專案 N...
應用層	聊天機器人	資料萃取, 搜尋	圖片/文字辨識與生成	語音/影像處理
MLOpS/ LLMOps	監控、微調及驗證	監控、微調及驗證	監控、微調及驗證	監控、微調及驗證
Model	大語言模型	大語言模型	大語言模型	大語言模型
AI 平台	AI 自建 Pipeline	AI 自建 Pipeline	AI 自建 Pipeline	AI 自建 Pipeline
資料層	企業資料來源	企業資料來源	企業資料來源	企業資料來源
硬體層	伺服器，儲存，GPU	伺服器，儲存，GPU	伺服器，儲存，GPU	伺服器，儲存，GPU

統一資源與模型運營，以支持應用服務快速部署，最佳化資源使用率

應用層

聊天機器人

資料萃取, 搜尋

圖片/文字辨識與生成

...

語音/影像處理

MLOpS/
LLMOps

監控、微調及驗證

Model

大語言模型整合

AI 平台

Private Cloud AI Platform

資料層

企業資料儲存與串接

硬體層

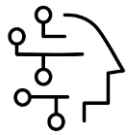
伺服器，儲存，GPU

結合台灣大語言模型與解決方案

HPE Gen AI Adoption Service

LMM App






LLM ISV providers
- 開源的大語言模型
- LLM 整解決方案
- Low Code/No Code Platform

HPE Private Cloud AI

HPE AI Essential


NVIDIA AI Enterprise


Unified Analytics

AI Optimized Infrastructure

Private Cloud Control Plane

AI Optimized Server and Switches

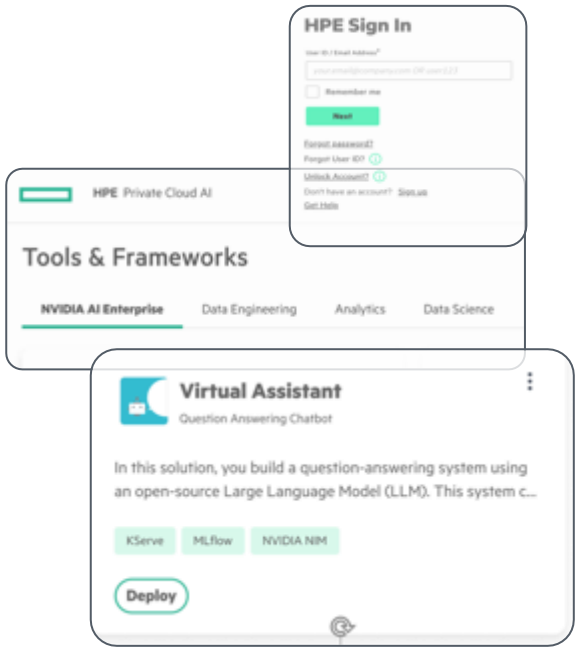
HPE GreenLake Cloud Platform

遵循開源架構
持續、快速的創新
廣泛的 Evergreen 工具
人才取得容易

企業級的整合
持續資安弱點改善
運行軌跡稽核

NVIDA 軟體框架
簡化生成式AI 應用程式的開發和部署

HPE Private Cloud AI 為AI 算力中心的核心基礎架構-四種配置規格



	AI Sandbox	Inferencing	Inferencing + RAG	Inferencing + RAG + Fine-tuning
	All-in-one Node	All-in-one Node	All-in-one Rack	Scalable by Rack
				
	Developer System (1 x DL380a Gen11)	Small (1 x DL380a Gen12)	Medium (2 x DL380a Gen12)	Large (2x DL380a Gen12)
Compute	2 NVIDIA H100 NVL GPU's	4 NVIDIA RTX 6000 GPUs	8 NVIDIA H200NVL GPUs	16 NVIDIA H200NVL GPUs
Storage	32 TB Integrated	32 TB integrated in node	109 TB file storage in rack	217TB file storage in rack
Networking	Customer Network	Customer Network	400GbE NVIDIA Networking	400GbE NVIDIA Networking
Power	Up to 2.2 kW	up to 2 or 4 kW	13 KW per rack	17 KW per rack

Optional 8~16 GPU expansion racks for Small, Med & Large

Unified experience through HPE GreenLake cloud

重新定義你的AI算力中心

Data Center

Edge

GB300 NVL72
GB200 NVL72

GB200 NVL72
GB300 NVL72
By HPE

模型建置、研發與
服務供應商

HGX B300

HPE Compute XD

大型 AI 模型的訓練、
微調與推論

H200 NVL

DL380a Gen12

DL385 Gen11

企業級推論、微調與
混合式工作負載

RTX Pro 6000

DL380a Gen12

DL380 Gen12

DL385 Gen11

企業級推論、微調與
混合式工作負載

GH200 NVL2

DL384 Gen12

記憶體密集型企業級
AI 與推論

L40S
L4

DL320 Gen12
DL325 Gen12

DL145 Gen11

視覺型 AI
邊緣推論

從模型建構到部署，HPE提供全方位AI運算解決方案

科學發現 | 個人化服務 | 內容創作生成 | 數位孿生 | 詐欺偵測 | 事件管理 | 虛擬助手 | 藥物研發 | 文件摘要 | 知識庫問答

Training
模型建構

Fine tuning
模型客製化

Inference
模型部屬

hours, days, weeks

minutes, hours, days

milliseconds, microseconds, seconds

HPE ProLiant Compute XD

HPE ProLiant Compute DL

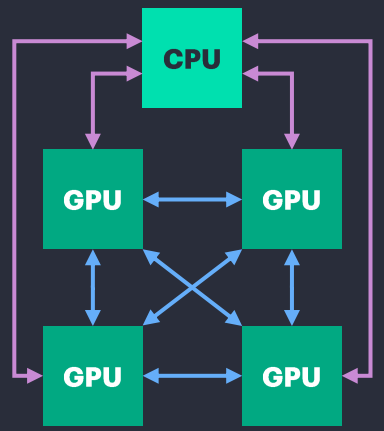
GPU Cluster 對應到企業在模型workload選擇，會有不同的算力配置

應用需求
- 選擇模型：模型核心能力
- 目標任務：推論 vs. 調優

效能需求
- 模型大小與類型
- 企業Fine-Tuning 策略

Training
模型建構

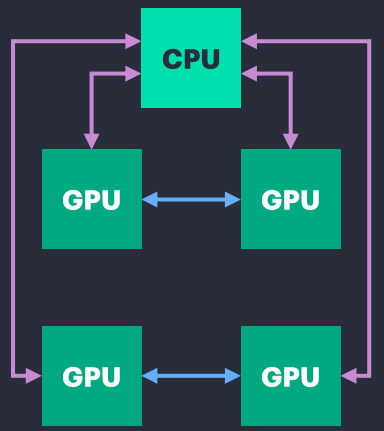
Modular SXM



GPU 間資料處理的效能最大化
透過NVLink switch 將主機上所有GPU做點對點的連接

Fine tuning
模型客製化

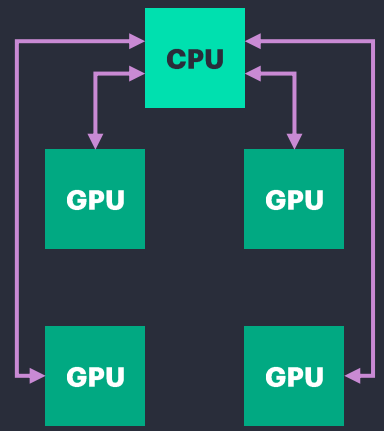
GPU-GPU Bridging



加速 GPU 間資料處理的效能
透過NVLink bridge將主機上的GPU 2-way或4-way點對點的連接

Inference
模型部屬

PCIe



適合推論與一般中小模型使用
透過PCIe做為主機上GPU和 GPU間資料處理的溝通橋樑此架構

HPE Compute for AI 精準對應不同階段的 AI 發展需求

Core — AI Applications — Edge

Training 模型建構	Fine tuning 模型客製化	Inference 模型部屬
8-way GPU	*4-way & 2-way GPU*	*2-way & 1-way GPU*
 HPE Cray XD	 HPE ProLiant Compute DL380a Gen12	 HPE ProLiant DL380a Gen11 DL380 Gen12 / DL385 Gen11
 HPE ProLiant Compute XD	 HPE ProLiant Compute DL384 Gen12	 HPE ProLiant Compute DL320 Gen12 / DL325 Gen12

HPE ProLiant Compute XD — **HPE ProLiant Compute DL**

AI Builder — AI Consumer

HPE Compute for AI 精準對應不同階段的 AI 發展需求

8-way GPU servers for accelerated AI

Training
模型建構

針對大型 **AI** 模型訓練與微調所打造的高效能平台

適用於 **LLM** 開發者與 **AI** 服務供應商

8-way GPU

HPE Cray XD670

8x NVIDIA H200

5U Air-cooled and liquid-cooled

HPE ProLiant Compute XD685

8x NVIDIA H200

8x NVIDIA B200 DLC

6U air-cooled or 5U liquid-cooled

HPE Compute XD690

8x NVIDIA B300

10U air-cooled

HPE Compute for AI 精準對應不同階段的 AI 發展需求

4-way & 2-way GPU servers for accelerated AI

Fine tuning
模型客製化

企業級 **AI** 的超大規模 **GPU** 加速

適用於需要靈活擴展 **GenAI** 工作負載的 **LLM** 使用者

針對高記憶體需求 **AI** 的效能

適用於需微調大型模型或進行 **RAG** 的 **LLM** 使用者

4-way & 2-way GPU

HPE ProLiant Compute DL380a Gen12

Up to **10x** NVIDIA H200 NVL GPUs
Up to **10x** NVIDIA RTX Pro 6000 GPUs

HPE ProLiant Compute DL384 Gen12

Dual NVIDIA GH200 **NVL2**

HPE Compute for AI 精準對應不同階段的 AI 發展需求

2-way & 1-way GPU servers for accelerated AI

Inference
模型部屬

可擴展且高效能的 **AI** 推論平台

驅動強化大型語言模型的推論應用

邊緣端電腦視覺 **AI**

專為邊緣 **AI** 打造

2-way & 1-way GPU

HPE ProLiant DL380a Gen11 / DL385 Gen11

Up to **4x** NVIDIA H100 NVL GPUs
*Up to **2x** NVIDIA H200 NVL GPUs
*Up to **2x** NVIDIA RTX Pro 6000 GPUs

*Only DL385 Gen11 can support

HPE ProLiant DL380 Gen12

Up to **3x** NVIDIA RTX Pro 6000 GPUs

HPE ProLiant Compute DL320 Gen12 / DL325 Gen12

Up to **2x** NVIDIA L40S GPUs
Up to **4x** NVIDIA L4 GPUs

“小而美
讓企業AI更靈活”

從小型應用開始，逐步推動企業AI全面落地

Thank You