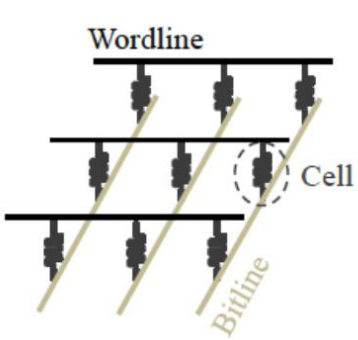
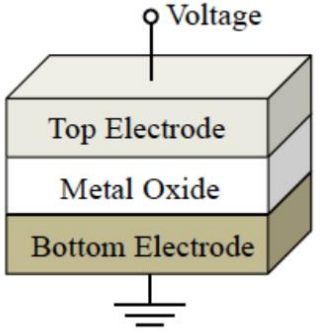
Thermal-Aware Layout Optimization and Mapping Methods for Resistive Neuromorphic Engines

Chengrui Zhang, Yu Ma, and Pingqiang Zhou

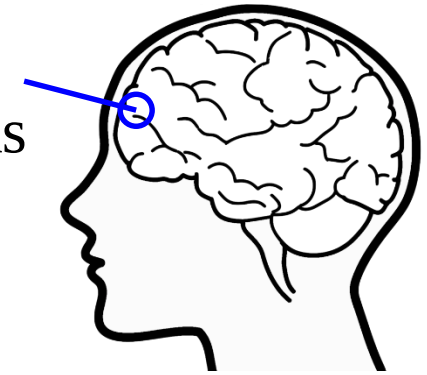
Outline



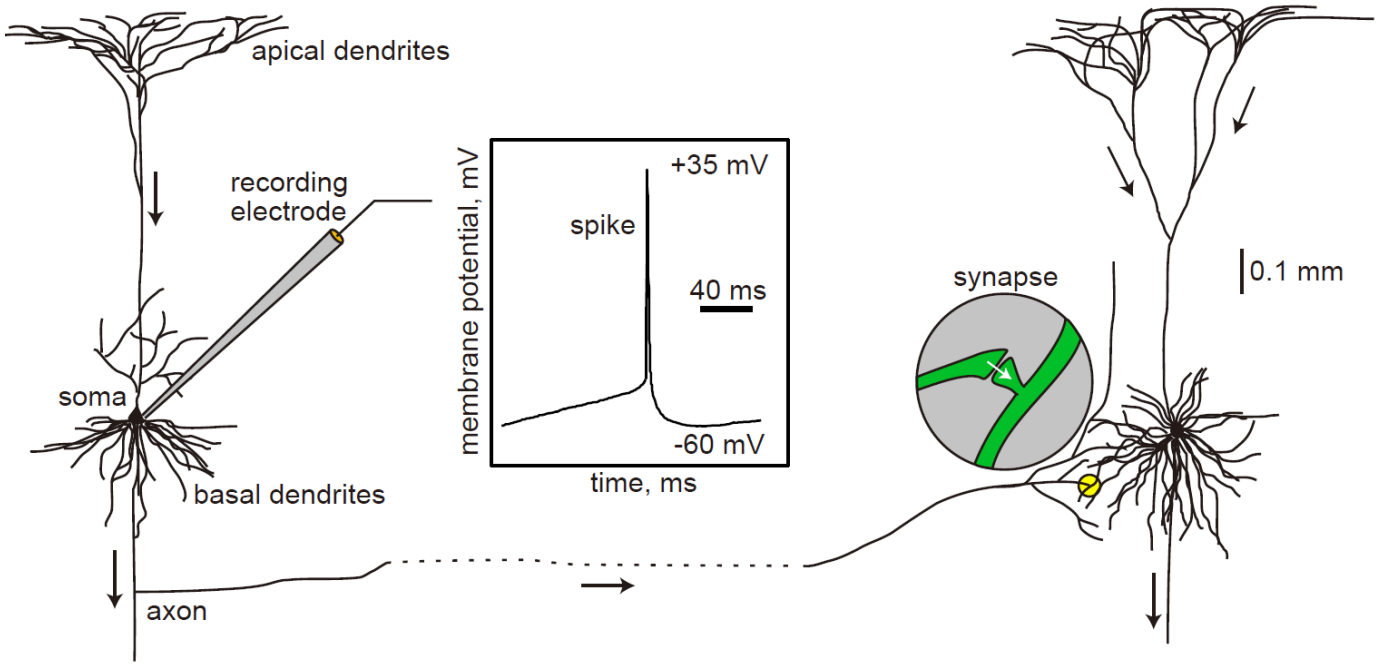


- Background
  - Spiking Neural Networks (SNNs)
  - Memristor Crossbar Array (MCA) based Accelerators
- Problem Formulation
- Proposed Techniques
  - Thermal-Aware Layout Design
  - Input-sensitive Cross-Array Mapping
- Experimental Results
- Summary

Spiking Neural Networks (SNNs)

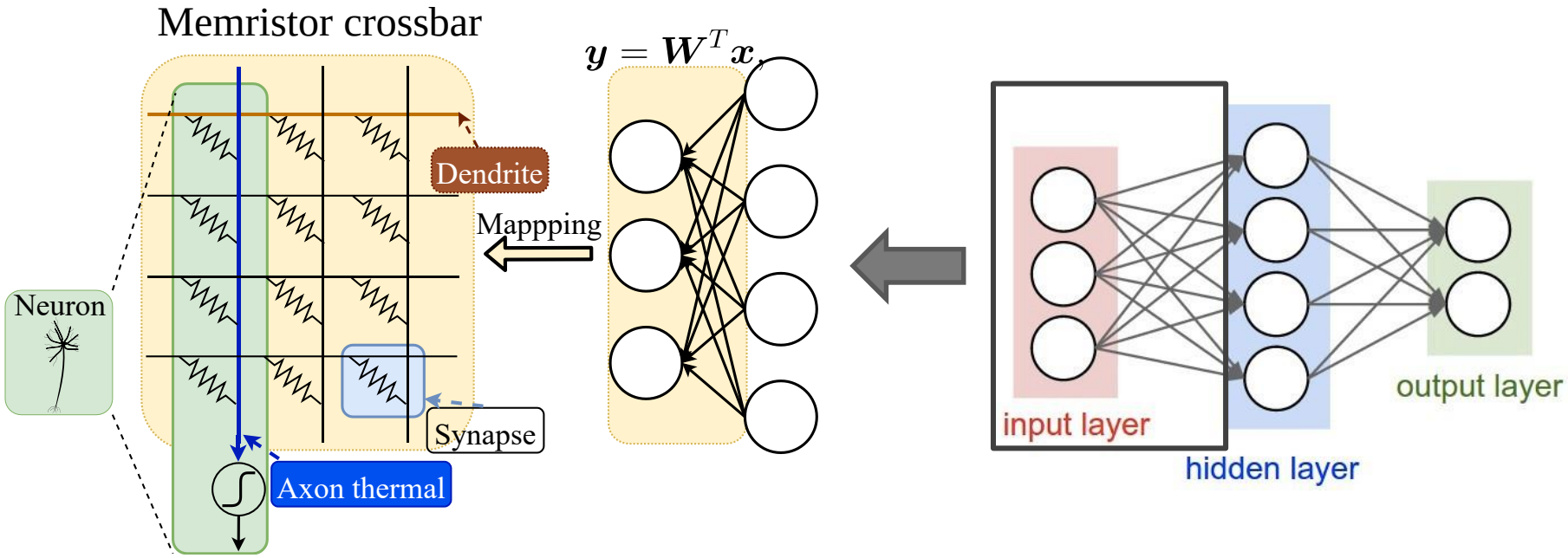


- The **third** generation of artificial neural networks (ANN)
  - Simulation of biological neurons
  - Data transmission – spikes
  - Neural imitation – neuron model



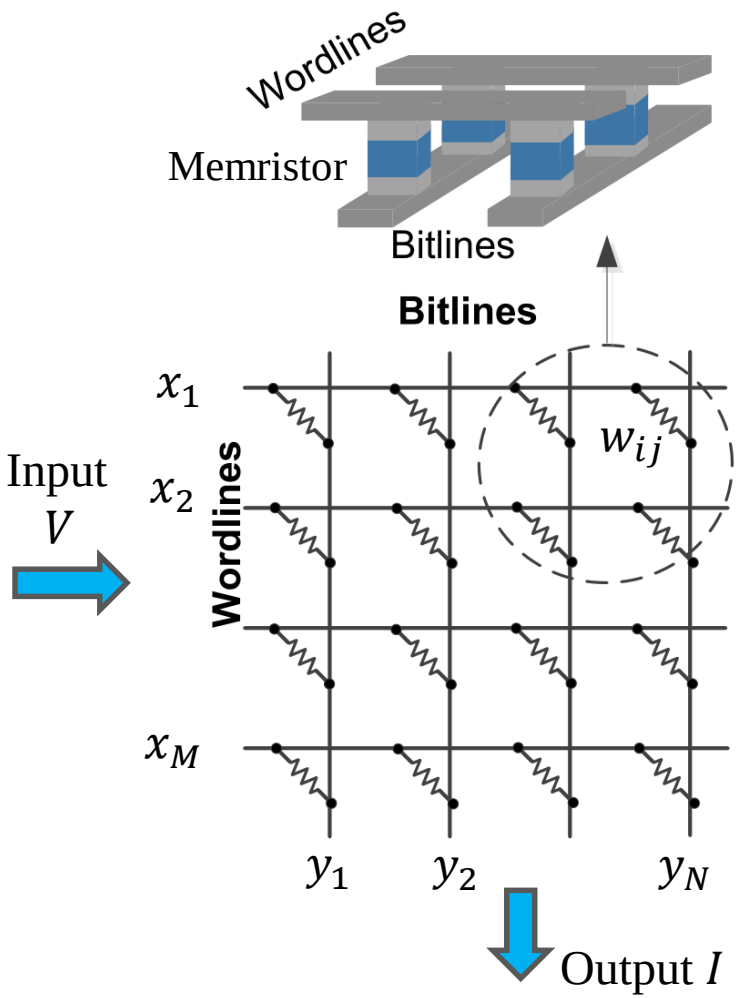
E. M. Izhikevich, Dynamical systems in neuroscience : the geometry of excitability and bursting, 2007.

Memristor-based Neural Network Accelerators



- Promising accelerator architecture for deep learning applications
- Weights stored in Memristor crossbar
- "Process in memory" manner
  - Less data movements
  - High density computation: Multiply Accumulate with time complexity O(1)

Memristor-based Neural Network Accelerators



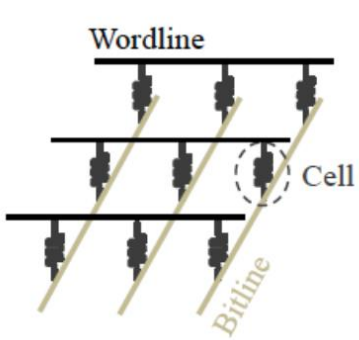
$$\boldsymbol{y} = \boldsymbol{W}^T \boldsymbol{x}, \quad W \in R^{M \times N}, x \in R^M, y \in R^N$$

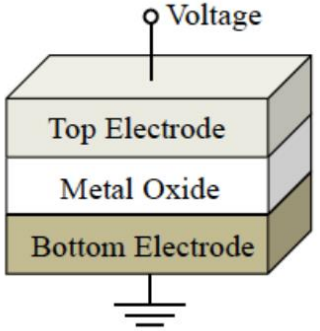
$$I_n = \sum_{m=1}^{M} G_{m,n} V_{in,m} \quad G \in R^{M \times N}, V \in R^M$$

Input vector	x
Weight matrix	W
Output vector	y
Input voltage vector	V
Conductance matrix	G
Output current vector	I

Xu, Cong, et al. "Overcoming the challenges of crossbar resistive memory architectures." HPCA 2015.

Outline



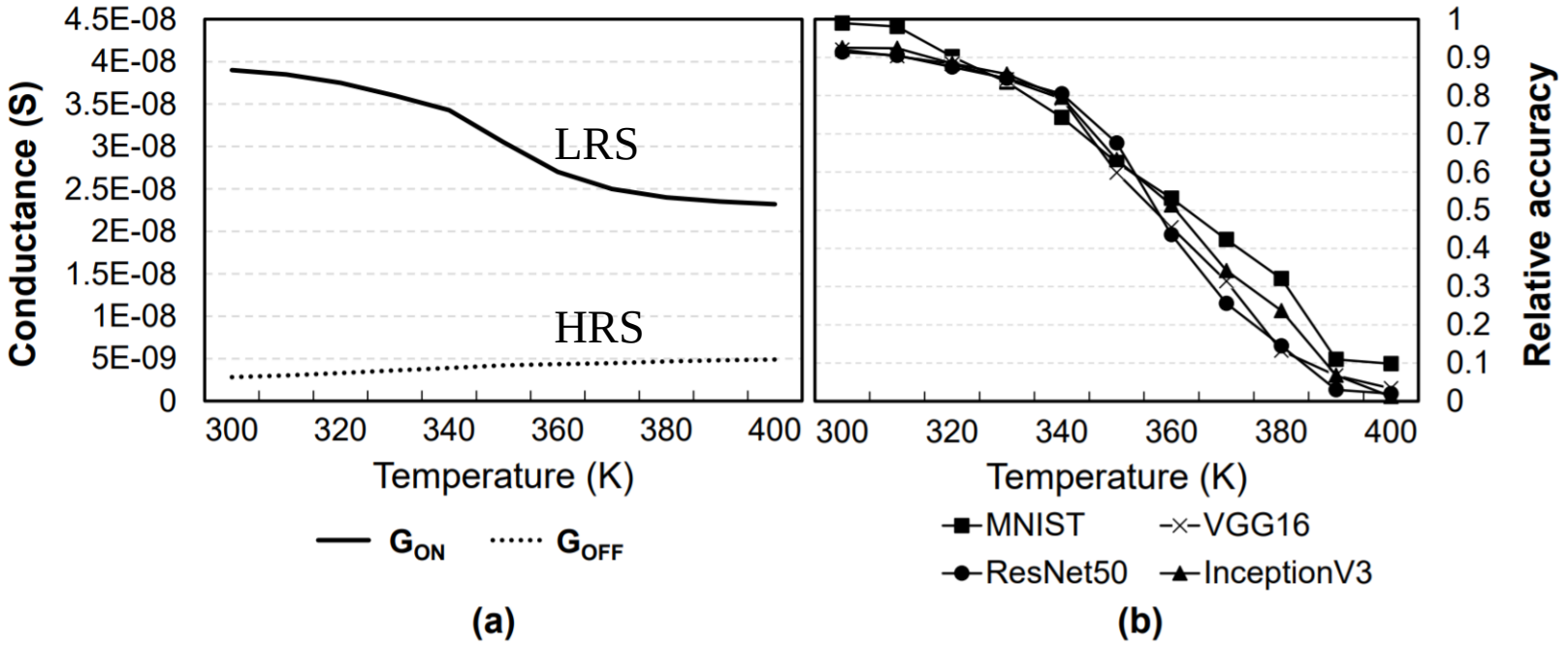


- Background
  - Spiking Neural Networks (SNNs)
  - Memristor Crossbar Array (MCA) based Accelerators
- Problem Formulation
- Proposed Techniques
  - Thermal-Aware Layout Design
  - Input-sensitive Cross-Array Mapping
- Experimental Results
- Summary

Reliability Challenge: Thermal Effect

- Conductance changes as temperature changes
- Accuracy & endurance degradation

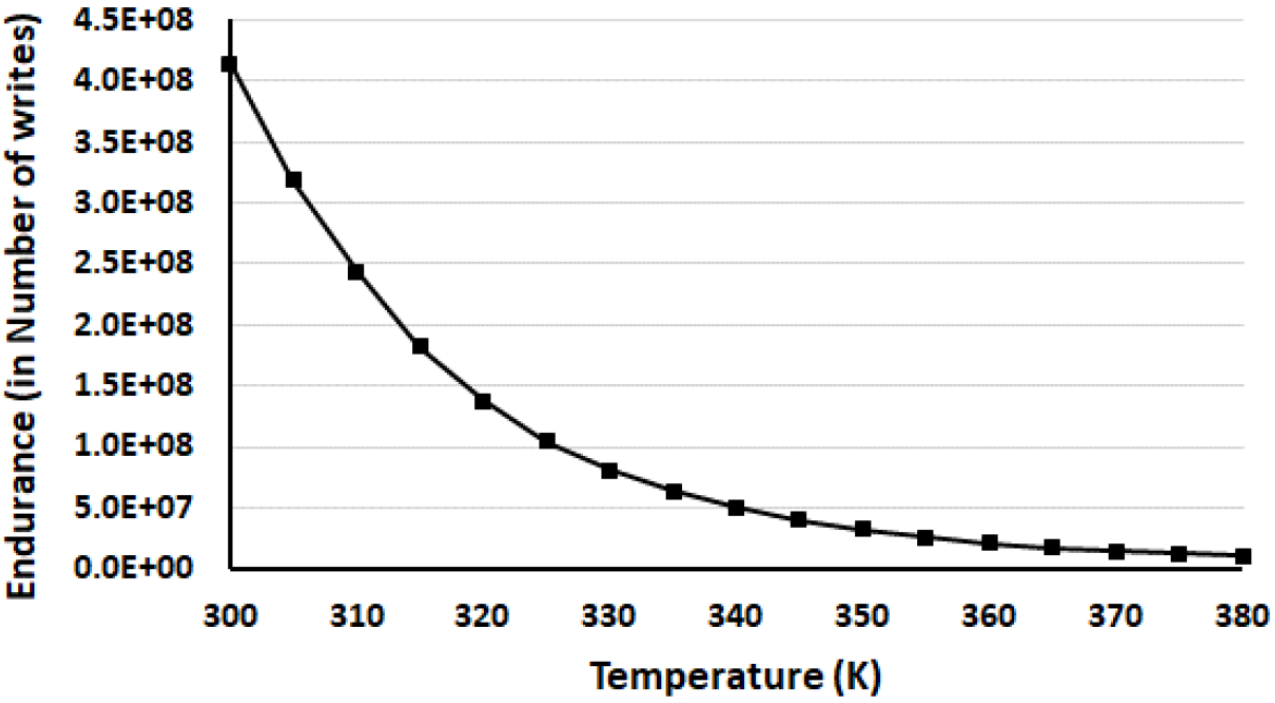
LRS/HRS: Low/High resistance state



X. Liu, M. Zhou, T. S. Rosing and J. Zhao, "HR3AM: A Heat Resilient Design for RRAM-based Neuromorphic Computing," ISLPED, 2019.

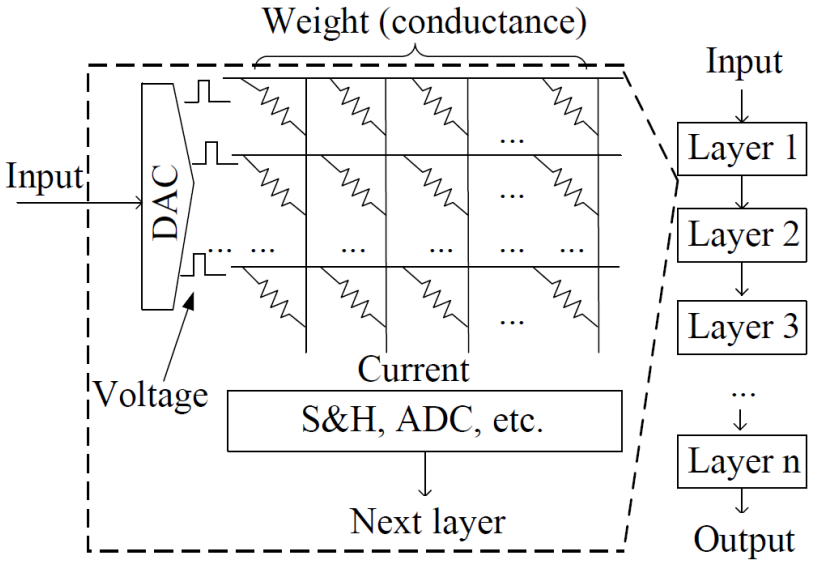
Reliability Challenge: Thermal Effect

- Conductance changes as temperature changes
- Accuracy & endurance degradation



Valad Beigi and M., Memik G, “Thor: Thermal-aware optimizations for extending ReRAM lifetime”, IPDPS, 2018.

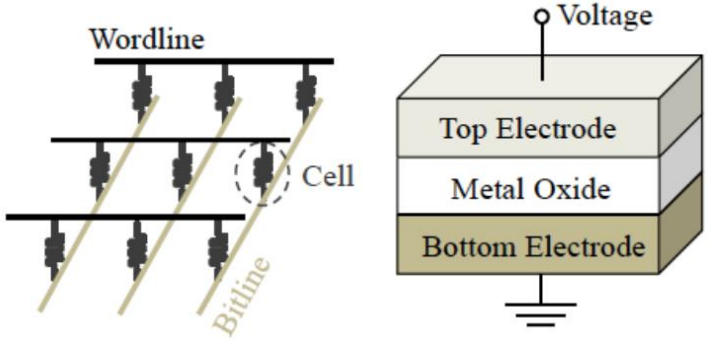
Related Works



$$P = V^2 * G$$
$$\propto Input^2 * Weight$$

- Compensate conductance decrease [X. Liu, ISLPED 2019]
- Weight training [S. Zhang, DATE 2019]
  - Reduce the total weight value to reduce total conductance
- Thermal-aware mapping [H. Shin, ICCAD 2020]
  - Even the power distribution to reduce the maximum temperature

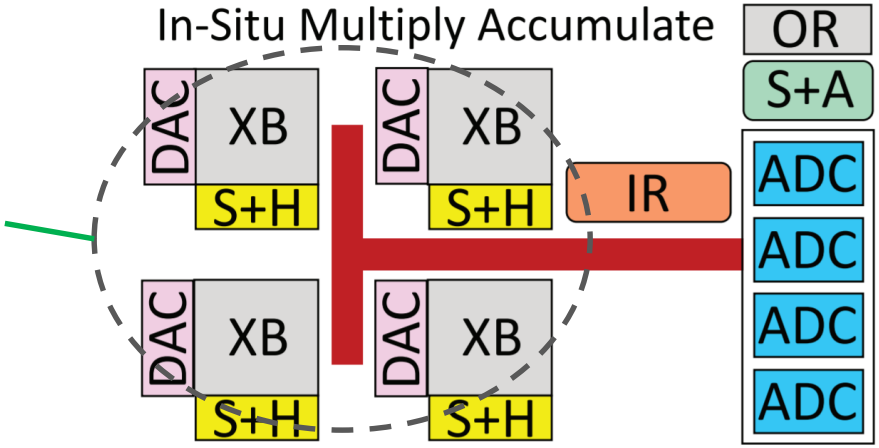
Outline



- Background
  - Spiking Neural Networks (SNNs)
  - Memristor Crossbar Array (MCA) based Accelerators
- Problem Formulation
- Proposed Techniques
  - Thermal-Aware Layout Design
  - Input-sensitive Cross-Array Mapping
- Experimental Results
- Summary

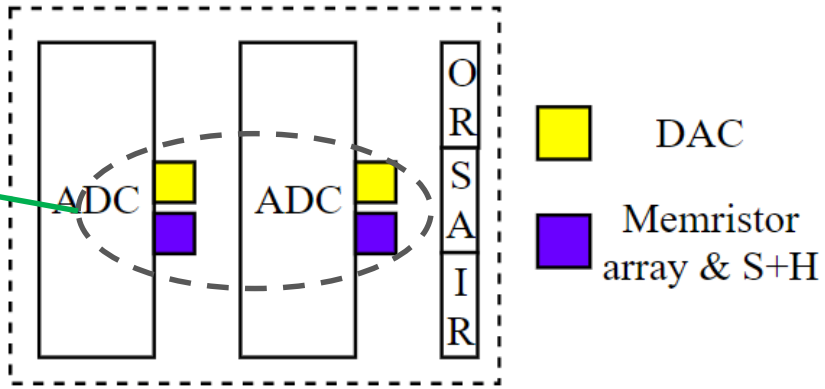
Thermal-Aware Layout Design

Hot
(largest power density)



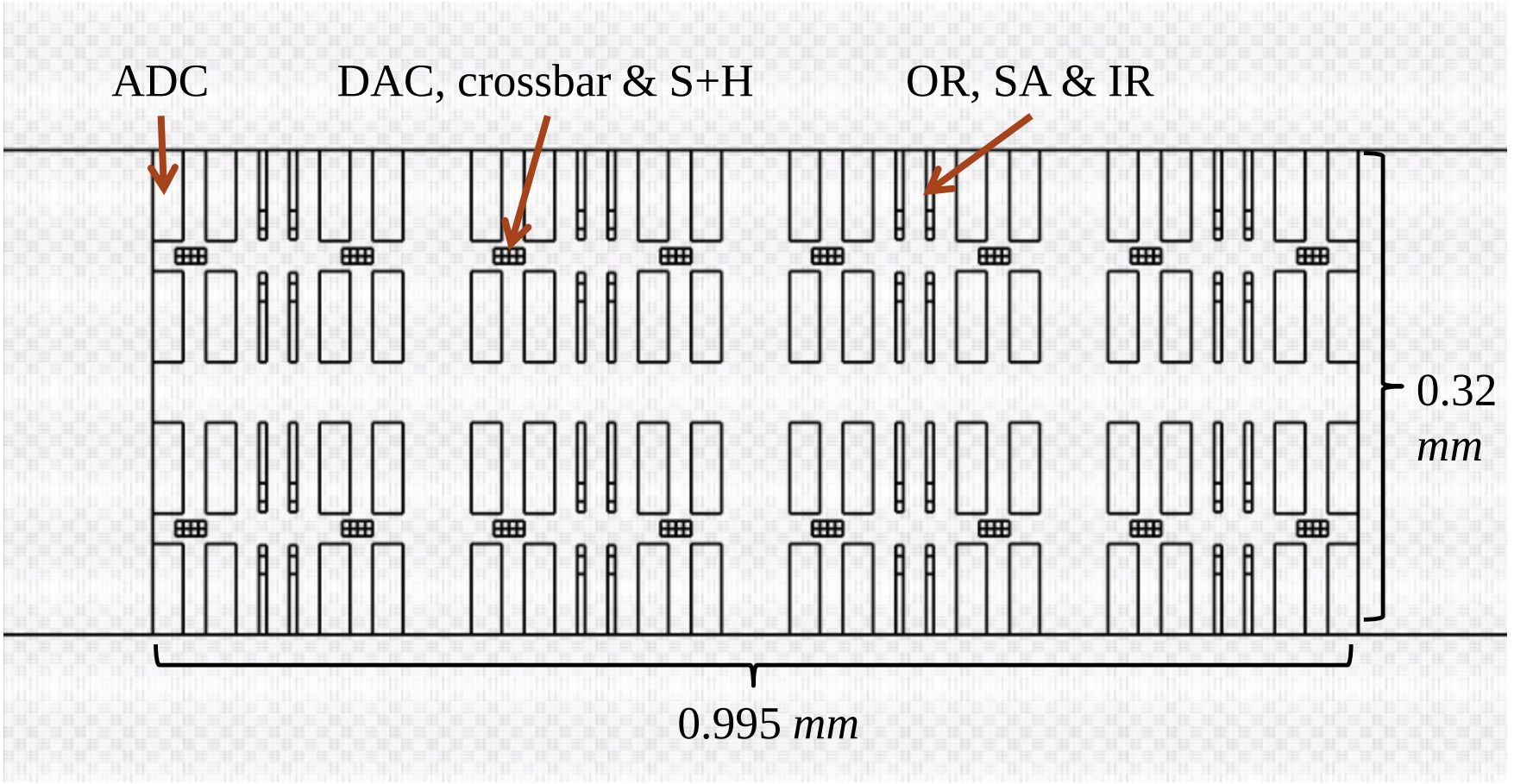
[A. Shafiee, ISCA 2016]

Depart the high power density parts



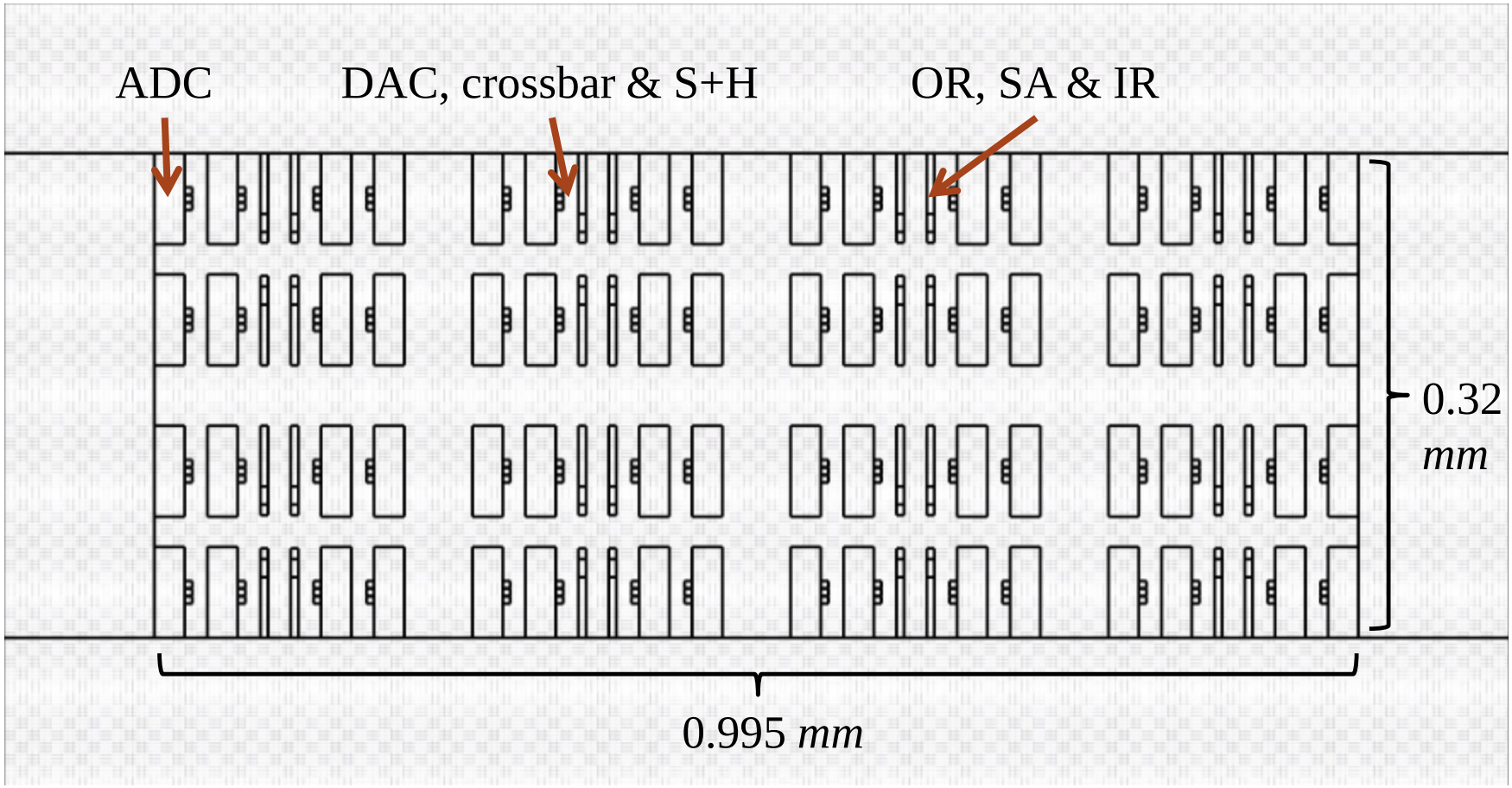
Ours

Thermal-Aware Layout Design



Possible layout of previous architecture

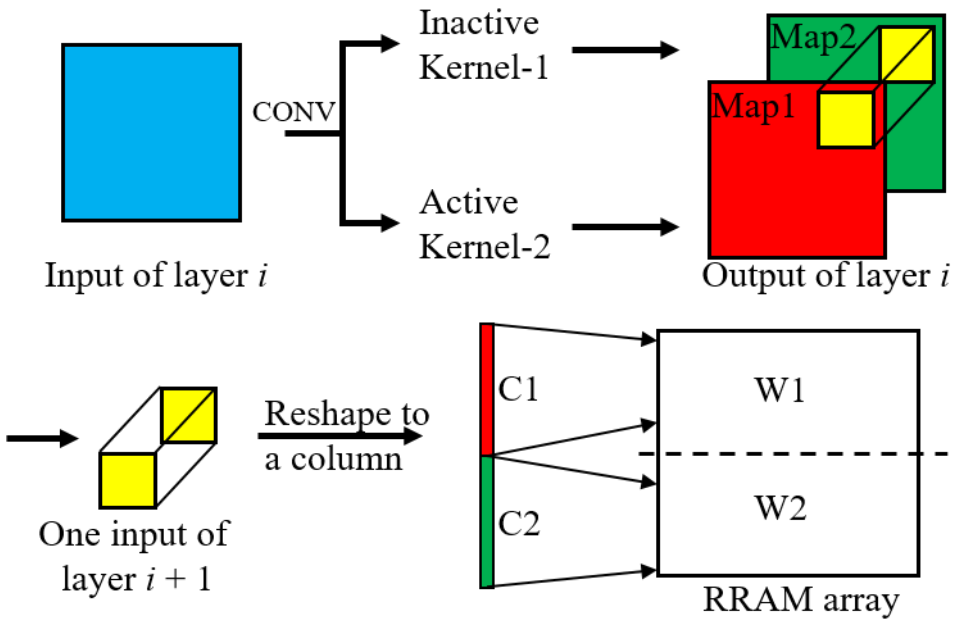
Thermal-Aware Layout Design



Layout of our method

Input-sensitive Cross-Array Mapping

$$P = V^2 * G$$
$$\propto \boldsymbol{Input^2} * Weight$$



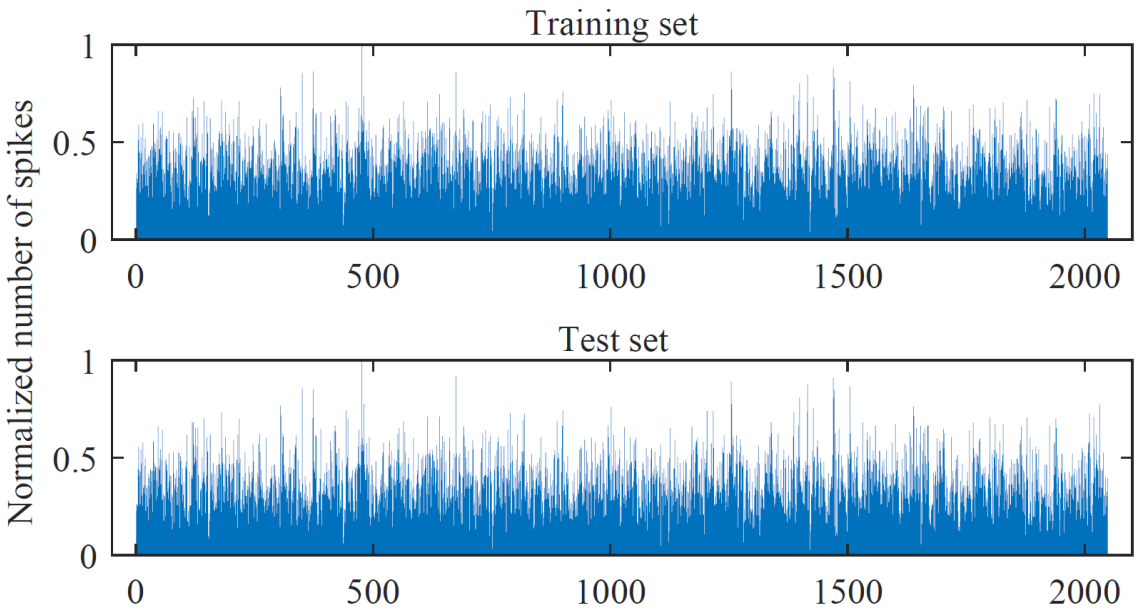
$Dist_i$: Input distribution of layer i

Using $Dist_i * Weight_i$ to guide our cross-array mapping method

One inference step in SNN from layer i to layer $i + 1$

Input-sensitive Cross-Array Mapping

$$P = V^2 * G$$
$$\propto \boldsymbol{Input^2} * Weight$$



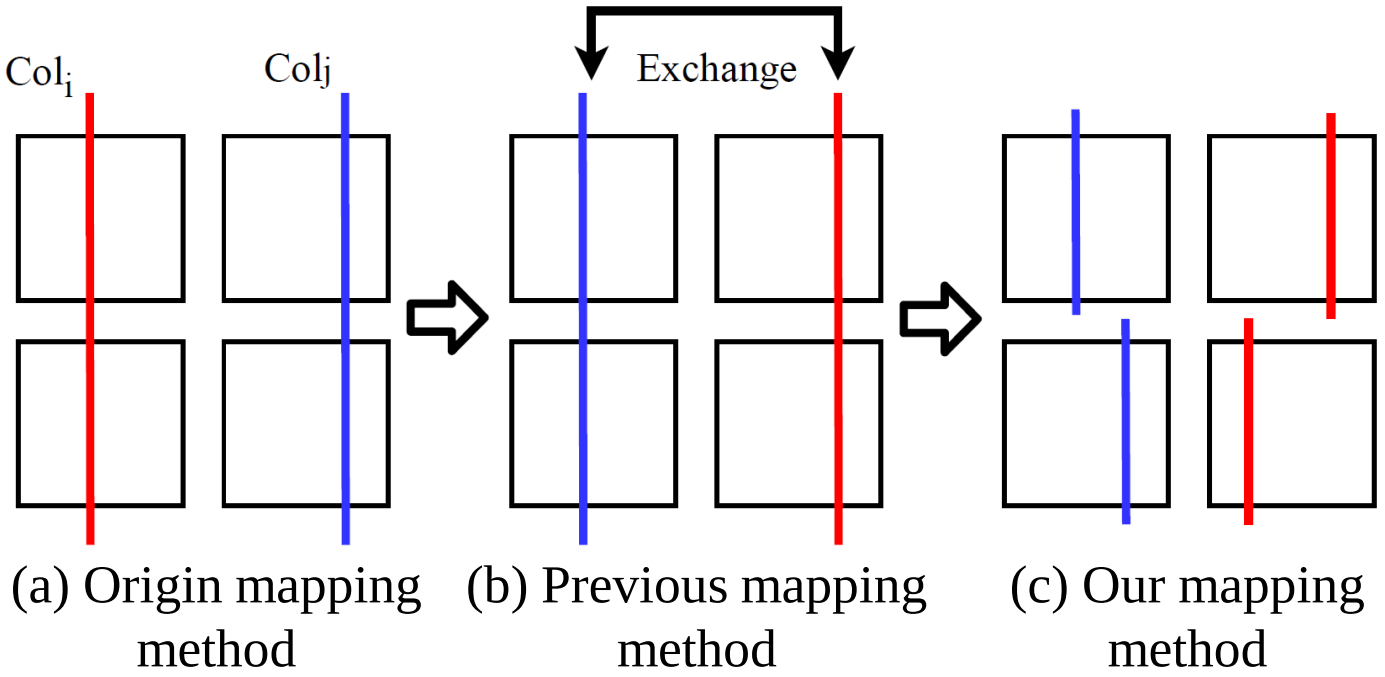
The normalized input distribution of the second fully-connected layer of VGG11 in training set and test set

$Dist_i$: Input distribution of layer i

Using $Dist_i * Weight_i$ to guide our cross-array mapping method

Input-sensitive Cross-Array Mapping

$$P = V^2 * G$$
$$\propto Input^2 * \boldsymbol{Weight}$$



(a) Origin mapping method (b) Previous mapping method (c) Our mapping method

Input-sensitive Cross-Array Mapping

$$\min(\max P - \min P)$$

$$s.t.\ every\ row\ (or\ column)\ in\ w_i\ must\ be\ in\ the\ same\ row\ (or\ column)\ of\ origin\ weight$$

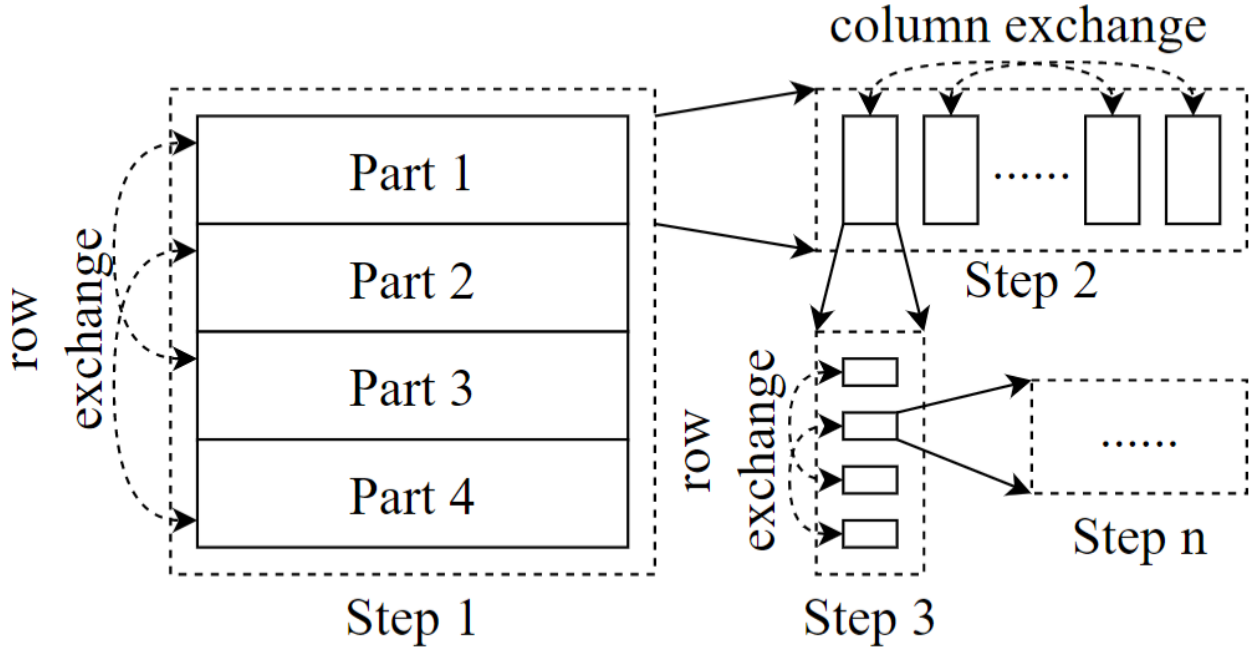
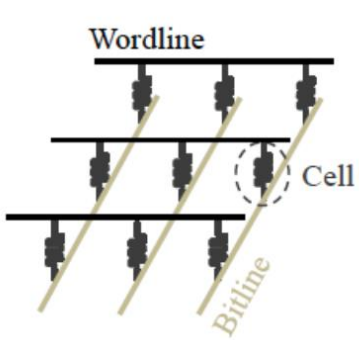
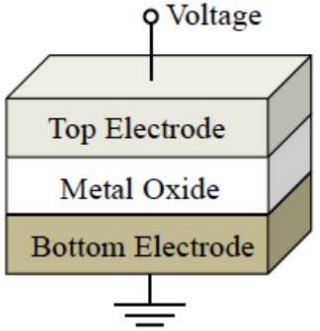


Illustration of proposed re-ordering method.

Outline

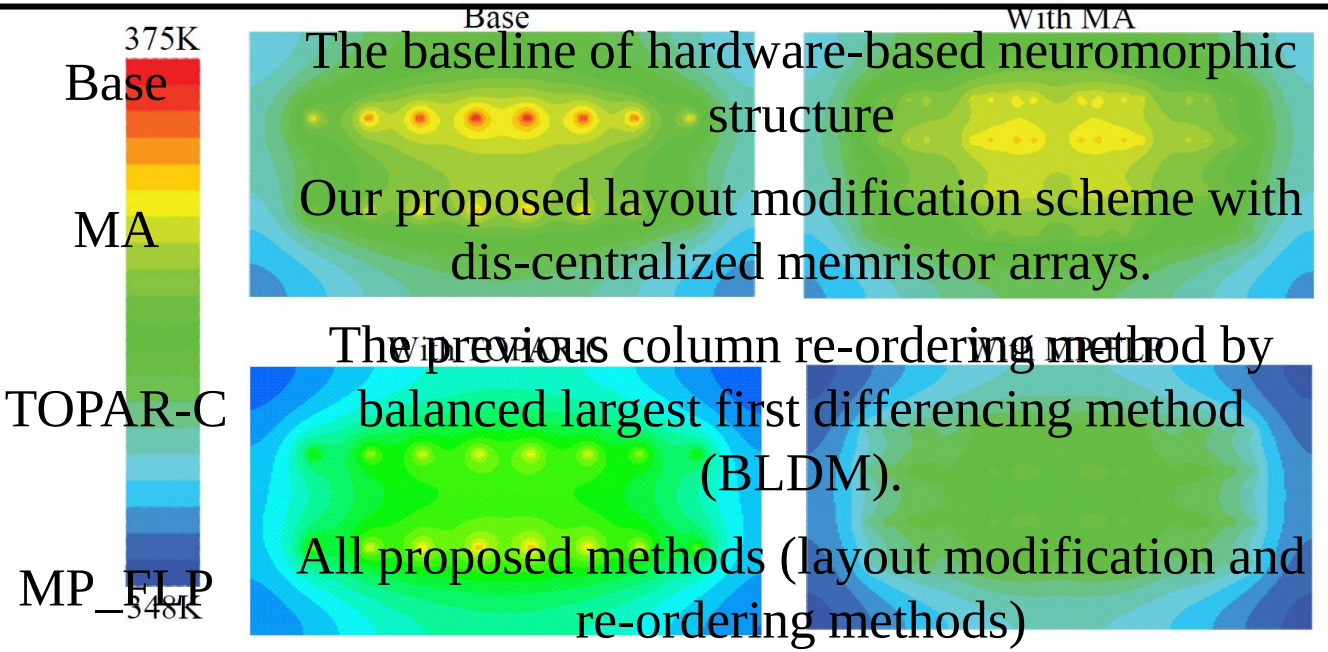




- Background
  - Spiking Neural Networks (SNNs)
  - Memristor Crossbar Array (MCA) based Accelerators
- Problem Formulation
- Proposed Techniques
  - Thermal-Aware Layout Design
  - Input-sensitive Cross-Array Mapping
- Experimental Results
- Summary

Experimental Results

Resistance range of memristor	Network	Crossbar size	VGG9 Resolution	VGG11 Input voltage	Working frequency
5kΩ ~ 500kΩ	Dataset	128x128	CIFAR-10 2 bits	CIFAR-100 0.9V	1.2 GHz



The baseline of hardware-based neuromorphic structure

Our proposed layout modification scheme with dis-centralized memristor arrays.

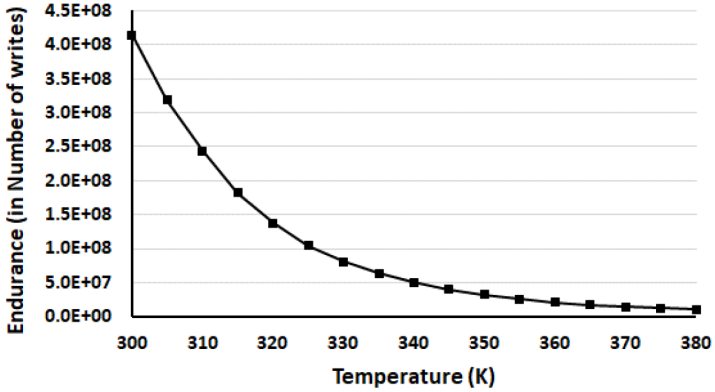
The previous column re-ordering method by balanced largest first differencing method (BLDM).

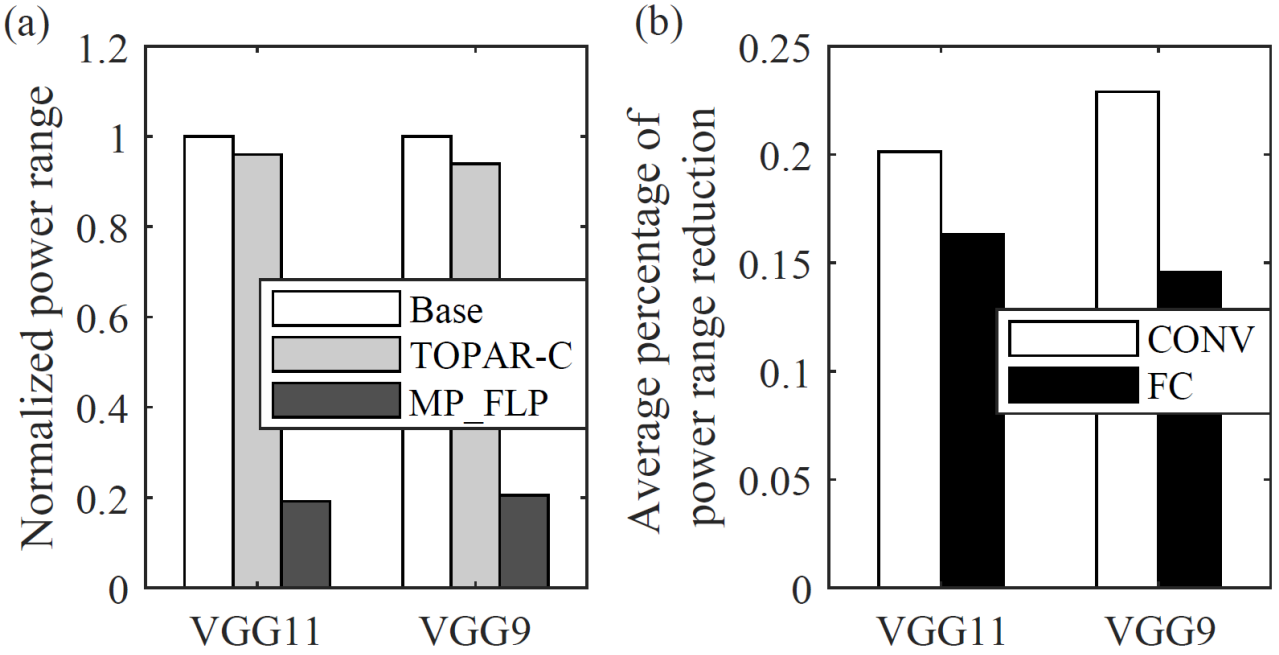
All proposed methods (layout modification and re-ordering methods)

The temperature distribution of the hottest tile in the third fully-connected layer of VGG9.

Experimental Results

- Peak temperature reduction: 10.4K
- Endurance improvement: 72%

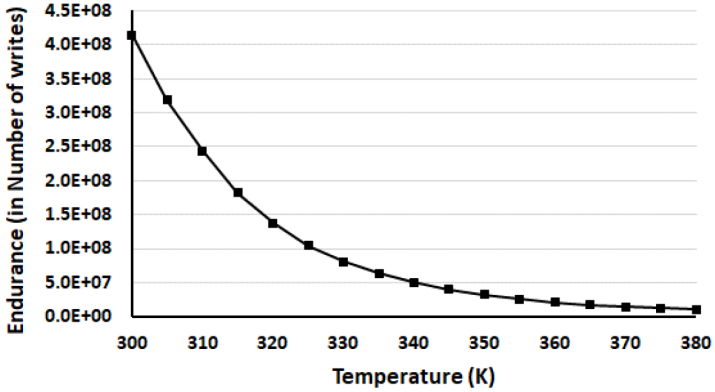


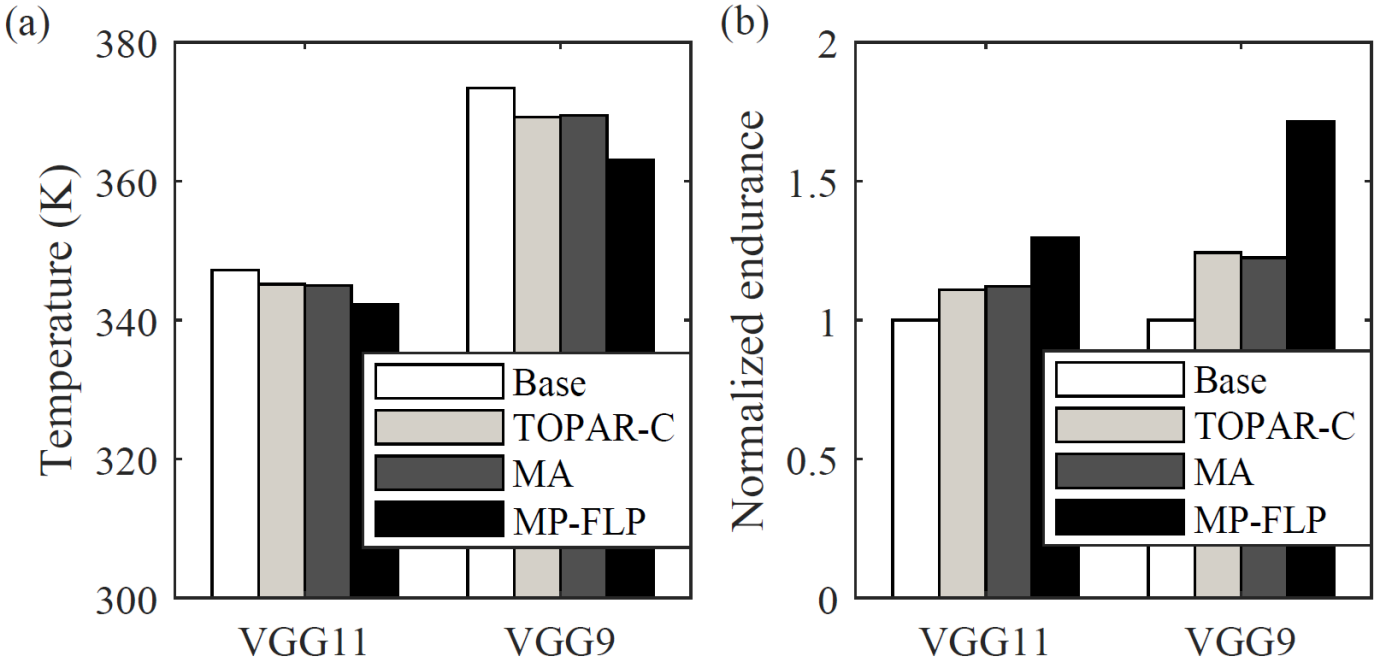


(a) The normalized power range. (b) The average percentage of power range reduction.

Experimental Results

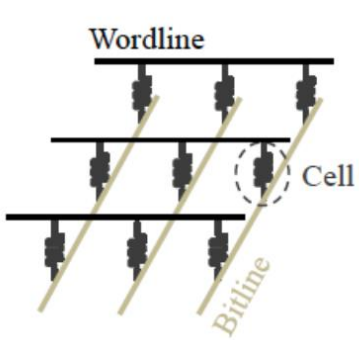
- Peak temperature reduction: 10.4K
- Endurance improvement: 72%

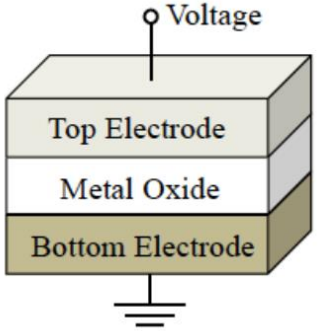




(a) Peak temperature. (b) Normalized endurance.

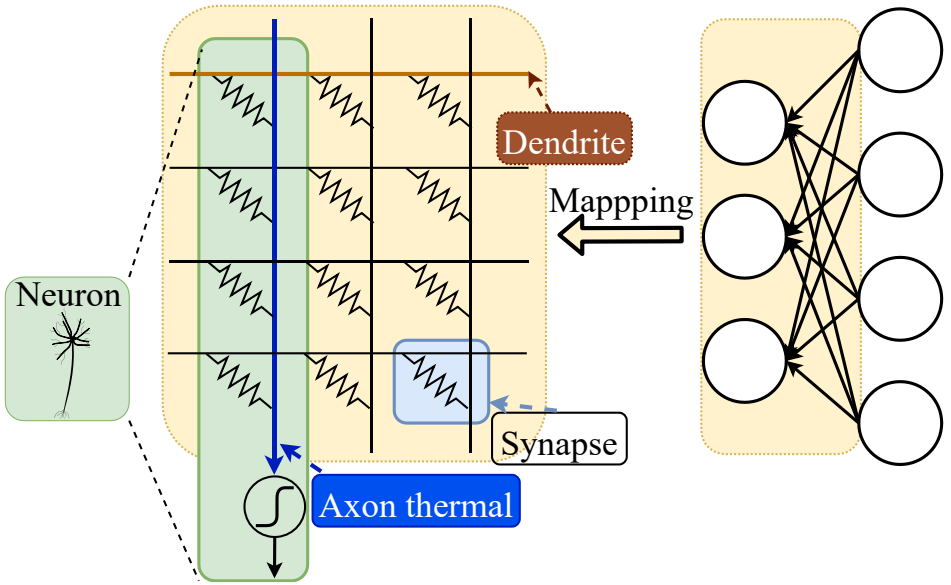
Outline

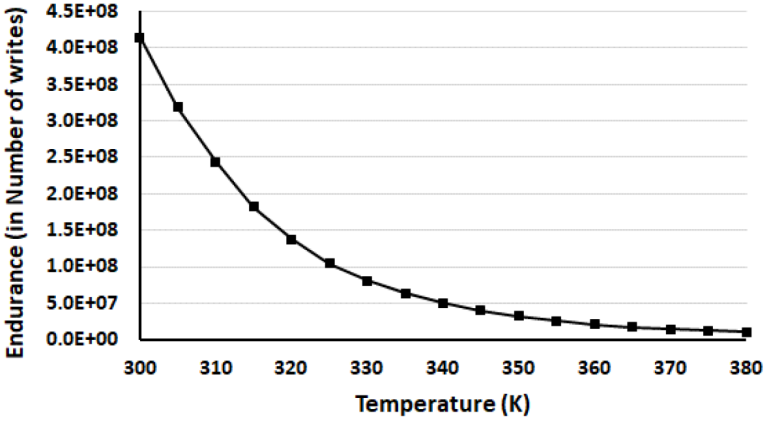


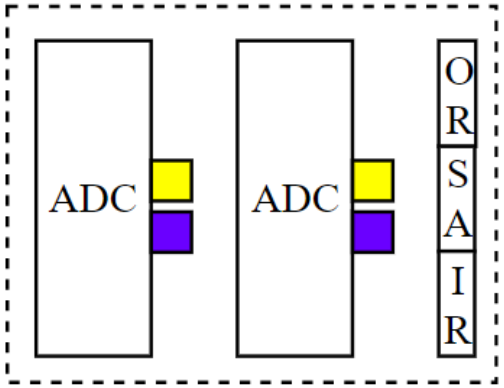


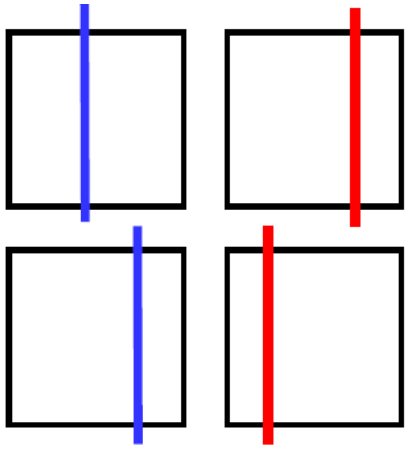
- Background
  - Spiking Neural Networks (SNNs)
  - Memristor Crossbar Array (MCA) based Accelerators
- Problem Formulation
- Proposed Techniques
  - Thermal-Aware Layout Design
  - Input-sensitive Cross-Array Mapping
- Experimental Results
- Summary

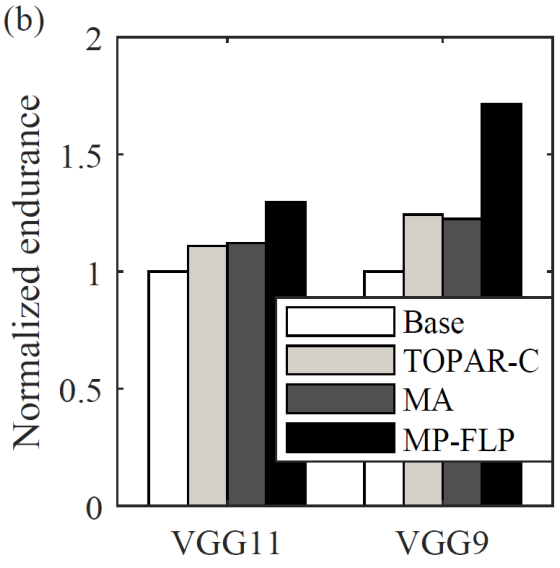
Summary


Dendrite
Mappping
Neuron
Synapse
Axon thermal


Endurance (in Number of writes)
4.5E+08
4.0E+08
3.5E+08
3.0E+08
2.5E+08
2.0E+08
1.5E+08
1.0E+08
5.0E+07
0.0E+00
300
310
320
330
340
350
360
370
380
Temperature (K)


ADC
ADC
OR
SA
IR




(b)
Normalized endurance
2
1.5
1
0.5
0
Base
TOPAR-C
MA
MP-FLP
VGG11
VGG9





Thanks!

zhangchr@shanghaitech.edu.cn