AI-Stack

突破AI開發瓶頸
GPU自動調度極大化算力使用率

AI-StacK: 新世代AI基礎設施管理與AI開發部署平台

2025-02-24

前言

人工智慧(AI)技術正迅速滲透到各行各業，從語音控制設備、自動駕駛汽車到藥物開發與疾病治療等，一系列改變遊戲規則的產品和服務不斷湧現。以人工智慧為驅動力的企業，正積極提升產品智能化，並優化營運與決策流程，這些能力不僅在改變各行各業，也將深刻改變企業的運作模式。

深度學習（Deep Learning）正是這場革命的核心，它屬於機器學習(Machine Learning)的一部分，是基於模仿人腦的複雜神經網絡模型。深度學習模型的開發極為依賴大量的計算資源，這使得新型硬體加速器（如 GPU 和 TPU）成為關鍵，幫助滿足龐大的計算需求。越來越多的企業正在擴大對 AI 計算資源的投資，將其引入內部資料中心或使用雲端服務，然而，如何以經濟高效的方式管理這些關鍵資源，仍然是許多企業在實施過程中需要解決的挑戰。

AI-Stack 解決了計算資源昂貴且有限的挑戰。它巧妙結合容器化技術、高性能計算（HPC）以及分散式計算的概念，將其應用於深度學習領域，實現了高效能與成本效益的最佳平衡。

當深度學習遇見企業

那麼，究竟有哪些行業正在導入人工智慧？ AI 在企業中蓬勃發展並影響各個領域，從半導體、能源、製造、交通，到醫藥、零售、汽車、金融，甚至科研與政府機構。

根據 statista 的研究報告，2024 年人工智慧的採用在全球各大企業中經歷了顯著的增長。將 AI 整合到至少一個業務功能中的公司比例大幅上升至 72%，相比 2023 年的 55%，增長顯著。更為引人注目的是生成式 AI 的爆炸性增長，全球有 65%的企業已經採用這項技術，顯示出這項技術從新興趨勢迅速轉變為主流商業工具。

資料科學的突破

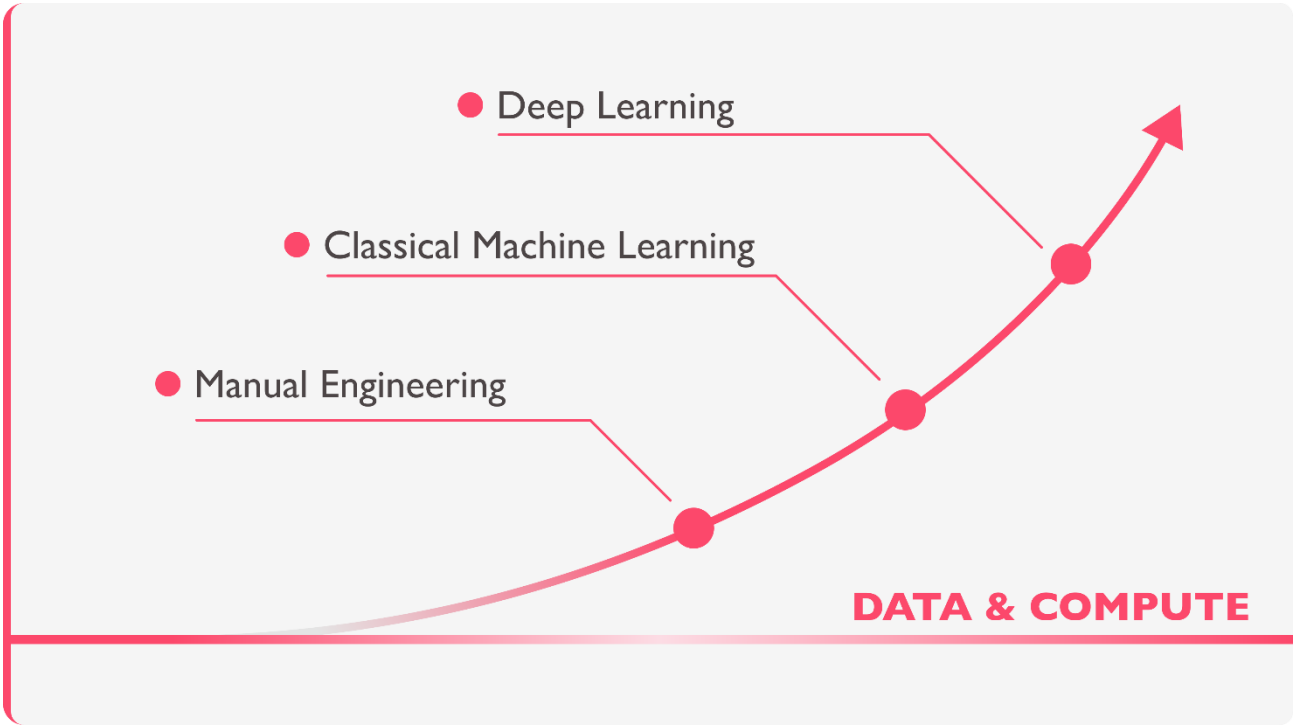

圖 1：支持深度學習的資料和計算需求呈指數增長

深度學習的發展主要由兩大因素驅動：

- **企業資料量大幅增加**：文本、圖像和影片等資料的激增，為深度學習模型提供了動力，推動自然語言處理和電腦視覺等領域的進步。
- **原生計算能力的可用性呈指數增長**：隨著 Nvidia GPU、Google TPU 等高性能硬體加速器的普及，企業能夠比以往更容易獲取強大的計算能力，加速 AI 應用的落地。

由於上述因素，深度學習在企業中正變得越來越普遍。

隨著越來越多的企業看到深度學習技術應用的成功，他們開始大膽的開發針對特定市場和消費者的新技術。然而，隨著企業對深度學習的興趣和採用不斷增加，在導入資料科學計劃的過程中仍存在諸多挑戰，導致部分企業無法順利實現預期成果。

深度學習不等於軟體工程

企業 IT 部門早已適應傳統軟體開發週期，而近年來興起的 DevOps 簡化了開發流程，並提升部署效率與規模。DevOps 將建構和部署程式的過程自動化，但在 AI 領域，尤其是深度學習與機器學習模型的開發過程卻大不相同。

深度學習作為機器學習的一個分支，和傳統軟體開發過程相比，其更具實驗性，需要頻繁的迭代和調整，這包含模型設計、問題探索、資料準備、模型訓練，最終再將模型部署至正式環境。雖然許多步驟最終將實現自動化，但目前仍仰賴資料科學家進行大量手動作業。軟體工程與深度學習的最大差異之一，正是它們對計算基礎設施的不同需求。

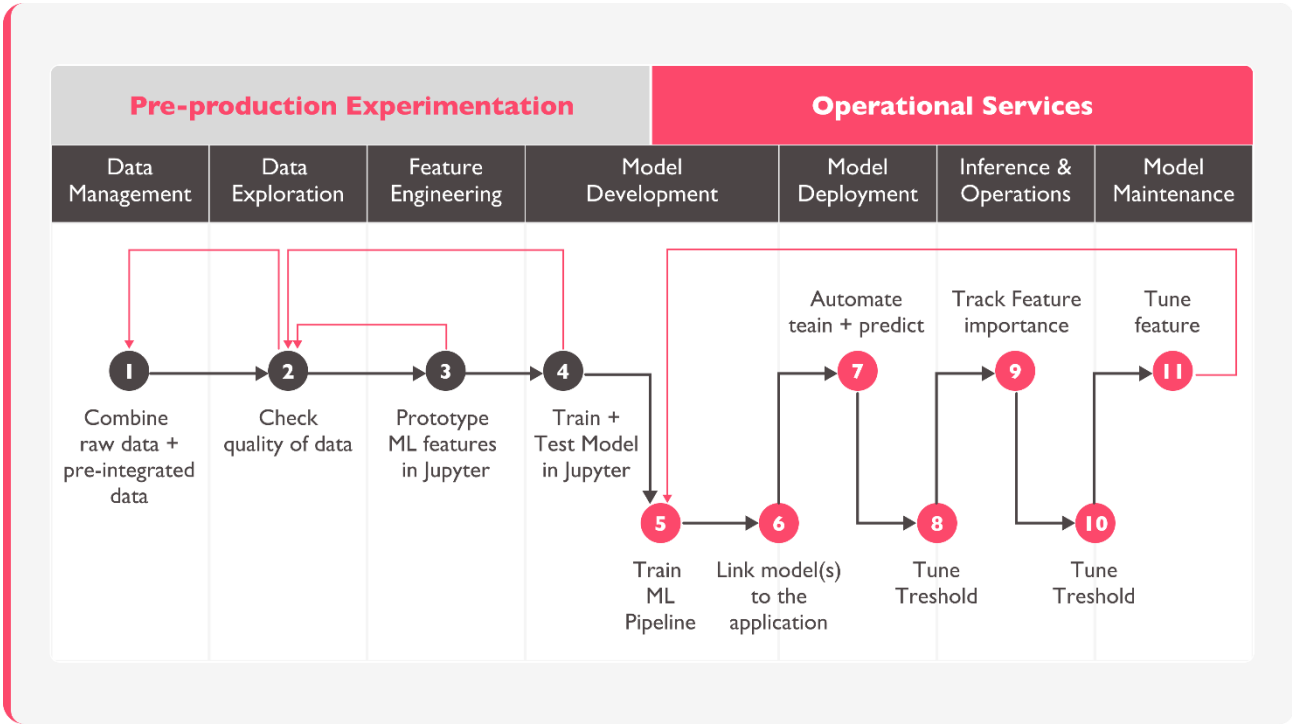

圖 2：機器學習與深度學習模型的開發過程與軟體開發不同

深度學習的開發始於實驗，不斷調整以尋找最佳模型參數正是這些實驗的核心目標。本質上它是在為特定問題挑選最合適的模型架構，並優化超參數、權重與價值組合。這一過程可能

持續數週，且通常需要同時進行多個實驗。因此，資料中心面臨前所未有的壓力——大量計算密集型工作負載並行運行，對基礎設施的彈性與效能提出了更高要求。

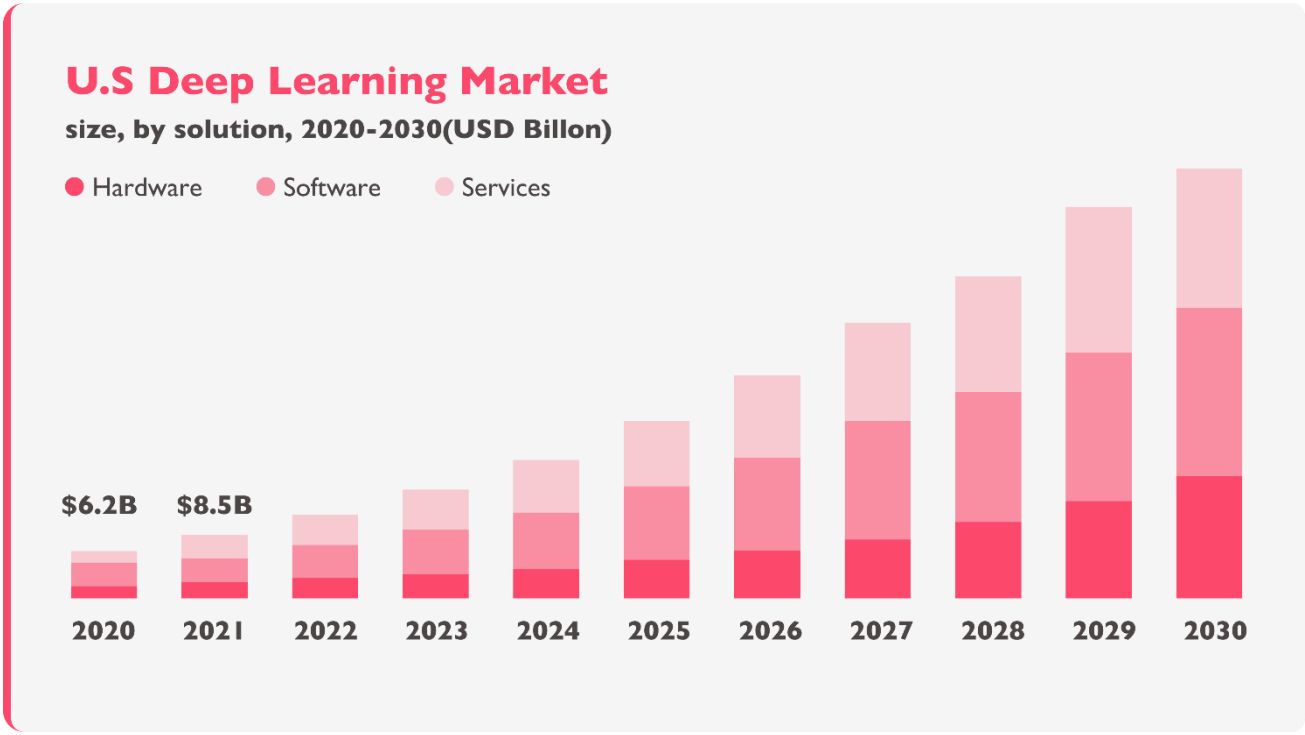

圖 3：不斷增長的硬體需求支持著蓬勃發展的深度學習活動

(來源：http://www.grandviewresearch.com/industry-analysis/deep-learning-market)

這些計算密集型的工作負載是在專用硬體上運行的，而這些硬體正變得日益普及。如上圖所示，企業以前所未有的規模投資於深度學習軟體、服務、以及 AI 專用硬體。Grand View Research 的數據顯示，全球深度學習市場預計將在 2030 年將達到 5,267 億美元，並在 2024 至 2030 年間保持 33.5% 的年均複合成長率（CAGR）。

以 AI 專用硬體加速器市場來說，NVIDIA 的 GPU 目前為領頭羊，但競爭格局正逐步變化。Google 的 TPU 等新型加速器已經上市，而 AMD、Intel 以及眾多新創公司（如 Cerebras、Graphcore）也積極投入這一領域，推出針對 AI 工作負載優化的專用晶片。部分 AI 專用晶片相較於 GPU 可能具備更佳效能，因為它們自設計之初便針對深度學習任務進行了硬體層面的優化。隨著更多 AI 專用晶片的推出，計算資源的成本將逐步降低，並更精準地滿足市場需求。

然而，計算能力的成本並非企業資料科學面臨的唯一挑戰。

IT 和資料科學還尚未完美結合

隨著機器學習計畫的推進，企業越來越重視運用資料科學與實驗來解決挑戰。根據 Precedence Research 的研究報告，全球 AI 基礎設施市場規模於 2025 年估計為 602.3 億美元，預計到 2034 年將達到 4993.3 億美元，從 2025 年到 2034 年的複合年均增長率 (CAGR) 為 26.60%。

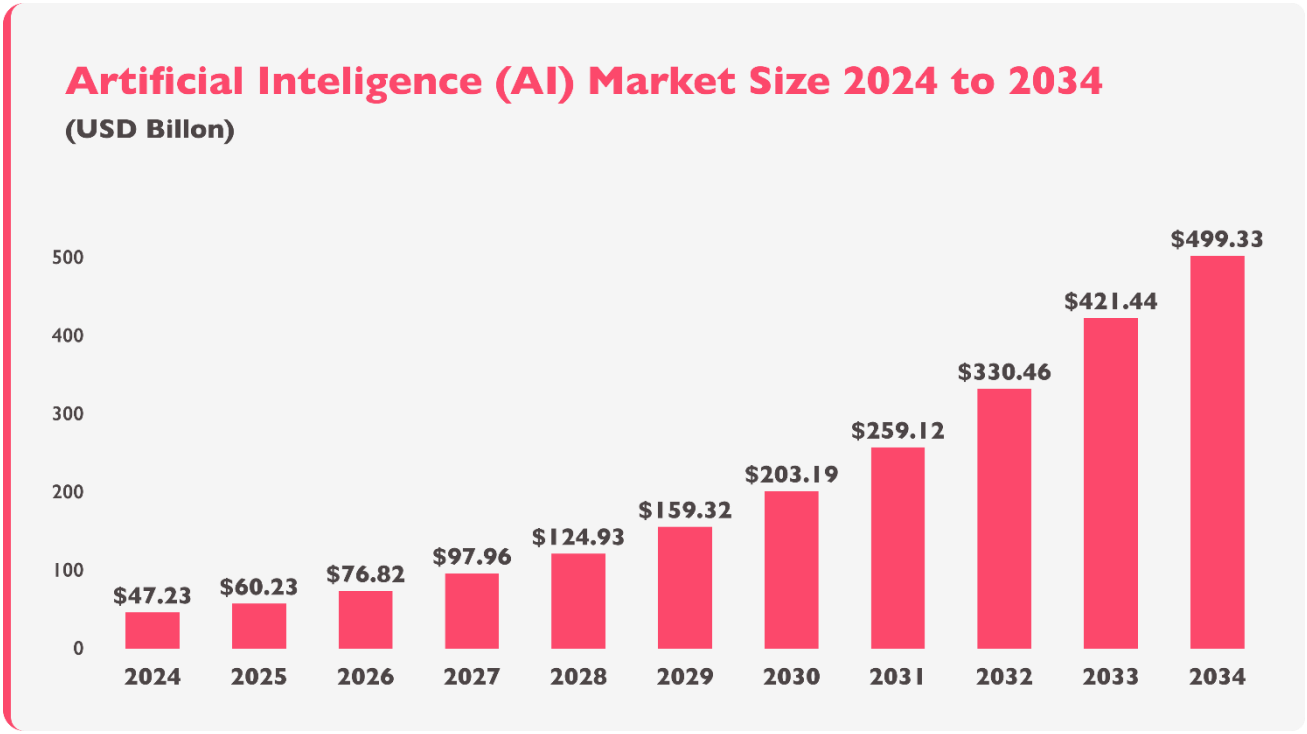

圖 4：AI 基礎設施的市場規模逐年擴大

(來源: https://www.precedenceresearch.com/artificial-intelligence-infrastructure-market)

隨著企業加大對 AI 基礎設施的投資並推動 AI 應用落地，IT 部門面臨更大的管理與運維壓力。隨著企業部署更多 AI 加速器（如 NVIDIA GPU），資源調度、維運與擴展的挑戰也隨之增加。

目前多數 GPU 以裸機方式部署於資料中心，並靜態分配給資料科學家。然而，這種分配方式往往降低 AI 研發效率，因為計算資源無法根據實際需求靈活調度。理想的 GPU 資源管理應確保高計算需求的任務獲得足夠的算力，而輕量任務則分配適量資源。若分配失衡，將導致研究人員在執行模型時遇到瓶頸，進而影響 AI 實驗的進度。

451 Research 在一項調查研究中證實了這樣的挑戰，「有近一半的企業表示，他們當前的 AI 基礎設施是無法滿足需求的。」除了需確保計算資源穩定供應，管理 GPU 也為 IT 維運團

隊帶來基礎設施管理上的挑戰。跨團隊與跨地點的基礎架構可視化、硬體維護、新資源配置等關鍵任務，皆變得更加複雜。

許多投資 GPU 的企業逐漸發現，硬體資源的管理與效益難以最大化，導致 AI 加速器叢集的利用率偏低。同時，由於基礎設施效率不彰、開發流程高度仰賴手動操作，加上冗長且昂貴的開發週期，企業生產力受限，用戶體驗也大打折扣。

再來談談 Kubernetes

到這裡，我們先來了解一下容器化技術與 Kubernetes。容器化技術在現今的資料科學領域中扮演著關鍵角色，它能夠建立一個完整的執行環境，包含所有所需的依賴軟體，使資料科學與 AI 實驗更加順暢且高效。而 Kubernetes 是新一代 IT 基礎設施軟體，專為管理與運行容器化應用而設計。類似於 VMware 在過去 20 年中對虛擬化技術的影響，Kubernetes 已成為容器編排的業界標準。透過 Kubernetes，企業能夠自動化部署、擴展與管理容器工作負載，目前已被廣泛應用於 IT 基礎架構中。

既然如此，Kubernetes 是否就能解決基礎設施低效的問題，並協調 AI 工作流程呢？透過它的調度程式（Scheduler），是否能有效調度 AI 工作負載？可惜，這些問題的答案是否定的。Kubernetes 無法完全解決 AI 工作負載的運行與管理所涉及的複雜問題。它缺乏 AI 擴展所需的關鍵批次調度能力，例如：

- **公平調度（Fair Scheduling）**：根據優先級與策略，動態管理多個調度佇列，確保資源合理分配。
- **整體調度（Gang Scheduling）**：管理多節點分散式工作負載，確保所有所需資源同時可用，避免部分資源閒置或運算中斷。
- **其他專為 AI 設計的調度機制**，用於優化計算資源使用效率。

除此之外，Kubernetes 不支援 Topology-aware scheduling（拓撲感知調度），這對深度學習工作負載的效能影響重大。Topology-aware scheduling 指的是調度程式根據硬體架構（如 CPU 拓撲、網路接口、GPU 互連）來優化計算資源分配，以提高運算效率。

簡而言之，Kubernetes 的調度機制並非專為 AI 工作負載設計，距離成為 AI 計算基礎設施的最佳調度工具，仍有很長的路要走。因此，企業若僅依賴 Kubernetes，將無法滿足 AI 計算環境的調度與管理需求，仍需搭配專為 AI 工作負載優化的資源調度與管理解決方案，以提升 GPU 利用率並優化計算效率。

在數位無限 INFINITIX Inc. 的核心產品【AI-Stack】中，我們提供了一套完整的解決方案，有效應對上述兩大挑戰——GPU 資源的管理與優化，以及透過 Kubernetes 簡化系統導入流程。我們將在下一章中進一步詳細說明這些功能與優勢。

AI-Stack 解決方案—簡化 GPU 資源調度和 AI 基礎設施管理

數位無限的產品 AI-Stack 核心設計理念是將硬體加速資源從資料科學家和深度學習工程師的開發流程中獨立出來，實現資源的動態分配與最佳化管理。基於 Kubernetes 架構，AI-Stack 有效解耦 AI 工作負載與底層硬體，使研究人員能夠專注於模型的開發與訓練，而無需關注基礎設施的配置與維護。透過自動化的資源調度與管理機制，AI-Stack 確保 AI 模型能夠高效地部署至適當的運算資源上，提升 AI 研發與應用的整體效能。

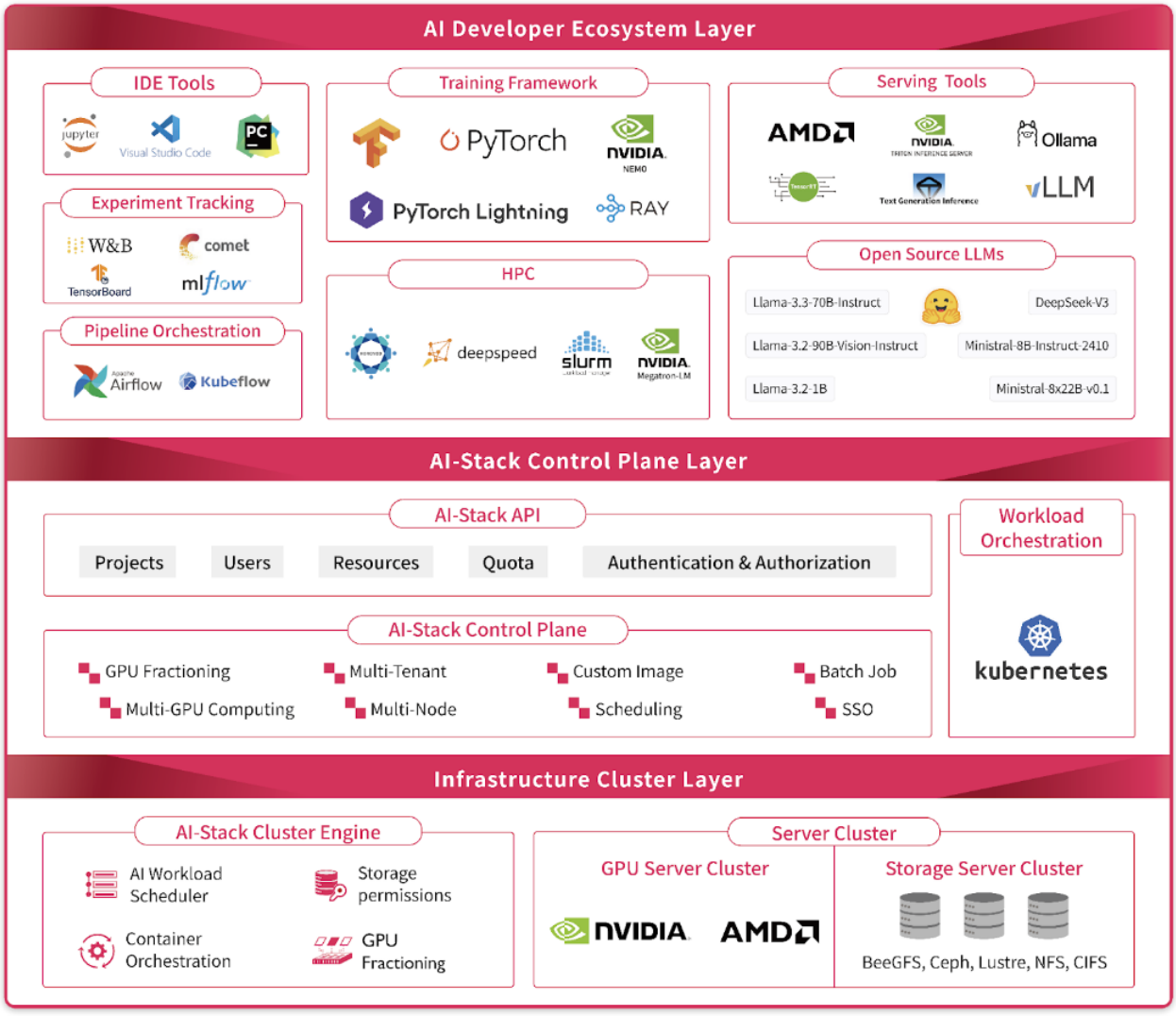

圖 5：AI-Stack 架構

AI-Stack 提供了理想的資料科學計算基礎設施，包括：

- **彈性計算資源**：為每位資料科學家提供所需的算力。
- **自動化排程運算**：一鍵排程模型訓練和運行大規模分散式計算。
- **GPU 資源最佳化**：提供高效 GPU 利用與可擴展性工具。
- **Kubernetes 整合**：透過簡單的插件，為資料科學家與 IT 團隊降低系統複雜度。

AI-Stack 具備自動化分散式計算技術，使 IT 管理者能夠更精確地分配資源，減少 GPU 閒置時間並提升叢集利用率。對於資料科學家而言，容器化 AI 基礎架構允許他們：

- **獲取更多 GPU 計算資源**，或執行更多實驗，提高生產力與研究質量。
- **透過多 GPU 進行分散式訓練**，大幅縮短模型訓練時間。

這些優化可將硬體利用率提升至少兩倍，並加速模型訓練速度超過兩倍，有效提升 AI 基礎設施的運行效率與投資回報。

資料科學工作流程的三種配置情況

AI-Stack 依據資料科學的三種不同類型的工作流程進行劃分，分別為 **模型建立（Build）**、**模型訓練（Train）** 和 **模型推論（Inference）**。在某些組織中，這些流程可能被視為一個單一的工作流程，但從計算角度來看，它們的需求與特性截然不同。

1. **模型建立（Build）**
   在模型建立階段，資料科學家需先開發、調整及測試模型，確保資料與程式碼順利對接，準備進入訓練流程。此階段涉及科學家、程式碼、資料與硬體資源的持續互動，需要反覆試驗與即時調整，計算需求較為零散，GPU 工作週期短且利用率低，對高效能計算的需求相對較低。

2. **模型訓練（Train）**
   訓練階段涉及大量計算資源，通常需要數天甚至數週的時間進行參數調整與優化。這類工作負載高度依賴 GPU 的運算能力，且記憶體需求極高，因此優化 GPU 資源的利用率至關重要，以確保計算效能與成本效益的最大化。

3. **模型推論（Inference）**
   在推論階段，已訓練完成的模型被部署至應用端，以處理實際應用場景的請求。此階段的模型對 GPU 資源的需求差異很大，且流量負載也可能隨時間波動。因此，該階段的關鍵在於根據模型需求動態分配 GPU 資源，透過部分切分或多 GPU 運行，以確保

效能最佳化。當流量高峰時，系統可以將同一個模型自動擴展部署，提高服務能力；而在流量低時，則自動降載至基本資源，維持運行效率。

資料科學工作流程各個階段的不同作業特性、工作負載以及 GPU 配置需求，可總結成下表。

模型建立（Build）	模型訓練（Train）	模型推論（Inference）
開發與調試	訓練模型、超參數優化	服務實際場景
互動式執行	無人值守的執行	依服務流量執行
短週期	運行時間長	運行時間長但零散
吞吐量不是那麼重要	吞吐量非常重要	吞吐量忽高忽低
GPU 利用率低	高 GPU 利用率	GPU 利用率時高時低

表 1：資料科學工作流程三種配置情況

這三個工作流程的差異已經清楚呈現。接下來，我們將在 AI-Stack 平台的 AI 模型開發生命週期背景下，探討這些工作流程的應用方式。

AI-Stack GPU 資源優化系統

隨著技術進步，單顆 GPU 的運算能力與記憶體空間大幅提升。NVIDIA GPU 記憶體規模已從早期僅 1.5 GB，發展至最新一代的 192GB，同時 GPU 核心數量亦不斷增加，顯著提升計算效能。

這種大量記憶體與高計算核心配置對於深度學習模型訓練等高強度運算應用至關重要。然而，許多應用場景（如模型推論）並不需要如此龐大的計算資源。若能針對不同類型的工作負載進行 GPU 切割，使單顆 GPU 被靈活分配至多個任務，將顯著提升 GPU 利用率，確保算力資源的最佳化分配與高效運行。

AI-Stack 為 Kubernetes 上的容器化工作負載提供了一套優秀的 GPU 資源優化系統。該系統支援運行基於 CUDA 指令集的工作負載，尤其適合處理模型推論與模型建立等輕量級 AI 任務。透過這套 GPU 資源優化系統，資料科學與 AI 工程團隊能夠在一片 GPU 上同時執行多個工作負載，讓企業得以在同一硬體設備上運行更多任務，例如電腦視覺、語音識別及自然語言處理等應用，從而大幅降低運營成本。

AI-Stack 的 GPU 資源優化系統能夠高效地在單一 GPU 上創建多個規格不同的邏輯 GPU，每個邏輯 GPU 都擁有獨立的 GPU 記憶體與計算空間，並可供容器獨立存取與運行，如同真

實 GPU 一樣。這使得多個工作負載能夠在同一個 GPU 上同時運行容器，且彼此隔離，不會互相干擾。

AI-Stack 的 GPU 資源優化系統具有透明、簡單且可移植的特性，無需對容器本身進行任何調整或修改，即可高效運行。典型應用場景中，可在單顆 GPU 上同時執行 10 個以上的工作負載，實現 10 倍以上的硬體利用效益。若使用 NVIDIA H100 80GB 等高階 GPU，在最小可切割至 1GB 情境下，甚至可切分高達將近 80 個 GPU 切片，支援近 80 個容器並行運行，大幅提升計算資源的靈活性與效率。

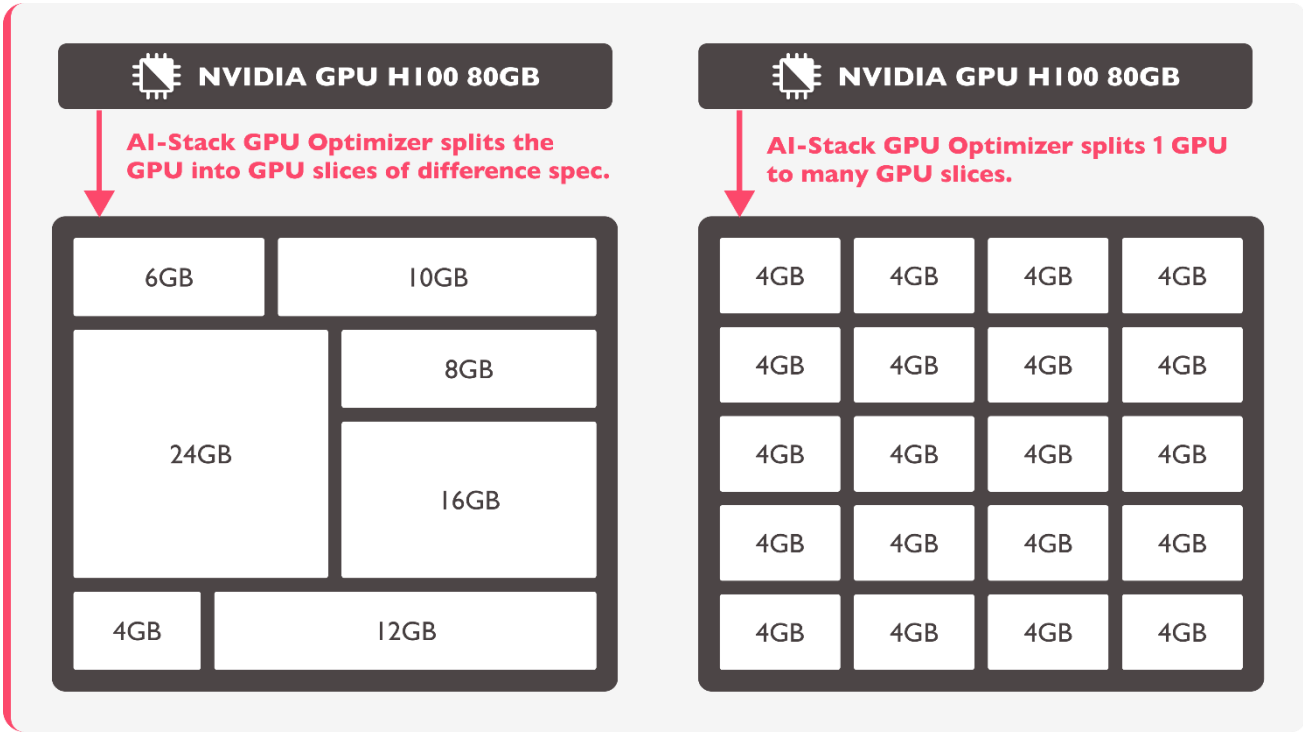

圖 6：GPU 資源優化系統依工作負載將 GPU 切分成多個邏輯 GPU

AI-Stack 的 GPU 切割功能，主要用於模型建立與模型推論等輕但多量的 AI 任務，滿足資料科學與 AI 工程團隊建立與部署模型需求。

透過整合異質資源，能夠在多個邏輯環境中靈活調度，從而原生支援模型建立、模型訓練與模型推論等不同階段的計算特性，並有效提升資源利用率。這樣的虛擬資源池可直接部署於 Kubernetes 叢集中，實現資源的統一管理與高效運用。

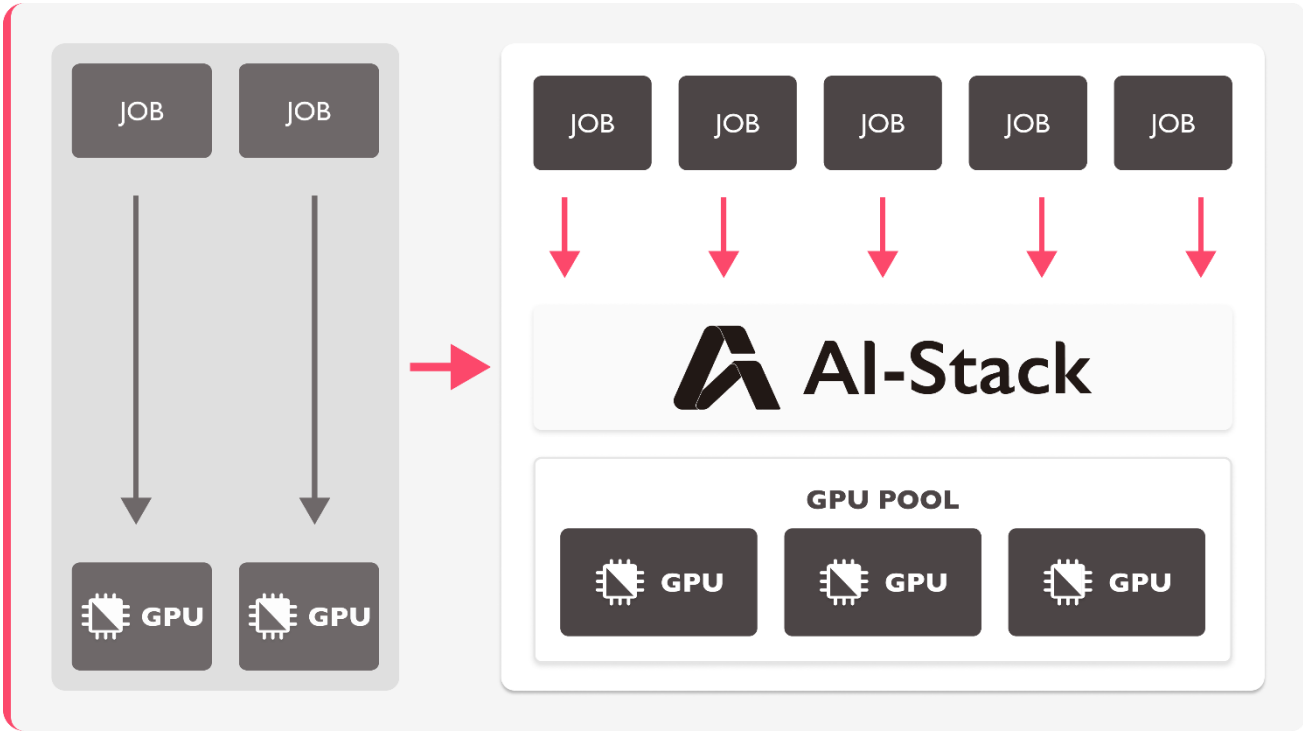

圖 7：AI-Stack 匯集異質資源池提高資源利用率

AI-Stack 工作負載調度程序

AI-Stack 獨特的工作負載調度系統運行於 Kubernetes 上，提供了關鍵的深度學習工作負載管理。其功能包括先進的多佇列機制、固定與保證配額、優先權與策略管理、自動暫停/恢復、多節點訓練支持等，為複雜的調度流程提供簡潔高效的解決方案。

透過匯集資源並搭配 AI-Stack 的調度程序進行管理，系統管理員可全面掌控資源配置，使其與業務目標保持一致，輕鬆新增用戶、維護與擴充硬體，並獲得 GPU 使用與利用率的完整視圖。同時，資料科學家可在無需依賴 IT 部門的情況下，自動配置所需資源。

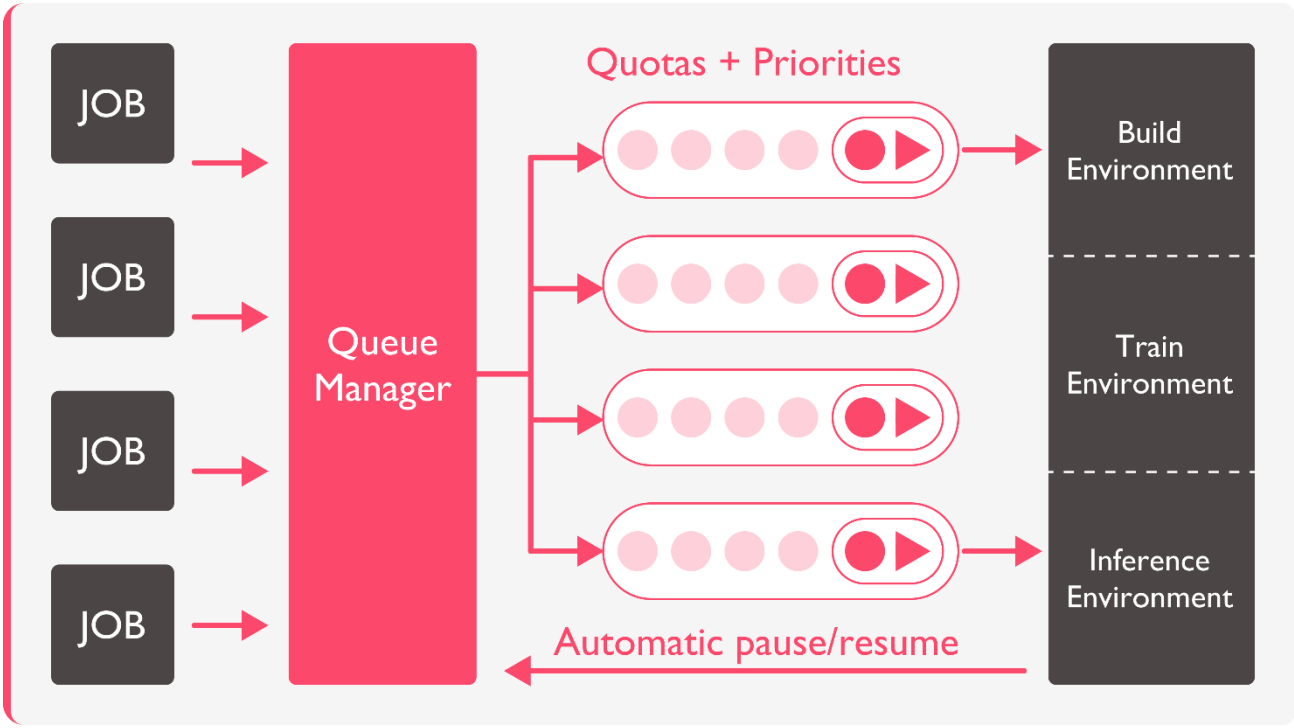

圖 8：獨到的 AI-Stack 工作負載調度程序

多個邏輯環境資源池與 AI-Stack 工作負載調度系統互動，支援模型建立、訓練與推論工作負載的高效管理。

GPU 資源分配策略

目前，大多數深度學習專案採用固定資源配額策略，即研究人員為專案預先分配特定數量的 GPU，稱為「定額分配」。與定額分配策略相反的，是「超額分配」策略，其允許專案在需要時使用超過原定配額的 GPU，最大化資源利用並減少閒置時間。

在超額分配策略下，系統會根據當前可用資源自動分配 GPU，即使該專案已超出原定配額。這種彈性機制突破了傳統定額分配的限制，使資料科學家能夠運行更多併發實驗，或在多 GPU 訓練中獲得更高效能，從而顯著提升整體 GPU 叢集的使用率。

定額分配策略	超額分配策略
適合大多數模型建立與推論的工作負載	適合模型訓練的工作負載
GPU 始終都要可以提供使用	用戶可以使用超過原定配額
	可以執行更多的併發實驗
	可以執行更多的多 GPU 訓練模型

表 2：GPU 資源分配策略

理想情況下，「模型建立」的工作流程中需要用的 GPU， 應始終可供資料科學家使用，以確保模型建立階段的開發與調試環境隨時可用。「模型訓練」的工作流程主要用於進行實驗，在實驗階段中也須要有能夠滿足模型需求的運算資源。當 GPU 閒置時，系統可將資源動態分配給當前有需求的實驗，提升整體資源利用率。

各專案依據組織需求，以佇列方式提交至 AI-Stack 調度程序。這些佇列可按照個人、團隊或特定業務活動劃分，每個佇列皆可設定專屬的優先級與資源配額，確保計算資源分配符合策略與運營目標。

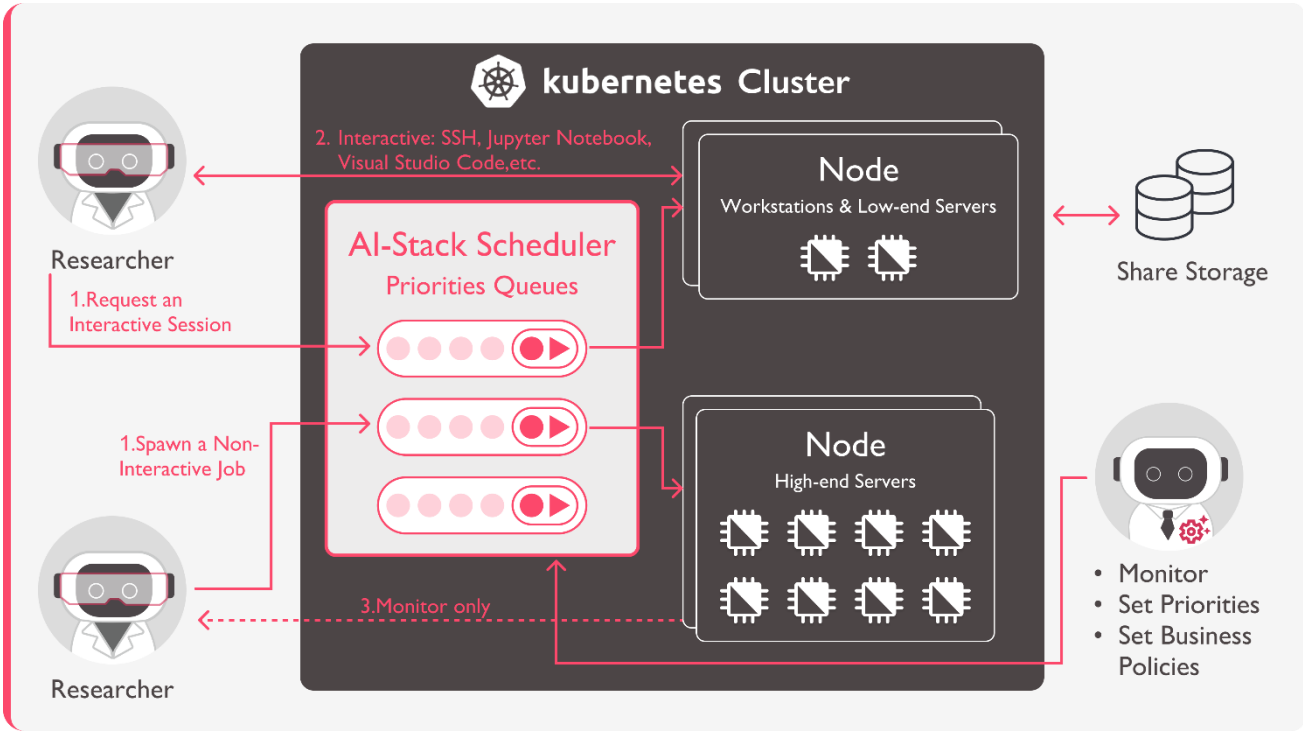

圖 9：在建立模型和訓練模型環境中，將 GPU 分配給不同研究人員

AI-Stack 的調度程序與超額分配策略如何優化資源利用？以上圖說明：圖中顯示研究人員提交兩種運算工作——互動式工作 (Interactive Session) 和非互動式工作 (Non-Interactive Job)。這兩種運算工作分別對應於資料科學工作不同流程中，互動式工作主要用於模型建立和推論，非互動式工作則涉及模型訓練。

- 在模型建立和推論環境中，建議採用定額分配策略。
    - GPU 資源固定分配，無法與別人進行共享。GPU 資源始終都要可以提供使用，確保研究人員的開發與調試隨時可用。
    - 互動式執行(如 SSH、Jupyter Notebook) 直接連接至伺服器的 GPU 節點，確保用戶可即時存取資源。

- 在模型訓練環境中，其規模較大，建議採用 GPU 池化與超額分配。
    - GPU 資源池化，不同工作負載可以靈活共享所有可用資源，提升利用率。
    - 用戶可以使用超過原定配額， 高優先級佇列可優先獲得 GPU 分配。若資源不足，調度程序會根據優先級與公平性原則，暫停部分低優先級的工作。當資源充足時，低優先級佇列仍可使用閒置 GPU。

GPU 資源分配案例

讓我們以實際的例子來說明。

資料科學家 Hunter 每天都會獲得固定的資源配額，例如兩個專屬 GPU，這些 GPU 只能由他使用。在短期內使用 GPU 通常沒有問題，且他的工作流程運行順暢（至少在模型構建階段是如此），此時他的重點放在模型開發和調試。然而，當模型訓練階段開始時，情況會有所改變。突然間，他的工作負載需要更多額外的 GPU 算力資源，但這些資源通常因為配額有限而無法獲得，即使同事沒有使用他們的資源，Hunter 仍無法超過配額。

通常，Hunter 需要更多計算資源來達成三個目標：

1. 進行更多實驗
2. 運行多 GPU 訓練以加速模型訓練並實驗更大的批次規模，從而提升模型精度
3. 在訓練一個模型的同時建立新的模型

然而，透過 AI-Stack，Hunter 能夠順利達成這些目標，實現更高效的工作流程。

如何做到的？全靠超額分配策略。Hunter 團隊的所有資源集中管理，這樣他可以獲得更大的 GPU 配額，並根據 AI-Stack 的超額分配策略，為他和其他每位用戶提供靈活的計算能力。

如前所述，Hunter 團隊中的每位資料科學家在模型建立階段都有固定的資源配額。系統會為模型開發分配固定的 GPU 資源，確保科學家始終能夠使用 GPU。然而，一旦他們希望超過配額，進行更多實驗、運行多 GPU 訓練、執行更大規模的工作或同時進行多個實驗，他們只需將工作發送到共享的更大資源池中，系統會進行協調。

如果 Hunter 不使用分配給他的 GPU，其他人可以使用這些資源。但當他需要這些 GPU 時，系統會確保他能夠使用它們，這就是他的「超額分配策略」。一旦資源再次可用，被搶占的工作會自動重新啟動。

當其他研究人員也加入會發生什麼事？

在下圖中，Hunter、Allen 和 Evan 共享一個包含八張 GPU 的叢集。在靜態分配或定額分配策略下，每位用戶會被分配兩個 GPU 的配額。然而，在超額分配策略下，每位用戶的 GPU 配額是虛擬分配的。Hunter 始終可以存取兩個 GPU，但如果他需要四個 GPU，且他的同事未使用，系統會允許他存取四個 GPU。一旦 Allen 或 Evan 需要這些 GPU，系統將回收並重新調度這些資源。

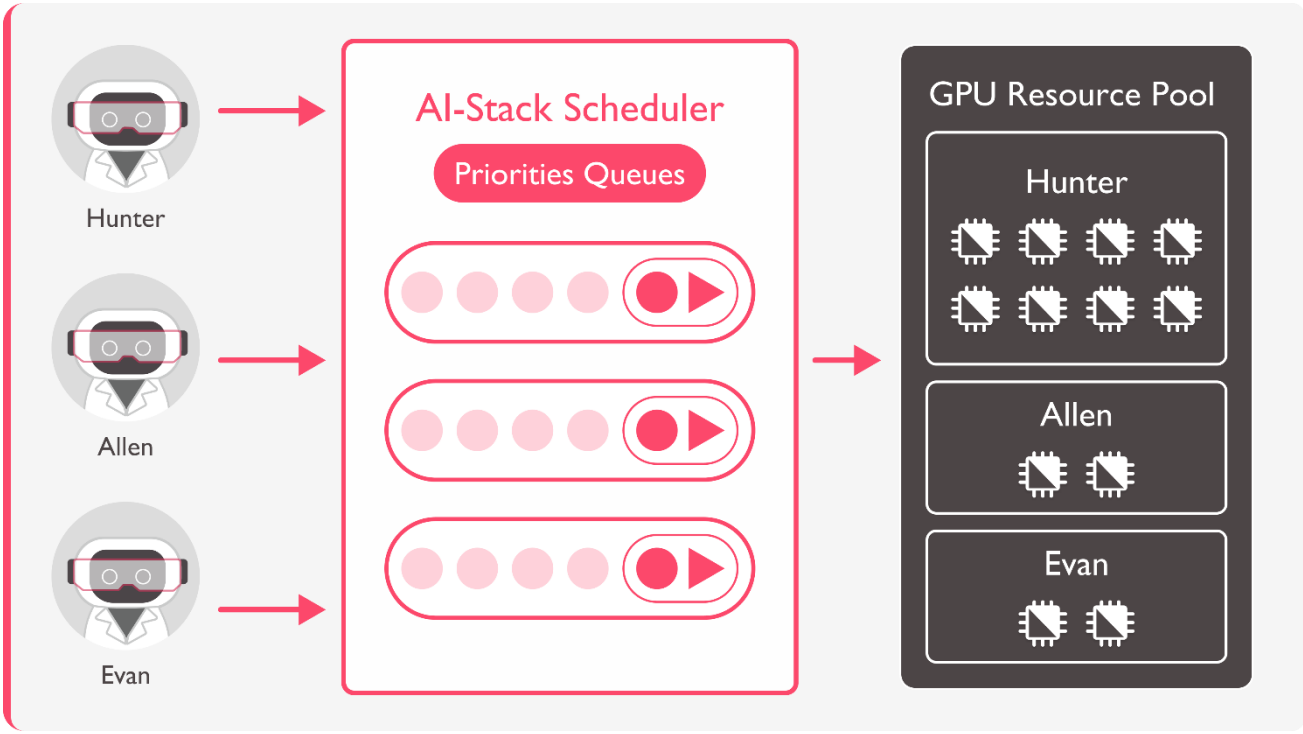

圖 10：共享一個 GPU 資源池進行訓練模型

總結來說，在超額分配策略下，如果用戶的資源閒置且有可用容量，系統會分配工作，即使這些工作超出了原定配額。當新工作進入系統時，AI-Stack 調度器會根據業務目標管理各個佇列，並提供一種機制，根據優先級和策略來暫停或恢復工作負載。這樣的策略不僅提高了整體叢集的資源利用率，還大大提升了資料科學團隊的生產力。

AI-Stack 實際客戶案例

AI-Stack 的其中一個客戶，其工程團隊在異質硬體環境中工作，包含各類 GPU，從搭載低階 Nvidia GeForce 或 Quadro 系列 GPU 的工作站，到配備高階 Nvidia Tesla A100/H100 GPU 的伺服器，以及 Nvidia DGX 伺服器。在導入 AI-Stack 之前，該客戶採用靜態的硬體分配方式。

最初，他們的資料科學家與工程師團隊規模較小，每位使用者均透過手動、靜態的方式分配專用資源，每人擁有不同數量與類型的 GPU。隨著團隊擴大、深度學習專案規模增長，IT 部門不斷採購更多硬體以應對需求。然而，預期的生產力提升並未實現，反而面臨愈來愈多的挑戰——新加入的使用者獲得了專屬資源，但整體團隊的運算效率仍然不佳。即使硬體投資持續增加，資料科學家仍因基礎設施瓶頸受限，導致生產力未能顯著提升，且運算成本不斷攀升，使投資回報率難以衡量。

AI-Stack 對該叢集的 GPU 初步分析後，迅速發現以下問題：

- GPU 規格多樣，包含高階及低階卡，難以統一列管
- 叢集內所有 GPU 的平均利用率僅 30% 以下。
- 多 GPU 訓練任務的使用率雖較高，但波動顯著。
- 建模相關的交互式工作負載利用率極低，GPU 長時間閒置，即便在運行時，利用率也低於 50%。

這樣的資源配置方式顯然未達最佳化——部分使用者間歇性使用 GPU，但大部分時間 GPU 均處於閒置狀態，導致整體利用率低於 30%。

使用 AI-Stack 提高利用率，加快實驗速度

當企業選擇 AI-Stack，他們的首要任務是將所有 GPU 節點納入 AI-Stack 管理，把所有資源集中整合。接著，GPU 依據不同用途劃分為兩個群組：

1. **定額分配策略的建立模型**：在此群組中，每位研究人員擁有一台工作站，並分配兩張低階 Nvidia GeForce 系列 GPU 的固定配額。研究人員在建立互動式模型時，可以交互使用兩張 GPU 的算力資源，但此群組內的資源無法與別人進行共享。

2. **超額分配策略的訓練模型**：此群組專為模型訓練設計。在該群組中，高階 GPU 資源專門分配給正在進行訓練的研究人員。使用高階 Nvidia DGX 伺服器，透過 AI-Stack 將所有高階 GPU 資源集中管理，所有研究人員皆可共享這些資源。該策略可以最大限度地提高計算能力，應對連續數天的工作負載，並加速結果產出。

在此基礎上，AI-Stack 提供全方位監控，讓客戶能夠：

- 掌握所有 GPU 的狀態與使用情況
- 讓管理員進一步優化資源利用率
- 自助創建 AI 開發環境
- 快速部署 AI 模型

導入 AI-Stack 後，客戶獲得顯著效益：

- **實驗數量大幅提升，成果產出更快**——以每 GPU 進行的實驗數來衡量，生產力提升超過 10 倍。
- **GPU 叢集的平均利用率超過 90%**，大幅優化硬體資源。
- **提高投資回報率（ROI）**，集群資源的高效運用降低了硬體支出成本。
- **簡化研究人員的工作流程**——資料科學家不再需花費大量時間在管理基礎設施與工具，而是能夠專注於提升實驗數量與質量，進一步推動創新。
- **IT 部門獲得更高的可見度與掌控度**，新用戶可輕鬆上手，硬體整合與資源配置更為流暢，並透過 AI-Stack 的分析功能協助未來資源規劃。

這一切，使企業的 AI 研發流程更加高效，實現更優異的資料科學成果。

結論

深度學習在各個行業領域的應用正急劇增長，隨之而來的是對計算能力的強烈需求。對於資料科學計劃日益增長的企業來說，僅僅購買和部署額外硬體並不足以解決問題。如果缺乏高效且具成本效益的資源利用策略，低效能和不明確的投資回報將繼續成為企業面臨的挑戰。

AI-Stack 針對不同工作負載定義了三種任務類型：較小的建立模型任務、較大的訓練模型任務以及隨時提供服務的推論模型任務。通過精確的資源分配，AI-Stack 確保較小的任務擁有足夠的 GPU 資源，較大的任務擁有無限制的資源支持，並靈活調度資源以支持推論模型的實際應用場景。這樣，資料科學的工作負載得以與硬體資源脫鉤，實現資源的最大化利用。

通過匯集資源並應用先進的調度機制來優化資料科學工作流程，AI-Stack 為客戶提供了保障優先工作的配額，使其能夠加速實驗運行、縮短結果產出時間，最終實現 AI 計劃業務目標。

了解更多 AI-Stack 相關資訊：https://ai-stack.ai/

