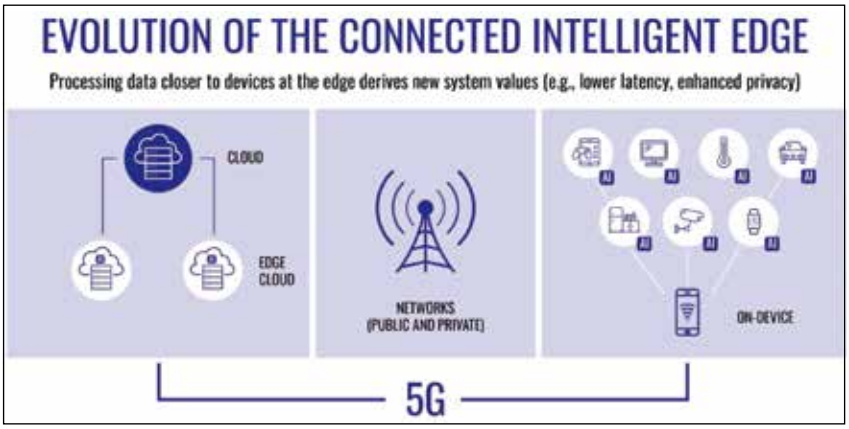
# AI Edge 核心誰作主？

■文：任芷萍

物聯網 (IoT) 已迅速普及日常生活，邊緣運算 (Edge Computing) 的重要性亦與日俱增。另一方面，研調機構 ResearchAndMarkets 指出，目前全球出貨的所有晶片組中有 85% 皆配備了 AI，到 2026 年，將有超過 63% 的電子產品將擁有某種形式的嵌入式智能。瑞薩電子 (Renesas) 在新近發佈的一篇文章中宣示：由物聯網、5G 連接和人工智慧 (AI) 所構成的 AIoT 趨勢，正在推動收集、儲存、處理、分發、保護和驅動數據的方式發生轉變，以便將其轉化為可從中學習的可操作情報，而互聯「智慧邊緣」(AI Edge) 將是大勢所趨。

圖 1：智慧邊緣聯網演進




資料來源：https://www.renesas.com/us/en/blogs/new-convergence-iot-ai-and-5g-bring-actionable-intelligence-factory-floor

## 瑞薩電子：分佈式邊緣運算看旺，工業物聯網動力足

瑞薩物聯網及基礎設施事業部執行副總裁兼總經理 Sailesh Chittipeddi 指出，2020 ～ 2022 年期間，其物聯網業務的年複合成長率 (CAGR) 大幅躍升了 79%，這與他們陸續收購 Intersil、IDT、Dialog Semiconductor 和 Celeno 等公司，集成感測器、連接、驅動和電源管理四大能力而大幅提升嵌入式處理器的實力功不可沒；尤其，工業物聯網 (IIoT) 更是潛力無窮。在靠近邊緣設備的地方處理數據將產生新的系統價值，如：更低的延遲、增強的隱私。這種巨變需從集中式、基於雲端的架構，轉向分佈式、基於邊緣的設計。

Chittipeddi 表示，使用微型機器學習 (ML) 的微控制器 (MCU) ／微處理器 (MPU) 節點來定義端點、加速數學模型並提高深度神經網路 (DNN) 性能。當 IoT 端點節點的創建以每年 85% 的 CAGR 呈爆炸式增長時，長線將在預測性維護、快速缺陷檢測、生物特徵識別和資產跟蹤等領域，開闢新的市場和收入來源。有鑑於此，瑞薩在 2022 年再併購以工業演算法聞名的 Reality AI 公司，以結合先進訊號處理、數學建模與 AI，建構能在 16 ～ 24 位元的嵌入式處理器上實施的機器學習模型。與此同時，亦投資 Syntiant 和 Arduino 等公司及贊助 Tiny ML Consortium，目前已在 AIOT 相關生態系統擁有 200 多個技術合作夥伴。

## 晶心科技：邊緣運算將是 RISC-V 與 AI 的舞台

2016 年起專注於開發 RISC-V 矽智財 (IP) 的晶心科技 (Andes) 認為，未來的邊緣運算將是 RISC-V 與 AI 的舞台。晶心科技總經理暨技術長蘇泓萌強調，對於上述應用，RISC-V 可擴充指令的這個特點，相當受到客戶重視。「可提供客製化串接機制，讓處

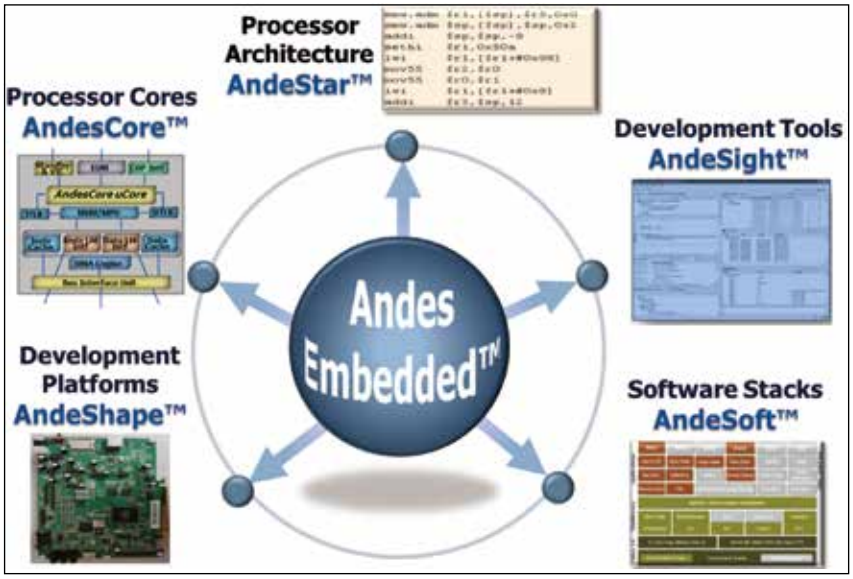
照片人物：晶心科技總經理暨技術長蘇泓萌 ( 左 )、晶心科技資深技術行銷經理王庭昭 ( 右 )

理器與加速器之間能有效溝通及傳輸資料，減少延遲，是可擴展 RISC-V AI 子系統的最大優勢」，他說。晶心科技資深技術行銷經理王庭昭表示，邊緣運算越多 AI 應用，包括：視覺、聲音／語音及經由感測器融合所產生的任何訊號，但其間所需運算力相差甚大。

考量到受限於運算力與電力有限，晶心除了面向物聯網的緊湊設計、安全及低功耗新推 32 位元控制器 AndesCore D23 以支援最新 RISC-V 擴充指令外，另在通用型處理器之外，鎖定深度 ML 推出可擴展的 AnDLA I 系列多心核心加速器。加速器旨在供邊緣設備做神經網路 (NN) 推論運算，可和自家 AndesCore 25/27/45/60 處理器系列搭配使用，用來擴充運算、資料交換及控制訊號的能力。王庭昭總結，RISC-V 數位訊號處理器 (DSP) ／單一指令多重資料 (SIMD) 和向量處理器 (Vector Processor) 可為不同 AI 應用和分衆市場提供高效運算力。

晶心另有深度學習 (DL) 加速器 AnDLA，為其可擴展 RISC-V AI 子系統的省電運算扮演關鍵推手。它具有用於矩陣乘法、卷積、池化功能等的硬連線處理單元，是一種適用於邊緣設備和端點的高效、成本敏感的 AI 解決方案，而 AndeSight IDE 和 Andes NN 軟體開發套件 (SDK) 開發工具則可為開發者帶來最終運行時間的效能及開發效率。特別一提的是，為因應神經網路所需，有些系統廠商亦會選擇現場可編程邏輯閘陣列 (FPGA) 上陣或作為運算加速器使用，MCU 供應商甚至已悄悄開啓內嵌網路處理器／神經處理器 (NPU) 的風潮。

圖 2：晶心科技產品與解決方案



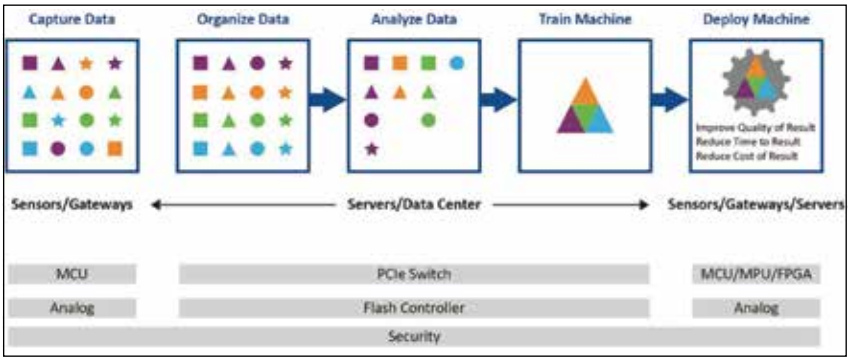

資料來源：http://www.andestech.com

## Microchip：資源受限的高負載運算，中檔密度 FPGA 是優選

擁有 MPLAB 開發生態系為強大奧援的微芯科技 (Microchip) 主張，智慧邊緣若考量到在最低的功率和熱餘量下、資源受限的環境中進行高負載運算，中檔密度 FPGA 將脫穎而出。Microchip 最新的視訊和圖像處理解決方案組合由「PolarFire」FPGA 視訊和成像套件提供支持。新套件使開發者能為利用 AI 和高解析度成像的邊緣應用實現功耗最低、外形尺寸最小的智能嵌入式視覺系統，還可協助使用基於 RISC-V 的 PolarFire 系統單晶片 (SoC) 和 MATLAB/Simulink 進行 FPGA 硬體在環 (Hardware-in-the-loop, HIL) 模擬。

機器視覺、熱成像、視訊監

圖 3：用於機器學習數據流的 Microchip 矽平台




資料來源：https://www.microchip.com/en-us/solutions/machine-learning

控、機器人技術、邊緣機器學習推理和人機介面 (HMI) 皆仰賴低功耗的相機和顯示器，且須支援高速接口以及數據和設計安全性，進而保護 IP。FPGA 固有的並行處理和高速 I/O 功能，使其成為高解析度成像和機器學習演算法所需的高數據吞吐量的理想處理平台。根據 Microchip 的說法，PolarFire SoC 和 PolarFire FPGA 的總功耗比競爭對手的中端 FPGA 低 30～50%、靜態功耗低五到十倍，使其成為「密集型邊緣」的理想方案，包括部署在熱和功率受限環境中的設備以及用於加速設計的工具。

PolarFire FPGA 亦可作為 Microchip 旗下 MCU／MPU 的加速器搭配使用。挾著強勢開發工具為後盾：MPLAB X 整合開發環境 (IDE)、MPLAB XC C 編譯器、MPLAB Data Visualizer—— 此 MPLAB 插件可追蹤應用程式的運行時間並分析功耗，且可使用集成 ML 插件來捕獲數據並傳輸到合作夥伴平台。加上 VectorBlox 加速器 SDK 相助，可對高能效 NN 進行編程；另附帶建構基於 PolarFire 視訊套件的 AI 相機平台，可評估不同的卷積神經網絡 (CNN)。其軟體工具包允許使用 TensorFlow、Keras、Caffe 和 ONNX，以及 TinyML 和 TensorFlow Lite 框架。

## NXP：MCU 內建 NPU & DSP，提高智慧運算預測性

恩智浦半導體 (NXP) 的 MCX N 系列，用於機械手臂、智慧電梯、智慧門鎖這類對智慧運算有更高預測性要求的高性能、低功耗微控制器，便已首次將 NPU 和 DSP 置入 MCU。N 系列中首先登場的是 MCX N94x 和 MCX N54x MCU 系列，多核架構設計在提高系統性能的同時還兼顧降低功耗，可實現性能與功耗的完美平衡。MCX N94x 適合工業應用，具有更廣泛的類比和電機控制外設，而 MCX N54x 則是聚焦消費和物聯網應用，集成帶 PHY 的高速 USB、SD 和智慧卡介面等諸多外設。兩者皆基於高性能雙核 Arm Cortex-M33。

開發人員可使用這些內核和加速器的任意組合來完成具體任務，無需提高 MCU 的時脈速度或增加功耗。雙核架構允許並行運作應用程式或根據需要關閉單個內核以降低總功耗，例如，在物聯網設備的空中 (OTA) 更新期間，主要 M33 內核可處理系統安全，而第二個從核執行控制功能。MCX N 系列的發佈，亦是恩智浦自主研發的 NPU 初次亮相，以實現邊緣的高性能和低功耗智慧。與只使用 CPU 內核相比，內置 NPU 的 ML 輸送量估可提高 30 倍，使 TinyML 在資源和功率受限的邊緣設備上展現超凡的運算力，以實現複雜的深度學習模型。

例如，為門禁控制添加人臉和語音辨識功能、為家庭安全系統創建電池供電的玻璃破碎探測器、為電機控制預測維護開發振動感測器、設計配備生物感測器的智慧穿戴裝置等。順帶一提，MCX N 系列 MCU 集成了 EdgeLock 安全子系統，可安全啓動不可變的信任根、實現硬體加速加密、主動和被動入侵偵測以及電壓和溫度篡改檢測，支援現場更新和線上傳輸，並可防止遠端原始設計製造商 (ODM) 過度生產。為更好地分析環境並實現本地智慧決策，恩智浦認為，先進的處理、機器學習能力並結合高速連接，是下波邊緣運算應用程式的關鍵要求。

於是，近日再發佈內嵌 NXP「eIQ Neutron」NPU 與圖像訊號處理器 (ISP) 的 i.MX 95，是其第一個 i.MX 應用處理器 (AP) 系

圖說：恩智浦推出 i.MX 95 系列應用處理器，提供安全可擴展的人工智慧邊緣平台


資料來源：恩智浦公司

列，擬於今年下半年為主要客戶供樣。其嵌入式 NPU 與 MCX N MCU 採用相同的 NPU 基本硬體架構，允許用戶根據需要進行擴展，支援從汽車連接、智慧駕駛座艙 (eCockpit)，到工業 4.0 和物聯網平台的邊緣應用。其中，eIQ Neutron NPU 便是作為連通多個攝影感測器或智能網路攝影機的視覺處理之用，而 eIQ ML 軟體開發環境則集成了用於建構機器學習的資料庫和開發工具。

## ST：分佈式邊緣運算勝在延遲、頻寬、數據主權、分析效率

選擇物聯網 MCU，還須考量連接協定、開發套件、資安機制與 AI ／ ML 等軟體支援。在 MCU ／ MPU 根基深厚的意法半導體 (ST) 算是很早投入智慧邊緣的先驅廠商之一，相關方案已完全集成到 STM32 生態系，感測器數據由處於系統核心的高能效 STM32 MCU 和 MPU 處理。相關產品包括：

- ●通用型無線、超低功耗、主流和高性能 MCU；
- ●具有 AI 加速器的 GP MCU，針對帶有嵌入式神經處理單元的低功耗推論運算進行優化 ( 型號定為 STM32N6，預估今年上市 )；
- ●適用於超低功耗應用、具有集成處理單元的 MEMS 感測器；
- ●用於 OpenSTLinux 上啓用 AI 的雙核 MPU 系列。

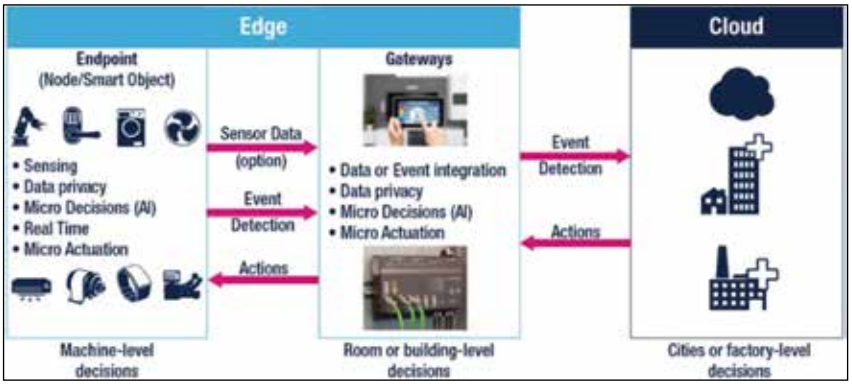
絕大多數 STM32 開發板皆配備感測器及連接到原型智慧邊緣的應用程式，可用於開發嵌入式 AI 解決方案。例如，基於 STM32L4 MCU，可建立具有擴展感測器和連接性物聯網節點的 STEVAL-STLKT01V1 SensorTile 開發套件——配備 3D 加速度計和陀螺儀、磁力計、飛時測距 (ToF) 趨近感測器、數位麥克風等微機電 (MEMS) 感測器，以及藍牙 (Bluetooth) ／近場通訊 (NFC) ／ Sub-GHz ／ Wi-Fi 等無線連接功能，適合音頻場景和活動識別應用。

另有適用於工業物聯網 (IIoT) 預測性維護和狀態監測應用之無線工業節點 (STEVAL-STWINKT1) 的 STWIN SensorTile；用於狀態監測的工業感測器評估套件、專為溫度和振動監測設計的 STM32 STEVAL-PROTEUS 1 開發板；以及專為穿戴裝置或運動感測應用設計的 STM32L5 MCU 開發板。ST 重申，分佈式邊緣運算可顯著降低傳輸延遲、所需頻寬和雲端伺服器的處理能力，且賦予用戶「數據主權」，將個人源數據經過預分析後再提供給具有更高級別解釋的服務商。

## STM32Cube.AI 新增預訓練 ANN ＋遠程基準測試模型

ST 為旗下市佔廣大的 STM32 MCU 開發映射和運行預訓練的人工神經網路 (ANN)，面向 Edge AI 開發人員提供免費工具 STM32Cube.AI，以便部署符合坊間主流 AI 框架、經

圖 4：智慧邊緣可提供更高效的端到端解決方案




資料來源：https://www.st.com/content/st_com/zh/about/innovation---technology/artificial-intelligence.html

過訓練的神經網路模型，可經由 STM32CubeMX 環境中的圖形介面和命令行使用，也可在 STM32Cube.AI Developer Cloud 線上取得。透過優化 AI 模型的主記憶體使用和推理時間，旨在打造最高效的 MCU 神經網路免費代碼生成器。近日，ST 再推出「支援線上訪問」補充版本——MCU AI Developer Cloud，方便開發者對 STM32 板上的智慧邊緣模型進行遠程基準測試。

開發者還能訪問「STM32 model zoo」可訓練的深度學習模型和演示儲存庫以加速應用程式開發，目前支援用例包括用於活動識別和跟蹤的人體運動感測、用於圖像分類或對象檢測的電腦視覺、用於音頻分類的音訊事件檢測等。它是預訓練 ML 的集合，可自動生成針對 STM32 優化的入門套件且可在 GitHub 使用。開發者可透過雲端取得 STM32 電路板並定期更新，以便遠程測量優化模型的實際性能。STM32Cube.AI Developer Cloud 已經過多家嵌入式開發客戶的測試和評估，現可供 MyST 註冊用戶免費使用 (https://stm32ai-cs.st.com)。

為擴大開發工具的功能並加速嵌入式 AI ／ ML 開發專案，ST 還推出 NanoEdge AI Studio 及 STM32Cube.AI 的升級版本：前者適合不需開發 NN 的應用，須與 STM32 MCU 及內建 ST 的嵌入式智慧感測器處理單元 (ISPU) 的 MEMS 感測器搭配使用；後者則是 STM32 人工智慧模型優化器與編譯器，適合 NN 研發，最新釋出的 7.3 版本已被完全整合至 STM32 生態系，將預先訓練好的 NN 轉換成能在 STM32 Arm Cortex 內核心 32 位元 MCU 上運行的 C 語言程式碼；升級版還可根據性能需求和記憶體容量調整現有神經網路，或平衡優化最佳效果。

## 「智慧邊緣」點火下一波處理器／控制器競逐

繼整合無線連接功能之後，AI ／ ML 絕對是下個 MCU 致勝點。芯科科技 (Silicon Labs) 就借助完整的多協定 SoC 產品組合、廣泛的開發工具選擇及跨無線標準的廣泛專業知識，將 ML 引入任何應用程式。ML 開發工具依開發人員的經驗，由淺到深依序有 ML Solutions、ML Explorer 和 ML Experts 三種；考慮到缺乏 ML 經驗的開發人員，Silicon Labs 還與 Sensory 合作開發關鍵字和喚醒詞應用程式，並與 Micro.ai 合作進行異常檢測。

此外，EFR32xG24 開發套件允許開發人員在目標設備上加載和運行示例項目，該開發套件運行使用 TensorFlow 引擎並具有集成 ML 模型的嵌入式應用程式。EFR32xG24 開發套件 (xG24-DK2601B) 是一個緊湊、功能豐富的開發平台，提供開發和原型無線物聯網產品的途徑；該開發平台支持高達 +10 dBm 的輸出功率，包括對 20 位元 ADC 的支持以及 xG24 的 AI ／ ML 硬體加速器等其他重點功能。上述所有軟體都將在此開發套件上運行。

邊緣設備最常見的用途之一就是：透過感測器監控環境變數，而位於設備核心的 MCU 功能也越見強大。現今的 MCU 多是以 SoC 方式呈現，內嵌一個或多個 MPU 以讀取數據並進行大量運算、儲存並呈現最終結果，當中每個內核都能獨立執行指令，可分別處理不同任務。基本上，多核系統由於執行指令快且內核可在不工作時中斷電源，功耗會顯著低於單處理器內核，因此當紅 MCU 幾乎皆採多核系統。

知名電子零件經銷商 Digi-Key 表示，多核心 MCU 常見配置有兩種：對稱 vs. 非對稱式。對稱式核心配置包含兩個或多個完全相同的處理核心，而非對稱式核心組合的方式相當多種，視應用及設計的需求而定。多核心 MCU 好處是能將應用分成多個執行域——不同執行域有助於精準控制應用的效能、功能和電力需求。( 參閱：https://www.digikey.tw/zh/articles/why-and-how-to-get-started-with-multicore-microcontrollers)。

智慧邊緣設備方興未艾，執掌帥印的控制／處理核心，無疑是下個兵家必爭之地。CTA