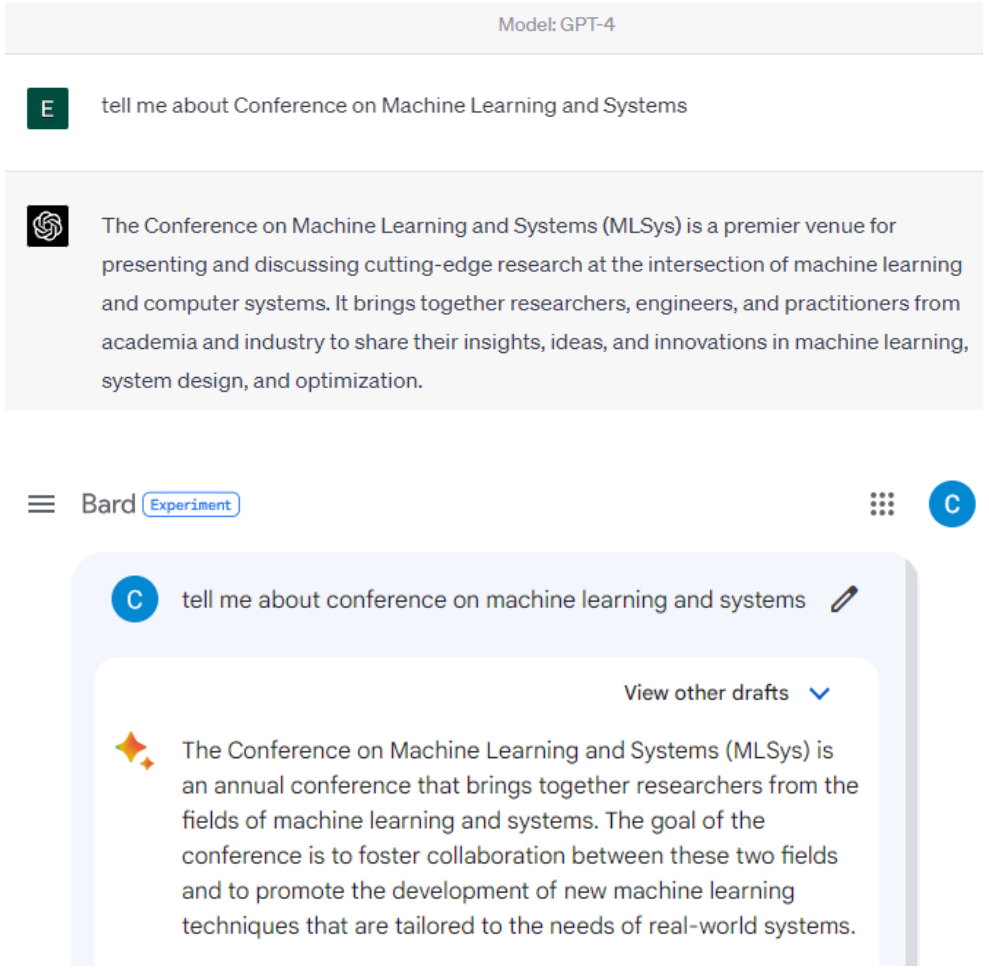
SPARSITY-AWARE MEMORY INTERFACE USING STACKED XORNET COMPRESSION

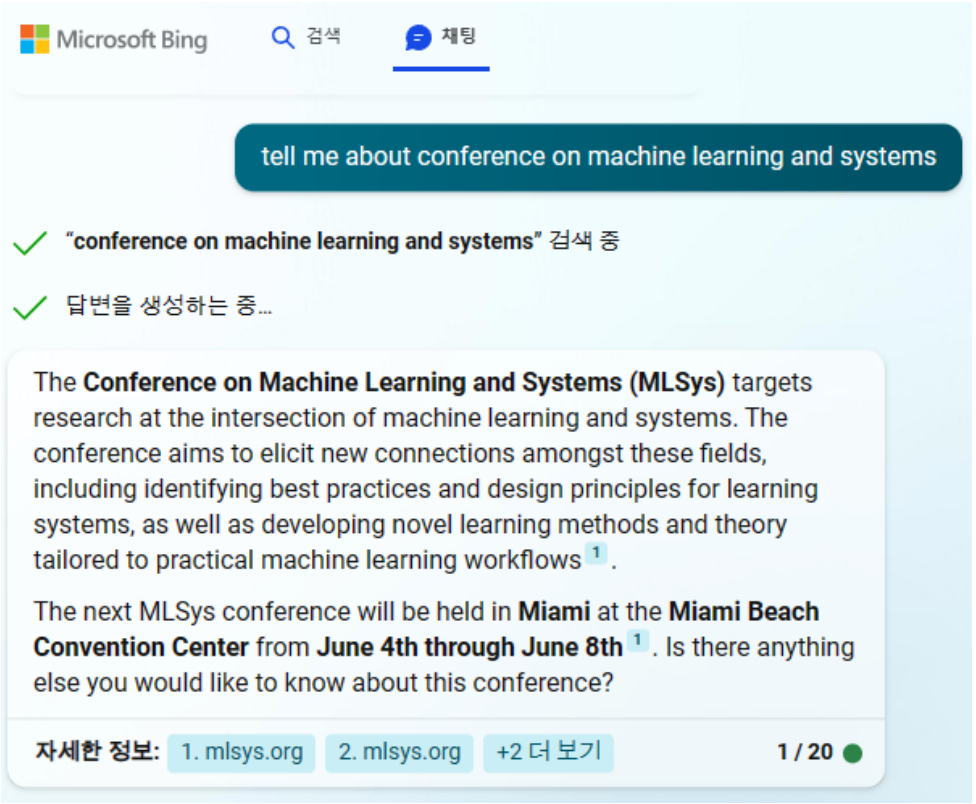
Younghoon Byun[1]*, Seungsik Moon[1]*, Baeseong Park[2], Se Jung Kwon[2], Dongsoo Lee[2], Gunho Park[1], Eunji Yoo[1], Jung Gyu Min[1], Youngjoo Lee[1]

[1]Pohang University of Science and Technology, [2]NAVER Cloud

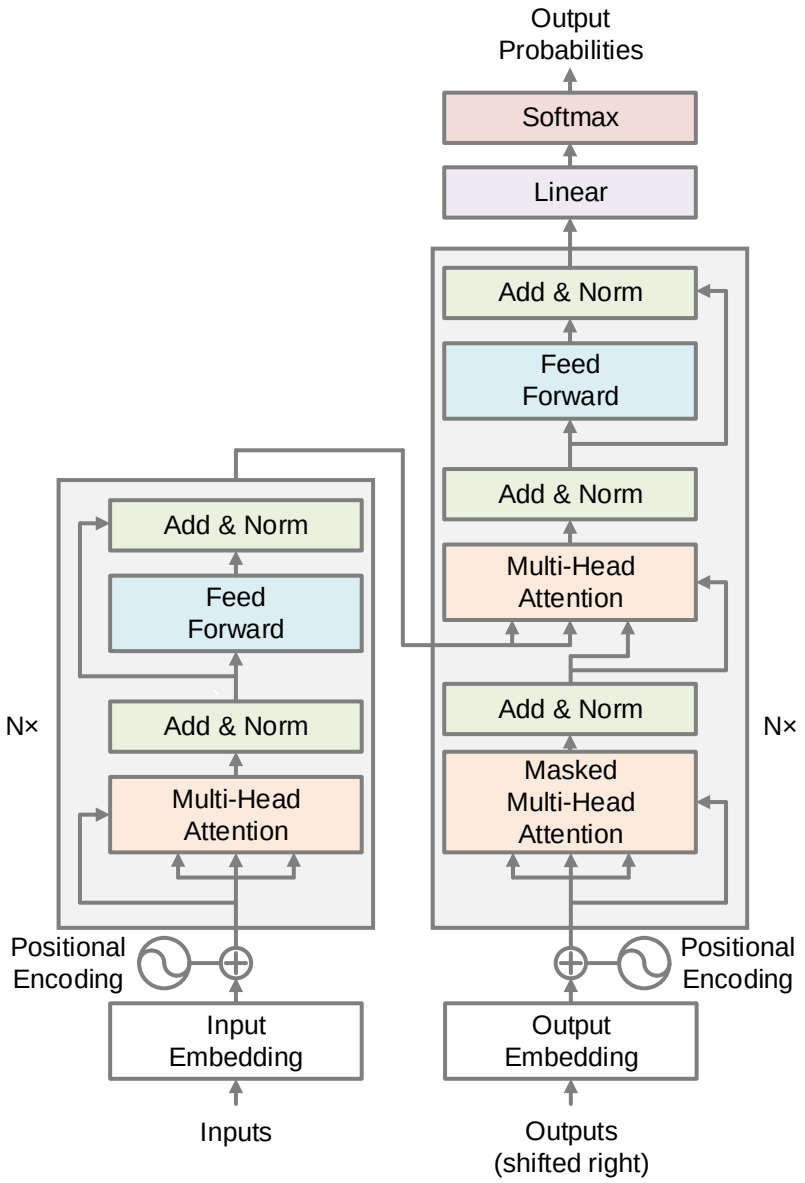
EPIC LAB

Motivation


Model: GPT-4
E
tell me about Conference on Machine Learning and Systems
The Conference on Machine Learning and Systems (MLSys) is a premier venue for presenting and discussing cutting-edge research at the intersection of machine learning and computer systems. It brings together researchers, engineers, and practitioners from academia and industry to share their insights, ideas, and innovations in machine learning, system design, and optimization.
Bard
Experiment
C
tell me about conference on machine learning and systems
View other drafts
The Conference on Machine Learning and Systems (MLSys) is an annual conference that brings together researchers from the fields of machine learning and systems. The goal of the conference is to foster collaboration between these two fields and to promote the development of new machine learning techniques that are tailored to the needs of real-world systems.


Microsoft Bing
검색
채팅
tell me about conference on machine learning and systems
"conference on machine learning and systems" 검색 중
답변을 생성하는 중...
The Conference on Machine Learning and Systems (MLSys) targets research at the intersection of machine learning and systems. The conference aims to elicit new connections amongst these fields, including identifying best practices and design principles for learning systems, as well as developing novel learning methods and theory tailored to practical machine learning workflows [1].
The next MLSys conference will be held in Miami at the Miami Beach Convention Center from June 4th through June 8th [1]. Is there anything else you would like to know about this conference?
자세한 정보: 1. mlsys.org 2. mlsys.org +2 더 보기
1 / 20

Motivation



Encoder (Context summarization)

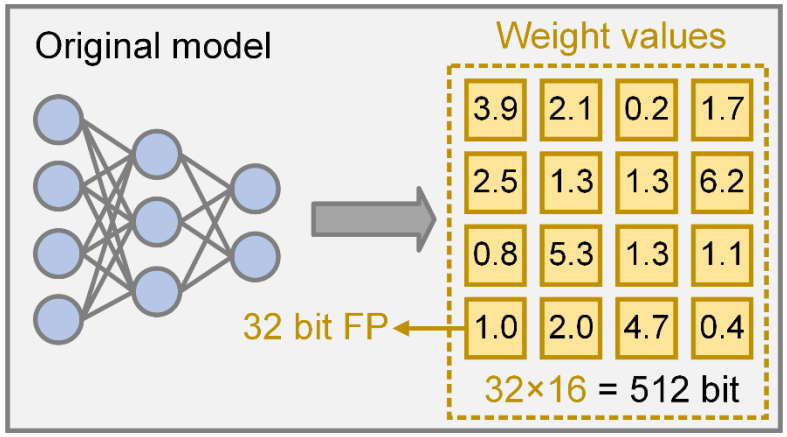
- ✓ Large batch size MM
  (seq_len × batch_size)
- ✓ Higher weight-reusability
- ✓ Computation-bound

Decoder (Generation)

- ✓ Small batch size MM
  (seq_len=1, autoregressive)
- ✓ Lower utilization
- ✓ Memory-bound

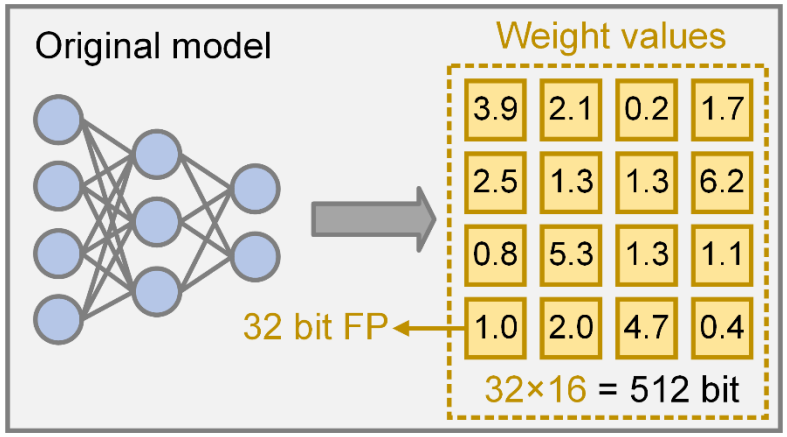
Motivation

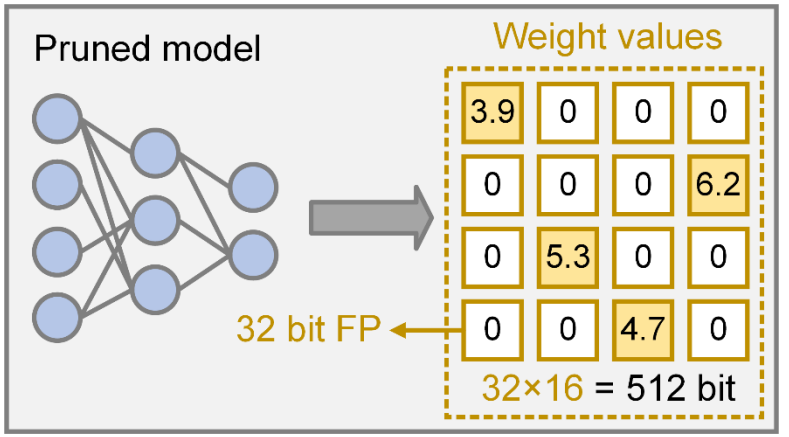
Matrix compression – pruning, Compressed Sparse Row (CSR)



Motivation

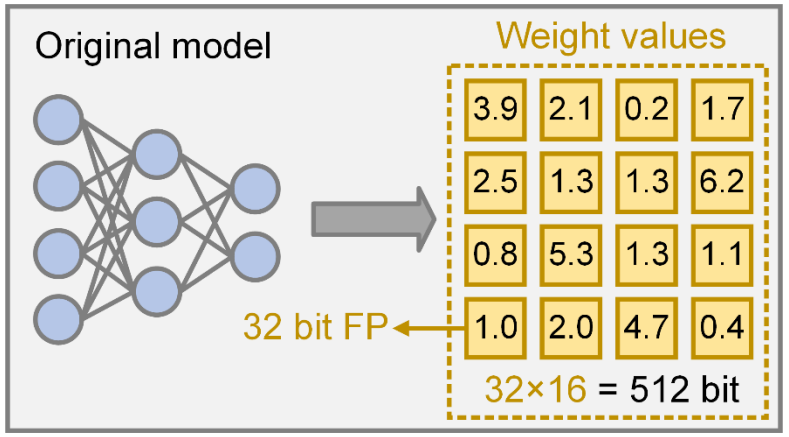
Matrix compression – pruning, Compressed Sparse Row (CSR)

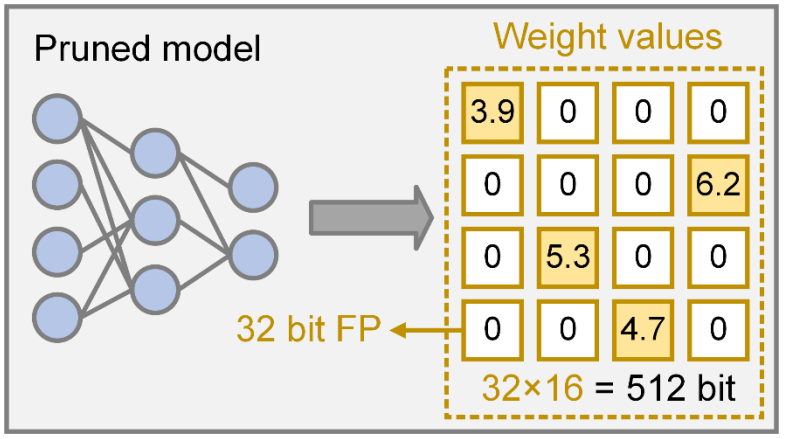




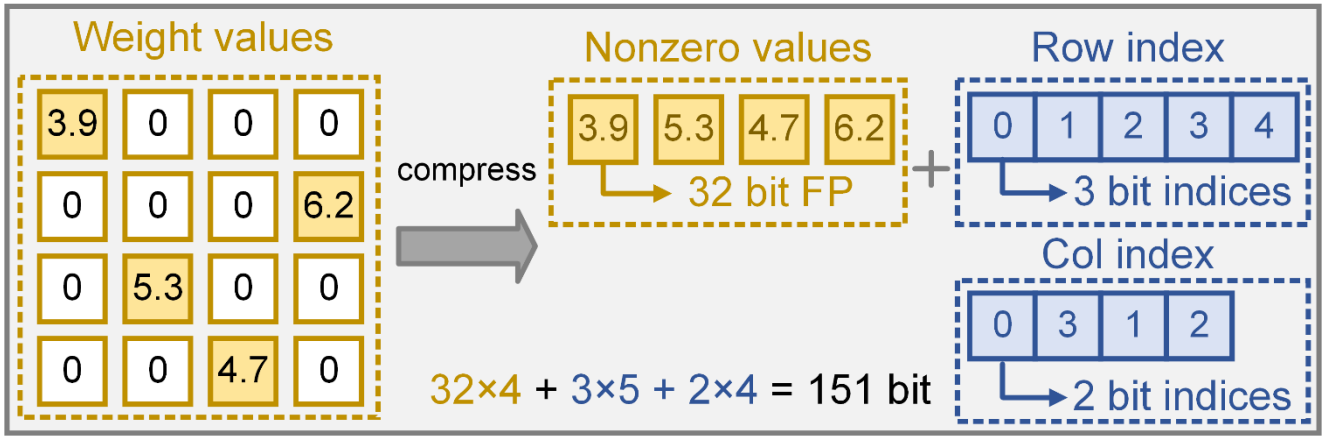
Motivation

Matrix compression – pruning, Compressed Sparse Row (CSR)





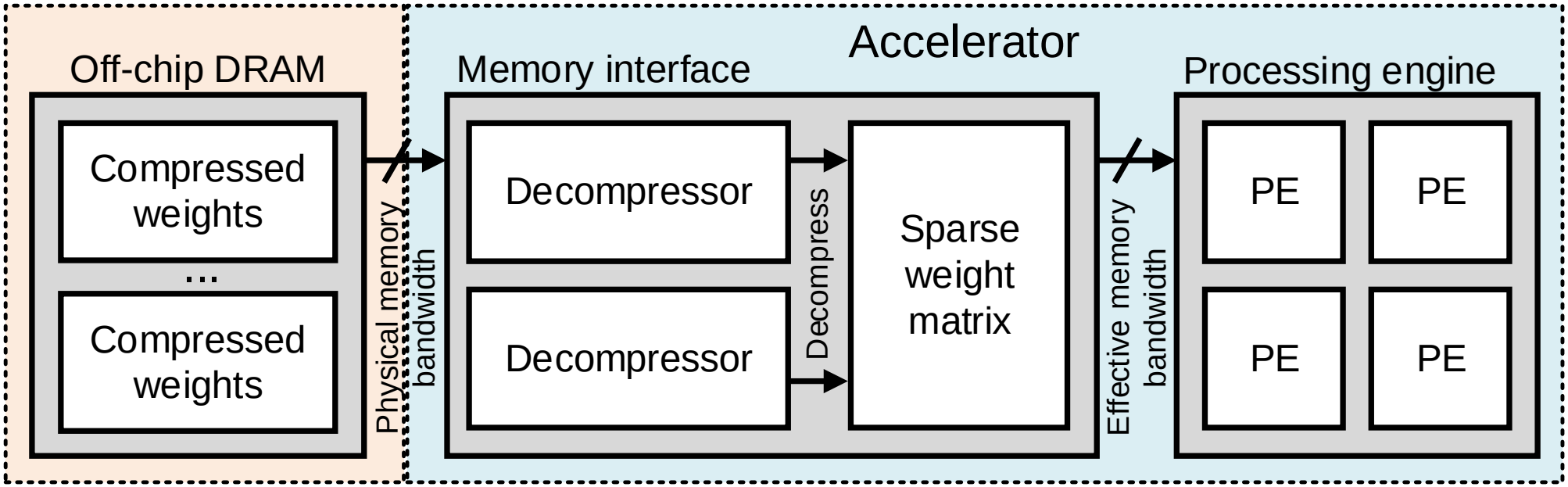
Compressed Sparse Row



Motivation

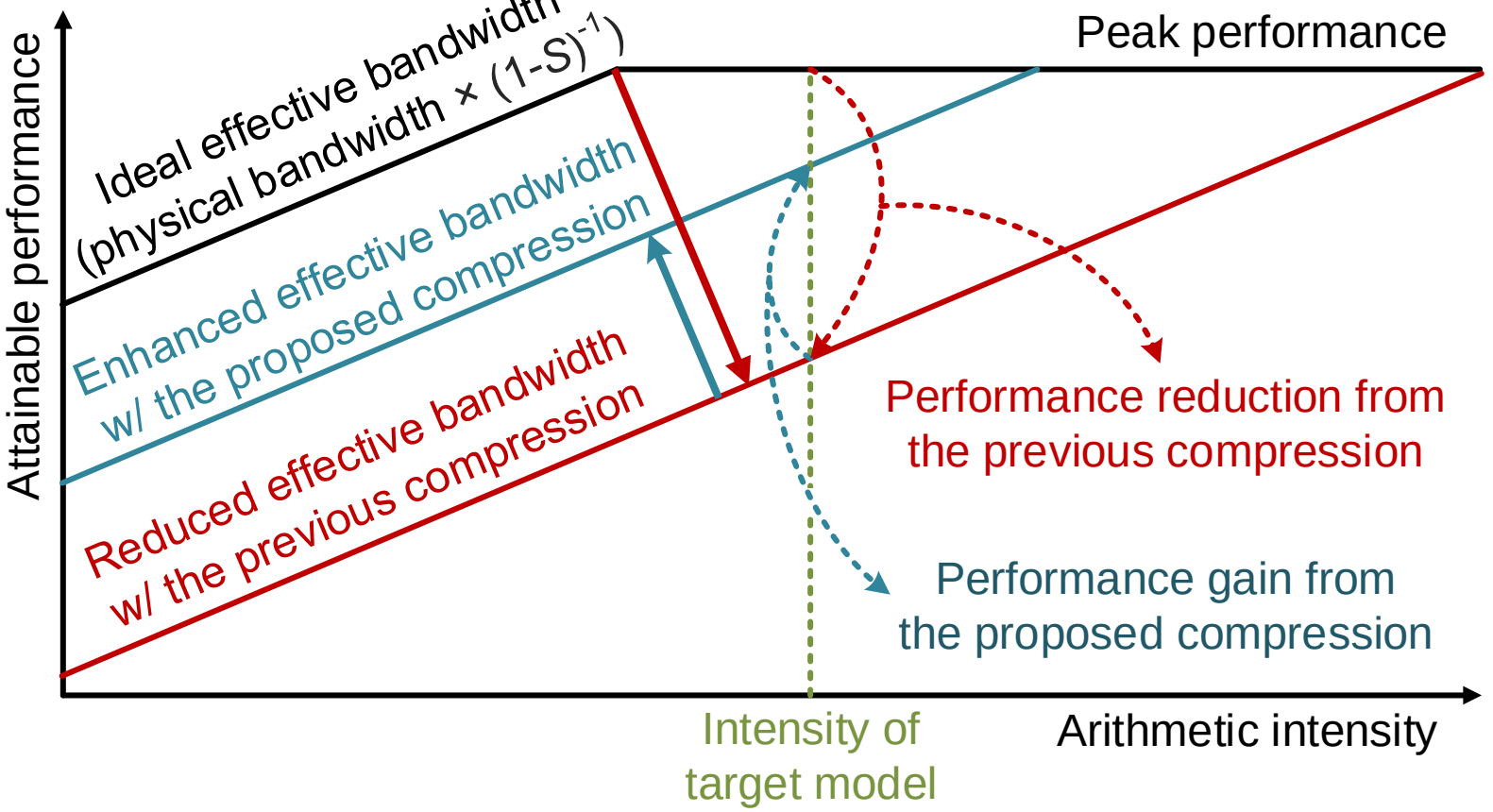
Effective bandwidth

✓ Effective bandwidth : bandwidth from the processing engine's perspective



Motivation

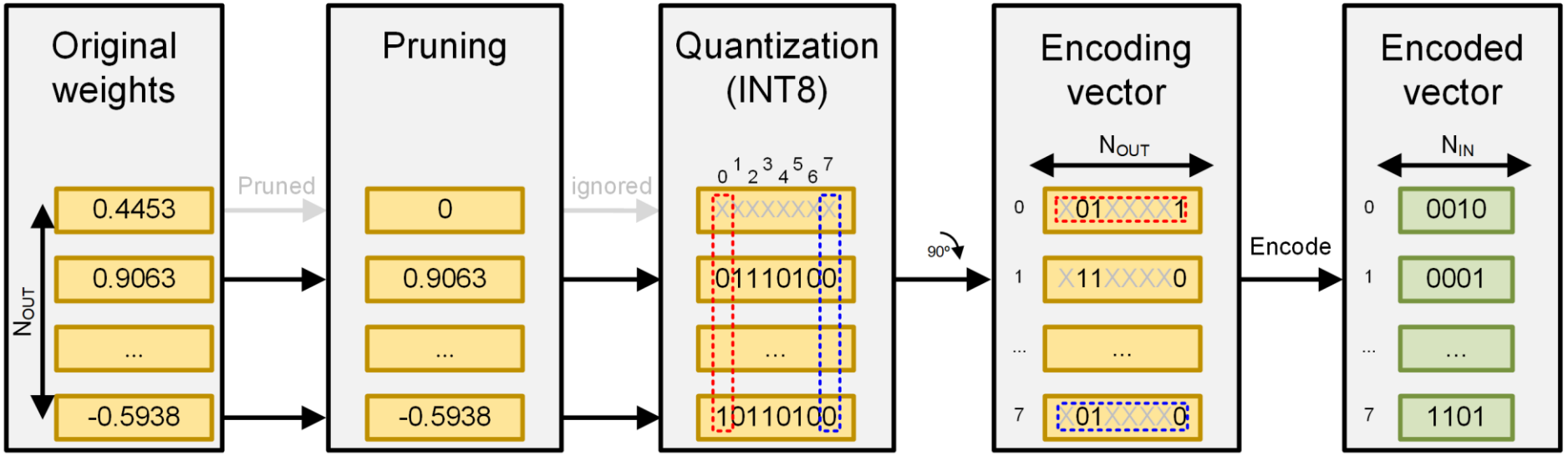
- ✓ Ideal compression ratio = $\frac{1}{1-S}$
- ✓ Ideal effective bandwidth = $\frac{1}{1-S}$ × physical bandwidth, S = sparsity



Backgrounds

XORNet compression process

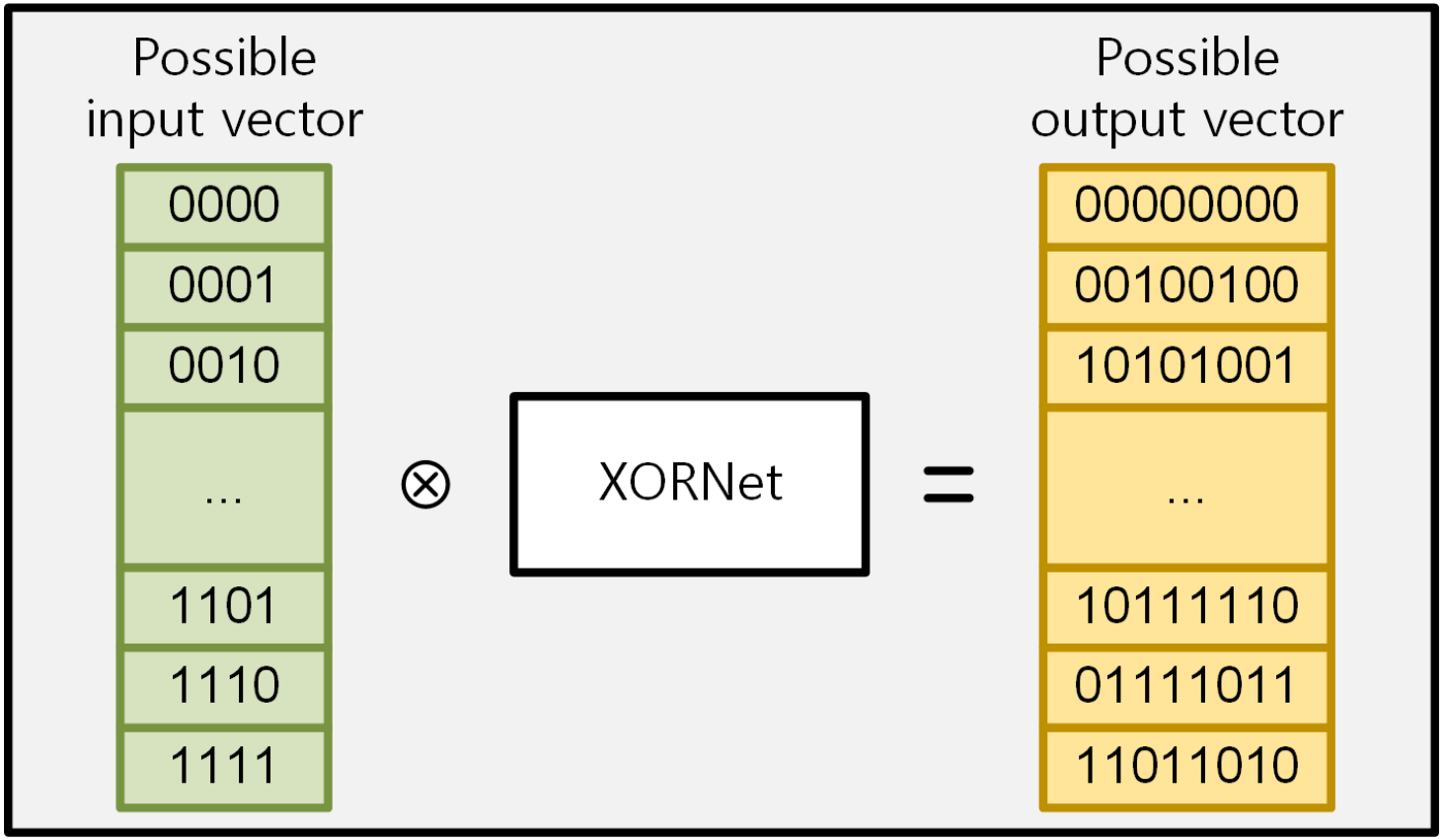
✓ Pruning, quantization, and XOR encoding



Backgrounds

XORNet encoding process

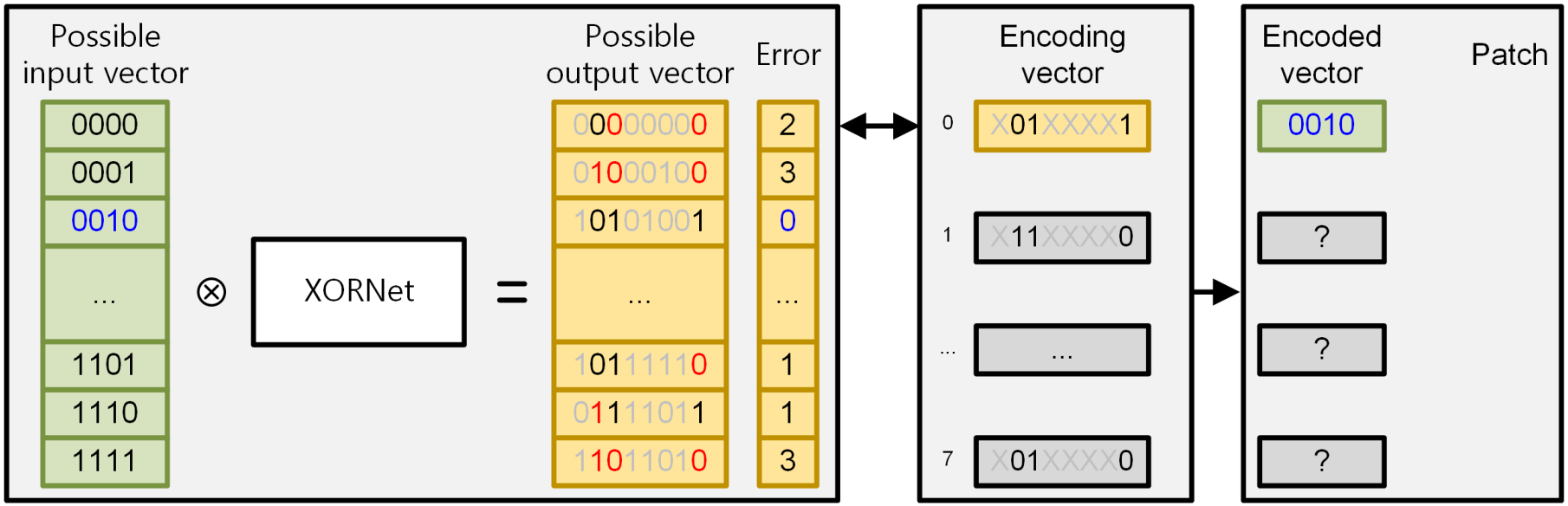
✓ Fixed length encoding with error patch



Backgrounds

XORNet encoding process

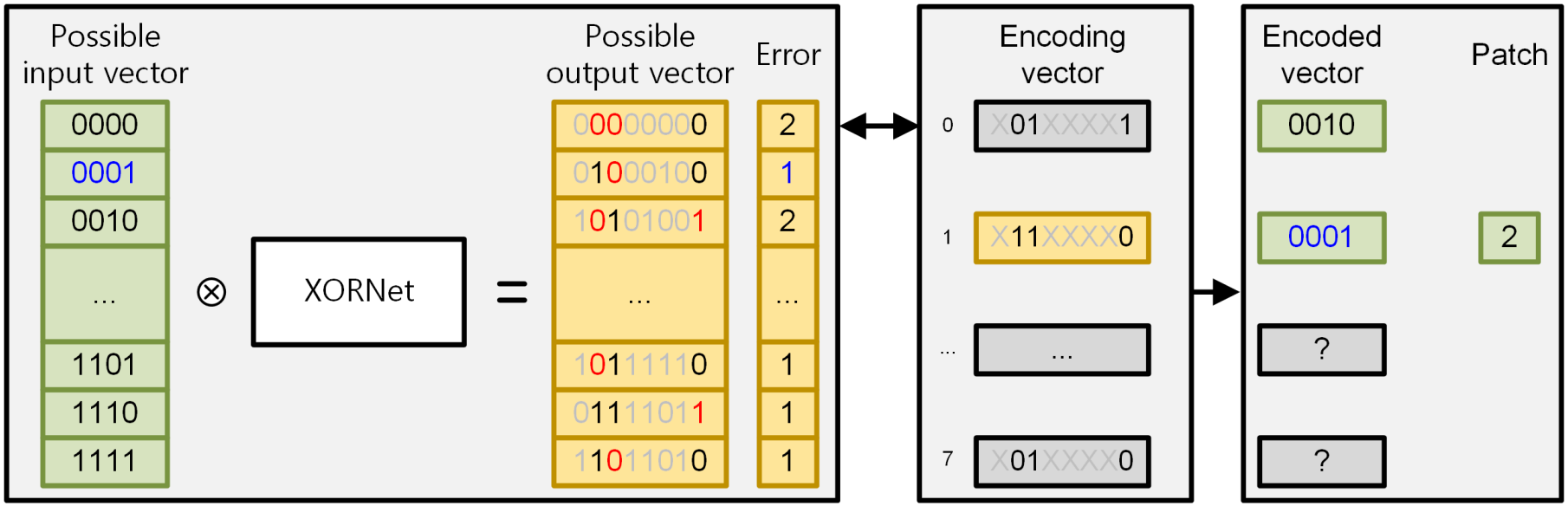
- ✓ Compare encoding sequence with possible output sequence
- ✓ Choose a minimum error sequence as an input (encoded) sequence



Backgrounds

XORNet encoding process

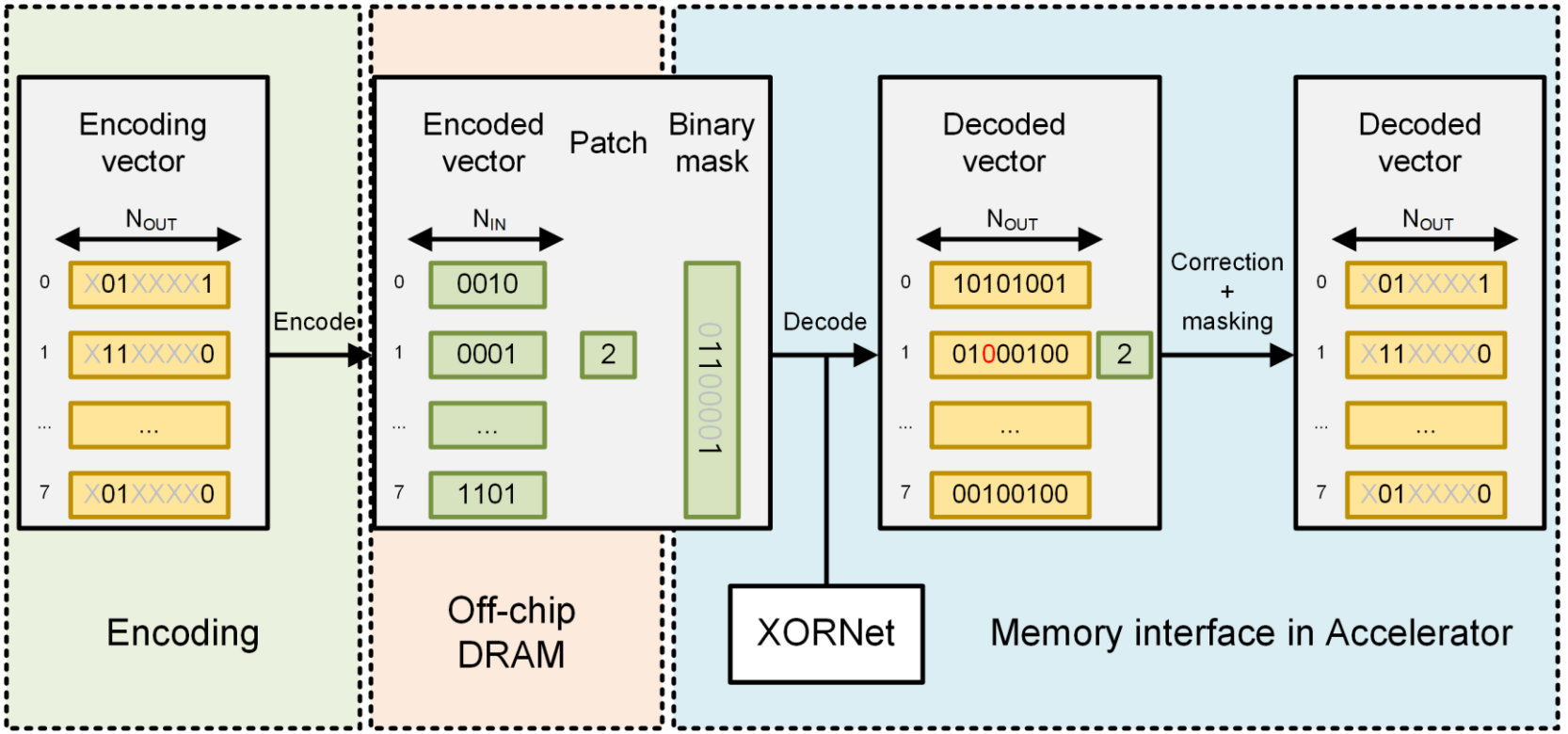
- ✓ Errorless matching is not guaranteed
- ✓ Add extra error index for correction (patch)



XORNet compression

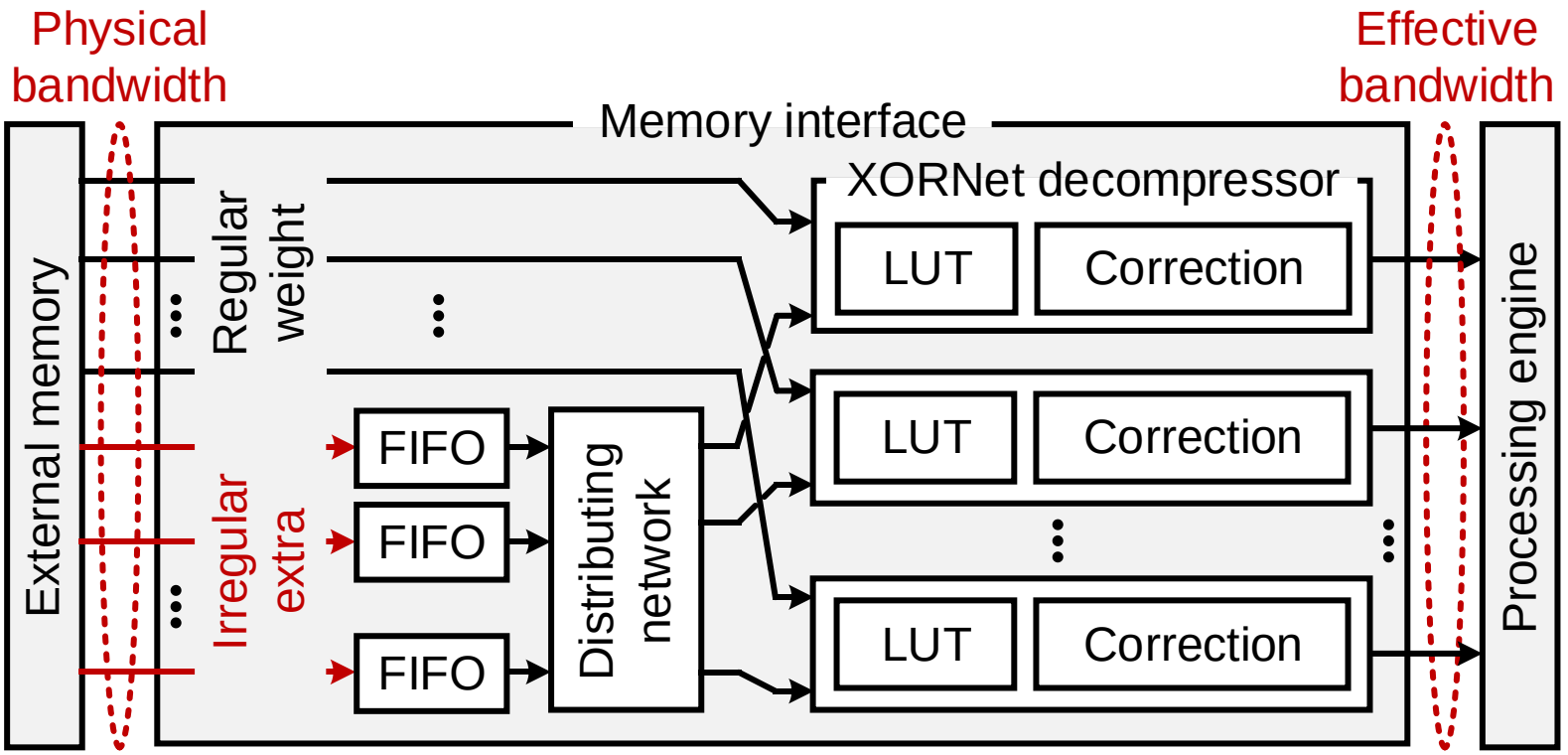
Detailed XORNet encoding and decoding process

✓ Encoded sequence and error position index (patch)



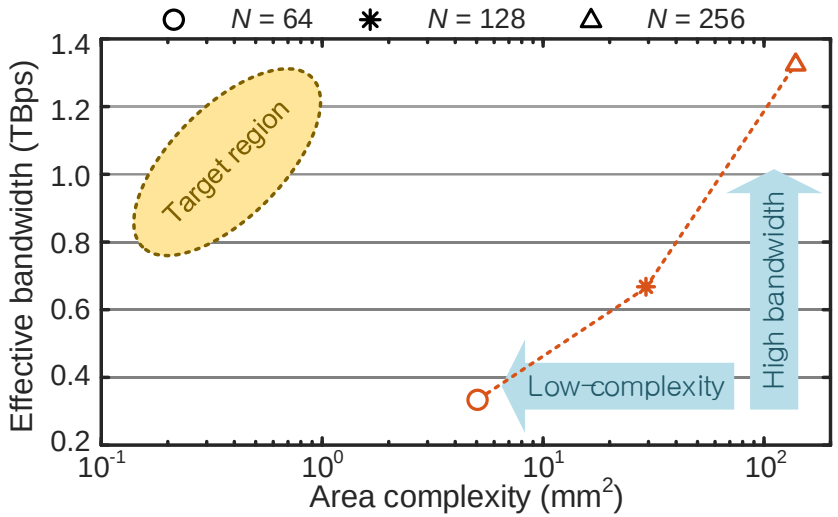
XORNet decoding hardware

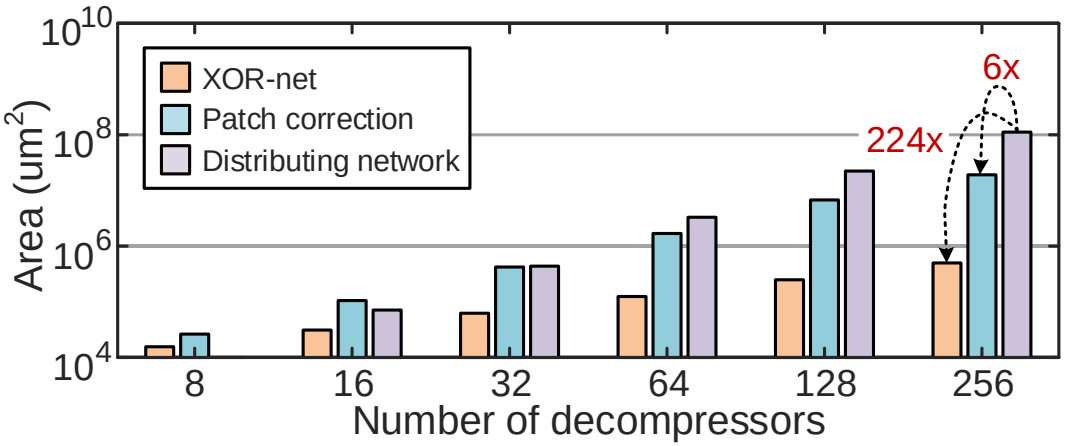
- ✓ XORNet decompressor with patch correction module
- ✓ Patch FIFO, distributing network



XORNet decoding hardware

- ✓ The number of decompressor (N)
- ✓ N=256 supports 960GBps memory bandwidth (RTX 3090 : ~940GBps)
- ✓ The area complexity increases exponentially with memory bandwidth
- ✓ Distributing network dominates area overhead in high-bandwidth system

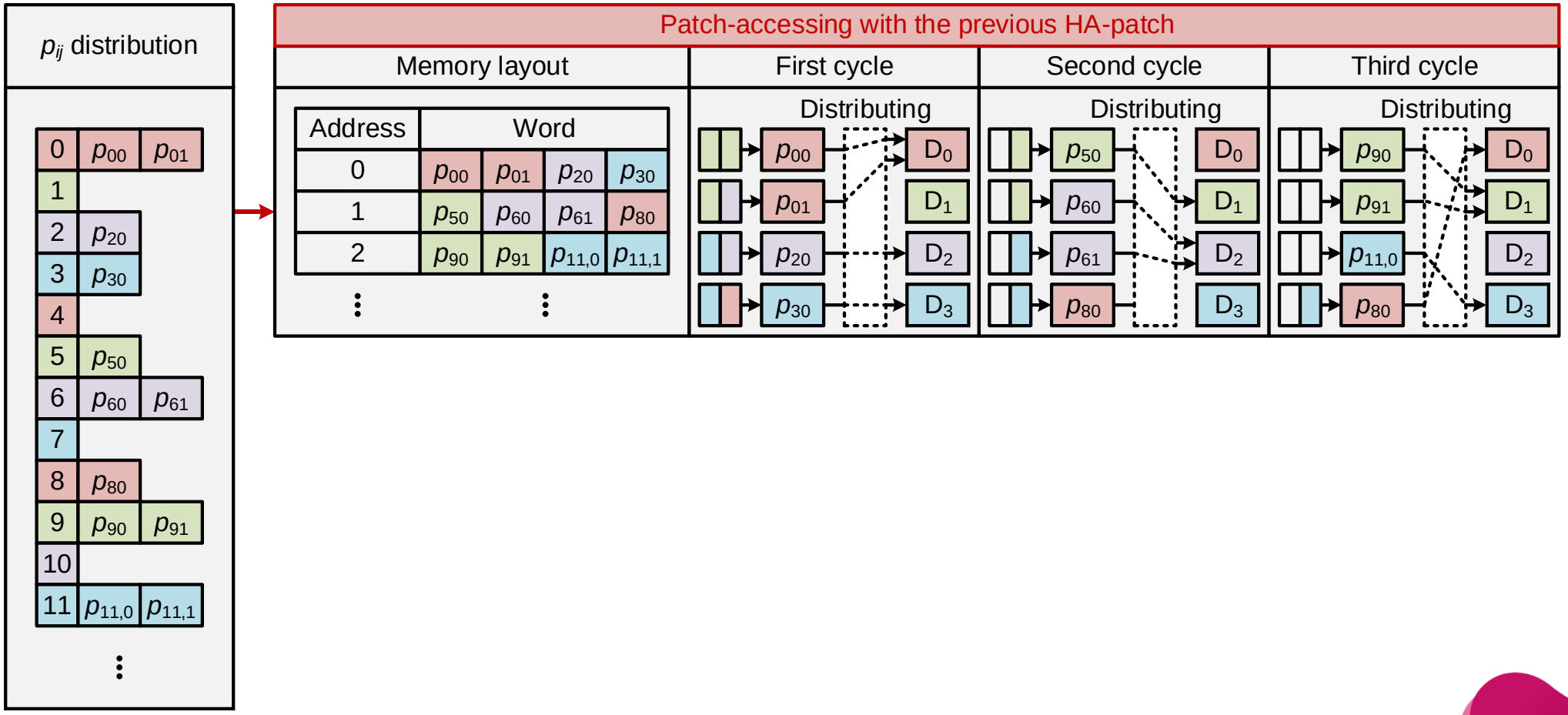




Interface for high-bandwidth system

Horizontally aligned (HA) patch memory

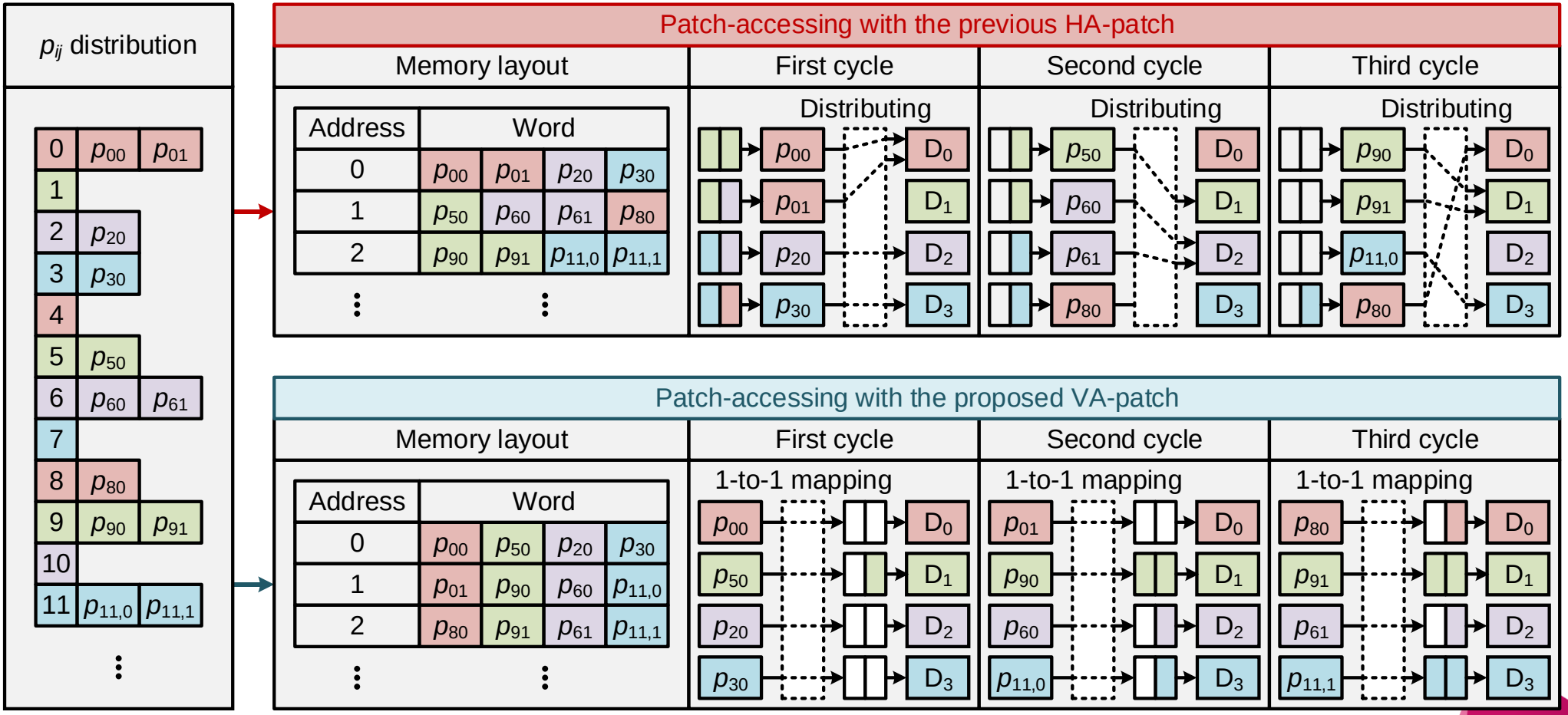
- ✓ p_{ij} represents j-th patch in i-th vector
- ✓ HA-patch save the patch sequentially



Interface for high-bandwidth system

Vertically aligned (VA) patch memory

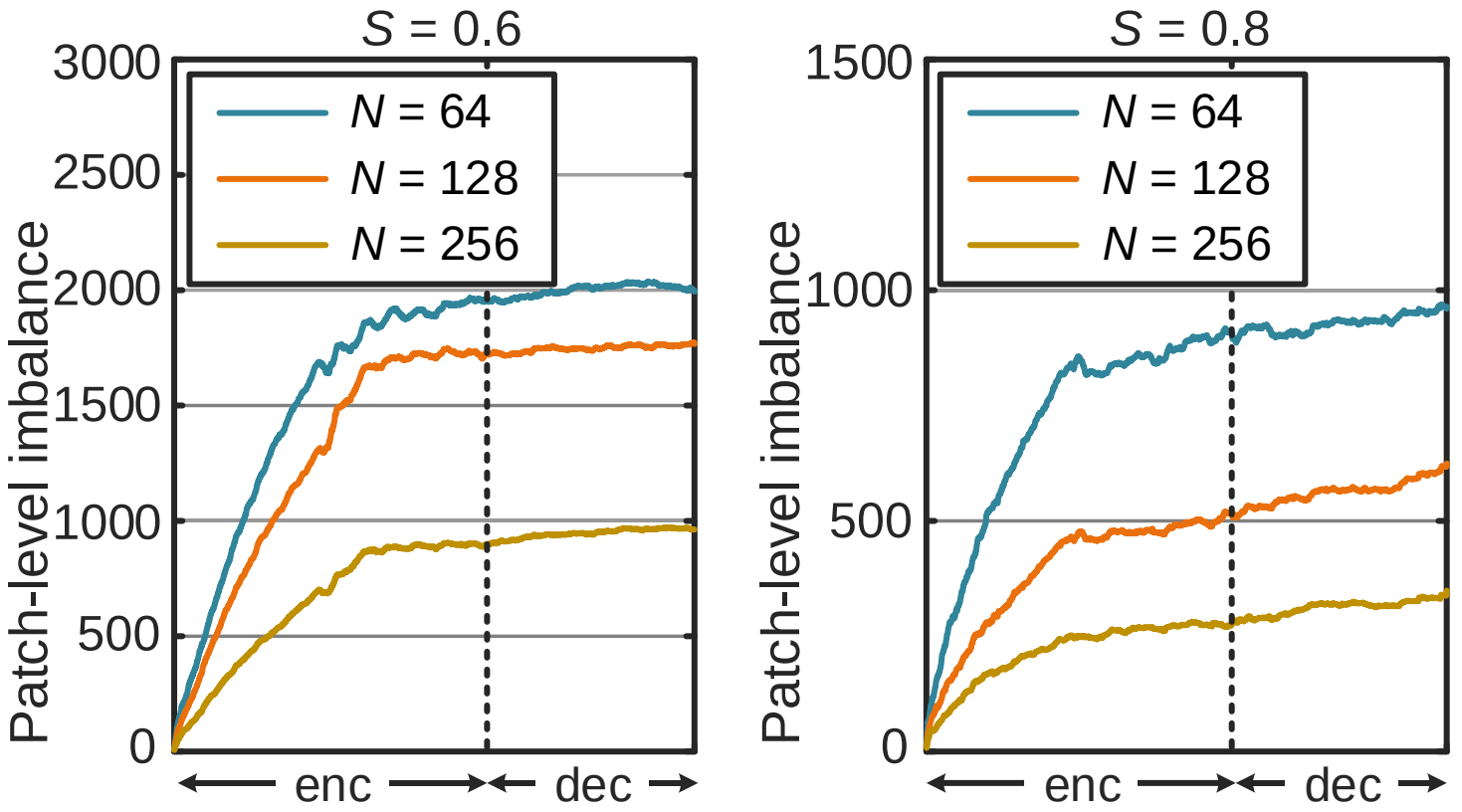
- ✓ VA-patch saves patch in the order of decompressor
- ✓ Large buffer size is required due to the patch imbalance



Patch imbalance along with decompressors

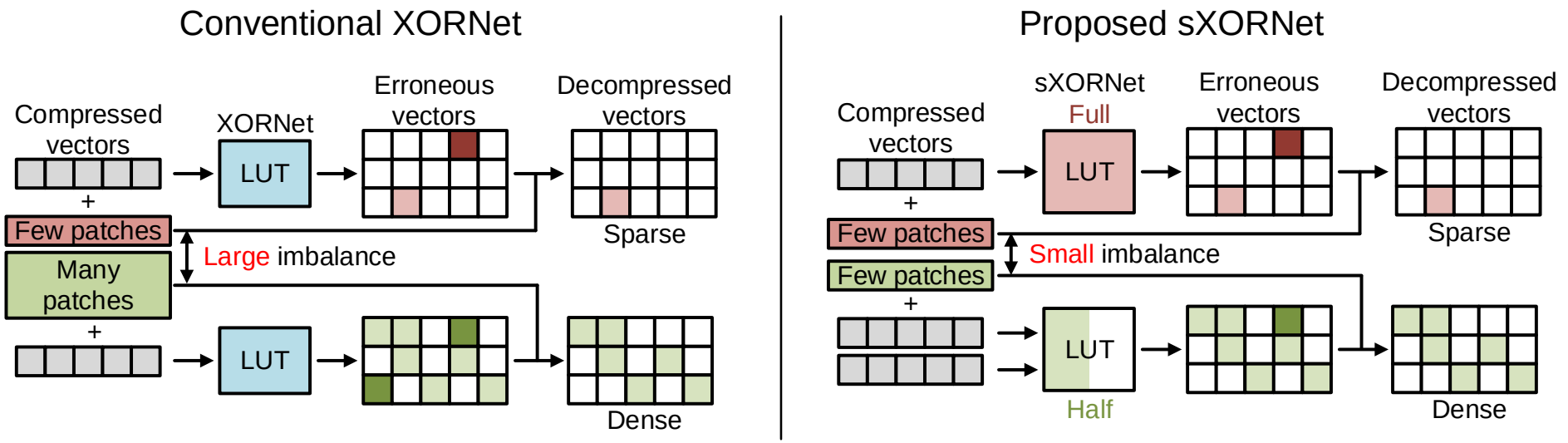
Vertically aligned (VA) patch memory

- ✓ Transformer model, WMT en-de task
- ✓ Large patch imbalance cause large patch buffer size



Stacked XORNet (sXORNet)

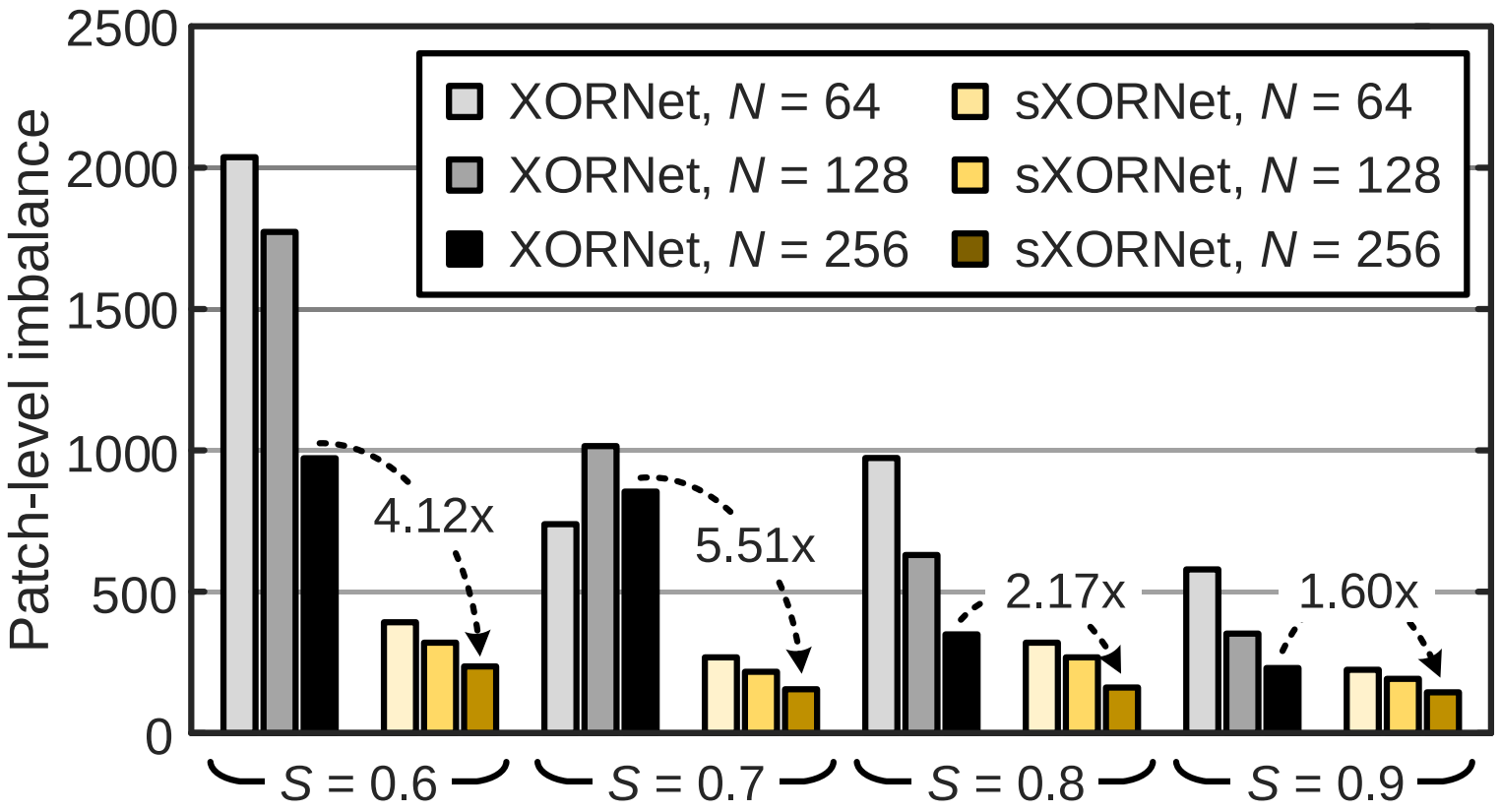
- ✓ Conventional XORNet uses single LUT independent with sparsity
- ✓ Proposed stacked XORNet adaptively uses LUT based on the sparsity
  - Generates fewer patches, leading to low patch imbalance



Experimental results

Patch imbalance comparison for VA-patch

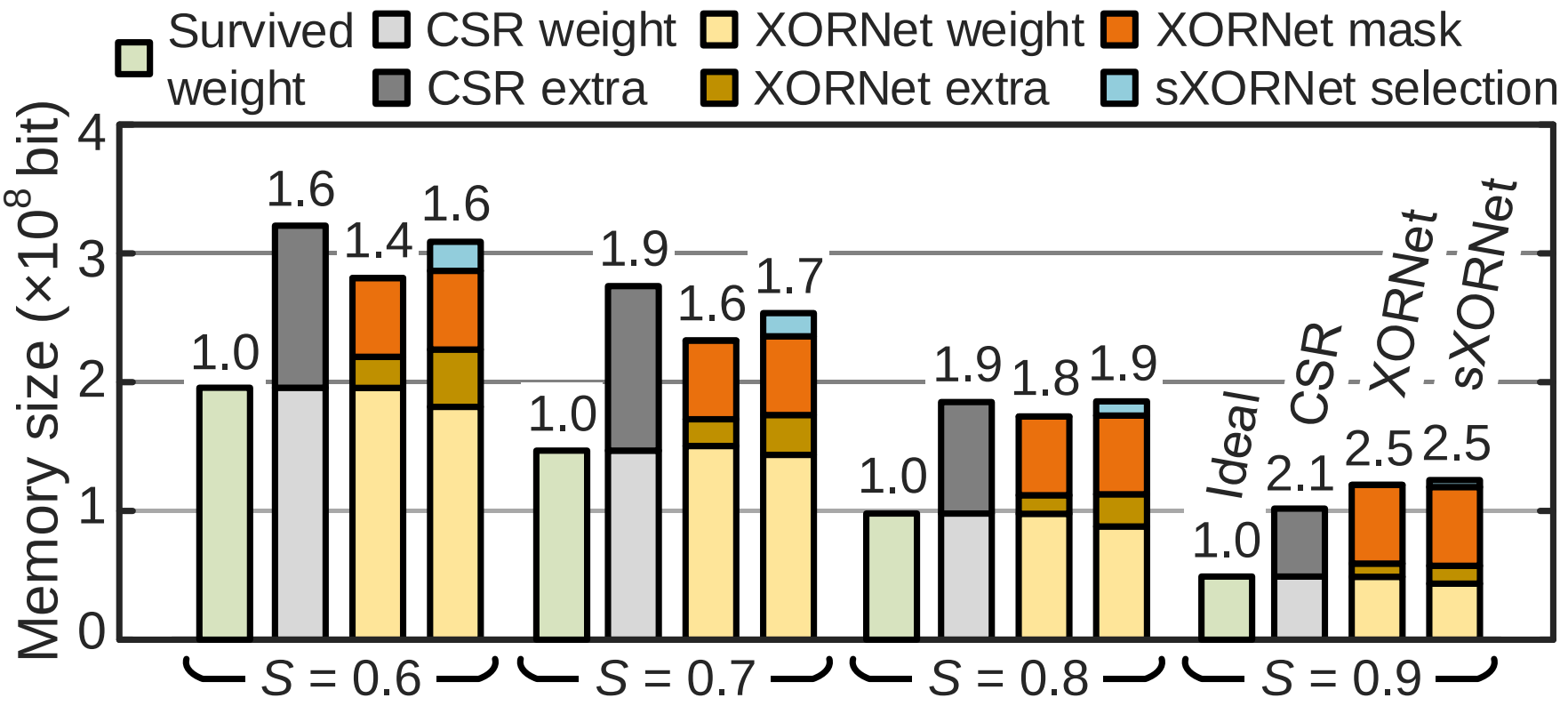
- ✓ Transformer model, WMT en-de task
- ✓ The patch imbalance decreased with the proposed sXORNet in all case



Experimental results

Compression quality for different compression techniques

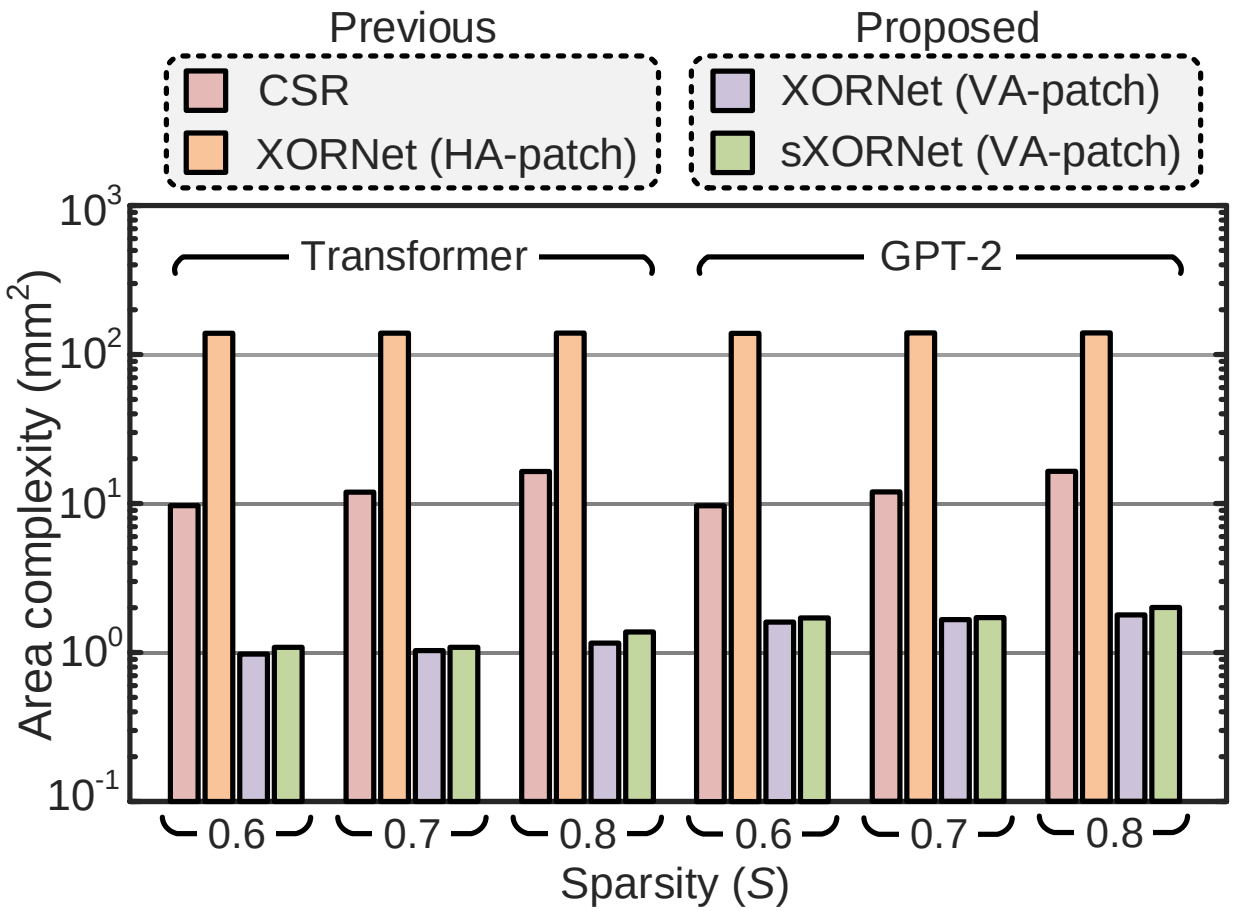
✓ 8-bit quantized transformer model



Experimental results

Sparsity and area complexity

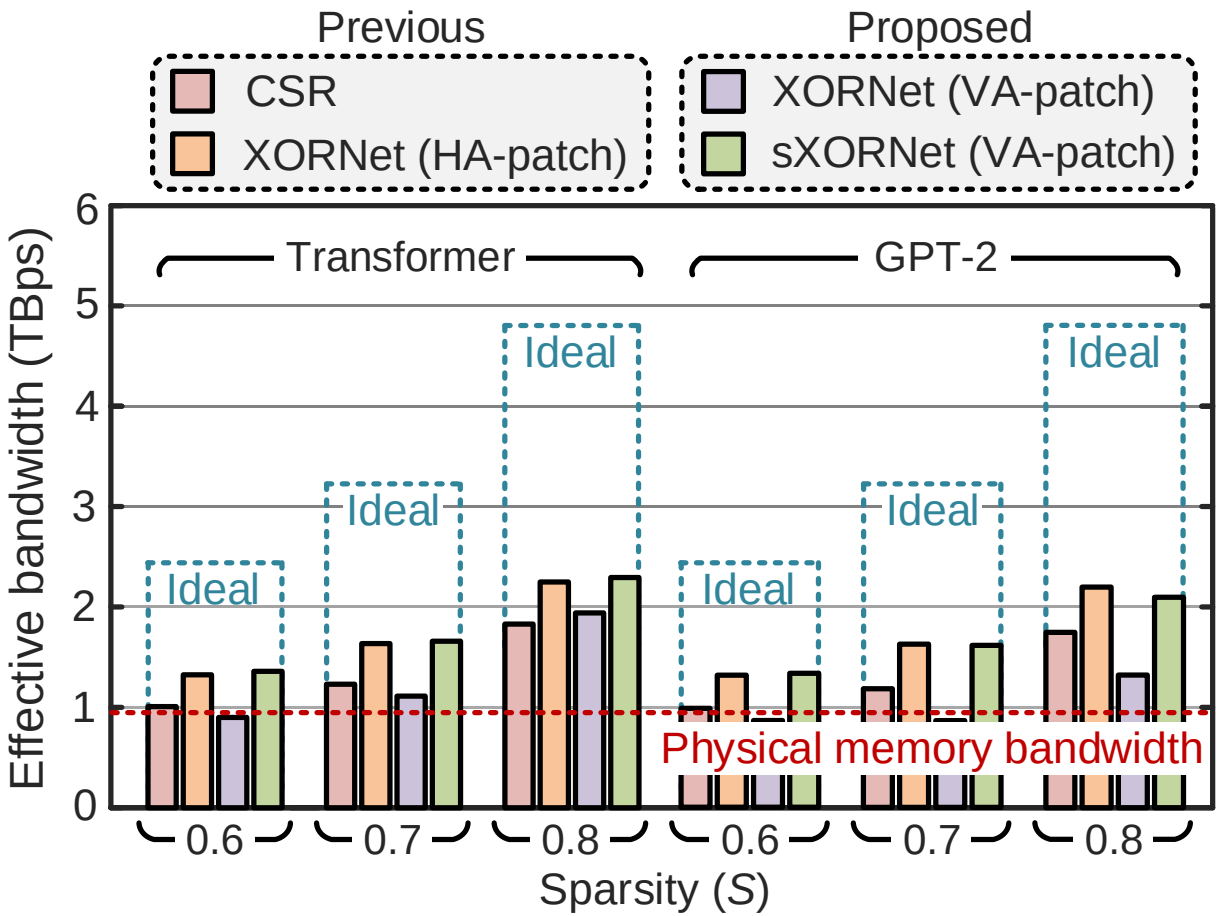
✓ N=256 (960GBps)



Experimental results

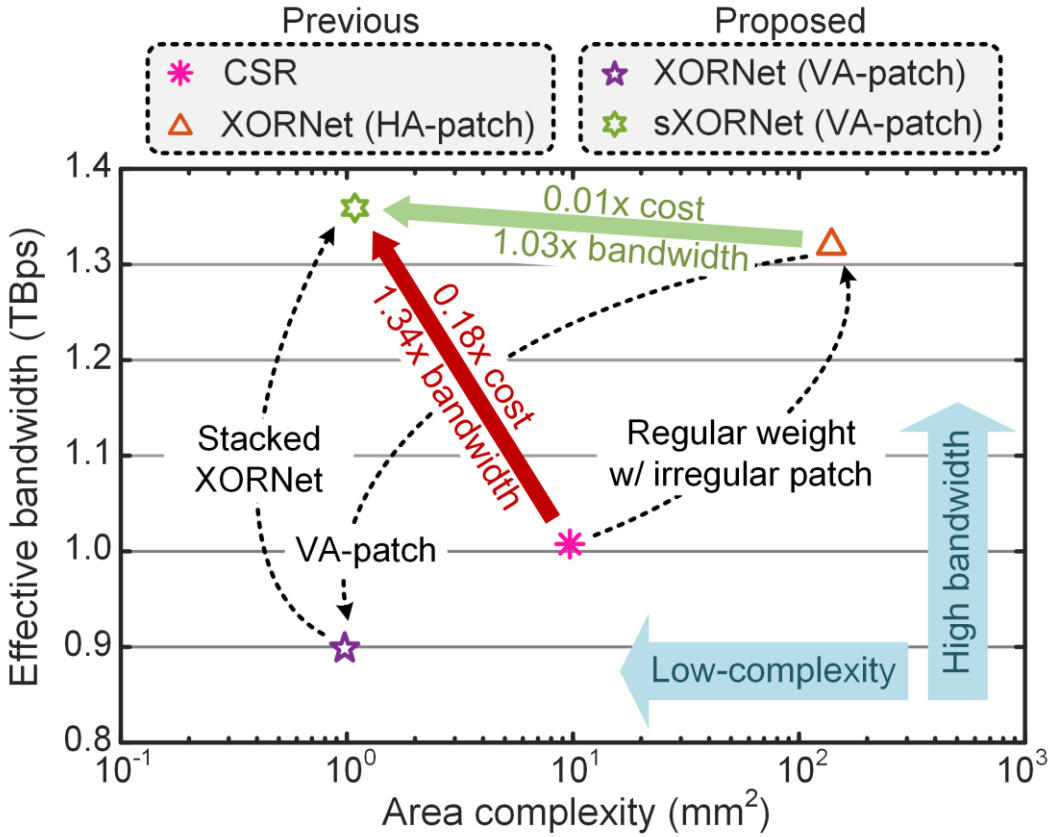
Sparsity and effective bandwidth

✓ N=256 (960GBps)



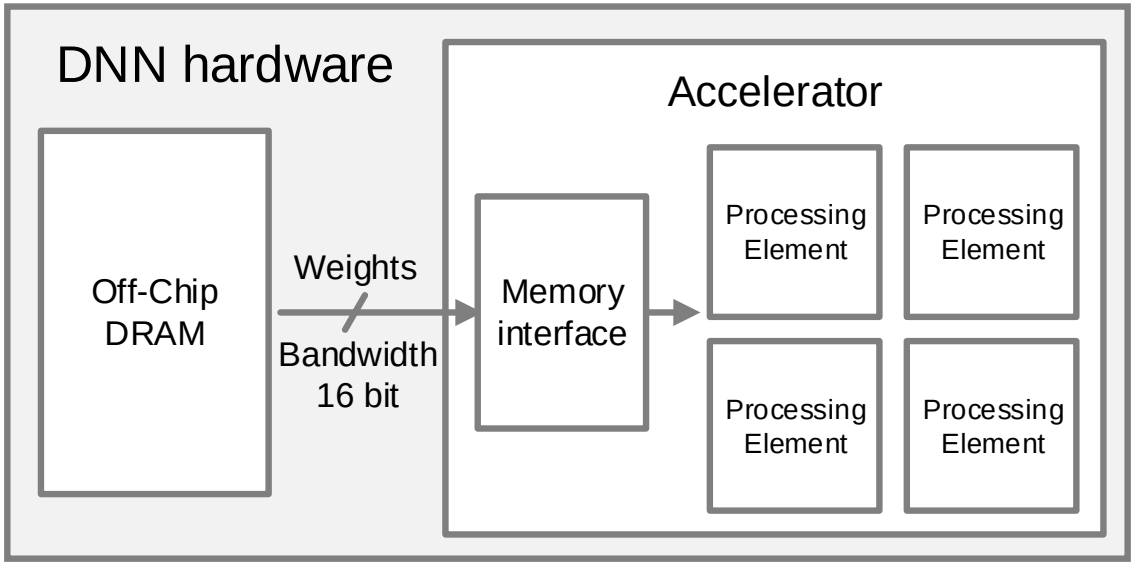
Conclusion

- ✓ Investigated interface-level overhead for different compression types
- ✓ Proposed XORNet optimized hardware patch architecture (VA-patch)
- ✓ Proposed imbalance considering algorithm (sXORNet)
- ✓ Achieved low-area complexity, high-throughput memory interface

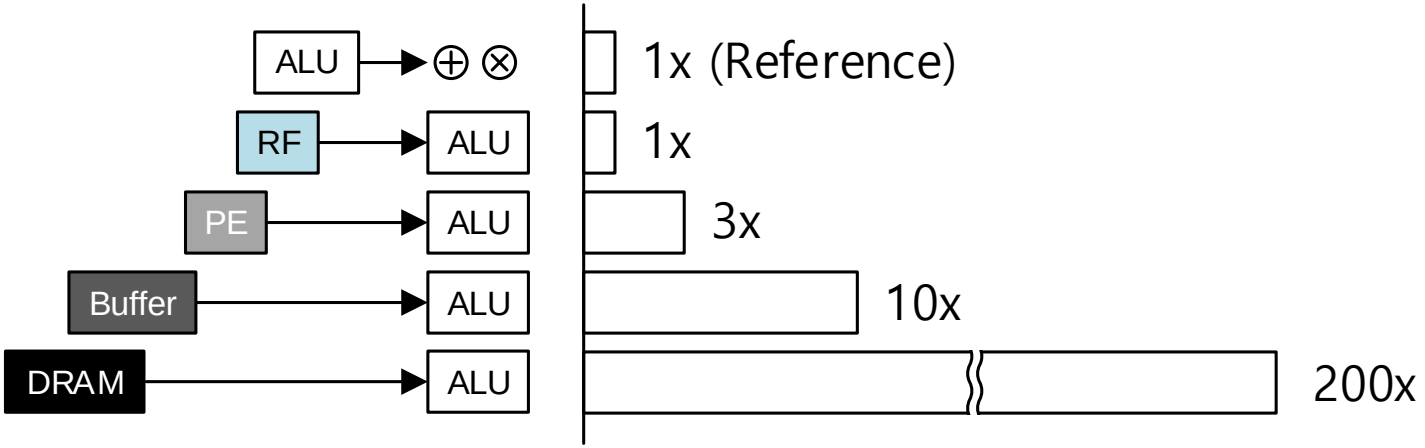


Thank you

Appendix A. DNN hardware with memory interface



Data Movement Energy Cost [1]



[1] Y. -H. Chen *et al*., "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," *ISCA*, 2016.

Appendix B. area complexity and effective bandwidth

N	Effective bandwidth (TBps)					Area complexity (mm^2)			
	Ideal	CSR	XORNet HA-patch	XORNet VA-patch ($B=256$)	sXORNet VA-patch ($B=256$)	CSR	XORNet HA-patch	XORNet VA-patch ($B=256$)	sXORNet VA-patch ($B=256$)
Transformer (Vaswani et al., 2017) ($S=0.6$)									
64	0.60	0.25	0.33	0.17	0.30	1.057	5.053	0.244	0.271
128	1.20	0.51	0.66	0.34	0.63	3.165	29.268	0.489	0.542
256	2.40	1.01	1.32	0.90	1.36	9.687	139.109	0.977	1.085
GPT-2 small (Brown et al., 2020) ($S=0.6$)									
64	0.60	0.25	0.33	0.09	0.27	1.057	5.053	0.244	0.271
128	1.20	0.50	0.66	0.23	0.58	3.165	29.268	0.489	0.542
256	2.40	0.99	1.32	0.61	1.21	9.687	139.109	0.977	1.085
ResNet-50 (He et al., 2016) ($S=0.7$)									
64	0.80	0.26	0.41	0.31	0.40	1.227	5.067	0.258	0.271
128	1.60	0.51	0.82	0.62	0.79	3.804	29.295	0.515	0.542
256	3.20	1.02	1.64	1.23	1.57	11.925	139.163	1.031	1.085