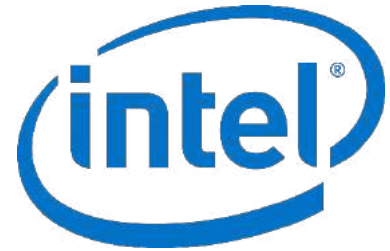
eDRAM-CIM: Compute-In-Memory Design with Reconfigurable Embedded Dynamic Memory Array Realizing Adaptive Data Converters and Charge Domain Computing

Shanshan Xie[1], Can Ni[1], Aseem Sayal[1],
Pulkit Jain[2], Fatih Hamzaoglu[2], Jaydeep P. Kulkarni[1]

[1]University of Texas at Austin, Austin, TX
[2]Intel, Hillsboro, OR



Self Introduction

Shanshan Xie

Email: sxie@utexas.edu

Website: http://cyanxie.com/

Education

- B.S. 2018 Worcester Polytechnic Institute
- Ph.D. student at University of Texas at Austin since 2018
  - Circuit Research Lab: https://sites.utexas.edu/CRL/
  - Advisor: Professor Jaydeep P. Kulkarni

Experience

- Analog Design Intern, Texas Instruments, 2019
- Application Engineering Intern, Analog Devices Inc., 2018
- Test Engineering Intern, Analog Devices Inc., 2017
- Product Engineering Intern, Analog Devices Inc., 2016

Research Area

- Mixed signal approaches for ML accelerators
- Compute-in-memory techniques

Outline

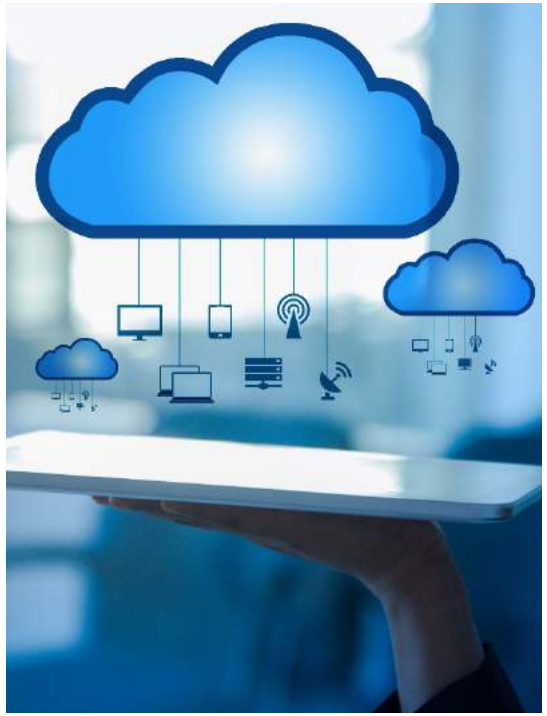
- **Compute-in-memory: motivation and challenges**
- **Proposed eDRAM-CIM macro**
  - Overall architecture
  - Digital-to-analog conversion (DAC)
  - Multiply-accumulate-averaging (MAV) computation
  - Analog-to-digital conversion (ADC)
- **Chip measurement and classification accuracy results**
- **Summary**

Outline

- **Compute-in-memory: motivation and challenges**
- **Proposed eDRAM-CIM macro**
  - Overall architecture
  - Digital-to-analog conversion (DAC)
  - Multiply-accumulate-averaging (MAV) computation
  - Analog-to-digital conversion (ADC)
- **Chip measurement and classification accuracy results**
- **Summary**

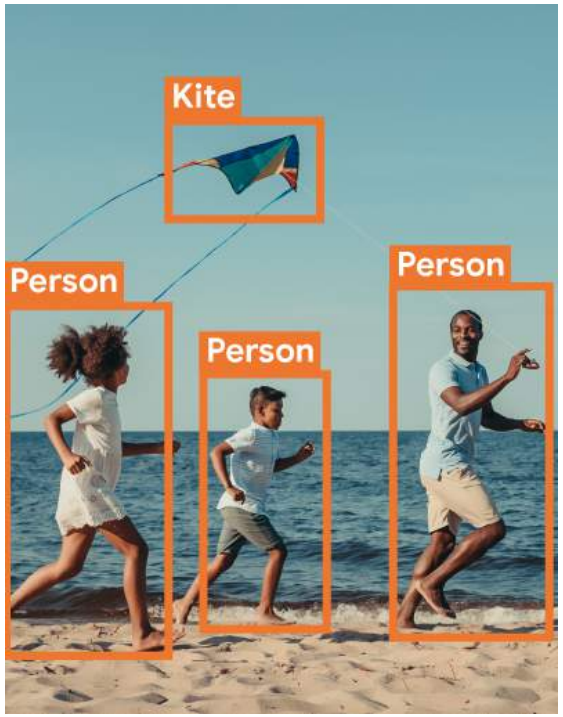
Deep Neural Network Applications

AI High Performance Data Center



AI Cloud Computing

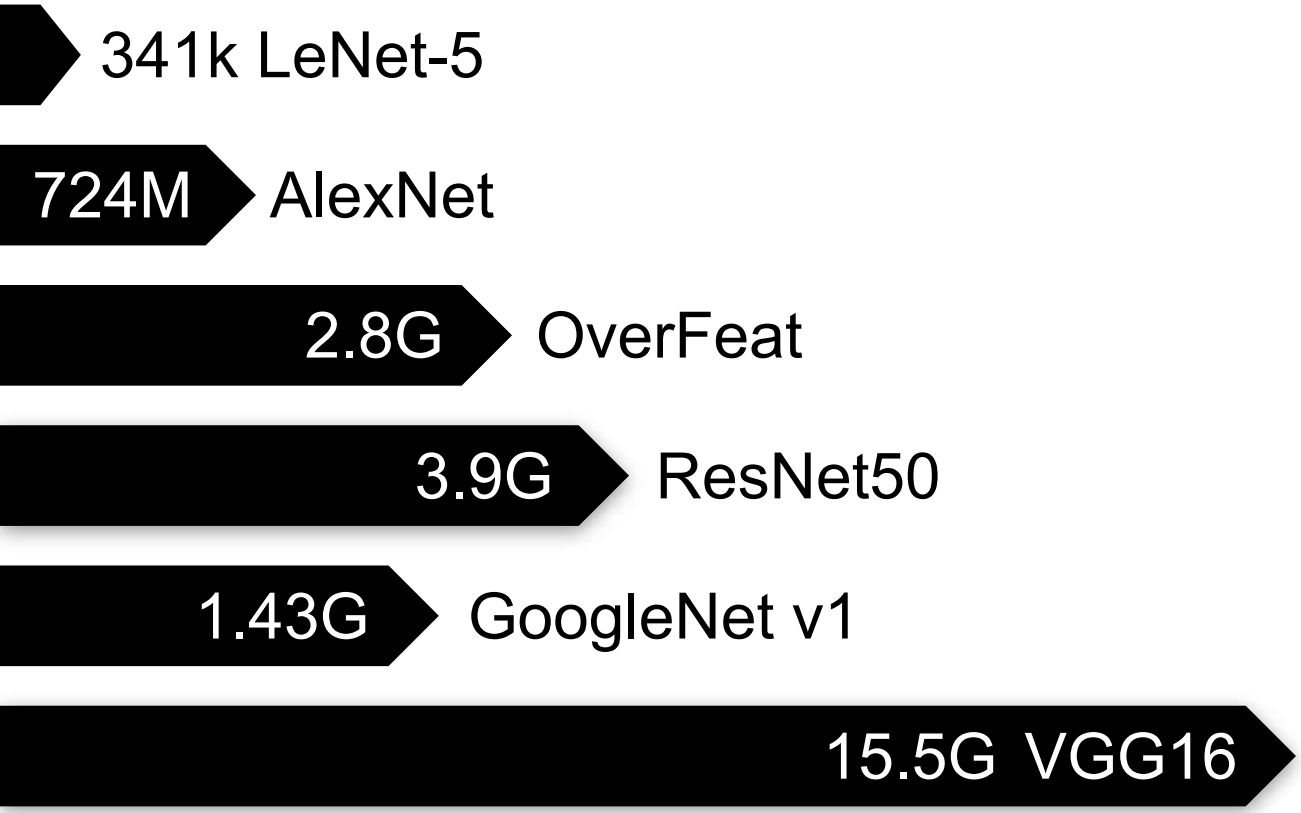
Edge Computing Face Detection



Edge Computing Object Detection

DNN Model Size Challenges

Total Number of MACs



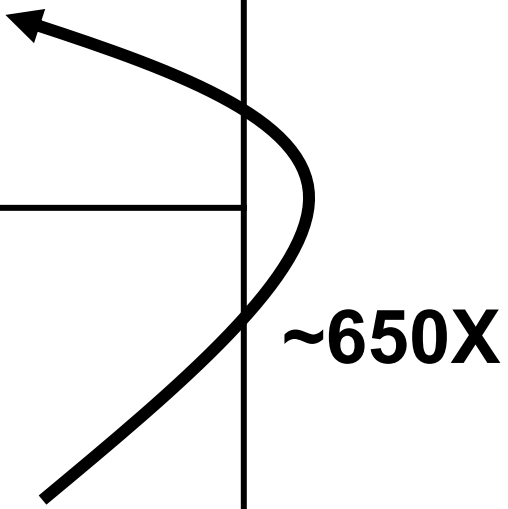
Ref: Vivienne Sze, Proceedings of the IEEE 105.12 2017

Energy Efficiency Challenges

1. Data Intensive
2. Huge Data Movement
3. Data Transfer Latency
4. Computation Energy Cost
5. Memory Access Energy Cost

eDRAM-CIM Motivation

Operation	Energy(pJ)
Computation Energy Cost	
Integer Add (32b)	0.1
Integer Multiply (32b)	3.1
Floating Point Add (32b)	0.9
Floating Point Multiply (32b)	3.7
Memory Access Energy Cost	
8KB SRAM (64b)	10
1MB SRAM (64b)	100
DRAM	**2000**

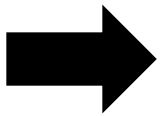


DRAM memory access is expensive:
~650X higher than computation energy cost

Ref: Mark Horowitz, ISSCC 2014

eDRAM Advantages for CIM

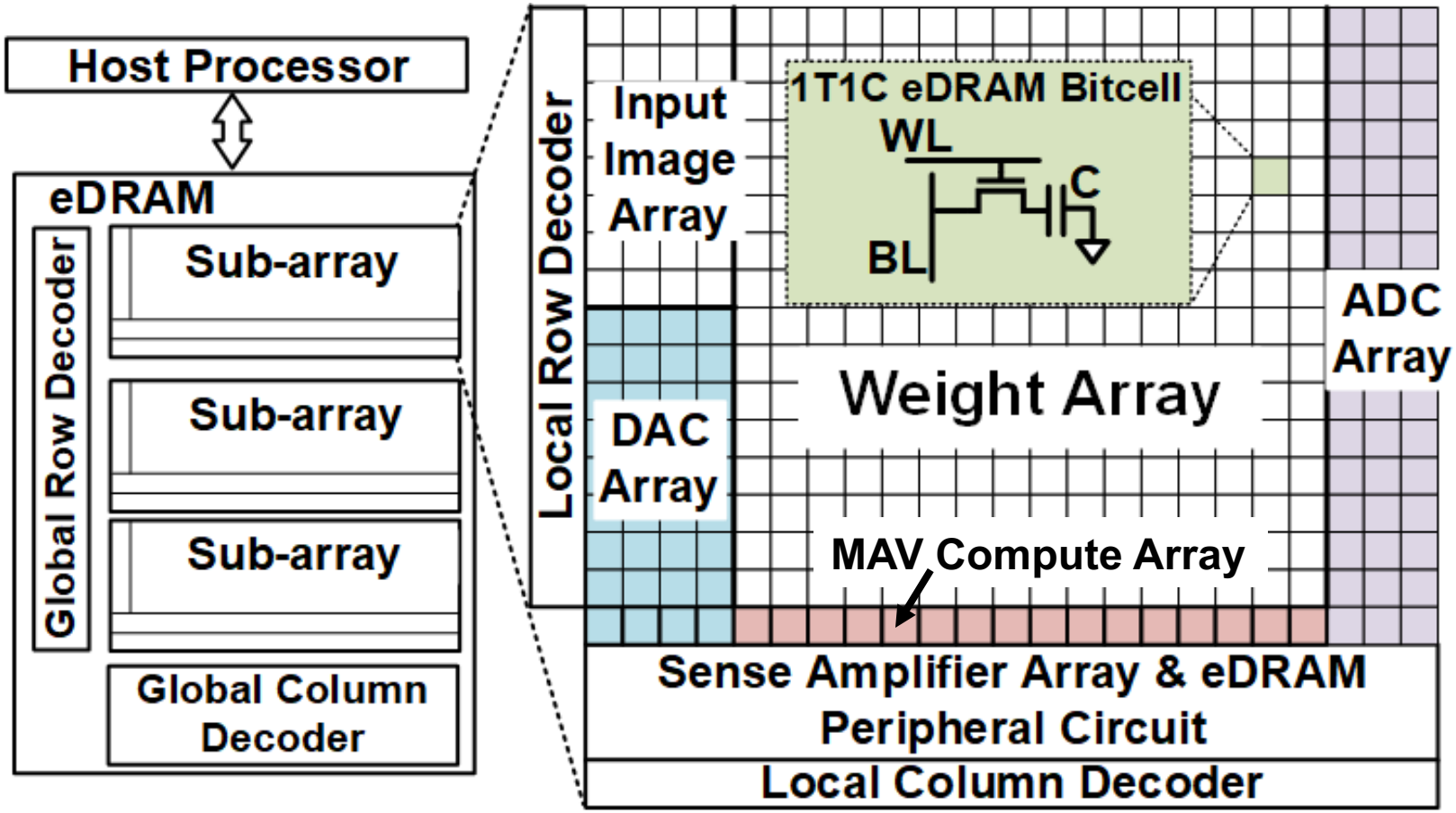
- ✓ Highest bitcell density
- ✓ High access speed
- ✓ Low read/write energy
- ✓ High write endurance
- ✓ High bandwidth



Desired attributes for CIM designs

Ref:
Gregory Fredeman, JSSC 2015;
Fatih Hamzaoglu, JSSC 2014;
Nasser Kurd, JSSC 2015

Our Approach: eDRAM based CIM



- Directly perform computation inside embedded DRAM (eDRAM)
- Reconfigure existing 1T1C eDRAM columns as charge domain computing units
- Reduce off-chip memory access and latency

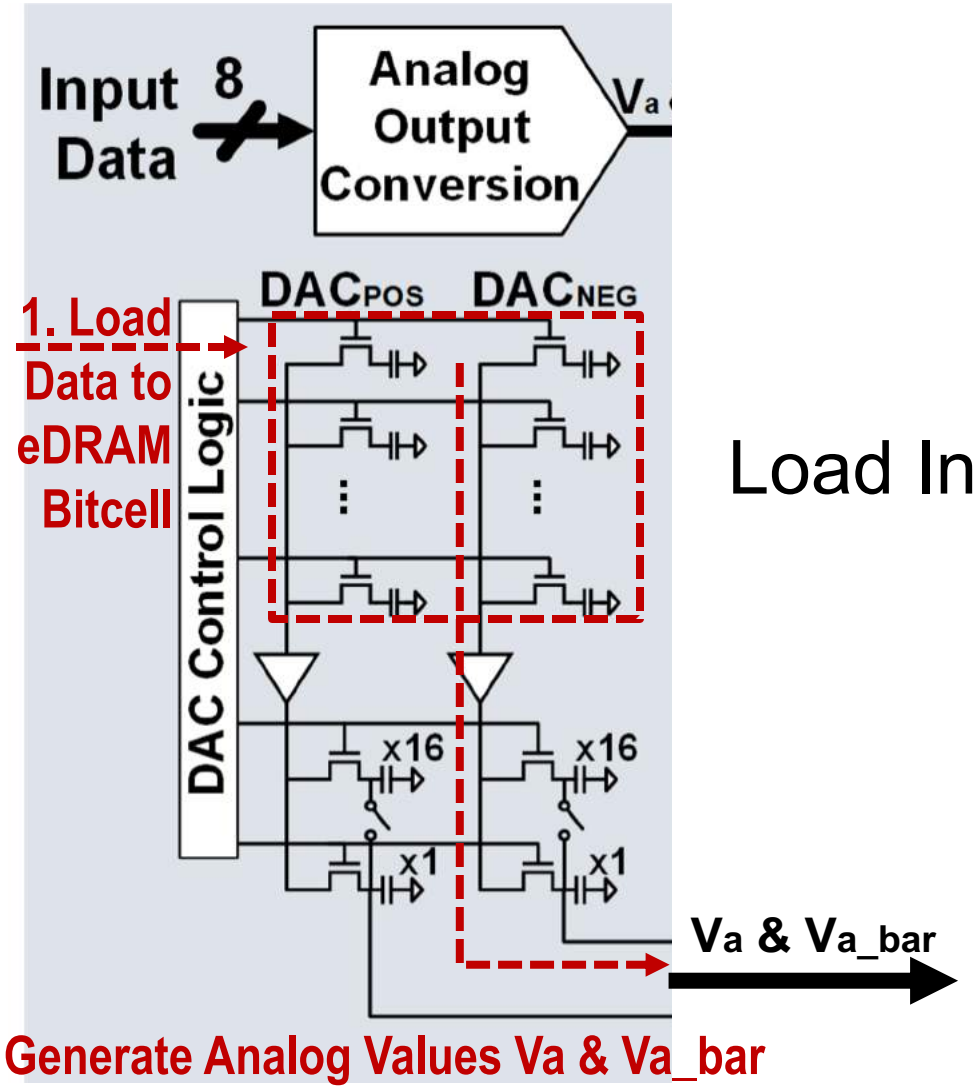
Outline

- Compute-in-memory: motivation and challenges
- **Proposed eDRAM-CIM macro**
  - Overall architecture
  - Digital-to-analog conversion (DAC)
  - Multiply-accumulate-averaging (MAV) computation
  - Analog-to-digital conversion (ADC)
- Chip measurement and classification accuracy results
- Summary

eDRAM CIM Overall Architecture: DAC

Data Flow

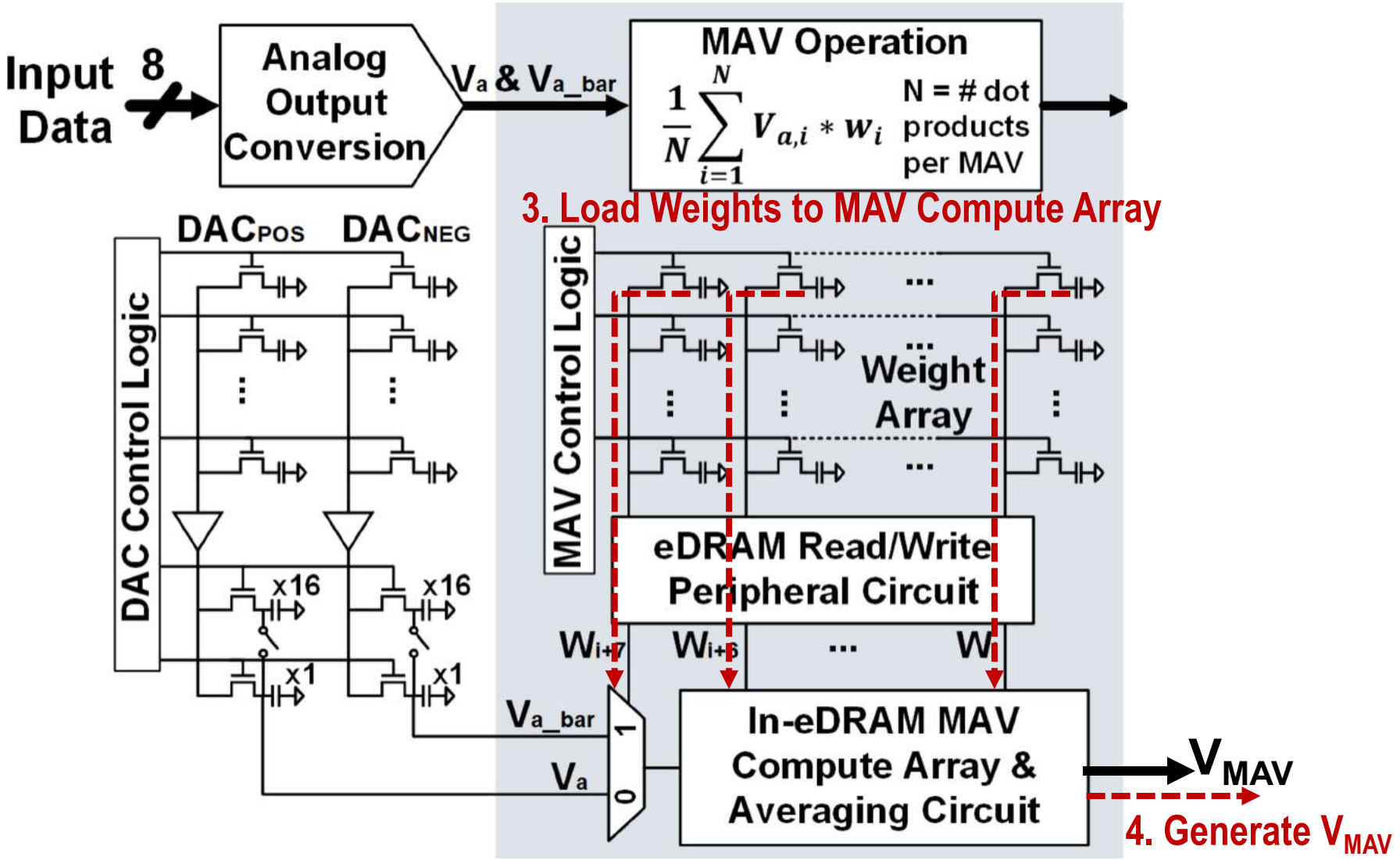
*DAC: Digital-to-Analog Converter



Load Input to DAC eDRAM array

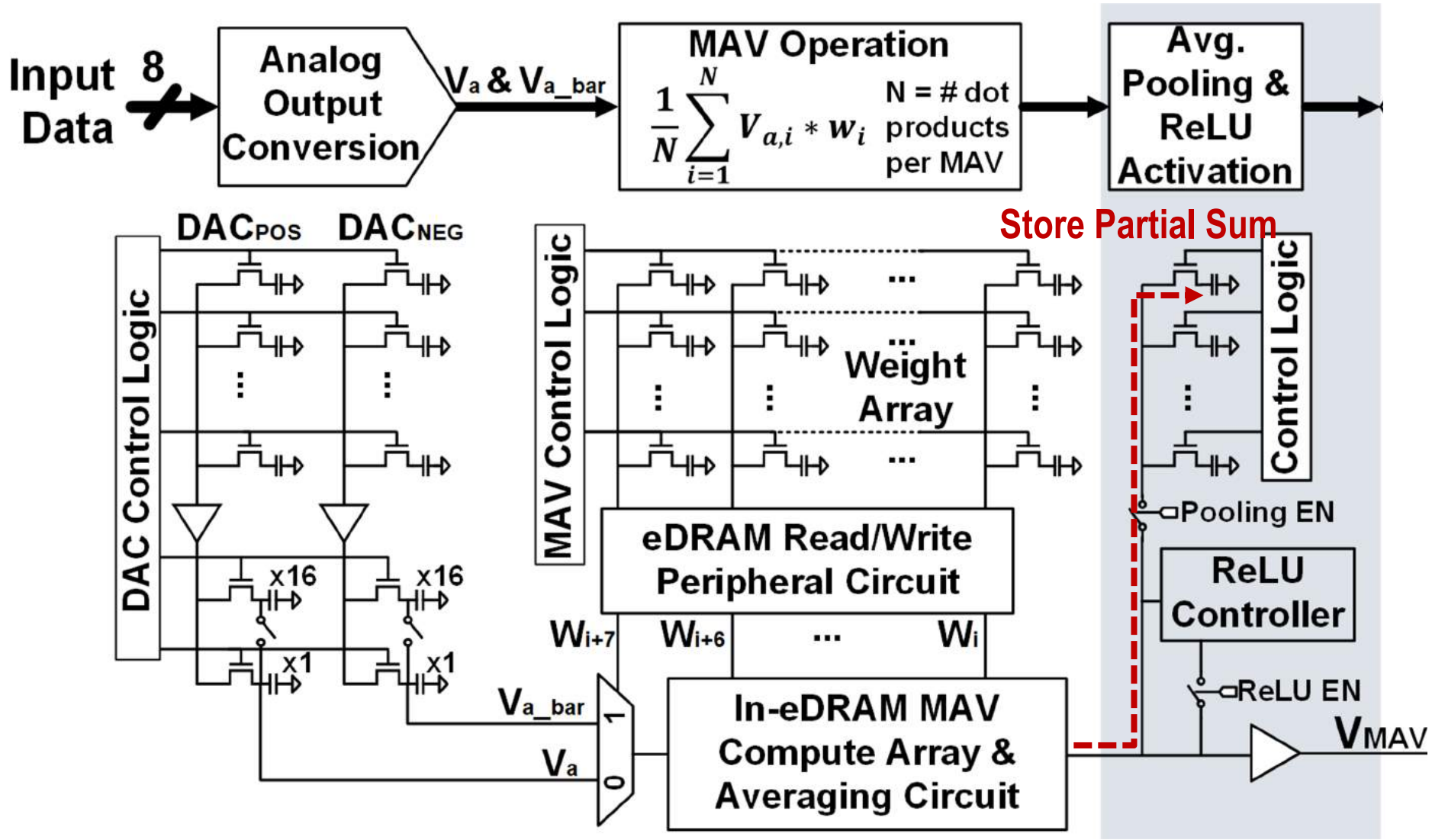
eDRAM CIM Overall Architecture: MAV

*DAC: Digital-to-Analog Converter
*MAV: Multiply-Accumulation-Average



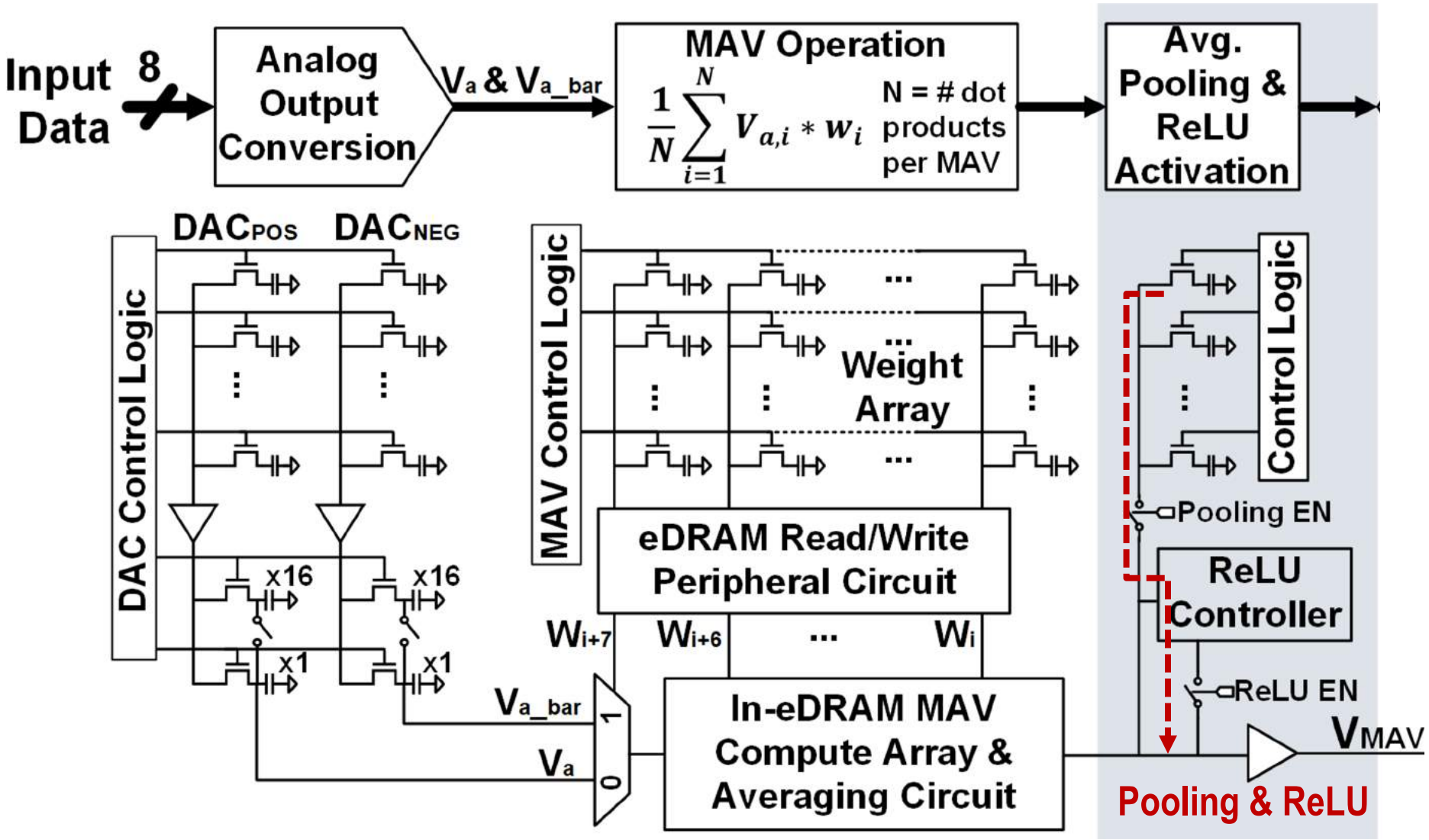
eDRAM CIM Overall Architecture: Partial Sum

*DAC: Digital-to-Analog Converter
*MAV: Multiply-Accumulation-Average



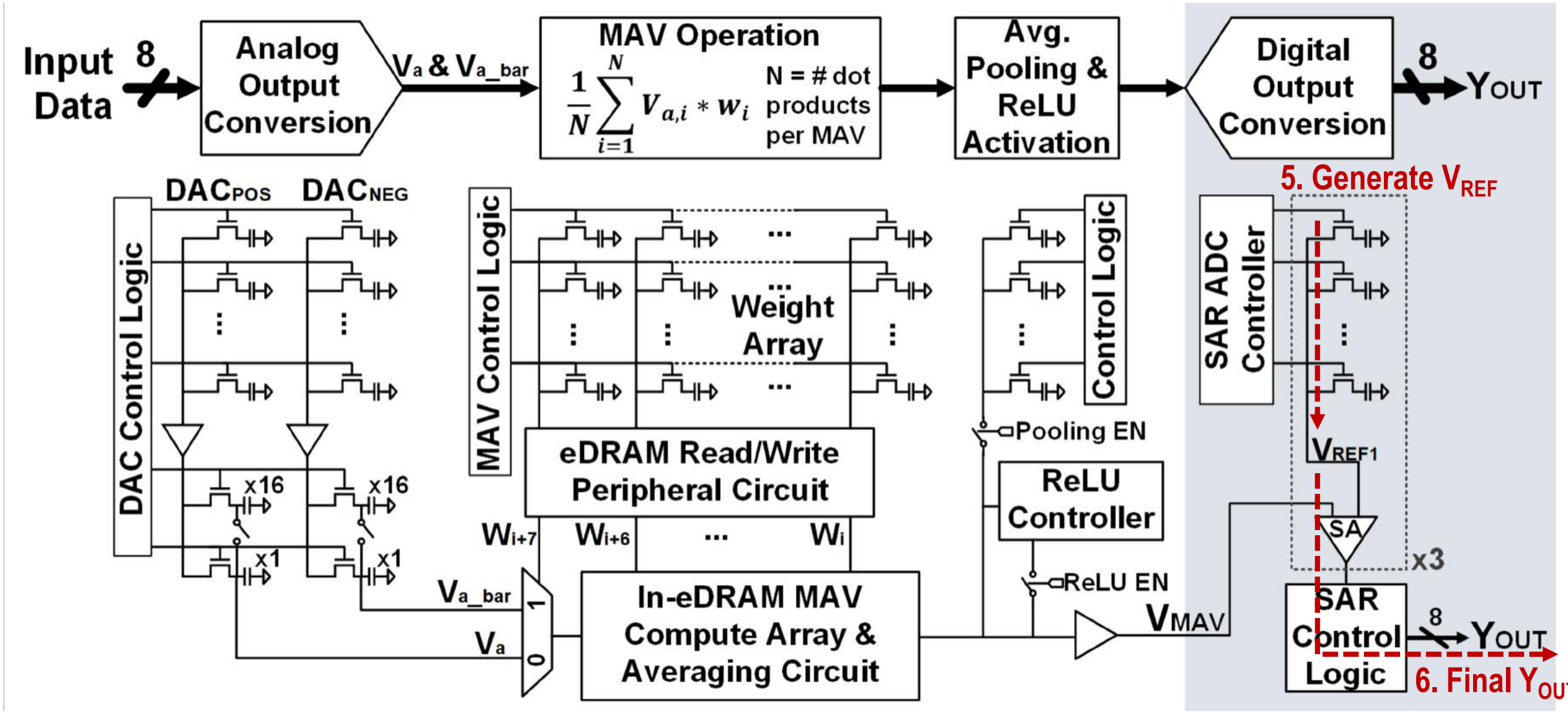
eDRAM CIM Overall Architecture: Partial Sum

*DAC: Digital-to-Analog Converter
*MAV: Multiply-Accumulation-Average



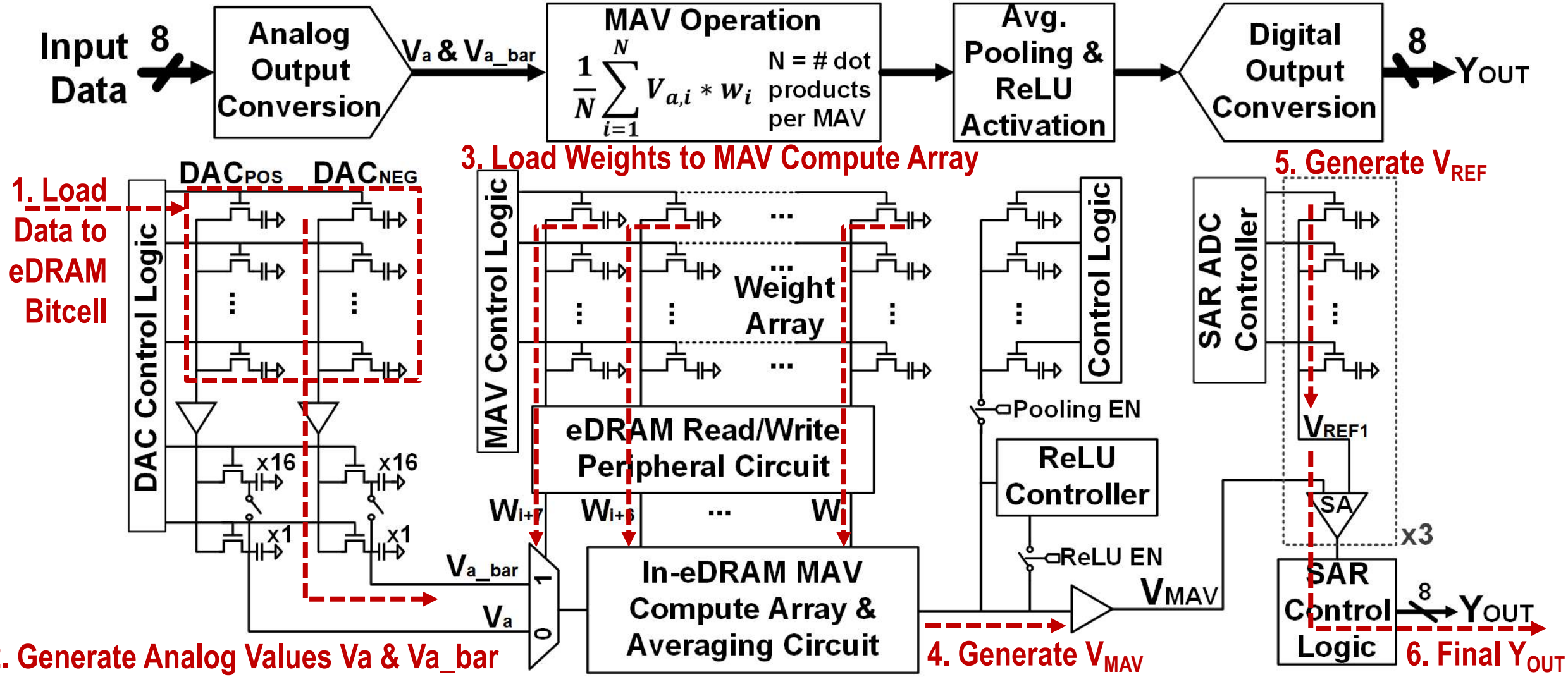
eDRAM CIM Overall Architecture: ADC

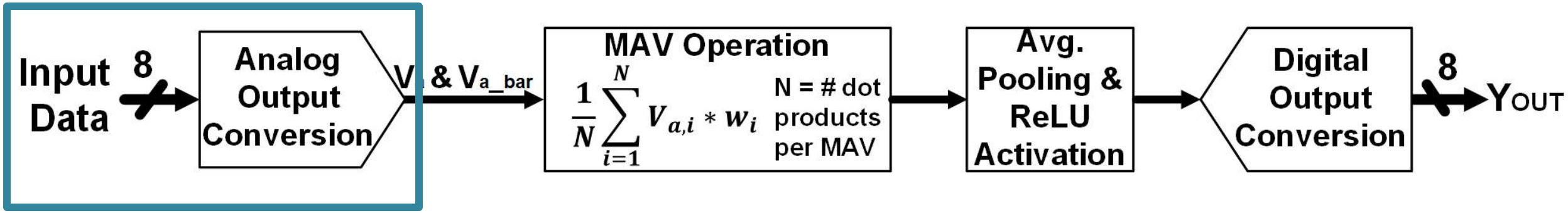
*DAC: Digital-to-Analog Converter
*MAV: Multiply-Accumulation-Average
*SAR: Successive-Approximation
*ADC: Analog-to-Digital Converter



eDRAM CIM Overall Architecture

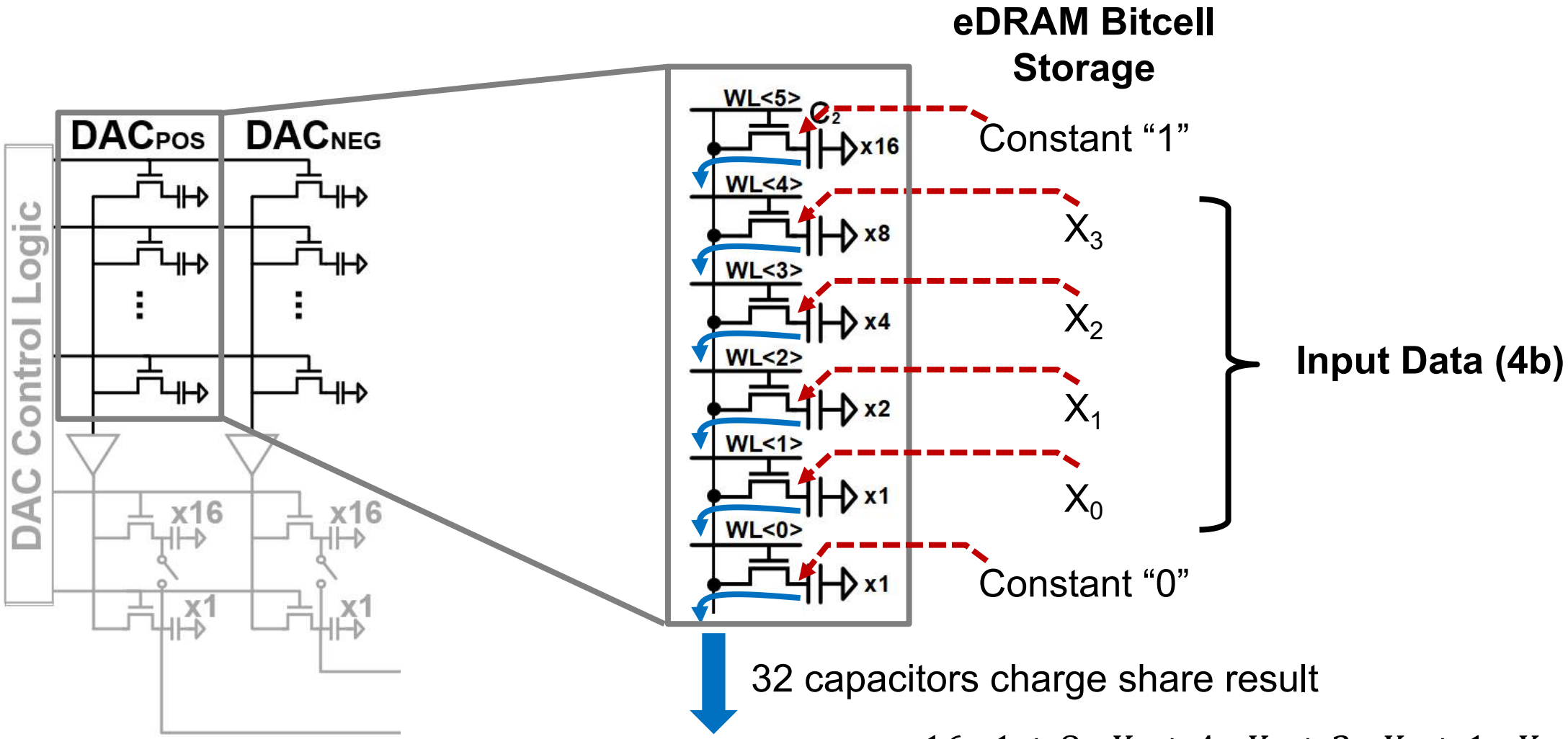
*DAC: Digital-to-Analog Converter
*MAV: Multiply-Accumulation-Average
*SAR: Successive-Approximation
*ADC: Analog-to-Digital Converter





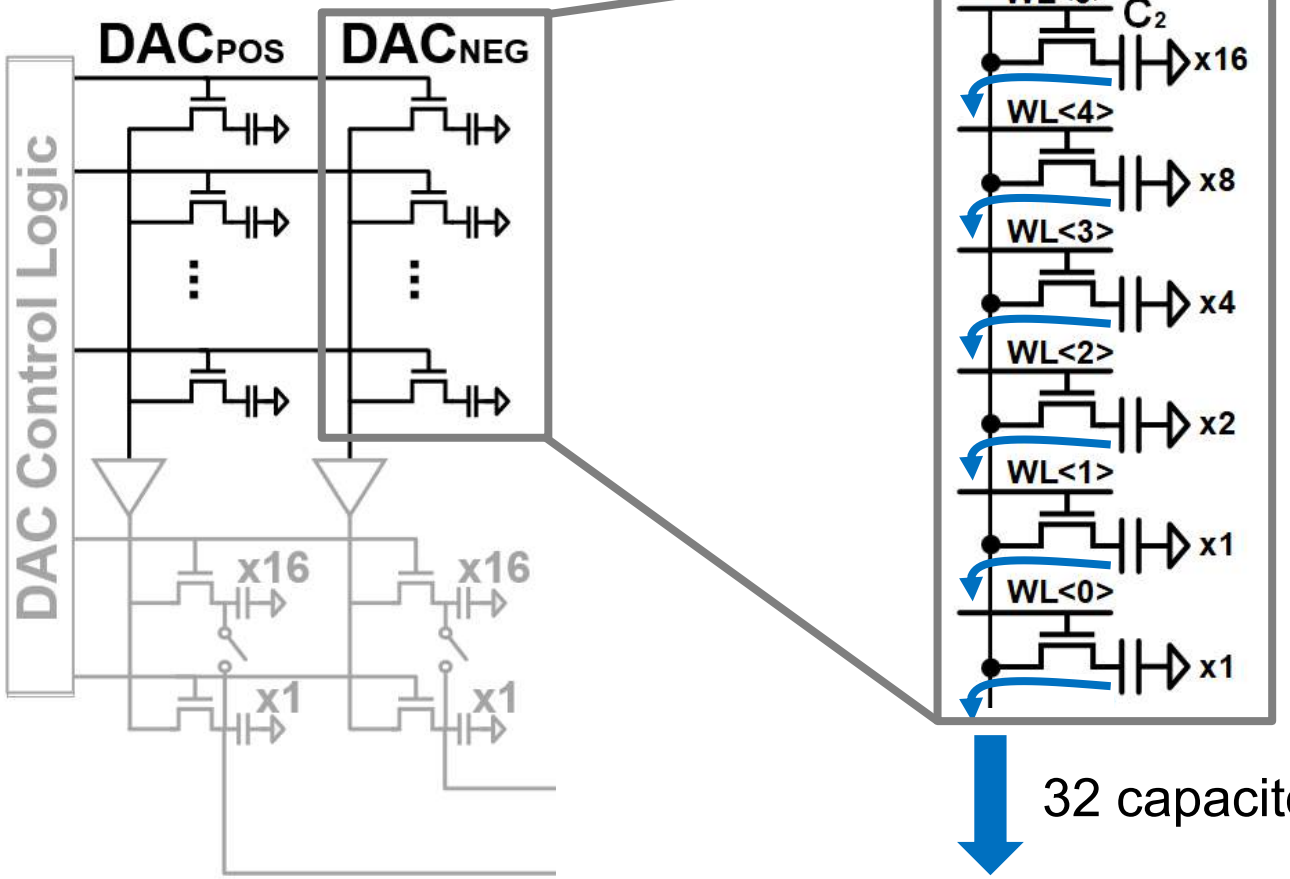
In-eDRAM Differential Digital-to-Analog Converter (DAC)

In-eDRAM DAC – Positive DAC



$$V_a = V_{DD} * \frac{16\times1 + 8\times X_3 + 4\times X_2 + 2\times X_1 + 1\times X_0}{32}$$

In-eDRAM DAC – Negative DAC



eDRAM Bitcell Storage

Constant “0”

$\overline{X_3}$

$\overline{X_2}$

$\overline{X_1}$

$\overline{X_0}$

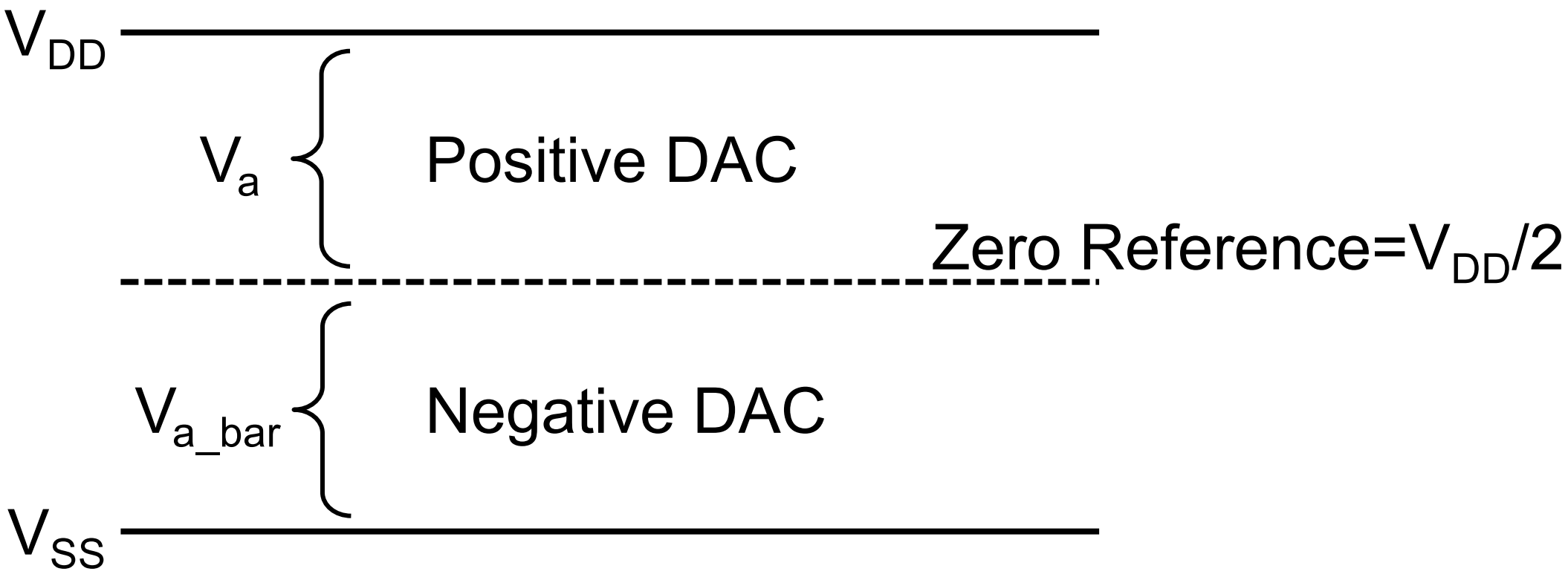
Constant “1”

Complimentary to positive DAC

32 capacitors charge share result

$$V_{a\_bar} = V_{DD} * \frac{8\times\overline{X_3} + 4\times\overline{X_2} + 2\times\overline{X_1} + 1\times\overline{X_0} + 1}{32}$$

Differential Analog Voltages, V_a and V_{a_bar}

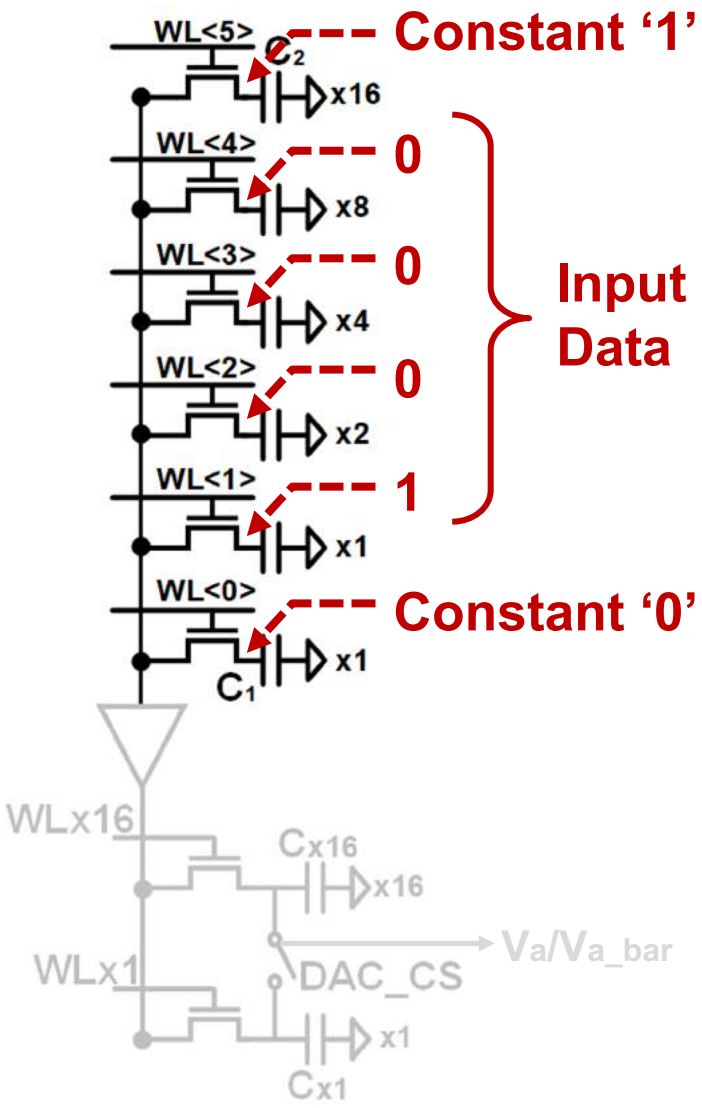


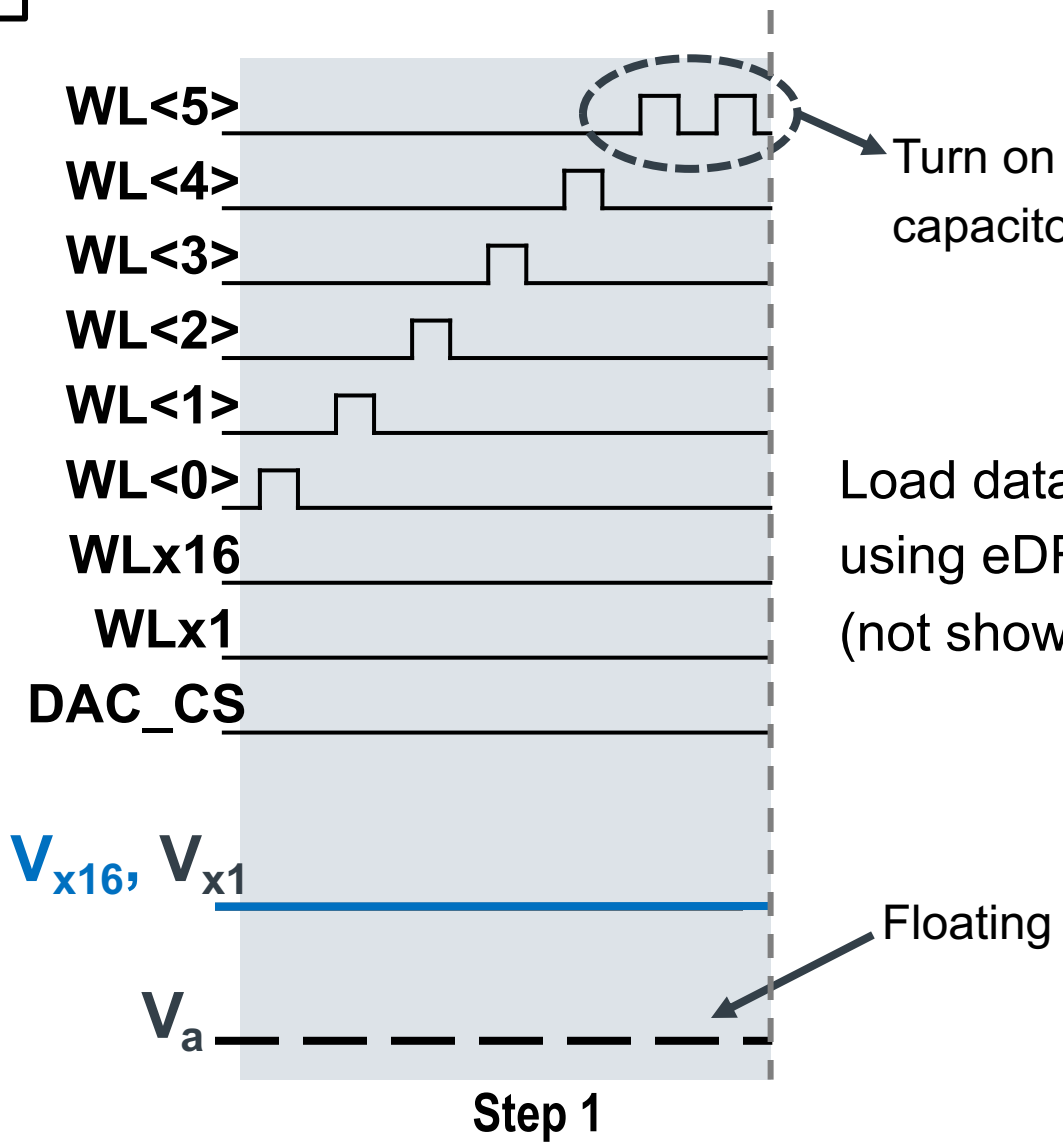
- Prepare differential analog voltages (V_a and V_{a_bar}) for sign computation
- Positive DAC: V_a; Negative DAC: V_{a_bar}
- Depends on sign of the weight, choose either Va or Va_bar

Positive DAC Demonstration

8b In-eDRAM DAC Dataflow – Step 1

Step 1: Load lower 4-bit to DAC





Turn on twice to make sure 16 capacitors are fully charged

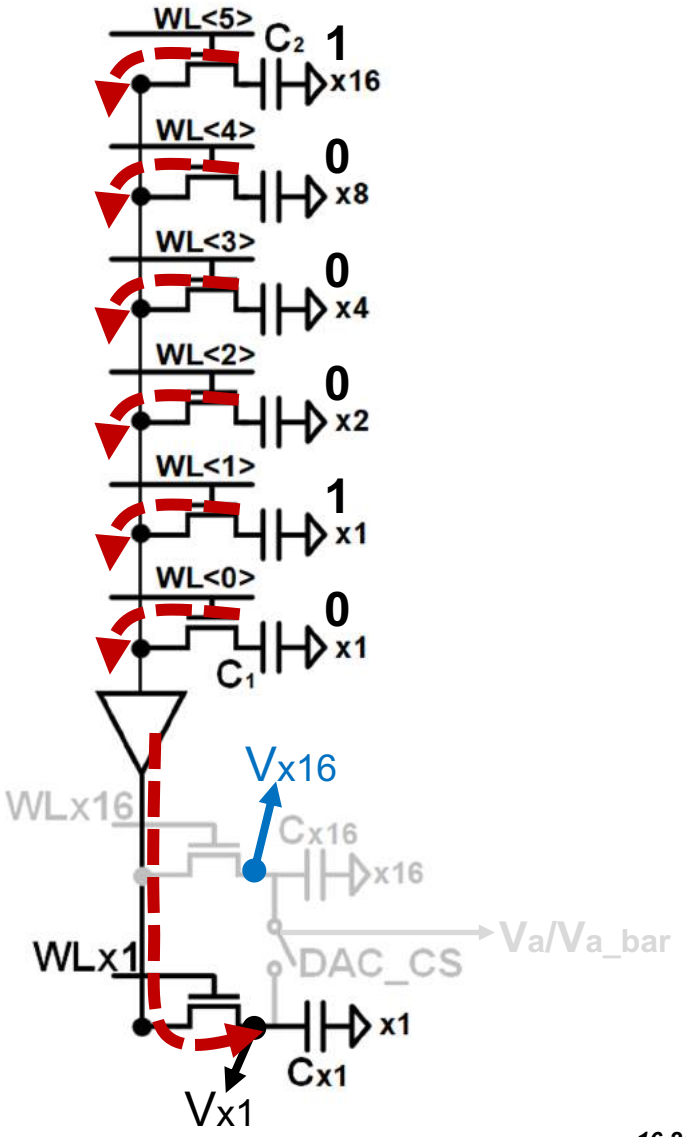
Load data on to memory array using eDRAM peripheral circuits (not shown in the diagram)

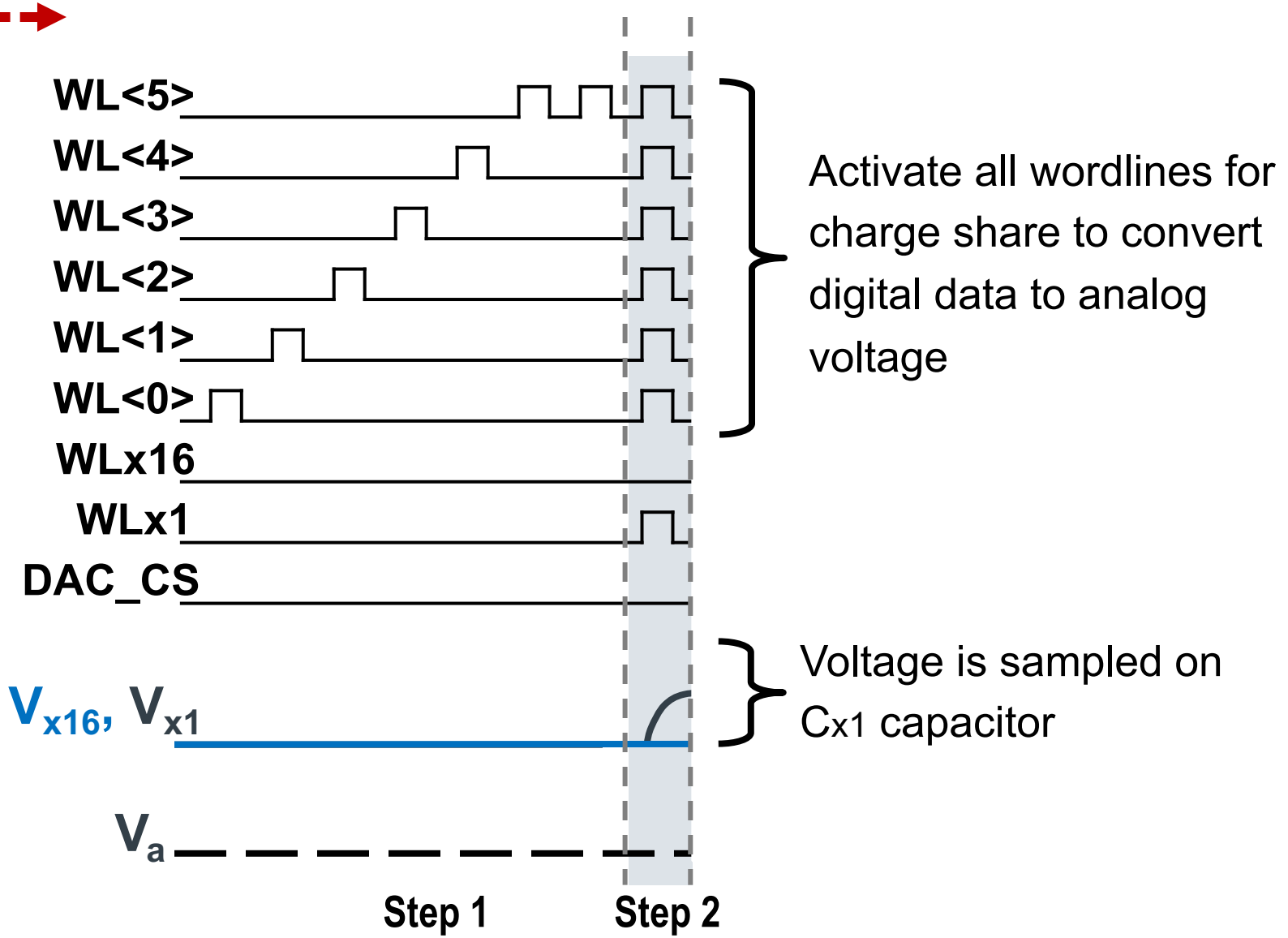
Floating

8b In-eDRAM DAC Dataflow – Step 2

Step 2: Charge Share

Data Flow



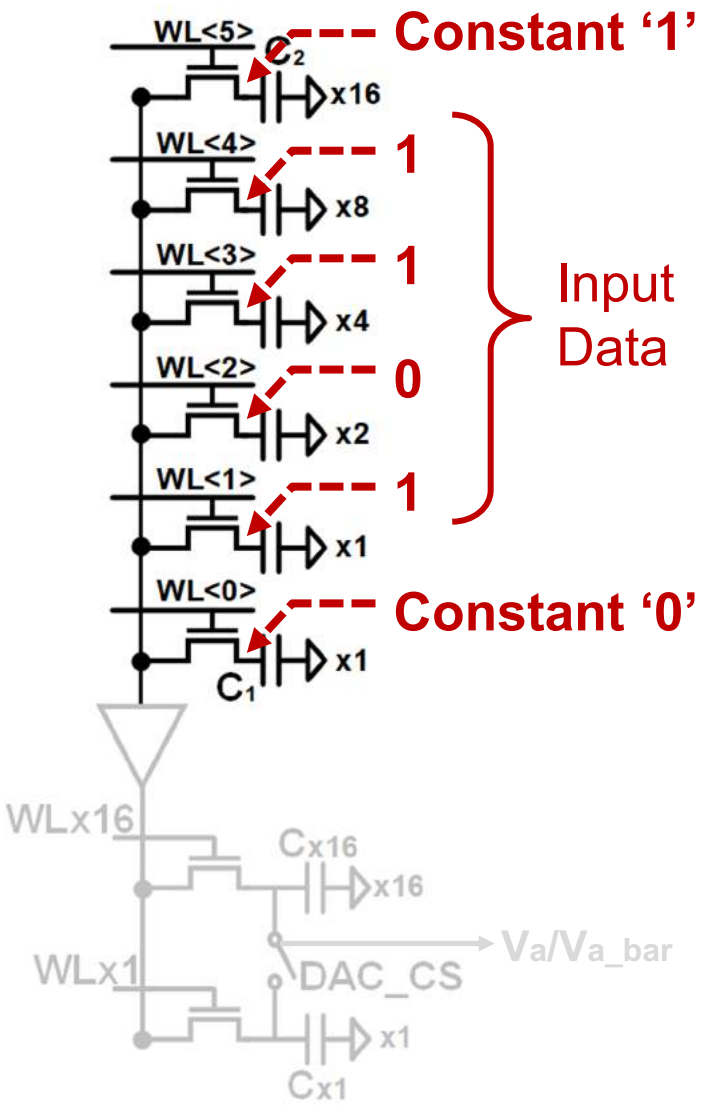


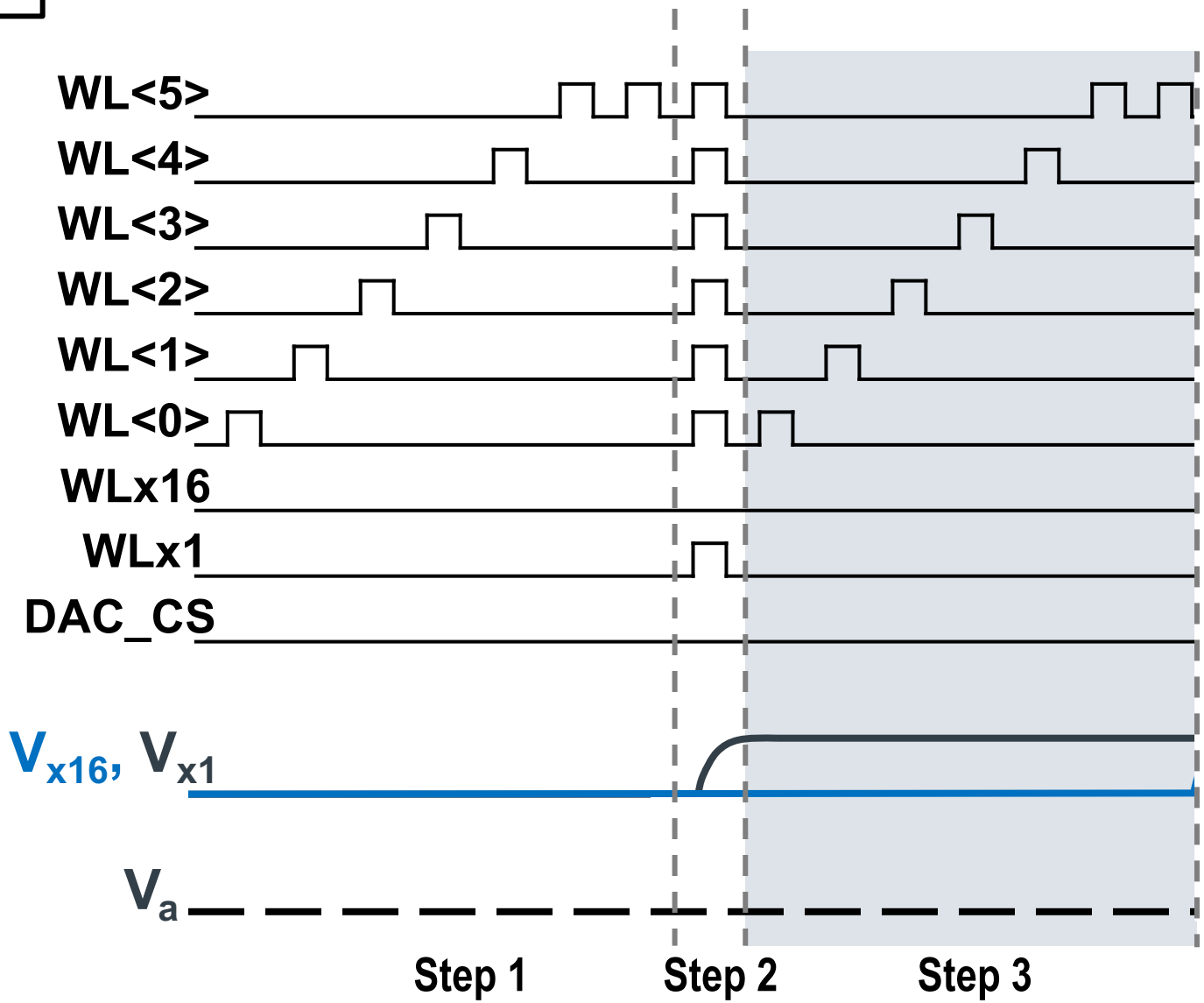
Activate all wordlines for charge share to convert digital data to analog voltage

Voltage is sampled on C_{x1} capacitor

8b In-eDRAM DAC Dataflow – Step 3

Step 3: Load higher 4-bit to DAC

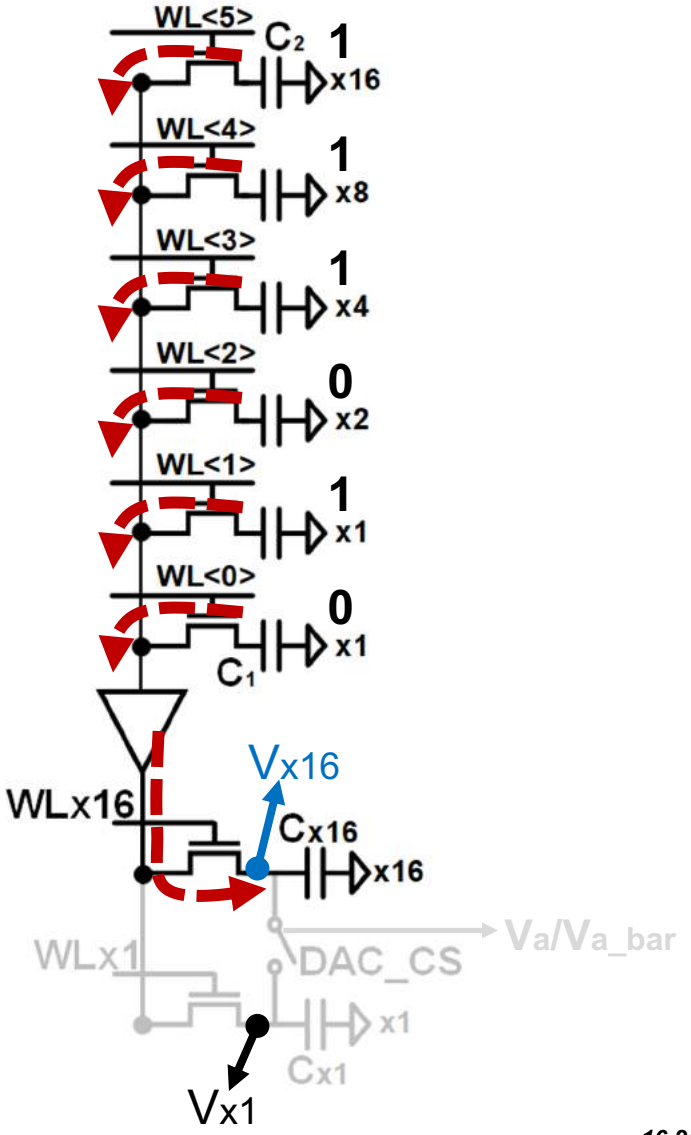


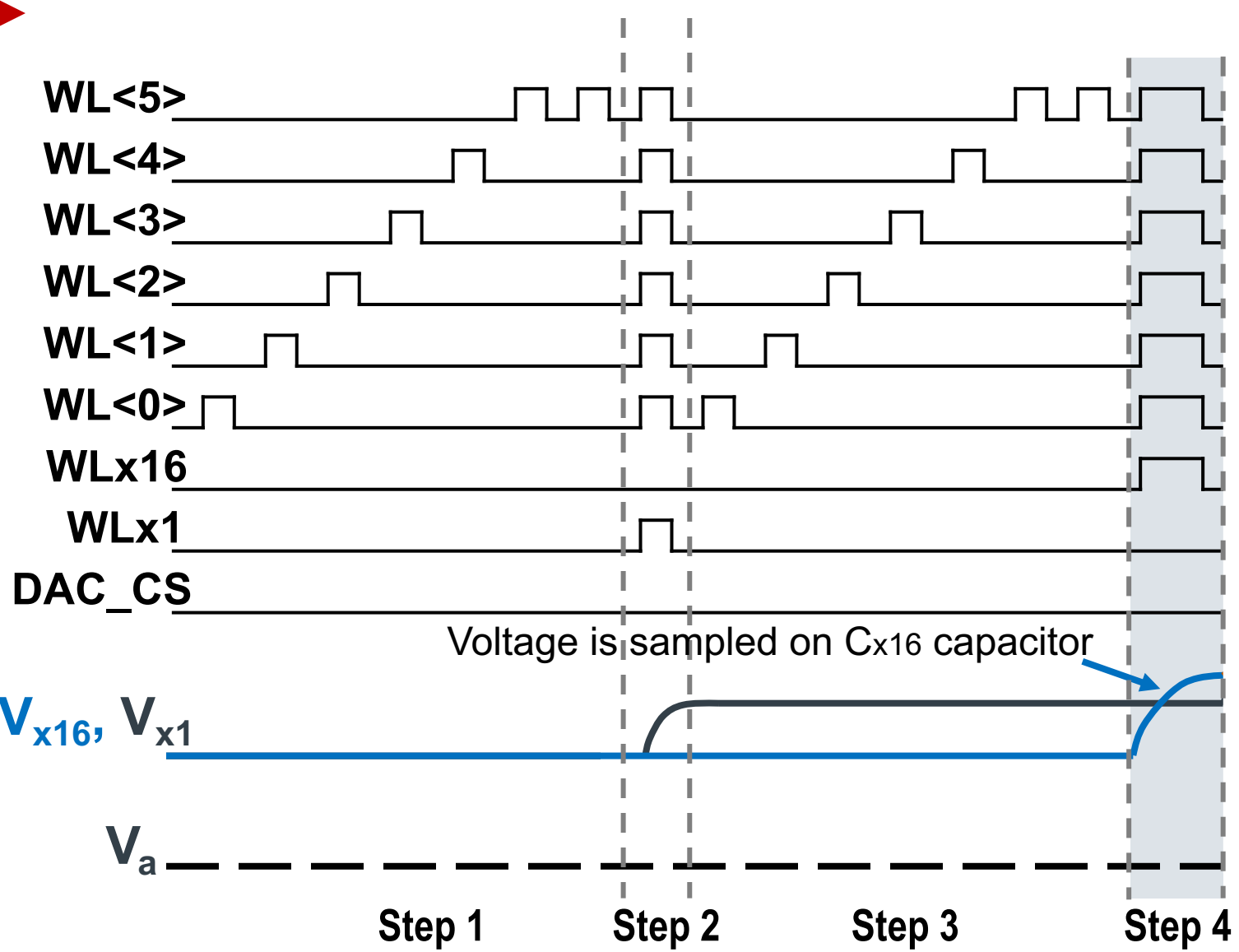


8b In-eDRAM DAC Dataflow – Step 4

Step 4: Charge Share

Data Flow

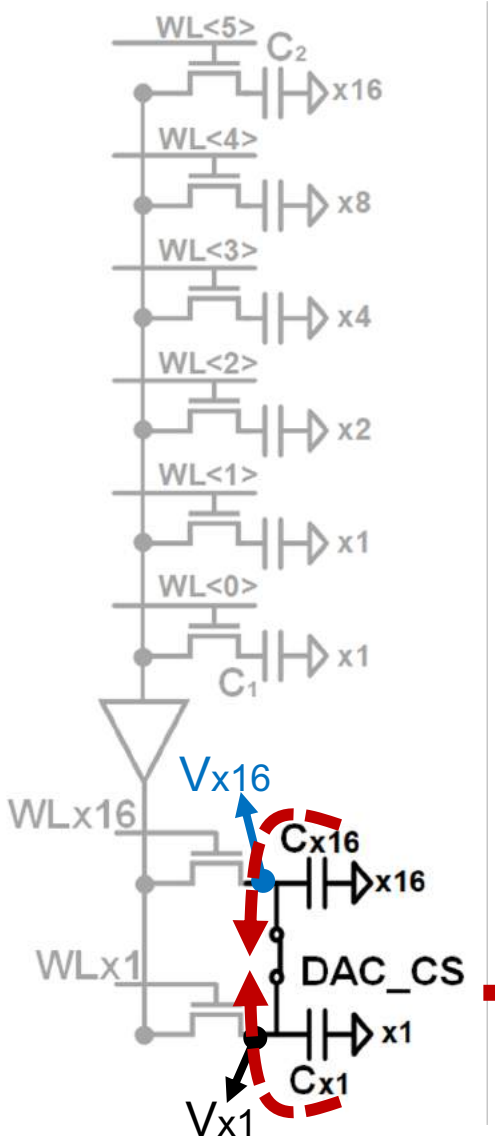


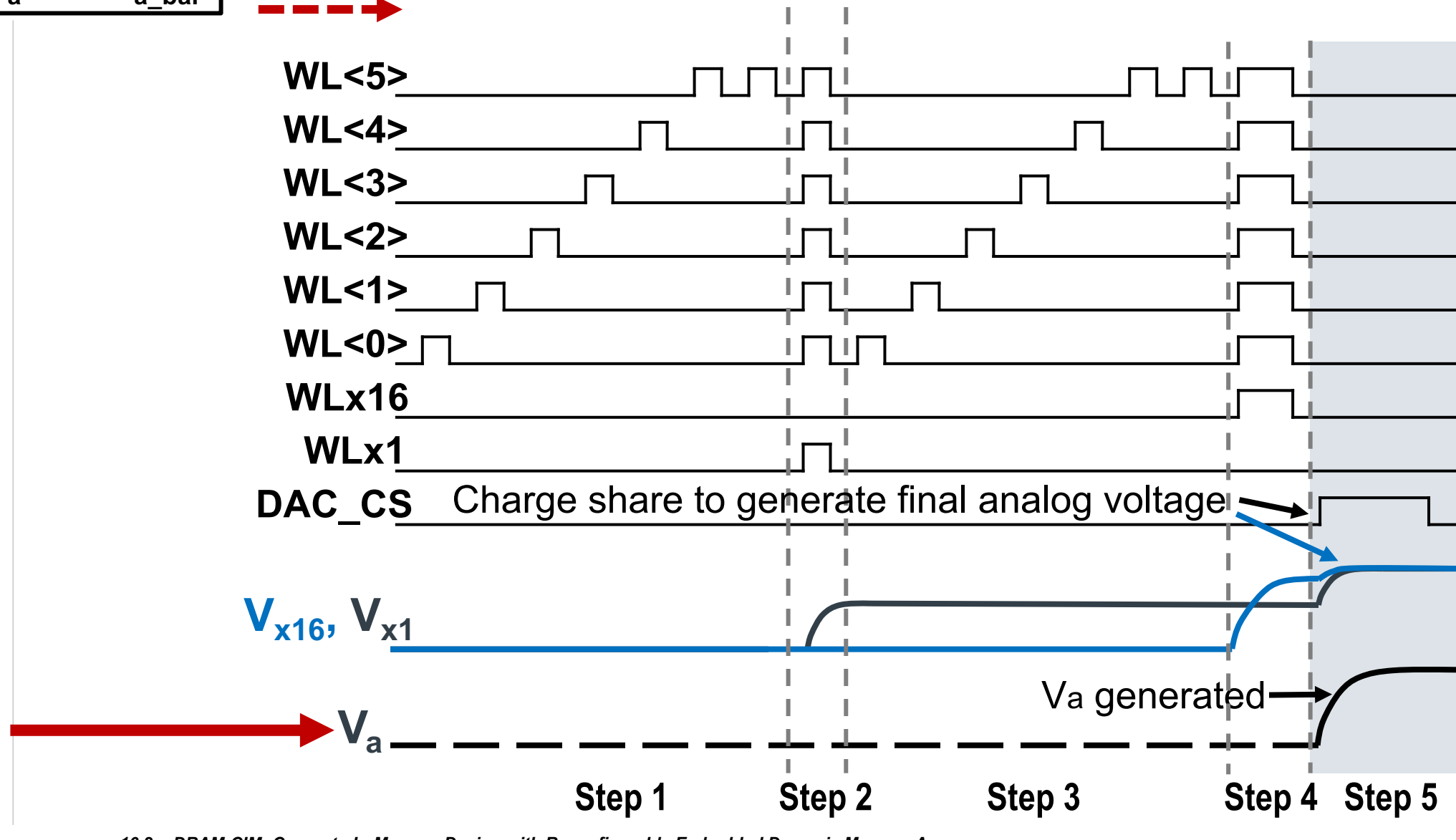


8b In-eDRAM DAC Dataflow – Step 5

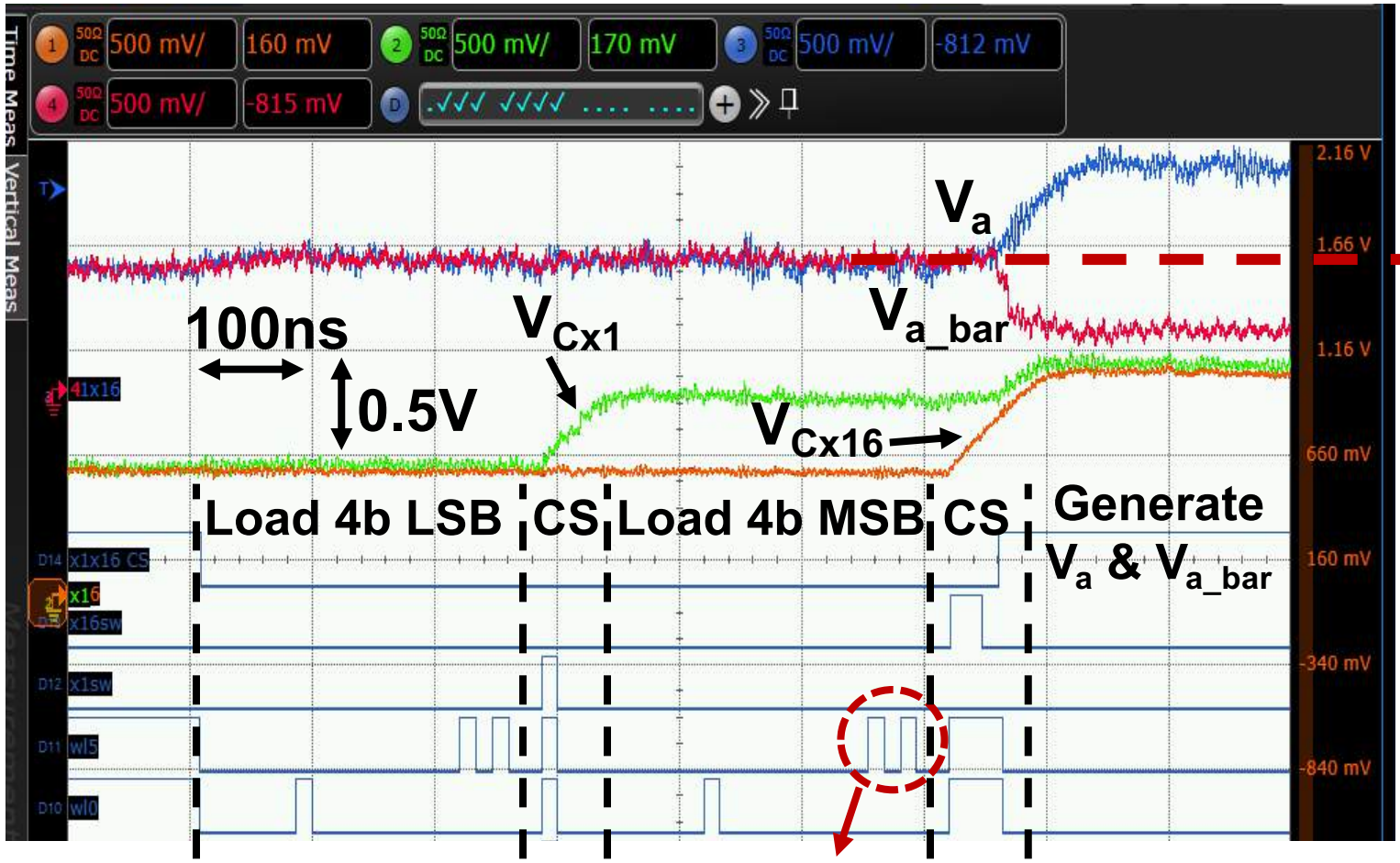
Step 5: Generate V_a and V_{a_bar}

Data Flow





1T1C In-eDRAM Differential DAC Demonstration



$V_{DD}/2$ = 600mV

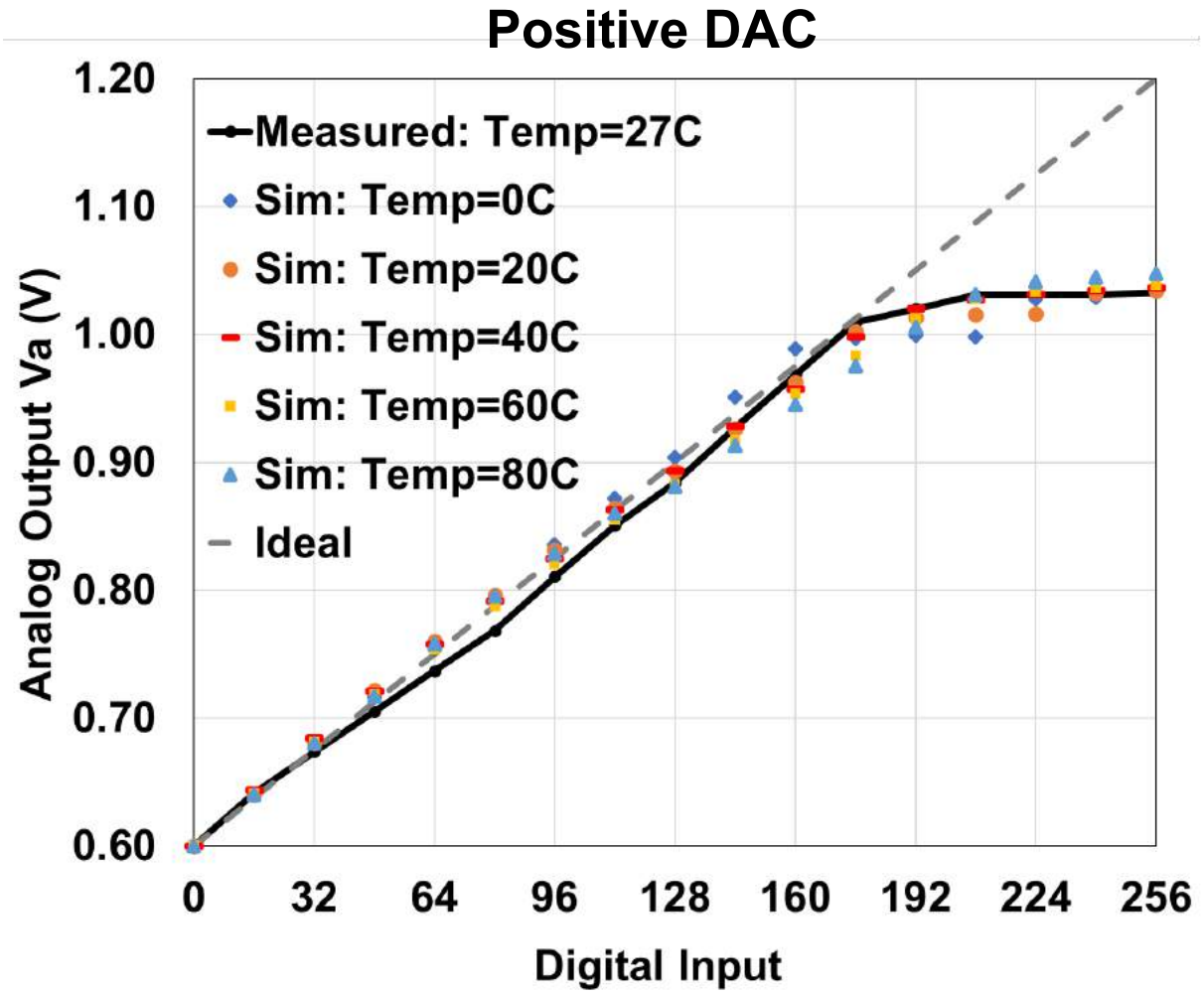
DAC_CS
WLx16
WLx1
WL<5>
WL<0>

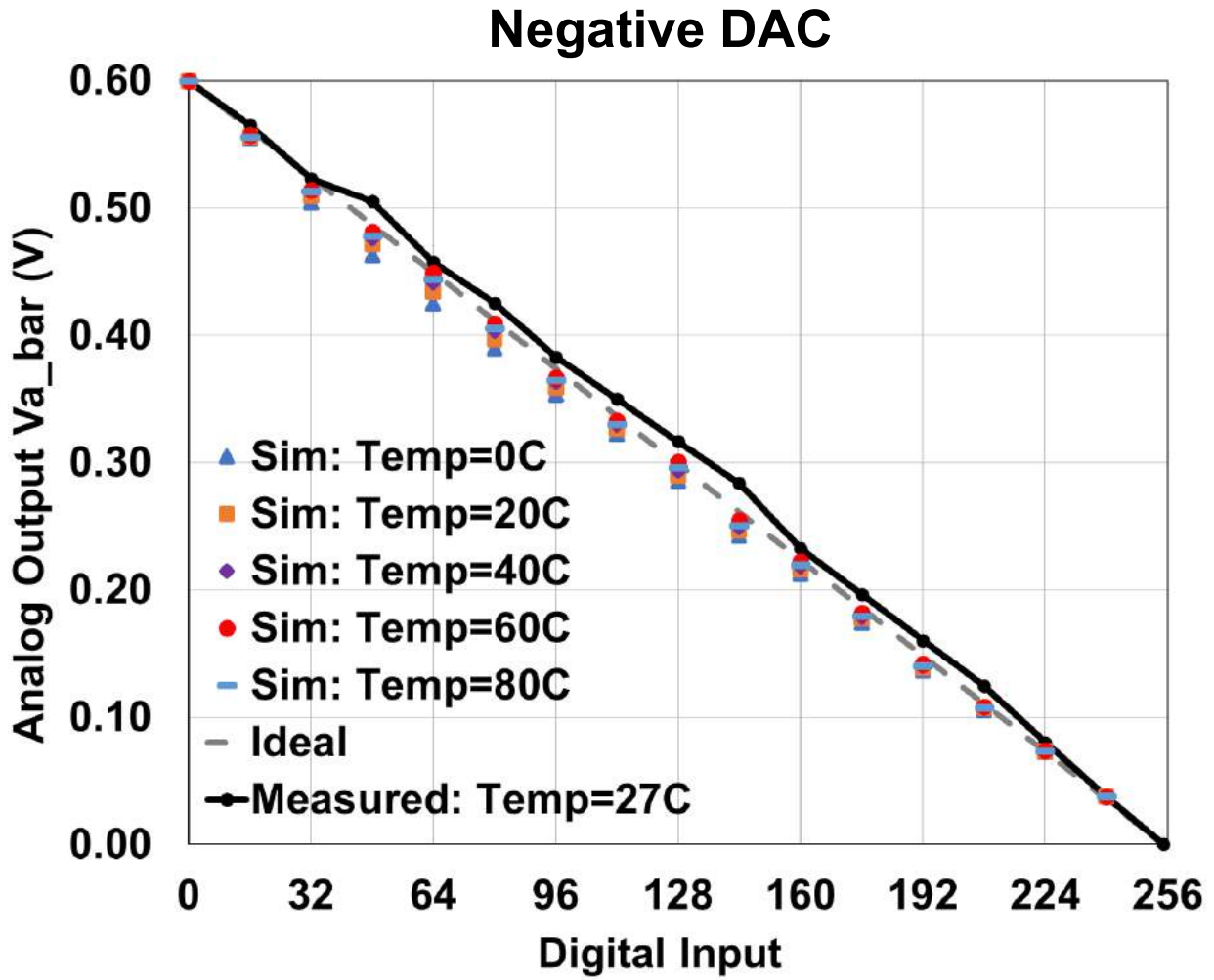
*CS=Charge Share

2 cycles: Make sure 16 capacitors are fully charged

- Signals match the expected waveforms
- V_a and V_{a_bar} are symmetrical around $V_{DD}/2$ (600mV)
- Higher 4 bits are loaded last: minimize capacitor leakage on final charge share value

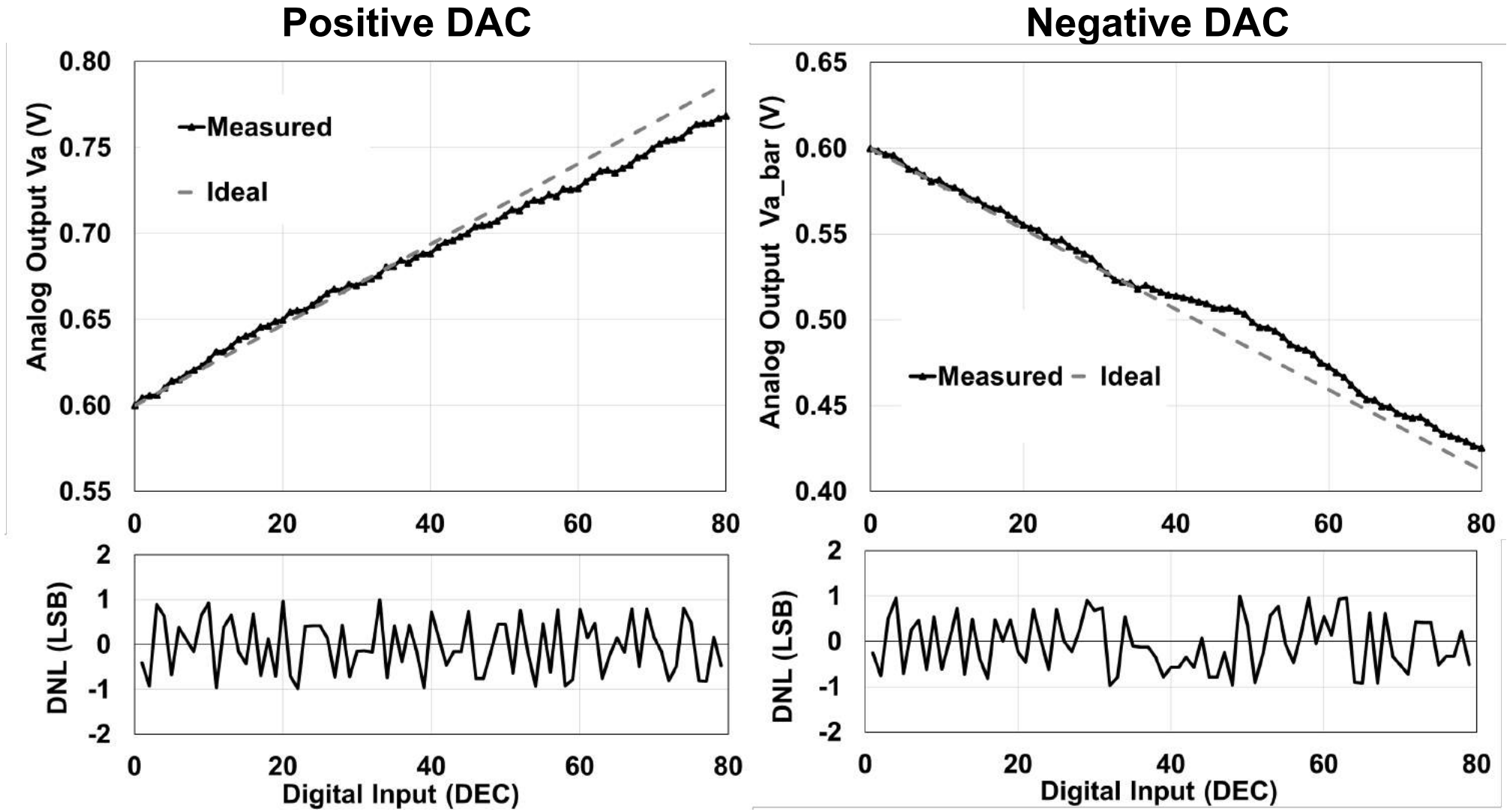
In-eDRAM DAC Characterization: Measured Results



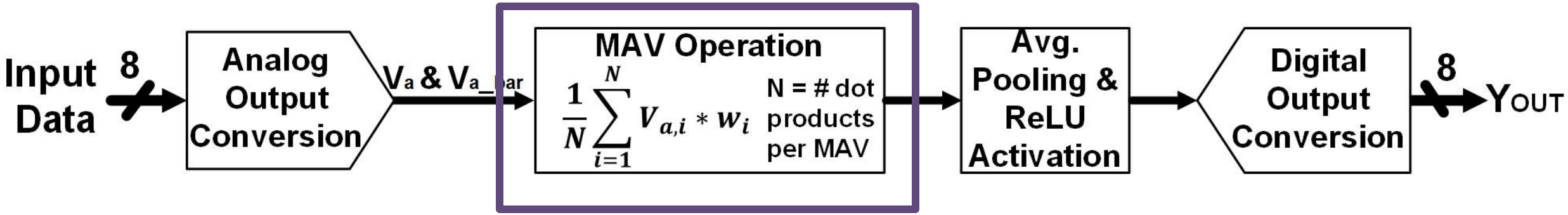


- Systematic deviation: positive DAC saturate at high value (can be correct during training)

In-eDRAM DAC Characterization: Zoomed In



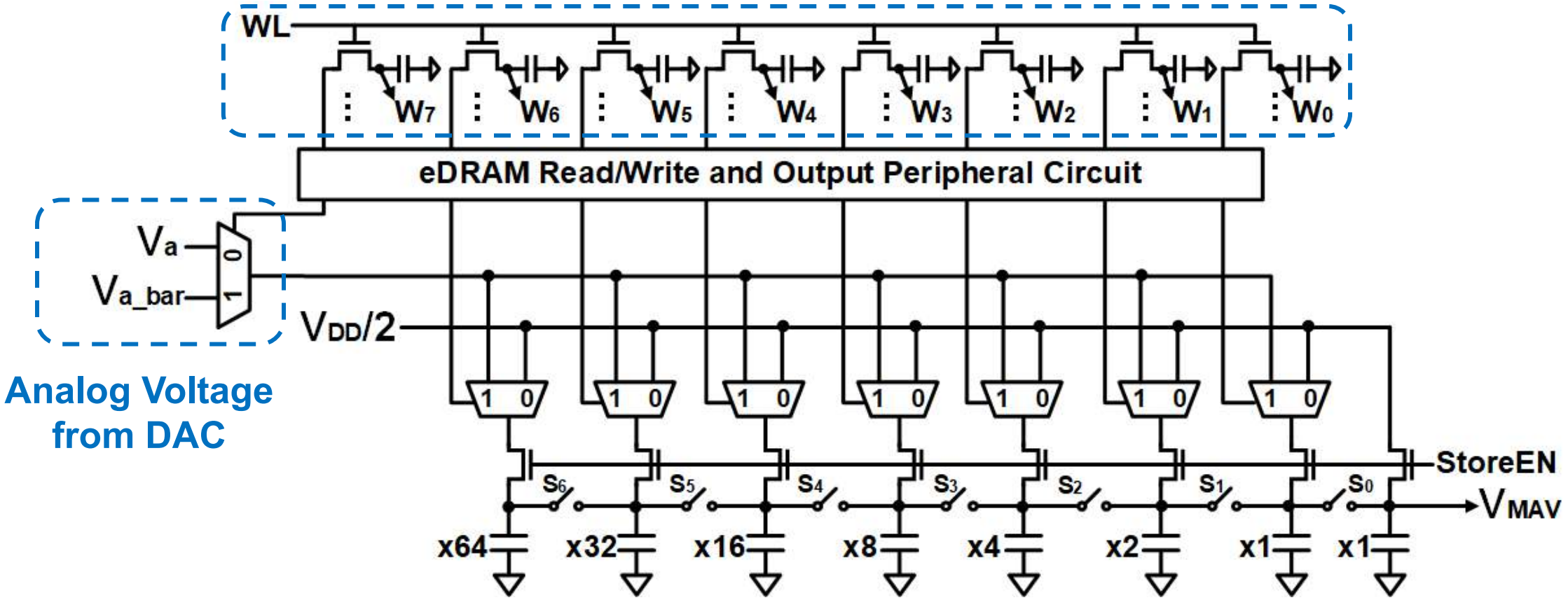
- Good linearity, DNL<1LSB



In-eDRAM Multiply-Accumulate-Averaging (MAV) Computation

eDRAM-CIM: MAV Computation

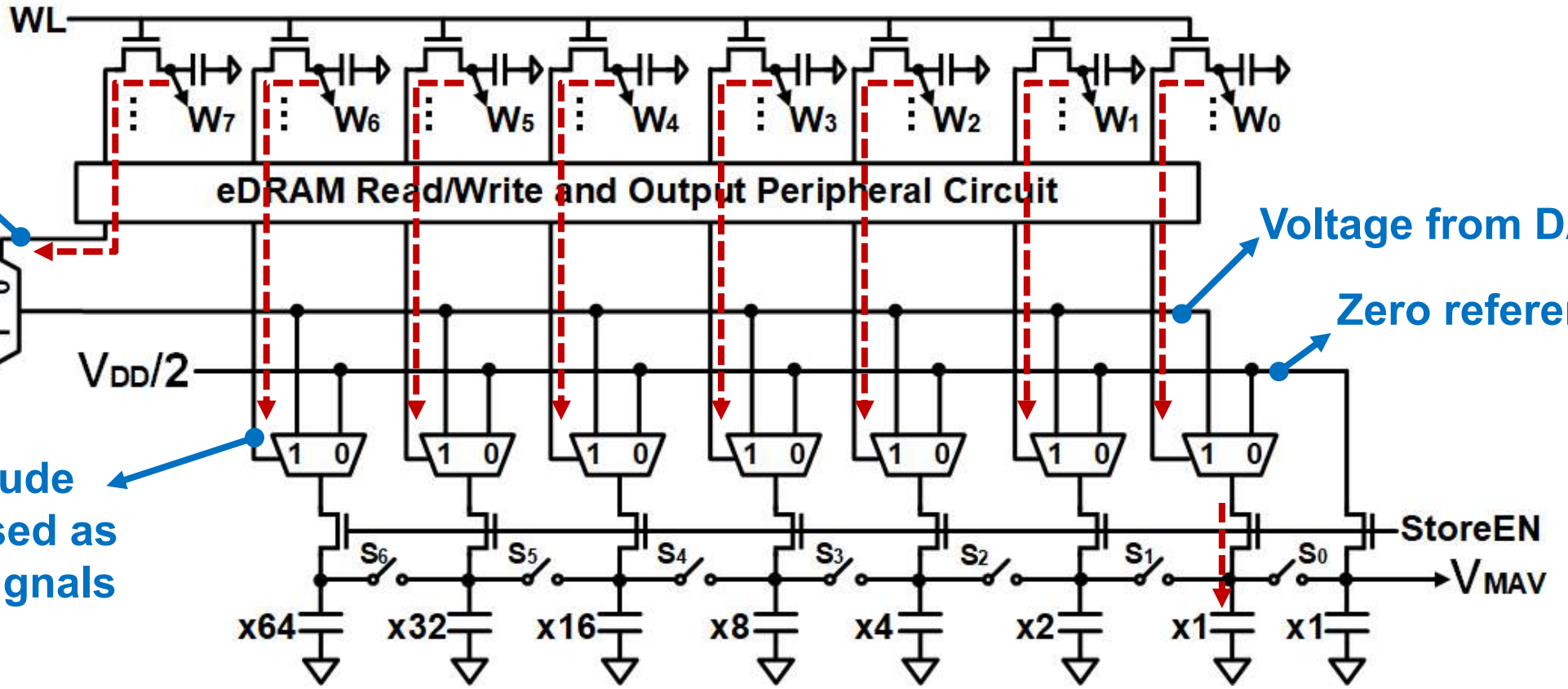
Weights are stored in 1T1C eDRAM bitcell: W_7 (sign bit); W_6~W_0 (magnitude bits)



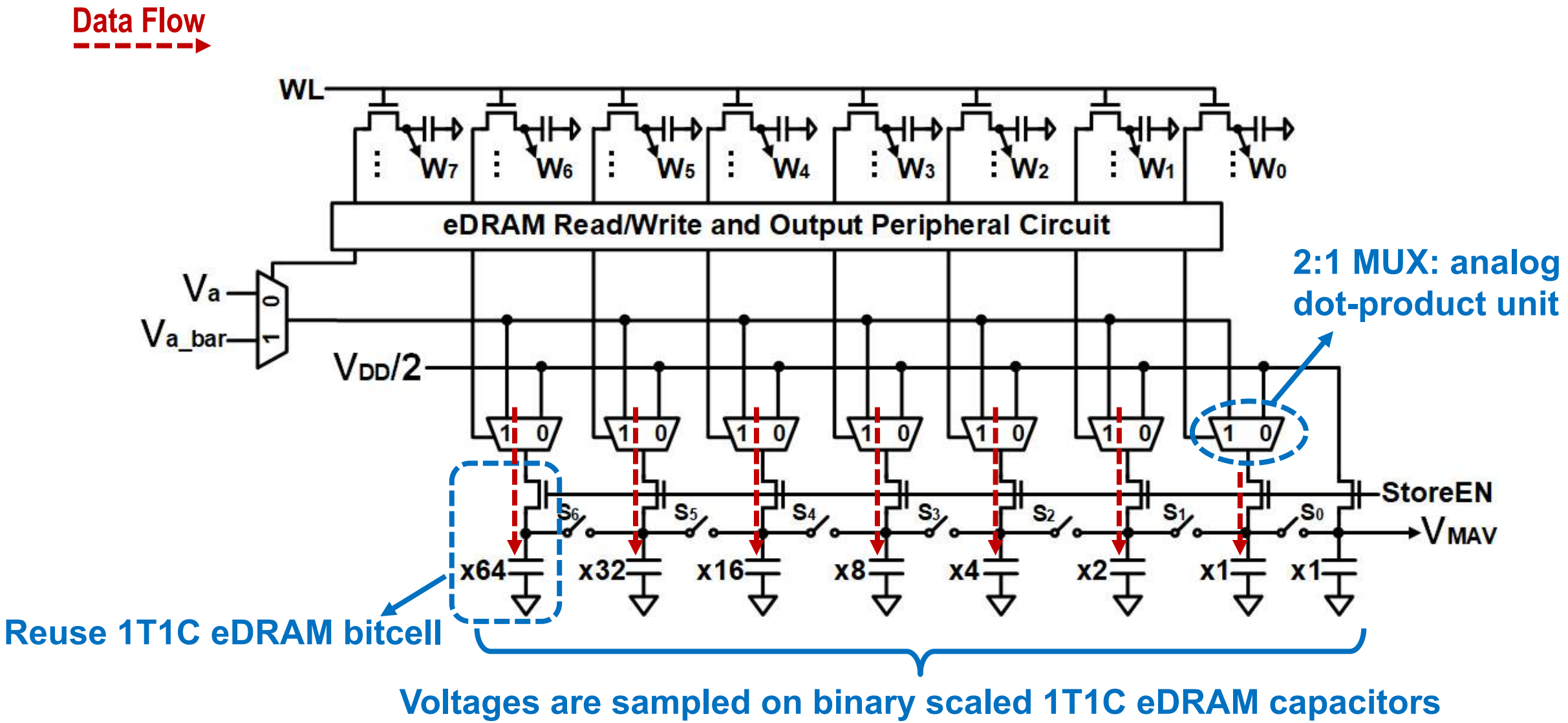
eDRAM-CIM: MAV Computation

Data Flow

W_7 (sign bit) is used to select between V_a and V_{a_bar}

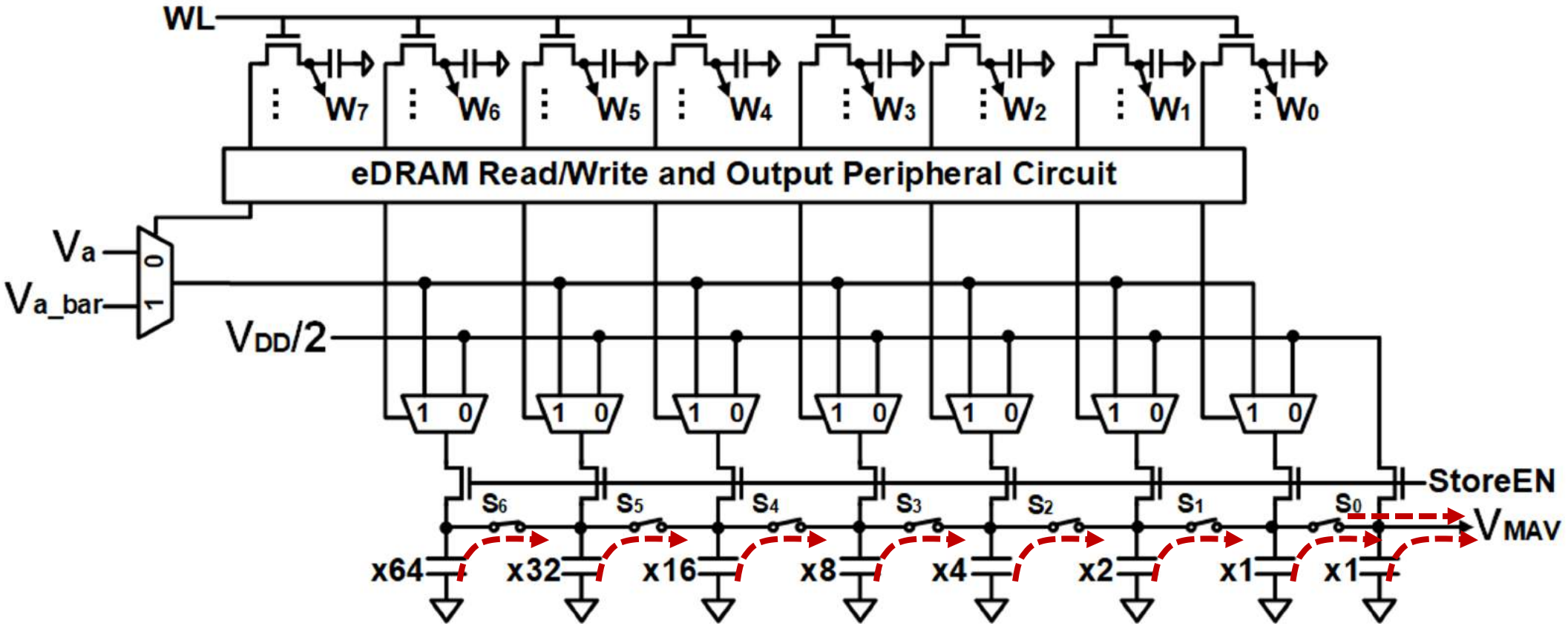


eDRAM-CIM: MAV Computation



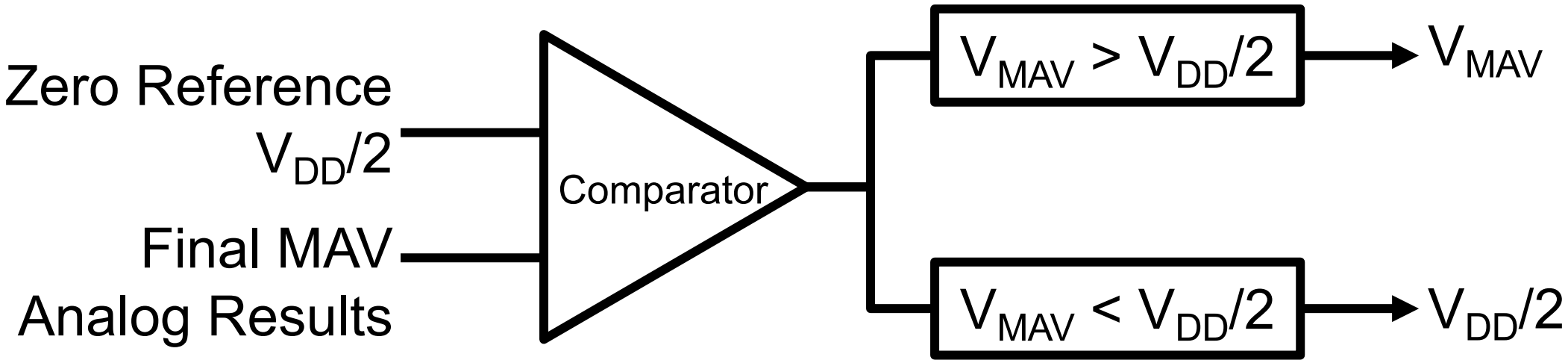
eDRAM-CIM: MAV Computation

Data Flow

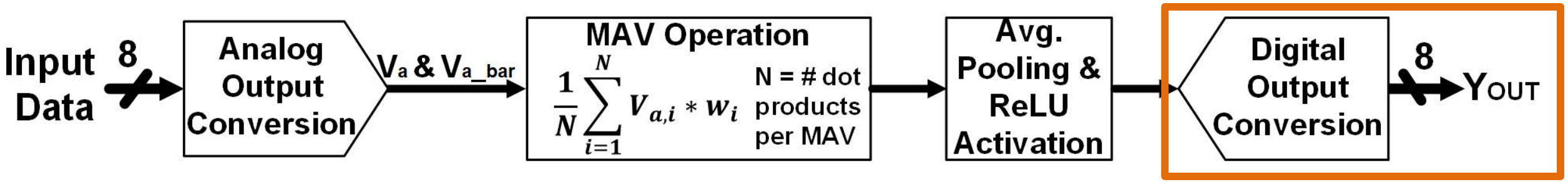


- Switches (S_0~S_6) are closed to charge share between all the 1T1C eDRAM cells
- V_{MAV} = final MAV computation result

ReLU

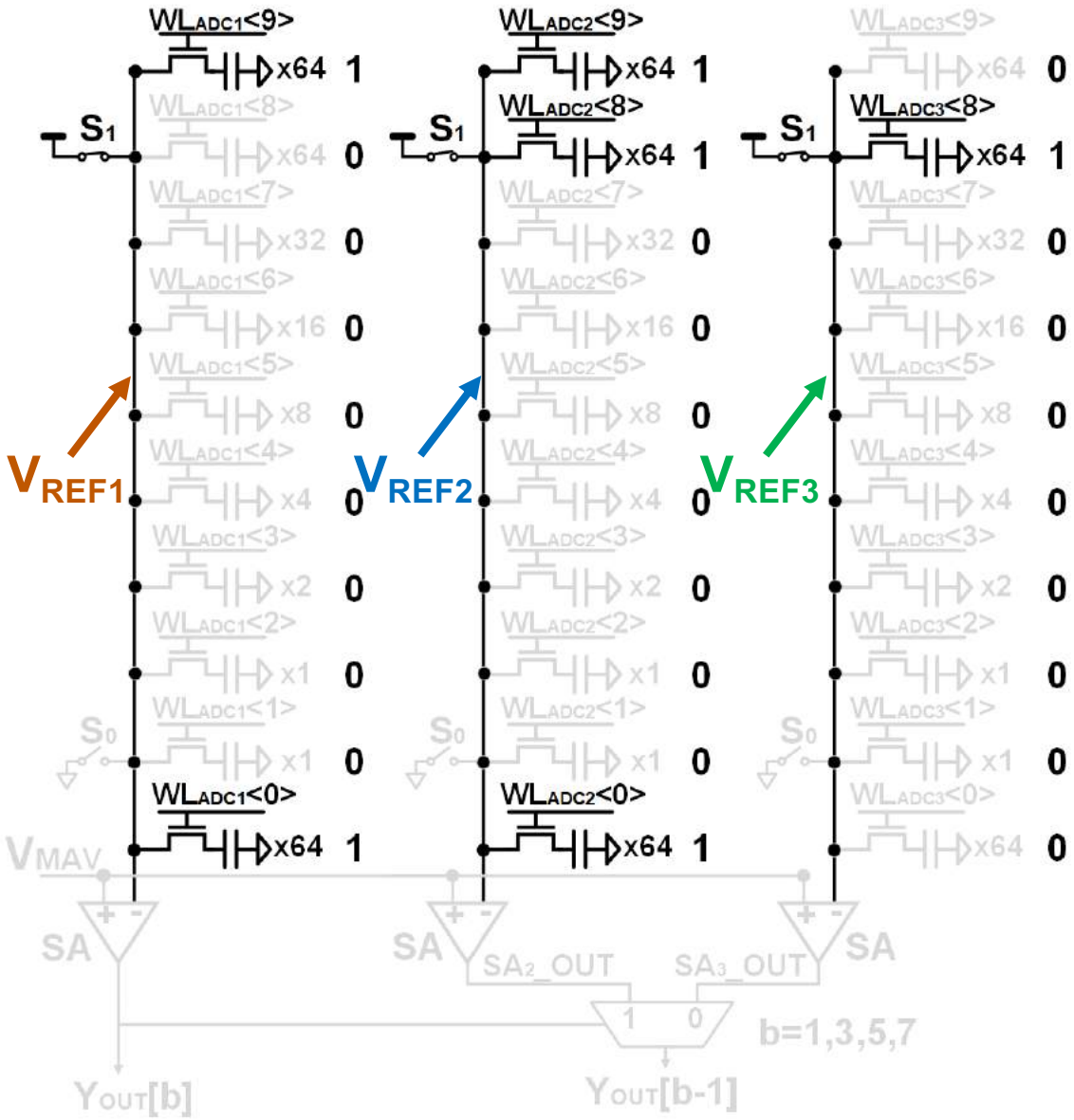


- Comparator compares VMAV and zero reference voltage VDD/2
- If $V_{MAV} > V_{DD}/2$ → output V_{MAV}
- If $V_{MAV} < V_{DD}/2$ → output $V_{DD}/2$

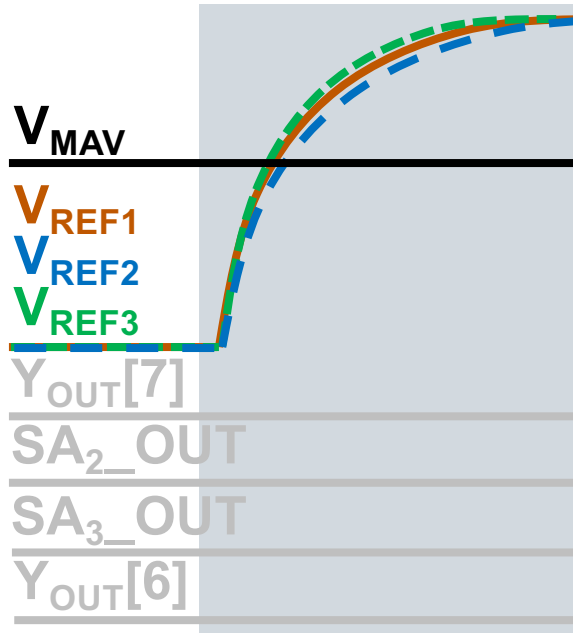


In-eDRAM Analog-to-Digital Converter (ADC)

eDRAM-CIM: ADC Operation

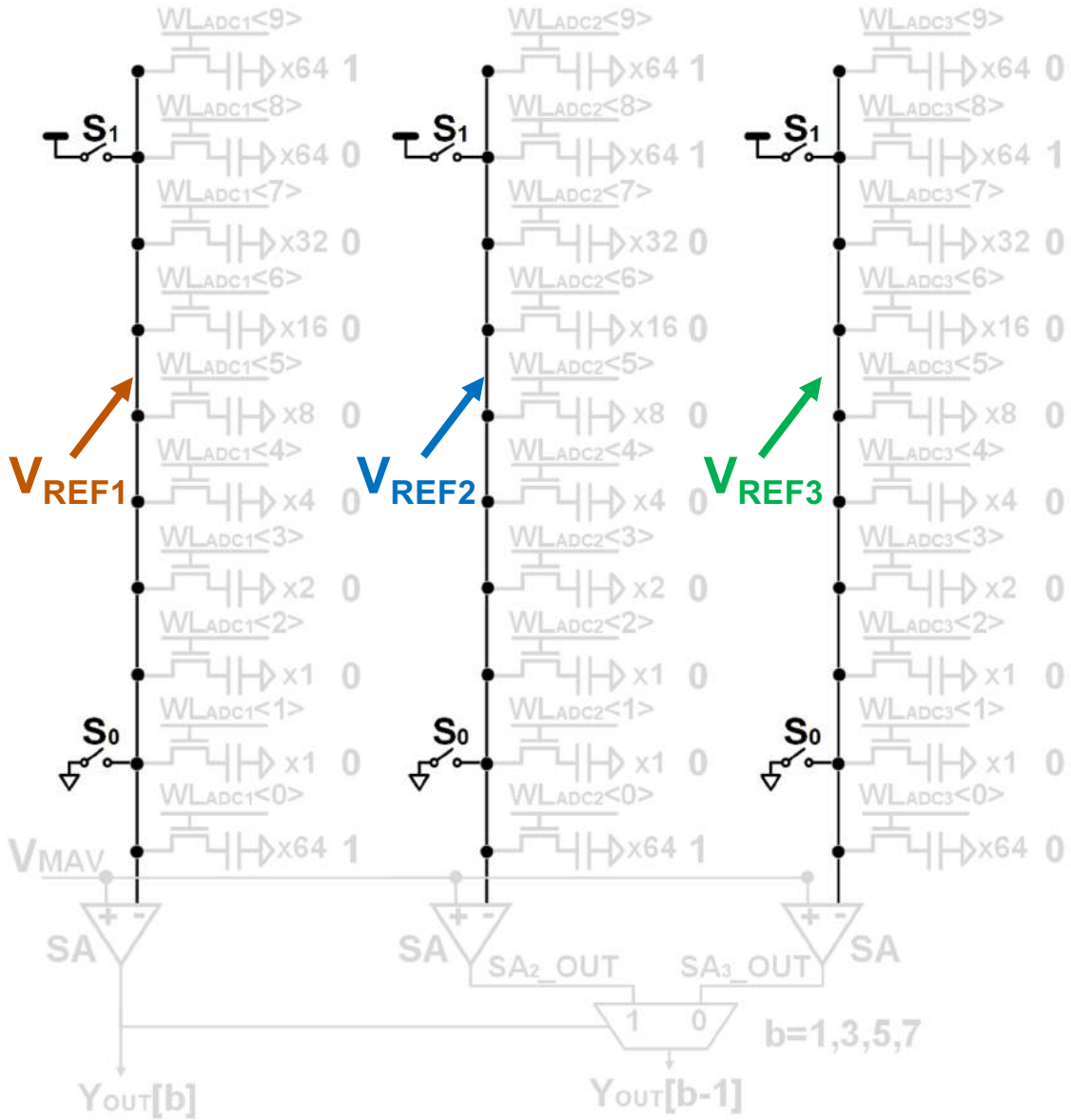


V_{REF} Generation Waveforms for 1st (b=1) conversion cycle

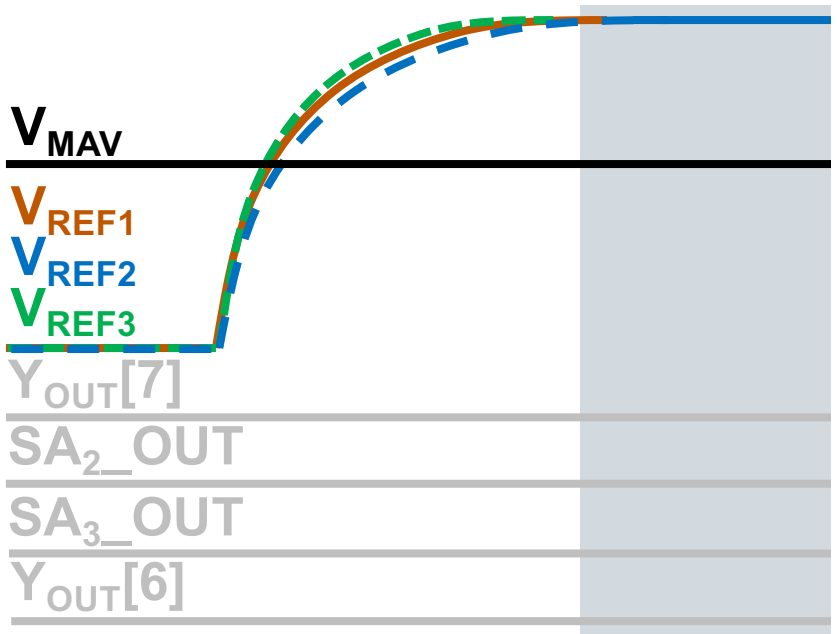


- **Step 1: Charge corresponding bitcells**

eDRAM-CIM: ADC Operation

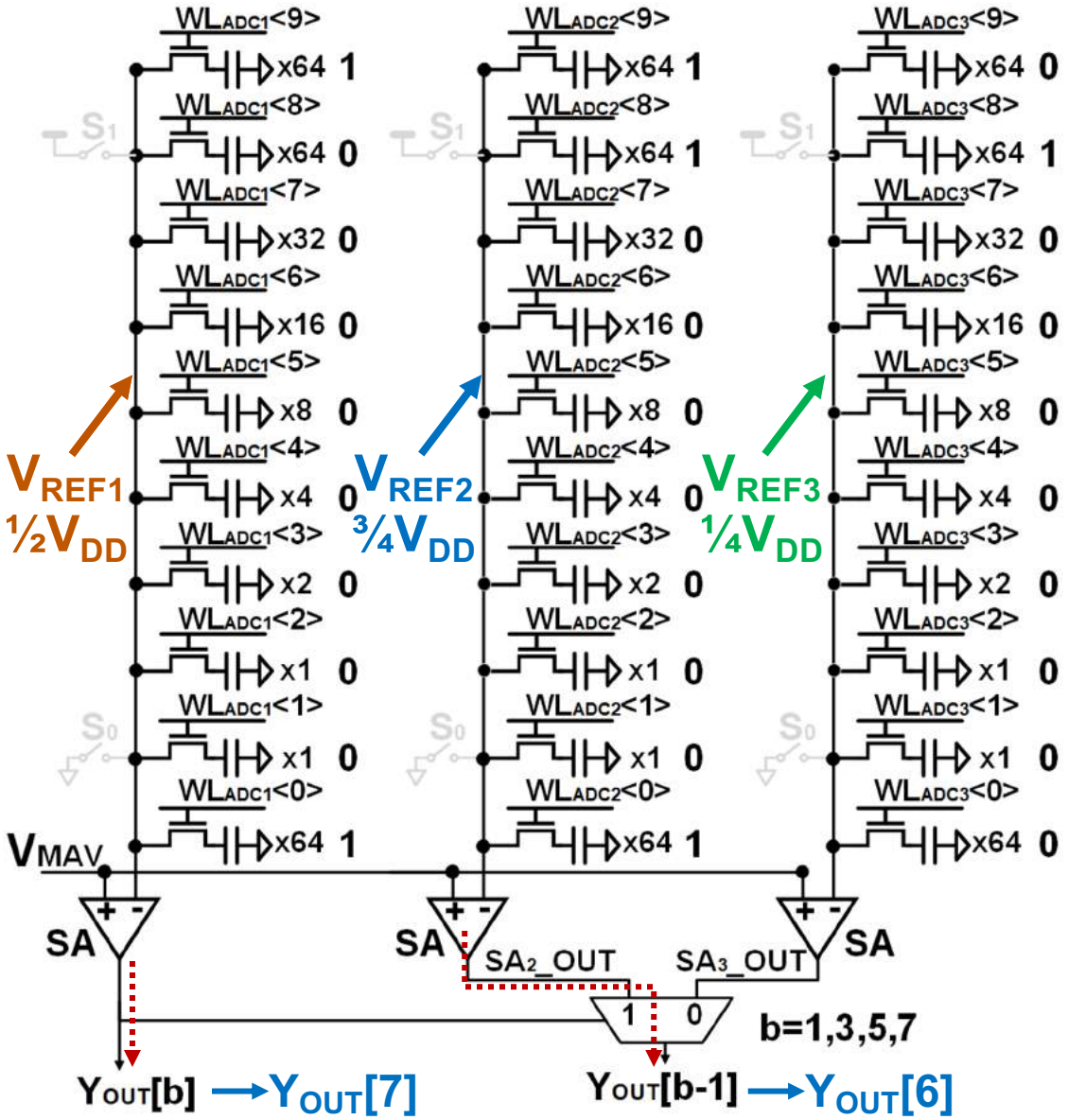


V_{REF} Generation Waveforms for 1st (b=1) conversion cycle

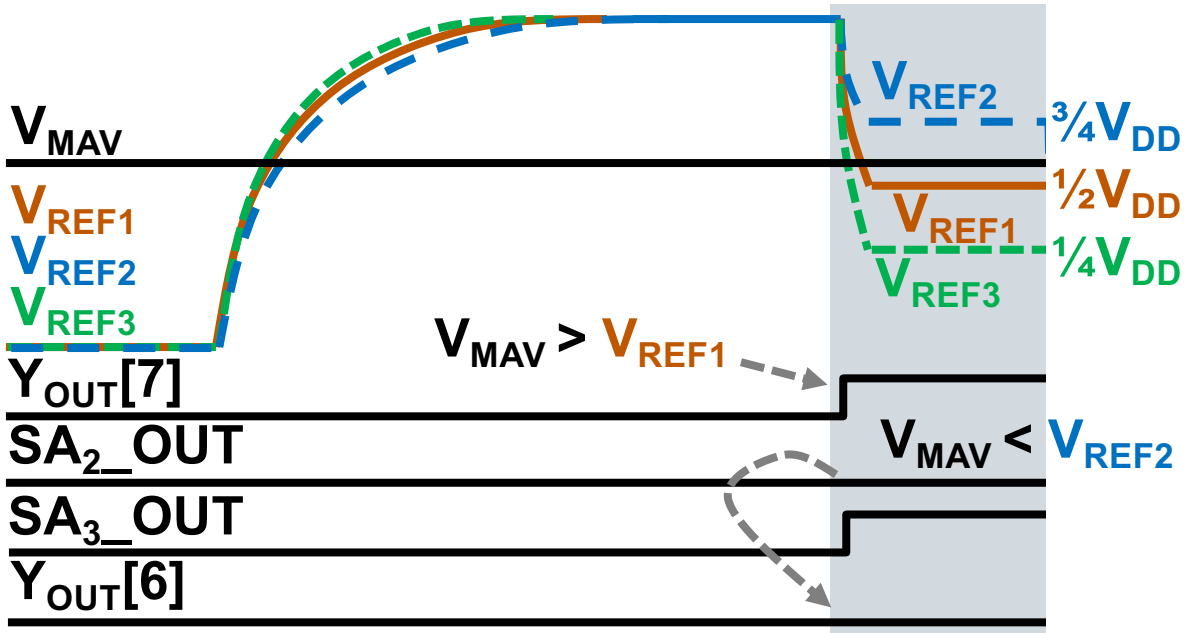


- Step 1: Charge corresponding bitcells
- **Step 2: IDLE**

eDRAM-CIM: ADC Operation

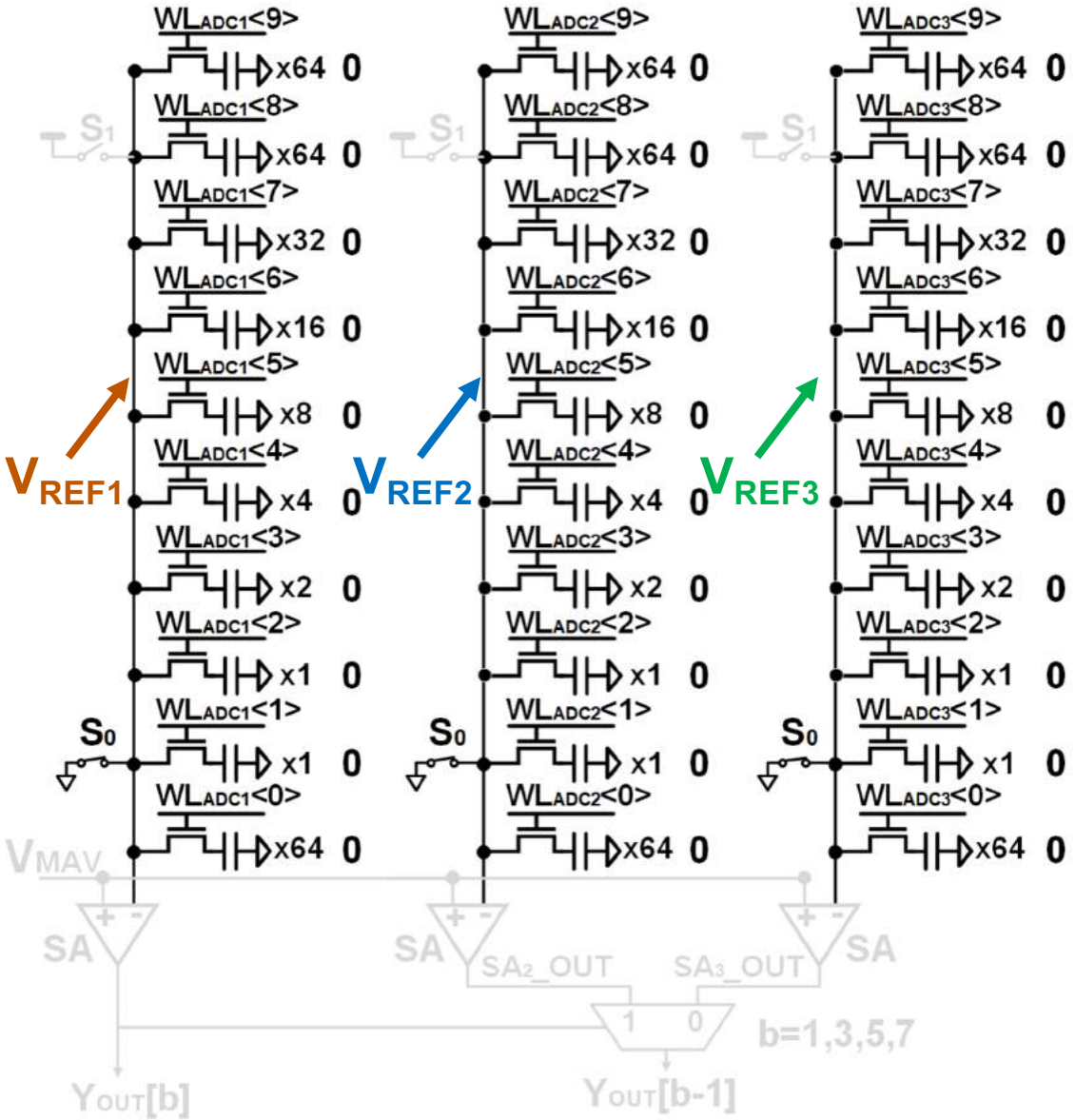


V_{REF} Generation Waveforms for 1st (b=1) conversion cycle

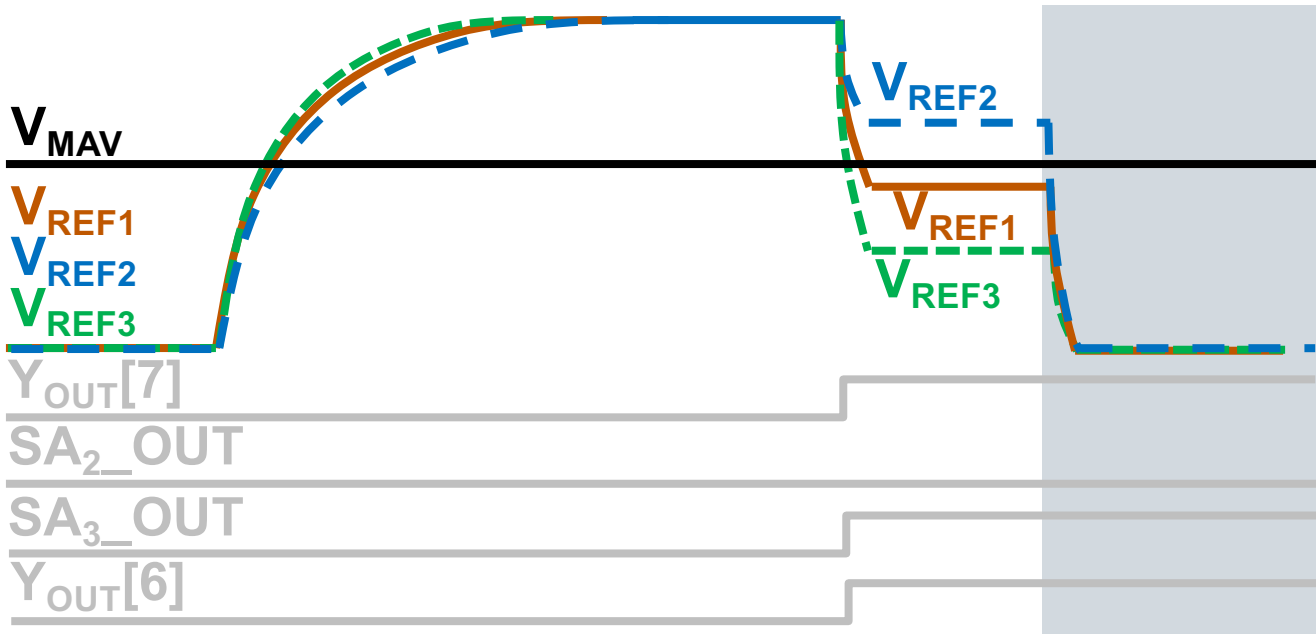


- Step 1: Charge corresponding bitcells
- Step 2: IDLE
- **Step 3: Charge Share to generate reference voltages**

eDRAM-CIM: ADC Operation

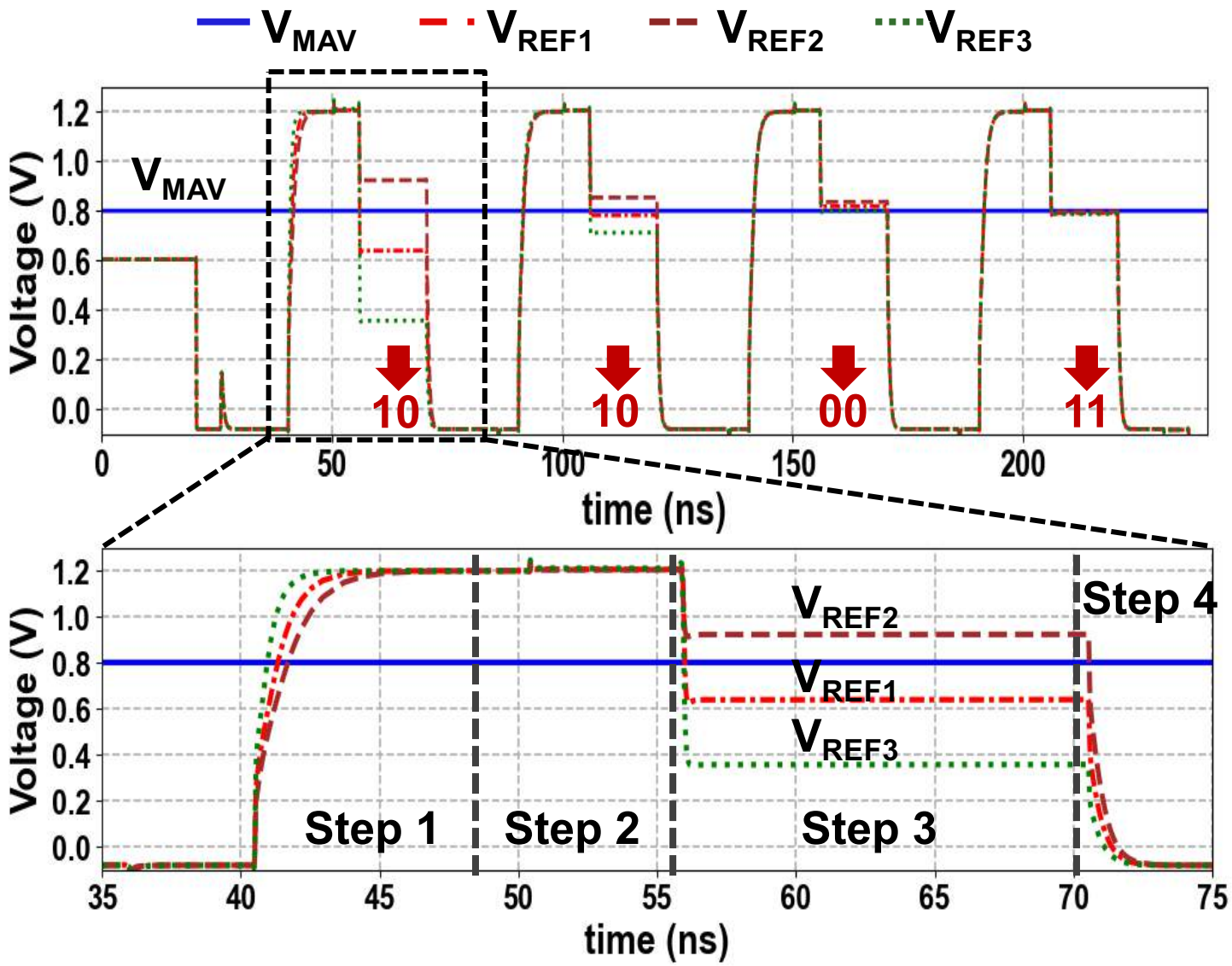


V_{REF} Generation Waveforms for 1st (b=1) conversion cycle



- Step 1: Charge corresponding bitcells
- Step 2: IDLE
- Step 3: Charge Share to generate reference voltages
- **Step 4: Discharge and reset 1T1C eDRAM capacitors**

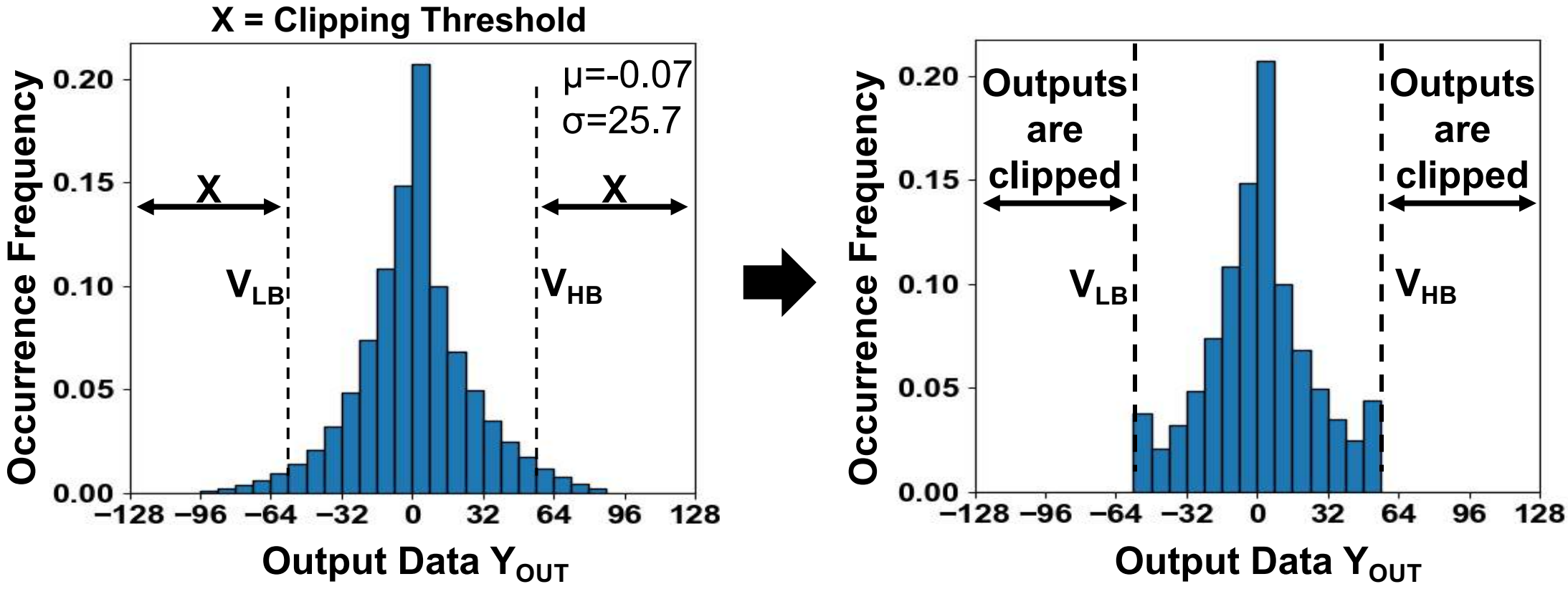
2b/Conversion In-eDRAM ADC Waveforms



- 2b/conversion cycle
- 2 extra eDRAM columns
- 3 reference voltages (V_{REF1}~V_{REF3})
- Minimize V_{MAV} capacitor leakage during conversion

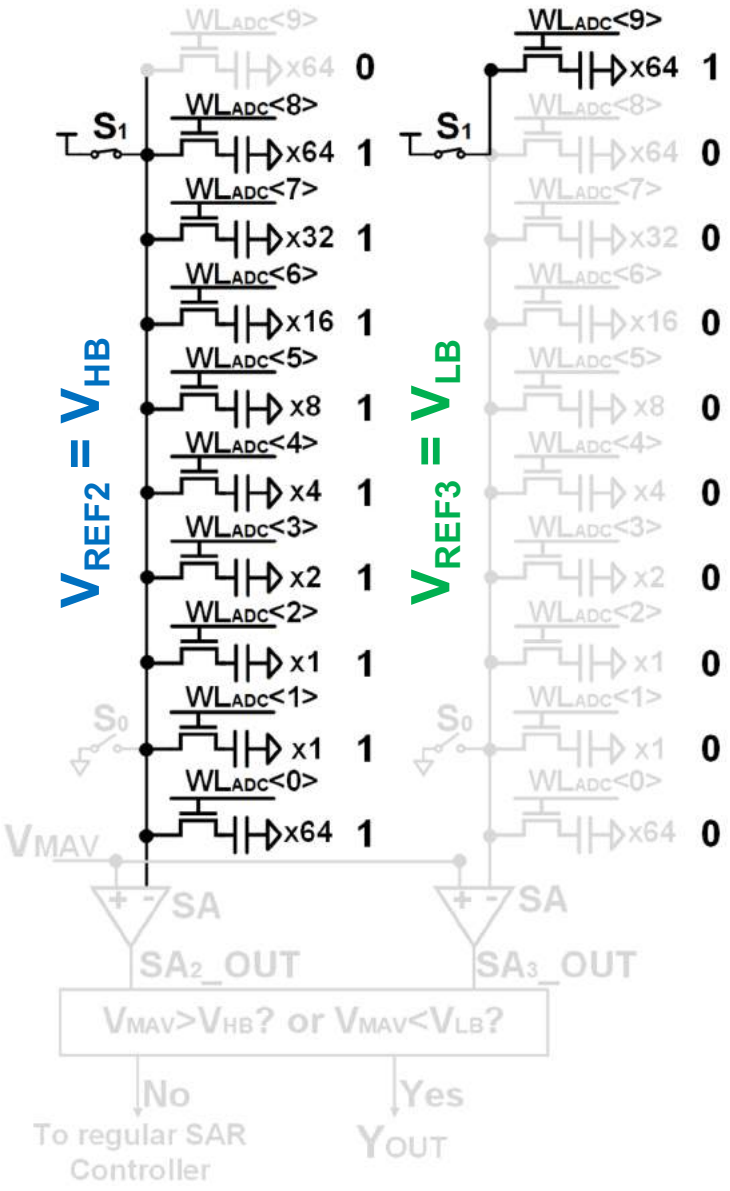
eDRAM-CIM: Adaptive ADC Motivation

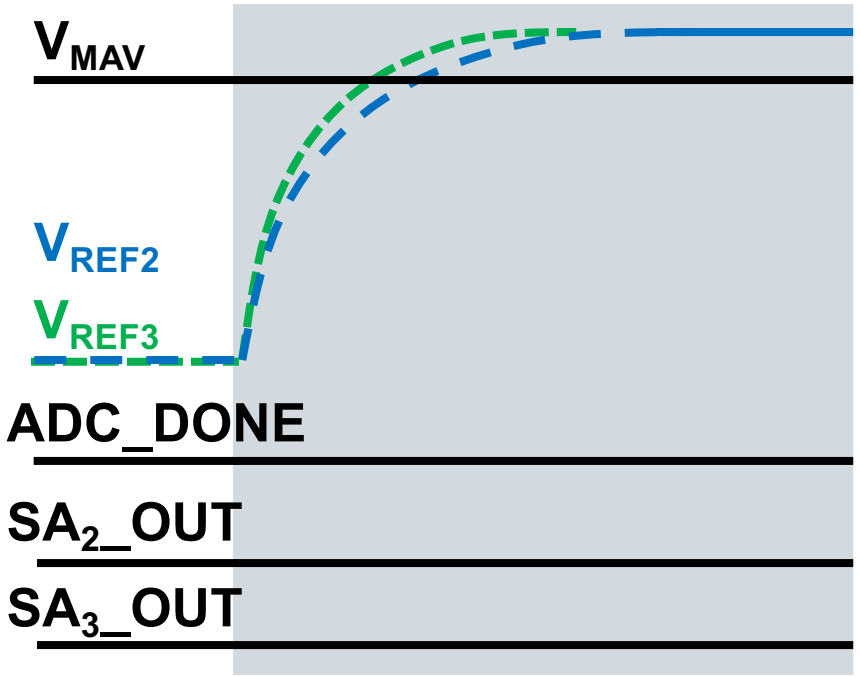
Convolution Outputs (CiFAR-10 dataset Layer 1)



Infrequent appearing MAV computation outputs (V_{MAV}) are trimmed to mitigate ADC energy and latency

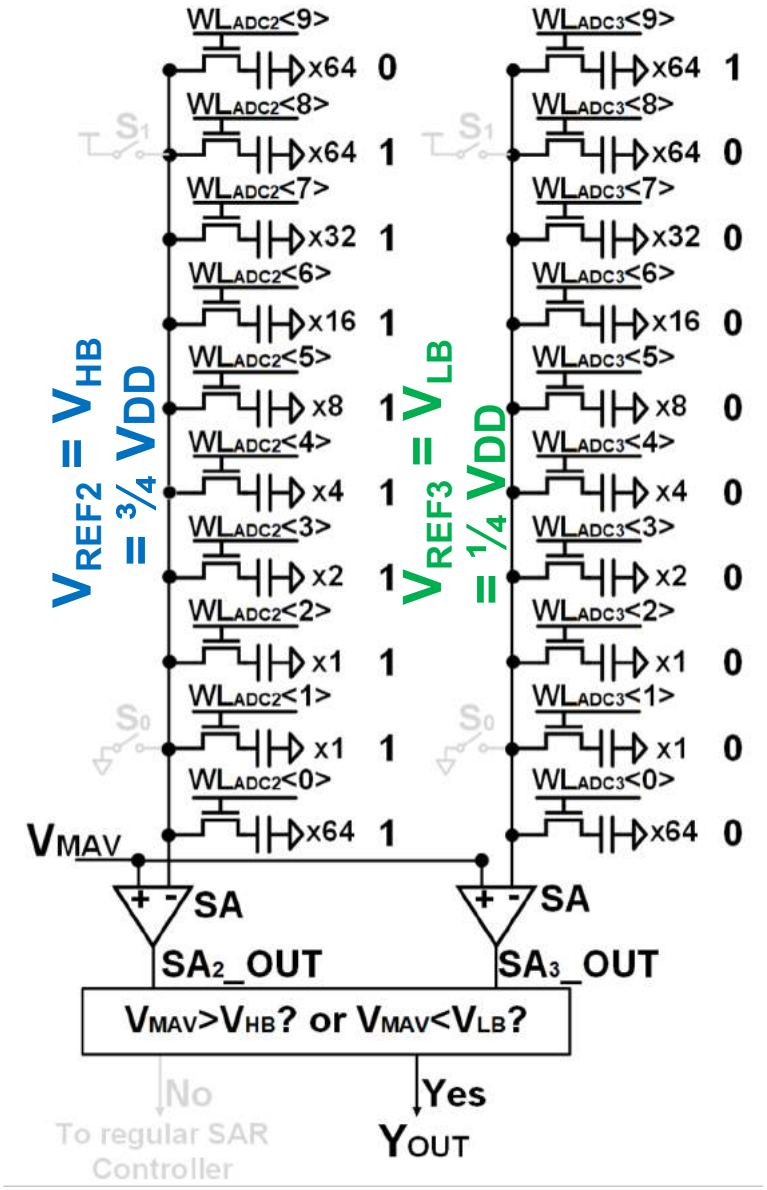
Clipping In-eDRAM ADC Data Flow

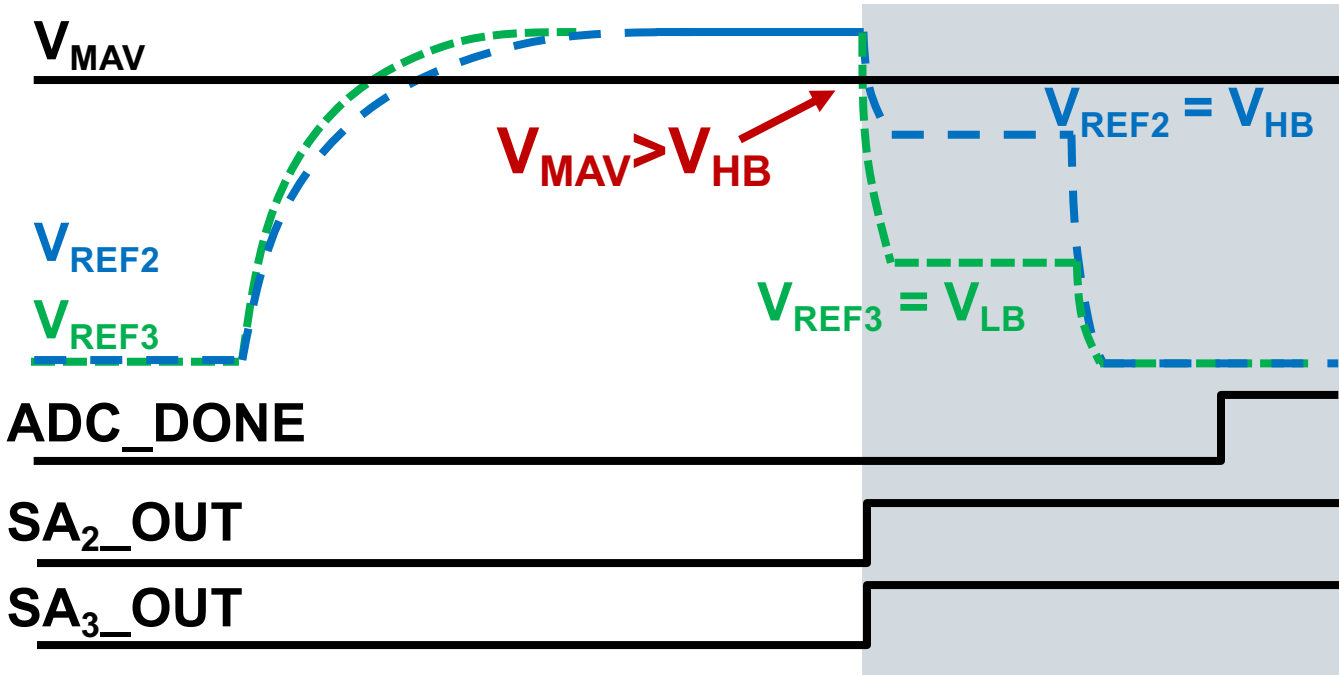




- **Step 1: Charge corresponding bitcells**
- **Step 2: IDLE**

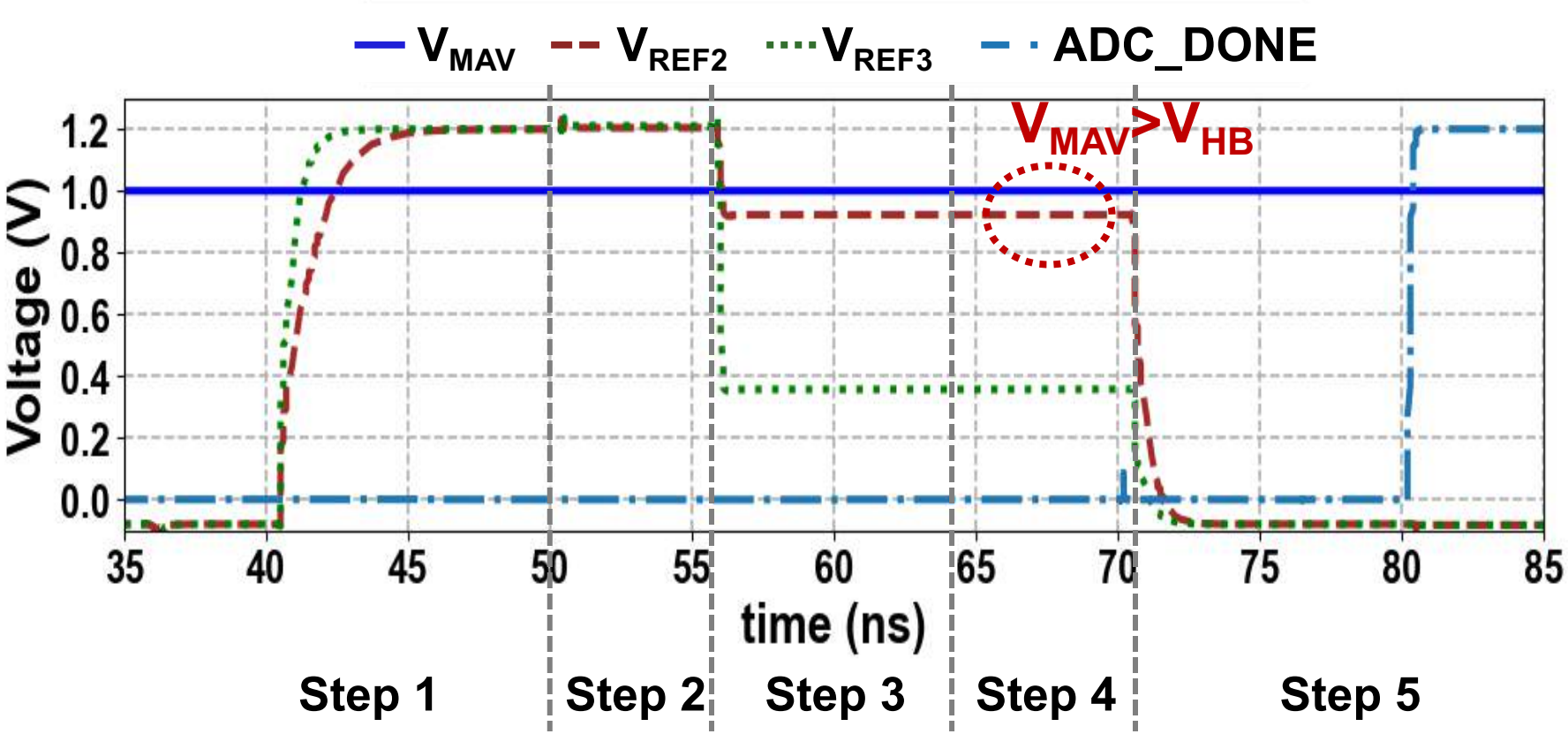
Clipping In-eDRAM ADC Data Flow





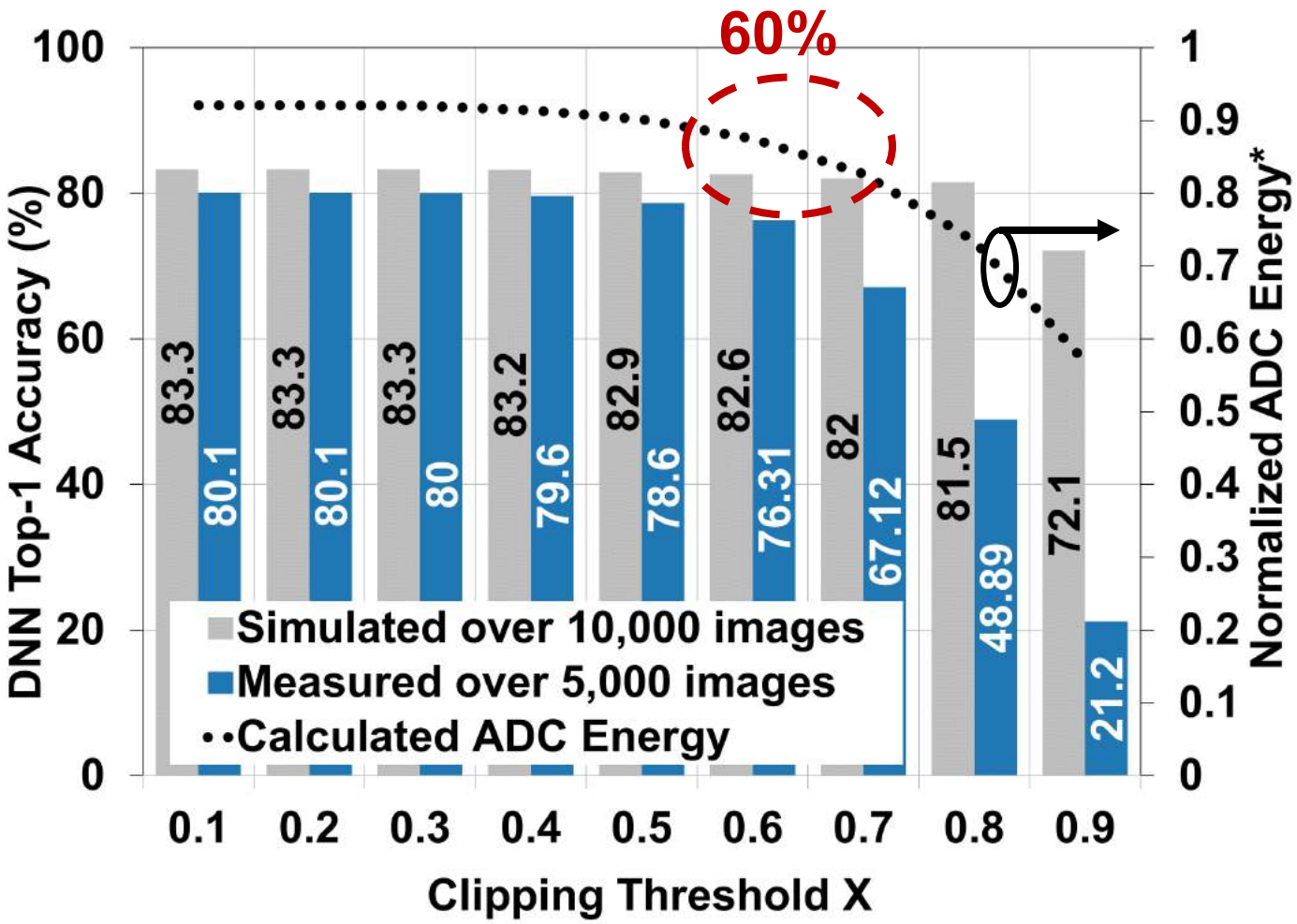
- Step 1: Charge corresponding bitcells
- Step 2: IDLE
- **Step 3: Charge share to generate reference voltages**
- **Step 4: Compare → $V_{MAV}>V_{HB}$? or $V_{MAV}<V_{LB}$? → Yes**
- **Step 5: Skip subsequent conversion & discharge BLs**

Clipping ADC Simulation Waveforms



- Skip subsequent ADC conversion cycles if V_{MAV} is higher than V_{HB} or lower than V_{LB}
- If V_{MAV} is in conversion range, in-eDRAM ADC will perform regular SAR conversion

Clipping Threshold vs Accuracy and Energy



Clipping threshold

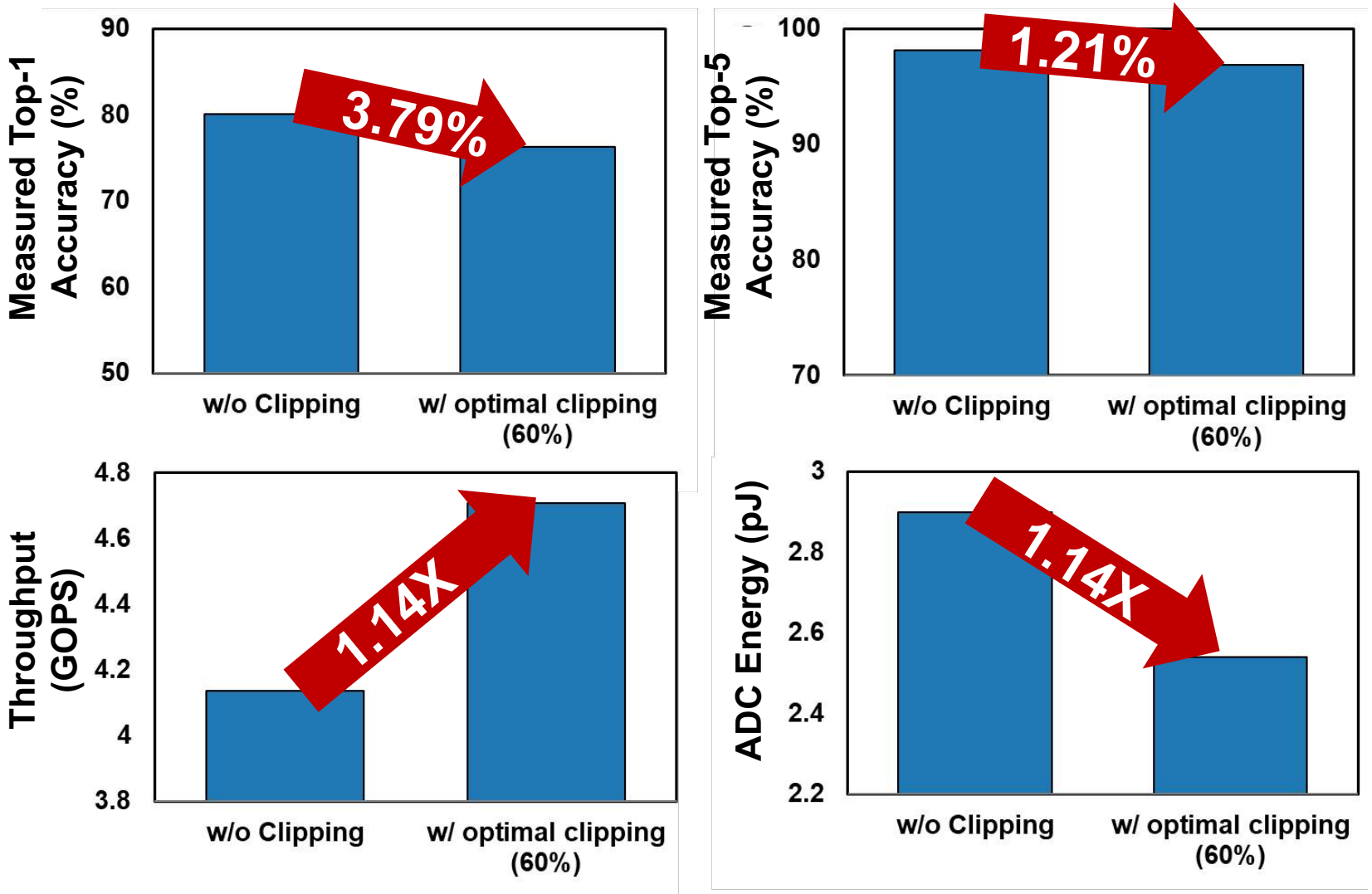
DNN accuracy

ADC Energy

Optimum clipping threshold = 60%

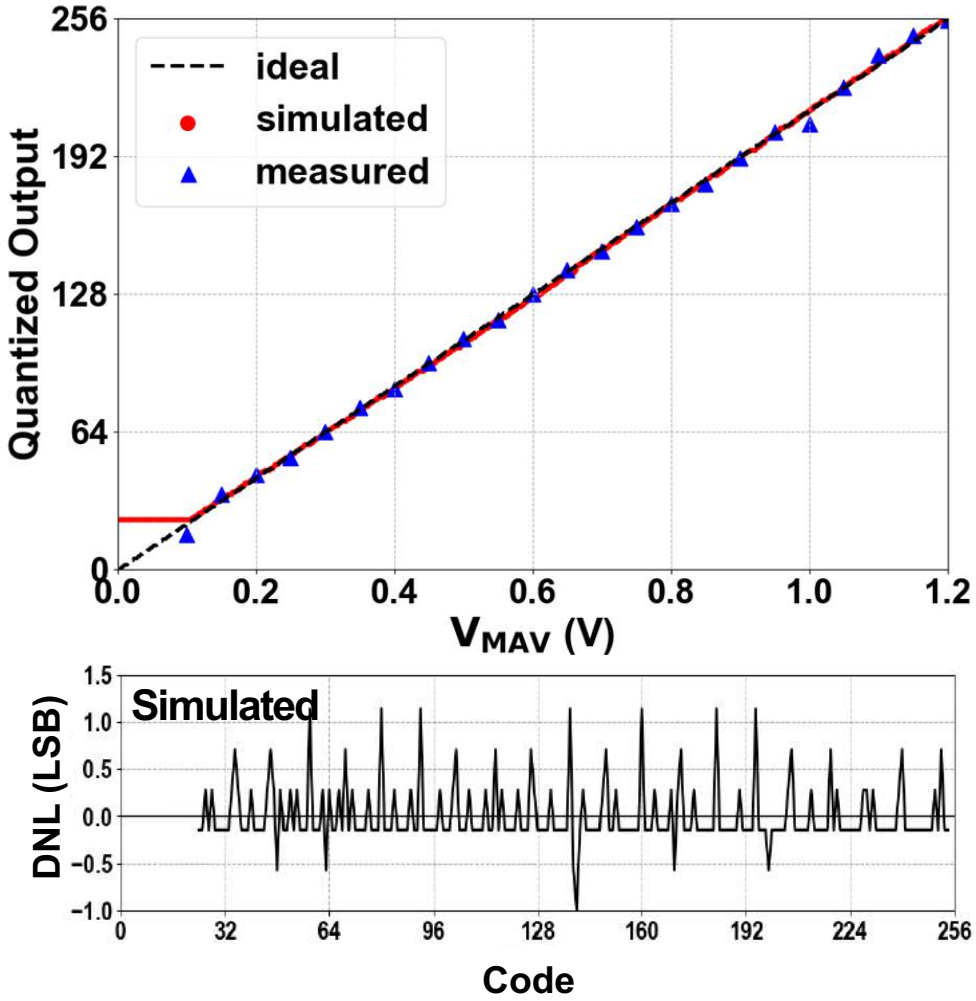
*ADC with clipping energy/without clipping energy

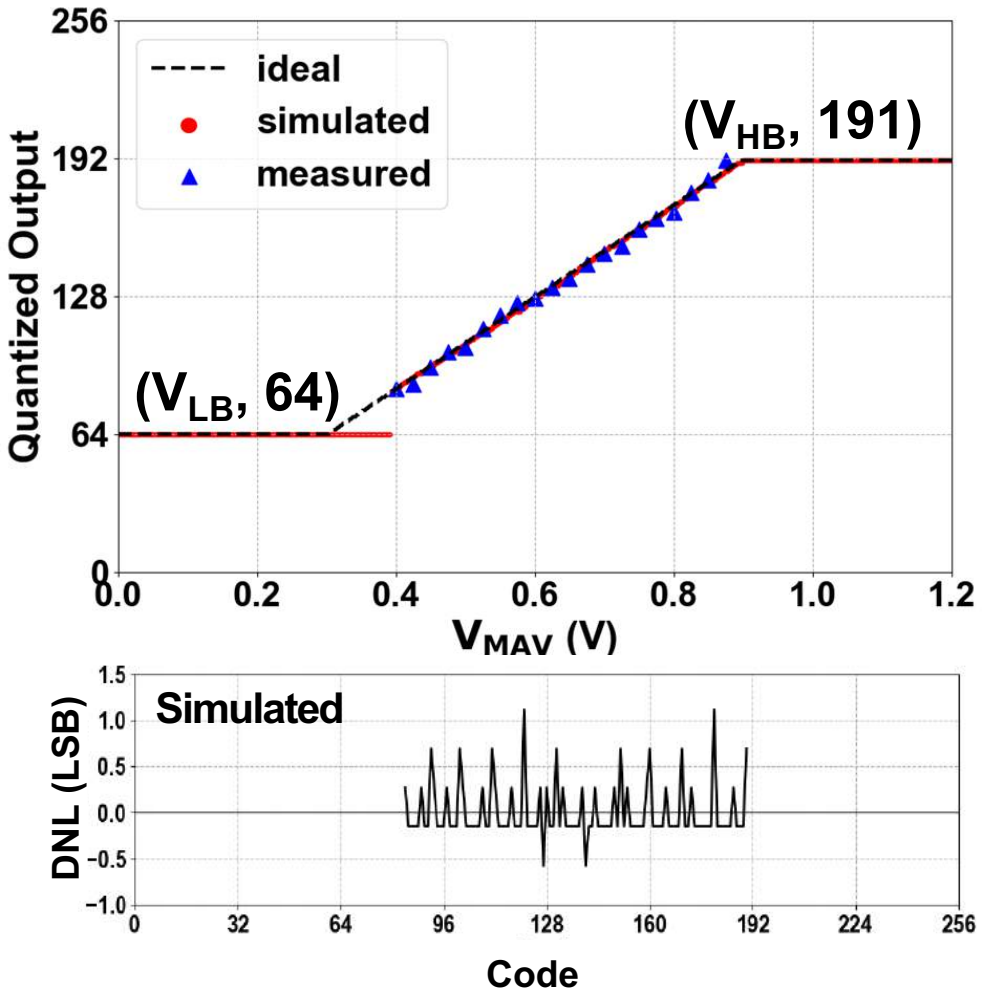
Clipping vs Non-Clipping ADC Tradeoff



- Throughput is improved by 1.14X while incurring 3.79% (1.21%) drop in Top-1 (Top-5) in CIFAR-10

In-eDRAM ADC Measurement Results





- Linear for both clipping and non-clipping in-eDRAM ADC

Outline

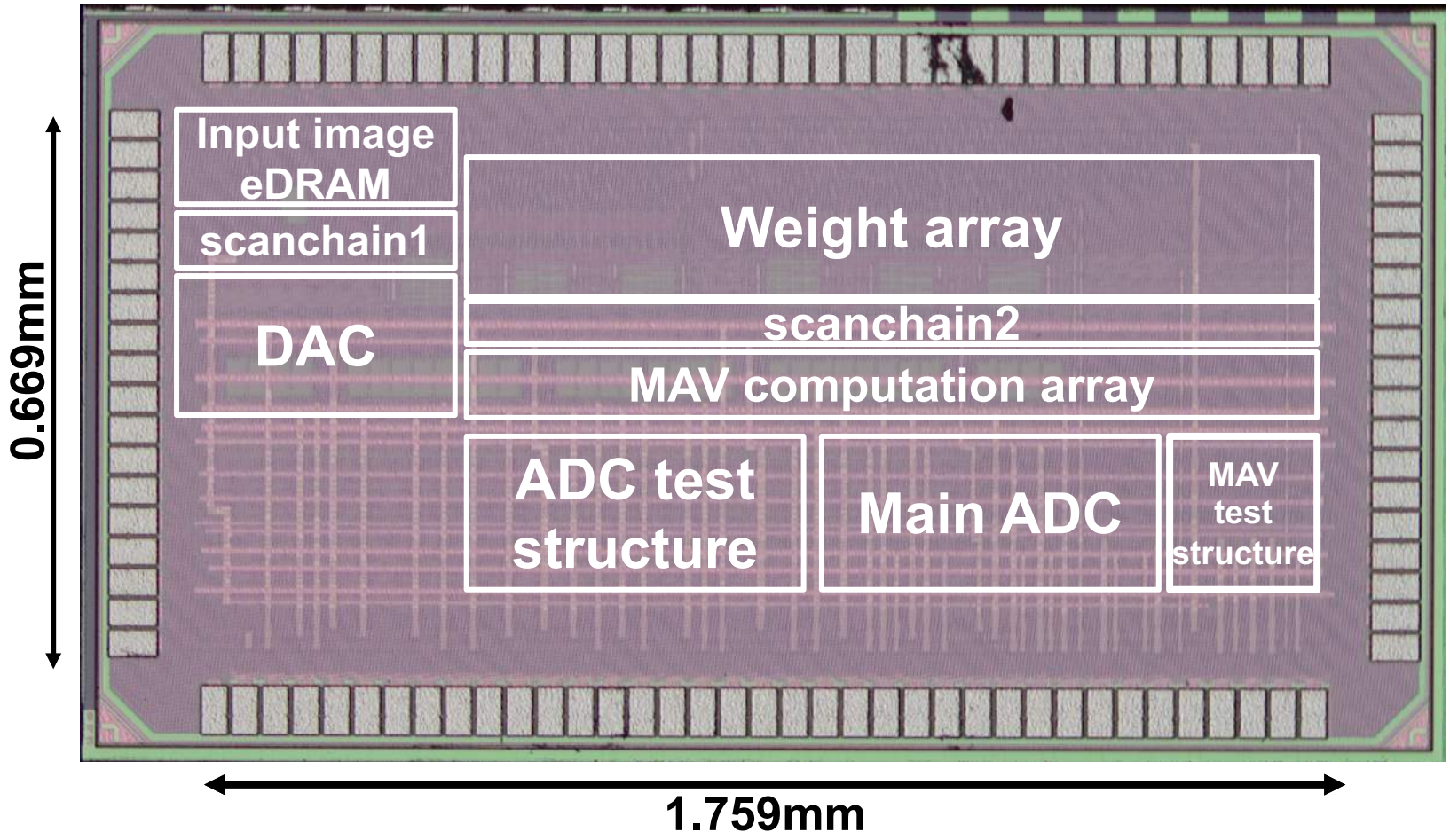
- Compute-in-memory: motivation and challenges
- Proposed eDRAM-CIM macro
  - Overall architecture
  - Digital-to-analog conversion (DAC)
  - Multiply-accumulate-averaging (MAV) computation
  - Analog-to-digital conversion (ADC)
- **Chip measurement and classification accuracy results**
- Summary

Chip Micrograph

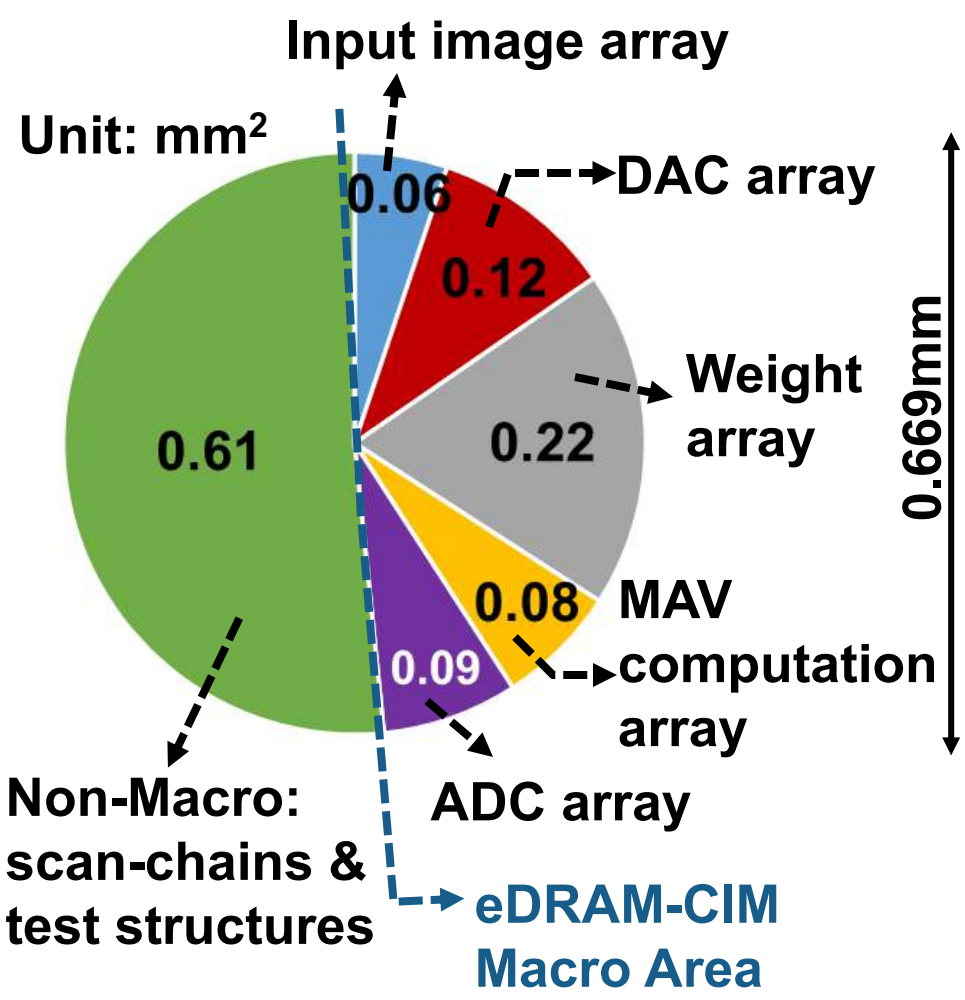
- 65nm native CMOS process
- Supply Voltage: 1~1.2V
- Input Clock Frequency:

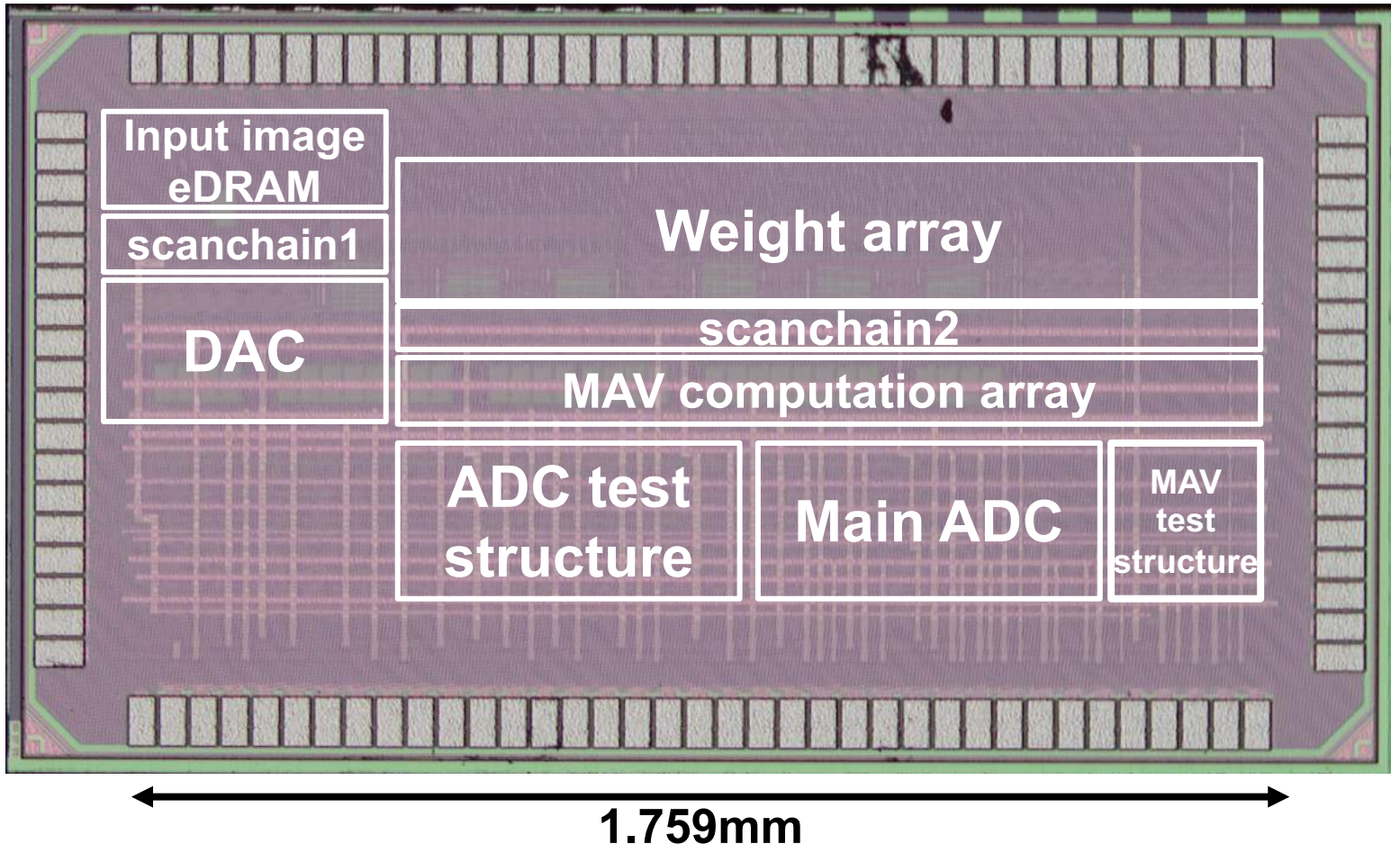
  50~200MHz
- Bitcell Size: 22.08um^2
- Macro Area: 0.57mm^2
- eDRAM Capacitor Type:

  Metal-Oxide-Metal (MOM)

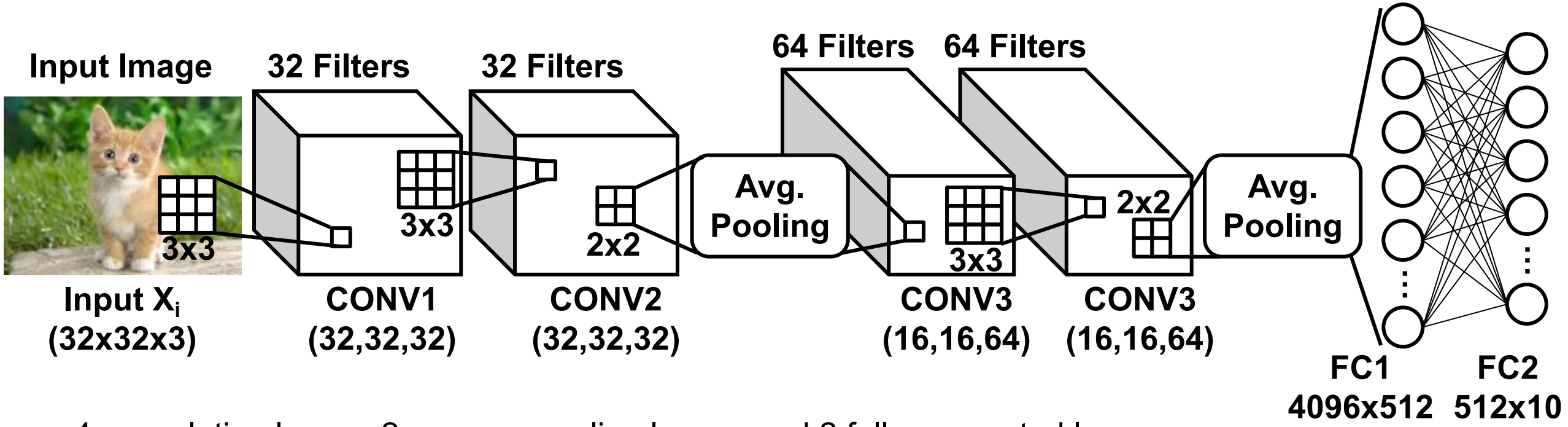


Test-chip Area Breakdown



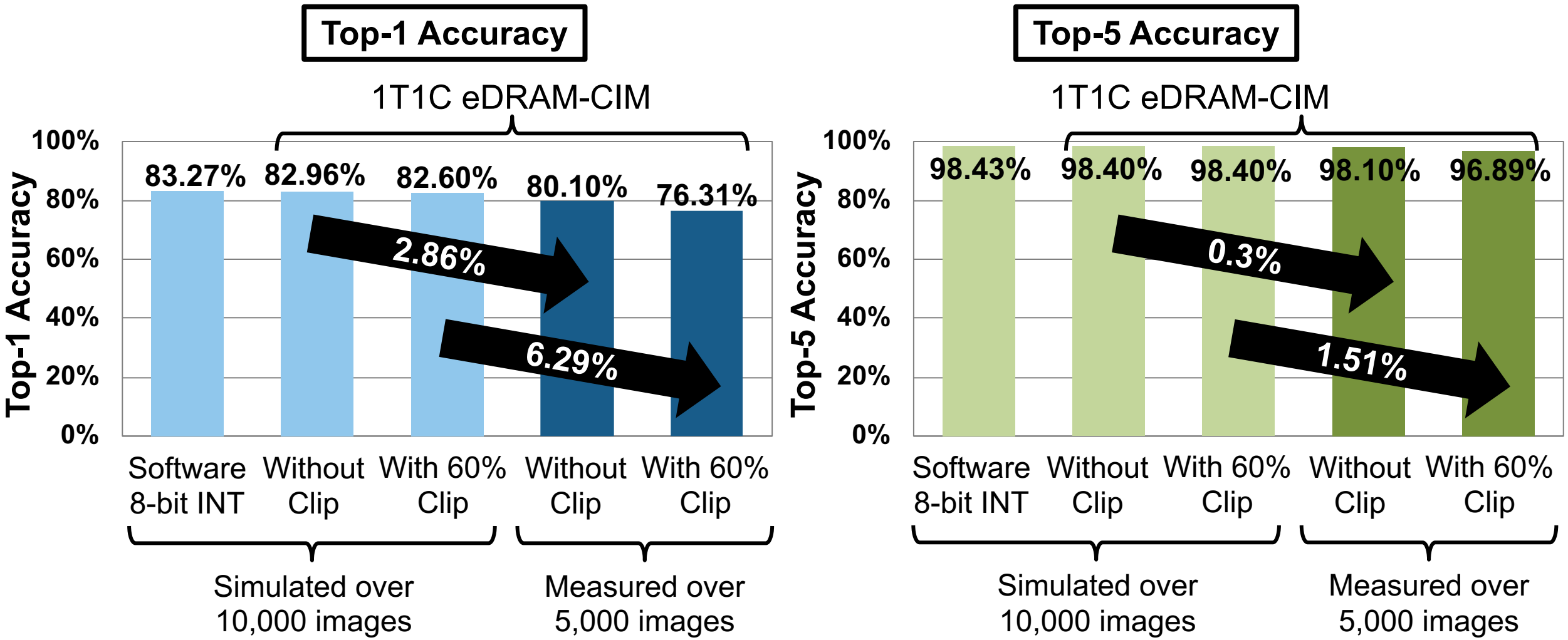


CIFAR-10 Classification Details



- 4 convolution layers, 2 average pooling layers, and 2 fully connected layers
- Input precision: 8 bits
- Weight precision: signed 8 bits
- Output precision: 8 bits

Classification Accuracy on CIFAR-10 Dataset



- Measured Top-5 accuracy is dropped 0.3% with non-clipping ADC and dropped 1.51% with clipping ADC
- Comparing with 8b software accuracy, eDRAM-CIM preserved good accuracy results

Comparison with Prior Works

	This work	ISSCC'20 [2]	ISSCC'20 [3]	ISSCC'20 [4]	ISSCC'18 [5]
Technology	**65nm**	28nm	28nm	22nm	65nm
Memory Cell Structure	**1T1C eDRAM**	6T SRAM	6T + Local Computing SRAM	1T1R SLC ReRAM	6T SRAM
Array Size	**16Kb**	64Kb	64Kb	2Mb	128Kb
Input Precision (bit)	**8**	8	8	4	8
Weight Precision (bit)	**8**	8	8	4	8
Supply Voltage (V)	**1~1.2**	0.85~1.0	0.7~0.9	0.8	1
Dataset	**CIFAR-10**	CIFAR-10			MNIST
Model	**CNN: 4 CONV + 2 Pooling + 2 FC**	CNN: ResNet-20	CNN: ResNet-20	N/A	SVM

[1]measured at 1.1V [2]Scaled to 65nm, assume energy ∝ (Tech.)^2 [7] [3] Limited by clocking infrastructure, chip size, technology and bit cell area

[4]FoM = input precision x weight precision x energy efficiency (scaled to 65nm) [5]Top-1 or Top-5 is not mentioned

Comparison with Prior Works

	This work	ISSCC'20 [2]	ISSCC'20 [3]	ISSCC'20 [4]	ISSCC'18 [5]
Technology	**65nm**	28nm	28nm	22nm	65nm
Memory Cell Structure	**1T1C eDRAM**	6T SRAM	6T + Local Computing SRAM	1T1R SLC ReRAM	6T SRAM
Array Size	**16Kb**	64Kb	64Kb	2Mb	128Kb
Input Precision (bit)	**8**	8	8	4	8
Weight Precision (bit)	**8**	8	8	4	8
Supply Voltage (V)	**1~1.2**	0.85~1.0	0.7~0.9	0.8	1
Dataset	**CIFAR-10**	CIFAR-10			MNIST
Model	**CNN: 4 CONV + 2 Pooling + 2 FC**	CNN: ResNet-20	CNN: ResNet-20	N/A	SVM
Measured Accuracy	**80.1% (Top-1), 98.1 % (Top-5)**	[5]91.91%	[5]92.02%	N/A	[5]83.27%

[1]measured at 1.1V [2]Scaled to 65nm, assume energy ∝ (Tech.)[2] [7] [3] Limited by clocking infrastructure, chip size, technology and bit cell area

[4]FoM = input precision x weight precision x energy efficiency (scaled to 65nm) [5]Top-1 or Top-5 is not mentioned

Comparison with Prior Works

	This work	ISSCC'20 [2]	ISSCC'20 [3]	ISSCC'20 [4]	ISSCC'18 [5]
Technology	**65nm**	28nm	28nm	22nm	65nm
Memory Cell Structure	**1T1C eDRAM**	6T SRAM	6T + Local Computing SRAM	1T1R SLC ReRAM	6T SRAM
Array Size	**16Kb**	64Kb	64Kb	2Mb	128Kb
Input Precision (bit)	**8**	8	8	4	8
Weight Precision (bit)	**8**	8	8	4	8
Supply Voltage (V)	**1~1.2**	0.85~1.0	0.7~0.9	0.8	1
Dataset	**CIFAR-10**	CIFAR-10			MNIST
Model	**CNN: 4 CONV + 2 Pooling + 2 FC**	CNN: ResNet-20	CNN: ResNet-20	N/A	SVM
Measured Accuracy	**80.1% (Top-1), 98.1 % (Top-5)**	[5]91.91%	[5]92.02%	N/A	[5]83.27%
Throughput (GOPS)	**[1,3]4.71**	N/A	N/A	N/A	4
Average Energy Efficiency (TOPS/W)	**[1]4.76**	7.3 [2](1.35)	14.08 [2](2.61)	28.93 [2](3.31)	3.125
GOPS/mm^2	**8.26**	N/A	N/A	N/A	2.78
[4]FoM	**304.6**	86.4	167	53	201.6

[1]measured at 1.1V [2]Scaled to 65nm, assume energy ∝ (Tech.)^2 [7] [3] Limited by clocking infrastructure, chip size, technology and bit cell area

[4]FoM = input precision x weight precision x energy efficiency (scaled to 65nm) [5]Top-1 or Top-5 is not mentioned

Test-chip Summary

		Test-chip Summary	Advanced Technology Node Estimation
Technology		65nm CMOS	22nm tri-gate CMOS
Total eDRAM Size		16Kb	[2]38.4MB
Bitcell Size (μm^2)		22.08	0.029
Unit eDRAM Capacitor Size (fF)		13	13
eDRAM Capacitor Type		Metal-oxide-metal (MOM)	Capacitor-over-bitline (COB)
Input Clock Frequency		50 MHz (measured) 200MHz (simulation)	1.6GHz
GOPS/mm^2		8.26	1305
MAV Operation Latency (ns)		10 (2 cycles)	1.25 (2 Cycles)
[1]Throughput @ 1.1V	**w/o ADC Clipping**	4.037 GOPS	19.5 TOPS
	with 60% ADC Clipping	4.71 GOPS	22.2 TOPS
[3]Average Energy Efficiency (TOPS/W) @ 1.1V		4.76	11.12
Energy/MAV (fJ/MAV) with 60% ADC Clipping @ 1.1V		30.6	[4]0.21

[1]Assume 1MAV = 2 operations

[2]Assume 30% utilization of 128MB (77mm^2) [6] eDRAM memory for CIM usage

[3]Excluded scan-chains power

[4]Increased parallelism in large-capacity eDRAM array

Outline

- Compute-in-memory: motivation and challenges
- Proposed eDRAM-CIM macro
  - Overall architecture
  - Digital-to-analog conversion (DAC)
  - Multiply-accumulate-averaging (MAV) computation
  - Analog-to-digital conversion (ADC)
- Chip measurement and classification accuracy results
- **Summary**

Summary

- 1T1C eDRAM bit cell: good candidate for compute-in-memory design
- Repurposing existing 1T1C eDRAM bitcell
- Intrinsic charge share operation in eDRAM bitcell
- Support 8b input and 8b signed/unsigned weight for MAV operations
- Measured Top-5 (Top-1) peak accuracy 98.1% (80.1%) on CIFAR-10
- Measured average energy efficiency 4.76 TOPS/W
- Measured throughput 4.71GOPS
- FoM metrics can exceed significantly in an advanced eDRAM technology
- Potential for scalability to advanced eDRAM technology node

Acknowledgements

- Clifford Ong for technical discussions
- Intel for funding support

Thank you for your attention!