GPUtw.ai

GPUtw.ai　算力方案簡介

高可用 AI 推論與訓練算力方案

以 RTX PRO 6000 、三座台灣資料中心與 ISO 27001 資安管理，
為您的團隊打造穩定、可擴充、資料不出境的 AI 基礎設施

簡報人

林永傑 Apple Lin
GPUtw.ai 硬體部門主管 Head of Hardware

林澈 Luke Lin
GPUtw.ai 共同創辦人暨執行長 Co-Founder & CEO

2026 年 11 月

常見需求與挑戰

GPUtw.ai

24/7 線上推論
長期 GPU 伺服器

- 30B 級 LLM 推論（vLLM）
- 翻譯、摘要等專用模型
- 語音辨識（離線與即時串流）

不定期模型訓練
On-demand 批次工作

- Fine-tuning 與模型訓練
- Container 環境即可
- H100 / H200 等級算力

資安與合規
視同自有地端環境

- 需符合 ISO 27001 等規範
- 資料涉及營業秘密或智財權
- 資料不外流、不作為訓練資料

常見痛點

共用或公有雲 GPU 資源無預警中斷，多層高可用設計使管理成本大增；高階 GPU 缺貨、價格飆漲，補助或專案經費到期後需尋求穩定可靠的付費方案。

GPUtw.ai： 台灣在地的 GPU 算力夥伴

GPUtw.ai

3 座

台灣中型資料中心

跨機房分散部署

> 99 %

服務可用性 SLA

UPS 不斷電保護

< 2 分鐘

環境部署時間

Container 即開即用

100 %

資料留在台灣

台灣機房、不出境

三種交付模式，依工作負載彈性組合

雲端 Container

按小時計費，適合不定期訓練與實驗

裸金屬獨享

整機獨享、無虛擬化損耗，適合 24/7 推論

實機落地租用

設備部署於客户機房，含安裝、保固與維修

主力算力： NVIDIA RTX PRO 6000 ，現貨充足

GPUtw.ai

NVIDIA RTX PRO 6000 Blackwell

96 GB

每卡 GDDR7 顯示記憶體

可大量供應

已有大量 RTX PRO 6000 上線服務中；以雙卡節點為單位，可依需求擴充節點數

兩種版本，依效能與能效選擇

	Workstation 版	Max-Q 版
功耗	600 W	300 W
相對算力	較 Max-Q 高約 20%	基準
顯示記憶體	96 GB	96 GB
適合情境	追求最大吞吐量	成本效益與能效優先

另有 2× RTX 5090 方案

頻寬與運算約為單張 PRO 6000 的兩倍，但每卡 32 GB ， 30B 級模型需改用 NVFP4 量化。

模型適配範例：每張卡即一個獨立服務單位

GPUtw.ai

模型	精度	權重估計	單卡 96 GB 部署	建議部署方式
gemma4:31b	FP8 / KV BF16	約 31 GB	✓ 可，約 60 GB 留給 KV cache	每卡一個 vLLM 副本，跨機房分散
translategemma:27b	FP8	約 27 GB	✓ 可	獨立副本，與主模型分開擴充
Whisper （離線）	FP16	< 5 GB	✓ 可	批次轉錄，可與其他服務共卡
Whisper streaming	FP16	< 5 GB	✓ 可	獨立卡確保即時低延遲

不需要跨卡 tensor parallel

單一模型即可放進一張 96 GB 卡，每張卡都是可獨立替換的服務單位：任一卡、任一節點、甚至整座機房離線，其餘副本照常服務。

註：權重大小為估算值，實際記憶體用量與吞吐量以雙方 PoC 實測為準。

高可用架構： UPS 不斷電 + 三機房分散部署

GPUtw.ai



負載平衡可由 GPUtw.ai 代管，或由貴單位團隊自行掌控；每節點 2 張 RTX PRO 6000 ， 4 節點共 8 卡；各機房節點數依流量配置。

配置範例： 4 個雙卡節點（共 8 卡），分散三機房

GPUtw.ai

	傳統： 4× H200 VM	建議： 4 節點 ×2 卡
總顯示記憶體	564 GB	**768 GB**
同時運行 31B 副本	4 個	**8 個**
部署位置	單一機構、 2 台 VM	**3 座資料中心、 4 個獨立節點**
穩定性	單點故障、無跨機房備援	**> 99% SLA 、 UPS 、跨機房備援**
供應狀況	全球缺貨、價格飆漲	**現貨充足，可快速擴充**

註： H200 為每卡 141 GB ； 31B 副本數以每卡一個副本計算。

關於 H200

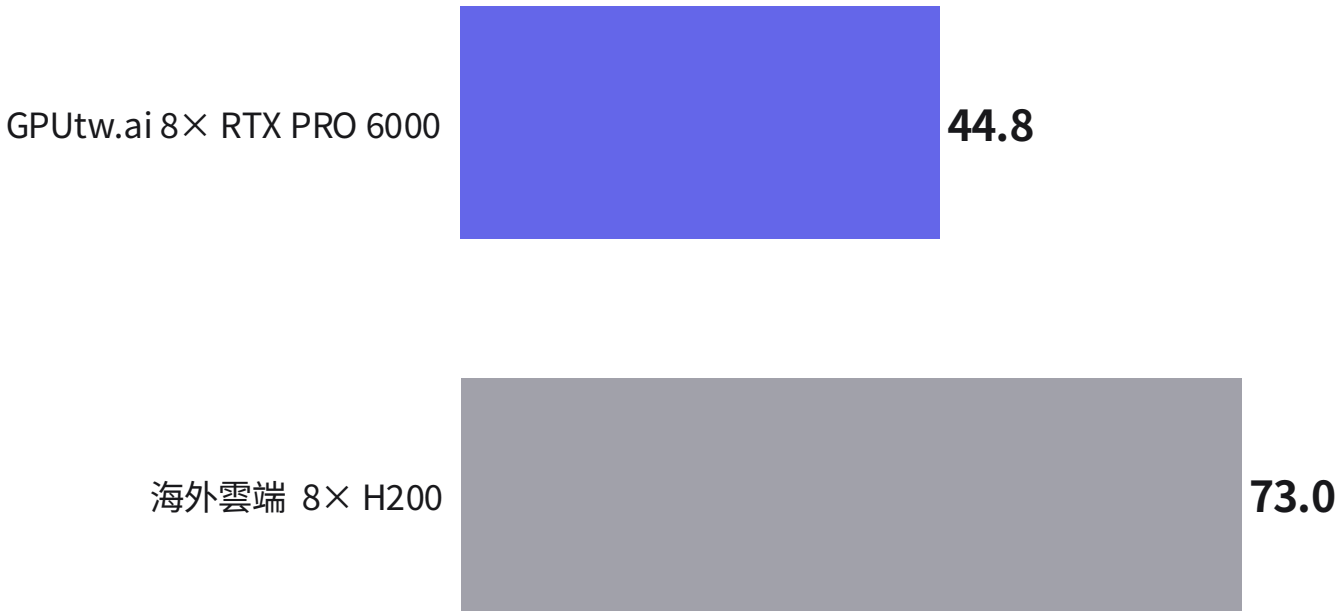
H200 等級算力全球需求極高、供應有限，價格與交期都不穩定。

建議以 RTX PRO 6000 承接 24/7 推論服務；訓練工作可使用按小時計費的 RTX PRO 6000 雲端環境。

成本效益： RTX PRO 6000 月費較 H200 省約 4 成

GPUtw.ai

8 卡月費比較（NT$ 萬／月；GPUtw.ai 為 4 節點 ×2 卡）



約 39% 月費節省

同樣 8 卡、8 個 31B 推論副本

NT$5.6 萬 每卡／月

海外雲端 H200 約 NT$9.1 萬

NT$583 每 GB 顯存／月

海外雲端 H200 約 NT$647

註：海外雲端 8× H200 月費約 NT$73 萬為參考估算（依匯率與方案而異）；GPUtw.ai 以單卡裸金屬牌價 NT$56,000 × 8（4 節點 ×2 卡）計。

同卡比價： RTX PRO 6000 較 GCP 低 34–47%

GPUtw.ai

	GCP 隨用隨付	GCP 1 年承諾	GPUtw.ai 按月	GPUtw.ai 年約
每卡月費	NT$10.5 萬	NT$7.3 萬	**NT$5.6 萬**	**NT$5.6 萬** 平均 NT$4.8 萬
8 卡月費	NT$84.1 萬	NT$58.0 萬	**NT$44.8 萬**	**NT$44.8 萬** 平均 NT$38.5 萬
每卡時價	US$4.50	US$3.11	**約 US$2.40**	**約 US$2.40** 平均 US$2.06
合約綁定	無綁約	綁約 12 個月	**按月預付**	**年約送 60 天**
交付型態	雲端 VM	雲端 VM	**裸金屬獨享**	**裸金屬獨享**

與 GCP G4 同卡比價

同一張 RTX PRO 6000 （96 GB），GPUtw.ai 按月 NT$56,000；年約加贈 60 天，平均約 NT$4.8 萬／月：

按月較隨用隨付低約 47%
年約較 1 年承諾低約 34%

年約平均每月再降約 14%；資料留在台灣、整機獨享。

註： GCP 為官網 g4-standard-48 （1 GPU、 48 vCPU、 180 GiB）牌價， 2026/11 查詢、未稅；以每月 730 小時、 US$1 = NT$32 換算。 GPUtw.ai 年約月費仍為 NT$56,000，平均月費 = NT$56,000 × 365 ÷ 425 天（含加贈 60 天）。

資料安全與隱私：視同自有地端環境

GPUtw.ai

ISO 27001

資訊安全管理標準

依循 ISO 27001 標準營運，可配合貴單位 ISMS 要求，共同擬定保密、個資與智財合約條款。

資料不出台灣

所有運算與儲存皆在台灣資料中心

單一租户獨享

裸金屬整機專用，不與其他客户共用

傳輸與儲存加密

資料於傳輸與儲存過程皆加密保護

不讀取、不作訓練

合約明訂不複製、不揭露、不作任何利用

存取需事先同意

維運存取須經同意，可要求客户人員陪同

退租資料抹除

租期結束後抹除儲存裝置並出具書面證明

合作實績

GPUtw.ai

信任 GPUtw.ai 的學研機構與企業（包括但不限於）

國立臺灣大學 National Taiwan University

國立清華大學 NATIONAL TSING HUA UNIVERSITY

國立陽明交通大學 NATIONAL YANG MING CHIAO TUNG UNIVERSITY

國立成功大學 National Cheng Kung University

中央研究院 ACADEMIA SINICA

國立臺灣科技大學 NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

TAIPEI TECH 國立臺北科技大學 National Taipei University of Technology

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

國立中興大學 NATIONAL CHUNG HSING UNIVERSITY

國立中正大學

國立東華大學 NATIONAL DONG HWA UNIVERSITY

國立聯合大學 NATIONAL UNITED UNIVERSITY

國立暨南國際大學 National Chi Nan University

工業技術研究院 Industrial Technology Research Institute

MEDIATEK

案例

國立大學研究實驗室：實機落地租用

- RTX PRO 6000 96 GB × 2 工作站， ECC 記憶體，到場安裝與驗收
- 租期內專人保固維修，回廠維修期間提供替代設備或遠端運算節點
- 價格保證：租期內不因零組件漲價調整租金；研究資料智財權歸客戶所有

方案與參考價格

GPUtw.ai

24/7 推論

裸金屬獨享

NT$56,000

每卡／月（含稅）

雙卡節點 NT$112,000 ／月

- 整機獨享、預裝 Ubuntu
- 按月預付，開立統編發票
- 長約加贈使用天數

不定期訓練

雲端 Container

約 NT$63

每卡／小時

依即時供應浮動

- 按實際用量計費
- 2 分鐘內完成部署
- Container 環境即開即用

地端部署

實機落地租用

另行報價

依規格與租期

雙卡工作站

- 設備部署於客户機房
- 含運送、安裝、保固維修
- 租期內價格保證

裸金屬長約加贈天數

滿 3 個月	滿 6 個月	滿 9 個月	滿 12 個月
+10 天	+25 天	+40 天	+60 天

以上為參考牌價。正式報價將於技術會議後提供，並可配合貴單位採購與評選標流程。

GPUtw.ai

下一步

01

技術會議

確認模型、 VRAM 、流量與上線時程，說明機房與網路架構

02

PoC 測試

提供測試點數，以您的目標模型實測吞吐量與延遲

03

合約與資安

依 ISO 27001 擬定保密、個資與智財條款

04

正式上線

依您的時程完成切換，與現有服務無縫接軌

聯絡窗口

林永傑 Apple Lin　硬體部門主管（網路、硬體與 GPU 部署）

林澈 Luke Lin　共同創辦人暨執行長 Co-Founder & CEO