Wafer-Scale 3D Photonic Chiplet Integration for AI System

Mitsu Koyanagi

Global INTegration Initiative (GINTI)
Tohoku University, Japan

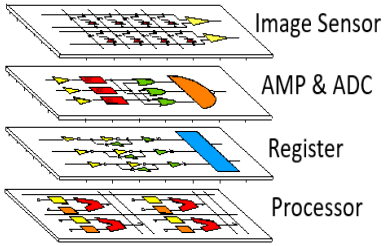
Tohoku-MicroTech. Co., Ltd, Japan

Outline

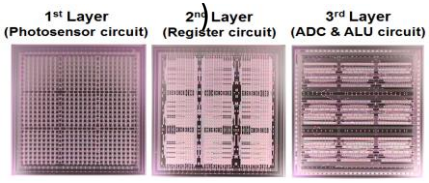
- Introduction
- Reconfigured Wafer-to-Wafer 3D Chiplet Integration Technology Based on Self-Assembly
- Silicon Photonics Chip for Visible Light Communication
- 3D AI Chip with CIM Analog/Digital Neuro Operation
- Wafer-Scale 3D Photonic Chiplet Integration for AI System
- Conclusions

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

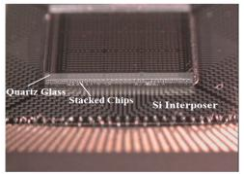
3D image sensor chip



(a)



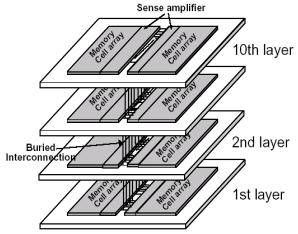
(b)



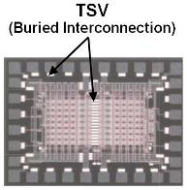
(c)

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

3D memory

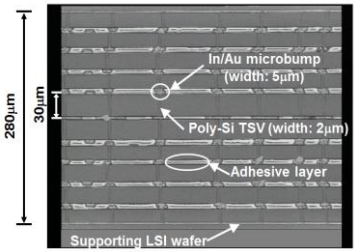


(a)

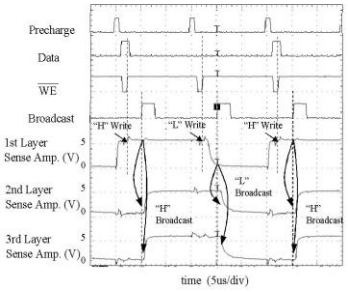


DRAM chip photo
(chip size: 2mmx1.6mm)

(b)



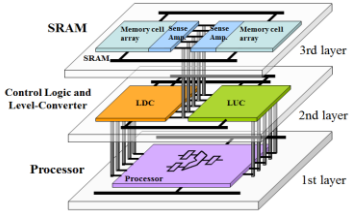
(c)



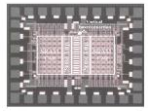
(d)

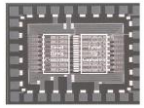
K. W. Lee, M. Koyanagi *et al.*
IEEE IEDM (2000)

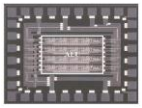
3D microprocessor chip

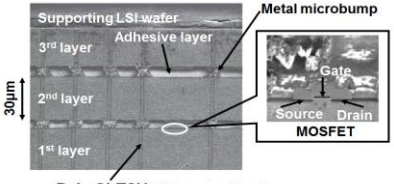


(a)


SRAM

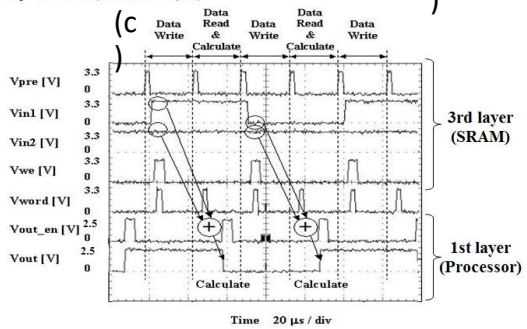

Level converter Control logic


Processor



(b)

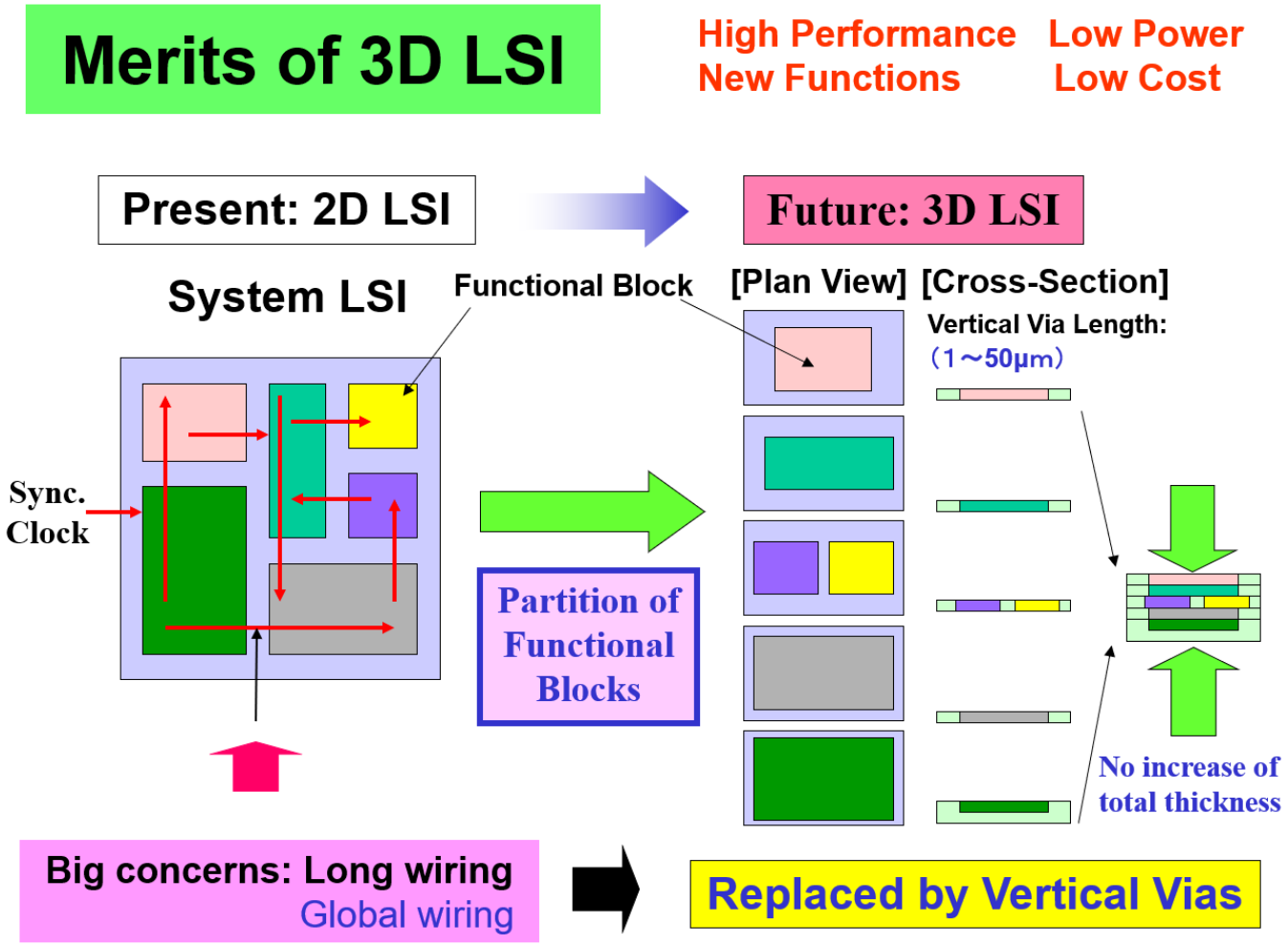
(c)



(d)

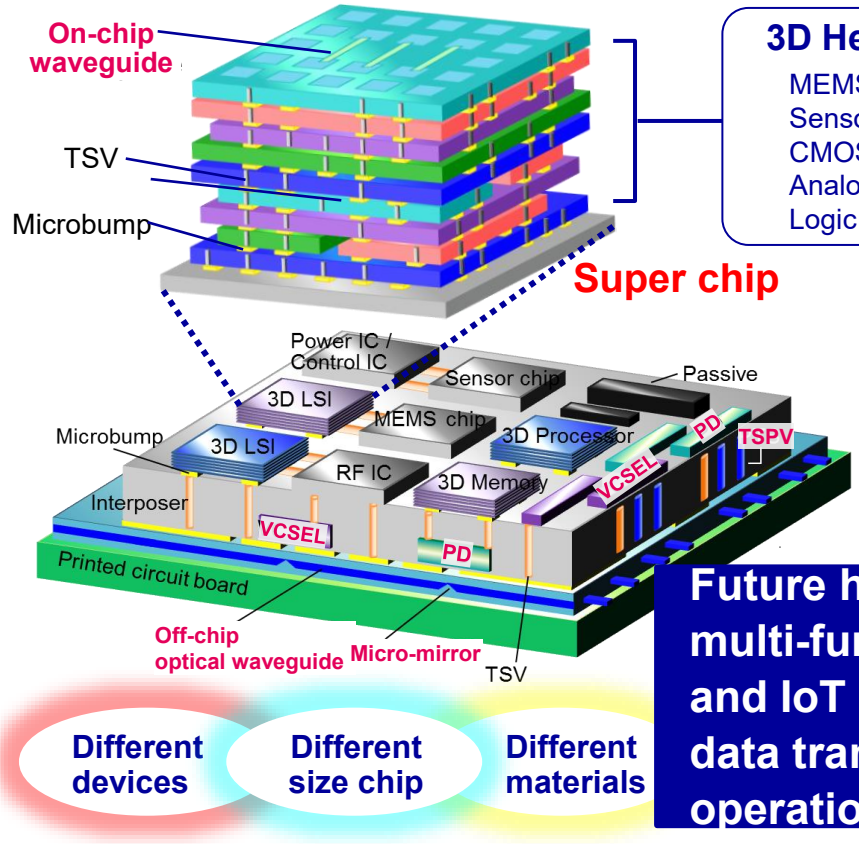
T. Ono, M. Koyanagi *et al.*, IEEE COOL Chips (2002)

Proposal of 3D Chiplet Integration by Tohoku University



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

3D Heterogeneous Integration Technology in Tohoku Univ.



3D Heterogeneous Integration

MEMS chip	Microprocessor
Sensor chip	Power IC/Control IC
CMOS RF-IC	Flash memory
Analog LSI	DRAM
Logic IC	SRAM

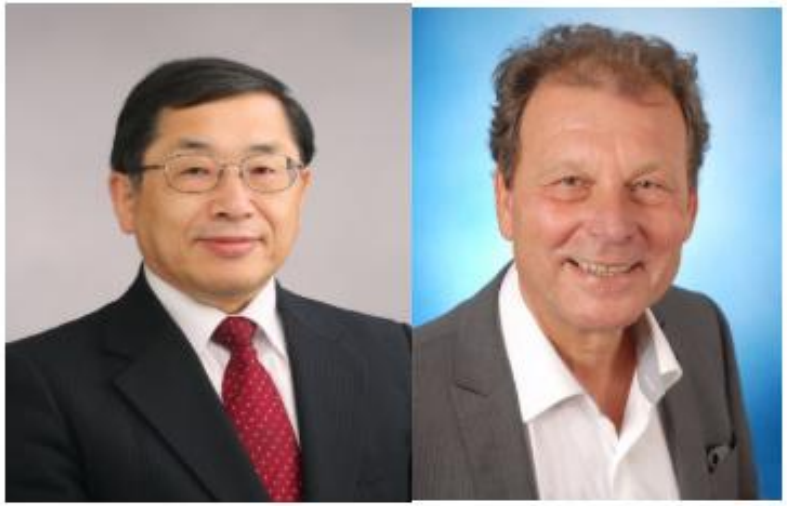
2.5D Heterogeneous Integration

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)

IEEE Rao Tummala (Electronics Packaging) Award



Recipients of the 2020

IEEE Electronics Packaging Award

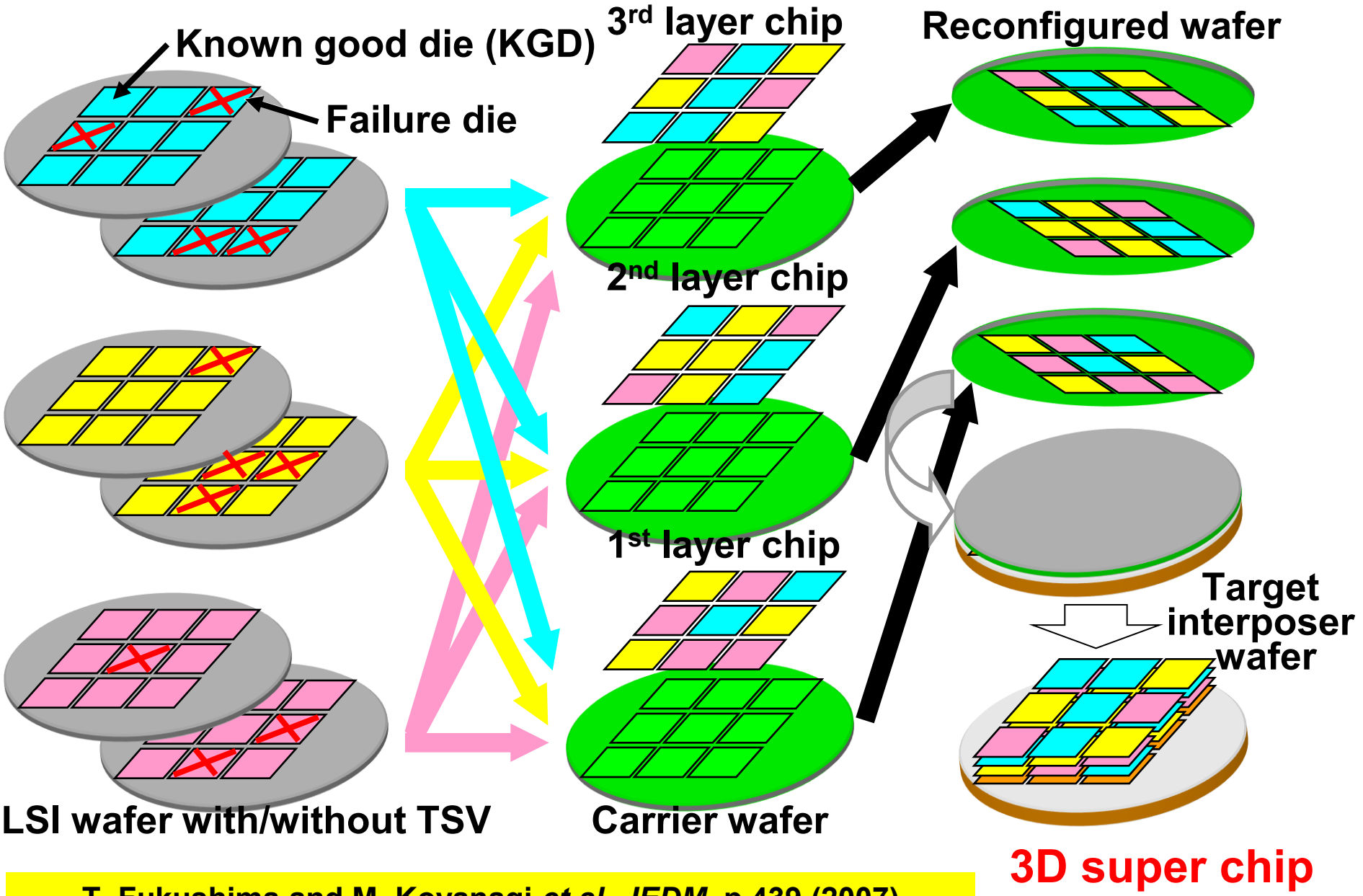
Mitsumasa Koyanagi and Peter Ramm

"For pioneering contributions leading to the commercialization of 3D wafer and die level stacking packaging"

Requirements for Wafer-Scale 3D Chiplet Integration

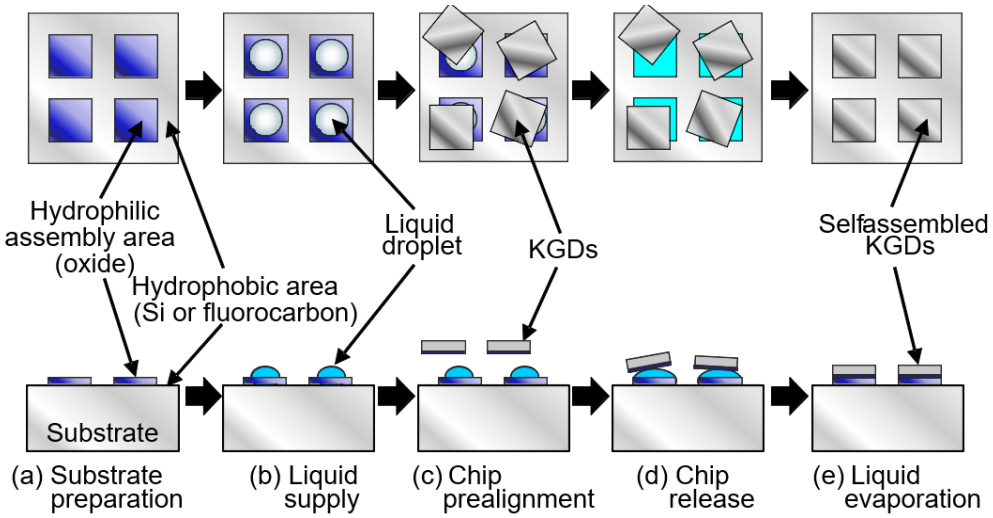
- High wafer yield
  - ➡ Reconfigured Wafer with KGDs
- Heterogeneous integration of different kinds of chiplets
  - ➡ Reconfigured Wafer with different kinds of chiplets
- Long distance interconnection with high data bandwidth and low power consumption
  - ➡ Silicon photonics and optical interconnection

New Reconfigured Wafer-to-Wafer 3D Integration



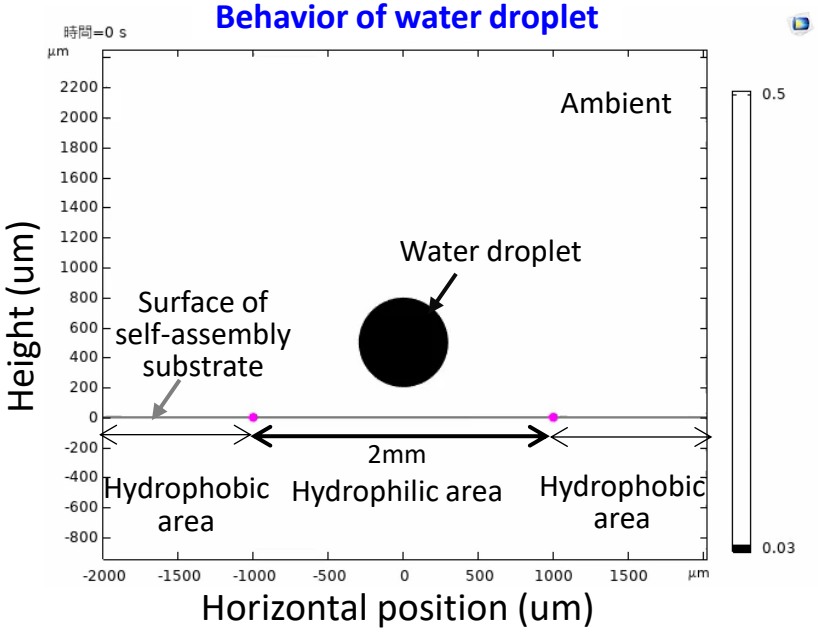
T. Fukushima and M. Koyanagi *et al.*, *IEDM*, p.439 (2007)

Self-Assembly Process Flow

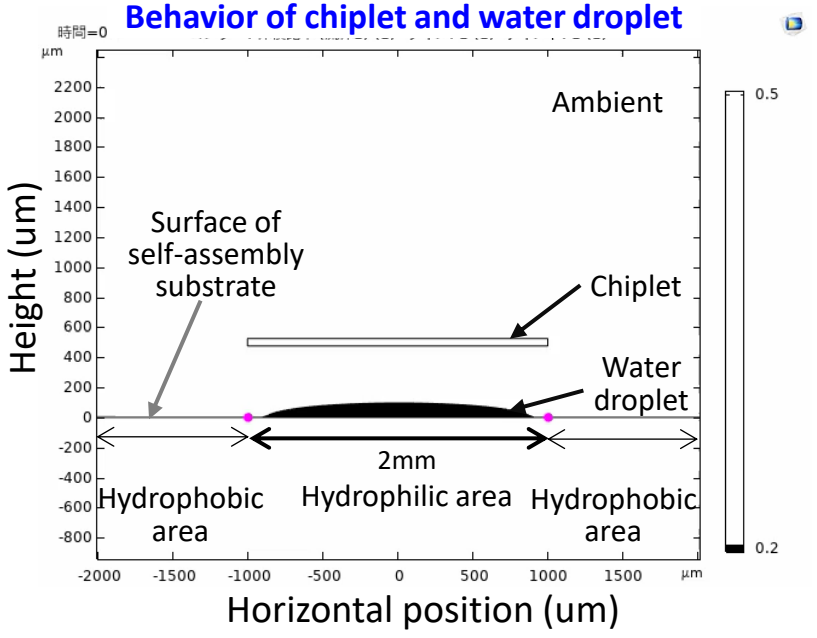


Two-phase Flow Simulation for Self-Assembly

Behavior of water droplet

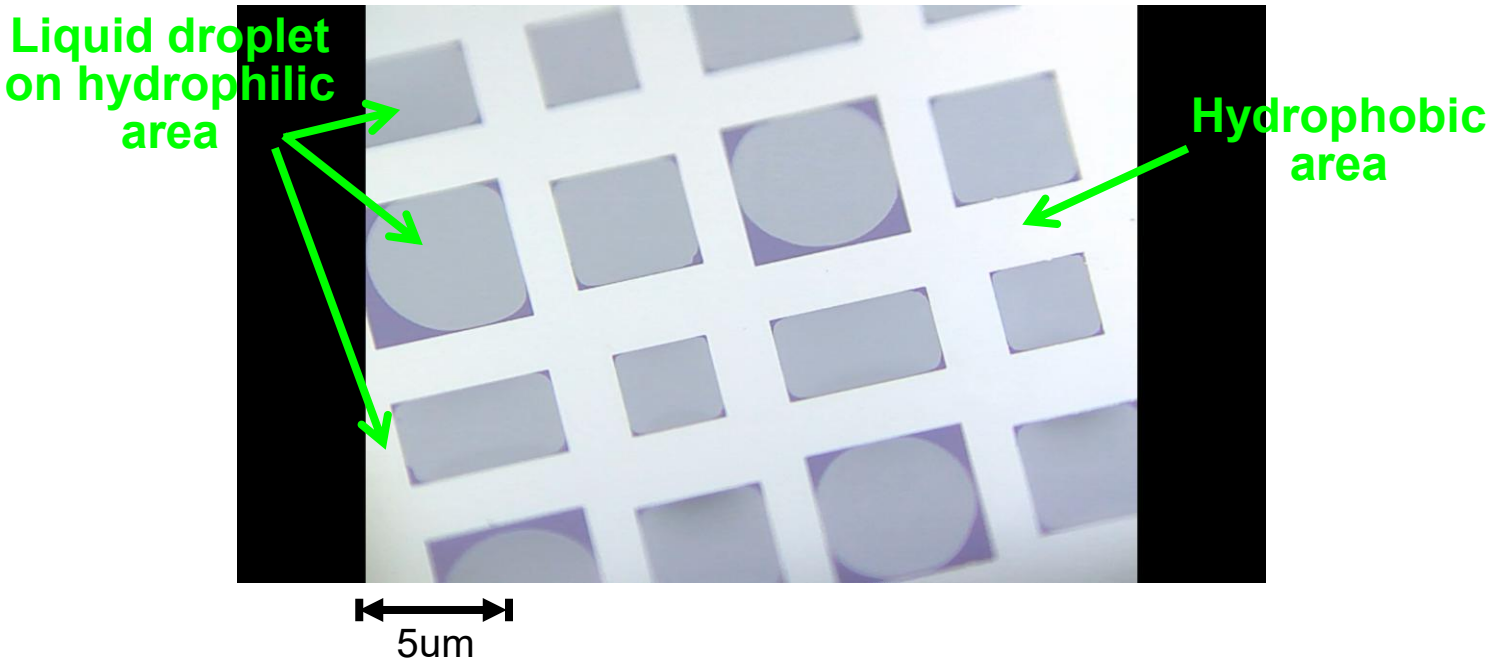


Behavior of chiplet and water droplet



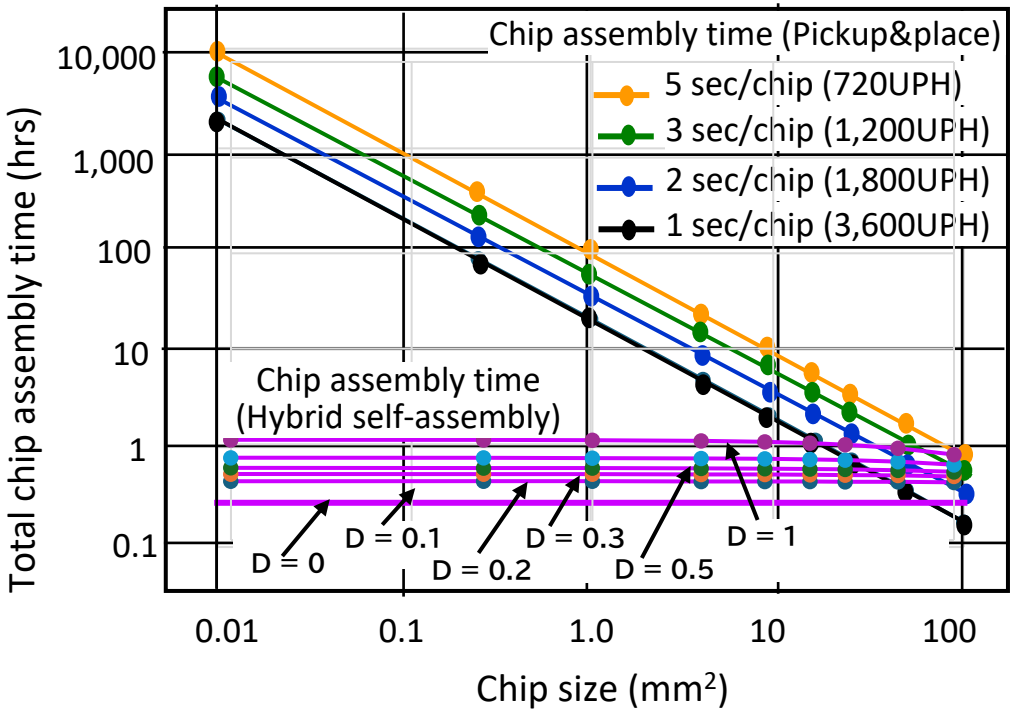
Simultaneous Bonding of Many Dies with Different Size by Self-Assembly

* This movie is real-time playing.



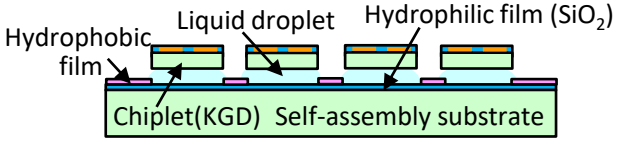
Throughput Comparison between Pick&Place and Self-Assembly for 12-inch Wafer

(Total Chip Assembly Time vs. Chip Size)

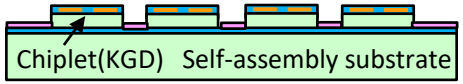


Reconfigured Wafer-to-Wafer 3D Chiplet Integration Using Self-Assembly

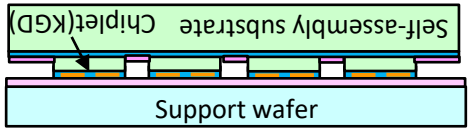
Fusion of 3D-IC Technology and Packaging Technology



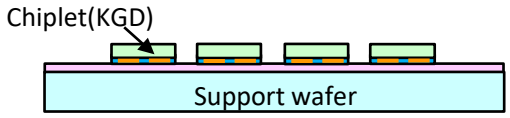
(a)Multi-chip self-assembly



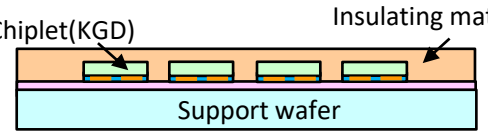
(b)Multi-chip temporary bonding



(c)Bonding to support wafer

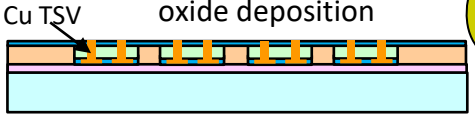


(d)De-bonding of self-assembly substrate
(Simultaneous transfer of many chiplets to support wafer)



(e)Wafer molding

(f)Thinning of chiplets/
oxide deposition

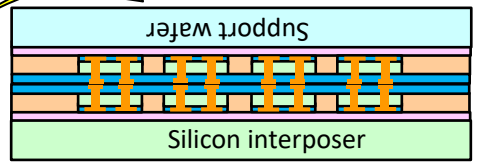


Reconfigured wafer

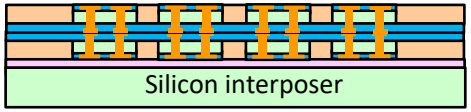
(g)Deep hole formation for TSV by plasma etching/Liner oxide deposition/ Barrier_seed layer sputtering/ Bottom-up Cu electroplating/ Removal of Cu overburden by CMP

Flip

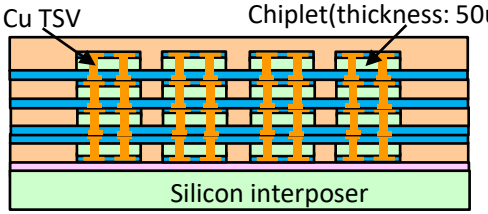
Key: Hybrid bonding of reconfigured wafers



(h)Reconfigured Wafer-to-wafer hybrid bonding



(i)De-bonding of upper support wafer

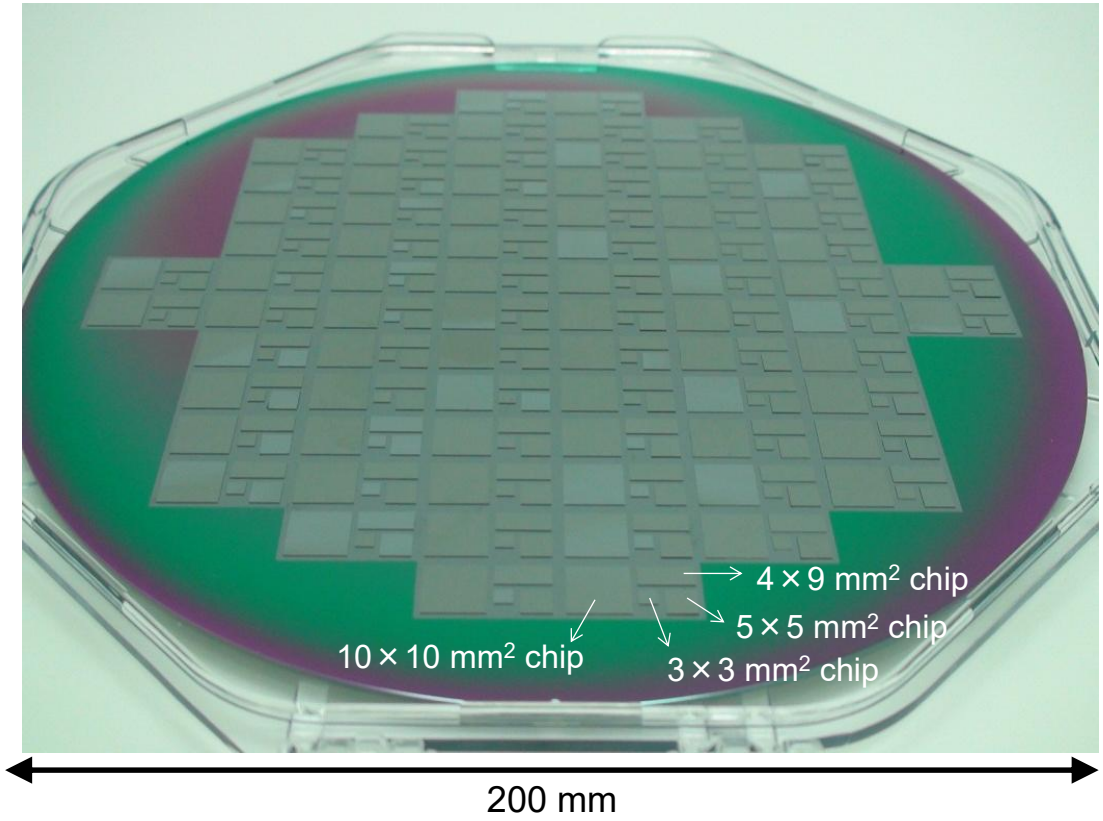


(j)Repeat of step(i) and step(j)
(Wafer level stacking of chiplets)

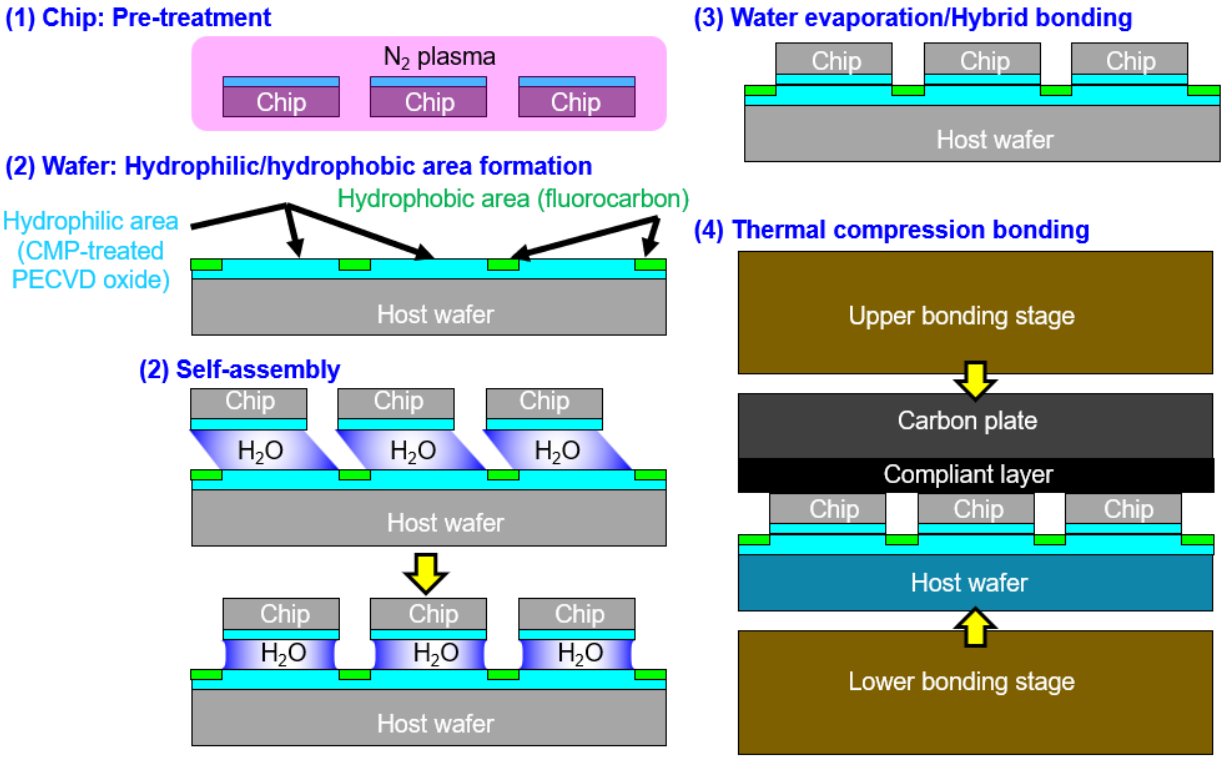
Key Technologies for Reconfigured Wafer-to-Wafer 3D Chiplet Integration

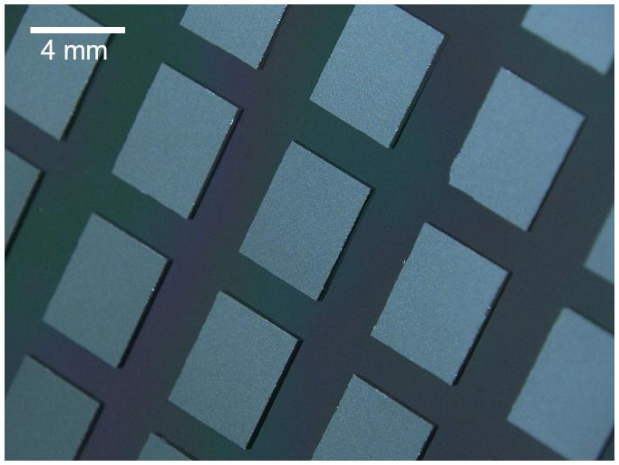
- Multi-chip self-assembly
- Multi-chip temporary bonding
- Multi-chip transfer to support wafer
- Simultaneous grinding and polishing of multi-chips
- Hybrid bonding of reconfigured wafers
- De-bonding of support wafer

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



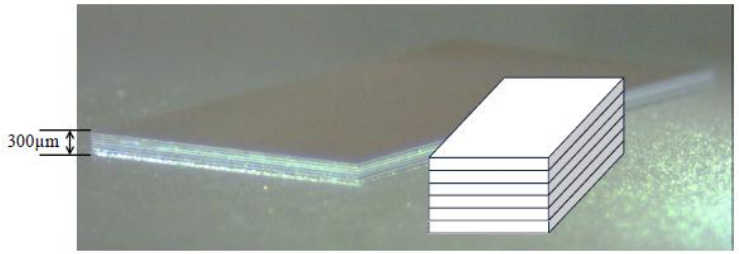
Process Flow of Multi-chip Self-Assembly and Hybrid Bonding





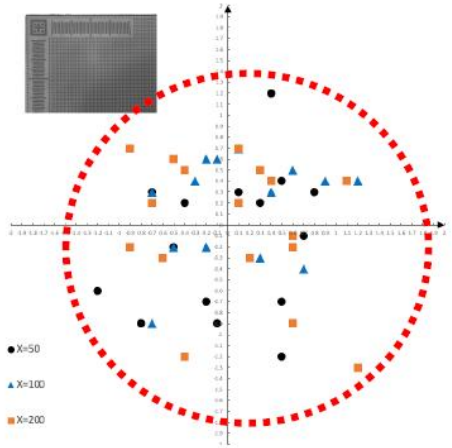
Combined Technology of Self-Assembly and Hybrid Bonding

①



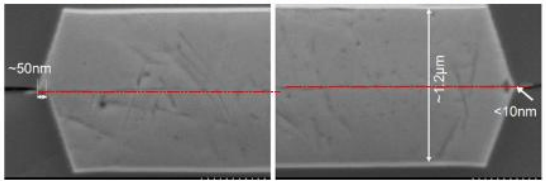
- Using this method, we have currently achieved an assembly of 6 layers. In the future, we will continue to explore the best conditions to achieve a structure of more than 12 layers.

②



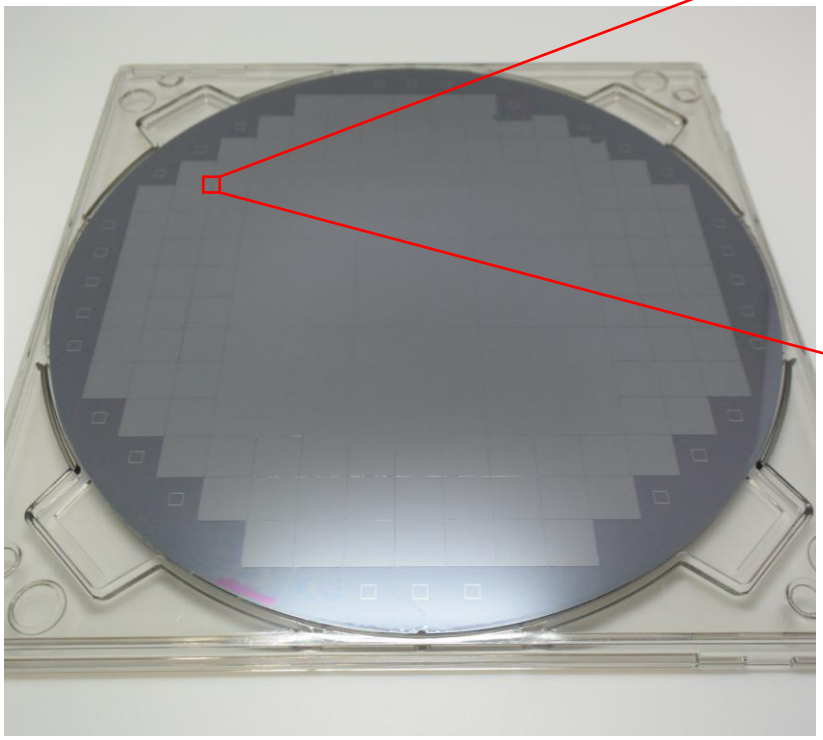
- The current average assembly accuracy has reached a level of less than 500nm but the data is still scattered, and efforts will be made to reduce the data scatter in the future.

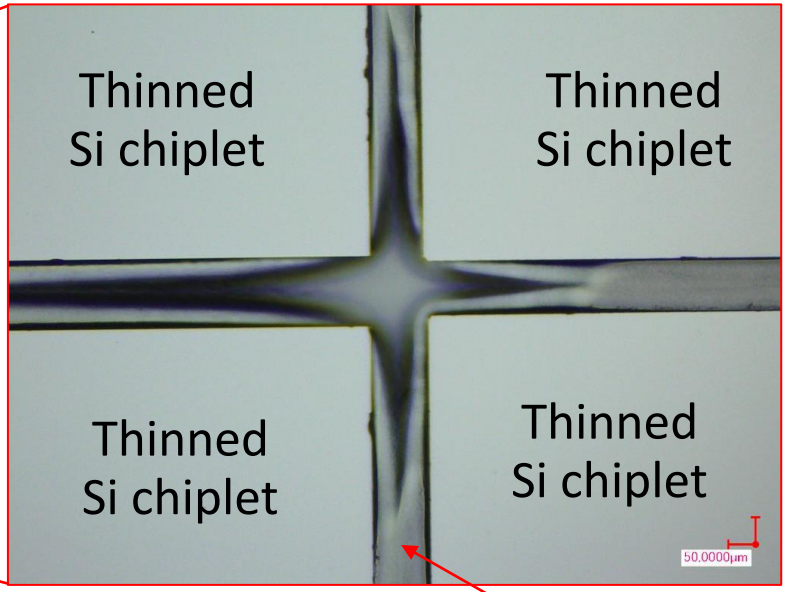
③

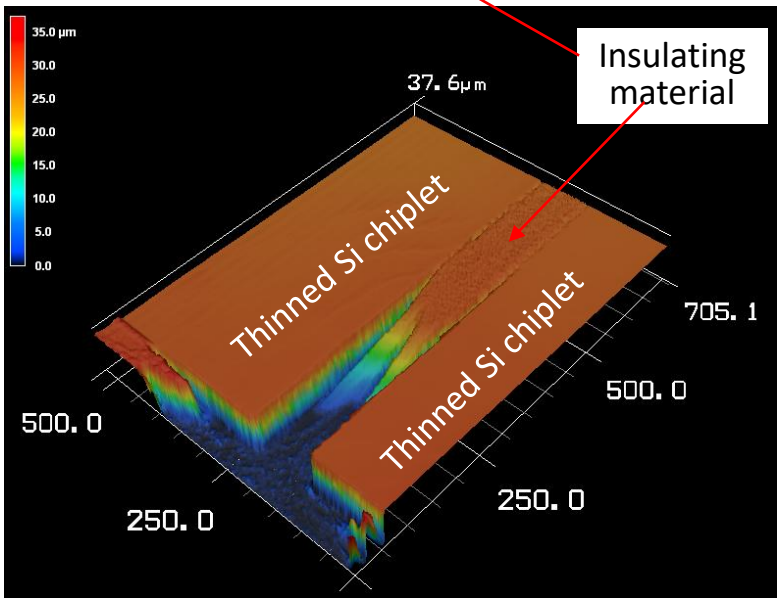


- The current thermal bonding can still see the bonding interface between Cu. In the future, we will continue to explore the impact of liquid on bonding and the optimal bonding conditions.

Simultaneous Grinding and Polishing of Multi-Chips

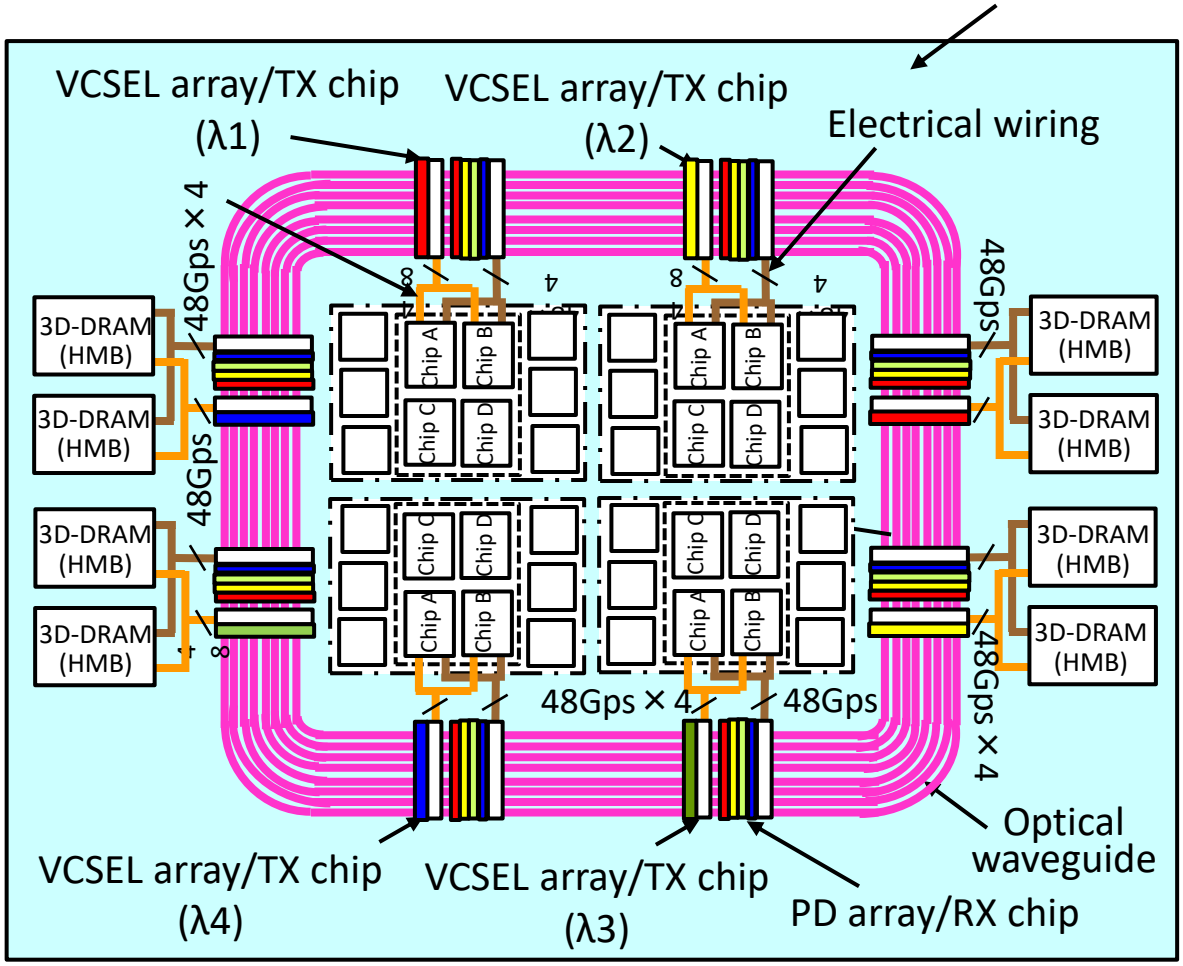






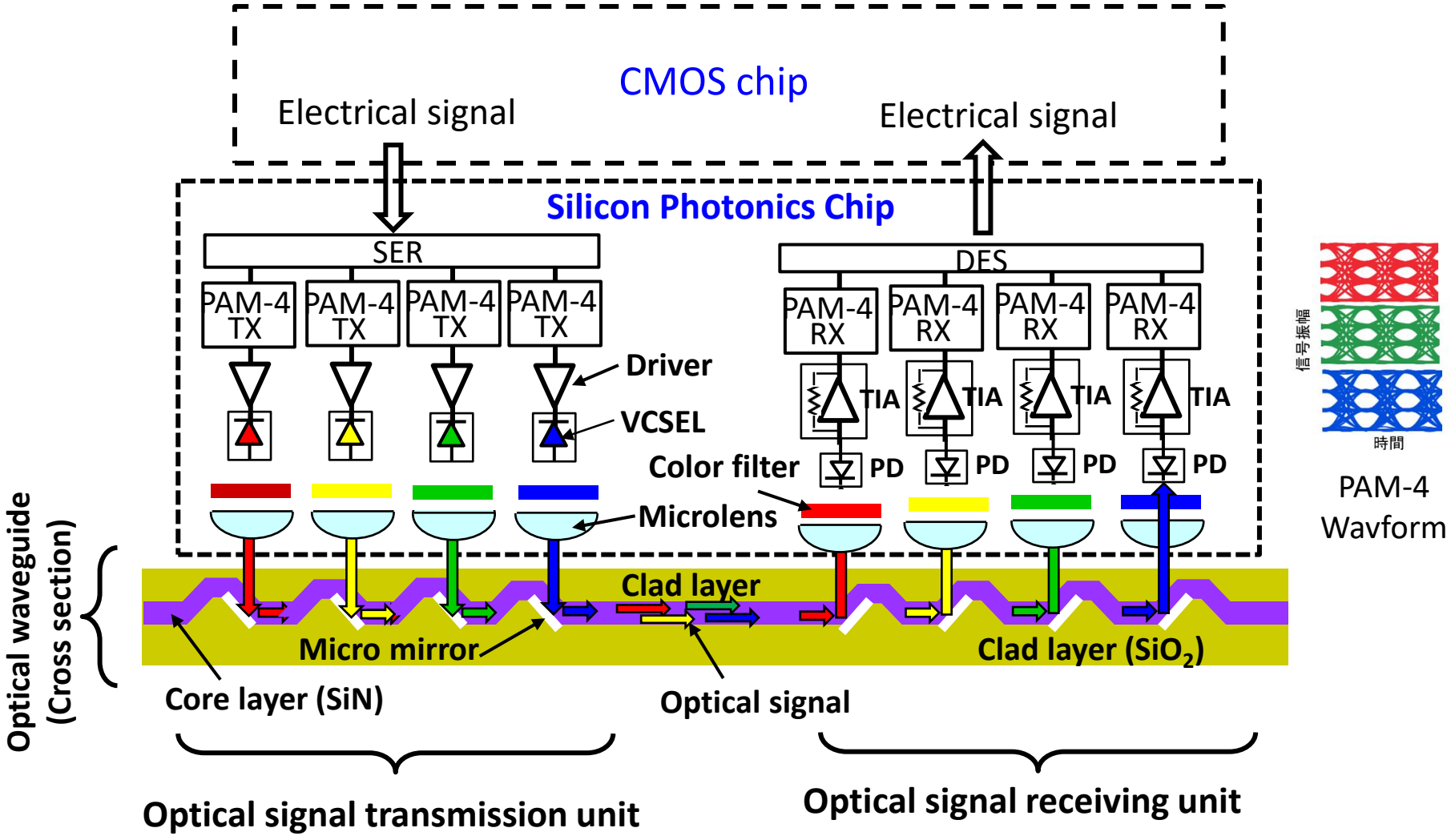
Multi-chip thinning down to 30um in thickness

Visible Light Communication for 3D Chiplet Integrated System



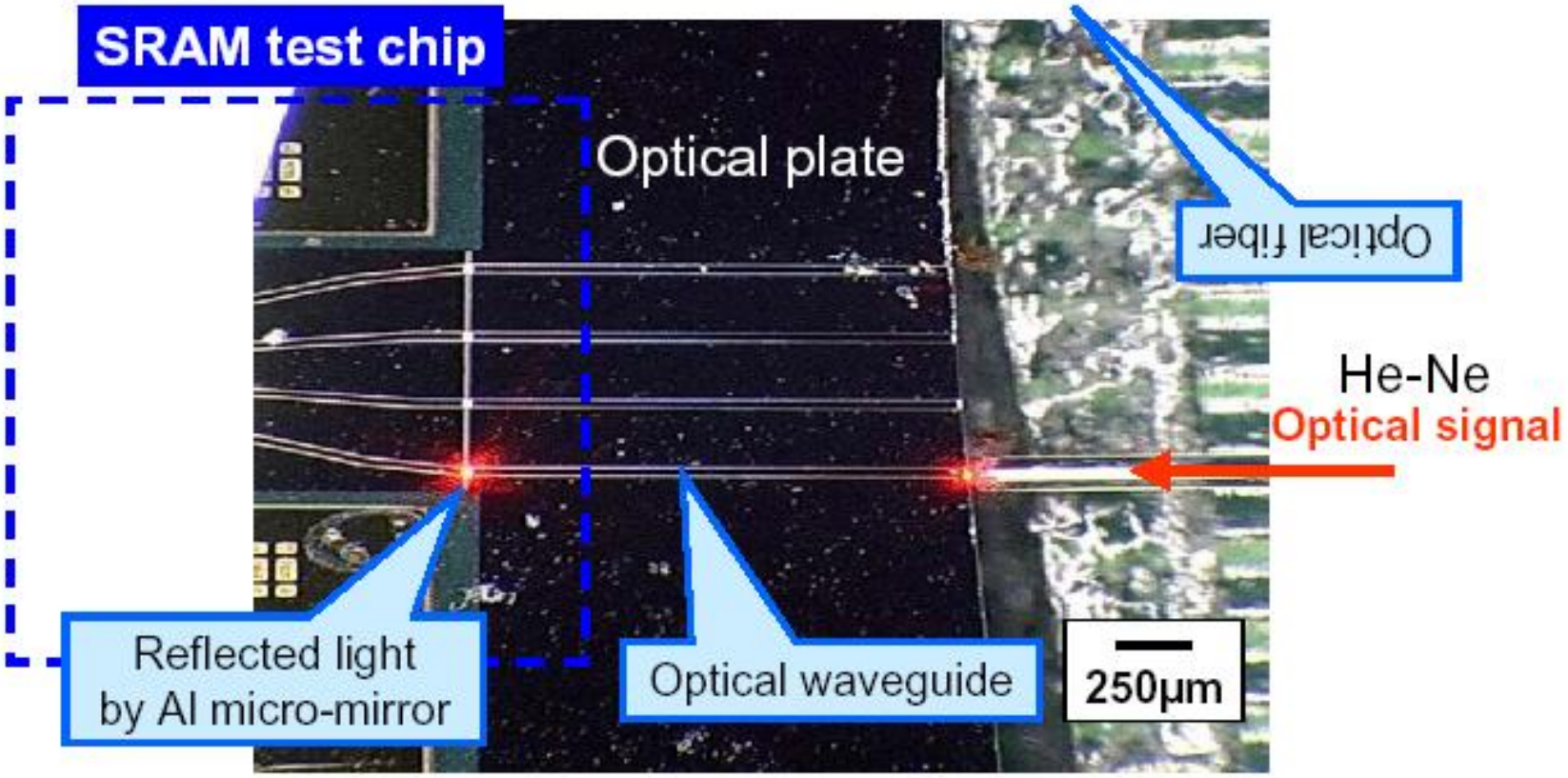
High Data-Rate Optical Interconnection by Visible Light WDM, Signal Mutiplexing and Parallel Date Transfer

Configuration of Silicon Photonics Chip for Visible Light Communication

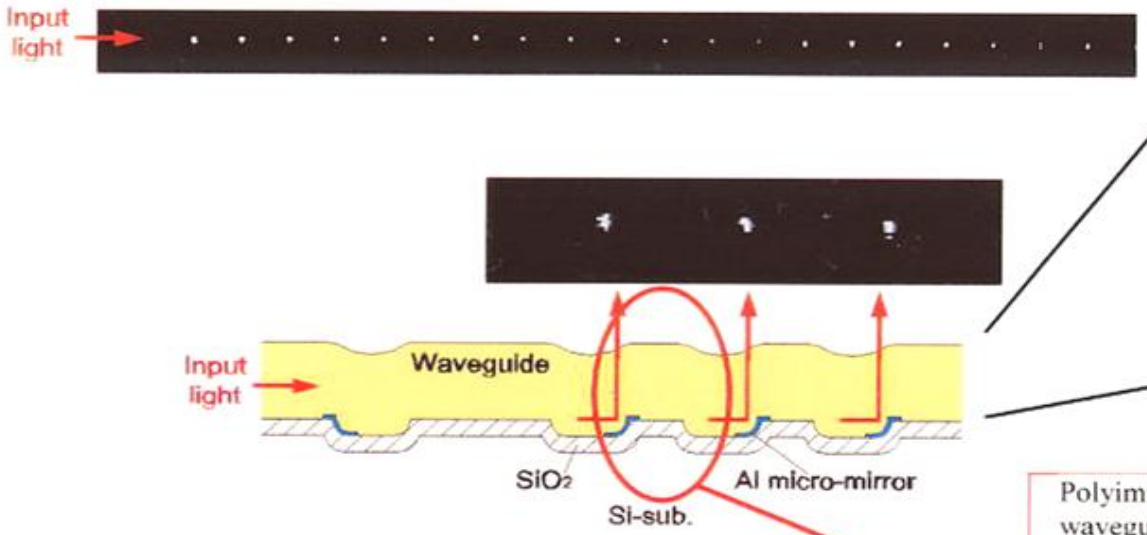


Visible Light WDM Immune to Wavelength Variation by Temperature Change

Optical writing operation to SRAM: Photograph of optical signal reflected at Al mirror

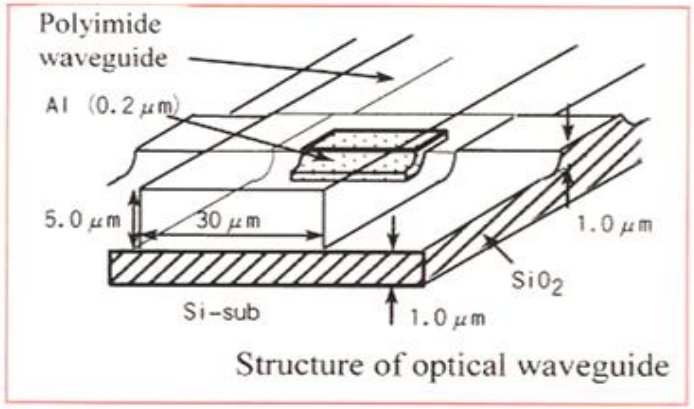


Polyimide Optical Waveguide with Micro-Mirrors



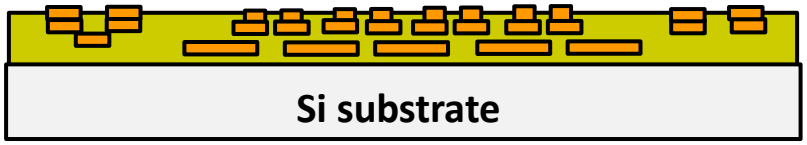
Cross-sectional structure of optical waveguide and output signal light

Output signal light emitted from the end of optical waveguide

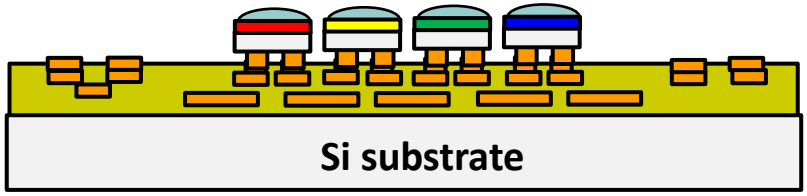


Structure of optical waveguide

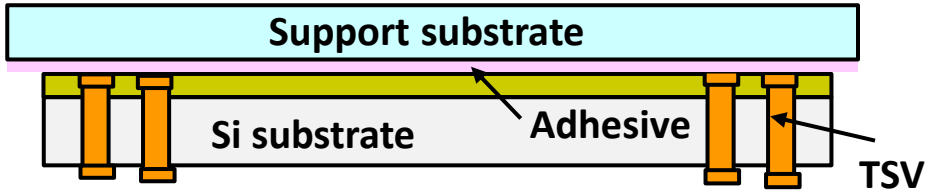
Fabrication Process of Silicon Photonics Chip



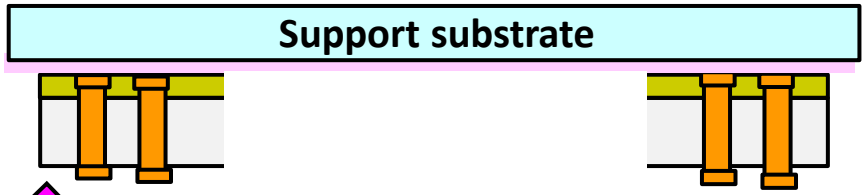
(a)Bump formation on silicon photonics chip wafer



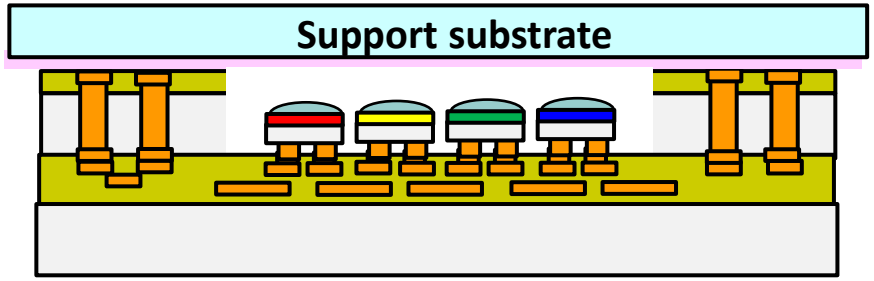
(b)VCSEL chips or PD chips bonded on silicon photonics chip wafer



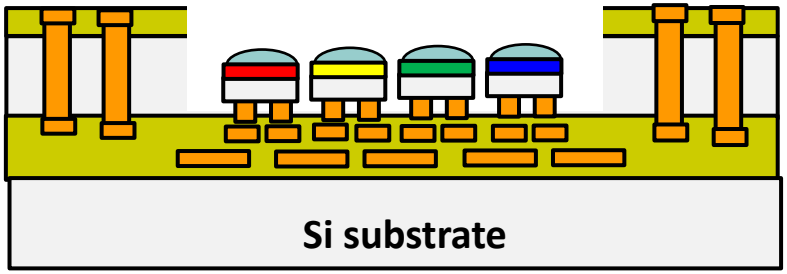
(c)Fabrication silicon interposer for silicon photonics chip



(d)Formation of silicon cavity



(e)Bonding of silicon interposer wafer with silicon cavities to silicon photonics chip wafer



(f)De-bonding of support substrate

Cross-sectional structure of silicon photonics chip

VCSEL Chiplet Integration on Glass Interposer by Self-Assembly

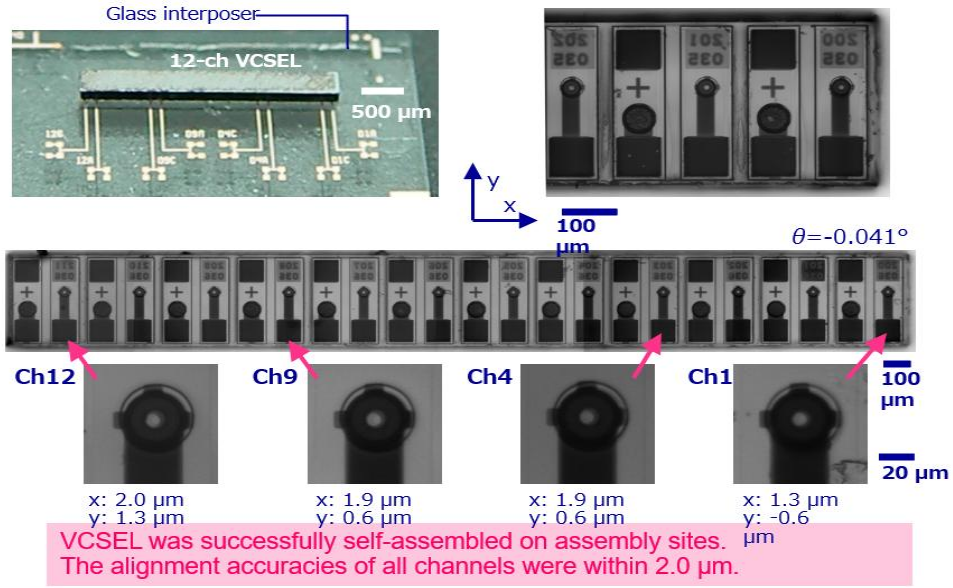
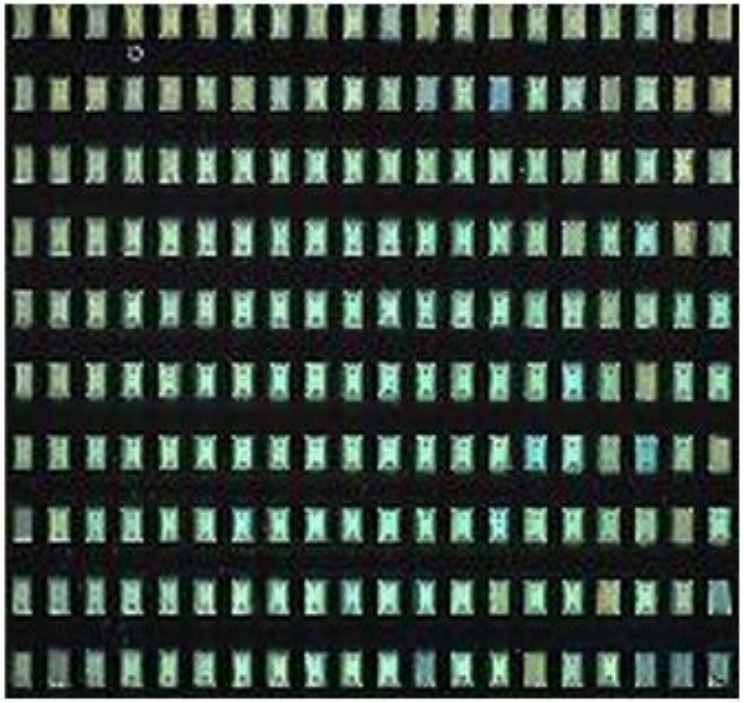
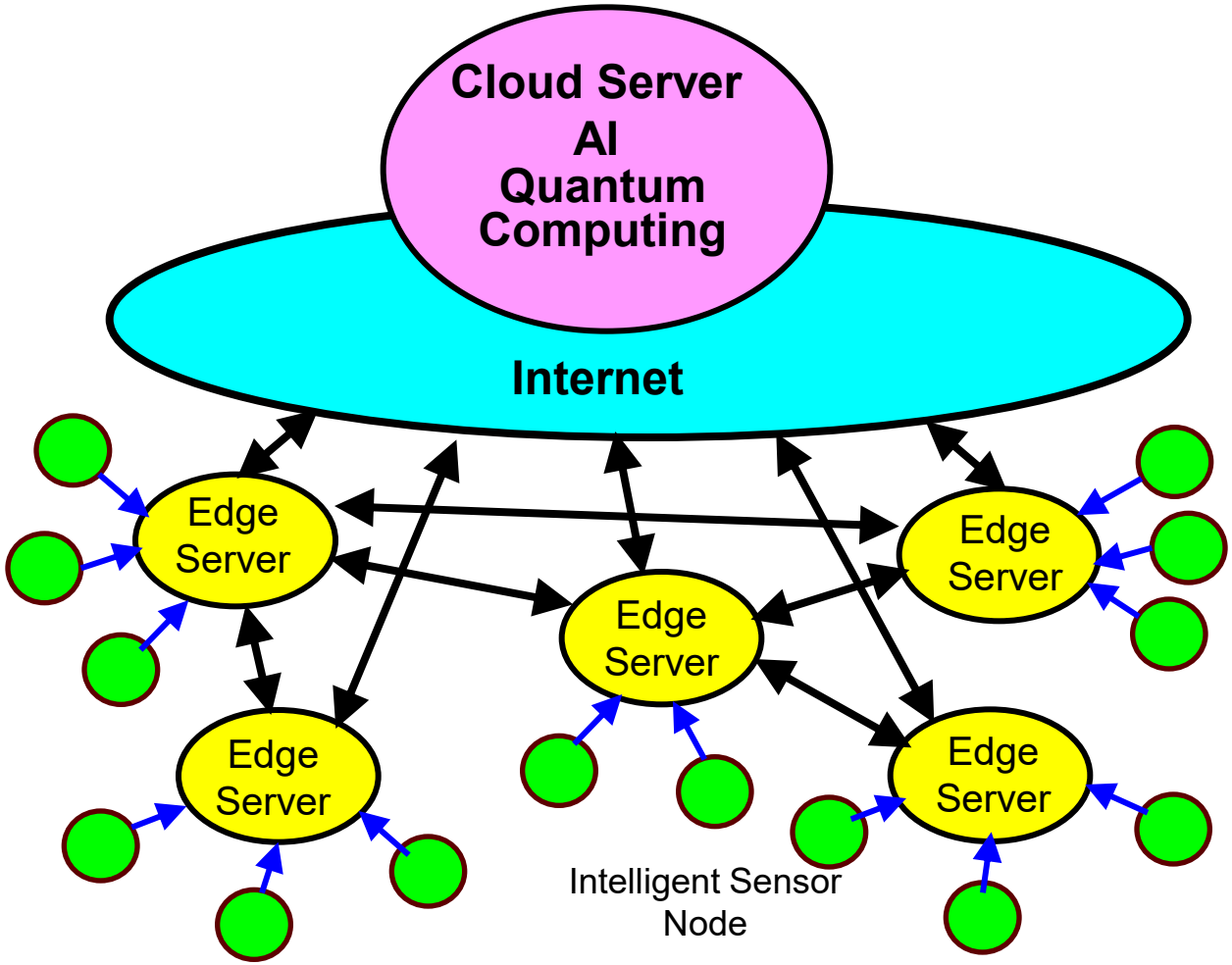


Photo of µLED Array Prepared by Self-Assembly

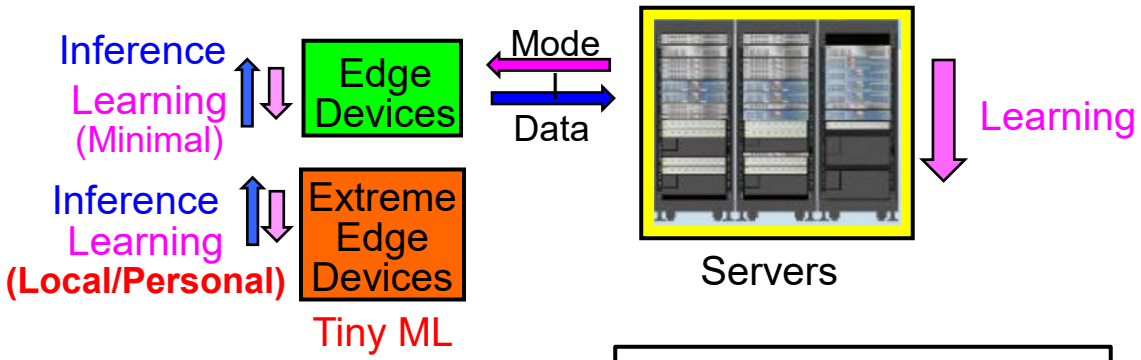


Self-Assembled Micro-LED chips
(75umx125um)

Global Network in IoT/AI/post-5G Era



Requirements for AI System and Technologies



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

Scalability of system

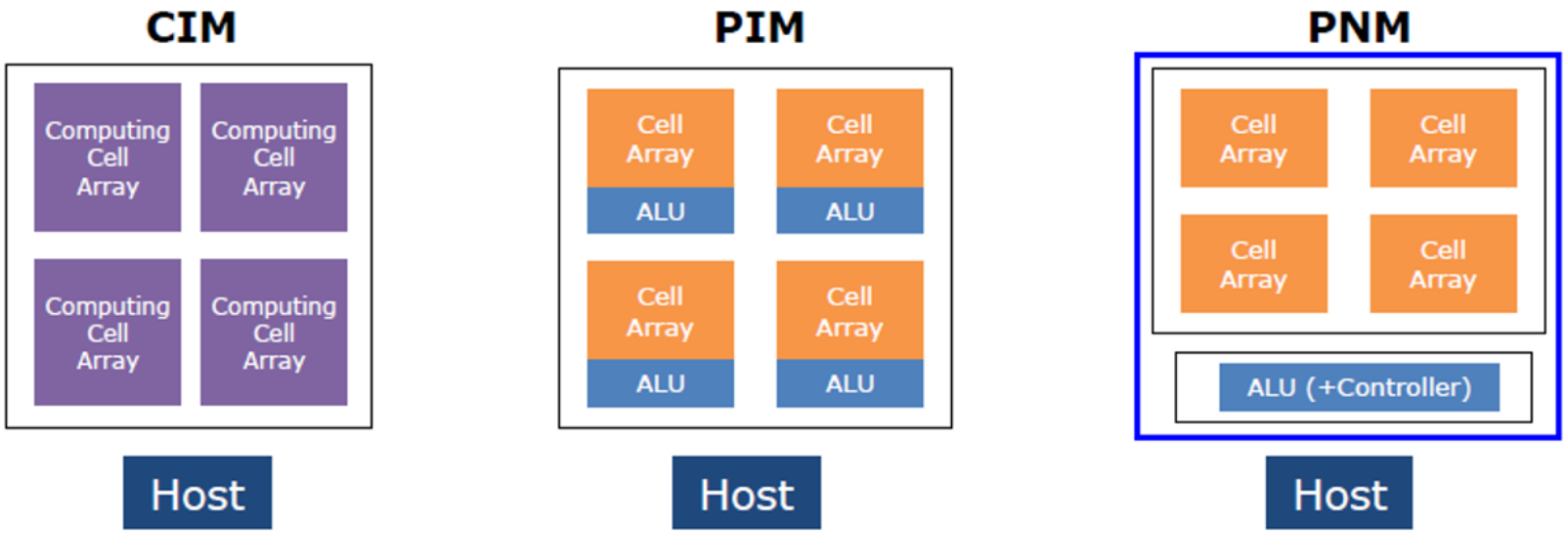
- More computing power: Highly efficient computing
- More memories: Large capacity memories with high bandwidth
- Connectivity with high data transfer capability

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

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

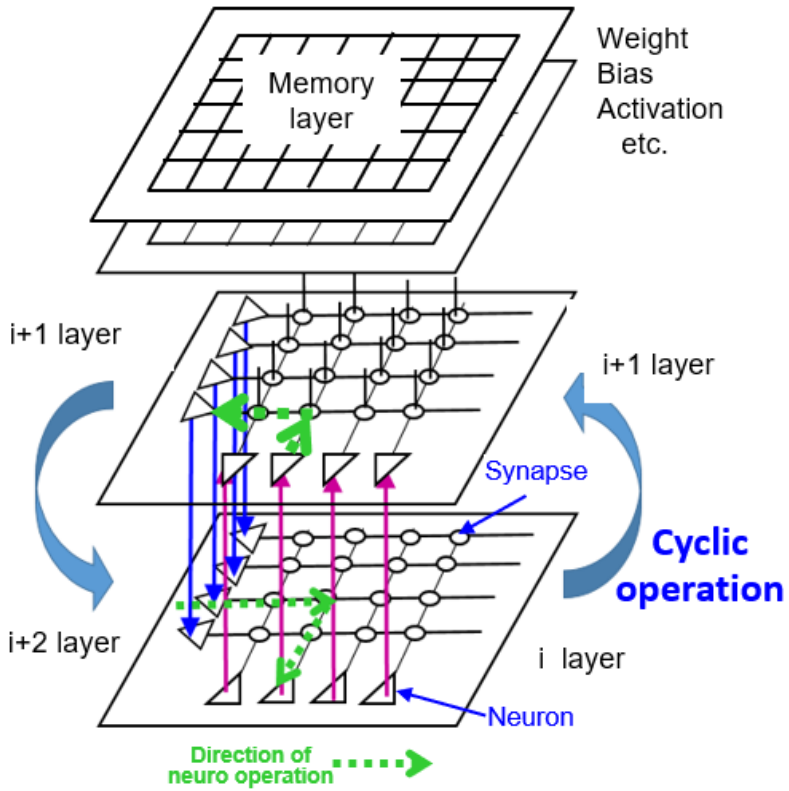
- CIM: use memory array as a processing unit
- PIM: use embedded logic near memory array as a processing unit
- PNM: use an additional chip for processing inside a memory package or a set



Kyomin Sohn, IEEE ISSCC Forum 2.6 (2024)

Cyclic Neuro Operation in 3D Stacked AI Chip

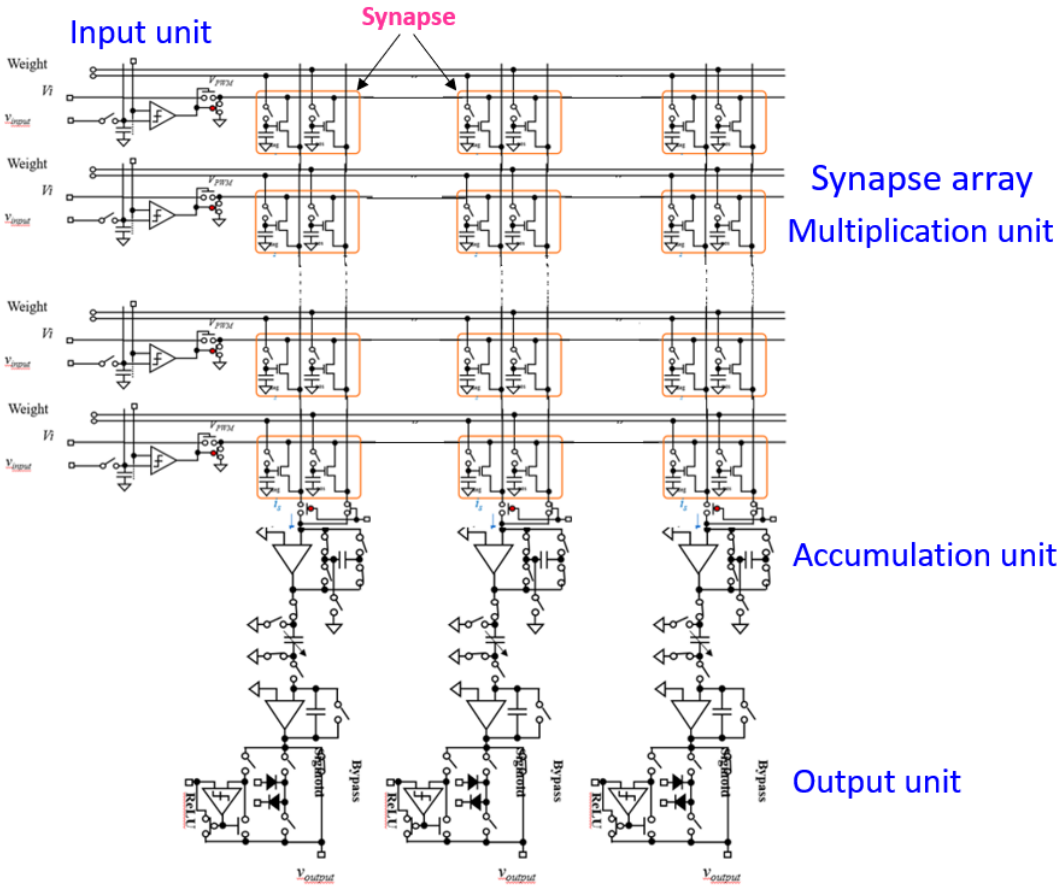
(Analog/Multivalued Neuro Operation)



3D Stacked AI Chip

NEDO AI Project (2018-2020)

MAC operation using CIM



Activation function implemented by analog circuits
ReLU, Sigmoid etc.

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

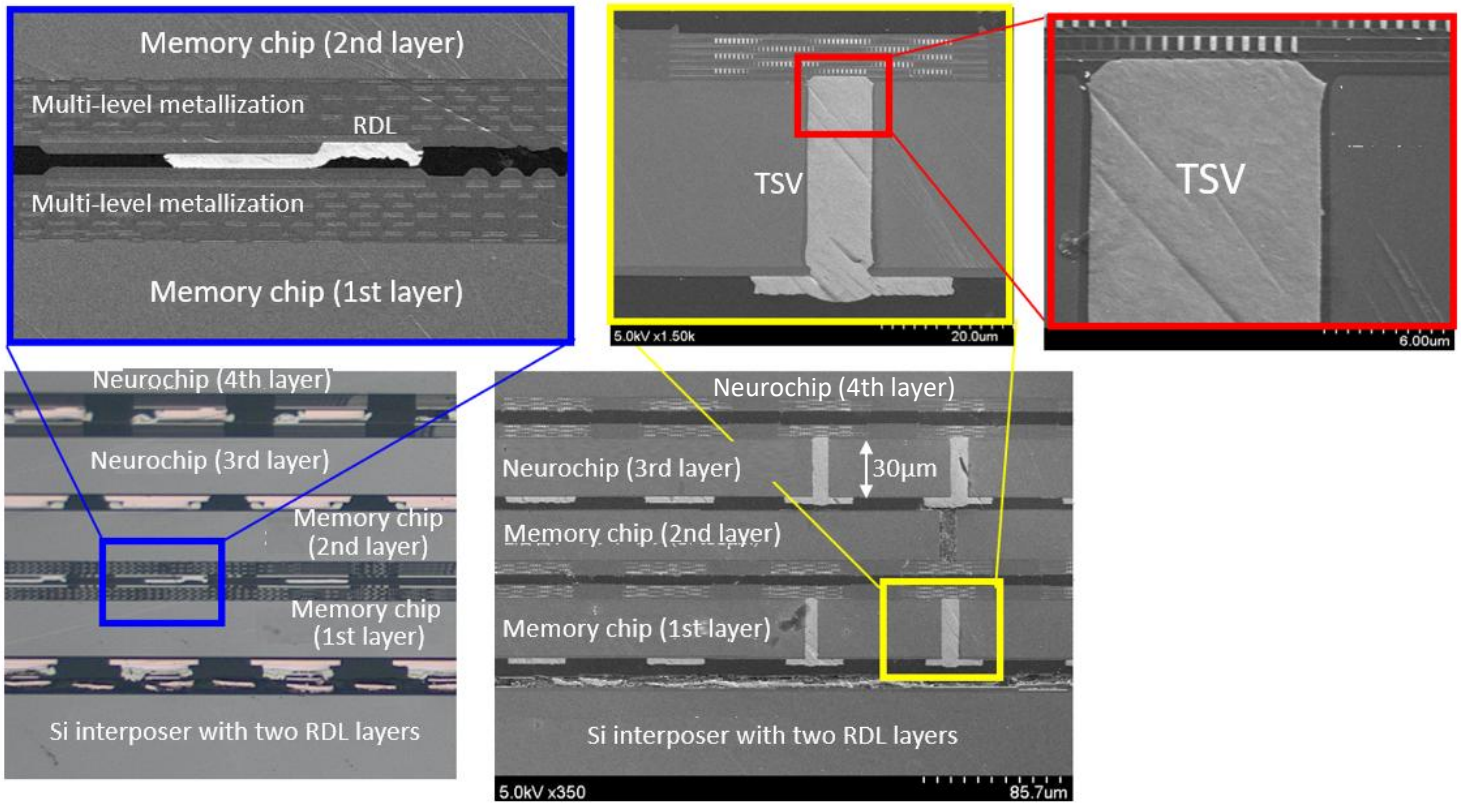
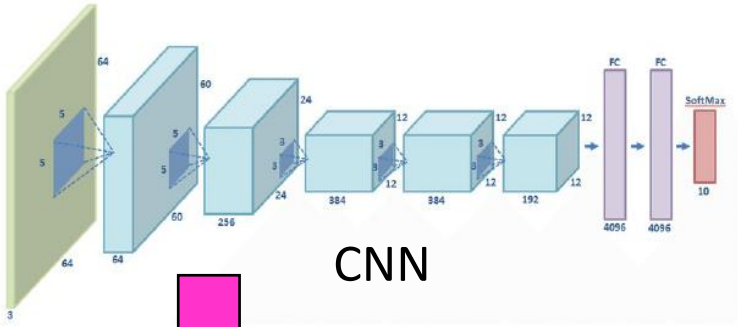
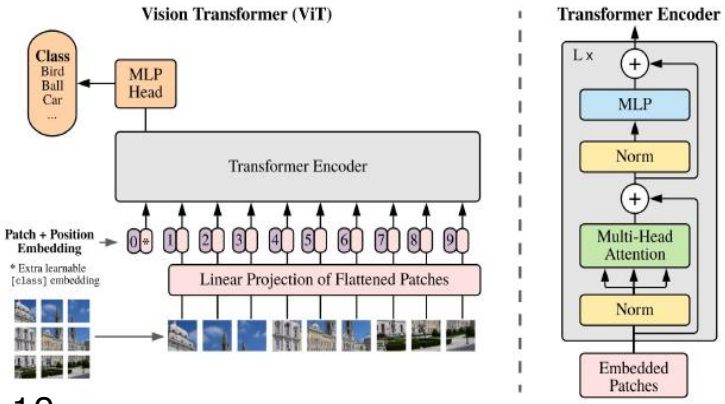


Image Recognition Using 3D Stacked AI chip

Paradigm Shift from CNN to ViT (Vision Transformer)



CNN



CIFAR-10

Vision Transformer (ViT)

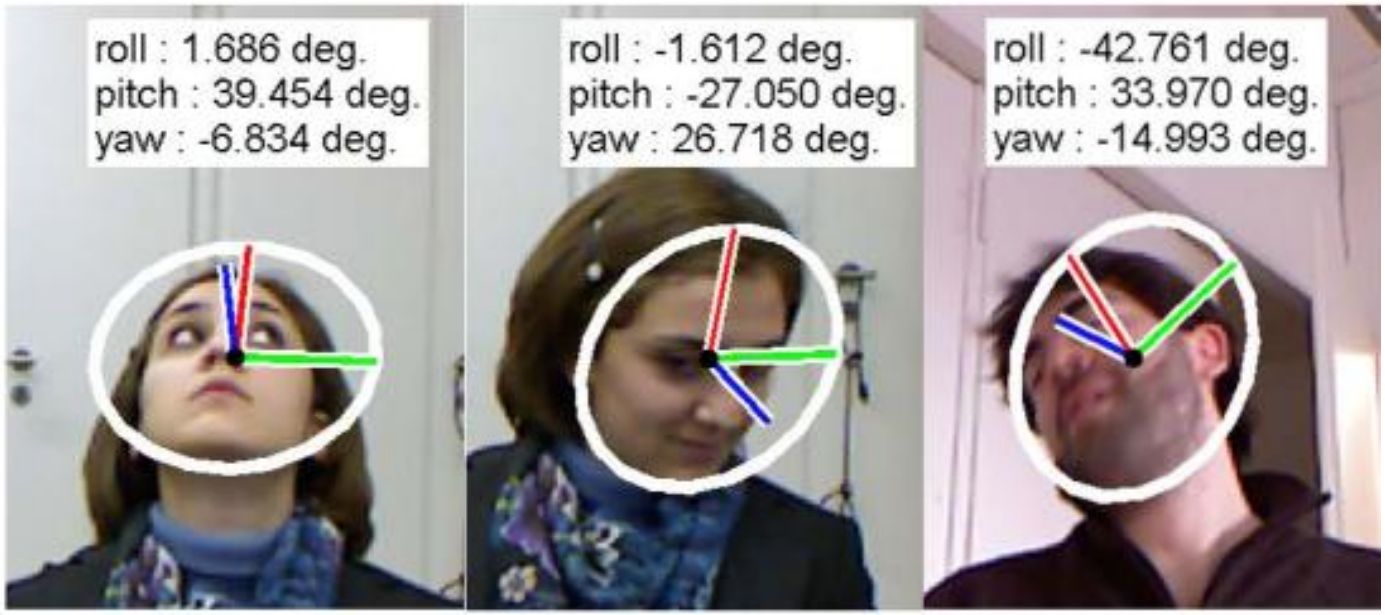
- We can significantly reduce the number of matrix product operation (MPO) in ViT.
- ViT is suitable for Edge application.

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

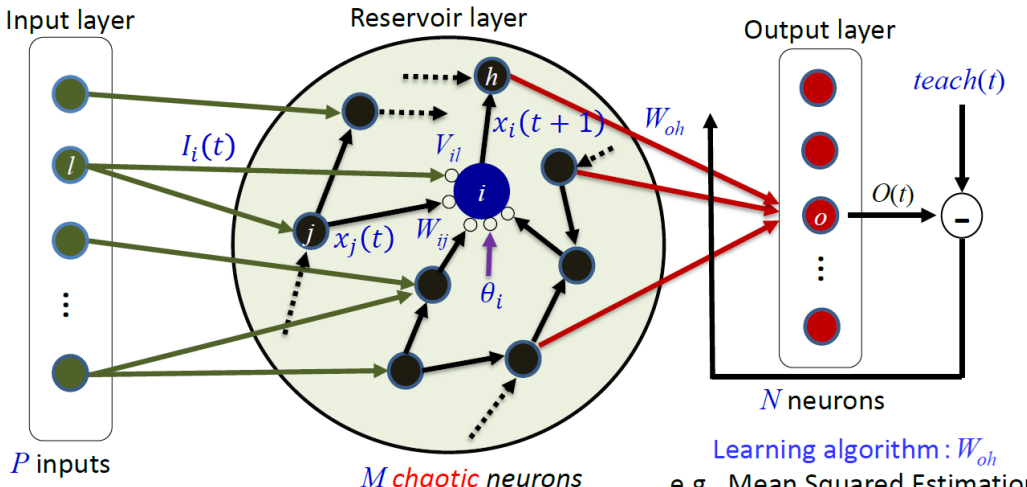
Face Recognition Using 3D Stacked AI chip

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



Learning algorithm : W_{oh}
e.g., Mean Squared Estimation
(linear regression)

Internal State of neuron i

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

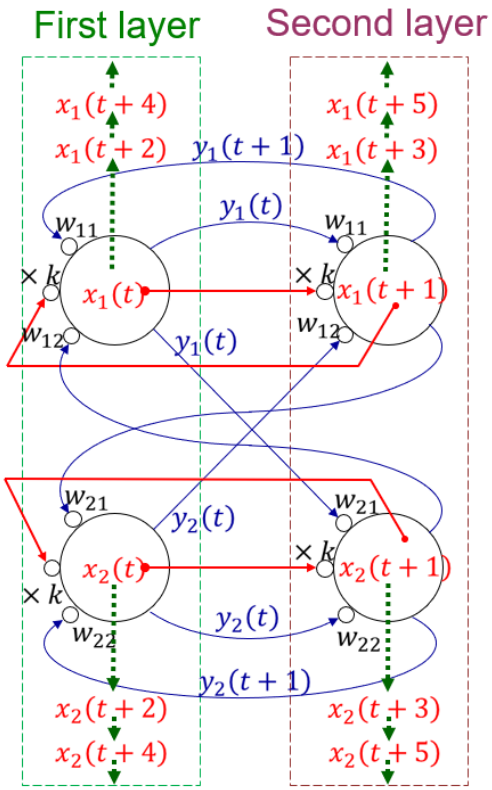
Output of neuron i

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

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

k: dumping factor of the refractriness
α: scaling factor θ_i: threshold
$f(\cdot) = 1/(1+\exp(-x/\epsilon))$

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)

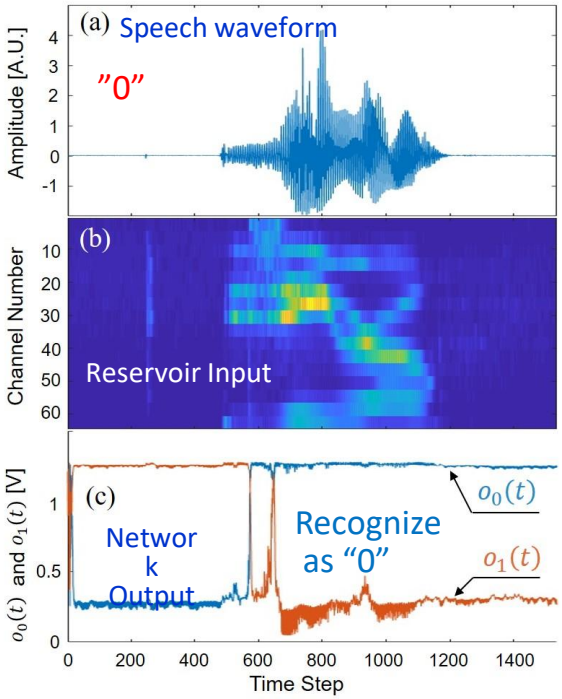
Voice Recognition Using 3D Stacked AI chip

Number of Reservoir Neuron	Connectivity within Reservoir
64	22 %

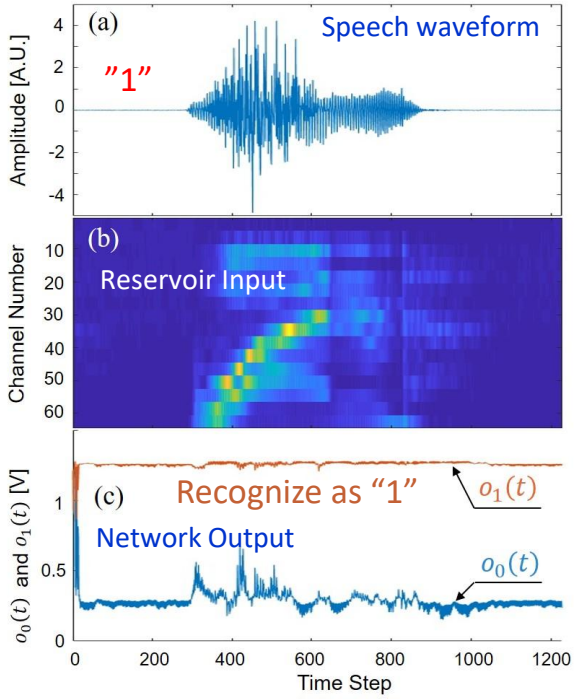
Input Connectivity	Output Connectivity
6.25 %	100 %

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

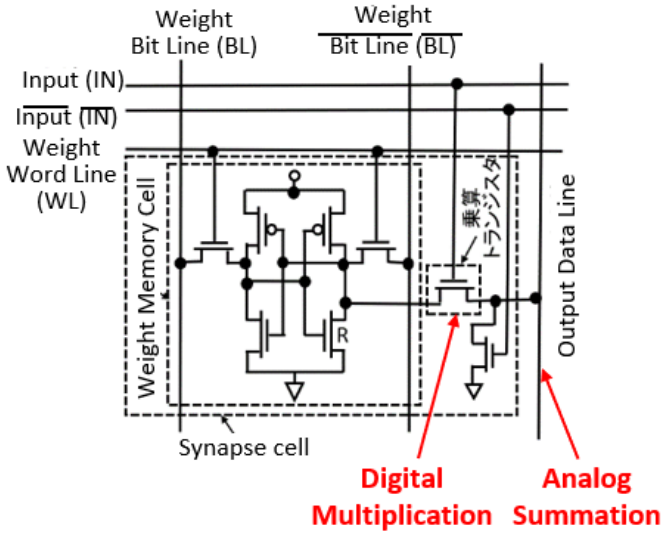
Example of network response after learning



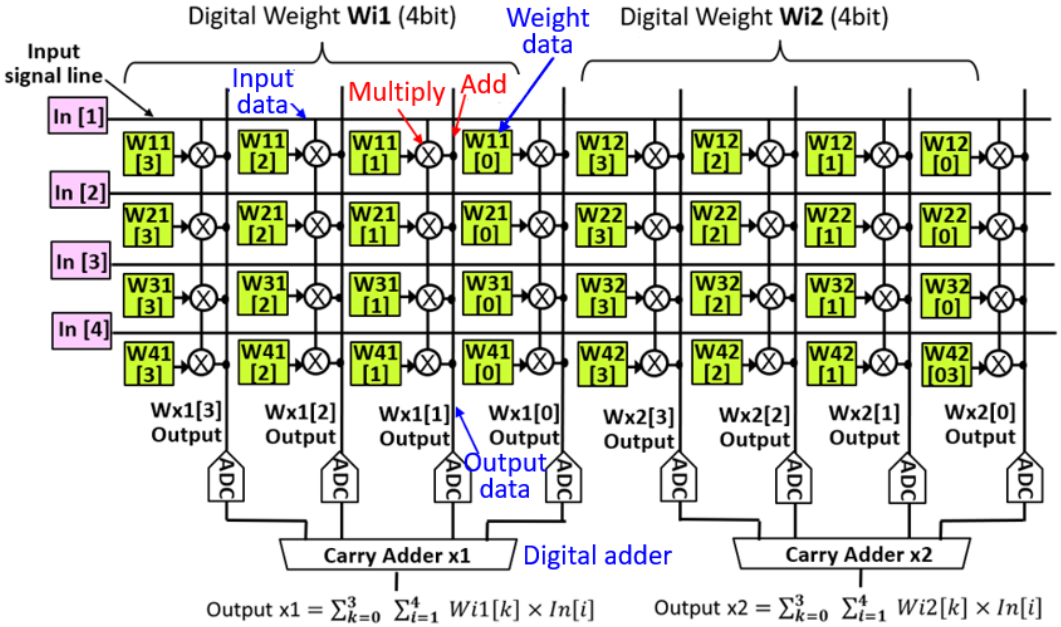
100 % recognition for ten different voices



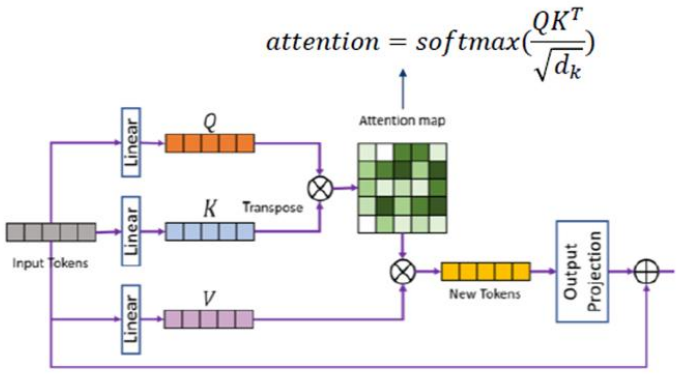
CIM Based AI Chip Using SRAM and Transformer Algorithm
(Mixed Analog-Digital Multivalued Neuro Operation)



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



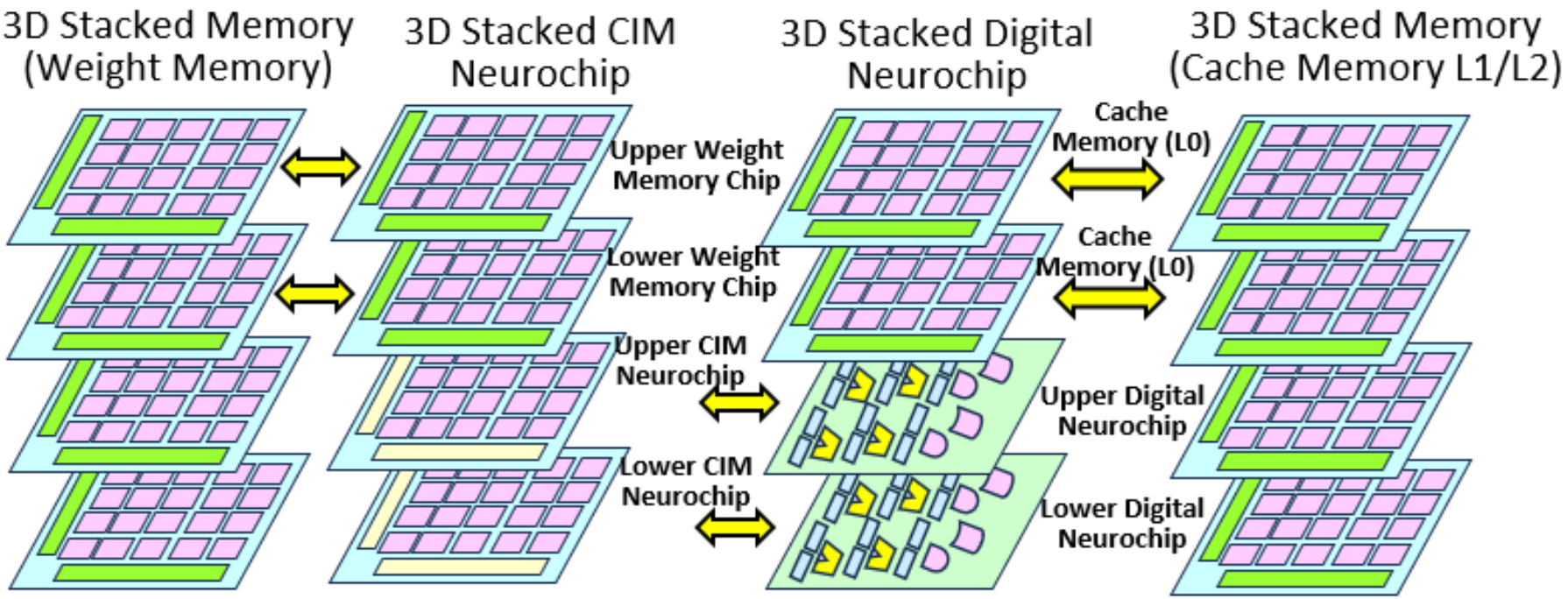
CIM Basic Synapse Circuit Array



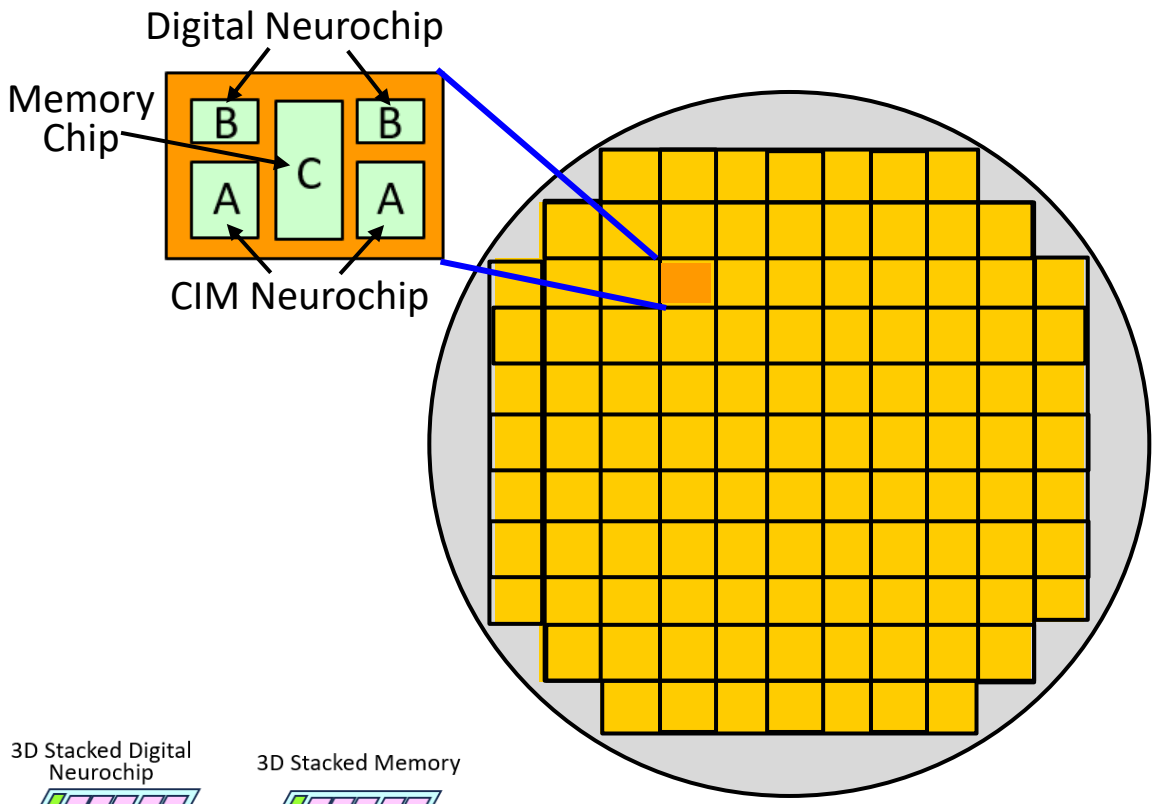
Transformer Algorithm

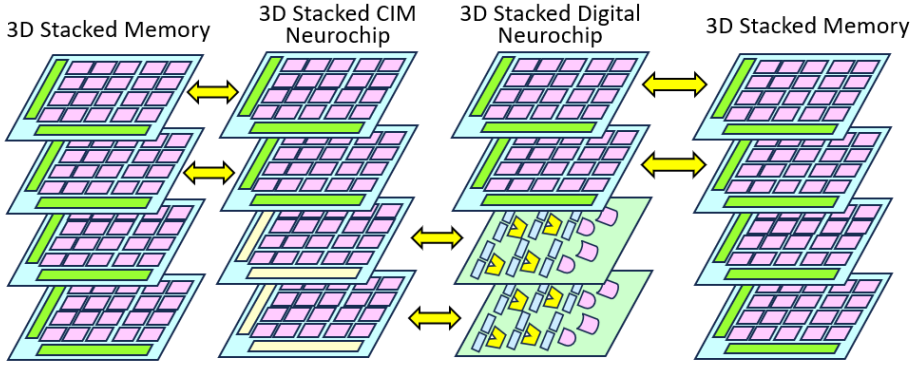
Configuration of 3D AI Chip

Direct Data Transfer between Adjacent 3D Stacked Neuro Chips

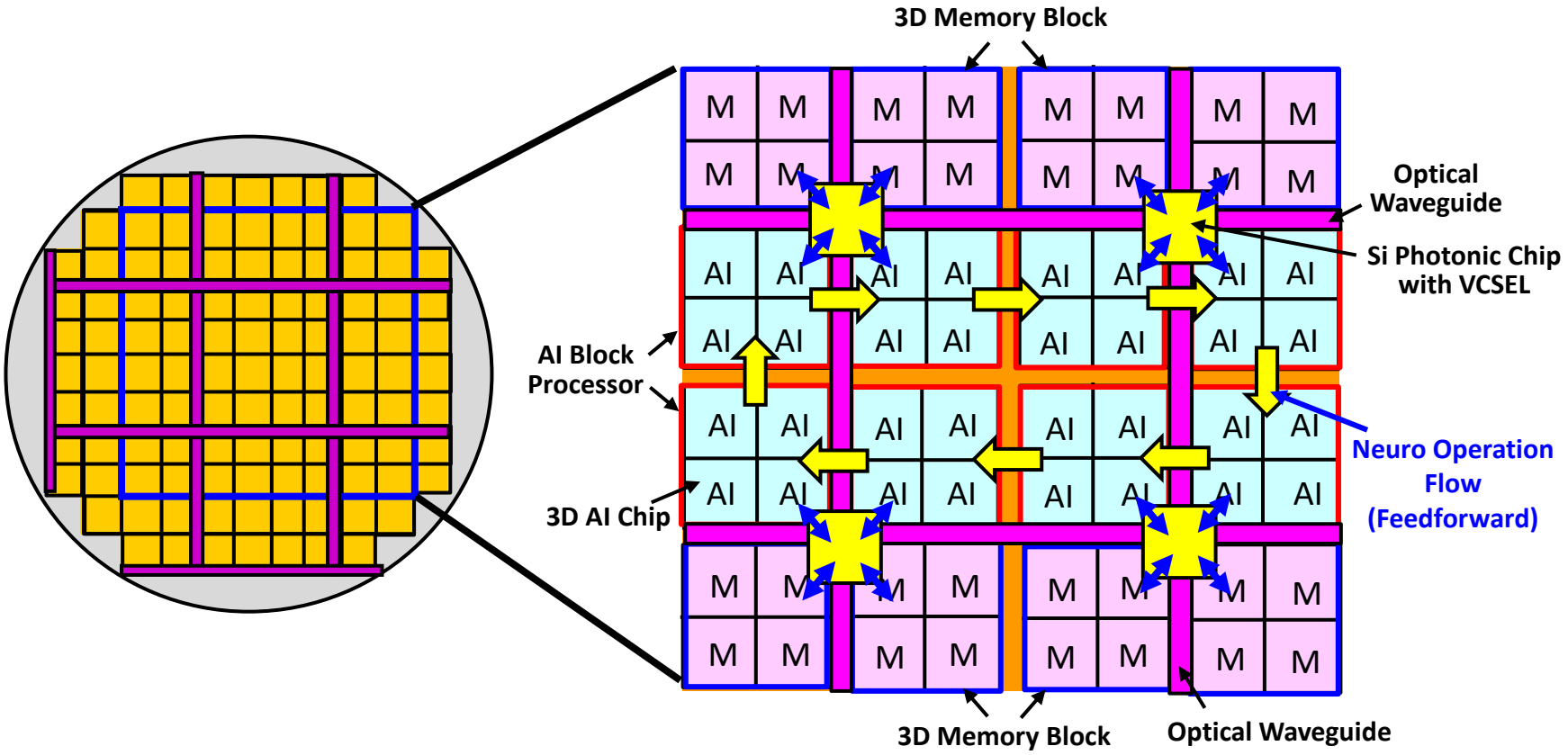


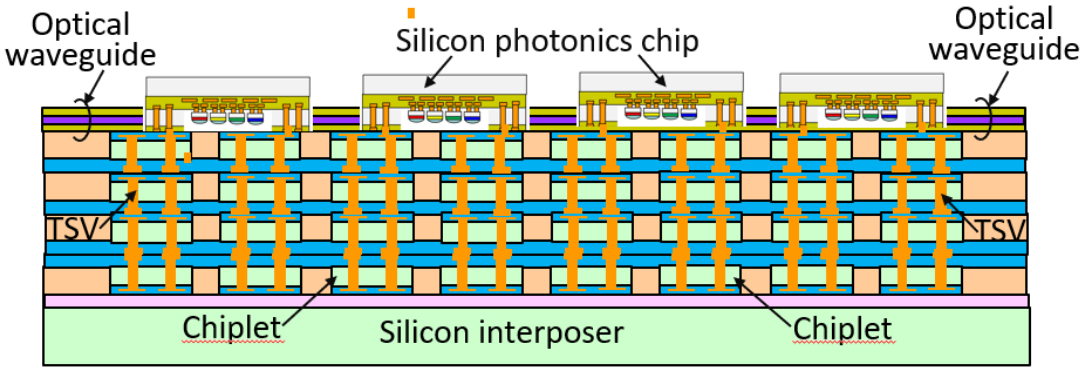
Wafer-Scale 3D Chiplet Integration





Wafer-Scale AI System by 3D Photonic Chiplet Integration

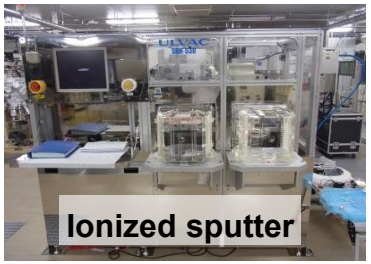




Conclusions

- Key technologies of reconfigured wafer-to-wafer 3D chiplet integration.
- Silicon photonics for visible light communication.
- Fabrication and evaluation of 3D stacked AI chip with cyclic neuro operation.
- Wafer-scale AI system by 3D photonic chiplet integration.

Thank you for your kind attentions !



12-inch 3D Production Line in Tohoku Univ. GINTI（Global INTegration Initiative）