

## 地端 LLM 導入系列報導之四：採購建置篇

# 硬體到底要怎麼選，才能「不花冤枉錢」？

在企業導入地端大型語言模型時，CIO 最常問的問題其實只有一句話：「我們建內部 LLM 系統，到底要買什麼樣的硬體，才能確保軟體真的跑得動，又能撐得住未來使用量？」

文／許旭安（未來巢科技董事長）

針對當前最熱門，想要在企業內部建置 LLM 系統，該怎樣規劃預算，會有哪些層面需要考量，這個問題看似單純，其實牽涉到四個層次的判斷：需求、模型、架構、治理。

### 你到底想讓 LLM 做什麼？

硬體選錯的根本原因，不是算力不夠，而是需求沒釐清。

在挑硬體之前，請先把這幾個問題問自己：

#### 1. 你要跑的是哪種應用？

- 如果只是內部搜尋、問答、摘要，重點在即時性與穩定性。
- 如果要結合多系統資料（如 ERP、CRM、知識庫），就需要更大的記憶體與儲存吞吐量。

#### 2. 要服務多少人？

- 不同部門、不同時間段的同時查詢量，才是規劃基準。
- 不需預測精確數字，只要界定「單機能支撐的尖峰狀況」與「何時要擴充」即可。

#### 3. 機器會是 LLM 專用，還是要共用？

若 GPU 也會被其他應用（如影像分析、數據視覺化、其他 AI 工具）使用，就要事先區分資源分配與調度權限。特別是在同時有多個應用並行（共現性高）的情境下，即使是小模型，也可能需要預留比模型本身更大的 GPU 記憶體作為快取空間。因此，預

先規劃 GPU 的使用與排程，是確保系統穩定運作的關鍵。

### 模型要「剛好會用」

過去企業容易以為，「模型越大、答案越好」。但實務上，模型規模應該根據應用情境和實際需求精準配對，而非一味追求最大。

首先，要先釐清業務問題本身的複雜度與所需語意深度。

- 大多數企業內部的知識問答、文件彙整、報表生成等場景，所需的模型水準還有準確率要求都不同，需要先評估怎樣可滿足合理場域需求。
- 在選擇模型時，還需注意商業授權與開源條款，不同模型家族（如 LLaMA、Qwen、Mistral 等）即使是同品牌不同尺寸，其授權條款、商業使用及再分發權限都可能不同。
- 建議以「足夠支撐需求、容易持續維護、資料治理能跟上」為原則，用「以目標回應品質與體驗反推所需模型」，再經小規模實測驗證，而不是盲目上最大型號。

### 硬體要從「整體架構」來思考

#### || GPU 是主角，但不是全部

它決定模型能不能載入、能不能在合理時間回

應，但效能的瓶頸往往在別處——記憶體不夠、磁碟太慢、CPU 處理 Token 時塞車，都是常見原因，而開發到一半才發現 GPU 的 VRAM 的不足更是常實務上發生的問題。

## || CPU 與記憶體的角色

每一次提問、搜尋、文字轉換，其實都先經過 CPU。

若 CPU 核心數太少、記憶體頻寬不足，就算 GPU 閒著，整體也跑不快。解法是用批次方式一次處理多筆資料，並把檢索、格式化等步驟分開同時進行，讓整體流程更有效率。

## || 儲存與 I/O

當模型需從知識庫即時擷取內容 (RAG 應用)，磁碟速度與資料通道會成為瓶頸。建議使用高速固態硬碟 (SSD/NVMe)，並讓資料與模型在同一台機器或低延遲網路下運行。

## || 網路與互連

多 GPU 或多台主機之間的傳輸，速度決定效能上限。互連規格 (如 NVLink、InfiniBand、百 G 等級乙太網) 要在規劃時一起評估。否則即使 GPU 再快，通訊延遲也會吃掉所有優勢。

## 性能以目標回應速度反推架構

很多公司在買機器時會問：「這樣的規格可以跑幾個人？」

事實上，沒有人能用表格回答這題。不同模型、上下文長度、提示內容都會讓結果差數倍。

正確做法是反過來：

1. 先定義「可接受的回應時間」與「使用高峰狀況」；
  2. 然後用你公司的真實語料，在幾個不同推論框架 (如 vLLM、TensorRT-LLM) 上測試；
  3. 觀察在那個目標延遲下，單機能穩定支撐多少請求；
  4. 再反推需要多少節點或硬體等級。
- 這樣，你買的不是「聽說能跑」，而是「確定能撐」。

## 用更聰明的方法榨出效能

想在有限硬體上支援更多人？

可以考慮量化 (Quantization) 與模型壓縮。這能降低記憶體占用、提高推論速度，但要注意：速度變快不代表答案不變。建議先用關鍵任務做對照，確認精度仍在可接受範圍，再全面採用。這樣既省資源，又不犧牲品質。

## 別讓伺服器成為「耗電怪獸」

LLM 伺服器不像一般資訊系統，可以長時間高負載運作。電力容量、冷卻能力、機櫃空間，都需要事前盤點。有些企業在硬體還沒上線前，冷氣就已經不夠用。因此：

- 預先計算整體耗能與散熱冗餘；
- 若電力有限，考慮地端+雲端混合部署，讓尖峰負載由雲端承接。

這樣能平衡投資，也避免營運風險。

## 沒有監控，就沒有優化

AI 硬體投資不是一次性。長期的維運、擴充、升級也都需要考慮。無論買多好硬體，若沒監控，就等於開車沒儀表板。以下需要上線後持續追蹤：

- 回應時長 (p95 response time)
- 字元 (Tokens) 處理速率 (吞吐量)
- 從輸入到輸出第一個字元的等待時間 (TTFT, Time To First Token)
- GPU/CPU/記憶體使用率
- 錯誤與超時比例

藉由持續觀測，才能在早期就發現效能下滑、資料異常，避免用戶體驗崩壞。

## 真正該買的不是「最強機器」，而是「最能支撐目標的系統」

導入地端 LLM，最怕的是「花了錢買硬體不知道要做什麼或是根本跑不動」。當我們完整的框架思考評估，硬體不再只是採購成本，而是企業智慧化轉型的長期基石。