Rethinking What AI Hardware Needs to Remember

Daichi Fujiki

Associate Professor, AI Computing Research Unit (ArtIC)
Institute of Science Tokyo

Institute of SCIENCE TOKYO

DoSSA-8 @ MICRO2026

About Myself

Institute of SCIENCE TOKYO

- 2016～2022 PhD student @ University of Michigan



Reetu (My advisor)



Me

- 2022～2023 Assistant Professor @ Keio University
  - In Yokohama
  - PIM followup work
- 2023～ Associate Proferssor @ Institute of Science Tokyo
  - Artificially Intelligent Computing Research Unit (ArtIC)
  - In Yokohama

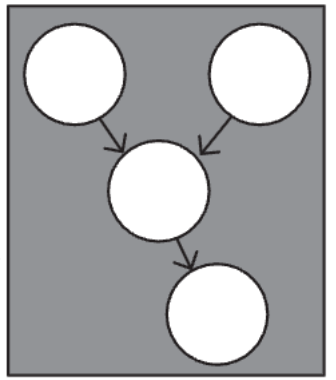


Why data representation matters

Institute of SCIENCE TOKYO

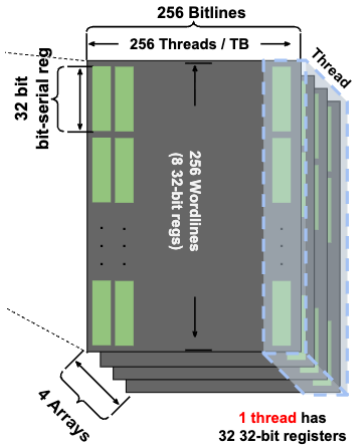
(My projects back in Michigan)

PIM for general applications



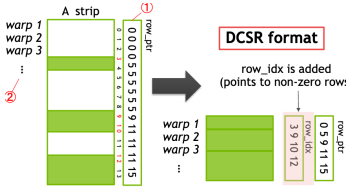
Data Flow Graph

ASPLOS 2018



Vector representation

MICRO 2019

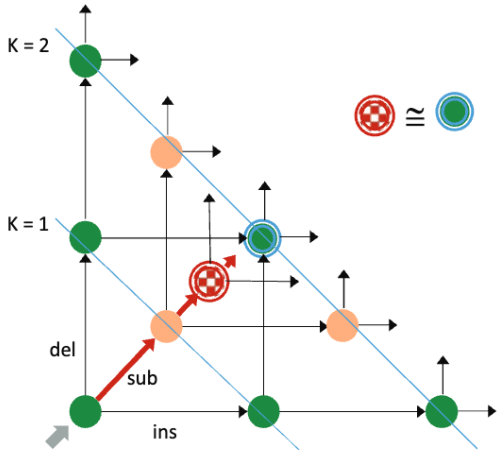


DCSR format

SC 2019

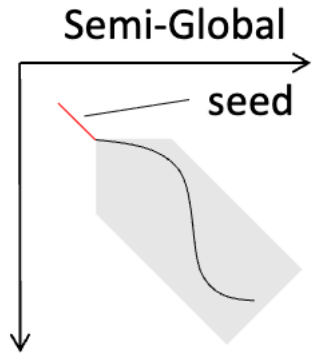
Facilitates implementing restrictions in PIM

Genomics accelerator



Universal Automata

ISCA 2018



DP Array

MICRO 2020

Facilitates implementing efficient dataflow

Why data representation matters

Institute of SCIENCE TOKYO

AI • ERIC SCHMIDT

Ex-Google CEO says superintelligence is tech's holy grail—but the U.S. is ill-equipped to address its 'natural limit'

By Dave Smith
Former Editor, U.S. News
July 18, 2025, 7:05 AM ET

Follow



Eric Schmidt, billionaire and co-founder of Schmidt Sciences, at the Raise summit in Paris, France, on Tuesday,

"AI's natural limit is electricity, not chips," Schmidt

"It's totally fair to say that AI is going to need a lot more energy," Altman

Altman Says We Still Don't Know How Much Energy AI W Need

Jan. 17, 2024, 3:25 AM GMT+9

Brad Stone
Bloomberg News

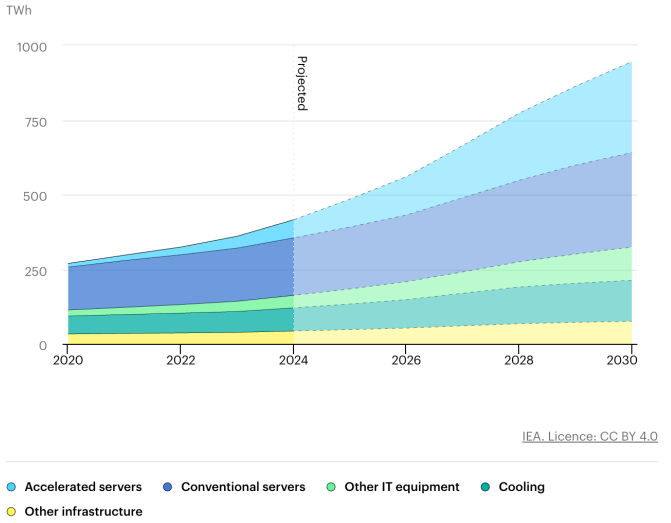
Josh Saul
Bloomberg News

- OpenAI chief says AI will force clean energy investment
- AI revolution has been accompanied by surging power demand

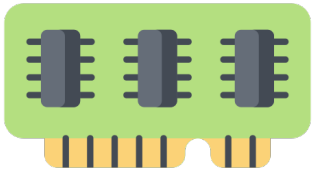
OpenAI Chief Executive Officer Sam Altman said massive demand for electricity to power artificial intelligence will require a breakthrough in clean energy generation, listing fusion or a combination of solar power and storage as two likely possibilit

"It's totally fair to say that AI is going to need a lot more energy," Altman said in an interview at Bloomberg House at the Wo Economic Forum in Davos on Tuesday. "That will force us to invest more in the technologies that can deliver this, none of v are the ones that are burning carbon."

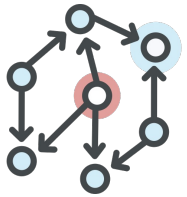




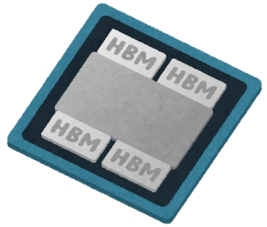
Weight Traffic / Limited Sparsity Configuration



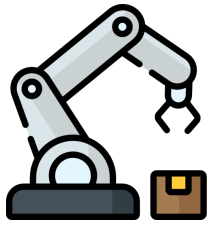
Sparse Irregular Memory Access



Activation / Growing Context



Repeated generation in Diffusion models

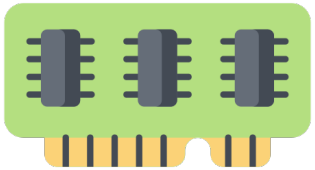


"Shuffling data is what costs energy, not the computation itself." Yann LeCun

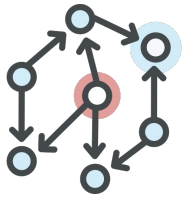
Why data representation matters

Institute of SCIENCE TOKYO

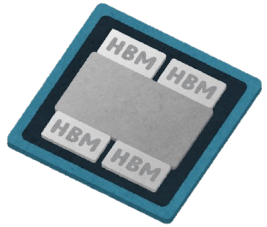
Weight Traffic /
Limited Sparsity Configuration



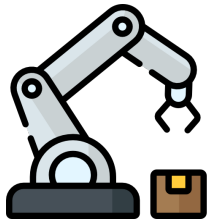
Sparse Irregular
Memory Access



Activation /
Growing Context



Repeated generation
in Diffusion models



Good Data Representation + Architecture

On-chip weight generation
with SLT

Access localization
with VQ

KV Compression
with PQ on PIM

Output Memoization
with Fuzzy Cache

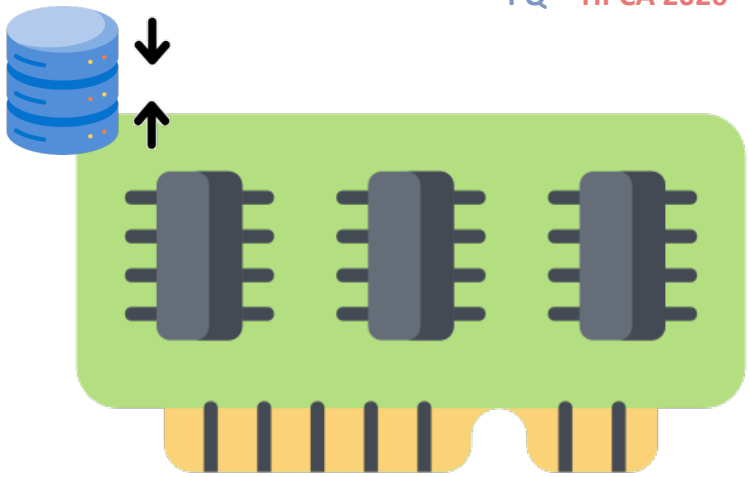
Data representations shapes what we compute

Institute of SCIENCE TOKYO

③ Capacity-Wall Problem of LLM on PIM

addressed by PIM-aware KV compression

PQ HPCA 2026

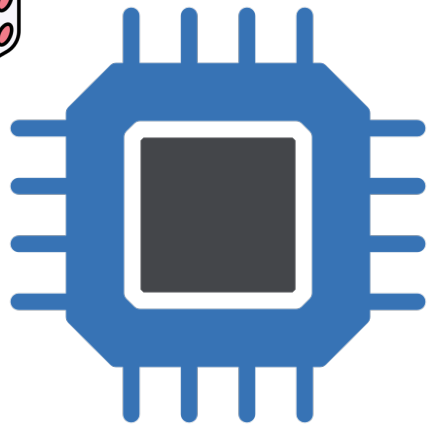


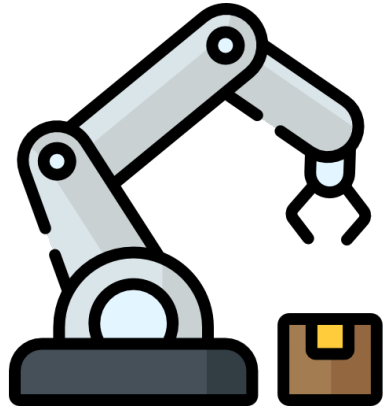
① On-Chip Weight Regeneration

through Strong Lottery Ticket Theory

SLT AAAI 2026 ISSCC 2022



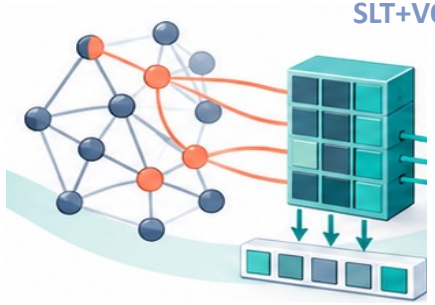




② Sparse Access Localization for GNNs

through X-partition message quantization

SLT+VQ ISCA 2025



④ Memory-Centric Approach for Physical AI

Action caching and local action generation

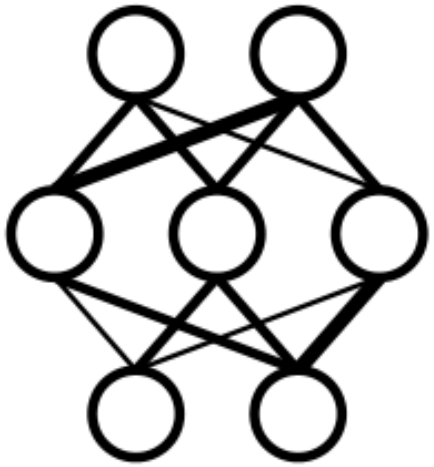
AC



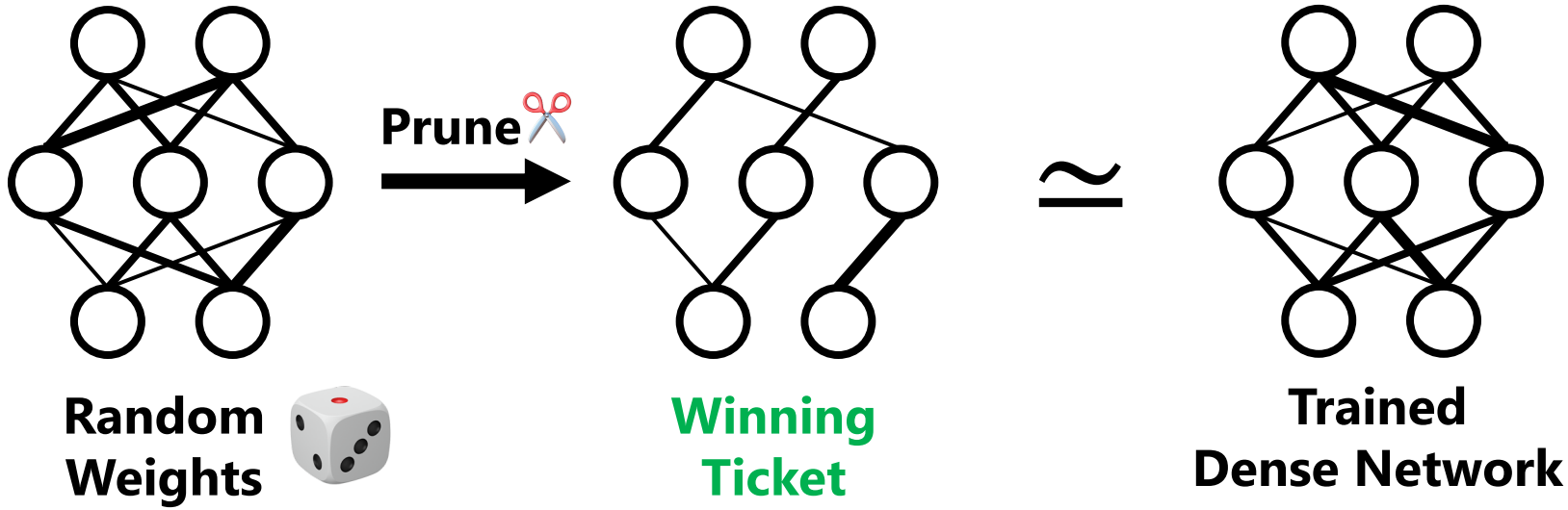
Can Random Weights Work for a Neural Network?

Freeze every weight at its random initial value.
Then, pick up edges.

Can the network still work?



Strong Lottery Tickets



Conjecture: Randomly-initialized networks contain an accurate subnetwork that, at initialization, match the performance of the full network.

Random weights stay fixed.
A learned mask (supermask) selects the subnetwork.

Random Weights Can Be Regenerated

Conventional network

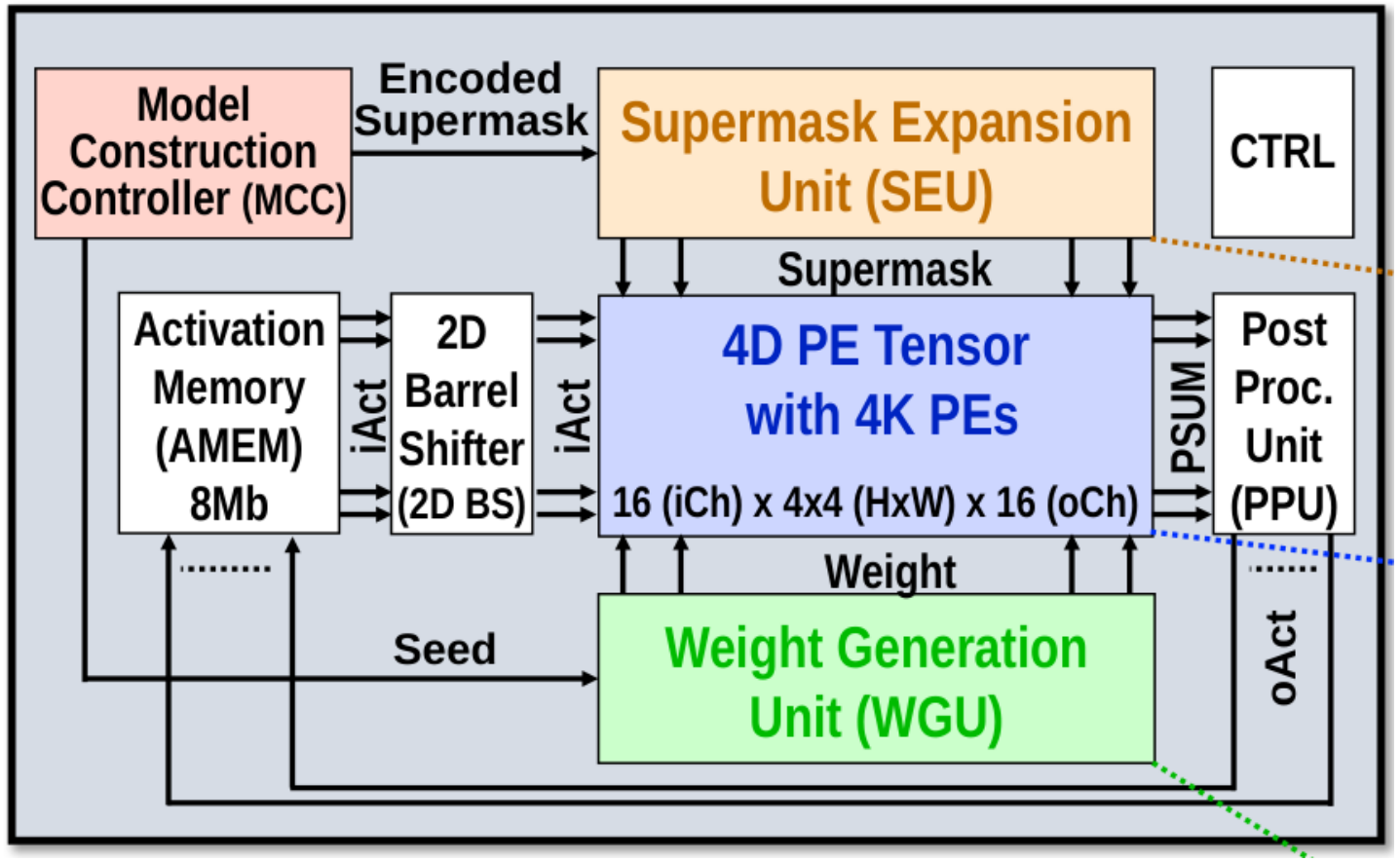


Strong lottery ticket



Weights can be regenerated locally with Random Number Generators (RNG).

Hiddenite: On-Chip Model Construction with SLT



Weight generation

Reproduce random weights from compact seeds.

Supermask expansion

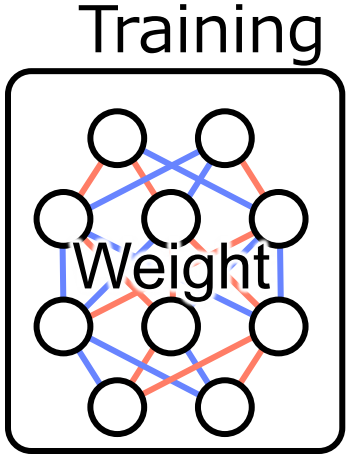
Decode connectivity applying supermasks.

ISSCC 2022 / SESSION 15 / ML PROCESSORS / 15.4

15.4 Hiddenite: 4K-PE Hidden Network Inference 4D-Tensor Engine Exploiting On-Chip Model Construction Achieving 34.8-to-16.0TOPS/W for CIFAR-100 and ImageNet

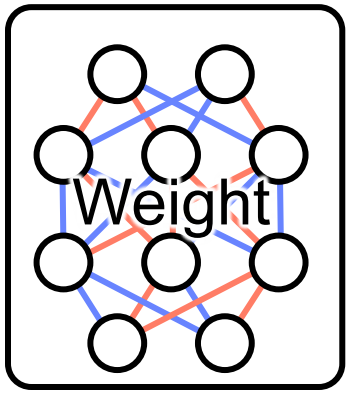
The MCC also delivers a Zero-Run-Length (ZRL) compressed supermask bitstream to the ZRL Memory (ZMEM; Fig. 15.4.3). Then, sixteen pairs of FIFO and ZRL decoder expand 16-oCh supermasks in parallel and store them in the Supermask Memory (SMEM). This SMEM is double-buffered to hide the SEU latency behind the PE Tensor

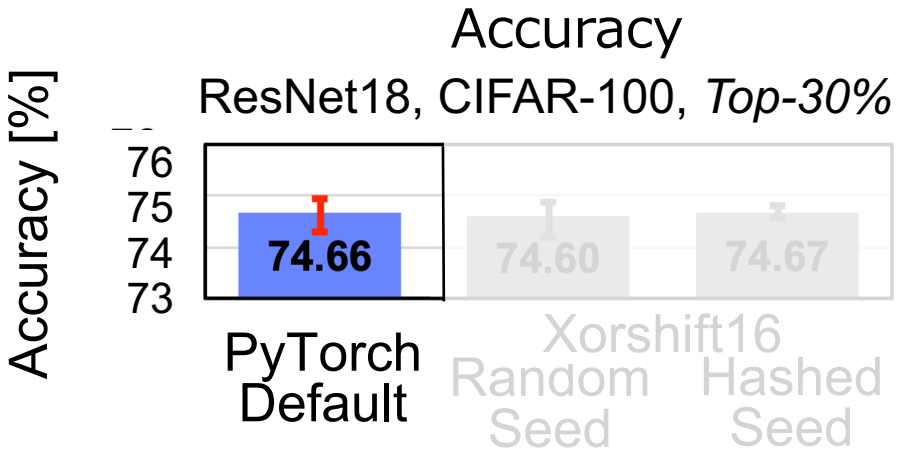
Hiddenite: On-Chip Model Construction with SLT

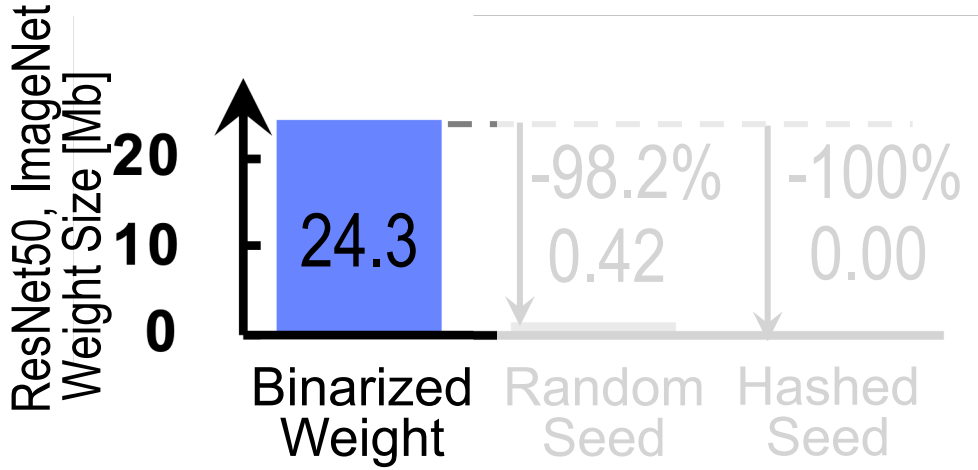


Update through training

Store and load

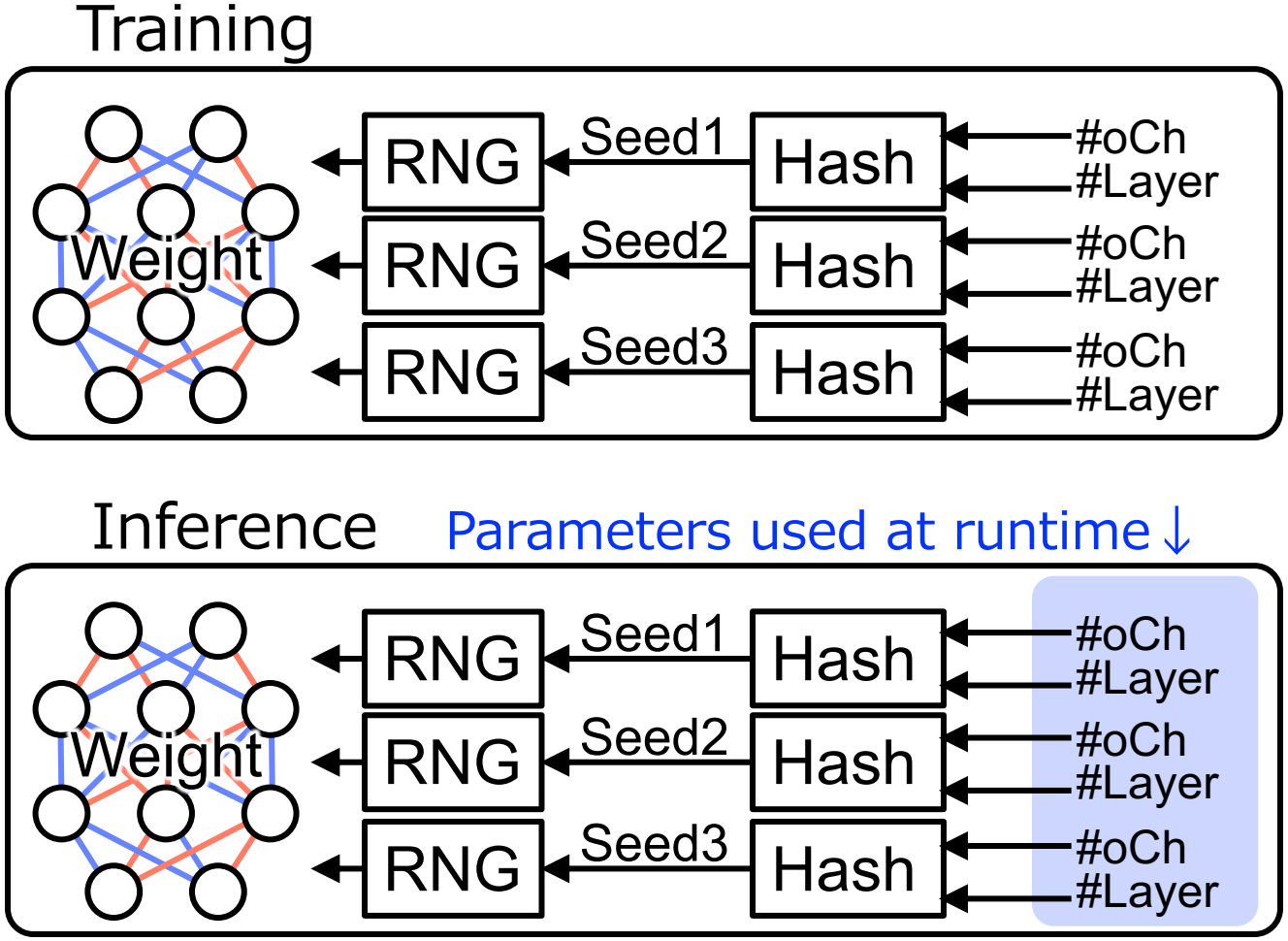


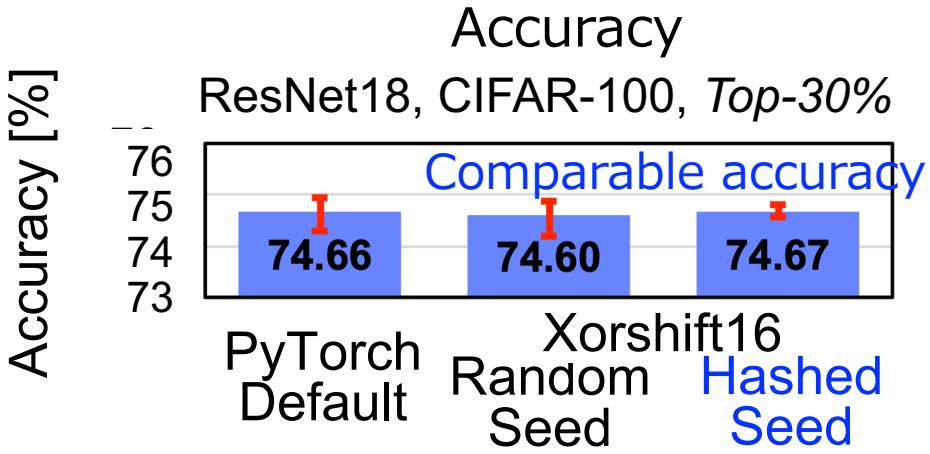


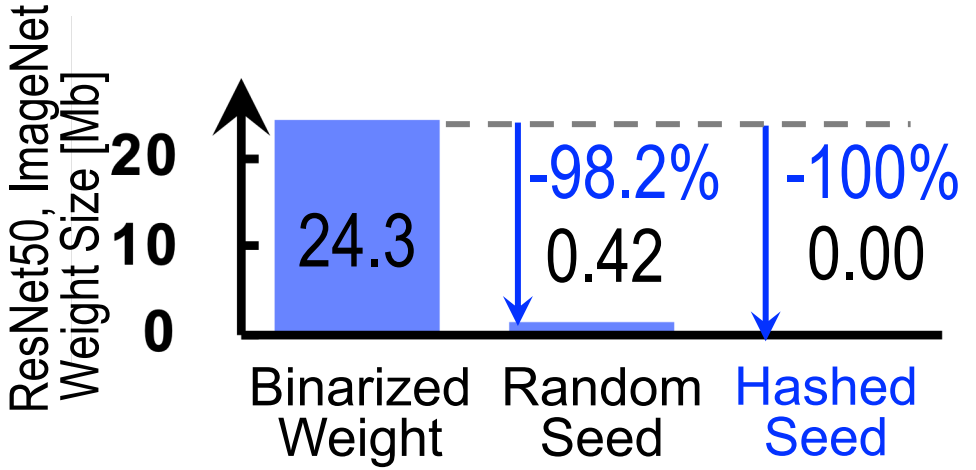


Conventional NNs require all weights to be loaded

Hiddenite: On-Chip Model Construction with SLT





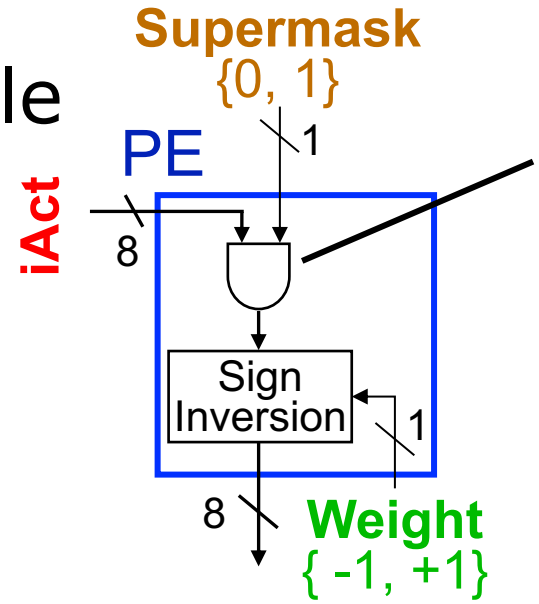


Eliminate all weight accesses without accuracy loss

Hiddenite: On-Chip Model Construction with SLT

- Computation is extremely simple

PSUM
= Σ if(Supermask){iAct x Weight}
= Σ{(iAct and Supermask) x Weight}
8 bit 1 bit 1 bit

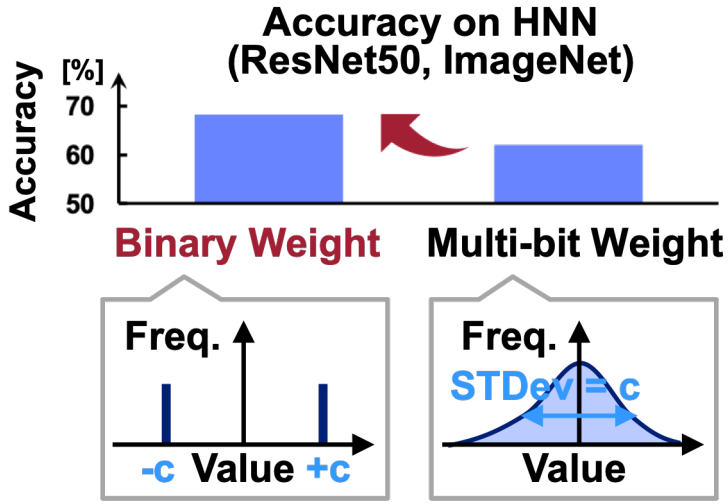


Supermask-based data gating
↓
☺ **Lower power**

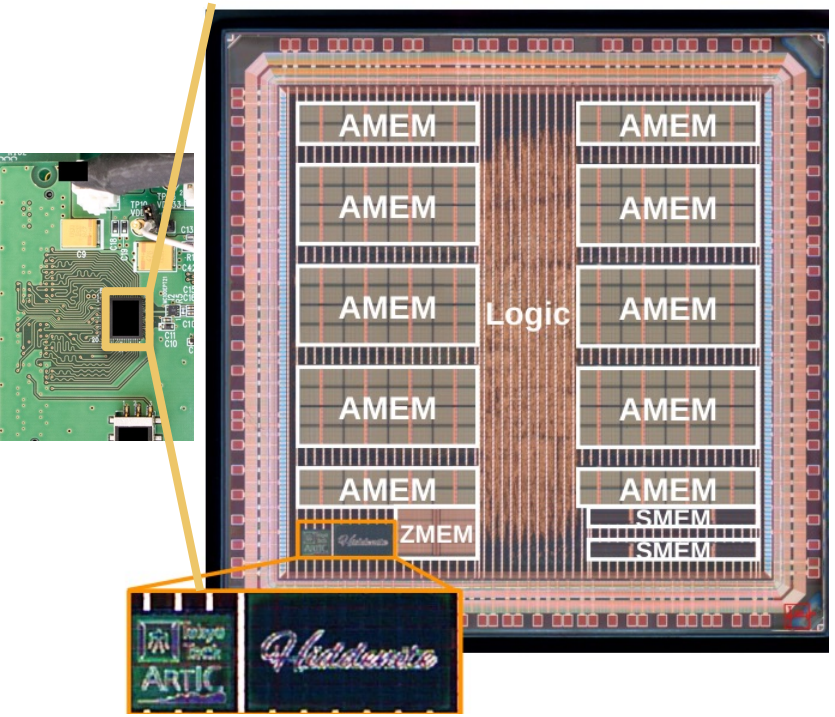
→ Area efficient PE (Processing Element)

- Binary weights {-1, +1} show better accuracy than multi-bit weights

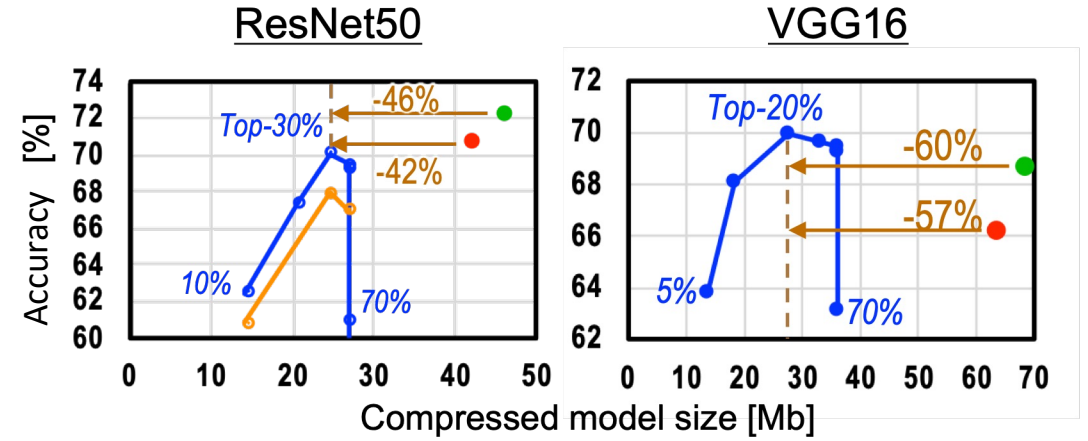
[V. Ramanujan+, CVPR2020]

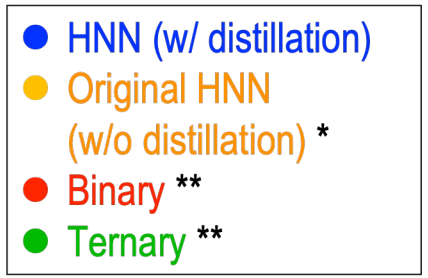


Hiddenite in Silicon

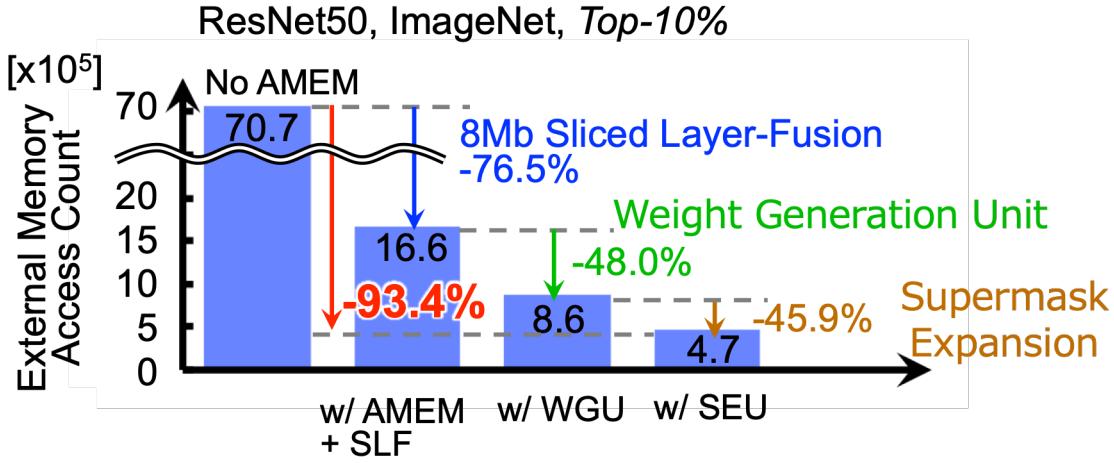


TSMC 40 nm CMOS





* [V. Ramanujan+, CVPR2020]
** [J. Faraone+, CVPR 2018]



On-chip Model Construction

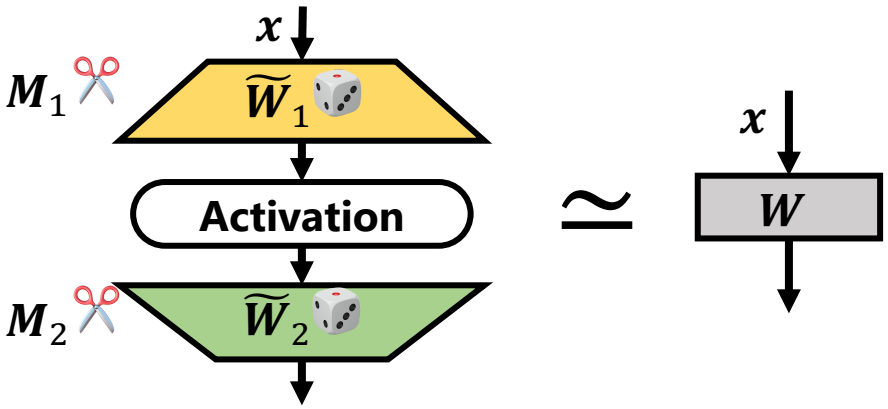
93.4% less memory accesses

16.0–18.2 TOPS/W on ResNet-50 at 0.77 V

Do Strong Lottery Tickets Exist in Attention?

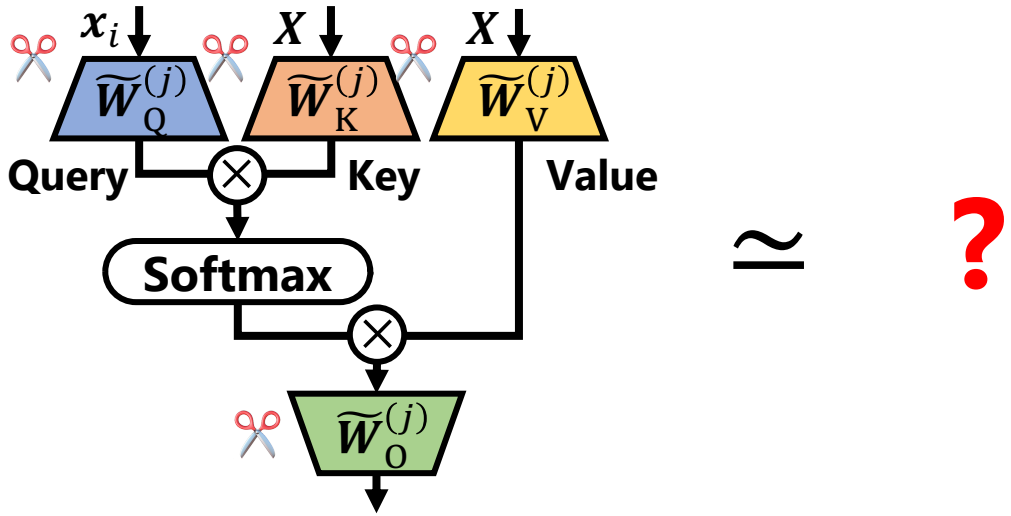
Prior SLTH Theory:

- **Scope:**
  MLPs, CNNs, Equivariant NNs [3-5]
- **Key proof tool:**
  Two-layers-for-one approximation



Toward Transformers:

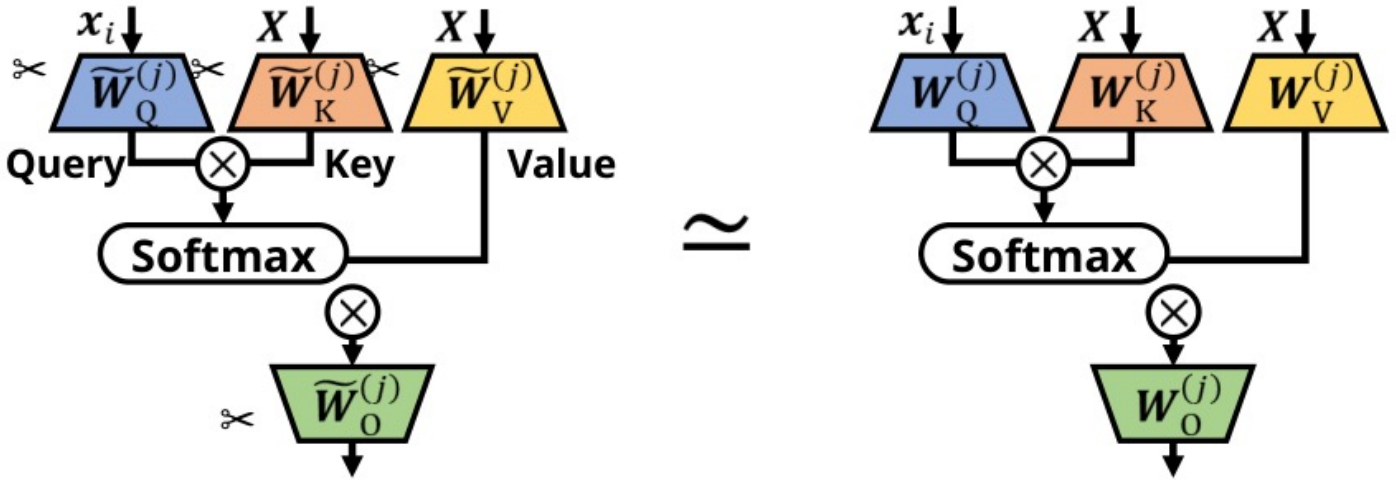
- **New component:**
  Multi-Head Attention (MHA)
- **Gap:**
  Pairwise query-key dot products (token-token interactions)



Multi-Head Attention (MHA)

Strong Lottery Tickets in Multi-Head Attention

A sufficiently wide random MHA contains a subnetwork
that approximates a target MHA under the theorem's assumptions.



Required Q/K/V width is independent of input sequence length and target Q/K/V width.

Agenda

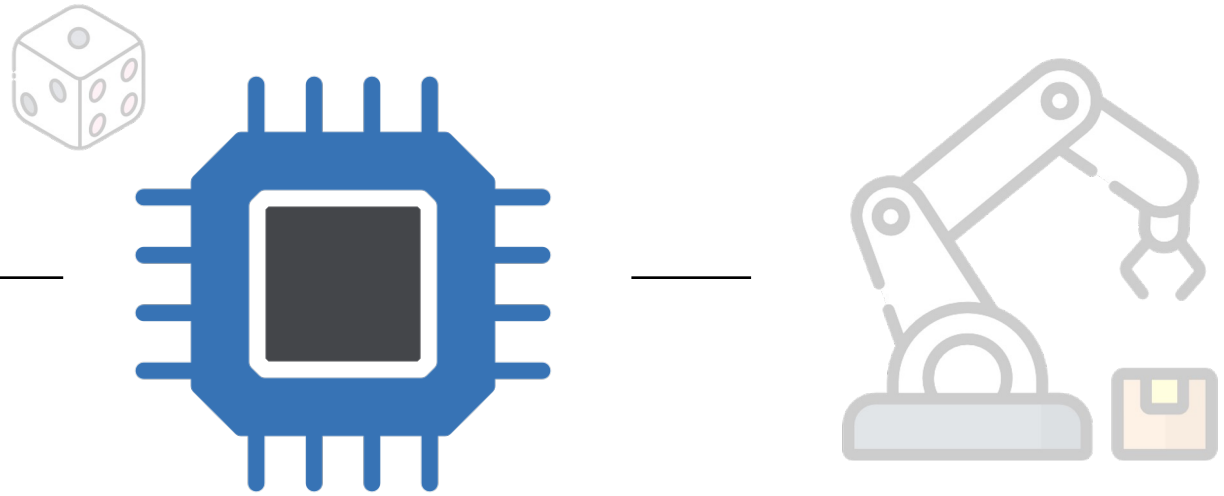
Institute of SCIENCE TOKYO

③ Capacity-Wall Problem of LLM on PIM
addressed by PIM-aware KV compression

PQ HPCA 2026

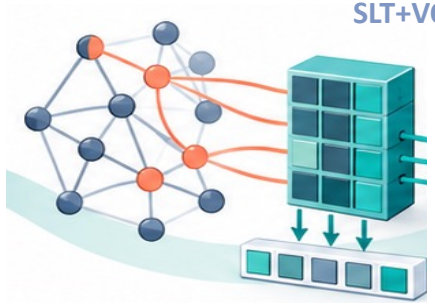
① On-Chip Weight Regeneration
through Strong Lottery Ticket Theory

SLT AAAI 2026 ISSCC 2022



② Sparse Access Localization for GNNs
through X-partition message quantization

SLT+VQ ISCA 2025



BingoGCN: Towards Scalable and Efficient GNN Acceleration with Fine-Grained Partitioning and SLT

Jiale Yan
yan.jiale@artic.iir.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Hiroaki Ito
ito.hiroaki@artic.iir.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Yuta Nagahara
nagahara.yuta@artic.iir.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Kazushi Kawamura
kawamura@artic.iir.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Masato Motomura
motomura@artic.iir.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

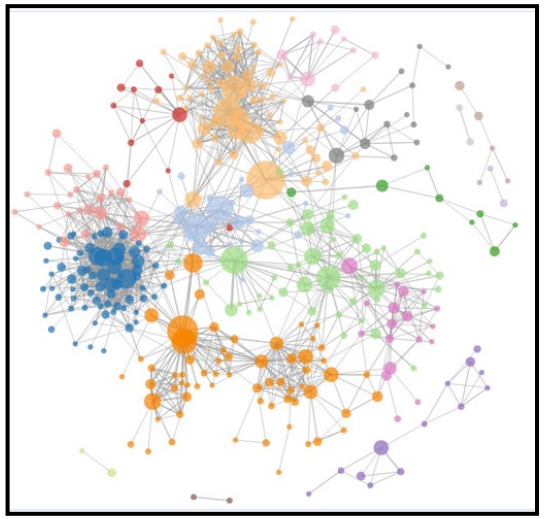
Thiem Van Chu
thiem@ict.eng.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Daichi Fujiki
dfujiki@artic.iir.isct.ac.jp
Institute of Science Tokyo
Tokyo, Japan

GNN Computation and Memory Access

GCNs handle irregular graph data and are widely used in multi-level applications.

Node-level Task



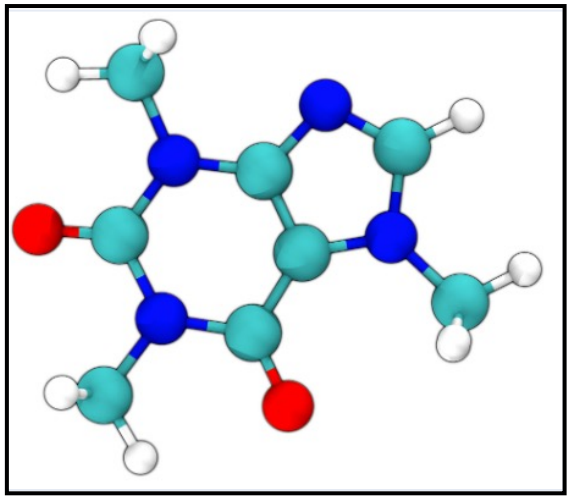
Academic network

Link-level Task



Social network

Graph-level Task



Molecular graph

Achieving real-time, large-scale GNN inference remains a significant challenge…

Challenges of GCNs

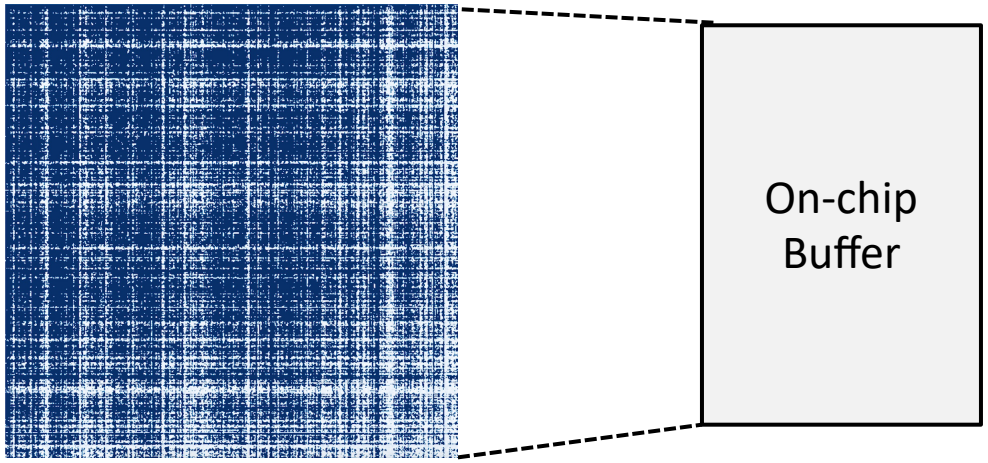
Institute of SCIENCE TOKYO

1. Excessive memory and communication demands

Prior work [FlowGNN HPCA'23, etc.]

assumes **small graphs** that entirely

fit in onchip buffers



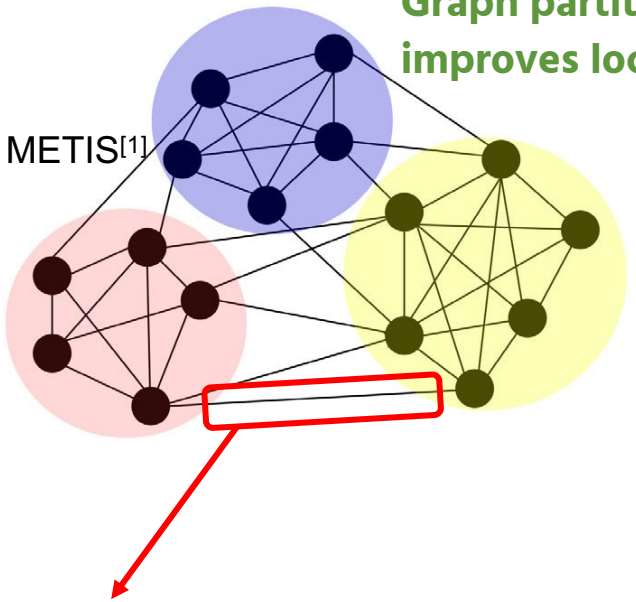
👎 **Not scalable!**

Challenges of GCNs

1. Excessive memory and communication demands

Solution

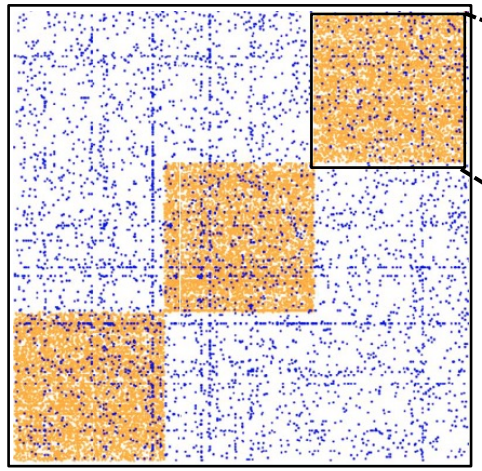
Graph partitioning improves locality!

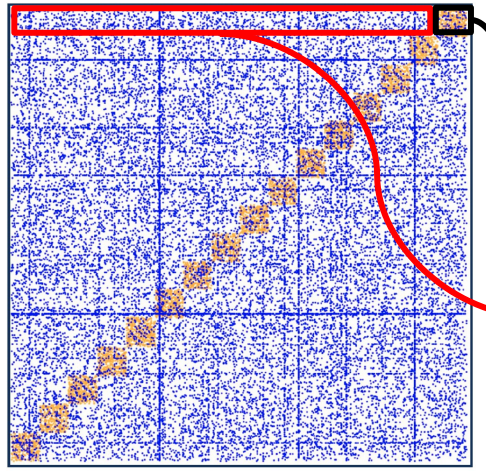


*Inter-partition edges grow **rapidly!***

#P = 16

OGBN-Arxiv (#P = 3)





On-chip Buffer

Off-chip DRAM

Buffer size ∝ (#N /#P)

👍 **Locality**

👍 **Area Efficiency**

Partitioning Dilemma!

👎 **Random off-chip access**

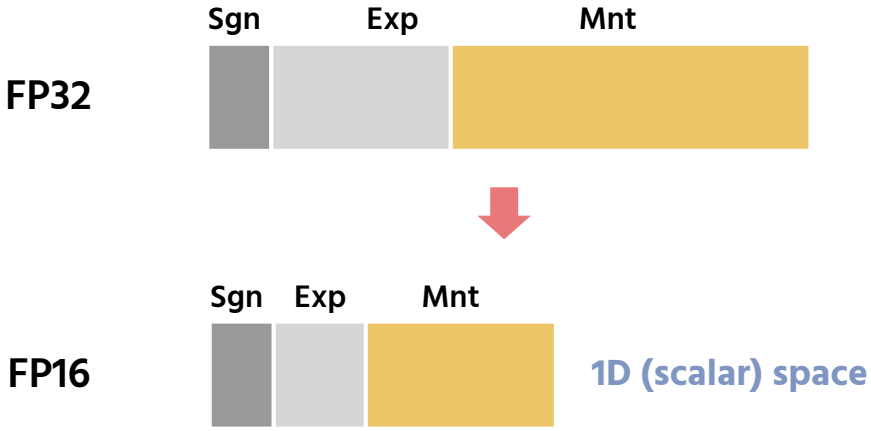
👎 **#Parts not scalable**

[1] George Karypis and Vipin Kumar. 1997. METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices.

Cross-Partition Message Quantization (CMQ)

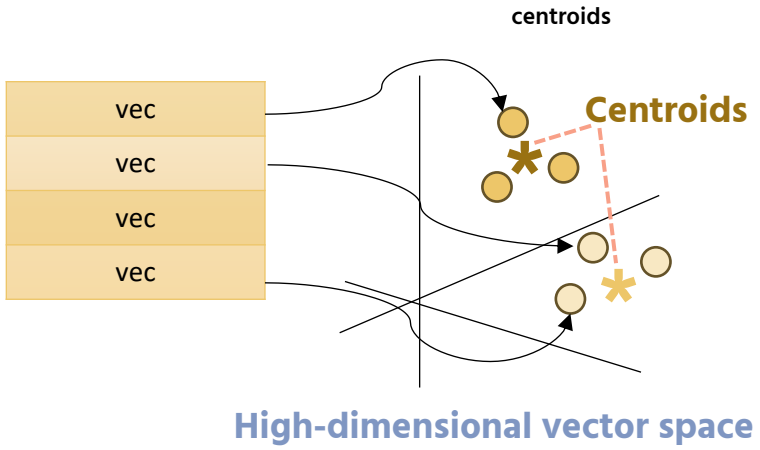
Institute of SCIENCE TOKYO

Traditional Quantization



- Quantizes value-by-value
- Low compressibility

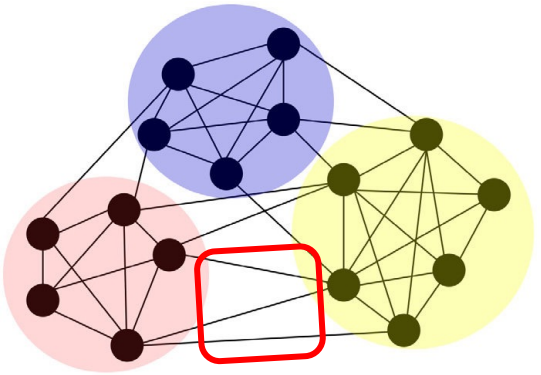
Vector Quantization

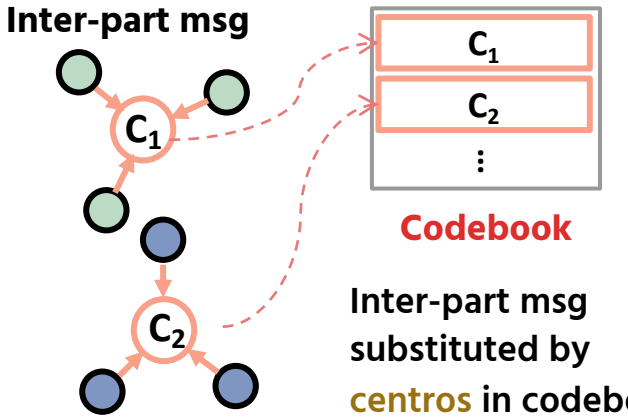


- Incorporates vector semantics
- Quantizes relative vector positions by **centroids**
- Context aware & High compressibility!

CMQ

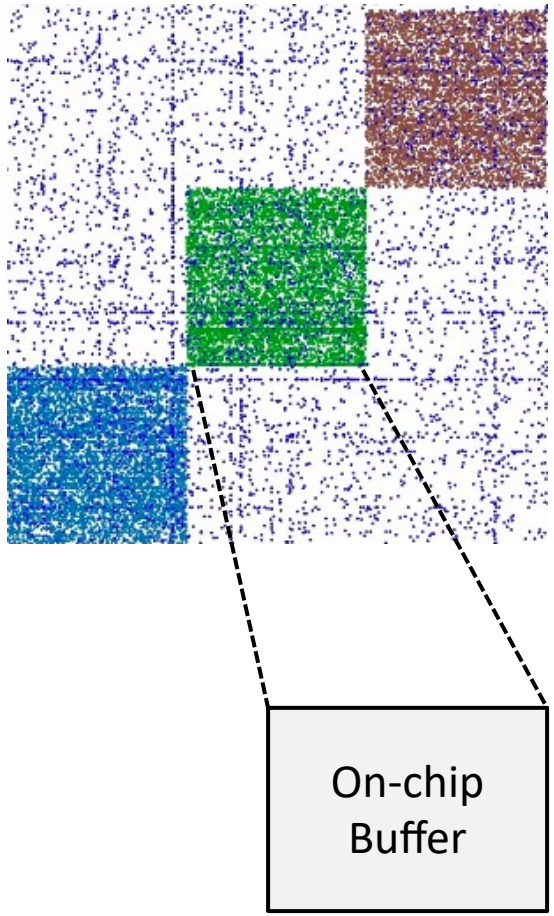
Cross-Partition **M**essage **Q**uantization





Inter-part msg substituted by **centros** in codebook

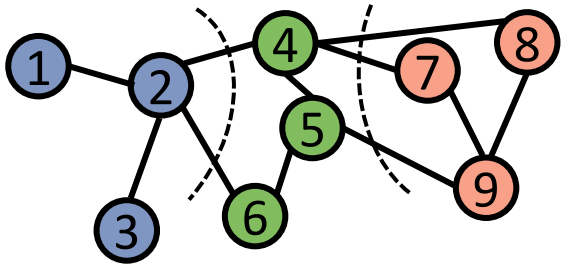
Cross-Partition Message Quantization (CMQ)



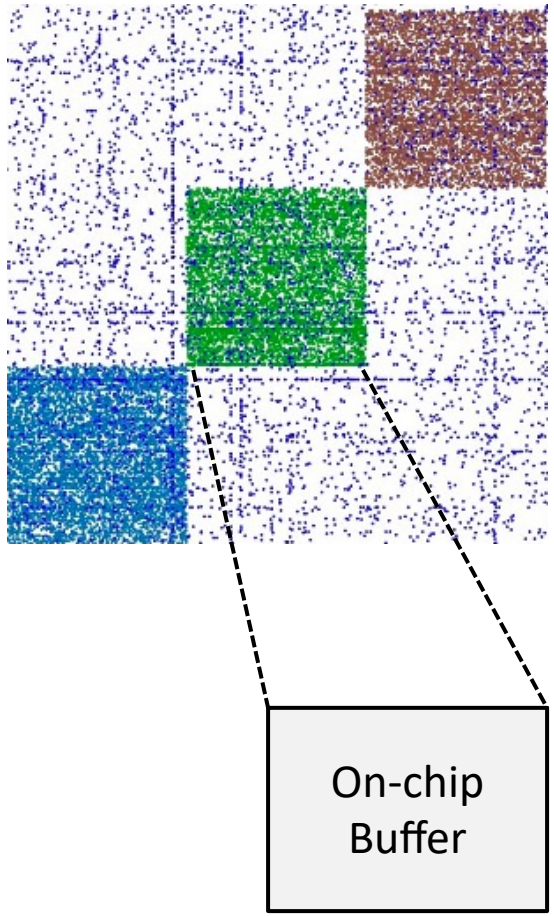
In the beginning of each layer…

Query the Code Books

3 Parts

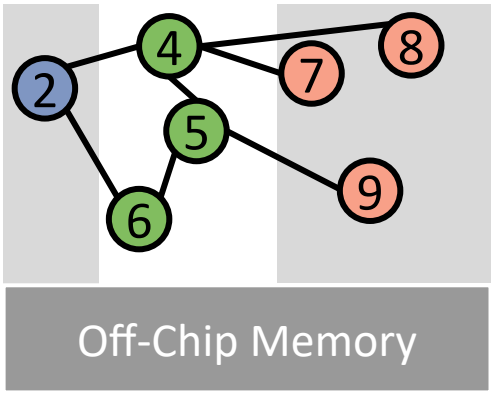


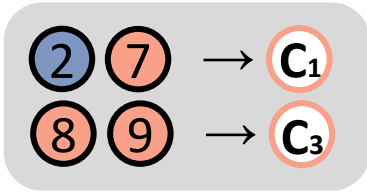
Cross-Partition Message Quantization (CMQ)



In the beginning of each layer…
Query the Code Books

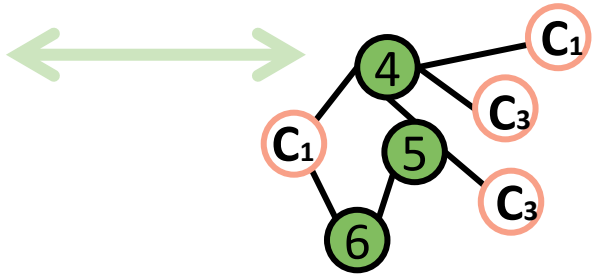
3 Parts

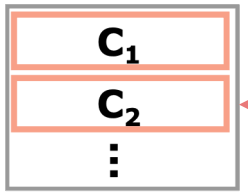




Cross partition
Message Quantizaiton

Online Msg
Quantization
&
Updates





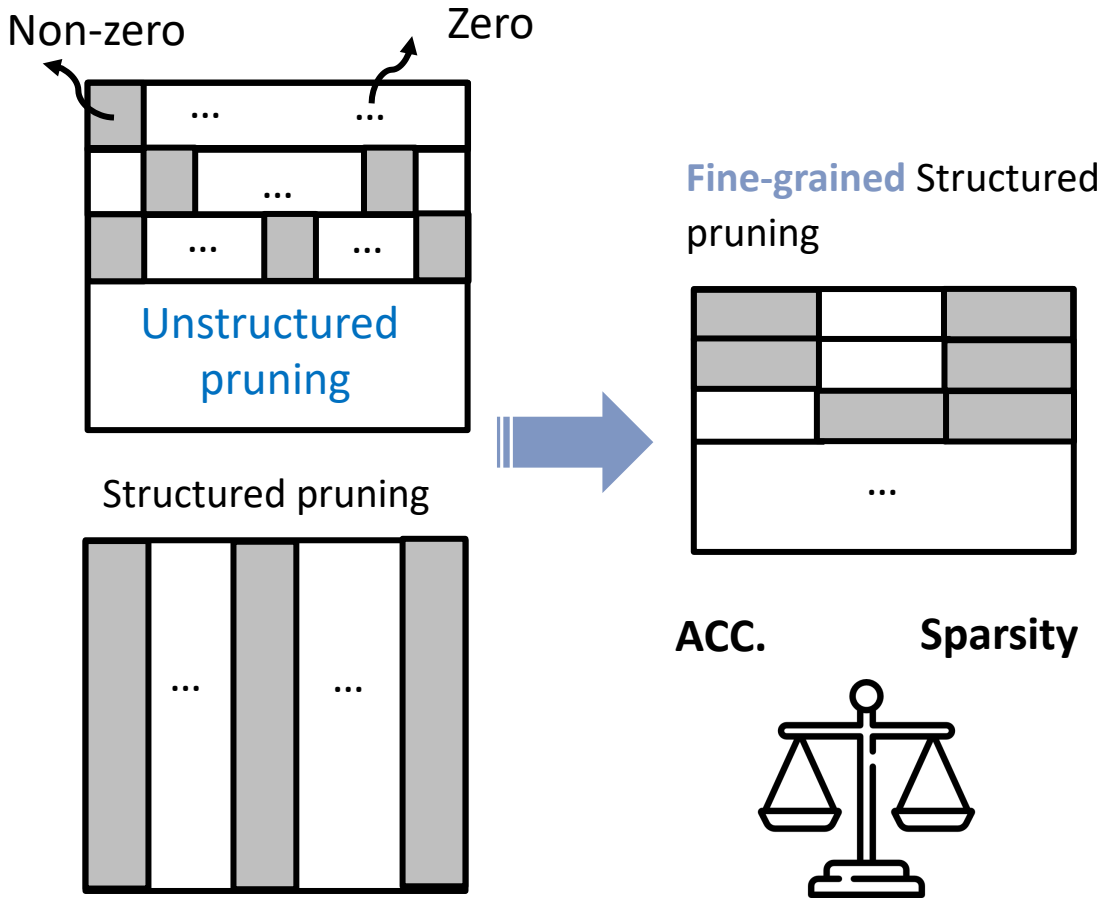
Codebooks
(On-Chip)

👍 No Off-Chip Access

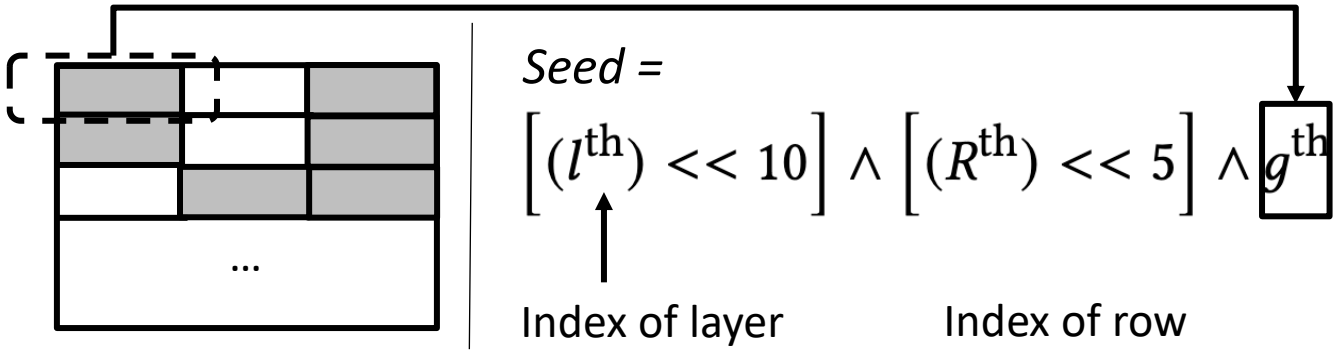
SLT for Sparsity and Efficient Computation

Institute of SCIENCE TOKYO

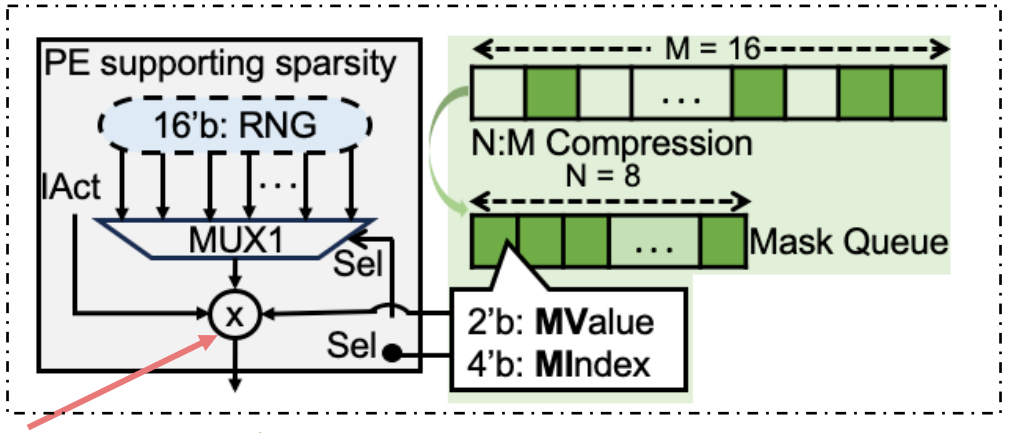
Pruning style



RNG for weights



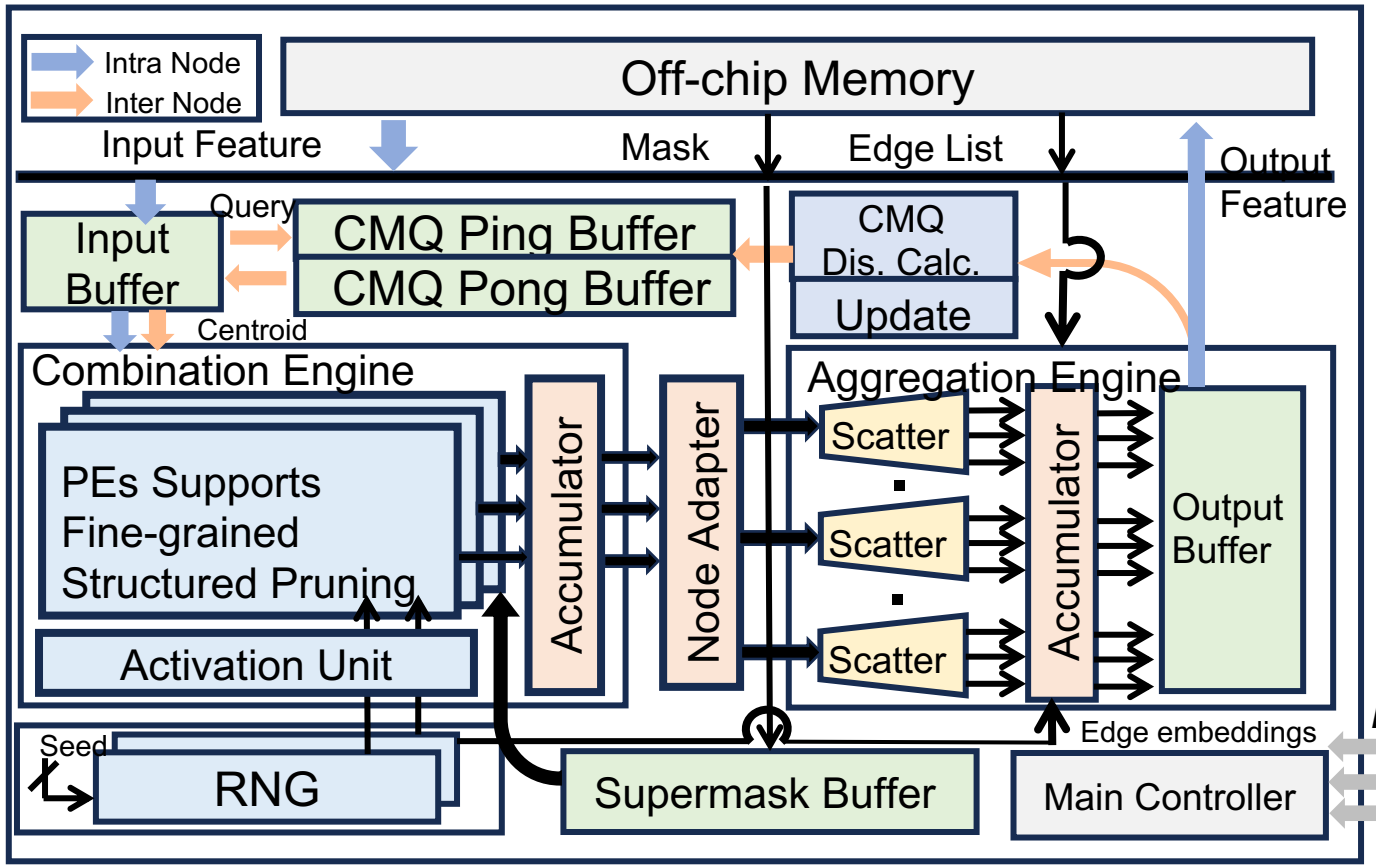
★Solve invalid generation problem

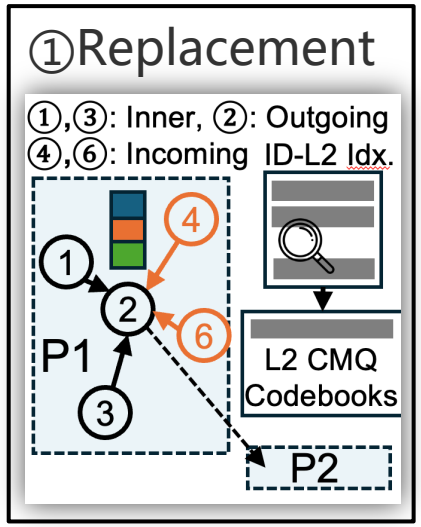


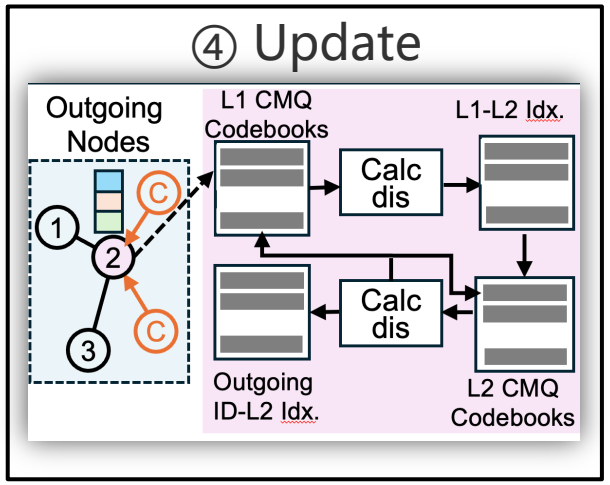
Add + Flip

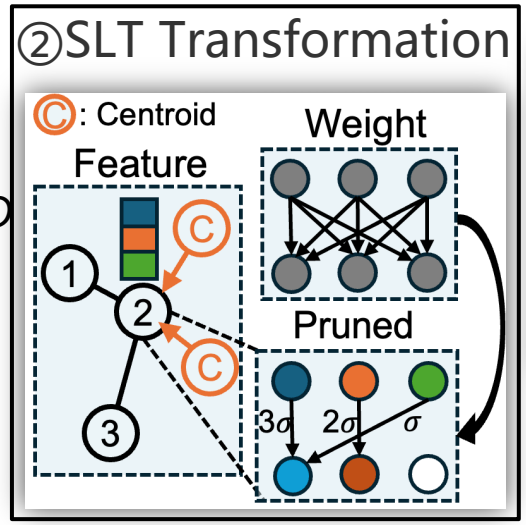
★Solve PE low utilization problem

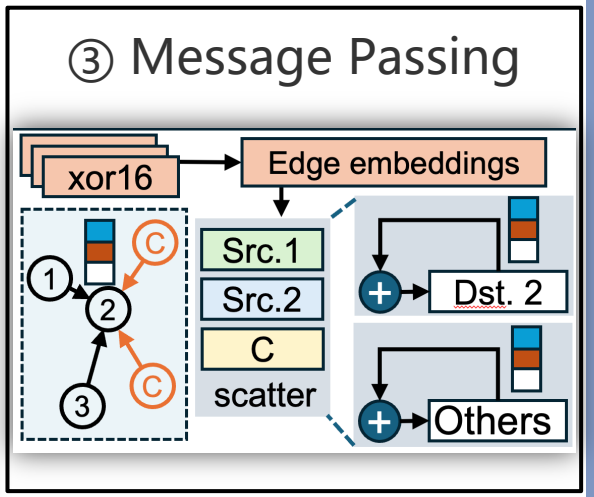
BingoGCN Architecture



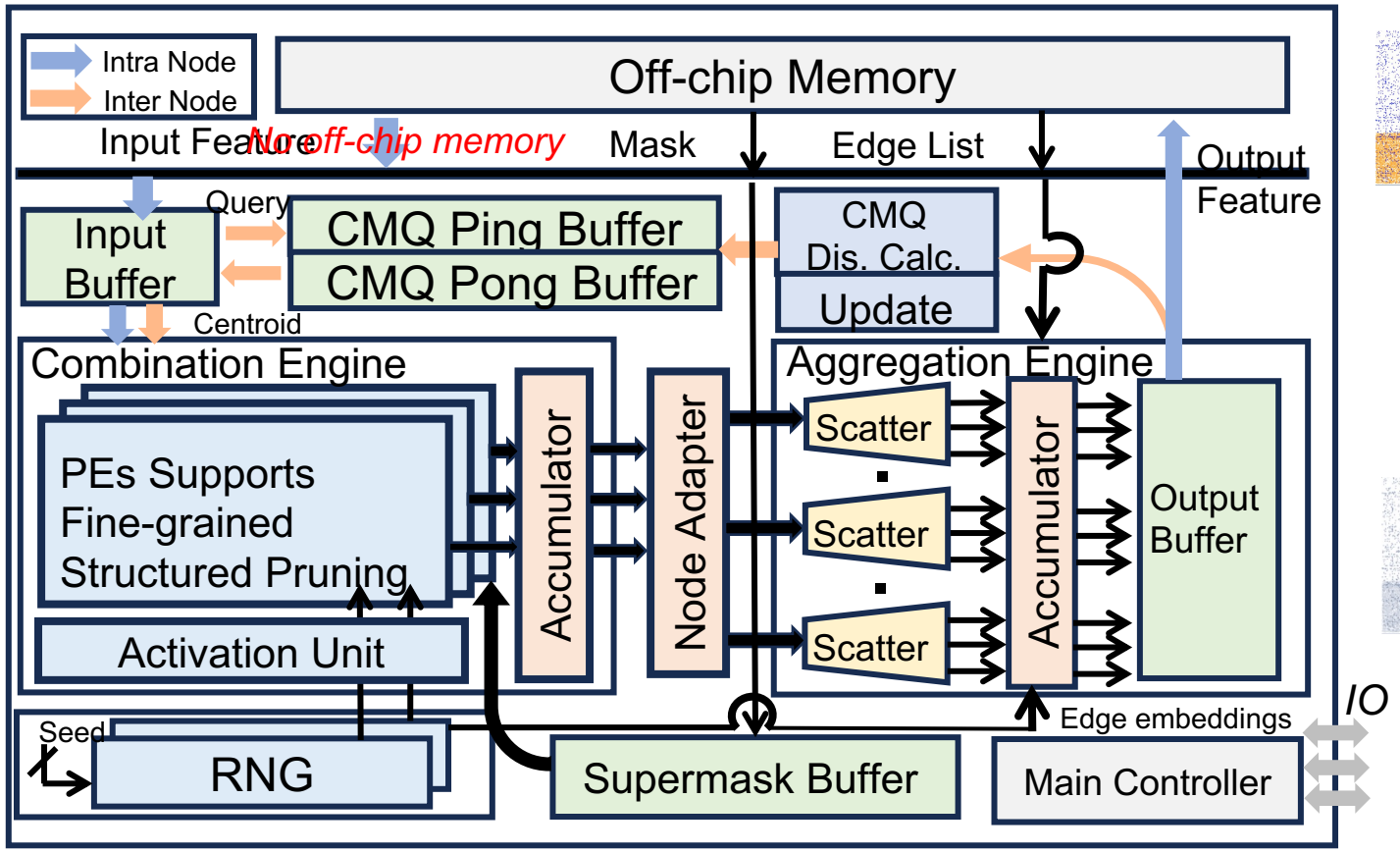


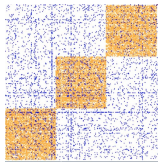






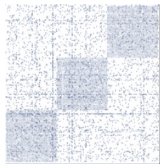
BingoGCN Architecture





Intra Node data access

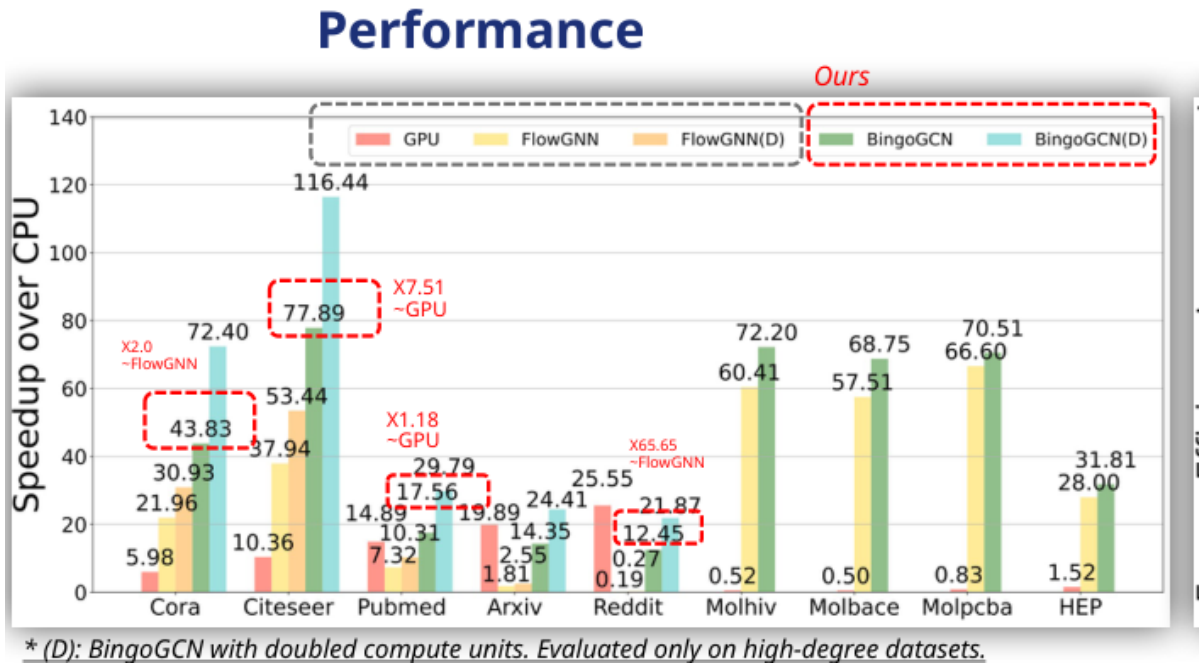
- Sequential access to DRAM
- Capture locality using buffers

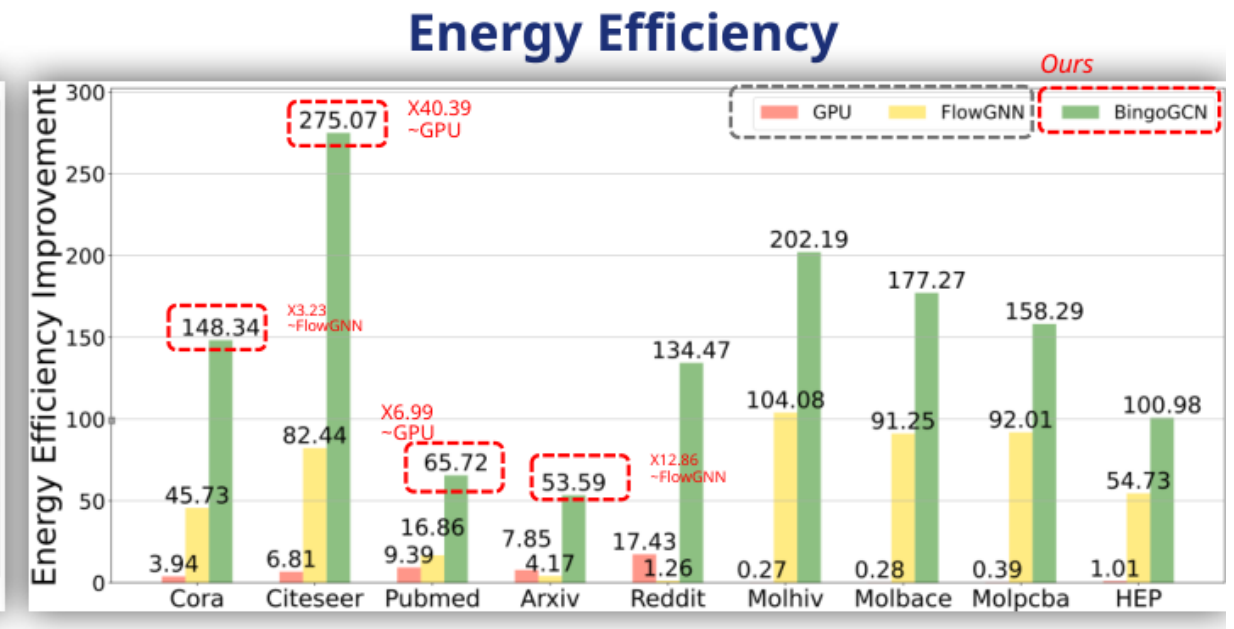


Inter Node data access

- Dynamically summarized by CMQ
- No off-chip memory access!

BingoGCN: Hardware Efficiency





** (D): BingoGCN with doubled compute units. Evaluated only on high-degree datasets.*

2.0–65.7× speedup

3.2–107× energy efficiency

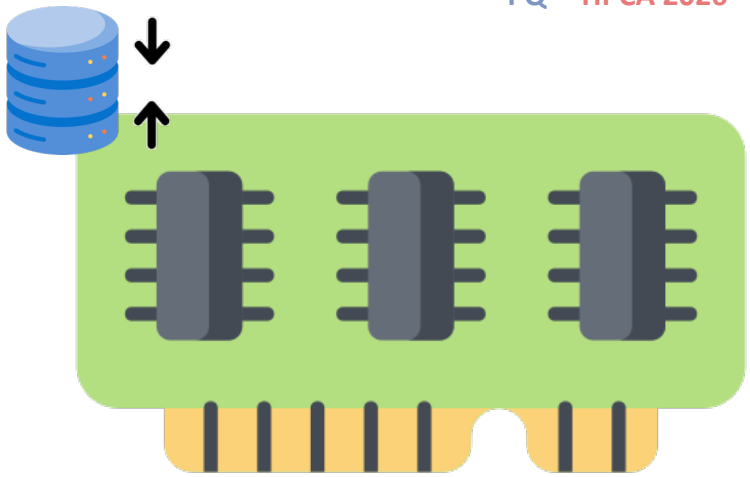
Agenda

Institute of SCIENCE TOKYO

③ Capacity-Wall Problem of LLM on PIM

addressed by PIM-aware KV compression

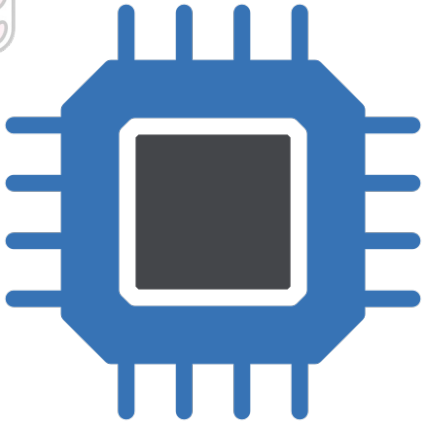
PQ HPCA 2026

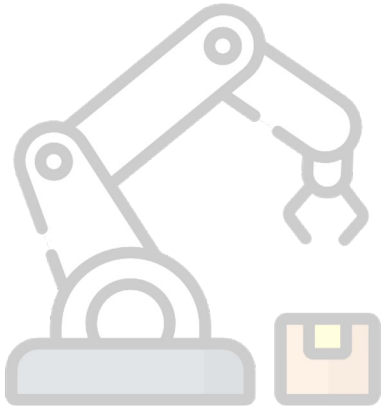


① On-Chip Weight Regeneration

through Strong Lottery Ticket Theory

SLT AAAI 2026 ISSCC 2022





2026 IEEE International Symposium on High-Performance Computer Architecture (HPCA)

AQPIM: Breaking the PIM Capacity Wall for LLMs with In-Memory Activation Quantization

Kosuke Matsushima†, Yasuyuki Okoshi†, Masato Motomura†, and Daichi Fujiki†
Institute of Science Tokyo†
{matsushima.kosuke, okoshi.yasuyuki, motomura, dfujiki}@artic.iir.isct.ac.jp

④ Memory-Centric Approach for Physical AI

Action caching and local action generation

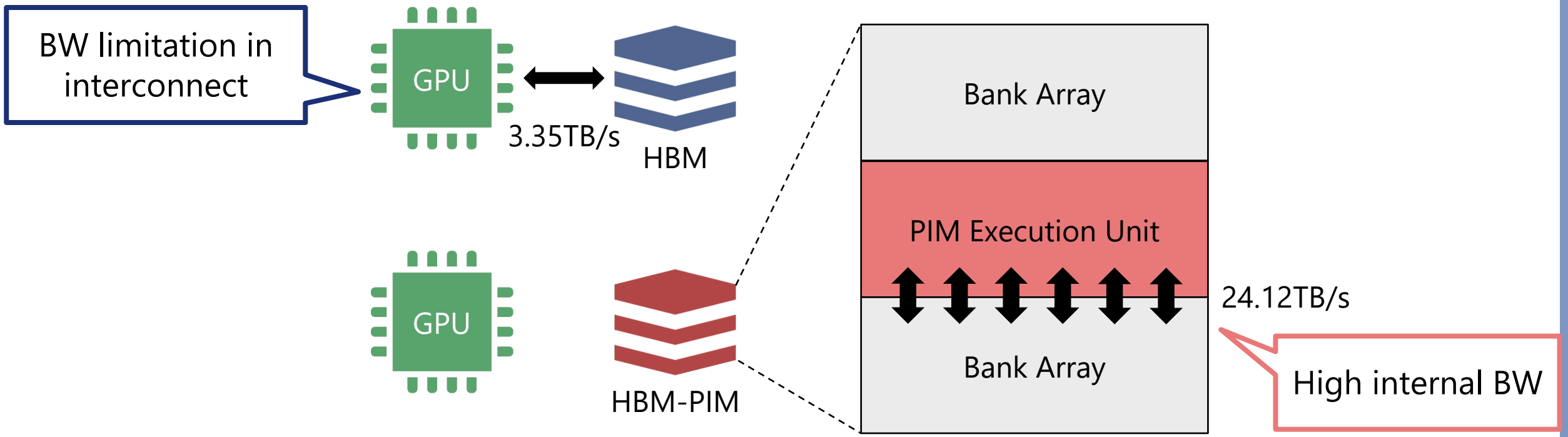
AC



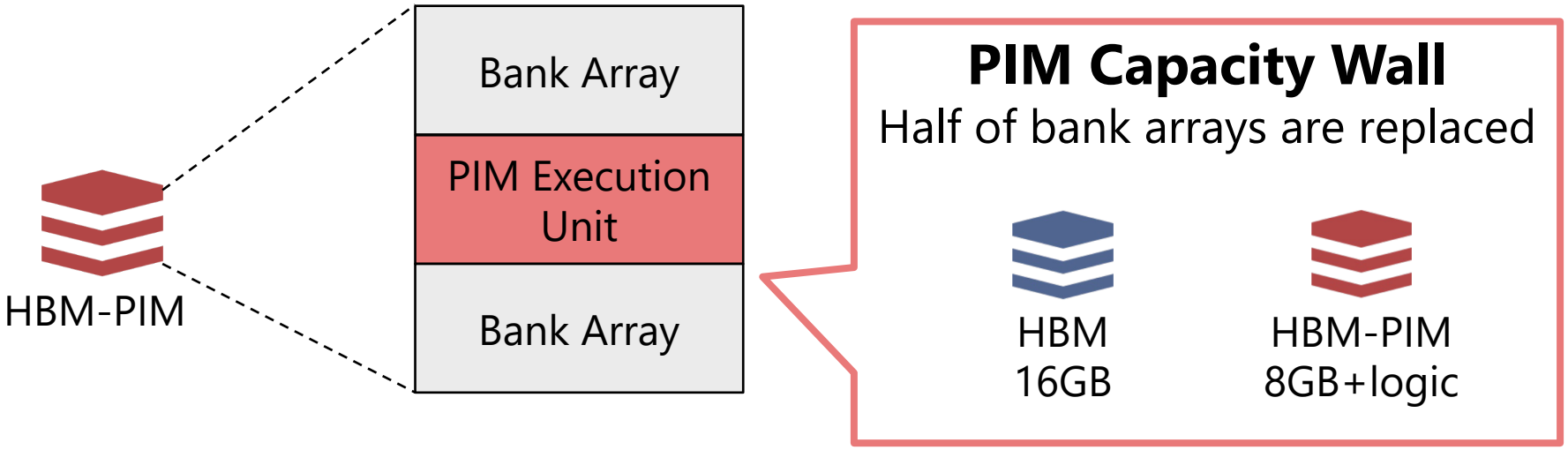
Addressing Memory-bound Computation in LLM

PIM (Processing-in-Memory) accelerates **memory-bound** computations

- Reduce data movement overhead by placing execution units within the memory



PIM Challenges: PIM Capacity Wall



How should we mitigate the PIM Capacity Wall?

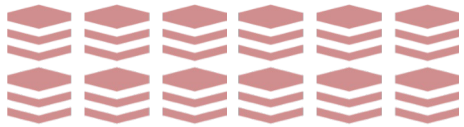
Naïve scaling | Quantization | Offloading

PIM Challenges: PIM Capacity Wall

Institute of SCIENCE TOKYO

Naïve scaling

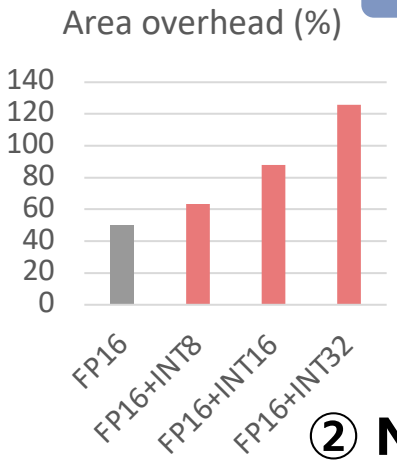
Massive memory consumption due to long context and CoT



10s to 100s of HBM-PIMs are required

① **Require data compression**

Quantization



Supporting multiple numerical formats in PIM increases area overhead

② **Need to minimize area overhead**

Offloading



HBM CPU memory

Offload KV cache to large CPU memory

- Low bandwidth causes a bottleneck
- Sparse access lower PIM utilization

③ **Desired to complete operations within PIM**

Proposal: AQPIM



Existing approaches are ill-suited for PIM

① **High compression ratio**

② **Minimal area overhead**

③ **Complete operations within PIM**

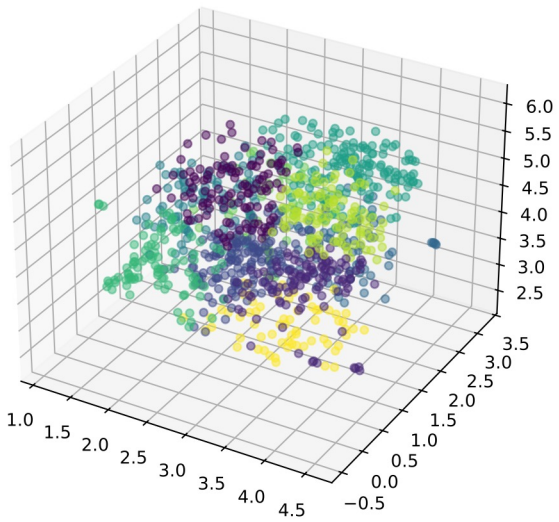
KV Cache Locality

KV cache exhibits context dependent
distribution biases

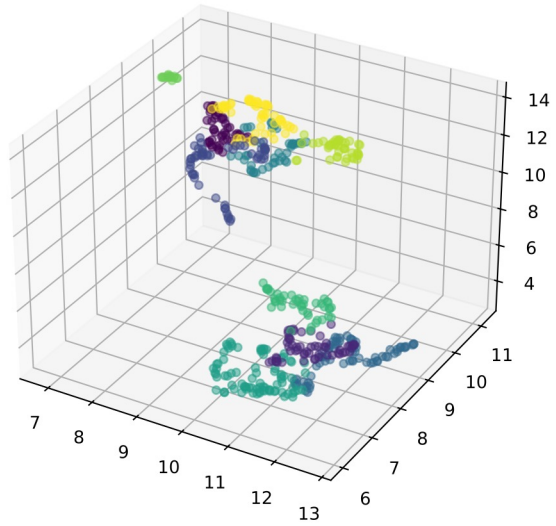
↓

Online clustering-based quantization
is suitable for compressing KV cache

Weight



Key



Visualization by UMAP [5]

Dimensionality reduction technique
preserving data's topological structure

[5] Leland McInnes, et al.,"UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv, 2018.

Product Quantization (PQ)

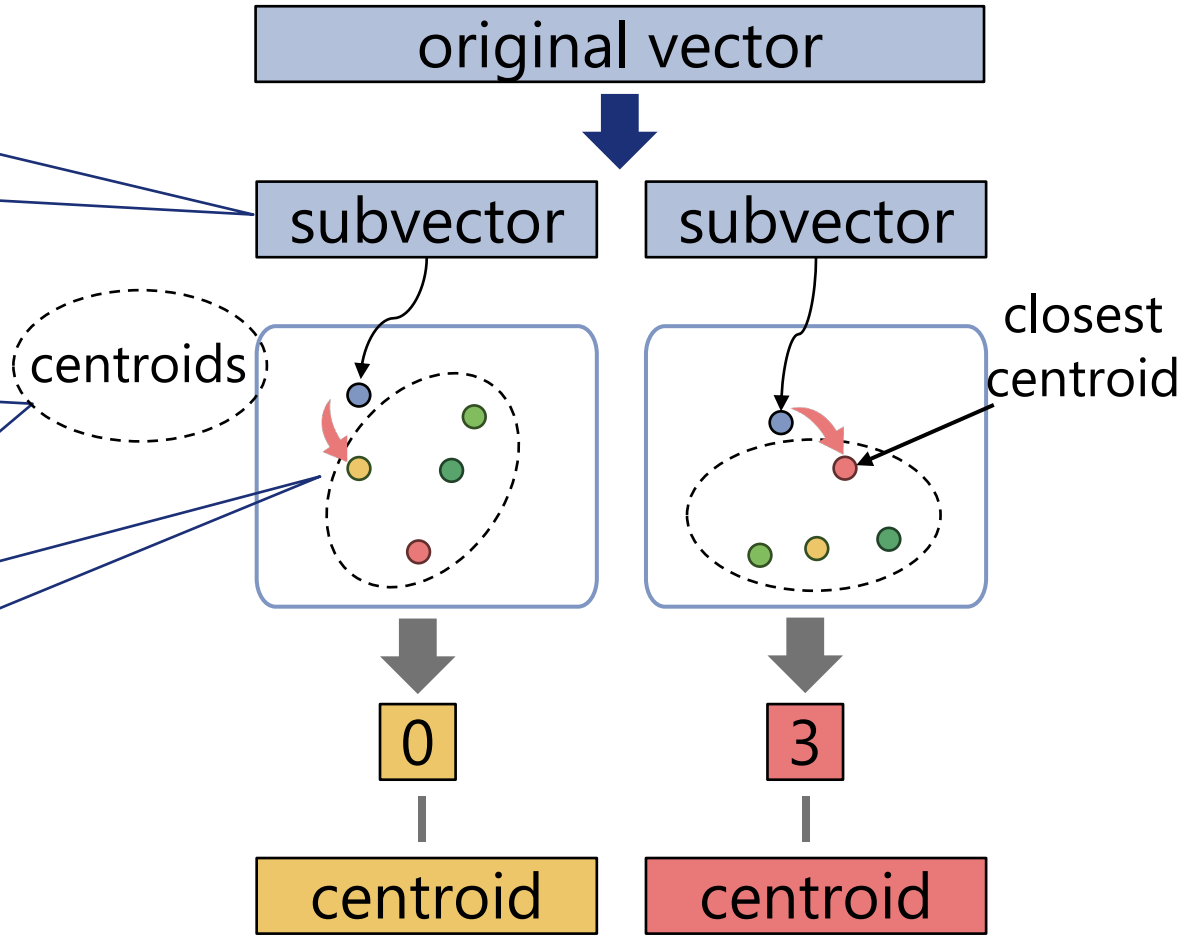
Institute of SCIENCE TOKYO

① **Split original vector into subvectors**
Expressiveness and parallelism

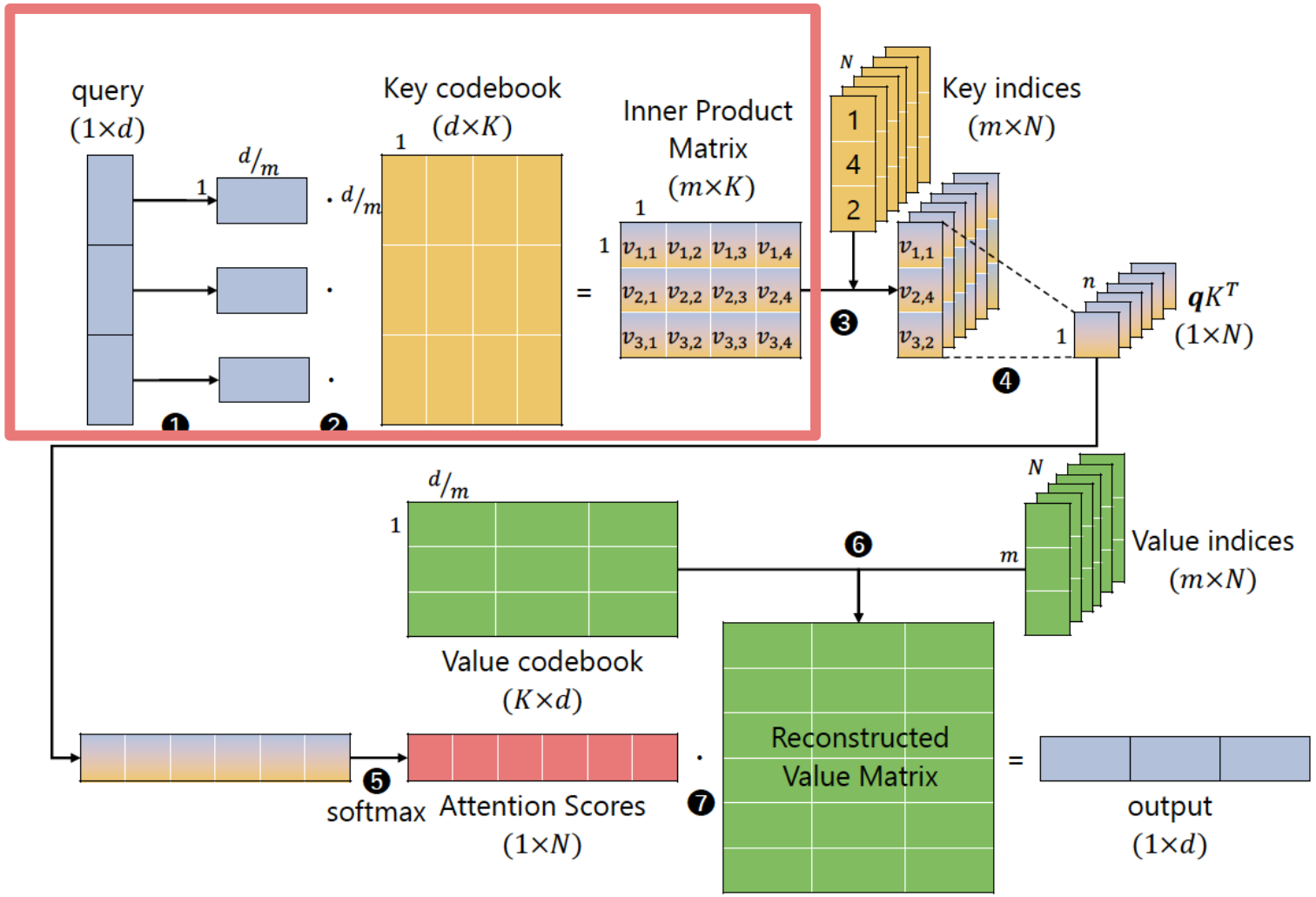
② **Perform clustering using k-means**
Calculate centroids online

③ **Assign each subvector to a centroid**
Represent each subvector by index

We further apply LLM-specific optimizations to preserve accuracy



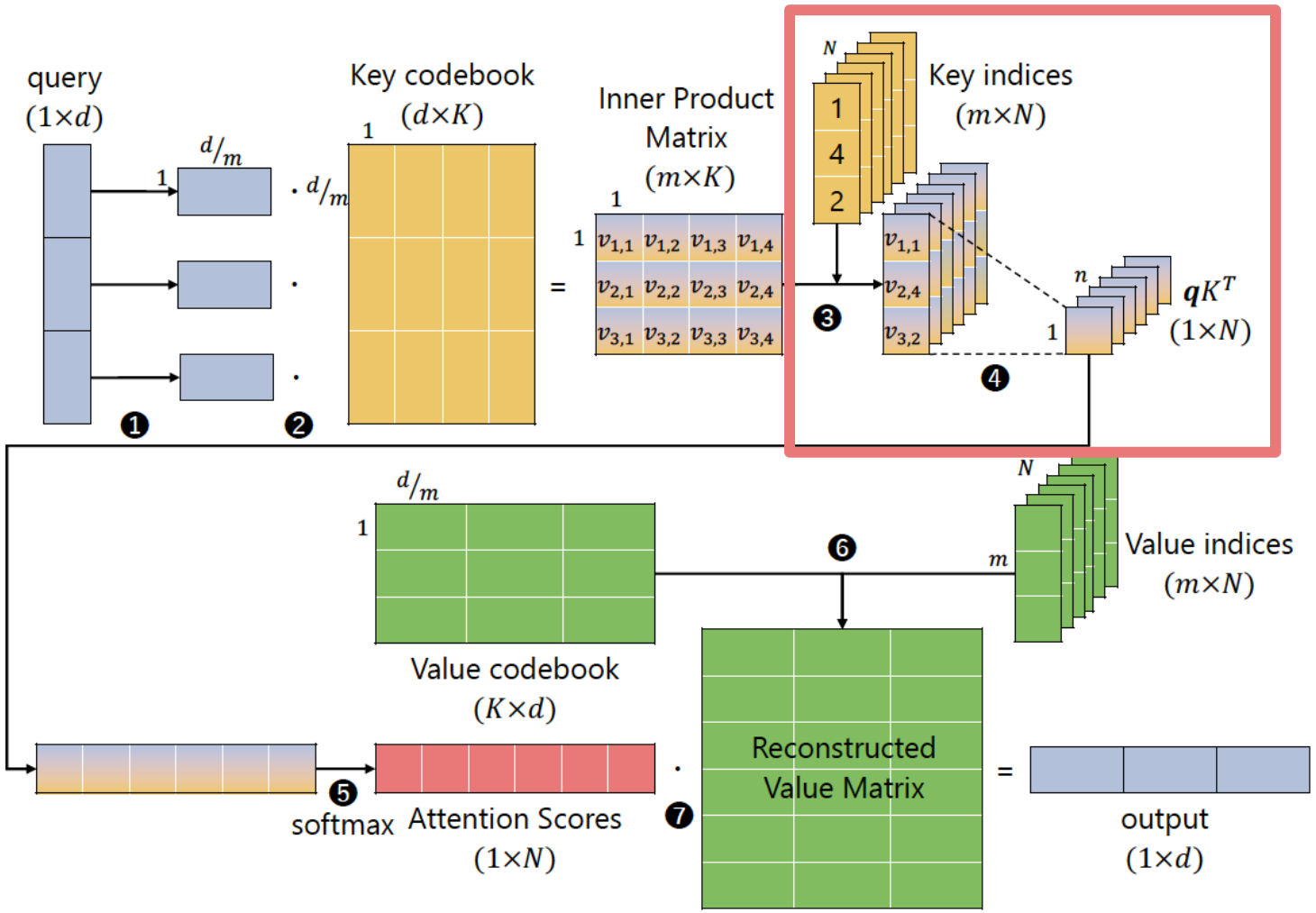
Attention Computation w/o Dequantization



01 QUERY × KEY CODEBOOK

Build small inner-product tables, not the full key matrix.

Attention Computation w/o Dequantization



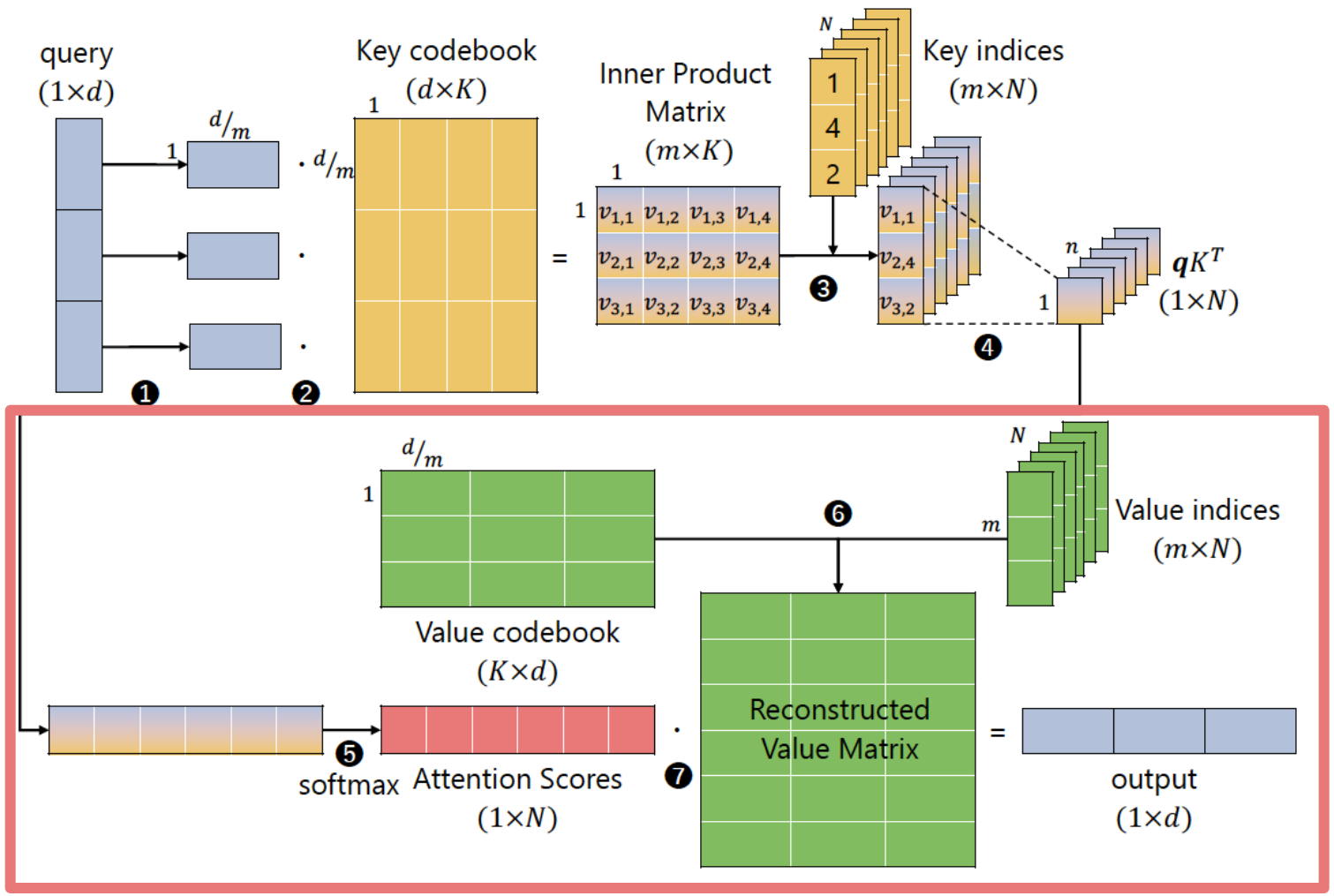
01 QUERY × KEY CODEBOOK

Build small inner-product tables, not the full key matrix.

02 INDEX LOOKUP + SUM

Stored key indices select partial products; sum across subvectors.

Attention Computation w/o Dequantization



01 QUERY × KEY CODEBOOK

Build small inner-product tables, not the full key matrix.

02 INDEX LOOKUP + SUM

Stored key indices select partial products; sum across subvectors.

03 SOFTMAX + VALUES

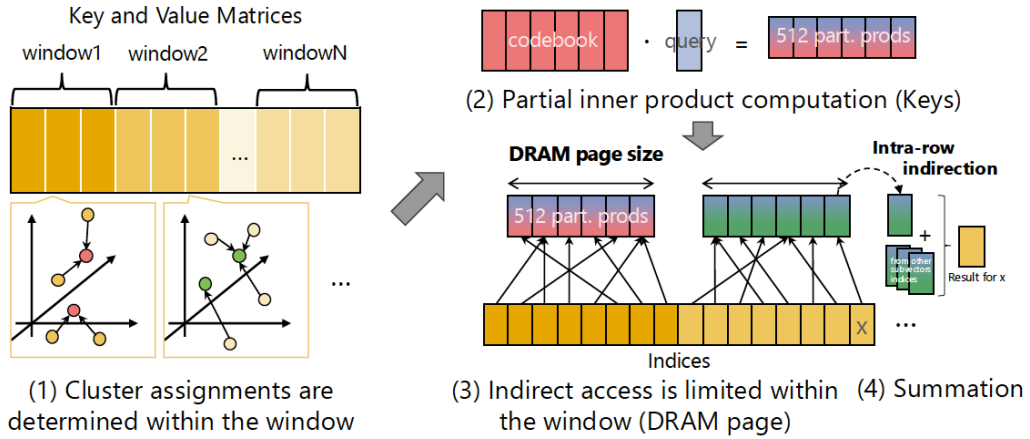
Apply softmax, then combine value codebooks and indices for output.

Compute directly on compressed PQ data.

Resolving Random Access to Inner Prod Matrix

Row-buffer locality

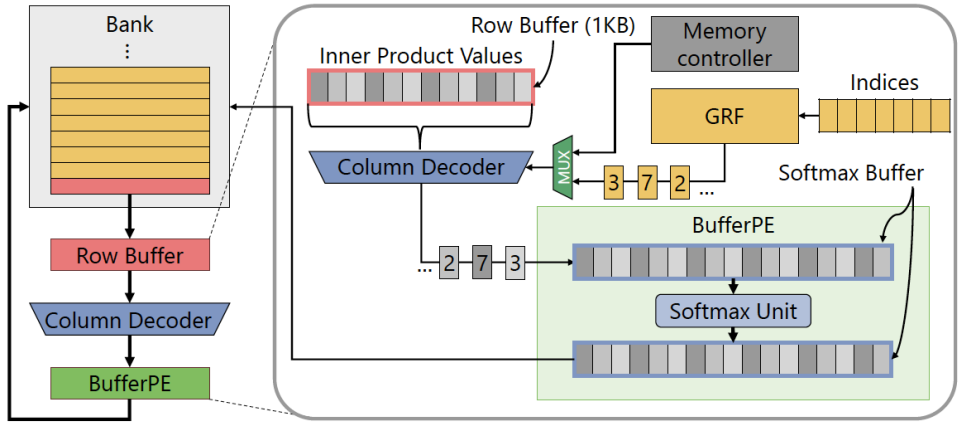
Keep the lookup table within one row buffer.



Page-aware clustering algorithm

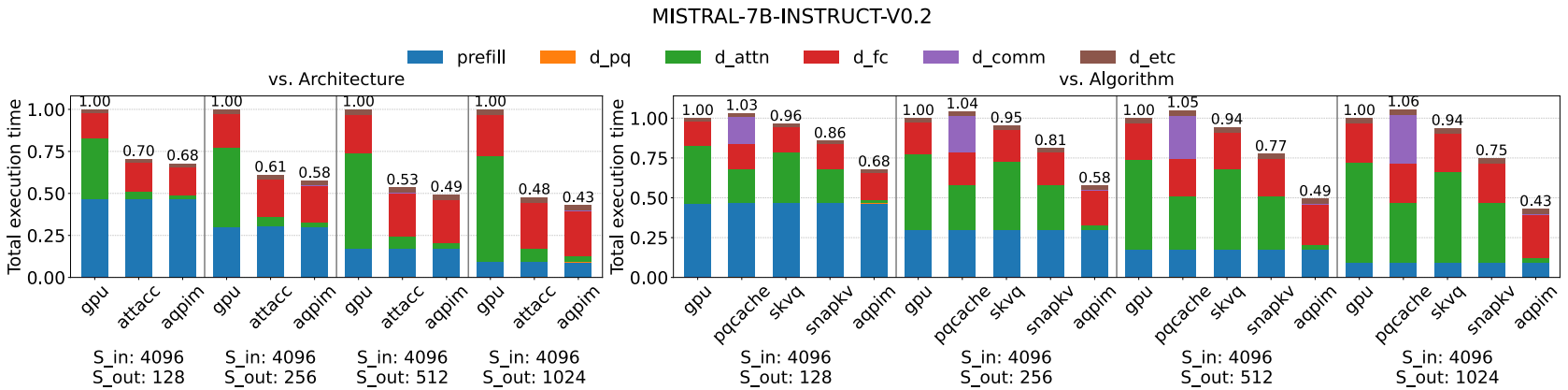
Indirect access

Route stored indices to the column decoder.

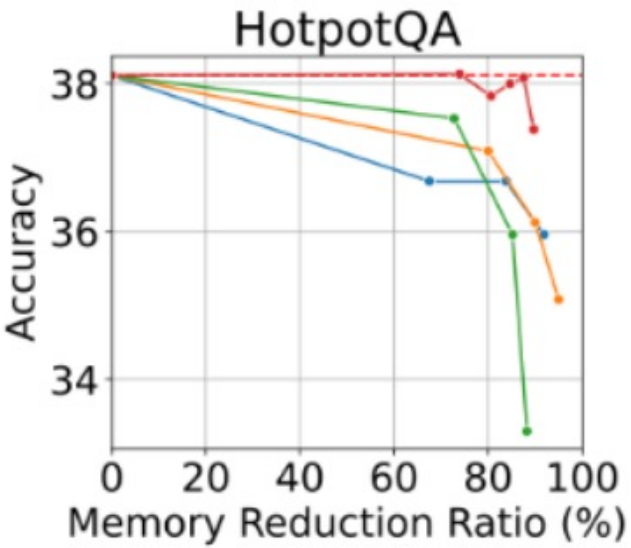


Intra-row indirection in PIM

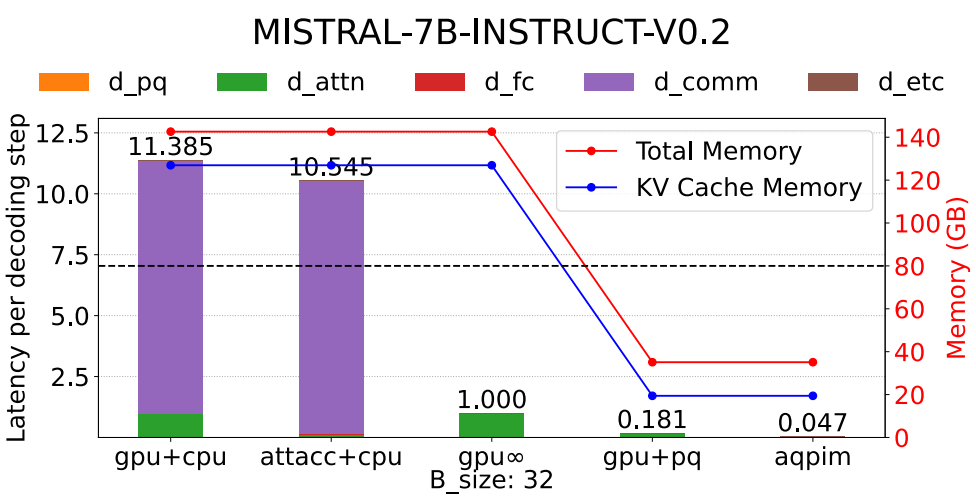
AQPIM: Capacity and Latency Benefits



AQPIM consistently achieves high compressibility and low latency



Red: AQPIM Blue: SnapKV
Orange: PQCache Green: SKVQ



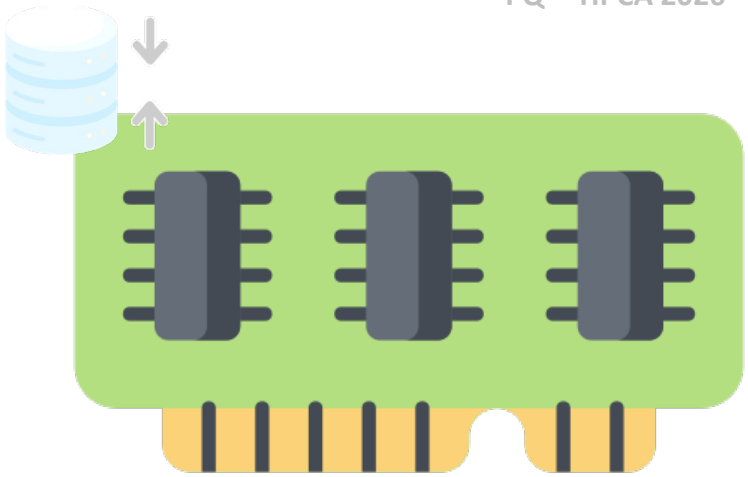
AQPIM shows superior performance especially when KV overflows GPU memory

3.4× vs. idealized AttAcc with sufficient capacity

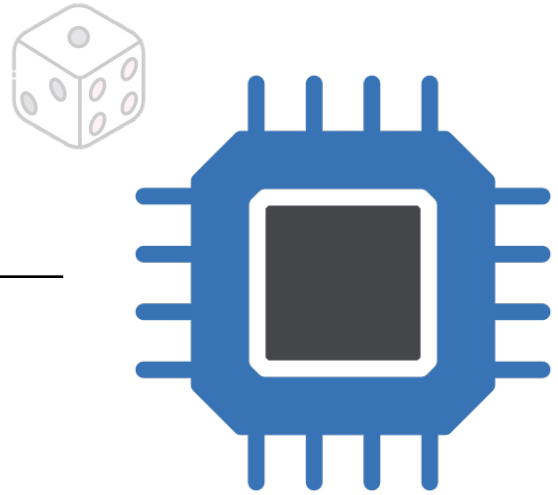
Agenda

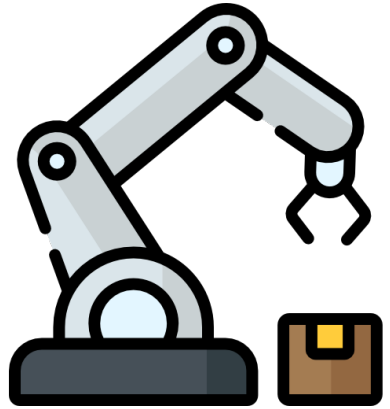
Institute of SCIENCE TOKYO

③ **Capacity-Wall Problem of LLM on PIM**
addressed by PIM-aware KV compression
PQ HPCA 2026



① **On-Chip Weight Regeneration**
through Strong Lottery Ticket Theory
SLT AAAI 2026 ISSCC 2022



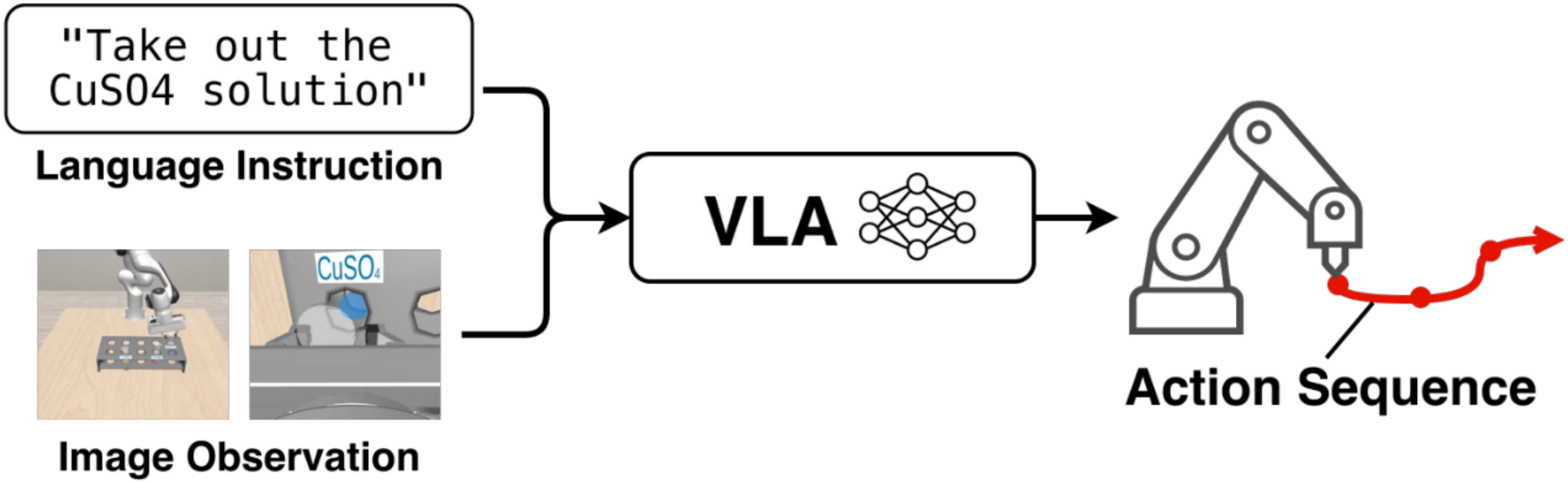


② **Sparse Access Localization for GNNs**
through X-partition message quantization
SLT+VQ ISCA 2025

④ **Memory-Centric Approach for Physical AI**
Action caching and local action generation
AC



Vision-Language-Action Models



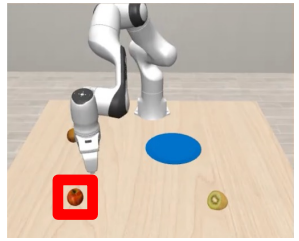


Repeated decisions create opportunities to reuse prior computation.

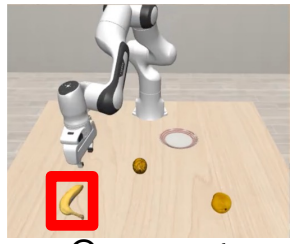
ActionCache: Caching Intermediate Actions

Reuse past actions as prior knowledge

Past Condition



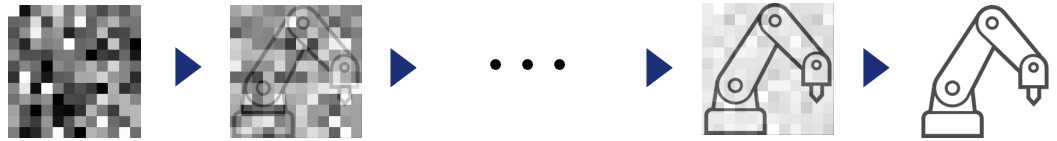
Current Condition



High similarity

Action Head

$\times N_{\text{full}}$ steps

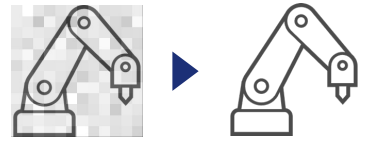


Reuse

$\times N_{\text{hit}}$ steps

Skip

$\times N_{\text{hit}}$ steps

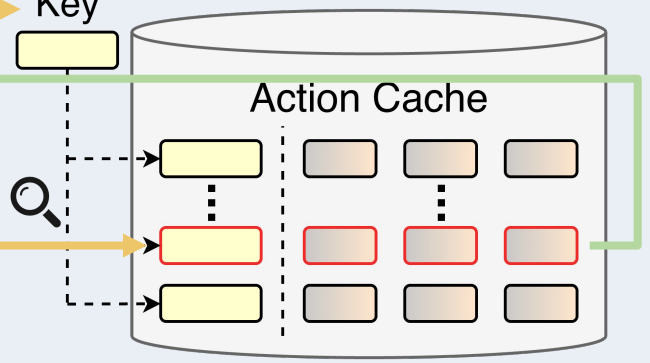


Action Cache

Similarity search with Action Condition (Key)

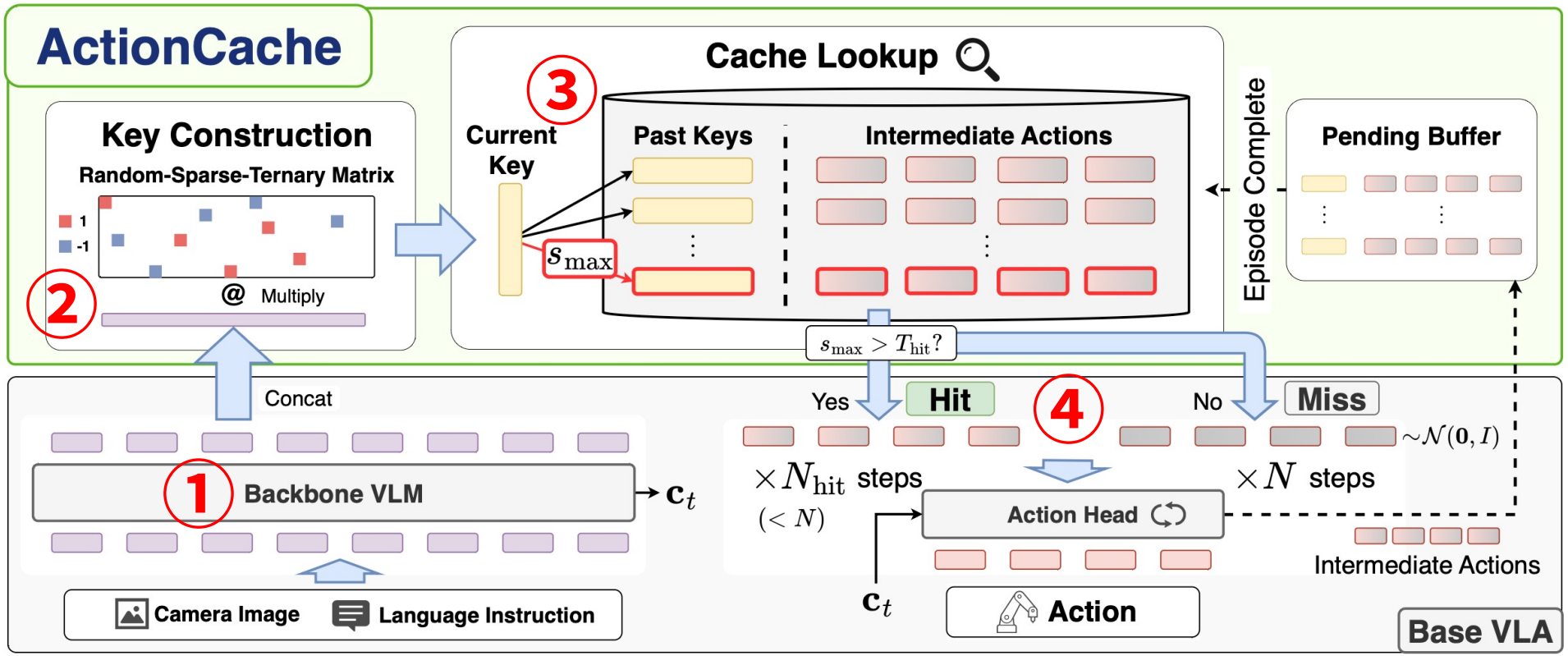
Key

Action Cache



✓ **Fewer denoising steps** accelerate **inference**

ActionCache



① Extract visual and language features

② Build the cache key: compress the VLM output with a random ternary sparse matrix

③ Cache lookup: Similarity search between keys ⇒ Hit if exceeds the threshold

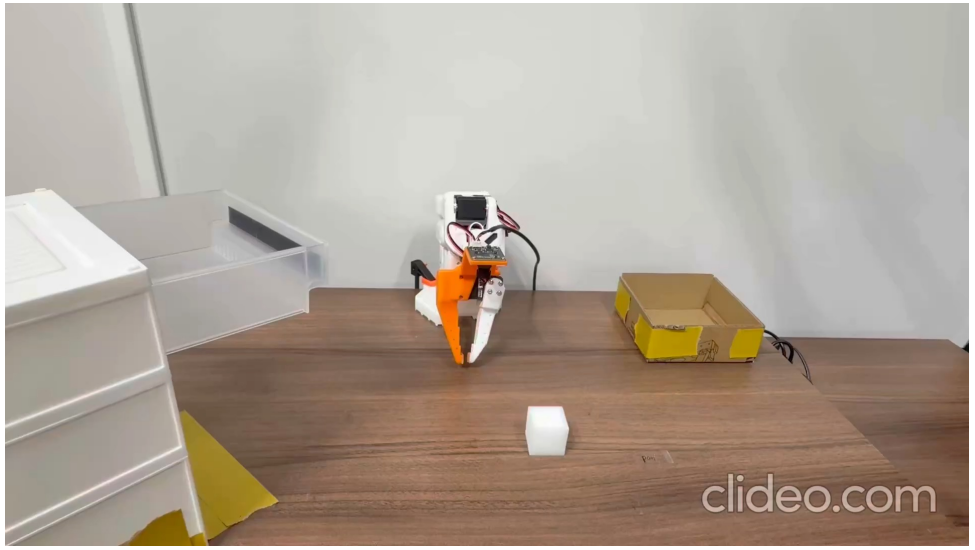
④ Action generation: Hit — fewer denoising steps / Miss — full denoising steps

Experiment: Real-Robot Evaluation

Institute of SCIENCE TOKYO

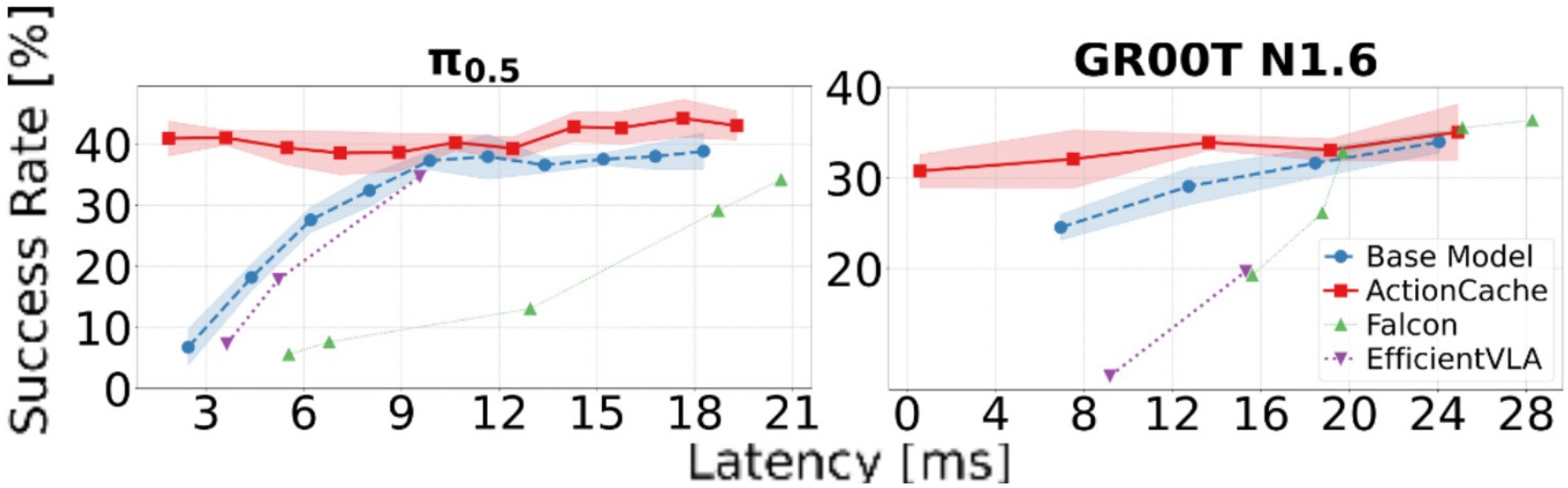
- SO-101 Robot-Arm Experiments

✓ **One-step generation**
matches base-model task success

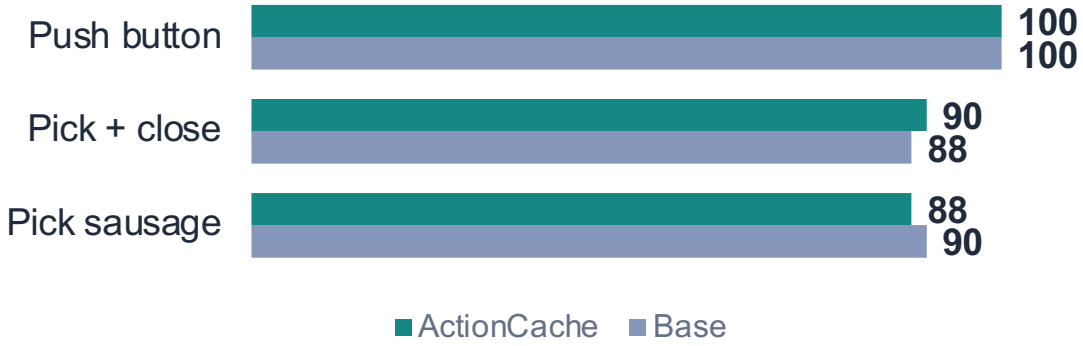


Task	Method	NFE	SR (%)	HR (%)	Latency (ms)					
					Emb.	VLM	Key proj.	Cache lookup	Action head	Overall
`pick_sausage`	Base Model	10	90	-	24.22	22.21	-	-	56.70 (1.00×)	103.13 (1.00×)
	ActionCache	1	88	41.8	24.26	22.24	0.17	0.18	35.06 (**1.62**×)	81.90 (**1.26**×)
`pick_and_close`	Base Model	10	88	-	24.23	22.22	-	-	55.79 (1.00×)	102.24 (1.00×)
	ActionCache	1	90	82.8	24.29	22.26	0.17	0.17	14.67 (**3.80**×)	61.56 (**1.66**×)
`push_button`	Base Model	10	100	-	24.31	22.29	-	-	59.65 (1.00×)	106.26 (1.00×)
	ActionCache	1	100	94.4	24.33	22.32	0.17	0.11	9.53 (**6.26**×)	56.46 (**1.88**×)

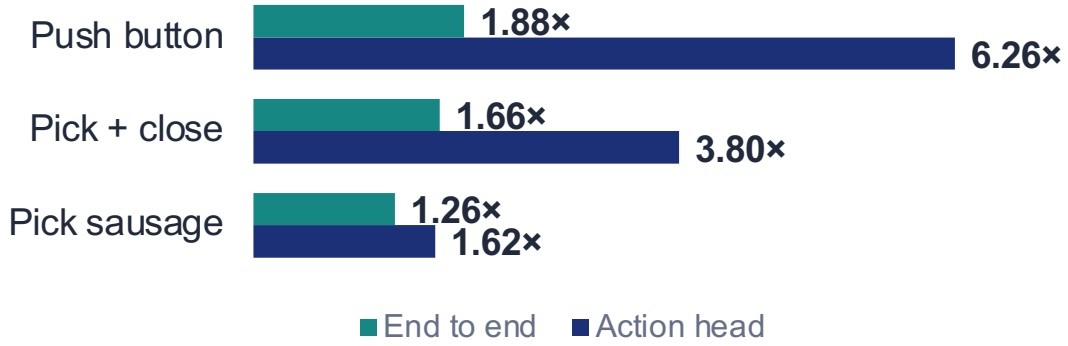
ActionCache: Success and Latency



SO-101 task success (%)



SO-101 speedup



10 base steps; 1 refinement step on a cache hit. Values from the original real-robot results.

Can We Skip the Entire VLA Call?

ActionCache



VLA-ULAP: selected local decisions



Local prediction removes selected full-model decisions.

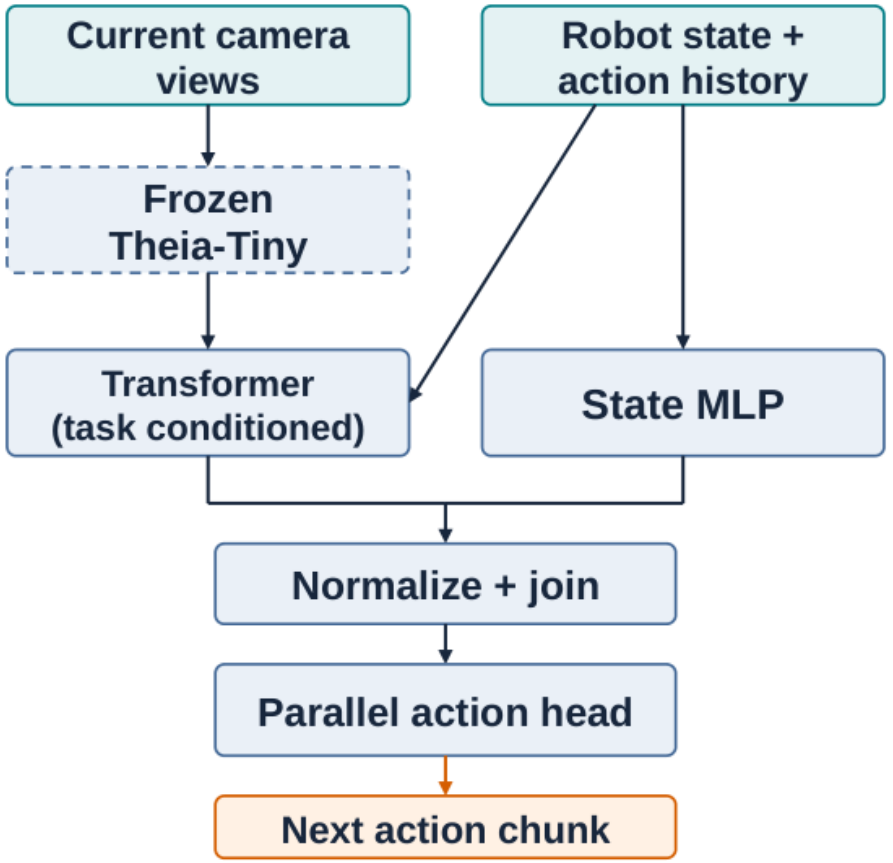
VLA-ULAP: Local Decisions with VLA Refresh

Fixed schedule at chunk boundaries



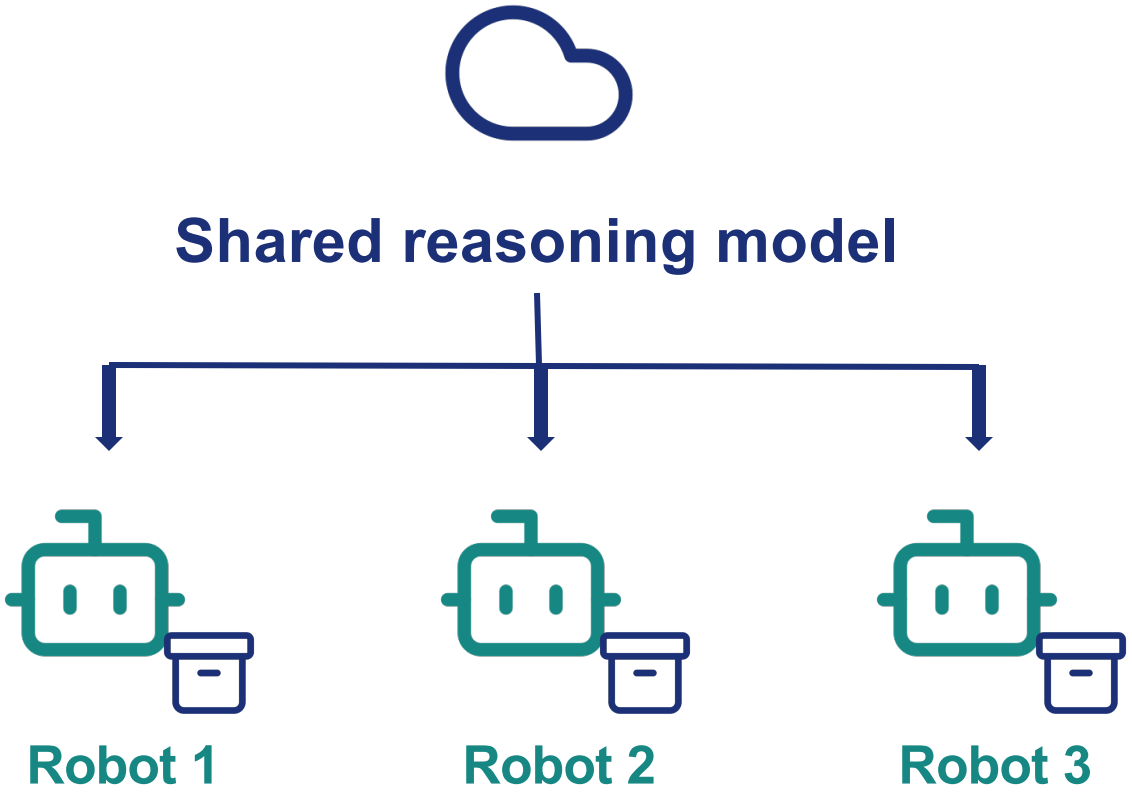
Full policy at selected times

Fresh local observations for each action chunk



Local decisions avoid round trips. VLA refreshes limit drift.

A Memory Hierarchy for Robots



What should be shared?

Semantic reasoning across robots

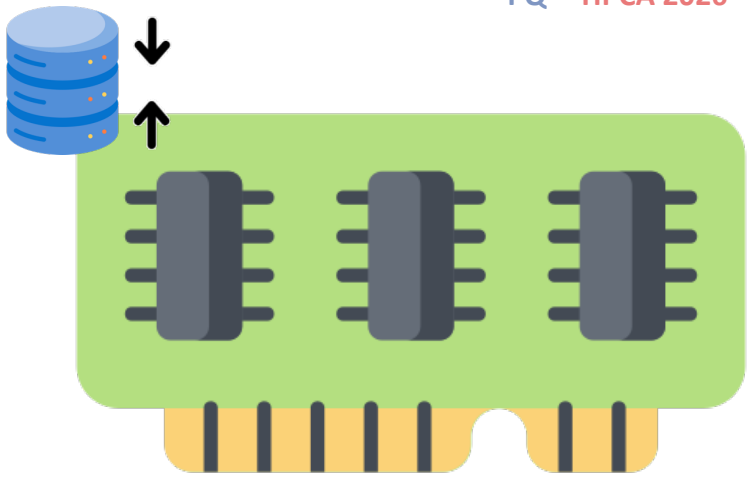
- Who queries, and when?
- What stays near the robot?
- How does it adapt?

Key Takeaway:
Data representations shapes what we compute

Institute of SCIENCE TOKYO

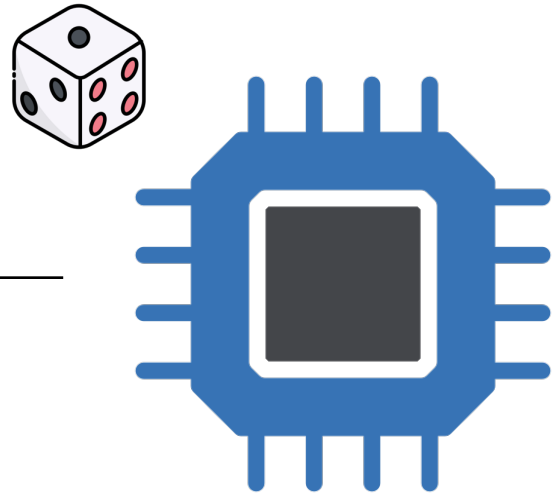
③ Capacity-Wall Problem of LLM on PIM
addressed by PIM-aware KV compression

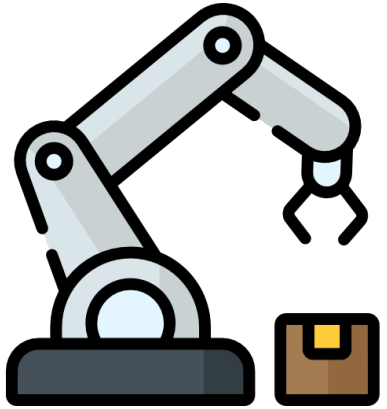
PQ HPCA 2026



① On-Chip Weight Regeneration
through Strong Lottery Ticket Theory

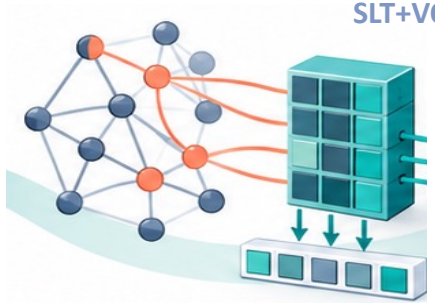
SLT AAAI 2026 ISSCC 2022





② Sparse Access Localization for GNNs
through X-partition message quantization

SLT+VQ ISCA 2025



④ Memory-Centric Approach for Physical AI
Action caching and local action generation

AC

