The 20th U.S.–Korea Nanotechnology Forum

Beyond Intelligent Computing

Dataflow-Aware Memory-Centric Computing: From Synaptic Devices to Practical AI Systems

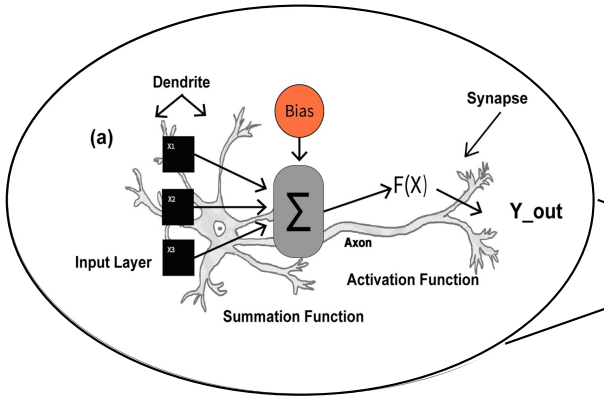
Bokyung Kim

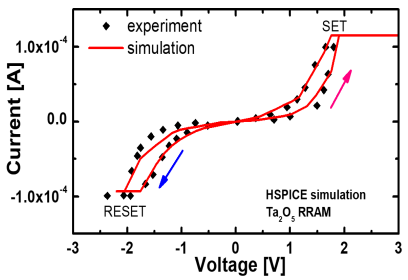
Assistant Professor

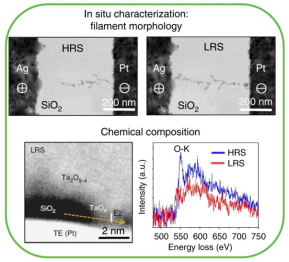
Dept. of Electrical and Computer Engineering

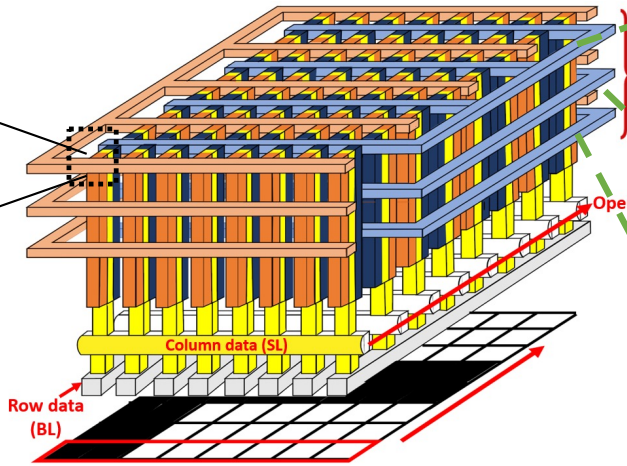
North Carolina State University

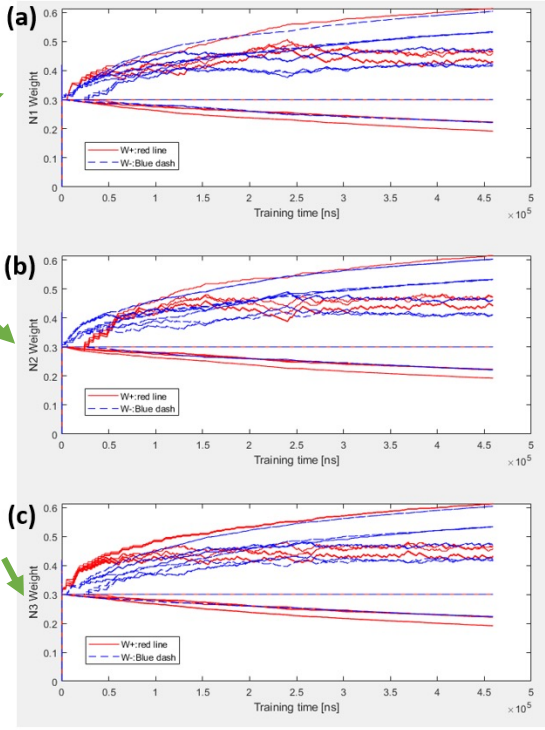
Neuromorphic Computing











1536F² → 256F² (6X reduced area)

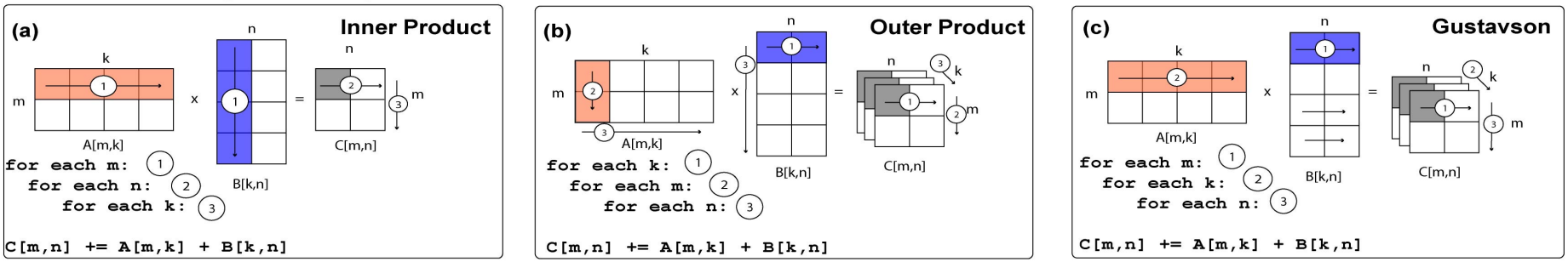
High Nonlinearity: Suppressing Sneak-Path Currents

B. Kim et al., ISVLSI, 2020

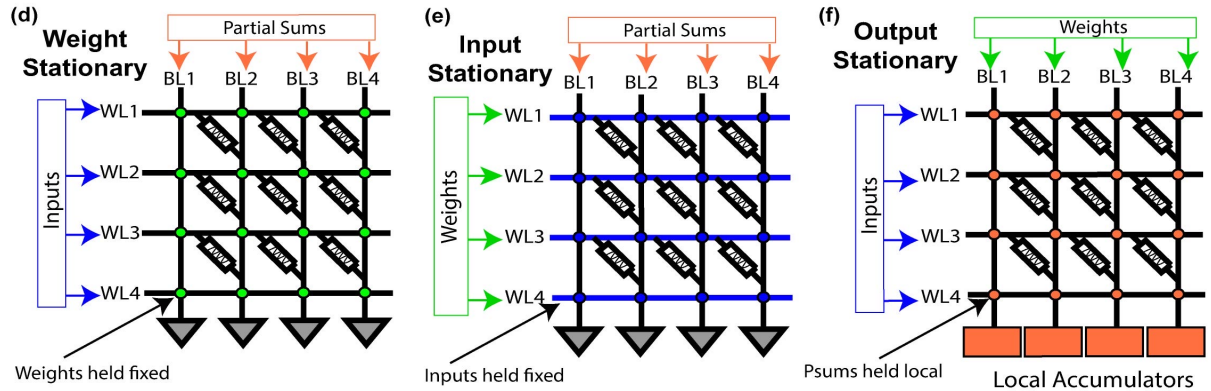
1 SYNAPTIC FUNCTION 2 VERTICAL ORGANIZATION 3 DATAFLOW REALIZATION

What Makes Dataflow?

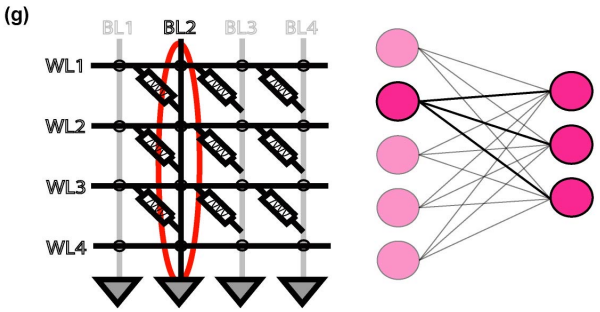
Dataflow Strategies: Computational Order



Dataflow Strategies: Mapping



ReRam crossbar array - Neural network application

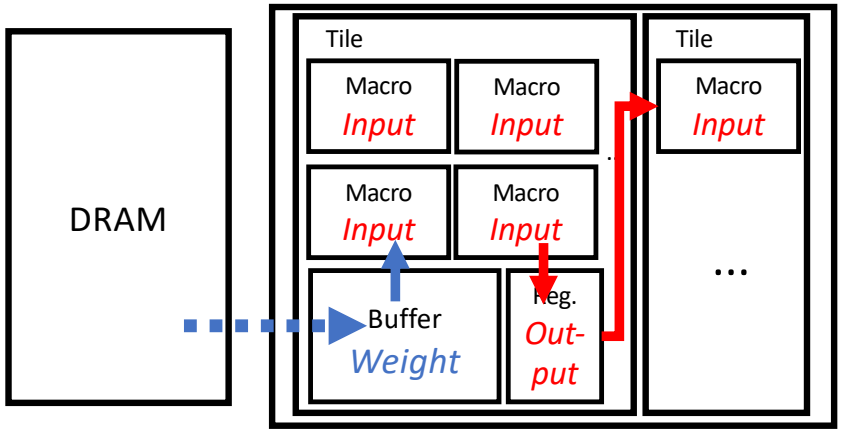


B. Kim et al., ICM, 2025

Why Not *Input-Stationary*?

- Inputs: Static / Weights: Dynamic
- Reducing Data Movement

  → High Efficiency & Scalability



Input-Stationary

WS Limit. 1. Excessive Memory Access

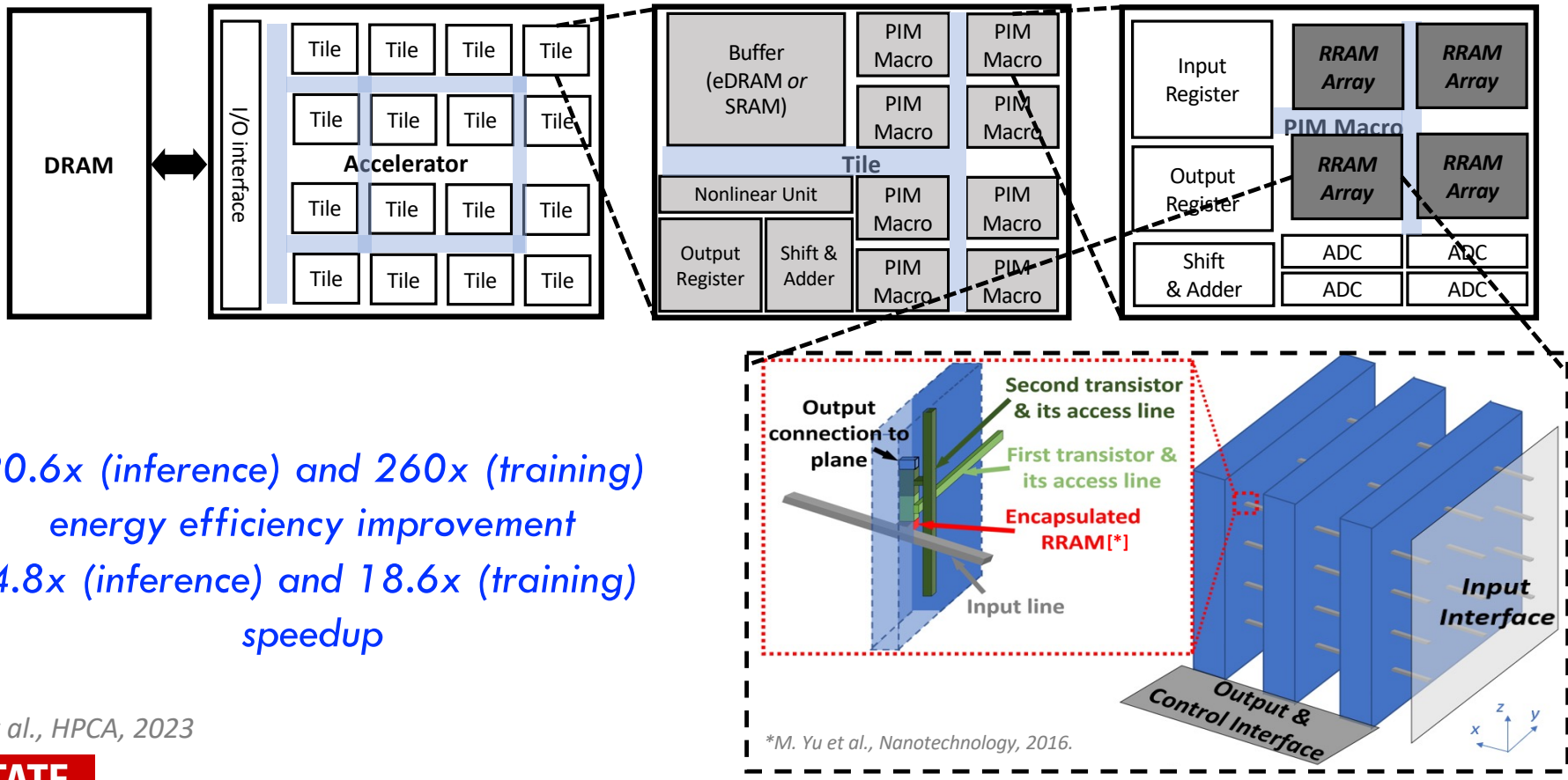
WS Limit. 2. Numerous RRAMs Consumption

WS Limit. 3. Necessity of Coarse-Grained Arrays

WS Limit. 4. RRAM Nonideality's Impact on Accuracy

B. Kim et al., HPCA, 2023

INCA & Key Results

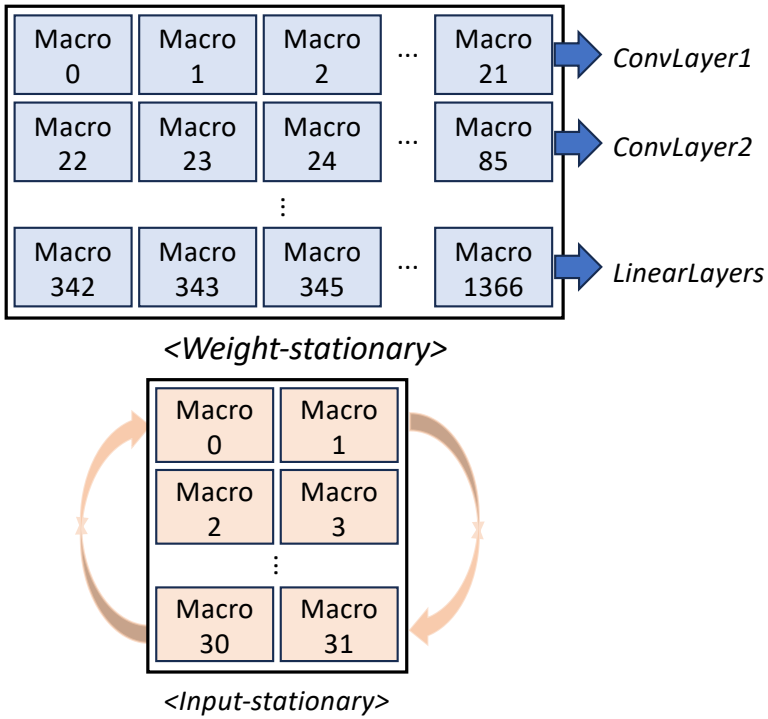


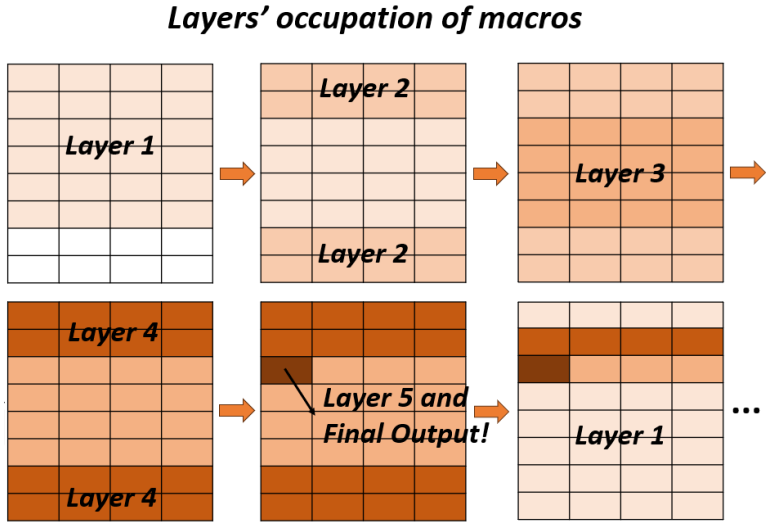
20.6x (inference) and 260x (training) energy efficiency improvement

4.8x (inference) and 18.6x (training) speedup

B. Kim et al., HPCA, 2023

Area Efficiency of Input-Stationary





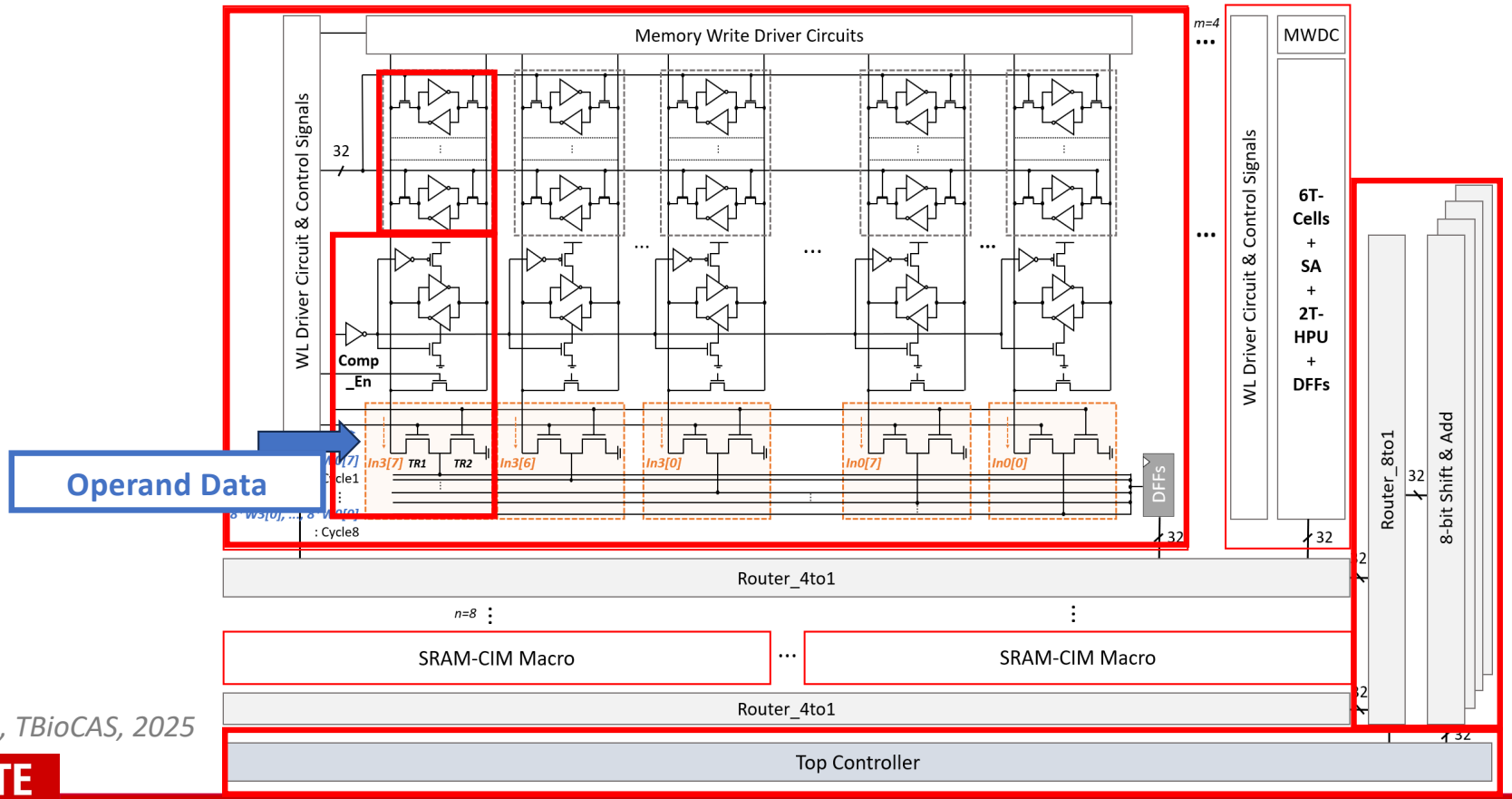
42.72X SRAM Capicity Saving

(WS vs. IS with recycling)

B. Kim et al., TBioCAS, 2025

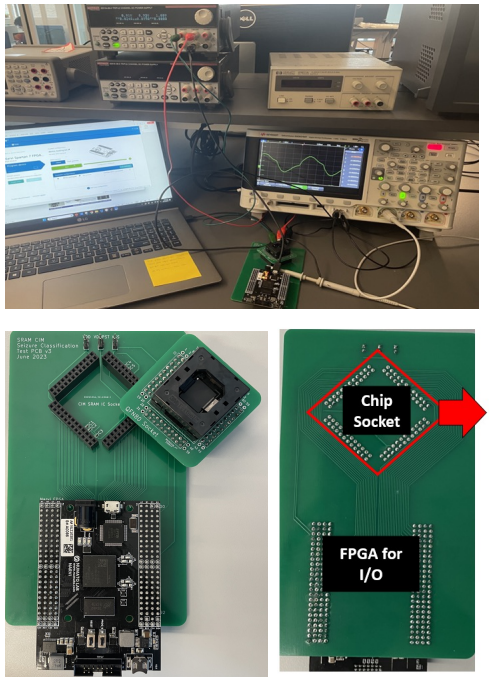
1 SYNAPTIC FUNCTION 2 VERTICAL ORGANIZATION 3 DATAFLOW REALIZATION

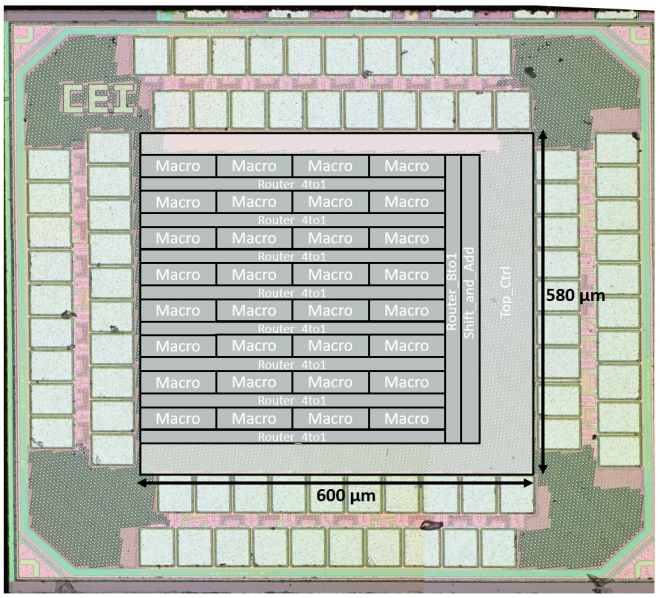
MulPi: Multi-Class & Patient Independent Seizure Classification



B. Kim et al., TBioCAS, 2025

MulPi Result



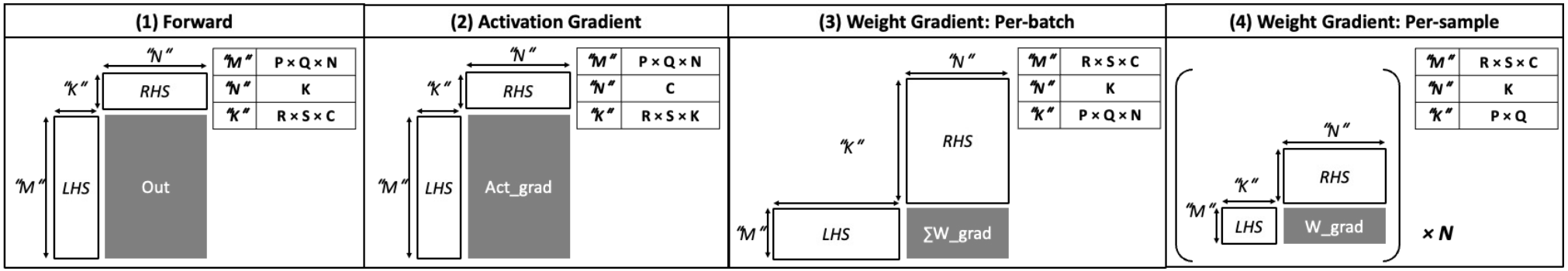


Technology	TSMC 65nm
Core Area	**0.348mm^2** << 1.96mm^2
Core V_{DD}	0.6V – 1V
I/O	2.5V - 3.3V
SRAM	32Kb
Power	0.28mW @ 0.6V
Latency	**0.12s** << 1s
Frequency	100MHz
Precision	FXP8
Sensitivity	98.5%
Specificity	99.2%
Accuracy	98.4%
PPV	98.4%

B. Kim et al., TBioCAS, 2025

Privacy: Differential Private SGD

- One forward and two back propagations with
- 1) Forward, 2) Activation Grad., 3) Per-Batch (Batch-wise) & 4) Per-Sample (Sample-) Weight Grad
- Norm along with Multiply-and-Accumulate (MAC)



B. Kim et al., DPIMA, 2026 (ISLPED 2025 Best Paper Awarded)

Customized Dataflow

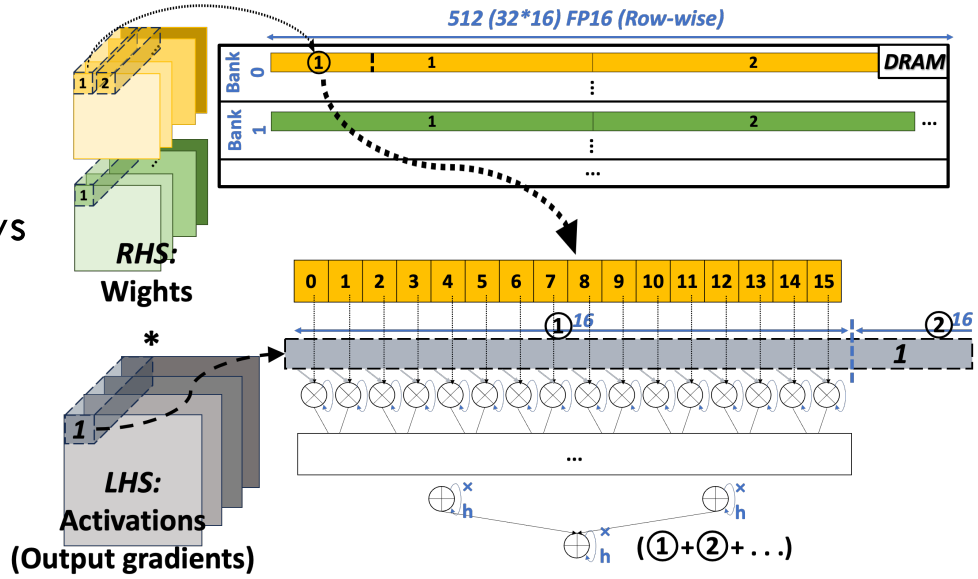
- To **Fully Utilize Available BW** of Bank
- *Commonality* in Operators & Output Size

1. ***FG_COMP*** for Fwd. & Act. Gradient
   - Channel-Mise Elements mapped to rows
2. ***WG_COMP*** for Weight Gradient

   (Per-Batch and Per-Sample)

   Column-Major Format

   mapped to rows of DRAM



*Achieving up to **13.8× Performance** and **123.8× Energy Efficiency***

B. Kim et al., DPIMA, 2026 (ISLPED 2025 Best Paper Awarded)

Beyond Intelligent Computing

Energy efficiency involves dataflow

– not a device property alone

The benefit of memory-centric computing depends on

- Which device and memory are used, *but also*
- Which data remain local, how intermediate results move, and
- Whether the architecture exposes the memory's native parallelism

What must remain local?
What movement dominates?
Where is parallelism exposed?