IEEE/ACM
2022 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN
ICCAD
41st Edition

COSIME: FeFET based Associative Memory for In-Memory Cosine Similarity Search

Che-Kai Liu[1,2,3], Haobang Chen[1], Mohsen Imani[3], Kai Ni[4], Arman Kazemi[2], Ann Franchesca Laguna[5], Michael Niemier[2], Xiaobo Sharon Hu[2], Liang Zhao[1], Cheng Zhuo[1], and Xunzhao Yin[1]

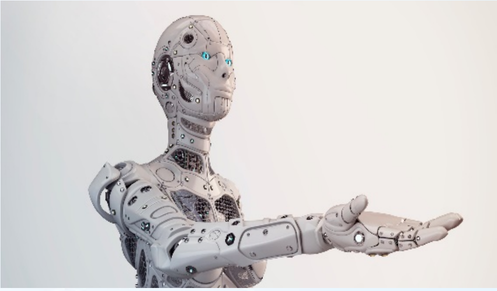
[1]Zhejiang University, Hangzhou, China, [2]University of Notre Dame, USA
[3]University of California, Irvine, USA, [4]Rochester Institute of Technology, USA,
[5]De La Salle University, Manilla, Philippines

Von-Neumman Paradigm

Self Driving Cars

Healthcare

Smart Robots



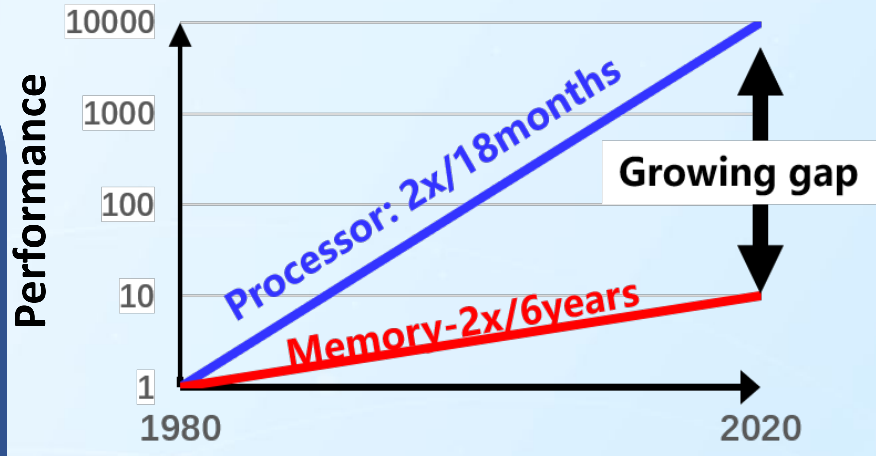
Finance

Machine Translation

Gaming

Operation	Energy (pJ)	Relative Cost
32-bit int ADD	0.1	1
32-bit int MULT	3.1	31
32-bit 32KB SRAM	5	50
32-bit DRAM	640	6400

Memory scaling is low

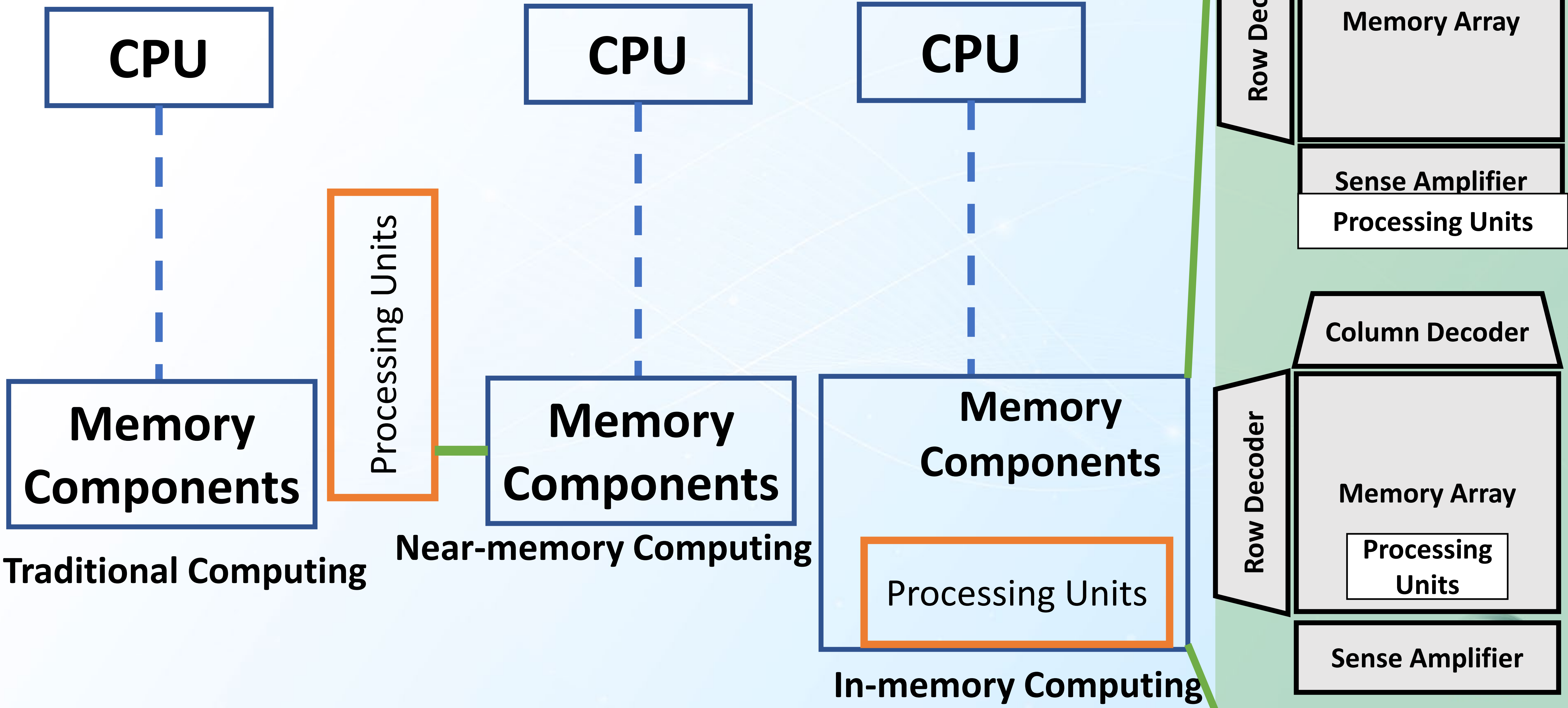


$$\left(\text{Performance} = \frac{1}{\textit{Execution time}}\right)$$

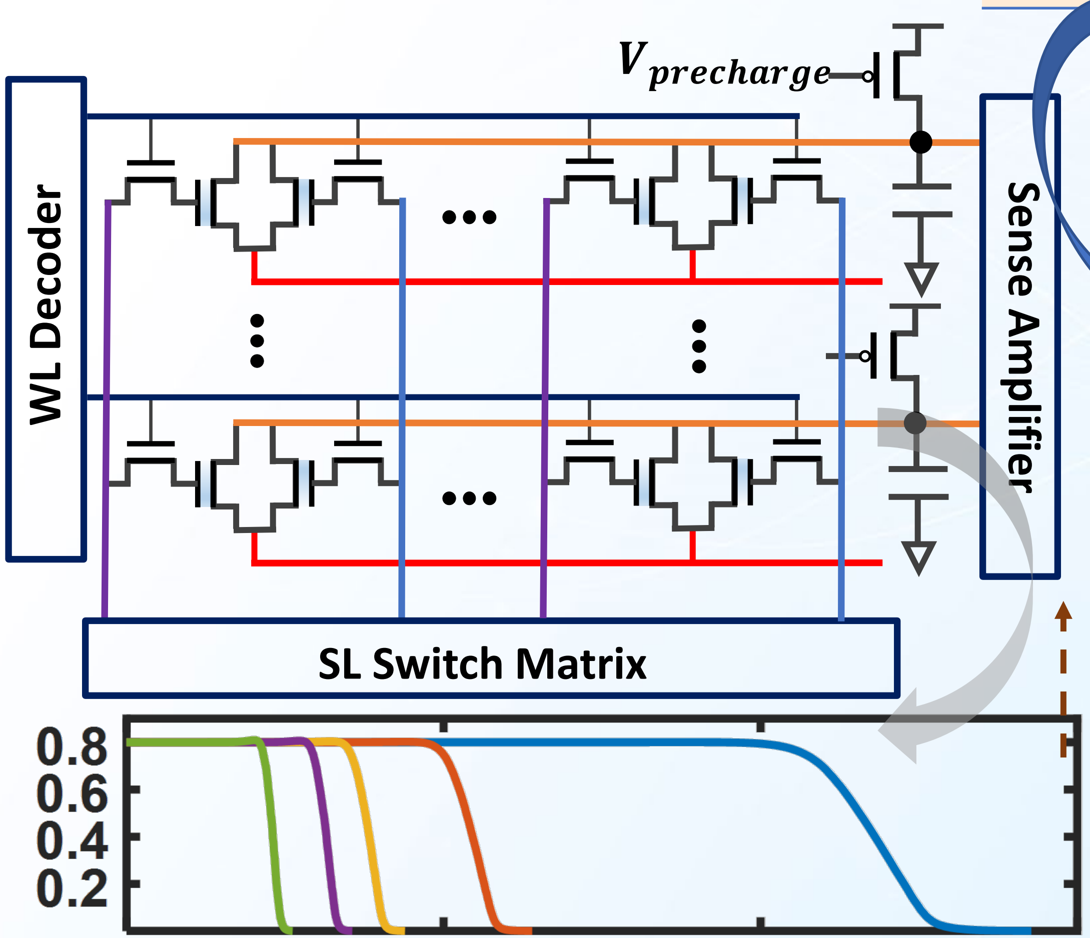
- **Concerns**
  - **Latency**
  - **Energy**
  - **Security**
  - **…**

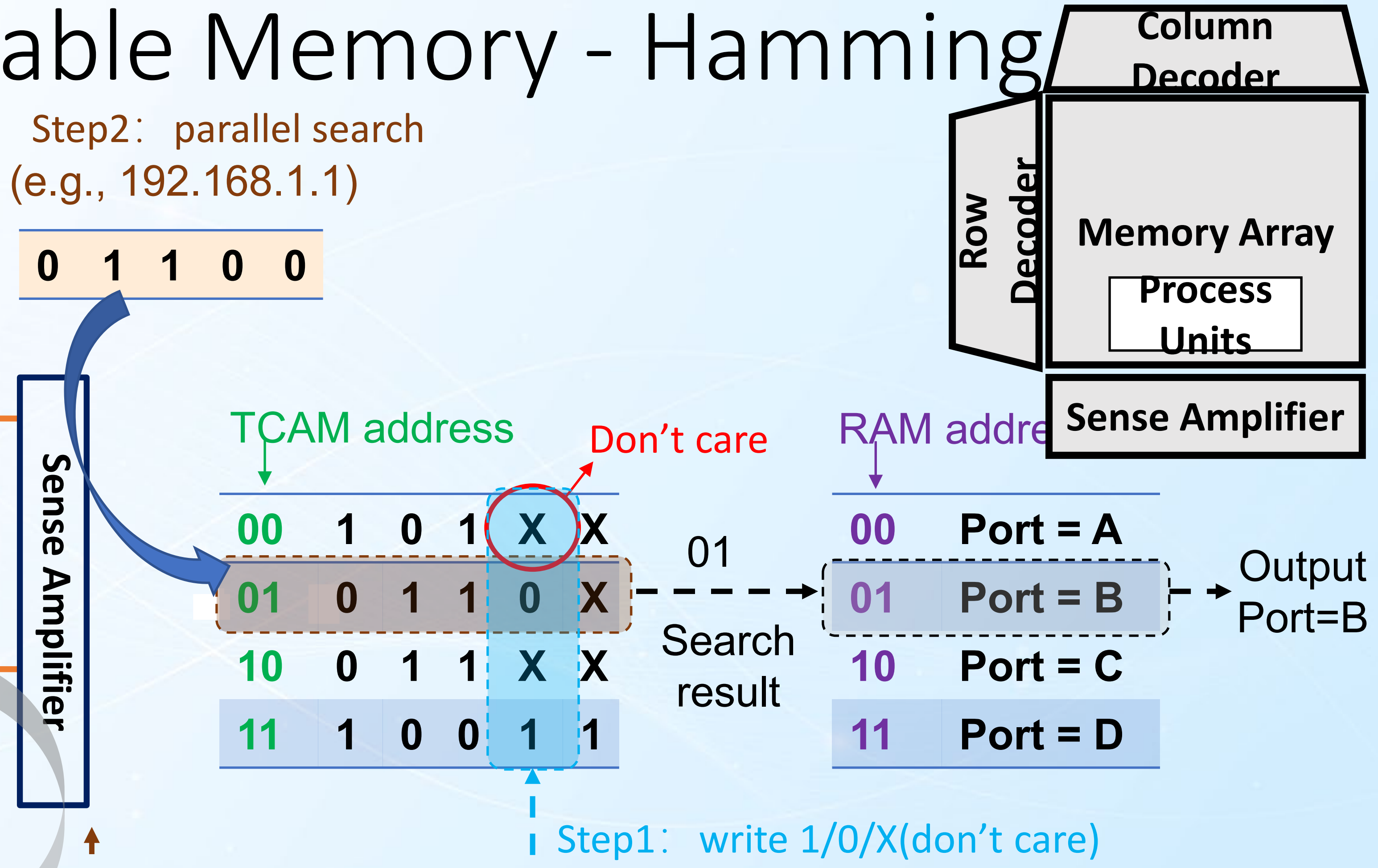
M. Horowitz. Energy table for 45nm process. Stanford VLSI wiki. [Online]. Available: https://sites.google.com/site/seecproject

Computing Paradigms



Content Addressable Memory - Hamming

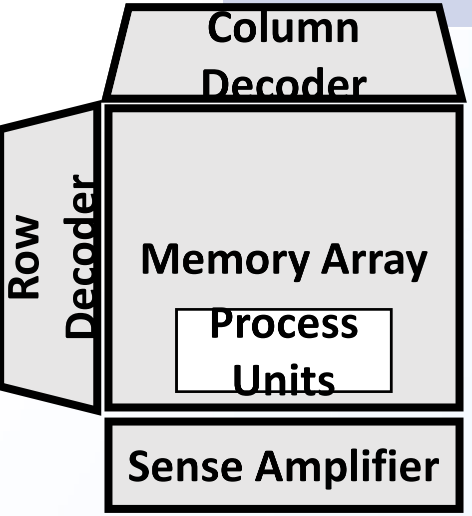


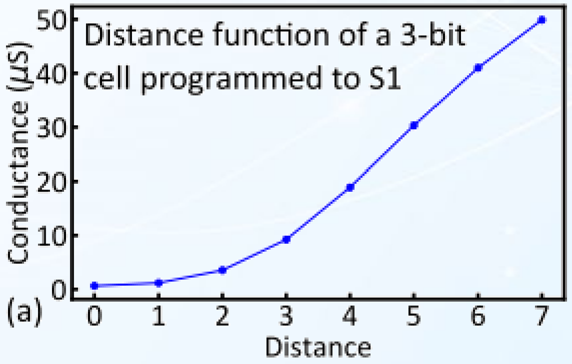


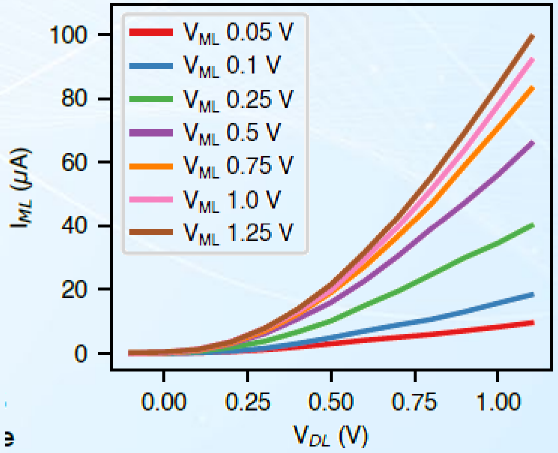
Hamming distance!

Content Addressable Memory – Sigmoid/L2

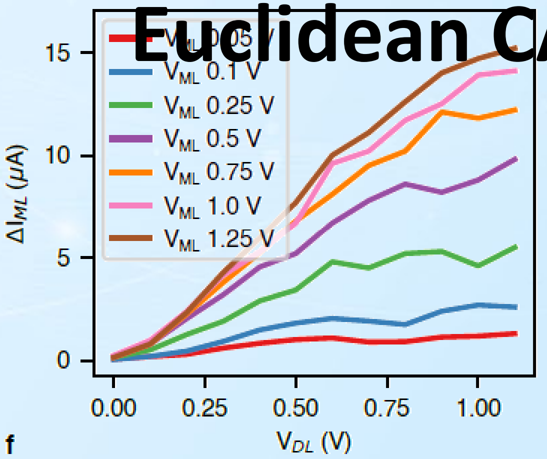
Region	Conditions	Expression	Distance
Cut-off	$V_{gs} \leq V_{th}$	$I_{ds} = 0$	NA
Linear (short)	$V_{gs} > V_{th}, V_{ds} \leq V_{gs}\text{-}V_{th}$	$I_{ds} \propto (V_{gs}-V_{th})\, V_{ds} - \frac{V_{ds}^2}{2}$	Sigmoid [1]
Saturation (long)	$V_{gs} > V_{th}, V_{ds} > V_{gs}\text{-}V_{th}$	$I_{ds} \propto (V_{gs}-V_{th})^2$	Squared Euclidean [2]







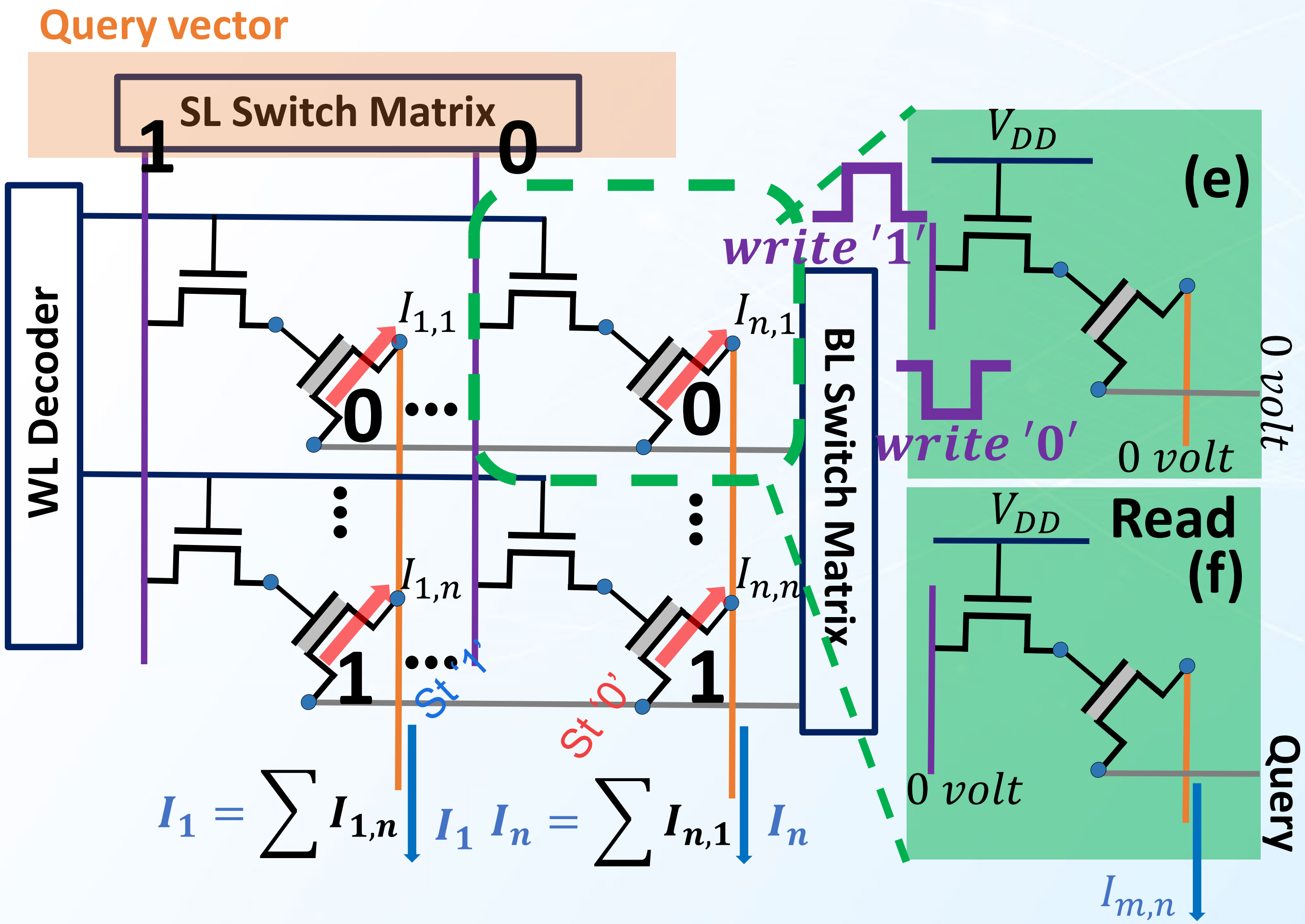
Euclidean CAM!

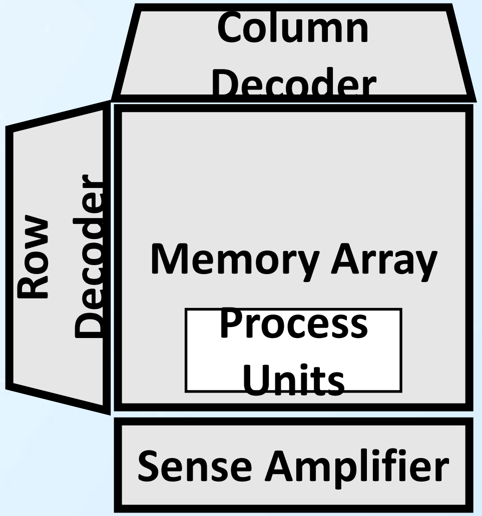


[1] FeFET Multi-Bit Content-Addressable Memories for In-Memory Nearest Neighbor Search, A. Kazemi et al., IEEE TC

[2] Software-equivalent accuracy in-memory hyperdimensional computing with ferroelectric devices, A. Kazemi et al., To appear in Scientific Report

Crossbar



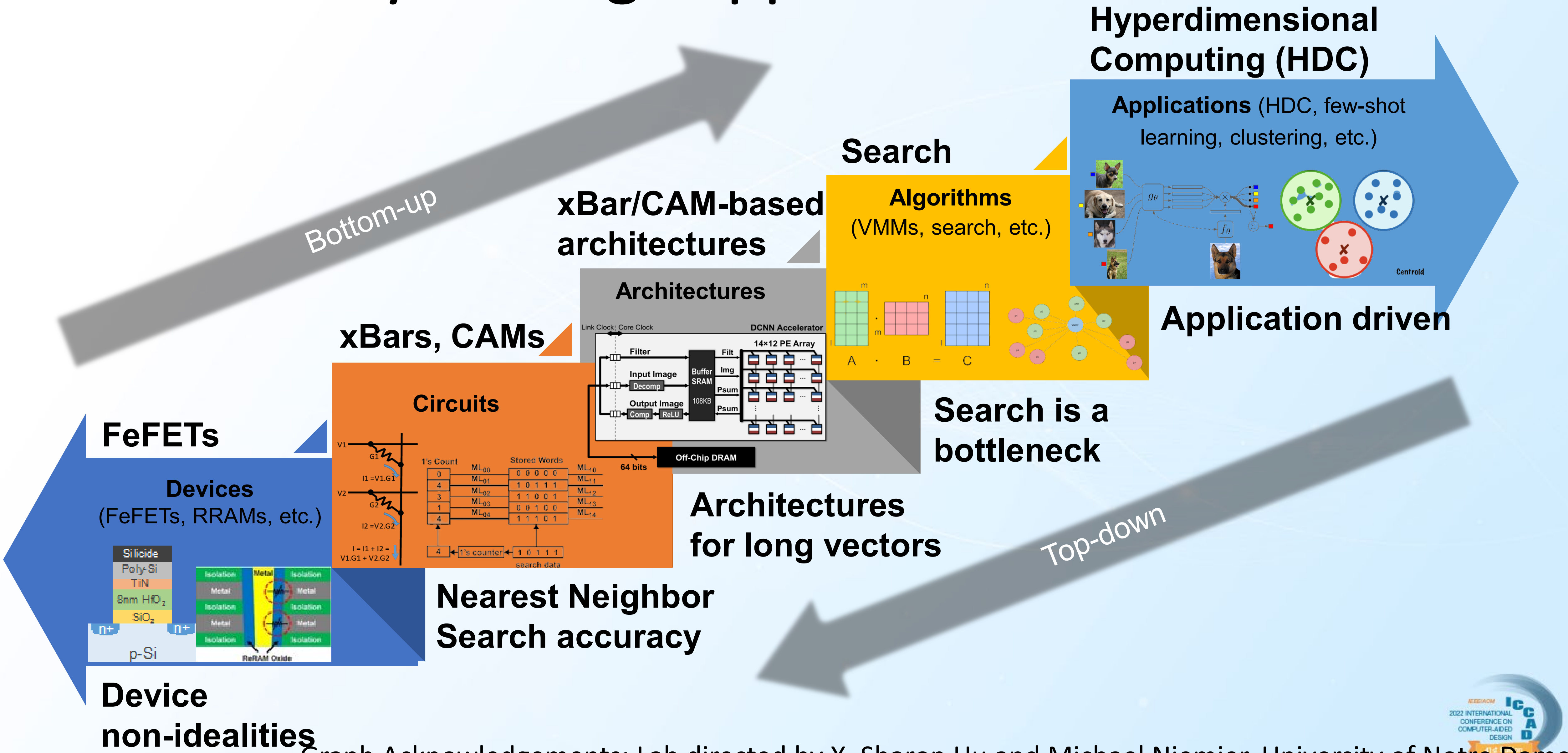


Inner product!

Where is Cosine?

Ferroelectric FET analog synapse for acceleration of deep neural network training, M. Jerry et al., IEDM'17

A cross-layer design approach



Graph Acknowledgements: Lab directed by X. Sharon Hu and Michael Niemier, University of Notre Dame

Possible solutions – AM for cosine similarity

- X-MANN, DAC'19: $\frac{a \cdot b}{||a||_1 + ||b||_1}$

- Nature communication, 2021: Quasi-orthogonal of hyper vectors

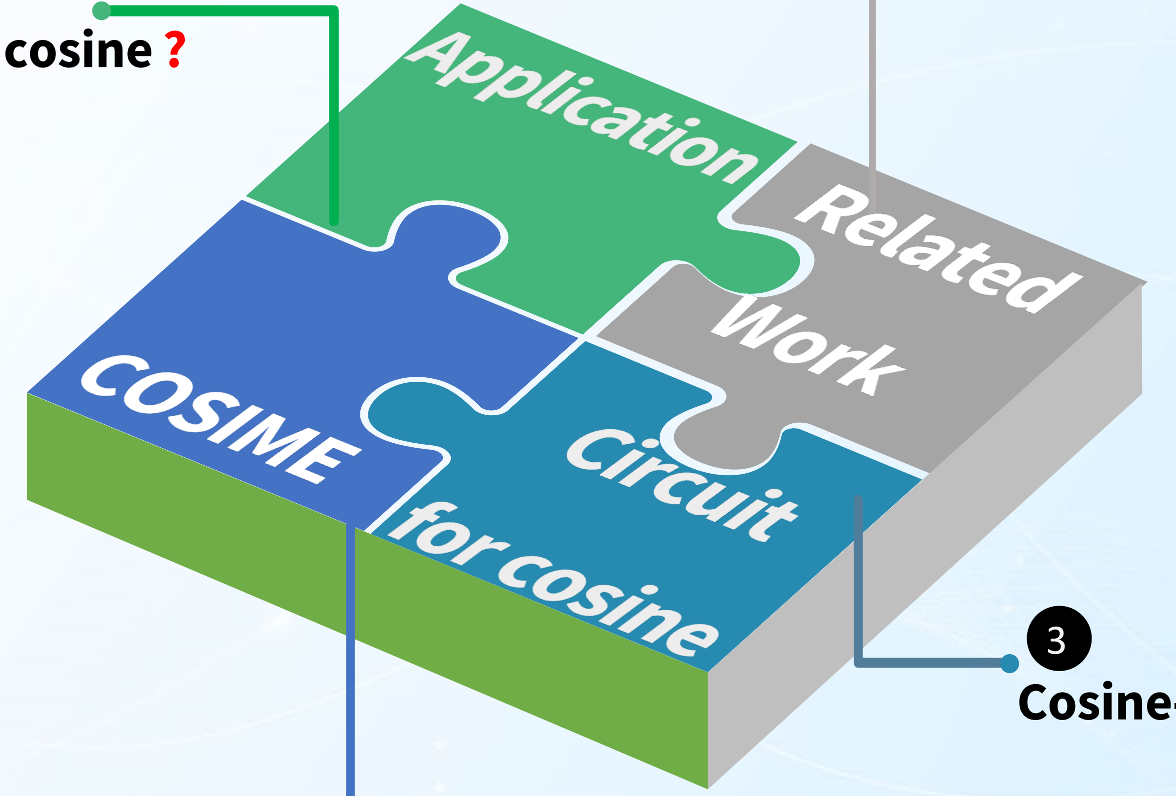
- Normalize the data first then cosine? GPU takes millisecond

- Bipolar vector eliminating division?

1

- CAM-friendly: Hamming √
- Few-shot: L2 on CAM √
- HDC, text-related app: cosine ?

2 **Approximation of cosine (related work)**



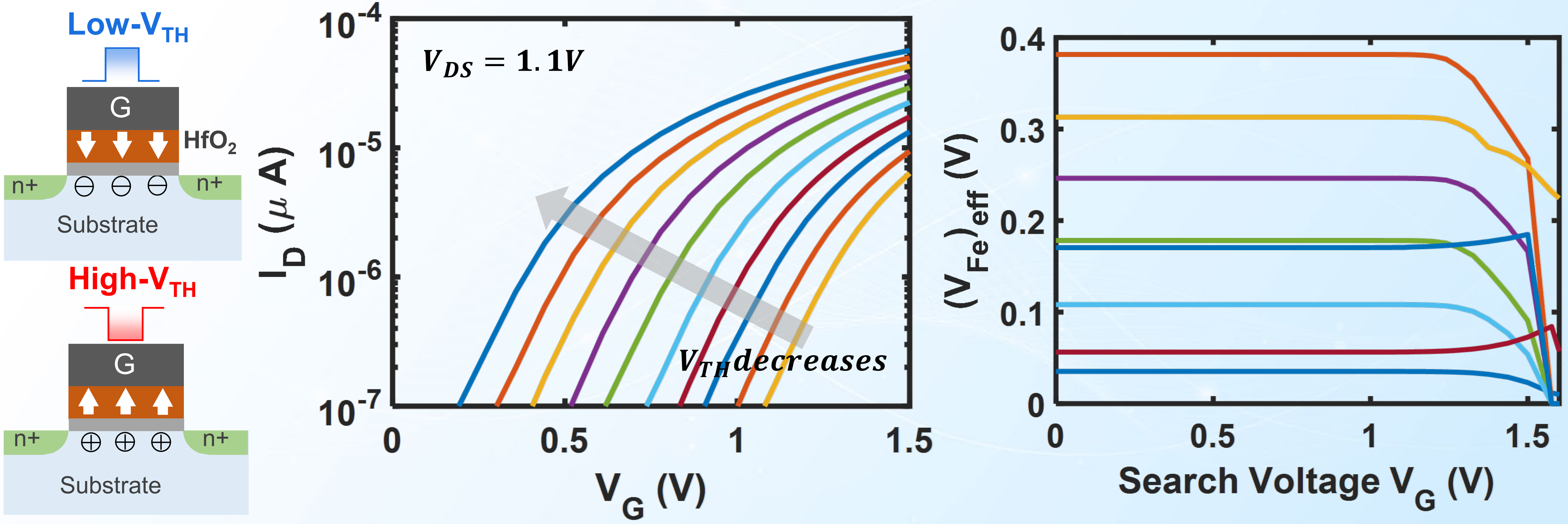
3 **Cosine-friendly analog circuit design**

$$\frac{a \cdot b}{||a||_2 \cdot ||b||_2} \rightarrow \frac{(a \cdot b)^2}{(\cancel{||a||_2} \cdot ||b||_2)^2} \rightarrow \frac{X^2}{Y}$$

L_2 norm of binary vector is counting “1”s

4 **Direct implementation of cosine similarity (this work, COSIME)**

Ferroelectric field-effect transistor (FeFET)



I thank Dr. X. Sharon Hu from the University of Notre Dame for providing this 22nm FeFET model

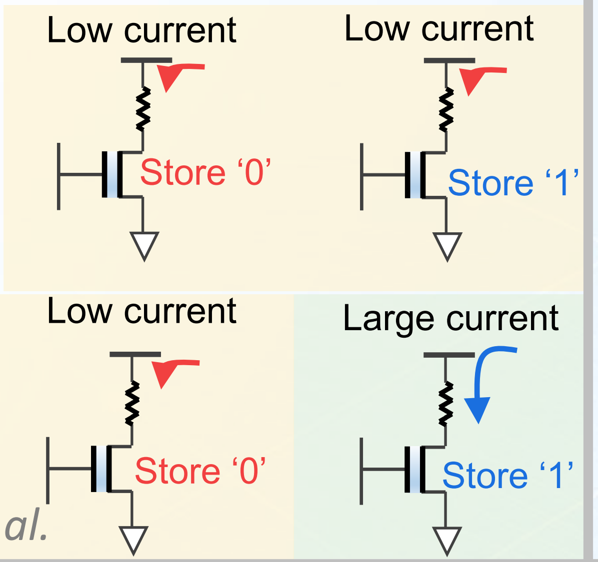
Breaking down the design

1FeFET1R Cell

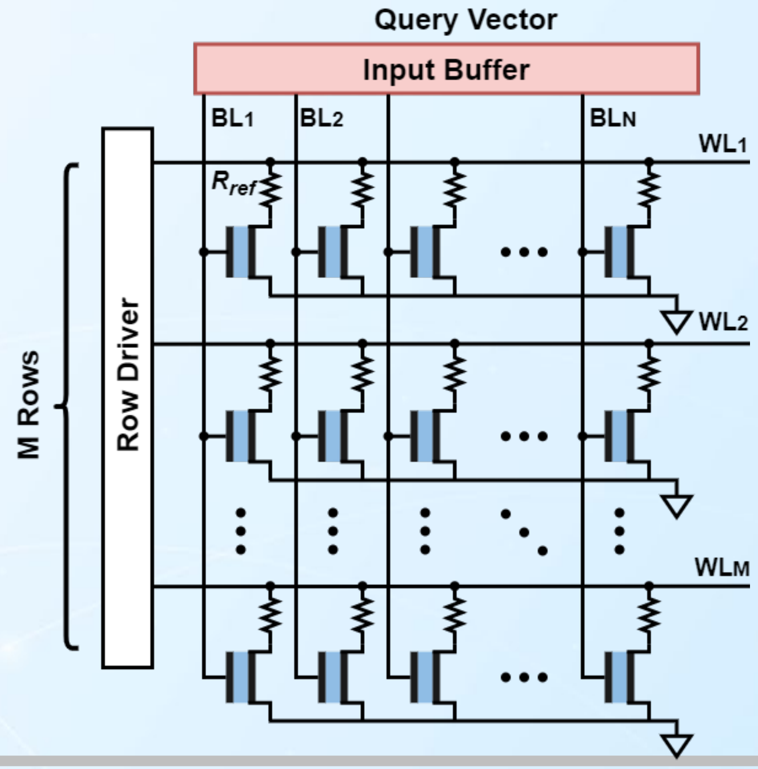
- Mitigate variations
- Scalable array design

$$\frac{(\frac{I_x}{N}\times N)^2}{\frac{I_y}{N}\times N} = I_z$$

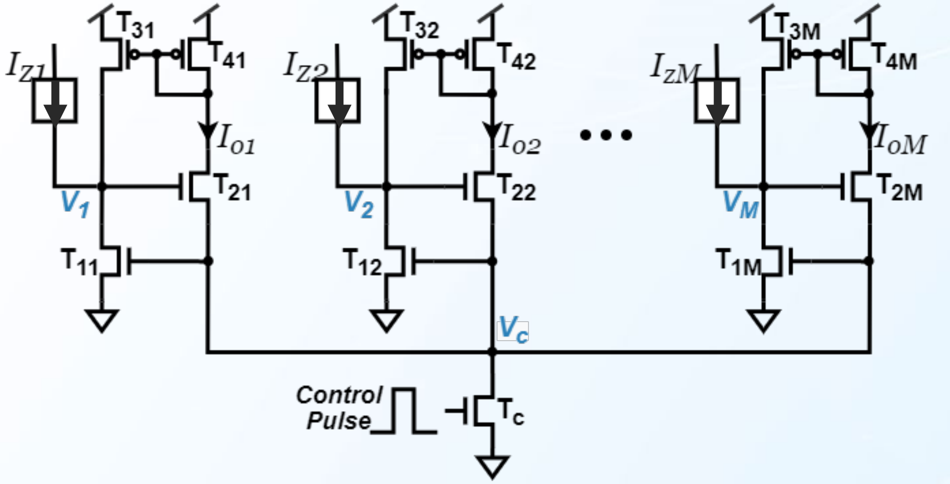
Process is demonstrated in Symp. on VLSI, 2021, D. Saito et al.



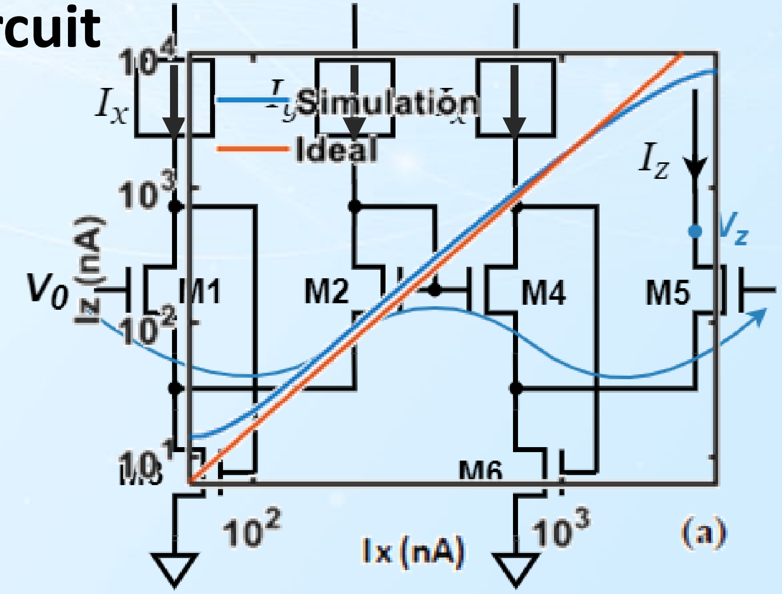
Crossbar



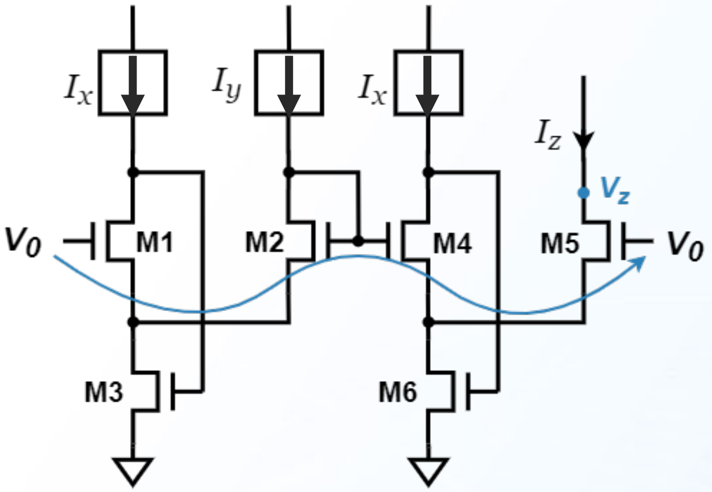
Winner-take-all circuit



$\frac{X^2}{Y}$ circuit



$$\frac{{I_x}^2}{I_y} = I_z$$



1

$$I_{DS} \approx I_0 \frac{W}{L} e^{\frac{V_{GS}}{\eta T}}$$

2

$$V_{GS} = V_T \eta \ln\left(\frac{I_{DS} L}{I_0 W}\right)$$

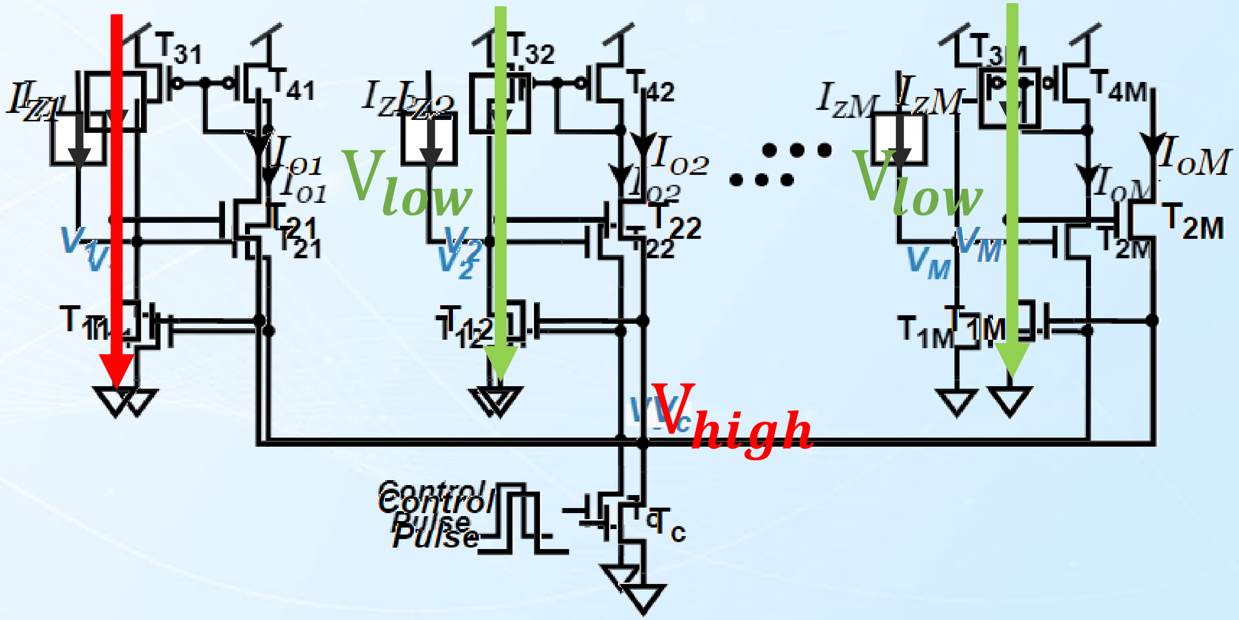
3

$$\sum_{ClockWise} V_{GS} = \sum_{counter\ CW} V_{GS}$$

(M1, M2, M4, M5)

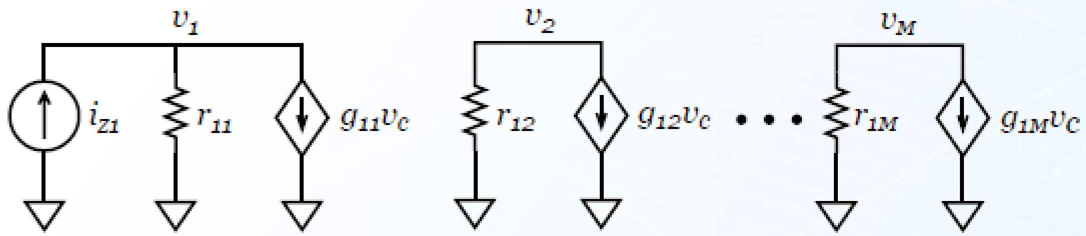
4

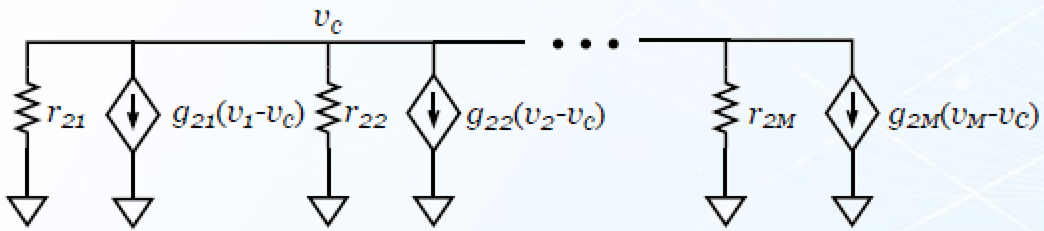
$$\frac{{I_x}^2}{I_y} = I_z$$

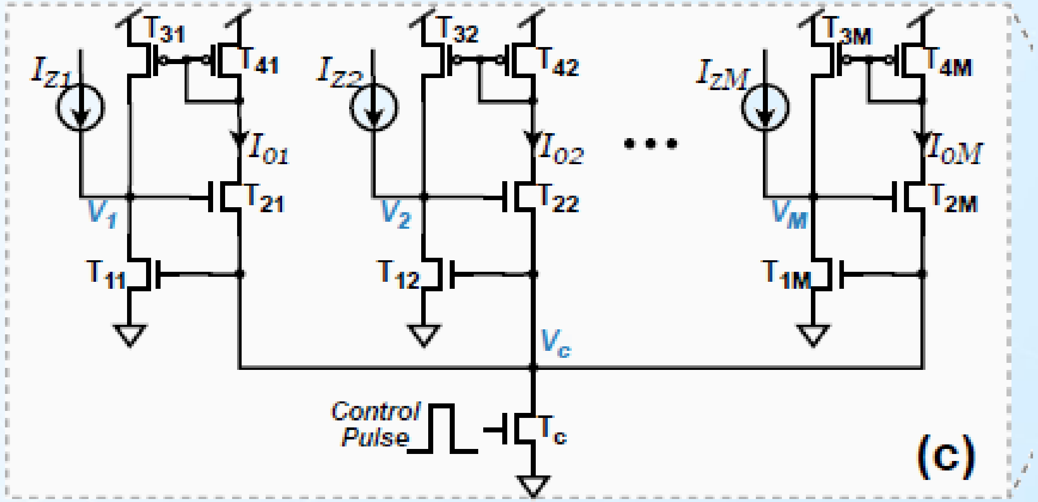


Scalability Analysis

$g_{11} = \frac{I_{Z1}}{V_T}$ and $r_{11} = \frac{V_A}{I_{Z1}}$ etc.







$$
\begin{cases}
i_{z1} = v_1 \dfrac{I_{z1}}{V_A} + v_c \dfrac{I_{z1}}{V_T} \\
v_j \dfrac{I_{zj}}{V_A} = -v_c \dfrac{I_{zj}}{V_T}, \quad \forall j \in [2, M] \\
\displaystyle\sum_{i=1}^{M} [\frac{I_{oi}}{V_T}(v_i - v_c) + v_c \frac{I_{oi}}{V_A}]
\end{cases}
$$

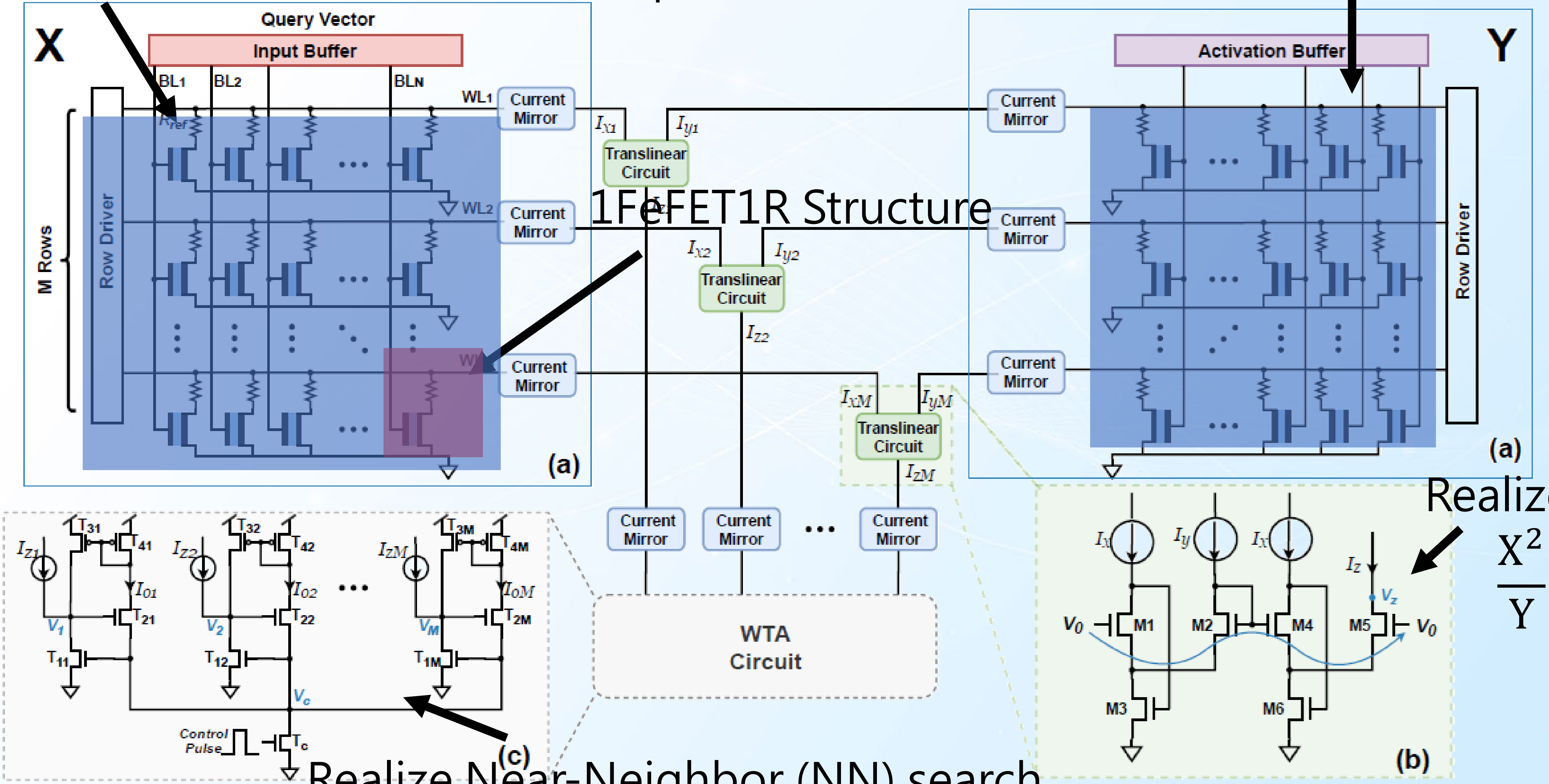
$$\frac{dV_1}{dI_{z1}} = \frac{M-1}{M}\frac{V_A}{I_{z1}}$$

$$\frac{dV_2}{dI_{z1}} = \frac{-1}{M}\frac{V_A}{I_{z1}}$$

Overall Design of "COSIME"

For cosine numerator, i.e dot product

For cosine denominator, i.e. L_2 $norm$



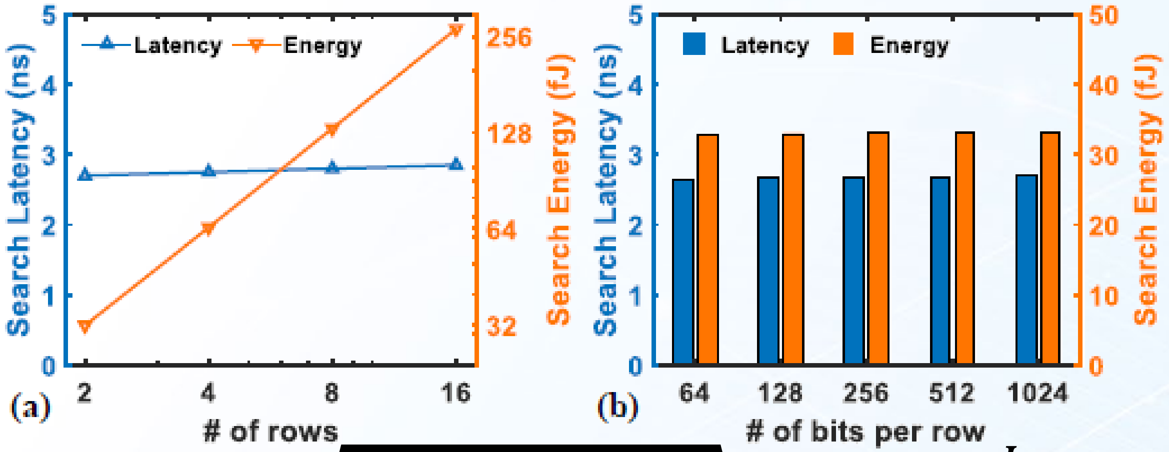
1FeFET1R Structure

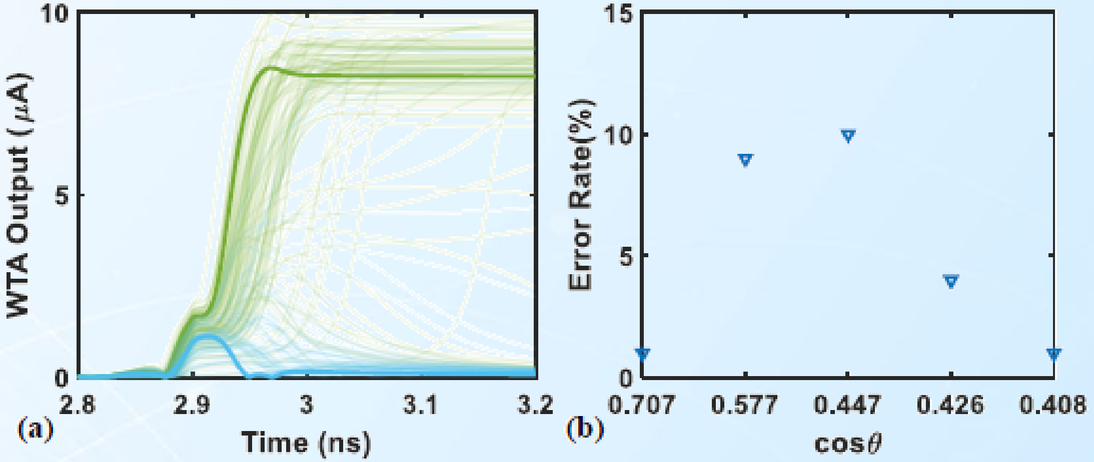
Realizes cos^2:

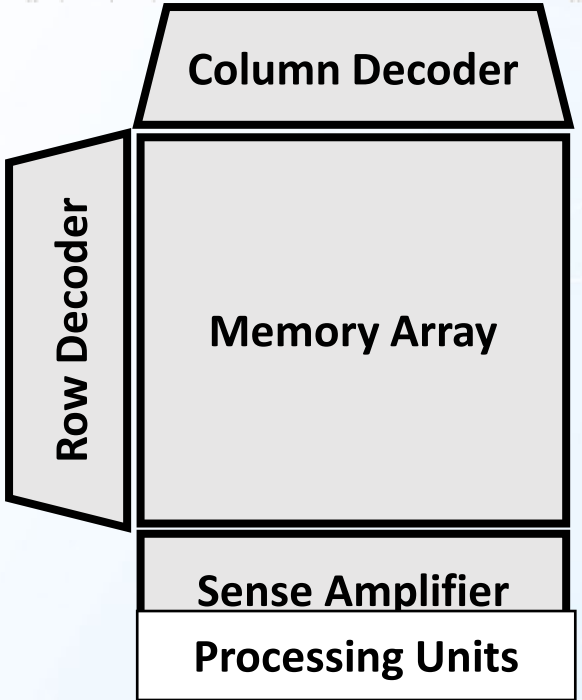
$$\frac{X^2}{Y} = Z$$

Realize Near-Neighbor (NN) search

Array-level Evaluation



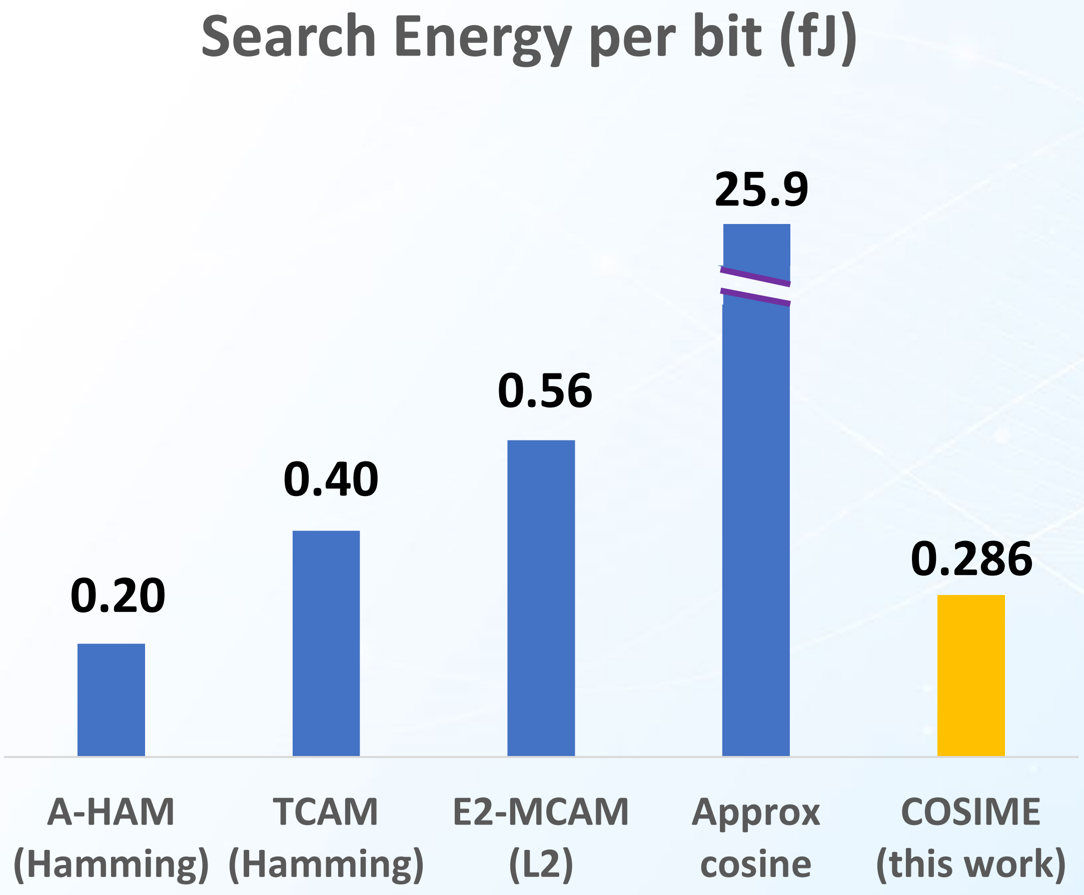


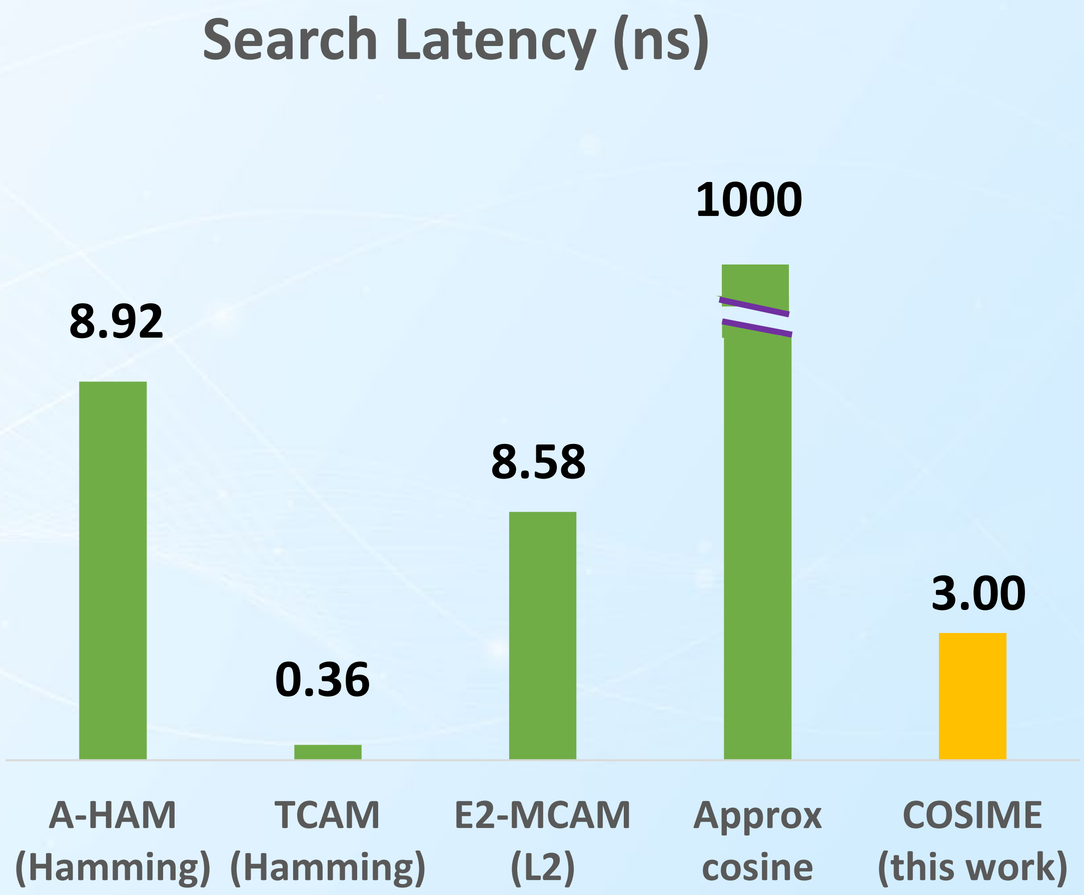


$$\frac{\left(\frac{I_x}{N}\times N\right)^2}{\frac{I_y}{N}\times N} = I_z$$

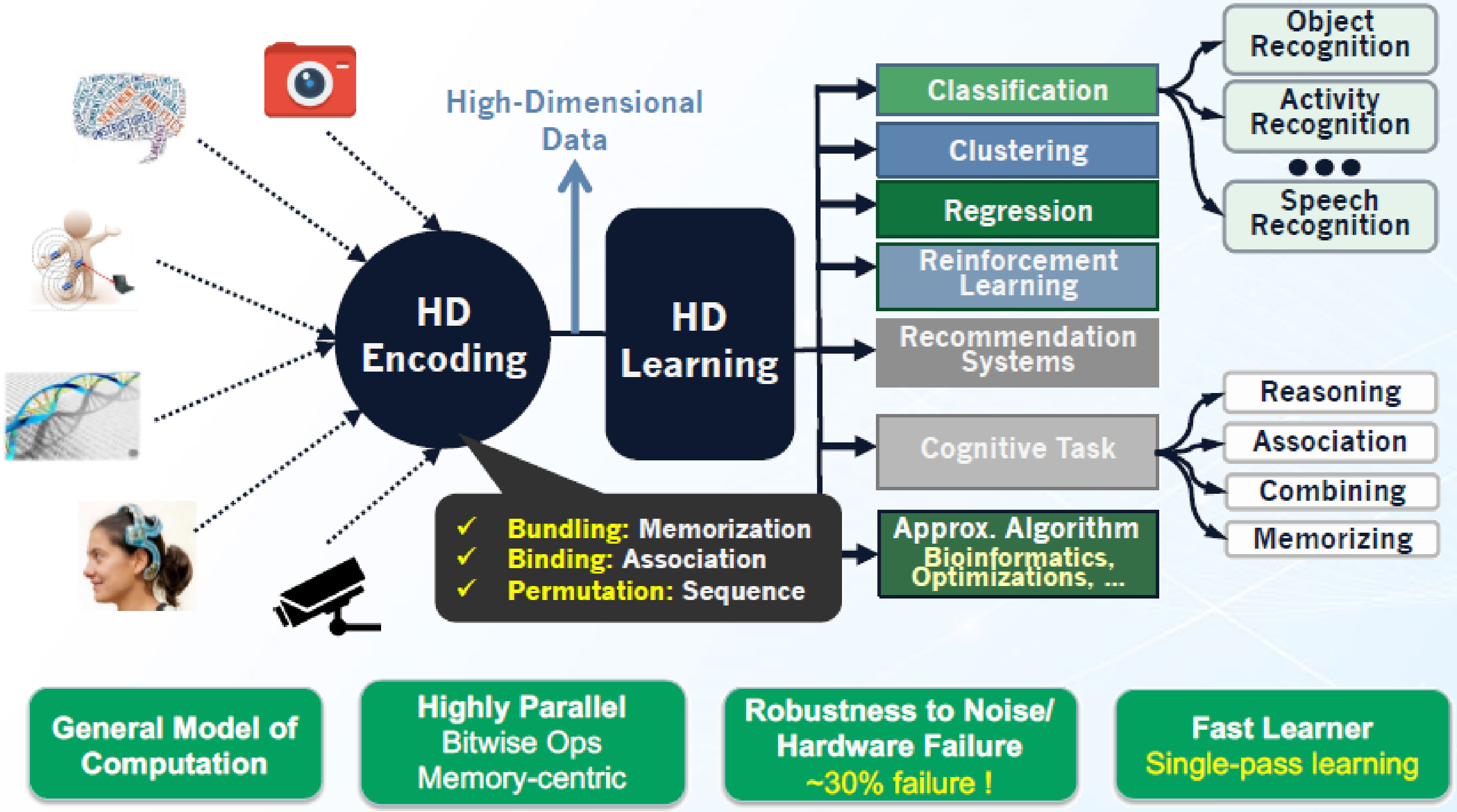
Variation including:
VDD, FeFET, Resistor in the cell, MOSFET

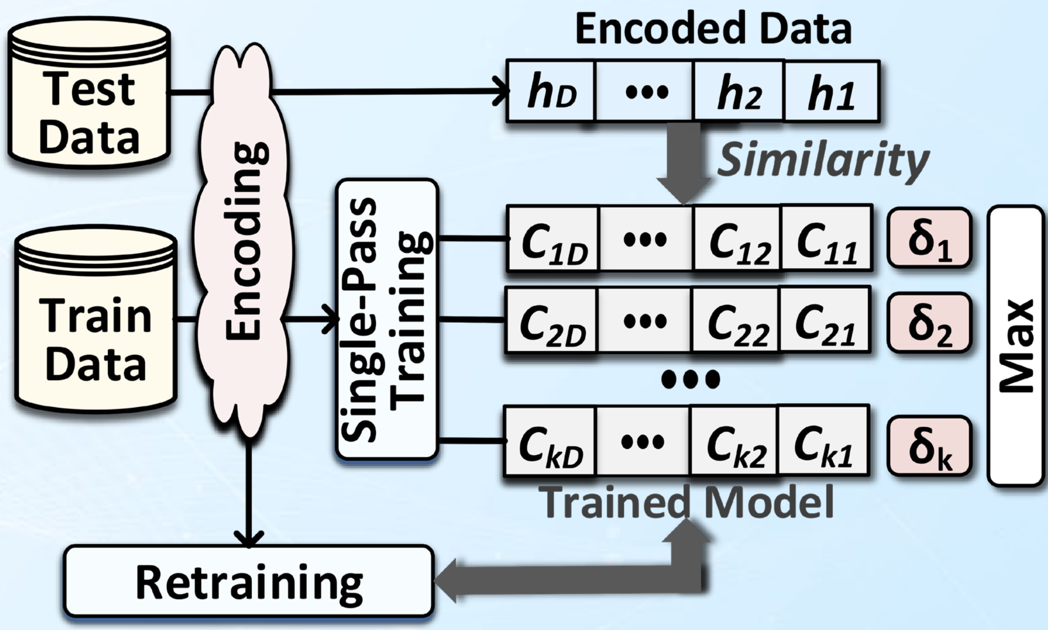
Existing Associative Memories





HDC Basics





- Limitation: Complex tasks such as Cifar-10 → 60%
- Robustness: What if Fp32 MSB get flipped in NN?

Detail operations in Python

- SOTA encoding: Inspired by kernel method [1].
  - $\cos\left(\vec{F} \cdot \vec{B} + \vec{b}\right) \sin(\vec{F} \cdot \vec{B})$
- Single-pass training:

$$\vec{C}_l = \sum \vec{H}_l$$

- Iterative training: Inspired by Perceptron learning algorithm (PLA).
  - MNIST: single-pass 86% Iterative: 93%

$$\vec{C}_l \leftarrow \vec{C}_l + \eta(1-\delta)\vec{Q}$$

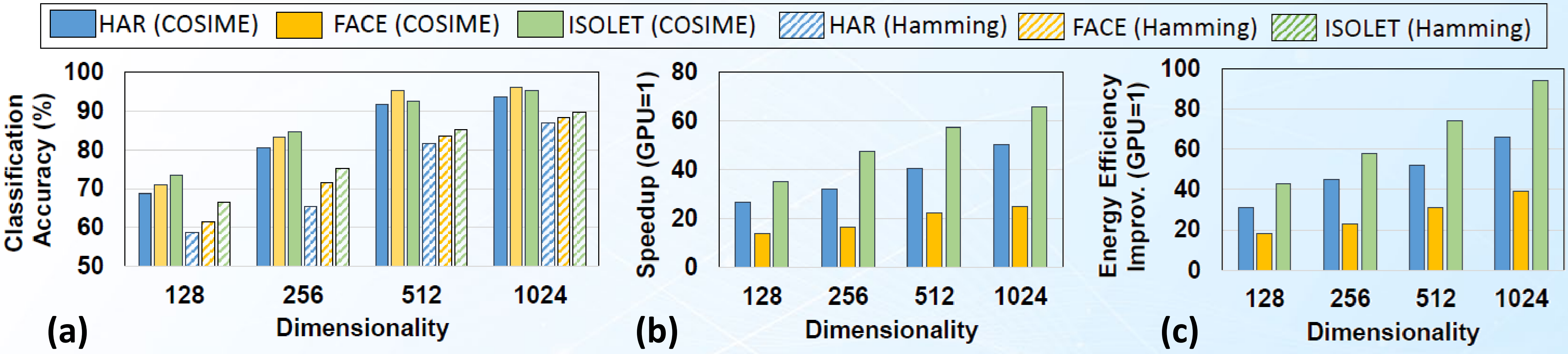
$$\vec{C}_{l'} \leftarrow \vec{C}_{l'} - \eta(1-\delta)\vec{Q}$$

- Inference

[1] Random features for large-scale kernel machines, A. Rahimi and B. Recht, NIPS 2007

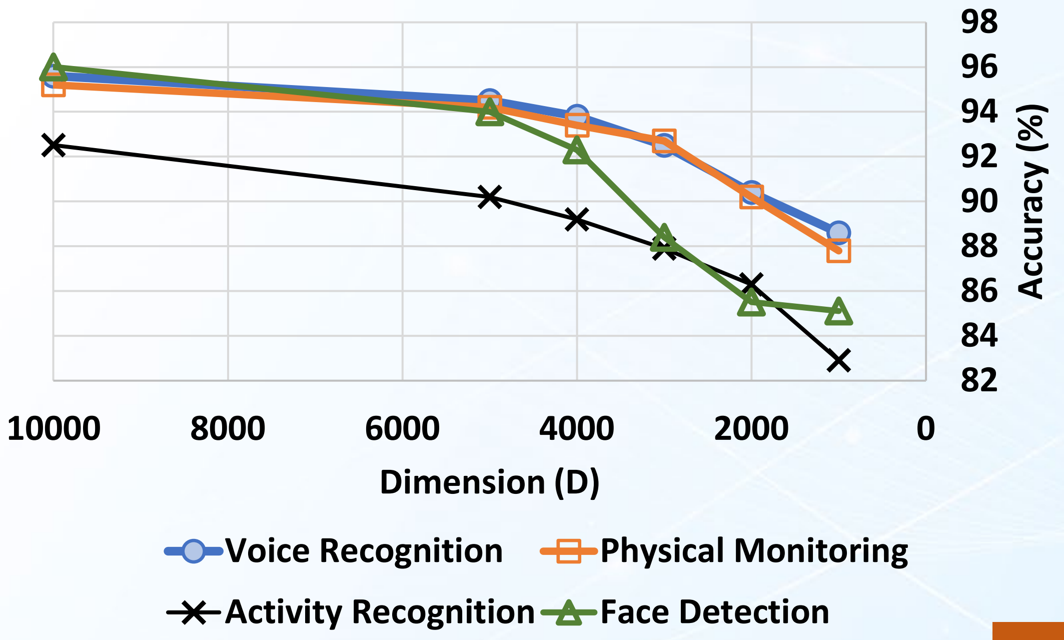
Benchmarking with HDC

	n	K	Train Size	Test Size	Description
UCIHAR	561	12	6,213	1,554	Activity Recognition[40]
FACE	608	2	522,441	2,494	Face Recognition[41]
ISOLET	617	26	6,238	1,559	Voice Recognition [42]



- On average 7% software acc improvement than Hamming distance.
- 47.1x faster/98.5x energy saving compared to 1080 GPU.
- Not so impressive? Because the dimension is low due to hardware.

A view from top-down



HD Computing loses accuracy on lower dimensionality

Scalable Edge-Based Hyperdimensional Learning System with Brain-Like Neural Adaptation. Zhuowen Zou et. al., SC 2021

Challenges and Research Opportunities

- Right memory hierarchy.
- Multibit associative memory.
- Low conducting current. Redesign the translinear operating region.
- Peripherals for NVMs

Conclusion

- A non-approximated design of cosine similarity search in-memory is presented.
- The evaluation and benchmarking results at array and application level show the robustness and superior accuracy to Hamming distance implementation.
- The proposed COSIME design is not limited to FeFET technology, but is rather general.

Thank you for your attention!

cliu29@nd.edu