

快速部署・好管理・可服務化

ARX OS (AI 邊緣運算作業系統) 是專為 Edge AI GPU Server 而生的地端 AI 管理平台。
當 AI 硬體進入辦公室、分公司、工廠與分散式場域，企業真正缺少的往往不僅只是算力，而是一套能讓本地端模型快速部署、集中管理、服務化發布，並持續接軌最新 AI 生態的 OS 層。

ARX OS 將冷冰冰的地端 AI 硬體，轉化為可立即使用、可管理、可共用的企業 AI 生產力服務，幫助組織更快將 AI 投資轉化為實際工作效率。

模型與應用部署，從數小時縮短到 15 分鐘內!

透過 ARX OS，企業不需要從零組裝 GPU 環境、推論引擎與模型服務，也不必仰賴少數專業 IT 工程師。讓企業與開發者的關注回到應用與商業流程的創新，而無須應付底層布署與維護的繁雜。使用者可透過 ARX Edge OS 的 No-code 圖形化介面，快速完成私有模型部署、API Key 建立、LLM Endpoint 發布與連線測試，並整合 RAG 知識庫、NAS 資料來源及遠端維運功能，大幅縮短 POC、系統整合與驗證週期，同時建立可輕鬆部署、管理與維運的地端推論環境。

核心效益



規模可控可升級
AI 開發團隊可快速完成 POC 與模型驗證，降低前期硬體投資與部署風險



快速試錯與創新
可自由切換與測試不同開源模型，加速找到最適合的模型組合



No-Code 方式部署與發佈
透過整合式 API Endpoint 與 API Key 管理，快速建立應用程式連線



釋放IT人員的繁複工作
降低 IT 團隊整合 GPU 基礎架構、Runtime、模型服務與權限控管環境所需的時間與人力

ARX OS 核心能力

將複雜的地端 AI 部署，簡化成可管理、可共用、可擴充的私有 AI 平台。



01 15分鐘快速部署

以 No-Code 圖形化介面單鍵部署模型、推論 Runtime，並串接 RAG 知識庫與 NAS 資料。

02 IT 好管理

集中管理 GPU、模型、API Key、模型存取、到期、RPM、使用記錄與遠端維運。

03 AI 可服務化

將不同地端模型發布為標準 LLM Endpoint，讓應用、團隊與不同場域安全共用。

圖形化操作控制台

使用者可透過瀏覽器遠端登入，並以圖形化介面管理模型、Endpoint、API Key、應用程式、GPU 資源與系統狀態。

單鍵模型部署

透過一鍵導入專業模型、企業自有模型或核備開源模型(Hugging Face)，並自動配置推論引擎與 GPU 資源的部署機制，大幅縮短導入時程、降低對專業 MLOps 人力的依賴。

模型平台與多 Runtime

支援多種推論引擎、地端部署與 GPU 資源配置的模型中台，並提供 GPU 資源配置、API Key 建立、撤銷、到期、model-level 權限控管、RPM 管理與使用量統計儀表板。使用者能夠在同一平台上靈活部署與調度不同類型的模型，進而免除自建異質推論環境的整合負擔。

模型生命週期管理

透過涵蓋模型上架、版本控管、更新與多版平測的全生命週期模型管理功能，使用者能夠有序管控每一個模型版本，並在必要時手動切換回先前的穩定版本。

Backup & Restore 備份與還原

支援 App、模型與向量資料庫並可跨機器（異機）還原的備份機制，使用者能夠在硬體故障或災難情境下快速重建服務。

專案級別模型隔離

透過依部門或專案獨立啟動模型實例（支援一對一或一對多等彈性部署）的隔離機制，使用者能夠為不同團隊配置專屬的模型服務與知識空間，進而落實資料分艙（data segregation）管理、降低跨部門資料混用與外洩風險。

安全存取控管

支援 allowlist、blocklist、異常存取封鎖與 API 存取控管，協助企業建立更安全的私有 AI 使用環境。

Agent / MCP / CLI 多重整合介面

將平台串接至既有內部系統、知識庫、資料庫與各類企業工作流程，進而讓企業自身或第三方廠商快速建構自動化應用、擴充平台價值，並保留未來延伸與生態整合的彈性。

AI邊緣運算平台的完整方案

為: SaaS 快速落地、開源生態整合、IT 開發後部署 而設計

同一套 Edge OS，可從快速 POC 延伸到企業規模化部署

可依不同階段之需求，選擇適合的算力與交付方式，同時保留一致的模型部署、管理與服務化體驗。

01 POC 驗證 → 02 Deploy 落地 → 03 Scale 多場域 → 04 Operate 持續維運







ARX-100

小型Edge AI 設備

Nvidia GB10

整合 GPU 硬體與 ARX OS、接近開箱即用的 Edge AI 一體機

適合

分公司、小型部門、POC 團隊與分散式場域。

關鍵價值

不需從零建置 GPU環境，即可快速啟動私有模型與 GenAI 應用。

預計終端售價 NT350,000起*

ISV 軟體開發夥伴，另有優惠

ARX-200 / 300

企業級 Edge AI 系統 & No-Code 私有模型伺服器

RTX Pro 6000/ GB300
DGX AI Station ; BTO

穩定、可管理、可治理的私有主權AI雲 (Sovereign AI Private Cloud)

適合

企業IT、AI團隊與需要私有模型治理的組織。諸如: 金融、政府、醫療、製造等。

關鍵價值

最快 15 分鐘建立模型服務，降低對 GPU, MLOps 與推論專家的依賴。

OS 軟體授權, 雙GPU節點計價* + MA維護

vARX

私有AI GPU運算租用與 PaaS 遠端開發環境

PaaS, by Subscription

月租即用，接近實機體驗的私有雲 PaaS。

適合

開發者、ISV、AI SaaS團隊與企業AI開發團隊。

關鍵價值

無需高額硬體前期投資，即可進行遠端開發、POC、模型部署與效能驗證。

依照算力(型號)、期間與專案需求報價 (詳見網頁)

把下一台 GPU Server，直接變成可用的企業 AI 服務

開端智能: www.ardge.com
業務合作聯繫: usasales@ardge.com

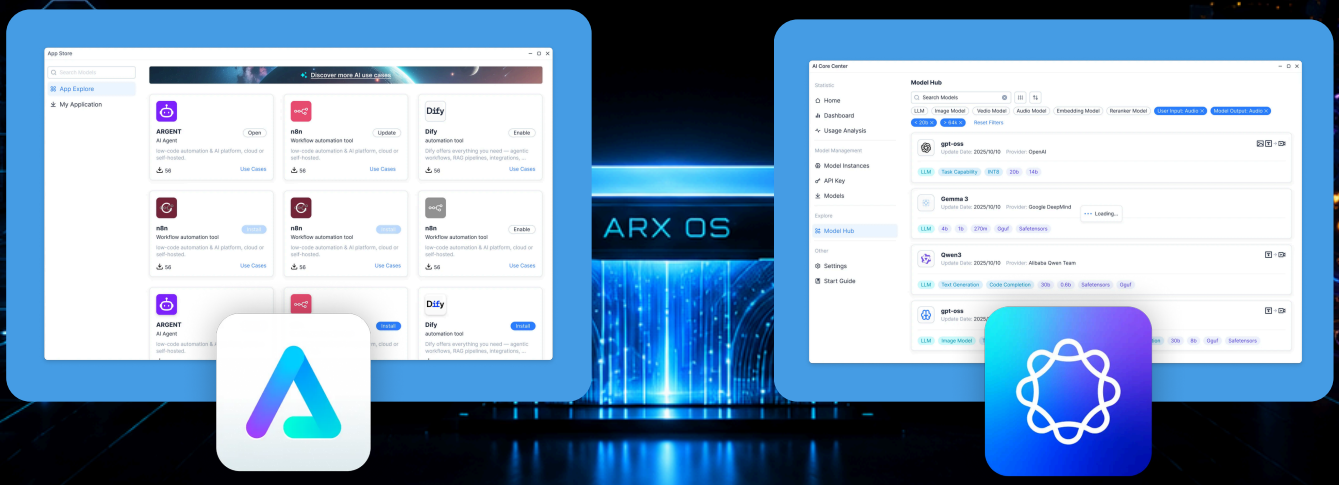
從資料到可執行的智慧

零門檻部署｜App Market

提供如行動裝置 App Store 般的操作體驗，一鍵安裝與管理通過驗證的 AI 工具，無須使用指令列。

AI 模型中心 | Model Hub

深度整合 Hugging Face 生態系，支援模型視覺化下載、推論管理與容器化部署，確保不同環境之間的一致性與部署穩定性。



ARGENT｜智慧知識代理建置平台

DIY 式知識庫與 AI Agent 平台
專為中小企業與非技術使用者設計。文件上傳後可自動完成向量化與結構化處理，以最少設定快速建立自然語言問答系統。

PreciseMD｜高精度文件辨識引擎

精準的OCR 資料前處理工具，支援批次地端處理非結構性資料，提供CLI/ MCP 等整合，ISV夥伴可不必重複開發。
支援複雜版面辨識與結構化資料擷取，提升 AI 回應的準確度與品質。

ScribeSTT｜語意影片搜尋

將影音內容轉化為結構化知識
將音訊轉換為高準確度逐字稿，並同步完成向量化處理，建立可搜尋的影片知識庫，讓使用者可透過自然語言快速查詢與存取內容。

讓地端 AI，成為可規模化的產品服務

當企業因合規、資料治理、智慧財產保護與主權 AI，要求模型部署於本地端，AI SaaS 與 ISV 團隊也迎來新的市場機會。但地端部署不能只是一次性的客製專案。面對不同硬體、GPU 平台與客戶環境，企業需要一套標準化、可管理、可持續支援的部署方式。

透過簡化且一致的地端邊緣基礎架構，軟體團隊無須投入複雜的底層整合，即可將每個客戶環境轉化為標準化、可控制的服務節點，持續專注於模型、Agent 與使用者體驗創新。



把下一台 GPU Server，直接變成可用的企業 AI 服務

開端智能: www.ardge.com
業務合作聯繫: usasales@ardge.com