

2025 Oct. 17 (Fri) - AI & Semiconductor Strategic Innovation



Dr. **Wei-Chung Lo**

Deputy Director General, EOSL, ITRI

PM 03:20 - 03:50

決勝AI邊緣運算：小晶片堆疊設計與異質整合

Edge AI: Chiplet-based Stacking Technology & Heterogeneous Integration

ITRI

Industrial Technology
Research Institute

決勝AI邊緣運算：

小晶片堆疊設計與異質整合

**(Edge AI: Chiplet-based Stacking Technology &
Heterogeneous Integration)**

Wei-Chung Lo 駱韋仲

Deputy General Director/Senior Principal Engineer/IMAPS Fellow
Electronic and Optoelectronic System Research Laboratory(EOSL)

Oct. 17th, 2025



Dr. Wei-Chung Lo

Electronic and Optoelectronic System Research Laboratories (EOSL)
of Industrial Technology Research Institute (ITRI)



Dr. Lo received his Ph.D. from National Taiwan University(NTU) and joined ITRI in 1997. Currently, He is deputy general director and senior principal engineer in EOSL of ITRI and visiting professor in Graduate School of Advanced Technology of NTU. He serves as Executive Secretary of AI on Chip Taiwan Alliance(AITA), and Chairman of 3D IC/Fan-Out consortium (Ad-STAC, Hi-CHIPS)

He has published more than 85 papers and 27 patent granted. His research interests focusing on Semiconductor Device, Wafer-level System Integration and Advanced electronic/opto-electronic packaging for more than 20 years, including Non-Volatile Memory(MRAM, Computing in Memory), 3D IC/ 3D SiP/ 3D Chiplets Integration, Silicon Photonics & Co-packaged Optics, WBG Power chip(GaN, SiC & Ga₂O₃), Power packaging and Power module, RF packaging(Antenna-in-package, AiP), and Integrated smart system and Heterogeneous Integration technology.

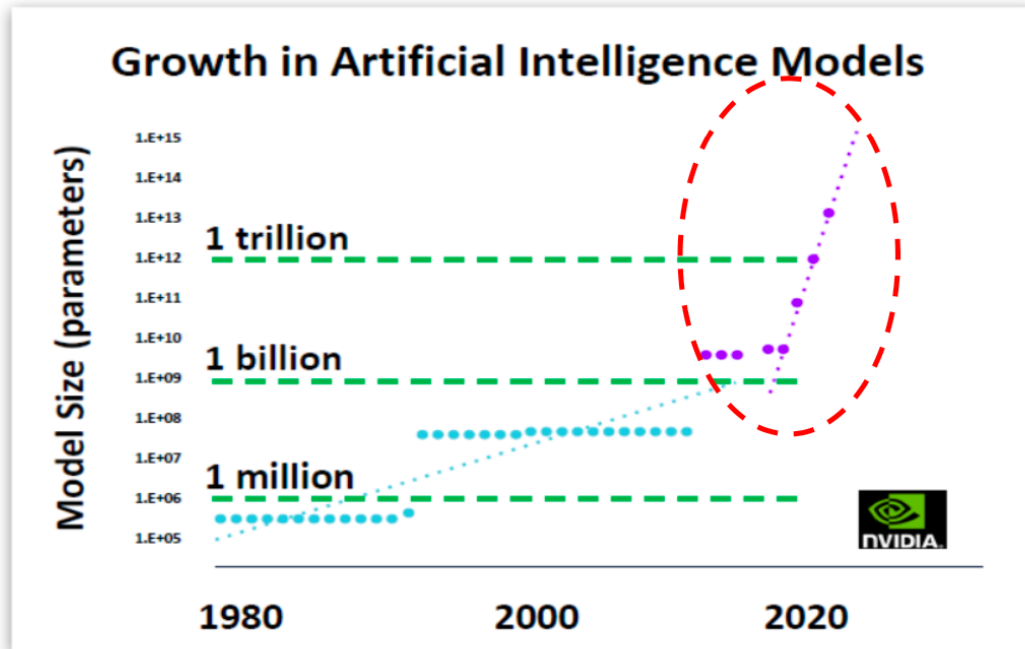
Awards:

- ✓ **“Fellow of IMAPS 2024”, “Chenming Hu Award-Semiconductor device, material and process Innovation Award@VLSI-TSA 2024”**
- ✓ **“Industry Contribution Award”** by SEMI Taiwan(2025/2024/2023/2022),
- ✓ **“Outstanding Engineer Award”(2021),**
- ✓ **“National Manager Excellence Award”(2018), “Innovation Elite Award”(2015) and “ITRI Elite”(2025 & 2013)”**

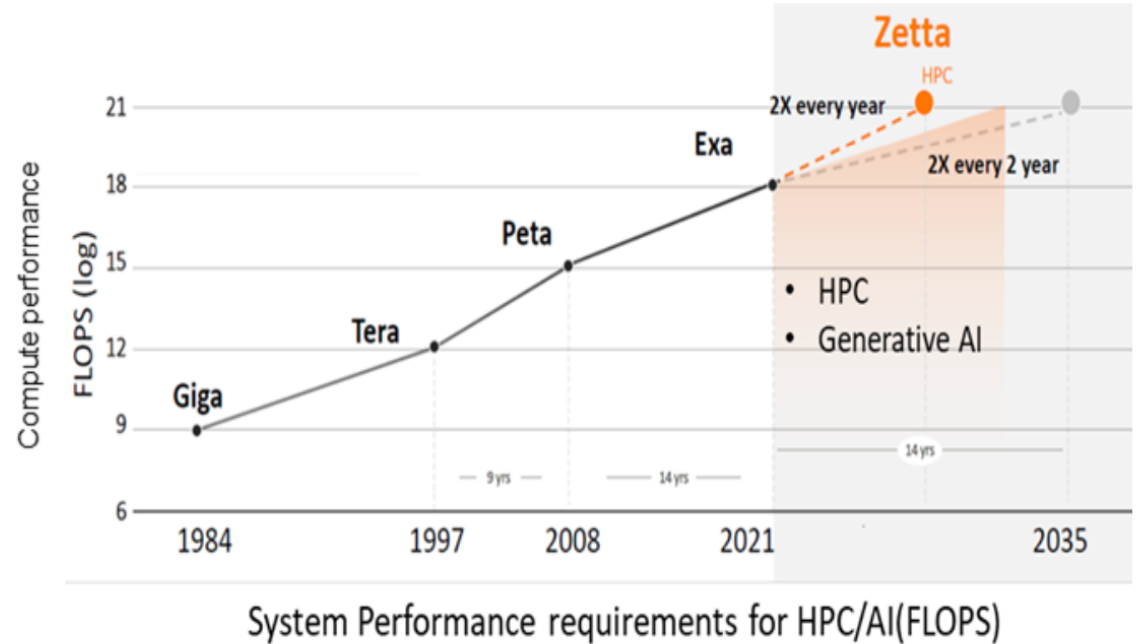
Outline

- **Semiconductor Trend and Challenges**
 - ✓ **Chip-level Tech Trend and System on Wafer(SoW)**
 - ✓ **Chiplets –Revolution of System Integration**
- **Chiplets and Heterogeneous Integration: Key Enabler for Edge AI**
 - ✓ **Design Technology and Service Platform**
 - ✓ **3D IC Technology and Service Platform**
- **Summary**

GenAI Enabling New Demands



Nvidia GTC March 2022



Source: "The patch to Zetta scale" by Wilfred Gomez, Intel Corp

- Digitalizing the real-world producing Zetta bytes of BIG DATA; 5G, 6G networks interconnecting and sensing at very high resolutions.
- AI Model parameters increased >5000X over 10 years to integrate information, analyze big data, and improve decision making in every walk of life.
- Generative AI demands increasing compute system performance (2X every year) to reach Zetta Scale Computing

The Challenge of Moore's Law

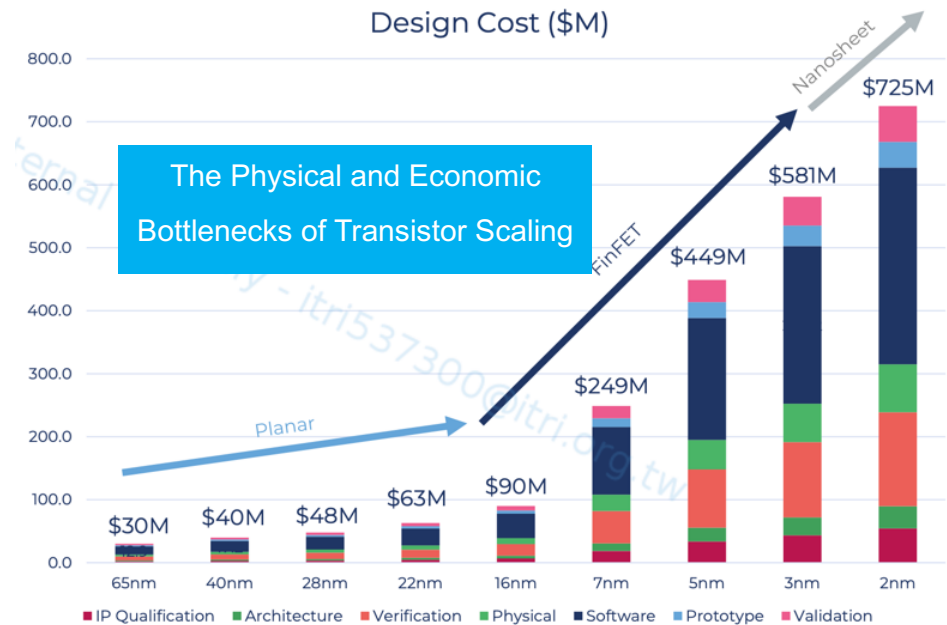
✓ Physical Limits:

As transistor sizes to atomic level, issues like quantum tunneling and heat density become increasingly severe.

Process Node	Key Physical Challenge	Economic Impact
< 28nm	Power Leakage (FinFET introduced)	Increased process complexity
< 7nm	Heat Density, Interconnect Resistance	Extreme litho costs (EUV)
< 3nm	Quantum Tunneling, Atomic-level defects	R&D costs exceed \$500M per chip

✓ Economic Limits:

The R&D and fabrication plant costs for each new process node (e.g., 5nm to 3nm) are growing exponentially.



Source: arm IPO, UCIE Consortium, modified by Yole

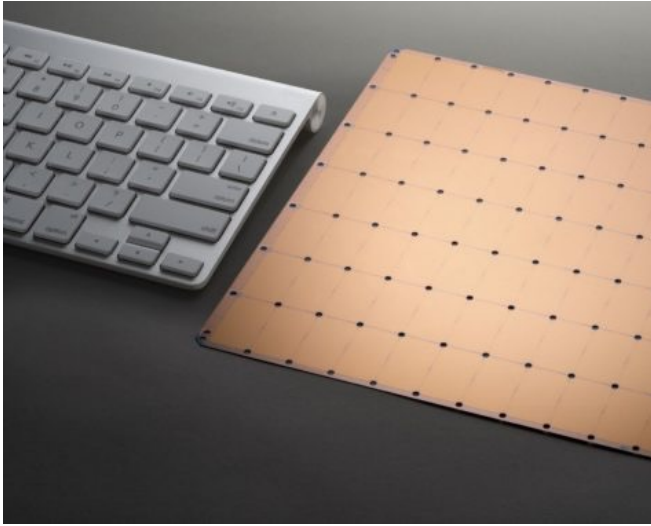
The Crossroads for SoC: (We Need a New Solution)

The SoC development model is facing a critical turning point.

- ✓ **The Cost Wall:** Prohibitive design and manufacturing costs.
- ✓ **The Yield Bottleneck:** Low production efficiency for large chips.
- ✓ **Integration Compromise:** Unable to balance performance and cost.
- ✓ **Market Inflexibility:** Lacks the flexibility to address fragmented demand.

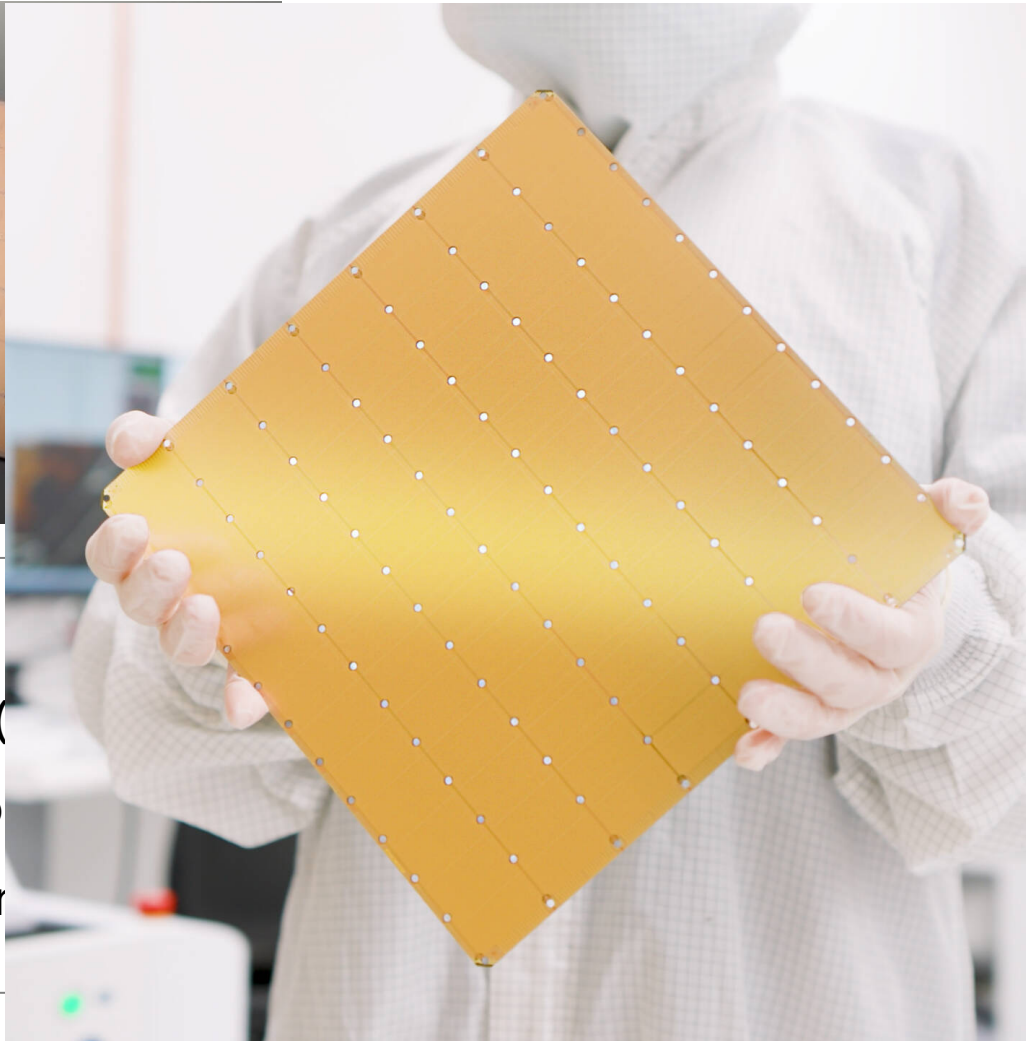
Challenge	Business Impact
Cost Wall	Reduced ROI, High barrier to entry for startups
Yield Bottleneck	Increased cost of goods sold (COGS), supply instability
Integration Compromise	Sub-optimal product performance and power
Market Inflexibility	Missed opportunities in niche markets, slow response time

System-on-Wafer



CS-1 vs. Newest GPU (CS-3 vs. H100)

- **X78** Computing Core (
- **x3,000** On-chip Memo
- **x33,000 BW** Interconn



The world's largest commercial chip

Cerebras WSE-3 in 03/2024

<https://www.cerebras.net/product-chip/>

**The Cloud AI
Supercomputer Built for
Generative AI**

- **215mm x 215mm**

- **46,225 mm²**, 4 trillion transistor chip

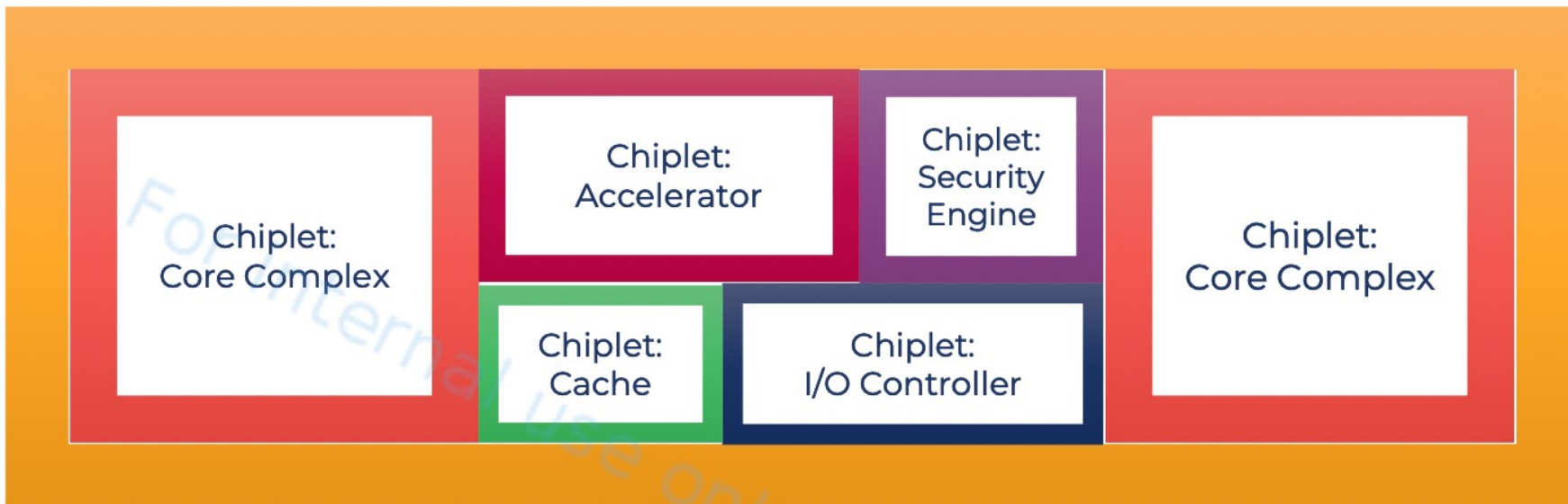
- **900,000** cores

- **44 GB** on-chip SRAM

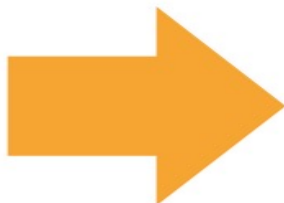
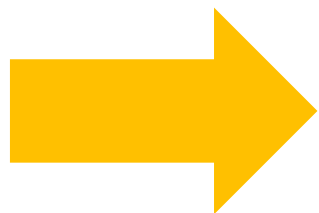
- **214 Pb/s** interconnect BW

- **TSMC 5nm process**

Chipselets



Advanced Packaging
Platforms integrating chiplets



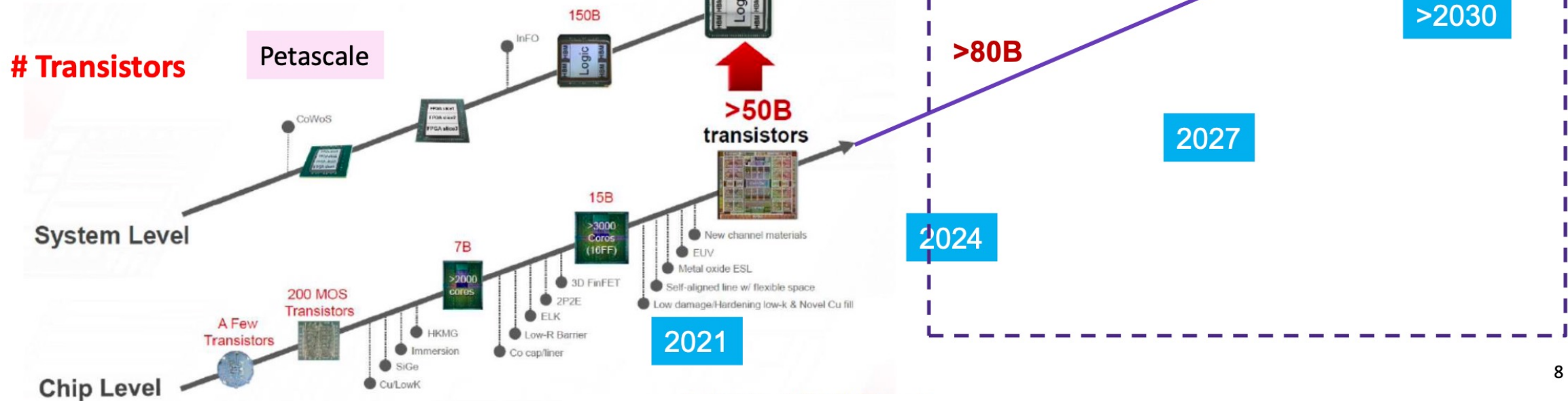
Fan-out

Flip-Chip

2.5D/3D

Chip Level vs System Level

- Exascale to Zetta-Scale Compute performance drives Package integration to reach >1T transistors in Advanced Package
- Ultra Large Multi-chiplet System in Package the key building block to achieve accelerated computing.
- Need Power-Performance-Formfactor-Cost Optimization



Source: Dr. Mark Liu, tsmc@ISSCC 2021 Keynote

The Revolutionary Solution of Chiplets

(A New Era of Modularity, Flexibility, and Innovation)

From Monolithic Stone to Flexible LEGO Bricks

- ✓ **Monolithic SoC:** Like a giant, intricately carved stone—integrated but inflexible.
- ✓ **Chiplet-based System:** Like LEGO bricks, freely combined and replaced to create endless possibilities.

Approach	Analogy	Key Characteristic
Monolithic SoC	Carved Marble Statue	All-in-one, fixed design
Chiplet System	LEGO Model	Modular, reconfigurable, customizable

Future of Edge AI: Chiplet Application Scenarios

Providing a Customizable Platform for Fragmented AI Applications

The modular nature of chiplets perfectly matches the diverse needs of Edge AI.

- ✓ **Smart Camera:** Base Processor Chiplet + AI Vision Accelerator Chiplet + Image Sensor Interface Chiplet.
- ✓ **Industrial Robot:** Base Processor Chiplet + Motor Control Chiplet + Real-time Communication Chiplet.
- ✓ **Autonomous Vehicle:** Base Processor Chiplet + Multiple AI Compute Chiplets + Sensor Fusion Chiplet.

Key Takeaway	Implication
Modularity is Key	Faster time-to-market and increased design flexibility.
Heterogeneous Integration	Optimal performance, power, and cost for every function.
Open Ecosystems	Drives innovation and lowers barriers for new market entrants.

Outline

- Semiconductor Trend and Challenges
 - ✓ Chip-level Tech Trend and System on Wafer(SoW)
 - ✓ Chiplets –Revolution of System Integration
- **Chiplets and Heterogeneous Integration: Key Enabler for Edge AI**
 - ✓ **Design Technology and Service Platform**
 - ✓ **3D IC Technology and Service Platform**
- Summary

Heterogeneous Computing/Integration: Design Technology and Service Platform

Need for Scalable AI Architecture

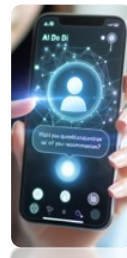
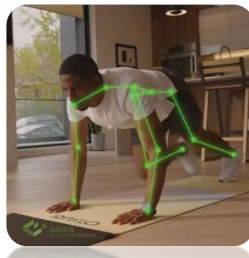
- Edge AI Applications vary widely, so do the performance requirements
- Need a scalable AI architecture to meet the diverse use cases



CNN Model



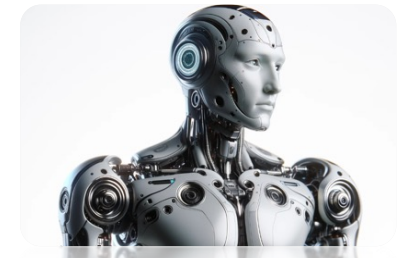
Vision Transformers



LLMs



MM-LLMs



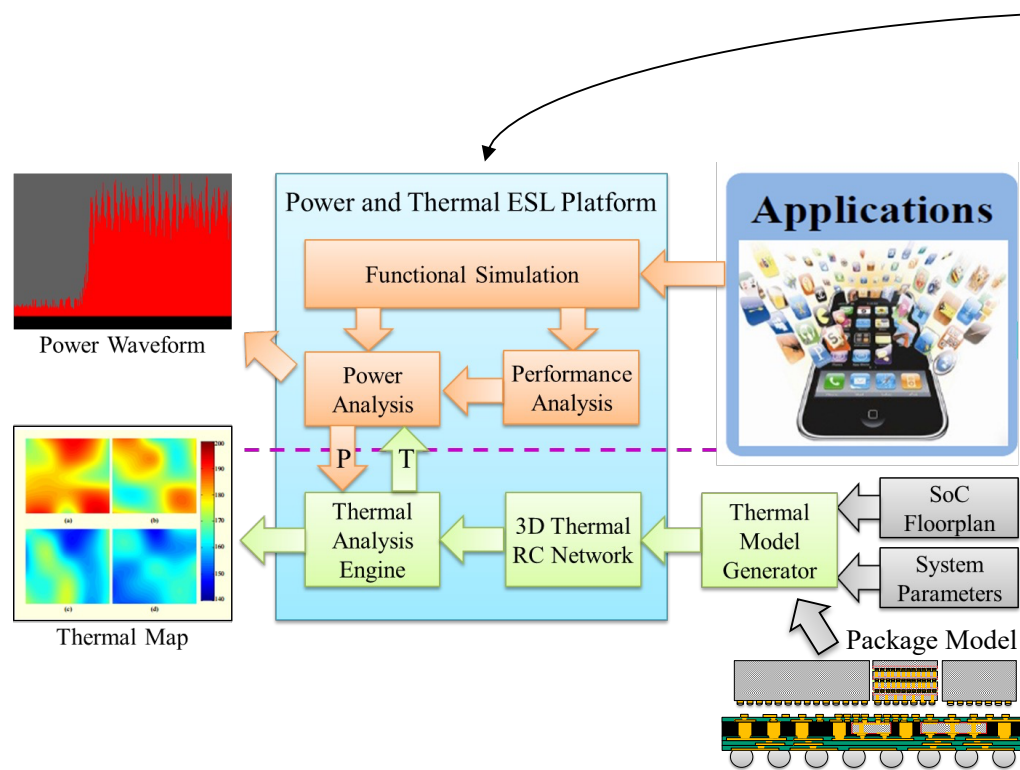
1 TOPS

10 TOPS

100 TOPS

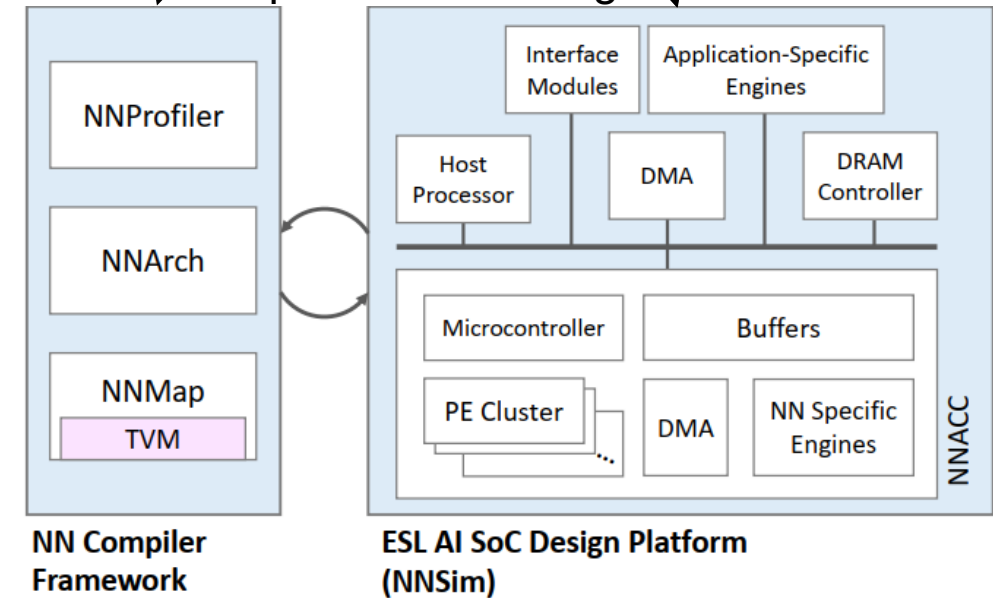
Hardware/Software Co-Design

Multi-Chip Package-Level Co-Design



Fast Steady-State Thermal Analysis
(ISOC'18)

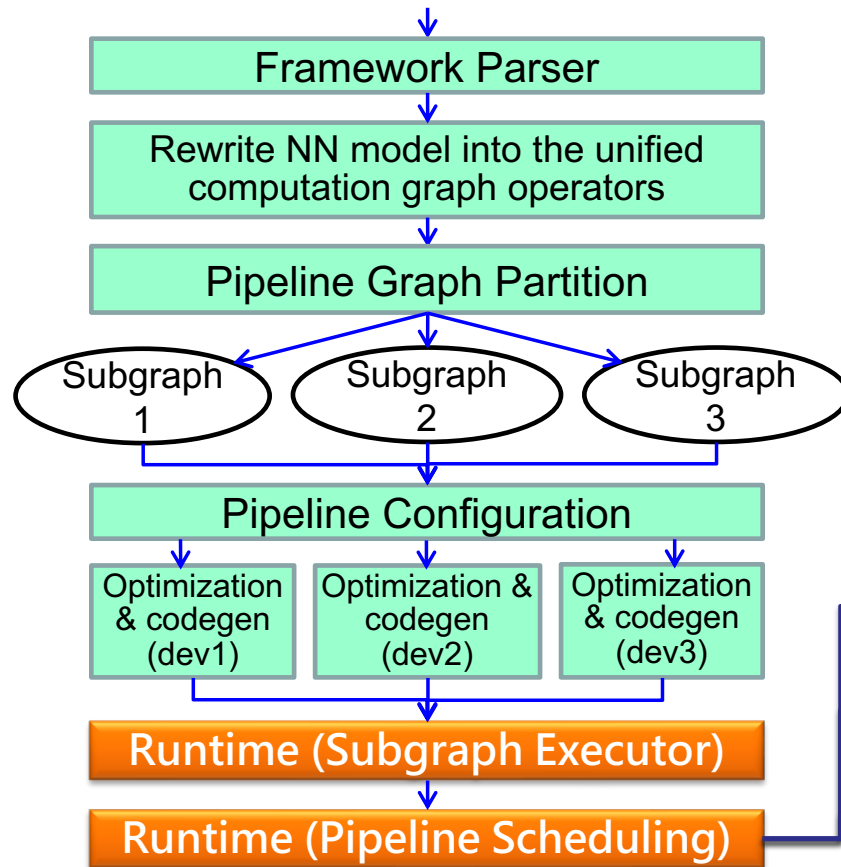
Chip-Level Co-Design



Optimization of AI SoC with Compiler-assisted Virtual Design Platform
(ISPD'23)

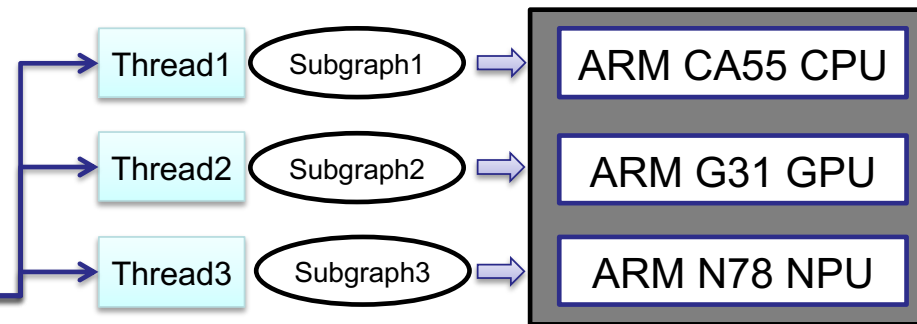
AI Compiler for Heterogeneous Computing

AI Model



Performance boost 1.46x on average with ITRI AI Compiler for the same HW

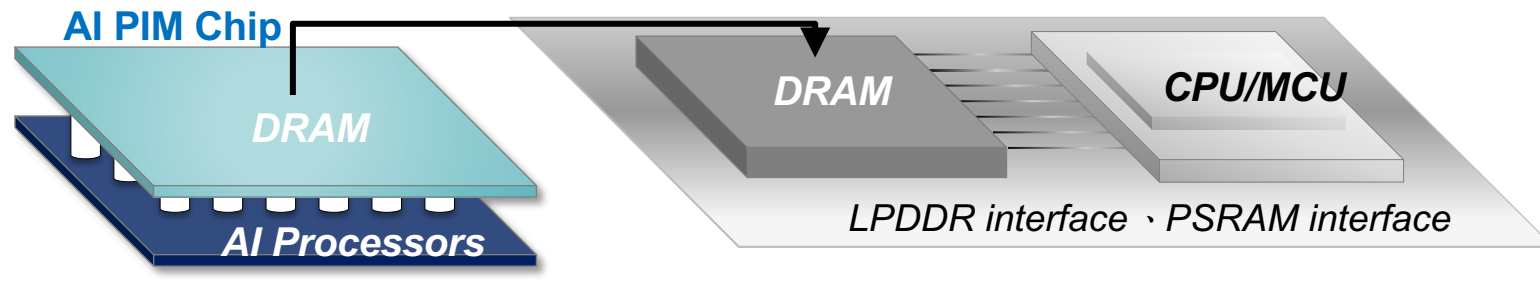
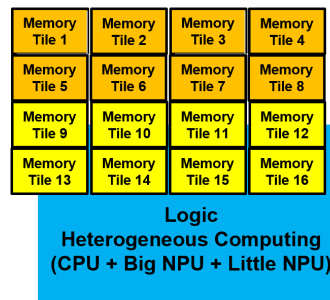
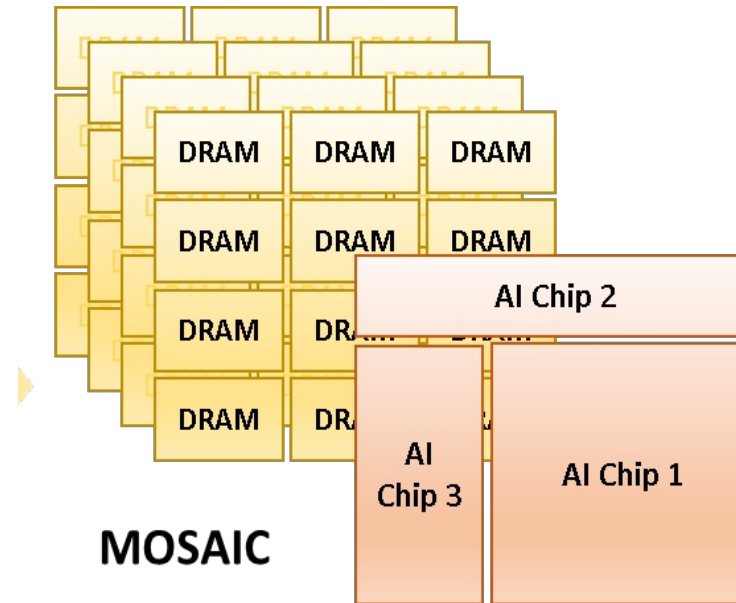
AI Model	NPU	CPU+NPU	Improvement	CPU+NPU+GPU	Improvement
ResNet50 (int8)	0.02337	0.017693	1.32x	0.016776	1.39x
SqueezeNet (int8)	0.013281	0.010944	1.21x	0.009354	1.42x
LeNet (int8)	0.002355	0.00175	1.35x	0.001632	1.44x
ESPCN Super Resolution (int8)	0.011476	0.008717	1.32x	0.007286	1.58x



Customer's AI Chip

MOSAIC 3D AI Chip

Memory-cube Operability in a Stacked AI Chip for Generative AI



<https://www.rdworldonline.com/mosaic-memory-cube-operability-in-a-stacked-ai-chip-for-generative-ai/>

Heterogeneous Integration:

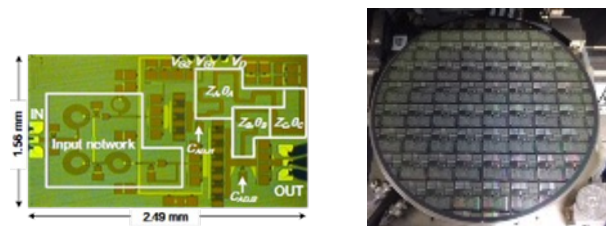
3D IC Technology and Service Platform

(Taiwan Cbl: WL-Chiplet Integration Shuttle)

Potential Collaboration

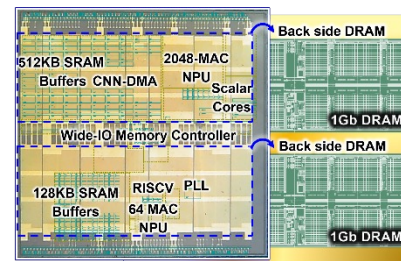
Device

High Frequency



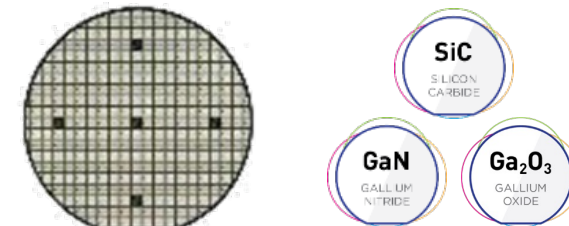
160 GHz PA with 8" GaN-on-Si

High Speed



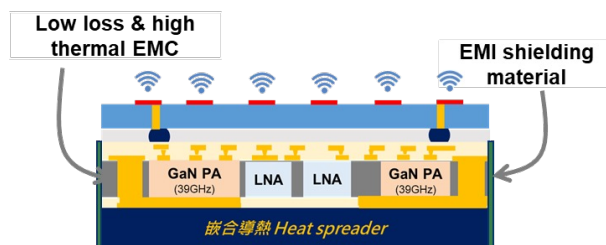
3D AI Chip

High Power

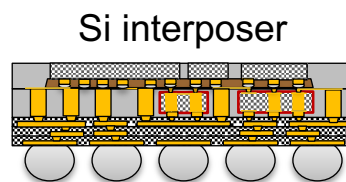


WBG Power Devices

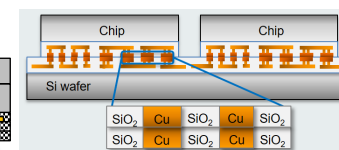
PKG



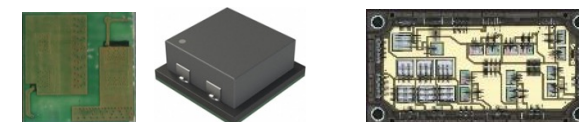
Antenna-in-Packaging by Fan-out



Si-bridge



hybrid bonding



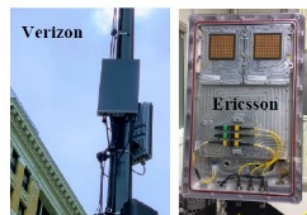
Power SiP

Power Module

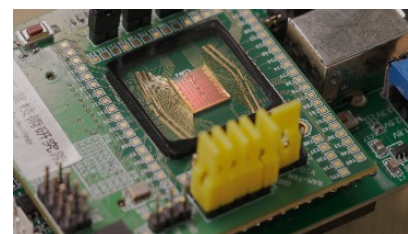
Sub-system



5G Router



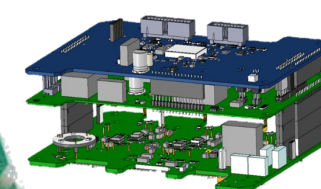
Infrastructure



AI Accelerator



Charger

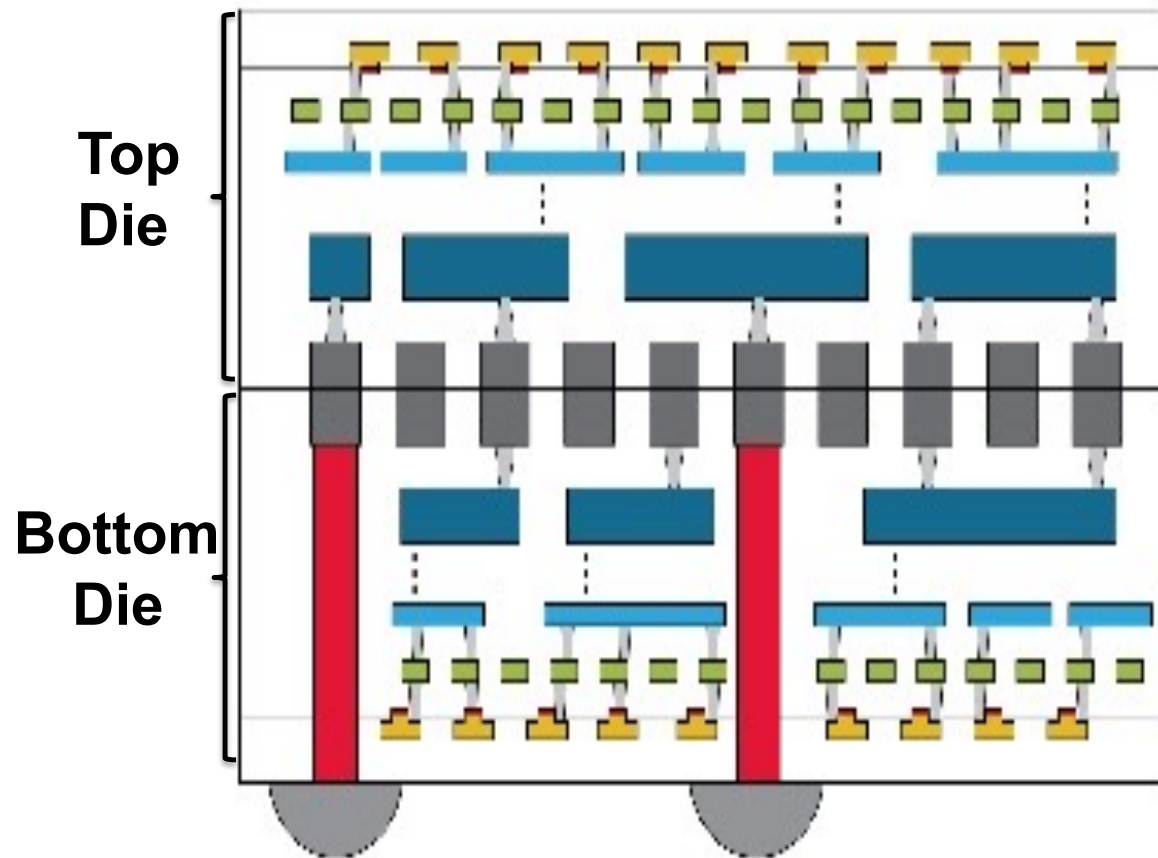


Motor Driver

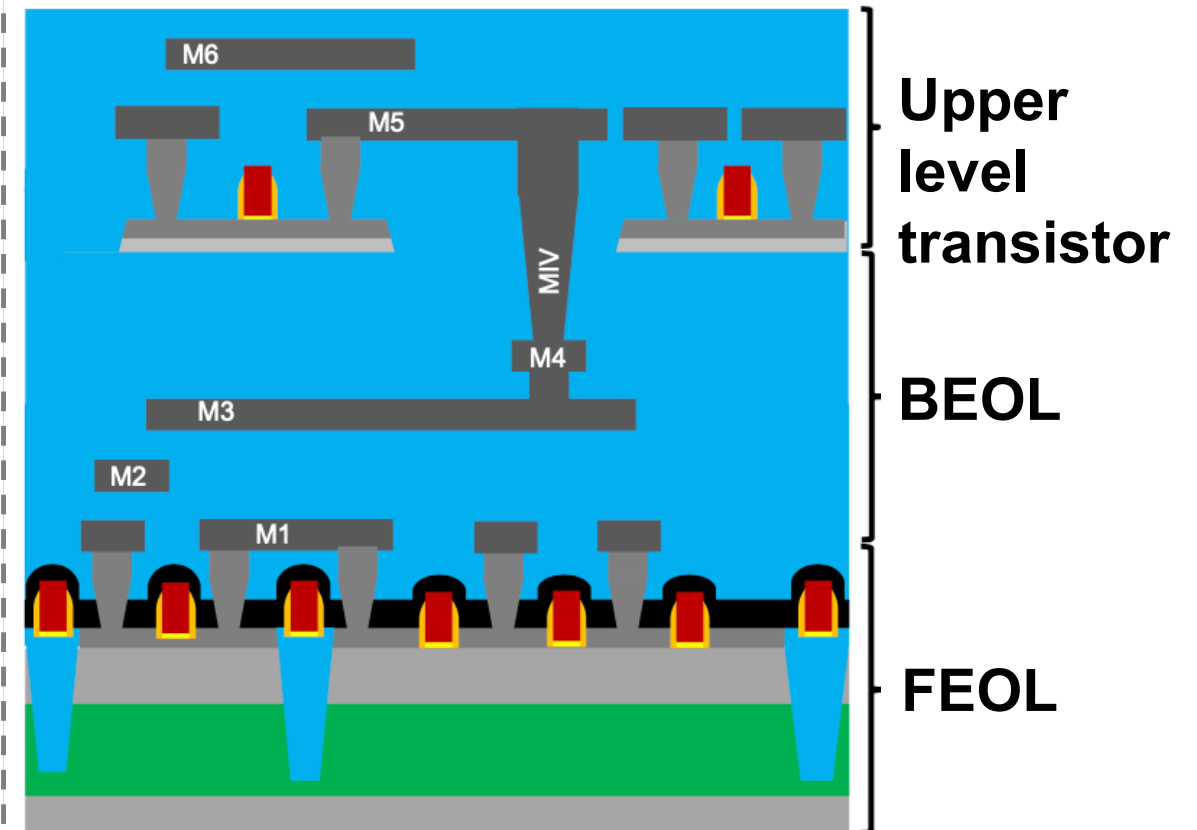
(I have a dream~)

System Integration/Scaling: H3D+M3D

Heterogeneous 3D(H3D)



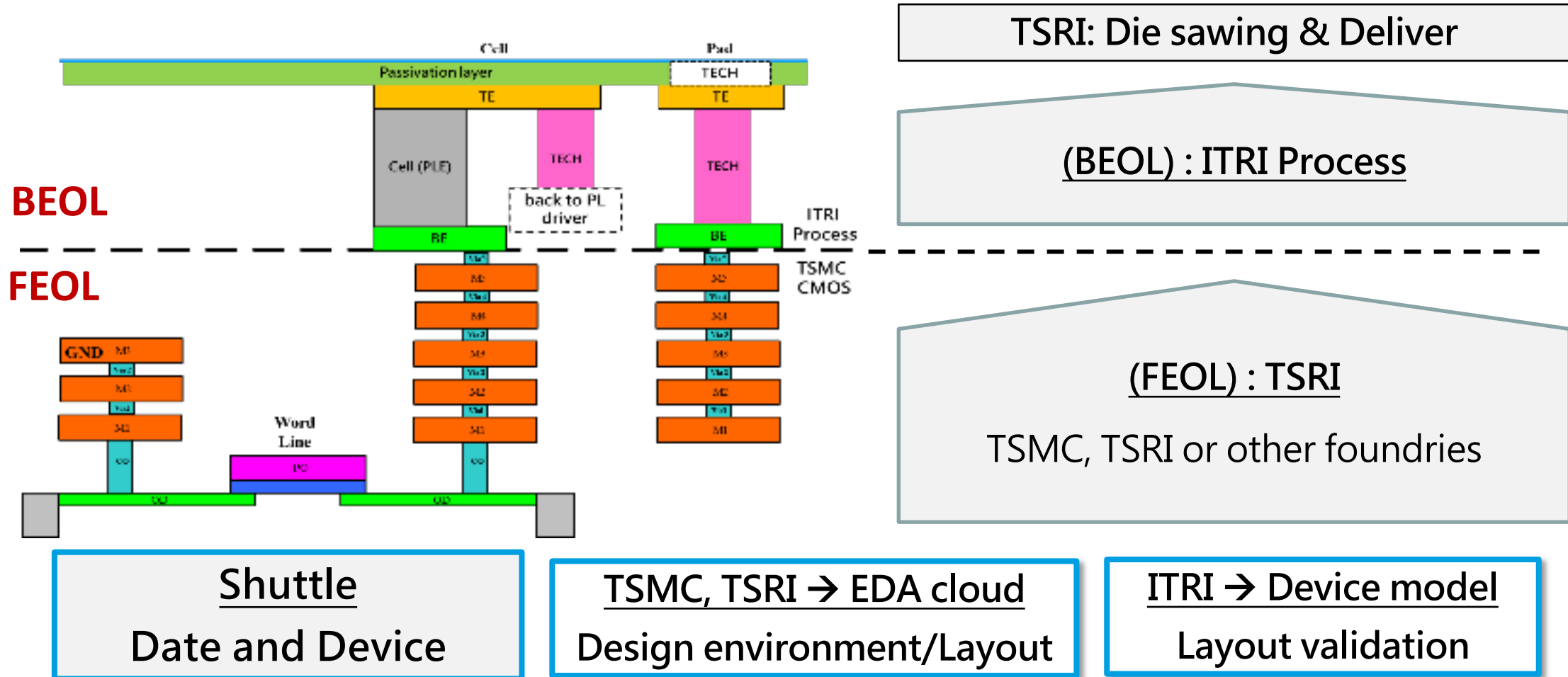
Monolithic 3D (M3D)



Taiwan Cbl: Semiconductor Device

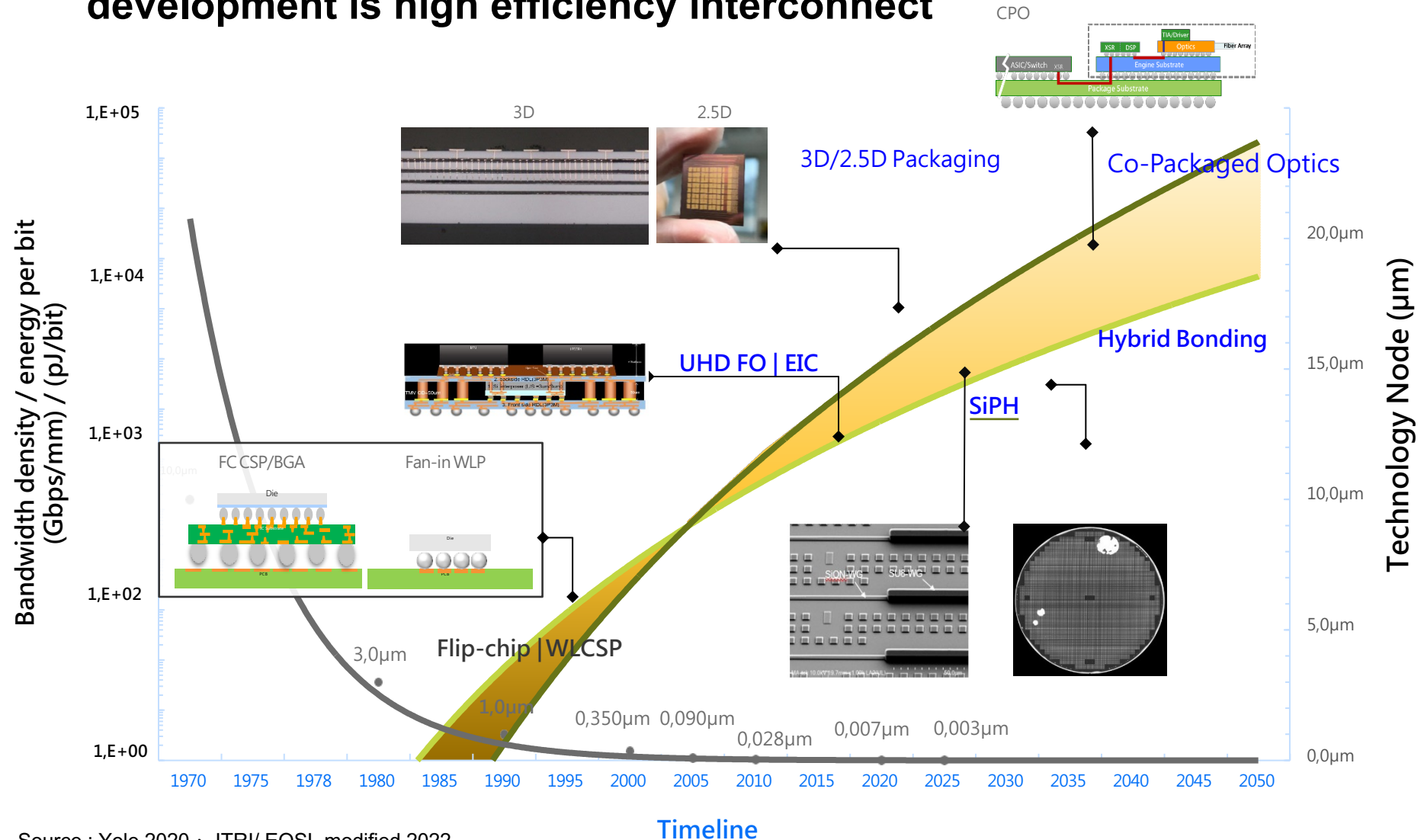
(Taiwan Chip-based Industrial Innovation Program)

BEOL: Collab with TSMC, TSRI & Universities



Semiconductor Packaging Trend

The technical trend of semiconductor high-bandwidth computing development is high efficiency interconnect

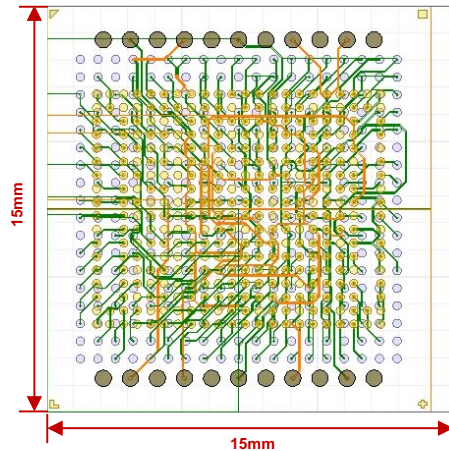


Source : Yole 2020 , ITRI/ EOSL modified 2022

Service Items

Design(Optional)

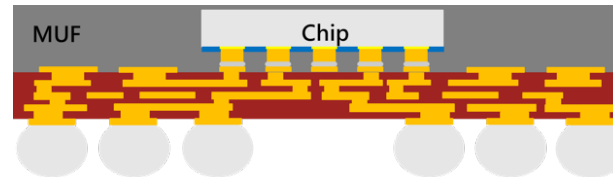
- Package RDL Design
- Multilayer Layout Design
- Simulation & Analysis



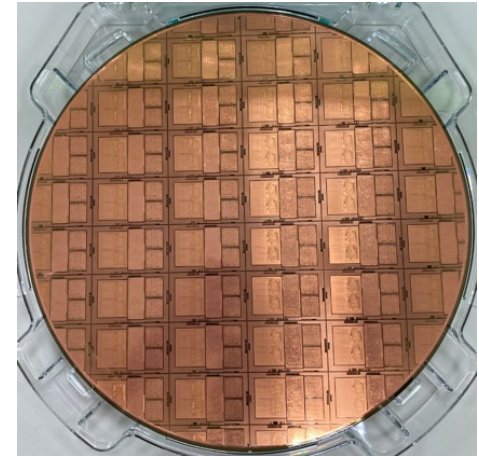
Fan-out RDL Shuttle Service

- Mask fabrication
- RDL process
- Bumping Process

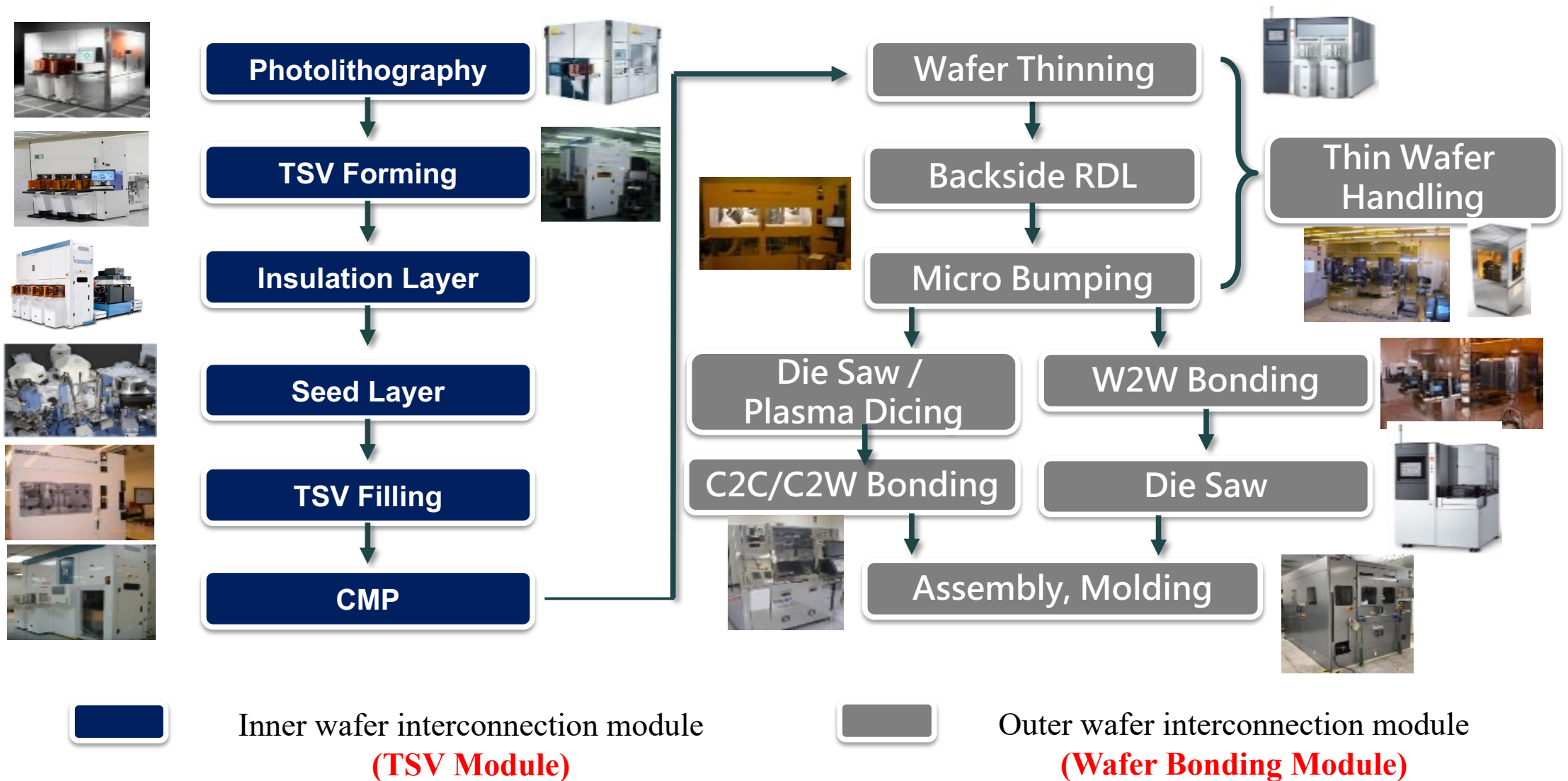
- Assembly
- Molding
- BGA



1. Structure: **5P5M**
2. Min. L / S : 5 μ m
3. Unit size : 7 mm*7 mm
(Max. available 50 mm*30 mm)



ITRI's 300mm 3DIC Platform



Achievement & Lessons Learned of Shuttle Service

Uniformity for RDL THK.

<10 % Design-in dummy bump/pad/metal

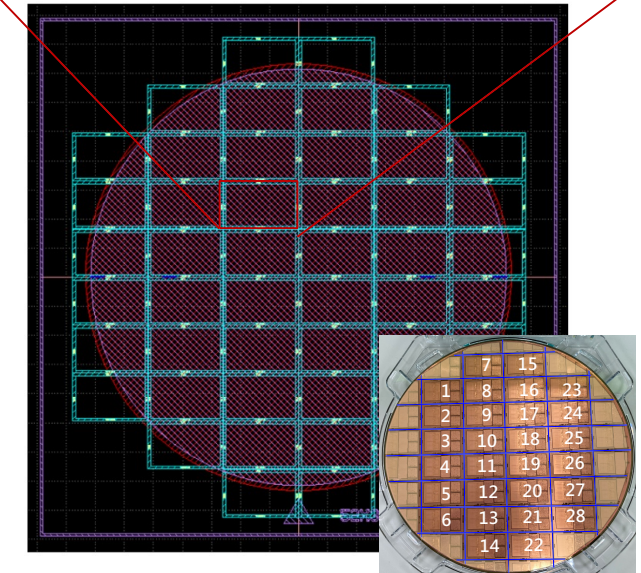
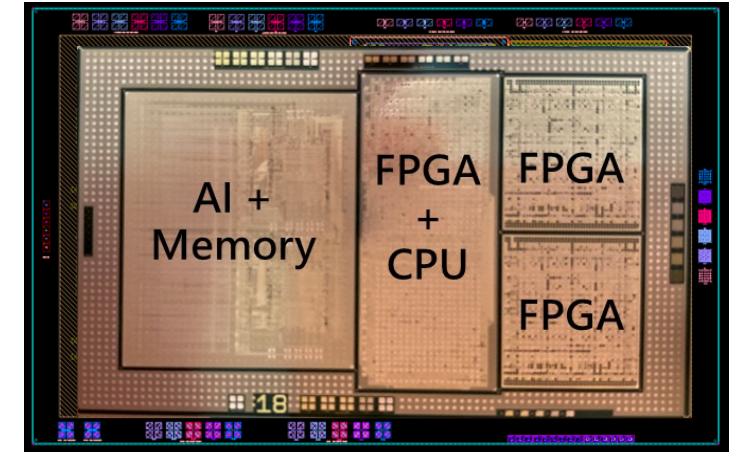
pA-level leakage control

Optimized integration/Ultra-Low Leakage(pA)

Yield control

Limited Design Rule (<30% to >80%)

- ✓ Different Cu trace density
- ✓ Fine Pitch L/S
- ✓ High Yield



A: FPGA AI Chip
B: Dual AI Accelerator
C: AI Accelerator+ 4 memories

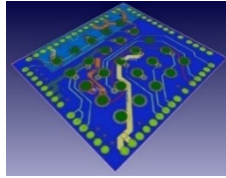
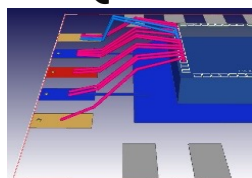
Design Template for Shuttle service

① RF chip+

WLP Fan-in package

QFN

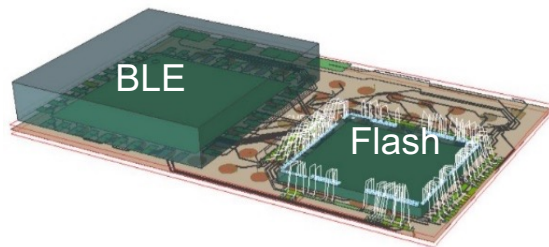
Fan-in



3x3 mm²

1.1x1.1 mm²

- Fan-in size 80% ↓

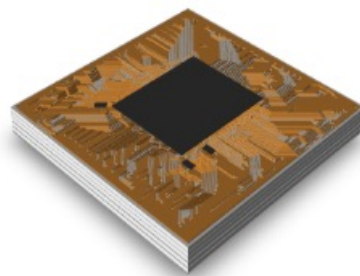
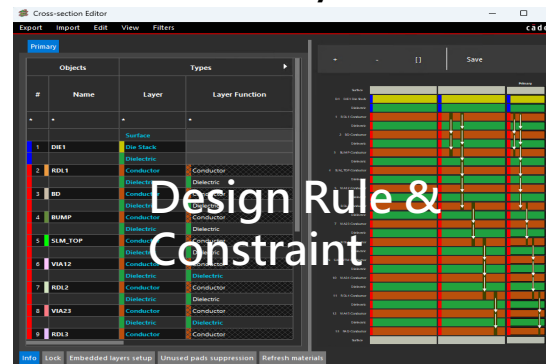


- BLE + Flash

BLE IO : 125; Flash IO : 11
FOP BGA : 56
Size : 7mm x 3mm

② SoC Process

5nm 1.6GHz /Dual Core



- 8K Vision SOC

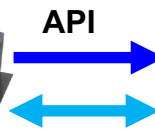
IO : 1,810
FOP BGA : 978,
Size: 16mm x 16mm

③ FPGA

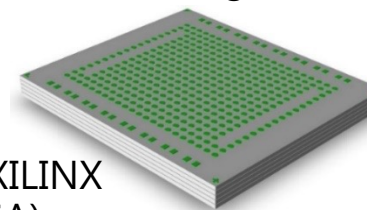
8K image Processing



Edge IPC



FPGA Fan-out Package



(AMD XILINX
FPGA)

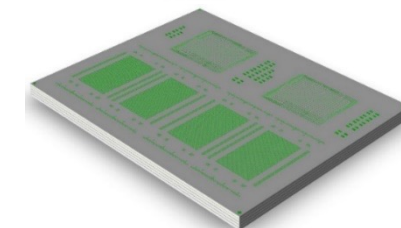
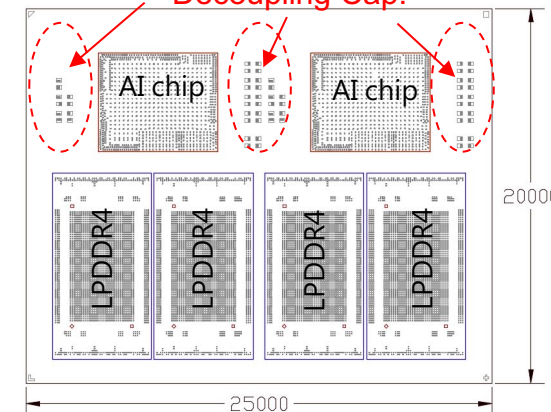
- FPGA

IO : 368
FOP BGA = 342
Size = 29mm x 12mm

④ <10TOPS AI

AI Chip*2ea+LPDDR4*4ea

Decoupling Cap.



- AI Vision Chip + Memory

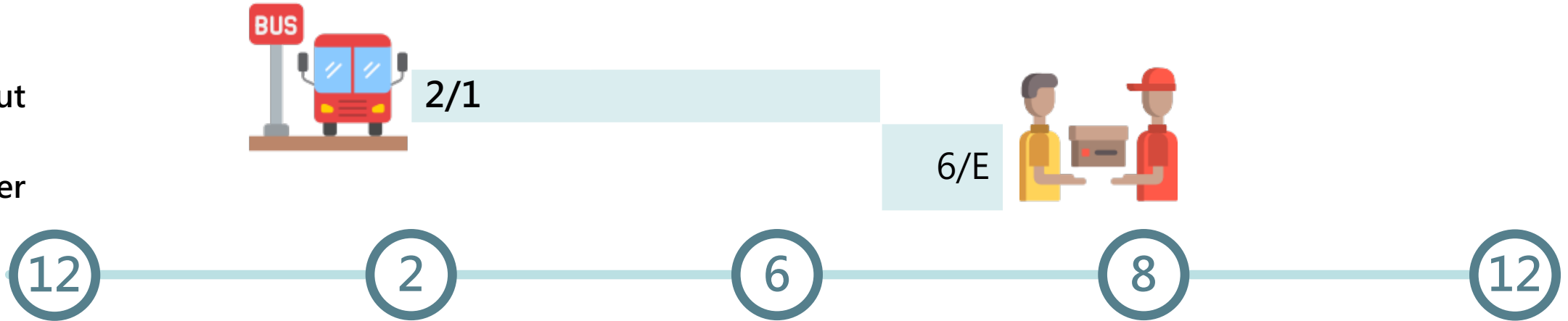
IO : ~2,474
FOP BGA : 1,240
Size: 25mm x 20mm

Schedule of Shuttle Service

5P6M · L/S 5um/5um
2/1 taped-out schedule

Taped-out

Deliver



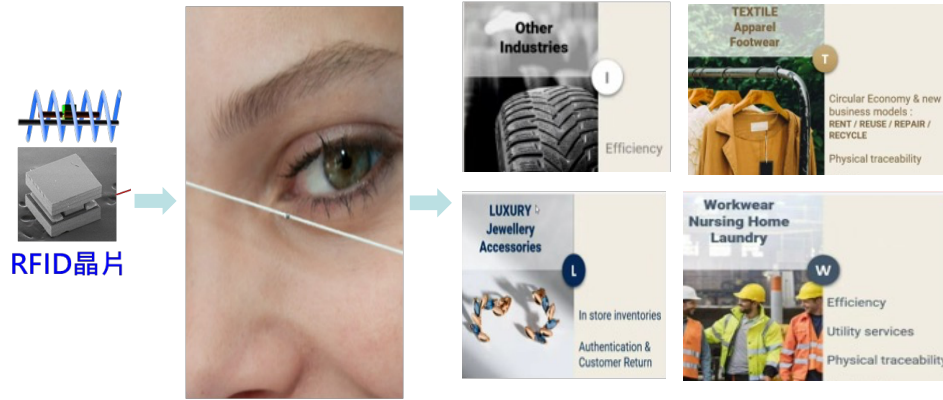
5P6M · L/S 5um/5um
8/1 taped-out schedule

Taped-out

Deliver

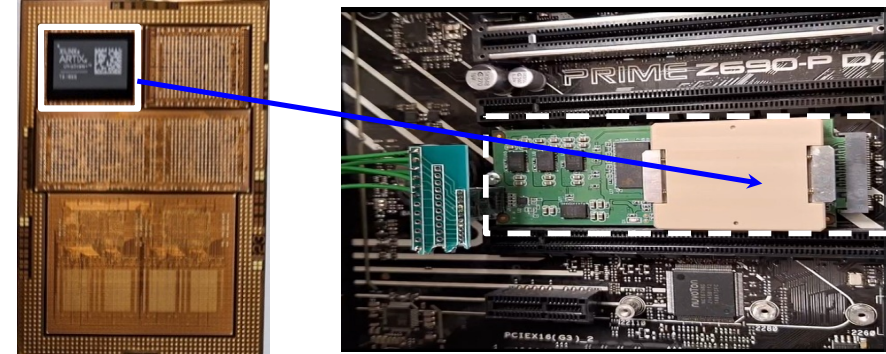


Empowering Startups and SMEs



Poou

- World' s Smallest RFID — Proven in Millions
- Trusted RFID for Luxury & Textiles
- Ultra-Mini RFID, Co-Validated with ASE

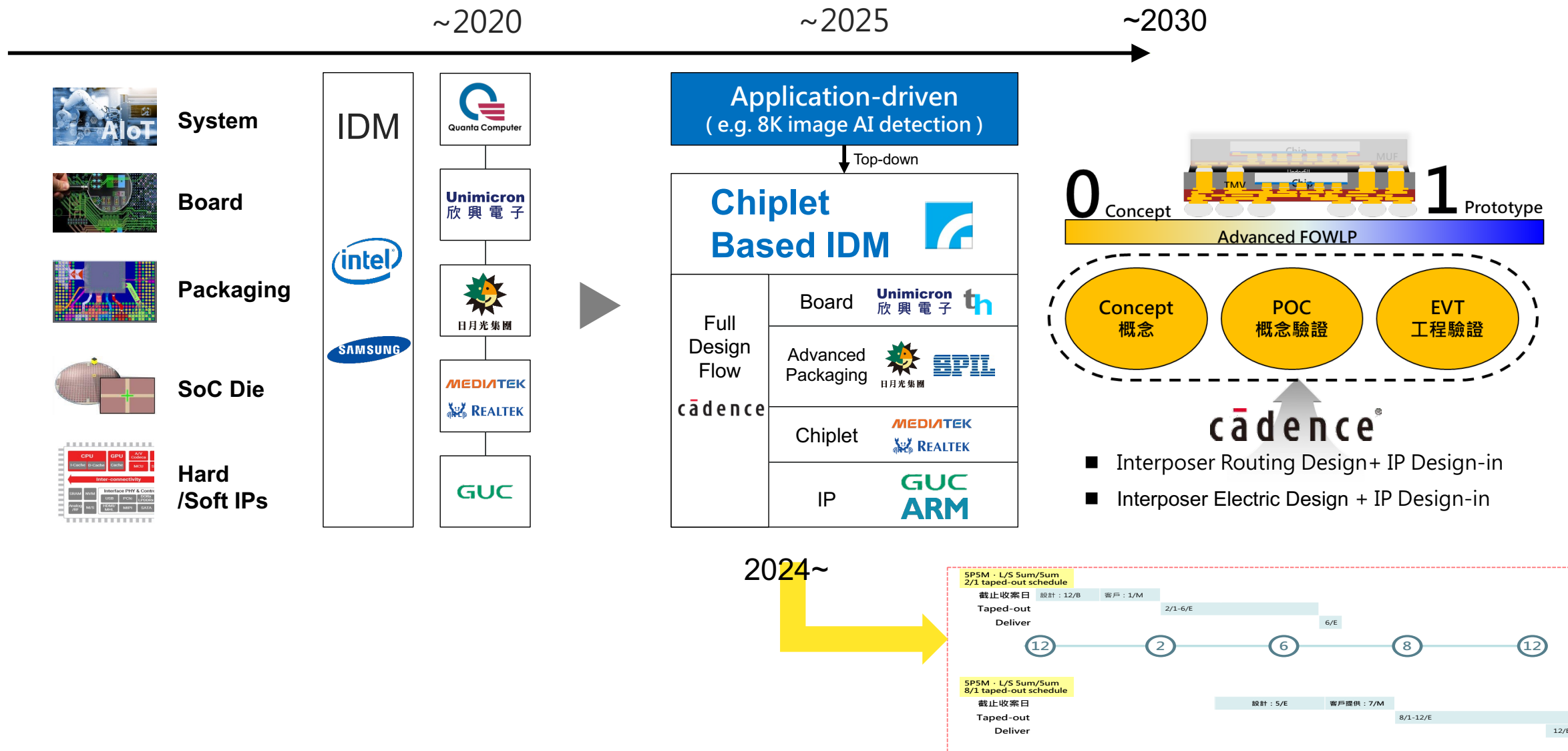


FPGA組裝後功能測試通過

Rockcore

- Seamless AI upgrade for existing IPC systems; in production.
- PCIe Gen3 (8 GT/s) compliant throughput.
- 80% footprint; 12× 8K video pre-processing.

Chiplet-Based Integration



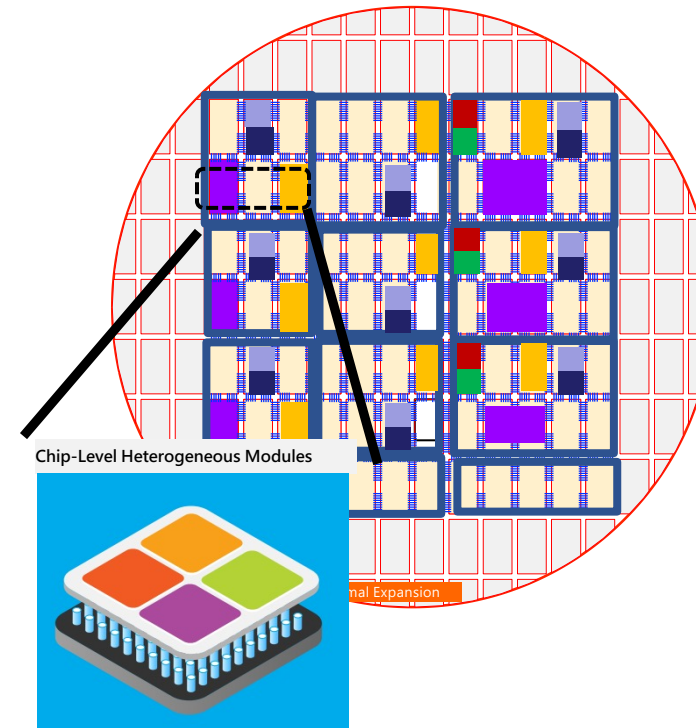
Chipselets Integration: Opportunities and Challenges

Design of connectivity

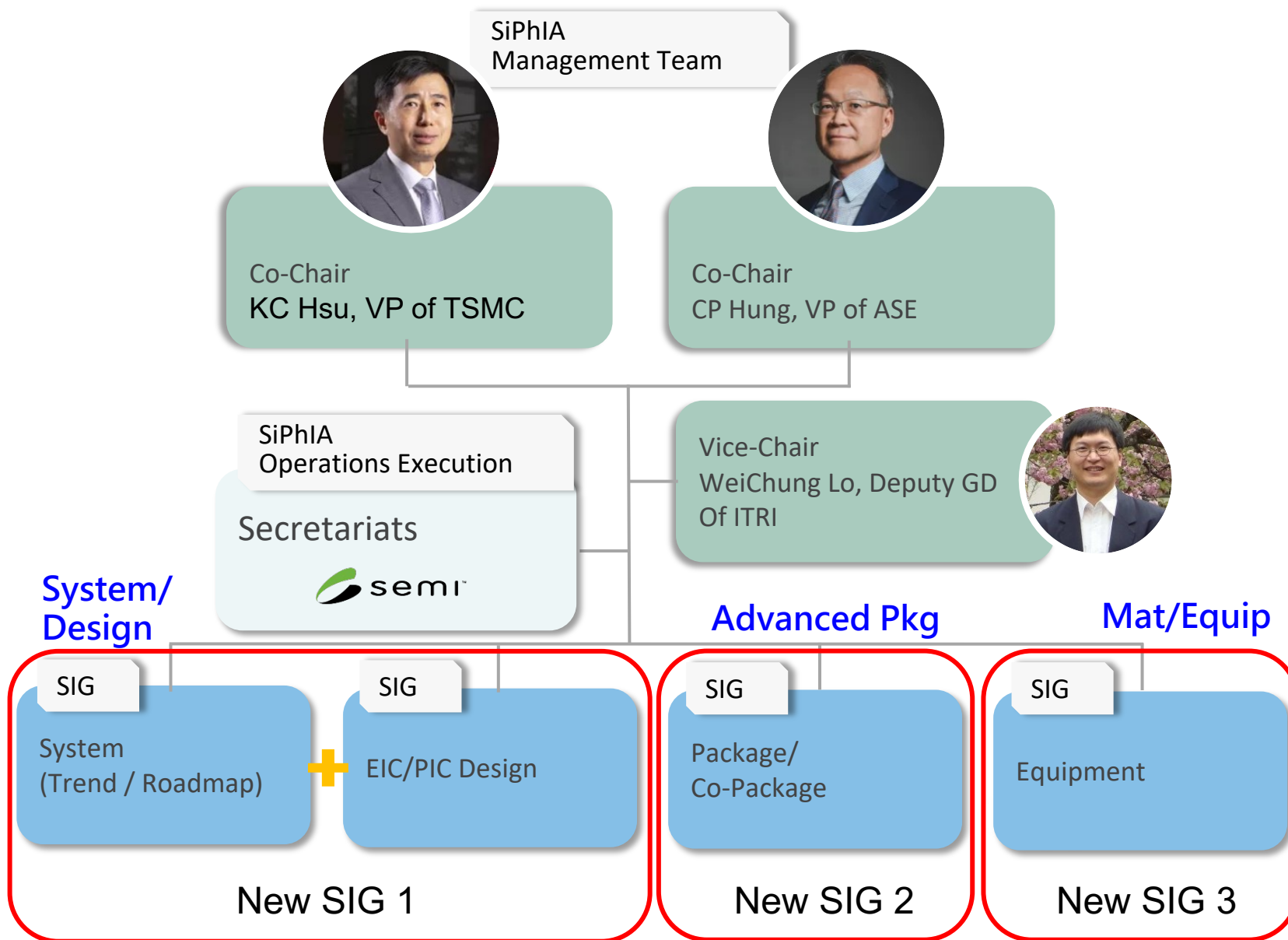
Chipselets integration

Integrated Yield

Reliability/Testing



SiGs of SEMI SiPhIA



SIG1 :

KC Hsu, VP of TSMC

CM Hung, Sen Director of MTK

SIG1: Aligning system companies on next-gen HPC, AI, and silicon photonics roadmaps. Advancing EDA tools, silicon photonics PDKs, high-speed modulators, and EPIC innovation.

SIG2 :

CP Hung, VP of ASE

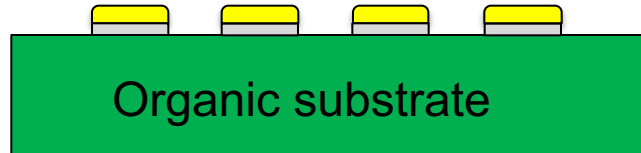
SIG2: Packaging/Co-Packaging: Driving 2D, 2.5D, and 3D heterogeneous integration and optical assembly for CPO applications.

SIG3 :

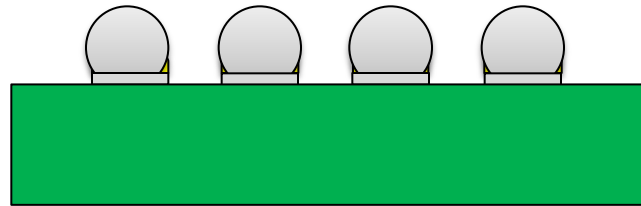
SIG3: Advancing automation in 2D/2.5D/3D processes, photonics assembly, and full-wafer optoelectronic inspection

3D SiPh Integration Platform

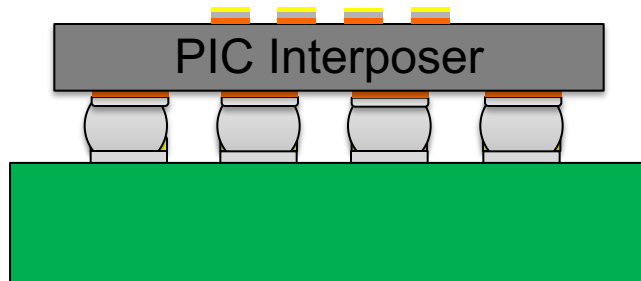
1. Organic substrate preparation with Ni/Au surface finish



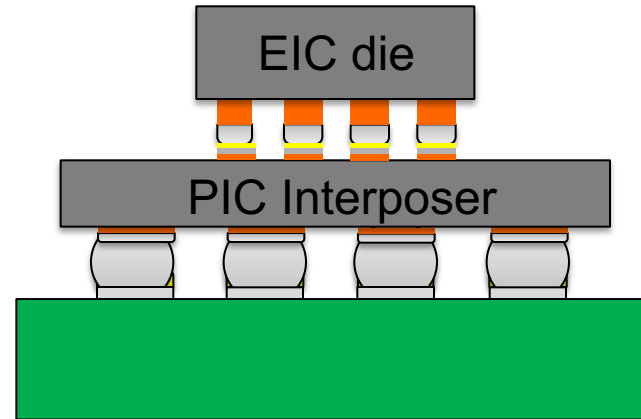
2. Solder ball attaching onto organic substrate and reflow



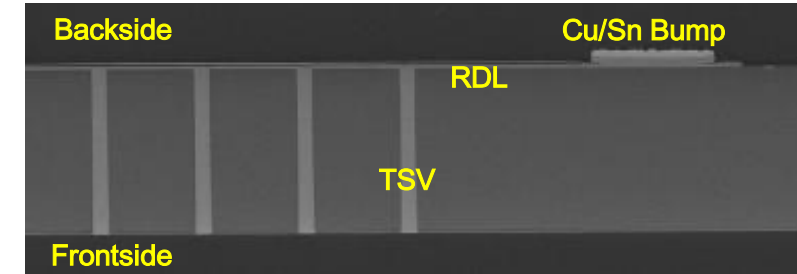
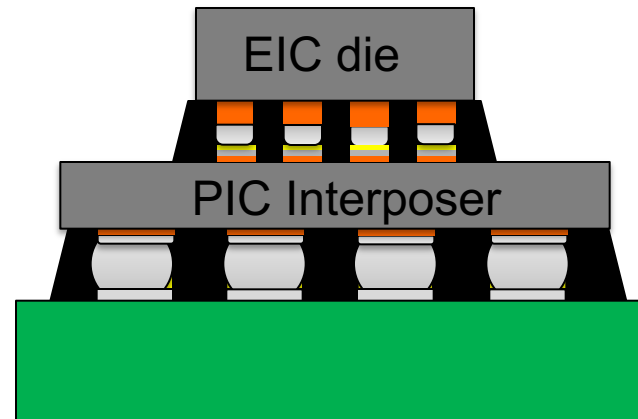
3. PIC Interposer attached on to organic substrate and reflow



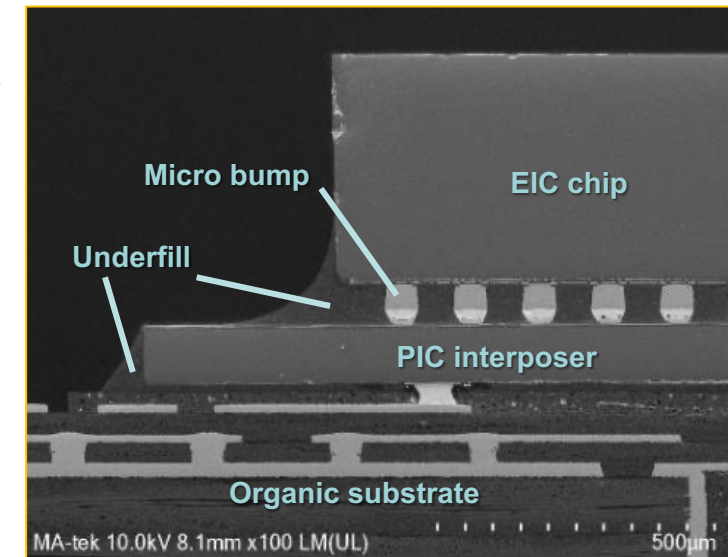
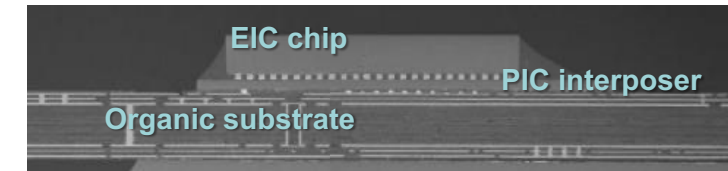
4. EIC die attached onto PIC Interposer and reflow



5. Underfill encapsulation and reflow



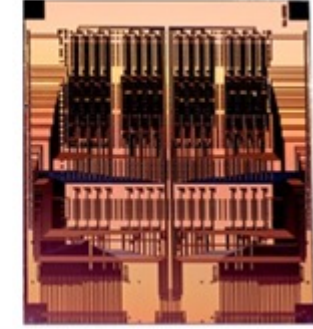
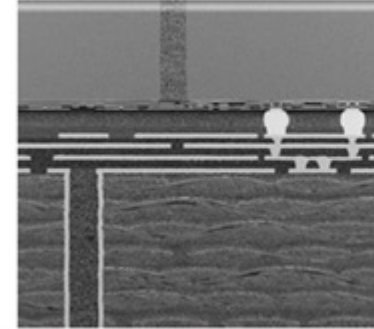
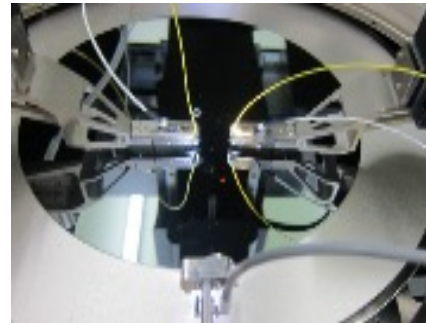
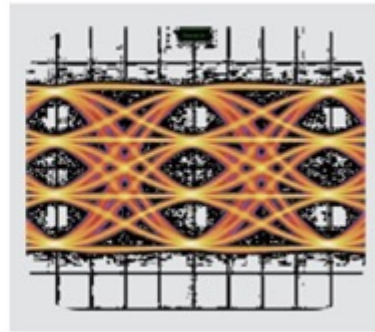
Overview



ITRI Core Competence on Silicon Photonics

■ ITRI/EOSL has the capability on **Advanced Packaging** and **Silicon Photonics**

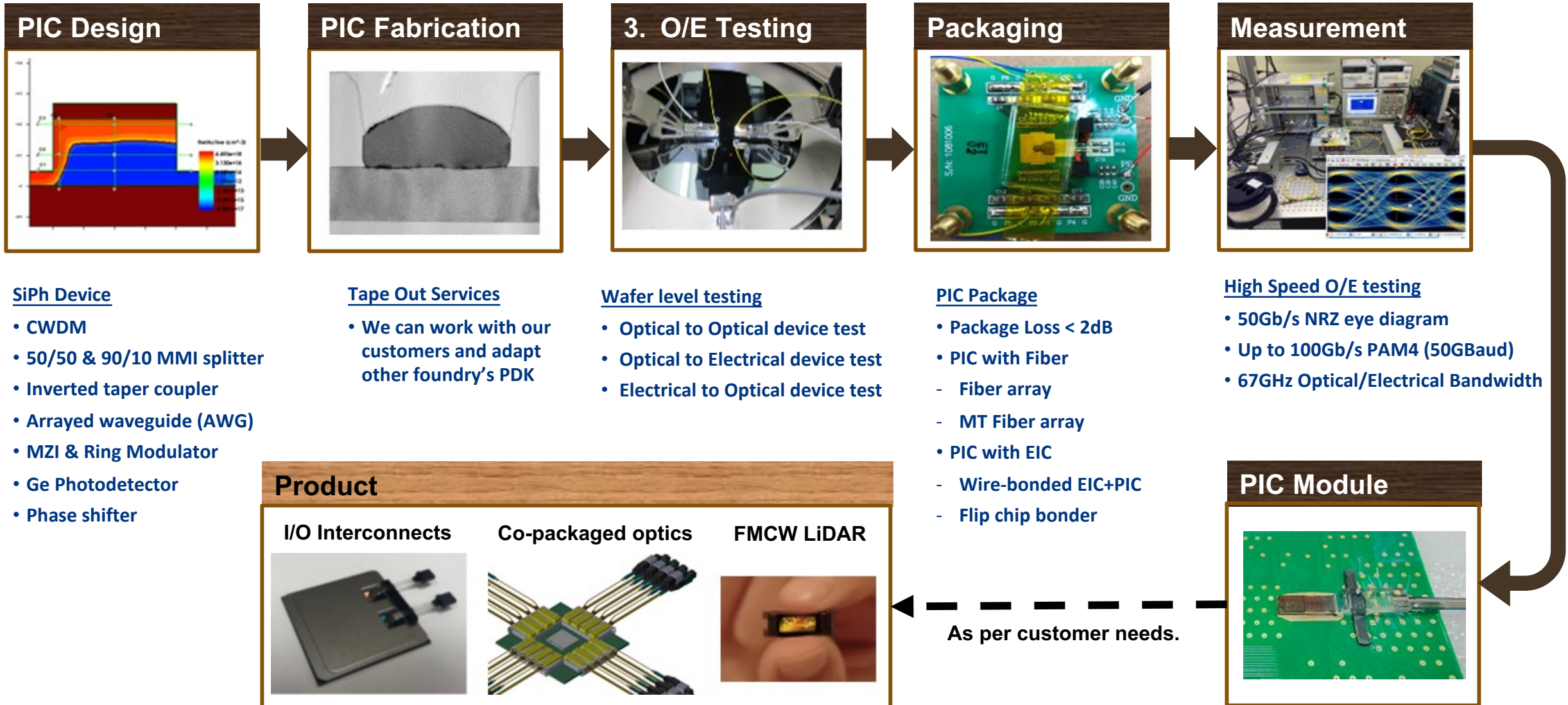
ITRI/EOSL



ASIC	Mixed Signal IC	Optical Devices & Fabs	Advanced Packaging & Test	Silicon Photonics
- Core switch, SerDes and DSP in leading node - Sustained generational differentiation	Power and performance optimized in both SiGe and CMOS	50M lasers/year from internal fabs - High-volume optical manufacturing - High-power, multi-wavelength sources	- Wafer-level test - TSV 2.5D/3D integration - Detachable optical connector	- High-density PIC design - Modulators and PDs in silicon - Low-loss SOI waveguides

Source: Broadcom

Silicon Photonic Platform



Summary

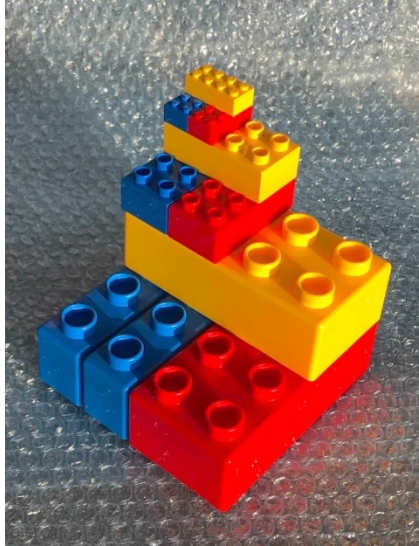
Chiplet-based System Integration (Material Innovation)

- **WL-Chiplets Integration Platform@ITRI:**
 - ✓ 12" WL-Chiplets integration: Full-flow design
 - ✓ LT bonding, EIC(bridging) and WL-chiplets shuttle service
- Monolithic 3D : ultra-low power/highest interconnect density
 - ✓ 3D-memory(NAND flash), emerging 3D-DRAMs/NVMs
 - ✓ A NEW “M3D-chiplet” integrated into 3d-integration platform

Future: Monolithic 3D + Heterogenous 3D

(Europractice, OpenHW, Google Silicon,...)

Thank You!



Any Questions?

New Wave of Semiconductor
for Microelectronics and Smart Systems