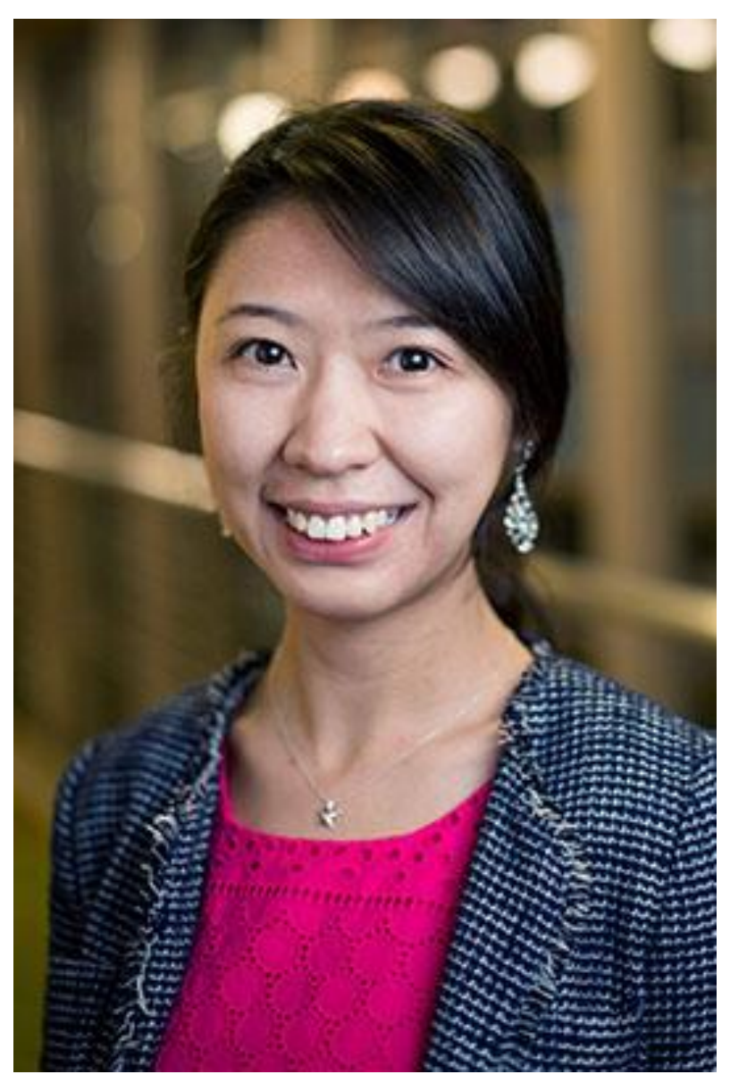


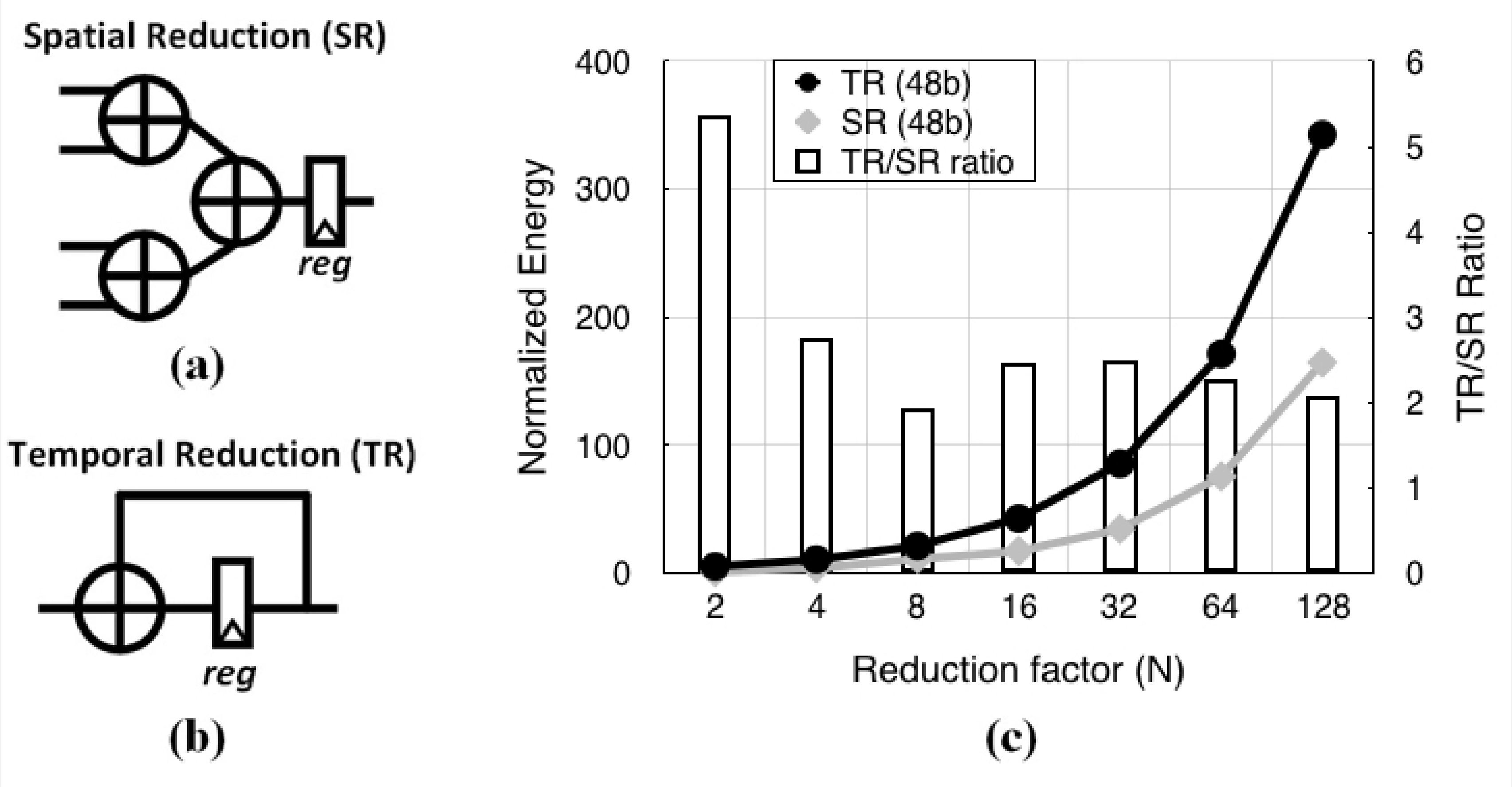
Stitch-X: An Accelerator Architecture for Exploiting Unstructured Sparsity in Deep Neural Networks

**Ching-En Lee, Yakun Sophia Shao, Jie-Fang Zhang,
Angshuman Parashar, Joel Emer, Stephen W. Keckler, Zhengya Zhang**

NVIDIA.
UNIVERSITY OF MICHIGAN

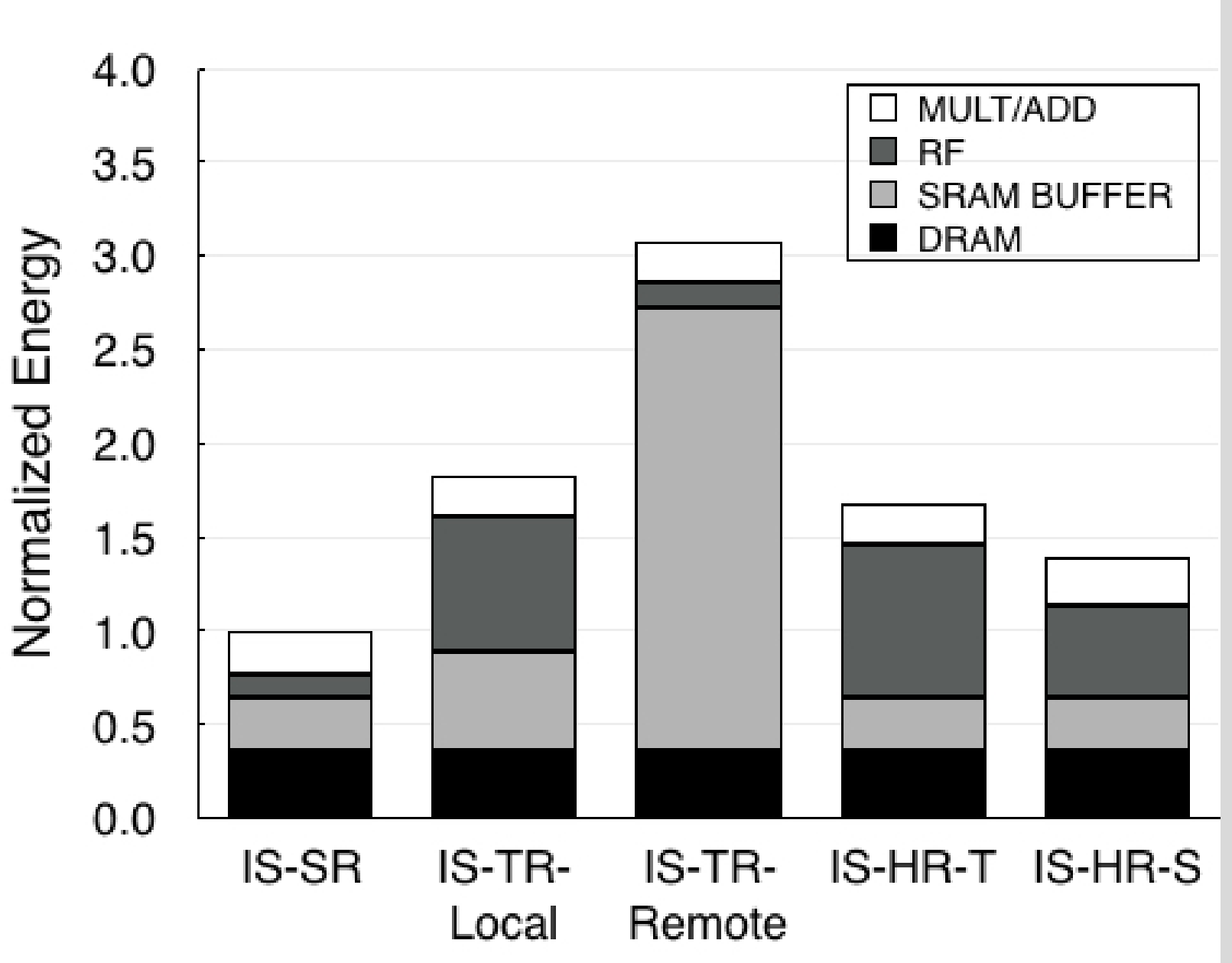
Spatial vs Temporal Reduction

- **Spatial Reduction (SR)** does partial-sum accumulation spatially with an adder tree without explicit storage.
- **Temporal Reduction (TR)** reduces over time by using a single adder to accumulate one partial sum per time.
- SR is *always* more energy efficient than TR, but TR is more flexible to support accumulation across different dimensions.

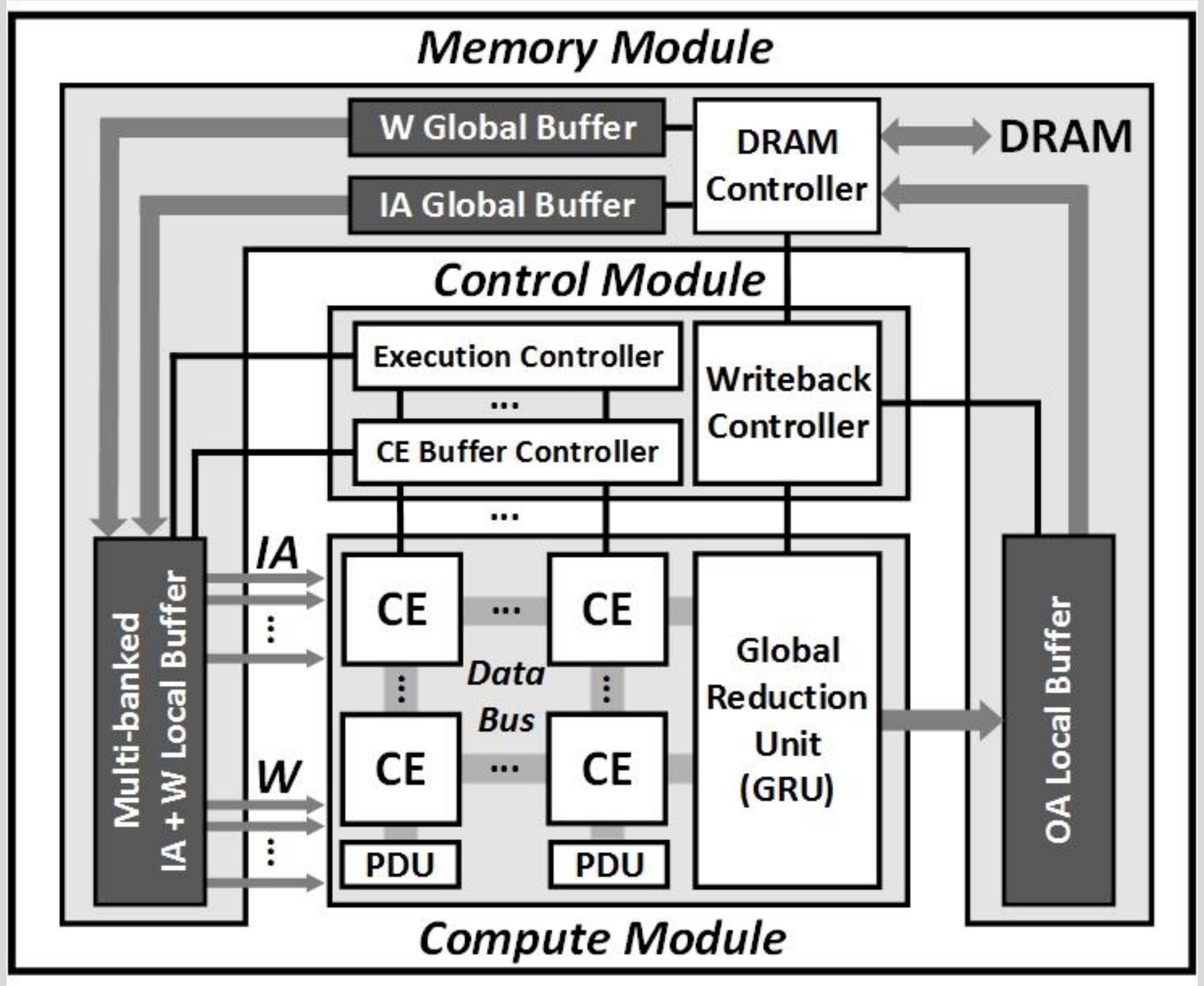


Dataflow Taxonomy		Spatial Reduction	Temporal Reduction	Hybrid Reduction
Data Reuse	Output Stationary		ShiDianNao DnnWeaver	
	Input (IA/W) Stationary	NVDLA BrainWave	SCNN EIE	**Stitch-X**
	No Local Reuse	DianNao DaDianNao Cambricon-X Cnvlutin	TPU Minerva	
	Row Stationary		Eyeriss	

There can be as large as **3x** energy difference for architectures of the same data reuse patterns but different reduction mechanisms.



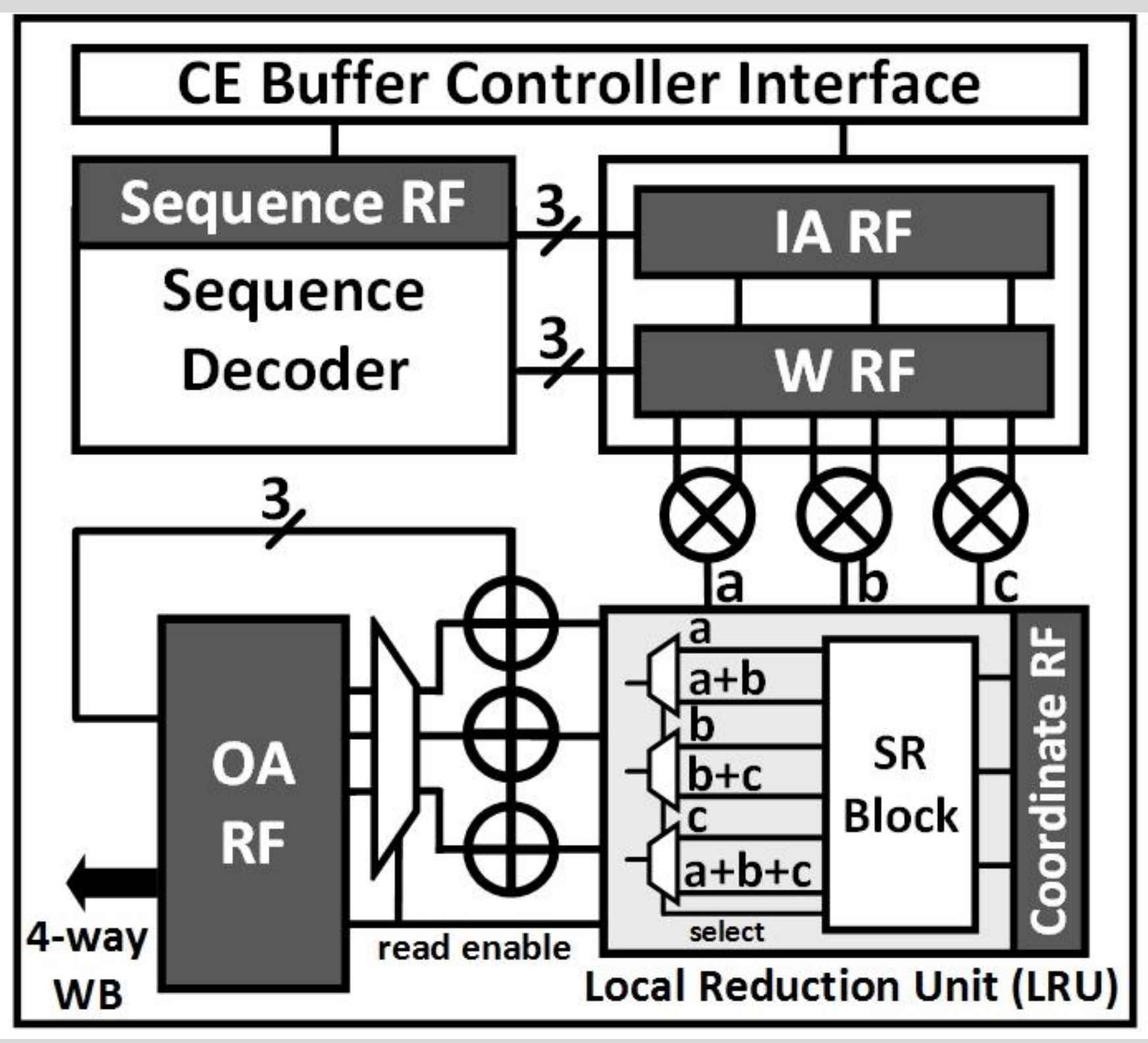
Stitch-X Architecture



- **Compute Module:**
  - Computing Elements
  - Parallelism Discovery Unit
  - Global Reduction Unit
- **Memory Module**
  - Global Buffer
  - Multi-backed IA and W Buffers
  - OA Buffer
- **Control Module**
  - Execution
  - CE Buffer
  - Writeback

Two-Level Hybrid Reduction:

- Local Reduction Unit
  - Flexible 3:1 Spatial Reduction Support.
  - Temporal Reduction with output register.
- Global Reduction Unit
  - Flexible Spatial Reduction Across CEs.
- Minimize memory bandwidth and access energy.

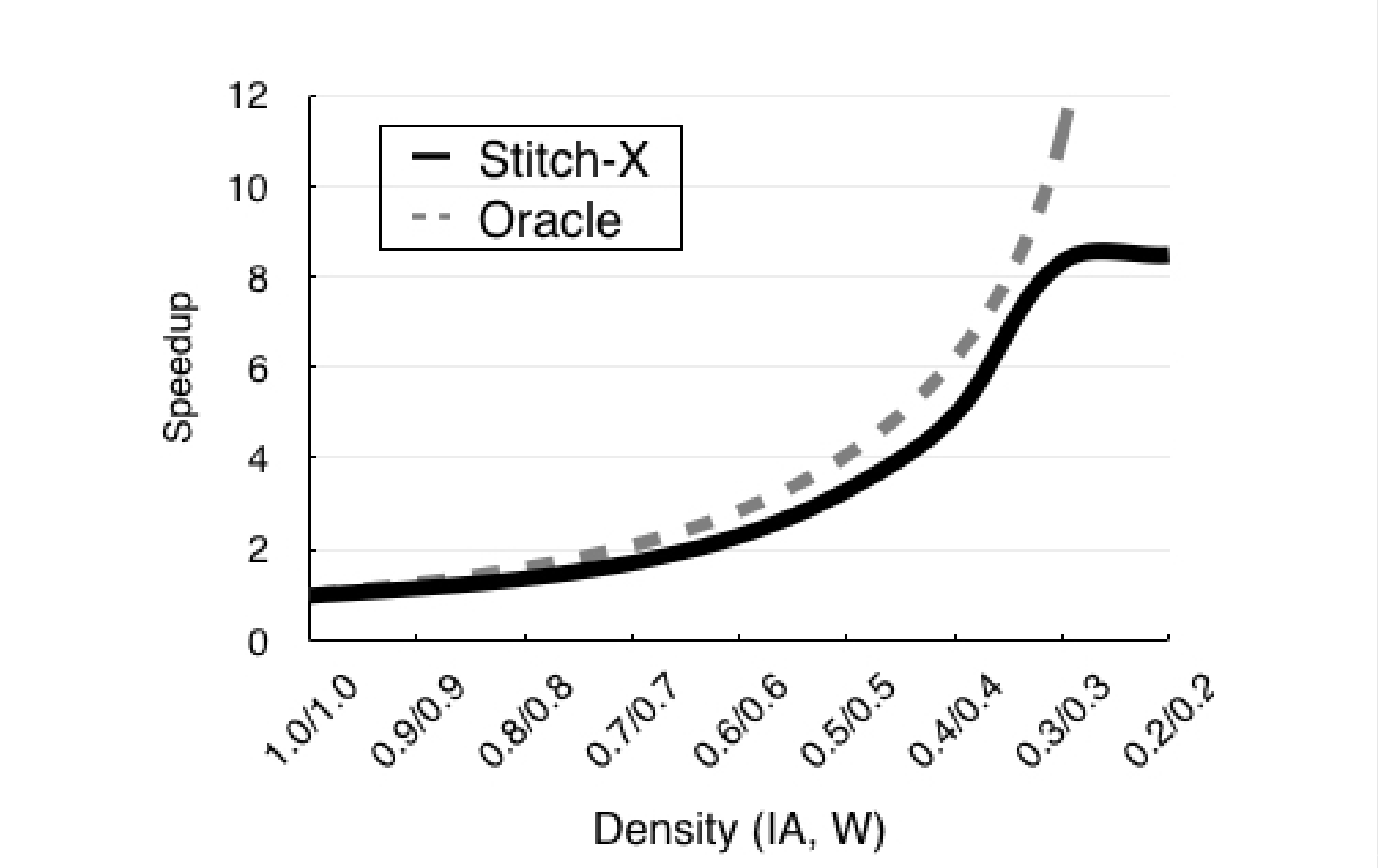


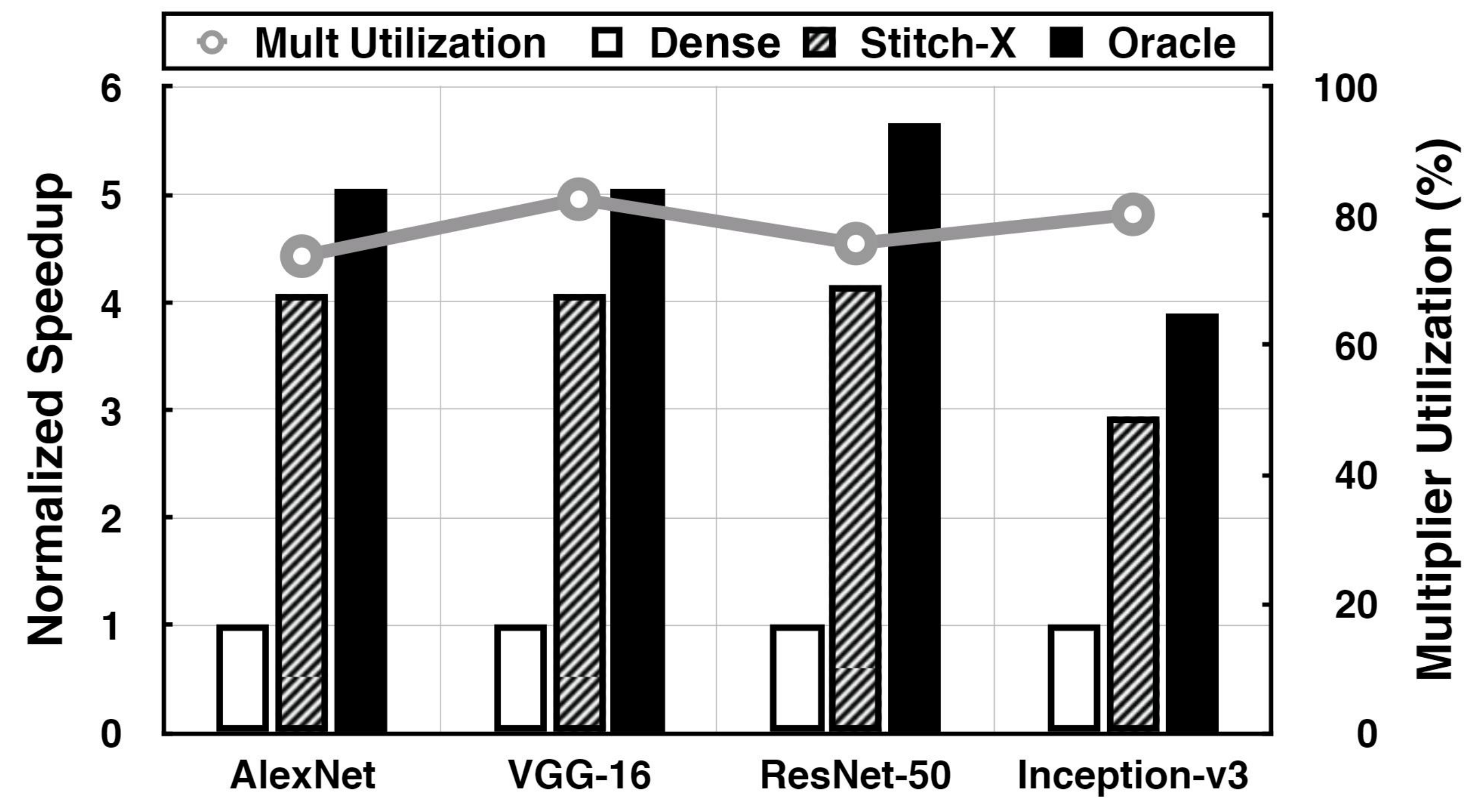
Parallelism Discovery Unit:

- Finds all reducible pairs of non-zero IA and W from compacted arrays dynamically.
- Performs a parallel search of IA and W indexes across multiple CEs.
- Improves multiplier utilization.

Evaluations

Stitch-X achieves a **3.8X** speedup and improves ED^2P by a factor of **10.3X** on average compared to an efficient, dense DNN accelerator. Compared to a state-of-the-art sparse DNN accelerator, Stitch-X delivers **1.6X** better performance.


Speedup
Stitch-X
Oracle
1.0/1.0
0.9/0.9
0.8/0.8
0.7/0.7
0.6/0.6
0.5/0.5
0.4/0.4
0.3/0.3
0.2/0.2
Density (IA, W)


○ Mult Utilization
□ Dense
▨ Stitch-X
■ Oracle
Normalized Speedup
Multiplier Utilization (%)
AlexNet
VGG-16
ResNet-50
Inception-v3