

# 釋放企業 AI 潛能

從創新 VMware Private AI 平台開始

Evan Huang

Broadcom 軟體事業群技術副總

2025/09/04

# Disclaimer

- Certain information in this presentation may outline Broadcom's general product direction.
- This presentation shall not serve to (i) affect the rights and/or obligations of Broadcom or its licensees under any existing or future license agreement or services agreement relating to any Broadcom software product; or (ii) amend any product documentation or specifications for any Broadcom software product.
- This presentation is based on current information and resource allocations and is subject to change or withdrawal by Broadcom at any time without notice.
- The development, release and timing of any features or functionality described in this presentation remain at Broadcom's sole discretion.
- Notwithstanding anything in this presentation to the contrary, upon the general availability of any future Broadcom product release referenced in this presentation, Broadcom may make such release available to new licensees in the form of a regularly scheduled major product release.
- Such release may be made available to licensees of the product who are active subscribers to Broadcom maintenance and support, on a when and if-available basis.
- The information in this presentation is not deemed to be incorporated into any contract.

ANNOUNCING

**vmware<sup>®</sup>**  
by **Broadcom**

**AMD** 

**Virtualize AMD Instinct  
MI350 Series GPUs**

**Leverage AMD  
Enterprise AI Software**

**Open  
Ecosystem Support**

The information in this presentation is for informational purposes only and may not be incorporated into any contract.  
There is no commitment or obligation to deliver any items presented herein.

## 議程

- 企業導入 AI 的挑戰及考量
- VMware Private AI 滿足企業 AI 需求
- Demo：輕鬆建立企業專屬 AI Agent

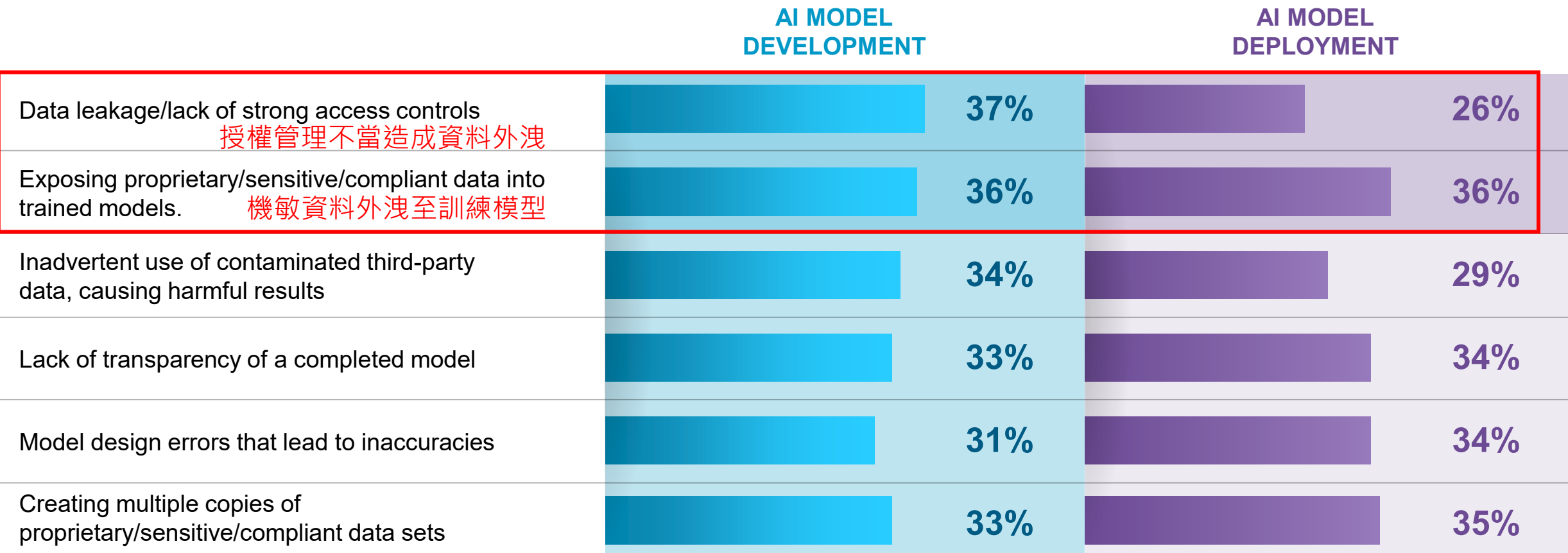
# 搭建企業與政府機構 Private AI 平台的關鍵考量



# 其中，隱私安全是導入 AI 的首要前提

## Top Risks Associated with AI Model Development and Deployment

What are the top risks that your organization associates with developing and/or deploying AI models?  
(Percentage of respondents)



Source IDC White Paper, sponsored by Broadcom, On-Premises AI Infrastructure Balances Innovation and Security, doc #US52747024, December 2024

# AI 機敏及隱私資料外洩案例頻傳，嚴重性不言可喻


**iThome** 新聞 專題 技術 AI Cloud 永續IT 醫療IT 資安 研討會 社群 IT

## 新興AI生產力平臺OmniGPT驚傳資料外洩，駭客聲稱握有3萬用戶個資、3,400萬筆聊天記錄

資安新聞網站HackRead指出，有人宣稱握有AI生產力平臺OmniGPT大批用戶資料及對話記錄，其中包含調查報告、大學文憑、警察刑事紀錄證明等各式的敏感資料，有可能對用戶帶來嚴重的資安風險

文/ 周峻佑 | 2025-02-14 發表 讚 5 分享

<https://www.ithome.com.tw/news/167370>



**Gloomer**

000-000

**GOD**

Yesterday, 10:04 AM (This post was last modified: Yesterday, 10:19 AM by Gloomer.)

This leak contains all messages between the users and the chatbot of this site as well as all also 30k user emails.

you can find alot of useful information in the messages such as API keys and credentials an are very interesting because sometimes they contain credentials/billing information.

Goodluck finding something and enjoy this leak! :

**Hidden Content**

Sample:

<https://limewire>



# 市面主流 AI 服務提供者也可能發生資料外洩

OpenAI 真實事件：用戶可看到他人資料

March 24, 2023 Product

## OpenAI March 20 ChatGPT outage: Here's what happened

An update on our findings, the actions we've taken, and technical details of the bug.

We took ChatGPT offline earlier this week due to a bug in an open-source library which **allowed some users to see titles from another active user's chat history**. It's also possible that the first message of a newly-created conversation was visible in someone else's chat history if both users were active around the same time.

The bug is now patched. We were able to restore both the ChatGPT service and, later, its chat history feature, with the exception of a few hours of history. As promised, we're publishing more technical details of this problem below.

<https://openai.com/index/march-20-chatgpt-outage/>

ChatGPT 提示：為避免資料外洩，用戶須自行控制承擔風險

iThome 新聞 專題 技術 AI Cloud 永續IT 醫療IT 資安 研討會 社群 IT

## 【當心提示注入、敏感資訊洩漏、錯誤資訊等問題】 生的LLM資安風險

生成式AI不當使用出現實際案例！臺灣有北捷AI客服可代寫程式碼，國外有三星員工將內部資料上傳至公用服務，以及美國律師誤用AI虛構判例打官司，均具體呈現LLM在不同層面的潛在風險

文/ 羅正漢 | 2025-04-18 發表

讚 45

分享

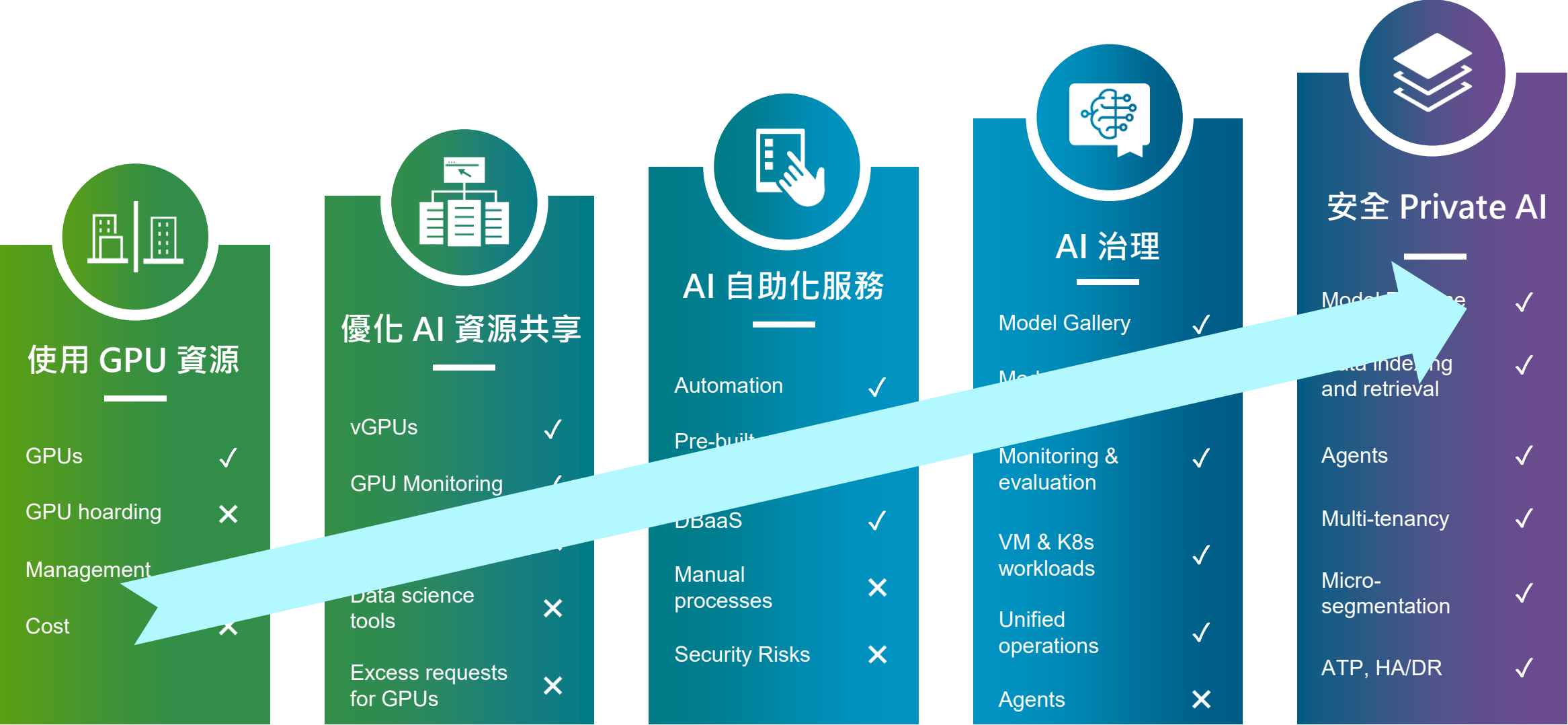


近年來大型語言模型爆紅，帶來新的機會，也帶來風險與挑戰，需要我們去注意，AI服務供應商也會不厭其煩，提醒使用者注意。例如，大家用熱門的ChatGPT服務時，AI在一開始就會宣告：請勿分享敏感資訊、查核事實，此舉就是希望用戶必須認知到相關風險。

<https://www.ithome.com.tw/news/167370>



# 一步到位搭建安全 Private AI 平台，才能避免風險並實現 AI 效益



# 典型組織內部，往往有多個業務或組織平行執行多項 AI 專案/系統

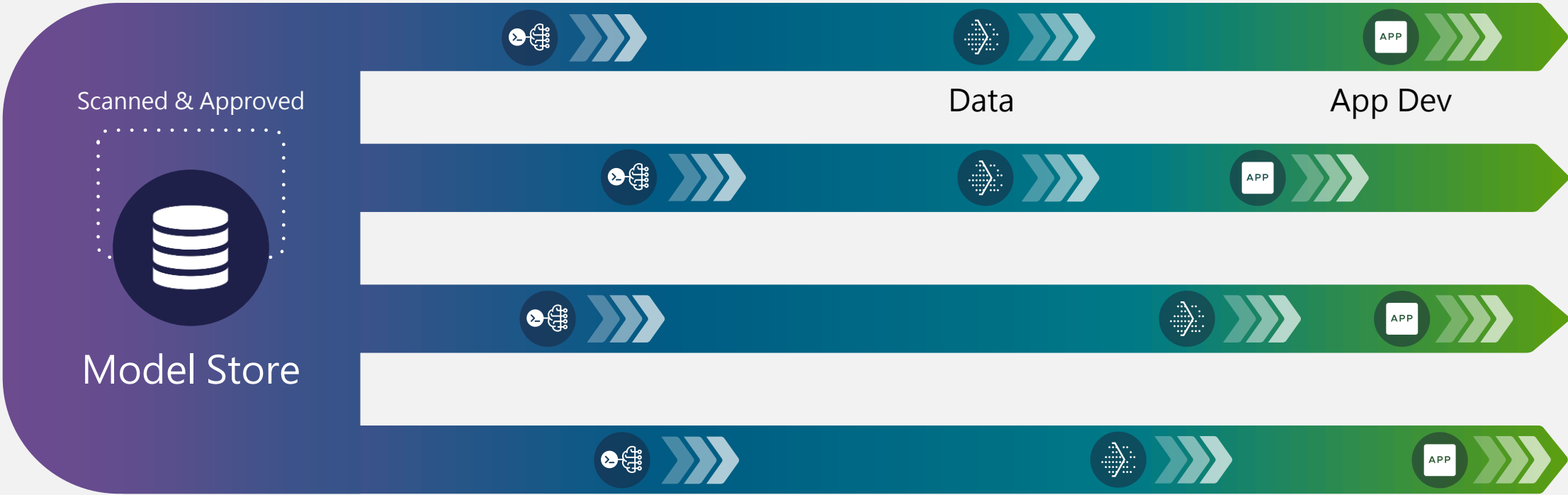
如何快速高效執行，避免相互影響，並且符合一致化的治理規範？



有哪些模型可以使用？哪些是符合規範？剩下多少可剩餘？如何處理、使用剩餘 GPU 與模型運行狀態？

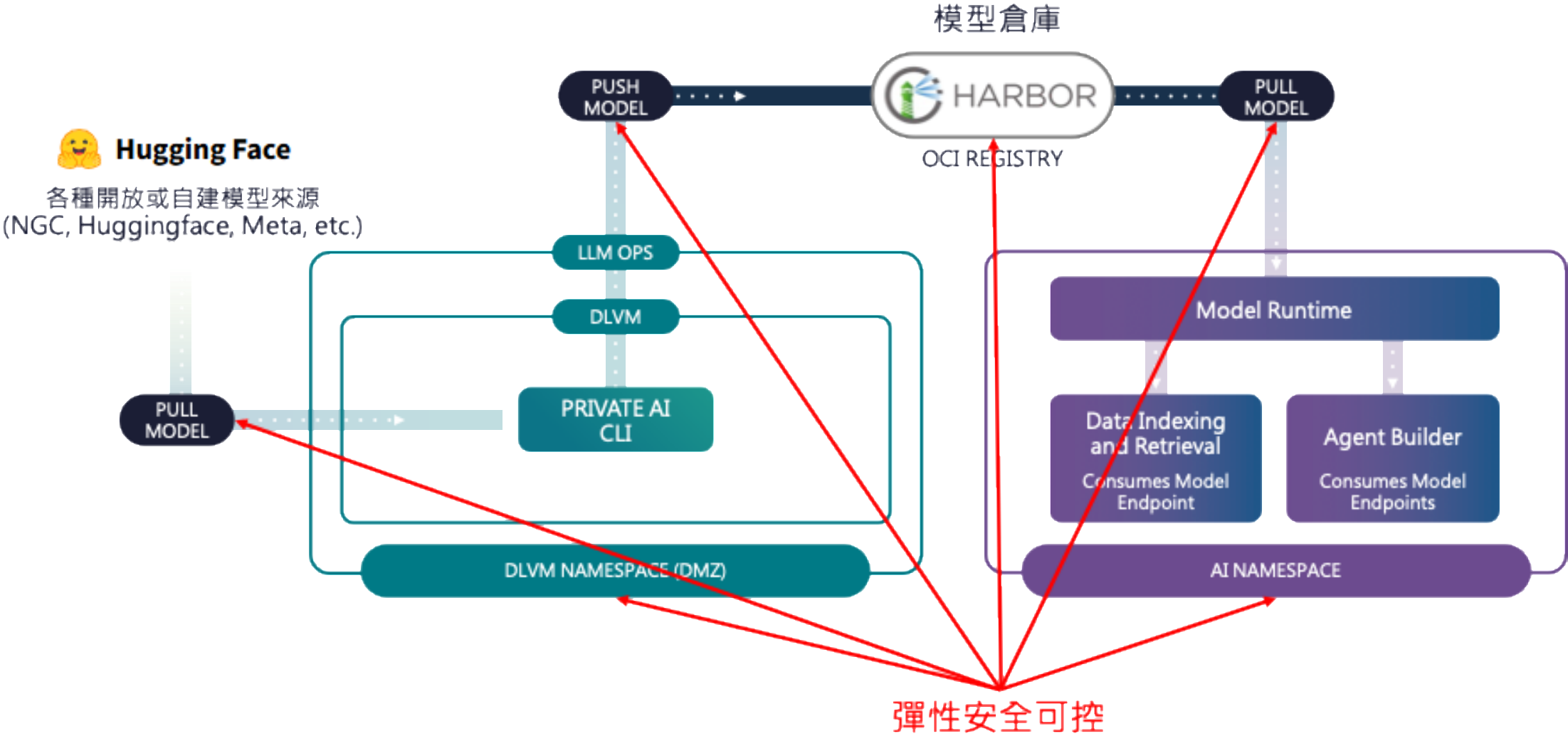
# 統一 Private AI 平台 – 模型倉儲 (Model Store)

降低 IT 團隊負擔，提高 AI 團隊生產力



選擇模型 >>> 搭建環境 >>> 部署模型 >>> 準備資料 >>> 開發應用（上線）

# 使用 Harbor 作為模型倉庫，管理模型

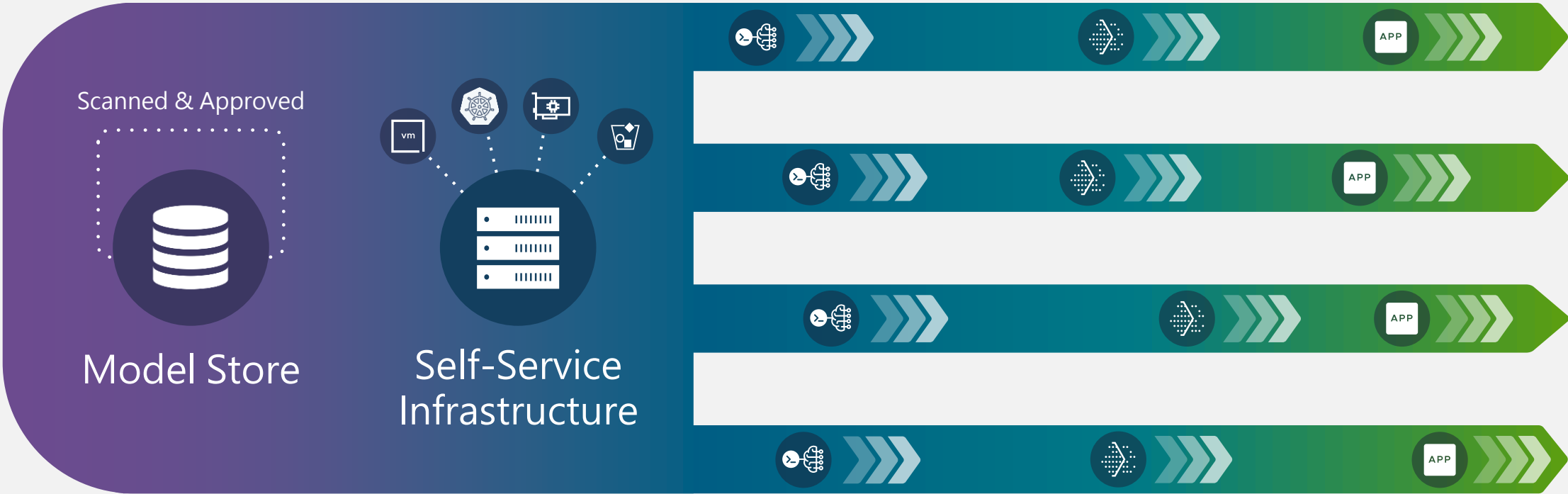


# 統一 Private AI 平台 – 自助自動依政策生成 AI 環境

降低 IT 團隊負擔，提高 AI 團隊生產力



成本 性能 合規



選擇模型



搭建環境



部署模型



準備資料



開發應用（上線）

# 使用目錄選單方式隨需自動化部署環境、設定政策、配置資源

Workloads

Hardware

Overview

Build & Deploy

Manage & Govern

Administer

Inbox

Instances

Catalog

Content Hub

Object Store

Services

paif29-ns-pc644

Overview

VMware Cloud Foundation Automation

Fritz Arbeiter

paif29-org

Catalog

The catalog offers ready-to-deploy compute resources, consisting of VMs, storage, networks, and so on, that you can provision to cloud regions associated with the projects they are members of. After a successful request, you can then manage the lifecycle of your deployed catalog instances.

Filter

Show items without a project

Search

Sort: Name (ascending)

AI Kubernetes cluster - e2e-sb-airgap-feb20

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - e2e-sb-cloud-dsm-feb21

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - e2e-sb-cloud-feb20

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - e2e-sb-proxy-feb20-r1

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - jf-airgap

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - jf-dsm-airgap

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - jf-dsm-cloud

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - jf-invalid-token

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - jf-proxy

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - paif29-ns-r4j36

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - swp-airgap

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - swp-cloud-public

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - zach-cloud-latest

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes cluster - zach-proxy-1

VMware Kubernetes cluster with GPU-capable worker nodes to run AI/ML cloud-native workloads.

Projects default-project

REQUEST ACTIONS

AI Kubernetes RAG Cluster - e2e-sb-airgap-feb20

VMware Kubernetes cluster with GPU-capable worker nodes to run a reference RAG solution.

Projects default-project

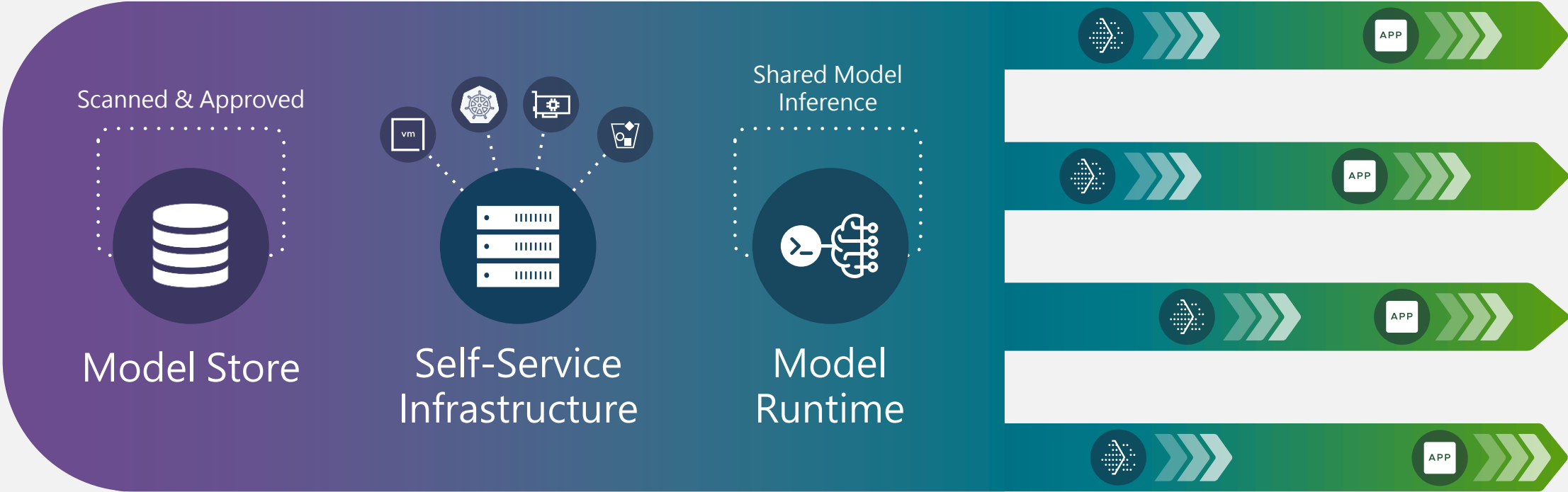
REQUEST ACTIONS

任選所需 AI 環境及資源配置，自動化供裝



# 統一 Private AI 平台 – 部署模型至 AI 運行平台

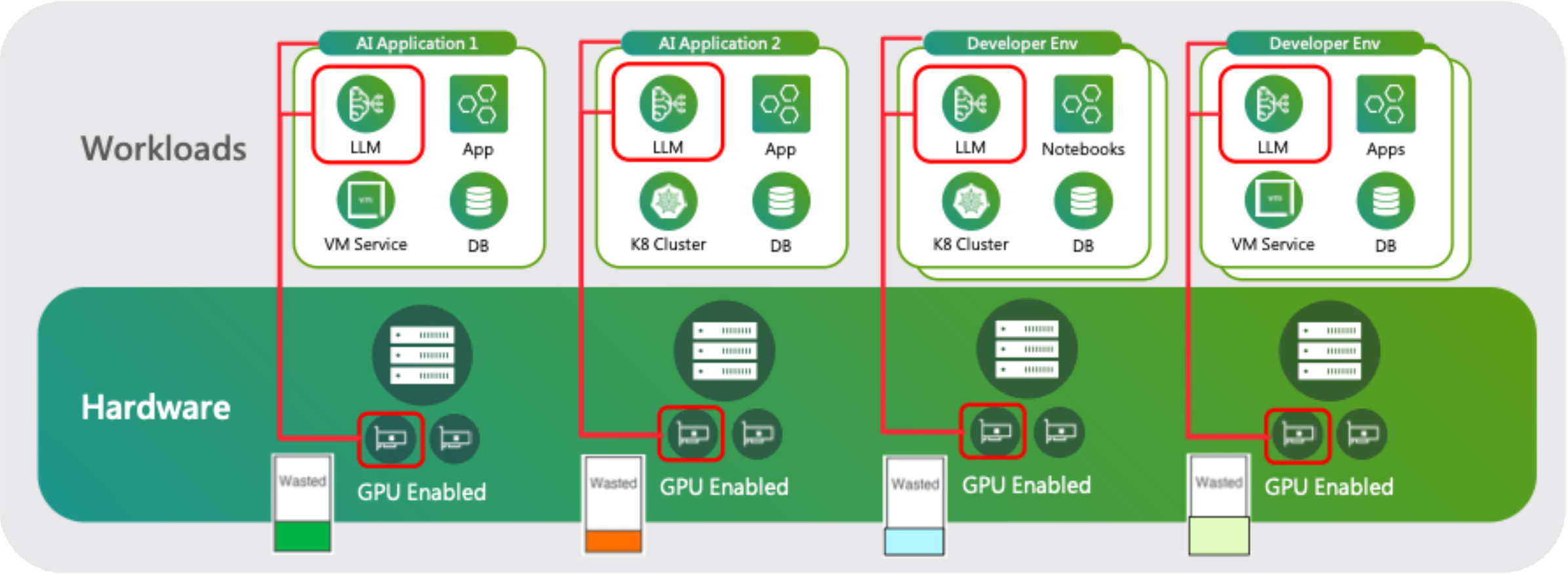
降低 IT 團隊負擔，提高 AI 團隊生產力



選擇模型 >>> 搭建環境 >>> 部署模型 >>> 準備資料 >>> 開發應用（上線）



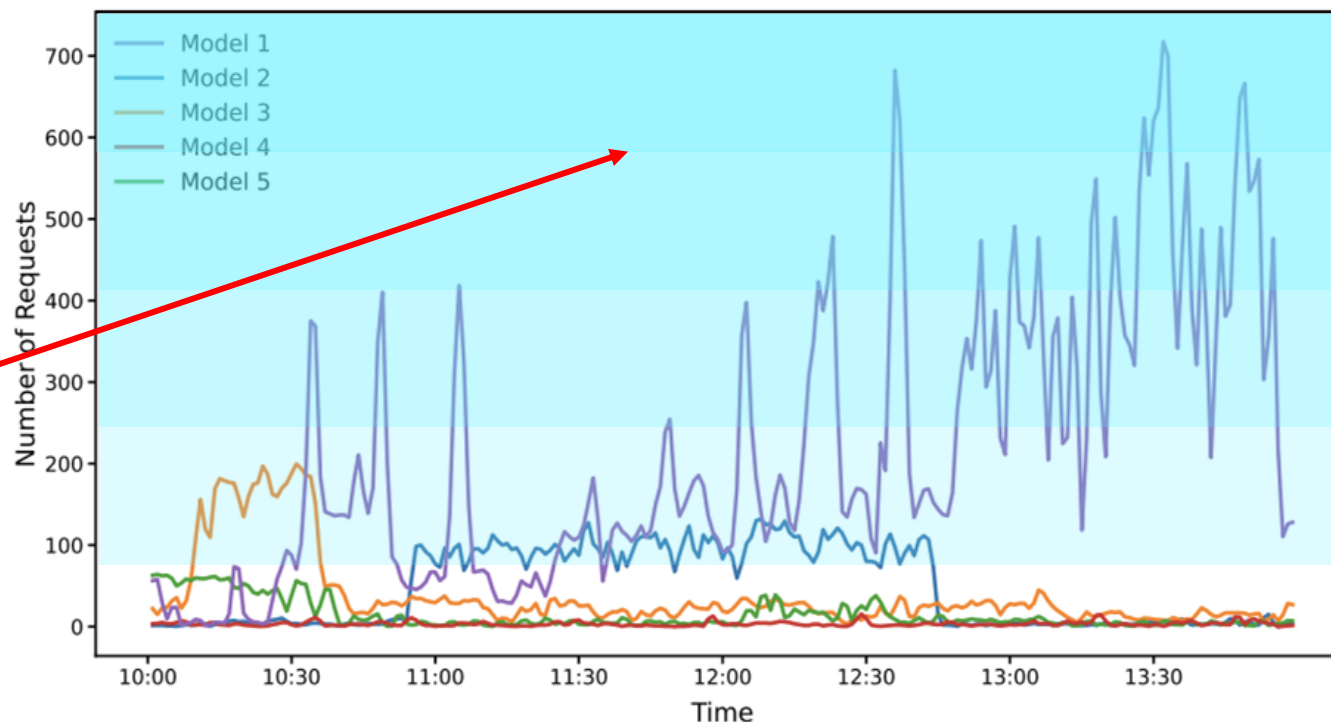
# 配置專屬 GPU 給特定模型及應用，容易浪費寶貴資源



# GPUs are significantly underutilized because

Inference loads change drastically with time

Request rates over a 4-hour period from an LLM inference service provider



寶貴 GPU 資源大量浪費

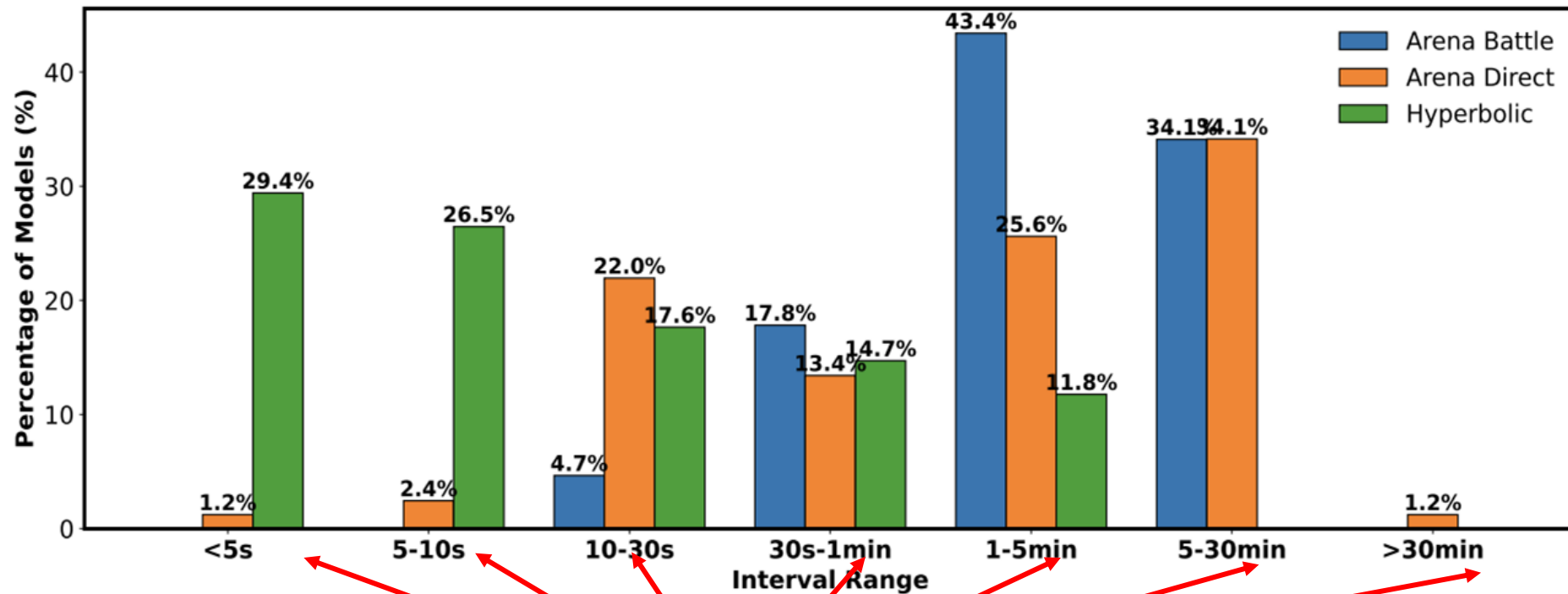
GPUs are overprovided for peak load,  
so underutilized during off peak

Jiarong Xing and the Prism team

# GPUs are significantly underutilized because

Long idle periods between request bursts

Median Request Interval Distribution for All Models

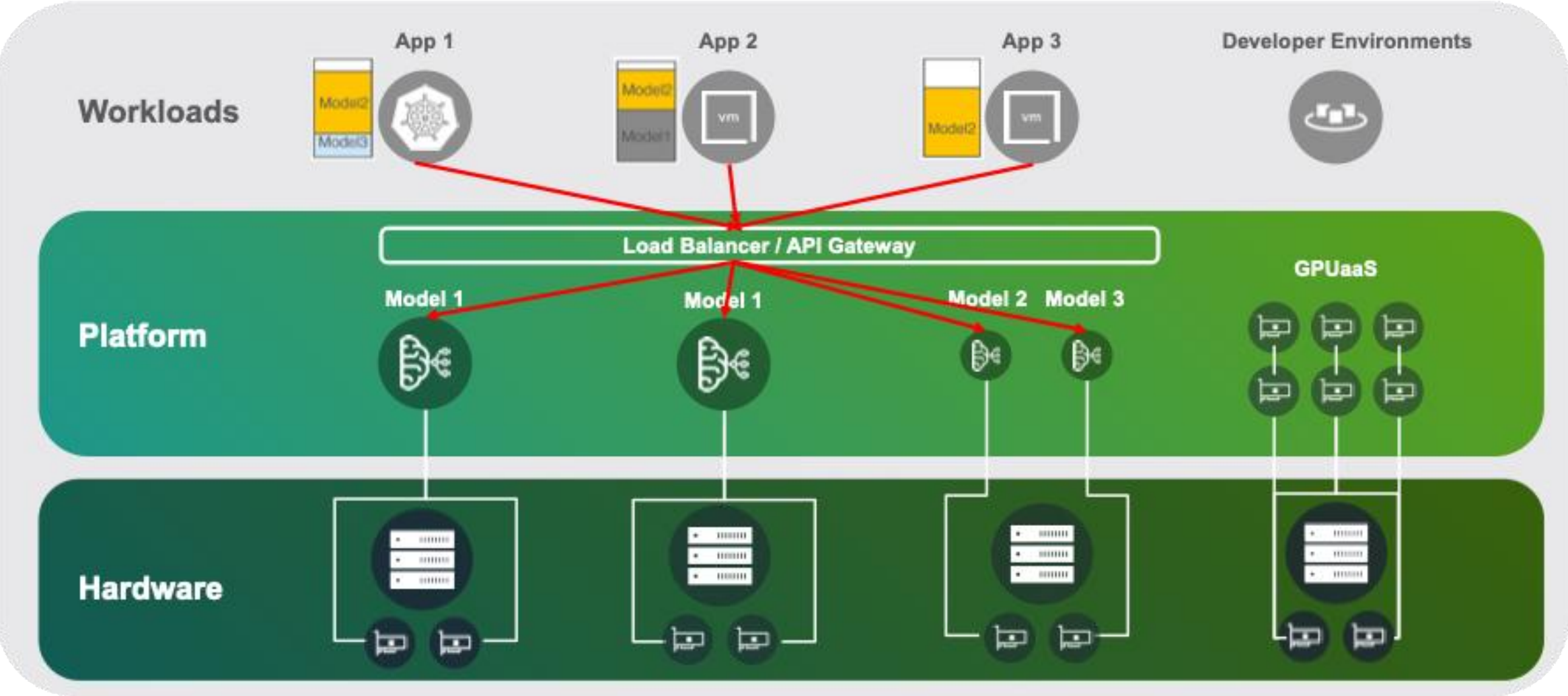


Models with no requests cause GPU idle

沒有處理 AI 服務請求的時候，已配置的 GPU 資源被閒置

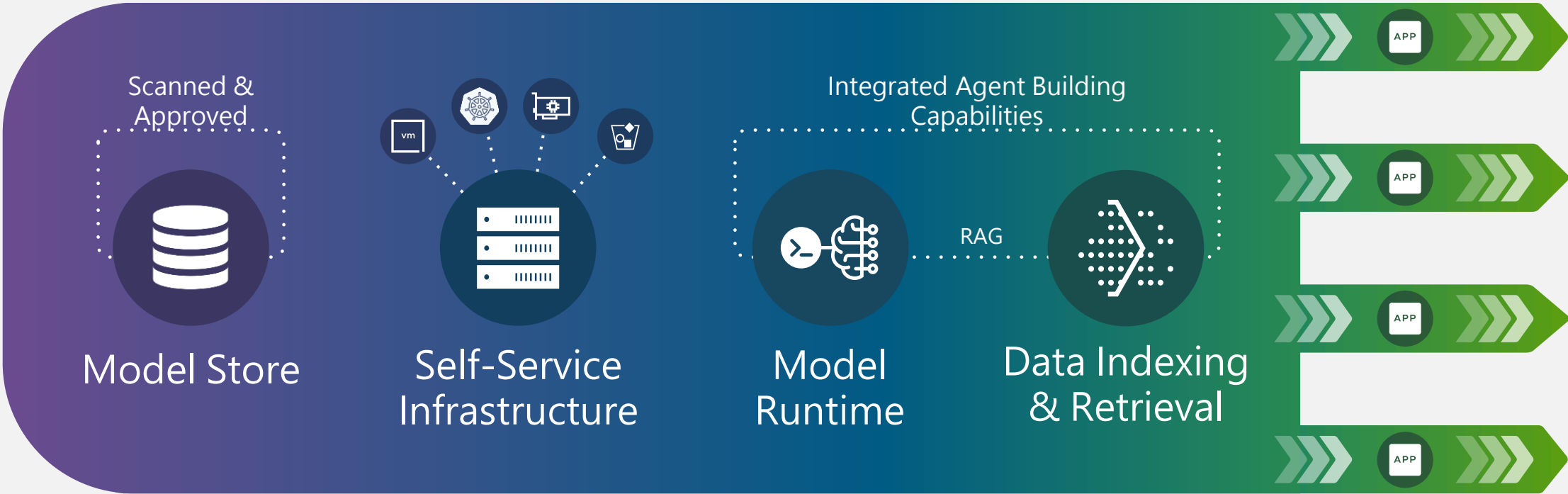
Jiarong Xing and the Prism team

# Private AI 平台彈性共享模型及 GPU，提高資源利用率



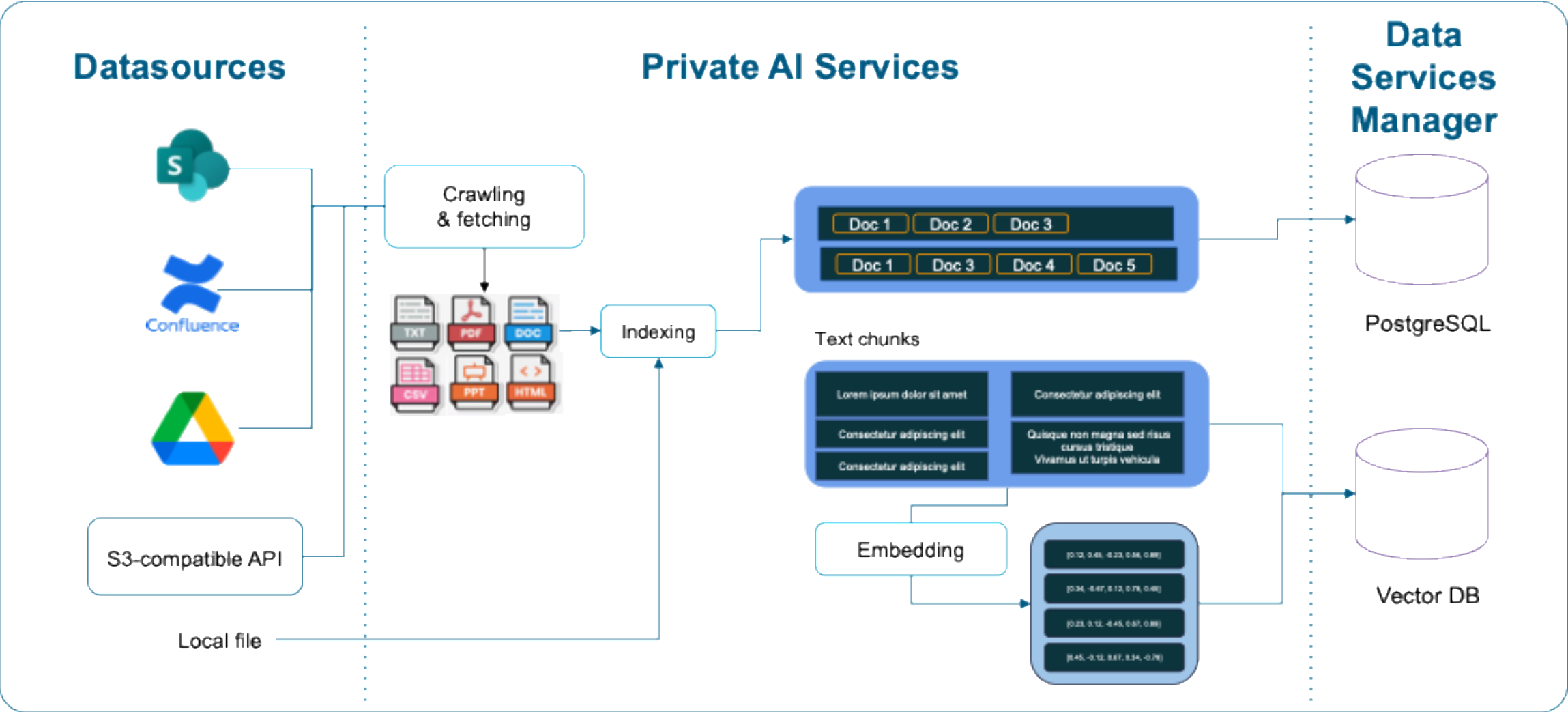
# 統一 Private AI 平台 – 資料索引及存取

降低 IT 團隊負擔，提高 AI 團隊生產力



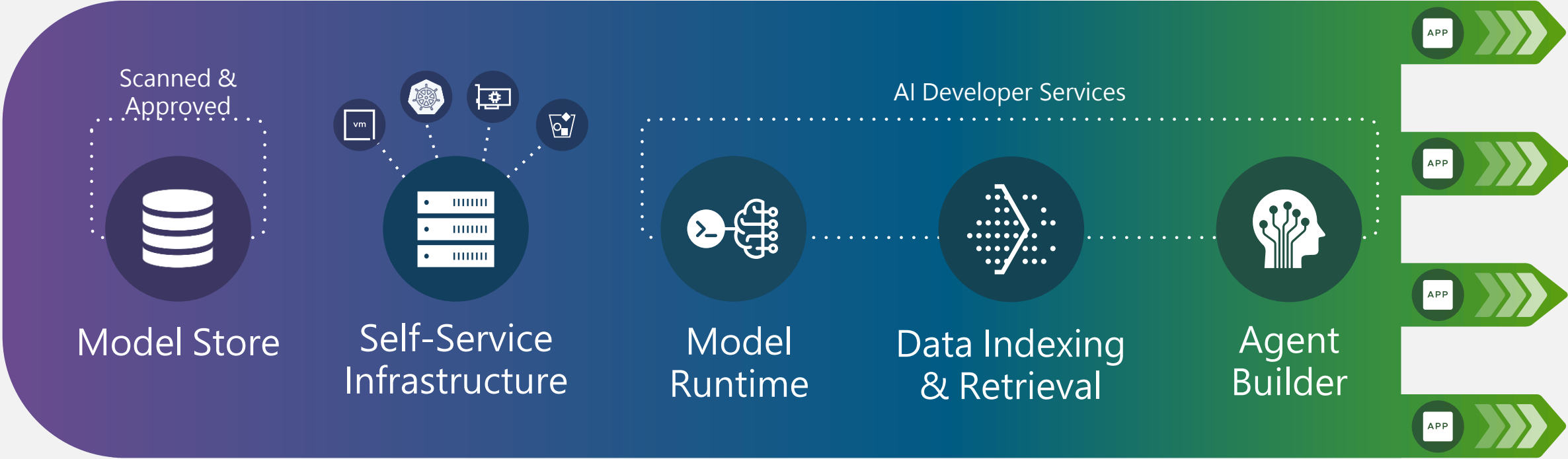
選擇模型 >>> 搭建環境 >>> 部署模型 >>> 準備資料 >>> 開發應用（上線）

# 快速建立向量資料庫 ( DBaaS ) ，建立資料索引並發佈



# 統一 Private AI 平台 – 整合建立 AI Agent，上線營運

降低 IT 團隊負擔，提高 AI 團隊生產力



選擇模型



搭建環境



部署模型



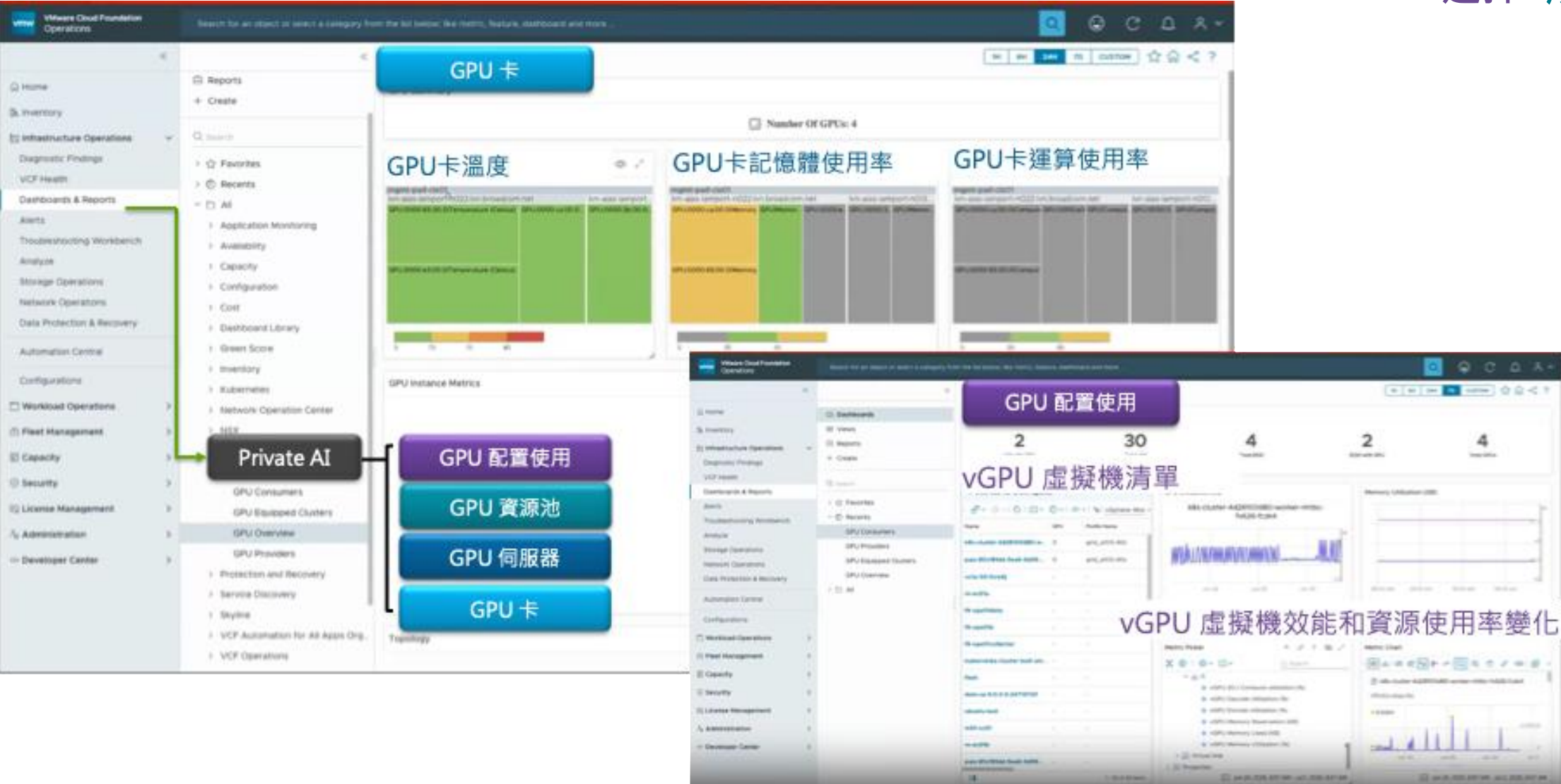
準備資料



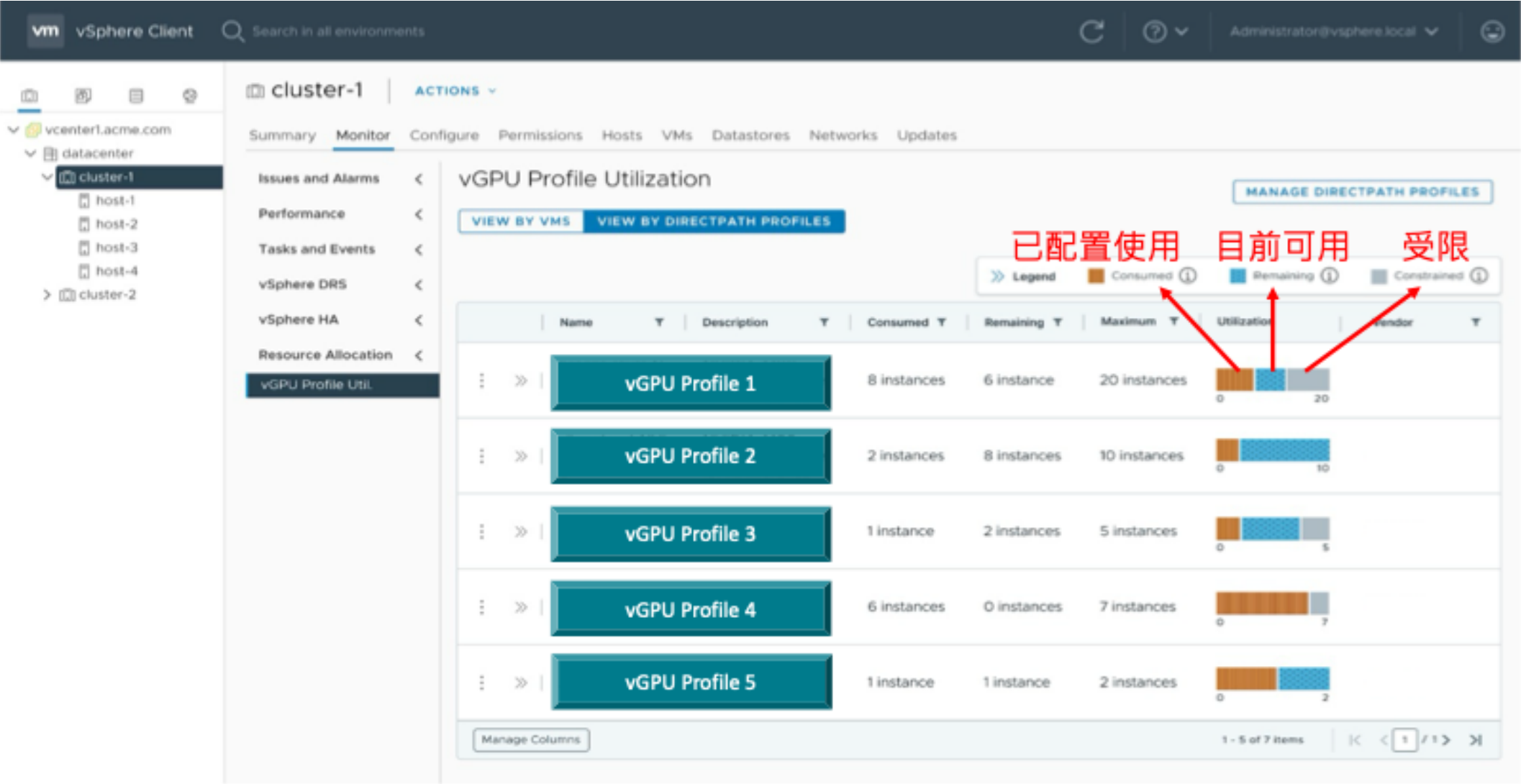
開發應用（上線）



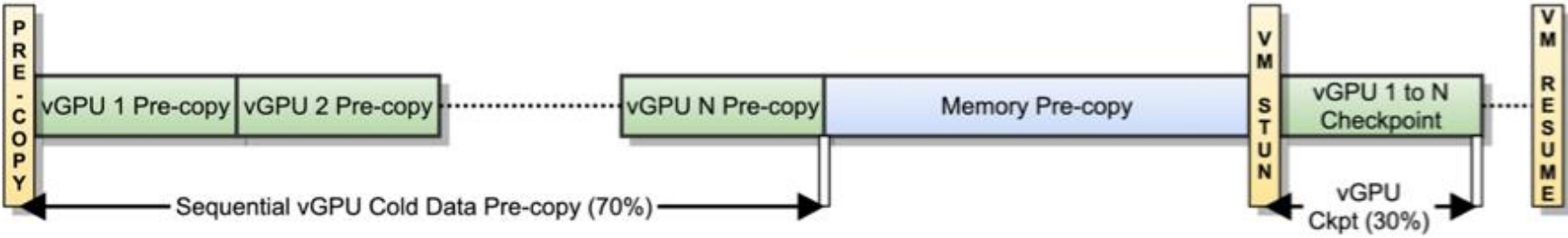
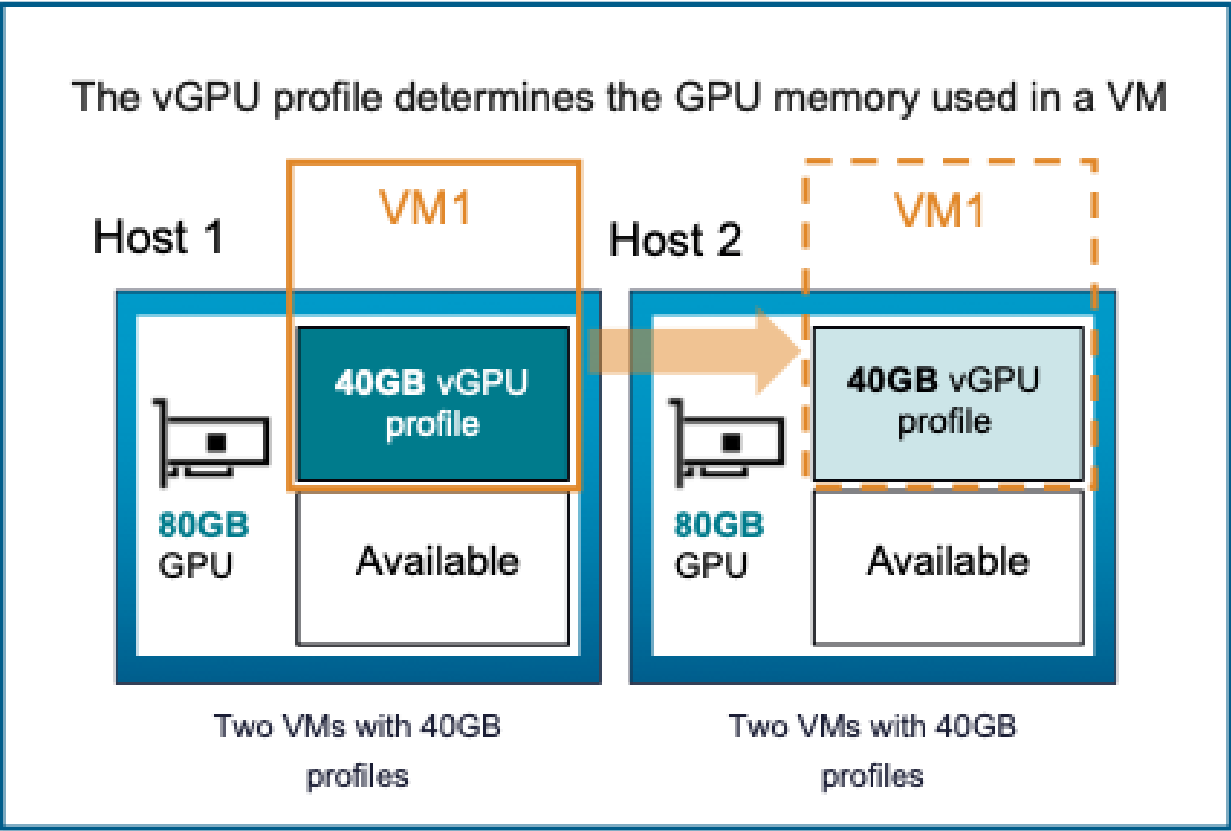
# 經由開箱即用的圖形化 Private AI 儀表板，掌控 GPU 使用



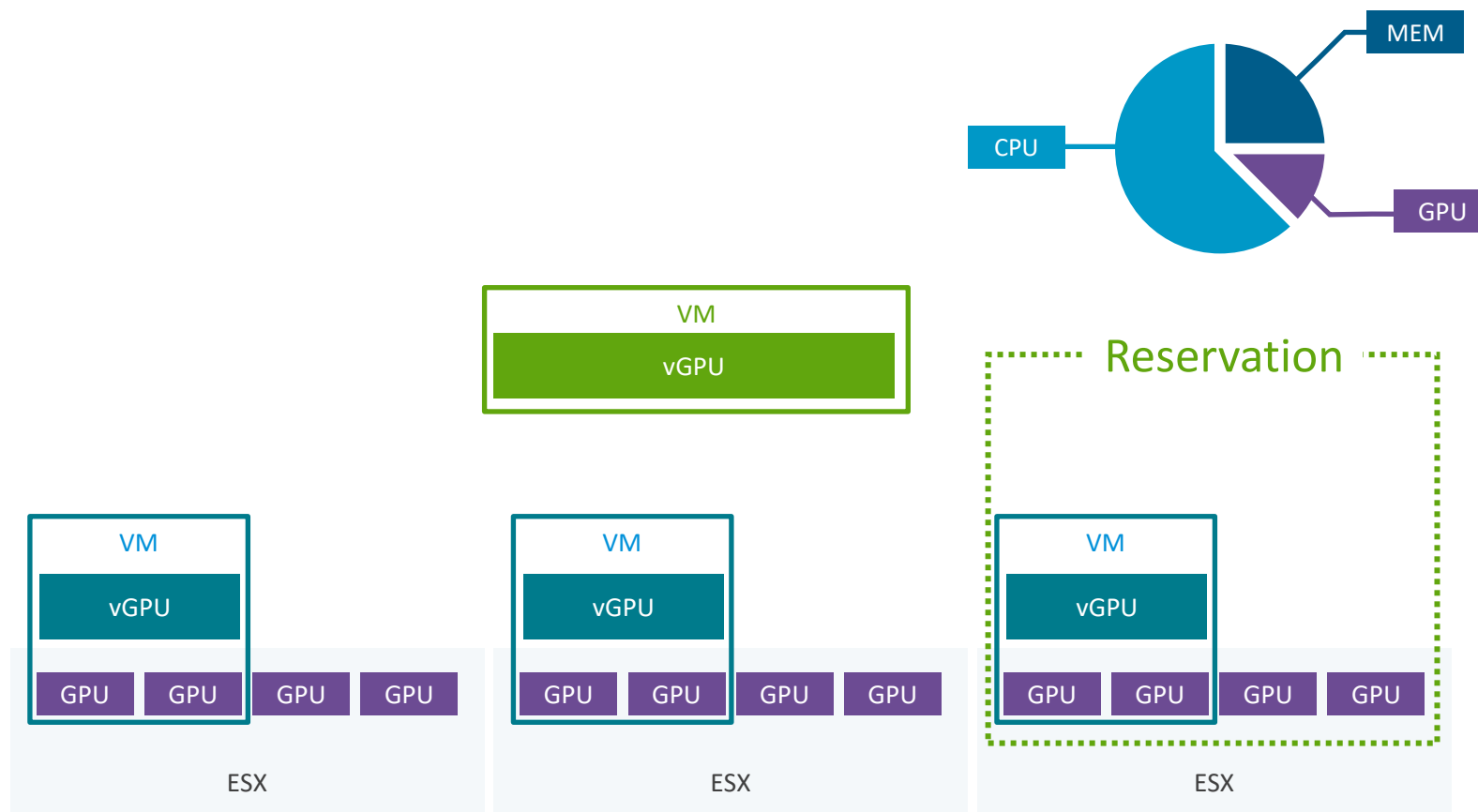
# 隨時了解 vGPU 配置狀態，動態根據需求規劃或預留資源



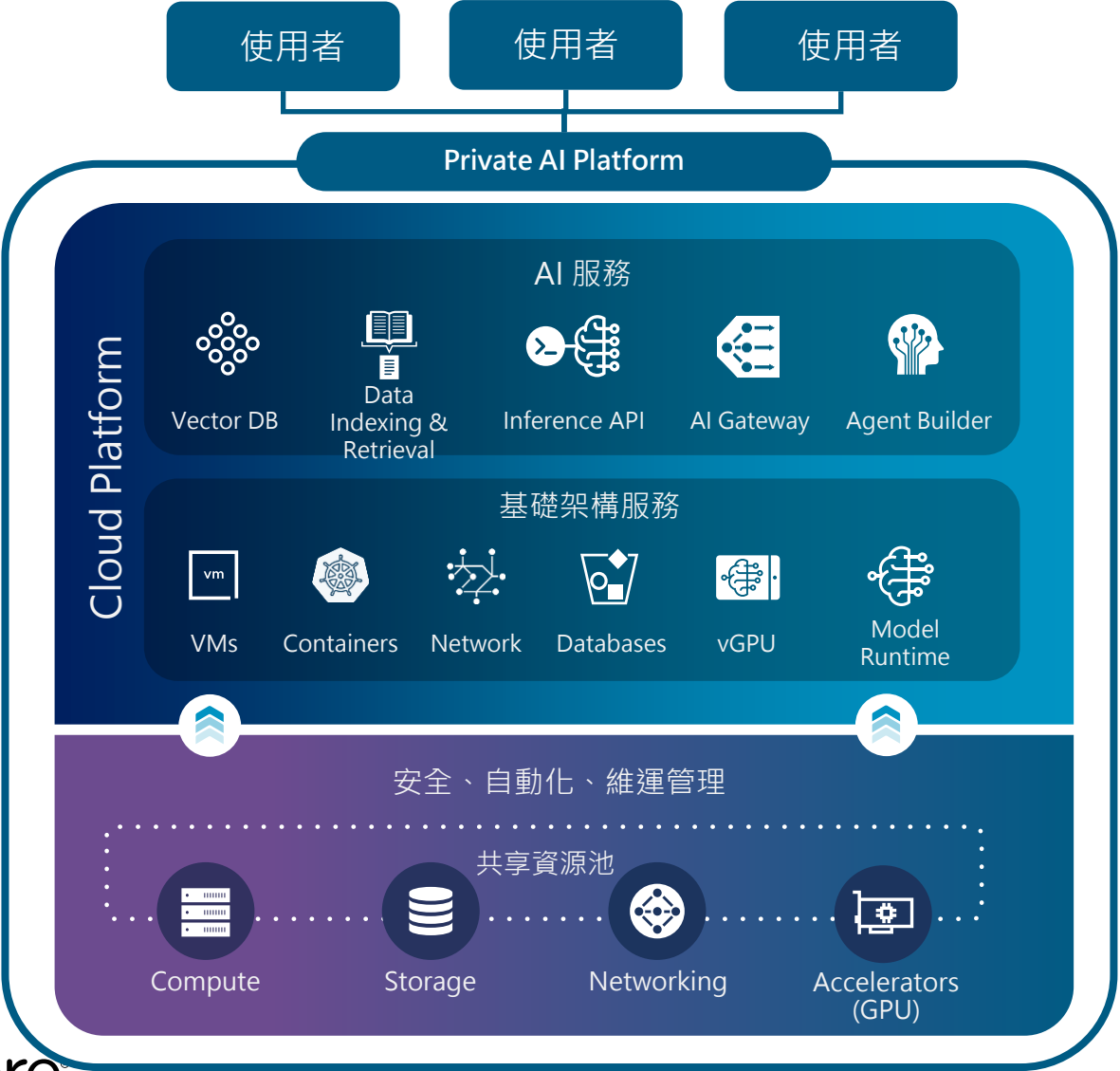
# 高效 vMotion 提高 GPU 利用率、可用性及使用彈性



# “資源保留” 確保高優先級 AI 應用的資源需求



# 小結：Private AI 關鍵能力充分滿足企業 AI 需求



## Private AI 關鍵能力

- > **最大化 GPU 利用率**  
Reduce the TCO of your AI platform by leveraging shared infrastructure model and GPU virtualisation to maximise utilization.
- > **內建 AI 自助服務**  
The Private cloud service portal expanded to include access to AI services and blueprints built to accelerate development and production AI workloads.
- > **整合模型部署**  
Seamlessly push models to production and provide a private OpenAI compatible API for your enterprise.
- > **便捷資料存取**  
Prepare private enterprise data and build custom AI Agents built for your enterprise.
- > **GPU即服務 (GPUaaS)**  
Extend existing PaaS operating model principles for private cloud, to deliver GPU as a service.
- > **一致性平台運維及治理管控**  
Extend existing private cloud controls & operations, to provide compliant, secure & resilient AI services.
- > **安全的 AI 基礎建設**  
Protect data by extending platform security policies and controls from private cloud into private AI.

# VMware Private AI 邏輯架構

以 VCF 企業私有雲為基礎，擴展即選即用的 AI 服務





The diagram illustrates the VMware Cloud Foundation Private AI architecture, showing the flow from data sources to the user interface.

- Data Sources:** Includes Google drive, Confluence, Sharepoint, and MinIO (S3-compatible API). These are connected to a **Data sources** block.
- Model Runtime (1):** Receives input from the **Data sources** block and the **Agent Builder**.
- Data Indexing and Retrieval (2):** Receives input from the **Data sources** block and the **Model Runtime**. It is connected to a **Vector DB**.
- Agent Builder:** A central component that interacts with the **Model Runtime** and the **Data Indexing and Retrieval** block.
- Agent API Endpoint (3):** Receives input from the **Agent Builder** and the **Data Indexing and Retrieval** block.
- User Interface:** Shows a chat interface with a user asking "Is there a VMware validated solution for Private AI?". The **Validated Solutions Helper** responds with a detailed answer about VMware's Private AI Ready Infrastructure for VMware Cloud Foundation.

VMware Cloud Foundation™





# Thank You