A 14-nm Energy-Efficient and Reconfigurable Analog Current-Domain In-Memory Compute SRAM Accelerator

Aya G. Amer[1], Maitreyi Ashok[1], Xin Zhang[2,3], John Cohn[3],
Anantha P. Chandrakasan[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. [2] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. [3] MIT-IBM Watson AI Lab, Cambridge, MA, USA.

Outline

- Motivation and background
- Conventional 8T IMC SRAM Challenges
- Proposed IMC 12T SRAM Design
- Proposed Analog IMC SRAM Macro Design
- Proposed Fully-Analog Classifier Architecture.
- Measurement Results
- Conclusions

Motivation: AI Applications & Edge Computing

Security & Surveillance

SmartTV

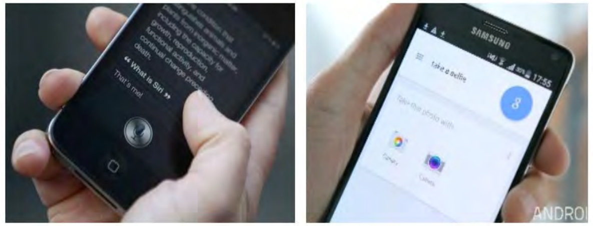
Autonomous Cars

Artificial Intelligence at home
Open Source Code Robot & Linux OS
Qbo

Home Robotics

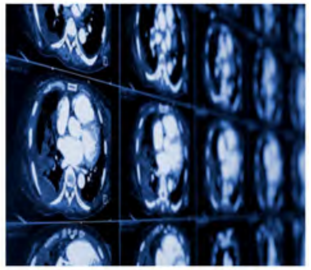
VR& Gaming

Face recognition



Speech Recognition

Wearable Medical Devices

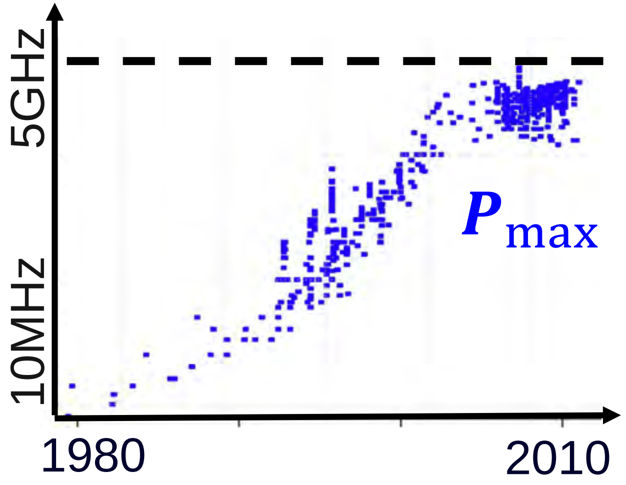


Medical Imaging

Huge amounts of data processing in real-time.

Motivation: Computing Limitations

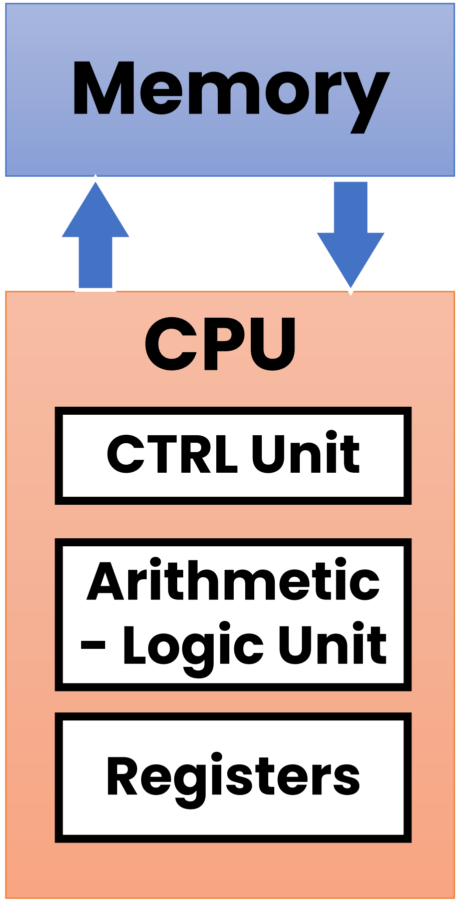
Power-Wall



Limited Power Budget

↓

Enhanced Transistors


Memory

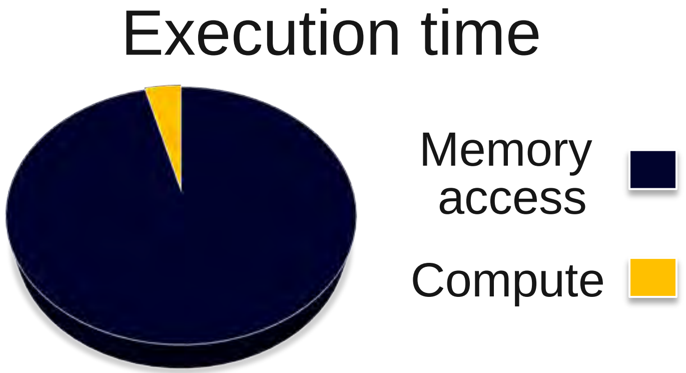
CPU

CTRL Unit

Arithmetic – Logic Unit

Registers

Memory-Wall



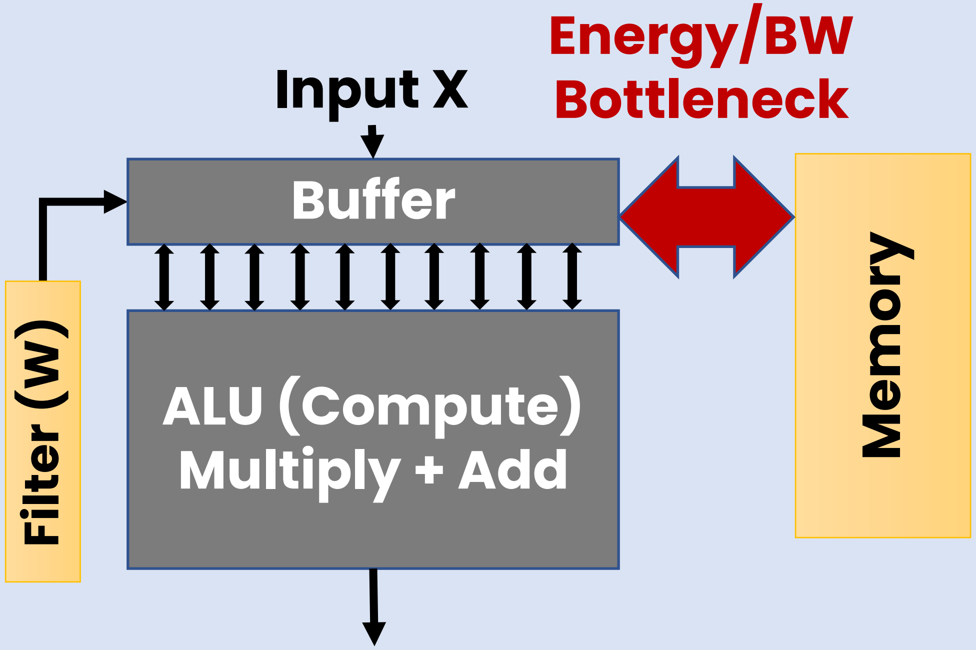
Limited Bandwidth

↓

Enhanced Architectures

Motivation: In-Memory Compute (IMC) Accelerator Architecture

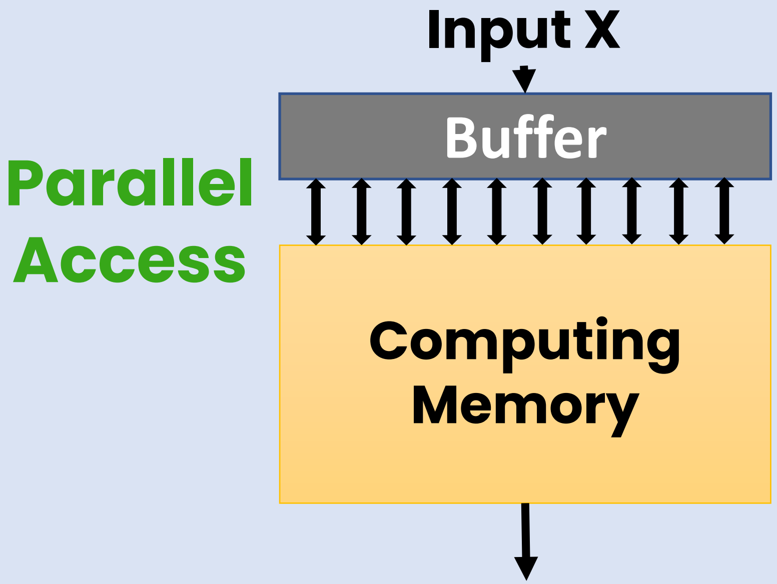
Von-Neumann Accelerator



Output Y = $\sum X_i * W_i$

Memory Wall
→ Energy, Delay Overhead

In-Memory Compute Accelerator



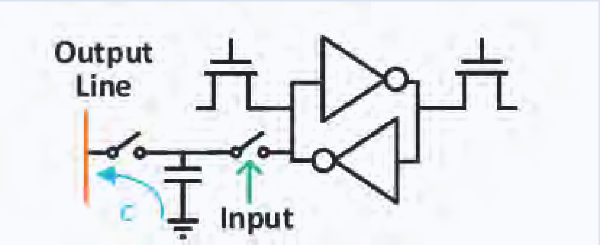
Output Y = $\sum X_i * W_i$

Parallel Processing
Reduce Data Movement
→ Lower Energy, Higher Speed

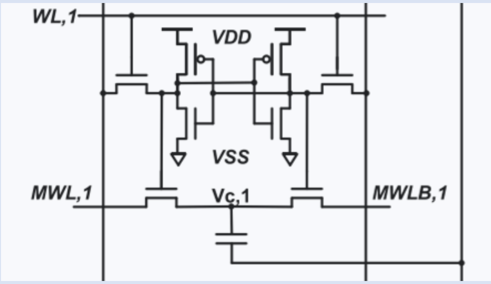
Compute in SRAM Cells

- **Charge Domain Computing**
  - ➢ Compute using capacitive coupling and charge sharing.

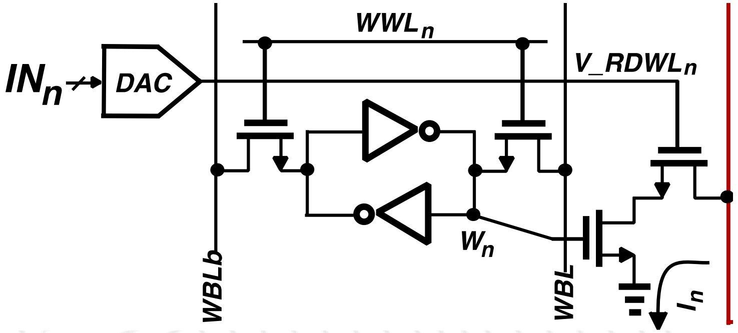
Bit-wise AND + Accumulate



Bit-wise X-NOR + Accumulate



- **Current Domain Computing**
  - ➢ Compute using discharging currents accumulation.

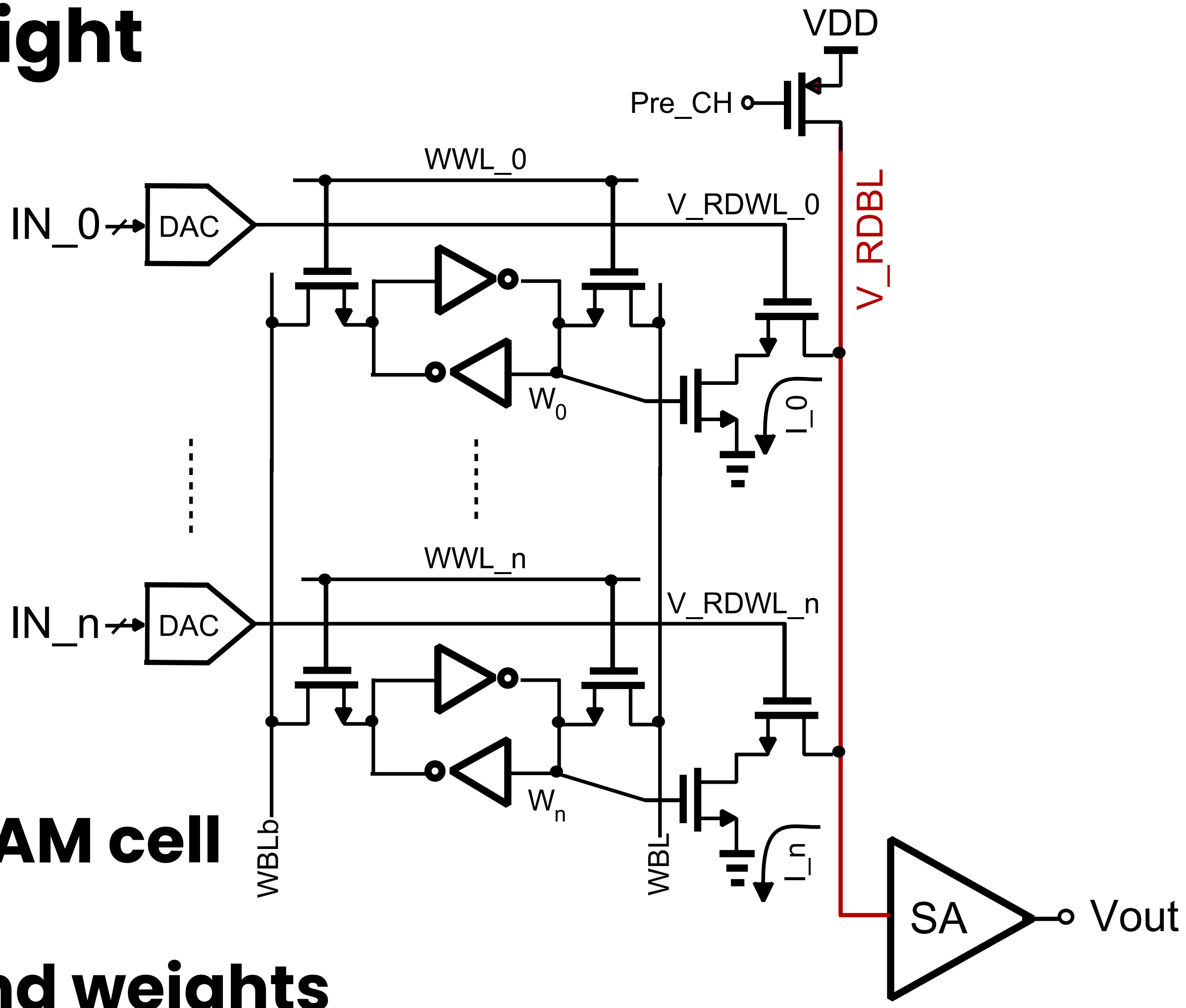


Conventional Current-Domain IMC 8T SRAM

- **Multiply analog IN by 1b-weight**
  - ➢ IN is applied to RDWL
  - ➢ W turns on/off SRAM read port.
  - ➢ $I_{out} = \sum I_n$ → MAC output
  - ➢ I_{out} discharges RDBL.
- **RDBL Voltage Sensing**

$$V_{out} \propto \sum_{i=0}^{n} IN_i * W_i$$

- **Energy-Efficient MAC**
- **Compatible with standard 8T SRAM cell**
- **Supports analog/ n-bit inputs and weights**



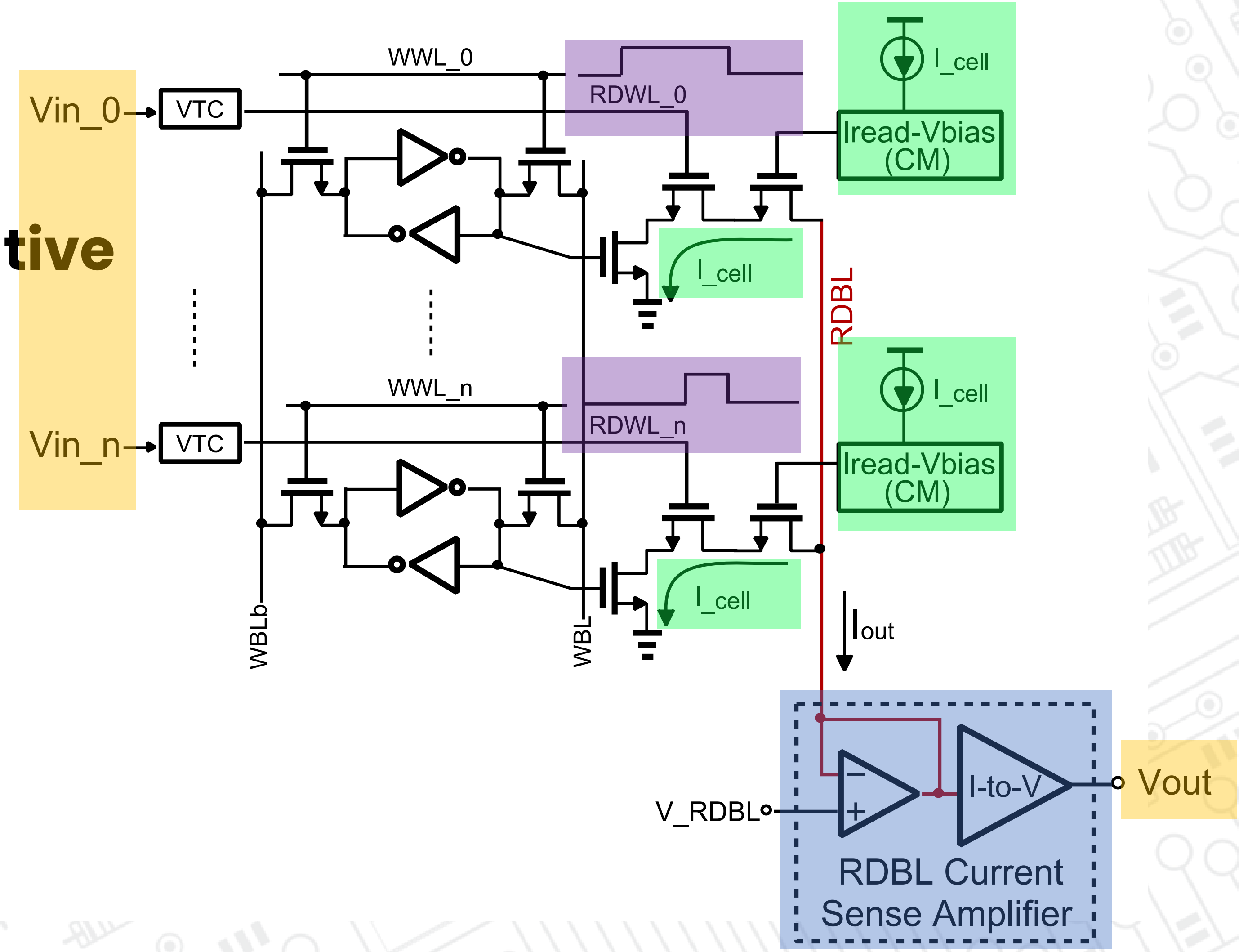
RDBL Voltage Sense Amplifier

Major IMC SRAM Challenges

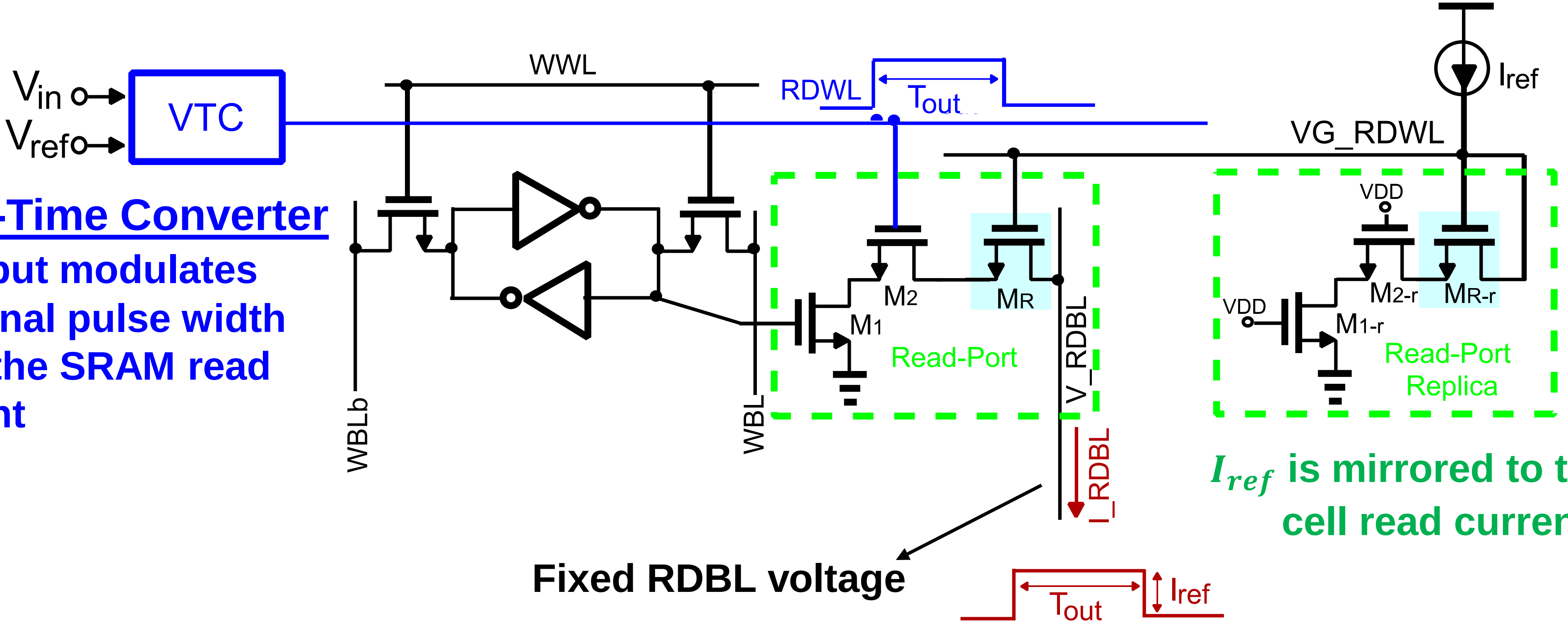
- **Limited Signal Margin at low VDD**
  - ➢ Limits IMC Parallelism.
- **Non-linearity with RDBL discharge**
  - ➢ Limits MAC accuracy.
- **Non-linearity with input/ output codes**
  - ➢ Limits IMC Parallelism and speed.
  - ➢ Limits MAC accuracy.
- **Process variations**
  - ➢ Limits MAC accuracy.

Proposed Solution

- **Current-Controlled SRAM Read**
  - Low-power subthreshold read
  - Higher IMC Parallelism
- **RDBL Current Sensing with Negative Feedback**
  - Improves MAC linearity
  - Lower mismatch and variations
- **Time-Domain MAC**
  - PWM digital RDWL signal
  - Improves linearity with input codes
- **Analog Inputs and Outputs**
  - Supports cascaded layers.
  - Higher MAC accuracy



Proposed IMC 12T SRAM Cell



Voltage-to-Time Converter

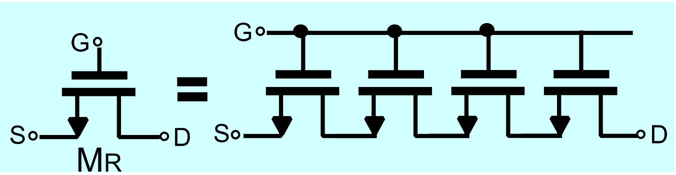
- Analog Input modulates RDWL signal pulse width
- Activates the SRAM read port current

Fixed RDBL voltage

I_{ref} is mirrored to the bit cell read current.

I_{out} pulse width increases with Vin

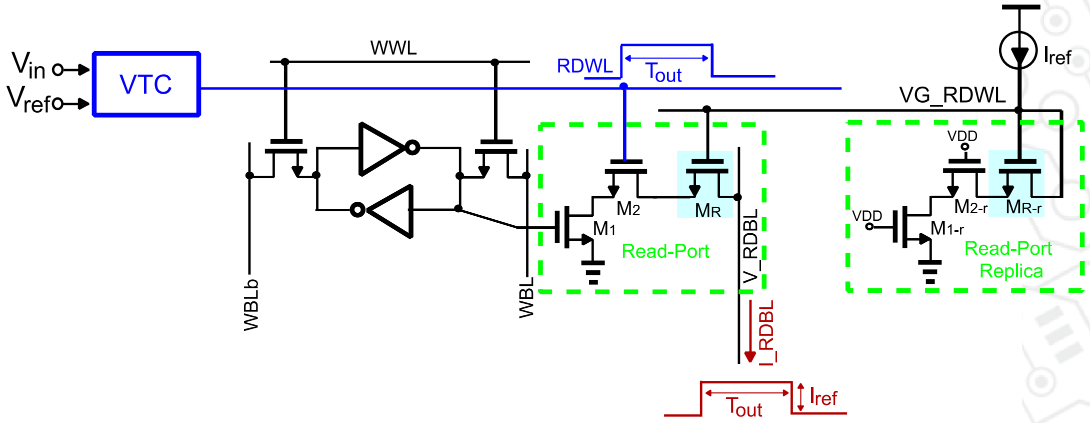
T_{out} α Vin*W

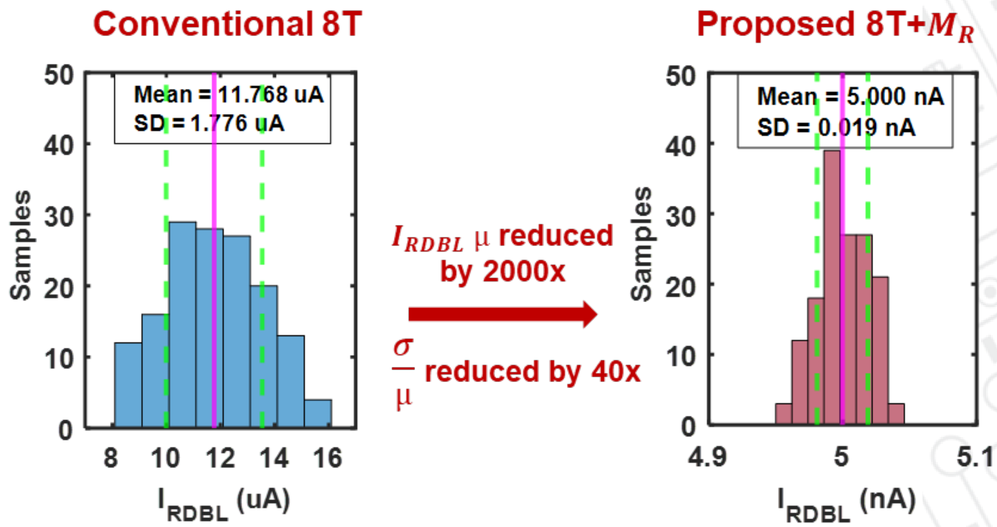


Effective bigger length MR for lower mismatches

Proposed IMC 12T SRAM Cell

- **Current-Controlled Read**
  - Current mirrored from a reference
  - Low-power mode vs high-speed mode
- **Low current mismatch & variation**
  - V_RDBL is regulated.
  - VG_RDWL is generated from current mirror.
  - M_R longer length and bigger area
- **Read current is independent of input/ output codes**
  - T_{out} changes linearly with MAC output
  - VG_RDWL is constant



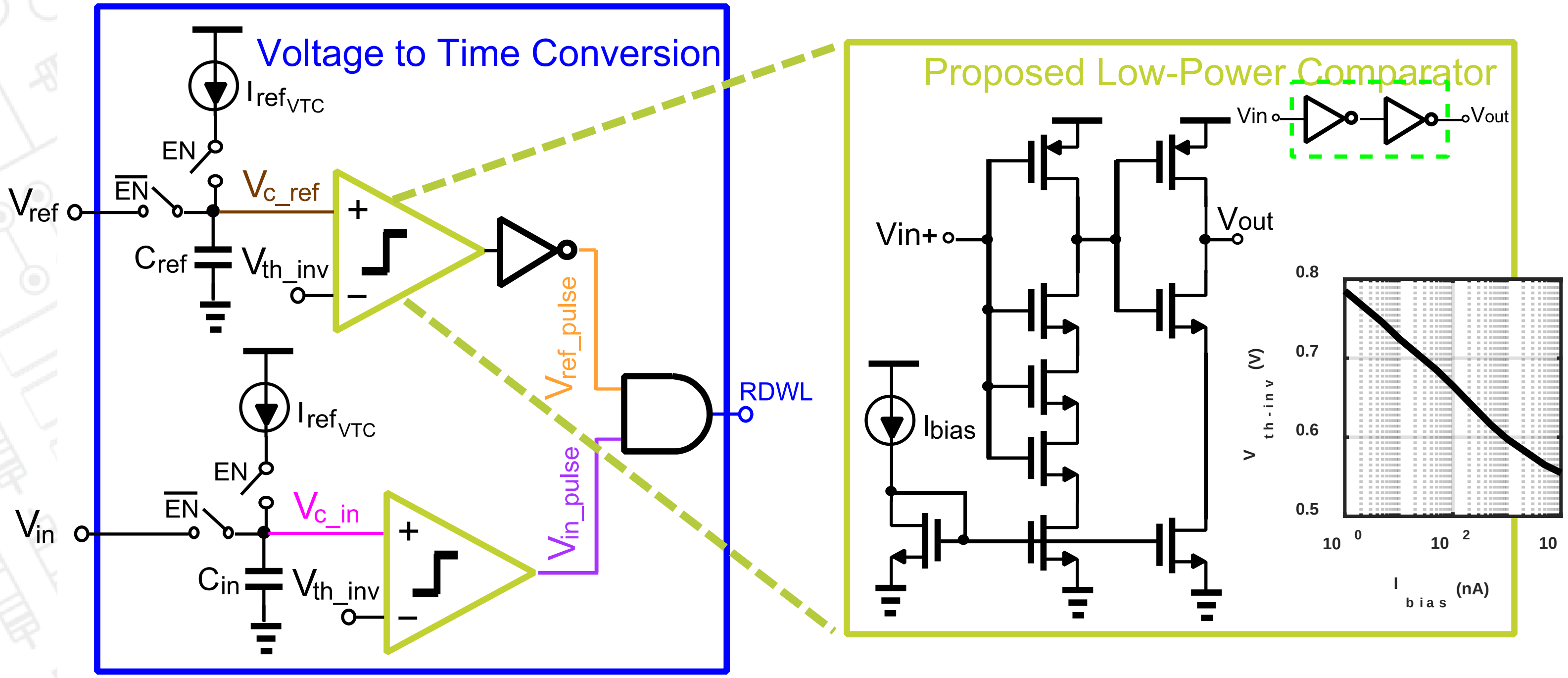

Conventional 8T

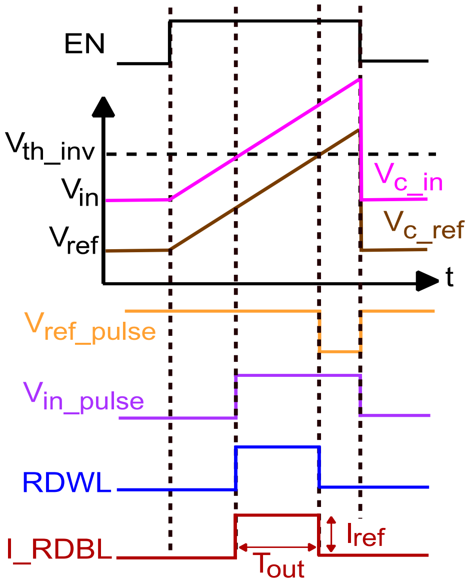
Proposed 8T+M_R

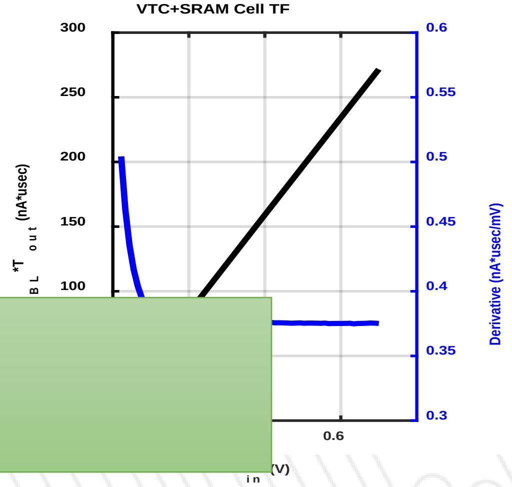
I_{RDBL} μ reduced by 2000x

$\frac{\sigma}{\mu}$ reduced by 40x

Voltage-to-Time Conversion (VTC)







- **VTC generates RDWL pulses**
  - **Analog Input modulates RDWL signal pulse w**

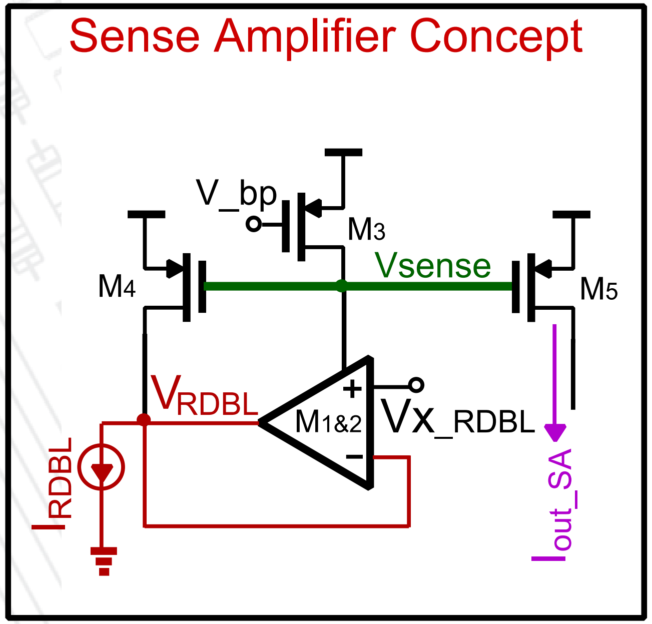
$$T_{pulse} = \frac{C_{in}}{I_{ref_{VTC}}} *$$

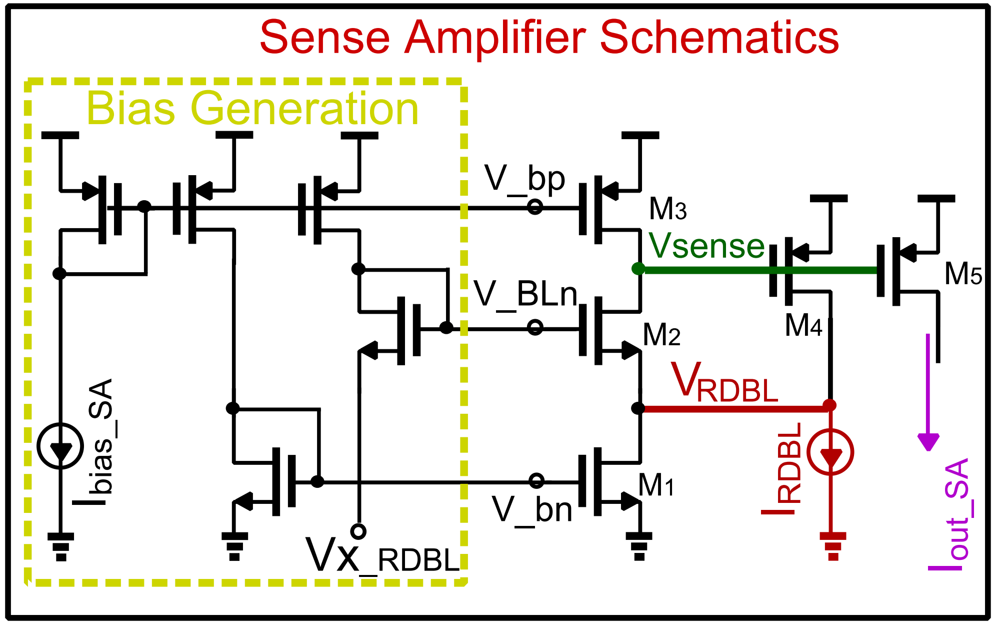
- **Low-Power Comparator**
  - **Current-controlled threshold**

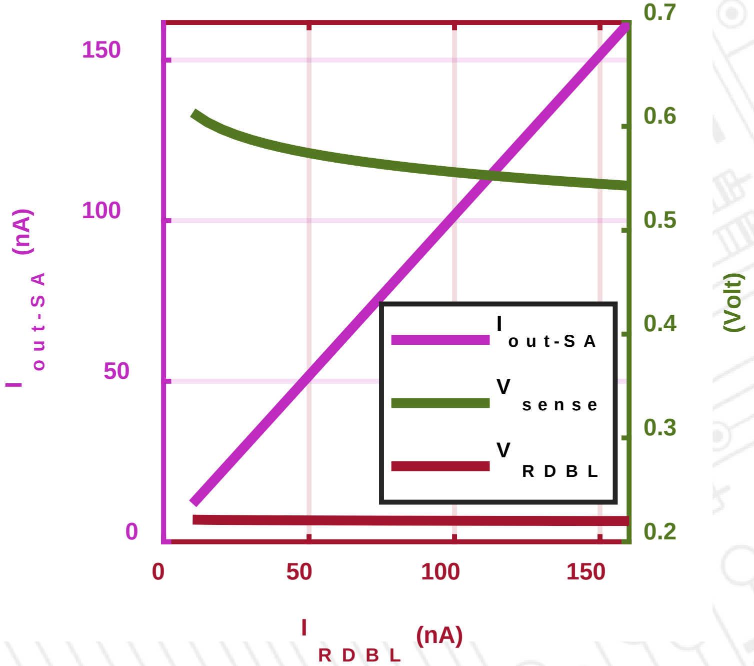
Linear Low-Power Operation
Robust against Process Variations

Current-Sense Amplifier (CSA)

- **Senses the SRAM output MAC current.**
- **Regulates the RDBL voltage.**
  - ➢ V_RDBL is constant through a negative feedback loop.
  - ➢ Improves MAC linearity



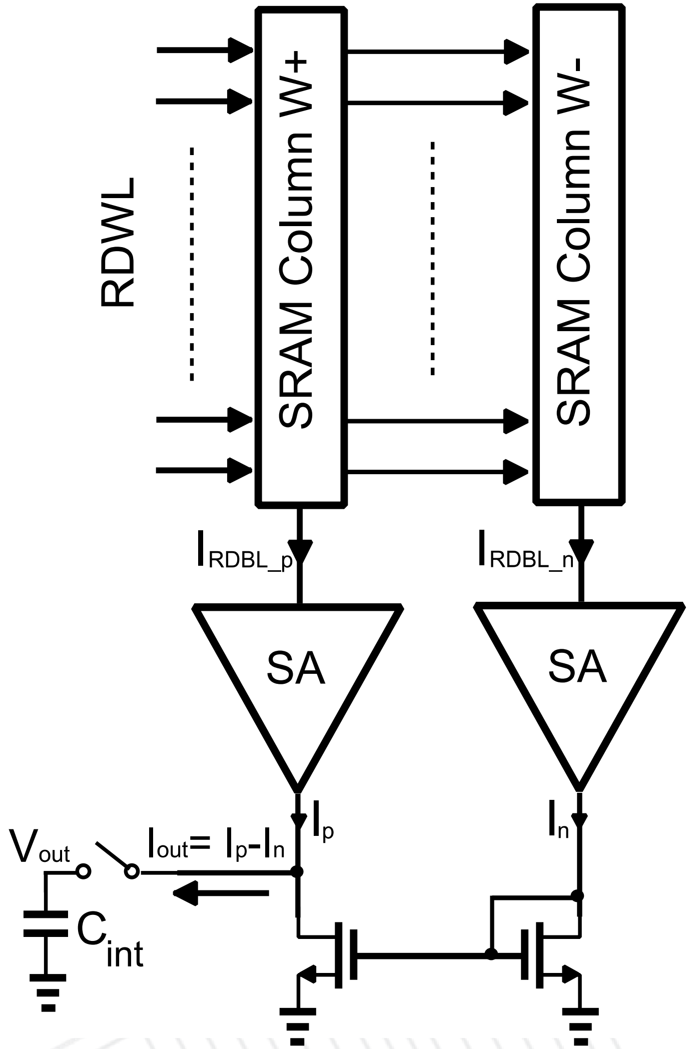




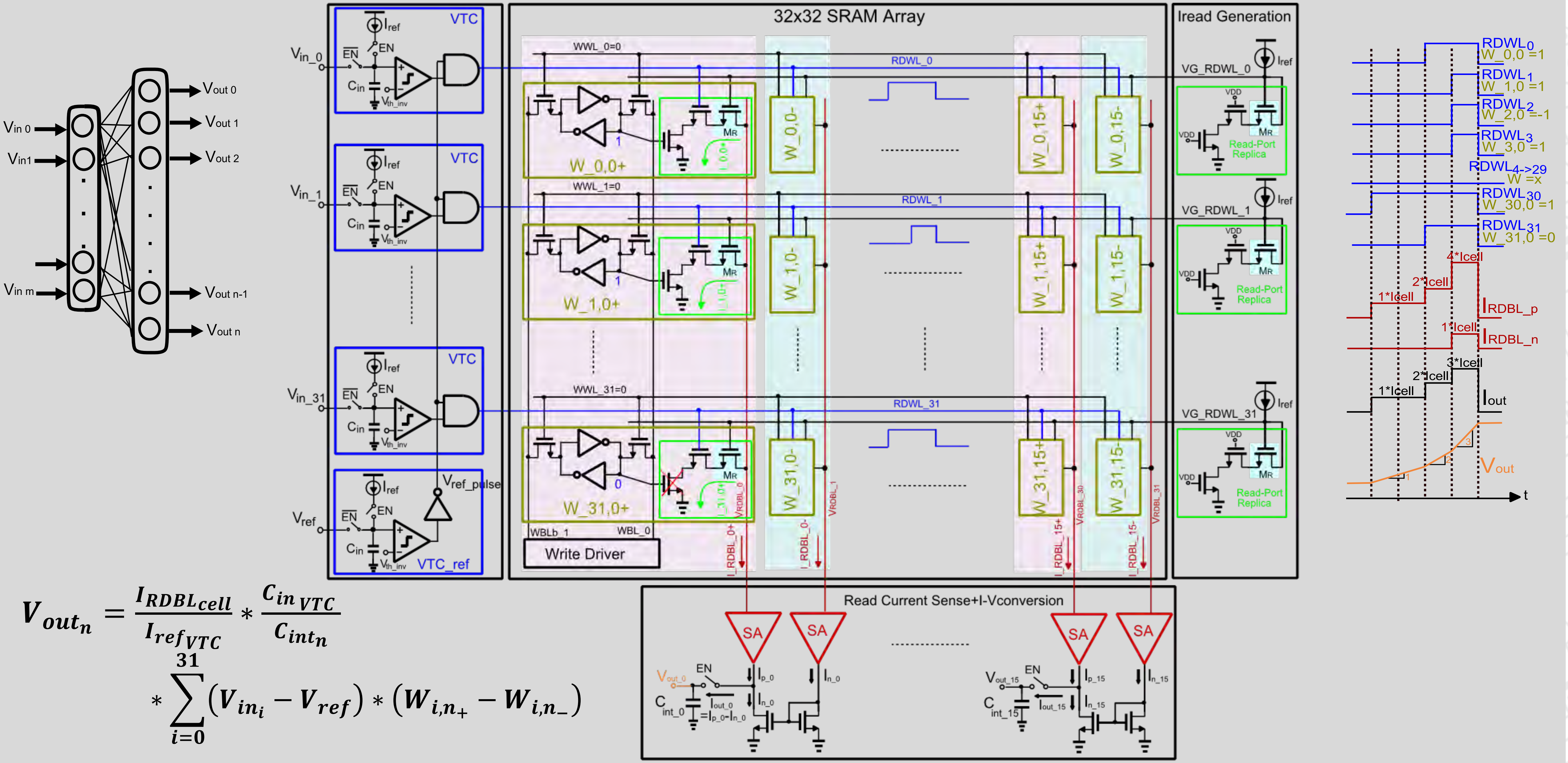
Differential Current-Sense Amplifier

- **Differential Current Sensing**
  - Senses and converts differential RDBL current to voltage on C_{int}.
  - Cancels mismatches.
  - Overcomes SRAM leakage currents.
  - Extends output signal margin.
  - Ternary weight representation {1,0,-1}

$$V_{out} \propto \sum_{i=0}^{31} (V_{in_i} - V_{ref}) * (W_{i,n_+} - W_{i,n_-})$$



Proposed Analog IMC SRAM Macro



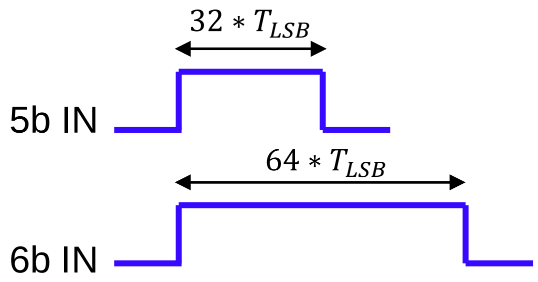
$$V_{out_n} = \frac{I_{RDBLcell}}{I_{ref_{VTC}}} * \frac{C_{in\,VTC}}{C_{int_n}} * \sum_{i=0}^{31} \left(V_{in_i} - V_{ref}\right) * \left(W_{i,n_+} - W_{i,n_-}\right)$$

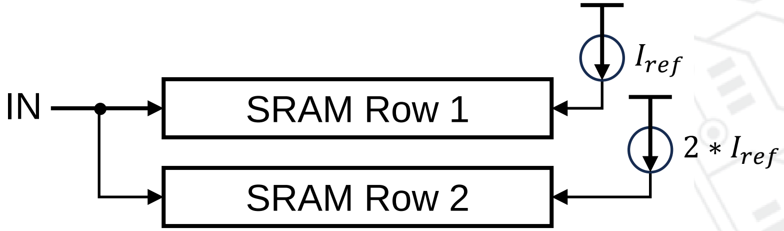
Reconfigurable IMC SRAM Operation

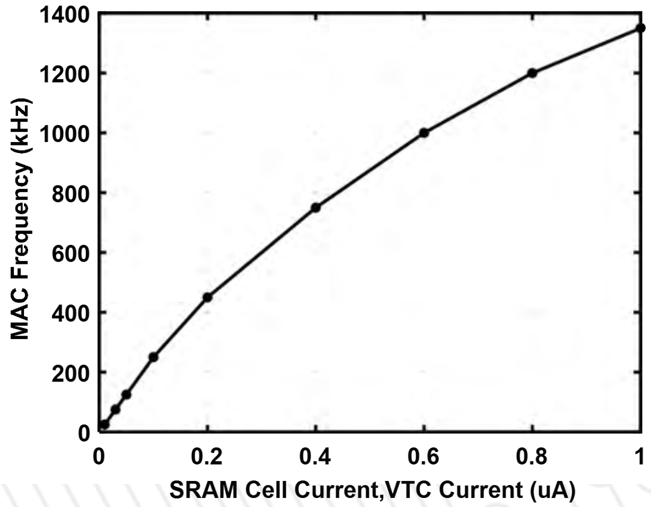
- **Supports multi-bit input precision**
  - Adjust $I_{RDBL_{cell}}$, $I_{ref_{VTC}}$ to change integration time.
  - Higher input precision requires slower operation.

- **Supports multi-bit weight precision**
  - Connect multiple SRAM rows to one IN for n-bit W.
  - Adjust SRAM rows’ currents for binary W representation.

- **Current Controlled MAC**
  - Control MAC Speed.
  - Adjust the outputs scaling/ dynamic range.



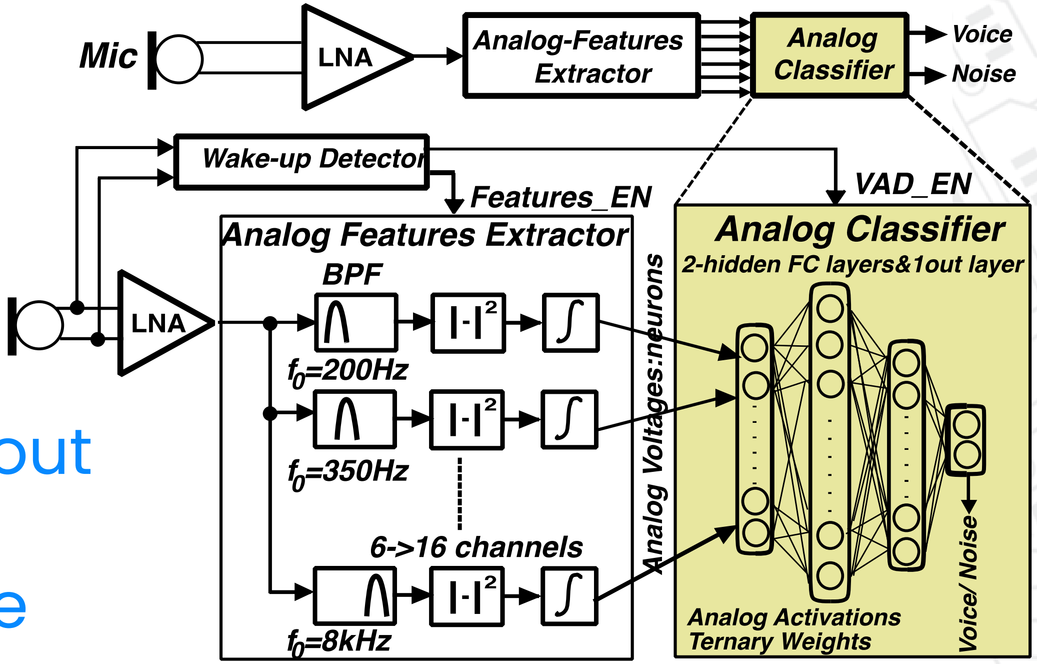




Fully Analog Systems

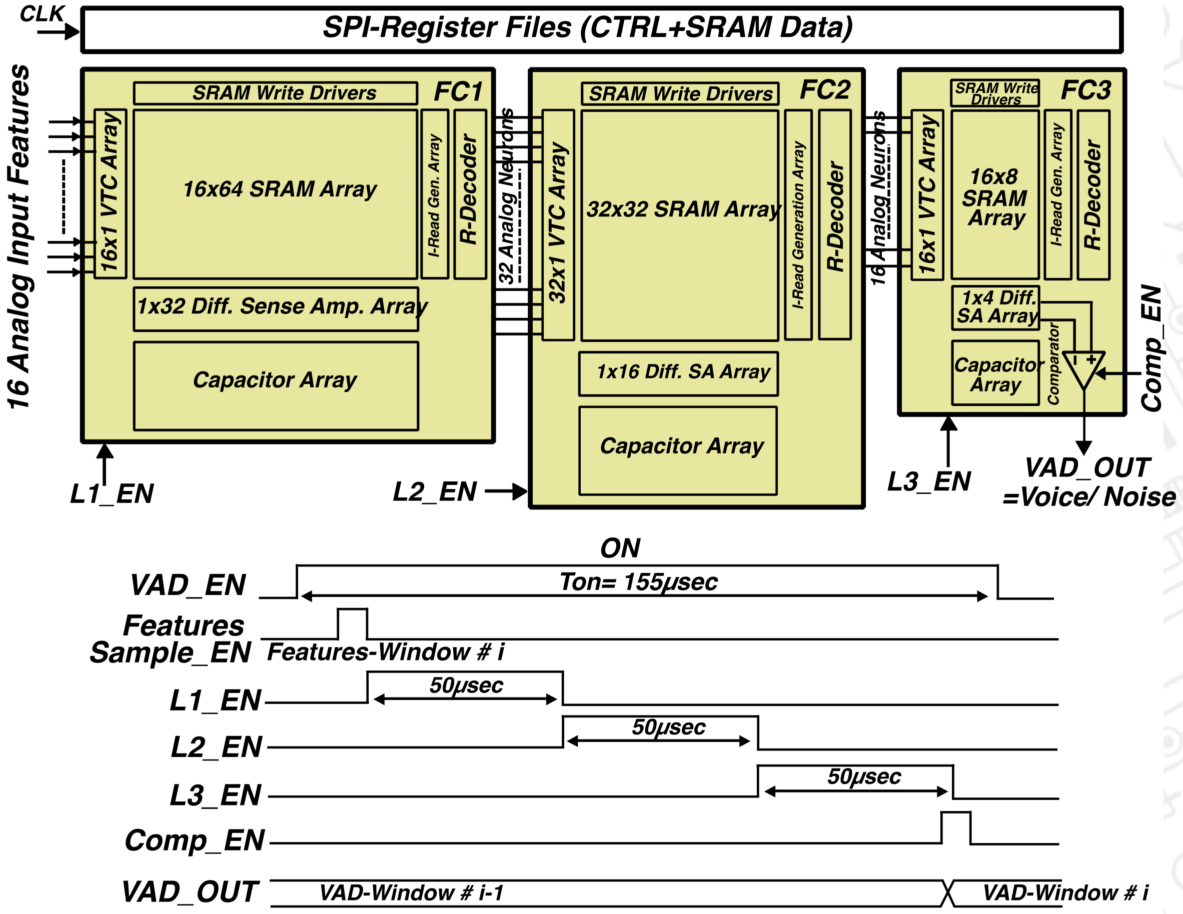
- **Proposed IMC architecture supports cascading macros without data conversion**
  - ➢ Analog inputs and outputs
  - ➢ Low power and higher MAC Accuracy

- **All-Analog Voice-Activity Detection (VAD)**
  - ➢ Process analog features directly without data conversion.
  - ➢ Improves overall system performance



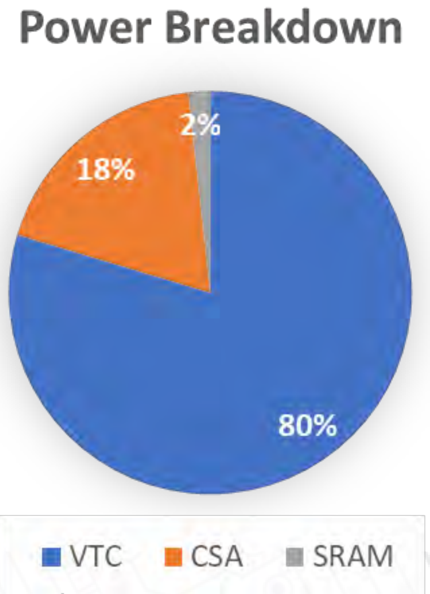
Analog VAD Classifier

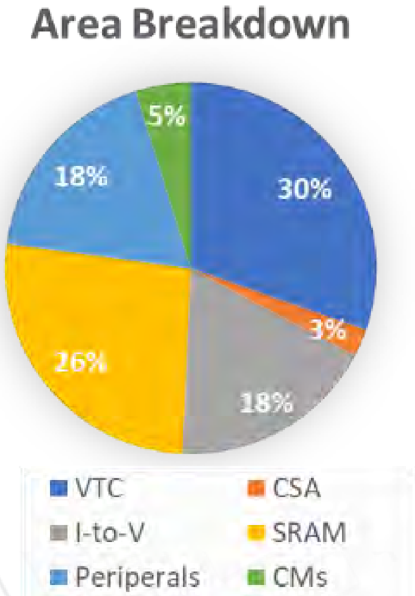
- **3-Layer MLP analog Classifier**
  - ➢ 3-Cascaded analog IMC macros
  - ➢ L_1=16X32, L_2=32X16, L_3=16X8
  - ➢ Voice input features are connected directly to the classifier inputs.
  - ➢ Reuse the output capacitors of one layer as the VTC capacitors of the following layer.

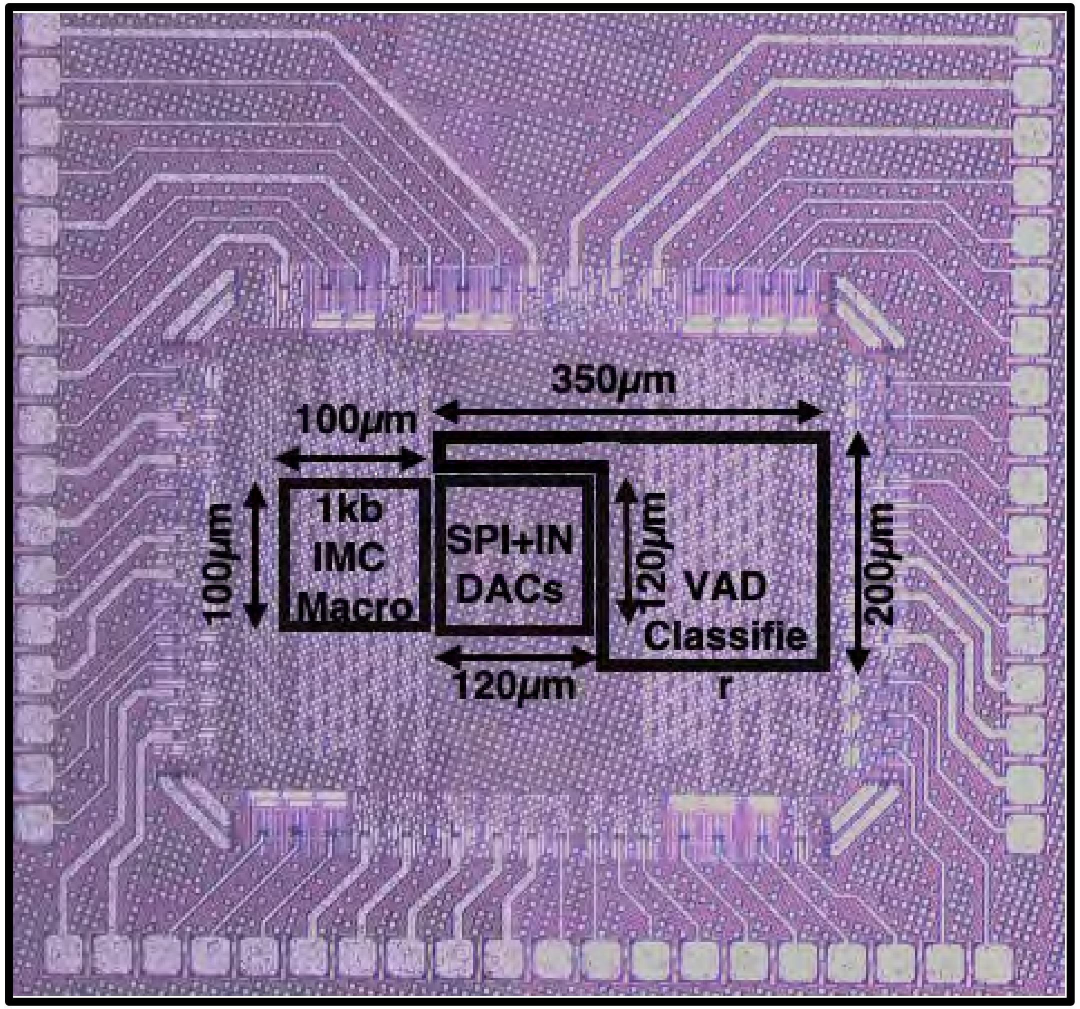


Die Photo

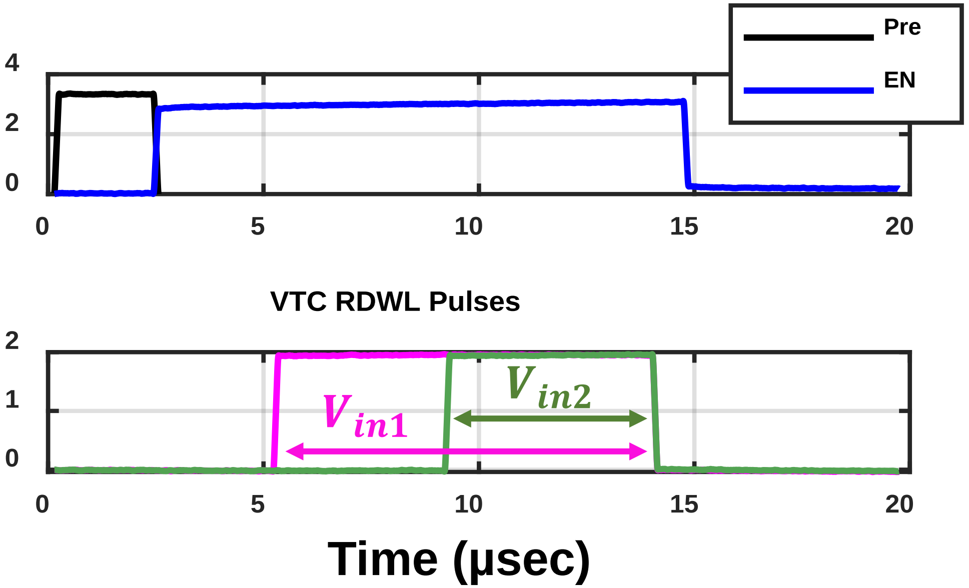
- **14-nm FinFET Technology**
- **1kb IMC SRAM Array**
  - 100μm x 100μm
- **VAD Classifier**
  - 230μm x 200μm
- **SPI + 6b input DACs**
  - 120μm x 100μm





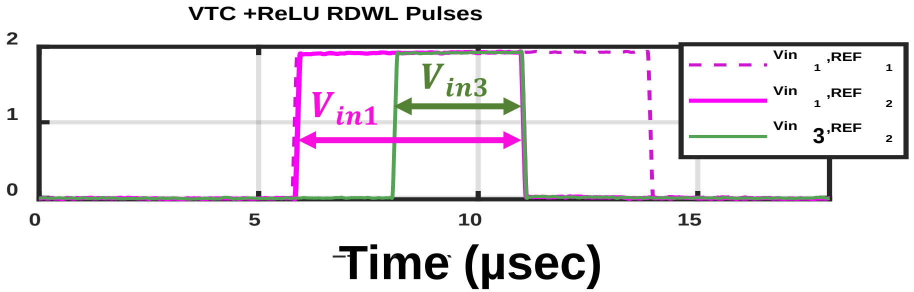


Measurements Results (VTC)



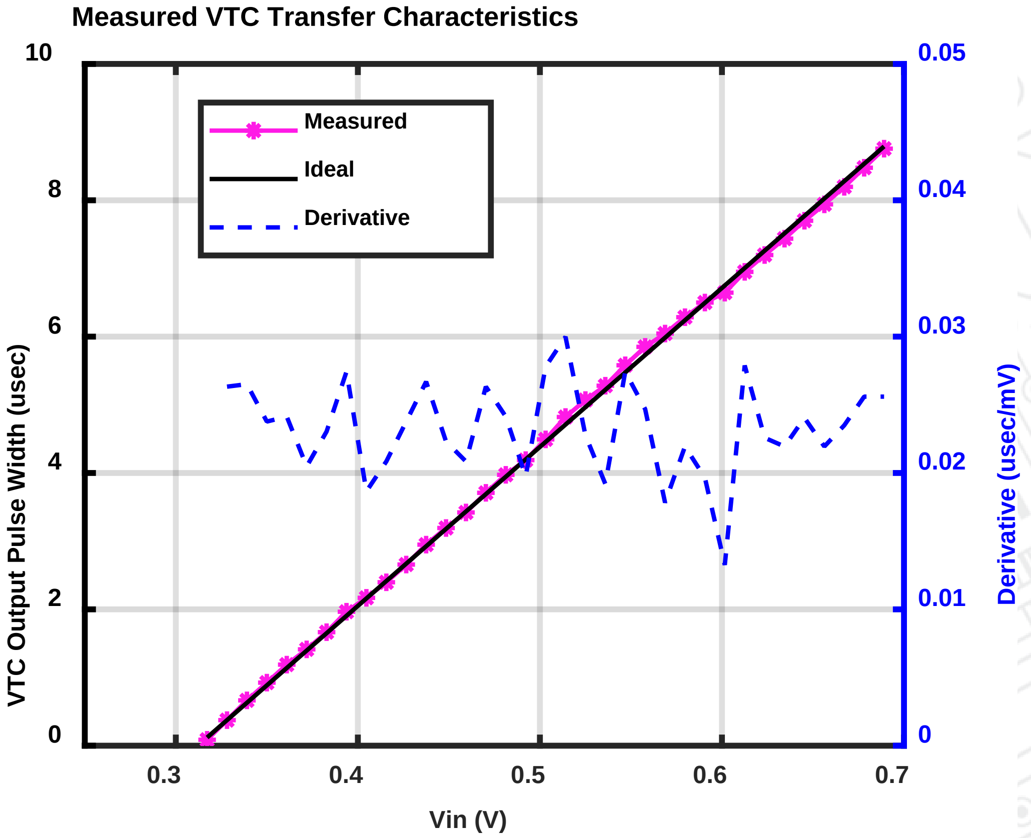
$V_{in1} = 2 * V_{in2}$

$V_{REF} = V_{REF1}$



$V_{in1} = \frac{4}{3} * V_{in3}$

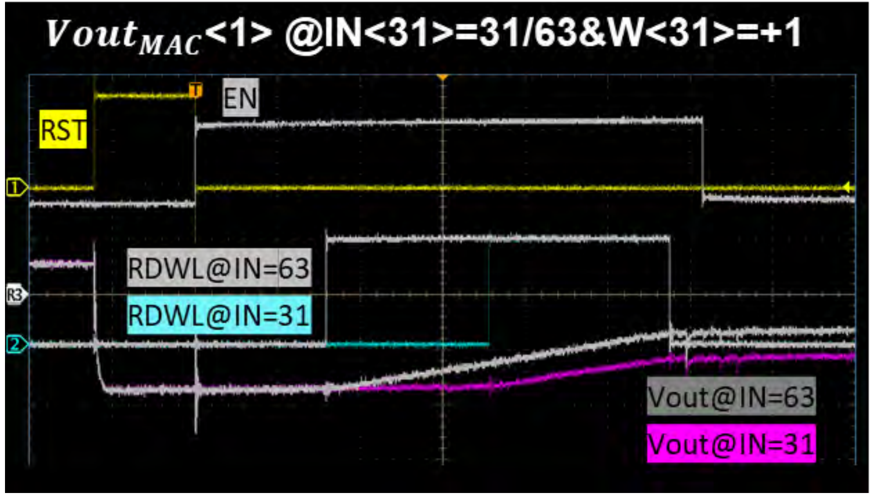
$V_{REF1} < V_{REF2}$



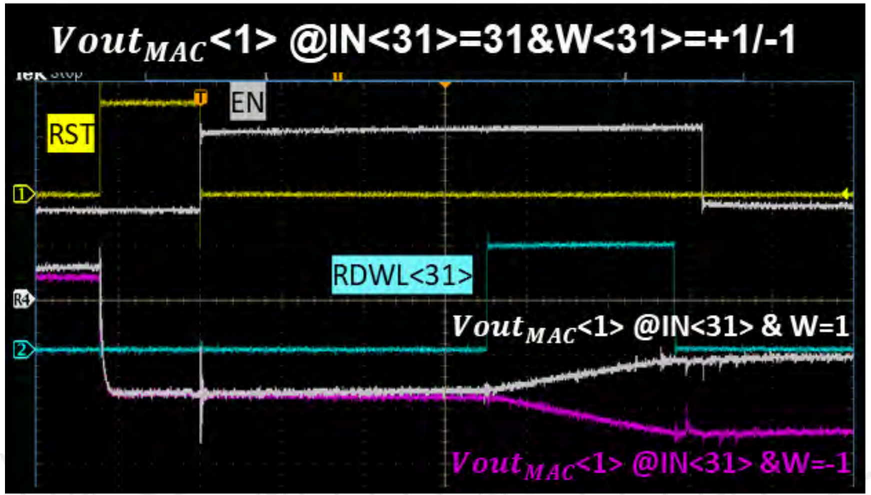
Linear VTC Operation
(-0.4*LSB<DNL<0.3*LSB)

Measurements Results (MAC)

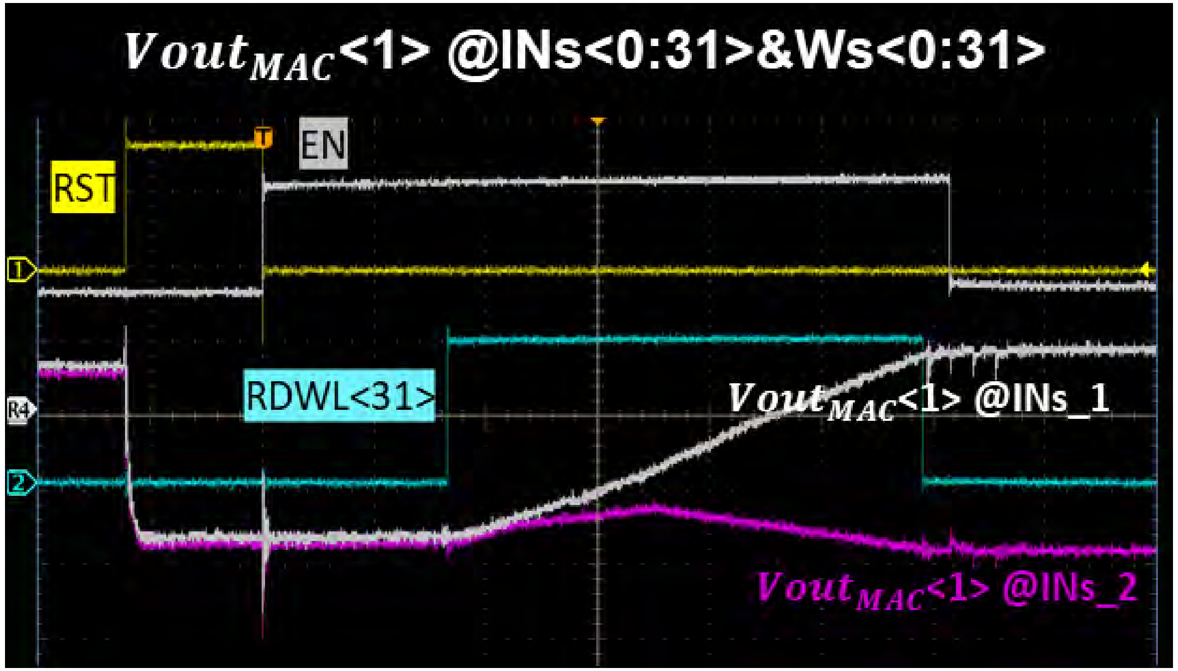
2 different IN Values, W=1



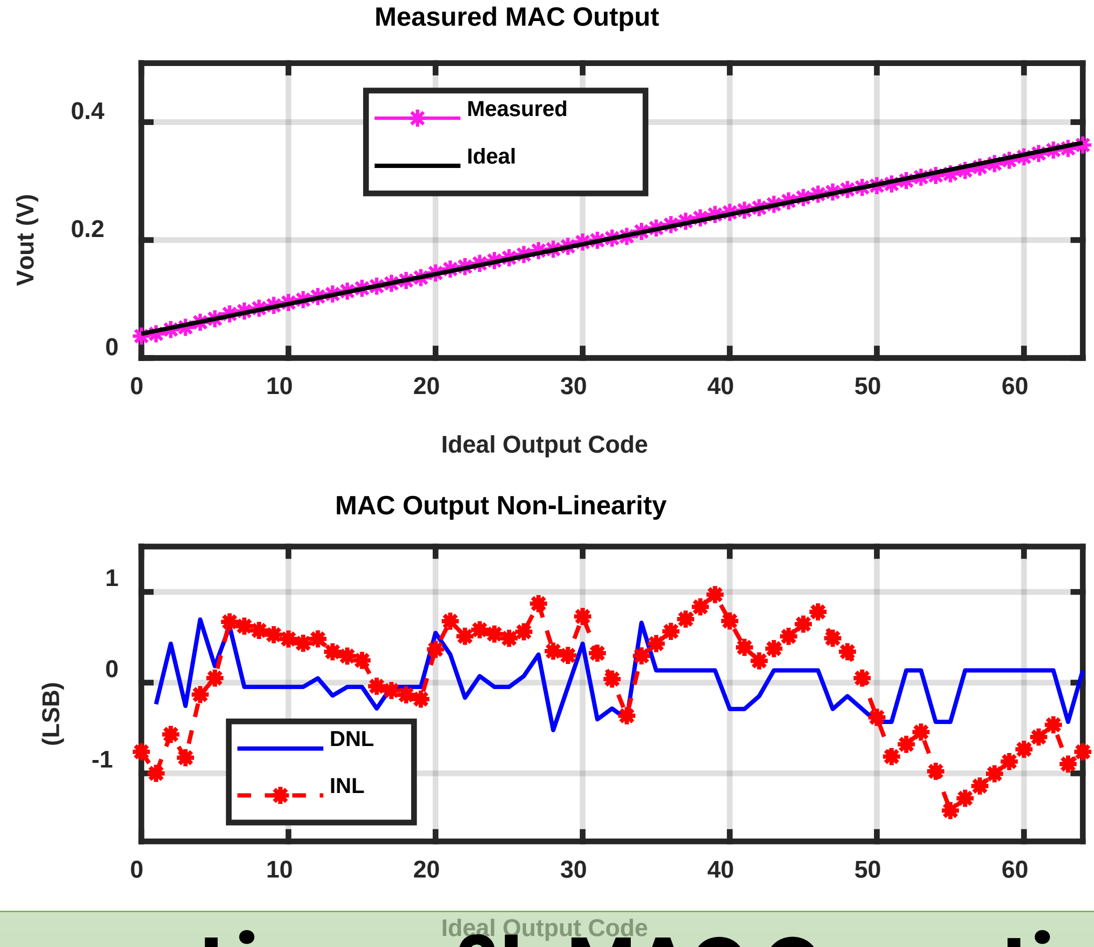
One IN, W=-1/+1



Different 32 IN/ W Values



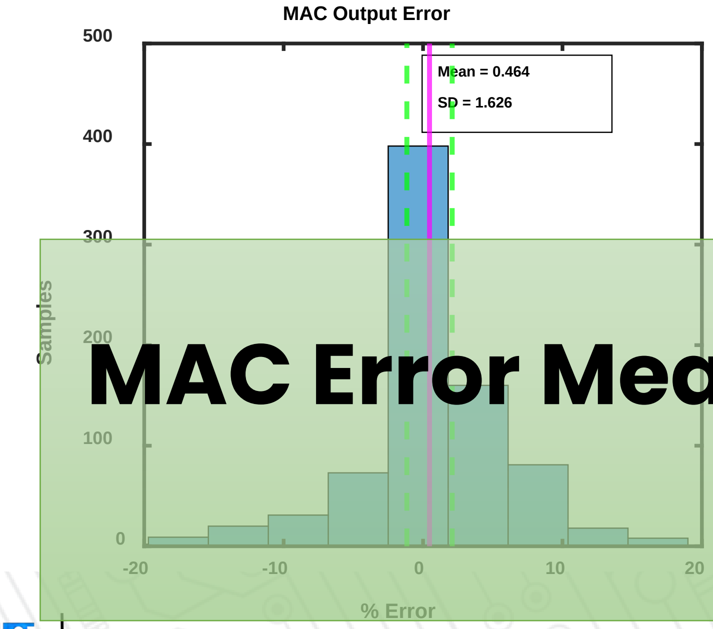
Measurements Results (MAC)

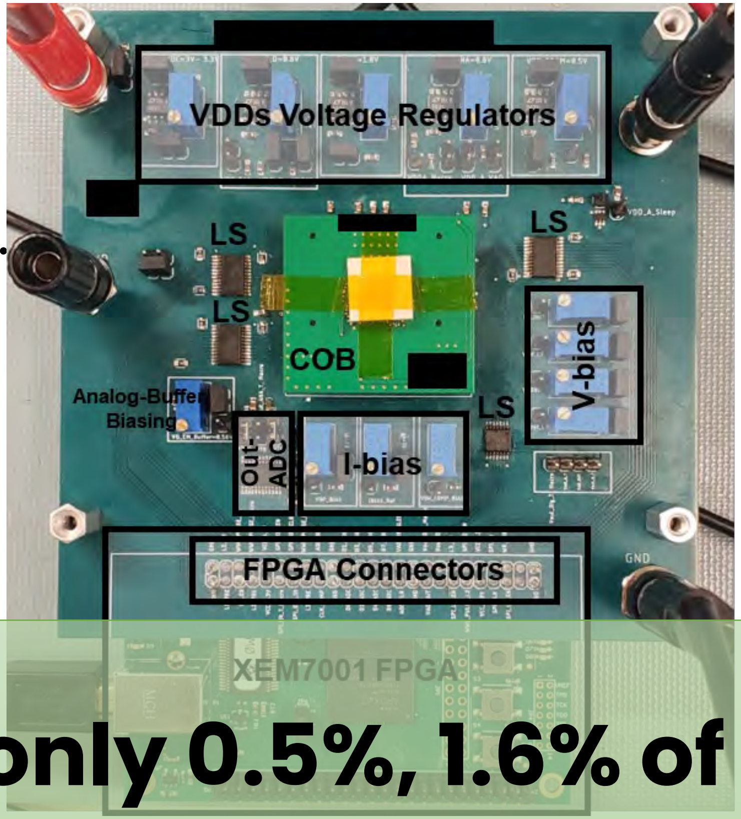


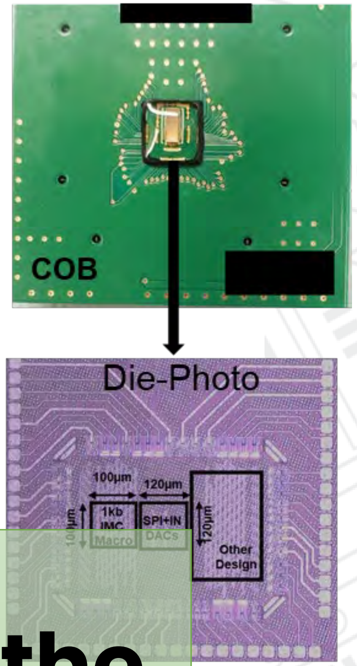
Linear 6b MAC Operation
(-1.4*LSB<INL<0.97*LSB) ,(-0.4*LSB<DNL<0.7*LSB)

Measurements Results (IMC Macro)

- $V_{DD} = 0.8\text{V}$ (Analog), $V_{DD} = 0.5\text{V}$ (Digital), $V_{DD} = 1.8\text{V}$ (Drivers & Buffers)
- $P_{avg} = 1.7\mu\text{W} - 2.4\mu\text{W}$ at $50\text{ kHz} - 100\text{ kHz}$
- 6b inputs, 2b ternary weights, 6b outputs
- An off-Chip ADC samples analog MAC values.
- XEM7001 FPGA checks the MAC accuracy.







MAC Error Mean/Sigma is only 0.5%, 1.6% of the Actual MAC Value

Measured IMC Macro Performance

	This Work			ISSCC'19 [3]	JSSC'22 [5]	VLSI'21 [1]	JSSC'21 [4]	VLSI'22 [2]
Technology	14-nm			55-nm	65-nm	14-nm	65-nm	22-nm
Cell Structure	12T (8T+4T(M_R))			Twin 8T	8T	PCM (8T +4R)	6T	6T
Bit-cell size	1µm x 0.7µm			2*(0.5µm x1.7µm)	1.8µm x 1.8µm	-	2.6 µm^2	-
Compute Mechanism	**Current**			**Current**	**Current**	**Current**	Charge	Charge
MAC Out Sensing	Diff. CSA + Integrating caps			SA+ 5b ADC	SA + 1-5b ADC	8b ADC	CS caps +7b SAR ADC	8b ADC
Macro Size	32 x 32		128 x 128	64 x 60	128 x 128	64 kb	512 x 128	128 kb
Macro area	100µm x 100µm		200µm x 250µm	229.5µm x 165.6µm	234µm x234µm	0.63 mm^2	330µm x 530µm	0.25 mm^2
Input/Output precision	**Analog (I:6b, O:6b)**	**Analog (I:5b, O:5b)**	**Analog (I:5b, O:5b)**	I:1b/2b/4b, O:3b/5b/7b	I:1b, O:1b-5b	(I:8b, O:8b)	Analog (I:4b, O:6b)	Digital (I:8b, O:8b)
Weight precision	**Ternary (2b)**			2b/5b	Binary (1b)	Analog (4b)	1b/ 2b	8b
Supply Voltage	0.5V (Digital) / 0.8V (Analog)			1V	0.45V/ 0.8V	0.8V	1.2V	0.7- 1.1V
Bitwise operations/cycle	12k (I:6b, W:2b, O:6b)	10 k (I:5b, W:2b, O:5b)	160 k (I:5b, W:2b, O:5b)	432 (I:1b, W:2b, O:3b)	32 k (I:1b, W:1b, O:1b)	32k	64 k (I:4b, W:1b, O:6b)	290k
Access Time	20µs [3] (50kHz)	10µs [2] (100kHz)	10µs [2] (100kHz)	3.2 ns	-	128 ns	14.3 ns	4.54 ns
Macro Energy/cycle	34pJ [3]	17pJ [2]	77pJ [2]	2.94 pJ	-	12.2 nJ	200.4 pJ	249 pJ
Throughput (GOPS)[1]	0.6 [3]	1 [2]	16.4 [2]	135	6132.7	32768	4587.2	38400-64000
Compute Density (TOPS/mm^2)[1]	0.06 [3]	0.1 [2]	0.328[2]	3.6	112 (w/. SA)	50.8	27.2	154-256
Energy Efficiency (TOPS/W) [1]	[illegible] [3] (measured)	**[illegible] [2] (measured)**	[illegible] (simulated)	144.06	400	336	815.2	992- 2060

High Energy-Efficiency with Good Linearity

[1] Bitwise metric: normalized to 1b IN & 1b W, 1MAC= 2 ops (1 mult + 1 add) [2] 100kHz (I:5b, W:2b, O:5b), [3] 50kHz (I:6b, W:2b, O:6b)

Measured VAD Analog Classifier Performance Summary

	This Work	**ISSCC'22 [6]**	**ISSCC'18 [7]**
Technology	14-nm	28-nm	180-nm
Network Size	3 FC (2.125 kb)	BNN- 4 FC (2.6 kb)	BNN- 4 FC (4.5 kb)
Area	0.052 mm^2	0.01 mm^2	0.6 mm^2
MAC Mode	Analog	Digital	Digital
Input precision	Analog (6b)	1b	Digital (9b)
Weight precision	Ternary (2b)	BNN (1b)	BNN (1b)
Neurons precision	Analog	1b	1b
Supply Voltage	0.5V/ 0.8V	0.6V	0.55V
Classification Rate	6.45kHz	100 Hz	100 Hz
Classification Power	3.71 μW	34.8 nW	620 nW*
Classification Energy	0.575 nJ	0.35 nJ	6.2 nJ
VAD Accuracy	**89.8%** **5-dB** SNR, White Noise	94% 10-dB SNR, NOISEX-92	85% 10-dB SNR, Restaurant Noise

* = Total power- (feature extraction / front-end power)

Low-Power and Accurate Implementation

Key Message

- **Current-Controlled SRAM Read**
  - Low power subthreshold MAC operation
  - Controlled MAC speed and extended signal margin
  - Robust against process variations

- **Reconfigurable Time-Domain MAC**
  - Support multi-bit inputs and weights
  - Improved MAC linearity

- **Differential Current Sensing**
  - Improved MAC linearity
  - Cancels mismatches and SRAM leakage currents.

- **Analog Current-Domain IMC**
  - Supports cascading layers.
  - Energy-efficient and accurate computing.
  - Can be used in all-analog classifier architectures.

Acknowledgement

- **This work is supported by the MIT-IBM Watson AI Lab.**
- **The authors would like to express our gratitude to the IBM teams:**
    - Thank Kevin Tien, Cliff Osborn, Seiji Munetoh, Kohji Hosokawa for tape-out support and valuable discussions;
    - Thank Cyril Cabral, Kai Schleupen, John Timmerwilke for packaging solutions;
    - Thank Dirk Pfeiffer, Daniel Friedman, Dan Dechene, David Cox, Mukesh Khare for management support
- **The authors thank members of the Energy Efficient Circuits and Systems Group at MIT for their valuable discussion and feedback.**

Please reach out to aya_amer@mit.edu **with any questions or feedback!**

Thank You!

Questions?

References

1. C. Yu, et al. (2022). A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks. IEEE Journal of Solid-State Circuits, 57(11), 3466-3476.
2. H. Wang, et al. (2022). A 32.2 TOPS/W SRAM compute-in-memory macro employing a linear 8-bit C-2C ladder for charge domain computation in 22nm for edge inference. In 2022 IEEE Symposium on VLSI Technology and Circuits (pp. 36-37). IEEE.
3. Xin Si, et al. "24.5 A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning." 2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2019.
4. Z. Chen, et al. (2021). CAP-RAM: A charge-domain in-memory computing 6T-SRAM for accurate and precision-programmable CNN inference. IEEE Journal of Solid-State Circuits, 56(6), 1924-1935.
5. C. Yu, et al. (2022). A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks. IEEE Journal of Solid-State Circuits, 57(11), 3466-3476.
6. F. Chen, et al. (2022, February). A 108nW 0.8 mm2 Analog Voice Activity Detector (VAD) Featuring a Time-Domain CNN as a Programmable Feature Extractor and a Sparsity-Aware Computational Scheme in 28nm CMOS. In 2022 IEEE International Solid-State Circuits Conference (ISSCC) (Vol. 65, pp. 1-3). IEEE.
7. M. Yang, et al. "Design of an Always-On Deep Neural Network-Based 1- $ \mu$ W Voice Activity Detector Aided With a Customized Software Model for Analog Feature Extraction," in IEEE Journal of Solid-State Circuits, vol. 54, no. 6, pp. 1764-1777, June 2019.