DATE 25
DESIGN, AUTOMATION AND TEST IN EUROPE
THE EUROPEAN EVENT FOR ELECTRONIC SYSTEM DESIGN & TEST
31 MARCH – 2 APRIL 2025
LYON, FRANCE
CENTRE DE CONGRÈS DE LYON



RoTA: Rotational Torus Accelerator for Wear Leveling of Neural Processing Elements

Taesoo Lim, Hyeonjin Kim, Jingu Park, Bogil Kim, and William J. Song

School of Electrical and Electronic Engineering
Yonsei University, Seoul, South Korea

COMPUTER ARCHITECTURE AND SYSTEMS LAB

Outline

- **Introduction**
- **Background and Motivation**
- **Methodology**
- **Evaluation Results**
- **Conclusion**

Introduction

- **Neural accelerators are increasingly incorporated into reliability-critical systems, such as automotive, aerospace, smart factories, etc.**
- **In this work, we focus on lifetime reliability issues.**



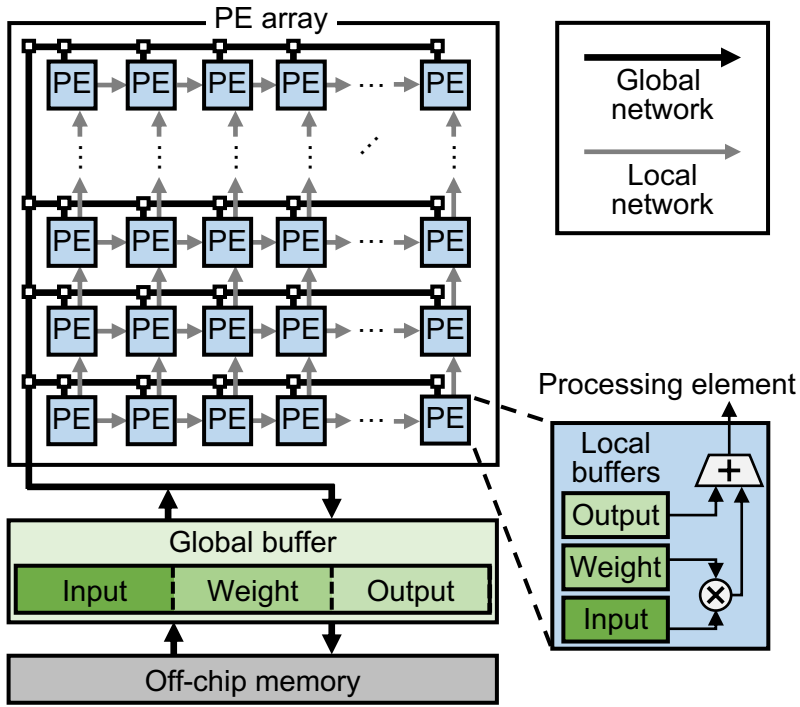


Images from electronicsera.in, linkedin.com, and wda-automotive.com

Background and Motivation

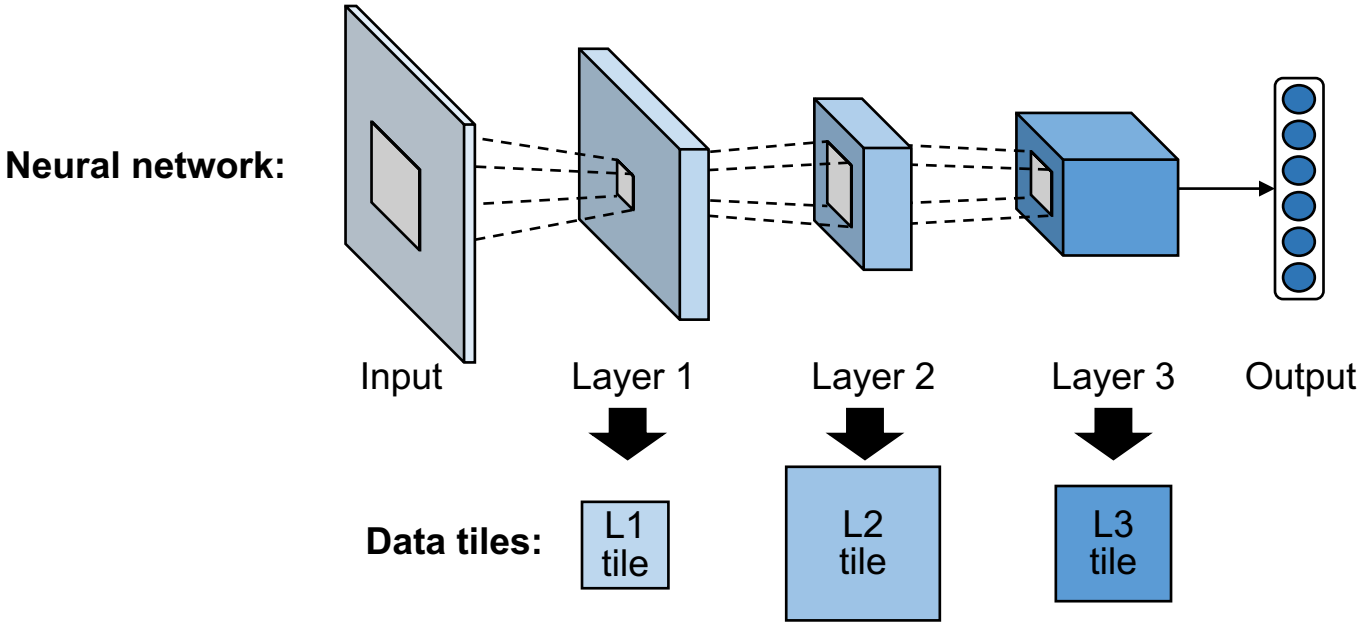
Background

- **Spatial neural accelerator architecture:**
  - **Processing elements (PEs):**
    2-D array of PEs for neural processing
  - **On-chip buffers:**
    Global & local buffers to store data tiles
  - **Interconnections:**
    Global & local networks to transmit data



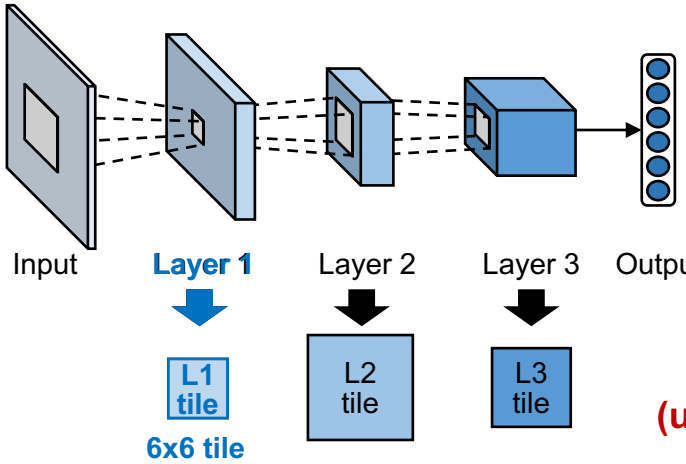
Motivation

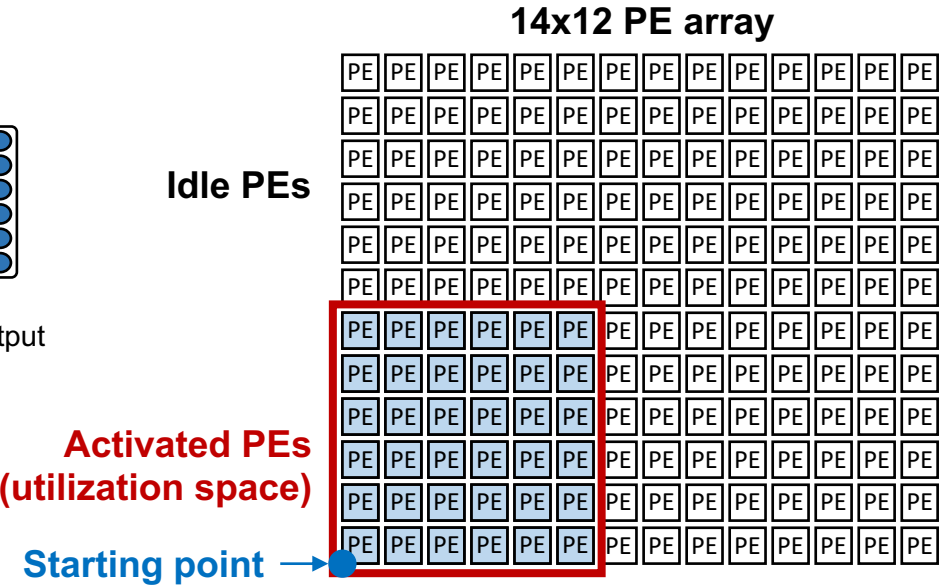
- **A neural network consists of a series of different-sized layers.**
- **Each layer is split into smaller data tiles fitting into on-chip buffers.**



Motivation

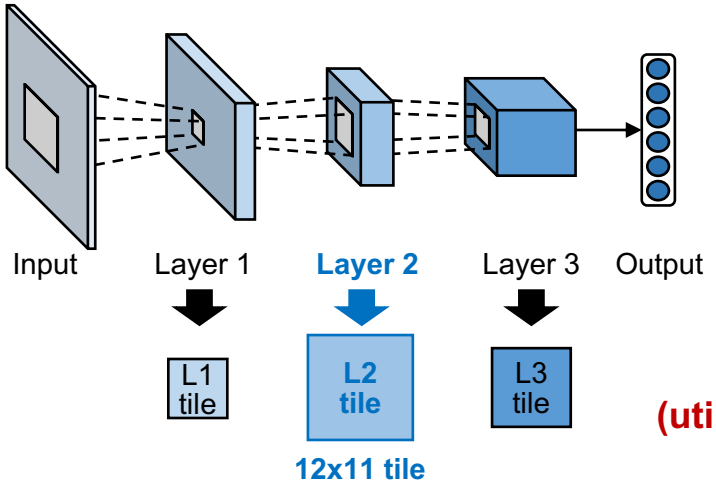
- **Processing a data tile starts at one corner of the 2-D PE array.**
- **However, different-sized data tiles have dimensional mismatches with the 2-D PE array.**

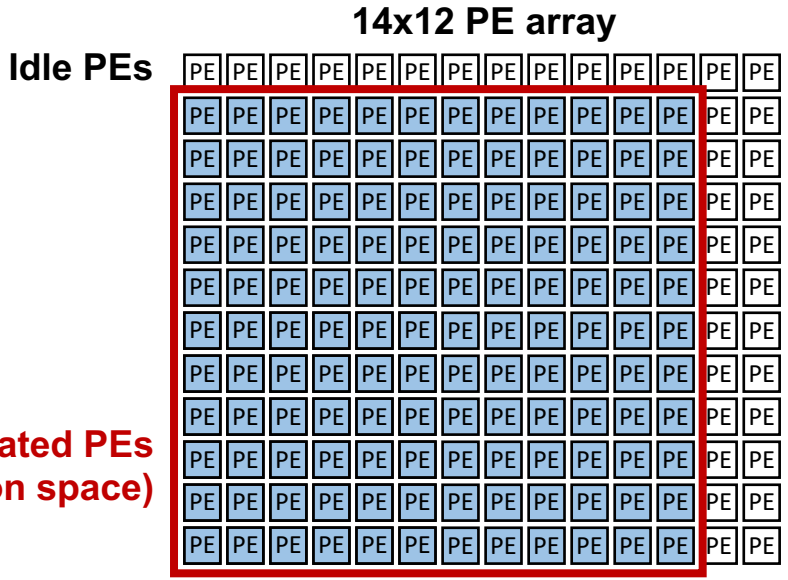




Motivation

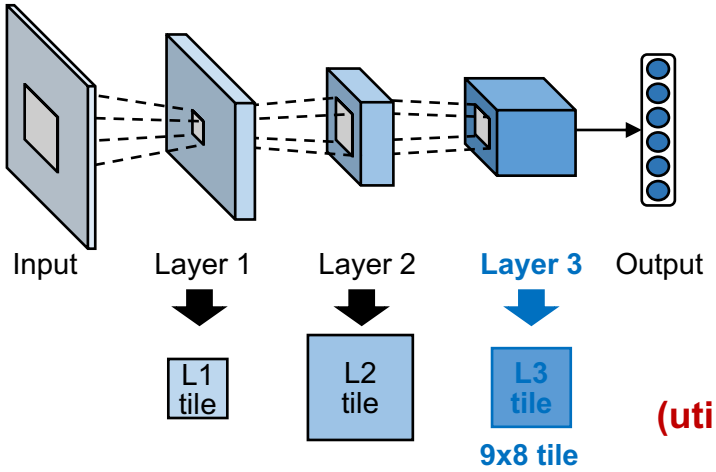
- **Processing a data tile starts at one corner of the 2-D PE array.**
- **However, different-sized data tiles have dimensional mismatches with the 2-D PE array.**

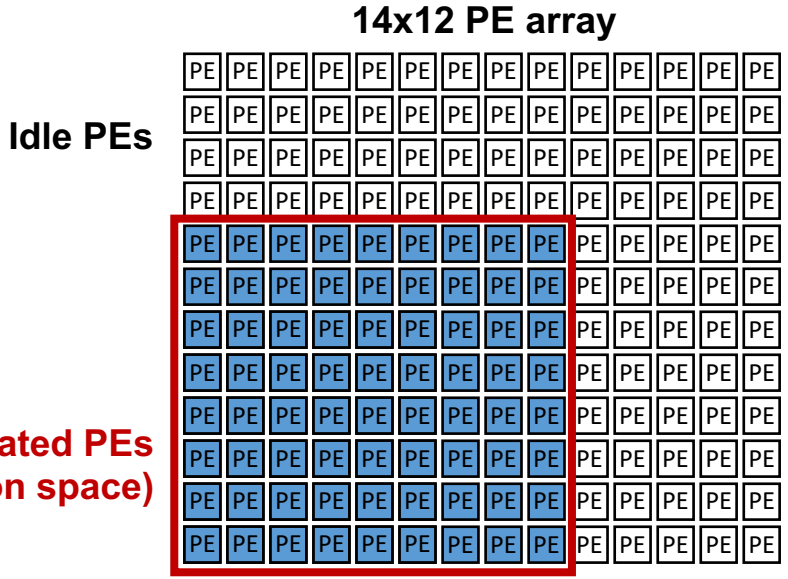




Motivation

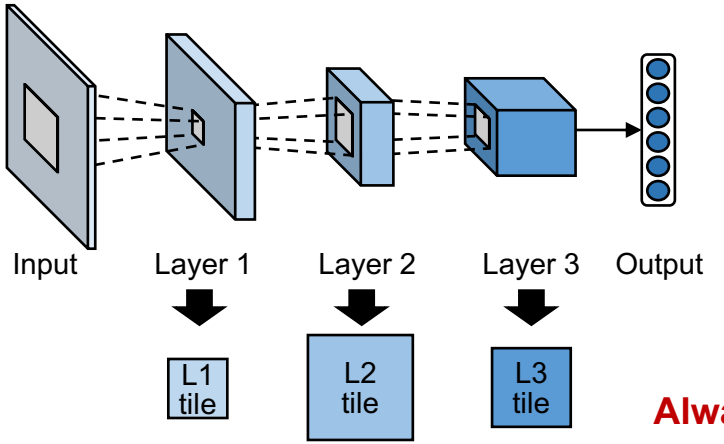
- **Processing a data tile starts at one corner of the 2-D PE array.**
- **However, different-sized data tiles have dimensional mismatches with the 2-D PE array.**

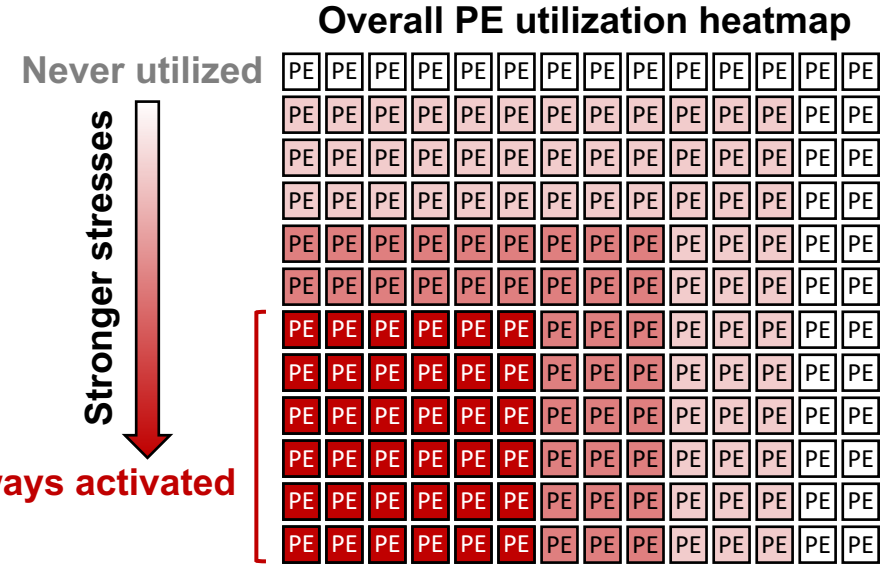




Motivation

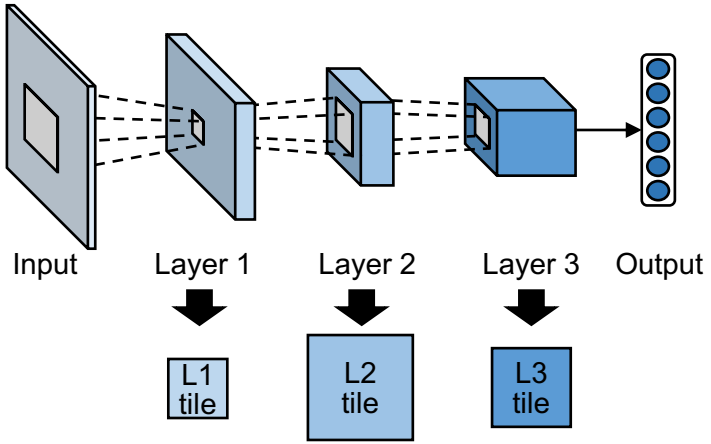
- **Processing a data tile starts at one corner of the 2-D PE array.**
- **However, different-sized data tiles have dimensional mismatches with the 2-D PE array.**

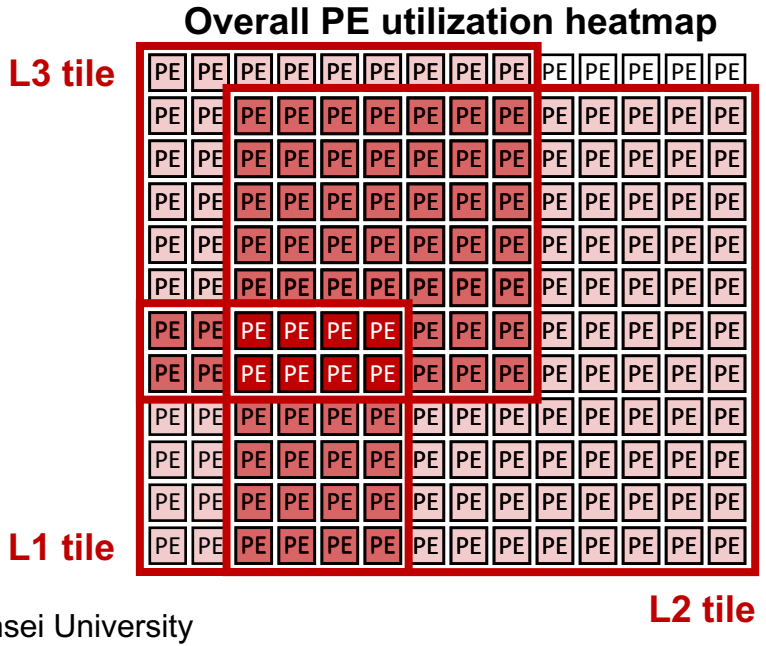




Motivation

- **Moving utilization spaces side-to-side on 2-D PE array only relocates the stress hotspot to the center and does not solve the problem.**





Lifetime Reliability Problems

- **Lifetime Reliability Problems in Neural Accelerators**
  1. **Different shapes of neural layers cause dimensional mismatches with the PE array.**

     ➡ **Optimized data mapping still results in dimensional mismatches.**

  2. **The fixed starting point of neural computations causes biased stresses depending on PE locations.**

     ➡ **We can "shuffle" the starting point of neural computations.**

Methodology

Lifetime Reliability Solutions

- **Two problems need to be addressed:**
  1. **How can we shuffle the starting point of neural computations?**
  2. **Does shuffling indeed resolve the lifetime reliability problems?**

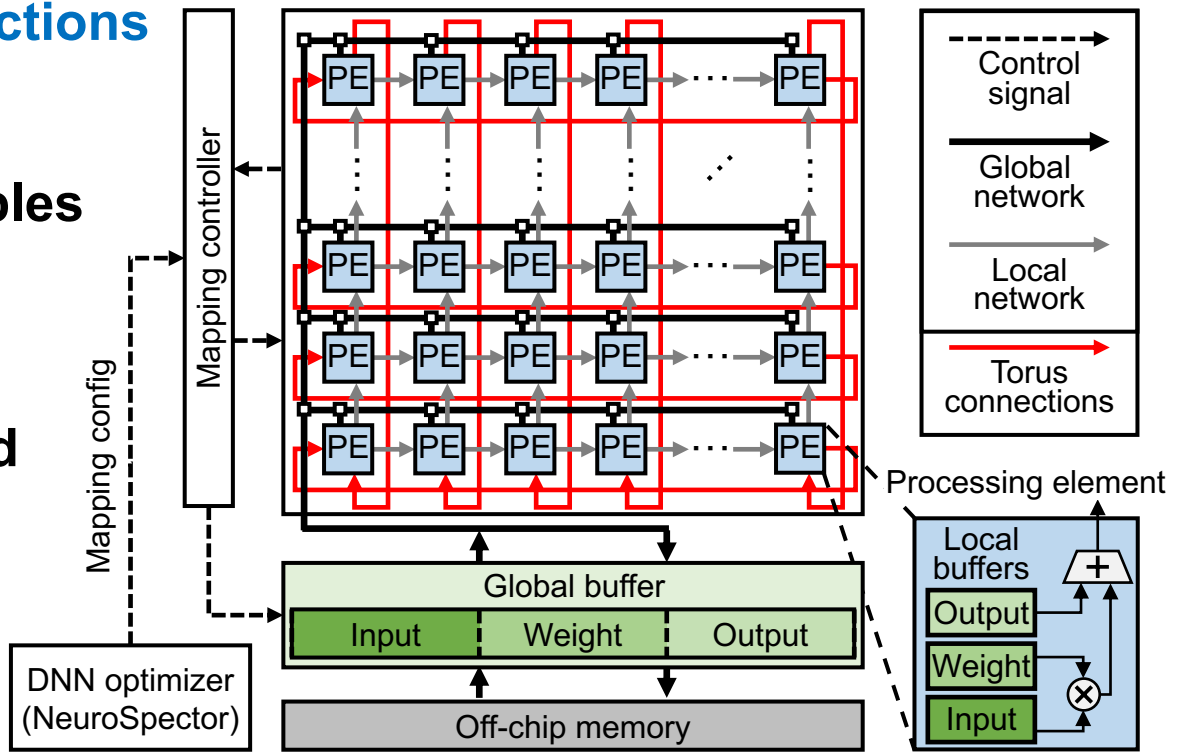
- **Our solutions:**
  1. **Rotational Torus Accelerator (RoTA)** ➡ **Accelerator design update**
  2. **Rotational Wear-Leveling (RWL)**
  3. **Residual Optimization (RO)**

  ⎫ ➡ **Wear-leveling schemes**

Rotational Torus Accelerator (RoTA)

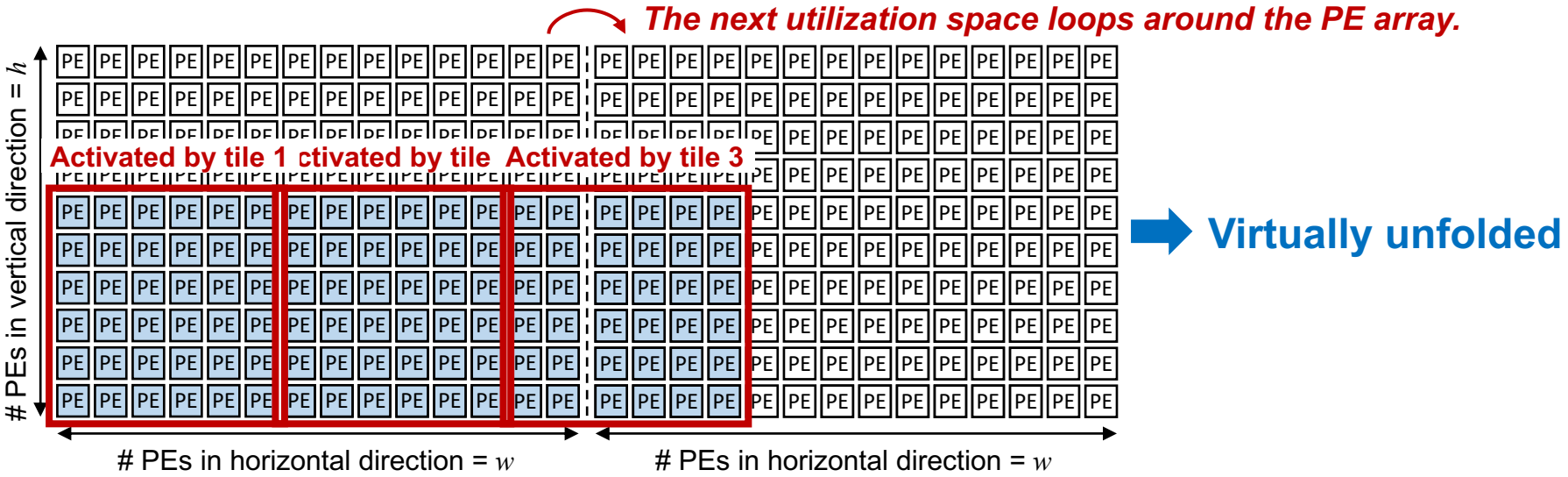
- **RoTA adds torus connections to the 2-D PE array.**
- **The torus PE array enables the utilization spaces to circulate.**
- **The torus can be formed without long wires by interconnecting every other PE.**

➡ **Small HW overhead**



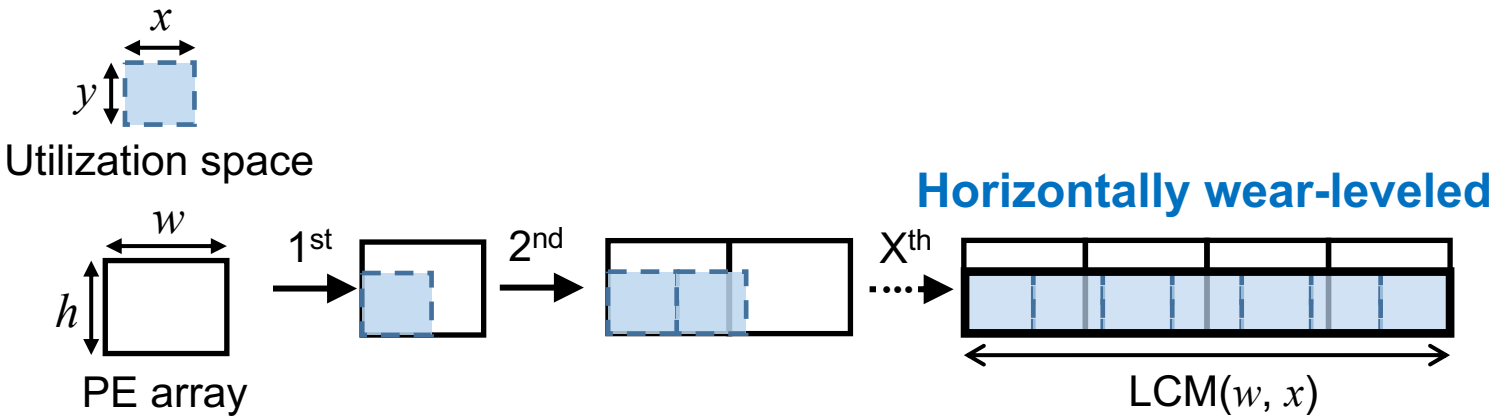
Rotational Wear Leveling (RWL)

- **A neural layer is split into many small tiles and repeatedly processed.**
- **The torus PE array can be virtually unfolded horizontally and vertically as if it had infinite connections.**



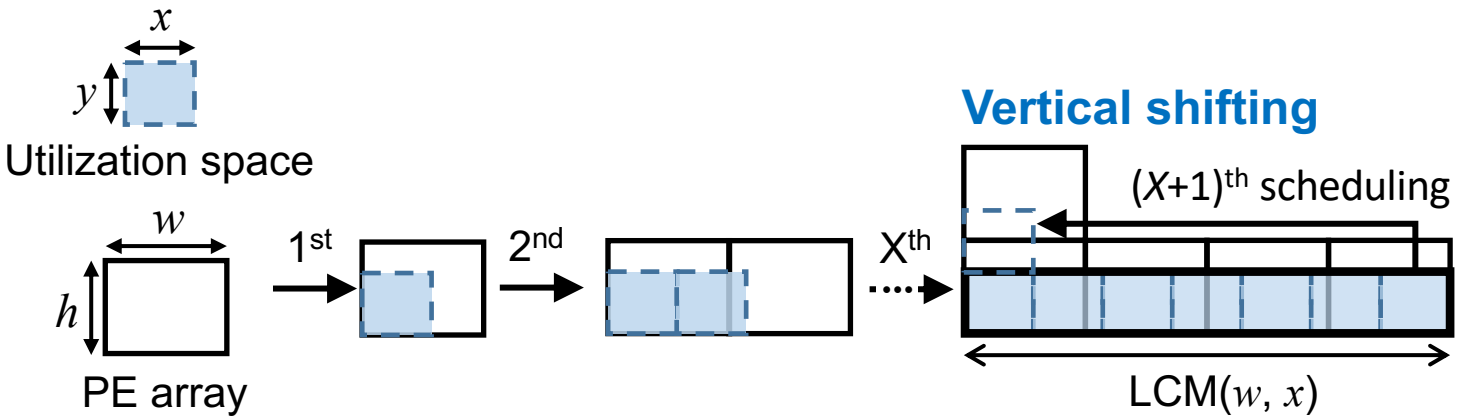
Rotational Wear Leveling (RWL)

- **The torus PE array is horizontally unfolded, and the utilization spaces keep shifting until their least common multiple (LCM) is met.**
- **Then, the row of horizontally unfolded PE arrays becomes perfectly wear-leveled.**



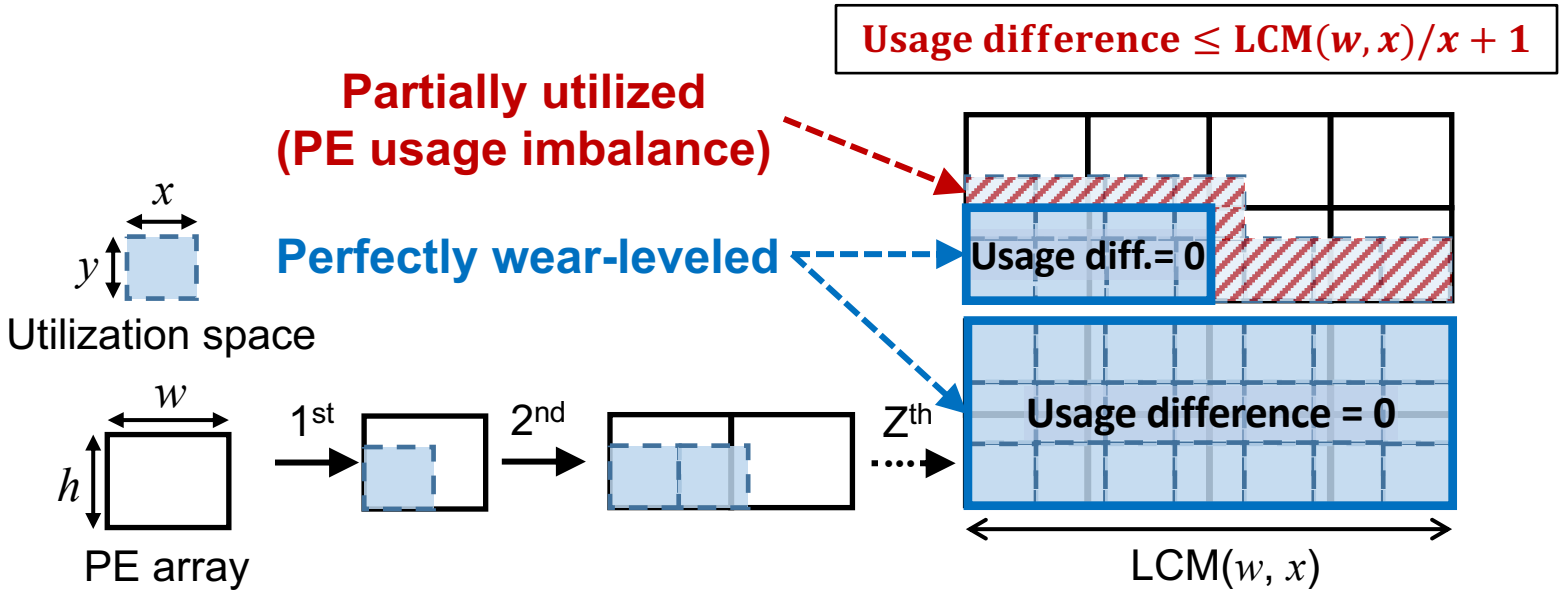
Rotational Wear Leveling (RWL)

- **The next utilization space shifts vertically, and then the same process is repeated.**



Rotational Wear Leveling (RWL)

- **At the end of a neural layer, virtually unfolded PE arrays are either i) perfectly wear-leveled or ii) partially utilized.**

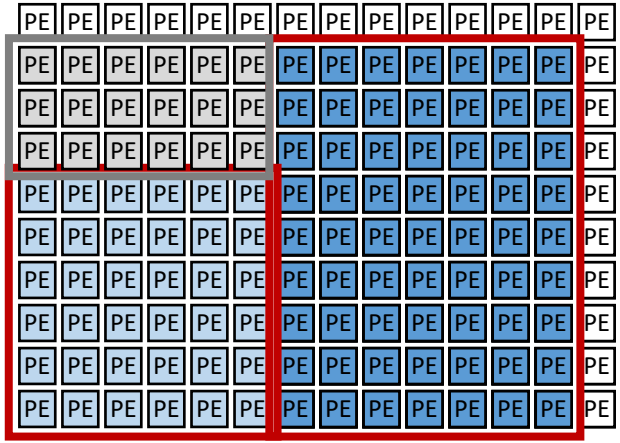


Residual Optimization (RO)

- When a neural layer ends without perfect wear leveling, the next neural layer resumes from the endpoint of the previous layer.
- This makes a few PEs duplicately used or unused, but in the long run, such imbalance is unbiasedly dispersed across the PE array.

These PEs are unused between layers, but OK!

The "last" utilization space of a neural layer ends with imperfect wear leveling.



The "first" utilization space of the next layer resumes from the previous endpoint.

Evaluation Results

Experimental Method

- **We used Weibull distribution to model the lifetime reliability of PEs.**

$$R_{\text{Weibull}}(t) = 1 - F_{\text{Weibull}}(t) = e^{-(\frac{t}{\eta})^{\beta}}$$

- **Mathematically, the mean-time-to-failure (MTTF) of the PE array is calculated as follows** *(refer to the paper for details)*:

$$\text{MTTF}_{\text{Array}} = \frac{\eta}{(\sum_{i,j}(\alpha_{i,j})^{\beta})^{\frac{1}{\beta}}}\Gamma(1+\frac{1}{\beta})$$

Experimental Method

- **We composed a simulator to track the usage count of PEs.**

Accelerator model

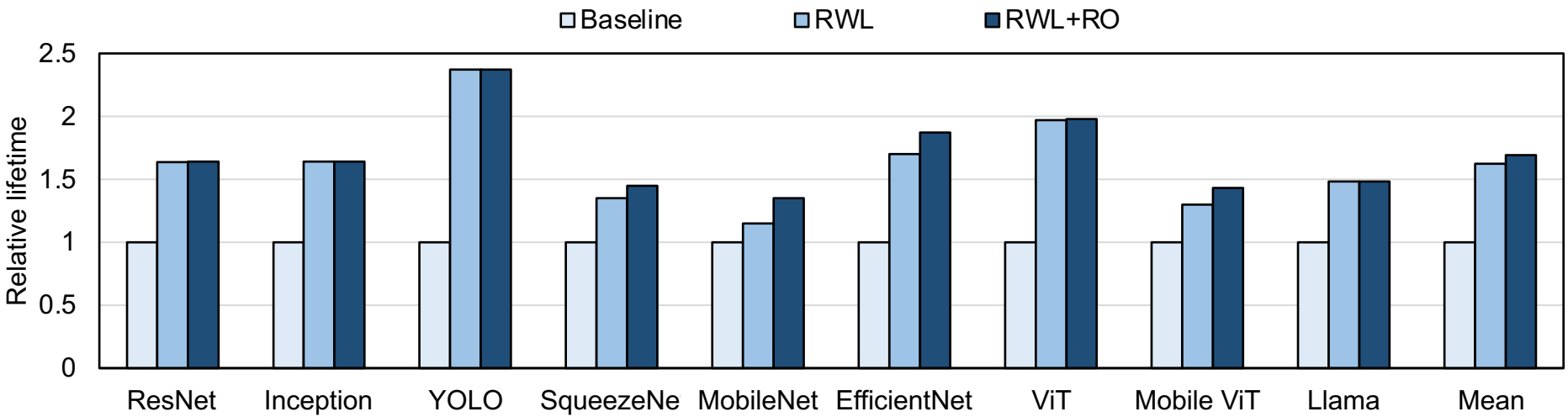
Configurations	Parameters
PE array size	14x12
Local buffers (input, weight, output)	24, 448, 48 bytes per PE
Global buffer	108 KB
Mapping optimization	NeuroSpector

Neural network models

DNN domains	Networks
Image classification	ResNet-50
	Inception v4
Object detection	YOLO v3
Lightweight network	SqueezeNet
	MobileNet v3
	Efficient Net
Transformer	ViT
	Mobile ViT
	Llama v2

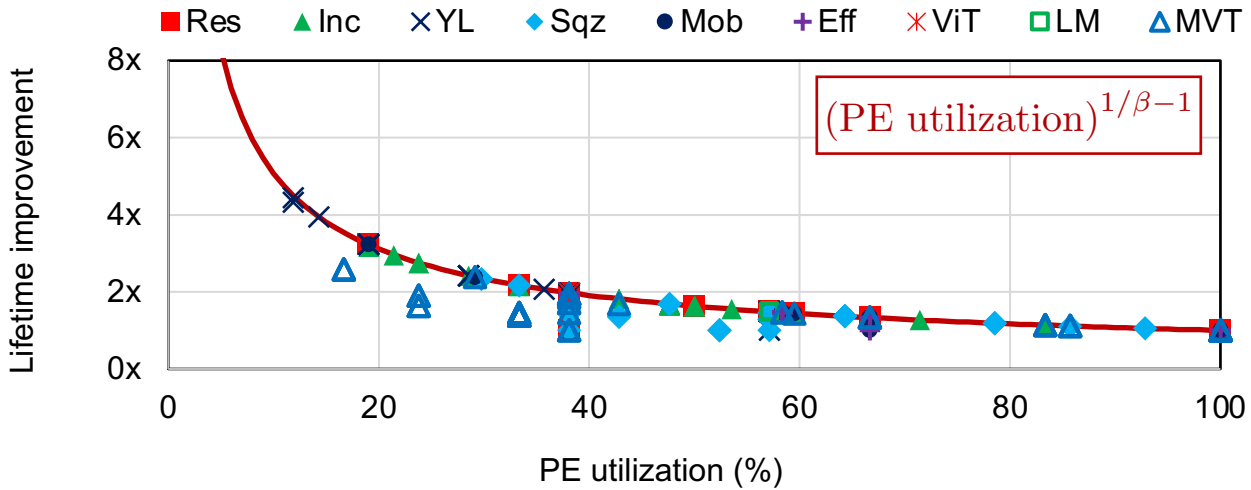
Evaluation Results

- **RoTA with RWL+RO schemes achieves a 1.69x lifetime reliability improvement on average.**



Experimental Results

- **The theoretically achievable lifetime reliability improvement is inversely proportional to the PE utilization ratio** *(refer to the paper)*.
- **The RWL scheme achieves near-max improvements.**



Conclusion

- **Neural accelerators are increasingly incorporated into reliability-critical systems, but their lifetime reliability has been overlooked.**
- **Lifetime reliability problems are due to i) dimensional mismatches between the PE array and DNN layers and ii) fixed utilization spaces.**

1. **How can we shuffle the starting point of neural computations?**

➡ **RoTA uses "torus" connections to enable RWL+RO wear-leveling schemes.**

2. **Does shuffling indeed resolve the lifetime reliability problems?**

➡ **RoTA with RWL+RO achieves "near-max" lifetime reliability improvements.**

Thank you!