

# Near Data Processing in AI Era : Challenges and Opportunities

Memory System Research  
Eui-cheol Lim

# Contents

---

- 1. SK Hynix & MSR introduction (3)**
- 2. AI era Architecture trend (5)**
- 3. Memory based Solution Projects (5)**
- 4. Data Hierarchy – ultimate near data processing architecture (4)**

## ● SK hynix's Product Portfolio



## ● SK hynix's innovative solutions

### Datacenter Solutions

**DRAM** DDRx, HBMx

**NAND** SATA eSSD, PCIe eSSD

### Computing Solutions

**DRAM** DDRx, LPDDRx

**NAND** PCIe cSSD, SATA cSSD

**CIS** VGA, HD, FHD~8Mp

### Mobile Solutions

**DRAM** LPDDRx

**NAND** eMMC, UFS

**MCP** eMCP, uMCP

**CIS** VGA ~ 20Mp

### Graphics Solutions

**DRAM** GDDRx, DDRx, LPDDRx, HBMx

### Consumer Solutions

**DRAM** DDRx, LPDDRx

**NAND** eMMC, UFS

**MCP** eMCP

**CIS** VGA, HD, 2Mp ~ 20Mp

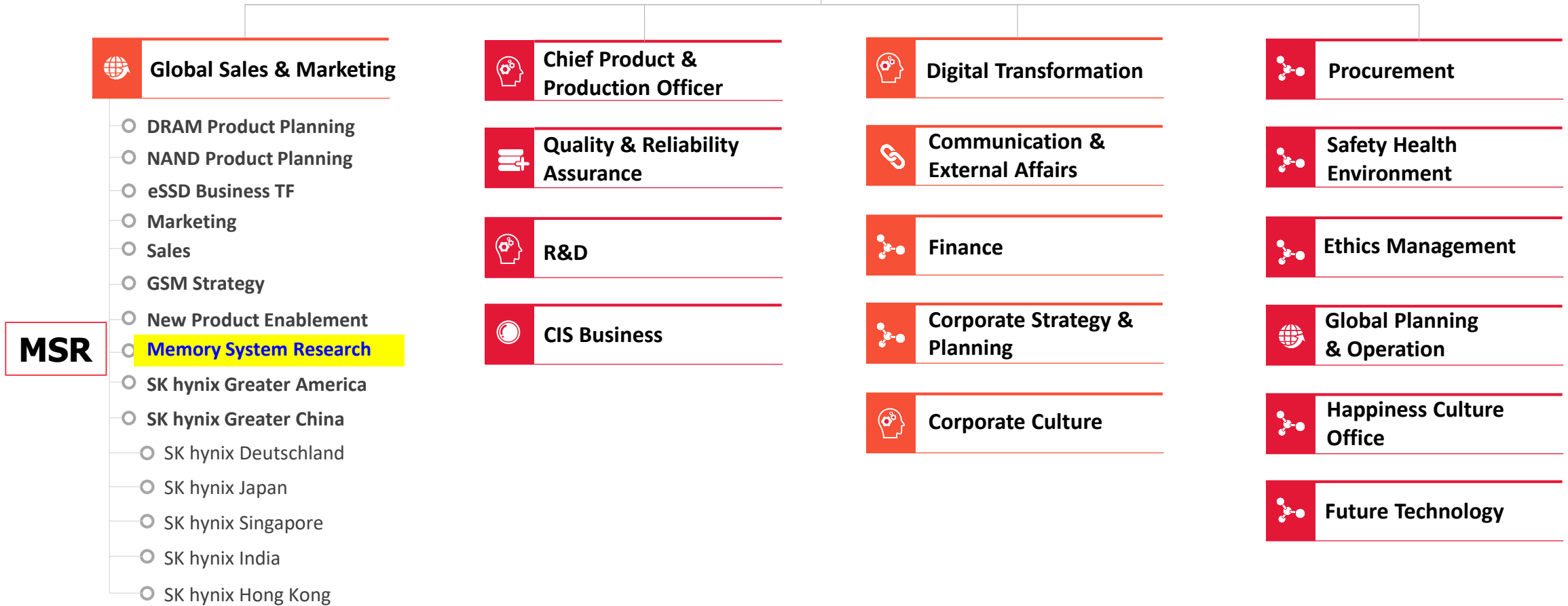
### Automotive Solutions

**DRAM** DDRx, LPDDRx, GDDRx

**NAND** eMMC

# MSR(Memory System Research) in SK hynix

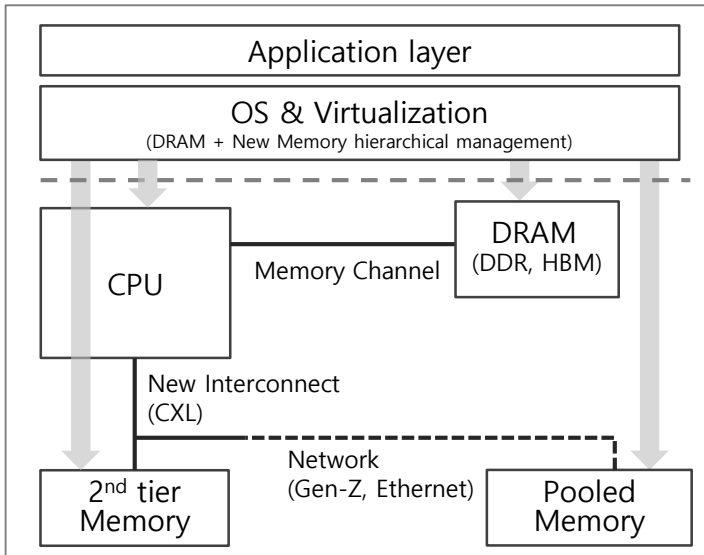
## SK hynix



1) Footnote : As of July, 2020

## Memory System for High Capacity Memory Server

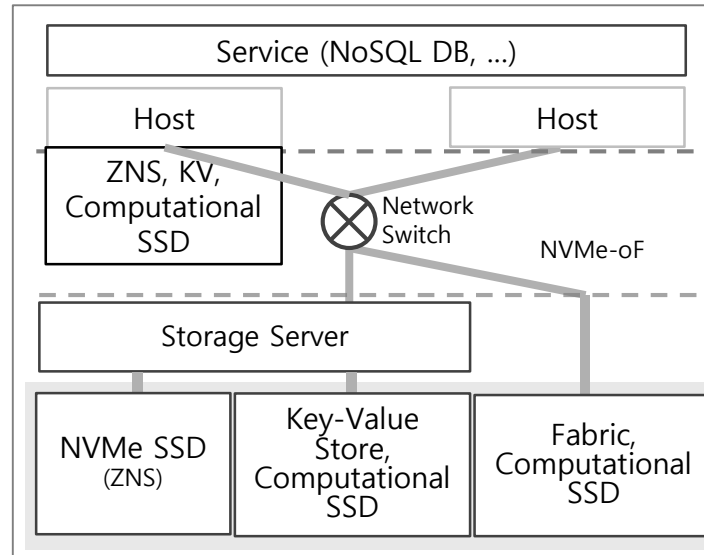
Advanced HW/SW solution for 2-tier memory (e.g. DRAM + MDS or PCM) to utilize high capacity and/or non-volatility



- 2<sup>nd</sup>-tier memory extension solution w/ new interconnect
- Management SW stack for new memory hierarchy
- Quantitative analysis of next-generation memory architecture

## Storage System for Next Cloud Storage

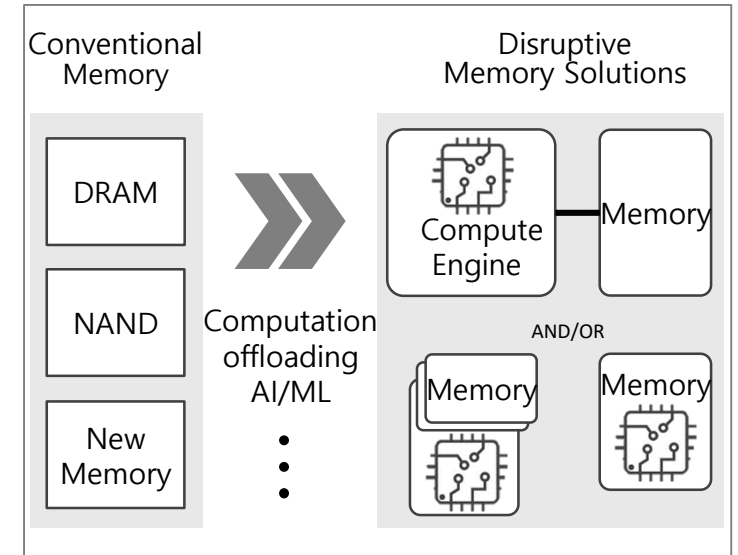
Core technology for next-generation storage devices (e.g. ZNS, KV SSD) suited for Software Defined Storage



- ZNS-based next generation SSD solutions
- Key-Value SSD
- Computational Storage:
  - ✓ In-line (Compression, Encryption) SSD
  - ✓ In-situ (Analytics) processing SSD

## Emerging Technology for AI/ML in Data Center

Memory-based deep-learning computing solutions for higher energy efficiency



- Near Data Processing for DLRM (Deep Learning Recommendation Model)
- AI Computing-in-Memory:
  - ✓ PNM (Custom HBM w/ an AI accelerator)
  - ✓ ANMAC (Analog MAC w/ resistive memory)
- AI SW stack for memory-based AI solutions

# Contents

---

1. SK Hynix & MSR introduction
2. AI era Architecture trend
3. Memory based Solution Projects
4. Data Hierarchy – ultimate near data processing architecture



# Computer vs. Human Brain


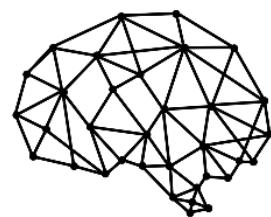
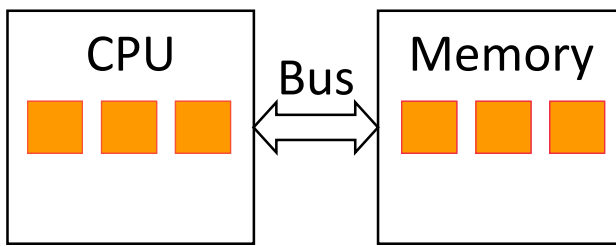
- AI has already proven that it is more capable than human in certain applications
- However in energy perspective, human brain is much more efficient than current computer system

## Go match: AlphaGo vs. Lee Sedol



# Human Brain vs. Neuromorphic vs. von Neumann

- Neuromorphic computing looks far better than von Neumann architecture, but it is in premature stage yet

	Human Brain	Neuromorphic	von Neumann
Architecture			
Device	Neuron / Synapse	Artificial Neuron / Synapse	CPU / Memories / Bus
Features	Event-driven	Event-driven / Asynchronous	Computing-centric / Compute, memory separated
Encoding Scheme	Spiking Signals	Analog/Digital/Mixed Spiking Signals	Binary Signals
Accelerator	-	Neuromorphic Processor	GPU / FPGA / ASIC
Power / Energy Consumption	Ultra Low	Low	Relatively High



# Energy Cost of Data Movement

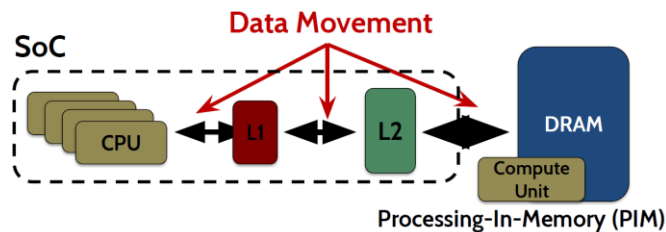
- Data movement consumes more than 60% of power in modern processors
- AI algorithm is more memory bounded than conventional app.

## Power Portion by Data Movement

- More than 60% of power is consumed by data movement
  - Analyze power consumption in popular google apps

### Energy Cost of Data Movement

1<sup>st</sup> key observation: 62.7% of the total system energy is spent on data movement



Chrome



TensorFlow



Video Playback

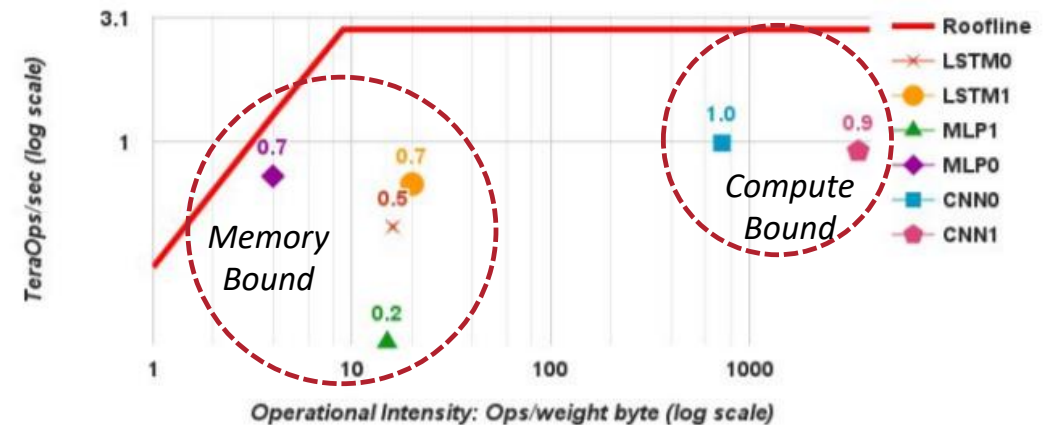


Video Capture

Amirali Boroumand, "GoogleWorkloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS'18

## Low Operational Intensity

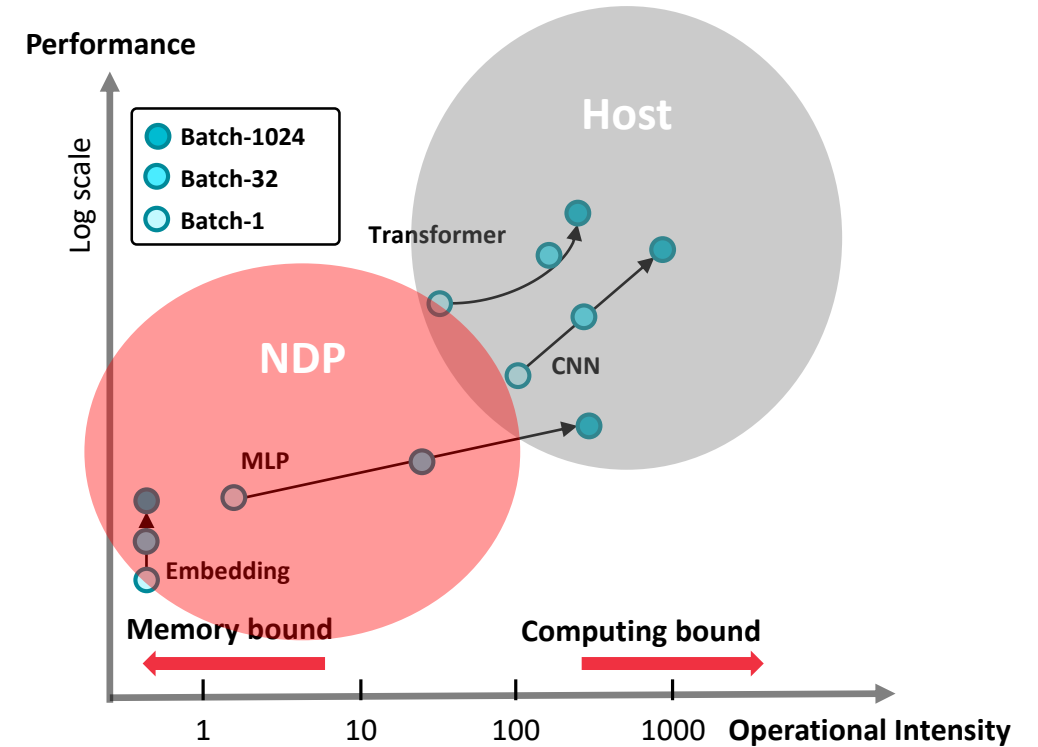
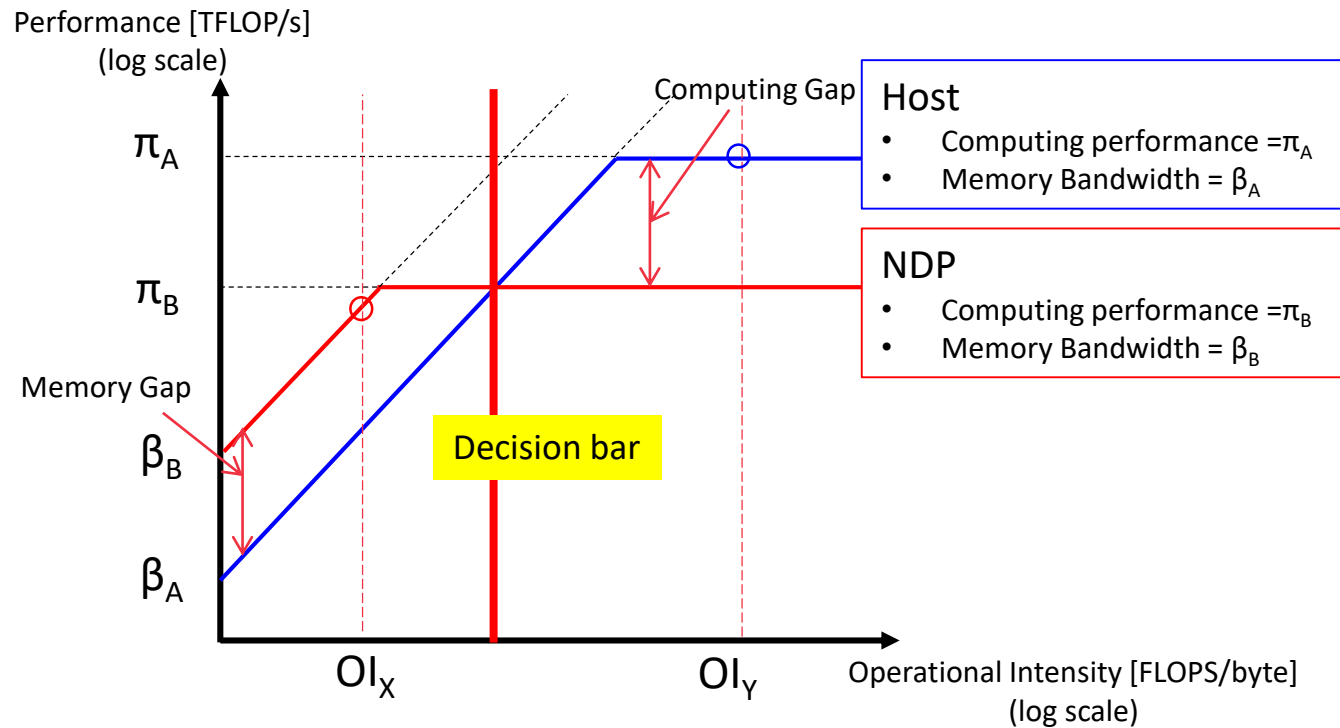
- In most cases, NN performances are bounded by BW
  - Recent NN Accelerators use internal memory to reduce bottleneck, but it is not sufficient



Norman P. Jouppi, "In-Datcenter Performance Analysis of a Tensor Processing Unit," ISCA'2017

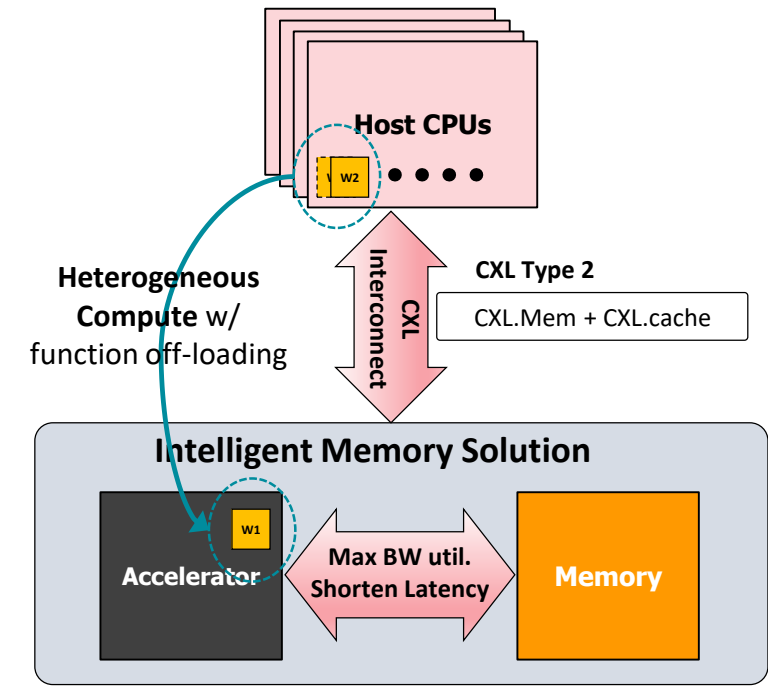
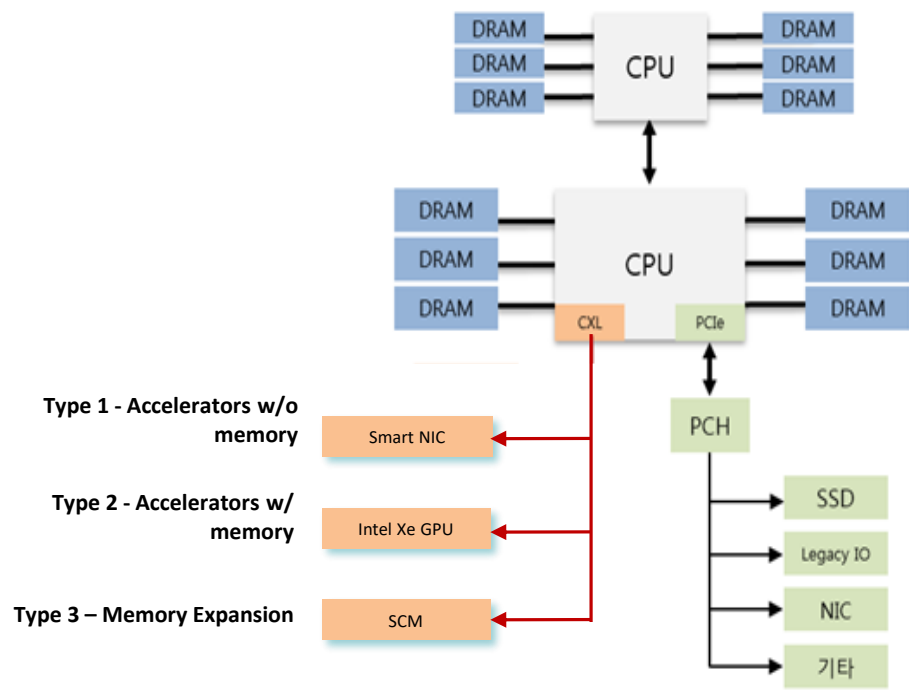
- **Operational Intensity of AI algorithm decides the appropriate computing architecture (Host vs. PIM)**

- High Operational Intensity: Host processing is better in performance & efficiency
- Low Operational Intensity: PIM is better in performance & efficiency



# Interconnect : good news for NDP

- Interconnect will be the core of the Server system which has heterogeneous architecture
- Intel introduces the CXL interconnect which enables Intel CPU will be the center of heterogeneous Computing System.
- Opportunity: Value added Memory solution
  - Unlike conventional memory system such as DIMM, CXL supports hand shaking protocol which enables us to build additional value on it. Ex) DRAM cache, Data processing engine...



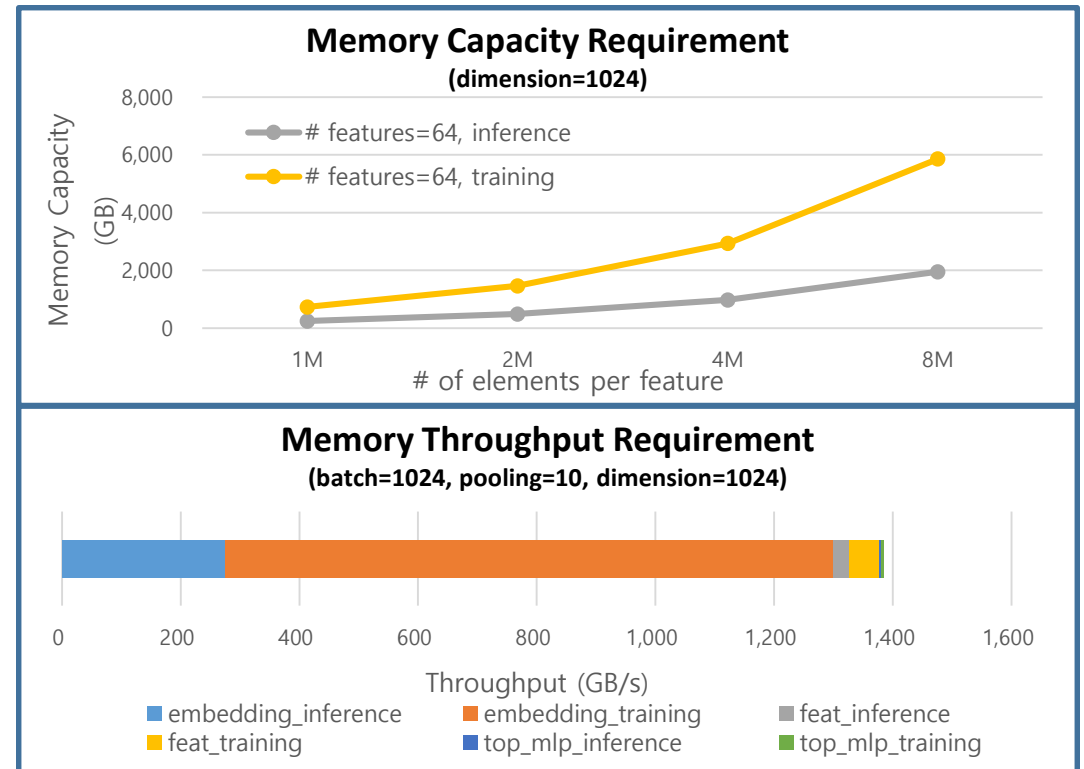
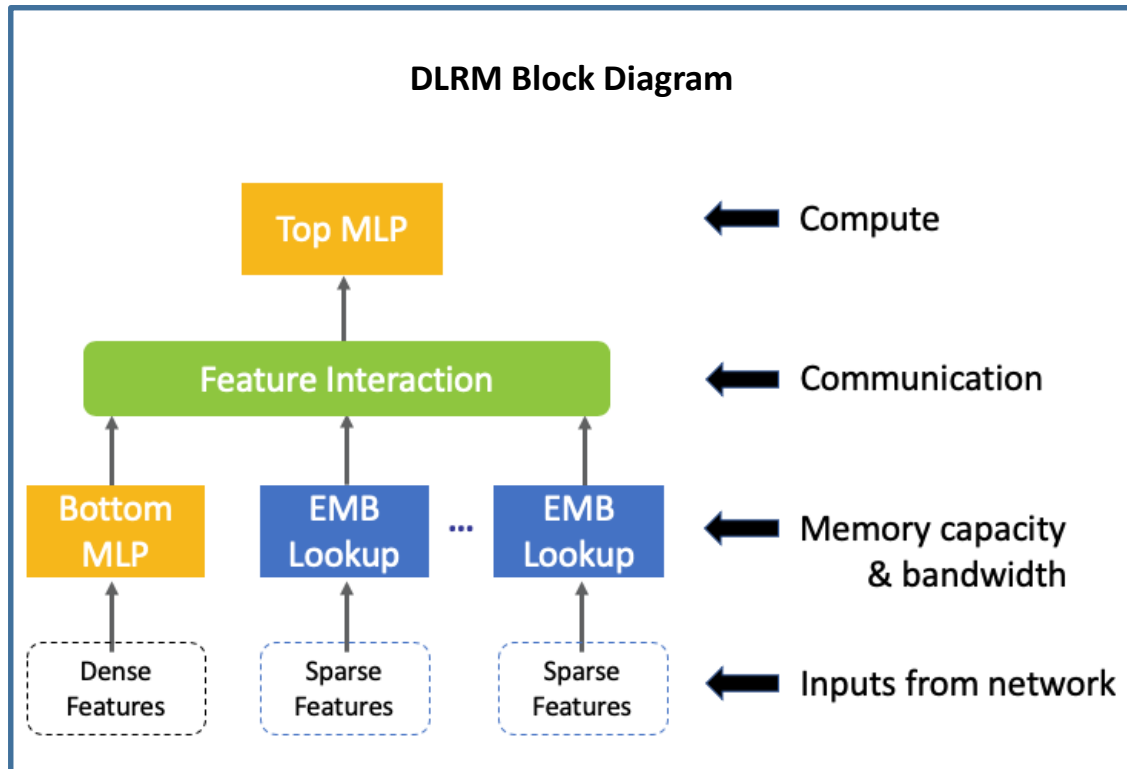
# Contents

---

1. SK Hynix & MSR introduction
2. AI era Architecture trend
- 3. Memory based Solution Projects**
4. Data Hierarchy – ultimate near data processing architecture

# Recommendation Model

- Recommendation is one of the key AI services that provides majority of revenue to datacenter companies
- DLRM is an advanced, open source Deep Learning Recommendation Model (from Facebook)
  - Composed of 3 layers (Embedding, Feature Interaction, Top/Bottom Multilayer Perceptron) and each layer has different computing characteristics
  - As the model size and # of data becomes large, the memory requirements for embedding goes higher

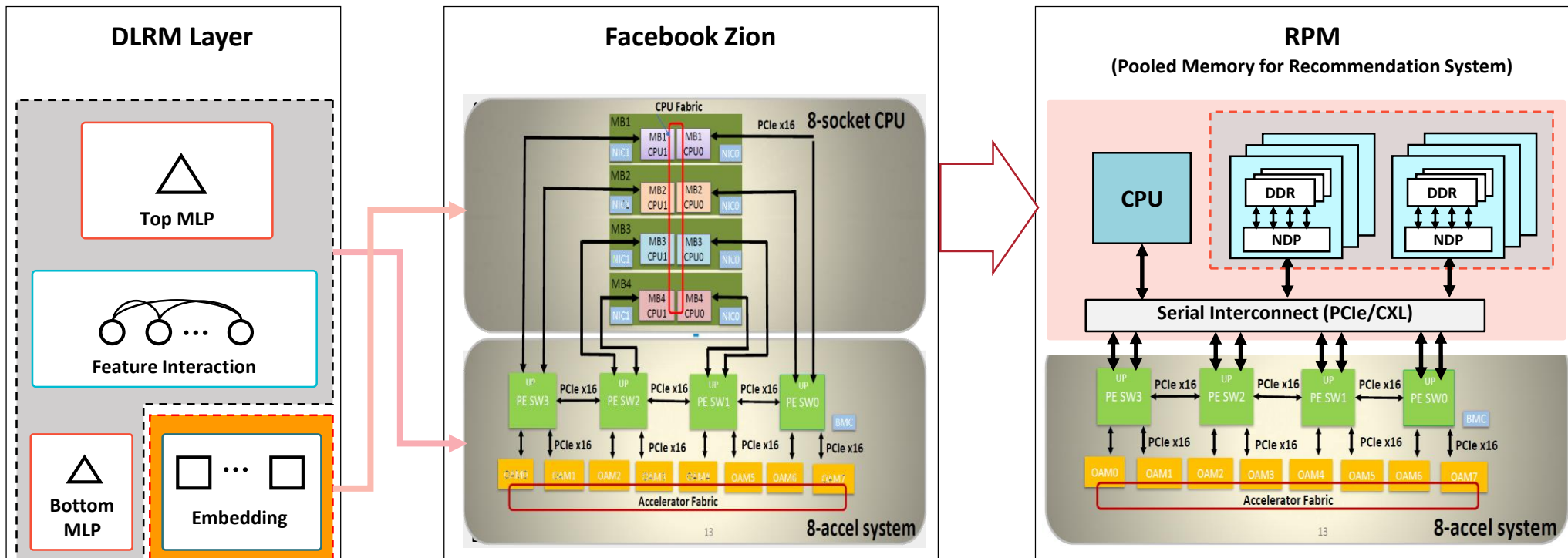


# RPM(Recommendation Pooled Memory) Project Overview

## ● Limitations in Conventional System

- Due to different compute characteristics of each layer, conventional system is using two different compute environment (Zion) – 8 socket CPU Server for Embedding and OAM (OCP Accelerator Module) for Feature Interaction, Top/Bottom MLP
- However, (1) CPU has a limited memory capacity and bandwidth to fulfill the ever increasing memory requirements for Embedding; (2) CPU is not adequate for small OI applications

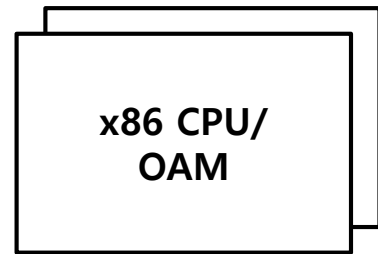
## ● RPM: Applying Near Data Processing concept for Embedding layer to resolve the limitations in CPU



# RPM(Recommendation Pooled Memory) Project Challenges

## ● Several challenges to overcome when we adopt Near Data Processing concept to Embedding layer

- Bandwidth limitation between the x86 CPU server (or OAM) and RPM
- Load balancing through the RPM modules
- Utilization for each RPM module
- RPM architecture for advanced RM algorithm

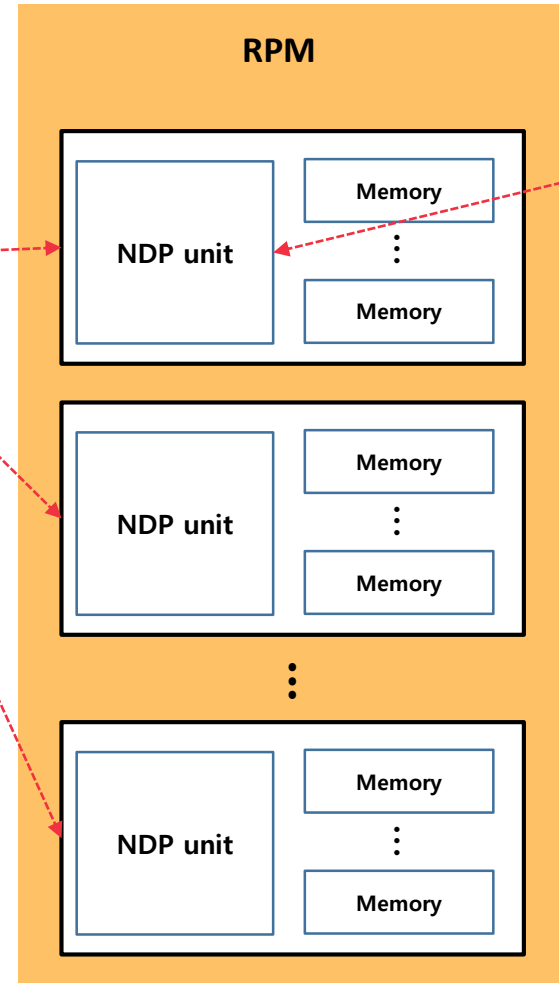


### Load Balancing?

- How we could fully utilize all the RPM?

### Bandwidth Limitation?

- How we could overcome the bandwidth limitation for interconnect?



### RPM Utilization?

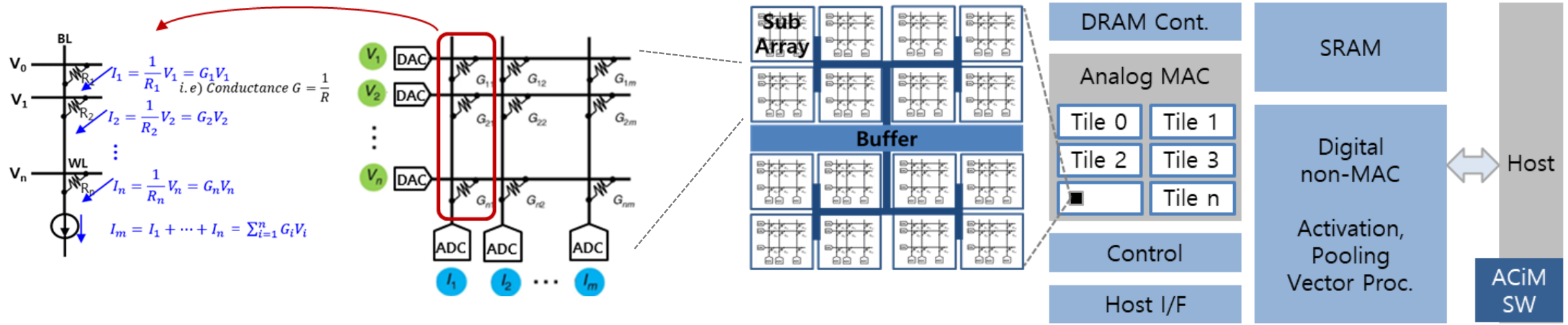
- How we could fully utilize the memory capacity and bandwidth of RPM?
- What is the best compute engine for NDP unit to support RM inference as well as training?

### RPM Architecture for Advanced RM?

- What is the best RPM architecture for advanced RM algorithm?

# ACiM (Analog Computing in Memory) Project Overview

- **Meaningful synergy between AI computing technologies and Memory technologies**
  - PNM & PiM - Computing digital logic located near or in memory.
  - **ACiM - Computing logic is made of memory cell (ReRAM, PCRAM and so on.)**
- **Target to achieve higher 'Perf./Power' and lower 'Energy consumption' than those of digital accelerators**
  - Perf./Power - Digital MAC: < 2TOPs/Watt vs. Analog MAC array: > 200 TOPs/Watt
  - Power consumption: < 1 Watt



[ Resistive memory cell array for MAC computations ]

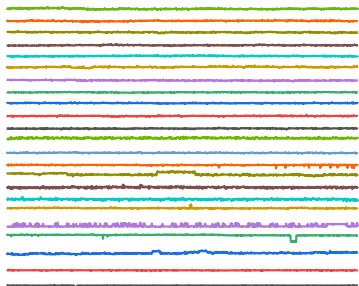
[ AI Accelerator using ACiM tech (Analog MAC) ]



# ACiM (Analog Computing in Memory) Project Challenges

- Need to consider different things in all levels from a device level to an application level
- Need a holistic approach and tight collaboration among various R&D participants

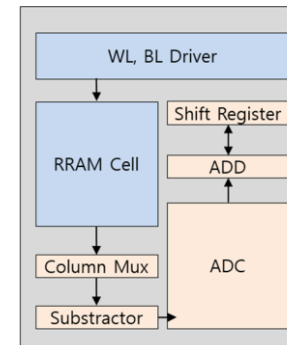
## Device



High-bit density cell

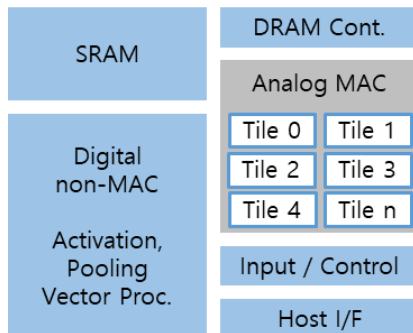
- Multi-level cell
- Retention & Reliability
- Read speed
- I-V linearity

## AMAC (Analog MAC)



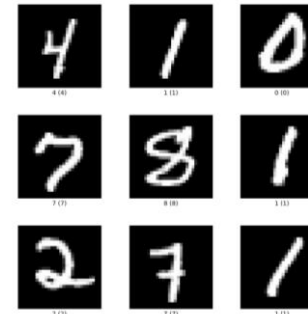
- ADC/DAC resolution
- Optimum sub-array size
- Ultra high Perf./Power in AMAC
- Analog aware inference

## Accelerator



- Energy efficiency in digital part
- SRAM minimization
- AMAC controller

## Application & SW



- Low precision algorithm
- Application specific
- ACiM specific SW toolkits (e.g.) TFlite, Edge TPU compiler

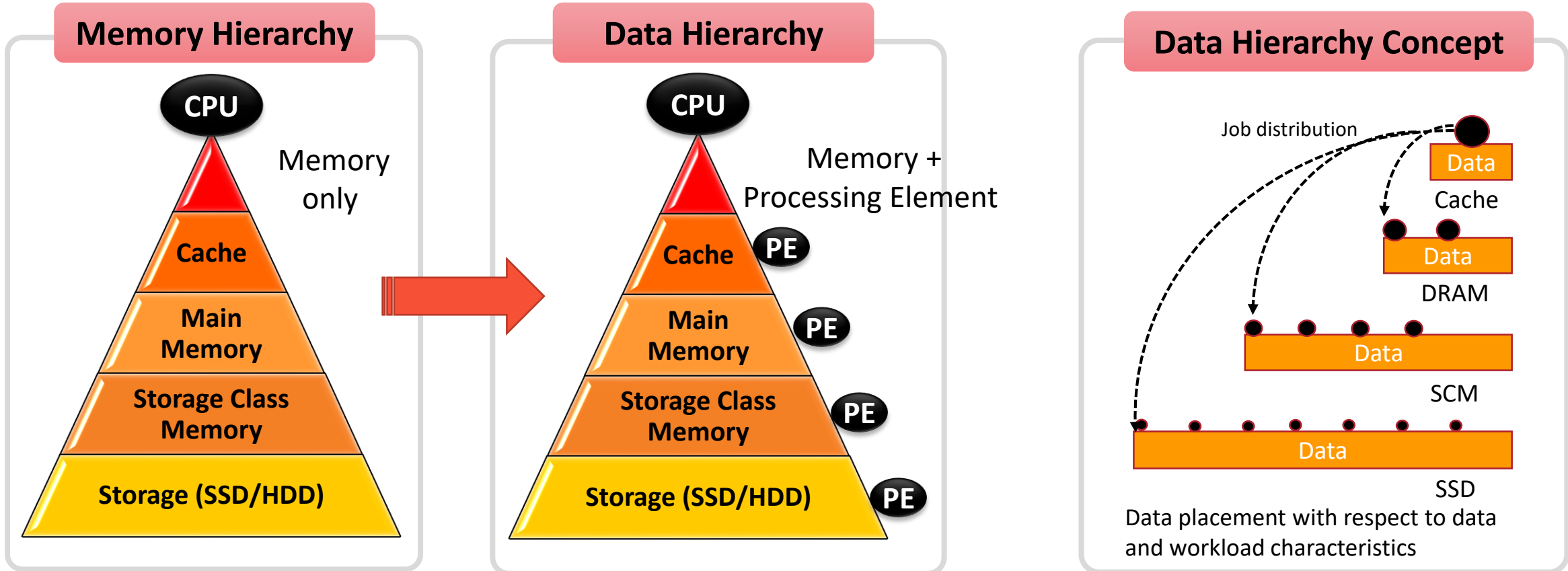
# Contents

---

1. SK Hynix & MSR introduction
2. AI era Architecture trend
3. Memory based Solution Projects
4. **Data Hierarchy – ultimate near data processing architecture**

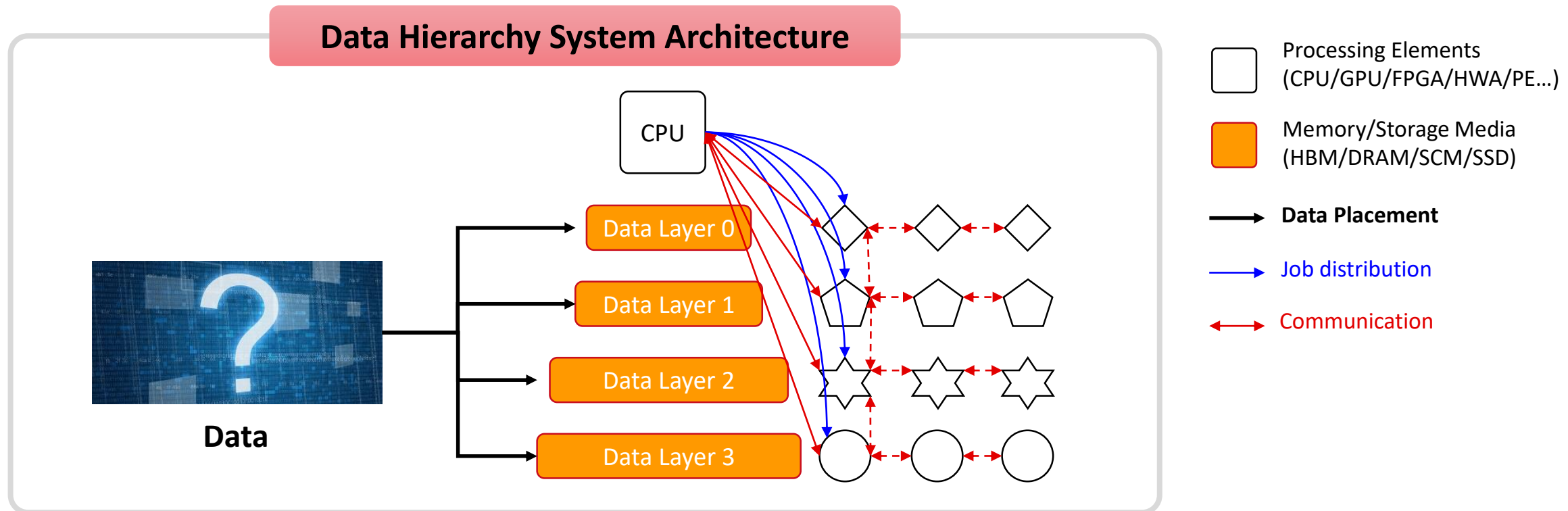
# Data Hierarchy – NDP in All Memory Layers

- **Architecture Concept – Every Memory layers have own processing element**
  - Minimizing data movement for the entire memory hierarchy
  - Data placed based on the data & workload's characteristics
  - Compute data where data reside



- **Data placement**

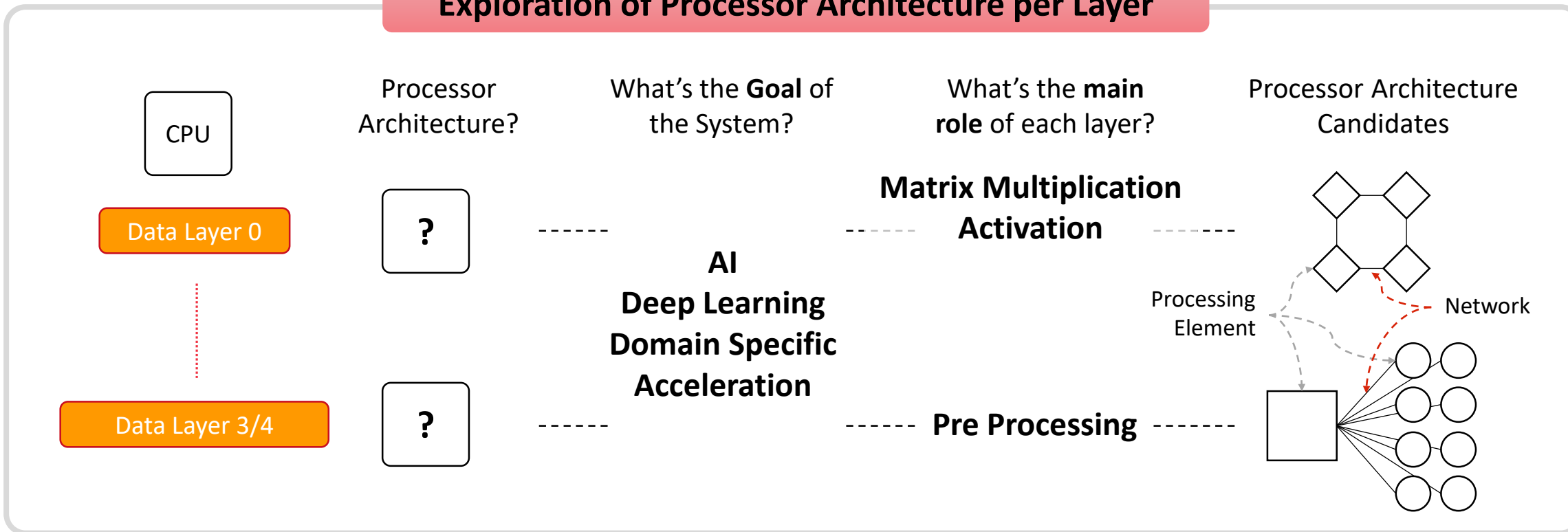
- How to **classify** the characteristics of data?
- How to **place** the data at the appropriate layer
- Whether the characteristics of data will be **determined only by the data itself**, or **by the algorithms** that utilize the data



- Processor architecture per layer

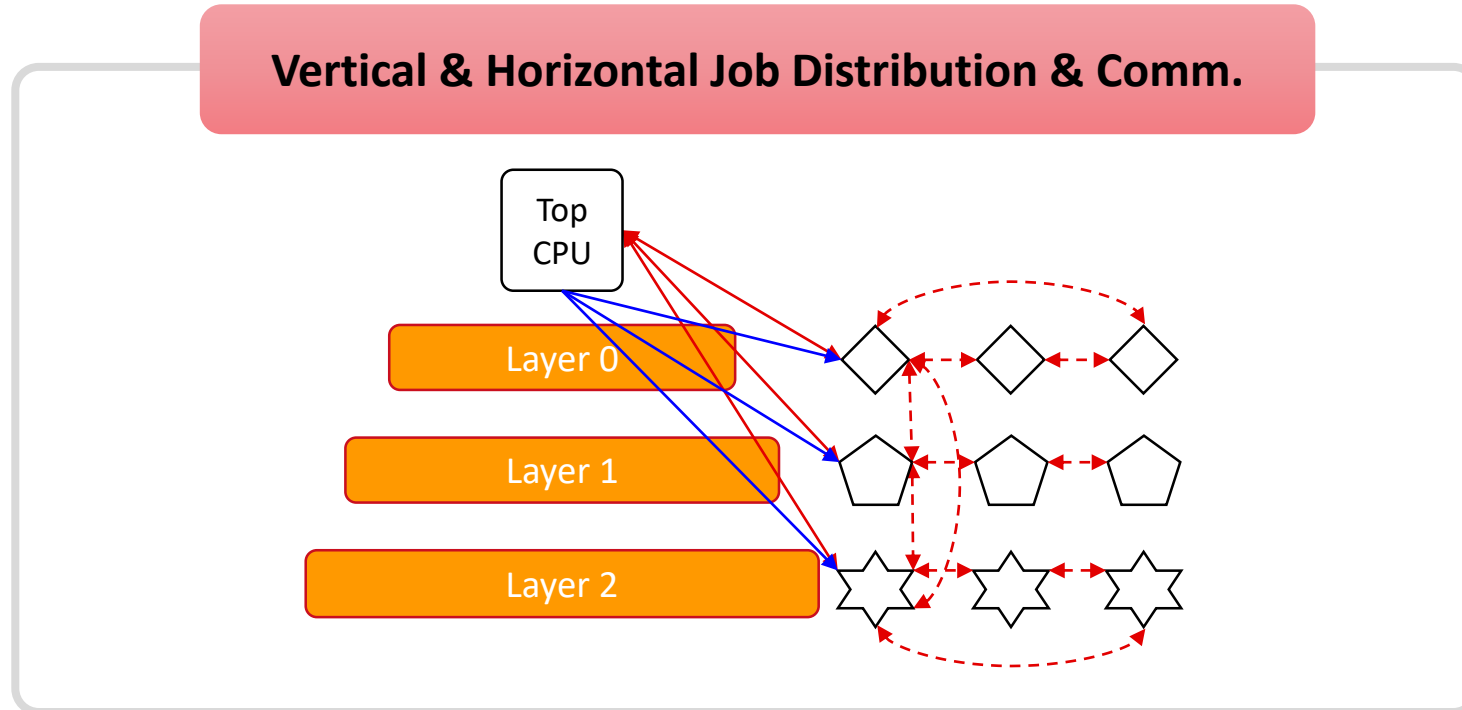
- What is the feasible structure of processing elements for each layer?
- **Domain specific architecture** is the key

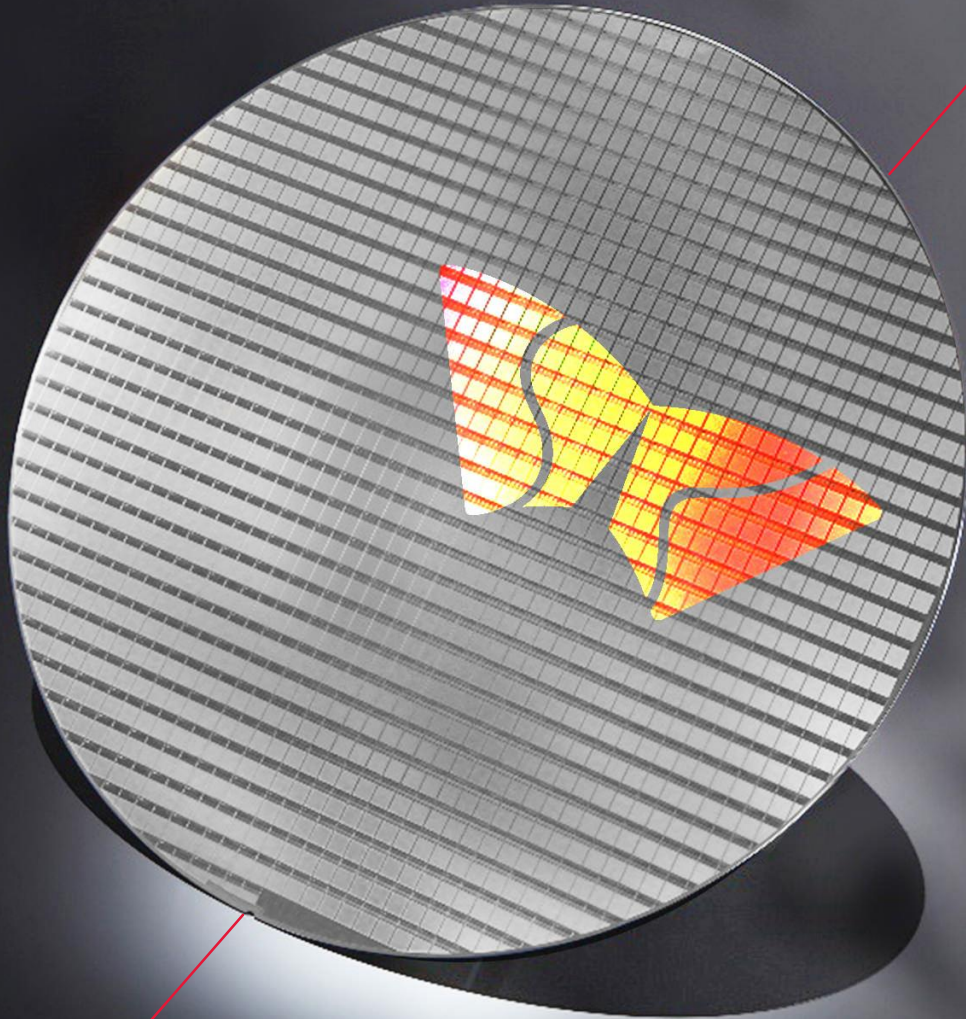
## Exploration of Processor Architecture per Layer



- **Job distribution & Communication**

- In terms of job distribution, Data Hierarchy is similar to heterogeneous computing
- Framework is required to assign a job to each layer and aggregate the results





# THANK YOU

[euicheol.lim@sk.com](mailto:euicheol.lim@sk.com)

