IDEC Chip Design Contest

Bit-Separable Transformer Accelerator Leveraging Output Activation Sparsity for Efficient DRAM Access

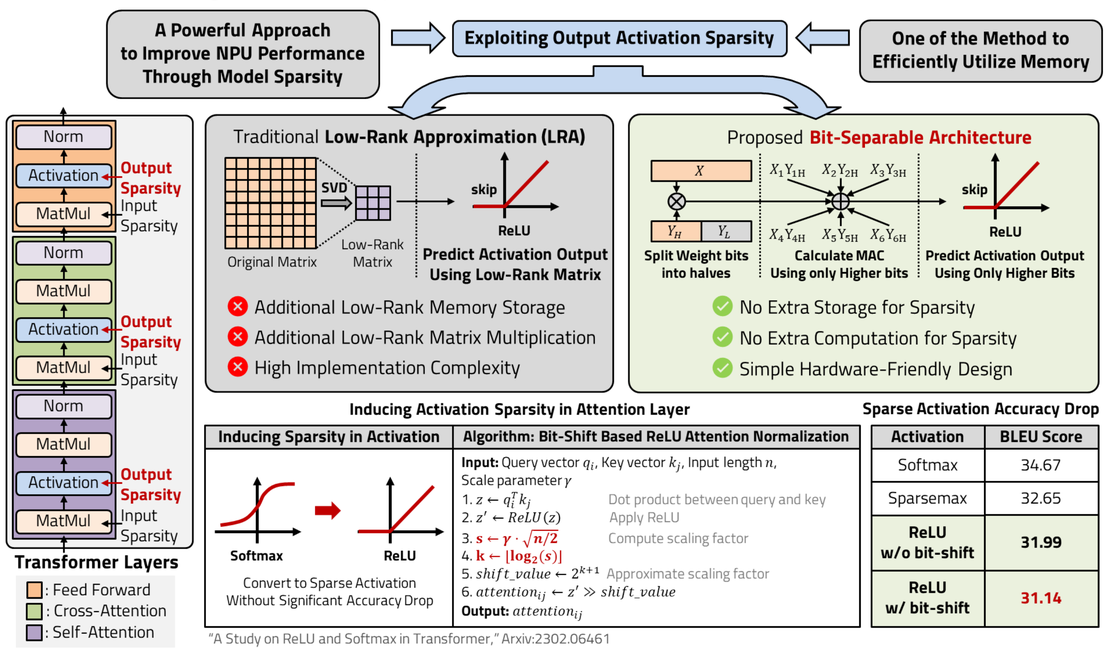
Seunghyun Park, Daejin Park
School of Electronic and Electrical Engineering, Kyungpook National University, South Korea
ijjh0435@knu.ac.kr

Abstract

We propose a bit-separable transformer accelerator that exploits output activation sparsity to skip lower-bit computations and optimizes DRAM access through subarray-level bit separation. The key idea is to predict activation results using only upper-bit operations, eliminating redundant processing without additional matrix operations. Subarray-level separation balances access rates, improves bank utilization, and reduces latency and power. Experiments demonstrate a 22.32% latency reduction, 64.3% increase in bank utilization, 19.4% bandwidth improvement, and 13.2% power reduction, achieving energy efficiency of 14.3–27.3 TOPS/W with small accuracy loss.

Exploiting Output Activation Sparsity in Transformer Models

We exploit output activation sparsity in transformer models to skip redundant lower-bit computations, reducing DRAM access and computation while maintaining high accuracy without extra storage or matrix operations.



Inducing Activation Sparsity in Attention Layer

Inducing Sparsity in Activation	Algorithm: Bit-Shift Based ReLU Attention Normalization
Softmax → ReLU Convert to Sparse Activation Without Significant Accuracy Drop	**Input:** Query vector q_i, Key vector k_j, Input length n, Scale parameter γ 1. $z \leftarrow q_i^T k_j$ Dot product between query and key 2. $z' \leftarrow ReLU(z)$ Apply ReLU 3. $s \leftarrow \gamma \cdot \sqrt{n/2}$ Compute scaling factor 4. $\mathbf{k} \leftarrow \lfloor \log_2(s) \rfloor$ 5. $shift\_value \leftarrow 2^{k+1}$ Approximate scaling factor 6. $attention_{ij} \leftarrow z' \gg shift\_value$ **Output:** $attention_{ij}$

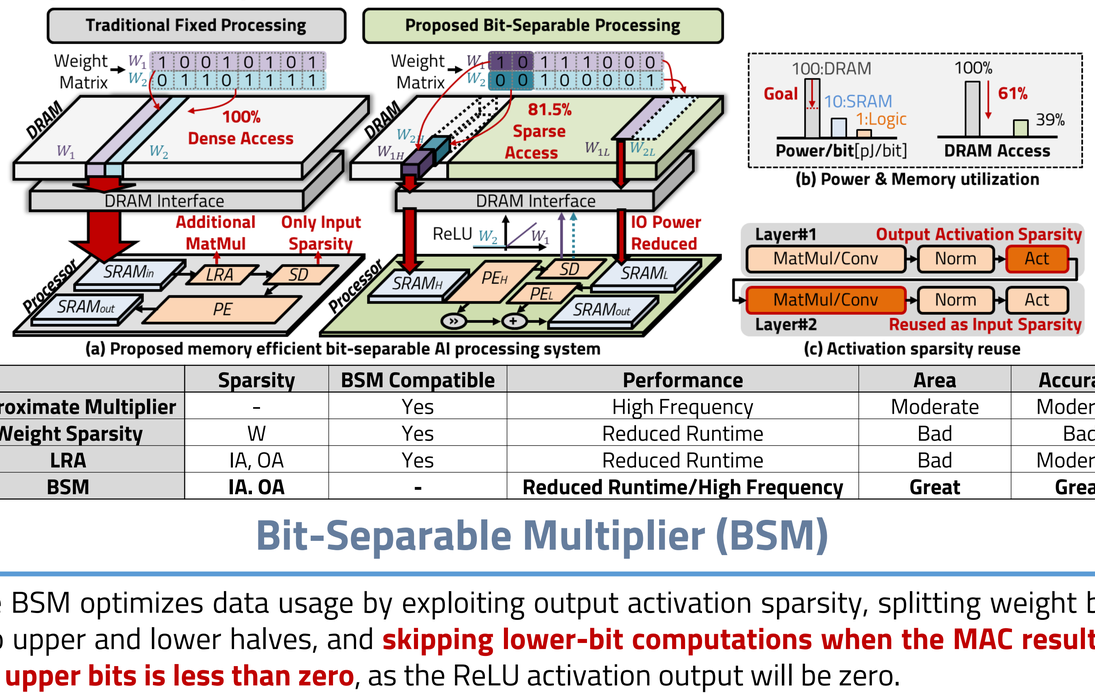
Sparse Activation Accuracy Drop

Activation	BLEU Score
Softmax	34.67
Sparsemax	32.65
ReLU w/o bit-shift	31.99
ReLU w/ bit-shift	31.14

"A Study on ReLU and Softmax in Transformer," Arxiv:2302.06461

Proposed Bit-Separable Transformer Accelerator

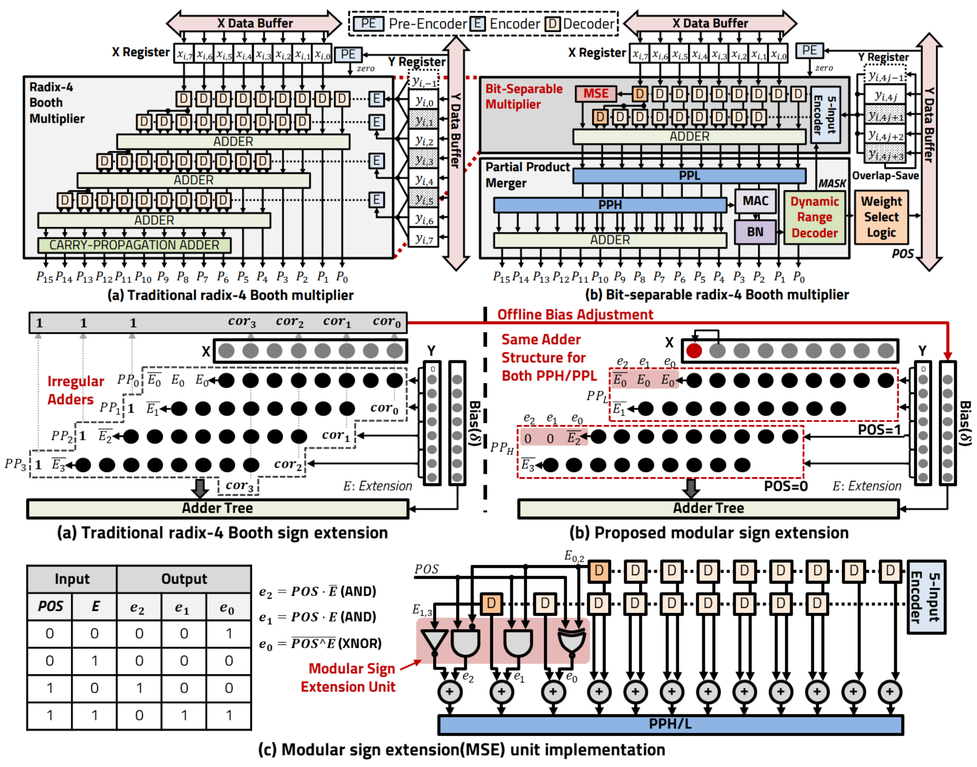
The proposed accelerator separates weights into upper and lower bits and predicts activation outputs using upper-bit results to skip unnecessary lower-bit computations. **The output activation sparsity detected in one layer can be directly reused as input activation sparsity in the next layer**, further reducing computation and DRAM access.



	Sparsity	BSM Compatible	Performance	Area	Accuracy
Approximate Multiplier	-	Yes	High Frequency	Moderate	Moderate
Weight Sparsity	W	Yes	Reduced Runtime	Bad	Bad
LRA	IA, OA	Yes	Reduced Runtime	Bad	Moderate
BSM	**IA. OA**	**-**	**Reduced Runtime/High Frequency**	**Great**	**Great**

Bit-Separable Multiplier (BSM)

The BSM optimizes data usage by exploiting output activation sparsity, splitting weight bits into upper and lower halves, and **skipping lower-bit computations when the MAC result of the upper bits is less than zero**, as the ReLU activation output will be zero.

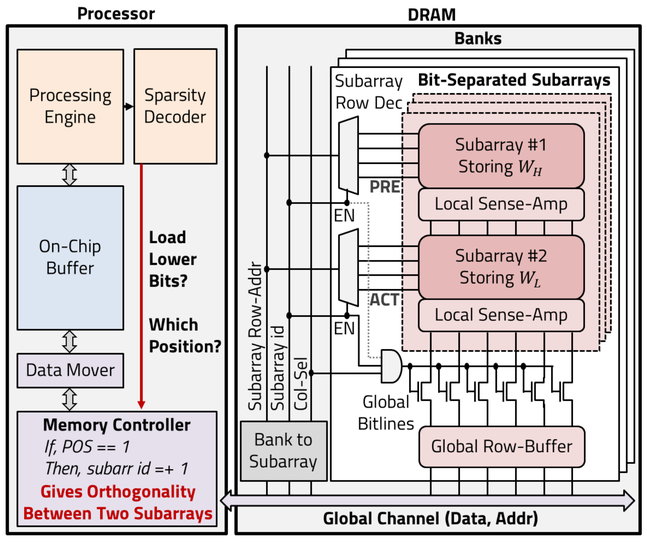


Input		Output		
POS	*E*	e_2	e_1	e_0
0	0	0	0	1
0	1	0	0	0
1	0	1	0	0
1	1	0	1	1

$e_2 = POS \cdot \overline{E}$ (AND)
$e_1 = POS \cdot E$ (AND)
$e_0 = \overline{POS \wedge E}$ (XNOR)

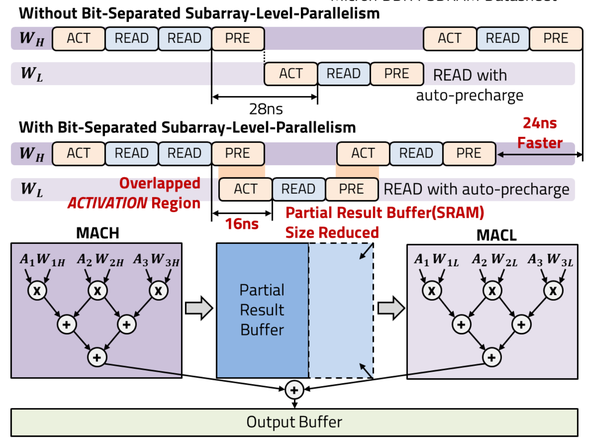
DRAM Access Optimization Using Subarray-Level Parallelism

The proposed method **stores upper and lower bits in different subarrays to balance access rates and improve bank utilization, loads lower bits only when needed** to reduce latency an d power, and achieves higher bandwidth and efficiency through simple subarray ID and positi on control without changing the column address.

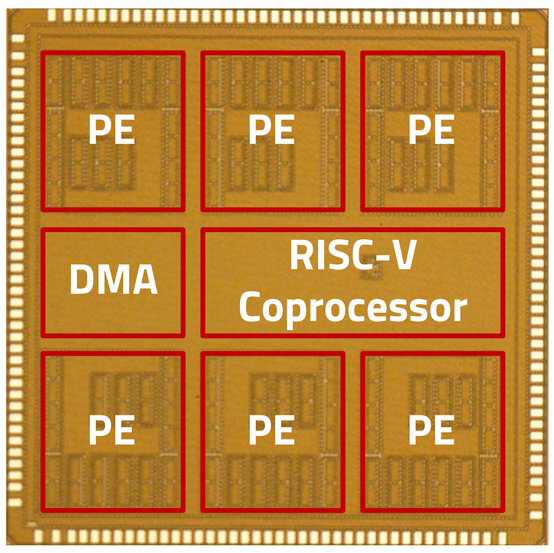


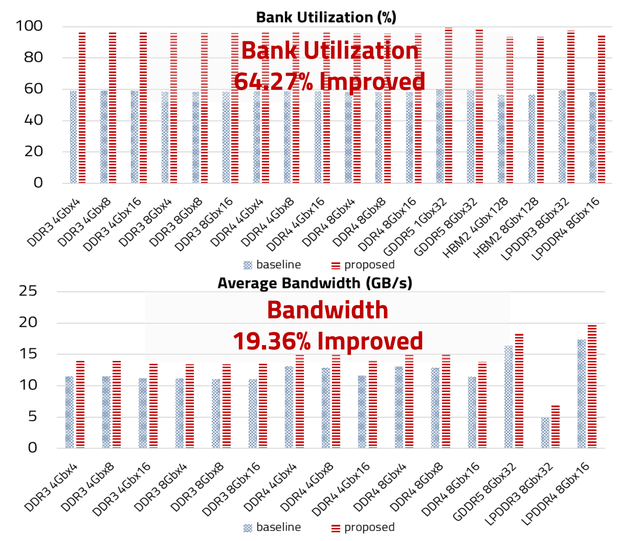
Command	ACT-READ	READ-READ	READ-PRE	PRE period
Delay	14ns	4ns	5.5ns	14ns

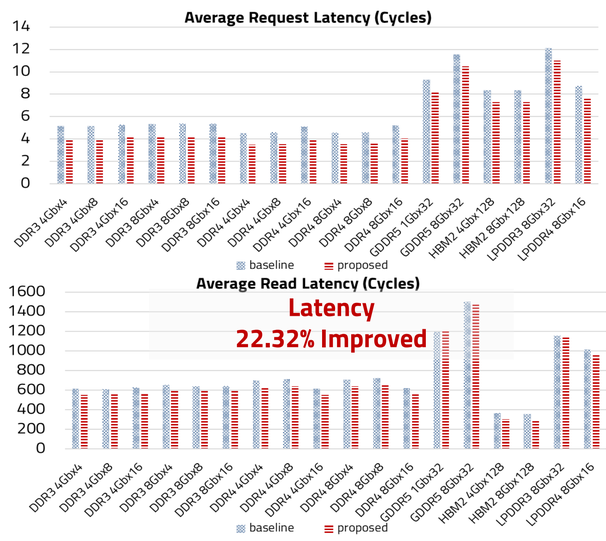
*Micron DDR4 SDRAM Datasheet

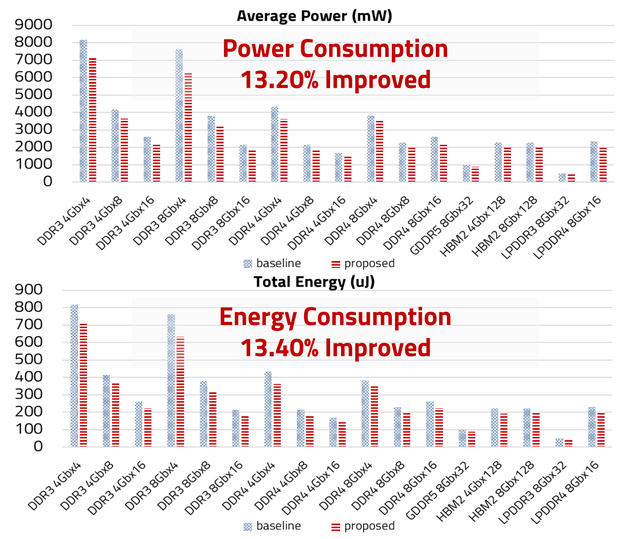


Experimental Results









Model	Dataset	Params	Samples	Baseline Acc. (%)	BSM Acc. (%)
DistilBERT	GLUE SST-2	66.96M	872	91.06	90.83
GPT-2	HellaSwag	124.44M	2,000	31.85	30.25
FLAN-T5-small	SuperGLUE	76.96M	2,000	64.20	63.90
ResNet18	CIFAR-10	11.69M	10,000	84.32	84.28
ResNet50	CIFAR-10	25.56M	10,000	84.59	84.52
MobileNetV2	CIFAR-10	3.50M	10,000	79.66	79.51
GooLeNet	CIFAR-10	13.03M	10,000	87.15	87.03

	SWPU TCAS'22	Trainer JSSC'22	CNN-DLA JSSC'23	DQ-STP TCAS'24	VersaDLA TECS'25	This Work
Sparsity Support	WS, IAS, OAS	WS, IAS, OAS	WS, IAS, OAS	WS, IAS, OAS	WS	**IAS, OAS(BSM)**
Precision	FP16	INT8/FP16	FP8/16	FXP16	BF16	**INT8**
Accuracy Drop [%]	0.31	0.4	0.07	1.02	0.5	**0.07**
Technology [nm]	28 Layout	28 Chip	28 Chip	65 Layout	28 Layout	**28 Chip**
Total Area [mm²]	6.80	20.96	16.40	21.50	7.90	**16.00**
Voltage [V]	0.56-1.0	0.58-1.0	0.6-1.1	1.1	0.8-1.1	**1**
Power [mW]	16-556	23-363	51-624	587	50-793	**56-235**
Frequency [MHz]	675	440	75-340	200	1066	**524**
TOPS	0.4-16.4	0.5-35.4	0.6-3.7	0.9-38.0	0.1-0.5	**0.8-1.5***
TOPS/W	2.7-126.0	2.1-173.28	5.3-16.4	1.5-90.6	0.7-2.1	**14.3-27.3***
GOPS/mm²	57.2-2404.4	21.5-1687.5	34.1-225.6	41.2-1768.4	16.5-56.9	**50.0-95.6***

WS: Weight Sparsity, IAS: Input Activation Sparsity OAS: Output Activation Sparsity *: @90% Sparsity Performance

Conclusion

The proposed accelerator leverages bit-separable computation and subarray-level DRAM opti mization to exploit output activation sparsity. By predicting results from upper-bit operations, it skips unnecessary lower-bit computations, reducing computation and memory access. Suba rray-level separation improves bank utilization while lowering latency and power. Experiments show notable speed and efficiency gains with minimal accuracy loss.

KNU KYUNGPOOK NATIONAL UNIVERSITY
AI-SoC Lab