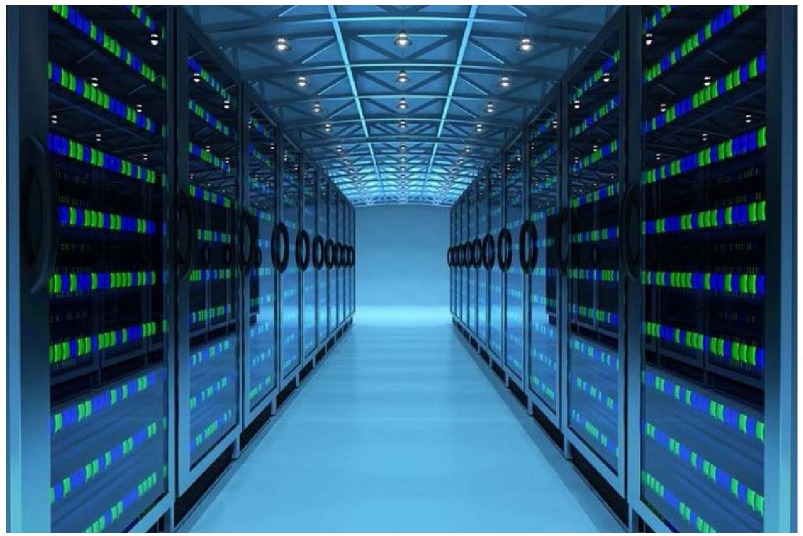
An All-Weights-on-Chip DNN Accelerator in 22nm ULL Featuring 24×1 Mb eRRAM

*Zhehong Wang[1], *Ziyun Li[1,2], Li Xu[1], Qing Dong[3],
Chin-I Su[4], Wen-Ting Chu[4], George Tsou[4],
Yu-Der Chih[4], Tsung-Yung Jonathan Chang[4],
Dennis Sylvester[1], Hun Seok Kim[1], David Blaauw[1]

[1]University of Michigan, Ann Arbor, MI
[2]Facebook, Seattle, WA
[3]TSMC, San Jose, CA
[4]TSMC, Hsinchu, Taiwan

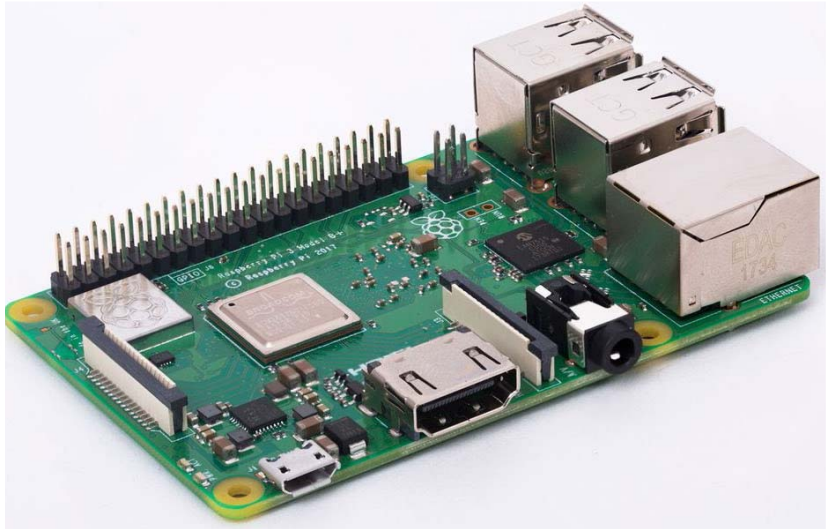
Motivation

- Machine Learning/Deep Neural Network applications exploded



Data Center

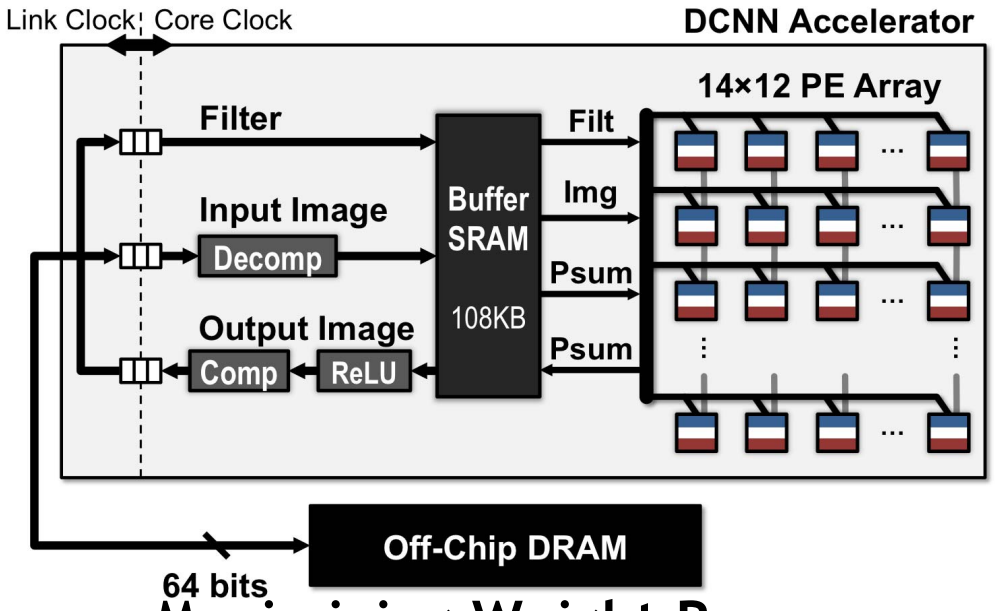
Mobile



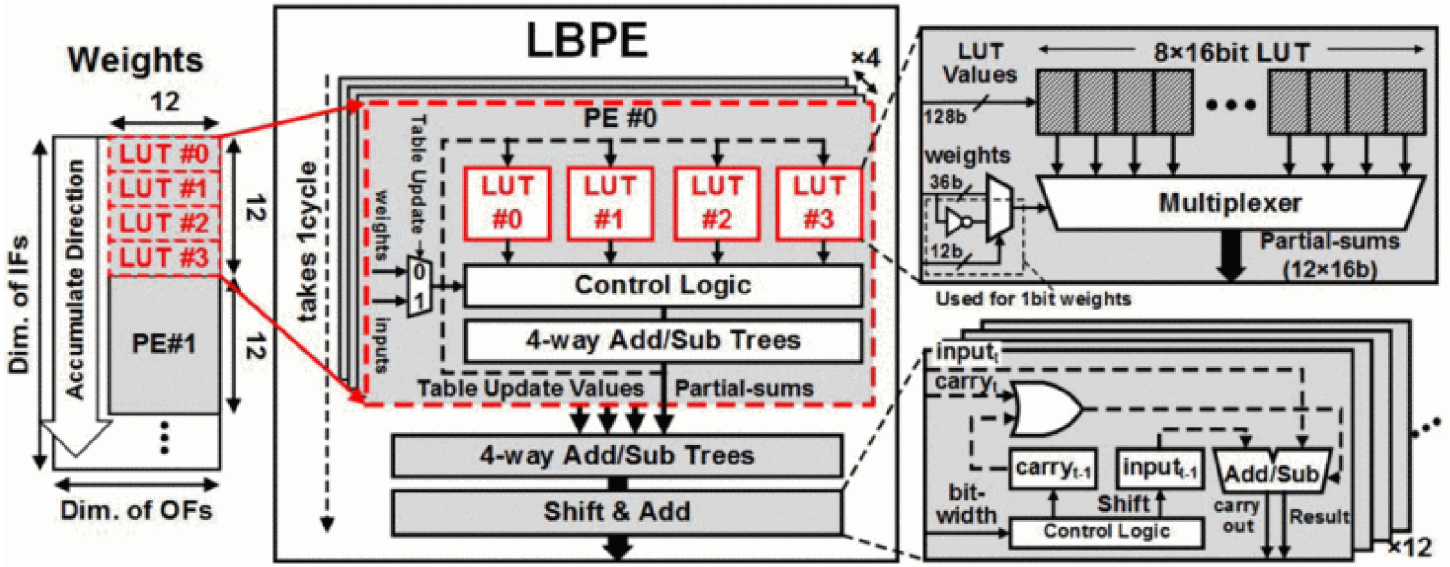
IOT

Motivation

- Various approaches to improve power efficiency



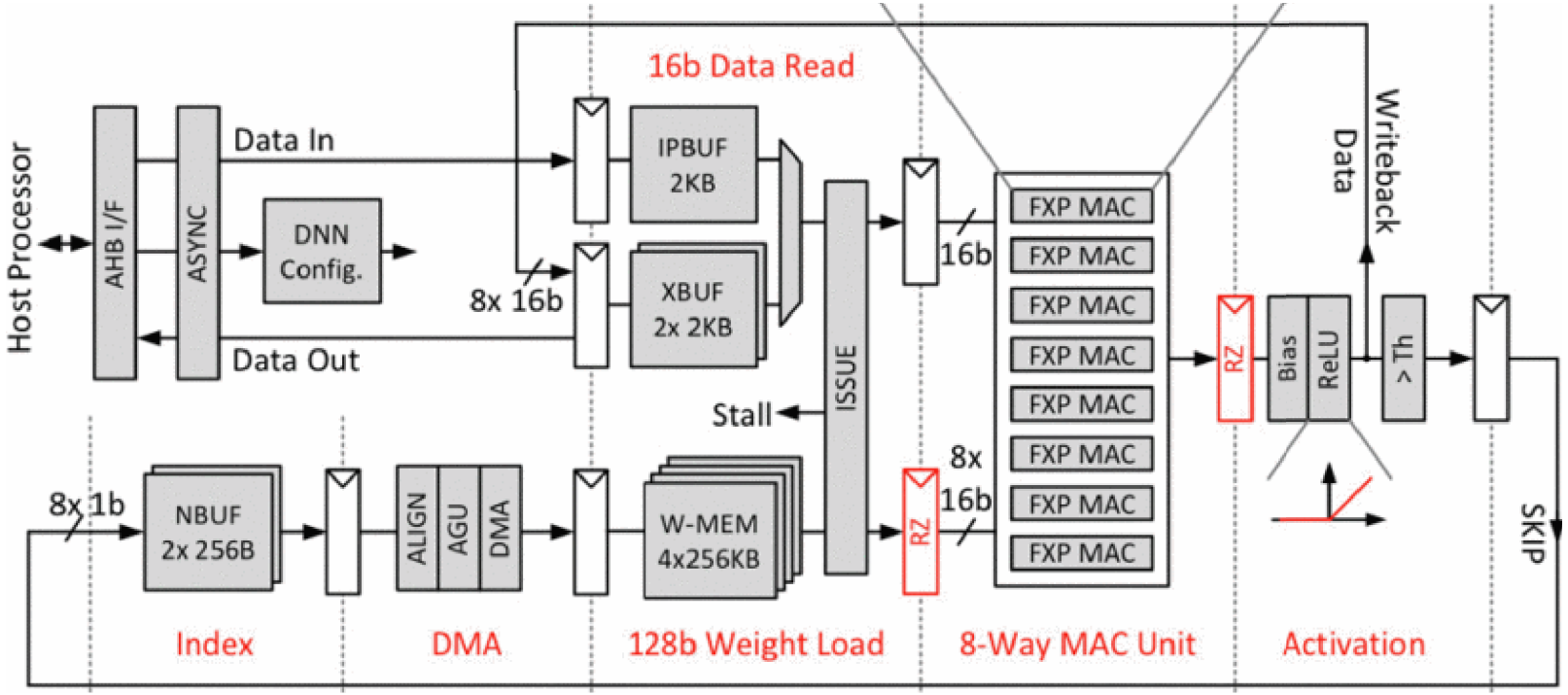
Maximizing Weight Reuse
[Y. Chen, ISSCC16]



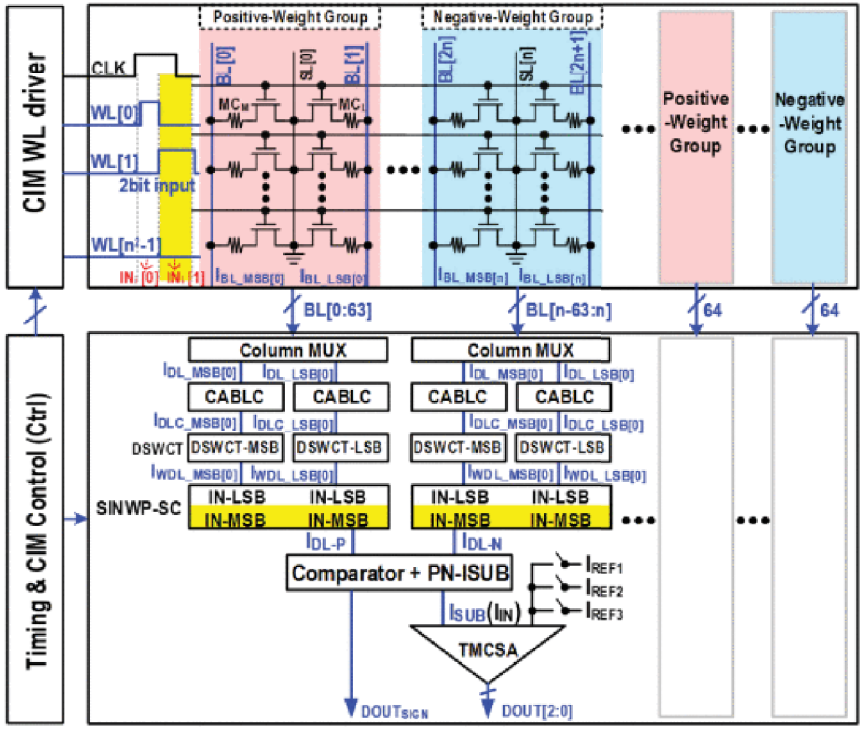
Bit serial
[J.Lee, ISSCC18]

Motivation

- Various approaches to improve power efficiency



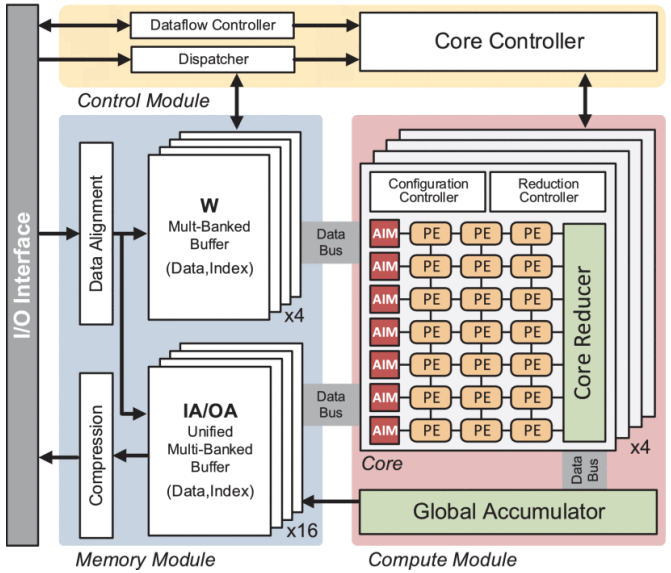
Sparsity Awareness
[P. Whatmough, ISSCC17]



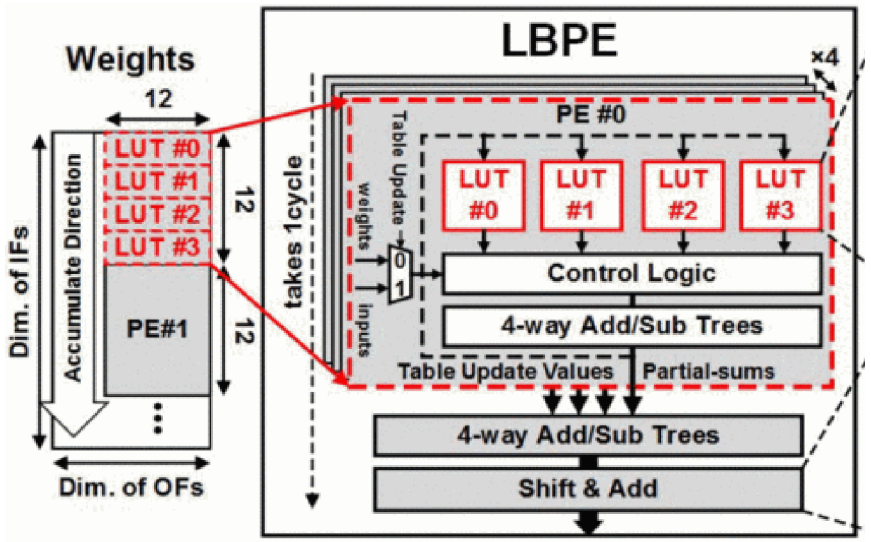
NVM CIM
[C. Xue, ISSCC19]

Motivation

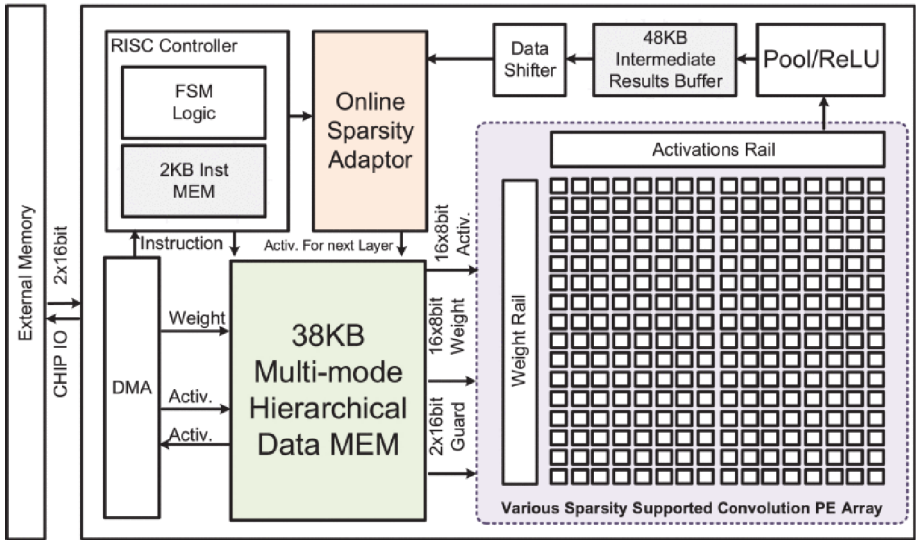
- Off-chip memory latency and power become bottleneck
- Large on-chip weight buffer



280.6KB
140.3K@16b
[J. Zhang, VLSI19]



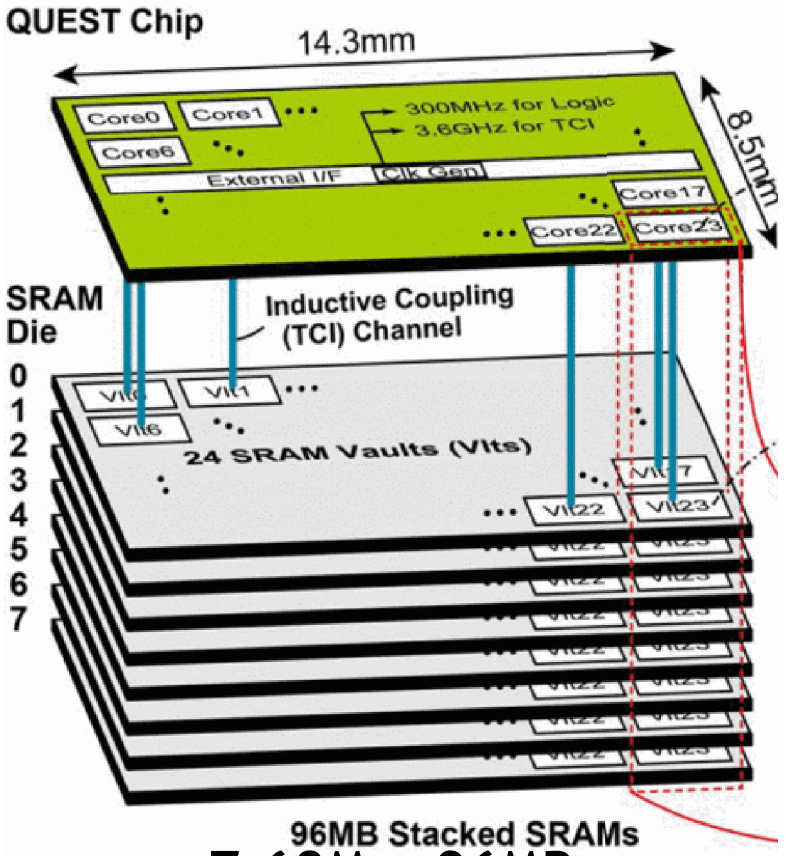
256KB
256K@8b
[J. Lee, ISSCC18]



170KB
170K@8b
[Z. Yuan, VLSI18]

Motivation

- Off-chip memory latency and power become bottleneck
- Large on-chip weight buffer
- Culminate in QUEST ISSCC18
  - 7.68MB on-chip SRAM
  - 96MB 3D stacked SRAM
  - 3.3W system power



7.68M + 96MB

15.36M@4b

[K. Ueyoshi, ISSCC18]

Motivation

- Off-chip memory latency and power become bottleneck
- Large on-chip weight buffer
- Culminate in QUEST ISSCC18
  - 7.68MB on-chip SRAM
  - 96MB 3D stacked SRAM
  - 3.3W system power
- Non-Volatile memory becomes an option

Our Contribution

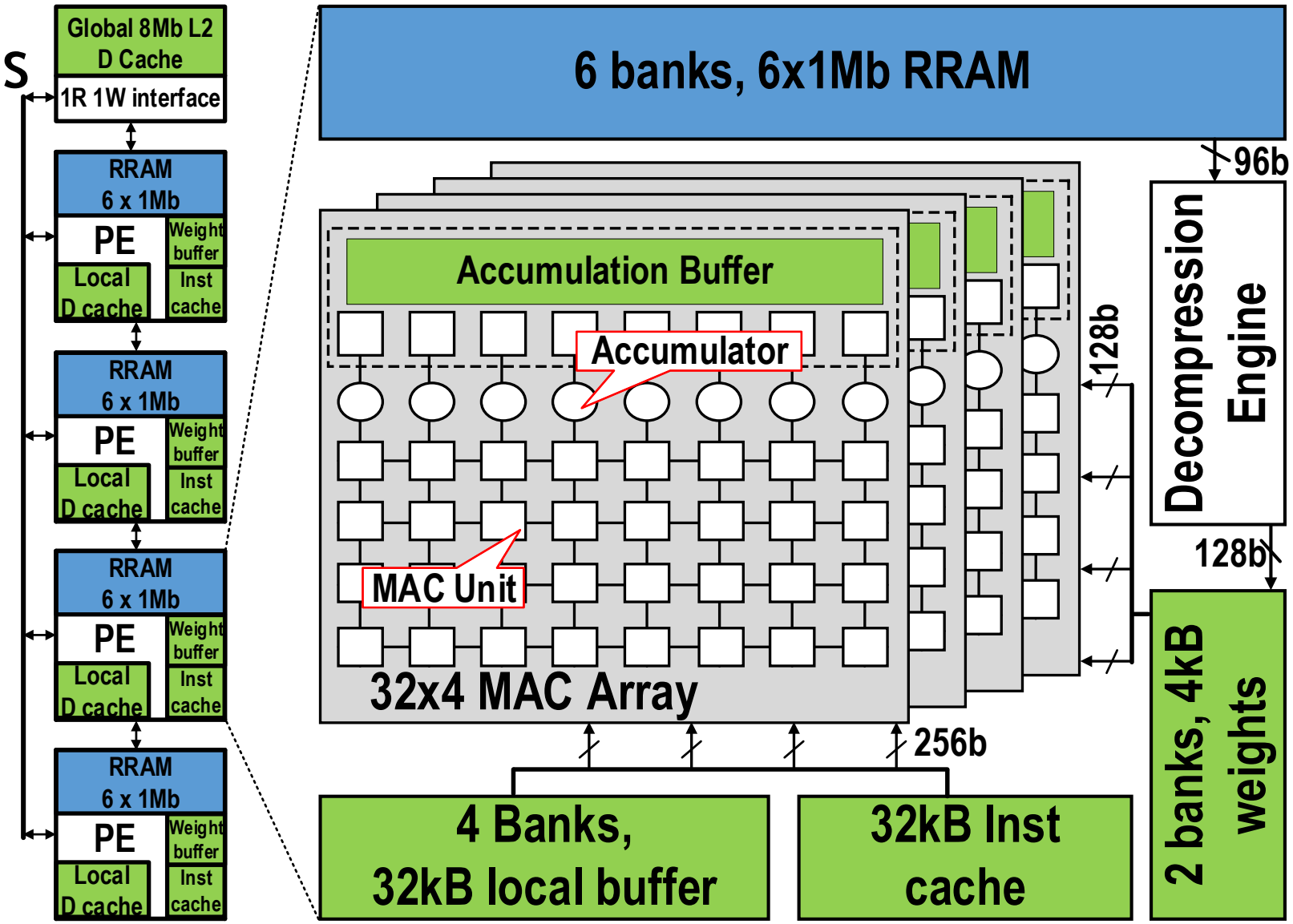
- The first digital DNN accelerator featuring 24 Mb eRRAM as dedicated weight storage to eliminate off-chip weight access
- Weight compression achieving 16 M 8-bit weights on-chip
- Dynamic clamping offset-canceling sense amplifier (DCOCSA) achieving sub-μA input offset

Outline

- Motivation
- All-Weights-on-Chip DNN Accelerator
- Test Results

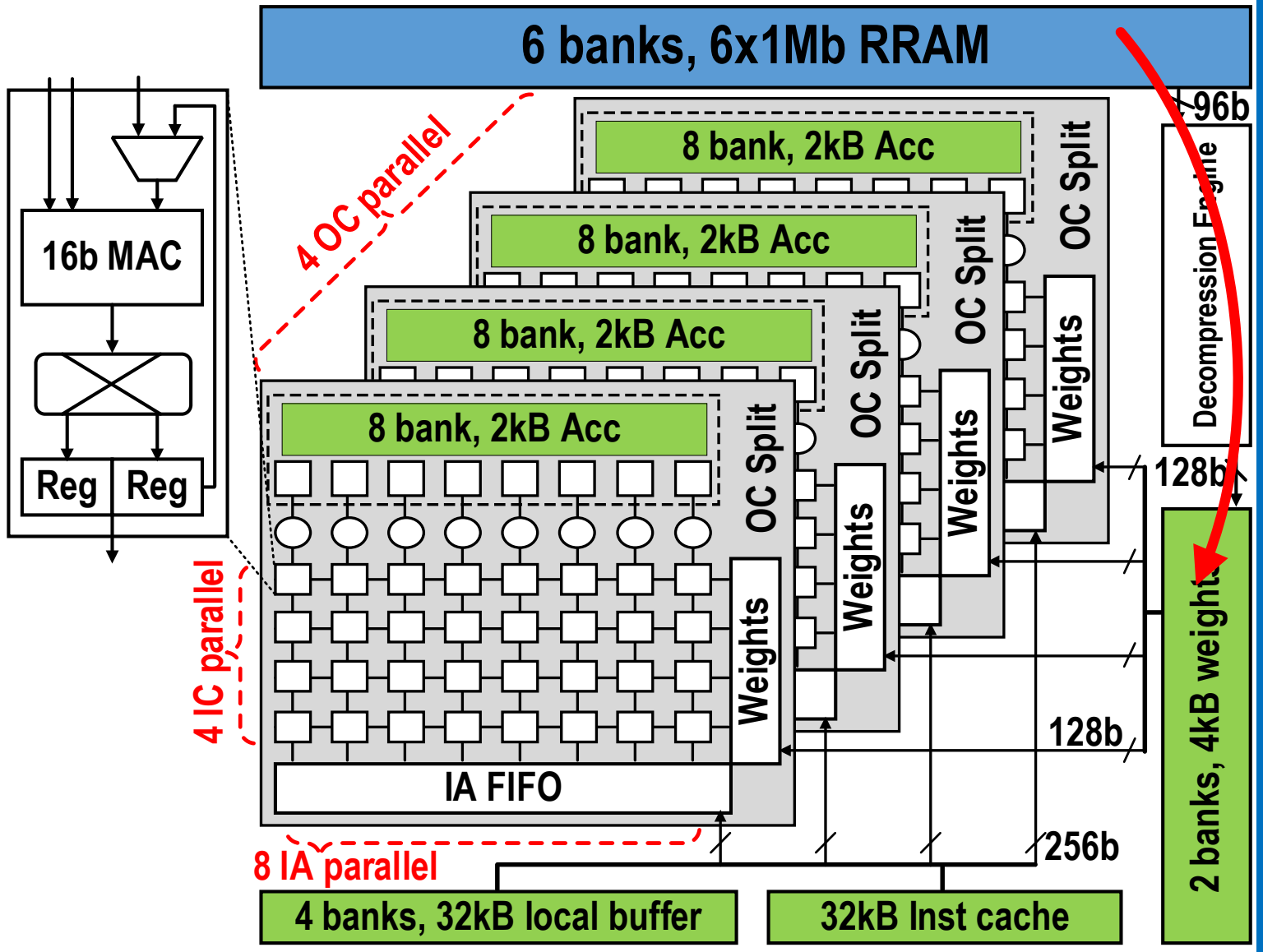
Chip and PE Architecture

- 4 mesh-connected PEs
- Each with
  - 128 MACs
  - Local 4KB weight
  - 32KB icache
  - 32KB input buffer
  - 6Mb RRAM
- 8Mb global buffer
- Overall 123 GOPS



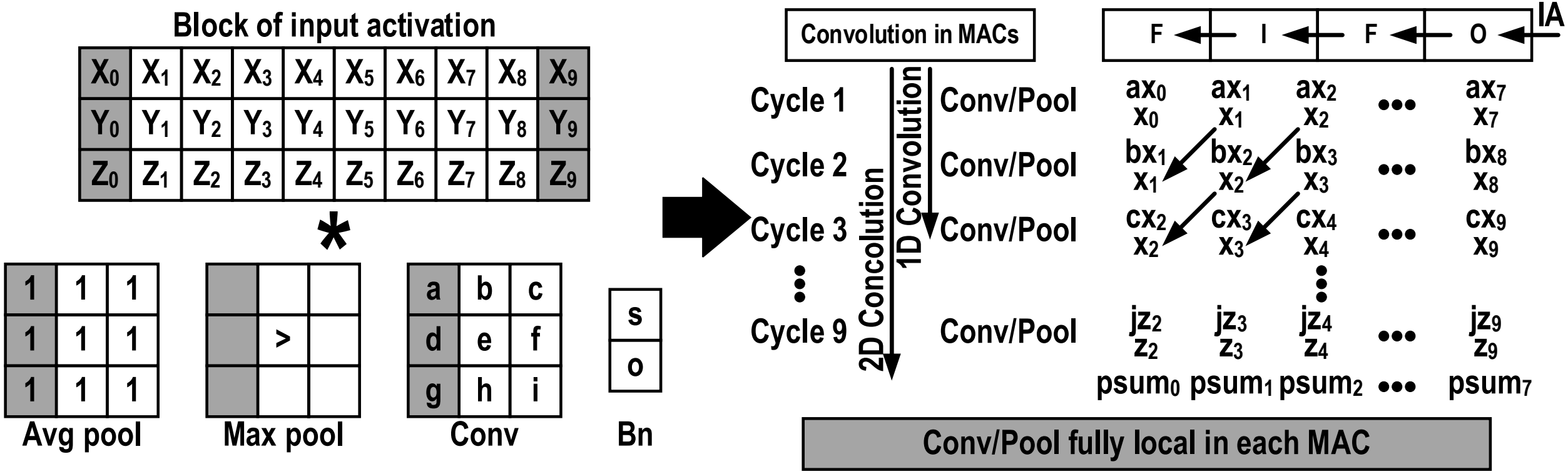
Chip and PE Architecture

- 4 clusters of 32 MACs
  - 8 IA parallel
  - 4 IC parallel
  - 4 OC parallel
- 256b VLIW ISA
- Compressed weights are read from RRAM and decompressed into weight buffer



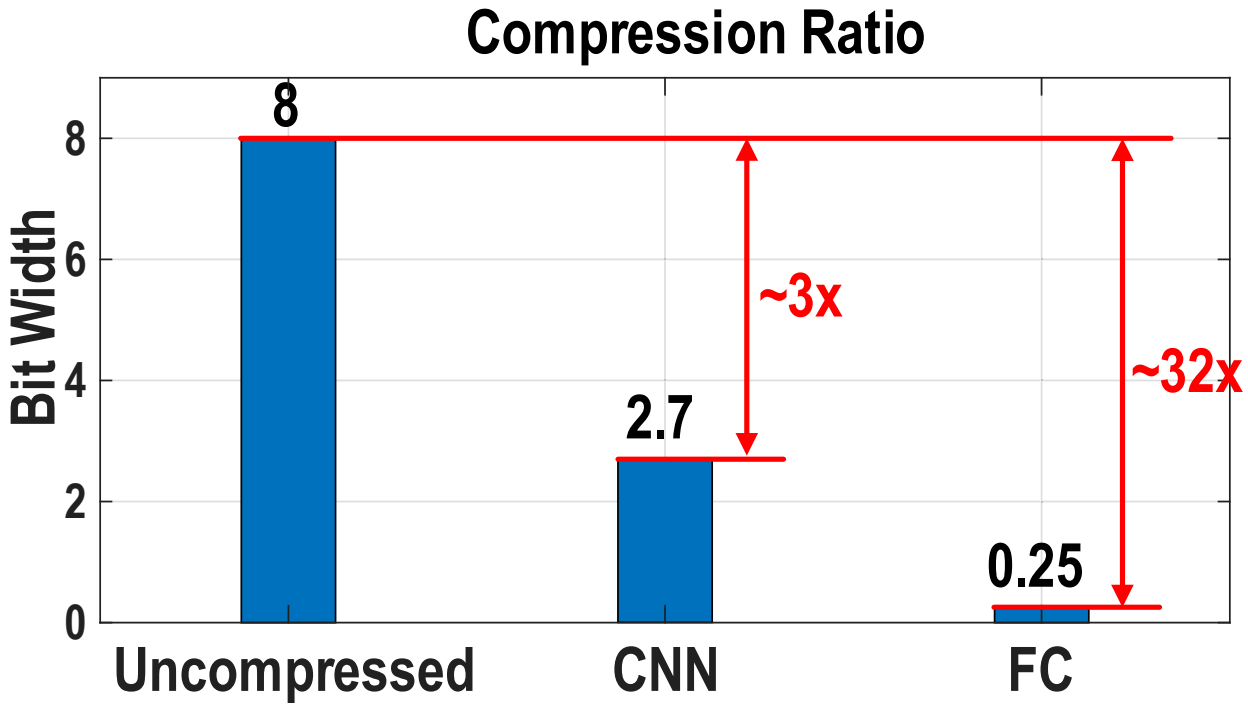
Neural Network Operations

- Similar to [Z.Li, ISSCC'19]
- Convolutional reuse & output reuse
- Combined Conv & BN & Nonlinear ops



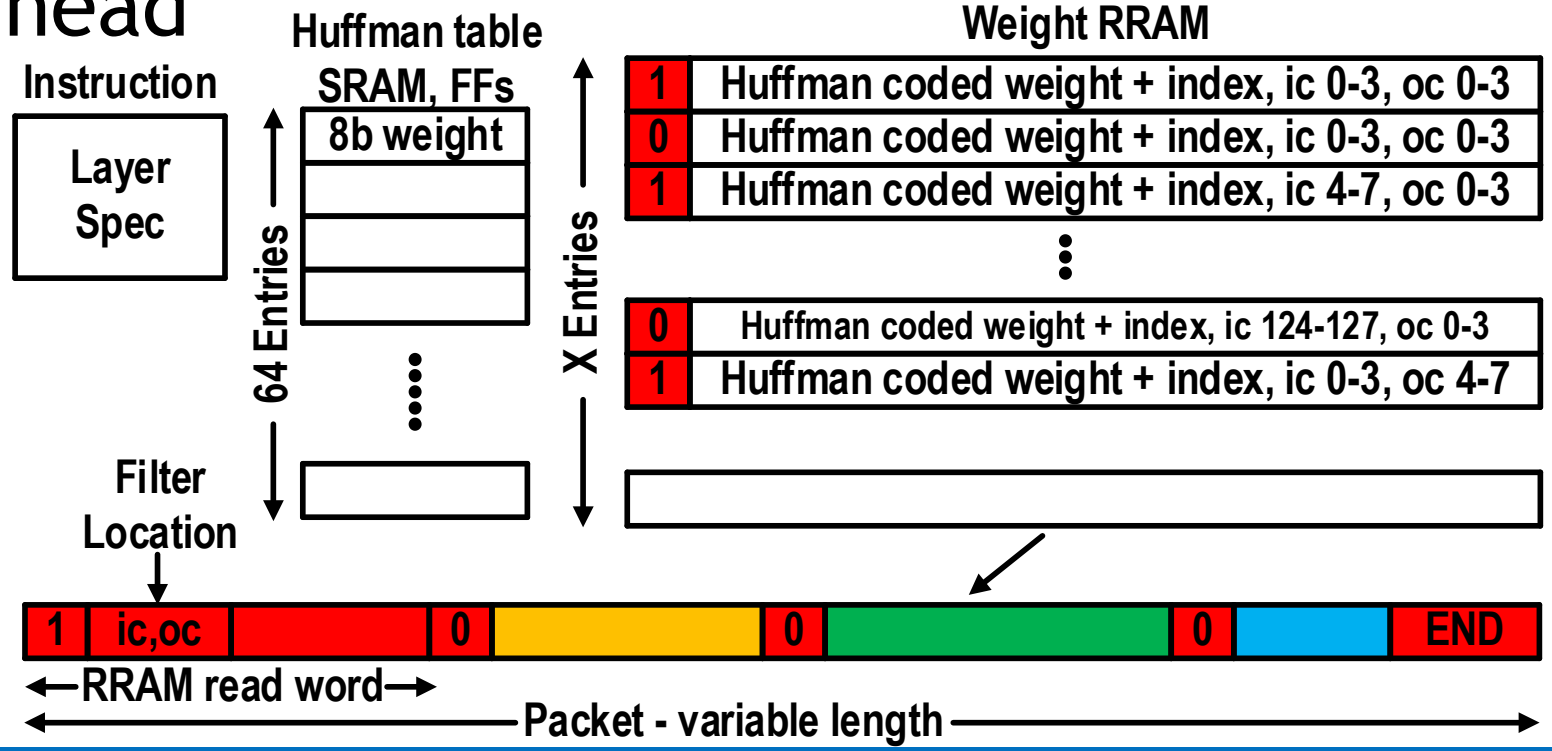
Weight Compression/Decompression

- Combining pruning, non-uniform quantization, run-length and Huffman encoding
- With 8b precision:
  - 2.7 bits per weight for CNN layers
    - 5.2 bits for non-zeros
  - 0.25 bit per weight for FC layers



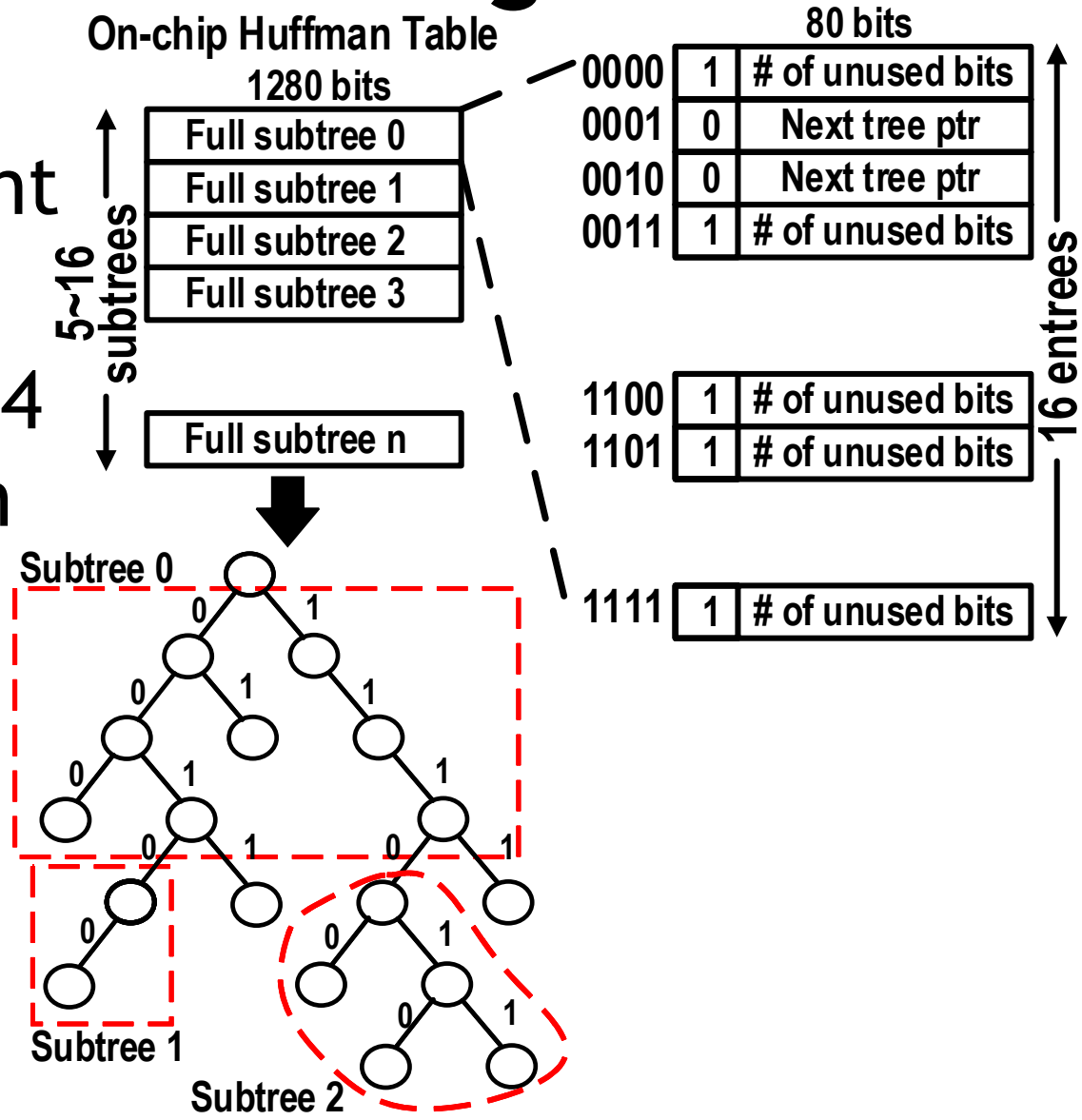
Weight Compression/Decompression

- Variable length weight packet with multiple 96b words
- Containing layer specification followed by Huffman-encoded weights and run-length coded indices
- ~5% total memory overhead



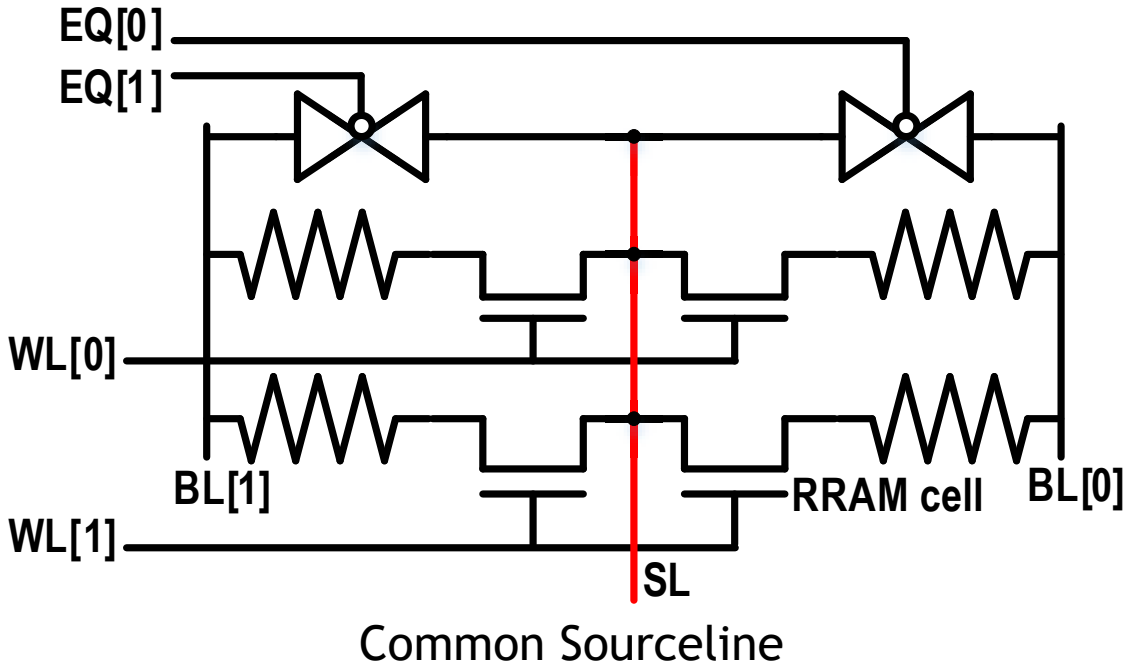
Parallel Decompression of Weights

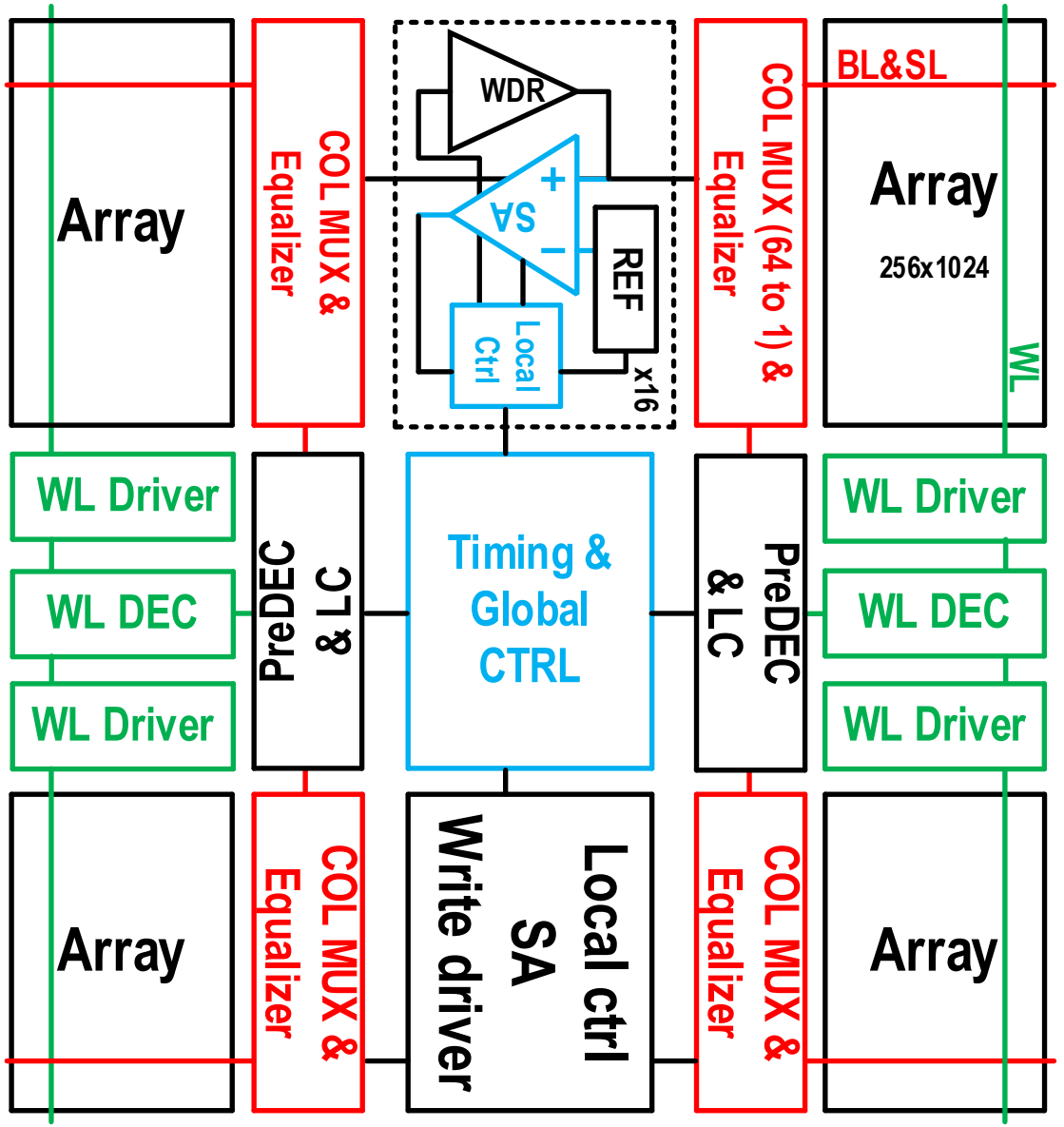
- Non-uniform quantization with Huffman encoding poses significant performance bottleneck
- Huffman codes are decoded with 4 bits in parallel → 2.7× faster than sequential decoding
- PEs decode in parallel, each maintaining ~10kb LUT
  - Updated via broadcast to all PEs



RRAM Bank Architecture

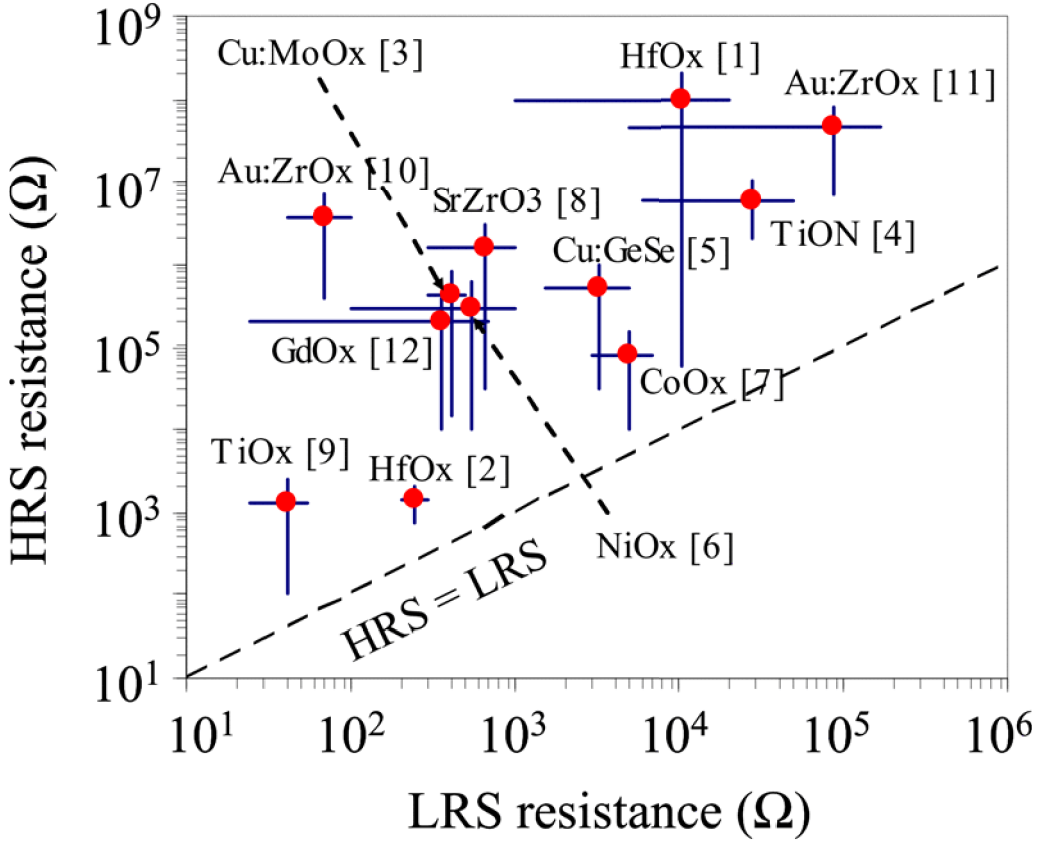
- 4 256x1024 RRAM arrays
- Common SL architecture
- 32-bit word length





Dynamic Clamping Offset-Canceling SA

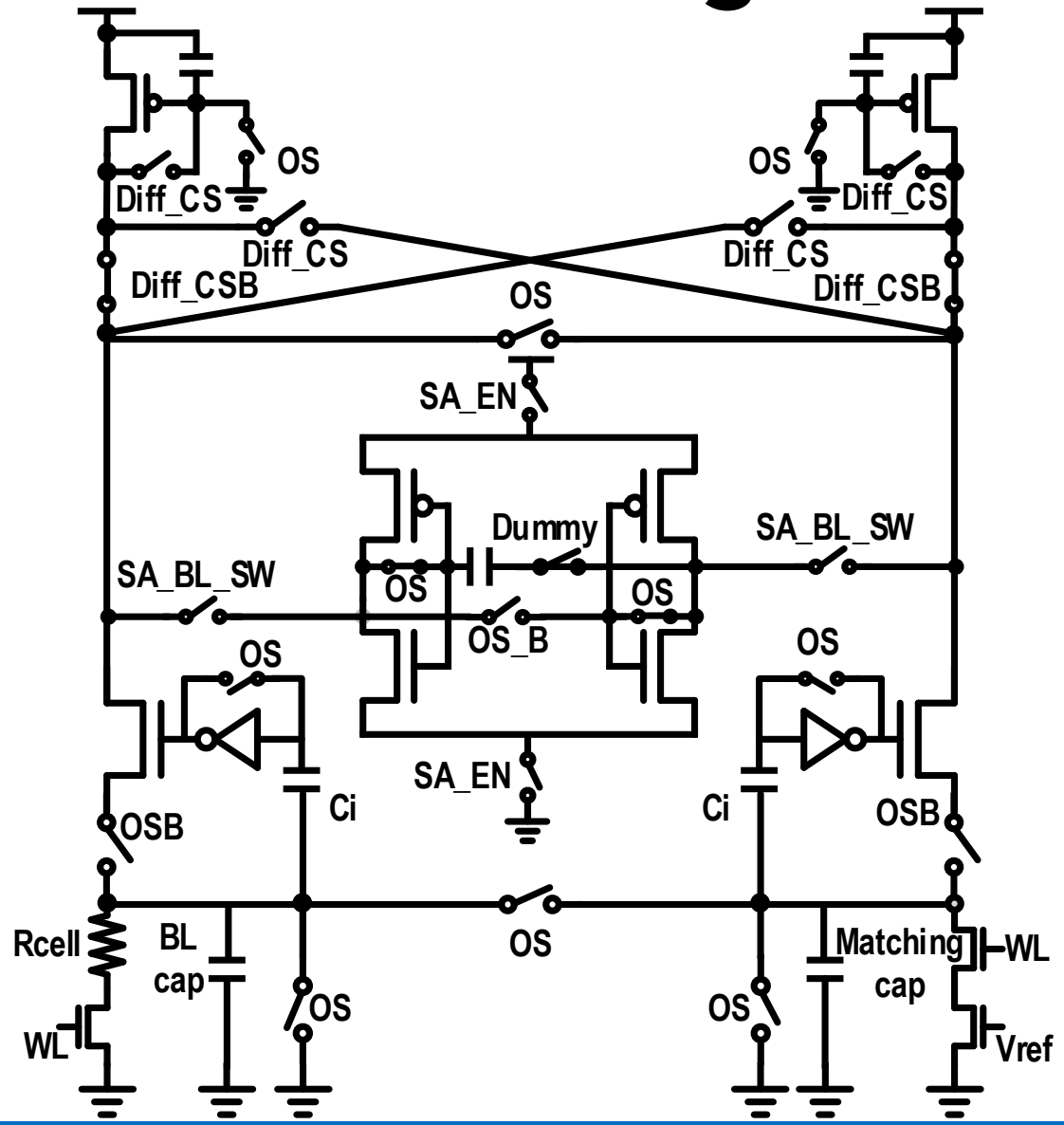
- High Variation of RRAM resistance
  - 2x~10x for low resistance state(LRS)
  - 5x~100x for high resistance state(LRS)
- Leaving small sensing margin on Sense Amplifier



[A. Chen, IRPS11]

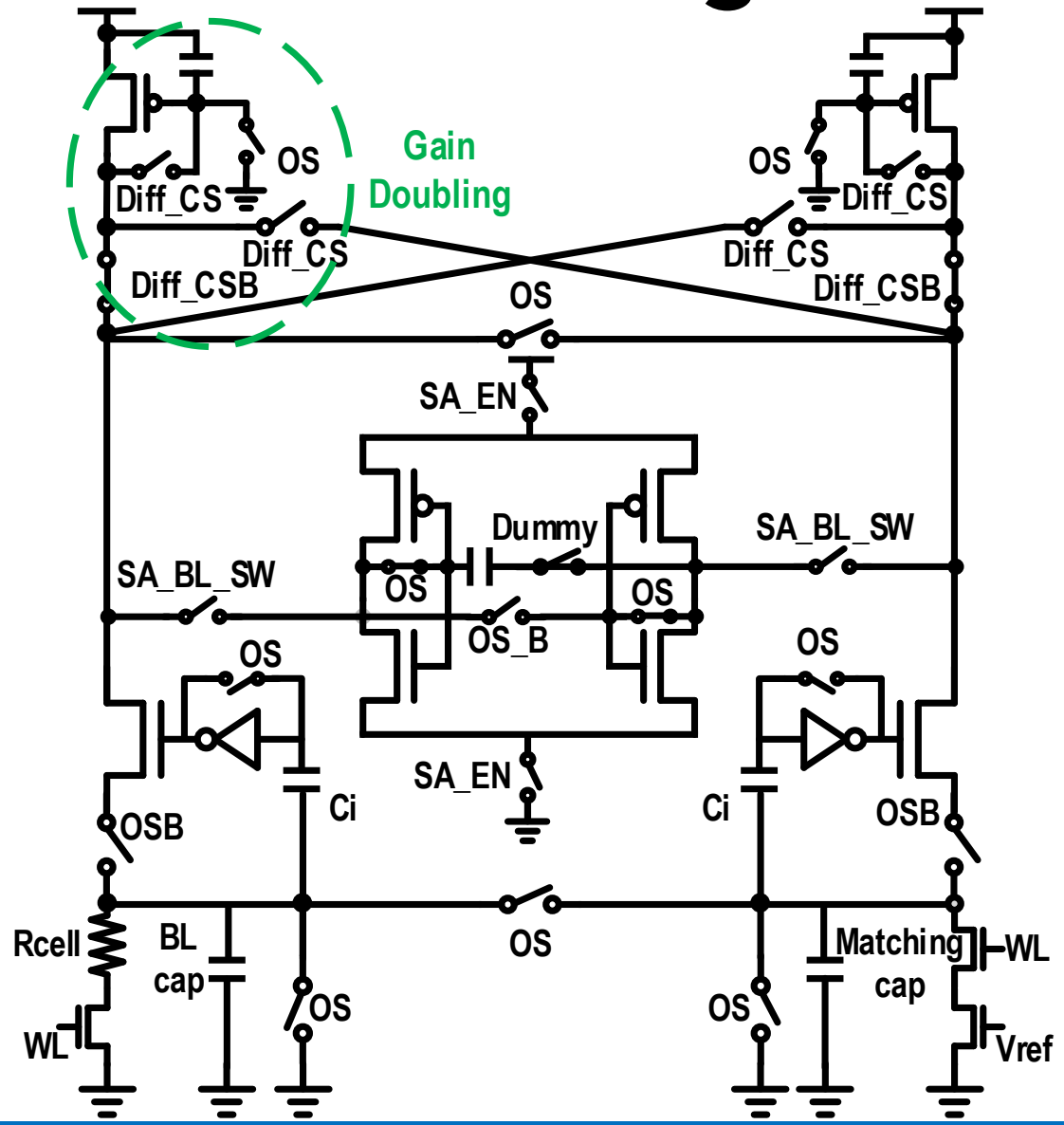
Dynamic Clamping Offset-Canceling SA

- Cross couple gain doubling
- Dynamic clamping
- Single cap auto-zeroing



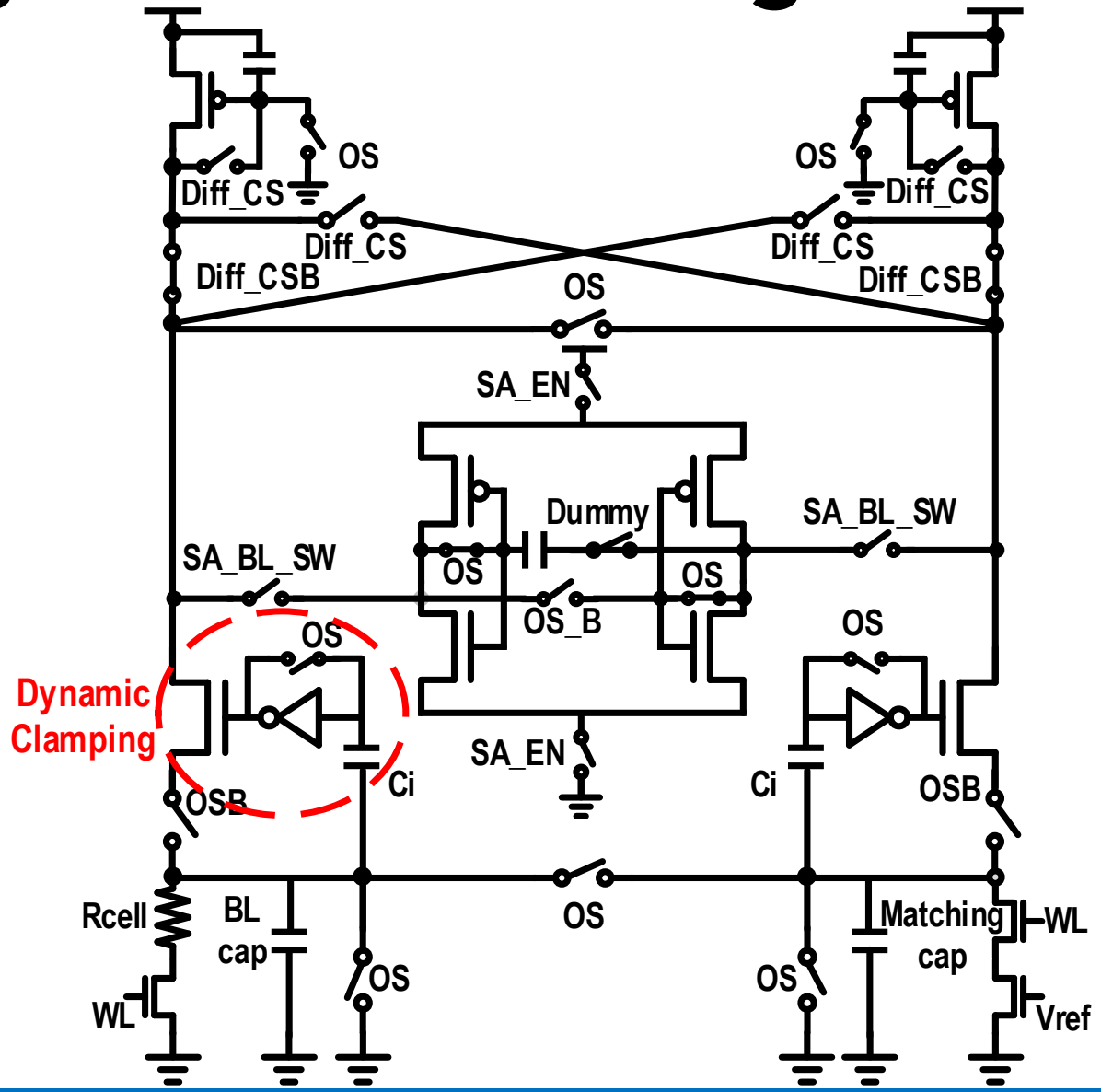
Dynamic Clamping Offset-Canceling SA

- Cross couple gain doubling
- Dynamic clamping
- Single cap auto-zeroing



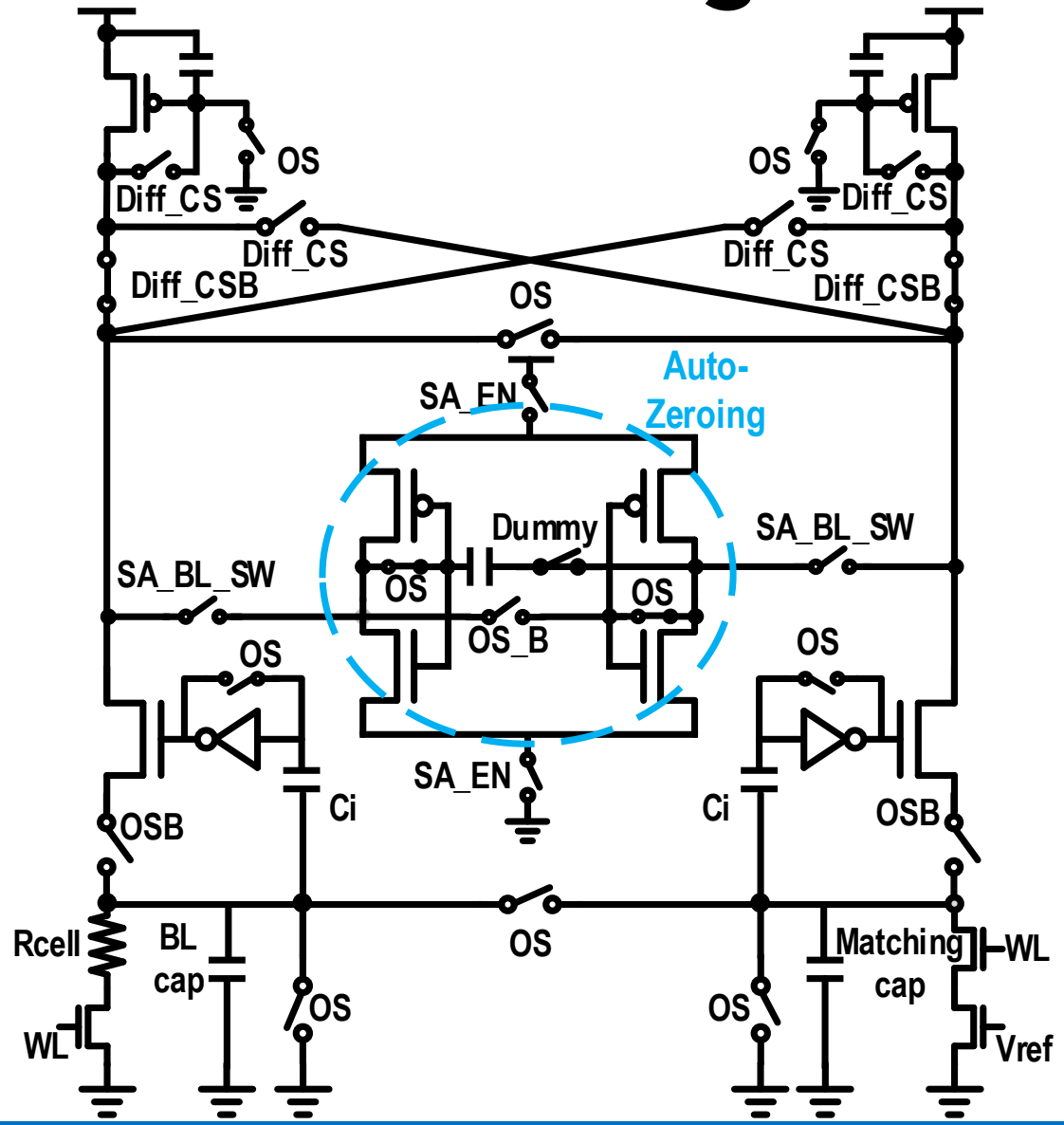
Dynamic Clamping Offset-Canceling SA

- Cross couple gain doubling
- Dynamic clamping
- Single cap auto-zeroing



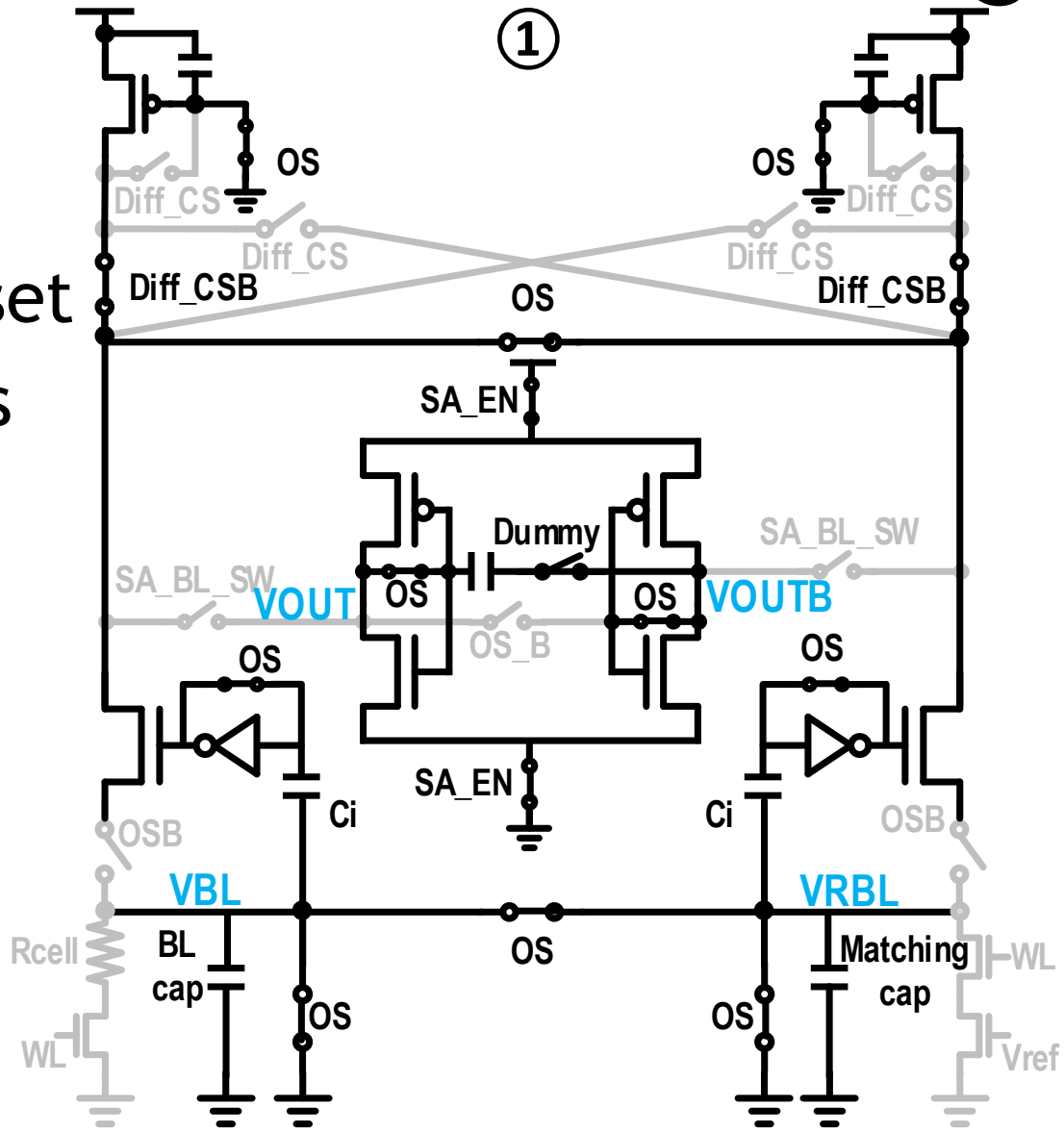
Dynamic Clamping Offset-Canceling SA

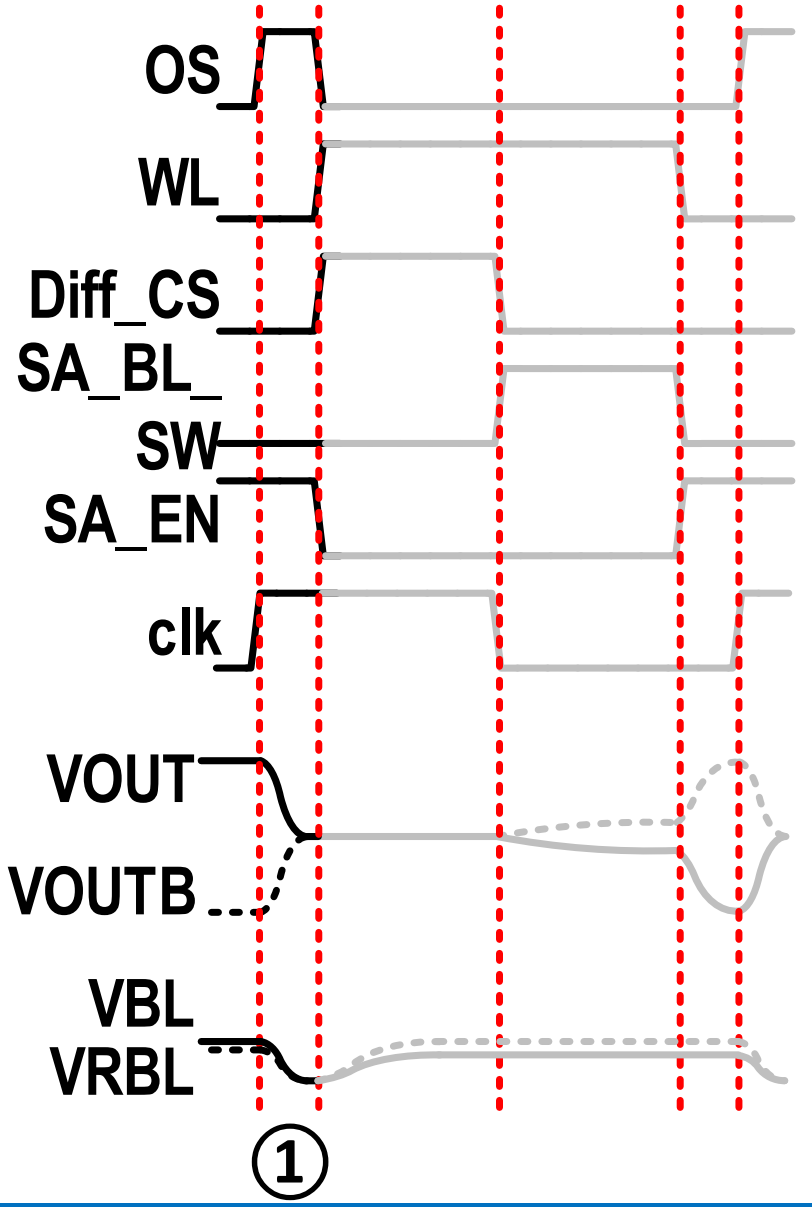
- Cross couple gain doubling
- Dynamic clamping
- Single cap auto-zeroing



DCOCSA Timing

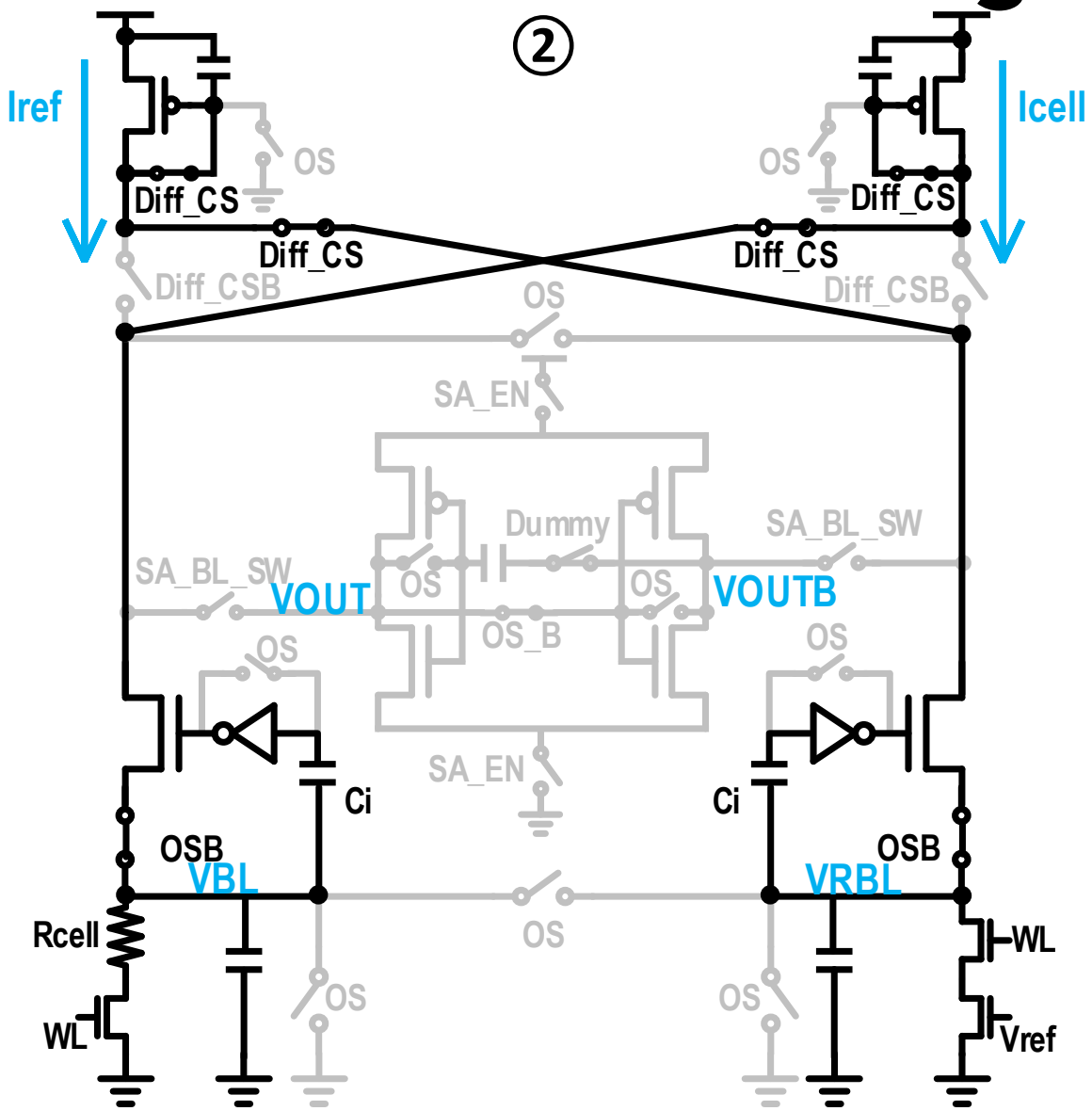
- Phase I
  - Pre-charge
  - Sample offset
  - Sample bias voltage

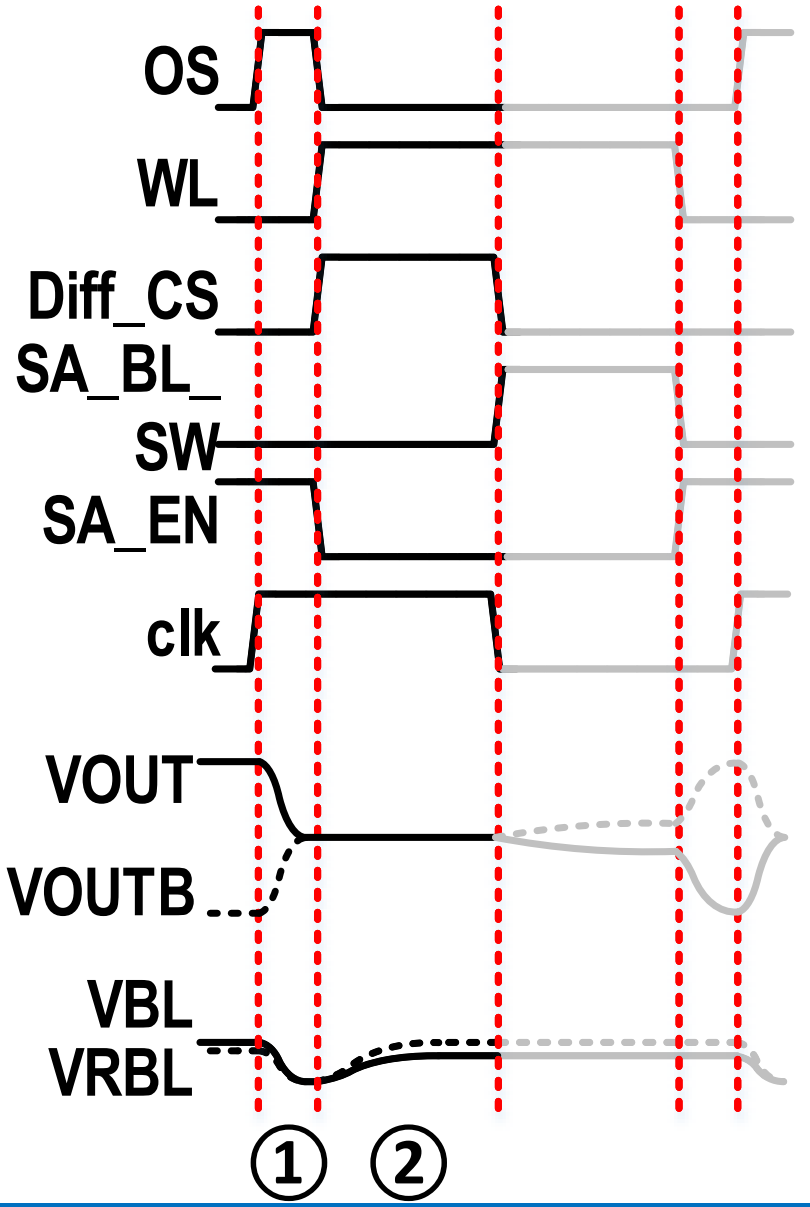




DCOCSA Timing

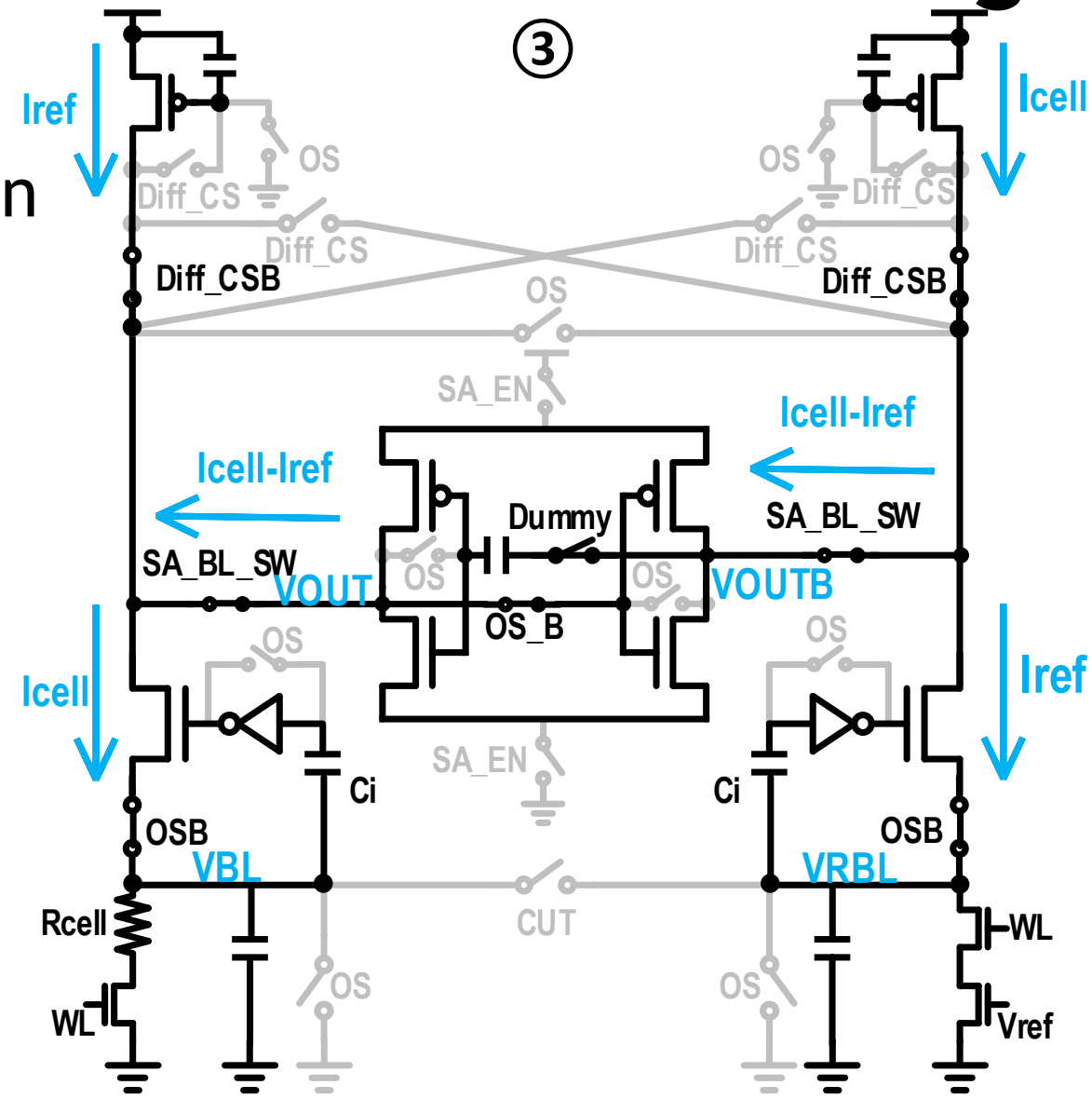
- Phase II
  - Sample current

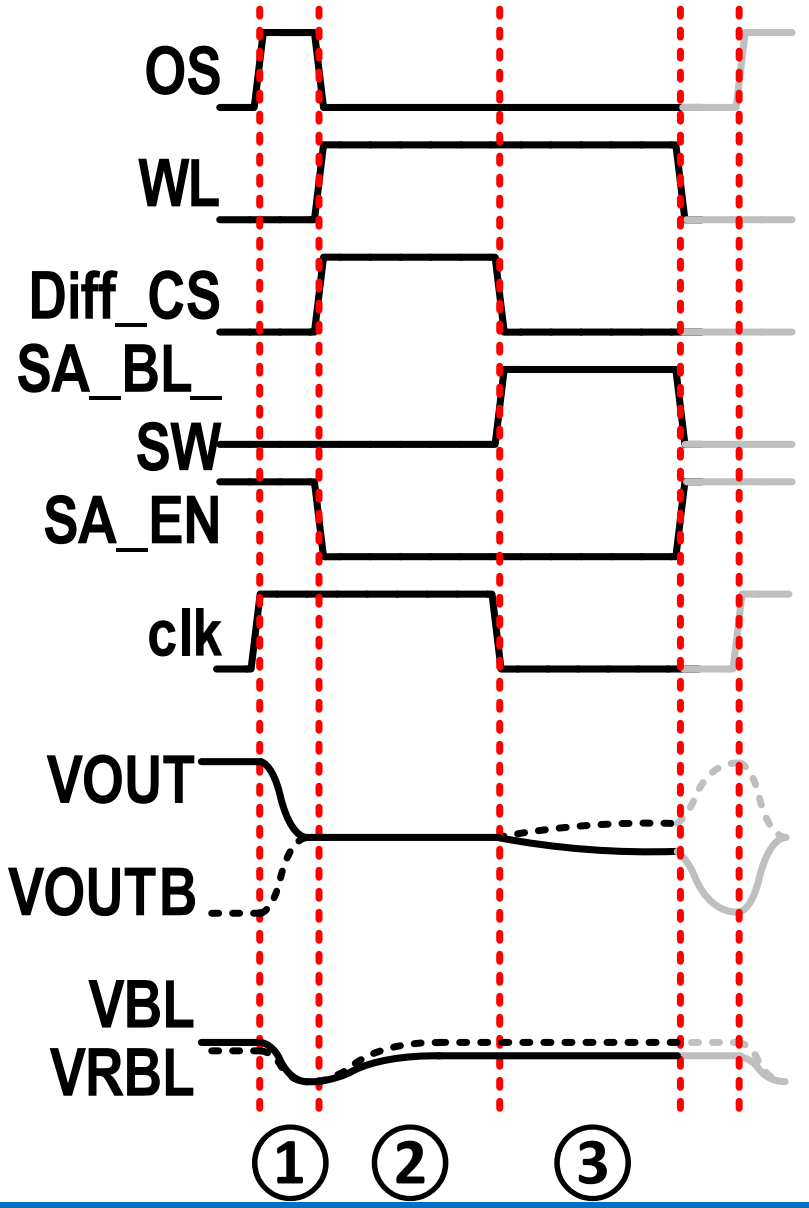




DCOCSA Timing

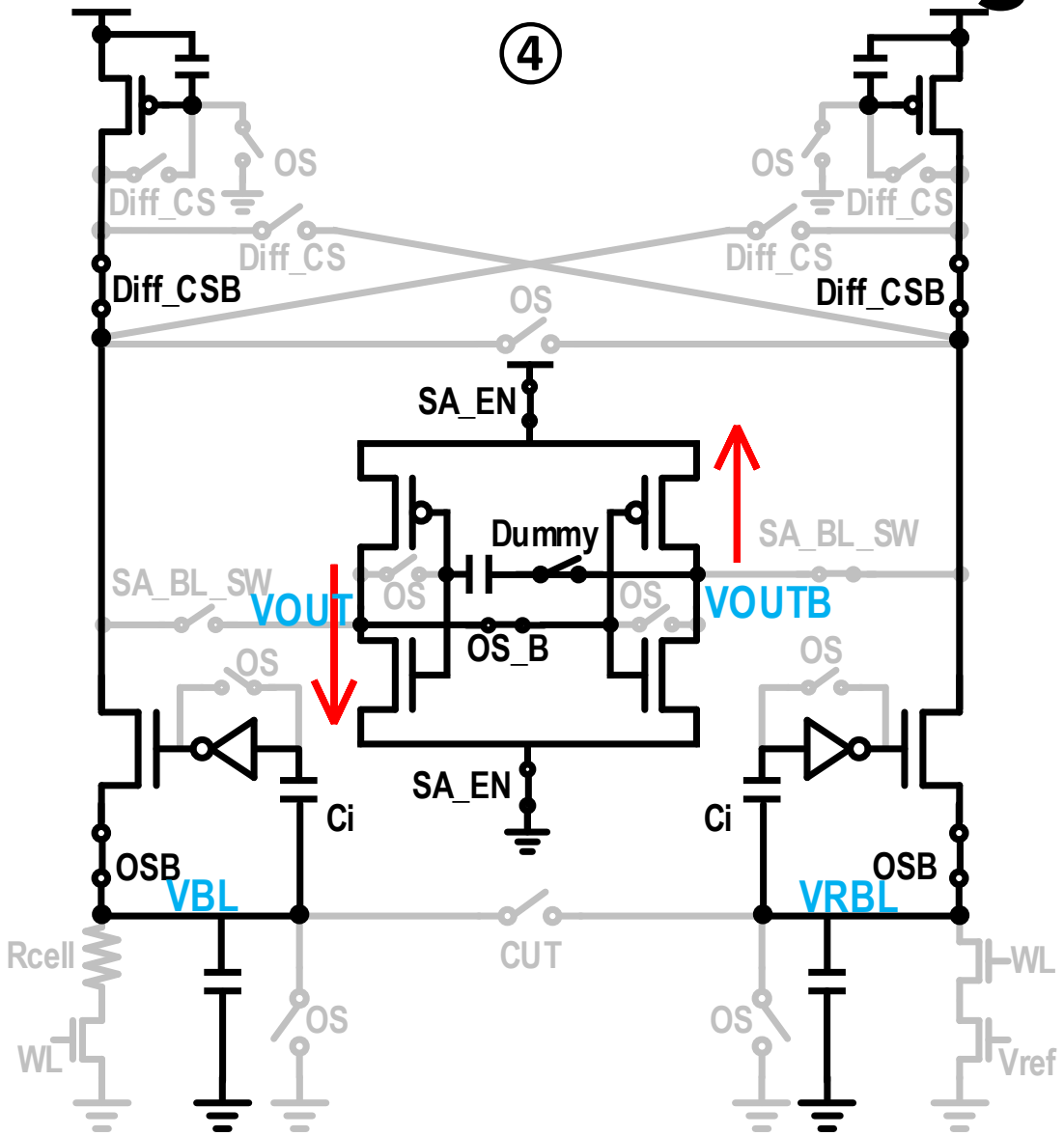
- Phase III
  - Double gain

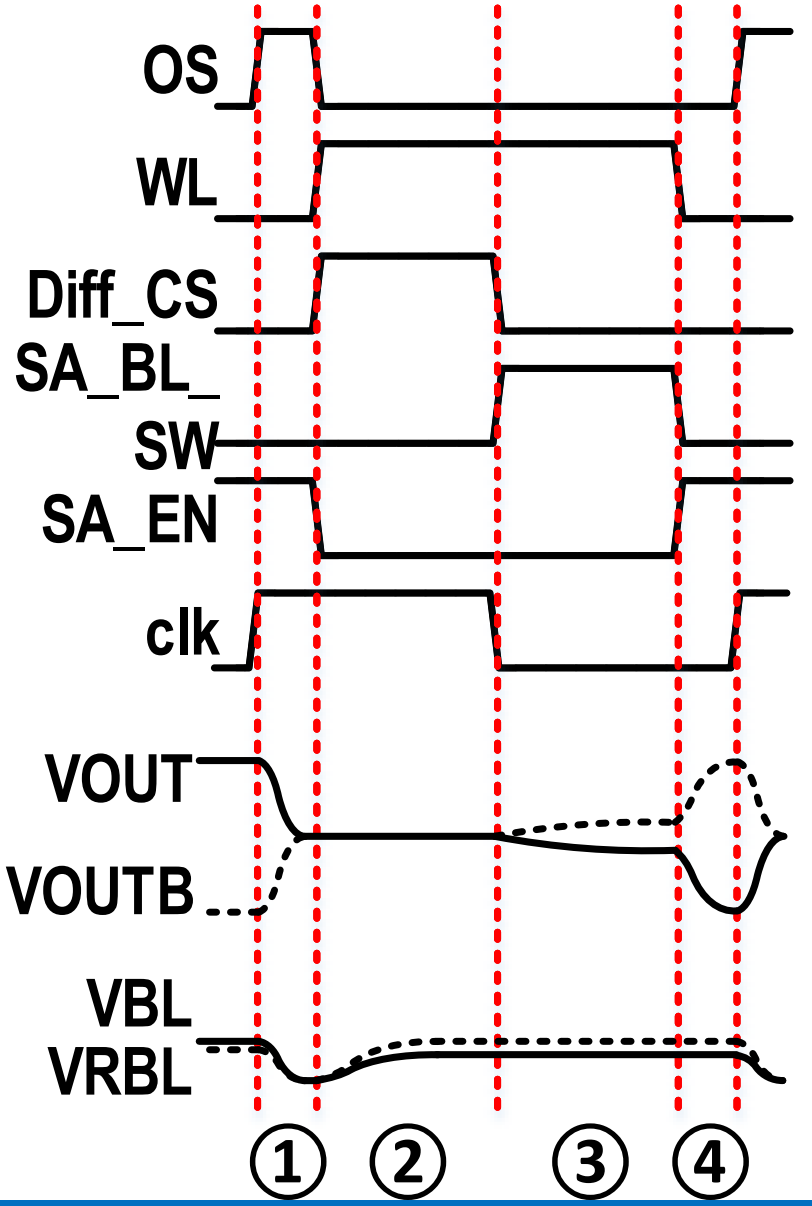




DCOCSA Timing

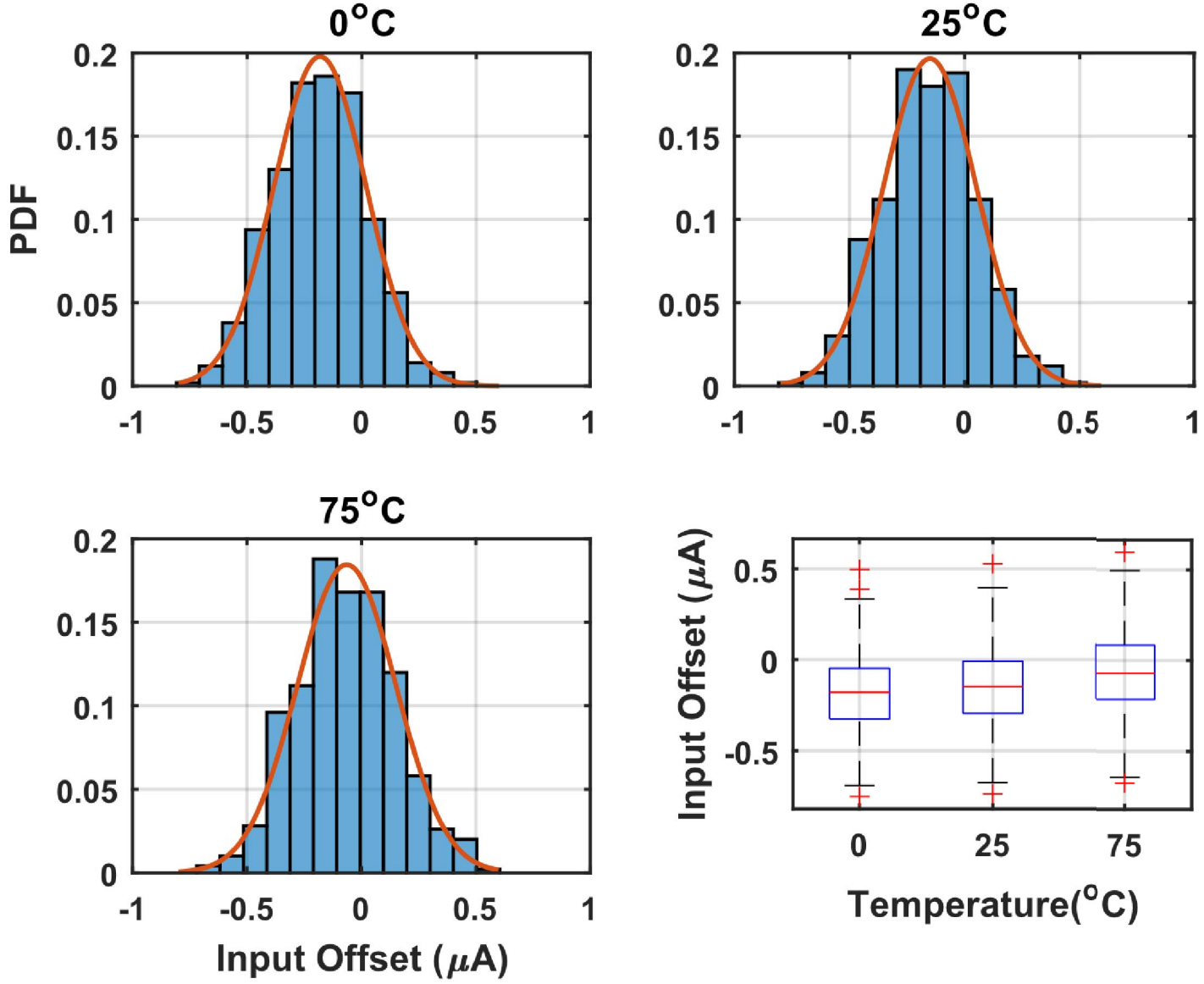
- Phase IV
  - Amplify





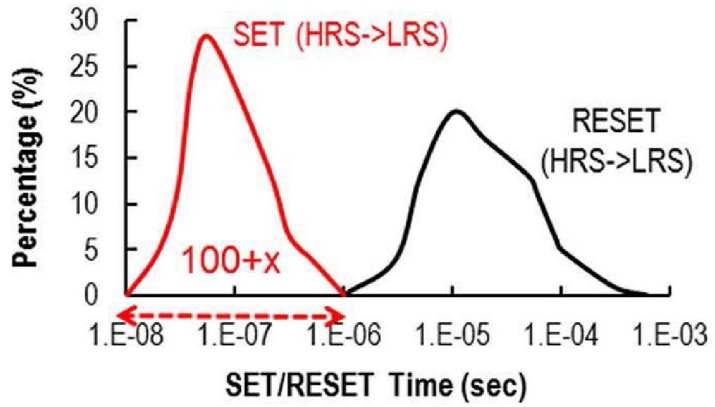
DCOCSA Monte Carlo Simulation

- Sub-μA input offset @21μA common mode

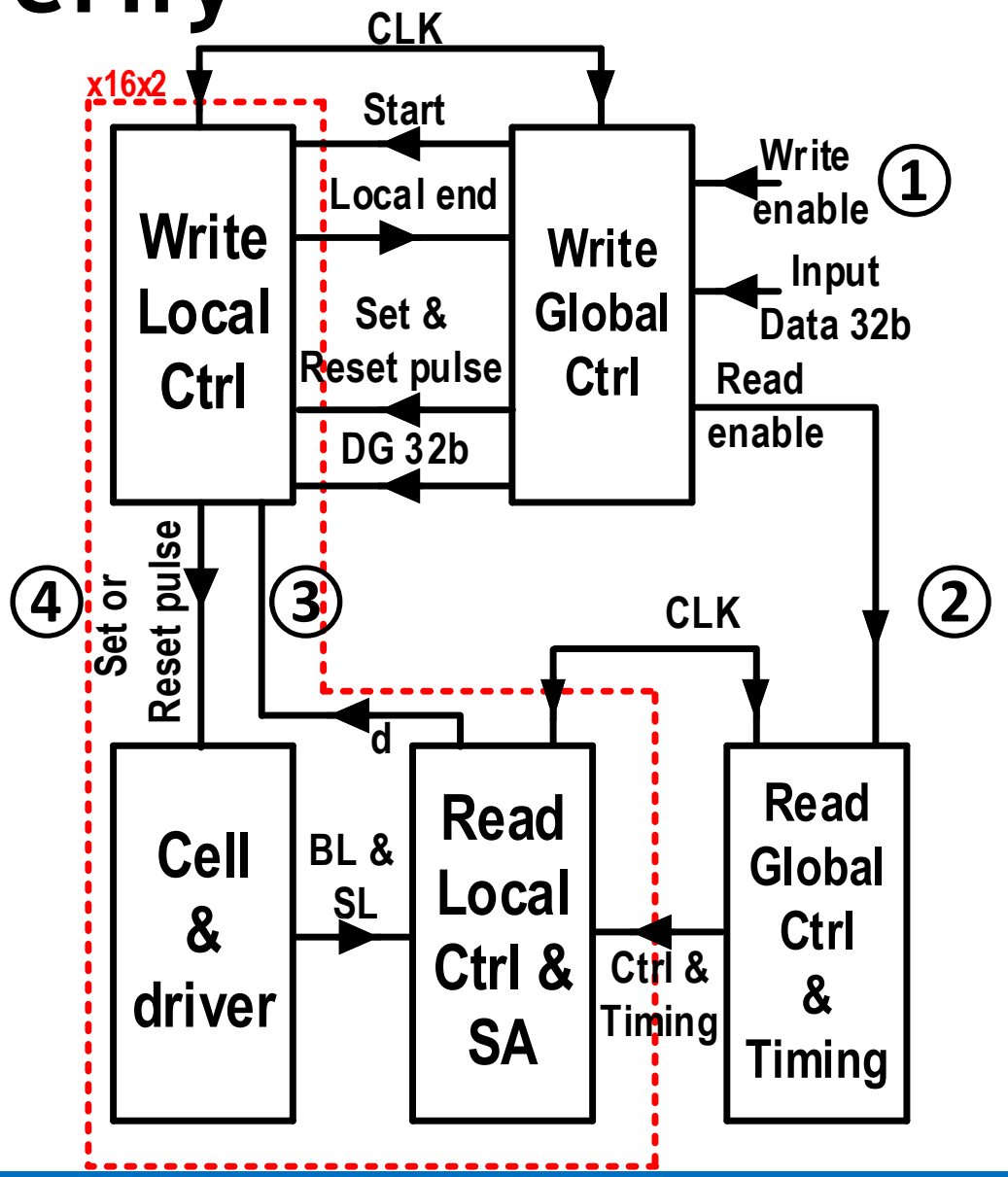


RRAM Write-Verify

- Fine-grained iterative Write-Verify
- Alleviates locality-dependent variation
  - Decouple fast and slow cells
  - Automatically adapts to the corresponding SA offset



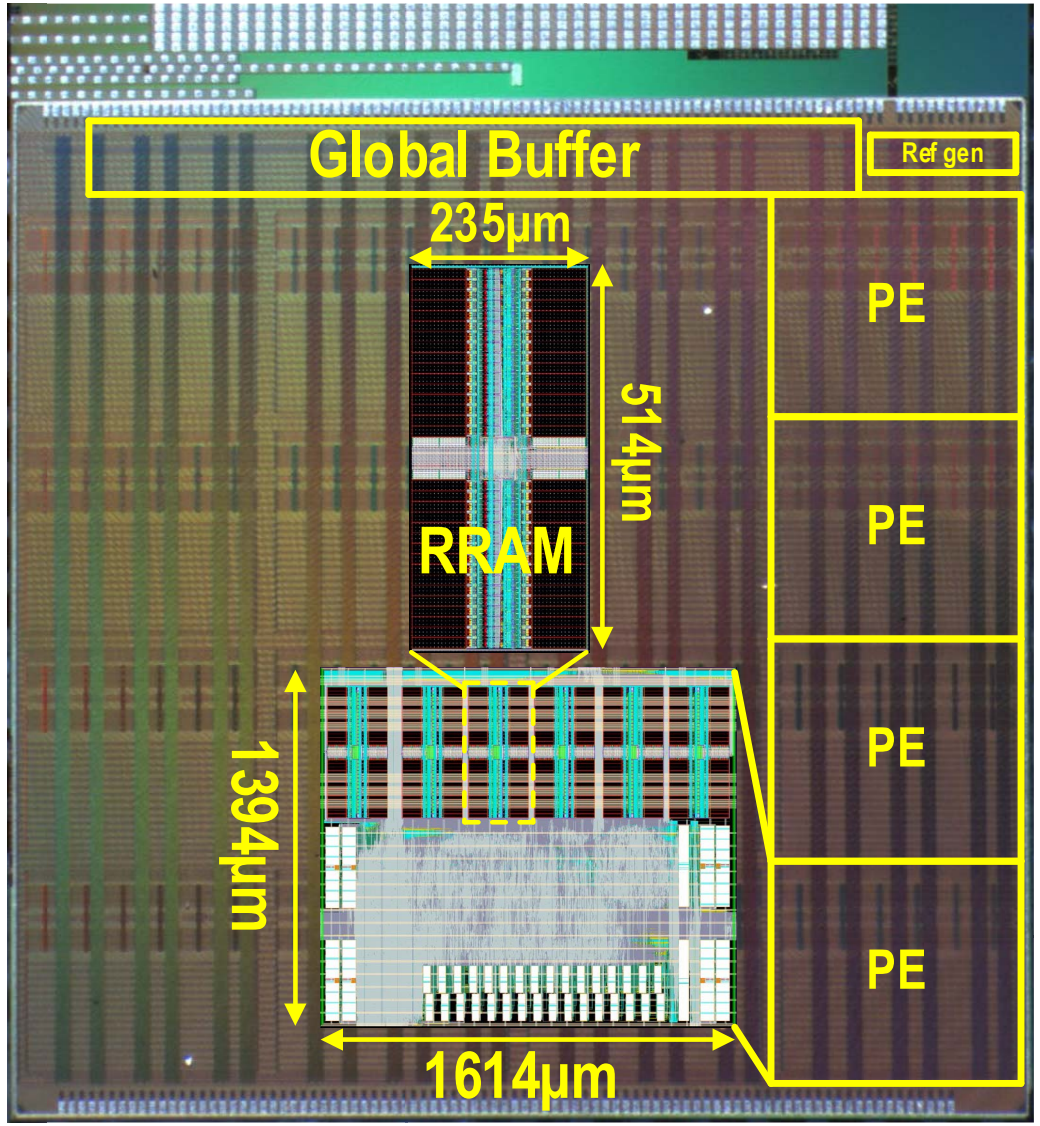
[M. Chang, JSSC]



Outline

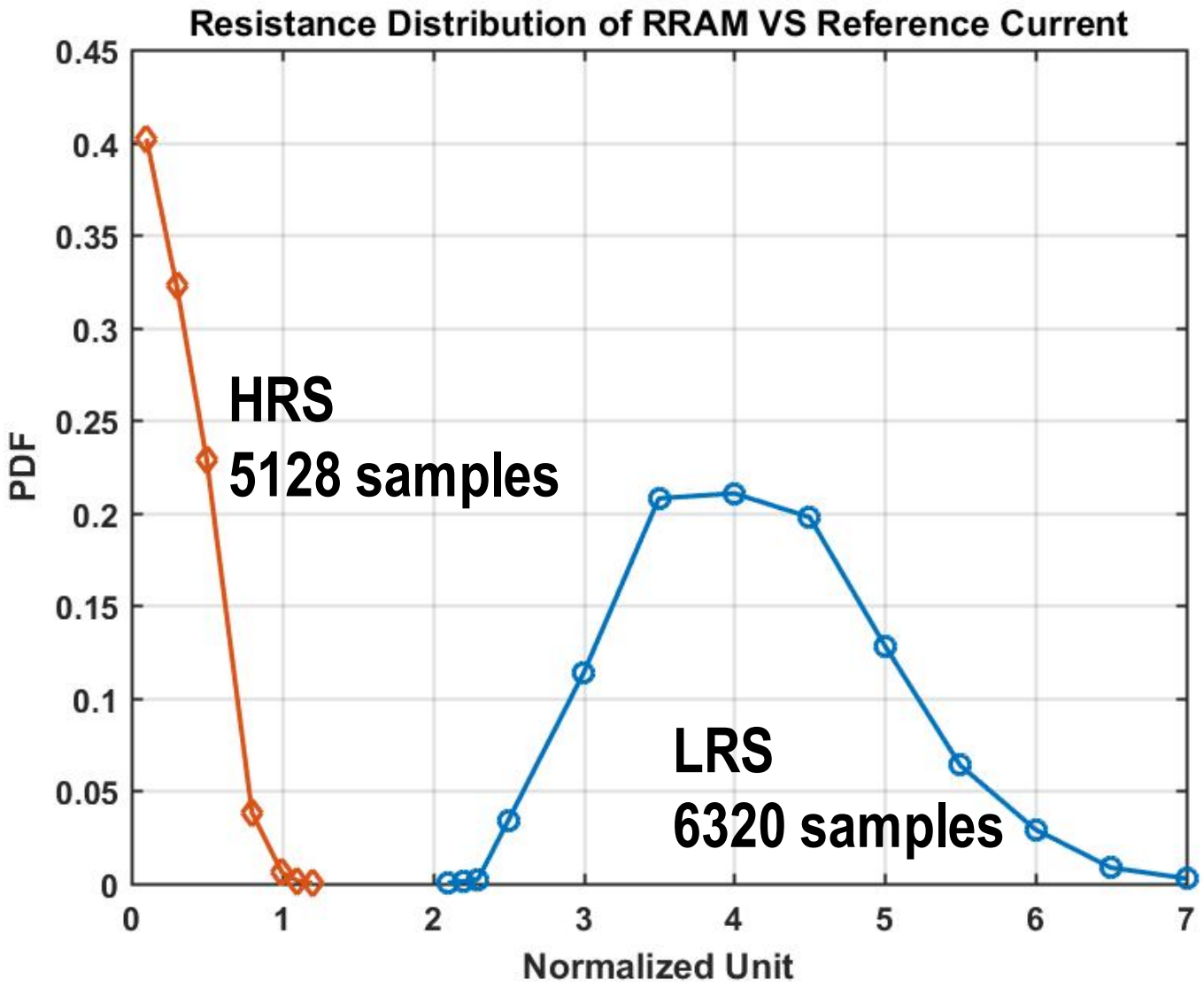
- Motivation
- All-Weights-on-Chip DNN Accelerator
- Test Results

Testing Setup and Die Photo



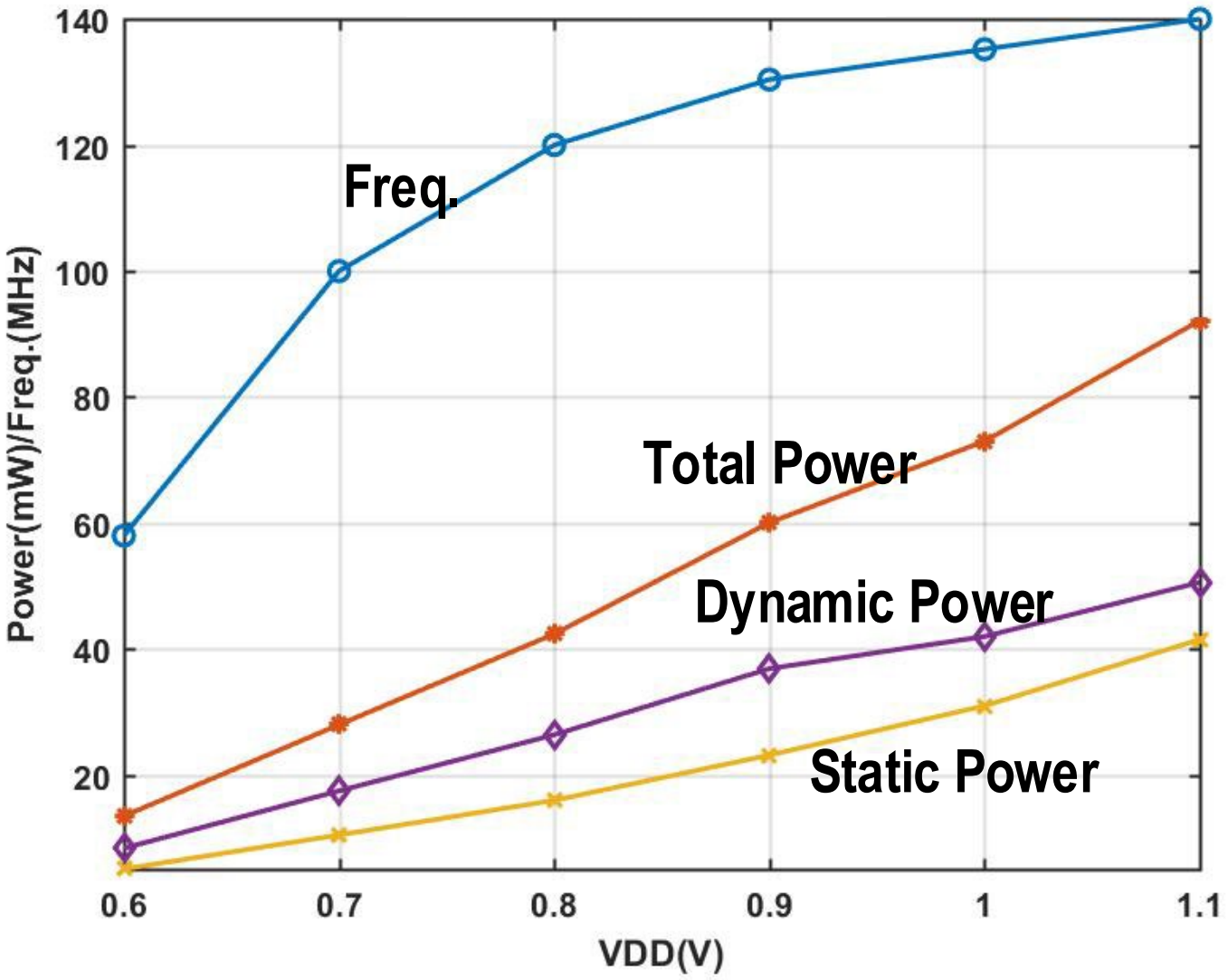
Measured RRAM Resistance Distribution

- Measured with ~10k random samples across 24 banks



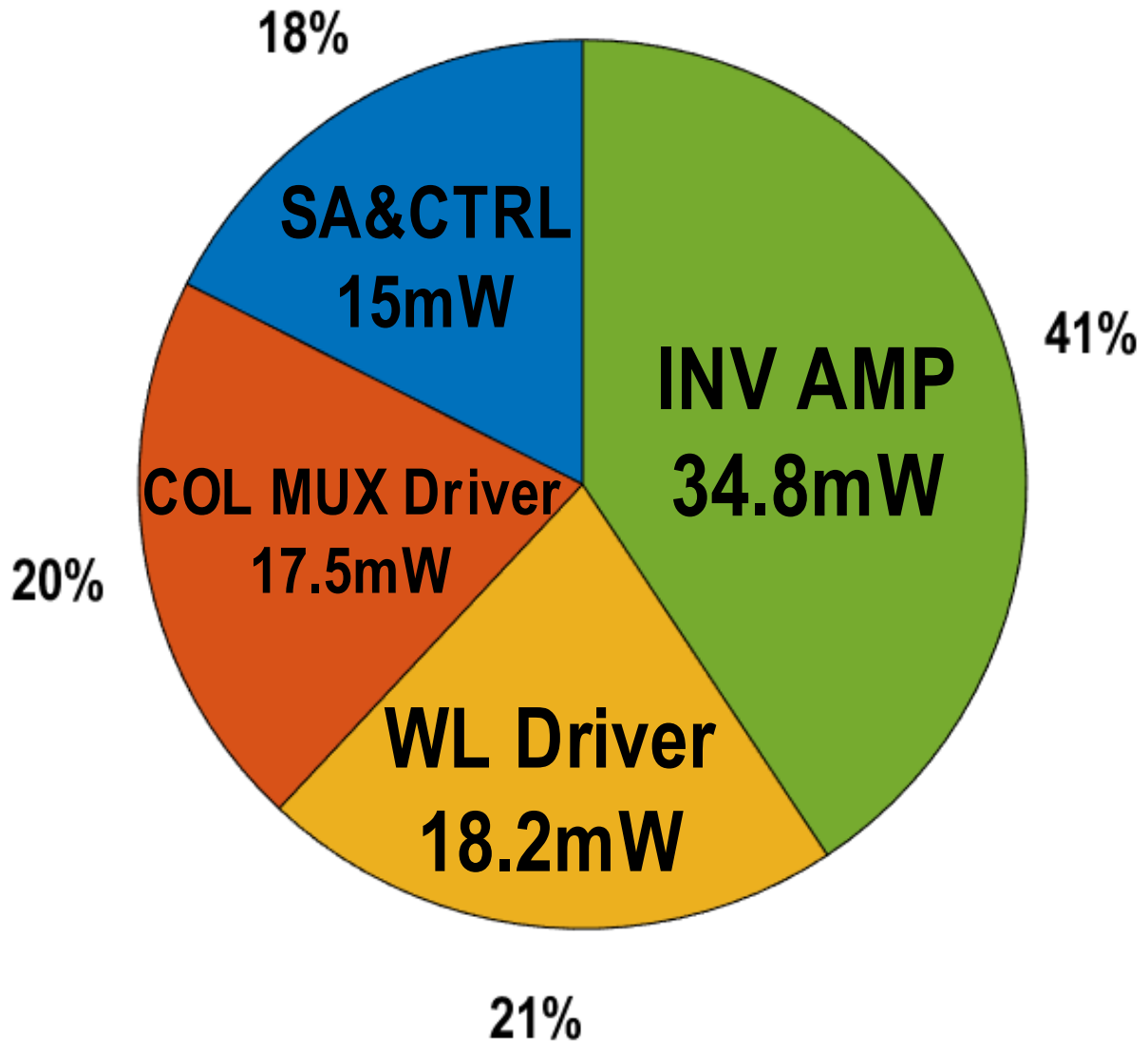
Measured VDD Scaling for Core Digital Logic

- 92.1mW, 140MHz @1.1V
- 42.4mW, 120MHz @0.8V
- 13.5mW, 60MHz @0.6V



Measured Power Breakdown of RRAM

- @60MHz RRAM clock
  - 1V SA and control: 15mW
  - 1.25V column mux: 17.5mW
  - 1.4V WL: 18.2mW
  - 1.1V inv amplifier: 34.8mW



Comparison table

	This Work	QUEST[11]	SNAP[12]	STICKER[13]	UNPU[2]
Technology	**ULL 22nm**	40nm	16nm	65nm	65nm
On-chip RAM(B)	**3M RRAM/1.3M SRAM**	7.68M/96M 3D SRAM	280.6K	170K	256K
Max On-chip Weight	**16M@8b Non-Volatile**	15.36M@ 4b Volatile	140.3K@16b Volatile	170K@8b Volatile	256K@8b Volatile
Off-chip Memory	**No**	Yes	Yes	Yes	Yes
MACs	**4x128 (8x8b)**	24x512 (1x1b log)	252 (16x16b)	256 (8x8b)	4x576 (1x16b)
Voltage (V)	**1.0-1.2 RRAM 0.6-1.1 Core**	1.1	0.55-0.8	0.67-1.0	0.63-1.1
Freq. (MHz)	**60 RRAM/120 Core**	300	33-480	20-200	200
TOPS/W	***0.96@8b**	**0.59@4b	***3.61@16b	***1.038@8b	***5.57@8b
GOPS	**123@8b**	1960@4b	65.52@16b	102@8b	690@8b
Power (mW)	**127.9@120MHz**	3300@300MHz	364@480MHz	284.4@200MHz	297@200MHz
Chip Area (mm^2)	**10.8**	122	2.4	12	16

- *Including power of loading weights from RRAM to SRAM and MAC arrays
- **Including power of loading weights from 3D SRAM to on-chip SRAM and MAC arrays
- ***Excluding power of loading weights from off-chip memory

Conclusion

- First digital DNN accelerator featuring 24 Mb eRRAM as dedicated weight storage to eliminate off-chip weight access
- Weight compression achieving 16 M 8-bit weights on-chip
- Dynamic clamping offset-canceling sense amplifier (DCOCSA) achieving sub-μA input offset

Acknowledgment

- TSMC University Joint Development Program and University Shuttle Program for chip fabrication and valuable advice

- ADA Joint University Microelectronics Program (JUMP) center for support

References

- [1] Y. Chen, et al., ISSCC, pp. 262-264, 2016
- [2] J. Lee, et al., ISSCC, pp. 218-220, 2018.
- [3] P. Whatmough, et al, ISSCC pp. 242-4, 2017
- [4] C. Xue, et al., ISSCC, pp. 388-390, 2019.
- [5] T. Wu, et al., ISSCC, pp.226-228, 2019.
- [6] Z. Li, et al., ISSCC, pp. 134-136, 2019.
- [7] S. Han, et al., ICLR, 2016.
- [8] C-C. Chou, et al., ISSCC, pp. 478-480, 2018.
- [9] P. Jain, et al., ISSCC, pp. 212-214, 2019.
- [10] Q. Dong, et al., ISSCC, pp. 480-482, 2018.
- [11] K. Ueyoshi, et al., ISSCC, pp.216-218,2018
- [12] J. Zhang, et al., VLSI, pp. 306-307, 2019.
- [13] Z. Yuan, et al., VLSI, pp. 33-34, 2018.

Thank you