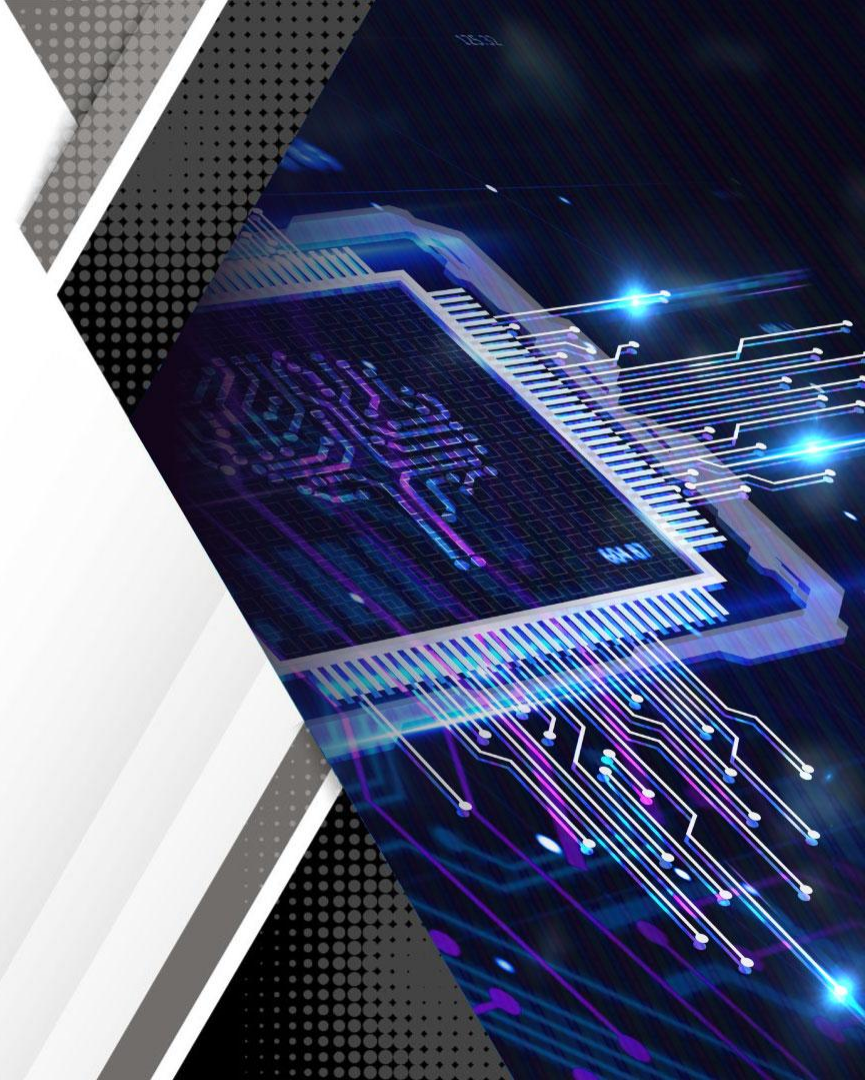
ALTOS
an Acer Group Company

# 生成式 AI 推論平臺佈署與安全治理

安圖斯科技 總經理
李占傑 (Jackie Lee)



# 個人介紹

ALTOS
an Acer Group Company

**工作經歷：**

2018 ~ 目前: 安圖斯科技總經理

2000 ~ 2018: 宏碁公司企業產品線資深處長

1999 ~ 2000: 友訊科技研發工程師

**學歷：**

1995 ~ 1997: 國立陽明交通大學資訊管理研究所

1991 ~ 1995: 國立陽明交通大學資訊工程學系

Agenda

# 大綱

01 生成式AI的發展歷史與軌跡

02 AI從模型訓練到推論

03 管理AI推論平臺佈署的挑戰（含資安）

04 Altos所提出的解決方案

# 01

# 生成式AI的發展歷史與軌跡

# 生成式AI發展階段：從實驗室到企業應用

ALTOS
an Acer Group Company

01

## 萌芽期（2014-2019）：學術研究奠基

2014-2019年為生成式AI萌芽期，學術界主導研究，GANs（Generative Adversarial Network 生成對抗網路）、VAE（Variational Auto-Encoder 變分自編碼器）等基礎架構問世，奠定生成模型理論基礎。

02

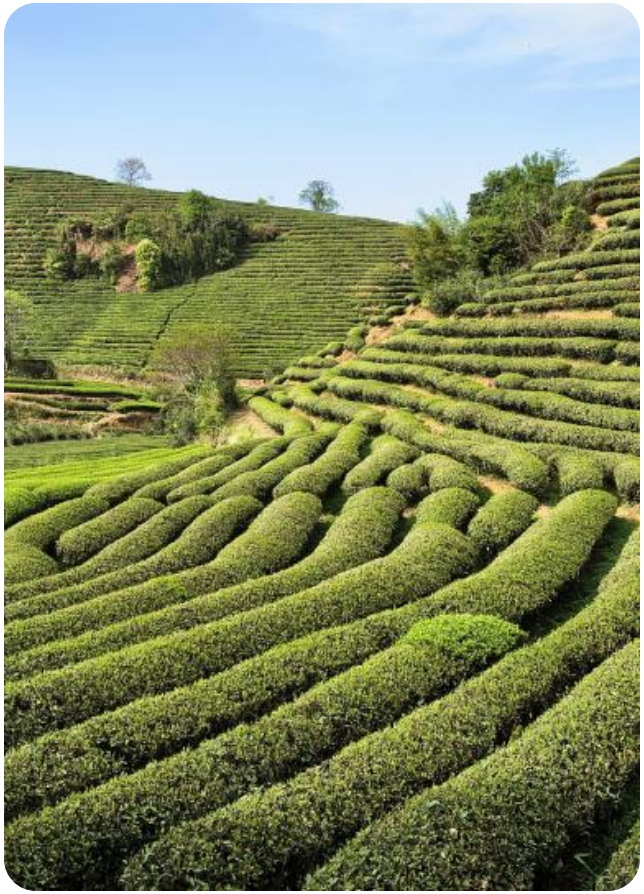
## 驗證期（2020-2022）：大模型爆發與PoC探索

2020-2022年進入驗證期，大語言模型快速反覆運算，GPT系列、Llama等問世，企業開始透過PoC（概念驗證）探索商業應用可行性。

03

## 應用期（2023-至今）：產業落地與價值釋放

2023年起邁入應用期，生成式AI從實驗室走向產業落地，早期佈署企業已將其融入產品研發、客戶服務等核心業務，獲得運營效率提升。



# 產業驅動力：為何現在行動?

## ROI窗口開啟：早期採用者快速獲益

ROI（投資報酬率）視窗已開啟，研究顯示早期採用生成式AI的企業，6個月內技術使用率超過70%，顯著提升生產力。

## 法規壓力升高：合規需求驅動**私有部署**

法規壓力持續升高，全球數據主權與隱私法規（如歐盟GDPR、個資法）趨嚴，公有雲API因數據出境風險難以滿足企業合規需求。

## 競爭加速：領導者已構建技術壁壘

產業競爭節奏加快，領導企業已構建私有生成式AI推論環境，在產品創新、成本控制等方面形成競爭壁壘，後進者需加速佈局。

# 02

# AI從模型訓練到推論

# 訓練 vs. 推論：企業的核心關注點

ALTOS
an Acer Group Company

## 成本結構：訓練的「一次性投入」vs 推論的「持續支出」

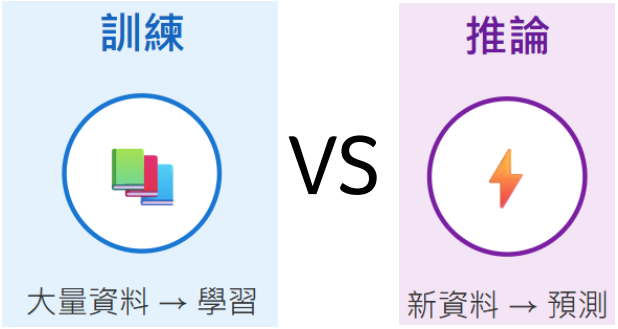
模型訓練需大量算力（如GPT-3訓練耗電量等同300輛車年耗電），屬前期高投入；推論則需長期運維，佔企業AI總成本60%以上。

## 資源需求：訓練的「算力密集」vs 推論的「效率優先」

訓練依賴大規模GPU集群（如1000+ A100）**追求精度**；推論需優化**回應速度**（毫秒級延遲），滿足實時業務（如客服機器人）。

## 業務目標：訓練的「模型研發」vs 推論的「商業落地」

訓練聚焦模型性能（參數量、準確率）由研究團隊主導；推論將模型轉化為業務工具（如行銷文案生成），直接影響營收與客戶體驗。




比喻:

- 訓練 = 學生讀書、準備考試
- 推論 = 學生快速答題

# 推論：企業AI價值的「最後一哩」


## 更快佈署：縮短「模型到應用」的落地週期

傳統推論佈署需2-4周環境配置，企業面臨業務需求與技術落地的時間差；**高效平臺可將佈署週期壓縮至小時級，快速響應市場變化。**


## 更低TCO：優化推論全生命週期成本

推論佔企業AI總成本60%-70%（已投入運營、使用量大的企業 AI 場景中），含算力、人力與維護費用；**需通過動態擴容與模型壓縮降低單位元推理成本**，提升投資回報率。


## 更強安全：確保「數據不出境」的治理要求

**企業推論面臨數據洩露風險，需符合歐盟GDPR、個資法等法規**；安全治理需實現模型隔離、數據加密與訪問審計，保障數據隱私與主權。

# 03

# 管理AI推論平臺佈署之挑戰（含資安）

# 挑戰1：效能與成本平衡




## 複雜查詢響應延遲的用戶體驗瓶頸

生成式AI複雜查詢響應延遲超過2秒時，用戶體驗顯著下降，影響業務流程連續性與使用意願。

## GPU資源利用率不足的浪費問題

企業獨立佈署時GPU利用率常低於30%，閒置資源無法有效調度，導致硬體投資回報率低下。

## 獨立佈署與雲端API的成本負擔

獨立佈署需承擔硬體採購與維護成本，雲端API按調用計費累積年成本可達百萬級，中小企業難以負荷。

# 挑戰2：合規與數據安全

**跨國數據存儲的法規遵從難題**

跨國企業使用公有雲API時，數據出境可能違反當地數據駐留法規（如歐盟GDPR），面臨罰款風險。

**未授權訪問的模型與數據濫用風險**

缺乏嚴格訪問控制時，未授權用戶可能竊取模型權限或篡改數據，導致商業機密洩露或錯誤輸出。

**操作軌跡追溯不足的監管審查挑戰**

模型調用與數據處理軌跡未完整記錄，難以滿足金融、醫療等行業監管審查的可追溯性要求。

# 挑戰3：佈署與管理複雜性

ALTOS
an Acer Group Company

## 多層技術棧整合的技術門檻

企業需整合GPU驅動、Kubernetes容器編排、Triton模型服務器等多層技術，需專業團隊維護，技術門檻高。

## 多租戶環境下的資源動態分配衝突

多租戶共用資源時，CPU/記憶體/GPU的動態分配易引發資源爭奪，導致部分任務延遲或失敗。

# 04

# Altos所提出的解决方案

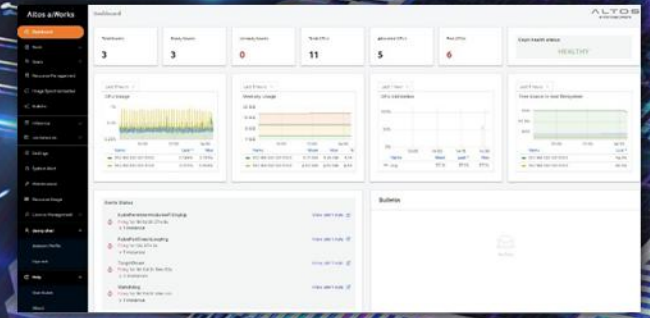
# Altos aiWorks – an AI Computing Platform

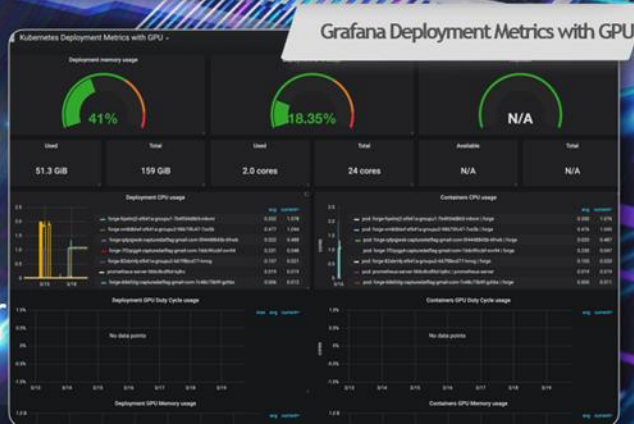
ALTOS
an Acer Group Company

One-Stop Hardware and Software Integrated Solution


ALTOS

Altos BrainSphere™
Server/Workstation




Grafana Deployment Metrics with GPU

Altos Accelerator Resource Manager

# Altos aiWorks Advantages

ALTOS
an Acer Group Company

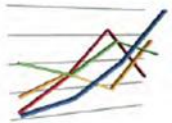
## Simplification
**Deployment Processing**

Developers can more focus on their development of products without facing any hardware resource matters.

## GPU Resource Mgmt.


Optimizing not only one GPU but multi GPUs, aiWorks can manage many clients using one GPU or a single job process on multi GPUs.

Providing AI Resource analysis console which is easy for managers to know AI cloud computing status.



## AI Resource Analysis

## AI Development Flow Integration

aiWorks solution provides not only data preprocessing and development & training, but also integrates candidate model and model validation to optimize developing flow and maximize ROI.

## Browser Interface

For managers who didn't have experience building AI environments, using the browser to set up is more convenient than coding brick by brick.

## Remote Access

Through browsers to remote access, providing a sustainable and utilizing AI developing environment everywhere.

## Rapidly deploy developing environments

Built-in common AI developing applications on aiWorks and can quickly establish a developing environment with one click.

TensorFlow jupyter python
NVIDIA CUDA
PyTorch ONNX Chainer
Keras theano MATLAB

## Private Cloud Mgmt.

Replacing AI GPU workstations to GPU Server not only centralize GPU resources, and user groups and scale-out when resources are facing limitation, but saving management cost is a major key factor.

# 搭配完整Altos AI伺服器與工作站產品線

ALTOS
an Acer Group Company

AI Model Size (Parameter)

Training (FP8/16)

Inference (FP4)

**Altos BrainSphere™ R880 F7 (8x B200)**
- 2.3TB HBM3e, **540~612B (FP16)**
- **72 petaFLOPS (FP8)** training and **144 petaFLOPS (FP4)** inference

**Altos BrainSphere™ R785 F6 (8x H200 NVL)**
- 1,128GB HBM3e, **420~480B (FP16)**
- **32 petaFLOPS (FP8)**

**Altos BrainSphere™ R680 F7 (8x H200 NVL)**
- 1,128GB HBM3e, **420~480B (FP16)**
- **32 petaFLOPS (FP8)**

**Altos BrainSphere™ R380 F7 / R385 F6 (2x RTX PRO 6000 Max-Q)**

**Altos BrainSphere™ GB10 F1**
- 128GB LPDDR5X, **200B** (FP4)
- **1000 AI TOPS (FP4)**

**Altos BrainSphere™ P150 F10 / P330 F6 SE (2x RTX PRO 6000 Workstation)**
- 192GB GDDR7, **300B ~ 360B** (FP4)
- **7022 AI TOPS (FP4)**

**Altos BrainSphere™ P130 F10 (RTX 5090)**
- 32GB GDDR7, **50 ~ 55B** (FP4)
- **3352 AI TOPS (FP4)**

**Computing Power (FLOP / AI TOP)**



Remark: All performance data are theoretical peak values for comparison purposes only.

# 核心定位：企業級生成式AI推論平臺

ALTOS
an Acer Group Company

## 定位：企業與學術界專屬解決方案

專注企業與學術界需求，針對生成式AI推論場景量身打造，兼具靈活性與穩定性。

## 目標：解決實務痛點，驅動業務創新

聚焦模型佈署效率、資源利用率與安全治理等痛點，實現AI技術落地與業務價值轉化。

## 核心：硬體+軟體生態一體化整合

整合大模型訓練架構、資源調度、多卡運算與安全治理，提供端到端完整解決方案。

# 解決方案1：高效部署與資源治理


## 12週快速部署路線圖

分三階段實現：W1-W4基礎設施就緒，W5-W8服務佈署整合，W9-W12上線優化，確保快速落地。

## GPU資源共用技術創新

透過Nvidia Multi-Instance GPU (MIG)將每個高階GPU拆分最多7個執行個體、Multi-Process Service (MPS)支援多任務共用，突破GPU資源瓶頸，提升利用率3-7倍。

## 佈署效率與資源利用率雙提升

縮短佈署週期同時最大化資源效益，解決傳統佈署耗時與資源浪費問題，降低總擁有成本。

# 解決方案2：多層次安全架構

ALTOS
an Acer Group Company

## 傳輸安全：端到端加密防護

採用TLS 1.3加密協議，確保數據在傳輸過程中全程安全，防範中間人攻擊與數據洩露。

## 身份管理：精細權限與統一認證

基於RBAC三級角色劃分，整合LDAP/AD，嚴格控制訪問權限，杜絕未授權操作。

## 操作審計：全程日誌與合規追溯

完整記錄所有操作日誌，滿足法規合規審計要求，實現安全事件可追溯、可調查。

# 解決方案3：企業級平臺架構

ALTOS
an Acer Group Company

## 用戶層：靈活接入與易用體驗

提供Web介面，支援多終端靈活接入，降低使用門檻。

## 模型層：多任務支援與高效推論

整合LLM推論引擎、Triton/Whisper/Stable Diffusion，支援文本、語音、圖像多任務推論。

## 基礎設施層：穩定可靠的資源底座

基於Kubernetes編排，深度優化GPU支持，確保高並發場景下的穩定性與擴展性。

## 治理層：全方位安全與資源管控

涵蓋身份認證、權限管理與資源配額，實現平臺全生命週期安全與資源高效治理。

# 成本與效能優勢




## 推論成本顯著降低

相比雲端API服務，推論成本降低60-80%，大幅減少企業長期運營支出。

## GPU資源效能倍增

GPU利用率提升3-7倍，資源浪費最小化，單位算力產出最大化，效能領先業界。


## 快速實現投資回報

典型配置下12個月即可收回初始投資，長期TCO優化，為企業創造可持續價值。

# 國立成功大學 精準運動科學研究專案

an Acer Group Company

| GPU Server | 3rd Gen. Intel Xeon CPU |
|---|---|
| 2 Server | 2 CPUs of 6326 |
| **Nvidia A100 GPU** | **GPU Memory** |
| 6 GPUs | 480 GB |


## 精準運動科學研究計畫第二期

彈性分配GPU運算能力,提供模型所需的高速即時運算。
提供遠端管理,實時監控平台GPU運算能力的運行狀態。
開發人員可專注於利用AI技術開發運動算法作為運動科學,並增強技術在各種運動項目中的應用。

**Altos aiWorks with Altos AI Servers**

# 台灣出發 · 全球布局 · 服務多國客戶經驗

ALTOS
an Acer Group Company

攜手全球合作夥伴生態系，提供整合且全面的應用解決方案，助力終端用戶突破效率瓶頸、實現無縫擴展，並加速創新落地。

## 政府單位


## 企業組織




## 高等教育





ALTOS
an Acer Group Company
THE BEST IS YET TO COME