



# ChatGPT引發的AI產業發展趨勢

福邦投顧 研究部  
2023.04

# 結論

- 1) 隨著AI運算能力逐漸上升，以及ChatGPT熱潮帶動市場對生成式AI的關注，使得在文字、影像和程式等領域的應用有望逐步實現商業化。預計未來10年市場規模將達27%的年複合成長率。
- 2) 隨著AI技術的普及和發展，對AI伺服器的需求持續增加，帶動相關產業快速成長。將為硬體製造商、資料中心營運商和AI應用開發商等相關企業創造更多商機。
- 3) 台灣廠商在AI發展方面的優勢主要在半導體領域。AI晶片包括GPU、FPGA和ASIC等類型。在雲端AI訓練中，GPU具有顯著優勢。然而，隨著AI應用的商業化逐漸成熟，Edge端對ASIC的需求將呈倍數成長。
- 4) 隨著AI晶片需求持續增加，相應地帶動了高頻寬記憶體、傳輸介面IP、晶圓代工先進製程及封裝需求的上升。在台灣廠商中，台積電與相關矽智財公司將受惠最為顯著。AI伺服器零組件規格也較高，其中散熱需求顯著提升，散熱模組廠商有望從中獲益。
- 5) 相關個股：台積電、創意、世芯、雙鴻、奇鋆、穎崴

# ChatGPT帶動AI發展熱潮

# AI深度學習技術演進，推動生成式AI技術發展

- 在AI技術中，深度學習（Deep Learning）是一種透過從原始資料中逐步提取更高層級特徵值的方法。最終，根據這些特徵值的加權組合，可以得到相應的結果。這個過程主要依賴演算法進行計算分析，從而降低人為介入的比例。
- 生成式AI（Generative AI）是在深度學習基礎上，結合語音辨識和自然語言生成技術而發展出來的。與決策/分析式AI不同的是，生成式AI不僅僅依賴已有數據進行判斷和預測，而是能夠利用數據進行新內容的創造生成式AI可應用於多個領域，如文字、圖像、影片、科學研究、人機互動以及醫療診斷等（決策/分析式AI主要應用在人臉辨識、廣告、金融等領域）。

圖1、人工智慧技術關聯圖

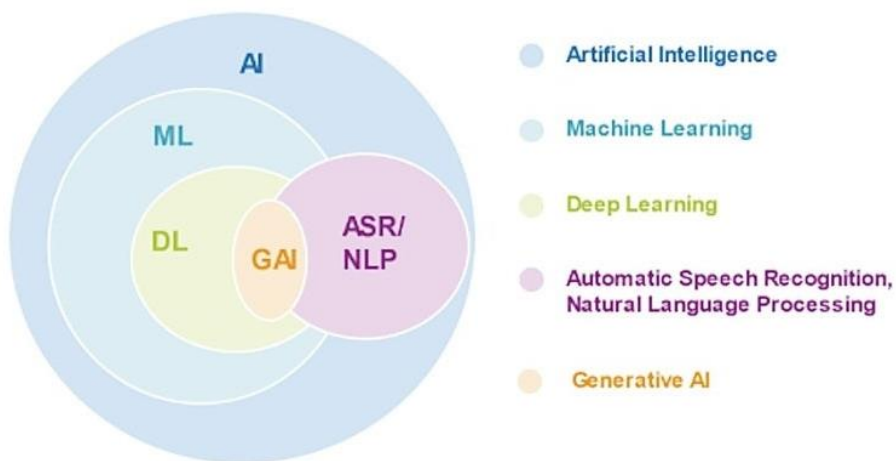
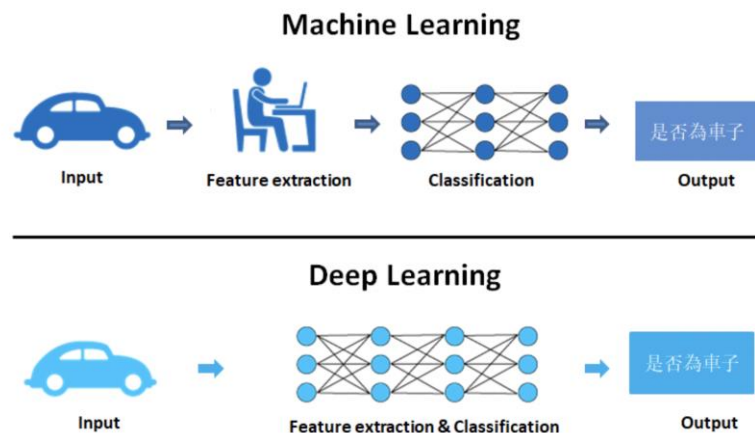


圖2、深度學習示意圖



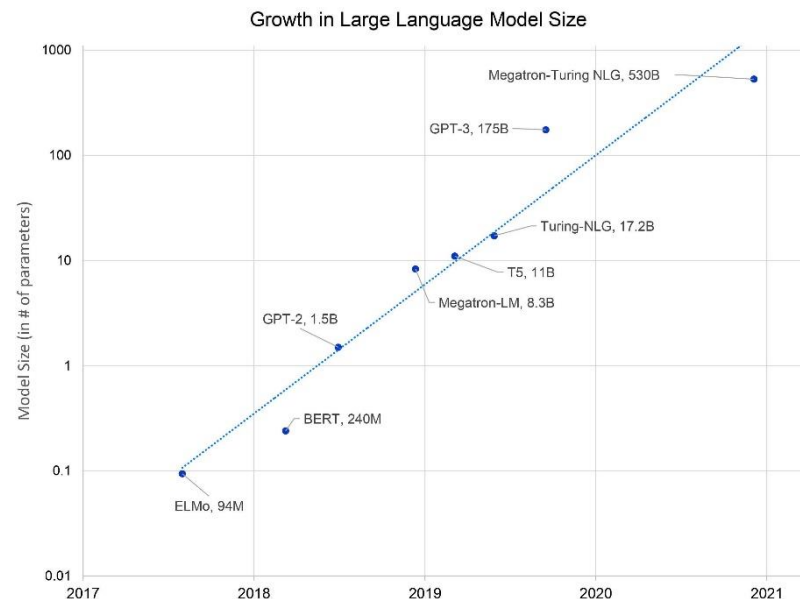
# ChatGPT引發熱潮，技術來自於Open AI的GPT-3

- ChatGPT技術來自於美國人工智慧研究公司OpenAI，微軟為其投資者之一，第一代GPT（生成式預訓練，Generative Pre-Training）技術發布於2018年，並採用Transformer模型的解碼器架構，透過預先輸入大量資料產生詞庫，並在同時處理前一輸入逐一生成輸出，加快生成速度
- 2020年OpenAI推出了GPT-3，該模型在規模方面有了顯著提升，並開放了API供開發者使用。2022年基於GPT-3的技術，OpenAI推出ChatGPT(可視作GPT-3.5)。相較於前代模型，ChatGPT在語言生成應用上有著明顯的進步，迅速吸引了市場的廣泛關注。

表1、GPT模型發展概況

MODEL	YEAR	參數(10億)	預訓練資料量
GPT	2018	0.11	5GB
GPT-2	2019	1.5	40GB
GPT-3	2020	175	45TB
GPT-4	2023	?	?

圖3、不同語言AI模型推出年份與參數



# ChatGPT 推進市場對AI模型之關注度

- 大型語言模型（LLM）是一種具有大量參數的自然語言處理（NLP）系統。它通過對大量文本進行訓練，使模型能夠理解問題並根據需求生成相應的回應，例如新的文本、程式碼或圖像。這種技術大幅提升了人工智能在語言理解和應用方面的能力。

圖4、全球主要LLM及其參數





# AI商業價值浮現，國際資金擴大對生成式AI投資

- 隨著AI訓練資料量的增加，其在商業應用中的價值逐漸顯現。這吸引了國際資金大量投資於AI新創公司。在2022年，投資金額超過了26億美元。在各種應用中，文字和影像在商業輔助使用上的應用相對成熟，因此獲得了較多的資金投入，但在程式、影音等應用也正快速成長中。

圖5、生成式AI投資金額(百萬美元)

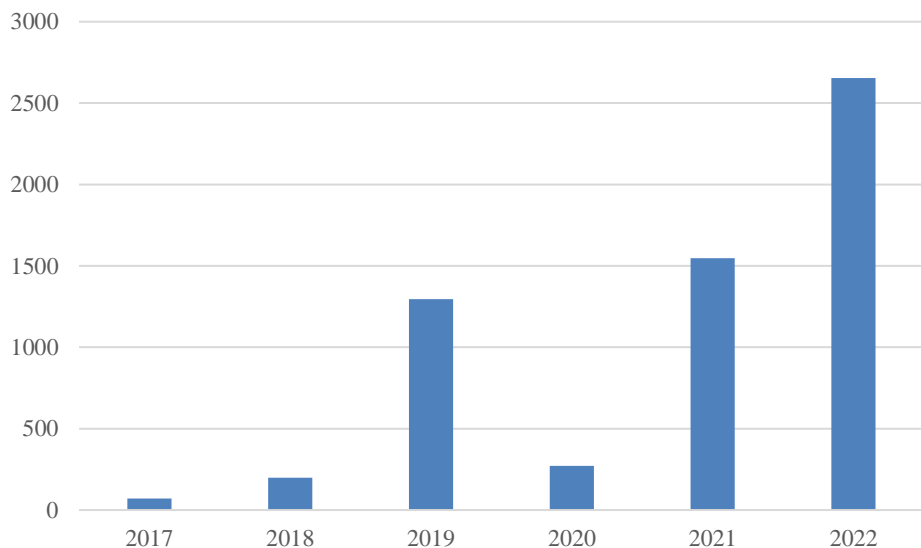
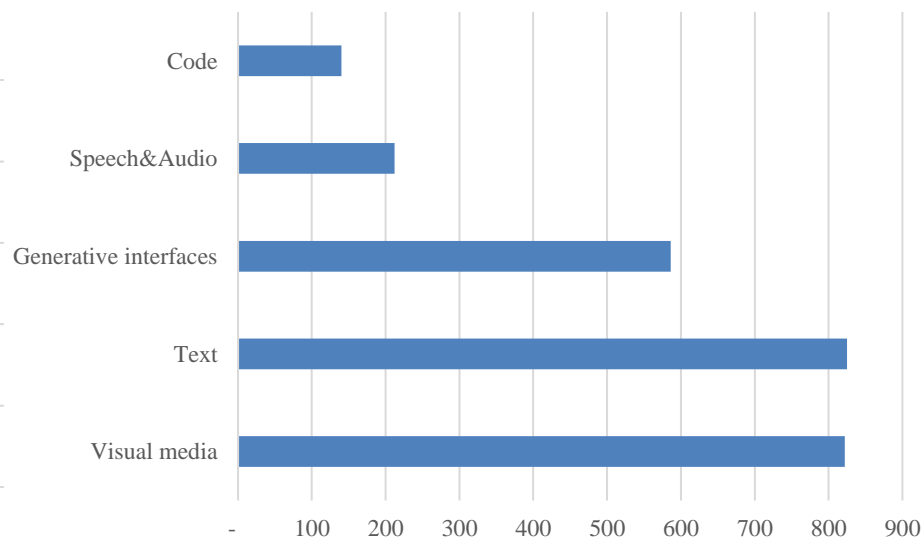
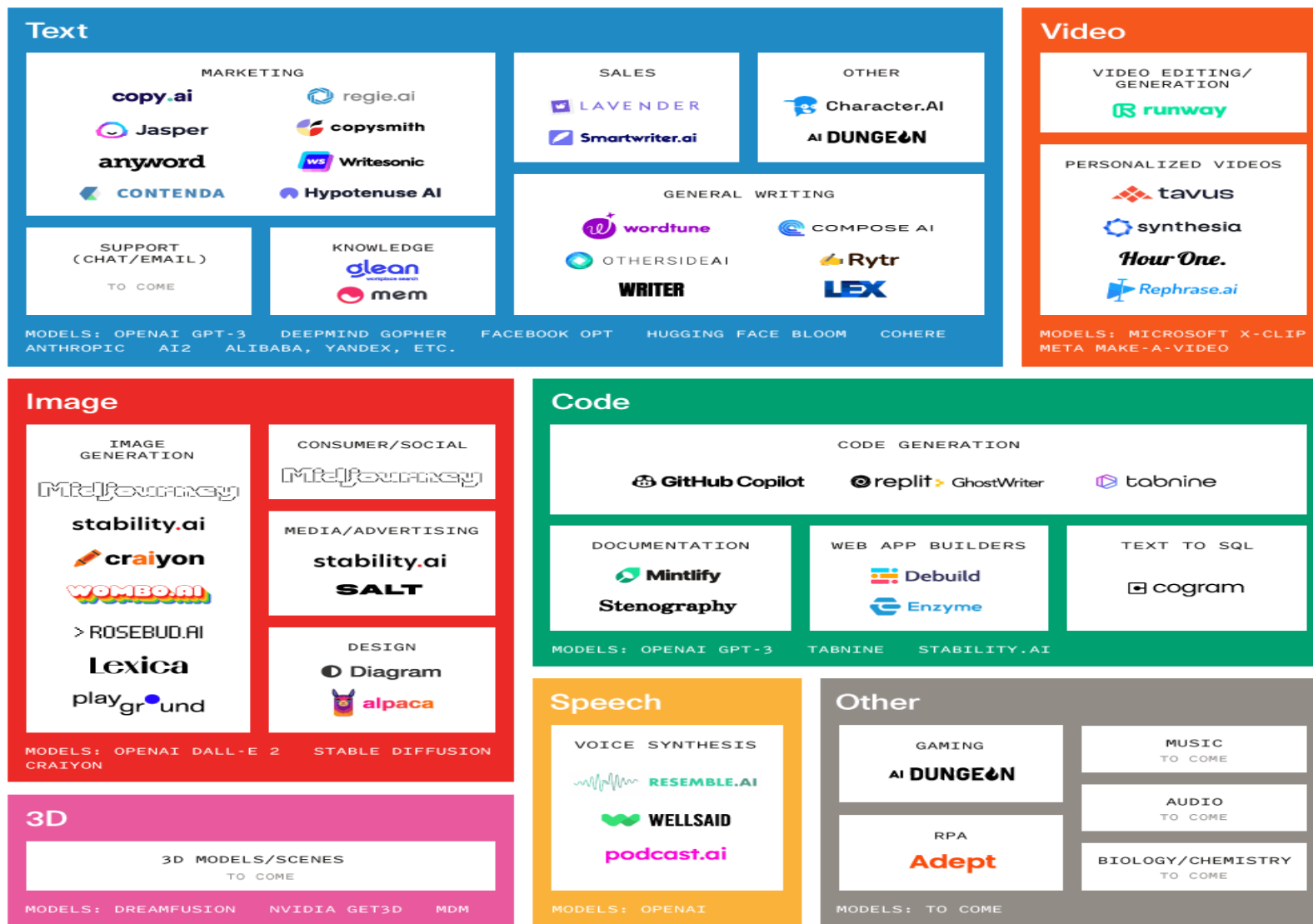


圖6、2022年各類型生成式AI新創投資金額(百萬美元)



# 生成式AI在文字、圖像、演講、程式多有應用

圖7、全球生成式AI公司





# ChatGPT可用於搜尋、文字整理、寫程式等應用

- ChatGPT可用於搜尋答案、整理和翻譯文本以及生成程式碼等應用，有助於提高用戶的工作效率。然而，該技術仍存在一些缺陷，例如解答不精確和無法進行精準預測等。因此，ChatGPT尚無法完全取代搜尋引擎和判斷分析等工作。

圖8、ChatGPT的各種應用

LE 請幫我寫一份圖表自動化的excel程式

為了提供您一個Excel自動化圖表的示例，我將建議使用VBA (Visual Basic for Applications) 編程。以下是一個簡單的示例，該示例將從一個數據範圍創建一個折線圖。

假設您的數據位於工作表“Sheet1”的A1:B6範圍內，以下是創建折線圖的VBA程式：

```
vba Copy code
Sub CreateLineChart()
    ' 定義數據範圍和圖表位置
    Dim DataRange As Range
    Dim ChartPosition As Range

    ' 設置數據範圍
    Set DataRange = Worksheets("Sheet1").Range("A1:B6")

    ' 在工作表上設置圖表的位置
    Set ChartPosition = Worksheets("Sheet1").Range("D1:K10")

    ' 創建一個新的圖表
    Dim LineChart As ChartObject
    Set LineChart = Worksheets("Sheet1").ChartObjects.Add( _
        Left:=ChartPosition.Left, Width:=ChartPosition.Width, _
        Top:=ChartPosition.Top, Height:=ChartPosition.Height)

    ' 使用數據範圍生成折線圖
    With LineChart.Chart
        .SetSourceData Source:=DataRange
        .ChartType = xlLine
        .HasTitle = True
        .HasLegend = False
        .ChartTitle.Text = "自動化折線圖"
    End With
End Sub
```

要運行此程式，您需要將其添加到Excel的VBA編輯器中：

1. 打開Excel，並打開您希望使用的工作簿。
2. 按`Alt + F11`鍵打開VBA編輯器。
3. 在VBA編輯器中，單擊菜單中的“插入”，然後選擇“模塊”。
4. 在新模塊中，將上面的程式代碼粘貼到模塊窗口中。
5. 按`Ctrl + S`保存工作簿。
6. 按`F5`運行程式。

在運行此程式後，您將在工作表“Sheet1”的D1:K10範圍內看到一個基於A1:B6範圍數據的折線圖。

LE 請幫我順稿

當然，我們很樂意幫您順稿。請問您需要哪方面的順稿幫助？請提供更多具體的信息，以便我們更好地理解您的需求。

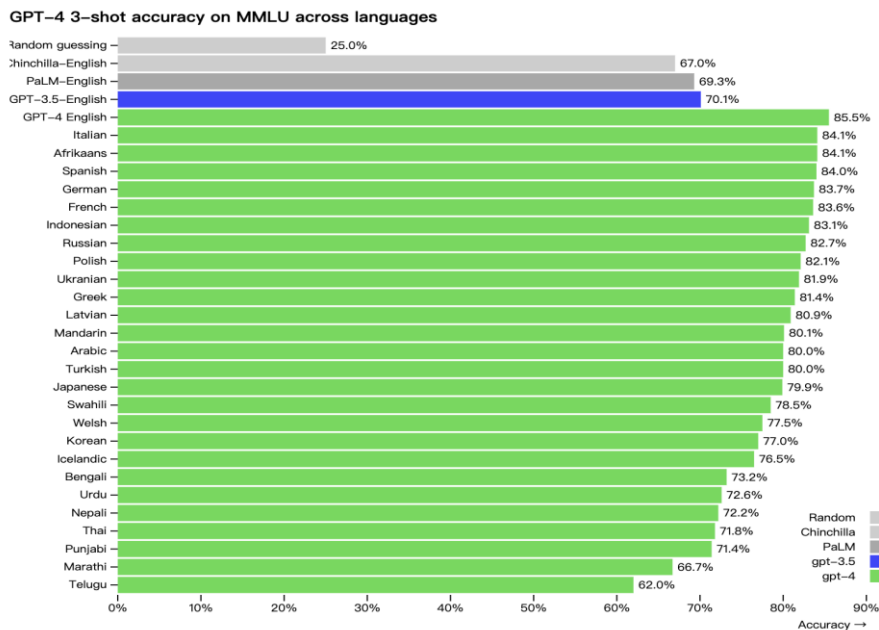
LE 根據金融時報報稱，ARM預計將禁止提供最新技術(Neoverse V系列架構授權) 給中國IC設計廠商平頭哥，平頭哥為中國伺服器與AI晶片的指標性IC設計公司之一，ARM的授權禁令將進一步限制中國AI發展，平頭哥在台積電投片生產14nm以下製程，但預計營收占比為低個位數，對營收影響不大，市場預估中國廠商在半導體禁令漸趨嚴格的狀況下，將加速發展RISC-V架構以提升自主性

根據金融時報的報導，ARM計畫禁止向中國IC設計公司平頭哥提供最新技術 (Neoverse V系列架構授權)。平頭哥是中國伺服器與人工智慧晶片領域的重要IC設計公司之一。這項授權禁令將進一步限制中國在人工智慧領域的發展。平頭哥目前在台積電生產14nm以下製程的晶片，但預計營收占比很低，對其營收影響不大。市場預測，在中國面臨半導體進口限制越來越嚴格的情況下，中國IC設計公司將會更加積極地開發RISC-V架構，以提高自主性。

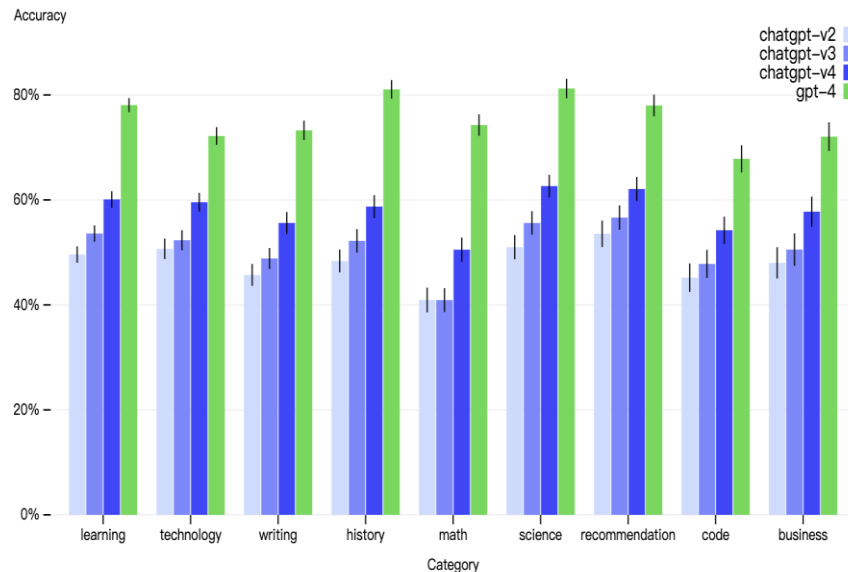
# GPT-4於2023年推出，進一步優化表現並提供多模型

- 2023年3月中旬，OpenAI推出了GPT-4，這款升級版本支援多模型應用，改善了GPT-3在生成Token數量不足的問題，並優化了回答能力。此外，GPT-4還支援圖像辨識等功能，進一步提高了其可用性。然而，GPT-4仍然存在一些局限性，例如資料庫尚未完全更新，回答可能不夠精確，以及潛在的偏見問題。儘管如此，GPT-4在多個應用領域仍展示出強大的潛力。

圖9、GPT-4相較其他語言模型改善幅度



Internal factual eval by category



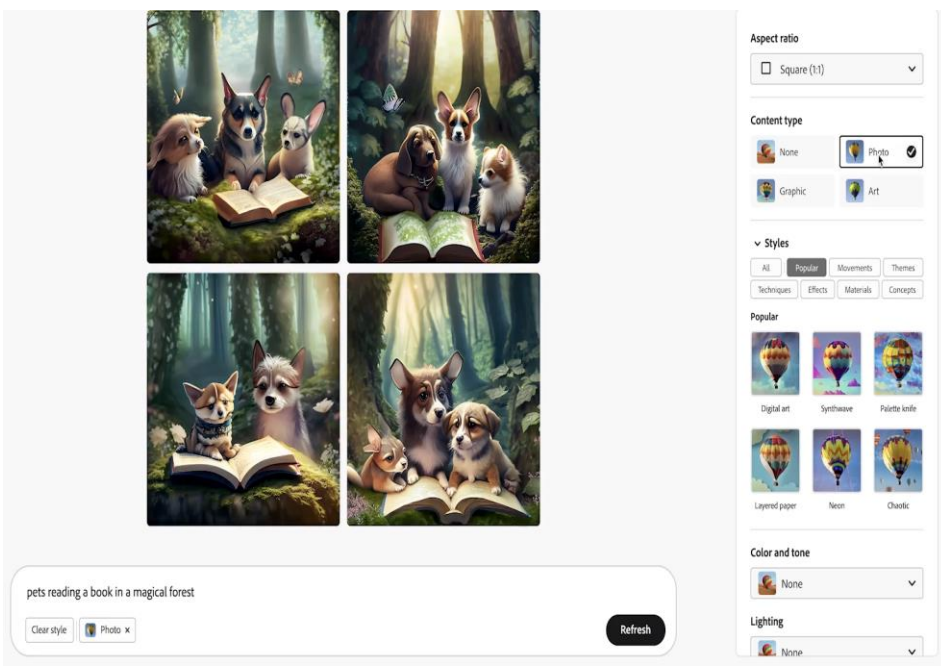
# 文字轉繪圖應用逐漸成熟

- Midjourney為AI繪圖軟體，可透過下關鍵字指令讓AI繪製圖片，並可來回修改得到理想圖片，例如左圖為科技產品之概念圖，指令為「Cyberpunk」、「bike」、「Future Style」、「Electric」
- Adobe也針對生成式AI市場推出Adobe Firefly，發展圖像生成與字型效果等應用，未來將大幅增進圖像市場的生產效率

圖10、使用AI繪圖軟體Midjourney繪製科幻載具



圖11、Adobe Firefly操作介面



# 微軟、谷歌等科技巨頭積極把AI導入產品線

- 微軟在3月中旬宣布推出Microsoft 365 Copilot，將AI應用整合到Office產品線中。Google也針對ChatGPT推出了基於LaMDA模型訓練的對話機器人Bard。根據內部訪談，ChatGPT與Bard在對話能力上差異不大，但ChatGPT具有更多的資料量和Coding能力，並在支援國際語言方面表現較好。目前Google正集中資源積極發展Bard以及其他AI專案，預計將其整合到Google Workspace和智慧家庭終端設備中。
- 通過導入AI技術，不僅可以優化搜尋路徑，還可以結合其他需要對話的場景應用，例如商業文件應用、智慧家庭裝置或車載系統的操作優化。這將使使用者能夠以更高效的方式處理各種任務，提升生產力和使用體驗。

圖12、Microsoft 365 Copilot示意圖

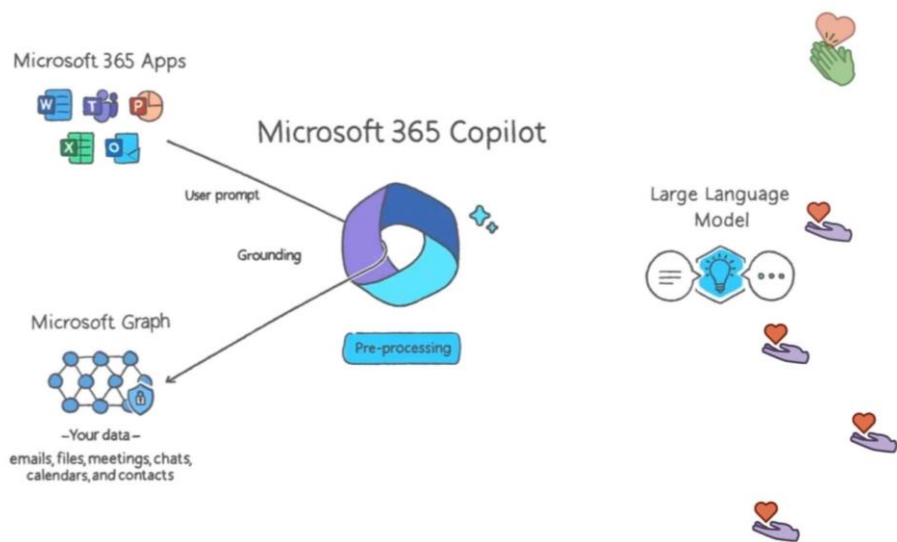
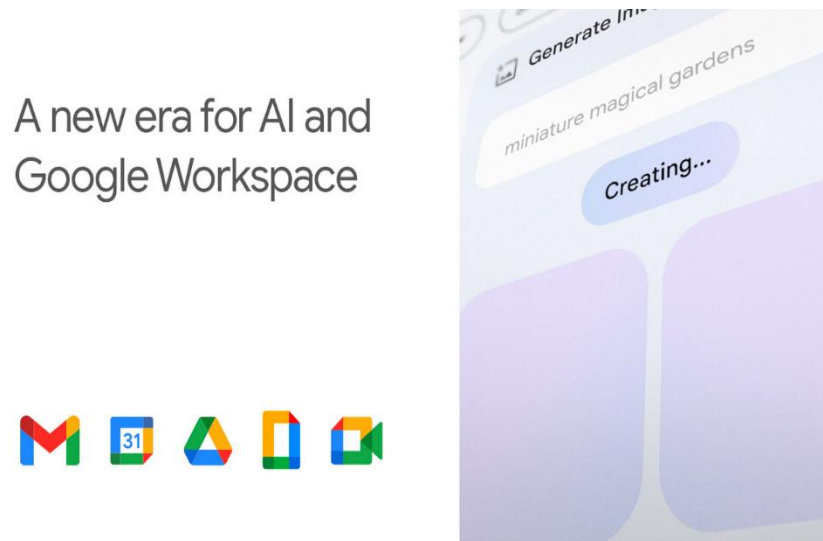


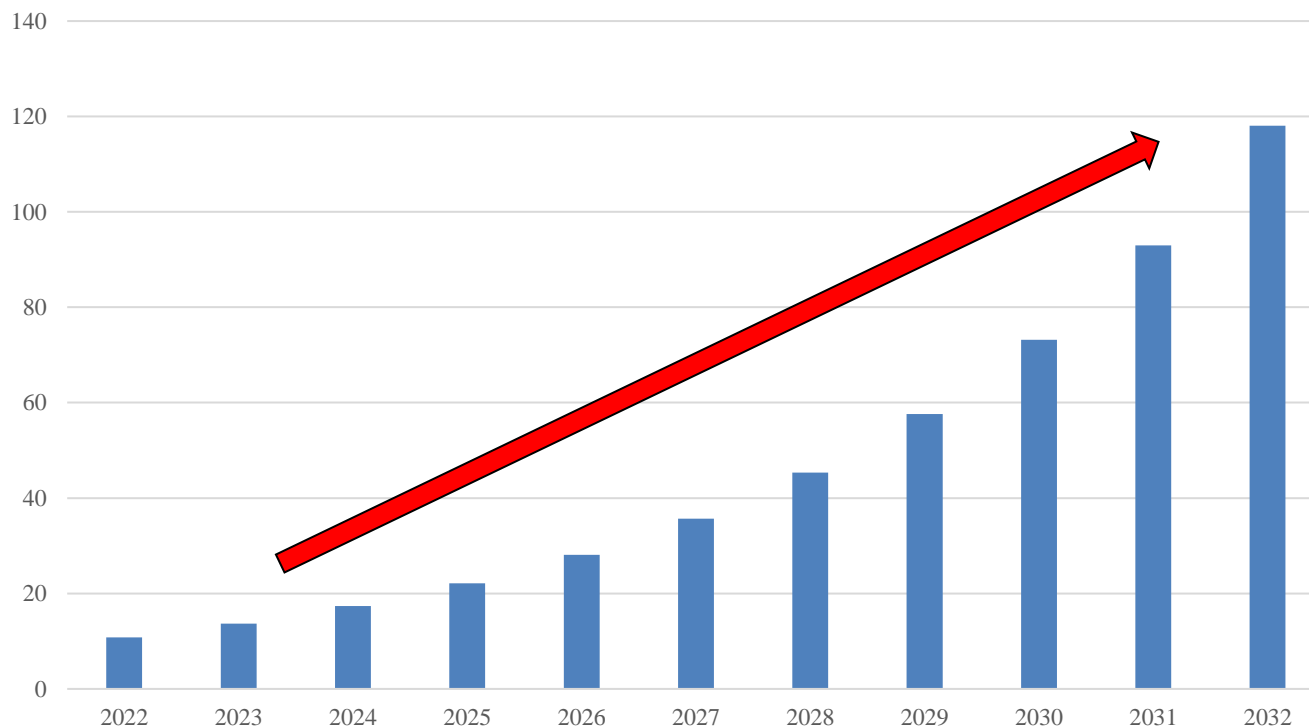
圖13、Google Workspace AI示意圖



# 2032年生成式AI市場規模將突破1000億美元

- 在眾多AI技術的推動下，市場將能更有效率地生產影片、圖像、文字和程式等內容。預計從2022年至2032年，整體市場規模將以27%的CAGR高速成長。
- 目前，北美地區以及媒體與娛樂部門在生成式AI市場佔比較高。然而，未來亞太地區、商業和金融領域的市場規模增長速度將超過平均水準，成為重要的市場發展趨勢。AI技術的應用將在全球範圍內逐漸擴展，並在各個行業和地區產生深遠的影響。

圖14、全球生成式AI市場規模(單位：10億美元)





# ChatGPT 對A100晶片需求推估將達2-3萬片

- 根據推算，ChatGPT在每日1300萬名活躍用戶的假設中，約需要2.6萬顆Nvidia A100晶片，未來在導入Bing之後，使用量將倍增
- Google瀏覽器每天有85億次瀏覽量/2.65億每日活躍用戶，若未來部分轉往使用AI對話應用，使用量將大幅成長

表2、Nvidia 晶片出貨量出貨量推估

	ChatGPT	BingChat
每日活躍用戶(百萬)	13	100
每日搜尋次數	10	10
每次搜尋處理字數	50	50
每日處理字數(百萬)	6500	50000
每顆 A100 每秒推論字數	3	3
需要 A100 顆數	26,331	202,546
假設H100/A100算力比例	9	9
每顆 H100 每秒推論字數*	27	27
需要 H100 顆數	2,925	22,504

\*根據NVIDIA 2023/3揭露，H100在LLM AI訓練效能最高為A100之9倍



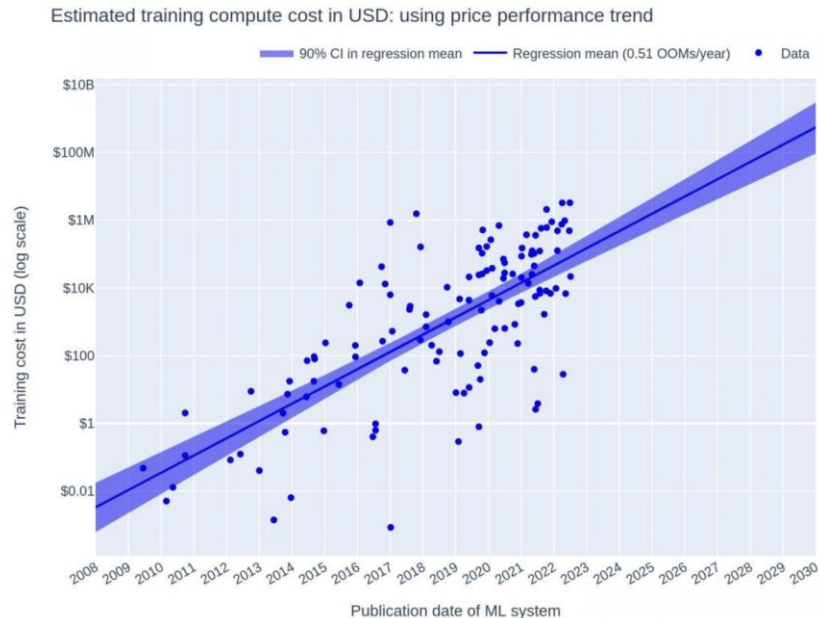
# AI模型訓練總成本高昂，科技巨頭發展具優勢

- Token為一段字詞，例如「ChatGPT is Great!」包含Chat、G、PT、is、Great、！六個Token，一般來說750個英文單字對應1000個Token，700個中文字對應1000個Token
- OpenAI收費模式除了每月20美元的ChatGPT Plus之外，也開放付費使用API，使用GPT3.5 turbo模型每1000個Token收費0.002美元，GPT-4輸入/生成收費0.03/0.06美元
- 在硬體方面，隨晶片運算能力提升，單位算力所需付出硬體與能源成本持續下降，但因訓練資料量增加，總成本仍將走升，雲端或科技業大公司較有能力投資軟硬體發展AI，提供串連API給一般企業發展自身業務

表3、各類語言模型訓練參數與Token數

Model	廠商	Parameters (Billion)	Training Tokens (Billion)
GPT-3	Open AI	175	300
LaMDA	Google	137	168
Gopher	DeepMind	280	300
MT-NLG	NVIDIA/Microsoft	530	270
LLaMA	META	65	1400

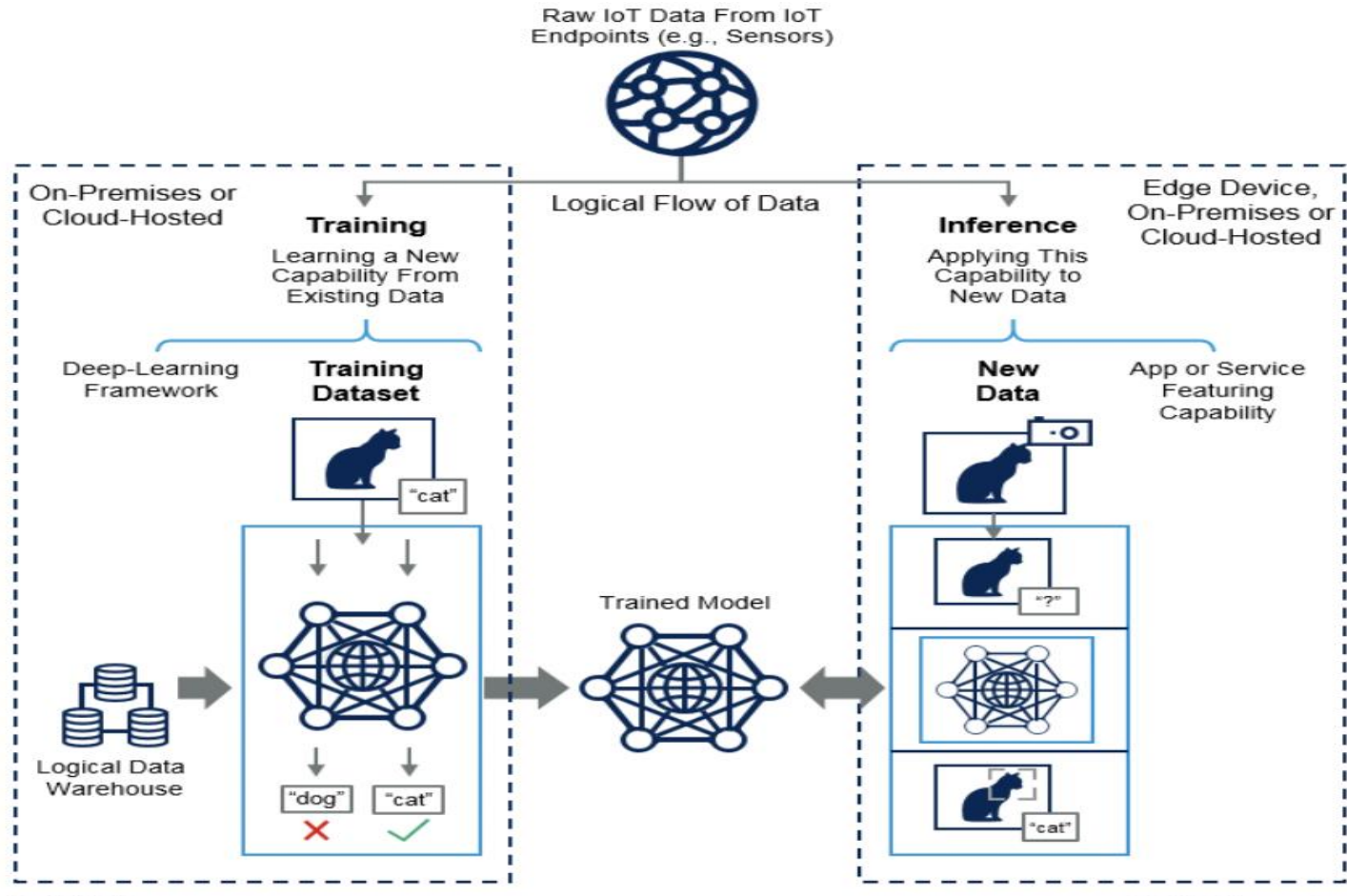
圖15、訓練AI成本趨勢



# AI晶片與相關供應鏈

# AI晶片分為「訓練」與「推論」兩類

圖16、AI訓練與推論示意圖



# AI晶片市場NVIDIA GPU普遍用於雲端AI訓練

- NVIDIA作為AI晶片市場的龍頭，得益於其早期投入AI發展以及在GPU軟硬體技術和生態系統方面領先競爭對手。在企業端GPU市場，NVIDIA的市佔率超過90%，其次是AMD。
- 目前，市面上主流的AI晶片為NVIDIA A100。在2022年，NVIDIA推出了下一代產品，基於Hopper架構的GPU H100，預計將繼續保持在市場上的主流地位。

圖17、全球企業端GPU市占率

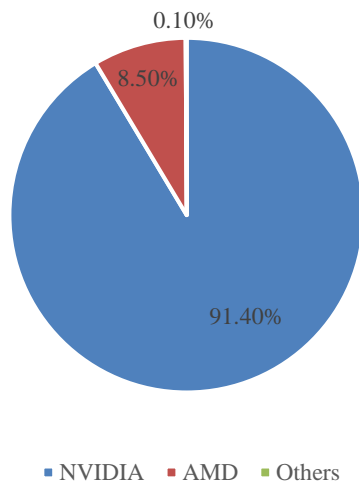


表4、NVIDIA AI晶片規格表

	H100	A100	Tesla V100
推出時程	2022	2020	2017
製程(nm)	TSMC 4	TSMC 7	12nm FFT
FP 16 Tensor	1000 TFLOPS	312 TFLOPS	125 TFLOPS
TDP(w)	350/700	250/300/400	250/300/450
GFLOPS/\$	30	21	13
GFLOPS/watt	1429	780	417
售價(\$)	>30000	15000-20000	10000

# ASIC應用於特定場景，用於AI推論具有優勢

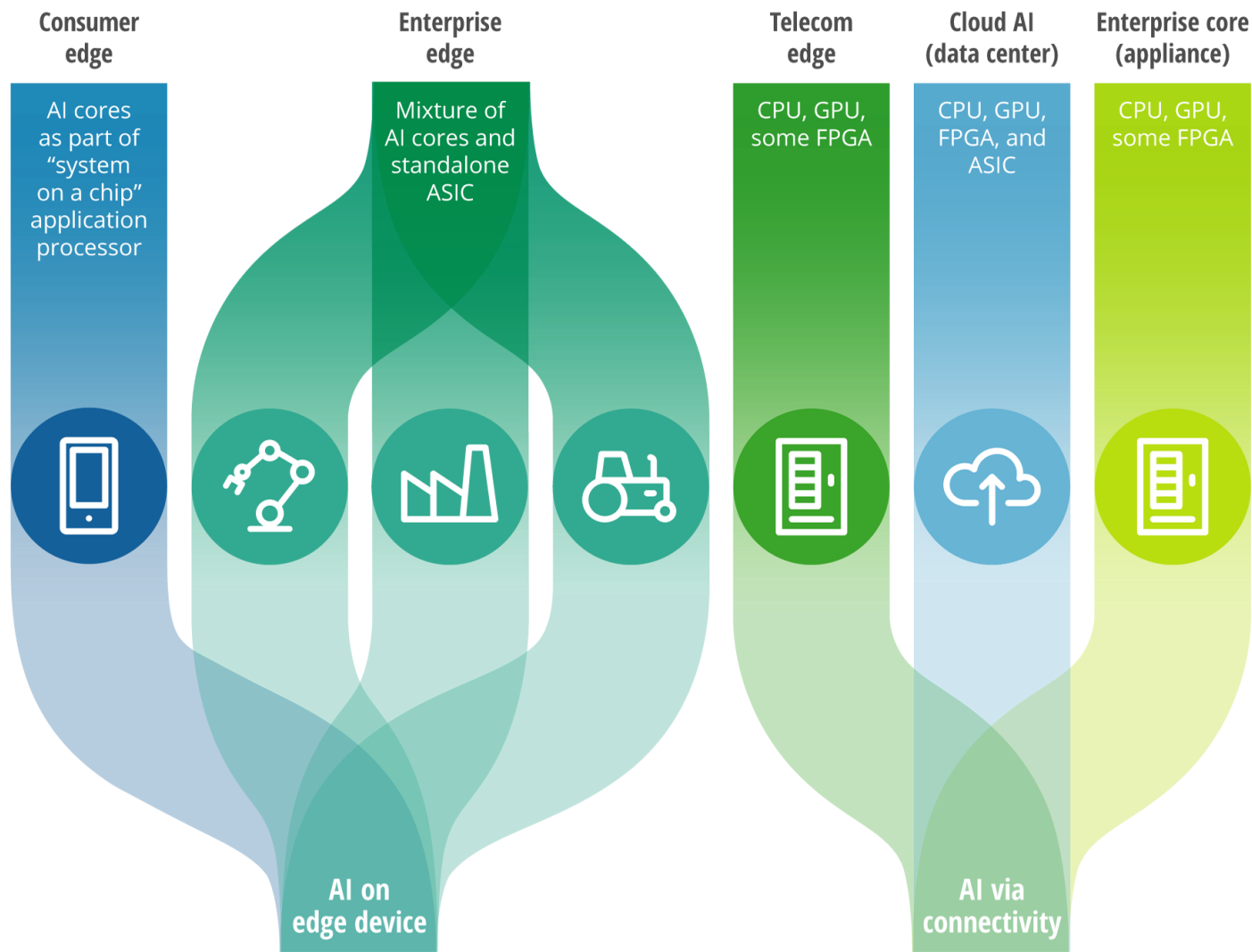
- ASIC 的「場景優化」大概可分為幾個點
  - 優化特定深度學習框架。Ex：PyTorch、TensorFlow
  - 優化特定應用。Ex：影像辨識、轉碼、虛擬主機
  - 優化推論能耗，以供邊緣運算的AI推論工作

表5、各類晶片於AI使用對比表

晶片種類	訓練		推論		通用性	推論準確度	用途
	效率	速率	效率	速率			
CPU	以CPU為一倍				非常高	98-99.7%	通用性最高，但針對AI應用效率不佳
GPU	10-100	10-10000	1-10	1-100	高	98-99.7%	通用性高，可大規模應用於AI訓練，但能耗高、晶片面積大，單顆GPU成本較高
FPGA	-	-	10-100	10-100	中	95-99%	效率較高，具可針對應用場景修改，但成本較高
ASIC	100-1000	10-1000	100-1000	10-1000	低	90-98%	針對特定應用效率高，量產後成本低，但通用性最低

# 不同晶片可應用於不同的AI使用場域

圖18、Cloud與Edge AI在不同場域應用

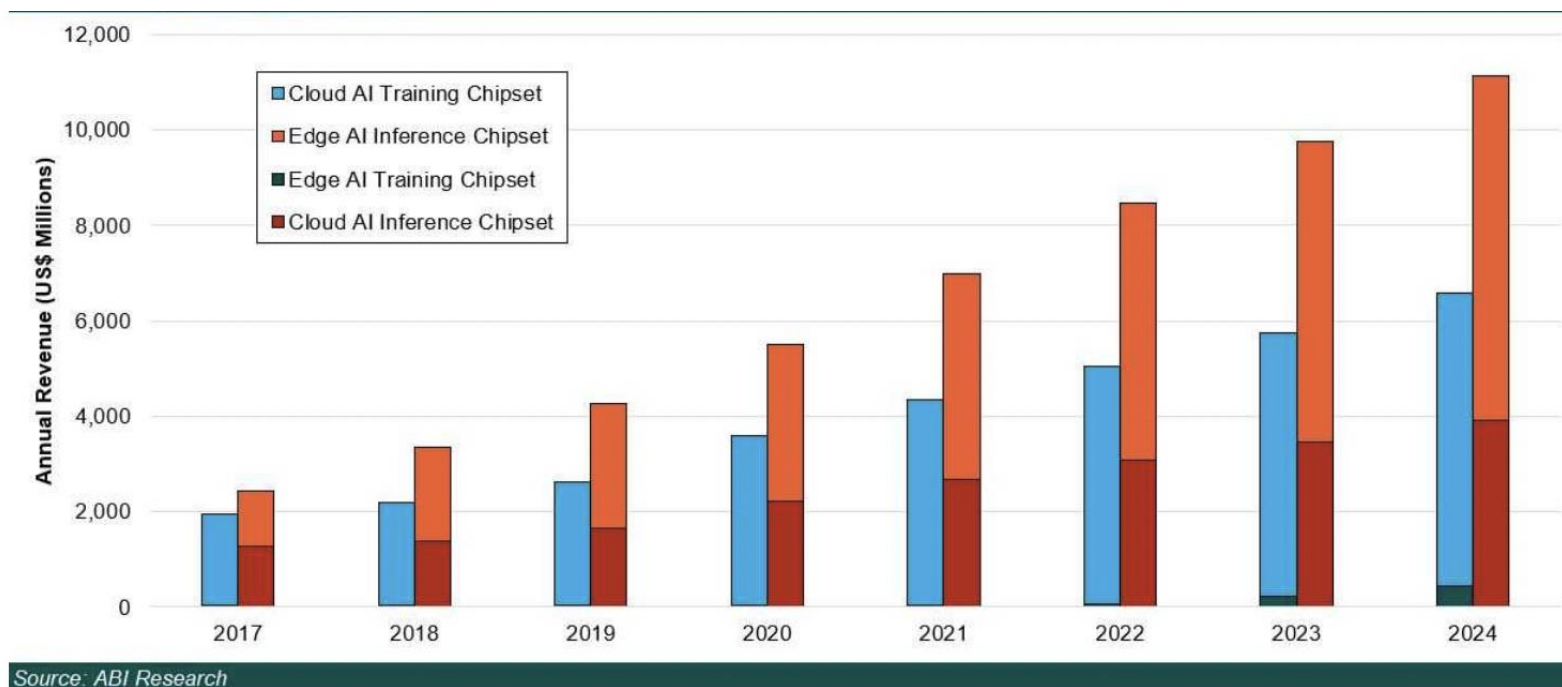




# Cloud AI主要用於訓練，未來Edge AI市場占比提升

- 雲端AI晶片相對於邊緣AI晶片，通常具有更高的運算能力、功耗、面積和價格。目前雲端訓練AI主要依賴GPU（如NVIDIA、AMD）和ASIC（如TPU）等晶片。
- 隨著邊緣AI應用的數量不斷增加，邊緣訓練和推理的需求將逐步上升。在這一趨勢下，ASIC和FPGA等晶片的需求將大幅增長。

圖19、全球AI晶片市場規模(百萬美元)



# 雲端服務業者積極布局晶片相關投資

表6、全球雲端業者AI晶片佈局

	名稱	晶片類型	用途	公司布局
Amazon	Nitro	ASIC	數據串流	除了使用A100做機器學習之外，也有推出自研AI晶片，目前正與新創公司Hugging Face合作，期望未來將AI導入在AWS產品線
	Graviton	CPU	AWS CPU	
	Trainium	ASIC	AI訓練	
	Inferentia	ASIC	AI推論	
Google	TPU	ASIC	AI訓練/推論	未來將多項AI技術導入現有產品線，包含Bard(LaMDA)與PaLM等，以建立新的資訊互動方式
Microsoft	NA	CPU	Azure CPU	加大投資OpenAI，未來將ChatGPT導入自家產品線，以to B為主，並自行研發AI晶片
	NA	ASIC?	AI推論?	
Meta	-	-	-	目前放棄自研晶片，採用NVIDIA GPU，使用AI發展社群、元宇宙應用
Alibaba	含光800	ASIC	AI推論	-
	倚天710	CPU	阿里雲CPU	
Tencent	紫霄	ASIC	AI推論	-
百度	崑崙芯2	ASIC	AI訓練/推論	3/16推出生成式AI「文心一言」，AI訓練將同時使用A100與寒武紀思元590晶片*

# 台積電製程優勢，主要AI晶片多投片於台積電

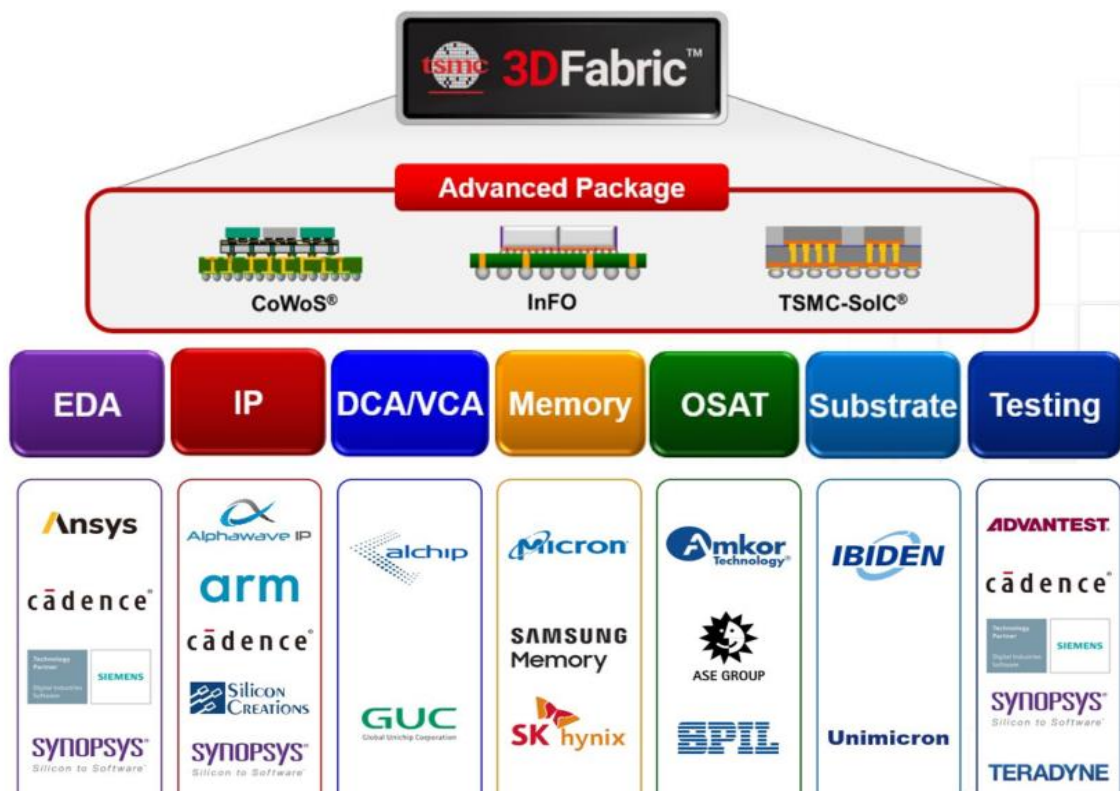
表7、全球AI晶片規格表

廠商	晶片	架構	推出時間	製程	記憶體	TDP(W)	備註
NVIDIA	A100	Ampere	2020	TSMC 7nm	40/80GB HBM2e	250/300/400	HBM2由三星提供
	H100	Hopper	2022	TSMC 4nm	80GB HBM3	350/700	HBM3由SK-Hynix提供
AMD	MI250	CDNA2	2021	TSMC 6nm	128GB HBM2e	500-560	Chiplet
	MI300	CDNA3	2023	TSMC 5/6nm	128GB HBM3	600+	3D Chiplet APU
Intel	Ponte Vecchio	Xe-HPC	2022	TSMC5/7nm +Intel 7nm	128GB HBM2e	600	下一代晶片遞延至2024-25年
	Gaudi2	Gaudi	2022	TSMC 7nm	96GB HBM2e	600	
Google	TPUv4	-	-	7nm	32GB HBM	175	
Amazon	Trainium/ Inferentia	-	-	TSMC 7nm	32GB HBM/HBM2e	-	下一代晶片可能於2025-26推出
百度	崑崙芯2/3	-	2021/2024	三星or中芯?	32GB/?	-	第一代為三星14nm製程

# 台積電先進封裝結合3D Fabric聯盟成員一同發展AI

- 台積電推出3D Fabric先進封裝平台，整合旗下CoWoS與InFO等封裝技術以及SoIC堆疊技術，並預計在2026年將先進封裝產能提升至2018年的20倍以上，隨AI/HPC/自動駕駛等應用發展，相關供應鏈將隨之成長

圖20、台積電3D Fabric聯盟



# 隨高速運算需求不斷提升，傳輸介面速率重要性上升

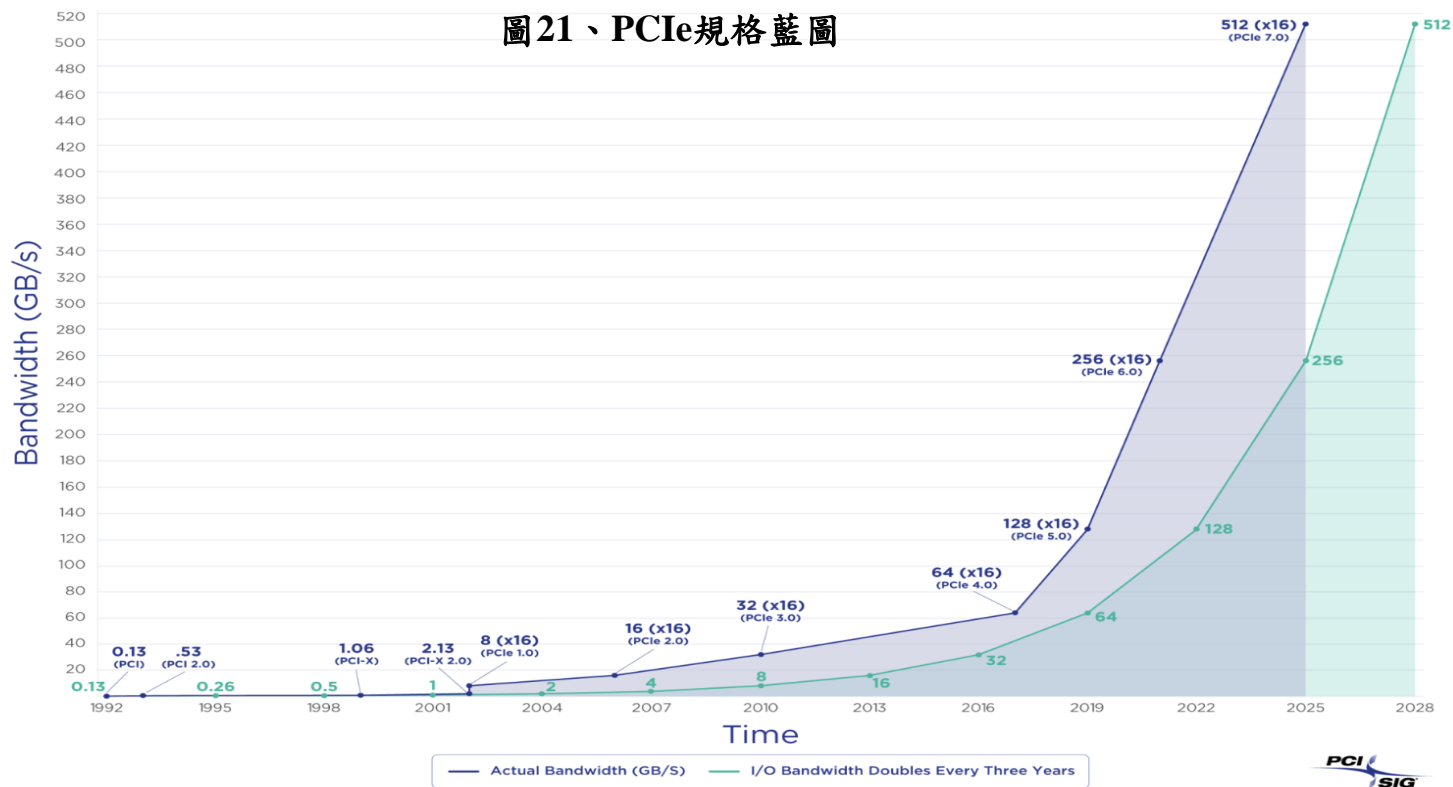
- AI晶片追求運算速度的需求，讓硬體廠商紛紛建立各種專門用於提高AI運算速度的技術。2014年，Nvidia因應當時的PCIe 3（頻寬32GB/s）無法滿足AI運算需求，推出了NVLink GPU互連技術。於2016年首次搭載在Tesla P100晶片上。為了進一步提高運算速度，Nvidia於2022年推出了NVLink C2C，以實現晶片間的互連功能。
- 與此同時，AMD也推出了自家的互連架構 - Infinity Fabric。該技術在AMD的產品線中不斷升級，最新的MI300晶片上將搭載第四代Infinity Fabric技術。

表8、傳輸介面比較表

	PCIe 6.0	PCIe7.0	AMD Infinity Fabric 3	AMD Infinity Fabric 4	Nvidia NVLink 3	Nvidia NVLink 4	Nvidia NVLink C2C
頻寬(GB/S)	16	32	100	-	50	50	-
連線數	16	16	8	-	12	18	-
總頻寬(GB/s)	256	512	800	-	600	900	900
晶片	-	-	MI250	MI300	A100	H100	Grace-Hopper Superchip

# AI高速運算刺激PCIe加速發展

- 根據產業標準組織PCI-SIG於2022年的宣布，預計將於2025年推出PCIe 7.0規格。該規格將延續每三年頻寬加倍的趨勢，為各種高速計算需求提供更強大的支援。
- PCIe 7.0的推出著重於對高速數據傳輸和低延遲有著極高的要求的應用，包括800G乙太網路、人工智能和機器學習（AI/ML）、雲端運算和量子運算等領域。

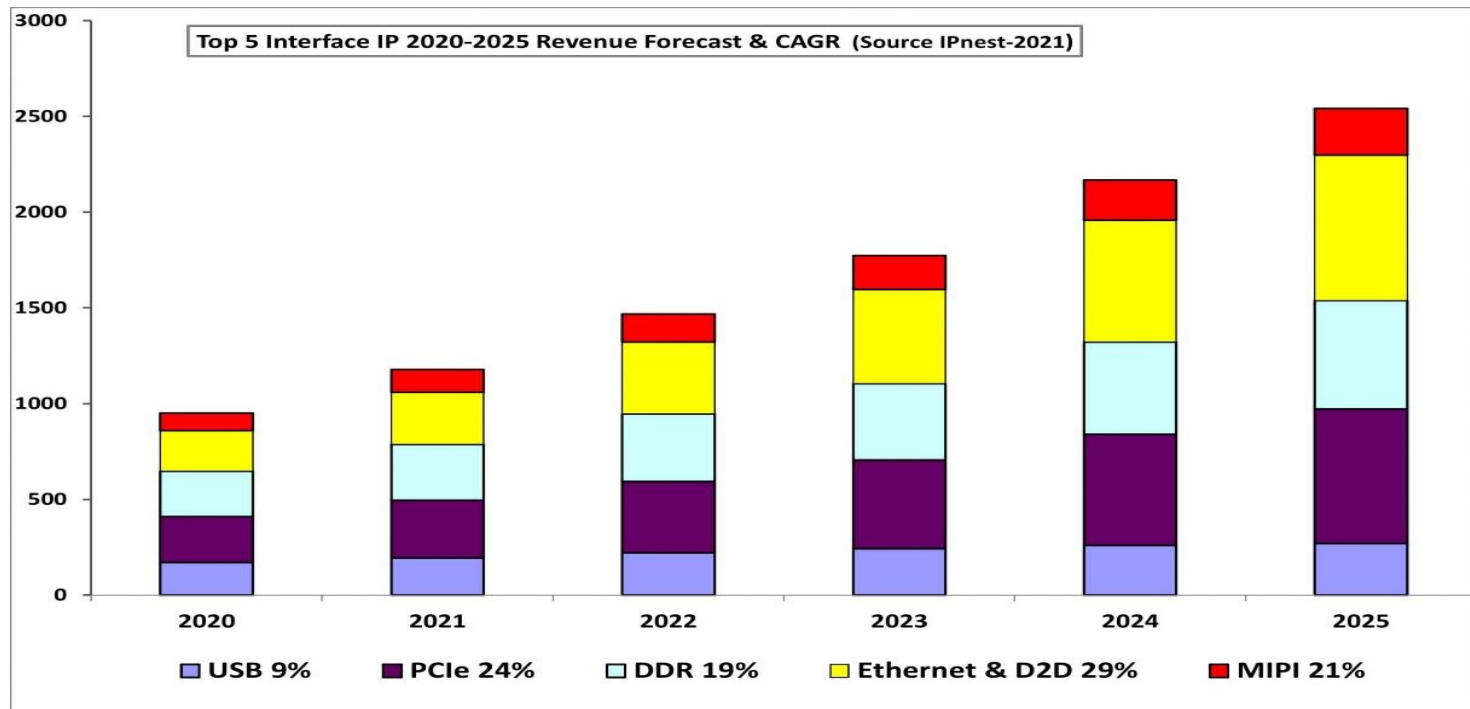




# AI GPU、ASIC 帶動周邊I/O傳輸介面同步升級

- AI晶片大多會採用如NVLink等專用傳輸介面做IC間的溝通，但因也需要與其他晶片與硬體傳輸，將帶動SSD、Memory、網路接口等裝置的傳輸介面加快升級腳步，減少AI運算時的頻寬瓶頸。

圖22、2021-2025高速傳輸介面IP營收(單位:百萬美元)



# CPU+GPU混合系統提高運算表現能力

- 儘管在GPU-GPU互連技術速度已經相當高，但在GPU-CPU間的主流互連技術仍然是PCIe Gen5。在涉及與CPU互動的場景中，由於頻寬受限，可能會出現性能瓶頸。因此AMD在其最新的AI GPU MI300中採用將CPU和GPU以chiplet（晶片組）的方式進行3D封裝，以提升晶片間傳輸能力。

圖23、AMD 3D chiplet示意圖

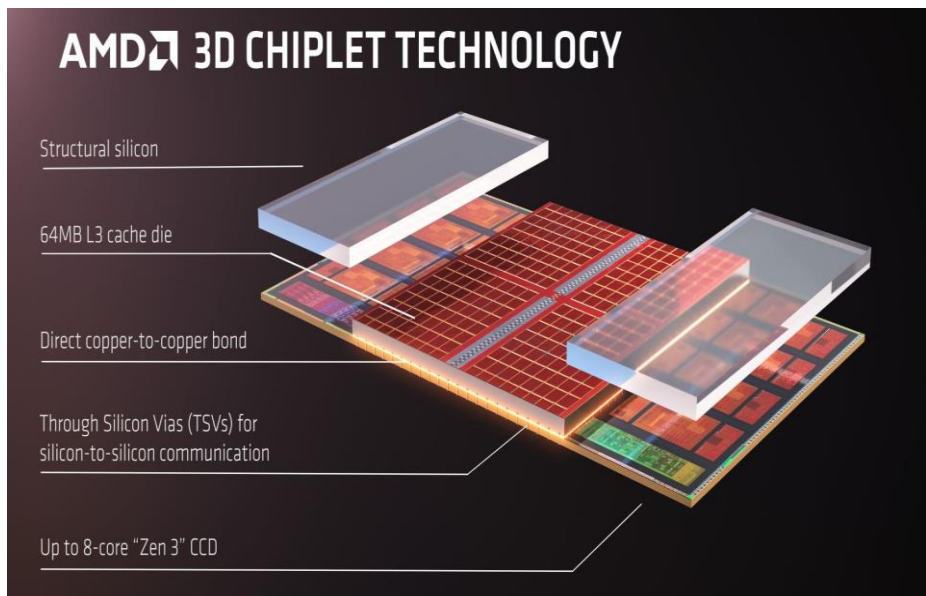
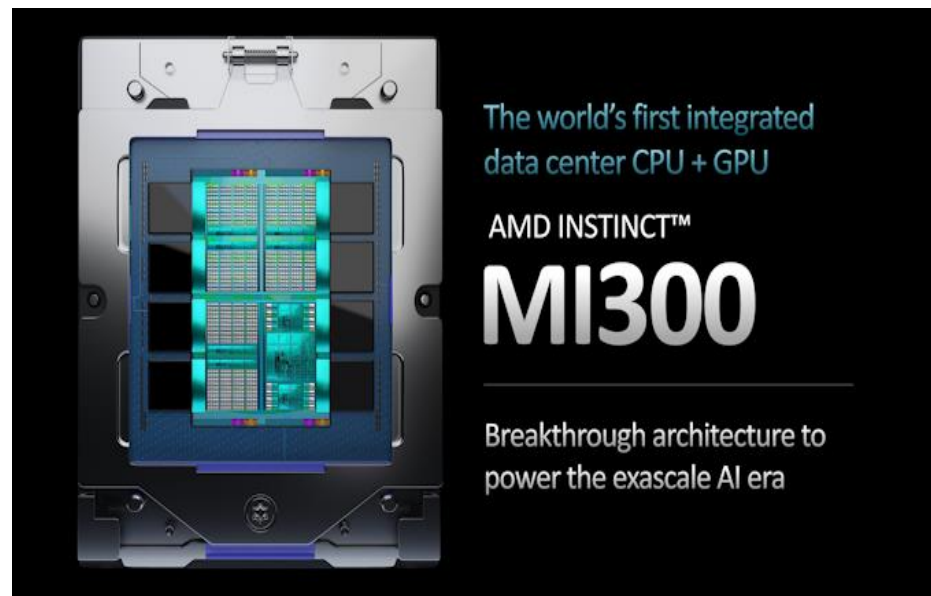
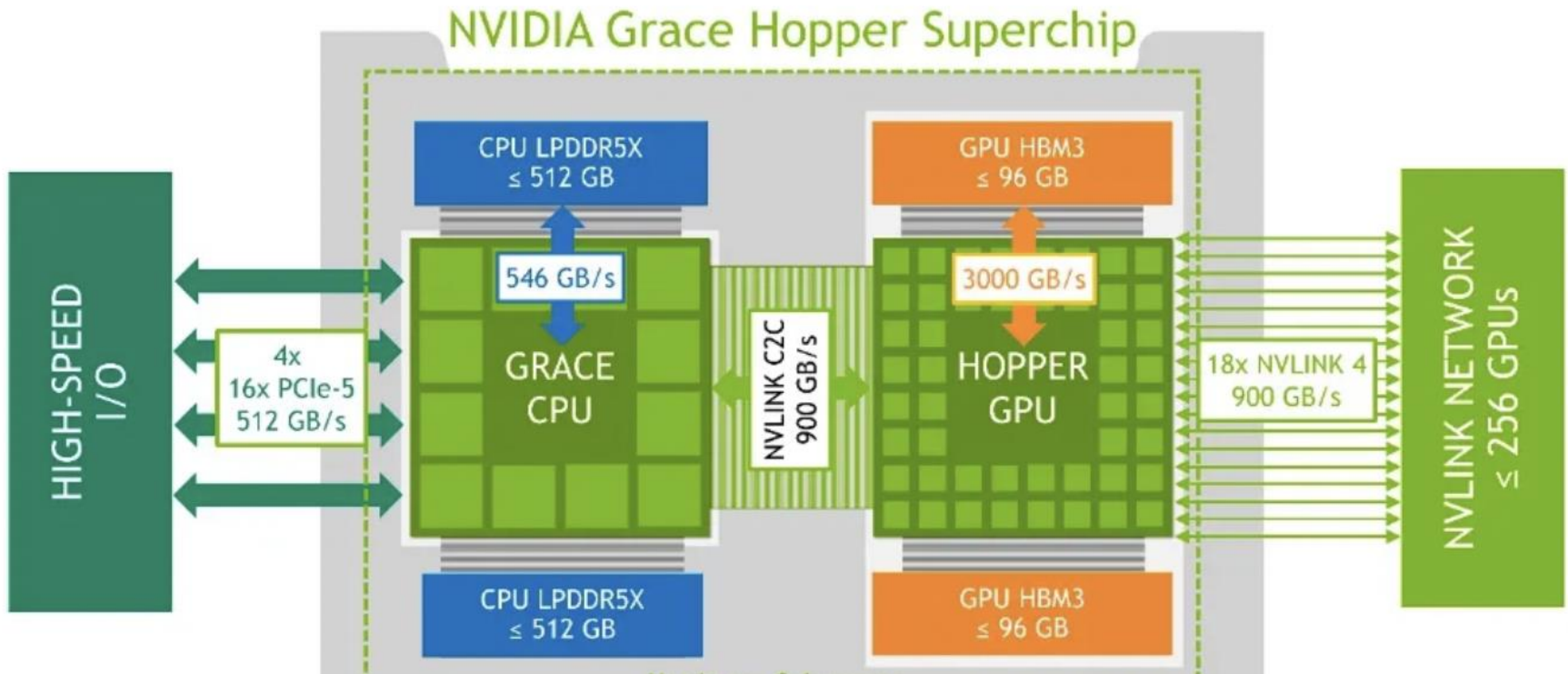


圖24、AMD MI300



# NVIDIA GRACE Hopper Superchip採用NVLink C2C

圖25、NVIDIA Grace Hopper Superchip結構圖



# 高寬頻記憶體(HBM)主要使用於AI晶片

- 高頻寬記憶體(High Bandwidth Memory, HBM)為基於3D堆疊工藝的高效能DRAM，適用於高記憶體頻寬需求的應用場合，例如AI、機器學習、大數據、高速運算與自動駕駛等，以GDDR5與HBM對比來看，HBM相對具有高頻寬、節省空間與低功耗之優勢
- 根據研調機構預估，隨AI/HPC市場需求成長，將推升HBM市場規模2026年達51億美金，2021-26年CAGR達30%以上

圖26、GDDR5 VS HBM

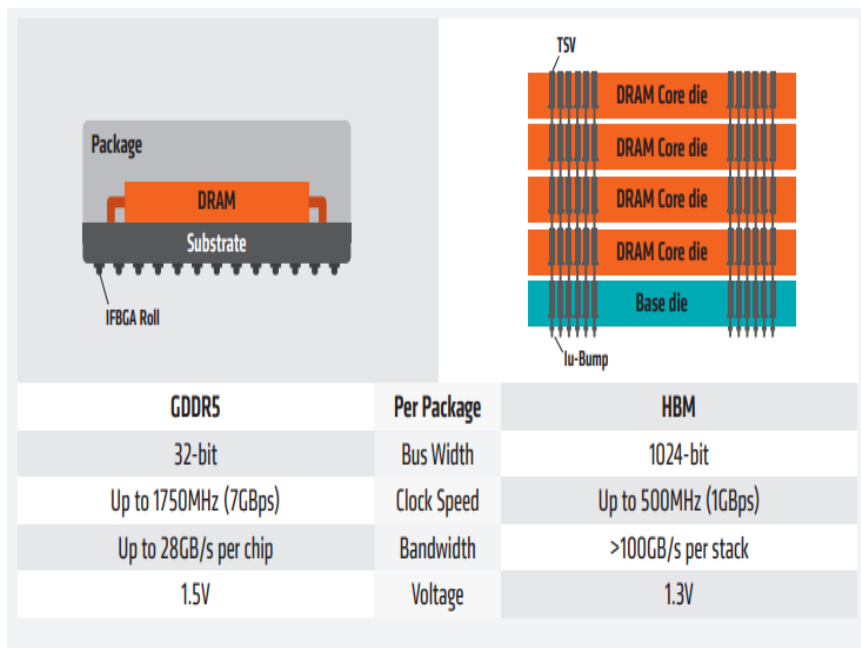
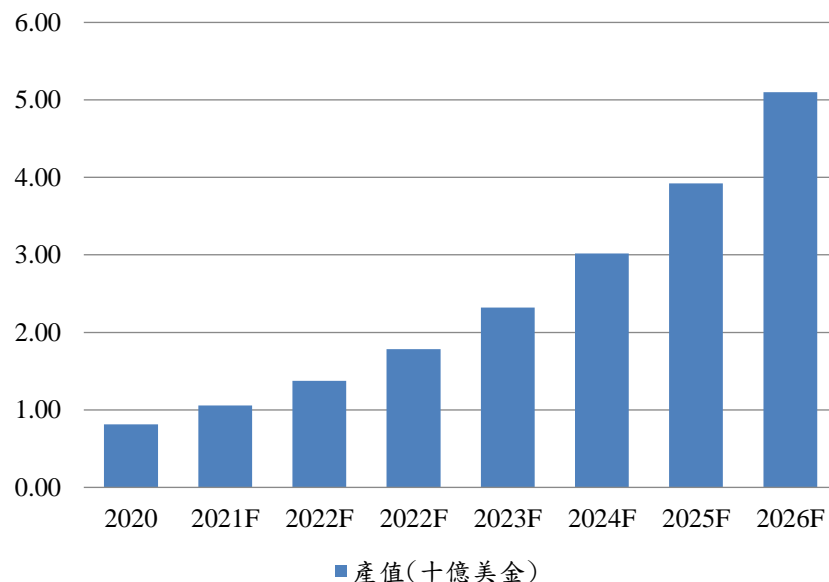


圖27、HBM市場規模(單位:十億美金)



# AI晶片將大幅推升HBM用量

- 根據Trendforce報告，2021年HBM位元需求僅佔整體DRAM不到1%，ChatGPT帶動AI熱潮，研究員預估對整體DRAM需求佔比仍然為低個位數，但需求成長速度優於整體，受惠廠商包含韓系與美國記憶體廠，以及相關IP與2.5/3D封裝供應商

表9、ChatGPT對HBM用量預估表

	ChatGPT	Bing Chat
A100 需求量(顆)	26,331	101,273
A100 HBM容量(GB) per Stack	16	16
Stack	5	5
單顆A100 晶片 HBM容量(GB)	80	80
A100 HBM總需求(GB)	2,106,480	8,101,856
佔總DRAM比例	<1%	1~2%

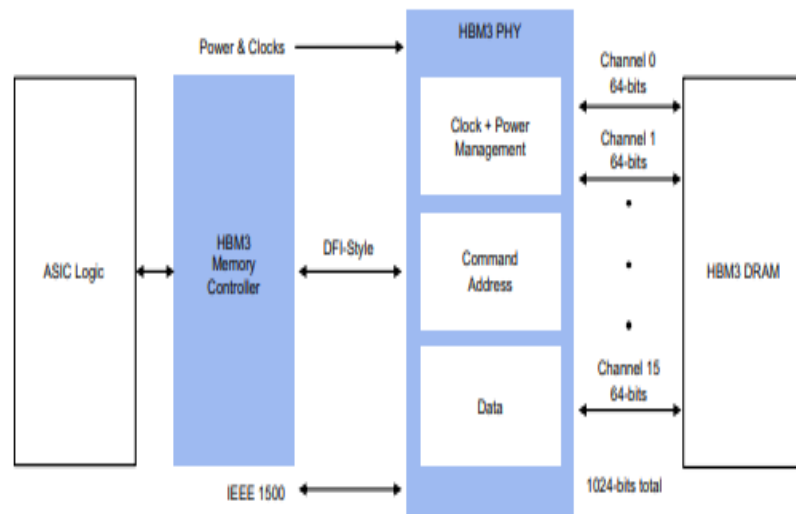
# 未來AI晶片將採用HBM3，台廠以ASIC/IP產業受惠較大

- 現有AI晶片記憶體主要採用HBM2e技術，下一代將採用HBM3技術，在頻寬與傳輸速率上近一步提升
- HBM記憶體需要控制器(Controller)與實體層(PHY)控管資料傳輸，創意與Rambus與SK-Hynix合作，皆可提供HBM3所需IP

圖28、HBM技術演進



圖29、HBM3傳輸介面示意圖





# 伺服器市場

# 2023 年伺服器出貨量低個位數衰退，下半年有望強勁回升

- 研究員預估23Q1 伺服器出貨量受去庫存影響，QoQ -10%，23Q2 拉貨回升，出貨量 QoQ +6.5%，但前面2季度均大幅季減，以拉貨量來看回升有限
- 預估23Q3 受惠 CSP業者重啟拉貨、新平台放量以及中國急單可能出現的狀況，預估 QoQ +13%、YoY 轉正，下半年拉貨動能回升

表10、研調機構及福邦伺服器出貨預估

		2017	2018	2019	2020	2021	1Q22	2Q22	3Q22	4Q22	2022	1Q23(F)	2Q23(F)	3Q23(F)	4Q23(F)	2023(F)	2024(F)
TrendForce	YoY						-7.9%	15.6%	6.5%	-5.7%	5.1%					1.9%	1.9%
Gartner	YoY										8.3%					6.9%	7.1%
Digitime	Shipment	13,910	15,238	15,060	16,255	17,008	4,406	4,551	4,750	4,415	18,122	3,966				18,901	20,135
	YoY	14.3%	9.5%	-1.2%	7.9%	4.6%	16.5%	10.2%	6.0%	-4.3%	6.5%	-10.0%				4.3%	6.5%
	QoQ						-4.5%	3.3%	4.4%	-7.1%		-10.2%					
IDC	Shipment	10,241	11,849	11,800	12,666	13,539	3,349	3,853	3,965	3,998	15,165	3,455	3,797	4,000	4,235	15,487	
	YoY	7.2%	15.7%	-0.4%	7.3%	6.9%	14.3%	14.3%	13.1%	7.1%	12.0%	3.2%	-1.5%	0.9%	5.9%	2.1%	
	QoQ						-10.3%	15.1%	2.9%	0.8%		-13.6%	9.9%	5.4%	5.9%		
	白牌YoY	48.0%	30.5%	11.8%	26.9%	11.8%					21.7%						
	品牌YoY	-0.7%	11.4%	-4.5%	-0.4%	4.4%					3.8%						
福邦	Shipment	13,910	15,238	15,060	16,255	17,008	4,406	4,551	4,750	4,415	18,122	3,966	4,224	4,773	4,964	17,926	19,629
	YoY	14.3%	9.5%	-1.2%	7.9%	4.6%	16.5%	10.2%	6.0%	-4.3%	6.5%	-10.0%	-7.2%	0.5%	12.4%	-1.1%	9.5%
	QoQ						-4.5%	3.3%	4.4%	-7.1%	0.0%	-10.2%	6.5%	13.0%	4.0%		

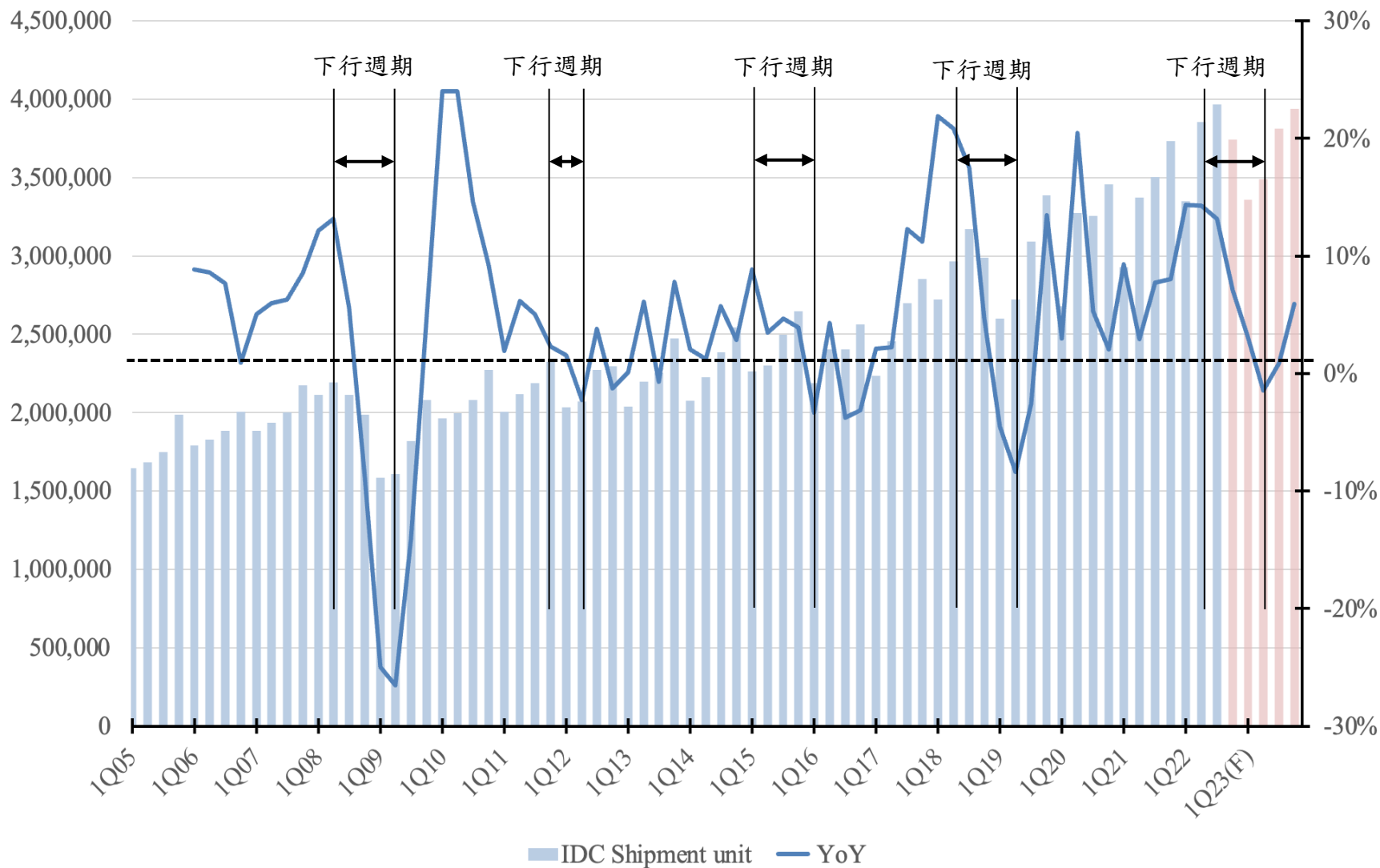
1. 單位：千台

2. Digitime 預估以主機板為基準，IDC 以機箱為基準。

3. 福邦預估以 Digitime 為基準預估

# 自高點滑落算起，過往伺服器下行週期大約維持維持4-5 季

圖30、伺服器出貨&年增率



# 2022 AI Server 出貨約 50-70 萬台，滲透率約 3.5-4%

- AI Server 如果以狹隘的定義（A100/H100才算），2022年出貨量在10萬台以內。較寬鬆的 AI Server 定義出貨量約 50-70 萬台，滲透率約在 3.5-4%。AI Server 的成長，將貢獻 2024 年伺服器出貨量 0.4-0.5%、產值約 2-4%。

圖31、中低階 AI 加速卡不一定會製作成 AI 專用 Server，可在普通 Server 的 PCIe 槽插上 2-8 張 AI 加速卡，升級為 AI Server，總量較難估計。

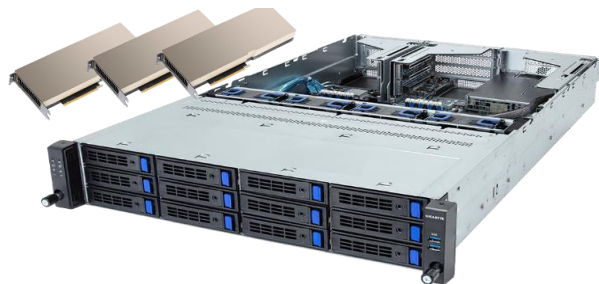


圖32、旗艦卡 A100 大多 8 張組成一台 AI 專用 Server



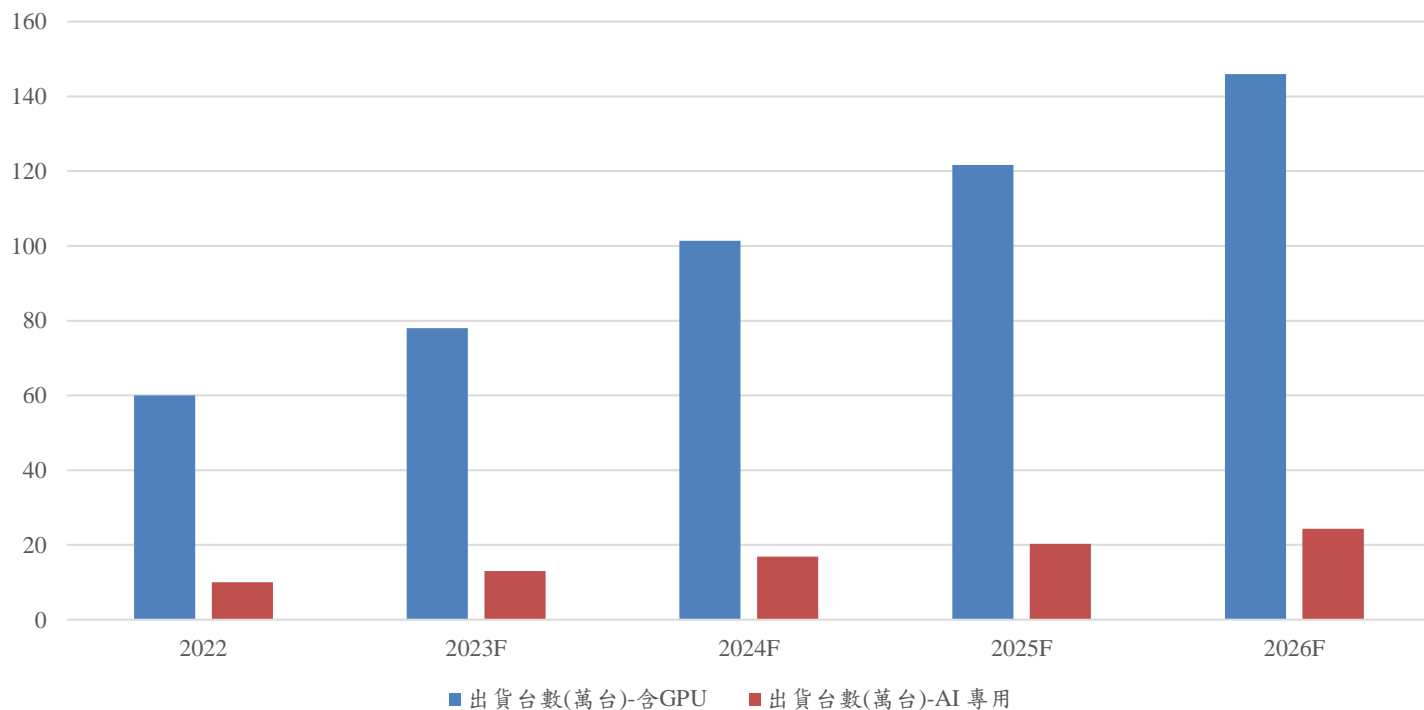
表11、Nvidia AI GPU出貨量、AI Server出貨量推估

位階	AI 加速卡型號	AI卡ASP	AI卡出貨量	對應伺服器數量	伺服器單價
旗艦卡	A100	18-20 K	50-70萬	7-9萬台	200K
中低階	A2/A10/A30/A40	2/4/5/6 K	140-160萬	50-70萬台	30-70K
通用伺服器	無			1400萬台	8-10K

# 大廠加碼投入AI，預估2024年AI伺服器出貨超過100萬台

- AI伺服器是專門用於訓練AI模型的伺服器，通常部署在大型資料中心。根據研究員的估計，2022年全球AI伺服器出貨量將達到60萬台。隨著各大科技公司加速研發AI技術，預估到2024年，AI伺服器出貨量將超過100萬台。
- 隨著AI技術的普及和發展，對AI伺服器的需求將持續增加，帶動相關產業的快速成長。將為硬體製造商、資料中心營運商和AI應用開發商等相關企業創造更多商機。

圖33、全球AI伺服器出貨預估



# AI晶片所需散熱能力逐步提升，散熱模組廠將受惠

- 隨著AI GPU的TDP（熱設計功耗）普遍超過500W，並預計未來將朝向1000W的方向發展，解決AI和伺服器晶片日益增加的功耗問題變得至關重要。更先進的散熱技術，如大型散熱模組、大體積氣冷（EVAC）和3DVC等方案的單價是普通散熱模組的數倍，未來也可能導入液冷，將帶動散熱模組市場規模的快速成長。

圖34、AI相關晶片推出年份與熱功耗瓦數

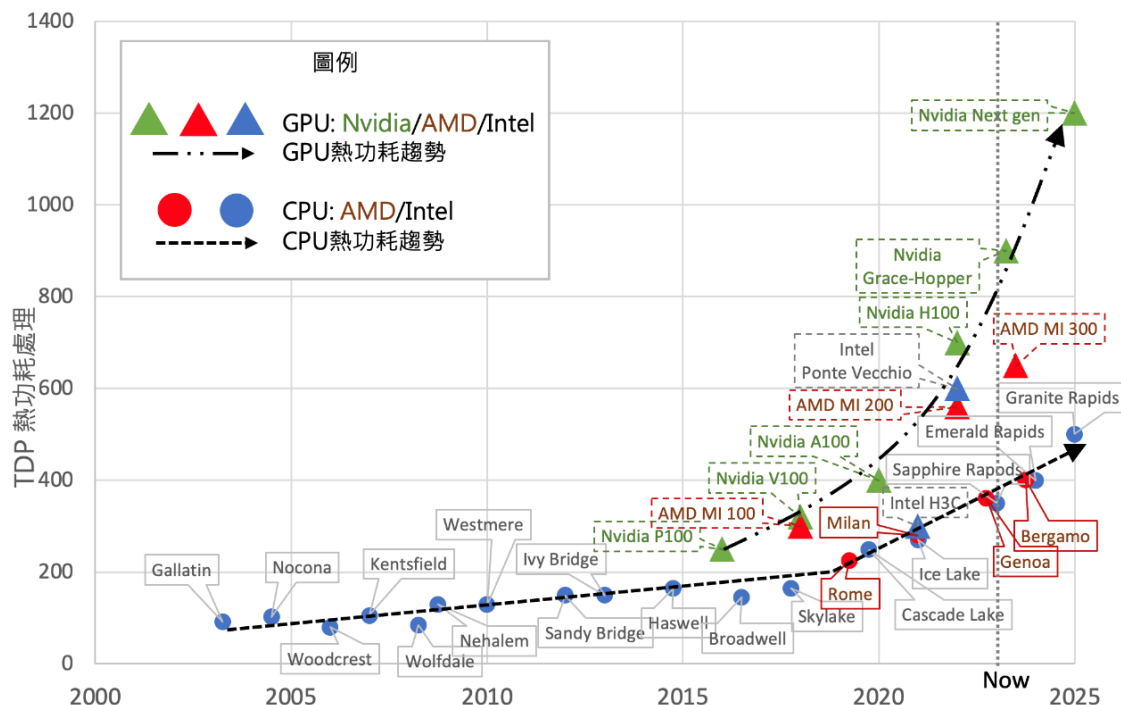


圖35、3DVC(上) & EVAC(下)



# AI伺服器相關零組件

表12、AI伺服器零組件與對應台廠

零組件	與AI伺服器之關係	相關個股
PCIe Retimer	AI 晶片必須搭配高速傳輸介面，高速傳輸介面經常伴隨訊號丟失問題，通常在PCIe I/O 處會有Re-timer 處理	譜瑞、群聯
CCL	PCIe5、400G Switch 需要使用Very/Ultra Low Loss 等級的CCL	聯茂、台耀、台光電
ABF	AI GPU 需要使用ABF載板，且面積與層數皆較一般應用有倍數成長	欣興、南電、景碩
散熱	AI 晶片超高的TDP需使用到大型散熱模組 /EVAC/3DVC，以及更高空氣流量的風扇	奇鋳、雙鴻、建準
機殼	AI Server 為少量客製化產品，通常可享有更高單價以及毛利率表現	勤誠、奇鋳
導軌	AI Server 通常為 4~10 U 特規設計，比普通伺服器重上數倍，有利單價提升	川湖
Switch	AI 訓練並連伺服器需要搭配 400G等級的交換器	智邦、眾達



# 相關個股

# 相關個股整理

相關個股	2022年預估EPS	2023年預估EPS	YoY	近5年PER
台積電 2330.TW	39.2	33.32	-15%	13-25
創意 3443.TW	27.69	34.5	24.6%	22-50
世芯 3661.TW	25.7	40.5*	57.5%	11-50
力旺 3529.TW	21.61	23.45	8.5%	28-100
雙鴻 3324.TW	14.68	16.01	6.7%	10-20
奇鋳 3017.TW	11.78	13.66	16.1%	10-18
穎歲 6515.TW	32.22	29.7	-7.8%	15-30

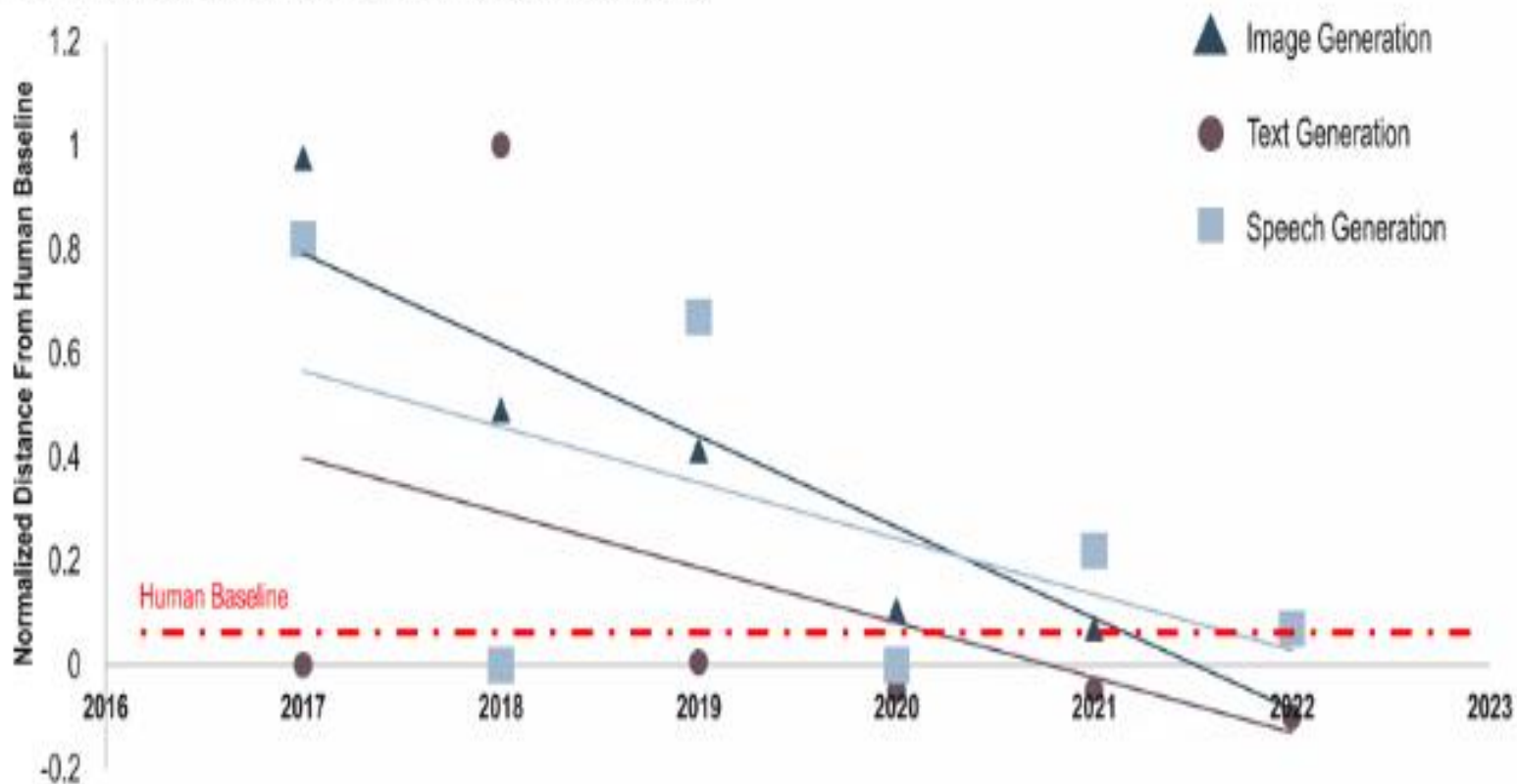
# 其他零組件個股評等

相關 個股	2022 年EPS	2023年預估EPS	YoY	近5年PER
台光電 2330.TW	15.24	13.27	-12.9%	9-18
欣興 3037.TW	20.08	14.02	-27%	8-22
南電 8046.TW	30.05	19.27	-35.9%	8-30
勤誠 8210.TW	8.32	7.55	-8.8%	9-13
川湖 2059.TW	42.56	30	-28.5%	9-25
緯穎 6669.TW	81.07	81.65	0.7%	10-20
英業達 2356.TW	1.71	1.81	5.9%	10-16
宜鼎 5289.TW	21.46	20*	-6.8%	8-15
群聯 8299.TW	27.71	22.62	-17.5%	8-16
譜瑞 4966.TW	61.6	40.2	-34.2%	10-30

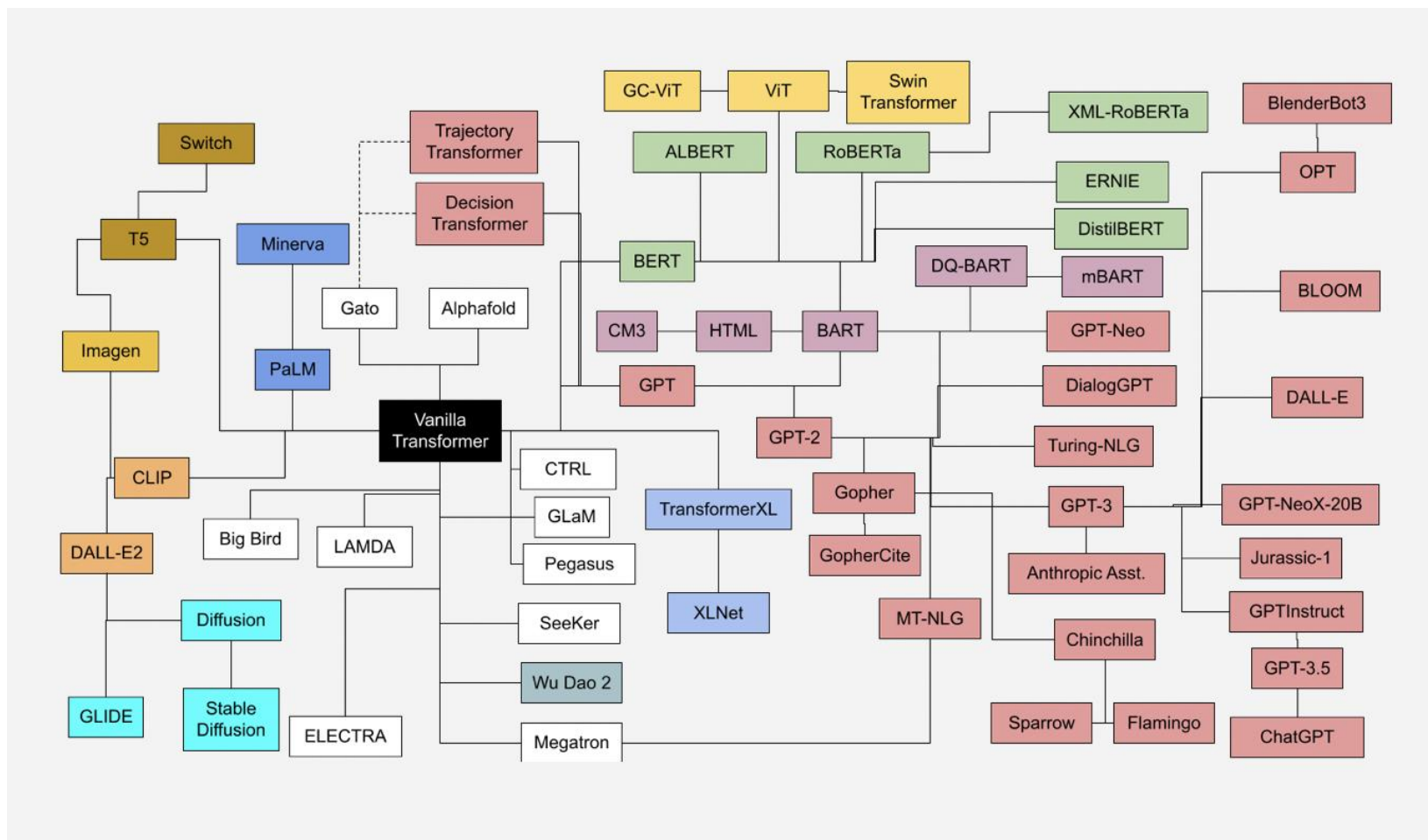
謝謝指教  
Q&A

# AI發展已經在部分領域超出人類表現標準

Near human-level performance across modalities and metrics



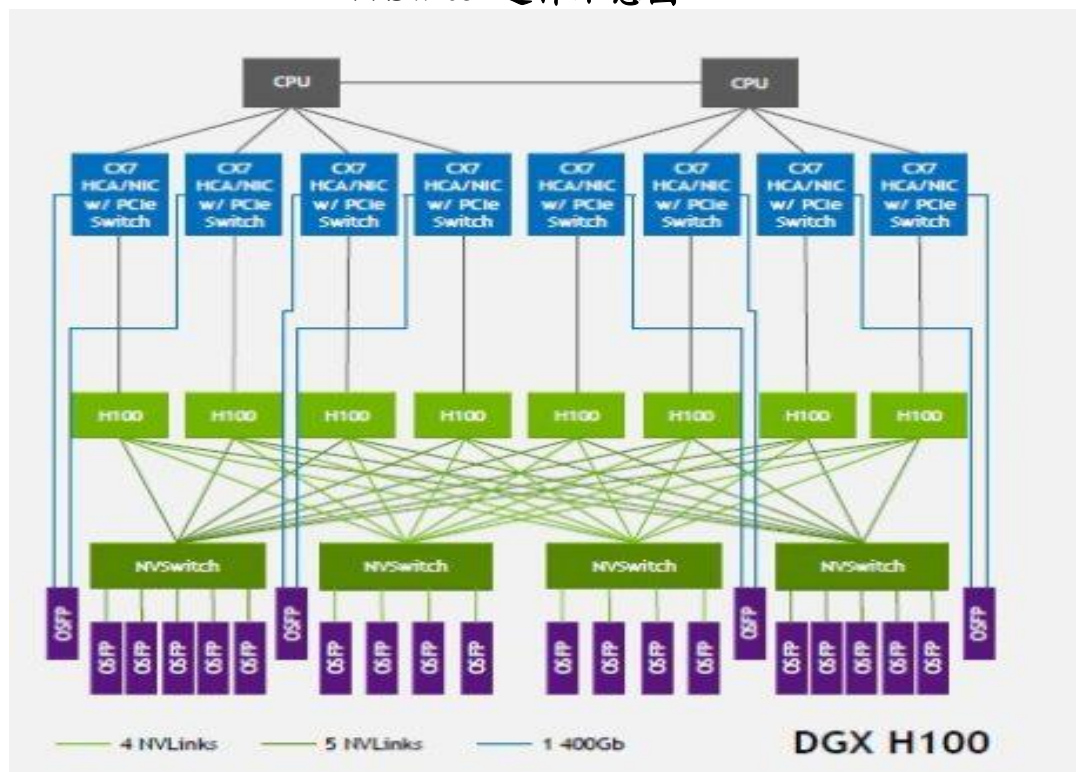
# 附錄:基於Transformer模型之Family Tree



# 大型語言模型「訓練」需跨伺服器並連，NVIDIA具技術優勢

- 大型語言模型進行「訓練」時，大多需要並連10~100台伺服器的GPU，例如GPT-3 並連了 124 台 AI Server，「推論」的算力要求相對低，不需並連多台伺服器，現有100G/400G Switch速度無法因應AI晶片需求，因此NVIDIA推出NVSwitch技術以提升傳輸速度

NVSwitch運作示意圖





# GPT模型訓練成本逐步上升

## GPT模型訓練成本推算

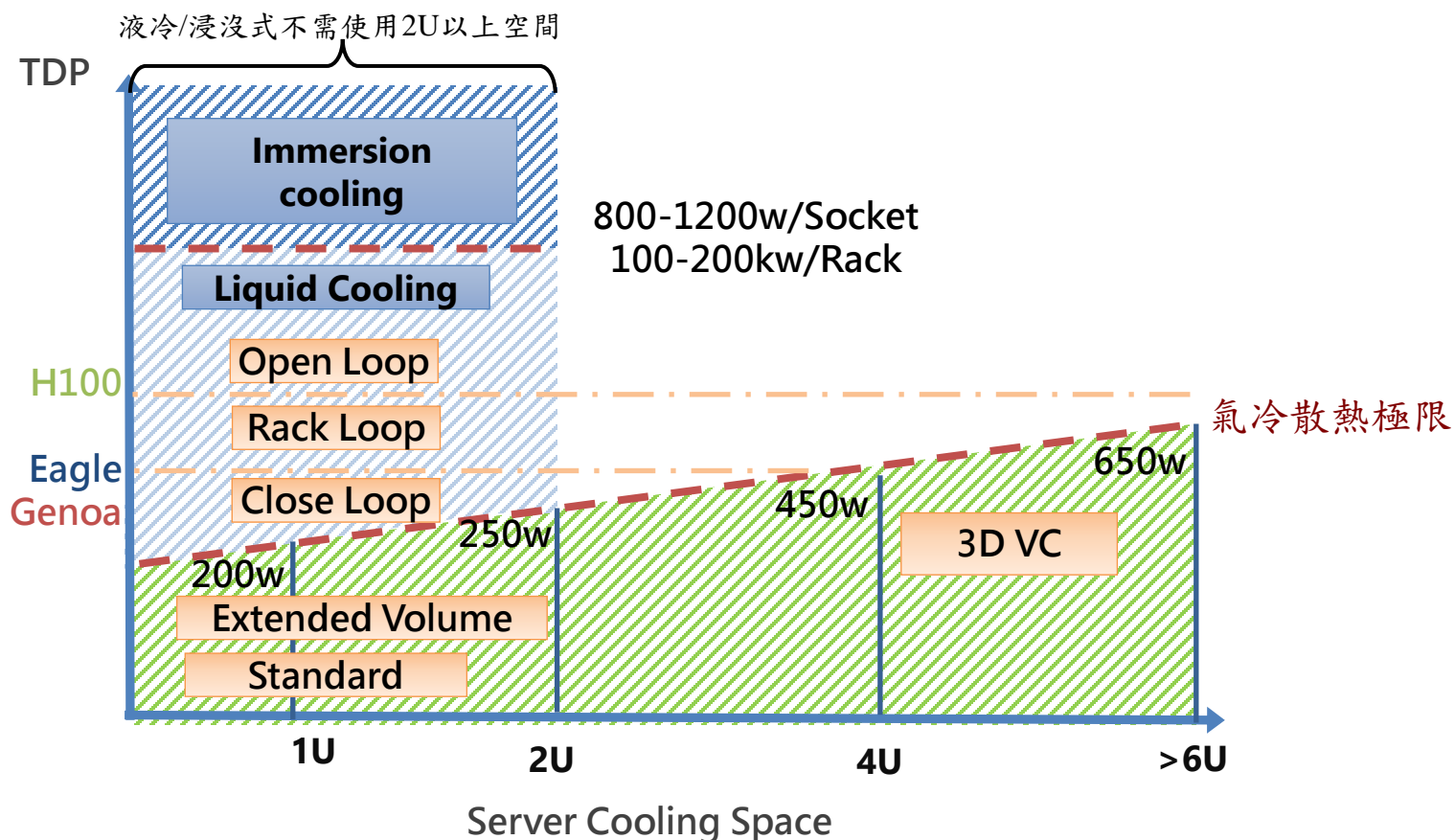
Model	Parameters	Tokens to Train to Chinchilla Point (B)	Cerebras Model Studio CS-2 Day to Train	Cerebras Model Studio Price to Train
GPT3-XL	1.3	26	0.4	\$2,500
GPT-J	6	120	8	\$45,000
GPT-3 6.7B	6.7	134	11	\$40,000
T-5 11B	11	34*	9	\$60,000
GPT-3 13B	13	260	39	\$150,000
GPT NeoX	20	400	47	\$525,000
GPT 70B	70	1,400	85	\$2,500,000
GPT 175B	175	3,500	Contact For Quote	Contact For Quote

\* - T5 tokens to train from the original T5 paper. Chinchilla scaling laws not applicable.

# 新一代伺服器已達氣冷散熱極限

- 2U空間下，250W大約是解熱極限；4U以上有可以解到600W。AI 晶片 TDP普遍超過 400W，不可能使用2U空間解熱。

圖11、熱散器解熱極限




# 散熱方案比較整理

表4、散熱方案比較整理

方案	Air Cooling 氣冷散熱	Liquid Cooling Cold Plate 液冷散熱		Single Phase Immersion Cooling 單項式 浸沒式液冷	Two Phase Immersion Cooling 兩項式 浸沒式液冷
熱交換原理		水對氣(L2A)	水對水(L2L)	水對水(L2L)	
水冷液成本(USD)/L	x	3.5-50	3.5-50	150-200	150-200
水冷液需求/Rack	x	30L	30L	1500L	1500L
最大解熱(Socket)	225W(1U)	<b>800W</b>	<b>800W</b>	800W	>800W
最大解熱(Rack)	10-20kW	10-22kW	<b>60-80kW</b>	160kW	250kW
PUE	Hyperscale:1.15 Enterprise:1.4-2	1.07	1.09-1	1.09	1.02-1.05
散熱硬體成本(USD) (不含空調設備)	3千美元	2-3萬美元	3萬美元	25-30萬美元	30-40萬美元
額外基礎建設需求		與氣冷相同	水冷塔 冰水管線	全新機房規劃 水冷塔、水管 線、機械手臂	全新機房規劃 機械手臂

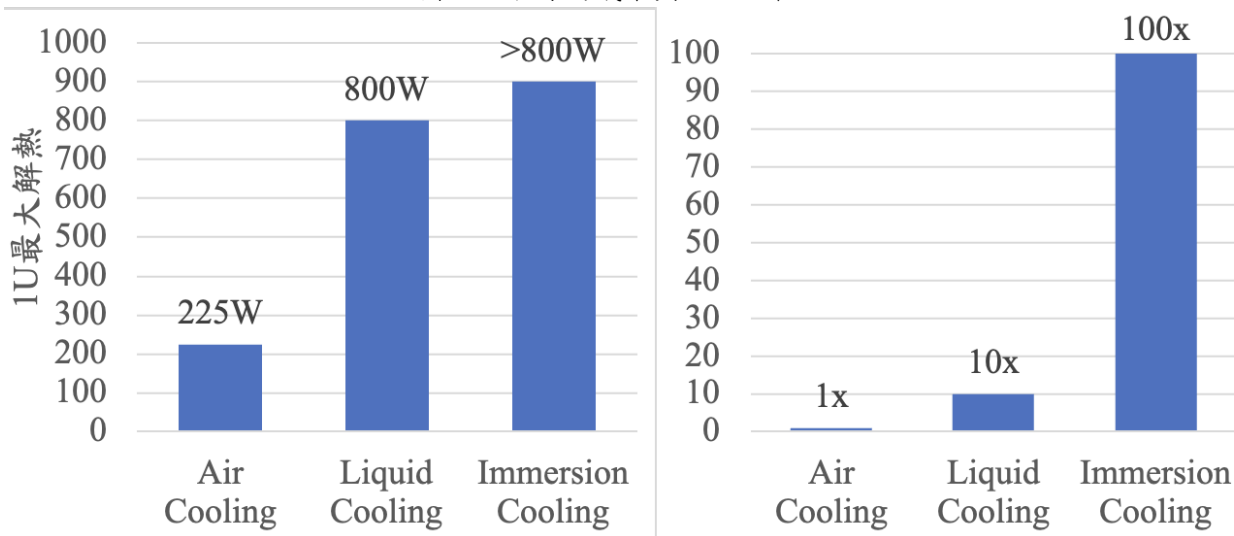
註：常見浸沒式Rack通常僅25U，為標準Rack一半。本研究為方便比較，各項數據皆轉換為42U標準。

 代表雖然不是最佳表現，但具有更好的成本效益。

# 液冷散熱「Liquid Cooling」明顯具有最佳的成本效益

- Liquid Cooling 在散熱能力、節能效果上，很明顯優於 Air Cooling；但 Immersion Cooling 多花費10倍的成本，卻只換來些微進步。
- 一些證據顯示 Liquid Cooling 現階段可行性遠高於 Immersion
  - 1. Liquid Cooling 已經存在幾十年，穩定性與技術相對成熟。
  - 2. 從 OPC、各加伺服器廠白皮書的份量，Liquid 篇幅明顯多於 Immersion。
  - 3. 阿里巴巴放緩 Immersion Cooling 的資料中心投資。
  - 4. Liquid Cooling 可在完全不更動基礎設施情況下進行部署。
- Immersion Cooling 須待成本下降後，才有機會被導入，研究員認為還要等3-5年時間。

圖26、液冷的成本與效益比較



## 【揭露事項與免責聲明】

本報告僅提供相關部門的內部教育訓練及相關人員之參考資料，並非針對特定客戶所作的投資建議，且在本報告撰寫過程中，並未考量讀者個別的財務狀況與需求，故本報告所提供的資訊無法適用於所有讀者。本報告係根據本公司所取得的資訊加以彙集及研究分析，本公司並不保證各項資訊之完整性及正確性。本報告中所提出之意見係為本報告出版當時的意見，邇後相關資訊或意見若有變更，本公司將不會另行通知。本公司亦無義務持續更新本報告之內容或追蹤研究本報告所涵蓋之主題。本報告中提及的標的價格、價值及收益隨時可能因各種本公司無法控制之政治、經濟、市場等因素而產生變化。本報告中之各項預測，均係基於對目前所得資訊作合理假設下所完成，所以並不必然實現。本報告不得視為買賣有價證券或其他金融商品的要約或要約之引誘。

可能個別基於特定目的且針對特定人士出具研究報告、提供口頭或書面的市場看法或投資建議（下稱“提供資訊”），鑑於提供資訊之單位、時間、對象及目的不同，本報告與本集團其他單位所提供資訊可能有不一致或相抵觸之情事；本集團各單位對於本報告所涵蓋之標的可能有投資或其他業務往來關係，各單位從事交易之方向亦可能與本報告不一致，讀者應審慎評估自身投資風險，自行決定投資方針，不應以前述不一致或相抵觸為由，主張本公司或本集團其他成員有侵害讀者權益之情事。