如何在 AI 終端應用中選擇合適的快閃記憶體

2016 年 3 月，Google AlphaGo 在與韓國九段棋手李世石進行的圍棋比賽中，以 4:1 的絕對優勢完勝；2018 年底，Google AlphaStar 與兩位世界頂尖遊戲玩家在《星際爭霸(StartCraft II)》中展開對決，最終以兩個 5:0 的成績橫掃對手。儘管早在 1997 年，IBM 開發的電腦程式“深藍”就戰勝了當時的國際象棋特級大師加里·卡斯帕羅夫，但考慮到國際象棋的下法難度遠遠低於圍棋，所以 AlphaGo 的勝利在某種程度上也被視作“AI 時代的真正來臨”。

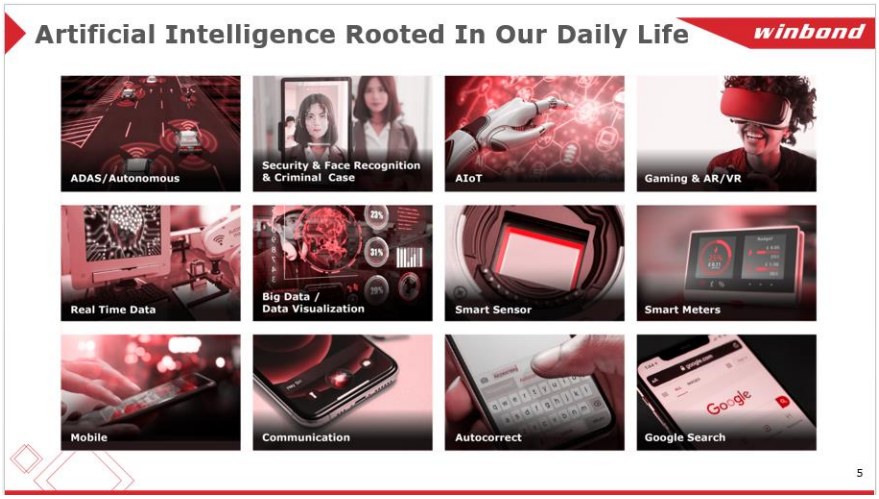
AI 的起源

1955 年到 1956 年間，時任 Dartmouth 學院助理教授的 John McCarthy，也是後來世界公認的 AI 教父，與來自哈佛大學的 Marvin Minsky、IBM 的 Claude Shannon、以及美國貝爾實驗室的 Nathaniel Rochester，首次共同定義了“人工智慧(AI)”的概念，即：

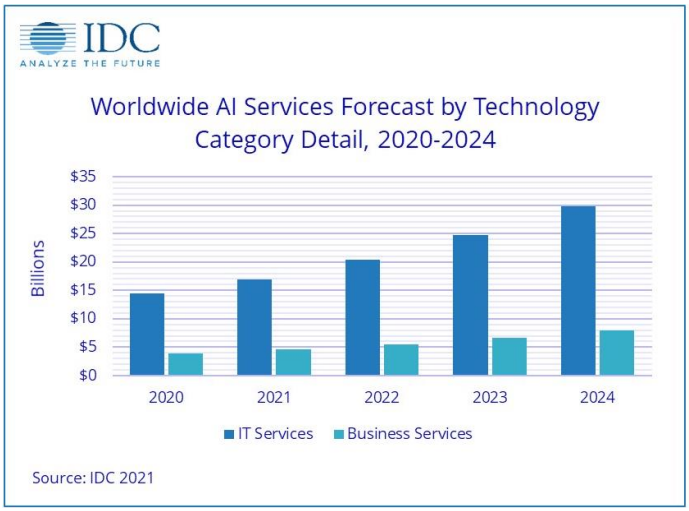
“如果任何的機器通過不一樣的語言，就可以將抽象的事物或概念表達出來，並且通過這個抽象的概念去幫助人們解決現有的問題，或者是它本身可以通過自主學習不斷地精進，我們就將其稱之為 AI。”

或者用更精煉的話語進行陳述，那就是：當任何機器的行為模式與人一樣時，我們就稱它是“智慧(intelligent)”的。

如此寬泛的定義自然帶來了與之對應的寬泛應用。除了下棋與遊戲，在自動駕駛領域，美國部分地區已經開放了 Level 4 級別的測試，相信真正的 Level 5 級別自動駕駛也是指日可待，而要保障車輛和行人的安全，我們依賴的除了法律法規，還有 AI 演算法的開發者；而在 IoT 應用領域，傳感、智慧手機、網路搜索、人臉/汽車牌照識別、智慧電表、機器視覺、工業控制……AI 正變得無處不在，讓工作和生活變得更加便捷與智慧。



IDC 的統計資料顯示，2020 年到 2021 年間，全球 AI 服務的年複合增長率達到了 17.4%，預計到 2024 年，這一數字將上升至 18.4%，市值約為 378 億美元。這其中包括了定制化的應用和針對定制化平臺所提供的相關支援與服務，例如一些深度學習架構、卷積神經網路、與 AI 相關的晶片產品(CPU、GPU、FPGA、TPU、ASIC)等。



同樣是來自 IDC 的資料，全球資料儲存量將從 2018 年的 33ZB 猛增到 2025 年的 175ZB，這其中超過 50%都來自 IoT 設備。如果考慮到 2025 年，全球將會部署約 140 億部 IoT 設備，那麼我們似乎就應該大量地增加雲端的計算單元數量與算力，才能應對海量的資料增長。

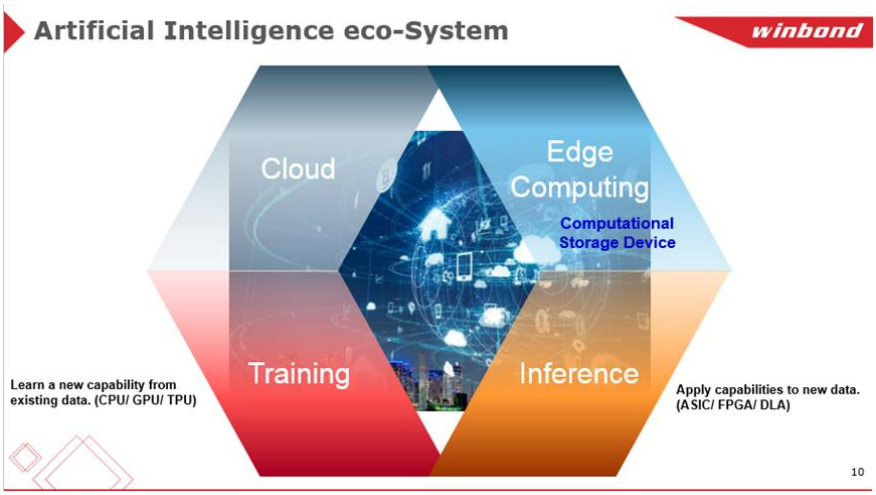
不過，事實並非如此，這一想法顯然沒有顧及到從終端到雲端的資料傳輸鏈條中面臨的頻寬、時間延遲等事實性挑戰，這也是為什麼“邊緣計算”能夠得以迅速崛起的原因。

因為我們最終發現，隨著 IoT 設備的快速增加，一味增加頻寬和伺服器數量的做法並非最優，不少應用完全可以在邊緣運算裝置中加以實現，而不必把所有資料都放到雲端去進行處理和傳輸、儲存和分析，這是不適合的。比如在工業自動化領域，資料儲存距離一定要近才有效率；5G 移動設備製造商如果不強化終端人工智慧裝置並進行計算與儲存架構更改，將會遭遇嚴重的電池壽命問題。

隱私安全，是另一個值得我們重視的環節，尤其是在當前萬物互聯的時代，機密資料/資料外泄或是遭到駭客入侵的事件屢有發生。如果我們能讓計算在邊緣側發生，節省“中央雲端至終端裝置”通路中資料傳輸的次數，那麼，在確保資料和網路安全的同時，也降低了功耗和系統總體成本。

不同 AI 晶片的比較

眾所周知，AI 技術根據應用不同被分為“訓練(Training)”和“推論(Inference)”兩大類，前者主要在雲端由 CPU、GPU、TPU 負責執行，目的是不斷增加資料庫資源以建立資料模型；後者則比較適合應用於邊緣裝置和特定應用，常由 ASIC、FPGA 類晶片進行處理，依託已經訓練好的資料模型進行推論。



如前文所述，與 AI 相關的晶片產品包括 CPU、GPU、FPGA、TPU、ASIC 等多種類型。華邦電子快閃記憶體產品企劃處黃仲宇從五個維度對上述不同的晶片類型進行了概略性的初步比較，包括算力(Computing)、軟體靈活性(Flexibility)、硬體相容性(Compatibility)、功耗(Power)和成本(Cost)。

- CPU

CPU 發展多年，運算能力強大，在軟硬體相容性方面首屈一指。但是，由於受到馮·諾曼架構的限制，資料需要在記憶體和處理器之間不斷反覆來回傳輸，所以限

制了處理的平均速度、且在功耗和成本表現上相較其他方案不是最好的，處於折衷的位置。

- GPU

以 Nvidia GPU 為例，由於採用了“Compute Unifiled Device Architecture”架構，且自身計算單元數量眾多，使得 GPU 不但能夠任意讀取記憶體的位置，而且還可通過虛擬記憶體的共用加速計算能力，儘管同樣受到馮·諾曼架構的限制，但其平均算力超過 CPU 幾十倍甚至上千倍。

GPU 同樣發展多年，軟硬體相容性好，但在功耗與成本表現上仍有改善空間，此外還須考慮硬體上的投資，如額外加裝冷卻系統以降低發熱。

- ASIC 晶片

ASIC 晶片是針對特定應用開發出來的產品，在經過驗證調整之後運算能力、整體功耗和成本可以達到最佳水準。

- FPGA

FPGA 同樣發展多年，軟硬體相容方面值得稱讚，整體算力、成本和功耗即便不是最佳水準，但也不失為一個不錯的折衷解決方案，對開發者來說，從 FPGA 入手切入 AI 晶片開發是個不錯的角度。

突破馮·諾曼架構瓶頸

在傳統計算設備廣泛採用的馮·諾曼架構中，計算和儲存功能不但是分離的，而且更側重於計算。資料在處理器和記憶體之間不停的來回傳輸，消耗了約 80%的時間和功耗。學術界為此想出了很多方法試圖改變這種狀況，例如通過光互連、2.5D/3D 堆疊實現高頻寬資料通信，或者通過增加暫存器數量、提高總暫存記憶體密度、縮短資料傳輸距離，來緩解資料傳遞時間延遲和總體功耗。

但試想一下，人類大腦有計算和儲存的區別嗎?我們是用左半球來計算，右半球做儲存的嗎?顯然不是，人腦本身的計算和儲存都發生在同一個地方，不需要資料移轉。

因此，學術界和產業界都希望儘快找到一種與人腦結構類似的創新架構的想法就不足為奇了，最好是能夠將儲存和計算有機地結合在一起，直接利用儲存單元進行計算，或者是將計算單元進行分類，使之對應不同的儲存單元，最大程度的消除資料移轉所帶來的功耗開銷，“計算存放裝置(Computational Storage Device)”的應用概念應運而生。

記憶體業界已有公司提出很值得借鑒的概念。NVM 不只儲存經過數位-類比轉換器轉換之後產生的類比信號，還可以將算力進行輸出，而輸入電壓和輸出電流則在 NVM 中扮演著可變電阻的角色，將類比電流信號經過類比-數位轉換器變為數位信號，從而完成數位信號輸入與數位信號輸出的全過程。這一做法的最大優勢在於它完全可以利用成熟的 20/28nm CMOS 工藝，而不用像 CPU/GPU 一樣去追求 7nm/5nm 這樣費用高昂的先進制程。

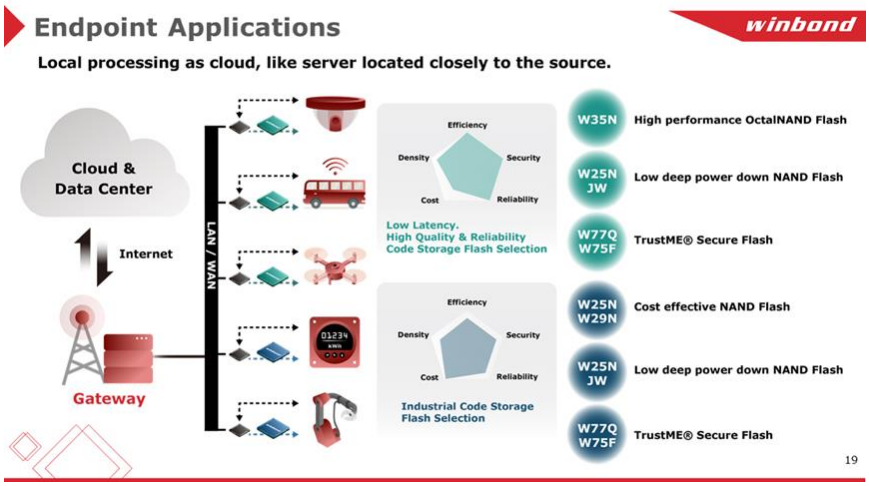
而伴隨成本和功耗開銷的降低，時間延遲特性也得到了顯著提升，這對無人機、智慧型機器人、自動駕駛、安防監控等應用來說都是至關重要的。

總體來說，終端推理過程計算複雜度低，涉及的任務較為固定，對硬體加速功能的通用性要求不高，無需頻繁變動架構，更符合邊緣運算裝置的運用。相關資料顯示，2017 年之前，人工智慧無論是訓練還是推理基本都在雲端完成，但到了 2023 年，在邊緣側設備/晶片上進行 AI 推理將佔據該市場一半以上的佔比，總額高達 200-300 億美元，這對 IC 廠商來說是一個非常龐大的市場。

AI 需要怎樣的快閃記憶體？

相信沒有人會反對高品質、高可靠性和低延遲快閃記憶體(Flash Memory)對 AI 晶片與應用的重要性。但黃仲宇提醒，面對不同的應用，需要從性能、功耗、安全、可靠性、高效能等多方面加以綜合考量，相比之下，成本考量雖然相當重要，但相較之下不是第一優先的考慮因素，不能顧此失彼。

高性能 OctalNAND Flash W35N 系列、適合低功耗應用的 W25NJW 系列、與安全相關的 W77Q/W75F 安全快閃記憶體系列，是華邦最具代表性的產品，例如，華邦 QspiNAND Flash 的資料傳輸率大概是 83MB 每秒，而 OctalNAND 系列最快的速度可以達到 240MB 每秒，幾乎是前者的 3 倍；在車載應用中，大量 AG1 125˚C NOR 系列和 AG2+ 115˚C NAND 系列 Flash 已經量產面世；而在智慧感測器或是產線機器人應用方面，華邦電子則可以提供具備成本、高效能的解決方案，比如 W25N/W29N NAND Flash 系列。



除了各式各樣的不同類型的 Flash 產品外，華邦 SpiStack®(NOR+NAND) 也很具特點。它將 NOR 晶片和 NAND 晶片堆疊到一個封裝中，例如 64MB Serial NOR 和 1Gb QspiNAND 晶片堆疊，使設計人員可以靈活地將代碼儲存在 NOR 晶片中，並將資料儲存在 NAND 晶片。此外，雖然是兩個晶片(NOR+NAND)的堆疊，但單一封裝的 SpiStack®，在使用上僅需 6 個信號引腳。

“華邦可以提供多樣化的 Flash 選型來滿足客戶各式各樣的記憶體代碼需求。就像在一場籃球比賽中一樣，晶片廠商扮演中鋒或前鋒，憑藉強大的算力和演算法不斷得分，而華邦就像後衛，在後場提供高品質、高性能的 Flash 產品，確保使用者在市場上不斷得分。”黃仲宇表示。

關於華邦

華邦電子為全球半導體記憶體解決方案領導廠商，主要業務包含產品設計、技術研發、晶圓製造、行銷及售後服務，致力於提供客戶全方位的利基型記憶體解決方案。華邦電子產品包含利基型動態隨機存取記憶體、行動記憶體、編碼型快閃記憶體和 TrustME® 安全快閃記憶體，廣泛應用在通訊、消費性電子、工業用以及車用電子、電腦周邊等領域。華邦總部位於台灣中部科學園區，在美國、日本、以色

列、中國大陸、香港、德國等地均設有子公司及服務據點。華邦在中科設有一座 12 吋晶圓廠，目前並於南科高雄園區興建新廠，未來將持續導入自行開發之製程技術，提供合作夥伴高品質的記憶體產品。

更多資訊，歡迎造訪： www.winbond.com