本課程為教育部109年度智慧聯網技術課程推廣計畫之補助課程
教材內容也自教育部智慧聯網技術重點模組教材改編

Jetson Nano Hardware Platform for AI Edge Computing

謝東佑
國立中山大學電機系

Office: 工EC-7038
07-5252000 Ext. 4114
tyhsieh@mail.ee.nsysu.edu.tw

人工智慧實現方式

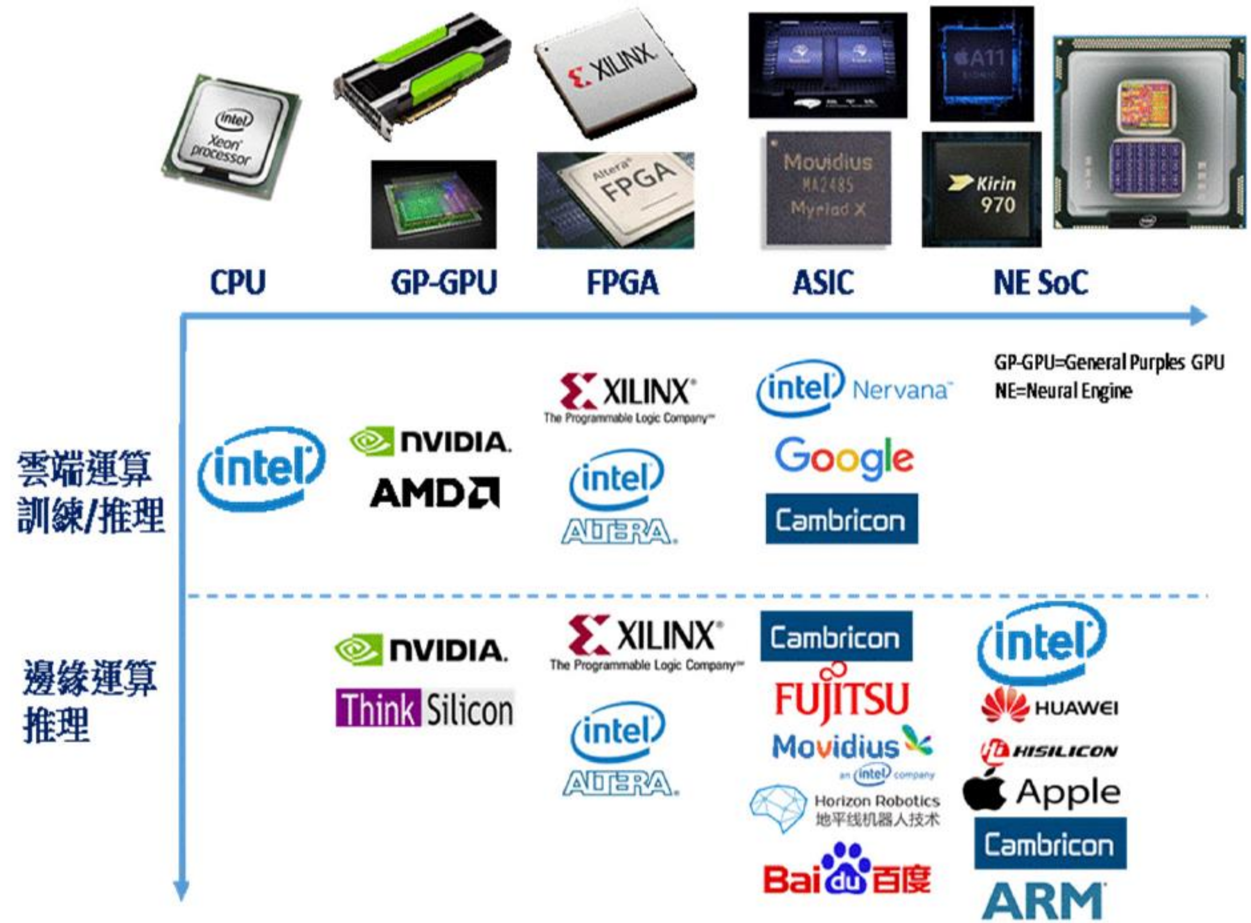
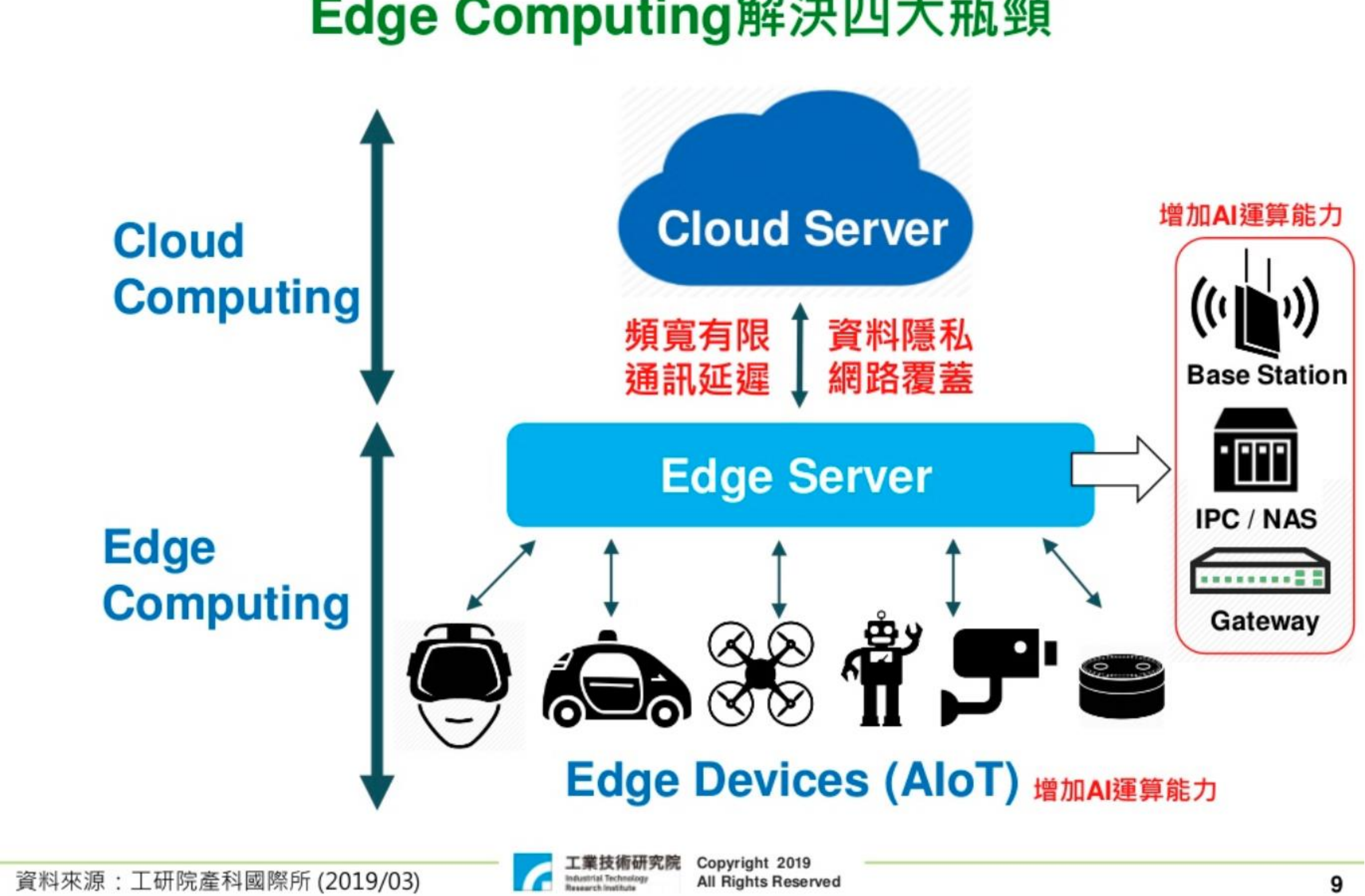


圖 5 分眾化 AI 邊緣運算專用晶片

資料來源：工研院 IEK Consulting（2018）

Source:蘇孟宗，陳右怡，"人工智慧驅策臺灣產業跨域創新"國土及公共治理季刊，6(4)，2018/12



Source: https://www.slideshare.net/ssuserb5d2d4/ss-144047424

AI+Edge Computing四大應用場域

建築(building-scale)

工廠/醫院內部影像辨識、機器人控制、機台數據分析、醫療診斷等....

家庭(room-scale)

家庭內如智慧音箱、智慧家電

個人(personal-scale)

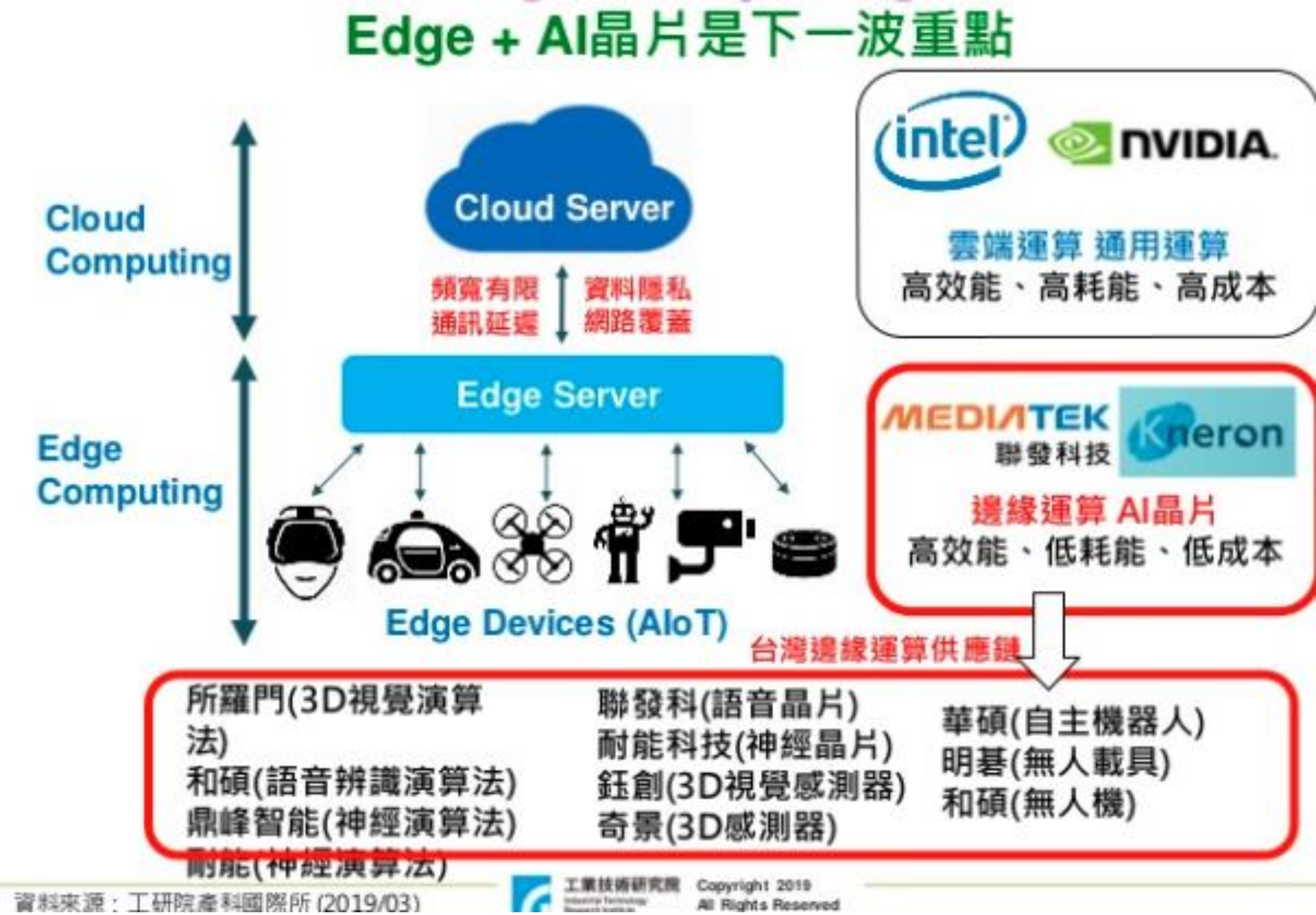
手機、平板、穿戴裝置、AR/VR頭盔....

城市(city-scale)

自駕車、無人機、自走載具、街頭監控系統....

資料來源：工研院產科國際所 (2019/03)

工業技術研究院 Industrial Technology Research Institute Copyright 2019 All Rights Reserved


Edge + AI晶片是下一波重點
Cloud Computing
Cloud Server
頻寬有限
通訊延遲
資料隱私
網路覆蓋
Edge Computing
Edge Server
Edge Devices (AIoT)
intel
NVIDIA
雲端運算 通用運算
高效能、高耗能、高成本
MEDIATEK
聯發科技
Kneron
邊緣運算 AI晶片
高效能、低耗能、低成本
台灣邊緣運算供應鏈
所羅門(3D視覺演算法)
和碩(語音辨識演算法)
鼎峰智能(神經演算法)
耐能(神經演算法)
聯發科(語音晶片)
耐能科技(神經晶片)
鈺創(3D視覺感測器)
奇景(3D感測器)
華碩(自主機器人)
明碁(無人載具)
和碩(無人機)
資料來源：工研院產科國際所(2019/03)
工業技術研究院
Copyright 2019
All Rights Reserved

邊緣運算需求分析

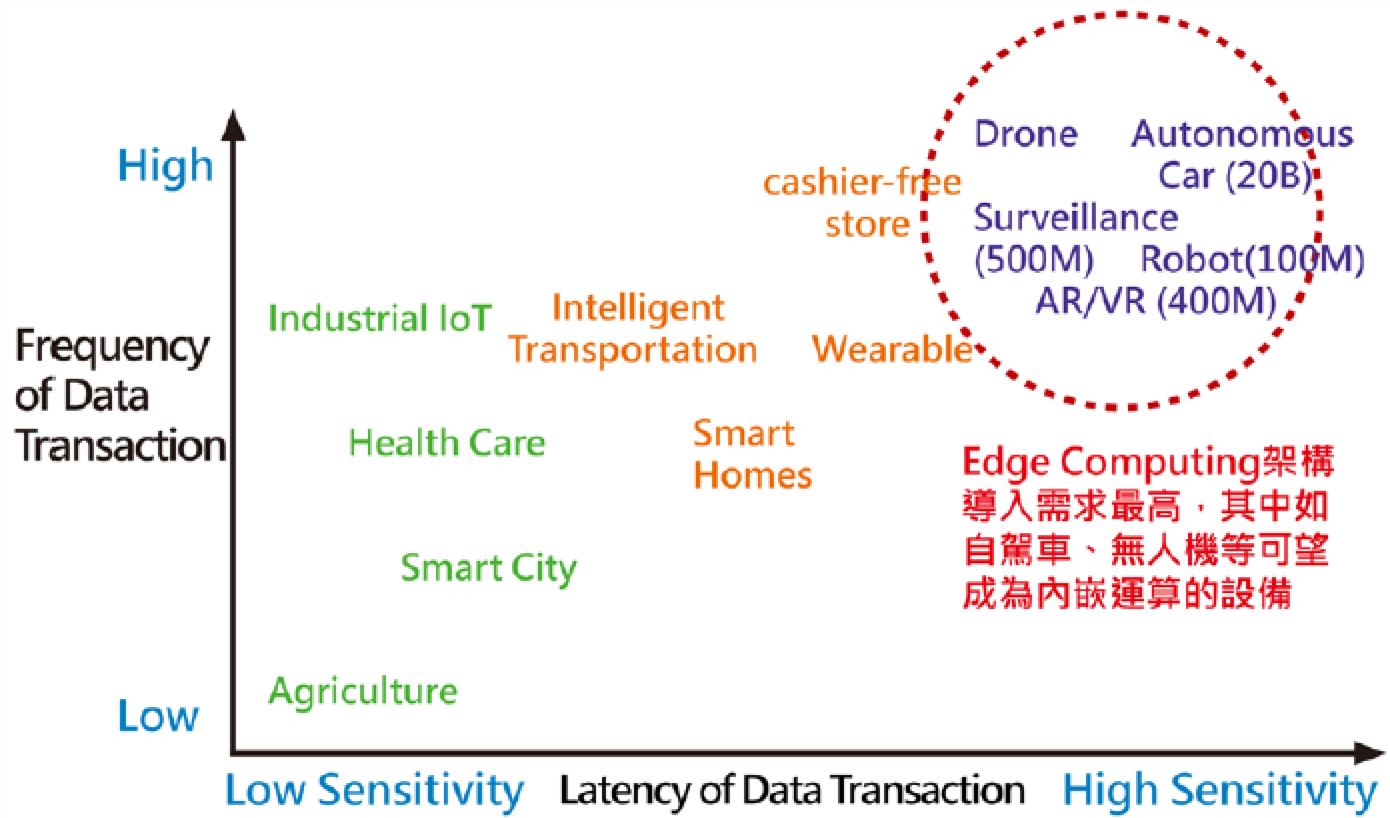
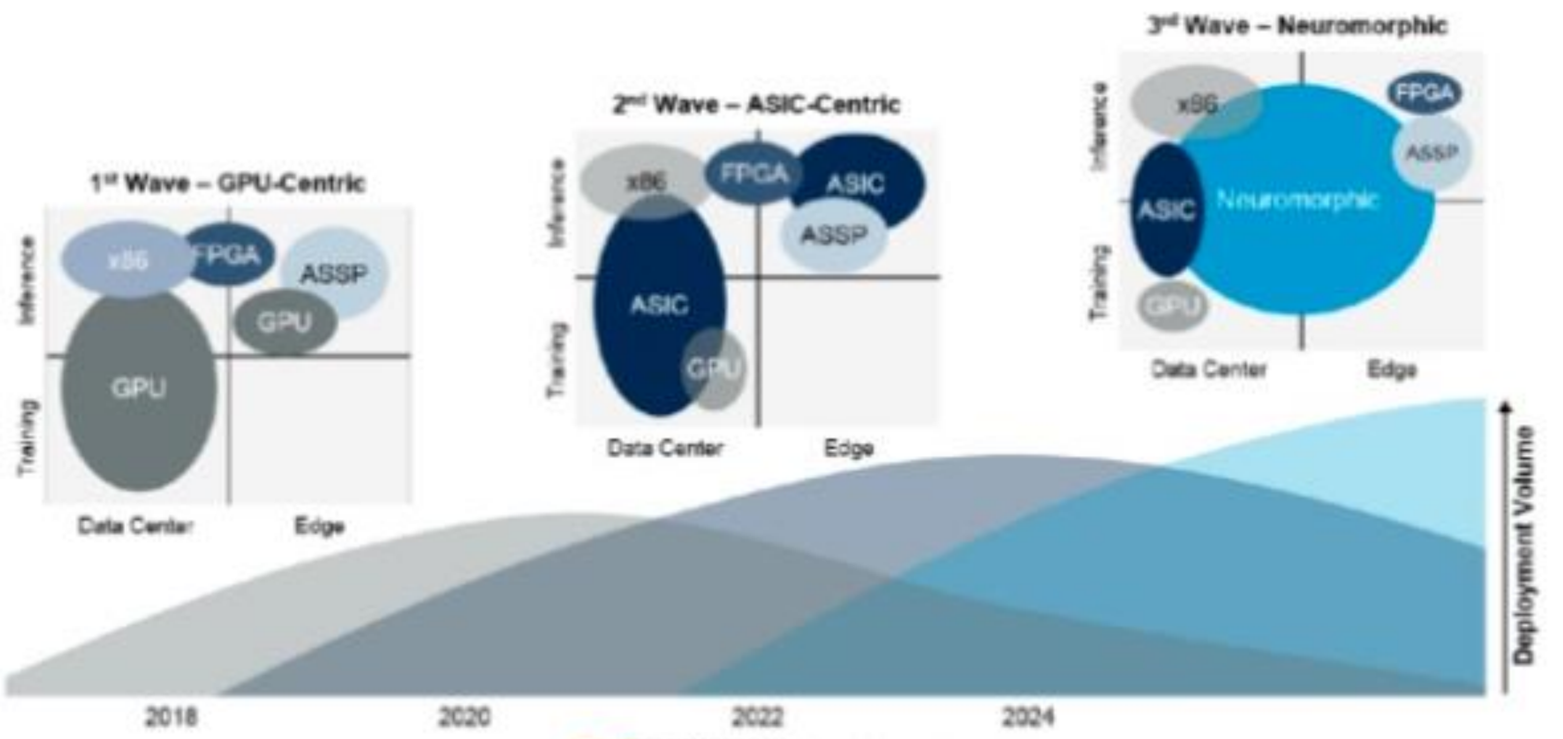


圖 3 五大終端載具對於 AI 邊緣運算需求分析

資料來源：NTT；工研院 IEK Consulting（2018）

AI專用運算晶片三波發展階段

- 第一波為NVidia為首，利用通用GPU進行AI運算，缺點為耗電與昂貴
- 第二波為專屬客製化ASIC晶片，缺點為研發成本高昂
- 第三波將為新架構晶片，目前以類腦架構(Neuromorphic)為主要發展方向



資料來源：Gartner (2018)

Copyright 2019 All Rights Reserved

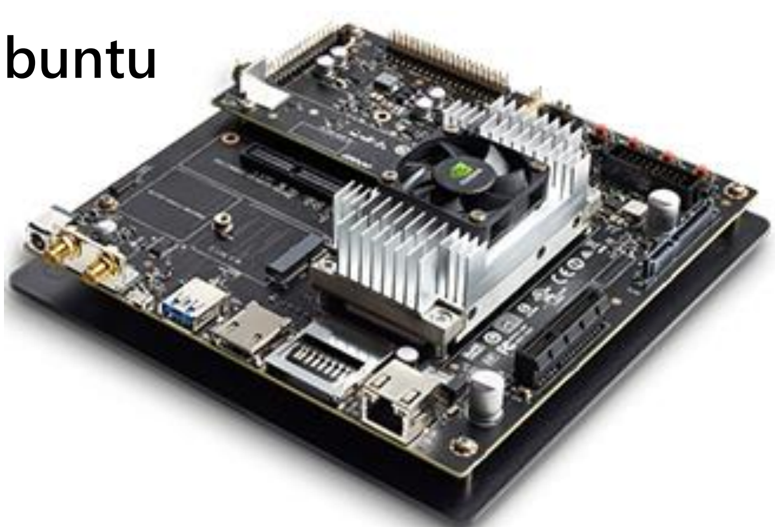
NVIDIA Jetson Nano 嵌入式平台 (472 GFLOPs)

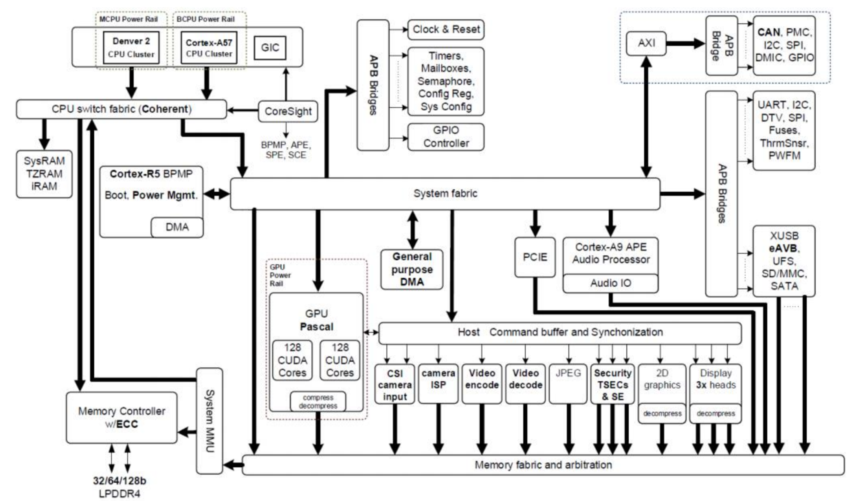
GPU	NVIDIA Maxwell 配備 128 個 核心
CPU	Quad-core ARM A57 @ 1.43 GHz
記憶體	4 GB 64-bit LPDDR4 25.6 GB/s
儲存空間	microSD (不包含)
影片編碼	4K @ 30 | 4x 1080p @ 30 | 9x 720p @ 30 (H.264/H.265)
影片解碼	4K @ 60 | 2x 4K @ 30 | 8x 1080p @ 30 | 18x 720p @ 30|(H.264/H.265)
相機	2x MIPI CSI-2 DPHY lanes
連線能力	Gigabit 乙太網路, M.2 Key E
顯示器	HDMI 和 DP
USB	4x USB 3.0, USB 2.0 Micro-B
其他接頭	GPIO, I2C, I2S, SPI, UART
機械規格	100 mm x 80 mm x 29 mm

圖 Jetson Nano

NVIDIA Jetson TX2 嵌入式平台

- Tegra X2 SOC
  - NVIDIA Pascal GPU，含 256 個 CUDA 核心
  - HMP Dual Denver + Quad ARM® Cortex-A57 CPU
- 8 GB 記憶體 搭配 128 位元寬度
- 32 GB eMMC
- OS
  - Ubuntu



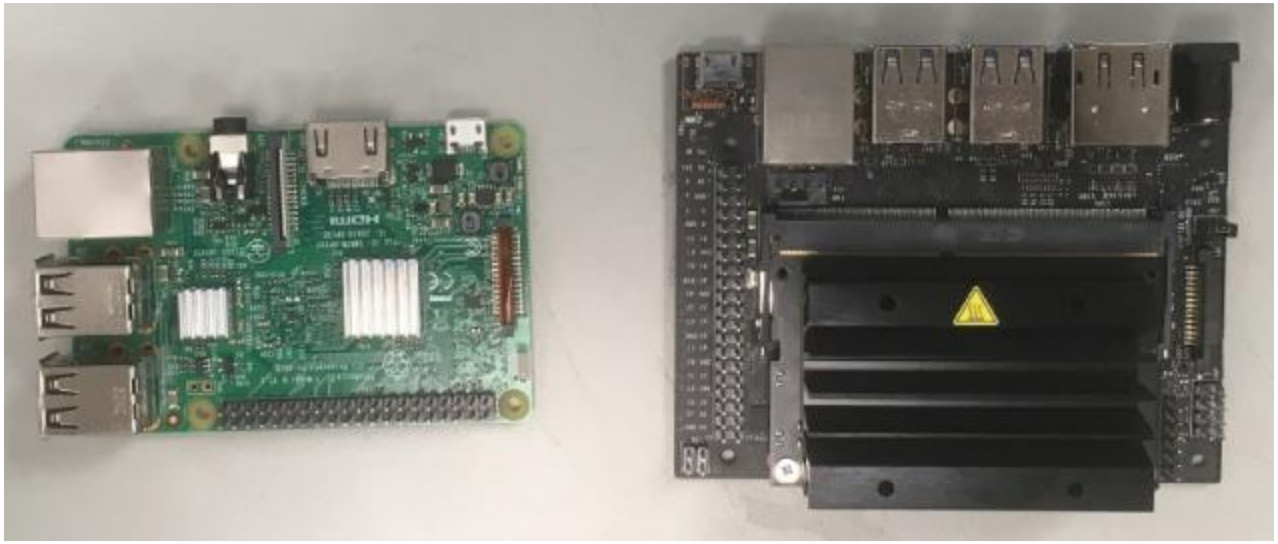


NVIDIA Jetson TX2

- 雙核心 Denver 2 64-bit CPU + 四核心 ARM® A57 Complex
- 8 GB L128 bit DDR4 記憶體
- 32 GB eMMC 5.1 Flash 儲存
- 可連接支持802.11ac WLAN 和 藍芽的裝置
- 10/100/1000BASE-T 乙太網路
- USB 3.0 Type A
- USB 2.0 Micro AB (支持recovery 與 host 模式)
- HDMI
- M.2 Key E
- PCI-E x4
- Gigabit Ethernet
- Full-Size SD
- SATA Data and Power
- GPIOs, I2C, I2S, SPI, CAN*
- TTL UART with Flow Control
- Display Expansion Header
- Camera Expansion Header

	TK1	TX1	TX2	AGX Xavier	Nano
GPU cores	192 Core	256 Core	256 Core	512 Core	128 Core
CPU	4 core	4 core	6 core	8 core	4 core
Memory	2GB DDR3	4GB DDR4	4GB DDR4 8GB DDR4 二種版本	16GB DDR4	4GB DDR4
Storage	16GB eMMC	16GB eMMC	32GB eMMC	32GB eMMC	SD(開發板) 16GB eMMC (二種版本)
Video Encode	1080p@30	4K@30	(3x) 4K@30	(4x) 4Kp60 / (8x) 4Kp30	4K@30
Video Decode	1080p@60	4K@60	(2x) 4K@60	(2x) 8Kp30 / (6x) 4Kp60	4K@60
WI-FI	O	O	O	O	X
Bluetooth	X	O	O	O	X
Power	5W	10W / 15W	7.5W / 15W	10W / 15W / 30W	10W
USB	USB 3.0	USB 3.0 + USB 2.0	USB 3.0 + USB 2.0	USB 3.0 + USB 2.0	USB 3.0 + USB 2.0

Jetson Nano VS Raspberry Pi



Source: https://chtseng.wordpress.com/2019/05/01/nvida-jetson-nano-初體驗：安裝與測試

	Jetson Nano	Raspberry Pi
CPU	64-bit Quad-core ARM A57 (1.43 GHz)	1.4 GHz 64-bit quad-core ARM Cortex-A53
GPU	128-Core Nvidia Maxwell	Broadcom VideoCore IV
RAM	4GB DDR4	1GB DDR2
WIFI	X	802.11ac
Bluetooth	X	LE 4.2
Ethernet	Gigbit	Gigbit (300Mbps max)
GPIO	40 pin	40 pin
USB	USB 2.0 x 3 USB 3.0 x 1	USB 2.0 x 4
Audio	X	Audio jack
Power	5~10W	400 mA (2.0W)
Price	$99	$35
算力	472 Gflops	24 Gflop

WHICH HARDWARE PLATFORM TO CHOOSE?

DBW Requirement for Computing CNNs

- **Because of the limited on-chip cache size, data such as IFMs, weights, and OFMs are necessary to be moved between DRAM and SRAM. This forms the DRAM bandwidth requirement.**
- **Taking Agilev3 for example, the data transferred between DRAM and SRAM for 30 fps of 416*416 input image resolution may be as high as 3.02 GB/s based on that 72kB kernel SRAM and 169kB IFM SRAM are equipped on-chip.**

Total Available DBW

- The total bandwidth is the product of
  - **Base DRAM clock frequency**
  - **Number of data transfers per clock**: Two, in the case of double data rate (DDR, DDR2, DDR3, DDR4) memory.
  - **Memory bus (interface) width**: Each DDR, DDR2, or DDR3 memory interface is 64 bits wide. Those 64 bits are sometimes referred to as a line.
  - **Number of interfaces**: Modern personal computers typically usetwo memory interfaces (dual-channel mode) for an effective 128-bit bus width.
- For example, a computer with dual-channel memory and one DDR2-800 module per channel running at 400 MHz would have a theoretical maximum memory bandwidth of:

400,000,000 clocks per second $\times$ 2 lines per clock $\times$ 64 bits per line $\times$ 2 interfaces = 102,400,000,000 (102.4 billion) bits per second (in bytes, 12,800 MB/s or 12.8 GB/s)

Suggested DRAM Type

Width	Height	fps	DBW needed (GB/s) Original/optimized	Suggested DRAM type
416	416	30	3.02/1.93*	DDR-400, PC-3200
1080	720	30	13.57/8.67*	DDR4-2400, PC4-19200/DR3-1600, PC3-12800
1920	1080	30	36.19/23.13*	No available DRAM/DDR4-3200, PC4-25600

* DBW of integer AgileNet, a light-weight CNN for mobile use.

Names	Memory clock	I/O bus clock	Transfer rate	Theoretical bandwidth
DDR-200, PC-1600	100 MHz	100 MHz	200 MT/s	1.6 GB/s
DDR-400, PC-3200	200 MHz	200 MHz	400 MT/s	3.2 GB/s
DDR2-800, PC2-6400	200 MHz	400 MHz	800 MT/s	6.4 GB/s
DDR3-1600, PC3-12800	200 MHz	800 MHz	1600 MT/s	12.8 GB/s
DDR4-2400, PC4-19200	300 MHz	1200 MHz	2400 MT/s	19.2 GB/s
DDR4-3200, PC4-25600	400 MHz	1600 MHz	3200 MT/s	25.6 GB/s

Mobile AI Platforms

Platform	CPU	GPU	Performance	DBW (GB/s)
Jetson Nano	4 cortex A57	128 CUDA cores	472 GOPs	25.6
Jetson TX2	2 Denver cores and 4 cortex A57	256 pascal gpu cores	1.33 TOPs	59.7
Jetson AGX Xavier	8 Carmel cores and ARM 8.2 64b CPU	512 volta gpu cores with 64 tensor cores	32 TOPs	136.5

Suggested Platform

Width	Height	fps	Operation required (GOPS)	Network	Suggested platform
416	416	30	981	AgileV3	Jetson TX2 or Jetson Nano for 14 fps
1080	720	30	4405		Jetson TX2
1920	1080	30	11752		Jetson AGX Xavier

Width	Height	fps	Operation required (GOPS)	Network	Suggested platform
416	416	30	1560	YOLOv3	Jetson TX2/Jetson AGX Xavier
416	416	30	1803	YOLOv4	Jetson TX2/Jetson AGX Xavier

Comparisons of Detection NNs

Network	Word length	No. of conv. layers	Model size (MB)	Conv. IO* (Mega)	Required GOPS*	Year
Agilev3	FP32	43	65.39	2023.8	480	2019
YOLOv3	FP32	74	241.78	3352.5	980	2018
YOLOv4	FP32	109	251.15	4795.8	900	2020
YOLOv4-tiny	FP32	21	23.10	707.1	100	2020
HarDNet+SSD	FP32	88	98.04	2361.3	770	2019
YOLOv5 - s	FP16	70	14.2	1075.8	110	2020

* for 416x416 @ 30 fps