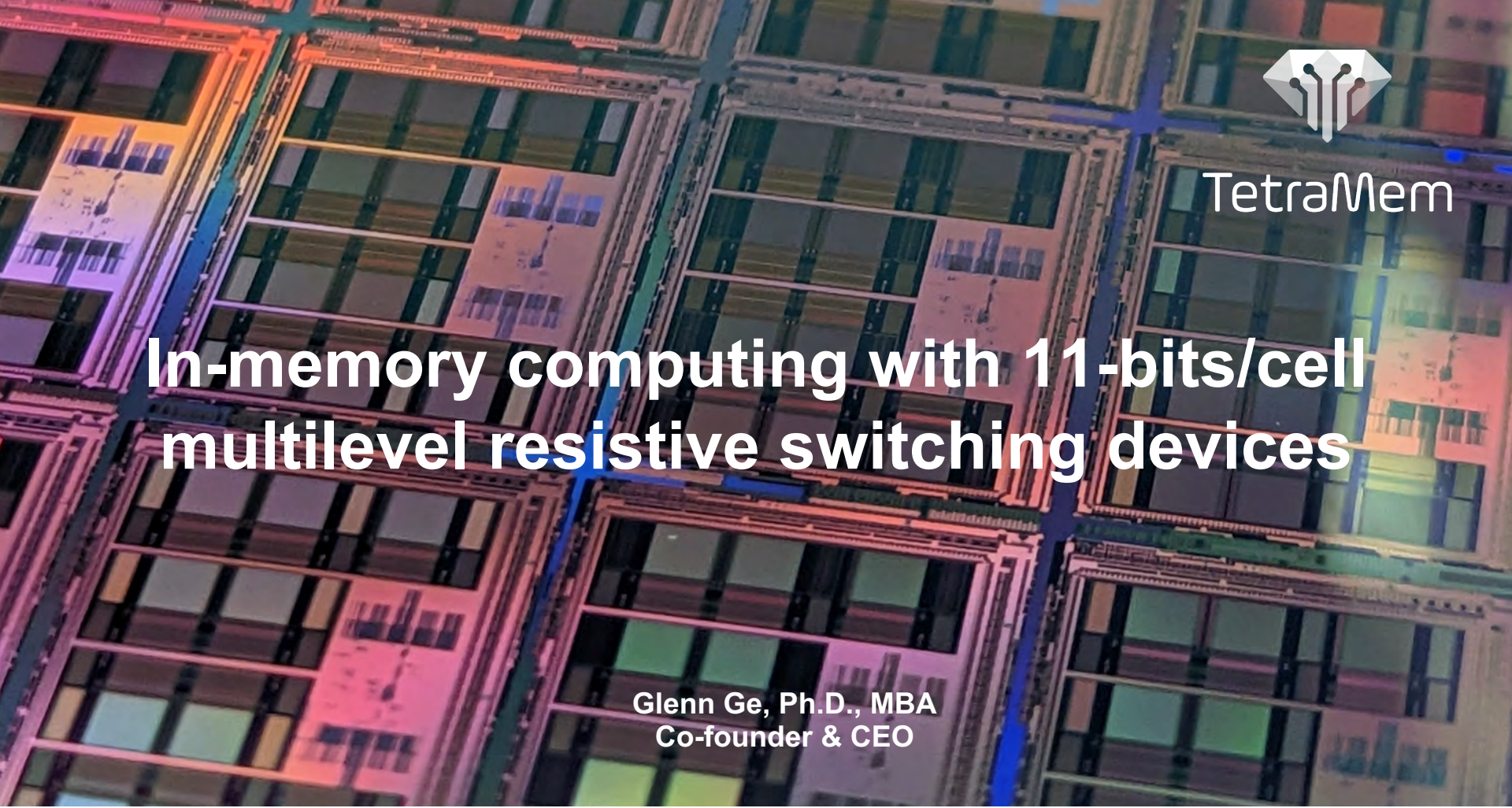


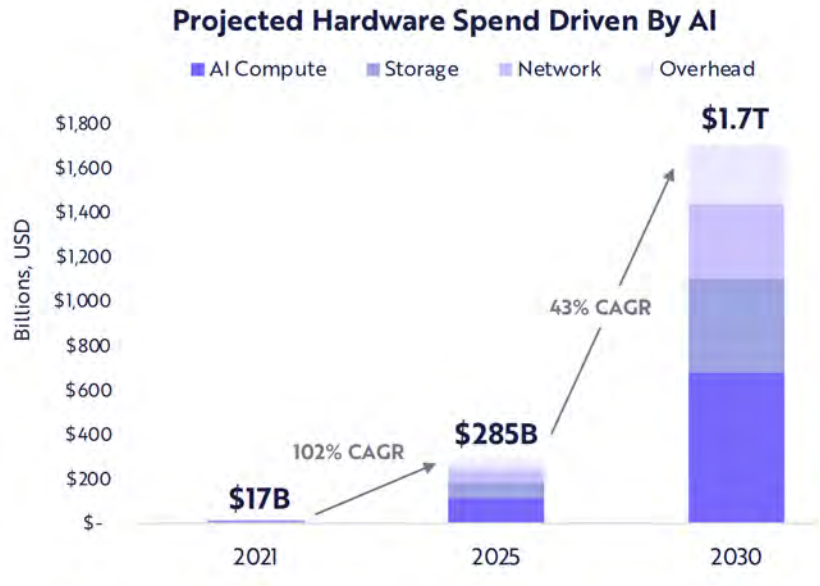
In-memory computing with 11-bits/cell multilevel resistive switching devices

Glenn Ge, Ph.D., MBA
Co-founder & CEO

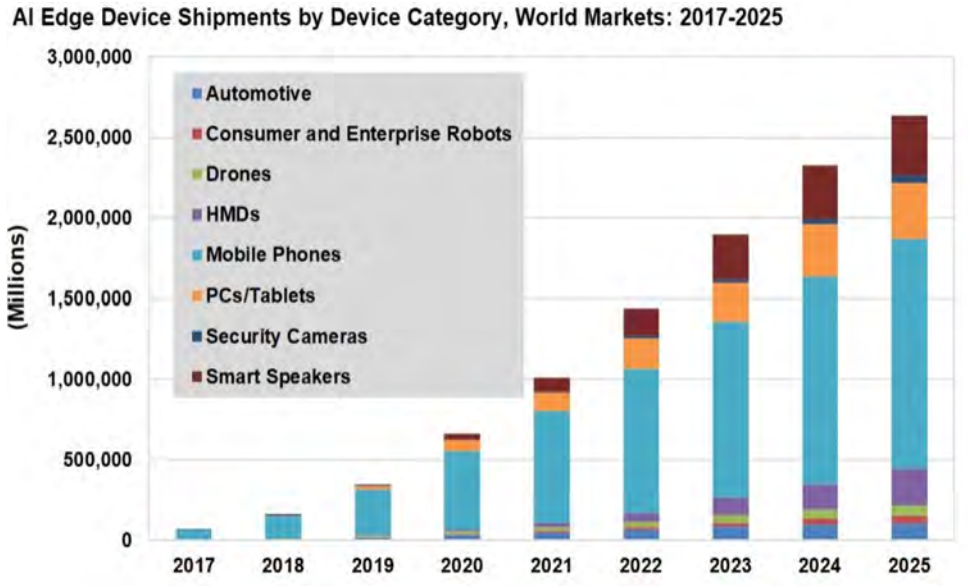
The AI and Edge AI Chip Market

Flash Memory Summit

AI applications will add $30 trillion to the global equity market capitalization during the next 2 decades and AI chip market will reach > $20 billion at 2025 with 33% CAGR



Source: ARK Big Idea 2022

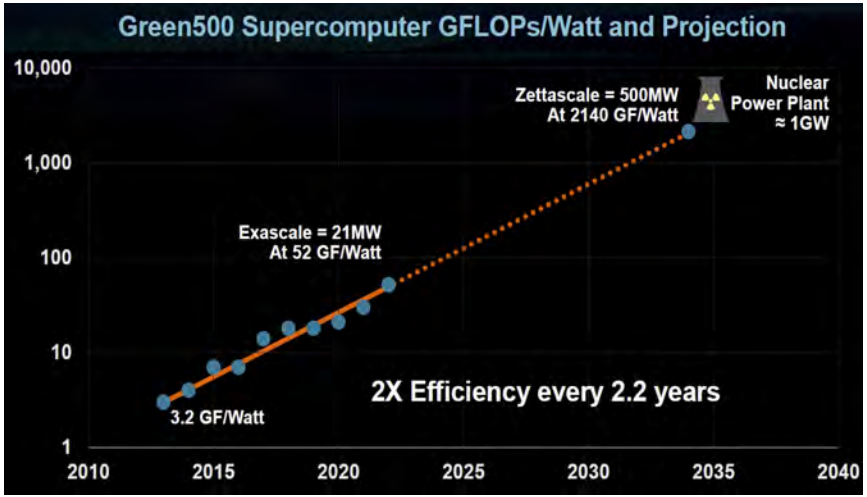


Source: Tractica

Customer Pain-points From Edge To Data Center

- Digital solutions **Could NOT** satisfy many edge AI applications
  - Example: Gaze-eye tracking for AR/VR headset, 200 FPS, <5ms latency, 1.0° accuracy, <10 mW power
    - Advanced nodes (<7nm) and SRAM are too expensive
    - Mature nodes (>28nm) could not meet performance
  - Example: Future Level 4 & Level 5 autonomous driving need > 300~4000 TOPS computing power
    - Scaling of SRAM as the primary impetus for technology advancement has experienced a significant slowdown for technology nodes below 5nm.
    - Automotive SoC power efficiency is capped ~ 3 TOPS/W for all the current digital solutions

- Digital solutions **Could NOT** fuel future Data Center applications as well



AMD's plenary talk, ISSCC 2023

One training = 5 mid-size cars life span CO_2 emission!!

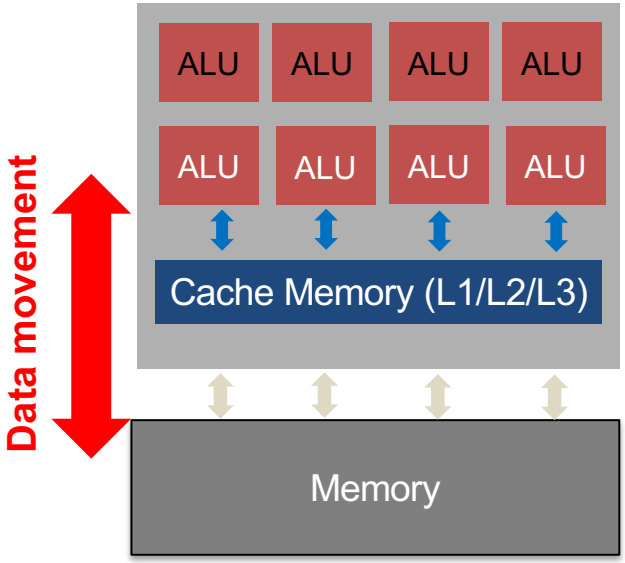
- Google, "The Evolved Transformer", 2019
- Estimated training energy: 656,347 kwh based on P100 x8 with 1.5kW and 274,120 hrs => **284 tons** CO_2 emission!

Emma S. et. al. Energy and Policy Considerations for Deep Learning in NLP, 2019

In-Memory Computing (IMC) – Most Fit Solution for AI Computing

Flash Memory Summit

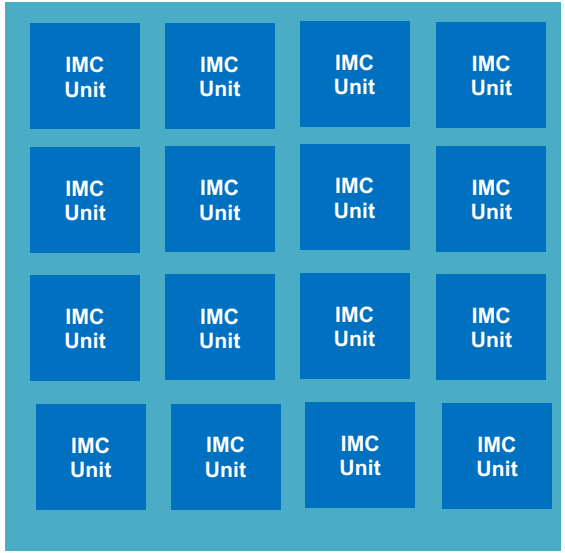
Traditional von Neumann Architecture



Limited by 0.1-5 TOPs/W

- ***Memory wall: data movement bottleneck***
  - ***SRAM: 10-100TB/s; DRAM: 40GB-3TB/s***
  - ***AI need: 1PB/s***
- ***Limited computer cores with high cost of memory access cost (30-1000X memory access energy than computing)***
- ***Clock limitation for computing speed***

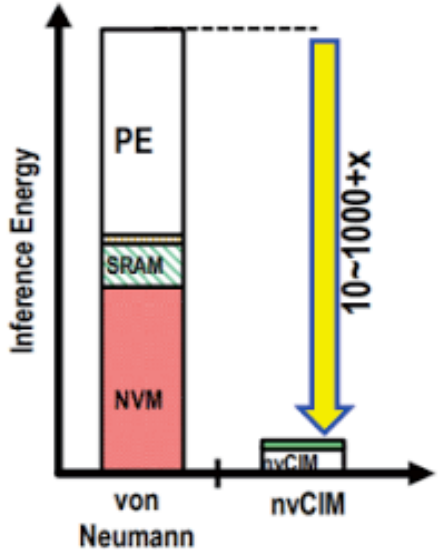
IMC Architecture



Demonstrated > 100 TOPs/W

- ***Data processed in the same physical location as it is stored with minimum intermediate data movement & storage => low power consumption***
- ***Massive parallel computing process by cross-bar array architecture with device-level grain cores => high throughput***
- ***Computing by physical laws (Ohm's law and Kirchhoff's current law) => low latency***

Superior architecture, but right device is the key



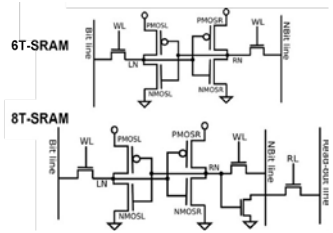
Tang etc. 2019 Symposium on VLSI Circuits

Note: nvCIM = nonvolatile compute in memory,
PE: processing element

Current Memory Device Main Limitations For Computing Applications

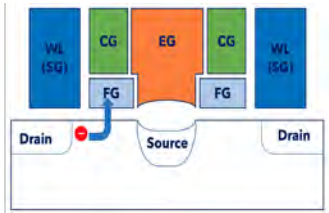
Flash Memory Summit

SRAM



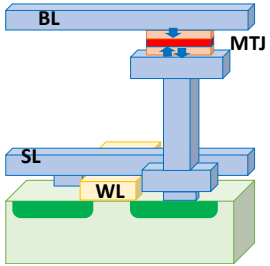
- Volatile binary device
- Accompany storage memory needed
- Min. 6T/cell, expensive
- Speed vs power tradeoff

NOR-FLASH



- Expensive flash memory process
- Scalability issue for <28nm
- High operation voltage
- Long program & erase time
- Retention issue for MLP

STT/SOT MRAM

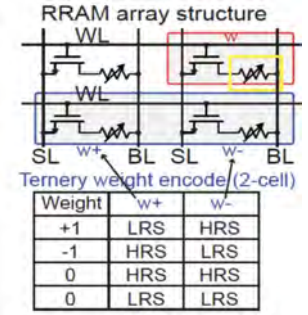


- Complicated Magnetic Tunnel Junction (MTJ) stack structure
- Yield and cost challenge
- Multi-level challenge
- Low off/on resistance ratio

PCRAM

- Wide distribution of the SET states
- Conductance drift
- Temperature-induced conductance variations
- Scaling disturbance issue

Other ReRAMs



VLSI, 2019

- Limited multi-level (Binary or <6 bits/cell)
- Retention issue for MLP
- Endurance issue
- I-V linearity issue

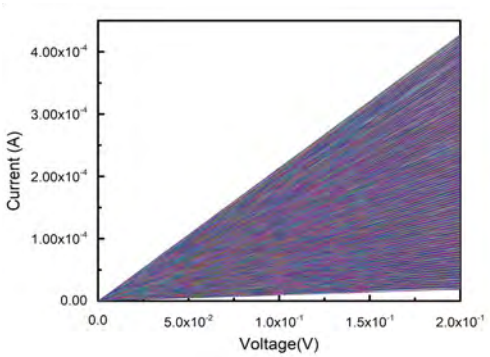
- Each memory device can find its own application space with its unique set of the device performance attributes.
- However, none of the current memory devices can meet the ideal requirements for the IMC. Therefore at TetraMem, we built the solution from material and device level.

Computing Memristor Optimized For Computing Application

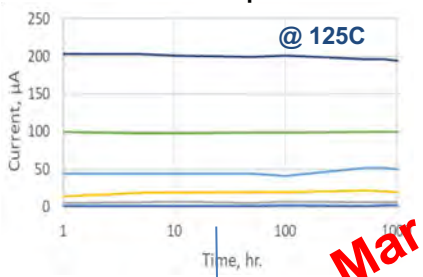
Devices designed for computing applications with superior multilevel, linearity, retention, endurance, uniformity, and etc. key metrics.

Flash Memory Summit

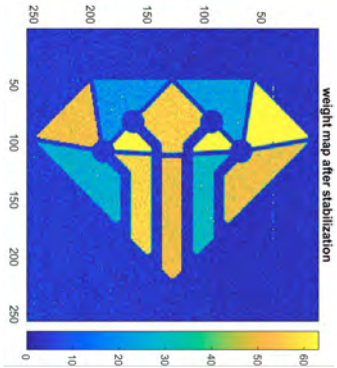
11 bit /cell multi-level programming

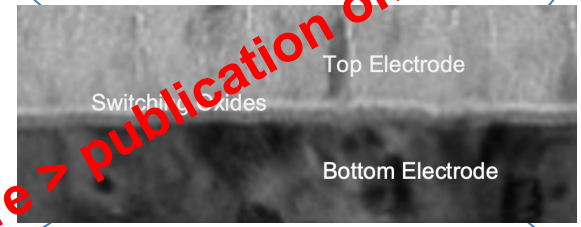


Superior retention and stability at elevated temperature

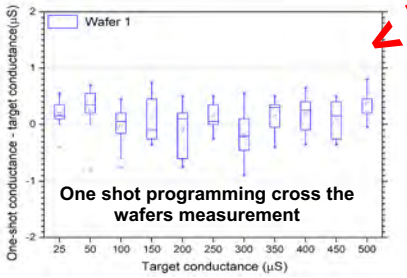


Neural network weight (memory conductance) mapping in 256x256 array

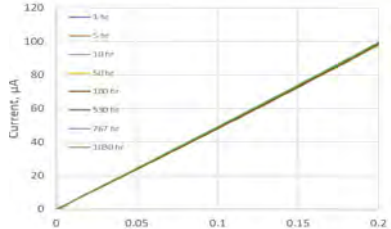




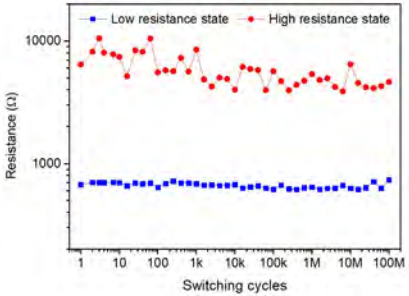
Excellent device uniformity



Excellent I-V linearity



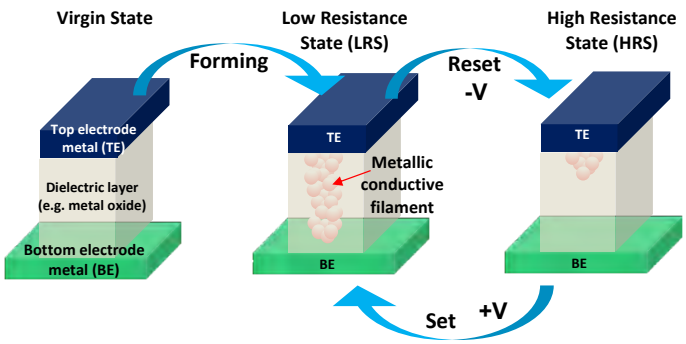
Superior memory grade endurance with > 100 million cycle (10^8)

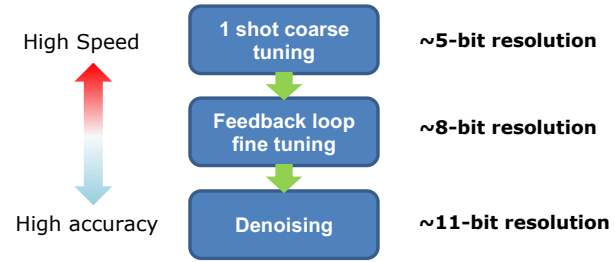


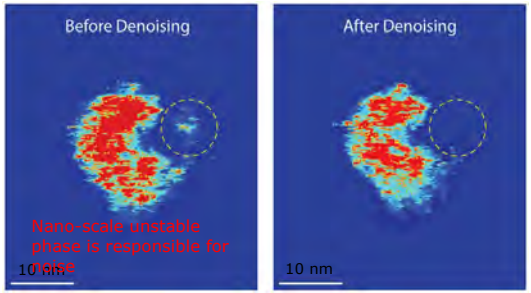
< Nature > publication on Mar 2023

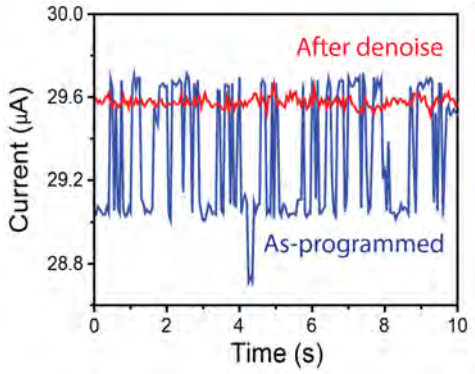
"Thousands of conductance levels in memristors monolithically integrated on CMOS", Nature, Mar 2023, https://rdcu.be/c8GWo

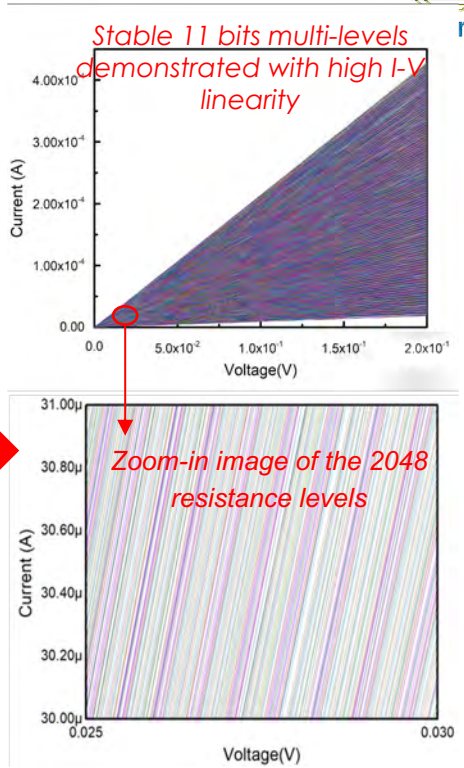
Memristor Working Principle & The Secret of World-record Accuracy On Top of Material & Device Engineering











- Just as the functioning of synapses and neurons is driven by the intricate movement of ions, so too is the main operation of memristors
- By control of the conductive filament size, ion concentration and height, different multi-levels for the cell resistance can be precisely achieved

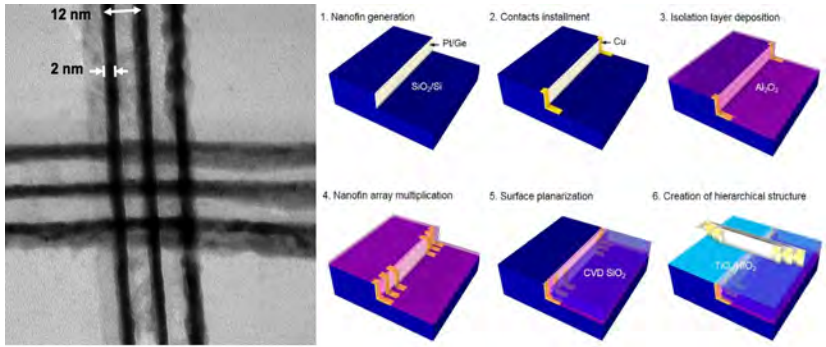
Scalable Technology For Scaling and Backend 3D Stacking

Flash Memory Summit

The device has very scalable device roadmap to support future sub-10nm and 3D integration development.

Go small

Scaling: 2×2 nm² Memristor in a Crossbar Array
- Smallest working devices in electronic circuits



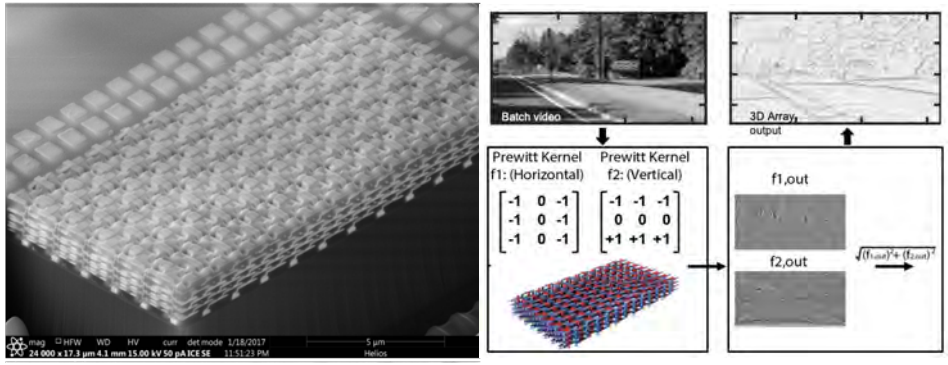
6-nm half-pitch; 4.5 Tbit/in² ; <100 Ω/μm wire resistance

'NANO' pattern written into the array

Nature Nanotechnology 14, 35-39 (2019) . Highlighted in the Editorial "*Testing memory downsizing limits*".

Go 3D

Stacking: Eight Layers of Memristor Crossbars
- A 3D AI processor



Eight Layers of Computing Memristor Crossbars with 300 nm linewidth

3D Convolutional Neural Network (CNN) video processor

Nature Electronics. Apr 2020, cover page of the issue

From Device To Chipset With Key Differences From Memory Applications

Flash Memory Summit

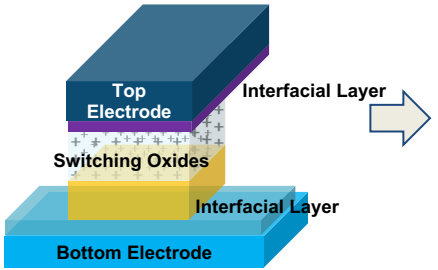
Memristor One Resistor (1R) Device
- Proprietary backend computing memory device
- 2-3 or a few more mask layers
- Commercial fab available, Gen-2 device scalable to support sub-10nm

One Transistor One Resistor (1T1R) Cell
- Standard CMOS frontend transistor & circuit
- IMC performance using mature technology node at backend of line (BEOL)
- All mature circuit and design, no need immature ***Selector*** device

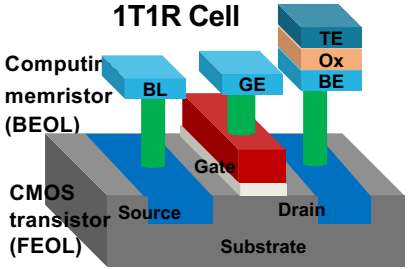
1T1R Crossbar Array (NPU MAC Function Engine core)
- Each cross-point includes a 1T1R cell
- 256x256 array size (-> 1k x 1k in future) with driving circuits
- Scalable IP core design

Chipset (SoC & Chiplet)
- Multiple NPUs with other functional blocks
- Packaged format (ASIC/SoC) or chiplet
- Comprehensive software support

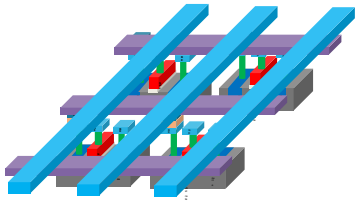
Computing memristor



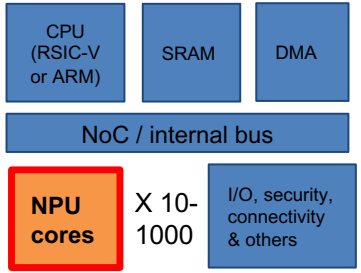
1T1R Cell



1T1R Crossbar Array (NPU IP core)



Chipset (SoC & Chiplet)



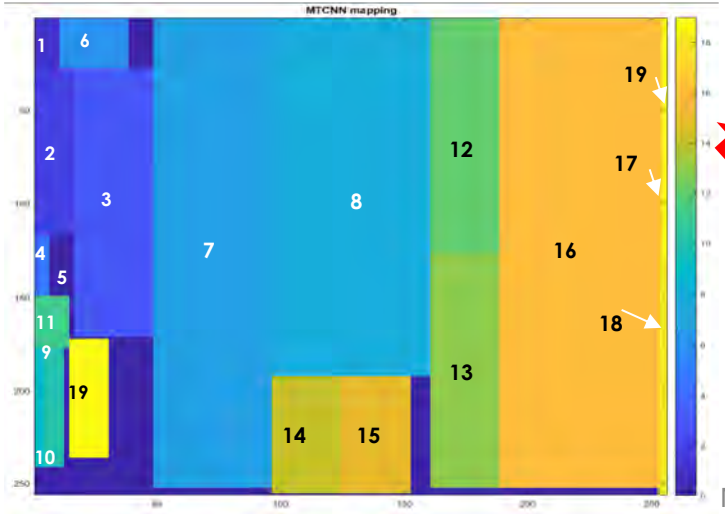
Key differences	Memory Memristor Not ready for large applications yet	Computing Memristor Ready to go!
Economical structure	Cost per cell sensitive with massive memory devices used	Cost per cell insensitive with limited devices for computing
Device technology maturity	• Very large 1S1R cross-bar array with immature selector • Stringent uniformity/variability/defect requirements • Need high programming speed, low energy and endurance	• 1T1R cross-bar with mature transistors • Less requirements on uniformity/variability/defect • Less demanding for programming speed, energy and endurance for AI inference applications
Competing technologies	Many strong contenders like DRAM, SRAM, Flash etc.	Other competing technologies have various limitations

Pratical Neural Networks Mapping Example

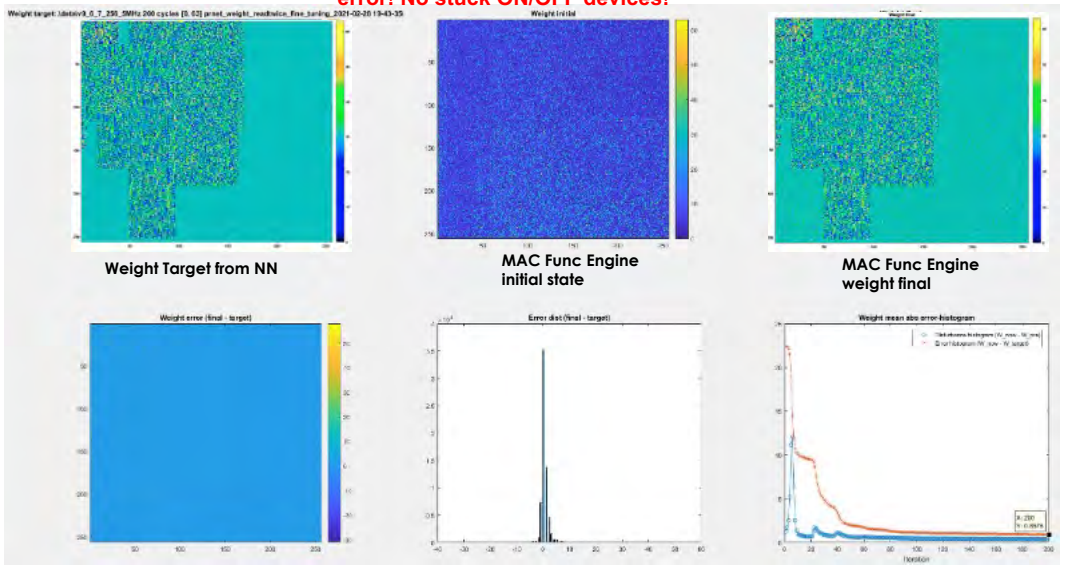
Multi-task Cascaded Convolutional Networks (MTCNN) Pnet & Rnet weight programming example for practical VMM

Index	type	W	L	Cin	Cout	Stride	Image _W	Image _L	Iteration	cycle per iter.	xbar. num	Time	xbar _WL	xba r. BL	Device	SW time (s)	Image Group	Digital time (s)
1	Conv	3	3	3	10	2	320	240	76800	62	8	5.95E-03	27	10	270	2.56E-06	1	4.80E-03
2	Conv	3	3	10	16	2	160	120	19200	131	8	3.14E-03	90	16	1440	2.56E-06	1	1.20E-03
3	Conv	3	3	16	32	2	80	60	4800	201	8	1.21E-03	144	32	4608	2.56E-06	1	3.00E-04
4	Conv	1	1	32	2	1	40	30	1200	59	8	8.85E-05	32	2	64	2.56E-06	1	7.50E-05
5	Conv	1	1	32	4	1	40	30	1200	61	8	9.15E-05	32	4	128	2.56E-06	1	7.50E-05
6	Conv	3	3	3	28	2	24	24	576	80	8	3.46E-03	27	28	756	2.56E-06	60	2.16E-03
7	Conv	3	3	28	48	1	12	12	144	325	8	3.51E-03	252	48	12096	2.56E-06	60	5.40E-04
8	Conv	2	2	48	64	1	12	12	144	281	8	3.03E-03	192	64	12288	2.56E-06	60	5.40E-04
9	FC	1	1	128	2	1	12	12	1	155	8	1.16E-05	128	2	256	2.56E-06	60	3.75E-06
10	FC	1	1	128	4	1	12	12	1	157	8	1.18E-05	128	4	512	2.56E-06	60	3.75E-06
11	Conv	3	3	3	14	2	48	48	2304	66	8	1.90E-03	27	14	378	2.56E-06	10	1.44E-03
12	Conv	3	3	14	28	1	24	24	576	179	8	1.29E-03	126	28	3528	2.56E-06	10	3.60E-04
13	*Conv	3	3	14	28	1	12	12	144	179	8	3.22E-04	126	28	3528	2.56E-06	10	9.00E-05
14		3	3	7	28	1	12	12	144	116	8	2.09E-04	63	28	1764	2.56E-06	10	9.00E-05
15		3	3	7	28	1	12	12	144	116	8	2.09E-04	63	28	1764	2.56E-06	10	9.00E-05
16	Conv	3	3	28	64	1	12	12	144	341	8	6.14E-04	252	64	16128	2.56E-06	10	9.00E-05
17	FC	1	1	128	2	1	40	30	1	155	8	1.94E-07	128	2	256	2.56E-06	10	2.56E-05
18	FC	1	1	128	4	1	40	30	1	157	8	1.96E-07	128	4	512	2.56E-06	10	5.12E-05
19	FC	1	1	128	10	1	40	30	1	163	8	2.04E-07	128	10	1280	2.56E-06	10	1.28E-04
Total												2.51E-02			61556			1.21E-02
				CD (b	50	CLK (s)	1.00E-0	Analog cy		25	FPS(P	39.8			65536		FPS(co	26.91

* Conv with size larger than 256, or encounter space limitation, need to be partitioned into two or more pieces and handle sequentially.



Actual 256x256 Crossbar Array Programming Results with < 2 mean error! No stuck ON/OFF devices!

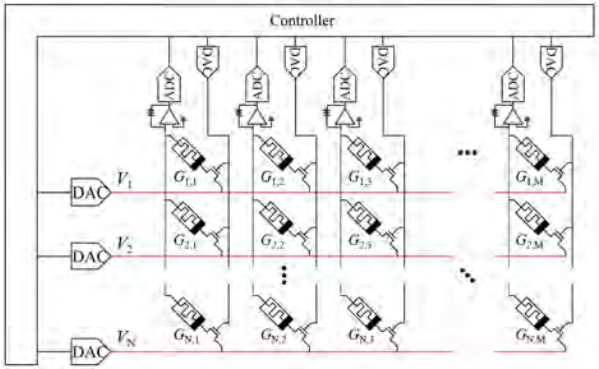


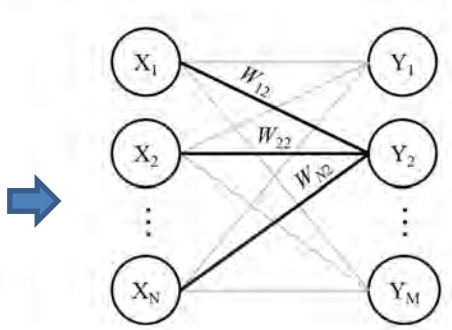
- Different from memory applications, computing applications are intrinsically more defect tolerant as the computing result is determined by the combined effect of a group of cells rather than just the reading result of a single cell.
- Variation will not be an issue with device<>circuit<> algorithm co-design
  - ✓ Optimized material and device design with very repeatable device switching behavior & small variation from foundry run to run
  - ✓ Closed-loop programming for very precise conductance level control
  - ✓ Auto calibration scheme can be used to mitigate the impact of defects, parasitics, thermal variations, and circuit variations

In Memory Computing with Analog Non-volatile Memory Crossbar

- Theory of operation using 1T1R crossbar
  - Programmable 8-bit NVM variable resistor
  - Configured as N x N (e.g. 256 x 256) crossbars
  - Multiply Accumulate (MAC) Operation
    - Voltage is input, Current is output
    - 1/R is weight, transistor for selection and current control
    - In Memory Computing (IMC) by Ohm's law and Kirchhoff's current law

- Efficient processing (>100X efficiency and speed improvement)
  - No need to load weights, no memory wall trade-off
  - Same input loaded to multiple filters in parallel
  - Massive operations in parallel
  - Minimize intermediate load stores
  - CNN = Many MAC operations





$$I_k = \sum_{i=1}^{N} G_{ik} \cdot Vi$$

$$Y_k = \sum_{i=1}^{N} W_{ik} \cdot Xi$$

One-step VMM (MAC) Calculation

- **In-memory computing!**
- **Parallel computing!**
- **Analog computing!**

Technology Evolution Through Series of Fab Tapeout Developments

Flash Memory Summit

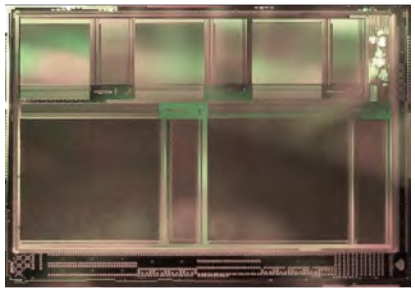
2019

- 6 bits
- 128x64 crossbar

2020

- 6 bits
- 256x256 crossbar

2021



- 8 bits
- 256x256 crossbar
- 5 crossbar arrays
- RSIC-V & digital functions

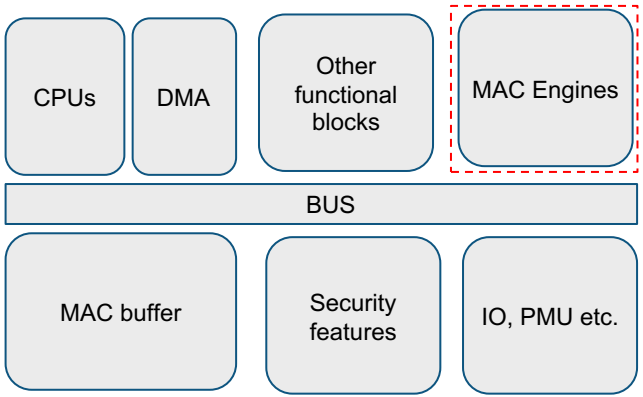
2022

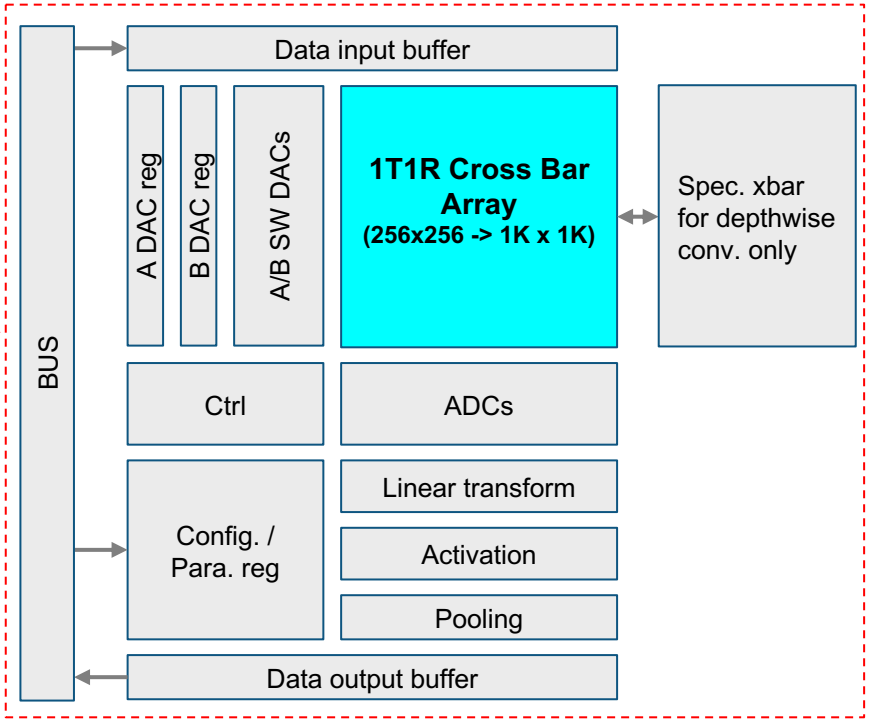
- 8 bits
- 256x256 crossbar
- 10 crossbar arrays
- Improved digital functions

Scalable System Architecture For Various Applications

Flash Memory Summit

Highly Scalable Design with MAC Func Groups:
Support Tiny Edge (e.g. CIS sensor) to
server/data center HPC accelerator



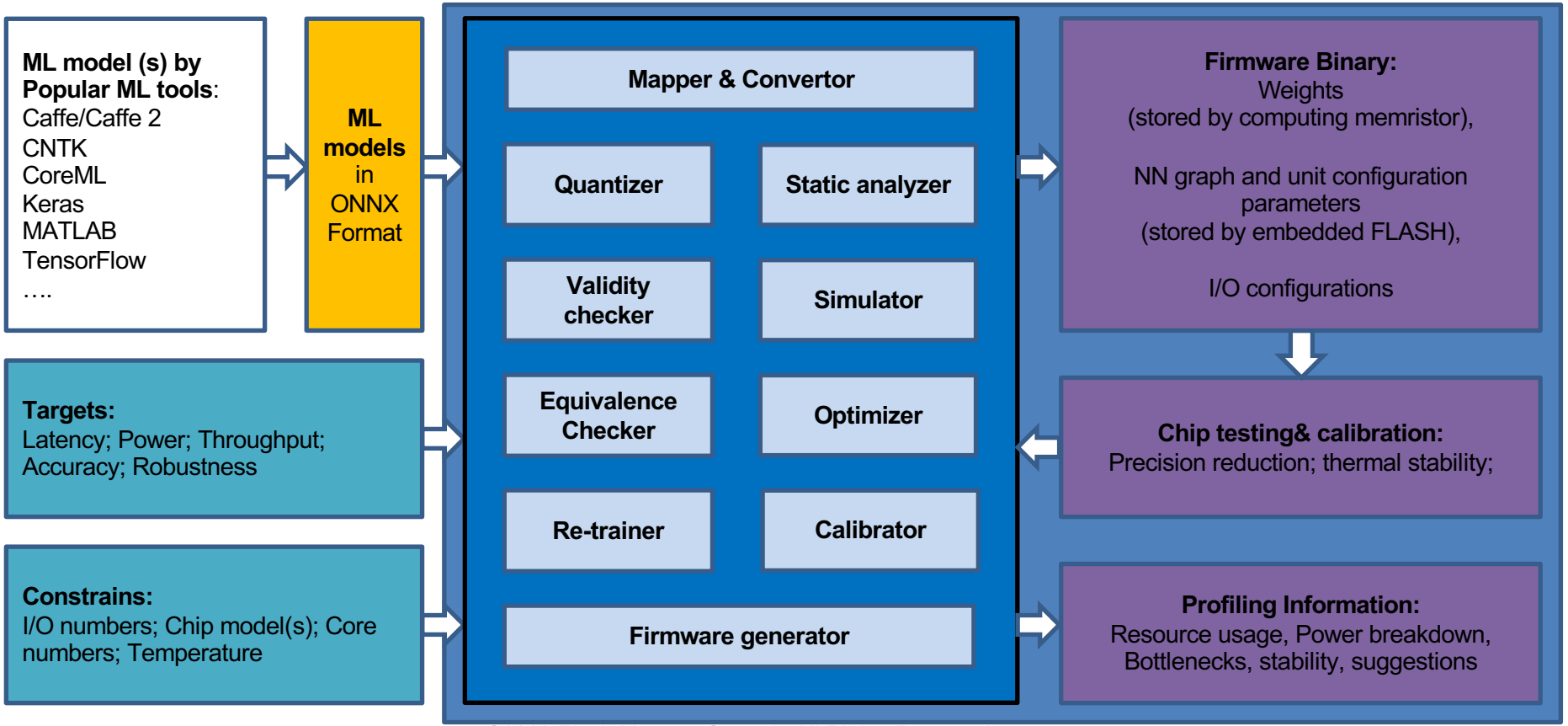


Single MAC Engine Core Architecture

Software Development Stack

Flash Memory Summit

SDK1.0 is available now



TetraMem Software Solution for the Very-edge AI Applications

Flash Memory Summit

ONE pre-trained model

*Overseen with system model to overcome the **analog loss***

Quantization

Optimization

Localization

MILLIONS of Chips with AI engines

Chip # 1
Analog IMC AI-engine
Other blocks

Chip # 2
Analog IMC AI-engine
Other blocks

Chip # 1,000,001
Analog IMC AI-engine
Other blocks

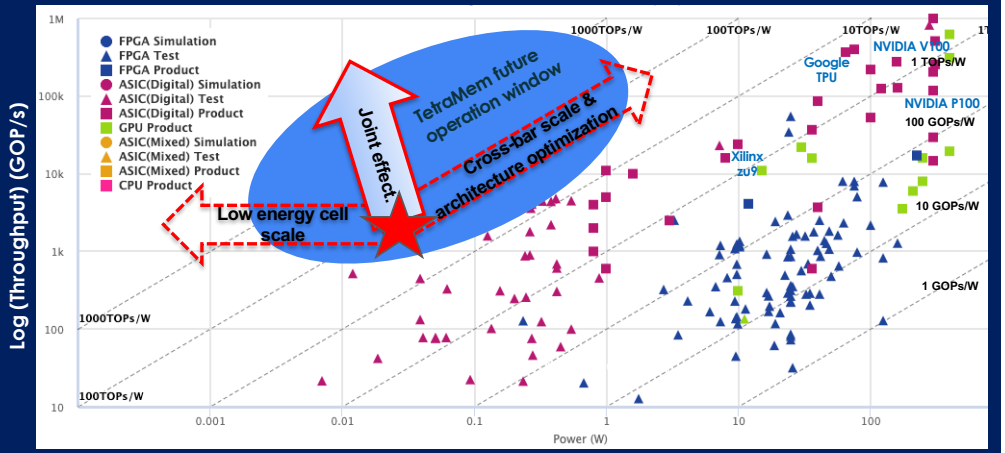
Unique advantages:

- *Near-zero boot-on time*
  - *Non-volatile 8-bit weight storage*
- *Ultra-low latency*
  - *Ultra-fast direct uint8->uint8 NN layer operation*
- *Ultra-low power*
  - *Sequential but fast layer execution*
  - *Minimize memory/peripheral circuit overhead*

Opportunity and the Market Through Technology Advances

Flash Memory Summit

TetraMem technology will offer both **Low-power** and **High-throughput** advantages



Adapted from https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

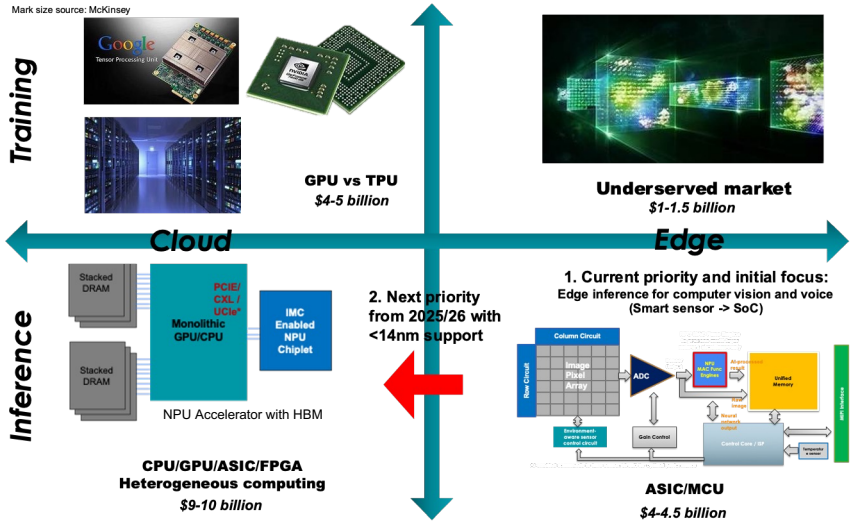
Low Power
Eliminate von-Neumann bottleneck

High Throughput
Massive parallel operations

Low Latency
Analog computing

High Adaptability
Supports all mainstream neural networks: RNN, CNN, LSTM etc.

Go-to-Market Strategies
1) IP & Soc for computing vision and voice for edge inference; 2) chiplet for data center

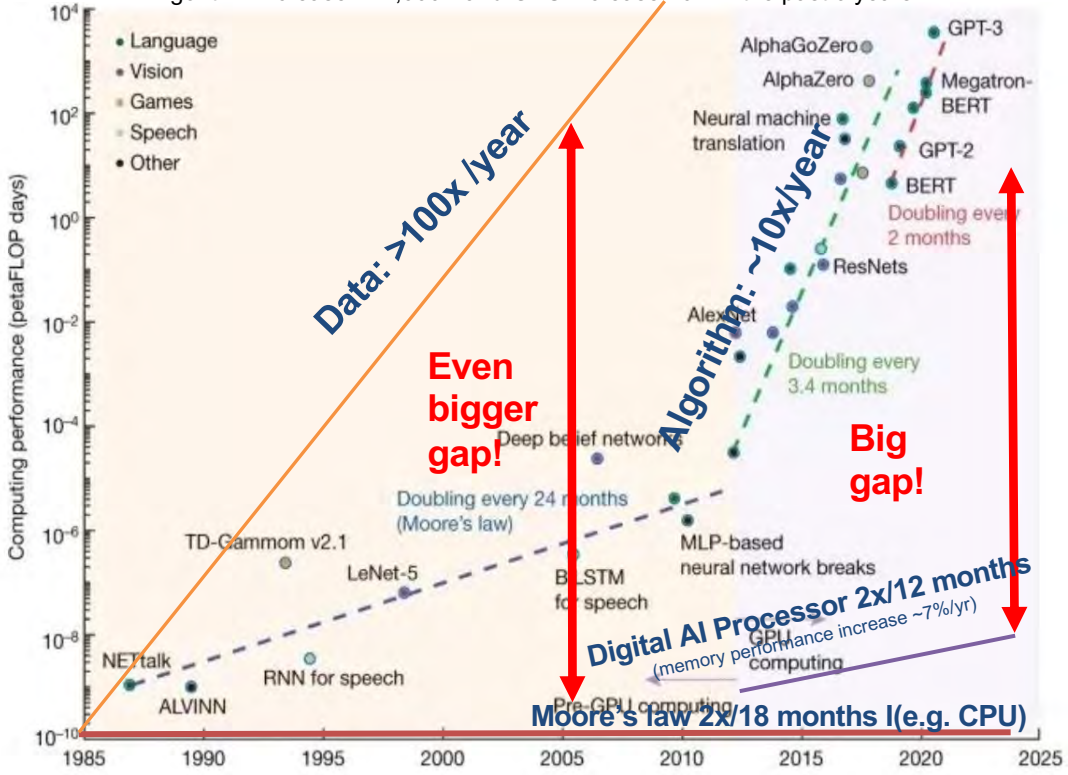


Appendix

The Problem - Unsustainable Path for Current Digital Approaches

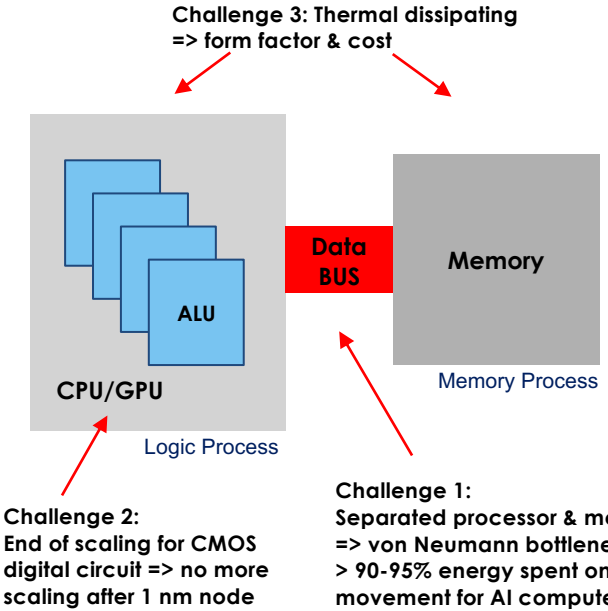
AI applications are driven by Data, Algorithm and computing hardware. **Hardware become the bottleneck**

Algorithm increase >17,000X and GPU increase 26X in the past 6 years

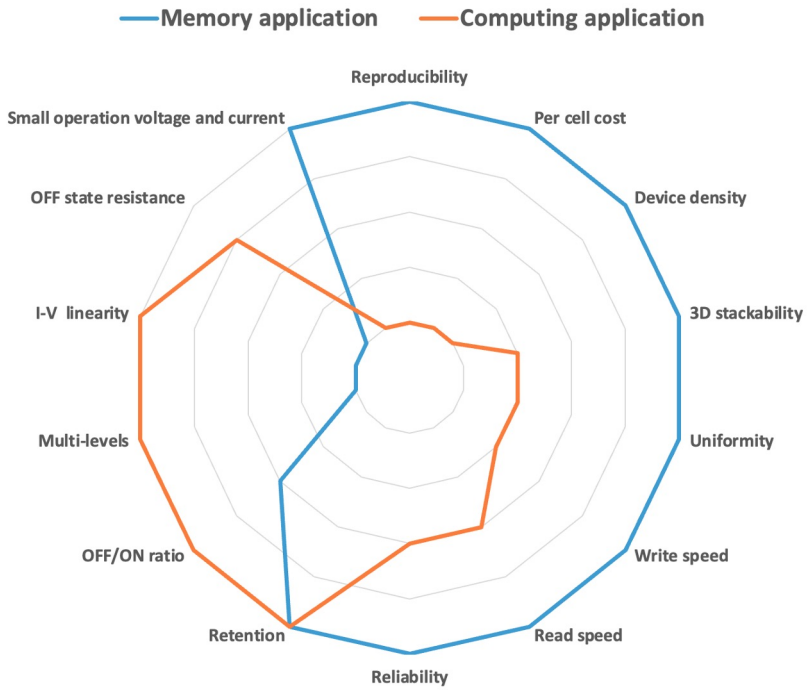


Source: OpenAI's 'AI and Compute'

Traditional von Neumann digital computing systems such as CPUs, GPUs, TPUs, and others encounter significant challenges when performing AI computations, primarily those based on matrix operations

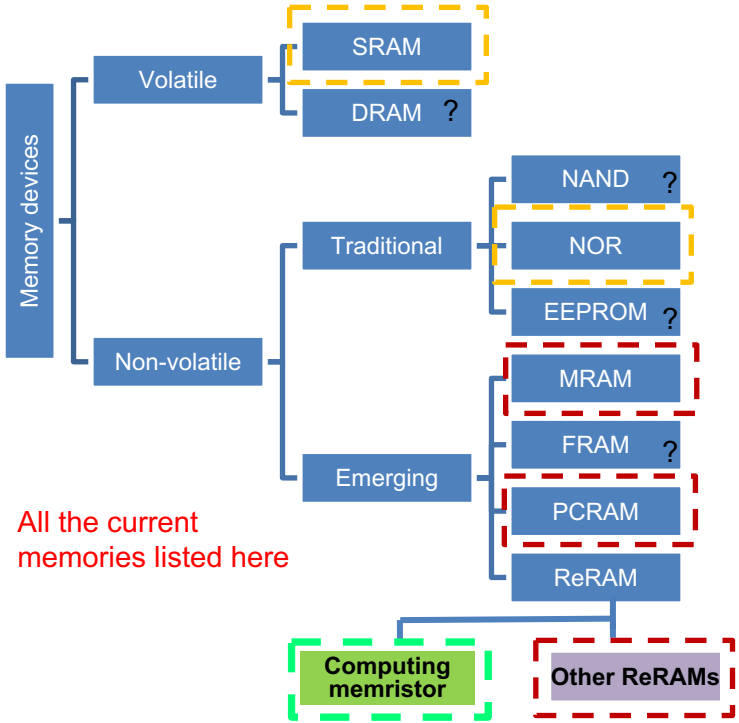


Computing Device: Memory with Special Attributes For Computing



J. Joshua Yang et al., "Memristive devices for computing" , Nature Nanotechnology 8, 13-24 (2013) (>3,200 citations).

- **IMC architecture is great! But suitable memory device is the key for the success. However, there are not many choices. And most of them are invented for memory application.**
- **Memory devices need decades of development**



All the current memories listed here

Analog In Memory Computing Working Mechanism Example

Flash Memory Summit

Legend: [X]: Vector of input information
A: Weight of neural network
G: Conductance of 1T1R cell to represent A
Y: Summation of vector-matrix multiplication

1) Training with features extraction

Image recognition example

2) Input data for inference

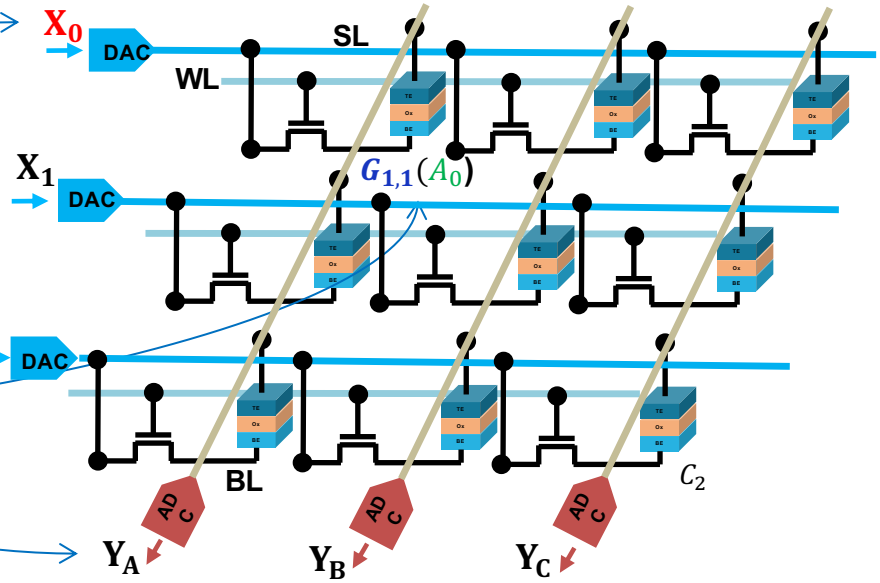
3) Output decision: Dog!

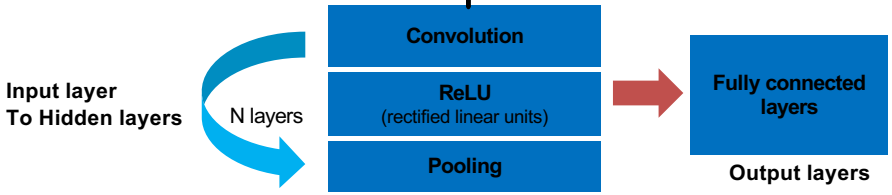
e.g. testing image

Neural Network

Output results used for inference

$$[X_0 \; X_1 \ldots X_n] * \begin{bmatrix} A_0 & \cdots & C_0 \\ \vdots & \ddots & \vdots \\ A_n & \cdots & C_n \end{bmatrix} = \begin{bmatrix} Y_A = X_0A_0 + X_1A_1 \ldots + X_nA_n \\ Y_B = X_0B_0 + X_1B_1 \ldots + X_nB_n \\ Y_C = X_0C_0 + X_1C_1 \ldots + X_nC_n \end{bmatrix}$$



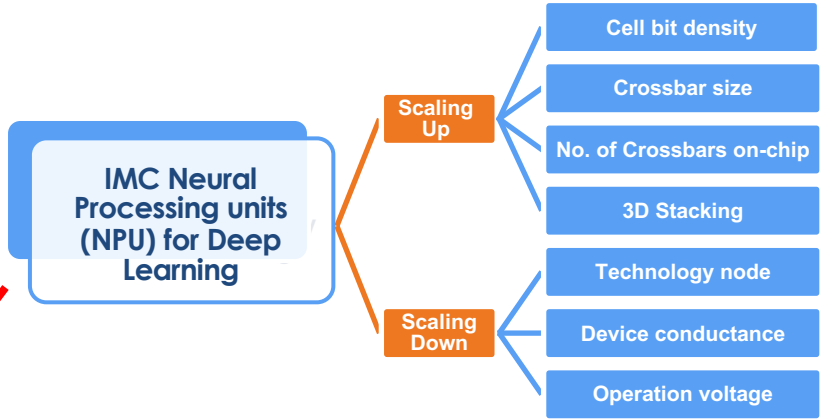


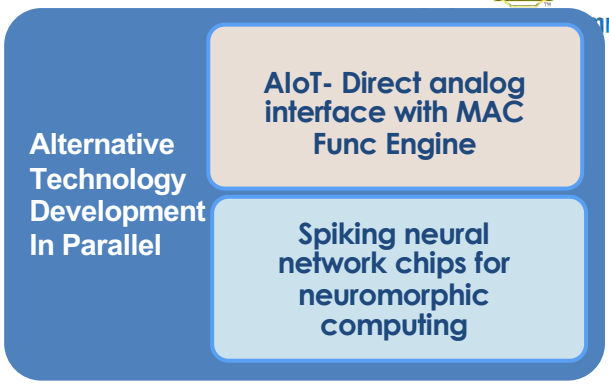
$$Y_j = \sum \underbrace{X_n G_{i,j}}_{\textit{Ohm's law}}$$

Kirchhoff's current law

Primary energy saving: weight access, matrix multiplication accumulation, data movement

Technology Roadmap





NPU Generation	Minimum Technology Node Support	# of Core	# of Crossbar Array /Core	Crossbar Size	Computing Memristor Device	Weight Capacity Support	TOPS Estimation	Core TOPS/W Estimation	Target Application
1st generation (2022)	65nm	1	10	256x256	Gen1 (8-bit/cell)	> 0.6M	~10	~30	Edge Inference
2nd generation (2024)	22nm	24	16	512x512	Gen1 (8-bit/cell)	> 110M	~100	~60	Edge Inference
3rd generation (2025/26)	14/12nm	64	32	1024x1024	Gen2 (10-bits/cell with low energy cell)	> 2G	~400	> 100	Edge/Cloud Inference