

嵌入式 AI 終端越來越“重”

■文：徐俊毅

早期的 AI 計算，大都依賴雲端強大的運算能力，終端或者邊緣節點主要的任務就是搜集資料。而大量 AI 應用的出現，資料中心的功耗以每年 1 倍的速度增加，雲端的資源正在變得緊張。

近些年，邊緣計算、AIoT 等一系列概念的提出，終端側的比重在逐年增加。

隨著嵌入式設備晶片功耗下降和計算能力不斷提升，包括 AI 語音和視覺功能正被越來越多地使用在終端設備，很多新應用正在被催生出來。

“AI 引領了許多應用發展，2020 年 AI/ML 在智慧家庭環境、語音處理、活動對象識別、人臉識別和虛擬實境等應用中開始採用。此領域包括工業自動化、深度學習在感測器遙測的應用、邊緣計算、智慧影片分析和認知計算



照片人物：恩智浦嵌入式處理器技術和業務戰略主管 Gowrishankar Chindalore

中的應用”恩智浦嵌入式處理器技術和業務戰略主管 Gowrishankar Chindalore 表示。

Gowrishankar Chindalore 表示，目前我們生活中的絕大部分應用都需要 ML 及邊緣計算解決方案，如語音輔助、人臉識別、符號偵測與定位、車牌偵測等。原因是，大多數這些應用場景需要即時的用戶互動與反應，需要本地能處理解決問題，而無需將數據傳遞到雲端並進行計算。這可解決延遲性、安全性和可靠性問題。

Arm 主任應用工程師張維良表示，在未來世界，end-point device（端點裝置）也能有資訊運算能力 (Intelligence)。Arm 的佈局就是在傳統 CPU 工作負載的效能以外，再多加一個象限，也就是 AI 與機器學習的表現。例如 Arm 近期發表的 Cortex-M55，它的 CPU 傳統工作負載表現大約在 Cortex-M4 等級，但其機器學習的表現則較 Cortex-M4 高出 15 倍。”

透過不同等級的單個 CPU，或者不同等級 CPU+NPU 的組合，可產出相當多樣的功能，包括從基本的震動、智慧感測、到關鍵字偵測、異常情況偵測、手勢、生物辨識到語音辨識，甚至到物件分類與即時辨識等功能。



照片人物：Arm 主任應用工程師張維良

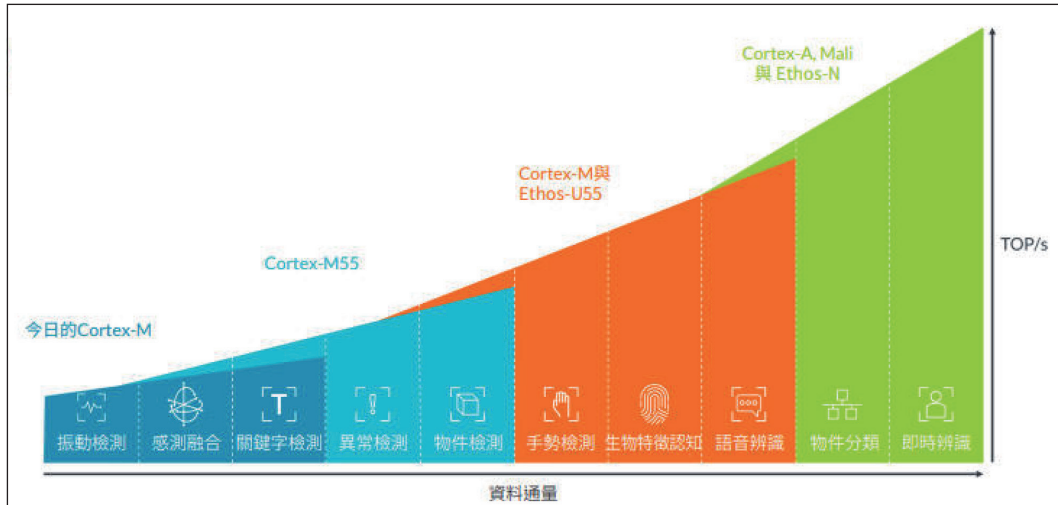
Arm 致力於以 AI 技術協助開發者、企業及產業，以創新的方式幫助改善人類生活。如圖，終端 AI 的應用在不遠的將來，將可應用在例如：

- 於手杖加上 360° 環景相機，解放視障者的行動限制
- 以耳道式裝置進行遠端健康監控
- 以聲控或動作控制進行機器操控
- 利用鐵路上的感測器預測與防止故障，並漸少誤點的可能性

萊迪思半導體企業行銷總監 Bob Nelson 表示，我們看到基於語音和手勢的應用成為了智慧家居領域的關鍵應用，因此在低功耗下支援語音人工智慧和計算機視覺變得非常重要。

除了節約流量和成本的考慮，

圖說：Arm 橫跨 AI 運算領域的解決方案



圖片來源：Arm

大量需要互動，需要多感測器資訊融合的應用需要在終端側即時完成，比如活體檢測這種，此外，出於保護隱私的意識，用戶希望儘量少地向雲端提交原始資料。這些因素都在促使雲端下的應用快速成長。

統計顯示，在 2017 年的 AI 應用中，雲端占了 80%，預計到 2024 年，終端側會到 60%，雲端則會降到 40%。需要說明的是，這個數字並非表示雲端的重要性下降，而是雲端寶貴的計算力將更多地留給那些計算密集型的應用。

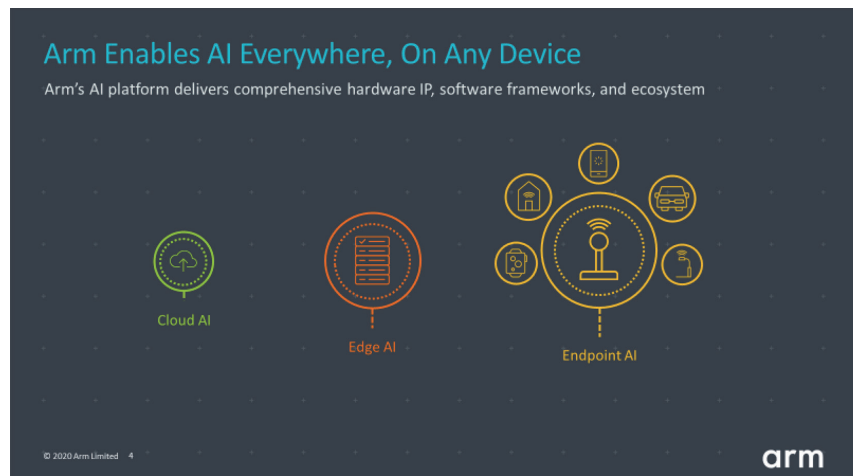
嵌入式 AI 終端進化加速

任何有網路的人都可以像學者們那樣進行最前沿的 AI 研究，做實驗像修改代碼那樣便捷，而租賃雲端的 GPU 僅需幾美分 / 小時，但並非所有的事情都要上雲端處理。

萊迪思企業行銷總監 Bob Nelson 認為：並非所有人工智慧

處理都需要在雲端進行。事實上，許多常見的人工智慧應用可以在網路邊緣執行，以減少數據延遲和與雲端相關的成本。例如，一隻貓經過智慧門鈴，門鈴將貓的圖像數據發送到雲端進行分析。但是，如果門鈴本身能夠識別出一隻貓，那麼它就不必將圖像上傳到雲端進行分析，從而節省成本並提高數據隱私。當門鈴確實註冊了需要進一步分析的人像或其他圖像時，門鈴就

圖說：從雲端到終端 Arm 提供完整 IP 和全生態系統



圖片來源：Arm

可以將數據上傳到雲端。

”AI 已完全改變運算的樣貌。AI 開始於雲端運算，以龐大的演算法與海量的資料集以建立複雜的訓練模型。隨著資料收集與累積越來越快、越來越多，運算已逐漸從雲端遷移至邊緣與端

點。“Arm 主任應用工程師張維良說。

超級運算巨擘如 Amazon，已運用雲端運算來預測在其網路中數以百萬計的伺服器的使用。伺服器端的運算，可在訓練與推論方面，卸載一下雲端運算的工作量。邊緣運算的演進相當快速，變化也相當大。而同樣的情況也將發生在端點運算。

在嵌入式 AI 終端，由於支持的設備數量不多，因此僅需少量的原始數據和簡單的算法，就能得到推論，相比雲端具有低遲、低功耗且能保護隱私的特點。

在安防監控應用中，以往監視器收集到影像必須在機房進行辨識，現在若在攝影機端具有 AI 功能，即可在影像收集的當下進行辨識。

在熱門的智慧音箱應用，音箱與喇叭具備 AI 語音助理功能後，帶動智慧家庭的呈現。透過全向性麥克風收集到的語音，即可即時判讀、回應與執行指令。

及時語言翻譯機 — 全球廣大語系眾多的詞彙過往也是透過雲端 AI 進行。譬如 Google 的離線翻譯引擎，僅需 20 多 M 的存儲空間，就能夠實現語音辨識和轉移，目前的正確率正在接近 90%。

應用場景是設計解決方案的基本

現在的 AI 晶片技術已有以下的重要進展，處理超高影像解析度 (Super Resolution)、準確的物件偵測 (Object Detection)、影像分類功能 (Images Classification)，快速的語音翻譯功能 (Speech Translation)。

張維良表示，Arm 認為應以應用場景作為設計解決方案之基本。以特定應用場景內，數據收集的狀況以及人們需要哪些具有 AI 裝置的協助，再發展以個人使用者終端的 AI 產品。

圖說：NXP 智慧家庭概念圖



圖片來源：NXP 網站

例如，智慧手機應用：AI 應用為使用者帶來更好的使用經驗。例如超高影像解析度能讓使用者掌握影像細節，大幅提升觀賞影片的體驗；Avatars 效果能讓使用者在自拍時使用有趣的即時動畫效果，增加人際溝通的樂趣；相機的智慧夜拍模式，則讓光源控制更輕鬆，在低光源下拍攝也能媲美日拍一般清晰；此外運用生物辨識技術的臉部辨識與解鎖功能，能讓消費者使用手機付款時更安全更便利。

智慧家庭 / 智慧電視應用：除了超高影像畫質能提升觀影體驗，AI 功能也能強化整體的聲光環境。例如能偵測影視內容所需，調整或放大聲量；或者對應室內光線環境，調校顏色對比與鮮明度；還能夠為家中有小朋友的家長把關，當偵測到電視前方有小朋友，電視會自動停止播放不適合的影視內容。

邊緣節點以及 AI 終端產品 IC 和方案優先需要考慮哪些的技術因素？

機器學習、決策能力，功耗和尺寸，低延時，快速回應能力，軟體演算法，隱私保護，資料安全，個性化設計都是設計人員需要考慮的重要因素，那麼應該為這些因素排序呢？

ARM：功耗和尺寸，資料安全性的首要

■ PPA(Power, Performance, Area) 功耗，在同樣的電力使用量下，能執行更多功能及運算，展現更高的續航力

● 算力 / 效能表現 — 也代表機器學習的應用以及其決策的準確性

● 尺寸 / 外型 — 更小、更輕巧的尺寸與外型，讓使用及部署都更為便利

■安全 (Safety & Security)

- 隱私保護與資料安全，是 AI 終端產品不可妥協的要素，更是物聯網 / 智慧物聯網能否大規模部署、促成數位轉型的關鍵。

Arm 提供一系列安全相關解決方案：

■Arm TrustZone 主要防護軟體攻擊。

■Arm Cryptocell/ Cryptoisland 主要防護物理攻擊，或者 Side Channel Attack。

■PSA(Platform Security Architecture) — 防護架構，這是一個架構能從 System 角度進行防護，循序漸進帶領夥伴如何從底層 IP、硬體開始直到軟體，建立防護合適的安全防護等級。從分析、定架構、實施、最後到裝置認證。

NXP：雲和端發揮各自優勢，保證資料安全性，提升 ML 和決策能力

AI 晶片發展趨勢是使用 NN/ML 加速度器的安全處理器，該處理器連接到各種感測裝置，在邊緣執行 AI/ML 推理，並連接到雲端以執行大數據分析、管理和操作。

雲端趨勢在於強大的計算和圖形處理能力，並提供安全的客戶解決方案，應用和設備管理環境。

邊緣端的趨勢則是保障隱私，更好的用戶體驗，高可靠性的處理器需求。降低雲端成本，包括頻

寬、處理和存儲。

■訓練 (如 TensorFlow、Caffe 等) 趨勢則是通過量化等為更精確的訓練模型生成提供動力。訓練模型是經過非常縝密計算而且時間非常密集。需要更多資料遷移和 AI 引擎被創造出來為 RTOS 或低功耗電池平臺提供輕量與靈活的訓練模型，以用於邊緣計算應用。

■推論 (如 Arm 推出的 Arm NN、Arm CMSIS-NN、OpenCV、TensorFlow 精簡版、Glow Compiler 等) 的趨勢則以提高卷積、匯集、動作等功能。

lattice：功耗和尺寸，機器學習和決策，資料安全性，低延時和快速回應

在網路邊緣運算應用中，功耗和尺寸是優先考慮的問題。機器學習和決策能力在關鍵應用發揮作用，特別是在工業領域。阻止未經授權的闖入者，保證資料安全同樣是不可或缺的功能。

在需要即時性的關鍵任務的應用中，保持低數據延遲非常重要。

不難看出，功耗和尺寸、安全性仍然是不變的主題。在未來的嵌入式 AI 應用中，平衡功耗和 AI 算力對設計人員將會是不小的考驗。越來越多的即時性任務出現，讓低延時特性變得十分重要性。設計人員不僅要充分理解應用場景，還要顧及諸多關鍵因素，才能使產品更具競爭力。

為複雜應用而生的 Chiplet

Chiplet 是半導體界近期的熱門話題，Chiplet 模式是通過 die-to-die 內部互聯技術將多個模組晶片與底層基礎晶片封裝在一起，構成多功能的異構 System in Packages (SiPs) 晶片的模式，一種很典型的搭積木方式。相比傳統的 SoC 方法，Chiplet 是一種短週期、低成本的集成協力廠商晶片 (例如 I/O、存儲晶片、NPU(神經網路處理器)等) 的技術，不僅讓晶片設計更加靈活，還可以大大降低晶片設計製造成本，並加快上市時間。

“Chiplet 概念就是把一顆 SoC 分成好幾個晶片，讓每顆晶片的效能維持，但成本可以更低，是 SoP(System on Packaging) 的概念。”台積電研究發展組織系統整合技術副總余振華在 SEMICON China 2020 的公開演講中說。

Bob Nelson 表示，萊迪思投入了當前基於 Chiplet 的設計趨勢。作為其開放運算專案工作的一部分，萊迪思開發了一種用於安全系統控制和管理的晶元板，它可以插入一個基板，從而實現安全 SiP 解決方案的原型設計。

現在，chiplet 模式還處於早期發展模式，有很多技術障礙需要跨越，目前主要是大廠在使用和驗證，但隨著技術的成熟，成本進一步可控，這項技術在嵌入式 AI 這樣的靈活應用中將有巨大的成長空間。

現階段設計人員仍然將圍繞應用場景和終端產品展開設計。CTA