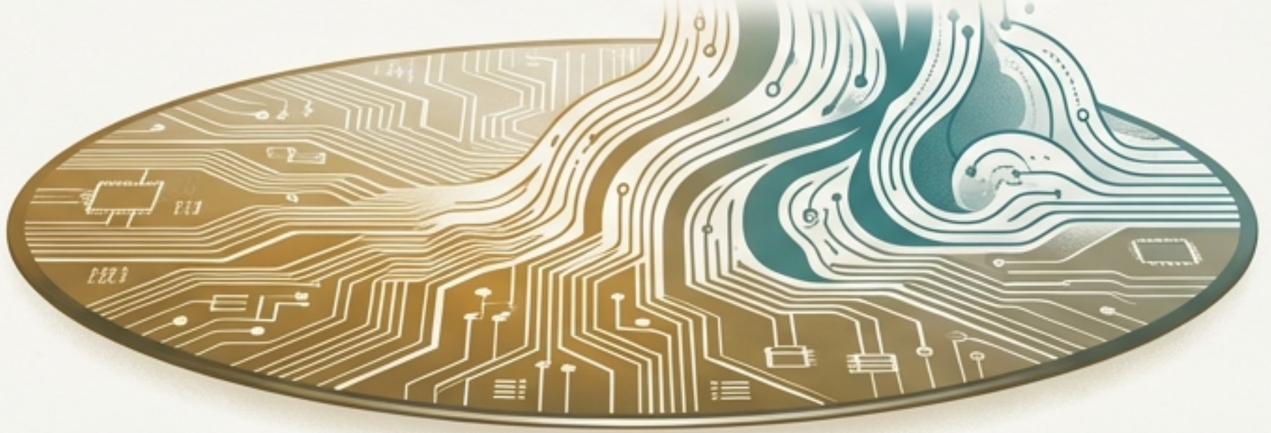
AWS 的全面整合 AI 策略：從客製化晶片到主權雲的垂直整合解析

一場深入分析，揭示 AWS 如何透過 Trainium 加速器、Neuron 生態系統及 AI 工廠服務，重塑 AI 基礎設施的競爭格局。



執行摘要：一場精心策劃的市場顛覆

AWS 正在執行一項從下到上的垂直整合策略，旨在透過提供無可比擬的性價比與部署彈性，挑戰 AI 基礎設施市場的現有領導者，並搶佔快速增長的主權 AI（Sovereign AI）市場。

市場切入點

核心武器

生態系統護城河

實戰驗證與未來藍圖

本簡報將深入解析：

1. **市場切入點 (The Market Play)**：AWS 如何利用「AI 工廠」服務，直接回應國家級數據主權與合規需求，開闢全新戰場。
2. **核心武器 (The Arsenal)**：新一代 Trainium3 晶片驚人的性能、效率與成本優勢，直接挑戰 Nvidia 的市場主導地位。
3. **生態系統護城河 (The Ecosystem)**：成熟的 Neuron 軟硬體堆疊如何降低客戶轉換門檻，並實現從底層晶片到上層框架的無縫整合。
4. **實戰驗證與未來藍圖 (Proof & Prophecy)**：以 Amazon Rufus 的超大規模部署證明其可行性，並透過 Trainium4 與 Nvidia 的結盟，揭示其未來更大的野心。

新的戰略前沿：主權 AI 的崛起

核心概念

主權 AI 指的是 AI 系統完全駐留在國家邊界內，滿足嚴格的數據落地（data residency）和監管要求。這已成為政府、國防單位及受監管行業（如金融、電信）的強制性要求。

市場驅動力

數據安全與國家安全
關鍵數據（公民、國防、關鍵基礎設施）必須免於外國管轄。

法規遵循
GDPR、數位主權法規等要求數據處理的地域性。

經濟自主
建立國家級 AI 能力，避免對外部技術供應商的過度依賴。

> *"For governments and enterprises operating under digital sovereignty mandates, the product offers the technological capabilities of public cloud regions in an isolated, controlled environment."*
> — **RCR Wireless News**

AWS 的解答：將雲端帶入客戶資料中心的「AI 工廠」

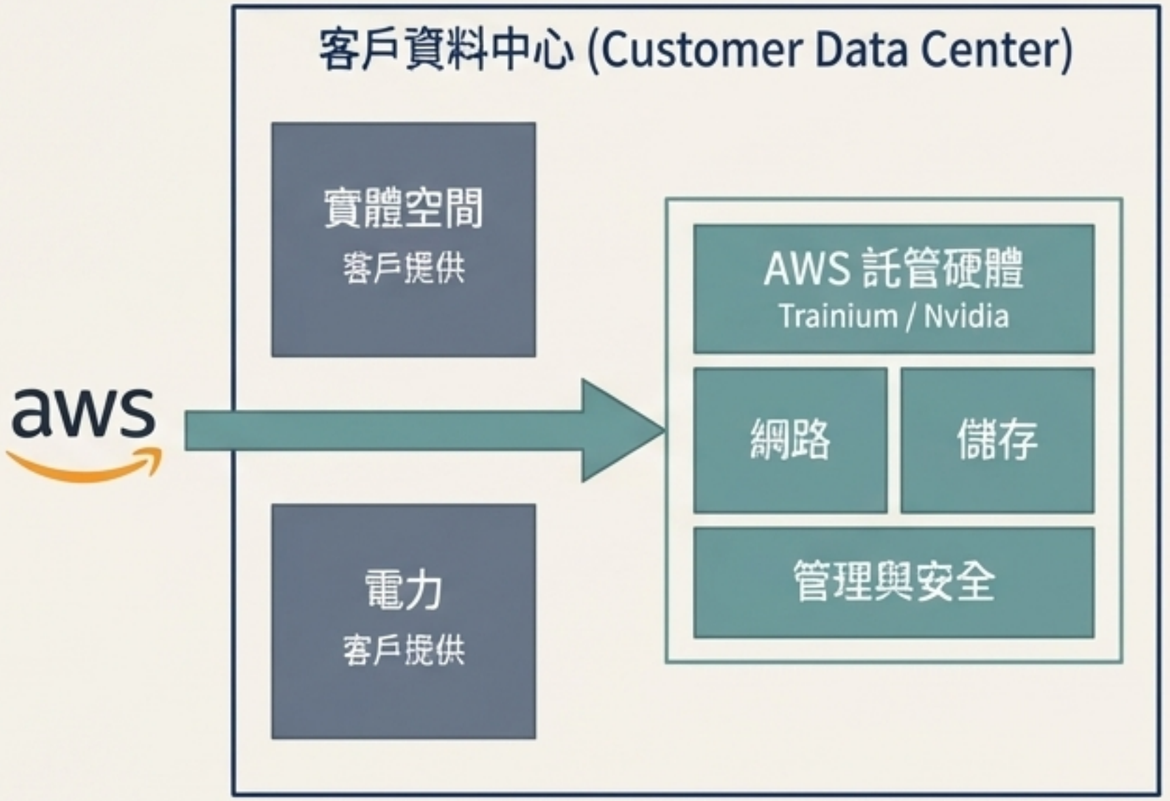
服務定義：
一項全新的託管服務，允許企業和政府在自有資料中心內部署由 AWS 管理的 AI 基礎設施。

運作模式：

- **客戶提供**：實體空間、電力。
- **AWS 負責**：部署、管理、維護所有底層系統，包括硬體、網路、儲存與安全。

核心價值主張：

- **兩全其美**：結合公有雲的操作簡易性與本地部署的數據控制權。
- **硬體選擇**：可選用最新的 AWS Trainium3 加速器或 Nvidia Blackwell GPU。
- **生態整合**：無縫接取 Amazon Bedrock 和 SageMaker，立即獲得基礎模型和 MLOps 工具。
- **加速部署**：AWS 聲稱可將 AI 基礎設施的建置時間從數年縮短至數月。



電信業的策略缺口：錯失的主權 AI 商機

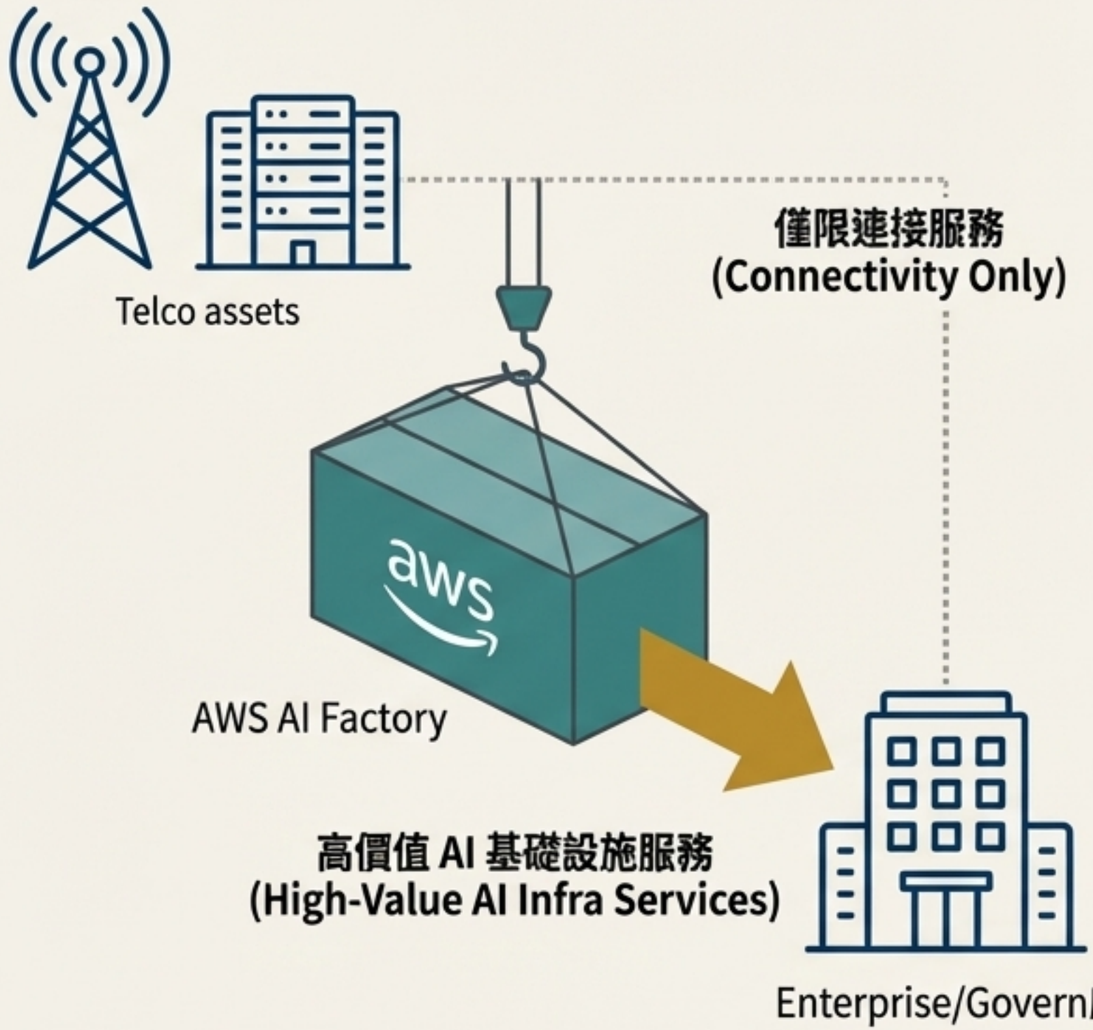
電信商的天然優勢

- **實體資產**：擁有遍布全國的資料中心和機房。
- **基礎設施**：已建立的電力和網路設施。
- **客戶關係**：與政府和大型企業有深厚的監管與商業關係。

正在發生的現實

- **錯失良機**：分析師指出，儘管擁有實體優勢，電信商並未將其轉化為有競爭力的整合式 AI 基礎設施服務。
- **AWS 積極搶佔**：AWS 正透過 AI 工廠模式，直接將高價值服務部署在潛在屬於電信商的客戶場域內。

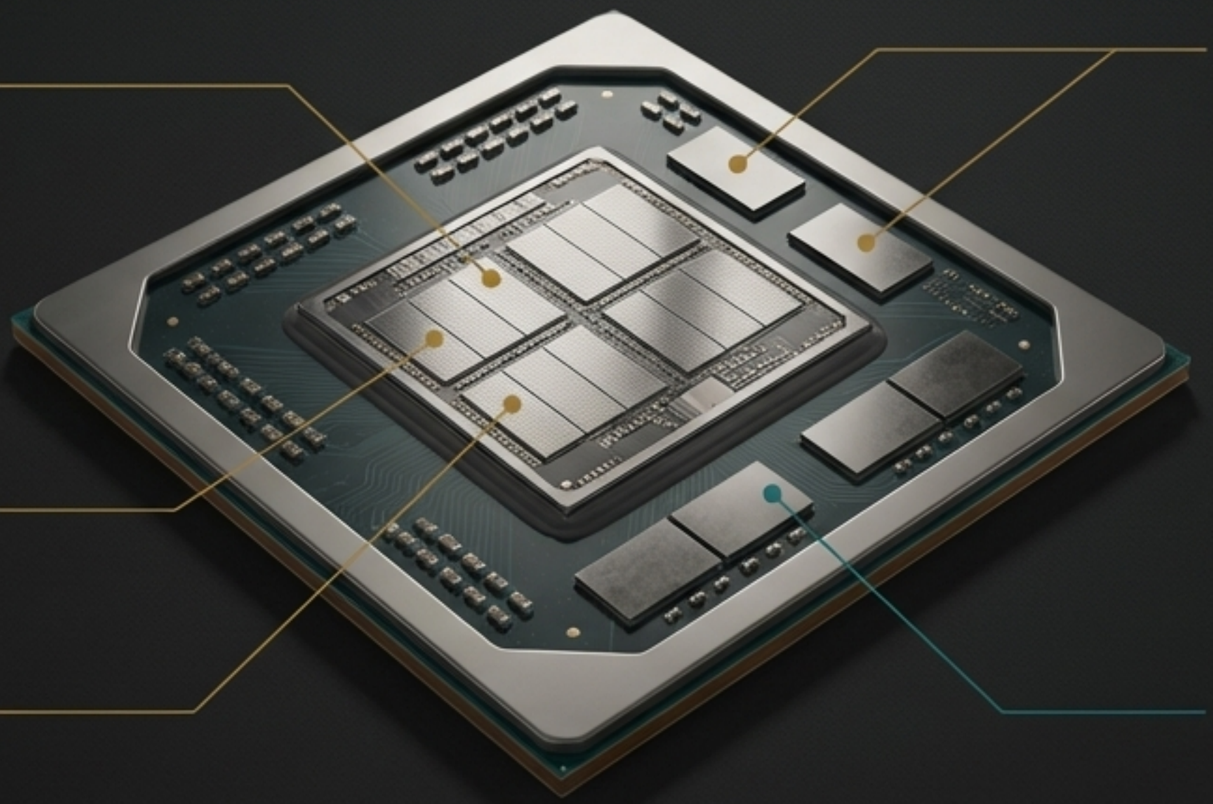
> "As Fierce Network noted... telcos appear to be missing the boat on the sovereign AI opportunity, even as AWS moves aggressively to claim it... Without a credible counter-strategy, telcos could find themselves relegated to providing connectivity while AWS captures the higher-value AI infrastructure opportunity."
>
> — 引述分析



策略的引擎：第四代 AI 晶片 Trainium3 登場

製程: 3nm

記憶體:
144 GB HBM3e



運算能力 (FP8):
2.52 PetaFLOPS (PFLOPs)

運算能力 (16:4 結構化稀疏):
高達 10 PetaFLOPS

記憶體頻寬:
4.9 TB/s

設計目標: 專為下一代代理式 AI (Agentic AI)、推理 (Reasoning) 及影片生成等
複雜工作負載而設計，針對密集型和專家混合模型 (MoE) 進行了優化。

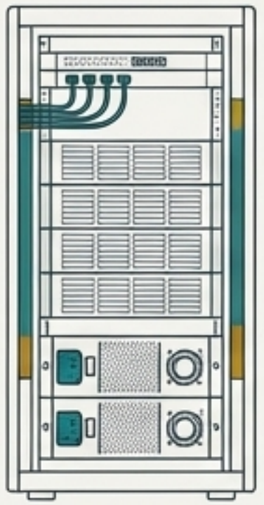
性能與效率的世代飛躍

高達 4.4 倍
性能提升

提升 4 倍
能效提升 (performance/watt)

降低高達 50%
AI 訓練成本

相較於上一代 Trn2 UltraServer

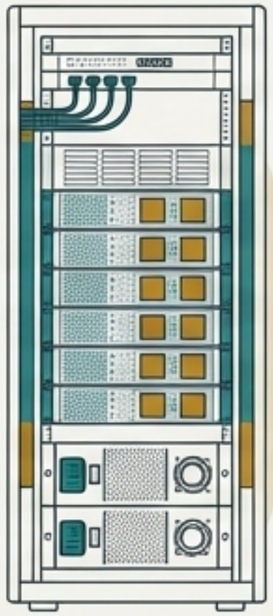


Trn2 UltraServer

加速器數量: 64 x Trainium2

總記憶體
20.7 TB
總記憶體頻寬
706 TB/s

記憶體頻寬提升: 提升 3.9 倍



Trn3 UltraServer 完整配置

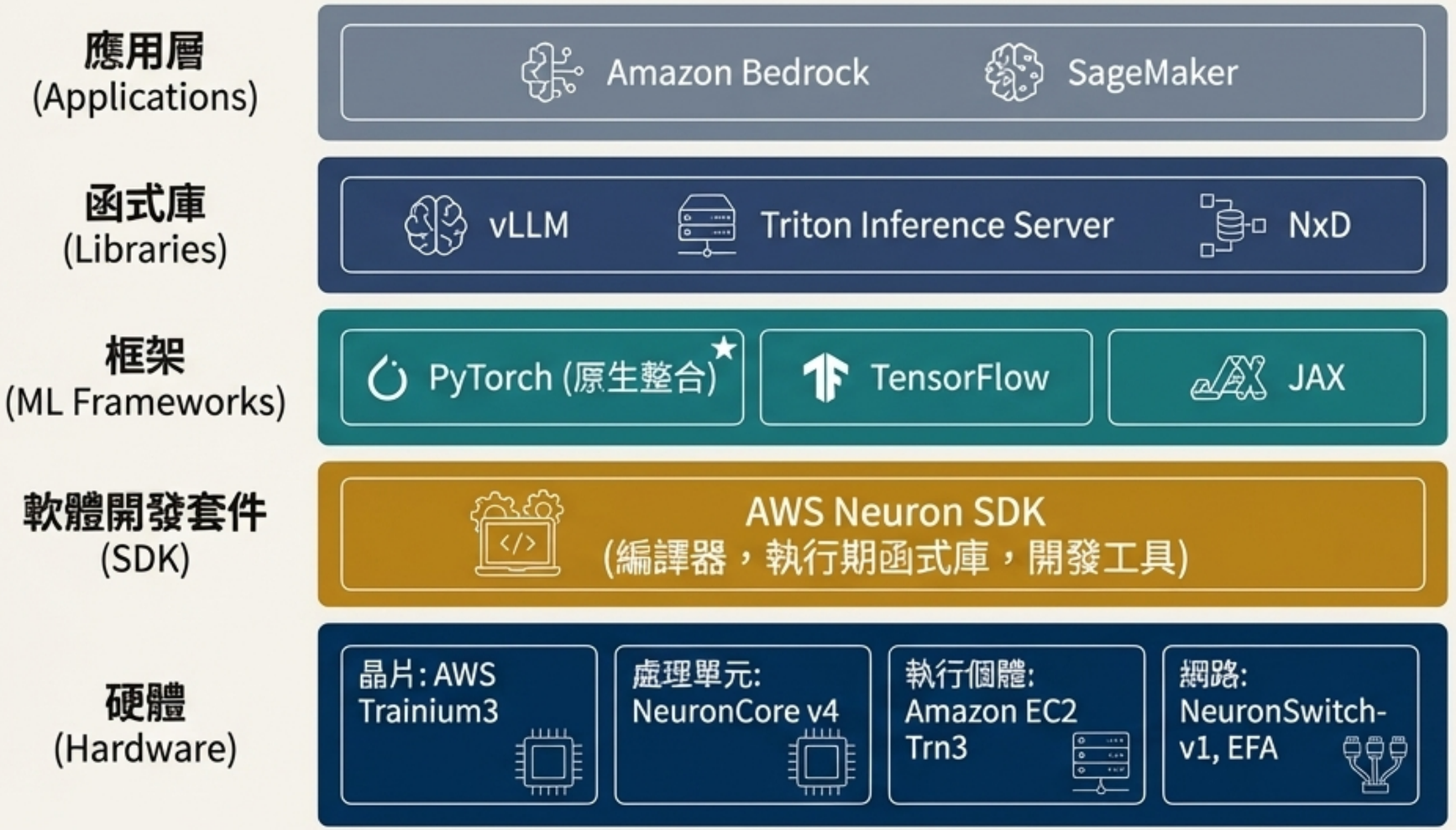
加速器數量: 144 x Trainium3 晶片

總運算能力 (FP8): 362 PFLOPs

總記憶體: 20.7 TB HBM3e

總記憶體頻寬: 706 TB/s

超越晶片：完整且成熟的 AWS Neuron 生態系統



核心訊息：

AWS 提供了從底層硬體到上層 ML 框架的完整解決方案，旨在簡化開發、部署與優化流程。

開發者體驗至上：無縫整合與深度優化

為主流框架打造

```
import torch
import torch_neuronx

device = torch.device("xla:0")
model.to(device)
```

無需修改程式碼

- **原生 PyTorch 整合**：開發者無需更改任何一行模型程式碼即可進行訓練和部署。這極大地降低了從 GPU 遷移的門檻。
- **廣泛框架支援**：同時支援 TensorFlow 和 JAX，為開發團隊提供靈活性。

為性能工程師賦能

AWS Neuron SDK：提供對 Trainium3 的底層存取能力，允許開發者進行效能微調、客製化核心（kernels），將模型性能推向極致。

開放的承諾：AWS 承諾透過開源工具和資源與開發者社群互動，促進生態系統的發展與創新。

進階工具鏈：提供全面的分析工具，如分析器（Profiler）和調試器（Debugger），幫助開發者深入了解和優化模型在 Trainium3 上的執行。

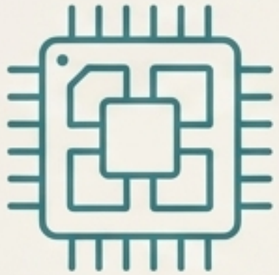
AI 加速器戰場：關鍵參與者規格對比

功能	AWS Trainium3	Google Trillium (TPU v6)	Nvidia Blackwell B200
製程節點	3nm	~3nm (est.)	4NP (Custom)
運算能力 (FP8)	2.52 PFLOPs	~3-4 PFLOPs (est.)	10 PFLOPs
運算能力 (FP4)	N/A	支援	20 PFLOPs
記憶體類型	HBM3e	HBM	HBM3e
記憶體容量	144 GB	192 GB	192 GB
記憶體頻寬	4.9 TB/s	4.9 TB/s	8 TB/s
能效提升	4x (vs Trn2)	67% (vs v5e)	25x (Inference vs H100)

分析摘要

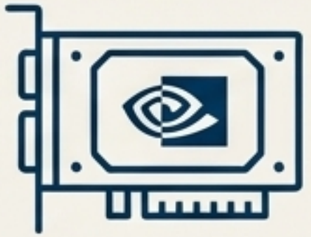
- **AWS Trainium3**：在性價比和訓練能效上極具競爭力，特別是在 FP8 精度。
- **Nvidia Blackwell**：在原始性能（尤其 FP4）和記憶體頻寬上保持領先，並擁有成熟的 CUDA 生態。
- **Google Trillium**：專為 Google Cloud 內部大規模模型優化，性能強勁。

務實的雙軌策略：自研 (Build) 與採購 (Buy) 並行



自研 (Build)

採購 (Buy)



為何自研 Trainium？

- **成本與效能優化**：針對 AWS 基礎設施進行垂直整合，以達到最佳性價比，降低內部和外部客戶的 TCO (總擁有成本)。
- **供應鏈控制**：減少對單一供應商的依賴，確保供應彈性與價格穩定性。
- **差異化競爭**：提供 Nvidia 之外的高性能選擇，吸引對成本敏感或尋求特定架構優化的客戶。

為何仍提供 Nvidia？

- **客戶選擇至上**：許多客戶的現有工作負載和技術堆疊深度依賴 Nvidia CUDA，AWS 必須滿足此市場需求。
- **頂尖性能需求**：Nvidia 仍在某些性能指標上處於領先地位，AWS 透過提供最新的 Blackwell 實例 (GB200/GB300) 來服務最頂端的性能追求者。
- **戰略夥伴關係**：維持與 Nvidia 的合作關係，確保能第一時間獲得最新的 GPU 技術。

結論：AWS 的策略並非取代 Nvidia，而是透過提供「最佳選擇」來贏得市場——無論是基於性價比的 Trainium 還是基於極致性能的 Nvidia。

實戰驗證：以 Amazon Rufus 擴展超大規模多節點推理

挑戰

- Rufus 是 Amazon 的生成式 AI 購物助理，由一個**巨大的客製化大型語言模型 (LLM)** 驅動。
- 單一加速器或執行個體的記憶體已無法容納整個模型，必須進行**多節點** (multi-node) 推理。
- 需要在服務**數百萬客戶**的巨大流量下，維持高品質互動、低延遲和高成本效益。

解決方案規模

- 成功地將一個比以往更大的模型，部署在 **數萬個 AWS Trainium 晶片** 上，為 Prime Day 的流量高峰提供支援。

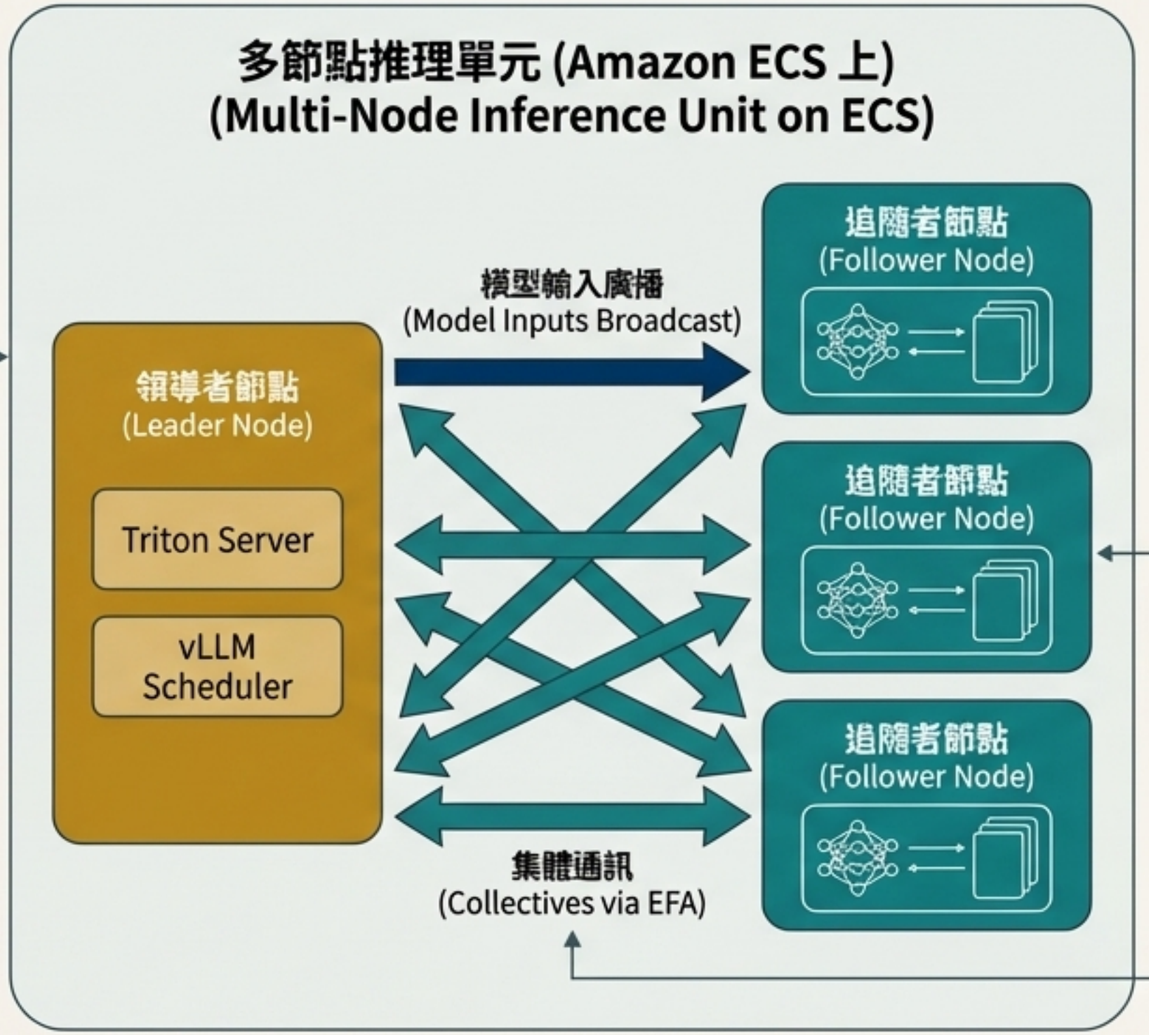
成果

- 實現了新的購物體驗，顯著提升了用戶參與度。
- 證明了 AWS Neuron 生態系統有能力在業界最大規模的生產環境中，支持關鍵任務的 LLM 部署。

Rufus 推理架構深度解析

1. 多節點推理架構

採用基於 vLLM 的 **領導者/追隨者** (Leader/Follower) 模式。Leader 節點負責請求排程、批次處理和協調，Follower 節點執行分散式模型運算。



2. 混合平行策略

透過 Neuron SDK 實現，在 **內容編碼 (Prefill)** 階段使用內容平行，在 **解碼** (Decoding) 階段使用資料平行，以最大化效率。

3. 基礎設施與網路

透過 EC2 API 進行 **網路拓撲感知部署**，確保節點物理位置相近，以利用 **EFA** 的低延遲、高頻寬 RDMA 網路。

在 **Amazon ECS** 上建立「多節點推理單元」抽象，實現可靠的滾動升級。

未來藍圖：Trainium4 與 Nvidia 的戰略結盟

下一代 Trainium4 預告

- **核心整合**：將成為首批採用 **Nvidia NVLink Fusion** 互連技術的非 Nvidia 晶片之一。
- **無縫通訊**：實現 AWS Trainium4 加速器、Graviton CPU 和 EFA 網路在 Nvidia MGX 機櫃內的無縫通訊。

性能目標

- 整體性能提升 **6 倍**
- FP8 浮點運算性能提升 **3 倍**
- FP4 浮點運算性能提升 **6 倍**
- 記憶體頻寬提升 **4 倍**

戰略意涵

- **打破壁壘**：這次合作標誌著過去封閉的 GPU 互連生態系統開始走向開放。
- **加速創新**：AWS 透過融合自家晶片設計與 Nvidia 領先的的互連技術，旨在打造下一代性能怪獸，進一步鞏固其在 AI 基礎設施領域的領導地位。

結論：一項全面、連貫且顛覆性的 AI 戰略

串連所有環節

AWS 的行動並非孤立的產品發布，而是一場精心策劃的戰略佈局。它鎖定一個高價值的主權 AI 市場（The Market），創造一個創新的 AI 工廠服務模式（The Delivery），打造一個具破壞力的 Trainium3 核心引擎（The Engine），建立一個成熟的 Neuron 軟硬體護城河（The Ecosystem），以 Rufus 的大規模應用證明其可行性（The Proof），並透過與競爭對手的合作展現其引領市場的野心（The Future）。

最終分析

AWS 正在從根本上重塑 AI 基礎設施的價值鏈。透過從晶片到雲端服務的垂直整合，它不僅為客戶提供了更多選擇，也為整個行業——從晶片製造商到電信運營商——帶來了深刻的結構性挑戰。