Aalto University
School of Electrical
Engineering

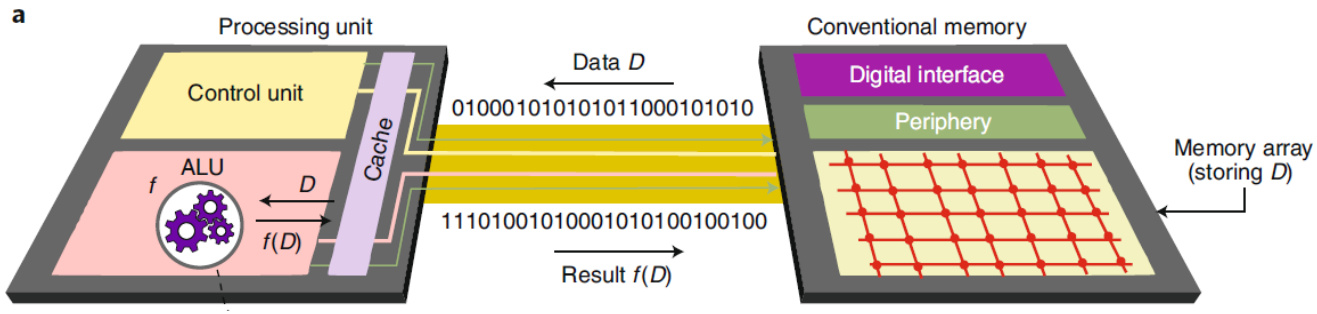
Postgraduate Course in Electronic Circuit Design II
ELEC-L351001

Agnimesh Ghosh

Department of Micro and Nanosciences
Aalto University, School of Electrical Engineering
agnimesh.ghosh@aalto.fi

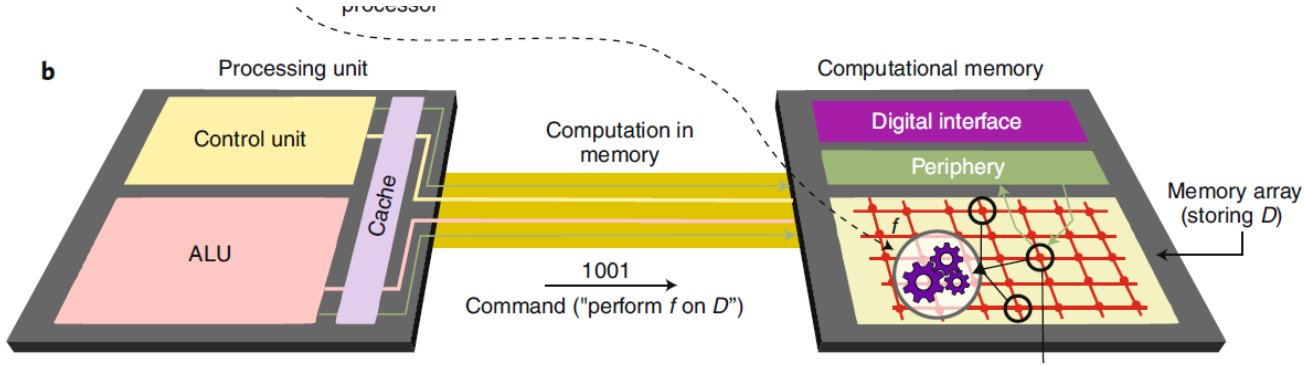
May 18, 2022

Traditional Computation



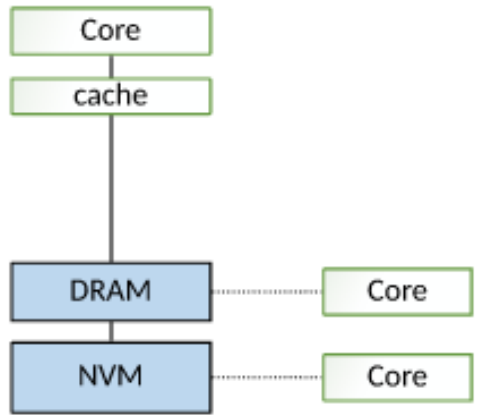
- Computing systems are primarily built based on the **Von-Neumann** architecture
  - During the execution of computational tasks, large amounts of data need to be shuttled back and forth between the CPU and memory units
  - Latency associated with accessing data from the memory units is a key performance bottleneck for a range of applications these days.
  - The energy cost of moving data is another significant challenge given that the computing systems are severely power limited due to cooling constraints and longevity in mobile computing devices.

Compute In-Memory (CIM)



- In-memory computing (CIM) is an alternate approach where certain computational tasks are performed in place in the memory itself.
- At no point during computation, the memory content is read back and processed at the granularity of a single memory.
- The massive parallelism required by signal processing (viz. AI or NN applications) can be achieved by a dense array of millions of memory devices performing computation.

Near-Memory Compute (NMC)



(d) Near-Memory Computing

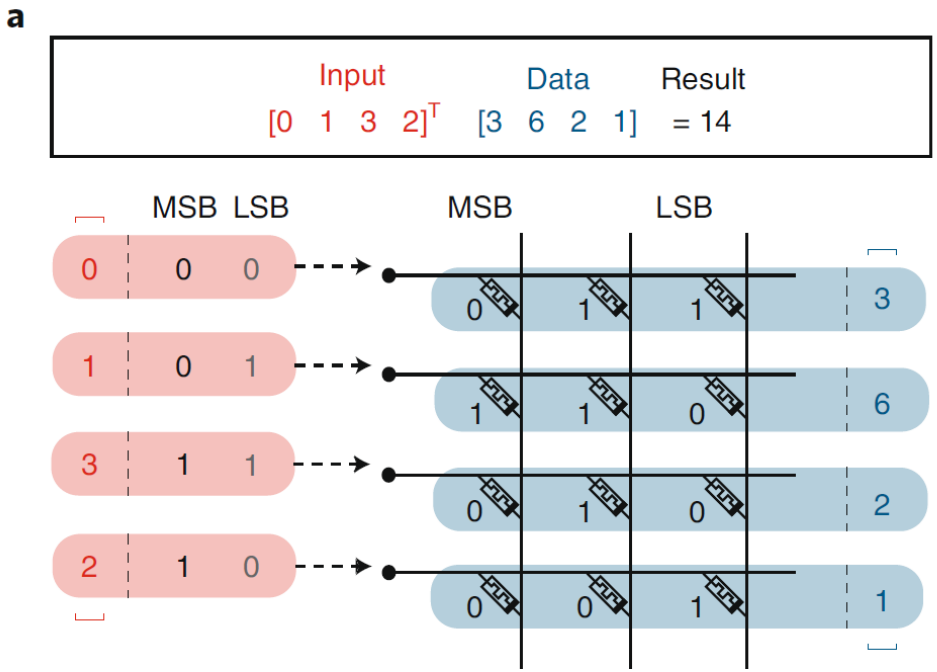
- System architects have been introduced to bridge the memory wall by introducing memory hierarchies that mitigated some of the disadvantages of off-chip DRAMs.
- However limited number of pins on the memory package is not able to meet today's bandwidth demands of multicore processors.
- NMC aims at processing close to where the data resides, which couples compute units close to the data and seek to minimize the expensive data movements, viz. 3-D stacking

Scientific Computing In-Memory

- Linear algebra computational kernels (MVM) are common to scientific computing applications.
- Both memristive (RRAM, MRAM) and charge-based memory (SRAM, DRAM) devices suffer from significant inter-device variability and inhomogeneity across an array.
- The precision of analog MVM operations with these devices is rather low.
- However the accuracy limitation can be mitigated to a certain extent, facilitated by an old technique called **bit-slicing**.
  - Each of the modules processes one bit field or ‘slice’ of an operand.
  - The grouped processing components will then have the capability to process, in parallel, an arbitrarily chosen full word-length of a particular task.

MVM using bit-slicing

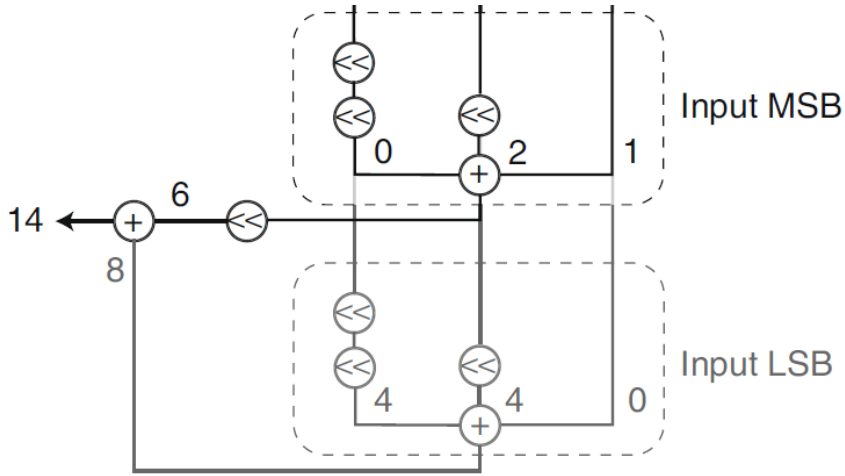
- An n-bit element of the matrix is mapped onto device conductance values of n binary crossbar arrays, that is, n bit slices so as each bit slice contains the binary values of the matrix elements in a particular bit position.



- Similaryly, bit slicing can also be applied to the input vector elements, where each bit slice is input to the crossbar arrays one at a time.

Drawbacks

- Inaccuracy arising from the analog summation along columns, potentially could be more detrimental in larger crossbar arrays.



- Extra peripheral circuitry of the shift and add external reduction networks could substantially increase the energy consumption and area.
- Mixed-precision computing can be used to aid the imprecise MVM to obtain an approximate solution, and then refine this solution based on the residual error calculated precisely through digital computing in an iterative linear solver.

Uses

- The crossbar-based analogue MVM can be used in many applications such as image compression, compressed sensing, combinatorial optimization.
- **Image Compression**
  - Analog image compression uses the idea of encoding a transform matrix, for example, a discrete cosine transform, as the conductance values of devices organized in a crossbar array.
  - Pixel intensities (represented as voltages) are applied to the crossbar first row by row and then column by column.
  - The compression is then performed by keeping only a certain ratio of the highest coefficients of the transformed image and discarding the rest.

Sparse Dictionary Learning

- It is a learning framework in which a sparse representation of input data is obtained in the form of a linear combination of basic elements, which forms the dictionary of features.
- Here both the dictionary and the sparse representation are learned from the input data.
- Dictionary learning requires updating the conductance values by exploiting the accumulative behaviour of the memristive devices, based on viz. stochastic gradient descent.
- However updating of the conductance values are challenging due to device stochasticity and nonlinear conductance change with the number of applied pulses.

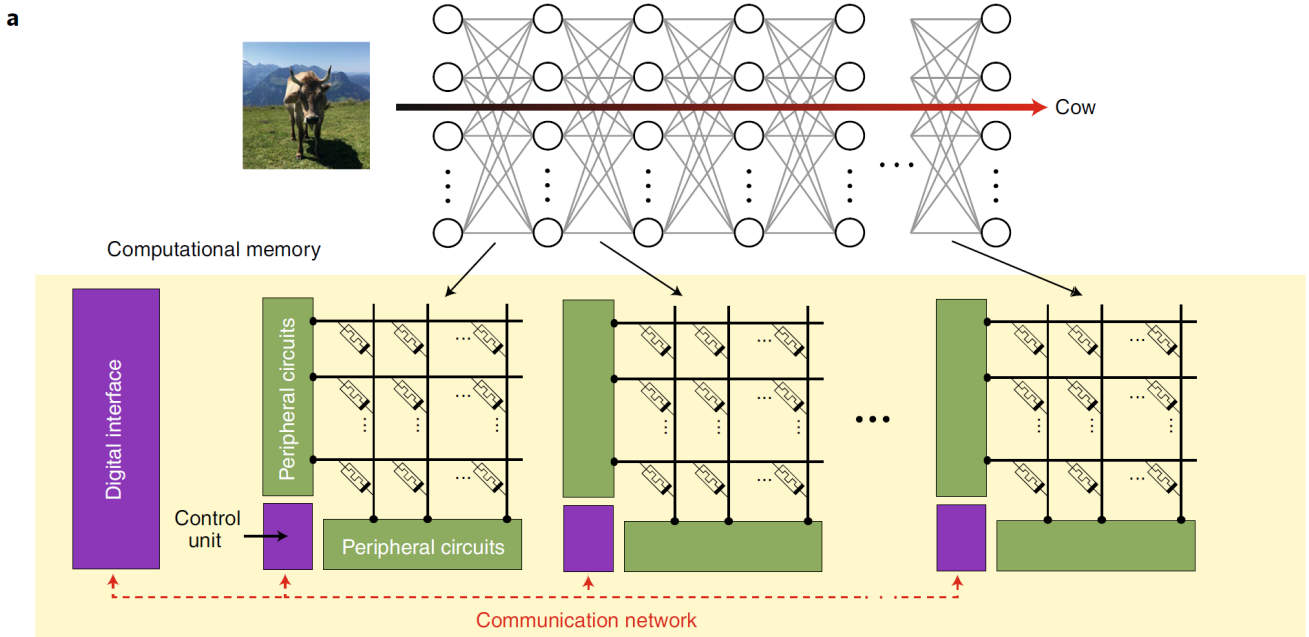
Principal Component Analysis

- It is a dimensionality reduction technique to reveal the internal structure of data by using a limited number of principal components.
- To acheive this, a linear feedforward neural network in which the weights are implemented in a crossbar array.
- The network is optimized via unsupervised learning using Sanger's rule to obtain the principal components, given by the weights connected to each output neuron representing the classes in which the data is clustered.

Deep Artificial Neural Networks

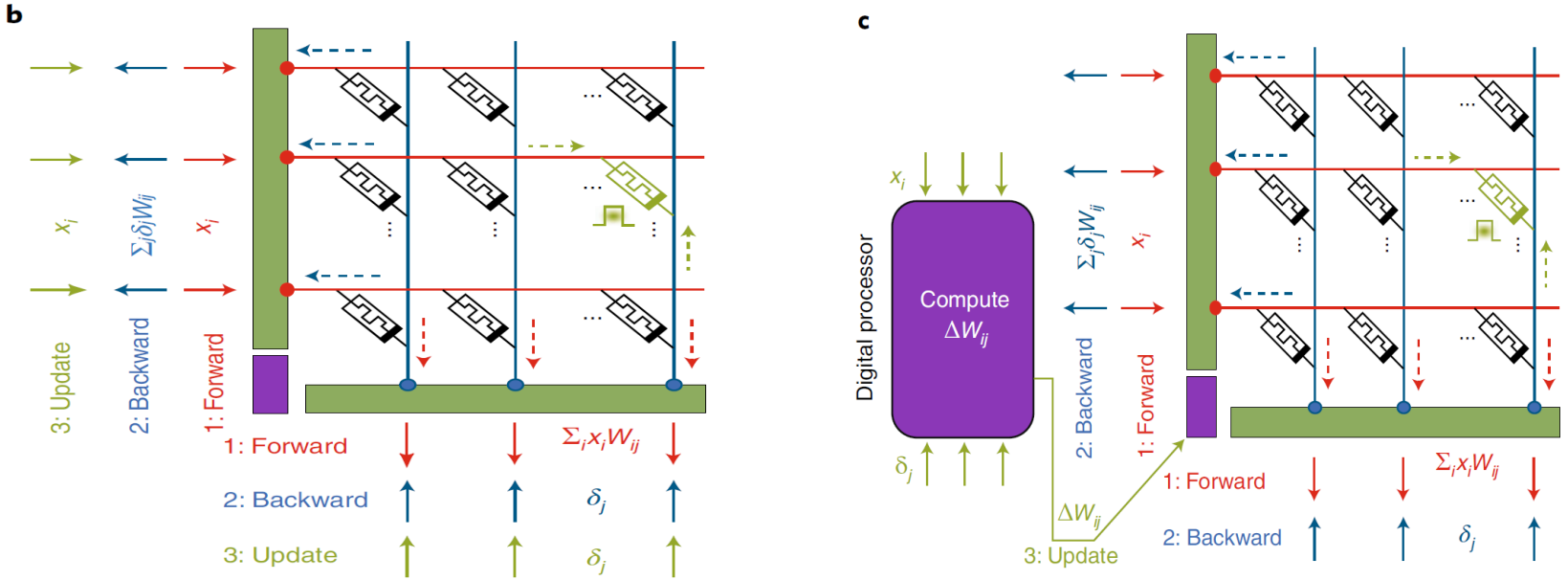
- DNN, loosely inspired by biological neural networks, have shown a remarkable human-like performance in tasks such as image processing and voice recognition.
- A deep neural network (DNN) consists of at least two layers of nonlinear neuron units interconnected by adjustable synaptic weights and by tuning the adjustable weights, for instance, optimizing them by using millions of labelled examples, these networks can solve certain problems really well.
- A DNN can be mapped onto multiple crossbar arrays of memory (The weights of the layers are stored in the charge or conductance state of the memory devices at the crosspoints) devices that communicate with each other.

Deep Artificial Neural Networks



- The propagation of data through that layer is performed in a single step by inputting the data to the crossbar rows and deciphering the results at the columns.
- The Neuron nonlinear function is typically implemented at the crossbar periphery, using analog or digital circuits.

Deep Artificial Neural Networks

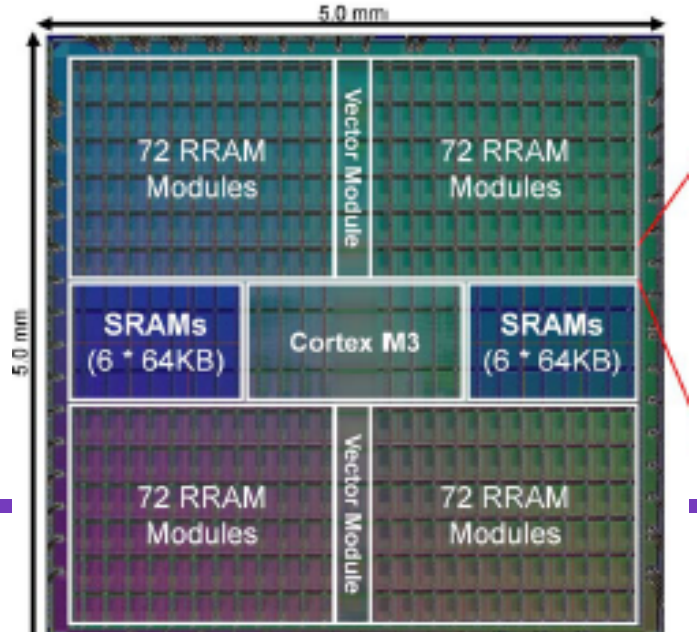


(a) (b)

- In-place weight update can be performed by sending pulses based on the values of activation and error from the rows and columns simultaneously (*a*).
- In (*b*), the weight update can be done in the digital domain and applied via programming pulses to the corresponding devices.

A 40nm 60.64TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems

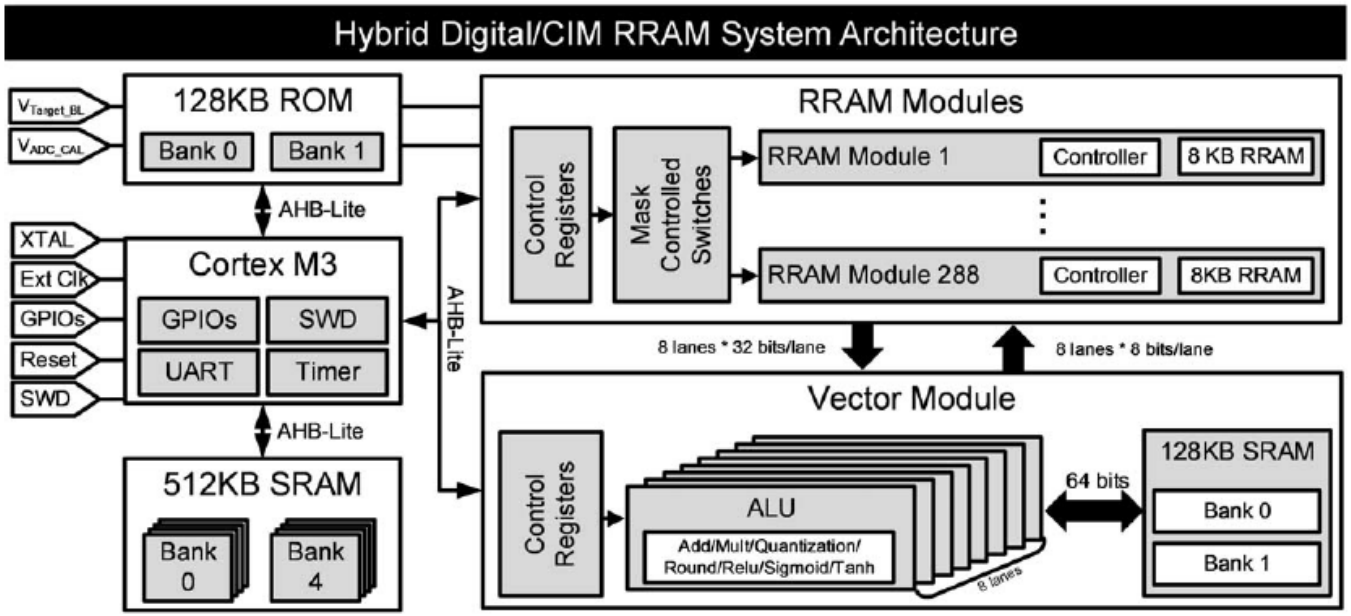
- Author: Muya Chang, Samuel D. Spetalnick, Brian Crafton et al.
- Georgia Institute of Technology, TSMC Corporate Research
- Presents a 2.25MB RRAM based CIM accelerator with 765kB of SRAM and an embedded Cortex M3 processor for edge devices. **9731675**



A 40nm 60.64TOPS/W ECC-Capable Compute In-Memory for Edge Recommendation System

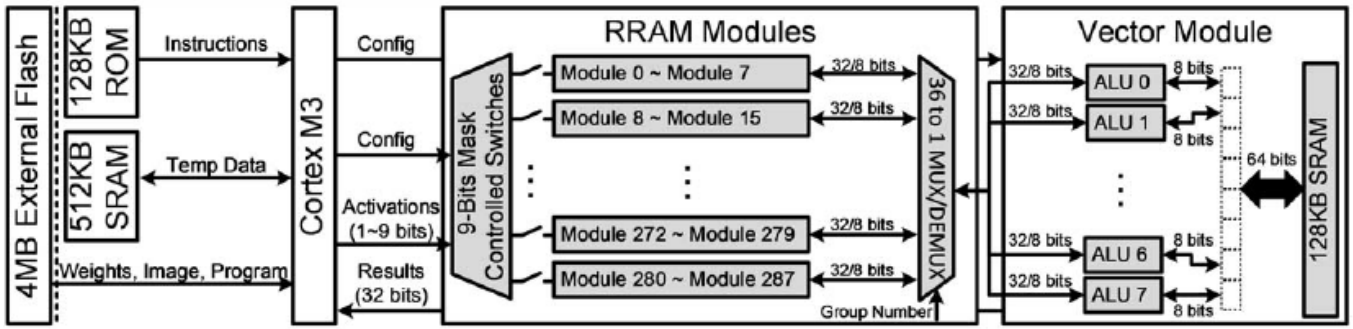
- Applications at the edge, requires higher memory capacity and bandwidth despite irregular data access patterns that prevent effective caching and data reuse.
- These applications rarely run continuously, but instead execution is triggered by events. Which can be facilitated by RRAM given its high density and non-volatility enabling near-zero leakage power and complete power down.
- The Cortex M3 receives events in the form of events and initiates inference on the RRAM processing elements.
- The neural network model is distributed across the RRAM (shallow layers) and the SRAM. The training is performed in SRAM using CMOS SIMD units in the last layer when feedback is received from the user to limit writes to the RRAM.

A 40nm 60.64TOPS/W ECC-Capable Compute In-Memory for Edge Recommendation System



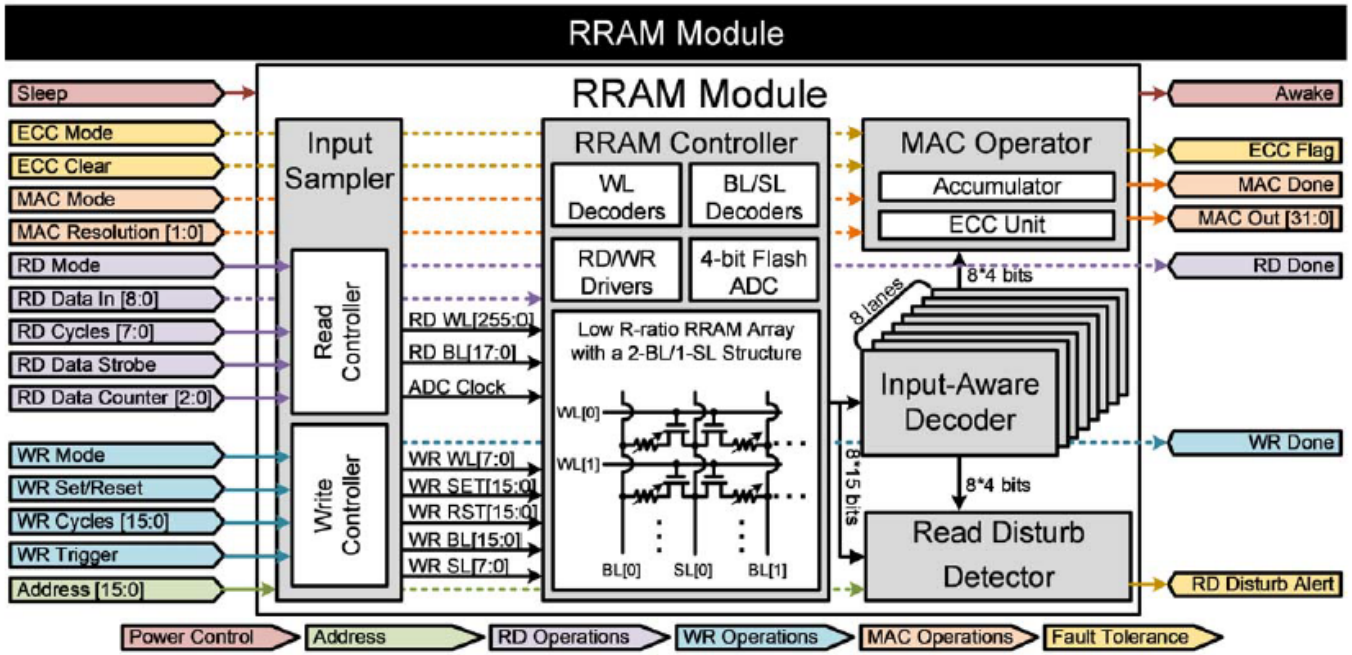
- The M3 sends the first set of inputs to the selected RRAM modules, afterwards the intermediate data is transferred between the RRAM modules and the Vector module to maximize throughput.

A 40nm 60.64TOPS/W ECC-Capable Compute In-Memory for Edge Recommendation System



- RRAM modules are selected based on a 9b mask and a 9b target index
- A fully integrated vector module via AHB-Lite which contains a 128KB SRAM inside to store intermediate results, and 8 sets of ALUs capable of various functions.

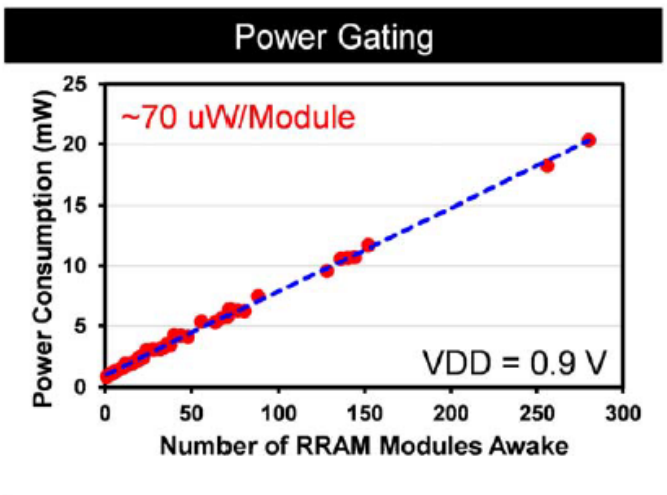
A 40nm 60.64TOPS/W ECC-Capable Compute In-Memory for Edge Recommendation System

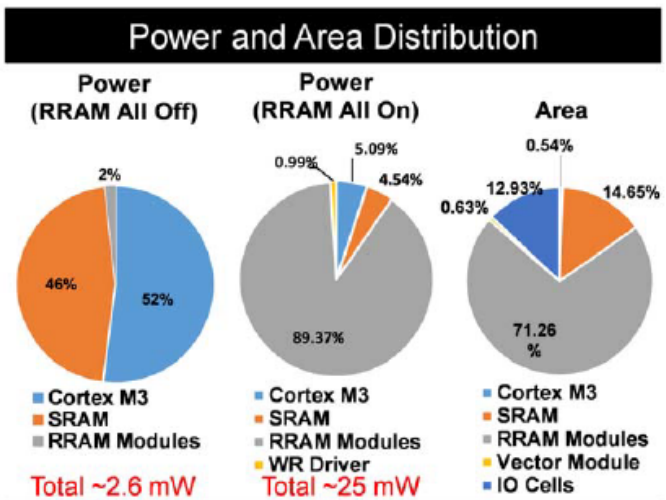


- RRAM module controls operates in different modes: (1) Power control. (2) Targeted address. (3) Read configurations. (4) Write configurations. (5) MAC configurations. (6) Fault-tolerance configurations.

A 40nm 60.64TOPS/W ECC-Capable Compute In-Memory for Edge Recommendation System

- To mitigate inherent device variation in RRAM that yields sum-of-products errors in CIM, a single error detection and single error correction (SECSED) ECC scheme is implemented in this.
- The error is minimized by reading one word line at a time to avoid SOP error at the expense of latency inclusion.
- A set of dedicated power gates are integrated inside each RRAM module to make the system low powered.



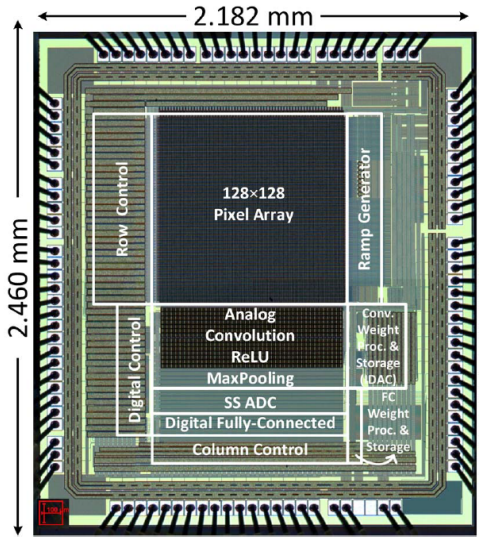


A 40nm 60.64TOPS/W ECC-Capable Compute In-Memory for Edge Recommendation System

	ISSCC 2018 [2]	ISSCC 2019 [3]	ISSCC 2020 [4]	ISSCC 2020 [5]	ISSCC 2019 [6]	ISSCC 2021 [7]	This work
Technology	65 nm	55nm	22nm	130nm	28nm	40nm	40nm
Memory	RRAM	RRAM	RRAM	RRAM	SRAM	RRAM	RRAM
Supply	1.0 V	1.0 V	0.7-0.9 V	1.8 V	0.6-1.1 V	0.9 V	0.9 V
Sensing Mode	Current	Current	Current	I&F	N/A	Voltage	Voltage
Low R-ratio CIM Arch.	No	No	No	No	N/A	Yes	Yes
Write Verification	No	No	No	No	N/A	Yes	Yes
Fault Tolerant	No	No	No	No	N/A	Yes	Yes
ECC	No	No	No	No	N/A	No	Yes
Resolution (IN/W/OUT)	N/A (OUT: 1-3b)	1-2b/3b/3b	1-4b/2-4b/6-11b	1b/analog/1b	Integer & floating point	1-8b/1-8b/20b	1-8b/1-8b/32b
Embedded Microprocessor	No	No	No	No	Yes	No	Yes
On-Chip RRAM/SRAM	128KB/0	128KB/0	256KB/0	8KB/0	0/128KB	8KB/0	2.25MB/768KB
External Flash Support	No	No	No	No	N/A	No	Yes
Energy Efficiency (TOPS/W)	19.2	53.17	121.38	148	0.55	56.67	60.64

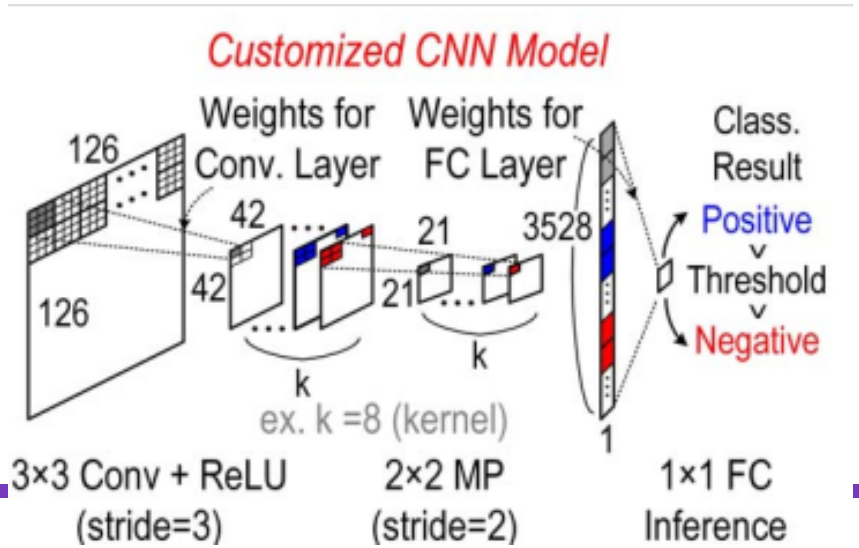
A 0.8V Intelligent Vision Sensor with Tiny Convolutional Neural Network and Programmable Weights Using Mixed-Mode Processing-in-Sensor Technique for Image Classification

- Author: Tzu-Hsiang Hsu, Guan-Cheng Chen et al.
- National Tsing Hua University
- Presents an intelligent vision sensor (IVS) with an embedded tiny CNN model and programmable weights to achieve configurable feature extraction and on-chip image classification using a mixed-mode processing-in-sensor (PIS) technique. **9731679**



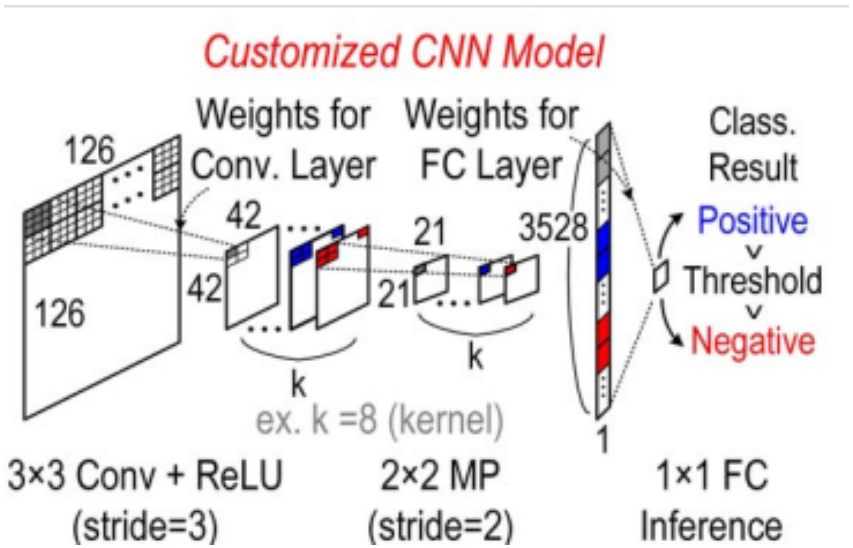
A 0.8V Intelligent Processing-in-Sensor for Image Classification

- Traditional convolutional CIS is inadequate for some tasks, due to the limitation on the numbers of layers/kernels (needs digital accelerator for Rectified Linear Unit: ReLU, Maximum- Pooling: MP, Fully-Connected layer: FC, etc.).
- This IVS consists mixed-mode PIS circuits which includes a 3×3 convolution layer with adjustable kernel number for feature map computation, a 2×2 MP layer for down sampling, and a 1×1 FC layer for inference.

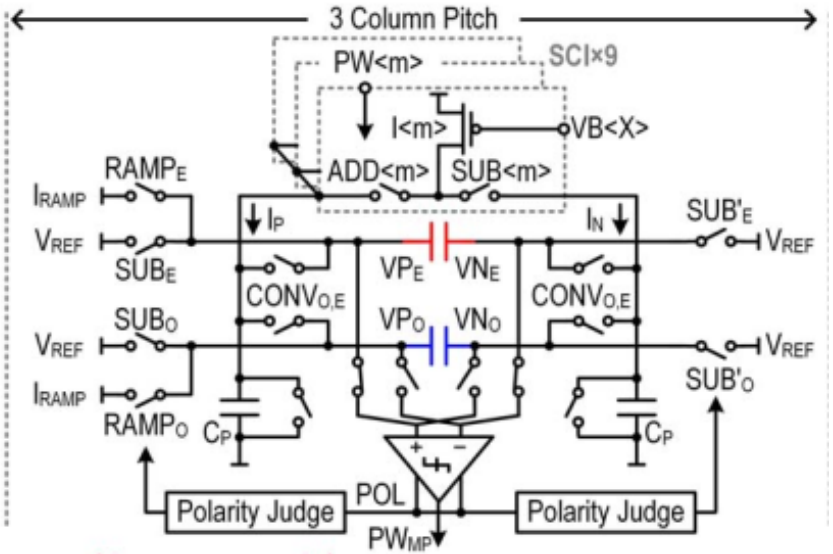


A 0.8V Intelligent Processing-in-Sensor for Image Classification

- Convoltuion operation is realized using the signal-dependent pulse width and weight-dependent current level. The result is realized by checking the polarity (POL) of VP and VN, followed by the corresponding level shifting on capacitor CM for signal subtraction.

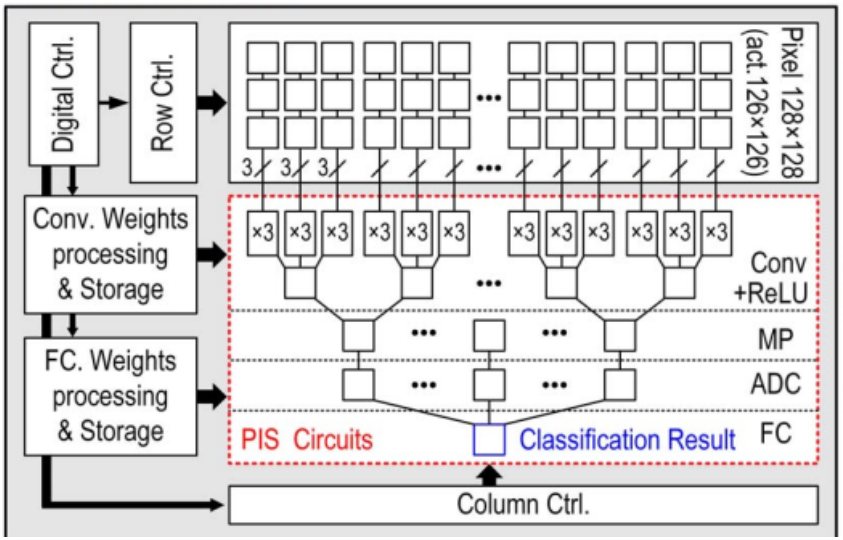


A 0.8V Intelligent Processing-in-Sensor for Image Classification



- ReLU can be realized by omitting level shifting to set CR = 0 when POL = 0 (VP < VN)

A 0.8V Intelligent Processing-in-Sensor for Image Classification



- With CNN processing architecture, multiple-scale image classification can be achieved using the same Conv weights with a hardware windowing operation. Resulting an accuracy of above 90% with detection windows of 84×84 and 66×66.

A 0.8V Intelligent Processing-in-Sensor for Image Classification

	[2] 2021 ISSCC	[3] 2021 VLSI	[4] 2017 ESSCIRC	[6] 2020 Sensors	This Work
Process	65nm CMOS	180nm CIS	65nm Logic CMOS	110nm CIS	180nm CMOS
Supply	0.8V ~ 1.2V	(Analog) 2.5V (Digital) 1.8V	(Analog) 2.5V (Digital) 0.5~0.8V	(Analog) 3.3V (Digital) 1.5V	(Analog) 0.8V (Digital) 0.8V
Die Area	2 mm x 2 mm	5.2 mm x 4.1 mm	3.3 mm x 3.3 mm	5.9 mm x 5.2 mm	2.46 mm x 2.18 mm
Pixel Size	9 μm	9.8 μm	7 μm	N.A.	7.6 μm
Pixel Type	40T log(I)-V Pixel + 1 MIMCAP	14T+4C pinned-PD	N.A.	4T APS	4T PWM
Fill Factor	12.9%	20.1%	N.A.	N.A.	36%
Array Size	160x128	240x240	320x240	160x120	126x126
Frame Rate	24~268 fps	120 fps (Global shutter)	1fps	120 fps	250 fps
In-sensor-processing Tasks	Haar-like filtering	Log-Haar-like filtering + Classifiers	Haar-like filtering (Integrate Digital Vision Processor on-chip)	Convolution, ReLU, MaxPooling, Fully-connected	Convolution, ReLU, MaxPooling, Fully-connected
Processing Algorithm	Viola-Jones 2-stage cascade classifier	25 Machine Learning feature classifier	Viola-Jones (3Ana.+20Dig.)-stage cascade classifier	5-layers CNN	3-layers CNN
Weight Precision	6 scales kernel 1.5b (+1,0,-1)	Multiscale kernel 1.5b (+1,0,-1)	3 scales kernel 1.5b (+1,0,-1)	(Conv.) 2x2 kernel (FC) 5b	(Conv.) 3x3 kernel, ±3b (FC) 1.5b (+1,0,-1)
FD Task *Acc. / Pre. / Rec.	> 80% / N.A. / N.A. window rejection (Need digital backend)	98% / N.A. / N.A. window rejection (Need digital backend)	> 90% / N.A. / N.A. (w/ on-chip digital processor)	89.3% / 94.7% / 72% (w/o digital backend)	93% / 90.4% / 97.2% (w/o digital backend)
Dataset	KODAK	FERET	N.A.	N.A.	LFW / Kaggle Oregon Wildlife
Power	42~206 μW	2.9 mW @ 120 fps	60 μW @1 fps (averaged)	0.96 mW @ 60 fps 1.12 mW @ 120 fps	80.4 μW @ 50 fps 104.1 μW @ 125 fps 134.5 μW @ 250 fps
iFoM	2.5~103.9 pJ/pix·fps	419 pJ/pix·fps	781.2 pJ/pix·fps	833 pJ/pix·fps @ 60fps 486 pJ/pix·fps @ 120fps	101.2 pJ/pix·fps @ 50fps 52.4 pJ/pix·fps @ 125fps 33.8 pJ/pix·fps @ 250fps

Conclusion

- There are multiple benefits in the applications of signal processing, AI, ML, DL that can be leveraged through in-memory/ near-memory computation in order increase syatem performance.
- In case of in-memory computing for MVM, it is preferable or the application to perform many MVMs on large squarish and dense matrices that stay constant throughout its execution.
- Drawbacks
  - Charge based analog computation is inherently subject to thermal noise, which sets an upper limit to the precision achievable for a given capacitor size and ambient temperature.
  - In case of memristive devices, these are write variability and conductance variations, which lies mostly in the stochastic nature of filamentary switching.
  - A critical issue is the need for digital-to-analogue (analog-to-digital) conversion every time data goes in to (out of) the crossbar arrays, which determines the precision of MAC operations.

Conclusion

- For computational tasks in resistive memories involving read-only operations, such as MVM, endurance is much less critical as long as the conductance states remain unchanged during their execution.
- Existing technologies in charged based memories consumes higher power in analog computation.
- Besides the conventional memory devices , several new memory concepts are being proposed for in-memory computing viz. PCM, Photonic memory devices.
- Even though promising, it is difficult to fully assess their benefits in the absence of large-scale demonstrations and/or integration with current CMOS technology.
- Due to plateauing of Moore graph, we might have to sustain older technology nodes but instead equip the chips with high performance.

References

- Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. et al. Memory devices and applications for in-memory computing. Nat. Nanotechnol. 15, 529–544 (2020).
- Singh, G., Chelini, L., Corda, S., Awan, A. J., Stuijk, S., Jordans, R., Corporaal, H., amp; Boonstra, A.-J. (2019). Near-memory computing: Past, present, and future. Microprocessors and Microsystems, 71, 102868. https://doi.org/10.1016/j.micpro.2019.102868
- H. Hsu et al., "A 0.8V Intelligent Vision Sensor with Tiny Convolutional Neural Network and Programmable Weights Using Mixed-Mode Processing-in-Sensor Technique for Image Classification," 2022 IEEE International Solid- State Circuits Conference (ISSCC), 2022
- M. Chang et al., "A 40nm 60.64TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems," 2022 IEEE International Solid-

Assignment

- Draw a citcuit diagram of the following MVM operation using bit-slicing

$$\begin{bmatrix}1 & 3 & 0 & 2\end{bmatrix}^T \begin{bmatrix}5 & 2 & 6 & 1\end{bmatrix}$$