ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

Discovering the In-Memory Kernels of 3D Dot-Product Engines

Muhammad Rashedul Haq Rashed[1], Sumit Kumar Jha[2], Rickard Ewetz[1]

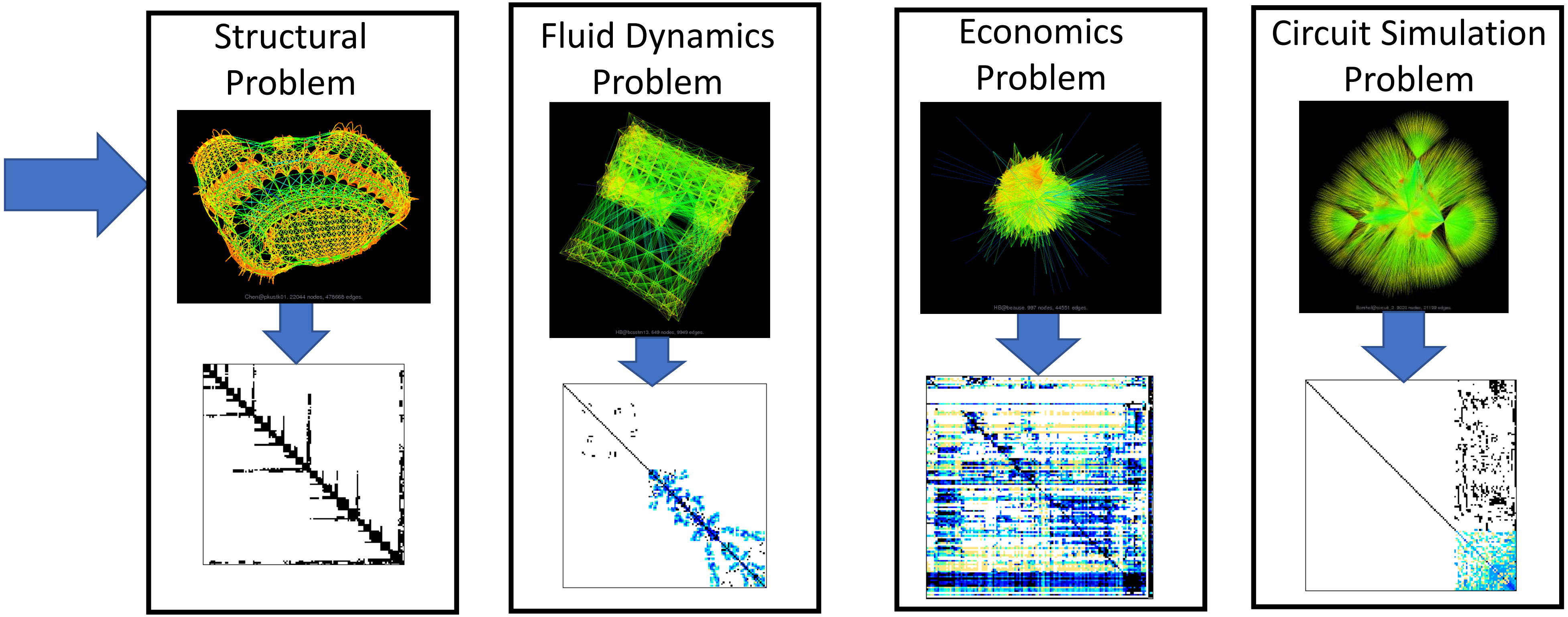
[1]University of Central Florida, [2]University of Texas at San Antonio

Outline

- Preliminaries
- 3D DPE Level
    - Sharing of hardware
    - Hardware mapping
- Application level
        - Methodology
- Evaluation
- Summary

Why In-Memory Computing?

Data-Intensive Systems

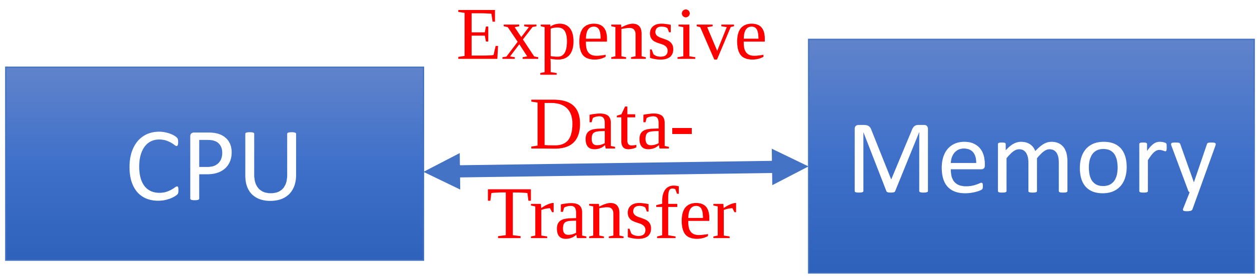


Structural Problem

Fluid Dynamics Problem

Economics Problem

Circuit Simulation Problem



CPU

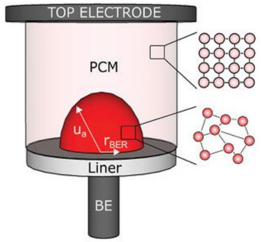
Expensive Data-Transfer

Memory

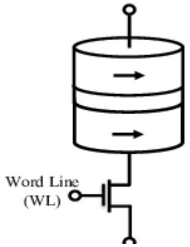
1. Davis, Timothy A., and Yifan Hu. "The University of Florida sparse matrix collection." *ACM Transactions on Mathematical Software (TOMS)* 38.1 (2011): 1-25.

NVM Technology and In-Memory Computing

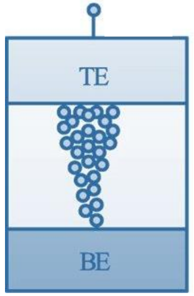
NVM Technologies



PCM[1]



STT-RAM[2]



ReRAM

Matrix Vector Multipication[3]

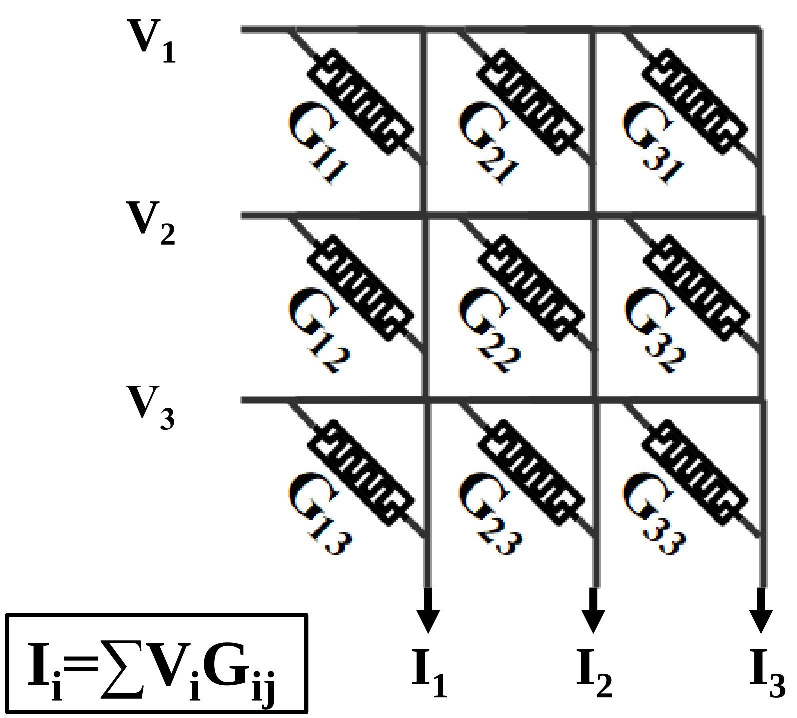
$$\begin{array}{cc|c|c|c|c|}
\cline{3-6}
V_1 & \rightarrow & G_{11} & G_{12} & \dots & G_{1n} \\
\cline{3-6}
V_2 & \rightarrow & G_{21} & G_{22} & \dots & G_{2n} \\
\cline{3-6}
\dots & \rightarrow & \dots & \dots & \dots & \dots \\
\cline{3-6}
V_n & \rightarrow & G_{n1} & G_{n2} & \dots & G_{nn} \\
\cline{3-6}
 & & \downarrow & \downarrow & \downarrow & \downarrow \\
 & & I_1 & I_2 & \dots & I_n
\end{array}$$

Pro

- In-situ computation
- Extremely energy efficient

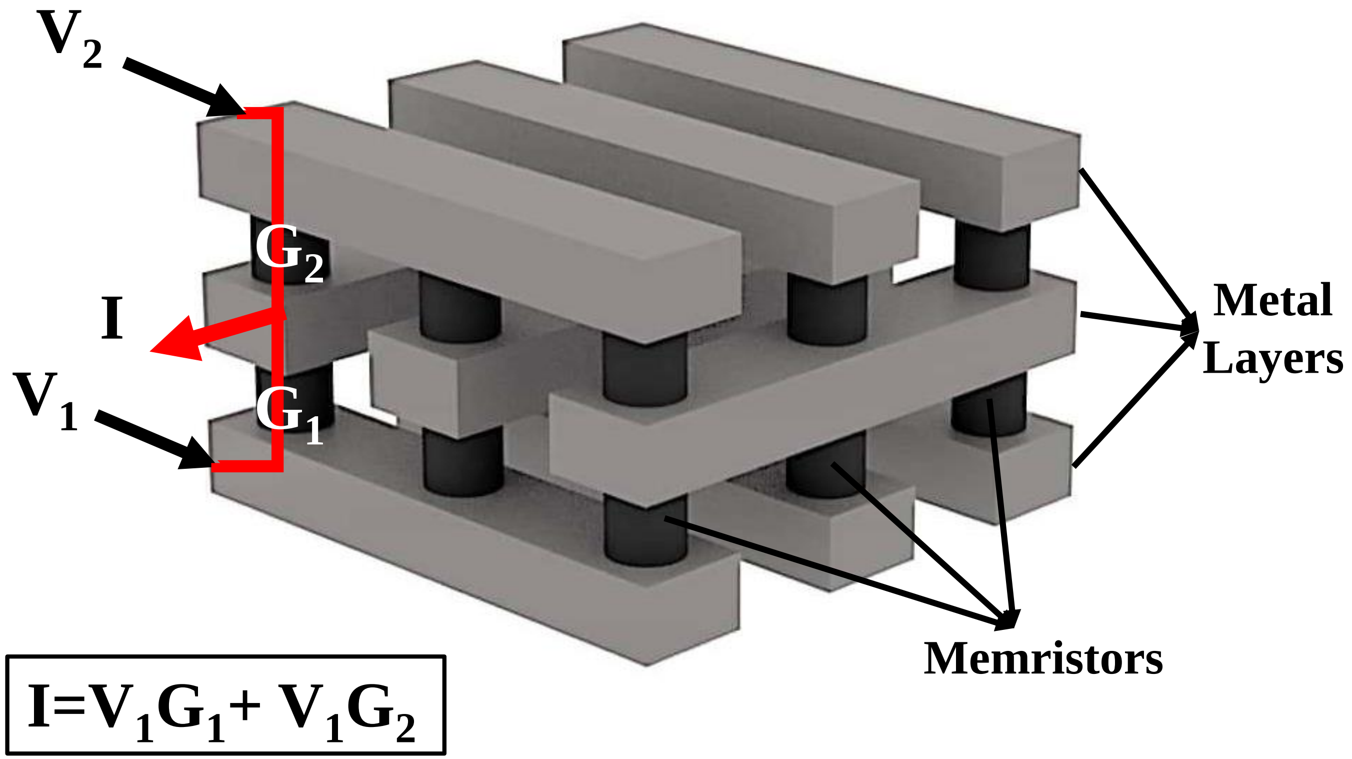
1. Ghazi Sarwat, Syed, et al. "Projected Mushroom Type Phase-Change Memory." *Advanced Functional Materials* 31.49 (2021): 2106547.
2. Seyedzadeh, Seyed Mohammad & Maddah, Rakan & Jones, Alex & Melhem, Rami. (2016). Leveraging ECC to Mitigate Read Disturbance, False Reads and Write Faults in STT-RAM.
3. 1. Wang, Feng, Guangyu Sun, and Guojie Luo. "SSR: A Skeleton-based Synthesis Flow for Hybrid Processing-in-RRAM Modes." *2021 ICCAD*. IEEE, 2021.

2D vs 3D DPE



$$I_i=\sum V_i G_{ij}$$

2D DPE



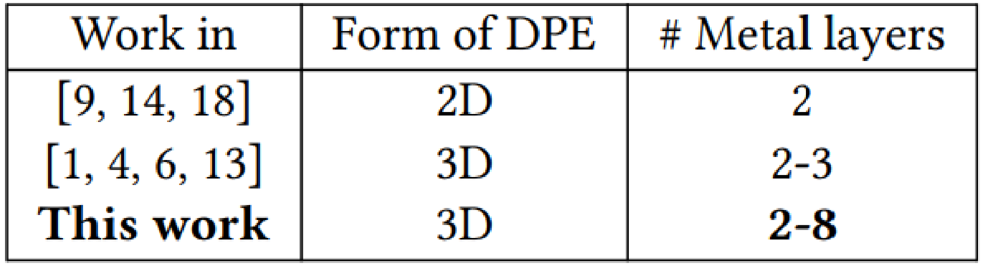
$$I=V_1G_1+ V_1G_2$$

3D DPE

Great Potential for hardware sharing

Previous Work

Table 1: Architectural comparison of dot-product engines



Work in	Form of DPE	# Metal layers
[9, 14, 18]	2D	2
[1, 4, 6, 13]	3D	2-3
This work	3D	**2-8**

- Previous works focused on mitigation of parasitics on the algorithmic level
- DAC and ADC operations are time multiplexed
- Efficient utilization of 3D ReRAM crossbars with more than three layers have not been explored

[1] G. C. Adam et al. 3d reram arrays and crossbars: Fabrication, characterization and applications. In *2017 Ieee-Nano*, pages 844–849. IEEE, 2017

[4] B. Chakrabarti et al. A multiplyadd engine with monolithically integrated 3d memristor crossbar/cmos hybridcircuit. *Scientific reports*, 7(1):1–10, 2017.

[6] F. Chen et al., A novel zero-free dataflow accelerator for generative adversarial networks in 3d reram. In *Proceedings of the 56th DAC 2019*, pages 1–6, 2019.

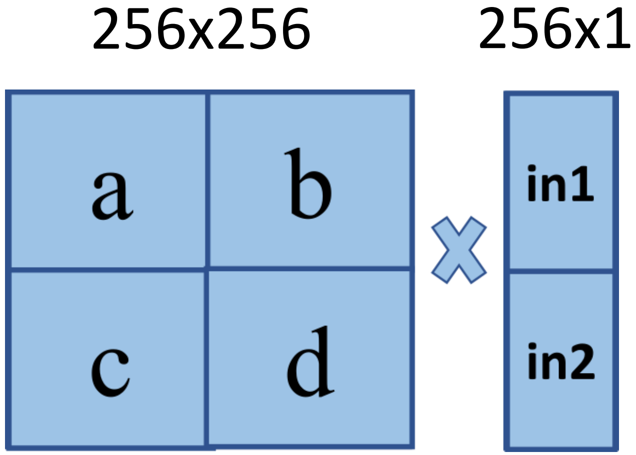
[9] B. Feinberg et. Al, Enabling scientific computing on memristive accelerators. In *2018 ACM/IEEE 45th ISCA*, pages 367–382. IEEE, 2018.

[13] M. A. Lastras-Montano et. Al, 3d-dpe: A 3d high-bandwidth dot-product engine for high-performance neuromorphic computing. In *2017 DATE*, pages 1257–1260. IEEE, 2017.

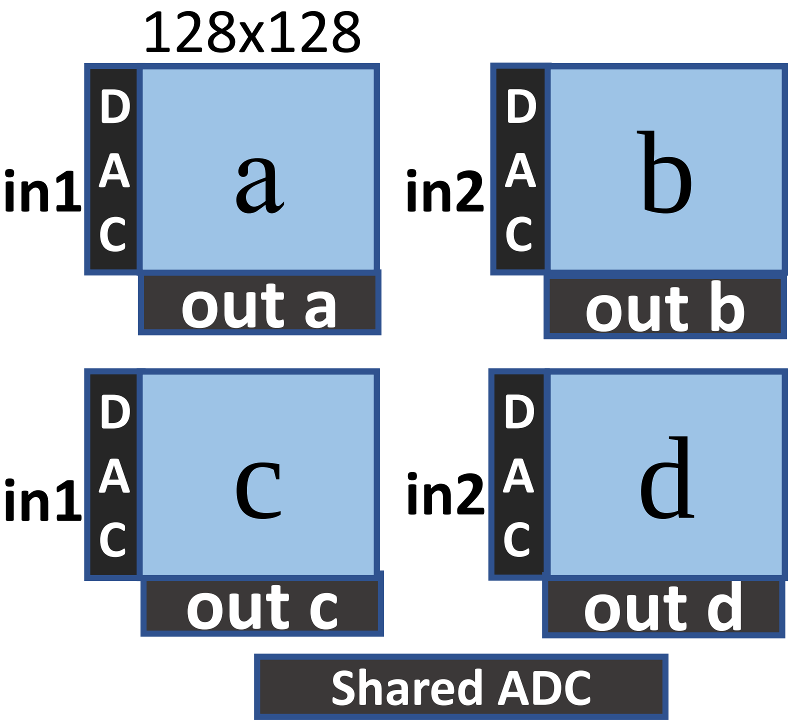
[14] C. Li et al. Analogue signal and image processing with large memristor crossbars. *Nature Electronics*, 1(1):52, 2018.

[18] A. Shafiee et. al., : A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Computer Architecture News*, 44(3):14–26, 2016.

MVM in In-memory DPE

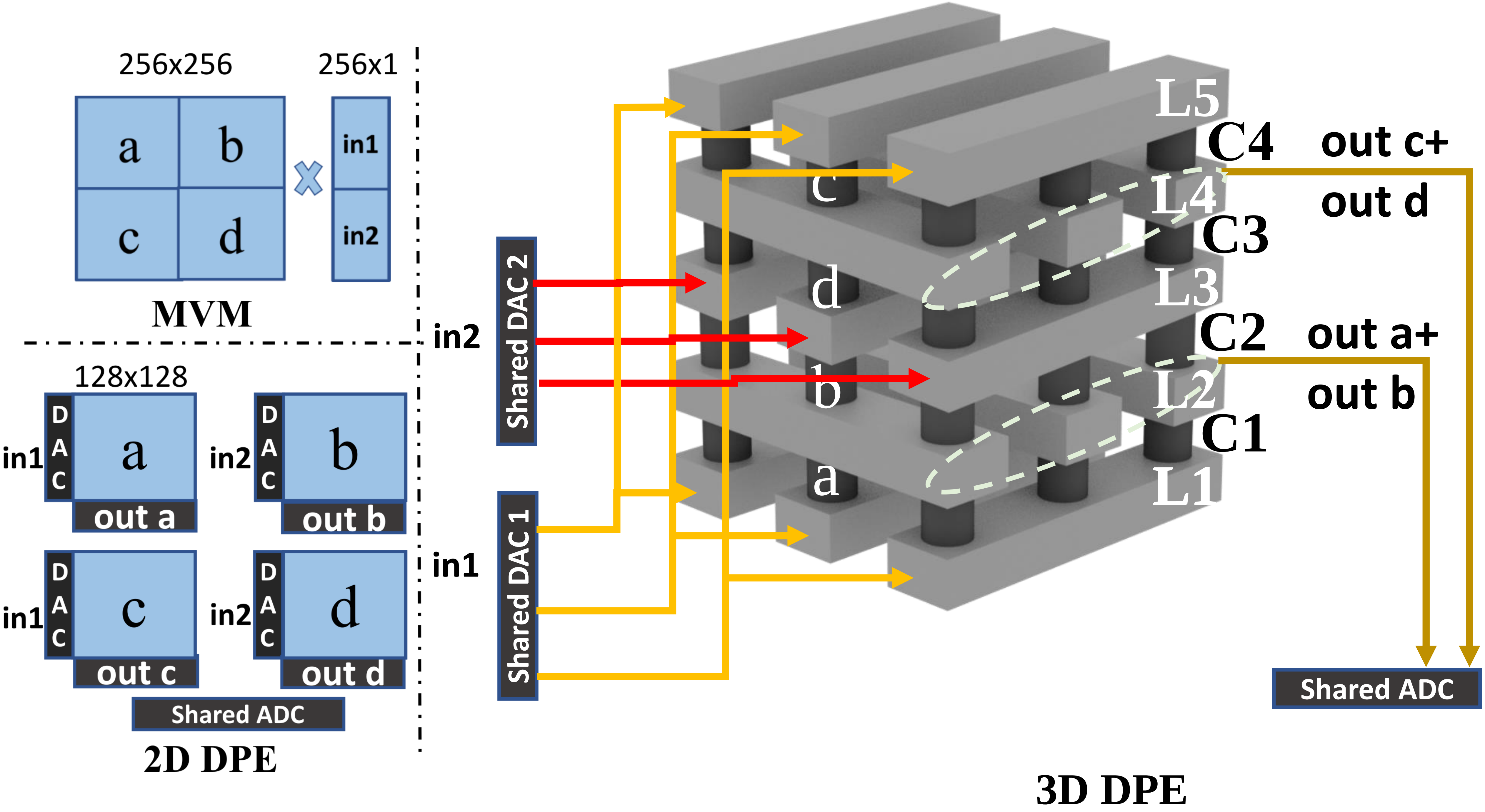


MVM



2D DPE

Hardware constraint:
Maximum Crossbar Dimension


256x256
256x1
a
b
c
d
in1
in2
MVM
128x128
in1
DAC
a
out a
in2
DAC
b
out b
in1
DAC
c
out c
in2
DAC
d
out d
Shared ADC
2D DPE
in2
Shared DAC 2
in1
Shared DAC 1
L5
C4
L4
C3
L3
C2
L2
C1
L1
c
d
b
a
out c+
out d
out a+
out b
Shared ADC
3D DPE

Overhead Comparison

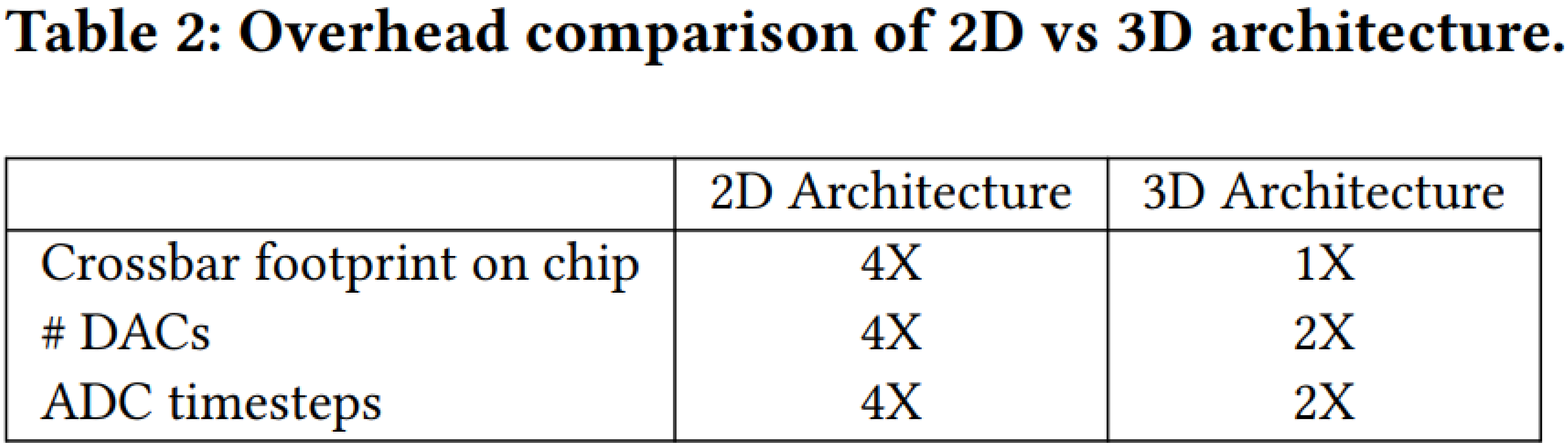
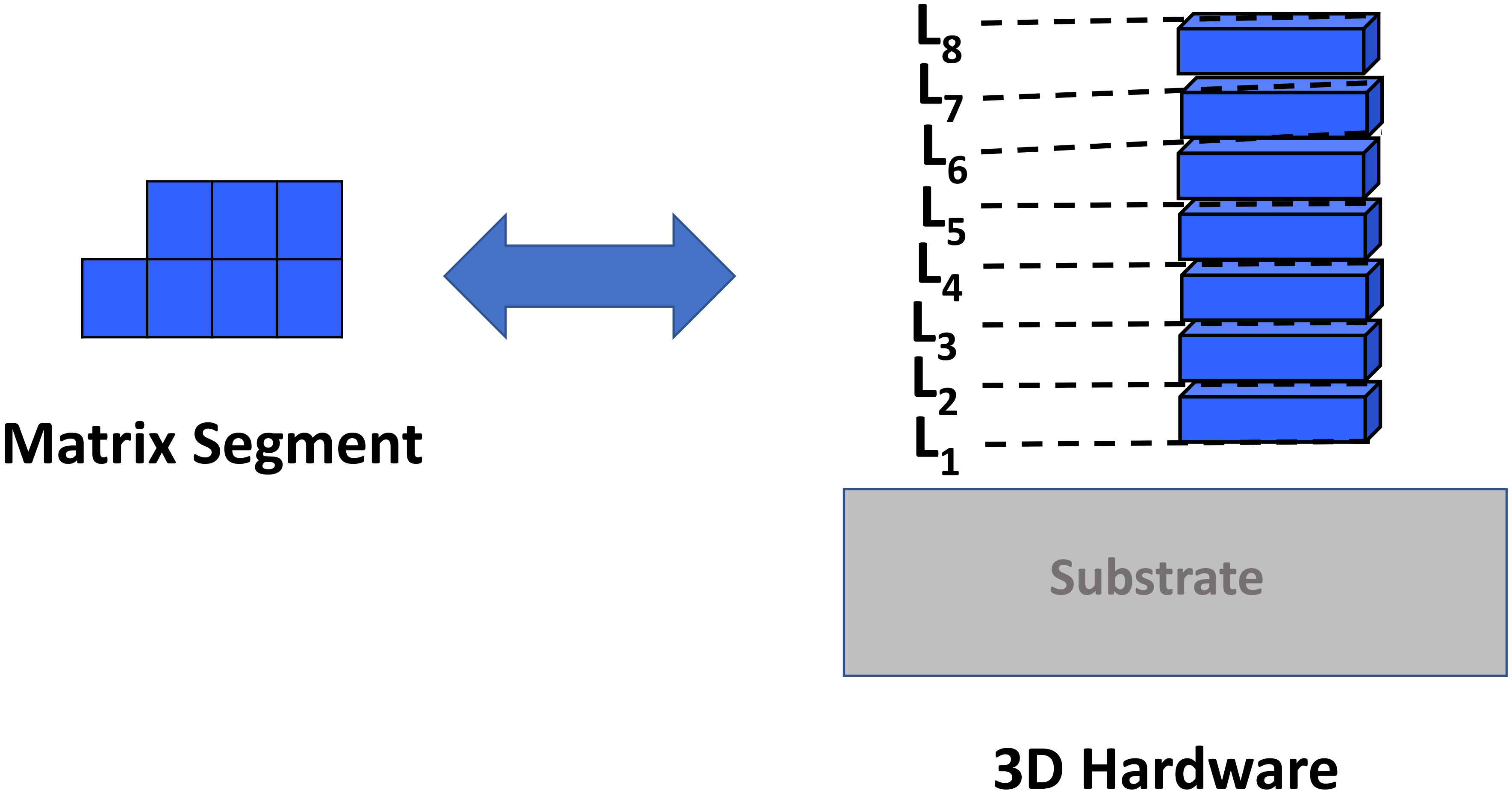


Table 2: Overhead comparison of 2D vs 3D architecture.

	2D Architecture	3D Architecture
Crossbar footprint on chip	4X	1X
# DACs	4X	2X
ADC timesteps	4X	2X

Problem 1: Hardware Mapping

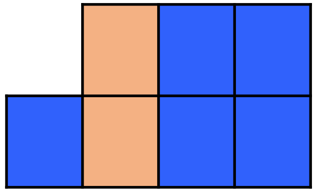


Matrix Segment

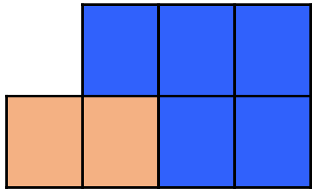
3D Hardware

Shareability Opportunity

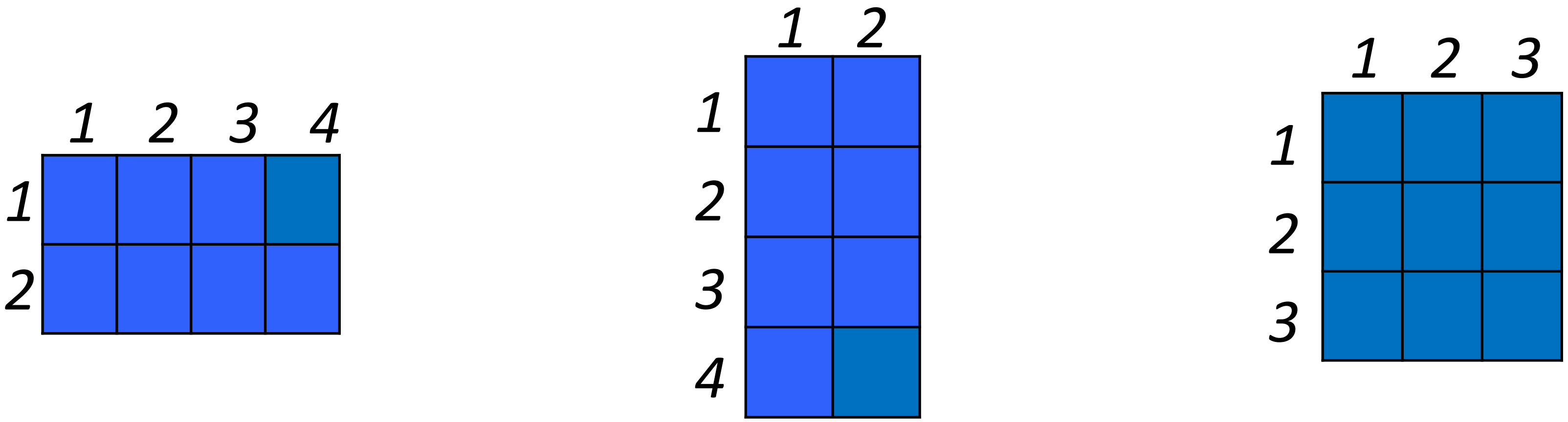
- Two adjacent crossbars in a 3D stack can share a set of DACs if the mapped matrix segments are in the same column i.e., to be multiplied with the same input vector



- The dot product output if two adjacent crossbars in a 3D stack can combined if the mapped matrix segments are in the same row i.e., contributing to the same output vector



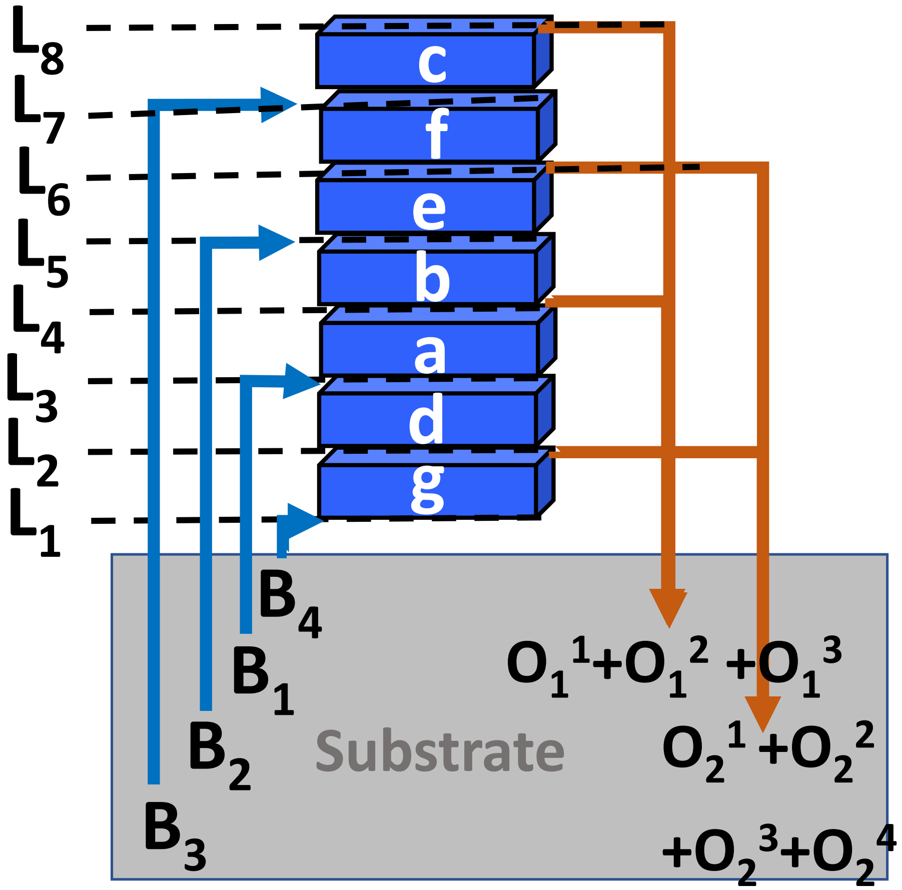
Problem 2: Discovering In-memory 3D DPE Kernel



Discovering In-memory 3D DPE Kernel

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & a & b & c & \\ 2 & d & e & f & g \end{array} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{bmatrix} = \begin{bmatrix} O_1^1+O_1^2+O_1^3 \\ O_2^1+O_2^2+O_2^3+O_2^4 \end{bmatrix}$$

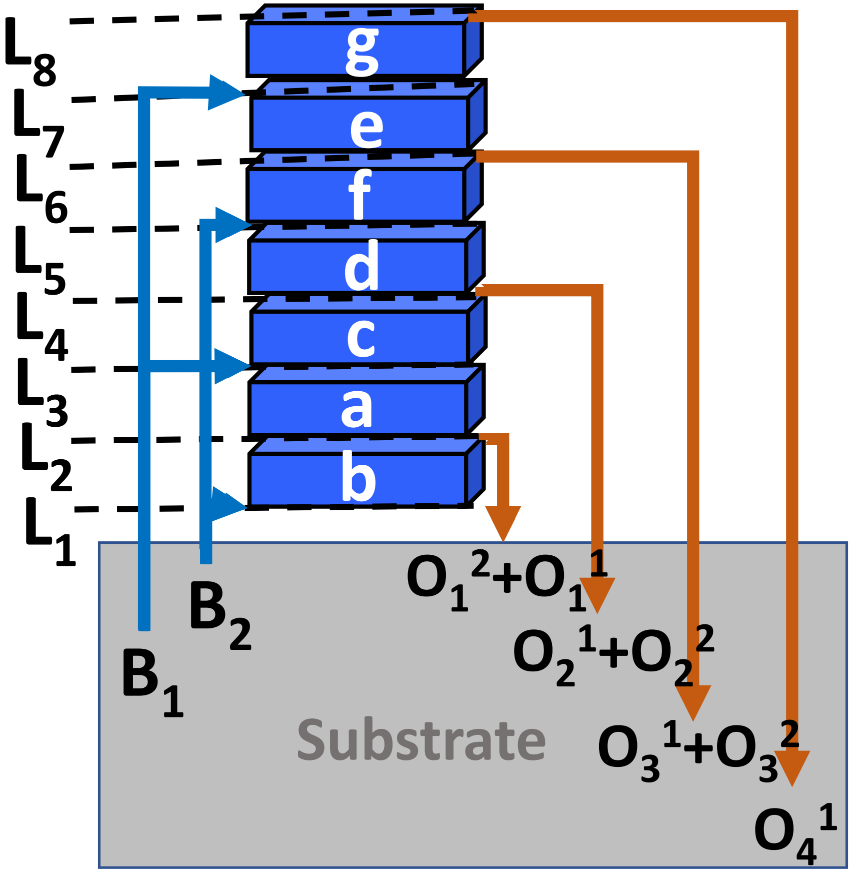
Kernel 1



Discovering In-memory 3D DPE Kernel

$$\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & a & b \\ 2 & c & d \\ 3 & e & f \\ 4 & g & \end{array} \times \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} O_1^2+O_1^1 \\ O_2^1+O_2^2 \\ O_3^1+O_3^2 \\ O_4^1 \end{bmatrix}$$

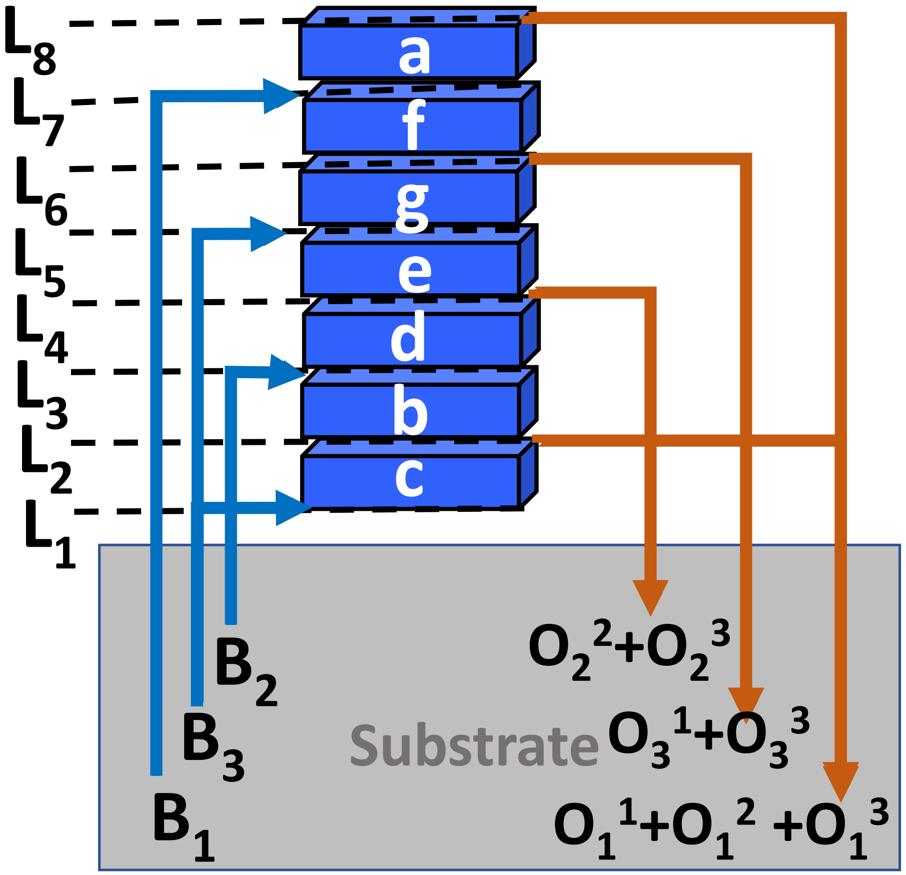
Kernel 2



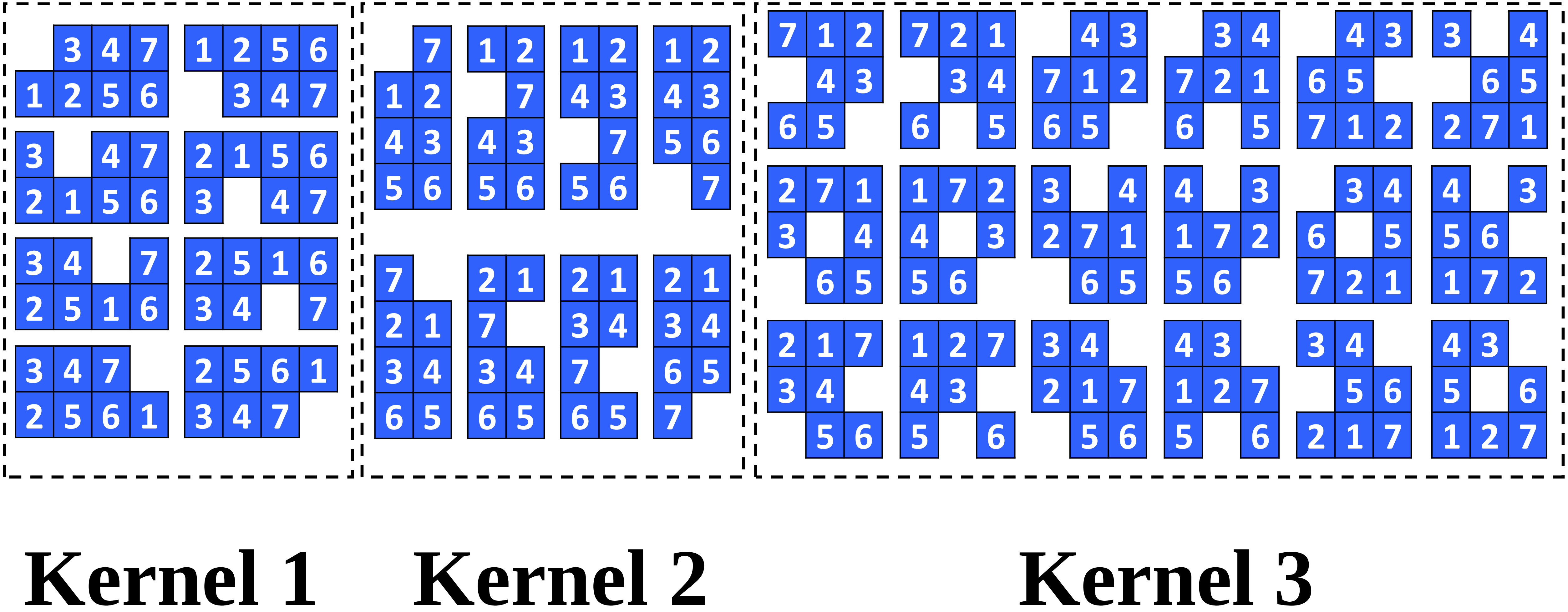
Discovering In-memory 3D DPE Kernel

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & a & b & c \\ 2 & & d & e \\ 3 & f & & g \end{array} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = \begin{bmatrix} O_1^1+O_1^2+O_1^3 \\ O_2^2+O_2^3 \\ O_3^1+O_3^3 \end{bmatrix}$$

Kernel 3

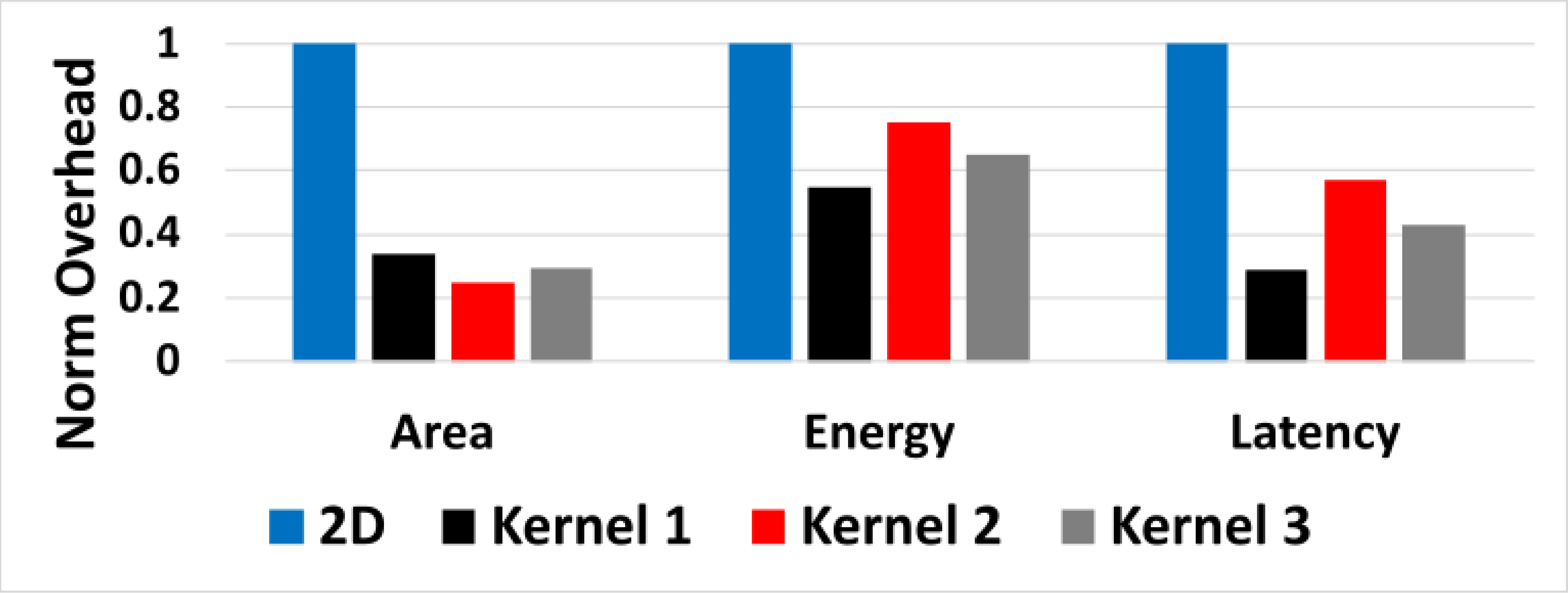


Library of In-memory 3D DPE Kernel

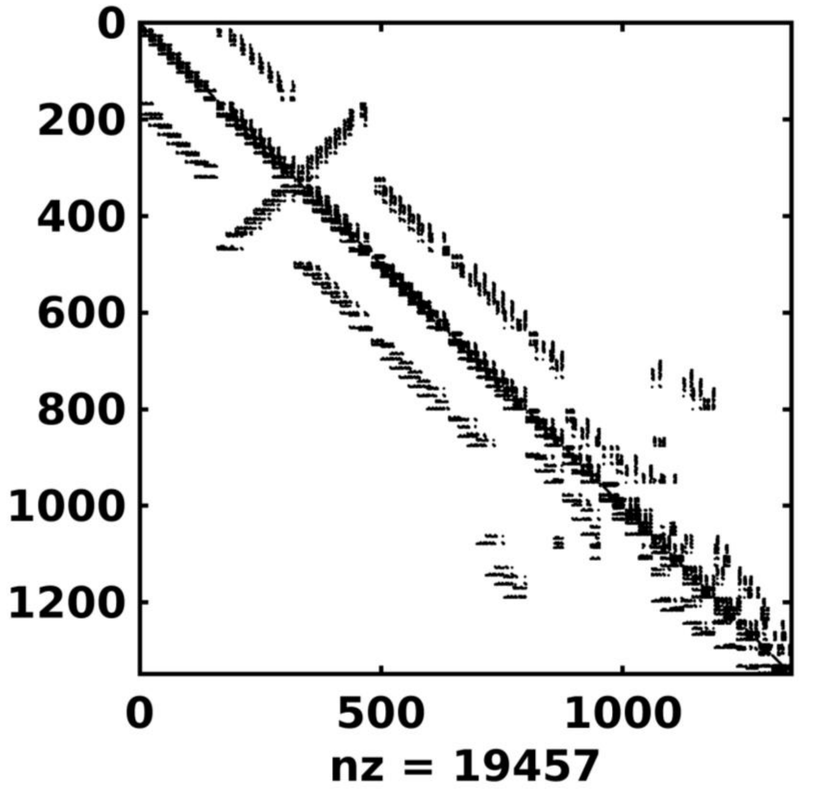


Kernel 1 **Kernel 2** **Kernel 3**

Overhead Comparison



Application level

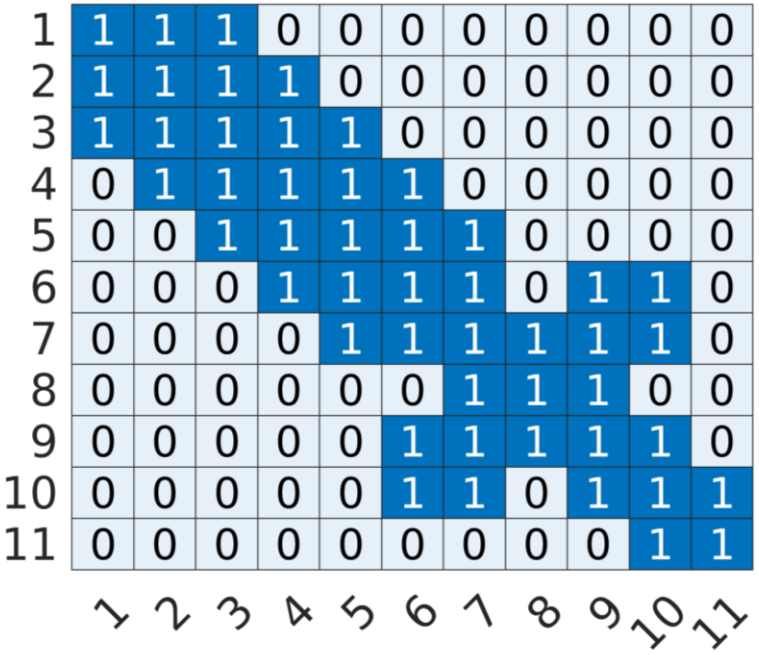


Problem:

- Matrix dimension might exceed crossbar dimension

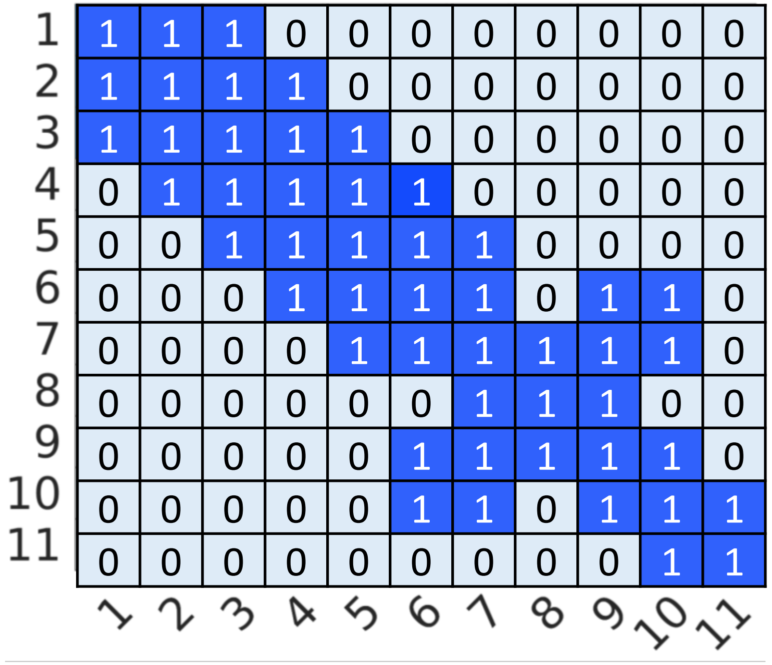
Solution:

- Partitioning Matrix into segments



	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0	0	0
3	1	1	1	1	1	0	0	0	0	0	0
4	0	1	1	1	1	1	0	0	0	0	0
5	0	0	1	1	1	1	1	0	0	0	0
6	0	0	0	1	1	1	1	0	1	1	0
7	0	0	0	0	1	1	1	1	1	1	0
8	0	0	0	0	0	0	1	1	1	0	0
9	0	0	0	0	0	1	1	1	1	1	0
10	0	0	0	0	0	1	1	0	1	1	1
11	0	0	0	0	0	0	0	0	0	1	1

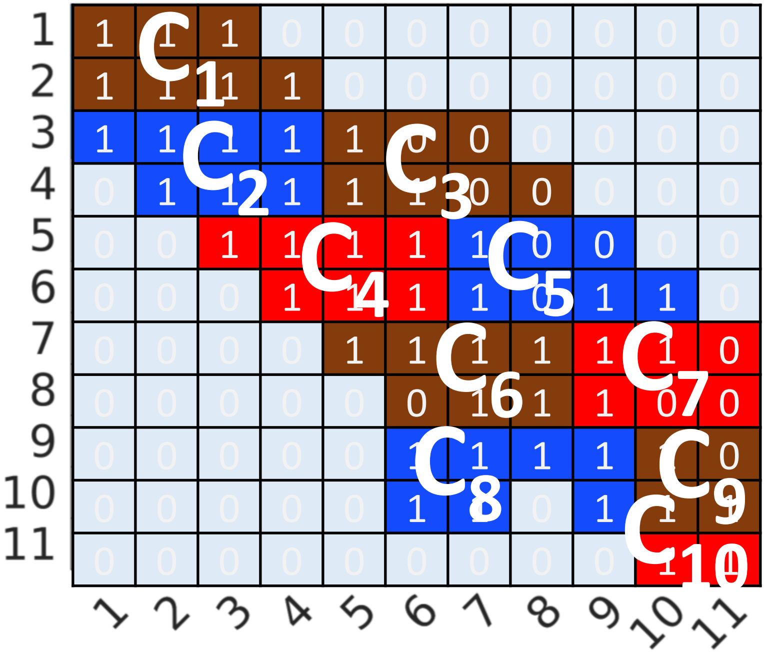
Greedy Kernel Cover of Sparse System



	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0	0	0
3	1	1	1	1	1	0	0	0	0	0	0
4	0	1	1	1	1	1	0	0	0	0	0
5	0	0	1	1	1	1	1	0	0	0	0
6	0	0	0	1	1	1	1	0	1	1	0
7	0	0	0	0	1	1	1	1	1	1	0
8	0	0	0	0	0	0	1	1	1	0	0
9	0	0	0	0	0	1	1	1	1	1	0
10	0	0	0	0	0	1	1	0	1	1	1
11	0	0	0	0	0	0	0	0	0	1	1

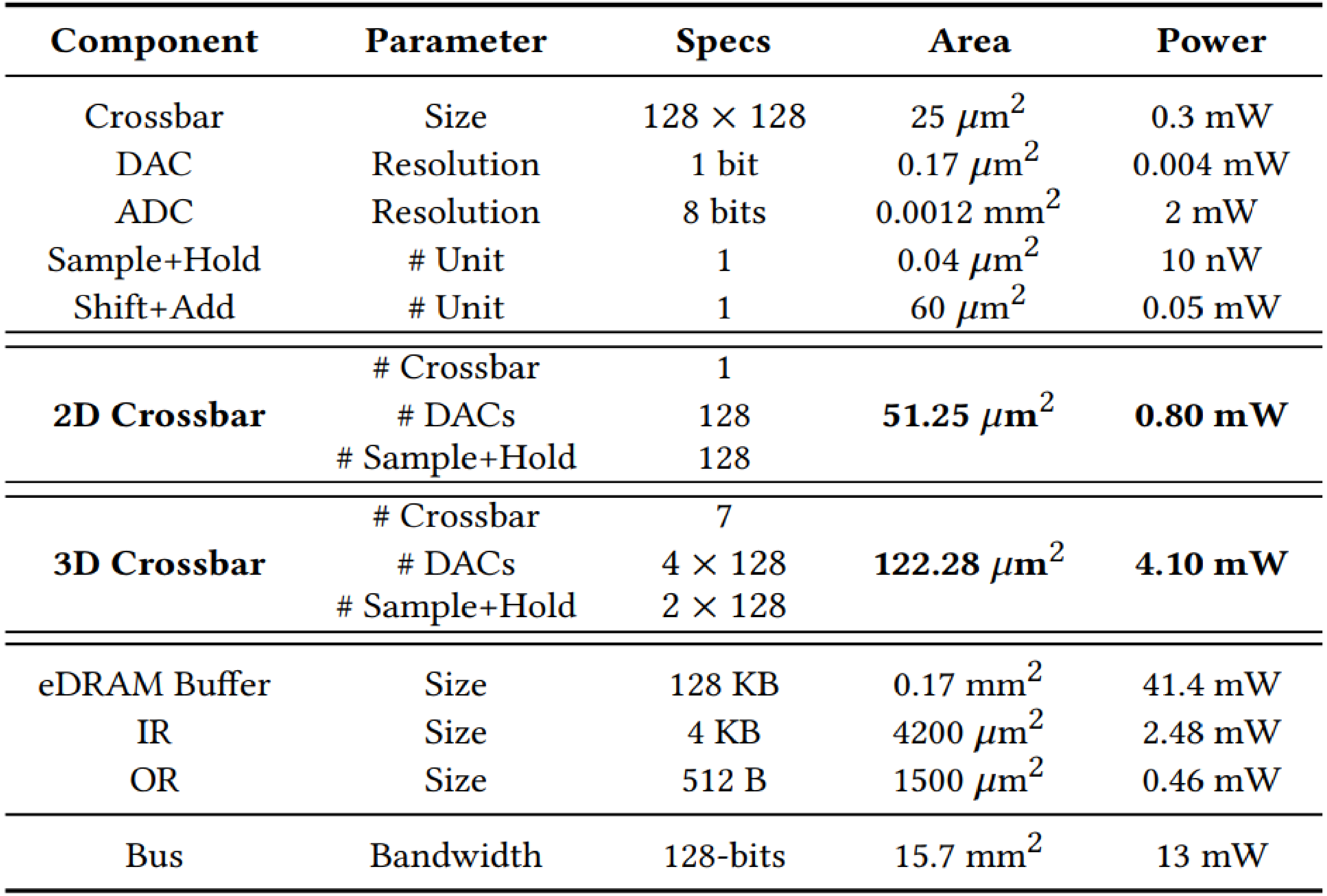
Target:

- Covering non-zero blocks using the kernel library



Evaluation

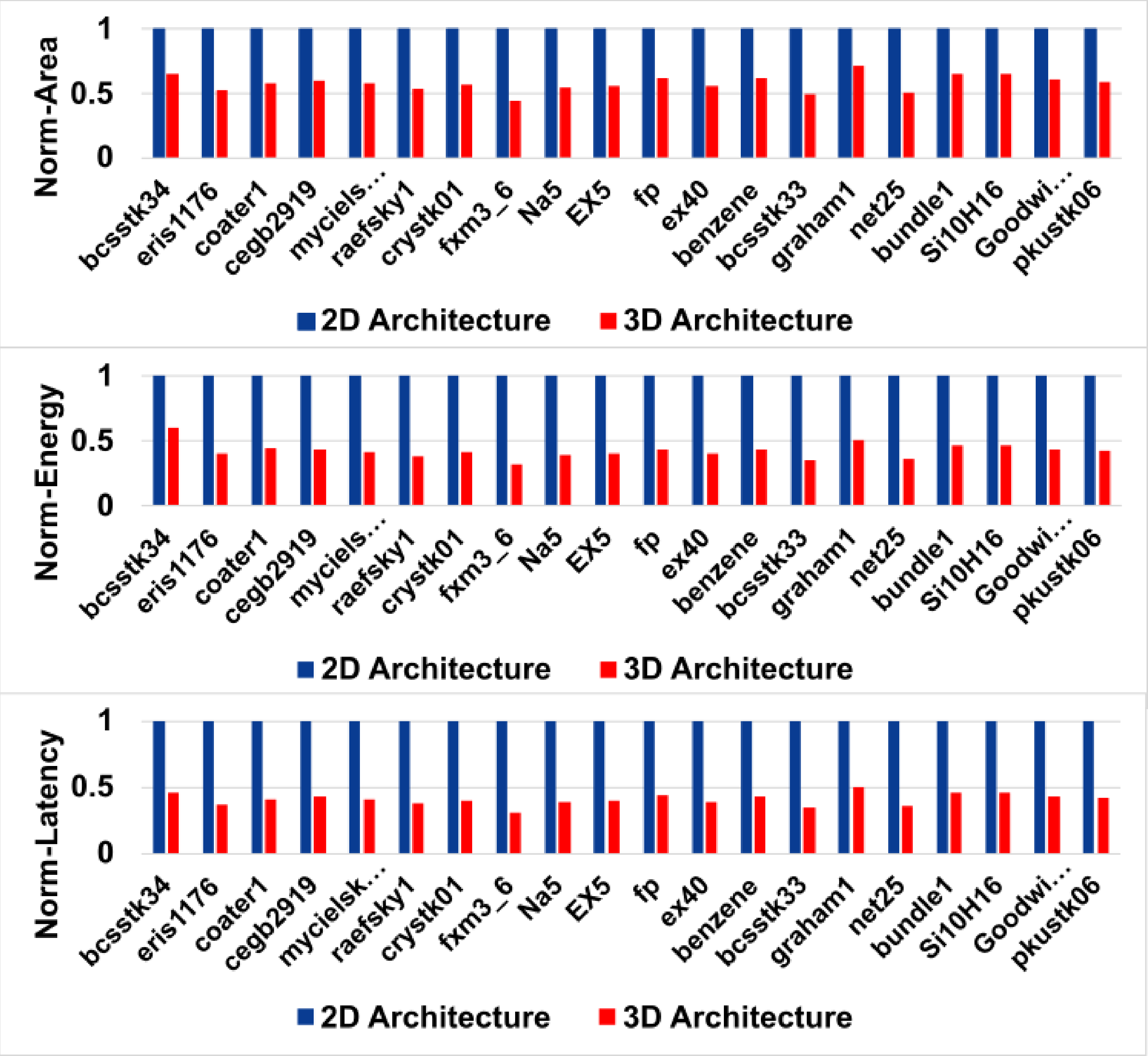
Area-power cost of architectural components



Component	Parameter	Specs	Area	Power
Crossbar	Size	128 × 128	25 μm^2	0.3 mW
DAC	Resolution	1 bit	0.17 μm^2	0.004 mW
ADC	Resolution	8 bits	0.0012 mm^2	2 mW
Sample+Hold	# Unit	1	0.04 μm^2	10 nW
Shift+Add	# Unit	1	60 μm^2	0.05 mW
2D Crossbar	# Crossbar	1	**51.25 μm^2**	**0.80 mW**
	# DACs	128		
	# Sample+Hold	128		
3D Crossbar	# Crossbar	7	**122.28 μm^2**	**4.10 mW**
	# DACs	4 × 128		
	# Sample+Hold	2 × 128		
eDRAM Buffer	Size	128 KB	0.17 mm^2	41.4 mW
IR	Size	4 KB	4200 μm^2	2.48 mW
OR	Size	512 B	1500 μm^2	0.46 mW
Bus	Bandwidth	128-bits	15.7 mm^2	13 mW

Table 4: Overview of benchmarks from the SuiteSparse Matrix Collection.

Applications	Systems	Matrix Dimensions	#Non-zeros
bcsstk34	Structural Problem	588 × 588	21418
eris1176	Power Network Problem	1176 × 1176	18552
coater1	Computational Fluid Dynamics	1348 × 1348	19457
cegb2919	Structural Problem	2919 × 2919	321543
mycielskian12	Undirected Graph	3071 × 3071	407200
raefsky1	Computational Fluid Dynamics	3242 × 3242	293409
crystk01	Materials Problem	4875 × 4875	315891
fxm3_6	Optimization Problem	5026 × 5026	94026
Na5	Theoretical/Quantum Chemistry	5832 × 5832	305630
EX5	Combinatorial Problem	6545 × 6545	295680
fp	Electromagnetics Problem	7548 × 7548	834222
ex40	Computational Fluid Dynamics	7740 × 7740	456188
benzene	Theoretical/Quantum Chemistry	8219 × 8219	242669
bcsstk33	Structural Problem	8738 × 8738	591904
graham1	Computational Fluid Dynamics	9035 × 9035	335472
net25	Optimization Problem	9520 × 9520	401200
bundle1	Computer Graphics/Vision	10581 × 10581	770811
Si10H16	Theoretical/Quantum Chemistry	17077 × 17077	875923
Goodwin_040	Computational Fluid Dynamics	17922 × 17922	561677
pkustk06	Structural Problem	43164 × 43164	2571768



Improvements:

- **Area by 2.02X**
- **Latency by 2.45X**
- **Energy by 2.37X**

Summary

Thank You

Contact us:

rashed09@knights.ucf.edu

sumit.jha@utsa.edu

rickard.ewetz@ucf.edu