

AI資料中心市場發展與關鍵議題

陳牧風

產業分析師

產業情報研究所

財團法人資訊工業策進會

2024.04.18





資料中心依設置地點可分為雲端本地與邊緣

以國際認證/ 設備完善 程度區分	Tier 1	Tier 2	Tier 3	Tier 4
	系統有足夠的功 能可以維持資料 中心運作	有備用基礎設備 、零件可以更換	可在不關閉IT設備 的情形下，維護 資料中心的能力	自動容錯機制， 發生故障時能夠 自主修復
以設置地點 與規模區分	企業資料中心 系統有足夠的功 能可以維持資料 中心運作		雲端資料中心 (超大規模資料中心) 資料是由雲端服 務供應商託管 可以容納成千上 萬個資料中心伺 服器，並可擴展	邊緣資料中心 較小且建立在最 終用戶的邊緣， 能減少延遲，增 加容量並改善連 接性
	託管資料中心 可以分成零售與 批發：零售僅提 供基礎設備出租 批發則出租整個 資料中心單元			
	本地資料中心 (On-premise) 傳統IT組織和新興雲端原生公司越來越 多將託管資料中心視為新的本地設施			

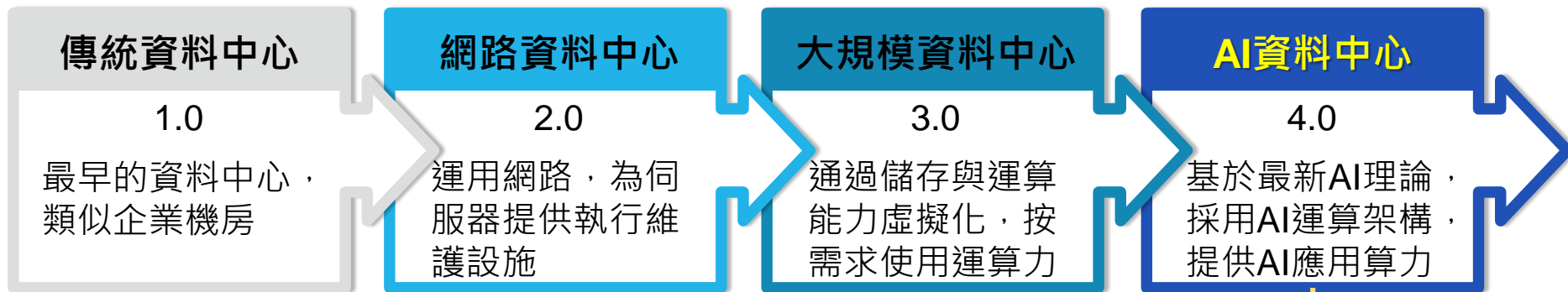
資料來源：各公司，MIC整理，2024年4月

- 資料中心若以設置地點與規模為標準，可分為企業、託管、雲端與邊緣資料中心
- 企業資料中心與託管資料中心存放企業數據，兩者均開始同樣被視為本地資料中心

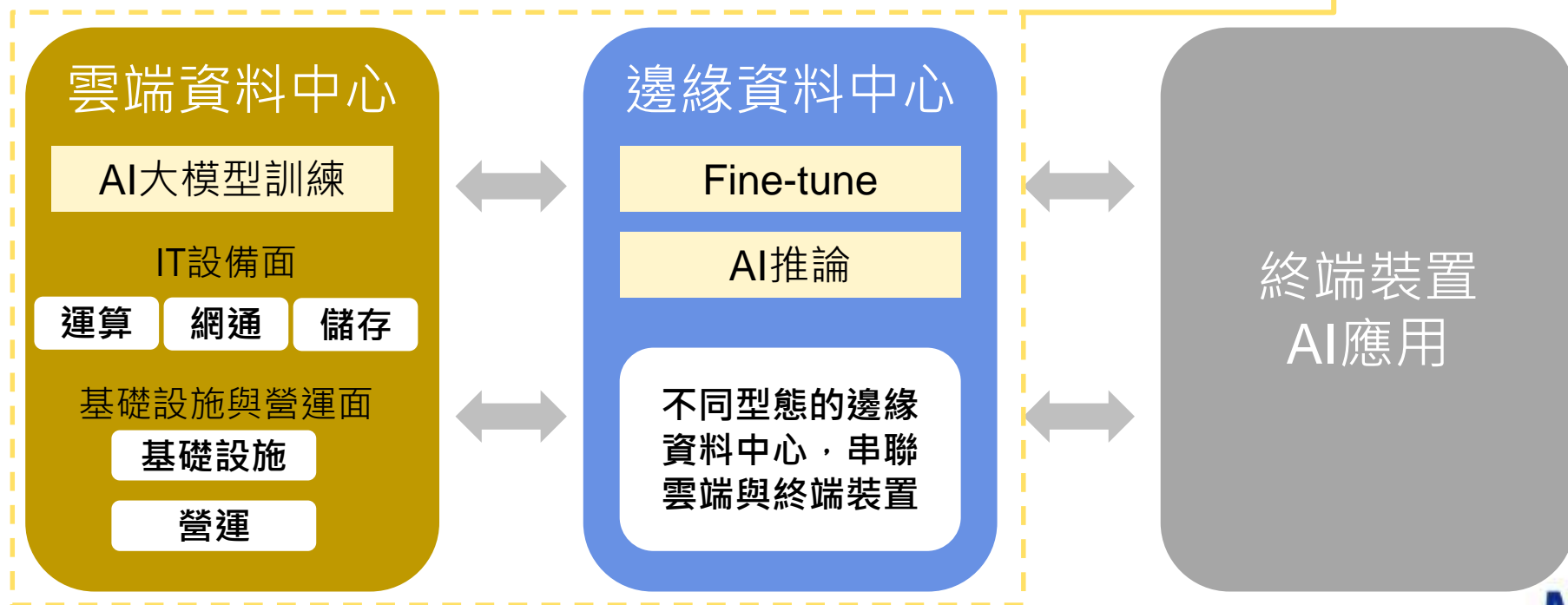


當前AI已同步導入雲端與邊緣資料中心

資料中心的演進過程



AI應用從雲端至終端的傳輸方式

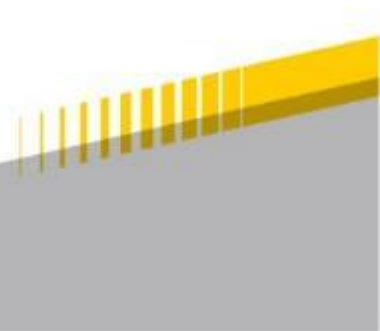


簡報大綱

- 全球資料中心發展現況與趨勢
- AI導入雲端資料中心關鍵議題
- AI導入邊緣資料中心關鍵議題
- 台灣產業發展機會
- 結論



全球資料中心發展現況與趨勢





2024年全球伺服器市場藉由AI伺服器重新回溫

影響全球伺服器市場因素

雲端服務商持續擴增東南亞、南亞的資料中心

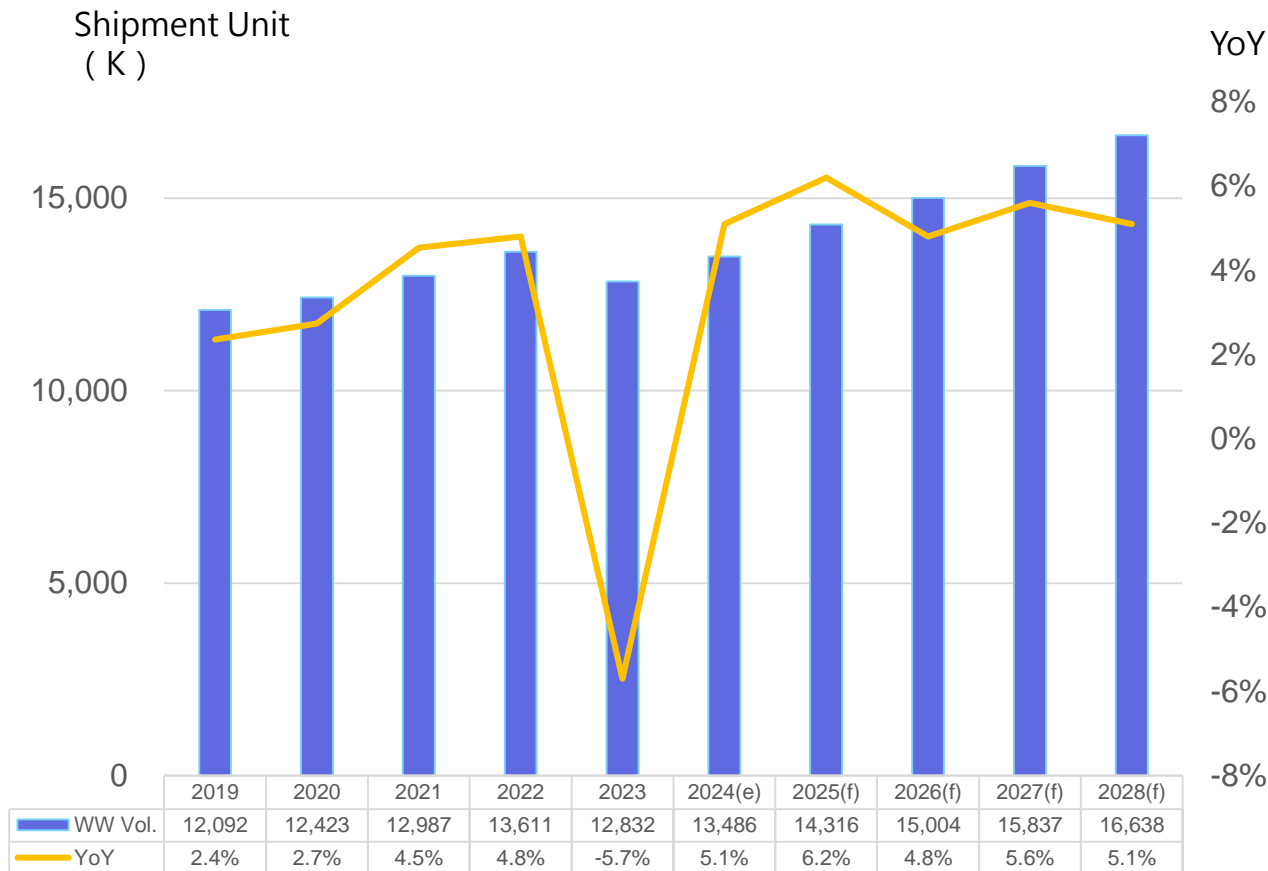
雲端服務商AI伺服器需求仍在延續

伺服器品牌商新AI伺服器產品量產

AI推論與中小模型微調需求

同樣資本支出下，一般伺服器採購率相對下滑

2019~2028年全球伺服器市場預測



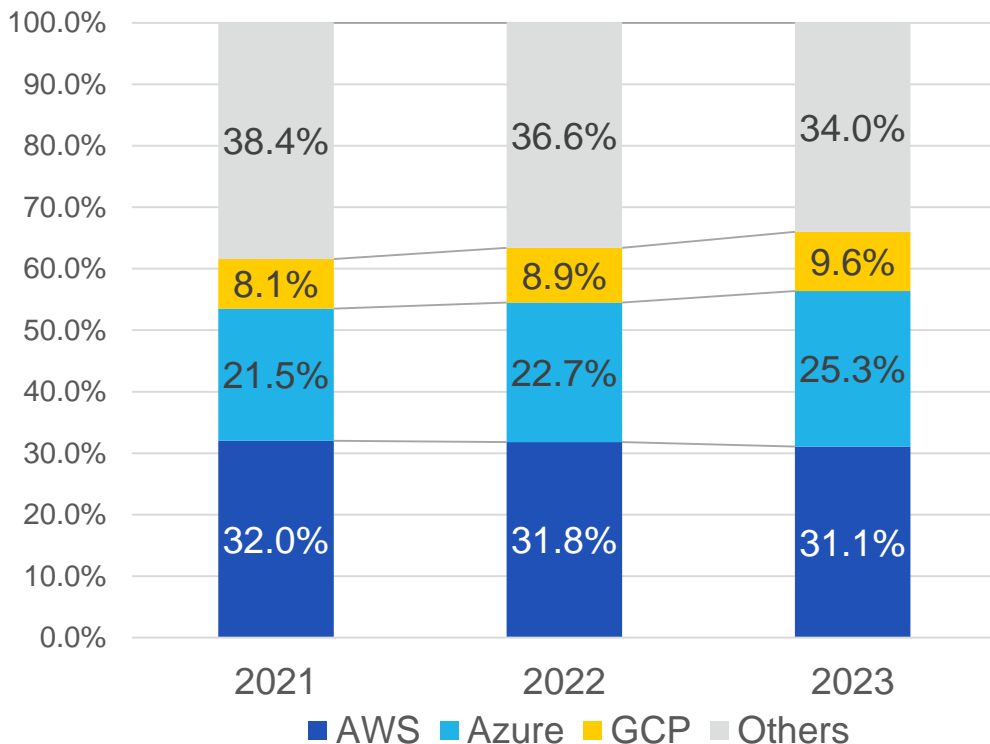
資料來源：MIC，2024年4月

- 全球伺服器市場2024年的主要驅動力為AI伺服器，不論雲端服務商、伺服器品牌商訂單均在上升
- 另外AI推論伺服器將因為企業端推論、中小模型微調（Fine-tune）需求而出貨量增加



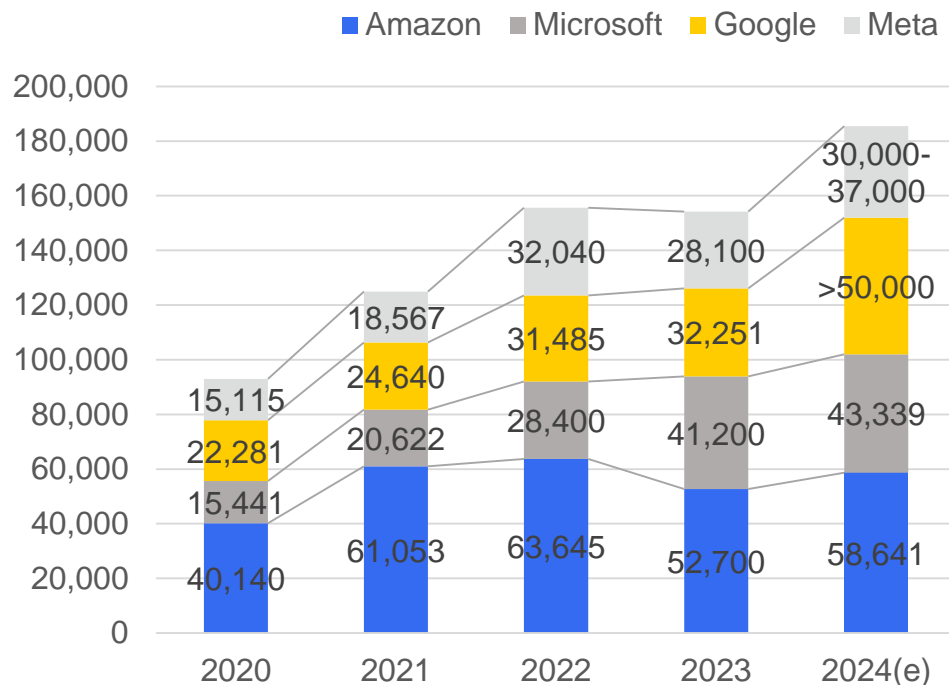
2024年四大雲端服務商將加大資本支出

2021年~2023年雲端服務商市占率



雲端資料中心前四大建造者資本支出

單位：百萬美金



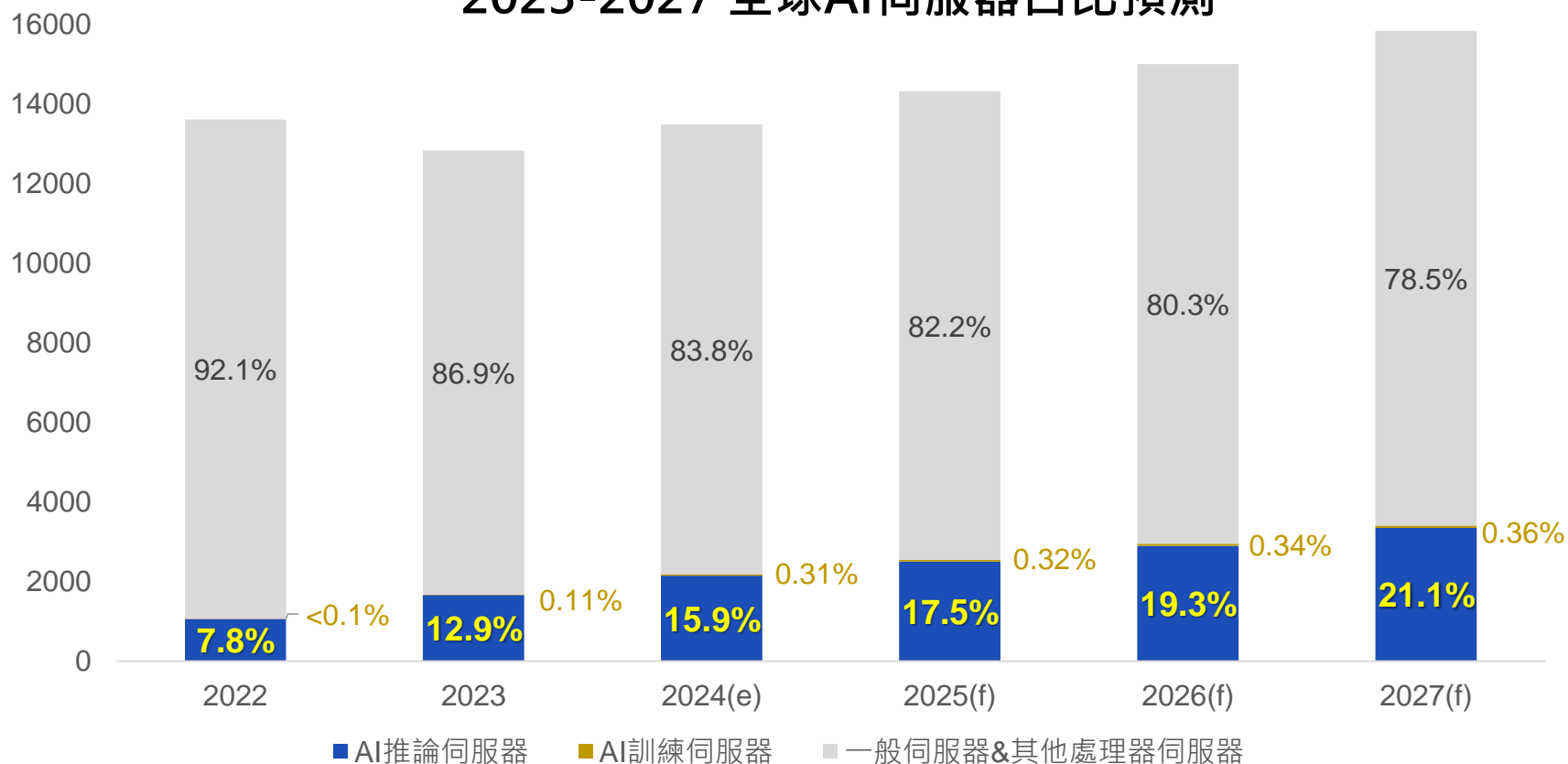
資料來源：各公司，MIC整理，2024年4月

- 2023年因生成式AI、大型語言模型需求，使Azure在市占率大幅提升，Google亦有小量成長
- 2024年預期四大CSP資本支出將進一步提升，主要支出於進行資料中心AI基礎建設



2024年全球AI伺服器出貨量與出貨占比遽增

2023-2027 全球AI伺服器占比預測

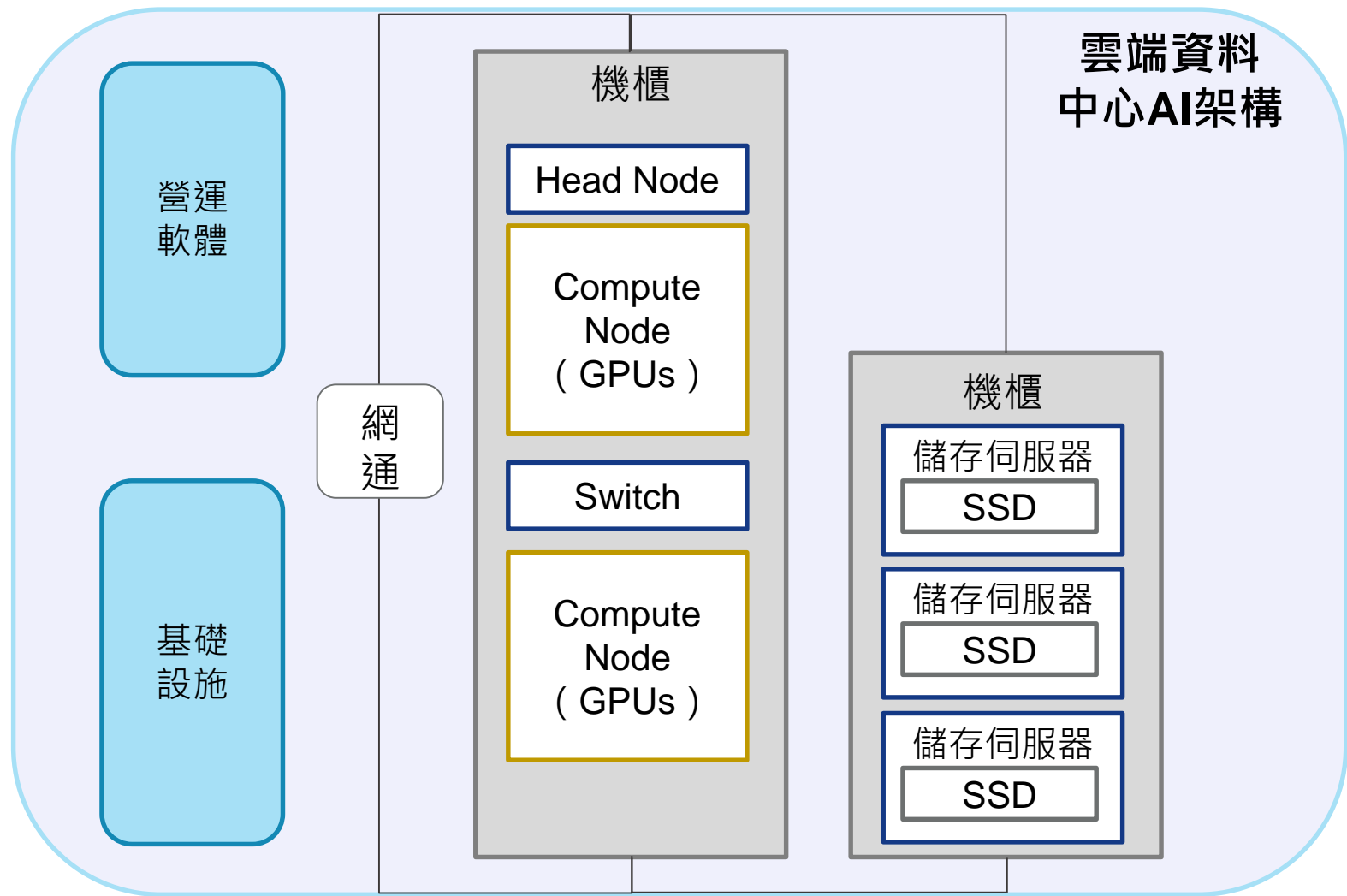


資料來源：MIC，2024年4月

- NVIDIA H100伺服器訂單，H200、GH200預計在2024Q2量產，將使AI訓練伺服器出貨進一步提升
- 搭載 AMD MI300x、MI300A、Intel Habana Gaudi 3的伺服器同樣將在2024年放量
- AI推論及中小模型訓練，NVIDIA中低階GPU伺服器、Xilinx FPGA同樣將帶動AI伺服器出貨成長



雲端資料中心架構因AI算力堆疊需要進行調整



- 雲端資料中心當中因為需要大量AI算力的堆疊，整體部署架構已經出現變化
- 當中運算節點、資料傳輸、資料儲存、基礎設施與營運軟體的升級，都是需要關注的重點



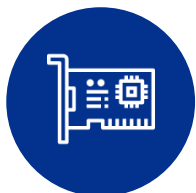
AI全面影響資料中心不同構面的建置

AI對資料中心IT設備的影響



運算

資料中心集成大量GPU、ASIC、FPGA等AI晶片。為符合AI運算的需求，AI伺服器需透過更緊密的方式進行叢集



網通

為符合AI訓練數據快速傳輸需求使用智慧網卡（NIC）、DPU來協助網路傳輸，並透過交換器、複雜的光纖結構來使用乙太網路（Ethernet）或Infiniband



儲存

AI訓練的過程中，除了AI伺服器中記憶體의暫存儲存熱數據外。大量的冷數據需要透過儲存伺服器來存放，增加HDD、SSD的需求

AI對資料中心基礎設施的影響



基礎設施

AI運算需耗用更多電力，因此UPS、配電裝置（PDU）、備援電池（BBU）均需要進行調整

冷卻方面包含氣冷風扇牆、冰水主機，通風管道，以及直接、沉浸式液體冷卻的導入



AI對資料中心營運的影響



營運

使用AI來營運資料中心，可以通過動態設置和自適應功能實現更好的工作負載。並透過機器學習（ML）來最佳化資料中心監測模型，進行預測性分析。

AI導入雲端資料中心關鍵議題

IT設備面：運算、網通與儲存





雲端服務商積極於資料中心堆疊AI算力



運算

自研晶片

超級電腦

GPU

Meta

MTIA (元訓練與推理加速器)：針對推理工作負載內部客製化加速器晶片

兩個各採用**24,576個Nvidia H100 GPU**的大型資料中心叢集

預計採購35萬顆H100 GPU，於2024年底打造60萬顆GPU規模的AI算力

Microsoft

Azure Cobalt：Arm架構的CPU，強化雲端運算

Maia 100：為AI負載設計的AI晶片，可執行大型AI模型

Azure Eagle：包含14,400個H100 GPU，超級電腦TOP 500排名第三

2024預計採購8萬台H100 AI伺服器（64萬顆），並將採購下一代B100 GPU

Google

TPUv5e：每個Pod擁有256個晶片，AI推論較高性價比

TPUv5p：每個Pod有8960個晶片，處理大型AI訓練

A3超級電腦：最多可堆疊26,000個H100 GPU

2023約採購5萬顆H100 GPU，預期2024將同步擴大GPU與TPU的堆疊

Amazon

AWS Inferentia：專注在深度學習與AI推論

AWS Trainium：將重點放在AI訓練與模型建構最佳化

P5 Ultra Scale GPU叢集：可以堆疊20,000個H100 GPU
Project Ceiba：包含16,384顆GH200超級晶片

2024年將採購大量GH200與H200 GPU使用於自身應用實例

Oracle

Oracle因自研晶片需要數年時間才能完成，決定不投入自研晶片，使用NVIDIA、AMD GPU

OCI Super Cluster：可以擴展數萬個NVIDIA H100 GPU

2024年除NVIDIA GPU外，AMD MI300X成重要採購目標

資料來源：各公司，MIC整理，2024年4月




- 雲端服務商積極部署自身的AI算力，透過自研AI加速晶片來符合自身雲端服務的客製化需求
- 此外，和伺服器品牌商、處理器廠商合作打造AI超級電腦，協助自身與客戶進行AI模型訓練



處理器廠商AI加速晶片進入全新的競爭態勢



運算

Vendor	GPU			超級晶片、APU	FPGA、ASIC
	AI訓練	AI推論	特殊規格		
	<ul style="list-style-type: none"> H100 H200 B200 	<ul style="list-style-type: none"> L40s B40 	<ul style="list-style-type: none"> H800 H20 L20 L2 	<ul style="list-style-type: none"> GH200：Arm架構推論 GH200 NVLink：Arm架構訓練與推論 GB200 GB200 NVLink 72 	<ul style="list-style-type: none"> 開發客製化晶片服務
	<ul style="list-style-type: none"> MI300 X 		<ul style="list-style-type: none"> 可能推出MI250降規版 	<ul style="list-style-type: none"> MI300A APU 	<ul style="list-style-type: none"> XILINX FPGA
	<ul style="list-style-type: none"> GPU Max Series (Ponte Vecchio) Habana Gaudi 2、3 		<ul style="list-style-type: none"> Gaudi 2 降規版 		<ul style="list-style-type: none"> Intel Agilex FPGA Intel 客製化 ASIC

資料來源：各公司，MIC整理，2024年4月

- 各處理器廠商為符合客戶各種應用場域的AI運算需求，開始擴增資料中心AI加速晶片的產品規格
- GPU包含AI訓練、AI推論專用及面對中國大陸市場限制的特殊規格
- 將CPU與GPU封裝在一起的超級晶片、APU亦成為新的模式，強調使CPU和GPU之間的互聯更加緊密



資料中心數據傳輸成為AI模型訓練的關鍵



開發AI模型的三個階段

資料準備：蒐集和彙整輸入AI模型資料集



AI訓練：透過大量資料提供給AI模型來訓練其執行特定任務



AI推論：根據新輸入、未見過的數據進行預測或決策

需要大量數據和運算資源來支援其更新過程，AI模型從不斷收集的數據中學習以細化其參數

使用數萬台的GPU伺服器進行叢集

AI資料中心網路必須100%可靠，才能最佳化完成時間、消除延遲導致的速度降低問題

AI資料中心網路的運作模式

網路拓撲架構設計 (fabric)

使用**CLOS架構**，由多個層級的網路組成，每個層級包含一組交換機 (switch)，具有高可靠性與擴展性。根據模型大小和GPU 規模，可以使用不同層級的結構

流量控制和擁塞避免

具有**最佳連接數量**的適當大小的結構互連，以及**檢測和糾正流量不平衡**以避免擁塞和資料包丟失，如明確壅塞通知 (ECN) 與資料中心量化壅塞通知 (DCQCN)

大規模建置與效能

乙太網路成為處理HPC和AI應用的首選開放標準解決方案，包括目前向**800 GbE** 和**資料中心橋接 (DCB)** 的進展

資料來源：Juniper Networks，MIC整理，2024年4月

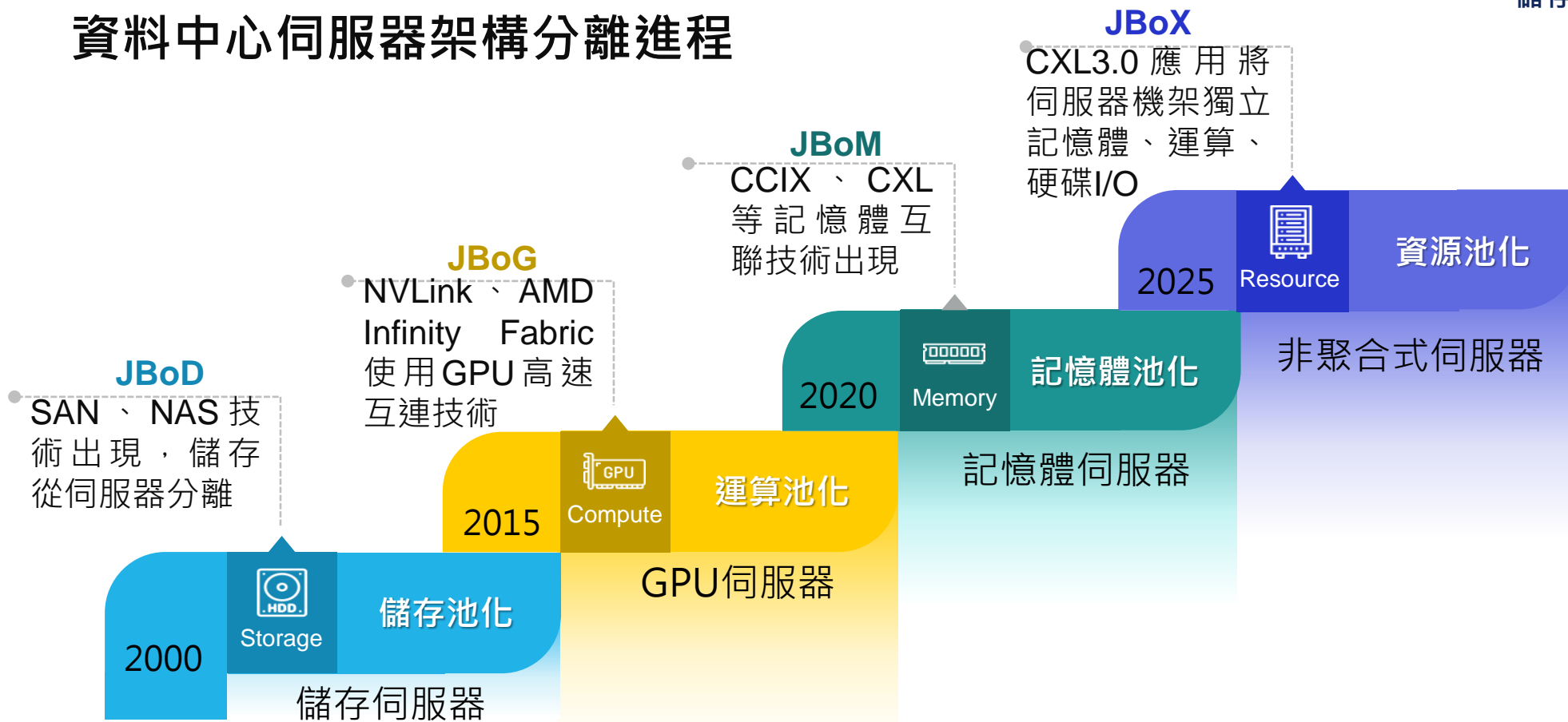
- 要建置完善的資料中心的網路傳輸才能最大限度提高GPU的利用率
- 透過不同的網路架構設計，才能符合AI訓練的高速傳輸需求
- 壅塞管理、最小化延遲與最佳化作業完成時間成為資料中心網路佈建時的關注重點



資料中心正朝「資源池」化發展



資料中心伺服器架構分離進程



資料來源：各公司，MIC整理，2024年4月

- 各關鍵零組件受到AI運算的影響，在運算需求大量提升之下，開始從伺服器中解構出來
- 最終目的是將GPU、FPGA、NIC、DPU等硬體資源徹底資源池化，將資源按需求組成動態虛擬伺服器
- 最早分離硬碟形成儲存伺服器，當前關注記憶體池化，未來於單一或多機櫃內形成非聚合式伺服器

AI導入雲端資料中心關鍵議題

基礎設施與營運面

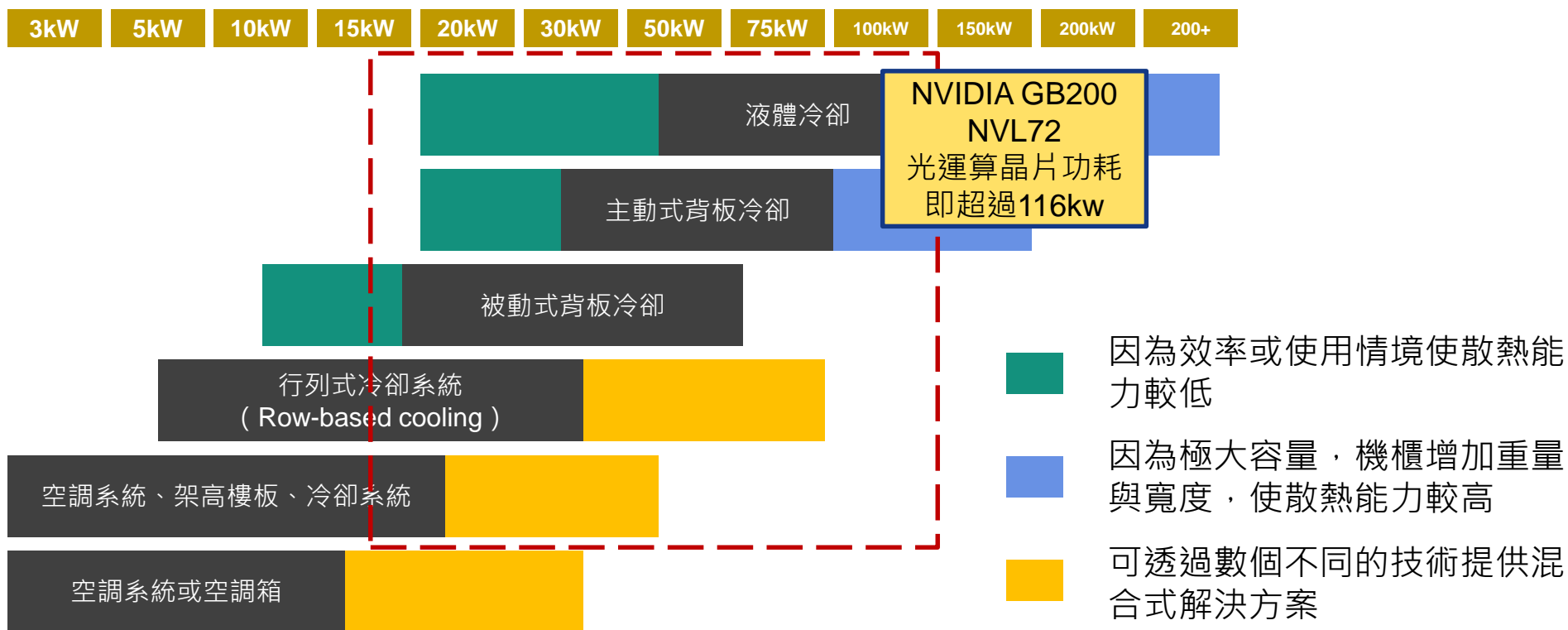




AI資料中心耗能增加使液體冷卻成為必備方案



基礎設施



資料來源：Vertiv，MIC整理，2024年4月

*紅色框為液體冷卻可行選項

- 當機架密度超過20kW 時，基於空氣的冷卻系統就會失去效力，此時液體冷卻就成為可行的方法
- 節能背板冷卻 (RDHx) 是一項成熟的技術，可為管理20 kW 以上的密度提供可行的解決方案
- 此技術不會將液體直接輸送到伺服器內部，但用液體高熱傳導性，採用相似於直接液體冷卻的做法

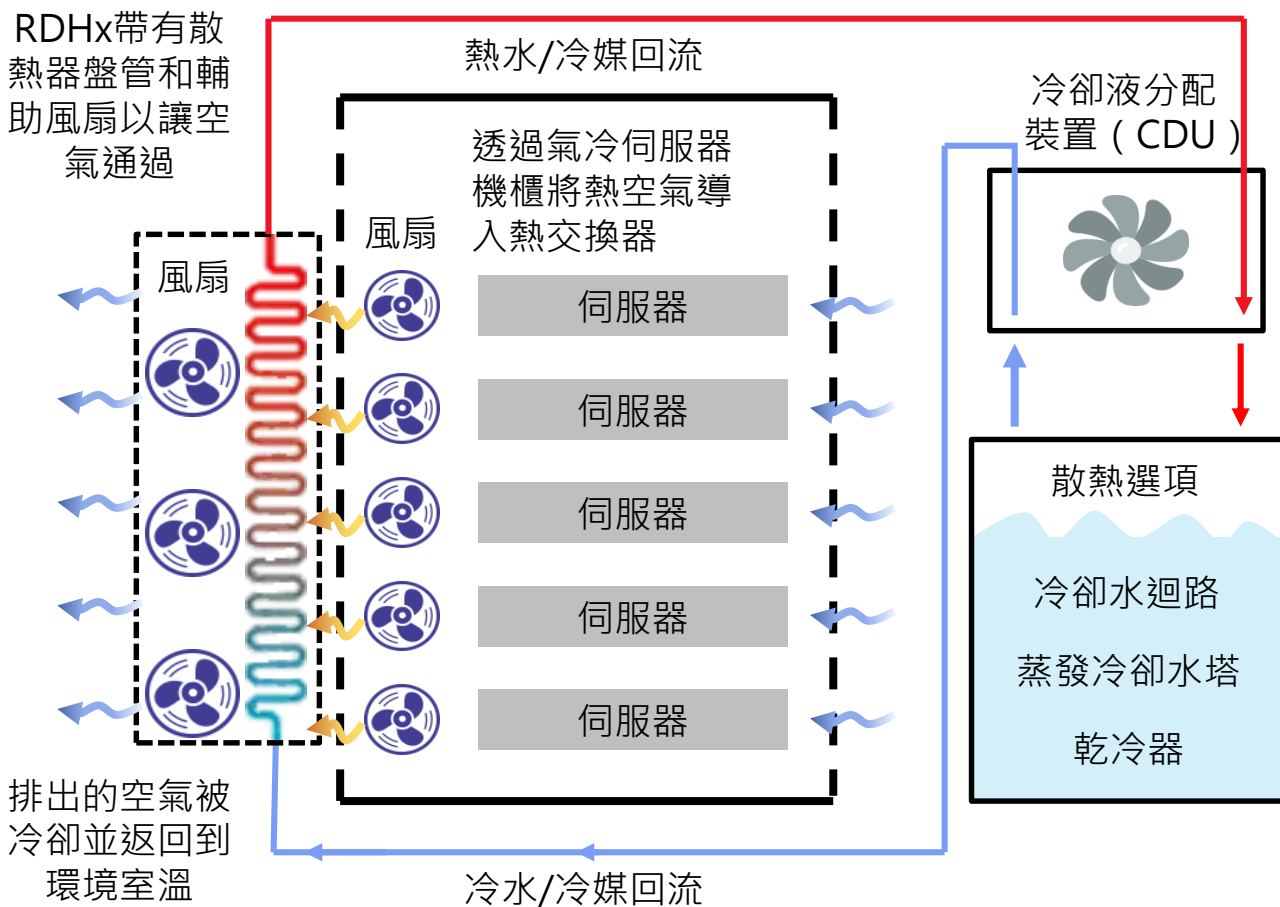


節能背板成為資料中心散熱的過渡產品



基礎設施

節能背板散熱流程圖



節能背板冷卻類型比較

	主動式	被動式
其他冷卻設備	不須任何設備即可提供冷卻	高度依賴機櫃內伺服器風扇提供的氣流
散熱方式	包含風扇，放置在機櫃背部，吸收熱量並透過冷凍水系統散熱	透過安裝在機架背門位置充滿液體的盤管排出熱空氣
問題	須避免伺服器風扇的空氣量大於背板	壓力難以控制需要即時監控

資料來源：Supermicro、AKCP、MIC整理，2024年4月

- 要將資料中心或企業機房轉換為液體冷卻需要考慮成本、建築結構、管線等因素
- 可以直接以現有機櫃進行安裝、導入相對容易，因此節能背板成為許多企業考慮的解決方案



直接式與沉浸式液冷應用場景有所差異



基礎設施

氣冷與直接式液冷成本比較表

	空氣冷卻	直接式液體冷卻 (D2C)
GPU伺服器 (包含 2xCPU, 8x H100 GPU)	7,000瓦特	6,300瓦特 (節省風扇功耗)
機櫃內伺服器數量	8	8
每櫃伺服器功耗	56,000瓦特	50,400瓦特
PUE	1.5	1.1 (減少空調成本)
每櫃所需總功耗	84,000	55,440
一年總耗電量 (度 Kwh)	485,654	735,840
每度價格 (以美國 平均電費為基準)	0.12美金	0.12美金
1年總耗電成本	88,301美金	58,279美金
3年總耗電成本	264,902美金	174,836美金
液冷安裝成本		30,000美金
3年每櫃節省成本		60,067美金

背板、直接式、沉浸式液冷優勢比較

散熱的方法	首要優勢	次要優勢
散熱背板	對既有規格影響最小	可以安裝在既有機櫃
直接式液冷	伺服器規格可適用 範圍較廣	降低風扇速度、噪音
沉浸式液冷	最具有散熱效率	最小化風扇速度、 靜音

資料來源：Supermicro、GRC、MIC整理，2024年4月

- 直接式液冷可運用機櫃方式提供，可由中小型資料中心、企業級資料中心導入，亦可導入大型資料中心
- 沉浸式液冷則適用於雲端資料中心，或單櫃使用於邊緣資料中心

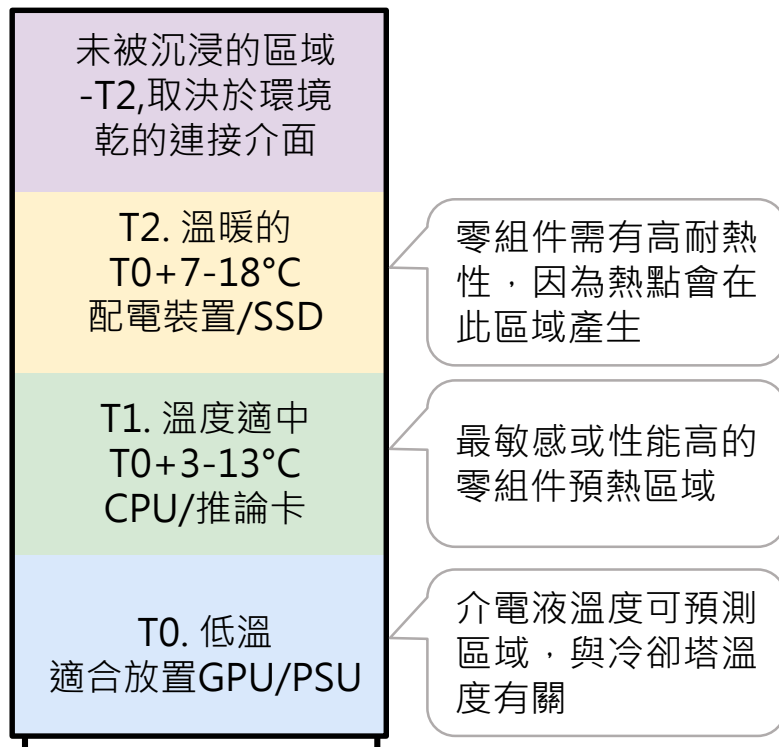


沉浸式冷卻關注槽內零組件配置與冷卻液



基礎設施

沉浸式液冷槽的溫度區間



沉浸式冷卻液體比較表

液體	導熱係數 (W/mK)	比熱 (J/kgK)	黏度 (cP)	密度 (kg/m3)	成本	沸點 (°C)
水	0.580	4186	1.00	1000	\$	100
去離子水	0.606	4200	1.00	997	\$+	100
50-50 水/乙二醇	0.402	3283	2.51	1082	\$\$	107
50-50 水/丙二醇	0.357	3559	5.20	1041	\$\$	106
甲酸鉀 Dynalene HC30	0.519	3100	3.70	1275	\$\$\$	112
全氟聚醚 (PFPE) Galden HT200	0.065	963	4.30	1790	\$\$\$	200
電子氟化液 Fluorinert FC72	0.057	1100	0.64	1680	\$\$\$	56
電器絕緣油 Shall Diala S4	0.142	2150	7.57	805	\$\$\$	>280
礦物油	0.136	1700 - 2100	10- 1000+	870	\$\$\$	218 -643

資料來源：OCP、Laird Thermal Systems、MIC整理，2024年4月

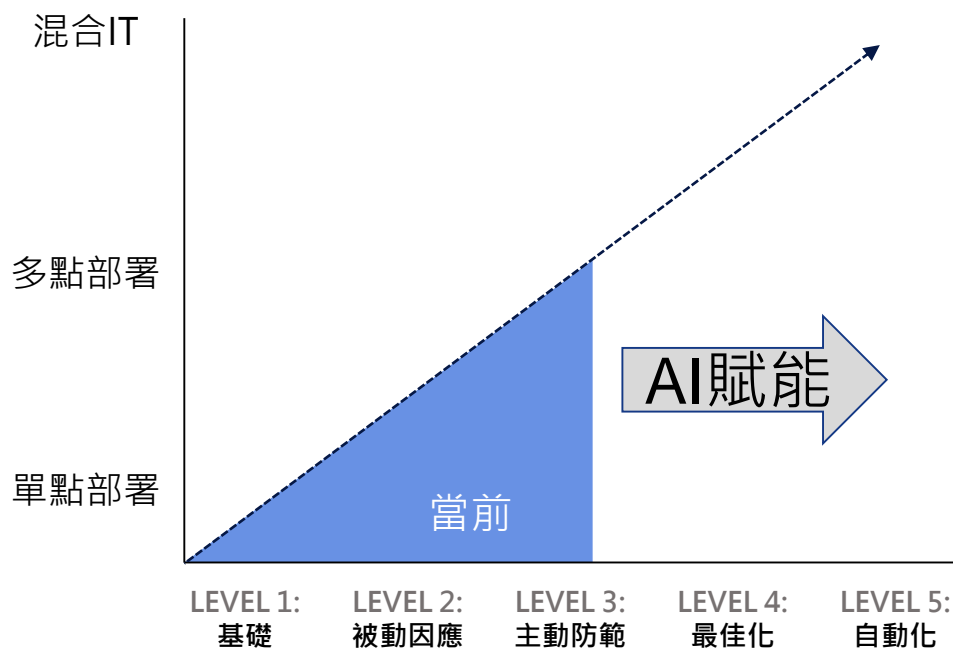
- 沉浸式冷卻當前眾所矚目的痛點，包含液冷槽內零組件的擺放與冷卻液的使用
- 冷卻液當中的電子氟化液符合雙相式液冷的需求，然而3M預計於2025年全面停產
- 當前其他化學品廠商開始切入此市場，研發氟化液或是其他可以替代的礦物油產品



AI資料中心營運管理由主動防範朝自動化發展



AIDCOPS成熟度模型

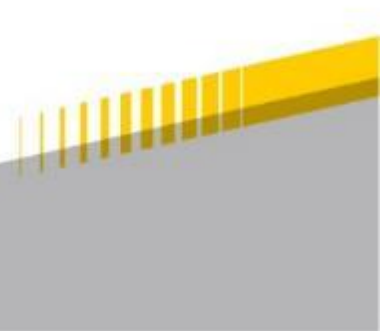


資料來源：uptime institute，MIC整理，2024年4月

- LEVEL 1:**
透過設備商軟體與BMS進行基礎監測
- LEVEL 2:**
監控環境和設備電力使用並調整基本控制（例如散熱）
- LEVEL 3:**
追蹤實體資料中心設備、能源和環境數據
- LEVEL 4:**
透過模型即時最佳化資料中心，AI應用在基於DCIM的資料湖以進行進階分析
- LEVEL 5:**
由AI驅動的整合管理系統根據資料中心整個生命週期的目標、規則和服務需求來最佳化資源

- 當前大部分的企業資料中心仍處於主動防範的第三階段，透過軟體實時監控資料中心數據
- 第二步需要進行AI模型的導入，才能將DCIM蒐集的數據進行最佳化的處理
- 未來則希望透過自動化的方式，將資料中心整個生命周期透過AI來進行規劃與因應調整

AI導入邊緣資料中心關鍵議題





邊緣資料中心透過多種樣態滿足低延遲需求

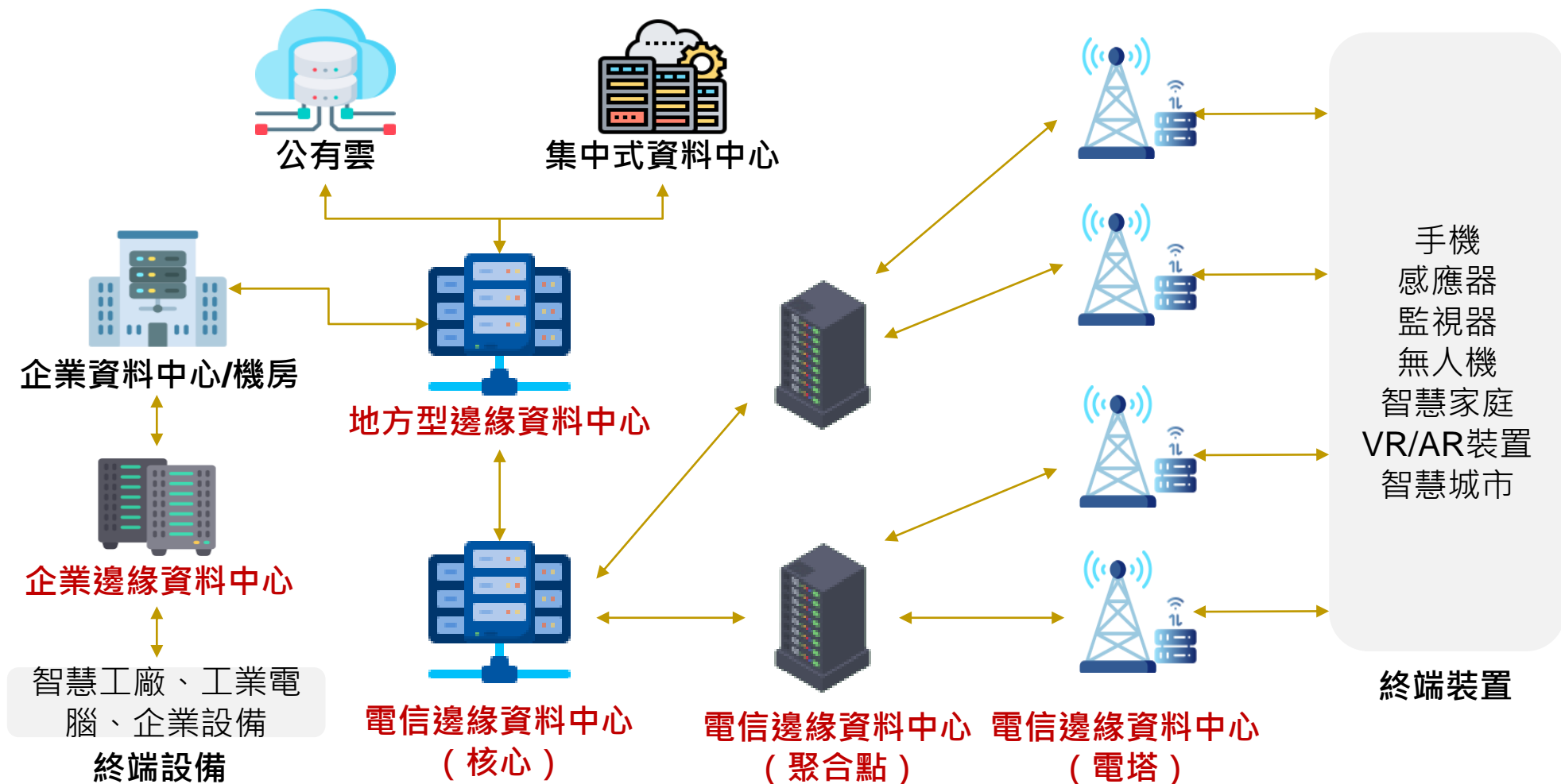
資料中心 類型	超大規模 資料中心 (HyperScale)	區域型 資料中心 (Regional)	中型 資料中心	地方型 資料中心 (Localized)	Edge			Edge	
					電信資料中心			企業資料中心	
類型	超大規模	區域型		Inner edge	Outer edge	Tower edge	企業 機房	企業 邊緣	
部署地點	偏遠地區	郊區		核心 (core)	聚合點	電塔	On-premise		
機櫃數量	>5,000	3,001- 5,000	801- 3,000	主要城市	主要 城市	小型 城鎮	遍布 全國	企業內	廠區
平均 延遲性				101-800	>10	2-6	最多 2櫃	11-200	1櫃
每1,000 萬人口的 DC數量				50 ms	40 ms	30 ms	10 ms	N/A	2-5 ms
				100座	10座	150座	3,000 座		

資料來源：STL Partners、MIC，2024年4月

- 邊緣運算可以存在於許多不同的位置，目的是比公有雲更接近用戶，因此邊緣資料中心的數量在不斷增加
- 地方型資料中心，於主要城市串接企業與城市內終端用戶；電信資料中心，透過不同形式串聯城鎮與偏遠地區
- 企業邊緣資料中心，放置於企業廠區內，透過最接近生產設備的方式，來降低整體的延遲性



終端設備與終端裝置需求不同類型的邊緣資料中心



- 終端設備透過企業邊緣資料中心，即時處理生產數據並反饋，大量且長時效數據存放至企業資料中心
- 終端裝置透過電信邊緣資料中心，使資料處理低延遲，預期部分小型推論亦可透過邊緣資料中心進行



藉由邊緣資料中心可以改善AI應用的執行

邊緣資料中心可以協助AI應用的重要因素

減少延遲

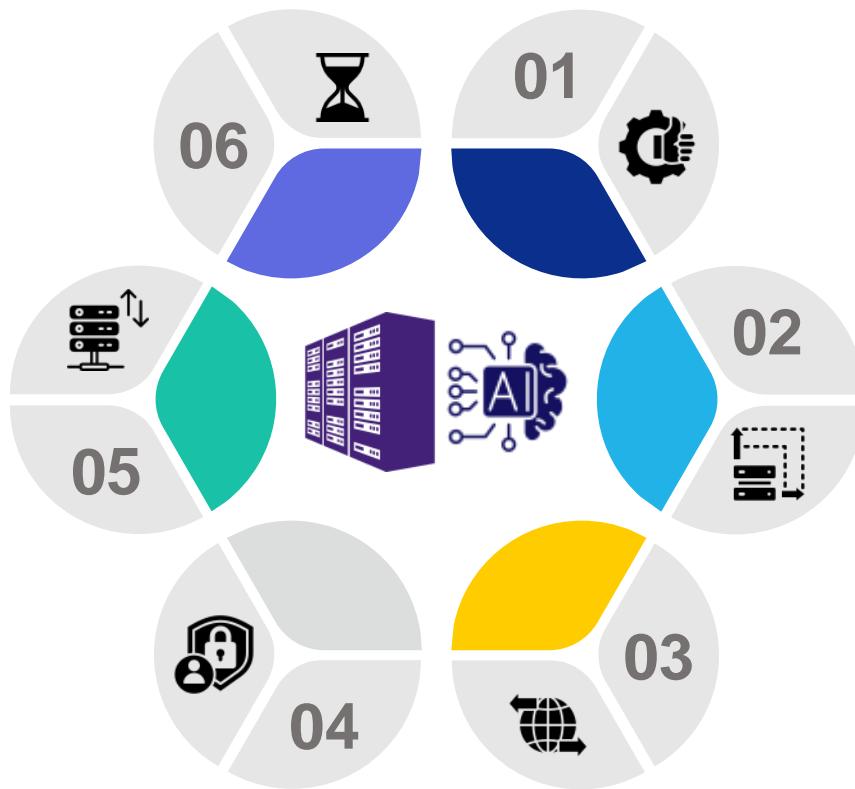
縮短傳輸距離，對需處理數據並即時決策的AI應用至關重要

提升頻寬效率

邊緣資料中心本地處理數據，AI應用可更有效地使用頻寬並減少必須透過網路發送的資料量

安全與隱私

AI應用需處理敏感資訊，邊緣資料中心可降低外洩風險並遵守資料主權法



可靠性與冗餘

即使與雲端資料中心網路斷線，邊緣資料中心也可以繼續運作，對需持續正常運作的AI應用十分重要

可擴展性

邊緣資料中心可根據需求增加容量，適用資料量遽增的場景，就如AI應用

網路交換中心

邊緣資料中心可串聯本地網路交換中心，提供更好的AI應用使用者體驗

資料來源：Proximity Data Centres，MIC整理，2024年4月

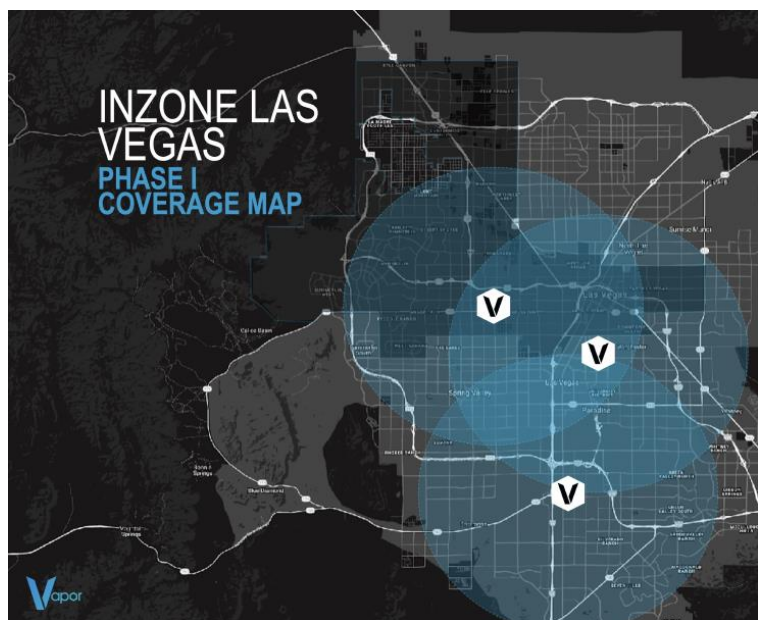
- 邊緣資料中心可以協助AI應用於各接近資料源的位置進行基礎資訊的迅速處理
- 主要效益包含降低AI應用的延遲性，並且提升頻寬效率、資料安全、可靠性、可擴展性等方面



案例：Vapor IO邊緣到邊緣的AI服務



Kinetic Grid 平台可以提供高度自動化系統所需的情境化遙測和程式控制，讓應用程式、操作工具和編排系統可以做出**智慧、演算法和人工智慧驅動的決策**



2023年8月，Vapor擴大與Les Vegas合作，從醫療區開始，在全市範圍內提供**普及的人工智慧**、專用5G、電腦視覺和其他物聯網應用

利用**人工智慧**來處理更接近資料來源的數據，減少延遲並確保即時回應。使市政府能夠部署需要即時決策的應用程式，從交通管理到公共安全

透過**AI應用**與電腦視覺集成，**即時分析影像**來源，從而實現人群監控、公共設施的預測性維護和增強的安全措施等應用

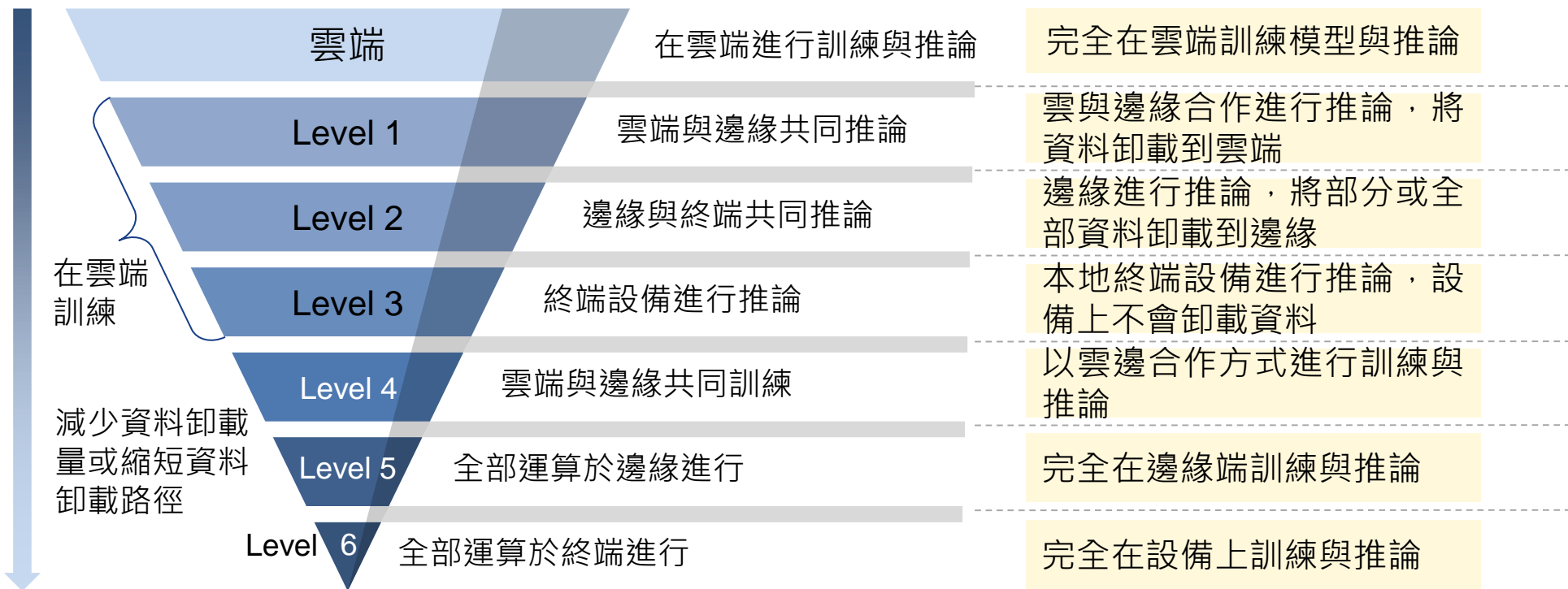
資料來源：Vapor IO · MIC整理，2024年4月

- Vapor的邊緣資料中心部署方式，可以讓AI應用在城市當中更有效率、低延遲的運行
- 協助城市當中的AI推論如即時影像分析，更有效的將資料傳輸到終端設備與終端用戶



AI訓練與推論從雲逐步朝邊緣擴展

AI運算從雲端到邊緣的六個層級

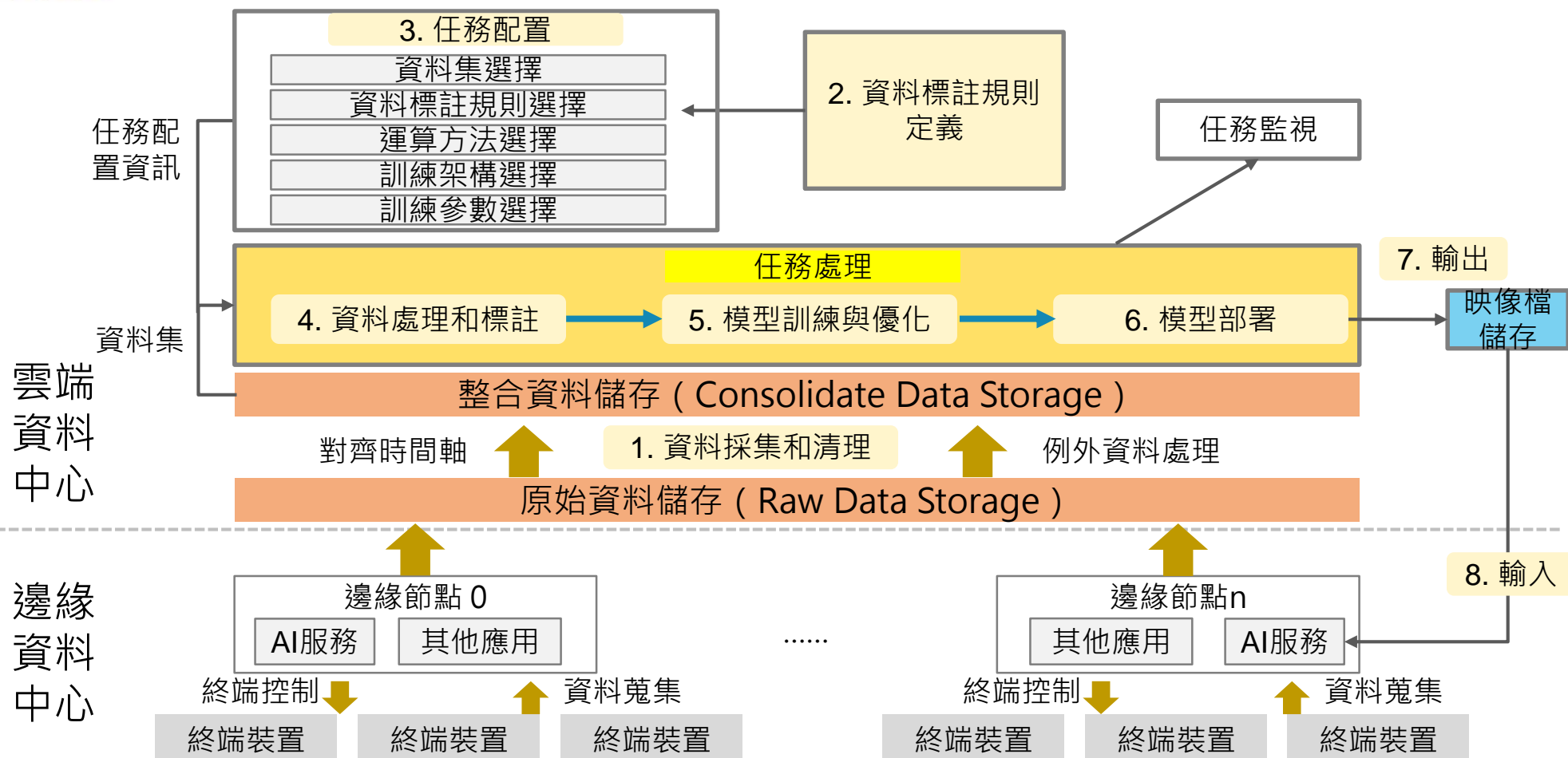


資料來源：《Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing》，MIC整理，2024年4月

- 當前AI訓練主要集中在雲端，由雲端服務商採購AI訓練伺服器進行訓練
- 推論則有部份轉移至邊緣，由企業自身或是邊緣資料中心採購AI推論伺服器進行協助
- 預計在未來幾年內不論AI訓練與AI推論均會往邊緣端發展



雲端與邊緣資料中心需協同進行AI訓練

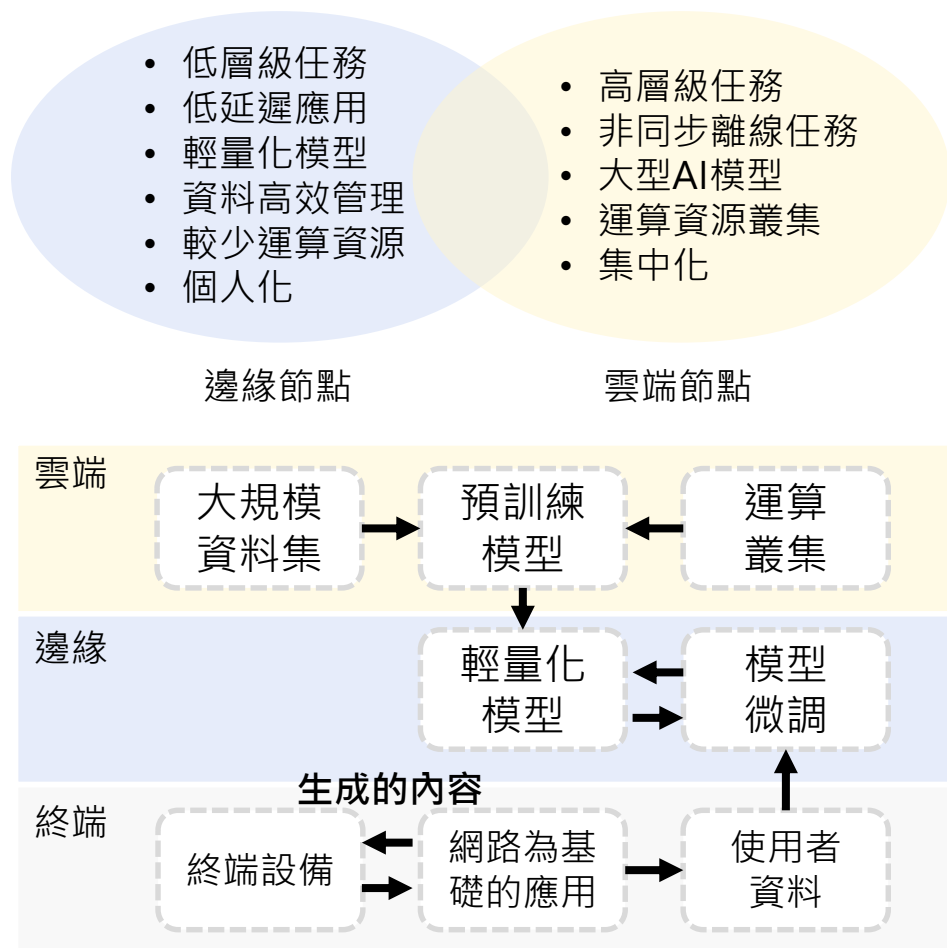


資料來源：《An AI Model Automatic Training and Deployment Platform Based on Cloud Edge Architecture for DC Energy-Saving》，MIC整理，2024年4月

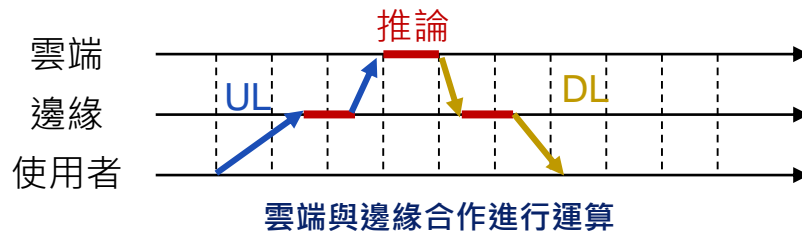
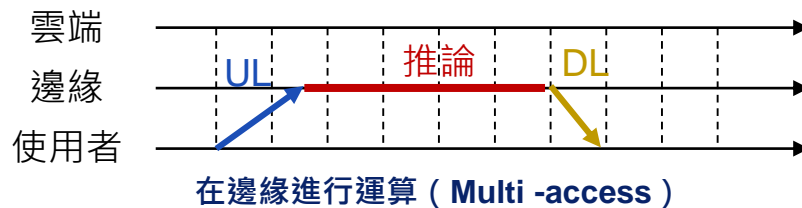
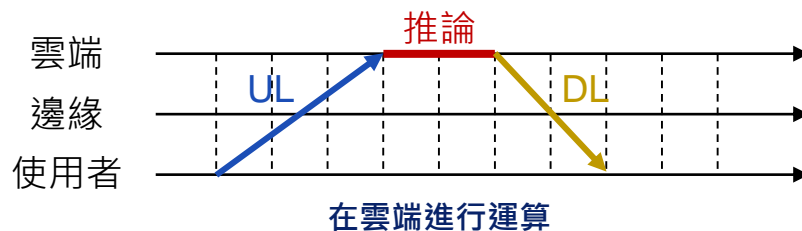
- 對於進行AI訓練而言，若要進行AI模型的即時更新，雲端與邊緣資料中心擁有不同的角色定位
- 藉由雲端與邊緣資料中心的共同協作，才能讓AI應用或服務更迅速的導入到終端裝置當中



生成式AI的即時性須透過雲端和邊緣共同加速



不同運算架構下的延遲性差距

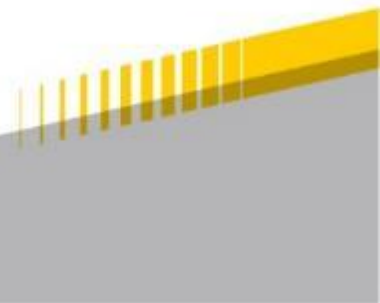


* UL=上行鏈結傳輸時間、DL=下行鏈結傳輸時間

資料來源：《An Overview on Generative AI at Scale With Edge-Cloud Computing》，MIC整理，2024年4月

- 雲端主要透過大規模資料集與運算叢集，訓練出預訓練模型，邊緣則使用輕量化模型來進行微調
- 在AI推論上同樣透過雲端與邊緣合作進行運算，可以使推論整體的傳輸效率更好

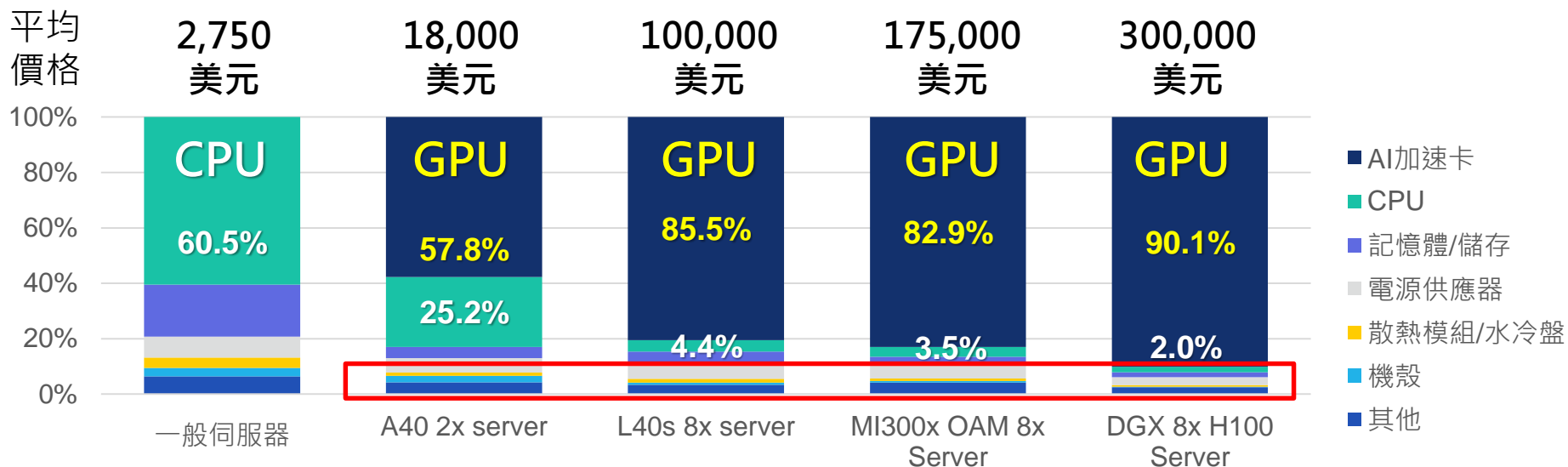
台灣產業發展機會





AI伺服器規格與成本占比更加複雜

一般伺服器 VS AI伺服器BOM表價格佔比



電源供應器

主流AI訓練伺服器搭載6顆鈦金級 3000wPSU，
當前廠商持續研發4000w~6000w的PSU

機殼

AI訓練伺服器機殼架構與高度調整，由1、2U
變為7U，使機殼整體單價增加

散熱模組/水冷盤

如氣冷所需的3DVC與直接式液冷所需的水冷盤

其他

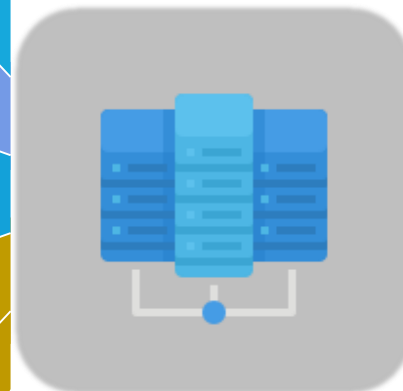
AI伺服器所需BMC數量增加、PCB單板層數增加

資料來源：MIC，2024年4月

- AI伺服器的規格逐漸多元化，在最終售價以及BOM表占比方面均有所不同
- 儘管伺服器零組件的單價均上升，伺服器當中GPU成本比重仍在提升



AI資料中心架構促使不同類型台灣廠商升級產品規格

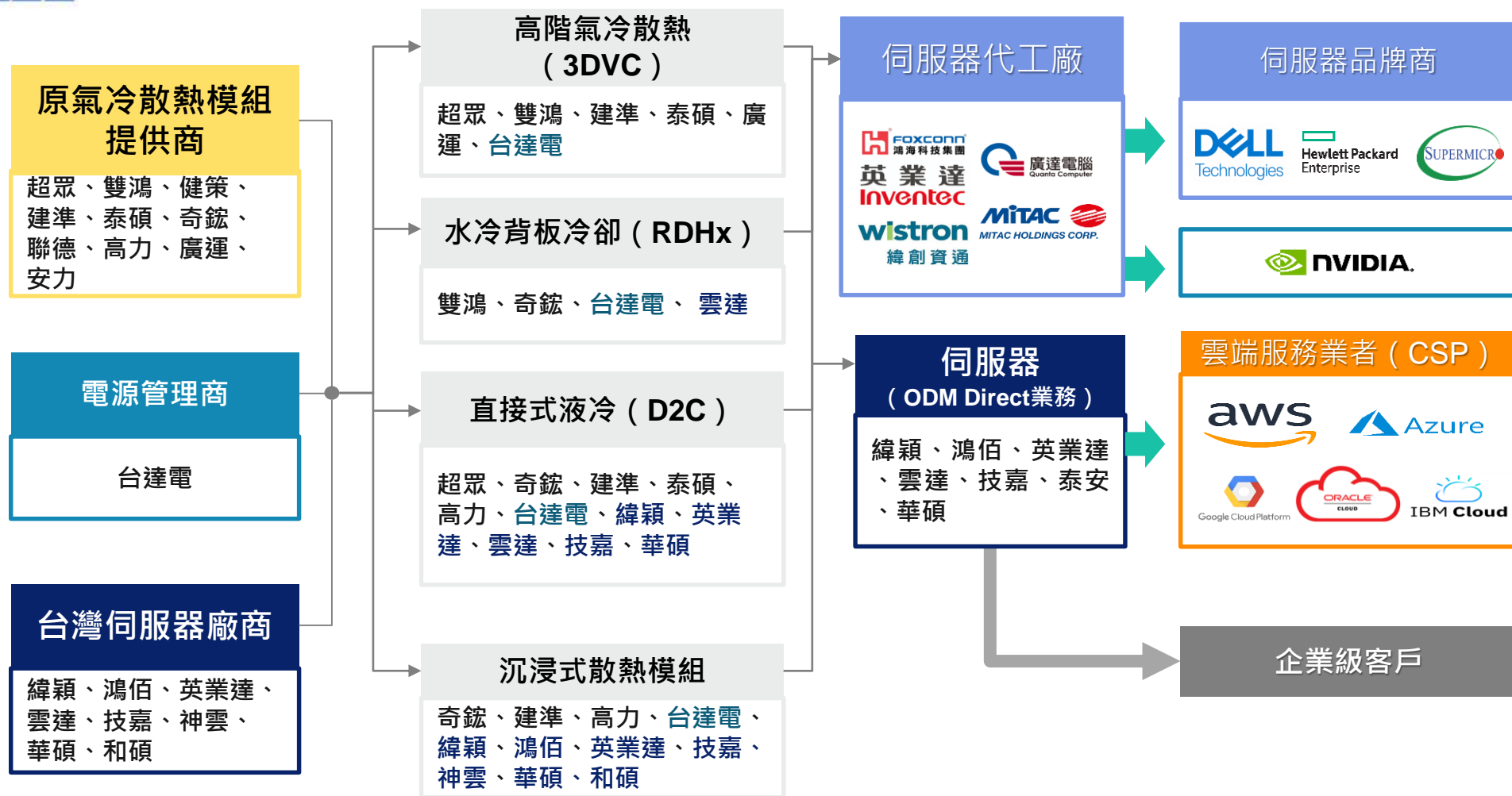


資料來源：MIC，2024年4月

- 台灣企業在AI硬體耕耘多年，AI資料中心架構主要影響AI伺服器發展
- 另外對於網通設備、儲存設備、散熱與基礎設施相關廠商均將造成影響



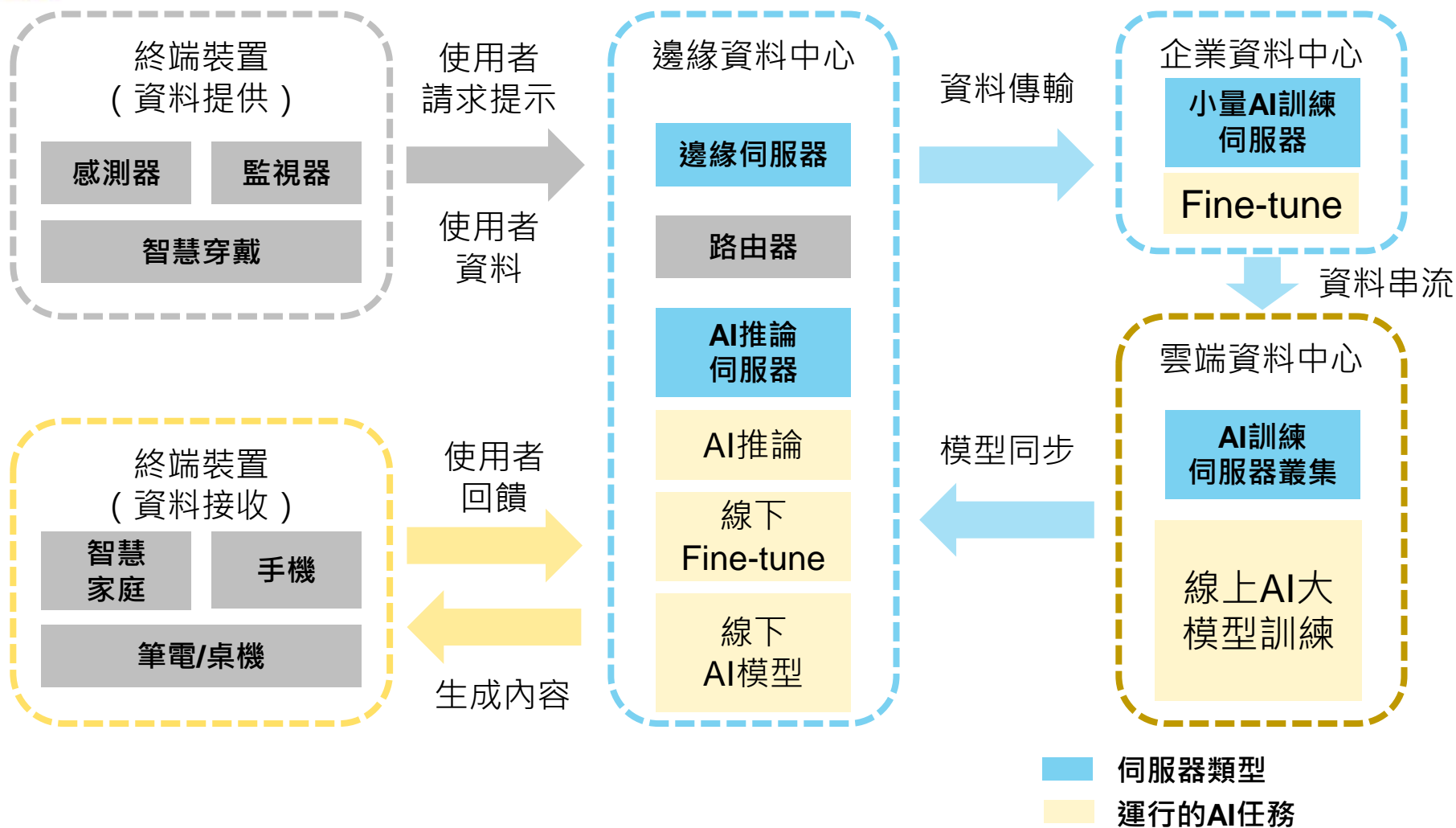
AI伺服器帶動液冷散熱吸引台廠進行布局



資料來源：各公司，MIC整理，2024年4月

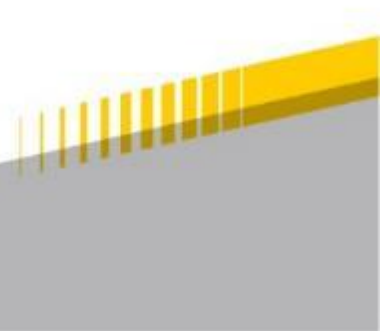
- 台灣既有的散熱模組提供商、電源管理商與伺服器廠商，紛紛開始布局液冷散熱
- 透過不同的散熱產品組合，滿足客戶需求並提供更完整的AI硬體解決方案

生成式AI應用將同步帶動AI訓練與推論伺服器需求



- 要使生成式AI應用運作最佳化，首先需要雲端資料中心中AI訓練伺服器叢集的AI大模型訓練
- 其次邊緣資料中心邊緣伺服器、AI推論伺服器的Fine-tune與AI推論共同協助，將同步推升產品需求

結論





結論 (1/2)

● 2024年雲端服務商建設AI資料中心，AI伺服器採購增加

- ◆ 因生成式AI帶動的大量AI算力需求，2024年雲端服務商持續採購AI伺服器，以**建造AI資料中心**。伺服器品牌商推出的**AI伺服器新品**上市，將使企業端提升對AI伺服器的採購，將帶動全球AI伺服器的需求

● AI算力需求使資料中心需要調整不同構面的建置

- ◆ AI改變資料中心設計架構可以從**IT設備**、**基礎設施**及**營運**來進行探討，AI算力的急遽增加促使**運算**、**通訊**、**儲存**的IT設備均須做出因應調整。基礎設施則是包含**液冷散熱系統**、UPS、配電裝置等需進行升級，營運方面**透過AI來智慧化監控資料中心**的運作，可望大幅提升運作效率

● 邊緣資料中心連接雲端與終端，改善AI應用的執行

- ◆ AI應用擁有許多**AI推論**的需求，邊緣資料中心可以在**接近資料源**的位置，協助進行**資料的處理**，藉此降低延遲、處理敏感資訊並提升可靠性



結論 (2/2)

- 生成式AI需透過雲端與邊緣資料中心合作來進行最佳化

- ◆ 透過在**雲端**資料中心的**AI模型訓練**、**大語言模型處理**，及**邊緣**資料中心的**微調 (Fine-tune)**、**輕量化模型**、AI推論，可以大幅改善生成式AI應用的延遲性

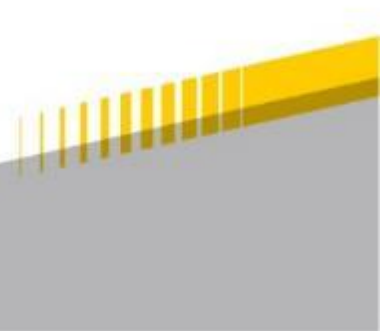
- AI硬體解決方案規格多樣化，帶動台廠布局液冷散熱系統

- ◆ 台系的伺服器廠商由代工廠擴展為**AI硬體解決方案提供商**，因此除AI伺服器之外對基礎設施進行更廣泛的布局。當前**液冷散熱**系統成為**原氣冷散熱提供商**、**台灣伺服器廠商**、**電源管理商**共同切入的目標

- 生成式AI應用同步帶動AI訓練與AI推論伺服器出貨

- ◆ 展望未來，要打造完整的AI應用體系需要**AI訓練伺服器**在**雲端進行叢集**來進行AI模型訓練，**AI推論伺服器**於**邊緣端進行AI推論、微調**，因此將會同步帶動高階AI訓練伺服器以及AI推論伺服器的需求

附件

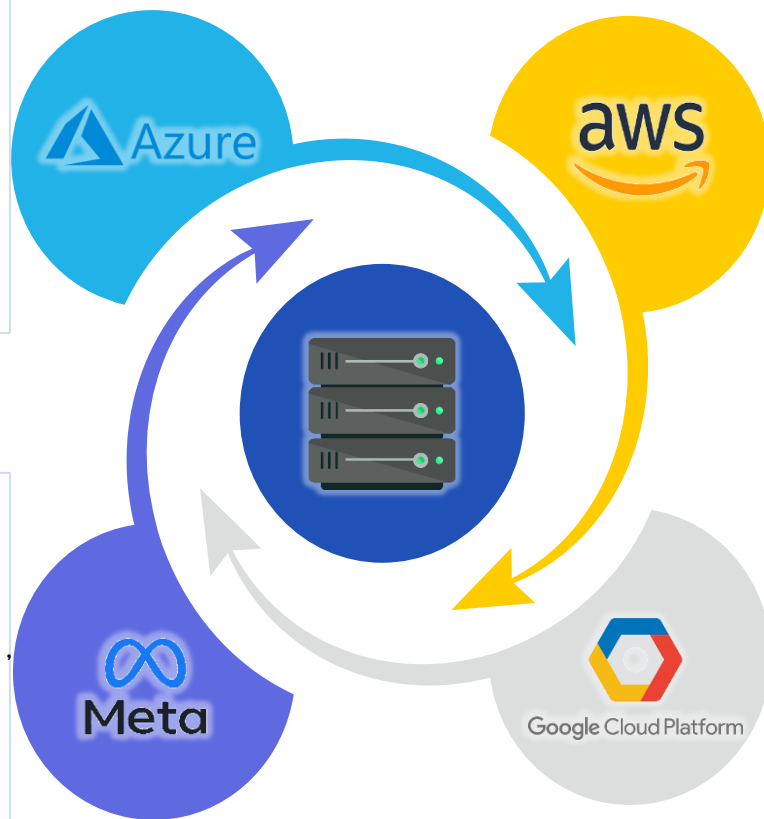




2024年AI資料中心為雲端服務商投資重點

為了打造支援**AI的資料中心**，預計2024財年**每個季度增加資本支出**。Azure營收增長速度有望在2024下半年重回逐季增加的態勢，**AI相關服務**佔比亦將逐季提高

預計2024全年資本支出將為**300-370億美元**，將於推出Llama 3、擴展Meta AI 助理等服務。預期2024年底，基礎設施將包括**35萬張NVIDIA的H100 GPU**



2024年**持續投資基礎設施**，以支援AWS客戶需求，包括**大型語言模型 (LLM)** 和**生成式AI**投資

預期 2024全年資本支出將高於2023年，其中至少**90%集中在技術基礎設施上**以滿足其雲端業務及人工智慧所需的基礎建設

資料來源：各公司，MIC整理，2024年4月

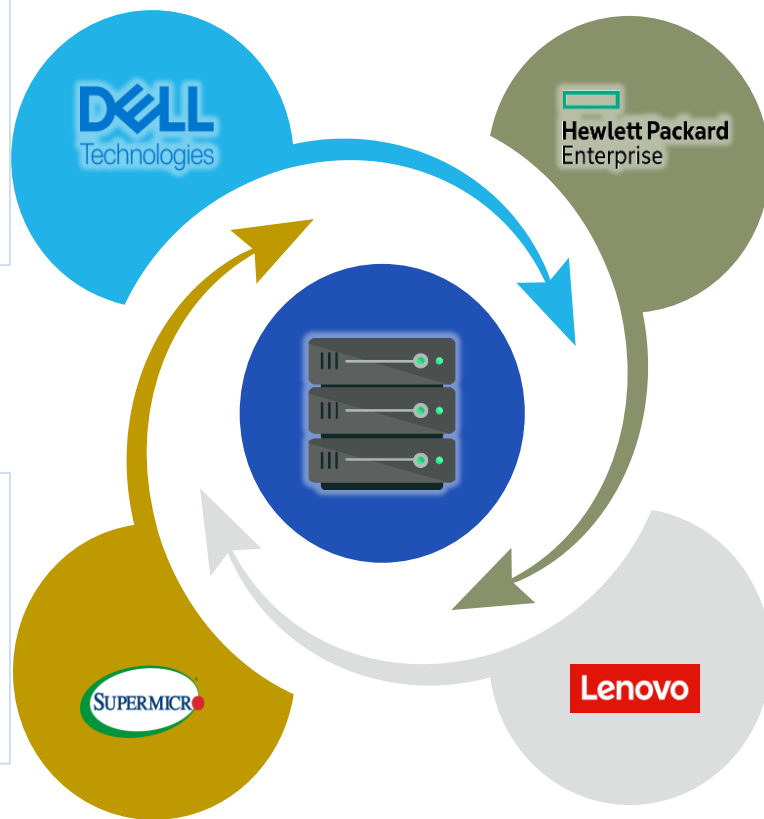
- 生成式AI與大型語言模型成為雲端服務商的投資重點，並且在2024年會將大部分的資本支出投資在建設AI資料中心、AI超級電腦方面，藉此來訓練新的模型提供新的服務



企業端AI需求促使伺服器品牌商2024年需求回穩

AI伺服器所帶動的出貨仍在繼續，有望帶動**訂單成長近40%**，積壓訂單有望翻倍，截至2023年底，AI伺服器的**積壓訂單為29億美元**

受惠AI伺服器訂單的需求，且目前客戶仍然需要更多AI伺服器和整體IT解決方案，2024上半年預期有**70-80億美元**的營收。



預計2024年**HPC**和**AI**的需求將持續改善，預計全年收入增長**4%至6%**。
HPE精簡產品組合和提供**雲原生數據**服務的戰略正在推動增長並提高利潤

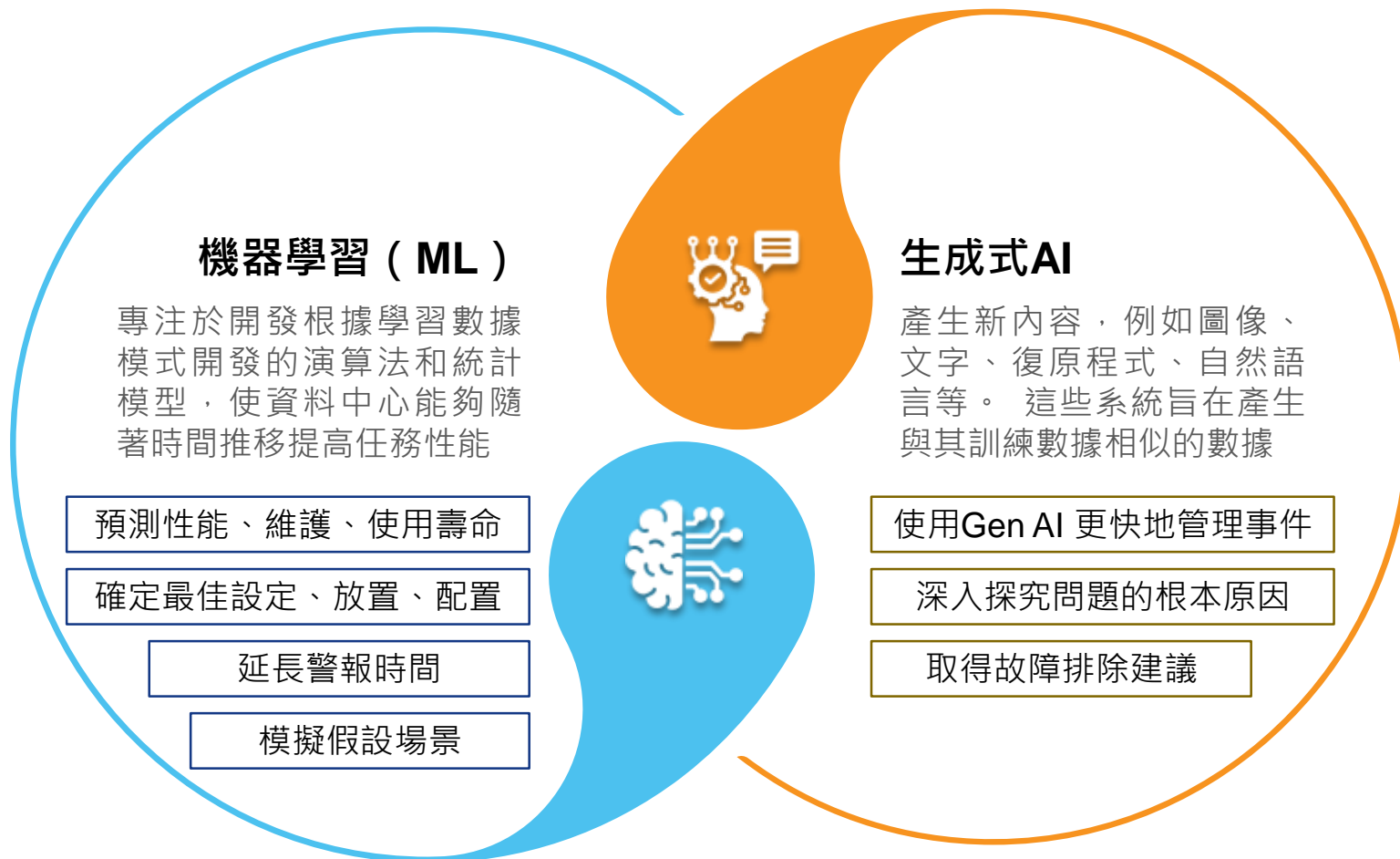
預估2024年**設備即服務市場**保持雙位數增速、**雲端運算解決方案**市場復合增長率**18%**

資料來源：各公司，MIC整理，2024年4月

- 儘管當前全球伺服器市場在通用伺服器仍未回復到既有水準，但是2024年AI伺服器的需求量擴展到企業端及中小型業者，促使品牌商新推出的AI伺服器規格前景看好



生成式AI與ML為資料中心營運帶來不同價值



資料來源：各公司，MIC整理，2024年4月

- 生成式AI與機器學習在資料中心內部營運產生不同作用，生成式AI主要更快的管理事件與生成故障建議
- 機器學習可對溫度、濕度、電力等基礎資訊蒐集與分析，強化預測性維護、延長警報等方面的性能



邊緣資料中心透過接近終端用戶串聯雲與地

邊緣資料中心最**接近設備和終端用戶**，旨在更快地**處理時間敏感的數據**

Hewlett Packard Enterprise

邊緣資料中心是位於**網路邊緣的小型設施**。可在更短距向設備提供運算資源，**減少延遲**並改善用戶體驗

Nlyte Software

邊緣資料中心是**較小的設施**，靠近其服務人群，向最終用戶提供雲端運算資源和快取內容

Sunbird

邊緣資料中心是一種**分散式設施**，具冷卻和電力基礎設施，可在**靠近資料產生處或資料使用處**提供儲存和運算

edgeuno

邊緣資料中心是一個**小型資料中心**，電力容量不超過2MW，放置在「邊緣」的許多位置：靠近人員、機器產生和使用數據

pwc



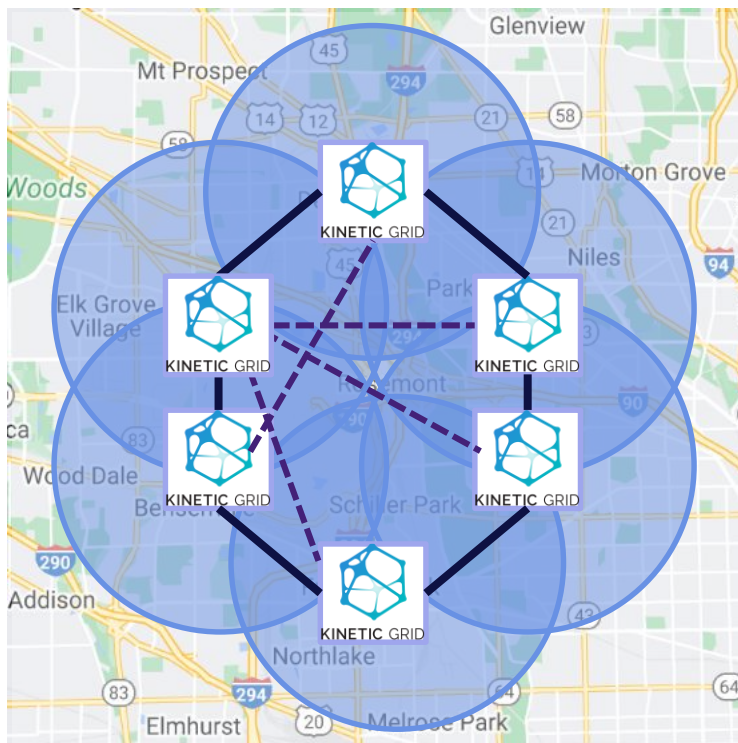
資料來源：各公司，MIC整理，2024年4月

- 綜整各家廠商的論述，可以發現邊緣資料中心為一個分散式IT架構
- 主要目的為透過接近設備與終端用戶等資料產生處或使用處，藉此處理低延遲的數據

案例：Vapor IO透過光纖網路打造Kinetic Grid平台



Vapor IO推出Kinetic Edge技術架構，將多個Vapor IO 邊緣資料中心組合成一個虛擬資料中心，該資料中心可以跨越整個城市，並提供超過10個可用區域



Vapor IO 的 Kinetic Edge 架構創建城市規模軟體控制虛擬資料中心，由許多透過**高速光纖**連接在一起的**邊緣資料中心**組成，涵蓋整個地理區域

Vapor IO 在美國佈建總長達**30,027公里**的光纖網路，擁有**252個邊緣互連站點**



Vapor IO**邊緣資料中心**專為極高容量和營運效率而開發，可在邊緣配置提供足夠的效能，並通常**位於環境條件惡劣的偏遠地區**

資料來源：Vapor IO，MIC整理，2024年4月

- Vapor透過大量部署小型、微型的邊緣資料中心，為客戶打造更低延遲、高傳輸效率資料中心架構
- 藉由將各個資料中心用光纖網路連接，形成智慧、靈活的邊緣基礎設施系統

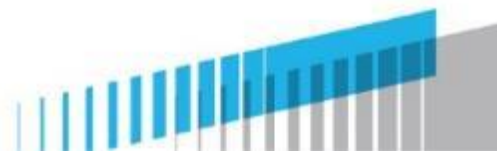


MIC 產業提昇的關鍵力量
Thank You


陳牧風 產業分析師

Stephenchen@iii.org.tw

產業情報研究所



智慧財產權暨引用聲明

- 本活動所提供之講義內容或其他文件資料，均受著作權法之保護，非經資策會或其他相關權利人之事前書面同意，任何人不得以任何形式為重製、轉載、傳輸或其他任何商業用途之行為
 - 本講義內容所引用之各公司名稱、商標與產品示意照片之所有權皆屬各公司所有
 - 本講義全部或部分內容為資策會產業情報研究所整理及分析所得，由於產業變動快速，資策會並不保證本活動所使用之研究方法及研究成果於未來或其他狀況下仍具備正確性與完整性，請台端於引用時，務必注意發布日期、立論之假設及當時情境
- 

AISP 情報顧問服務

Advisory & Intelligence Service Program

產業情報顧問服務AISP為資策會MIC最核心的產業情報資料庫服務，運用最先進數位平台服務技術，提供產業在資訊與通訊（ICT）領域最完善的新知識、新技術、新方向的產業情報資訊服務平台。服務內容包括「產業情報資訊、突發事件觀察剖析、關鍵議題焦點評論、產業議題深度研究、國際大展情報蒐集分析、前瞻趨勢」等。隨時觀察產業發展動態與趨勢，觀測掌握全球重要的產業發展動態，並依據產業需求規劃研究範疇與議題，開展符合產業需求的產業情報資料庫。

推薦資料庫



Application IC & Components 應用IC與關鍵零組件

本產品研究範疇包含應用IC與關鍵零組件於新興應用之發展，聚焦新興應用晶片與零組件技術、應用晶片與技術發展、晶片大廠競合與熱門議題等。透過技術分析、應用分析、產品分析與市場動態等不同面向，探討半導體晶片與關鍵零組件導入新興應用之相關發展。

研究範疇

- 應用IC與關鍵零組件於新興應用之發展分析

研究重點

- 新興應用晶片與零組件技術
- 應用晶片與技術發展
- 晶片大廠競合與熱門議題

研究構面

- 技術分析
- 應用分析
- 產品佈局分析
- 市場動態分析

Performance Computing 運算系統

本產品針對電腦主機板、桌上型電腦與伺服器等資訊系統產品，並新增高效能運算、資料中心、邊緣運算與雲端服務大廠之重要議題，除原本產銷訪查與趨勢分析，另針對重要議題之產業發展、產品動態進行研究剖析，以協助上下游業者掌握運算系統產業未來商機。

研究範疇

- 一般資訊運算暨高效能運算系統產品之產業趨勢與市場前景

研究重點

- 桌上型個人電腦與其主機板
- 伺服器與企業資訊運算系統
- 資料中心技術與應用發展
- 邊緣運算與分散式架構
- 雲端運算產業與政策研析

研究構面

- 市場分析
- 產銷分析
- 產品發展分析
- 關鍵晶片分析
- 產業競爭分析

AISP情報顧問服務網
<https://mic.iii.org.tw/aisp>

瞭解更多

Artificial Intelligence 人工智慧

本產品以「AI產業化」及「產業AI化」兩大主軸進行研究，在「AI產業化」上探討各式新興AI及生成式AI演算法之應用、AlaaS服務、人工智慧硬體及晶片、軟體框架等議題；「產業AI化」則探討不同行業AI及生成式AI技術於場域之議題、產業AI化動態及新創應用案例等內容。

研究範疇

- AI產業化之相關軟硬體、治理議題，以及產業AI於不同垂直領域之應用發展

研究重點

- 新興AI軟硬體及平台
- AI新興算法與服務
- AI大廠領域佈局動向
- 重點應用領域趨勢
- 可信任AI與AI評測

研究構面

- 技術趨勢前瞻
- 標竿廠商動向
- 國際重點政策
- 產品發展分析
- 標竿應用案例

MIC到府簡報服務

趨勢洞察力 決定 企業競爭力

MIC 協力為您促進 組織 / 人才 再升級

組織人才前瞻力的提升，儼然已成為現今企業突破轉型的新顯學。為成功協助企業菁英掌握瞬息萬變的市場趨勢，特別針對產業熱門議題以及MIC重點研究，提供研究顧問至貴公司「到府簡報」之服務，期盼能將MIC多年凝聚累積的研究能量及專業精闢的情報服務，深耕企業內部員工，加速提升組織競爭力，共創企業新價值，與企業組織人才攜手找出迎向新經濟的解方



15 大
議題精選

- 產經趨勢
- 資訊產業
- 半導體產業
- 5G/B5G
- 數位經濟
- 金融科技
- 科技應用
- 電動車
- 人工智慧
- 數位轉型
- 資安防護
- 智慧城市
- 智慧製造
- 智慧醫療
- 能源與環境

點擊詳閱
MIC到府簡報議題



欲瞭解詳情，請洽MIC會員服務中心，由專人為您服務

☎ (02)2378-2306 ✉ members@iii.org.tw

MIC 資策會 | 產業情報研究所