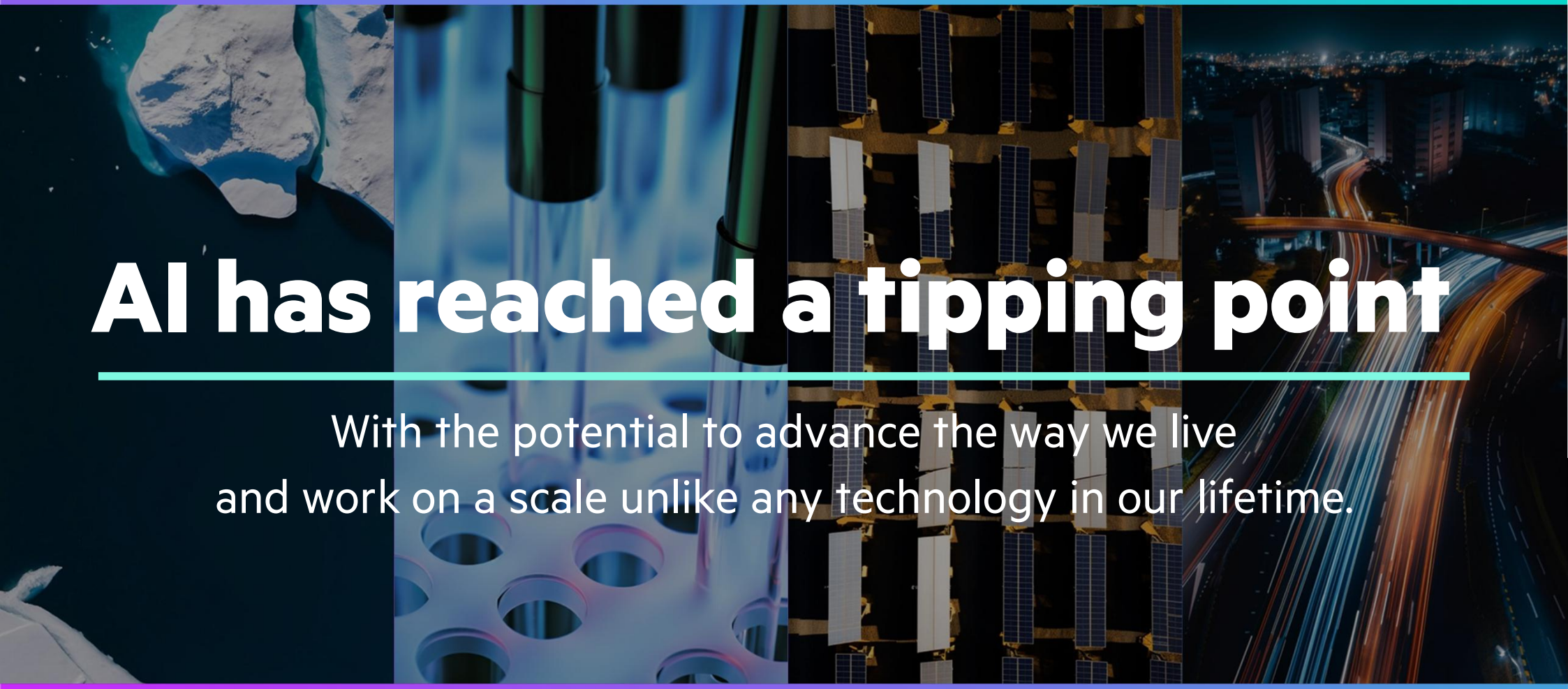

Hewlett Packard
Enterprise
加速AI運算革命
HPE 與 AI Labs 的創新合作
黃光平
HPE Compute Specialist
2025-04



AI has reached a tipping point

With the potential to advance the way we live
and work on a scale unlike any technology in our lifetime.

HPE Compute for AI inference, fine tuning, and training

part of NVIDIA AI Computing by HPE

Scientific Discovery | Personalization | Content Creation | Fraud Detection | Incident Management | Virtual Assistants

Training	Fine tuning	Inference
Build new AI models	Customize existing models	Deploy models
hours, days, weeks	minutes, hours, days	milliseconds, microseconds, seconds

HPE ProLiant XD

HPE ProLiant DL


Future vision of
AI powered business
Innovation
with Digital Twin
Cybersecurity
with Fraud Detection
Quality control
with Vision AI
Customer service
with Conversational AI
Employee productivity
with Generative AI
Personalization
with eCommerce AI

Top AI use cases for transforming business efficiency and driving growth

各行業的主要應用案例

客戶服務自動化	預測性維護	供應鏈優化	網路安全
個人化行銷	詐欺檢測與預防	財務規劃與分析	數據驅動決策

產業特定應用案例

金融服務	醫療保健	製造業	零售業
詐欺檢測與預防 自動化交易 信用評分/風險評估 監理合規與報告	個人化治療方案 藥物研發 醫學影像分析 虛擬健康助理 預測疾病爆發	品質管制 機器人流程自動化 數位孿生 調度和勞動力分配 庫存最佳化	店內分析 需求預測 客戶情緒分析 自動結帳系統 個人化體驗

HPC and AI systems portfolio

Leadership-class supercomputing

100% fanless direct liquid cooling

The next frontier of supercomputing systems redesigned for HPC, AI, and converged workloads

HPE Cray Supercomputing EX4000

HPE Cray Supercomputing EX2500

HPE Slingshot combines the performance of a supercomputing interconnect with the cost-effectiveness of Ethernet

Accelerated AI

70% direct liquid cooling, liquid to air cooling

Purpose-built for AI model training, tuning and inference

8-ways NVLink GPU servers

HPE ProLiant Compute XD680

HPE ProLiant Compute XD685

HPE Cray XD670

HPE Cray XD675

HPE ProLiant Compute accelerating AI applications for Enterprises

4-way & 2-way NVLink GPU servers

HPE ProLiant Compute DL380a Gen12

HPE ProLiant Compute DL384 Gen12

2-way NVLink & 1-way GPU servers

HPE ProLiant DL380a Gen11 / DL385 Gen11

HPE ProLiant Compute DL320 Gen12

Mainstream HPC/AI

Density-optimized, scale-out compute for HPC and AI workloads

HPE Cray XD2000

HPE Cray XD665

Purpose-built storage

Unprecedented data storage price/performance for HPC, AI, and converged workloads

HPE Cray Supercomputing Storage Systems E2000

Cray ClusterStor E1000 Storage Systems

HPE Cray Storage Systems C500

HPE AI Compute **Each tailored to different AI journey maturity**

Core — **AI Applications** — **Edge**

Training	**Fine tuning**	**Inference**
Build new AI models	Customize existing models	Deploy models
8-way GPU servers	***4-way & 2-way GPU servers***	***2-way & 1-way GPU servers***
	 New	
HPE Cray XD	**HPE ProLiant Compute DL380a Gen12**	**HPE ProLiant DL380a Gen11 / DL385 Gen11**
 New	 New	
HPE ProLiant Compute XD	**HPE ProLiant Compute DL384 Gen12**	**HPE ProLiant Compute DL320 Gen12**

HPE ProLiant XD — **HPE ProLiant DL**

AI Builder — **AI Consumer**

Portfolio of purpose-built servers for AI model training, tuning and inferencing

8-way NVLink GPU servers for accelerated AI

Training

High performance for large AI model training and tuning

For LLM builders and AI service providers

8-way GPU servers

HPE Cray XD670	HPE Cray XD675	HPE ProLiant Compute XD680 New	HPE ProLiant Compute XD685 New
8x NVIDIA H200	**8x** AMD MI300X	**8x** Intel Gaudi 3	**8x** NVIDIA H200 **8x** NVIDIA Blackwell GPUs **8x** AMD MI325X

HPE will be time to market to support the NVIDIA GB200 NVL2, and the new Blackwell, and Rubin platform

Portfolio of purpose-built servers for AI model training, tuning and inferencing

4-way & 2-way GPU servers for accelerated AI

Fine tuning

Ultra-scalable GPU acceleration for enterprise AI

For LLM consumers that need flexibility to scale GenAI workloads

Next-level performance for memory intensive AI

For LLM consumers fine tuning large models or RAG

4-way & 2-way NVLink GPU servers

HPE ProLiant Compute DL380a Gen12 New

Up to **8x** NVIDIA H200 NVL GPUs

HPE ProLiant Compute DL384 Gen12 New

Dual NVIDIA GH200 **NVL2**

HPE will be time to market to support the NVIDIA GB200 NVL2, and the new Blackwell, and Rubin platform

Portfolio of purpose-built servers for AI model training, tuning and inferencing

2-way & 1-way GPU servers for accelerated AI

Inference

Scalable and high performing for AI inferencing

Powering large language models
Speech AI, Fraud detection

Computer Vision AI at the Edge

Purpose-built for AI at the edge
Loss prevention, Smart spaces

2-way NVLink & 1-way GPU servers

HPE ProLiant DL380a Gen11 / DL385 Gen11

Up to **4x** NVIDIA H100 NVL GPUs

HPE ProLiant Compute DL320 Gen12

Up to **2x** NVIDIA L40S GPUs
Up to **4x** NVIDIA L4 GPUs

FedGPT

標準版 STD
文本知識應答

- 文本/ 文字表格理解
- 預訓練產業專家模型

專業版 Pro+
知識推論整合

- 產業知識整合推論應答
- 預訓練產業專家模型
- 訓練與優化特定任務模型

企業版 Enterprise
處理企業複雜問題

- 處理企業複雜問題
- 懂影音的產業專家模型
- 人臉識別、物件辨識
- 時事追蹤
- 訓練與優化特定任務模型

Key Takeaways

HPE提供完整的AI應用產品線 - 選擇合適的GPU Compute Platform

Training	Fine-tuning	Inferencing
8-Way GPU server	4-Way & 2-way GPU server	2-Way & 1-way GPU server
Cray XD	***DL380a Gen12***	***DL380a Gen11***
- 8x H200/ MI300X	- Up to 8x H100 NVL/ H200 NVL/ L40S	- Up to 4x H100 NVL/ L40s
Compute XD	***DL384 Gen12***	***DL320 Gen12***
- 8x H200/ MI325X/ Gaudi 3	- Dual GH200/ GB200	- Up to 2x L40S/ 4x L4

Thank you