Heterogeneous 3D Chiplet Integration for AI application

Mitsu Koyanagi

Global INTegration Initiative (GINTI)
Tohoku University, Japan

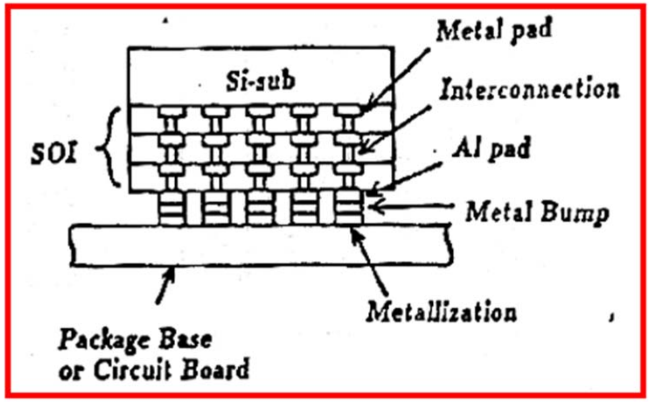
Tohoku-MicroTech. Co., Ltd, Japan

Outline

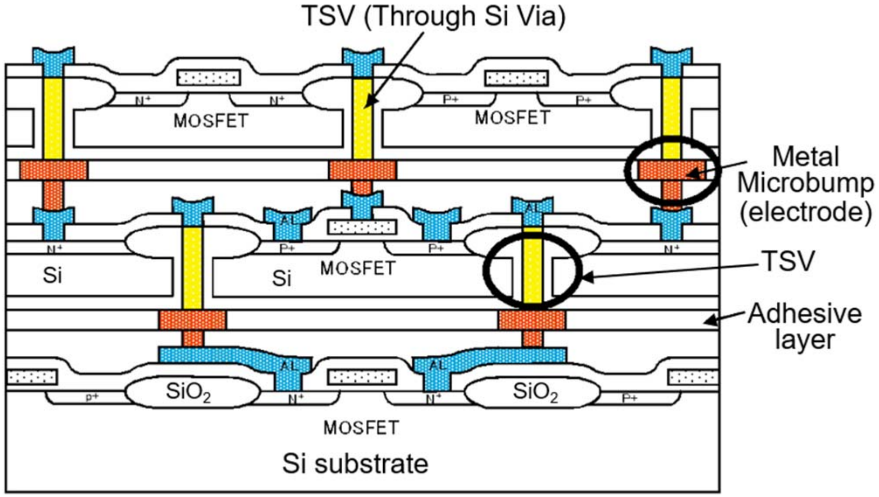
- Introduction
- 3D Chiplet Integration
- 3D Heterogeneous Integration
- Neuro/AI System by 3D Integration Technology
- Conclusions

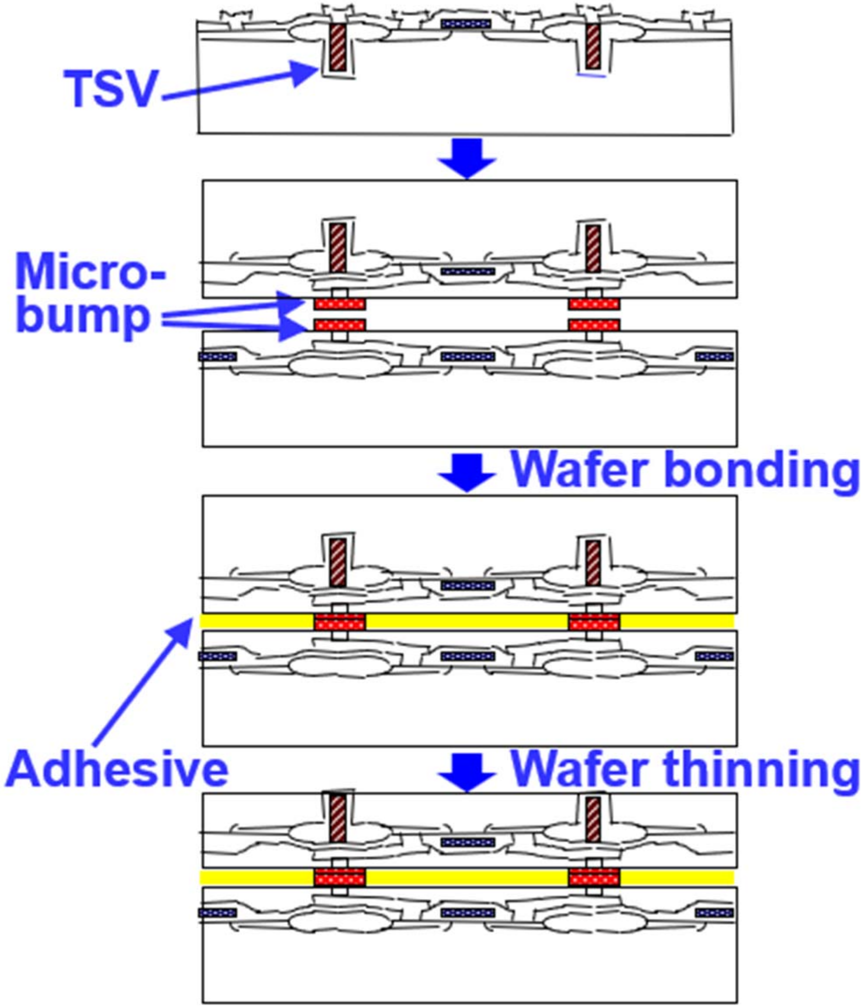
First Proposal of 3D Integration Technology in Tohoku University

The 1st proposal of 3D integration using wafer bonding in 1989



M. Koyanagi, Proc. 8th Symposium on Future Electron Devices, pp.50-60 (Oct. 1989)

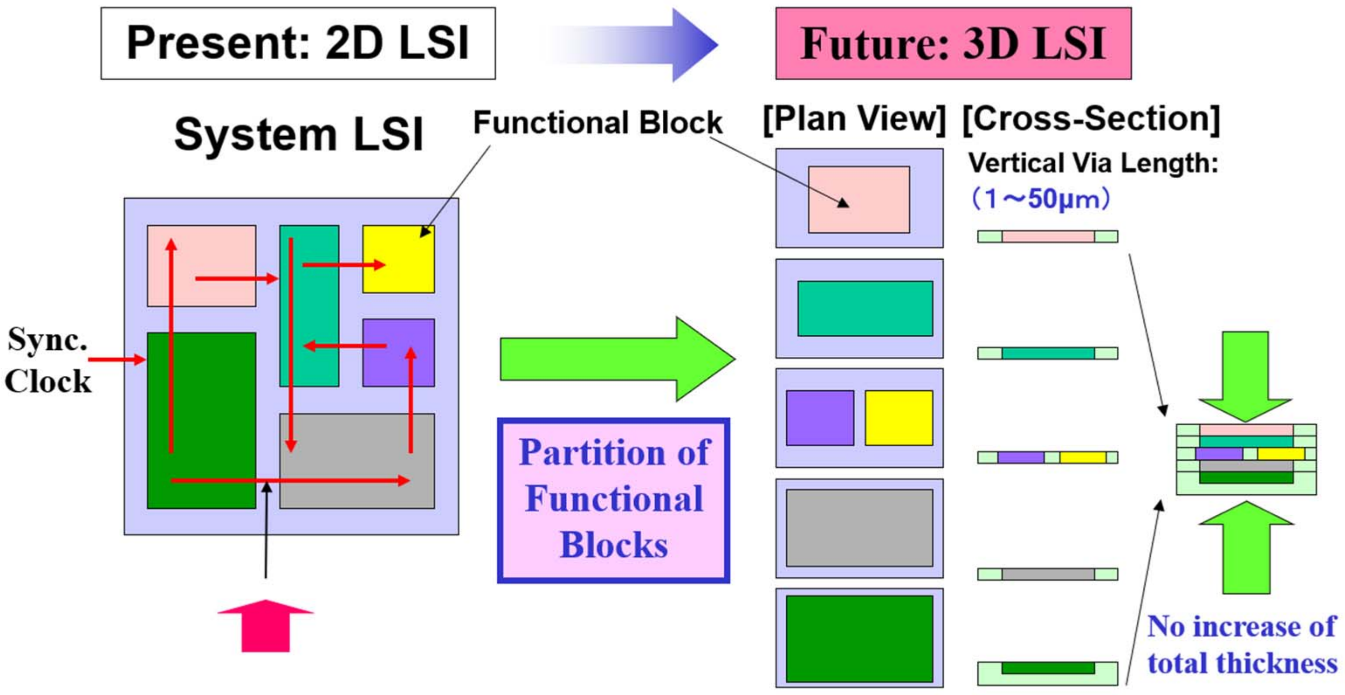




3D Chiplet Integration in Tohoku University

Merits of 3D LSI

High Performance Low Power
New Functions Low Cost



Present: 2D LSI

Future: 3D LSI

System LSI

Functional Block

[Plan View] [Cross-Section]

Vertical Via Length:
（1～50μm）

Sync.
Clock

Partition of Functional Blocks

No increase of total thickness

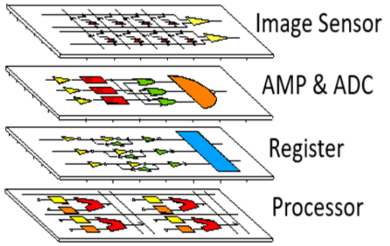
Big concerns: Long wiring
Global wiring

Replaced by Vertical Vias

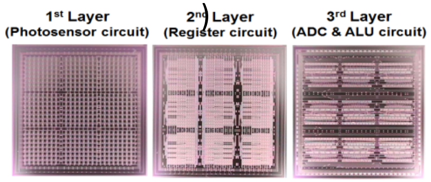
M. Koyanagi, Stanford University Workshop (CIS Round Table) (2005)

First 3D-IC Test Chips with TSVs Fabricated in Tohoku Univ.

3D image sensor chip

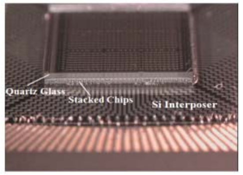


(a
)



1st Layer (Photosensor circuit)
2nd Layer (Register circuit)
3rd Layer (ADC & ALU circuit)

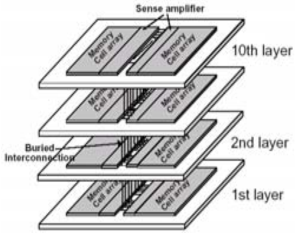
(b)



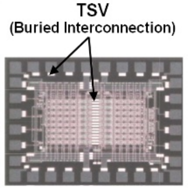
(c
)

H. Kurino, M. Koyanagi *et al.*
IEEE IEDM (1999)

3D memory

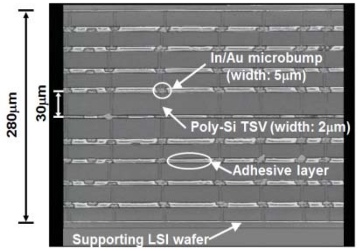


(a
)

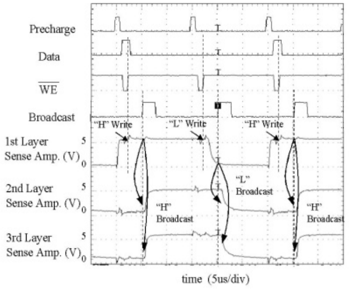


DRAM chip photo
(chip size: 2mmx1.6mm)

(b)



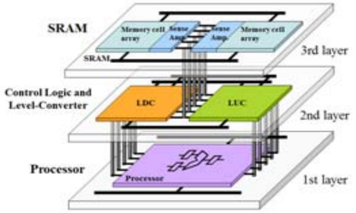
(c
)



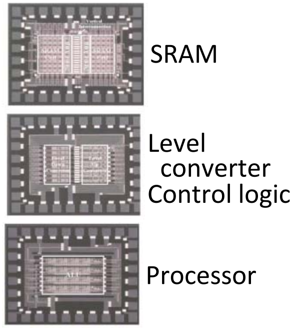
(d)

K. W. Lee, M. Koyanagi *et al.*
IEEE IEDM (2000)

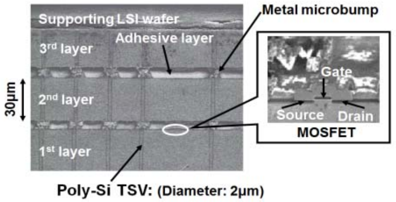
3D microprocessor chip



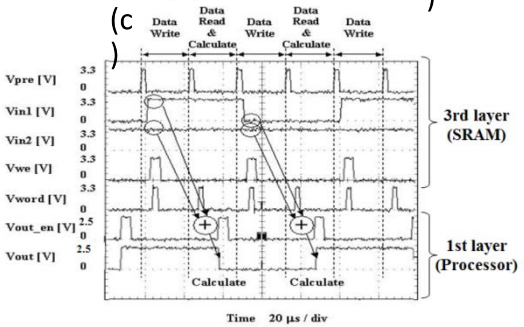
(a



(b
)



(c
)

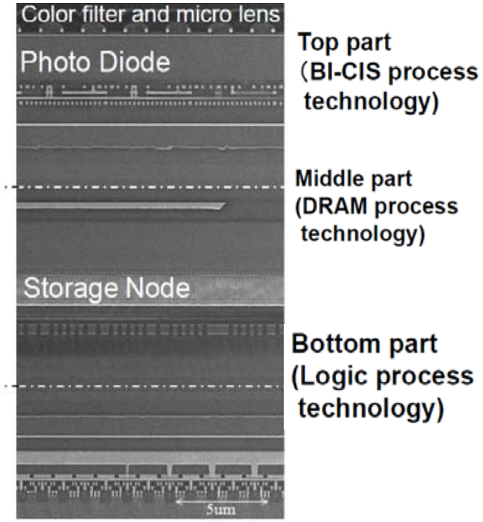


(d
)

T. Ono, M. Koyanagi *et al.*, IEEE COOL Chips (2002)

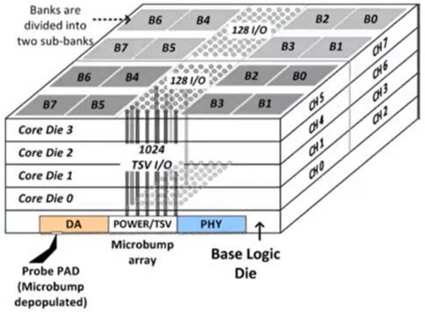
Practical Implementation of 3D-IC Chips with TSVs in Semiconductor Manufacturers

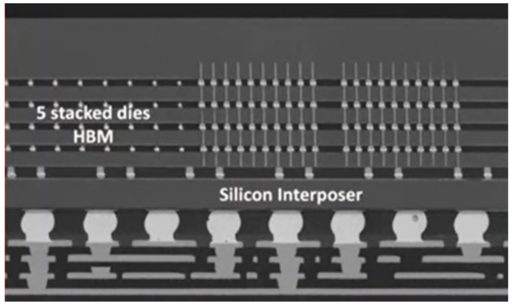
3D image sensor chip



T. Haruta (Sony)
IEEE ISSCC (2017)

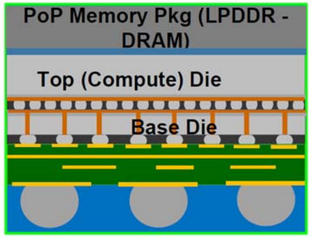
3D memory

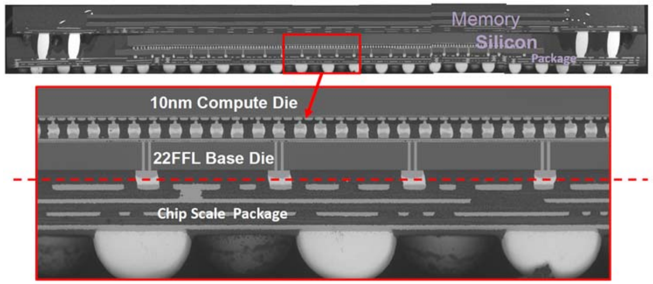




J. C. Lee et al. (SK Hynix)
IEEE ISSCC (2016)

3D microprocessor chip

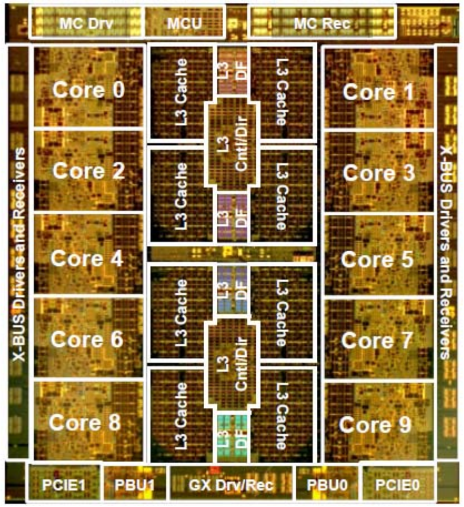




Wilfred Gomes et al. (Intel)
IEEE ISSCC (2020)

Chiplet Integration

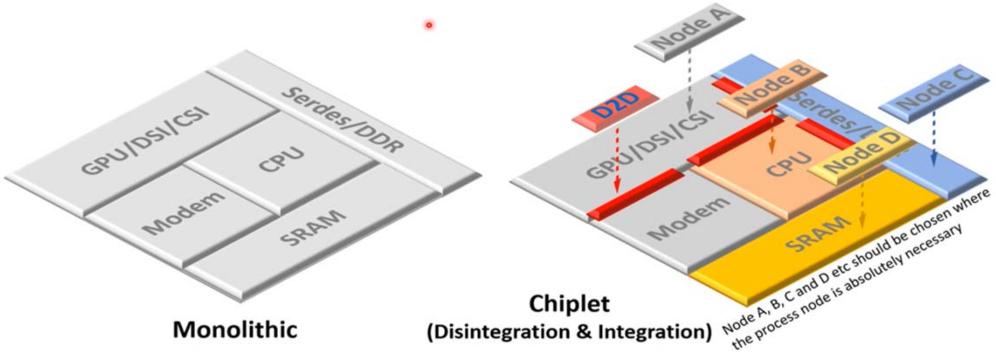
UCIe (Universal Chiplet Interconnect Express)



Monolithic chip

IBM z14 Processor
14nm technology node
6.1B transistors
Chip size: 696 mm2
(ISSCC 2018)

▶ **Disintegration SOC first and integration with best/optimized node**


GPU/DSI/CSI
Serdes/DDR
CPU
Modem
SRAM
Node A
Node B
Node C
Node D
D2D
Monolithic
Chiplet
(Disintegration & Integration)
Node A, B, C and D etc should be chosen where the process node is absolutely necessary

Source: S-H You (IEDM2020 Short Course)

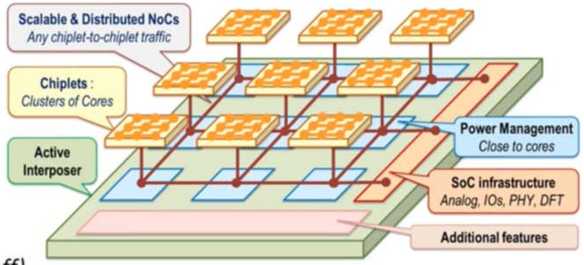
Chiplets & Active Interposer : Concepts

- Chiplet for:
  - Lower cost
  - Higher modularity
  - From IP-reuse to circuit-reuse

✓ Using passive interposers (2.5D) or organic substrate:

❖ But limitations regarding
- ➢ Chiplet connectivity (scalability),
- ➢ Less scalable function (heterogeneity)

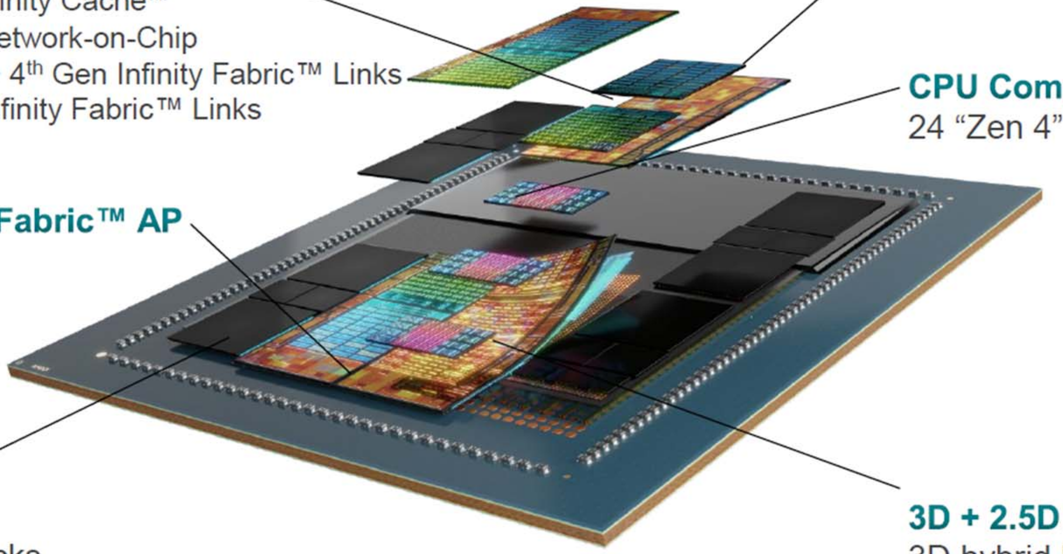
- Active Interposer, the « *Smart Hub* » for :
  - Scalable System Interconnects
  - PHYs for off-chip communication
  - Power Management
  - DFT, thermal, etc



In a mature CMOS technology (cost-performance trade off)

D. Dutoit et al. (CEA-LETI), IEDM2020

AMD Instinct™ MI300A Modular Chiplet Package



I/O Die (IOD) x4
128 Channel HBM3 Interface
256MB AMD Infinity Cache™
Infinity Fabric Network-on-Chip
2 x16 PCIe® 5 + 4th Gen Infinity Fabric™ Links
6 x16 4th Gen Infinity Fabric™ Links

Accelerator Complex Die (XCD) x6
228 AMD CDNA™ 3 Compute Units

CPU Complex Die (CCD) x3
24 “Zen 4” Cores [ISSCC23]

AMD Infinity Fabric™ AP Interconnect

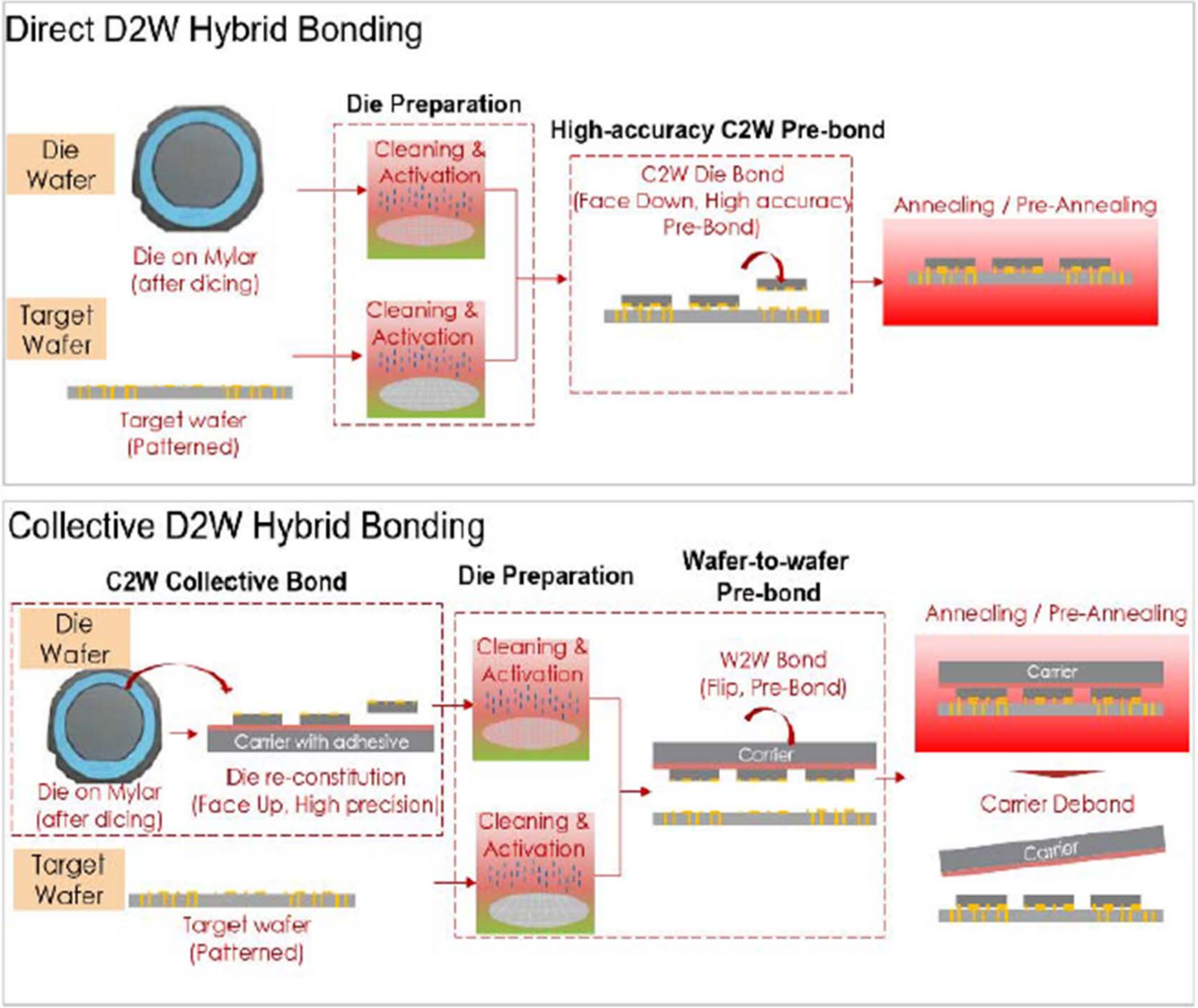
HBM3
8 physical stacks
AMD Instinct™ MI300A: 128 GB (8-high)

3D + 2.5D Advanced Package
3D hybrid bonding
2.5D silicon interposer

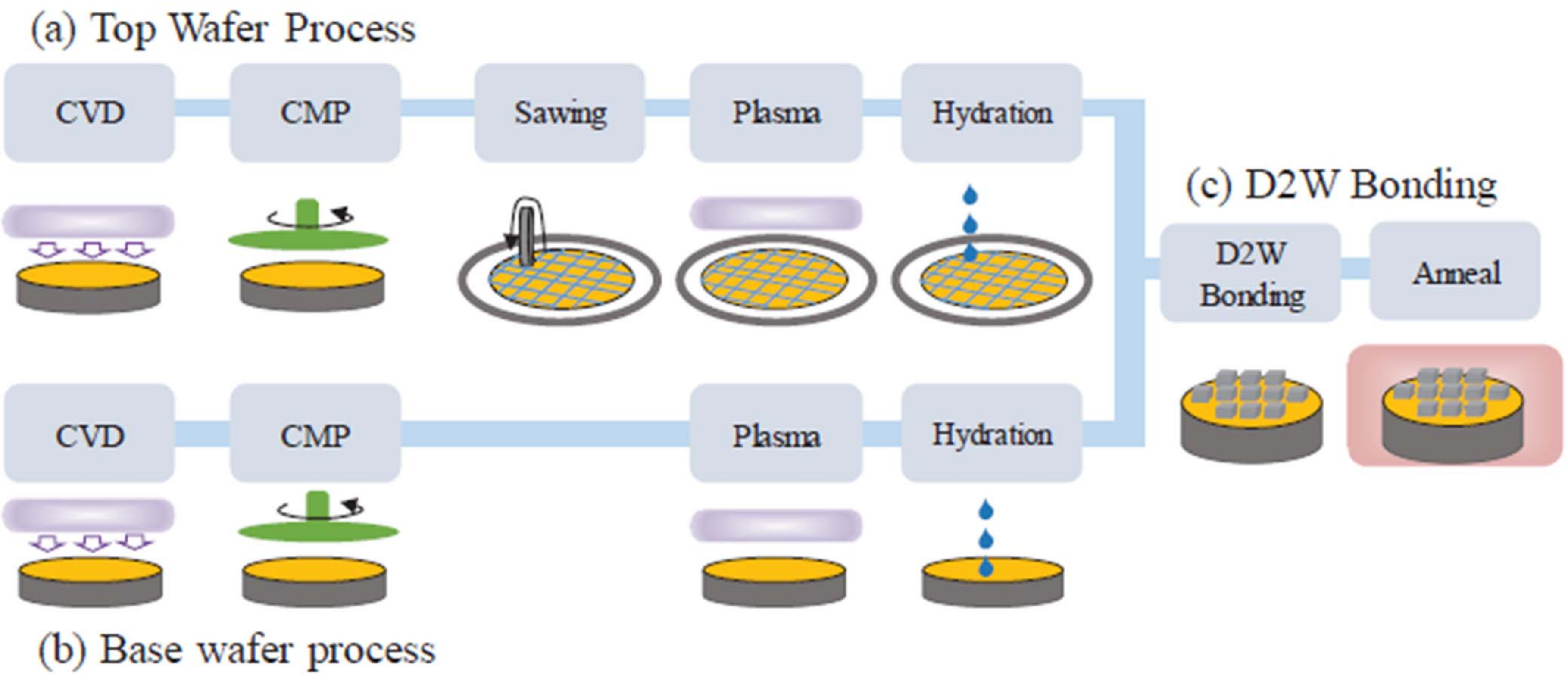
Alan Smith et. al.. IEEE ISSCC, pp.208-209 (2024)

Direct D2W Hybrid Bonding and Collective D2W Hybrid Bonding

(Reconfigured W2W Hybrid Bonding)

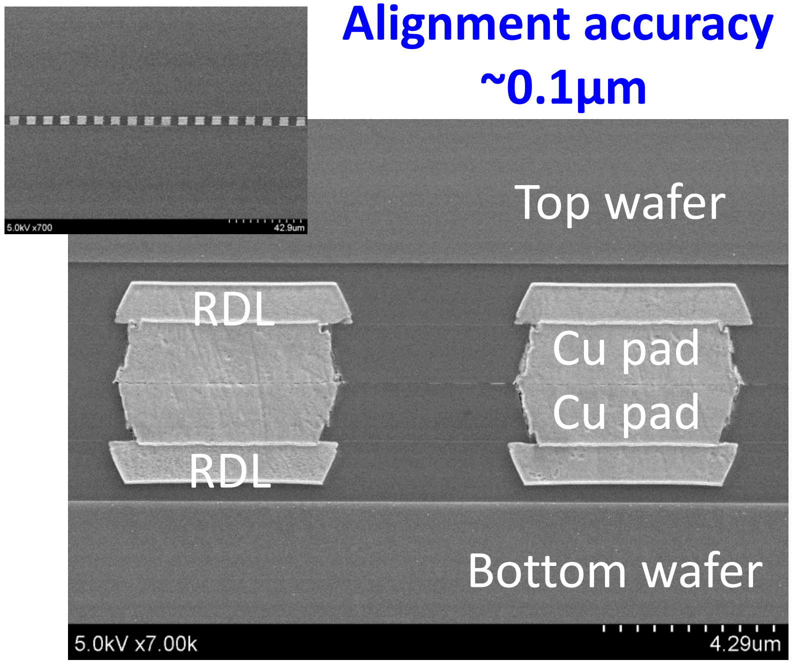

Direct D2W Hybrid Bonding
Die Wafer
Die on Mylar (after dicing)
Target Wafer
Target wafer (Patterned)
Die Preparation
Cleaning & Activation
Cleaning & Activation
High-accuracy C2W Pre-bond
C2W Die Bond (Face Down, High accuracy Pre-Bond)
Annealing / Pre-Annealing
Collective D2W Hybrid Bonding
C2W Collective Bond
Die Wafer
Die on Mylar (after dicing)
Carrier with adhesive
Die re-constitution (Face Up, High precision)
Target Wafer
Target wafer (Patterned)
Die Preparation
Cleaning & Activation
Cleaning & Activation
Wafer-to-wafer Pre-bond
W2W Bond (Flip, Pre-Bond)
Carrier
Annealing / Pre-Annealing
Carrier
Carrier Debond
Carrier

Process Flow of D2W Hybrid Bonding



S. Lee et al., IEEE ECTC, pp.1085-1089 (2022)

Hybrid Bonding in Tohoku University

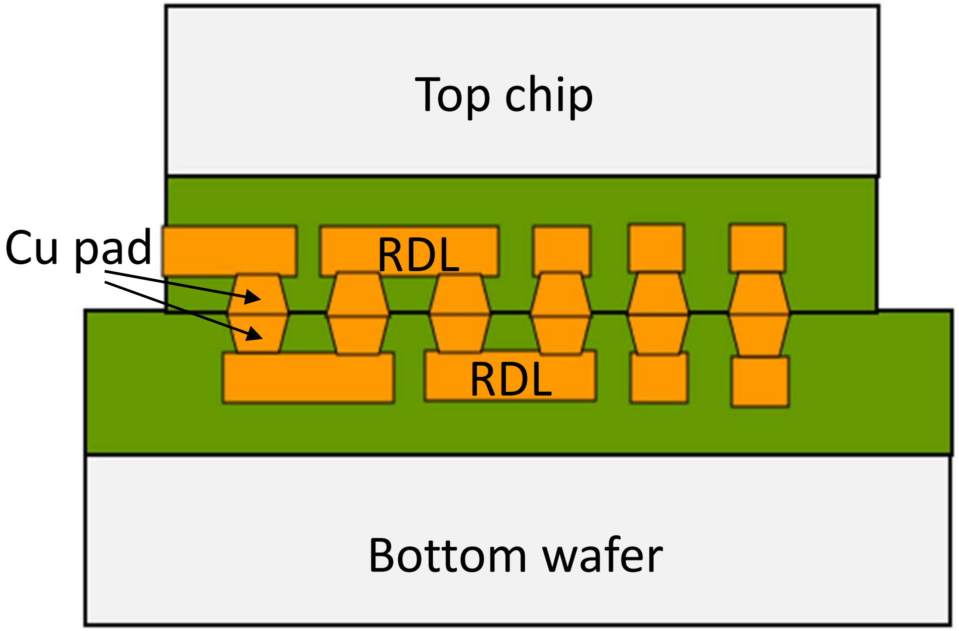


Alignment accuracy ~0.1µm

Cu post size 5µm; pitch 10µm

(a) W2W

M. Murugesan, M. Koyanagi, T. Fukushima, IEEE ECTC (2022)

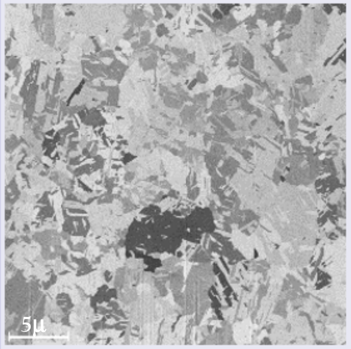
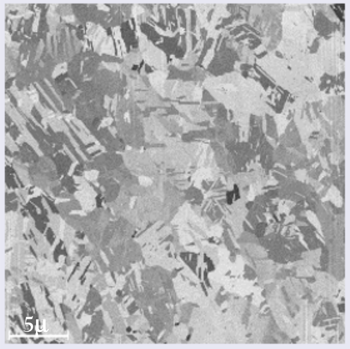
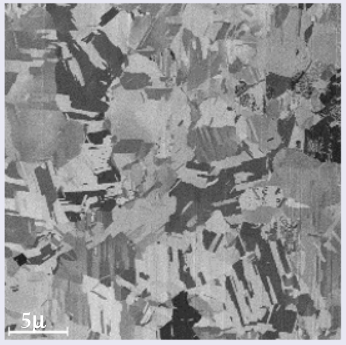
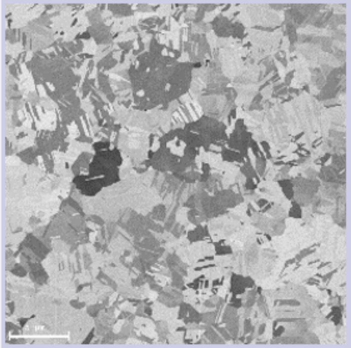
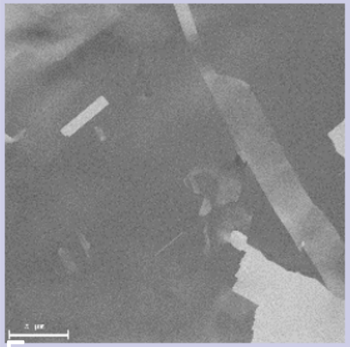
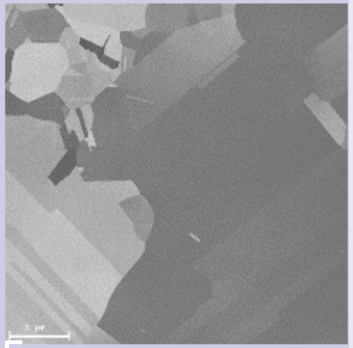


Cu post size 1.5µm; pitch 2.5µm

(b) C2W

To be published

Cu Grain Morphology _ SEM

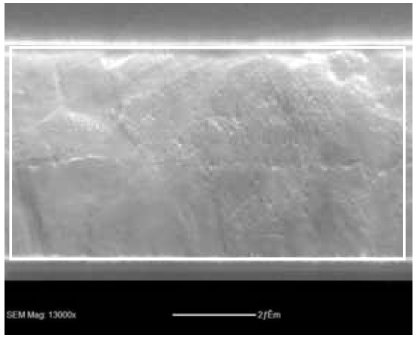
	(250℃*1hr)	(300℃*1hr)	(350℃*1hr)
Plating Chemistry A			
Plating Chemistry B			

Grain size ~2μm
(tiny and extremely random oriented)

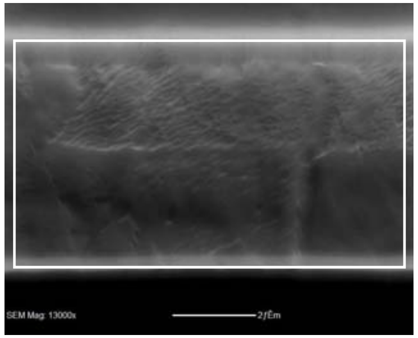
Grain size >10μm
(very large and relatively oriented)

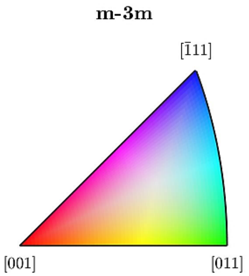
Cu Grain Crystallographic Orientation _ EBSD

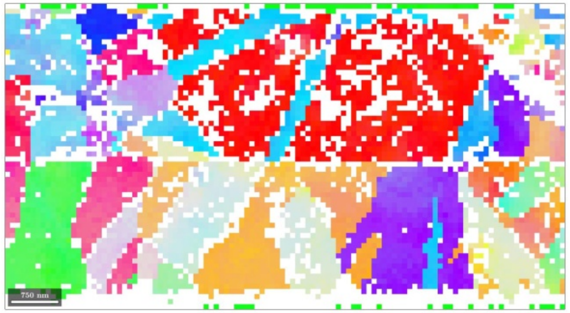
Plating chemistry A

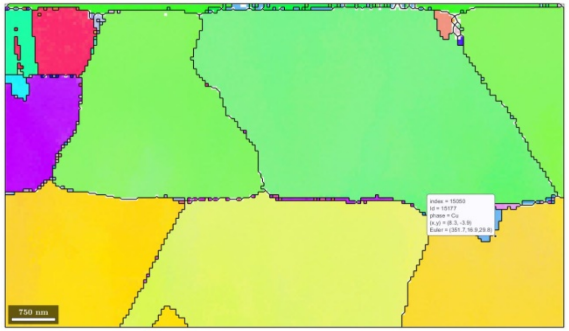


Plating chemistry B



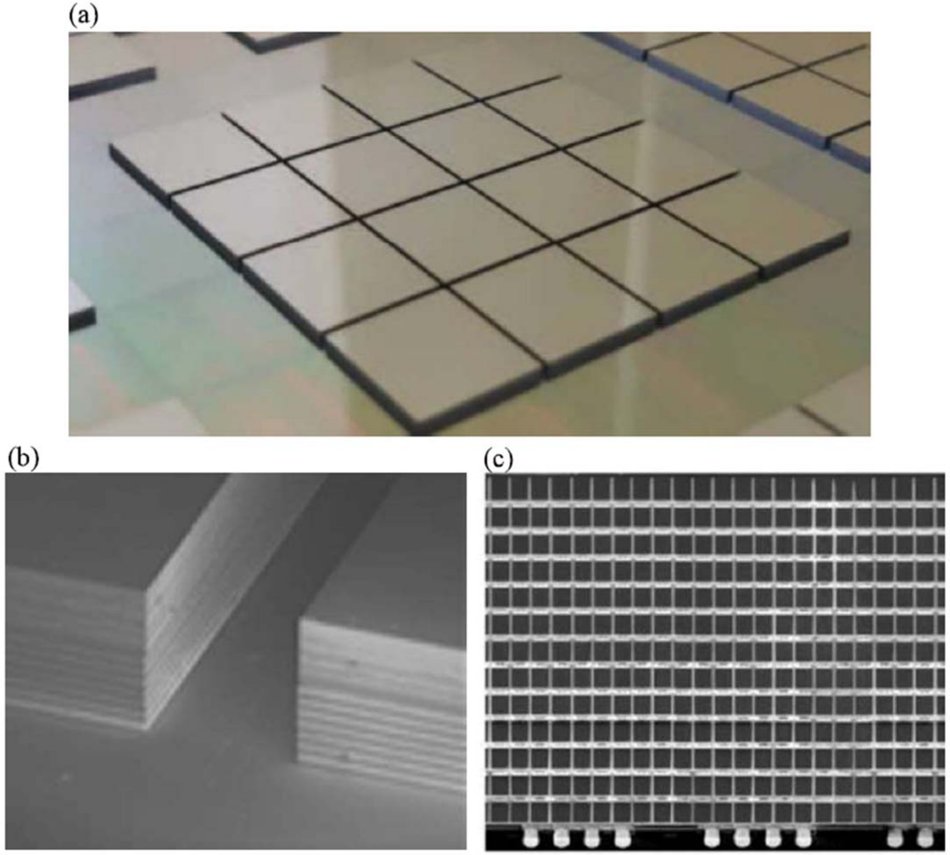






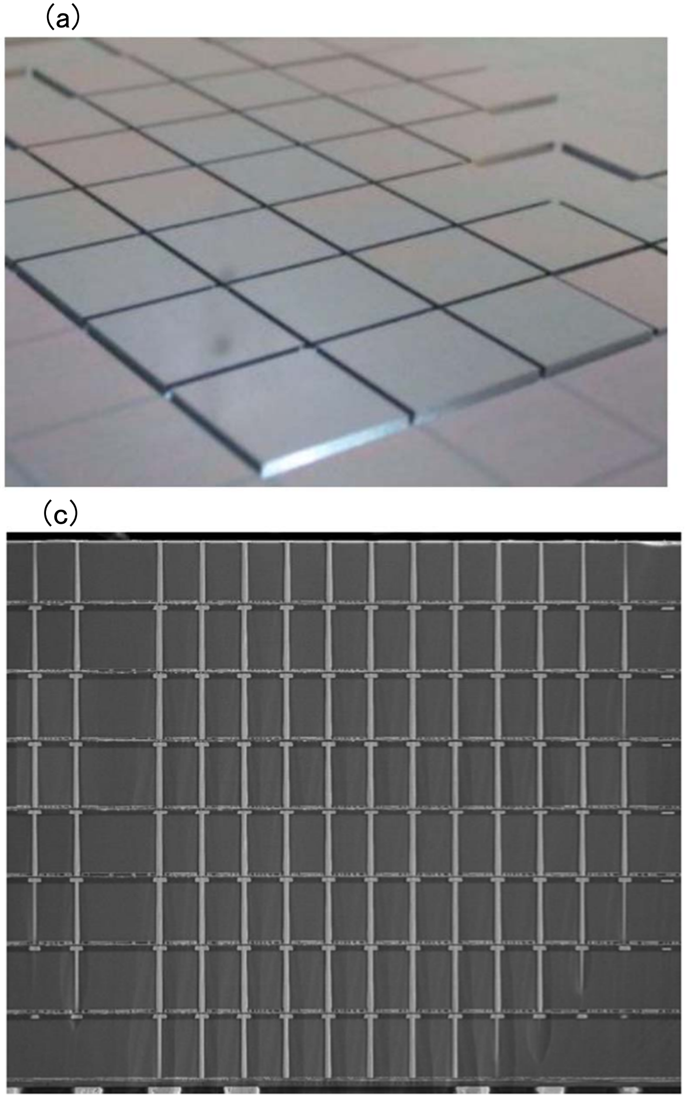
M. Murugesan, M. Koyanagi, T. Fukushima, IEEE ECTC (2022)

3D Chiplet Integration on Wafer by D2W Hybrid Bonding



(a) Low magnification image (b) Tilted SEM image, and (c) Cross-sectional SEM image

S. Lee et al., IEEE ECTC, pp.1085-1089 (2022)

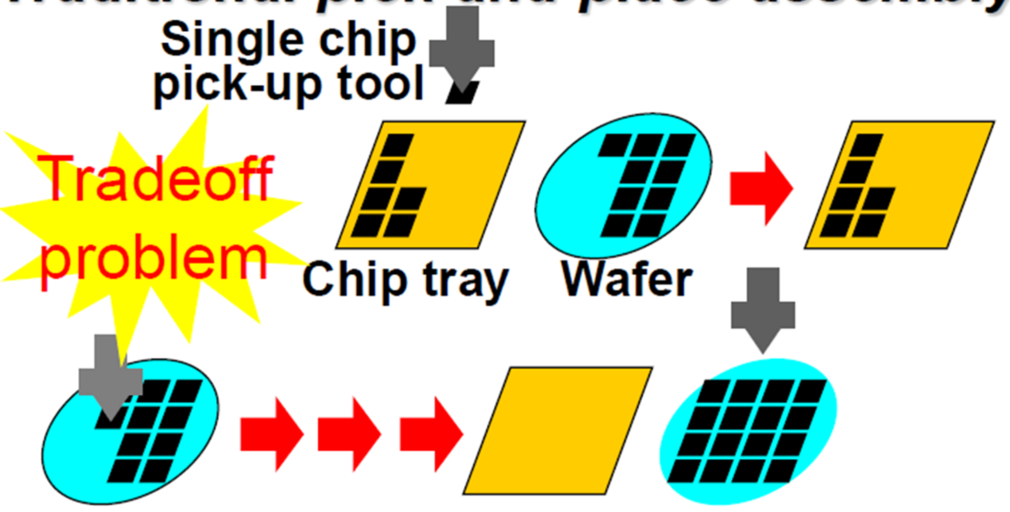


HBM with 8 memory layers

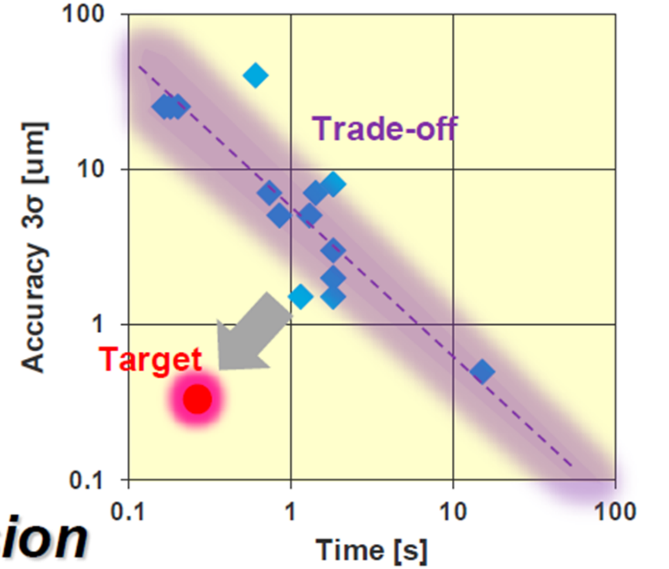
Jaesik Lee, IEEE IEDM, SC2.2 (2023)

Pick-and-Place vs Self-Assembly

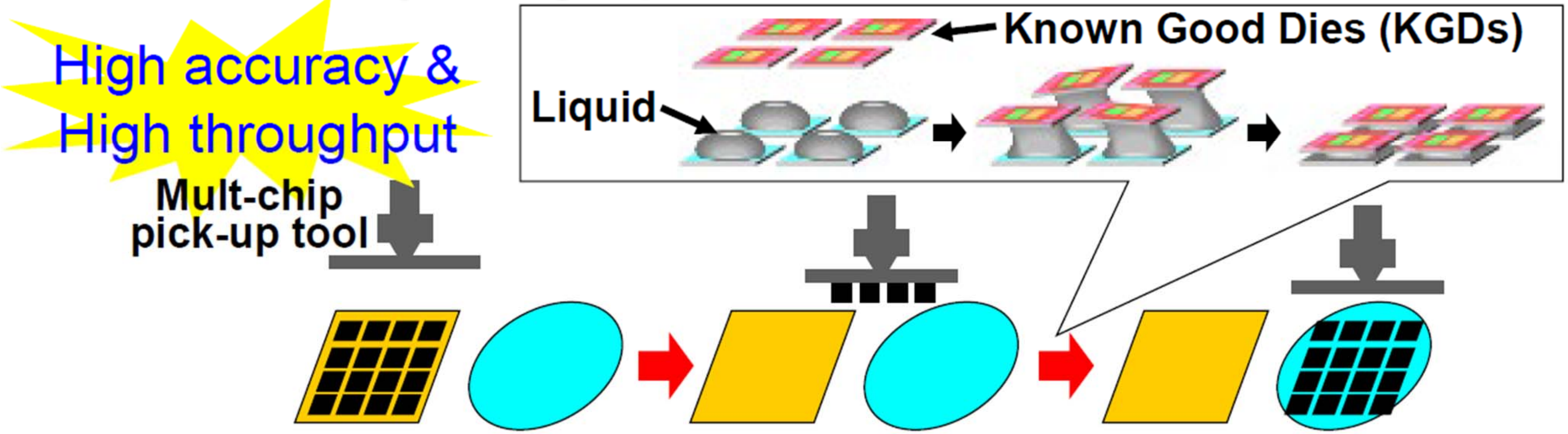
- ***Traditional pick-and-place assembly***



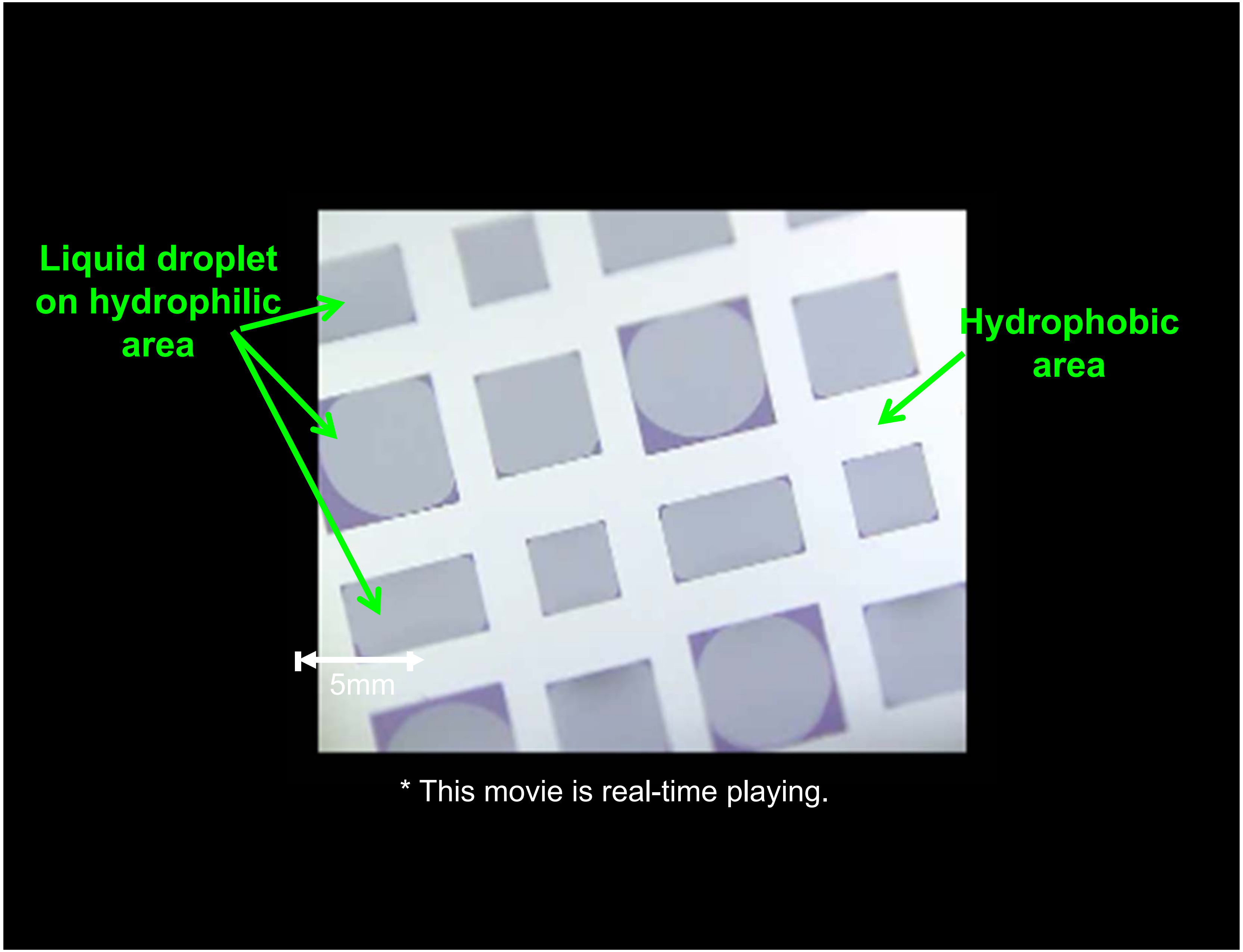
Alignment accuracy and time with commercial bonders



- ***Self-assembly using liquid surface tension***

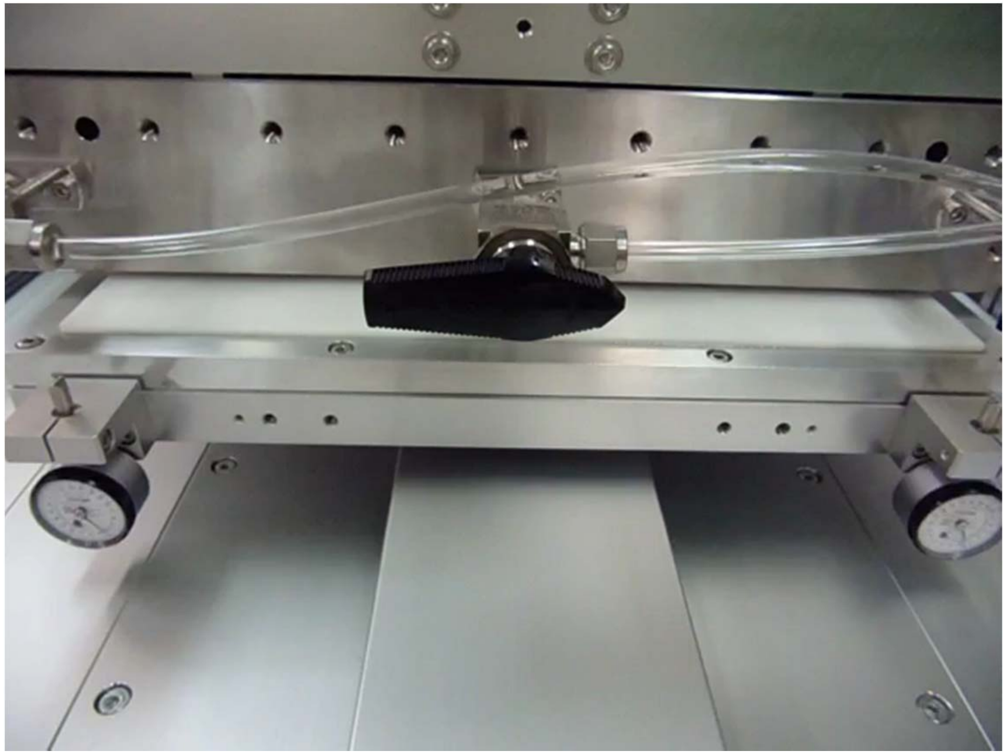


Simultaneous Bonding of Many Dies with Different Size by Self-Assembly

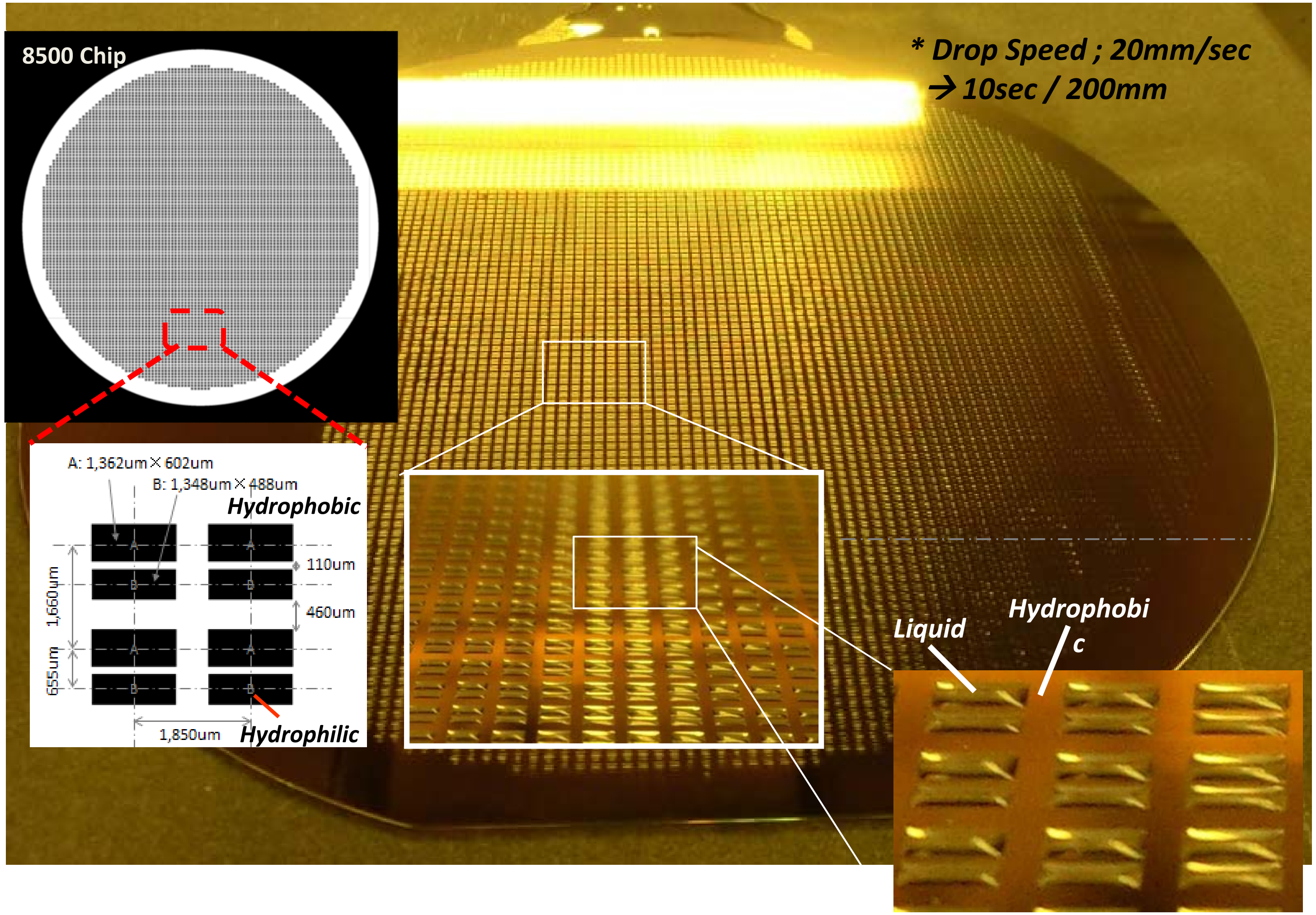


* This movie is real-time playing.

High-Speed Water Droplet Spray for Self-Assembly



Water Droplets Supplied on Small Hydrophilic Area



** Drop Speed ; 20mm/sec*
→ 10sec / 200mm

Photo of Various-size Self-Assembled Chips on 8-inch Wafer Prepared by Hybrid Self-Assembly

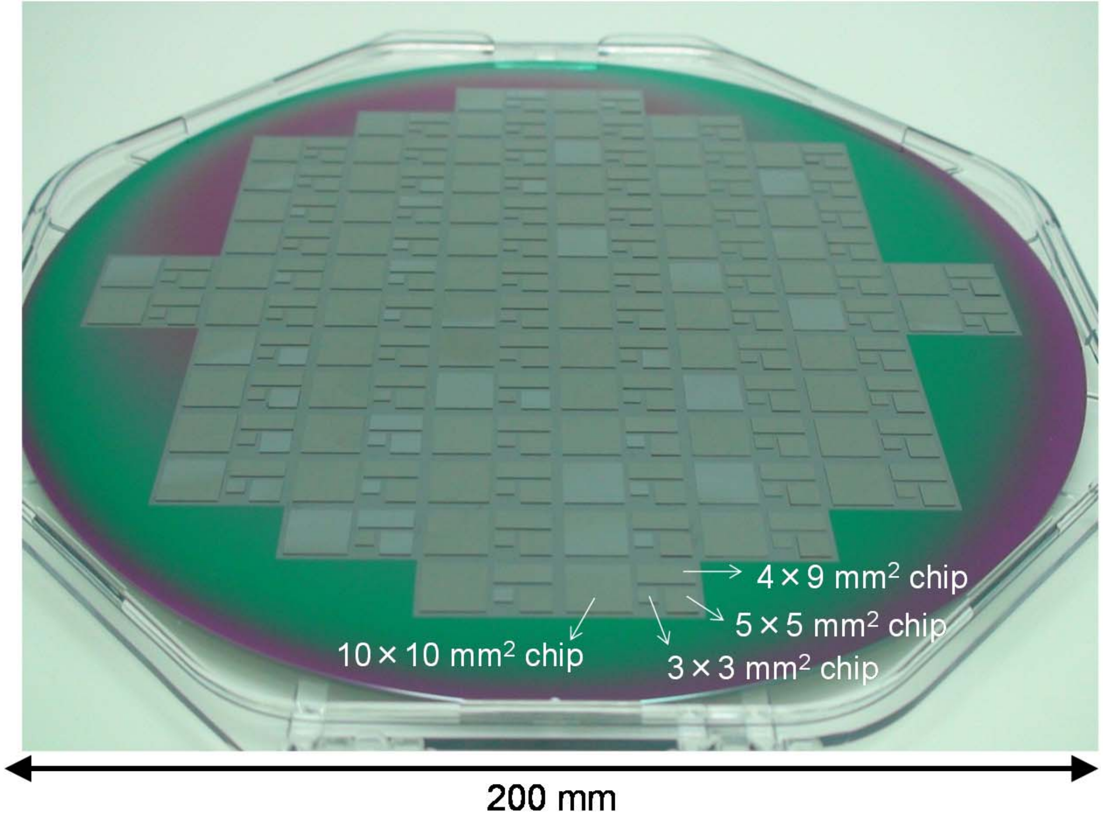
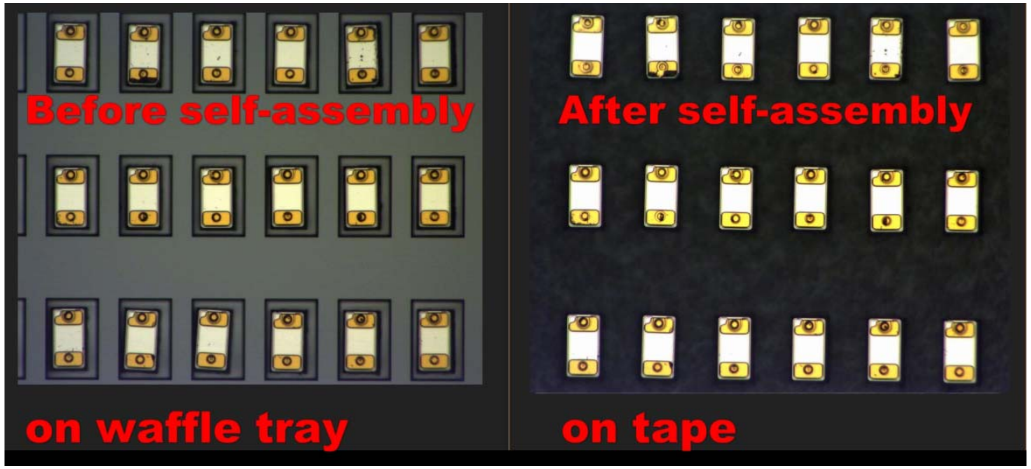
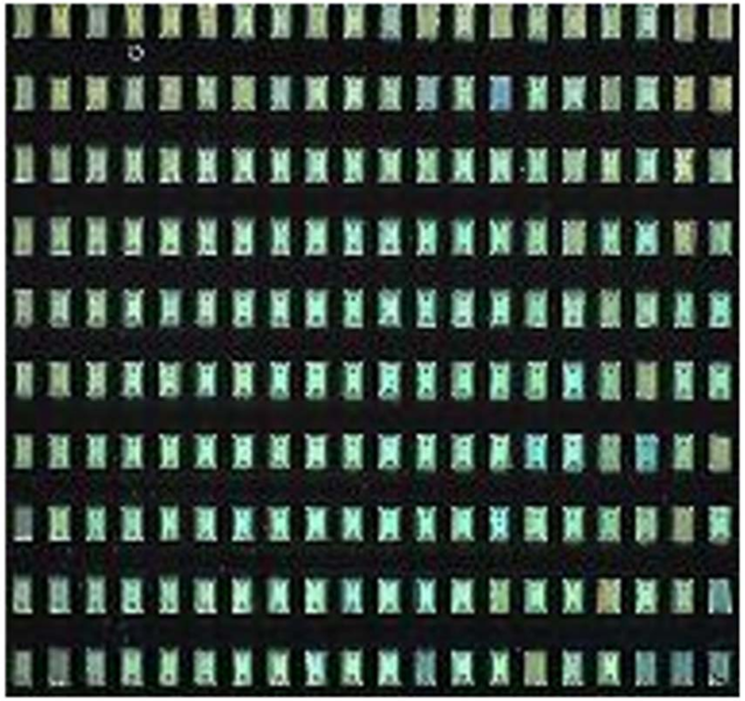


Photo of µLED Array Prepared by Hybrid Self-Assembly

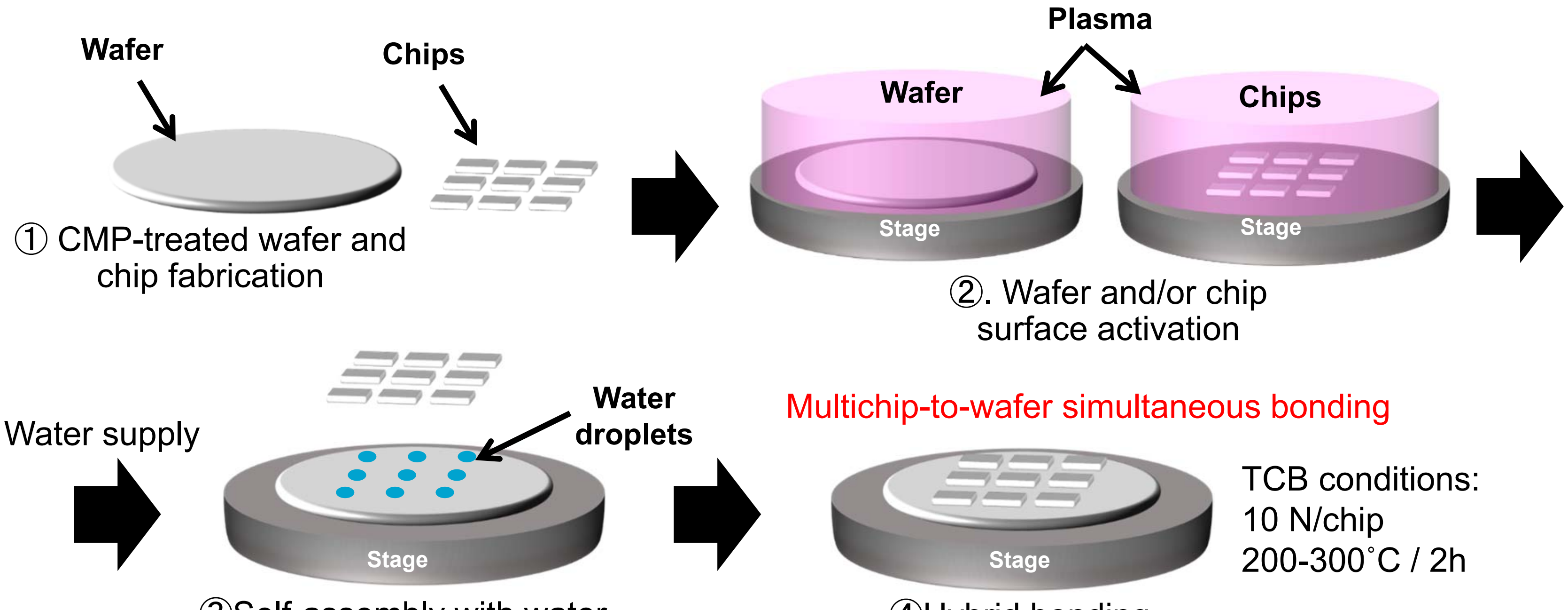




Self-Assembled Micro-LED chips
(75umx125um)

Combined Process Sequence of Self-Assembly and Hybrid Bonding

(SA-Hybrid Bonding)



① CMP-treated wafer and chip fabrication

②. Wafer and/or chip surface activation

Water supply

③Self-assembly with water for multichip-to-wafer bonding

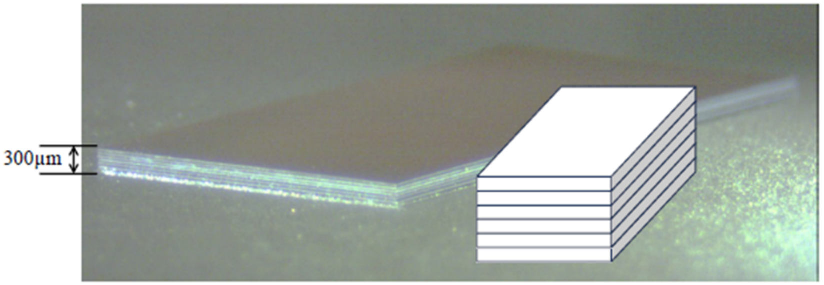
Multichip-to-wafer simultaneous bonding

④Hybrid bonding

TCB conditions:
10 N/chip
200-300˚C / 2h

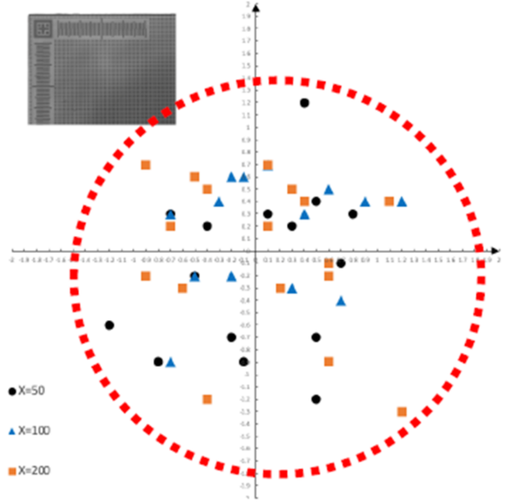
Results of SA-Hybrid Bonding

①



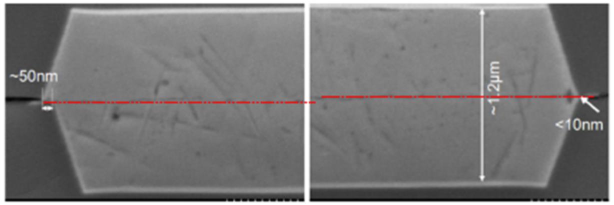
- Using this method, we have currently achieved an assembly of 6 layers. In the future, we will continue to explore the best conditions to achieve a structure of more than 12 layers.

②



- The current average assembly accuracy has reached a level of less than 500um, but the data is still scattered, and efforts will be made to reduce the data scatter in the future.

③



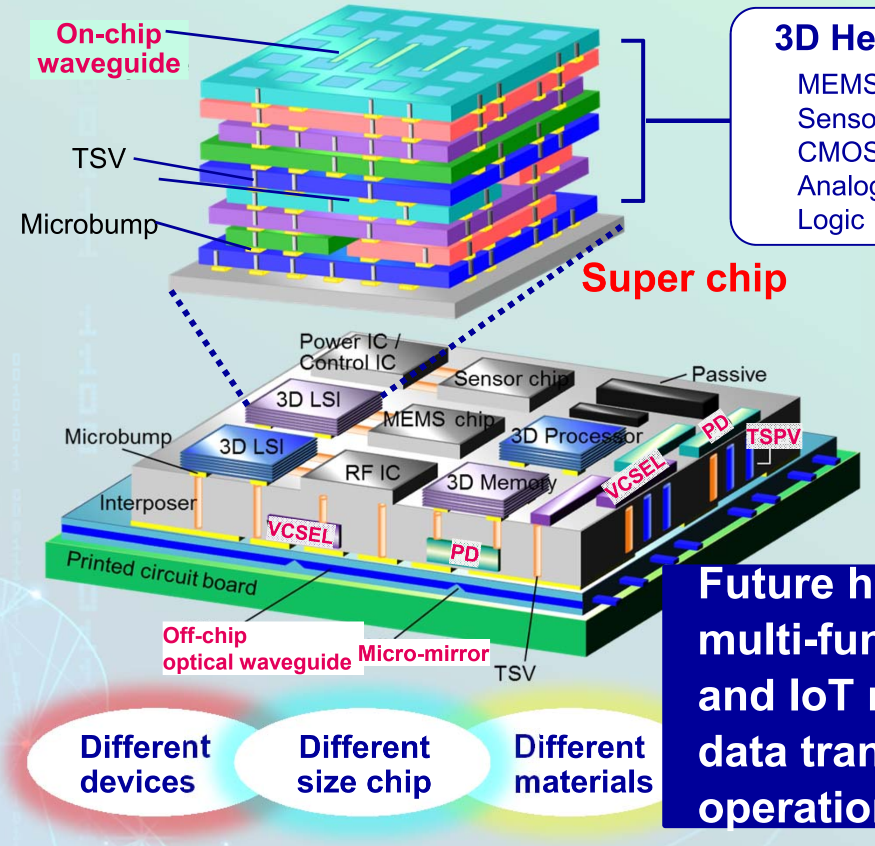
- The current thermal bonding can still see the bonding interface between Cu. In the future, we will continue to explore the impact of liquid on bonding and the optimal bonding conditions.

Heterogeneous Integration

- Device Level
- Architecture/System Level

Device Level Heterogeneous Integration

3D Heterogeneous Integration Technology in Tohoku Univ.



3D Heterogeneous Integration

MEMS chip
Sensor chip
CMOS RF-IC
Analog LSI
Logic IC

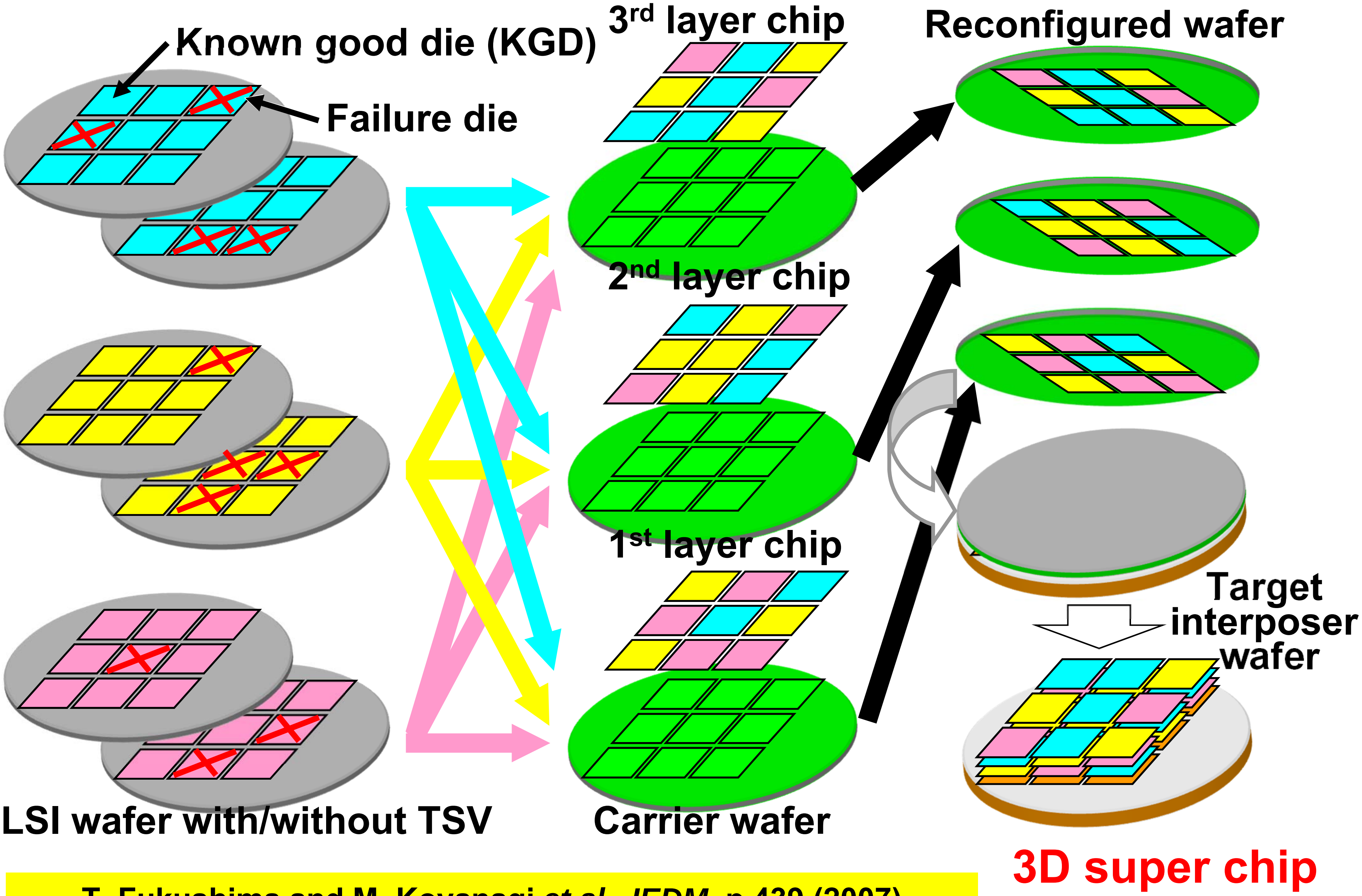
Microprocessor
Power IC/Control IC
Flash memory
DRAM
SRAM

2.5D Heterogeneous Integration

Future high-performance / multi-functional trillion sensors and IoT require high-speed data transfer & low power operation.

T. Fukushima, M. Koyanagi et. Al., IEEE IEDM, p.359 (2005)
K-W Lee, M. Koyanagi et. al., IEEE IEDM, p.531 (2009)

New Reconfigured Wafer-to-Wafer 3D Integration

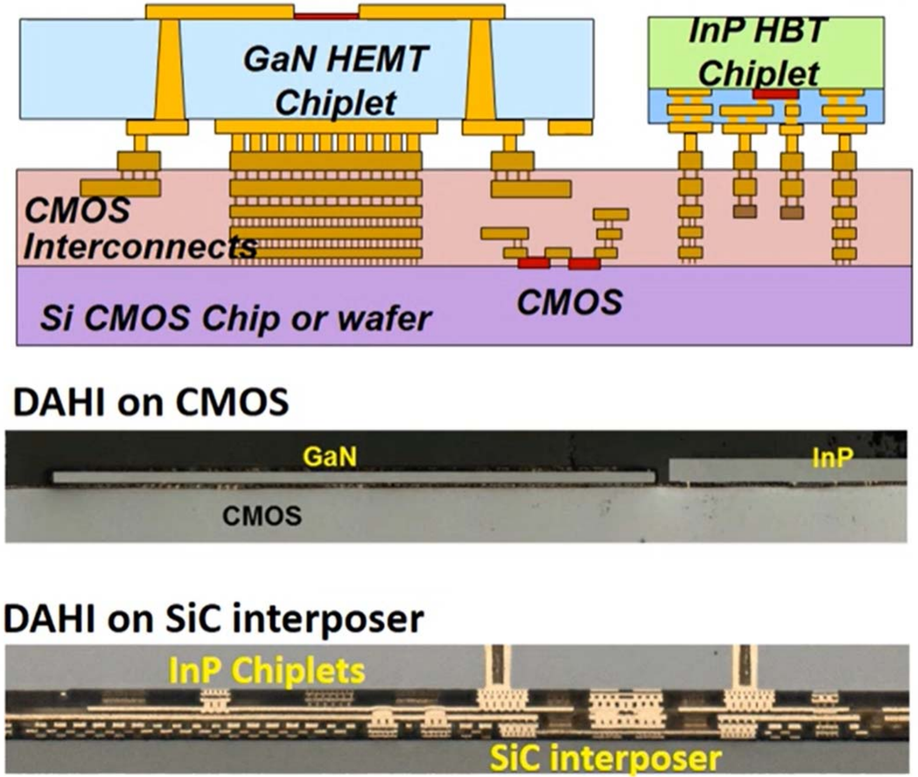


T. Fukushima and M. Koyanagi *et al*., *IEDM*, p.439 (2007)

Various Kinds of Heterogeneous Integration

- **Heterogeneous Integration with Non-Si devices**
- **Heterogeneous Integration with Sensor/MEMS**
- **Heterogeneous Integration with Photonics/Optics**
- **Heterogeneous Integration with Bionics**
- **etc.**

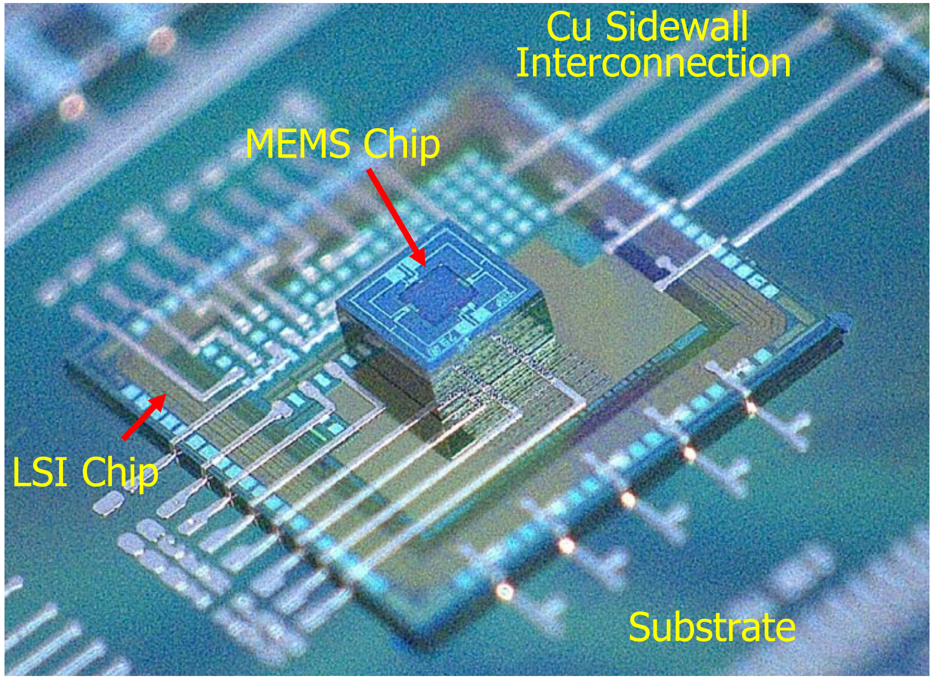
Heterogeneous Integration with Compound Semiconductor Chip (DAHI)



DAHI on CMOS

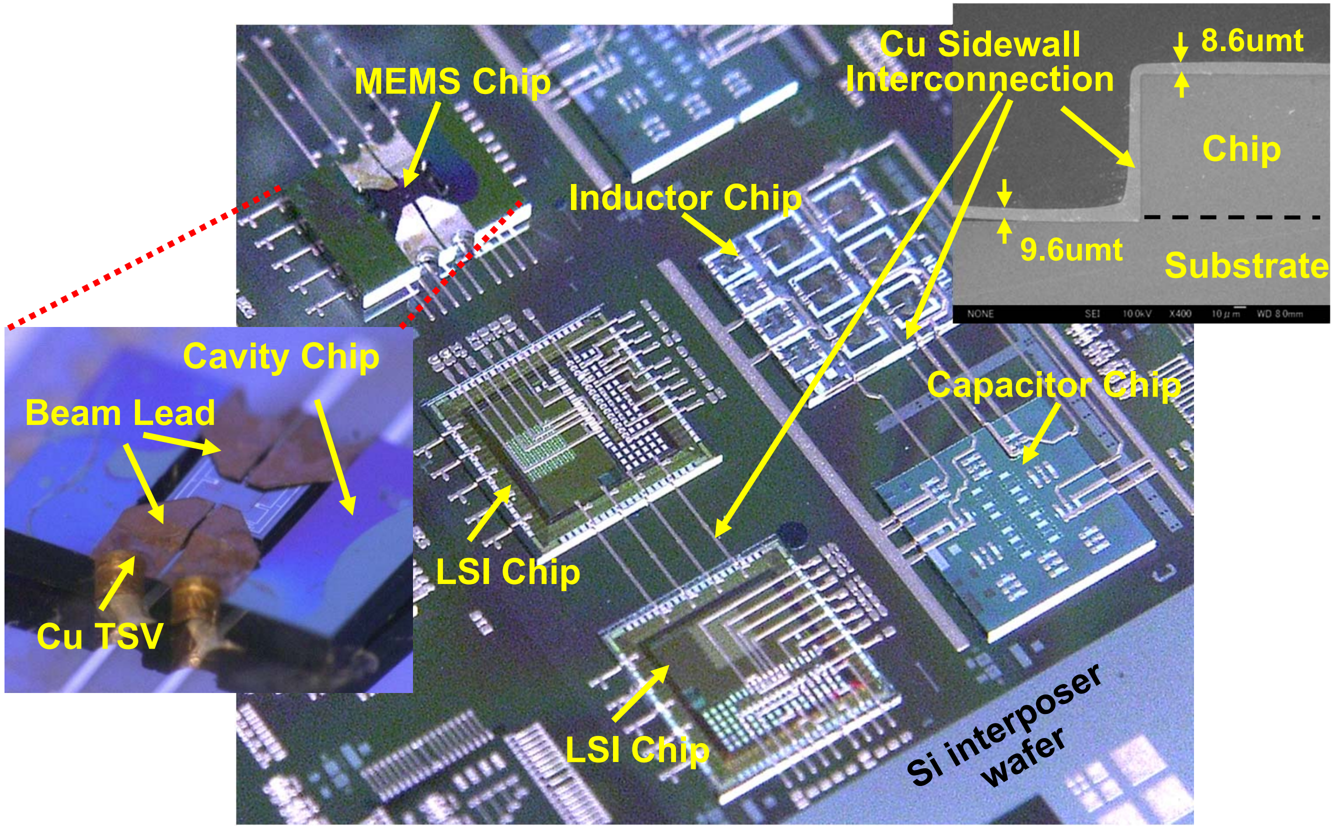
DAHI on SiC interposer

3D Heterogeneous Integration with MEMS Using Self-Assembly



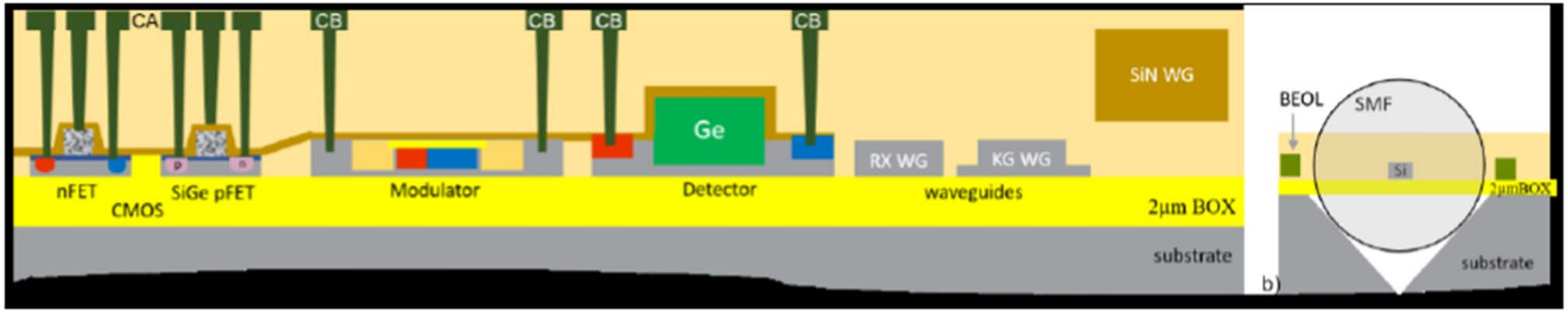
K-W Lee, M. Koyanagi et al., 3D-IC, Sept. 28, 2009

2.5D/3D Heterogeneous Integrations of CMOS, MEMS and Passive Device Chips on Si Substrate



Electronic-Photonic Systems-on-Chip for Compute, Communications and Sensing

45SPCLO process

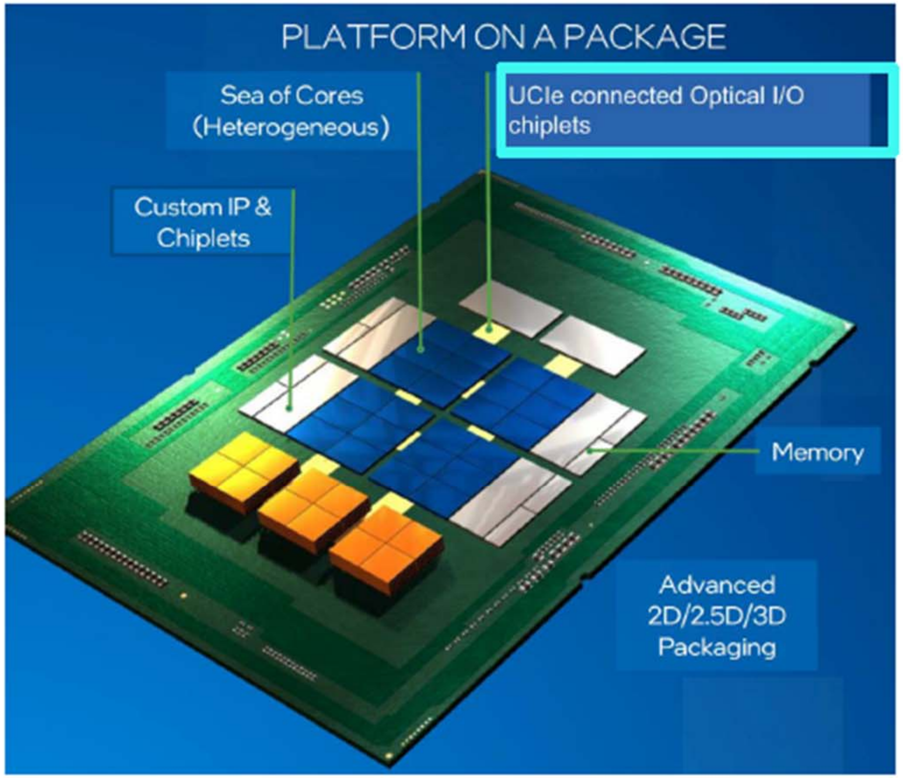


Rakowski *et al* OFC 2020

- Same transistors as in 45nm SOI
- Number of features optimized for photonics
  - Ge photodetectors, Si dopings, Si partial etch, SiN, V-groove couplers etc.

Vladimir Stojanović, IEEE ISSCC Forum 6.8 (2024)

Future Systems-In-Package with Optical I/O

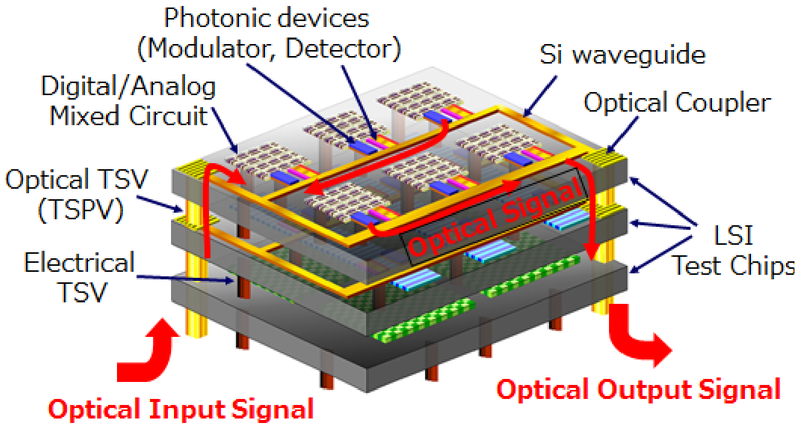

PLATFORM ON A PACKAGE
Sea of Cores (Heterogeneous)
UCIe connected Optical I/O chiplets
Custom IP & Chiplets
Memory
Advanced 2D/2.5D/3D Packaging

Gen	Electrical I/F (Advanced Package)				Optical I/F (CW-WDM)			Optical Chiplet BW (Tx+Rx)	Off-package IO BW (4-8 chiplets per package)
	I/F	Modules	Tx / Rx IOs	Data Rate [Gbps/IO]	Ports	λs / Port	Data Rate [Gbps/λ]		
1	AIB	24	20 / 20	2	8	8	16	2 Tbps	8-16 Tbps
2	AIB	16	80 / 80	2	8	8	32	4 Tbps	16-32 Tbps

- Gen 1 and Gen 2 already built and hardware validated
- 16-32 Tbps off-socket optical I/O bandwidth possible today

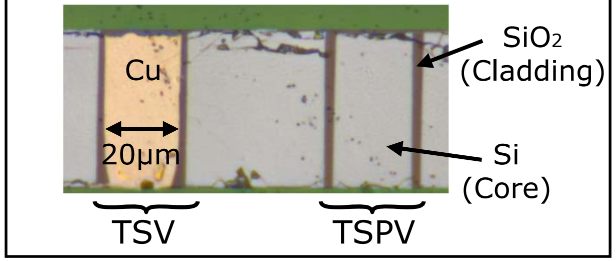
(Source: Wade *et al* HotChips 2023)

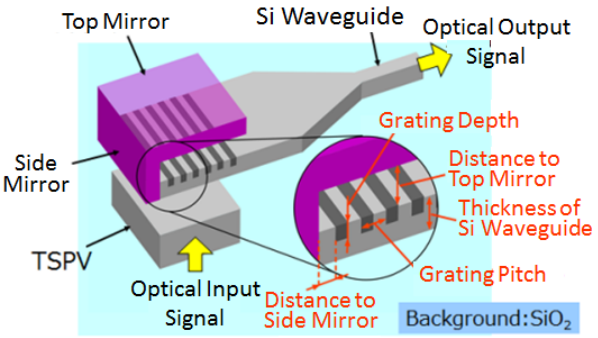
3D Heterogeneous Integration with Photonics
(Photonic 3D Integration)

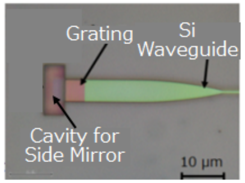


Vertical optical interconnection (TSPV)

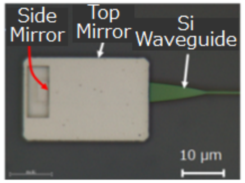
- TSPV (Through Si photonic via)



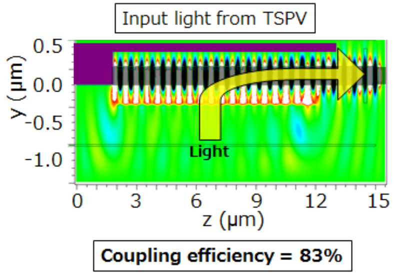




Before mirror formation

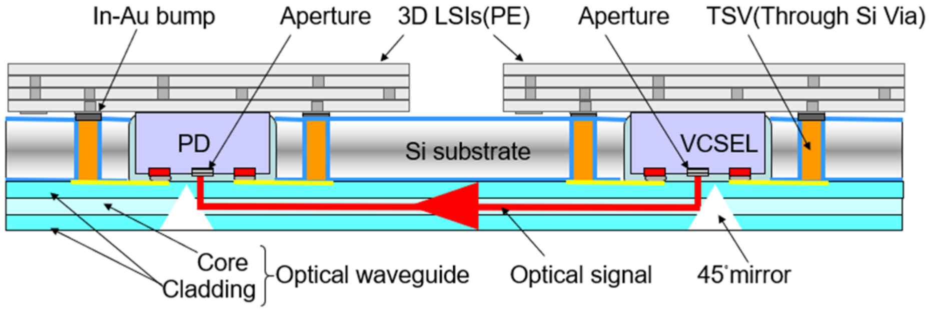


After mirror formation

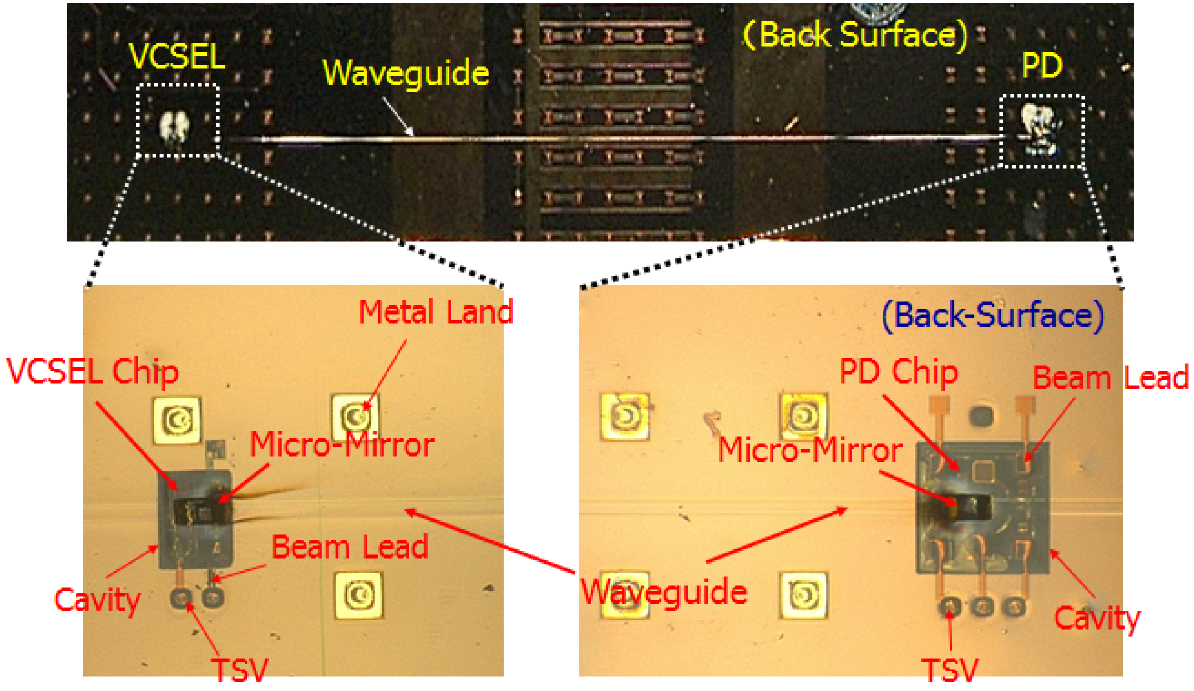


K-W Lee, M. Koyanagi et. al., IEEE Trans. on Electron Devices, vol.58, p.748 (2011)

Photonic 2.5D/3D Heterogeneous Integration
(Optical Interposer Embedded with VCSEL/PD Chips)

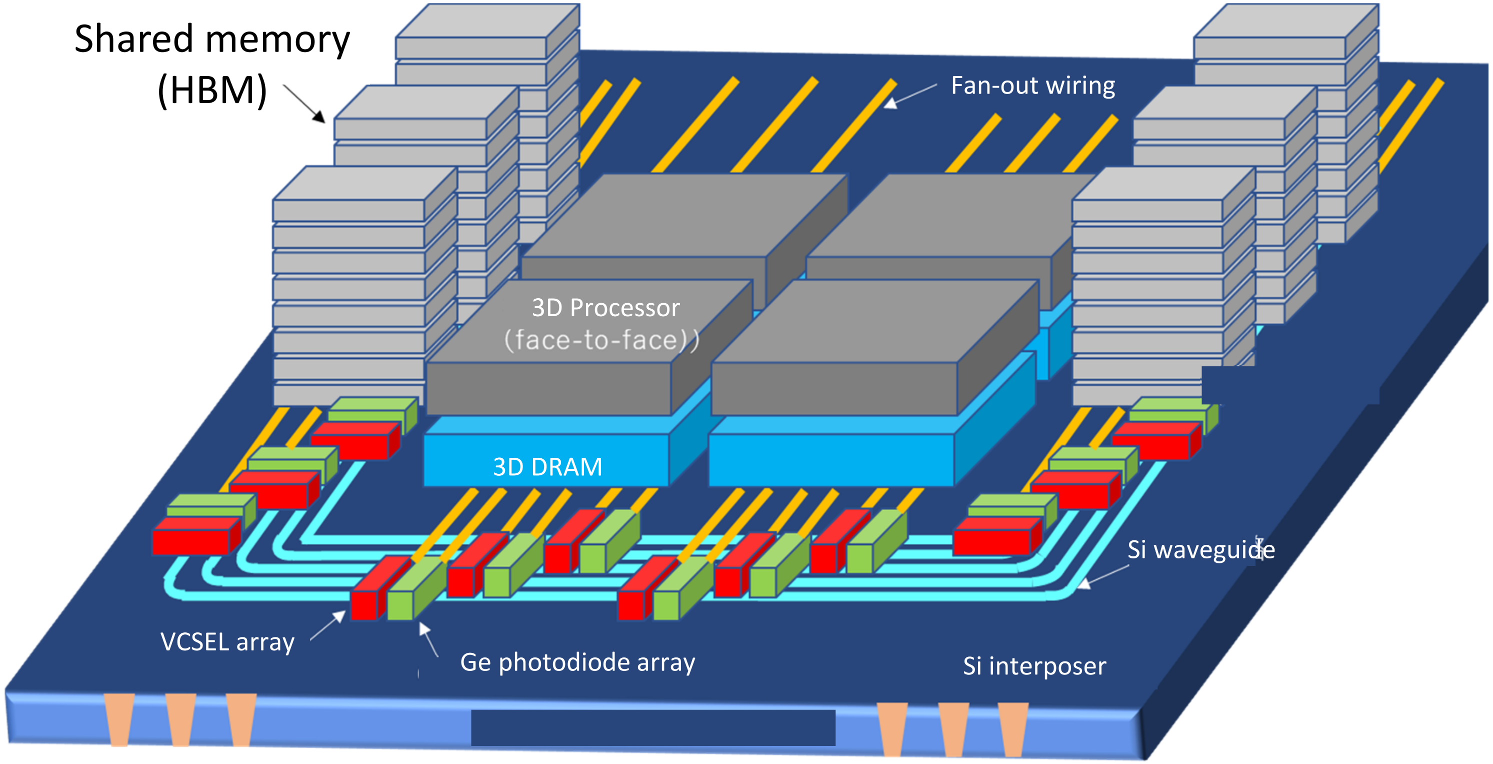


(Photograph of optical interposer captured from back surface)

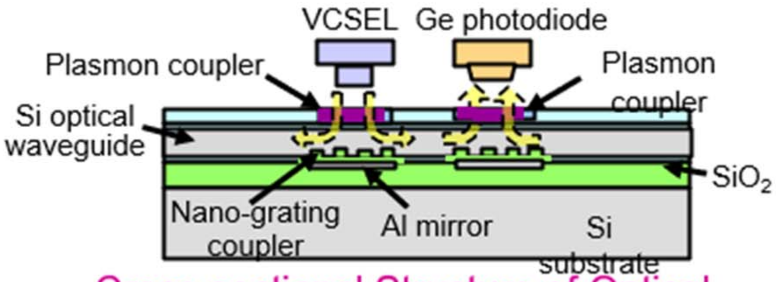


A.Noriki, M. Koyanagi et. al., Jpn. Jour. Appl. Phy. Vol. 48, p.C113-1 (2009)

MEC System Module with Optical Interconnection by Heterogeneous Chiplet Integration



Fabrication Flow of Optical Interconnection with Grating Coupler and Plasmon Coupler



Cross-sectional Structure of Optical Interconnection with Grating Coupler and Plasmon Coupler

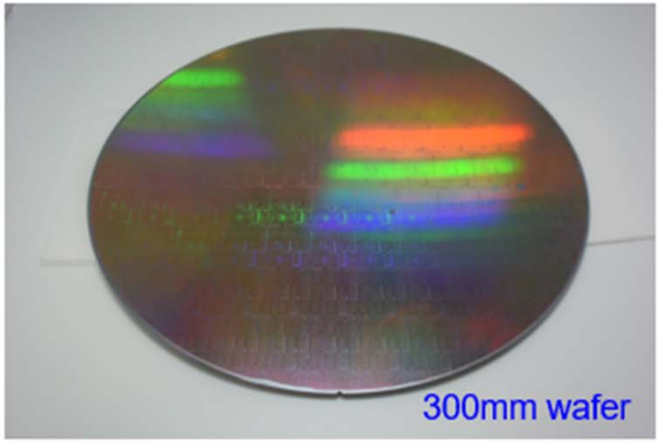
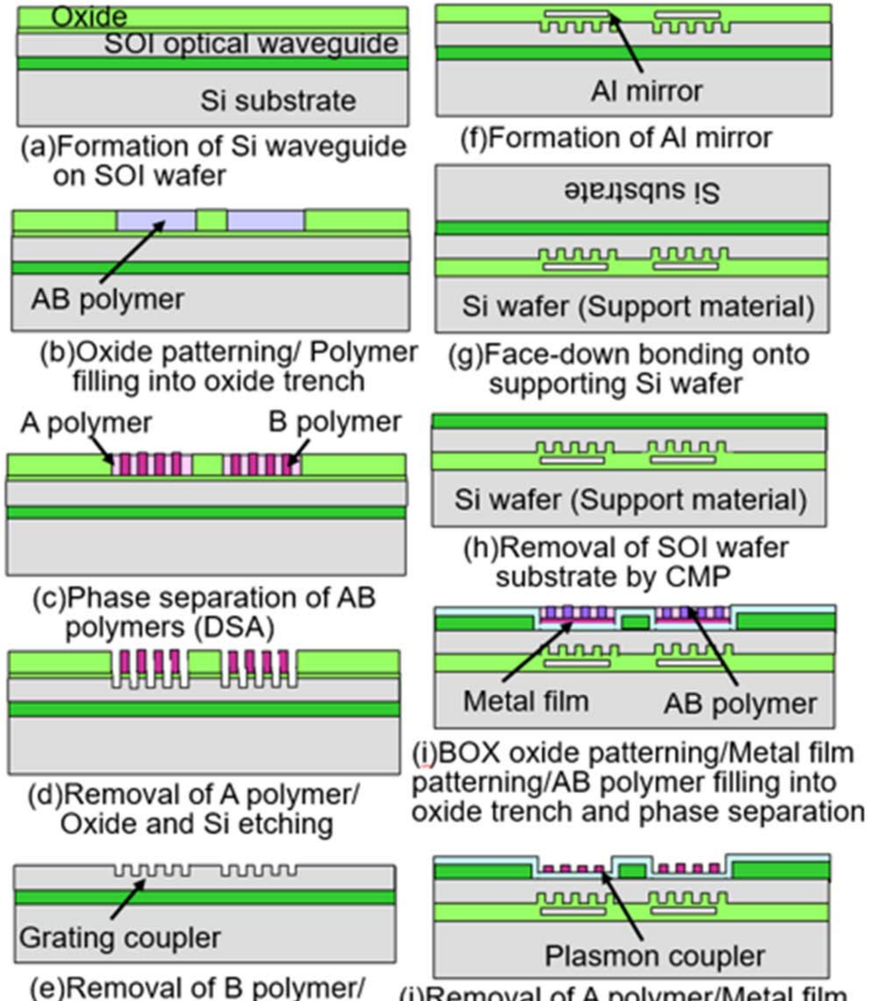


Photo after transferring Si optical waveguide patterns onto a Si interposer wafer



(a)Formation of Si waveguide on SOI wafer

(b)Oxide patterning/ Polymer filling into oxide trench

(c)Phase separation of AB polymers (DSA)

(d)Removal of A polymer/ Oxide and Si etching

(e)Removal of B polymer/ Oxide etching

(f)Formation of Al mirror

(g)Face-down bonding onto supporting Si wafer

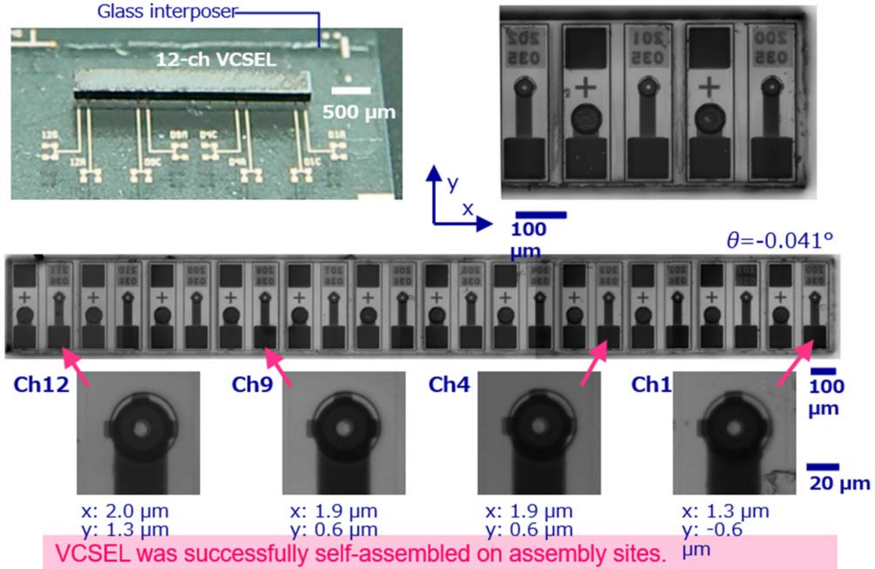
(h)Removal of SOI wafer substrate by CMP

(i)BOX oxide patterning/Metal film patterning/AB polymer filling into oxide trench and phase separation

(j)Removal of A polymer/Metal film etching/Removal of polymer B

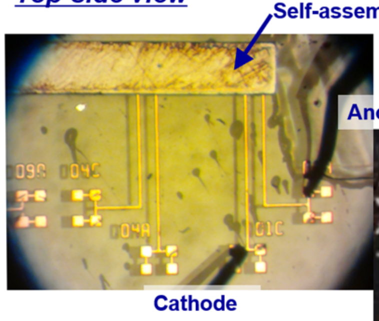
VCSEL Chiplet Integration on Glass Interposer by Self-Assembly

Self-Assembly of 12-ch VCSEL Chip on Glass Interposer



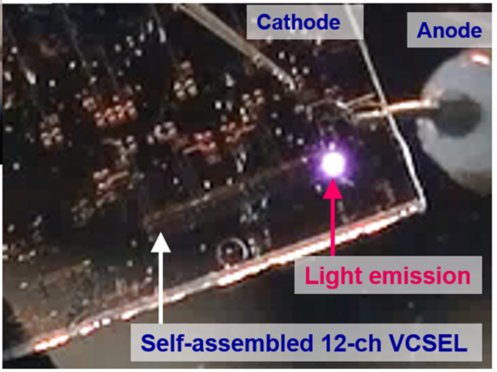
VCSEL was successfully self-assembled on assembly sites.
The alignment accuracies of all channels were within 2.0 µm.

Top-side view



Applied voltage : 2 V
Measured current value : 5.0 mA

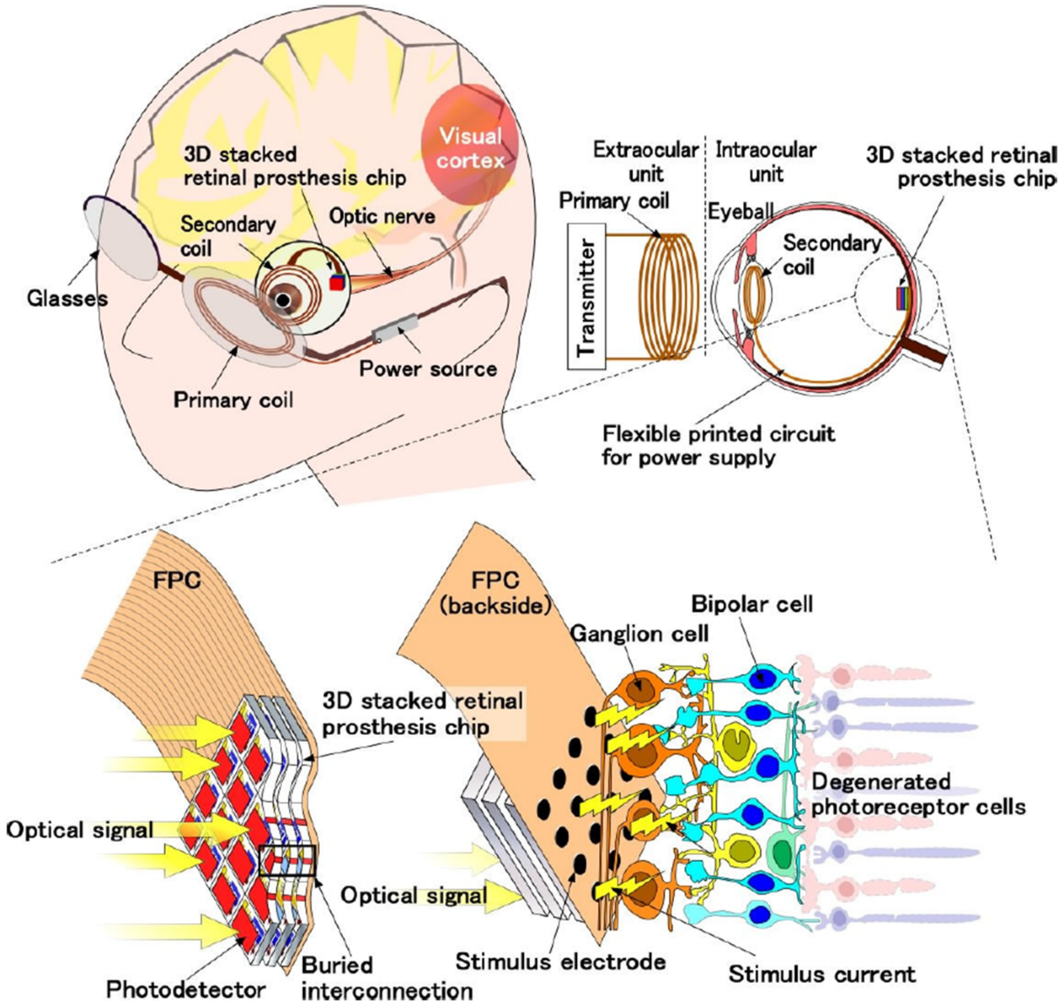
Bottom-side view

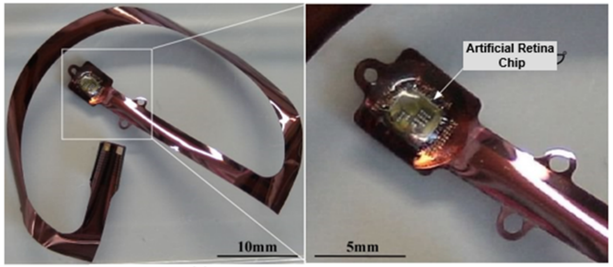


3D Heterogeneous Integration with Bionics

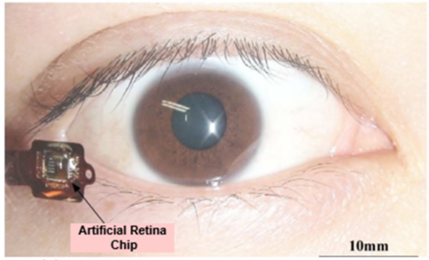
(Bionic 3D Integration)

3D-Retina Chip Implantation into Human Eye (Retinal Prothesis)





(a) Retinal Prosthesis Module

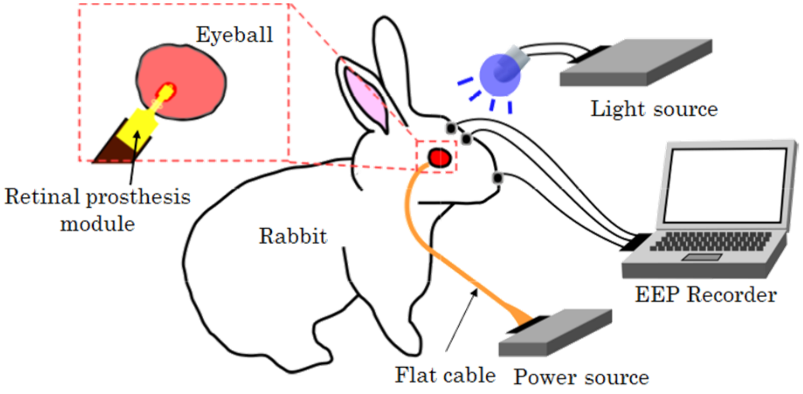


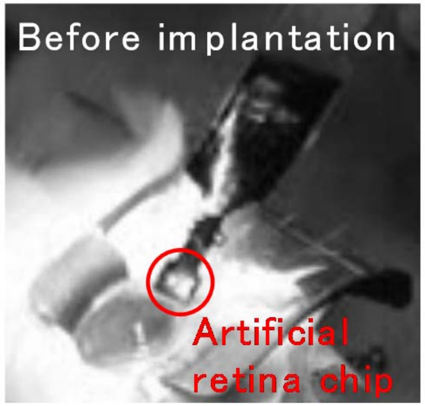
(b) Retina Prosthesis Module and Human Eye

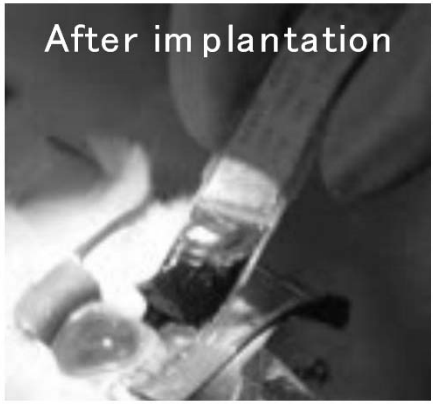
T. Tanaka, M. Koyanagi et. al.,
IEEE IEDM, p.1015 (2007)

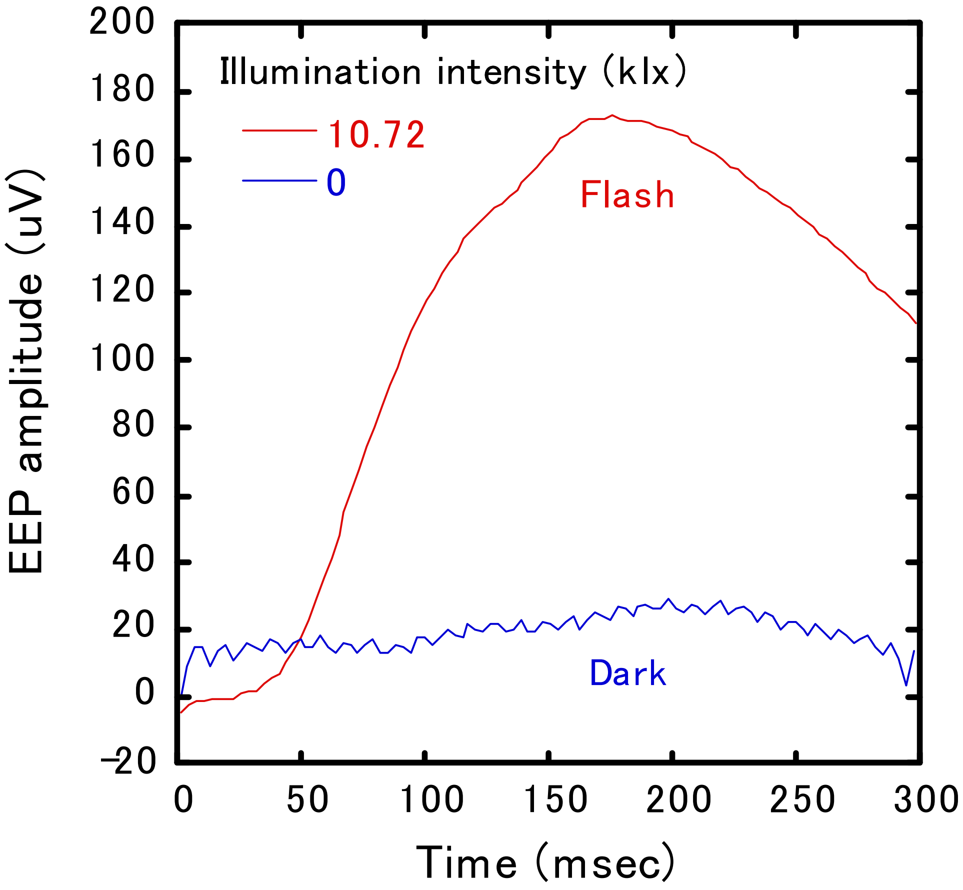
EEP Waveforms with/without Flashlight

Animal experiment setup




Before implantation

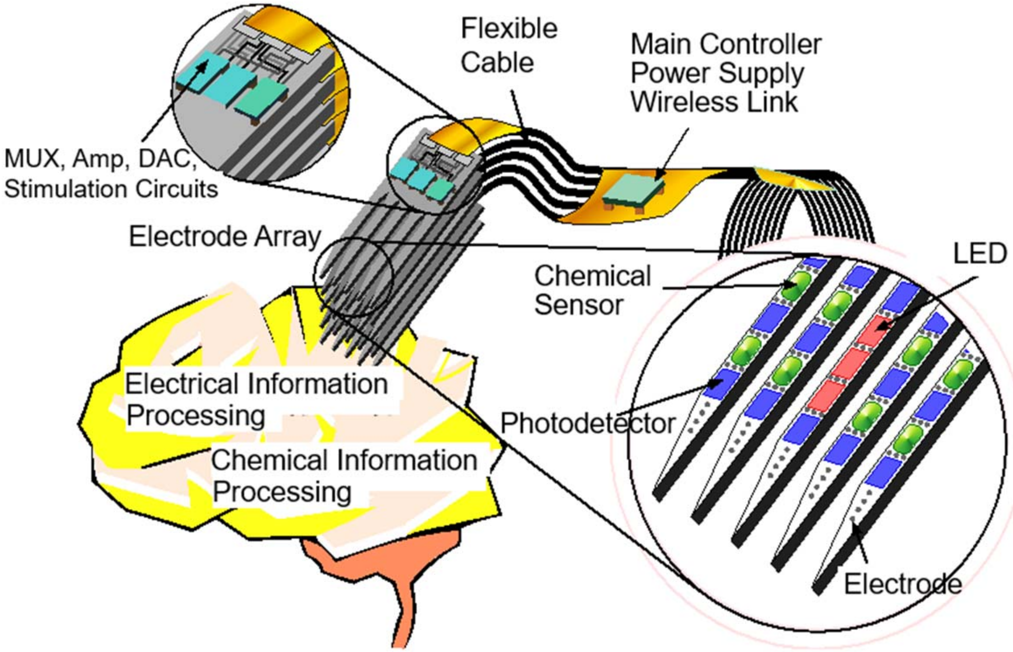

After implantation

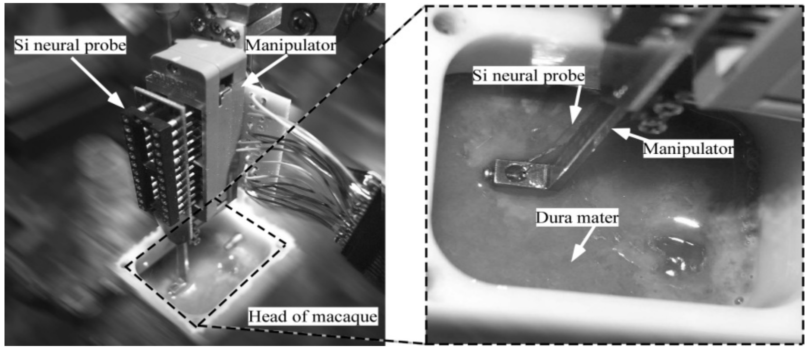


3D Heterogeneous Integration with Bionics

(Bionic 3D Integration)

Brain-Machine Interface (BMI) and Intelligent Si Neural Probe with Multi-electrodes and Sensors



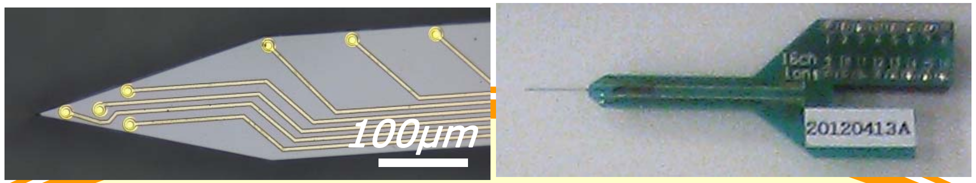


Recording of Neuron Potential in a Brain Using Si Neural Probe

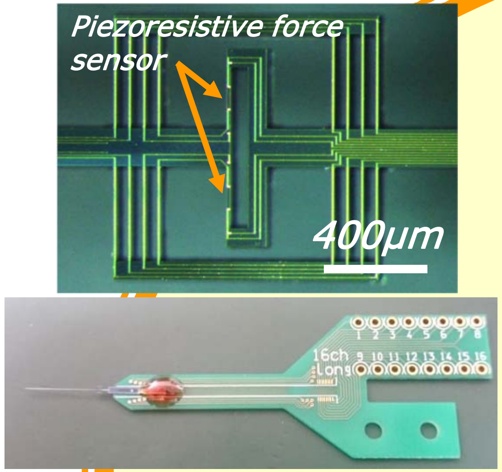
S. Kanno, T. Tanaka, M. Koyanagi et. al., Jpn. Jour. Appl. Phy., Vol. 48, p. C189-1 (2009)

Tohoku Univ. Intelligent Neural Probe family

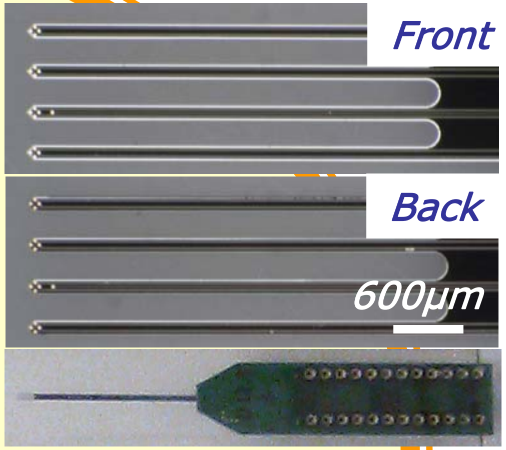
(Prof. Tetsu Tanaka)



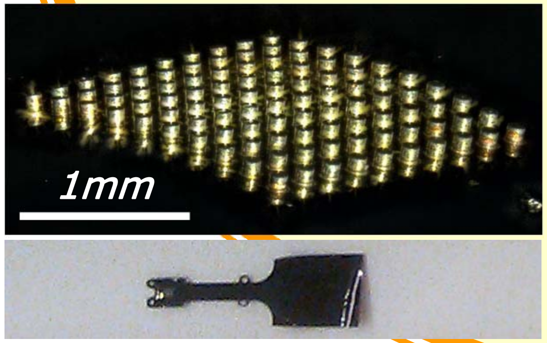
Si neural probe mounted on PCB



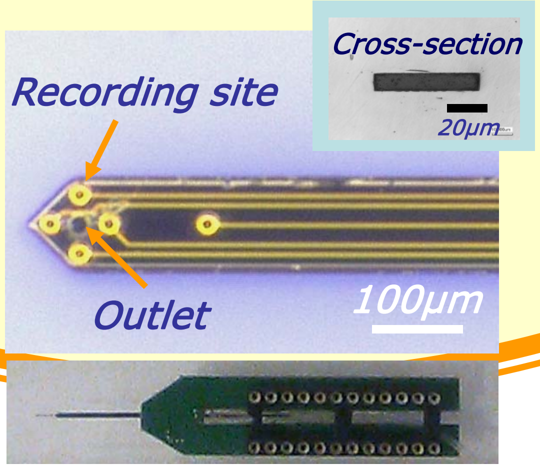
Si neural probe with piezoresistive force sensor



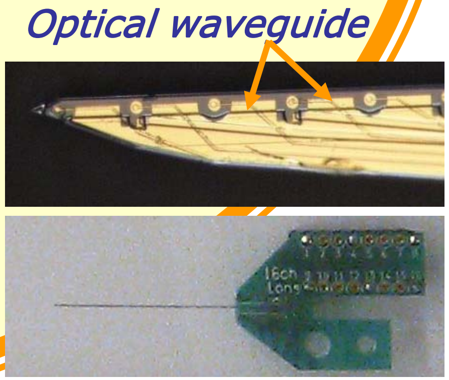
4-shunk double-sided Si neural probe



Pillar-type electrode array (10x10)



Si neural probe with microfluidic channel



Si neural probe with optical waveguides

Neural stimulation

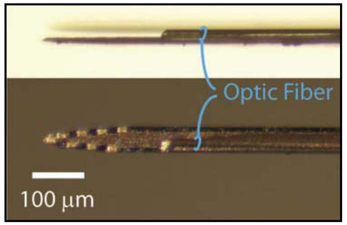
	Electrical	Chemical	Optical
Stimulus speed	Fast ☺	Slow	Fast ☺
Neural activities	Excitation	Excitation Inhibition ☺	Excitation Inhibition ☺
Cell selectivity	No	No	Yes ☺

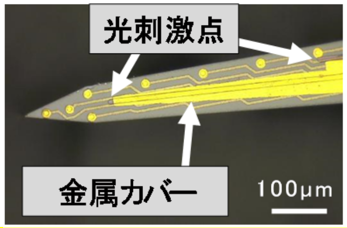
■Optogenetics

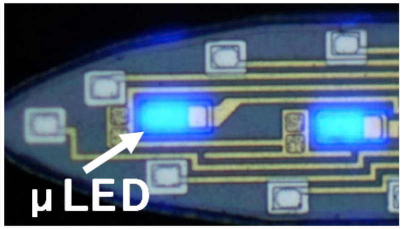
Gene transfer of a protein molecule that responds to light of a specific wavelength enables the control of neuronal excitation and inhibition by light.

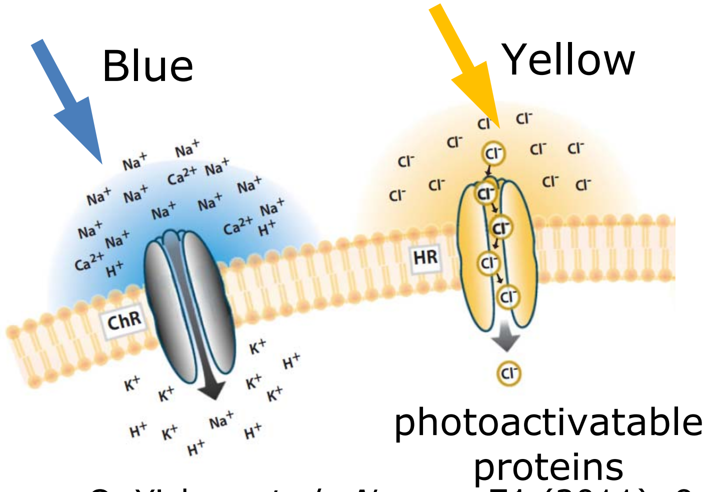
■Neural probe w/ light control func.

Optical fiber/Optical waveguide/μLED









O. Yizhar *et al., Neuron,* **71** (2011), 9-34

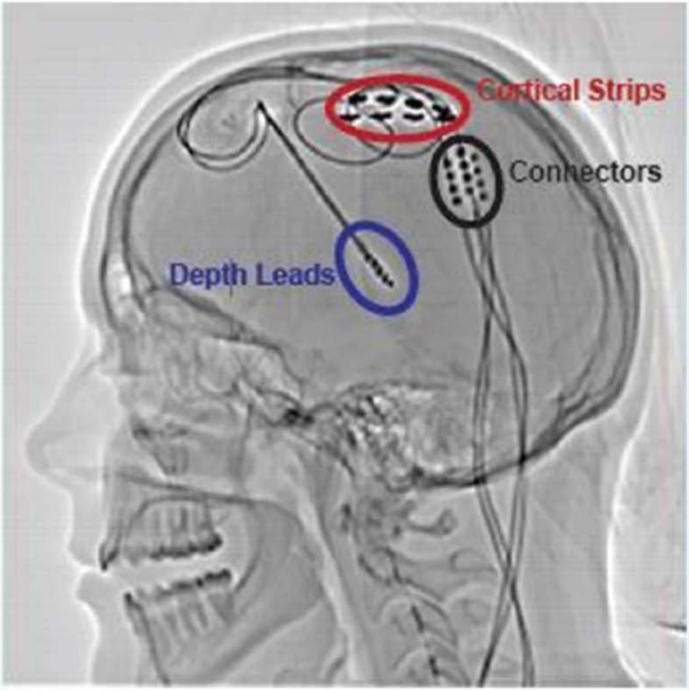
Realization of neural stimulation with high spatial and temporal resolution and cell selectivity.

By Courtesy of Prof. Tetsu Tanaka (Tohoku University)

3D Heterogeneous Integration with Bionics

(Bionic 3D Integration)

Review Bi-directional Brain Interface from Risk Management Perspective



UF UNIVERSITY of FLORIDA | J. Crayton Pruitt Family Department of Biomedical Engineering

Brain Mapping Laboratory

Closing the Loop on Tremor: A Responsive Deep Brain Stimulator for the Treatment of Essential Tremor

PIs: Aysegul Gunduz, Kelly Foote D.

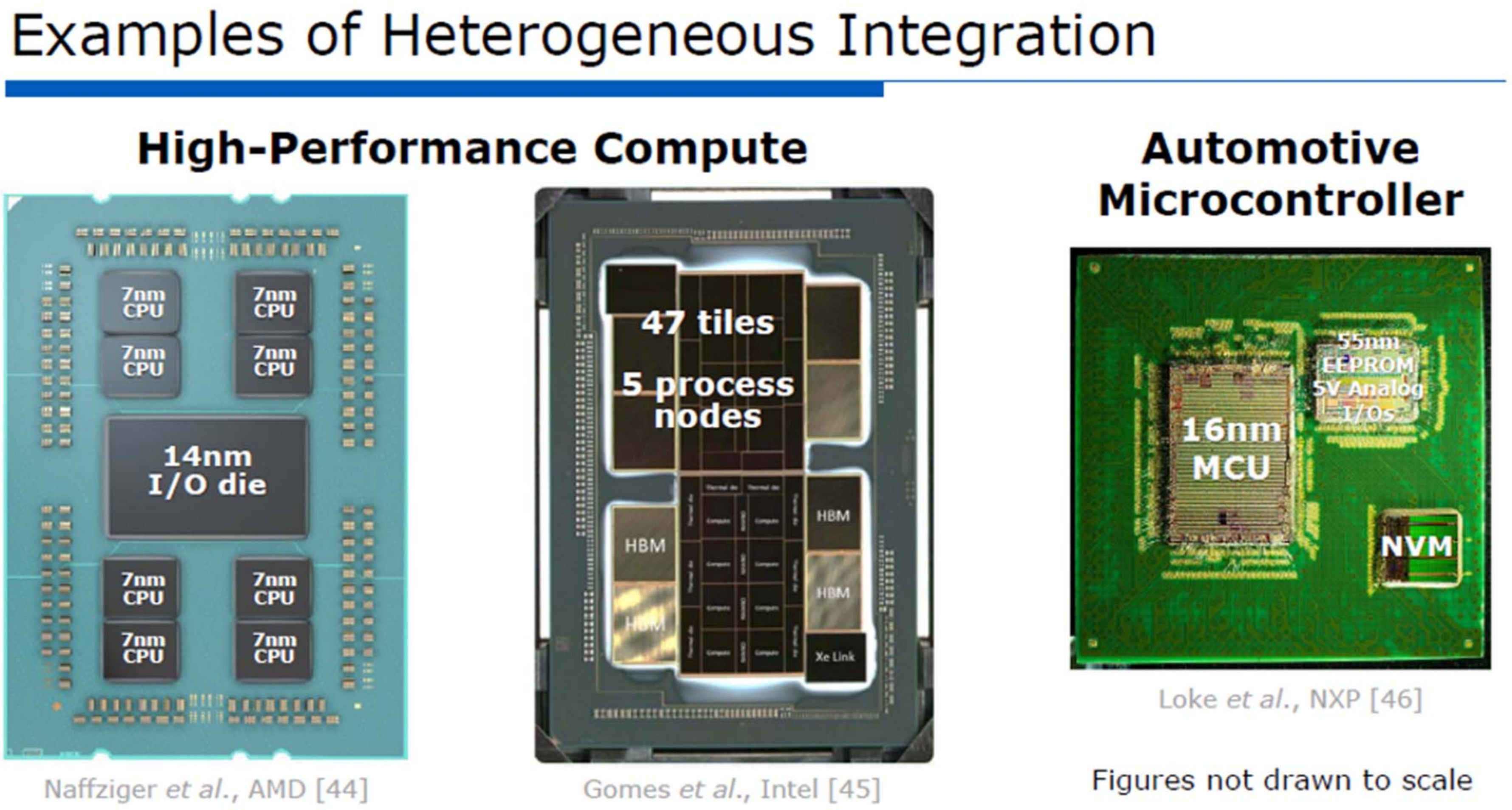
Analysis: Enrico Opri

University of Florida, Gainesville, FL, United States

UF UNIVERSITY of FLORIDA Opri et al, Chronic embedded cortico-thalamic closed-loop deep brain stimulation for the treatment of essential tremor, Science Translational Medicine, 2020

Tim Denison, IEEE ISSCC, Forum 4.1 (2024)

Architecture/System Level Heterogeneous Integration



Examples of Heterogeneous Integration

High-Performance Compute

Naffziger *et al.*, AMD [44]

Gomes *et al.*, Intel [45]

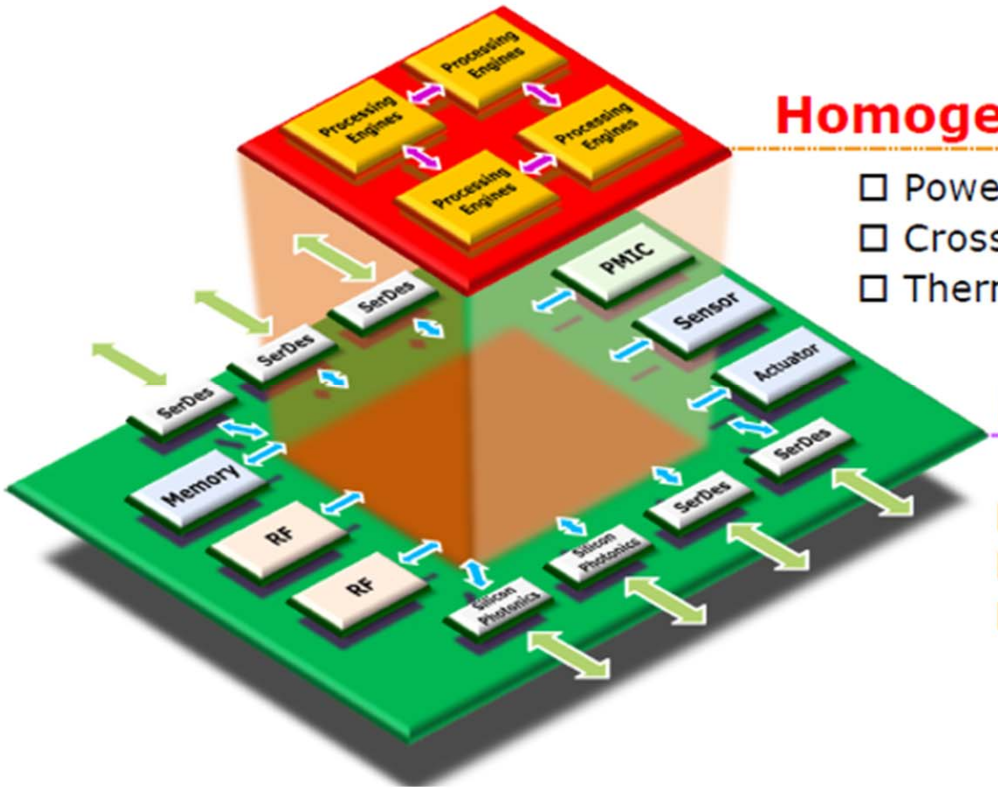
Automotive Microcontroller

Loke *et al.*, NXP [46]

Figures not drawn to scale

Alvin Loke, IEEE ISSCC, Forum 3.1 (2024)

Architecture/System Level Heterogeneous Integration



Homogeneous Integration

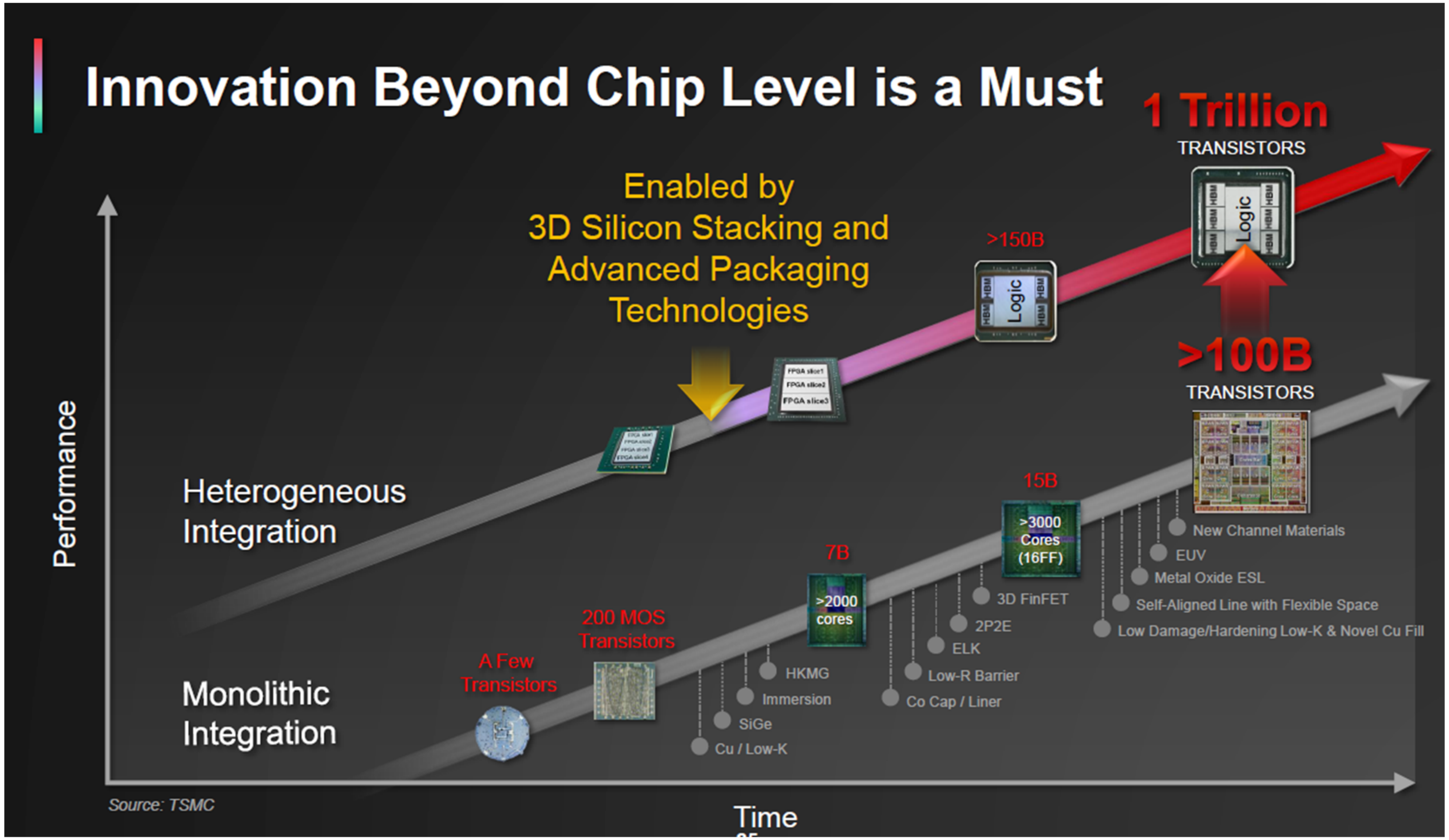
- □ Power, performance, area and chip cost
- □ Cross-IP data link and latency
- □ Thermal- and IR-aware die floorplanning

Heterogeneous Integration

- □ System cost and time-to-market
- □ Cross-Die data link and latency
- □ Multiphysics-aware chip integration

Lawrence Loh, IEEE ISSCC, Plenary (2020)

Evolution of IC from Device Integration to System Integration

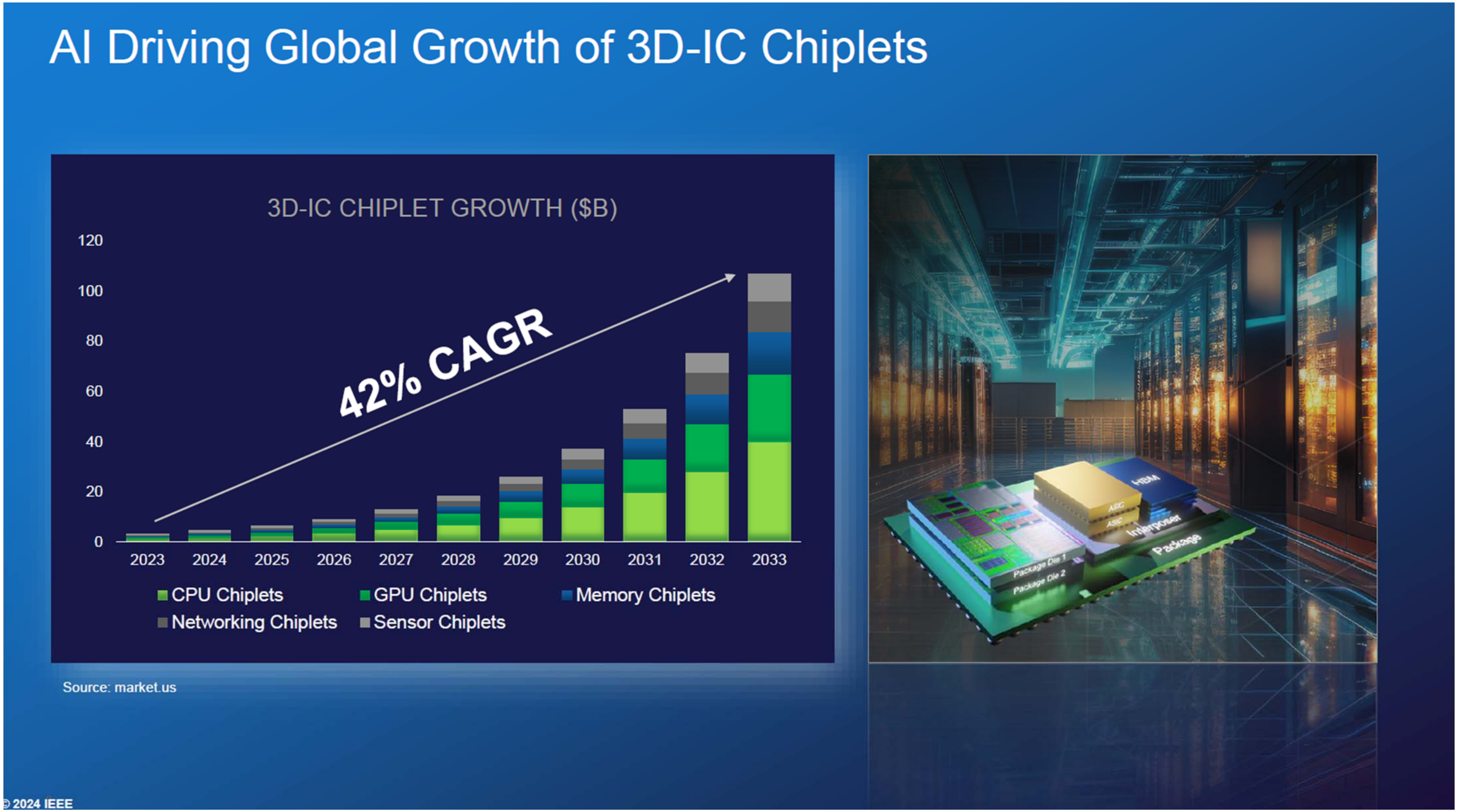

Innovation Beyond Chip Level is a Must

Source: TSMC

Kevin Zhang, IEEE ISSCC, Plenary 1.1 (2024)

Key Technologies for AI

3D Integration, Chiplet Integration, Heterogeneous Integration



AI Driving Global Growth of 3D-IC Chiplets

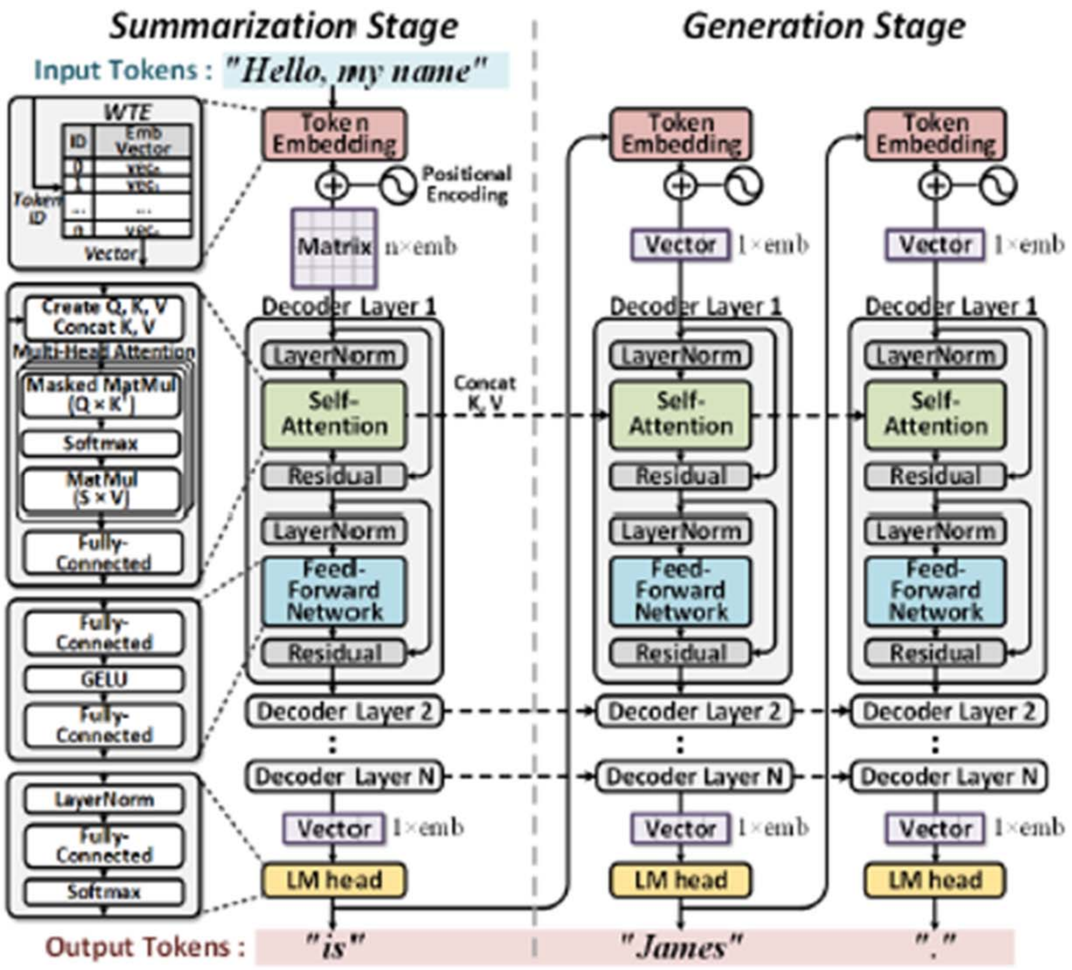
3D-IC CHIPLET GROWTH ($B)

Source: market.us

© 2024 IEEE

Lip-Bu Tan, IEEE ISSCC, Plenary 1.4 (2024)

LLM Computations (Training vs. Inference)



Training (model making)

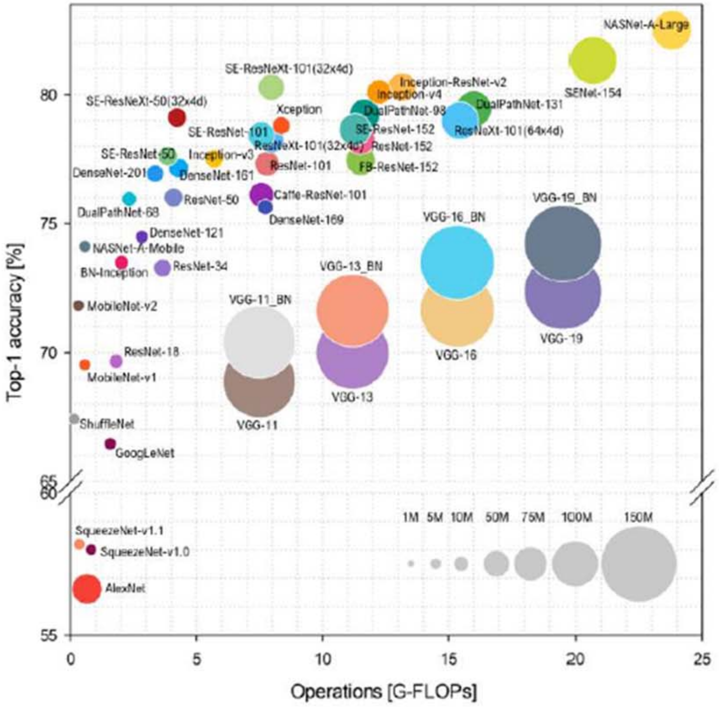
- □ Large batch
- □ Compute-intensive
- □ Throughput-oriented HW

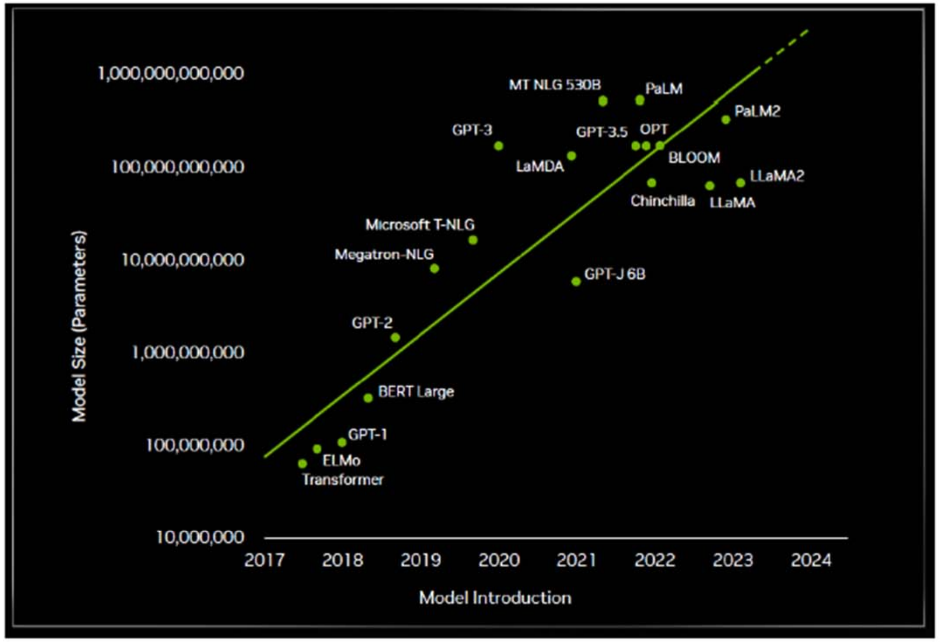
Inference (service)

- □ Small batch
- □ Memory-intensive
- □ Latency-oriented HW

Joo-Young Kim, IEEE ISSCC Forum 2.5 (2024)

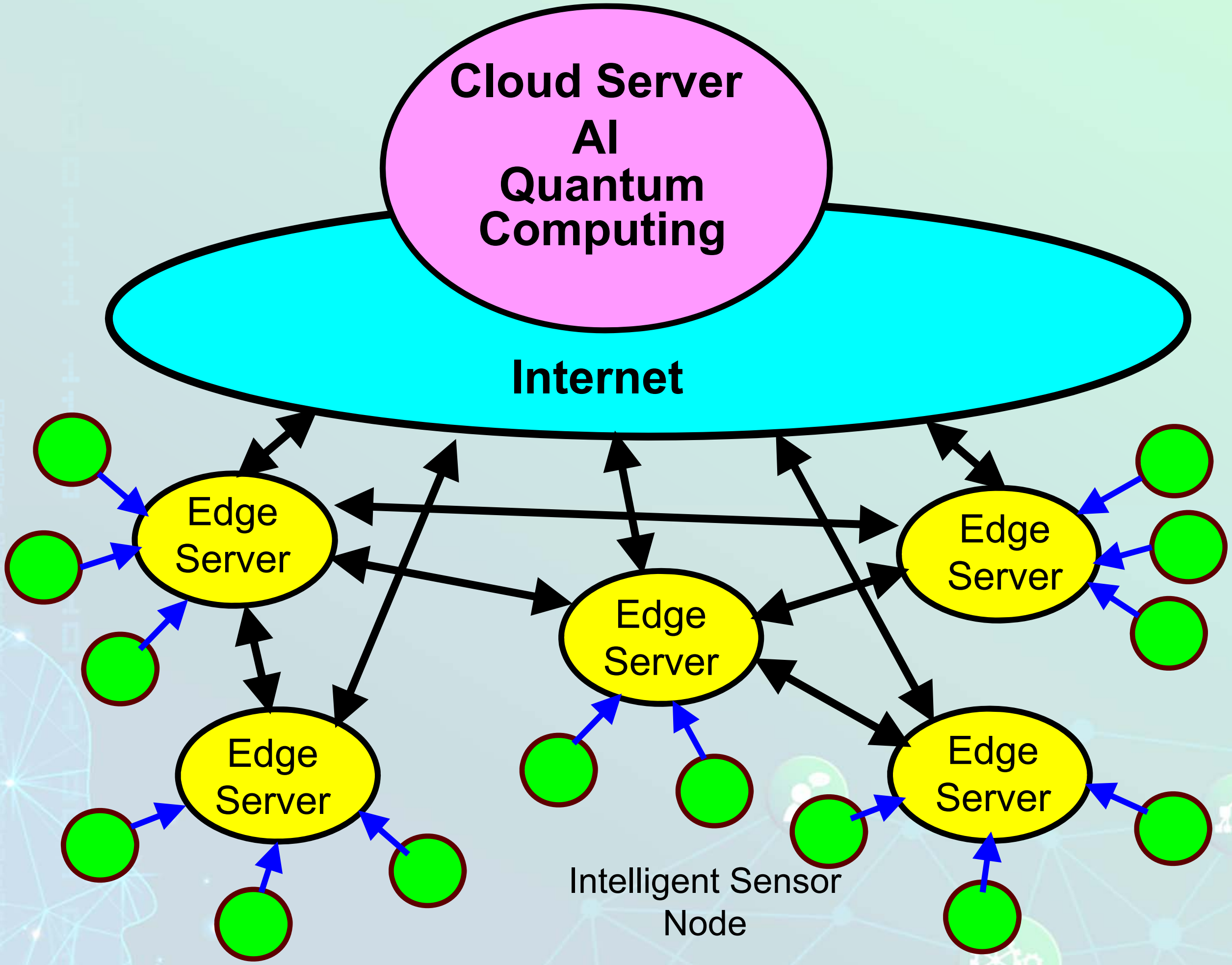
Growth in application complexity



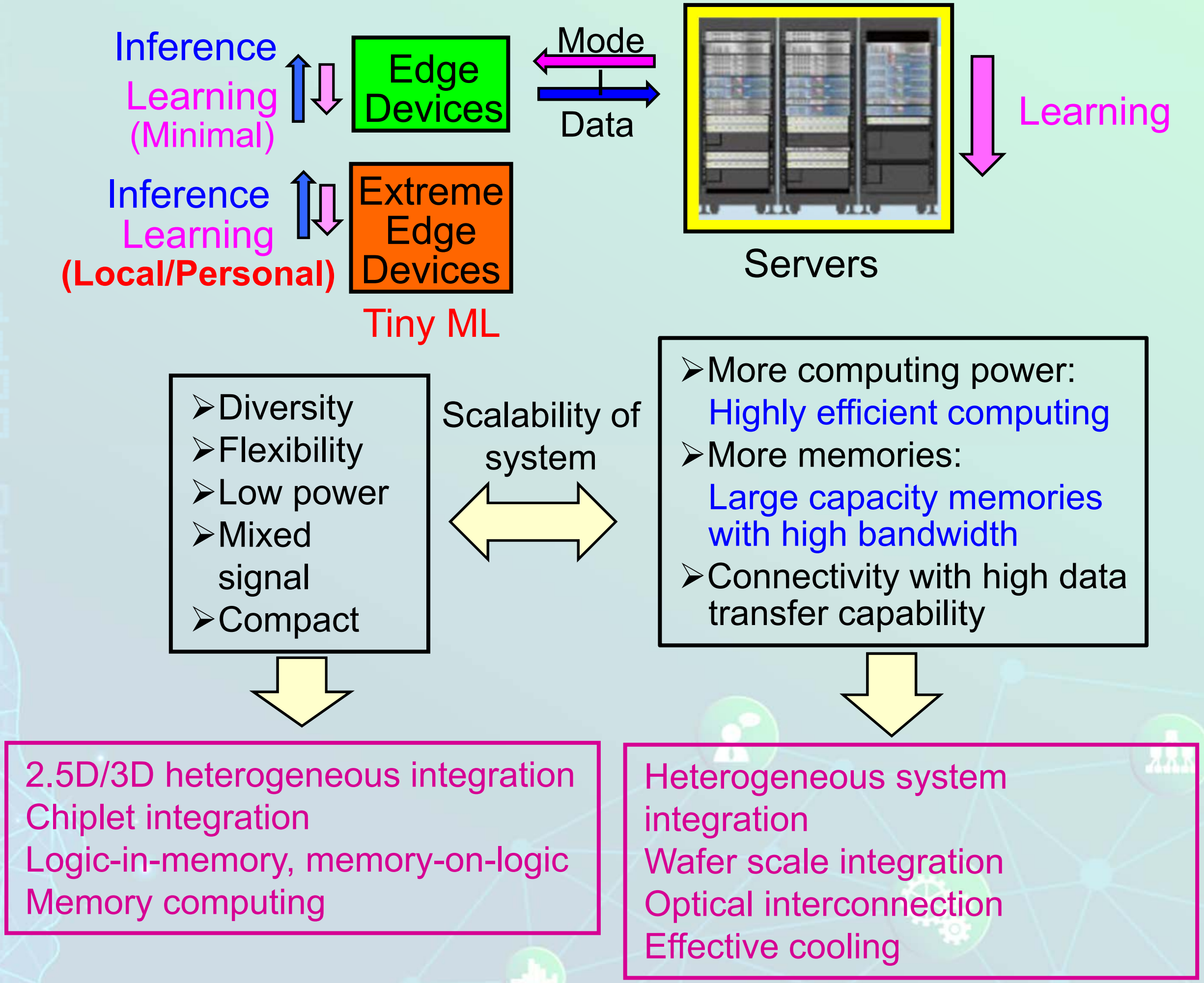


S. Bianco et al., "Benchmark Analysis of Representative Deep Neural Network Architectures," IEEE Access, 2018

Global Network in IoT/AI/5G Era



Requirements for AI System and Technologies



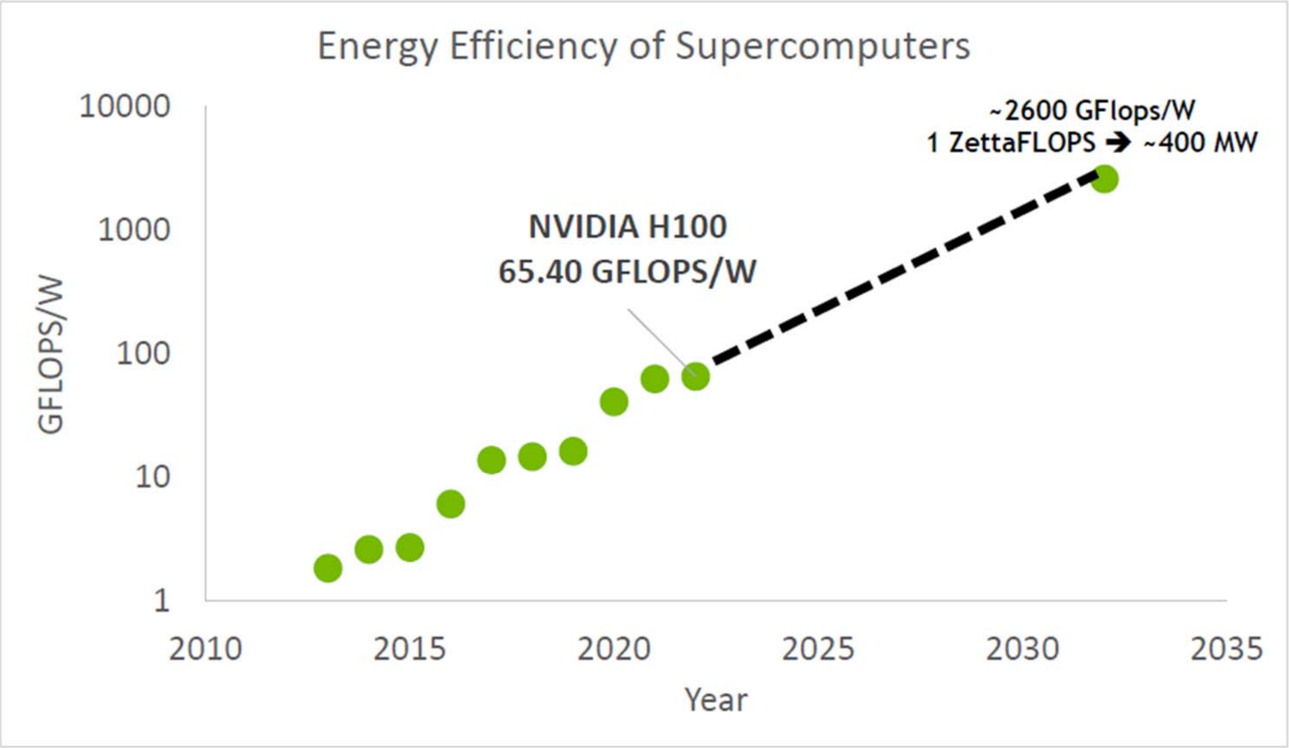
- ➢Diversity
- ➢Flexibility
- ➢Low power
- ➢Mixed signal
- ➢Compact

- ➢More computing power: Highly efficient computing
- ➢More memories: Large capacity memories with high bandwidth
- ➢Connectivity with high data transfer capability

2.5D/3D heterogeneous integration
Chiplet integration
Logic-in-memory, memory-on-logic
Memory computing

Heterogeneous system integration
Wafer scale integration
Optical interconnection
Effective cooling

Energy efficiency challenge



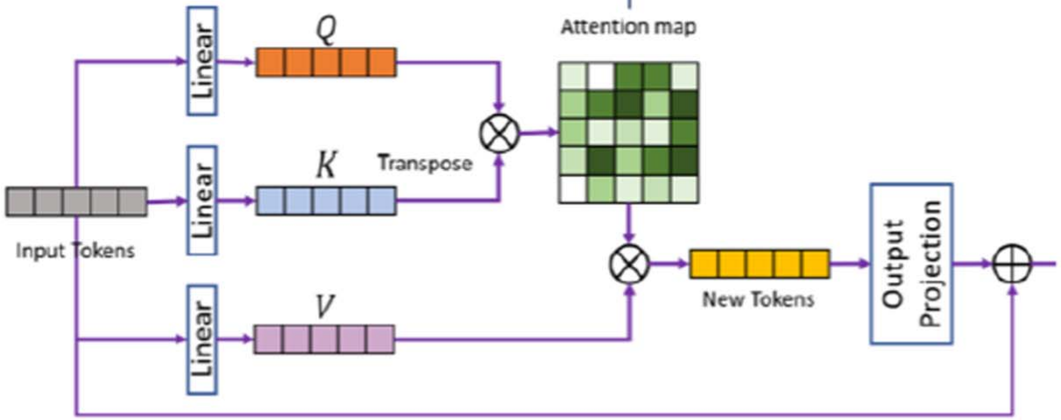
https://www.top500.org/lists/green500/

Rangharajan Venkatesan, IEEE ISSCC SC1 (2024)

Optimizing Networks (for the Edge)

Every token is compared to all other tokens to compute attention map → Quadratic complexity

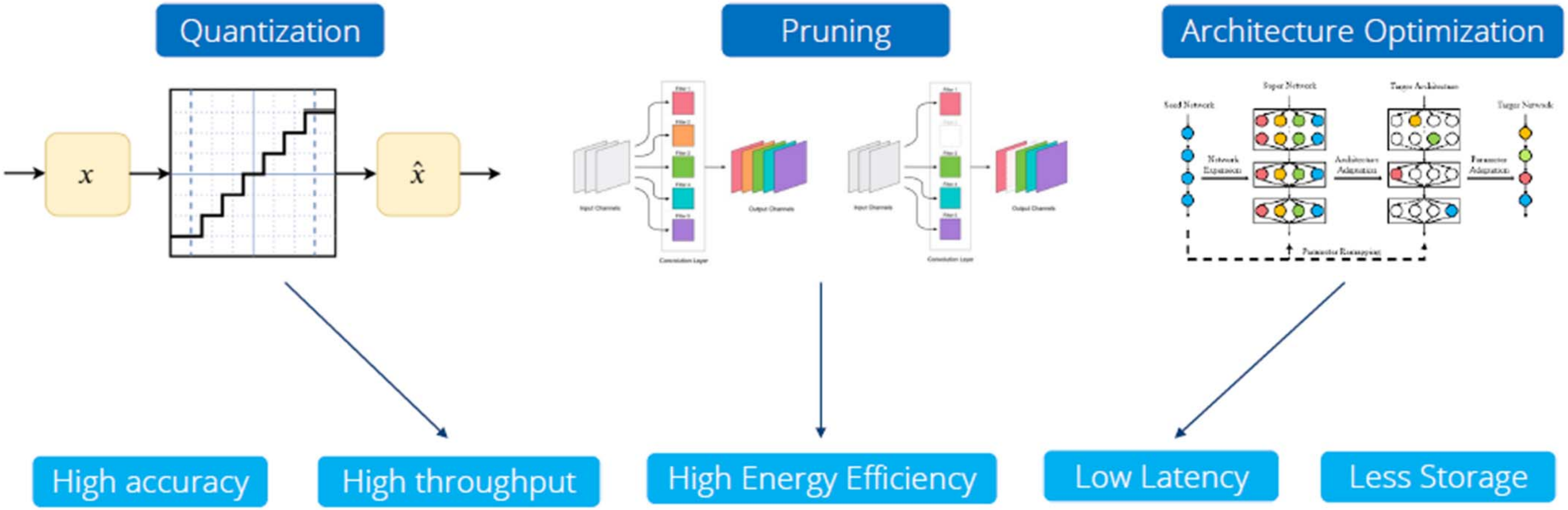
$$attention = softmax(\frac{QK^T}{\sqrt{d_k}})$$



Bram Verhoef, IEEE ISSCC Forum 2.7 (2024)

Optimizing Networks (for the Edge)

Overall Goals



Quantization

Pruning

Architecture Optimization

High accuracy

High throughput

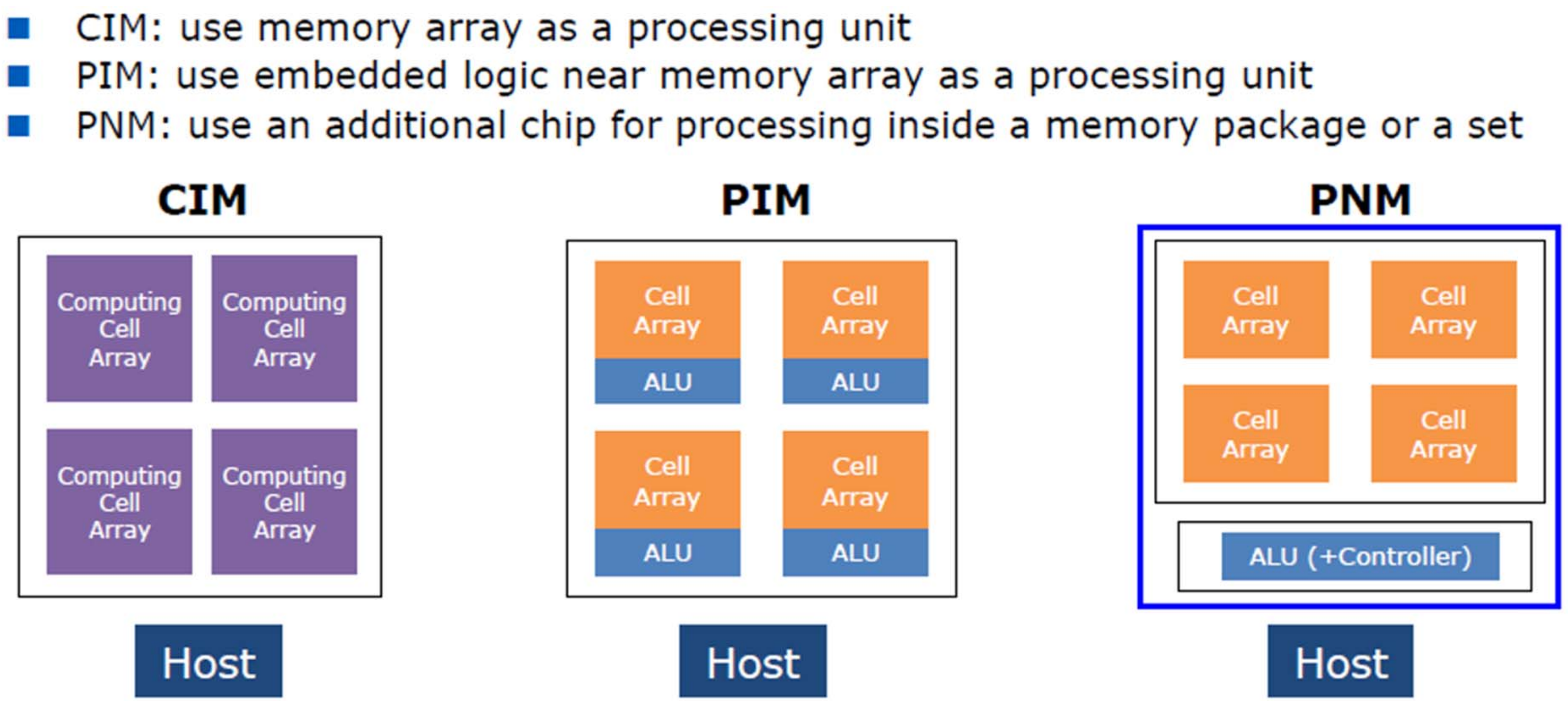
High Energy Efficiency

Low Latency

Less Storage

Bram Verhoef, IEEE ISSCC Forum 2.7 (2024)

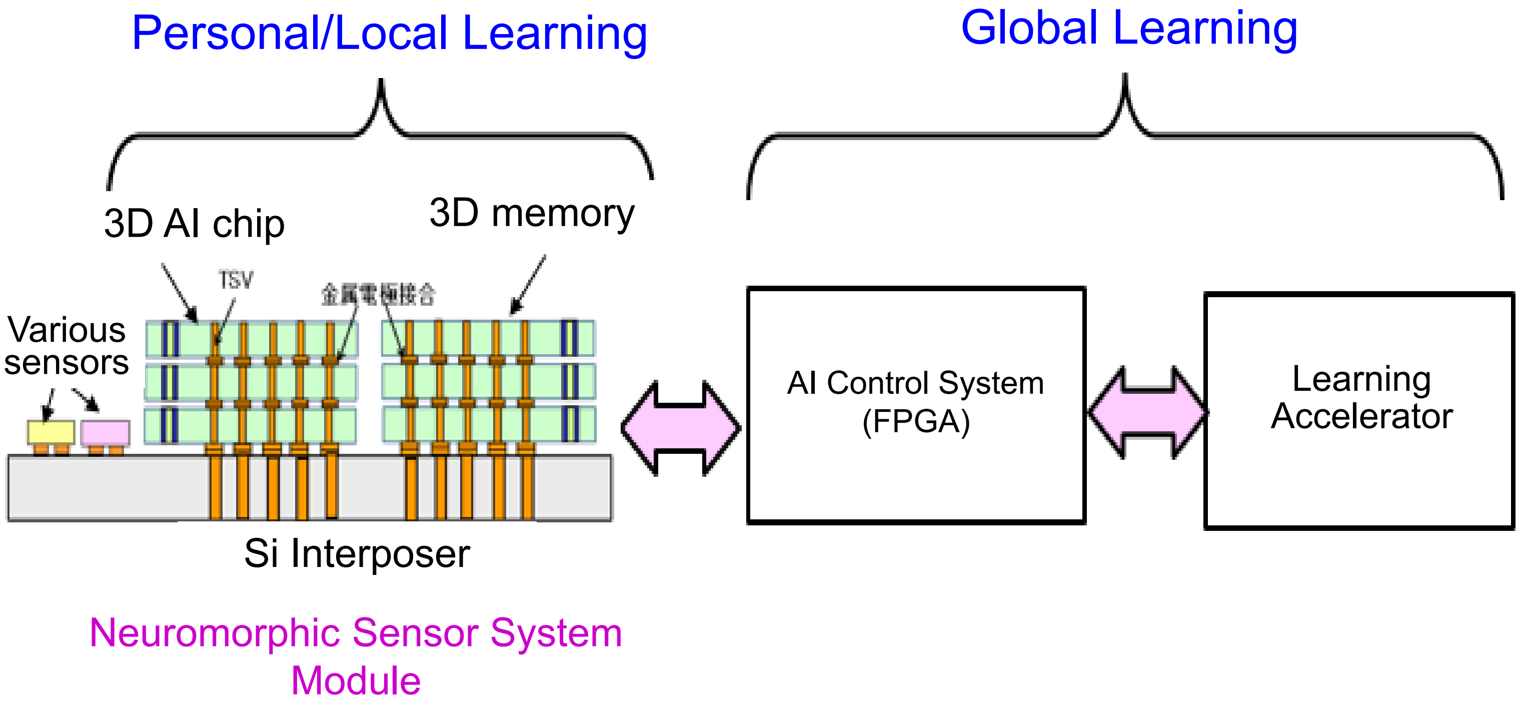
Memory-Base AI Processor to Achieve High Energy Efficiency and Compactness



- CIM: use memory array as a processing unit
- PIM: use embedded logic near memory array as a processing unit
- PNM: use an additional chip for processing inside a memory package or a set

Kyomin Sohn, IEEE ISSCC Forum 2.6 (2024)

Neuro-Centric Sensor System Project in Tohoku Univ.



Personal/Local Learning

Global Learning

3D AI chip

3D memory

TSV

金属電極接合

Various sensors

Si Interposer

Neuromorphic Sensor System Module

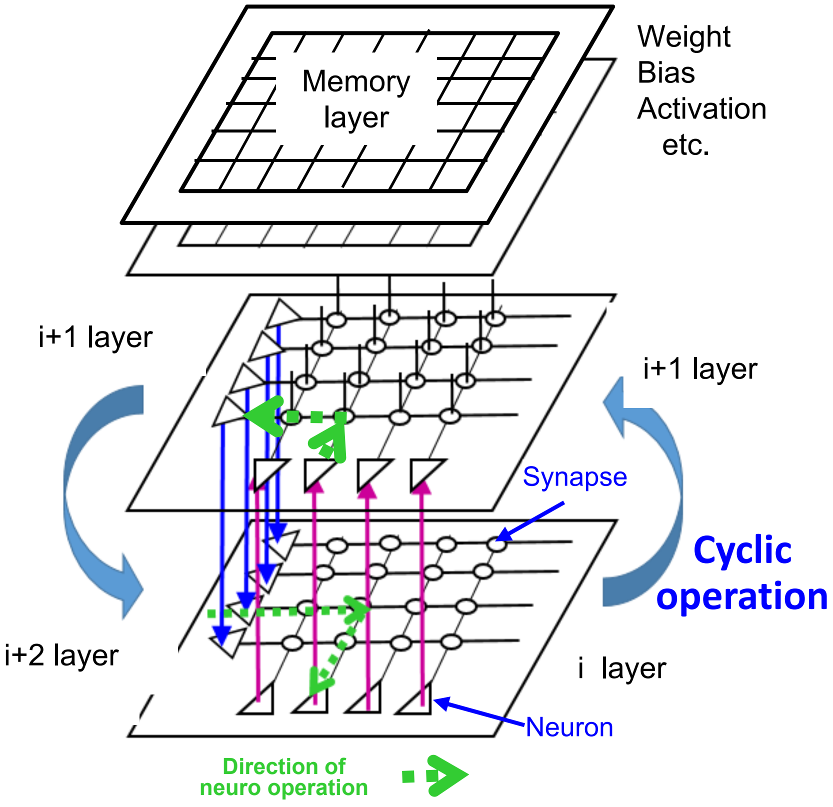
AI Control System (FPGA)

Learning Accelerator

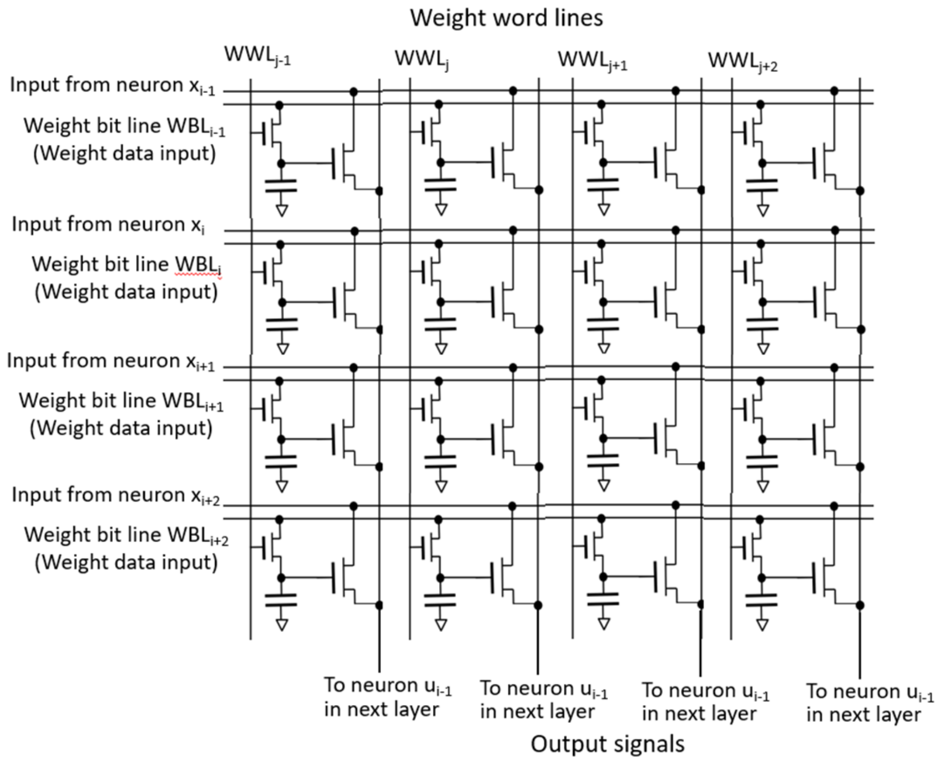
National Project in Japan (NEDO) of AI Chip (2019-2020)

Cyclic Neuro Operation in 3D Stacked AI Chip

(Forward propagation/ Backward propagation)

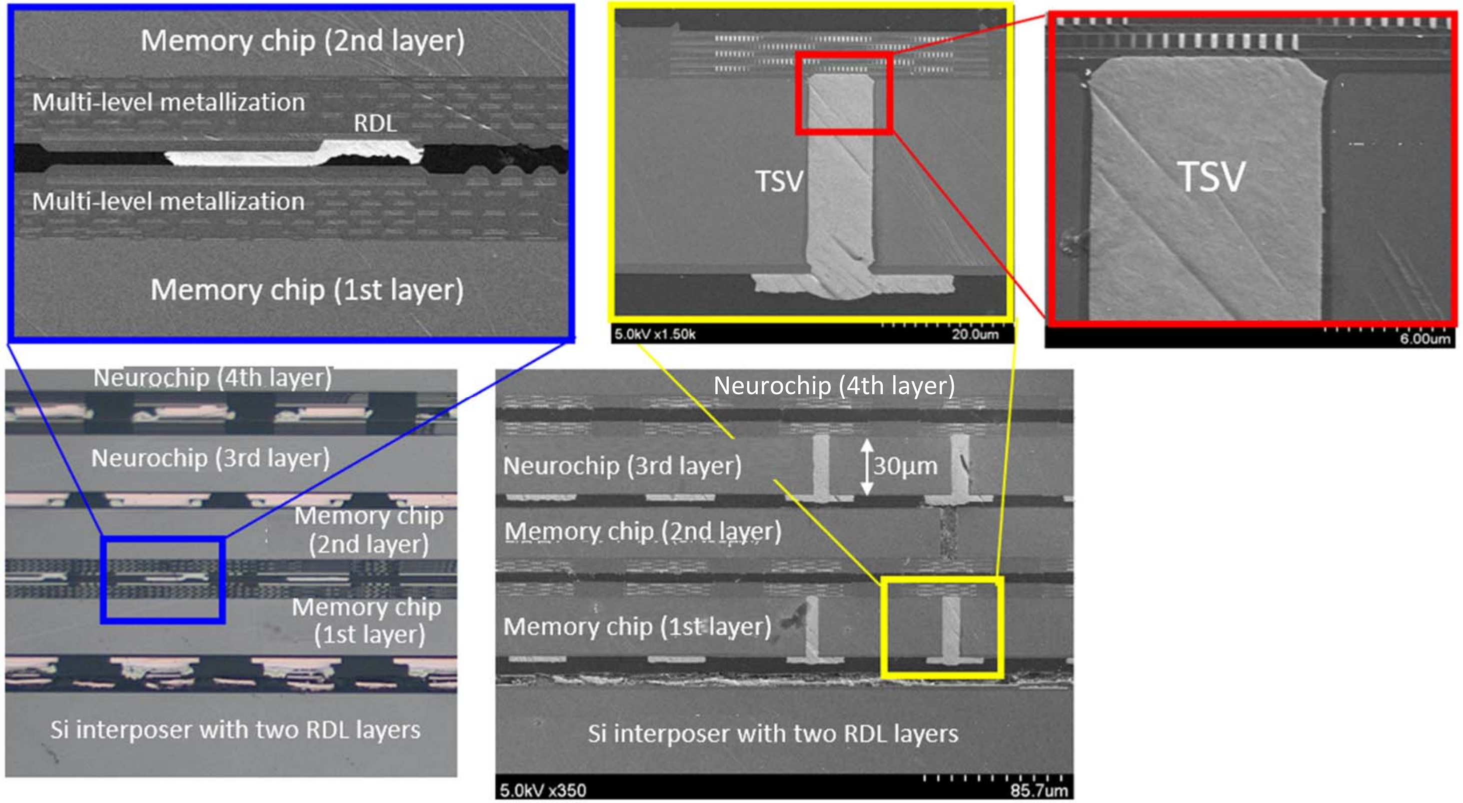


3D Stacked AI Chip


Weight word lines
WWL_{j-1}
WWL_{j}
WWL_{j+1}
WWL_{j+2}
Input from neuron x_{i-1}
Weight bit line WBL_{i-1}
(Weight data input)
Input from neuron x_{i}
Weight bit line WBL_{i}
(Weight data input)
Input from neuron x_{i+1}
Weight bit line WBL_{i+1}
(Weight data input)
Input from neuron x_{i+2}
Weight bit line WBL_{i+2}
(Weight data input)
To neuron u_{i-1} in next layer
To neuron u_{i-1} in next layer
To neuron u_{i-1} in next layer
To neuron u_{i-1} in next layer
Output signals

Synapse cell array

Cross-sectional View of 3D Stacked AL chip with Four Stacked Layers



Neuron-Block Partition to Densify Sparse Synaptic Connections in CNN

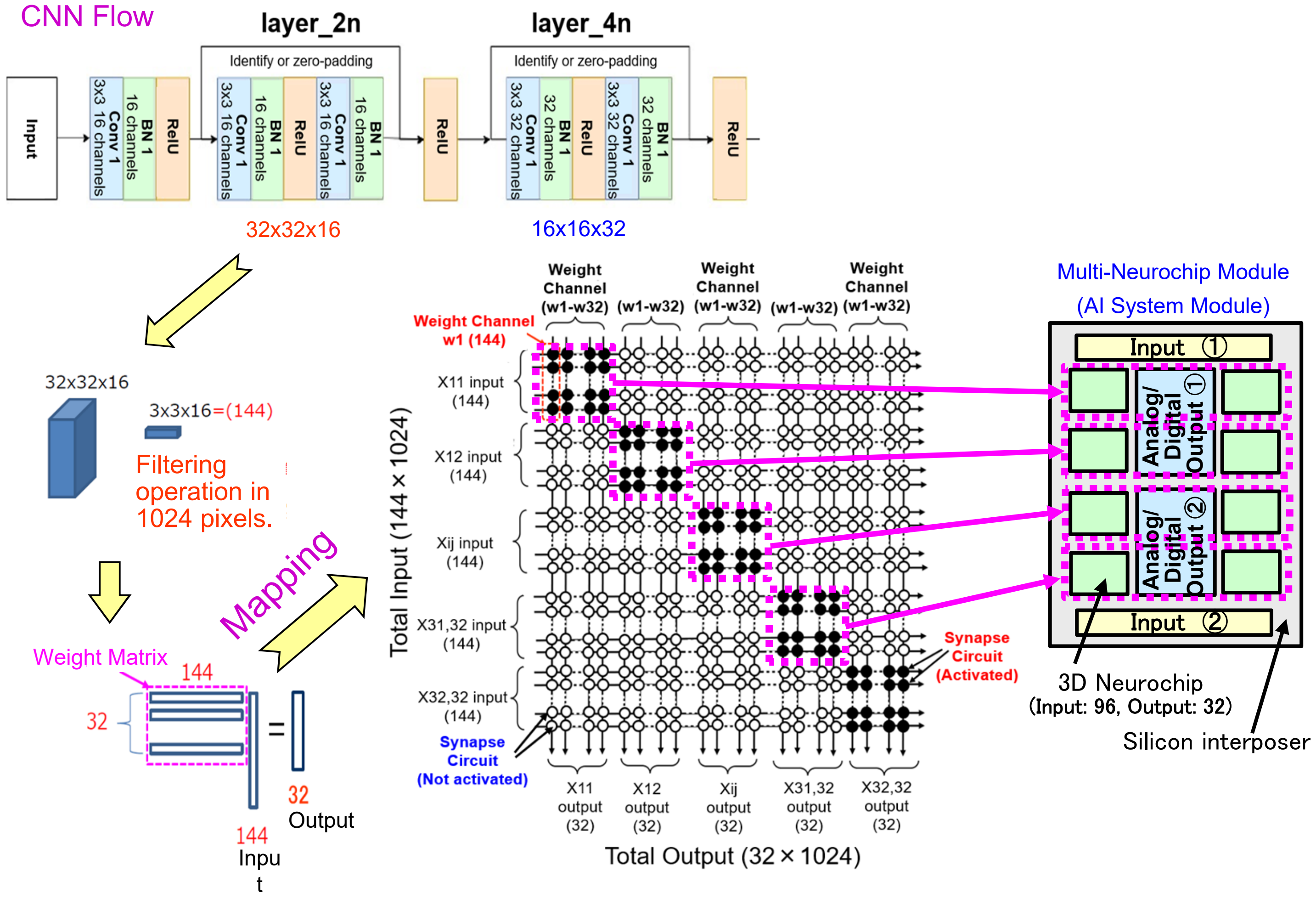
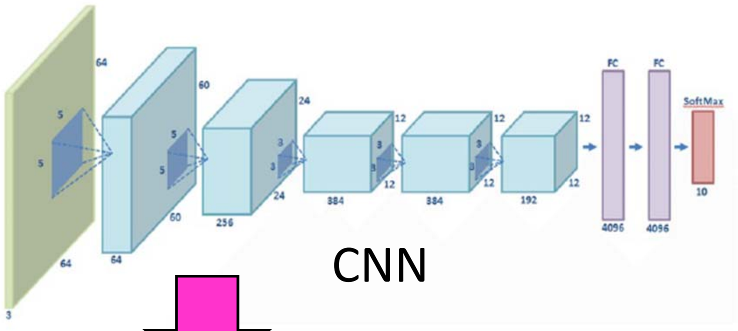
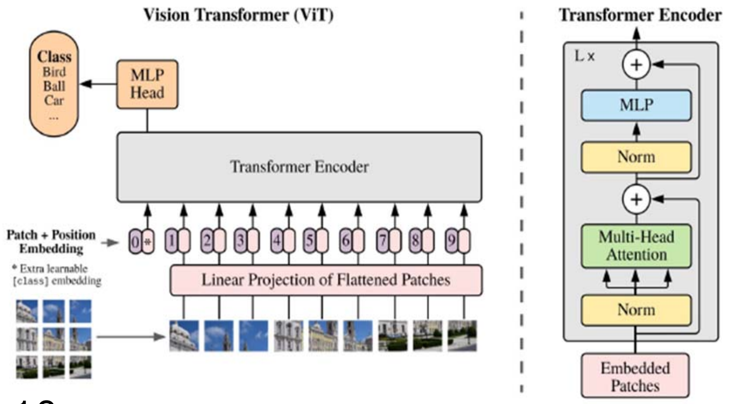


Image Recognition Using 3D Stacked AI chip

Paradigm Shift from CNN to ViT (Vision Transformer)



CNN



CIFAR-10

Vision Transformer (ViT)

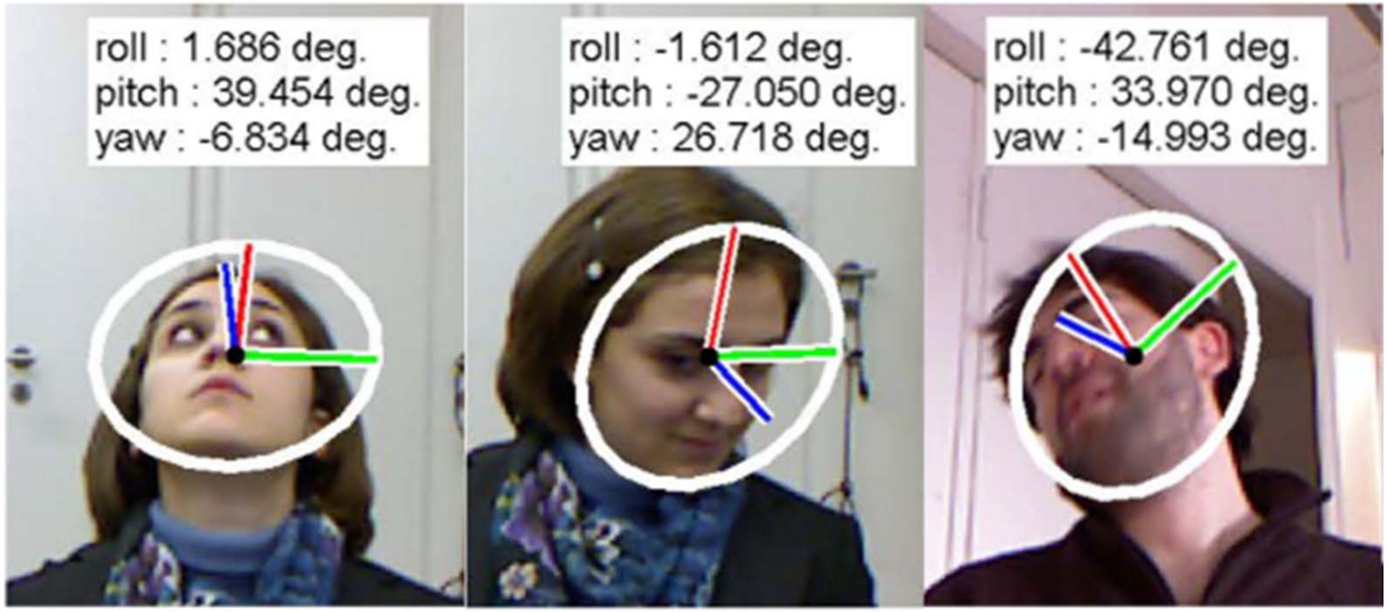
- ➢We can significantly reduce the number of matrix product operation (MPO) in ViT.
- ➢ViT is suitable for Edge application.

Network	Number of Matrix Product	Accuracy
CNN（Optimized）	6,212	71.4%
Tiny ViT-V1	1,169	71.3%
Tiny ViT-V2	150	76.6%

By Courtesy of Prof. T. Okatani, Tohoku University

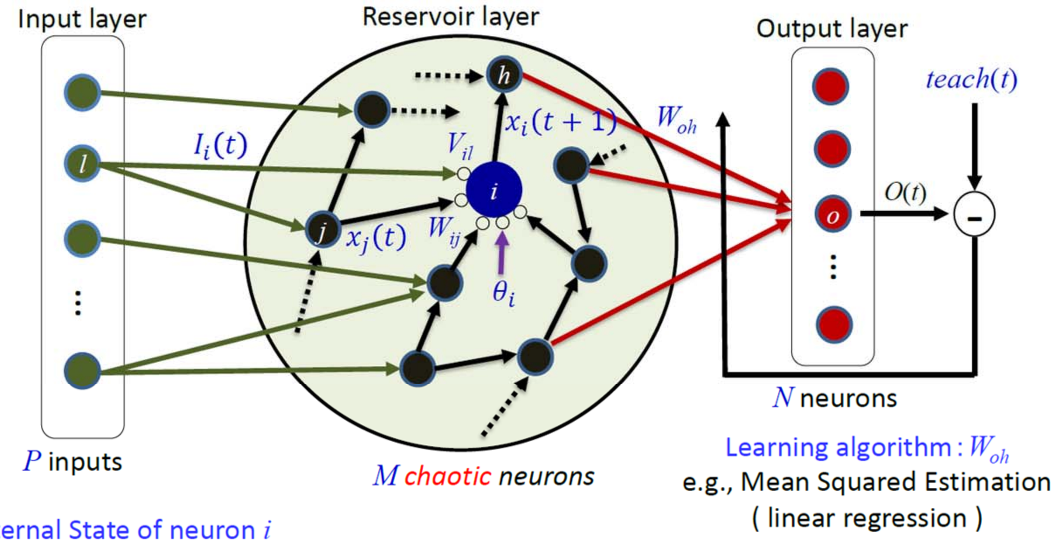
Face Recognition Using 3D Stacked AI chip

Average Error＝Yaw angle: 8.0 deg., Pitch angle: 8.7 deg., Roll angle: 7.6 deg.



By Courtesy of Prof. T. Okatani, Tohoku University

Implementation of Reservoir Neural Network in 3D Stacked AI Chip with Cyclic Neuro Operation



Internal State of neuron i

$$y_i(t+1) = k y_i(t) + \sum_{j=1}^{M} W_{ij} f(y_j(t)) + \sum_{l=1}^{P} V_{il} I_i(t) - \alpha x_i(t) - \theta_i(1-k)$$

Output of neuron i

$$x_i(t+1) = f\big(y_i(t+1)\big)$$

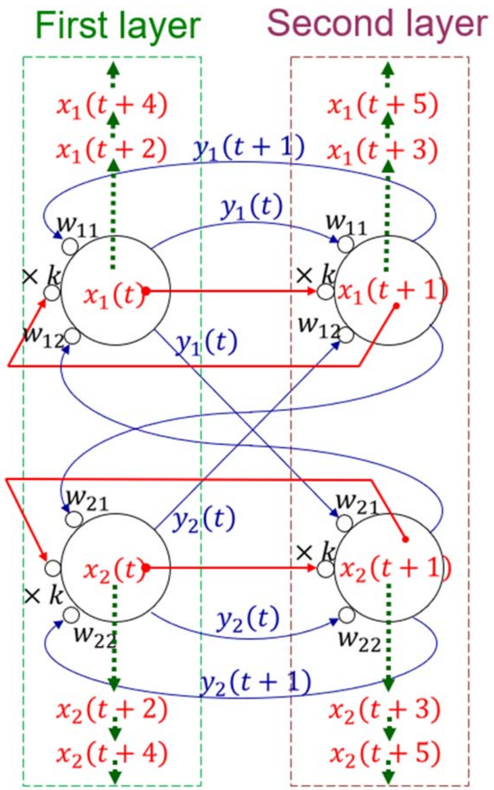
k: dumping factor of the refractriness

α: scaling factor θ_i: threshold

$f(\cdot) = 1/(1+\exp(-x/\epsilon))$

K. Aihara et al., Phys. Lett. A, vol. 144, 1990.

Configuration of Reservoir Neural Network with Simple Learning



Mapping of Reservoir Neural Network to 3D Stacked AI Chip with Cyclic Neuro Operation

K. Fukuda, Yoshihiko Horio et al. , NOLTA, IEICE (2021)

Voice Recognition Using 3D Stacked AI chip

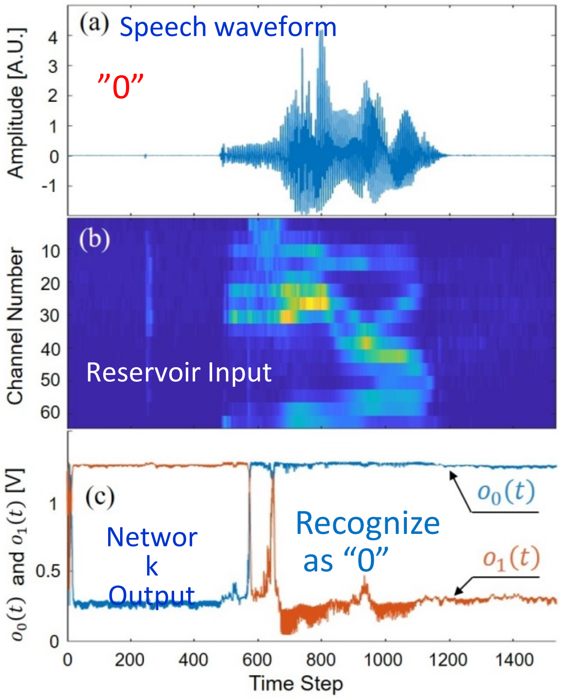
Number of Reservoir Neuron	Connectivity within Reservoir
64	22 %

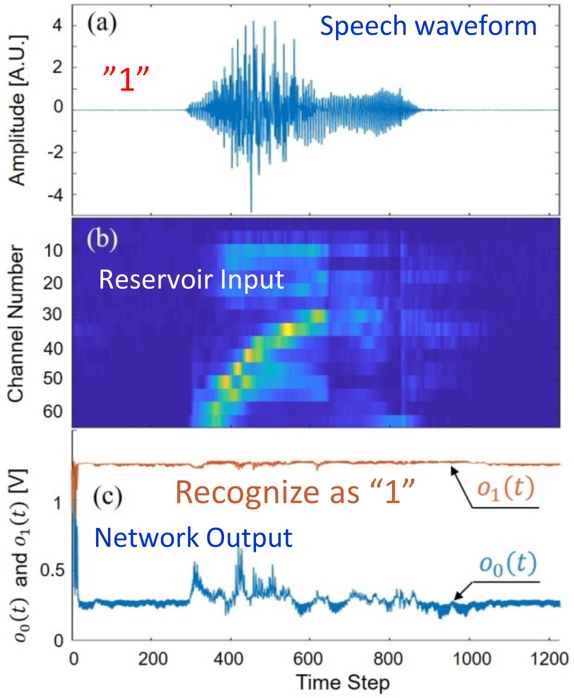
Input Connectivity	Output Connectivity
6.25 %	100 %

Learning (ten times) of “zero” and “one” by Linear Regression

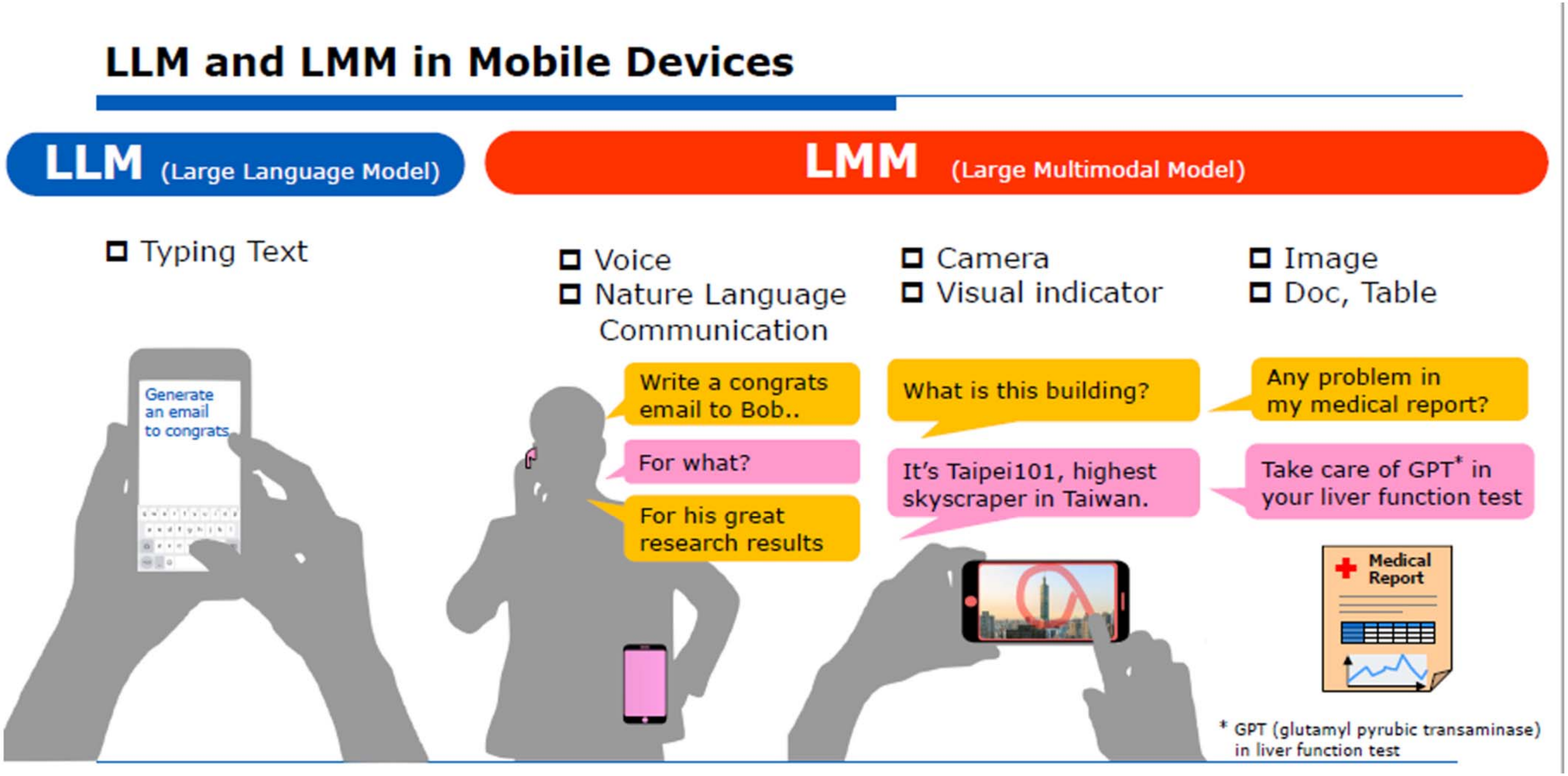
Example of network response after learning

100 % recognition for ten different voices





Next Generation AI



LLM and LMM in Mobile Devices

LLM (Large Language Model)

LMM (Large Multimodal Model)

□ Typing Text

□ Voice
□ Nature Language Communication

□ Camera
□ Visual indicator

□ Image
□ Doc, Table

Generate an email to congrats

Write a congrats email to Bob..

For what?

For his great research results

What is this building?

It's Taipei101, highest skyscraper in Taiwan.

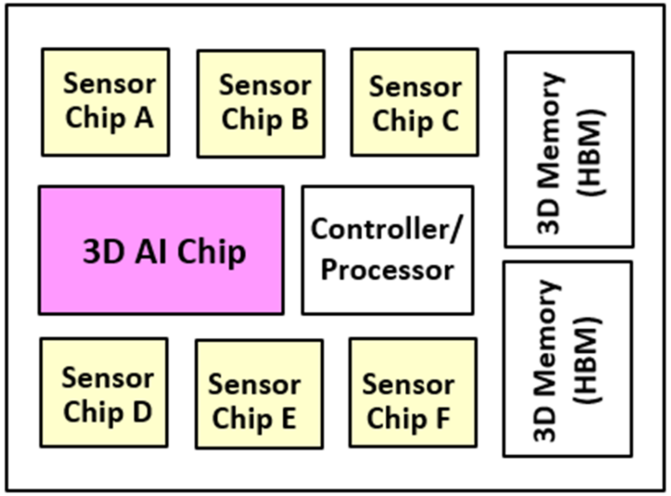
Any problem in my medical report?

Take care of GPT* in your liver function test

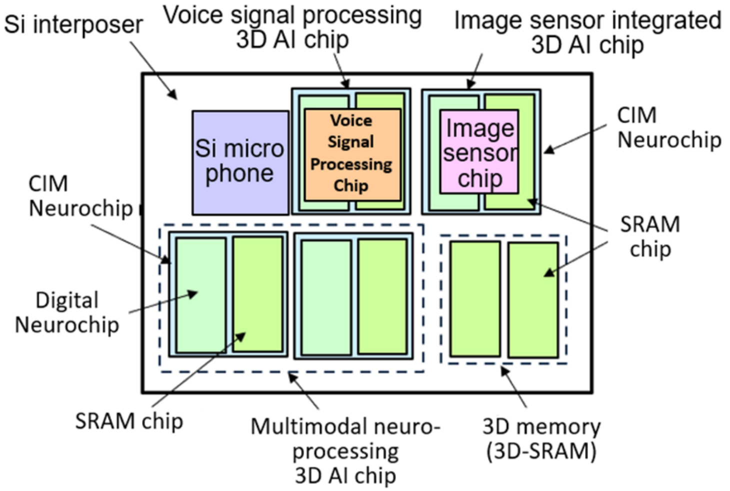
* GPT (glutamyl pyrubic transaminase) in liver function test

Bor-Sung Liang, IEEE ISSCC Forum 2.8 (2024)

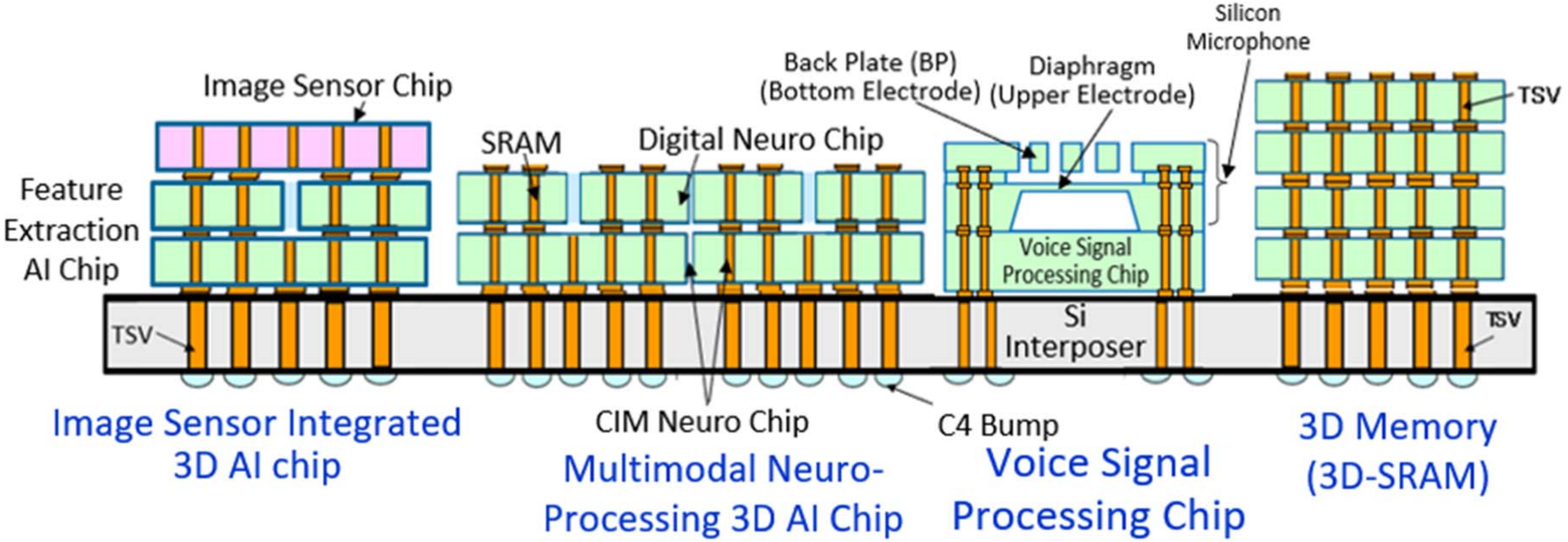
LMM AI Chip Project in Tohoku University (AI−Sensor Fusion)



Sensor Integrated AI System Module
(Final Target)

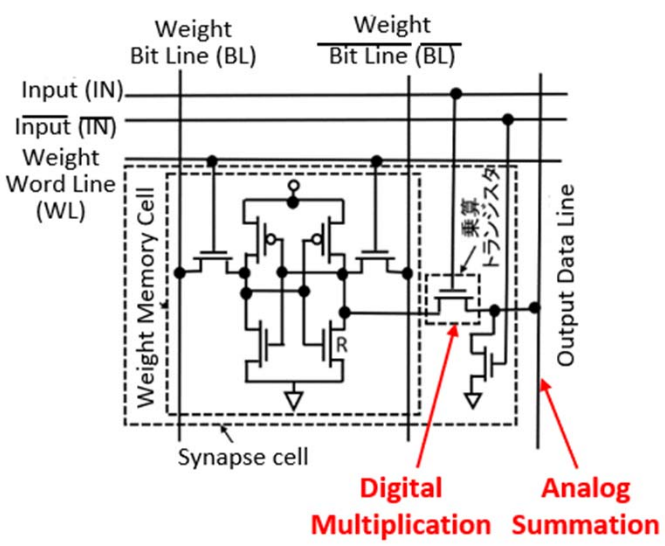

Si interposer
Voice signal processing 3D AI chip
Image sensor integrated 3D AI chip
Si micro phone
Voice Signal Processing Chip
Image sensor chip
CIM Neurochip
CIM Neurochip
SRAM chip
Digital Neurochip
SRAM chip
Multimodal neuro-processing 3D AI chip
3D memory (3D-SRAM)

Sensor Integrated AI System Module
by Heterogeneous 3D Chiplet Integration

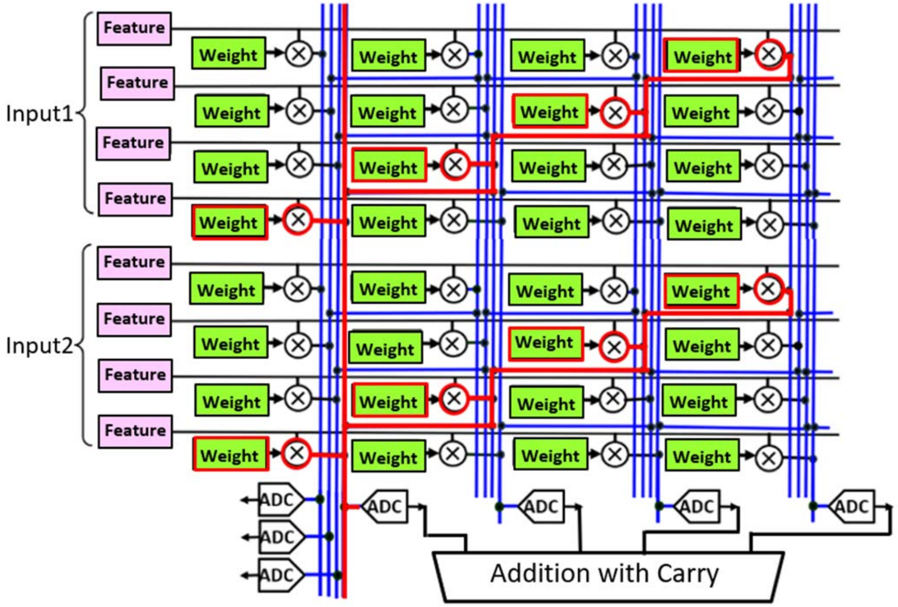

Silicon Microphone
Back Plate (BP) (Bottom Electrode)
Diaphragm (Upper Electrode)
Image Sensor Chip
SRAM
Digital Neuro Chip
TSV
Feature Extraction AI Chip
Voice Signal Processing Chip
Si Interposer
TSV
TSV
CIM Neuro Chip
C4 Bump
Image Sensor Integrated 3D AI chip
Multimodal Neuro-Processing 3D AI Chip
Voice Signal Processing Chip
3D Memory (3D-SRAM)

Cross-sectional Structure of Sensor Integrated AI System Module

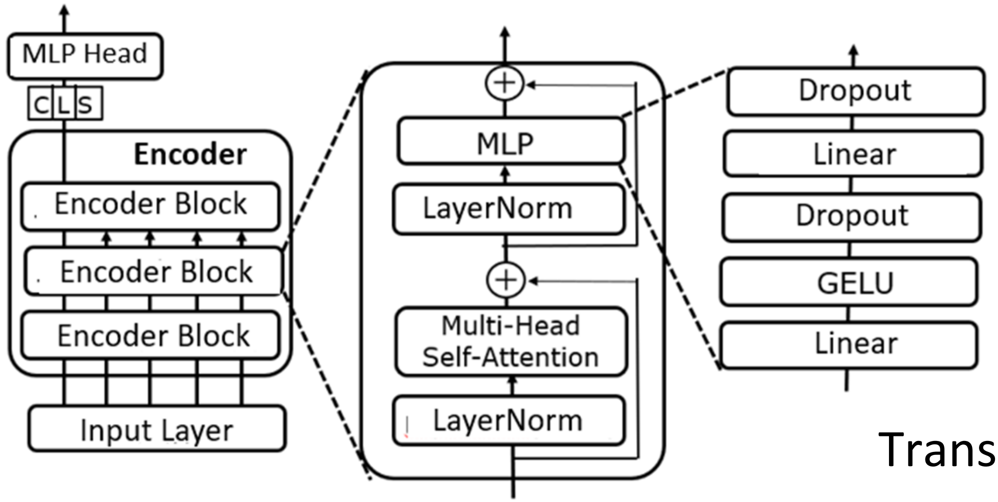
AI Chip Based on Memory-in-Computing (CIM) with SRAM and Transformer Algorithm



CIM (Memory-in-Computing)
by SRAM Memory Cell



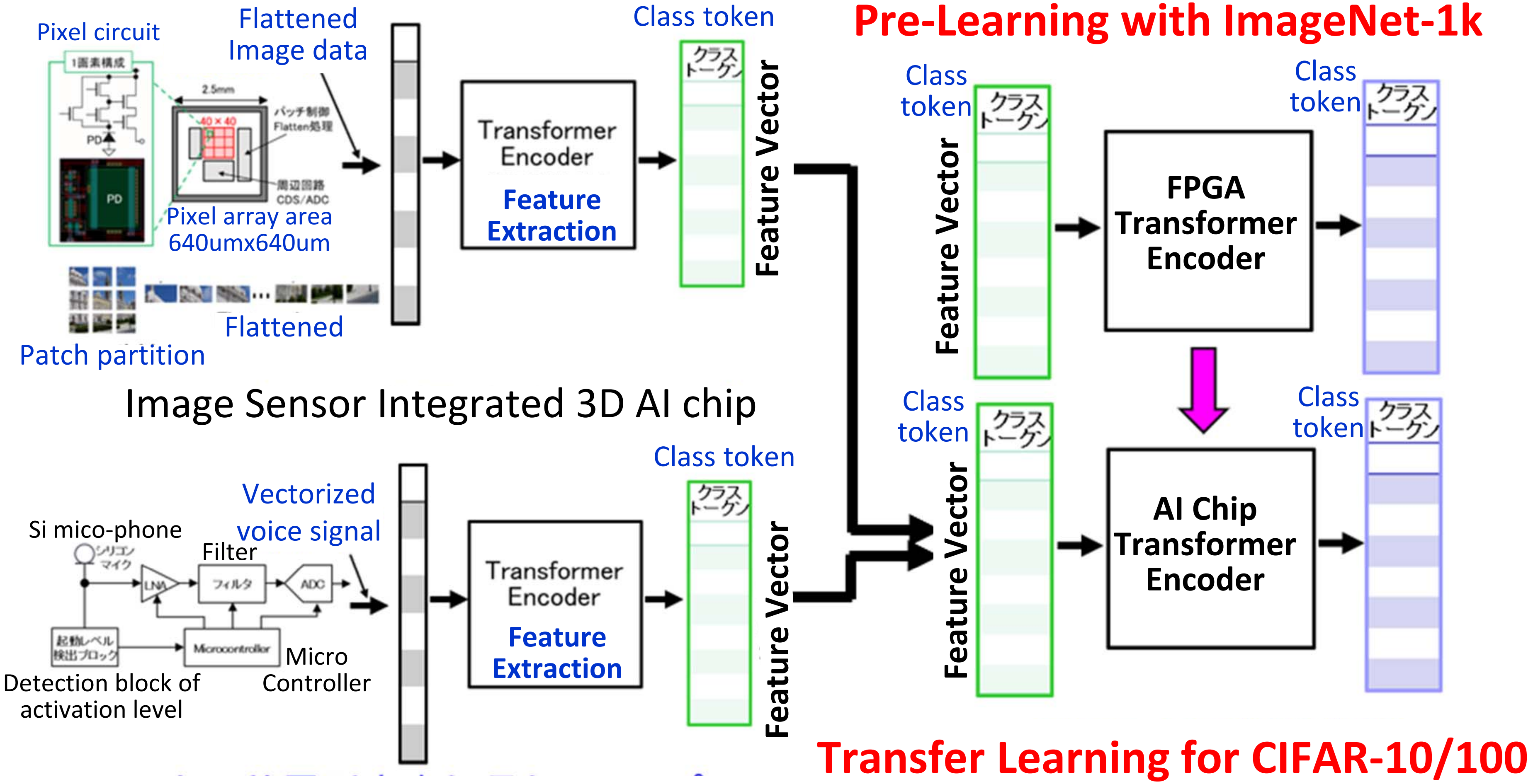
CIM Basic Synapse Circuit Array



Transformer Algorithm

LMM AI Chip Project in Tohoku University (AI−Sensor Fusion)

Energy Efficient Neuro Processing and Reduced Data Transfer by Integrating Sensor Devices with Feature Extraction Function



Pixel circuit

Flattened Image data

Pixel array area 640umx640um

Patch partition

Flattened

Transformer Encoder

Feature Extraction

Class token

Feature Vector

Image Sensor Integrated 3D AI chip

Si mico-phone

Filter

Vectorized voice signal

Detection block of activation level

Micro Controller

Transformer Encoder

Feature Extraction

Class token

Feature Vector

Pre-Learning with ImageNet-1k

Class token

Feature Vector

FPGA Transformer Encoder

Class token

Class token

Feature Vector

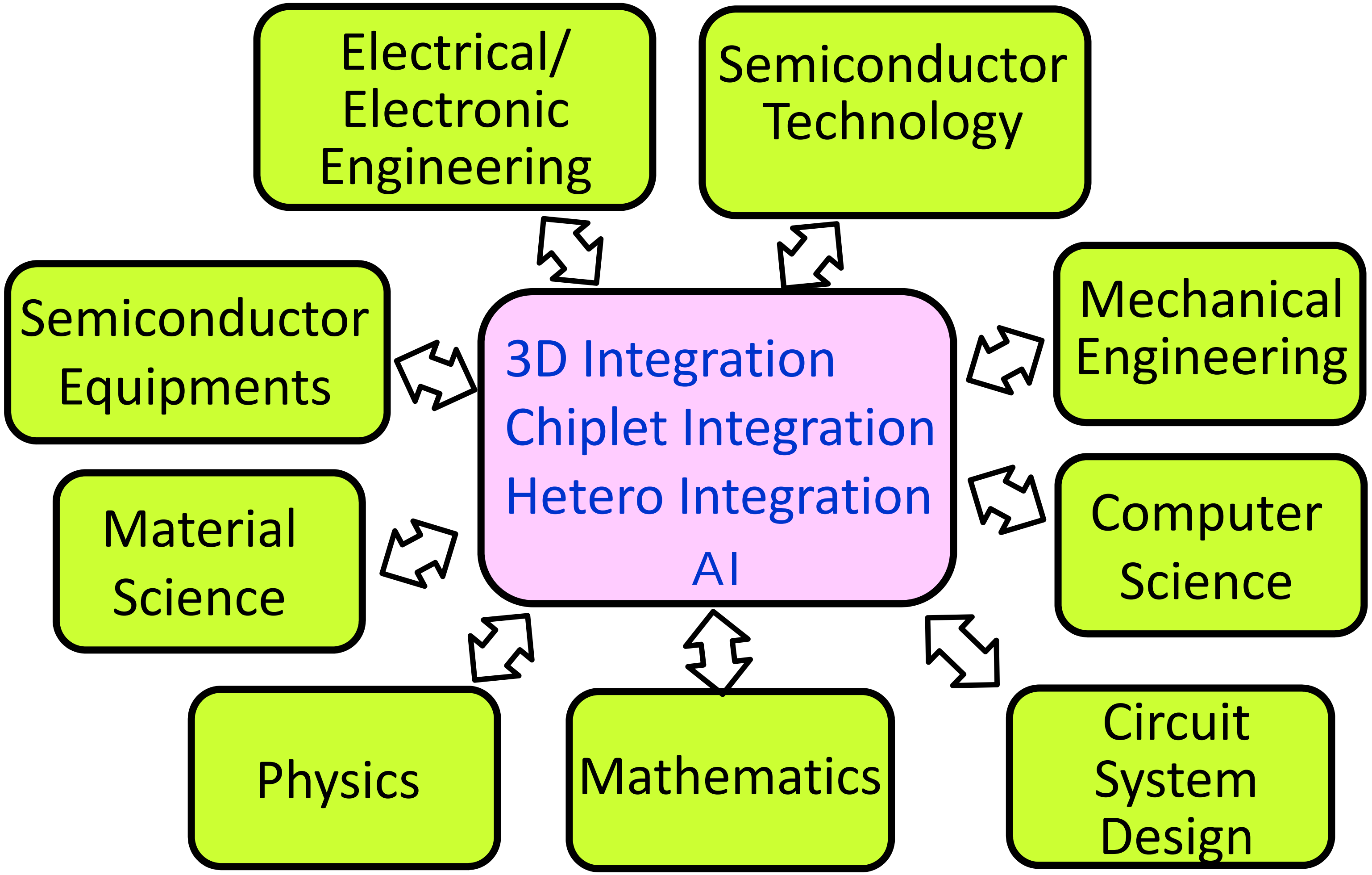
AI Chip Transformer Encoder

Class token

Transfer Learning for CIFAR-10/100

Voice Signal Processing Chip

Heterogeneous Chiplet Integration/AI Need Various Kind of Knowledges and Skills



3D Integration
Chiplet Integration
Hetero Integration
AI

Electrical/ Electronic Engineering

Semiconductor Technology

Semiconductor Equipments

Mechanical Engineering

Material Science

Computer Science

Physics

Mathematics

Circuit System Design

12-inch 3D Production Line in Tohoku Univ.

GINTI（Global INTegration Initiative）

Bonder / Debonder

Wafer edge trimming

Wafer thinning

patterning

Deep Si etch

SiO2 etch

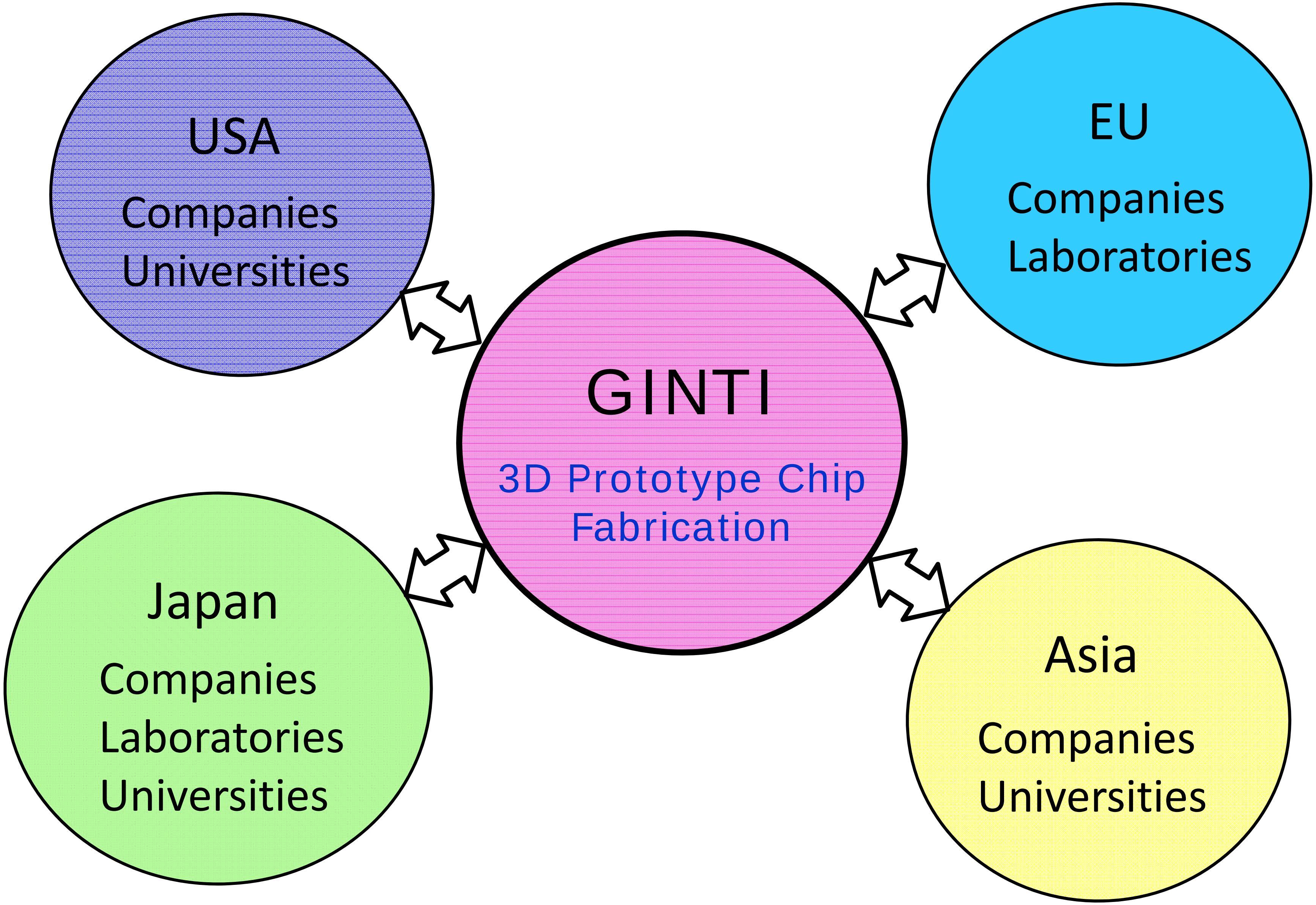
Asher

Insulator CVD

Ionized sputter

plating

International Cooperation IN GINTI



USA

Companies
Universities

EU

Companies
Laboratories

GINTI

3D Prototype Chip Fabrication

Japan

Companies
Laboratories
Universities

Asia

Companies
Universities

Conclusions

- 3D heterogeneous integration and chiplet integration are the key for future intelligent systems such as HPC, AI/ML, post-5G/6G systems and quantum computer systems.
- 3D heterogeneous integration and chiplet integration need various kinds of knowledges and skills. Therefore International collaborations are indispensable.