AI on Chip 軟硬整合

黃立仁

資訊與通訊研究所

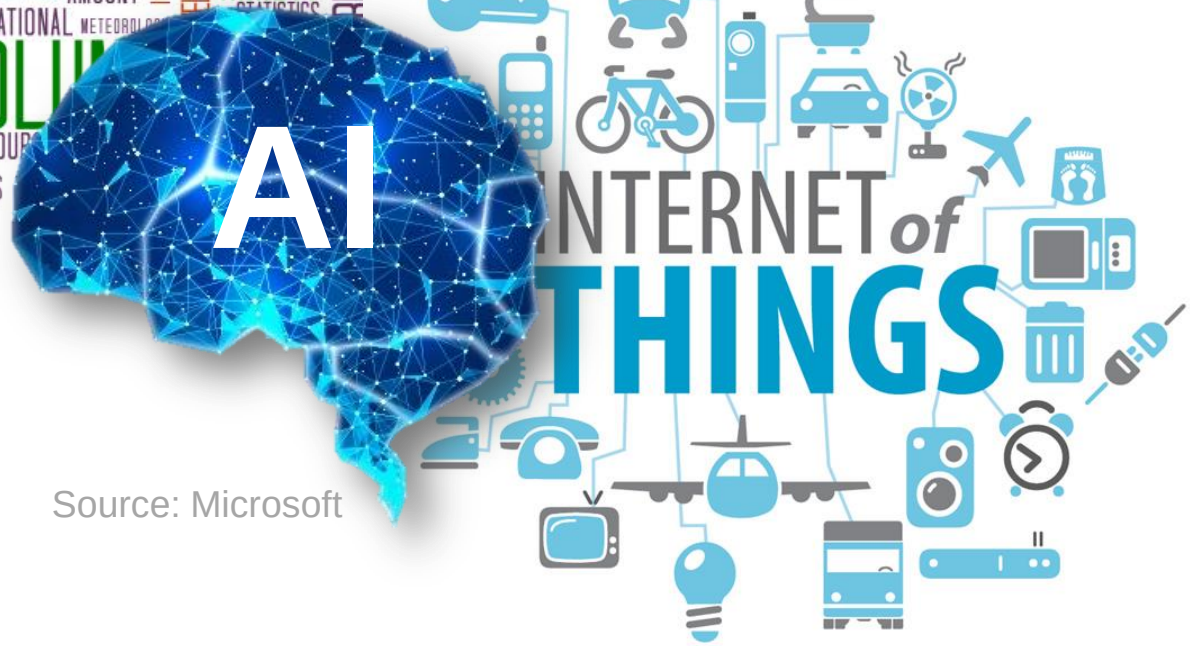
Information and Communication Research Lab (ICL)

ITRI
Industrial Technology
Research Institute

AI圓滿了大・人・物

遍在智慧

從雲到端的智慧裝置



見多識廣

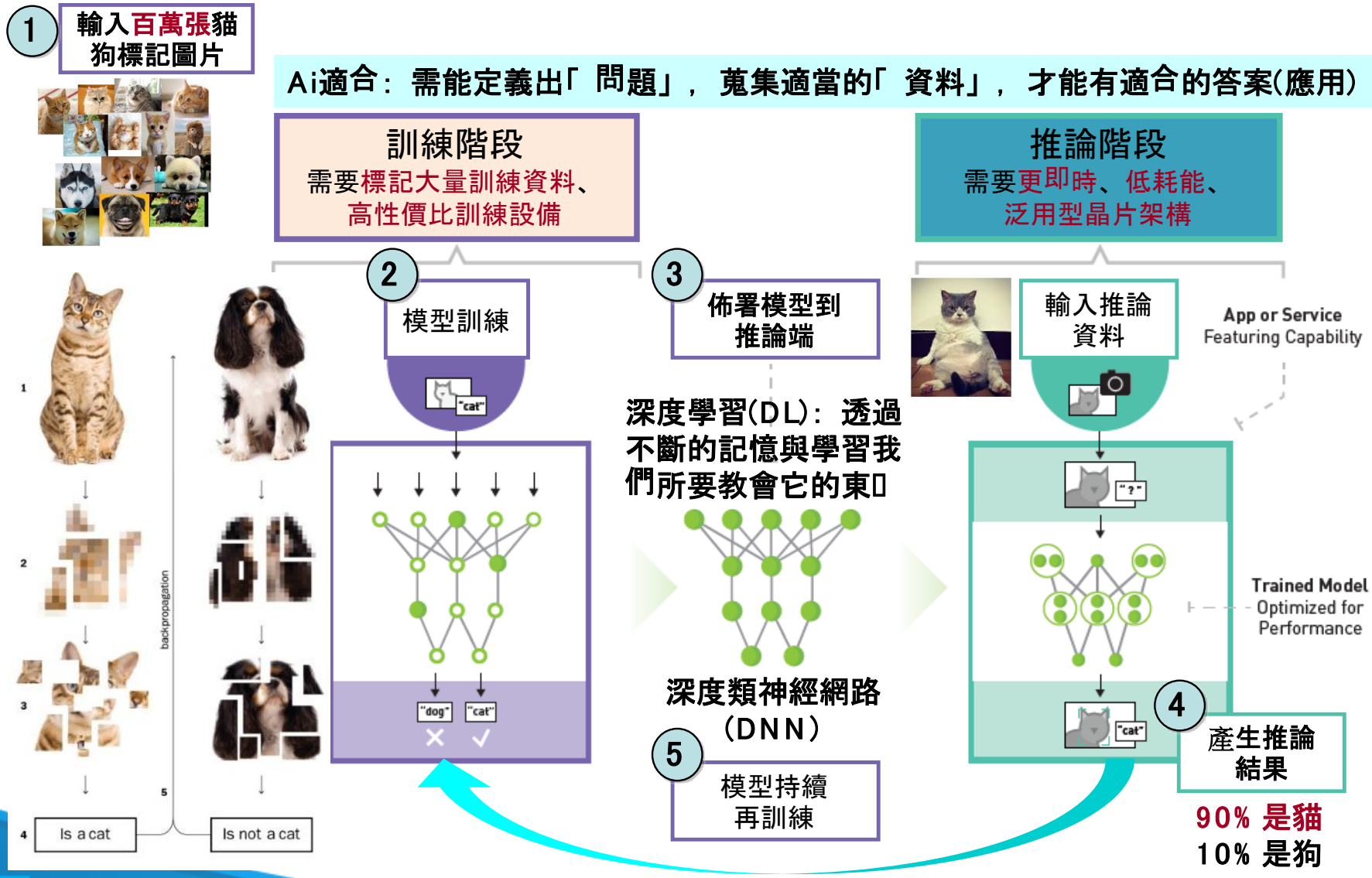
有效的AI來自大數據訓練

Source: Microsoft

遍在智慧: Ubiquitous Intelligence

Copyright 2018 ITRI 工業技術研究院

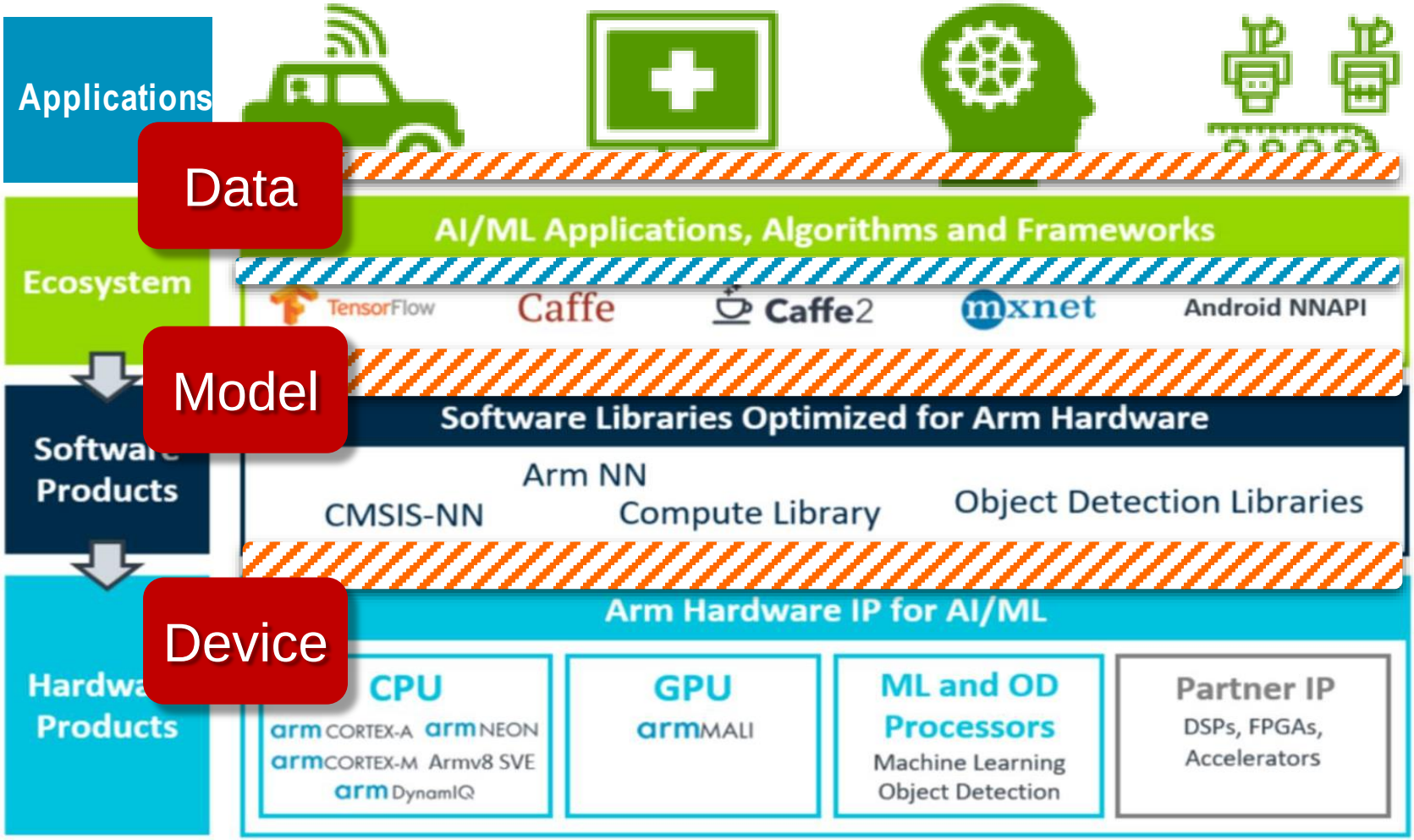
從數據到系統應用



圖片來源: nVidia；工研院資通所整理

Copyright 2018 ITRI 工業技術研究院

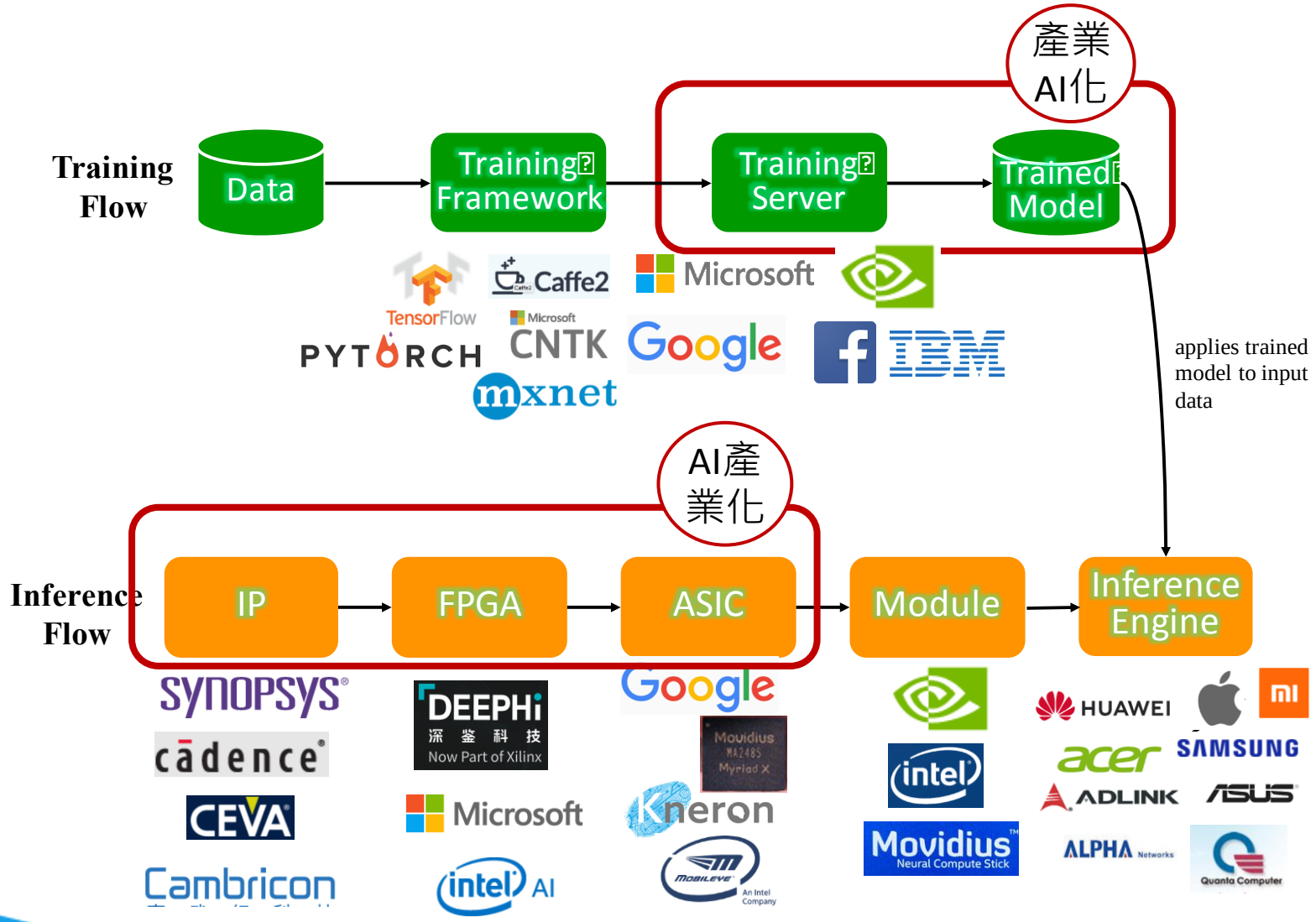
其實軟體比硬體重要



Source: arm

Copyright 2018 ITRI 工業技術研究院

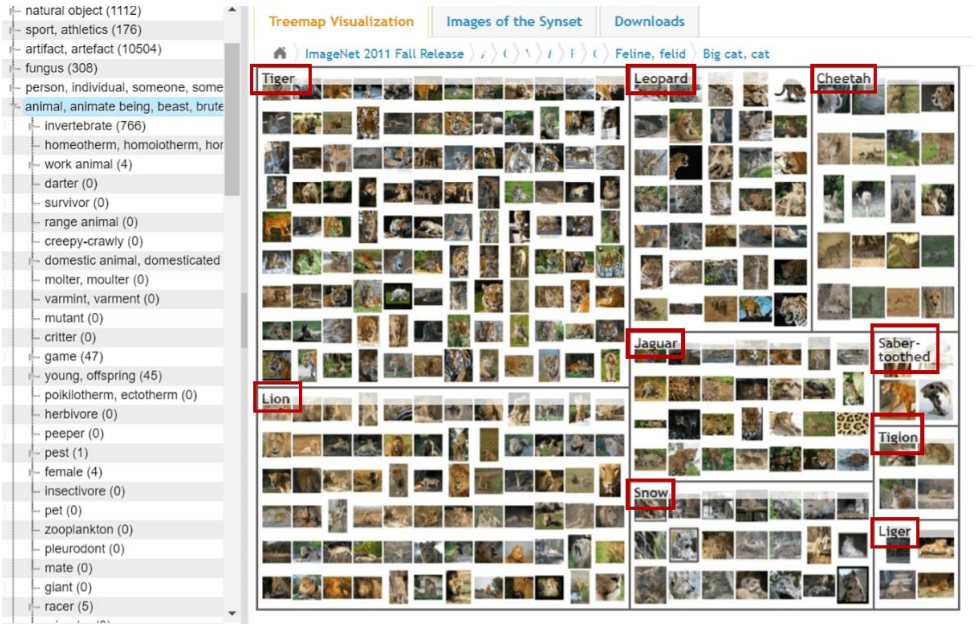
從國際生態鏈找出台灣的機會



Copyright 2018 ITRI 工業技術研究院

如何有效地訓練出AI模型

- 不同的資料會訓練出不同的結果
  - 資料收集、清理、標註
- AI模型訓練是黑箱作業
  - 訓練監督、分析、除錯
- 訓練得又快又好
  - 減少訓練會回數，以及每回合的訓練時間
- 複雜的大模型在裝置端落不了地
  - 模型壓縮



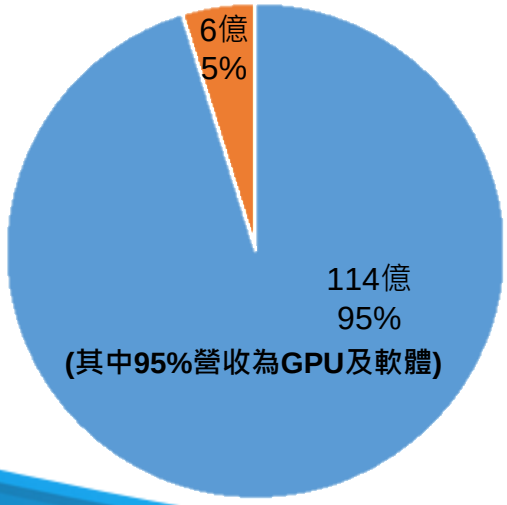
Source: imagenet

Copyright 2018 ITRI 工業技術研究院

從伺服器代工到AI訓練系統

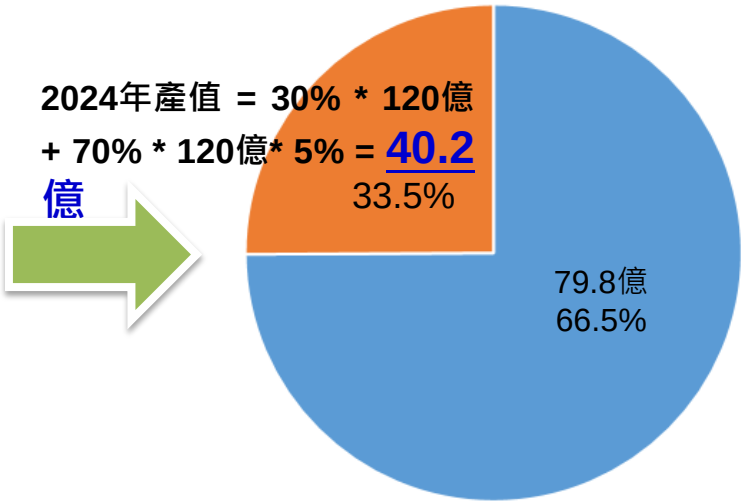
- Merrill Lynch 預估全球深度學習訓練系統市場將從目前6.5億美金，成長至2020年70億美金
- Gartner 預估企業或專網佔比60%，至2024年複合成長率130%，產值達120億美金。(企業專網深度學習訓練系統 = x86伺服器 + GPU + **AI軟體工具**)

只做x86伺服器代工
(5%, 6億美金)



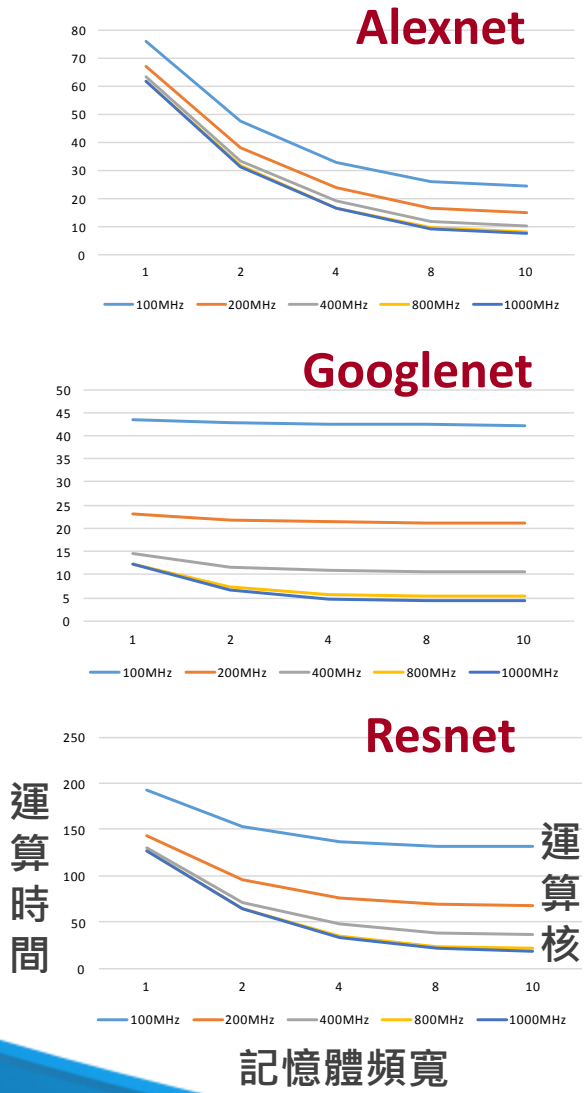
提供企業深度學習訓練系統
(33%, 40億美金)

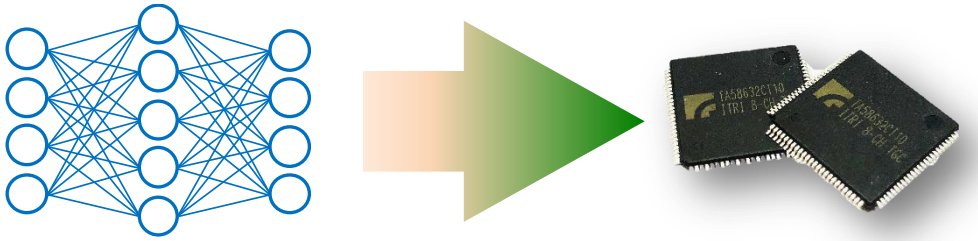
2024年產值 = 30% * 120億
+ 70% * 120億* 5% = 40.2億



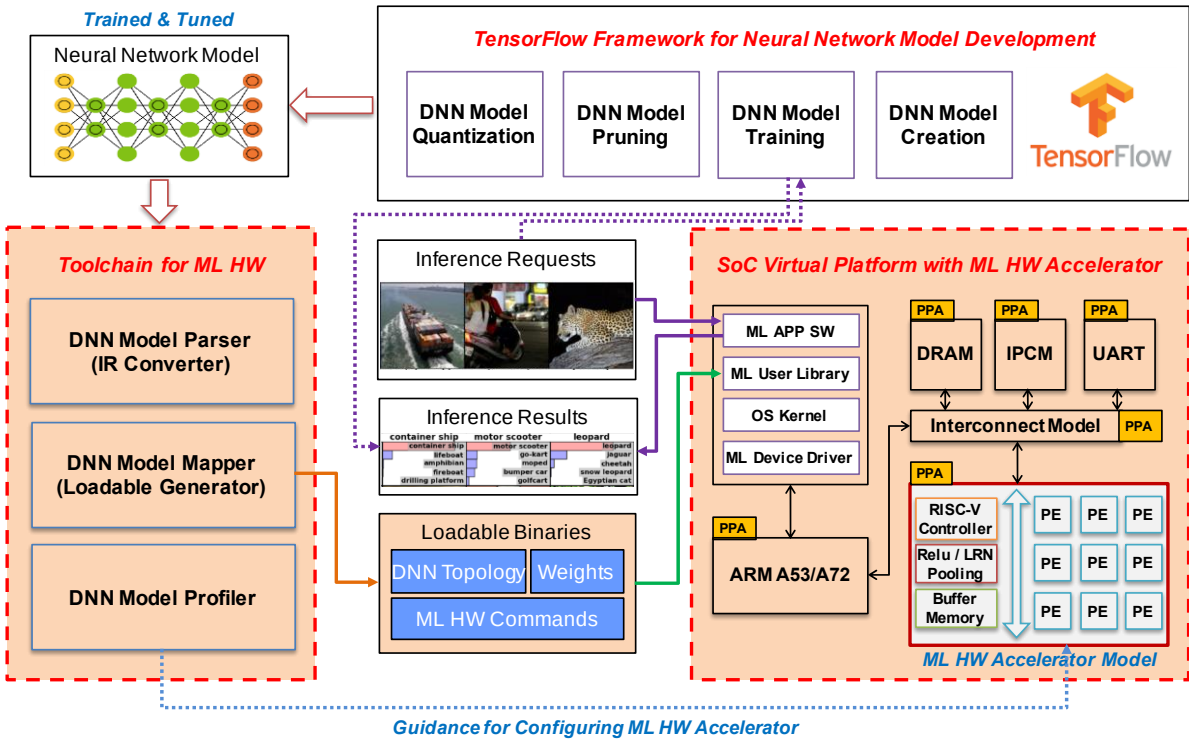
Copyright 2018 ITRI 工業技術研究院

AI晶片軟硬整合開發環境



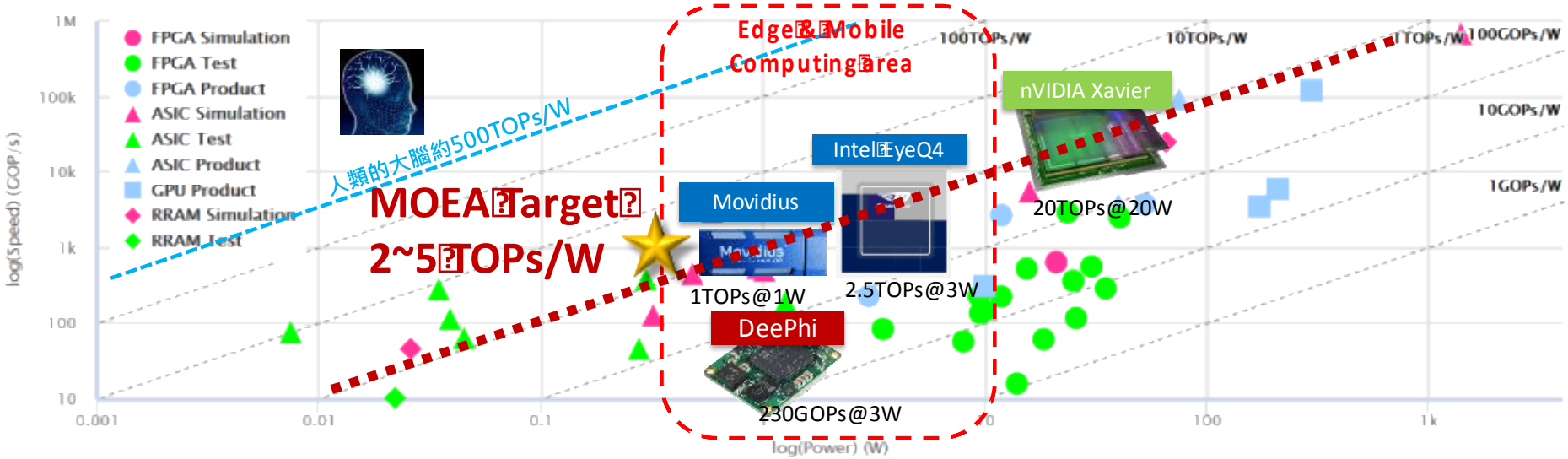


ITRI | Open Source

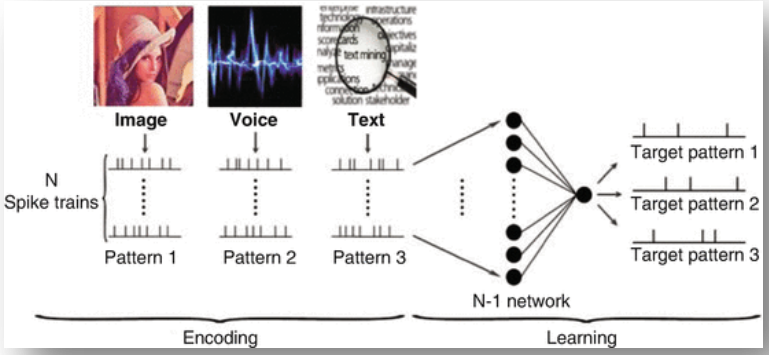


Copyright 2018 ITRI 工業技術研究院

基於台灣半導體強項的AI



- 記憶體內運算
  - 32位元->8位元->1位元，又是一個軟體的問題
- 仿生神經網路，一個未來的大戰場
  - 下世代非揮發記憶體
  - 類比式Spiking Neural Network



Copyright 2018 ITRI 工業技術研究院

Wrap Up

Domain Scenario, Big Data

物聯網嵌入式AI解決方案

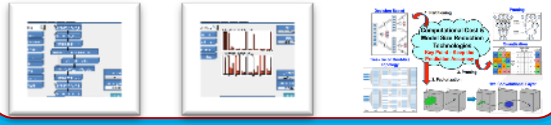
- 聚焦即時影像分析相關應用的邊緣運算
- 智慧監控、照護人機的互動

DNN Training Tool Kit

產業AI化的加速工具

- 讓AI訓練更簡單快速
- 降低產業導入AI的成本
- 讓應用的AI模型更輕量化

Compiler and Compressor



邊緣運算最佳化

- 支援共通DNN模型交換標準
- 善用運算資源、最佳運算排程，消除記憶體頻寬瓶頸

DNN inference Processor

嵌入式AI系統晶片

- AI on Chip核心矽智財
- 可依據應用合成不同AI晶片
- 可以FPGA或ASIC實現

Copyright 2018 ITRI 工業技術研究院

THANK YOU!

QUESTIONS AND COMMENTS?

Lrhuang@itri.org.tw

ITRI
Industrial Technology
Research Institute

Copyright 2015 ITRI 工業技術研究院