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

Seunghyun Park and Daejin Park

AI-**E**mbedded **S**^{2}oftware **o**n **C**hip Lab.

School of Electronic and Electrical Engineering, KNU, KOREA

2025.08.26

AI-**E**mbedded **S**^{2}oftware **o**n **C**hip Lab

KYUNGPOOK NATIONAL UNIVERSITY

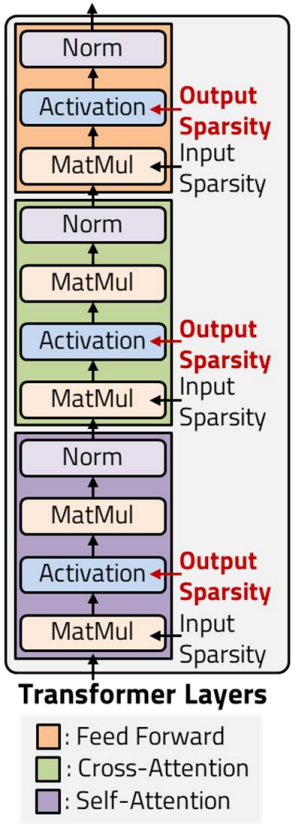
Abstract

- We propose a **Bit-Separable Transformer Accelerator** to cut compute and DRAM cost in AI inference.
- The key idea is to **split weight bits, compute with upper bits, and skip lower bits when ReLU output is zero.**
- This conservative approximate method removes redundant work while preserving accuracy.
- For memory, we use subarray-level separation to balance banks, reduce latency, and save power.
- A Samsung 28nm chip shows up to **27.3 TOPS/W** with very small accuracy loss.

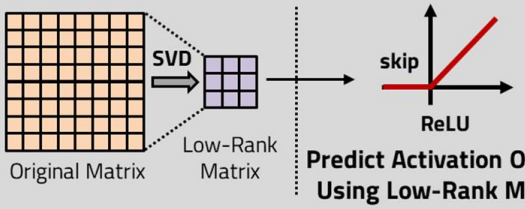
Exploiting Output Activation Sparsity

- We exploit output activation sparsity to skip lower-bit work, cutting compute and DRAM access while preserving accuracy without extra overhead.

A Powerful Approach to Improve NPU Performance Through Model Sparsity → **Exploiting Output Activation Sparsity** ← **One of the Method to Efficiently Utilize Memory**

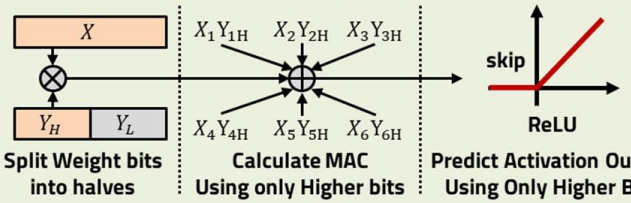


Traditional Low-Rank Approximation (LRA)



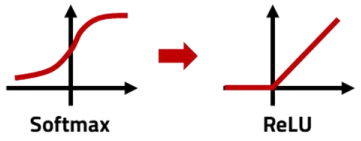
- Additional Low-Rank Memory Storage
- Additional Low-Rank Matrix Multiplication
- High Implementation Complexity

Proposed Bit-Separable Architecture



- No Extra Storage for Sparsity
- No Extra Computation for Sparsity
- Simple Hardware-Friendly Design

Inducing Activation Sparsity in Attention Layer

Inducing Sparsity in Activation	Algorithm: Bit-Shift Based ReLU Attention Normalization
 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. $\mathbf{s} \leftarrow \boldsymbol{\gamma} \cdot \sqrt{n/2}$ 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

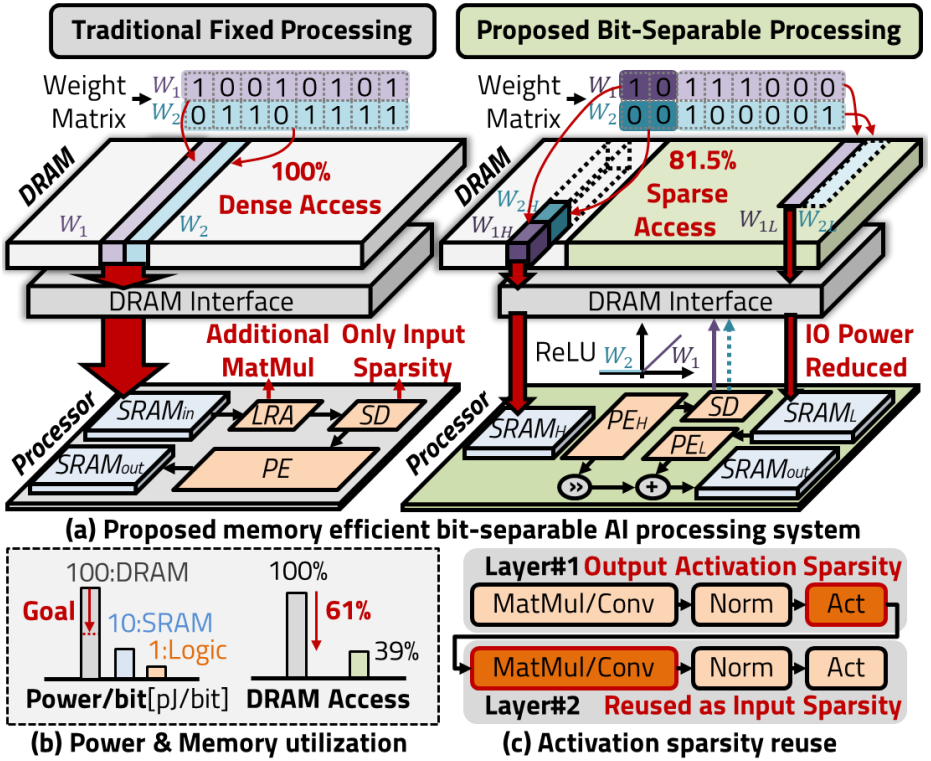
Leveraging Output Activation Sparsity

- **World's first ReLU-based Transformer Accelerator.**
  - Attention layer was normalized by standard deviation for model convergence.*

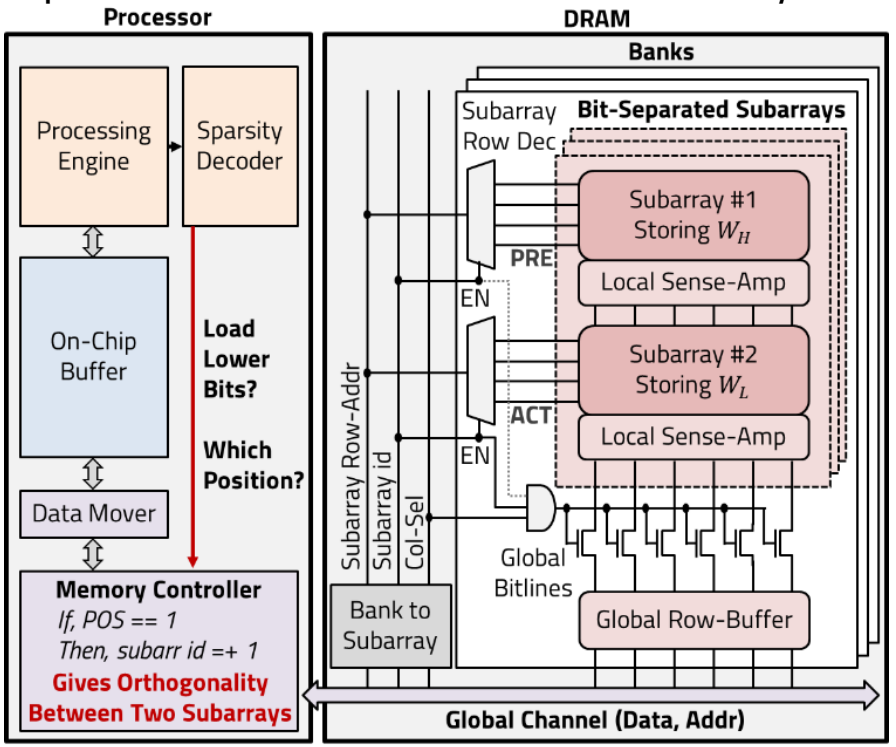
* "A Study on ReLU and Softmax in Transformer," Arxiv:2302.06461

- **Subarray-level parallelism** was applied to minimize data access.*
  - Overlap the timing of DRAM operation.
  - Store separated bits into each subarrays. => prevents collisions between subarrays.

"A Case for Exploiting Subarray-Level Parallelism(SALP) in DRAM," ISCA, 2012



BSM-Based ReLU-Transformer NPU

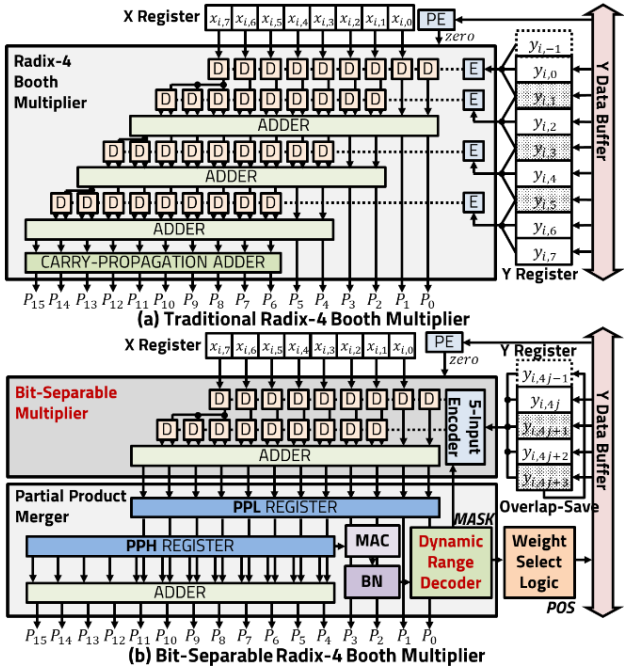


Subarray-Level Parallelism

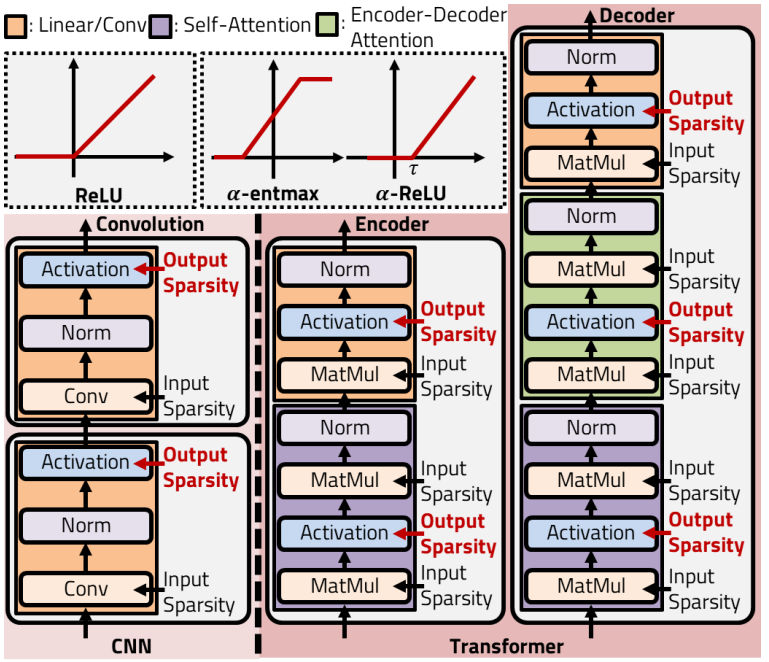
Bit-Separable Multiplier (BSM)

* "Bit-Separable Radix-4 Booth Multiplier for Power-Efficient CNN Accelerator," COOLChips, 2024.

- In our previous study, the **bit-separable multiplier (BSM)** was introduced.*
  - BSM optimizes **data usage** by exploiting **output activation sparsity**.
  - BSM **split weight bits into upper and lower halves.**
    - If MAC result of upper bits < 0,
      - ReLU activation outputs 0.
      - **Skip computation of lower bits!**

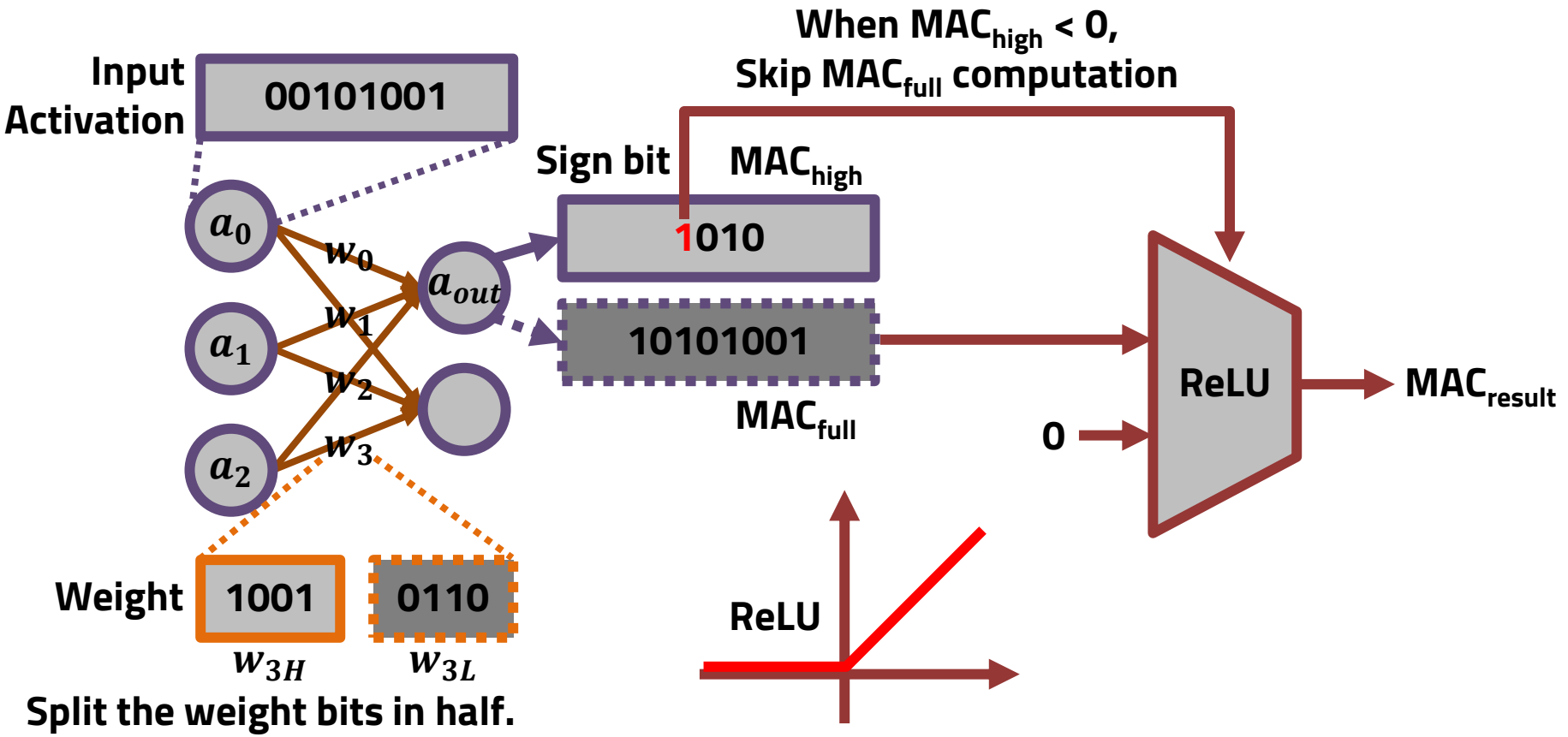


Bit-Separable Multiplier Architecture



Leveraging Output Activation Sparsity

Bit-Separable Multiplier (BSM)



In this way, the Bit-Separable Architecture can predict Activation Output Using Only Higher Bits.

DRAM Simulation Results

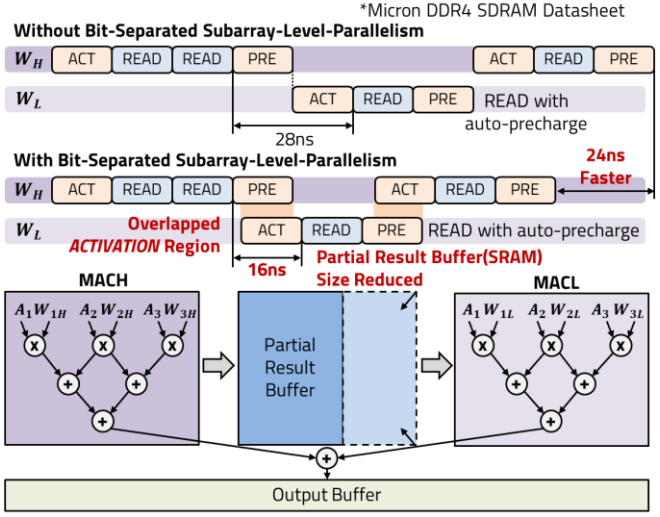
- **22.3% memory access latency** reduced.
- **35.7% bank utilization** increased.
- **19.4% bandwidth** improved.
- **13.2% power** saved.

  => memory-efficient operation!

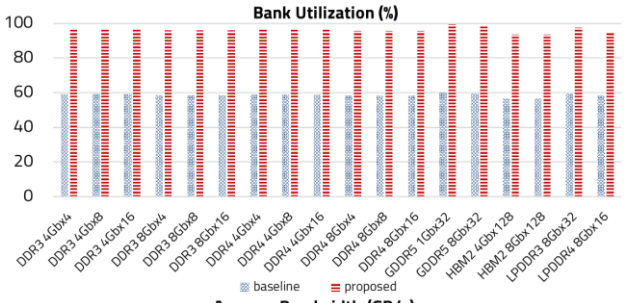
Bit-Separable Subarray-Level Parallelism

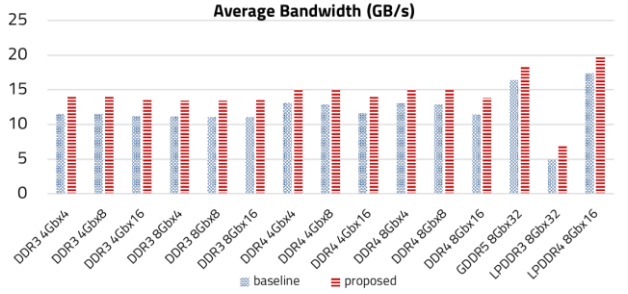
Command	ACT-READ	READ-READ	READ-PRE	PRE period
Delay	14ns	4ns	5.5ns	14ns

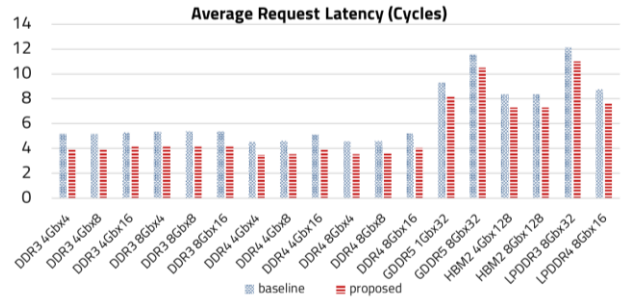
*Micron DDR4 SDRAM Datasheet

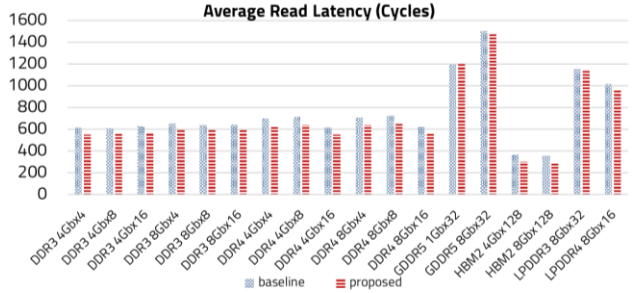


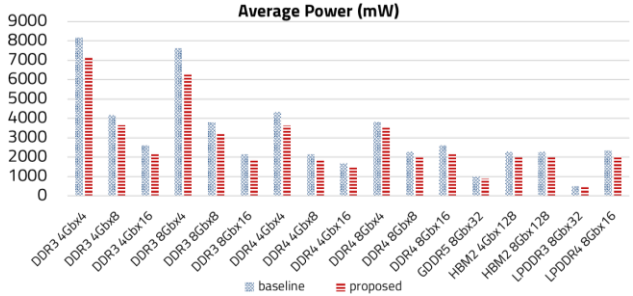
DRAMSim3 Results

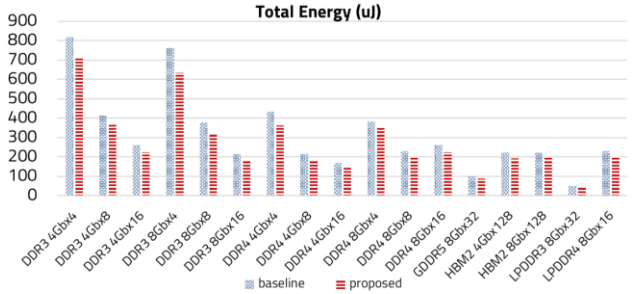






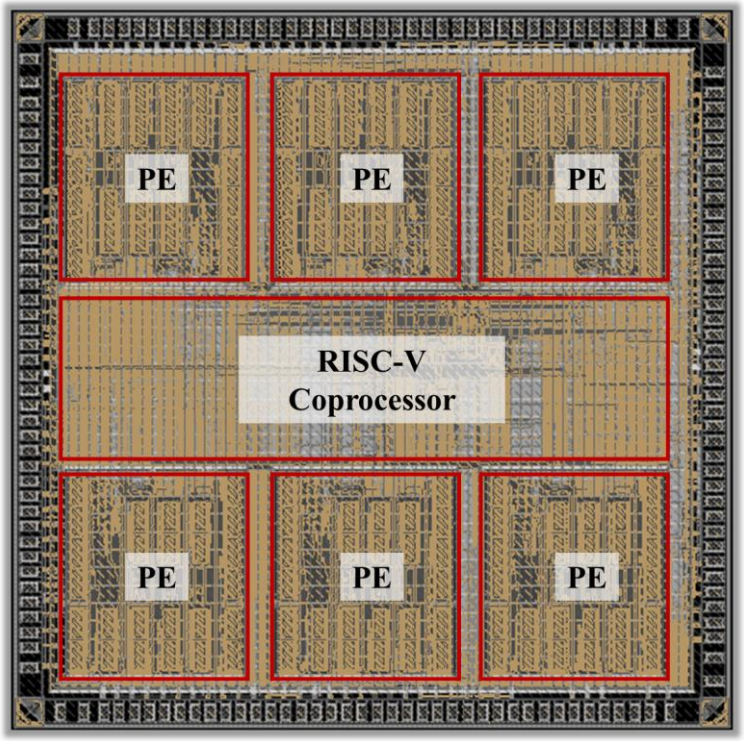






Bit-Separable Transformer Accelerator

- Bit-separable multiplier skips unnecessary lower-bit computations
- Optimizes DRAM access using subarray-level bit separation.
- Outperforms existing sparsity-aware accelerators with 14.3~27.3 TOPS/W



	INT8			INT16		
Sample Size	100,000					
Kernel Size	9	25	128	9	25	128
Error Probability	1.91E-02	1.96E-02	1.99E-02	1.19E-03	1.08E-03	9.00E-04
Max. Error	4.15E-03	1.43E-03	2.70E-04	1.68E-04	6.07E-05	5.20E-06
Mean Error	7.36E-04	2.66E-04	5.22E-05	4.92E-05	1.85E-05	3.28E-06
Error Expectation	1.40E-05	5.20E-06	1.04E-06	5.86E-08	2.00E-08	2.95E-09
MRED	1.98E-02	1.97E-02	2.22E-02	1.16E-03	1.09E-03	7.98E-04
NMED	**2.81E-05**	**1.04E-05**	**2.08E-06**	**1.17E-07**	**4.00E-08**	**5.90E-09**

	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(LRA)	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 [mm^2]	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/mm^2	57.2-2404.4	21.5-1687.5	34.1-225.6	41.2-1768.4	16.5-56.9	**50.0-95.6***

Conclusion

- The proposed accelerator uses bit-separable computation and subarray-level DRAM optimization.
- It predicts outputs with upper bits and skips unnecessary lower-bit operations.
- This reduces both computation and memory access with no accuracy penalty.
- Subarray-level separation balances bank usage and lowers latency and power.
- Overall, the design achieves strong efficiency gains with minimal accuracy loss.