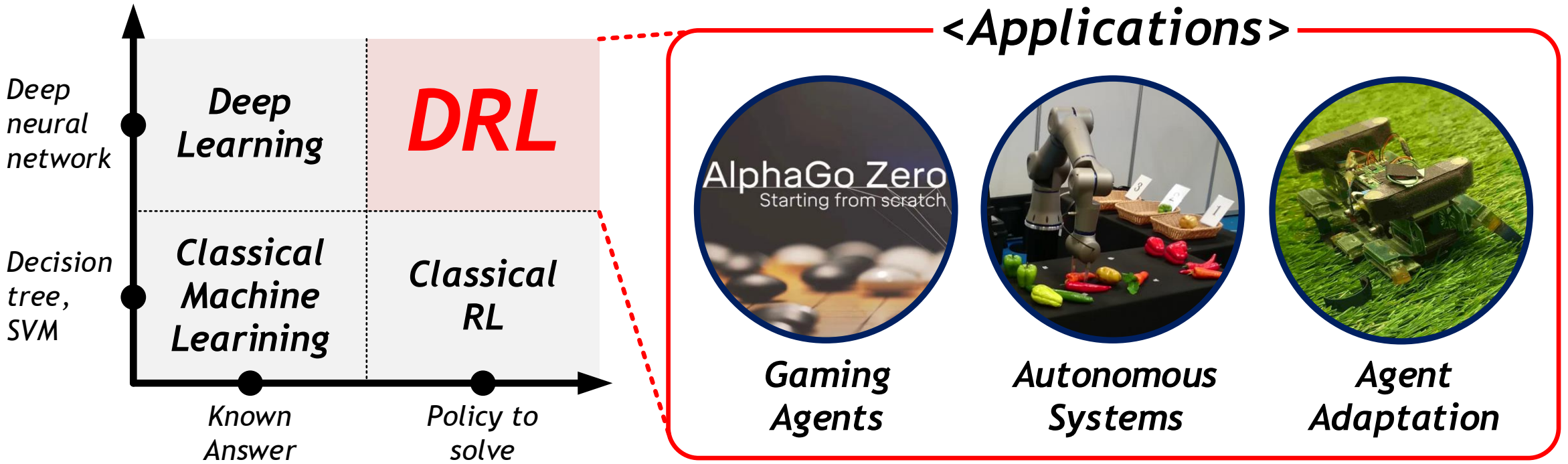
# OmniDRL: An Energy-Efficient Mobile Deep Reinforcement Learning Accelerators with Dual-mode Weight Compression and Direct Processing of Compressed Data

**Juhyoung Lee,** Sangyeob Kim, Ji-Hoon Kim, Sangjin Kim,
Wooyoung Jo, Donghyeon Han and Hoi-Jun Yoo

**Semiconductor System Lab.**
**School of EE, KAIST**

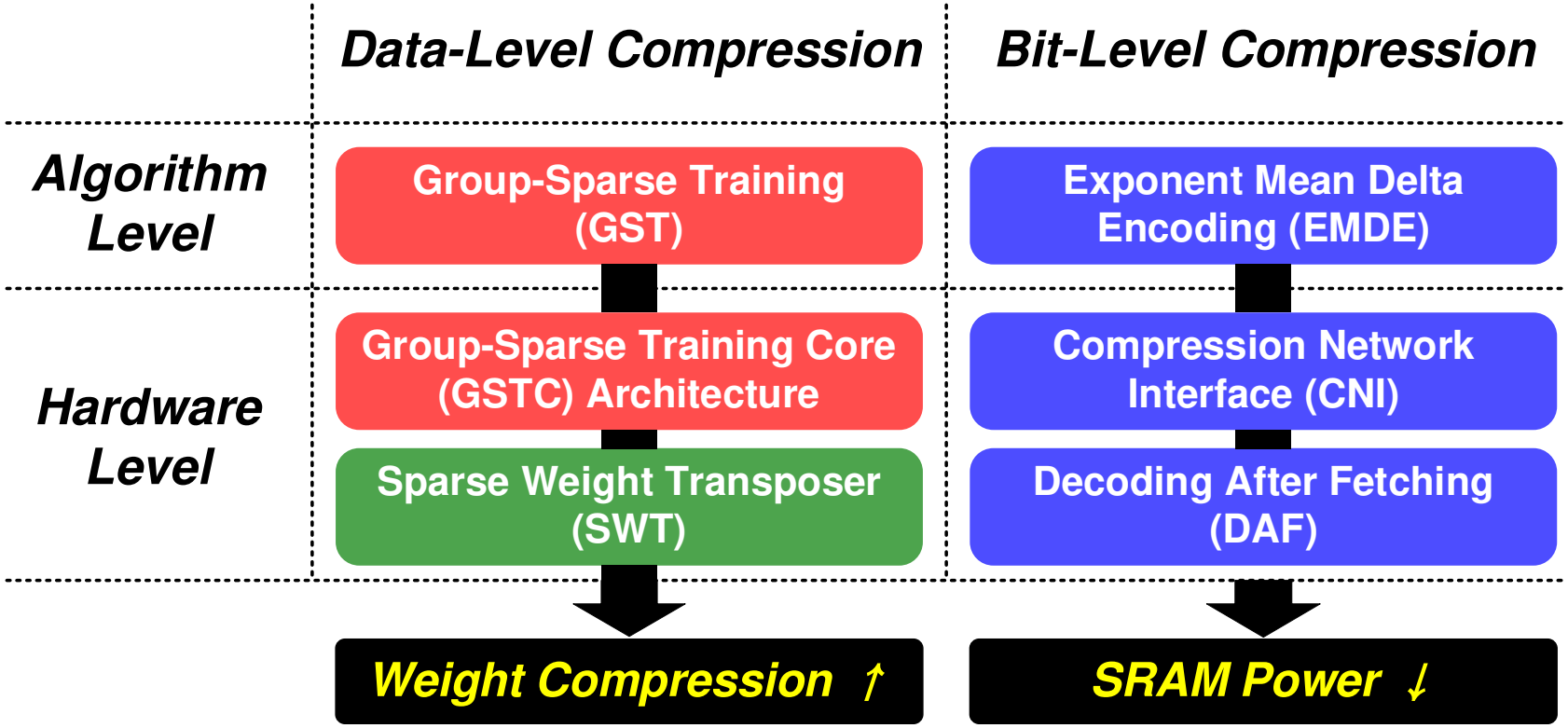
# Deep Reinforcement Learning (DRL)

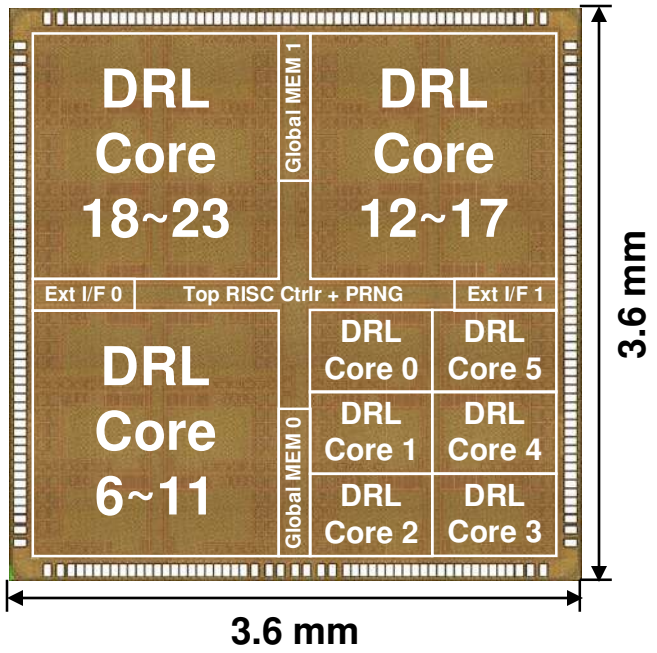
- **No Pre-labelled Data → Training with Trial-and-errors!**
  - Sequential decision making problems @ Unknown environments
  - Applications: gaming agent, autonomous systems, agent adaptation



# Abstract – OmniDRL: DRL Processor

- **Minimizing both External & Internal Memory Access!**
  - *Concept: compress data & direct utilization of compressed data*

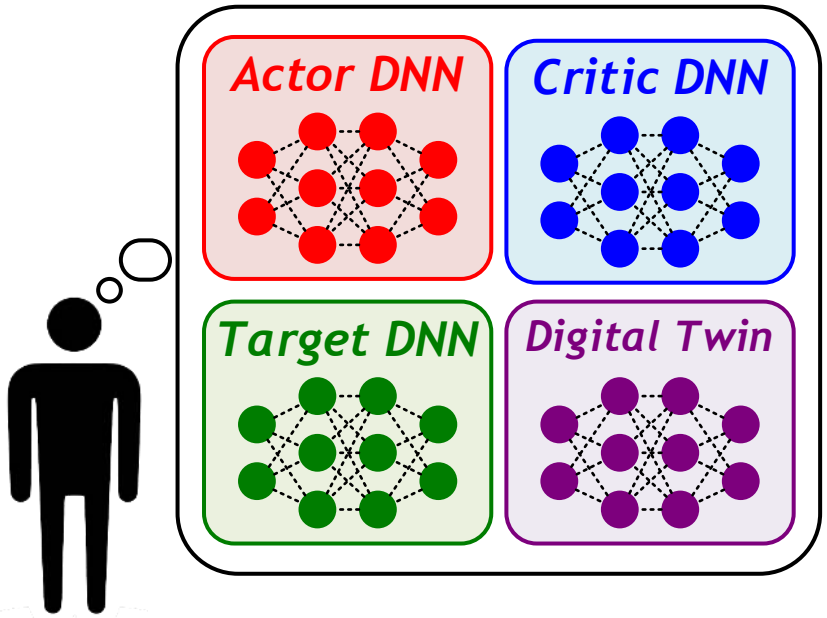


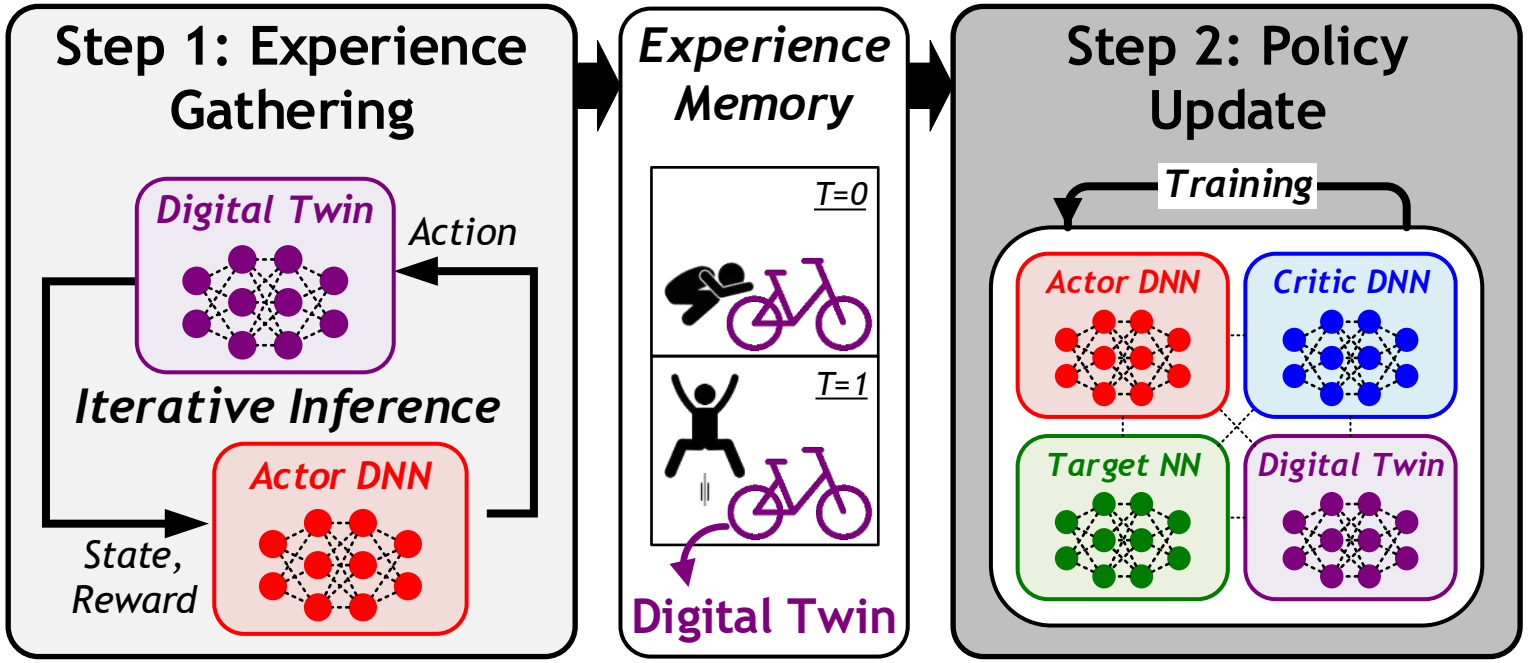
# DRL with Multiple DNNs

- **Trends: Utilizing Multiple DNNs (>3) for DRL**
  - Environment modelling w/ "Digital Twin" gathers experiences for training

**<Multiple DNNs in DRL>**

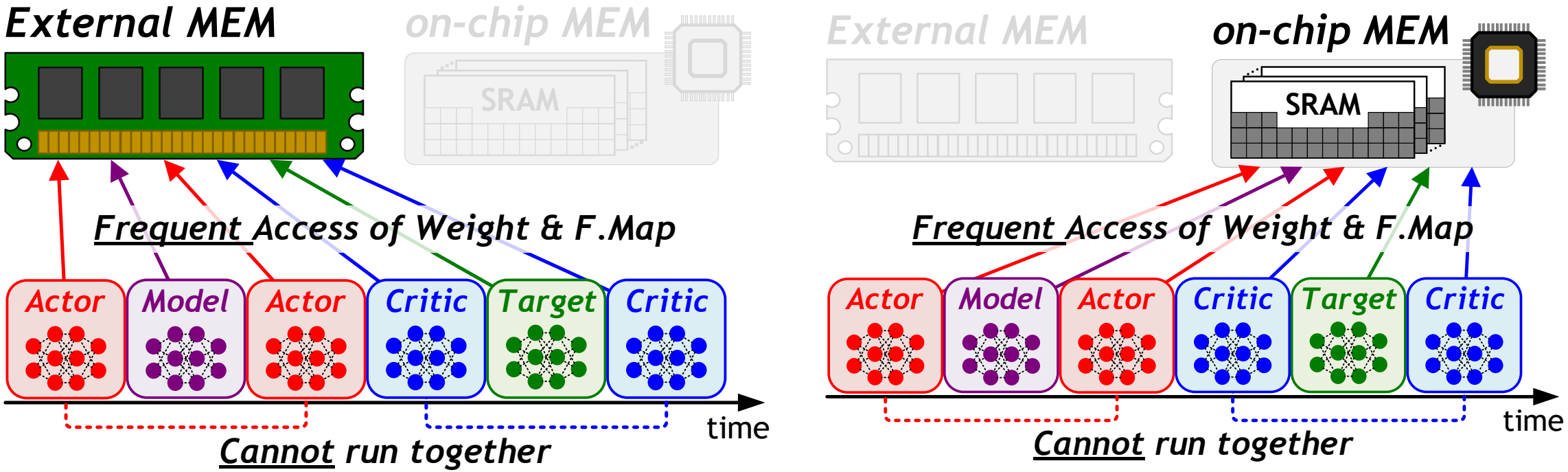



**<Overall flow of DRL training scenario>**



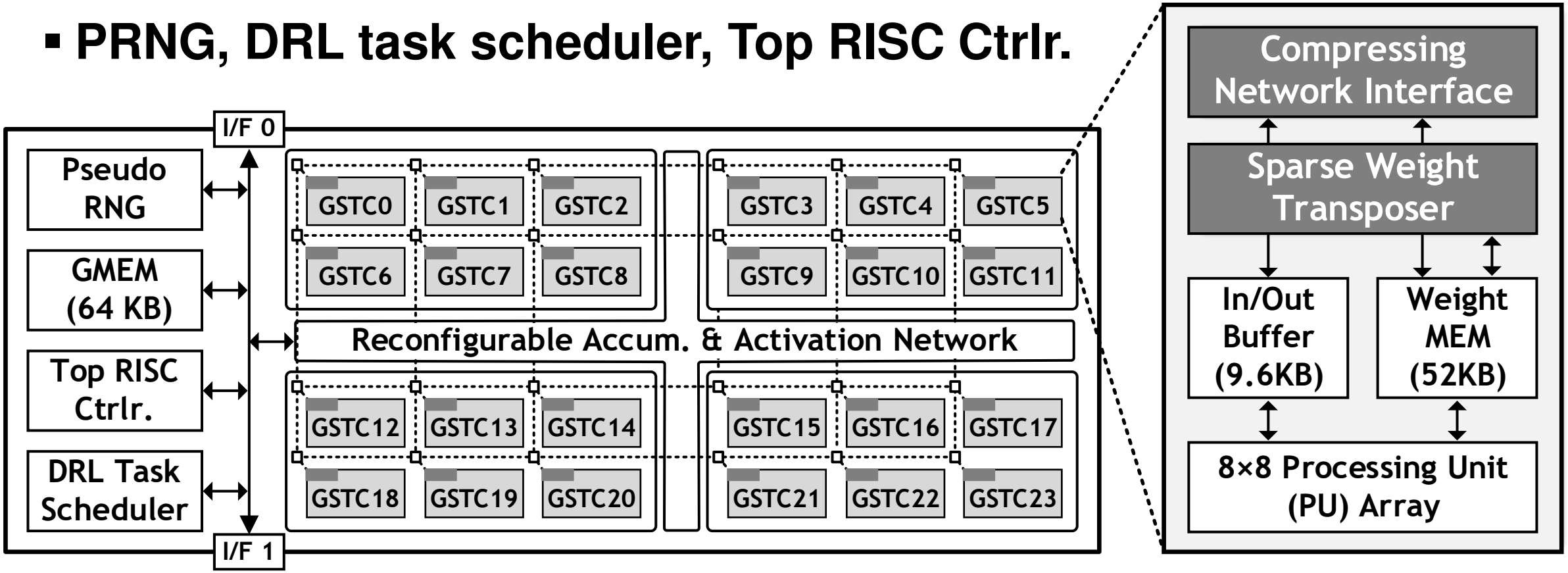
# Challenges of DRL Processor

- **1. Limited Computational Intensity (~60 Ops/byte)**
- **2. Dominant SRAM Power Consumption (> 53.6 %)**
  - Reason: Multiple DNNs (mainly FC) require complex, sequential execution



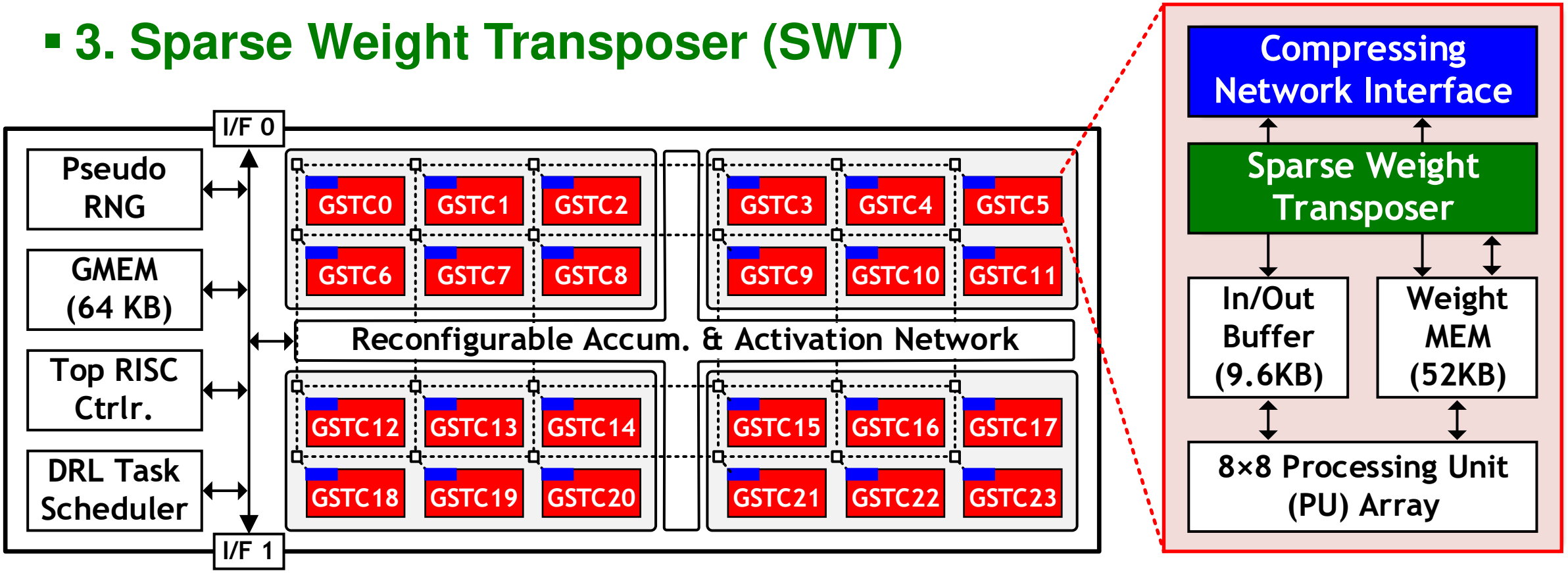
# Overall Architecture of OmniDRL

- 24 Group-Sparse Training Core (GSTC)
- 2-D Mesh Network-on-chip w/ 2 External I/F
- PRNG, DRL task scheduler, Top RISC Ctrlr.



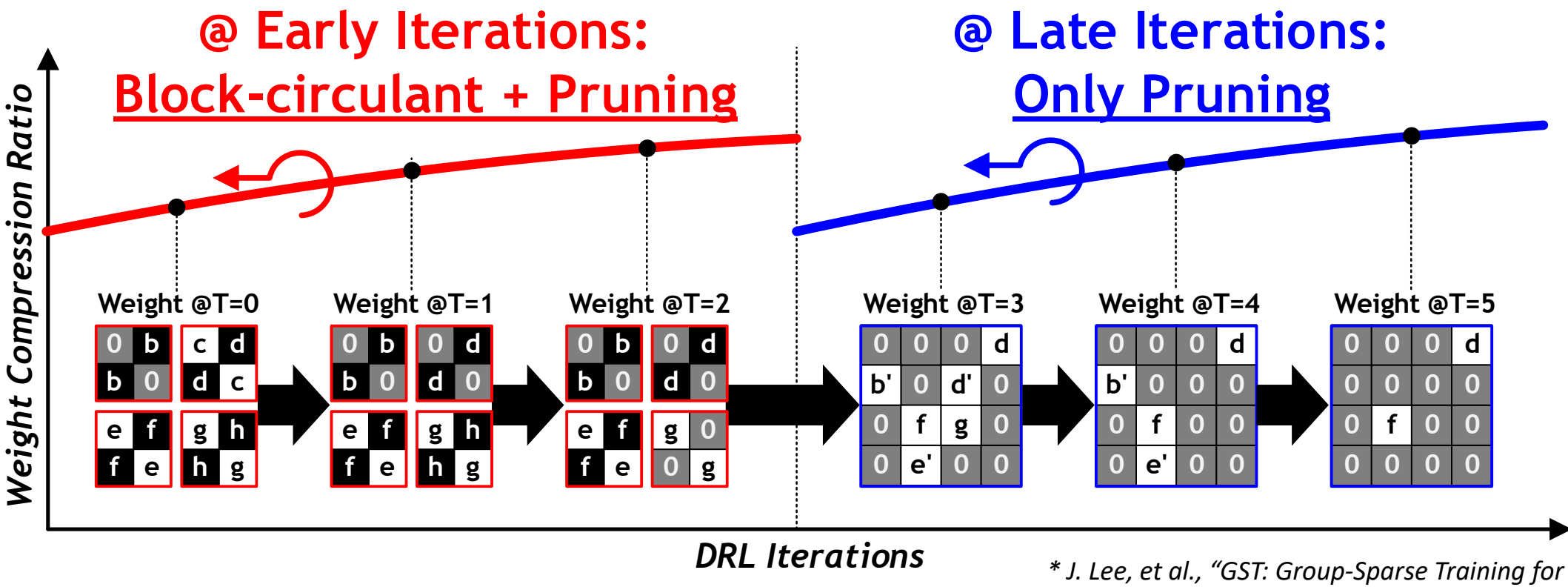
# Key Features for Energy-Efficiency

- 1. 24 Group-Sparse Training Core (GSTC)
- 2. Exponent Mean-Delta Encoding (EMDE)
- 3. Sparse Weight Transposer (SWT)



# Proposed Data-Level Compression

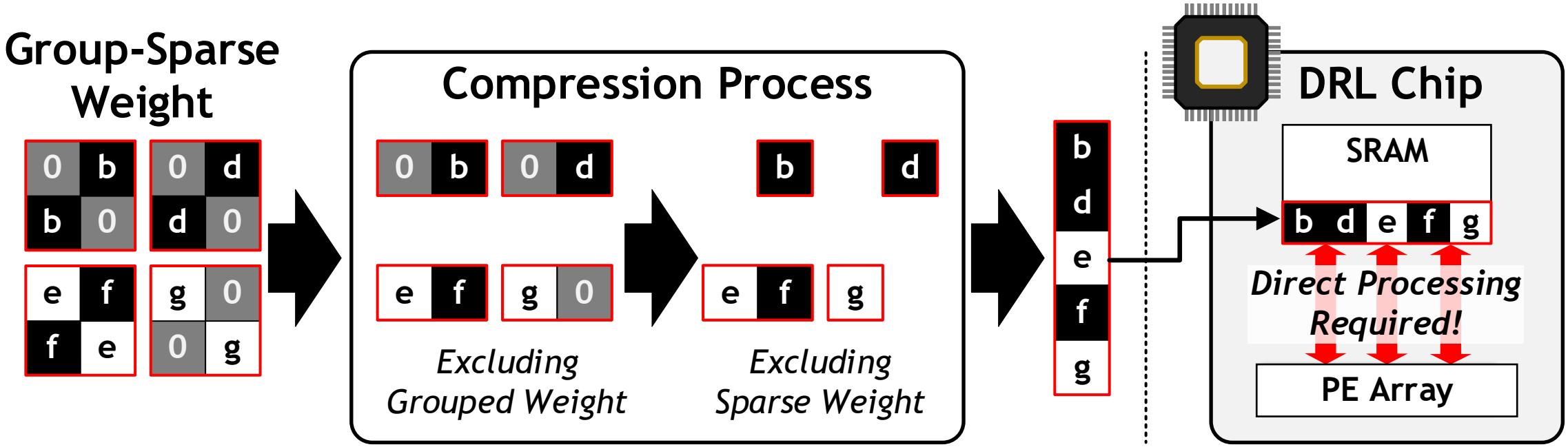
- **Group-Sparse Training (GST)***
  - ➔ ~41.5 %p higher weight compression ratio than iterative pruning ☺




* J. Lee, et al., "GST: Group-Sparse Training for Accelerating Deep Reinforcement Learning," arXiv:2101.09650, 2021.

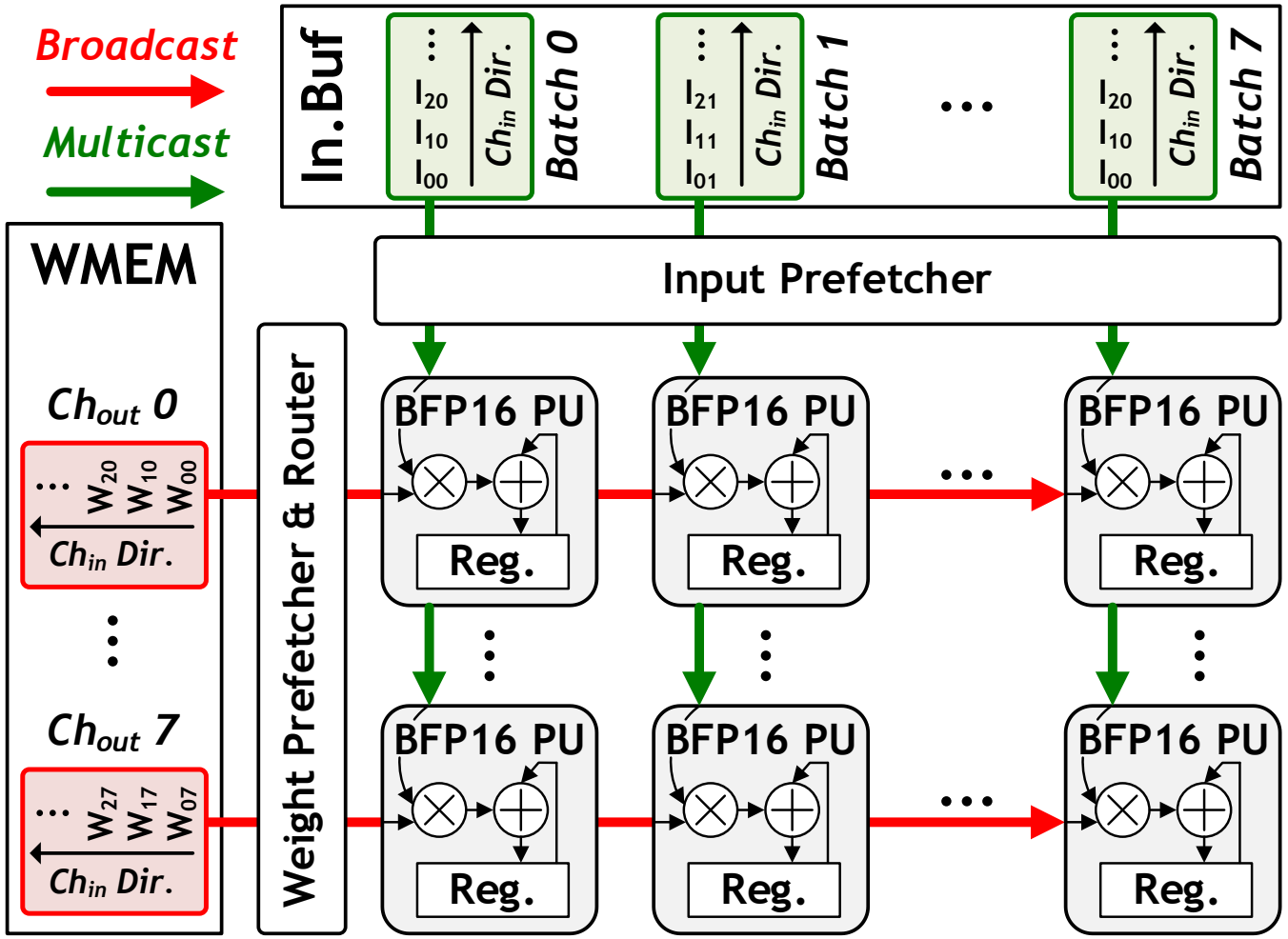
# Opportunity of GST

- **Direct Utilization of Compressed Weight is Required!**
  - Opportunity 1: ~×4 additional data reuse for grouped weight
  - Opportunity 2: ~×10 speed-up by skipping computation of sparse weight



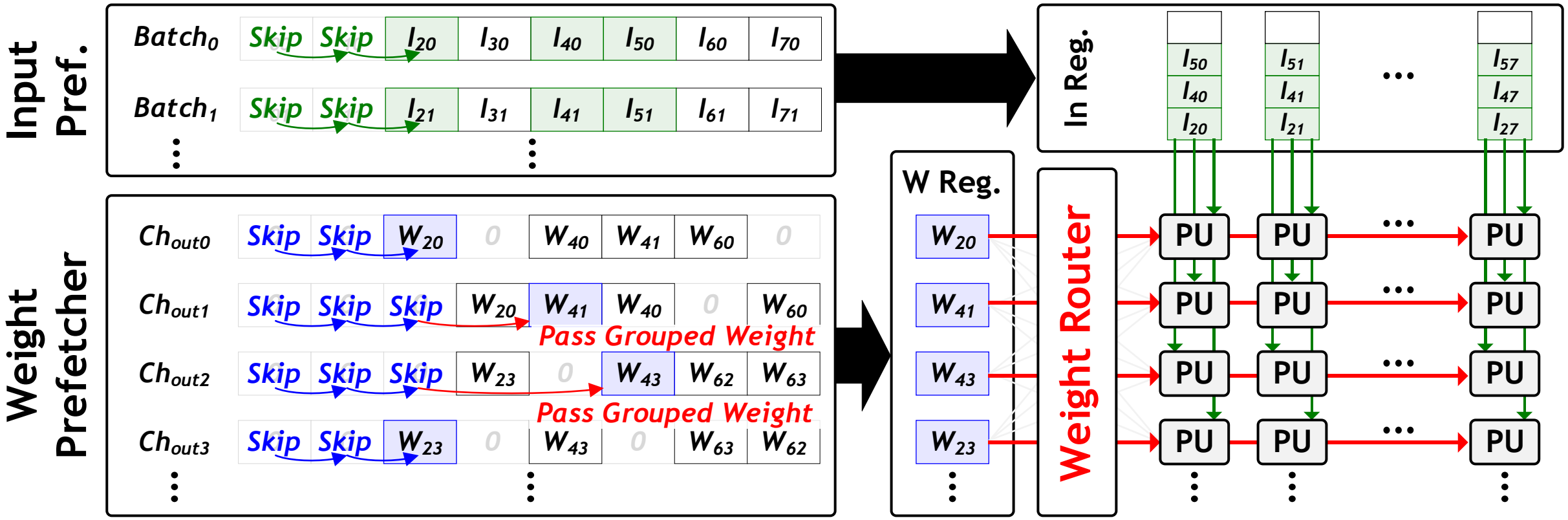
# Group-Sparse Training Core (GSTC)

- **Component**
  - 8×8 Bfloat16 PU array
  - Prefetchers & Router
- **Output Stationary**
- **Row-wise Allocation**
  - $Ch_{out}$ parallelism
  - Weight broadcasting
- **Collumn-wise Allocation**
  - Batch parallelism
  - Input multicasting



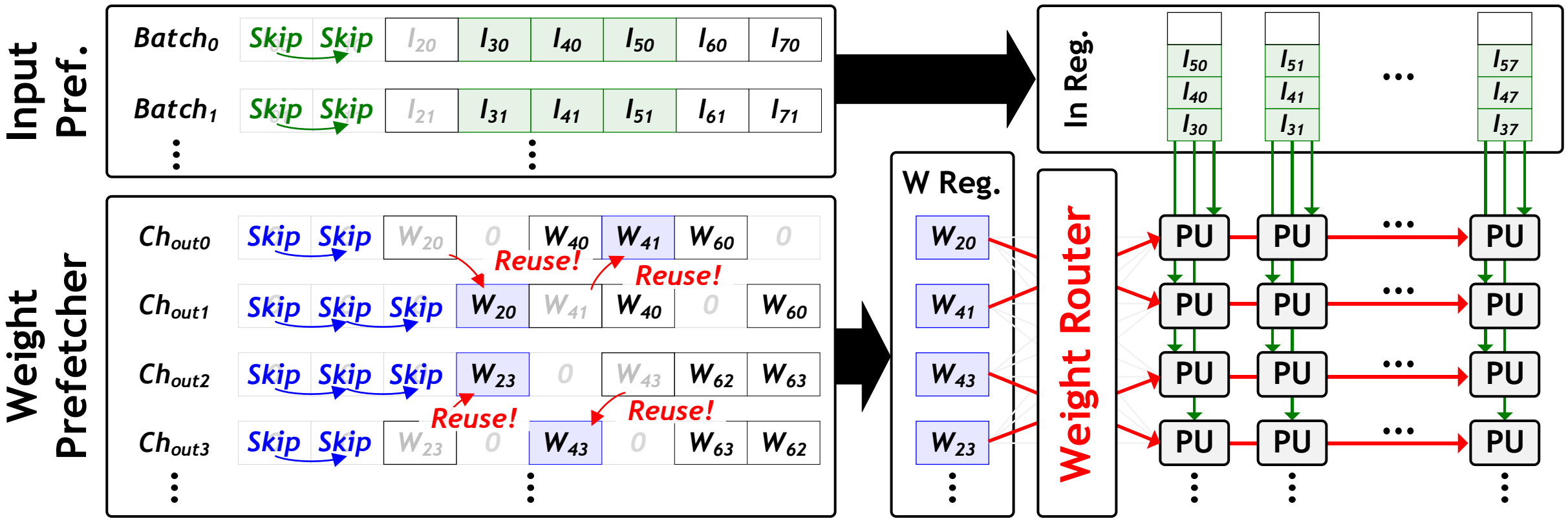
# GSTC Operations

- 1. Weight Zero Skipping @ Prefetchers (CLK 0)



# GSTC Operations

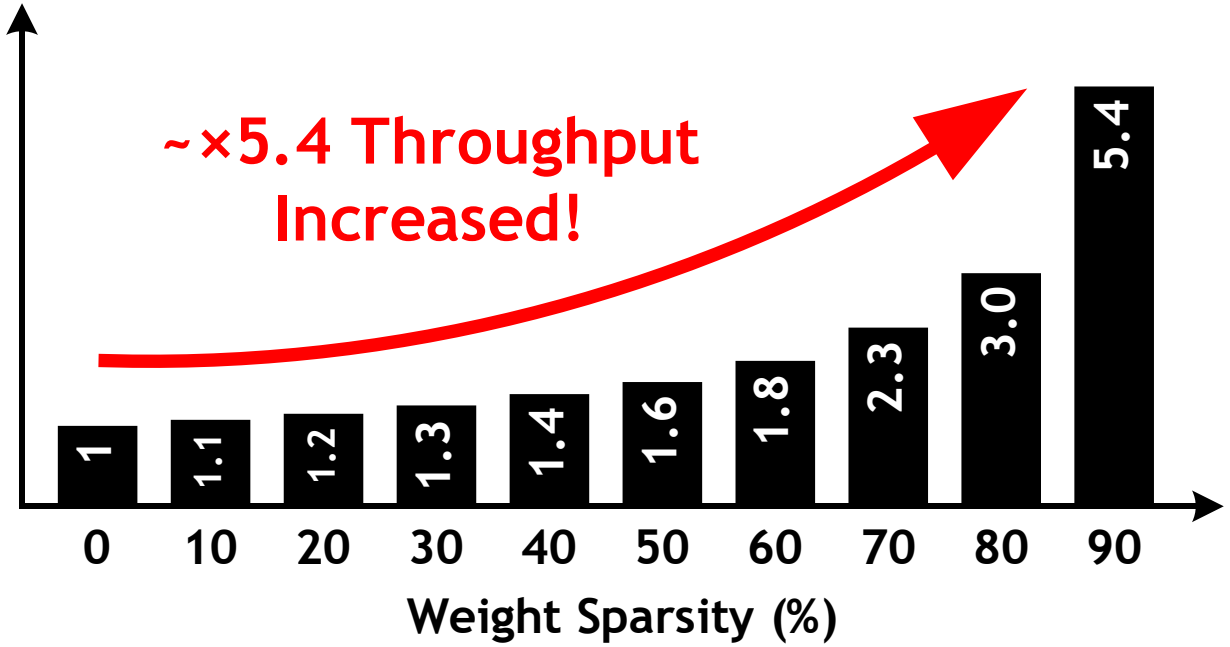
- 2. Weight Group Reuse @ Weight Routers (CLK 1)



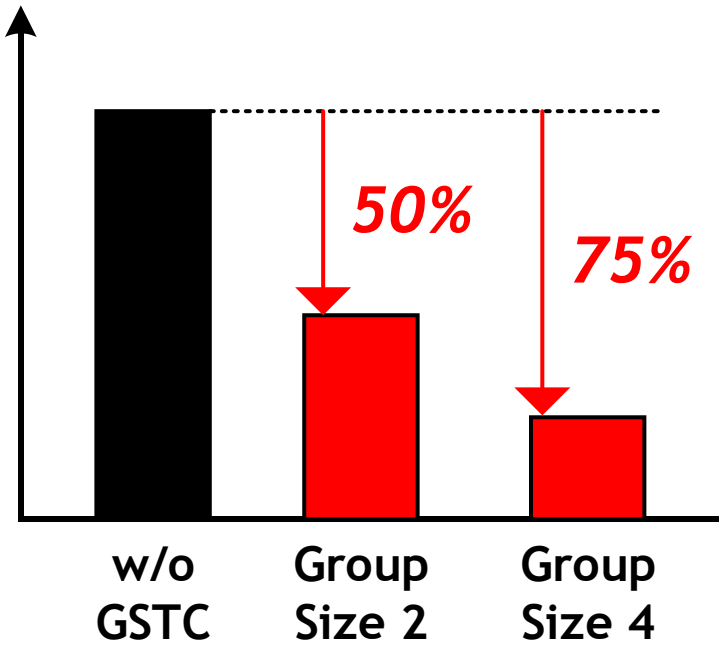
# Results of GSTC

- Throughput Increase with Weight Sparsity Exploitation
- WMEM Access Decrease with Weight Group Exploitation

<Normalized GSTC Throughput>

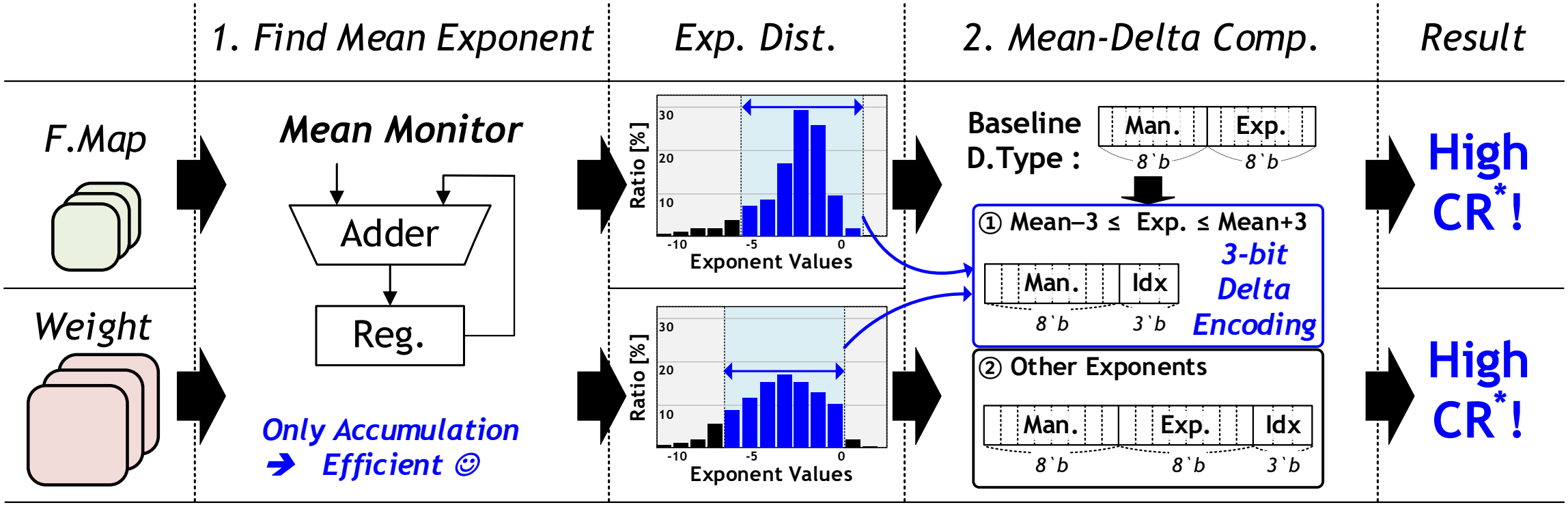



<Normalized WMEM Access>



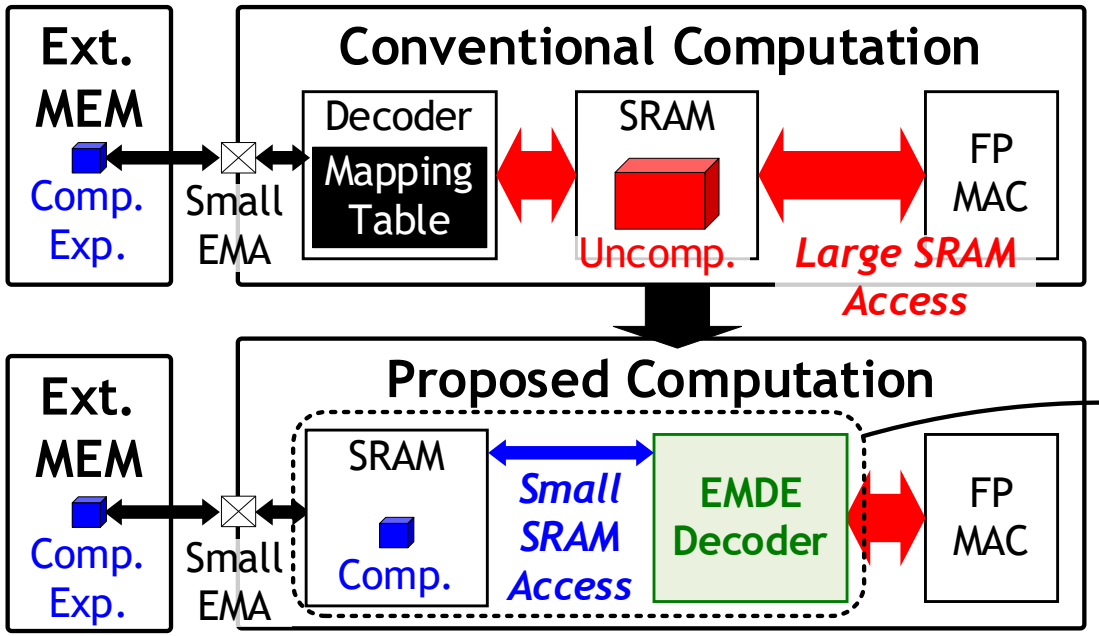
# Proposed Bit-level Compression

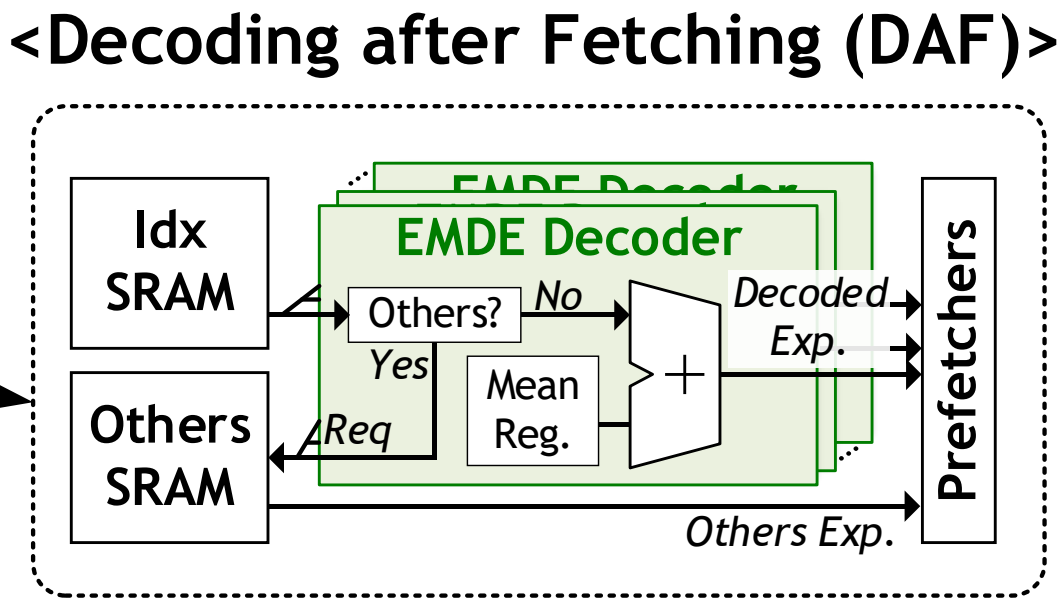
- **3-bit Exponent Mean-Delta Encoding (EMDE)**
  - Efficient searching, ×1.6 times high CR for both weight and f.map



# Decoder of the EMDE

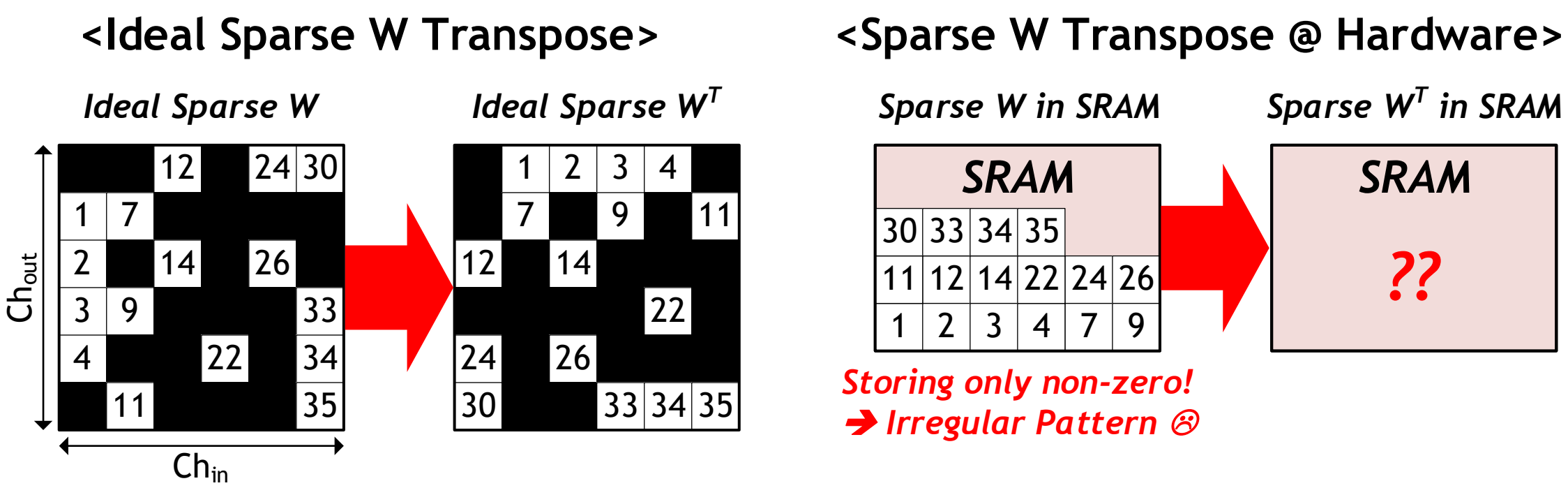
- **Decoding after Fetching (DAF)**
  - Simple adder based decoder ➔ DAF ➔ Memory access power 23.3 % ↓



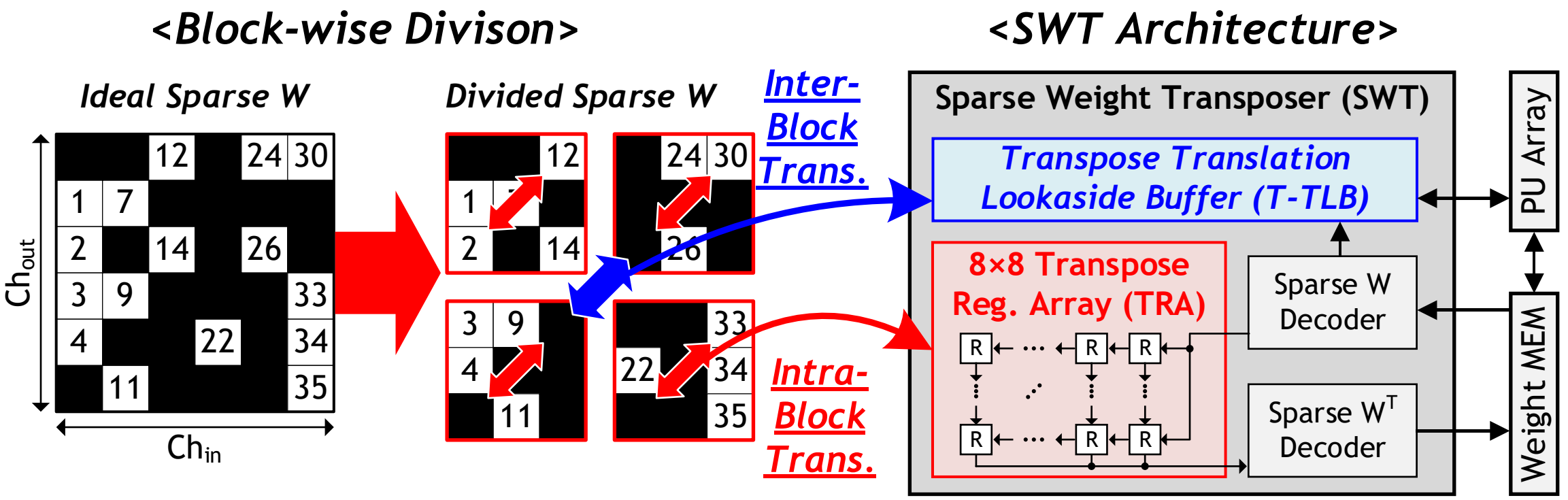
# Motivation: Training w/ Sparse Weight

- **Back-propagation Stage: Transposed Weight Required**
- **Irregularity Due to Storing Only Non-zero Values**



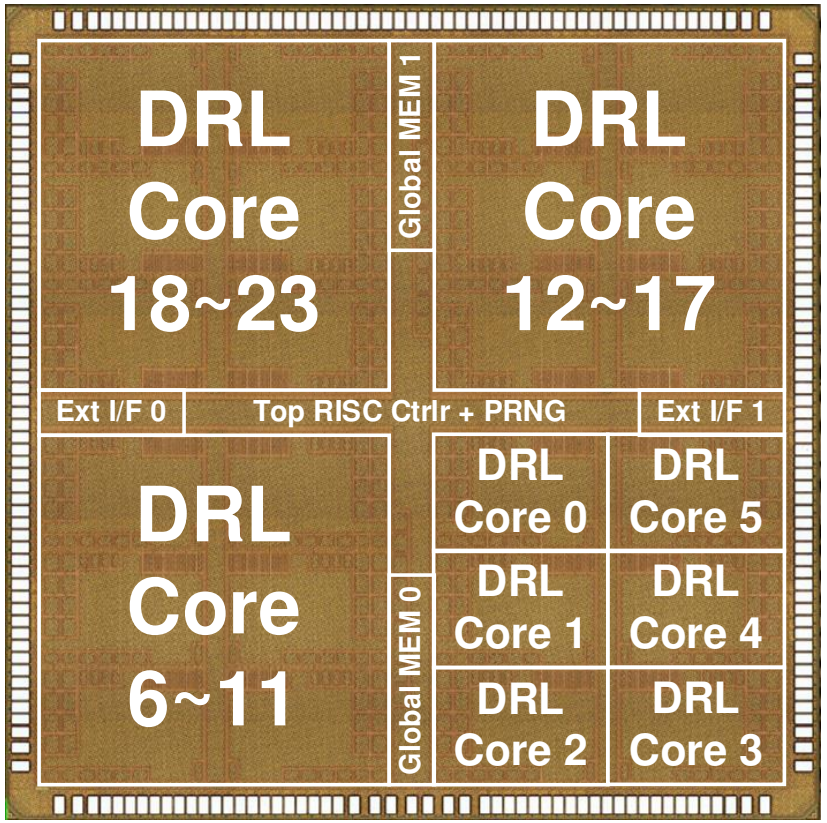
# Sparse Weight Transposer (SWT) Arch.

- **Block-wise Division of Sparse W + Hierarchical Transpose**
  - Intra-block transpose @ Transpose Register Array (TRA)
  - Inter-block transpose @ Transpose Translation Lookaside Buffer (T-TLB)



# Chip Performance and Summary

- ×2.5 Higher DRL Training Efficiency than Previous SOTA

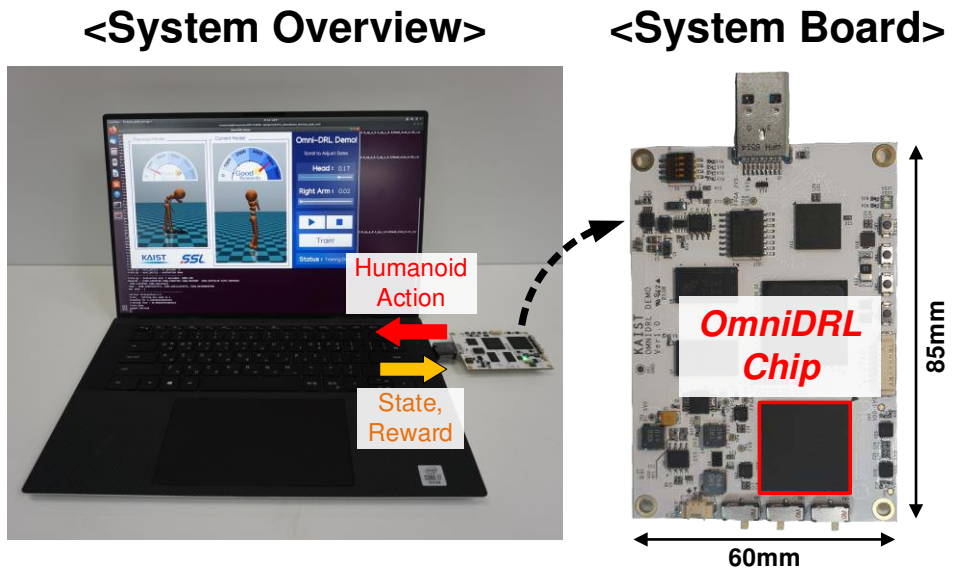


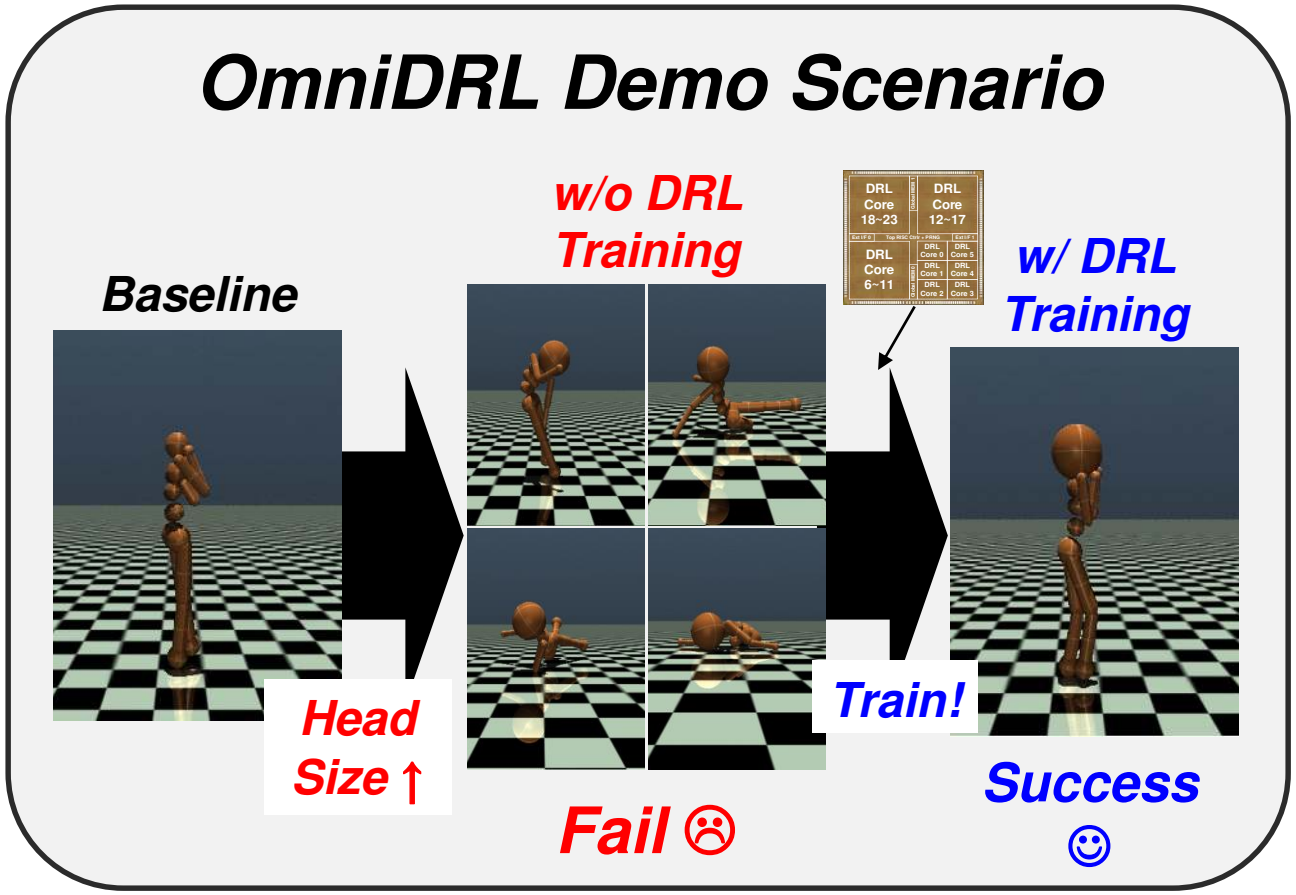

| | Chip Specifications |
|---|---|
| Technology | 28nm Logic CMOS |
| Die Area | 3.6 mm × 3.6 mm |
| SRAM | 1.5 MB |
| Supply Voltage | 0.68 V ~ 1.1 V |
| Frequency | ~ 250 MHz |
| Peak Performance | 0.77* – 4.18** TFLOPS @ 250MHz |
| Power Consumption [mW] | 3.1** – 3.9* @ (5MHz, 0.68V) |
| | 231.6** – 282.8* @ (250MHz, 1.1V) |
| Energy Efficiency [TFLOPS/W] | 4.2* – 29.3** @ (10MHz, 0.68V) |
| | 2.7* – 18.0** @ (250MHz, 1.1V) |

** = Weight Sparisty 0%, No Weight Group, Exp Comp 0%,*
*** = Weight Sparisty 90%, Weight Group 4, Exp Comp 90%*

# Demonstration System

- **Humanoid Adaptation to Sudden Environment Change***




* https://www.youtube.com/watch?v=qpnu1k8jqSQ

# Conclusion

- **1. For Low Memory Bandwidth**
  - Group-sparse training ➔ weight compression ~41.5 %p ↑
  - Exponent-mean-delta-encoding ➔ exponent compression ~1.6× ↑
  - World-first on-chip sparse weight transposer ➔ weight EMA 22.8 % ↓
- **2. For Low Memory Power Consumption**
  - Group-sparse training core ➔ Energy-efficiency ~4.4× ↑
  - Decoding-after-fetching ➔ SRAM power consumption 23.3 % ↓

**OmniDRL: A 29.3 TFLOPS/W DRL Training Processor for Mobile Applications**

# Thank You!

- **Questions? Feel Free to Contact Me!**
  - E-mail: juhyoung@kaist.ac.kr
  - LinkedIn: https://www.linkedin.com/in/juhyounglee
  - Zoom Meeting: https://us02web.zoom.us/j/3466650389?pwd=c1RWaXFTWGIjaVU1MINtcDhKaGg0dz09 (Password: HC_JHLEE)
  - System Demonstration Video Link: https://www.youtube.com/watch?v=qpnu1k8jqSQ