

Oaken: Fast and Efficient LLM Serving with Online-Offline Hybrid KV Cache Quantization

Minsu Kim*

Seongmin Hong*†

RyeoWook Ko

Soongyu Choi

Hunjong Lee†

Junsoo Kim†

Joo-Young Kim†

Jongse Park

KAIST

† HyperAccel

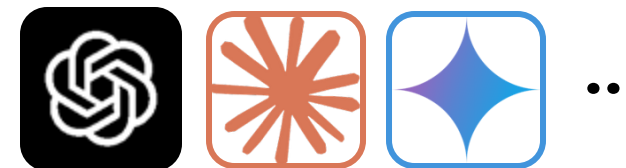
** Co-first authors who contributed equally to this work*



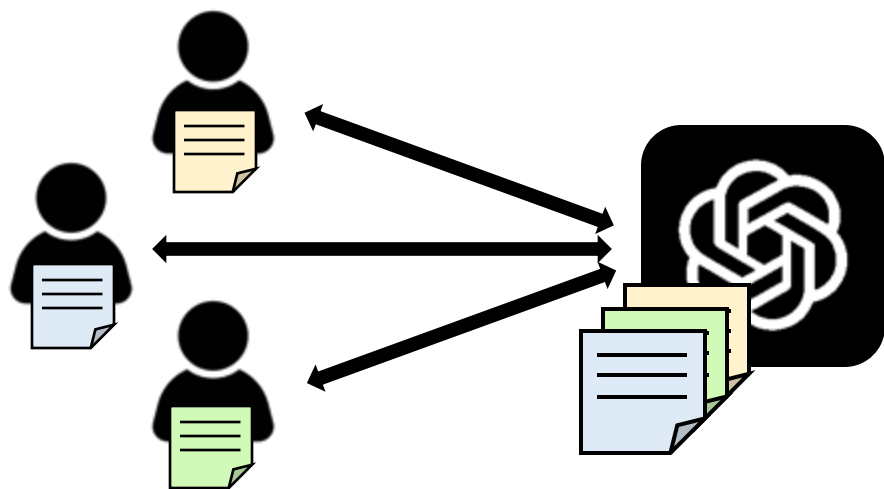
ISCA 2025

LLM Serving at Scale

- LLM serving system should simultaneously handle **a large number of, long-context requests**

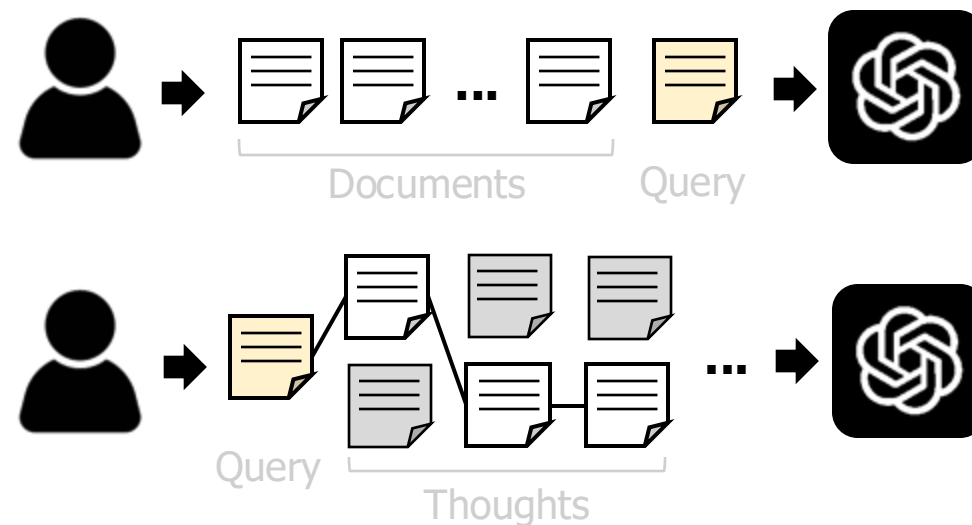


Large Batch Size



LLM serving system batches multiple requests (+10,000) from users

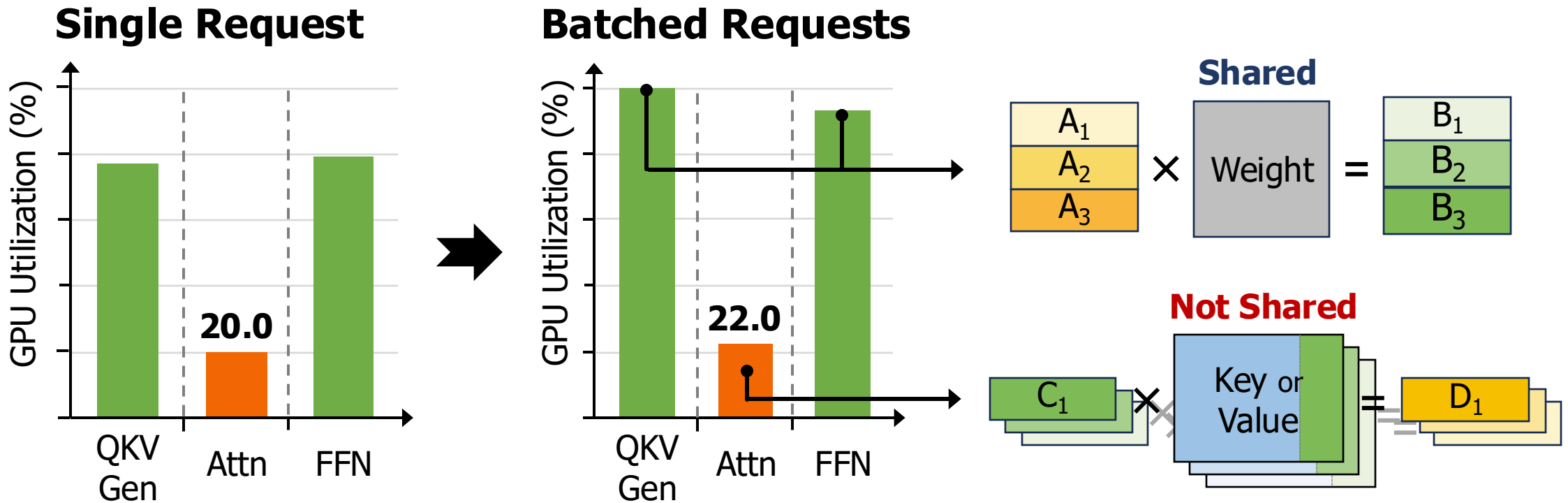
Long Context Length



Recent LLM tasks (*e.g., RAG, reasoning*) involve over tens of thousands of tokens

**Larger Batch & Longer Context put
pressure on Memory Capacity & Bandwidth**

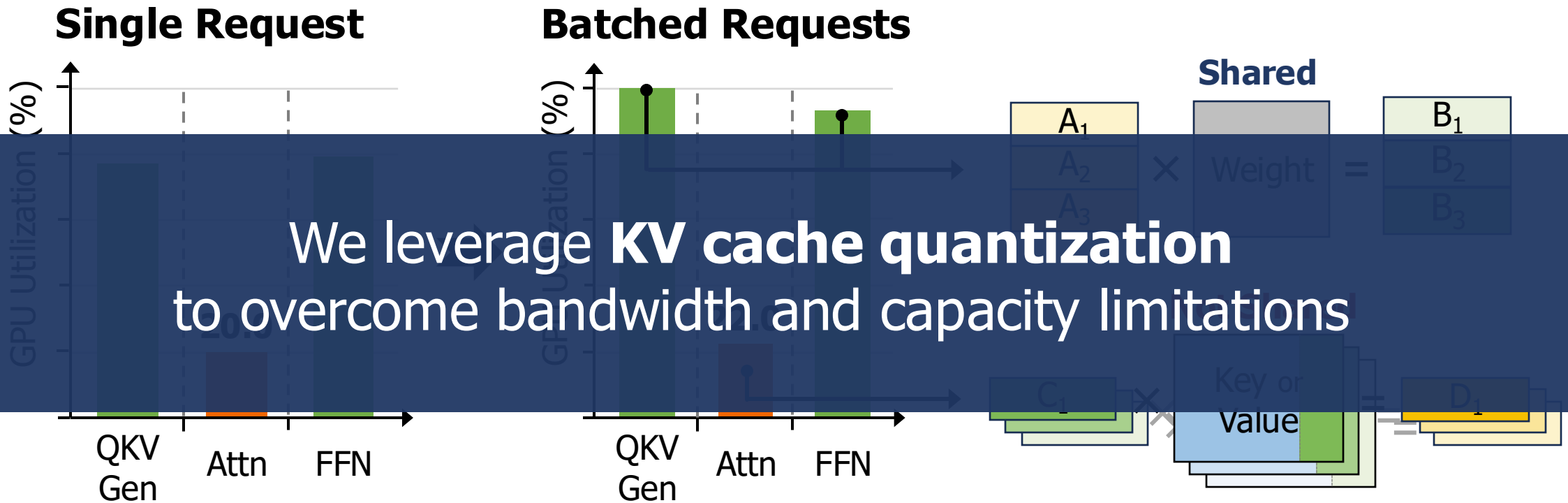
KV Cache Matters for “Bandwidth”



* NVIDIA A100, Llama2-13B, context length: 1K

- Increasing batch size improves utilization **except for attention operation**
- Attention operation is **bandwidth-bound** due to un-sharable KV cache

KV Cache Matters for “Bandwidth”



* NVIDIA A100, Llama2-13B, context length: 1K

- Increasing batch size improves utilization **except for attention operation**
- Attention operation is **bandwidth-bound** due to un-sharable KV cache

Per-Vector Mixed-Precision

**Oaken achieves both high performance & accuracy
through co-designing
quantization algorithm & hardware modules**

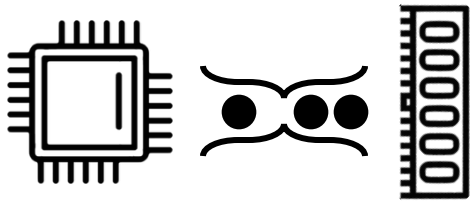
- ✗ accuracy loss
- ✗ Mixed-precision increases complexity
- ✗ Online KV profiling incurs overhead
- ✗ KV profiler & quant. needs 2x

- ✗ Hardware-friendly algorithm with minimal overhead
- ✗ Coarse-grained grouping leads to large accuracy loss
- ✗ 100% 16B, 24, 160B, 180B, 240, 270B, 300B, 360B, 400B, 450B, 512B, 560B, 600B, 672B, 720B, 768B, 800B, 896B, 960B, 1024B, 1120B, 1200B, 1280B, 1344B, 1440B, 1536B, 1600B, 1728B, 1800B, 1920B, 2048B, 2112B, 2240B, 2304B, 2400B, 2560B, 2688B, 2800B, 2880B, 3000B, 3136B, 3200B, 3456B, 3600B, 3744B, 3840B, 4000B, 4096B, 4300B, 4480B, 4608B, 4800B, 4960B, 5120B, 5376B, 5600B, 5760B, 6000B, 6144B, 6400B, 6528B, 6720B, 7008B, 7200B, 7360B, 7680B, 7840B, 8000B, 8192B, 8448B, 8640B, 8960B, 9216B, 9600B, 9728B, 10000B, 10240B, 10496B, 10800B, 11008B, 11200B, 11520B, 11712B, 12000B, 12288B, 12480B, 12800B, 13056B, 13280B, 13440B, 13824B, 14000B, 14336B, 14400B, 14784B, 15000B, 15232B, 15360B, 15744B, 16000B, 16128B, 16512B, 16800B, 17024B, 17280B, 17536B, 17600B, 17984B, 18000B, 18368B, 18560B, 18720B, 19104B, 19200B, 19584B, 19680B, 20064B, 20160B, 20544B, 20800B, 21024B, 21120B, 21504B, 21600B, 21984B, 22080B, 22464B, 22560B, 22944B, 23040B, 23424B, 23520B, 23904B, 24000B, 24384B, 24480B, 24864B, 24960B, 25344B, 25440B, 25824B, 25920B, 26304B, 26400B, 26784B, 26880B, 27264B, 27360B, 27744B, 27840B, 28224B, 28320B, 28704B, 28800B, 29184B, 29280B, 29664B, 29760B, 30144B, 30240B, 30624B, 30720B, 31104B, 31200B, 31584B, 31680B, 32064B, 32160B, 32544B, 32640B, 33024B, 33120B, 33504B, 33600B, 33984B, 34080B, 34464B, 34560B, 34944B, 35040B, 35424B, 35520B, 35904B, 36000B, 36384B, 36480B, 36864B, 36960B, 37344B, 37440B, 37824B, 37920B, 38304B, 38400B, 38784B, 38880B, 39264B, 39360B, 39744B, 39840B, 40224B, 40320B, 40704B, 40800B, 41184B, 41280B, 41664B, 41760B, 42144B, 42240B, 42624B, 42720B, 43104B, 43200B, 43584B, 43680B, 44064B, 44160B, 44544B, 44640B, 45024B, 45120B, 45504B, 45600B, 45984B, 46080B, 46464B, 46560B, 46944B, 47040B, 47424B, 47520B, 47904B, 48000B, 48384B, 48480B, 48864B, 48960B, 49344B, 49440B, 49824B, 49920B, 50304B, 50400B, 50784B, 50880B, 51264B, 51360B, 51744B, 51840B, 52224B, 52320B, 52704B, 52800B, 53184B, 53280B, 53664B, 53760B, 54144B, 54240B, 54624B, 54720B, 55104B, 55200B, 55584B, 55680B, 56064B, 56160B, 56544B, 56640B, 57024B, 57120B, 57504B, 57600B, 57984B, 58080B, 58464B, 58560B, 58944B, 59040B, 59424B, 59520B, 59904B, 60000B, 60384B, 60480B, 60864B, 60960B, 61344B, 61440B, 61824B, 61920B, 62304B, 62400B, 62784B, 62880B, 63264B, 63360B, 63744B, 63840B, 64224B, 64320B, 64704B, 64800B, 65184B, 65280B, 65664B, 65760B, 66144B, 66240B, 66624B, 66720B, 67104B, 67200B, 67584B, 67680B, 68064B, 68160B, 68544B, 68640B, 69024B, 69120B, 69504B, 69600B, 69984B, 70080B, 70464B, 70560B, 70944B, 71040B, 71424B, 71520B, 71904B, 72000B, 72384B, 72480B, 72864B, 72960B, 73344B, 73440B, 73824B, 73920B, 74304B, 74400B, 74784B, 74880B, 75264B, 75360B, 75744B, 75840B, 76224B, 76320B, 76704B, 76800B, 77184B, 77280B, 77664B, 77760B, 78144B, 78240B, 78624B, 78720B, 79104B, 79200B, 79584B, 79680B, 80064B, 80160B, 80544B, 80640B, 81024B, 81120B, 81504B, 81600B, 81984B, 82080B, 82464B, 82560B, 82944B, 83040B, 83424B, 83520B, 83904B, 84000B, 84384B, 84480B, 84864B, 84960B, 85344B, 85440B, 85824B, 85920B, 86304B, 86400B, 86784B, 86880B, 87264B, 87360B, 87744B, 87840B, 88224B, 88320B, 88704B, 88800B, 89184B, 89280B, 89664B, 89760B, 90144B, 90240B, 90624B, 90720B, 91104B, 91200B, 91584B, 91680B, 92064B, 92160B, 92544B, 92640B, 93024B, 93120B, 93504B, 93600B, 93984B, 94080B, 94464B, 94560B, 94944B, 95040B, 95424B, 95520B, 95904B, 96000B, 96384B, 96480B, 96864B, 96960B, 97344B, 97440B, 97824B, 97920B, 98304B, 98400B, 98784B, 98880B, 99264B, 99360B, 99744B, 99840B, 100000B

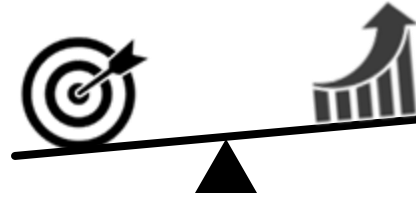
Overview of Oaken

Design Objectives

- ① Address memory bottleneck in LLM serving



- ② Find sweet spot between accuracy & performance



- ③ Maximize hardware utilization & performance

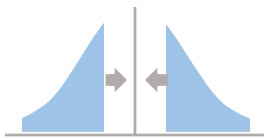


Algorithm Design

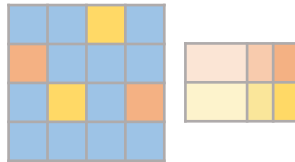
Threshold-based hybrid grouping



Group shift quantization

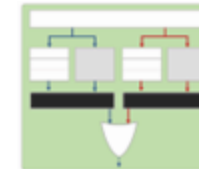


Dense-and-sparse encoding

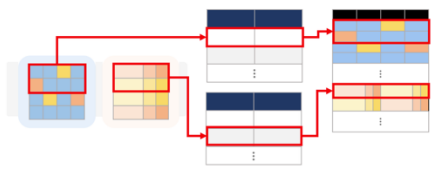


Hardware Design

Streamlined module architecture



Page-based memory management



Key Observations on KV Distribution

Observation 1

KV distribution **varies** across models and decoder layers



Insight 1

Oaken should determine quantization scale **for each model and decoder layer**

Observation 2

KV distribution is **consistent** across input datasets



Insight 2

Oaken can use shared quantization scale **regardless of model inputs**

Observation 3

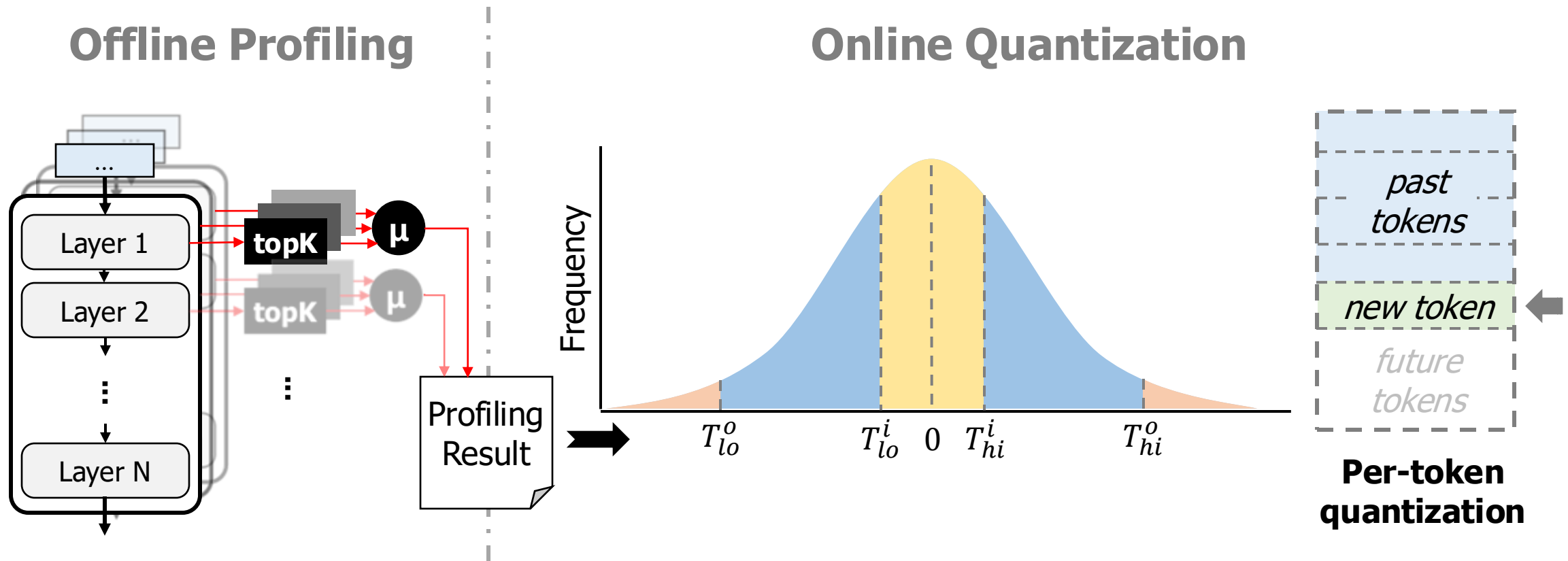
KV distribution has **exceptions** to channel-wise pattern



Insight 3

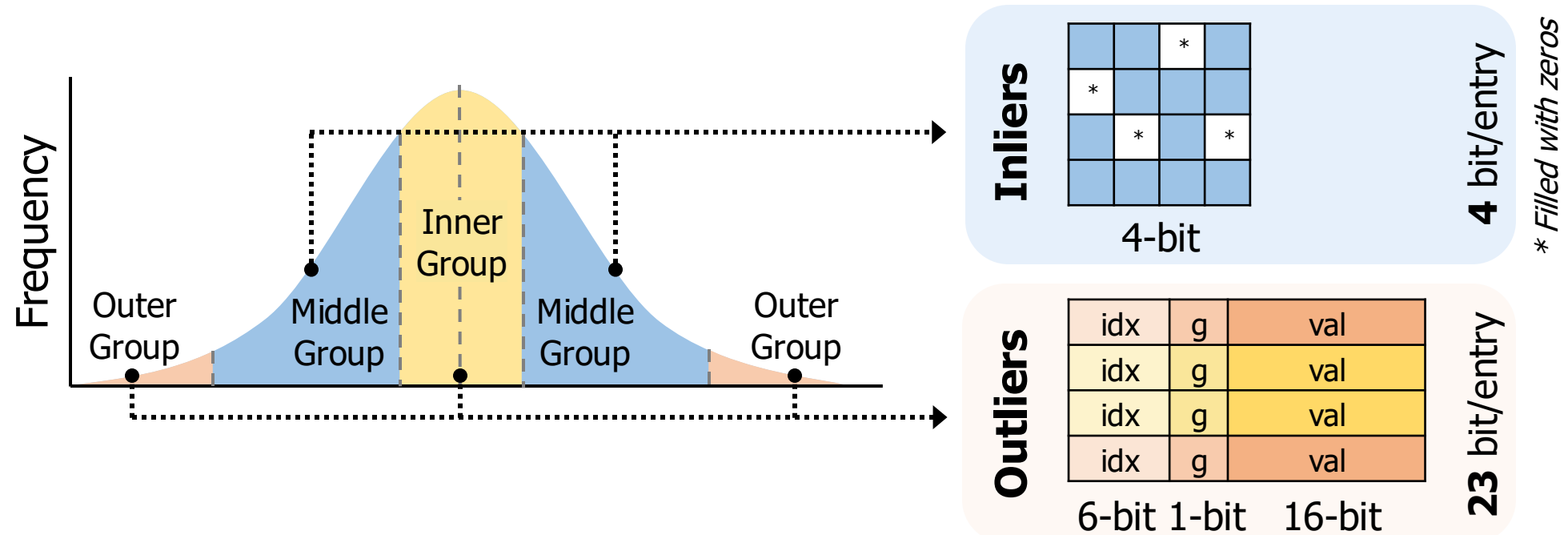
Oaken should use **multiple quantization groups** segmented by magnitude

Threshold-based Online-Offline Quantization



- Offline profiling requires **one-time cost** for each model (~ 100 inferences, ~ 10 min)

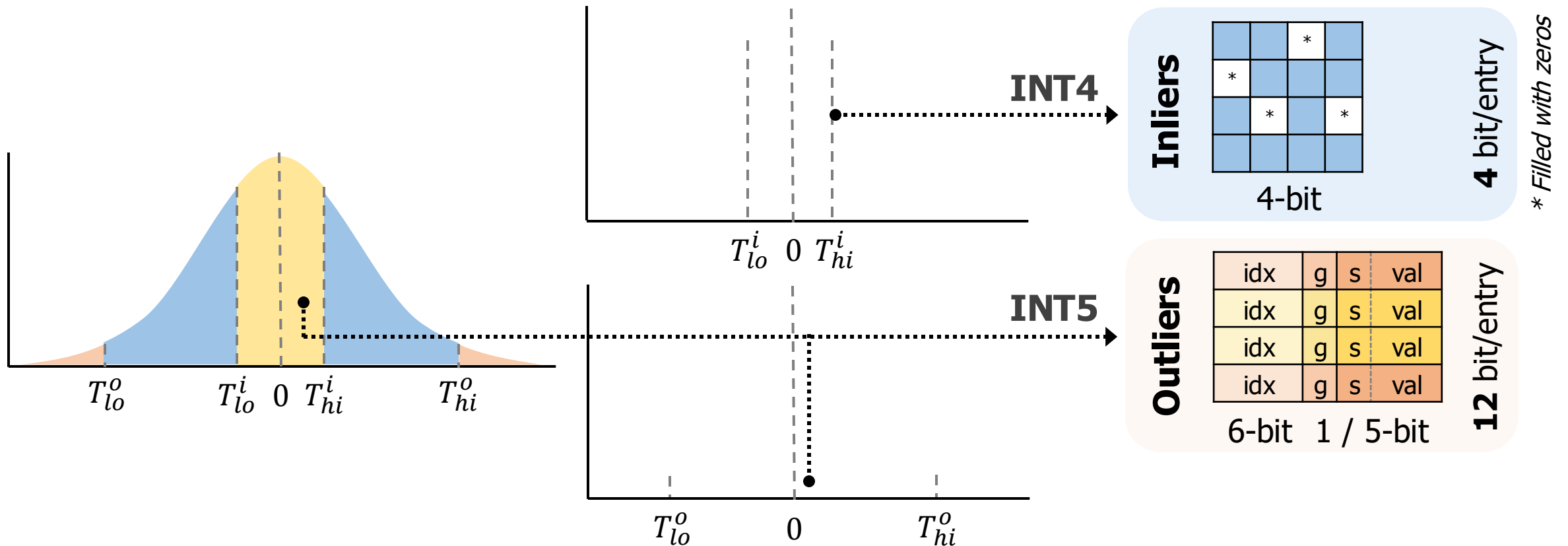
Threshold-based Online-Offline Quantization



Challenges:

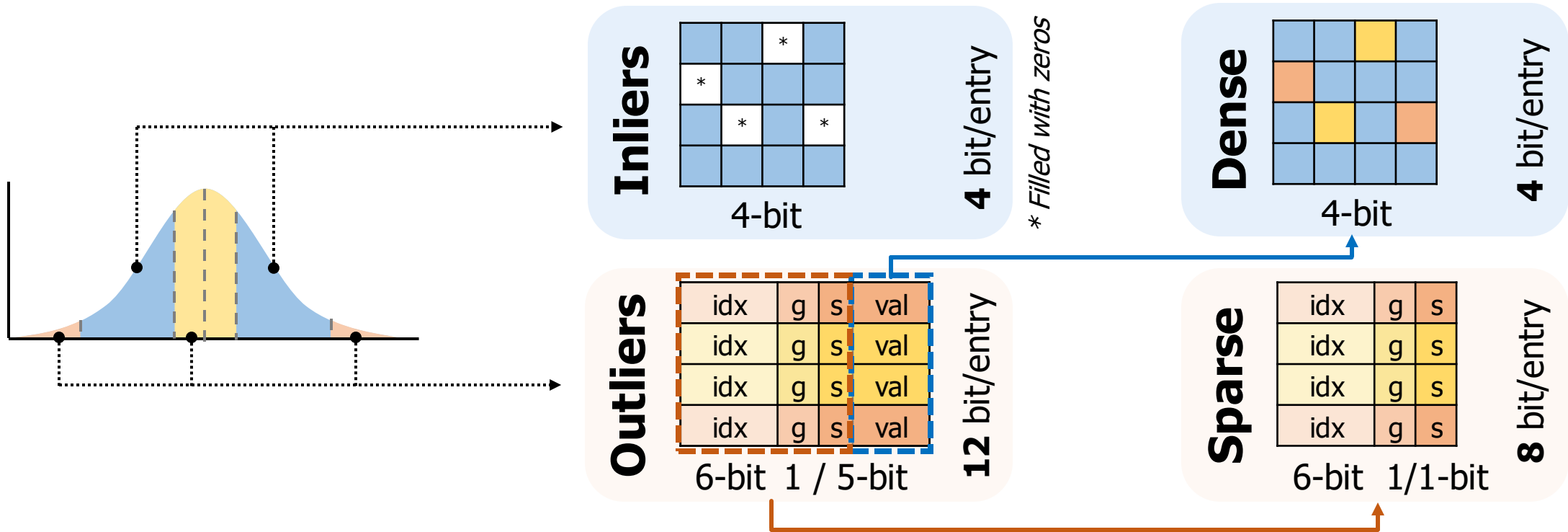
- Outliers add **storage and hardware costs**
- Outliers are **hard to quantize** due to large magnitude

Group Shift Quantization



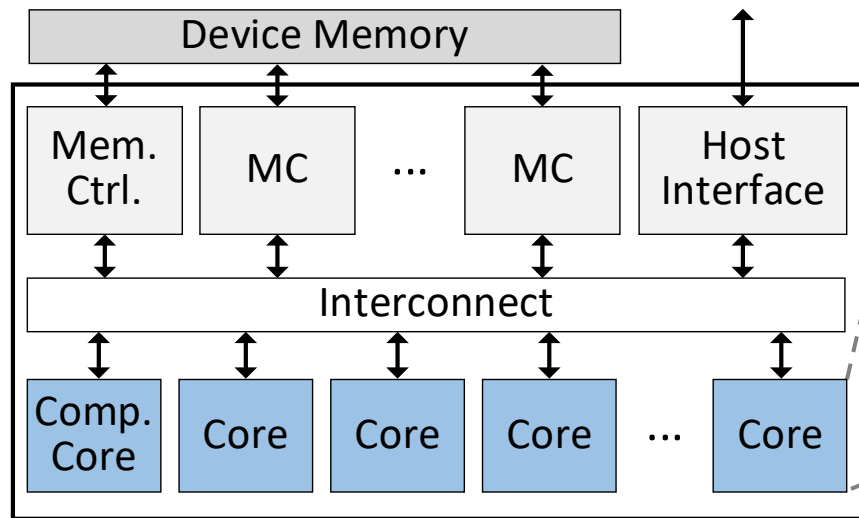
- **Group shift algorithm** reduces average bitwidth **from 5.9 to 4.8** * 10% Sparsity

Fused Dense-and-Sparse Encoding

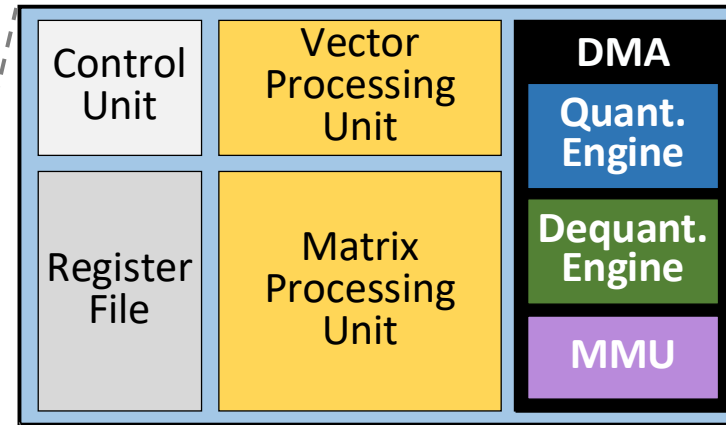


- 8-bit sparse matrices are **hardware-efficient** and **memory-aligned**
- **Fused encoding** reduces average bitwidth **from 4.8 to 4.4** * 10% Sparsity

Oaken Accelerator Integration



Oaken Accelerator



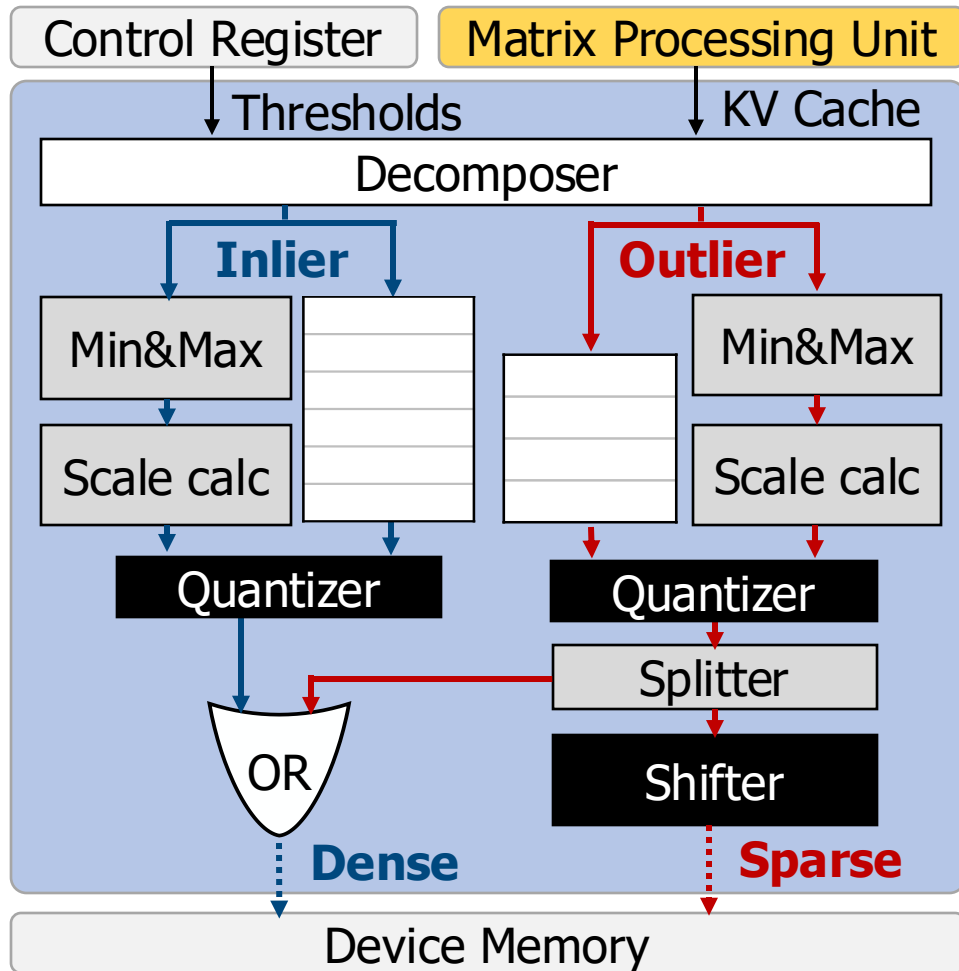
Oaken Compute Core

Module	Area
VPU	22.86%
MPU	6.03%
Quant. Engine	1.86%
Dequant. Engine	6.35%
Total	100%

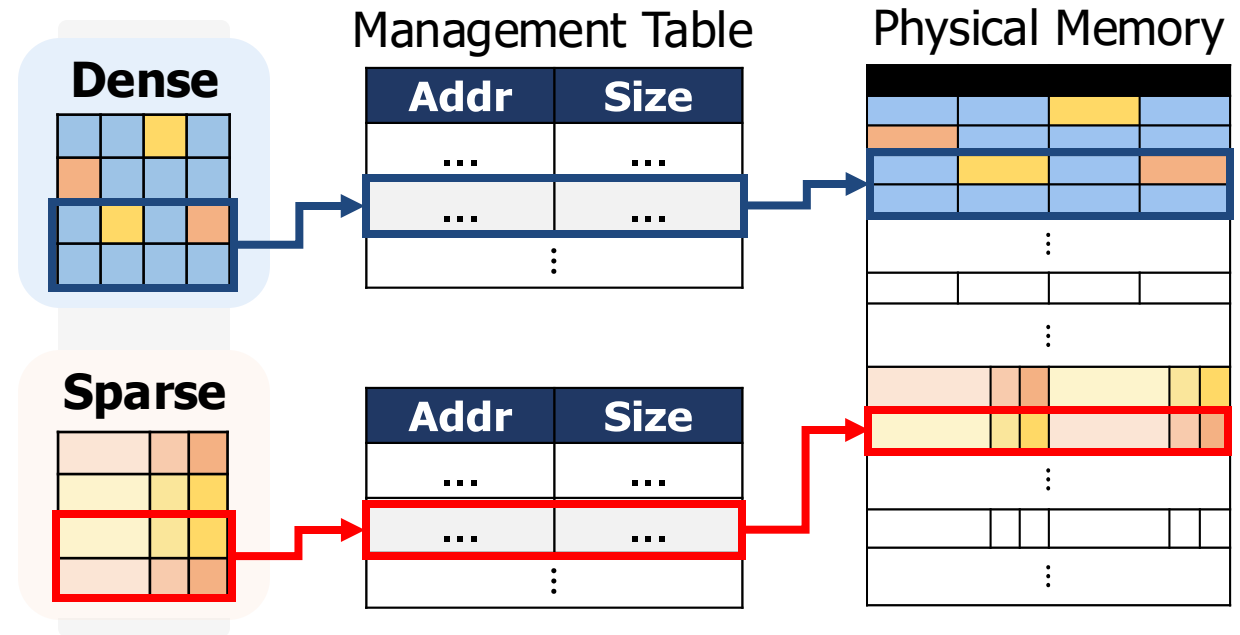
** Synthesized on TSMC 28nm*

- Oaken modules do **not modify** the existing compute logic in the accelerator
- Oaken modules are integrated into existing accelerator **with low overhead**

Oaken Hardware Modules



Quantization Engine



Memory Management Unit

- Oaken modules are designed to maximize **hardware** and **memory utilization**

Evaluation Methodology

▪ Models

- Llama2 – 7B, 13B, 70B*
- OPT – 6.7B, 13B, 30B*
- Mistral – 7B
- Mixtral – 8x7B*

** Used **2 GPUs** with pipeline parallelism*

▪ Baselines

- Tender (ASIC)
- Atom (GPU)
- QServe (GPU)
- KIVI (GPU)
- KVQuant (GPU)

▪ Datasets

- WikiText2, PIQA, WinoGrande, and HellaSwag

▪ Group Configuration

- **4%, 90%, 6%** for outer, middle and inner group

▪ Hardware Specification

	NVIDIA A100	Oaken-HBM	Oaken-LPDDR
FP16 TFLOPS	312	270	270
Memory type	HBM	HBM	LPDDR
Memory capacity	80 / 160* GB	80 GB	256 GB
Memory bandwidth	2.0 TB/s	2.0 TB/s	1.1 TB/s

** Used **2 GPUs** with pipeline parallelism*

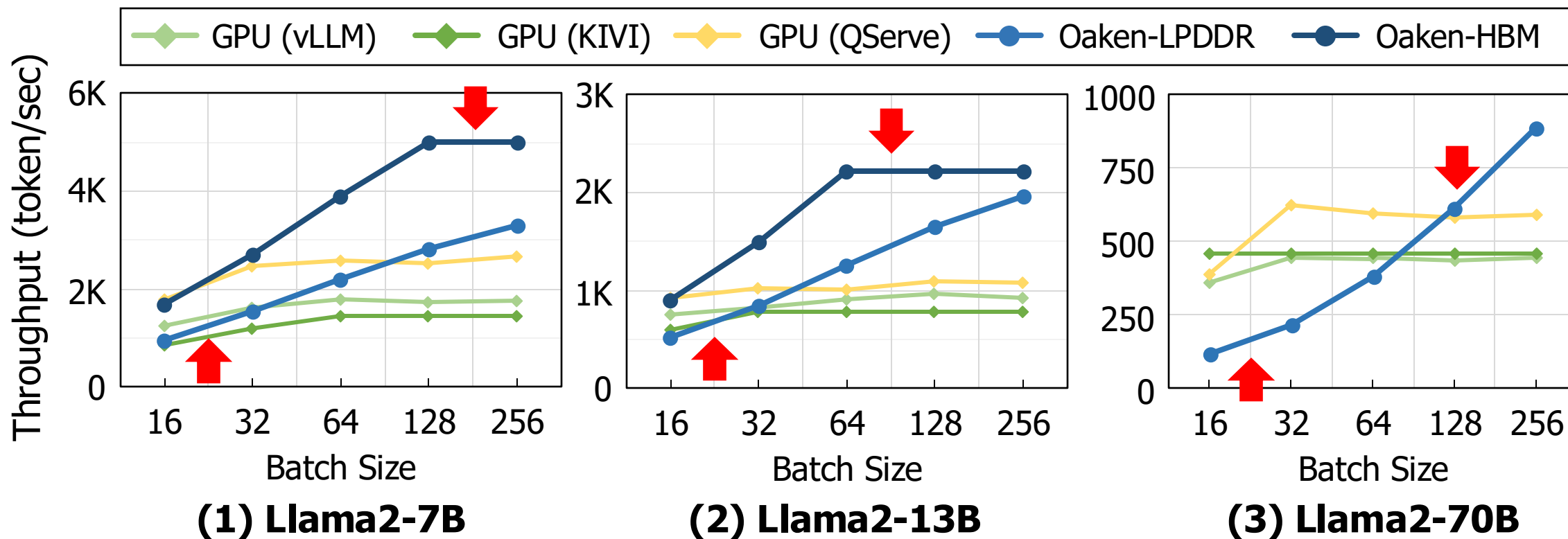
Evaluation Results

Throughput

Oaken-HBM achieves performance improvement of **1.79X** over vLLM (FP16)

Oaken-LPDDR is also a competitive option for **larger models** and **larger batches**

** Context length : 2K*



Evaluation Results

Accuracy

Model	Llama2							
	13B	70B	13B	70B	13B	70B	13B	70B
Dataset	WikiText2		PIQA		WinoGrande		HellaSwag	
Metric	Perplexity (↓)		Accuracy (%)		Accuracy (%)		Accuracy (%)	
Original	4.88	3.32	80.52	82.70	72.80	80.20	79.38	83.82
→ KIVI	4.90	3.33	79.05	78.07	70.96	76.81	78.97	83.47
→ QServe*	5.12	3.36	77.48	81.77	66.80	76.09	76.69	83.24
→ Oaken	4.93	3.34	79.71	82.59	70.56	76.64	78.24	83.50

* Activated KV quantization feature only

Oaken incurs **0.87%** and **0.32% accuracy loss** compared to FP16 and KIVI

Oaken achieves **1.38% higher** accuracy compared to QServe

Additional Results in Our Paper

- Performance evaluation using other LLMs and baselines
- Accuracy and effective bits with varying group configurations
- End-to-end latency breakdown
- Sensitivity study to total sequence length
- Performance evaluation using real-world benchmark
- Synthesized area and power

Conclusion

- **Oaken**

- Acceleration solution for LLM inference serving including algorithm-hardware co-designed KV cache quantization technique

- **Contributions**

- Addresses memory bandwidth and capacity bottlenecks in modern LLM serving
- Finds sweet spot in accuracy-performance trade-off of KV cache quantization

- **Future works**

- Extending Oaken to handle recent attention architectures (*e.g., latent attention, linear attention*)
- HyperAccel's high efficiency LLM accelerator with broad quantization support