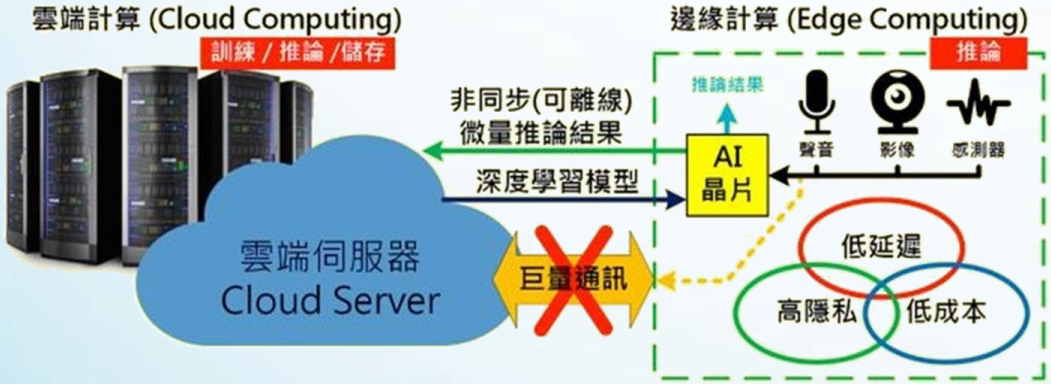
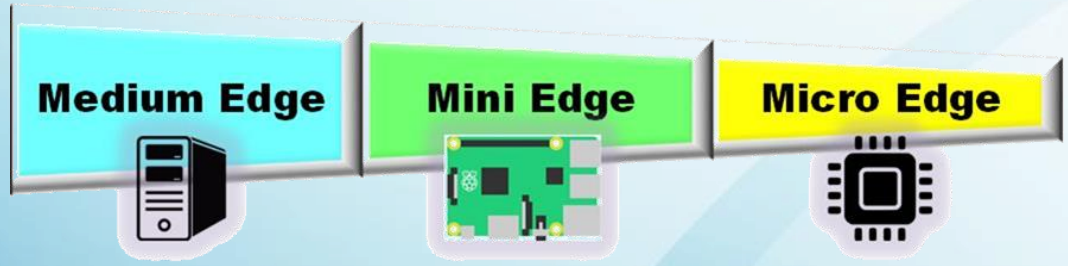
OmniXRI's Edge AI & TinyML 小學堂





【第2講】
邊緣運算硬體

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

簡報大綱

- 2.1.基本運算原理
- 2.2.加速運算晶片
- 2.3.開發板類型
- 2.4.硬體選用評估

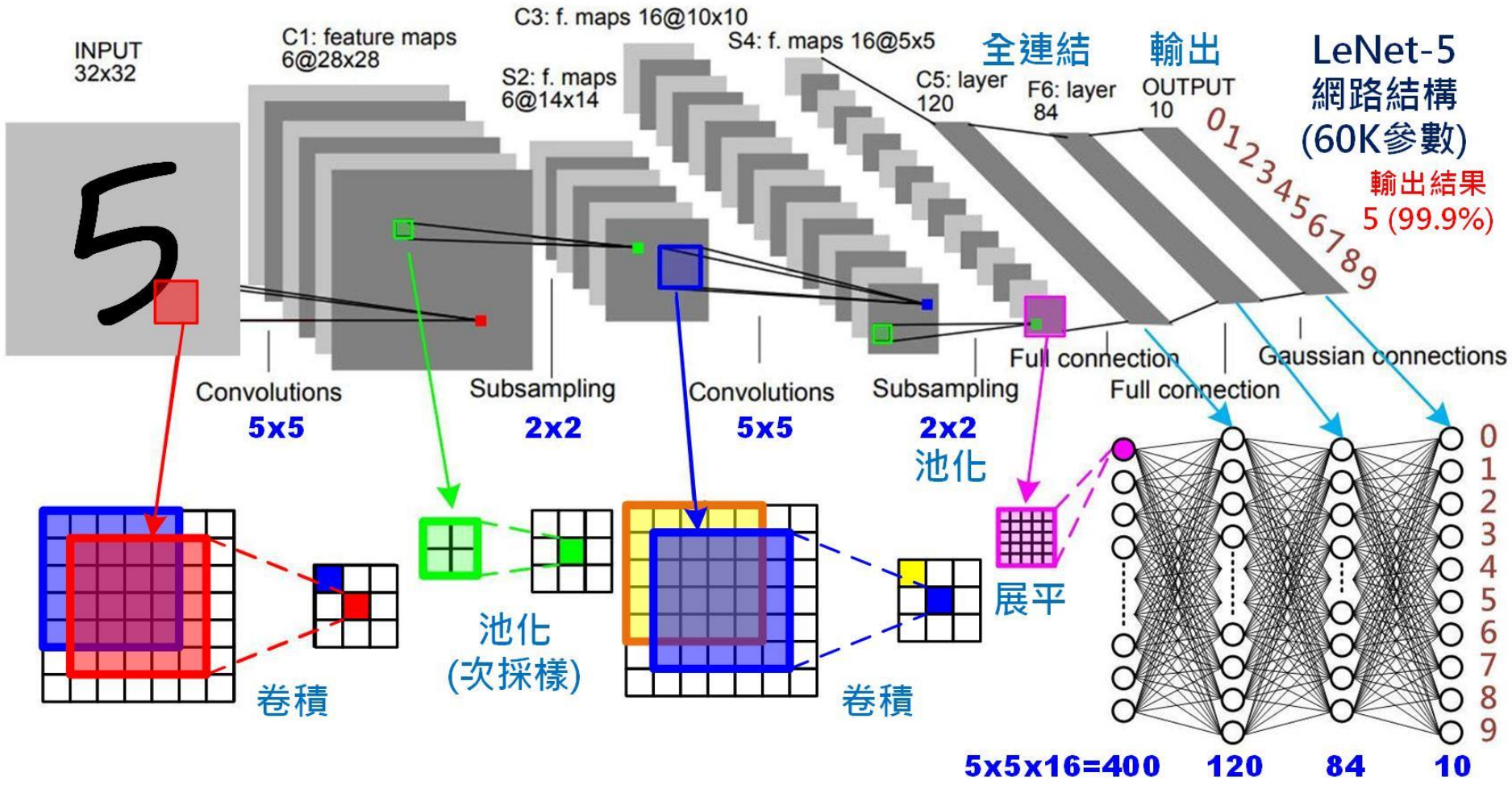
本課程完全免費，請勿移作商業用途！
歡迎留言、訂閱、點讚、轉發，讓更多需要的朋友也能一起學習。

完整課程大綱： https://omnixri.blogspot.com/2024/02/omnixris-edge-ai-tinyml-0.html
課程直播清單： https://www.youtube.com/@omnixri1784/streams

2.1.
基本運算原理

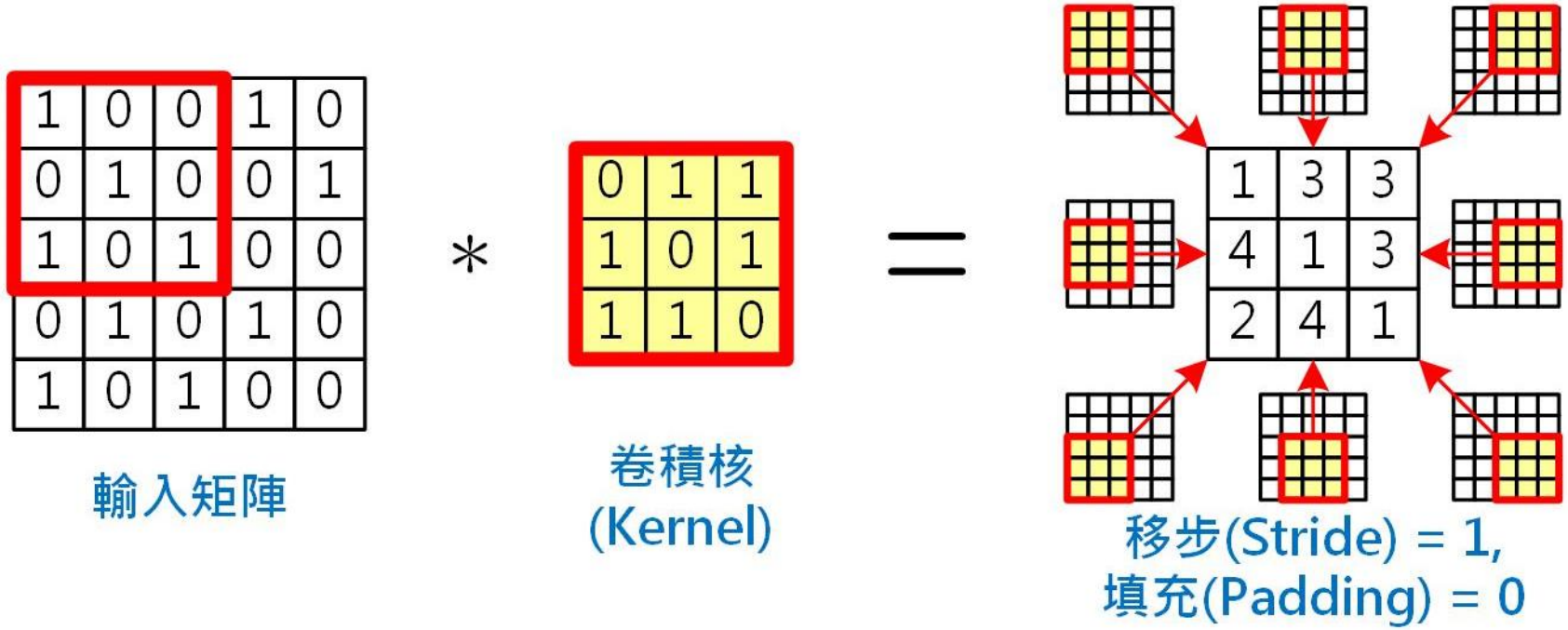
- 卷積神經網路
- 數字表示系統
- 數字表示系統
- 平行 / 並行運算

卷積神經網路



資料來源：http://omnixri.blogspot.com/2018/06/blog-post.html

卷積神經網路—卷積



1	2	3
4	5	6
7	8	9

*

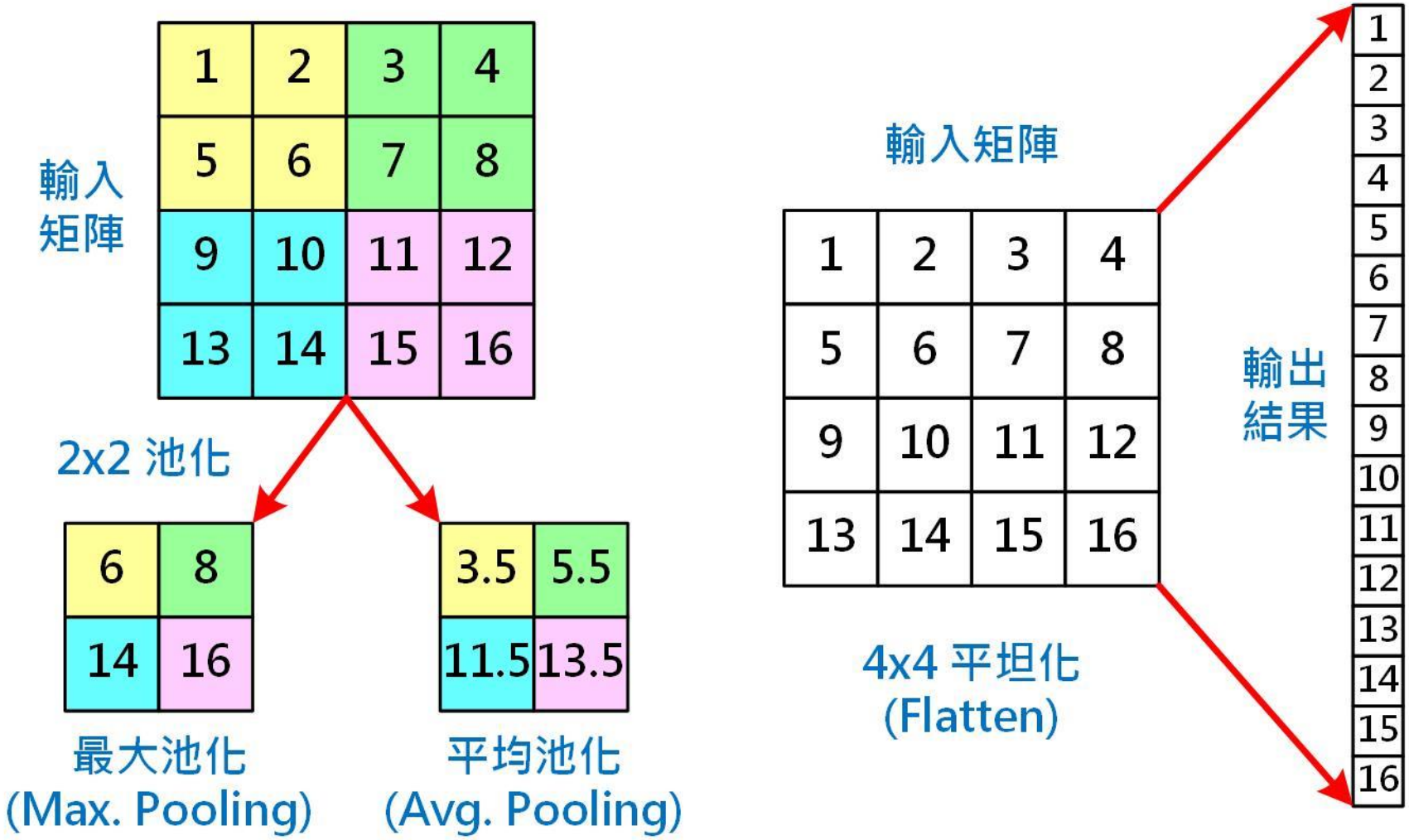
A	B	C
D	E	F
G	H	I

=

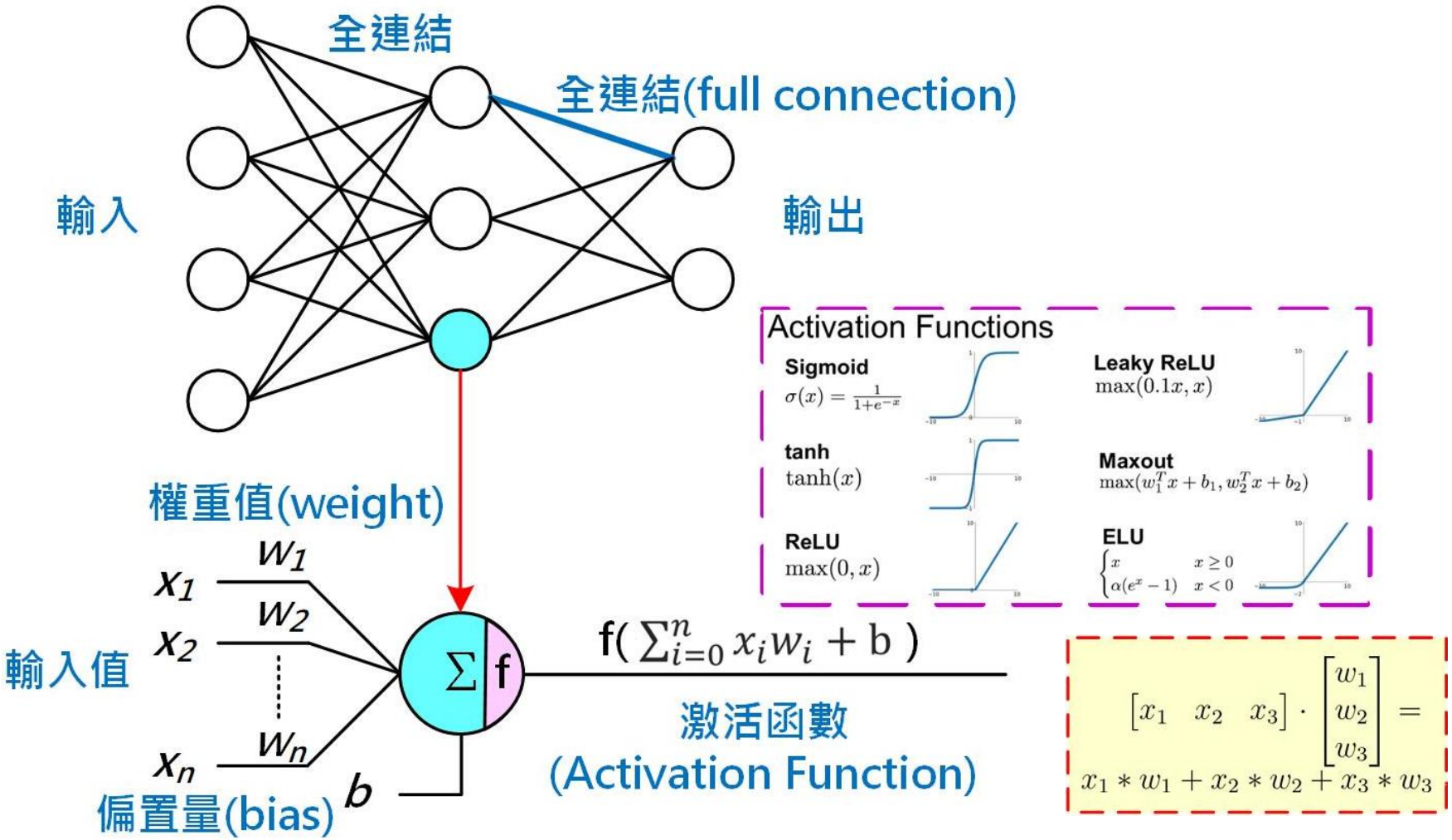
1*A + 2*B + 3*C +
4*D + 5*E + 6*F +
7*G + 8*H + 9*I

*** 相當於Python numpy 的矩陣各元素相乘 a*b 而非矩陣點乘 a.dot(b)**

卷積神經網路—池化、平坦化



卷積神經網路—全連結

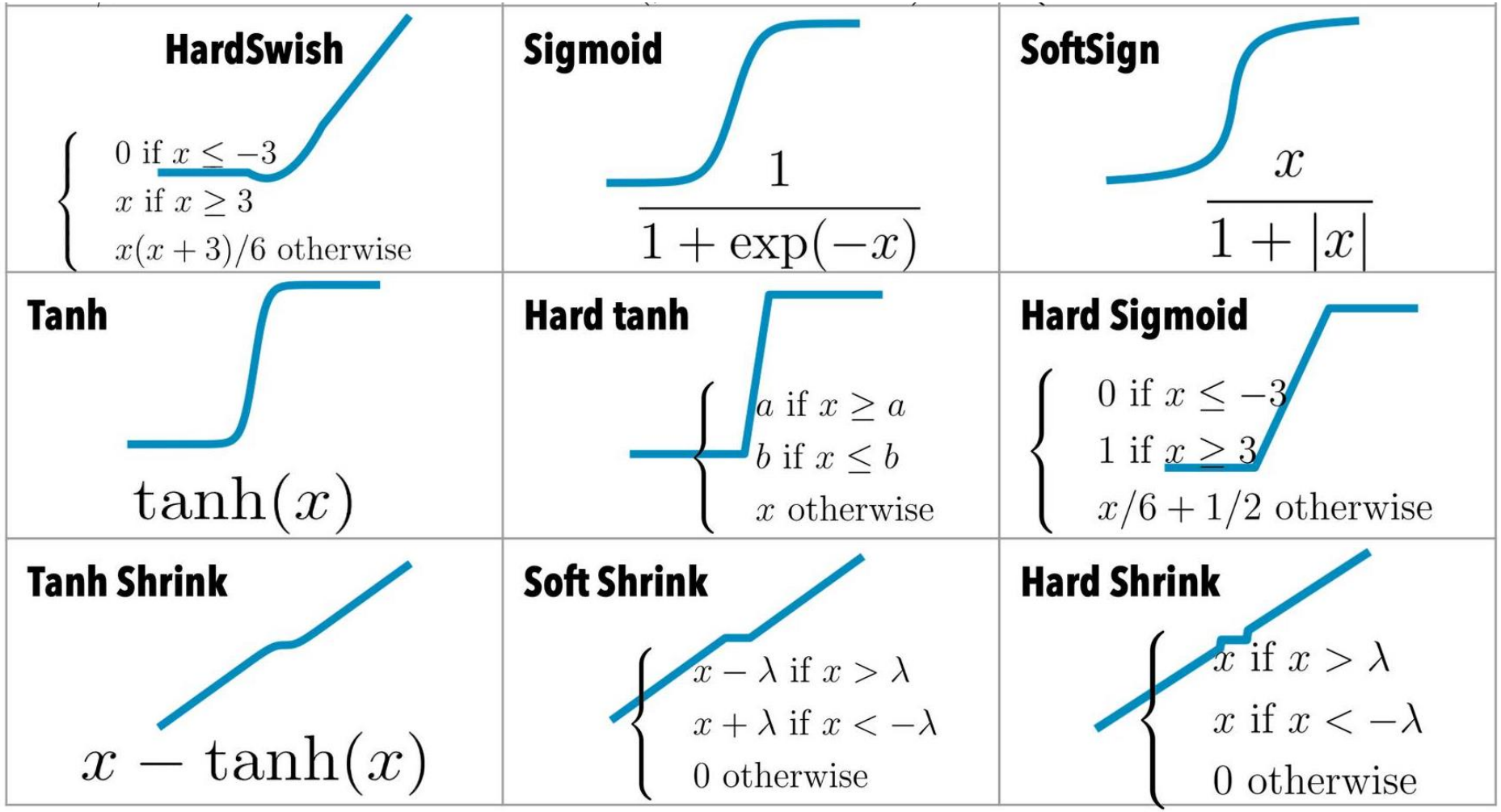


*** 相當於Python numpy 的矩陣點乘 a.dot(b) 而非矩陣各元素相乘的 a*b**

卷積神經網路—激活函數(1/2)

ReLU $\max(0, x)$	**GELU** $\frac{x}{2}\left(1+\tanh\left(\sqrt{\frac{2}{\pi}}\right)(x+ax^3)\right)$	**PReLU** $\begin{cases} y = x & if\ x > 0 \\ y = ax & if\ x \leq 0 \end{cases}$
ELU $\begin{cases} x \text{ if } x > 0 \\ \alpha(x \exp x - 1) \text{ if } x < 0 \end{cases}$	**Swish** $\frac{x}{1+\exp -x}$	**SELU** $\alpha(\max(0, x)+ \min(0, \beta(\exp x - 1)))$
SoftPlus $\frac{1}{\beta}\log(1+\exp(\beta x))$	**Mish** $x\tanh\left(\frac{1}{\beta}\log(1+\exp(\beta x))\right)$	**RReLU** $\begin{cases} x \text{ if } x \geq 0 \\ ax \text{ if } x < 0 \text{ with } a \sim \Re(l, u) \end{cases}$

卷積神經網路—激活函數(2/2)



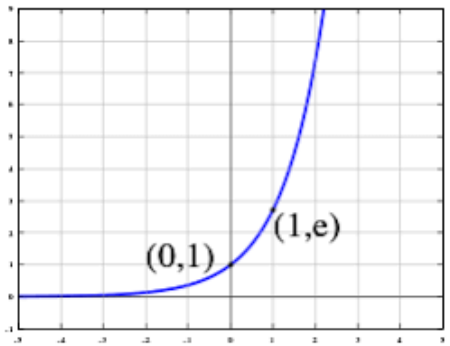
卷積神經網路—輸出函數

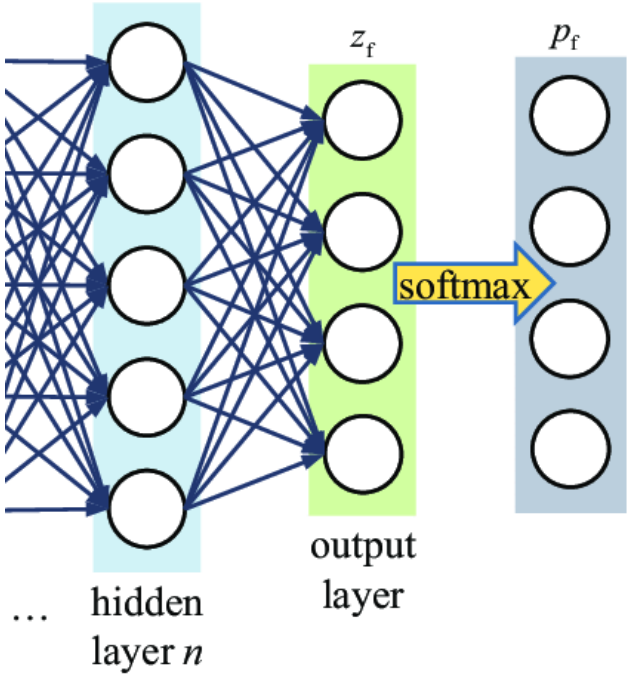
softmax()

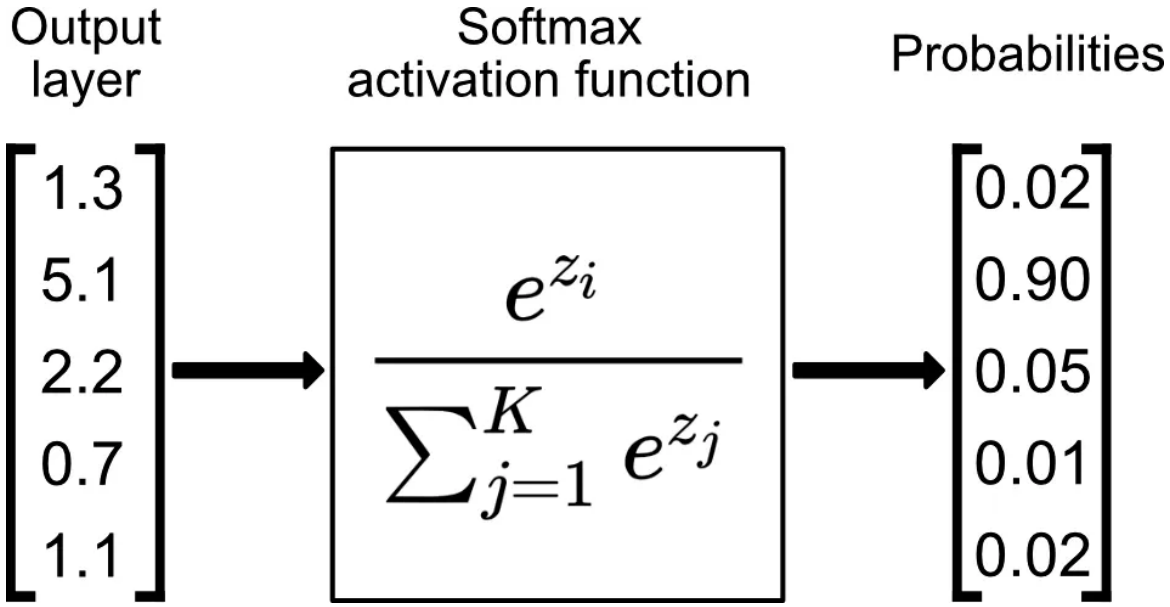
歸一化指數函數
將所有發生機率
壓縮到0~1之間

exp()

以自然常數e為底
的指數函數







卷積神經網路—權重與計算量

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

LeNet-5 模型結構

圖片來源：
https://www.researchgate.net/publication/329970273_Image_recognition_by_knowledge_transfer_using_Deep_Convolutional_Neural_Network

數字表示系統 – 整數

整數常用名稱及表示範圍

AI用名	C/C++名稱	Byte	數值範圍
INT8	char	1	-128 ~ +127
	unsigned char	1	0 ~ +255
INT16	Short (Short int)	2	-32,768 ~ +32,767
	unsigned short	2	0 ~ +65,535
INT32	Int (long int)	4	-2,147,483,648 ~ +2,147,483,647
	unsigned int	4	0 ~ +4,294,967,295
INT64	long long	8	-9,223,372,036,854,775,808 ~ +9,223,372,036,854,775,807
	unsigned long long	8	0 ~ +18,446,744,073,709,551,615
INT4		1/2	0 ~ 15 or -8 ~ +7

負（整）數表示法

最高位元

BIN	DEC
00000010	2
00000001	1
00000000	0
10000001	-1
10000010	-2

1的補數

BIN	DEC
00000010	2
00000001	1
00000000	0
11111110	-1
11111101	-2

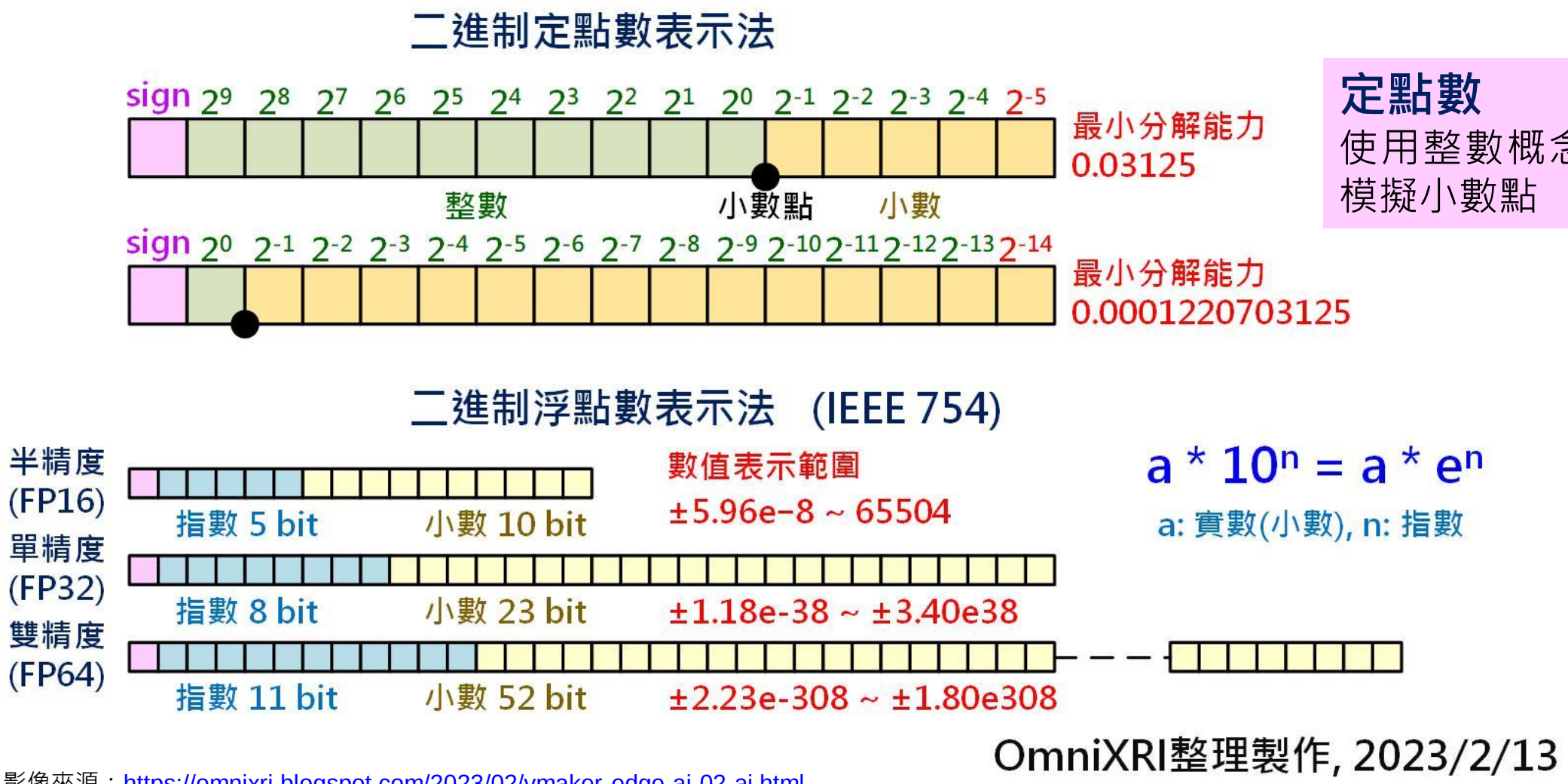
2的補數（1的補數加1）

BIN	DEC
00000010	2
00000001	1
00000000	0
11111111	-1
11111110	-2

OmniXRI整理製作, 2023/02/13

影像來源：https://omnixri.blogspot.com/2023/02/vmaker-edge-ai-02-ai.html

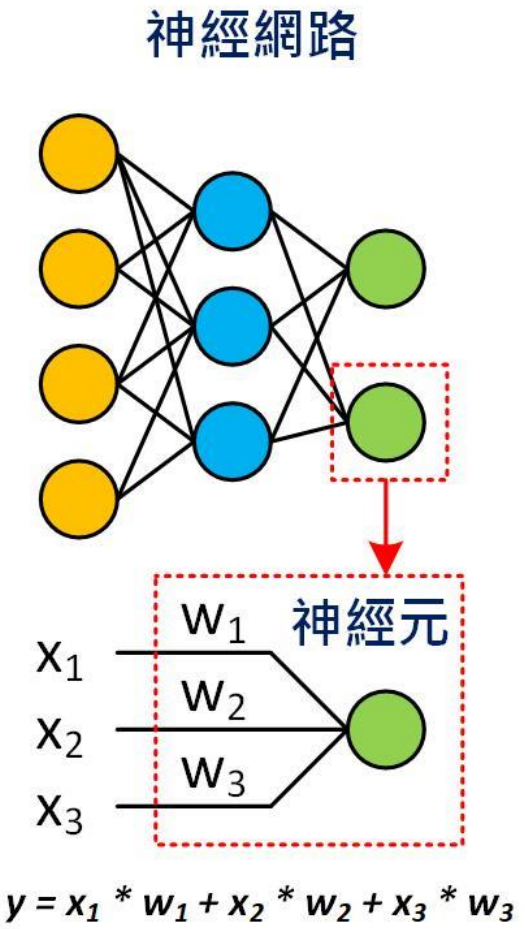
數字表示系統 — 小數

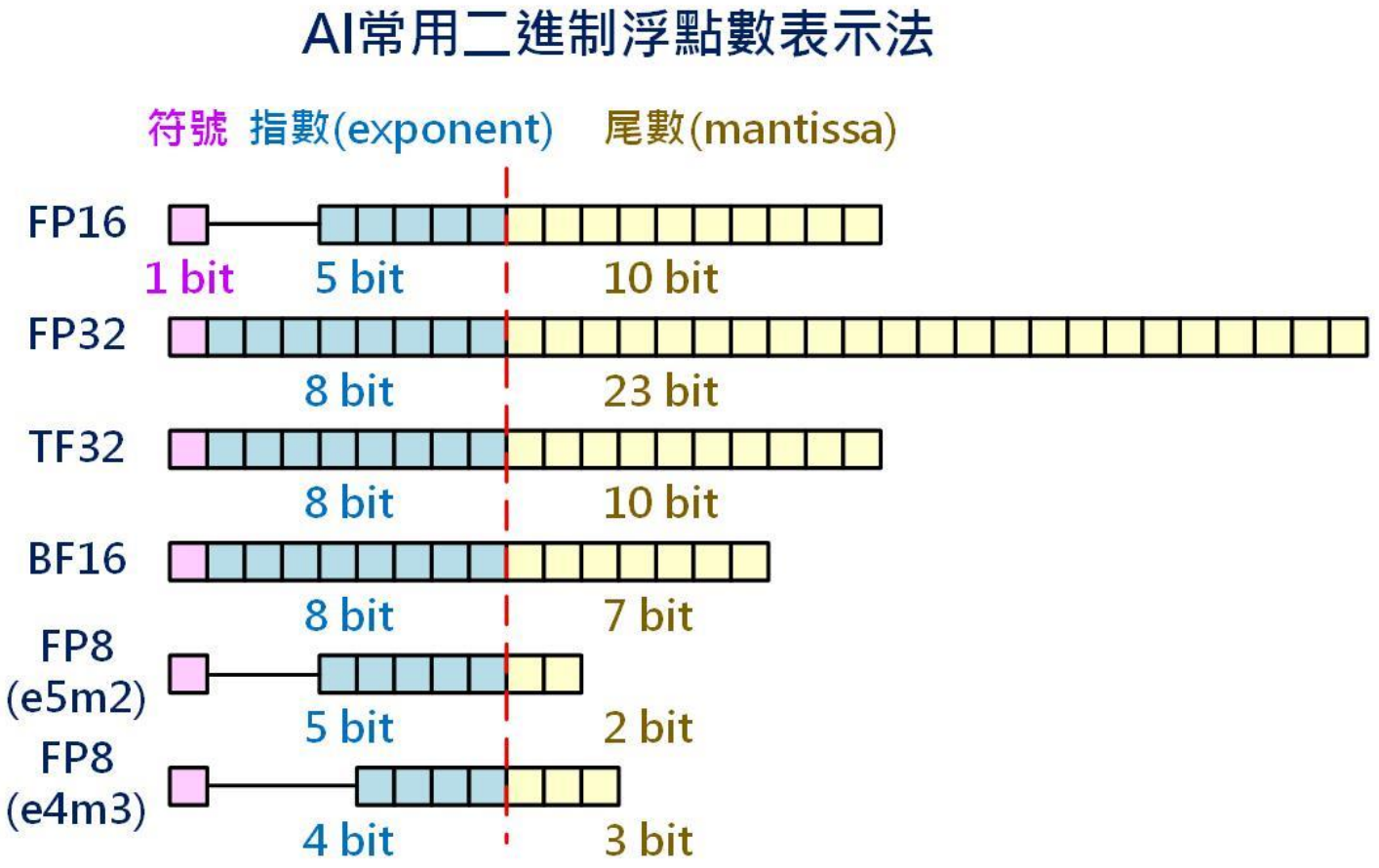


定點數
使用整數概念
模擬小數點

影像來源：https://omnixri.blogspot.com/2023/02/vmaker-edge-ai-02-ai.html

數字表示系統－AI常用浮點



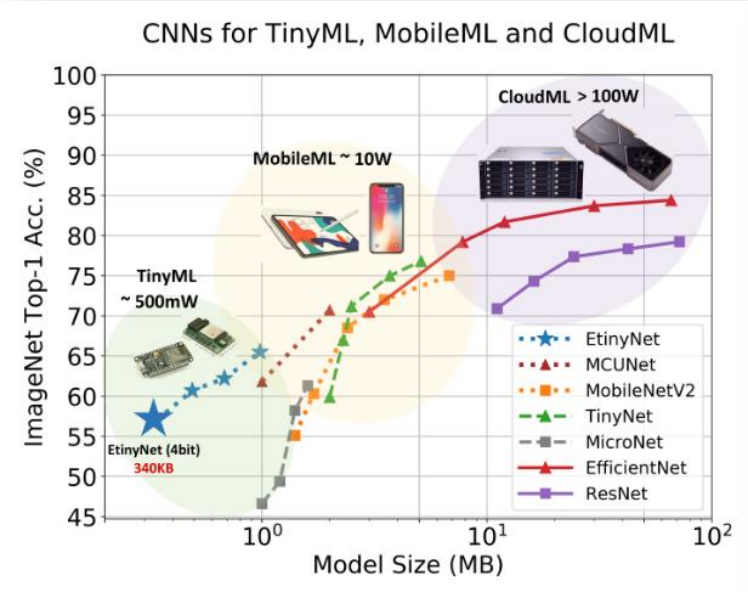


OmniXRI整理製作, 2023/2/13

影像來源：https://omnixri.blogspot.com/2023/02/vmaker-edge-ai-02-ai.html

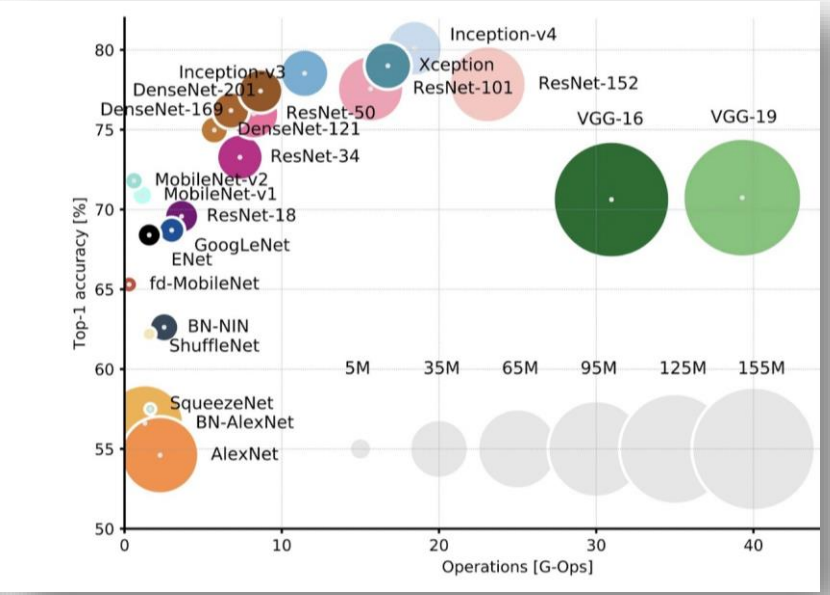
AI模型參數量及位元數演進

數字系統演進： FP32 → FP16 → INT8 → BF16 → FP8 → INT4



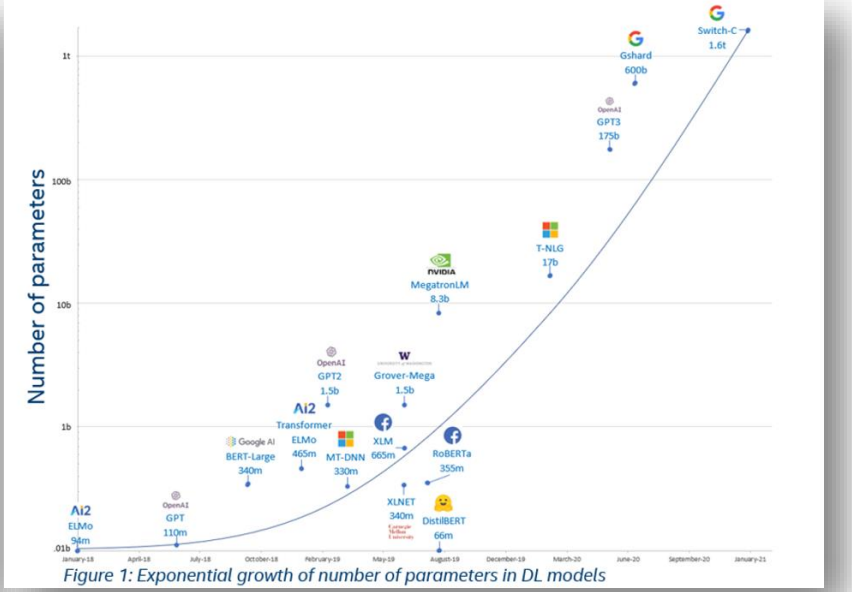
常見單晶片模型
(KB~MB)

https://ojs.aaai.org/index.php/AAAI/article/view/20387



常見影像辨識模型
(MB~GB)

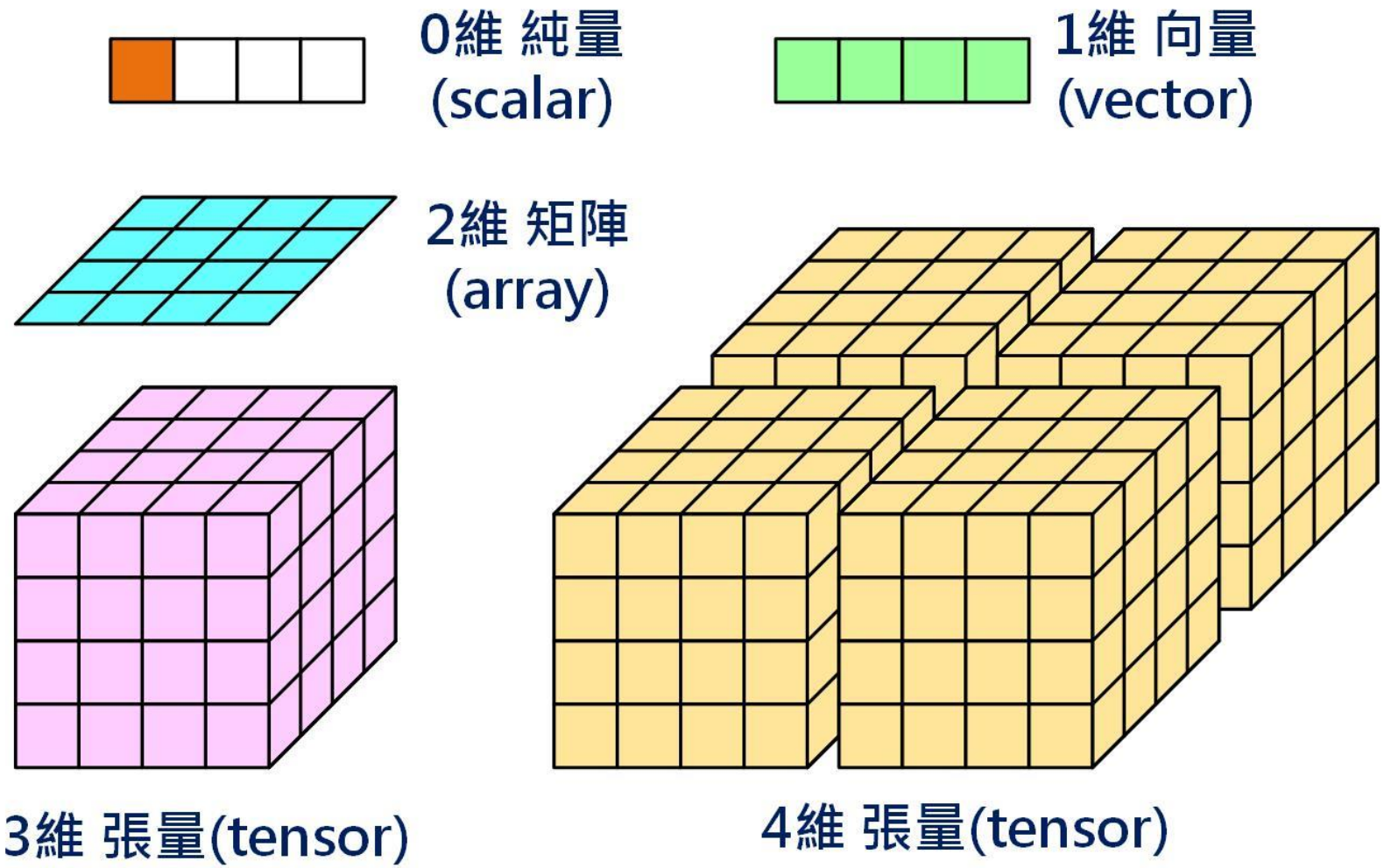
https://www.it120.vip/yq/6142.html



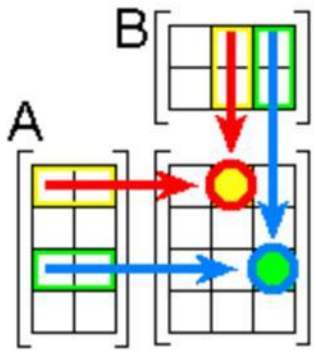
常見大型語言模型
(GB~TB)

https://www.netadmin.com.tw/netadmin/zh-tw/trend/16482C89018F47719FE3DD037AF7CC9D

矩陣 / 張量運算一名詞定義



矩陣 / 張量運算—矩陣乘法



$$(AB)_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}$$

$$(AB)_{1,2} = \sum_{r=1}^{2} a_{1,r} b_{r,2} = a_{1,1} b_{1,2} + a_{1,2} b_{2,2}$$

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & \cdots \\ b_{2,1} & b_{2,2} & b_{2,3} & \cdots \\ b_{3,1} & b_{3,2} & b_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} B_1 & B_2 & B_3 & \ldots \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \end{bmatrix} \times \begin{bmatrix} B_1 & B_2 & B_3 & \ldots \end{bmatrix} = \begin{bmatrix} (A_1 \cdot B_1) & (A_1 \cdot B_2) & (A_1 \cdot B_3) & \cdots \\ (A_2 \cdot B_1) & (A_2 \cdot B_2) & (A_2 \cdot B_3) & \cdots \\ (A_3 \cdot B_1) & (A_3 \cdot B_2) & (A_3 \cdot B_3) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

平行 / 並行運算

平行(Parallelism)

- 將工作分割成多份不同參數但內容相同工作內容，同時執行，待全部完成，再進行下一動作。
- 如矩陣乘法是由很多個AxB+C組合而成，傳統上須依序計算再組合結果，此時就很適合分割成很多小工作一起計算，再合併結果得到所需結果。

並行(Concurrency)

- 可將工作(task)拆分成多個獨立工作，內容可不相同，可同時運行，互不干擾，亦可不同時完成。
- 如編譯一個由很多函式組成的程式，利用多處理器（或多執行緒）方式同時編譯，待全部編譯完成後再連結成執行檔。

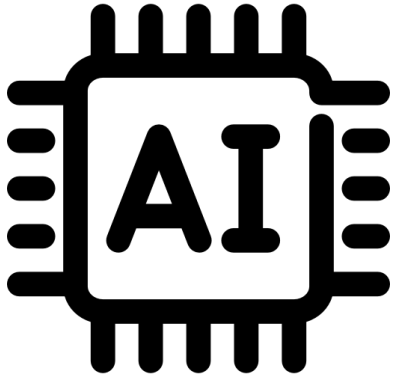
2.2.
加速運算晶片

- 發展方向
- CPU
- DSP
- GPU
- FPGA
- NPU
- ASIC
- LPU
- 類腦晶片
- 光子晶片
- 記憶體內計算
- 可重構晶片

加速運算晶片－發展方向

彈性

通用



效能

專用

CPU	DSP	GPU	FPGA	NPU	ASIC
多核心 高工作頻率	多乘加 飽和運算	超多乘加 高速平行	專用IC前 驗證	神經網路 專用	特定AI 專用IC

加速運算晶片－CPU

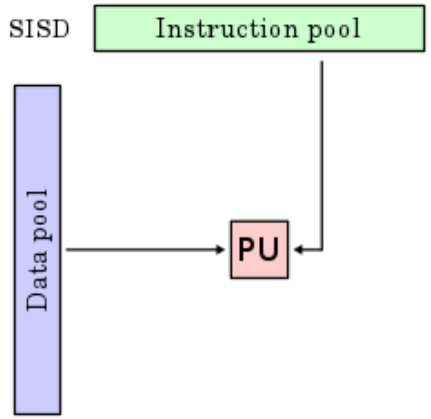
複雜指令集 (CISC)	精簡指令集 (RISC)
• INTEL(x86) • AMD • VIA	• ARM • RISC-V • MIPS

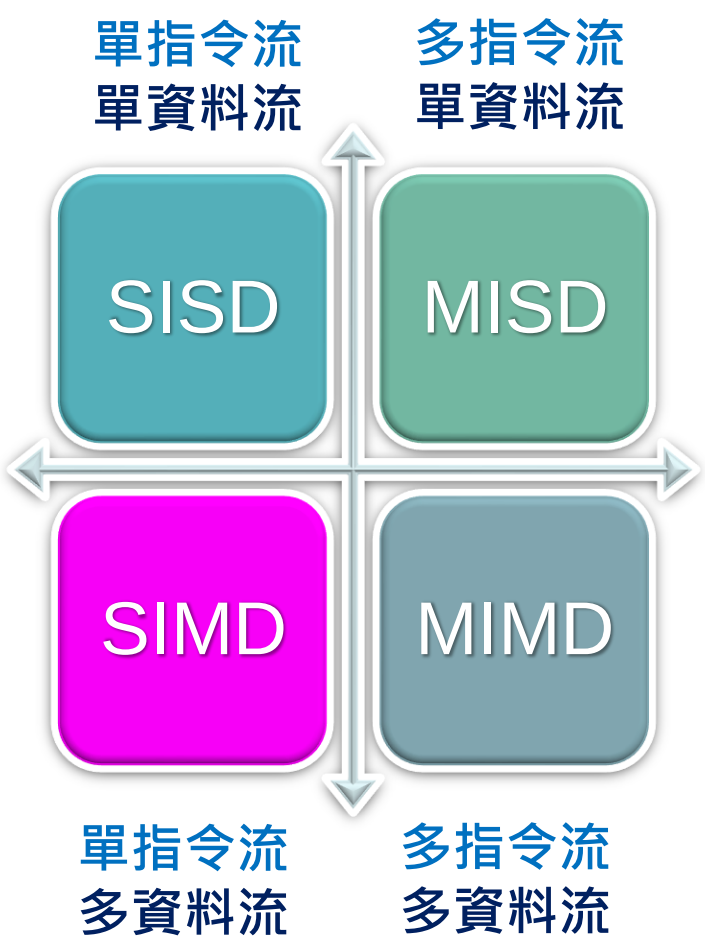
X86 SIMD
MMX (8*64bit)
SSE (128bit)
AVX(256/512bit)

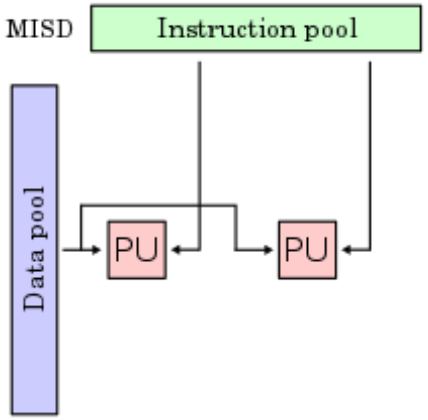
ARM SIMD
Cortex-A
NEON
Cortex-M
DSP / MVE

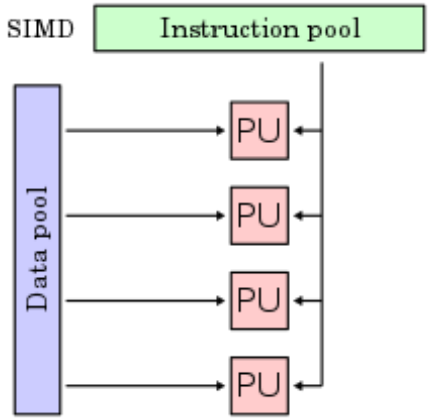
加速方式：
增加工作頻率
增加運算平行度
多核並行運算

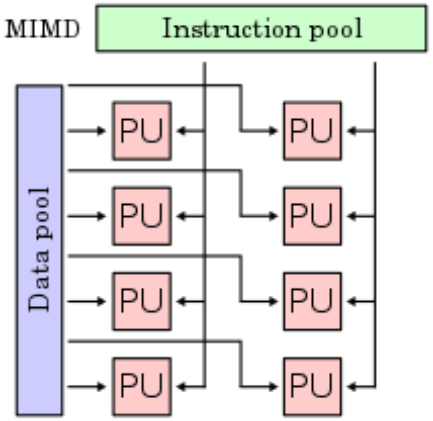
RISC-V
P指令集 (SIMD)
V指令集 (Vector)











加速運算晶片－DSP

數位信號處理器(Digital Signal Processor)

TI	Qualcomm	Cadence	Synopsys
OMAP C6000/5000/2000 DaVinci AM62x	Hexagon	Tensilica * Vision P5/ P6 / C5 * Xtensa Lx6 / Lx7	Arc * EMxD * HS4xD * VPX

ESP32家族

常見聲音及影像應用運算

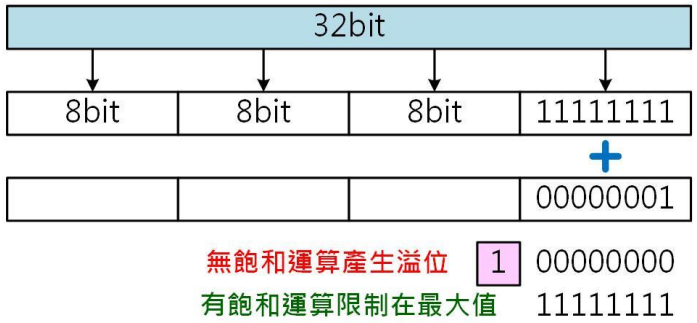
FIR、IIR、DFT、DCT、FFT、Dot product、Convolution及其它矩陣運算等。可進行整數、定點數、浮點數計算。

乘積累加 (Multiply Accumulate, MAC)

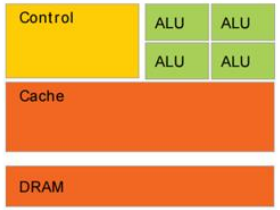
乘法和加法一次作完，如 y = ax + b

飽和運算 (Saturation Arithmetic)

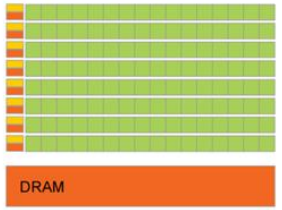
執行SIMD指令，將多位元指令分拆為少位元運算時，避免溢位產生。



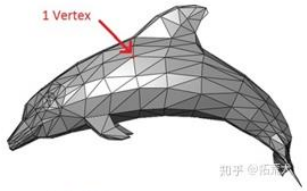
加速運算晶片－GPU



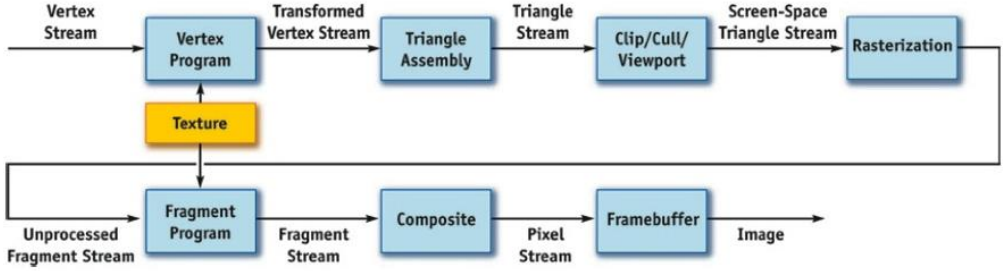
CPU



GPU



渲染(Render)



繪圖處理單元(Graphic Process Unit)

Radeon

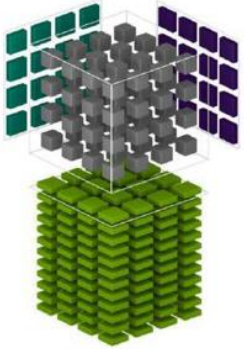
GeForce
Quadro/Tesla
Jetson/DGX

HD Graphic
Iris Xe / Arc

MALI

Pascal

Volta Tensor 核心



加速運算晶片－Nvidia GPU（訓練）

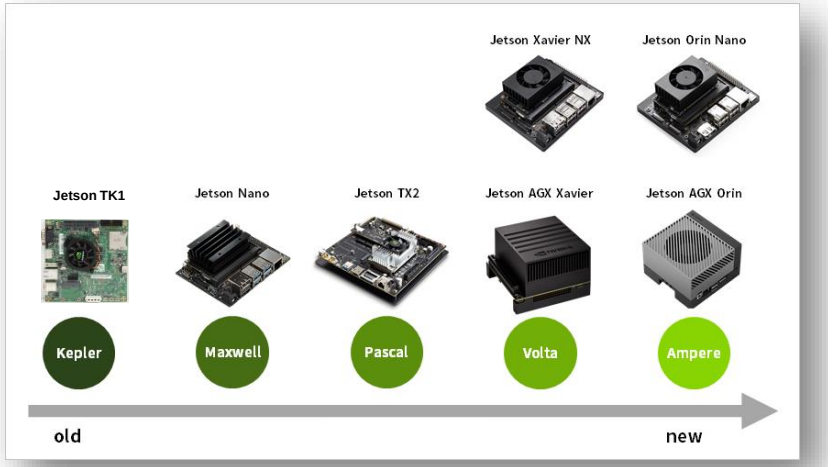
訓練用

NVIDIA GPU	A100	TU102	GV100	GP100
GPU架構	Ampere	Turing	Volta	Pascal
製程	TSMC 7nm	TSMC 12nm	TSMC 12nm	TSMC 16nm
晶片 面積	826mm2	754mm2	815mm2	610mm2
電晶體數目	542億	186億	211億	153億
FP32 CUDA Cores	6912	4608	5120	3584
GPU Boost Clock	1.41GHz	1770MHz	1530MHz	1480MHz
記憶體時脈	2.4Gbps 5120bit	14Gbps 384bit	1.75Gbps 4096bit	1.4Gbps 4096bit
記憶體	40GB HBM2	24GB GDDR6	16/32GB HBM2	16GB HBM2
Single Precision (FP32)	19.5 TFLOPs	16.3 TFLOP	15.7 TFLOPs	10.6 TFLOPs
INT8 Tensor	624 TOPs	261 TOPs	N/A	N/A
FP16 Tensor	312 TFLOPs	130.5 TFLOPs	125 TFLOPs	N/A
TF32 Tensor	156 TFLOPs	N/A	N/A	N/A
TDP	400W	260W	300/350W	300W

加速運算晶片－Nvidia GPU（推論）

推論用

CUDA Core
Tensor Core
NVDLA

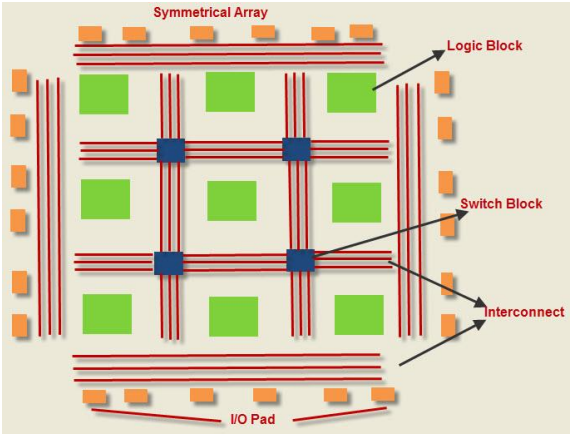


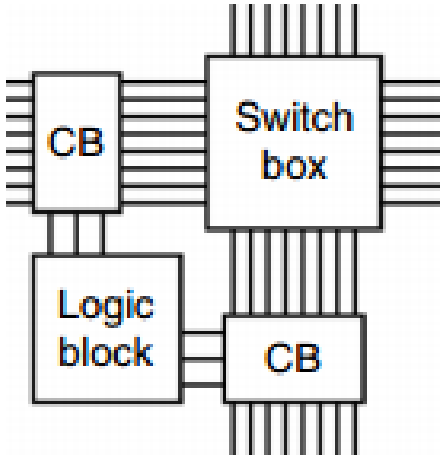
	NANO	TX2 NX	XAVIER NX*	ORIN NX 8 GB	ORIN NX 16 GB	AGX XAVIER**	AGX ORIN 32GB	AGX ORIN 64GB
GPU	128 Core Maxwell 0.5 TFLOPs (FP16)	256 Pascal cores 1.3 TFLOPS (FP16)	384 Core Volta 21 TOPs (INT8)	1024 Cores Ampere NVDLA V2 70 TOPS (INT8)	1024 Cores Ampere 2x NVDLA V2 100 TOPS (INT8)	512 Core Volta + NVDLA 10 TFLOPS (FP16) 32 TOPS (INT8)	1792 Cores Ampere 2x NVDLA V2 200 TOPS (INT8)	2048 Cores Ampere 2x NVDLA V2 275 TOPS (INT8)
CPU	4 core ARM A57	6 core Denver and A57 (2x) 2MB L2	6 core Carmel ARM V8 (3x) 2MB L2 + 4MB L3	6 cores A78 ARM V8 1.5MB L2 + 4MB L3	8 cores A78 ARM V8 2MB L2 + 4MB L3	8 core Carmel ARM V8 (4x) 2MB L2 + 4MB L3	8-core A78 ARM V8 2MB L2 + 4MB L3	12 core A78 ARM V8 3MB L2+ 6MB L3
Memory	4 GB 64-bit LPDDR4 25.6 GB/s	4 GB 128-bit LPDDR4 51.2 GB/s	8GB or 16GB 128-bit LPDDR4x 51.2 GB/s	8 GB 128-bit LPDDR5 102.4 GB/S	16 GB 128-bit LPDDR5 102.4 GB/S	32GB or 64 GB 256-bit LPDDR4x 137 GB/s	32 GB 256-bit LPDDR5 204.8 GB/s	64 GB 256-bit LPDDR5 204.8 GB/s
Storage	16 GB eMMC	16 GB eMMC	16 GB eMMC	- (Supports external NVMe)	- (Supports external NVMe)	32 GB eMMC	64 GB eMMC	64 GB eMMC
Encode	4K @ 30 (H.265)	4K @ 60 (H.265)	2x 4K @ 30 (H.265)	1x 4K @ 60 (H.265)	1x 4K @ 60 (H.265)	4x 4K @ 60 (H.265)	1x 4K @ 60 (H.265)	2x 4K @ 60 (H.265)
Decode	4K @ 60 (H.265)	2x 4K @ 60 (H.265)	2x 4K @ 60 (H.265)	2x 4K @ 60 (H.265)	2x 4K @ 60 (H.265)	6x 4K @ 60 (H.265)	2x 4K @ 60 (H.265)	3x 4K @ 60 (H.265)
Camera	12 (3x4 or 4x2) MIPI CSI-2 D-PHY 1.1 lanes (18 Gbps)	12 lanes (3x4 or 5x2) MIPI CSI-2 D-PHY 1.2 (30 Gbps)	12 lanes (3x4 or 6x2) MIPI CSI-2 D-PHY 1.2 (30 Gbps)	8 lanes MIPI CSI-2 D-PHY 1.2 (20 Gbps)	8 lanes MIPI CSI-2 D-PHY 1.2 (20 Gbps)	16 lanes MIPI CSI-2 | 8 lanes SLVS-EC D-PHY (40 Gbps) C-PHY (59 Gbps)	16 lanes MIPI CSI-2 D-PHY 2.1 (40 Gbps) C-PHY 2.0 (164 Gbps)	16 lanes MIPI CSI-2 D-PHY 2.1 (40 Gbps) C-PHY 2.0 (164 Gbps)
Mechanical	69.6mm x 45mm 260 pin edge connector	69.6mm x 45mm 260 pin connector	69.6mm x 45mm 260 pin edge connector	69.6mm x 45mm 260 pin edge connector	69.6mm x 45mm 260 pin edge connector	100mm x 87mm 699 pin connector	100mm x87 mm 699 pins connector	100mm x87 mm 699 pins connector
Software	JetPack SDK - Unified software release across all Jetson products							

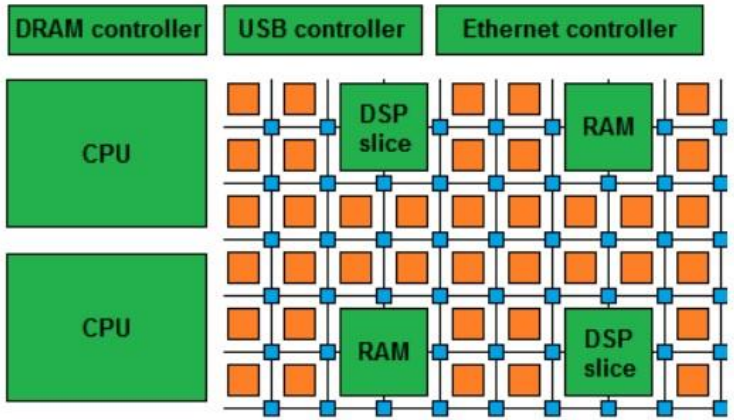
資料來源：https://www.seeedstudio.com/blog/nvidia-jetson-comparison-nano-tx2-nx-xavier-nx-agx-orin/

加速運算晶片－FPGA

現場可程式化邏輯陣列(Field - Programmable Gate Array)







*** Intel(Altera)**
Agilex 7
Cyclone 10 GX
Arria 10

*** AMD(Xilinx)**
Kria K24 / K26

*** Lattice**
Avant G/X

AMD (Xilinx)

Intel (Altera)

Microsemi

Lattice

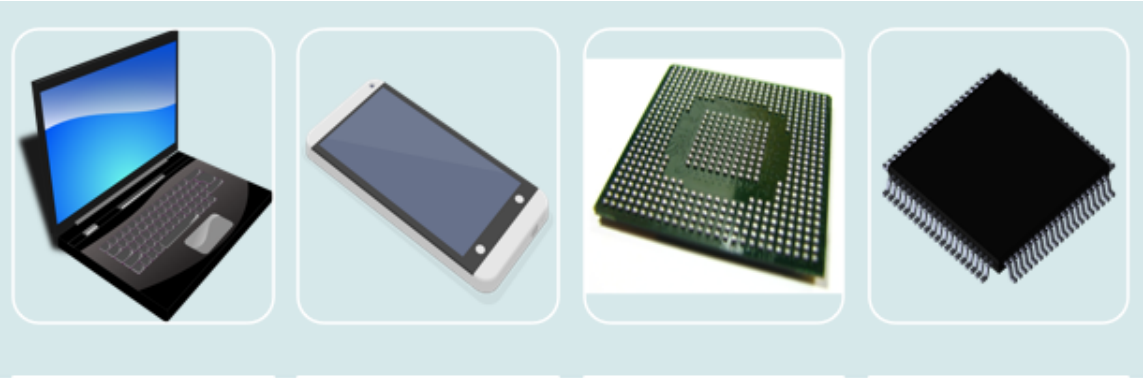
加速運算晶片－NPU

以神經網路運算為主要工作的晶片，屬於ASIC的一種，各廠家稱呼方式略有不同。

常見相關產品

- Google TPU
- Intel(Movidius) MA2485
- 耐能Kneron KL520 / KL720
- ARM Ethos U55 / U65
- 嘉楠勘智 K210 / K510
- 瑞芯微 RK1808 / RK3588

常見組合



CPU+NPU	SoC+NPU	MPU+NPU	MCU+NPU
* **Intel** Core Ultra * **AMD** Ryzen 8040 * **APPLE** M1/M2/M3	* **聯發科** 天璣9300 * **Qualcomm** Snapdragon 8 Gen 3	* **ALIF** Ensemble Arm Cortex-A32+M55+U55 * **Rockchip** RK3588 Arm Cortex-A76+A55+NPU	* **Arm** Cortex-M55 + Ethos U55/U65(**奇景, 新唐, ALIF, Infineon**) * **STM**32N6 * **NXP** NCX N * **ADI** 78000

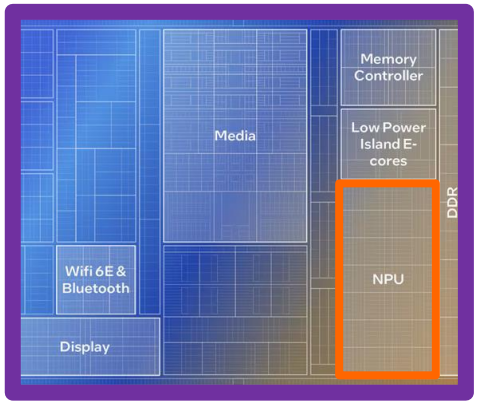
加速運算晶片－Intel AI PC CPU 結構

Intel 14代CPU

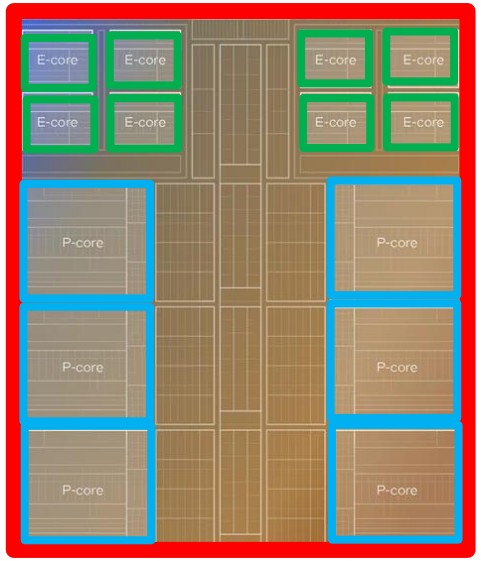
Core Ultra
(Meteor Lake)

CPU + iGPU + NPU

OpenVINO™

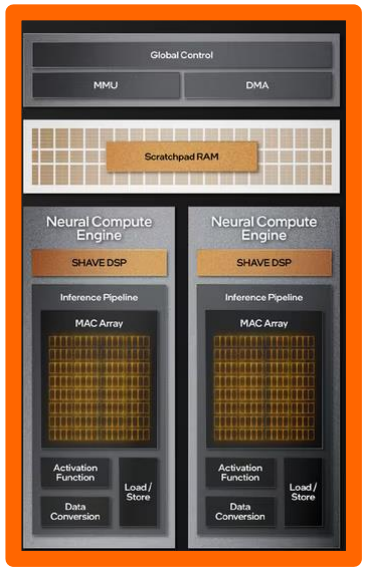


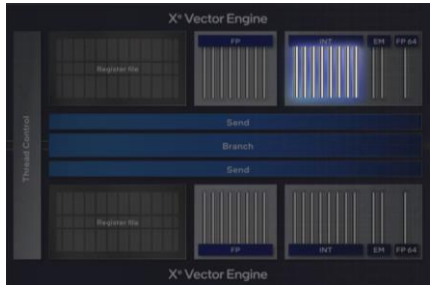
SoC tile



Compute tile

NPU （類似原Movidius神經加速棒 VPU 結構）





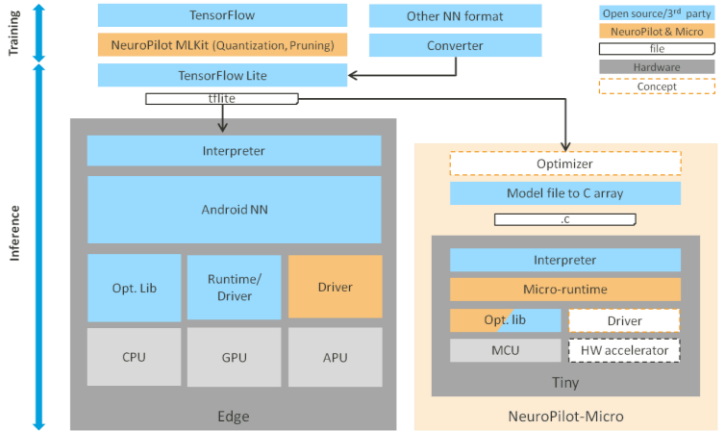
DP4A (GPU)

P-Core x6

E-Core x8

影像來源：https://www.4gamers.com.tw/news/detail/59826/intel-meteor-lake-architecture-overview

加速運算晶片 – 聯發科AI手機SoC結構



CPU : (4nm)
4 x Cortex-X4 CPU @ 3.25GHz
4 x Cortex-A720 @ 2.0GHz
8+10MB Cache Buffer

Memory:
LPDDR5T 9600Mbps
UFS 4.0 Flash, MCQ

GPU : (Immortalis-G720)
12 Cores, 2x MSAA

AI : (APU 790)
Support 33B LLM, LoRA Fusion

ISP : (Imagiq 990)
Video Encoder / Decoder
Semantic Segmentation

Display : (MiraVision 990)
WQHD @ 180Hz, 4K @ 120Hz
AI HDR

資料來源：https://www.mediatek.tw/products/smartphones-2/mediatek-dimensity-9300

加速運算晶片－ASIC（大陸）

華為海思 昇騰(Ascend) 310 / 910

紫光展銳 虎賁T710

地平線 征程2/3, 旭日2/3

寒武紀 思元 290/270/100/200

比特大陸 算丰BM1680 / 1682 / 1684 / 1880

遂原科技 燧思, 雲燧T10 / T11 / i10

雲天勵飛 DeepEye1000

全志科技 R329, R818

瑞芯微 RK3399Pro / RK1808

- **鯤雲科技** CAISA
- **依圖科技** 求索
- **啟英泰倫** CI100X / 110X / 112X
- **知存科技** MemCore001 / 001P / 101 / 201
- **黑芝麻智能** A1000
- **深聰智能** 太行TH1520
- **肇觀電子** N / D / V系列
- **天數智芯** BI
- **探境科技** 音旋風611 / 612 / 621 / 631 / 711 / 712, Imagist851

嘉楠科技 堪智K210 / 510

雲知聲 蜂鳥

清微智能 TX101 / 210 / 510 **(可重構芯片)**

酷芯微 AR9000 / 9200

杭州國芯 GX8002 / 8010 / 8009 / 8008 / 8001

耐能科技 KL520 / 720

平頭哥 含光800

百度 崑崙1 / 2

北京君正 T02 / T31

應用情境

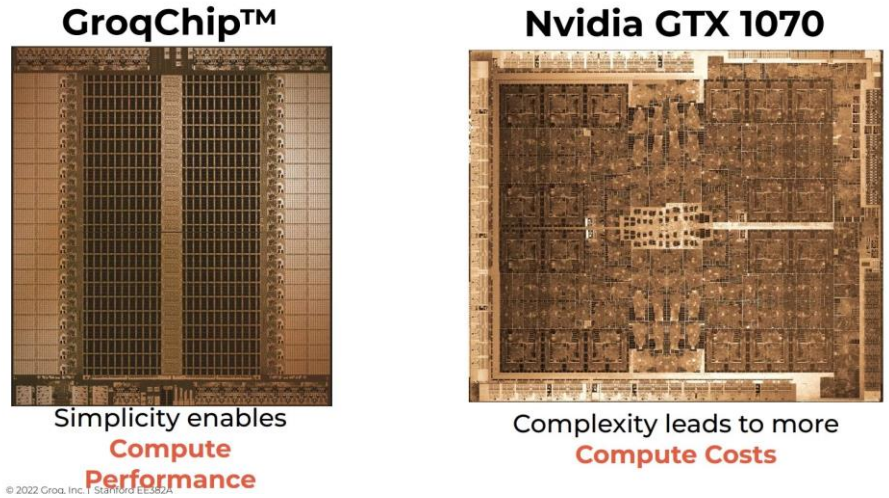
- 智慧音箱（語音命令）
- 自駕車
- 安全監控
- 人臉 / 語音辨識
- 智慧製造 / 倉儲 / 教育
- 智慧城市 / 家庭 / 生活
- 智慧醫療 / 交通 / 金融
- **雲端機房訓練 / 推理**
- 視頻分析
- 物聯網(AIoT)
- 機器人視覺
- 可穿戴裝置

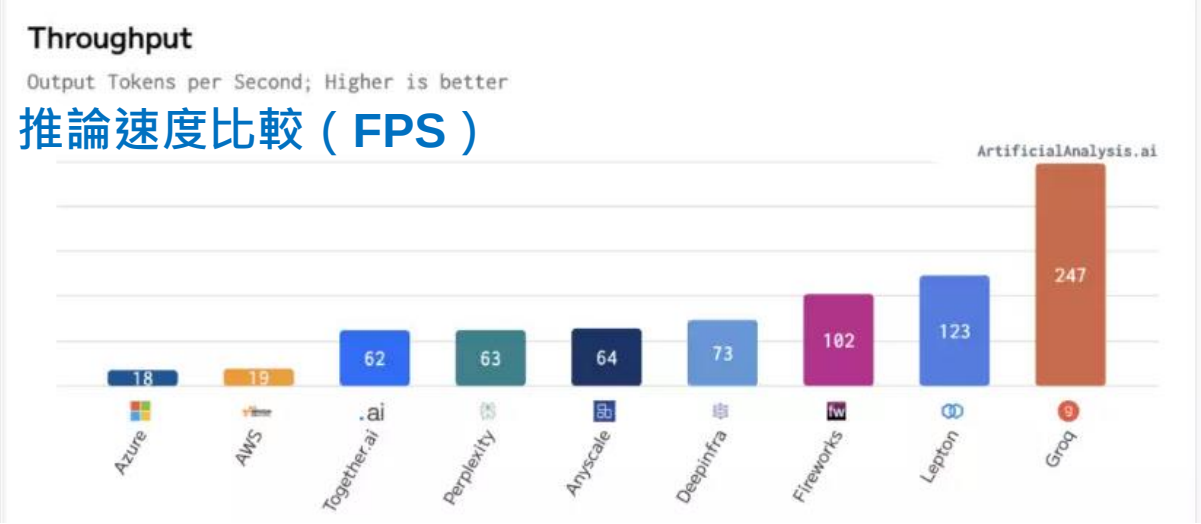
資料來源：https://www.esmchina.com/news/7491.html

加速運算晶片－ASIC（LPU）

Groq LPU（推論用）
(Language Processing Unit)

擁有230MB靜態隨機存取記憶體（SRAM）相當於每秒80TB的記憶體頻寬。實測大語言模型Llama 2 Chat (70B) 每秒能輸出247個tokens，比Azure 18個tokens快14倍多，推理速度是英偉達GPU的10倍。





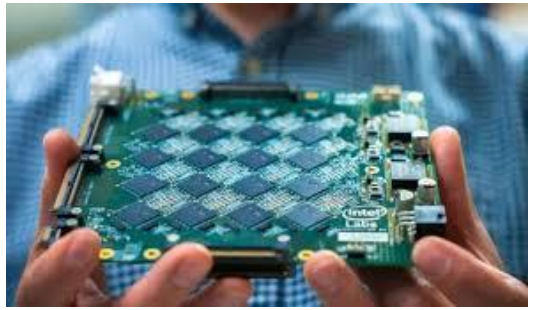
資料來源：https://www.edigest.hk/%e7%a7%91%e6%8a%80/ai%e6%99%b6%e7%89%87groq-chatgpt-nvidia-gpu-nvda-amd-intel-1250142/5/

加速運算晶片 – 類腦晶片

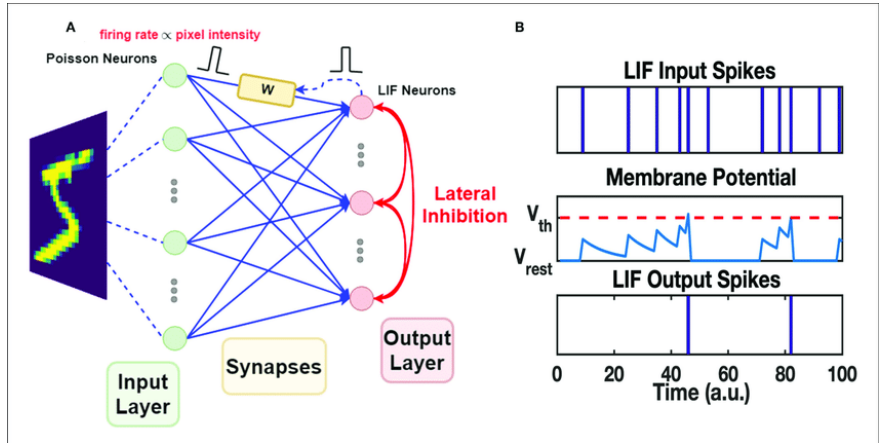
Neuromorphic 神經形態

- **數位式**
  - IBM TrueNorth
  - INTEL Loihi
  - SpiNNaker
  - BrainChip
- **類比式**
  - Neurogrid
  - BrainScales
  - ROLLS
- **新材料**
  - Memristor

Brainchip
Akida AKD1000

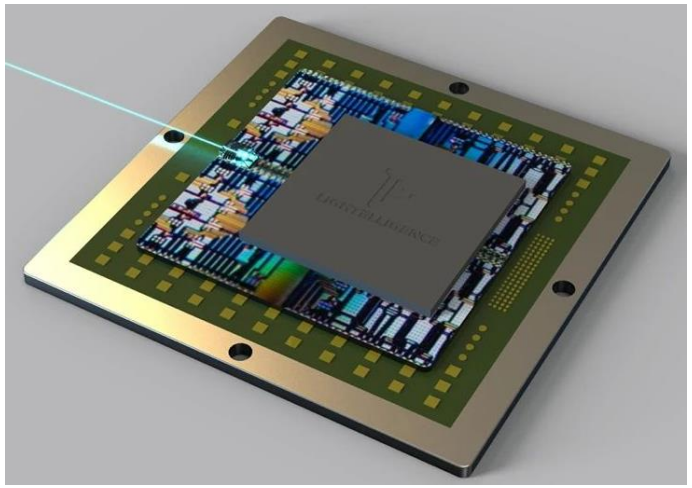


INTEL Pohoiki Beach
(64 x Loihi Chip)



脈衝神經網路
Spiking Neural Networks

加速運算晶片－光子晶片



Lightmatter

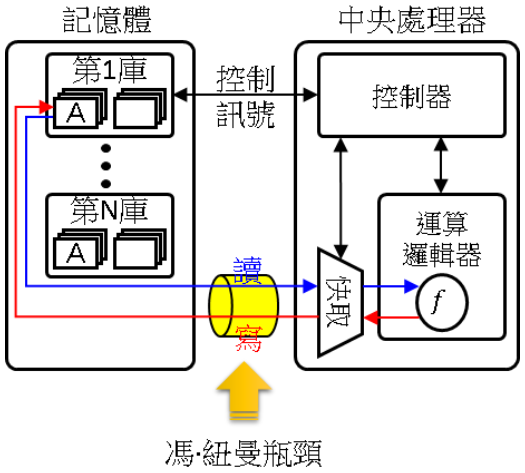
- MIT獨立出新創公司
- 利用矽光子和微機電製程技術，以毫瓦等級雷射光供電，能以接近光速執行矩陣向量乘法。

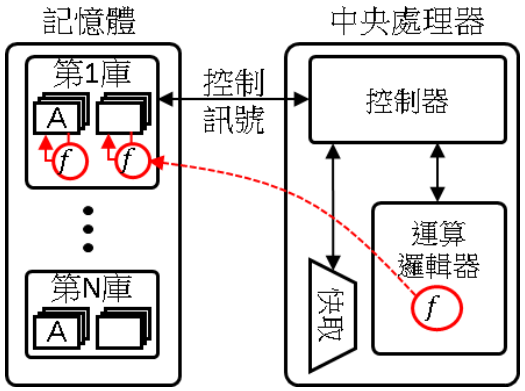
曦智lightelligence

- 大陸團隊，百度投資
- 和Lightmatter皆為MIT技術，技術類似。

資料來源：https://www.eettaiwan.com/20201006nt31-optical-compute-promises-game-changing-ai-performance/
資料來源：https://www.lightelligence.co/

加速運算晶片 – 記憶體內計算



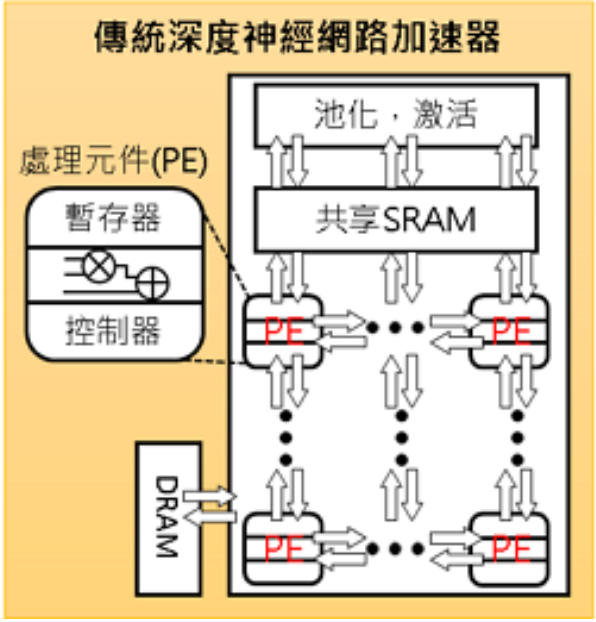


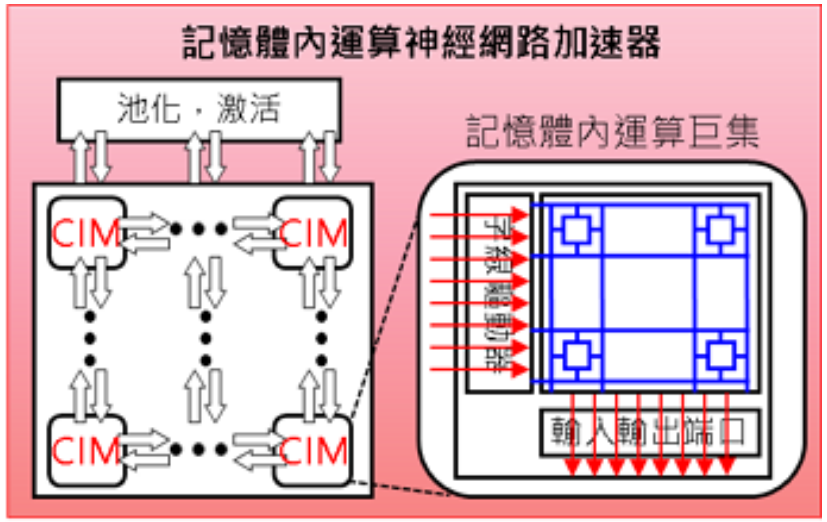
In-Memory Computing 存算一體、存內計算

減少記憶體到計算單元距離來加速運算

*** SRAM (靜態隨機存取記憶體 / 電晶體式)**

*** MRAM (磁阻式) * RRAM (可變電阻式)**





資料來源：
https://ictjournal.itri.org.tw/content/Messagess/contents.aspx?PView=1&SiteID=654246032665636316&MmmID=654304432061644411&SSize=10&MSID=1037365734470512074

加速運算晶片－可重構晶片

粗粒度可重構架構（Coarse Grain Reconfigurable Architecture, CGRA）

- 可以使用軟體動態變更硬體架構，包括乘加器(MAC)位元寬度、數值精度、矩陣處理結構、運算處理單元(PE)等，以適應不同神經網路運算效能。

常見供應商及產品

- 耐能（Kneron）－ KL520 / 720 / 530
- 清微智能（大陸清華魏少軍教授）－ TX210 / 510
- 雲天勵飛 － DeepEye 1000
- 燧原科技 － 邃思 T10 / T20

2.3.
開發板類型

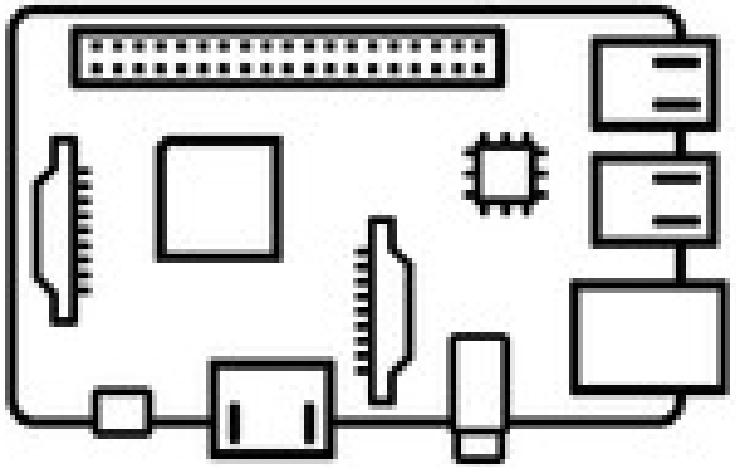
- 單板微電腦
- USB加速棒
- 模組板卡
- 整合型裝置
- TinyML開發板

單板微電腦類型

Intel

Nvidia

Google



ARM

- Raspberry Pi
- Asus Tinker
- Rockchip
- Arduino
- Banana Pi

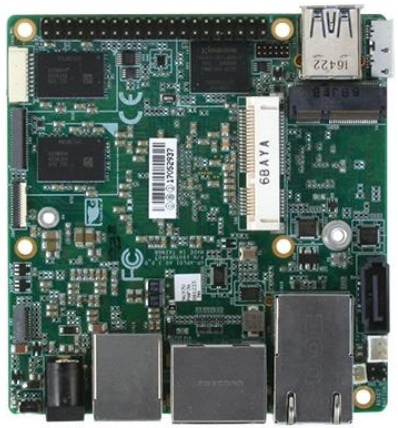
FPGA

- Avnet
- Microsemi
- Xilinx

RISC-V

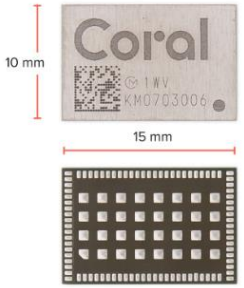
- Renesas
- Asus Tinker V

單板微電腦 (1/2)

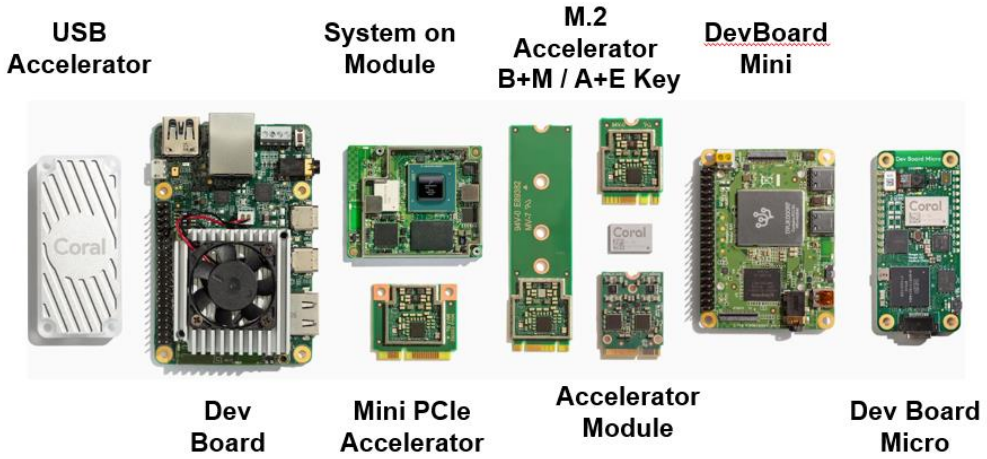


AAEON UP Squared
(Pentium N4200 / Celeron N3350)

Nvidia Jetson



Google Edge TPU



Asus Tinker Edge T

單板微電腦 (2/2)

樹莓派家族

ARM Based CPU
+Broadcom GPU
VideoCore IV

Pi 0	ARMv6z(32bit)
Pi 3B	ARMv8-A(32/64bit)
Pi 3A+	ARMv8-A(32/64bit)
Pi 4	ARMv8-A(32/64bit) (VFPv4 + NEON)

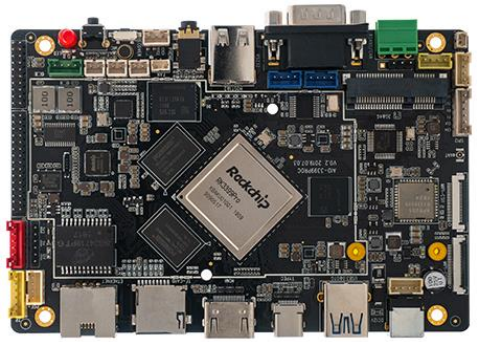
Pi 3A+ (4 * Cortex-A53 1.4GHz)

Pi 0 (Arm11 1GHz)

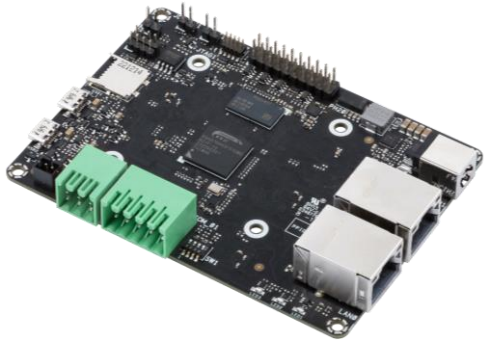
Pi 3B (4 * Cortex-A53 1.4GHz)

Pi 4B (4 * Cortex-A72 1.5GHz)

Pi 5 (4 * Cortex-A76 @2.4GHz)
with mPCIe 2.0

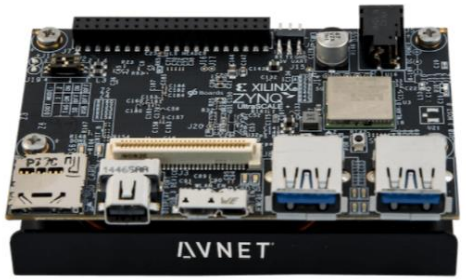


Rockchip RK3399Pro (with NPU)



Asus Tinker V (RISC-V)

AMD(Xilinx) FPGA Kria KR260



AVNET FPGA Ultra96-V2

USB加速棒

Intel NCS 2

- Movidius VPU MA2485
- USB 3.0 / 2.0
- 4 TOPS
- Support OpenVINO

耐能KL520 AI Dongle

- KL520 (Arm Cortex-M4 x2 + NPU)
- 345GFlops
- 576MAC/Cycle (@INT8)

模組板卡

SOM

- **RockChip SOM**
  - RK3399Pro
  - Cortex-A72 * 2 + Cortex-A53 * 4
  - NPU 3 TOPS(INT8/16)
  - GPU Mali-T860MP4

- **Google Coral SOM**
  - Edge TPU
  - NXP i.MX 8M SoC (quad Cortex-A53, Cortex-M4F)
  - 4 TOPS(INT8), 2 TOPS/W
  - 1GB RAM / 8GB eMMC

M.2 / mPCIE

- **AI CORE XM 2280**
  - Intel Movidius Myriad X VPU, MA2485 * 2
  - Support Tensorflow, Caffe, MXNET
  - M.2 2280 B+M Key
  - Support Intel OpenVINO toolkit

AAEON

- **AI CORE X**
  - Intel Movidius Myriad X VPU, MA2485 * 1
  - Support Tensorflow, Caffe, MXNET
  - mini PCIE (mPCIe)
  - Support Intel OpenVINO toolkit

AAEON

整合型裝置

家用智慧音箱

工業智慧攝影機

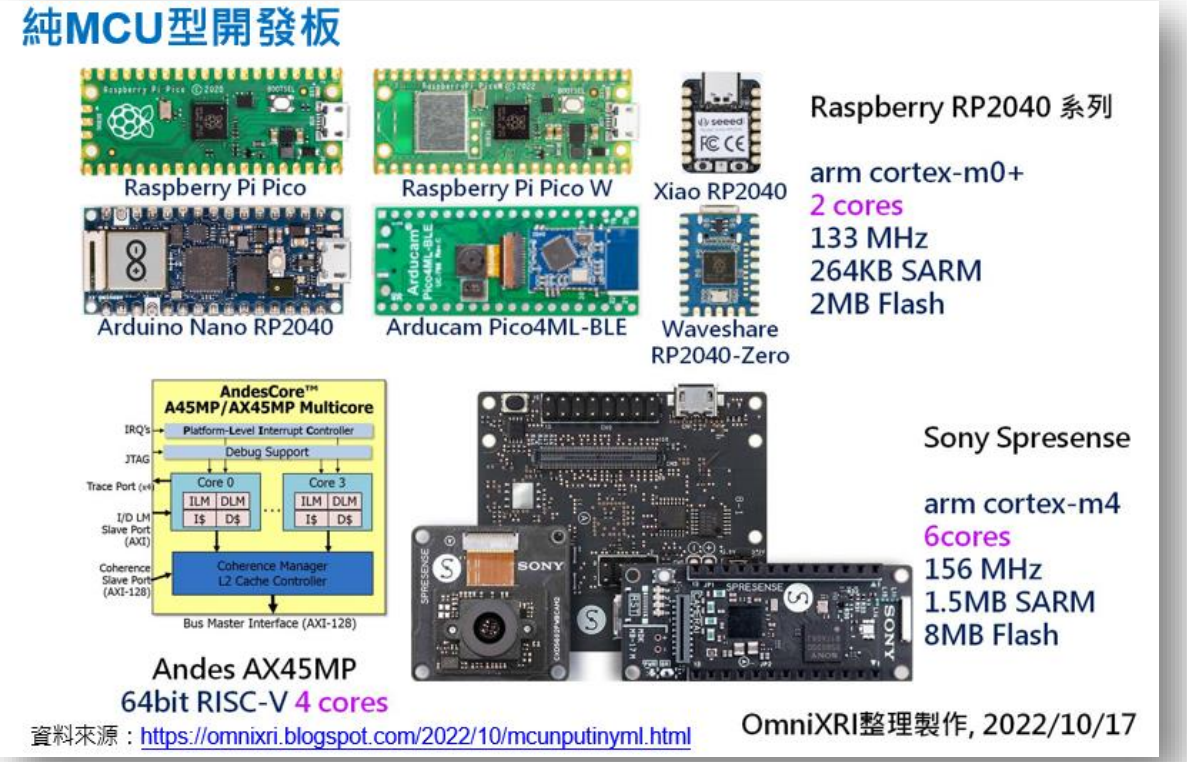
ADLINK NEON-1000-MDX

Intel Atom E3930 + Movidius VPU

ADLINK NEON-2000-JNX

NVIDIA Jetson Xavier NX-based

TinyML開發板



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

資料來源：https://omnixri.blogspot.com/2022/12/tinymlmcu.html

2.4.
硬體選用評估

- 運算效能
- 開發工具
- 應用情境
- 週邊擴展

運算效能 – 主要指標

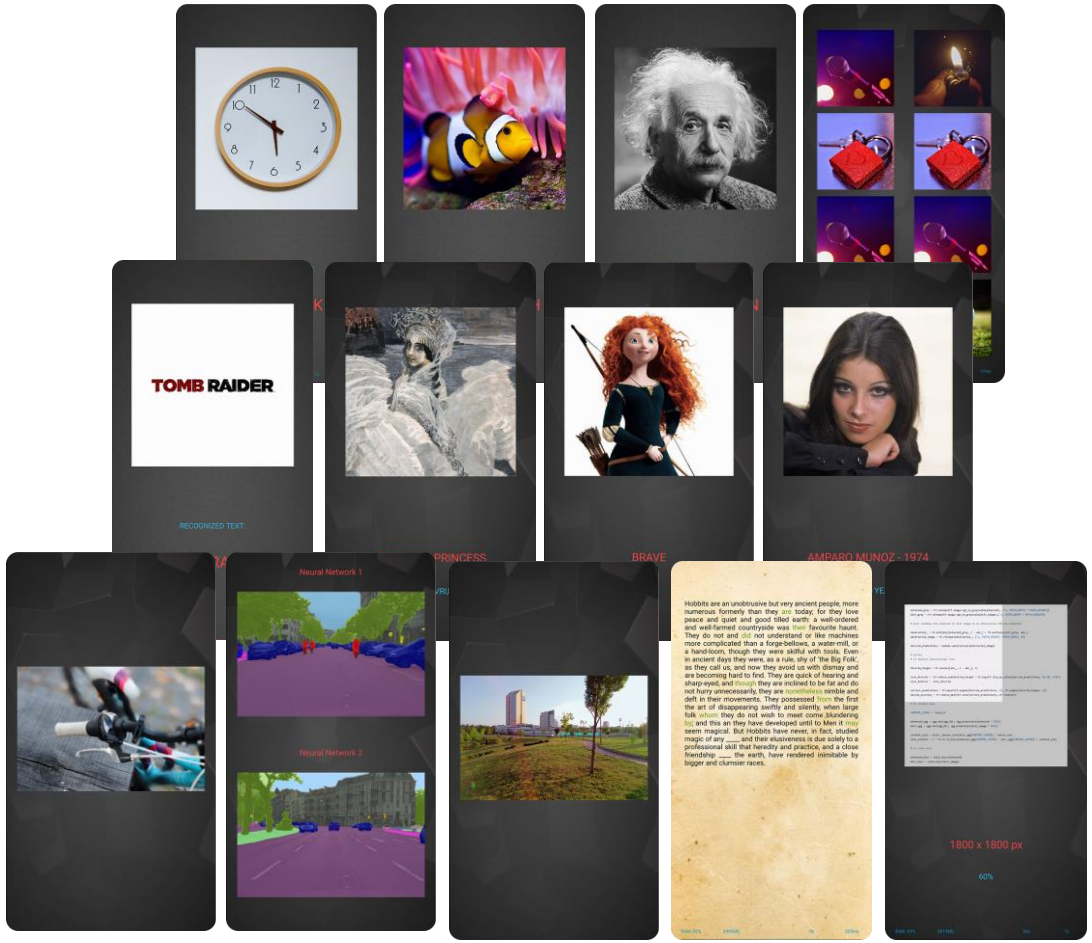
主要指標

- TOPS（每秒運算次數, 單位Tera,10^{12}）
- TOPS / W （每瓦運算次數）
- 算力成本（每塊錢得到之TOPS）
- 數值精度（INT8 / INT16 / FP16 / FP32）
- Top 1 (Top 5)運算精確度（Accuracy）
- 均值平均精確度(mean Average Precision, mAP)
- 推論速度（Frame per Second, FPS）
- 批量/延遲（Batch Size / Delay）
- 整體功耗（Power Consumption, W）

運算效能－AI-Benchmark

ETH 手機晶片AI測試

- Object Classification (MobileNet-V2 / Inception-V3 / MobileNet-V2)
- Face Recognition (MobileNet-V3 Large-M)
- Optical Character Recognition(CRNN / Bi-LSTM)
- Image Deblurring (PyNET-Mini)
- Image Super-Resolution(VGG-19 / SRGAN)
- Bokeh Simulation (U-Net)
- Semantic Segmentation (DeepLab-V3+)
- Photo Enhancement (DPED-ResNet)
- Text Completion (Static RNN / LSTM)
- Memory Limits (SRCNN 9-5-5)



資料來源：https://ai-benchmark.com/

運算效能 – ML COMMONS

Training

MLPerf

已併入

ML Commons

Area	Benchmark	Dataset	Quality Target	Reference Implementation Model
Vision	Image classification	ImageNet	75.90% classification	ResNet-50 v1.5
Vision	Object detection (light weight)	COCO	23.0% mAP	SSD
Vision	Object detection (heavy weight)	COCO	0.377 Box min AP and 0.339 Mask min AP	Mask R-CNN
Language	Translation (recurrent)	WMT English-German	24.0 Sacre BLEU	NMT
Language	Translation (non-recurrent)	WMT English-German	25.00 BLEU	Transformer
Language	NLP	Wikipedia 2020/01/01	0.712 Mask-LM accuracy	BERT
Commerce	Recommendation	1TB Click Logs	0.8025 AUC	DLRM
Research	Reinforcement learning	Go	50% win rate vs. checkpoint	Mini Go (based on Alpha Go paper)

Inference: Edge
Inference: Tiny

Area	Task	Model	Dataset	QSL Size	Quality	Multi-stream latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32 (76.46%)	50 ms
Vision	Object detection (large)	SSD-ResNet34	COCO (1200x1200)	64	99% of FP32 (0.20 mAP)	66 ms
Vision	Object detection (small)	SSD-MobileNets-v1	COCO (300x300)	256	99% of FP32 (0.22 mAP)	50 ms
Vision	Medical image segmentation	3D UNET	BraTS 2019 (224x224x160)	16	99% of FP32 and 99.9% of FP32 (0.85300 mean DICE score)	N/A
Speech	Speech-to-text	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 - WER, where WER=7.452253714852645%)	N/A
Language	Language processing	BERT	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 (f1_score=90.874%)	N/A

資料來源：https://mlcommons.org/

開發工具 – 語言與工具

主要程式語言

- C / C++
- Python
- ASM

平行處理語言

- GPGPU
- CUDA
- OpenCL
- OpenVX

作業系統

- Linux (Raspbain, Ubuntu)
- Windows / Mac
- Android / iOS
- ROS, Mbed, RTOS….

開源工具

- OpenCV
- Google Colab
- Intel OpenVINO

開發工具－AI框架

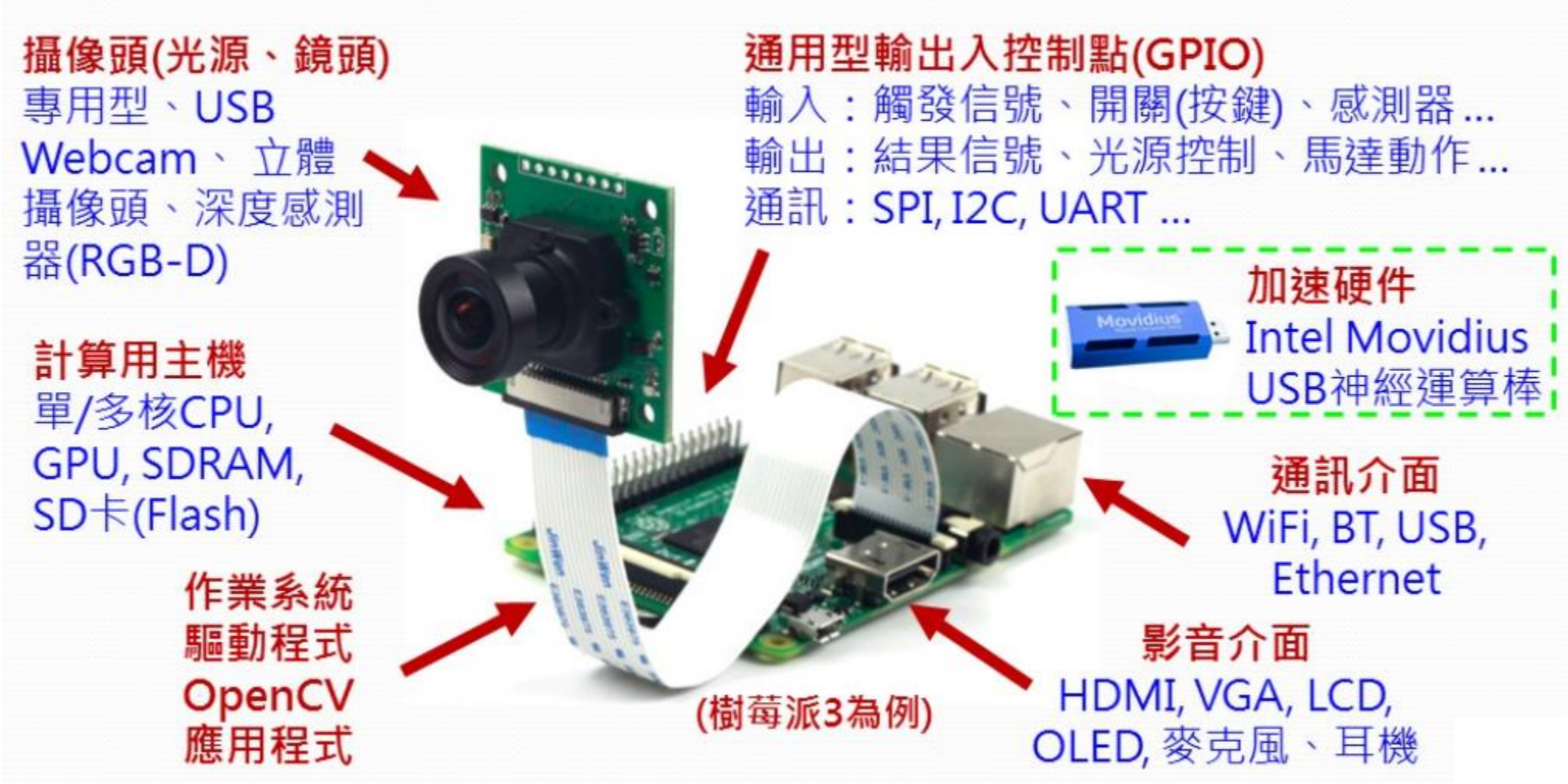
TensorFlow 2.0
Keras
Chainer
Marvin
neon
tiny-dnn
Caffe
USI/SUPSI IDSIA
Microsoft CNTK
CoreML
Torch
Paddle
DL4J
ONNX
theano
BigDL APACHE Spark
Caffe2
dy/net
PYTORCH
mxnet
TensorFlow Lite
TensorRT
OpenVINO
arm CMSIS-NN

應用情境

單晶片	單板微電腦	桌機/伺服器
MCU	SBC	PC
感測分析 語音命令 運動偵測	影像分類 物件偵測 語音辨識	模型訓練 影像分割 自然語言

運算速度 /
平均功耗 /
單位成本

週邊擴展


攝像頭(光源、鏡頭)
專用型、USB Webcam、立體攝像頭、深度感測器(RGB-D)
計算用主機
單/多核CPU, GPU, SDRAM, SD卡(Flash)
作業系統
驅動程式
OpenCV
應用程式
通用型輸出入控制點(GPIO)
輸入：觸發信號、開關(按鍵)、感測器...
輸出：結果信號、光源控制、馬達動作...
通訊：SPI, I2C, UART ...
加速硬件
Intel Movidius USB神經運算棒
Movidius
通訊介面
WiFi, BT, USB, Ethernet
影音介面
HDMI, VGA, LCD, OLED, 麥克風、耳機
(樹莓派3為例)

小結

- **基本運算原理**小節中介紹了基本運算原理，包括 卷積神經網路、 數字表示系統、矩陣 / 張量運算及 平行 / 並行運算 。
- **加速運算晶片**小節中介紹各種晶片型式，包括CPU, GPU, VPU, FPGA, ASIC及其它類型晶片。
- **開發板類型**小節中介紹各種單板微電腦、USB加速棒、不同介面模組板卡及智慧音箱和智慧攝影機。
- **硬體選用評估**小節中分別介紹和運算效能相關的名詞、各種開發工具框架，並延至應用情境及未來如何將週邊擴展。

參考文獻

- 許哲豪， “NTUST Edge AI 2023 人工智慧與邊緣運算實務 2.邊緣運算硬體”
https://github.com/OmniXRI/NTUST_EdgeAI_2023/tree/main/Ch2_Hardware
- 許哲豪， “人工智慧AI晶片與Maker創意接軌”
http://omnixri.blogspot.com/2018/05/aimaker.html
- 許哲豪， “AI 邊緣運算即將掀起風潮”
http://omnixri.blogspot.com/2018/07/aicolumnai.html
- 許哲豪， “AI晶片如何評比效能”
http://omnixri.blogspot.com/2019/03/ai.html
- 許哲豪， “【vMaker Edge AI專欄 #02】 要玩AI前，先來認識數字系統”
https://hackmd.io/@OmniXRI-Jack/vMaker_EdgeAI_02_Digital
- 許哲豪， “【vMaker Edge AI專欄 #03】 AI晶片發展歷史及最新趨勢 ”
https://hackmd.io/@OmniXRI-Jack/vMaker_EdgeAI_03_AIChip
- 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， “有了TinyML加持MCU也能開始玩電腦視覺了 ”
https://omnixri.blogspot.com/2022/12/tinymlmcu.html

延伸閱讀

- 許哲豪， “樹莓派GPU的算力釋放”
http://omnixri.blogspot.com/2018/07/gpu.html
- 張臣雄，AI硬體專屬晶片最新技術未來創新發展
https://www.books.com.tw/products/0010910226
- 許哲豪，【vMaker Edge AI專欄 #07】TinyML (MCU AI) 運行效能誰說了算？
https://omnixri.blogspot.com/2023/07/vmaker-edge-ai-07tinyml-mcu-ai.html
- MLCommons – Inference: Edge
https://mlcommons.org/benchmarks/inference-edge/
- MLCommons – Inference: Tiny
https://mlcommons.org/benchmarks/inference-tiny/
- AI-Benchmark
https://ai-benchmark.com/

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