Halcyon Investment Advisor
Sophistication. Credence. Performance.

Alpha Helix Asset Management

投資備忘錄 (六十四)

2024.06.30

微軟跟蘋果的新戰場--EdgeAI 將如何影響全球對記憶體的需求

「殺雞焉用牛刀」：大模型執行複雜任務，小模型處理簡單問題：

AI的應用是近期以來最流行的議題。大模型需要的推論成本很高，通常需要耗費數百個 GPU 進行推論。不過，不是所有的問題都要依靠大型的 LLM，有些日常生活中的小任務，可以輕易地被能夠運行在電腦/手機上的小模型解決。如果不依據任務的難易度分流，將所有的推論需求都運行在數十兆個參數組成的超大語言模型之上，其所耗費的電力、土地和硬體成本，會讓算力提供者無法負荷。適時的將某些計算需求讓渡到手機/電腦上，才是更為經濟的選項。

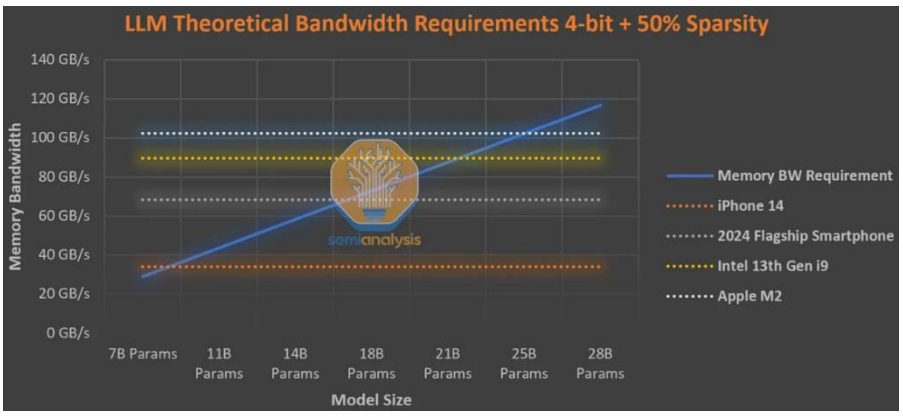
因此，除了訓練運行在資料中心，專門處理複雜任務的超大型LLM 以外，我們也需要發展能運行在手機/電腦上的小模型，也就是 EdgeAI。

EdgeAI 的優點：低成本、有隱私、易取得：

除了低成本以外，EdgeAI 在進行推論時，因為不需要將資料傳輸到雲端，使用者資料外洩的可能性大幅降低，提供了更高的隱私與資安保障。同時，因為不用將資料丟到與數據中心等待模型的計算，EdgeAI 可以在沒有網路連線的場景中應用，在特定場景（如飛機上、網路不好的郊區等）更為吃香。

EdgeAI 遭遇的瓶頸 -- 記憶體牆：

在大型模型的訓練/推論過程中，大部分的時間並不是花在計算矩陣相乘上，而是等待數據傳輸到計算資源上。因此，為了降低數據傳輸的等待時間，運行這些 AI 模型需要大量的記憶體頻寬。可以說記憶體頻寬的大小限制了 AI 的使用者能夠使用的模型參數多寡，同時，為了儲存模型的參數，使用者也需要相對應的記憶體容量。**簡單來說，為了加入 LLM 的功能，幾乎所有的硬體都需要更多的記憶體容量和更大的記憶體頻寬。**



EdgeAI 將大幅帶動 DRAM 的需求：

根據試算，要在一台電腦上運行一個 14B 參數大小的 LLM，至少需要 16GB 的記憶體容量，與微軟最新公佈的，AI PC需要的最低記憶體容量大小不謀而合。根據花旗銀行的研究報告，截至 2023 年底，平均一台電腦的記憶體容量僅約為 8.9GB，這表示，每台 AI PC所需要的 DRAM 容量會是過去的 1.8倍！

此外，根據 Apple 2024 WWDC，Apple Intelligence 只有在 iPhone 15 pro 上才能順利運行，而 iPhone 15 pro 是唯一擁有 8GB 記憶體容量的手機，因此可以推測，未來的 iPhone，為了要能夠順利運行Apple的 AI 模型，至少都會裝載8GB 以上的 DRAM。結合三星推出的 AI 手機也搭載了至少 12GB 以上的 DRAM來看，AI 手機對於 DRAM 的需求將比舊款高出至少 33% 以上！

在 EdgeAI 浪潮下受益的記憶體公司：

EdgeAI 對於 DRAM 的需求，結合 2025 微軟結束對 Windows 10的支援而帶來的 PC 換機潮，以及 Apple iPhone 16 將帶來的換機潮，有望持續推動 DDR5 的價格，三大記憶體廠商（三星、SK Hynix、美光）將持續受益。

另外，因為記憶體廠商的產能都被拿去生產 Nvidia GPU 所需要的 HBM，DRAM供給端緊繃，廠商開始將成熟製程（DDR3、DDR4）的產能轉移至 DDR5的生產，成熟製程的供給減少，成熟記憶體製造商也有望間接受惠。

結論：

雖然生成式 AI 的發展仍在早期階段，隨著模型的持續進步，市場已經開始針對不同的應用場景推出不同大小的基礎模型，原因很簡單：減少未來的推論成本。能在手機/電腦上運行的 EdgeAI，參數量較小、容量更輕盈、更適合處理較為簡單的任務，因此開始成為眾人關注的焦點。

不過，為了運行 EdgeAI，我們需要跨越「記憶體牆」-- 手機和電腦都需要有更大的記憶體容量與記憶體頻寬，才能保證最低階，參數最小的 LLM 能在端裝置上順利運行。我們估計， AI Phone 與 AI PC 需要增加的記憶體容量，至少會是現行裝置的30% 以上。有鑑於 AI PC和AI手機的推出勢在必行，未來全球對於 DRAM的需求將有跳躍式的成長。