VMware Private AI
雲平台與案例分享

Evan Huang
Broadcom

2024/11/05

Agenda

- Private AI 雲平台 – Why & What?
- Private AI 案例分享

AI 浪潮來襲

"We need to make AI a priority in our business."

- Everyone

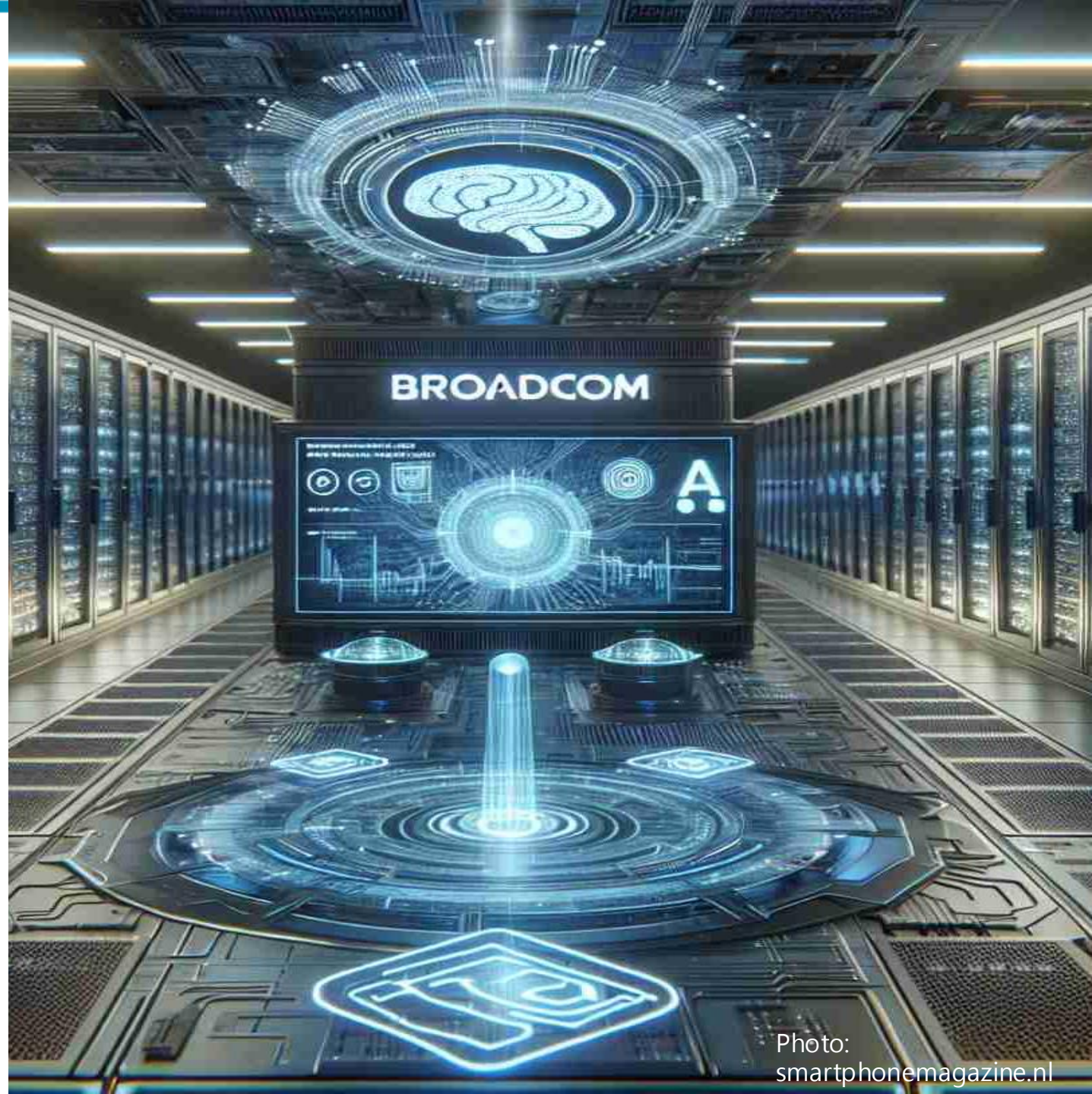


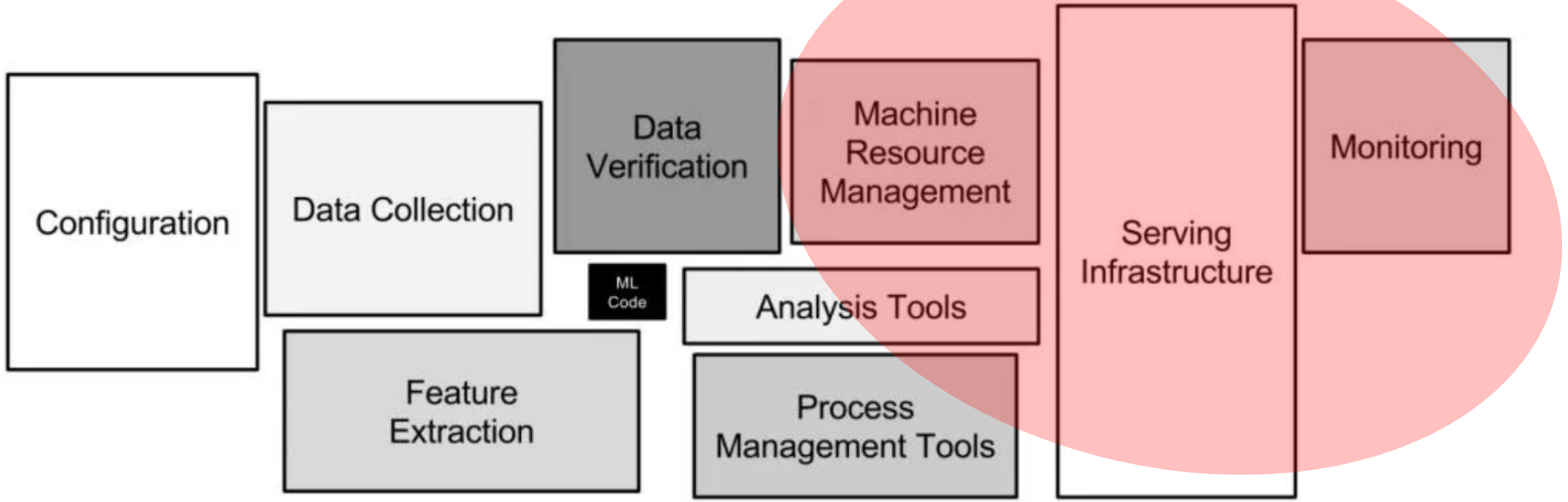
Photo: smartphonemagazine.nl

為何 AI 需要雲平台？

所有資料處理、AI 運算、AI 應用都需要在安全、彈性、高效的平台環境運行

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.




All software will run on cloud
+ " AI"
高自動化
彈性伸縮
自助服務
安全合規
自我優化
高韌性

AI 在可靠性及維運角度方面的挑戰 – 透過 " 雲維運模式" 優化

典型 AI 訓練場景，近 8 成非預期系統異常及服務中斷，可歸因硬體故障及維運流程的影響

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

Source: The Llama 3 Herd of Models, Meta, 2024/07/23

為何企業 AI 需要 “自主可控” 的雲平台？

對企業而言，“安全隱私” 是重中之重

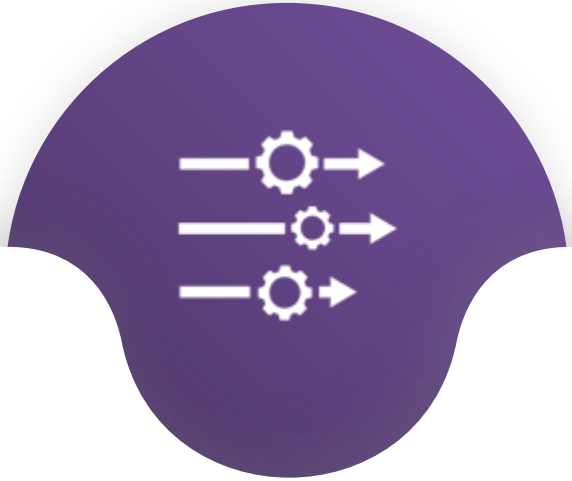
智慧財產權

“須降低智慧財產權外洩的風險，否則再大再強的企業也可能毀於一旦”

(企業/客戶) 機敏資料

“須確保內部及客戶機敏資料絕不外洩，避免監管處罰、鉅額交易損失、及天價損害賠償”

存取權限



“須嚴格控管高價值 AI 模型及資料的權限，保障業務營運

全球金融業對 AI 安全隱私戒慎恐懼

JPMorgan Chase Restricts Staffers' Use Of ChatGPT

Siladitya Ray Forbes Staff

Siladitya Ray is a New Delhi-based Forbes news team reporter.

Follow

Feb 22, 2023, 07:21am EST

TOPLINE JPMorgan Chase has restricted the use of ChatGPT by its staff, Bloomberg and the *Telegraph* reported, becoming the latest organization to limit the use of OpenAI's chatbot in the workplace following the likes of Amazon and several U.S. universities.

Workers' ChatGPT Use Restricted At More Banks—Including Goldman, Citigroup

Brian Bushard Forbes Staff

Brian is a Boston-based Forbes breaking news reporter.

Follow

Feb 24, 2023, 12:14pm EST

TOPLINE CitiGroup, Bank of America, Deutsche Bank, Goldman Sachs and Wells Fargo have restricted employees' use of ChatGPT, Bloomberg and Financial News reported Friday, joining JPMorgan Chase, as well as Amazon and multiple major public school districts to limit the use of OpenAI's new chatbot, which has taken the internet by storm and raised concerns about sensitive information sharing.

全球高科技業對 AI 安全隱私嚴密控管

iThome

K8s & 雲原生講者徵求　2024 臺灣雲端大會熱烈招生中　iThome 線上研討會免費收看

新聞

員工外洩內部機密！三星開放ChatGPT後出事緊急限縮使用

The Register、Tom's Hardware等媒體引述南韓當地媒體Economist的消息，指出三星員工在不清楚ChatGPT使用規範下，為了工作之便直接將半導體設備、程式碼相關資訊上傳給ChatGPT進行處理，導致三星的內部機密資料外洩

文/ 林妍溱 | 2023-04-07 發表

讚 223　分享

三星的資料外洩事件是由南韓媒體《Economist》首先報導，Economist宣稱三星員工逕自將公司機密資訊輸入ChatGPT，導致該公司半導體設備以及內部會議等資料外洩。(圖片來源 / 三星)

防機密外洩！蘋果限制員工用ChatGPT等工具

蘋果公司　機密外洩　ChatGPT　生成式AI

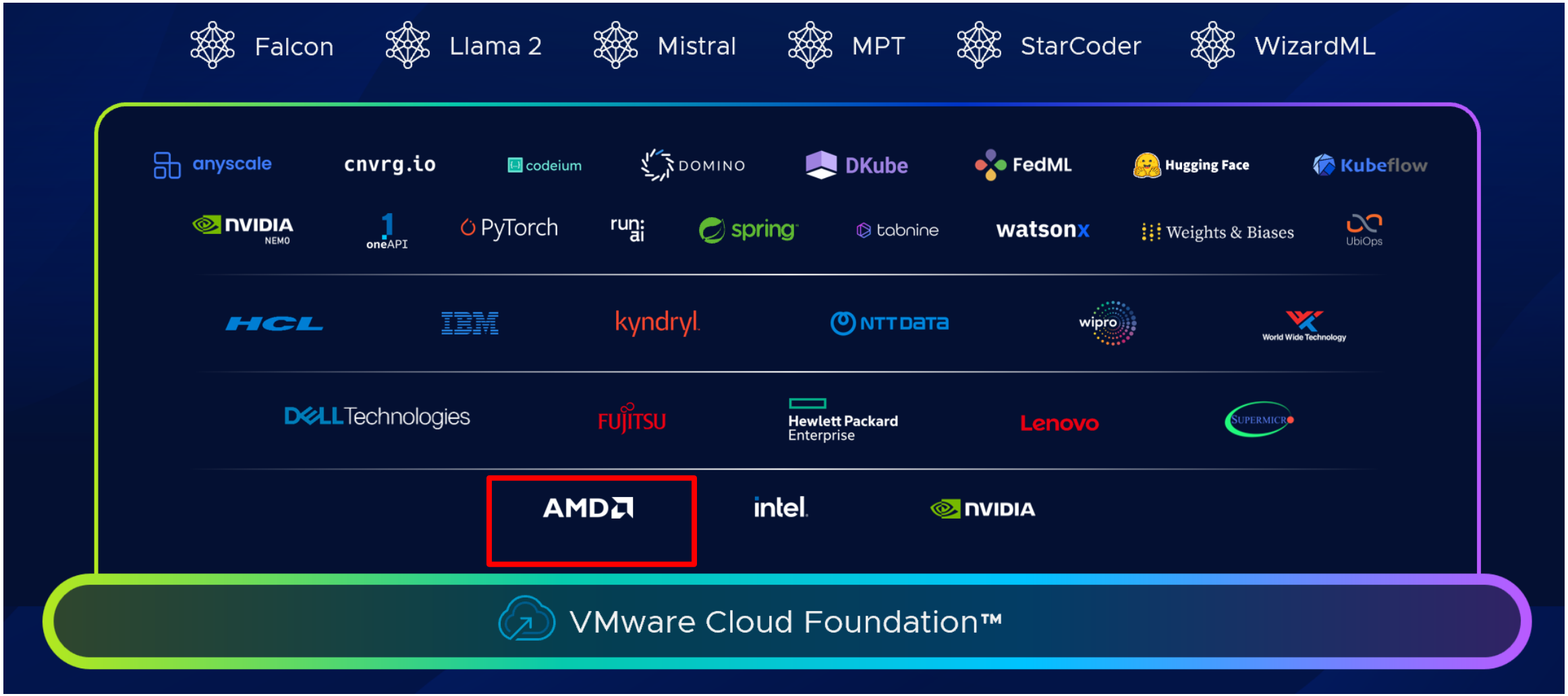
時間：2023-05-19 16:20　新聞引據：採訪、華爾街日報　撰稿編輯：陳文蔚

讚 5　分享

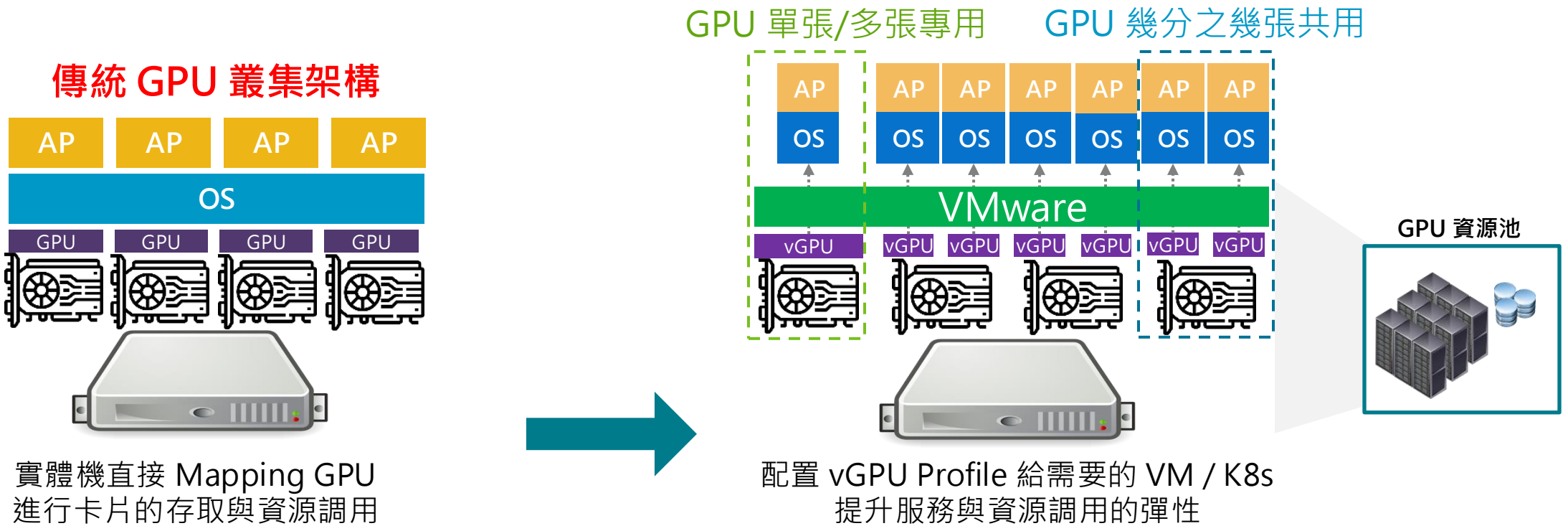
為防範機密外洩，蘋果公司(Apple)已限制內部員工使用ChatGPT以及其他AI工具。(unsplash圖庫)

解決方案：建立在 VCF 雲平台上的開放 VMware Private AI 生態系

AMD 是生態系中的重要基礎



實體 GPU 資源採虛擬化 vGPU 形式提供給使用者



實體機直接 Mapping GPU
進行卡片的存取與資源調用

配置 vGPU Profile 給需要的 VM / K8s
提升服務與資源調用的彈性

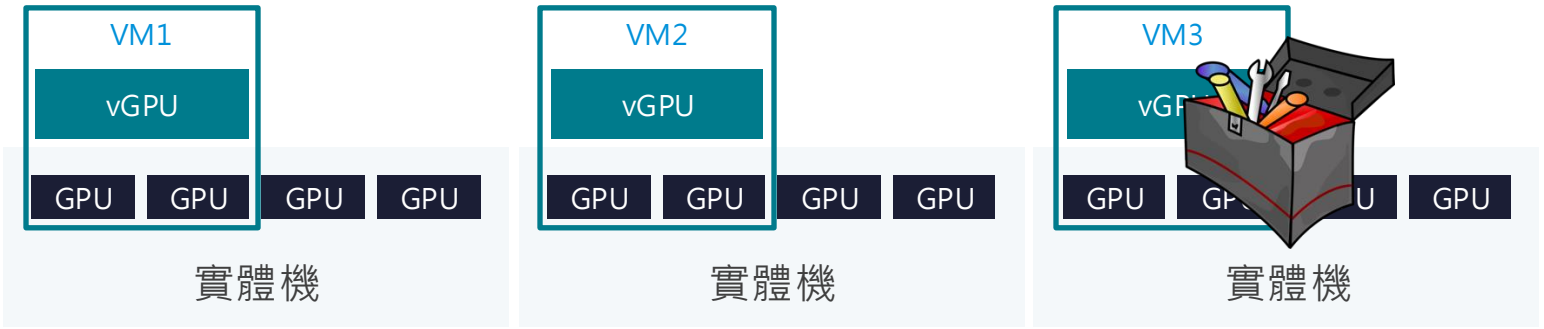
- ✗ 部署服務費時費工
- ✗ 較難分析 GPU 資源使用率
- ✗ GPU 算力資源調用彈性低
- ✗ 缺乏叢集層級的高可用保護

- ✓ 透過 VM 或 K8s 部署環境**簡單快速**
- ✓ vCenter 即支援 vGPU 與 實體 GPU 資源使用率，以**圖形化介面**展示
- ✓ GPU 算力資源調用**彈性高** (可採取Time-Sliced或MIG方式切分)
- ✓ ESXi 高可用叢集提供服務**高可用**保護，可快速重啟服務

vmware by Broadcom

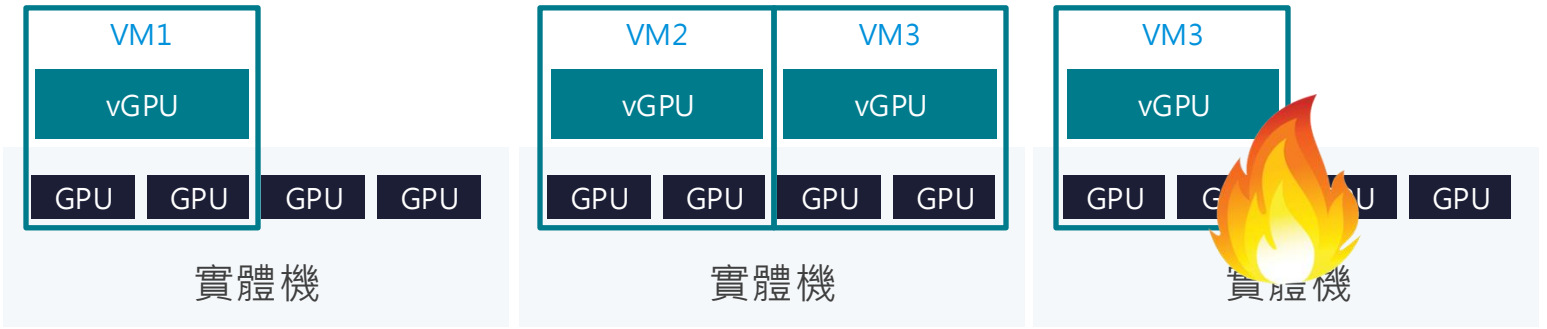
場景：軟硬體維護更新，GPU 負載可藉 vMotion 無縫移轉到其他機器

- vMotion (動態遷移) 可隨作業需求轉移 GPU負載運行位置
- VM HA (高可用) 可在異常狀況發生時，自動切換到其他機器並重啟
- DRS (分散資源配置) 可平均配置 GPU 工作負載，提高整體效能及使用率
- 當平均配置GPU資源造成碎片化時，可自動動態調度，優化資源配置



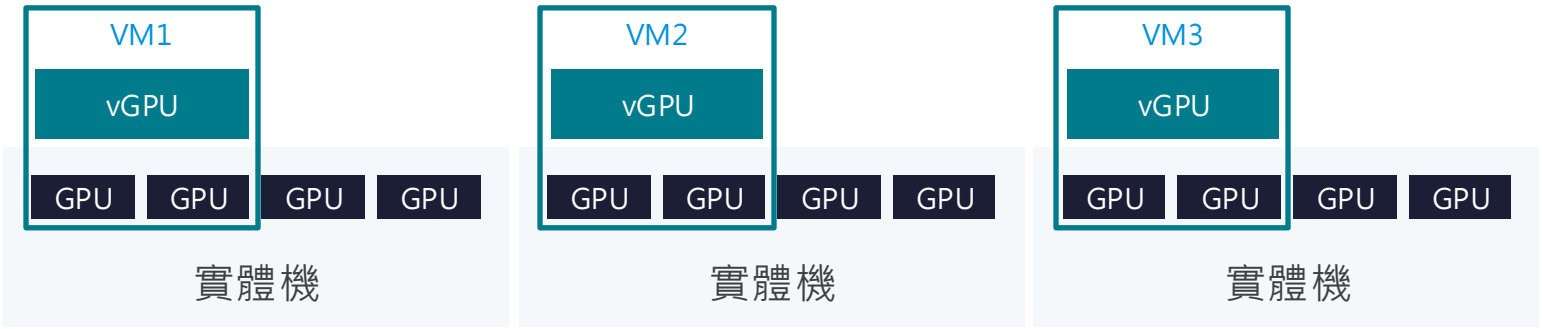
場景： 軟硬體異常，GPU 負載可藉 VM HA 切換運行

- vMotion (動態遷移) 可隨作業需求轉移 GPU負載運行位置
- VM HA (高可用) 可在異常狀況發生時，自動切換到其他機器並重啟
- DRS (分散資源配置) 可平均配置 GPU 工作負載，提高整體效能及使用率
- 當 GPU 配置碎片化時，可動態調度，優化資源配置



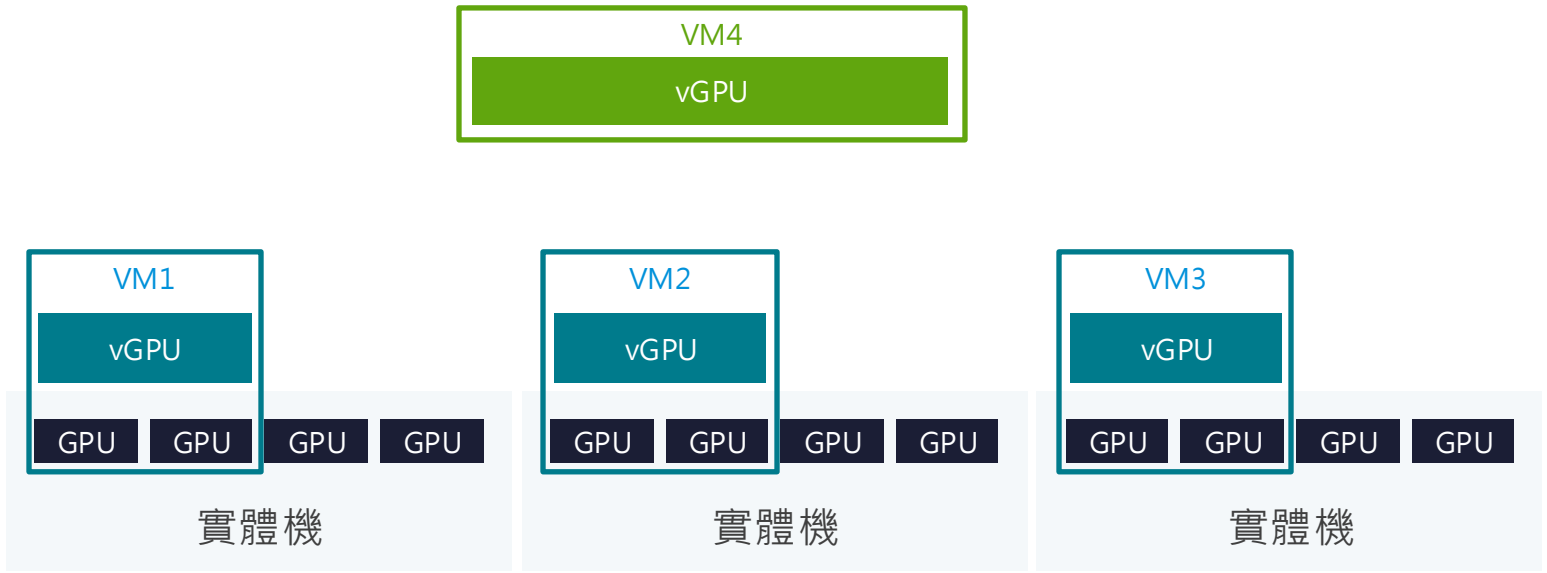
GPU 負載可藉 DRS 最適調配，提高性能與資源利用率

- vMotion (動態遷移) 可隨作業需求轉移 GPU負載運行位置
- VM HA (高可用) 可在異常狀況發生時，自動切換到其他機器並重啟
- DRS (分散資源配置) 可平均配置 GPU 工作負載，提高整體效能及使用率
- 當 GPU 配置碎片化時，可動態調度，優化資源配置



GPU 資源配置碎片化時，可動態調度滿足運算需要

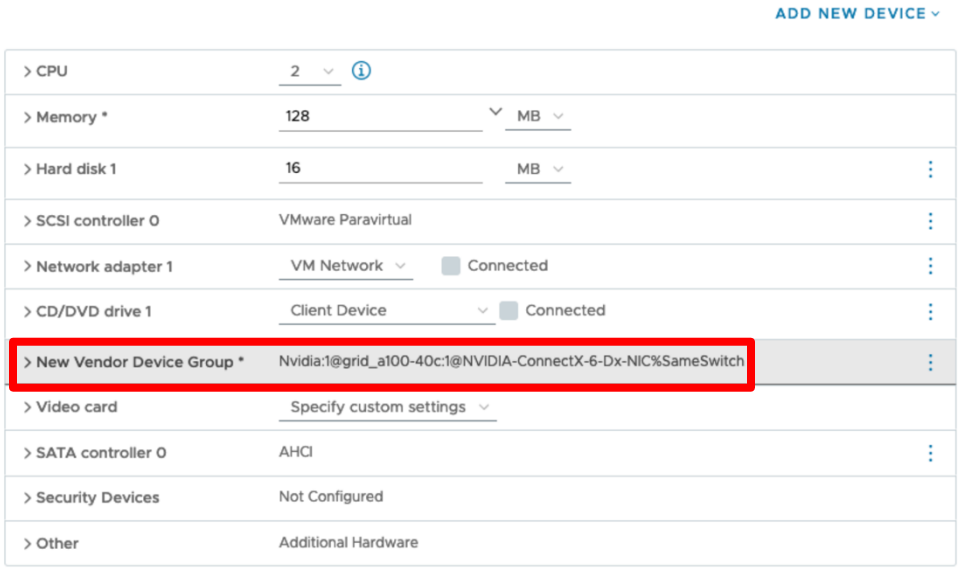
- vMotion (動態遷移) 可隨作業需求轉移 GPU負載運行位置
- VM HA (高可用) 可在異常狀況發生時，自動切換到其他機器並重啟
- DRS (分散資源配置) 可平均配置 GPU 工作負載，提高整體效能及使用率
- 當 GPU 配置碎片化時，可動態調度，優化資源配置

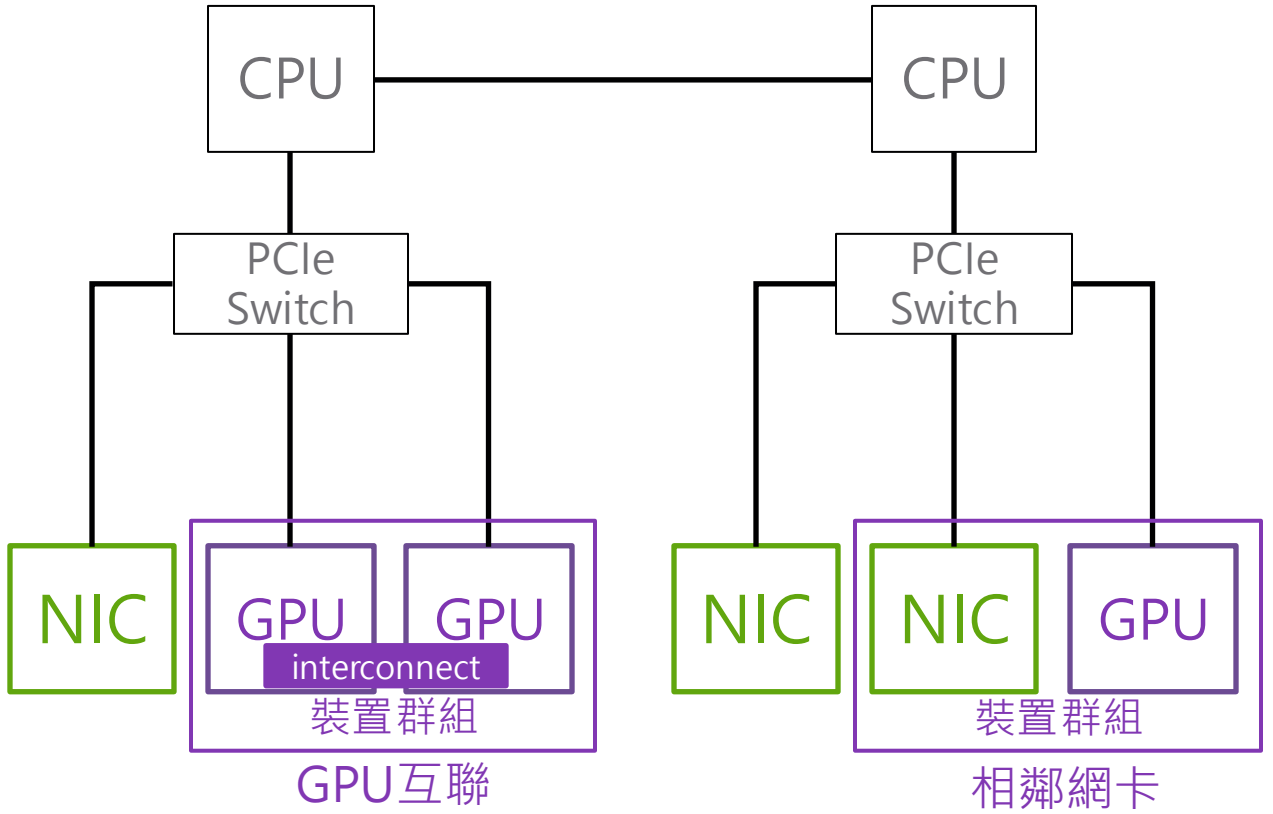


可配置“ 裝置群組” 加速 AI 運算

vGPU 裝置群組可識別配置特殊優化硬體，提升GPU與GPU間，及網卡及GPU間的傳輸性能

- 將鄰近或硬體互聯的多個GPU或網卡，視作一個單元提供給AI/ML 工作負載
- 以裝置群組為單位在vMotion, vSphere DRS 與 vSphere HA 的場景下最佳化調度





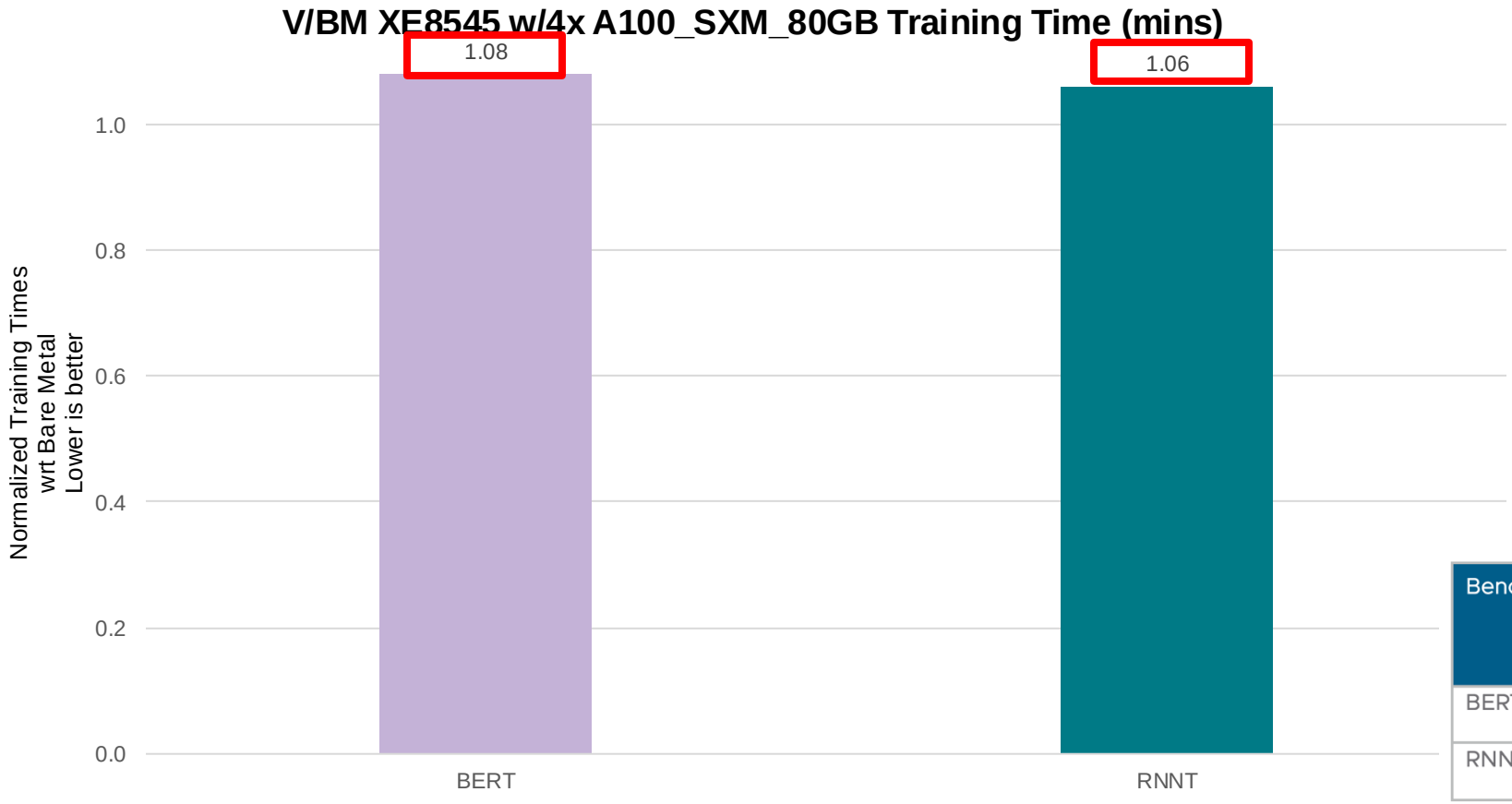
AMD 平台實體機 A100-SXM 80GB 與虛擬化 A100-SXM 80C 性能比較

對照組 (A) AI 實體機配置	對照組 (B) AI 虛擬化配置
Dell PowerEdge XE8545	Dell PowerEdge XE8545
2x AMD EPYC 7543	2x AMD EPYC 7543
128 cores with hyperthreading	推論場景 • 16 allocated to the VM for inference (112 available for other VMs/workloads) 訓練場景 • 88 allocated to the VM for training (40 available for other VMs/workloads)
1TB memory	推論場景 128GB (for inference VM) 訓練場景 900GB (for training VM)
4x A100_SXM_80GB with NVLinks	4x A100_SXM_80C with NVLinks
3 TB NVME Disk	3 TB NVME Disk
Ubuntu 20.04	Ubuntu 20.04
CUDA 12	CUDA 12
—	VMware NVIDIA VIB: 525.85.07

在本次兩種訓練及推論測試場景下，VM環境都只使用實體機的一部分資源，保留剩餘資源可做其他運用

https://blogs.vmware.com/performance/2023/07/vsphere8-performance-ai-ml.html

MLPerf 訓練場景：實體機 A100-SXM 80GB 與虛擬化 A100-SXM 80C 性能比較



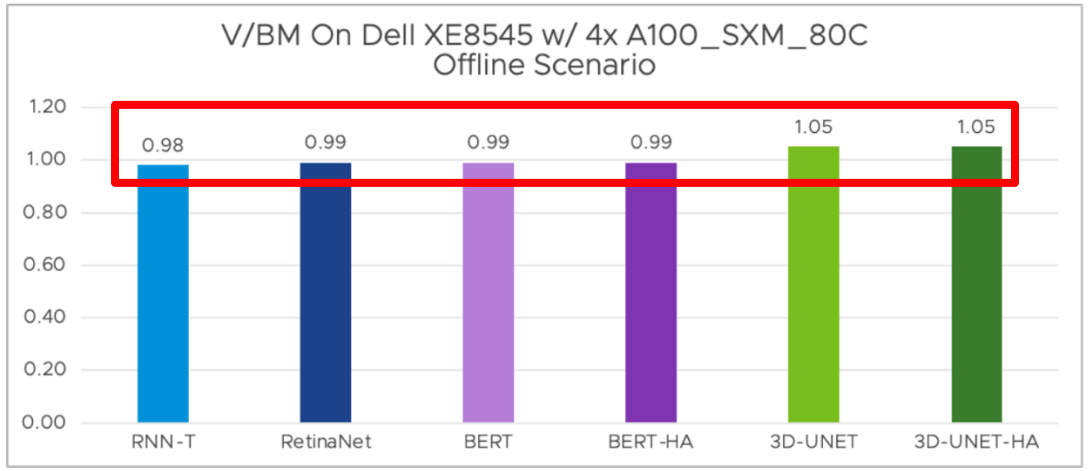
Benchmark	Bare metal 4x A100 training times (mins)	vGPU 4x A100-80c training times (mins)	vGPU/BM
BERT-large	32.792	35.28	1.08
RNNT	55.086	58.447	1.06

虛擬環境只使用實體機的一部分資源，就可提供接近的效能（加上所有 VCF 雲平台優勢）

https://blogs.vmware.com/performance/2023/07/vsphere8-performance-ai-ml.html

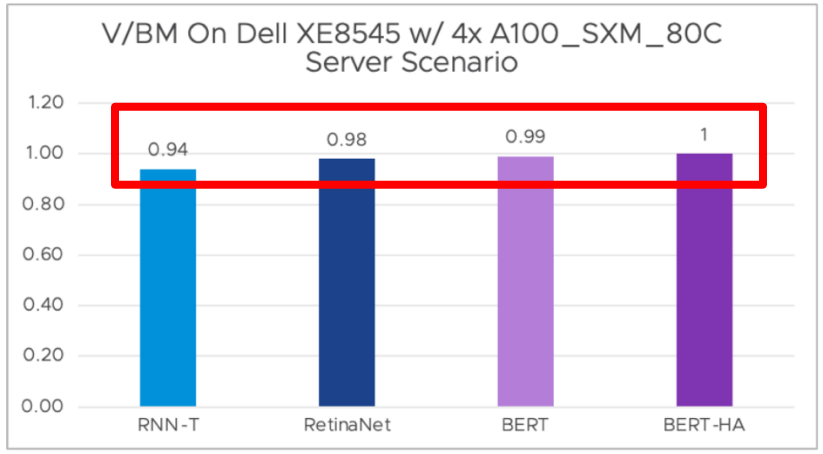
MLPerf 推論場景：實體機 A100-SXM 80GB 與虛擬化 A100-SXM 80C 性能比較

Figure 6. Normalized throughput for OFFLINE scenario (qps): vGPU 4x A100-80c vs bare metal 4x A100-80GB



Benchmark	Bare Metal 4x A100	vGPU 4x A100-80c	vGPU/BM
RNN-T Offline	57084.00	56174.00	0.98
RetinaNet Offline	2910.78	2876.56	0.99
BERT Offline	15090.00	14923.10	0.99
BERT HighA Offline	7880.00	7767.84	0.99
3d-UNET-99 Offline	14.44	15.10	1.05
3d-UNET-99.9 HA Offline	14.44	15.10	1.05

Figure 7. Normalized throughput for SERVER scenario (qps): vGPU 4x A100-80c vs bare metal 4x A100

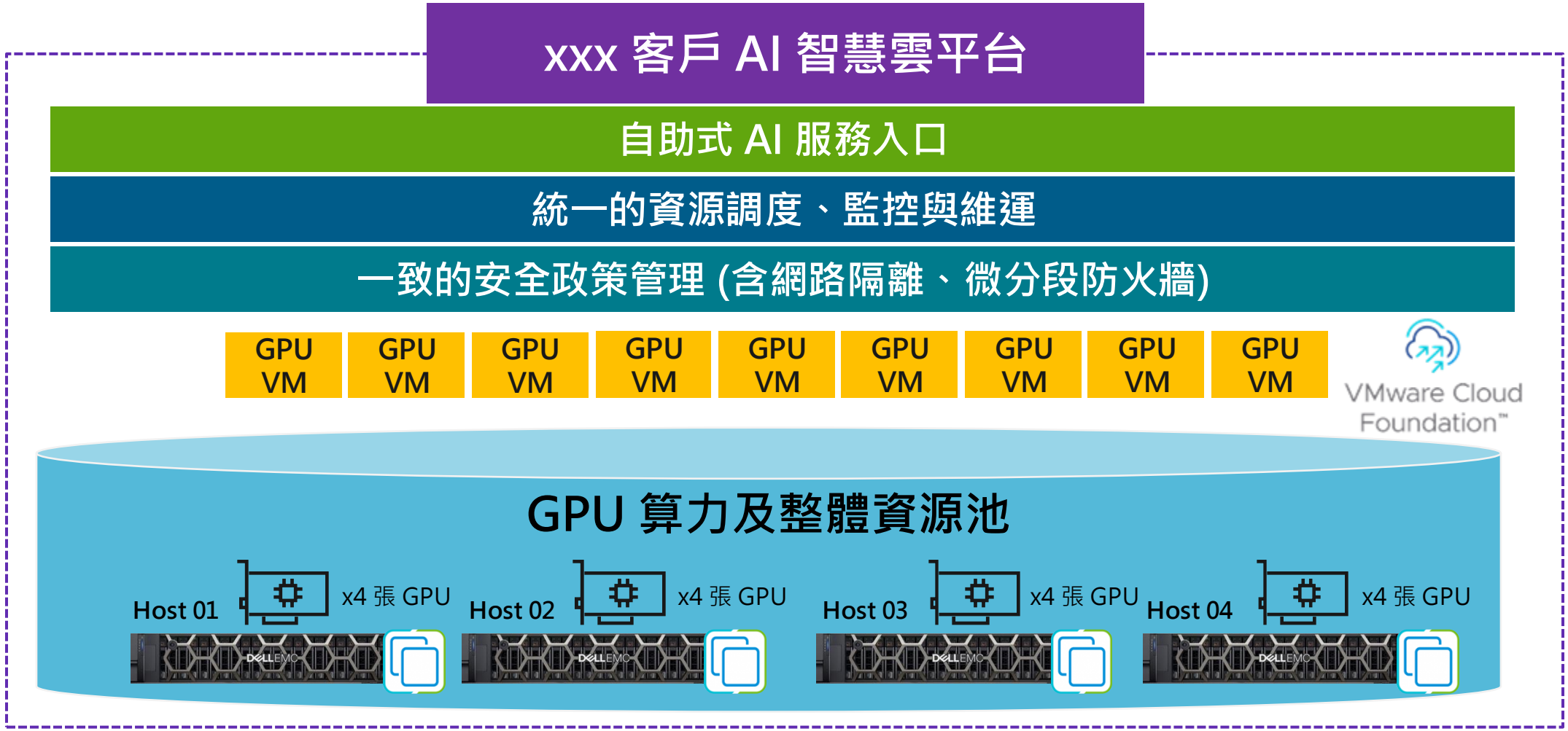


Benchmark	Bare Metal 4x A100	vGPU 4x A100-80c	vGPU/BM
RNN-T Server	54000.40	51001.80	0.94
RetinaNet Server	2848.84	2798.93	0.98
Bert Server	13597.00	13497.90	0.99
Bert HighA Server	7004.00	7004.02	1.00

虛擬環境只使用實體機的一部分資源，就可提供接近的效能（加上所有 VCF 雲平台優勢）

https://blogs.vmware.com/performance/2023/07/vsphere8-performance-ai-ml.html

某財星全球500大企業實現 Private AI 的案例 – 建置



實現 Private AI – 依照企業業務需求，提供 AI 資源

自動化、自助化、標準化 GPU 虛擬機交付

多租戶/團隊
自助選取GPU算力藍圖

A事業群
產線推論 AI 團隊

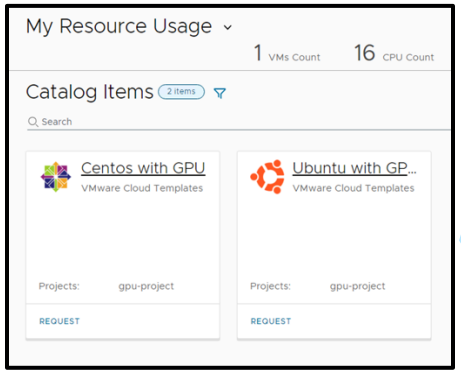
B 事業群
產線推論 AI 團隊

研究中心
訓練模型 AI 團隊

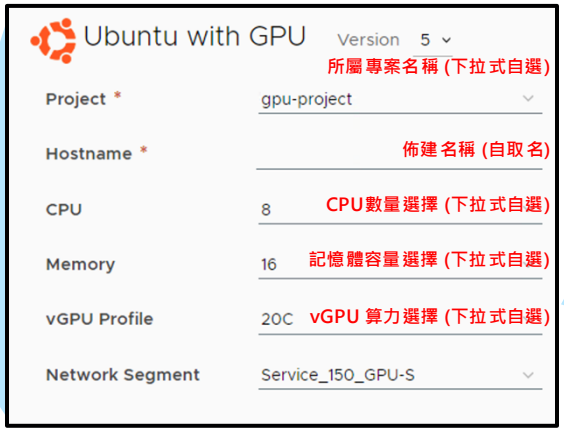
使用單位

目錄式選單

GPU 虛擬機
自動供裝藍圖

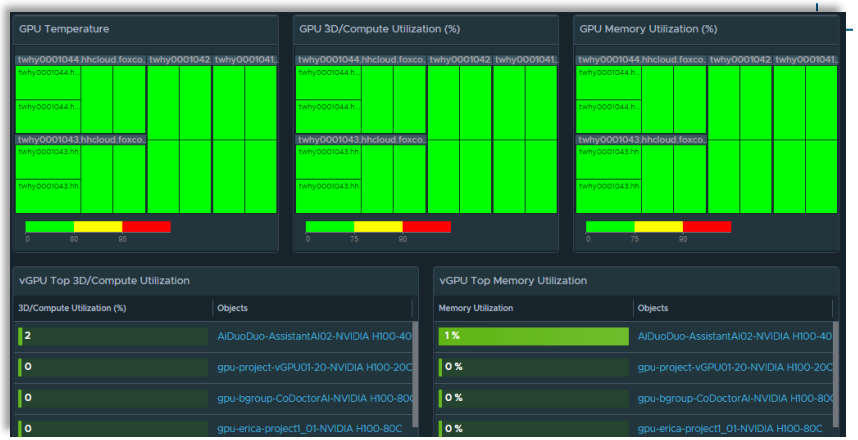


細部配置



GPU 算力選擇

20C	0.25張算力
40C	0.5張算力
80C	1張算力



GPU 監控管理

GPU共用 (單張/ 多張 / 幾分之幾張)

Ubuntu VM

Ubuntu VM

Ubuntu VM

Ubuntu VM

GPU

GPU

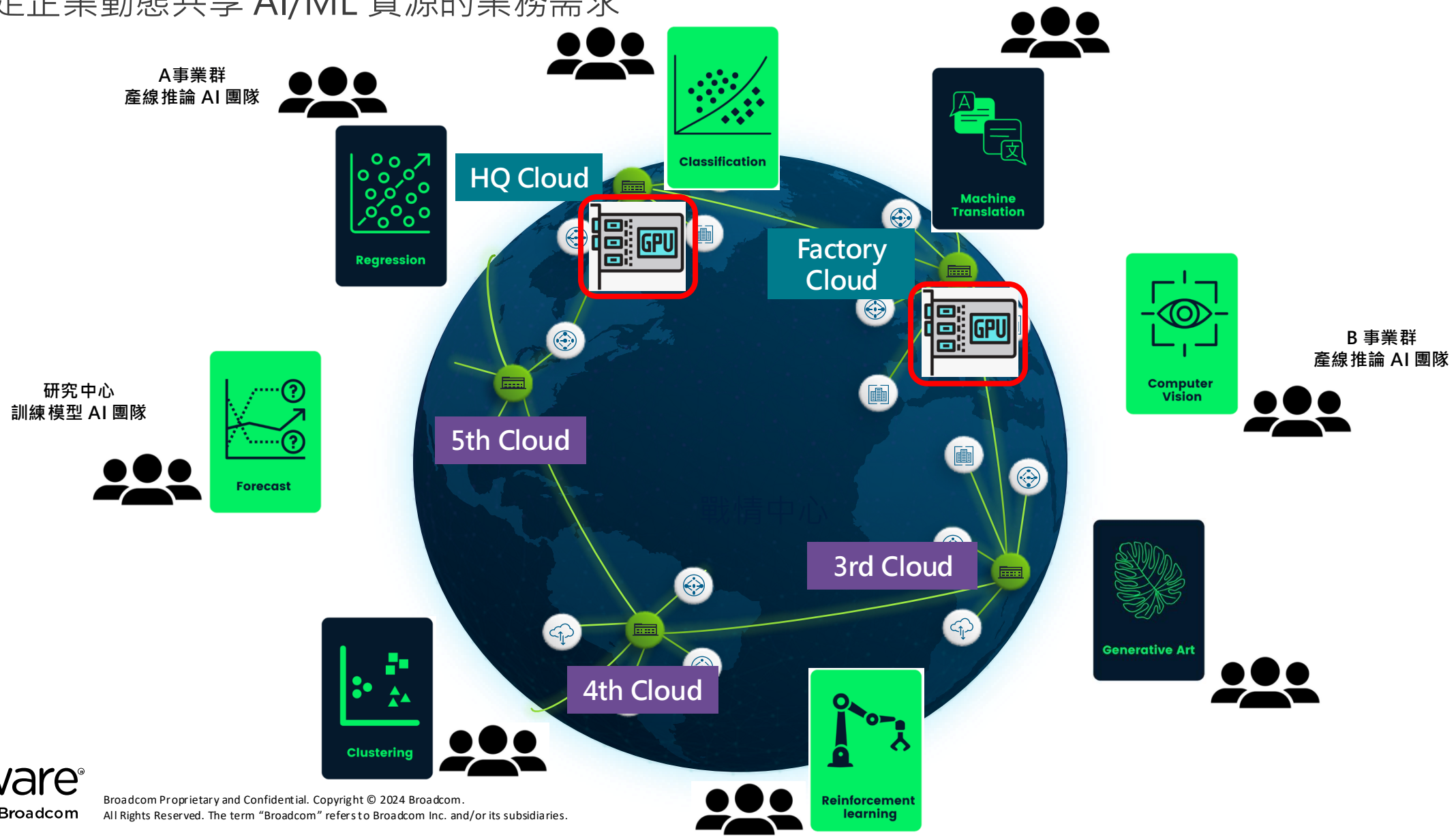
VMware Cloud Foundation (VCF)

AI & GPU 算力資源池

vmware by Broadcom

AI 資源跨區域、跨業務、跨組織，統一使用、監控、調度及管理

滿足企業動態共享 AI/ML 資源的業務需求



Thank You