On Endurance of Processing in (Non-Volatile) Memory

Salonik Resch, Husrev Cilasun, Zamshed Chowdhury, Masoud Zabihi, Zhengyang Zhao, Jian-Ping Wang, Sachin Sapatnekar, Ulya R. Karpuzcu

University of Minnesota - Twin Cities
Department of Electrical and Computer Engineering

International Symposium on Computer Architecture (ISCA) June 18, 2023

Background

- The *memory wall* limits performance and energy-efficiency
- Processing-in-Memory (PIM) comes to rescue

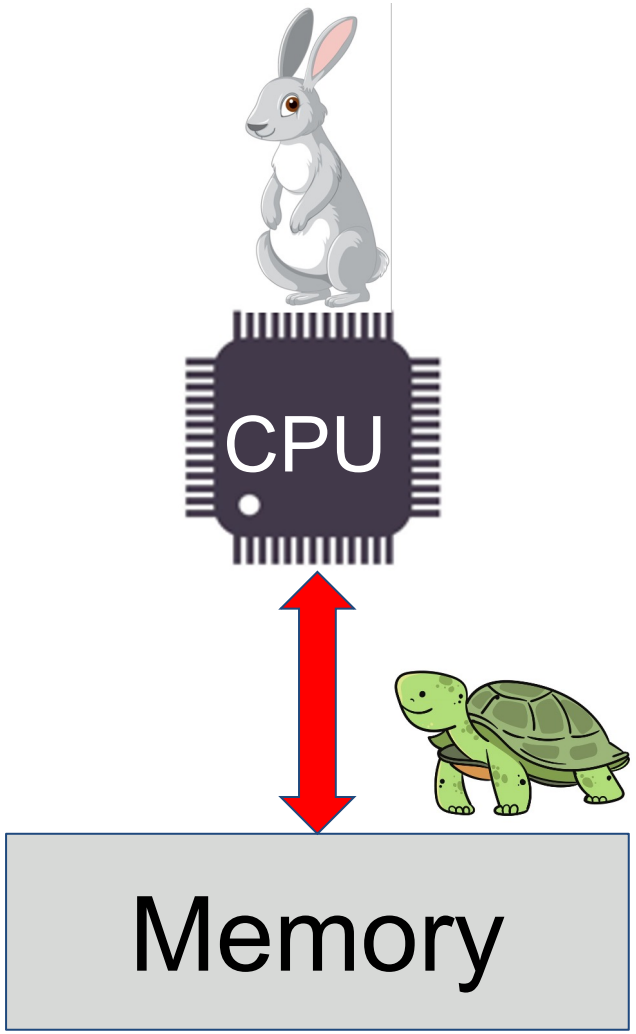
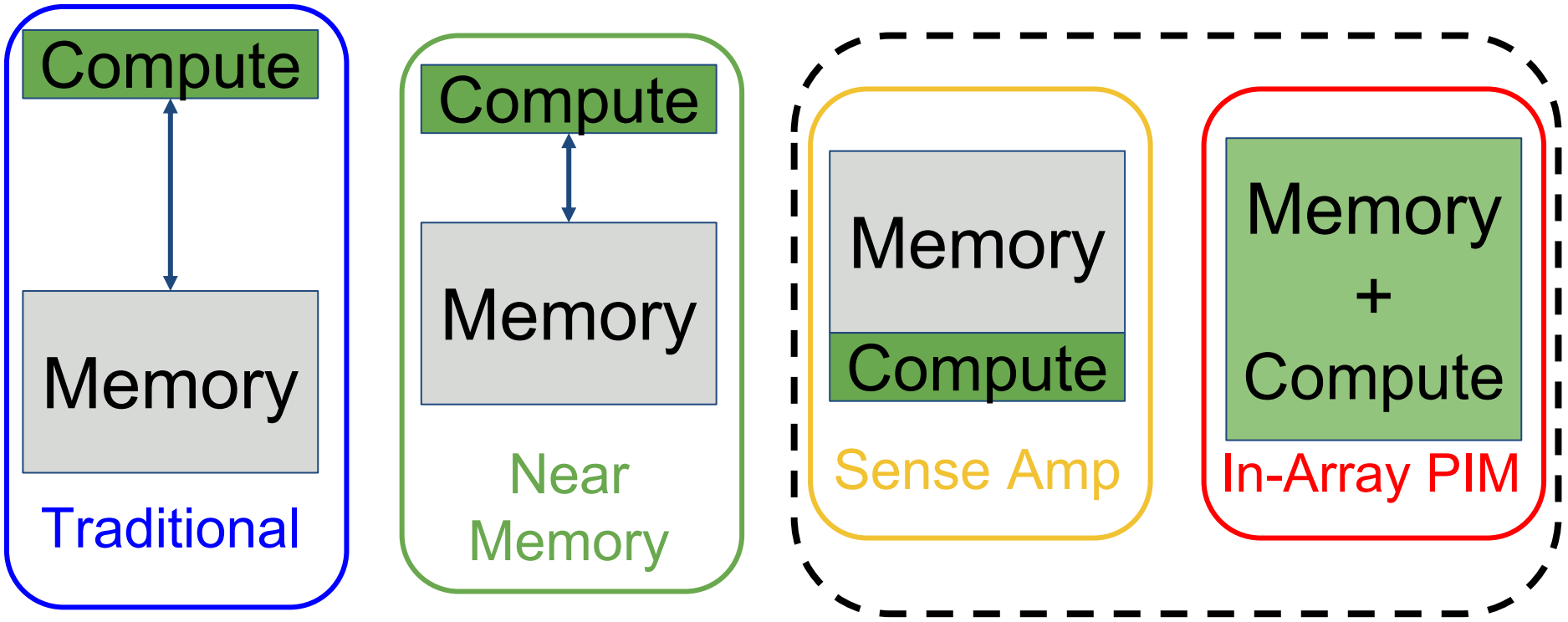
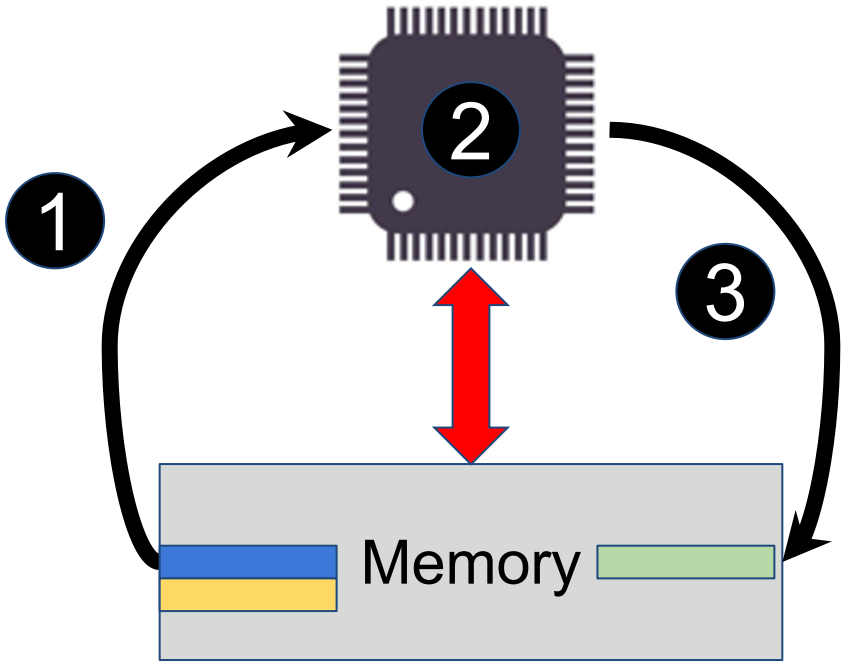


Image by brgfx on Freepik

PIM Taxonomy

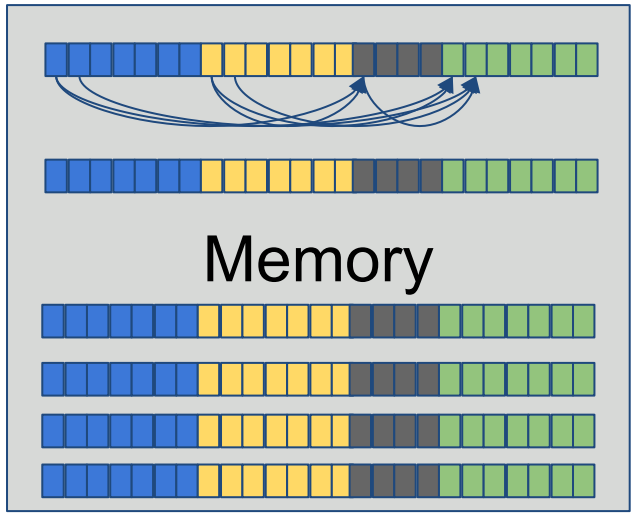


Traditional



1. Read
2. Compute
3. Write back

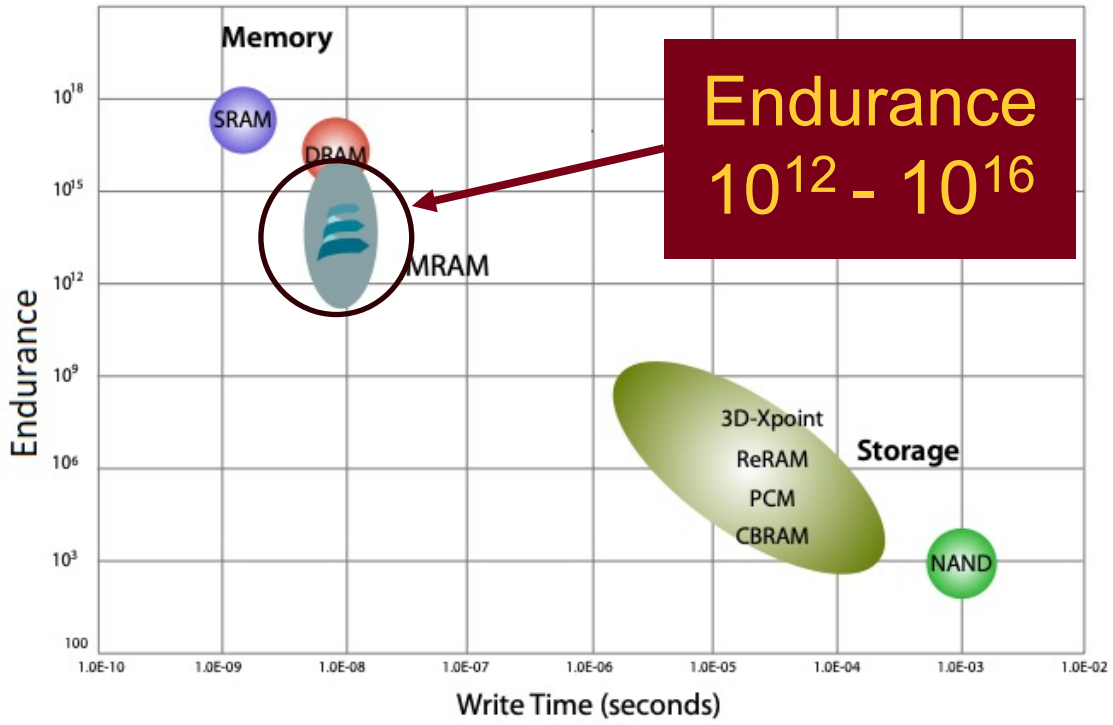
PIM



Perform sequence of logic gates

Non-Volatile PIM

- Non-volatile PIM is energy-efficient
  - Resistive RAM (RRAM)
  - Magnetic RAM (MRAM)
- Non-volatile devices fail after too many writes



Everspin

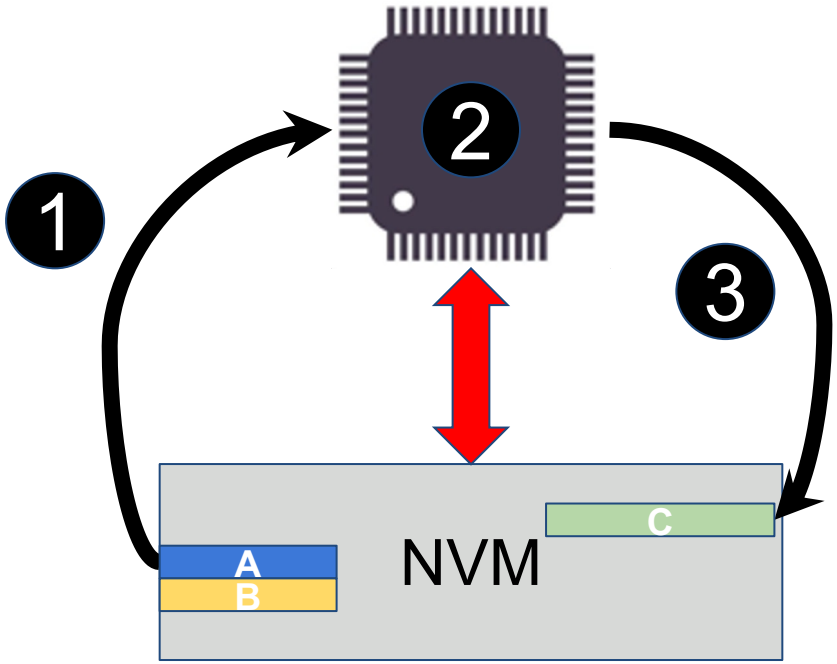
Endurance Challenge

- Under finite endurance we must be very careful with:
  - Write count
  - Load Imbalance
- Strategies for non-volatile memories (NVM) include:
  - Write cancellation (avoid unnecessary writes)
  - Load-Balancing
- PIM poses new challenges
  1. Greatly increases write count
     - Makes write cancellation impossible
  2. Restricts load-balancing

Write Count Increase

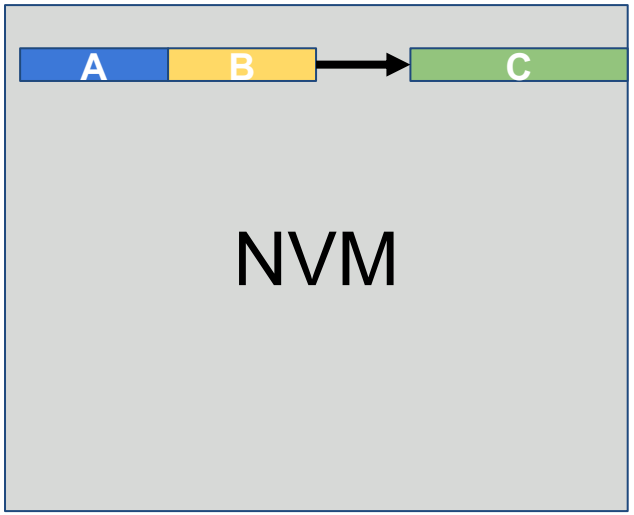
Example:
A (32-bit) x B (32-bit) = C (64-bit)

Traditional



1. Read (64-reads)
2. Compute (0 writes)
3. Write back (64-writes)

PIM



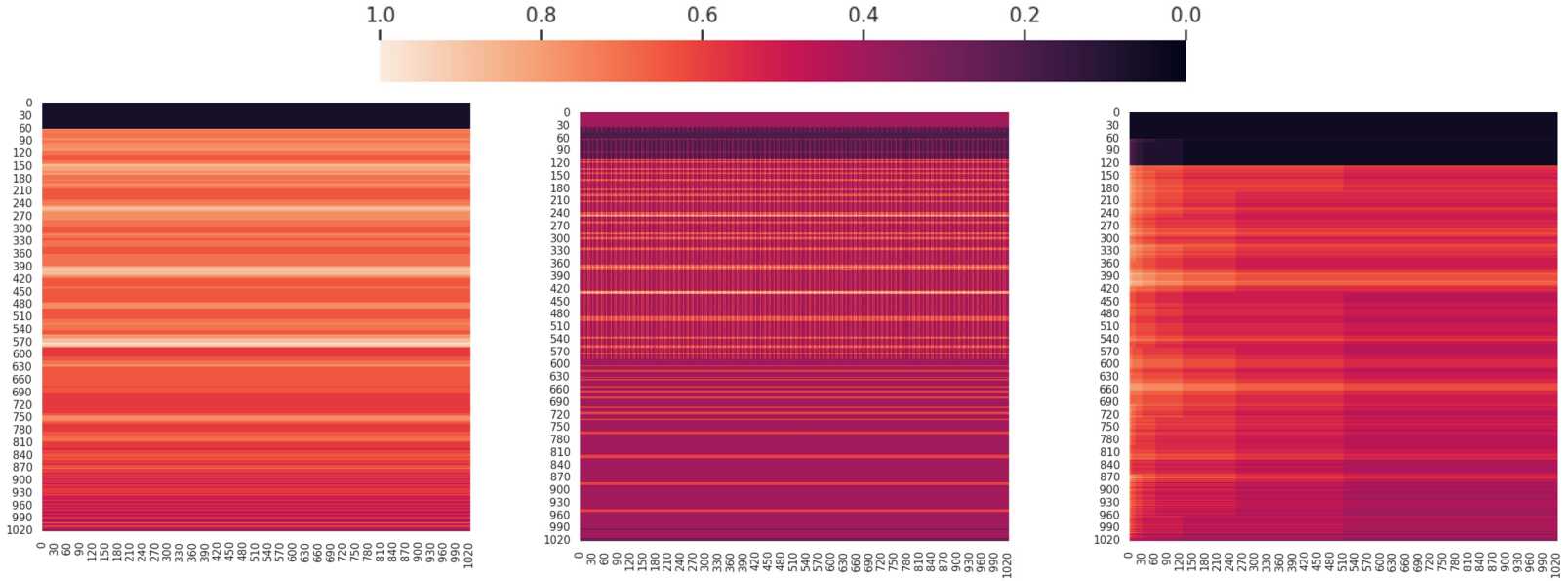
Compute (~10,000 writes)

> 150x increase

Writes cannot be cancelled

Load-Imbalance

Some cells are used more than others

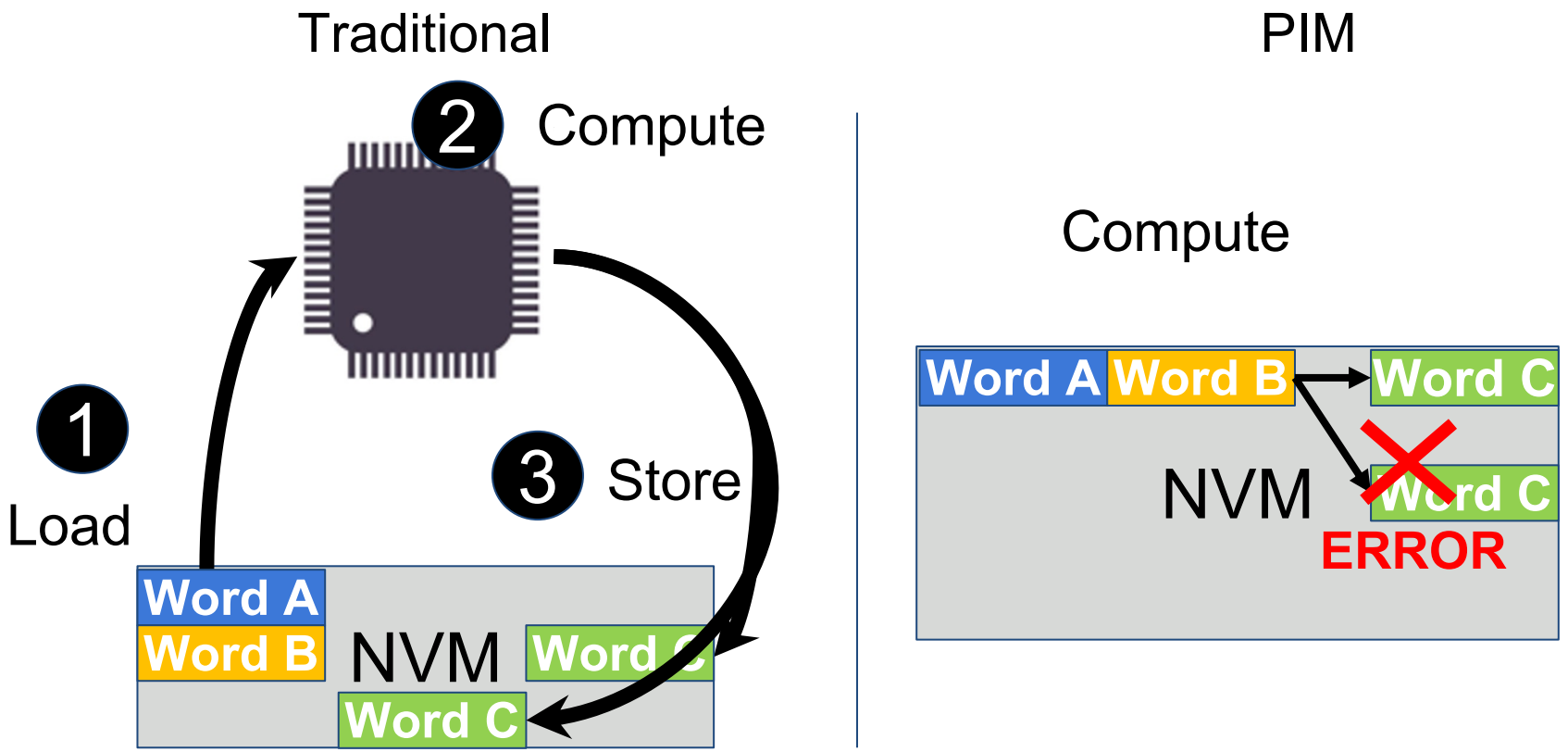


Multiplication

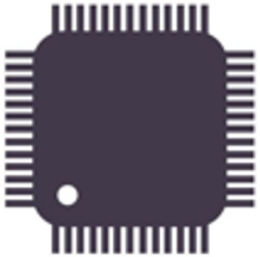
Convolution

Dot-Product

Load Balancing



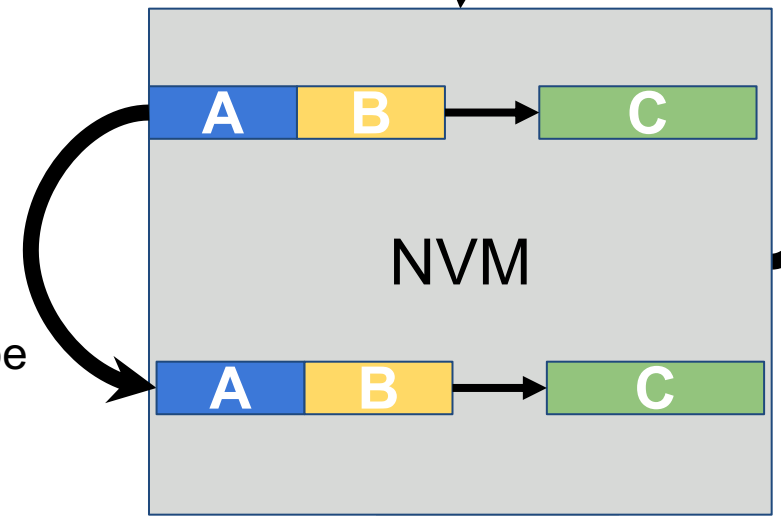
What can be done for PIM?



Option A

Statically re-map (infrequently)

- High cost
- Can be amortized



Option B

Dynamically Swap Rows

- Entire row must be swapped

Lifetime Limit Study

Benchmark	Maximal Lifetime Improvement
Multiplication	1.5x
Vector-Dot Product	1.5x
Convolution	2.8x

Software and architectural optimization alone is not sufficient

Lifetime Limit Study

Benchmark	Lifetime Endurance = 10^{12} (Pessimistic)	Lifetime Endurance = 10^{16} (Optimistic)
Multiplication	22 days	602 years
Dot Product	26 days	712 years
Convolution	19 days	520 years

Lifetime critically depends on device-level endurance

Conclusion

- Endurance limitation comes with all NV architectures
- NV PIM increases endurance requirements
- Projected device improvements can enable sufficiently long lifetimes

On Endurance of Processing in (Non-Volatile) Memory

Salonik Resch, Husrev Cilasun, Zamshed Chowdhury, Masoud Zabihi, Zhengyang Zhao, Jian-Ping Wang, Sachin Sapatnekar, Ulya R. Karpuzcu

University of Minnesota - Twin Cities
Department of Electrical and Computer Engineering

International Symposium on Computer Architecture (ISCA) June 18, 2023