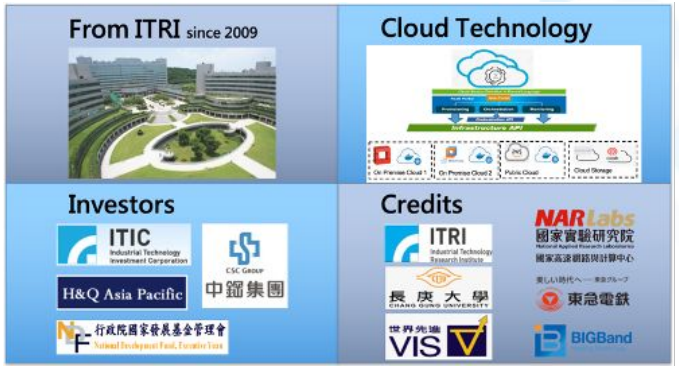
geminiopencloud
雙子星雲端運算

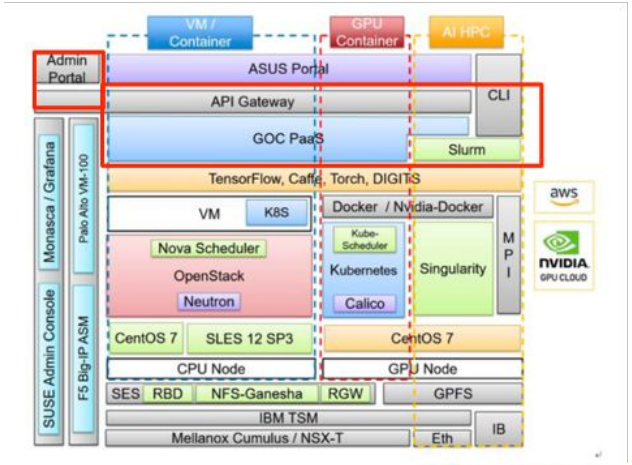
如何透過 AI Console 同時滿足校園裡的教學與研究

Roger Fang

關於雙子星雲端

- 2015 年成立，前身為***工研院資通所雲端部門*** 開發 CloudOS (IaaS) 系統，是亞洲少有IaaS 底層開發的團隊
- 為 OpenStack 基金會聯合創始董事會成員2017 年開始支援 Kubernetes(K8S)，並成為 CNCF (雲原生基金會)的 KCSP 夥伴
- 2018 年起協助建置***TWCC 台灣AI 雲，為台灣最大公有雲平台*** 提供資源管理平台(PaaS) 與 API 管理平台，目前管理高達1500 片 GPU，以及CPU/GPU Server 至少 200 台以上
- 2019年，將上述技術輕量化成為三項產品：Gemini AI Console 、GOC API Gateway、Gemini Management Console
- 2020年，推廣AI Console SE單機版並提供完整AI/ML Platform
- 2021年，持續產品研發、提供TWCC/TWS 維運、與多家公有雲策略合作進行***多雲(Multi-Cloud)架構並提供專業服務*** 利用產品及專業服務提供多雲管理。至今基於三大產品推廣雲原生服務優化策略，協助企業數位轉型並達到IT現代化





geminiopencloud

雙子星雲端產品

GEMINI MANAGEMENT CONSOLE

一站式雲原生應用開發與混合多雲管理

透過主流雲原生工具與 K8s 容器管理技術，搭配軟體敏捷開發流程 (CI/CD)，能讓企業專注在業務邏輯與服務開發，加速因應市場變化！

- 加速企業導入雲原生技術：容器、K8S 與微服務技術
- 加速持續整合與持續部署(CI/CD)
- 提供資源管理、無痛升級，與充分利用多雲部署
- 減低維運成本，增加開發滿意度

GEMINI AI CONSOLE

GPU 共享與加速，提升 AI 開發效率

基於獨家 GPU 共享技術，以及 K8s 管理平台，能有效分配 GPU 資源，最大化運算效率。搭配隨開即用的 AI 開發編輯器，大幅縮短 AI 開發時程！

- 簡化 IT 複雜性，優化GPU 管理效益
- 提升研究人員研發效率，縮短開發時程
- 支援不同運算架構與異質環境，滿足未來各種架構可能性
- 從管理到開發，完美AI 運算體嚴

GOC API GATEWAY

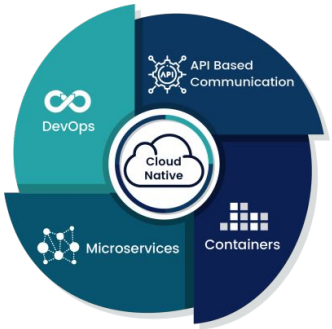
微服務最佳利器

協助企業快速導入微服務關鍵技術 API Gateway，以多重安全控管機制，有效管理 API 生命週期，快速增加外部合作夥伴和商業機會。

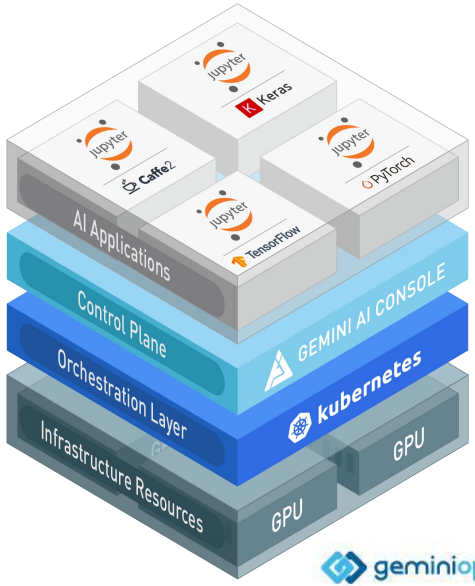
- 簡化對外API 服務的整合與控管流程
- 加速網路服務進入市場的速度
- 搭配為服務架構，能夠有效提升整體服務的迭代與更新
- 幫助企業快速對應市場變化的需求

Who are Our Customers

以雲原生技術支持 AI 研發已成為主流



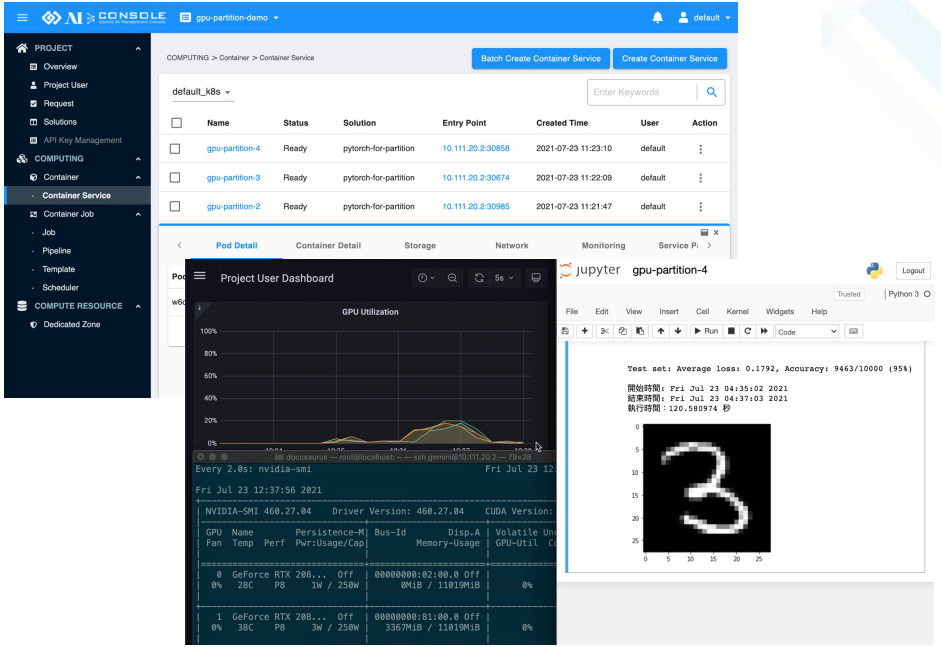
- 數位轉型 - IT 現代化
- 雲原生:***<u>容器、微服務、DevOps、雲端技術</u>***
- 採用輕便的容器架構，替代傳統主從式的單體式架構
- 快速部署、移轉、自動化
  - ***<u>部署:</u>***公有雲、本地部署、邊緣(Edge)
  - ***<u>移轉:</u>***服務不中斷的方式進行移轉
  - ***<u>自動化:</u>***滿足 IT 管理人員與開發人員之間的資源、服務自動化管理
- 以容器及 Kubernetes 技術，簡化IT 與開發工作負擔
  - 提供通用平台 (Write once, run everywhere)
  - 快速部署 AI 開發所需的 GPU 容器環境



geminiopencloud

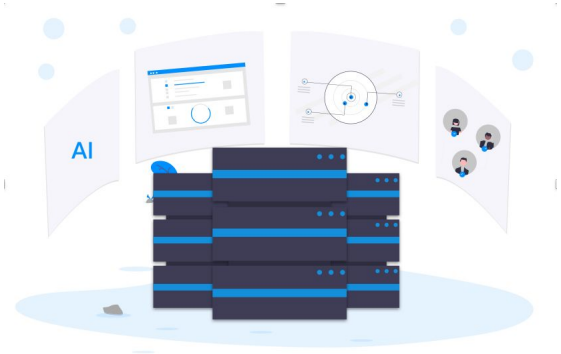
Gemini AI Console 產品簡介

- 以 Kubernetes 為技術核心，可同時管理混合異地 K8s 叢集，秒級啟動 AI 運算容器
- 可多容器共享 GPU，且資源獨立
- 啟動後內建 AI 開發使用的 Jupyter Notebook 編輯器
- 內建多種 AI 開發框架與私有鏡像庫，可客製化專有模板
- 內建 CSI，可串接多種儲存設備，並將既有資料掛載到容器內進行運算
- 完整資源監控介面
- 多租戶/多角色管理架構
- 容器服務/任務分開運行



geminiopencloud

Gemini AI Console 三大優勢

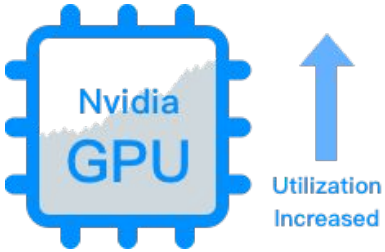


建立 AI 平台戰略

Gemini AI Console 是專門為

多用戶、多團隊、多種工作負載

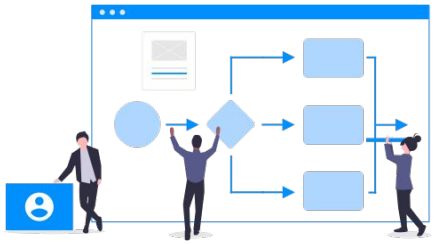
協同共享而設計的人工智慧管理平台



GPU 利用率最大化

在 Kubernetes 中，***最大限度地利用 GPU 資源。***進行深度學習訓練和推論預測

讓企業從中獲得價值



AI 工作調度管理自動化

具有***自動化特性和功能***

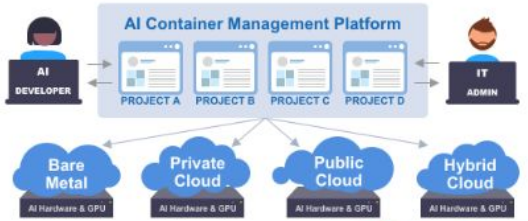
解放 IT 架構師和資料科學家

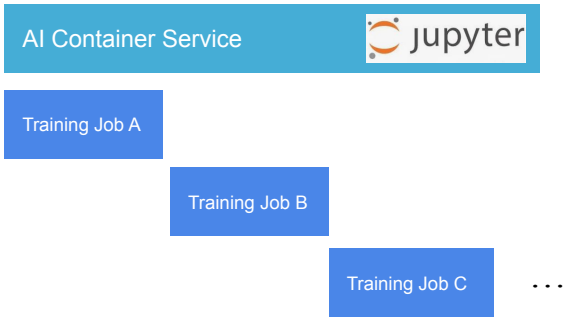
人工安排調度管理的窘境

geminiopencloud

建立 AI 平台戰略

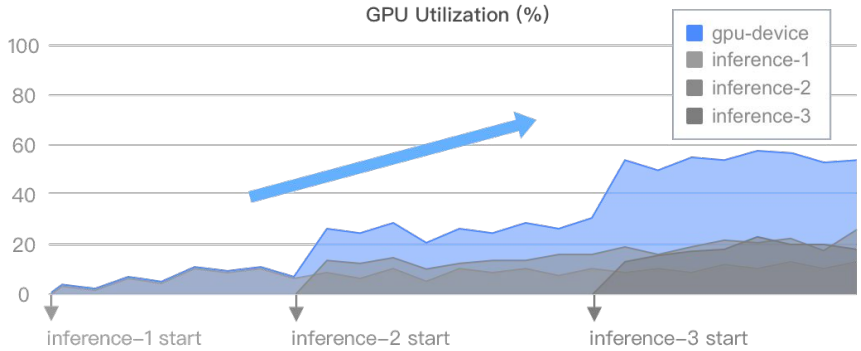
- **雲原生 Kubernetes 資源池**
  - 支援多雲、多 Kubernetes 叢集管理
  - 可串接公有雲 K8s，與管理地端私有 K8s 叢集
  - 協助整合為單一資源池，有效集中控管
- **多租戶 / 多專案管理**
  - 有效控管單一專案資源使用量
  - 適合企業組織架構：多專案多用戶/多角色管理
  - 內建軟體市集（Marketplace），可控管各專案的軟體使用權限
- **支援不同工作附載**
  - 同時支援不同大小工作附載的使用情境
  - 同時支援任務型（Job）與服務型（Service）的工作附載
  - AI 開發服務內建 Jupyter Notebook 編輯器





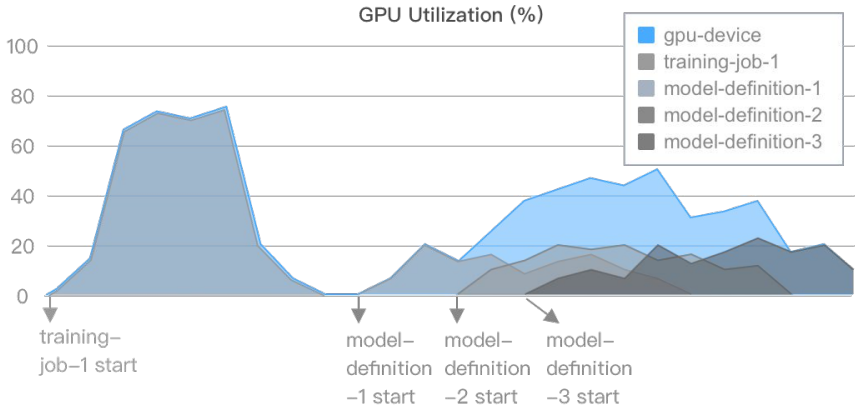
GPU 利用率最大化：GPU Partitioning

GPU-1



一顆 GPU 可同時分配給許多 low workload 的推論服務容器同時執行

GPU-2



一顆 GPU 可依據機器學習容器 的 workload 做適當的資源分配。Training Job 結束後，可分給其他 low workload 的容器一同運行。

geminiopencloud

GPU Partitioning 分割共享功能特性

CUDA GPUs

純軟體解決方案，只要是支援 CUDA 的 GPUs，皆可使用

無需修改程式

使用者無須修改任何程式，就可以使用分割後的GPU

有效資源隔離

有能力控制容器資源的隔離獨立，保證個別容器的資源，不受其他容器的干擾

彈性調度

使用者可規範的最小/最大QoS. 還可以彈性地自動調大GPU額度，也減少人員管理 overhead

提高容器使用量

Kubernetes可以執行同時更多機器學習容器，減少搶佔，更減少排隊時間

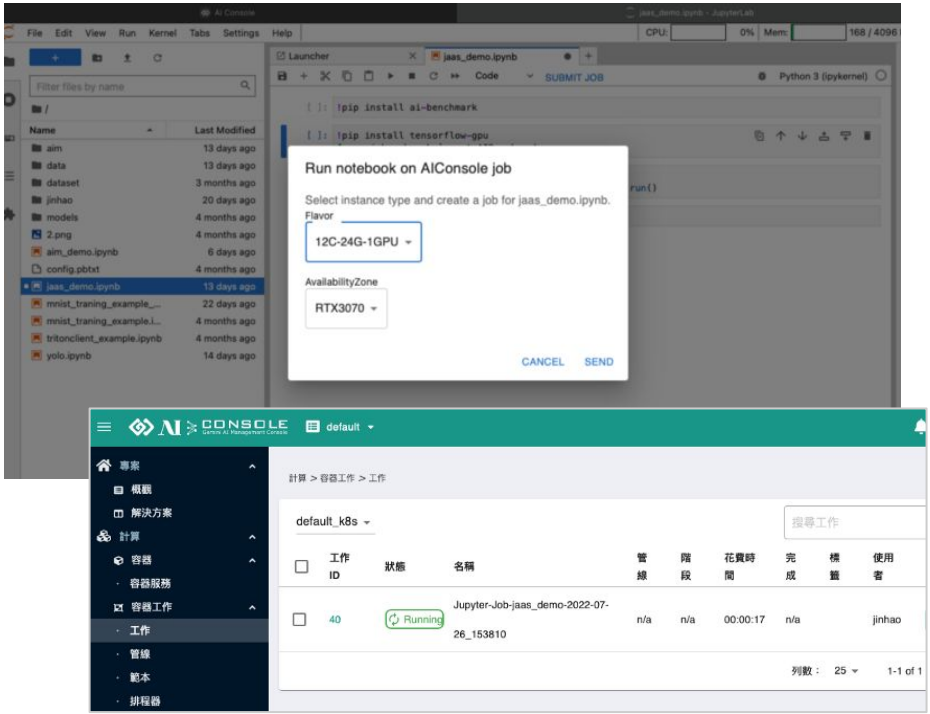
資源利用率提高

總利用率可接近個別容器利用率之和

總任務時間下降

假設原先各自任務時間為 x 跟 y，使用後兩個任務全部結束的時間小於 (x+y)

GPU 利用率最大化：Jupyter to Job

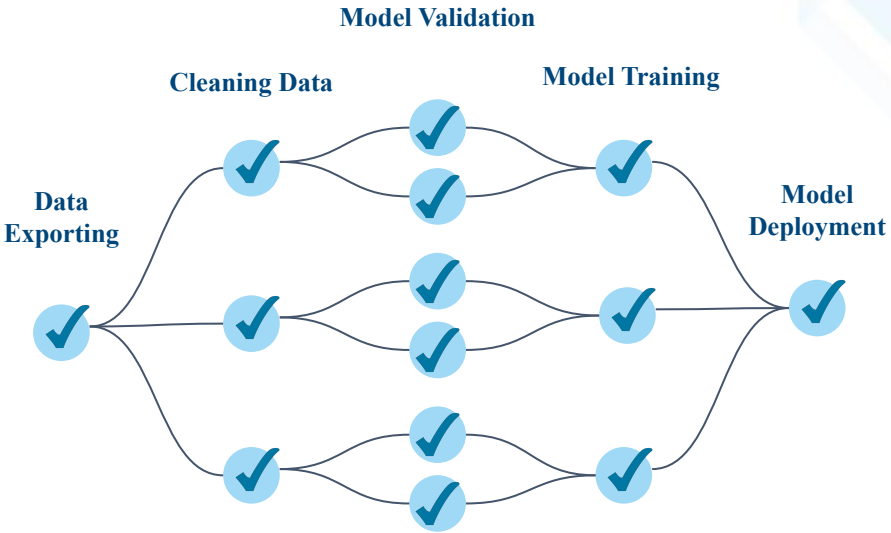


- 開發的筆記本服務，與 GPU 運算任務分開執行
- 保留完整訓練日誌檔，作為 AI 開發的除錯與追蹤

geminiopencloud

AI 工作調度管理自動化

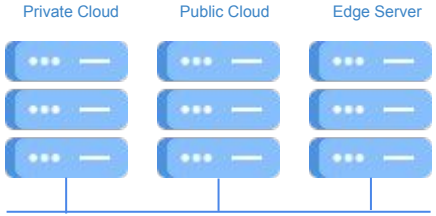
- 單一Job 測試用任務
- Pipeline 串接 Job 任務
  - 可透過 Stage 依工作順序串接
  - 單一 Stage 可同時執行多個 Job
  - 任何 Job 發生錯誤, 即中止任務節省資源
- 搭配 Pipeline Template 重複執行訓練
- 搭配 Scheduler, 可指定時間訓練
  - 指定特定時間或區間
  - 可搭配終止條件



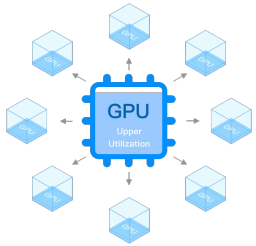
不同 AI 開發階段，機器學習負載不同

	建立模型	訓練	推論
使用情境	開發與除錯	訓練與超參數 (Hyperparameter) 調整	模型需要即時地服務資料
運作方式	高度的人機交互操作	通常在背景反覆持續執行	做為服務，長時間運行
著重點	重視使用者體驗與易用性	重視大輸出	重視短延遲時間
運作時間	週期時間短	週期時間長	長時間運行
效率	Low GPU utilization	High GPU utilization	Low GPU utilization，但偶有短時間內的爆量需求 (burst)
AI Console 推薦功能	Jupyter Container Services + Jupyter to Job **容器服務 + Jupyter to Job**	Jobs/Pipelines with Scheduler **Job任務、流水線與工作排程器**	Container Services + GPU Partitioning **容器服務 + GPU 分割共享**

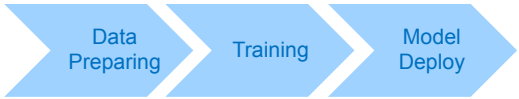
Gemini AI Console 六大功能回顧



異質 / 混合多雲平台



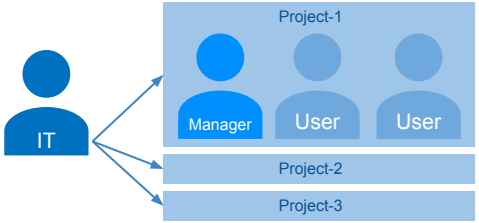
Gemini GPU Partitioning
多容器共享GPU



Job / Pipeline建置
自動化 MLOps 環境

AI 框架與開發軟體市集

API / Web UI 操作與監控



三層角色與多租戶管理

雲原生核心技術 - 協助企業數位轉型

國際認證與市場口碑

多種國際技術認證，大型企業實戰經驗與口碑

TWCC TAIWAN COMPUTING CLOUD

kubernetes CERTIFIED SERVICE PROVIDER

Google Cloud Certified PROFESSIONAL CLOUD ARCHITECT

NCHC NARLabs 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing

公有雲串接

與公有雲服務商合作，提供技術與遷移服務

Microsoft Azure

Google Cloud

aws

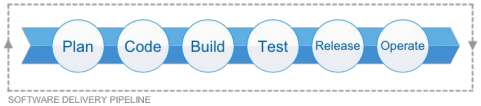
linode

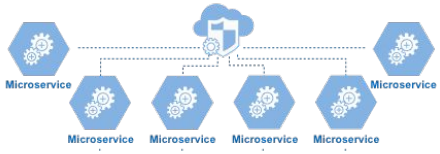
docker

Virtual Machines

雲原生/微服務架構

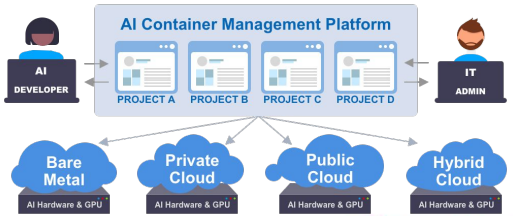
基於雲原生架構提供IT 現代化與 AI / MLOps 解決方案





多雲與系統整合

異業合作，提供客戶最佳解決方案



geminiopencloud

geminiopencloud

雙子星雲端運算

以簡馭繁・直上雲端

Thank you

www.geminiopencloud.com

contact@geminiopencloud.com

03-6590698