



CEVA TECHNOLOGY
SYMPOSIUM SERIES

拓璞TRI
TOPOLOGY RESEARCH INSTITUTE



Machine Learning Inference Chips for Edge Computing: Overview and Forecast

Yao Chia Yang, Senior Analyst
Topology Research Institute

www.ceva-dsp.com



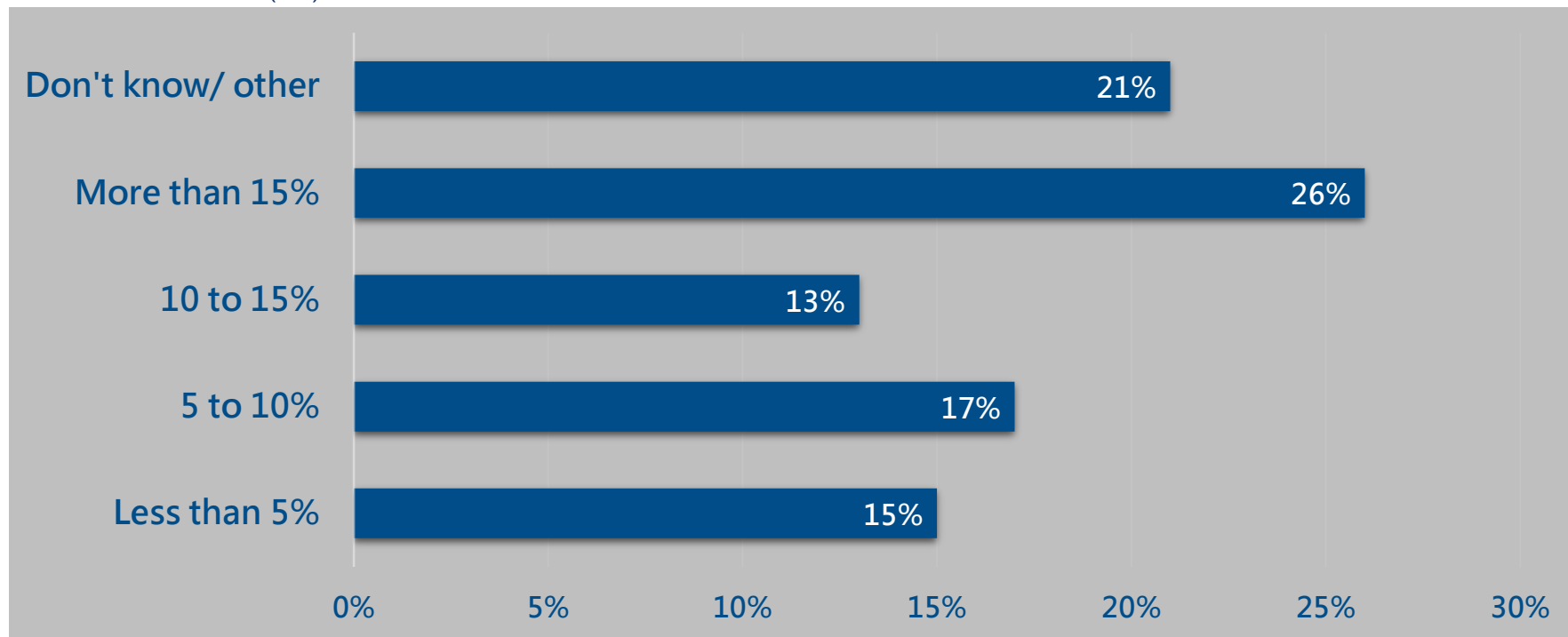
Outline



- ▶ Worldwide AI Market and Ecosystem
- ▶ AI Deep Learning in Cloud and 5G
- ▶ Edge Computing-Inference
 - ◆ Smartphone
 - ◆ Automotive
 - ◆ Low Power Device-Base on MCU
- ▶ Chip Company Strategy-Xilinx , NXP
- ▶ China AI Industry Overview & Start-up AI Chip Companies
- ▶ Summary

IT Budget for 2017 Is Earmarked for ML

Machine Learning
在IT预算的比例分配 (%)



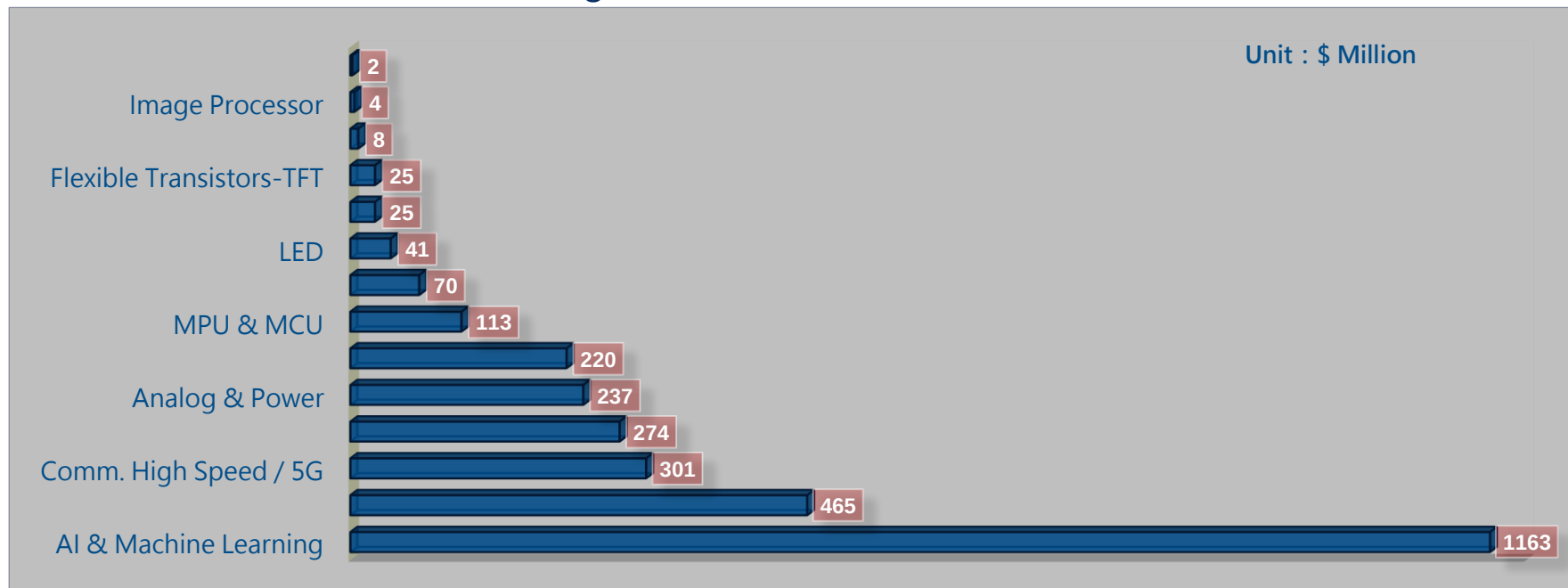
Source : MIT Technology Review & Google Cloud

参与调查企业的比重 (%)

Worldwide Fabless Company Venture Capital Funding

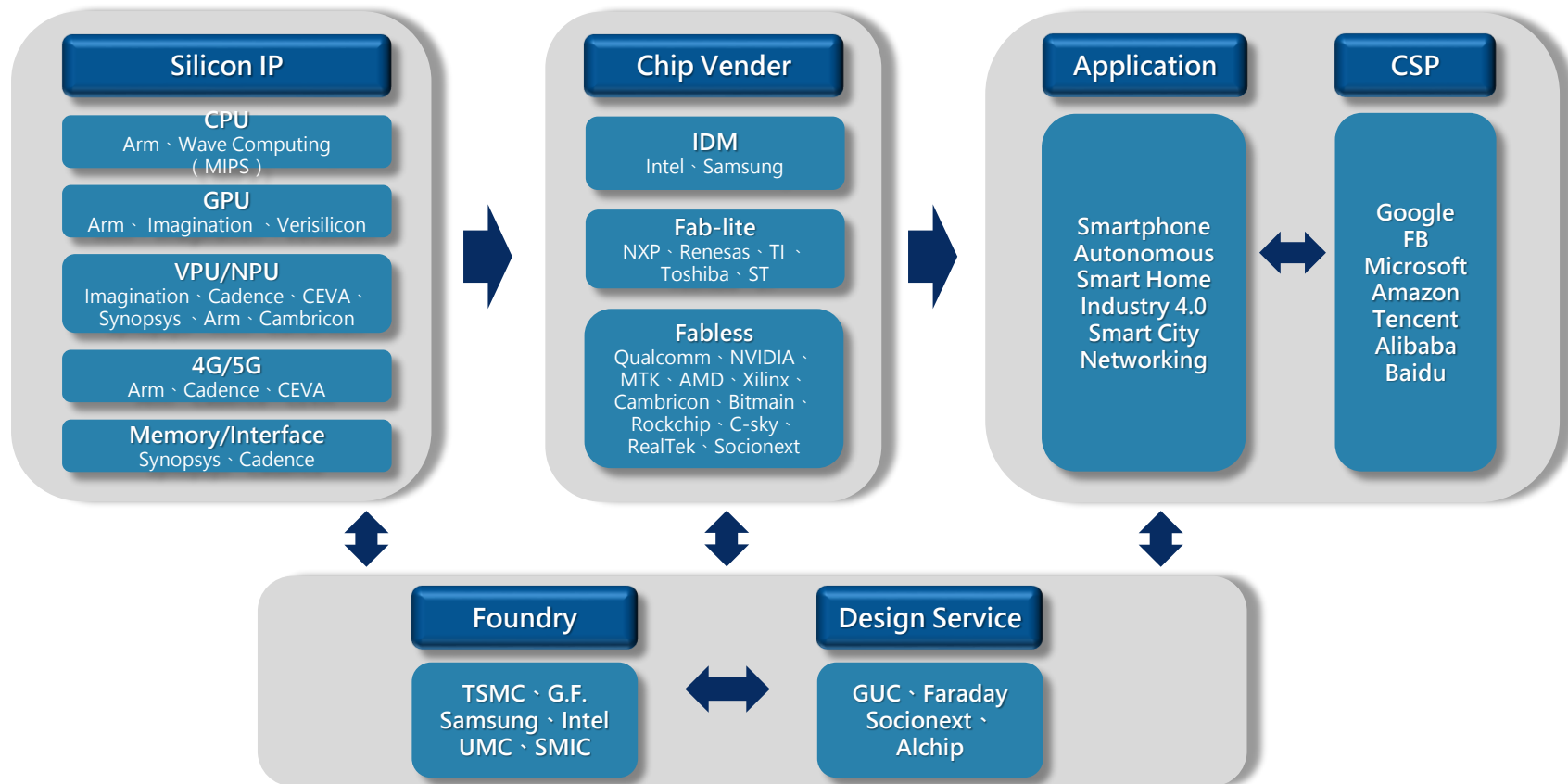


Market Segments Funded 2012 – 2018 YTD



Source : Mentor Graphic ; 整理 : 拓璞产业研究院 (2018/8)

AI Chip Ecosystem Overview



Cloud Inference-More Accelerator Cards



2018.08.13

2018.08.20

2018.09.12



Design &
Visualization
QUADRO RTX

产品 型号	RTX 5000	RTX 6000	RTX 8000
CUDA Cores	3072	4608	4608
Tensor Cores	384	576	576
RT Cores	Yes	Yes	Yes



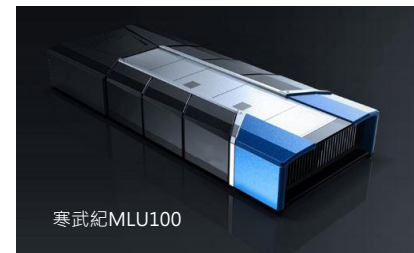
Gaming
GeForce RTX

产品 型号	RTX 2070	RTX 2080	RTX 2080 Ti
CUDA Cores	2304	2944	4352
Tensor Cores	288	368	544
RT Cores	Yes	Yes	Yes



DC AI
Tesla T4

- CUDA Cores : 2560
- Tensor Cores : 320
- 传输接口 : PCIe Gen3 x16
- 功耗 : 75W
- 应用 : Inference
for voice, video, image
and recommendation
services



寒武纪MLU100

Networking-AI in 5G

"Automation & AI are going to be game-changers for the telecom industry"



— Börje Ekholm,
President and CEO,
at MWC 2017

Automation, driven by machine intelligence, has the potential to change today's workplaces as dramatically as the machines of the industrial revolution changed the factory floor



5G Systems: Network Slicing and Artificial Intelligence as a Service for Big Data Value

Dr. David Soldani
Chief Technology Officer Huawei (Australia)

May 2018

ReefShark – Concentrated power for RF and baseband processing



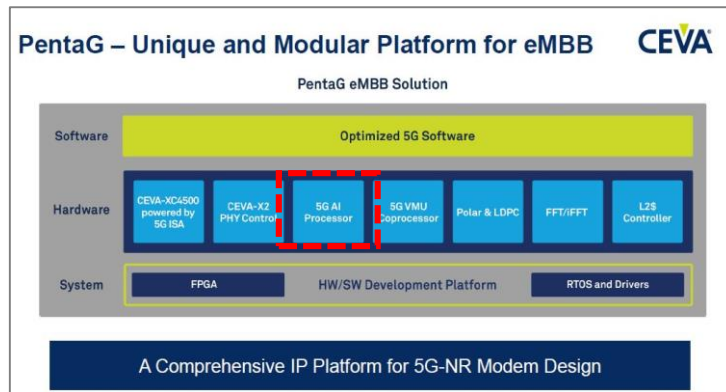
Pushing the limits with in-house silicon innovation. Some nodes as small as 10nm

AI capabilities embedded within radio and baseband processing

Compute optimized for all layers of the network edge

- RFIC and transceiver: massive MIMO Adaptive Antenna solution
- Digital Front End for LTE and 5G radio systems supporting massive MIMO
- Baseband Processor

© 2018 Nokia The TAO to make 5G move NOKIA



Edge Computing-Inference

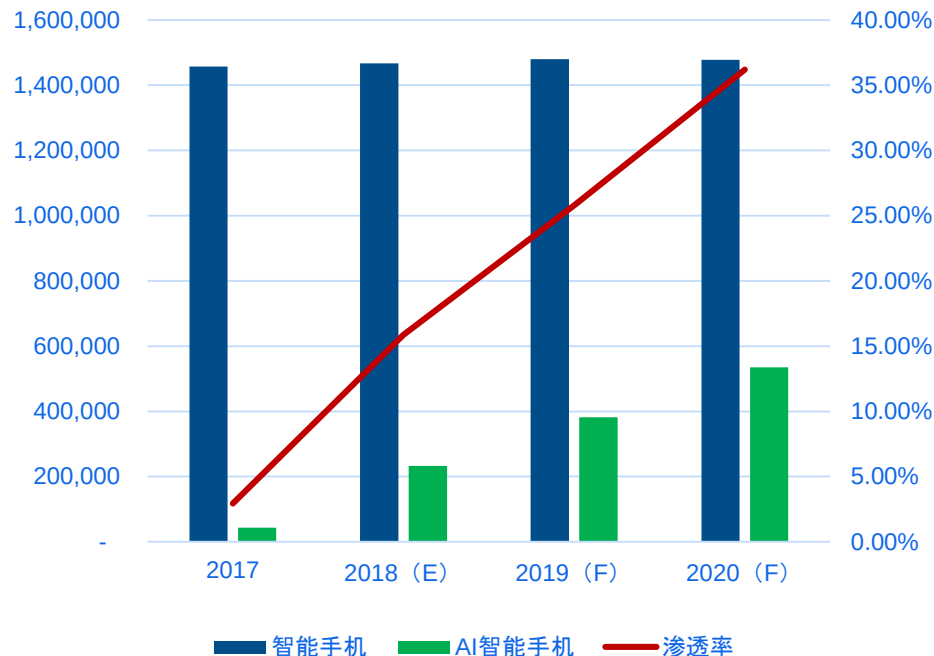


- Smartphone
- Automotive
- Low Power Device-Based on MCU



AI Smartphone Market and Premium AP in 2019

单位：千台



source : TrendForce and TRI



A11 Bionic

- 六核CPU
- 三核GPU
- Neural Enginex2
- 10nm

2017/09/13

A12 Bionic

- 六核CPU
- 四核GPU
- Neural Enginex8
- 7nm

2018/09/09



Kirin 970

- A73x4 + A55x4
- Mali-G73 x12
- NPU / Image DSP
- 10nm

2017/09/01

Kirin 980

- A76x2 + A76x2 + A55 x4
- Mali-G76 x10
- NPU x2
- 7nm

2018/09/02



SDM 845

- Kryo 385 CPU x8
- Adreno 630 GPU
- Hexagon 685
- 10nm

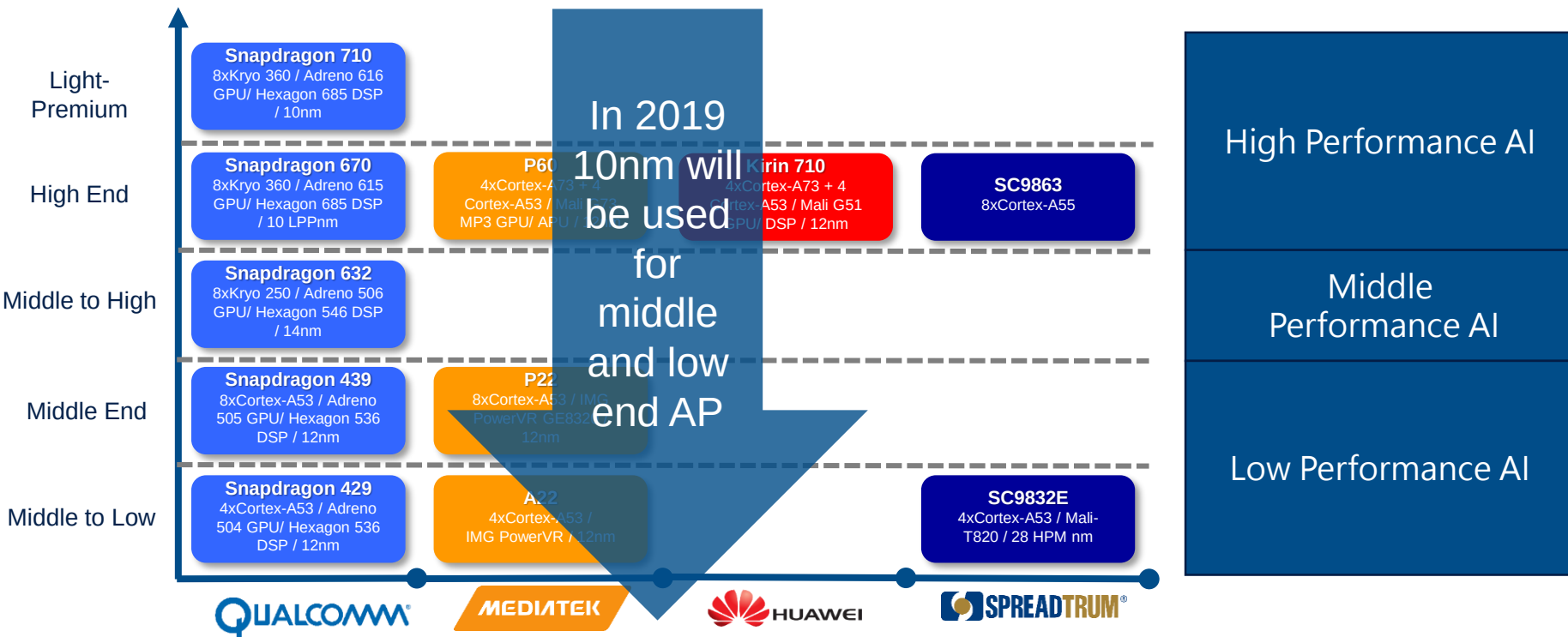
2017/12/06

SDM 855

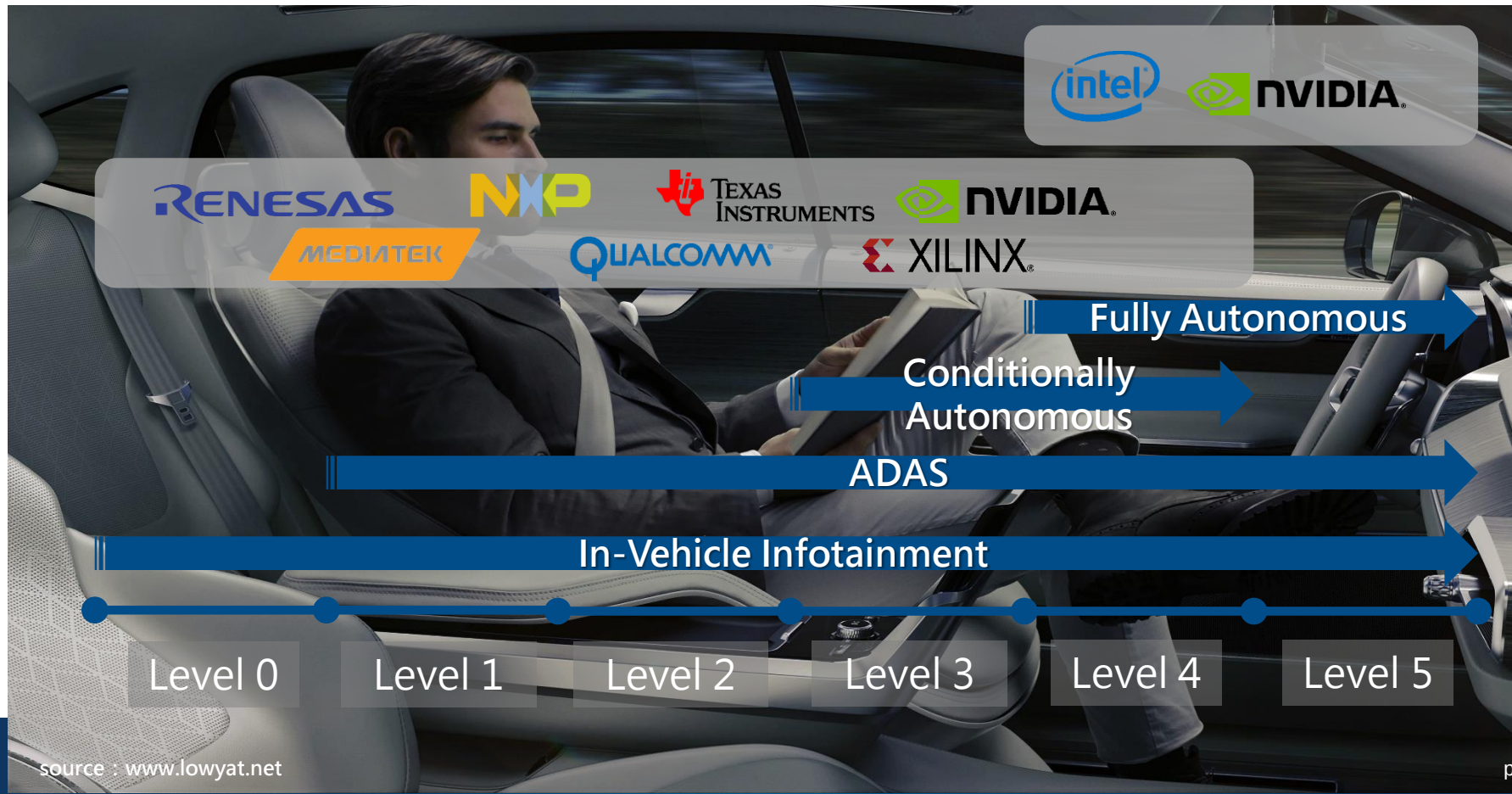
- ??
- ??
- ??
- 7nm

2018/12

Smartphone New AP in 2018

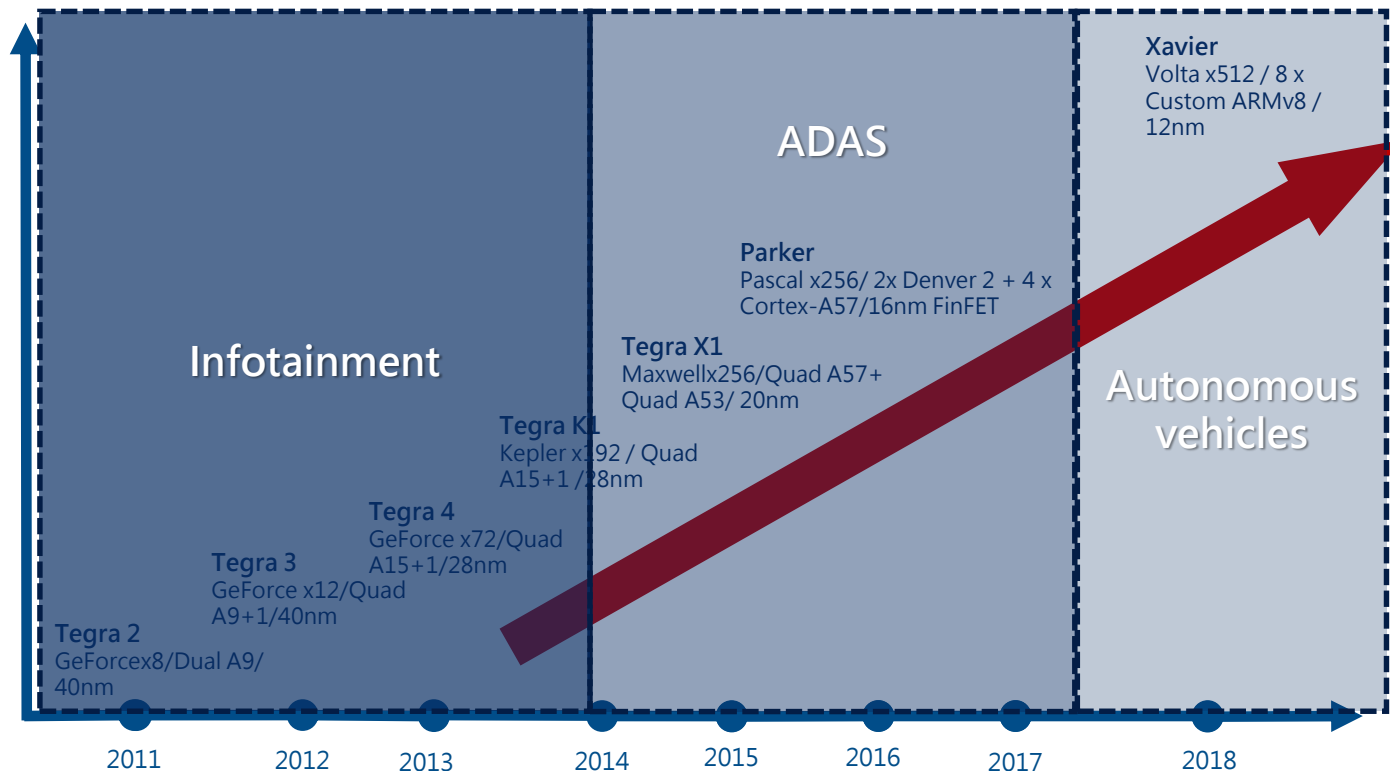


Automotive- Central Processor

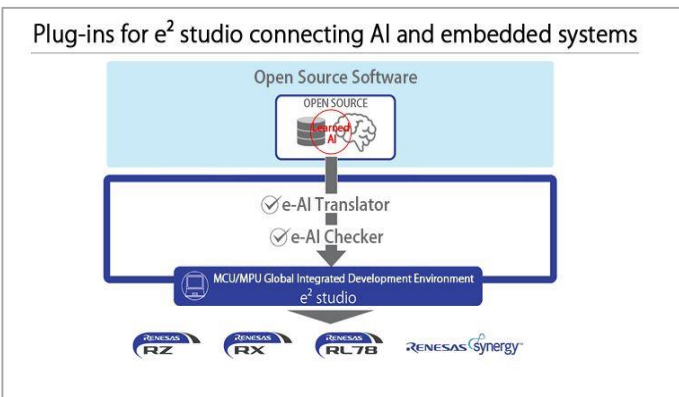
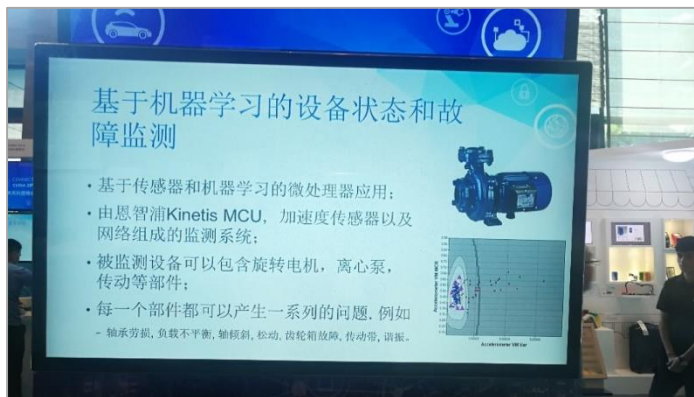
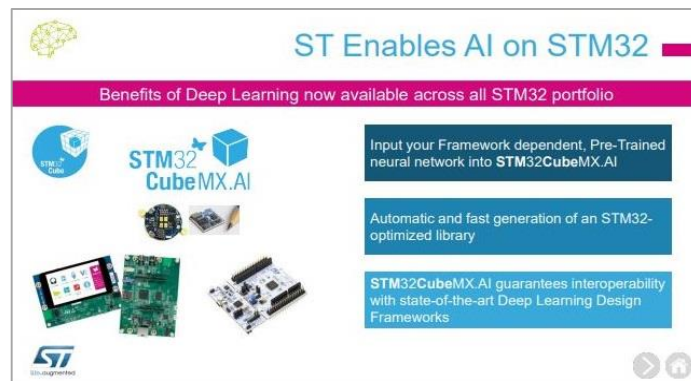
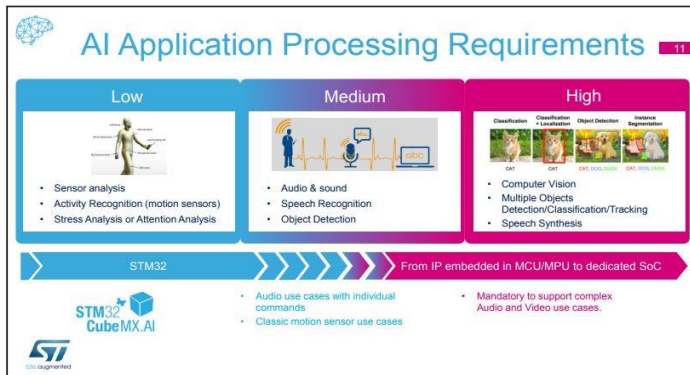


Automotive- Central Processor (1)

Performance



Low Power Device-Base on MCU



先进制程与统一核心硬件架构

7nm
FinFET



Core Architecture

Scalar Engines

Dual A72+Dual R5

Adaptable
Engines

FPGA

Intelligent
Engines

AI Engines +
DSP Engines



Versal
Portfolio

Markets

AI Core

Data Center, Wireless

AI Edge

Automotive, Wireless,
Broadcast, A&D

从云到端，
满足各类推论应用场景

Prime

Data Center, Wired ,
Multiple Market

Premium

Wired, Test and
Measurement

HBM

Data Center, Wired,
Test and Measurement

大会重要讯息

物联网/
人工智能

富士康工业
互联网

打造奠基于「云端+行动终端+硬设备」的先进工业物联网平台，并进行充分的客制化开发，以满足制造业错综复杂的需求。

中科虹霸

为中科虹霸提供完整的虹膜辨识芯片解决方案，让广大用户更快、更好地享受人工智能带来的乐趣和便利。

车用电子/
自驾车

毫米波雷达

频段为76-81GHz，在新的参考平台上将S32R处理器、TEF810x CMOS射频收发器及天线设计组合在一起，并提供创新的雷达软件开发套件。

逆变器
开发工具包

与VEPCO Technologies (Tire 1)、Fuji Electric (IGBT 开关组件) 合作，开发逆变器与BMS参考平台

收购
OmniPHY

OmniPHY是高速汽车以太网IP领域的领导者，提供涵盖100BASE-T1及1000BASE-T1标准的15个IP系列

大会站台厂商

阿里云

在物联网产品分成两大块：AliOS Thins (2G-5G/NB-IoT/LoRa) 以及Link Edge (Wi-Fi/ZigBee/BLE)

京东

尽可能利用无人载具实现无人货仓应用，另外也有智慧巡检机器人，用于夜间巡检，结合相当多元的感测技术与无线传输技术可以辅助甚至是替代保安人员在夜间巡仓的工作

国际芯片大厂与CSP/当地应用端业者合作，
打造AI应用场景已成必然趋势

吉利汽车

智能驾驶的发展规划：预计2020年为G-Pilot 3.0，2024年为G-Pilot 4.0；G-Pilot 2.0已经具备智慧自适应巡航功能与行人自动紧急制动

百度

百度不做车，但采取开放策略，与车厂合作，已有不少车厂采用百度平台，全球已有119个合作伙伴，有90个以上的项目正在进行当中，相关的开发者，超过1万名以上

China AI Industry Overview

北京市政府
行政区域图

全中国共有4040家AI企业，
北京则有1070家

第一步

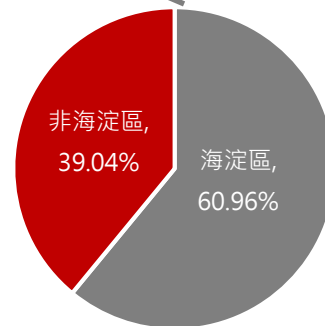
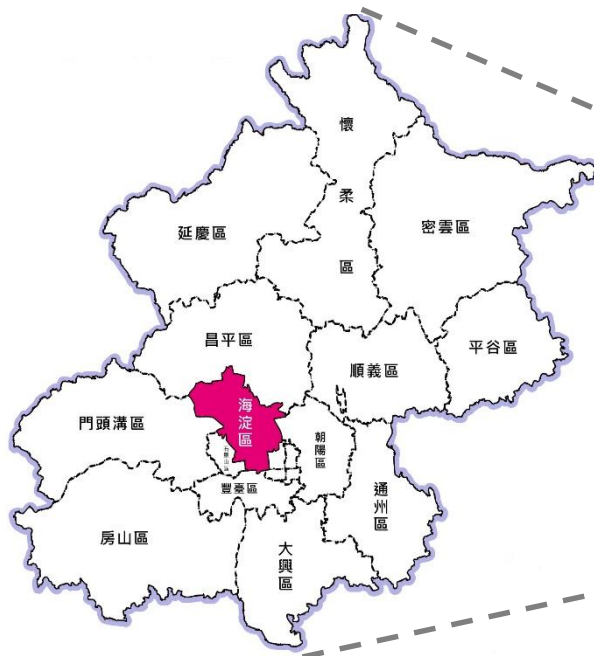
- 到2020年人工智能总体技术和应用与世界先进水平同步
- 人工智能核心产业规模超过1500亿元，带动相关产业规模超过1万亿元。

第二步

- 到2025年人工智能基础理论实现重大突破，部分技术与应用达到世界领先水平
- 人工智能核心产业规模超过4000亿元，带动相关产业规模超过5万亿元。

第三步

- 到2030年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心
- 人工智能核心产业规模超过1万亿元，带动相关产业规模超过10万亿元。



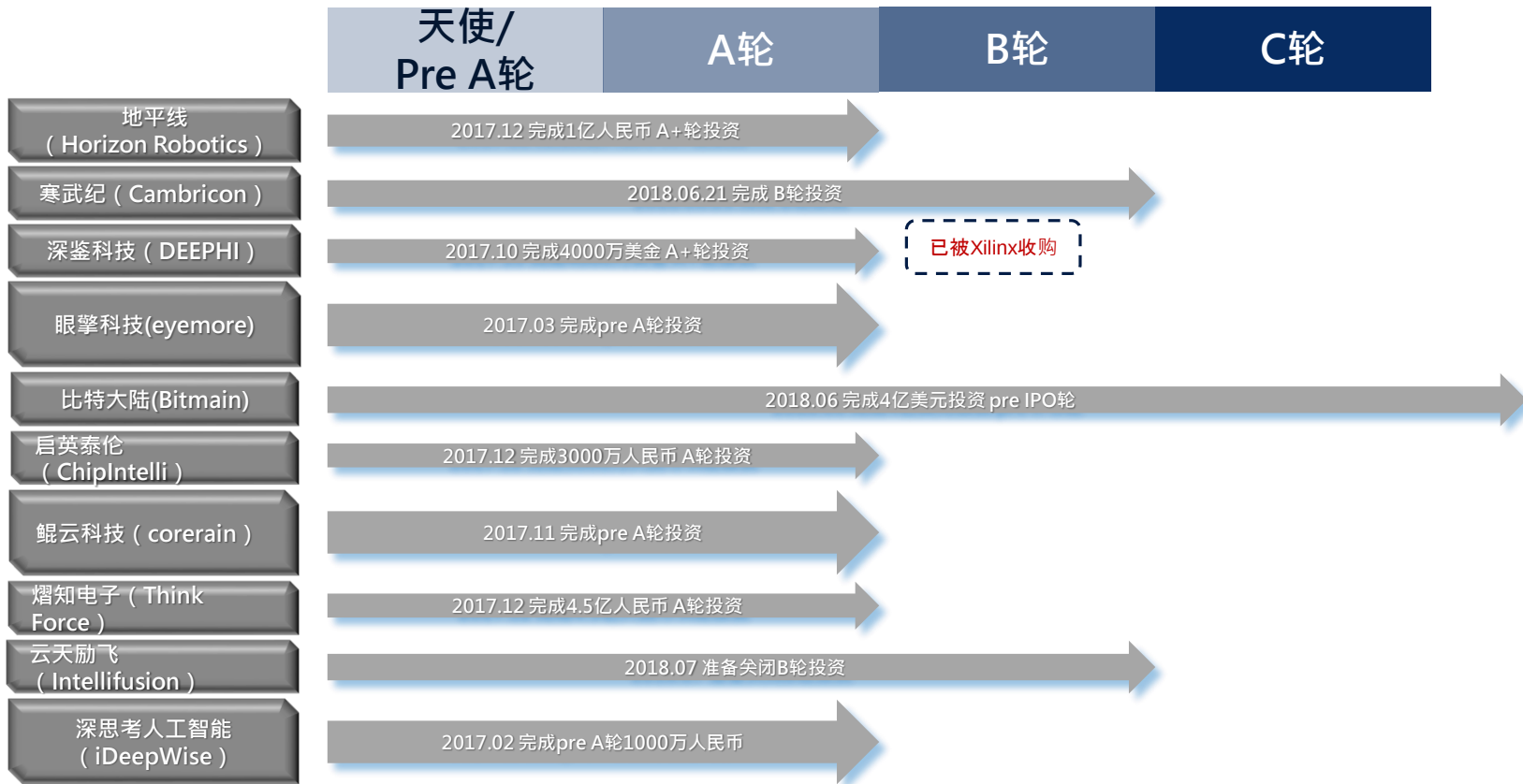
Source : 中国国务院 ; 2017/07

Source : 北京人工智能產業發展白皮書 ;
整理 : TRI (2018.10)

China Start-up AI Chip Companies-1

公司名称	主要投资方	成立日期	主要产品/应用
地平线 (Horizon Robotics)	2015/10：晨兴资本、高瓴资本、红杉资本、金沙江创投、线性资本、创新工厂、真格基金、 2016/04：Yuri Milner 2016/07：双湖投资、青云创投、祥峰投资 2017/10：Intel、嘉实投资	2015 (北京)	自动驾驶、智能城市、智能零售
寒武纪 (Cambricon)	2016：元禾原点、科大讯飞、涌铎投资 2017：国投创业、阿里巴巴、联想创投、国科投资、中科图灵、元禾原点、涌铎投资	2016成立 (北京) (2008已有研究团队)	行动处理器IP、MLU芯片/版卡 (For Server)
深鉴科技 (DEEPI)	Xilinx、蚂蚁金服、联发科、Samsung、清华控股、方和资本、华创资本、招商局创投、金沙江创业投资、BANYAN Capital、阿里巴巴	2016 (北京)	智能安防、智能辅助驾驶、智能大数据
眼擎科技(eyemore)	京基资本(京基集团)、柔宇科技	2014 (深圳)	智能视觉方案
比特大陆(Bitmain)	红杉资本、IDG Capital、美国对冲基金Coatue、投资基金EDB	2013 (北京)	Training/Inference处理器/板卡/Server
启英泰伦 (ChipIntelli)	上海辰韬资产管理	2015 (成都)	人工智能深度学习语音识别专用芯片
鲲云科技 (corerain)	星瀚资本、深圳云创、珠海拓金	2016 (深圳)	Image Deep Learning AI芯片
耀知电子 (Think Force)	依图科技、云锋基金、红杉资本、高瓴资本	2017 (上海)	AI芯片
云天励飞 (Intellifusion)	真格基金等投资机构首笔天使轮投资	2014 (深圳)	AI芯片、图像识别、前端Camera
深思考人工智能 (iDeepWise)	昕朴投资、九阳股份	2015 (北京)	硬件芯片、软件 智慧医疗、语义交互

China Start-up AI Chip Companies-2



China Start-up AI Chip Companies-3



	硅IP	算法/SDK	模块/板卡	System
地平线	O	O	X	X
寒武纪科技	O	O	O	X
深鉴科技	X	O	O	X
眼擎科技	O	O	O	O
比特大陆	X	X	O	O
肩英泰伦	X	X	O	X
鲲云科技	X	X	X	X
熠知电子	X	O	X	X
云天励飞	X	O	O	O
深思考人工智能	X	O	X	X

Summary-2018年发展与2019年展望



2018

- 手机旗舰处理器AI运算效能持续提升，持续聚焦“视觉”相关应用
- NVIDIA居自驾车中央处理器技术领先地位
- MCU AI功能开始落地
- 大陆新创AI Chip公司勃蓬发展，
- 统一硬件运算架构蔚为AI芯片产业风潮

2019

- 手机处理器AI运算功能将会逐渐下放至低阶市场
- 车用中央处理器业者短期内应无太多动静，可观察车厂导入状况
- 系统业者须从「应用场景」层面考虑MCU系统开发
- 大陆新创 AI Chip产品逐渐导入终端应用
- 统一硬件运算架构有助生态系统发展

运算架构与先进制程演进带动AI芯片效能持续提升与进入门槛降低

晶圆代工业者

硅智财业者

Thank you



The TRI logo is contained within a white rectangular box. It features the Chinese characters "拓璞" in a bold, black, sans-serif font, followed by "TRI" in a similar font. Below this, the words "TOPOLOGY RESEARCH INSTITUTE" are written in a smaller, black, sans-serif font. A small red square is positioned above the "I" in "TRI".

Yao Chia Yang, Senior Analyst ,TRI

Email: Cyyao@topology.com.tw

www.ceva-dsp.com

www.topology.com.tw/