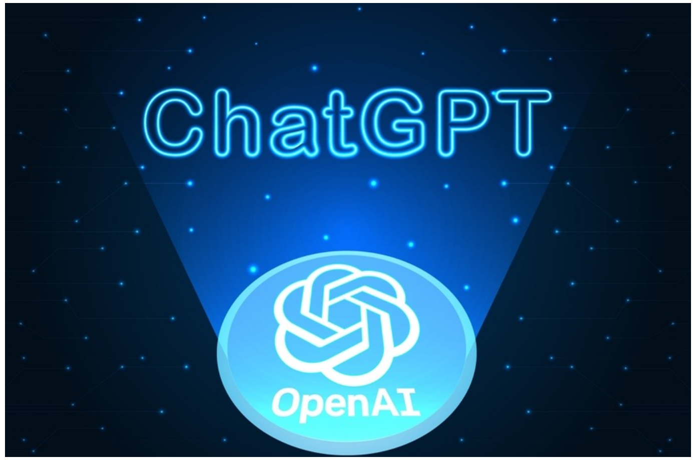
HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity

Yannan Nellie Wu[1], **Po-An Tsai**[2], Saurav Muralidharan[2], Angshuman Parashar[2], Vivienne Sze[1], Joel S. Emer[1,2]

[1]MIT, [2]NVIDIA

http://emze.csail.mit.edu/highlight

Many Applications Involve DNNs



Natural Language Processing

Autonomous Navigation

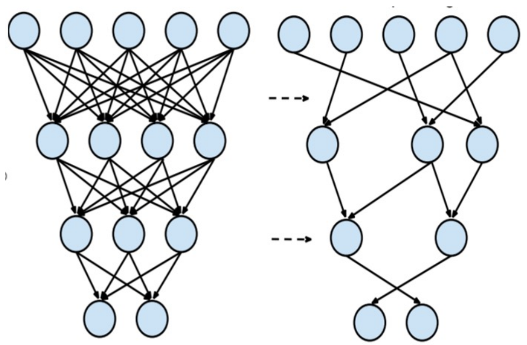
Medical Imaging & Diagnostics

data and computation intensive
subject to prediction accuracy & latency requirements

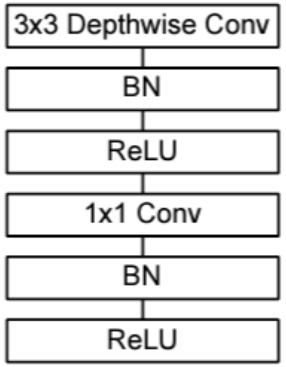
In great need of optimizations and accelerations

Different DNN Optimizations Introduce Different Sparsity

Optimizations to Reduce Model Size



Pruning Techniques
[Han, NeurIPS15]



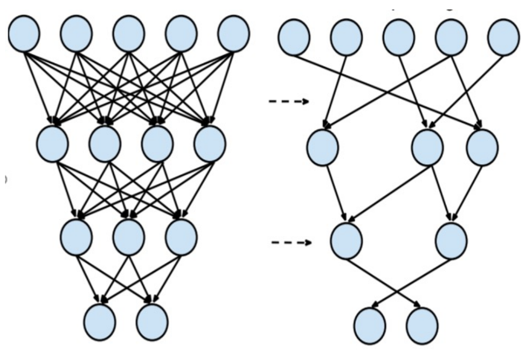
Depth-wise Separable Layers
[Howard, CVPR17]

Introduces Sparse Weights

Introduces Dense Weights

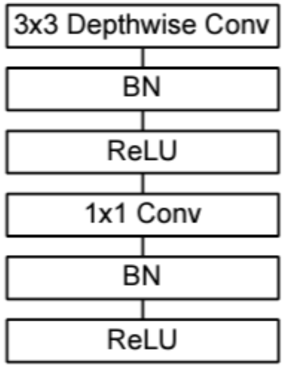
Different DNN Optimizations Introduce Different Sparsity

Optimizations to Reduce Model Size



Pruning Techniques
[Han, NeurIPS15]

Introduces Sparse Weights

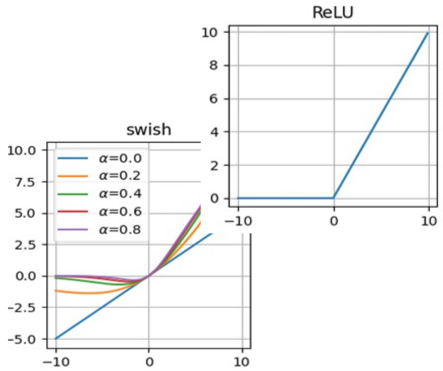


3x3 Depthwise Conv
BN
ReLU
1x1 Conv
BN
ReLU

Depth-wise Separable Layers
[Howard, CVPR17]

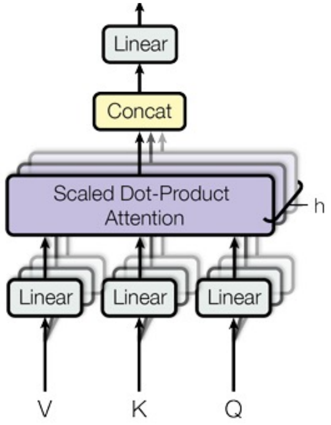
Introduces Dense Weights

Optimizations to Improve Accuracy



Activation Functions
[Apicella, NN21]

Introduces Dense/Sparse Activations

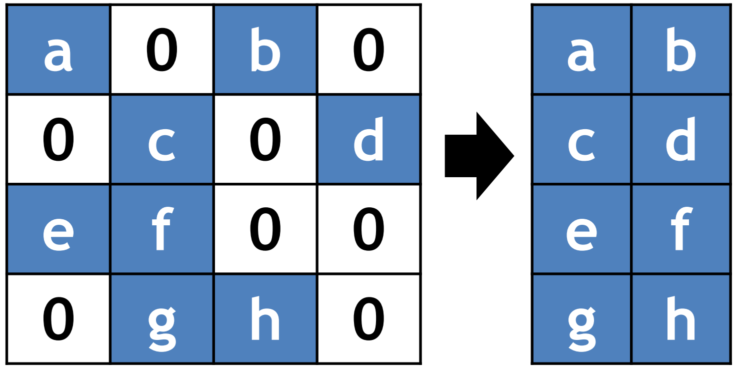


Attention-based Modules
[Vaswani, NeurIPS17]

Introduces Dense Act./Weights

Modern DNNs can weights and activations that are either dense or sparse with various sparsity degrees

High-Level Opportunities for Sparse DNNs



a	0	b	0
0	c	0	d
e	f	0	0
0	g	h	0

a	b
c	d
e	f
g	h

Zero Values
Can be Compressed Away

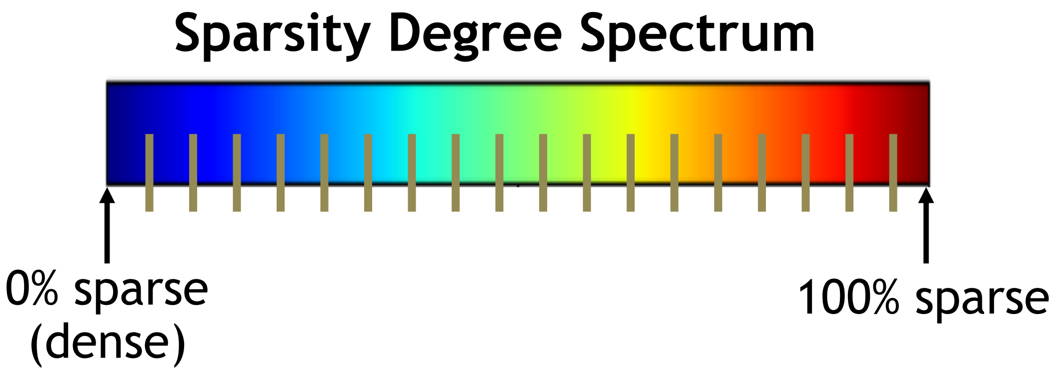
$$x \times 0 = 0$$

$$x + 0 = x$$

Ineffectual Operations
Can be Eliminated

Important to design sparse DNN accelerators to exploit such opportunities

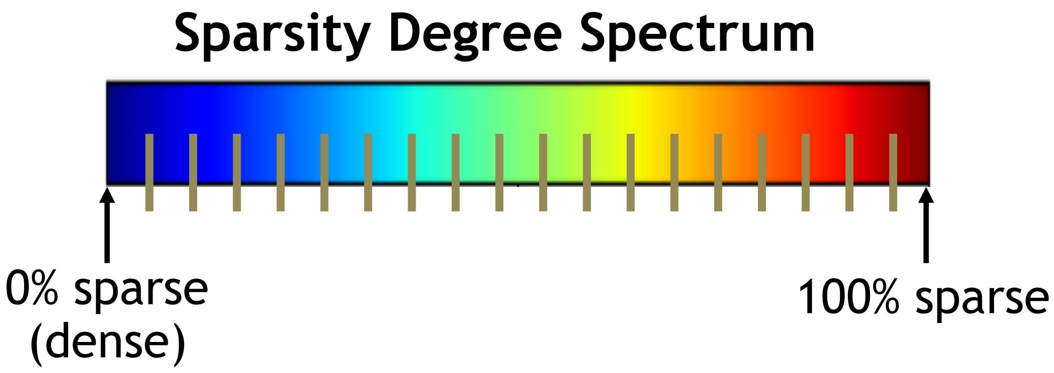
Requirements for an Ideal Sparse DNN Accelerator



Flexible

exploit many sparsity degrees

Requirements for an Ideal Sparse DNN Accelerator



Flexible

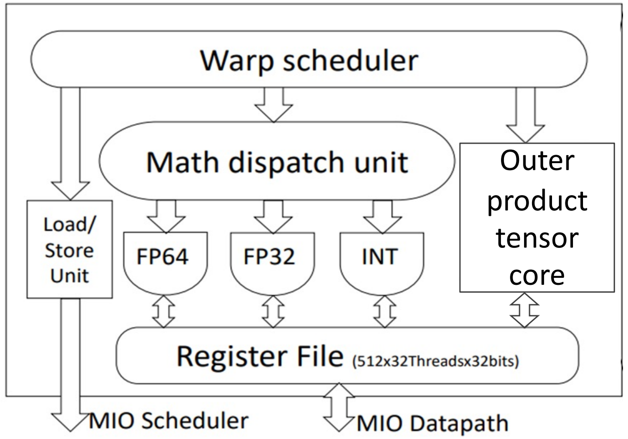
exploit many sparsity degrees

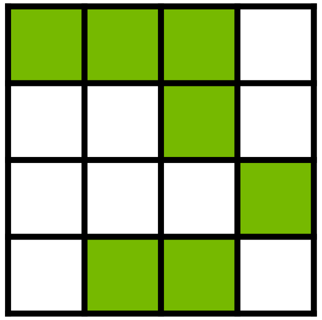
Efficient

*low **sparsity tax** for eliminating ineffectual operations*

Existing Works Do Not Meet Such Requirements

Unstructured Sparse Accelerators





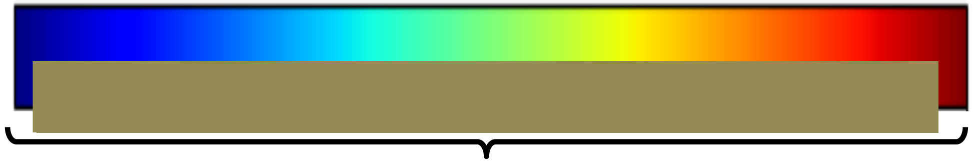
🟩	🟩	🟩	
		🟩	
			🟩
	🟩	🟩	

Unstructured sparse

Dual-Side Sparse Tensor Core (DSTC)

[Wang, ISCA21]

Sparsity Degree Spectrum

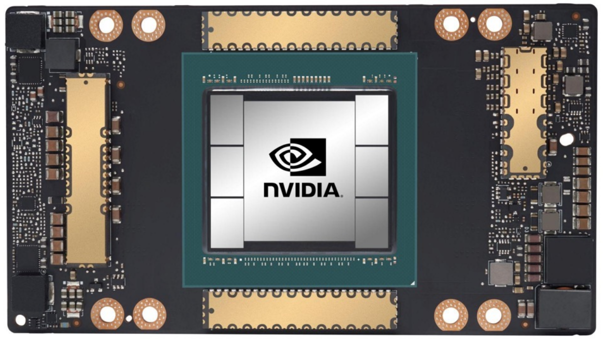


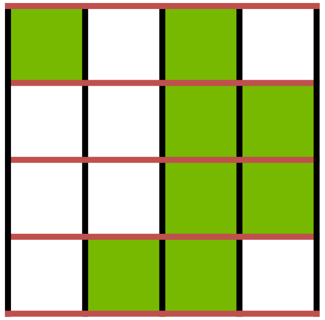
Continuously Translated into Savings

− High Sparsity Tax

+ Flexible

Structured Sparse Accelerators





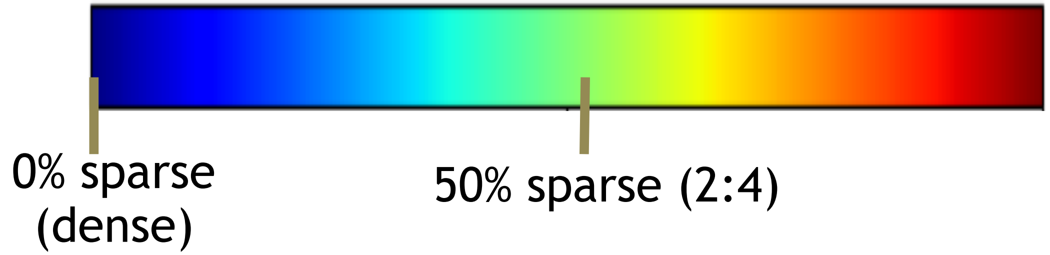
🟩		🟩	
		🟩	🟩
		🟩	🟩
	🟩	🟩	

Per-row 2:4 structured sparse (G:H pattern)

NVIDIA Sparse Tensor Core (STC)

[NVIDIA, TechReport20]

Sparsity Degree Spectrum

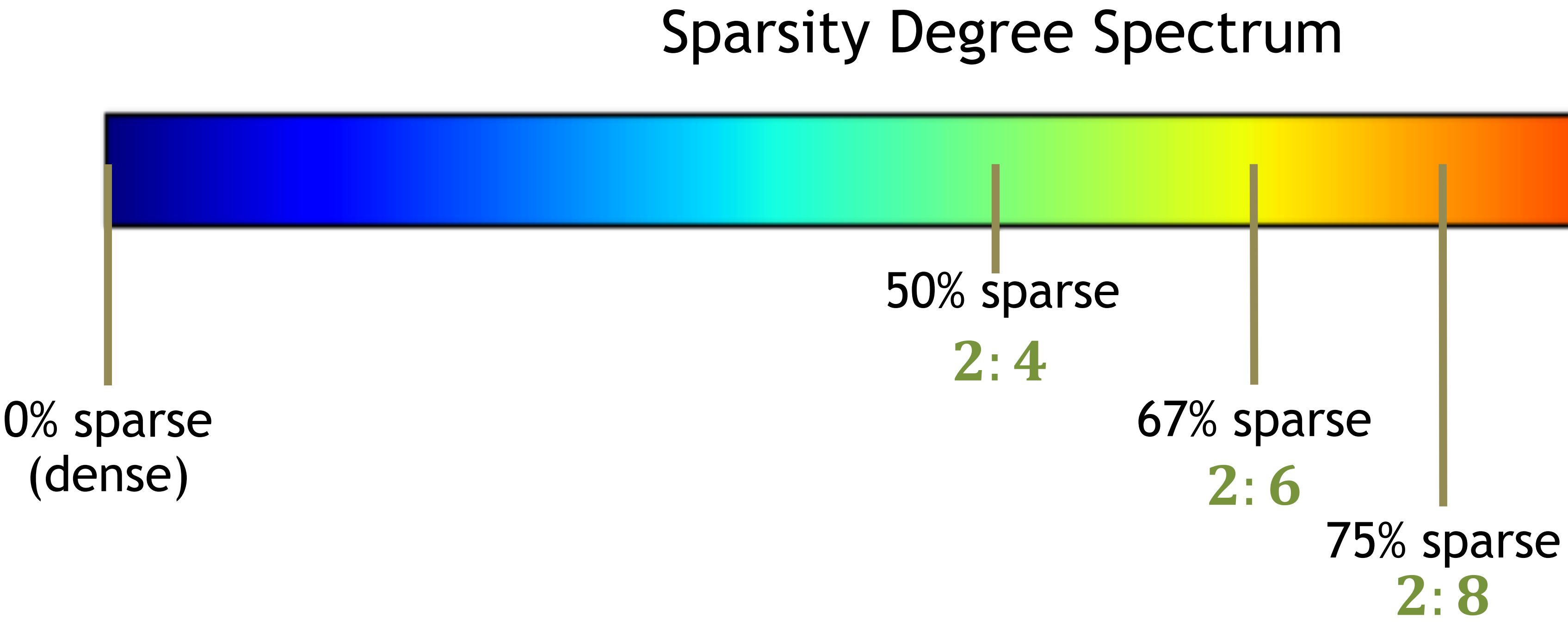


+ Low Sparsity Tax

− Inflexible

Naïve Way to Increase Flexibility Structured Sparse Designs

Extend the Number of G:H Ratios Supported



Not Scalable
Sparsity tax increases approximately in proportion to the number of sparsity degrees

Our Proposal

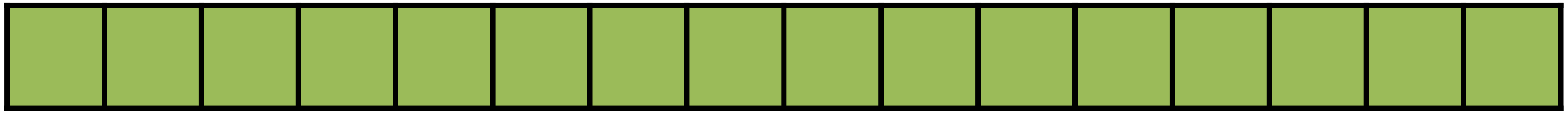
Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity

Hierarchical Structured Sparsity (HSS)

Compose G:H sparsity patterns in a hierarchical fashion

N-Rank HSS: $\underset{\textit{Rank N-1}}{G{:}H} \rightarrow \underset{\textit{Rank N-2}}{G{:}H} \dots \rightarrow \underset{\textit{Rank 0}}{G{:}H}$

What does a 3:4→2:4 pattern look like?



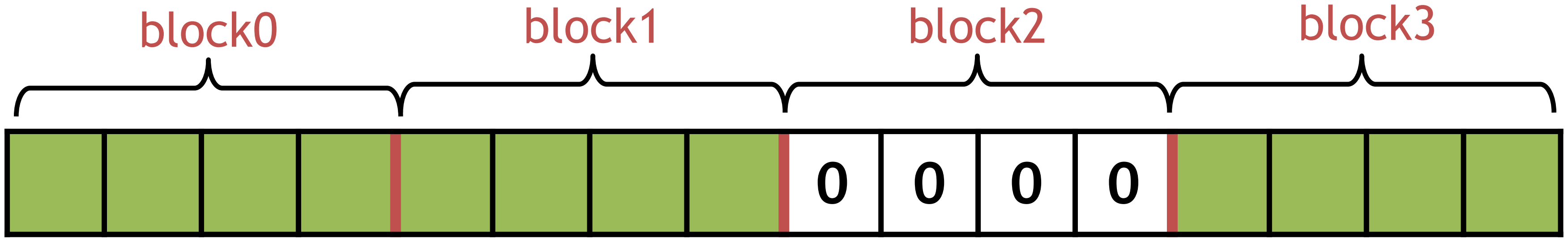
Dense Vector

Hierarchical Structured Sparsity (HSS)

Compose G:H sparsity patterns in a hierarchical fashion

What does a 3:4→2:4 pattern look like?

Rank1: 3 nonempty blocks out of the 4 blocks



Vector with Rank1 Sparsity Applied

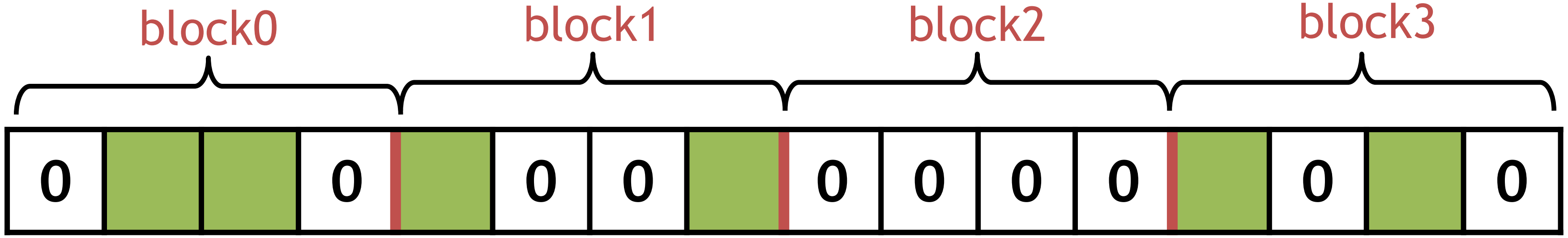
Hierarchical Structured Sparsity (HSS)

Compose G:H sparsity patterns in a hierarchical fashion

What does a 3:4→2:4 pattern look like?

Rank1: 3 nonempty blocks out of the 4 blocks

Rank0: 2 nonzero values out of 4 values within the block

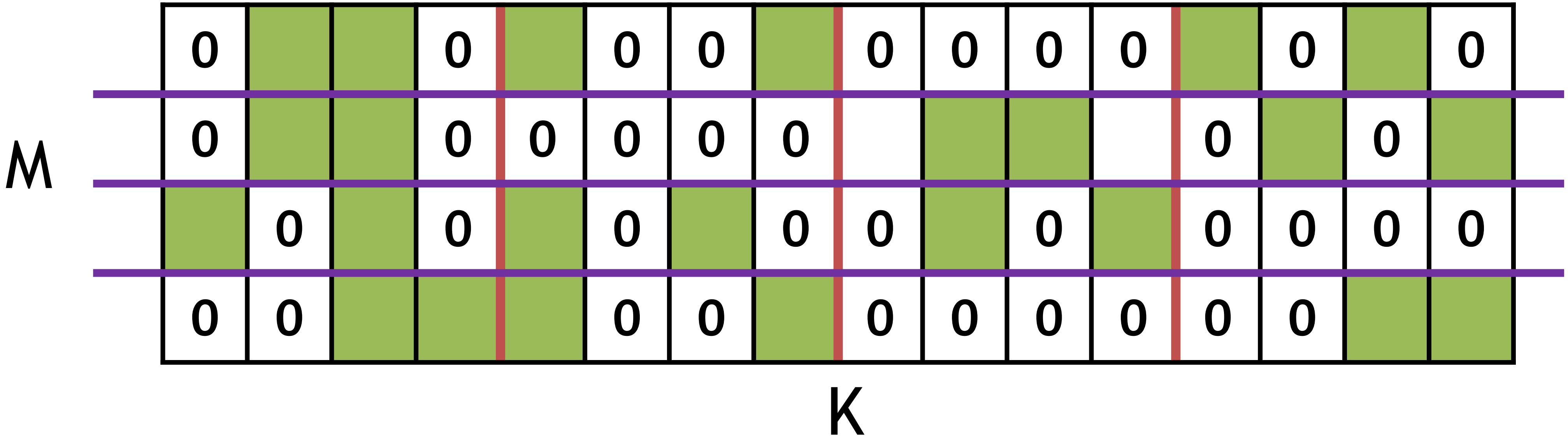


0	🟩	🟩	0	🟩	0	0	🟩	0	0	0	0	🟩	0	🟩	0

Vector with Both Ranks' Sparsity Applied

Hierarchical Structured Sparsity (HSS)

DNN Workloads Often Have Tensors with Multiple Dimensions



0	🟩	🟩	0	🟩	0	0	🟩	0	0	0	0	🟩	0	🟩	0
0	🟩	🟩	0	0	0	0	0		🟩	🟩		0	🟩	0	🟩
🟩	0	🟩	0	🟩	0	🟩	0	0	🟩	0	🟩	0	0	0	0
0	0	🟩	🟩	🟩	0	0	🟩	0	0	0	0	0	0	🟩	🟩

Per-Row 3:4-2:4 Tensor

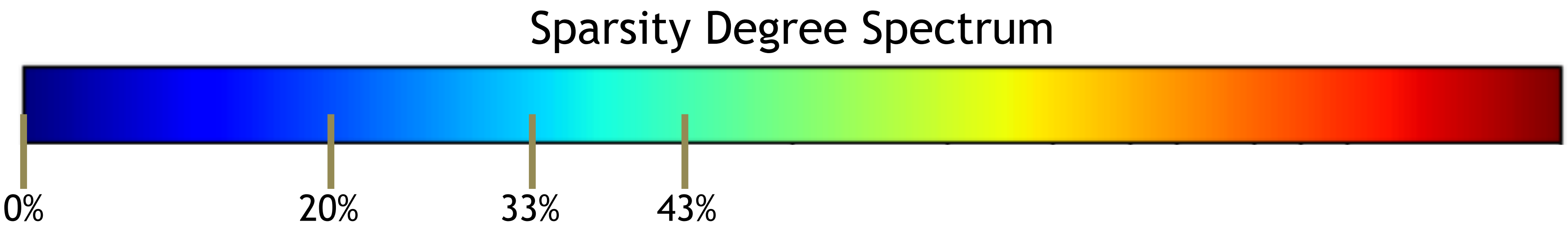
HSS can be applied to an arbitrary dimension in a multi-dimensional tensor

HSS Introduces A Flexible Way to Express Sparsity Degrees

4 sparsity degrees

Rank 1

4:4 (0%)	4:5 (20%)	4:6 (33%)	4:7 (43%)



HSS Introduces A Flexible Way to Express Sparsity Degrees

4 sparsity degrees

Rank 1

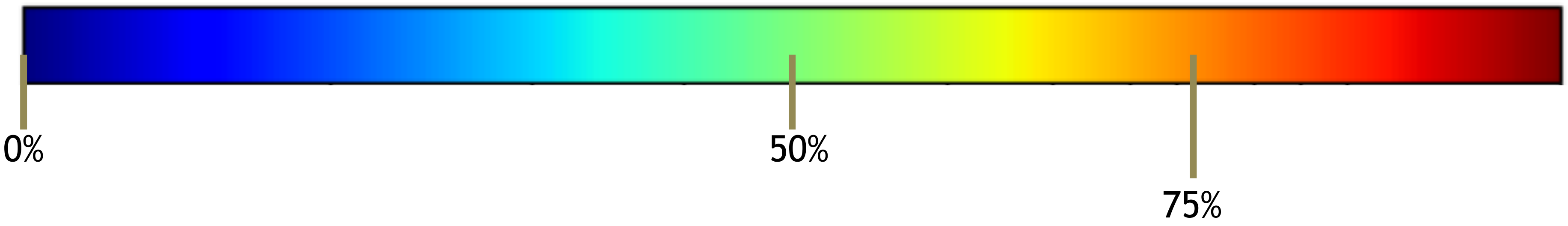
4:4	4:5	4:6	4:7
(0%)	(20%)	(33%)	(43%)

Rank 0

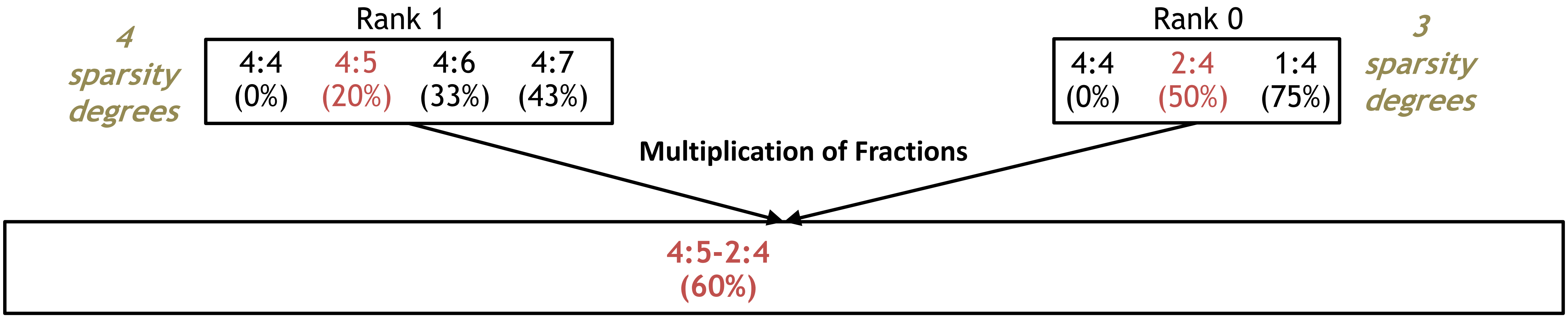
4:4	2:4	1:4
(0%)	(50%)	(75%)

3 sparsity degrees

Sparsity Degree Spectrum



HSS Introduces A Flexible Way to Express Sparsity Degrees



4 sparsity degrees

Rank 1

4:4 (0%) | 4:5 (20%) | 4:6 (33%) | 4:7 (43%)

Rank 0

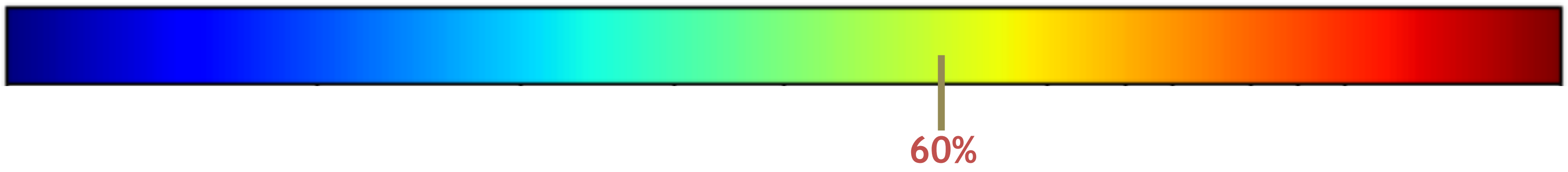
4:4 (0%) | 2:4 (50%) | 1:4 (75%)

3 sparsity degrees

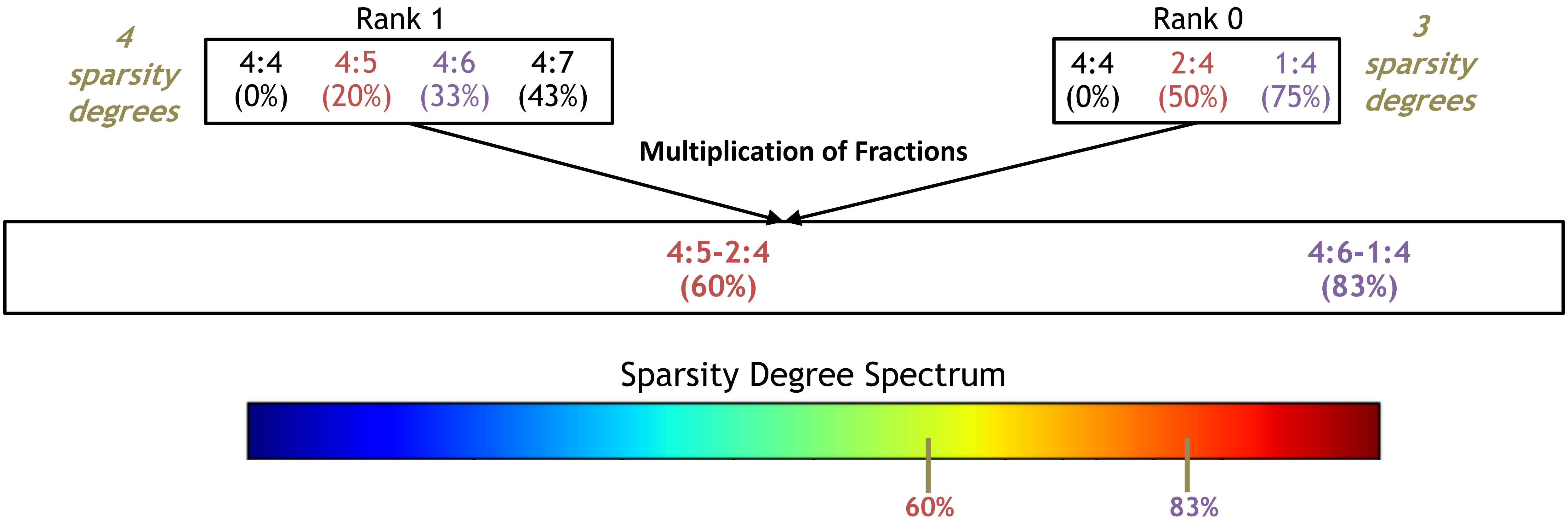
Multiplication of Fractions

4:5-2:4 (60%)

Sparsity Degree Spectrum



HSS Introduces A Flexible Way to Express Sparsity Degrees



4 sparsity degrees

Rank 1

4:4 (0%) 4:5 (20%) 4:6 (33%) 4:7 (43%)

Rank 0

4:4 (0%) 2:4 (50%) 1:4 (75%)

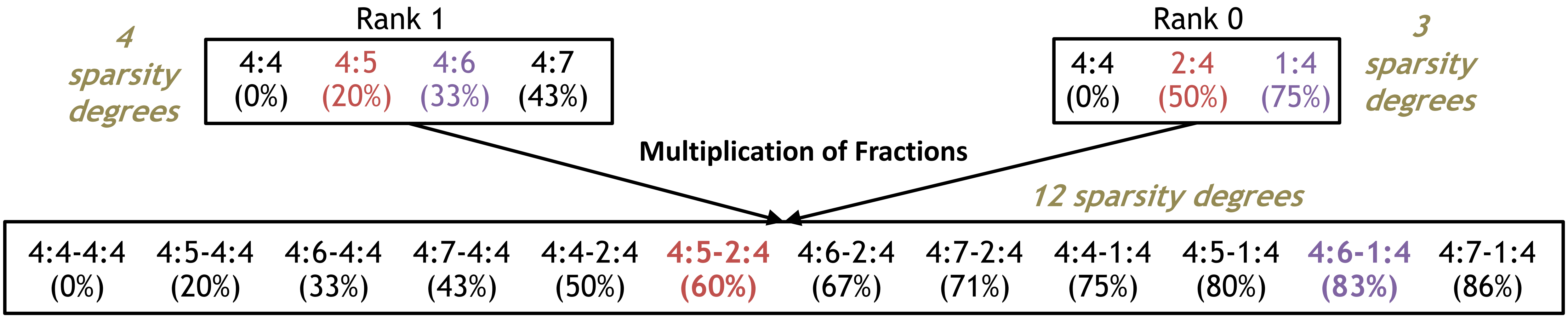
3 sparsity degrees

Multiplication of Fractions

4:5-2:4 (60%) **4:6-1:4 (83%)**

Sparsity Degree Spectrum

HSS Introduces A Flexible Way to Express Sparsity Degrees



4 sparsity degrees

Rank 1

4:4	4:5	4:6	4:7
(0%)	(20%)	(33%)	(43%)

Rank 0

4:4	2:4	1:4
(0%)	(50%)	(75%)

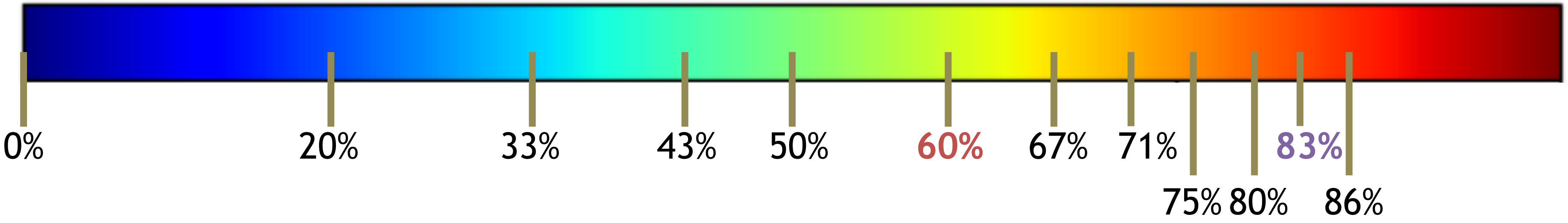
3 sparsity degrees

Multiplication of Fractions

12 sparsity degrees

4:4-4:4	4:5-4:4	4:6-4:4	4:7-4:4	4:4-2:4	4:5-2:4	4:6-2:4	4:7-2:4	4:4-1:4	4:5-1:4	4:6-1:4	4:7-1:4
(0%)	(20%)	(33%)	(43%)	(50%)	(60%)	(67%)	(71%)	(75%)	(80%)	(83%)	(86%)

Sparsity Degree Spectrum

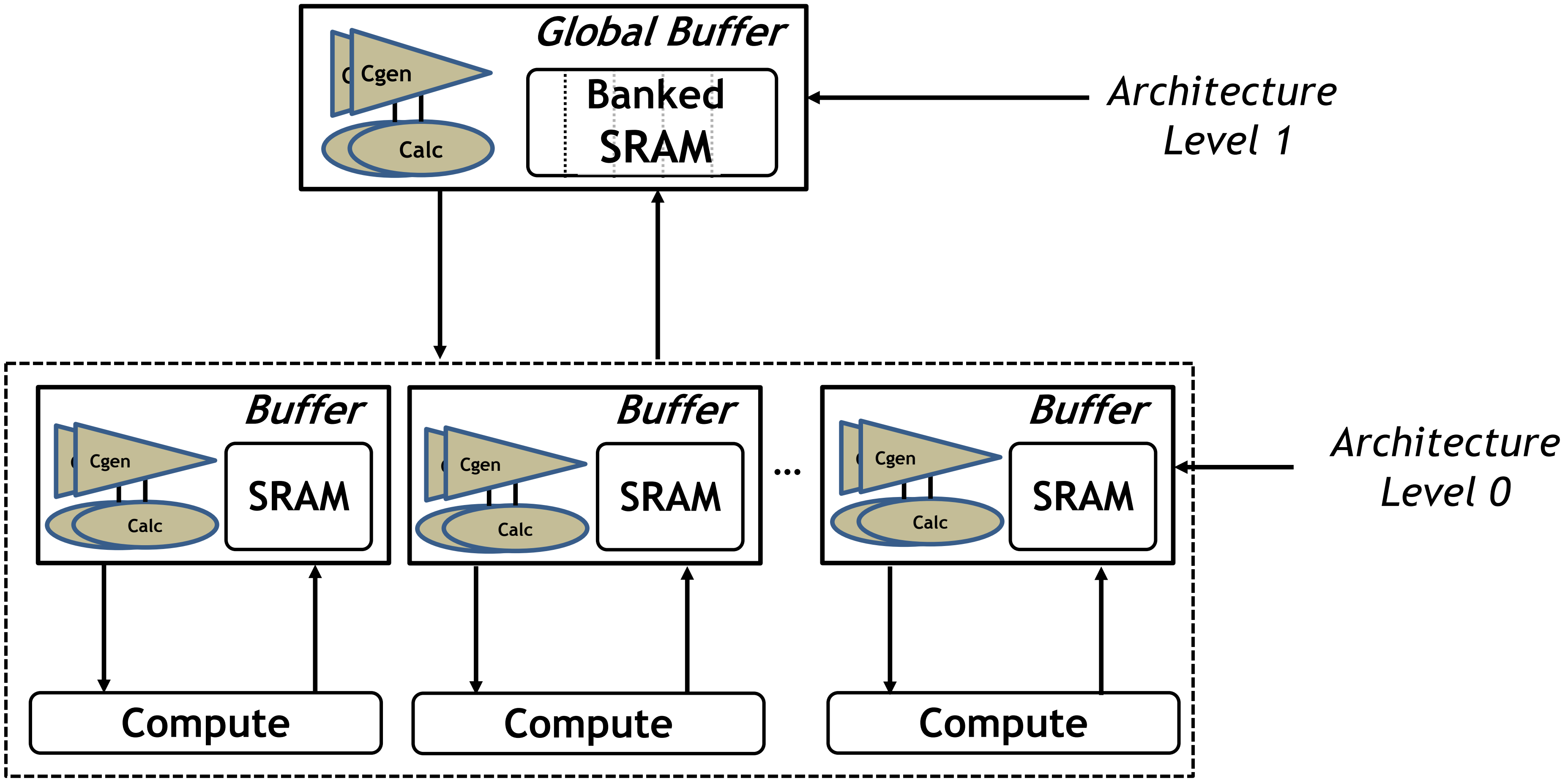


Fraction multiplication allows flexible representation of many sparsity degrees in a wide range

HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

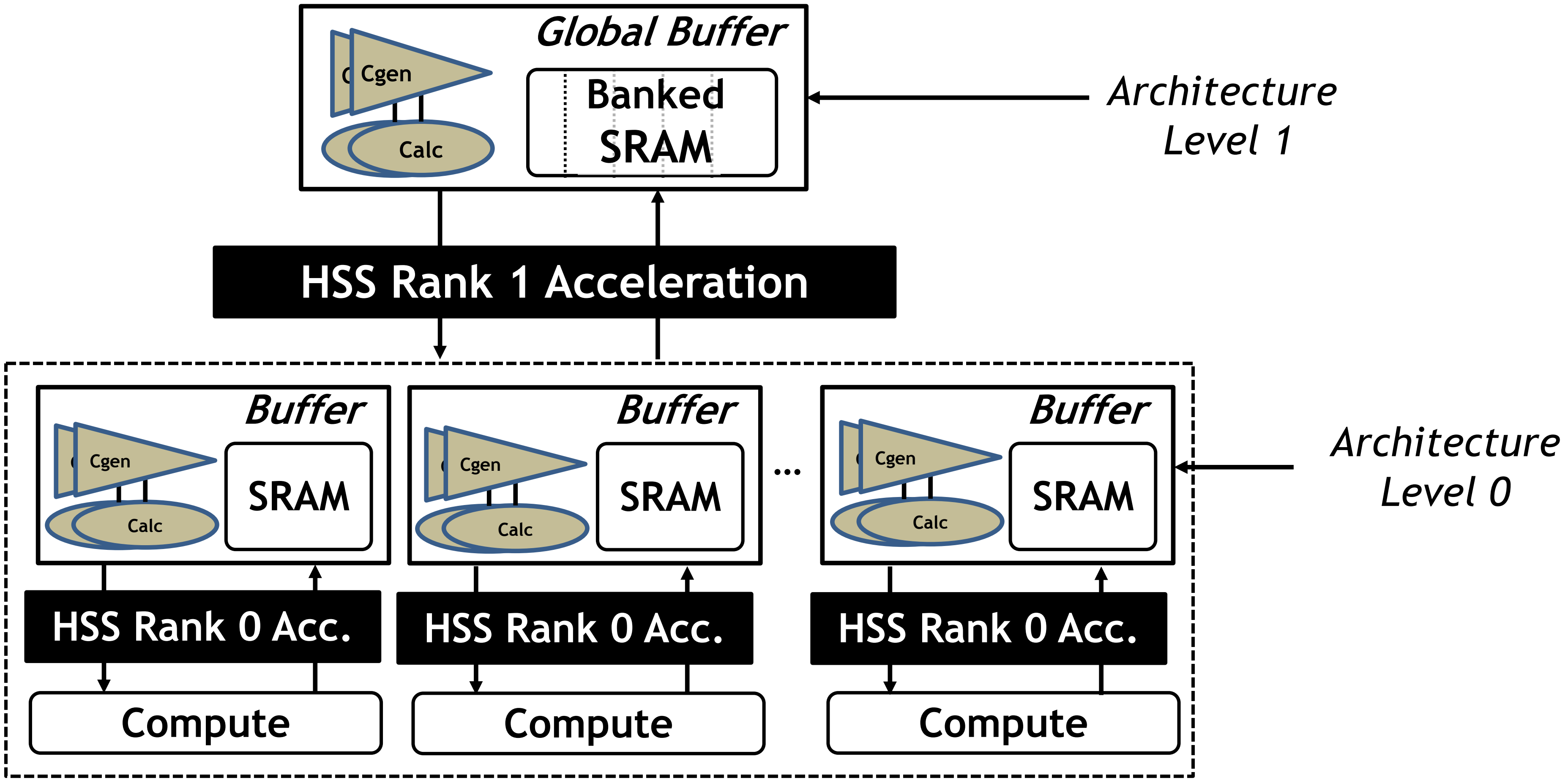
Example Accelerator Architecture Organization



HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

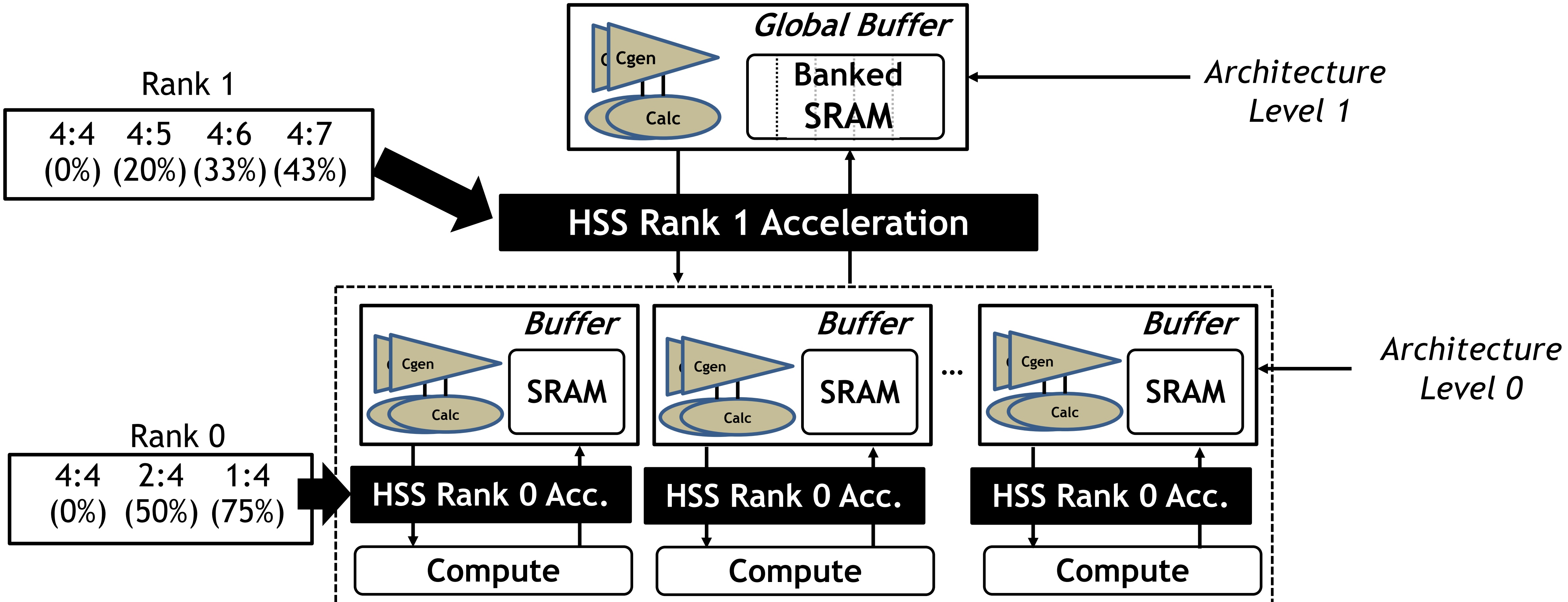
Example Accelerator Architecture Organization



HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

Example Accelerator Architecture Organization



Rank 1

4:4	4:5	4:6	4:7
(0%)	(20%)	(33%)	(43%)

Global Buffer

Cgen

Calc

Banked SRAM

Architecture Level 1

HSS Rank 1 Acceleration

Rank 0

4:4	2:4	1:4
(0%)	(50%)	(75%)

Buffer

Cgen

Calc

SRAM

HSS Rank 0 Acc.

Compute

...

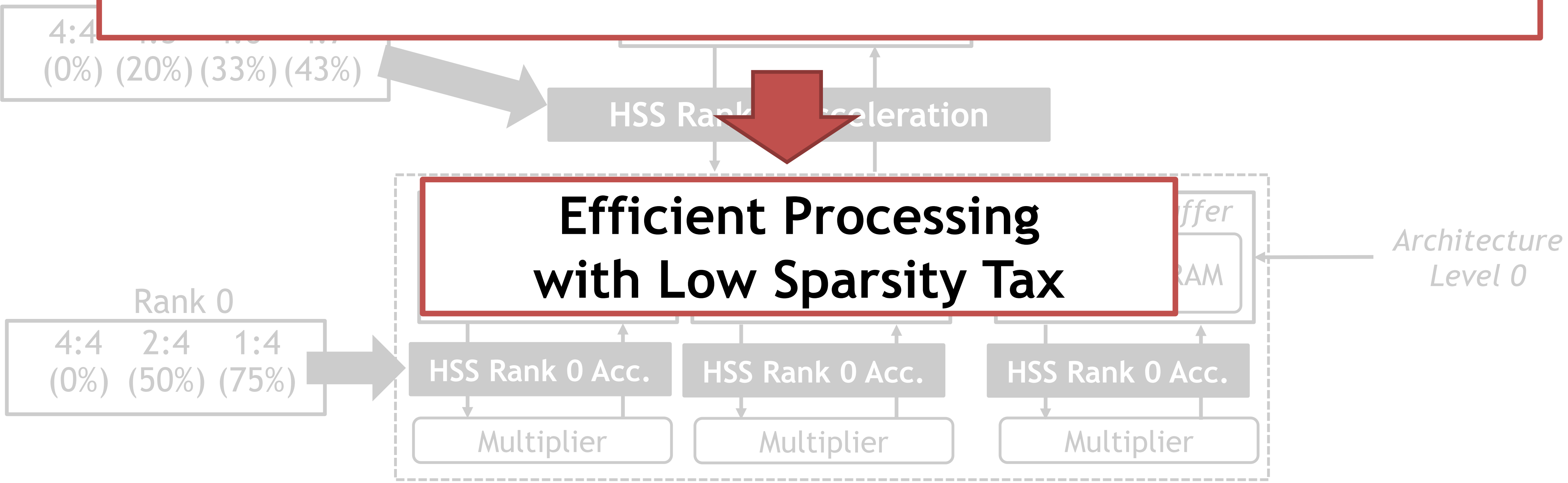
Architecture Level 0

Each level only needs to accelerate for a few sparsity degrees

HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

Simple Acceleration at Each Architecture Level Leads to Low Hardware Overhead



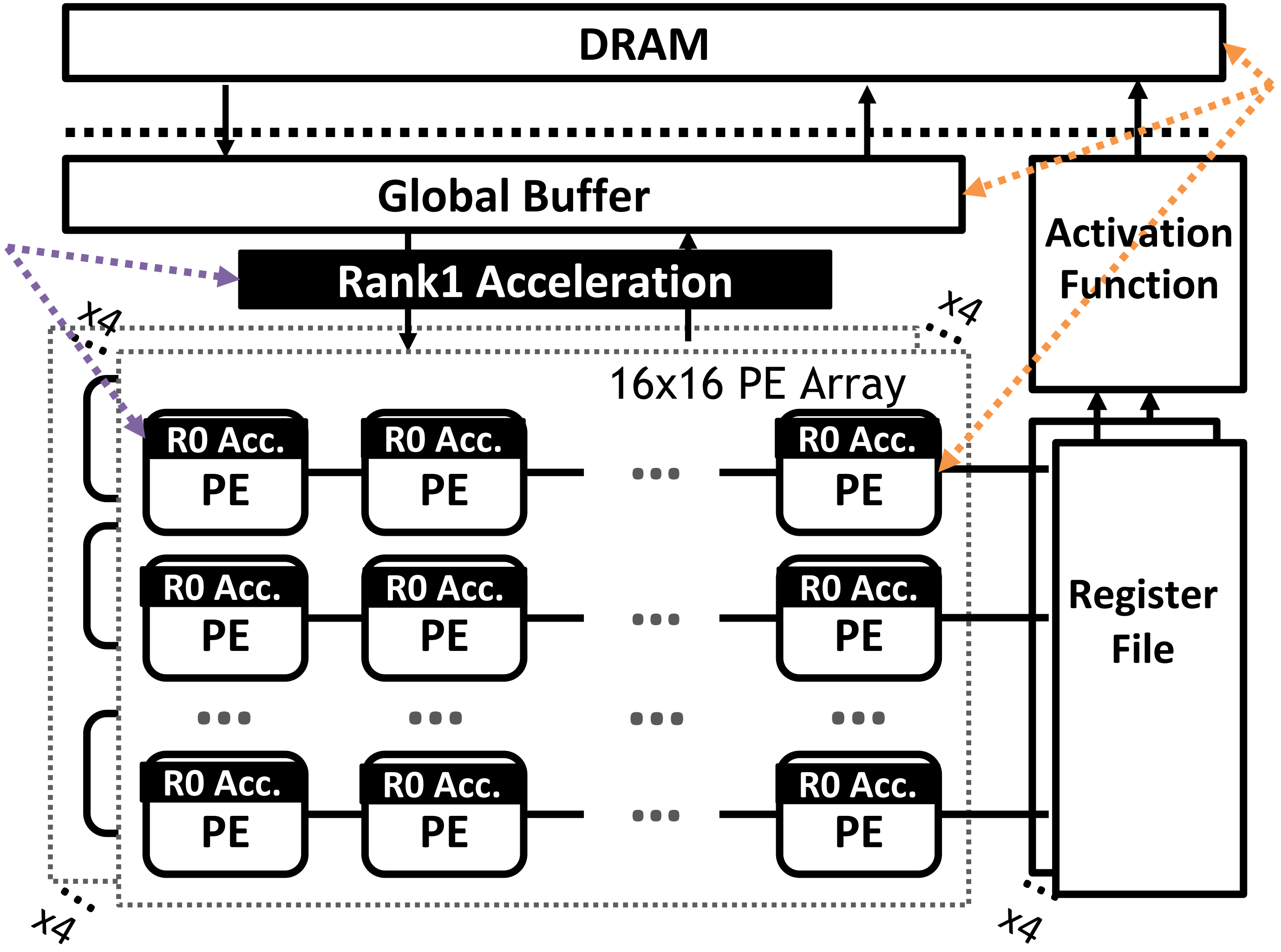
Efficient Processing with Low Sparsity Tax

Each level only needs to accelerate for a few sparsity degrees

HighLight: Flexible and Efficient Sparse DNN Accelerator

Hierarchical Skipping

Reduce Latency and Energy Consumption



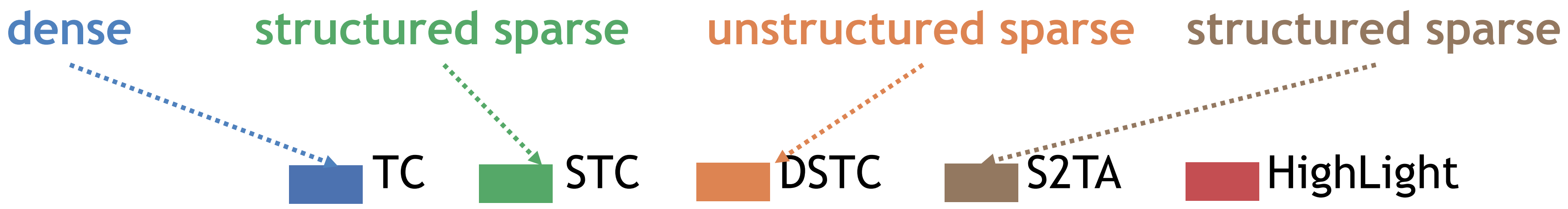
HSS-based Compressed Representation Format

Reduce Storage Requirement and Energy Consumption

HighLight translates two-rank HSS into reduction in latency and energy

Experimental Results

We Compare HighLight with Representative Designs



dense

structured sparse

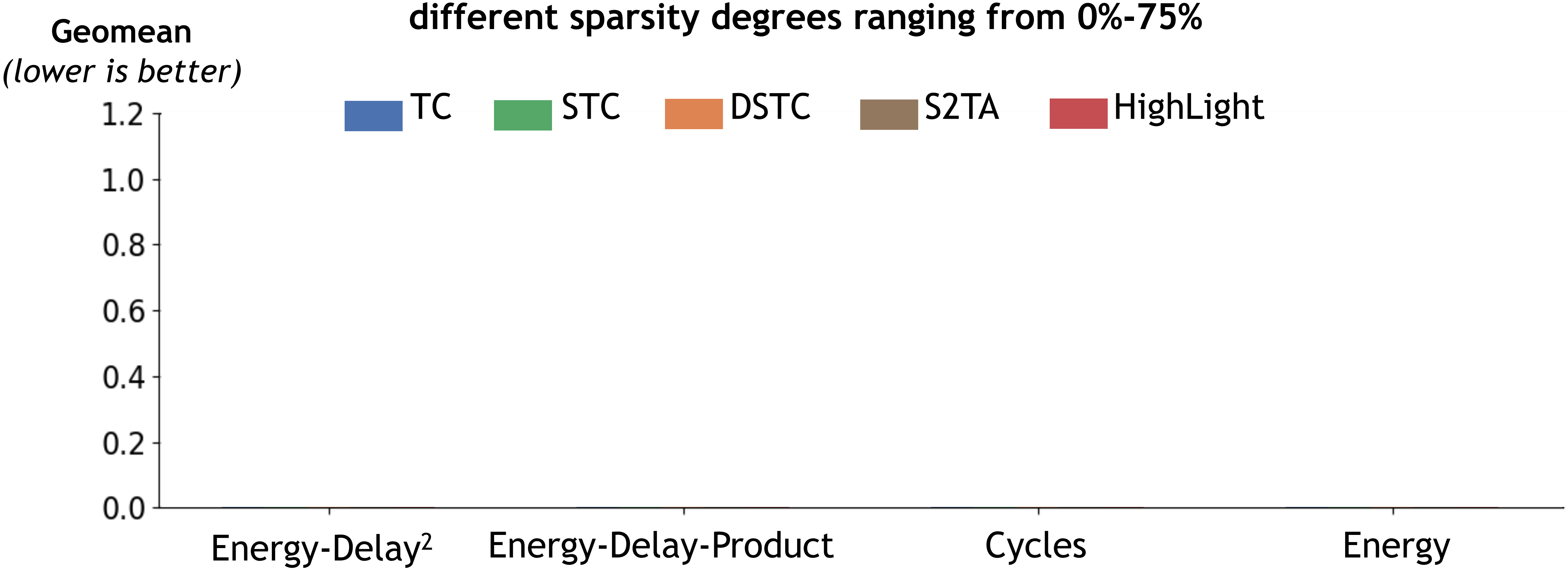
unstructured sparse

structured sparse

🟦 TC 🟩 STC 🟧 DSTC 🟫 S2TA 🟥 HighLight

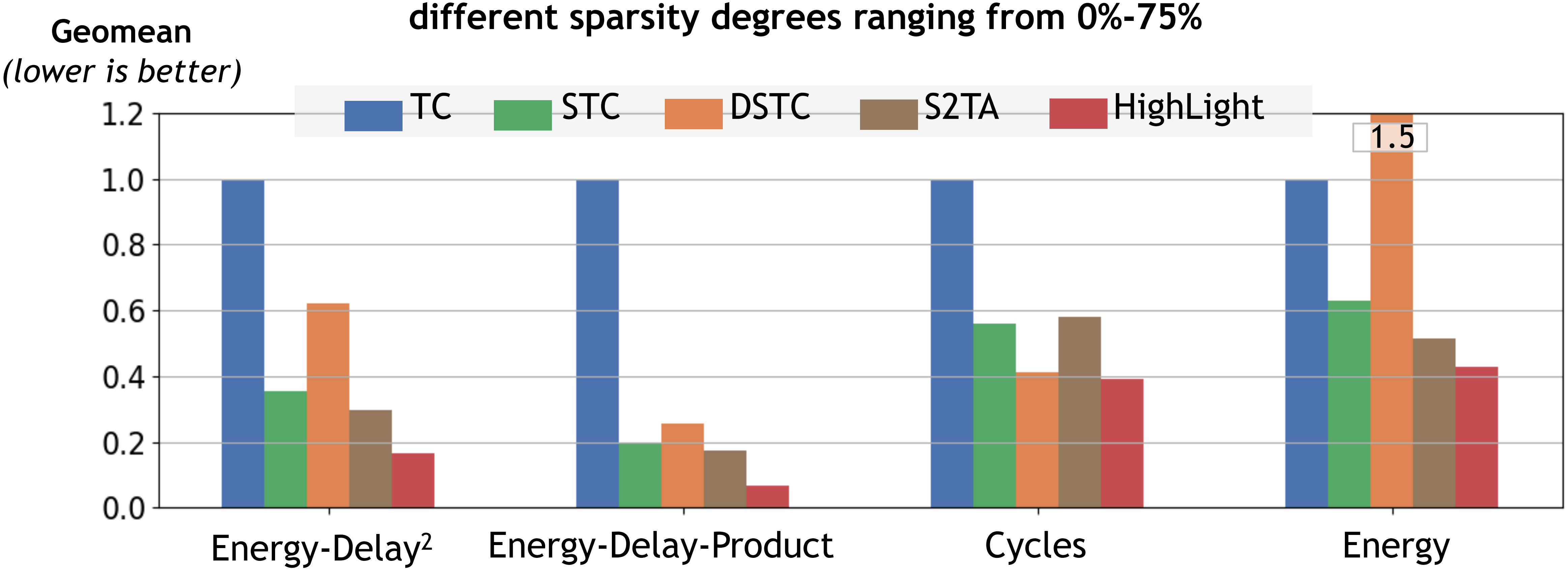
Geomean Across Various Hardware Performance Metrics

We evaluate the designs with synthetic workloads with different sparsity degrees ranging from 0%-75%



Geomean Across Various Hardware Performance Metrics

We evaluate the designs with synthetic workloads with different sparsity degrees ranging from 0%-75%



HighLight is efficient across evaluated metrics

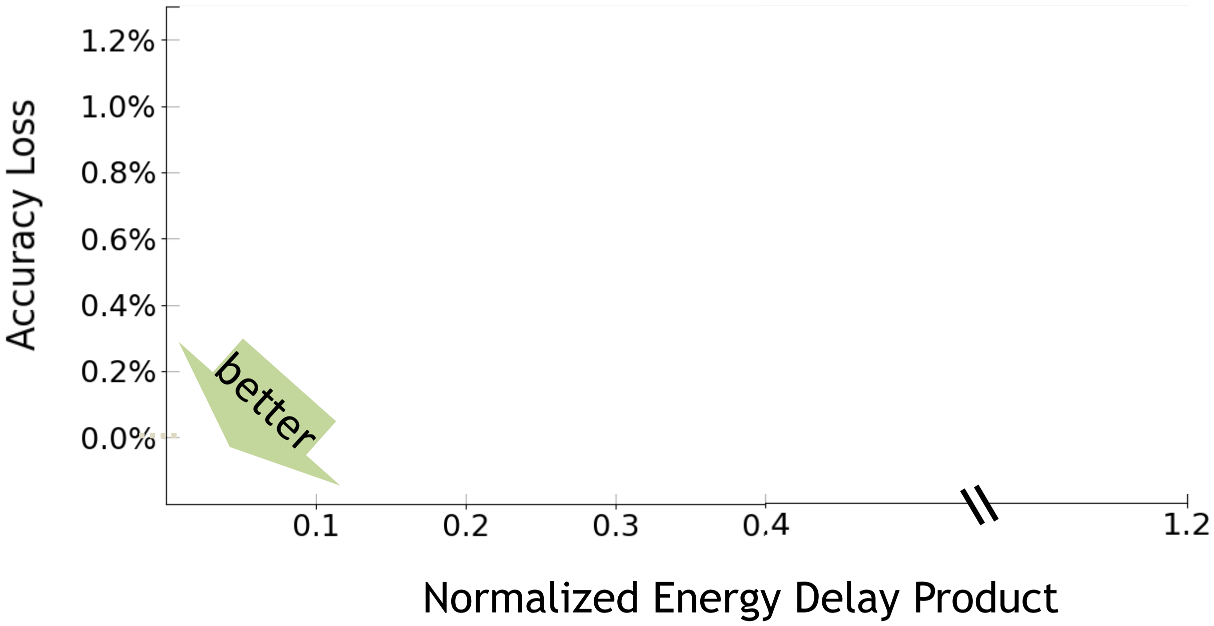
Accuracy-Energy Delay Product Pareto Frontier

We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any)

Accuracy-Energy Delay Product Pareto Frontier

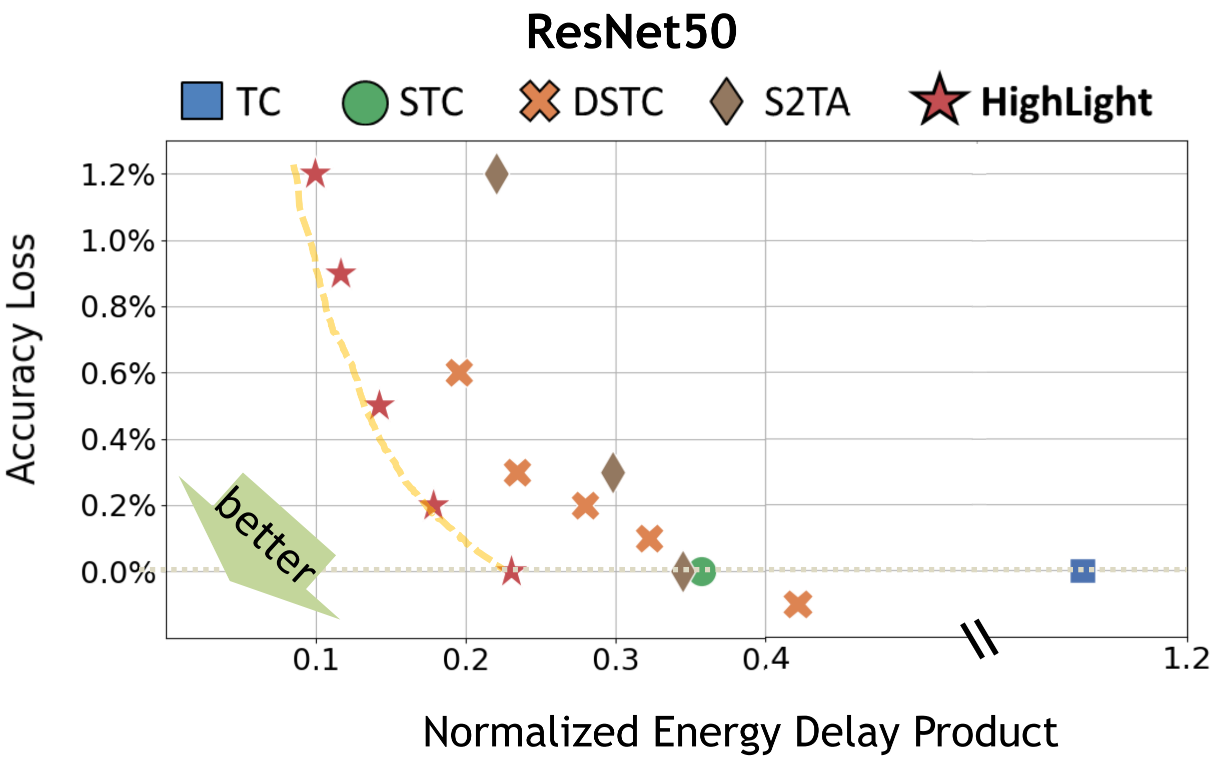
We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any)

ResNet50



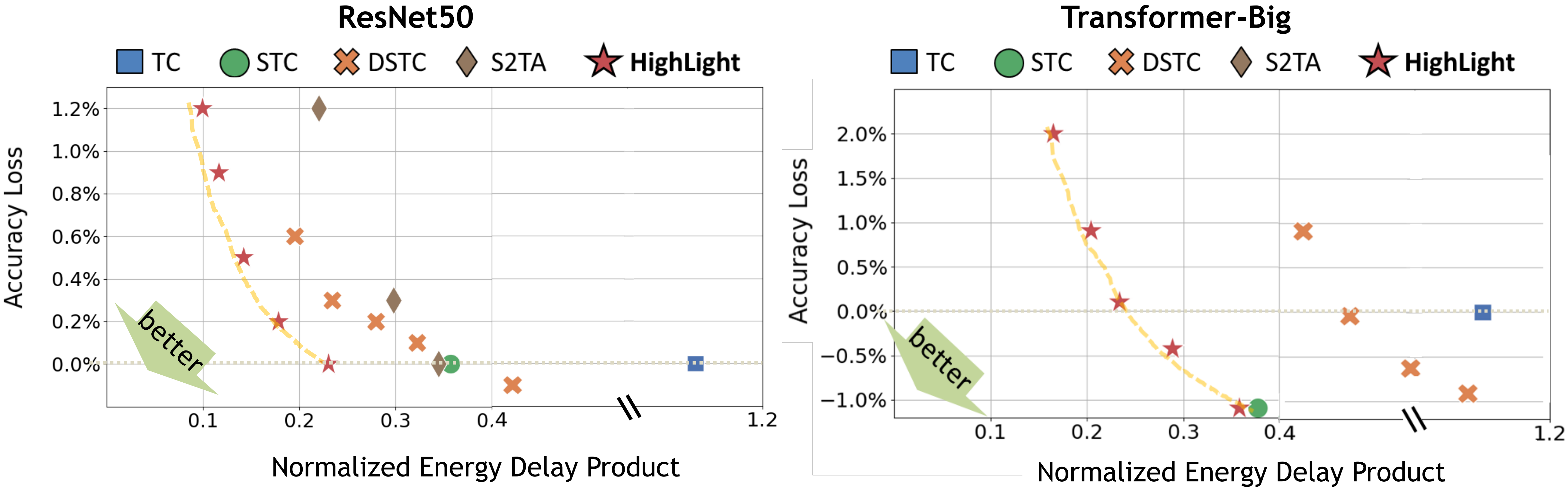
Accuracy-Energy Delay Product Pareto Frontier

We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any)



Accuracy-Energy Delay Product Pareto Frontier

We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any)



HighLight sits on the accuracy-energy delay product pareto frontier

More Details in Paper!

- How to systematically represent the diverse sparsity patterns in DNNs?
  - Short answer: sparsity specification via fibertree abstraction.

- What does HighLight's energy and area sparsity tax breakdowns look like?
  - Short answer: low sparsity tax as HighLight independently accelerates simple sparsity patterns at different architecture levels.

- …

Summary

http://emze.csail.mit.edu/highlight

SCAN ME

Hierarchical Structured Sparsity (HSS)

- Composed of multiple levels of simple sparsity patterns
- Allows flexible expression of diverse sparsity degrees

HighLight Accelerator

- Supports two-rank HSS for a few degrees at each level
- Implements low-overhead support for each rank at different architecture levels
- Ensures both efficiency and flexibility

Acknowledgement: NVIDIA Research, MIT AI Hardware Program