

# AI Semiconductor (On-Device AI) Present and Future

**2024. 03. 26**

**KAIST**

**Hoi-Jun Yoo**  
**ICT Chair Professor**  
**Director, AI-PIM Center**  
**Dean, AI Semiconductor Graduate School**

**KAIST** 

# Contents

---

- 1. What is AI Semiconductor**
- 2. Present AI Semiconductors**
- 3. Future of AI Semiconductor**

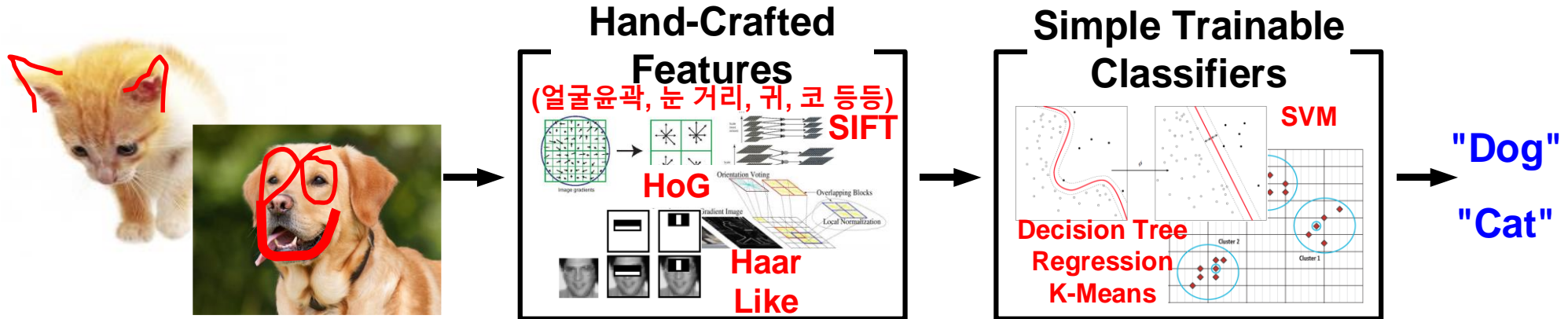
# Contents

---

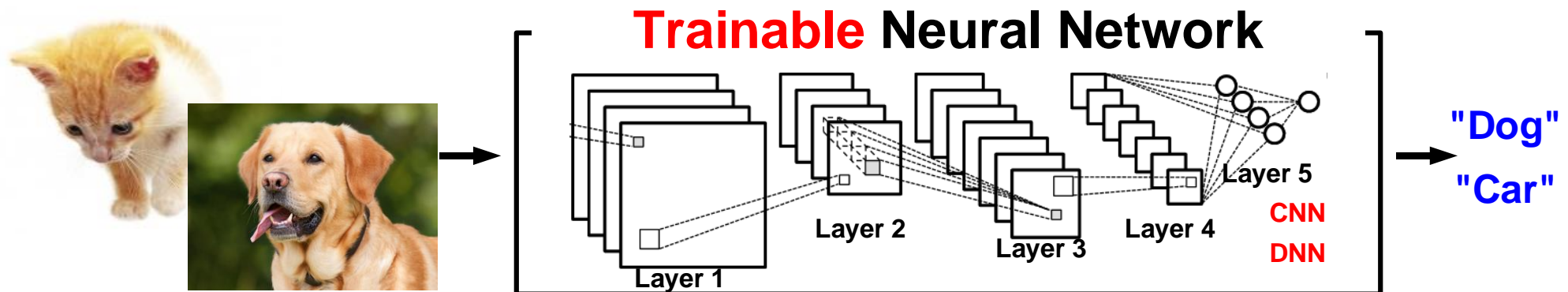
- 1. What is AI Semiconductor**
2. Present AI Semiconductor
3. Future of AI Semiconductor

# AI & Deep Learning

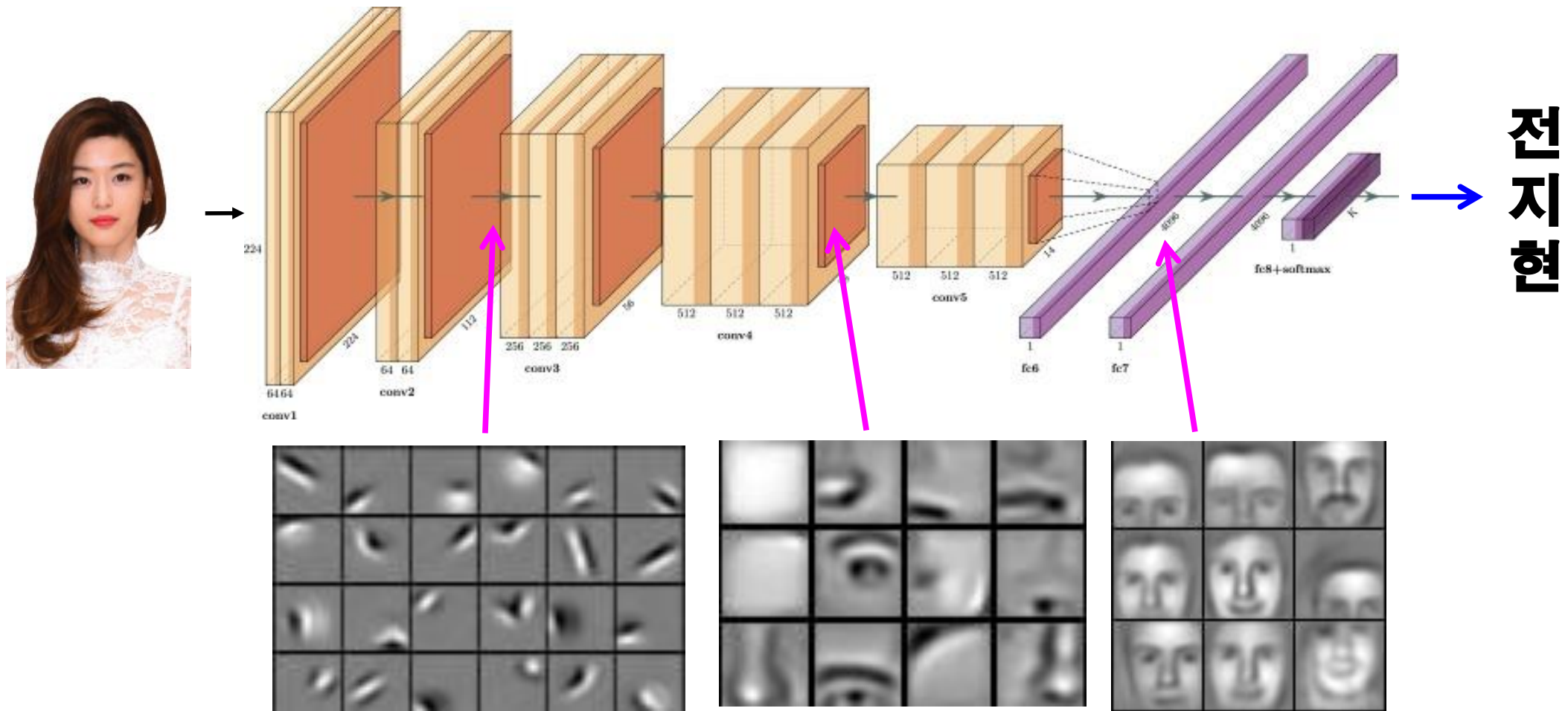
- Conventional AI



- Deep Learning (DNN): Learn by Data



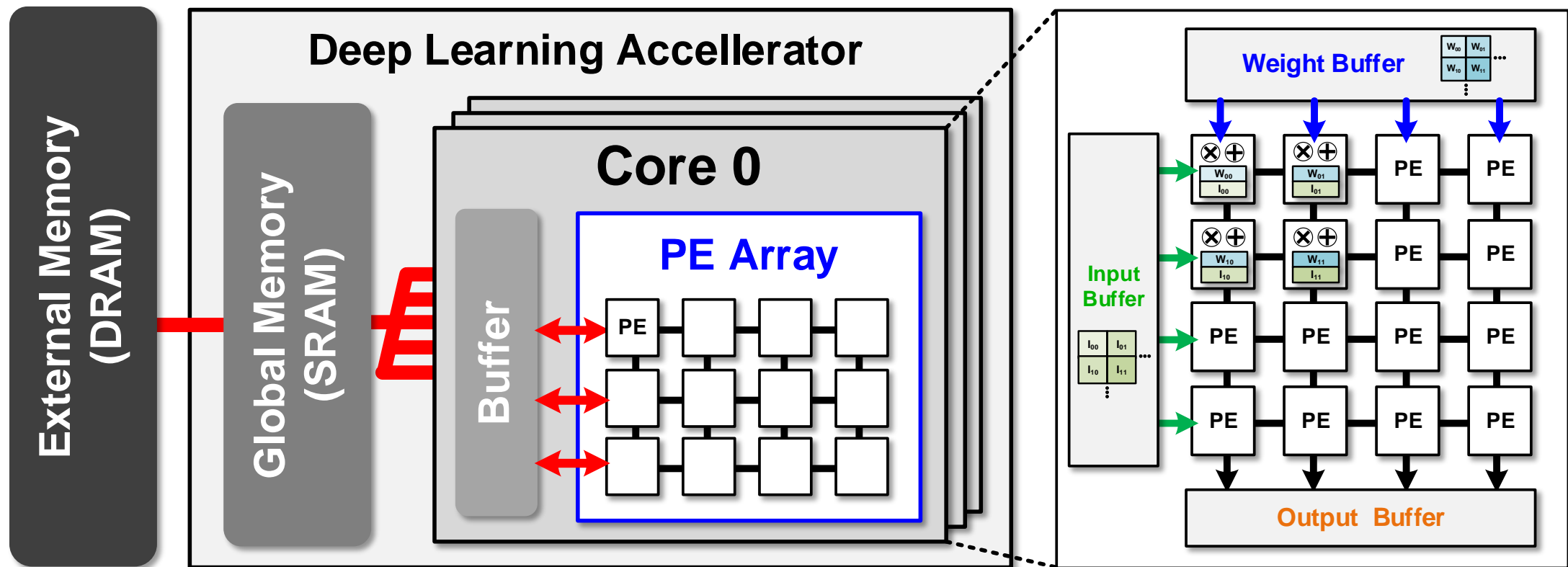
# Deep Neural Network (DNN)



# Neural Processing Unit (NPU)

- **Basic NPU Architecture**

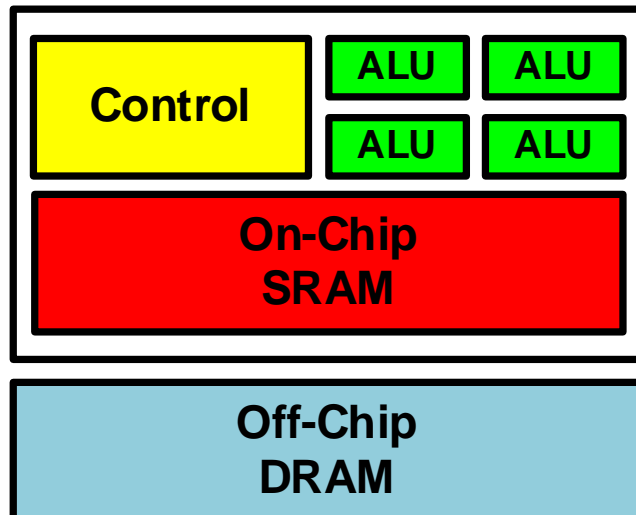
- Fetch Inputs and Weights from DRAM or SRAM
- Matrix Multiplication and Addition in PE Array



# Evolution of DNN Processors

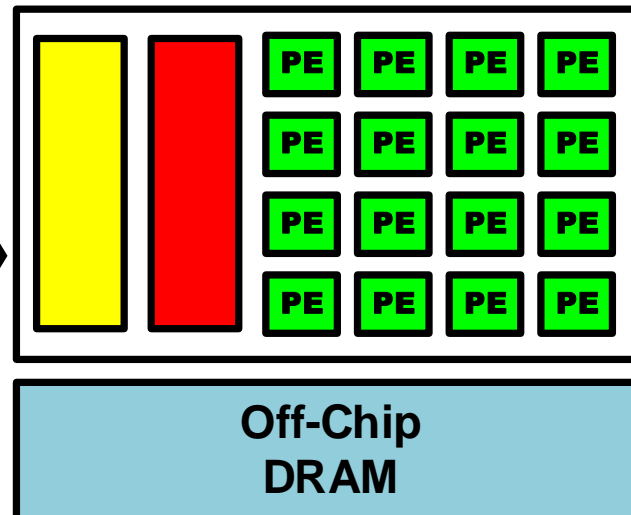
- CPU : Low Performance of DNN
- GPU : High Power
- NPU : Optimized for DNN Operation

## CPU



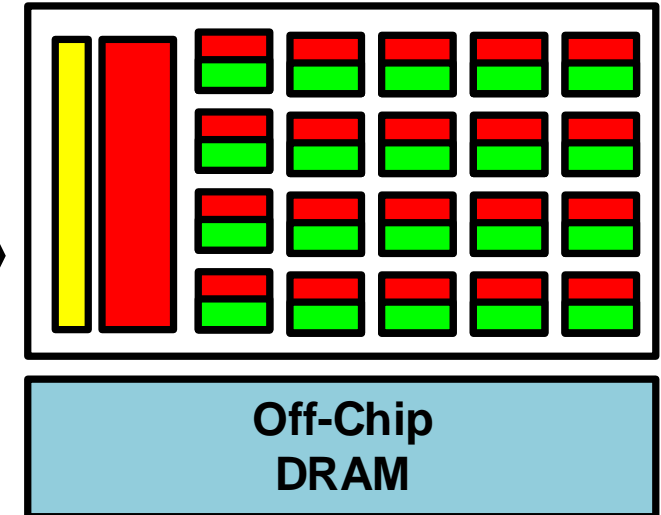
- <10 Processing Cores
- General Purpose
- Floating Point Operation
- SW Programmability

## GPU



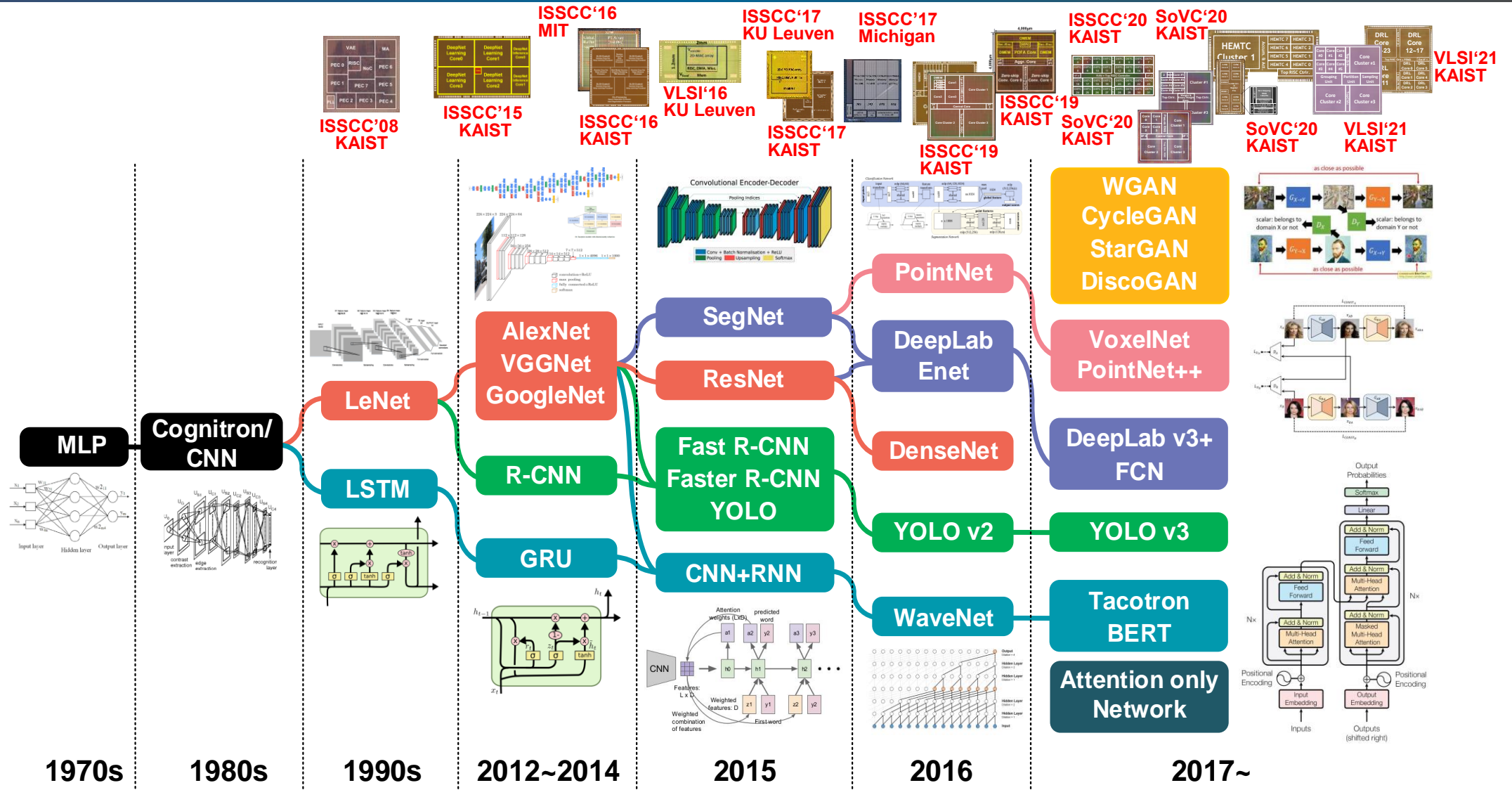
- ~1K FP PEs
- Floating Point Operation
- Matrix Computation
- CUDA Programmability

## NPU



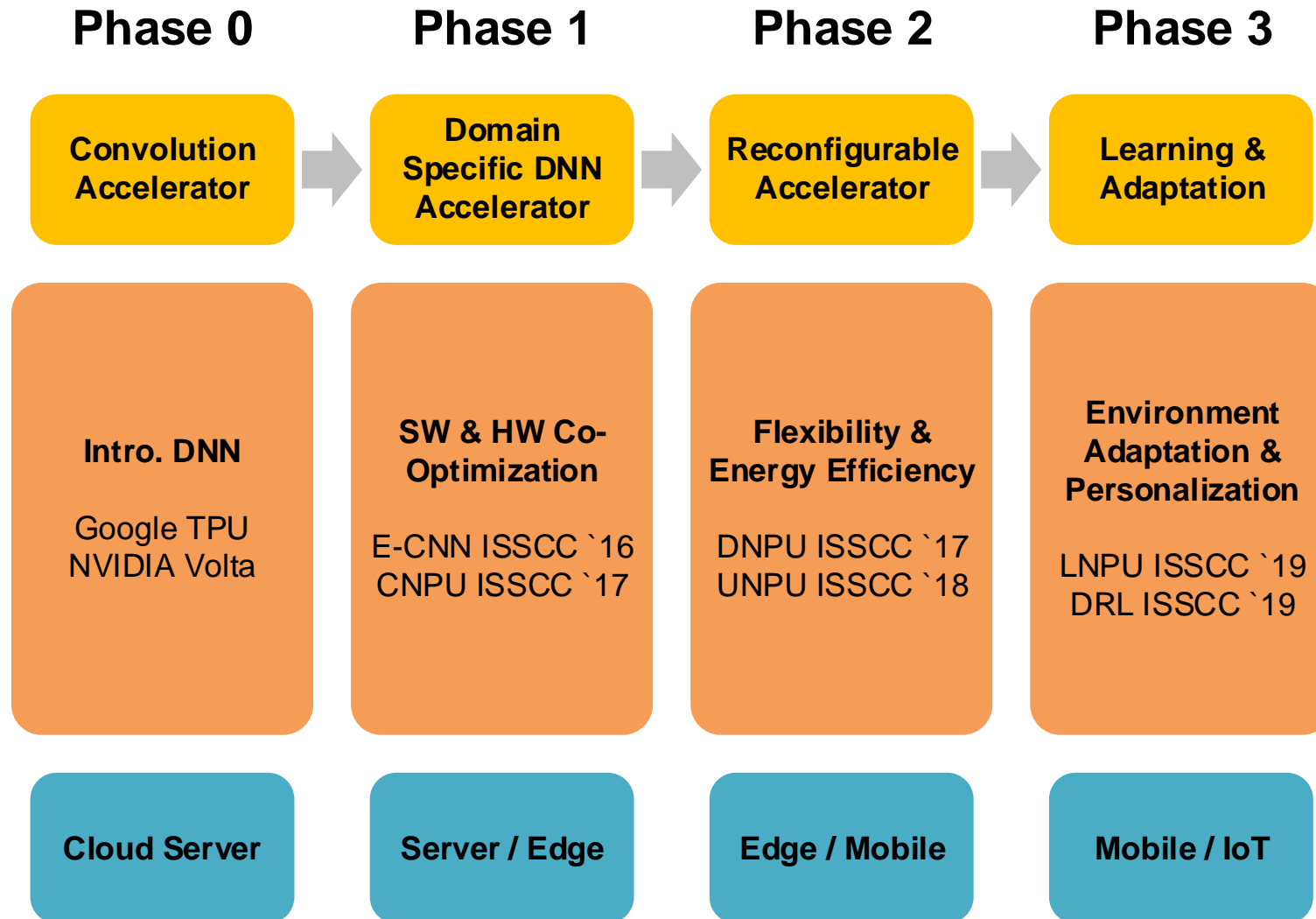
- >10K Integer PEs
- FP/Integer Operation
- Convolution Operation
- Data Reuse

# Developments of DNN

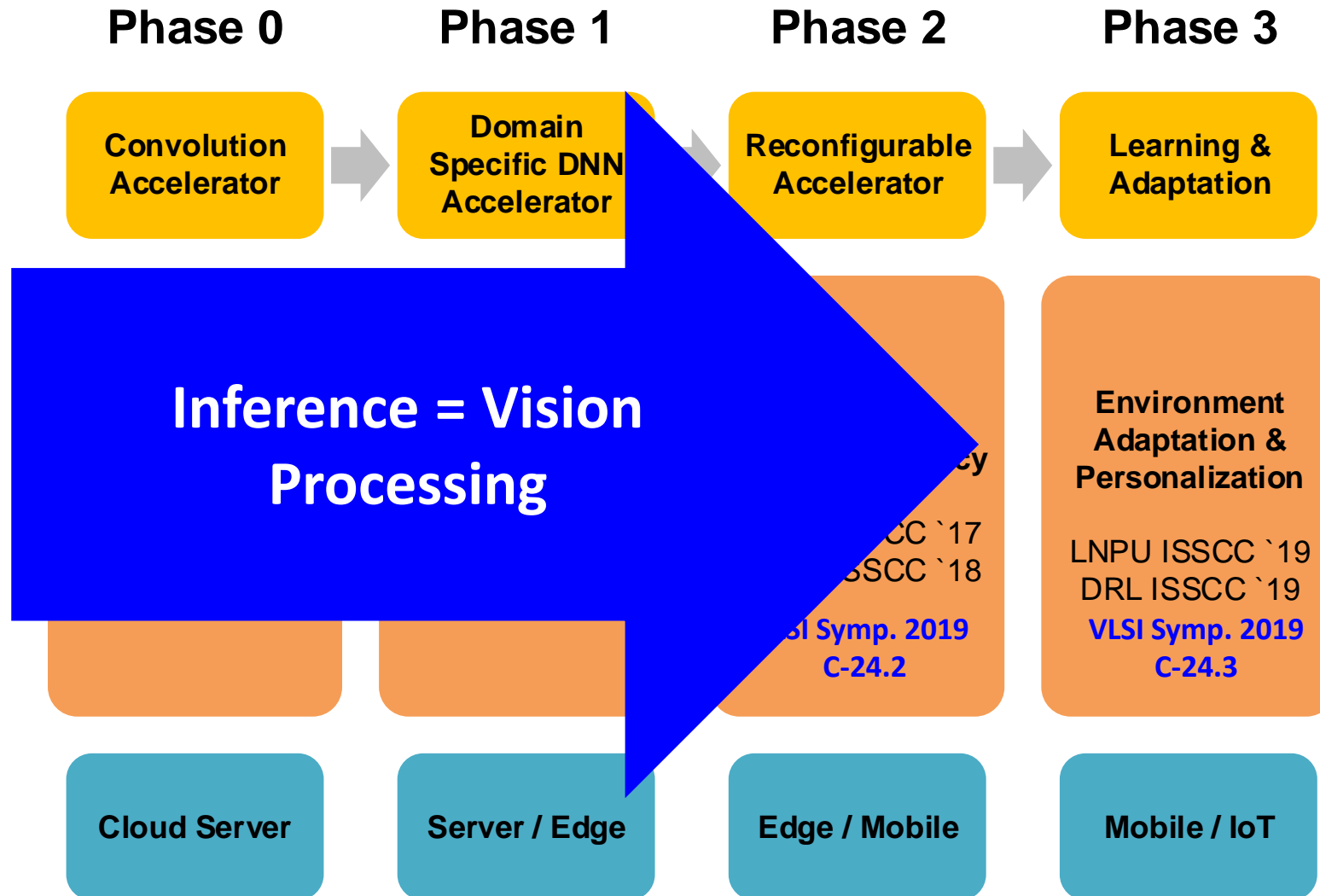




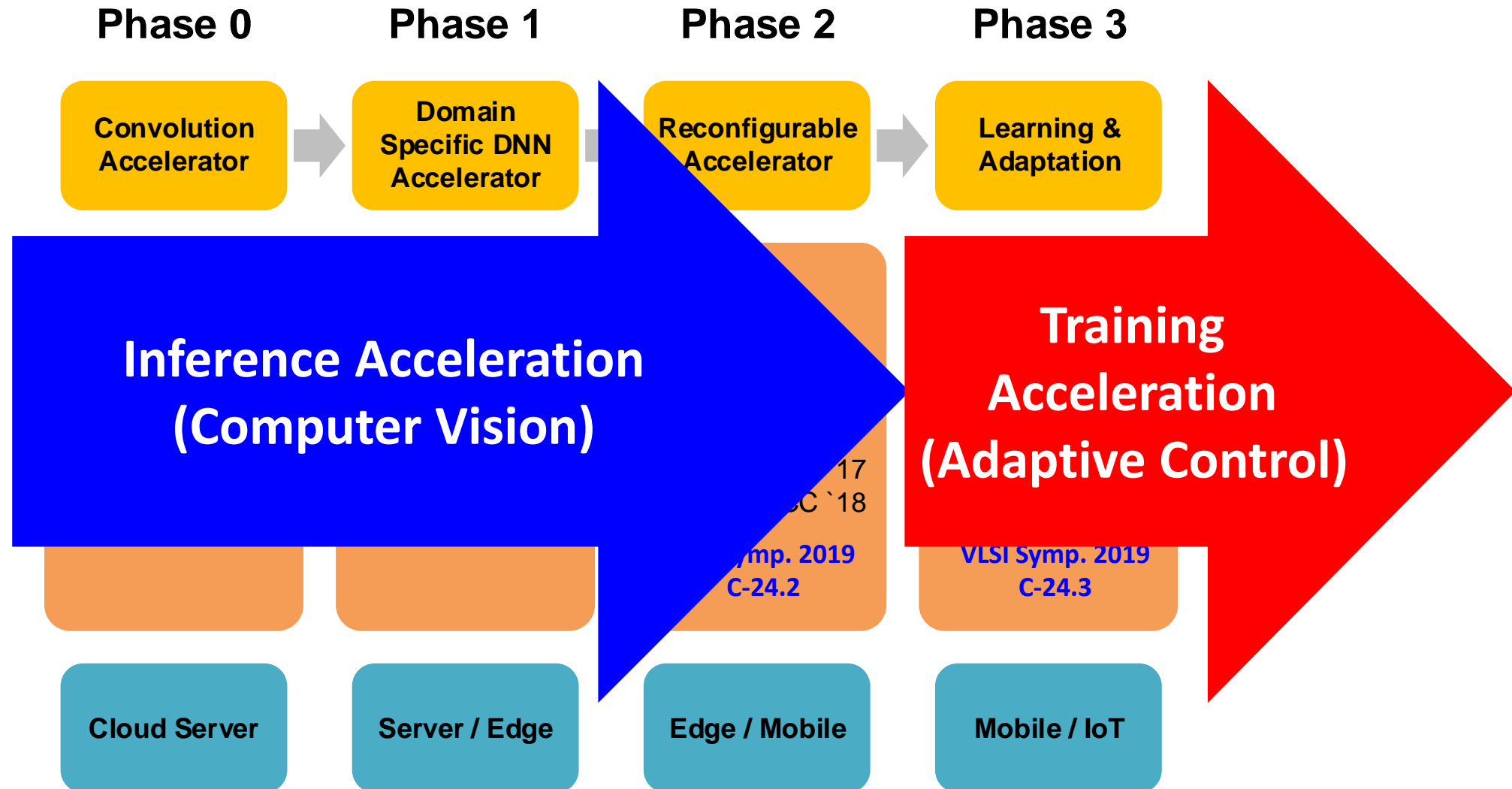
# Evolution of DNN Accelerators



# Intelligence Revolution



# Intelligence Revolution

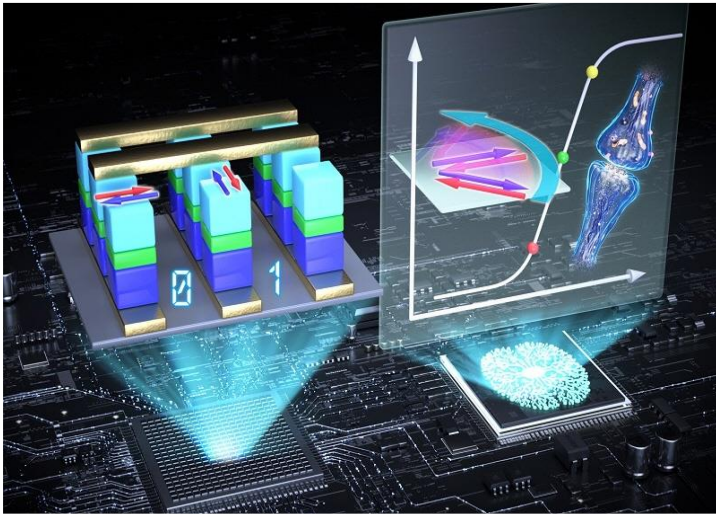


# Trends in AI Semiconductor

- Large Model with Low Power Consumption → On-Device AI
- Co-Optimization of SW, HW, and Domain Specific Application

## 1. Basic

- DRAM PIM and NVM PIM
- Neuromorphic & SNN



## 2. Domain Specific App.

- DRL, NeRF, Gen. AI NPU
- 6G, Metaverse, DigitalTwin



## 3. Large Model

- LLM (chatGPT) Acceleration
- LMM Optimization



# Contents

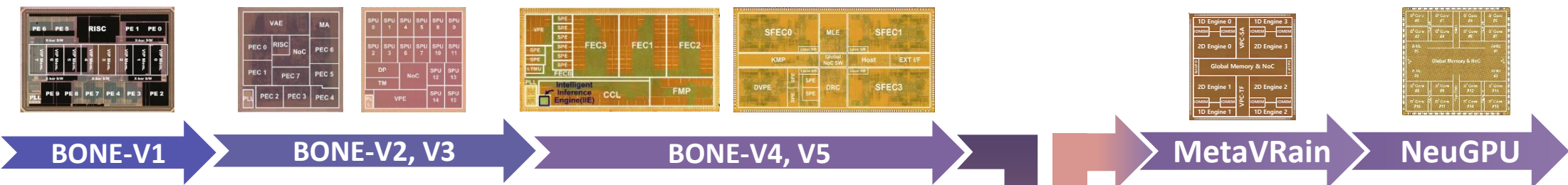
---

1. What is AI Semiconductor
- 2. Present AI Semiconductor**
3. Future of AI Semiconductor

# History of AI Semiconductor

## Recognition Processors with Attention

## 3D Intelligence SoC

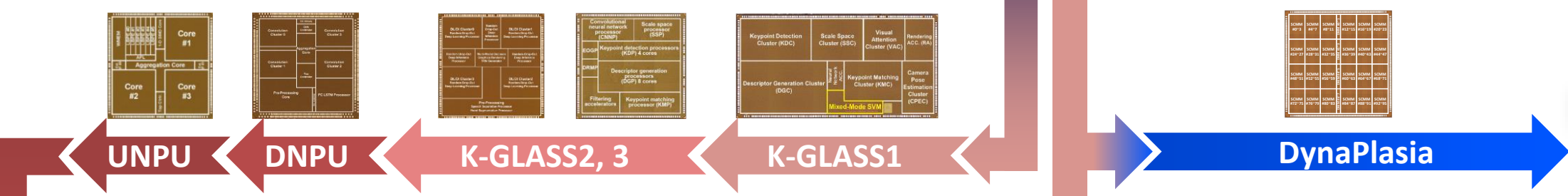


### DNN Inference Processors

### AR/VR Processors

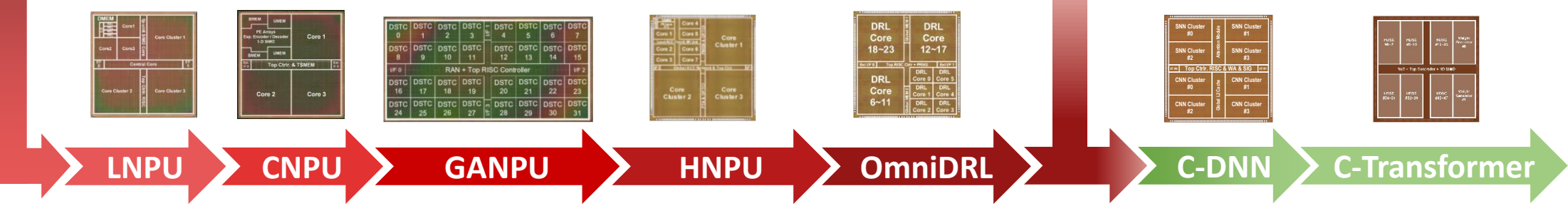
### Processing-In-Memory (PIM)

## Future Envision of AI NPU



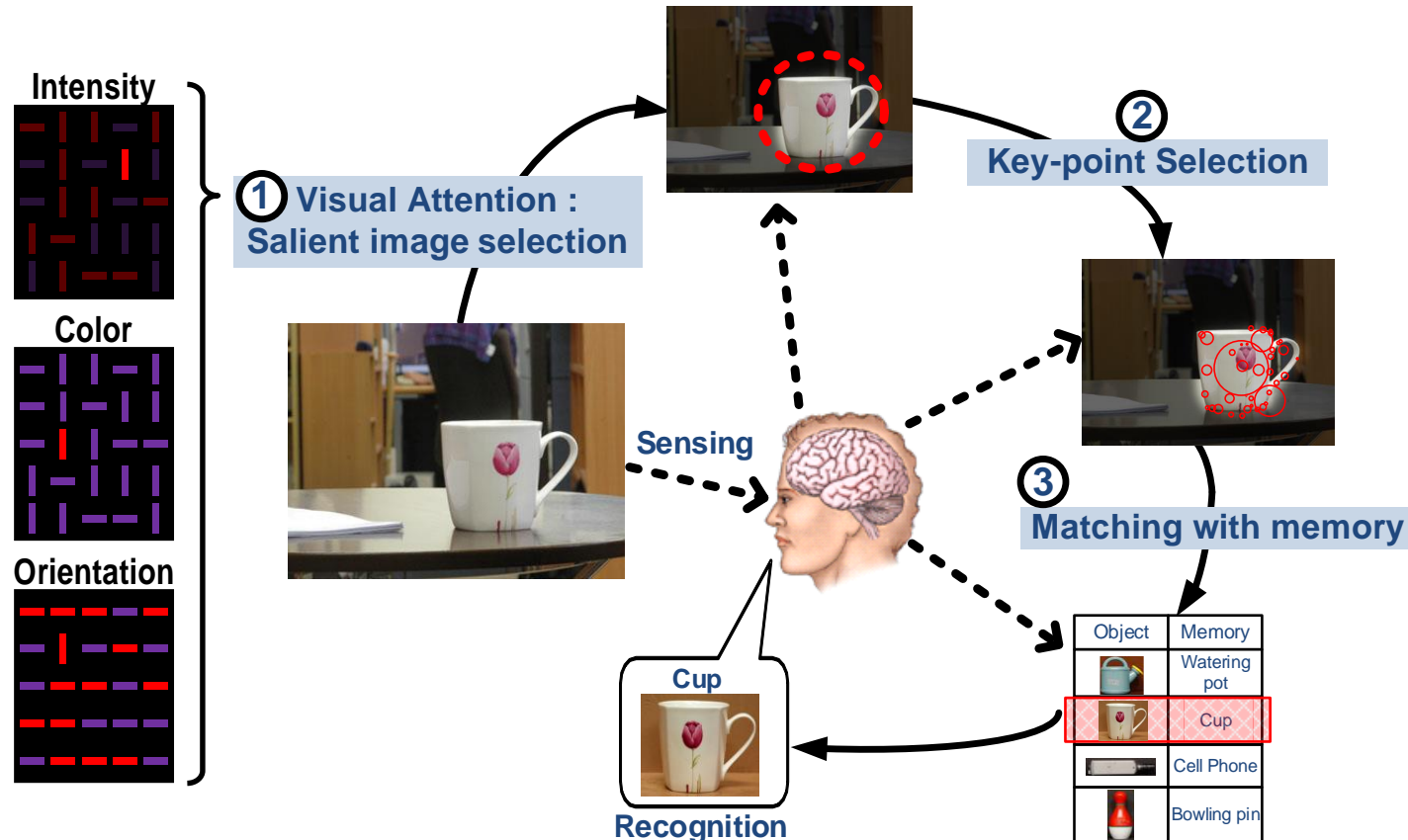
### DNN Training Processors

### Neuromorphic Processor



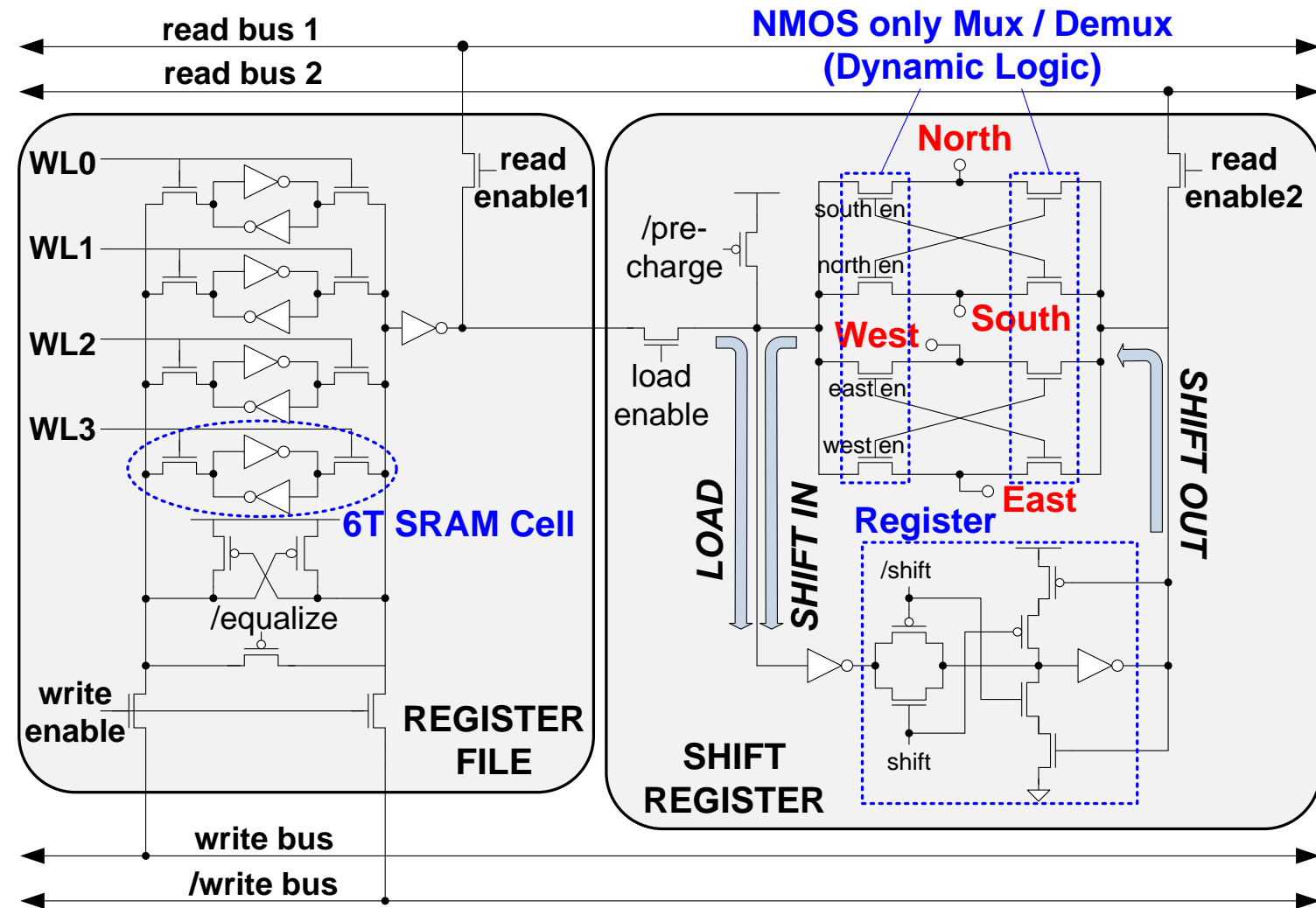
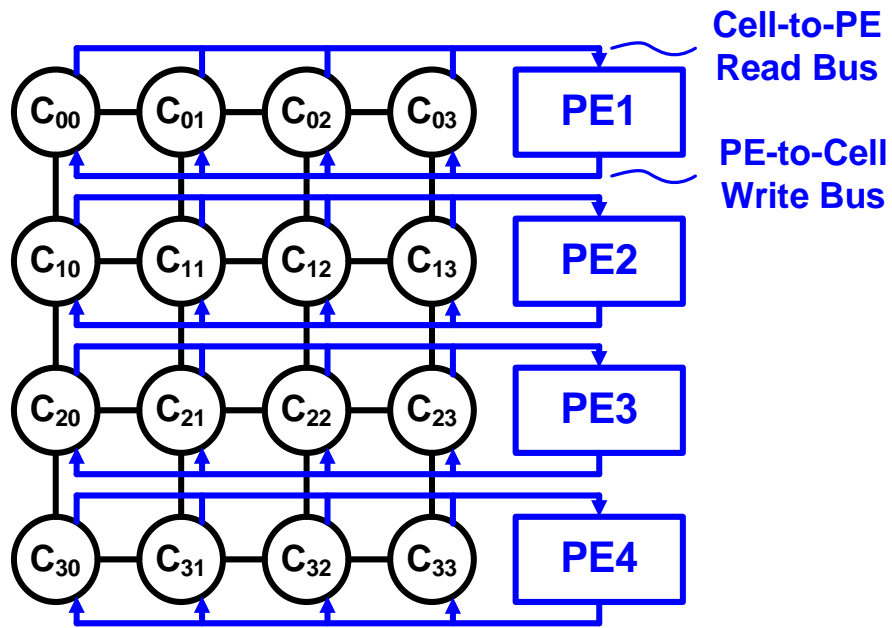
# 2007 BONE-V2 : Visual Attention

- Implementation of “**Visual Attention**” on silicon chip
  - 이미지 상의 중요 키폰트 강조
  - 키폰트 필터링을 통한 **Pixel-level visual attention**



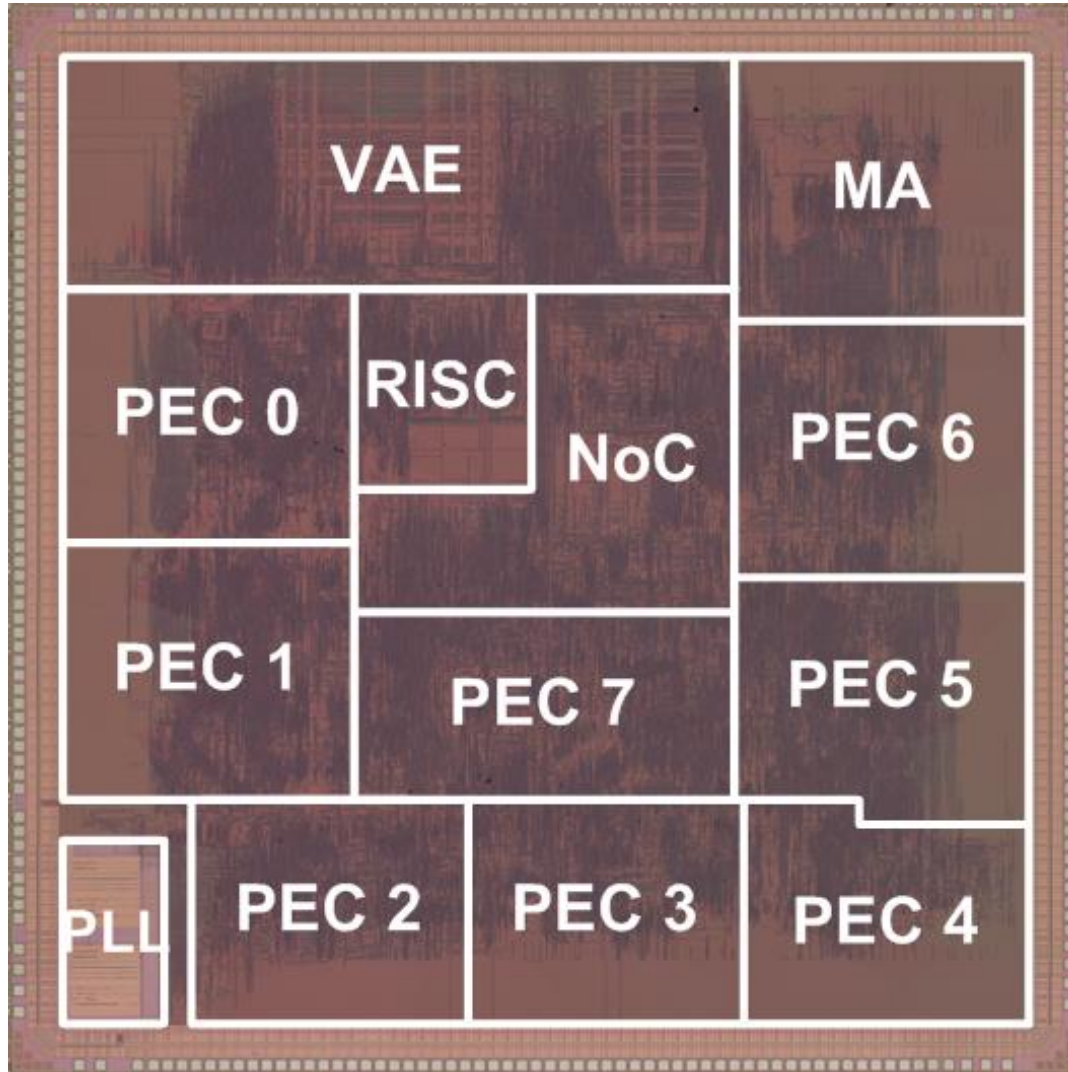
# 2007 World First CNN Accelerator

## 2D Cell + Shared PE Digital CNN





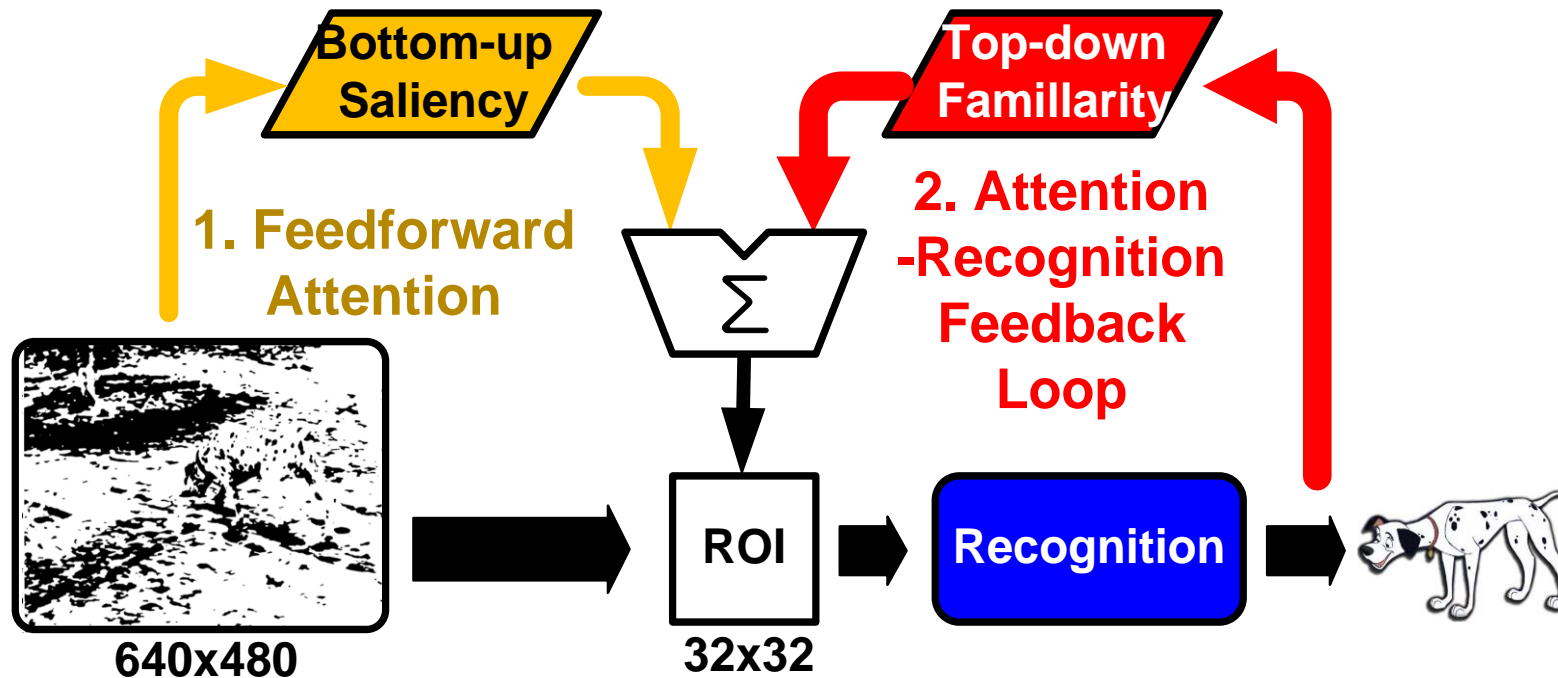
# BONE-V2 (2007 ISSCC)



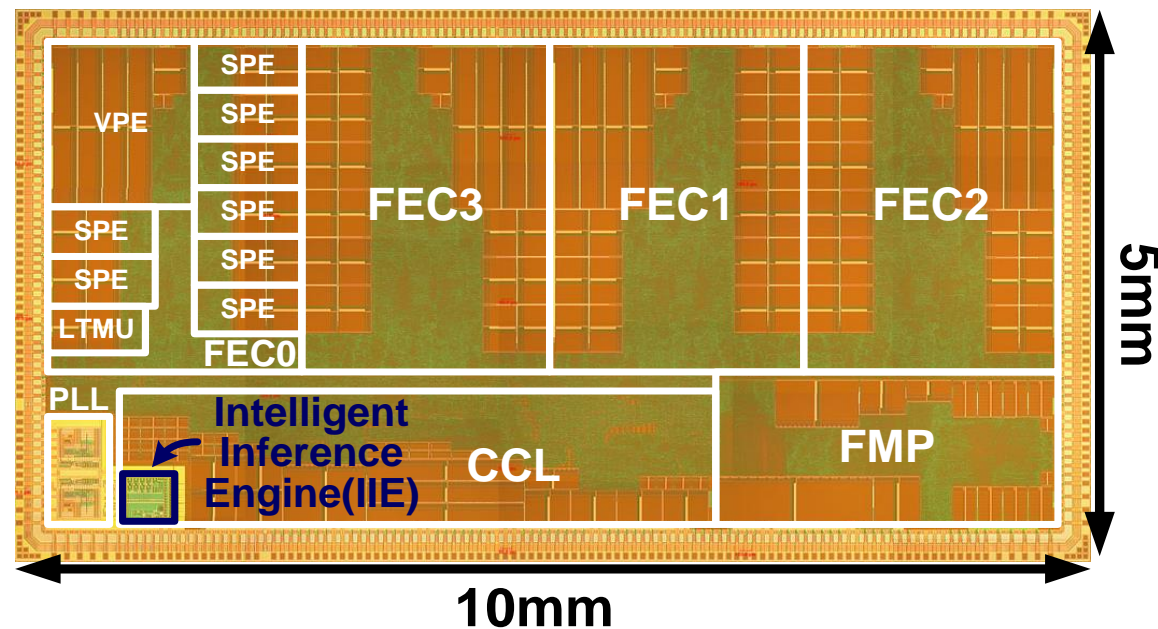
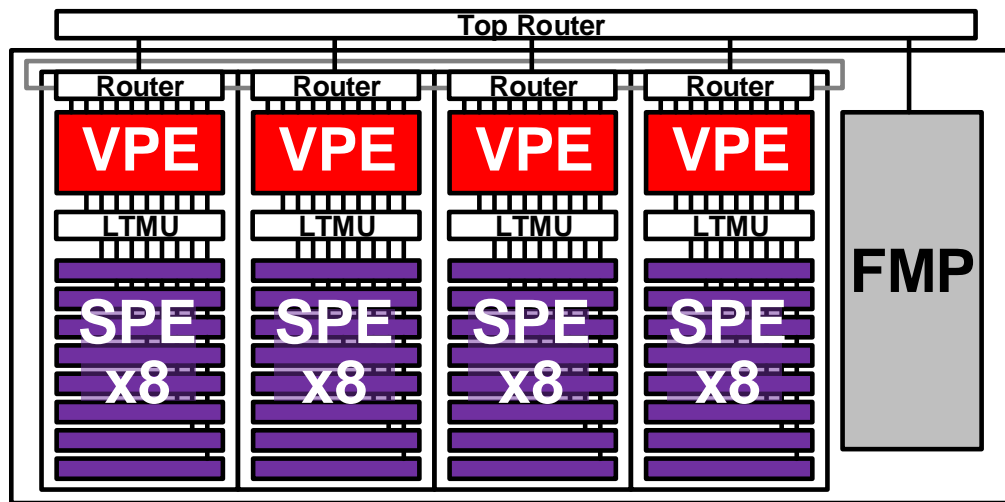
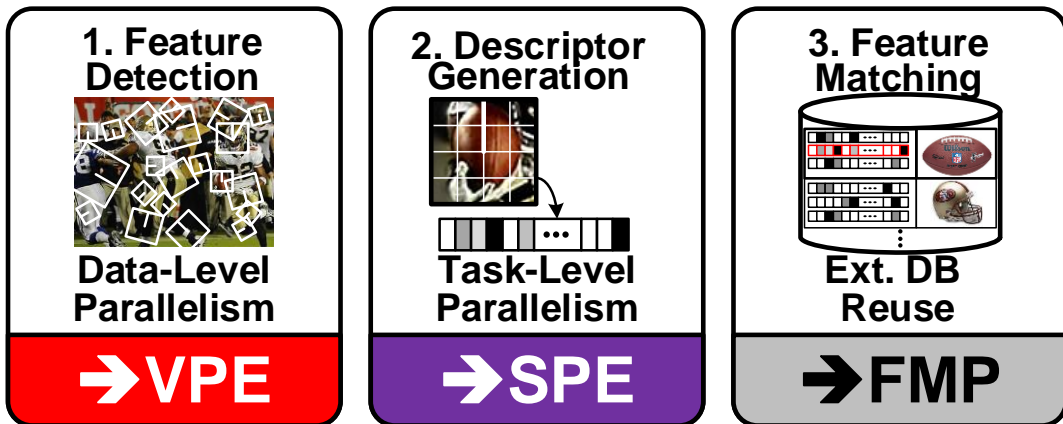
- 0.13 $\mu$ m 8M CMOS Tech.
- 6mm x 6mm
- Power Supply
  - 1.2V: Core
  - 2.5V: I/O
- Operating Frequency
  - 200MHz for IPs
  - 400MHz for NoC
- # of Transistors
  - 1.9M gates
  - 228kB SRAM
- Power Consumption
  - Less than 583mW (Object recognition)

# 2009 BONE-V4: Unified Attention Model\*

- Feedforward Attention
  - Bottom-up : Salient Image Features
- Attention-Recognition Feedback Loop
  - Top-down : Familiar Objects



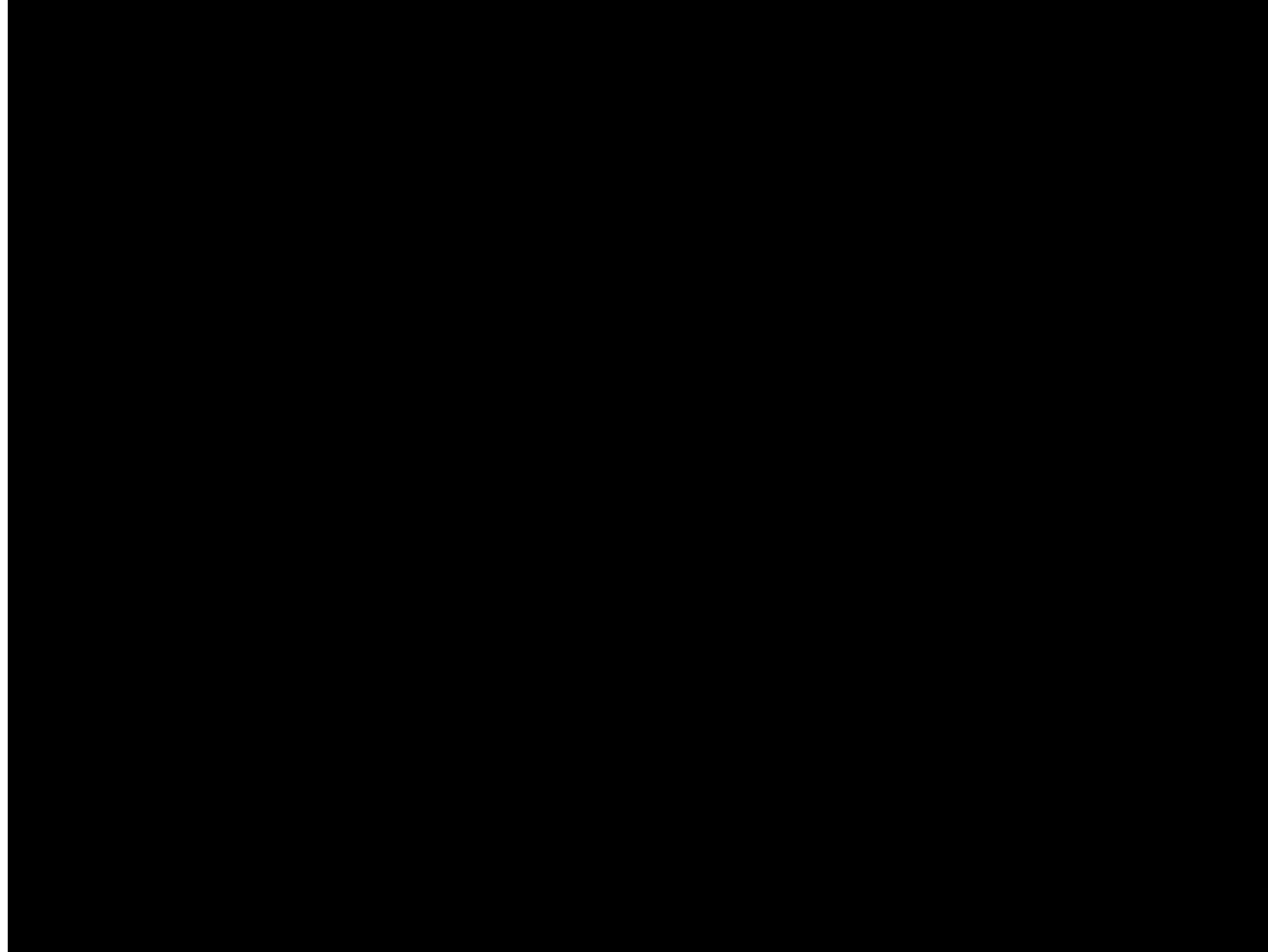
# BONE-V4



Technology	0.13um 1P8M Logic CMOS	
Die Size	50mm <sup>2</sup> 10.0mm x 5.0mm	
Gates / SRAM	2.92M Gates / 612 kB	
NoC IPs	51	
Power Supply	CCL & NoC	1.2 V
	PPL	0.65 ~ 1.2 V
Operating Frequency	Global NoC	400MHz (45FO4)
	CCL	200MHz (90FO4)
	PPL	50 ~ 200MHz (90FO4)

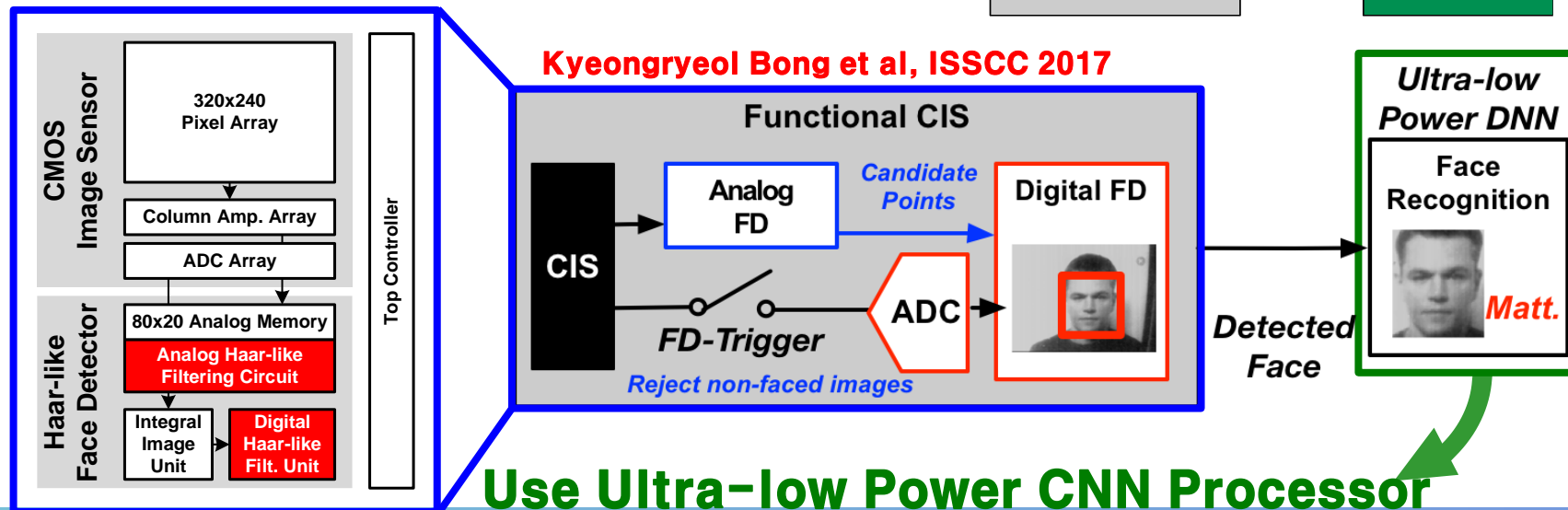
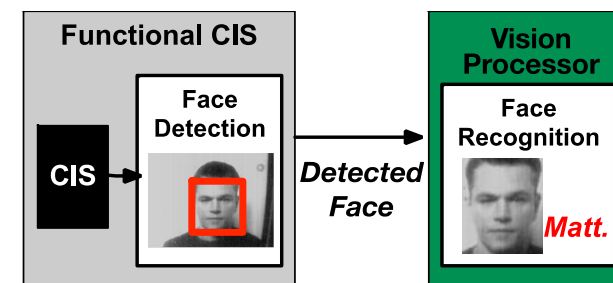
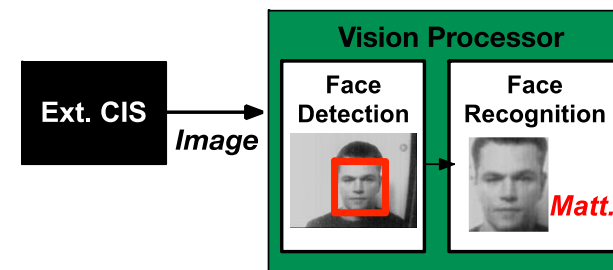
# 2009 BONE-V4: Demonstration

---



# 2017 Low Power Face Recognition SoC

- **Always-On → Ultra Low Power**
  - 0.6mW Full CNN Operation
- **Hybrid Face Detector**
  - Face detection by CMOS image sensor
  - Combine analog & digital face detector



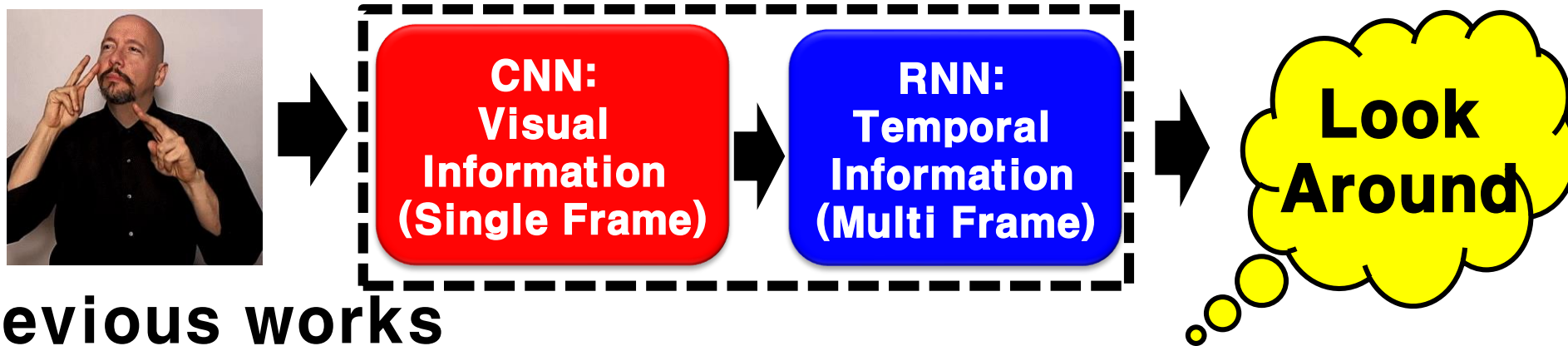
# 2017 Face Recognition Demo Video



# 2017 CNN + RNN Deep Neural Network

Dongjoo Shin et al,  
ISSCC 2017

- **CNN: Static Picture Recognition**
  - Face recognition, image classification...
- **RNN: Temporal Video Recognition**
  - Translation, speech recognition...
- **CNN + RNN: CNN-extracted features → RNN input**



- **Previous works**

- Optimized for convolution layer only: [6], [3]
- Optimized for FC layer and RNN only: [5]

[3] B. Moons, SOVC 2016  
[5] S. Han, ISCA 2016  
[6] Y. Chen, ISSCC 2016

# 2017 DNPU : Pet Robot Demonstration



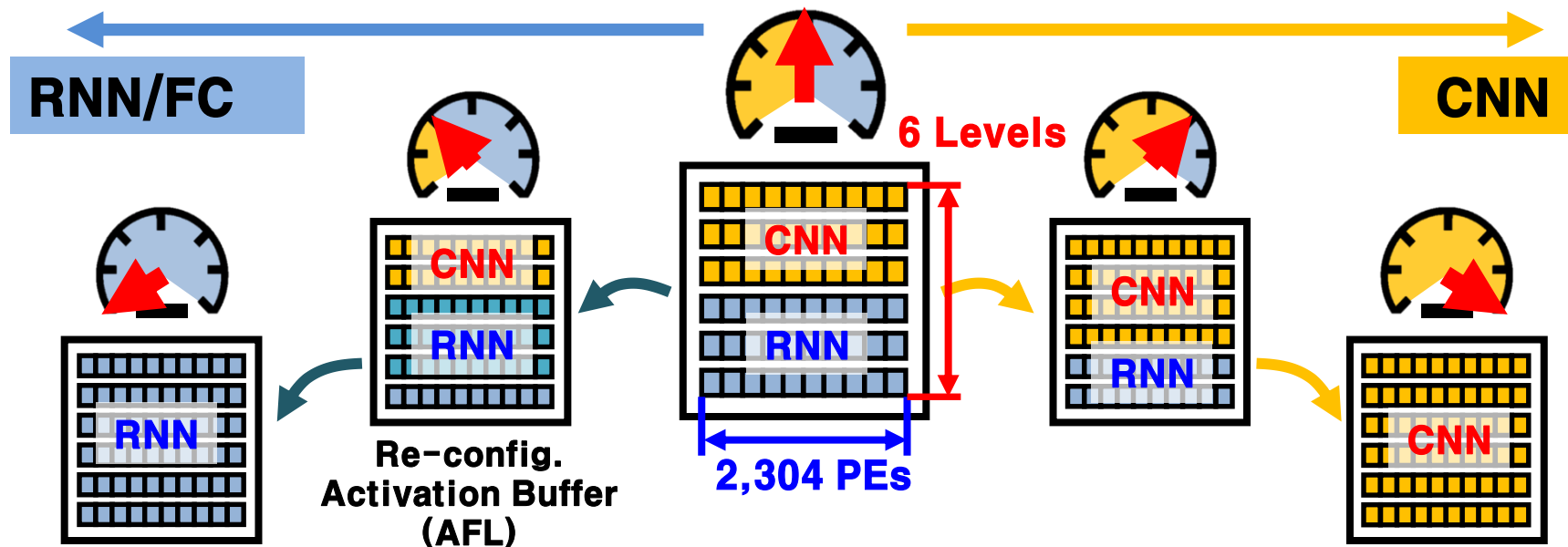
D. Shin, et al. "14.2 DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks." *ISSCC 2017*



# 2018 Unified NPU: Programmable DNN Arch.

- **Unified Data Path**
  - Dynamically Programmable for CNN, RNN/FC

- **Support Various CNN & RNN Workload**

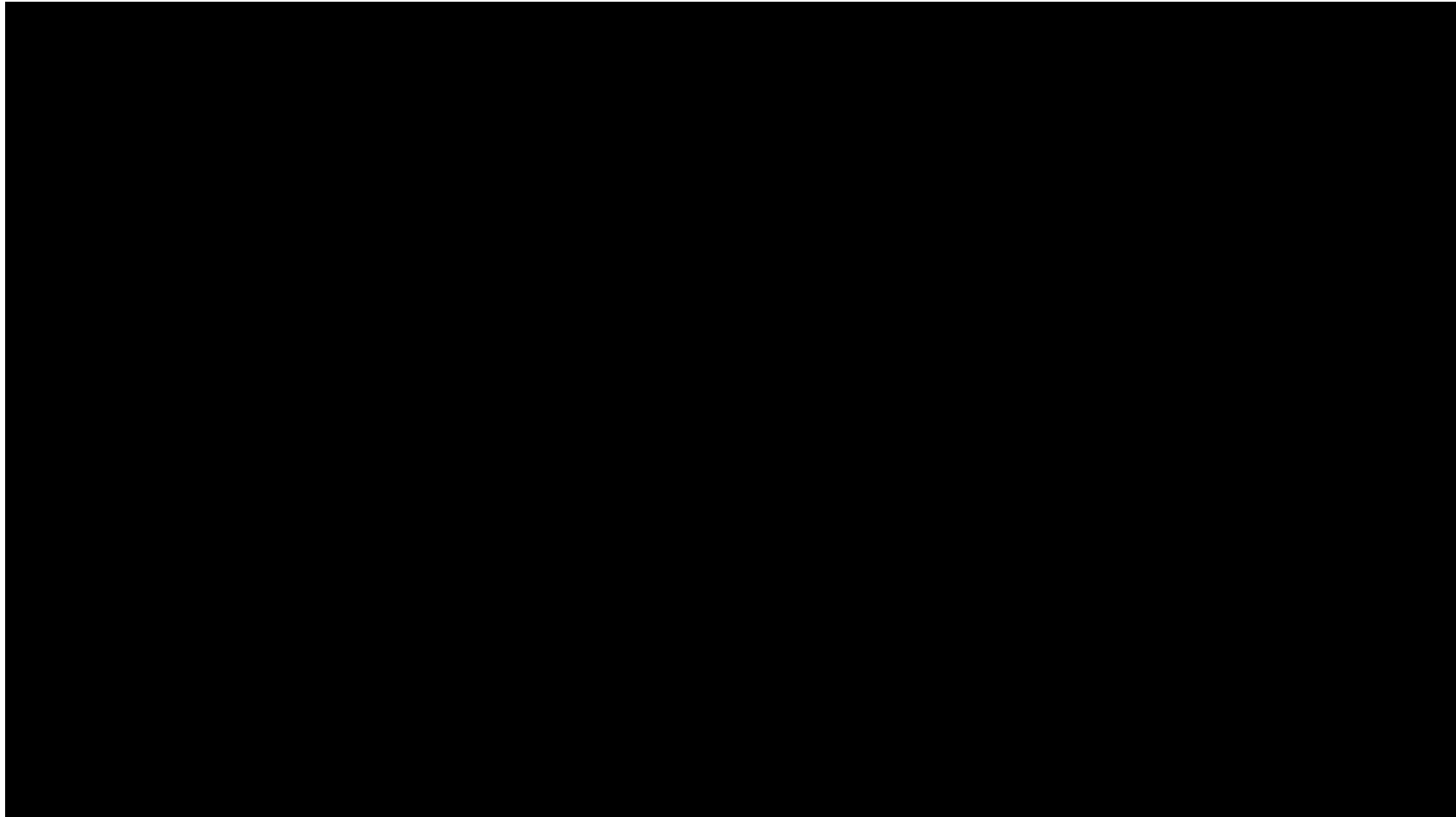


Lee, Jinmook, et al.

"UNPU: A 50.6 tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision." *ISSCC 2018*

# 2018 UNPU : Emotion Recognition

---

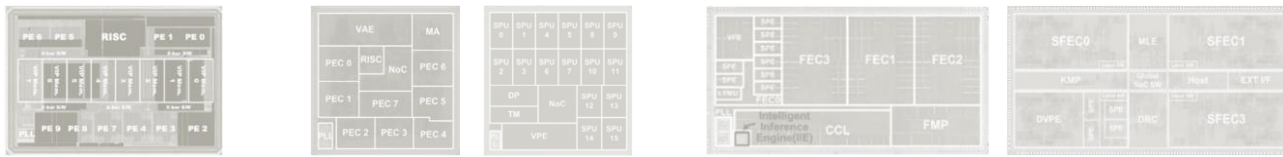


Lee, Jinmook, et al.

"UNPU: A 50.6 tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision." *ISSCC 2018*

# History of AI Semiconductor

## Recognition Processors with Attention



BONE-V1

BONE-V2, V3

BONE-V4, V5

MetaVRain

NeuGPU

## DNN Inference Processors

## AR/VR Processors

## Processing-In-Memory (PIM)



UNPU

DNPU

K-GLASS2, 3

K-GLASS1

## DNN Training Processors

DynaPlasia

## Neuromorphic Processor



LNPU

CNPU

GANPU

HNPU





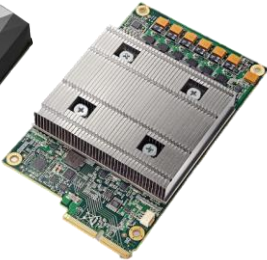

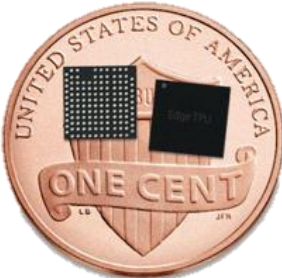
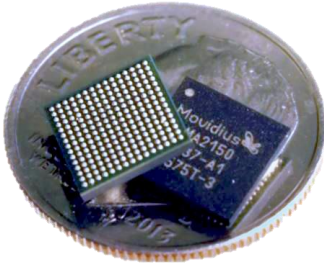

OmniDRL

C-DNN

C-Transformer

Future  
Envision  
of  
AI NPU

# Inference & Training

	Cloud / Data Center	Edge / Embedded
Training	 P100  TPU 2.0	
Inference	 GTX 1080 Ti  TPU  Arria 10 FPGA	 Edge TPU  Movidius  Hailo-8™

# Inference & Training

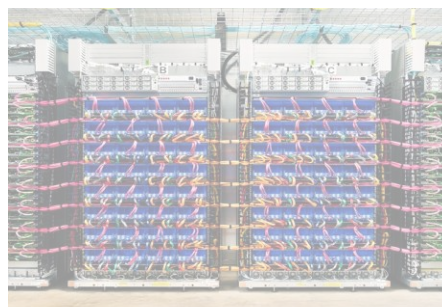
## Cloud / Data Center

## Edge / Embedded

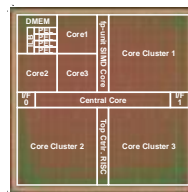
Training



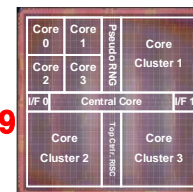
P100



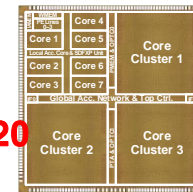
TPU 2.0



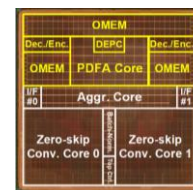
**LNPU**  
ISSCC' 19



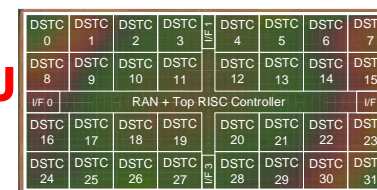
**PNPU**  
S. VLSI' 20



**HNPu**  
JSSC' 21

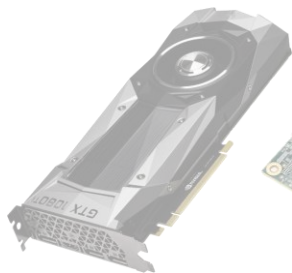


**DF-LNPU**  
S. VLSI' 19

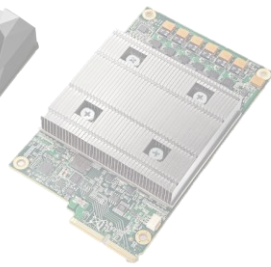


**GANPU**  
ISSCC' 19

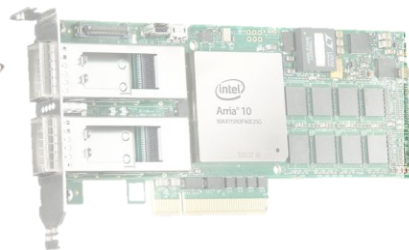
Inference



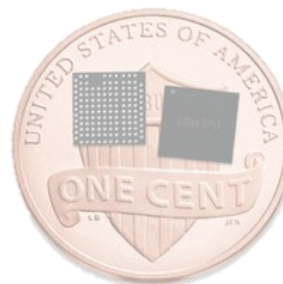
GTX 1080 Ti



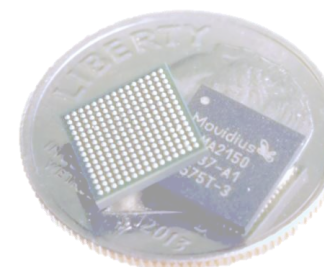
TPU



Arria 10 FPGA



Edge TPU



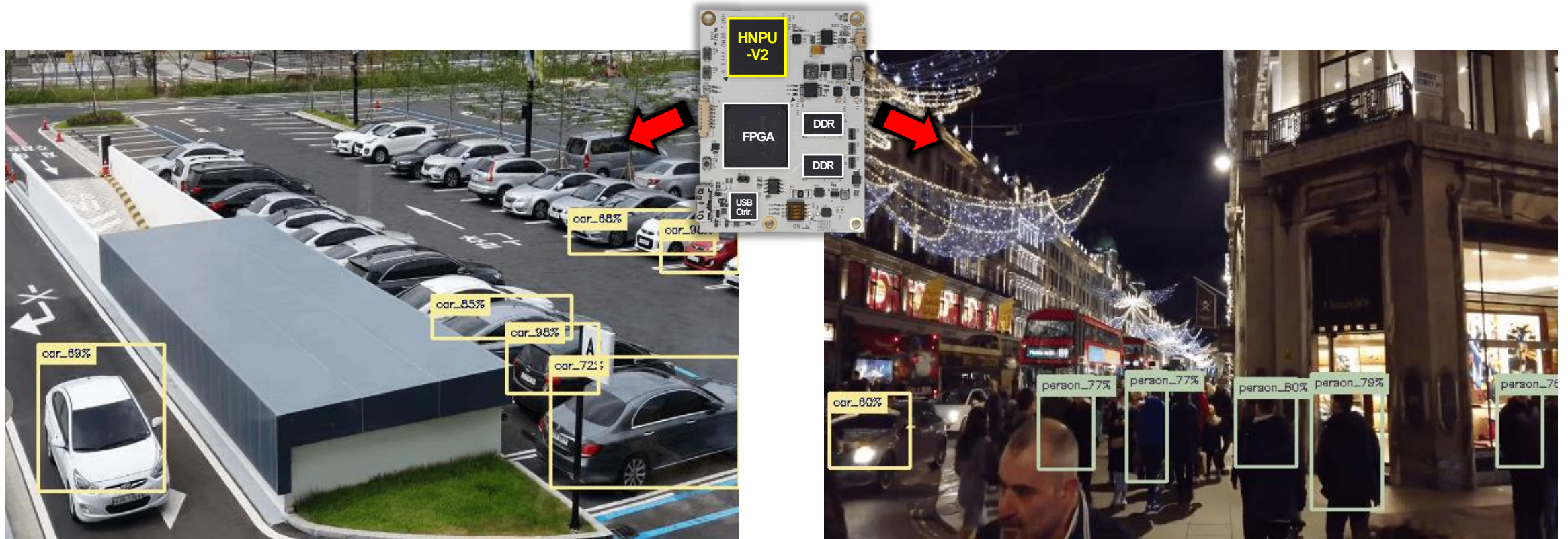
Movidius



Hailo-8™

# Robust Object Detection w/ DNN Training

- HNPU-V2: 정확도 보상을 위한 Online DNN Tuning
- 예상치 못한 상황에서 **자동으로 정확도 회복 가능**



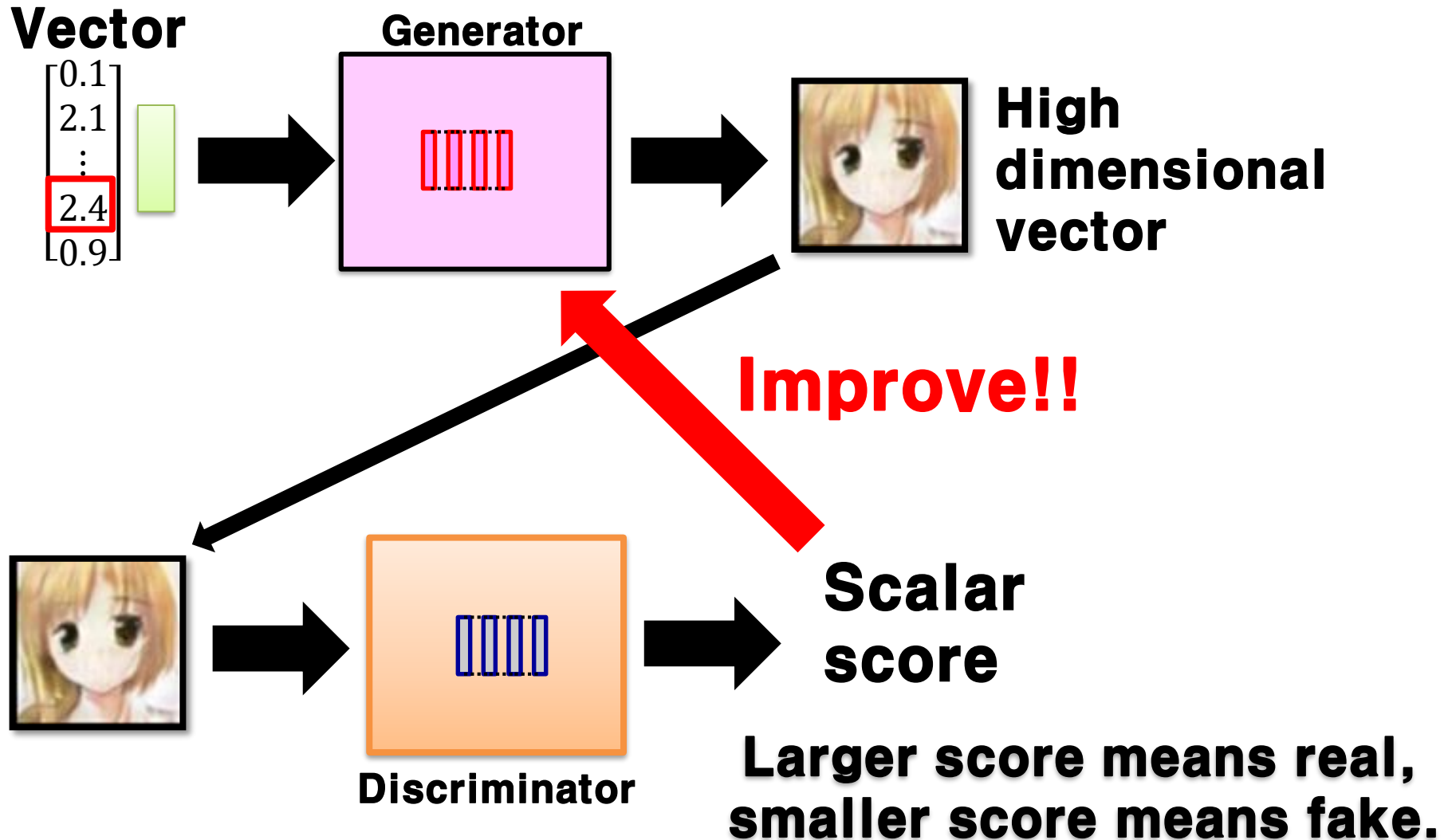
# 2021 HNPU-V2 Demonstration Video

---

# HNPU-V2

Mobile On-chip Training  
System Demonstration

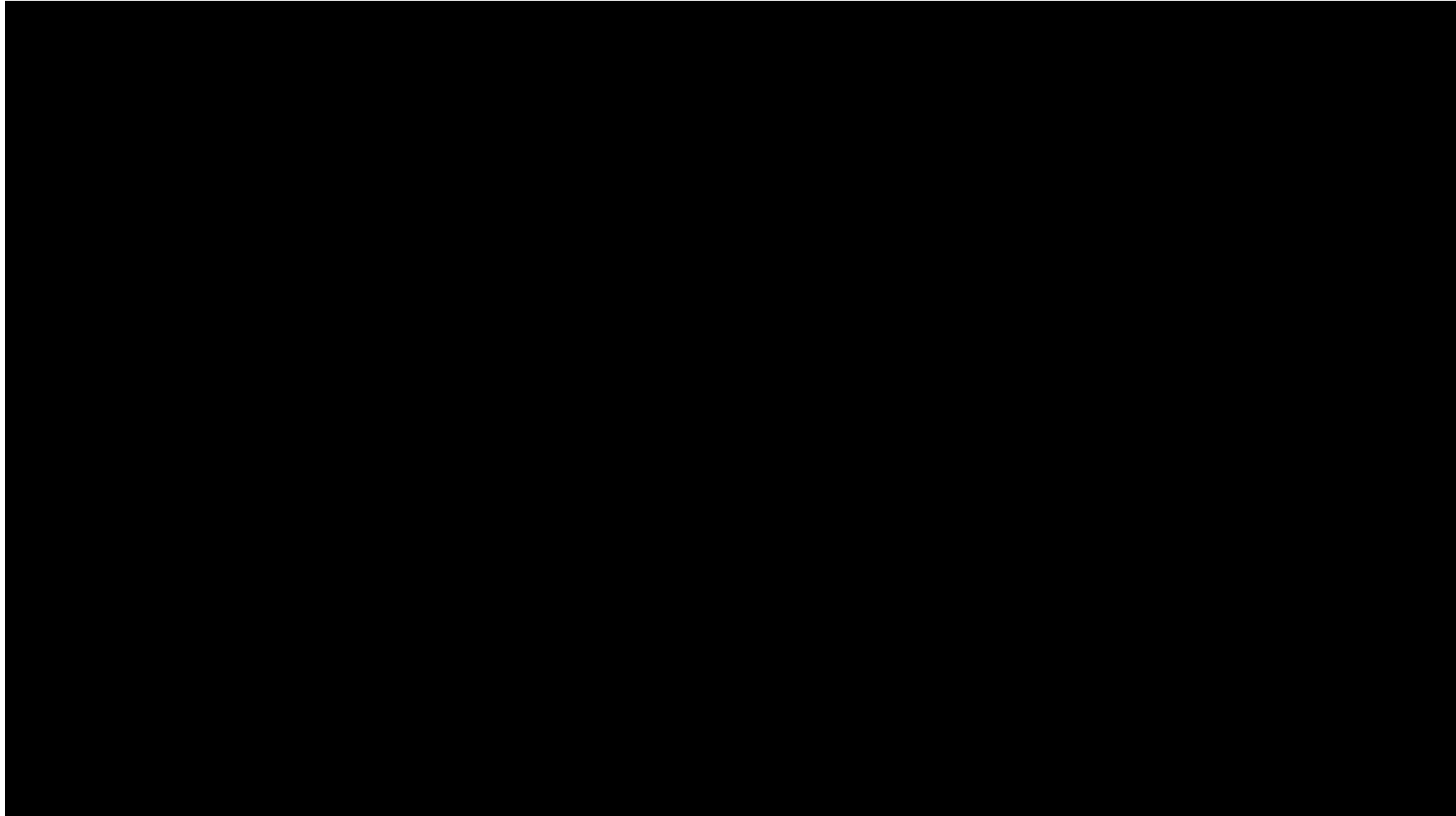
# Generative Adversarial Network



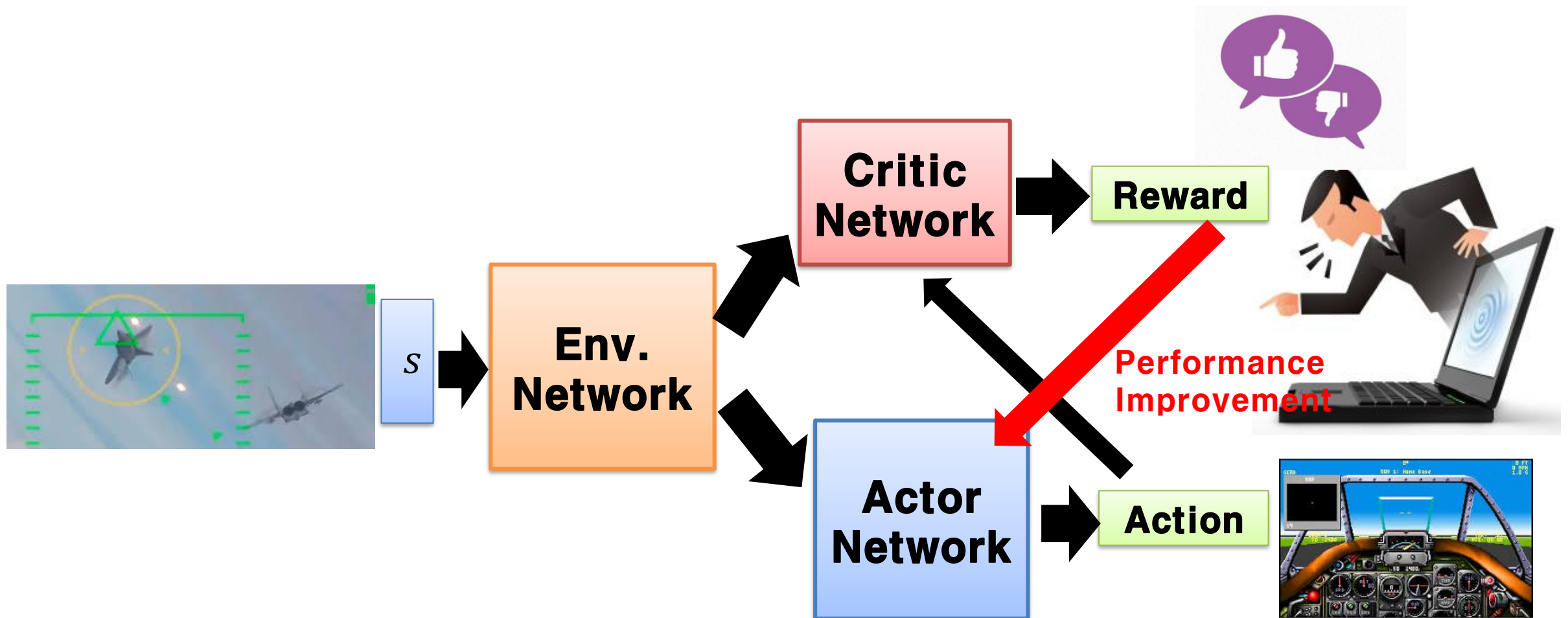


# 2020 GANPU

---

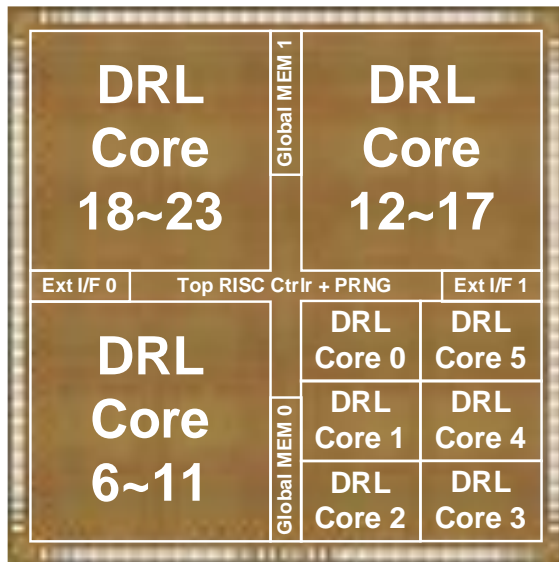


# Deep Reinforcement Learning

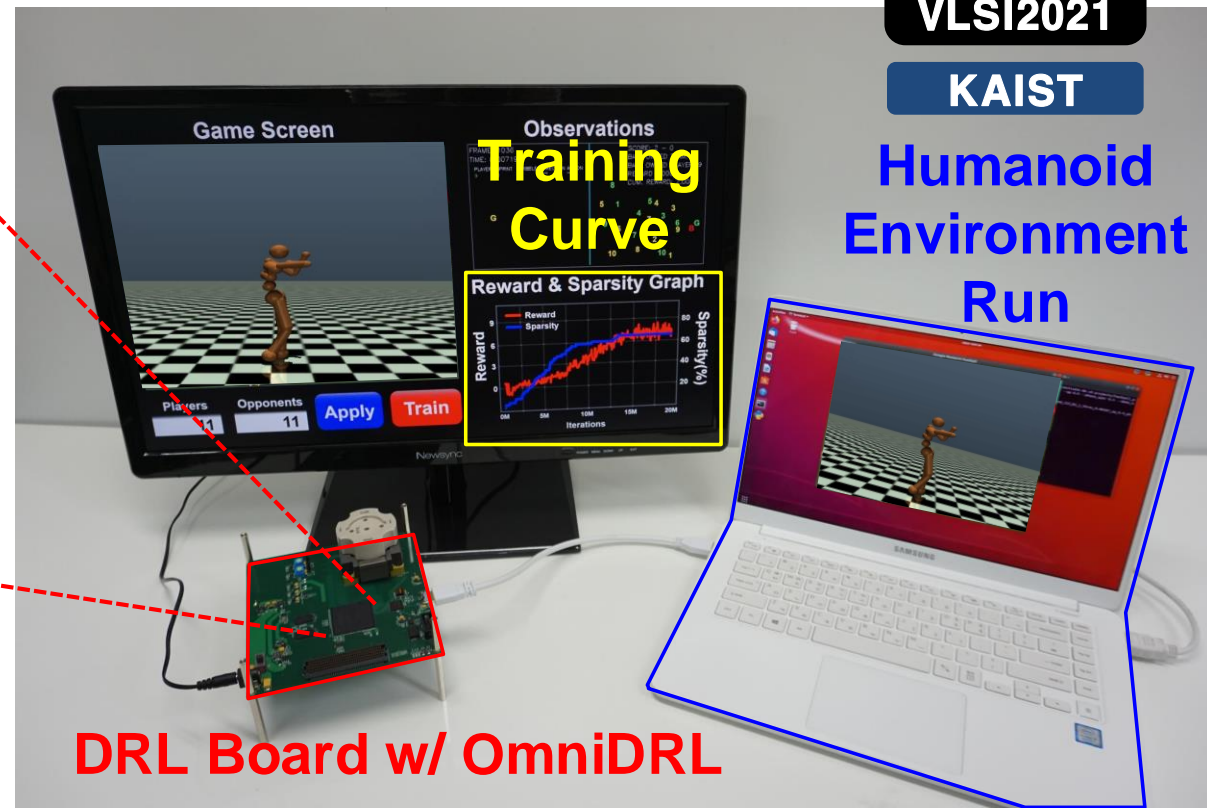


# OmniDRL : Advanced DRL Processor

- Humanoid Robot Agent Training w/ DRL processor



**4.18TFLOPS,  
29.3 TFLOSP/W  
Advanced DRL Processor**



J. Lee, et al. "OmniDRL: A 29.3 TFLOPS/W Deep Reinforcement Learning Processor with Dual-mode Weight Compression and On-chip Sparse Weight Transposer," , VLSI 2021

# 2021 OmniDRL : Demonstration Video

---

## Omni-DRL

### Mobile DRL Training System Demonstration

J. Lee, et al. "OmniDRL: A 29.3 TFLOPS/W Deep Reinforcement Learning Processor with Dual-mode Weight Compression and On-chip Sparse Weight Transposer," , VLSI 2021

# Contents

---

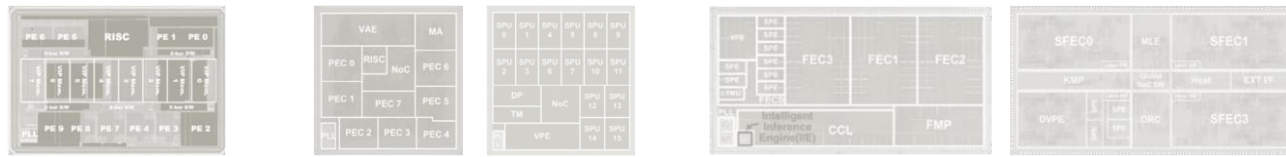
1. What is AI Semiconductor

2. Present AI Semiconductor

**3. Future of AI Semiconductor**

# History of AI Semiconductor

## Recognition Processors with Attention



BONE-V1

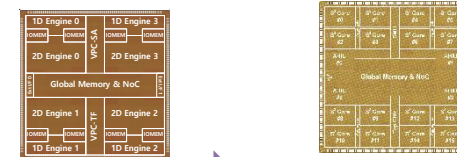
BONE-V2, V3

BONE-V4, V5

DNN Inference Processors

ARVR Processors

## Spatial Computing SoC



MetaVRain

NeuGPU

Processing-In-Memory (PIM)



UNPU

DNPU

K-GLASS2, 3

K-GLASS1

DNN Training Processors



LNPU

CNPU

GANPU

HNPU

OmniDRL



C-DNN

C-Transformer

DynaPlasia

Neuromorphic Processor

Future  
Envision  
of  
AI NPU

# Spatial Computing

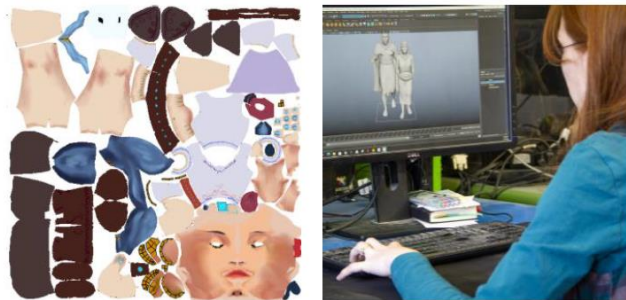
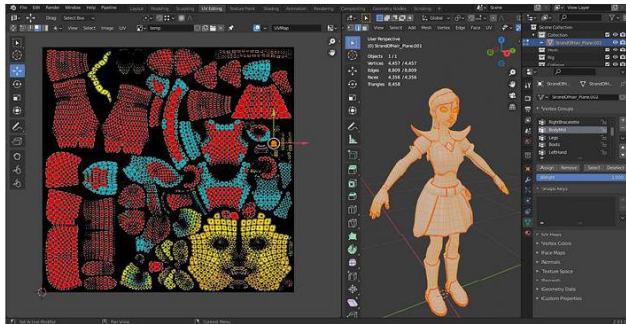
- ❑ **Human + CPS (Cyber Physical System)**
- ❑ The **digitization** of **activities** of **machines**, **people**, **objects**, and the **environments** in which they **enable and optimize actions and interactions**.



# Conventional 3D Modelling

## Manual Design w/ 3D Graphics Tool

- Expert-only 😞
- 70-110h to design 😞



## Specialized 3D Scanning Studio

- High-cost equipment 😞  
(~150 DSLR Cameras)



## Photogrammetry w/ Mobile Camera

- Requires feature extraction
- Fail for featureless surface



Distorted Surface 😞

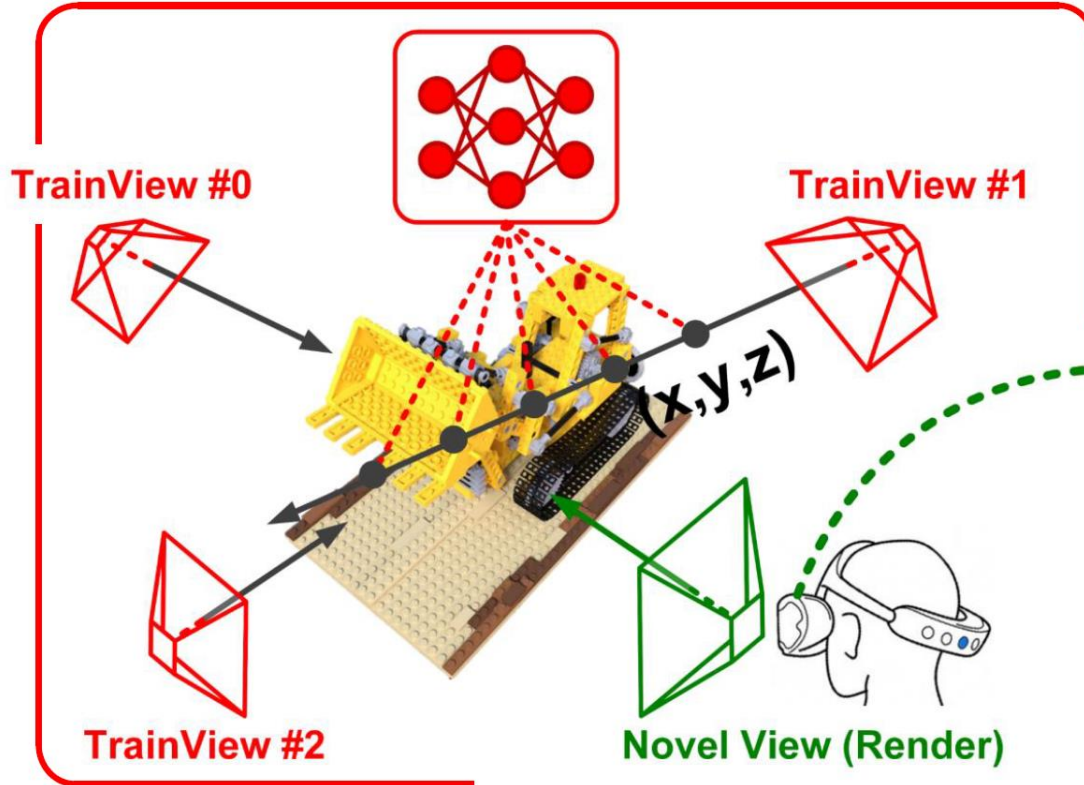


# NeRF 3D Modelling

## Training 2D Images



## 3D Modeling (Training)



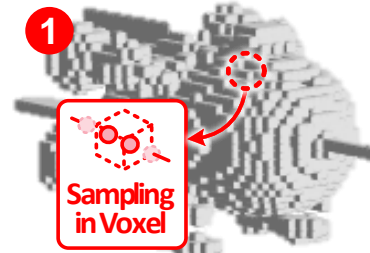
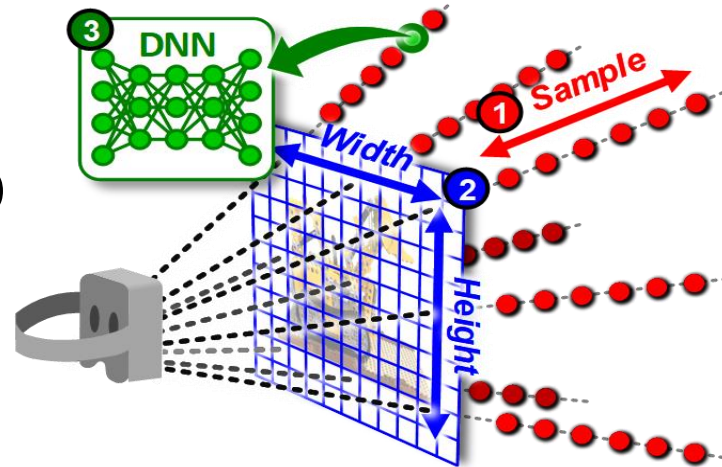
## 3D Rendering (Inference)



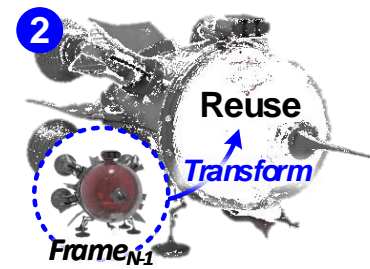
# 2023 MetaVRain: 3D NeRF Processor

- Mobile AR/VR 기기를 위한 AI 기반 Real-time Rendering

NeRF:  
DNN Based 3D  
Rendering  
Algorithm



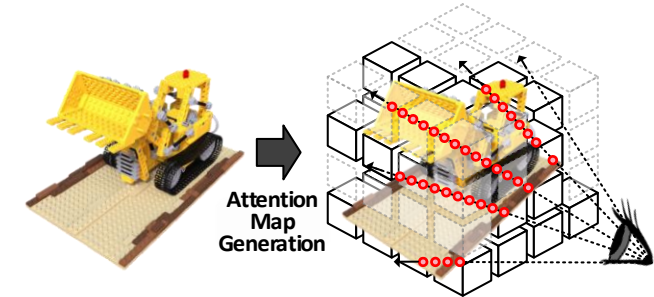
Spatial Attention(SA)



Temporal Familiarity (TF)



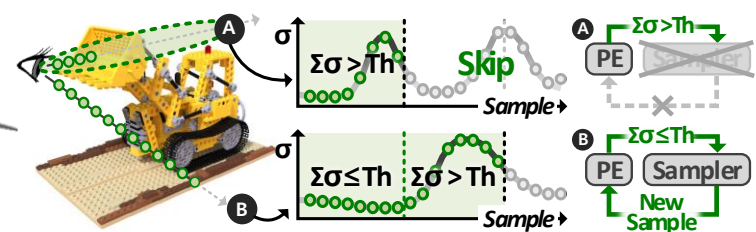
Top-down Attention (TDA)



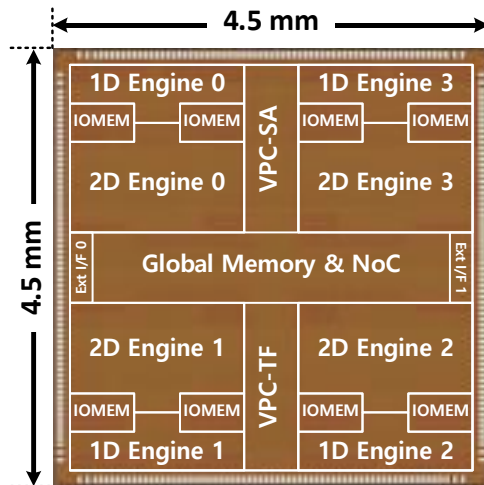
Need New RGB-D      Reuse RGB-D from Previous frame



3D Transform



Chip  
Photo



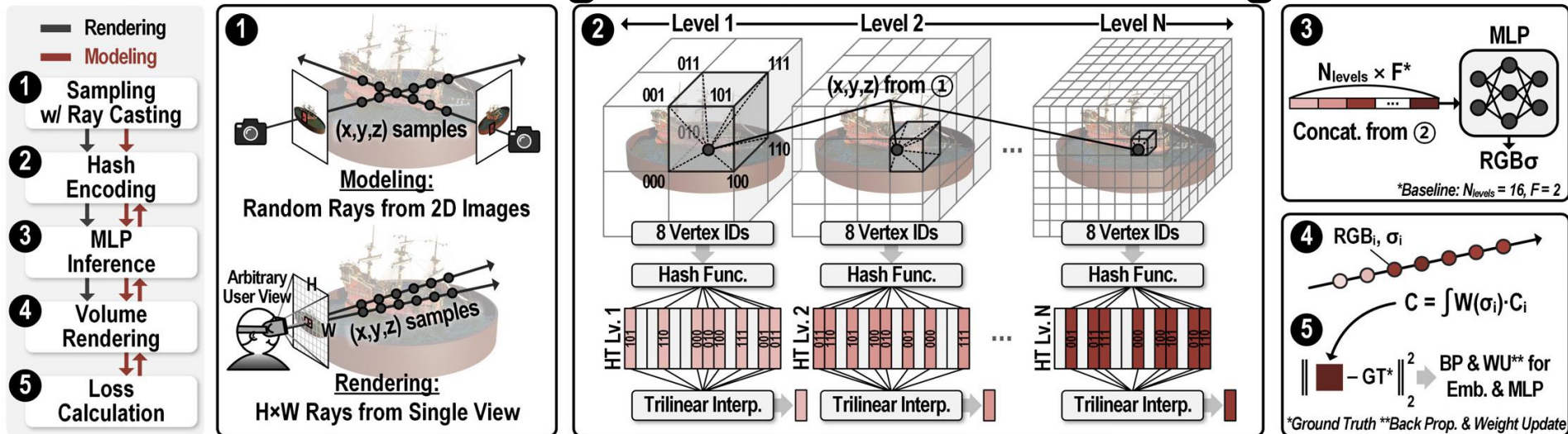
# 2023 MetaVRain : YTN



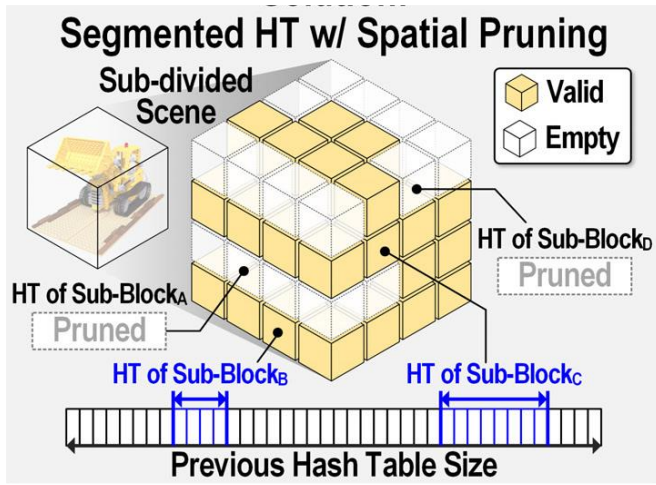
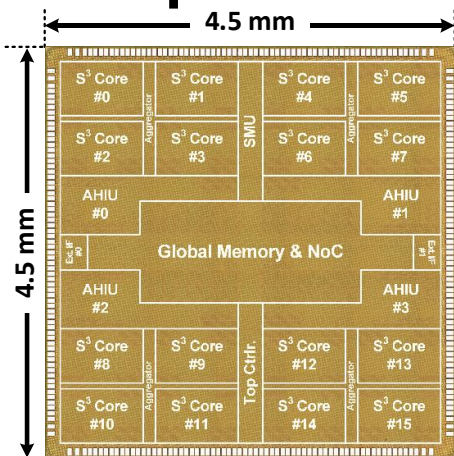
# 2024 NeuGPU: 3D NeRF Processor

## • NeRF-based Instant Modeling & Real-time Rendering Processor

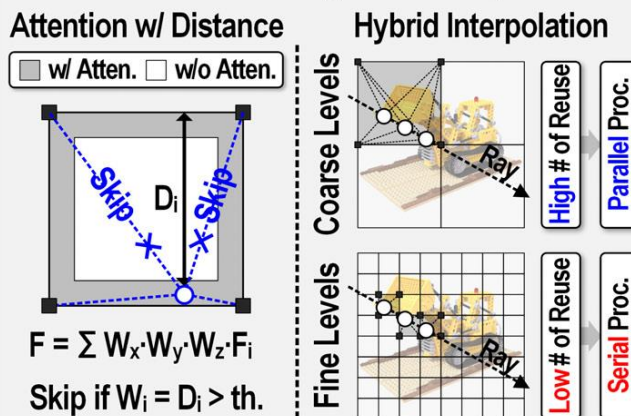
**NeRF :**  
Hash Encoding & MLP  
based  
3D Modeling/Rendering  
Algorithm



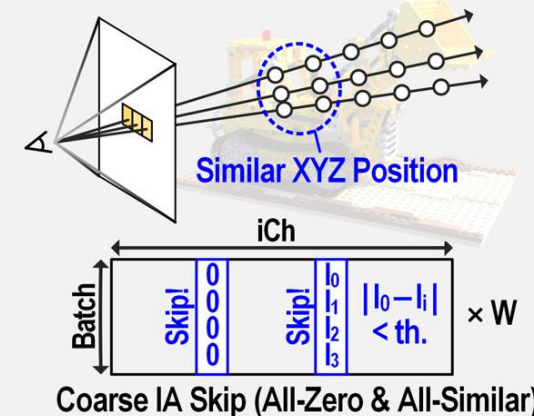
### Chip Photo



### Attention-based Hybrid Interpolation

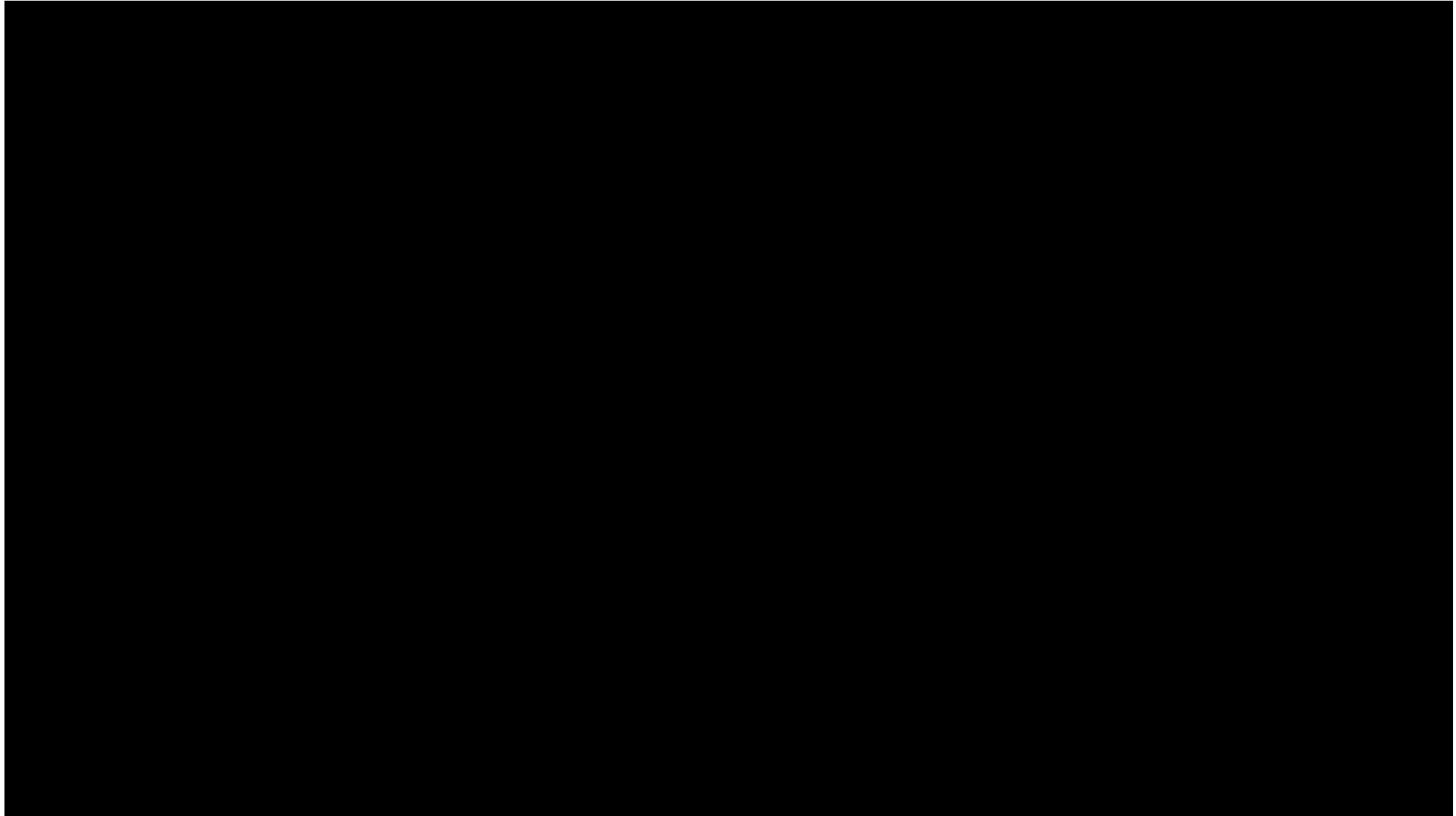


### Sparsity-Similarity Skipping



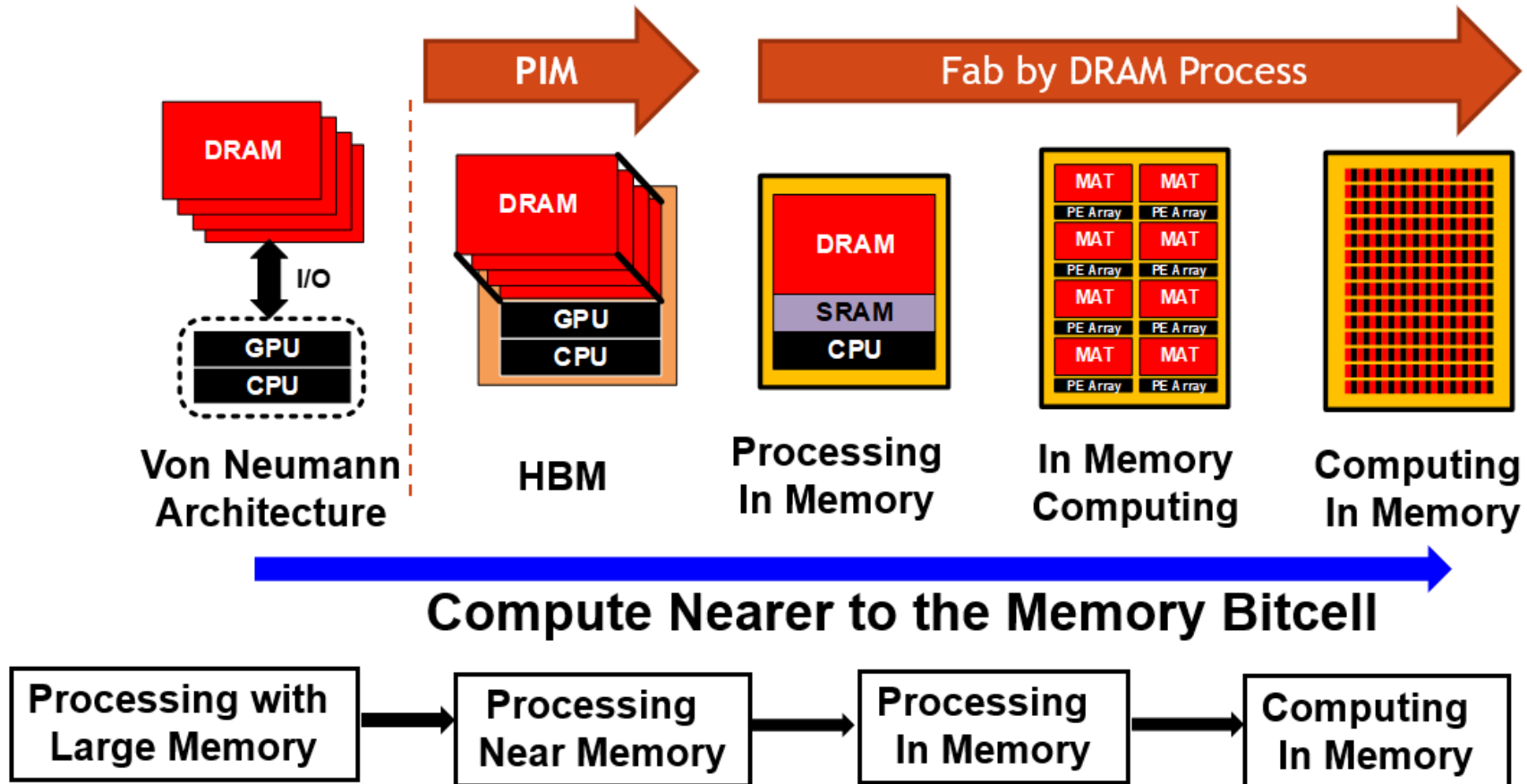
# 2024 NeuGPU: Demonstration Video

---



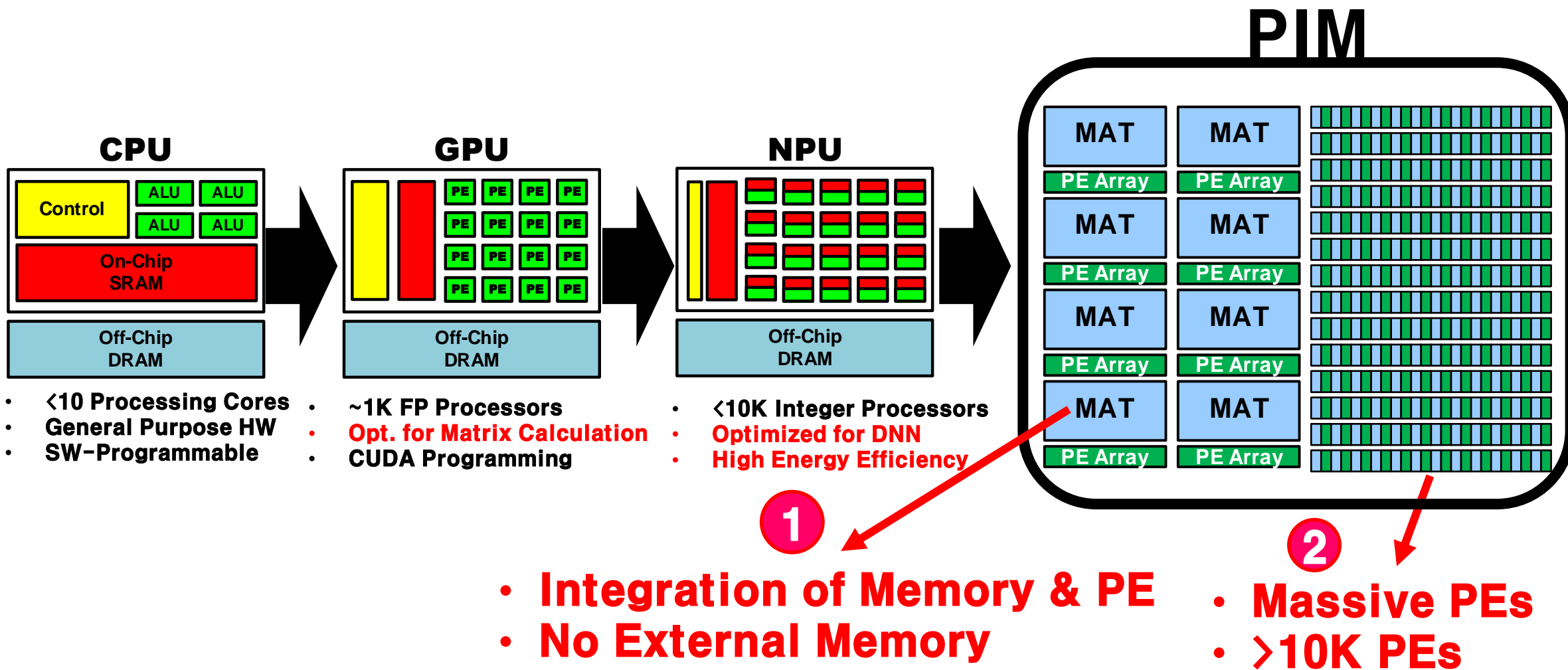
# Evolution of PIM Architecture

- 메모리와 연산기의 융합성 증가
  - Near Memory Processing → Processing in Memory

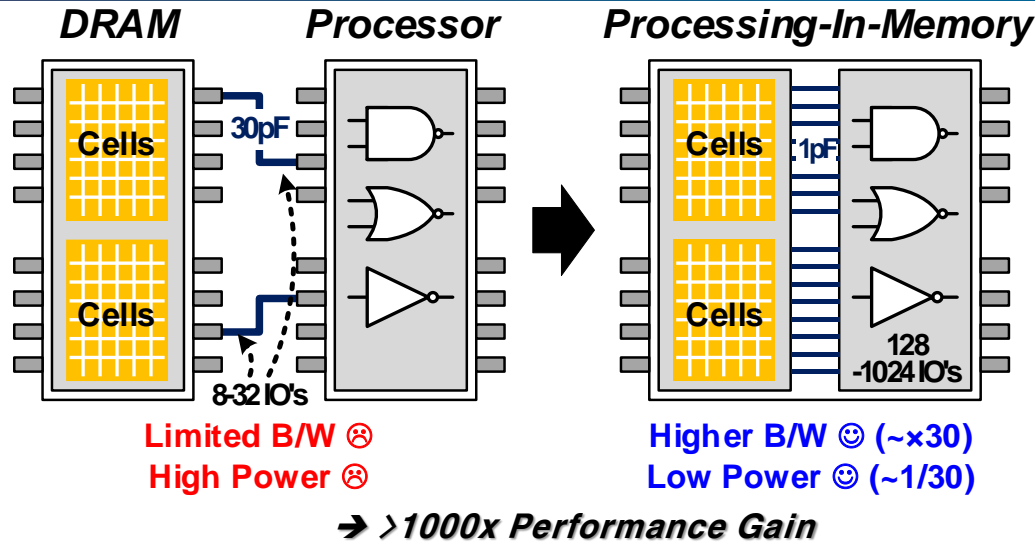


# Evolution of DNN Processor

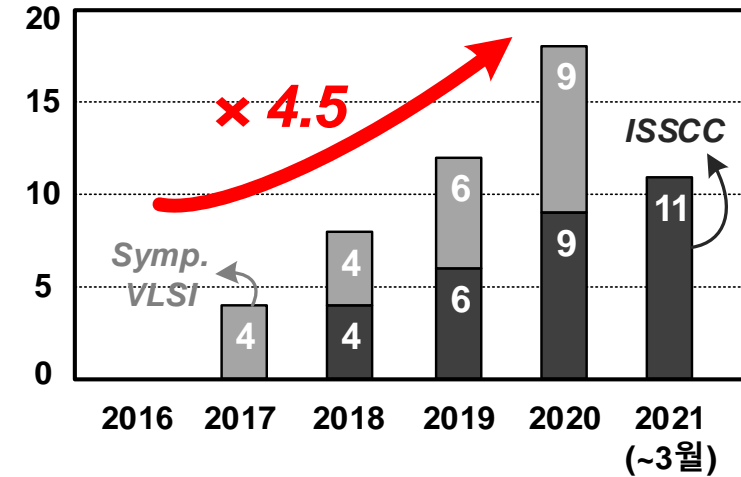
- Evolved to Memory Centric Computing



# Advantages of PIM

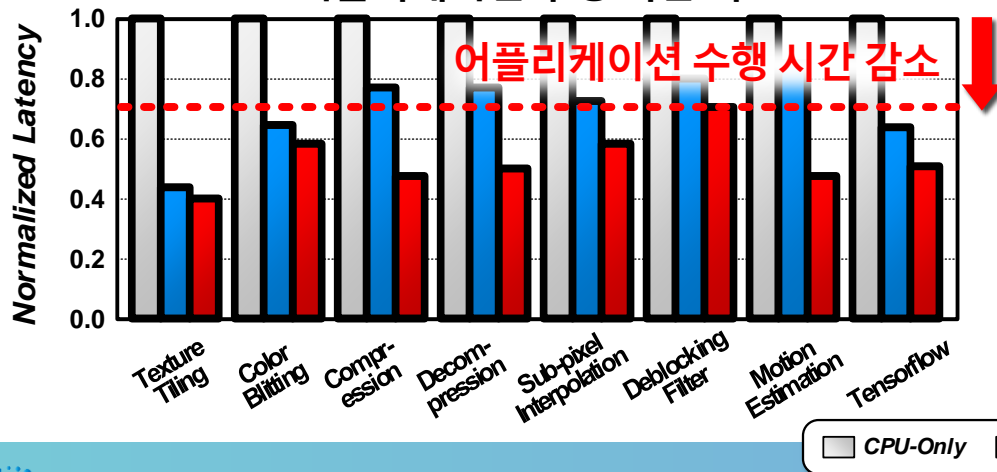


< PIM 관련 논문 통계 >

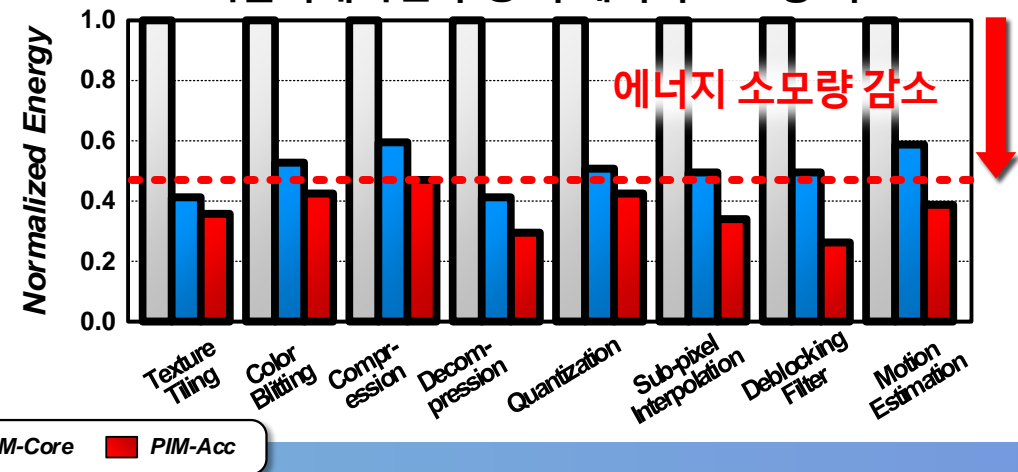


< PIM을 통한 폰-노이만 아키텍처의 한계 극복 >

< 어플리케이션 수행 시간 비교 >



< 어플리케이션 수행 시 에너지 소모량 비교 >



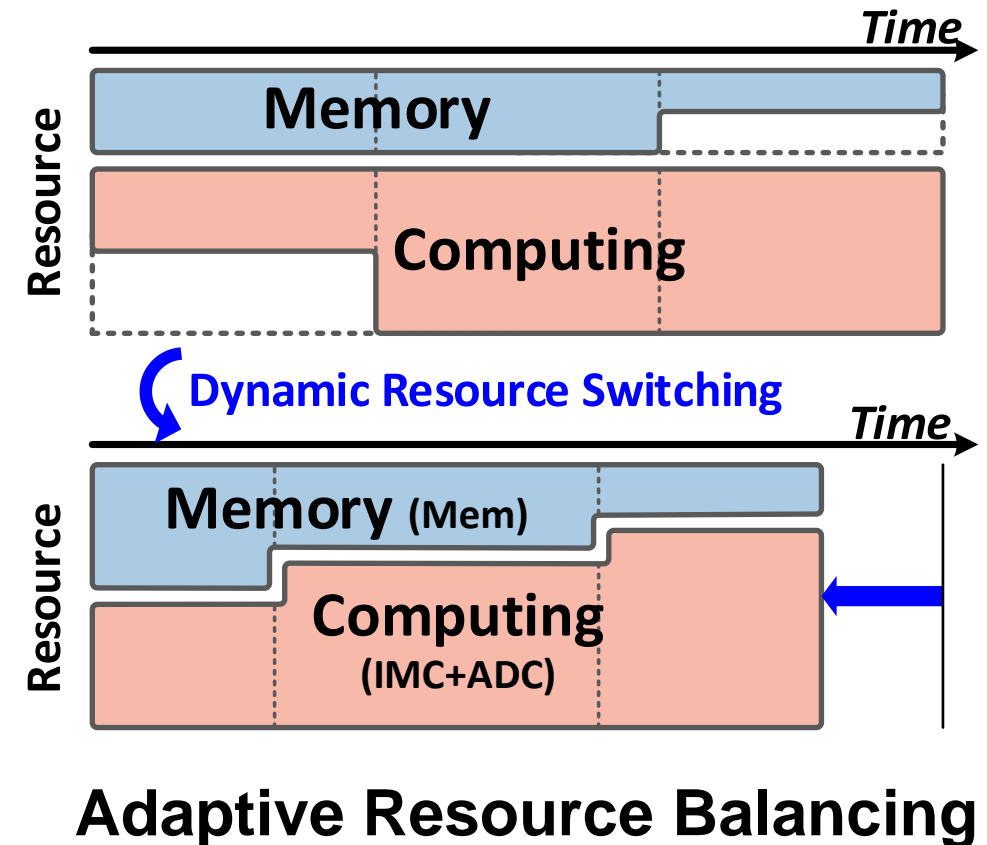
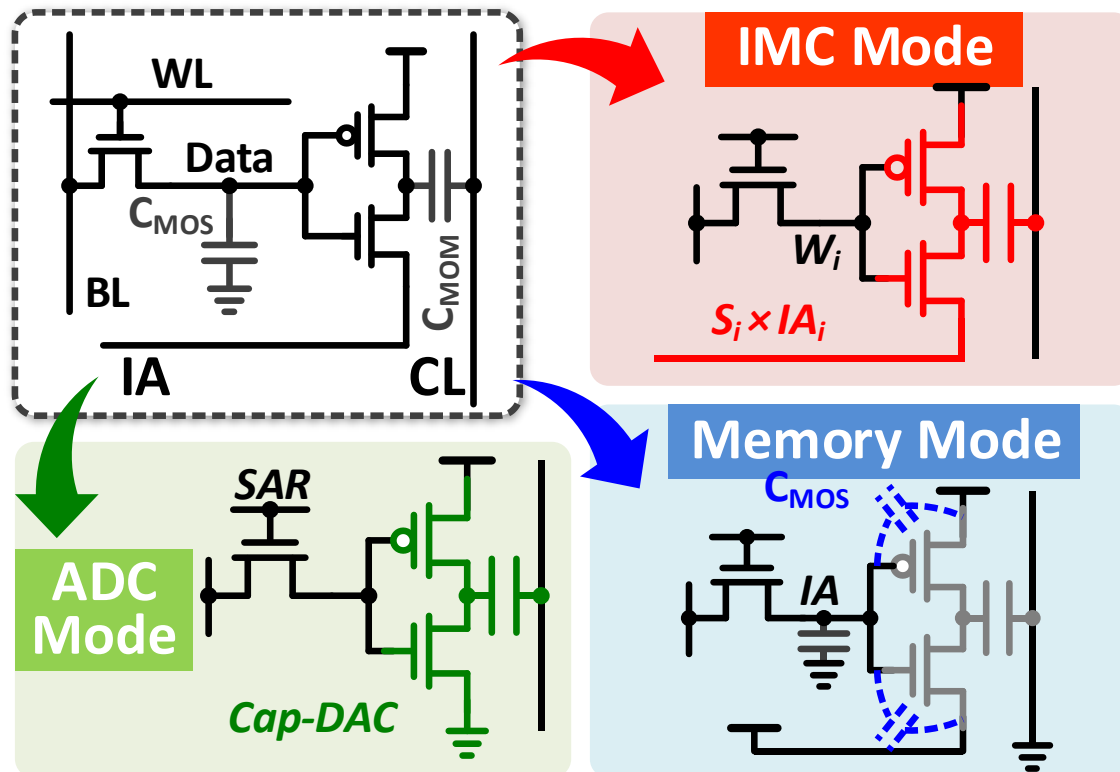
☐ CPU-Only    ☐ PIM-Core    ☐ PIM-Acc



# KAIST PIM: Triple Mode Cell

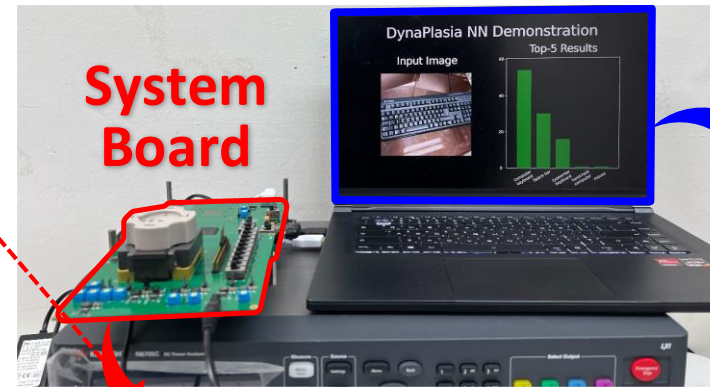
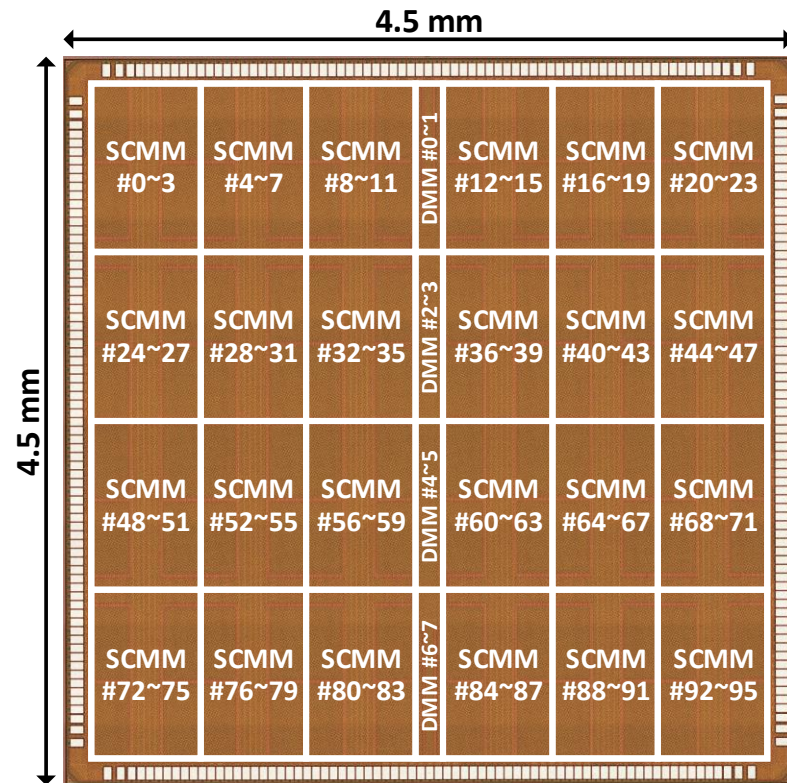
- **Multi-functional 3T-2C Cell**

- Support dynamic resource switching (Computing ↔ Memory)



# DynaPlasia

- DynaPlasia (ISSCC'23) : Reconfigurable IMC



S. Kim, et al. "DynaPlasia: An eDRAM In-Memory-Computing-Based Reconfigurable Spatial Accelerator with Triple-Mode Cell for Dynamic Resource Switching," , ISSCC 2023

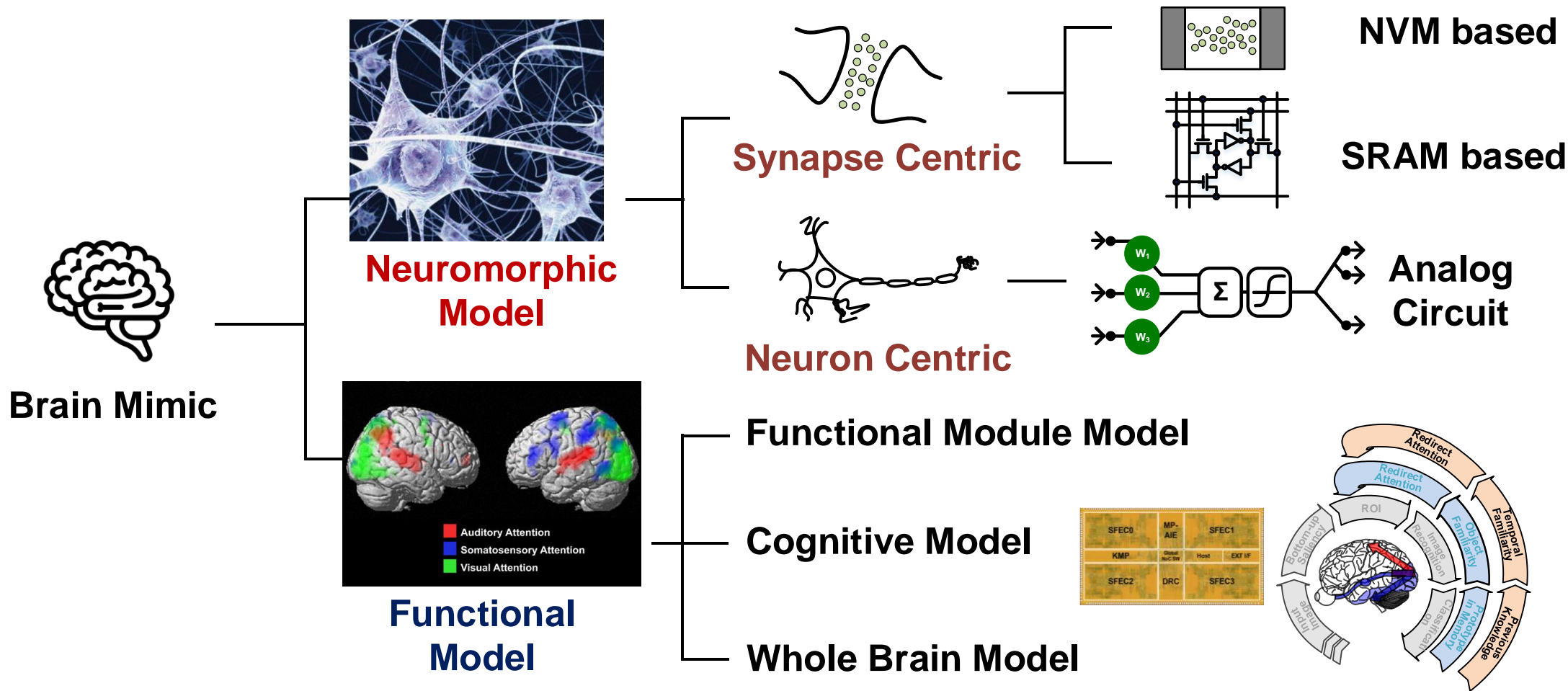
# 2023 DynaPlasia

AM IMC-based Reconfigurable NPU  
Chip Die  
Measurement Board  
카메라  
연합뉴스TV  
건조주의보  
물체 인식 결과  
1번 analog clock  
2번 wall clock  
3번 digital clock  
DRAM-PIM 코어 구조  
Layer 2-10  
Layer 11

06:37 월요일에 대규모 파업...공무원·전공의·교사 등 ▶수령 600년 천연기념물 제주 9.3°C

# Neuromorphic/Spiking NN

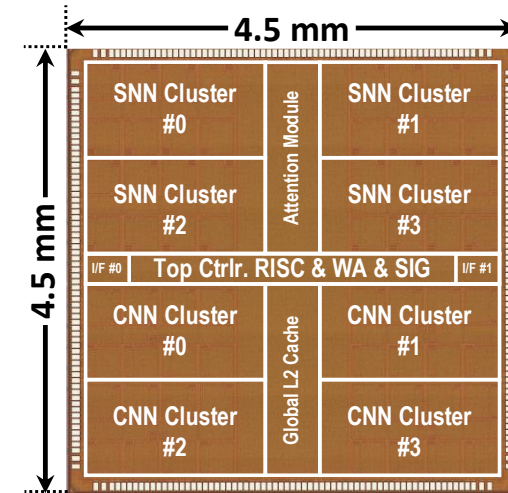
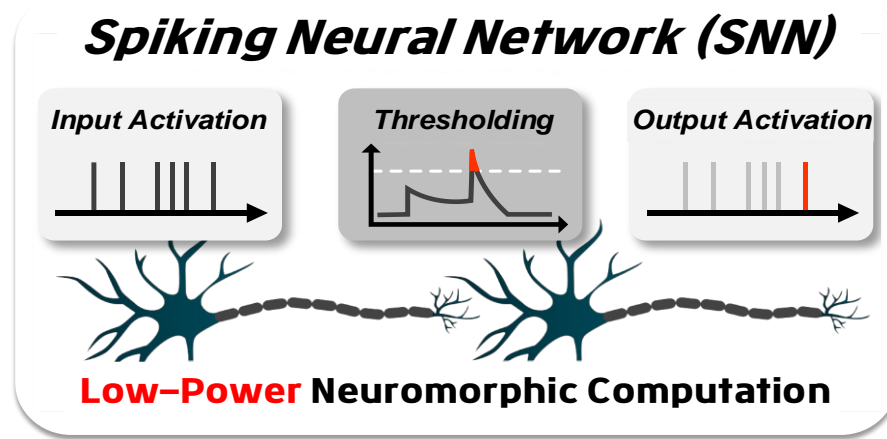
## □ Microscopic Brain Structure or Macroscopic Brain Function



# 2023 C-DNN: Complementary-DNN Processor

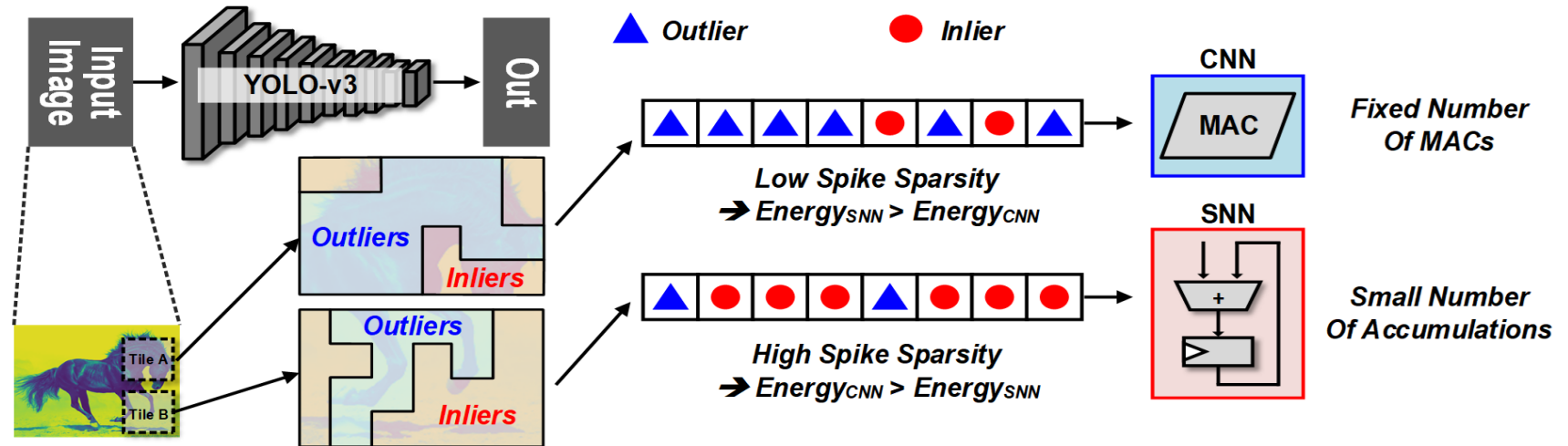
- Energy Efficient CNN/SNN Hybrid Processor

Brain-mimic  
Computation



Chip  
Photo

CNN-SNN  
Workload Division

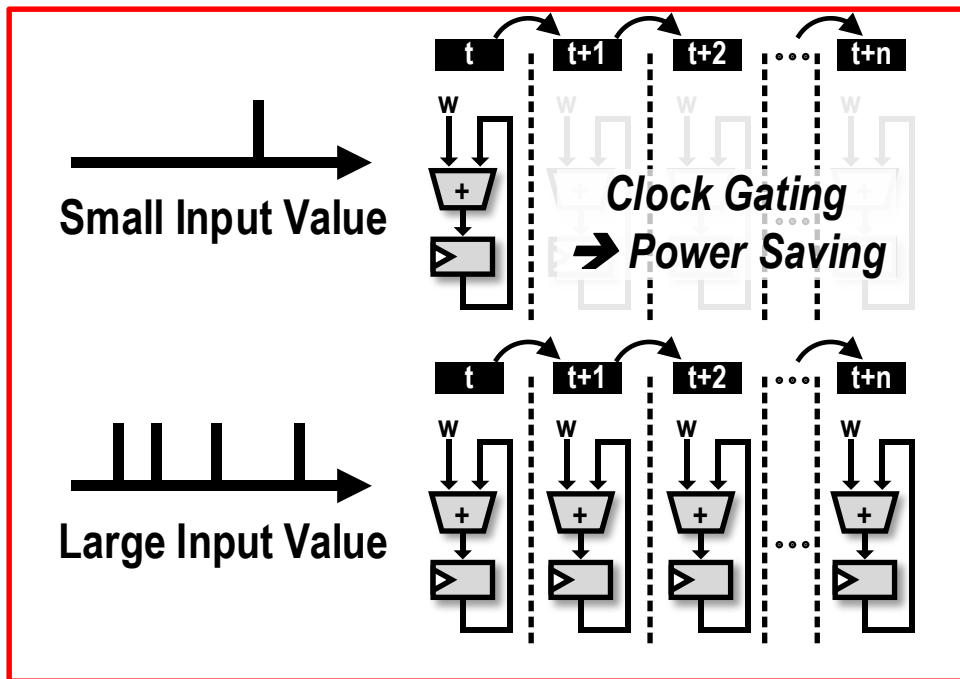


# KAIST C-DNN: Neuromorphic Accelerator

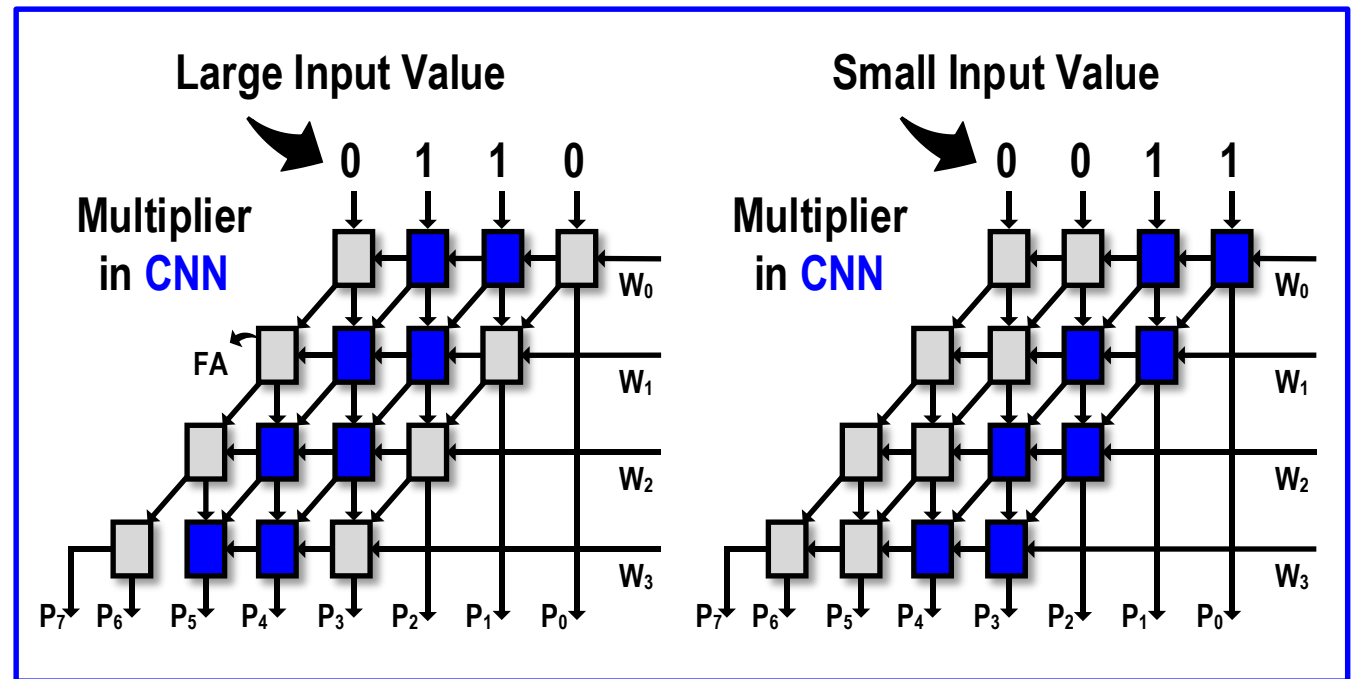
## □ Input magnitude incurs small performance variation in CNN

- Small magnitude input data  $\uparrow$   $\rightarrow$  **SNN domain efficient**
- Small magnitude input data  $\downarrow$   $\rightarrow$  **CNN domain efficient**

### SNN: Wide Power Variation

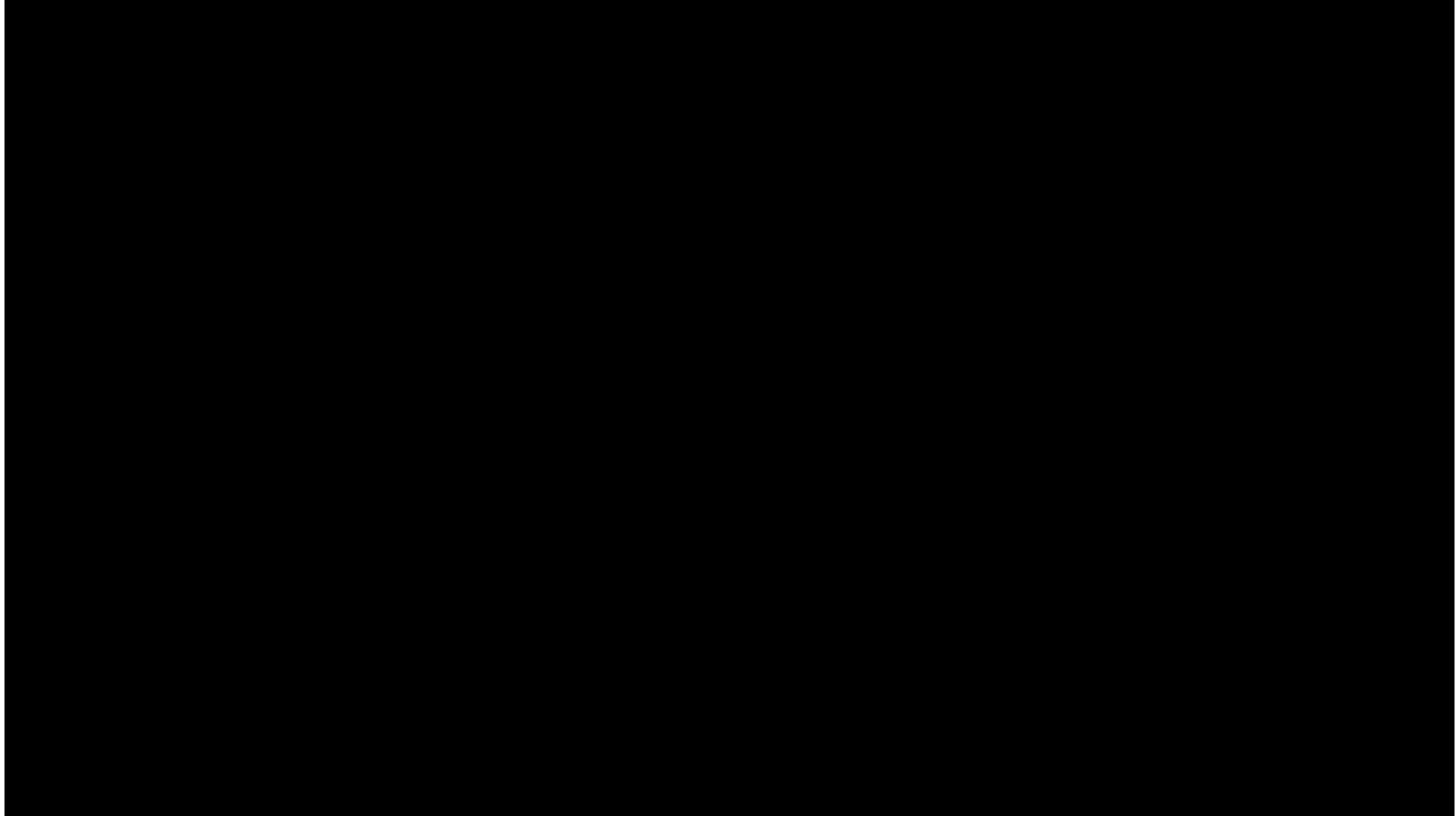


### CNN: Small Power Variation



# 2023 C-DNN: Demonstration Video

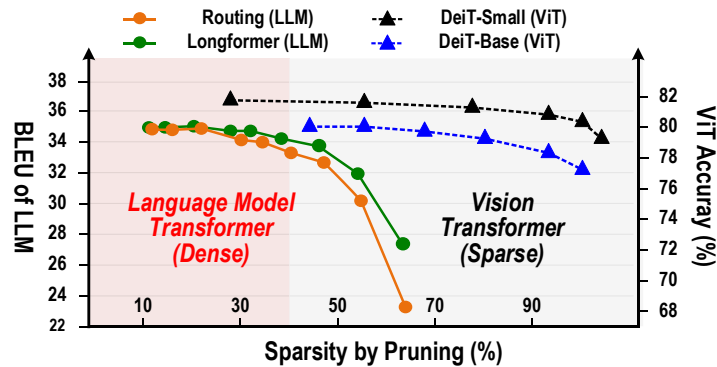
---



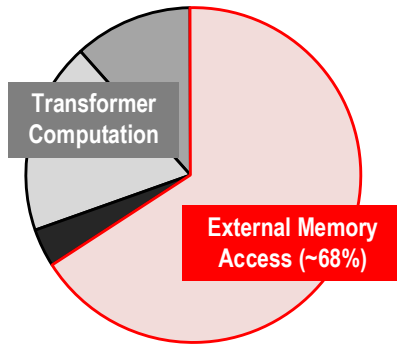
# 2024 C-Transformer : DNN/Spiking Transformer processor

- Motivation

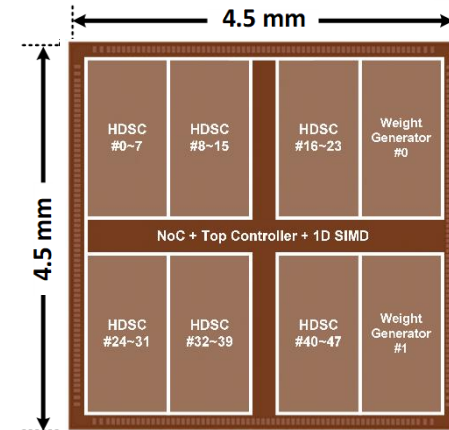
- Large External Memory Access



Sparsity-Accuracy Graph for Transformer Models

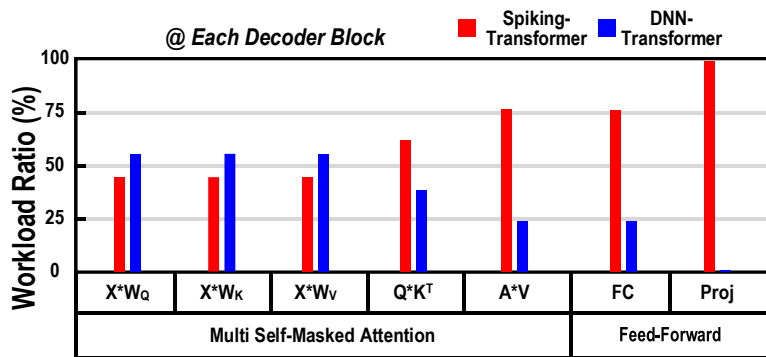


Energy Breakdown of LLM system

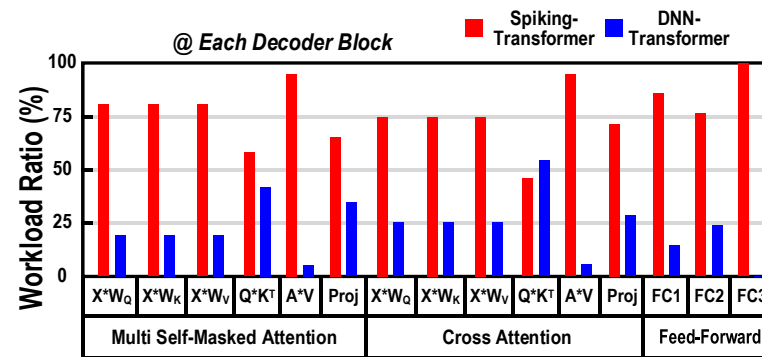


Chip Photo

- High Reconfigurability is Required



<Language Modeling with GPT-2 and Wikitext>



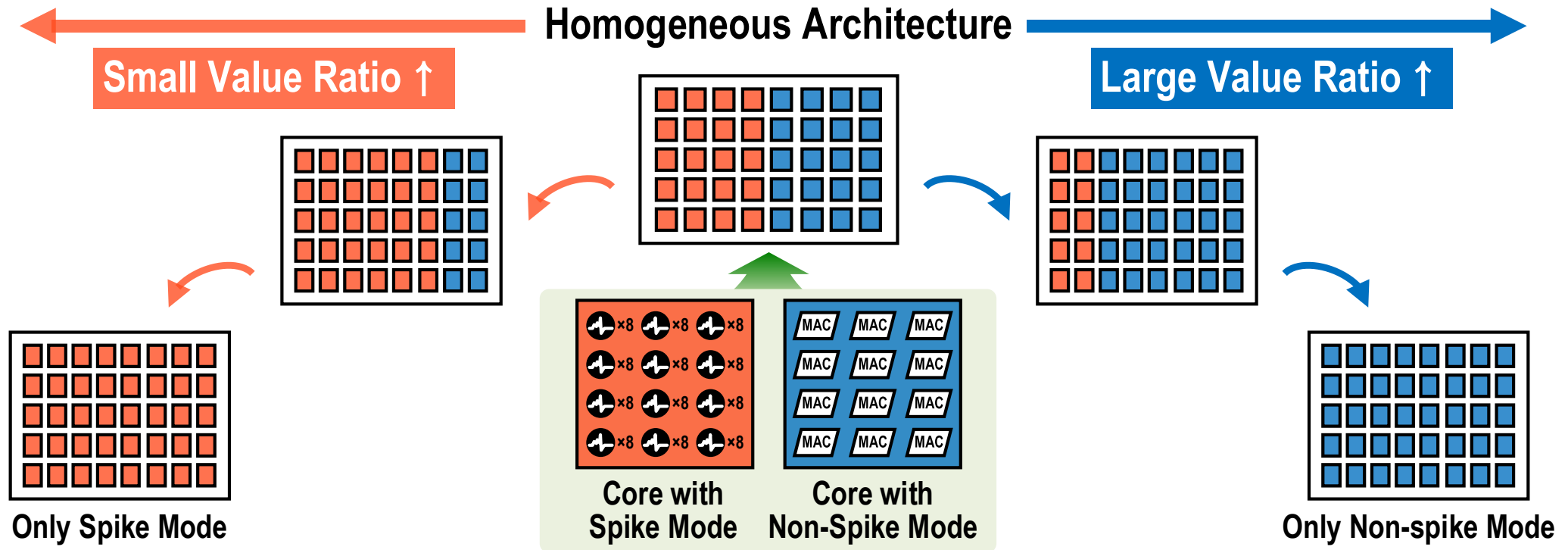
<Translation with mT5 and IWSLT>



# 2024 C-Transformer Architecture

- **Homogeneous DT/ST Core**

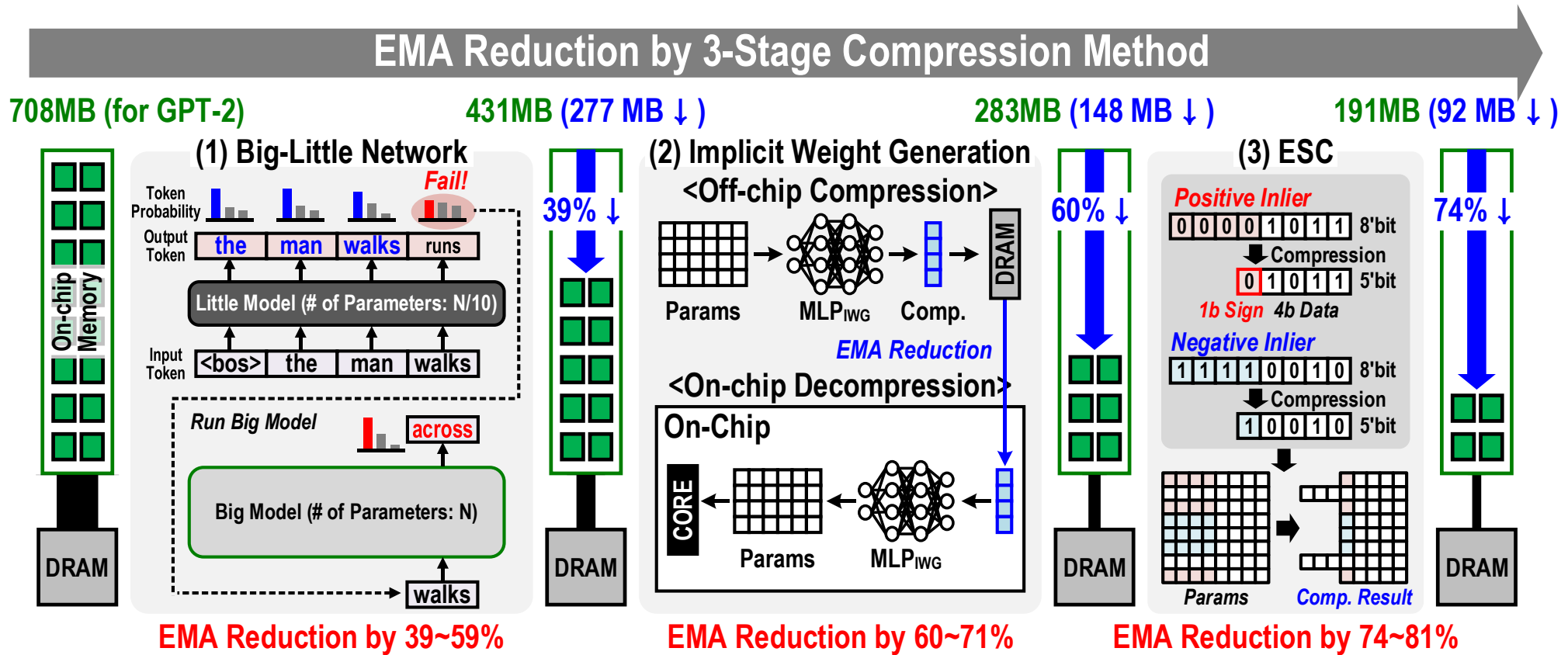
- Dynamically changed ratio of spike and non-spike domain  
→ Hybrid multiplication/accumulation unit (HMAU) is proposed!



# 2024 C-Transformer Architecture

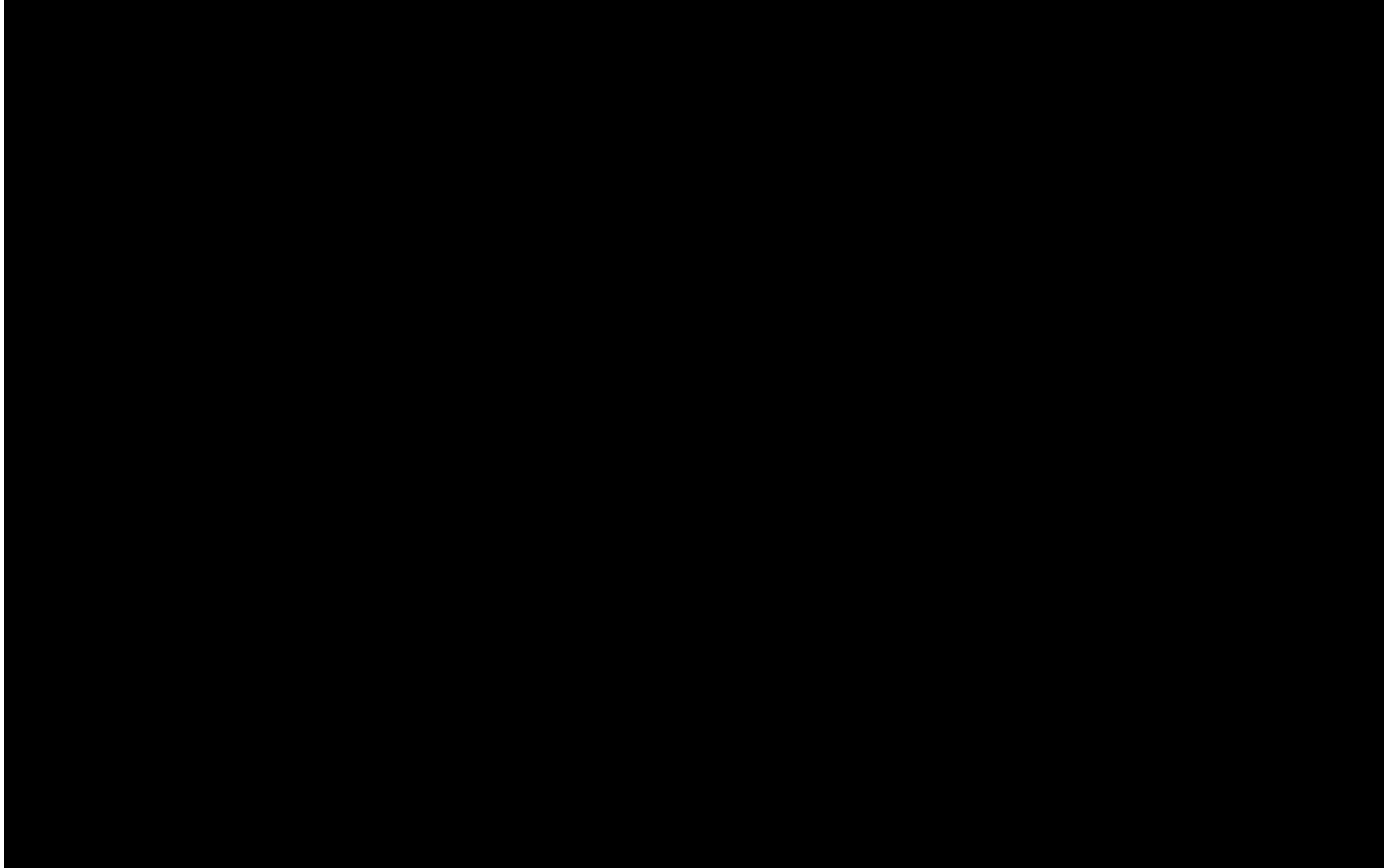
- Results of 3-Stage Compression

- Extended Sign Compression: 74~81% parameters are reduced



# 2024 C-Transformer: Demonstration Video

---



**KAIST**



**Thank you!**

