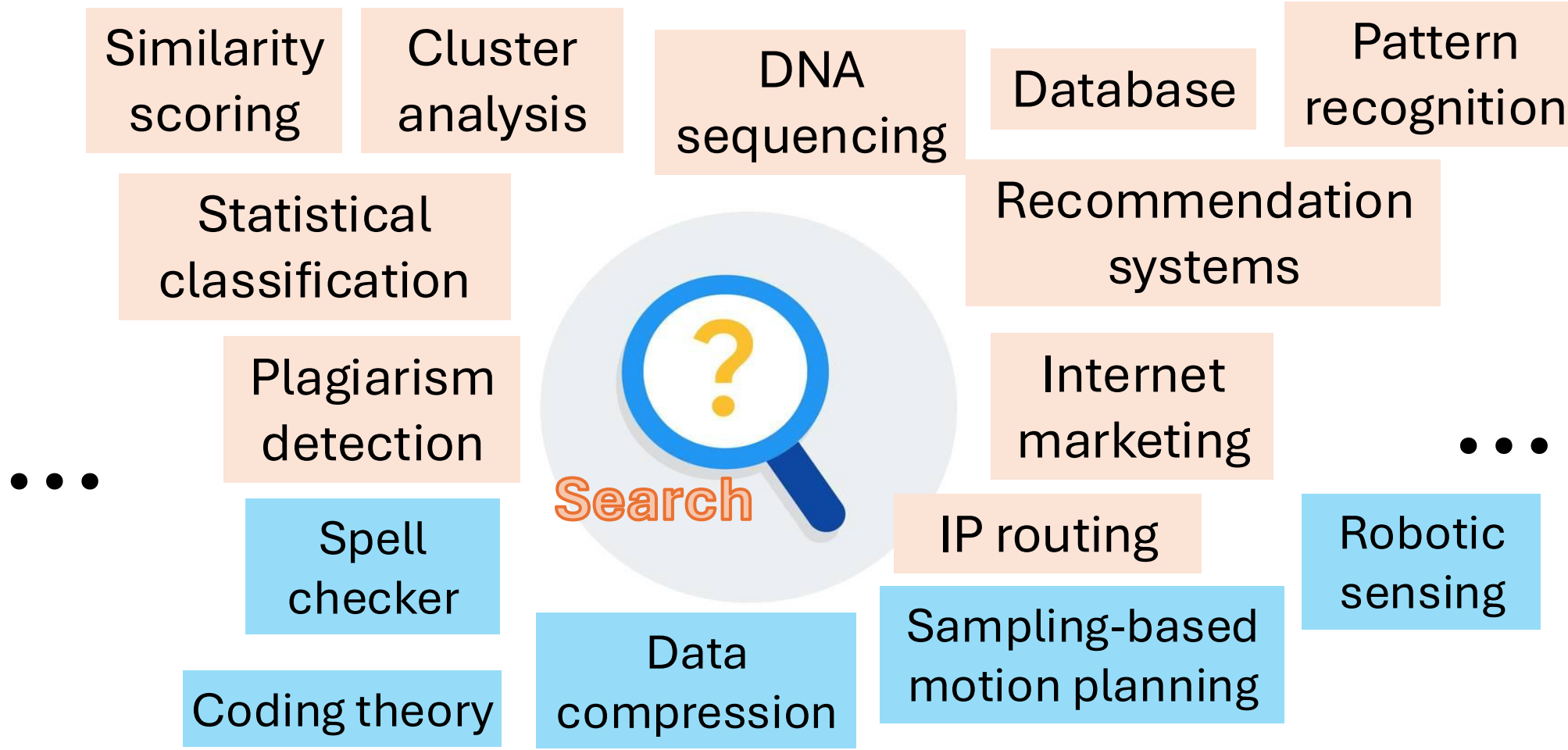
Challenges and Opportunities in Accelerating Large-Scale Search Using NVM-based In-Memory Computing

X. Sharon Hu (shu@nd.edu)

Department of Computer Science and Engineering

University of Notre Dame

Search as an important class of computation kernels



Growing use of associative search in AI/ML workloads

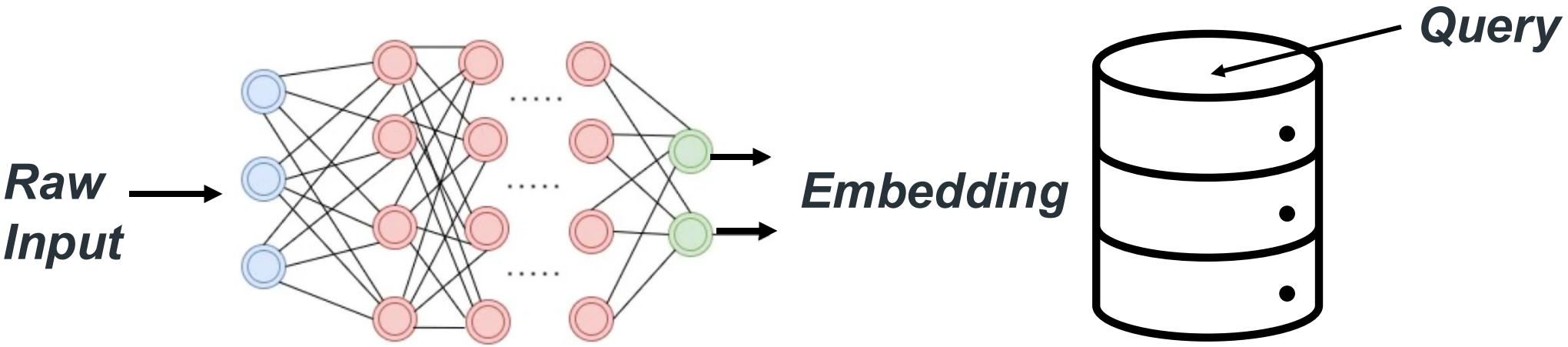
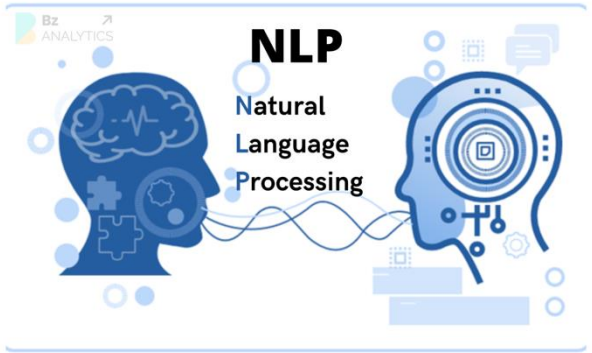
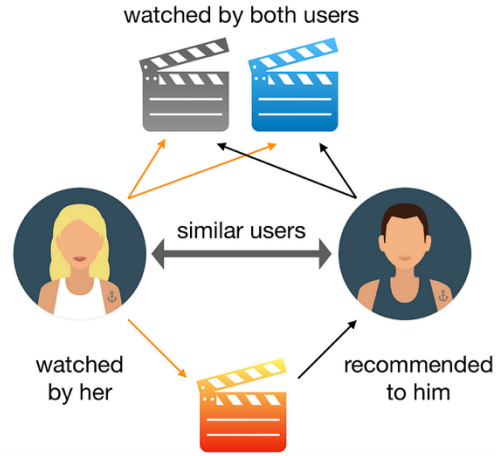


Image processing

Semantic search

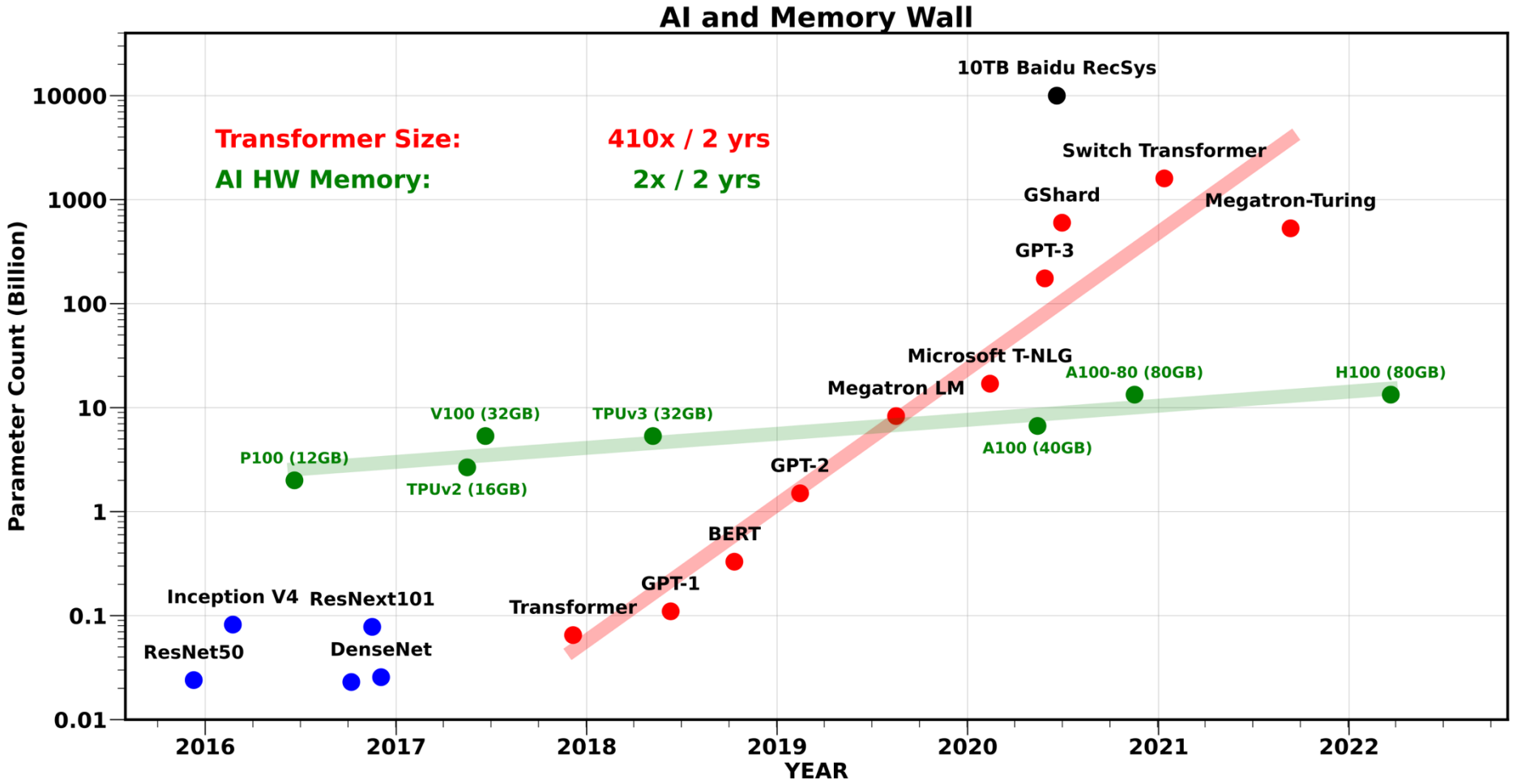


Natural language processing



Content-based recommendation systems

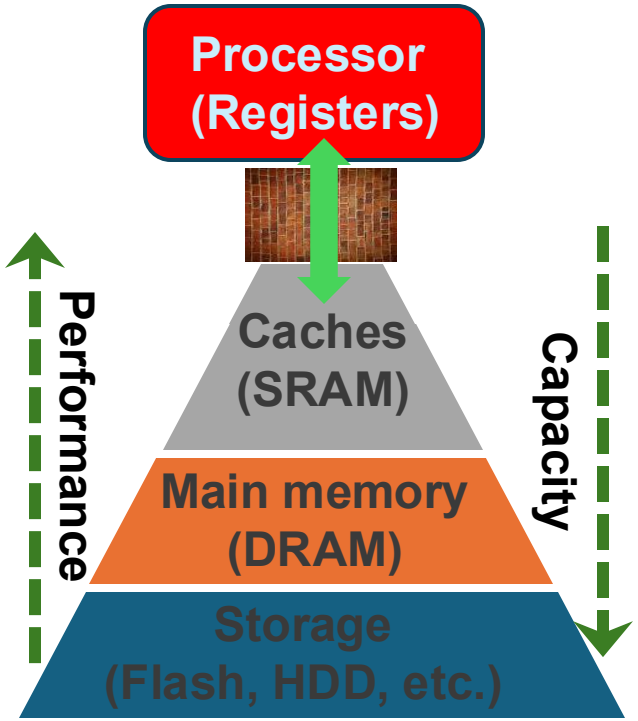
AI and memory wall



'The evolution of the number of parameters (410x) of SOTA AI models over the years, along with the AI accelerator memory capacity (2x)'

A. Gholami, et.al, IEEE Micro Journal, 2024

Memory wall in conventional computing platforms



- **Energy**
- **Latency/ throughput**
- **Security**
- **…**

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

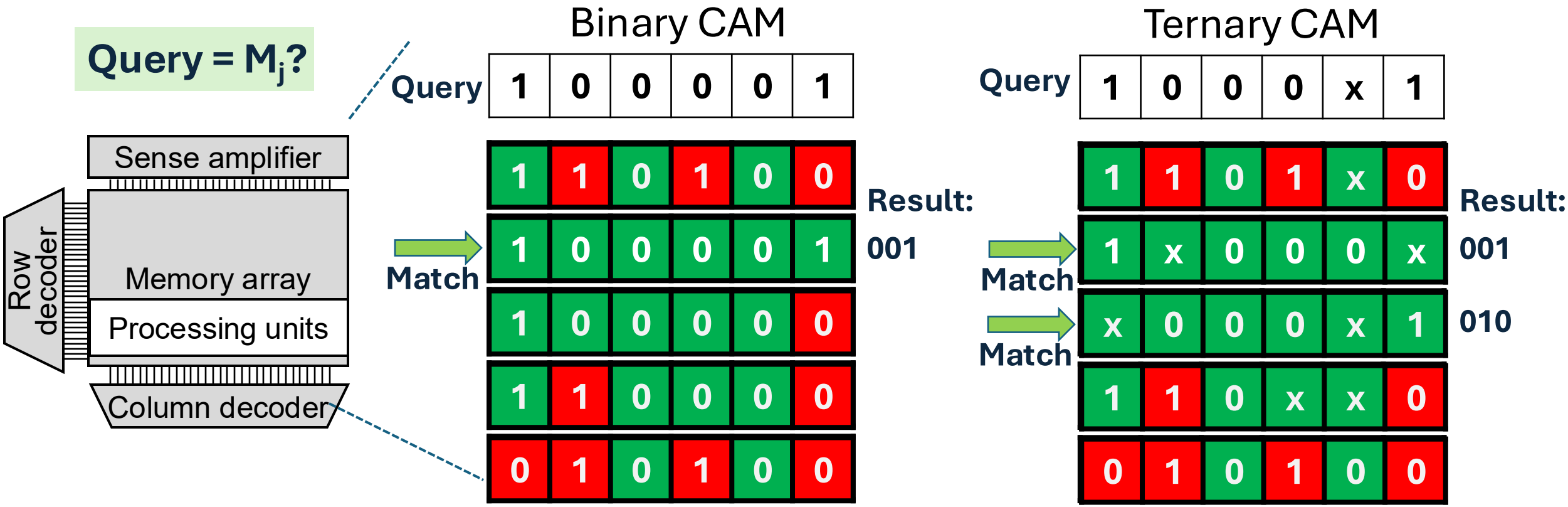
M. Horowitz. Energy table for 45nm process, Stanford VLSI wiki. : https://sites.google.com/site/seecproject

Reduce memory transfer overhead is critical!

In-memory search via content addressable memory

In-memory search: Content-addressable memory (CAM)

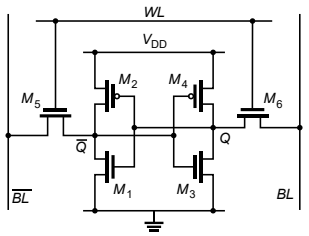
- **CAM => Associative memory (AM)**
  - Support parallel search
  - Widely used in high-associativity cache, network router, etc.
  - **Growing use in data analytics, machine learning, vector processing**



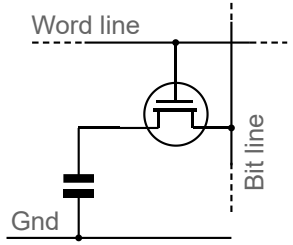
Choices of memory device technologies for CAMs

Memory Technologies

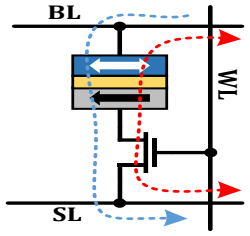
SRAM



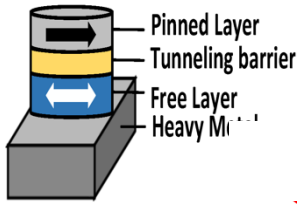
DRAM



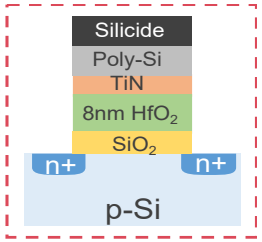
STT-MRAM



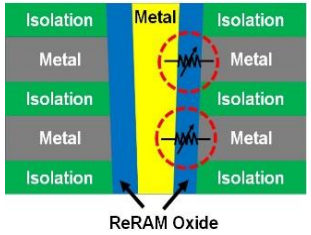
SOT-MRAM



FeFETs



RRAM



...

What memory devices should be used to construct CAM cells?

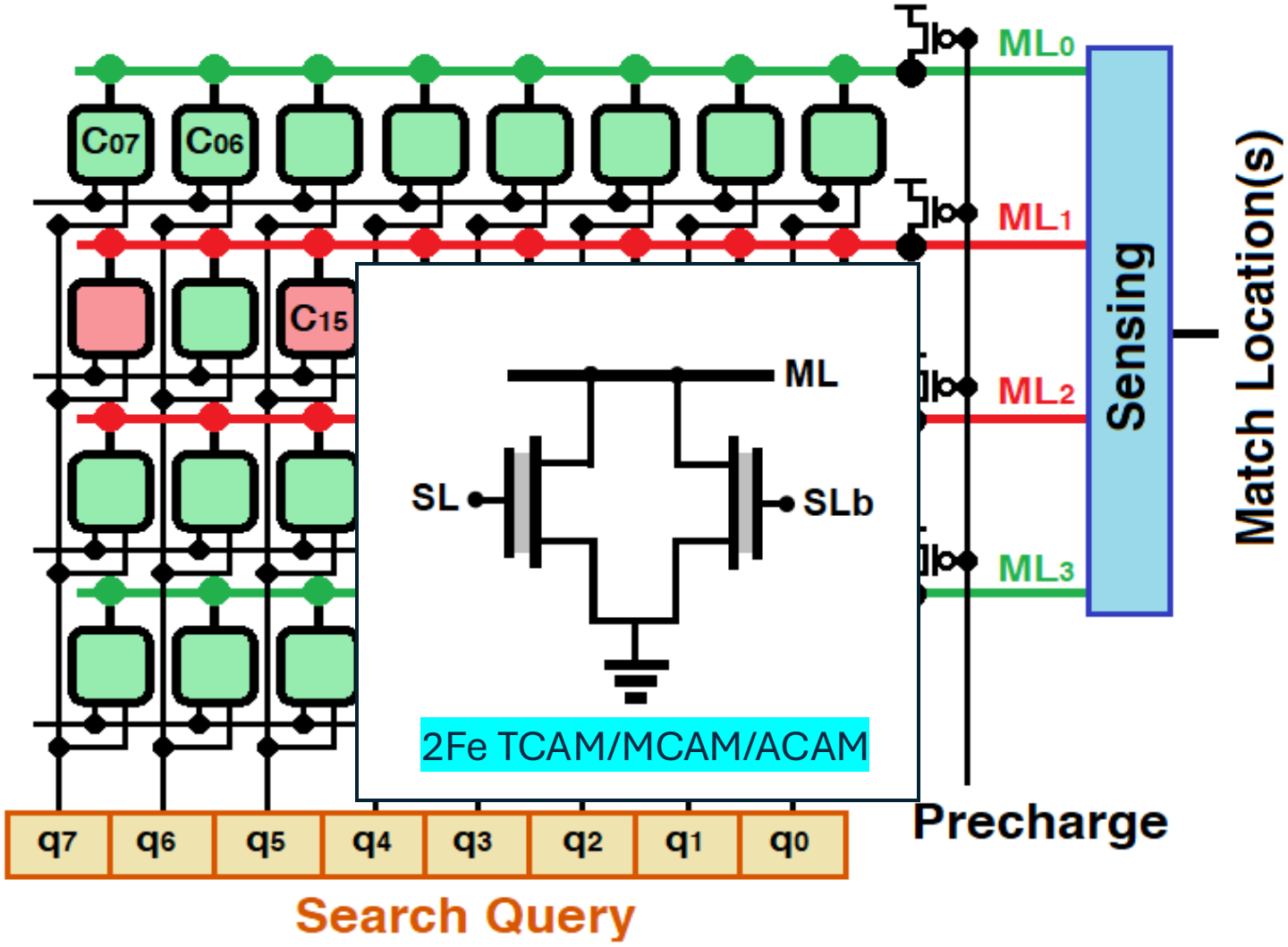
Desired properties:

- Large memory windows
- Multi-level data
- Tunable resistance levels
- Robustness
- Endurance
- Retention
- Speed, power, area
- ...

NVM opens doors for designing more sophisticated CAM functions

CAM data representations

Generic CAM array structure



Binary CAM (BCAM) 0 1

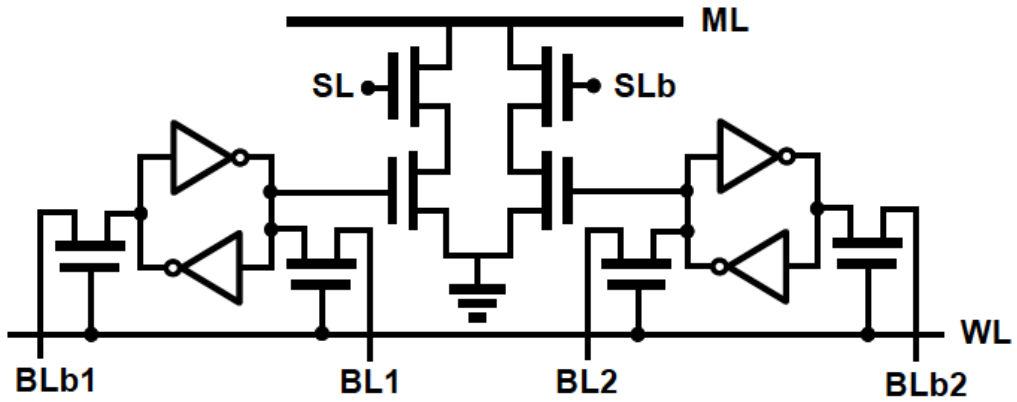
Ternary CAM (TCAM) 0 1 X

Multi-bit CAM (MCAM) S0 S1 S2 S3

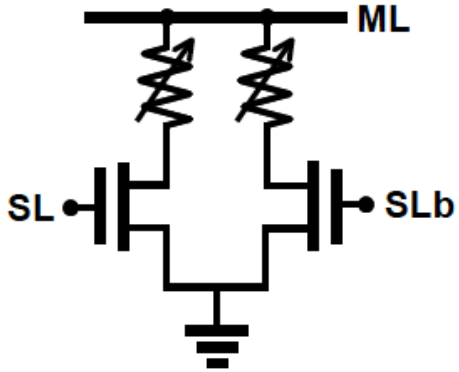
Analog CAM (ACAM)

X. Sharon Hu, et al., IEDM, 2022
Xunzhao Yin, IEEE TED, 2020

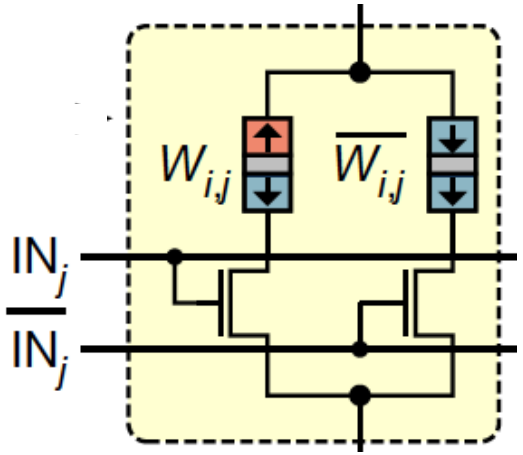
CAMs based on different memory technologies



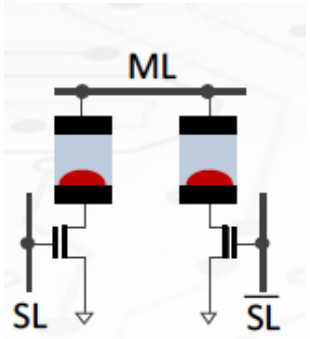
16T CMOS TCAM



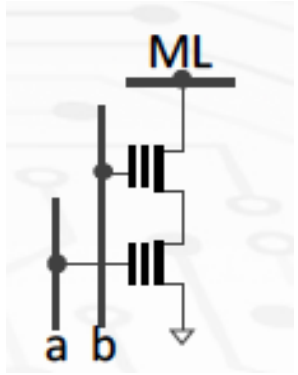
2T2R TCAM [1]



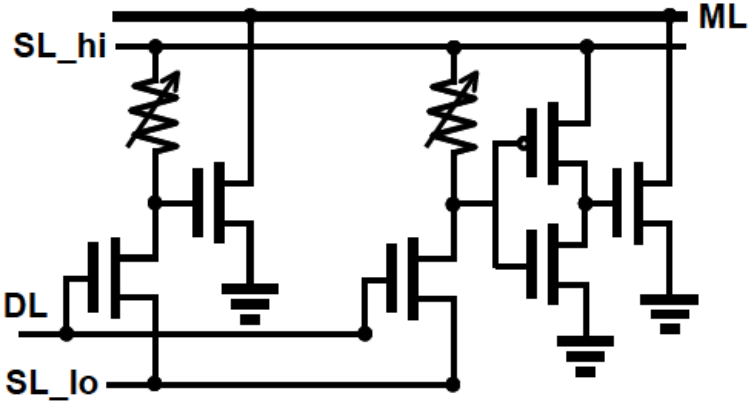
2T2MTJ TCAM [2]



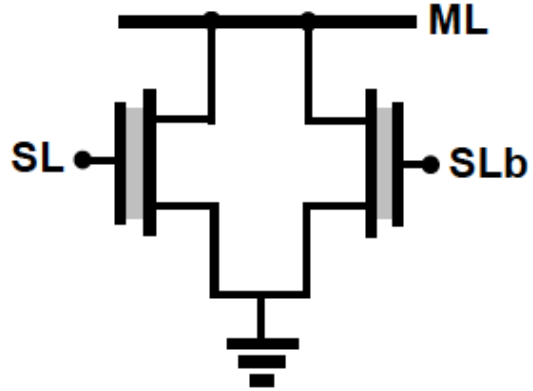
2T2PCM TCAM [3]



2Flash TCAM/MCAM [4,5]



6T2R ACAM [6]

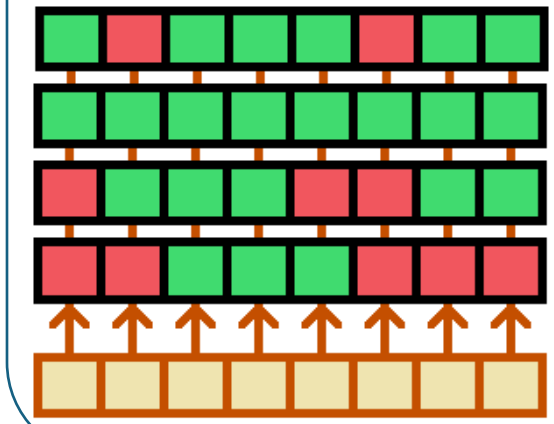


2Fe TCAM/MCAM/ACAM [7]

[1] Haitong Li, et al., IEEE TED, 2021. [2] Seungchul Jung, et al., Nature, 2022. [3] Jing Li, et al.,IEEE JSSC 2013. [4] Viacheslav V. Fedorov, et al., IEEE TC, 2016. [5] Arman Kazemi, et al., IEEE ISLEP, 2021. [6] Giacomo Pedretti, et al., Nat. Comm, 2021. [7] Xunzhao Yin et al., IEEE TED, 2020

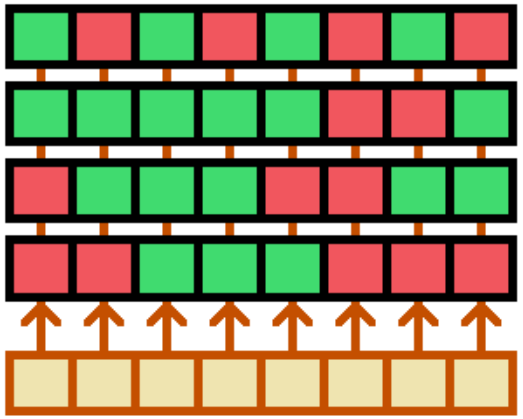
Different types of match for T/MCAM

Exact (EX) Match



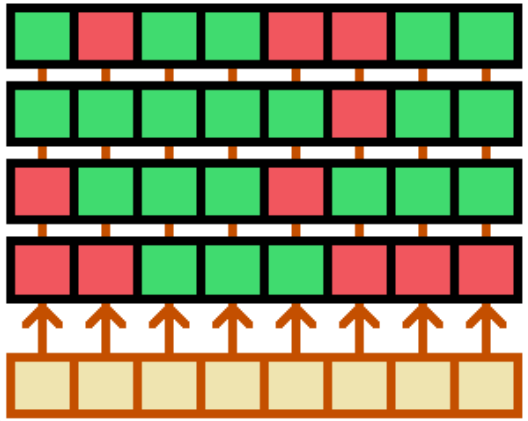
Mismatch
Match
Mismatch
Mismatch

Best (BE) Match



Mismatch
Match
Mismatch
Mismatch

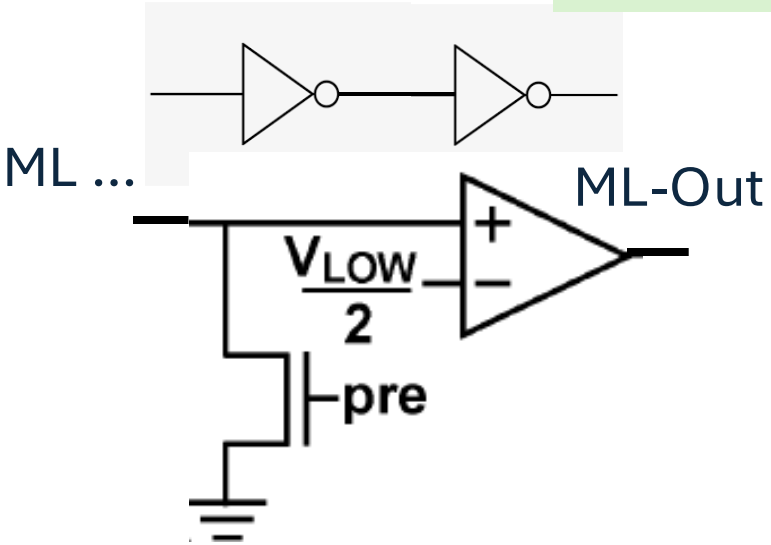
Threshold (TH) Match



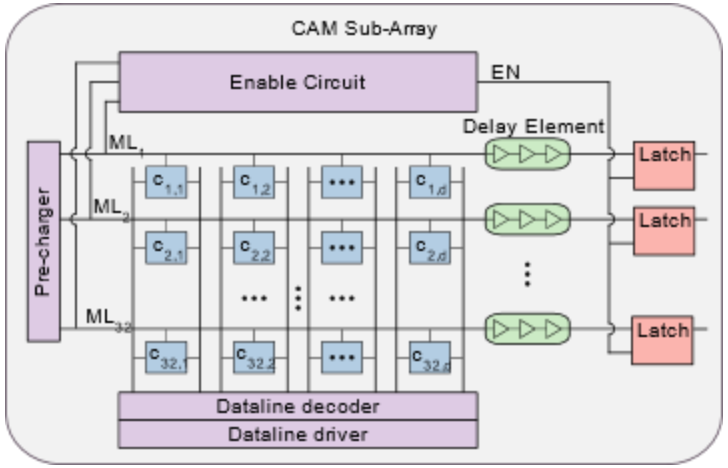
Mismatch
Match
Match
Mismatch

Thd =2

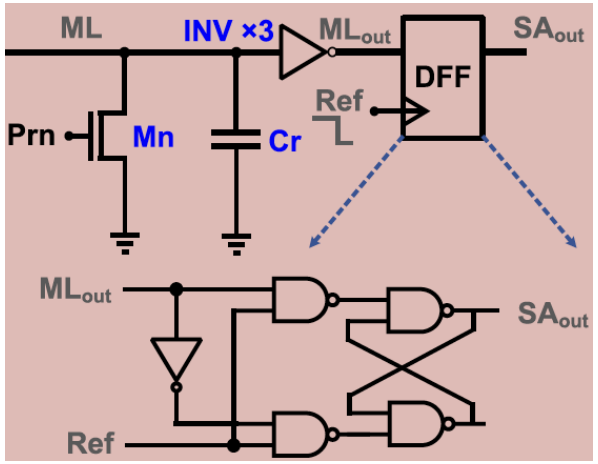
Same CAM cell design but different sensing circuits



Kasai, et al., CICC 2003



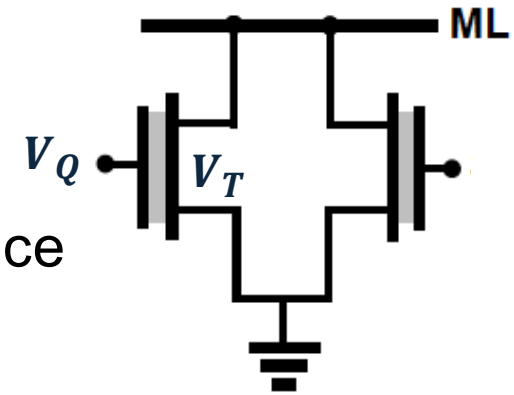
Arman Kazemi, et al., Sci. Report, 2022



Liu Liu, et al., TCAS-1, 2023

2-FeFET MCAM design implement other distances

- **Conventonal B/TCAM uses Hamming distance; can other distances be implemented?**
  - Yes!
- **V_T and V_Q assignment problem:**
  - Find the optimal V_T and V_Q for each state
  - Minimize the error between each CAM distance and the desired distance
  - Define necessary constraints



CAM distance desired distance

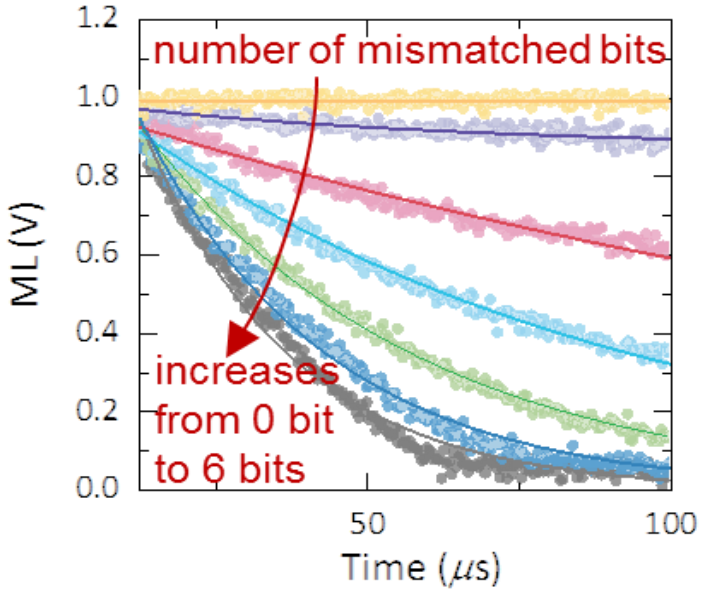
$$\min_{V_Q, V_T, \alpha} I(V_{Qi}, V_{Tj}) - \alpha \times d_{ij} - \beta$$

Subject to $0 \le V_{Qi} \le V_{DD}, V_{Tmin} \le V_{Tj} \le V_{Tmax}, V_{Ti} \le V_{Qi}, \forall i, j$

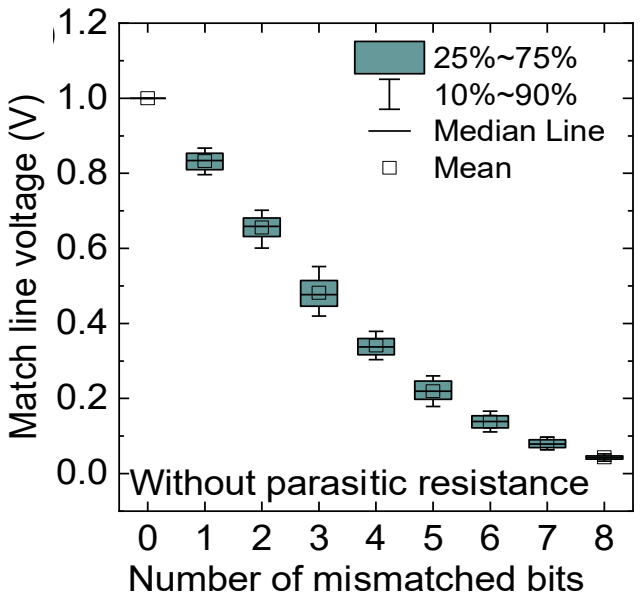
CAM for complex and large-scale search

How big can a single CAM array be?

❑ **Sub-array size, particularly # of columns, cannot be too big!**



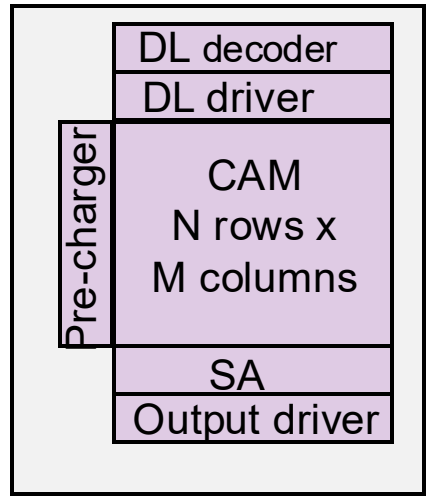
Maximum detectable distance



6-to-8 bit mismatch detectable … in 64 column array

Kai Ni, et al., Nature Electronics, 2020.

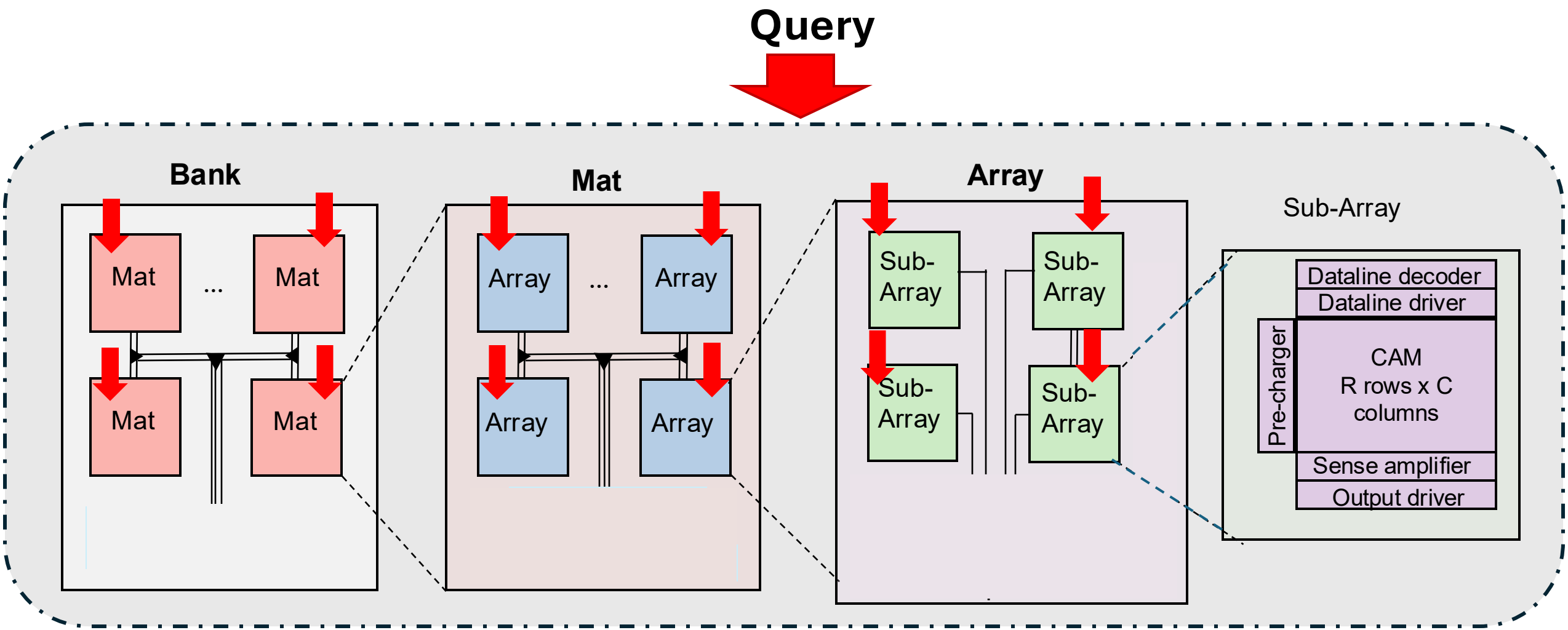
Sub-Array



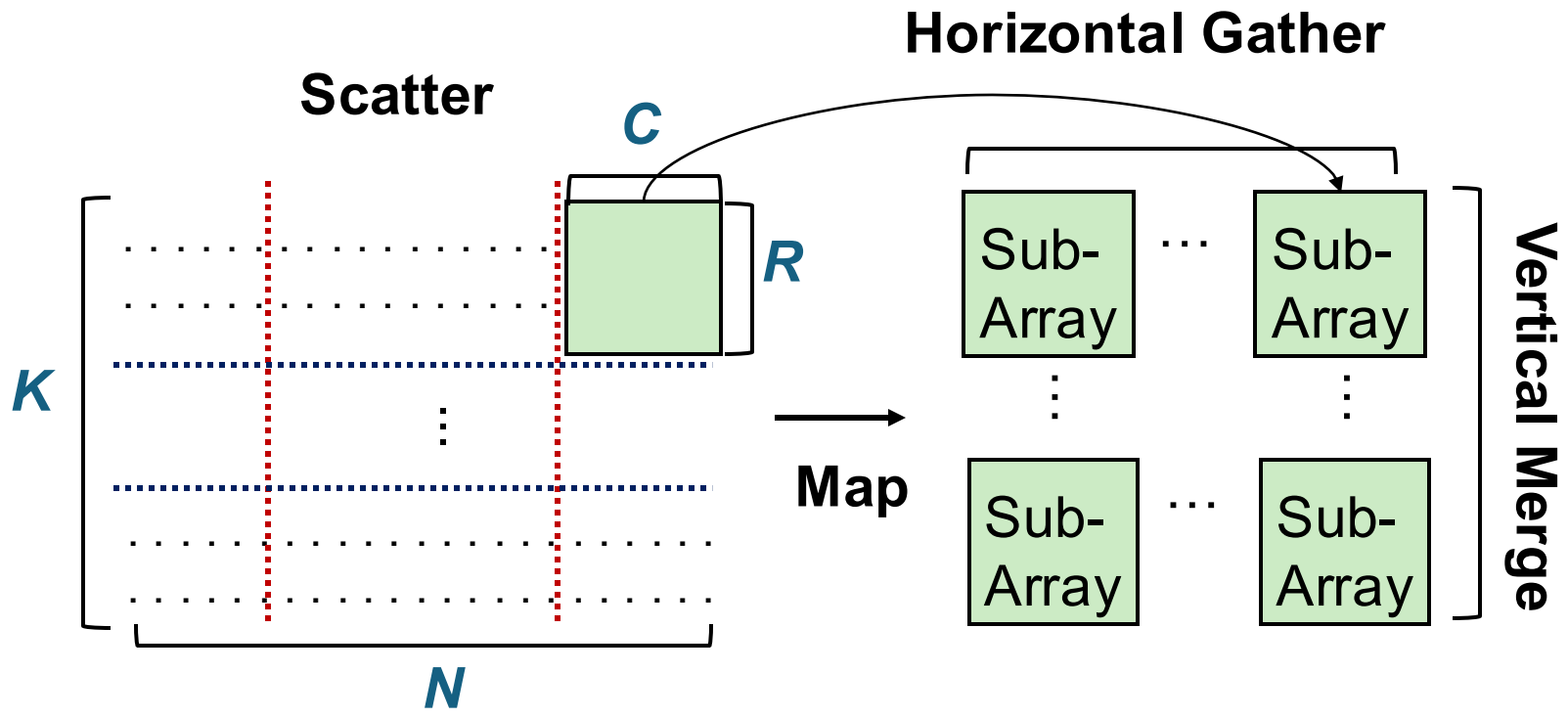
❑ **How to support high dimensional vectors with small CAM subarrays?**

❑ **How do architectural solutions impact accuracy, energy, latency?**

Hierarchical CAM architecture



Scatter and gather



Horizontal Gather
When N>C
How to combine partial results on different dimensions together?

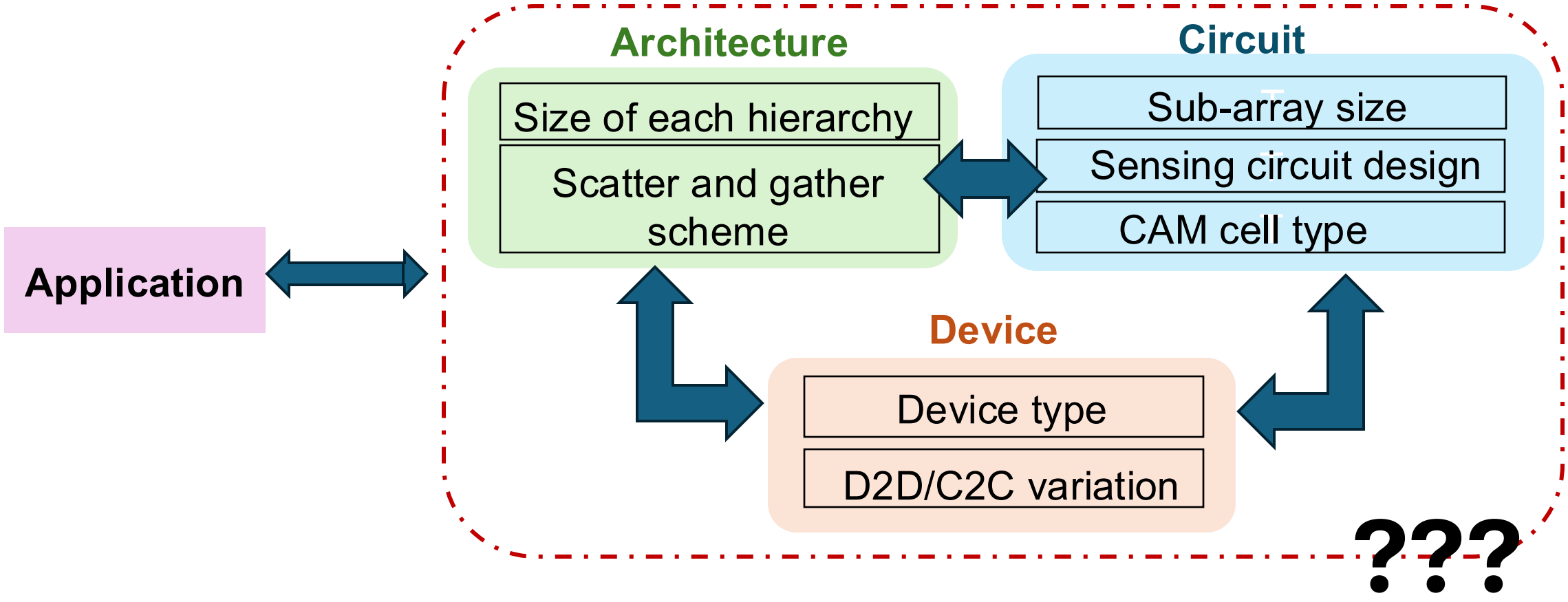
Vertical Merge
When K>R
How to find the match results across different groups of entries?

Gather schemes and limitations

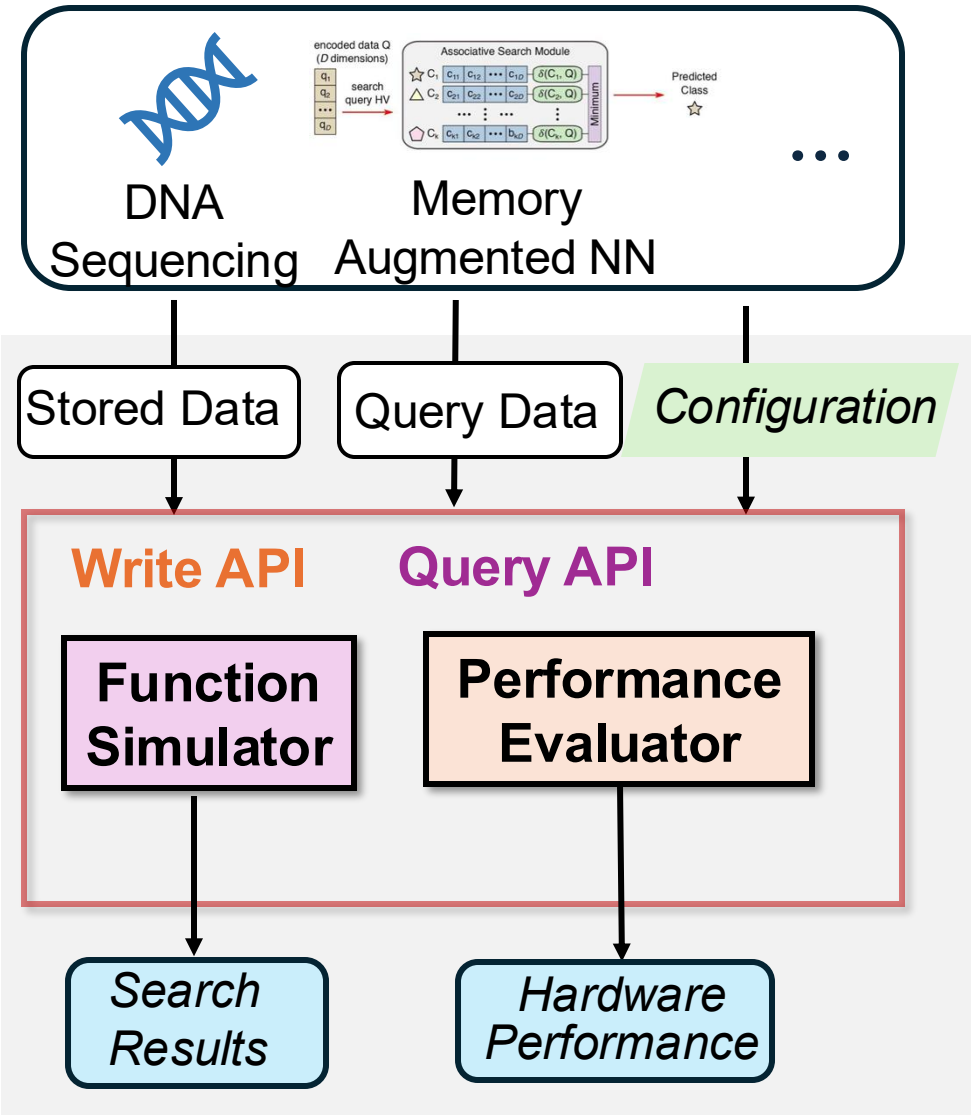
Match Type	Horizontal only	Vertical only	Horizontal & Vertical
Exact	AND	Merge	AND and Merge
Best	Voting	Comparator	N/A
Threshold	N/A	Merge	N/A

- Merge without error
- Gather with approximation error!
- No effective gather schemes proposed yet!!

Cross-layer design choices in CAM-based accelerators



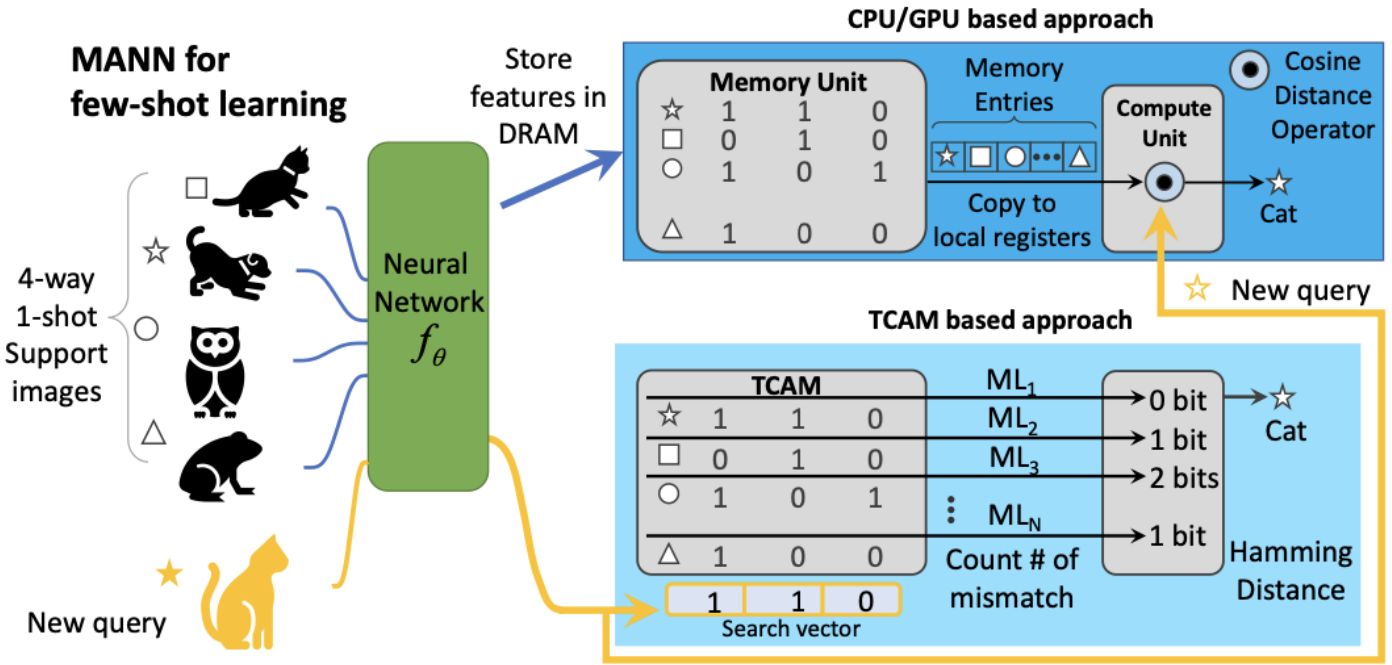
CAMASim: CAM-based accelerator simulator



https://github.com/ND-IMC/CAMASim-V1.0

Case study and look forward

Case study: Memory-augmented neural network (MANN)

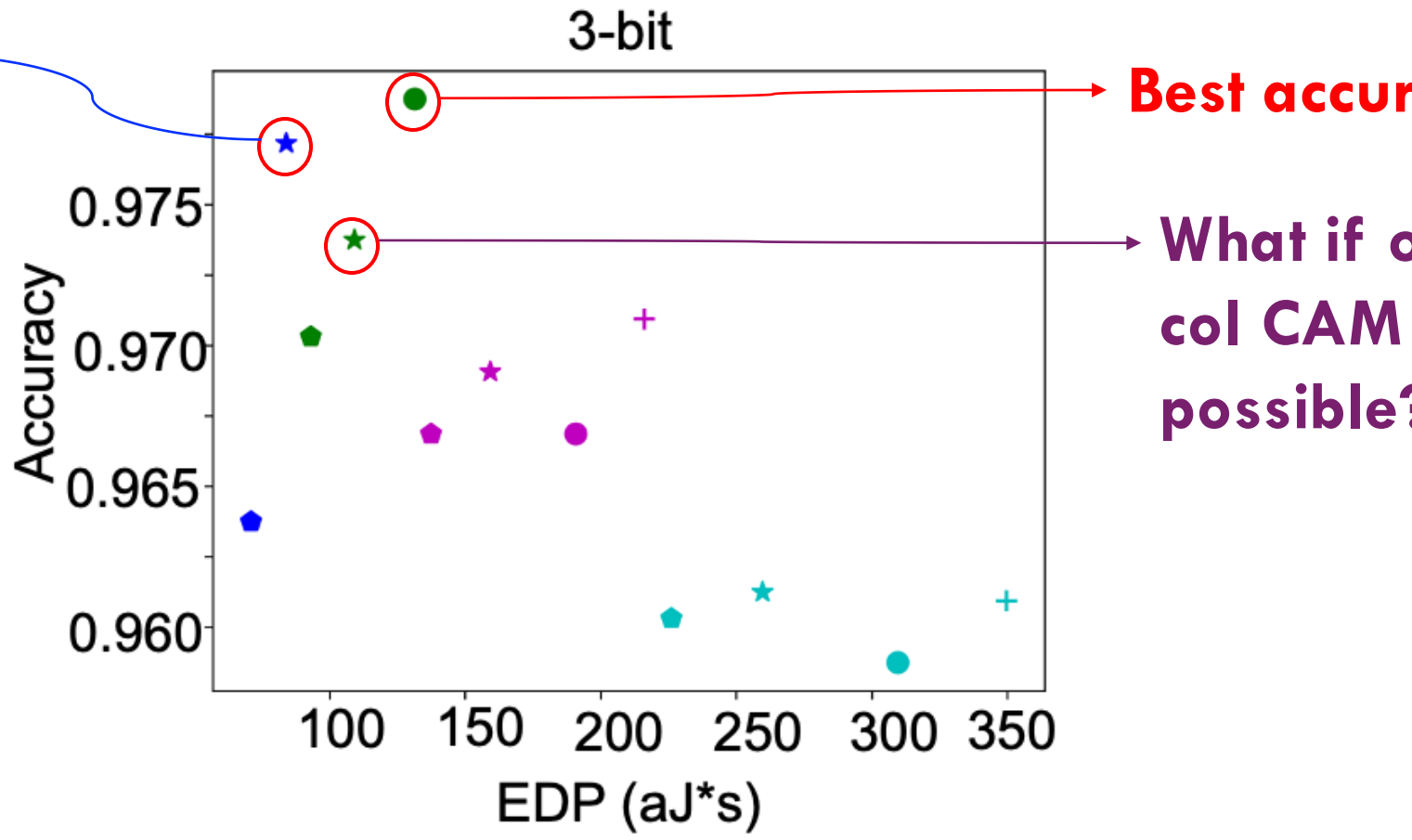


- ❑ A **CNN** for generating high-dimension embeddings
  - ➢ Embedding dimensions ranges from 64 to 512
- ❑ A **CAM**-based accelerator for classification

- **High-dimension embedding → Horizonal gather is required**
  - ➢ **Majority voting used here**
- **What is the best design configuration considering gather errors?**

Impact of embedding and subarray size

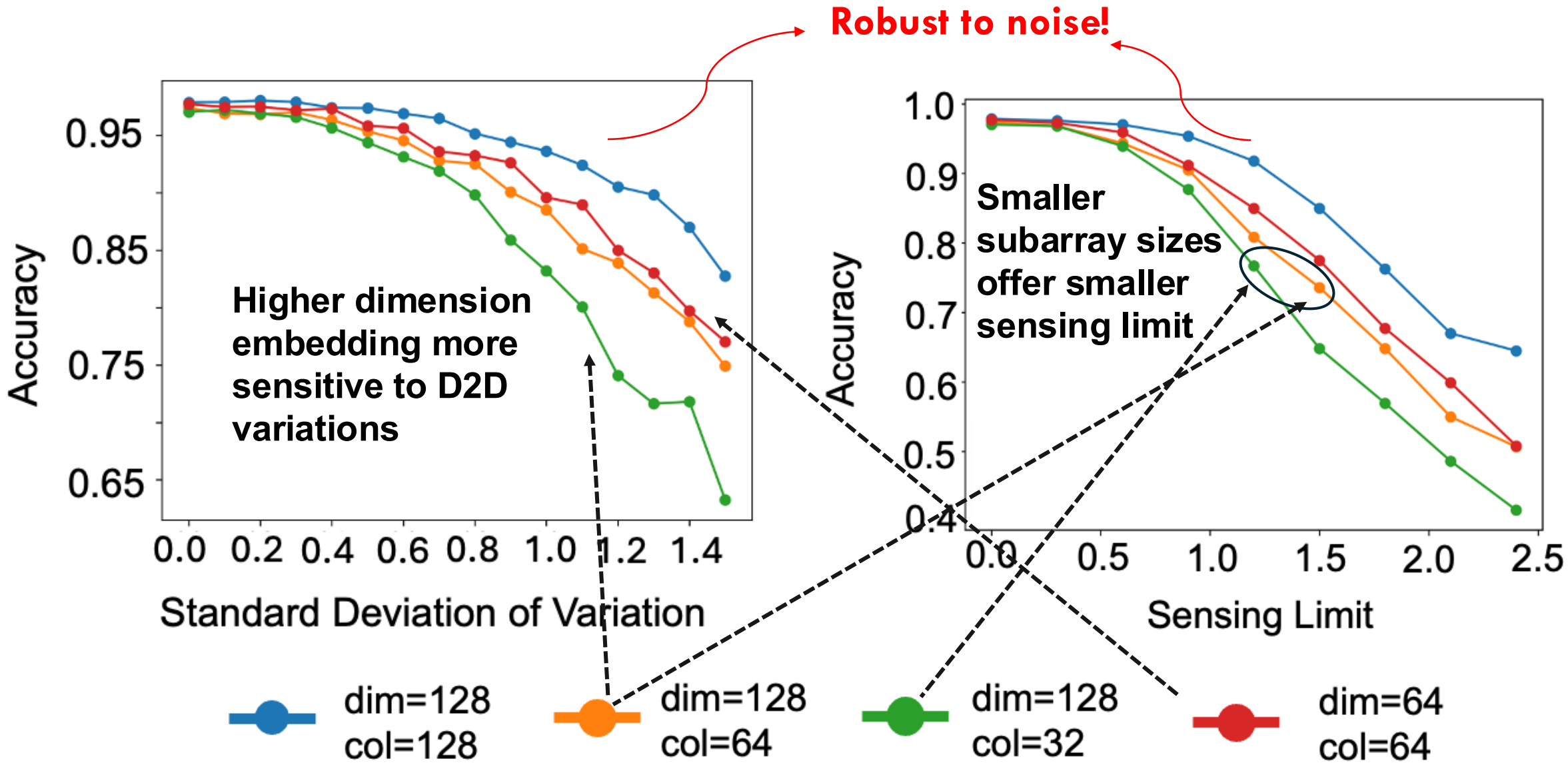
How about this one?



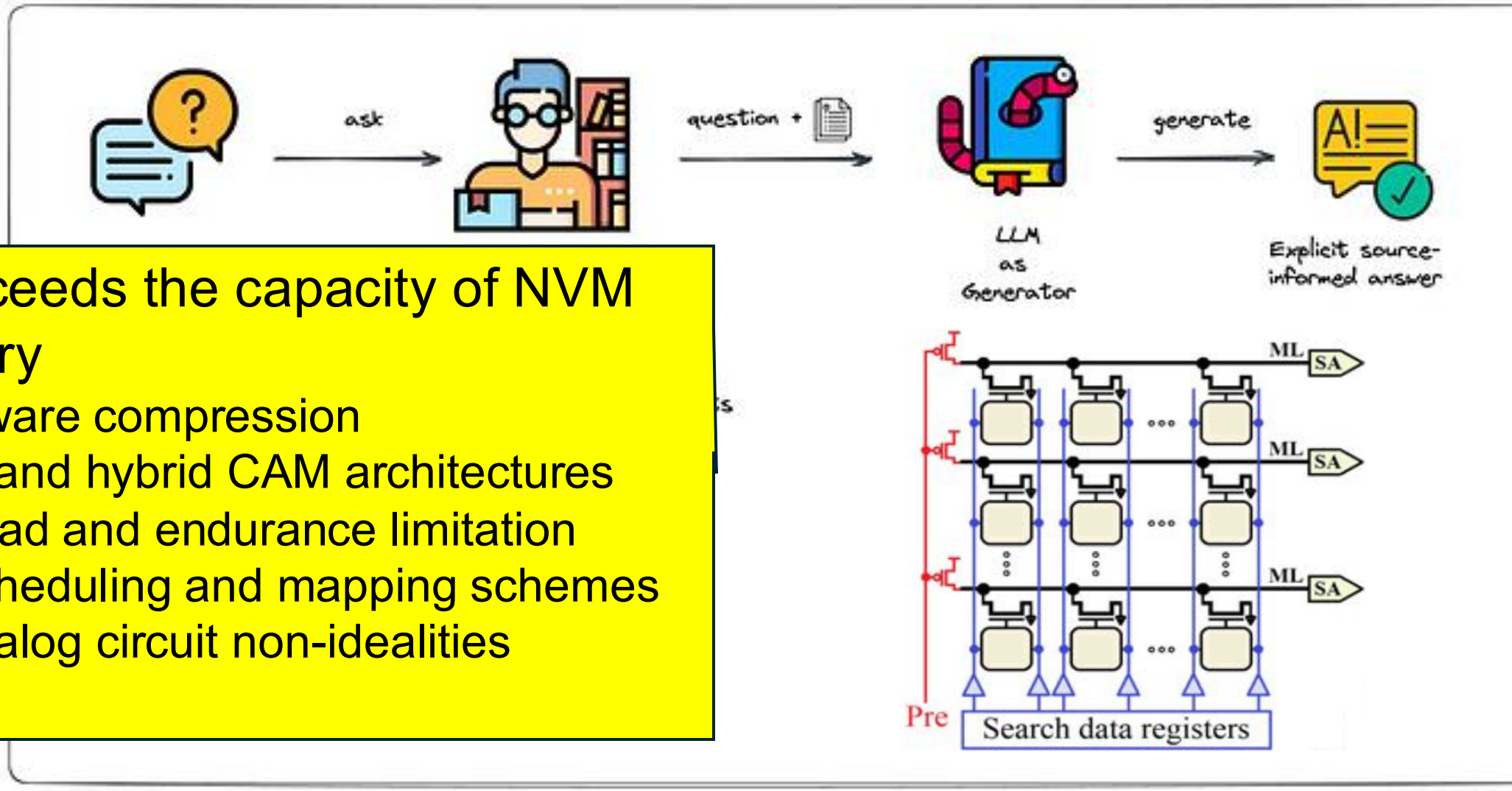
Best accuracy!

What if only 64-col CAM is possible?

Impact of device variation and sensing limit



Looking forward



Data size exceeds the capacity of NVM based memory

- Hardware-aware compression
- Hierarchical and hybrid CAM architectures
- Write overhead and endurance limitation
- Intelligent scheduling and mapping schemes
- NVM and analog circuit non-idealities
- …

Use of associative search in retrieval augmented generation for Large Language Model (LLM)