A COMPUTING-IN-MEMORY ENGINE FOR SEARCHING ON HOMOMORPHICALLY ENCRYPTED DATA

Dayane Reis, Michael Niemier and X. Sharon Hu
University of Notre Dame

Arm Research Summit, September 2019

ASCENT
Applications and Systems Driven Center for Energy-Efficient Integrated Nanotechnologies

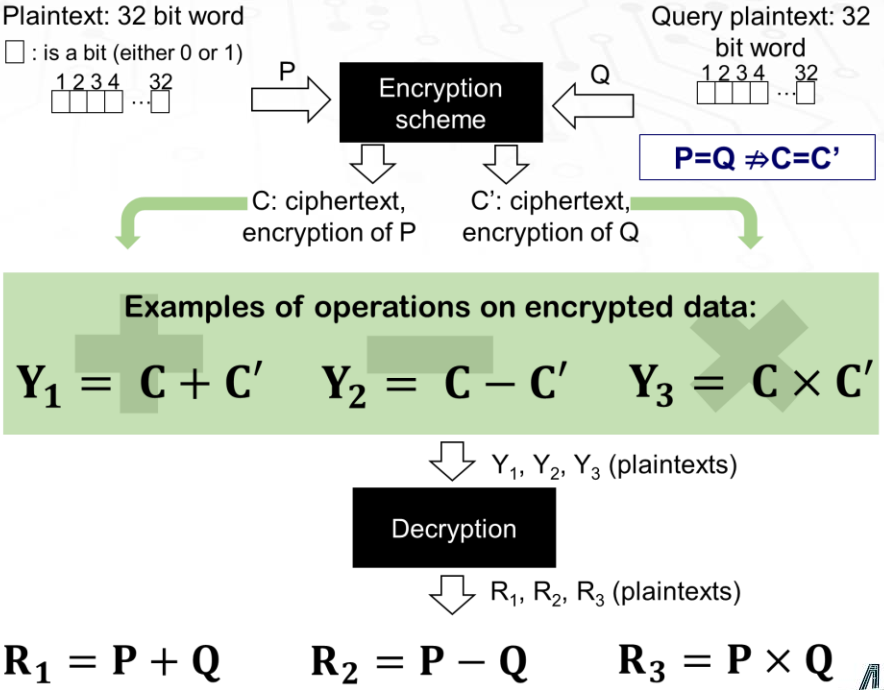
This work was supported in part by the Semiconductor Research Corporation (SRC) and DARPA.

LET’S LOOK AT HOMOMORPHIC ENCRYPTION!

Homomorphic encryption (HE)

- Enables computation to be performed on encrypted data
  - No prior decryption is necessary!

Daniel J. Solove, 2015



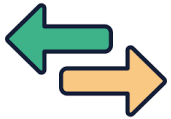
HE challenges

- Ciphertext sizes: 10s or even 100s of KB for a 32 or 64-bit word in plaintext

Long latency of computations

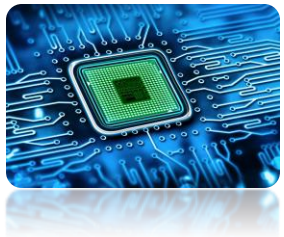
How long does it take to run?

Memory transfers



How long is to fetch ciphertexts?

Memory density



Where to store all these ciphertexts?

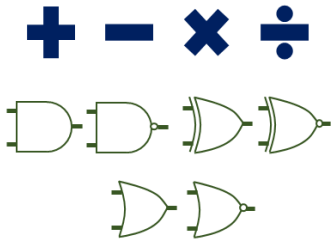
HOW TO ADDRESS THESE ISSUES?

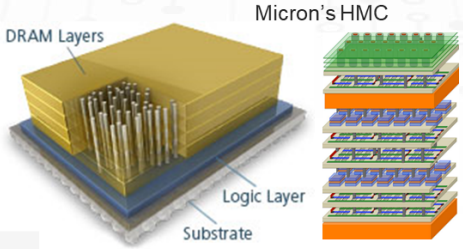
Computing-in-memory (CiM)

- Mixing **logic** and **memory** on same chip
  - Energy and performance improvements due to reduced data movement!

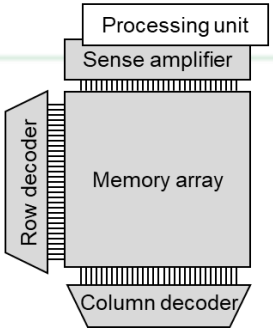
Brings computation to memory

Possible operations:
Arithmetic & Logic





Integration of logic layers in 3D memory



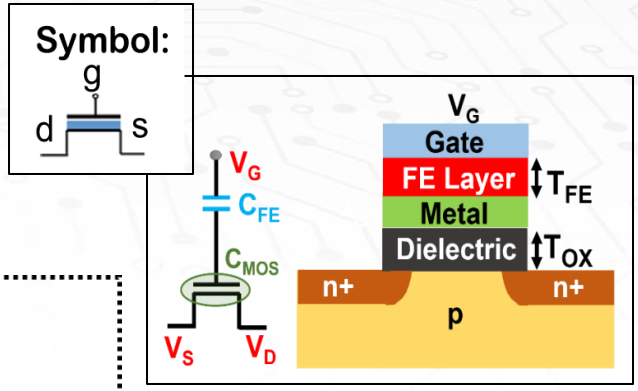
- **FeFET-CiM**[2]
- **STT-CiM**[3]
- **Compute caches** [4] …

Modifications in the peripheral circuits

[2] Reis D, Niemier M, Hu XS. Computing in memory with FeFETs. ISLPED 2018 (SRC P094461)
[3] S. Jain, A. Ranjan, K. Roy and A. Raghunathan, "Computing in Memory With Spin-Transfer Torque Magnetic RAM," in IEEE TVLSI, 2018
[4] S. Aga, S. Jeloka, A. Subramaniyan, S. Narayanasamy, D. Blaauw, and R. Das, "Compute caches," in *2017 IEEE HPCA*, 2017

Ferroelectric FETs and their applications for memory

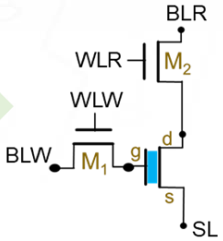
- **Construction resembles MOSFETs**
  - Ferroelectric material added in the gate stack
- **FE material responsible for hysteretic behavior**
  - Device works as both a switch and a memory
- **Fabrication prototypes available**
  - By Dr. Suman Datta (ND), NamLab, GF.



Adapted from X. Yin, 2017

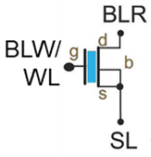
FeFET-based memory cells used in CiM

2T+1FeFET



	BLW	WLW	BLR	WLR	SL
Write "1"	V_{write}	V_{DD}	0	0	0 → V_{write}
Write "0"	0	V_{DD}	0	0	V_{write} → 0
Read	0	0	V_{read}	V_{DD}	0

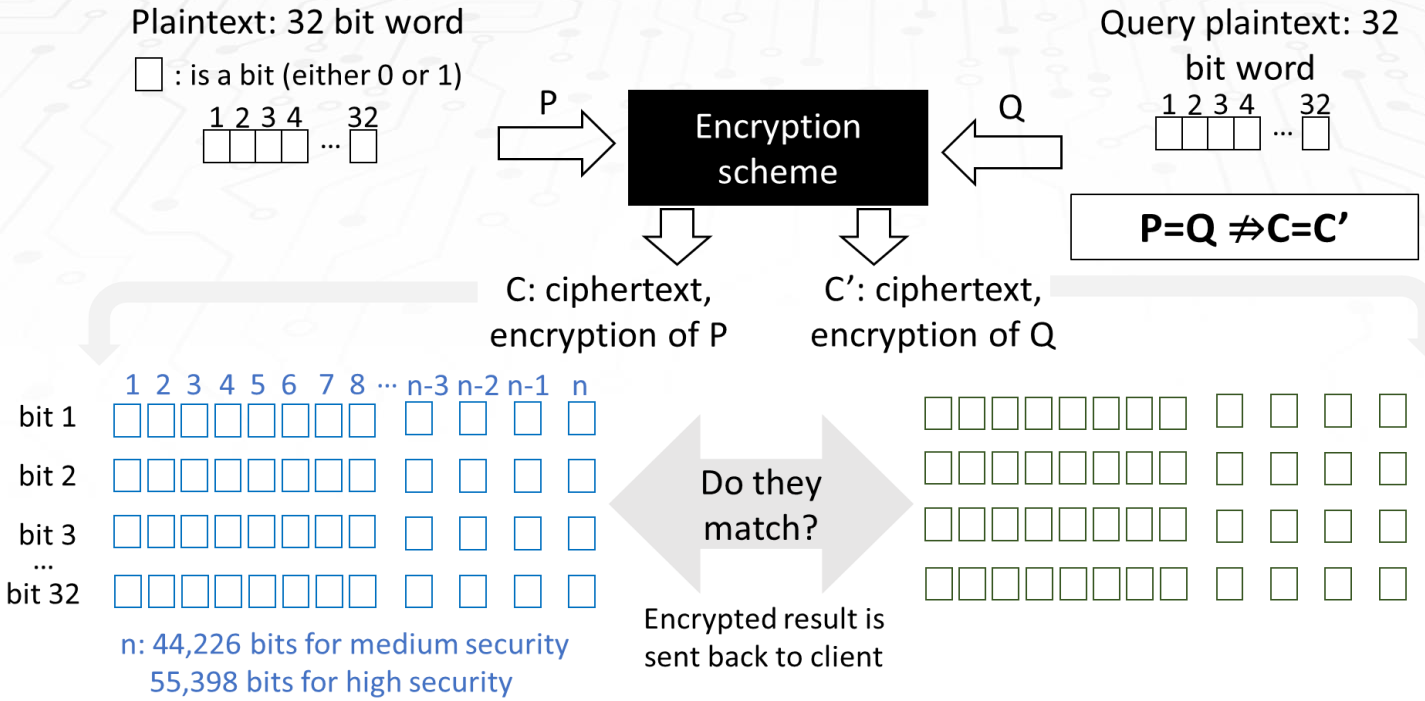
1FeFET



	BLW/WL	BLR	SL
Write "1"	V_{write}	0	0
Write "0"	$-V_{write}$	0	0
Read	V_{read}	V_{read}	0

Vwrite = ±4V. Vread =0.5V
(pre-charged bitline)
Preisach-based model

SeCAM: Secure content addressable memories High level view



SeCAM: Secure content addressable memories
Fully-additive search function

Conventional search function performed with a CAM:

$$\prod_{i=1}^{w} \overline{C_i \oplus C'_i}$$

XNOR-AND operations in HE context require computationally intensive multiplications

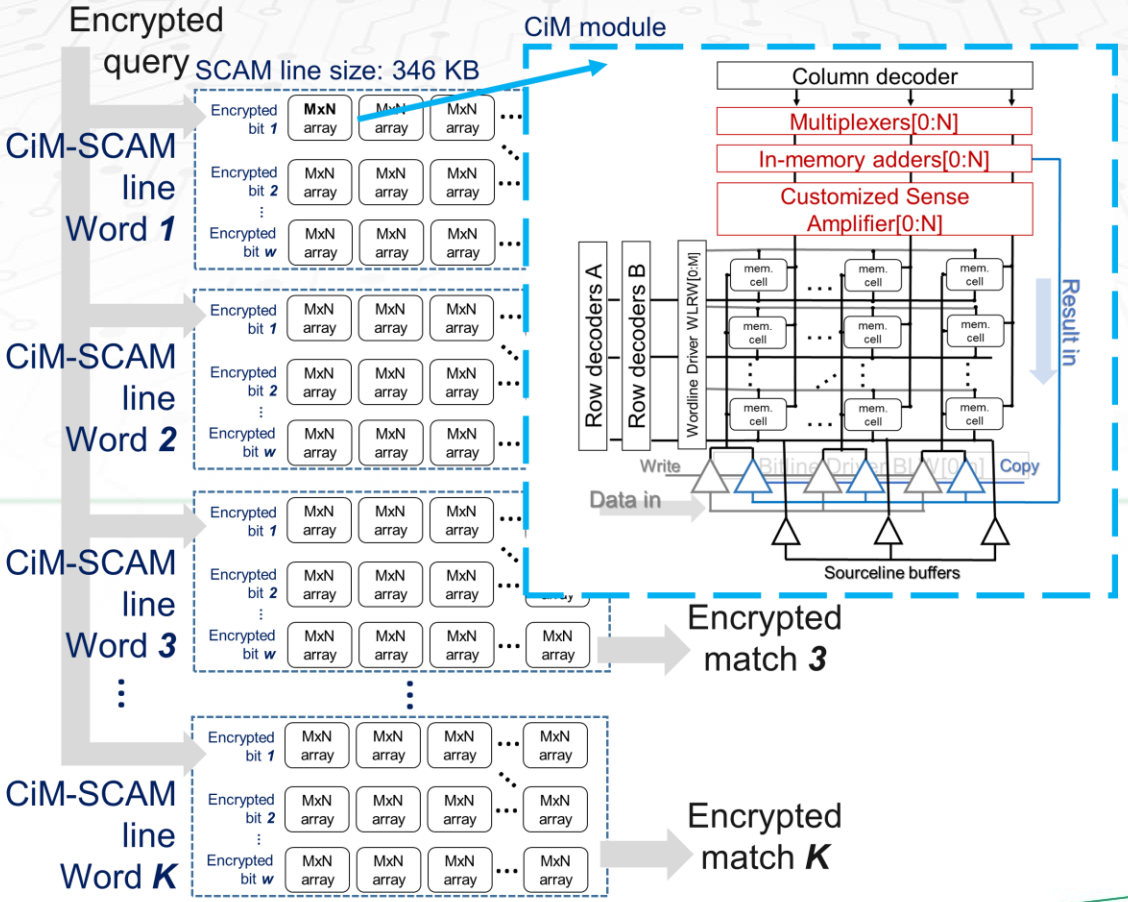
Too much noise is introduced into the ciphertexts

Modified search function performed with secure CAM (SCAM[1]):

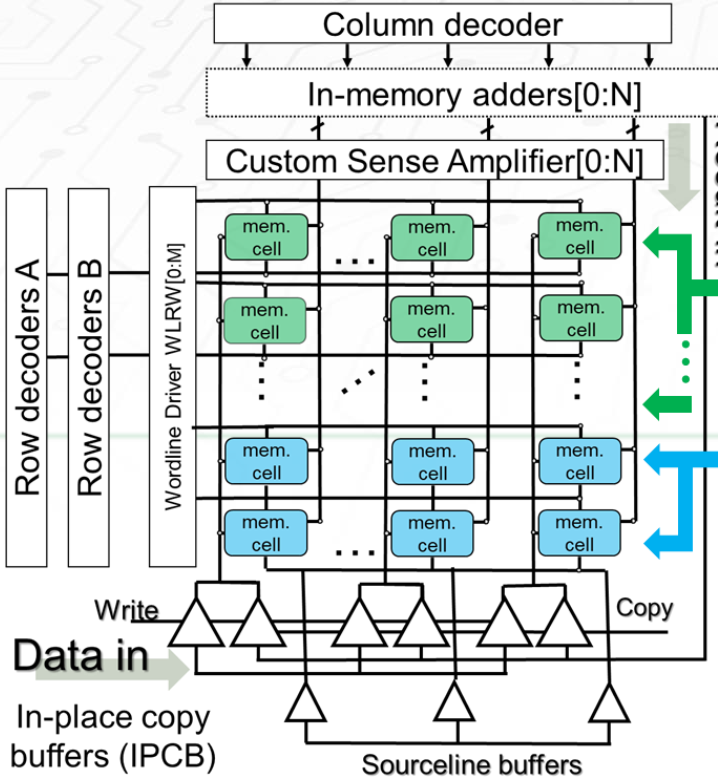
$$\sum_{i=1}^{w} C_i + (-C'_i)$$

[1] S. Bian, M. Hiromoto, and T. Sato, "SCAM: Secured content addressable memory based on homomorphic encryption," in DATE, 2017

CiM-SeCAM: A CiM engine to facilitate search in HE



CiM-SeCAM: Sequence of operations




```
CiM_SCAM(){
    accumulator[1] = ciphertext[1];
    for(i=1 to w){
        if (i%2 ==0)
            src=1, dest=2;
        else
            dest=1, src=2;

        result=ciphertext[i]+
                accumulator[src]

                accumulator[dest] =
                result;
    }
    Encrypted match = result
    }
```

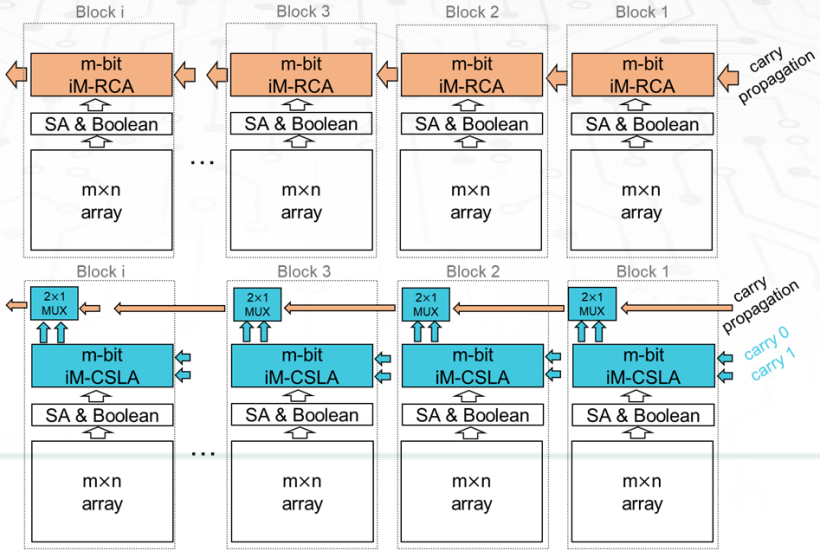
CiM module: in-memory (iM) adders

iM-RCA
(ripple carry adder)
Serial carry computation

1024-bit iM addition
(SRAM-CiM)
43.22 ns

iM-CSLA
(carry select adder)
Parallel carry computation

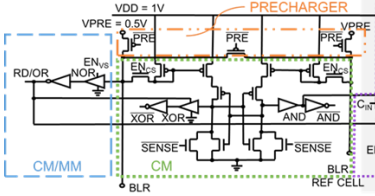
1024-bit iM addition
(SRAM-CiM)
11.13 ns



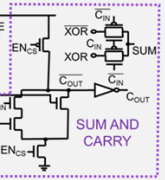
Schematics of iM adders:

Customized sense amplifier

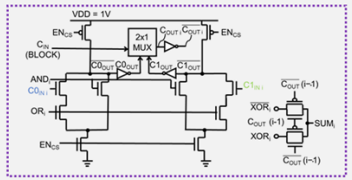
D. Reis, et al, ISLPED 2018 (SRC P094461)



Ripple carry adder (RCA)



Carry select adder (CSLA)



RESULTS AND PERSPECTIVES

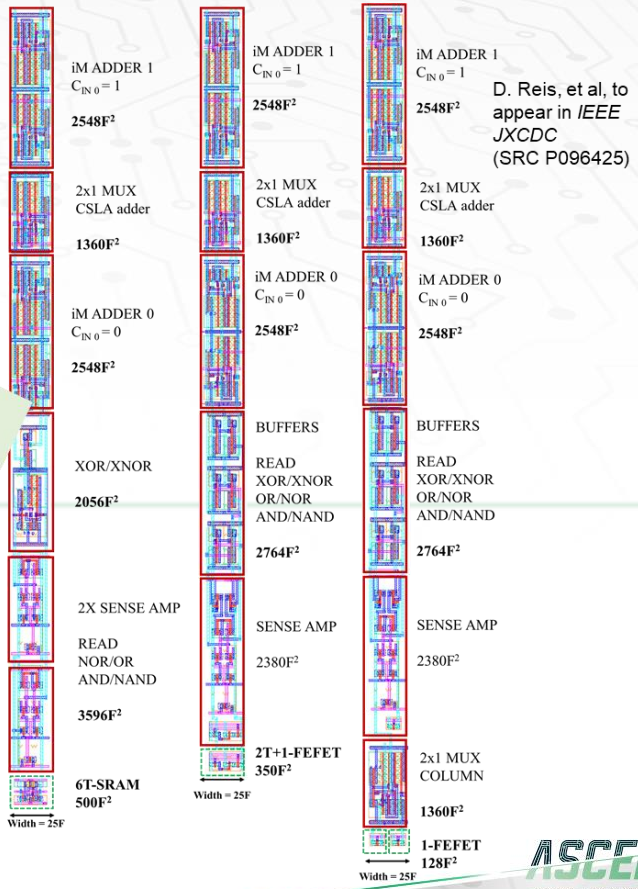
Evaluation and results

- To **measure** search **energy/time**:
  - Spice simulations of a 64x64 array and peripherals

  (see “Figure of Merit” for results)

- To **estimate** area:
  - Layout for CiM-SCAM implemented with iM-CSLA

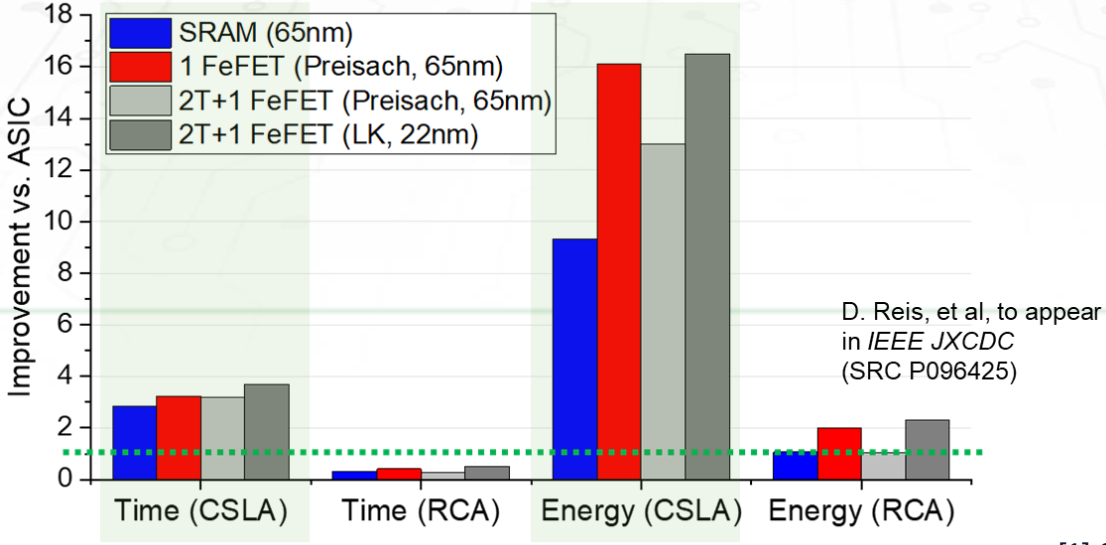
SRAM, 2T+1-FeFET and 1-FeFET memory cells are considered



Figures-of-Merit (FOM)

- Search on homomorphically encrypted data

CiM approaches vs. an ASIC[1]



D. Reis, et al, to appear in *IEEE JXCDC* (SRC P096425)

With SRAM: 9.5X energy and 2.8X search time reduction
With FeFET: up to 16X energy and 3.2X search time reduction

[1] S. Bian, M. Hiromoto, and T. Sato, "SCAM: Secured content addressable memory based on homomorphic encryption," in DATE, 2017

Toward CIM-based Fully homomorphic encryption

- Focus on FHE based on **Ring Learning-With-Error (RLWE)**
- Operations are computed over polynomial rings (e.g., $\mathbb{Z}_q[X]/\Phi_M(X)$)
  - I.e., $\{a_n x^n + \cdots a_1 x + a_0 | a_0, \cdots, a_n \in \mathbb{Z}_q\}$ where $n = 2^d$ for some d
- Core ops – homomorphic addition, multiplication, and re-linearization (noise reduction) –facilitated with 3 basic primitives

Division with rounding: $\left\lfloor \frac{x}{y} \right\rceil$

- CIM friendly if $y = 2^k$ for some k

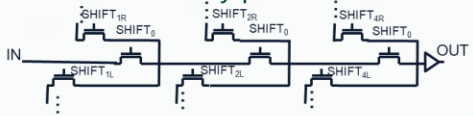
Modular reduction: $[x]_q = x - q\left\lfloor \frac{x}{q} \right\rceil$

- CIM friendly if $q = 2^k$ for some k

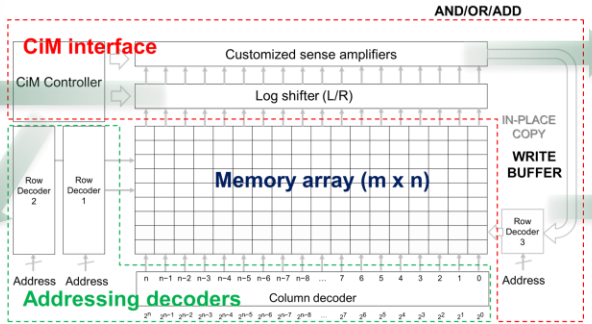
Modular arithmetic: $[x \pm y]_q, [x \times y]_q$

- CIM friendly if $q = 2^k$ for some k

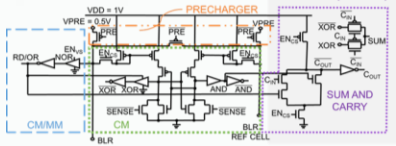
Performs direct div/mult by powers of 2



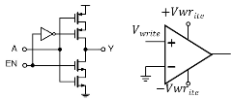
Coordinates sequence of CiM operations



Performs bitwise logic/ ADD between memory words



Write buffers (for SRAM or FeFETs)



ASCENT
Applications and Systems Driven Center for Energy-Efficient Integrated Nanotechnologies

THANK YOU!
QUESTIONS?

JUMP

Joint University Microelectronics Program

www.src.org/program/jump

Semiconductor Research Corporation

@srcJUMP