NVIDIA GPU 產品線

Hopper 到 Blackwell

20251128 Allen Sun

🧠 NVIDIA 業界慣用語分層整理

層級	常用稱呼	範例	說明
消費級 GPU	GeForce RTX	RTX 5090 / RTX 5080	個人玩家、創作者、AI 開發入門
專業工作站 GPU	RTX Professional	RTX 6000 Ada / RTX PRO 6000	專業繪圖、工程設計、AI 開發環境
資料中心 GPU	NVIDIA Data Center GPU或簡稱「A/H/B 系列 GPU」	A100 / H100 / B200	AI 訓練、推理、大模型服務
GPU 模組平台	HGX	HGX H100 / HGX B200	伺服器用 GPU 主板（SXM 版本）
整機系統 / 超算平台	DGX / GB200 / NVL72	DGX H100 / GB200 NVL72	完整 AI 超算解決方案

功率區別 (TDP)：SXM 版本（HGX）的功耗通常遠高於 PCIe 版本（例如 H100 SXM 是 700W，PCIe 則是 300-350W），B200 SXM: **~1000W** (比 H100 高 40%)。

🧱 實際業界溝通中常見的用語分層

常見說法	實際含義	對應技術層級
RTX 泛指工作站或消費級 GPU	RTX 5090 / RTX PRO 6000 等	PCIe Slot
PCIe GPU	指標準主機板插槽 GPU	RTX 系列 A100 PCIe H100 PCIe B200 PCIe
HGX 平台	一塊裝多顆 SXM GPU 的伺服器模組	HGX H100 HGX B200
DGX 系統	NVIDIA 自家整機伺服器 （含 CPU、儲存、網路）	DGX H100 DGX B200
NVL72 / GB200 叢集	超大規模 AI 系統 （多 DGX 結構）	36 CPU: Grace 72 GPU: Blackwell

📊 業內實際講法舉例（工程師之間的口語）

語句	含意
這台伺服器插兩張 RTX 6000 Ada	工作站 / 小型 AI 推理用機，PCIe 介面
我們訓練在 H100 HGX 上跑	使用 SXM + NVLink 平台進行分散式訓練
我們在跑 8卡 A100	常見的 DGX A100 配置
客戶要上 B200 PCIe	想在一般伺服器主板上做推理 / 加速
那是 DGX H100 機櫃	整機 NVIDIA 官方系統
GB200 NVL72 cluster	指整合 Grace CPU + Blackwell GPU 的大型叢集架構

🧭 歸納成一句話：

- 「RTX」是工作站與消費級的統稱
- 「A/H/B 系列 (Data Center GPU)」是伺服器級 GPU 的統稱
- 「HGX」是伺服器 GPU 主板
- 「DGX」是完整伺服器系統
- 「SXM GPU」才有支援 NVLink，「PCIe GPU」不支援 NVLink
- 在 Hopper 與 Blackwell 世代，僅 SXM 版本支援高速 NVLink 互連

🔬 核心規格快速對照

項目	A100	H100	B200
架構	Ampere	Hopper	Blackwell
製程	7nm	4nm	4nm (TSMC)
記憶體	40/80GB HBM2e	80GB HBM3	192GB HBM3e
FP8 效能	-	~2000 TFLOPS	~4500 TFLOPS
NVLink	600GB/s	900GB/s	1800GB/s
典型價格(USD)	~$15K	~$30K	~$40K (預估)

🎯 如何選擇GPU?

- 訓練 **GPT-4** 等級模型 → GB200 NVL72
- 訓練 **7B-70B** 模型 → HGX H100 (8卡)
- **LLM** 推理服務 → H100/B200 PCIe
- 微調 **/ RAG** 應用 → A100 或 RTX 6000 Ada
- 開發測試 → RTX 4090 / 5090

🚀 未來展望

掌握了這些核心概念與術語層次後，未來在接續研究 Rubin 架構及其後的技術路線圖時，將能更加輕鬆上手，快速理解 NVIDIA 在 AI 運算領域的演進藍圖。