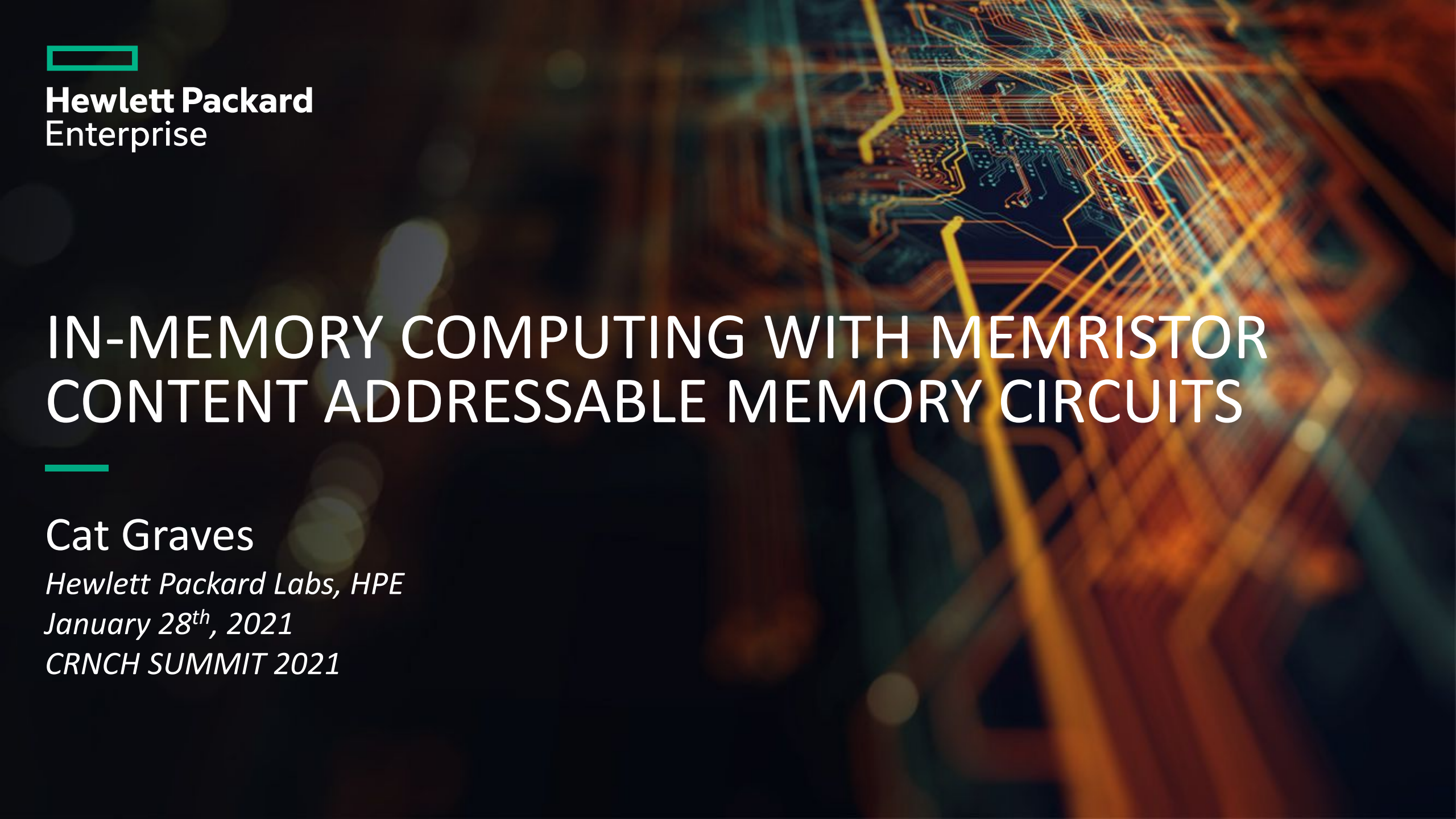


Hewlett Packard Enterprise

IN-MEMORY COMPUTING WITH MEMRISTOR CONTENT ADDRESSABLE MEMORY CIRCUITS

Cat Graves

Hewlett Packard Labs, HPE

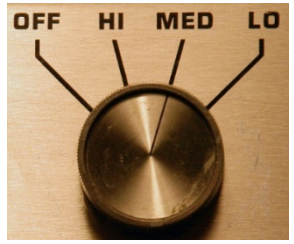
January 28th, 2021

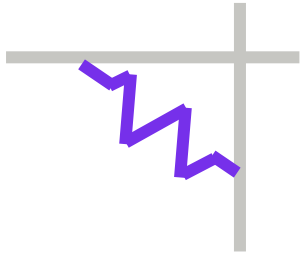
CRNCH SUMMIT 2021

CROSSBARS ACCELERATE MATRIX OPS WITH IN-MEMORY COMPUTE

Non-volatile, analog tunable memristors

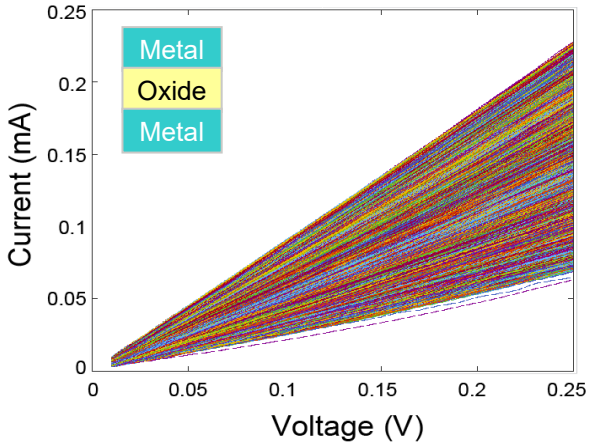
States



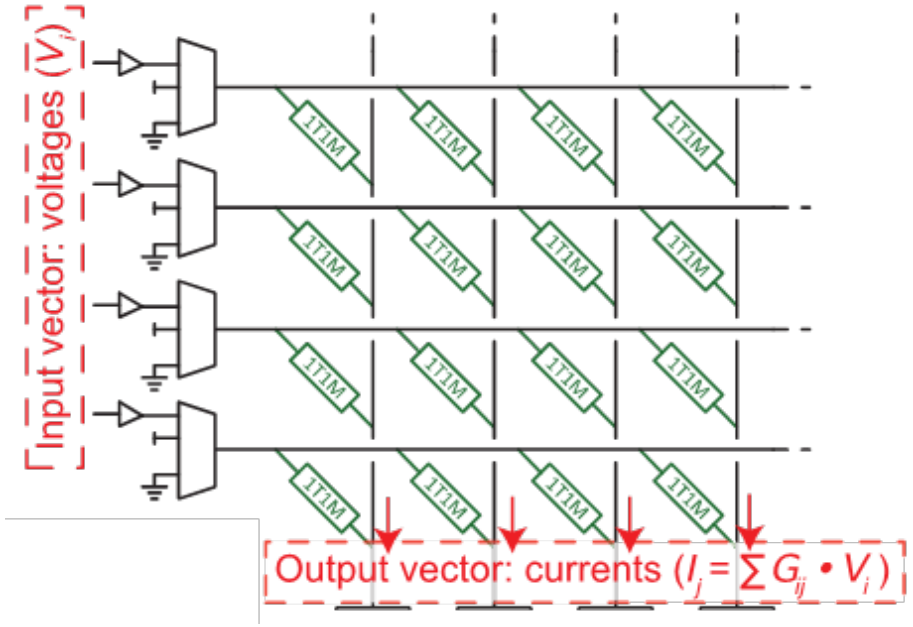


Memristor = Memory + Resistor

Tunable Resistance



Computing: Crossbar circuit of memristors

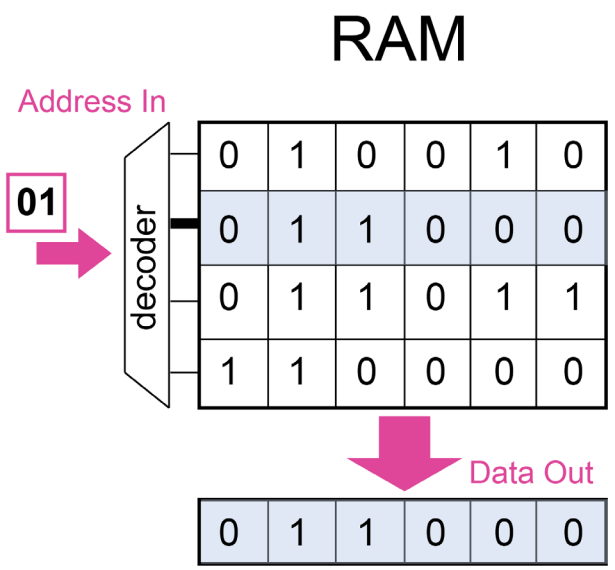


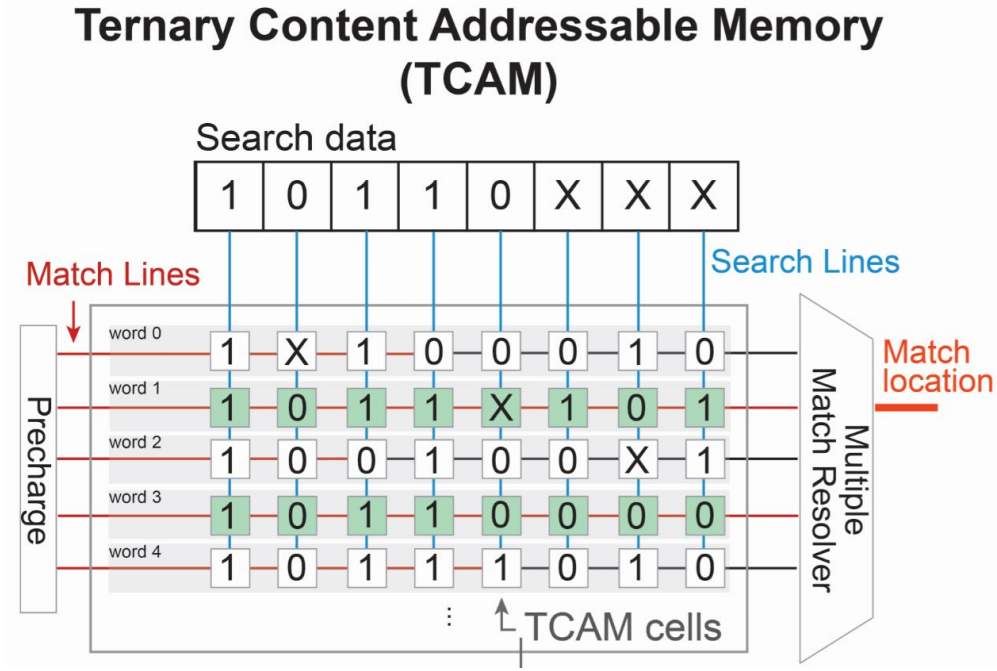
Note: many non-volatile technologies may be used (PCM, STT-MRAM, Ferroelectric, Flash)

Function: vector-matrix multiplication

- Machine Learning
  - Fully connected layers in neural networks
- Scientific computing
  - PDE solver
  - Linear equation solver
- Signal processing
  - FFT, DCT, compression
- Optimization
  - Core for Hopfield NN

ANOTHER IN-MEMORY COMPUTING PRIMITIVE:
Content Addressable Memories (CAMs)





Traditional Random Access Memory (RAM)

- **Input**: Address, **Output**: Contents at this address
- Exact and precise storage of information

Content Addressable Memories (CAM) and Ternary CAMs

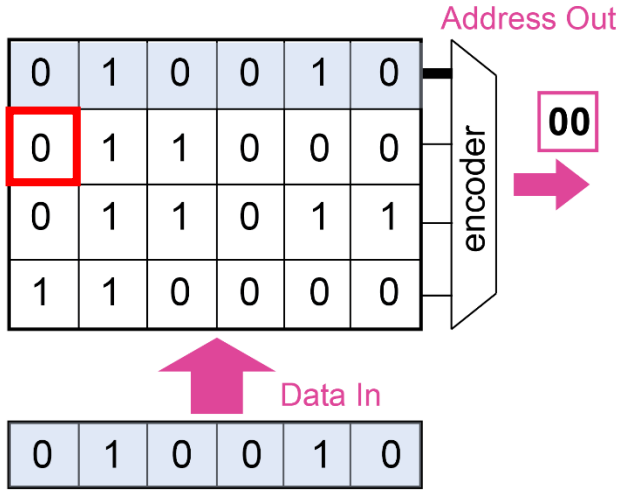
- **Input**: Search word (content), **Output**: address of a match
- Ternary CAMs store and search 0,1, and X = ‘don’t care’
- High performance: search word compared to all entries simultaneously
  - Rapid search time, No collisions
  - Used in network routing, highly associative caches, translation lookaside buffers (TLBs), microarchitectural queues etc.

USE NON-VOLATILE MEMORY INSTEAD OF SRAM IN TCAM CELL

TCAMs cost **30 times more $ / bit** than DDR SRAMs, and consume **150 times more power / bit**

Y. Tsukamoto, et al. *Symp. on VLSI Circuits, 2015*

CAM/TCAM cell

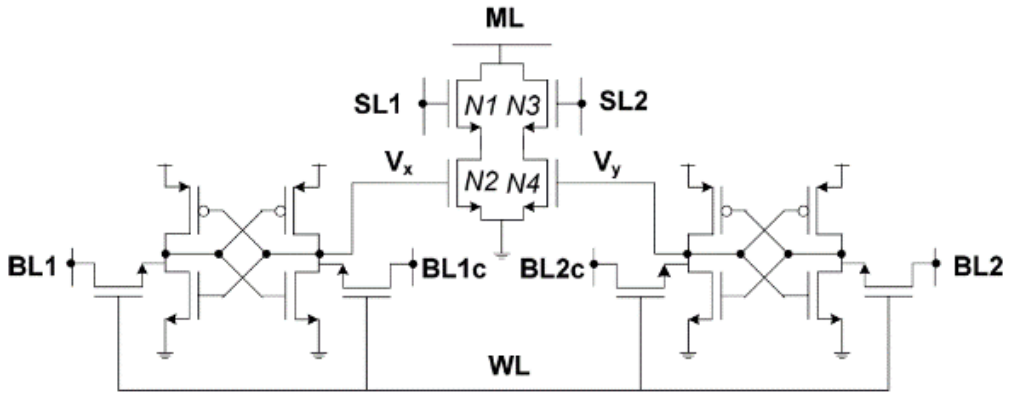


- Utilize non-volatile memristors with no leakage power
- 2-3.5x lower area than SRAM-TCAM
- 35x lower power that SRAM-TCAM
- However, much slower re-programming time (us rather than ns)

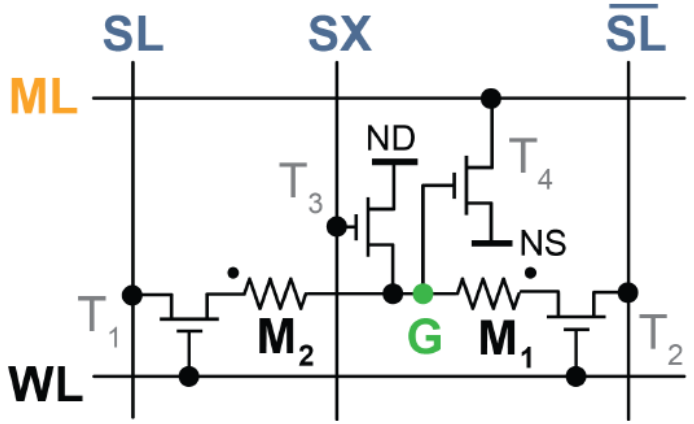
→ Good match for in-memory computing

C. E. Graves, *et al*, **NANOARCH** (2018);
C. E. Graves, *et al*, **ICRC** (2018);
C. E. Graves, *et al*, **IEEE TNano**, (2019)

16T SRAM-based TCAM cell



4T-2Memristor TCAM cell

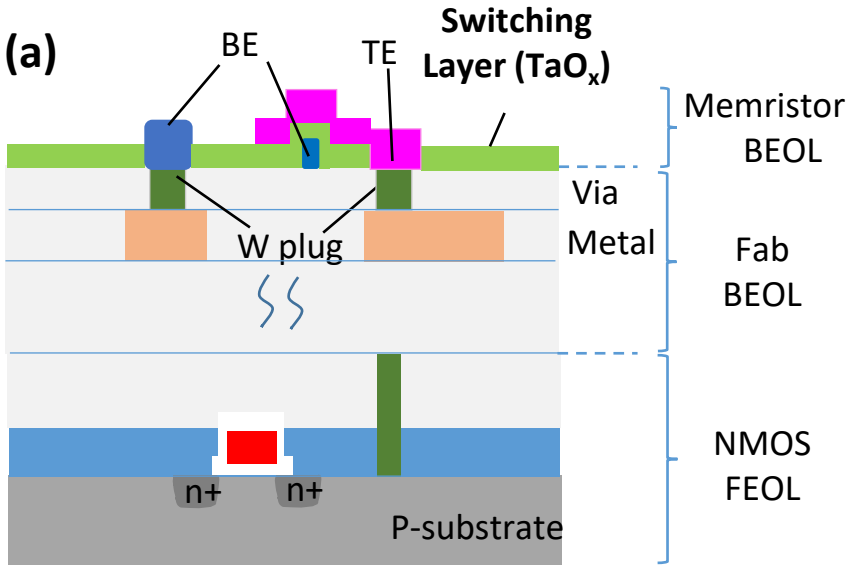


TCAM content	(M_1, M_2)
0	(0,1)
1	(1,0)
X	(0,0)

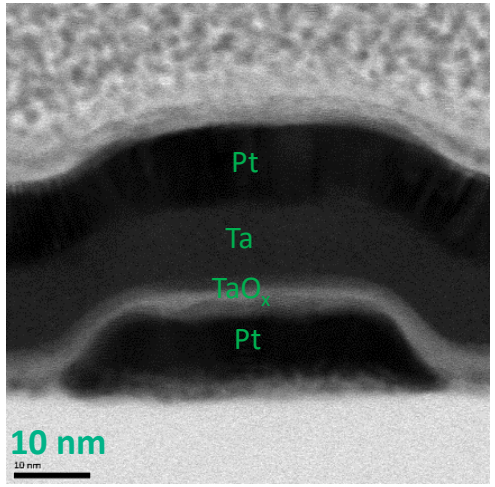
0 = HRS, 1 = LRS

ML = Match Line
WL = Word Line
SL/$\overline{SL}$ = Search Line

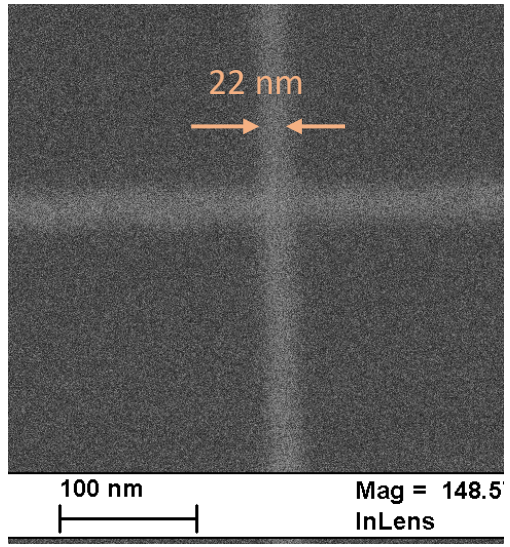
NANOSCALE MEMRISTOR INTEGRATION WITH CMOS

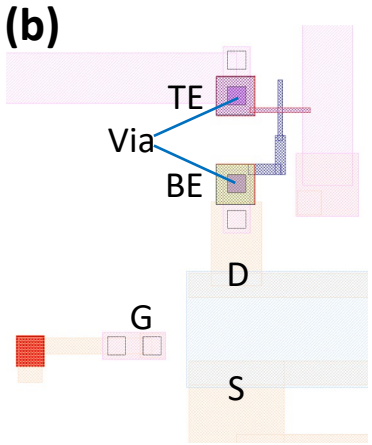


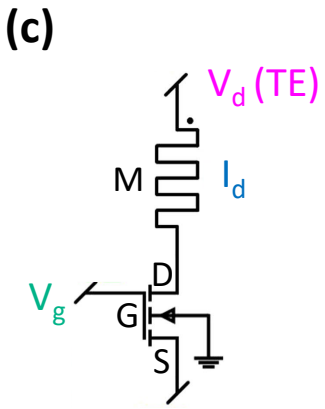
Cross Section TEM



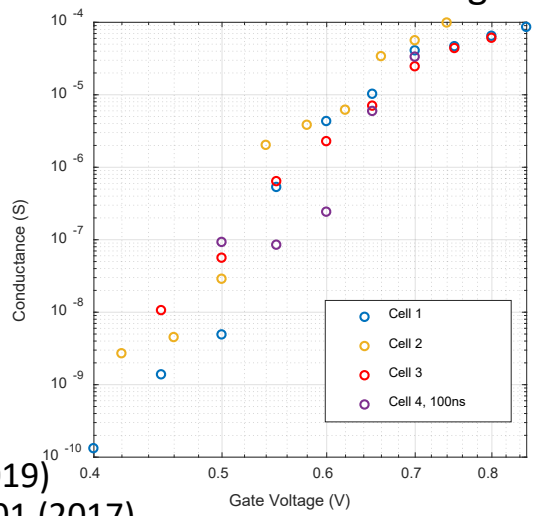
< 25nm memristors







Analog programmable, >10^5 Conductance range



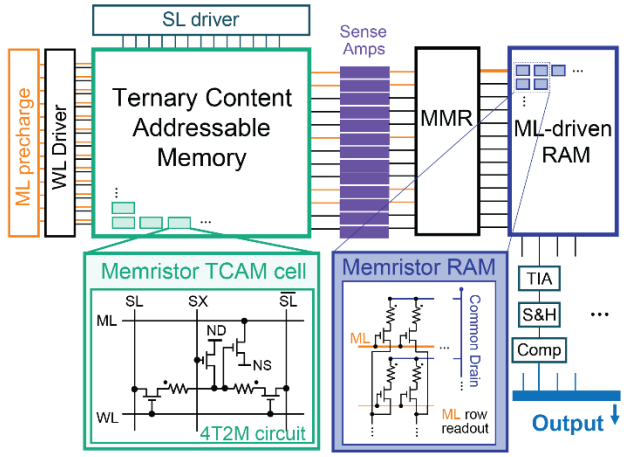
How many bits? ($2\sigma_{max}$ *spacing)*

G Range (S)	σ_{max} (S)	Number of States	Number of Bits
10^{-7}-10^{-6}	2.5×10^{-7}	4	2
10^{-6}-10^{-5}	3×10^{-7}	20	>4
10^{-5}-3×10^{-4}	1×10^{-6}	150	>7
3×10^{-4}-6.2×10^{-4}	3×10^{-7}	532	>9

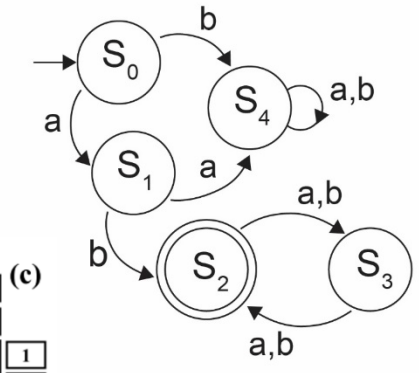
Trade-off between #bits and current levels

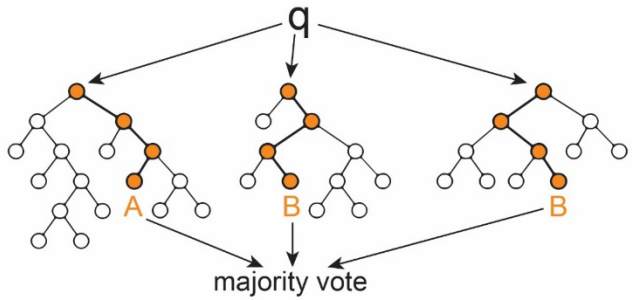
X. Sheng, *et al*, **Adv. Electron. Mater**, 5, 1800876 (2019)
C.E. Graves, *et al*, **Applied Physics Letters** 110, 123501 (2017)

IN-MEMORY COMPUTING OPPORTUNITIES WITH CAMS
Lower power and flexible NVM designs drive new applications

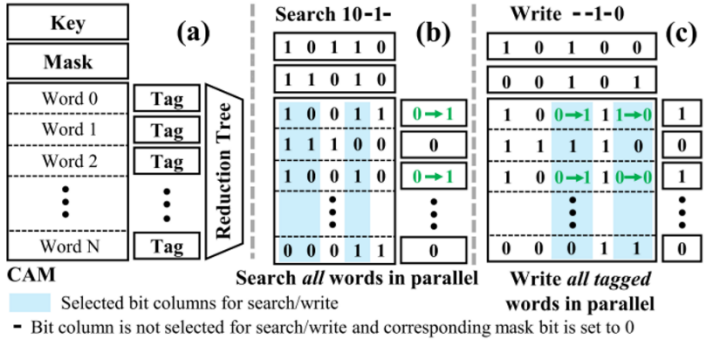


Finite State Machines
(Regular Expression matching for network security, branch prediction and cache controllers for HPC)





Associative Computing



Tree-based Models
(Random Forests, Business Rule Management Systems)

$$f(X_1, X_2, X_3) = X_1X_2X_3 + X_1\overline{X_2}X_3 + \overline{X_1}X_2X_3$$

$X_1X_2X_3$	Z
000	0
001	0
010	0
011	1
100	0
101	1
110	0
111	1

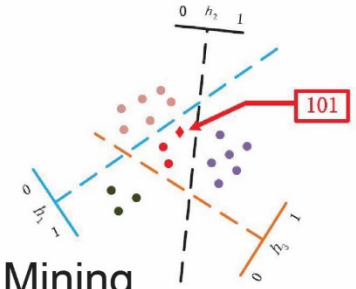
LUT

$X_1X_2X_3$
011
101
111

BCAM

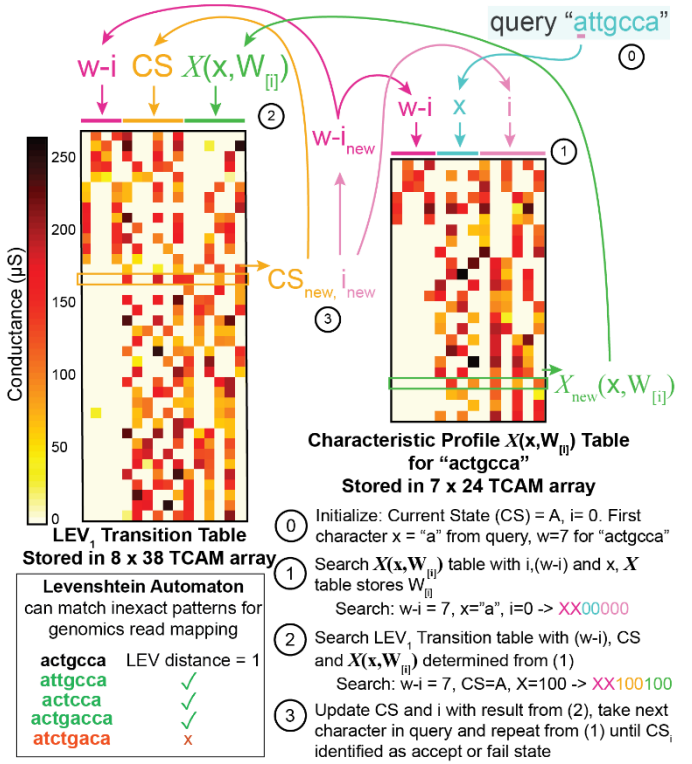
$X_1X_2X_3$
011
1X1

TCAM



Data Mining
(Locality sensitive hashing (LSH), TF-IDF, data-intensive workloads)

Reconfigurable Computing
(Arbitrary logic functions, associative computing)

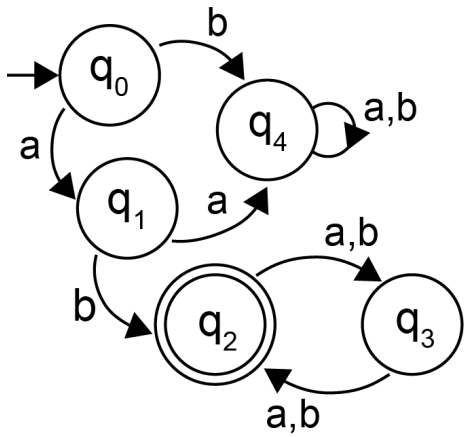


Edit distance or Levenshtein Automaton
(Genomics)

USING CAMS FOR COMPUTING: FINITE AUTOMATA

Deterministic Finite Automaton to accept **w** when **w** is of even length and begins with ‘ab’

Transition Diagram



OR

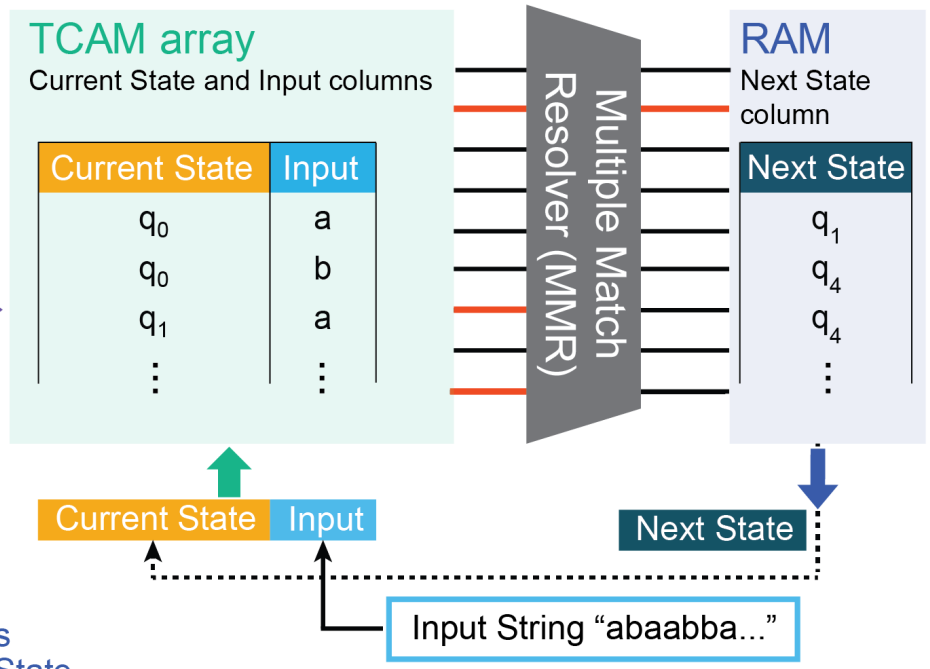
State Transition Table

Current State	Input	Next State
* q_0	a	q_1
* q_0	b	q_4
q_1	a	q_4
q_1	b	q_2
q_2	a	q_3
q_2	b	q_3
q_3	a	q_2
q_3	b	q_2
q_4	a	q_4
q_4	b	q_4

Search word = Current State + Input

Returns Next State

State Transition Table directly encoded in TCAM + RAM hardware

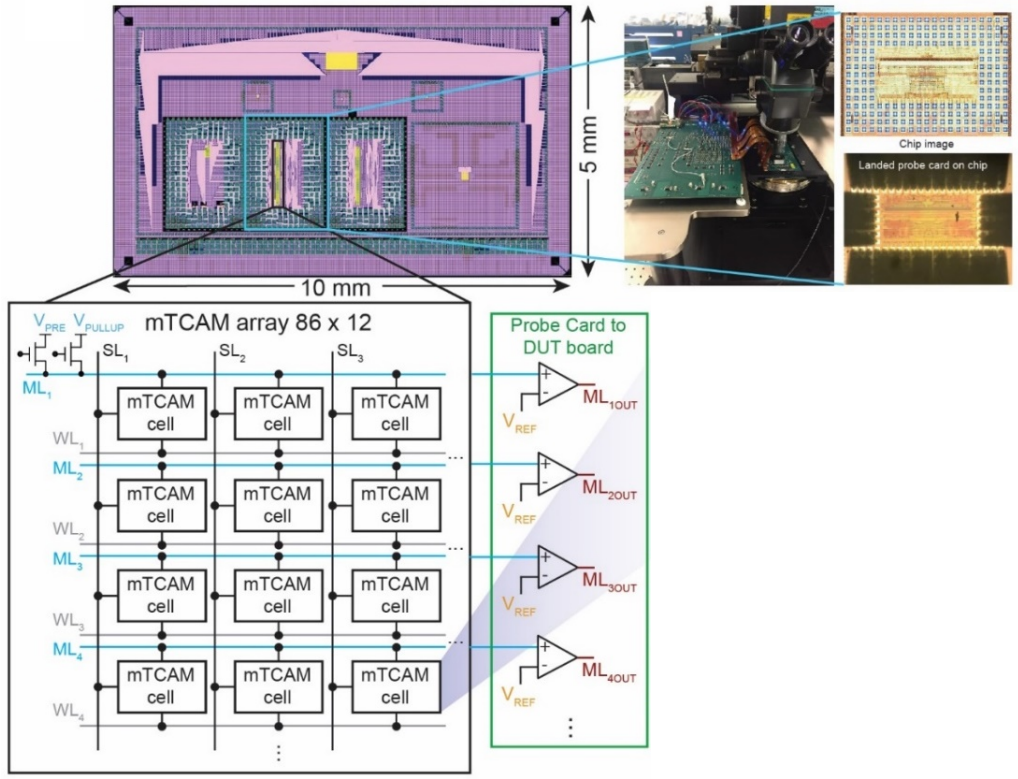


Finite Automata have applications in Bio-informatics/Genomics and Security

C. E. Graves, *et al*, **ICRC** (2018);
C. E. Graves, *et al*, **IEEE Transactions on Nanotechnology**, 18, 963 (2019)

EXPERIMENTAL DEMONSTRATIONS FOR NETWORK SECURITY

Non-volatile TCAM arrays taped-out
180nm CMOS + 50 nm Integrated memristors



Regular Expression matching on chip

RegEx: “sid=[0-9A-F]{32}”

Naive STT: 37 states, 516 rows

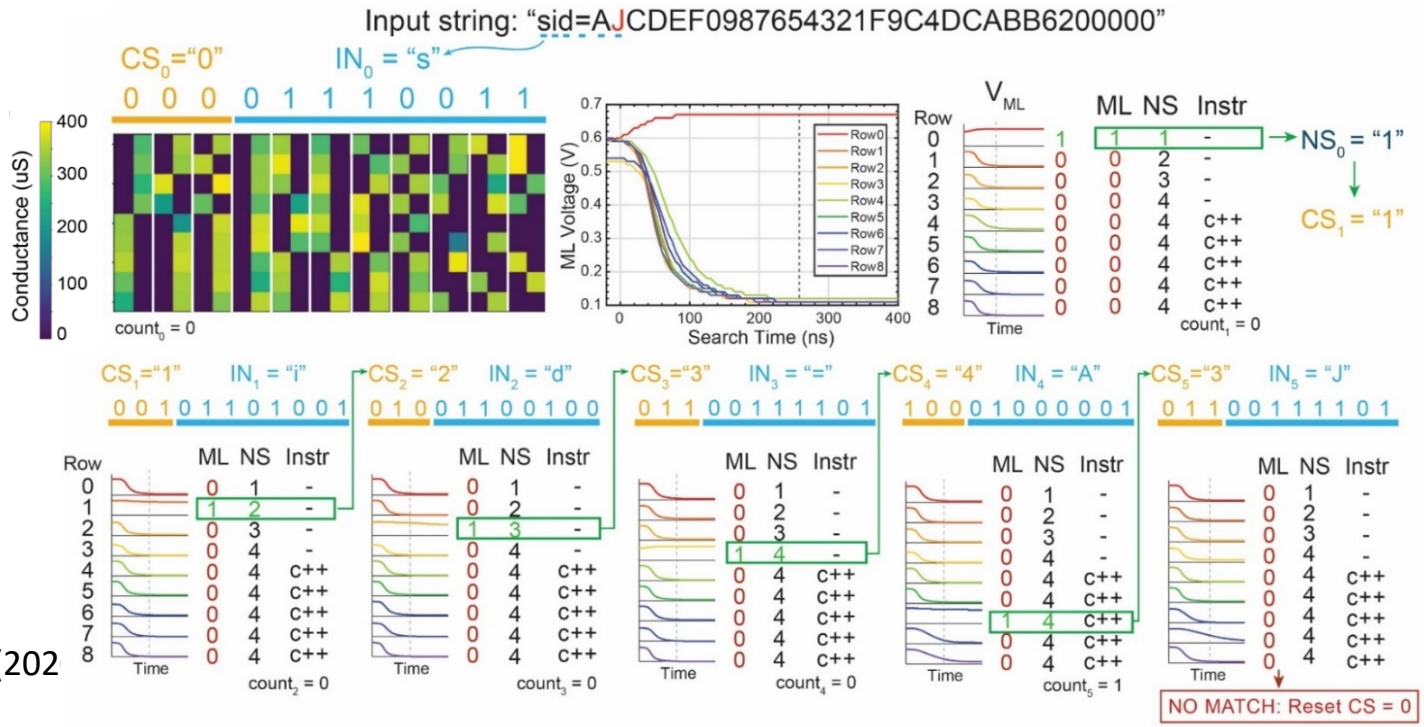
CS	Input	NS
0	01110011 (s)	1
1	01101001 (i)	2
2	01100100 (d)	3
3	00111101 (=)	4
4	[0-9A-F] 16 entries	5
5	[0-9A-F] 16 entries	6
.....	.....	.....
.....	.....	.....
35	[0-9A-F] 16 entries	36

Compressed STT: 5 states, 9 rows

CS	Input	NS	Inst.
000	01110011 (s)	1	
001	01101001 (i)	2	
010	01100100 (d)	3	
011	00111101 (=)	4	
100	00110XXX	4	Count++
100	0011100X	4	Count++
100	010000X1	4	Count++
100	01000X10	4	Count++
100	0100010X	4	Count++

Compressed STT for TCAM array

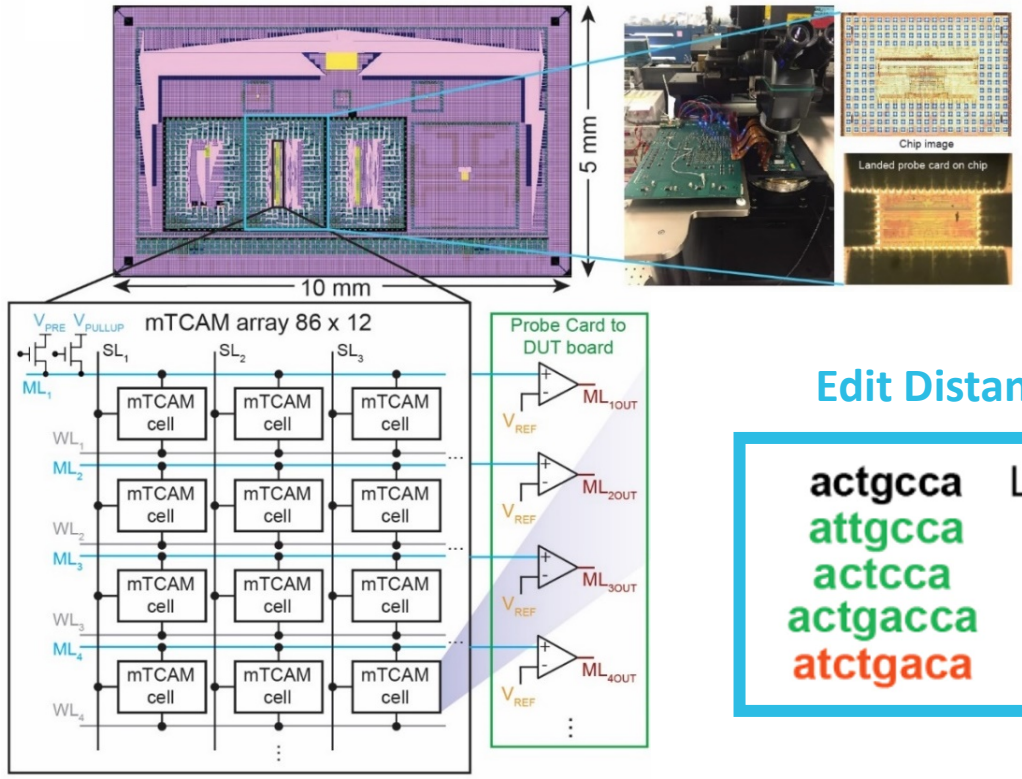
0	0	0	0	1	1	1	0	0	1	1
0	0	1	0	1	1	0	1	0	0	1
0	1	0	0	1	1	0	0	1	0	0
0	1	1	0	0	1	1	1	1	0	1
1	0	0	0	0	1	1	0	X	X	X
1	0	0	0	0	1	1	1	0	0	X
1	0	0	0	1	0	0	0	0	X	1
1	0	0	0	1	0	0	0	X	1	0
1	0	0	0	1	0	0	0	1	0	X



C.E. Graves, *et al*, **Advanced Materials,** 32, 2003437 (202

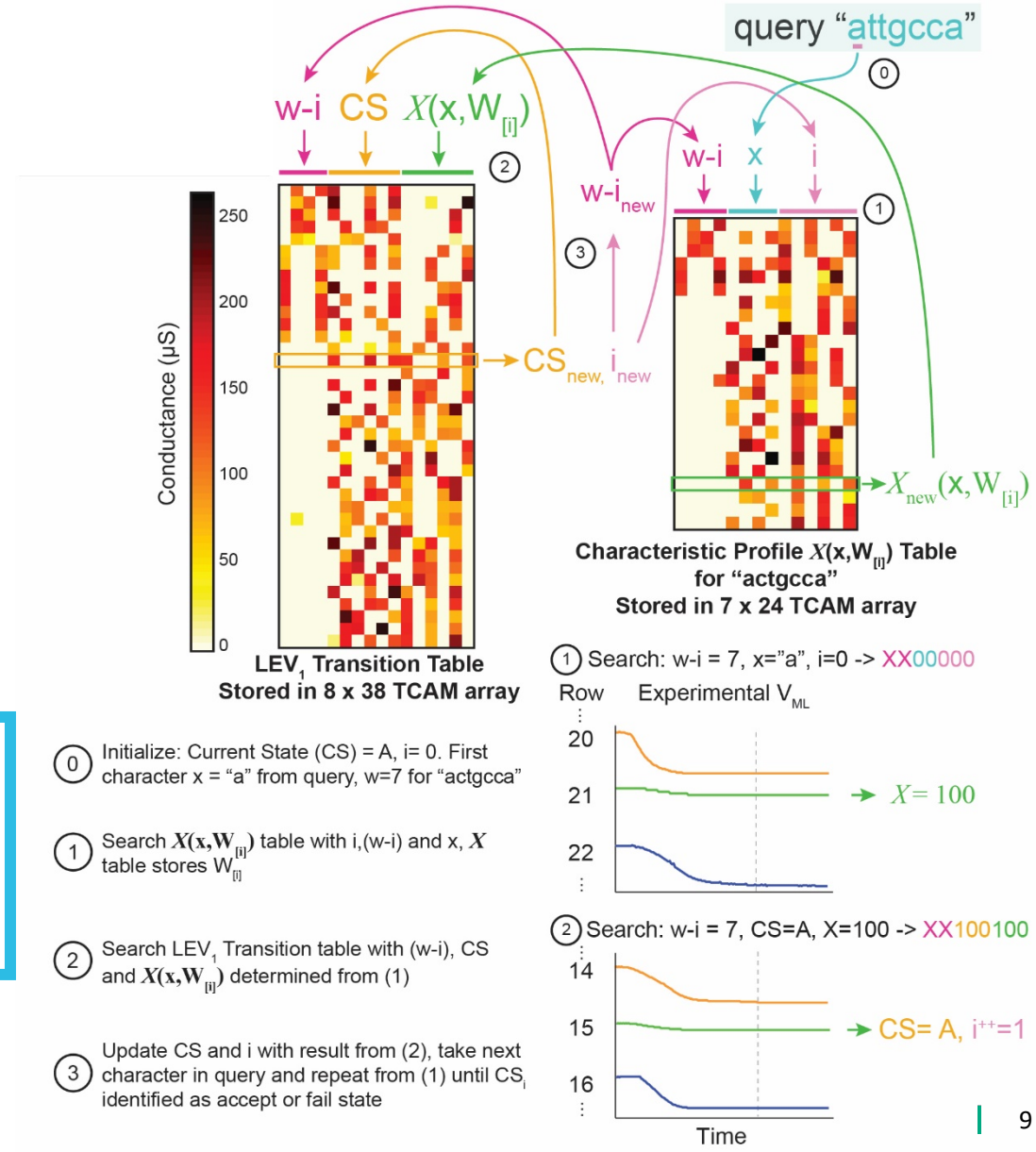
EXPERIMENTAL DEMONSTRATIONS FOR GENOMICS APPLICATIONS

Non-volatile TCAM arrays taped-out
180nm CMOS + 50 nm Integrated memristors



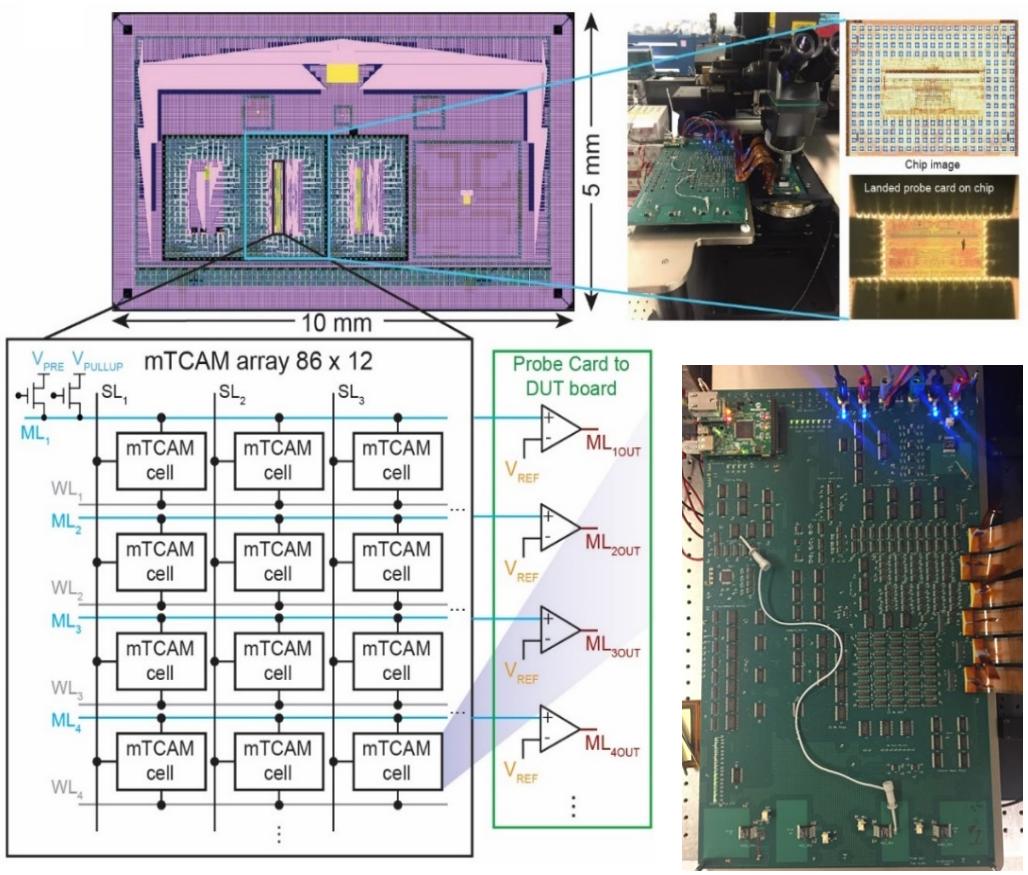
Edit Distance Automata

actgcca	LEV distance = 1
attgcca	✓
actcca	✓
actgacca	✓
atctgaca	x



C.E. Graves, *et al*, **Advanced Materials,** 32, 2003437 (2020)

PERFORMANCE: 25X IMPROVEMENT VS STATE-OF-THE-ART FPGA



Performance on a Network Security ("SNORT") ruleset:

47.2 Gbps at **0.3W**
Memristor-TCAM system
7-var stride [1]

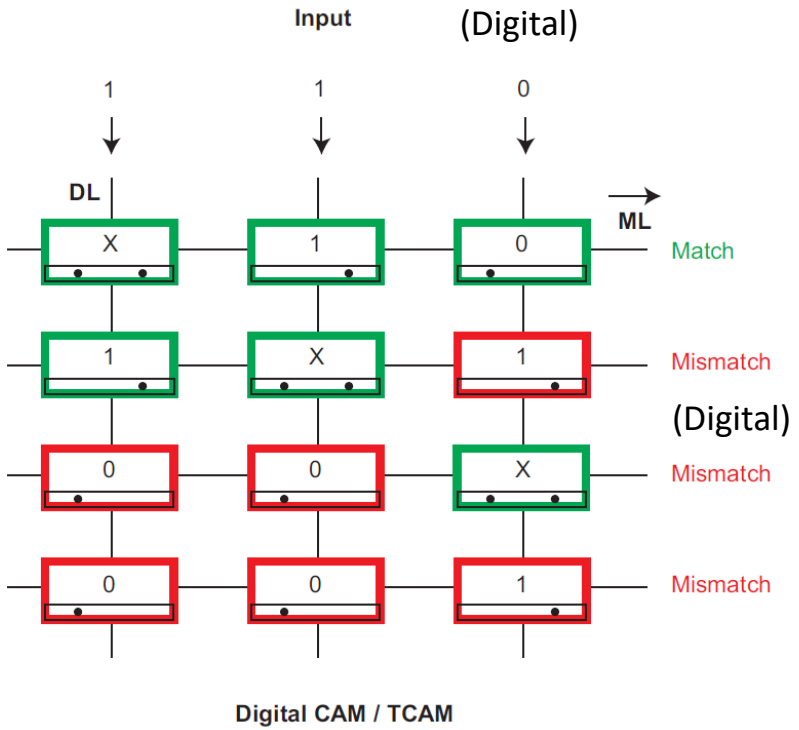
3.9 Gbps at **0.63W**
FPGA, Best reported
4-stride [2]

25x improved Throughput/Watt (with smaller area and greatly improved scalability)

C.E. Graves, *et al*, **IEEE Transactions on Nanotechnology**, 18, 963 (2019)
C.E. Graves, *et al*, **Advanced Materials,** 32, 2003437 (2020)

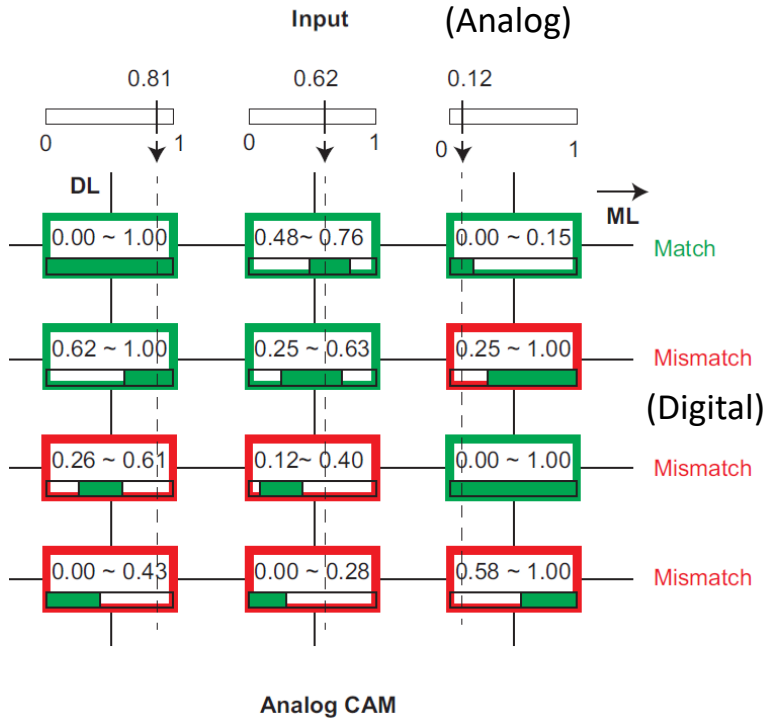
NEW INVENTION: ANALOG CONTENT ADDRESSABLE MEMORY (ACAM)

Previous Digital TCAM



Stores: 0,1,or 'X'
Input Search: 0,1, or 'X'

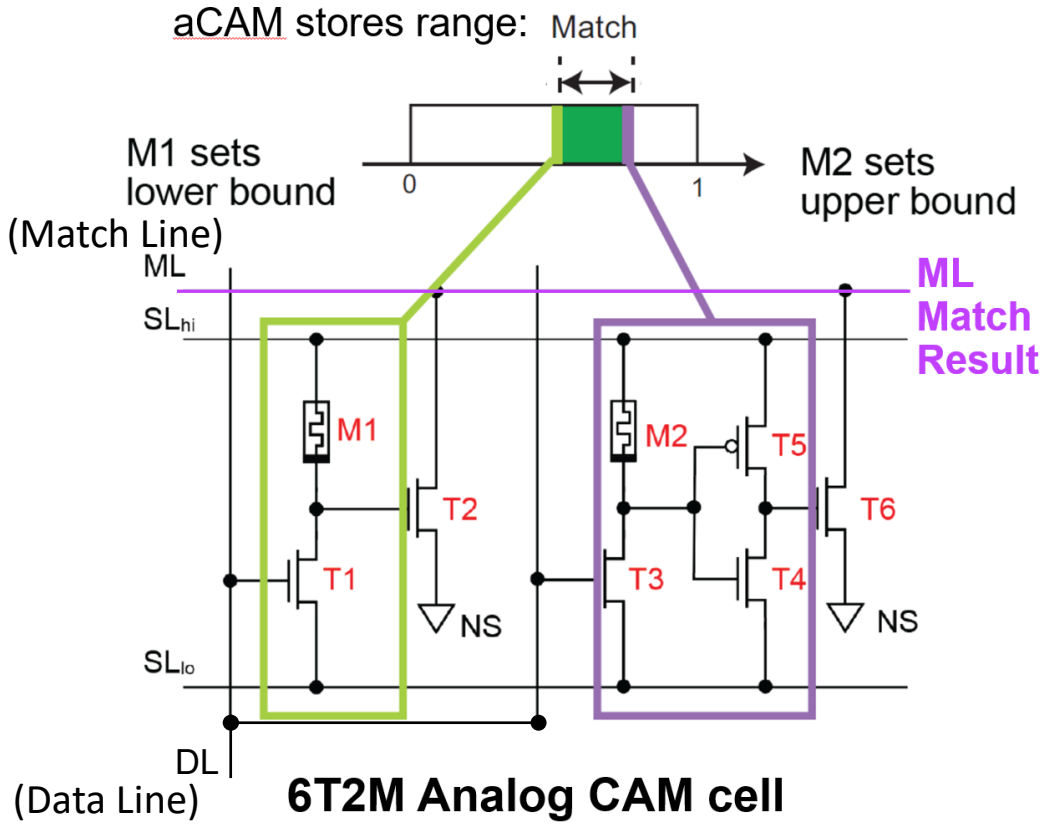
New Analog CAM

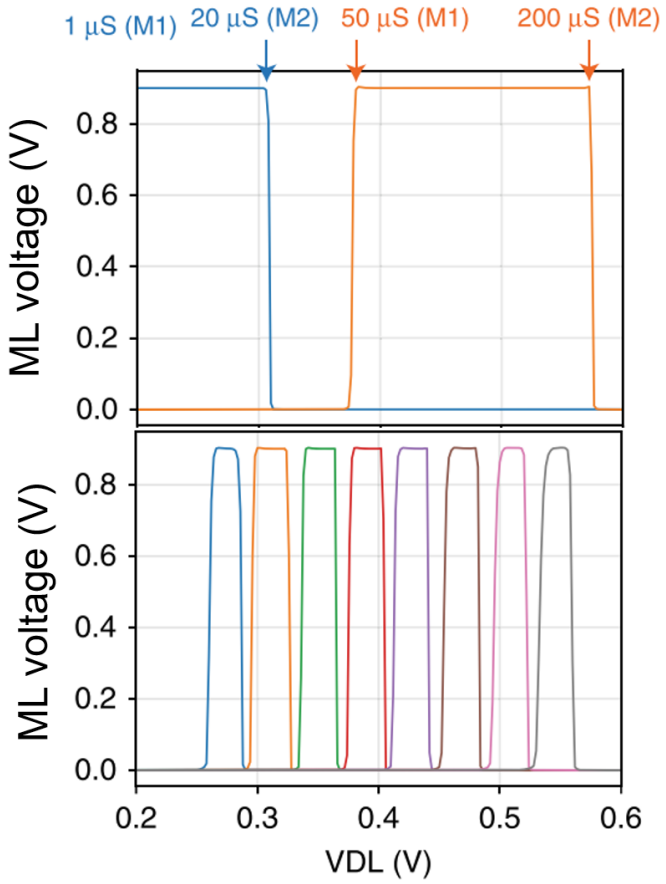


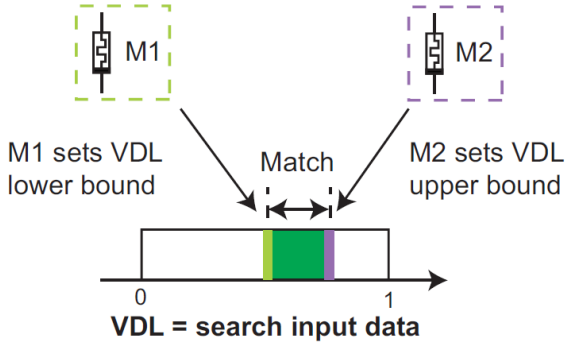
Stores: analog ranges
Input Search: analog values (multi-bit)

Can Li *et al*, **Nature Communications** 11, 1 (2020)

ANALOG CAM (A-CAM) CIRCUIT USING MEMRISTOR TUNABLE CONDUCTANCE







Memristor a-CAM	SRAM TCAM
6 transistors	16 transistors
3-4 bits	1-ternary-bit

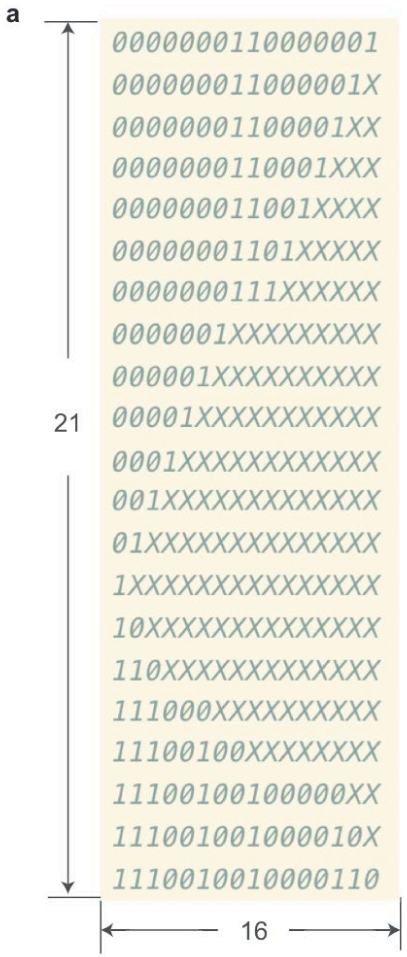
+ a-CAM is non-volatile
+ allows range encoding

Can Li *et al*, **Nature Communications** 11, 1 (2020)

AREA/ENERGY COMPARISON TO HIGHLY OPTIMIZED SRAM-TCAM

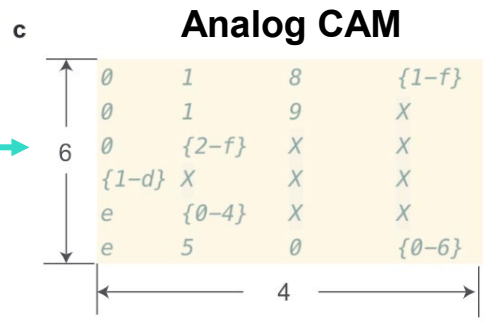
Class B IP address routing table

Digital CAM (TCAM)



Reduced rows due to range-capability

Reduced columns due to multi-bit storage



Over 4x lower energy/bit/search

18.8x reduction in Area (@16nm)

and a-CAM is non-volatile

Supply	Energy/search	Energy/search/ equiv. TCAM bit
ML precharging	102.9 fJ	0.007 fJ
SLhi driver	298.5 fJ	0.021 fJ
Others	86.4 fJ	0.006 fJ
DAC	52.1 fJ	0.004 fJ
Total	**539.9 fJ**	**0.037 fJ**

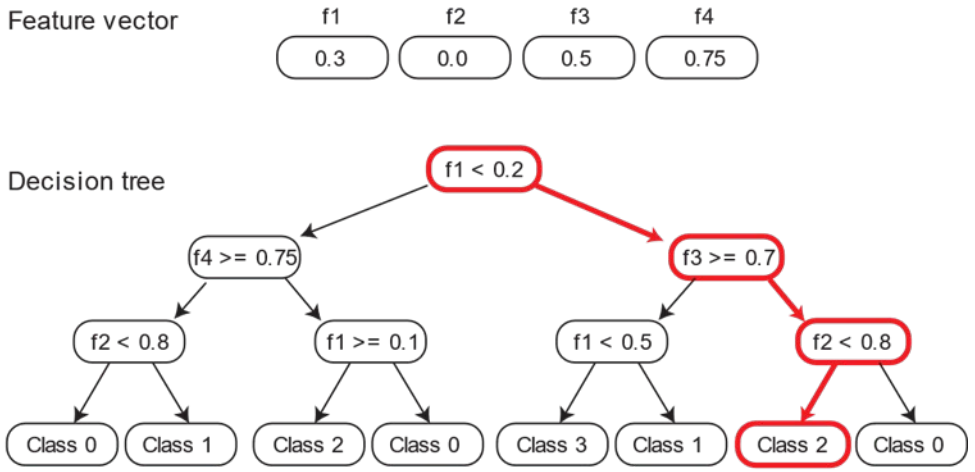
vs 0.165 fJ with highly optimized SRAM-TCAM

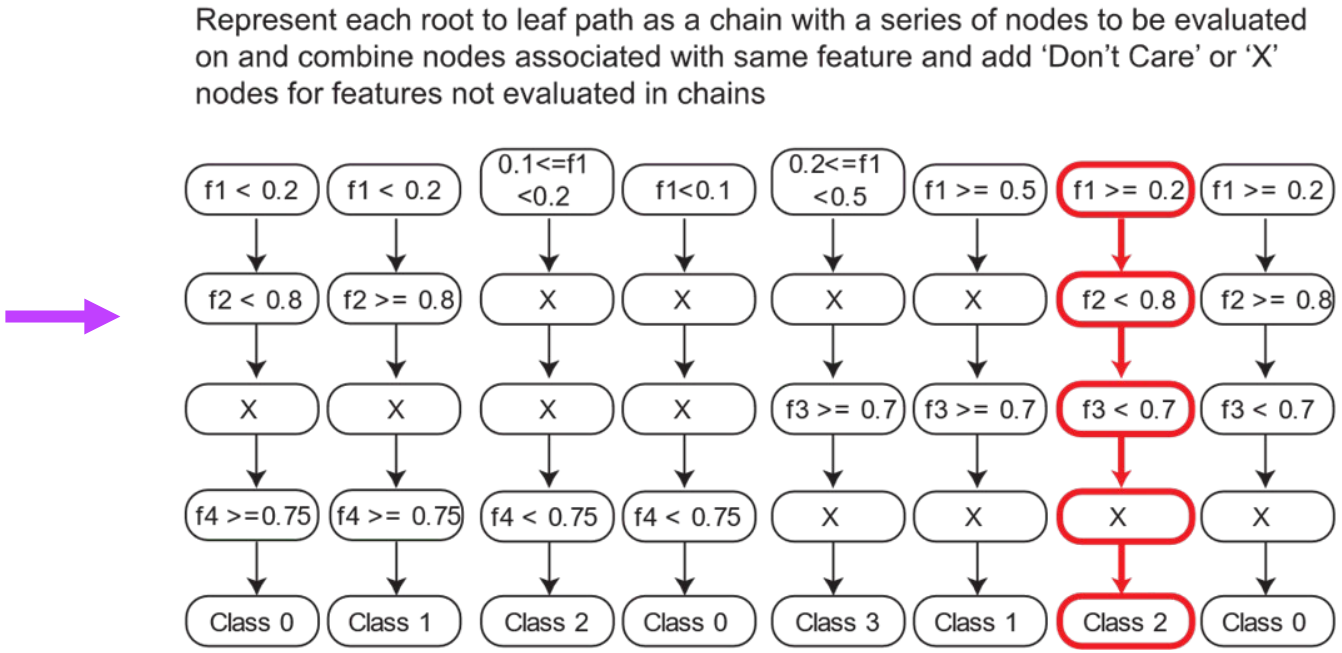
(SRAM has >10x power saving techniques not implemented in our a-CAM)

Can Li *et al*, **Nature Communications** 11, 1 (2020)

NEW APPLICATIONS ENABLED: DECISION TREES WITH A-CAM

Decision Trees & Random Forests: Easy to train; support smaller training sets; Highly interpretable and verifiable

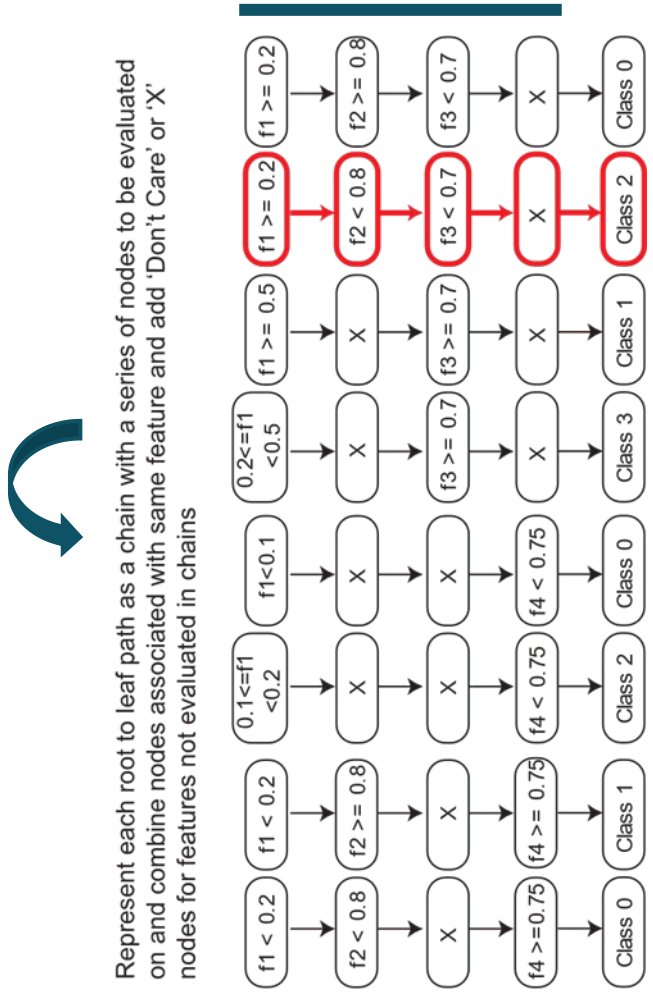


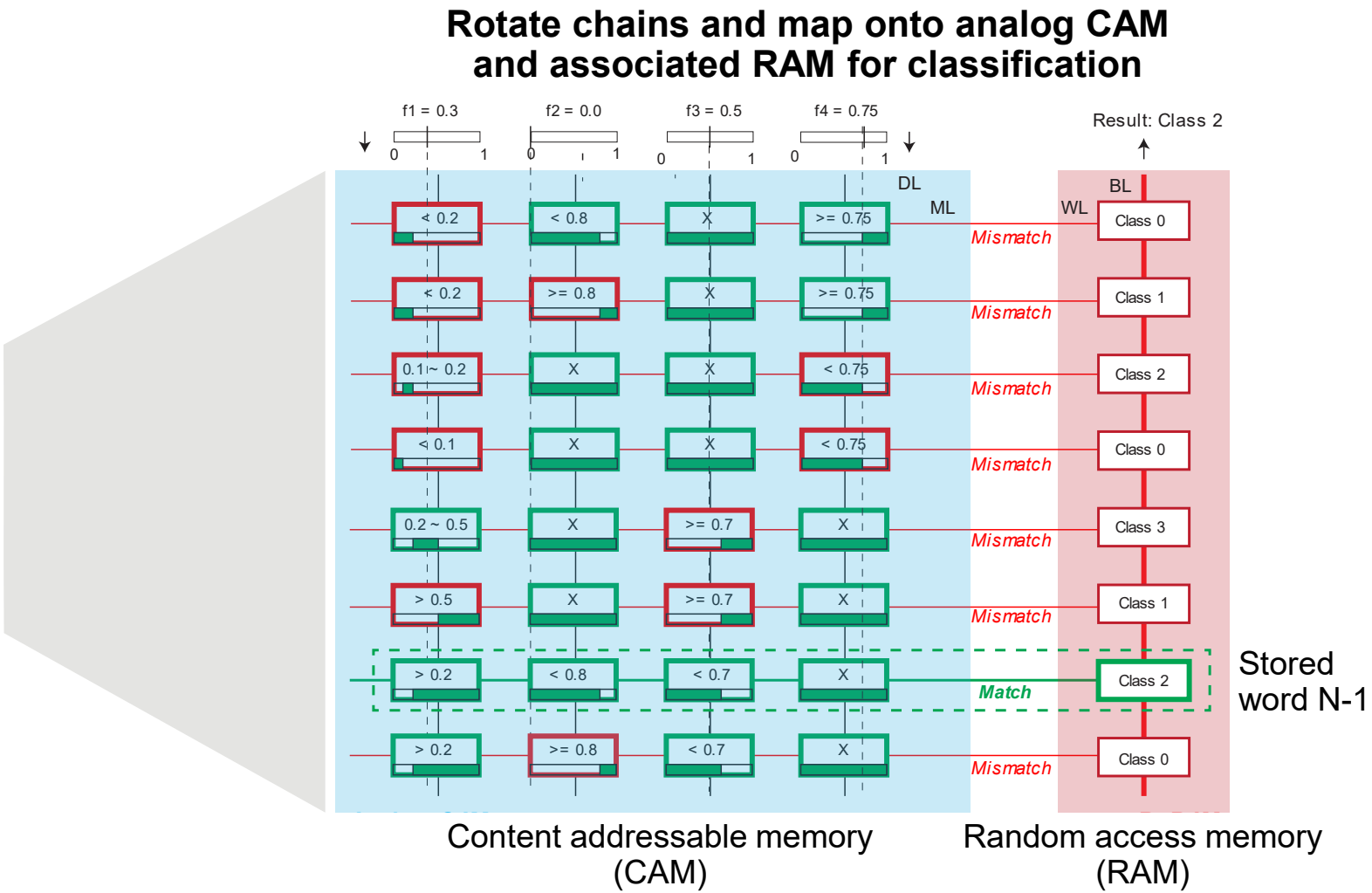


Figures adapted from T. Tommy, et al "Towards machine learning on the automata processor." *International Conference on High Performance Computing*, p. 200 Springer, Cham, 2016.

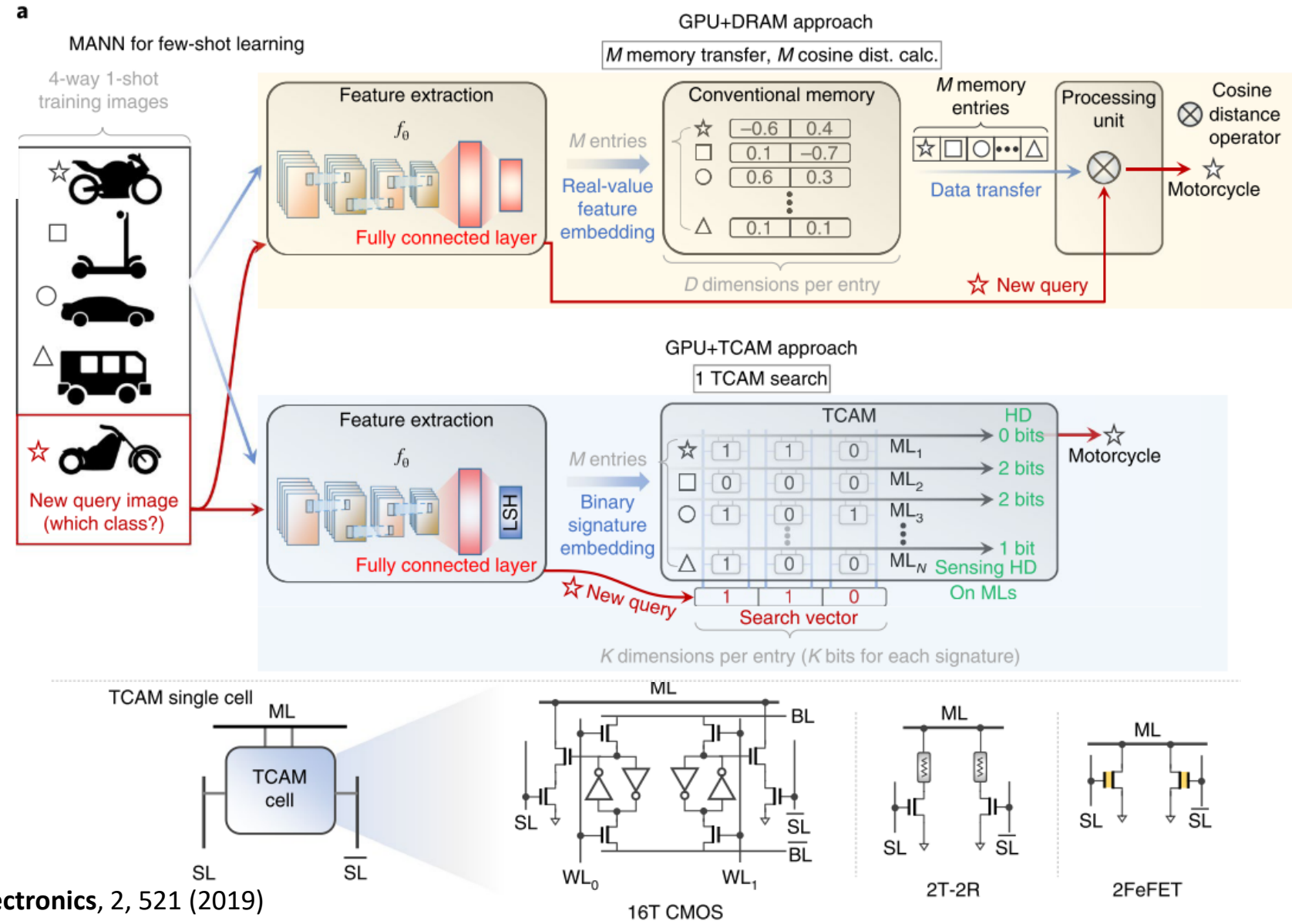
NEW APPLICATIONS ENABLED: DECISION TREES WITH A-CAM

Search feature vector on cols



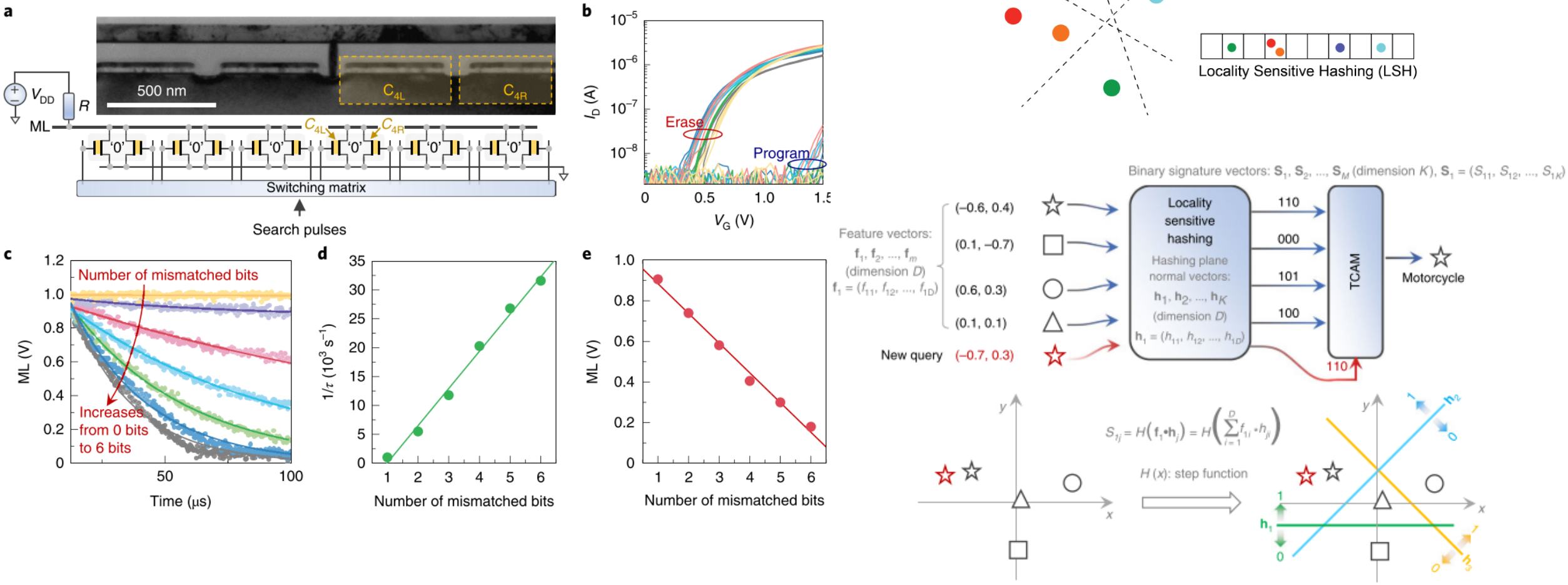


OTHER RECENT CAM COMPUTING WORK: LOCALITY SENSITIVE HASHING WITH FE-CAM



K. Ni, *et al*, **Nature Electronics**, 2, 521 (2019)

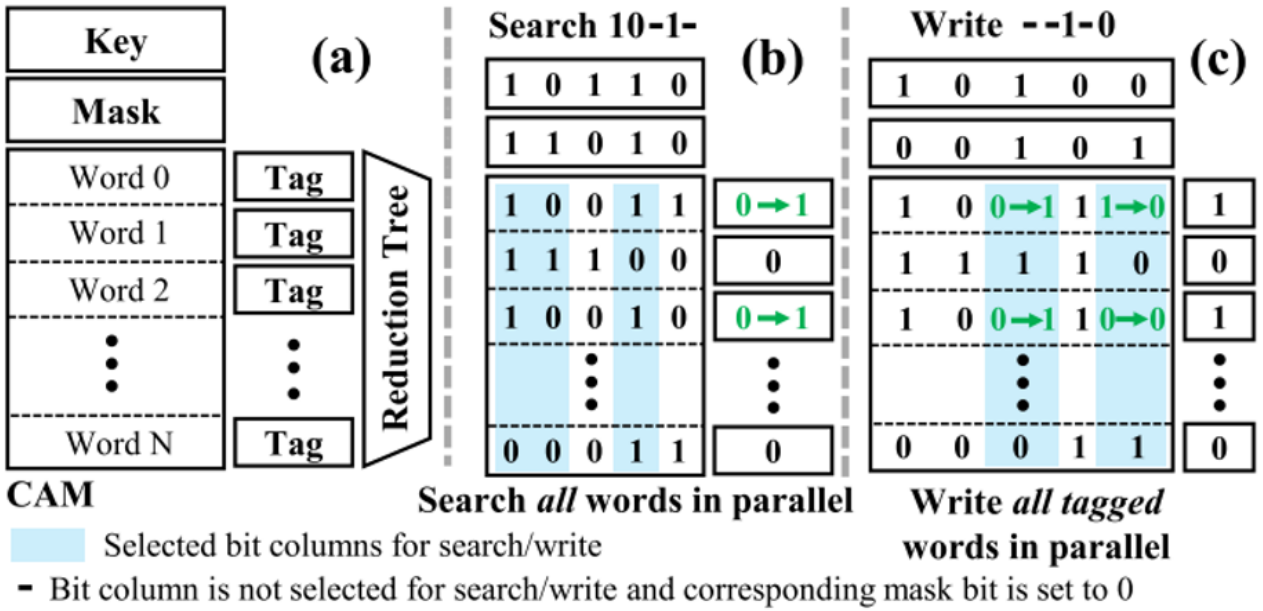
OTHER RECENT CAM COMPUTING WORK: LOCALITY SENSITIVE HASHING WITH FE-CAM



K. Ni, *et al*, **Nature Electronics**, 2, 521 (2019)

OTHER RECENT CAM COMPUTING WORK: ASSOCIATIVE PROCESSING

AP operates concurrently on every word where the data is stored, eliminating the need for data movement.
Word-level parallelism points to applications with high data-level parallelism



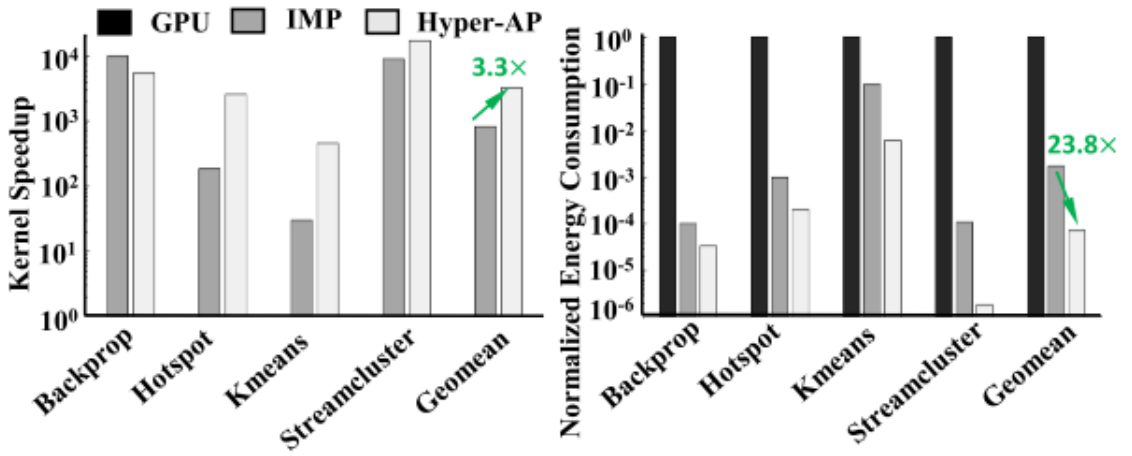


Fig. 18. The speedup (left) and energy consumption (right) results for the Rodinia kernels. The energy consumption of *Hyper-AP* and IMP are normalized to that of GPU.

Y. Zha and J. Li, "Hyper-Ap: Enhancing Associative Processing Through A Full-Stack Optimization" **ISCA**, 846 (2020)

OTHER RECENT CAM COMPUTING WORK: GRAPH PROCESSING

- Scatter Gather Apply operations mapped
- CAM + Crossbar architecture efficiently maps sparse vector matrix multiplication

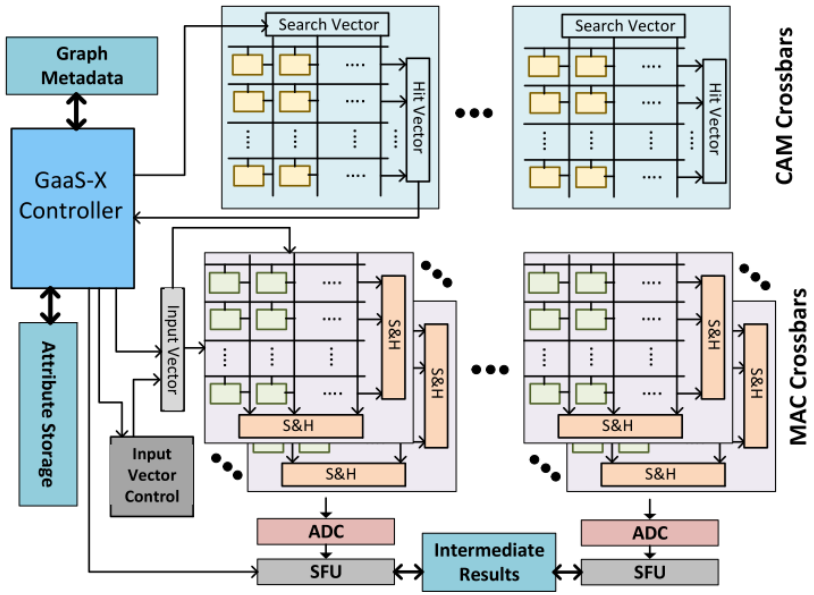


Fig. 6. GaaS-X architecture overview

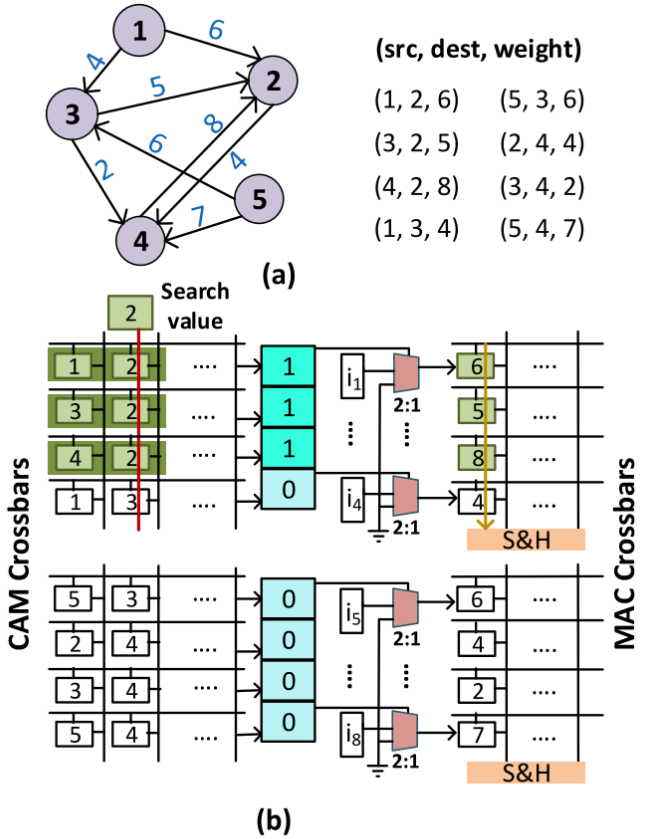


Fig. 7. GaaS-X computing operation

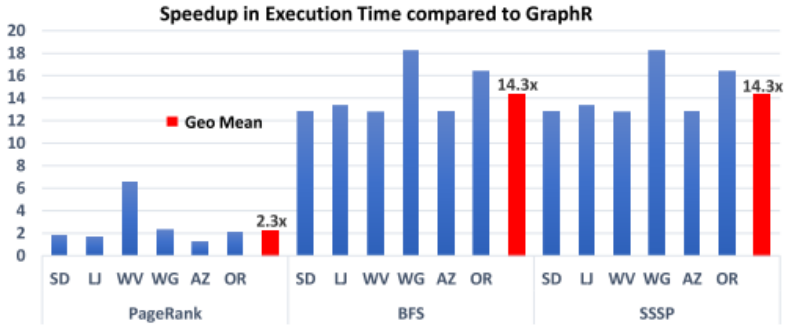


Fig. 11. Speed up in execution time of GaaS-X accelerator for various graph datasets compared to GraphR [33] accelerator.

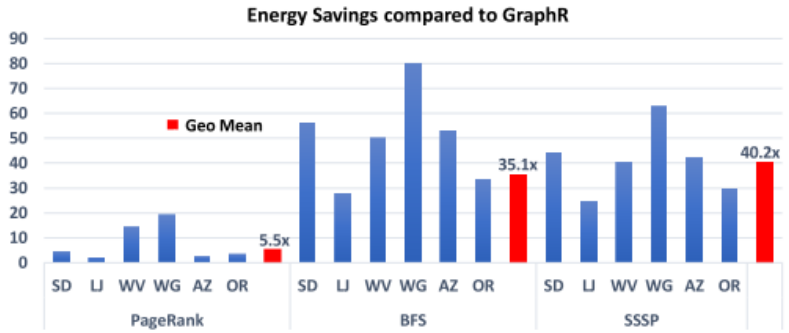


Fig. 12. Energy savings of GaaS-X accelerator for various graph datasets compared to GraphR [33] accelerator.

N. Challapalle, *et al,* "GaaS-X: Graph Analytics Accelerator Supporting Sparse Data Representation using Crossbar Architectures" **ISCA**, 433 (2020)

SUMMARY

- A “new” computing primitive Content Addressable Memory (CAMs): analog and digital CAMs for applications in Security, Genomics, Decision Trees, and more

- CAMs can map a diverse range of computing models
  - Finite State Machines (Reg Ex matching, Edit Distance)
  - Hamming distance calculation, Locality Sensitive Hashing
  - Associative Processing
  - Graph Analytics

- In-memory analog computing platform using non-volatile memristors integrated with CMOS, low current memristors down to 25nm, tuned for CAM properties with application demonstrations

- Analog CAM circuit provides new opportunities
  - ‘Almost matching’ capabilities or range matching

Thank you

Hewlett Packard Labs

John Paul Strachan
Can Li
Xia Sheng
Martin Foltin
Lenny Kiyama
Xuema Li
Jim Ignowski
Rob Wessel
Jacqui Ingemi
Cullen Bash

Program support

Jeffrey Weinschenk (IARPA)
Richart Slusher
Karl Roenigk
John Daly (ARO)
Chad Meiners (MIT LL)
Matthew Hardy (LTS)

Hewlett Packard Enterprise