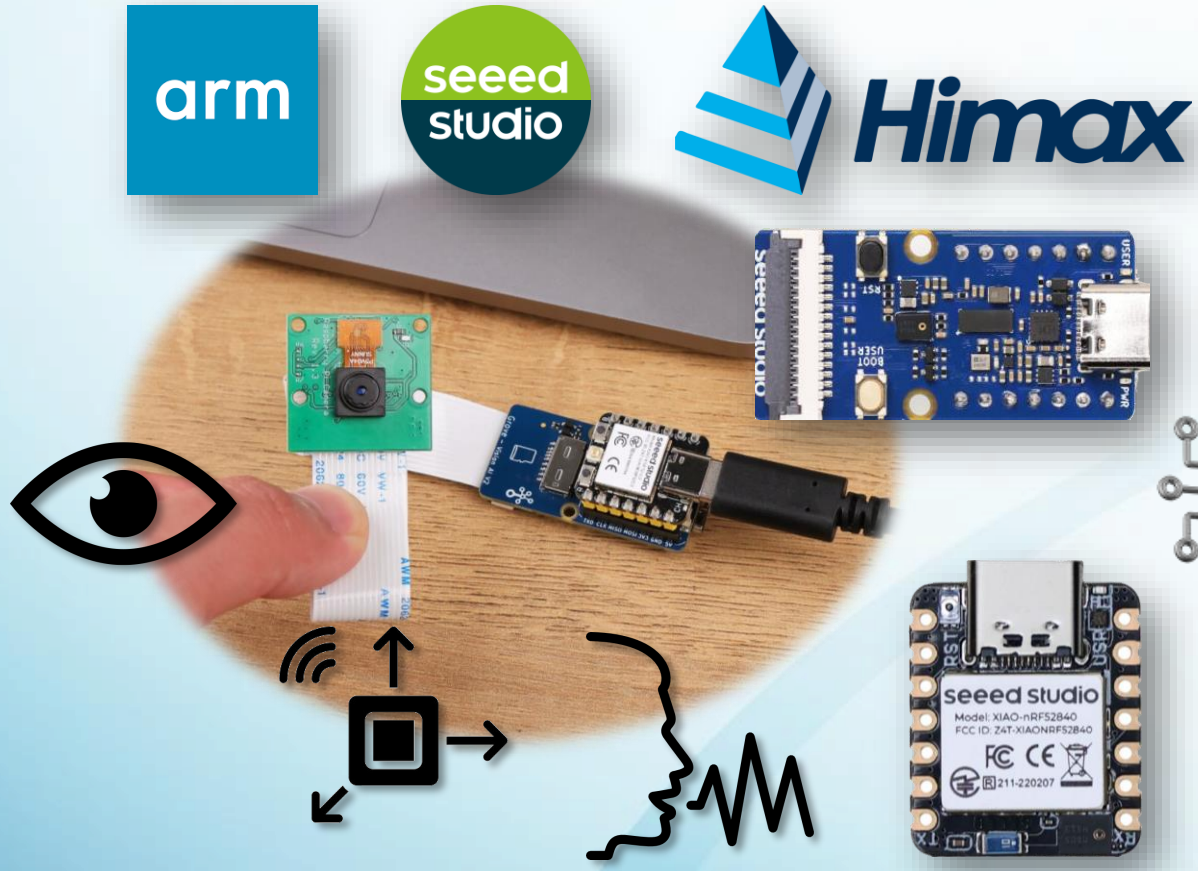
OmniXRI TinyML 小學堂 2025



Cortex-M Processor

Ethos-U MicroNPU

【第 4 講】

arm 單晶片加速運算

歐尼克斯實境互動工作室 (OmniXRI Studio)
許哲豪 (Jack Hsu)

簡報大綱

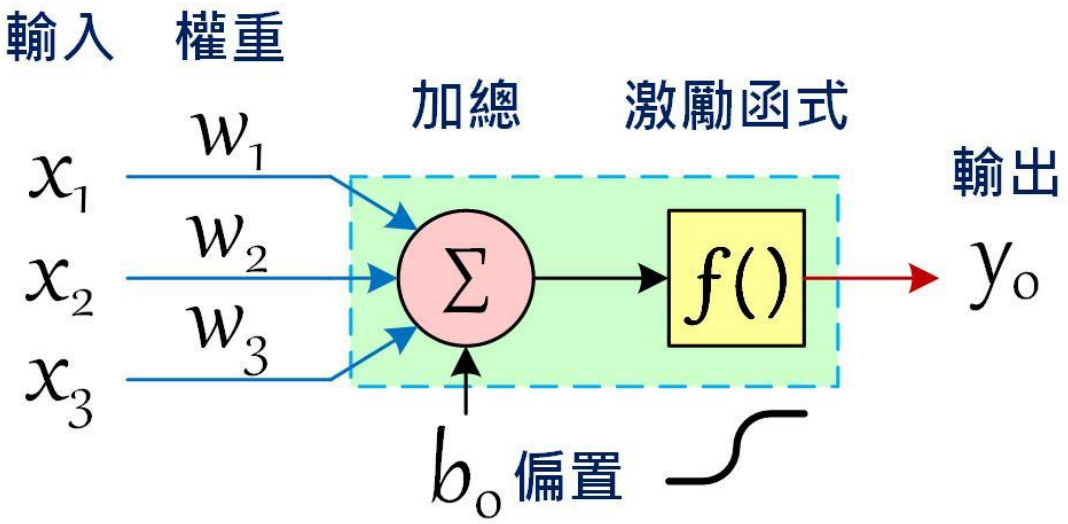
- 4.1. 常見硬體加速手法
- 4.2. DSP (SIMD) 指令集
- 4.3. Helium (MVE) 指令集
- 4.4. Ethos-U MicroNPU

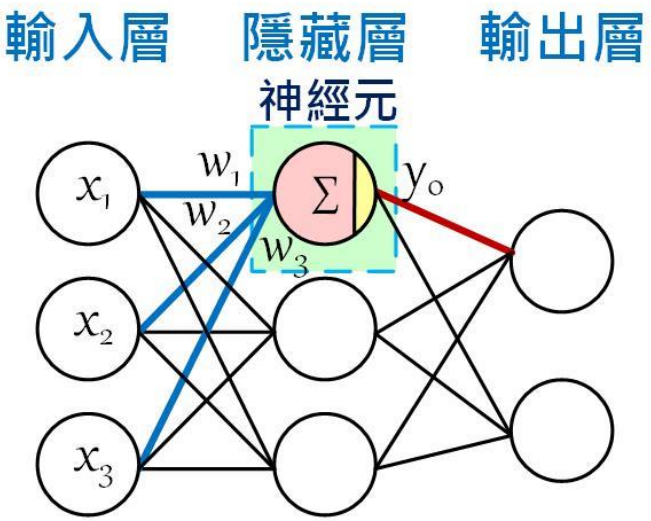
本課程完全免費，請勿移作商業用途！
歡迎留言、訂閱、點讚、轉發，讓更多需要的朋友也能一起學習。

完整課程大綱： https://omnixri.blogspot.com/2025/03/omnixri-tinyml-2025-0.html
課程直播清單： https://www.youtube.com/@omnixri1784/streams

4.1. 常見硬體加速手法

神經網路基本計算量

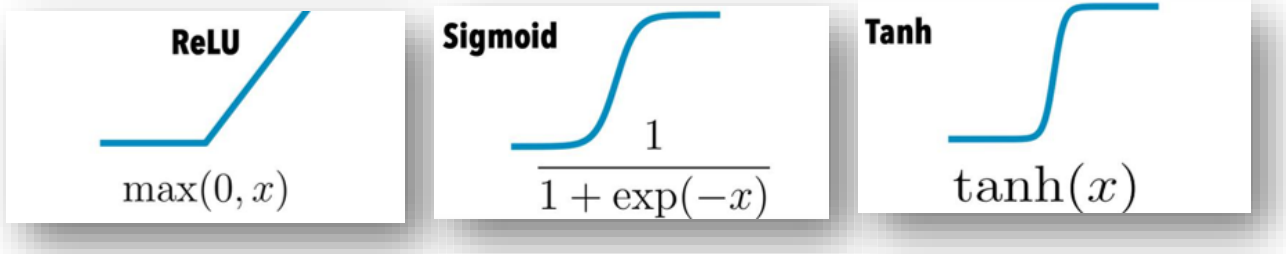




神經網路
(Neural Network, NN)

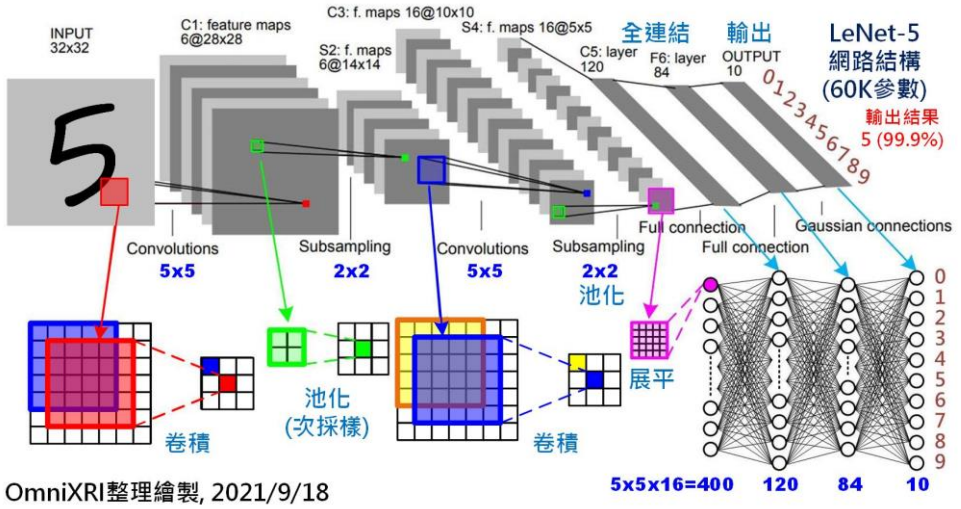
$$y_o = f(x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + b_o)$$

激勵函式



資料來源：https://omnixri.blogspot.com/2022/10/mcunputinyml.html

為什麼要加速？



Layer Name	Input W×H×D	Kernel W×H×D/S	Output W×H×D	Params	Mults
C1: conv2d	32×32×1	5×5×6	28×28×6	1×5×5×6+6 =156	28×28×1×5×5×6 =117,600
S2: pool/2	28×28×6	2×2/2	14×14×6	0	0
C3: conv2d	14×14×6	5×5×16	10×10×16	6×5×5×16+16 =2,416	10×10×6×5×5×16 =240,000
S4: pool/2	10×10×16	2×2/2	5×5×16	0	0
C5: conv2d	5×5×16	5×5×120	1×1×120	16×5×5×120+120 =48,120	1×1×16×5×5×120 =48,000
F6: conv2d	1×1×120	1×1×84	1×1×84	120×1×1×84+84 =10,164	120×84 =10,080
F7: conv2d	1×1×84	1×1×10	1×1×10	84×1×1×10+10 =850	84×40 =840
			Total	61,706	416,520

以CNN LeNet-5 為例

- 輸入影像尺寸單色 32x32 pixel，模型參數6萬個，推論一次需416,520次乘加。(整數 or 浮點數)
- 訓練模型至少數千次到數十萬次推論及修正。
- 主要計算集中在**乘加（MAC）**及激勵函式複雜度。

$$y = a * x + b$$

- 整體推論速度還會卡在資料搬移速度。

硬體加速手法 – 提高工作時脈

工作時脈(Clock) MHz, GHz

工作週期(Cycle) uS, nS

1MHz (10^6) ⇔ 1us(10^{-6})

- 一個指令週期不一定等於一個工作週期。
- 不同指令所需工作週期不同。
- 浮點數計算通常工作週期會較長。
- 執行速度 MIPS 表示每秒可執行百萬指令

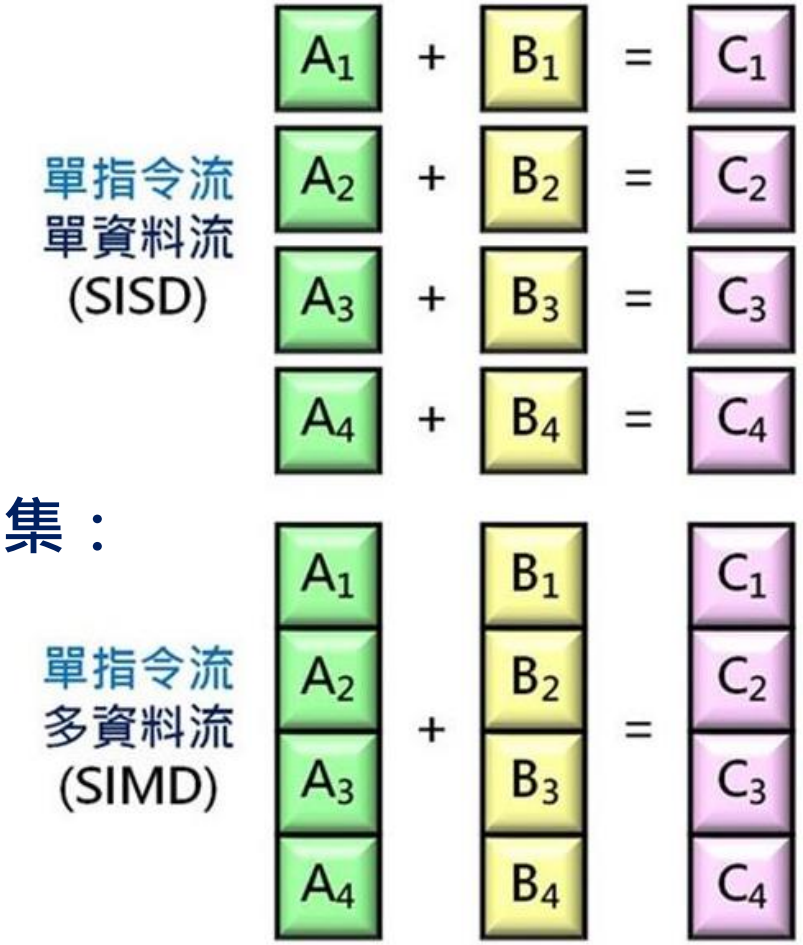
以早期8048 / 8051為例，時脈6MHz，一個指令6個週期，所以只有 1MIPS。

STM32F103x (Cortex-M3) 大多數指令為1個週期，其速為 1.25 MIPS/MHz，故在 72MHz 下可達 90MIPS。

以新唐 Cortex-M4 為例

- M460系列 200MHz
- M480系列 192MHz
- M471系列 72/120MHz
- M451系列 72MHz

硬體加速手法 – 平行/向量指令集加速

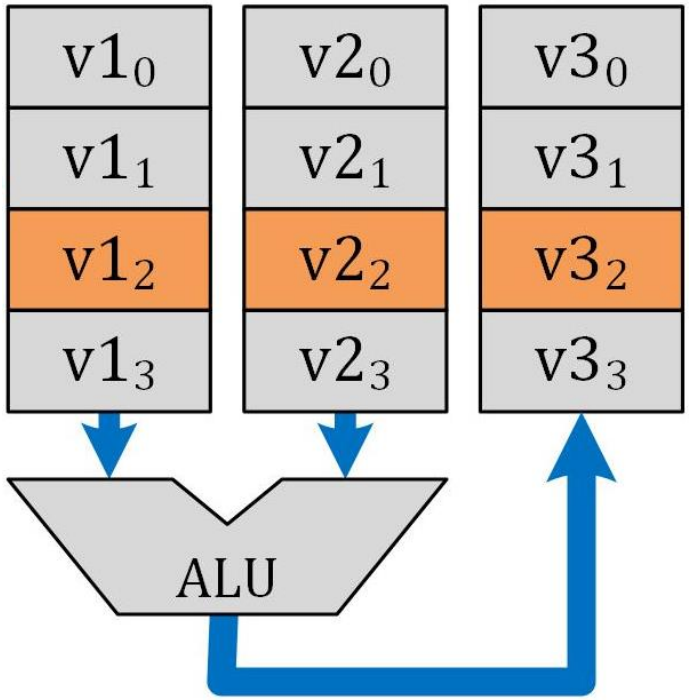


SIMD指令集：

arm DSP,
RISC-V P

向量指令集：

Intel AVX, arm MVE, NEON,
RISC-V V



資料來源：https://omnixri.blogspot.com/2022/10/mcunputinyml.html

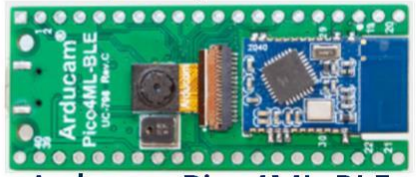
硬體加速手法 – 多核心加速


Raspberry Pi Pico


Raspberry Pi Pico W


Xiao RP2040

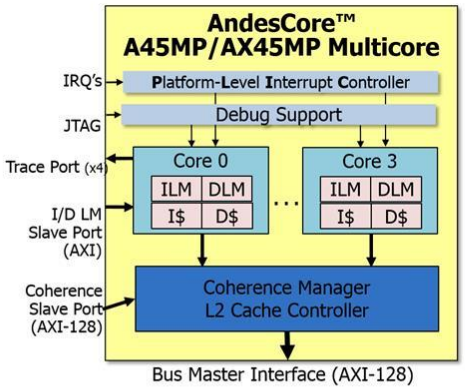

Arduino Nano RP2040


Arducam Pico4ML-BLE

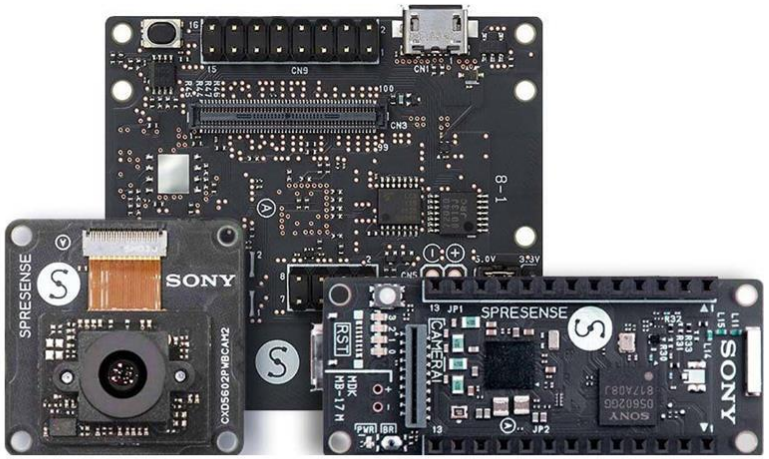

Waveshare RP2040-Zero

Raspberry RP2040 系列

arm cortex-m0+
2 cores
133 MHz
264KB SARM
2MB Flash



Andes AX45MP
64bit RISC-V 4 cores



Sony Spresense

arm cortex-m4
6cores
156 MHz
1.5MB SARM
8MB Flash

多核心組成

- 同質多核
  - 同頻時脈
  - 異頻時脈
- 異質多核
  - 分工作業
  - 功耗調配
- MPU+MCU
  - 兼顧高效
  - 作業系統

OmniXRI整理製作, 2022/10/17

資料來源：https://omnixri.blogspot.com/2022/10/mcunputinyml.html

硬體加速手法 – NPU神經網路加速器

- **arm Ethos-U55**

ALIF Ensemble, Nuvoton M55, Himax WE2, Infineon PSoC Edge, Synaptics Astra SR

- **arm Ethous-U65**

NXP i.MX93

- **arm Ethous-U85**

ALIF E8

- **ST Neural-ART Accelerator**

STM32N6

- **NXP eIQ Neutron NPU**

i.MX95, MCX-N54/94

- **Renasas DRP-AI**

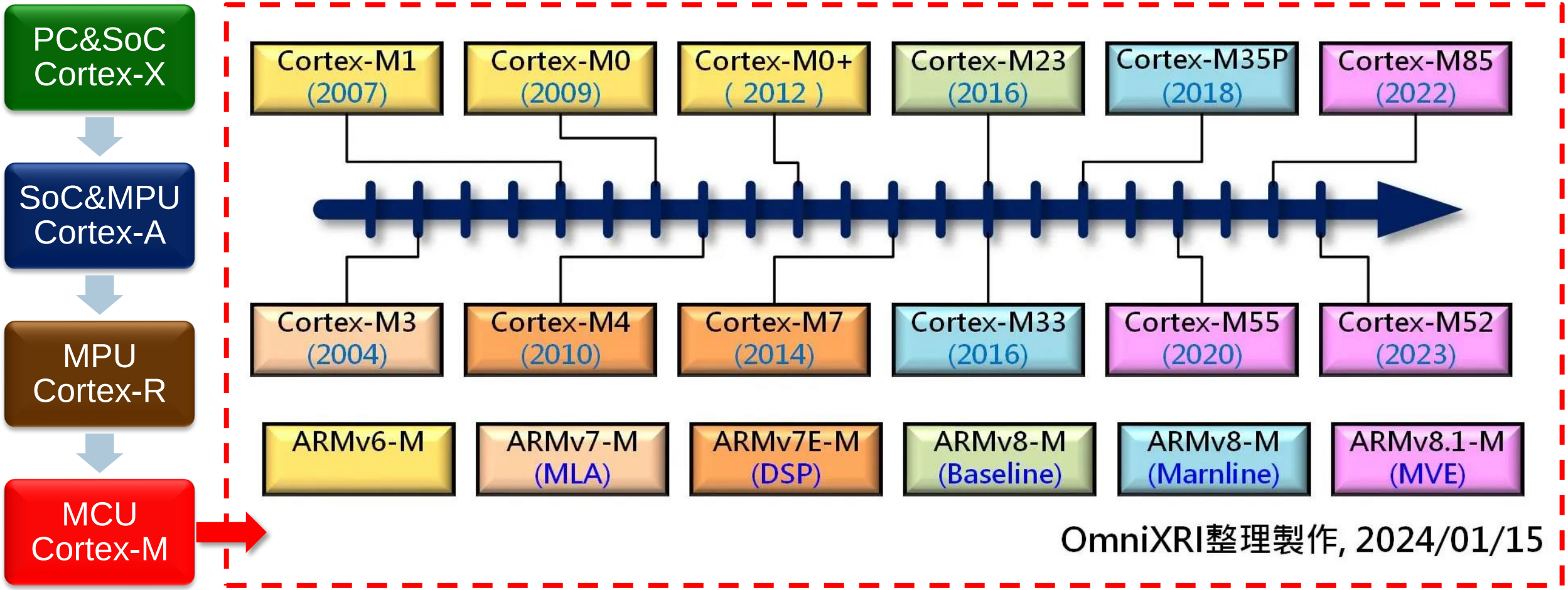
RZ/V2M, RZ/V2L, RZ/V2H, RZ/V2N

- **Others**

Ceva, Tensilica, VeriSilicon, Synopsys, 耐能(Kneron), 英業達(Inventech) …

4.2. DSP (SIMD) 指令集

常見 Arm 晶片CPU等級及指令集



資料來源：https://omnixri.blogspot.com/2024/01/vmaker-edge-ai-13-npuai.html

arm Cortex-M 指令集（v6m, v7m)

指令集主要功能

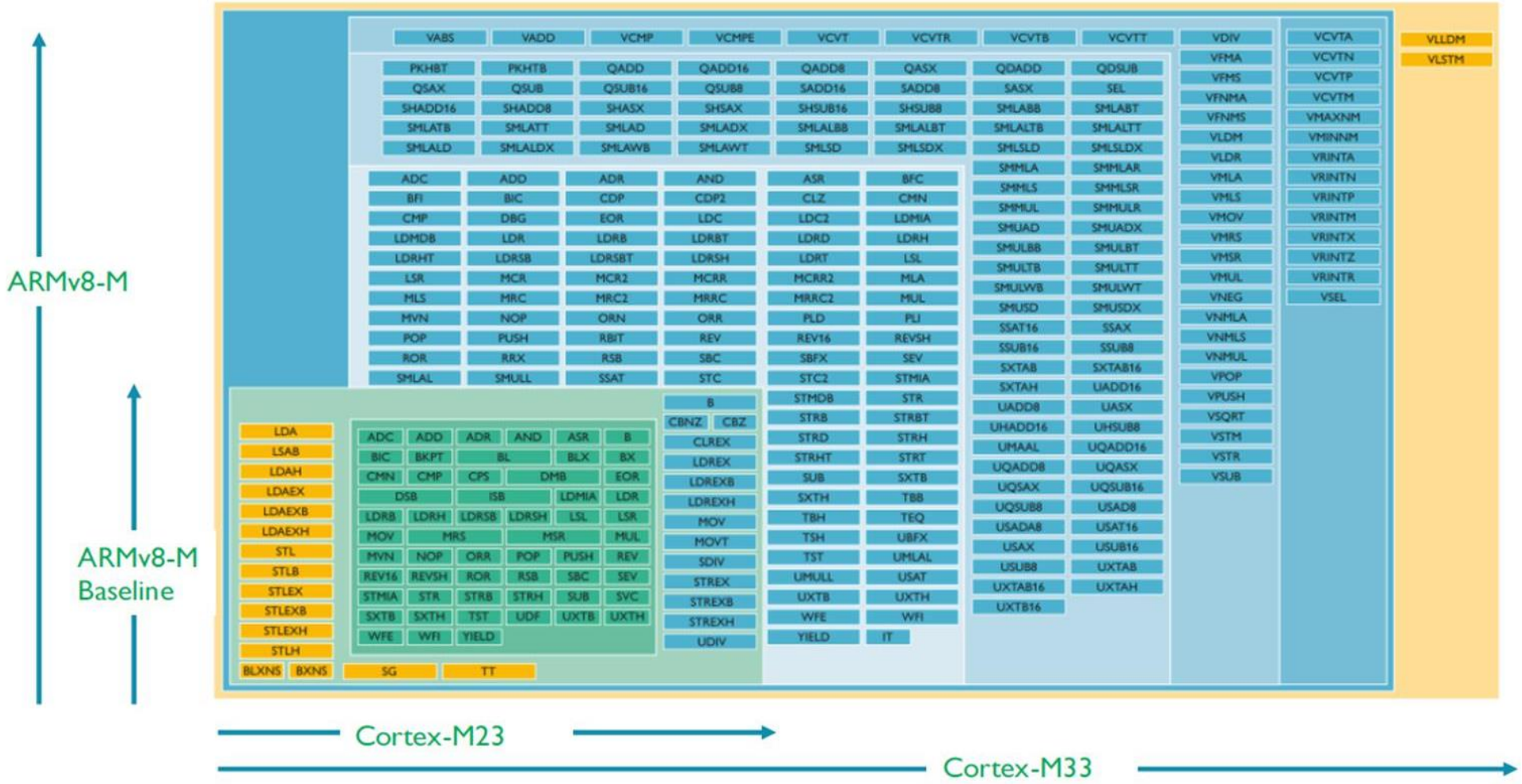
- 算術指令
- 邏輯指令
- 分歧、控制指令
- 記憶體處理指令
- 中斷處理指令
- 浮點運算指令

Thumb-1 v6m 純16bit

Thumb-2 v7m以上 16 / 32 bit混合

資料來源：https://ithelp.ithome.com.tw/m/articles/10267487

arm Cortex-M 指令集（v8m)



資料來源：https://ithelp.ithome.com.tw/m/articles/10267487

基本乘法指令 MUL

32x32 位元乘法

Cortex M0 / M0+ / M1 / M23 取低32位元輸出。

Cortex M3 / M4 / M7 / M33 / M35P 得64位元輸出。

指令週期

Cortex M0 / M0+ / M23 ，1或32個週期。

Cortex M3 ，3 ~ 5個週期。

Cortex M4 / M7 / M33 / M35P，1個週期。

基本乘加指令 MLA

乘積累加運算（Multiply Accumulate, MAC）

基本加法 **ADD** r0, r1, #5 (r0 = r1 + 5)

基本乘法 **MUL** r0, r1, r2 (r0 = r1 x r2)

基本乘加指令 **MLA** r0, r1, r2, r3 (r0 = r1 x r2 + r3)

基本乘減指令 **MLS** r0, r1, r2, r3 (r0 = r1 x r2 - r3)

Cortex M3 ，3 ~ 5個週期。

Cortex M4 / M7 / M33 / M35P，1個週期。

OP，操作，即加減乘除。
OPS，每秒多少操作。
GOPS，每秒10億(10^9)次操作。
TOPS，每秒1兆(10^12)次操作。

常見 SIMD 並行指令

SIMD 指令 (Single Instruction Multiple Data)

SADD16 兩個16位元有號數相加

SADD16 Rd, Rn, Rm (Rd[15:0] = Rn[15:0] + Rm[15:0], Rd[31:16] = Rn[31:16] + Rm[31:16])

SMUAD 兩個16位元有號數乘法後相加

SMUAD Rd, Rn, Rm (Rd = Rn[15:0] * Rm[15:0] + Rn[31:16] * Rm[31:16])

SMLAD 兩個16位元有號數乘法後累加

SMLAD Rd, Rn, Rm, Ra (Rd = Rn[15:0] * Rm[15:0] + Rn[31:16] * Rm[31:16] + Ra)

SMAX16 2個16位元有號數取最大值

SMAX16 Rd, Rn, Rm (Rd[15:0] = max(Rn[15:0], Rm[15:0]), Rd[31:16] = max(Rn[31:16], Rm[31:16]))

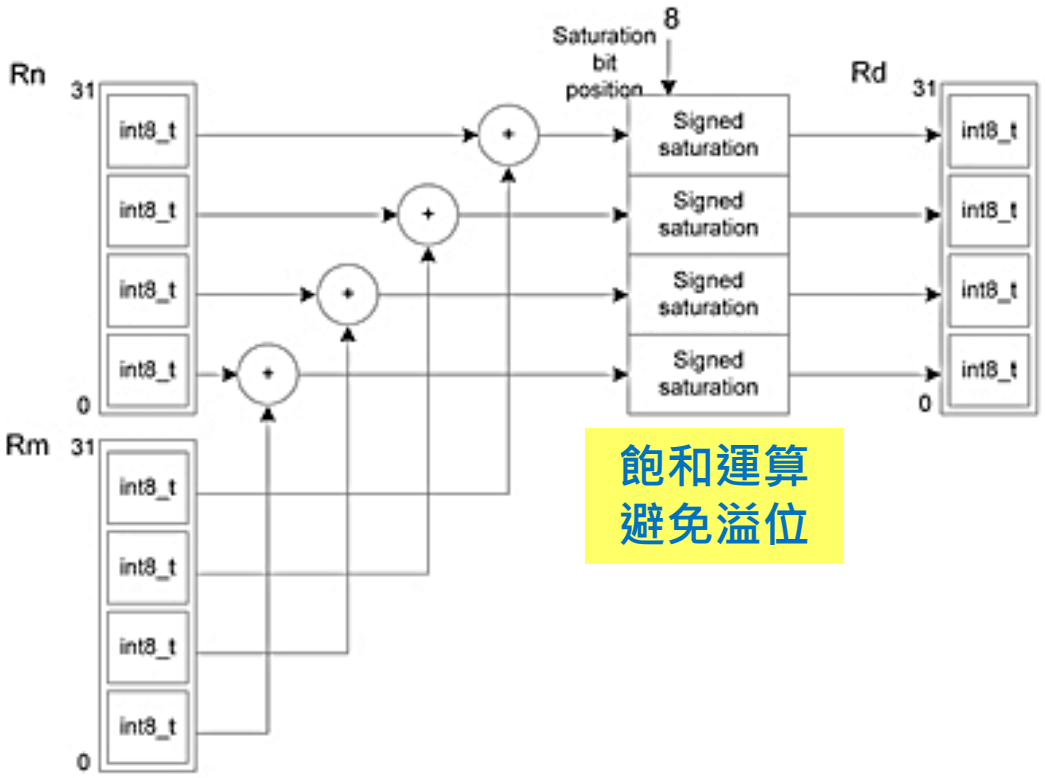
資料來源：https://www.elec4.co.kr/article/articleView.asp?idx=17179

SIMD 飽和加法指令 QADD

飽和指令 (Saturation Instructions)

QADD8 {<Rd>,} <Rn>, <Rm>

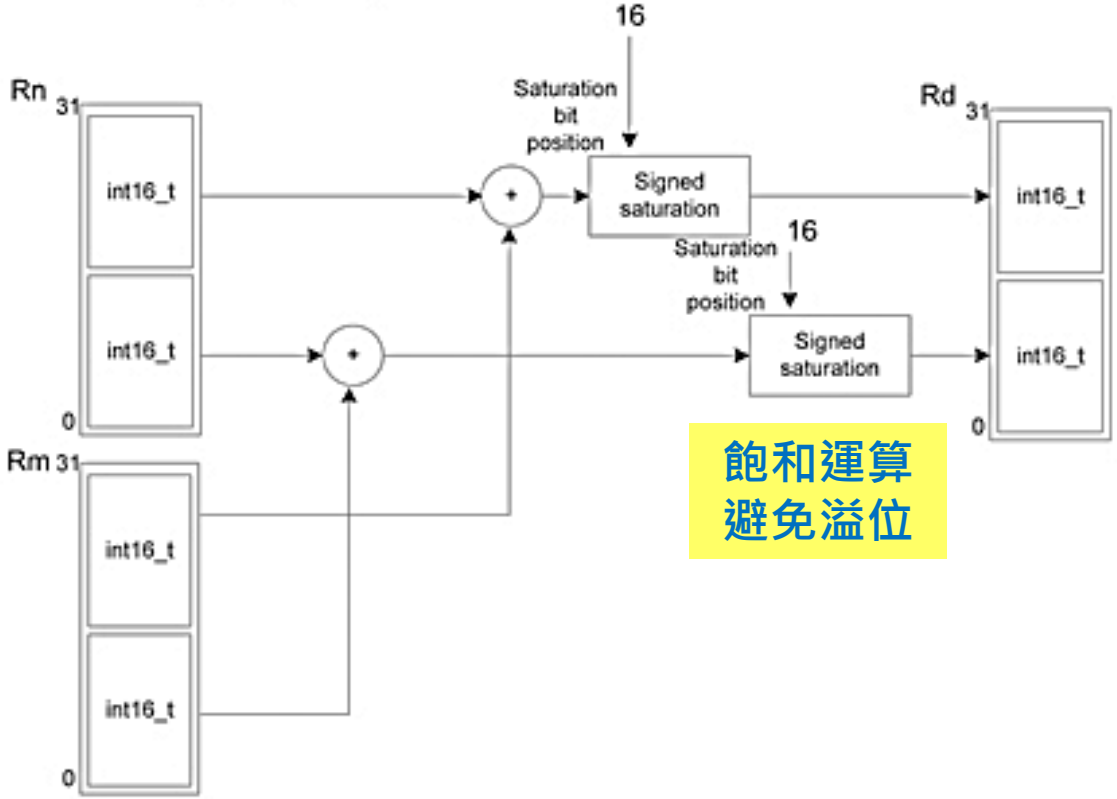
4個8位元相加



飽和運算
避免溢位

QADD16 {<Rd>,} <Rn>, <Rm>

2個16位元相加



飽和運算
避免溢位

資料來源：https://www.elec4.co.kr/article/articleView.asp?idx=17179

浮點運算指令

Cortex-M4 不帶硬體浮點計算， M4F 有帶硬體浮點計算

指令以**V**開頭表示

基本算術運算： VADD.F32, VSUB.F32, VMUL.F32, VDIV.F32 …

乘加乘減運算： VMLA.F32, VMLS.F32, VFMA.F32, VFMS.F32 …

數學函式運算： VSQRT.F32, VABS.F32 …

比較指令： VCMP.F32, VCMPE.F32 …

數值存取： VLDR.F32, VSTR.F32, VMOV.F32 …

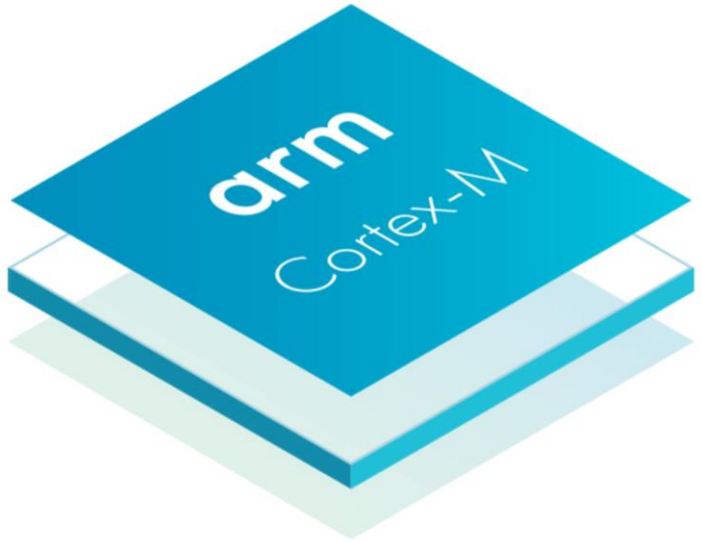
數值轉換： VCVT.F32.S32, VCVT.S32.F32, VCVT.F32.U32 …

加法、乘法通常為1個週期，除法和平方通常要14個週期

4.3. Helium (MVE) 指令集

Cortex-M 指令加速操作比較

處理器	指令集	OPs at 100MHz
Cortex-M3	多功能指令（MLA，2個OPS）需要2個週期。 由於處理器還需要執行NN處理的記憶體負載操作，假設Mac vs Load的比率為1：1，因此平均操作/週期為0.6。	0.06 GOPs/sec
Cortex-M4, Cortex-M33	DSP/SIMD指令支持雙MAC操作（4個操作）。 由於處理器還需要執行NN處理的記憶體加載操作，假設Mac vs Load的比率為1：1，因此平均操作/週期為2。	0.2 GOPs/sec
Cortex-M7	該處理器支持DSP/SIMD和記憶體負載的雙重問題， 因此平均操作/週期為4。	0.4 GOPs/sec
Cortex-M52	使用Helium技術，這些處理器可以與數據負載並聯處理4個MAC/週期。結果，平均操作/週期為8。	0.8 GOPs/sec
Cortex-M55, Cortex-M85	使用Helium技術，這些處理器可以與數據負載並聯處理8個MAC/循環。結果，平均操作/週期為16。	1.6 GOPs/sec

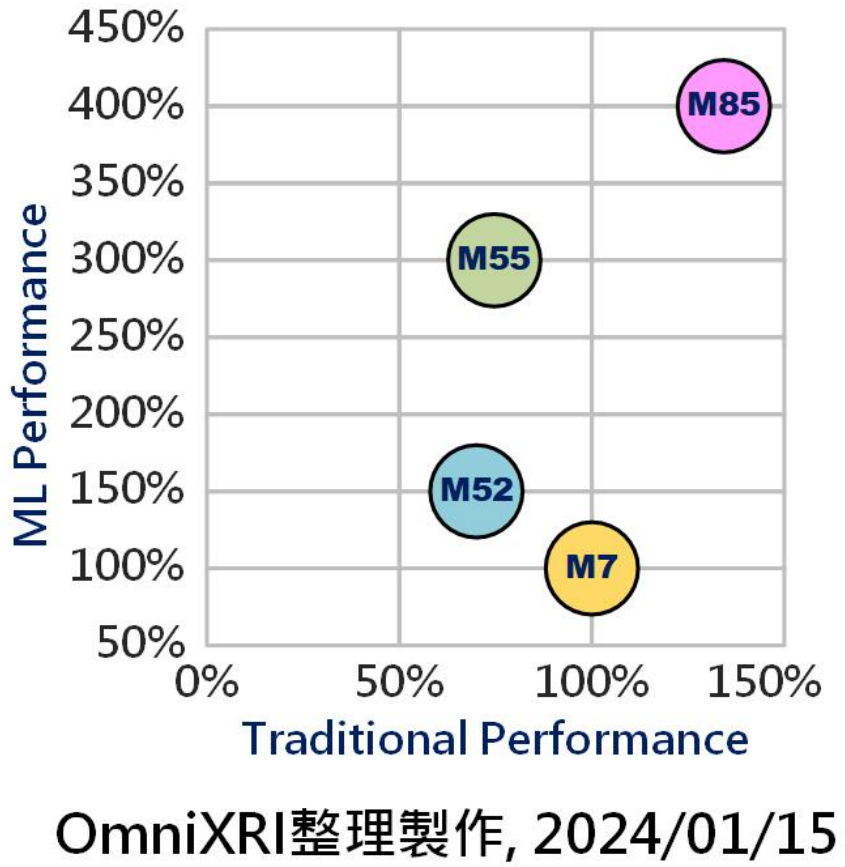


arm v8.1m (Helium) 指令集
M-Profile Vector Extension, MVE

資料來源：https://hackmd.io/@OmniXRI-Jack/arm_developer_cortexm55_ethosu55_guide

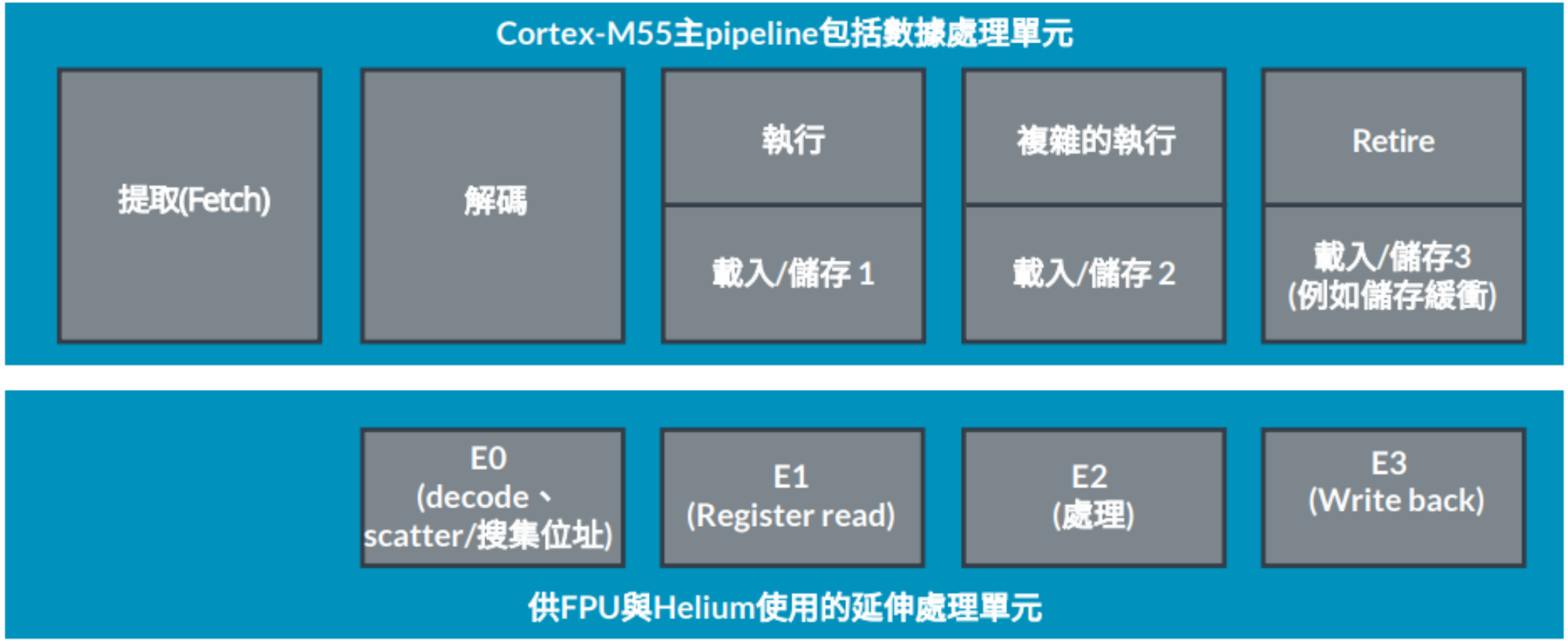
Cortex-M 指令集與加速計算

IP	指令集	Helium	DMIPS/MHz	CoreMark/MHz
Cortex-M7	v7E-M	X	2.31	5.29
Cortex-M55	v8.1-M	Dual-beat	1.69	4.40
Cortex-M85	v8.1-M	Dual-beat	3.13	6.28
Cortex-M52	v8.1-M	Single-beat	1.60	4.30



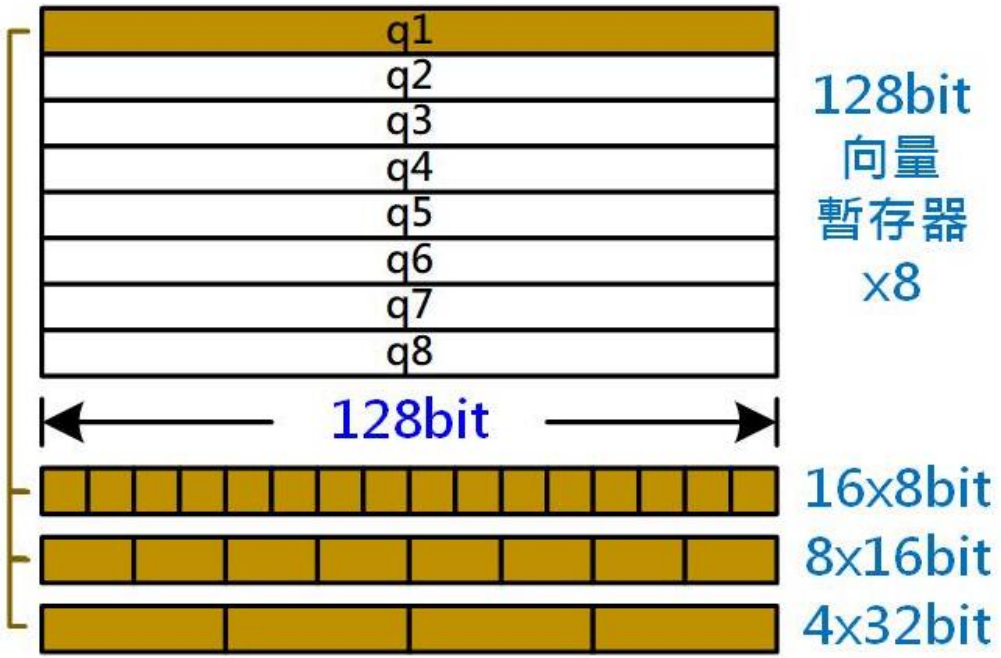
資料來源：https://omnixri.blogspot.com/2024/01/vmaker-edge-ai-13-npuai.html

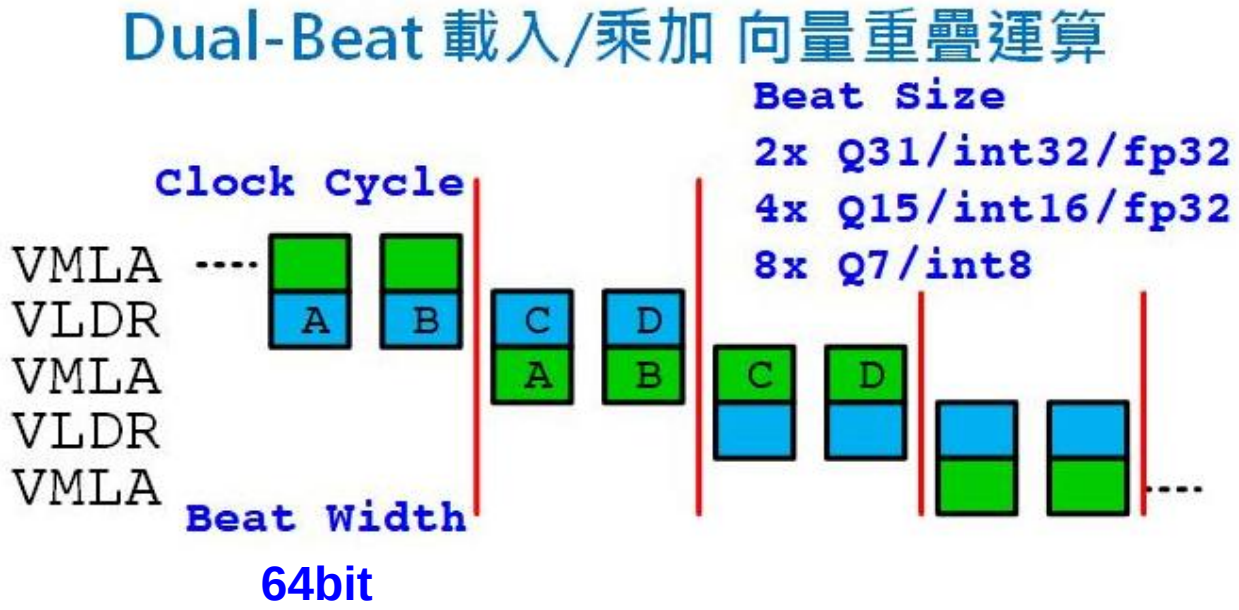
Cortex-M55 工作流水線



資料來源：https://armkeil.blob.core.windows.net/developer/Files/pdf/white-paper/arm-cortex-m55-processor-wp-tw.pdf

Cortex-M55 雙節拍工作模式





資料來源：https://omnixri.blogspot.com/2024/01/vmaker-edge-ai-13-npuai.html

DSP / Helium 指令集比較

特性	Helium (ARMv8.1-M)	Cortex-M4 DSP	Cortex-M4F 浮點
向量寬度	128 位	32 位	無（純量）
數據類型	8/16/32 位整數, FP16/FP32	8/16/32 位整數	FP32
指令數量	~150 種	~30 種	~20 種
主要應用	ML, DSP, 向量處理	傳統 DSP	浮點控制算法
硬件需求	ARMv8.1-M + 可選 FPU	ARMv7-M + MAC	ARMv7-M + FPU

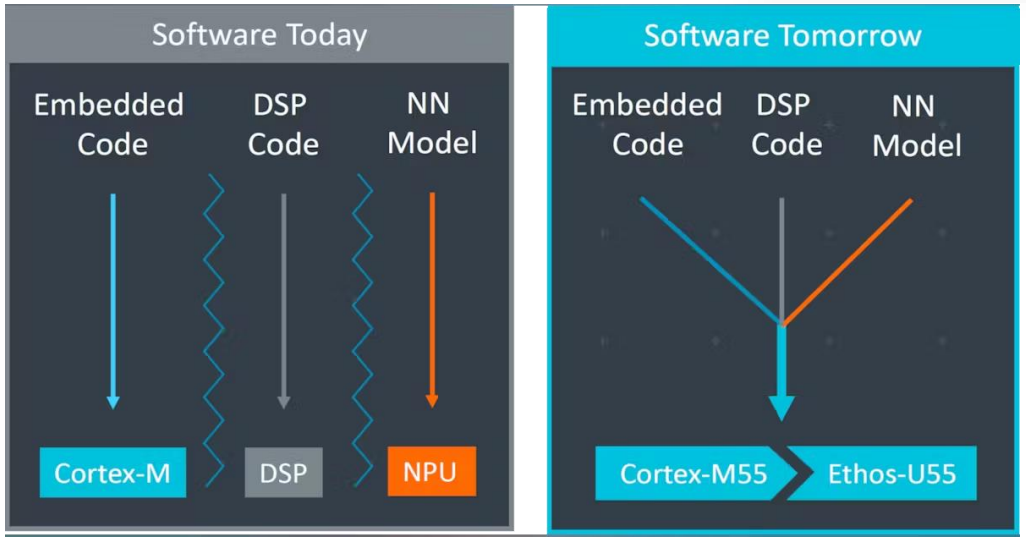
4.4. Ethos-U MicroNPU

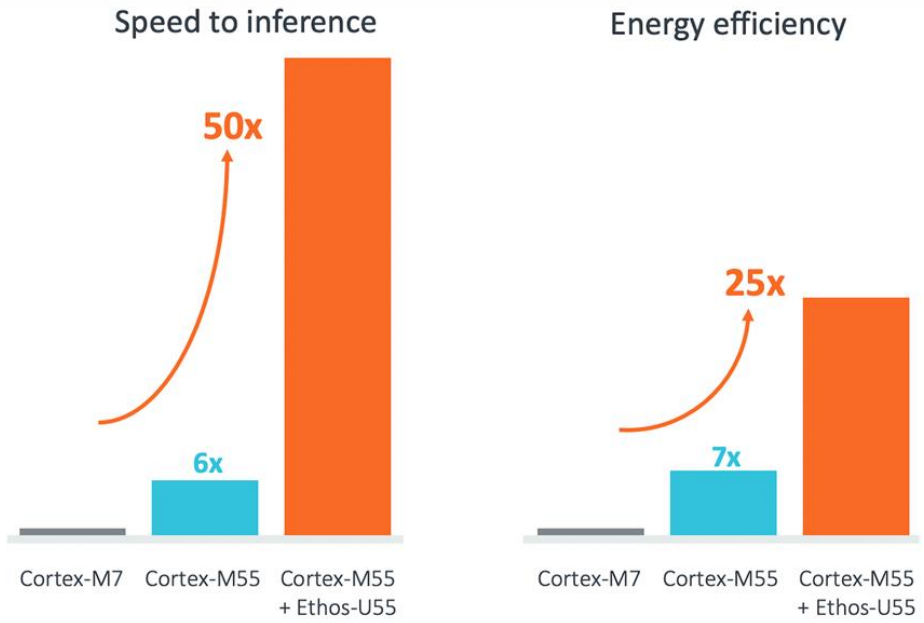
arm Micro NPU

arm Micro NPU

- Ethos-U55
- Ethos-U65
- Ethos-U85

Key Features		Ethos-U55	Ethos-U65	Ethos-U85
	Performance (At 1 GHz)	64 to 512 GOP/s	512 GOP/s to 1 TOP/s	256 GOPS/s to 4 TOP/s
	MACs (8x8)	32, 64, 128, 256	256, 512	128, 256, 512, 1024, 2048

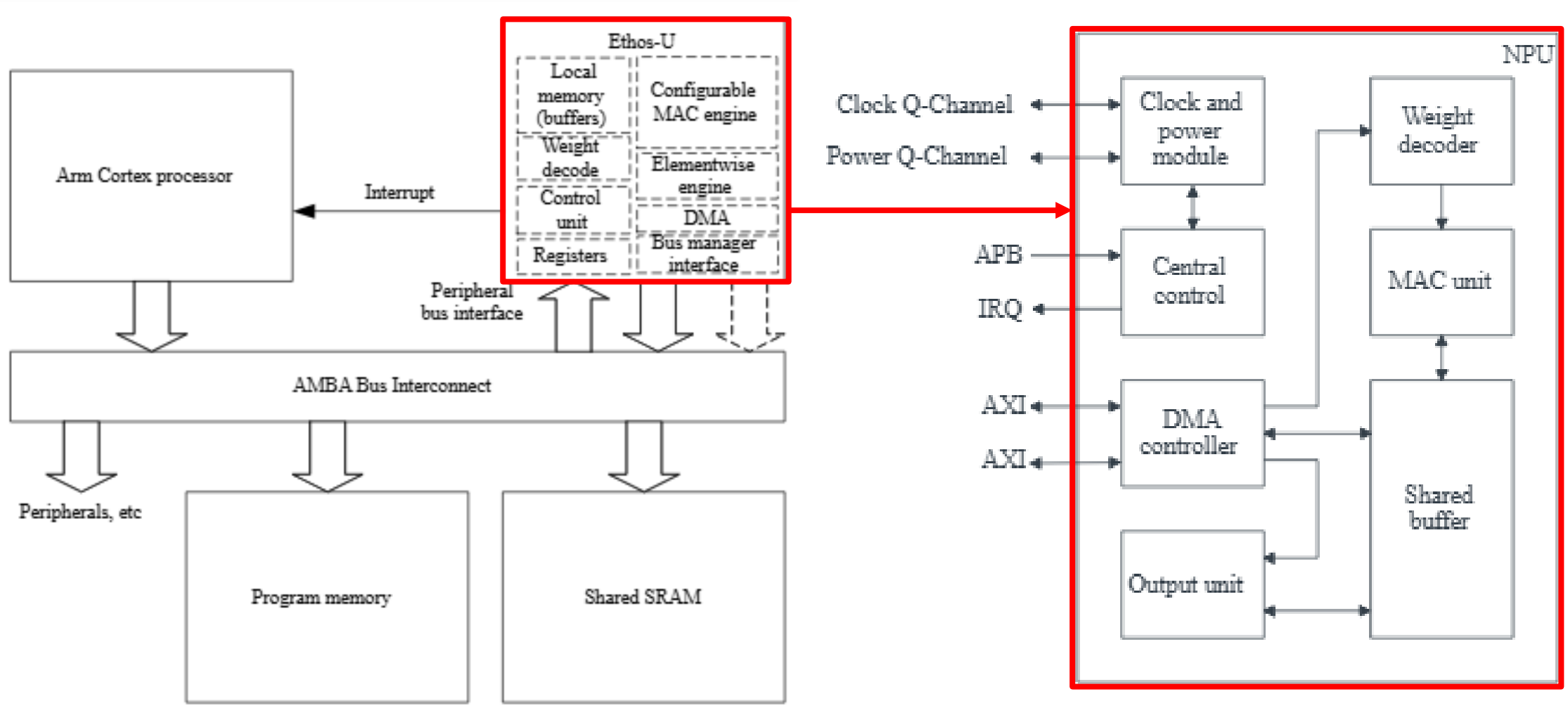




Cortex-M55 + Ethos-U55

資料來源：https://armkeil.blob.core.windows.net/developer/Files/pdf/white-paper/arm-cortex-m55-processor-wp-tw.pdf

Ethos-U55 系統架構圖



資料來源：https://hackmd.io/@OmniXRI-Jack/arm_developer_cortexm55_ethosu55_guide

Ethos-U55 可支援算子及操作

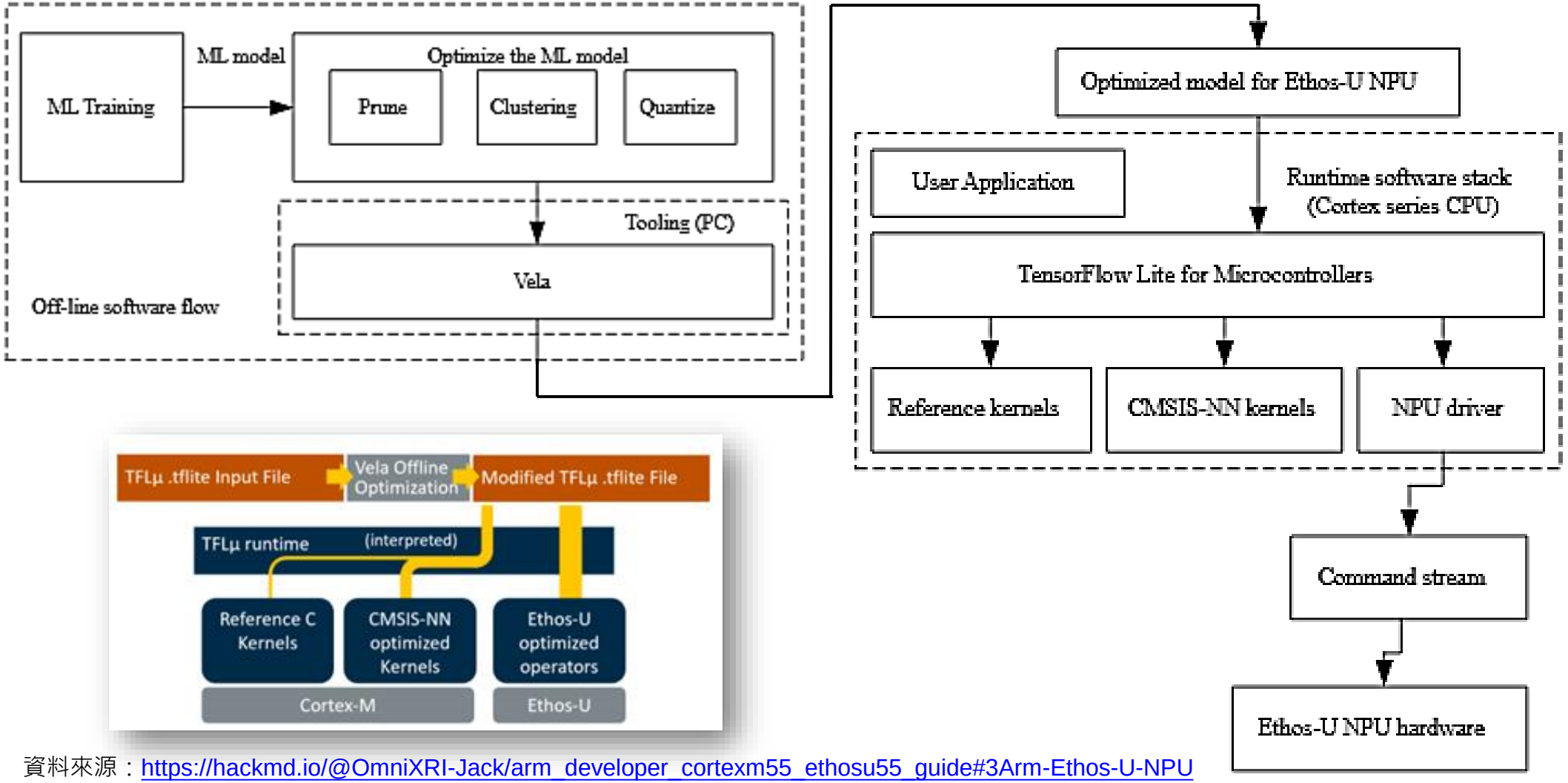
算子(Operators)

- Concat
- EspandDims
- GRU
- Identity
- Logistic
- LSTM
- Pack
- Reshape
- Split
- Squeeze
- Stack
- Unpack
- Unstack
- Resize_Bilinear
- BatchRenorm
- StridedSlice
- 1-Strides Only

操作(Operations)

- Convolution
- Depth-wise Convolution
- Pooling
- Vector-Product
- Elementwise
- Reduction

Ethos-U55 開發流程



資料來源：https://hackmd.io/@OmniXRI-Jack/arm_developer_cortexm55_ethosu55_guide#3Arm-Ethos-U-NPU

參考文獻

- 許哲豪，臺灣科技大學資訊工程系「人工智慧與邊緣運算實務」（2021~2023）

https://omnixri.blogspot.com/p/ntust-edge-ai.html

- 許哲豪，OmniXRI's Edge AI & TinyML 小學堂 Youtube 直播課程總結

https://omnixri.blogspot.com/2024/06/omnixris-edge-ai-tinyml-youtube.html

- 許哲豪，歐尼克斯實境互動工作室系列發文—TinyML(MCU AI)系列

https://hackmd.io/1PK1URhIQ7GutcWgpgsWbg#TinyMLMCU-AI%E7%B3%BB%E5%88%97

- Wiki – ARM Cortex-M

https://zh.wikipedia.org/wiki/ARM_Cortex-M

參考文獻

- 許哲豪，MCU攜手NPU讓tinyML邁向新里程碑

https://omnixri.blogspot.com/2022/10/mcunputinyml.html

- 許哲豪，誰說單晶片沒有神經網路加速器NPU就不能玩微型AI應用？

https://omnixri.blogspot.com/2024/01/vmaker-edge-ai-13-npuai.html

- Arm Cortex-M & Ethos-U55 ML開發者指南

https://hackmd.io/@OmniXRI-Jack/arm_developer_cortexm55_ethosu55_guide

- Arm Cortex-M55 處理器介紹

https://armkeil.blob.core.windows.net/developer/Files/pdf/white-paper/arm-cortex-m55-processor-wp-tw.pdf

歐尼克斯實境互動工作室
(OmniXRI Studio)
許哲豪 (Jack Hsu)

Facebook : Jack Omnixri
FB社團： Edge AI Taiwan邊緣智能交流區
電子信箱：omnixri@gmail.com
部落格：https://omnixri.blogspot.tw
開　源：https://github.com/OmniXRI

YOUTUBE 直播： https://www.youtube.com/@omnixri1784/streams