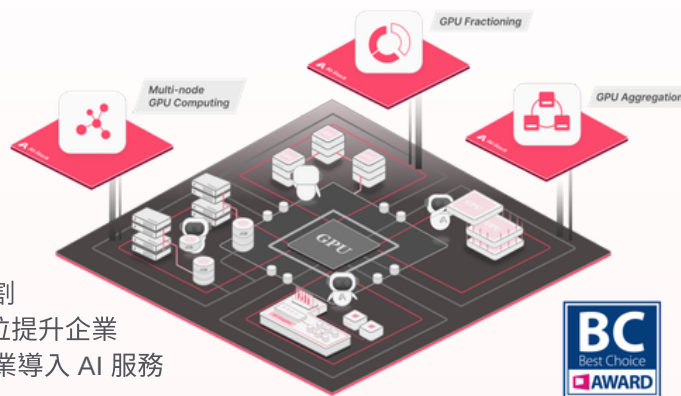


產品介紹

異構算力調度及 AI 基礎設施管理平台

業界頂尖的AI基礎設施管理軟體，整合異構算力管控、GPU 切割與聚合、跨節點運算、容器化與MLOps流程等核心功能，全方位提升企業 GPU 使用率與工作流程效率，有效降低維護與部署成本，是企業導入 AI 服務不可或缺的關鍵工具。



● 極致化 GPU 利用率

● 視覺化監控儀表板

● 精準算力資源調度

● 大幅降低IT營運成本

核心功能 / Core Capabilities

1 叢集運算

- 使用Kubernetes與Docker容器化技術管理。
- 支援已發行的全系列 NVIDIA 與 AMD GPU 型號進行管理以確保 GPU 之資源整合與利用。
- 同時管理多台GPU伺服器與混合GPU卡環境。
- 支援多節點運算。

2 快速建立研發環境

- 開發環境支援主流 AI 框架，包含 PyTorch、TensorFlow；推論工具包含 vLLM、Ollama；並可部署主流開源大型語言模型。
- 提供彈性接口，於容器建立時支援鏡像內建之各類開發服務，包含 SSH、Jupyter、JupyterLab、TensorBoard、WebTerminal 及 Code Server。
- 支援NGC AI深度學習框架，支援自定義鏡像。

3 帳號控管、群組資源配額與權限管理

- 可提供 AD、LDAP 整合機制，平台登入頁面提供 LDAP 登入與本機帳號登入選項。
- 提供彈性資源配置，針對不同專案計畫，分別設定資源使用上限；涵蓋範圍包括 GPU 數量、顯存量、CPU 核心數及系統記憶體。
- 具備基於角色之存取控制機制（RBAC）。針對不同使用者職能，定義並區分系統管理者、專案管理者及一般使用者層級之操作權限，以落實權責分立。

4 批次建立、資源共享、儲存服務

- 支援容器批次建立，使用者可一次建立多個相同配置之容器，以利大規模訓練或團隊開發環境快速部署。
- 支援多人使用同一張 GPU 卡；或是單一工作負載分配多片GPU資源，可於同一容器內使用多張GPU進行運算。
- 平台可支援外部儲存後端整合，包含 NFS 網路檔案系統、S3 相容物件儲存（如 MinIO）。

模組	功能		描述
Web 介面	可視化GUI介面		使用者透過瀏覽器即可操作，無需安裝額外用戶端軟體
Dashboard	即時總覽	資源使用監控	查看叢集整體即時資源使用狀況，監控指標包含：GPU用量、CPU用量、RAM用量、磁碟用量、網路用量、節點耗電量
		健康狀態監控	查看叢集整體健康狀態，包含節點健康度統計、鏡像倉庫健康狀態、鏡像倉庫磁碟使用率、GPU溫度分布、整體事件發生次數統計等
機器學習專案	機器學習服務MLS	容器管理	1. 呈現該專案下之所有容器 2. 操作容器的新增與刪除（可由此建立GPU共享的容器，支援 Nvidia 以及 AMD GPU 切割） 3. 檢視容器部署詳細資訊 4. 透過容器啟動服務 5. 檢視容器監控資訊 6. 檢視容器事件記錄
		鏡像管理 - 公用鏡像	查詢所有專案可用的公共鏡像，選擇所需鏡像，用於建立容器、任務、任務樣板
		鏡像管理 - 自定義鏡像	1. 選擇將使用中之容器以快照方式儲存 2. 查詢已儲存之自定義鏡像、可操作鏡像之編輯與刪除
		分佈式訓練叢集	使用者可將模型訓練工作負載分散至多個容器或節點，可手動擴縮容器數量，簡化大規模模型訓練與資源管理流程
		任務管理	1. 呈現該專案下之所有任務 2. 操作任務的新增與刪除 3. 檢視任務詳細資訊、日誌、事件 4. 設定例行執行之任務
	快速容器服務RCS	部署應用程式	透過運用自定義之部署、pod、服務、Secret、ConfigMap、磁碟區、Network Policy、Ingress 來部署Instance，生產快速連結獲取服務。
設定	儲存叢集	儲存叢集管理	1. 建立儲存叢集，支援儲存類型包含NFS、MinIO 2. MinIO儲存裝置：支援單一儲存體容量上限設定、可依據專案設定儲存容量使用上限
	身分來源驗證設定		管理員可於後台設定 LDAP/AD 連線參數，使用者即可使用企業內部帳號登入 AI-Stack 使用者平台
	專案設定	自動審核設定	1. 設定自動審核專案，使專案不須手動審核直接自動開通 2. 自訂自動審核數量上限，當到達數量便關閉自動審核 3. 設定自動審核所需條件，含GPU額度及型號，符合條件的專案可自動核准
	進階設定	系統參數管理	1. 於MLS專案的非共享模式下，可以設定instance預計回收時間，決定服務的GPU使用率為0%多久會自動刪除 2. 可以設定instance預計刪除時間，決定容器或任務最長只能使用多久
系統	報表		將平台各種資訊以報表形式匯出，如：容器用量表、訂單報表以及單一專案報表