INFINITIX

基礎設施再進化
AI-Stack 驅動智慧應用落地

數位無限執行長 陳文裕

AI的發展快速
企業紛紛加速AI落地

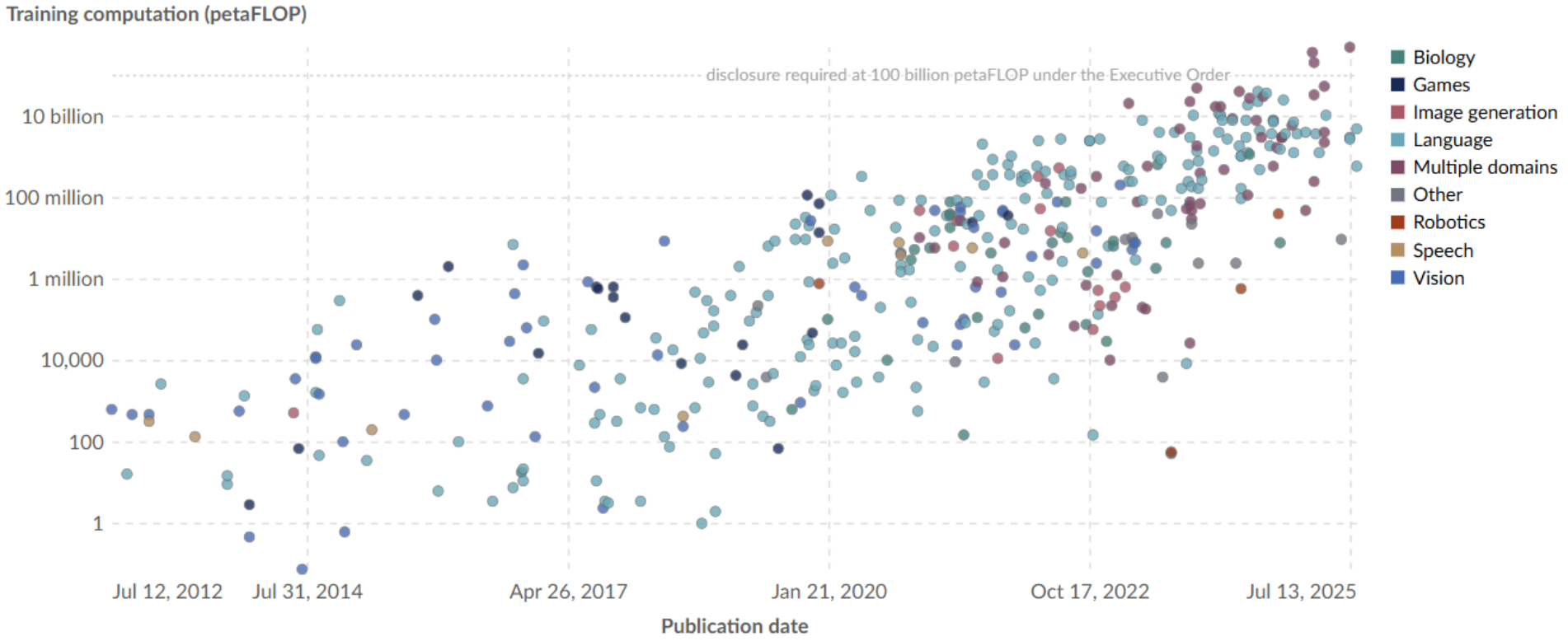
AI運算需求快速成長

Computation used to train notable artificial intelligence systems, by domain

Our World in Data

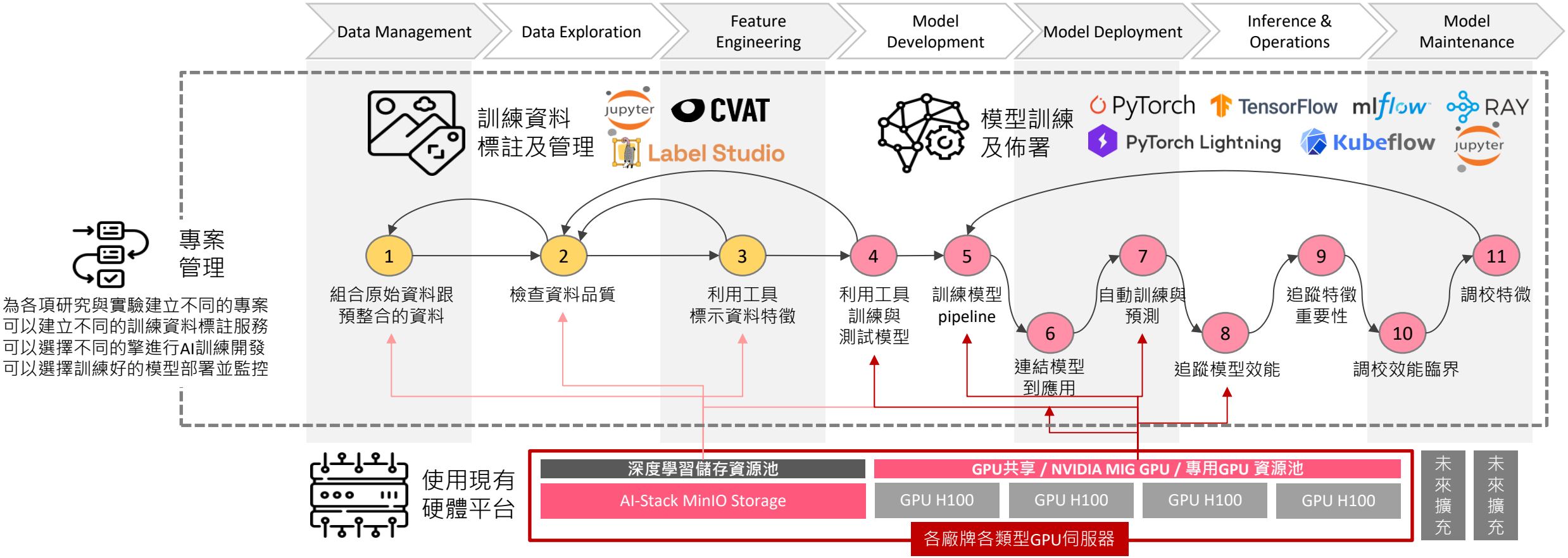
Computation is measured in total petaFLOP, which is 10^{15} floating-point operations. Estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.

Table | Chart | Settings

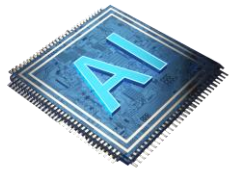


AI 開發全生命週期

圖形化介面
使用及管理



人工智慧應用演進



Step One

模型訓練

Step Two

推論服務

辨識、預測

Step Three

GenAI

生成式AI是人工智慧中的一個分支，主要用於創造性的工作，例如文章生成、影像生成、音樂生成等。

Step Four

Agent AI

AI Agent，指的是能夠自主做出決策、完成行動，而且毋需人類介入的人工智慧。

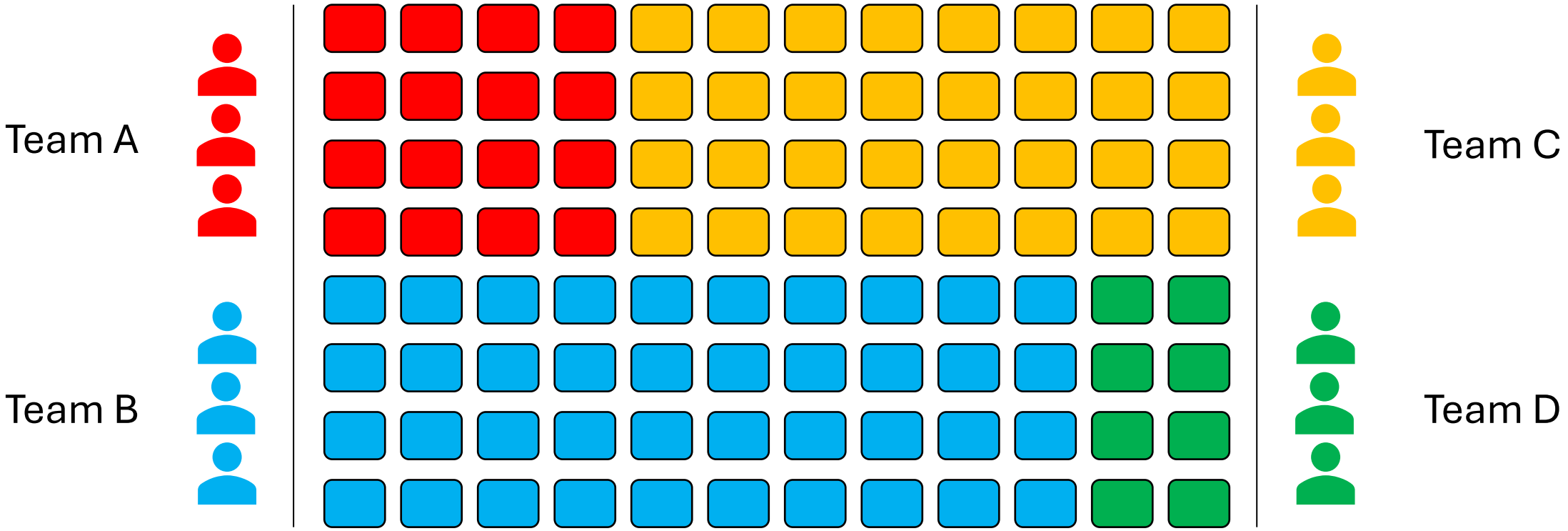
GPU算力對AI導入至關重要

AI 與GPU 算力

模型訓練　　**推論服務**

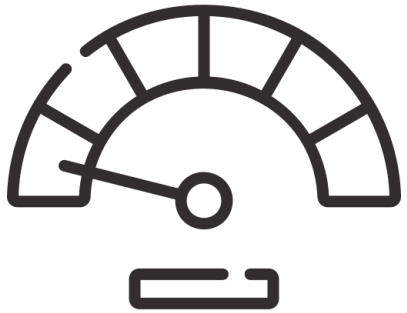
GenAI　　**Agent AI**

GPU 的使用各自為政



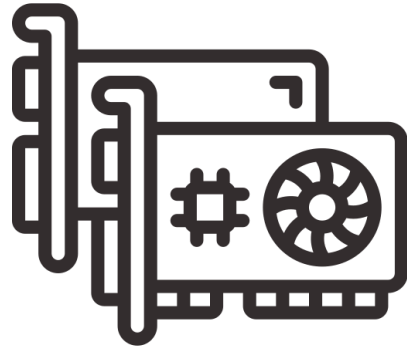
人工智能基礎設施管理面臨的挑戰

缺乏控制和優先次序



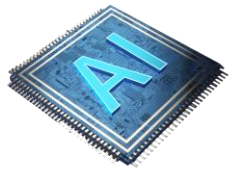
利用率低，成本高

難以獲得能見度和更好的決策



用戶仍然需要更多 GPU

GPU算力調用情境



Step One

模型訓練

Step Two

推論服務

辨識、預測

GenAI

生成式AI是人工智慧中的一個分支，主要用於創造性的工作，例如文章生成、影像生成、音樂生成等。

Four

nt AI

，指的是能夠出決策、完成行動，而且毋需人類介入的人工智慧。

- ✓ GPU 記憶體容量(高)
- ✓ GPU 聚合
- ✓ 分散式訓練

GPU算力調用情境

Step Two

推論服務

辨識、預測

- ✓ GPU 記憶體容量(小)
- ✓ GPU 分割

GenAI

生成式AI是人工智慧中的一個分支，主要用於創造性的工作，例如文章生成、影像生成、音樂生成等。

自主做出決策、完成行動，而且毋需人類介入的人工智慧。

GPU算力調用情境

- ✓ GPU 記憶體容量(中)
- ✓ GPU 分割

Step One

模型訓練

Step Two

推論服務

辨識、預測

Step Three

GenAI

生成式AI是人工智慧中的一個分支，主要用於創造性的工作，例如文章生成、影像生成、音樂生成等。

Step Four

Agent AI

AI Agent，指的是能夠自主做出決策、完成行動，而且毋需人類介入的人工智慧。

GPU算力調用情境

- ✓ GPU 混合行記憶體容量配置
- ✓ GPU 分割+聚合
- ✓ 算力調度策略

Step One

模型訓練

Step Two

推論服務

辨識、預測

生成式AI是人工智慧中的一個分支，主要用於創造性的工作，例如文章生成、影像生成、音樂生成等。

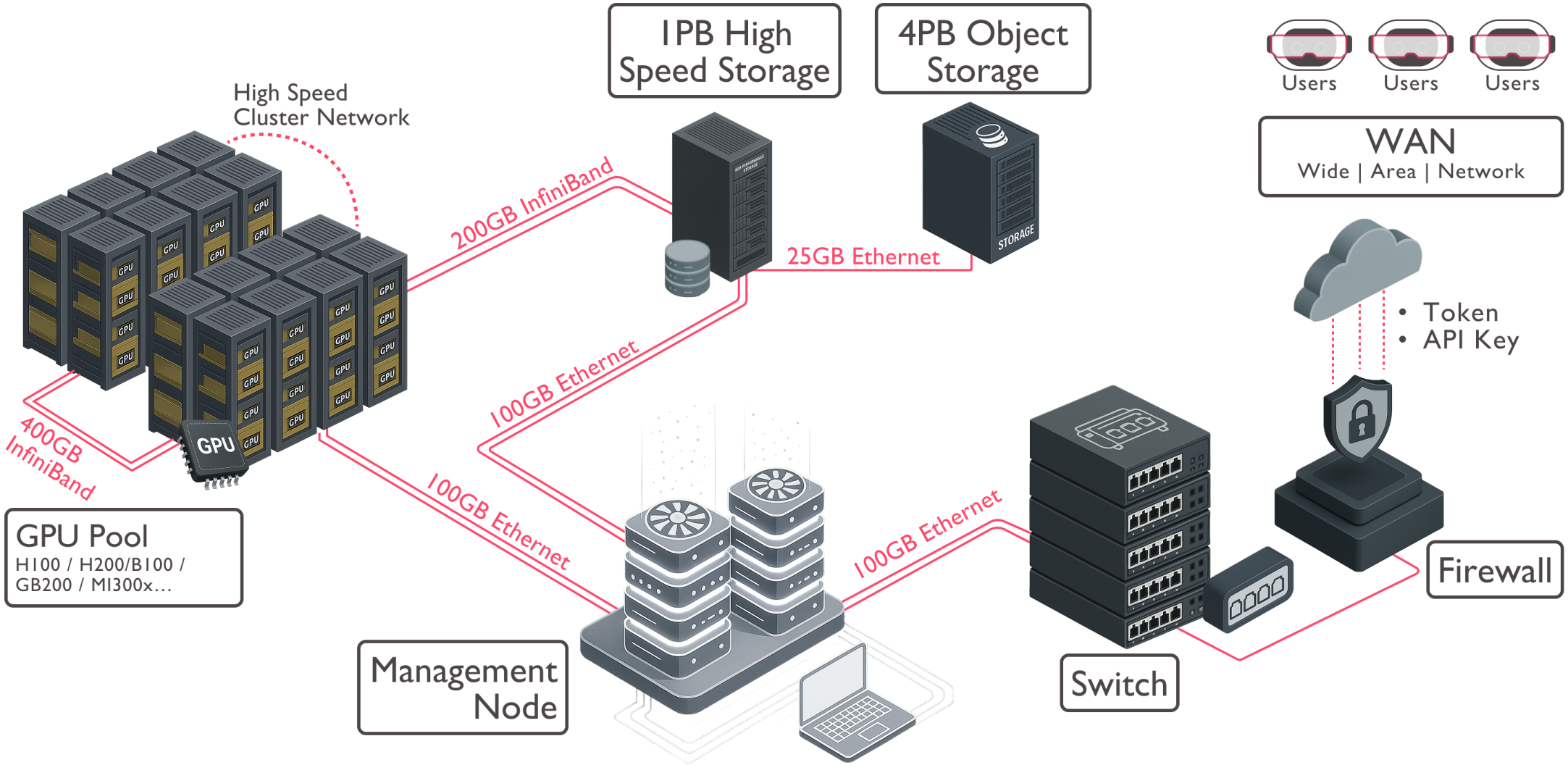
Step Four

Agent AI

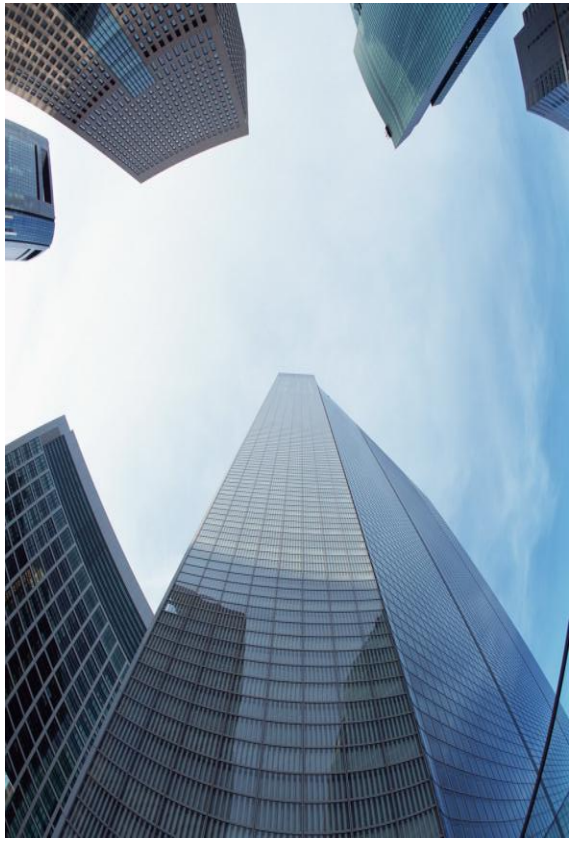
AI Agent，指的是能夠自主做出決策、完成行動，而且毋需人類介入的人工智慧。

AI算力管理全面看向AI基礎設施管理

High-Performance Computing Architecture



AI基礎設施建置的考慮要點



混合 (AI+HPC)

同一平台支持AI(人工智慧機器學習訓練)與HPC (High Performance Computing)

算力高低配置

同一平台支持各種算力組合，滿足不同情的算力需求。例如：使用H200進行模型訓練，L40S作為推論服務。

跨平台(NV+AMD)

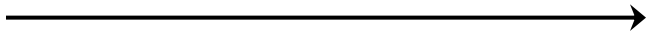
同一平台支持不同廠牌GPU卡片，降低艱深的學習曲線。

GPU 共享從各自為政到協同工作

筒倉式
基礎設施

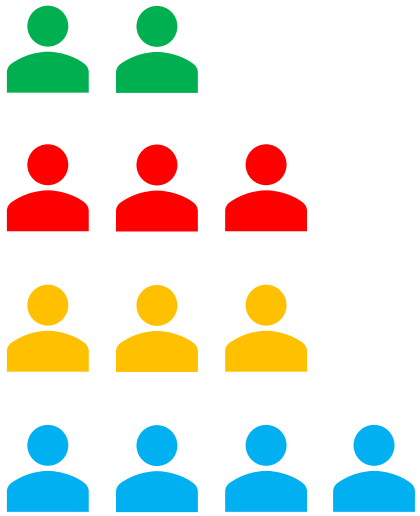
共享
叢集

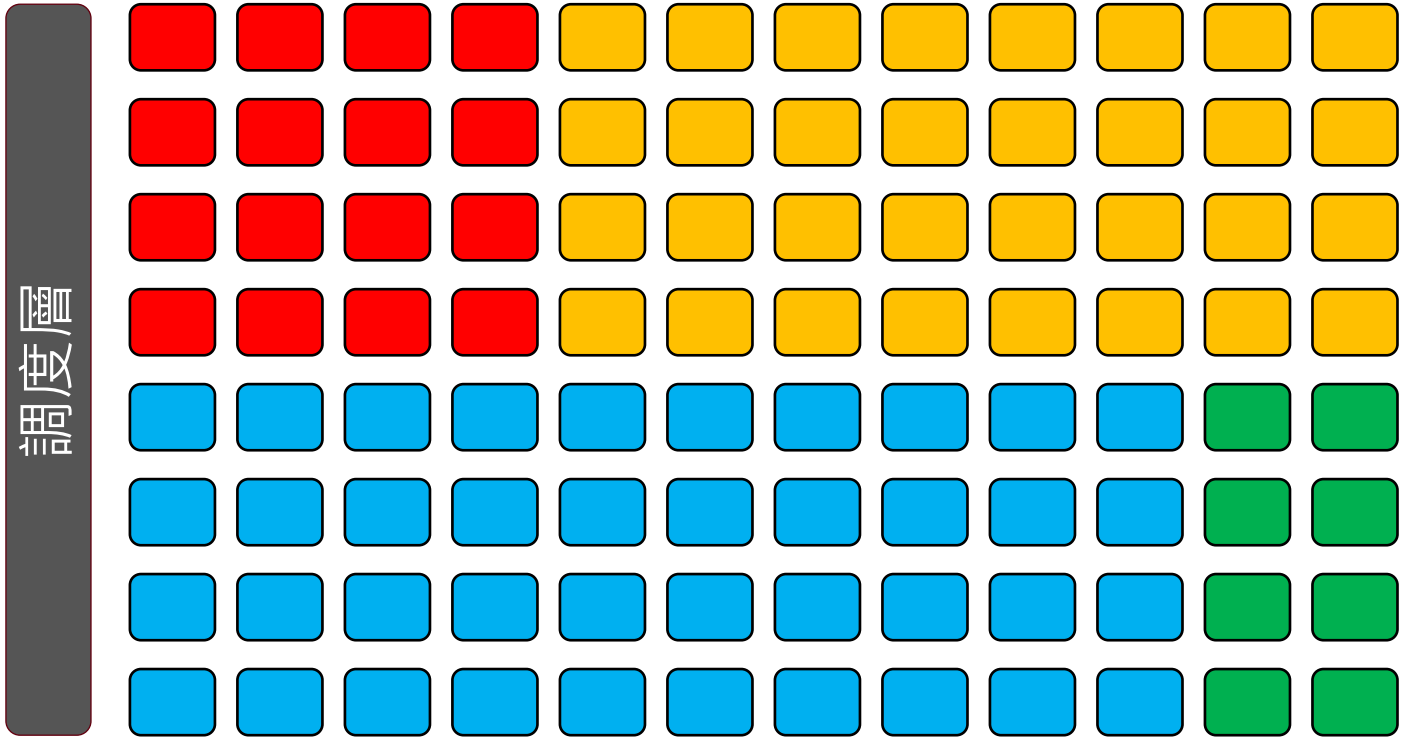
按需
計算



預留
叢集

GPU 池化 + 調度

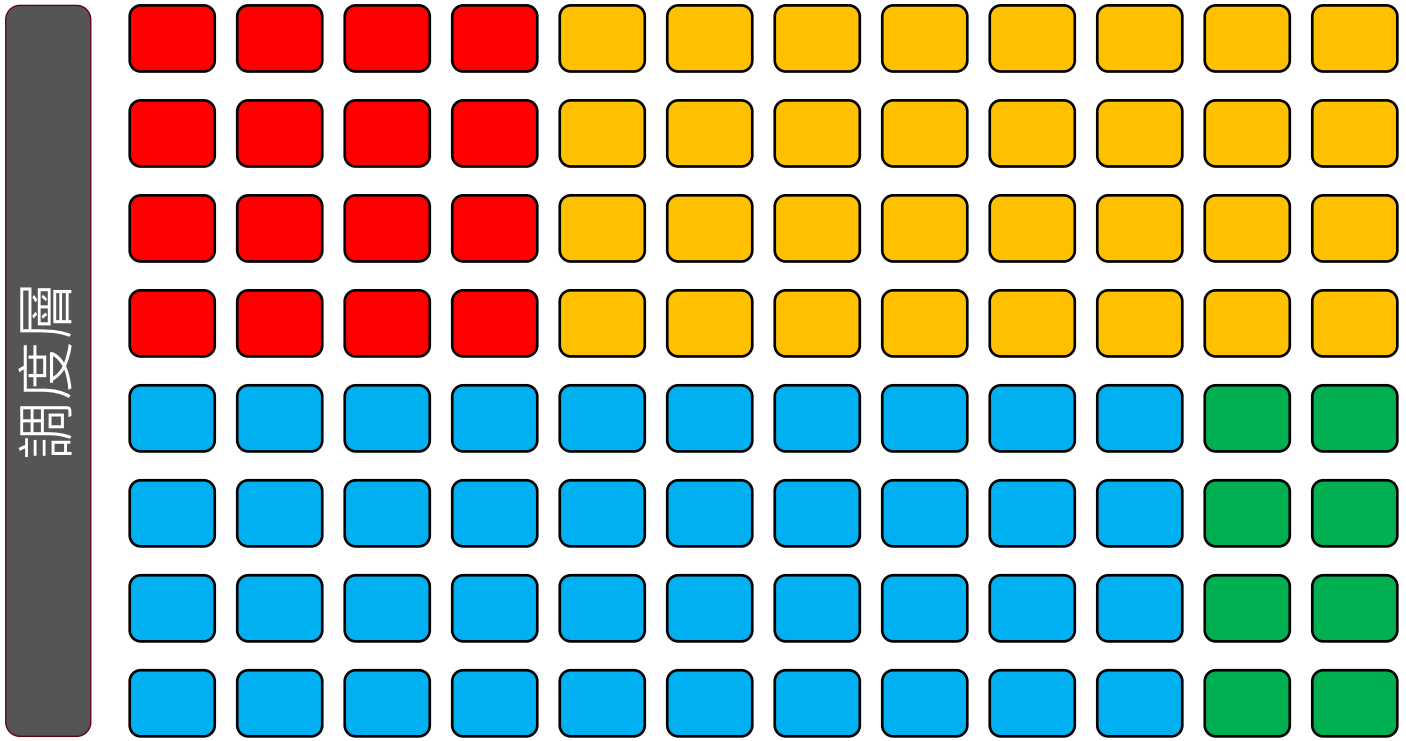




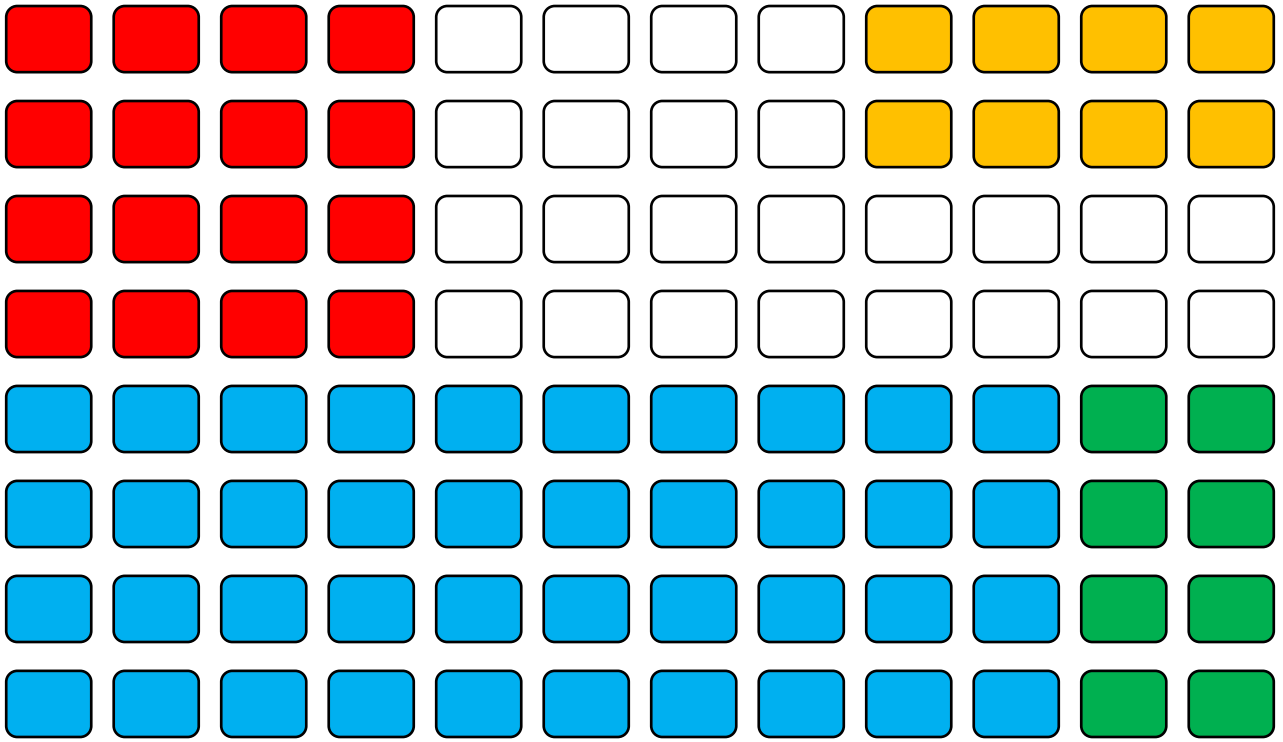
GPU 池化 + 調度

調度層的能力

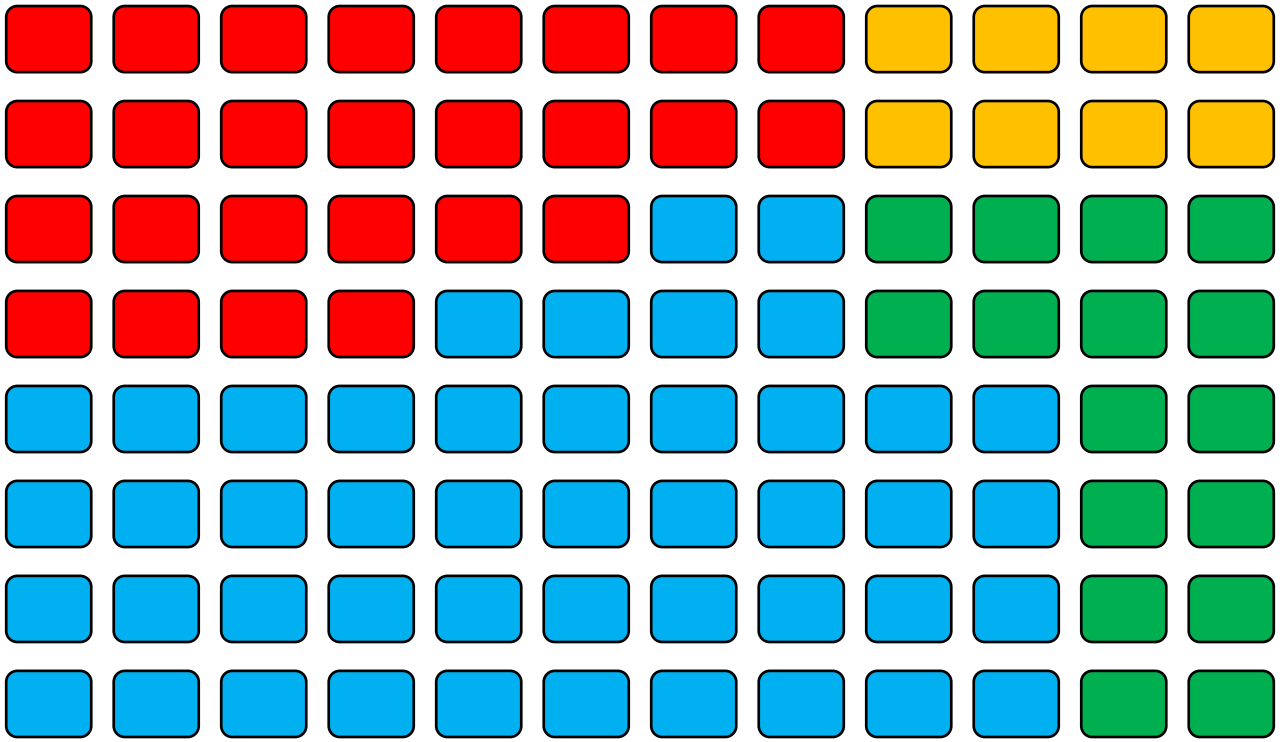
- 資源分配控制
- 工作負載優先級
- 工作排隊
- 工作負載監控和執行
- 審計和報表



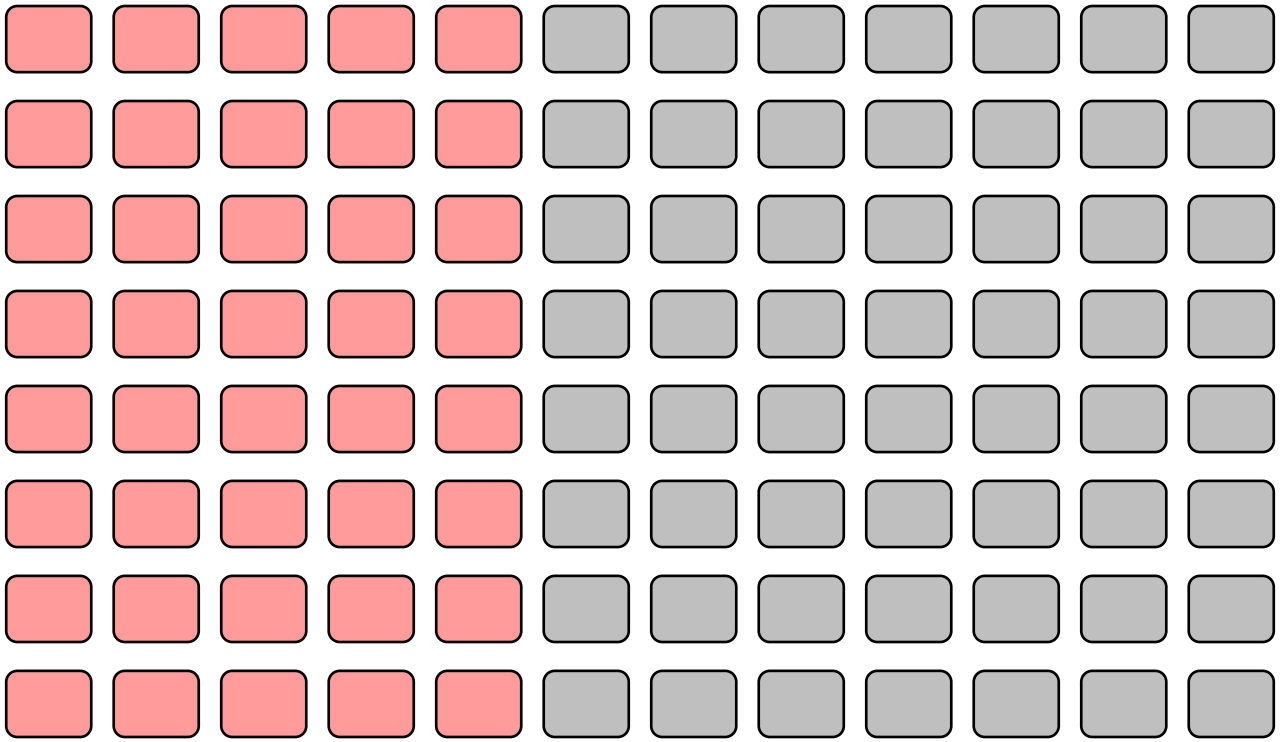
在不同團隊之間重復使用資源



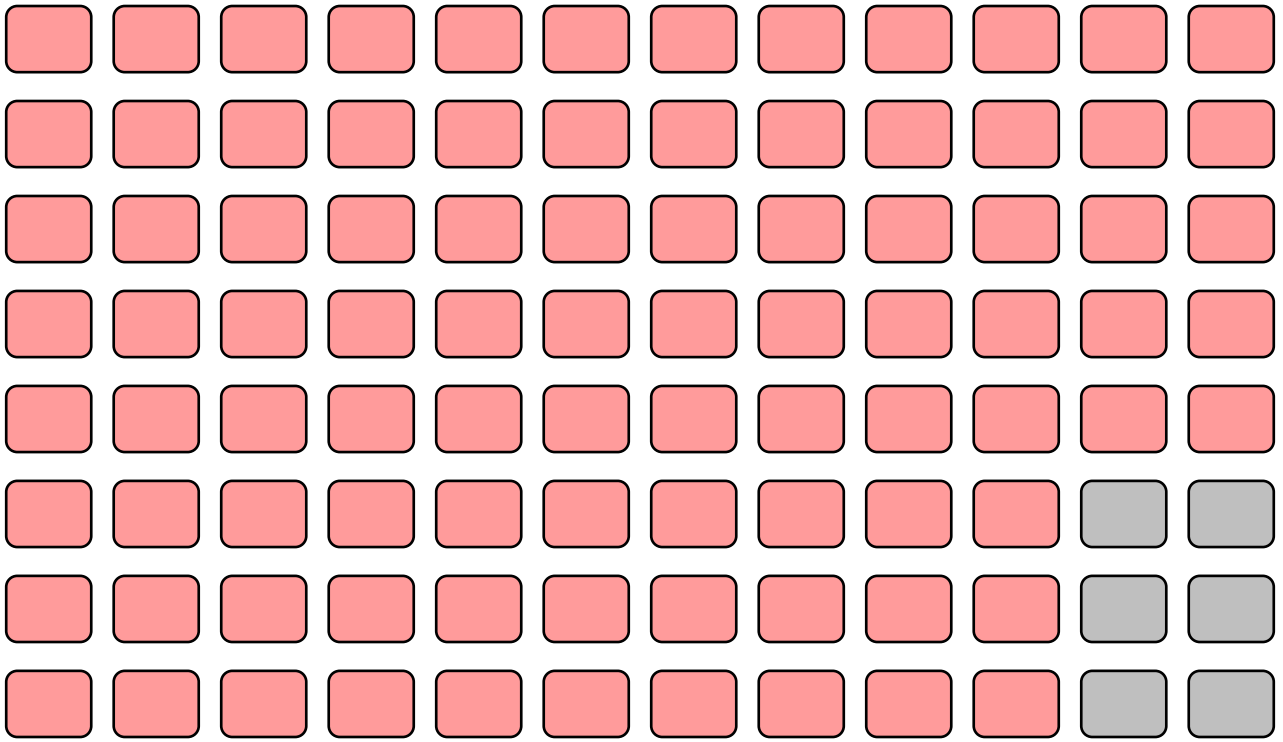
在不同團隊之間重復使用資源



在不同工作負載之間重新利用資源



在不同工作負載之間重新利用資源



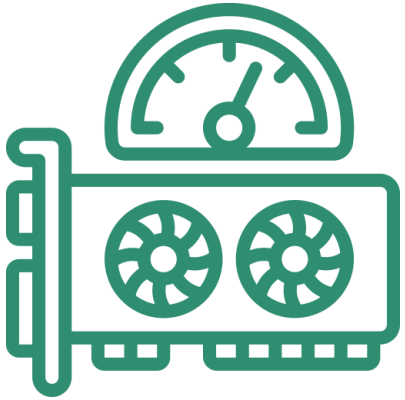
訓練

推論 @ 夜晚

效益

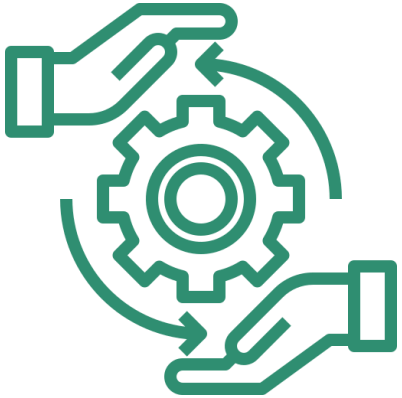
更高效率

通過資源共享和再利用



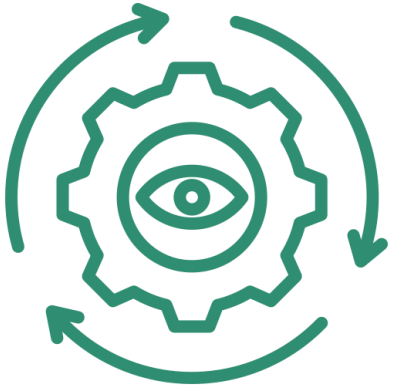
更多 GPU 可存取

用戶更容易獲得更多 GPU，從而提高工作效率



控制與管理

根據業務目標調整資源的能力

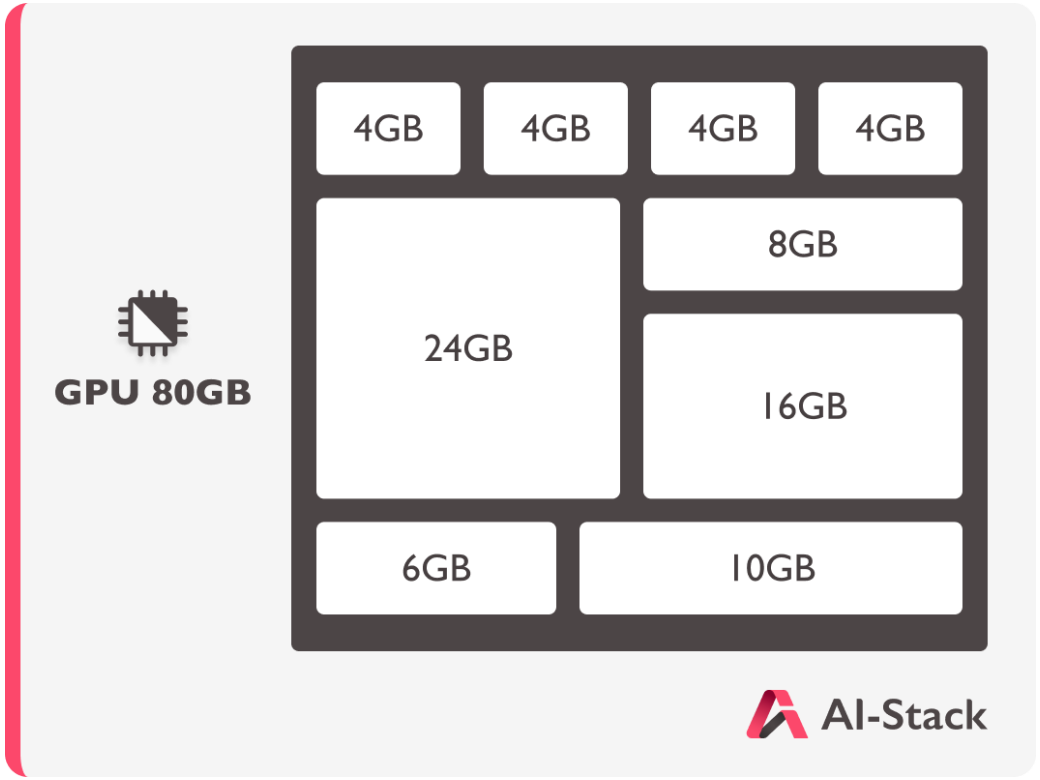


集中可見度

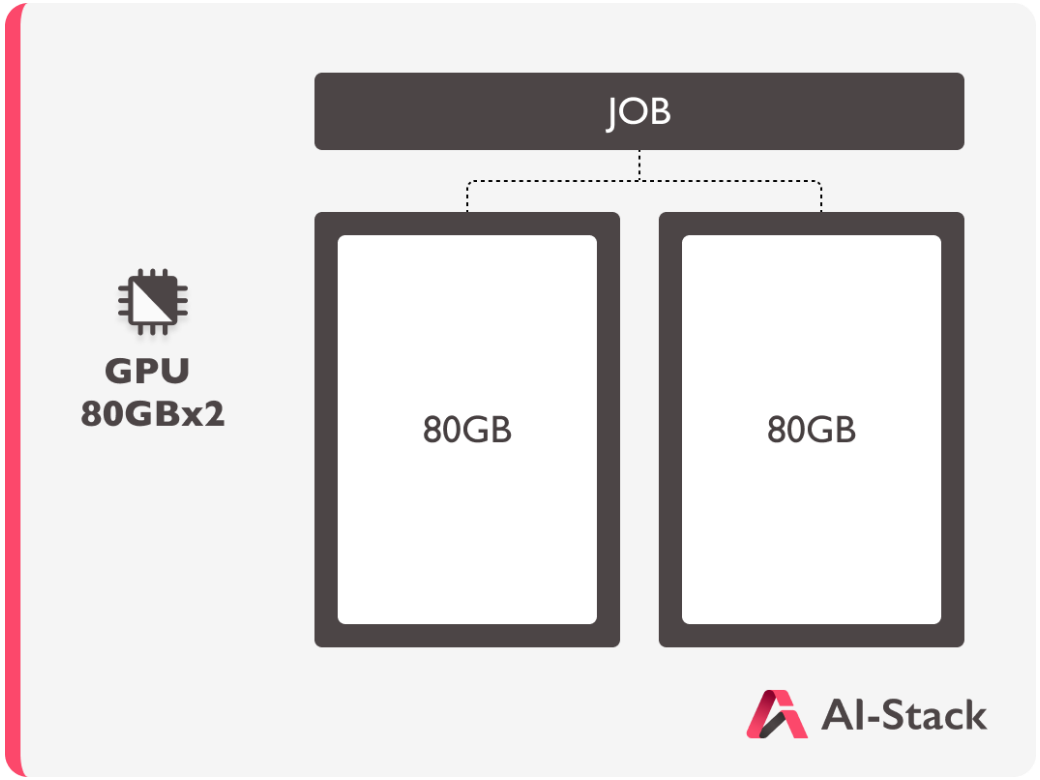
更好的規劃和決策

AI-Stack 加速企業AI應用落地

AI-Stack：GPU切割及聚合技術，效率最大化



One GPU → 使用率 ↑ → Multi GPU Fractions



Multi GPUs → 運算效能 ↑ → One GPU Aggregation

GPU Slices Example

INFINITIX

AI-Stack Slicing

GPUs X 3

1	2	10
10	20	16
10	6	6
5	4	
6		
V100 32GB	V100 32GB	V100 32GB
Container X5	Container X4	Container X3

12 Containers For Training and Inference

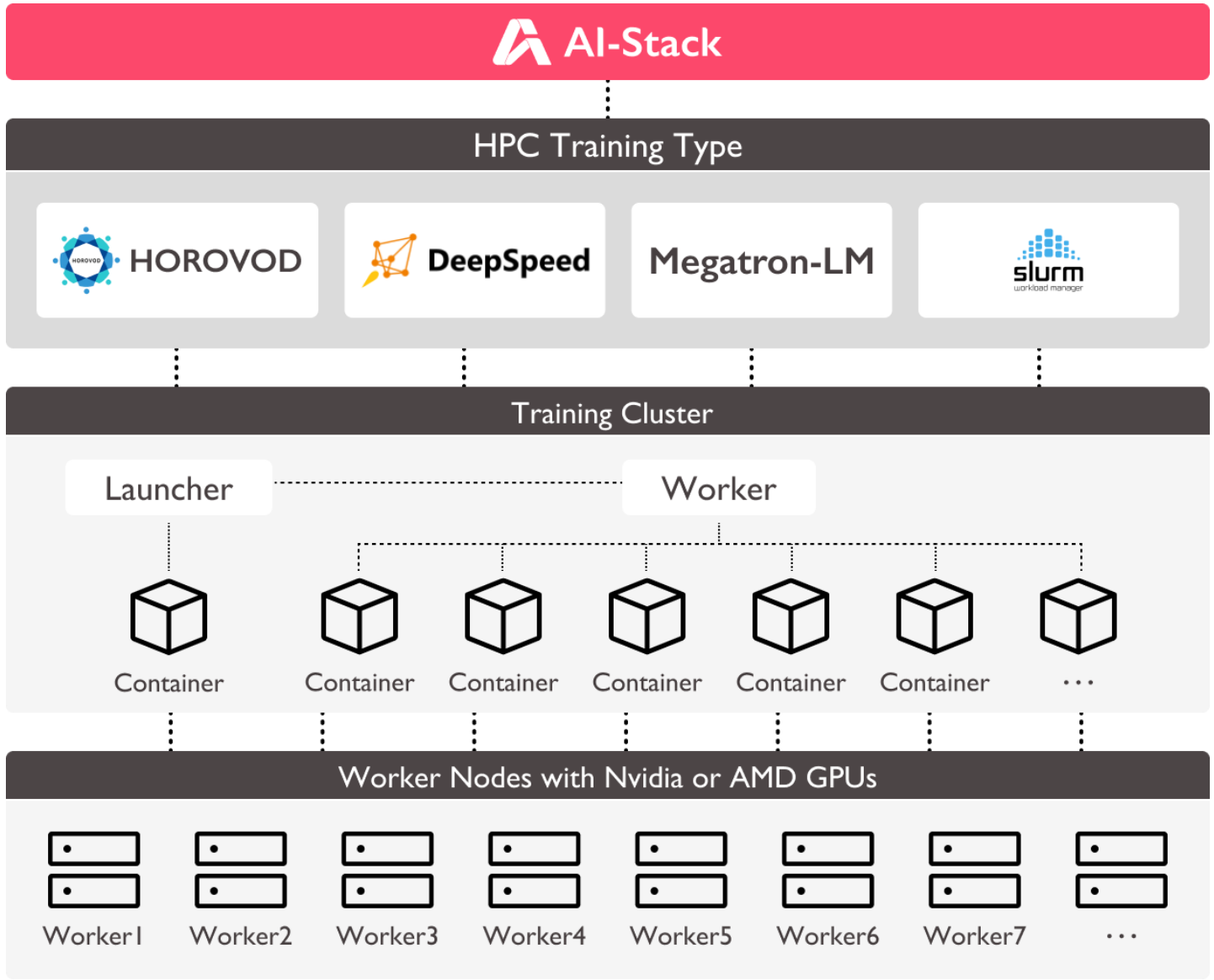
Without AI-Stack

GPUs X 12

AI-Stack 高效能運算

High-performance computing, HPC

平台可根據需求將訓練任務**分配至多個節點運算**，並利用**彈性分布式訓練 (Elastic Distributed Training)**，將多個容器組織成訓練群組，平行分散處理巨量數據，有效**縮減模型訓練時間**，提升運算效率和資源利用率。



AI-Stack Data Center

算力中心整合性解決方案

GPU 伺服器

AI 雲端服務與營運解決方案

營運管理中心

營運管理者

商品管理
集中管理所有模型與服務，組織與維護更輕鬆。

資源與樣板配置
客製運算、儲存及範本，彈性支援各類業務場景。

用戶管理
便捷用戶/租戶註冊、分組及授權。

運營儀表板
視覺化即時監控平台健康狀態與用量。

計費與支付
自動記錄用量、計費與支付，收支一目了然。

稽核記錄
詳細追蹤平台操作，確保資安與合規。

開發者控制台

AI 開發者

隨選即用 AI 運算
即時啟動與管理 AI 運算資源，彈性計費，成本最佳化。

生成式 AI 工作室
探索、測試與整合頂尖 GenAI 模型，支援OpenAI API 及互動工具。

一鍵應用部署
一鍵部署與更新應用，簡化從開發到上線的全流程。

協作與帳號管理
多人協作與帳號權限靈活控管，安全高效。

AI-Stack 分層架構

使用者

資料科學家與人工智慧研究員

- ✓ 專注於研究與模型訓練而非開發維運
- ✓ 僅花 1 分鐘即刻開始AI容器佈署

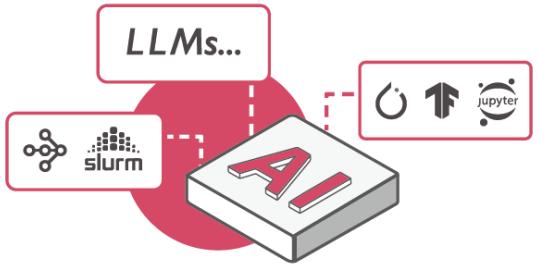
管理者

IT 管理員

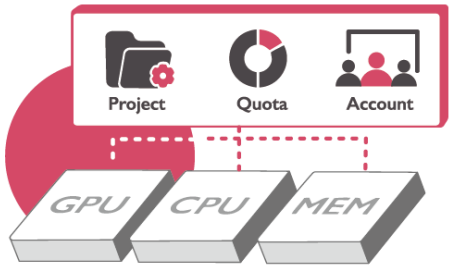
- ✓ 企業級安全防護，資源有效運用
- ✓ 易於監控及完善資源管理

AI-Stack 提供AI開發 最核心的基礎建設

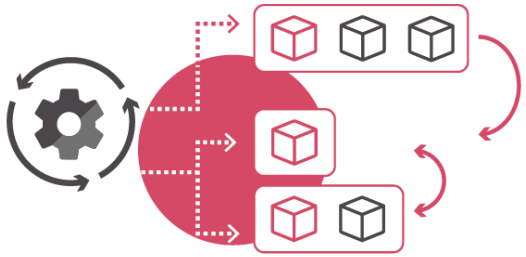
GPU資源管理 + MLOps



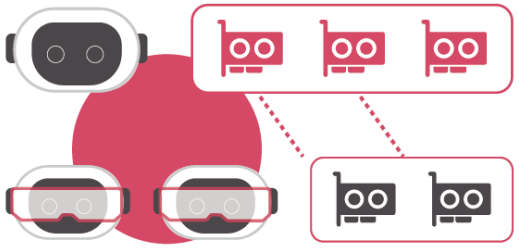
無縫對接常用AI開發工具



資源隔離、權限管理、配額限制

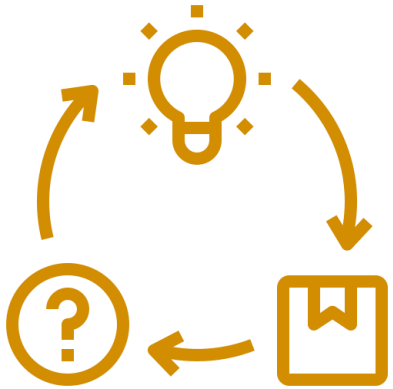


自動執行，預設單次或批次任務



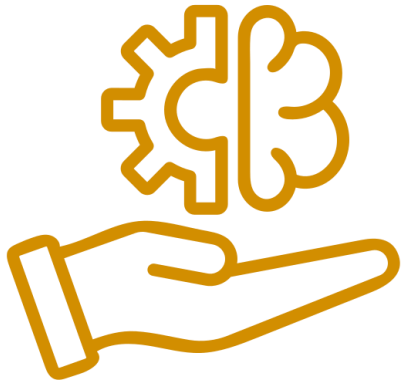
協助團隊更精準、有效管理GPU資源

支持整個人工智能生命週期



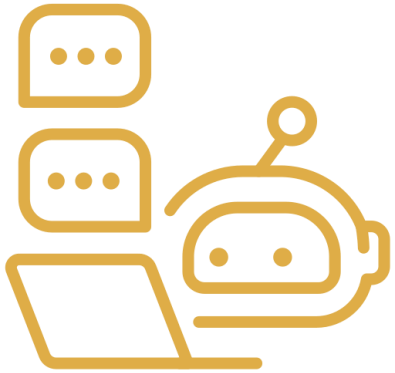
模型開發

使用 Jupyter notebook、VSCode、PyCharm 等集成開發環境工具進行開發和調試



微調與訓練

以批處理作業的方式運行長時間的模型調整或訓練工作負載



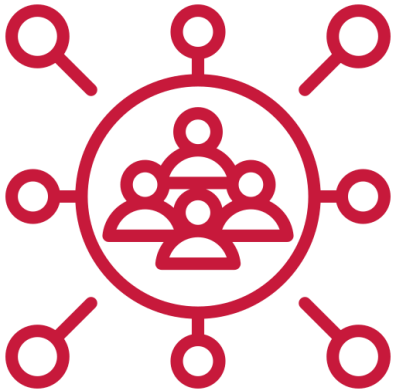
提示工程

通過提示工程實驗語言和 GenAI 模型

服務生產環境

在生產環境中部署模型，為業務應用提供服務

運營人工智能基礎設施平台的關鍵



資源共享

將 GPU 集中到一個集群中，以簡化管理並提高效率

工作負載調度

根據業務目標重新利用資源並確定工作負載的

GPU 切片

在同一基礎設施上運行更多程式編輯器或推理服務器

工具和整合

支持整個人工智能生命週期，並保持開放性和靈活性，以支持新的工具

AI-Stack 協助企業導入AI時， GPU 效益加乘

10倍↑

GPU 利用率

切割技術讓使用率

30% → 90%

90%↑

工作負載運行

多人多項任務讓效率

↑10倍

1分鐘↓

開發環境建置

縮短開發時間

2週 → 1分鐘

10倍↑

提升投資效益

增加ROI效益

↑10倍

Q & A

35

Connect with us

FB

LinkedIn