

Driving AI Efficiency through Heterogeneously Integrated Data-Centric Computing

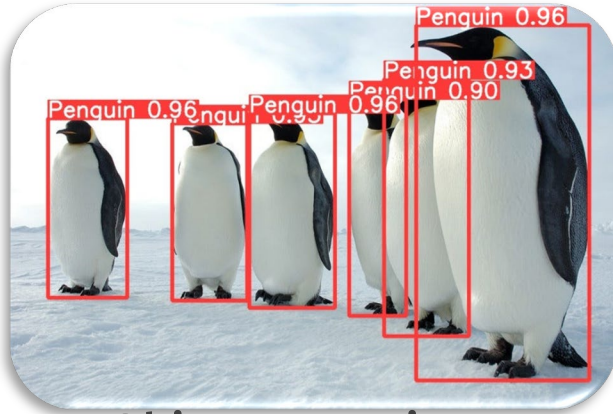
Dr. Wantong Li

Assistant Professor, Department of Electrical and Computer Engineering
Cooperating Faculty, Department of Computer Science and Engineering

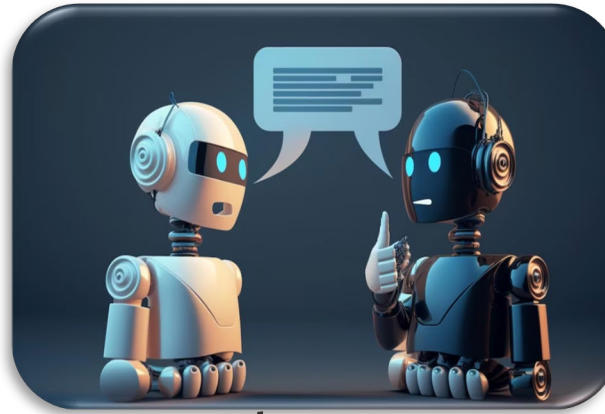
Bourns College of Engineering

Jul 21, 2025

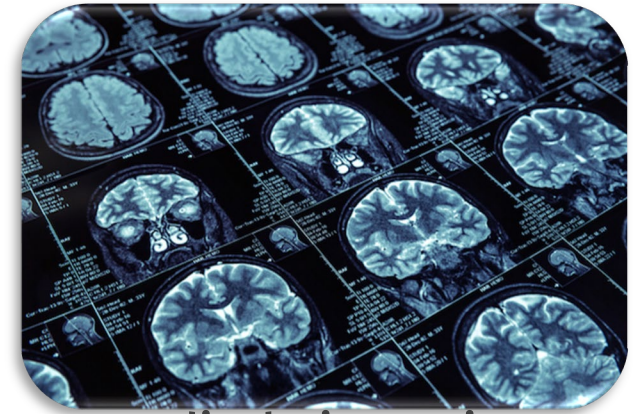
Wide-Ranging Applications of AI



Object Detection



Natural Language



Medical Diagnosis



Autonomous Navigation

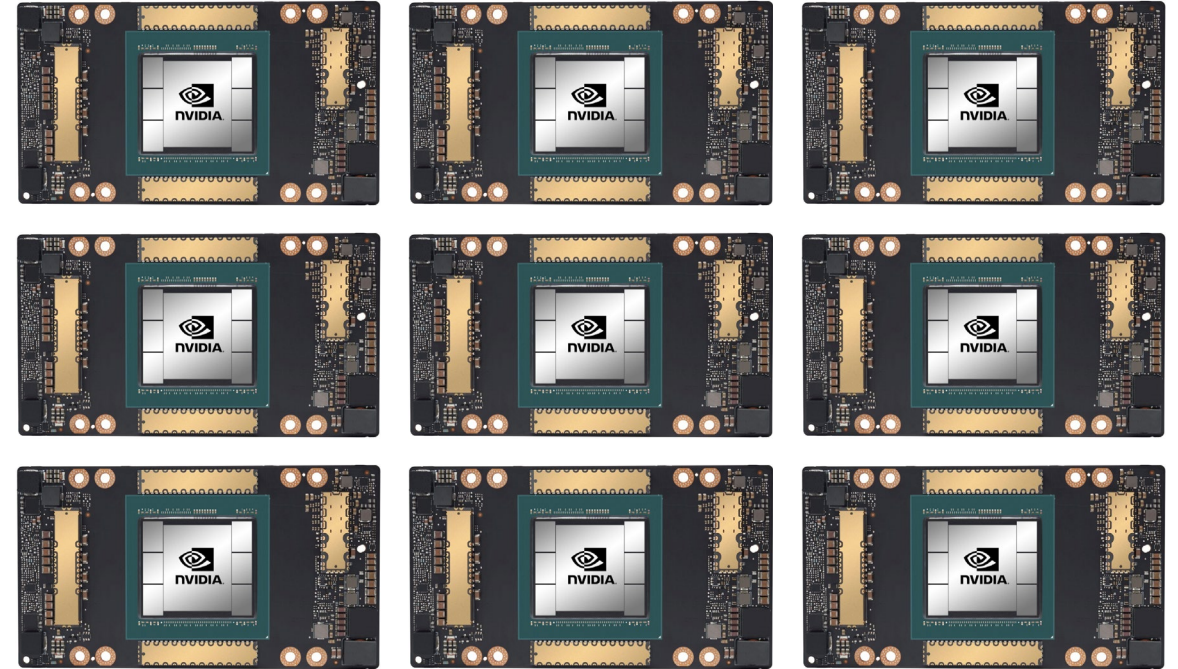
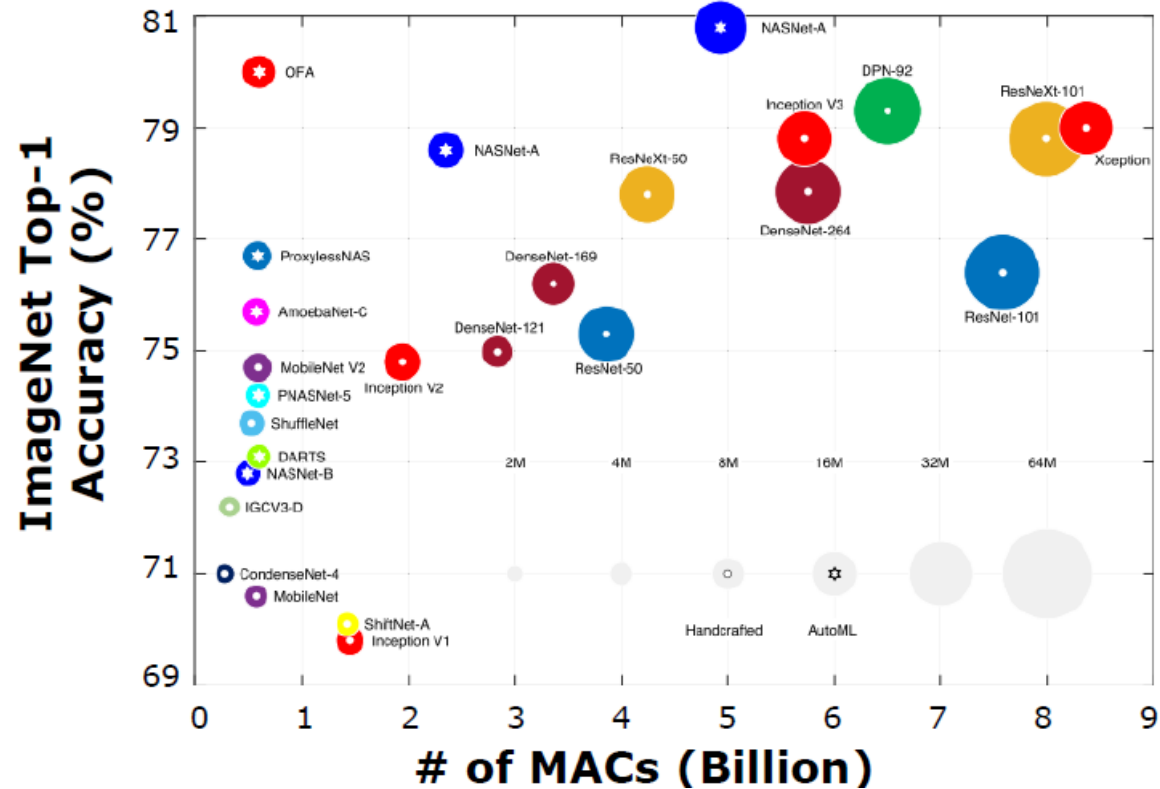


Malware Detection

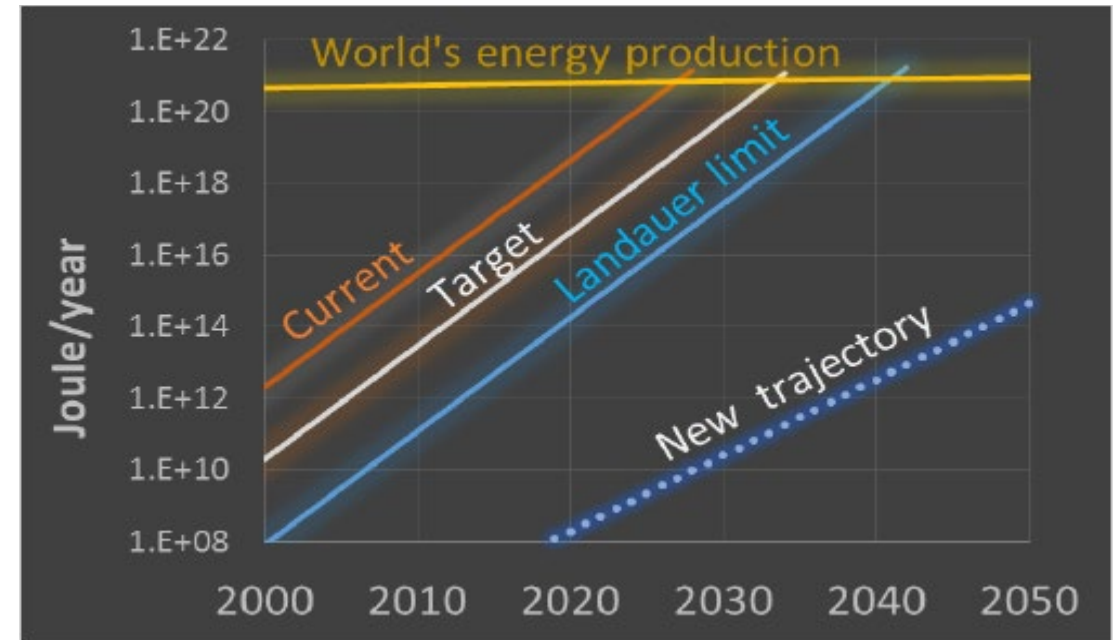
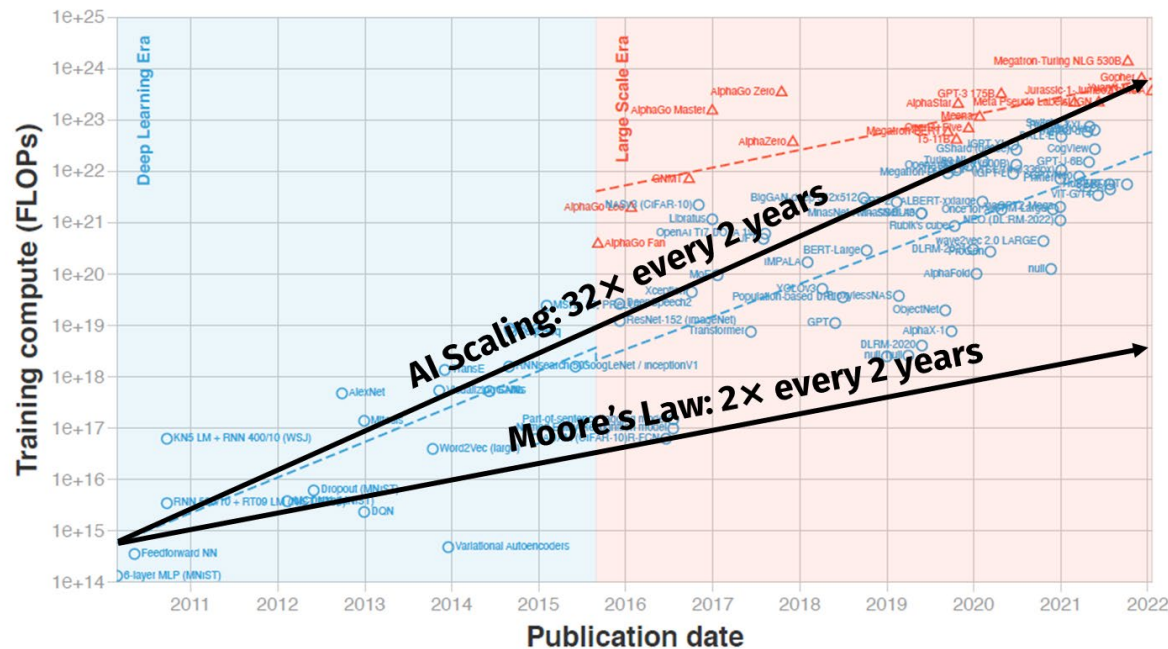


Drug Discovery

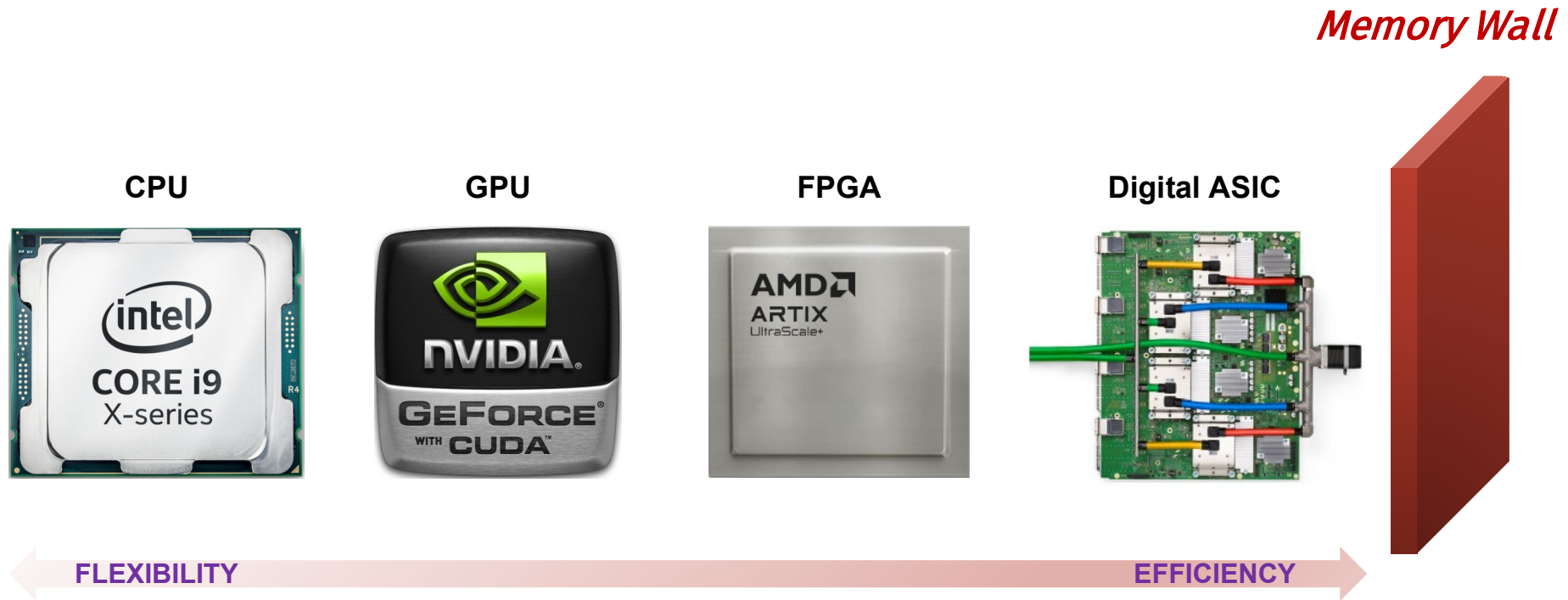
Recent Triumphs of Deep Learning



Grand Challenge: Computational Efficiency

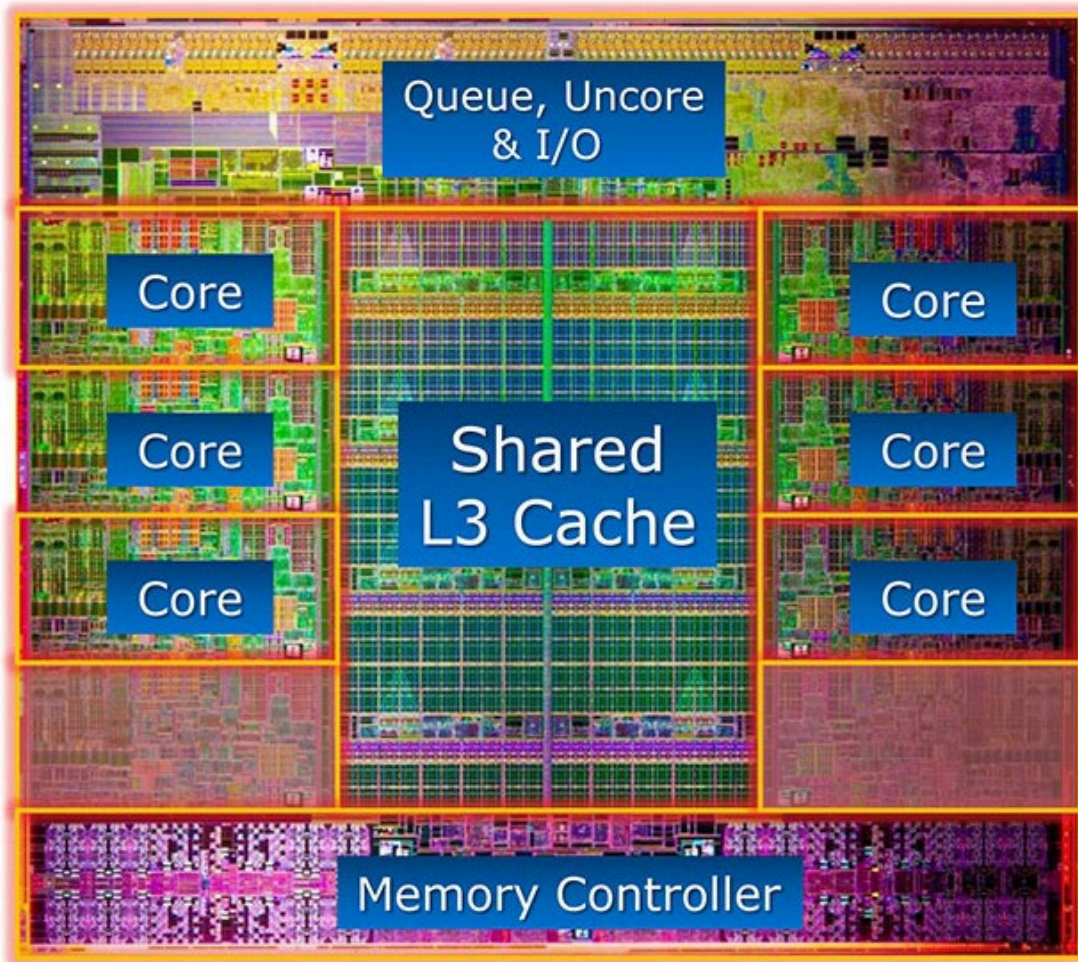


Landscape of Today's Computing Systems



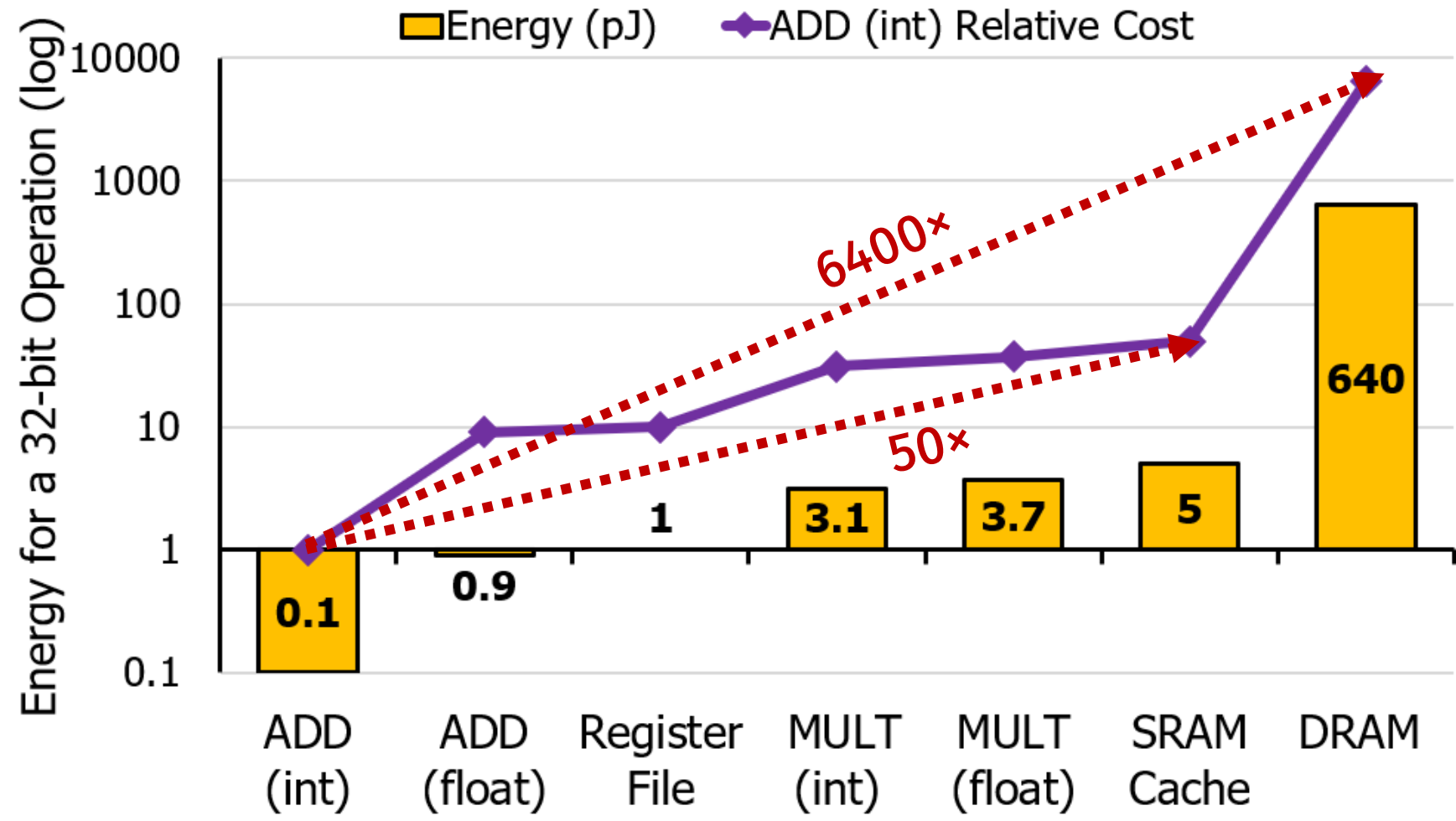
They are all **processor-centric** platforms

Processor-Centric Systems



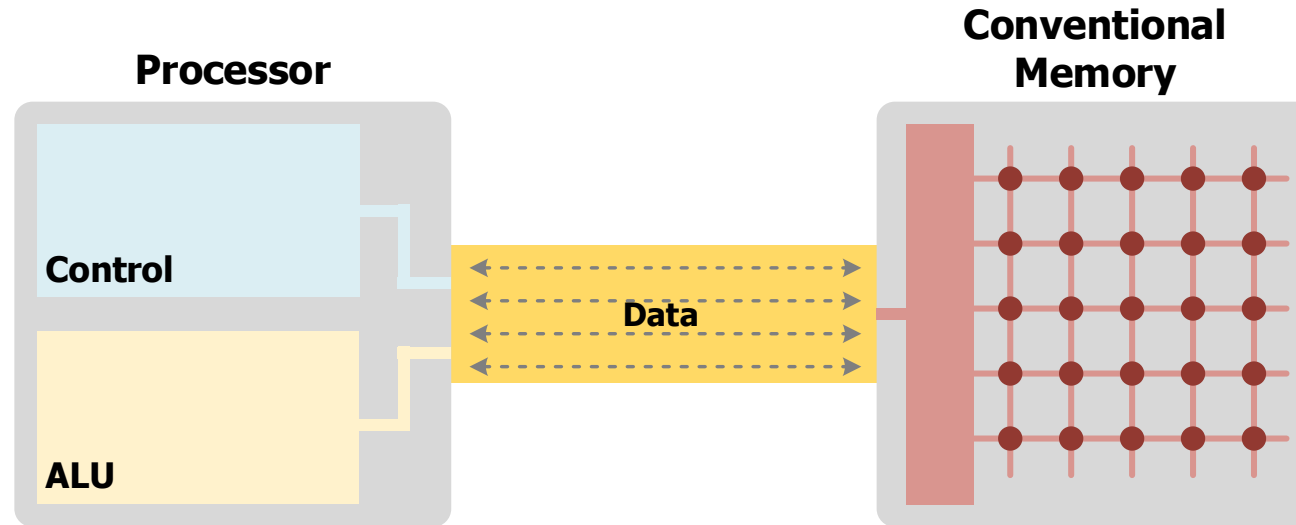
- Imbalanced computing systems
 - Most of the system is dedicated to storing and moving data
 - Processing is done only in processors
- Yet, system energy and latency are still **bottlenecked by data**

Cost of Data Movement



Paradigm Shift of Computing

- **Challenge:** Processing is performed far away from where data is stored



- Paradigm shift towards **data-centric computing** to:
 - ✓ enable computation close to or in memory
 - ✓ shorten distance of data movement
 - ✓ reduce dimensionality of data

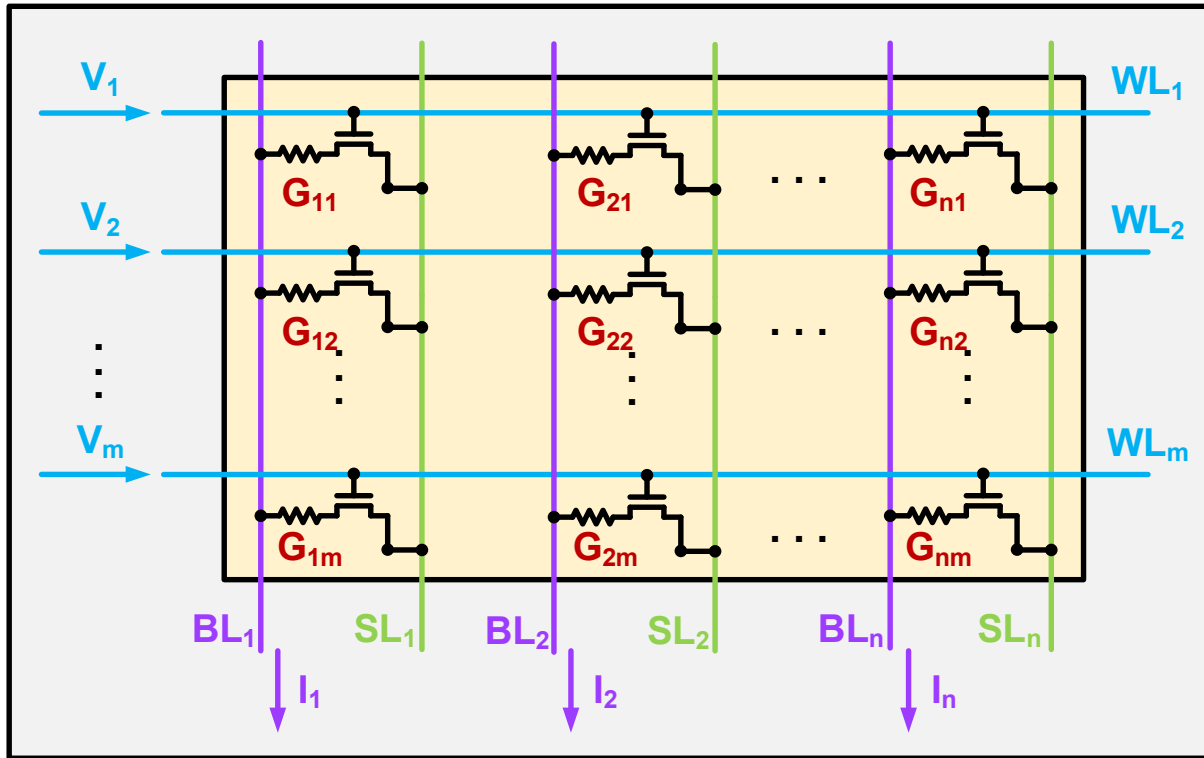
Outline

- Theme 1: Emerging Computing Paradigms
 - Enable computation inside memory
- Theme 2: More-than-Moore Heterogenous System
 - Shorten distance of data movement
- Theme 3: Algorithm/Hardware Co-Design
 - Reduce volume of data movement

Outline

- Theme 1: Emerging Computing Paradigms
 - Enable computation inside memory
- Theme 2: More-than-Moore Heterogenous System
 - Shorten distance of data movement
- Theme 3: Algorithm/Hardware Co-Design
 - Reduce volume of data movement

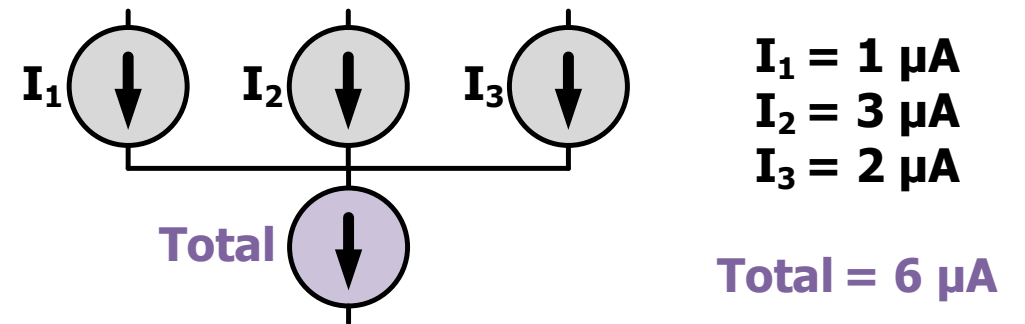
Compute-in-Memory as MAC Accelerator



- Multiply: logical AND

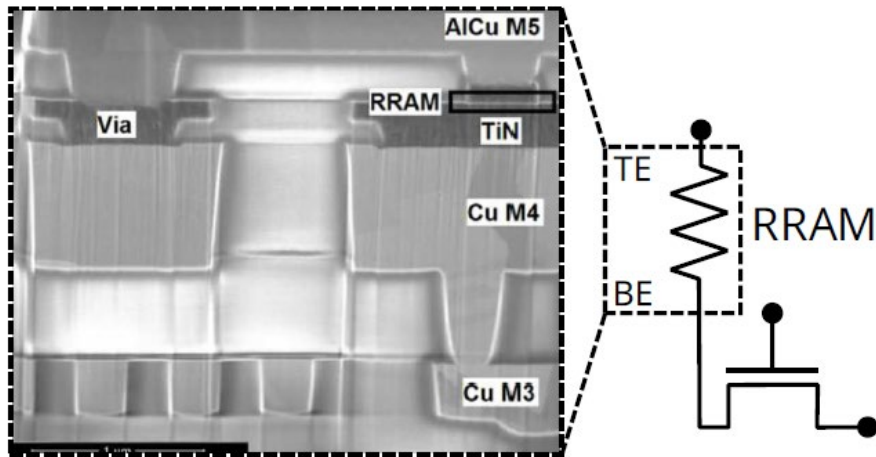
V (input)	G (conductance)	Out (bitline)
0	0	0
0	1	0
1	0	0
1	1	1

- Accumulate: current summation

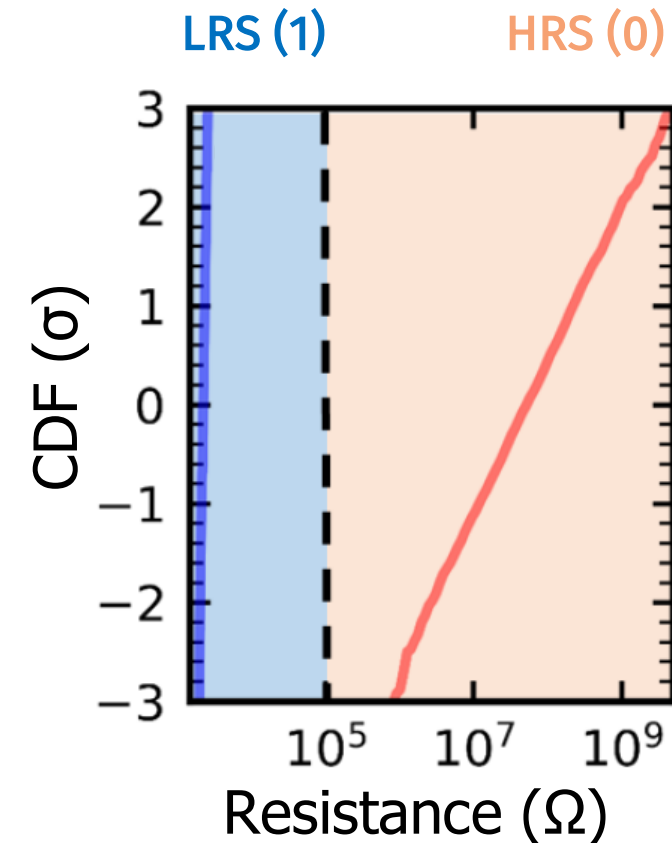


RRAM-based CIM for AI

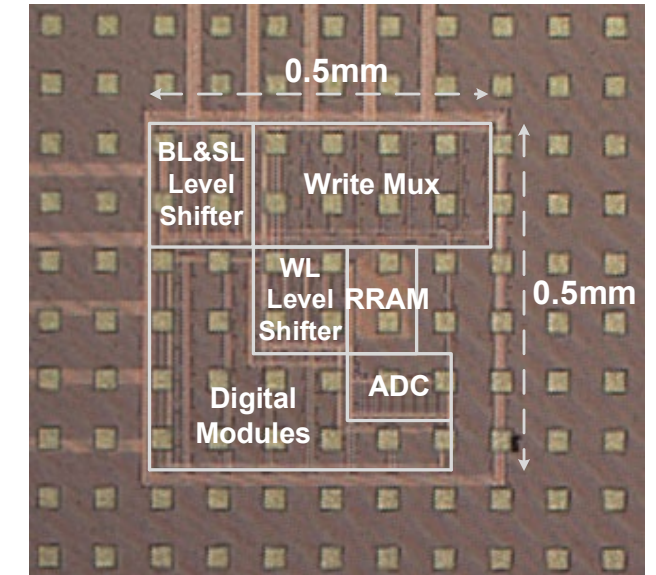
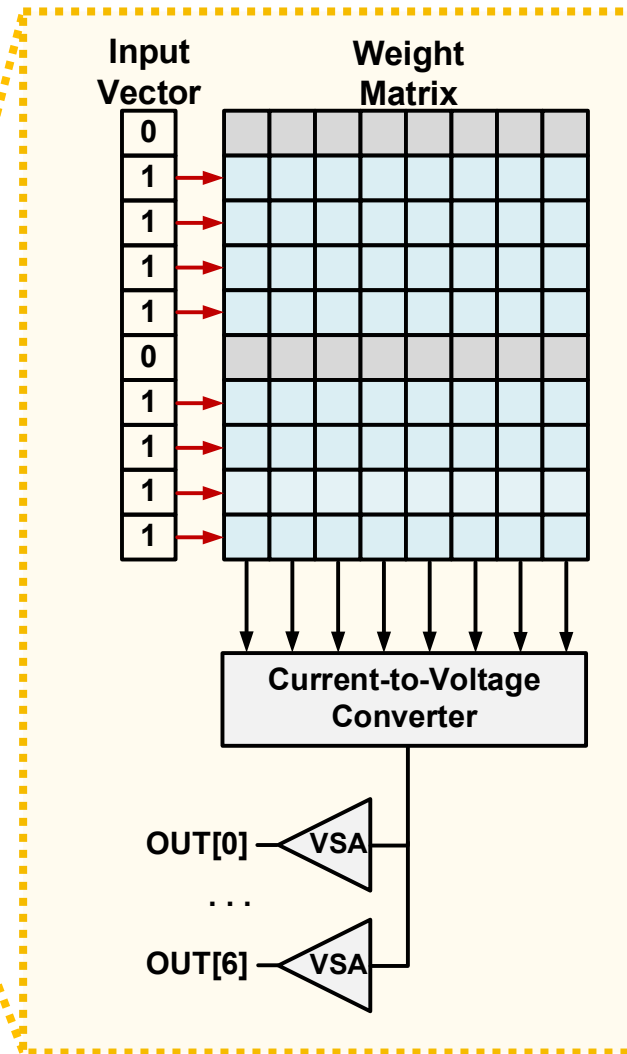
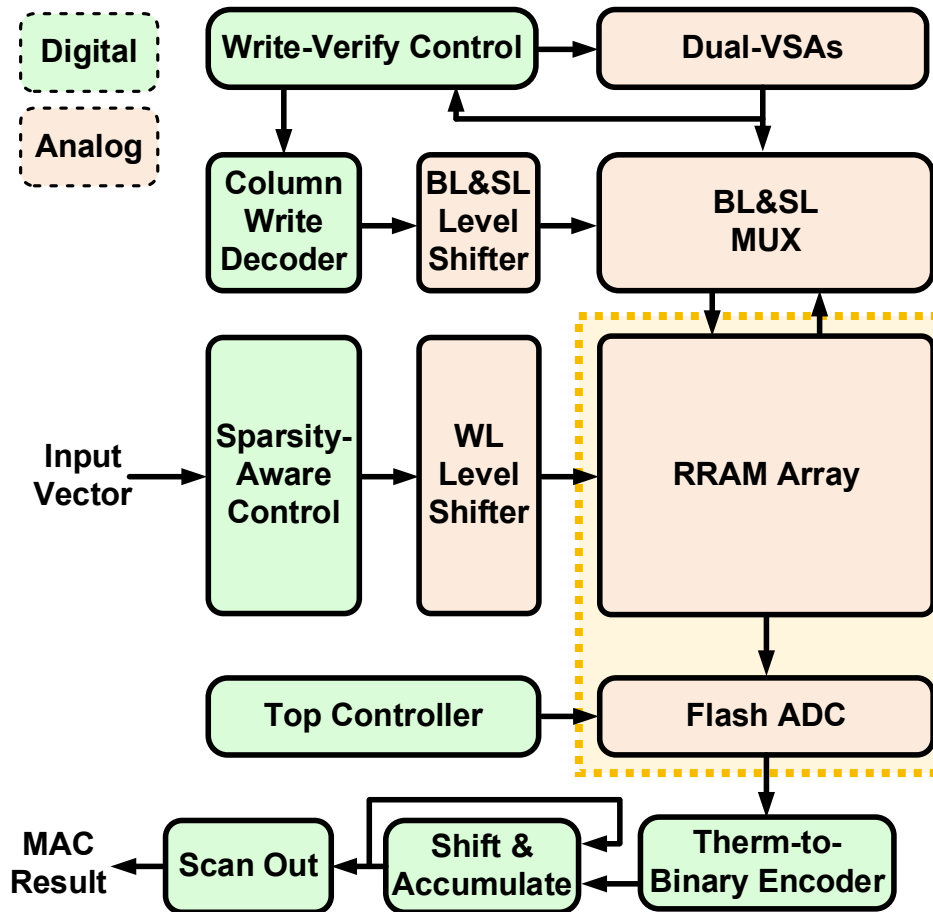
- Benefits of resistive RAM (RRAM) for CIM
 - Non-volatile on-chip storage of model parameters
 - High capacity with possible multi-level cell
 - Low leakage during standby



- Low-power and low-cost AI inference in embedded electronics

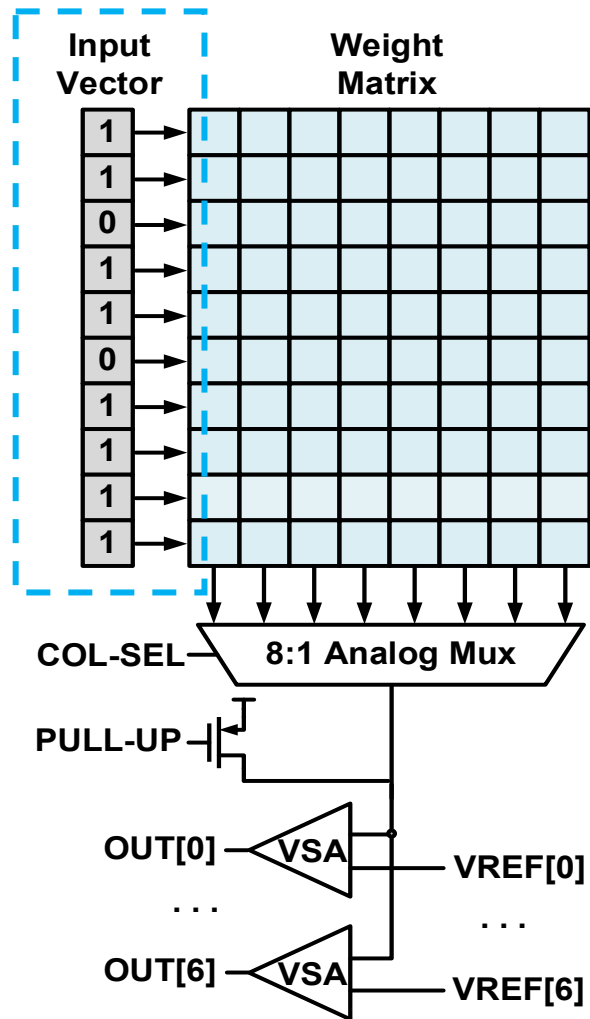


RRAM-CIM Tape-out



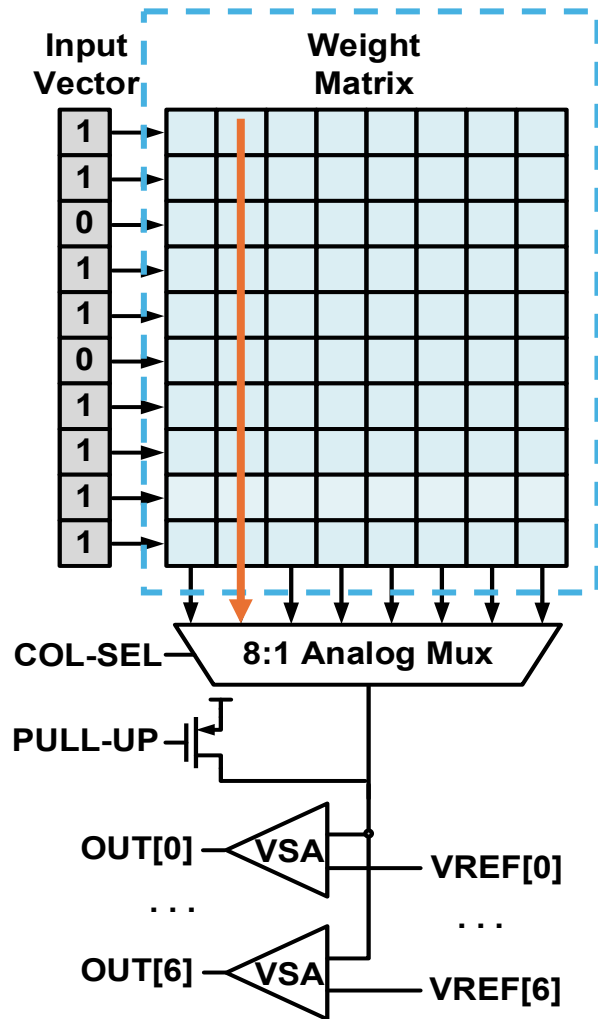
Test chip taped-out in TSMC 40-nm node with embedded 1T1R RRAM

Compute Scheme: Input Activations

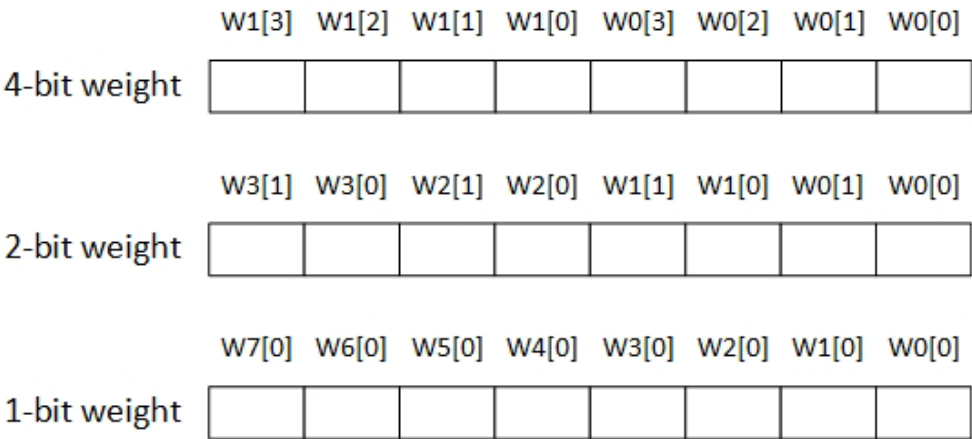


- Opening multiple wordlines (WLs) at once enable parallelized MAC operations
- Each input bit asserts one WL
- Input controller ensures 7 activated WLs for each set of computation

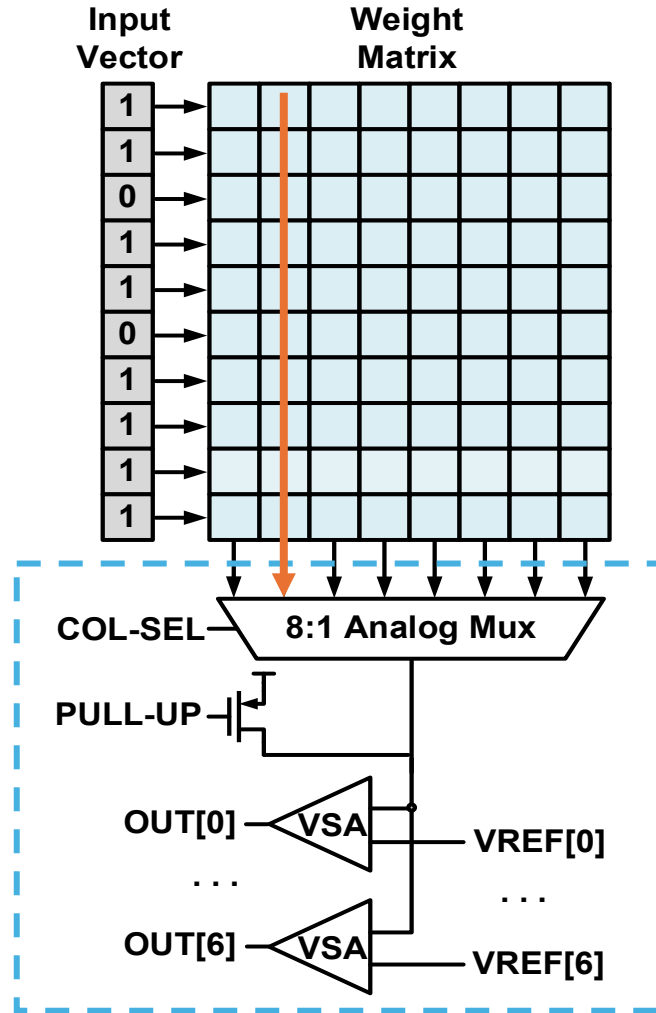
Compute Scheme: Weight Storage



- Weights are stored in 1T1R cells
 - HRS denotes '0' and LRS denotes '1'
 - Reconfigurable weight precision

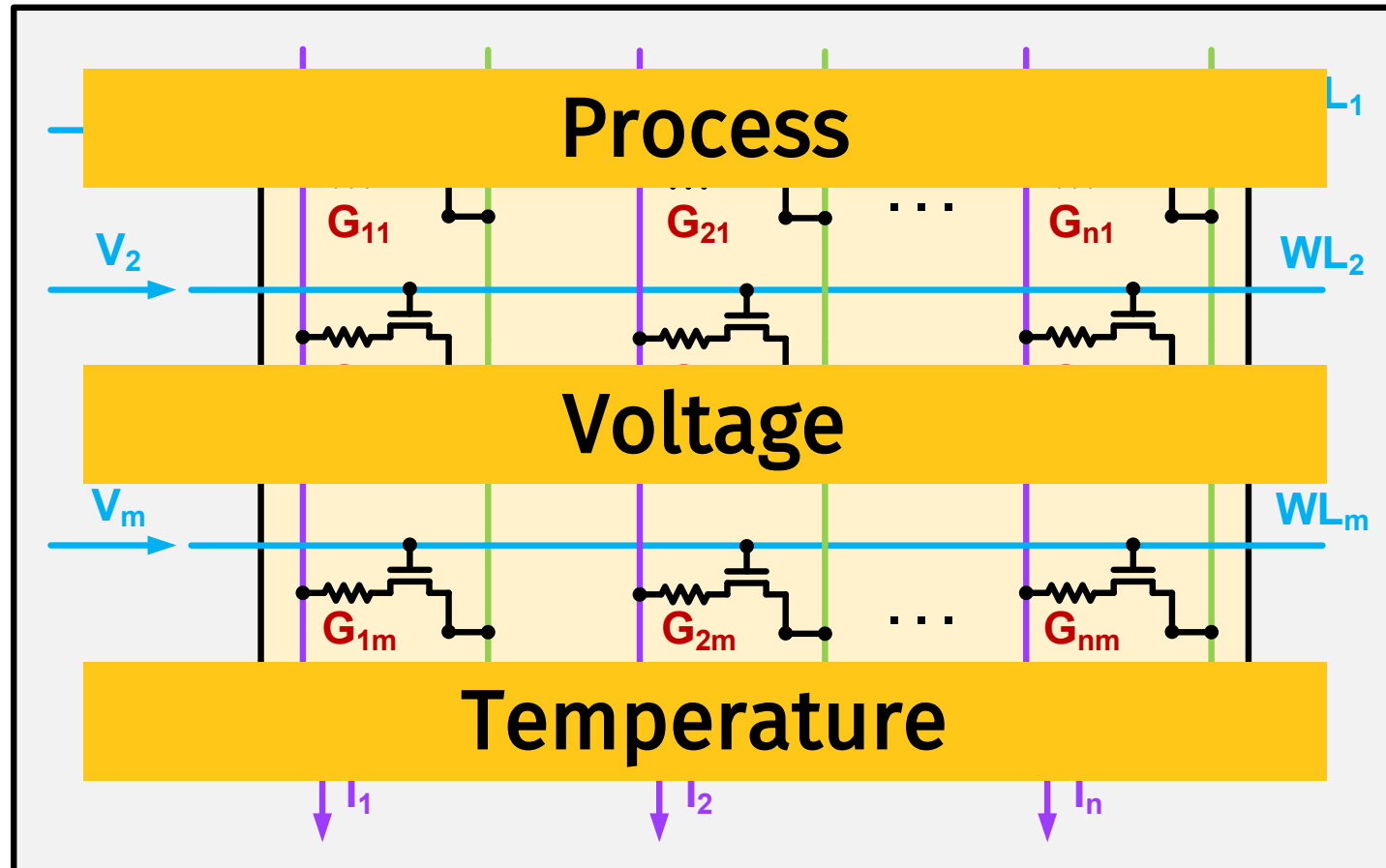


Compute Scheme: Peripheral Sensing

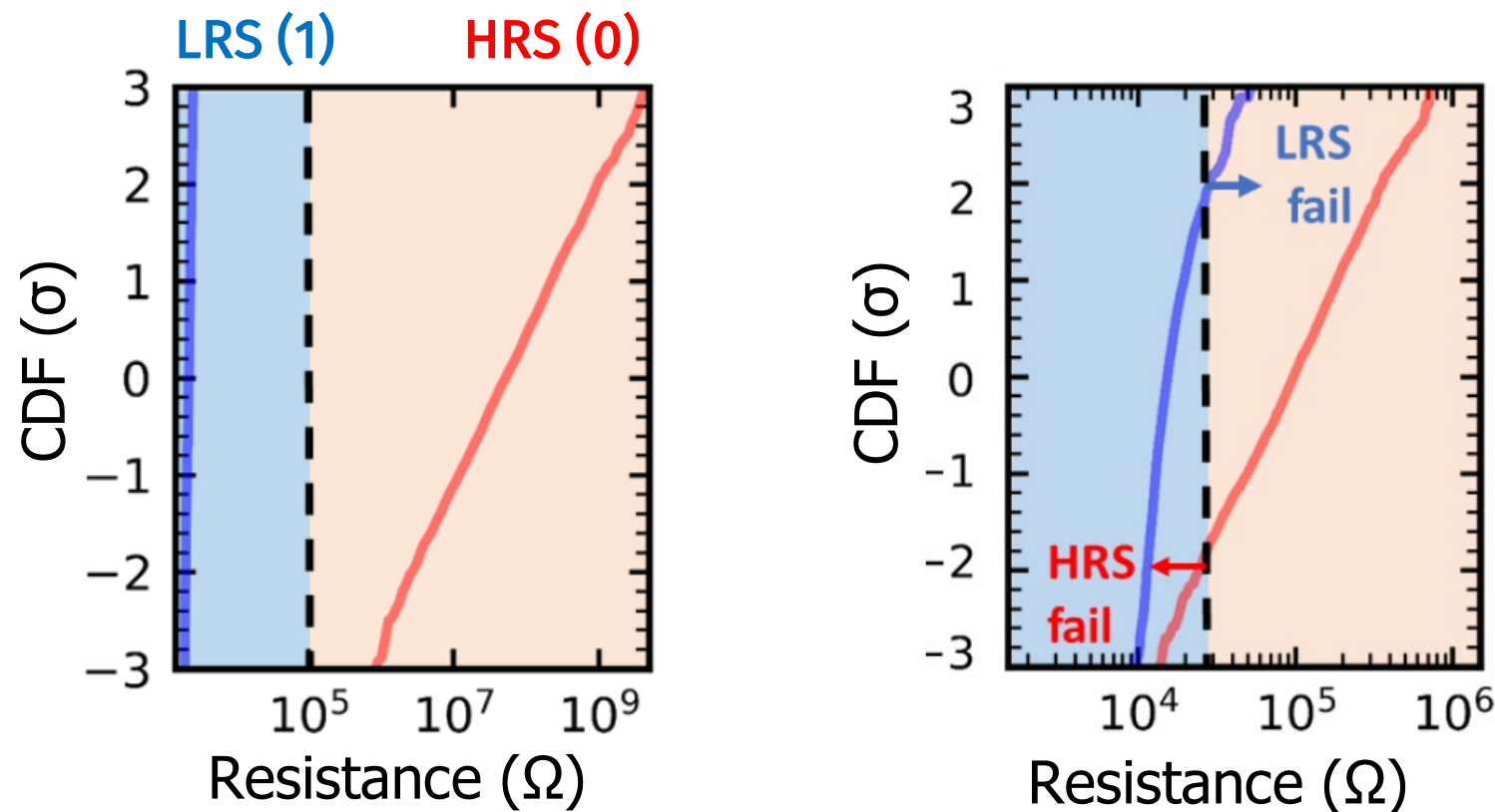


- 8:1 mux for time-multiplexing
- Resistive divider between RRAM cells and pull-up PMOS
- BL voltage sensed by 3-bit flash ADC (7 voltage sense amplifiers)
- 7-bit ADC thermometer output is encoded as 3-bit binary

Design Challenges in RRAM-based CIM

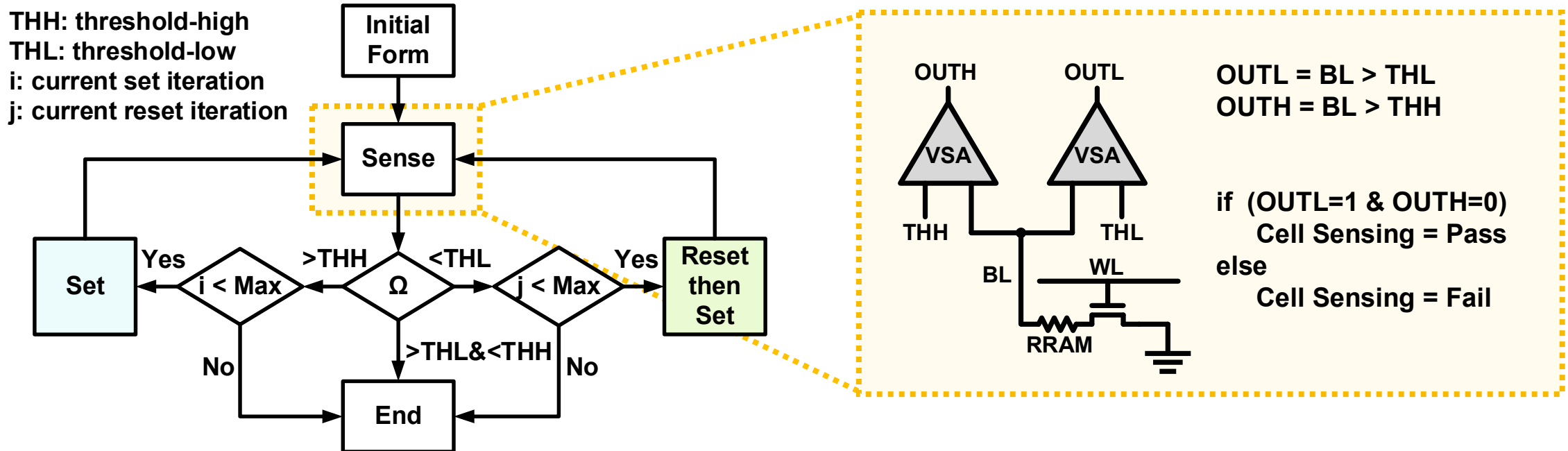


Challenge: Device Process Variation



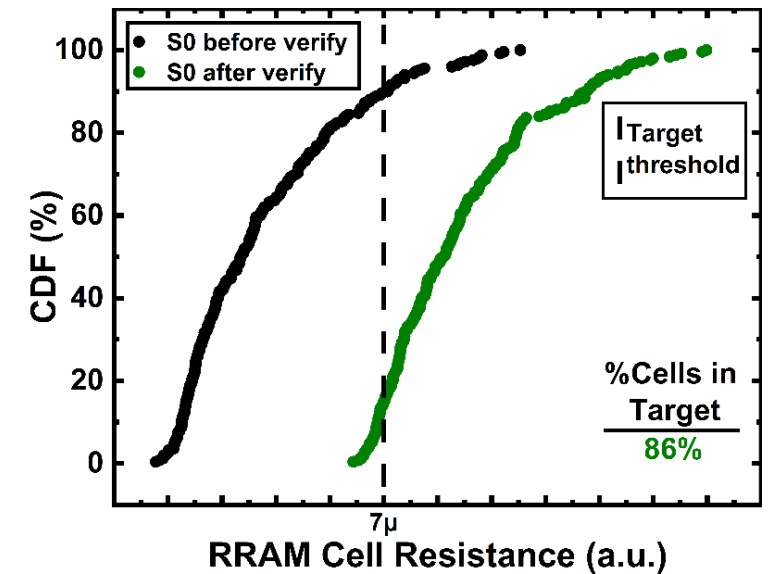
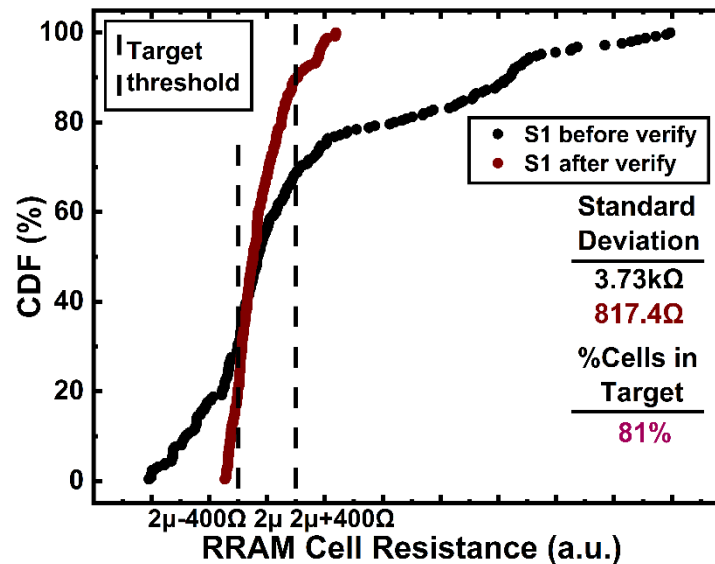
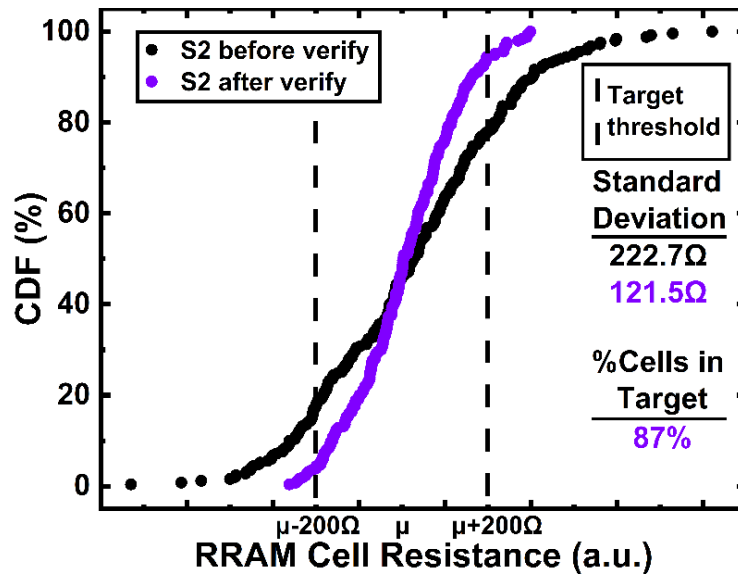
On-Chip Write-Verify for CIM

- High-speed on-chip programming of RRAM cells for CIM applications
 - Two thresholds (THH & THL) make up a resistance window



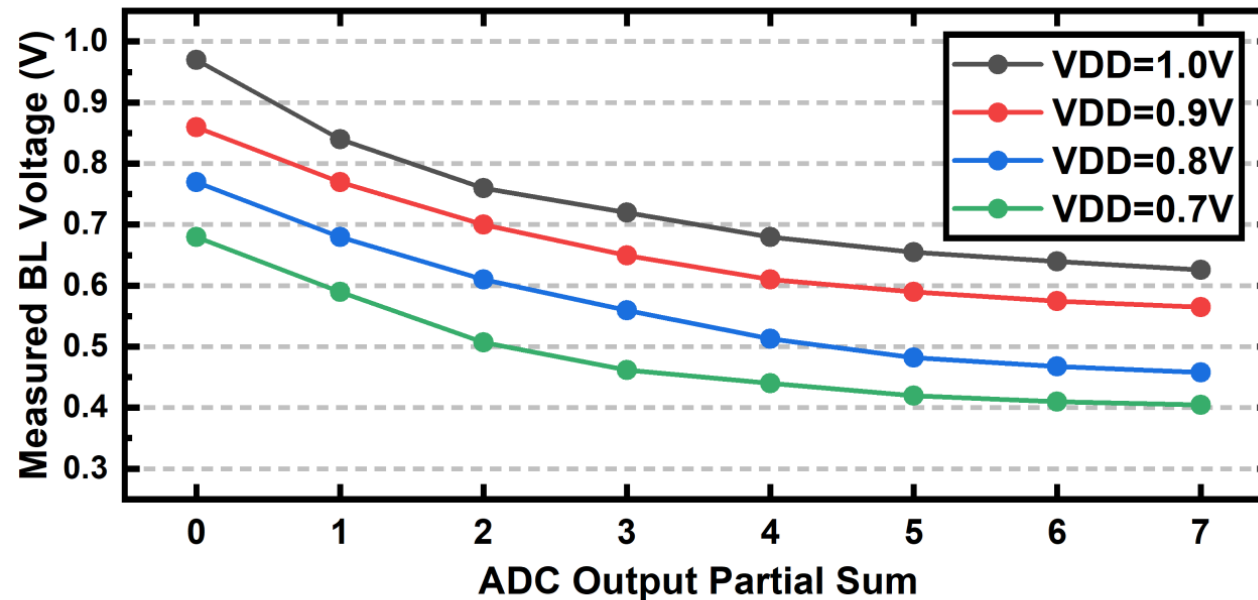
Write-Verify Measurement Results

- **>80% cells** can be programmed in target with on-chip write-verify
- Achieve **10^5 programming speedup** compared to off-chip programming equipment



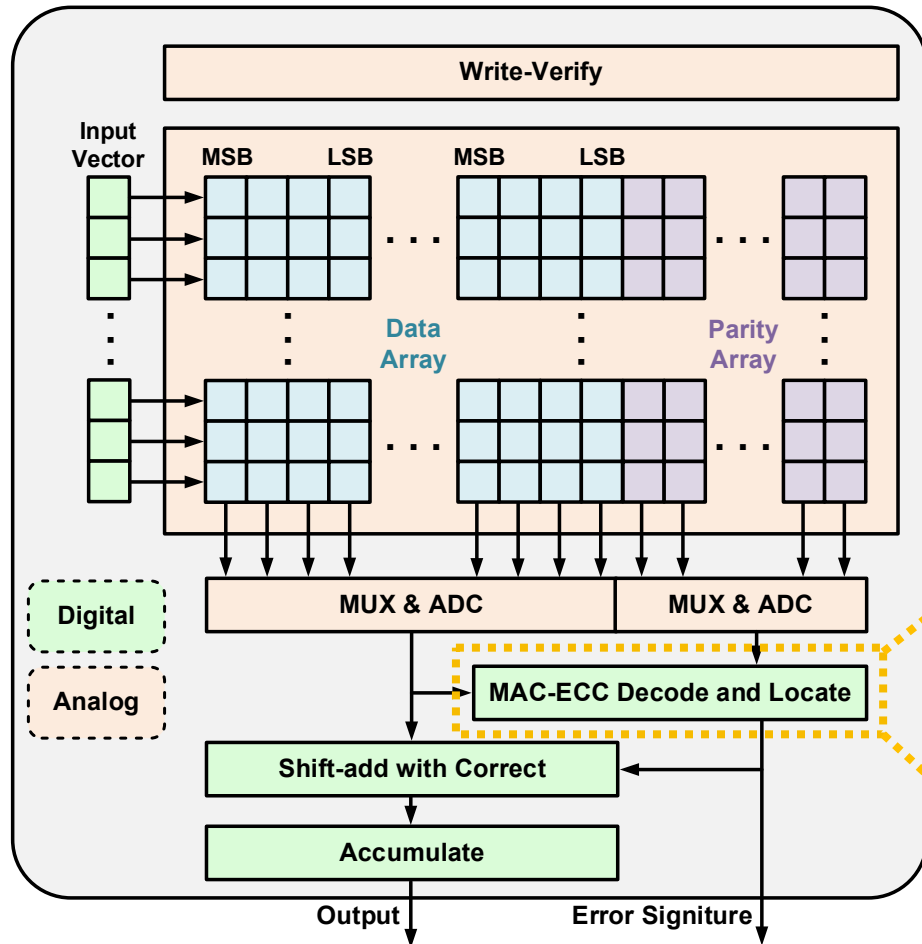
Challenge: Supply Voltage Toggling

- **Supply voltage (VDD) scaling:** popular method to toggle systems between different modes of operations

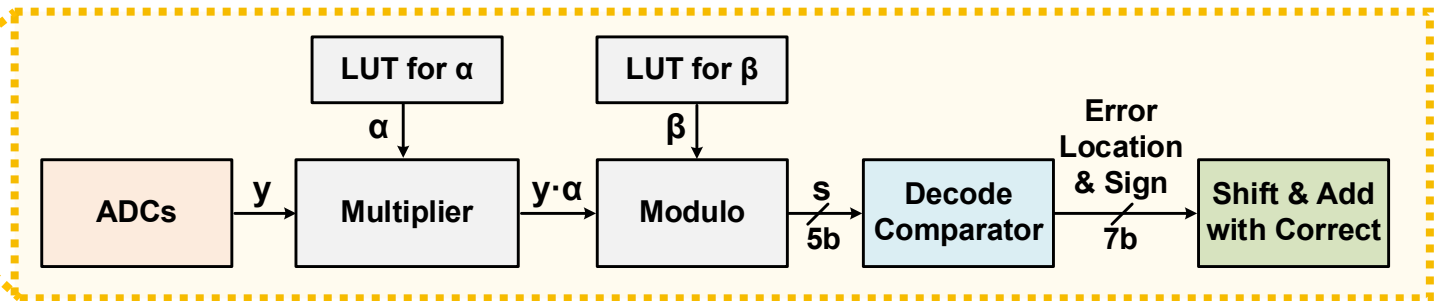


- Scaling VDD from 1.0V to 0.7V **reduces ADC sense margin** by >50%
 - Reduced sense margin induces **more ADC errors**

MAC-ECC: In-Situ ECC for CIM




- MAC-ECC **corrects analog and mixed-signal errors** in the digital domain (after ADC)
- Correct errors based on **arithmetic distance**
 - Example: $(0111)_2$ and $(1000)_2$
 - Hamming distance: 4 Arithmetic distance: 1



Iso-Accuracy Voltage-Scaling

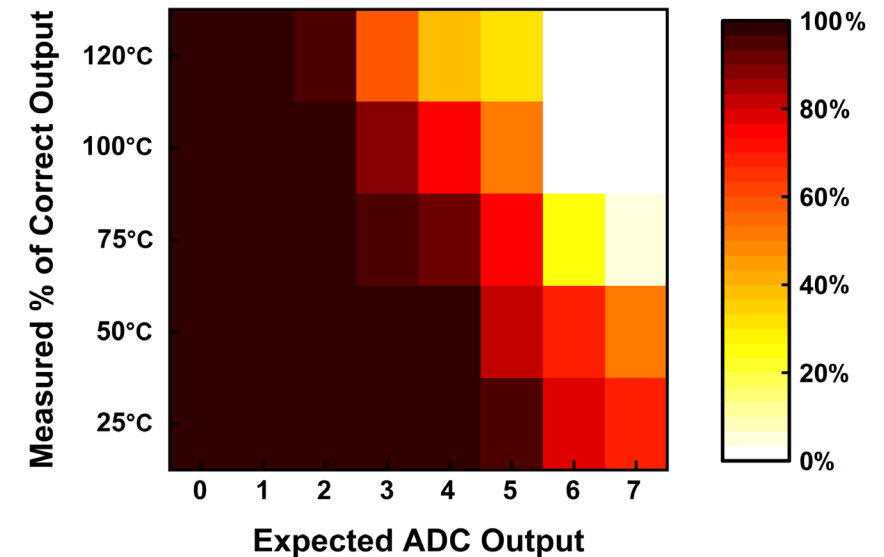
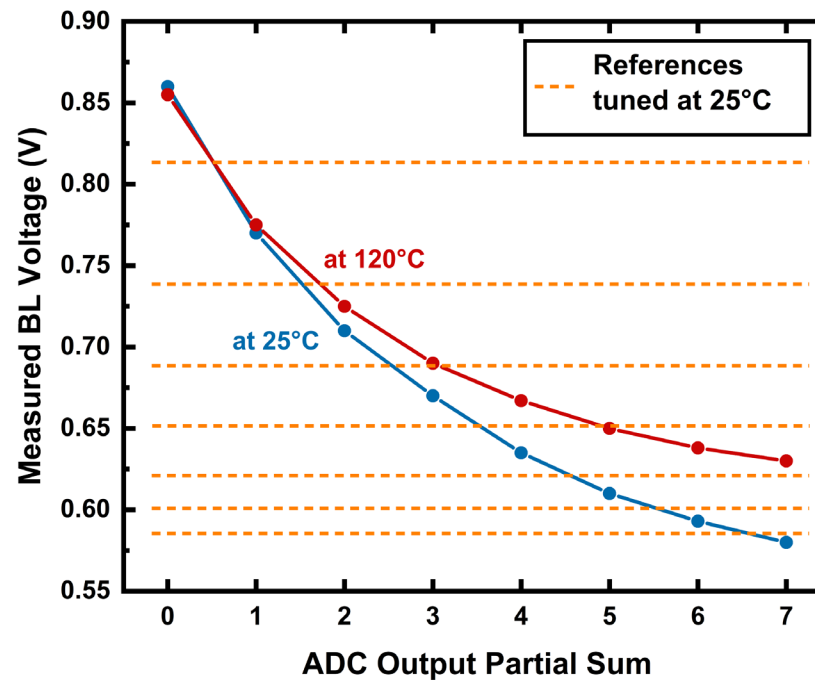
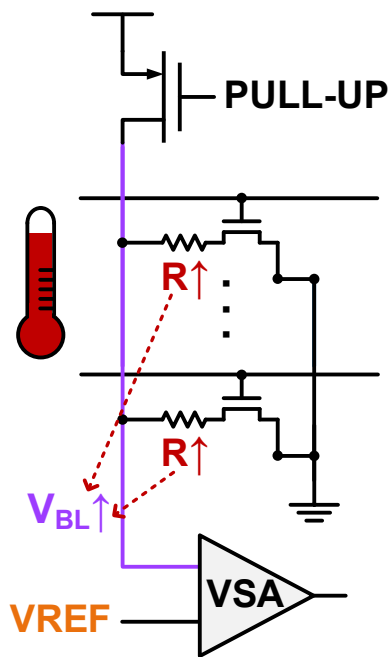
- Voltage scaling allows toggling between **high performance** and **low power** modes
- Higher error rates at low VDDs are compensated by allocating more MAC-ECC columns
 - Achieve **iso-accuracy** voltage scaling (<1% CNN accuracy loss under all tested VDDs)



VDD	Least costly MAC-ECC for <1% loss	Frequency	Energy Efficiency (TOPS/W)	Compute Efficiency (GOPS/mm ²)	Energy Overhead	Area Overhead
1V	No ECC	115MHz	43.0	112.5	0%	0%
0.9V	(31, 25)	100MHz	46.2	93.1	3.73%	3.1%
0.8V	(25, 19)	90MHz	52.4	82.8	4.95%	4.11%
0.7V	(16, 10)	80MHz	59.1	70.9	7.48%	6.06%

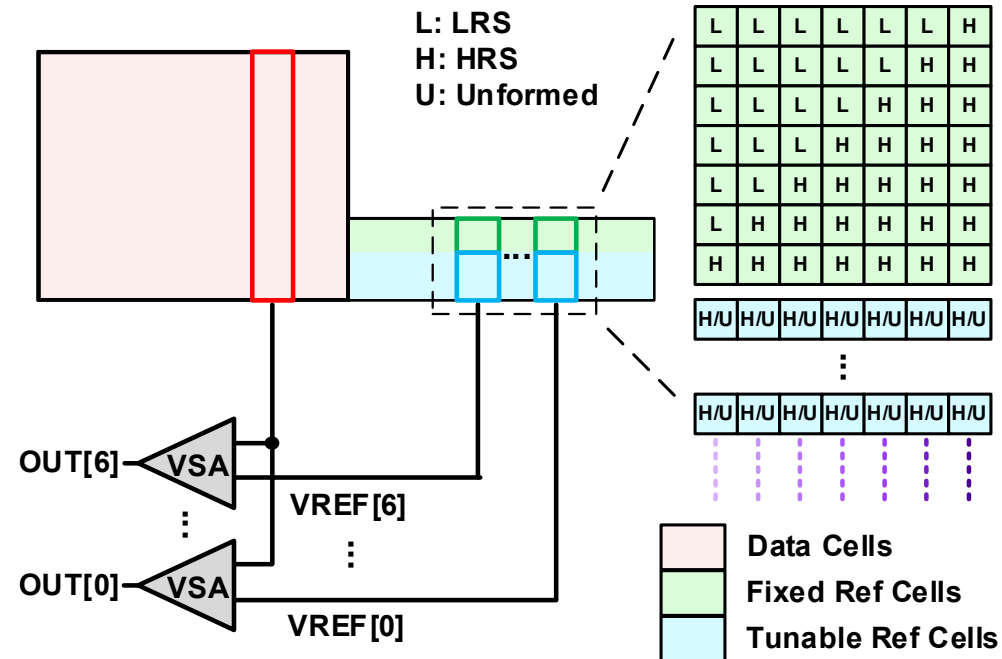
Challenge: Temperature Fluctuation

- RRAM resistance tends to increase with rising temperature
 - Thus, sensed V_{BL} is dependent on temperature
 - **Rigid references** tuned at **one temperature** (e.g. 25 °C) works poorly at **others** (e.g. 120 °C)
 - Mismatch between references and BL voltage curve causes missing ADC codes



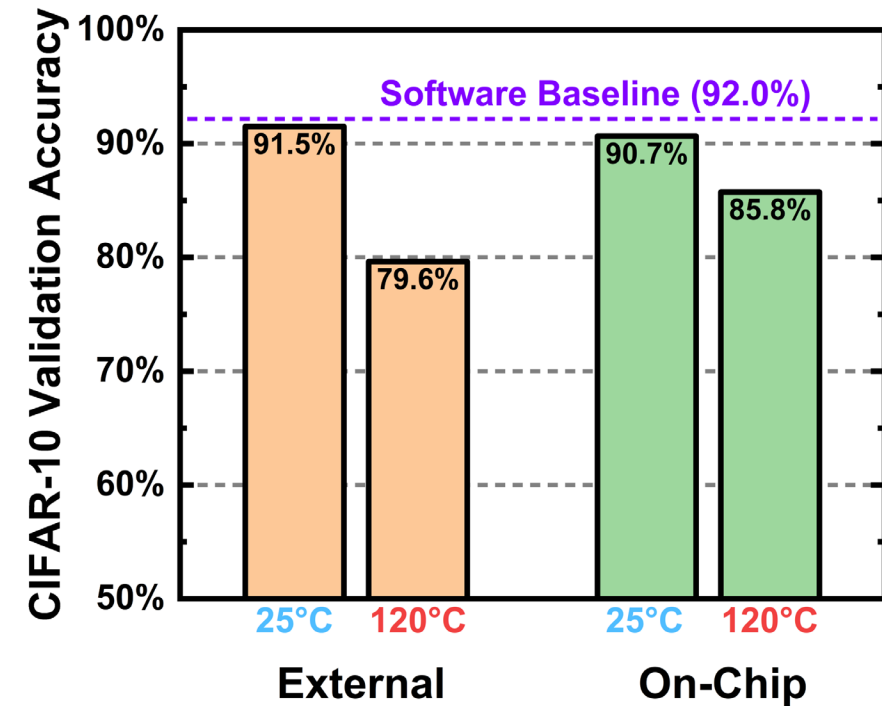
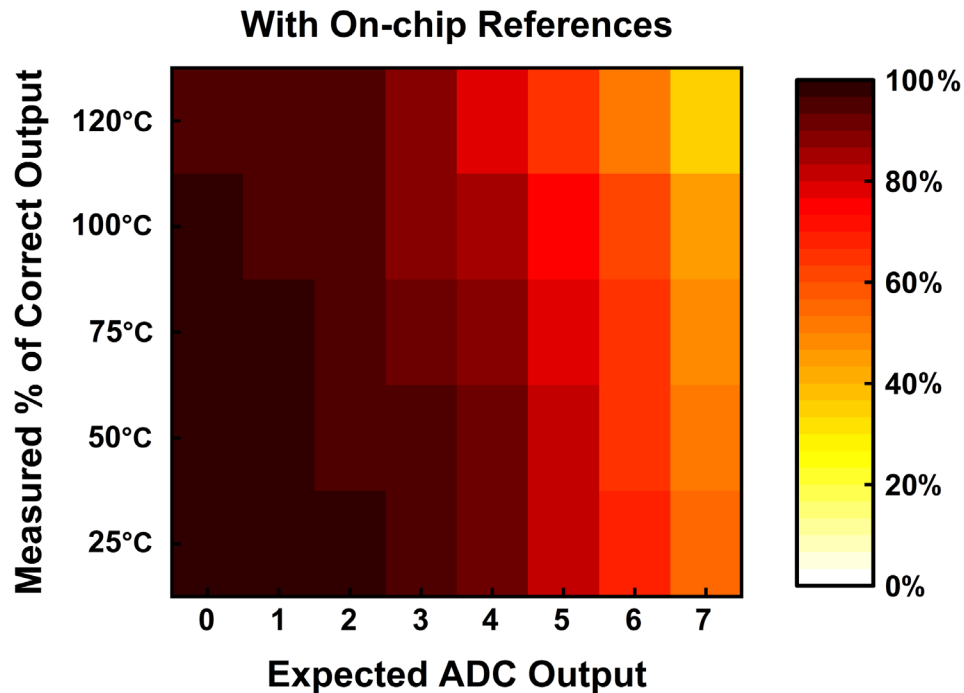
Temperature-Independent ADC References

- Use a separate section of RRAM cells for on-chip ADC reference generation
 - RRAM provides self-tracking to temperature
 - Tunability (to different RRAM states) helps cancel ADC offsets
 - Requires ~5% overhead of RRAM cells



Evaluation of ADC References

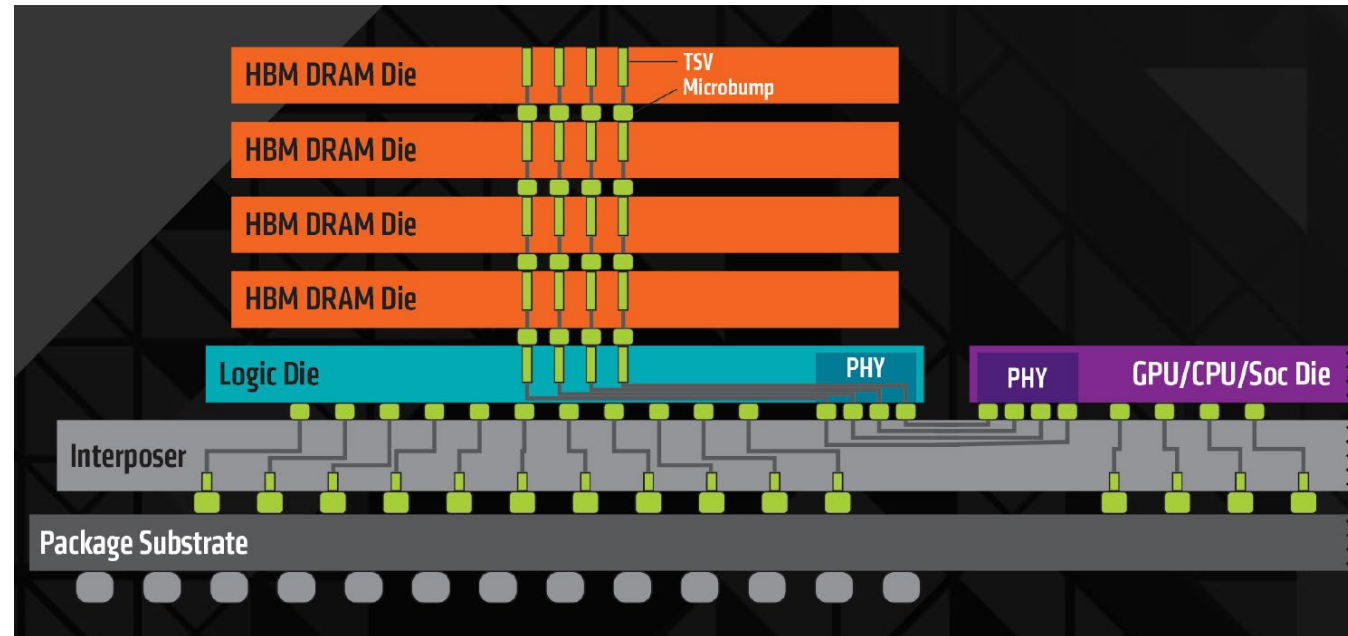
- Measured ADC outputs with random test vectors at various temperatures
 - RRAM-based references can **retain all ADC codes**
 - Network accuracy at 120 °C is **recovered by ~6%** using temperature-independent references



Outline

- Theme 1: Emerging Computing Paradigms
 - Enable computation inside memory
- Theme 2: More-than-Moore Heterogenous System
 - Shorten distance of data movement
- Theme 3: Algorithm/Hardware Co-Design
 - Reduce volume of data movement

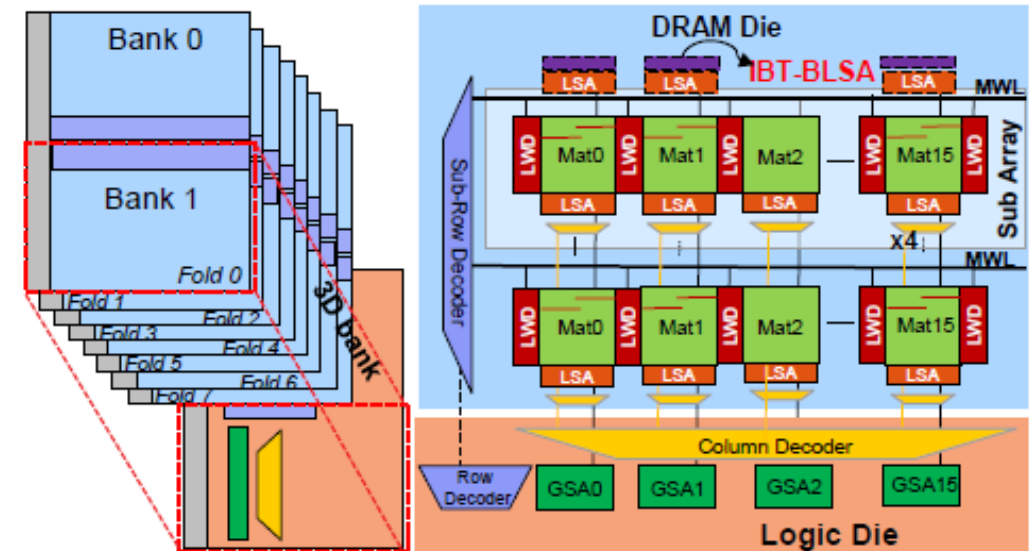
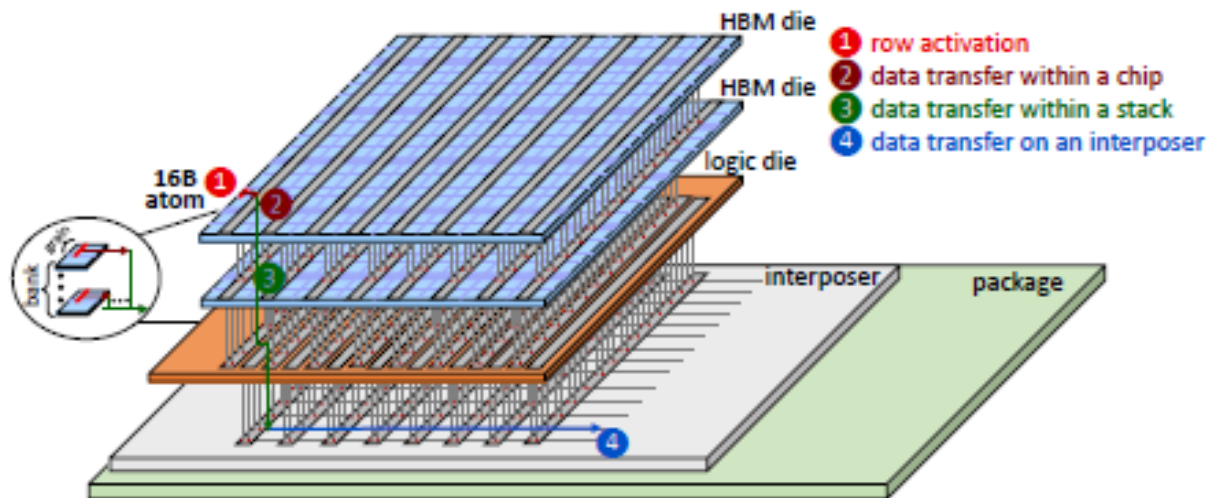
Heterogeneous 3-D Integration (H3D)



- **Advantages:**
 - Compact form factor to store data
 - Flexibility to stack different technologies and fabrics into the same package
- State-of-the-art H3D is memory-on-memory (NAND Flash, HBM, 3D V-Cache etc.)

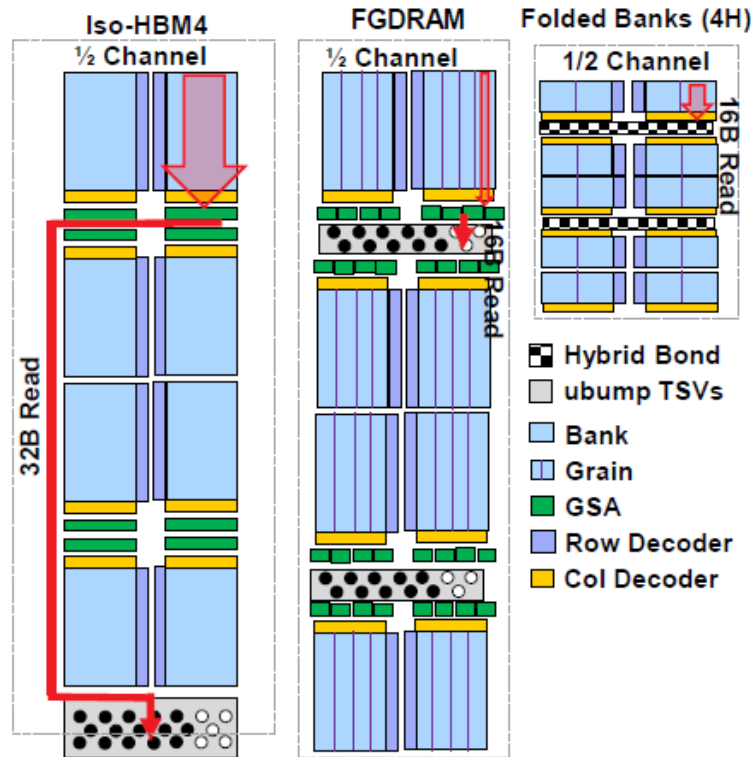
Folded-Banks HBM

- Alternative version of HBM4 optimized for random accesses
 - Redistributes bank subarrays (“folds”) across multiple dies and relocates command, control, and global sense amplifiers to an additional base layer



Folded-Banks HBM

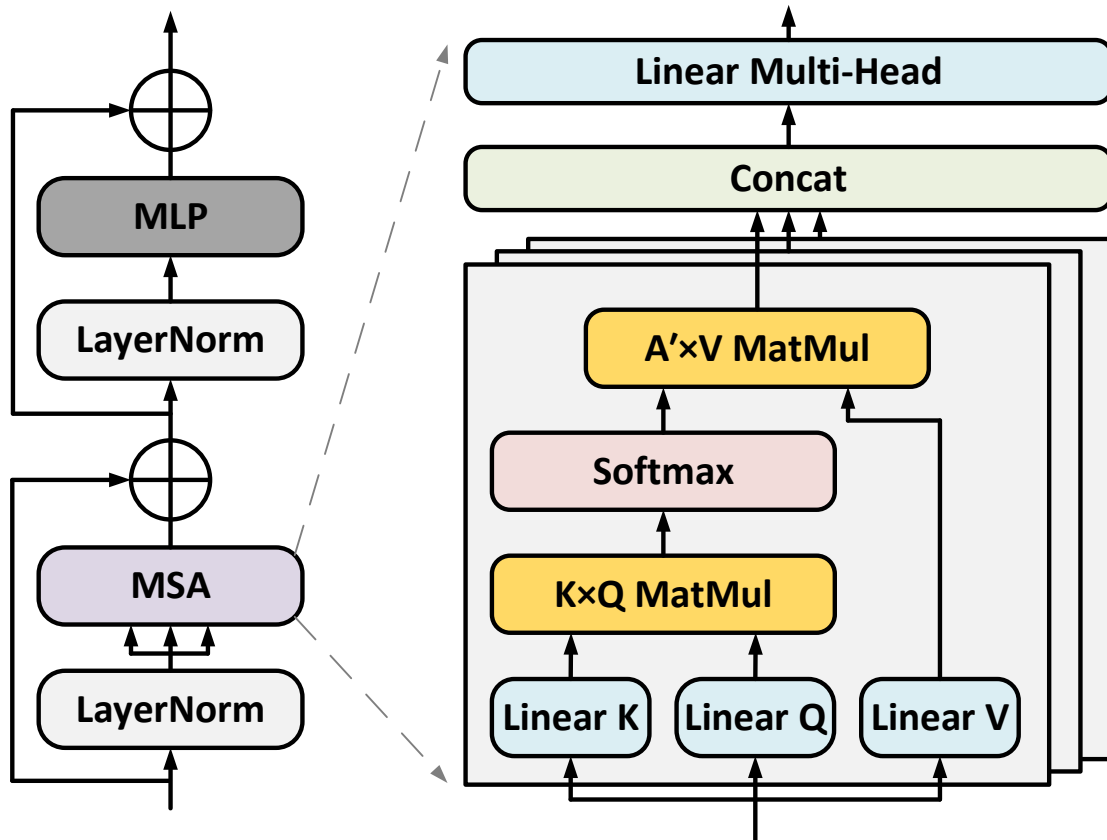
- Novel architecture leads to significant reduction in t_{CCD} and t_{CL} delays
- Compared to a projected HBM4 design, FB-HBM achieves a **6.7× improvement in random-access performance**
- Translates to a **2.28× speedup** across HPC and sparse machine learning applications



Parameter	HBM3	Iso-HBM4	FB-HBM
Technology Node (nm)	16	9	9
Bank Height (μm)	918.4	635.5	95.8
Driver Enable Delay (ns)	0.2	0.2	0.2
CSL Driver Resistance (Ω)	250	300	300
CSL Load Capacitance (fF)	8	8	8
SSA Pre Delay (ns)	0.2	0.2	0.2
Wire Resistance (Ω/mm)	2670	4000	4000
Wire Capacitance (fF/mm)	180	180	180
MDL Driver Resistance (Ω)	200	300	300
TSV Resistance ($m\Omega/\text{TSV}$)	–	–	154.9
TSV Capacitance (fF/TSV)	–	–	11.6
CSL Resistance (Ω)	2450	2550	381.8
CSL Capacitance (fF)	173.1	122.7	91.6
MDL Resistance (Ω)	2450	2550	381.8
MDL Capacitance (fF)	165.1	114.7	83.6
t_{CCD} (ns)	2.55	2.07	1.044
t_{CL} (ns)	15.8	13.4	10.04

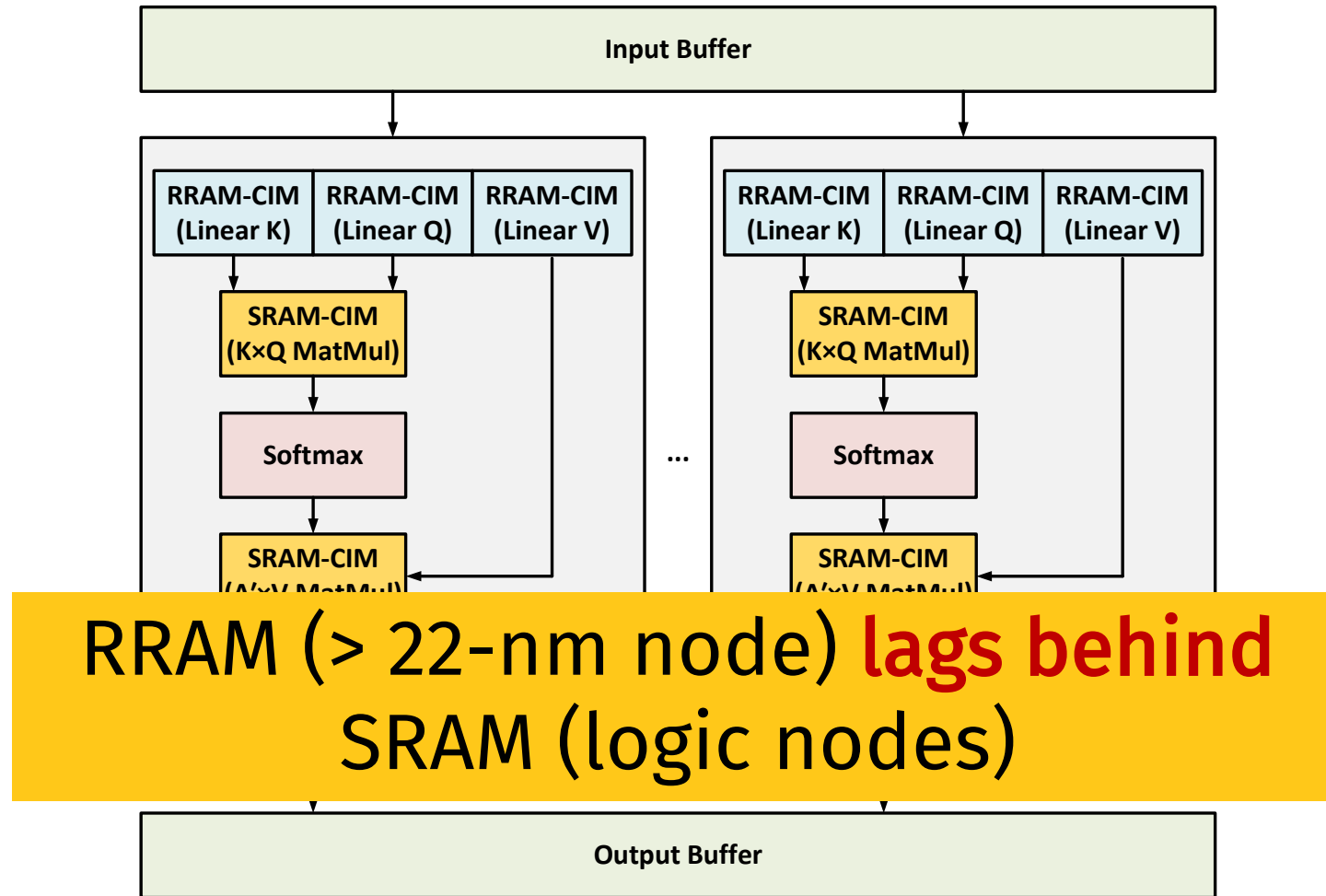
Choice of Memory Technologies for Transformers

Multi-head Self-Attention (MSA) in Transformer Model

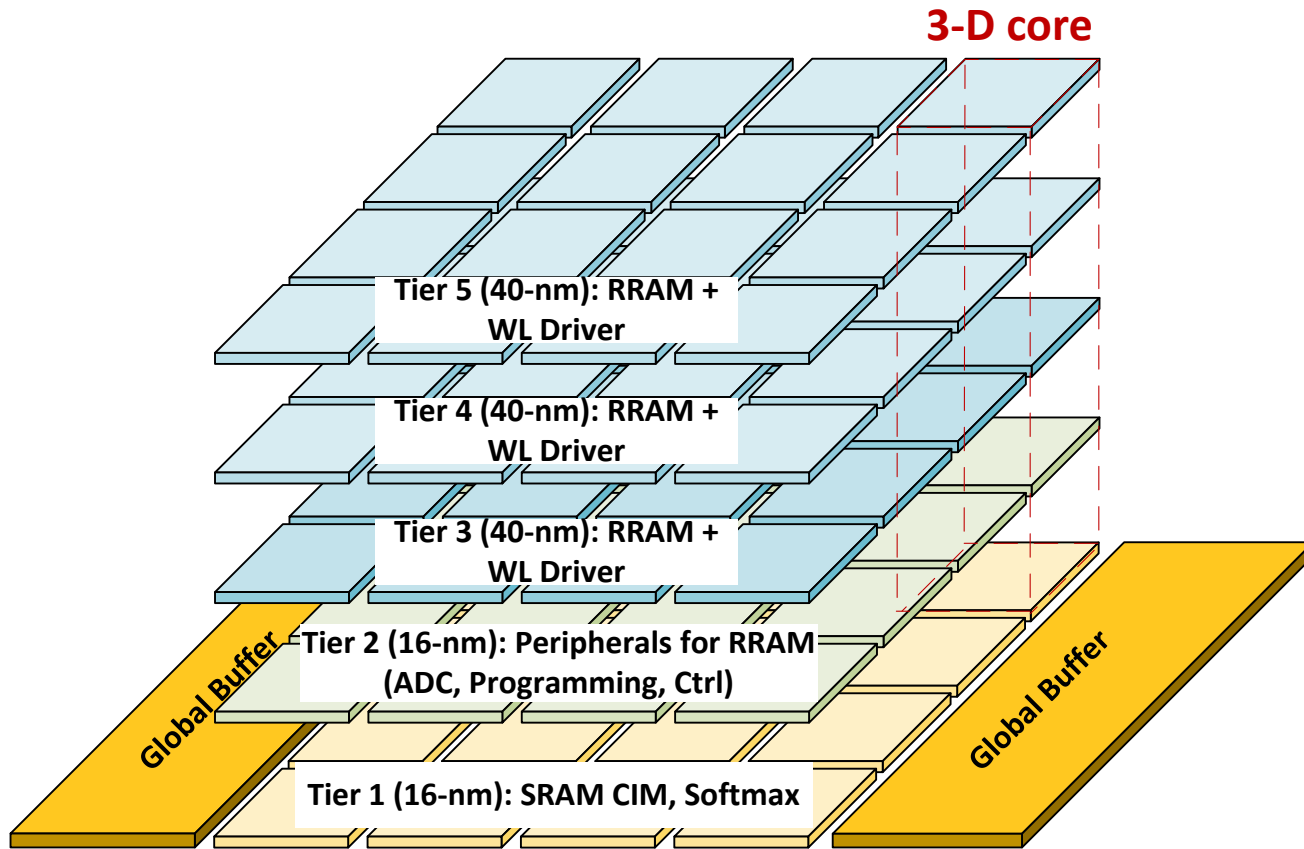


- MSA is the critical module in transformers
 - Matrix multiplication workloads in MSA involve different properties
- Linear (fully-connected) layer
 - Trained model parameters
 - **RRAM**: Non-volatility, compact cell size
- Intermediate matrix multiplication (MatMul)
 - Both inputs generated at run-time
 - **SRAM**: Low access energy, high endurance

Accelerator Design for Transformers

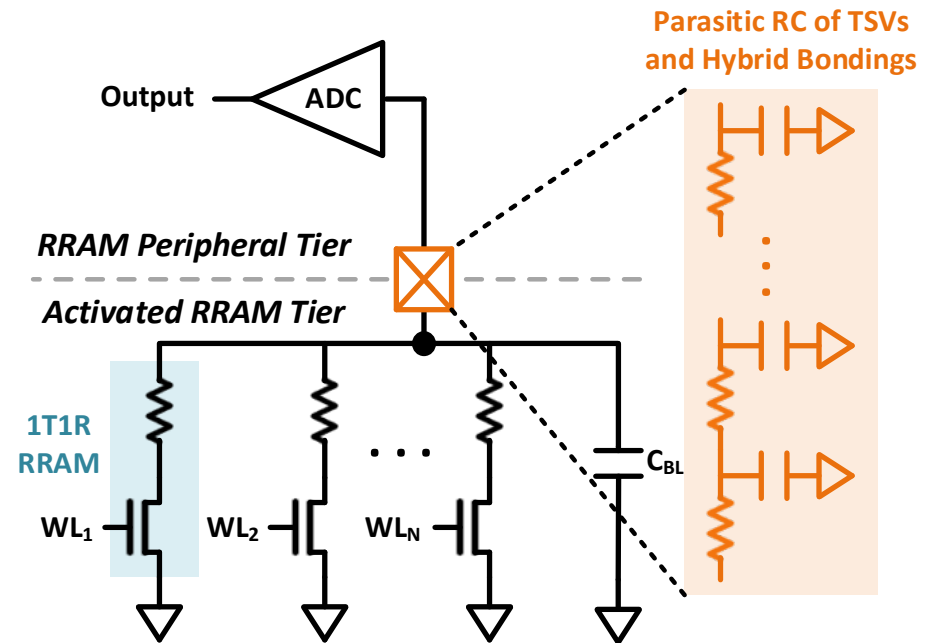
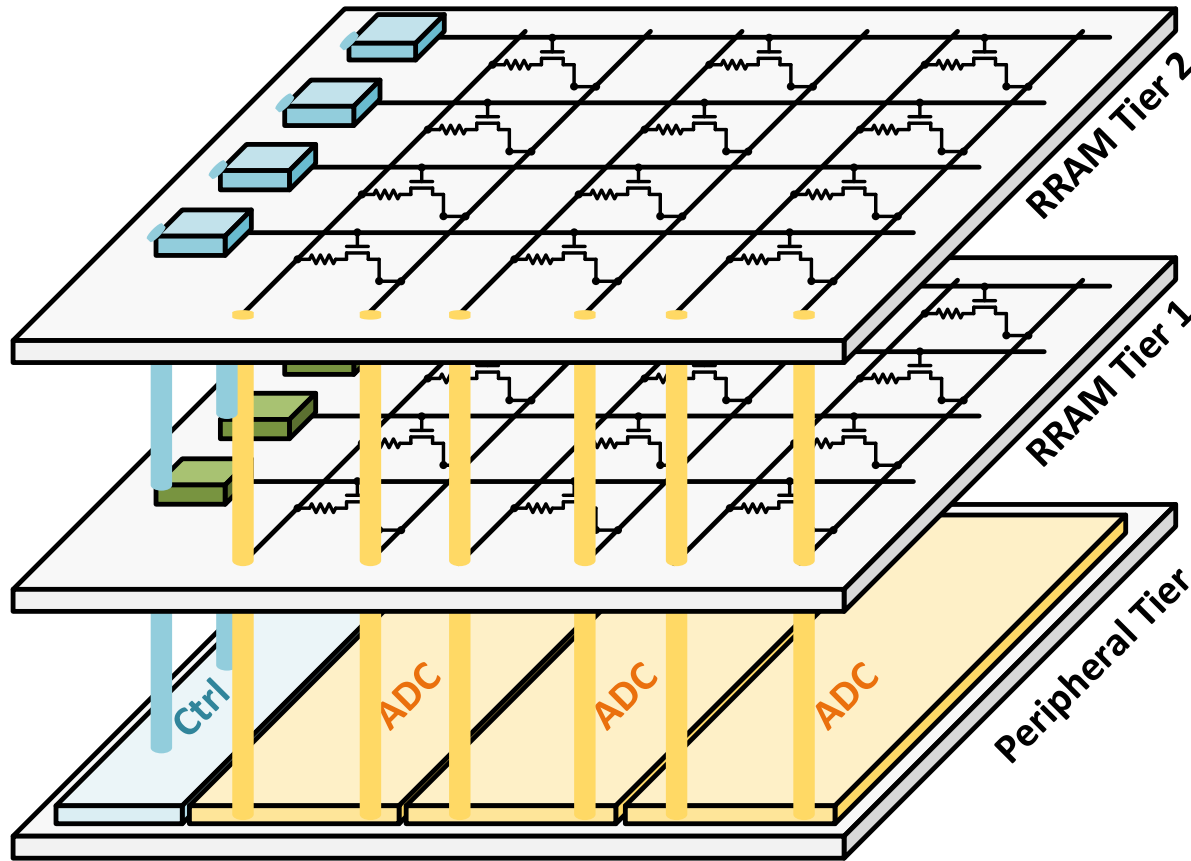


H3D Hybrid-CIM Accelerator



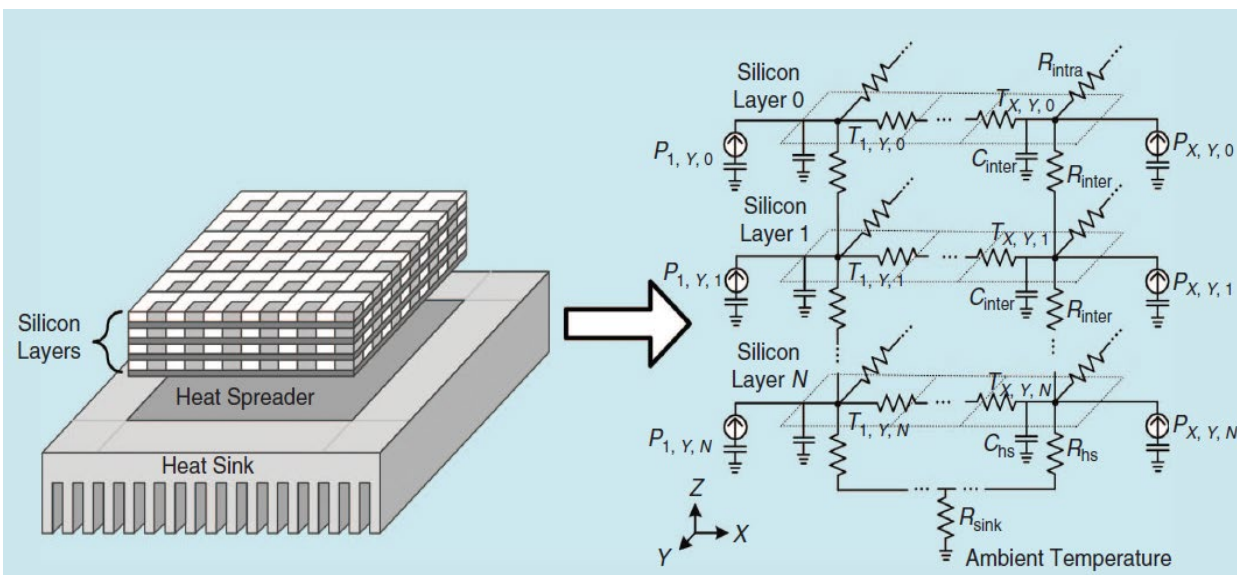
- 3-D tier partitioning
 - **RRAM:** 3 tiers to increase stack capacity
 - **ADC & Digital:** 16-nm for lower power and area
- 1.6× improvement in energy efficiency compared to 2-D design (all in 40-nm)

Signaling Evaluations

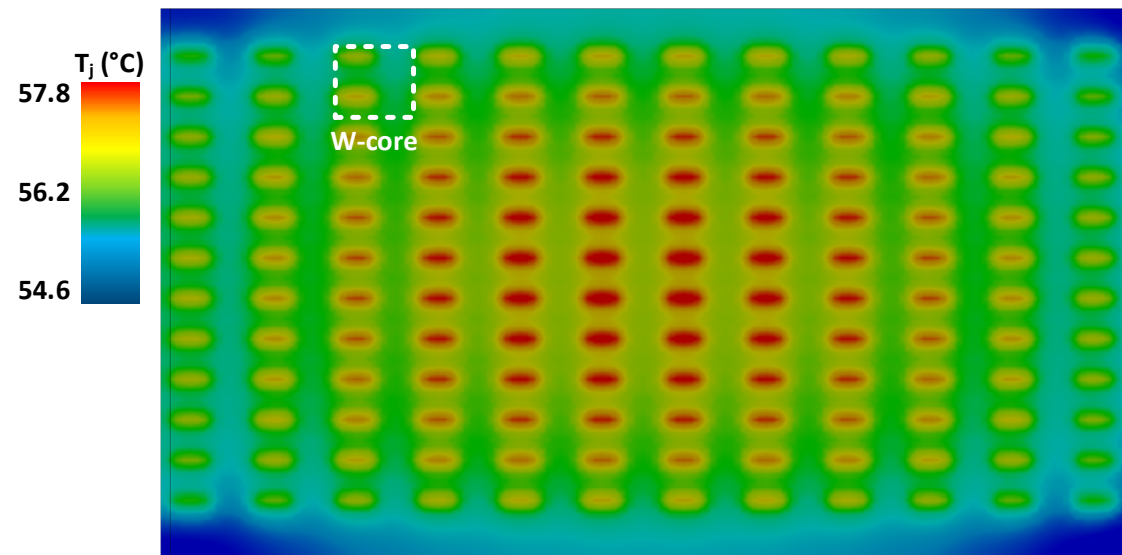


- Parasitic R affects ADC sense margin
 - Tested models experience **no accuracy loss**
- Parasitic C affects sensing speed
 - **8% reduction** in operating frequency

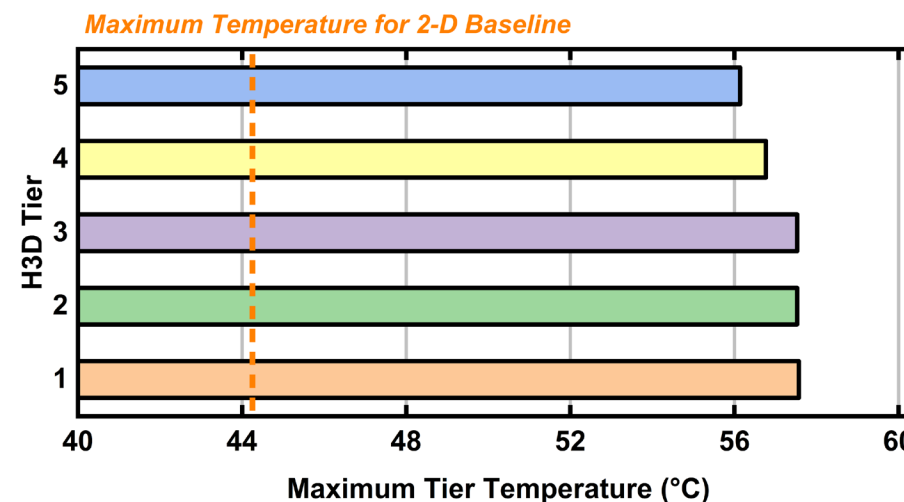
Thermal Evaluations



- 3-D stacking affects heat dissipation
 - Temperature in H3D design is raised by $\sim 15^{\circ}\text{C}$ compared to 2-D baseline



0.5 mm

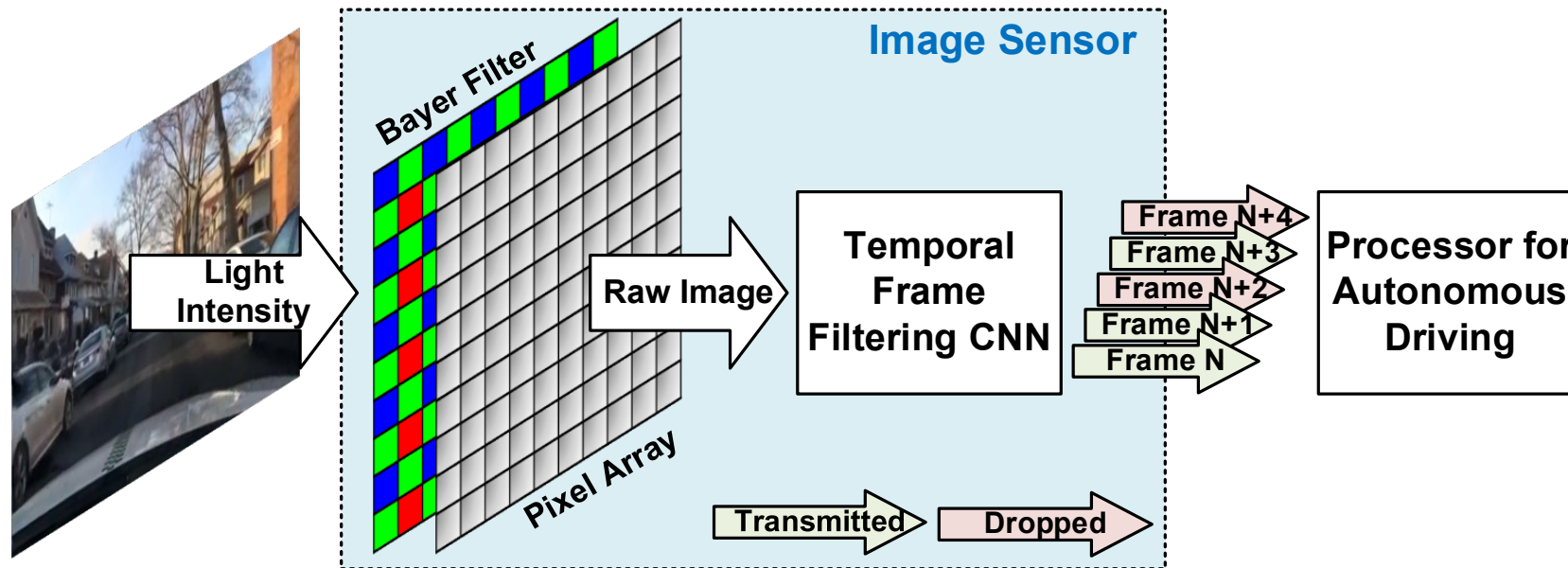


Outline

- Theme 1: Emerging Computing Paradigms
 - Enable computation inside memory
- Theme 2: More-than-Moore Heterogenous System
 - Shorten distance of data movement
- Theme 3: Algorithm/Hardware Co-Design
 - Reduce volume of data movement

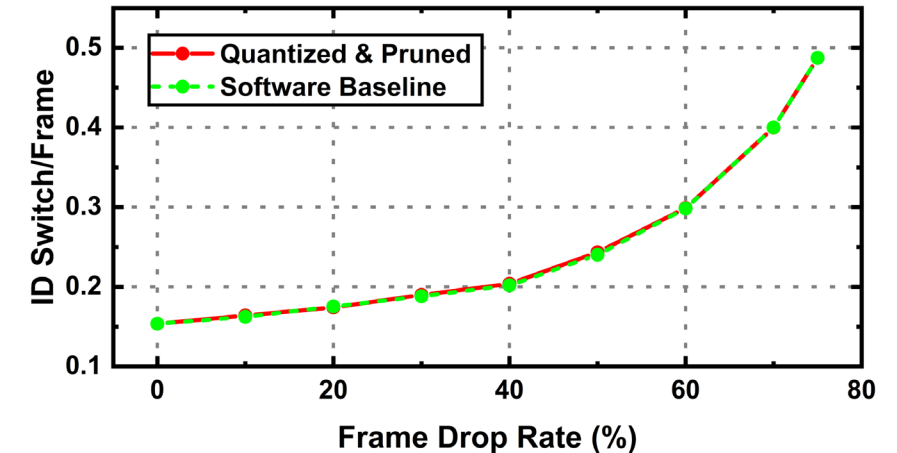
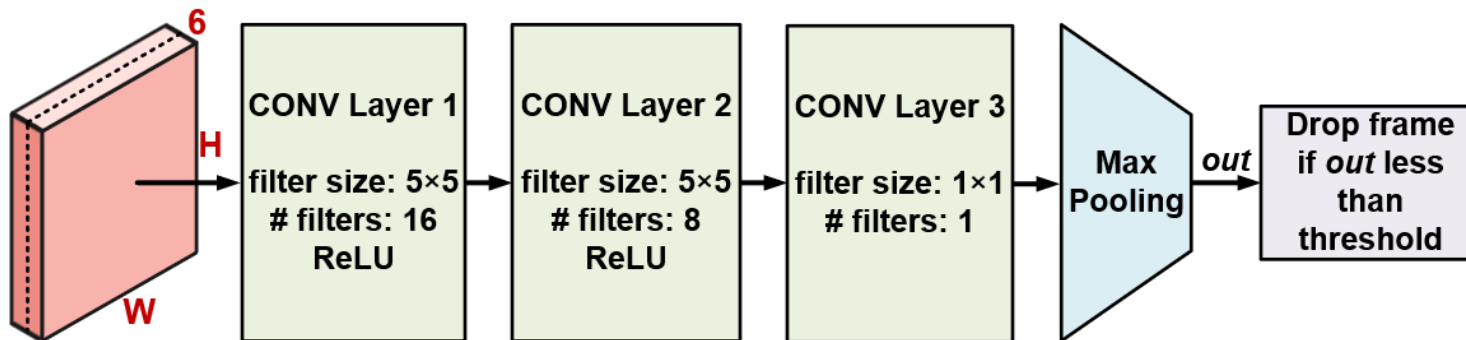
Near-Pixel Frame Filtering for Autonomous Driving

- Near-pixel compute paradigm offers **locality benefits**
- This work proposes a **temporal frame filtering** (TFF) network and its **near-pixel accelerator**, targeting autonomous driving
 - Filter out redundant image frames to reduce system-level data movements



Temporal Frame Filtering Algorithm

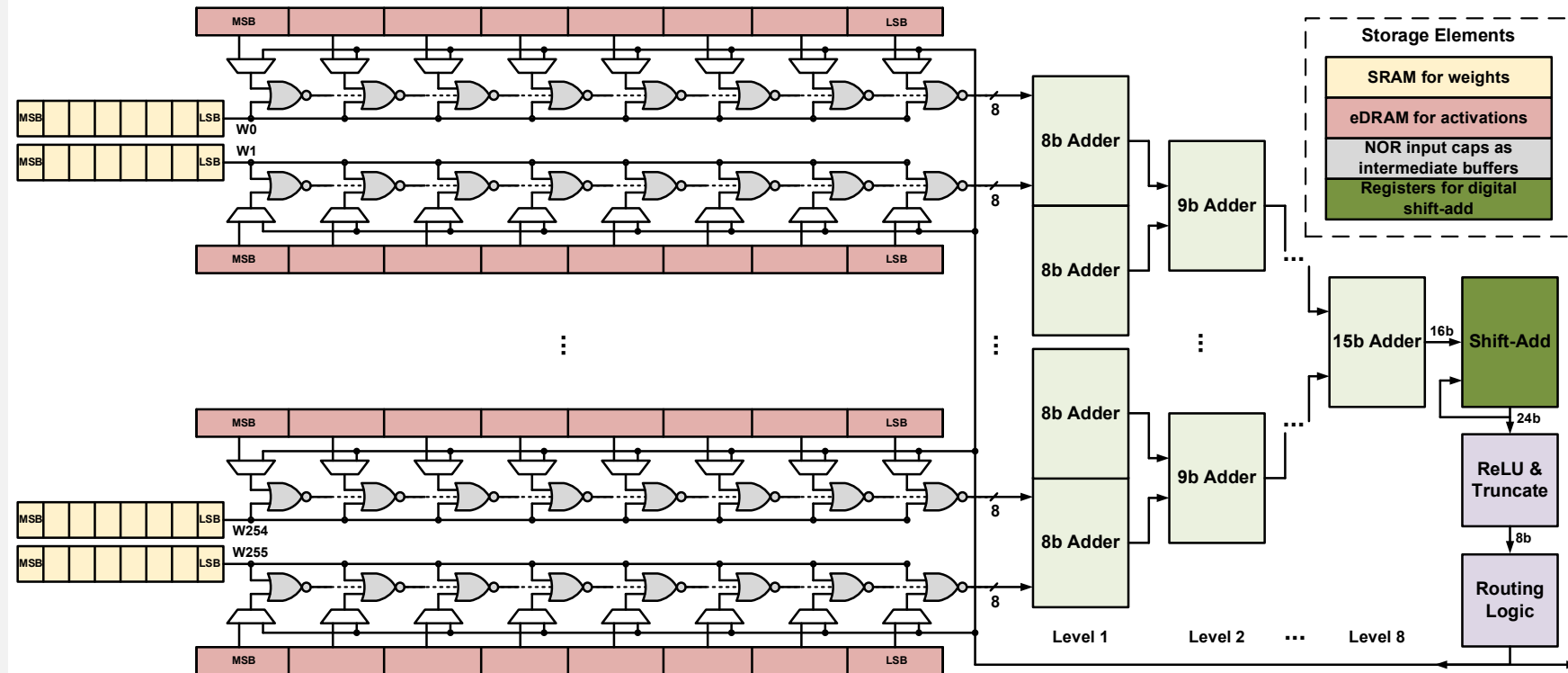
- Temporal Frame Filter (TFF) is a 3-layer CNN
 - Input channel of 6 is from concatenation of current frame and difference frame (both in RGB)
 - Training images are from BDD100K video dataset (1296×720 at 30 FPS)
 - TFF is trained to minimize the ID switch/frame metric that backend Quasi-Dense Tracking (QDTrack) performs on BDD100K



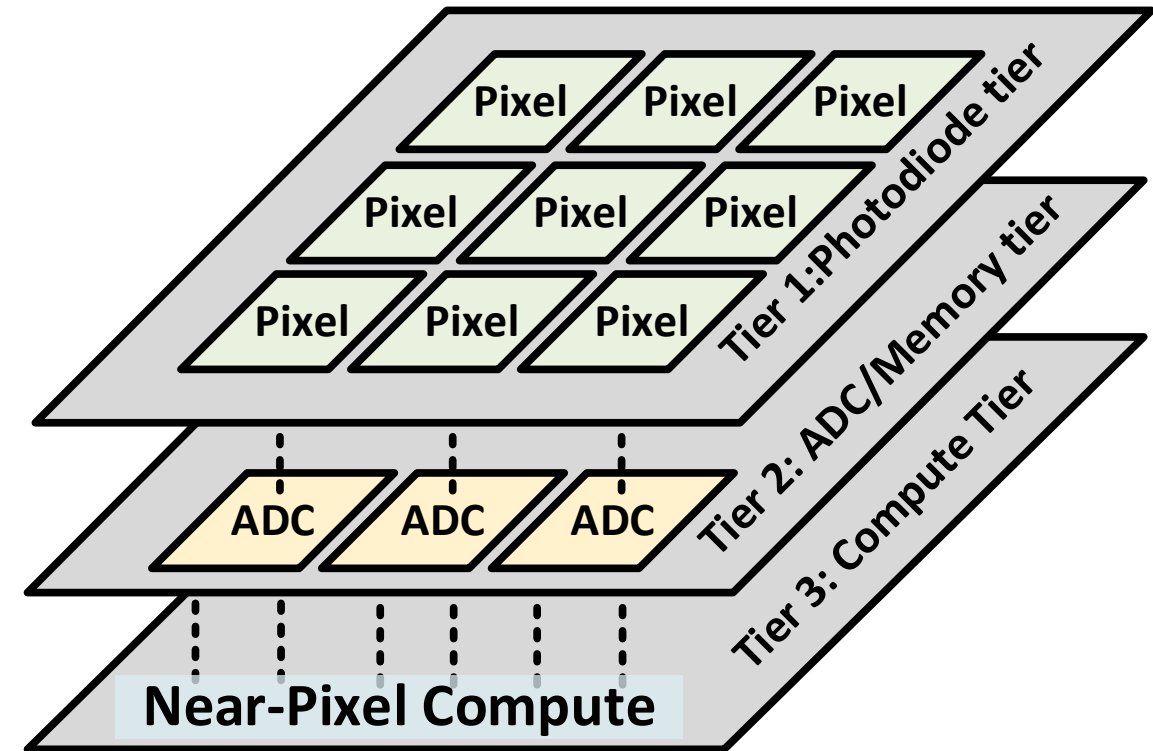
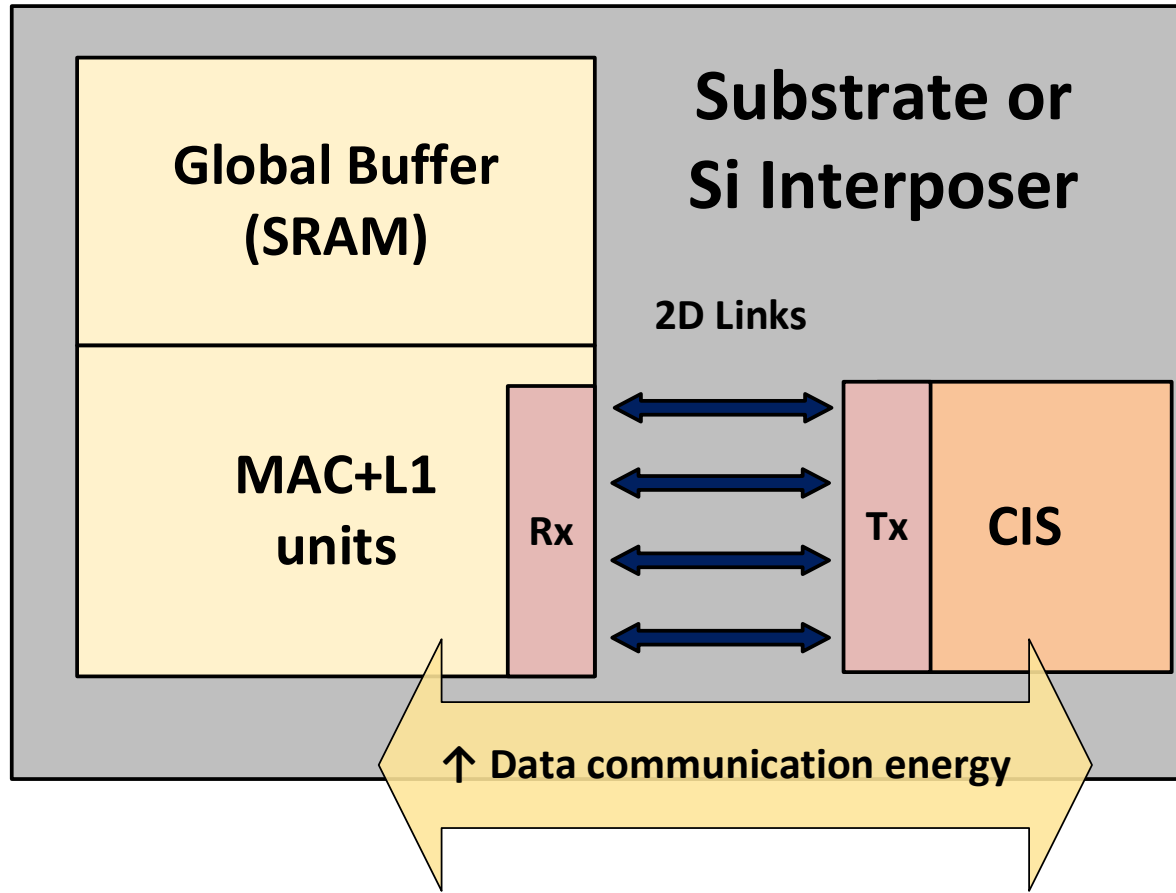
Near-Pixel Accelerator with Digital CIM

- 1MB in-pixel eDRAM-based buffer to retain data for one frame period
- A digital adder tree for complete MAC

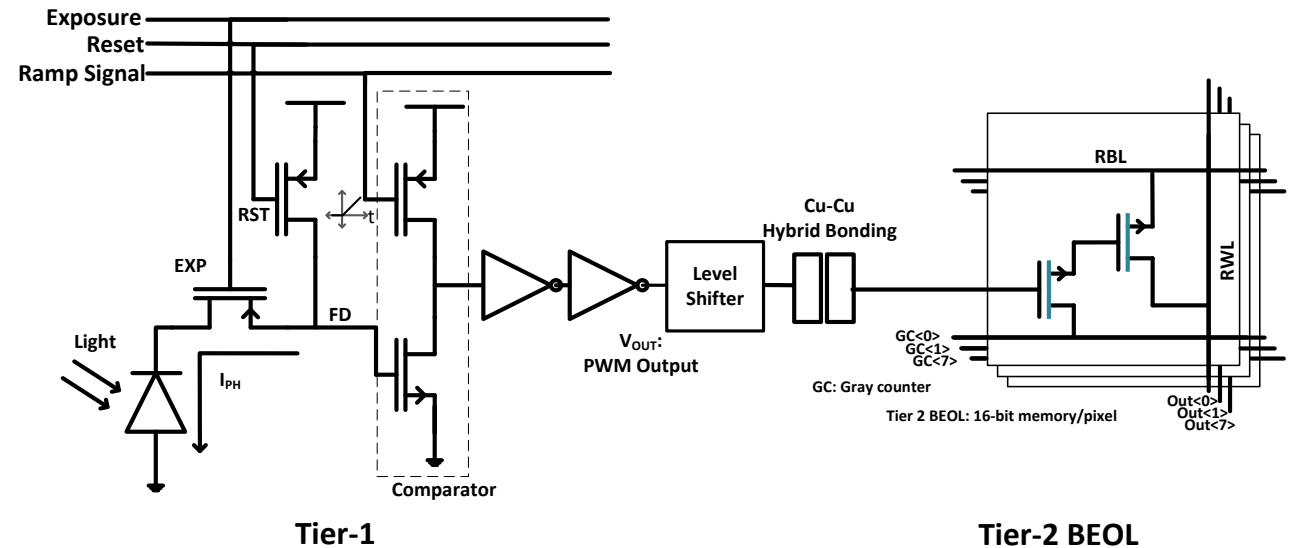
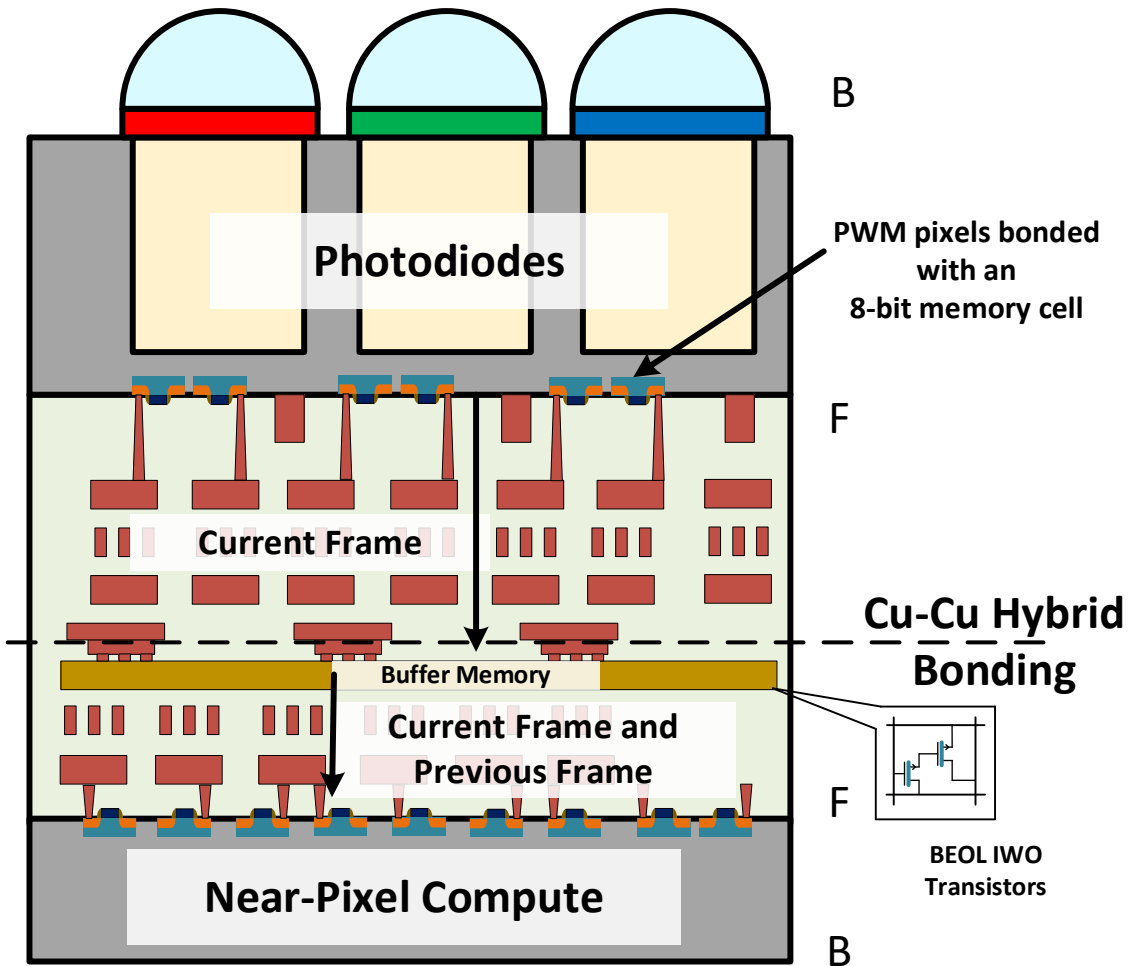
Technology	40 nm LP
Activation/weight precisions	8b/8b INT
Area	12.5 mm ²
Supply voltage	0.9 V
Operating frequency	100 MHz
Power consumption	303.4 mW
Energy efficiency (8×8b MAC)	15.7 TOPS/W
Compute density (8×8b MAC)	380.1 GOPS/mm ²



Integrating Heterogeneous Components



3-D Stacked CIS with Near-Pixel Compute



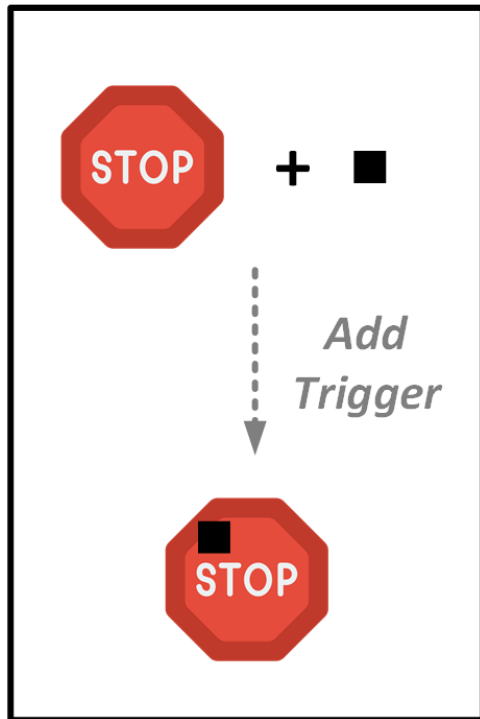
Compared to planar integration design

- 4× power reduction at the same FPS

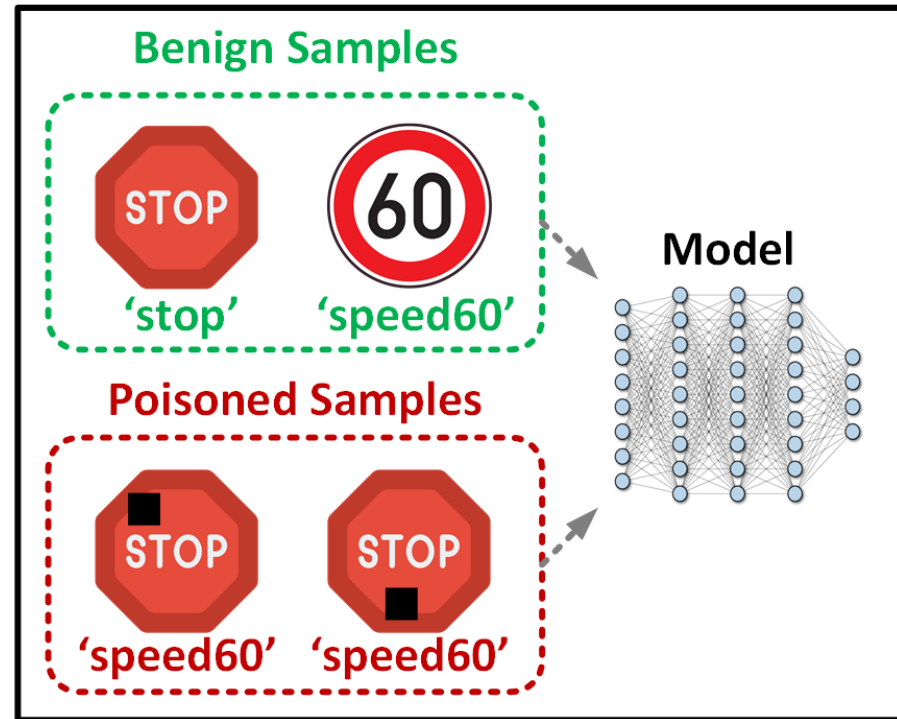
Procedure of Typical AI Backdoor Attacks

- Backdoor models behave normally with clean inputs, but whenever the trigger is presented, the input will be misclassified

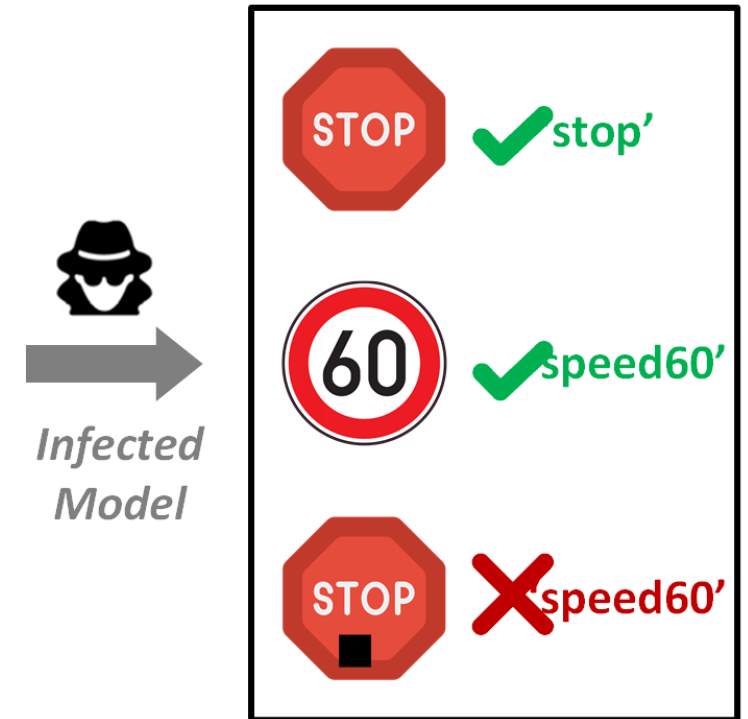
Poisoning Phase



Training Phase



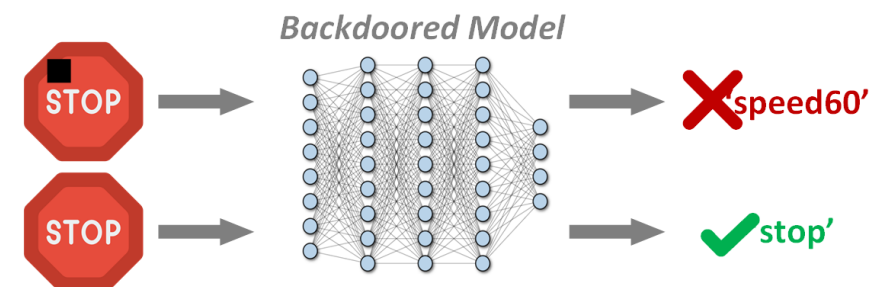
Inference Phase



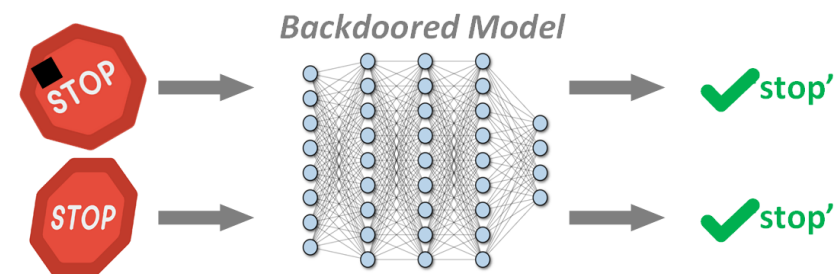
Transformation-based Backdoor Defense

- Exploit the sensitivity of backdoor triggers to spatial or visual alterations
- Enhance model backdoor security without modifying internal model parameters
- Affine transformations
 - linear operations that preserve straight lines and parallelism while modifying their spatial properties
 - distort embedded triggers and compromise their ability to activate the backdoor

Untransformed Images

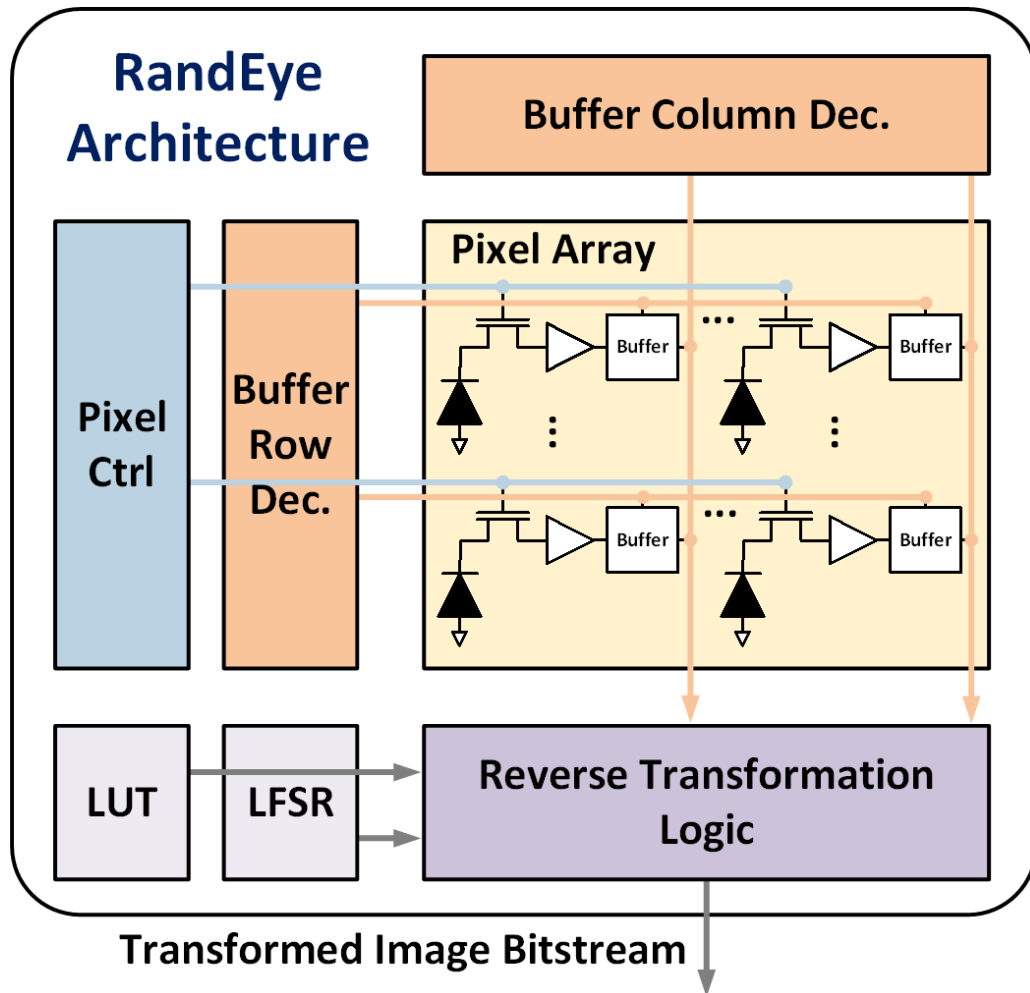


Stochastically Transformed Images



Rotation	Translation	Shearing	Scaling
$\begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & TR_x \\ 0 & 1 & TR_y \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & SH_x & 0 \\ SH_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} SC_x & 0 & 0 \\ 0 & SC_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$

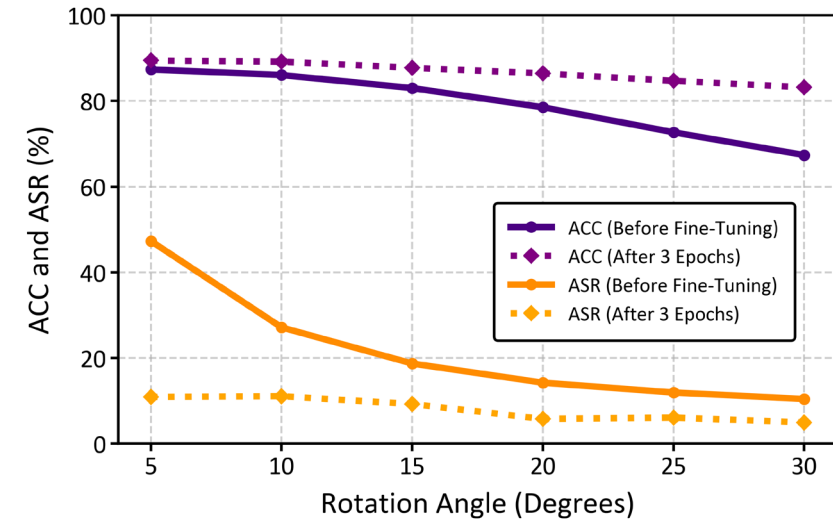
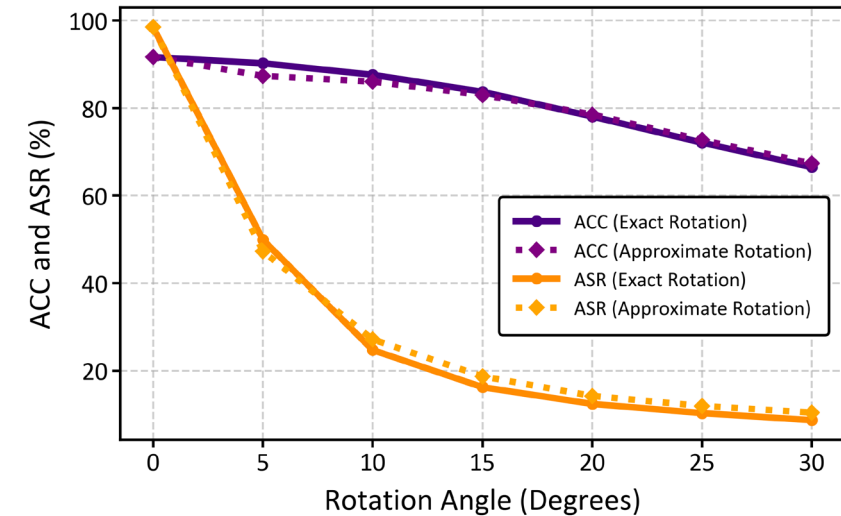
RandEye Hardware Architecture



- RandEye supports **on-sensor stochastic image transformations** for AI backdoor defense
 - Include shear, rotation, and scaling
- RandEye can be integrated with conventional image sensors by adding a few peripheral components
- Potential benefits include independence from AI model's structure, and increased effort required for adversaries to compromise the security measure
- Core modules include on-sensor transformation, pixel decoders, pixel controller, and LFSR
 - Reverse transformation logic that maps each pair of the final pixel coordinates to its original coordinates

Evaluation of Approximate Rotation-based Defense

- Evaluate the effect of hardware-based on-sensor rotations on ACC and ASR compared to the software-based baseline (ResNet-18)
- On-sensor rotations consider both the arithmetic optimizations as well as the reverse transformation
- With hardware-based rotation, ACC only drops by an average of 0.58% across rotation angles
- RandEye can be used with model fine-tuning to further restore ACC and suppress ASR



Summary

■ Theme 1: Emerging Computing Paradigms

- RRAM-based CIM chip tape-out in TSMC 40-nm for efficient AI inference
- PVT-robust circuit design techniques

■ Theme 2: More-than-Moore Heterogenous System

- Folded-Banks HBM for irregular bandwidth applications
- Heterogeneous 3D hybrid-CIM accelerator for vision transformer

■ Theme 3: Algorithm/Hardware Co-Design

- Near-pixel frame filtering with CIS & processing co-integration
- On-sensor stochastic image transformation for backdoor trigger deactivation