

VMware Cloud Foundation 9

標準化 × 自助 × AI 驅動的未來校園

Broadcom, Commercial & Partners, API

Nov 2025

校務應用場境 – AI 落地難

自助服務

AI 應用團隊/師生透過
入口網站一鍵存取AI
工具，無需IT申請。



自建 RAG 最
佳支援

整合大學知識庫，生
成更精準回應。

VCF Private AI Service

真正可落地、安全的「AI教室」或「AI自助服務」方案

師生/研究員/AI應用團隊無需處理複雜的 VM 或 K8s 叢集，
僅需選擇模型與校內資料，即可快速建構基於 RAG 的 AI
應用，專注教學、研究與創新。

資安優先

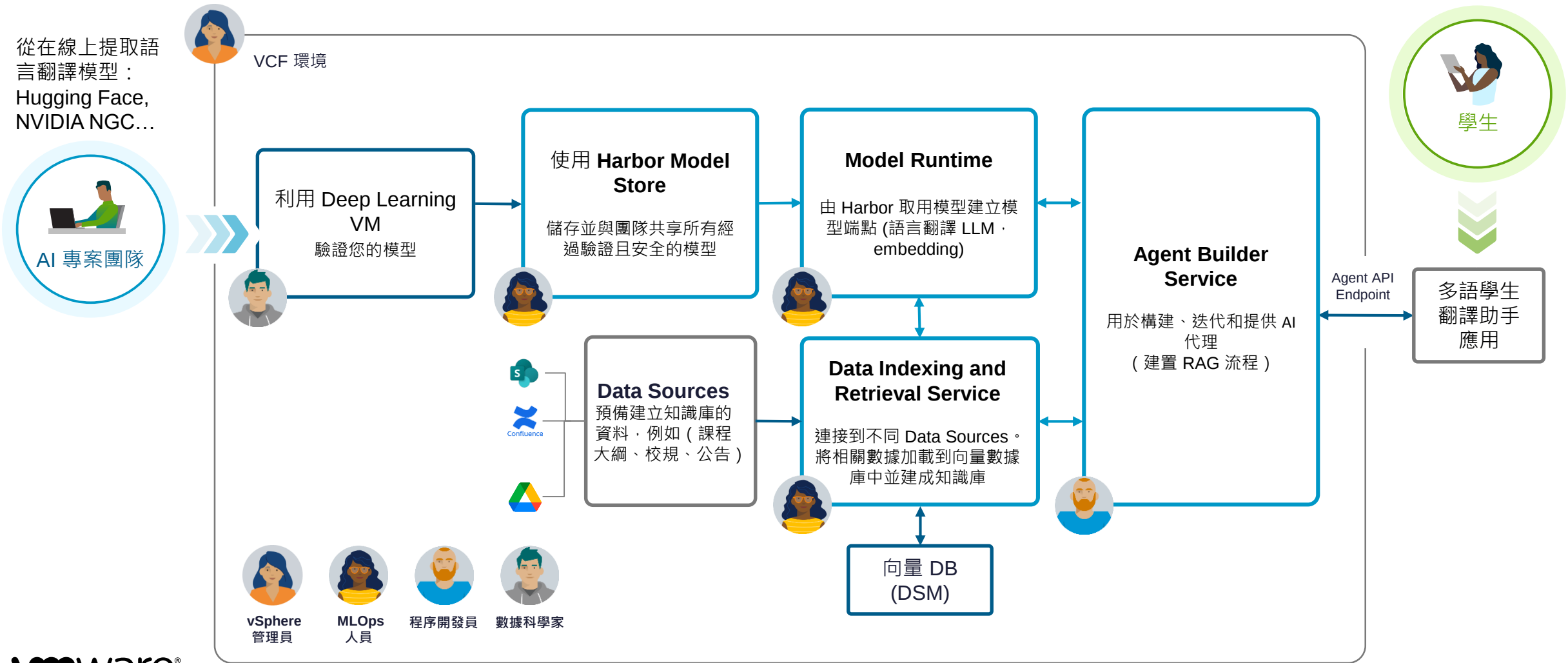
私有雲環境，資料不
離校園，亦保障校園
IP。

快速部署

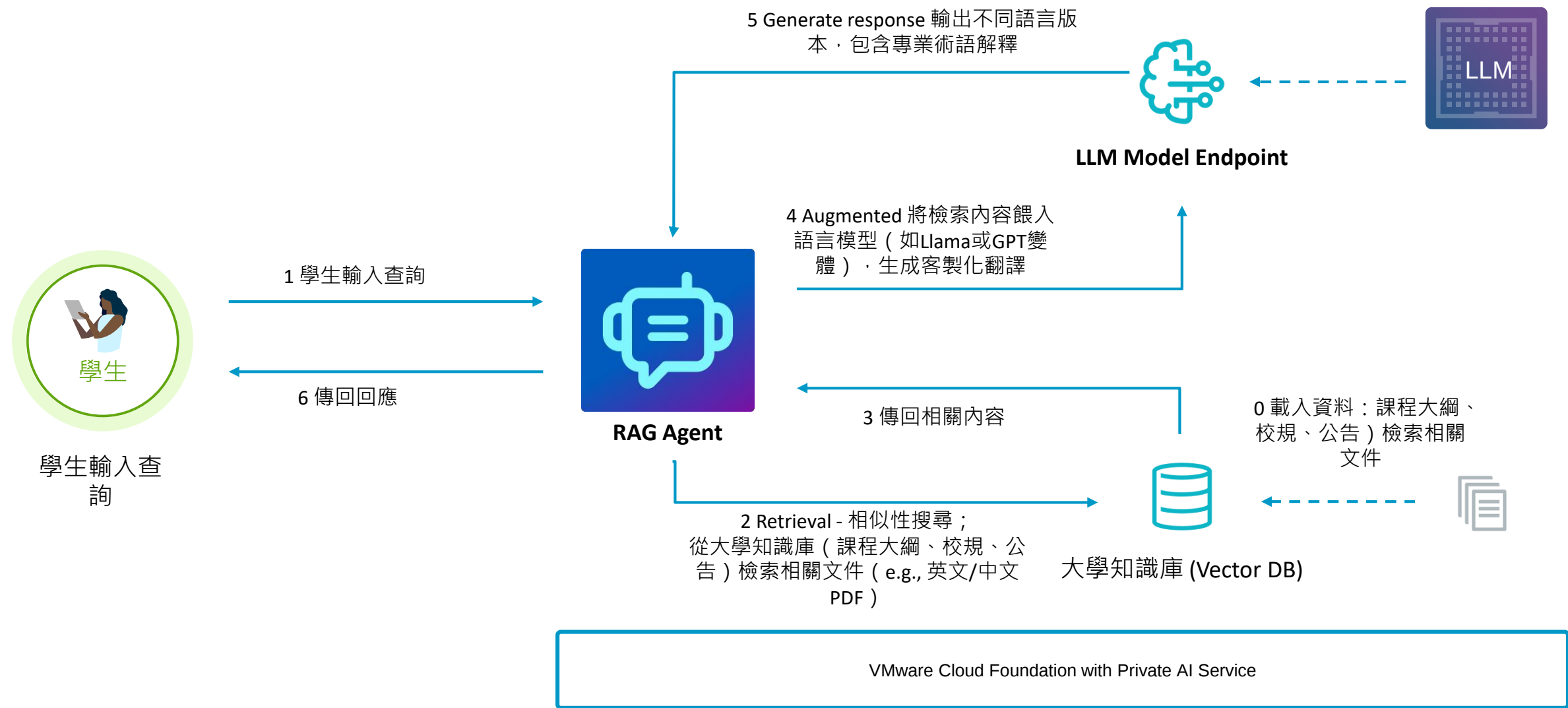
模型運行時與端點數
分鐘內上線，支援
NVIDIA GPU加速。

PAIS 內建 RAG 工作流程，輕鬆建立 AI 驅動的多語學生翻譯助手

AI驅動的多語學生翻譯助手示意圖



RAG 如何提升翻譯效能頁：脈絡智慧生成

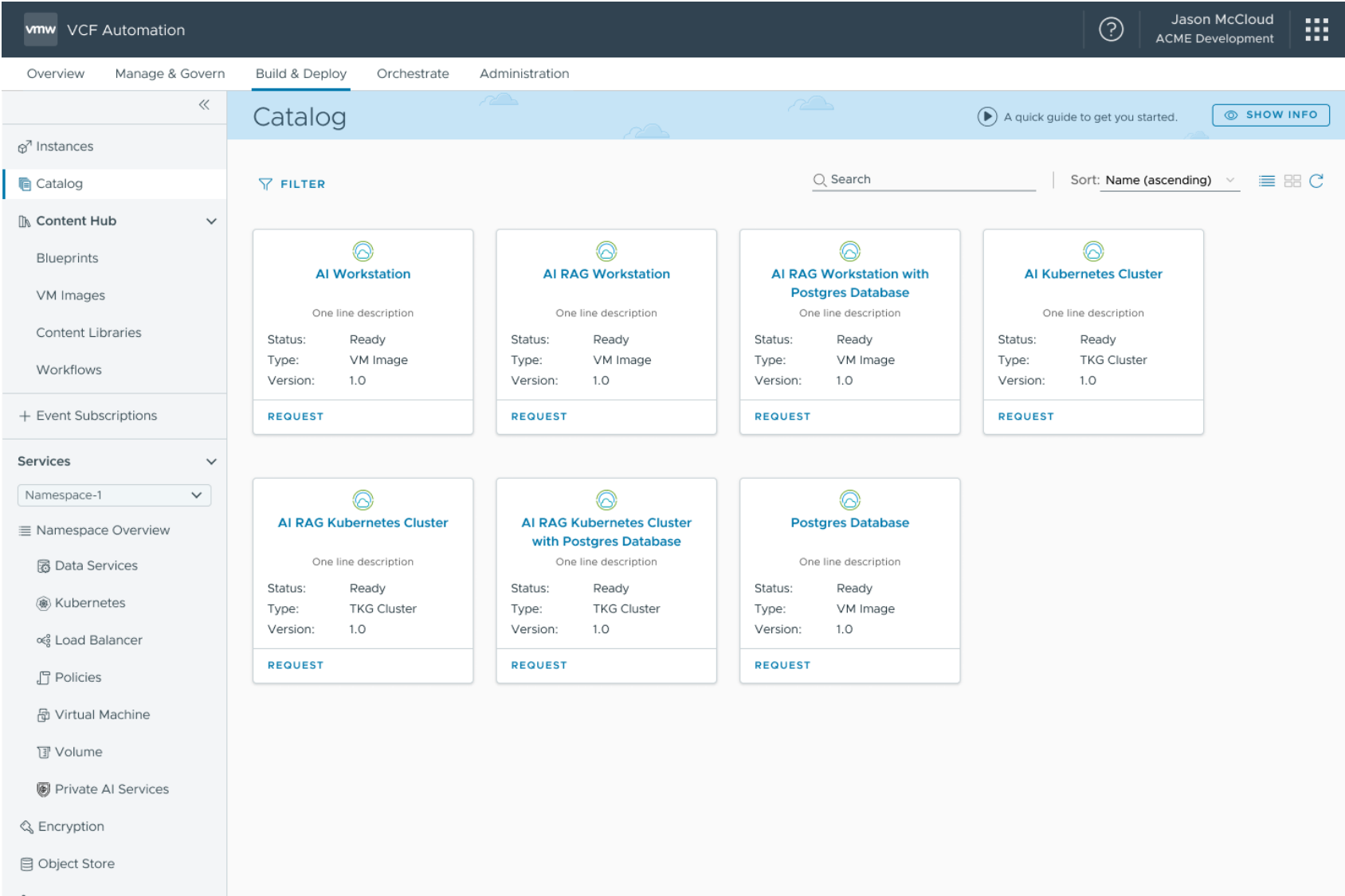




導入「AI教室」或「AI自助
服務」，AI 應用團隊或師生
可以...

PAIS 用戶體驗 – 自助 AI 負載部署及設定

Deep Learning
VM



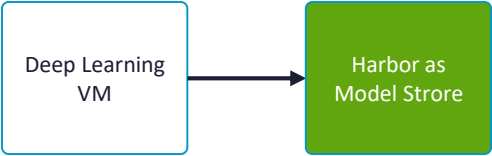
MLOps

PAIS 提供自助 AI 工作負載目錄
以支援 AI 應用程式的設定

MLOps 工程師可以部署 AI 工作
站 (DLVM) ，並預先配置
GPU

驗證我們將上線的模型的安全
和效能

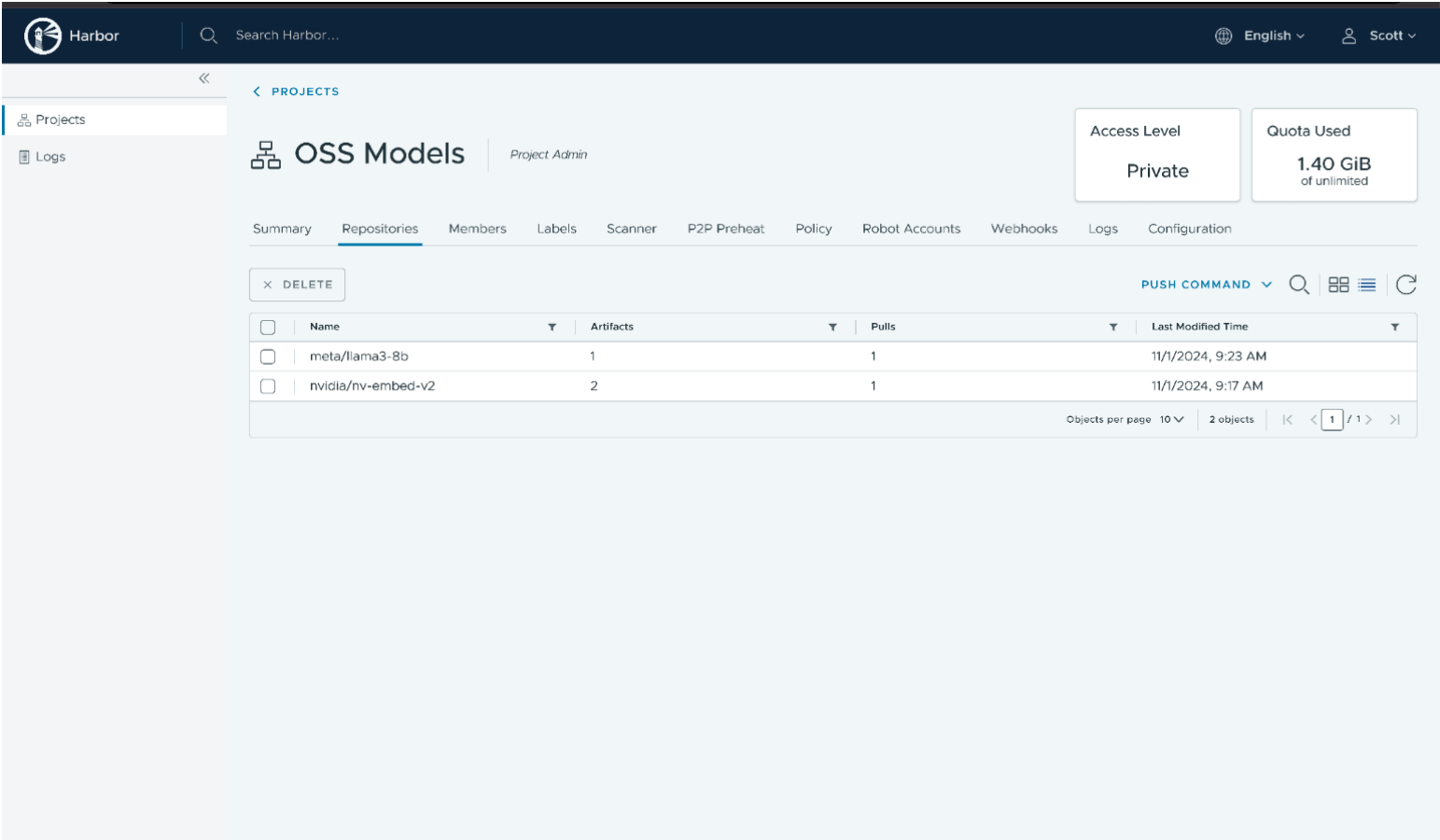
PAIS 用戶體驗 - 模型被推送到 Harbor 模型商店



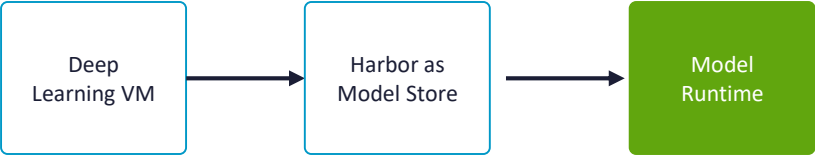
MLOps

MLOps 利用 Harbor 進行

- 管理模型
- 新增出處的 metadata
- 新增標籤以簡化管理
- 建立專案和共用模型



PAIS 用戶體驗 – 自助部署 Model Runtime 及 Endpoint





MLOps

MLOps 工程師可以使用私有 AI 服務來部署模型端點

vmw VCF Automation

Service Broker

Consume Content & Policies Infrastructure Inbox

Project: ALL PROJECTS

Catalog Deployments

Supervisor namespaces

Private AI Services

DATA INDEXING AND RETRIEVAL AGENT BUILDER

Documentation Show Details

Model Runtime Service Configuration



Run your first large language model

To get started with the Model Runtime service, create a model endpoint.

1. Upload model binaries to Harbor

Harbor is an OCI registry where AI/ML model files are stored.

Use the PAIS CLI client to upload validated model files to a project in Harbor identified as a central model gallery for your organization.

LEARN MORE

2. Prepare connection to Harbor

Create a robot account for each Harbor project that is identified as a model gallery, and store those credentials as Kubernetes secrets in this VCF namespace.

Model Runtime will use the secret to pull files from the project to deploy a model endpoint.

LEARN MORE

3. Deploy a model endpoint

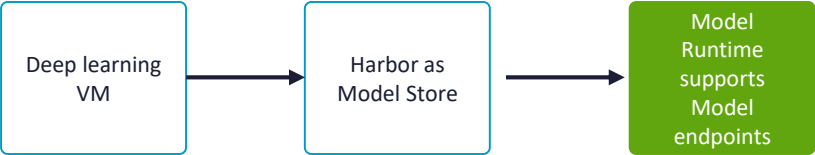
Select a model from a Harbor project, define the model endpoint specification, and select the VM and storage classes, and start the deployment.

DEPLOY MODEL ENDPOINT

vmware
by Broadcom

Broadcom Proprietary and Confidential. Copyright © 2024 Broadcom.
All Rights Reserved. The term “Broadcom” refers to Broadcom Inc. and/or its subsidiaries.

PAIS 用戶體驗 – Model Runtime Service





MLOps

此服務支援熱門的推論引擎
為完成和內嵌模型建立端點。

vmw VCF Automation

Service Broker

Consume Content & Policies Infrastructure Inbox

Project: ALL PROJECTS

Catalog Deployments Supervisor namespaces

Deploy Model Endpoint

Model Information

2. Endpoint Specification

Routing Name

meta/llama-3-7B

Display Name

meta-llama3-7B

Model Type

Completions

Engine

NVIDIA Triton vLLM

Flags or Arguments for Engine Customization

Argument/Flag

Enter a flag name or argument for the engine

ADD ARGUMENT

Override Environment Variables

Variable Name

Value

var2_name

var_value_2

ADD VARIABLE

Override Configuration

01

02

03

04

05

NEXT

YAML

Model Endpoint

object1

01

02

03

04

05

06

07

08

09

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

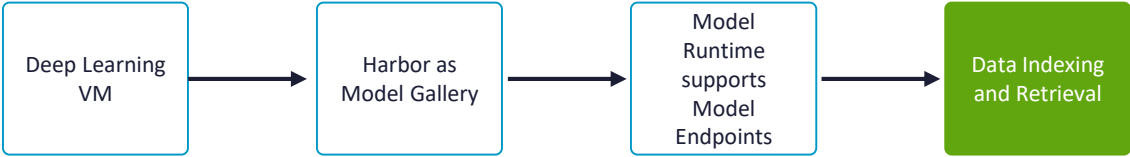
27

28

29

DOWNLOAD .ZIP

PAIS 用戶體驗 – 自助校園知識庫建立



vmw Private AI Services

Context: ml-project/prod-namespace

Scott Davis

Data Indexing and Retrieval

Knowledge bases

Data sources

Agent Builder

Agents

Data Indexing and Retrieval > Knowledge Bases > Customer Support Data

Customer Support Data

DELETE KNOWLEDGE BASE

Knowledge Base Details

Knowledge base type

Data source

Created on

7/1/2024

Description

Knowledge Base for Support Data.

...

...

Index refresh policy

Weekly

(Every Thursday at 12:00PM)

Last index

--

Next index

--

EDIT DETAILS

Data Sources

Customer Support

Indexed

Accounting Sheet

Indexed

ACTIONS

Index Settings

Index status

Completed (10/1/2024, 11:33AM)

Embedding model

nvidia/neva-22

Chunk size

xxx

Token overlap

0

Text splitter

Sentence

RE-INDEX KNOWLEDGE BASE

Document List

Document name	Status
doc 1	Indexed
doc 2	Indexed
doc 3	Indexed
doc 4	Indexed
doc 5	Indexed
doc 6	Indexed
doc 7	Indexed

vmw Private AI Services

Namespace: default

Data Indexing and Retrieval

Knowledge Bases

Data Sources

Agent Builder

Agents

Data Indexing and Retrieval > Data Sources > Add Data Source

Add Data Source

Required fields are marked with *

1. General Information

Enter name and description

2. Data Source Details

Select data source type and provide credentials

Select the type of data source you want to create and enter data location information and credentials. Learn more about data source types

Supported file types: .pdf, .html, .md, .pptx, .docx, .txt, .csv file formats.

Data Source Type

Select source type

Google Drive Folder

Confluence Space

Amazon S3 Compatible Store

Sharepoint Enterprise Site

NEXT

3. Review and Complete

Review details



MLOps

資料服務系統管理員

MLOps 或資料服務系統管理員會連線到知識庫，匯入資料、將其分塊，並產生 embedding 以用於 RAG 應用程式。

vmware

by Broadcom

Broadcom Proprietary and Confidential. Copyright © 2024 Broadcom.

All Rights Reserved. The term “Broadcom” refers to Broadcom Inc. and/or its subsidiaries.

PAIS 用戶體驗 – 自助知識經濟建立



MLOps



Data Service
Admin

MLOps 工程師或資料服務系統
管理員會使用不同的
Embedding 模型和區塊化
(Chunking) 策略，自訂/試
驗其知識庫。

Index Knowledge Base



Select an embedding model, configure index settings, and index the knowledge base.

Embedding model endpoint * ⓘ

Select



Text splitter * ⓘ

Sentence

Chunk size * ⓘ

Enter maximum number of tokens

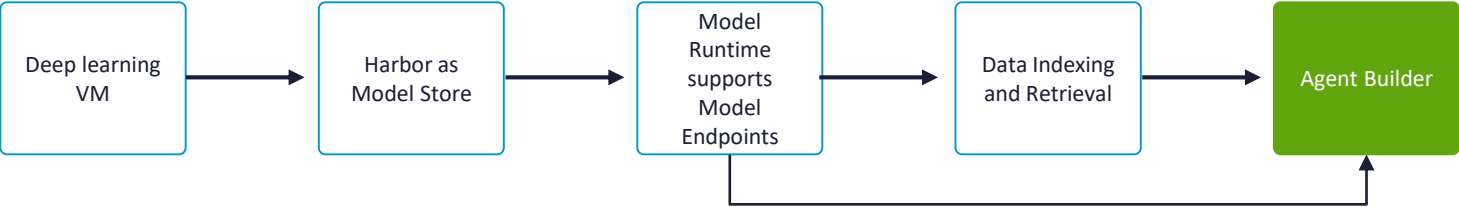
Token overlap * ⓘ

0

CANCEL

SAVE SETTINGS

PAIS 用戶體驗 – 自助 Agent 建立及測試場



Developer

開發人員建置、測試和評估代理程式

vmw Private AI Services

Context: Default/ai-namespace

Agent Builder > Agents > Customer Support Agent

Customer Support Agent

EDIT ACTIONS

Data Indexing & Retrieval

Knowledge Bases

Data Sources

Agent Builder

Agents

Agent Details

Agent ID: 123ABC

Status: Available

Description: A chatbot to answer customer support requests

Model Endpoint: meta/llama3

Created on: 11/4/2024

Instruction

Instruction: You are a friendly and professional customer support agent. Always answer questions in the follow format:

Delivery Method: Send as system message

Data Indexing & Retrieval

Knowledge Base: customer support (Available)

Chunks to Retrieve (Top K): 5

Data Reference

Playground

CLEAR NEW SESSION Use Session

You: What is VMware Private AI Foundation with NVIDIA?

Customer Support Agent: VMware Private AI Foundation is a collaborative solution developed by VMware and NVIDIA. It is an architectural approach for AI services that enables privacy, security, and control of corporate data, as well as integrated security and management. It offers several benefits, including:
1. Data-center scale: Multi-GPU IO pathing allows AI workloads to scale across up to 16 vGPUs/GPUs in a single virtual machine and across multiple nodes to speed up generative AI model customization and deployment.

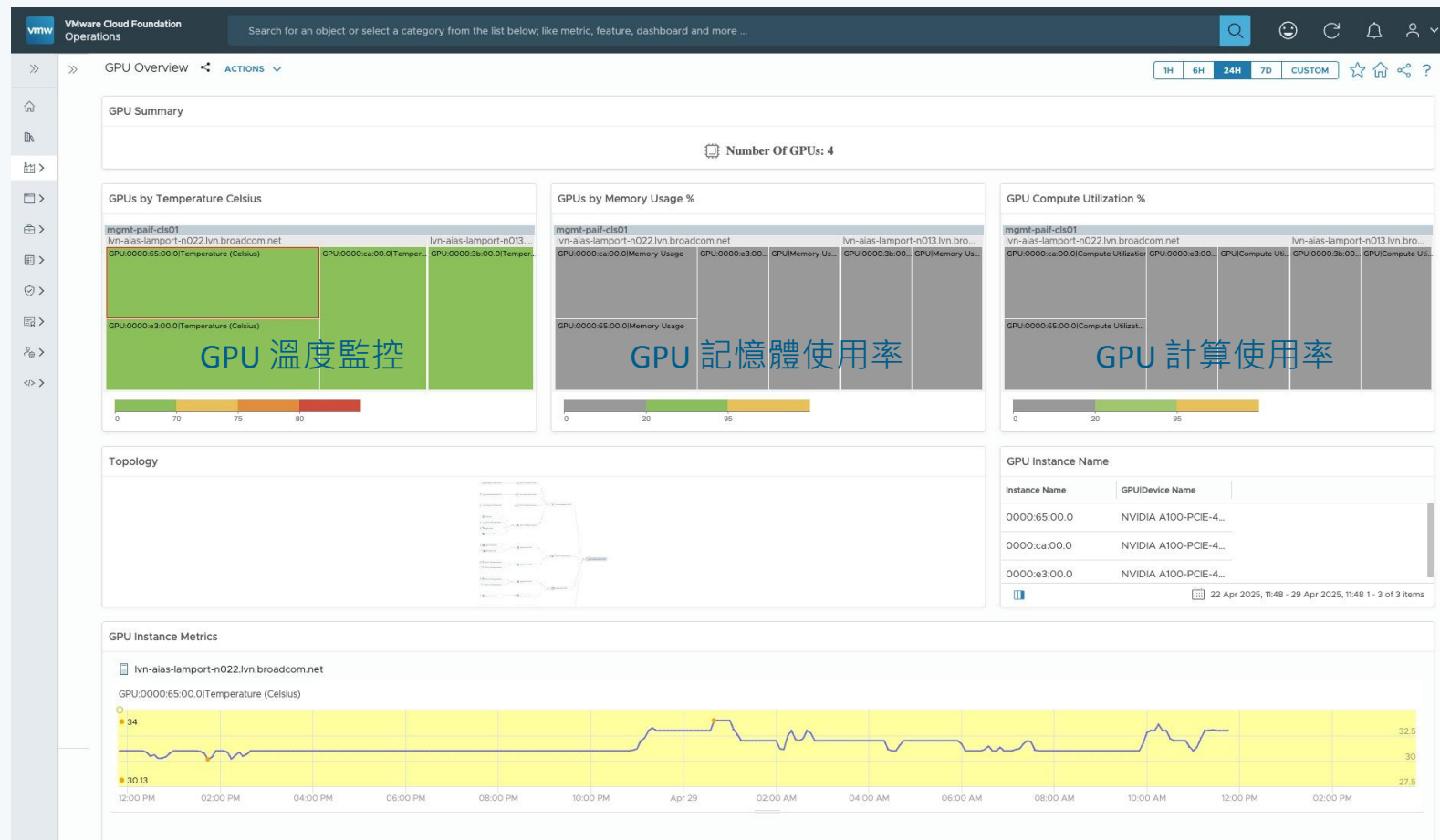
Generating response

Type to chat with the agent

SEND

PAIS 用戶體驗 – 一站式 GPU 使用監控

監控用於專用 AI 的 GPU 資源



- 即時詳細檢查 GPU 資源
- GPU 溫度
- 影格緩衝區記憶體使用量
- 計算使用率
- 發現問題及早遷移 AI 負載



客戶成功案例

德州高等教育聯盟 × VMware Private AI

客戶

- 德州最大高等教育集團（多校園、數萬師生）

需求與挑戰？

- 希望提供 AIaaS 給各大高等院校足進教學創新
- 需求：安全、私有、快速、低成本的「AI 即服務」
- AI 熱潮爆發，大家都在喊但做不出來

怎麼做到的？

- 選用 VMware Private AI Foundation + NVIDIA
- 既有 VCF 直接升級，3 天就跑通 PoC（用 Open Web UI 工具測試）
- 預裝 GPU 伺服器 + 一鍵部署 Llama 3 70B
- Open WebUI + RAG + NIM 微服務，數分鐘上線新模型

客戶效益

- 專案成功，達成「AI 即服務」目標
- 引起州內校園看到成果後都希有興趣加入聯盟 Private AI 生態



美國德州

高等教育行業



PAIS

德州高等教育機構透過

VMware Private AI Foundation

與 NVIDIA 重塑教育的未來

校方原話：「我們對成果非常滿意！沒有 Broadcom 與 NVIDIA 的全力支援，絕對不可能這麼快做到這麼多！」
(原文：We are very pleased with the outcome... Has Broadcom and NVIDIA been very supportive partners through it all? Absolutely!)

PAIS 客戶效益



簡化模型管理及部署：

師生或研究團隊從 Hugging Face 或 NVIDIA NGC 取模型
→ 校內驗證 → 推至 Harbor 集中管理 (含掃描、存取控制、版本追蹤) → 部署為高效能端點，供課程與研究使用。



靈活整合校內數據，打造系所知識庫：

支援 Google Drive、SharePoint、S3 或圖書館系統 → 自動分塊轉向量、定期更新 → 建立系所專屬 RAG 知識庫，提升教學與研究準確性。



快速構建校園 AI 代理：

師生或研究團隊透過簡易介面選擇模型端點 + 知識庫，快速部署 RAG 基礎的 AI 代理 (如課程問答機器人、實驗輔導助手)。提供簡單的界面和指令設定，允許開發者測試與優化代理行為，無需深入技術細節。



提供開放標準且安全的模型端點存取：

提供 OpenAI 規格相容 API，內建負載平衡與校園 SS 身份驗證，確保高效、安全的模型存取，符合大學資安與治理規範。



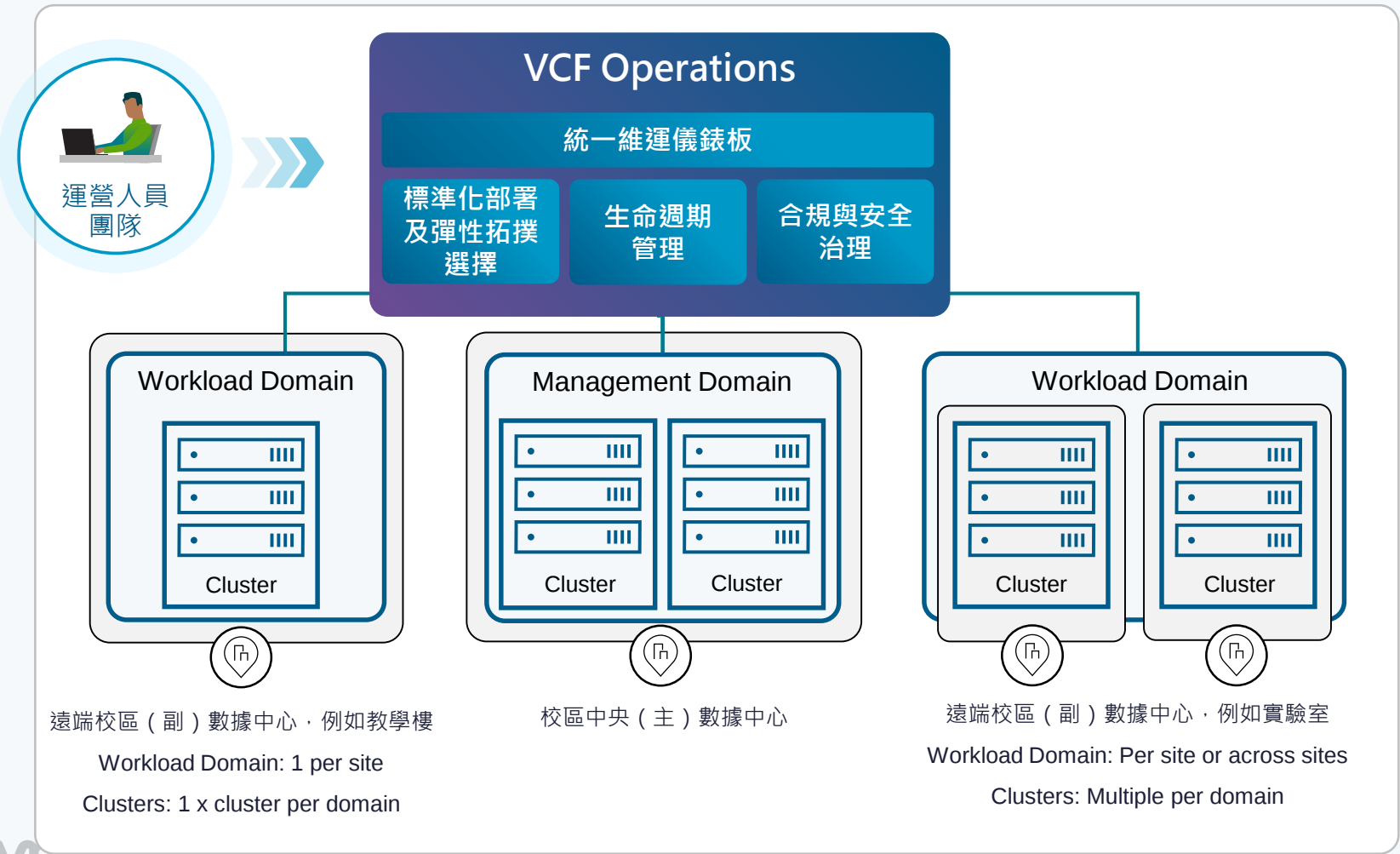
Private AI Service PAIS 讓師生專注 AI 創新

師生/研究員/AI應用團隊無需處理複雜的 VM 或 K8s 叢集，僅需選擇模型與校內資料，即可快速建構基於 RAG 的 AI 應用，專注教學、研究與創新。

資訊服務應用場境 - 資安合規壓力大



各校區都在同一標準上進行佈建、拓展，提升運營效率及資安



客戶效益

- **VCF 標準化佈建**
 - 各校區統一標準，無需重複設計，快速拓展校區 IT。
- **VCF Operations Fleet Management 遠端生命週期管理與監控**
 - 提供統一儀表板遠端管理多站點，IT 無需出差即可完成升級與監控，確保資安及預防問題於未然

基於 VPC 的自助網路服務及安全隔離



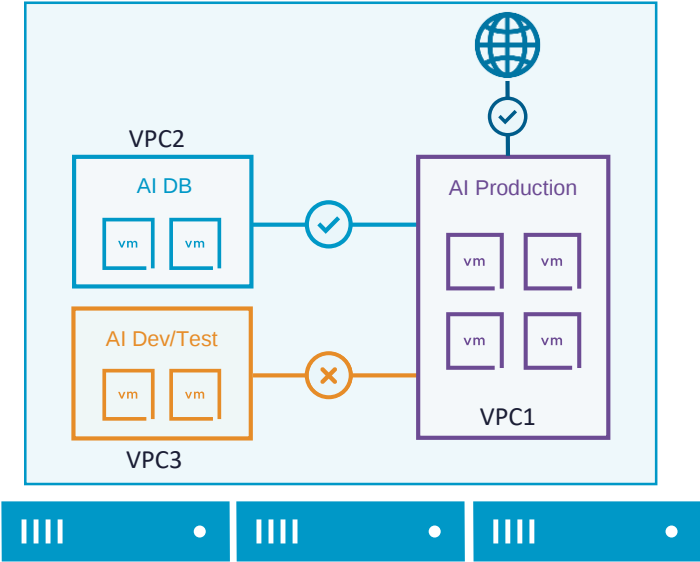
運營人員
團隊



部署應用程式在已具備適當連線與獨立分隔的VPC內



應用/師生/研究
員團隊



識別主機	X
建立VLAN	X
設定路由、防火牆規則與負載平衡	X

避免複雜的 ticketing 處理與基礎設施配置 (已包含在VPC配置中)

客戶效益

VPC 內建連線與自助網路服務 (NAT、防火牆規則和負載平衡。)

應用團隊與學生專案無需等待網路設定，數分鐘內完成部署，應用快速上線。

VPC 最佳安全實踐

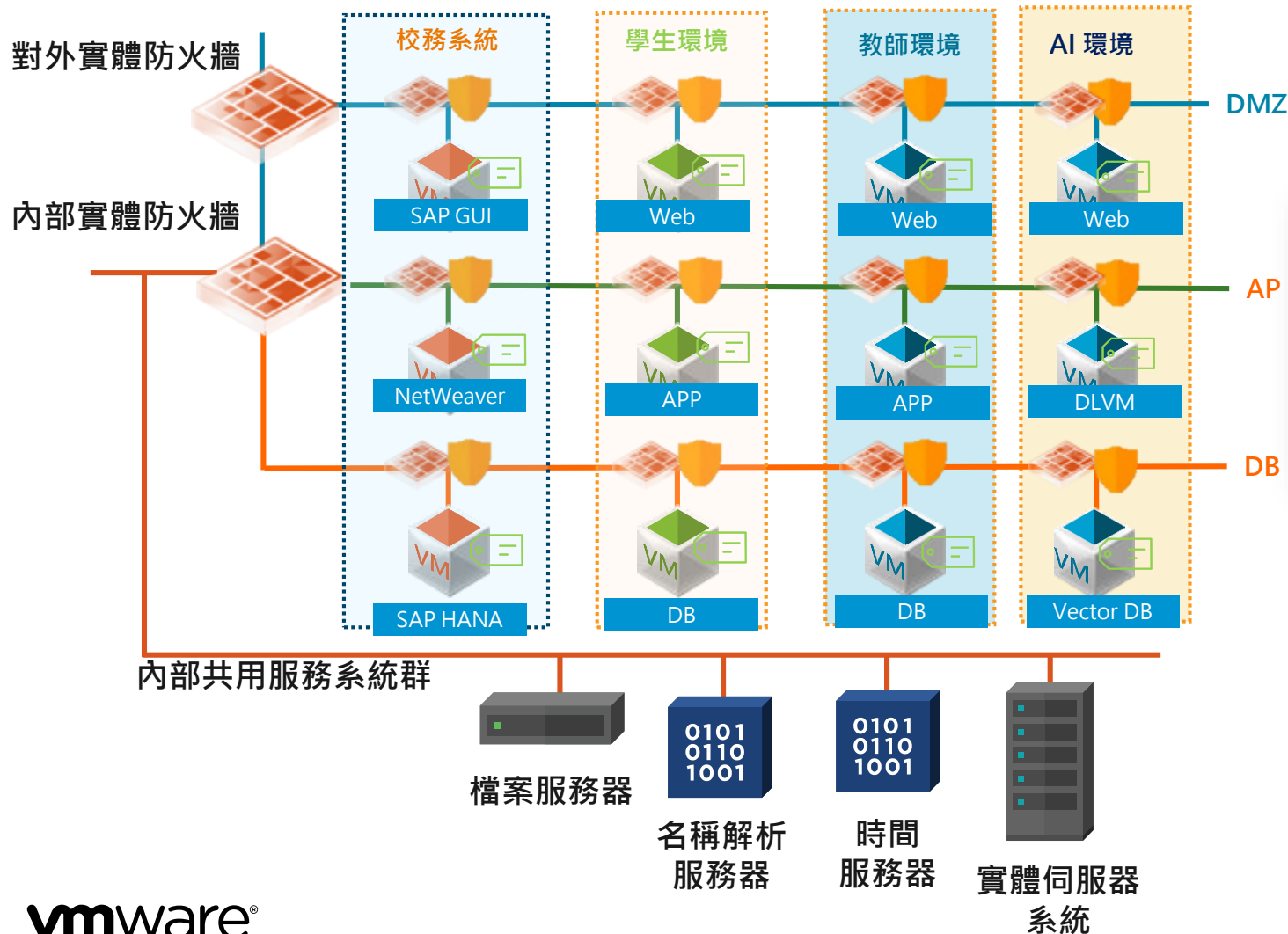
VPC 提供更簡易的分隔工作流程為不同 IT 及學系專案建立獨立 VPC (如資工 vs. 醫科)，避免相互干擾，打造更安全網絡環境。

無 ticketing 自助運維

應用團隊自主管理 VPC 資源，免除繁瑣票證流程，實現規模化部署與資安合規，顯著減輕 IT 運維負擔。

基於政策微分段安全防護，透過【軟體定義安全標籤】分類管理

僅同業務系統間，進行最低限度必要溝通



微分段防護技術是【零信任網路安全】概念的實作

- 1 定義網路微分段安全防護規則
- 經VMware網路流量分析系統，蒐集業務系統流量後，提供防護規則制定建議



客戶效益

- 2 透過【軟體定義安全標籤】管理網路微分段安全防護

- 將安全標籤附加到特定負載後，網路微分段安全防護規則自動套用，省時省心，降資安等險。
- 如: 僅附加【校務系統】安全標籤的虛擬機，彼此才可互相溝通

VMware Tech Summit- 2025/12/9

報名網址：<https://ems-event-tw.neocities.org/20251209/>



by Broadcom



VMware Tech Summit 2025

台北文創 | 12月9日

誠摯邀請您參加將於 12 月 9 日在台北文創舉辦的 **VMware Tech Summit 2025**，一場技術願景與實踐創新相結合的盛會。活動的重點包含以 **VMware Cloud Foundation (VCF)** 為核心的現代化私有雲、應用現代化，以及 **Private AI** 方案。



2025 年 12 月 9 日 (二)



09:00 - 17:00



台北文創
(台北市菸廠路 88 號 6 樓)

立即報名

參加 VMware Tech Summit 的理由

- 

深入瞭解現代化私有雲和 Private AI 的專業知識
- 

向 VMware 專家、同儕和業界先進吸取技術能量和經驗
- 

取得優化成本、安全和營運的實用技能
- 

建立與 IT 從業者的交流與人脈

活動亮點

- 

各行業用戶現身說法分享眾多關於 VCF 及 Private AI 的經驗與案例
- 

介紹在 VMware Explore Las Vegas 大會上發佈的創新成果
- 

深入探討 VMware Cloud Foundation 9.0 的主要功能
- 

有關準備和運行現代私有雲和 Private AI 的實用指南
- 

深入探討多種技術主題和最佳實作

不要錯過此難得的機會，強大您的 VMware 產品技能，
並掌握私有雲和 AI 的未來！
請即刻報名，期待與您在 VMware Tech Summit 2025 相見！

聯絡資訊

Account Director

姓名: Edison Chen, 陳奕廷

電話: 0912910257

Line ID: edd0820

Email: edison.chen@broadcom.com





Thank You