AI 搶灘陣式①：加強定義問題能力＆前瞻基礎研究

AI 大潮逼近！瀟灑衝浪或……狼狽滅頂？

■文：任芷萍

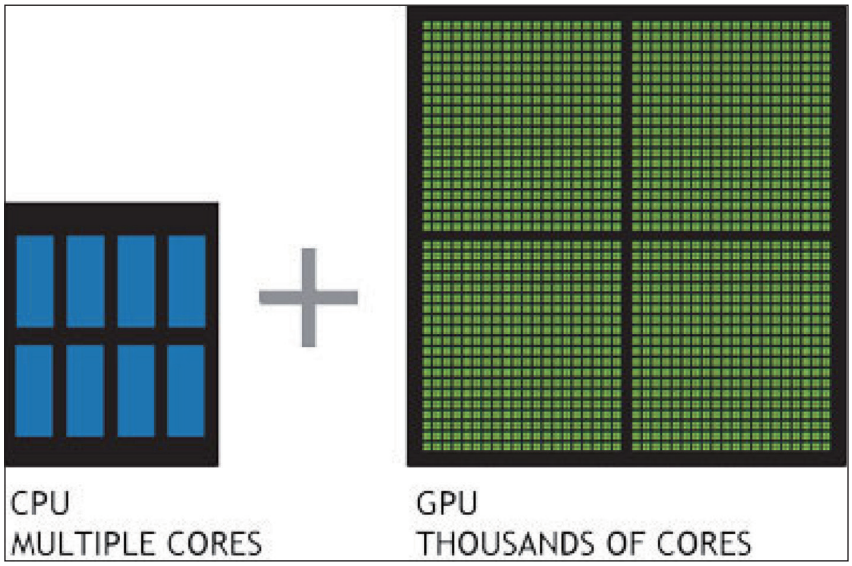
人工智慧 (AI) 狂潮來襲，有人盲從跟風、見獵心喜，也有人視若洪水猛獸，愁腸百結；關於強／弱 AI 的道德爭議、軟／硬運算技術、實用性、與人類競合矛盾的辯證不絕於耳，對於 CPU(中央處理器)、GPU(繪圖處理器)，乃至不斷推陳出新的各式特定應用積體電路 (ASIC)——APU(加速處理器)、VPU(視覺處理器)、TPU(張量處理器)，以及現場可編程邏輯陣列 (FPGA) 的討論亦不斷躍上版面。在走馬看花眾多演進史、產品規格與零散案例後，有志屹立 AI 浪頭的開發者在實際應用時，到底該考量哪些面向並選擇符合所需的元件？

GPU 以「浮點運算」名動江湖

CPU 是由運算邏輯單元 (ALU)、記憶體和控制器所組成，可一力完成訊號處理、指令編碼、發號施令等動作；然而在訊號鏈傳遞的過程中，若事無輕重緩急皆由 CPU 出面，未免不夠「知人善任」。為加速繁雜且重複性高的資料處理 (如：影像圖檔)，人們變通地將 IC 大部分的電晶體挪給超大陣列 ALU 做平行運算，因而促成 GPU 興起。若將整個 IC 資源配置用人類大腦潛能開發類比，有人善於思考、決策，一如 CPU；有人精於邏輯推理、就像 GPU，在「高精度浮點運算」尤其相形見長。因此，若將兩者硬性比較，似乎有失公允。

近年拜 AI 遍地開花之賜，讓 GPU 忽集萬千寵愛於一身，目前已有超過 2,000 家的 AI 新創公司建構於龍頭廠商輝達 (NVIDIA) 的產品之上，連帶使其身價暴漲。眼見 GPU 風華正茂，英特爾 (Intel) 與超微 (AMD) 也著手開發新一代指令集以提升 CPU 的浮點運算能力；前者於近期正式發佈業界首款「類神經網路處理器」(NNP)——Nervana，省略標準快取記憶體 (Cache)，避免記憶不同步、須清除快取的麻煩，以提高運算效率，後者在收購 ATi 公司後啓動「Fusion」專案，欲借助「異質系

圖 1：CPU 含有數顆核心，適用於為循序的序列處理最佳化；GPU 含有數千個更小型且高效率的核心，適用於同時處理多重任務



資料來源：http://www.nvidia.com.tw/object/what-is-gpu-computing-tw.html

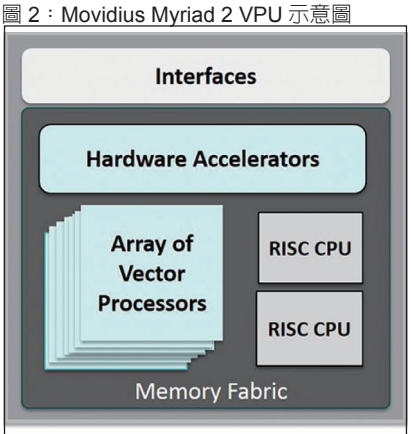
統架構」(HSA) 的 APU，以 CPU ＋ GPU「二打一」之姿奮力一搏。

APU 最大特點為：使用 HyperTransport (HT) 匯流排將 CPU 和 GPU 兩個不同運算架構整合在同一個 IC 上，使其協同運作；開發人員可透過免權利金 OpenCL (開放運算語言) 的應用程式介面 (API)，加速 CPU、GPU 和 APU 進程。APU 的最終目標是：將 CPU 與 GPU「完全融合」，根據任務類別自動分派運算任務，以降低裝置的功耗和發熱。特別一提的是，AMD 於 2007 年發佈的 SSE5 指令集，已意外被 Intel 吸收、優化為「乘法及加法融合指令」(Fused Multiply-Add, FMA)，以簡化運算步驟、使浮點運算的峰值倍增。

VPU、TPU 爲「特定應用」分衆市場而生

VPU 主要專注於影像運算、不須負責輸出或處理其它應用，旨在作為特定 AI 工作的獨立加速器或「推論」(inference) 引擎，頗具「特務」意味；真正捧紅 VPU 一詞的推手，當屬電腦視覺晶片先驅 Movidius 公司。Movidius Myriad 2 VPU 晶片組已獲聯想 (Lenovo) 虛擬實境 (VR) 產品採用，而 Google 也承諾用它作為平台的神經運算引擎，增進行動裝置的機器學習能力。2016 年 9 月收歸 Intel 旗下後，今年更為大疆創新科技的首款迷你無人機 Spark 加強物體檢測、實現 3D 製圖和基於深度學習演算法的環境感知。

圖 2：Movidius Myriad 2 VPU 示意圖



資料來源：https://www.movidius.com/solutions/vision-processing-unit

隨後，Intel 又陸續推出內建 Movidius Myriad 2 VPU 的神經運算棒 (NCS) 與 Myriad X VPU 系統單晶片 (SoC)——前者每秒超過千億次浮點運算可即時執行深度神經網路 (DNN)；後者訴求「全球首個配備專用神經網路運算引擎的 SoC」。同樣看好影像處理商機，日商索思未來 (Socionext) 亦搶在今年初發表首款符合 OpenVX 開放 API 標準的圖像顯示控制器暨加速器 SC1810，內建專利 VPU 除配有 OpenVX 硬體加速器，還具備可編程資料平行加速器；免授權金的 API 及專屬 VPU、H.264 編解碼器，可降低嵌入環景監控、物體偵測等視覺功能成本。

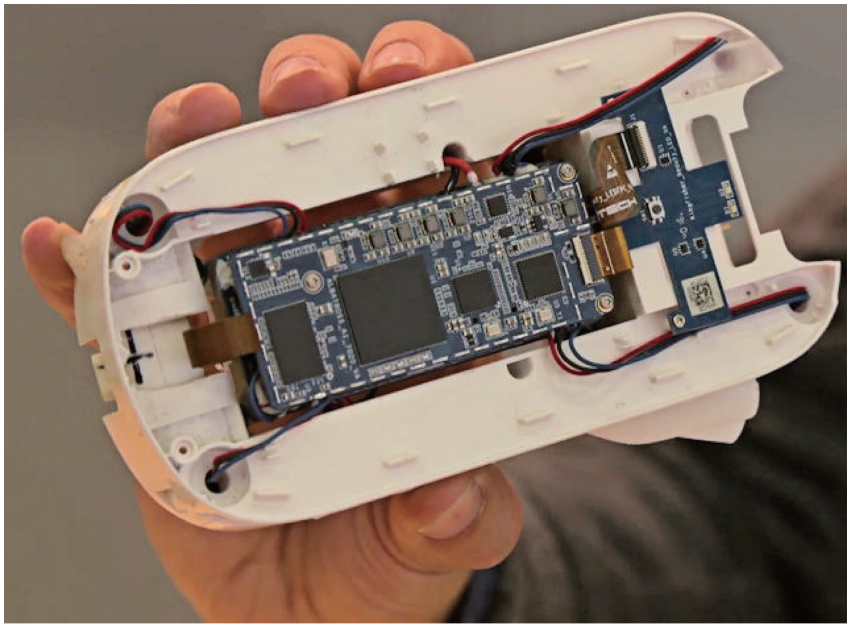
TPU 是 Google 針對自家開源深度學習框架 TensorFlow 所發展的 ASIC，兼具 CPU、GPU 可編程優點，能在不同網路模型執行複雜指令集 (CISC)，在深度學習 (Deep Learning) 的「推論」，擁有更好的「單位功耗效能」。即使個別浮點運算能力不如 GPU，但團結力量大，今年發佈的第二代 TPU2(又稱 Cloud TPU) 集成四個 16 位元處理器，每組效能達 180 TFLOPS (TeraFLOPS)；現已佈署在 Google Compute Engine 平台上，與 CPU、GPU 協作加速。若將 64 個 TPU2 串聯升級為「TPU Pods」超級電腦，效能更上看 11.5 PFlops (PeraFLOPS)。

「靈活」是 FPGA 最大優勢，異質運算蓄勢起飛

上述一衆 ASIC 產品，都是電路佈局既定的硬體系統；但嚴格來說，用於圖像處理 (訓練) 及統計運算 (推論) 的技術要求仍有些許差異。若想要用一塊電路板靈活滿足兩種需求，就得為開發人員保留部分編程空間，於是，類似「半成品」的 FPGA，因為可一舉解決全定製電路的不足與 FPGA 邏輯閘電路有限的缺憾，也趁勢搭上這波 AI 順風車。開發人員可運用 Verilog 或 VHDL 等硬體描述語言定義邏輯電路，根據當下需要快速佈線、連接內部邏輯區塊後，再燒錄至 FPGA；惟因運算速度有限、功耗相對大，無法因應過於複雜的設計。

賽靈思 (Xilinx) 繼今年 3 月在 2017 嵌入式世界大會上展示視覺導向智慧系統後，5 月更宣佈投資專精於深度壓縮、編譯工具鏈，以及系統層級最佳化的機器學習應用領域的深鑒科技公司，擬推出各種從終端到雲端的推論平台。整體論之，FPGA 適用於初期出貨量偏

圖 3：零度科技 (Zerotech) 的只有口袋大小的無人機 Dobby AI，採用 Xilinx Zynq Z-7020 SoC 運作深度學習功能偵測人的手勢



資料來源：https://forums.xilinx.com/t5/Xcell-Daily-Blog/Zerotech-s-palm-sized-Dobby-AI-drone-uses-DeePhi-machine/ba-p/774764

小的產品或供原型設計驗證之用，Cadence、Synopsys 等電子設計自動化 (EDA) 供應商亦積極提供相關解決方案；以 NVIDIA 為首的 GPU 受惠於先天的平行運算架構，格外適合捕捉「瞬間」的高效運算 (HPC)；但欲將 AI 進階到「人腦決策」，CPU 亦須與時俱進、加入協作行列。

另一方面，AMD 正快馬加鞭衝刺 AMD Radeon Instinct GPU 加速器，加上自有 x86 CPU 專利、鎖定伺服器開發的多核心 EPYC 處理器加持，是否能讓「異質融合運算」再上一層樓、與 Intel、NVIDIA 一爭長短？值得關注。此外，由於多數機器學習不需從 cache 讀取數據，若資料來源高度本地化、側重就地執行的專職應用，則 VPU、TPU 等「術業有專攻」的分衆 ASIC，更便於「挑重點」嵌入、集中火力專攻某一項職能——這也是為何 AlphaGo 從 CPU + GPU 架構轉成 TPU 後，對弈實力如秋風掃落葉、大獲全勝之故。

AI 要能付諸實際應用，才有獲利機會

事實上，系統商自行開發的 ASSP 預估有增多趨勢，也是衆家矽智財暨 EDA 工具商所念茲在茲、殷切期盼的潛力市場；例如，微軟 (Microsoft) 的擴增實境 (AR) 顯示器 HoloLens，即內建「全像處理器」(Holographic Processing Unit, HPU)。歸根結底，我們究竟要如何在這波滔天巨浪中安身立命？誠如波士頓顧問公司 (BCG) 合夥人暨董事總經理徐瑞廷日前在

照片人物：BCG 合夥人暨董事總經理徐瑞廷

《2017 arm 科技論壇高峰座談會》中所提及：「時至今日，光談技術及對行業的影響是不夠的，如何落實到應用層面才重要，而不同應用場景的使用案例 (Use Case) 更是焦點。」

徐瑞廷表示，台灣科技業過去多只著重於自己的產品且過度分工，如果不能了解終端使用者怎麼應用產品，今後將很難使得上力。除了商業模式的改變、滿足客戶對產品規格的需求並降低成本外，還須考慮到與不同生態系的夥伴合作。企業內部也要重新思考組織策略，不是只有賣產品、而是要更深入理解客戶，故人才需求也不同；從銷售單一元件、產品到整體解決方案，既有員工能力未必跟得上。為獲取新能量，導致投資併購、企業轉型與是否引入新血的議題十分熱烈，例如，日本某家感測器廠商為取得終端資料便成立顧問團隊，改以「提升良率」為賣點。

物聯網 (IoT) 先驅之一的研華科技技術長楊瑞祥指出，千禧年以前的第一波數位革命，締造了台灣 PC 盛況；但 2000 ～ 2015 年、

特別是 2007 年智慧手機的爆量，迎來了以聯網為基礎的第二波數位經濟，成就了全球市值最高的五大企業——Amazon、Google、Facebook、APPLE 和 Microsoft，可惜台灣卻未能從中蒙受真正大利益。2015 年後，無所不在的聯網將涵蓋設備與人，從雲端平台衍生的服務與實體企業轉型，將是另一波突圍契機。例如，日本地震頻繁，為確保道路、橋樑等基礎設施安全，研華與 arm 合作以 IoT 取代人工量測振動幅度。

我們未必活在演算法當中，但可能被 AI 包圍

楊瑞祥主張，科技導入是提升效能的有效方法。導入運算平台和無線感測器後，可提供機器學習 (Machine Learning) 建模資訊，借助端到端連接提升產業效能。他強調，IoT 的價值在產業應用，有專業知識 (Domain Know-how) 的門檻，不能單兵作戰、須與不同產業跨域協作「共創」(co-create)。AI 與資料是密不可分的，需要足夠量體及高品質的資料做訓練，不是單靠演算法就能解決問題；雖然不需太多訓練的通識認知也是 AI 一個支脈，但如何操作複雜設備或備料預估就需專業訓練，以便收集資料後經由機器學習、建立數據模型，進而產生商業效益。

照片人物：研華科技技術長楊瑞祥

楊瑞祥以進展相對快速的機器視覺為例，原先精密運動控制生產線還須作業員目視檢測是否有不良品，但經研華收集三個月資料、與學界及研究單位合作兩週並訓練三天後，利用 52 層類神經網路做「轉移學習」(Transfer Learning) 做出來的第一版成果，就比人工目測還準確。AI 只是泛用的方法，關鍵在挑對題目，並給出成批的資料或相對應的規則；產業應用須導入專業知識，資料收集與標籤分類需要與專家合作。另一個例子是用深度學習收集工業控制的虛擬量測資料，以建立模型、模擬實際作業站點的溫度，將溫度誤差精準控制在正負 1.5°C 內。

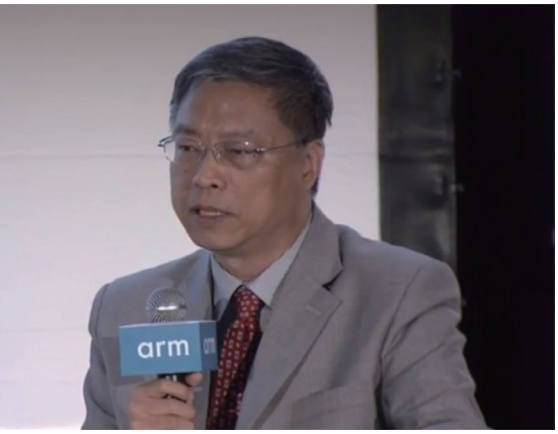
「AI 帶來的衝擊不亞於當初導入電力。它不是純題材或資訊題目，而是需要端到端軟、硬系統的共襄盛舉」，楊瑞祥說。中央研究院資訊科技創新研究中心副主任修丕承剖析，技術引導破壞式創新將跳脫現有軟、硬體框架，將為 IoT 帶來革命性發展。例如有電就走、沒電就停的「非揮發性處理器」(non- Volatile Processor)，因為資料不會掉、不必重做，即使電源微弱或不穩定仍可運作，這是 CPU 做不到的。目前仍在學術研究階段，日後可應用在醫療植入式生理監測；透過環境震動或太陽能／無線供電，不需電池、將體積最小化，也不須過幾年還要再度開刀更換。

爭取與雲端服務商一同學習，培育人才刻不容緩

照片人物：中央研究院資訊科技創新研究中心副主任修丕承

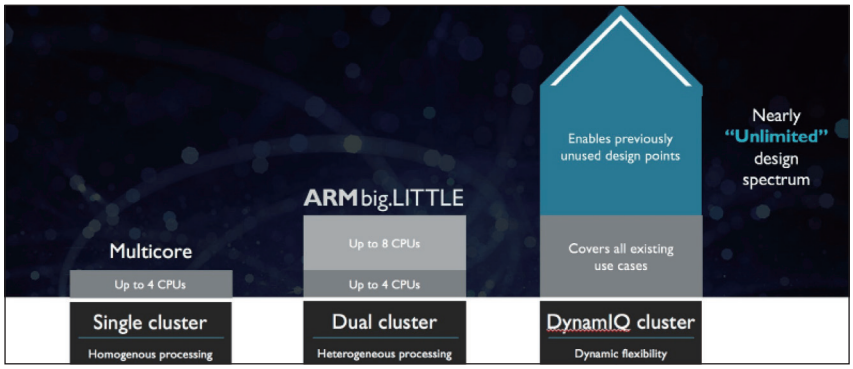
這樣的方式可降低維護成本及環境污染，一旦商用化，可望複製先前非揮發性記憶體為手機產業寫下一頁輝煌的模式。修丕承認為，台灣擅長軟硬整合，更適合講究客製化的市場；不過，最有價值不是軟體或硬體本身，而是已經標示分類的資料。Google、Amazon 等雲端服務商之所以終端裝置只賣 30、40 美元，就是意在獲得使用者資料，好據以提供高獲利的應用服務。他建議，台灣的強項在終端裝置，與其將資料白白送給雲端服務商，不如爭取一起參與學習的過程，而這需要分散式或嵌入式深度學習等前瞻研究的協助。

曾與教育部合作人才培育計畫長達二十年的台大系統晶片設計中心副主任暨台大電子所所長吳安



照片人物：台大系統晶片設計中心副主任暨台大電子所所長吳安宇

圖 4：DynamIQ 是 arm Cortex-A CPU 的新技術，透過在叢集 (Cluster) 內部集成智能電源來提高 AI 運算能源效率



資料來源：https://community.arm.com/processors/b/blog/posts/arm-dynamiq-technology-for-the-next-era-of-compute

宇則提到「大軍未動、糧草先行」，人才儲備先於一切。以往行動裝置的成功建立在技術水平分工上，現在應設法弭平學用落差，找到垂直應用去串接所有技術。吳安宇透露，教育部今年已開始執行為期四年的「智慧聯網應用與技術人才培育計畫」，將加強產學互動並翻轉教育方式，由偏重解決問題及單向訓練，轉向「界定問題」(Problem Formulation) 與「問題導向學習」(Problem-based Learning)，讓學生學會定義問題，並順利與產業界無縫接軌。

吳安宇認為，台灣地小並非弱點，反而利於串接資源；加上善用雲端教學，在邊緣運算會有不錯的發展機會。他舉例，工廠或家庭中有些緊急狀況，待上傳雲端運算完成，一來一回可能為時已晚；如何就近判別、處置，結合元件與顧問專業為技術創造價值，是不錯的發展方向。 CTA

Arm 全新顯示解決方案 將使用者體驗提升到下一個世代

為解決顯示技術在虛擬實境 (VR)、高動態範圍 (HDR)、多重視窗模式以及眾多面板與介面上所帶來的挑戰，全球 IP 矽智財授權廠商 Arm 推出全新顯示架構 Komeda，包含 Mali-D71、CoreLink MMU-600 及 Assertive Display 5 等產品。

- **Mali-D71**：針對 VR 應用確保 4K 120 高解析度效能
  - —降低 30% 系統功耗：使用固定功能的硬體元件，透過疊加、旋轉、高品質縮放及其他影像處理，在管線的最後階段進行、即傳送最後輸出訊號到螢幕前，以降低 GPU 工作量。
  - —節省兩倍空間：驅動單一顯示時、能使用第二顯示的資源，讓更多全螢幕圖層同時被處理，效能翻倍。
  - —四倍延遲容忍度：要呈現 4K 解析度，幀率必須至少輸出 120 fps，顯示處理器須最佳化使用系統匯流排的時間；沒有顯示的時候，須以毫秒方式預先擷取像素，才能在緩衝區一直保有充分內容。歸功於記憶體子系統優化，Mali-D71 能在相同資料吞吐量的情況下，較前一代產品容許系統匯流排有四倍的延遲。
  - —兩倍畫素吞吐量：當在新的並列 (side-by-side) 模式中運行時，Mali-D71 能達到前所未有的雙倍畫素吞吐量，呈現極致的 VR 4K120 高解析度並維持極低功耗曲線。
- **CoreLink MMU-600**：新興行動顯示的挑戰幾乎都與資料管理有關。將 CoreLink MMU-600 專屬渲染版本與 Mali-D71 緊密結合，能節省 55% 面積；透過在即時路徑上將 MMU 的延遲隱藏，可降低 50% 的延遲度。
- **Assertive Display 5**：延續在螢幕上反映接近人類視覺感知極限的獨特能力，進一步改善在強光下的顯像效果。除了提供 HDR 管理功能，還改進色彩與色域管理功能及省電效能。