AI算力管理與服務

建構跨域 AI 沙盒，落實永續人才培育

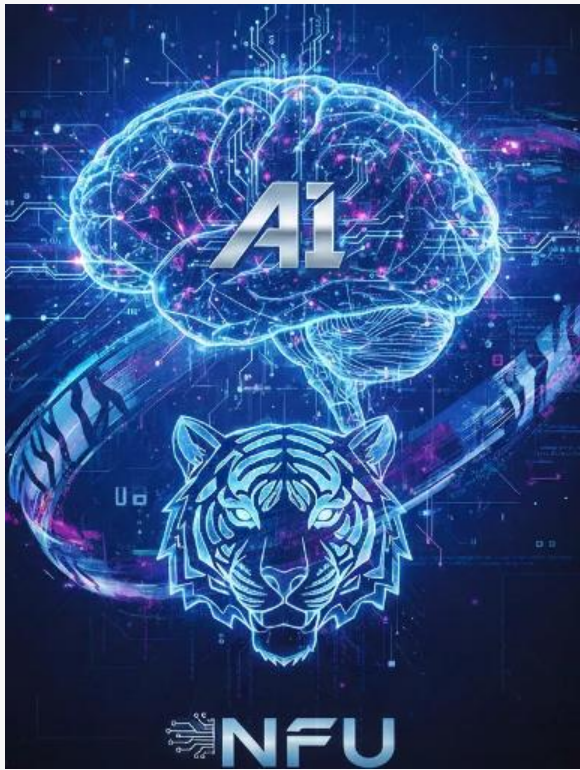
國立虎尾科技大學 (NFU)

自我介紹

google scholar

- 講師姓名：徐茂修
- 現職：國立虎尾科技大學光電工程系副教授
  兼電子計算機中心網路組組長
- 專長：

1.人工智慧
2.深度學習
3.影像訊號處理
4.AI 機器人自動化
5.嵌入式系統
6.物聯網
7.邊緣運算
8.擴充實境/虛擬實境



基礎設施：突破算力瓶頸

現況挑戰

- 設備分散於各系所，資源難以共享
- 電力供應與散熱系統面臨極限
- 單機算力不足，難以支撐大型模型訓練

解決方案

集中化管理

整合 NVIDIA GPU 資源，建立統一調度平台。

高速骨幹網路

優化節點間資料傳輸速率，降低運算延遲。

基建升級

強化機房電力供應與智慧冷卻系統，確保穩定運行。

軟硬體設備組成

核心AI伺服器 (5台)

搭載4張NVIDIA H200 NVL GPU、8張NVIDIA RTX PRO 6000、8張Nvidia RTX 6000 Ada 、4張NVIDIA A100，提供方案的計算核心。

集中式儲存系統 (3台)

Synology 1台 RS3G21RPXS、2台RS2423RP+ ，提供大容量、高效能的AI資料儲存服務。

高速網路交換器 (3台)

Ruckus ICX 系列3台8200，構建高速、低延遲、備援的網路架構。

管理軟體 (1套AI-Stack)

整合GPU切割技術(NVIDIA/AMD)、GPU多片聚合技術、 跨節點運算、直覺性的使用者介面、容器化與MLOps流程、開源深度學習工具、環境部署功能

伺服器及儲存系統規格

CPU/Memory/Storage

Dell PowerEdge XE7745(FOR H200*4)

- CPU: AMD EPYC 9555 (G4C) *2
- RAM: 1152GB DDR5
- HD: 7.68TB NVMe Gen5 (SSD) *8=40T
- M.2 : 960GB (RAID 1) *2=2T

Dell PowerEdge XE7745(RTX PRO 6000*8)

- CPU: AMD EPYC 9555 (G4C) *2
- RAM: 1152GB DDR5
- HD: 3.84TB NVMe Gen5 (SSD)*4=16T
- M.2 : 960GB (RAID 1) *2=2T

CPU/Memory/Storage

WinFast WS2040(Nvidia 6000 ada*8)

- CPU: Intel(R) Xeon(R) Gold 5415+*2
- RAM: 754GB DDR5
- HD: 3.5T*4+1.7T=16T(SSD)

WinFast WS2050(Nvidia A100*4)

- CPU: Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz*2
- RAM: 1.2T DDR4
- HD: 1T*2+14T(HDD)=15T

Storage

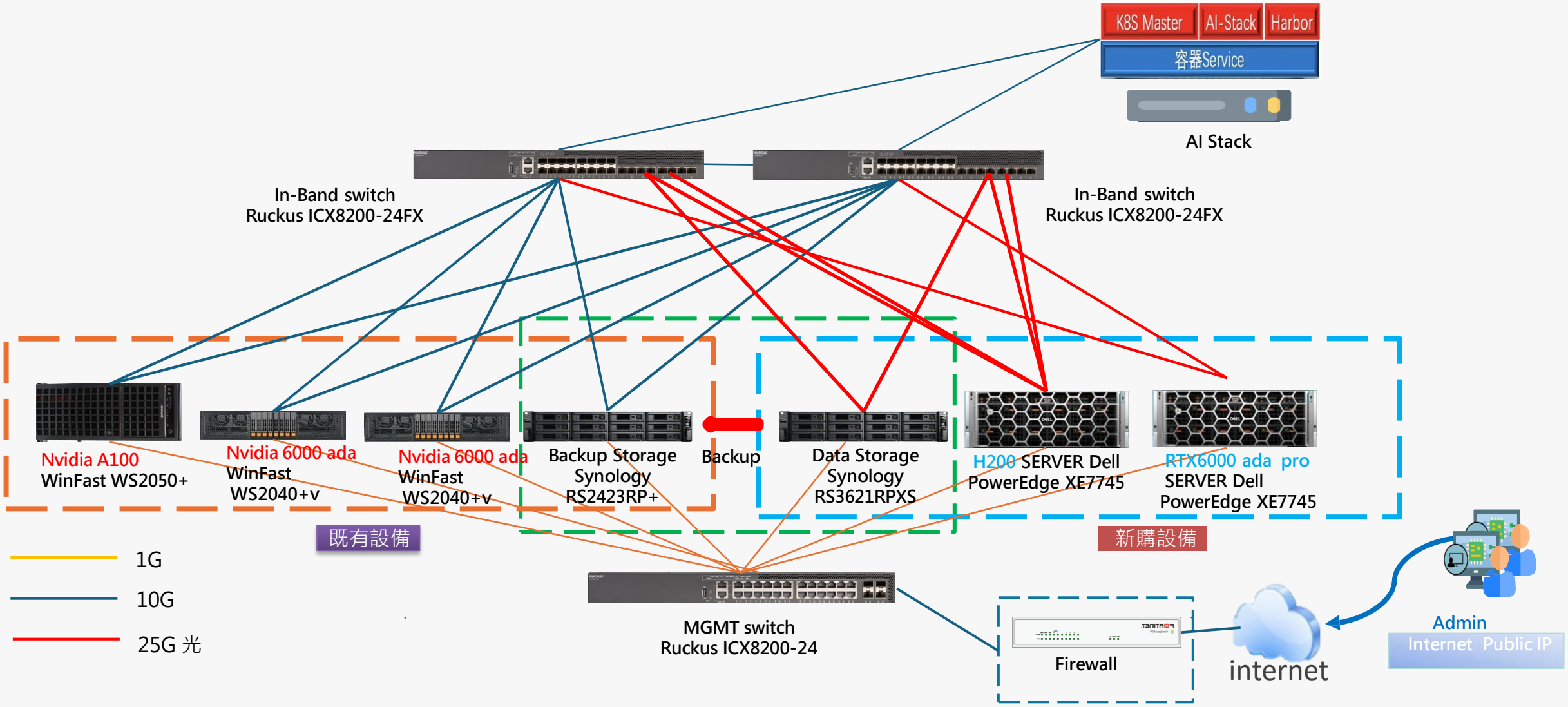
Synology RS3621RPXS

- CPU: Intel Xeon D-1531 (2.2GHz up to 2.7GHz)6核心
- RAM: 16G*2
- SSD: 1.5TB
- HDD: 20T*8=160T

Synology RS2423RP+

- CPU: AMD Ryzen V1780B*1 (64bit/4 core/3.35GHz up to 3.6GHz)
- RAM: 16G*2
- HDD: 8T*10=80T

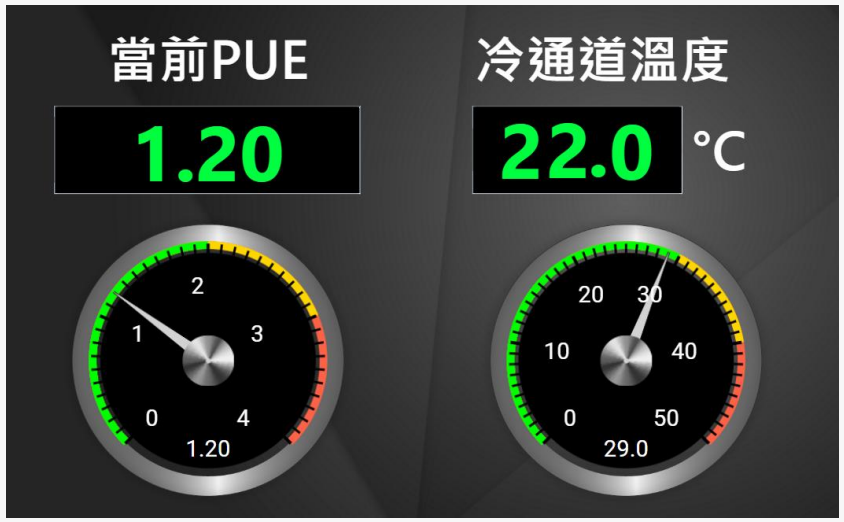
網路架構圖

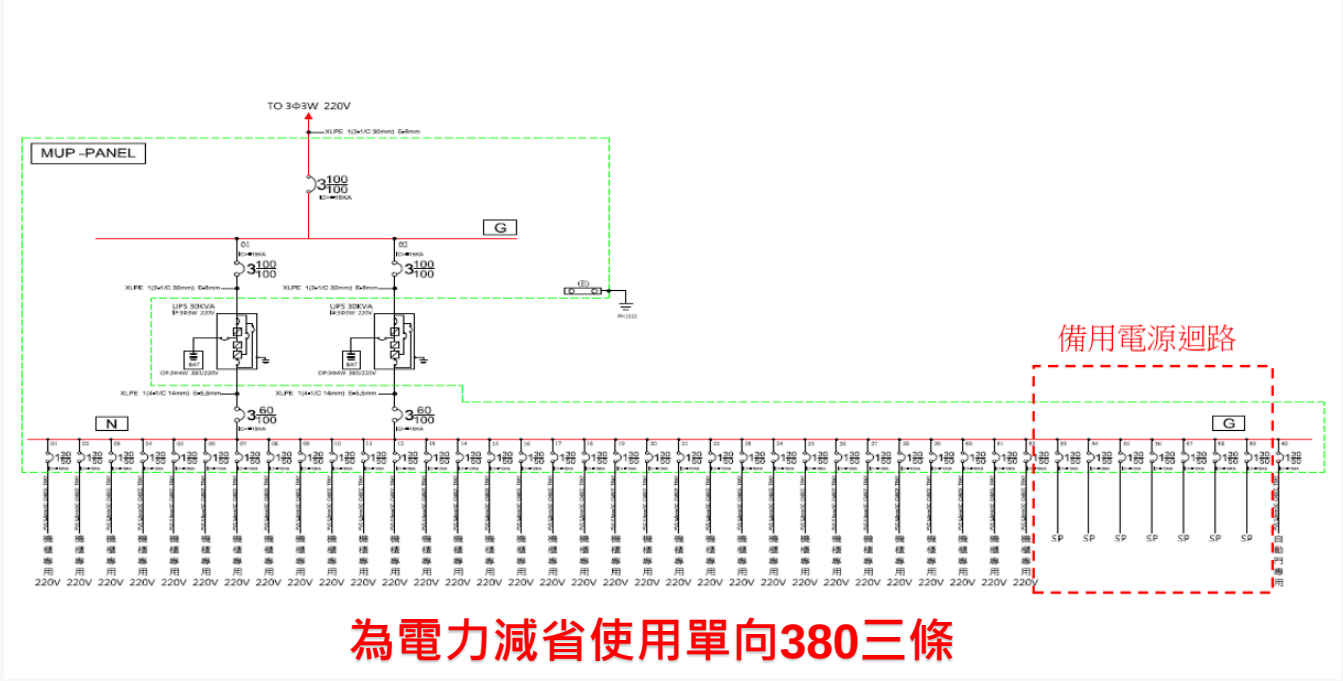

K8S Master
AI-Stack
Harbor
容器Service
AI Stack
In-Band switch
Ruckus ICX8200-24FX
In-Band switch
Ruckus ICX8200-24FX
Nvidia A100
WinFast WS2050+
Nvidia 6000 ada
WinFast
WS2040+v
Nvidia 6000 ada
WinFast
WS2040+v
Backup Storage
Synology
RS2423RP+
Backup
Data Storage
Synology
RS3621RPXS
H200 SERVER Dell
PowerEdge XE7745
RTX6000 ada pro
SERVER Dell
PowerEdge XE7745
既有設備
新購設備
1G
10G
25G 光
MGMT switch
Ruckus ICX8200-24
Firewall
internet
Admin
Internet Public IP

設備機房示意圖及電力

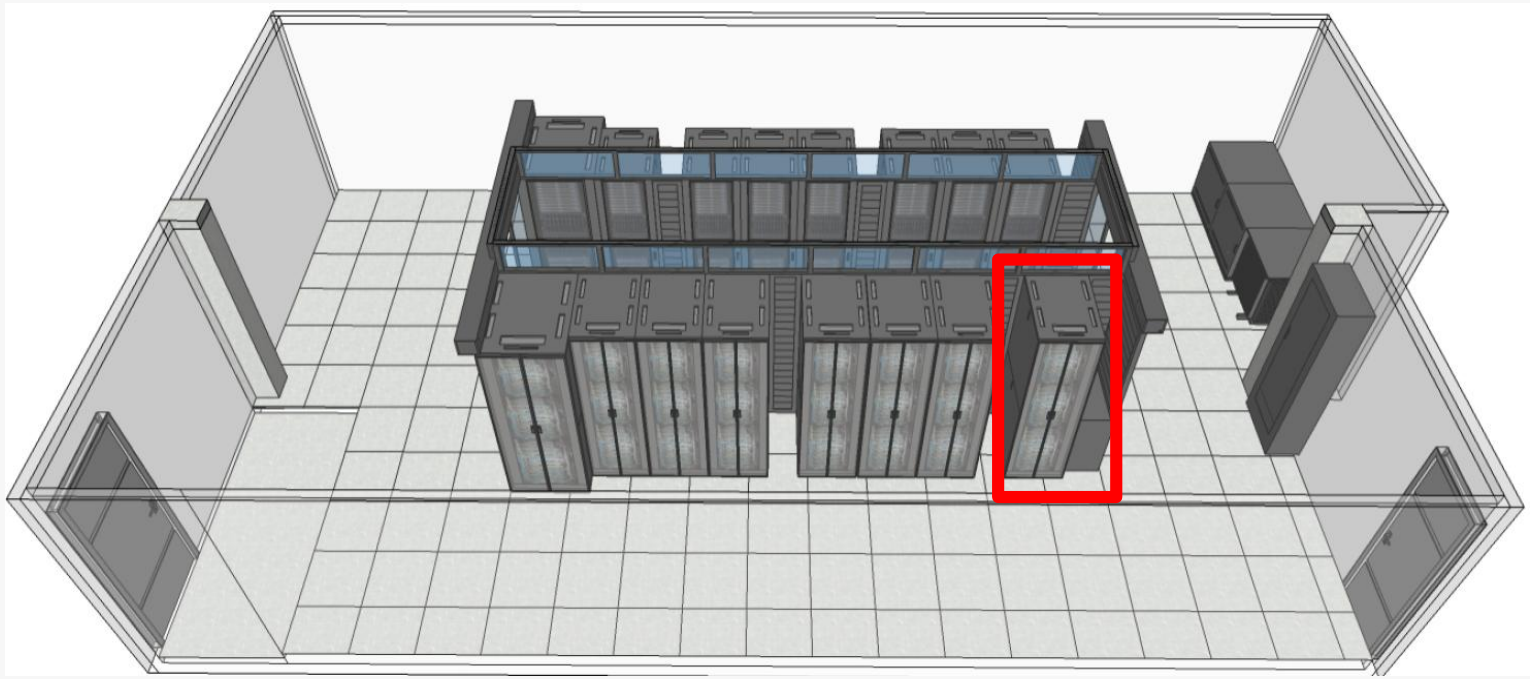
PUE分級

等級	PUE
白金級	＜1.25
黃金級	1.25~1.43
銀級	1.43~1.67
銅級	1.67~2
尚可	2~2.5
不佳	＞2.5





為電力減省使用單向380三條



AI-Stack開發核心基礎架構

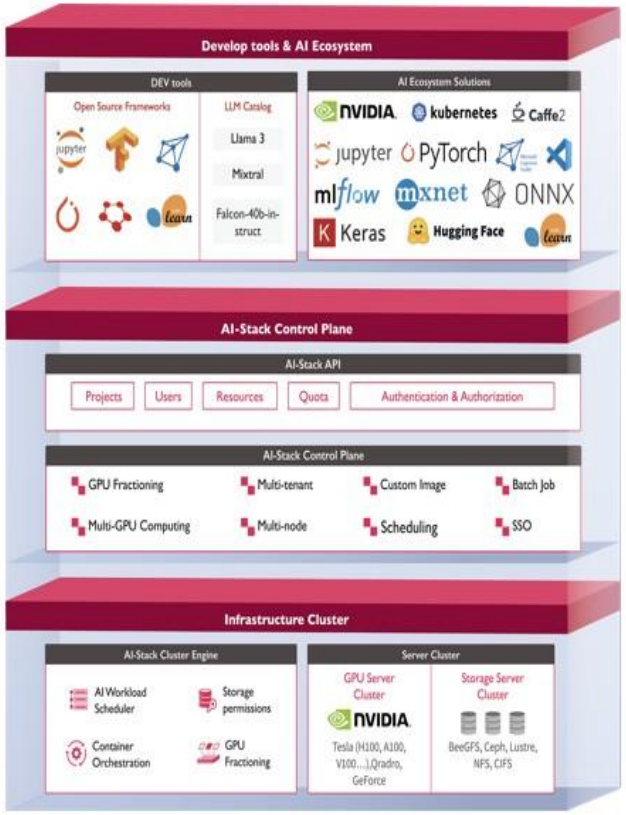
面臨的挑戰與同點

整合困難
與常用的 AI 開發工具整合困難

資源使用不明確
資源使用情況缺乏清晰度，難以辨識使用者及追蹤使用時間

環境建置繁瑣
無法一次性建立相同環境，需手動逐一設置

GPU 排程不足
缺乏靈活的 GPU 排程機制，無法依據工作負載調整 GPU 使用率



AI-Stack 解決方案

無縫工具串接
可無縫串接常用的 AI 開發工具

資源控管
資源隔離、權限管理與配額限制功能

自動化作業
支援單一或批次任務的自動執行

更佳的 GPU 資源管理
協助團隊更精確、高效地管理 GPU 資源

使用者

資料科學家與人工智慧研究員

Data Scientist/ Developer/ AI Researcher

- ✓ 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

IT Administrator

- ✓ 企業級安全防護，資 源有效運用
- ✓ 易於監控及完善資源 管理

教學賦能：AI 學習零門檻

環境優化

即開即用：無需安裝，瀏覽器登入即可開發。

全方位支持: 預載 TensorFlow, PyTorch 等框架。

教學應用

跨學科普及：文管、農學、藝術師生輕鬆上手。

教學協作：統一分發環境，確保進度同步。

雲端沙盒

JupyterHub 整合：提供一致性的開發體驗。

資源隔離：確保每位學生擁有獨立運算空間。

讓 AI 成為每一位師生的基礎工具

動態調度：極大化資源利用率

動態配額 (Quota)

根據課程需求、專題進度實時調整 GPU 分額，確保資源分配與教學目標精準對齊。

排程系統

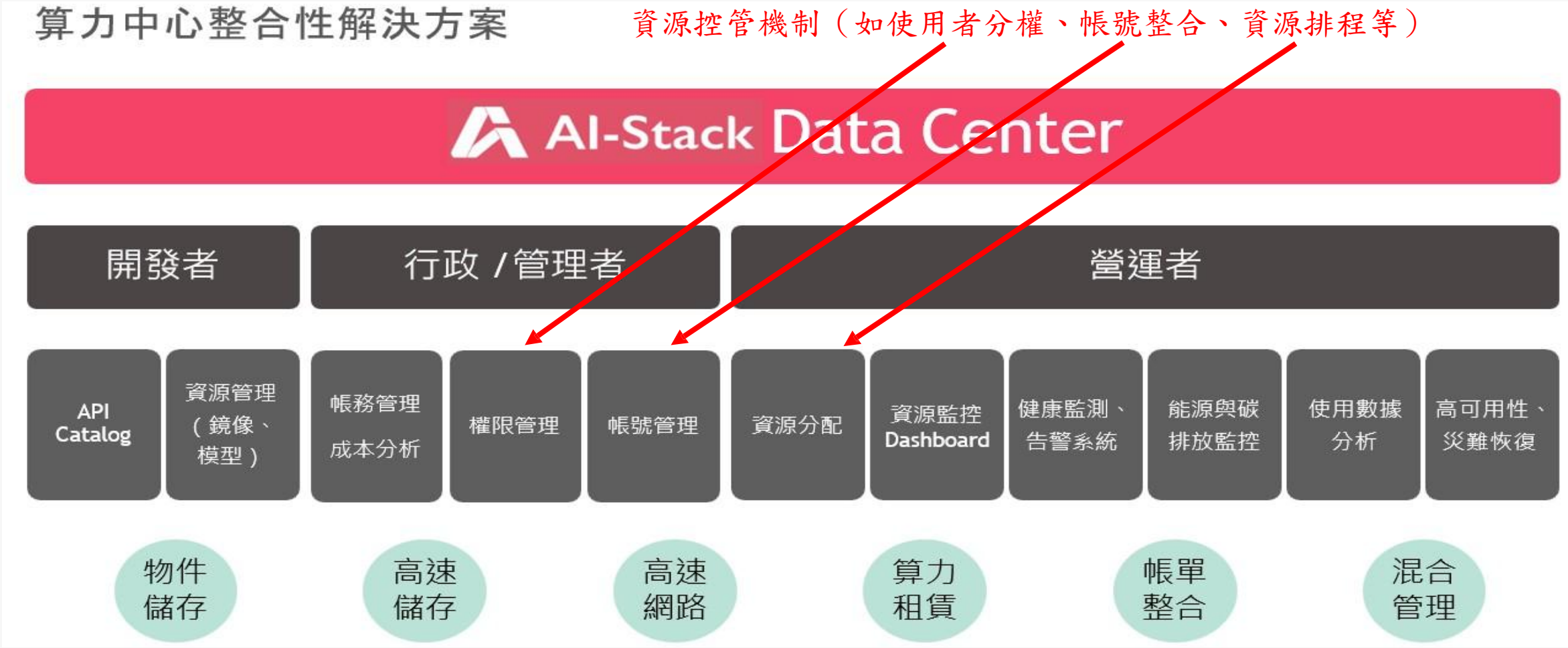
智慧化排隊機制，避免少數使用者長期佔用，確保全校師生使用權利的公平性。

實例應用

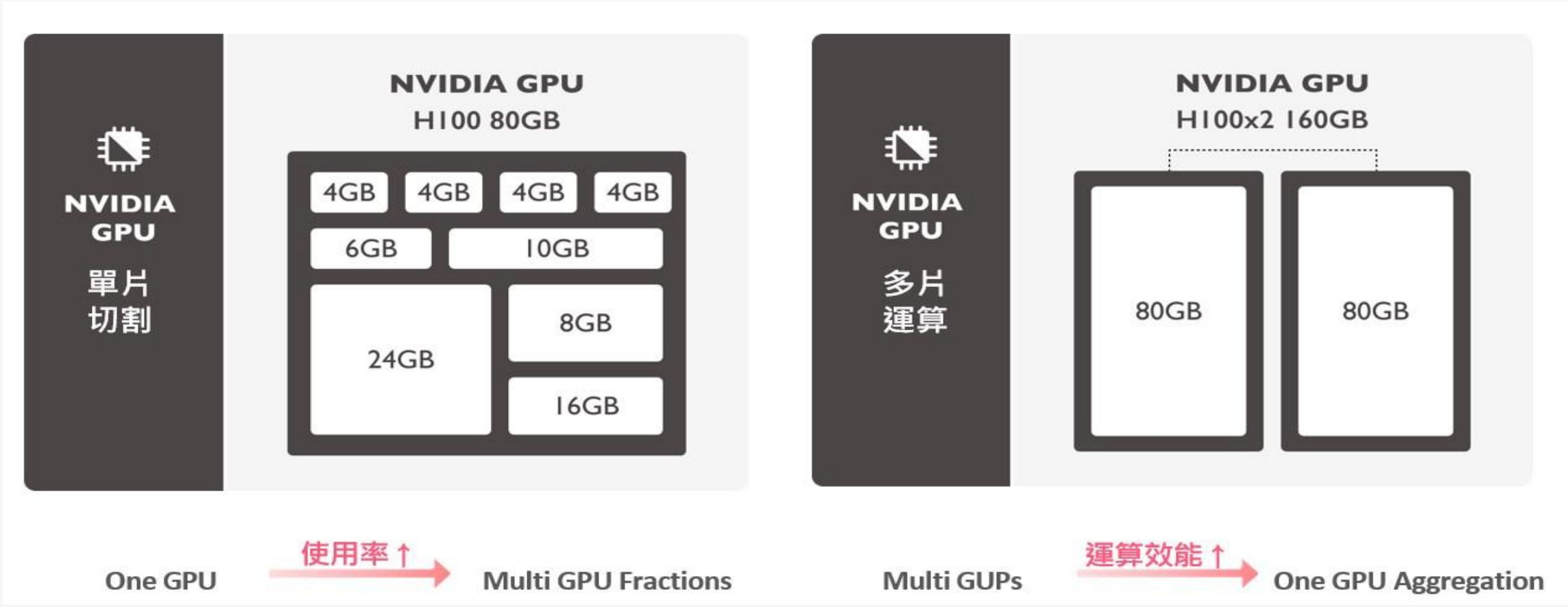
在 500 人大型通識課高峰期，系統自動回收閒置算力，支援高併發運算任務。

核心價值：達成資源「不浪費、不短缺」的精準治理

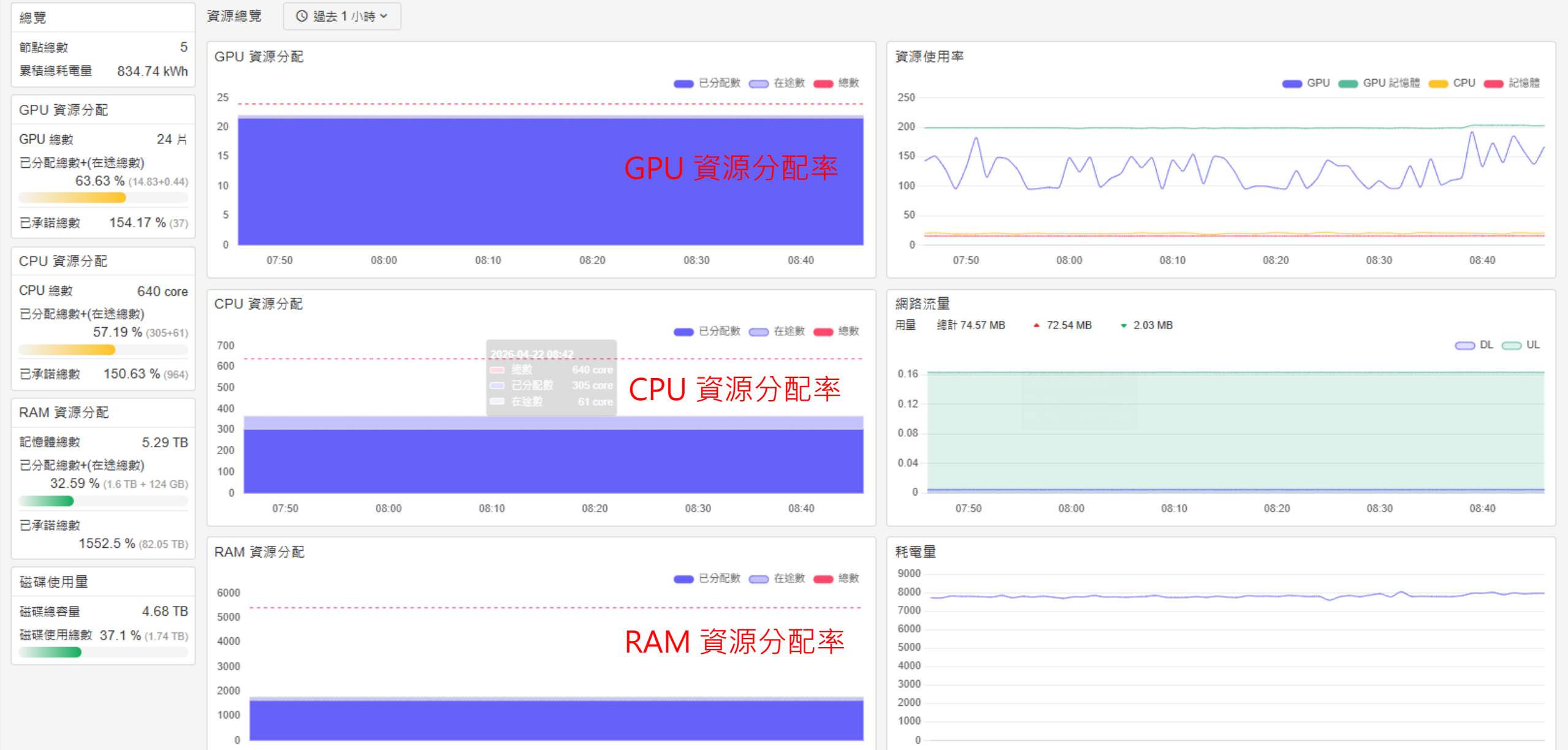
AI-Stack 整合性解決方案



AI-Stack GPU記憶體切割及聚合技術(以H100為例)



叢集資源使用監控


總覽
節點總數 5
累積總耗電量 834.74 kWh
GPU 資源分配
GPU 總數 24 片
已分配總數+(在途總數)
63.63 % (14.83+0.44)
已承諾總數 154.17 % (37)
CPU 資源分配
CPU 總數 640 core
已分配總數+(在途總數)
57.19 % (305+61)
已承諾總數 150.63 % (964)
RAM 資源分配
記憶體總數 5.29 TB
已分配總數+(在途總數)
32.59 % (1.6 TB + 124 GB)
已承諾總數
1552.5 % (82.05 TB)
磁碟使用量
磁碟總容量 4.68 TB
磁碟使用總數 37.1 % (1.74 TB)
資源總覽
過去 1 小時
GPU 資源分配
已分配數 在途數 總數
GPU 資源分配率
資源使用率
GPU GPU 記憶體 CPU 記憶體
CPU 資源分配
已分配數 在途數 總數
CPU 資源分配率
網路流量
用量 總計 74.57 MB 72.54 MB 2.03 MB
DL UL
RAM 資源分配
已分配數 在途數 總數
RAM 資源分配率
耗電量

目前申請使用狀況

AI賦能上課使用

AI賦能上課使用

狀態	使用中
起始日	2026/02/25
結束日	2027/12/31
成員	19 位 (含 2 位管理員)
容器	17 個
任務	0 個

11463110

狀態	使用中
起始日	2026/03/03
結束日	2027/12/31
成員	3 位 (含 3 位管理員)
容器	1 個
任務	0 個

wjlab_test

狀態	使用中
起始日	2026/01/09
結束日	2026/07/09
成員	12 位 (含 7 位管理員)
容器	2 個
任務	0 個

醫學萃取訓練

本專案專注於開發鼻實 CBCT 影像之 3D 特徵萃取模型。透過深度學習（Deep Learning）技術，對高解析度之 3D 醫療影像進行特徵標記與模型權重訓練，以達成自動化醫學影像辨識之目的。

狀態	使用中
起始日	2026/03/03
結束日	2026/05/31
成員	5 位 (含 5 位管理員)
容器	1 個
任務	0 個

RTET火箭隊

我們團隊想要借用RTX 6000 PRO進行火箭的應力分析及模擬，以便後續火箭的應力分析及設計。

狀態	使用中
起始日	2026/03/06
結束日	2026/08/31
成員	5 位 (含 2 位管理員)
容器	0 個
任務	0 個
GPU 額度	

LLM測試

LLM測試

狀態	使用中
起始日	2026/02/25
結束日	2049/12/31
成員	3 位 (含 3 位管理員)
容器	1 個
任務	0 個
GPU 額度	

生成式指紋

進行生成式指紋模型行之相關研究

狀態	使用中
起始日	2026/03/18
結束日	2026/12/31
成員	2 位 (含 2 位管理員)
容器	0 個
任務	0 個

ReasonDrive-VLA

利用nvidia ar1實現在emulator carla上 會使用到參數量大的模型以及使用到不管是nvidia 提供的數據集做微調 以及使用實驗室的電動巴士實車錄製的資料集做台灣道路場景的訓練以及微調（...

狀態	使用中
起始日	2026/03/27
結束日	2027/03/22
成員	2 位 (含 2 位管理員)
容器	1 個
任務	0 個
GPU 額度	

總算力統計方式

為了方便評估，我們將效能分為單精度 (FP32) 以及針對 AI 訓練最關鍵的半精度 (FP16/Tensor) 來進行計算。

型號	數量	單卡 FP32	單卡 FP16 (Tensor)	總 FP32 算力	總 FP16 算力	顯示記憶體 (VRAM)
H200	4	67 TFLOPS	1,979 TFLOPS*	268 TFLOPS	7,916 TFLOPS	564 GB (141GB×4)
A100 (80GB)	4	19.5 TFLOPS	624 TFLOPS*	78 TFLOPS	2,496 TFLOPS	320 GB (80GB×4)
RTX 6000 Ada	8	91.1 TFLOPS	1,457 TFLOPS*	728.8 TFLOPS	11,656 TFLOPS	384 GB (48GB×8)
RTX 6000	8	16.3 TFLOPS	130.5 TFLOPS*	130.4 TFLOPS	1,044 TFLOPS	192 GB (24GB×8)
總計	24			1,205.2 TFLOPS	23,112 TFLOPS	1,460 GB

行政面應用

知識整合

將歷年法規、SOP 手冊、會議紀錄全面數位化，打破資訊孤島。

餵養 AI

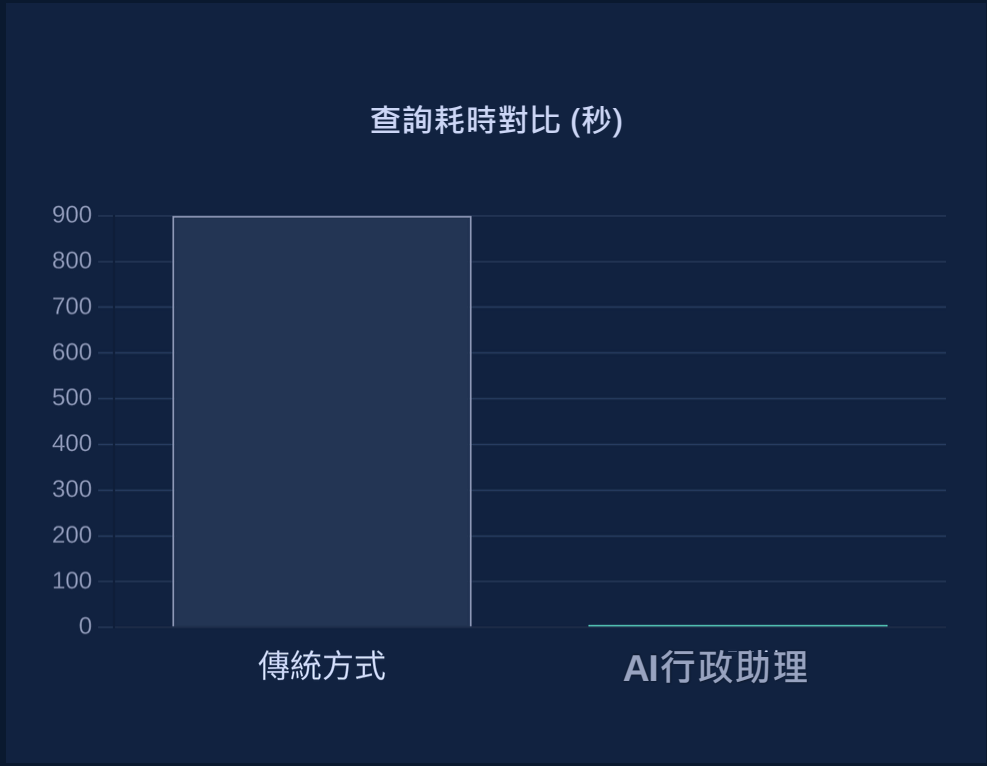
將結構化與非結構化資料導入 LLM，建立專屬行政知識大腦。

智慧檢索

從關鍵字搜尋進化為語意詢問，數秒內獲得精確答案與條文引用。

實戰案例：差旅規定查詢

項目	傳統方式	AI行政助理
查詢動作	翻閱厚重 PDF/紙本	**自然語言詢問**
定位精準度	需人工逐頁比對	**精確定位條文**
處理時間	約 15 - 20 分鐘	**約 5 - 10 秒**
結果產出	手動節錄資訊	**自動生成摘要**



註：數據基於行政庶務常見文件檢索場景估算

技術架構：在地化與高效能

Open WebUI

直覺友善的操作介面
降低行政人員使用門檻
支援多樣化外掛擴充

Ollama

確保資料隱私不外流
在內部環境穩定運行
快速切換多種開源模型

本地 Gemini 4

強大的多模態理解力
處理超長文本與複雜文件
提供精準的邏輯推理

隱私安全　混合部署　極速回應

導入 AI 的預期效益：行政轉型

效率飛躍

70% +

處理繁瑣事務的速度顯著提升，大幅縮短作業週期。

錯誤率降低

精準無誤

減少人為翻閱漏看或記憶偏差，確保行政作業準確性。

知識傳承

數位資產

將資深經驗轉化為 AI 知識庫，降低人員交接與培訓成本。

價值重塑

專業升級

行政人員專注於流程優化與服務品質，提升職涯價值。

永續產學：培育 AI 實戰人才

產學對接

業界規格開發環境

提供與企業對標的算力資源，吸引高品質產學合作專案。

企業場域實作

學生直接在 AI-Stack 平台進行實戰演練，縮短學用落差。

永續發展

人才永續

建立「學習-研究-實作」閉環，培育具備即戰力的 AI 人才。

資源永續

透過平台化營運延長硬體生命週期，達成資源效益最大化。

學習 → 研究 → 實作 永續循環

將算力轉化為國立虎尾科技大學的核心競爭力

高效　易用　公平　永續

—Q&A—

持續深化產學合作，打造頂尖 AI 人才搖籃