Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLMs

Yeonhong Park
ilil96@snu.ac.kr

Jake Hyun
jakehyun@snu.ac.kr

SangLyul Cho
chosanglyul@snu.ac.kr

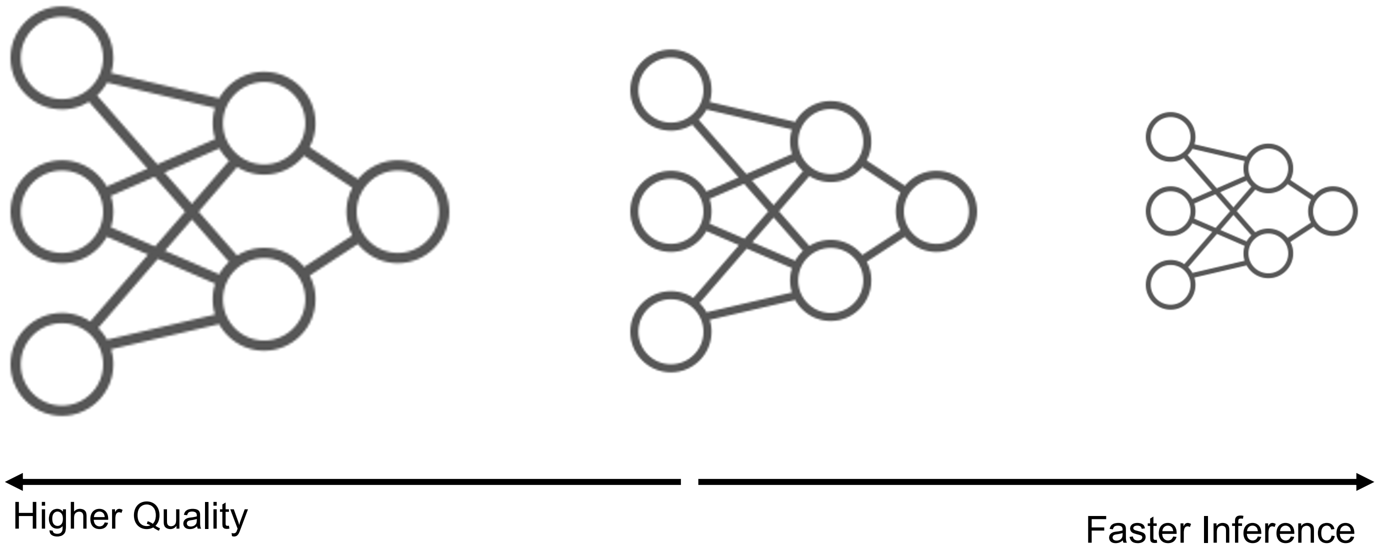
Bonggeun Sim
bg.sim@snu.ac.kr

Jae W. Lee
jaewlee@snu.ac.kr

Department of CSE
Seoul National University

41st International Conference on Machine Learning (ICML 2024), Vienna, Austria

Often, We Need Multiple LLMs of Different Sizes

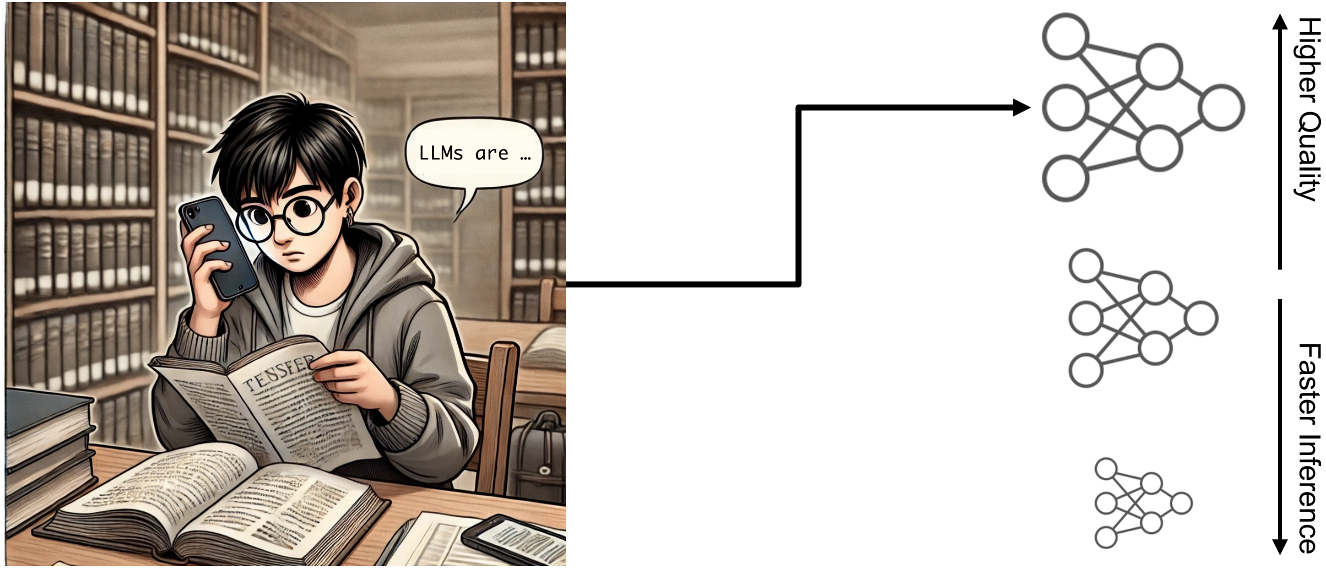


Often, We Need Multiple LLMs of Different Sizes

Scenario #1: Different queries with different latency requirements

Often, We Need Multiple LLMs of Different Sizes

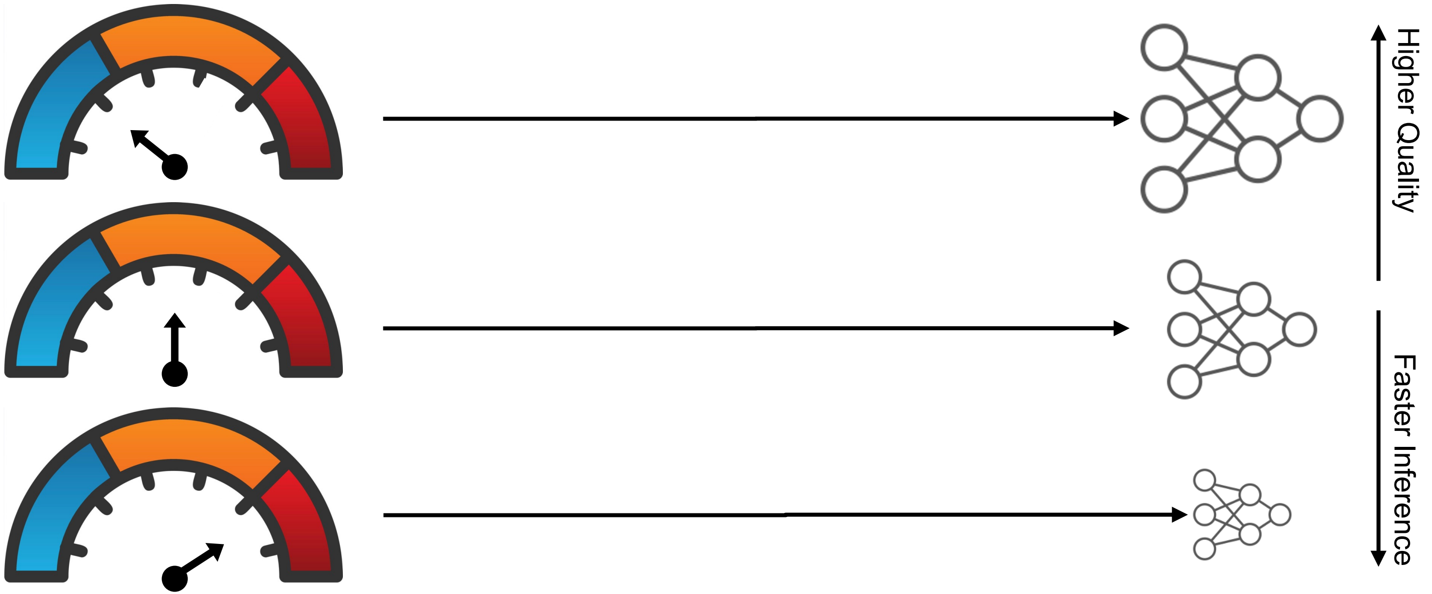
Scenario #1: Different queries with different latency requirements



Department of CSE
Seoul National University

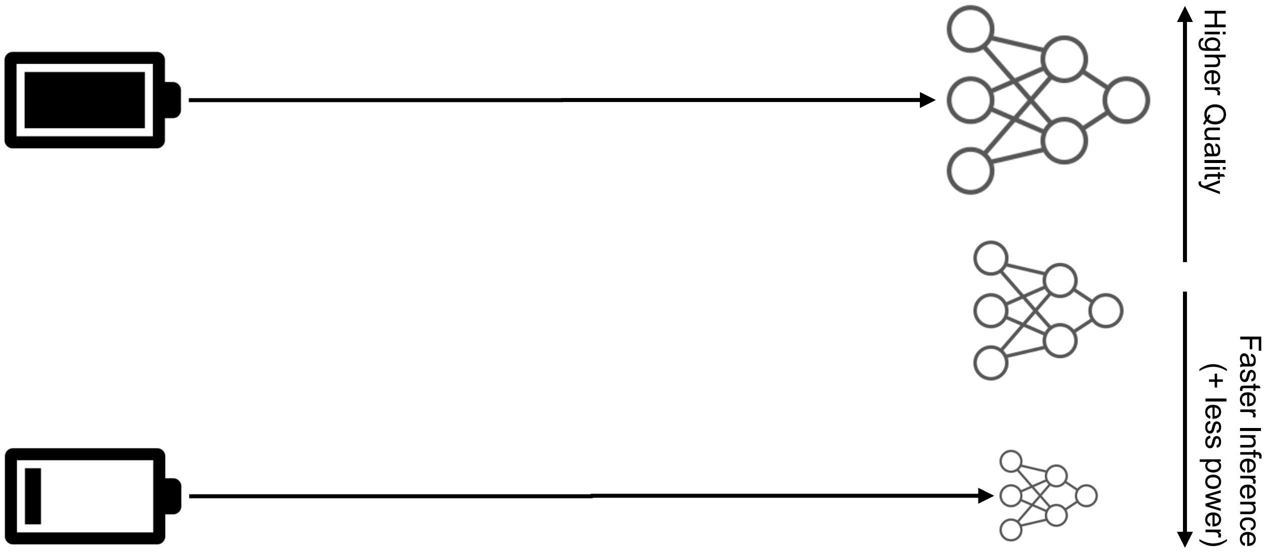
Often, We Need Multiple LLMs of Different Sizes

Scenario #2: Dynamic fluctuation of system load and power budget



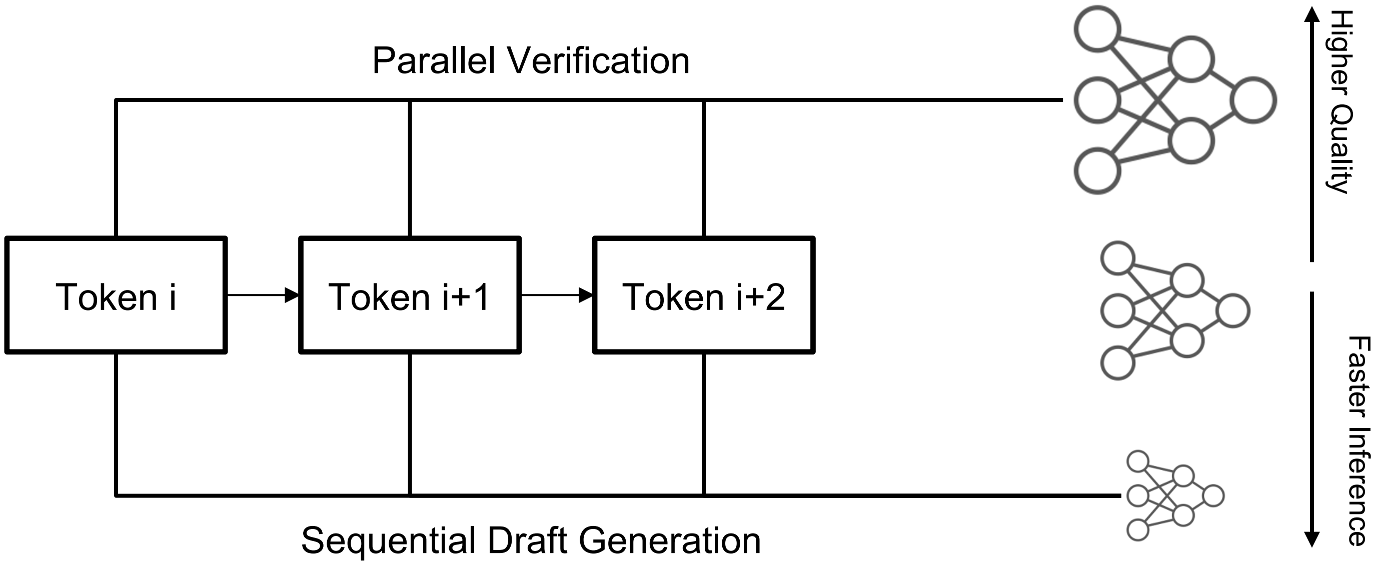
Often, We Need Multiple LLMs of Different Sizes

Scenario #2: Dynamic fluctuation of system load and power budget



Often, We Need Multiple LLMs of Different Sizes

Scenario #3: Speculative decoding

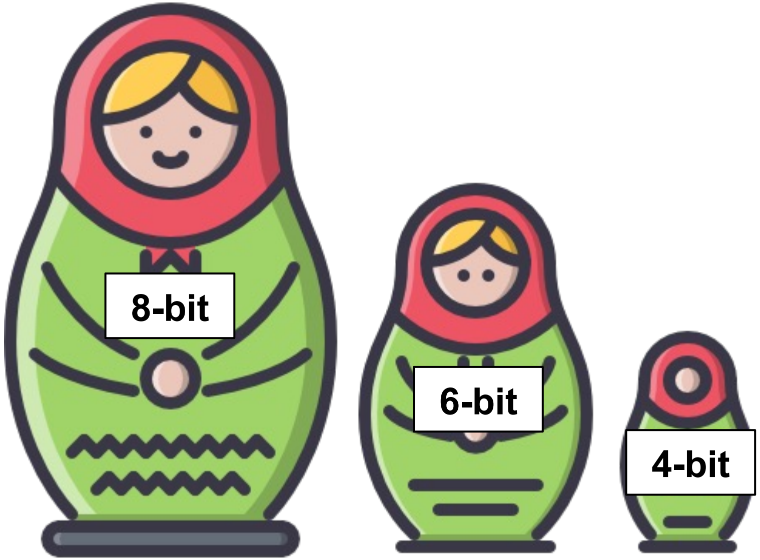


Y. Park et al., “Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM”

Research Question

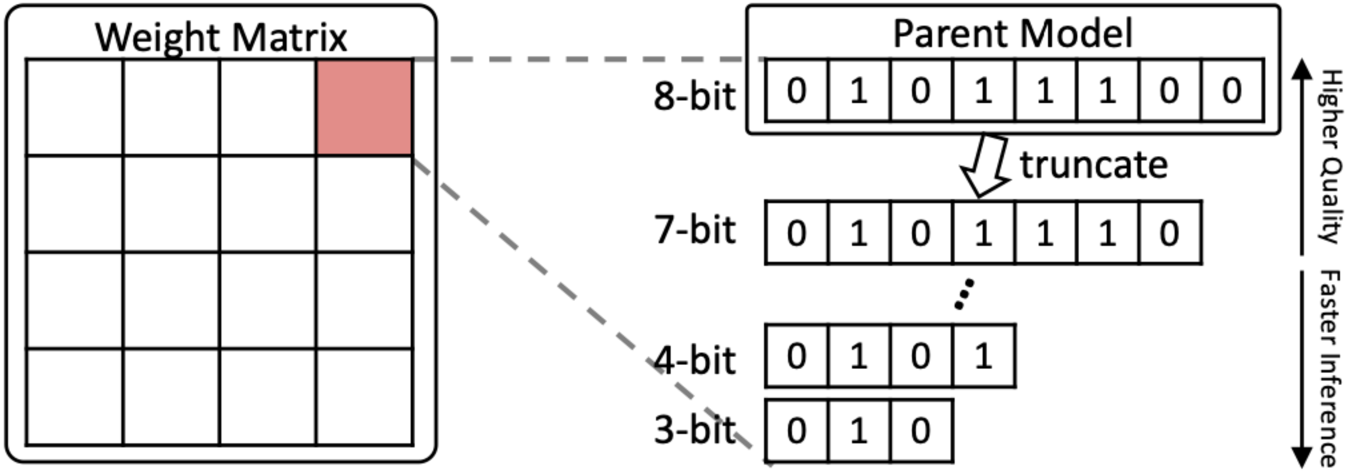
How can we deploy multiple, different-sized LLMs in a memory-efficient way?

Solution: Any-Precision Quantization (Yu et al., 2021[1])



[1] Yu et al,. "Any-Precision Deep Neural Networks", AAAI'21

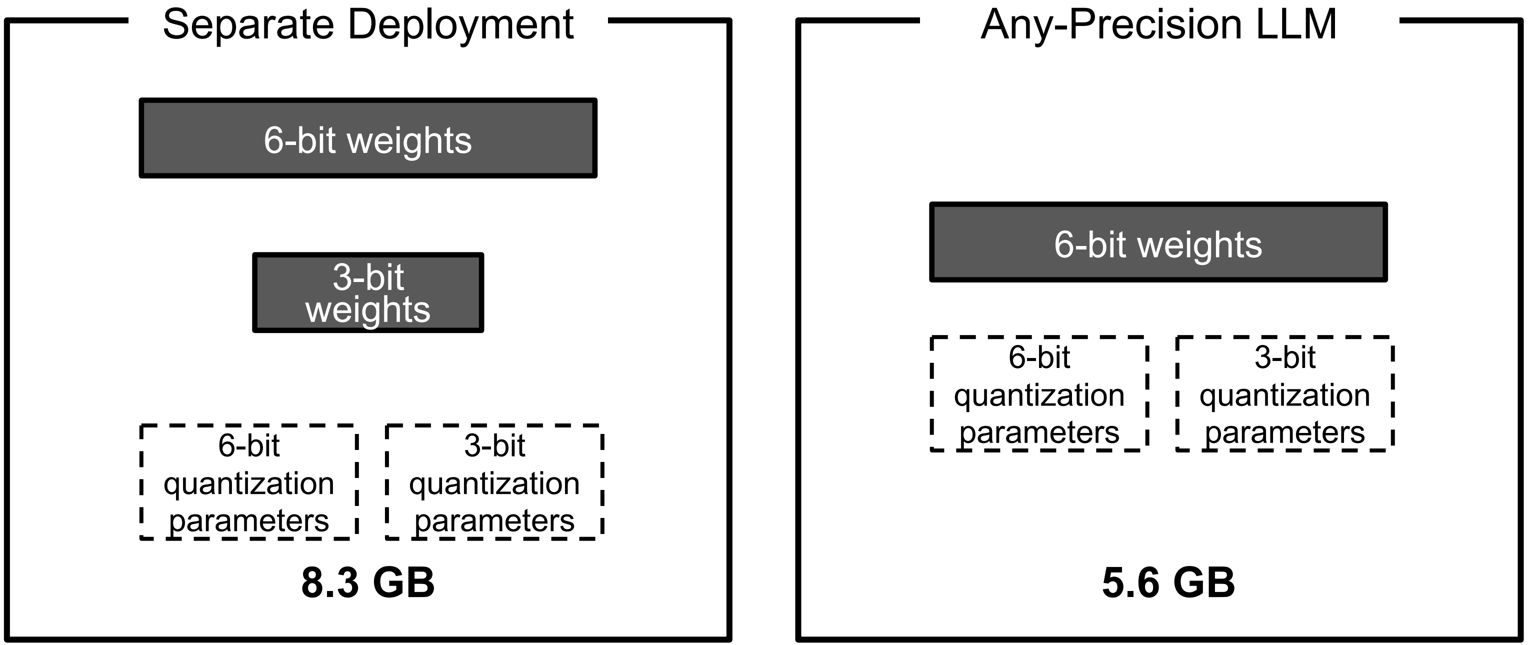
Solution: Any-Precision Quantization (Yu et al., 2021[1])



[1] Yu et al,. "Any-Precision Deep Neural Networks", AAAI'21

Potential Memory Saving of Any-Precision LLM (Llama-2-7B)

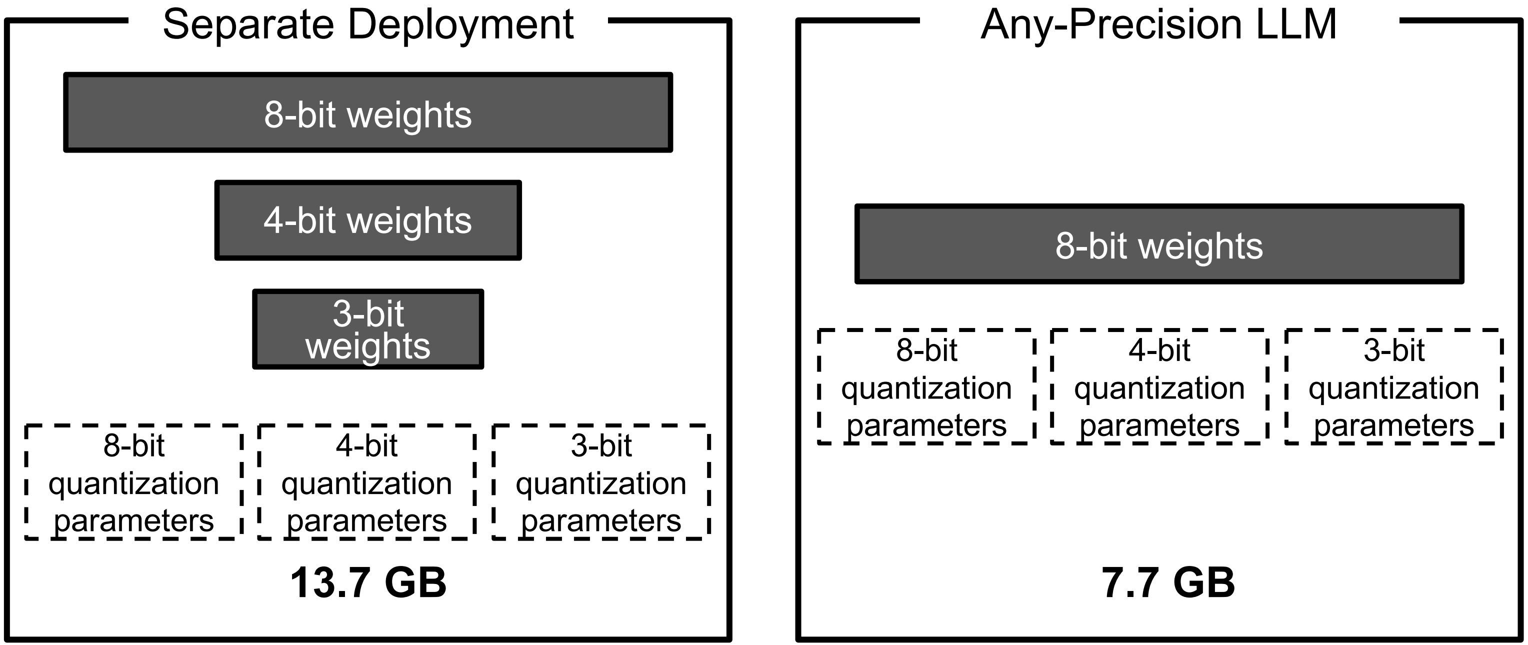
Supported bit-widths: {3, 6}



1.49X saving

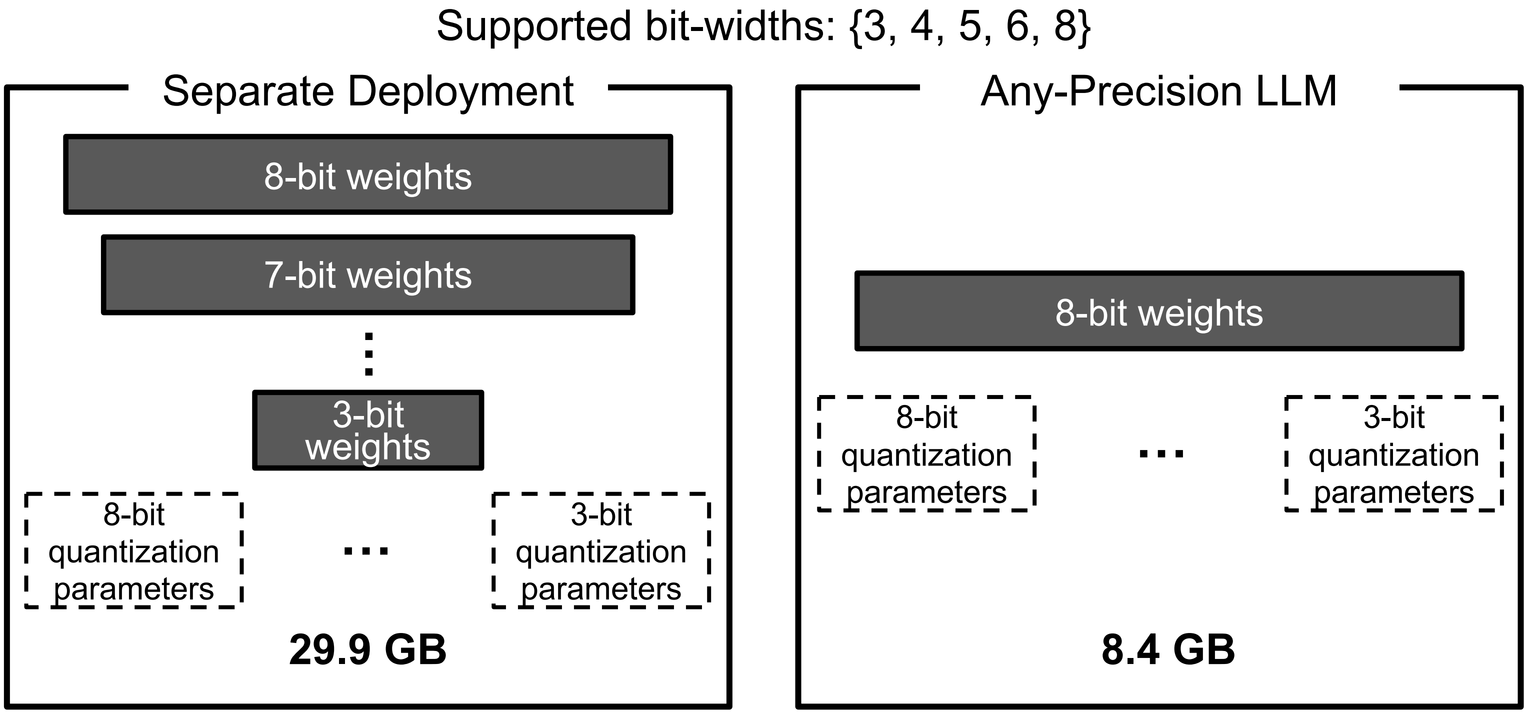
Potential Memory Saving of Any-Precision LLM (Llama-2-7B)

Supported bit-widths: {3, 4, 8}



1.76X saving

Potential Memory Saving of Any-Precision LLM (Llama-2-7B)



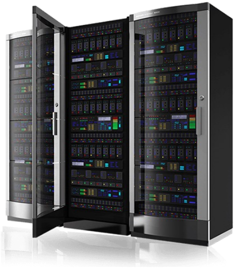
3.56X saving

Challenges of Any-Precision LLM

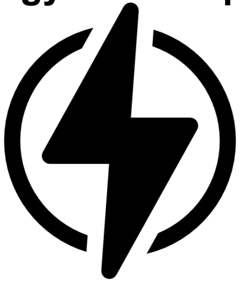
① High Training Cost

Original work adopts **QAT (quantization-aware training)** scheme for any-precision quantization

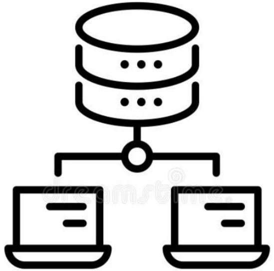
Compute Cluster



Energy Consumption



Dataset



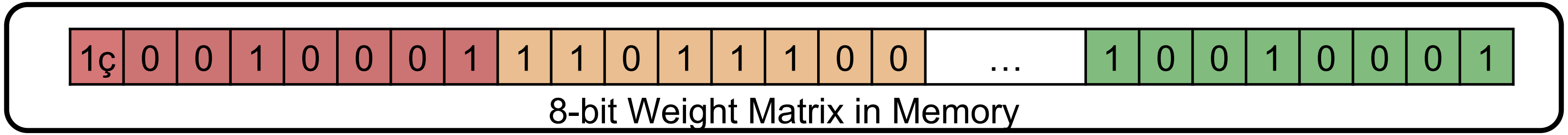
Not affordable to most users

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Challenges of Any-Precision LLM

② Memory Bandwidth Saving

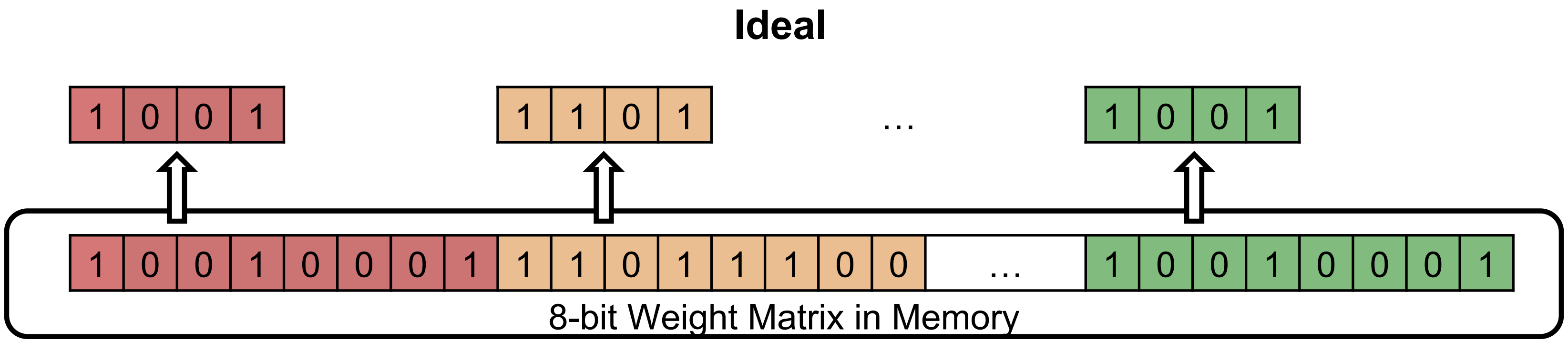
Original work does not save any memory bandwidth, resulting in **no performance improvement for LLM**



Challenges of Any-Precision LLM

② Memory Bandwidth Saving

Original work does not save any memory bandwidth, resulting in **<u>no performance improvement for LLM</u>**

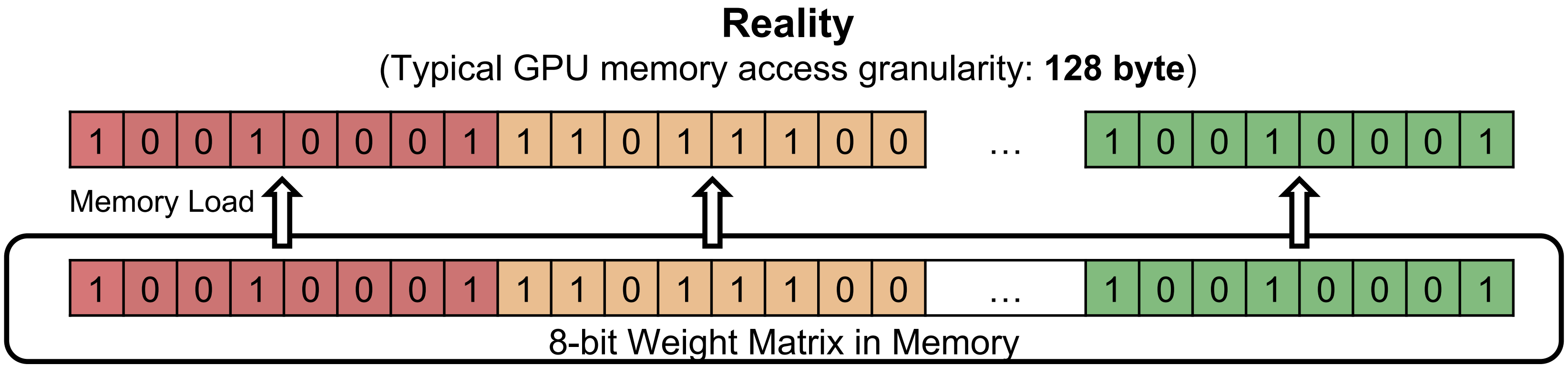


Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Challenges of Any-Precision LLM

② Memory Bandwidth Saving

Original work does not save any memory bandwidth, resulting in **no performance improvement for LLM**



Department of CSE
Seoul National University

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Low-Cost Deployment of Multiple, Different-Sized LLMs

We make a strong case for any-precision quantization of LLM that **does not require training** and leads to **real end-to-end inference speed-ups**.

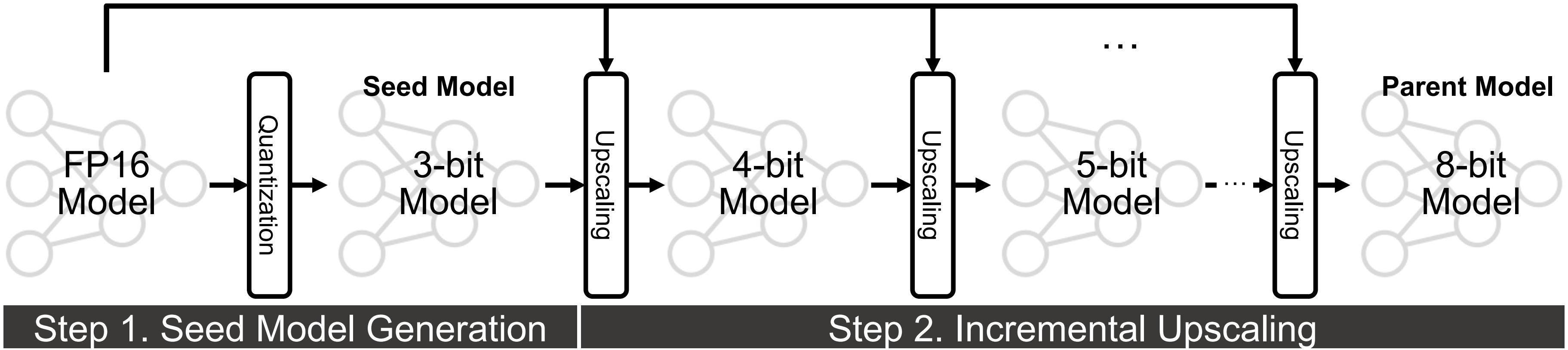
- **Lightweight method for any-precision quantization of LLMs** leveraging post-training quantization (PTQ) framework, called incremental upscaling
- **Specialized software engine** for efficient serving of any-precision LLM adopting a bitplane-based memory layout

Low-Cost Deployment of Multiple, Different-Sized LLMs

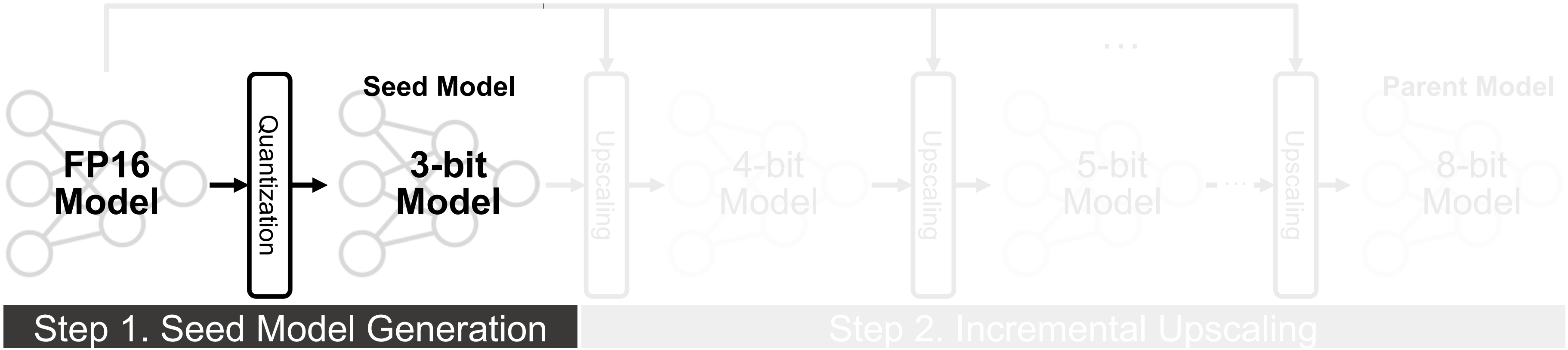
We make a strong case for any-precision quantization of LLM that **does not require training** and leads to **real end-to-end inference speed-ups**.

- **Lightweight method for any-precision quantization of LLMs** leveraging post-training quantization (PTQ) framework, called incremental upscaling
- **Specialized software engine** for efficient serving of any-precision LLM adopting a bitplane-based memory layout

Any-Precision Quantization with Incremental Upscaling



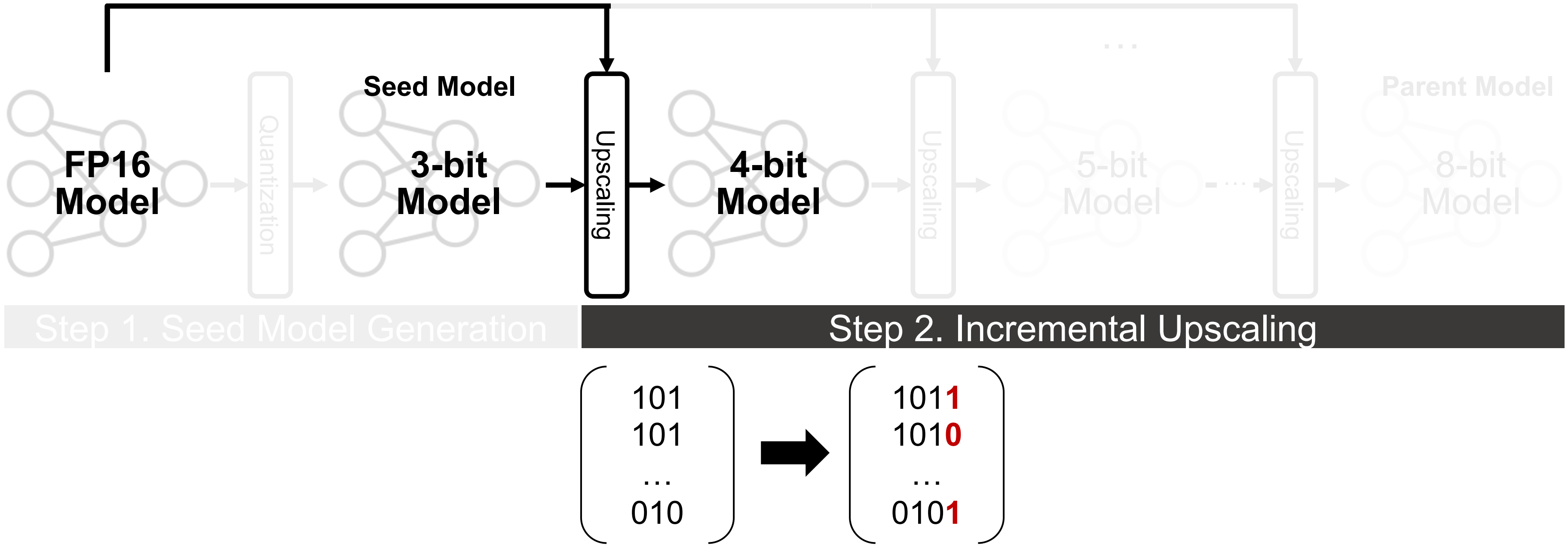
Any-Precision Quantization with Incremental Upscaling



$$\begin{pmatrix} 0.58 \\ 0.47 \\ \cdots \\ -0.12 \end{pmatrix} \rightarrow \begin{pmatrix} 101 \\ 101 \\ \cdots \\ 010 \end{pmatrix}$$

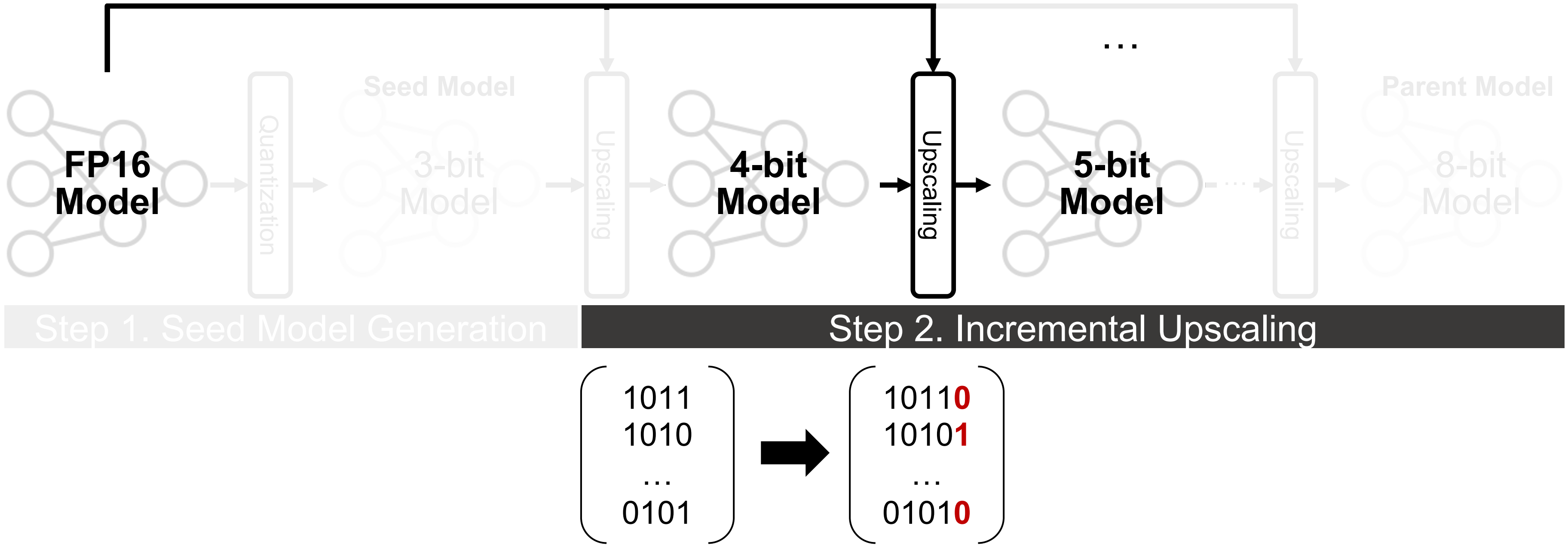
Quantize the model to the minimum supported bit-width (3-bit)

Any-Precision Quantization with Incremental Upscaling



Upscale 3-bit model to 4-bit model

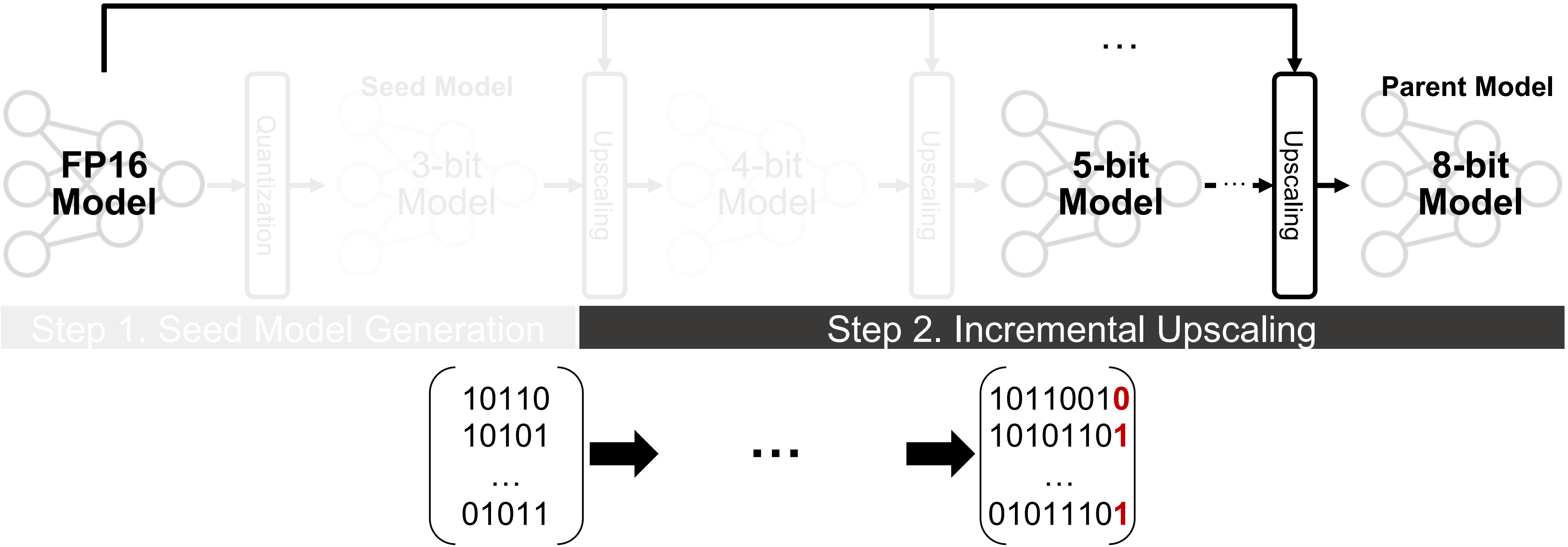
Any-Precision Quantization with Incremental Upscaling



Upscale 4-bit model to 5-bit model

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Any-Precision Quantization with Incremental Upscaling



Upscale upto 8-bit model (parent model)

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Selection of Backbone Quantization Method

- A particular quantization method must be adopted as a backbone for **seed model generation** and **incremental upscaling (IU)**

Which Quantization Method to Use?

- A particular quantization method must be adopted as a backbone for **seed model generation** and **incremental upscaling (IU)**

Requirement #1. Low-Cost
No training; PTQ (post-training quantization) methods

Which Quantization Method to Use?

- A particular quantization method must be adopted as a backbone for **seed model generation** and **incremental upscaling (IU)**

Requirement #1. Low-Cost
No training; PTQ (post-training quantization) methods

Requirement #2. High-Performance
Demonstrate SOTA results in terms of PPL/downstream task evaluation

Which Quantization Method to Use?

- A particular quantization method must be adopted as a backbone for **seed model generation** and **incremental upscaling (IU)**

Requirement #1. Low-Cost
No training; PTQ (post-training quantization) methods

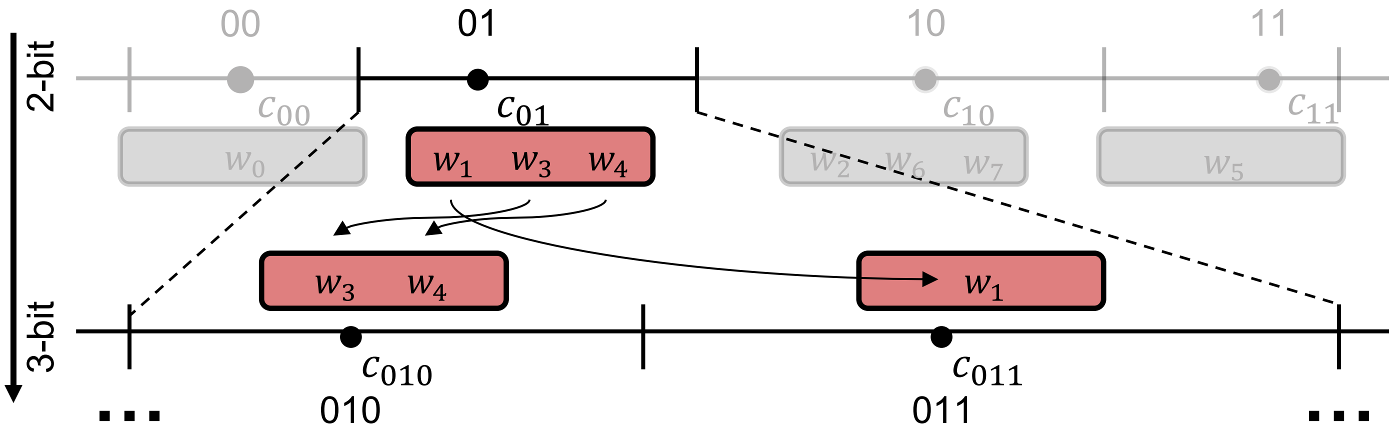
Requirement #2. High-Performance
Demonstrate SOTA results in terms of PPL/downstream task evaluation

Requirement #3. Compatible with IU
Easily extended to support IU

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Incremental Upscaling with SqueezeLLM

* SqueezeLLM: **non-uniform** quantization by **weighted K-means clustering**



Divide each cluster into two sub-clusters by **weighted K-means clustering**

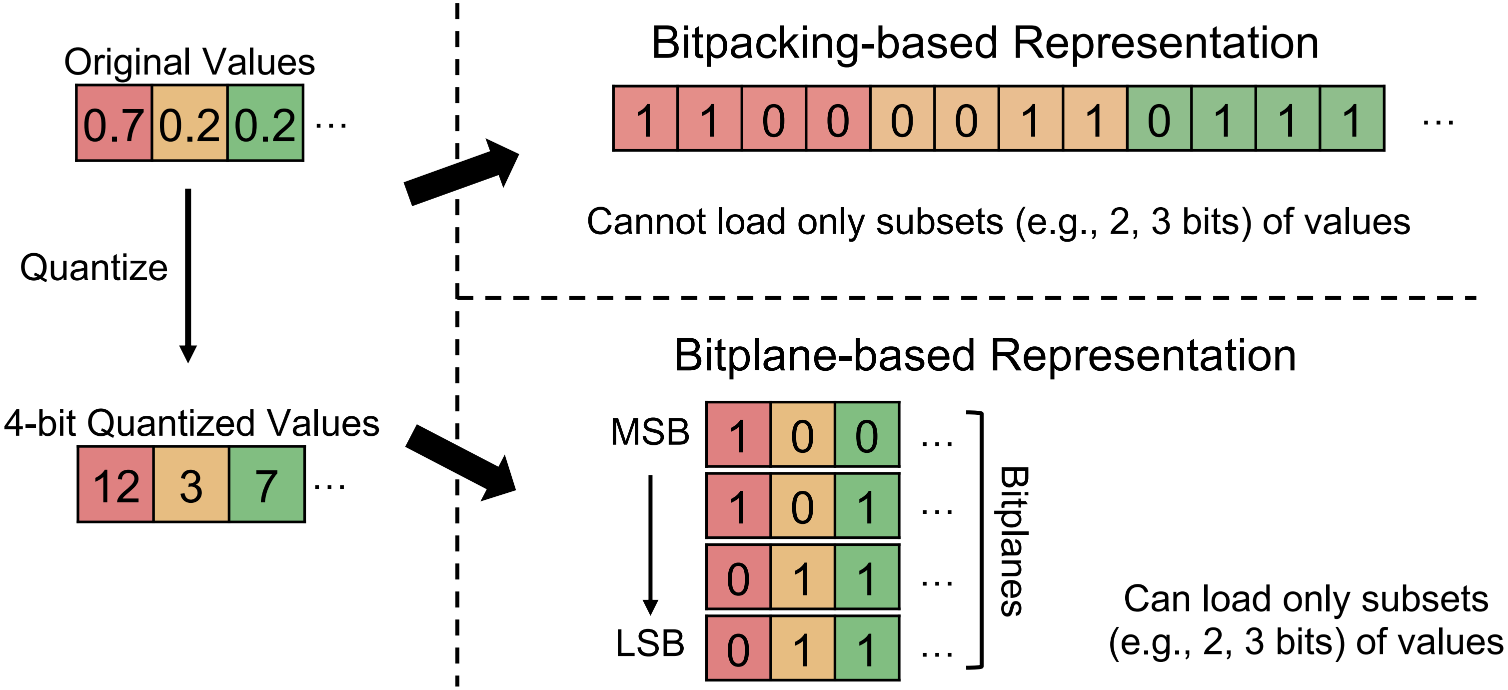
Y. Park et al., “Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM”

Low-Cost Deployment of Multiple, Different-Sized LLMs

We make a strong case for any-precision quantization of LLM that **does not require training** and leads to a **real end-to-end inference speed-up**.

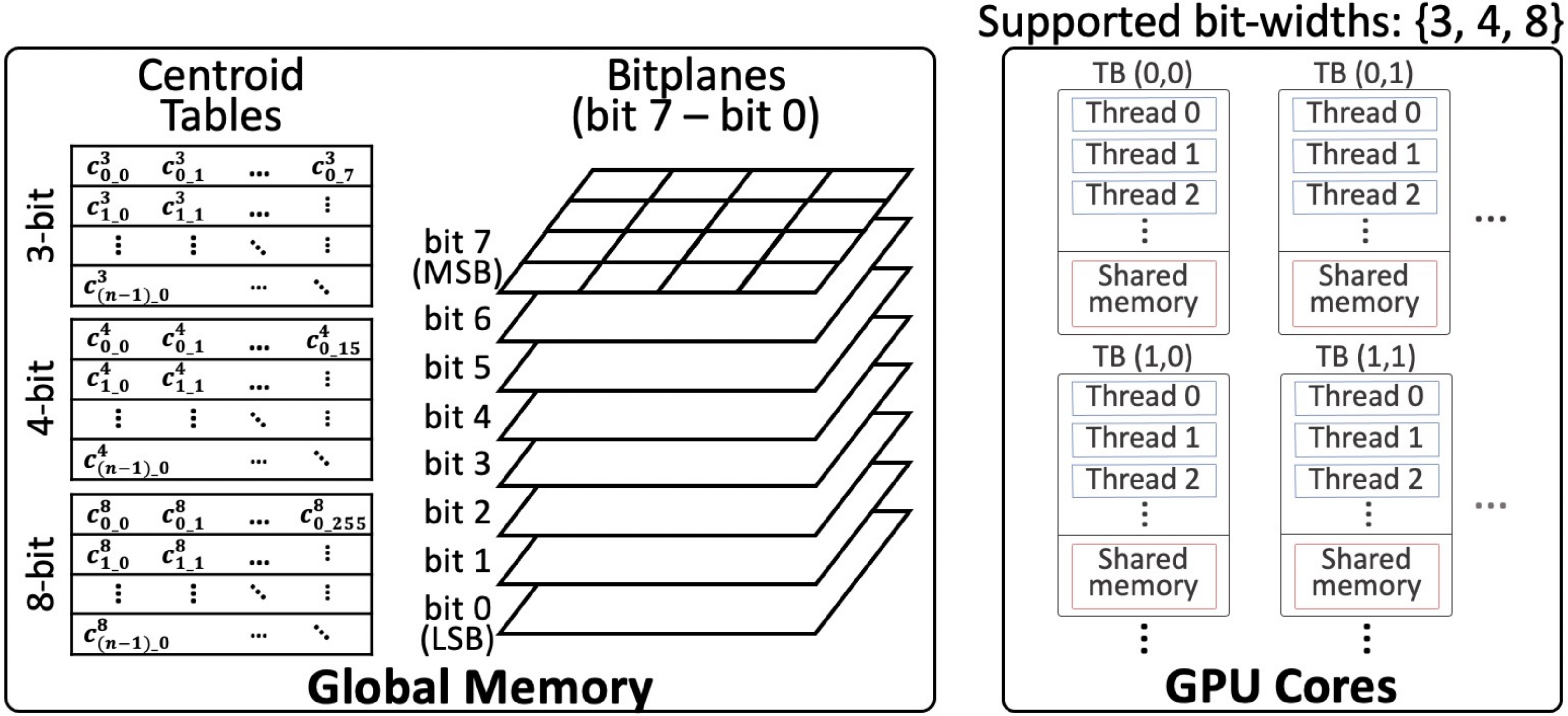
- **Lightweight method for any-precision quantization of LLMs** leveraging post-training quantization (PTQ) framework, called incremental upscaling
- **Specialized software engine** for efficient serving of any-precision LLM adopting a bitplane-based memory layout

Bitpacking-based vs Bitplane-based



Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Overview of Our Specialized Engine

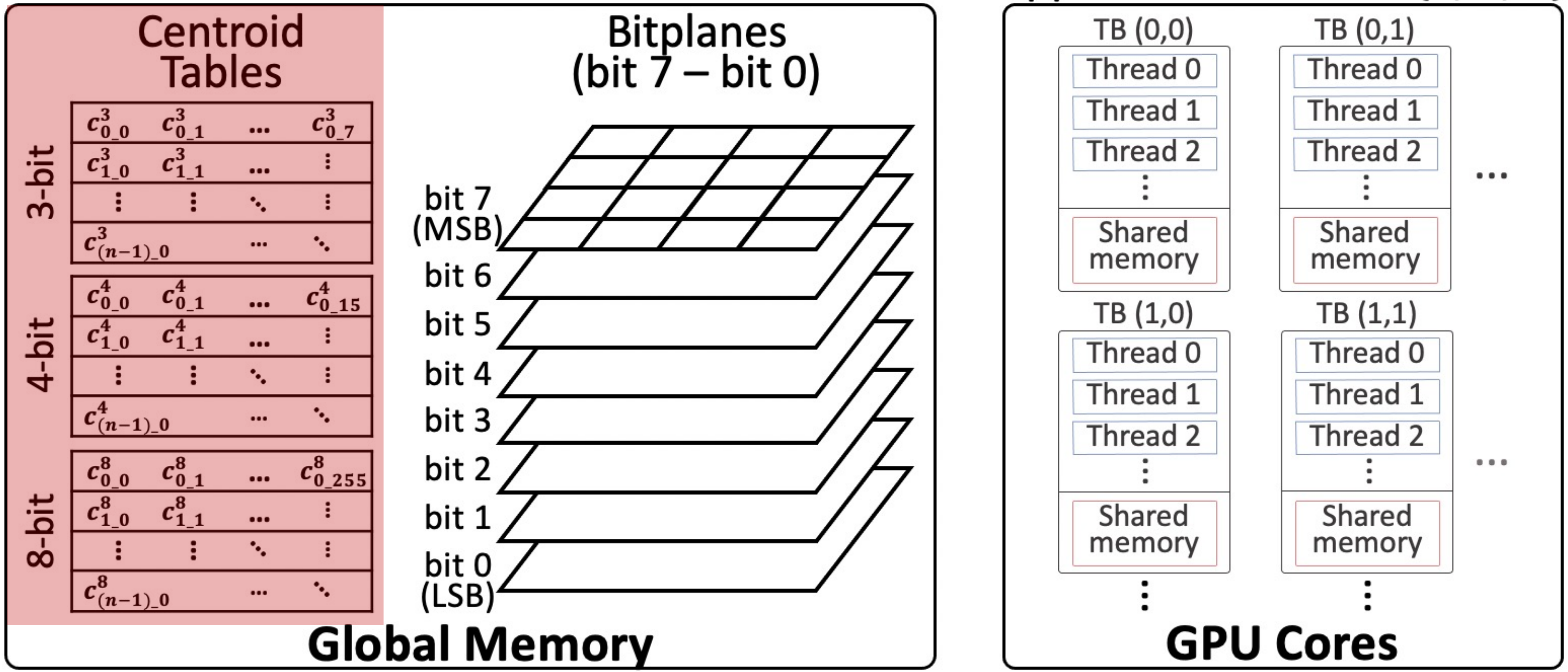


Department of CSE
Seoul National University

Y. Park et al., “Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM”

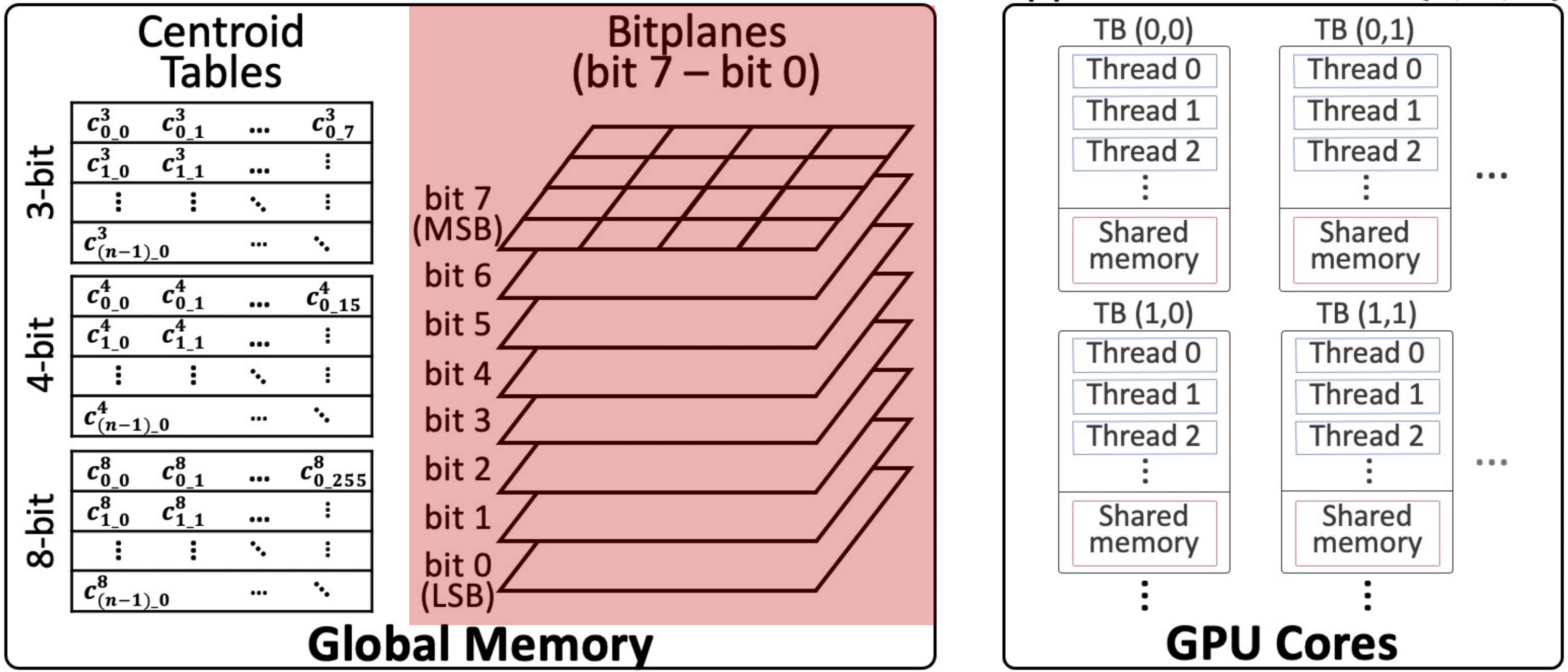
Overview of Our Specialized Engine

Supported bit-widths: {3, 4, 8}

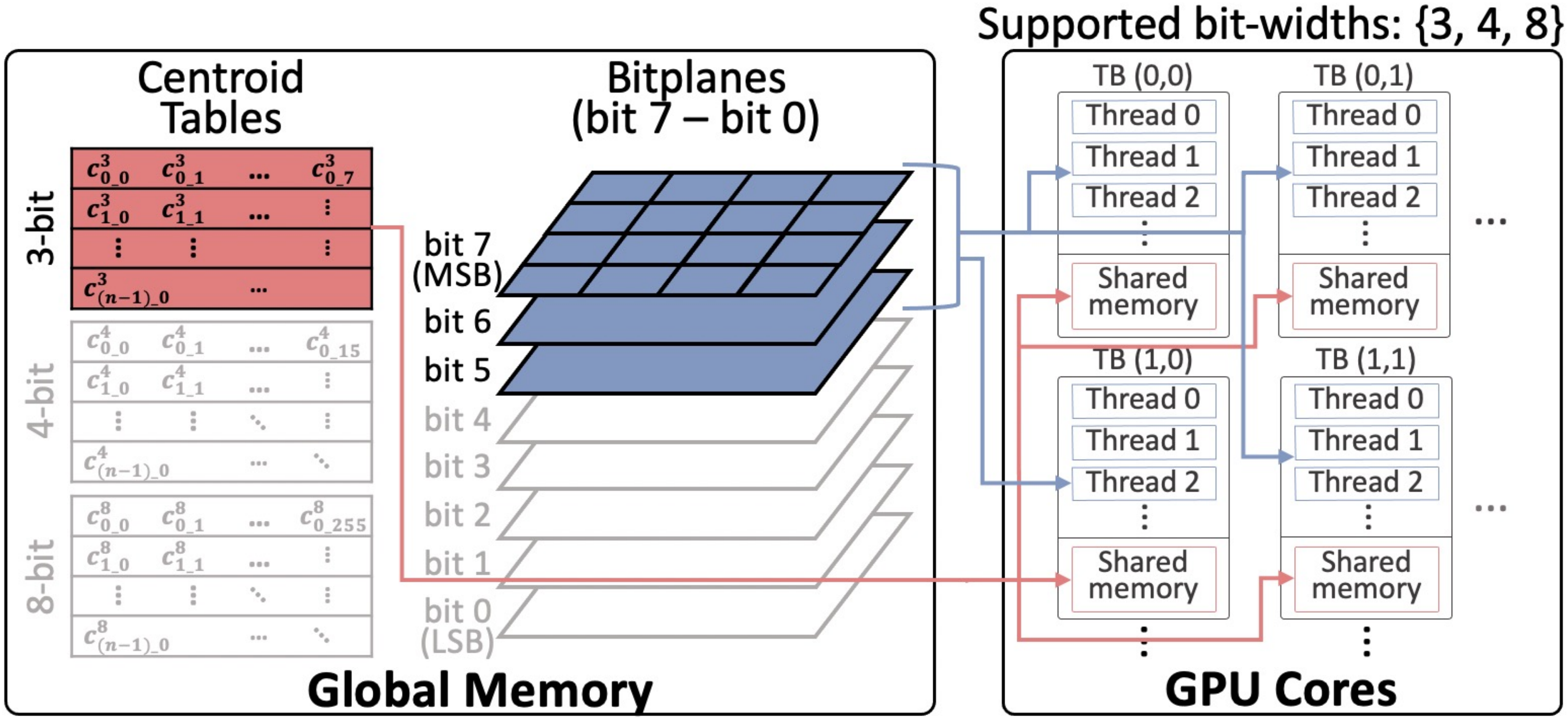


Overview of Our Specialized Engine

Supported bit-widths: {3, 4, 8}



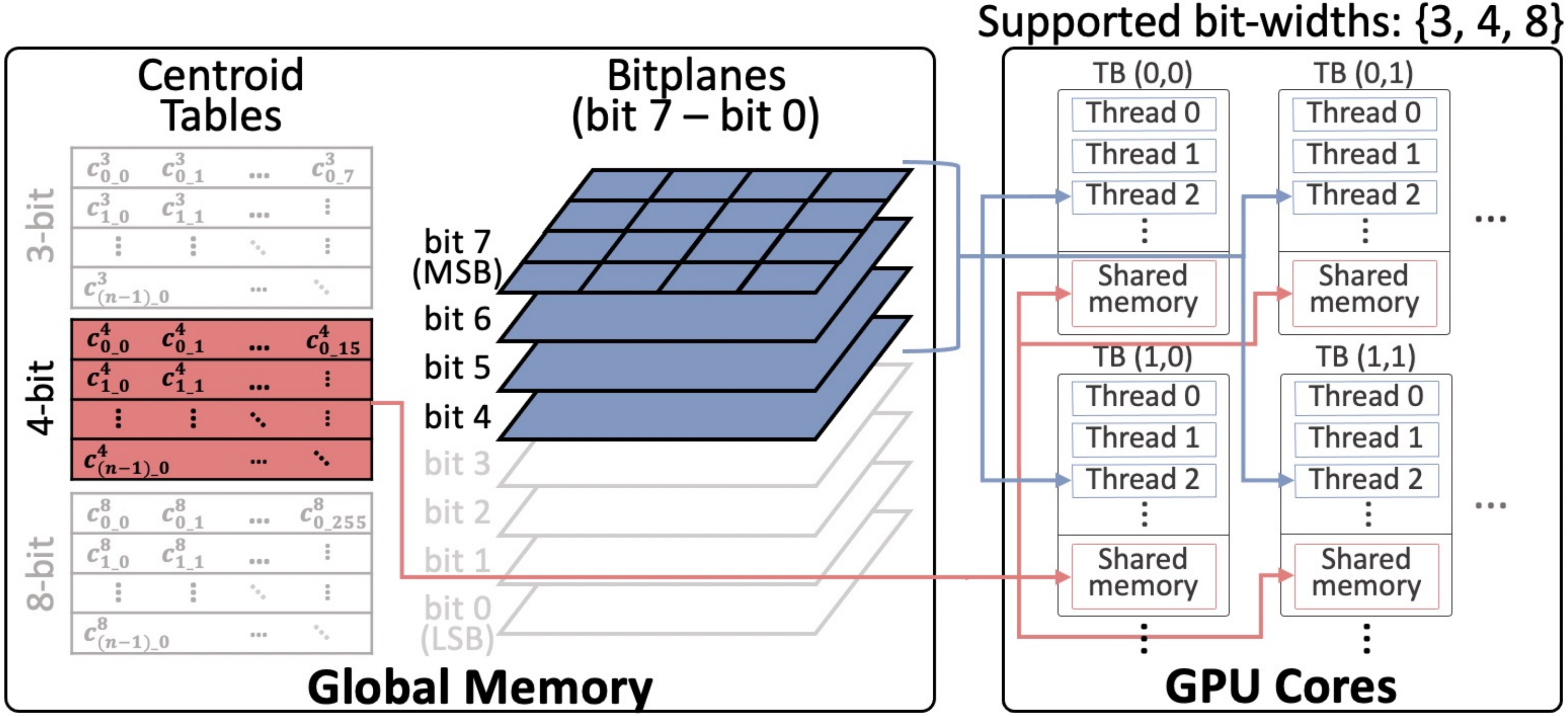
Overview of Our Specialized Engine



Selected bit-width: 3

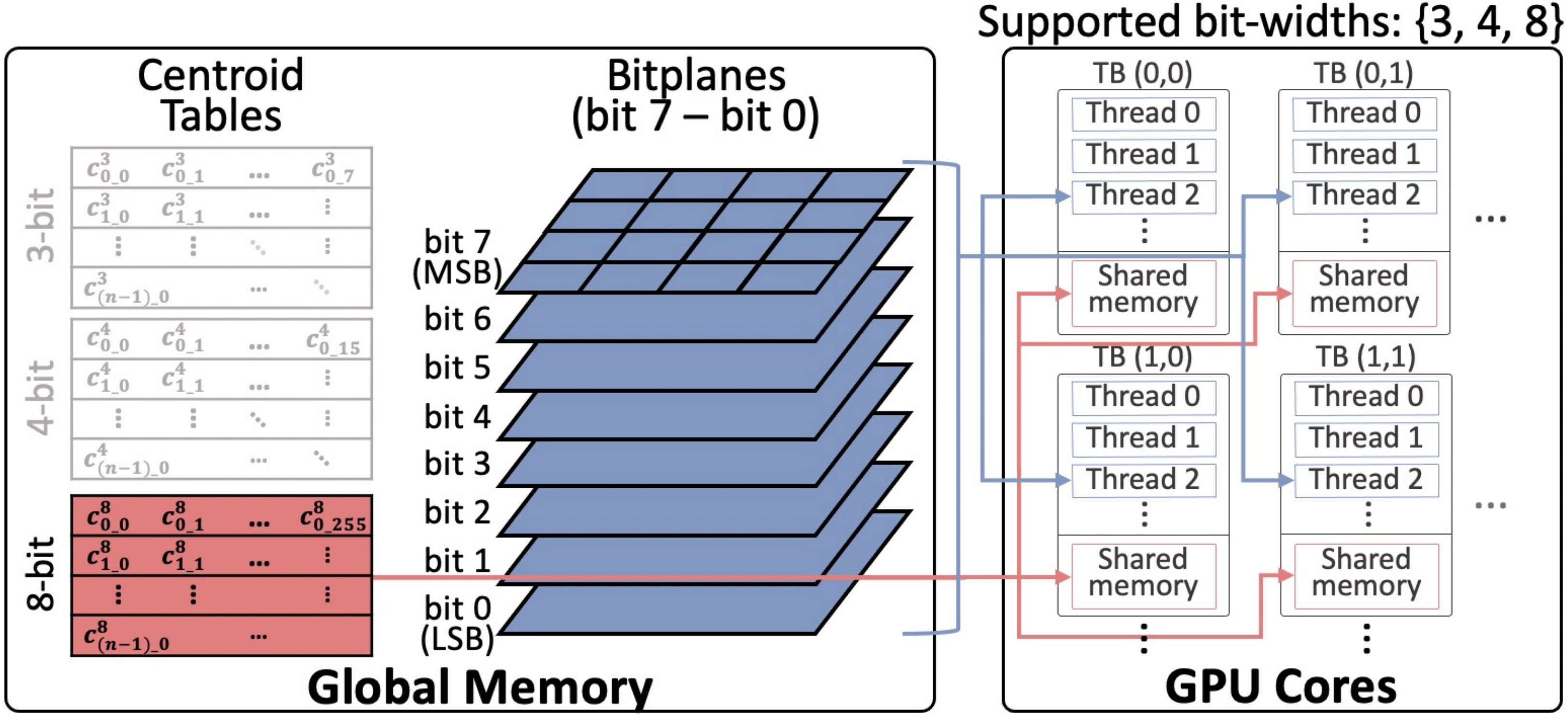
Department of CSE
Seoul National University

Overview of Our Specialized Engine



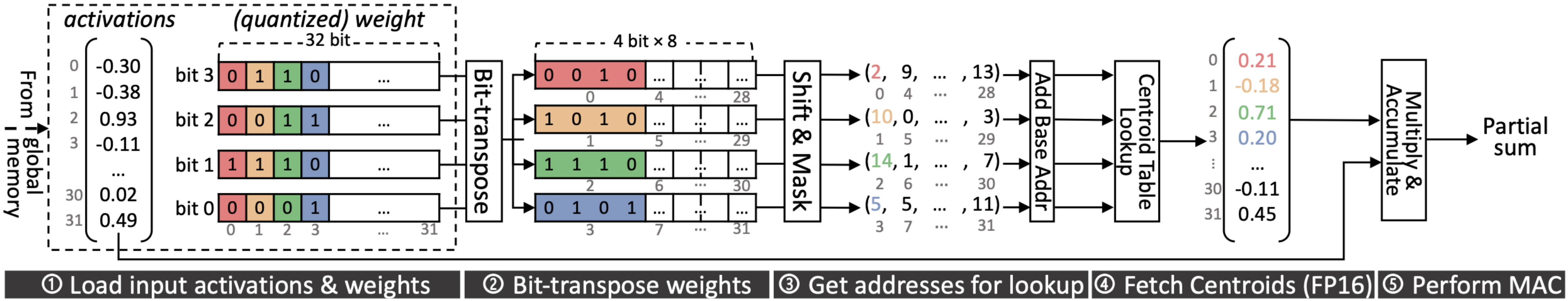
Selected bit-width: 4

Overview of Our Specialized Engine



Selected bit-width: 8

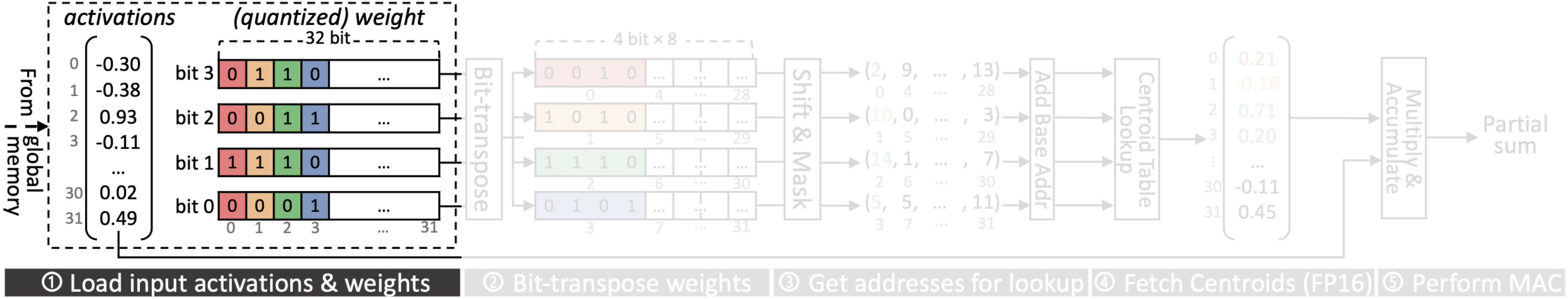
Thread-Level Operations



Five step thread-level operations assuming a bit-width of 4

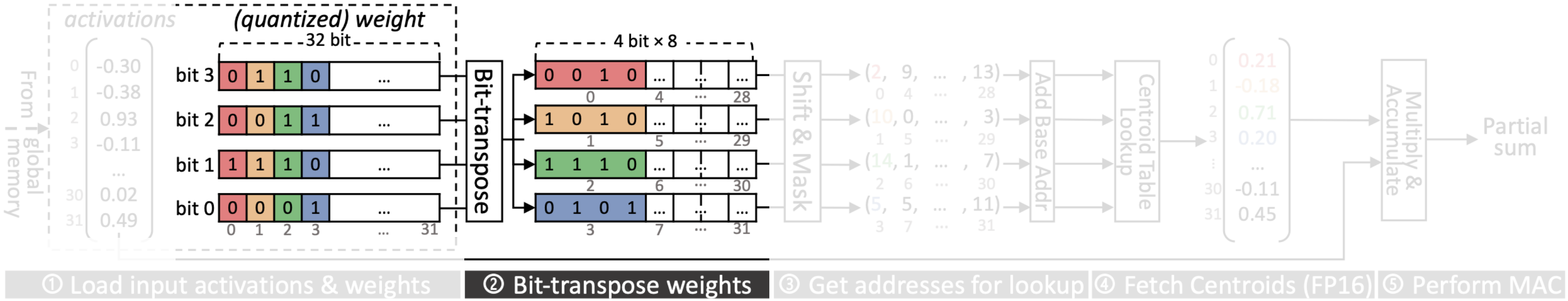
Y. Park et al., “Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM”

Thread-Level Operations



Load **input activations and weights** from global memory

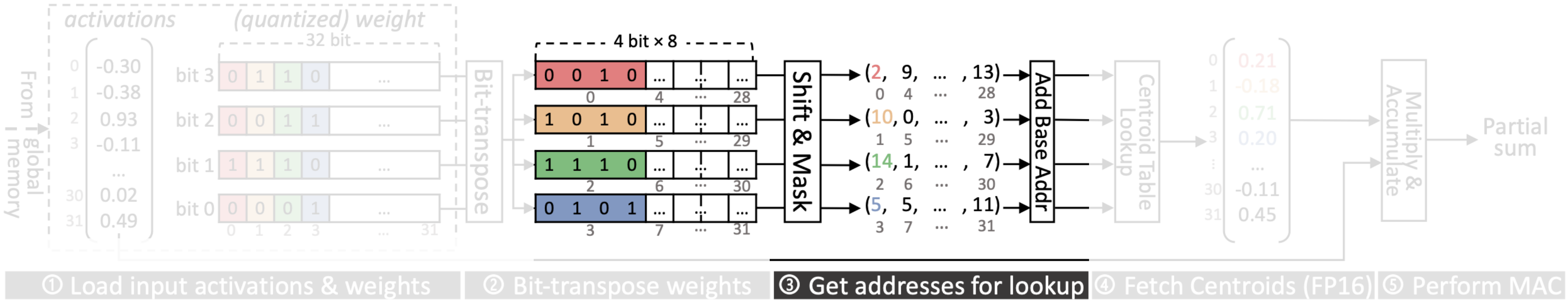
Thread-Level Operations



Align the bits of each weight contiguously
(bit-level transpose)

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Thread-Level Operations



Get addresses for centroid table lookup by

1) Shift & Mask
2) Add offset (base address)

Department of CSE
Seoul National University

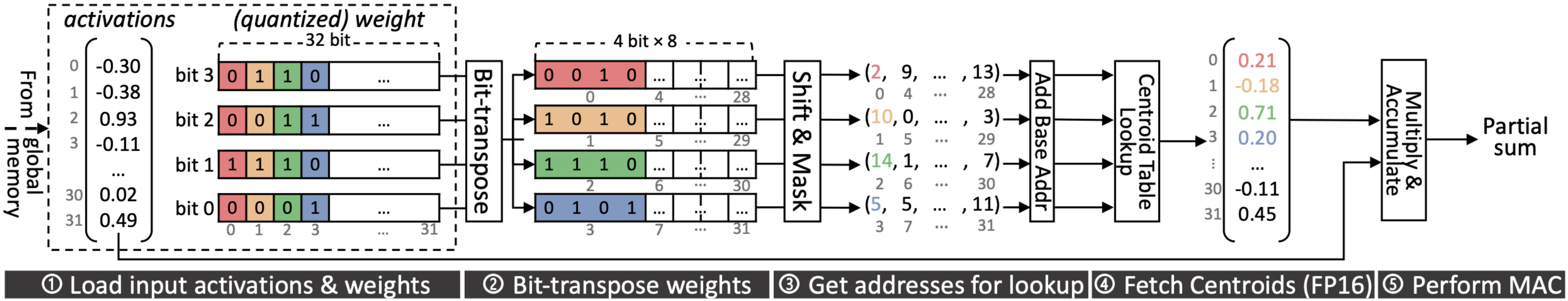
Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Thread-Level Operations

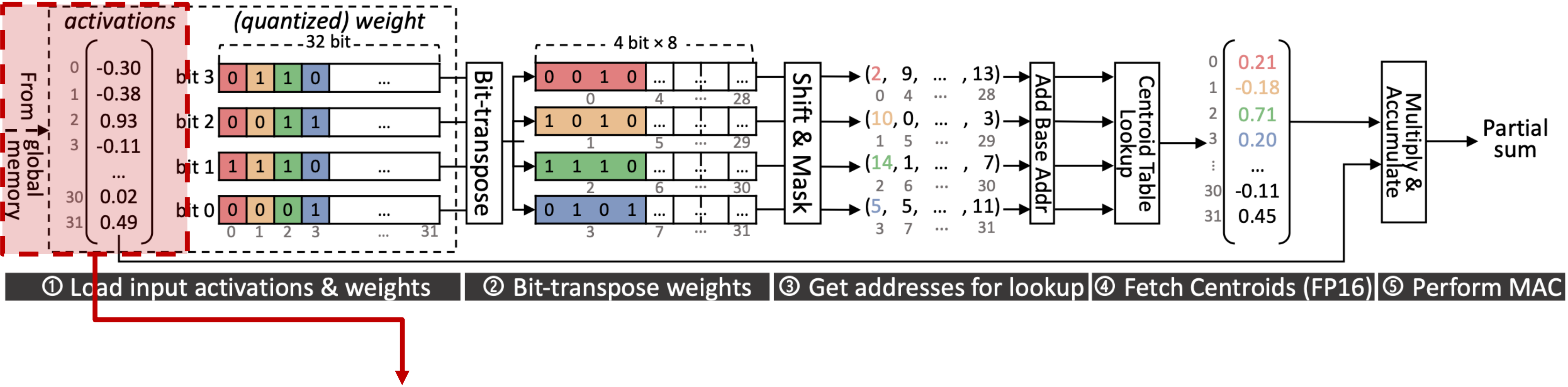
Fetch centroids and perform MAC

Department of CSE
Seoul National University

GPU Kernel Optimization



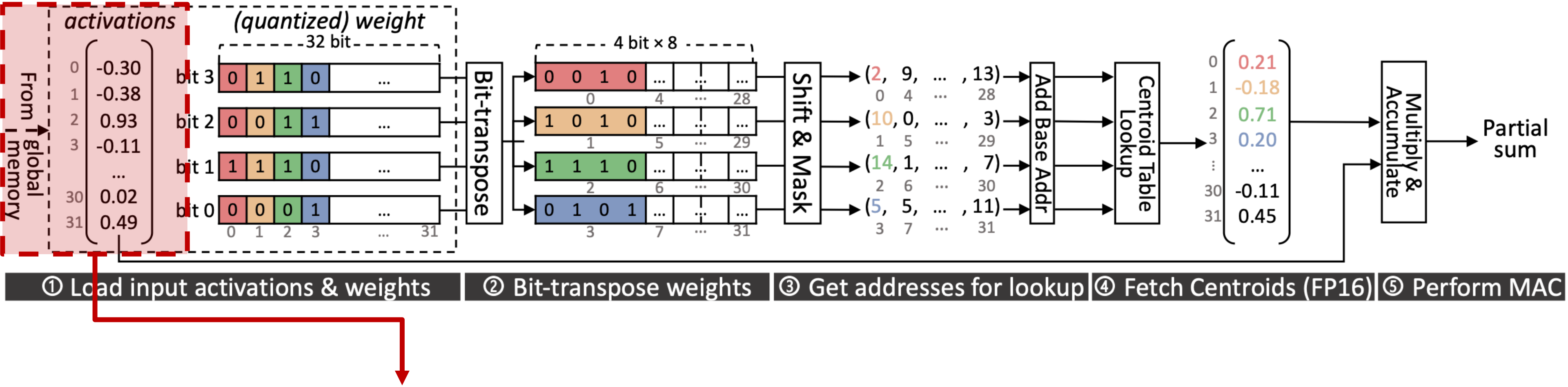
GPU Kernel Optimization



Challenge ①

Uncoalesced memory access

GPU Kernel Optimization

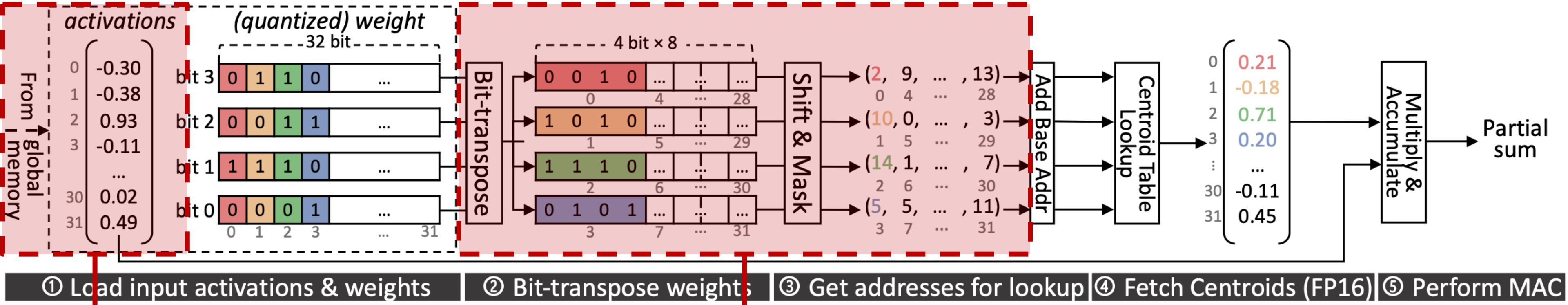


Challenge ①

<u>Uncoalesced memory access</u>

-> *Weight bitplane layout optimization*

GPU Kernel Optimization



Challenge ①

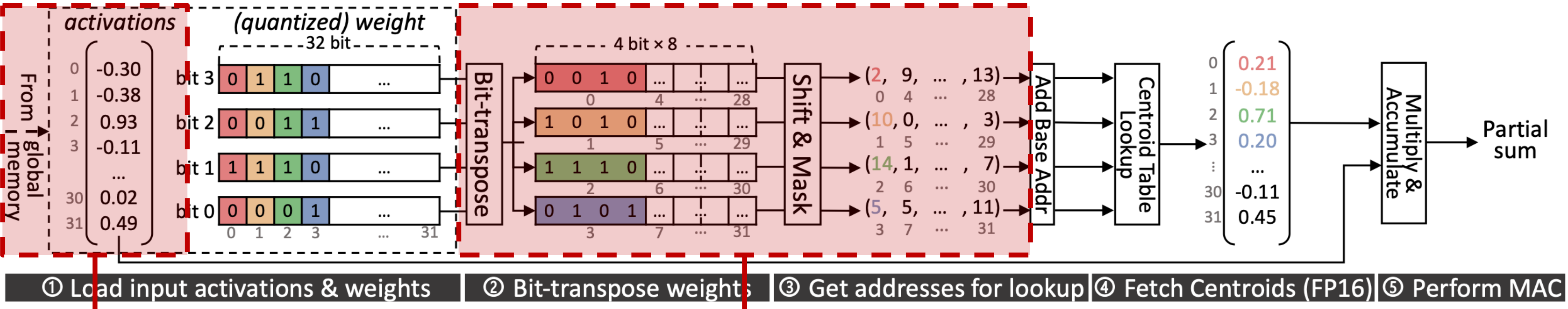
<u>Uncoalesced memory access</u>

-> *Weight bitplane layout optimization*

Challenge ②

<u>Bitwise operation overhead</u>

GPU Kernel Optimization



Challenge ①

Uncoalesced memory access

-> *Weight bitplane layout optimization*

Challenge ②

Bitwise operation overhead

-> *Efficient bit-transpose*

-> *Table lookup merge*

Evaluation

Evaluation

- **Key Result #1:**

  A set of quantized models generated from incremental upscaling match the SOTA quantization results

- **Key Result #2:**

  Our engine matches or even outperforms existing engines while providing memory-efficient any-precision support

Evaluation

- **Key Result #1:**
  A set of quantized models generated from incremental upscaling match the SOTA quantization results

- **Key Result #2:**
  Our engine matches or even outperforms existing engines while providing memory-efficient any-precision support

Perplexity on Wikitext-2

		3-bit	4-bit	5-bit	6-bit	7-bit	8-bit	FP16
	SqLLM	6.13	5.61	5.50	5.47	5.47	5.47	
Llama-2-7B	SqLLM+IU		5.62	5.50	5.47	5.47	5.47	5.47
	Δ		↑ **0.01**	-	-	-	-	

Upscaled models (**SqLLM+IU**) match independently quantized models (**SqLLM**)

Similar trends observed for *other datasets (PTB, C4)* and *other models (Mistral, OPT)*

Zero-shot Task Accuracy

* Average accuracy across five tasks (ARC-easy/challenge, HellaSwag, PIQA, WinoGrande)

		3-bit	4-bit	5-bit	6-bit	7-bit	8-bit	FP16
	SqLLM	66.2	68.3	68.6	68.8	68.9	68.9	
Llama-2-7B	SqLLM+IU		68.2	68.8	68.9	68.9	69.0	69.0
	Δ		↓ **0.01**	↑ **0.02**	↑ **0.01**	-	↑ **0.01**	

Upscaled models (**SqLLM+IU**) match independently quantized models (**SqLLM**)

Similar trends observed for *other models (Mistral, OPT)*

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Evaluation

- **Key Result #1:**
  A set of quantized models generated from incremental upscaling match the SOTA quantization results

- **Key Result #2:**
  Our engine matches or even outperforms existing engines while providing memory-efficient any-precision support

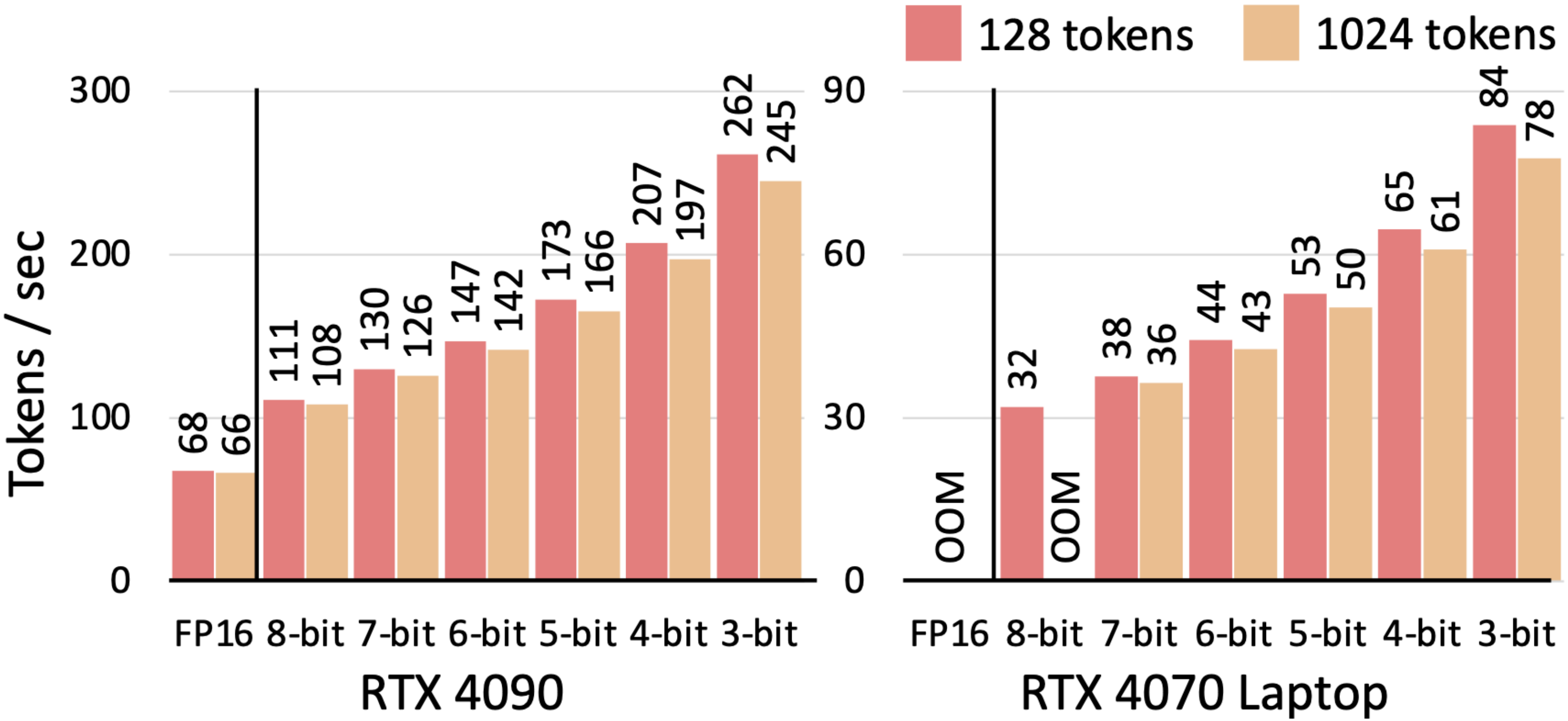
Kernel Microbenchmark (GEMV)

* Speedup over cuBLAS FP16 baseline

		(1, 4096) x (4096, 4096)						(1, 11008) x (4096, 11008)					
		3-bit	4-bit	5-bit	6-bit	7-bit	8-bit	3-bit	4-bit	5-bit	6-bit	7-bit	8-bit
RTX 4090	Ours	3.99×	3.03×	2.61×	2.18×	1.87×	1.56×	4.45×	3.42×	2.74×	2.28×	2.08×	1.81×
	SqLLM	3.69×	3.07×	-	-	-	-	4.22×	3.17×	-	-	-	-
RTX 4070 Laptop	Ours	4.97×	3.73×	3.01×	2.51×	2.10×	1.76×	5.29×	3.66×	3.05×	2.52×	2.13×	1.87×
	SqLLM	4.74×	3.66×	-	-	-	-	5.29×	3.93×	-	-	-	-
Jetson AGX Orin	Ours	3.84×	3.02×	2.56×	2.33×	2.10×	1.78×	4.35×	2.96×	2.54×	2.52×	2.16×	1.86×
	SqLLM	3.15×	1.94×	-	-	-	-	3.36×	2.04×	-	-	-	-

Our kernel outperforms for the most cases even with any-precision support

End-to-End Speedup



Generation throughput of Llama-2-7B using our kernel integrated with TensorRT-LLM

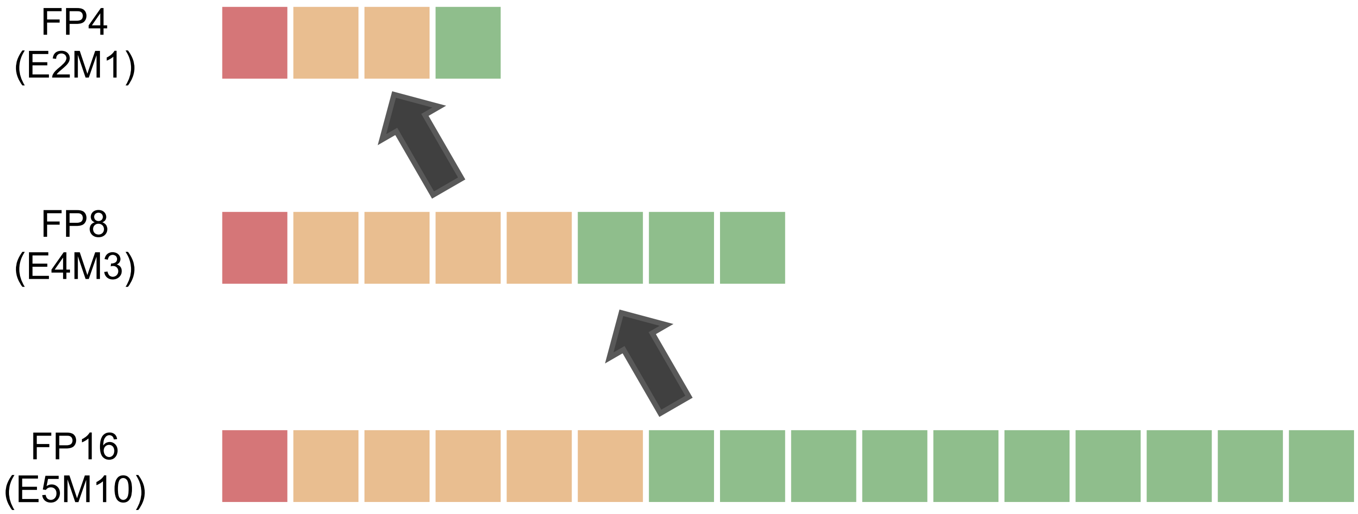
What's Next?

Next Step: Any-Precision LLM for Datacenter Settings

Single Batch	→	Large Batch
Memory-Bound	→	Compute-Bound
Custom Number Formats	→	FP16, FP8, FP4

Y. Park et al., "Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM"

Next Step: Any-Precision LLM for Datacenter Settings



Y. Park et al., “Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLM”

Summary

Any-precision LLM is a memory-efficient and cost-effective solution for deployment of multiple, different sized LLMs.

- **Reduce memory cost** of deployment of multiple LLMs by any-precision quantization
- Propose **lightweight any-precision quantization scheme** for LLMs which produces quantized models with SOTA results
- Propose **specialized software engine** for efficient serving of any-precision LLMs
- Currently exploring ways to extend this concept to datacenter settings

Paper

Code