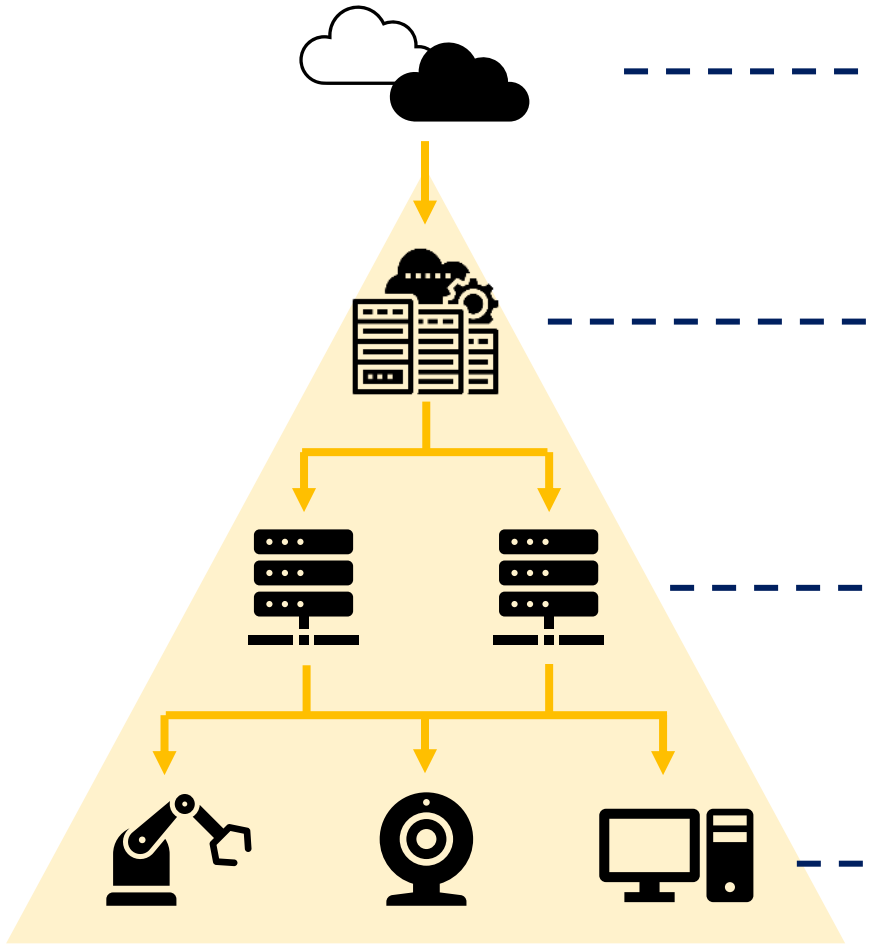
Edge AI浪潮下的硬體業者發展觀察

申作昊 Zouhao Shen

DIGITIMES研究中心

April.2024

「邊緣」運算定義：運算節點較靠近終端



雲端層

邊緣運算

地區資料中心

小型/微型資料中心

邊緣伺服器

Server、IPC...etc

感測/終端

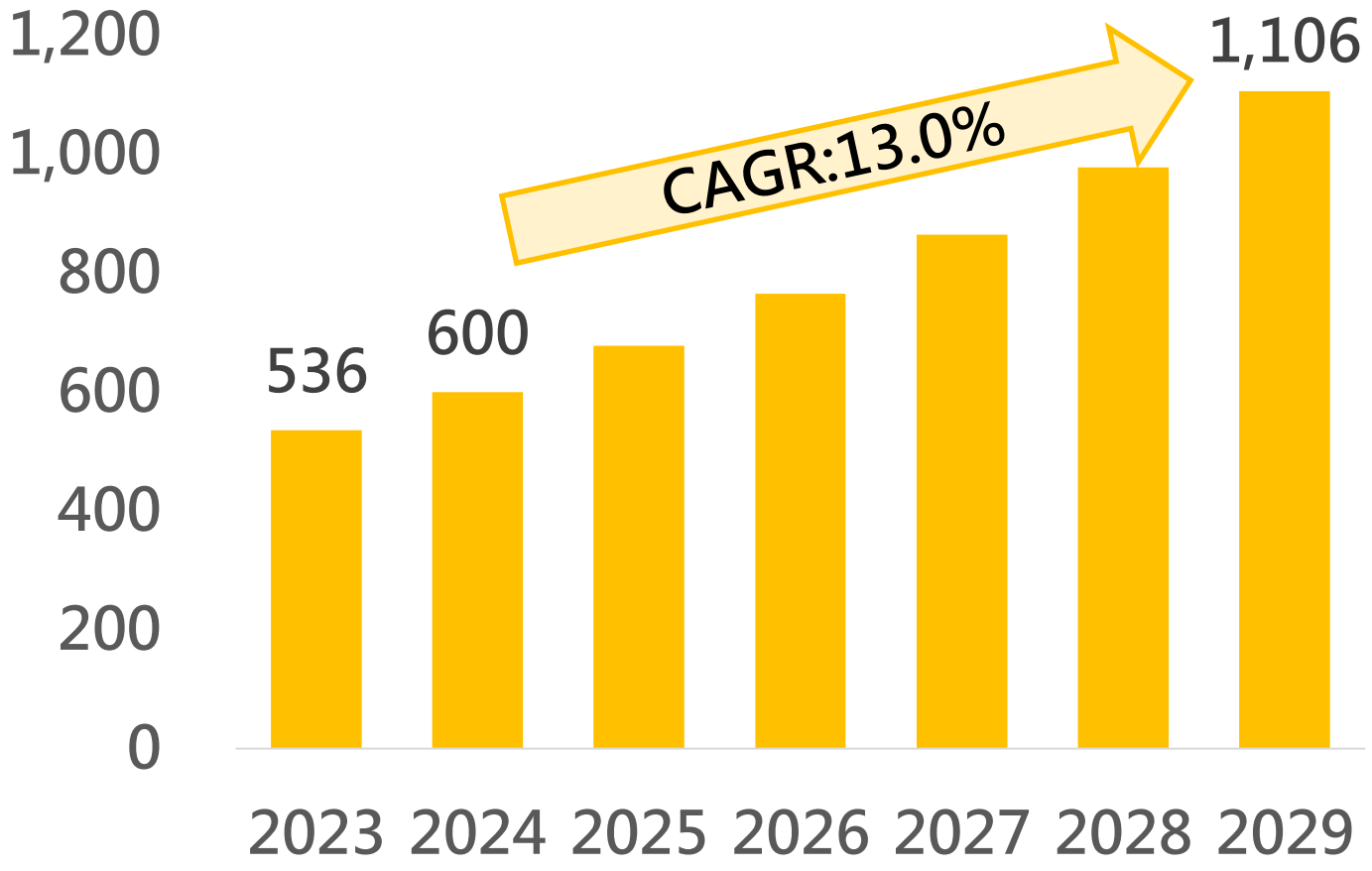
AI手機、AI 筆電、
智慧感測器...etc

	運算資源	儲存空間	延遲程度
邊緣運算	較多	較大	較高
雲端層	較少	較小	較低

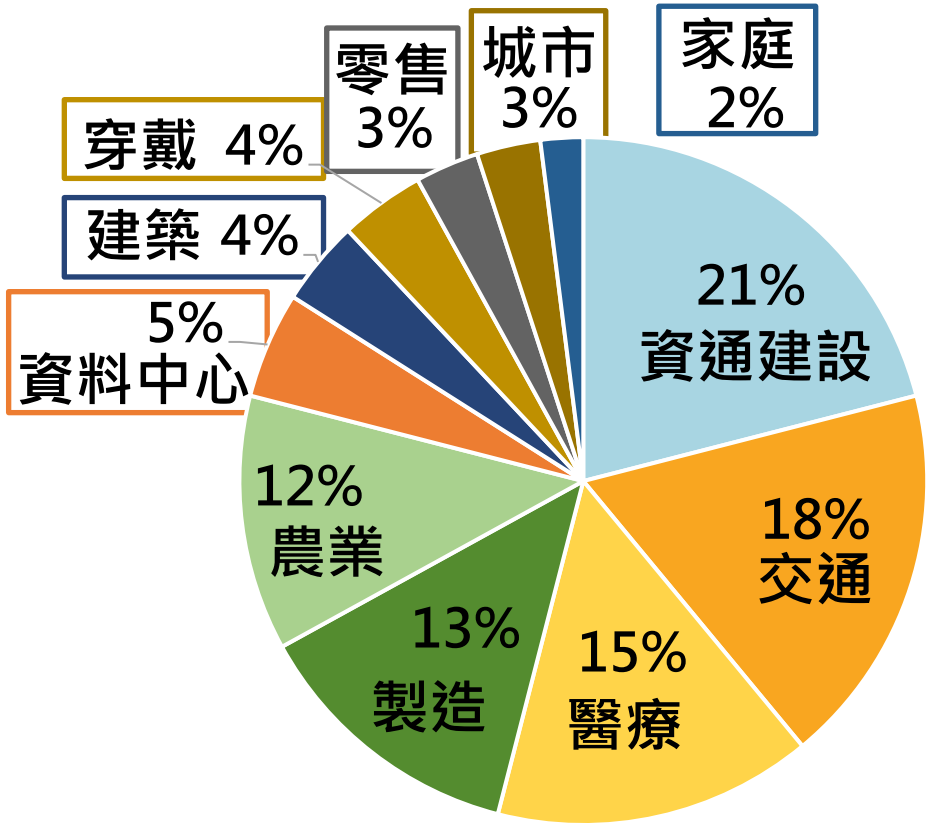
全球運算市場穩定成長　年增長率估約15.7%

單位：億美元

全球邊緣運算市場成長預測



邊緣運算各場域應用分布



資料來源：MarketsandMarkets、IIC，DIGITIMES Research整理，2024/4

邊緣市場構成與重要參與者列舉

Edge AI 軟體

整合式開發平台

EDGE IMPULSE
Deeplite
Plumerai
NotaAI
CLIKA.
alwaysAI
LatentAI
EDGENEURAL.ai

預訓練模型

deci. Break the AI Barrier
LatentAI
Deeplite
imagimob
Micro.ai
EDGENEURAL.ai
SensiML
RealityAI

Edge AI晶片

通用晶片-GPGPU

NVIDIA
AMD
intel

MCU

Infineon
RENESAS
life.augmented
nuvoTon

NPU(XPU)

Ambarella
Kneron
AXELERA
HAILO
brainchip
DEEPX
CEVA
groq

邊緣設備/邊緣終端

邊緣伺服器

SUPERMICRO
DELL
Quanta Computer
wistron
FOXCONN 鴻海科技集團
英業達集團 Inventec
MITAC
Hewlett Packard Enterprise
Lenovo

消費型終端

SAMSUNG
MEDIATEK
vivo
oppo
acer
ASUS
DELL
Lenovo

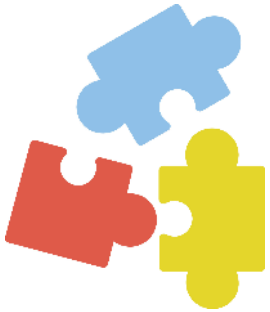
邊緣IPC

ENNOCONN Steering the Industrial Metaverse
ADVANTECH
SIEMENS
BECKHOFF
DFI
ADLINK

智慧終端

智慧鏡頭
感測器
智慧電錶
...etc

邊緣端產業生態兩大發展重點

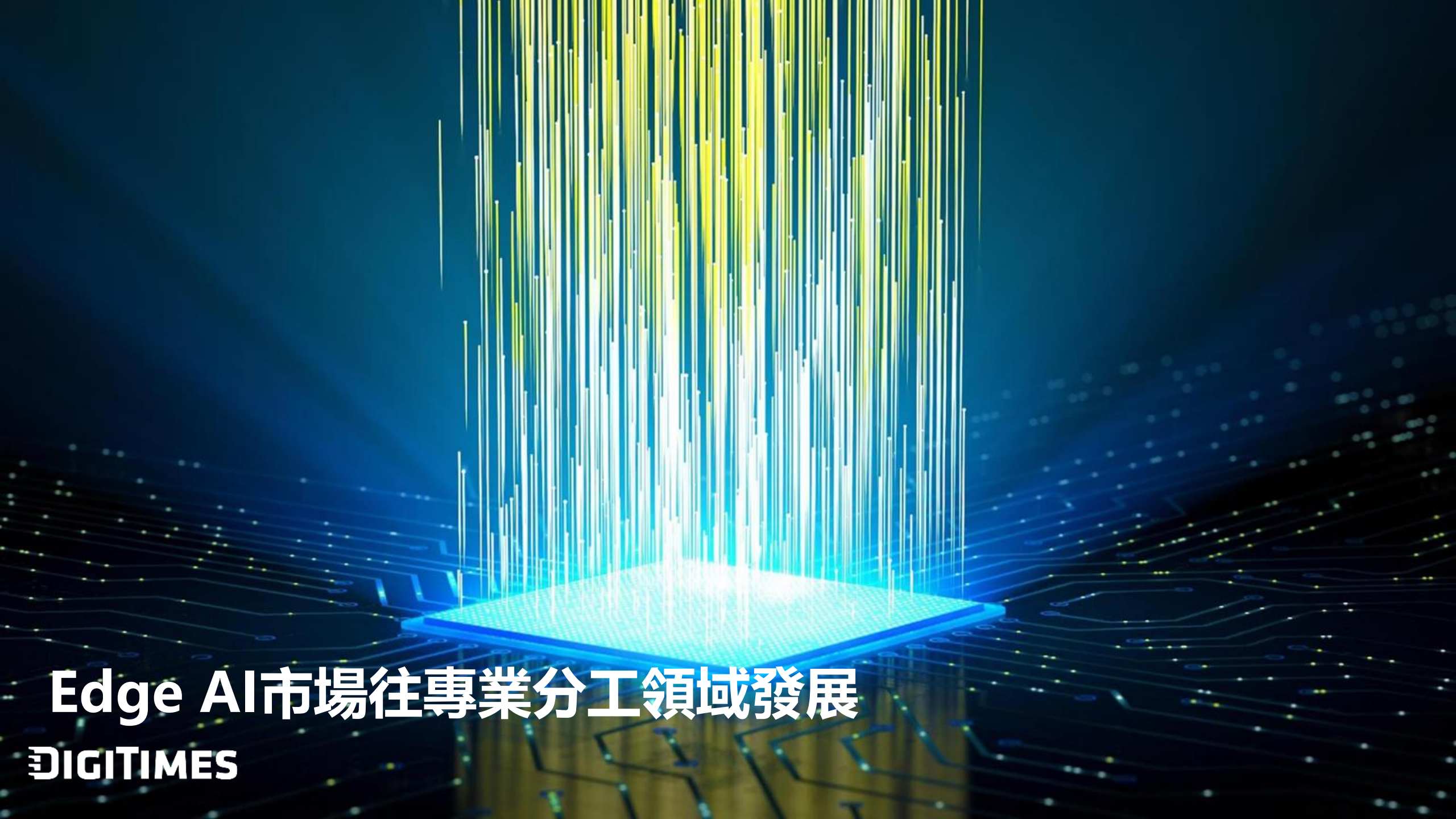


市場需求

應用專業化

供應商策略

強調軟硬整合


Edge AI市場往專業分工領域發展
DIGITIMES

邊緣端市場趨向應用專業化

車載電腦

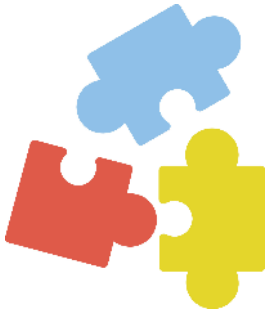
工業應用

安防監控

城市交通

智慧家居

...etc



市場需求

供應商策略

應用專業化

強調軟硬整合

邊緣架構依情境處理不同類型AI任務

雲端 | 邊緣層：資料中心、伺服器 | 邊緣層：物聯網感測器

任務複雜度高

- 各式推理任務 e.g.文字生成、圖像生成...etc
- 複雜種類模型訓練任務

地端LLM應用

- 各領域專家系統
- 專家模型訓練

機器視覺類應用

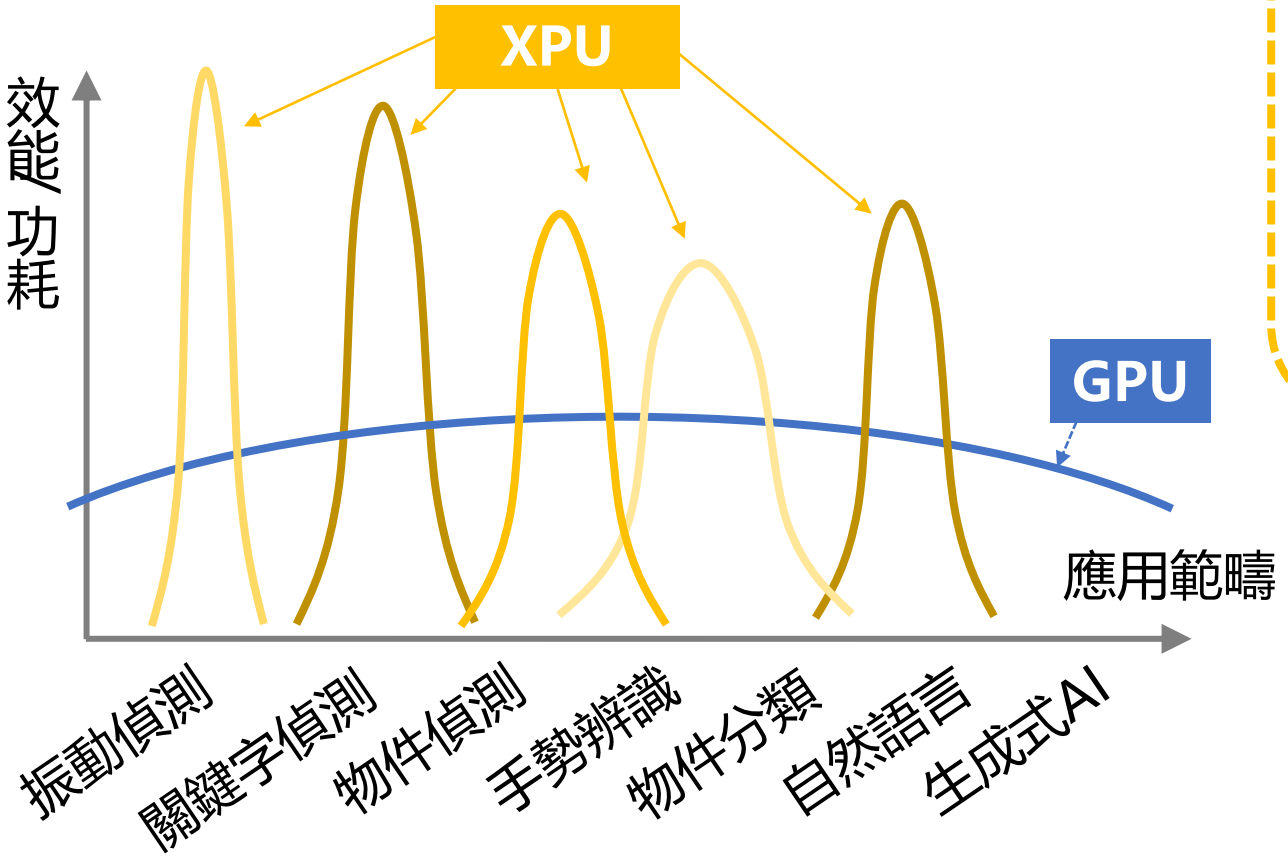
- 物件分類
- 手勢識別
- 身分識別

非影像感測應用

- 震動偵測
- 語音喚醒
- 異常偵測

大模型、通用類晶片 → 輕量模型、專用晶片

邊緣端AI晶片類別(GPU、XPU)與差異



圖片來源：Hailo

NPU(XPU) 專用度較高

Ambarella　Kneron
AXELERA　HAILO
brainchip　DEEPX
CEVA　groq

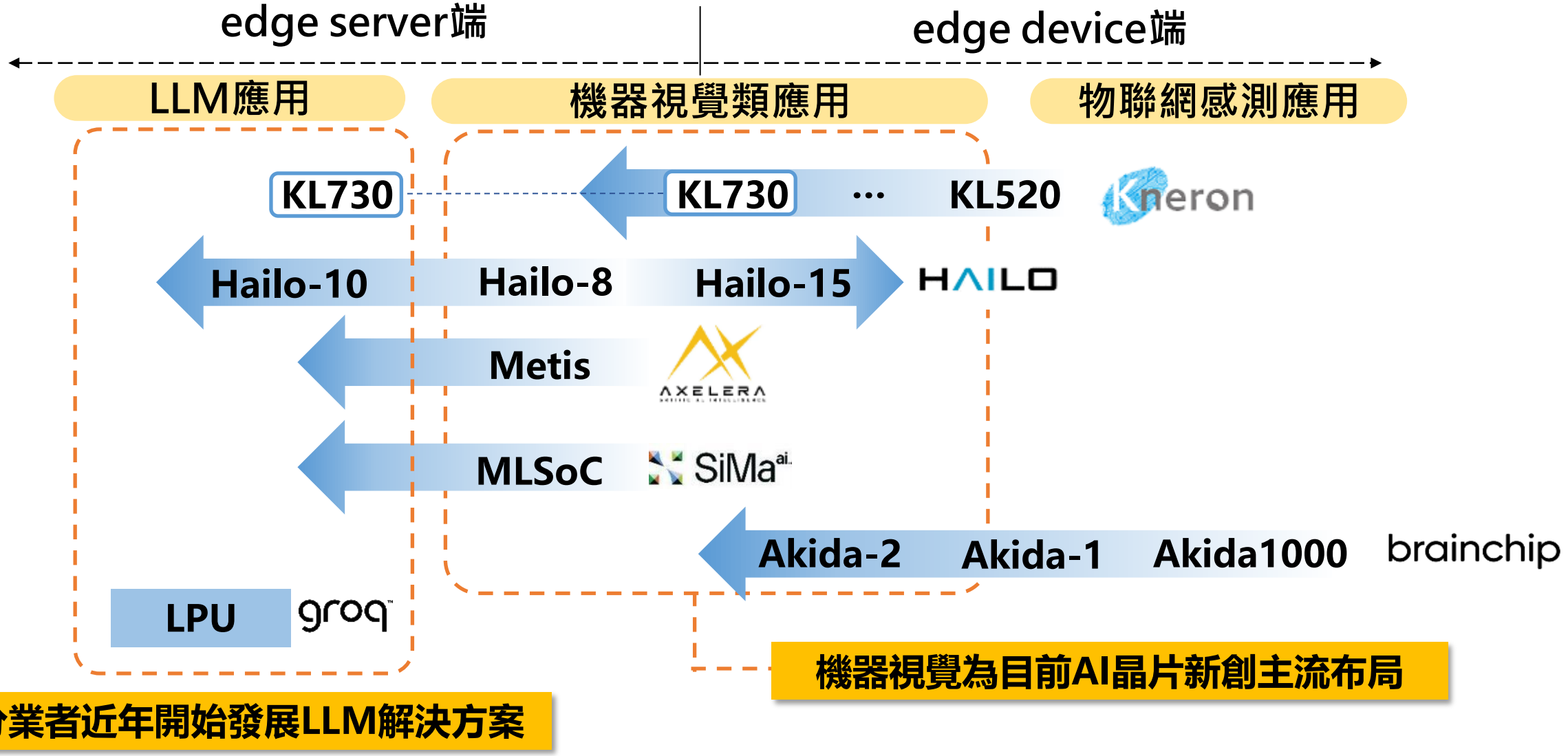
- ✓ 多針對特定AI任務開發
- ✓ 特定情境下效能、功耗等較佳
- ✓ 須一併考量AI演算法開發
- ✓ 較適合邊緣端單一應用情境

GPU 通用度較高

NVIDIA.
AMD
intel

- ✓ 可處理多種、複雜AI任務
- ✓ 較適合用於AI模型訓練
- ✓ 軟體開發生態較開放、完善
- ✓ 較適合雲端多樣任務情境

機器視覺為當前AI晶片業者主流應用方向



邊緣運算各環節AI加速硬體需求不同　解決方案各異

雲端/資料中心

- AI任務種類多樣
- 可擴充性要求高
- 訓練模型需求高

GPU　通用性高 處理各種任務

ASIC　雲端業者自研 配合自家算法

Edge server

- 可擴充性要求高
- 需視應用調整規格

GPU　通用解決方案

NPU　專用解決方案 保留可擴充性

SoC　整合運算單元 具性能優勢

Edge device

- 應用單一、專用度高
- 成本、能耗敏感

MCU + NPU　外接運算能力 保留設計彈性

SoC　整合運算單元 具性能優勢

MCU — NPU　微控制器 內嵌運算能力

NPU業者多布局機器視覺技術　應用領域各不相同

機器視覺晶片各應用領域產品

業者逐漸從外掛NPU轉向發展整合度較高產品

資料中心/伺服器

- 運算力需求高
- 大量記憶體空間
- 可擴充性高

SoC

SiMa.ai

外掛NPU

HAILO

車載系統

- 運算力需求高
- 高可靠性
- 延遲程度低

SoC

HAILO

Ambarella

智慧鏡頭

- 延遲程度低
- 低耗能要求
- 成本敏感度高

SoC

Kneron

Ambarella

物聯網終端

- 低耗能要求
- 成本敏感度高
- 保留設計彈性

外掛NPU

brainchip

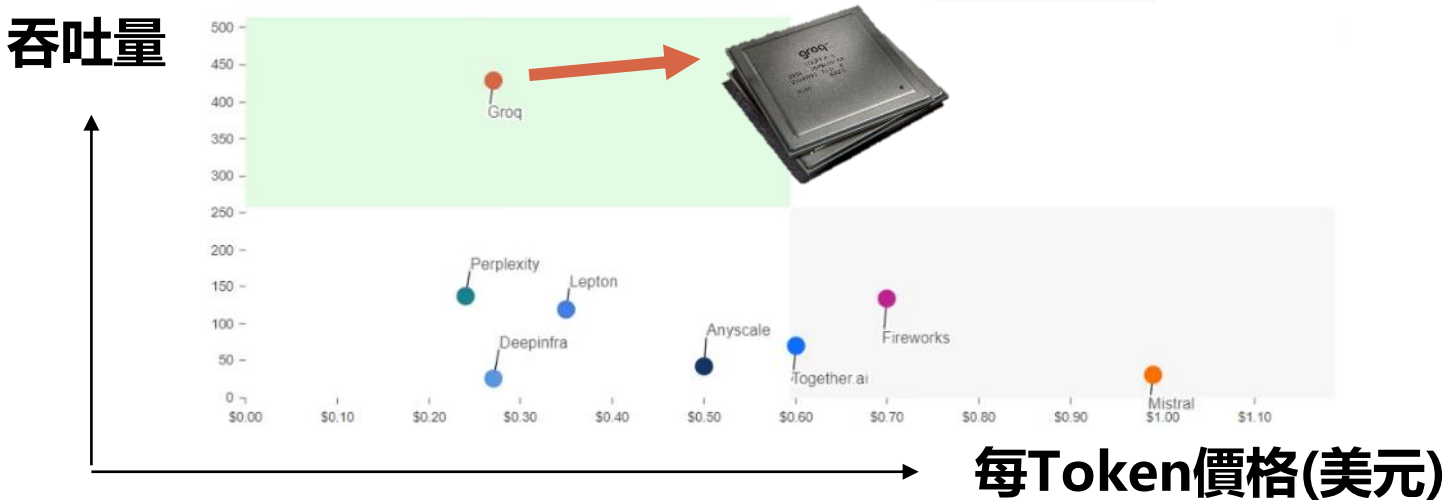
NPU+MCU

SYNTIANT

部分NPU新創業者近期推出LLM專用AI加速晶片

LLM專用晶片

groq™ **LPU**



推廣Groq cloud服務

- 購併 AI 解決方案公司 Definitive Intelligence
- 以高吞吐量、低收費為優勢

推廣Groq 晶片至資料中心

- 成立Groq Systems推銷晶片
- Edge LLM尚無特殊用例
- 建置成本為一大劣勢

HAILO **Hailo-10**

- 以M.2加速卡形式推出，支援邊緣server或PC
- Hailo-10H提供40TOPS運算力
- 功耗低於 3.5W

AXELERA

SiMa.ai

- 表達近年將推出LLM專用開發軟體或硬體

Source: Artificial Analysis、Groq、Hailo， DIGITIMES Research整理，2024/4

處理器、MCU大廠新產品趨向內嵌NPU增加運算力

embeddedworld Exhibition&Conference

處理器、MCU大廠展出新品內嵌NPU

AMD Versal AI Edge Gen 2

- 整合NPU的系統單晶片，共6款產品
- 標榜低功耗與低延遲適於嵌入式應用
- 以車電、工業、醫療等為目標市場

AMD邊緣AI軟硬體解決方案

軟體	AMD Vitis AI 開發環境 (具模型壓縮優化工具)
硬體	 Versal AI Edge SoC 運算力20~200 TOPS

Infineon PSoC Edge MCU系列

- 以物聯網應用為目標市場
- E83與E84 MCU內建Ethos-U55 NPU
- 搭配ModusToolbox軟體開發工具
- 2025年量產，預估單價6~8美元

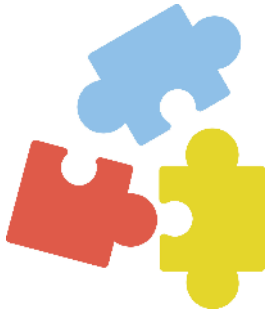
NXP FRDM開發板內嵌N947 MCU

- 具兩顆Cortex-M33核心
- 內建自研eIQ Neutron NPU
- 搭配MCUXpresso開發工具

Edge AI硬體業者走向軟硬整合

DIGITIMES

邊緣端產業生態兩大發展重點



市場需求

應用專業化

供應商策略

強調軟硬整合

產業應用軟體合作

軟體開發工具

客製化程度高
開發導入時間拉長

+

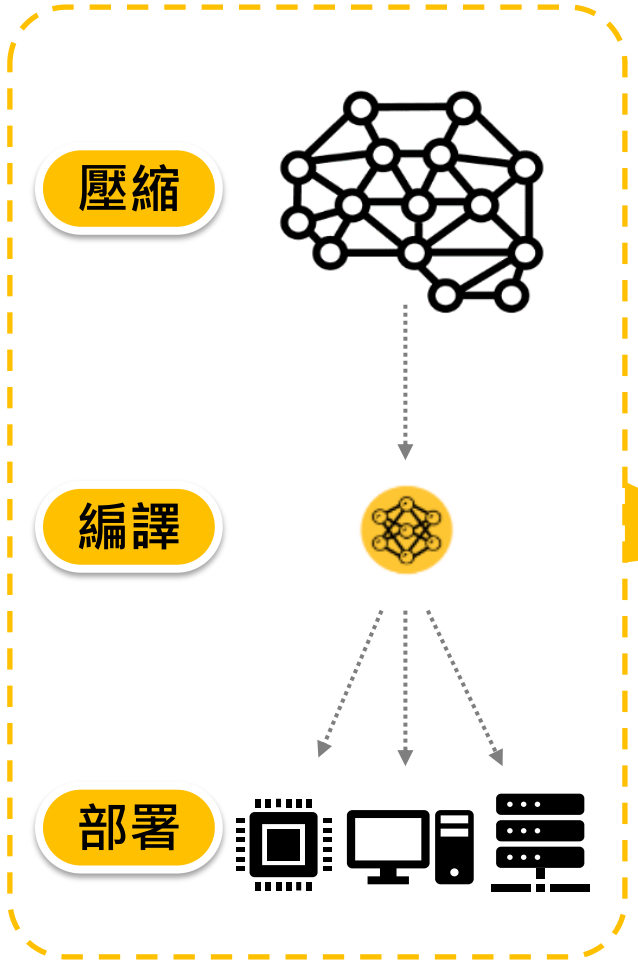
雲端通用AI模型
不適用邊緣硬體

客戶獨立開發軟體解決方案面臨挑戰

邊緣整合式開發業者將雲端開發工具串聯至邊緣端

Edge AI整合式開發平台業者

EDGE IMPULSE
Plumerai
Deeplite
CLIKA.
NotaAI
LatentAI
alwaysAI
deci. Break the AI Barrier
EDGENEURAL.ai



串聯雲端、邊緣開發生態

NVIDIA TAO 預訓練模型與開發工具

模型壓縮優化工具

infineon
life.augmented
NXP
RENESAS

串聯MCU業者開發工具

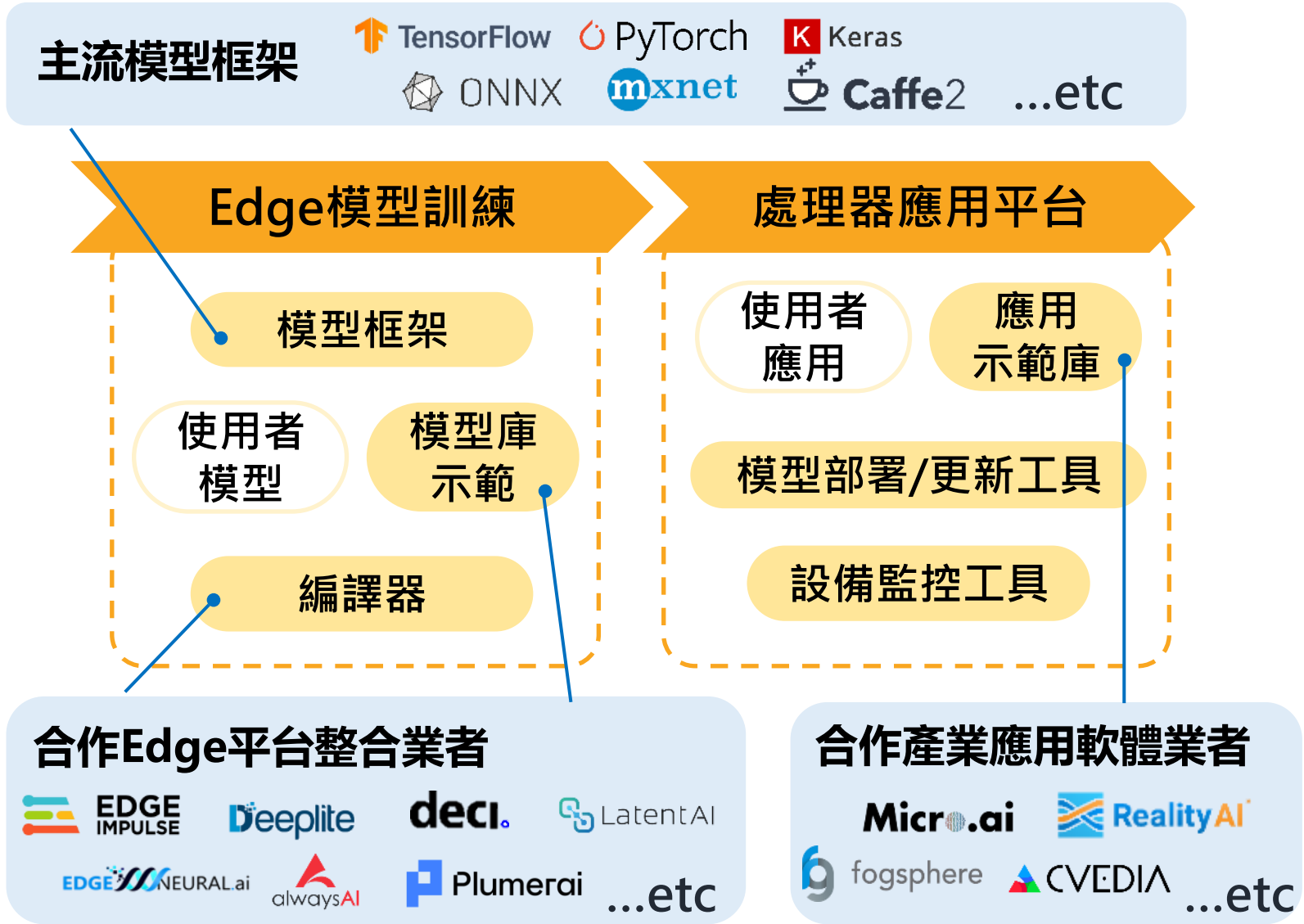
NVIDIA OMNIVERSE 生成數據模擬驗證

NVIDIA JETSON 導入邊緣硬體模組

AI晶片業者提供軟體開發套件支援客戶開發相容模型

AI晶片業者提供軟體開發套件

- NXP ▶ MCUXpresso
- Infineon ▶ ModusToolbox
- ST life.augmented ▶ ST Edge AI Suite
- SiMa.ai ▶ Palette
- HAILO ▶ AI Software Suite
- AXELERA AI ▶ VOYAGER SDK



AI晶片業者提供預訓練模型加速客戶開發

預訓練邊緣模型

外部合作取得資源

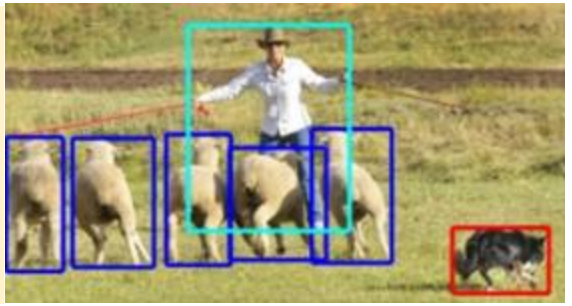
- deci (Break the AI Barrier)：物件偵測、物件分類、語意分割
- Deeplite：物件偵測、物件分類
- LatentAI：物件偵測、物件分類
- imagimob：聲音偵測
- EDGE NEURAL.ai™：物件偵測

內部研發

NXP、RENESAS、ST life.augmented、HAILO、AXELERA

晶片業者本身開發預訓練模型

物件偵測



物件分類

語意分割

AI晶片業者與場域軟體業者合作提供應用解決方案

製造場域應用

SECO 訂單管理系統

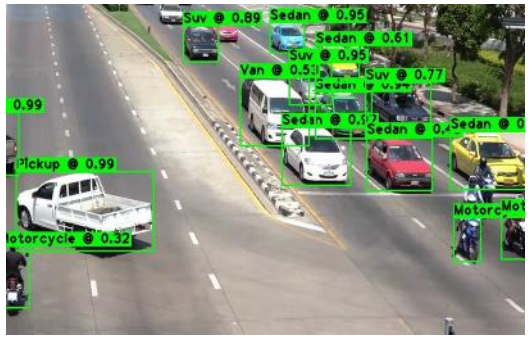
- 訂單管理系統
- 機器視覺：機器手臂抓取點、拾取物體識別
- 連線倉儲管理系統，達成自動化訂單管理

fogsphere 工安偵測系統

- 機器視覺：安全裝備配備狀況
- 行為安全、緊急通報、個人防護設備檢測

交通場域應用

CVEDIA 智慧交通系統

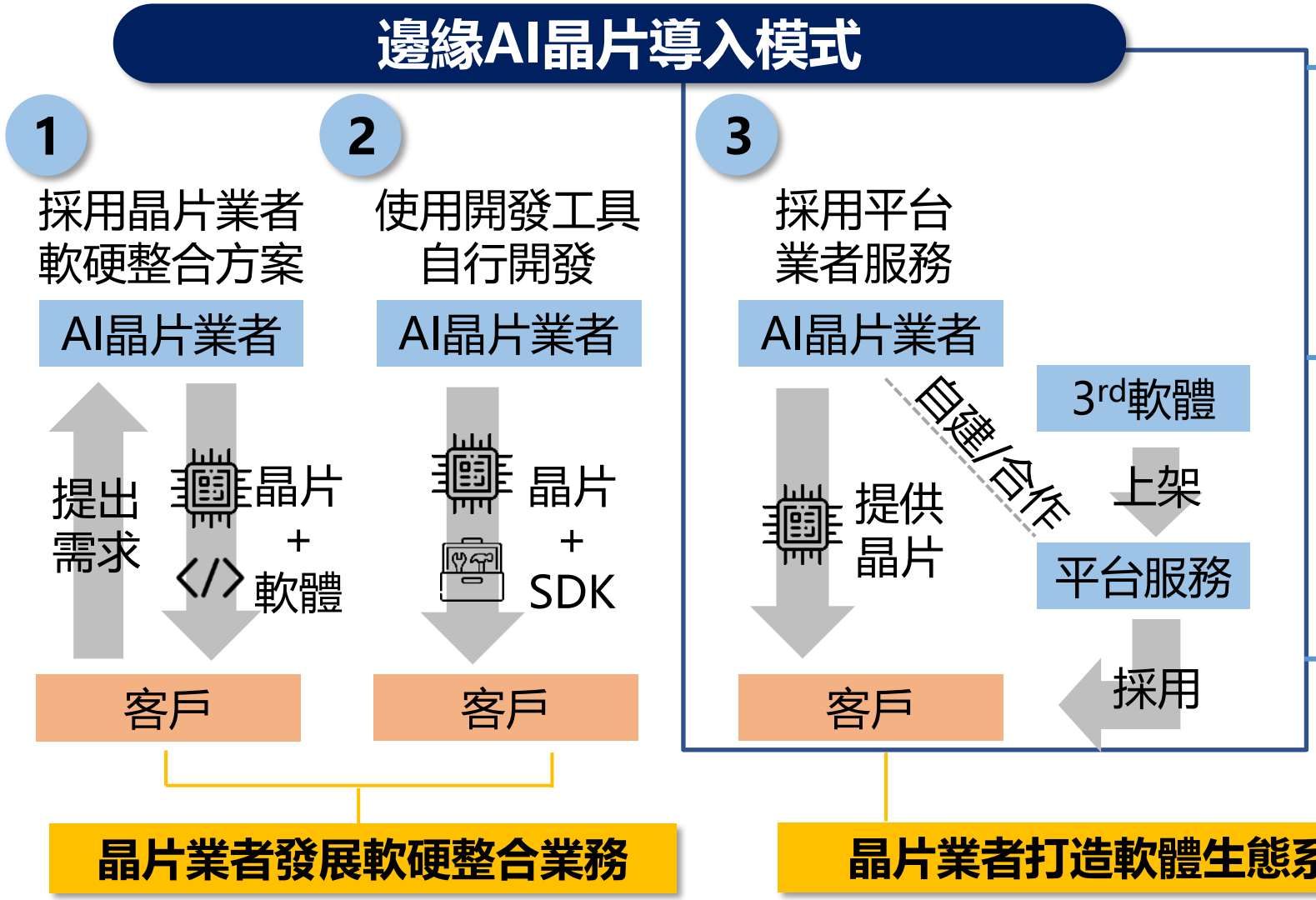


- 智慧交通系統
- 機器視覺：車輛識別、行人識別
- 車流計算、分類、人流分析等應用

Nota AI 路況偵測系統

- 機器視覺：車輛識別、行人識別
- 路口監控、事故偵測、智慧停車場等應用

AI晶片硬體業者逐步打造自家軟體生態體系

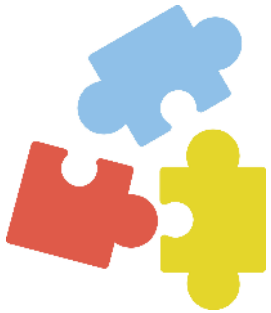


Nvidia TAO
工具套件平台

合作3rd業者
預模型平台

自建預模型平台
引入3rd業者軟體

結論



市場需求

應用專業化

穠纖合度 **適得其所**

- 硬體求取成本與效能平衡
- 通用性與專用性的取捨

供應商策略

軟硬整合方案

串聯生態系 **降低採用門檻**

- 不再固守硬體業務
- 發展微垂直領域解決方案

THANK YOU

申作昊 Zouhao Shen

DIGITIMES研究中心

April.2024