AI 搶灘陣式③：從裝置、主機到雲端，軟、硬全面加速

「深度學習」堆疊不簡單！NVIDIA 讓機器自己寫軟體

■文：任芷萍

照片人物：NVIDIA 創辦人暨執行長黃仁勳

GPU(繪圖處理器) 已成大數據運算、分析、排序等平行運算的骨幹。GPU 鼻祖可追溯至輝達 (NVIDIA) 於 1999 年首推、專職幾何轉換並改進光影表現的 GeForce 256；2008 年，受惠於資訊大廠力拱 OpenCL 跨平台應用程式介面 (API)，終將 GPU 推上「超級電腦」舞台。2015 年，「JETSON TX1」模組的每秒浮點運算能力達陣 Tera 等級，為 NVIDIA 成功打開深度學習 (Deep Learning)、電腦視覺等嵌入式人工智慧 (AI) 市場。

CUDA 爲 ASIC 編程奠定里程碑

日前 GTC Taiwan 年會上，NVIDIA 創辦人暨執行長黃仁勳一上台就表示：台灣是 PC 革命的中心，將電腦帶入每一個家庭；到了行動世代，電腦被帶進每一個人的口袋；邁入雲端紀元，每個行動裝置都將成為一部超級電腦；今後的 AI 世界，電腦將可自行編程軟體、自主學習，為數以億萬計的運算裝置注入智慧，為產業帶來空前盛況，而軟體與運算是電腦科技的兩大驅動因素。他認為，軟體發展將因深度學習而有巨大改變，能自動偵測、學習，從大數據取得所需並設計成可被理解的架構與知識，進而找出共通點和規則，做出預測、判斷。

然而，深度學習有一個重要前提：強大的運算能力。可惜受限於半導體物理特性，每年增加 50% 電晶體及效能的摩爾定律已近尾聲；取而代之的是，另一股新興力量正在崛起——基於 GPU 的全新演算法和運算架構，讓它與 CPU 並行協作、加速運算。黃仁勳推估，在新的微處理器、軟體堆疊、演算法及應用程式開發者的攜手合作下，2025 年的運算能力將增加千倍。他並提到，軟體與運算平台兩者其實互為雞生蛋、蛋生雞的微妙關係：運算平台需要軟體配合，才能解決過去無法處理的問題；另一方面，軟體開發也要考慮到硬體是否有能力支撐。

「這也是為何近年並無太多新運算平台橫空出世的原因」，黃仁勳說。他回顧，誕生已滿十個年頭的革命性運算架構 CUDA，即

結合了高效能的特定應用積體電路(ASIC)及可編程模式，才能讓開發者輕鬆應對大量、複雜的平行運算；近五年間，CUDA 開發者數量已狂增十五倍，迄今累積逾 64.5 萬人，下載次數大於 600 萬次，單是去年就有 180 萬次。黃仁勳自豪宣示，2017 年諾貝爾物理、化學兩個獎項的得主——前者證明愛因斯坦重力波理論，後者藉低溫電子顯微鏡的高傳真原子尺度研究分子，皆得力於 NVIDIA GPU 的匡助。

VR 商業價值漸顯，「Holodeck」爲終端創建運作環境

有鑑於電腦繪圖也是虛擬實境 (VR) 的重要推手，NVIDIA 特為 hTC VIVE 等 VR 裝置創造名為「Holodeck」的高度逼真、可遵循物理原則之 VR 環境，以描繪真實場景；使用者可用它分享數位內容、邀請真人穿梭其中並分派 AI 角色；場中人員的轉頭、揮手動作皆能忠實呈現，且可感覺到觸碰或疼痛。借用這樣的虛擬會議室召開產品會議，可直接將汽、機車等設計圖匯入「Holodeck」，讓分散各地的與會人員彷彿置身同一個約定空間商討；不僅能透視產品內部構造、獲悉全部細節，還可即時調整參數或變更設計外觀、材質。

黃仁勳強調，AI 可解決以往軟體編程無法解決的問題，例如，長時間做光線追蹤，而 NVIDIA 卻能運用深度學習來訓練自動編碼器，完成局部呈現的寫實影像；每一次的光粒子與表面撞擊到進入眼睛的過程需要許多數學運算，若不夠完整，根本無從察覺差異。NVIDIA 與 Remedy 創建一種神經網路，可透過觀看影片、從說話者的語態模擬當時的 3D 面部表情，做成動畫；另與加拿大新創公司 WRNCH 訓練網路、推論 2D 影片中的人物在 3D 空間裡的位置及姿勢，一個典型應用是：只要有攝影機對著人、物拍攝，就能瞬間將其轉換到 VR 環境。

圖 1：在「Holodeck」虛擬環境進行產品會議，有身歷其境之感

資料來源：翻攝於 NVIDIA GTC Taiwan 螢幕展示

此外，愛丁堡大學的研究人員訓練網路模擬一個能適應不同環境與地形的虛擬角色，它會自行規劃行進路徑並聰明地避開障礙物；而加洲大學柏克萊分校與 OpenAI 發明的「一次性模仿學習」，只須寥寥數次的示範，就能成功教導機器人執行新任務。諸如此類，都是人類編程不容易做到的，也呼應了黃仁勳稍早「AI is eating Software」的說法。為協助培植台灣本土 AI 產業，NVIDIA 將與科技部合作，提供包括網路實驗室與研討課程等實作訓練課程，學習如何使用開源框架與 NVIDIA GPU 加速深度學習平台，擬於未來四年培訓 3,000 位開發人員。

Tesla P100 GPU 爲伺服器加速，Jetson TX2 聚焦邊緣裝置

與此同時，國家高速網路與計算中心將組建全台第一部專為 AI 打造、搭載 NVIDIA DGX AI 運算平台與 Volta GPU 的超級電

腦，期於明年達到 4 petaflops 的效能、躋身全球五百大排行榜的前二十五名，預計四年內上看 10 petaflops。事實上，NVIDIA 在去年推出搭載 Tesla P100 GPU 加速器、由 124 部 DGX-1 伺服器組成的 DGX SATURNV 超級電腦，一上市就在 TOP 500 勇奪第二十八名，每秒可執行 1 quintillion (10 的 18 次方) 次運算，鎖定高效燃油引擎、完全燃燒核融合反應器模型及醫藥研究等超精密大型應用，包括 NVIDIA DRIVE PX 2 自駕車。

DGX-1 整合了深度學習軟體、開發工具及八顆 Pascal 架構的 Tesla P100 GPU，可謂是連接 AI 與超級電腦的橋樑。延續 Pascal 架構、於今年第一季新登場的 Jetson TX2 借助「六核異構運算」，將整套 AI 系統集成在信用卡大小的電路板上，且耗電不到 7.5 瓦，特別適用於小體積、低功耗的邊緣裝置 (Edge Device)，可在商用無人機、工業機械、智慧攝影等終端實現進階導航、影像分類與語音識別的神經網路運算，例如：多合一電腦分享、自動偵測商品存量、接合 360° 影片做 4K 高畫質直播串流服務，或支援小型無人機的視覺演算。

伺服器、超級電腦和終端佈局完成後，NVIDIA 的研發腳步並未停歇；為追求更高效運算 (HPC)，今年第二季，NVIDIA 再發佈第七代 GPU 架構 Volta——採用台積電 12nm 製程、集成 210 億顆電晶體以及新的數字格式和 CUDA 指令，可執行 4×4 矩陣運算、支援 250 個應用程式，首款 GPU 代表作是 Tesla V100；一部搭載 Tesla V100 GPU 的伺服器效能，號稱足以頂替市售搭載數百顆 CPU 的傳統 HPC 電腦，跨越深度學習的 100 TFLOPS 效能「天塹」(官方數據為 120 TFLOPS)，為 AI 訓練 (Training) 和推論 (Inference) 提供更高的浮點運算效能。

Volta 應 HPC 而 生，Tesla V100 未演先轟動

黃仁勳指出，傳統 HPC 目前只有不到 10% 有搭載加速器，市場成長空間極大，而 CUDA 已成 HPC 的基礎核心。Tesla V100 能支援語音助理、個人化搜尋與建議系統等高度精準 AI 服務，還能加速 HPC 與繪圖作業且具備擴充性；若將 DGX-1 AI 超級電腦連接八個 Tesla V100 GPU，可使深度學習能力再翻倍。此外，NVDIA 還另行開發名為「NVLink」的高速互連通訊介面以加快多個 GPU 之間、或與 CPU 的溝通，並與三星共同開發資料傳輸率達 900 Gbps 的 HBM2 DRAM 記憶體，為大型資料中心組建「HGX 參考架構」以推動 AI 雲端運算。

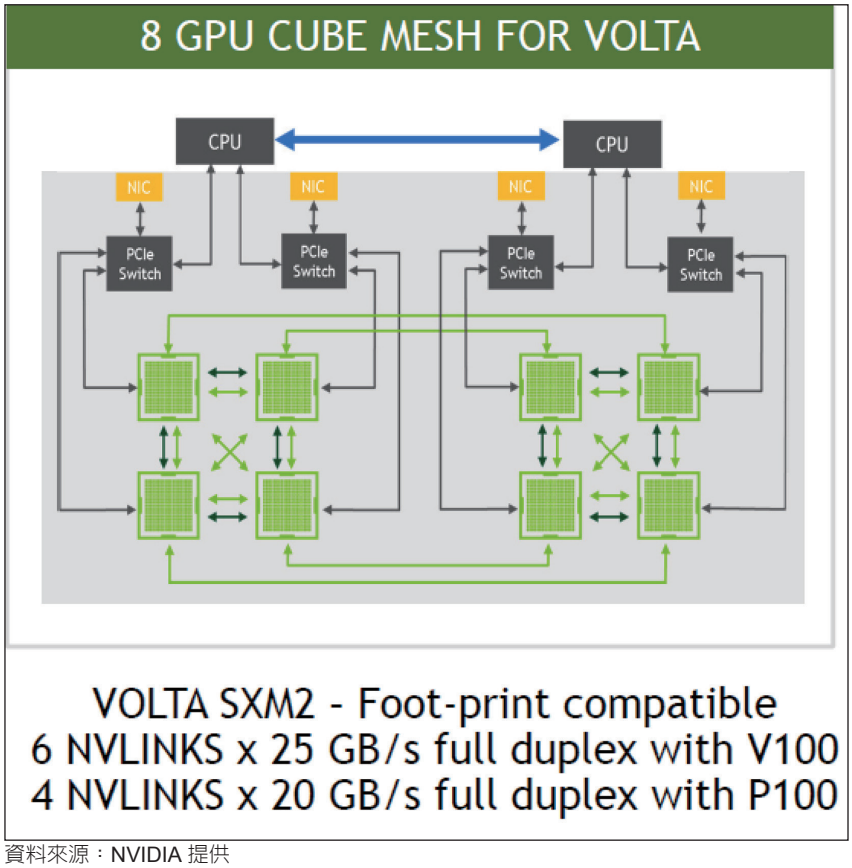
意識到公有雲市值已達 250 億美元，蘊含驚人商機；NVIDIA 早在 2010 年便與 Amazon AWS 推出首款針對 GPU 進行優化的雲端執行個體；Facebook 的 Caffe 2、PyTorch 與 Google 的機器學習、高效運算和資料分析，以及 Microsoft 的 Azure N 系列、Project Olympus 和 Cognitive 工具套件，亦見 NVIDIA 蹤跡，並相繼宣示將升級至 Volta。美國橡樹嶺國家實驗室將於明年問世的科研超級電腦 Summit，也以 Volta GPU 作為運算核心。中國市場亦大有斬獲，百度雲及智慧駕駛、騰訊的語音／相片／視訊及騰訊雲的深度學習平台，也是 NVIDIA 夥伴。

黃仁勳預期，AI 會將資訊注入 2,000 萬個雲端伺服器、上百億萬計的車輛及工業機器人，最終，高達一兆個物聯網 (IoT) 裝置與各種感測器將會智能監控一切，從人體心跳、血壓，到工廠設備的振動；

圖 2：DGX SATURNV 的運算能力可更快速訓練深度神經網路，創建更智能的 AI

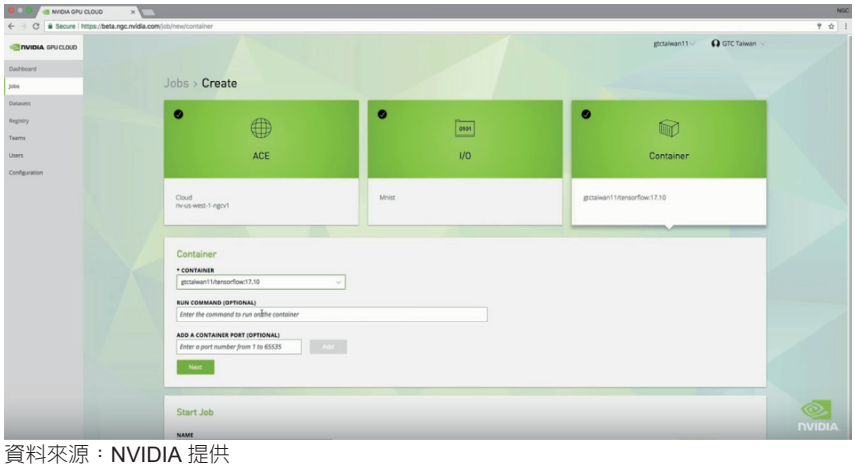
資料來源：NVIDIA 官網

圖 3：HGX 伺服器設計將八個 Tesla V100 GPU 加速器透過 NVLink 互連技術，組成混合式立方網絡 (hybrid cube mesh)



資料來源：NVIDIA 提供

圖 4：NGC 是 GPU 加速雲端平台，開發者可透過本地網路、Amazon EC2 或其他雲端平台提供商的深入學習框架，快速入門



資料來源：NVIDIA 提供

就算不上公共網域，也會在局網中運行，將收集到的資料匯入神經網路裡。雖然百度、騰訊和京東的資料中心也是 AMD EPYC 處理器的用戶，不久前更傳出阿里巴巴和百度有意採用 AMD Radeon Instinct GPU 加速器做深度學習的消息，對 NVIDIA 的 Tesla P100 GPU 可能造成威脅，不過 NVIDIA 回應，挾著多年生態系耕耘優勢，對自家產品仍信心十足。

雲端 NVDocker 容器，免去開發者編譯困擾

「深度學習的軟體堆疊相當複雜，包括運算、系統、網路資源分配、中介軟體、函式庫到各式演算法，且以等比級數的速度成長」，黃仁勳說。考慮到開發者有各自偏好的架構及工具，在編譯 (Compiler) 時或將遭遇版本相容性問題，NVIDIA 特針對「訓練」必要的模型及網路架構，為每個堆疊創造 NVDocker 容器；透過 CUDA 加速，將經過優化、測試的堆疊全數儲存於 NVIDIA GPU 雲端容器註冊表 (NGC)。開發者上網註冊就能下載、導入至內建 CUDA 晶片的資料中心、工作站或個人電腦。最重要的是，NVIDIA 承諾會永久維護，讓內容保持在最新狀態。

黃仁勳剖析，深度神經網路 (DNN) 結合多個演算式及上百萬個參數，是非常龐雜的檔案，可能應用在大型資料中心、也可能是小型機器手臂或單純的程式堆疊指令，而未來大部分的節點都將用於「推論」。推論裝置大爆炸的結果是：整個地球將被神經網路包圍，無遠弗屆；這些堆疊須在大大小小的不同應用平台運行，例如：麥克風、機器人、自駕車或超級電腦中心，將迫使各式網路不斷成長，故推論平台須具備可編程特性及擴展性以因應多元且不斷升級的網路架構，

而新架構、更深層的網路與神經網路層設計，又將持續增進卷積神經網路 (CNN) 效能。

AI 推論大爆發，運算之外……編譯也要加速

用於分類的循環神經網路 (RNNS) 以及長短期記憶 (LSTM) 的辨識語句與翻譯效能已超越人類，而生成對抗網路 (GAN) 利用一個可被訓練用於偵測的判別器、以及另一個用於製造事例來欺騙判別器的生成器，可完美產生影像、聲音並除噪。為方便不同裝置的編譯加速，NVIDIA 另推全球首款可編程推論加速器 TensorRT，會依據目標應用適度移除神經網路上不必要的元素，並善用指令將複雜架構重新堆疊融合 (Fusion)，執行多串流 (Multi- Stream)。第三代 TensorRT3 便是衝著 CUDA GPU 神經網路而來，利用 CUDA 深度學習指令創建運行時間並優化數值精準度、分層和張量。

黃仁勳認為，深度學習模型須先行將電腦最佳化，才能在低延遲、高吞吐量、低功耗、少量記憶體的情況下進行推論。不久的將來，資料中心內每個查詢與工作負載都將包含一個或多個 DNN 推論，而推論的吞吐量會直接反應在資料中心的總持有成本 (TCO) 上。例如，CPU ＋ TensorFlow 每秒可處理 140 個圖像，但 Tesla V100 ＋ TensorRT 每秒可處理 5,700 個、足足快了 40 倍，語音更快上 150 倍，意謂每個 V100 伺服器可省下約 50 萬美元的成本。

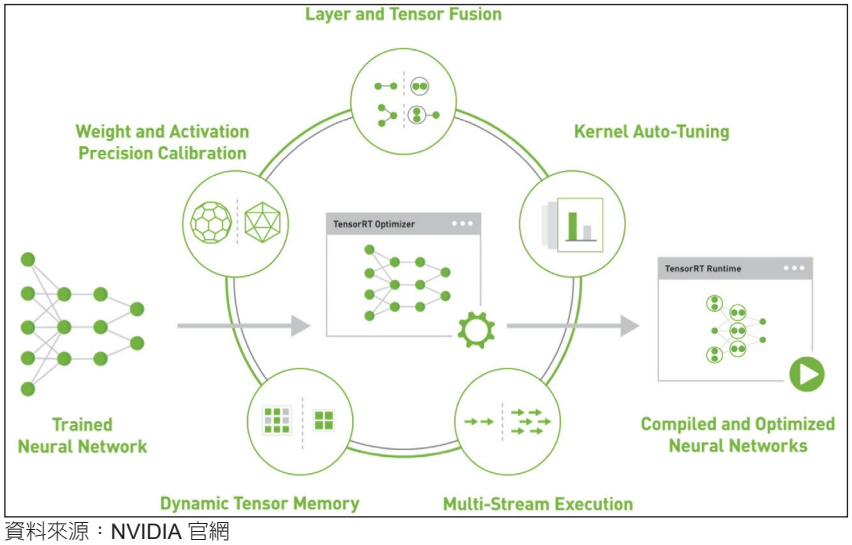
AI 下個重頭戲：自駕車

黃仁勳主張，透過價值功能取向的試誤及獎懲做強化學習，在不斷嘗試後，最終機器人一定能把任務學好；而將深度學習軟體與服務結合，NVIDIA 有信心做出史上第一輛沒有駕駛和方向盤的真正無人車。他預告：「自主機器世代即將來臨！自駕車就是第一個自主機器人。感測器、人工智慧與節能的 CUDA GPU 將為自主機器打開一個新世界」。為符合自駕車「Fail-Safe Operator」要求 (即使當機仍要正常運作)，NVIDIA 的策略是在資料中心裡訓練神經網路，用超級電腦模擬所有哩程——DRIVE PX PEGASUS 是全球首部專為量產自駕計程車所打造的電腦。

「這個超級運算資料中心僅有車牌大小，可放在後車箱做 ASIL 安全等級測試，320 TOPS 運算效能的功耗只有 500W」，黃仁勳介紹。最後，他總結 AI 趨勢：繼產業自動化後，將迎向「自動化的自動化」(機器自己寫軟體)；為此，NVIDIA 備有五大策略因應：

- JETSON 自主機器平台：超級電腦的 baby，用於終端設備；
- JetPack SDK：專為 JETSON 設計的整合軟體開發套件；
- DIGITS 應用程式：專為訓練神經網路、或導入預先訓練過的網路所設計；
- Isaac 虛擬實驗室：機器人在其中被創造並學習將任務做到盡善盡美；
- 深度學習機構：機器人模擬環境平台，可訓練打高爾夫球等高難度動作。透過機械原理、感測器與傳動裝置，搭配精準的環境模型與物理模擬，訓練過後的類神經網路可被下載並導入真實世界使用。 CTA

圖 5：TensorRT 可用於快速優化、驗證和部署訓練有素的神經網路，包括大型資料中心、嵌入式系統或汽車平台



資料來源：NVIDIA 官網