工業技術研究院
Industrial Technology
Research Institute

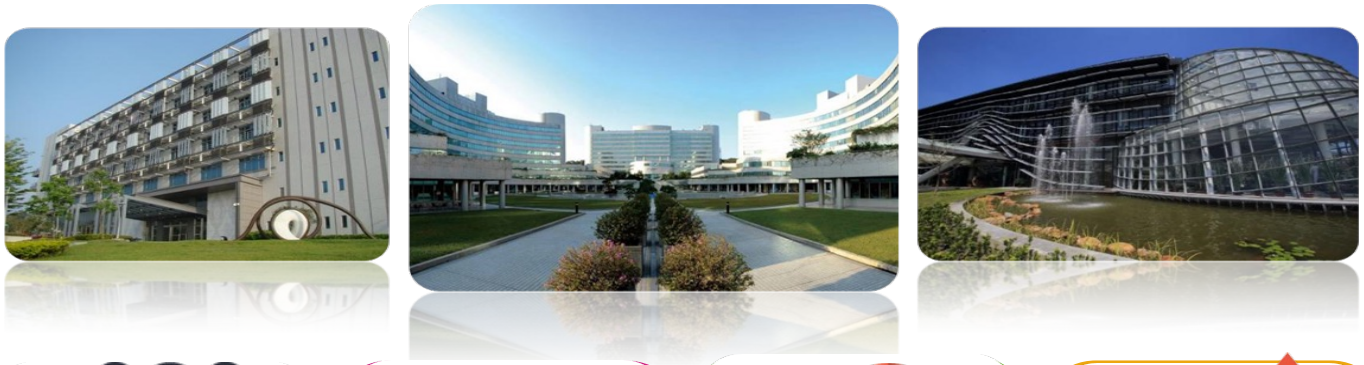
AI on Chip Technology and Heterogeneous Integration

Wei-Chung Lo,

Deputy General Director/Senior Principal Engineer
Electronic and Optoelectronic System Research Laboratories (EOSL) of ITRI

Feb. 24th, 2023@HIR, 6th annual conference

ITRI Overview



Total Staff

5,952

Ph.D	1,288
Master	3,637
Bachelor	1,027
Alumni	27,052

Total Patents

31,382

Startups(~2022/Q2)

153

Incubatees(~2022/Q2)

210

Industry Service (~2021)

Provided Services

18,392

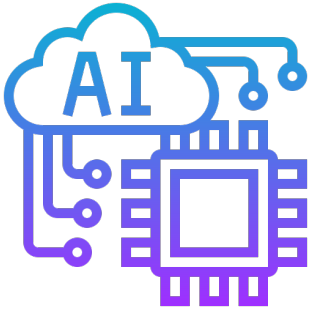
Transferred Technologies

672

The Trend of Chip Technology

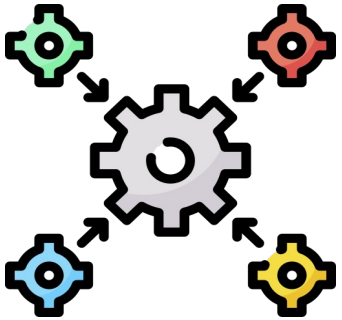
B5G/6G

High Freq Device
(>100GHz
GaN PA)



AI

High speed/AI Chip
(Heterogeneous
Integration)



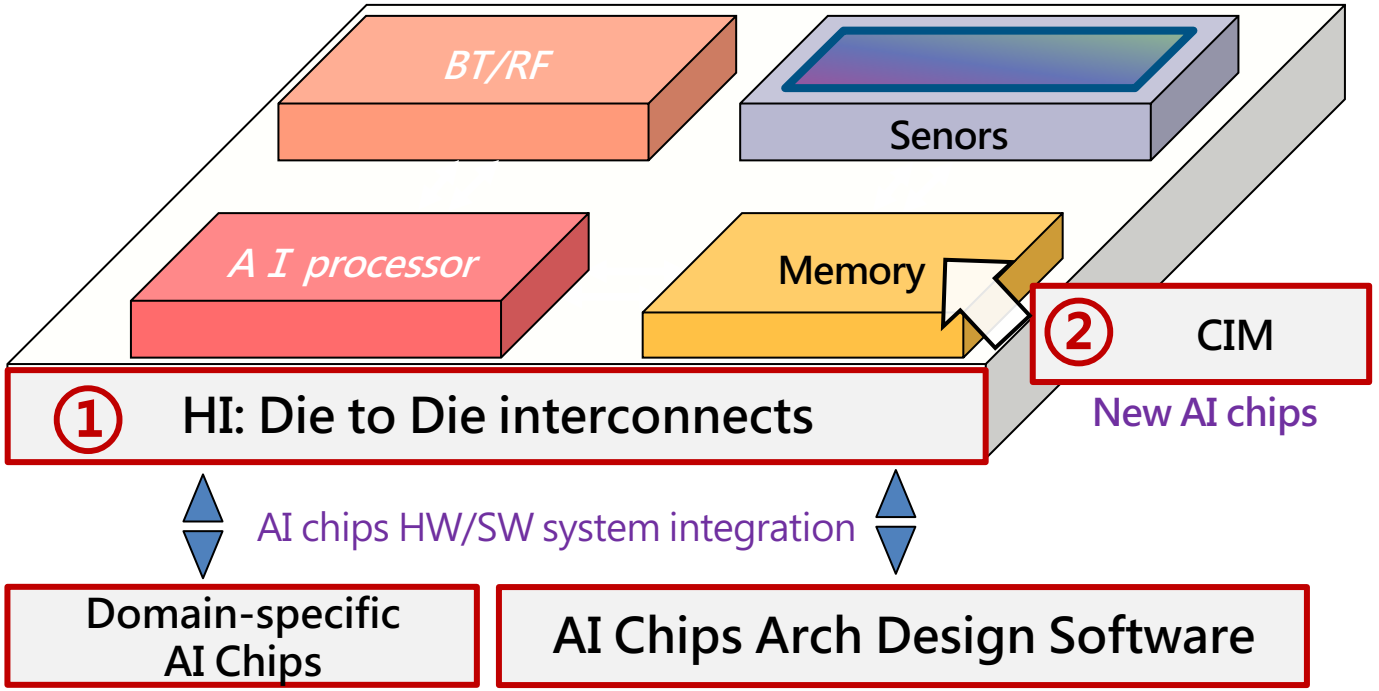
Power

High Power/WBG device
(1.7kV/3.3kV
SiC module)

- Near Memory
  - CoW, WoW, SoIC, ...
  - HPC: thermal solution, ...
- Computing in Memory
  - Innovative Architecture, New non-Volatile Memory(MRAM)
  - Low power(KWS): CiM (ISSCC 4 year in a row)
- Platform of HI
  - Programmable pkg
  - Wafer-level Chiplets integration Shuttle service

AI on Chip @ITRI

AIoT : Highly integrated, Innovative Architecture & Ultra-low Power



Near Memory

Categories of Heterogeneous Integration

- on Silicon Substrates (TSV Interposers) ➔ 2.5D
- on Silicon Substrate (TSV-less Interposers) ➔ Si bridge
- on Fan-Out RDL-Substrates ➔ FanOut/Substrate-less
- on Organic Substrates ➔ Flip Chip, 2.1D

Category-I: 2.5D w TSV

Nvidia Tesla P100 Hardware Architecture

- One 8 Gb HBM2 die contains over 5,000 through-silicon via holes
- The combination of HBM2 stack, GPU die, and Silicon interposer are packaged in a single 55mm x 55mm BGA package
- Tesla P100 accelerators will ship with four 4-die HBM2 stacks, for a total of 16 GB of HBM2 memory.

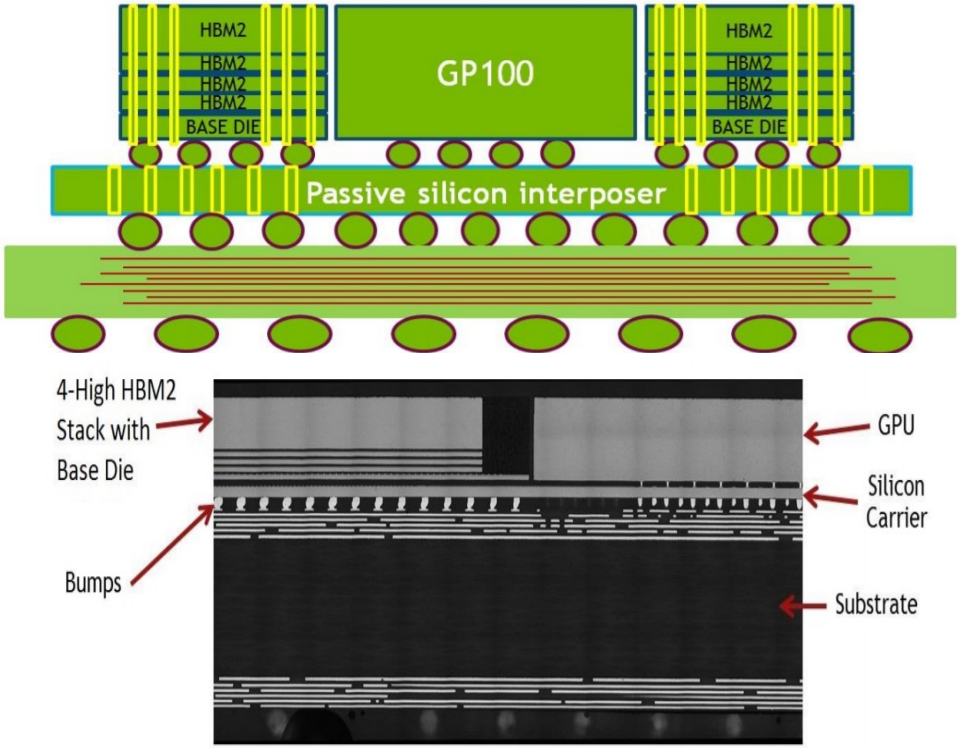
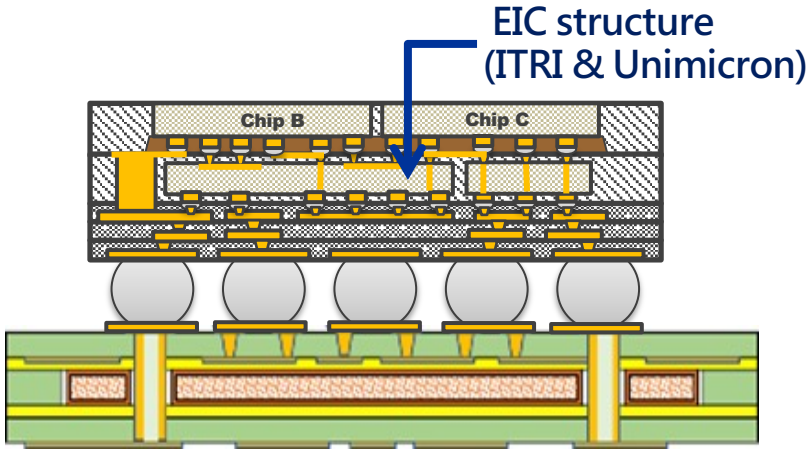


Image Source: Nvidia 2016

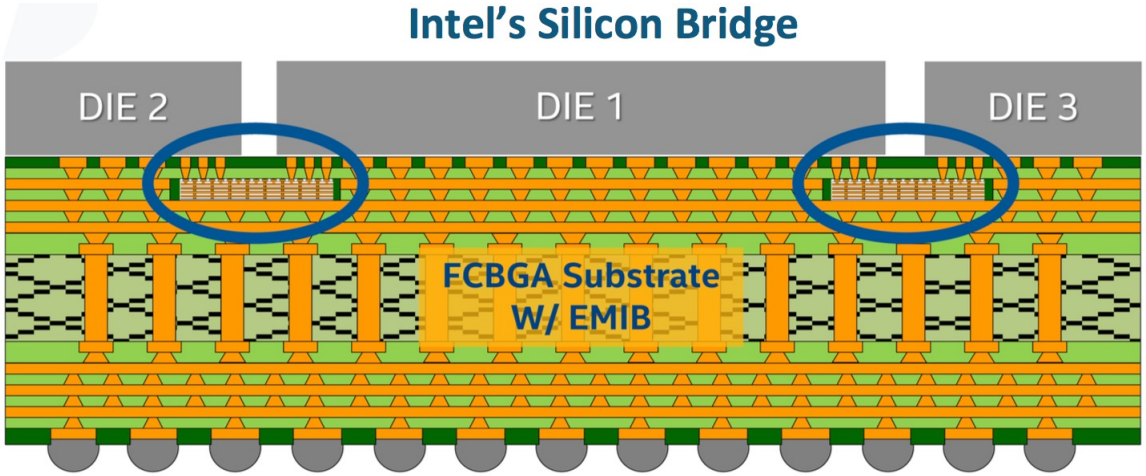
Category-II: Si Bridge

- Interposer embedded into the substrate (or package): to serve as a bridge between two chips
- Provide flexible architecture for interconnection
- Compatible to fan-out process
- Achieve close to 3D IC performance with lower cost



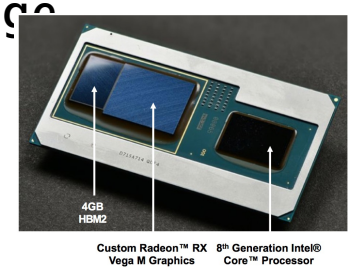
EIC–Embedded Interposer Carrier

Category-II: Si Bridge



Embedded Multi-die Interconnect Bridge (EMIB)

- Intel's CPU (Kaby Lake) and AMD's GPU (Radeon)
- Intel/AMD/Hynix Heterogeneous Integration using Intel's EMIB

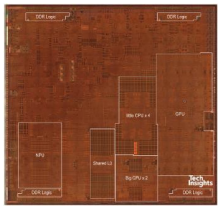


Source: Semi Taiwan 2018, John Lau

Category-III: FanOut/Substrate-less

Package on Package (PoP) for the Mobile DRAMs & Application Processor (AP) of iPhone XS

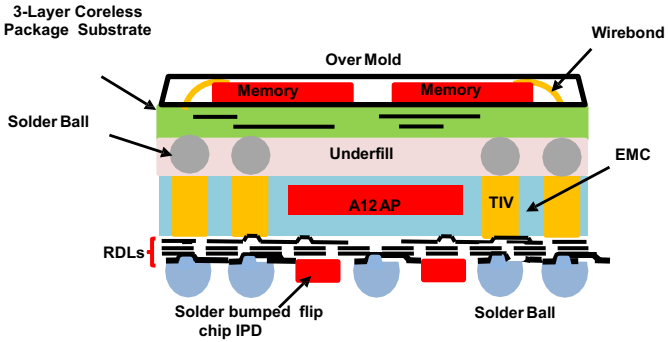
A12 AP SoC



9.9mm x 8.4mm

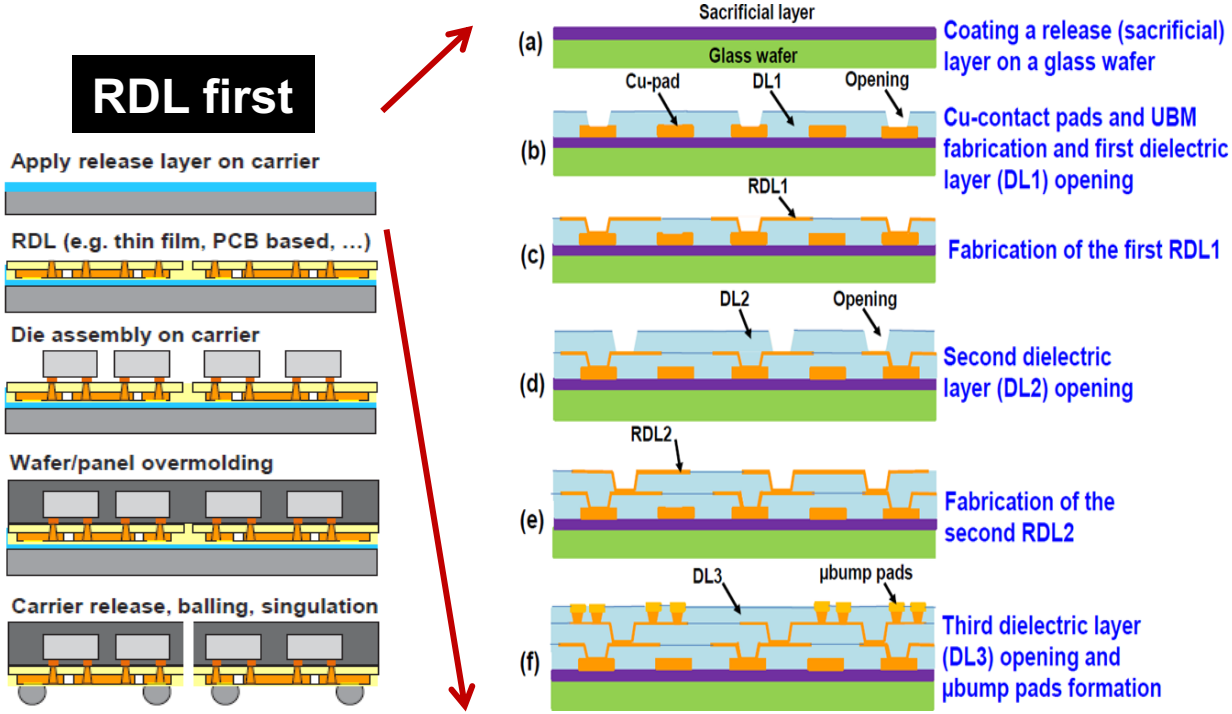
PoP

13.4mm x 14.4mm x 0.815mm



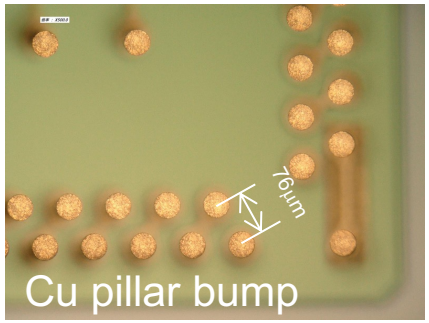
Fan-Out Package Type _ Chip First

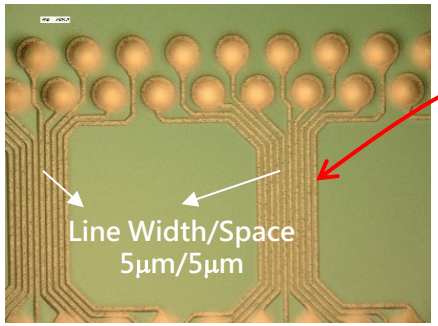
Category-IV: Flip Chip, 2.1D



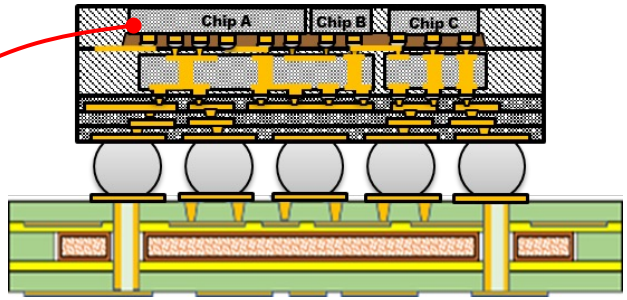
AI Chip System Integration

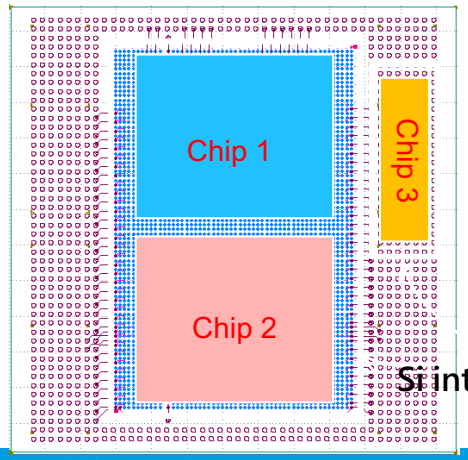
HPC/ AIoT	Heterogeneous Integration: Chiplets & Flexibility • Modulized: interposer layers≧ 2 • System scaling: I/O pitch 40~80μm • Multi-chip pkg ≧ 3 chips with AI chip





2.5D stacking with EIC package module





Si interposer

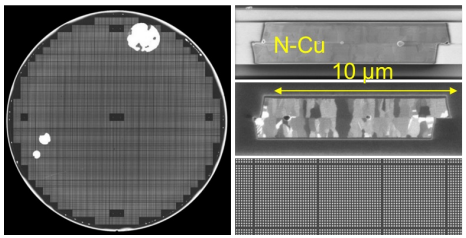
- Si interposer
- Density: 5μm / 5μm
- Challenges: warpage and coplanarity control

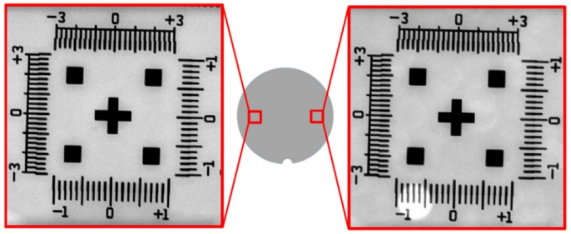
Low Temp Bonding_1/2

Hybrid Bond with Nt-Cu
EP/sputter
stable

Alignment accuracy
≦ ± 0.2 μm
(with new EVG Bonder)

- Bonding temperature：350°C→200°C
- Electro migration Lifetime：3X (compared with standard copper)
- Specific contact resistance：1.2×10^{-9} Ω·cm²
- Bonding Area > 90 %

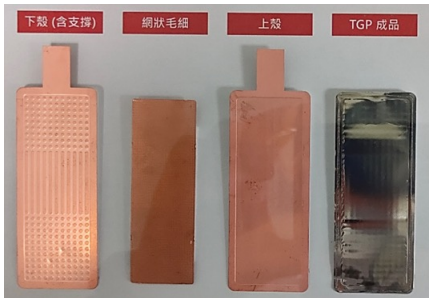


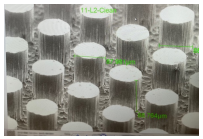


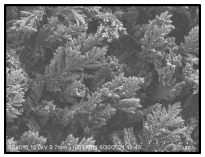
Left & Right Post-bond Alignment (um) w/ offset			
Left X	Left Y	Right X	Right Y
0.2	0.4	0.1	0.0

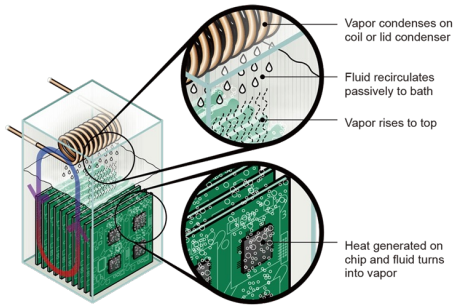
Achievements： thermal solutions

< 10W	0 - 300W	300W – 2000W
Low power thermal solutions，for mobile equipment	Medium power，for ICT equipment	Ultra-high power thermal solutions, for HPC system
● Ultra-thin vapor chamber (UTVC)，t <0.35mm	Mature technologies	● VC Lid，immersion cooling，P>500W







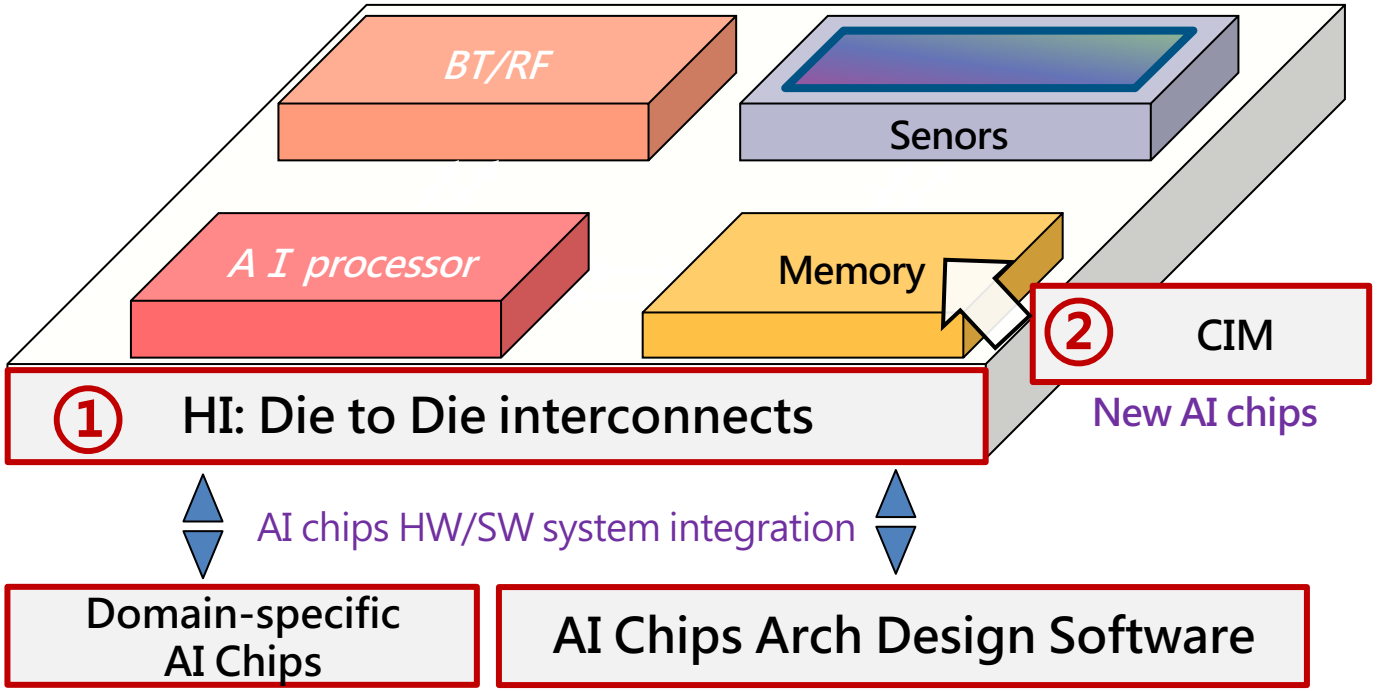


Developing UTVC and high performance VC lid ,

also extend to integrate VC in liquid cooling for HPC high heat dissipation solutions

AI on Chip @ITRI

AIoT : Highly integrated, Innovative Architecture & Ultra-low Power

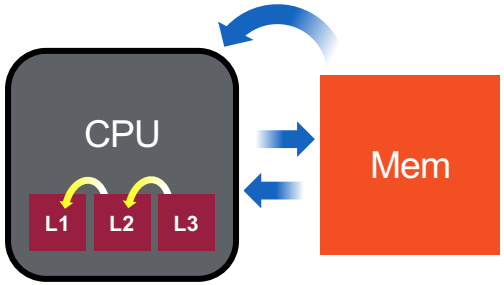


- Near Memory
  - CoW, WoW, SoIC, ...
  - HPC: thermal solution, ...
- **Computing in Memory**
  - **Innovative Architecture, New non-Volatile Memory(MRAM)**
  - **Low power(KWS): CiM (ISSCC 4 year in a row)**
- Platform of HI
  - Programmable pkg
  - Wafer-level Chiplets integration Shuttle service

In-Memory Computing

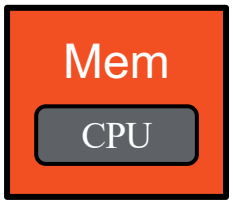
Up to 10x~100x Energy Efficiency

AS-IS



- Data movement
- Memory bandwidth
- MAC operation for AI neural network

To-Be



In-Memory-Computing

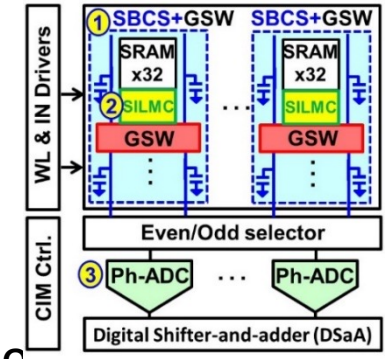
Computing-in-Memory (CIM)

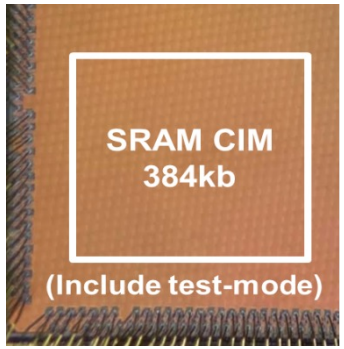
Worldwide leading 8 bits precision CIM w 20TOPs/W

Low power charge sharing R/W architecture:

ISSCC 2021

- energy efficiency: 20TOPs/W @ 8b design
- achievement: 90TOPs/W @ 4b





Prioritized-hybrid-ADC (Ph-ADC)

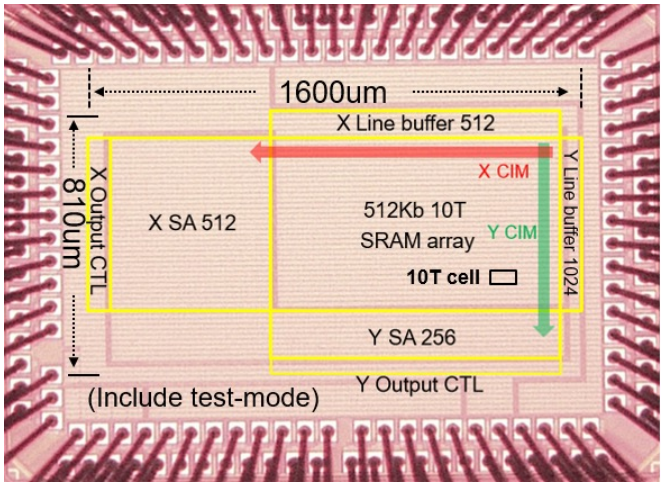
- Energy consumption: 1.44~2.63x reduction
- small area and power overhead for analog readout

	tsmc ISSCC2020	NTHU ISSCC2020	ITRI ISSCC2020	★ ITRI ISSCC2021
Tech	7nm	28nm	28nm	**28nm**
Precision	4/4/4	4/4/12	4/4/12	**4/4/12**
TOPs/W	351	47.85	30.40	**60.28**
FoM	733.9	462.8	258.2	**3585.7**

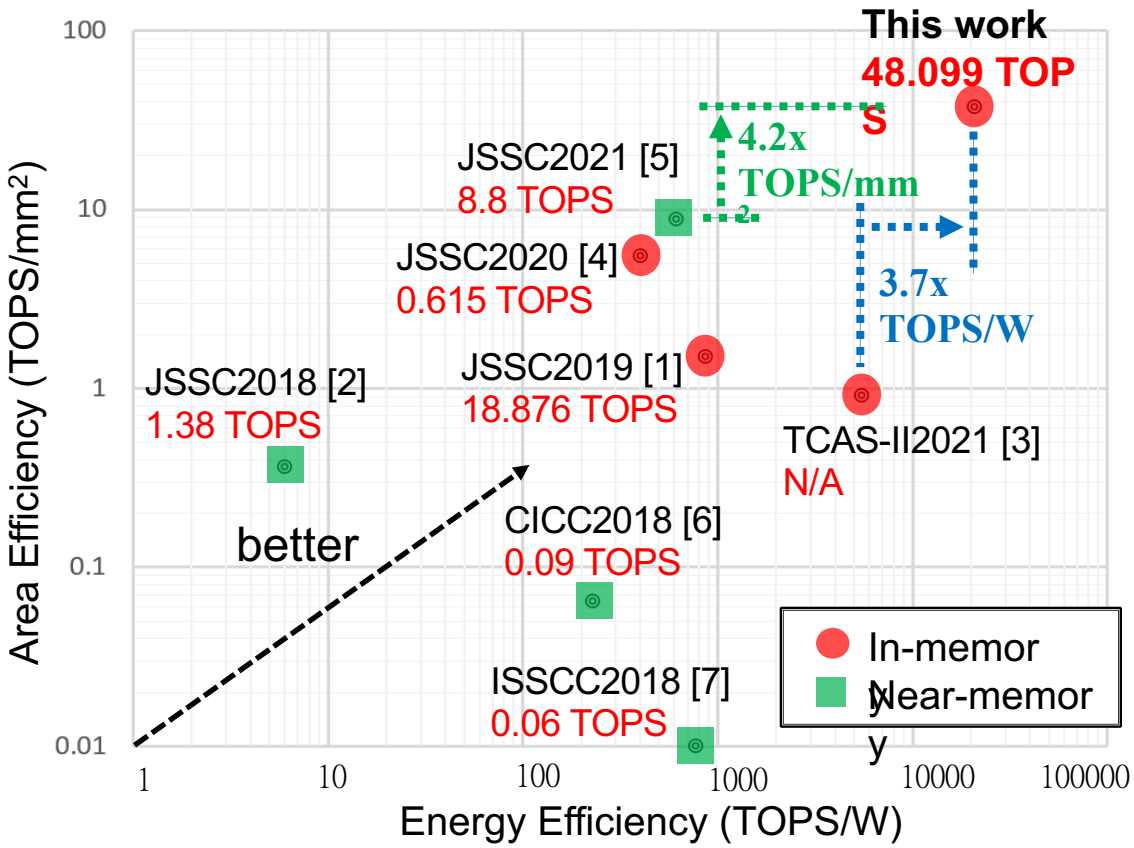
*: Input/ Weight/ Output (bit)

A 48 TOPS and 20943 TOPS/W 512kb Computation-in-SRAM Macro for Highly Reconfigurable Ternary CNN Acceleration

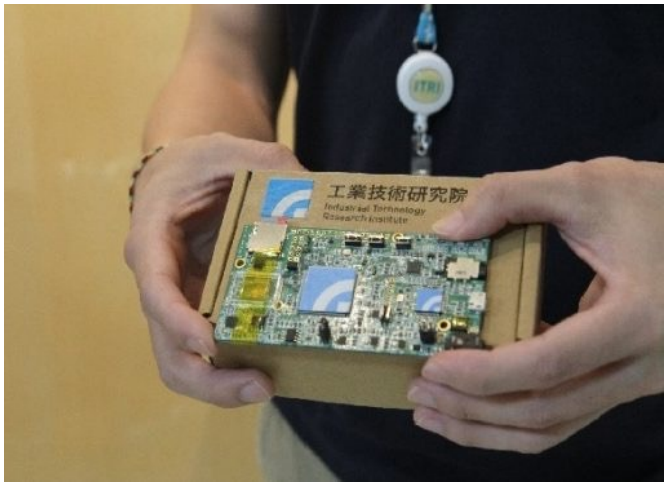
28nm 10T SRAM
512kb IMC macro
20943 TOPS/W
37.1 TOSP/mm^2



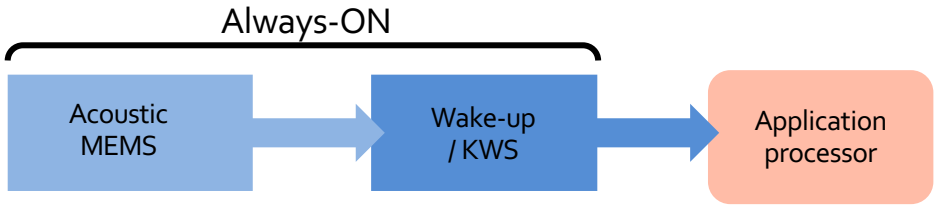
J.-S. Lin et al., ASSCC 2021



Always-On Key-Word Spotting (KWS)



KWS demo system: SRAM-IMC testchip + MCU host

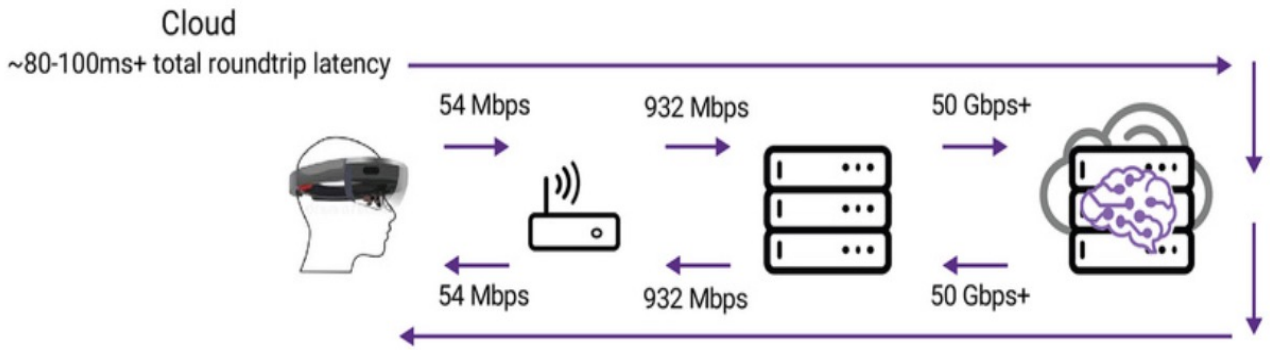


KWS Data set	Number of class	Memory (kb)	accuracy
GSCD	12	652	94.02
GSCD	36	804	88.5

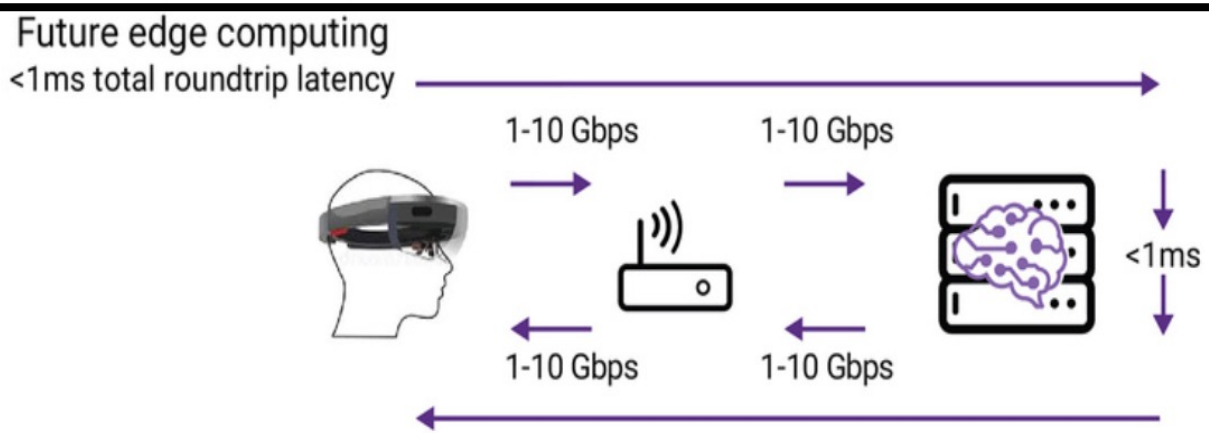
Source:TH, Hou. NYCU

Edge Computing Benefits the Future

Current:
m-sec



Future:
μ–sec



Source: https://www.synopsys.com/designware-ip/technical-bulletin/ai-edge-computing-5g-iot.html

SOT-MRAM(1/5)

Faster speed & longer duration of ITRI SOT-MRAM

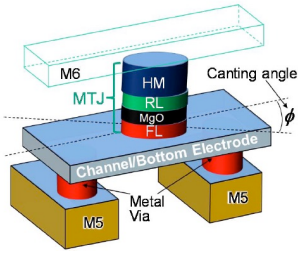
Current

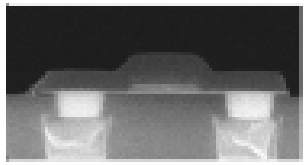
SOT-MRAM

Single device

►Writing: 10 ns

►Duration: 0.1B cycles





Source: IEDM2019

State-of-Art: ITRI

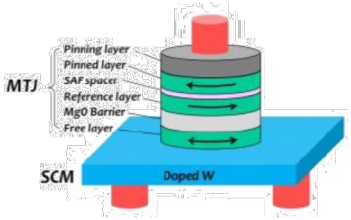
SOT-MRAM@ITRI

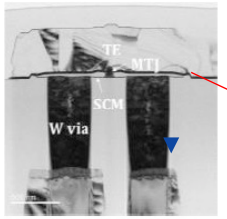
Symposium on VLSI 2022

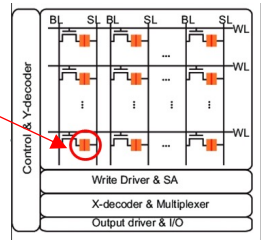
8Kb Memory Area

High speed：10 ns→1 ns

Duration：0.1B→7000 B







Embedded memory area

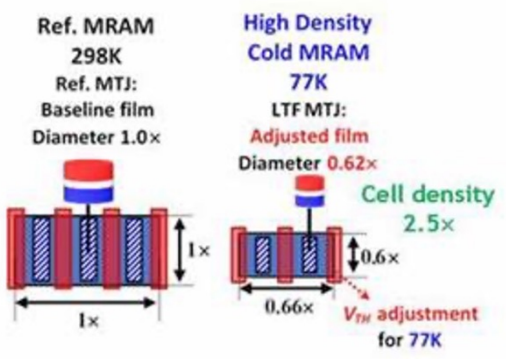
STT-MRAM(2/5)

Lowest temp(4K)、high duration STT-MRAM

Current

MRAM

►Operation: 77K
►Temp range: 77K~300K
►Duration: 1E9



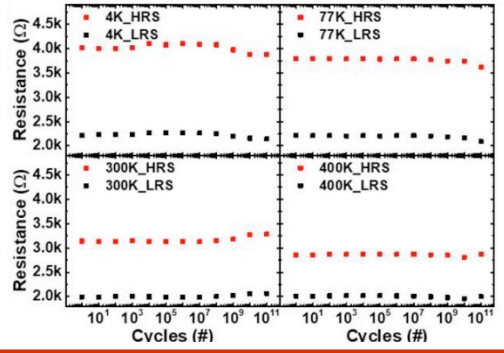
Source: Symp.VLSI2021

State-of-art: ITRI

Wide range MRAM

Symposium on VLSI 2022

- CoFeB / Mg / CoFeB free layer，working temp range：4K~400K
- High reliable and writing (1E11 duration)



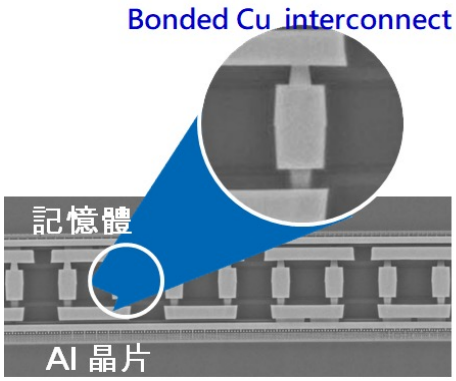
Full temp range (4K~400K) STT MRAM

Wafer Bonding(3/5)

High density hybrid bonding: Memory & CMOS Image Sensor

Wafer-level bonding

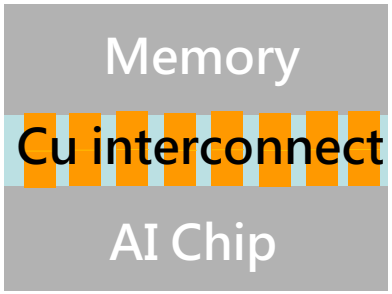
Worldwide



I/O >1,600
Bonding
350℃

- Application: CIS & memory
- Existing 350 ℃
- I/O>1,600

Chiplets integration



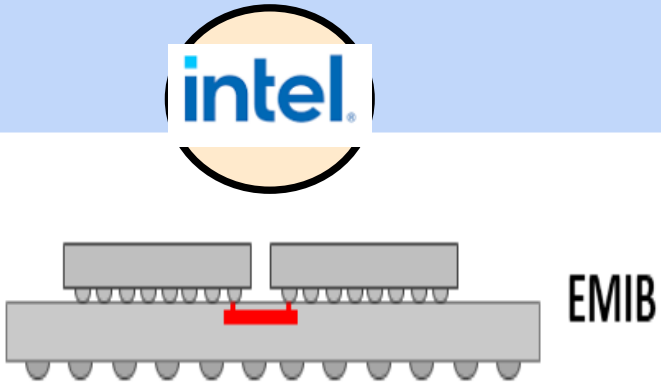
I/O >15,000
Bonding
temp. 180℃

- Memory or Sensor <350℃ bonding
  - **Bonding temp.：350℃→180℃**
- **I/O：>1,600→>15,000**
- α –SITE with industrial partner

Embedded Interposer Carrier (EIC) (4/5)

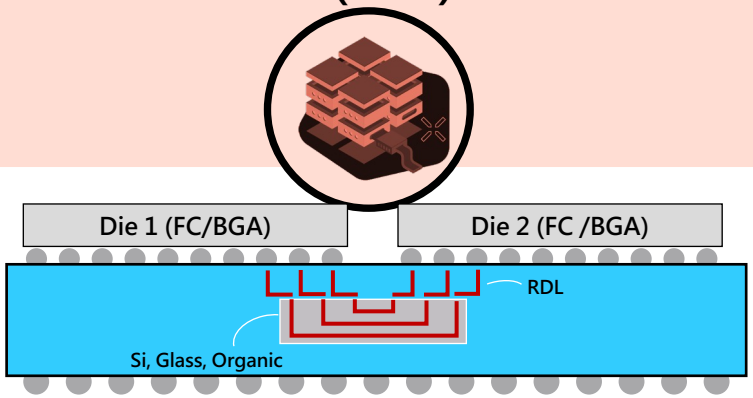
- Embedded chip technology combined by substrate(carrier)
- Co-owned patent with Unimicron

State-of-art(Intel)



- EMIB: Local bridge, different bump size
- Apply for Intel CPU
- In production

State-of-art (ITRI)



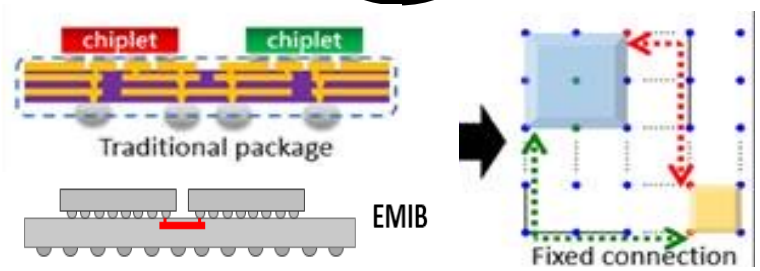
- Embedded Interposer Carrier (EIC): local bridge between two or more chips with or without TSV
- Same intra-structure as FCBGA

Programmable Packaging(5/5)

- Pre-fabricated substrate (size/function scalable)
- Patented embedded switching chips
- One-layer redistribution to meet different applications (infrastructure)

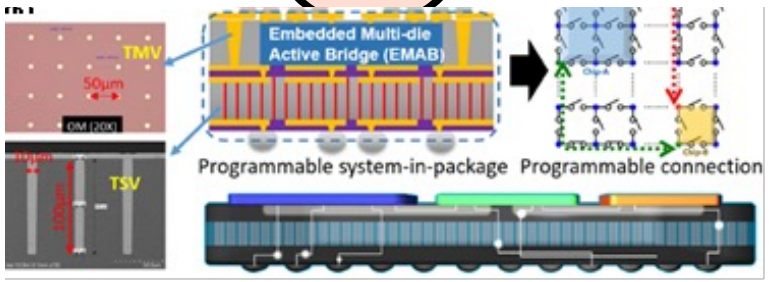
Current/State-of-Art

Intel
Amkor



- regular design, fixed interconnect (w/o programmable) and MP
- less flexibility for same approach

Innovative design/structure



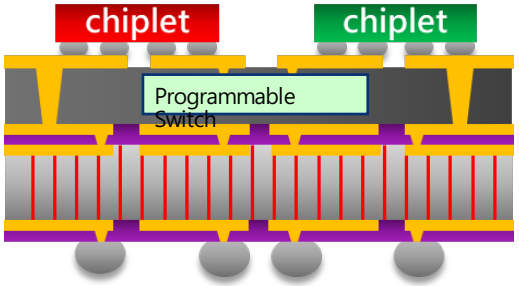
- Innovative achievement for AIoT
- 2022 IEEE Symposium on VLSI
- Prototyping and time-to-market

- Near Memory
  - CoW, WoW, SoIC, ...
  - HPC: thermal solution, ...
- Computing in Memory
  - Innovative Architecture, New non-Volatile Memory(MRAM)
  - Low power(KWS): CiM (ISSCC 4 year in a row)
- Platform of HI
  - **Programmable pkg**
  - Wafer-level Chiplets integration Shuttle service

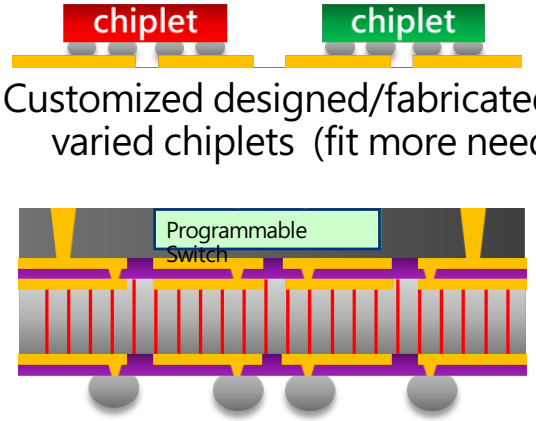
Programmable Package(with Switching Chip)

Multi-modal Package
The best solution to fulfill a variety of fast-prototyping

Programmable SiP
Variety of interface: GPIO, I2C, SPI, UART, SAR ADC



Patent Pending

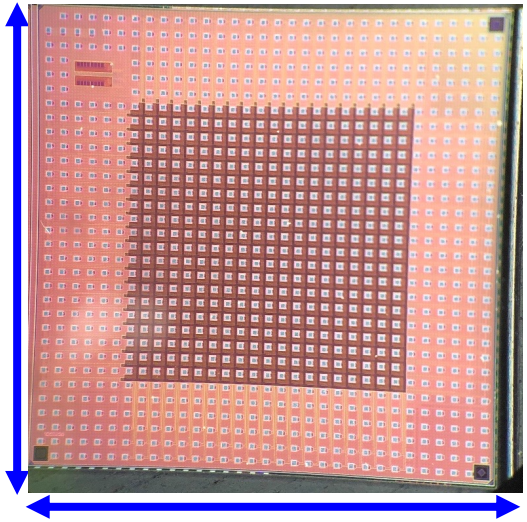


Customized designed/fabricated for varied chiplets (fit more need)

Pre-designed/fabricated main package body (one design)

Source: VLSI Symposium, Hawaii 2022

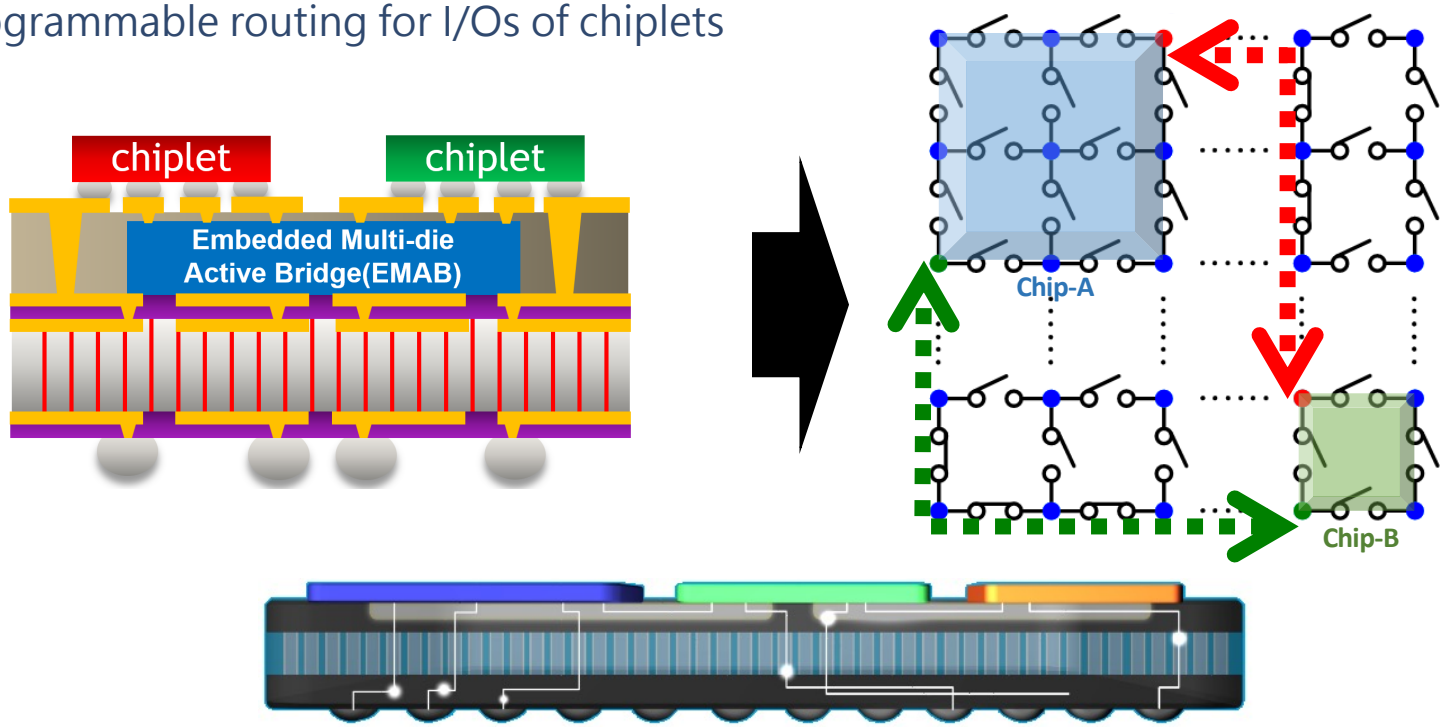
Summary Table of the EMAB Chip



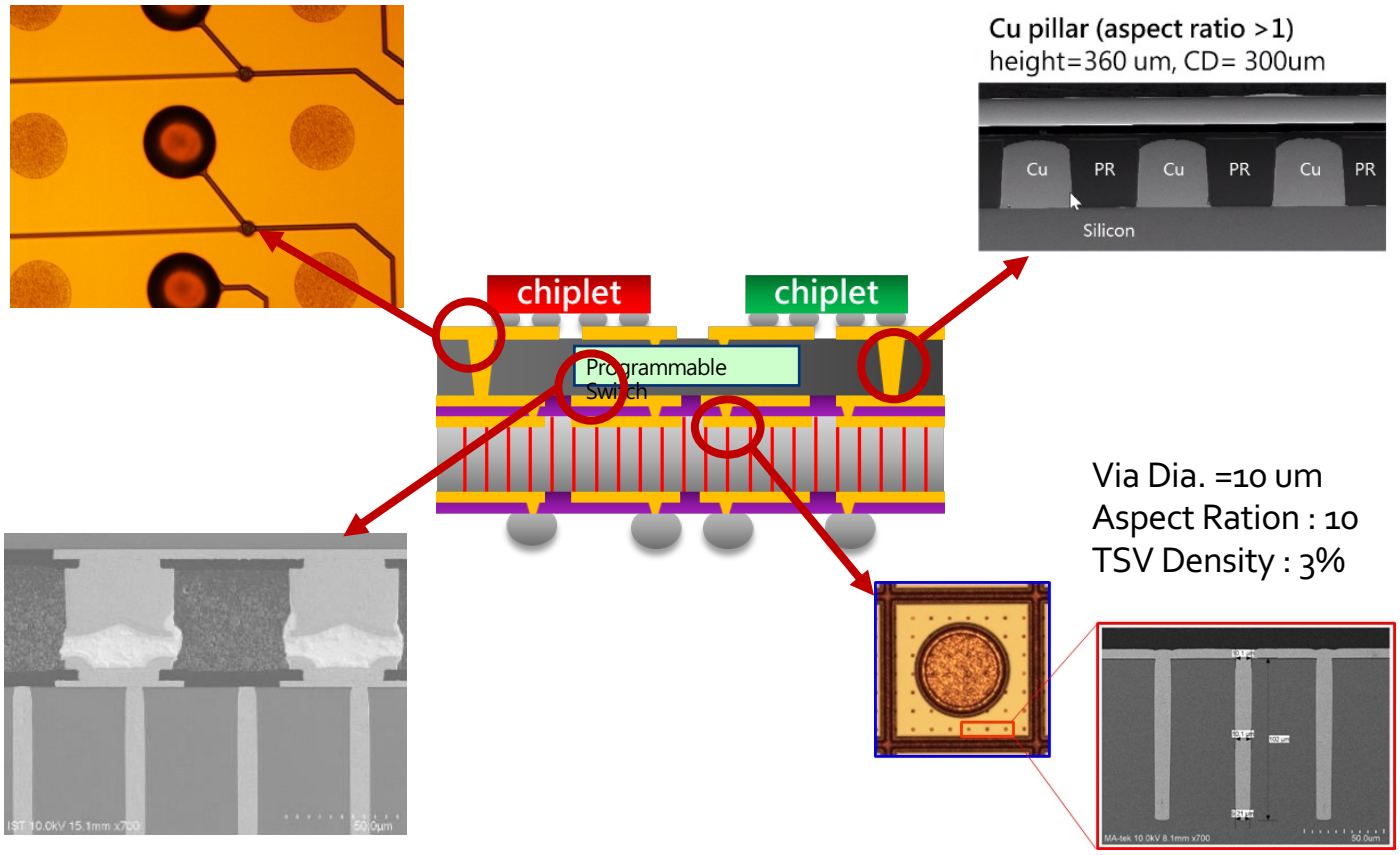
Process	TSMC 130nm
Number of I/O blocks	400
Num. of Checkerboard path per I/O block	4
Num. of Highway path per I/O block	2
Max. voltage of I/O blocks (V)	1.5
Max. current of each I/O block (mA)	5
Power consumption (static)	7.74 uW
maximum data rate of checkerboard path (1 I/O block)	5 Gbps/switch
maximum data rate of checkerboard path (20 I/O block)	100 Mbps
maximum data rate of super highway path	1 Gbps

Switching Chip@Programmable Package

- Embedded Multi-die Active Bridge (EMAB) Chip
  - Programmable routing for I/Os of chiplets



Achievements for Integration

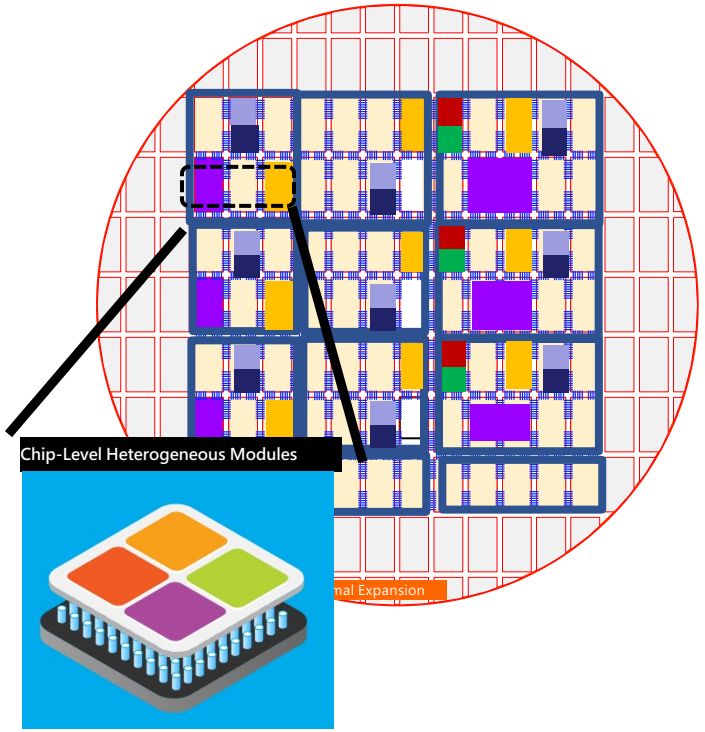


Source: VLSI Symposium, Hawaii 2022

- Near Memory
  - CoW, WoW, SoIC, ...
  - HPC: thermal solution, ...
- Computing in Memory
  - Innovative Architecture, New non-Volatile Memory(MRAM)
  - Low power(KWS): CiM (ISSCC 4 year in a row)
- Platform of HI
  - Programmable pkg
  - **12’ Wafer-level Chiplets integration/Shuttle service**

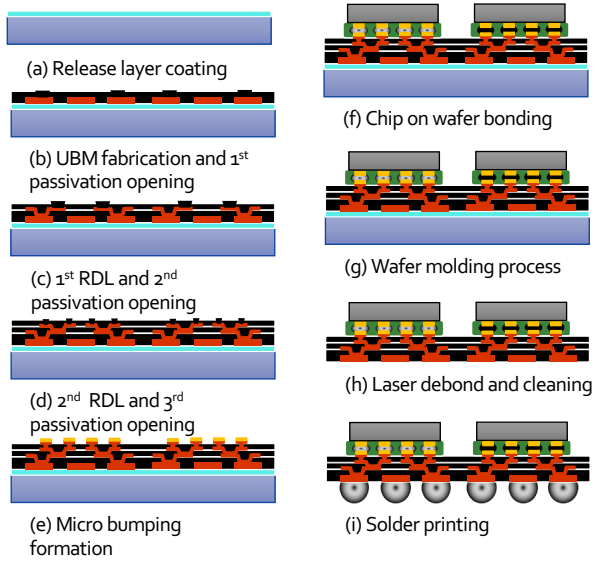
Chiplets Integration: Opportunities and Challenges

- Design of connectivity
- Chiplets integration
- Integrated Yield
- Reliability/Testing



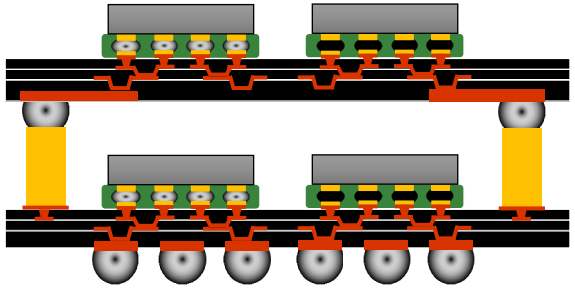
300mm Integrated Fan-Out Line@ITRI

Fan-Out Platform



Year 2021~2023

Fan-Out Stacking



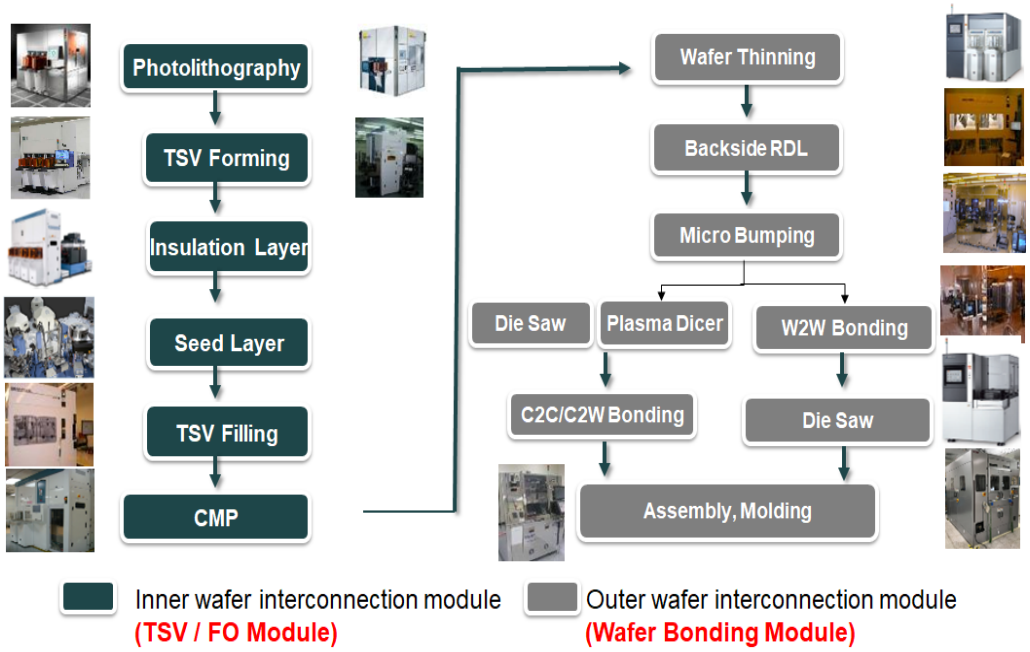
Year 2023~2024

Hi-CHiP

異質整合系統級封裝開發聯盟

Heterogeneous Integration and Chiplet System Package Alliance

ITRI 300mm Back-end Platform



SIG 1

Intelligent System

- Chiplet Interface Design
- Design For Automotive Reliability
- Thermal Dissipation Solution

SIG 2

3D/FO Process Integration

- 3D Chiplet Stack
- Fan-Out Pilot Line
- CPO Integration

SIG 3

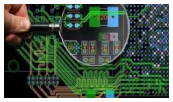
Equipment/Materials

- Research Center construction
- Materials and process development in the ITRI pilot line
- Data Bank

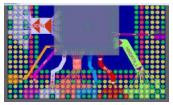
Chiplet-Based Integration

~2020 ~2025

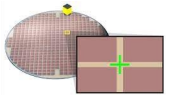
System



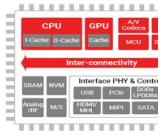
Board



Packaging



SoC Die



Hard
/Soft IPs

IDM

intel

SAMSUNG

Quanta Computer

Unimicron 欣興電子

日月光集團

MEDIATEK
REALTEK

GUC

Application-driven
(e.g. 8K image AI detection)

Top-down

Chiplet Based IDM

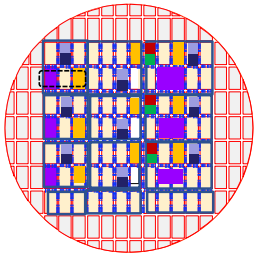
Full Design Flow
cadence

Board — Unimicron 欣興電子 tn

Advanced Packaging — 日月光集團 SPIL

Chiplet — MEDIATEK REALTEK

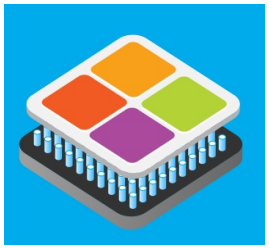
IP — GUC ARM



CHIPLETS SHUTLLE

WLP, FanOut, 3D stacking

- Chiplet foundry
- Full flow co-design service



NEW ARCHITEC

Software define chiplets, Programmable pkg

- SMEs/startups
- Customized integrated platform

DIGITAL MANUFACTURING

Equipment, Material, Design service

- Optimized flow for chiplets
- Yield improvement

Summary

- High Frequency- GaN on Si/AiP/mmWave
  & High Power- SiC module design/fabricating/test
- High speed, Project "**AI on Chip**": Ultra-low power CIM
  - ✓EIC, low temp. wafer bonding, adv. thermal
  - ✓Ultra-low power AI accelerator, wide temp range (4K~400K)、highest speed (8Kb aray of SOT-MRAM ，1ns & duration7000B)
- **Heterogeneous Integration**:
  - ✓Programmable packaging
  - ✓12" WL-Chiplets integration