

Device Requirements and Challenges of Mixed-Signal Neuromorphic Hardware

Dmitri Strukov
UC Santa Barbara

National Academies of Sciences, Engineering, and Medicine Workshop on Frontiers for Memristive Materials for Neuromorphic Processing Applications

February 28, 2020
Washington, DC

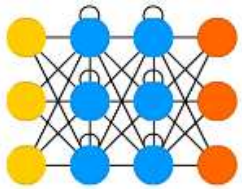
Outline

- Introduction to models, applications & digital hardware
- Neurocomputing with memory devices
- Device requirements & challenges
- Examples of recent mixed-signal accelerator prototypes
- Concluding remarks

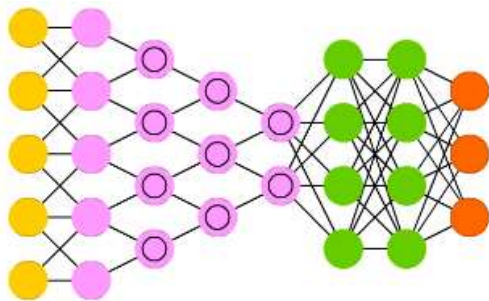
Part I.
**Neural Network Models,
Applications, and Digital
Hardware**

Neural Network Zoo

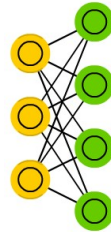
Recurrent Network (RNN)



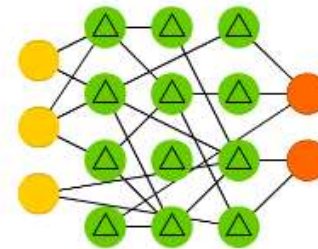
Deep Convolutional Network (CNN)



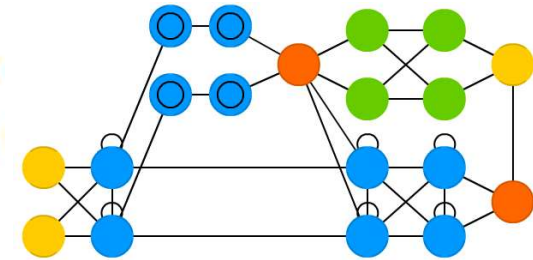
Restricted BM (RBM)



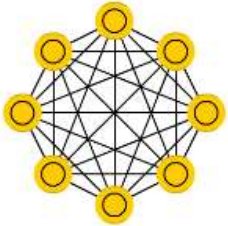
Liquid State Machine (LSM)



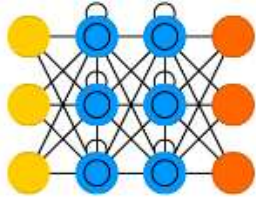
Attention Network (AN)



Hopfield Network (HN)



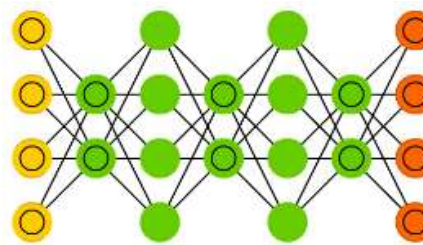
Long / Short Term Memory (LSTM)



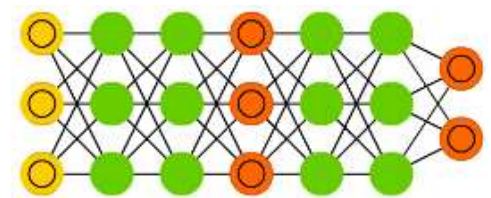
Boltzmann Machine (BM)



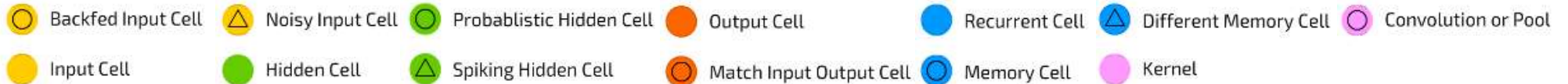
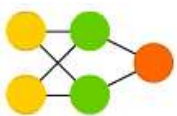
Deep Belief Network (DBN)



Generative Adversarial Network (GAN)



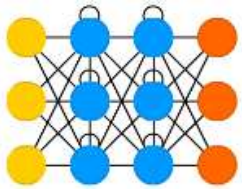
Multilayer Perceptron (MLP)



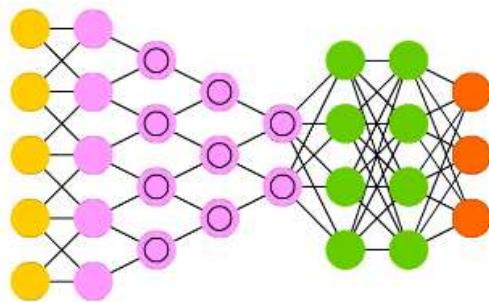
- Networks differ mostly by neuron connectivity & neuron functionality (more complex towards the right)

The Most Important Operation in Neural Networks

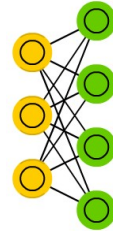
Recurrent Network (RNN)



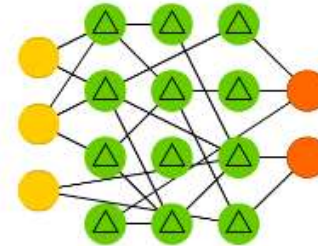
Deep Convolutional Network (CNN)



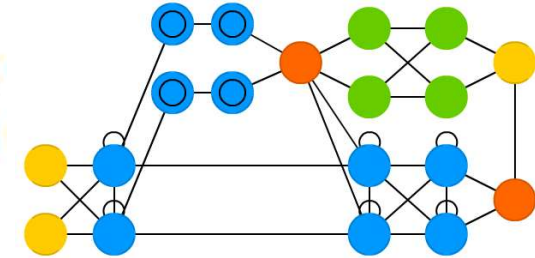
Restricted BM (RBM)



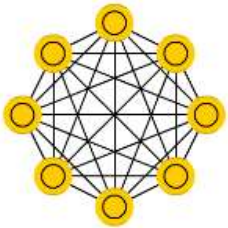
Liquid State Machine (LSM)



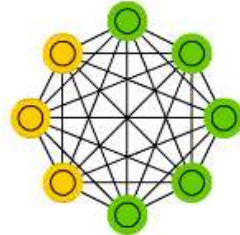
Attention Network (AN)



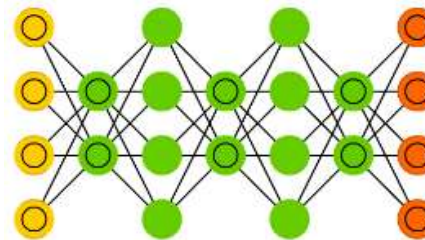
Hopfield Network (HN)



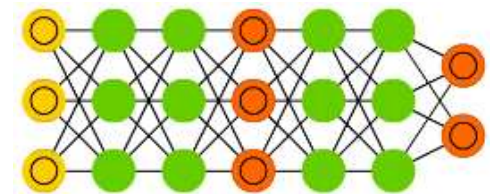
Boltzmann Machine (BM)



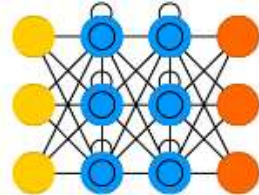
Deep Belief Network (DBN)



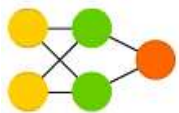
Generative Adversarial Network (GAN)



Long / Short Term Memory (LSTM)



Multilayer Perceptron (MLP)

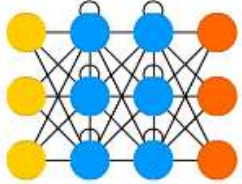


- Vector-by-matrix multiplication is the most common operation

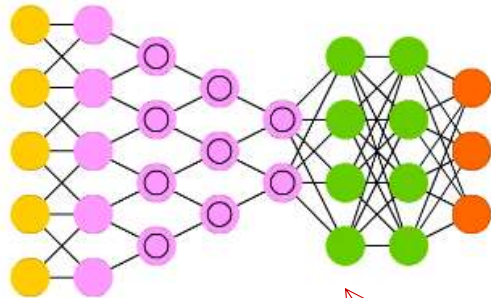
$$x_i = f\left(\sum_{j=1}^N w_{ij} y_j\right)$$

Practically Useful Neural Network Models

Recurrent Network (RNN)



Deep Convolutional Network (CNN)



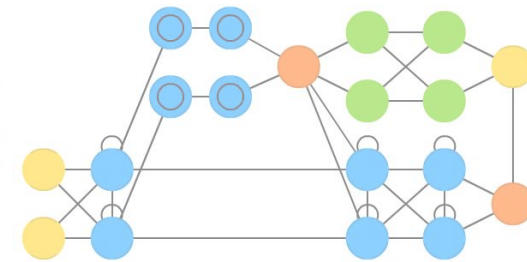
Restricted BM (RBM)



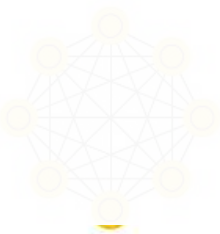
Liquid State Machine (LSM)



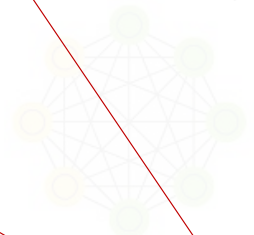
Attention Network (AN)



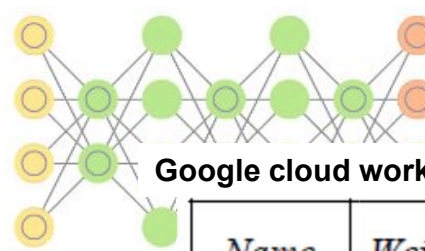
Hopfield Network (HN)



Boltzmann Machine (BM)



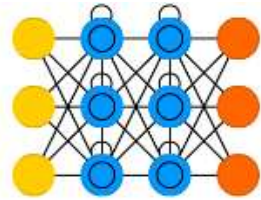
Deep Belief Network (DBN)



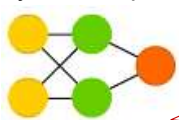
Generative Adversarial Network (GAN)



Long / Short Term Memory (LSTM)



Multilayer Perceptron (MLP)



Google cloud workload from N. Jouppi et al., ISCA'17

Name	Weights	% of Deployed TPUs in July 2016
MLP0	20M	61%
MLP1	5M	
LSTM0	52M	29%
LSTM1	34M	
CNN0	8M	5%
CNN1	100M	

Speech
Image

- Only firing-rate (non-spiking) feedforward & recurrent networks used commercially
- Weight sharing for CNN and LSTM

Largest Neural Network Models

Recurrent Network (RNN)



Deep Convolutional Network (CNN)



Restricted BM (RBM)



Liquid State Machine (LSM)



Example of talking to 1.5B-weight GPT-2 transformer (<https://talktotransformer.com/>):

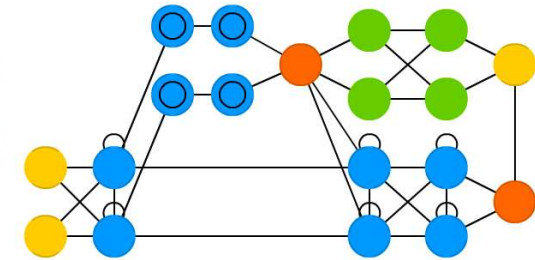
The main challenges for memristors are ...

... their low power consumption (unlike micro-transistors) and high initial operating voltage. They have a low dynamic range due to low switching frequency and high currents. The current flow in a memristor will decrease at higher switching frequencies than low-voltage transistors. Some memristors are expected to have lower resistance than the values that are currently used. High dynamic range capability will eventually be overcome with improvements in energy dissipation. Typical memristors can currently be specified to more than 2 nanovolts (billionths of a volt)...

} custom prompt

} ANN- generated completion

Attention Network (AN)

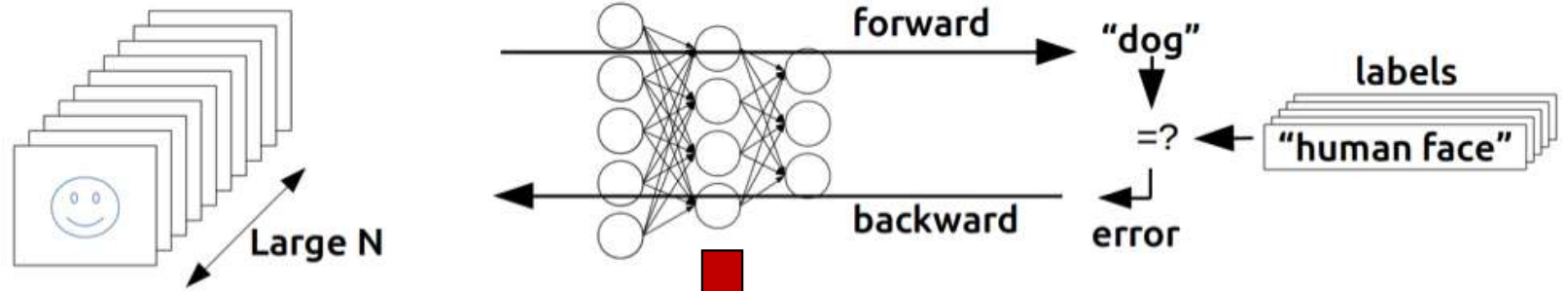


Generative Adversarial Network (GAN)

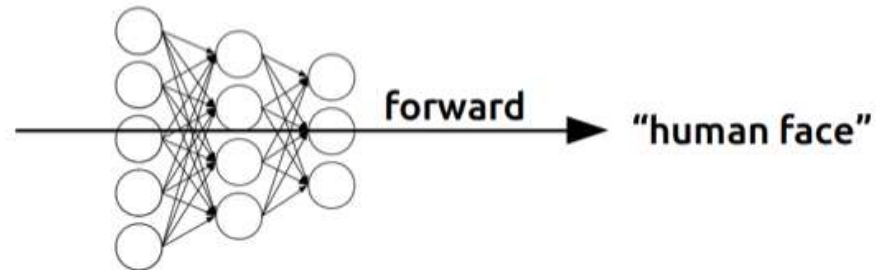




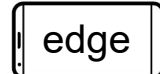
- >8B weights / >10¹⁹ FLOPs training cost in the state-of-the-art attention (transformer) networks
- Larger network → better quality

Training



Inference



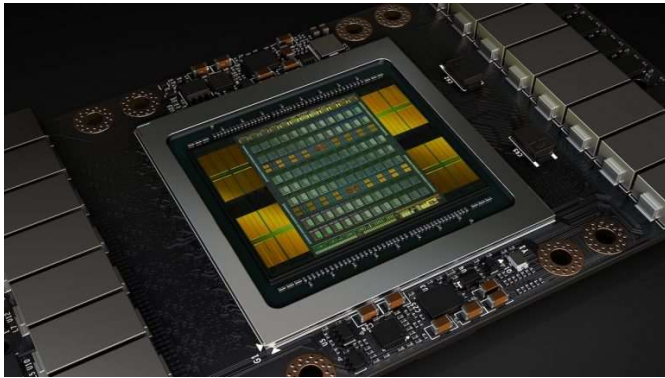
	Importance (world total in 2020, Source: NVIDIA)	Key differences in HW specs	Where deployed	Primary HW metric
Training	~55 ExaFLOP/s	Medium-to-high (8÷32 bit) computing precision	 cloud	Throughput per chip area
Inference	~450 ExaIOP/s FLOP = floating point op IOP = integer op Exa = 10 ¹⁸	Low-to-medium (4÷8 bit) computing precision Persistent (nonvolatile) weights	 cloud  edge	Throughput per chip area Energy efficiency

State-of-The-Art Deep Learning Hardware

commercial digital systems for high performance and mobile applications

100s W, \$\$\$\$

few W, \$\$

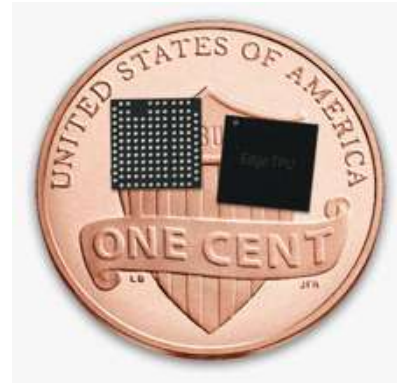


NVIDIA's Turing (12 nm)

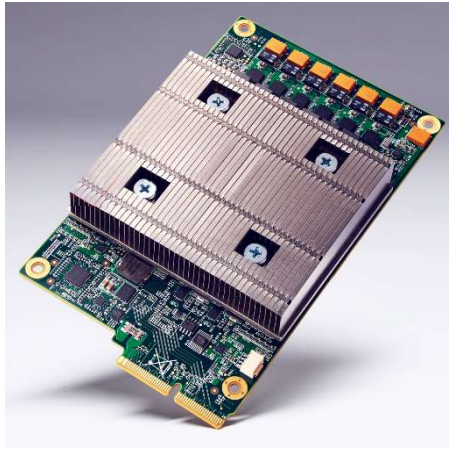


Intel's Movidius

Google Edge TPU



Google's Tensor Processing Unit



NVIDIA's Jetson

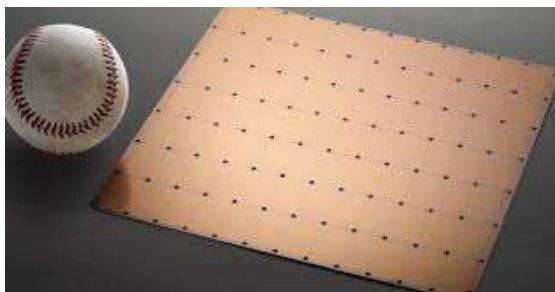


State-of-The-Art Deep Learning Hardware

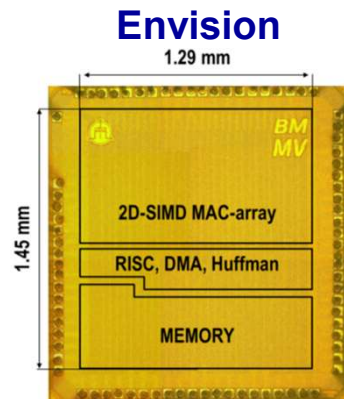
custom experimental digital systems

Startups: Cerebras, Habana, Wave Computing, Graphcore, Groq ...

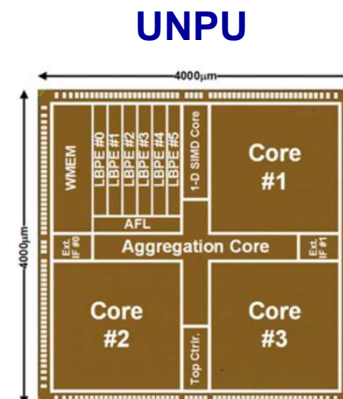
Cerebras (largest chip ever built)



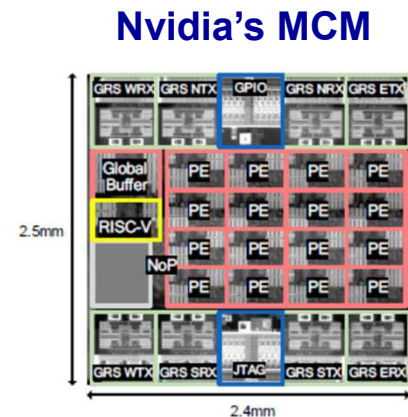
46,225 mm² 1.2T trans @ 16 nm
18 GB on-chip memory



B. Moons et al., ISSCC'17



J. Lee et al., ISSCC'18



B. Zimmer et al., VLSISymp'19

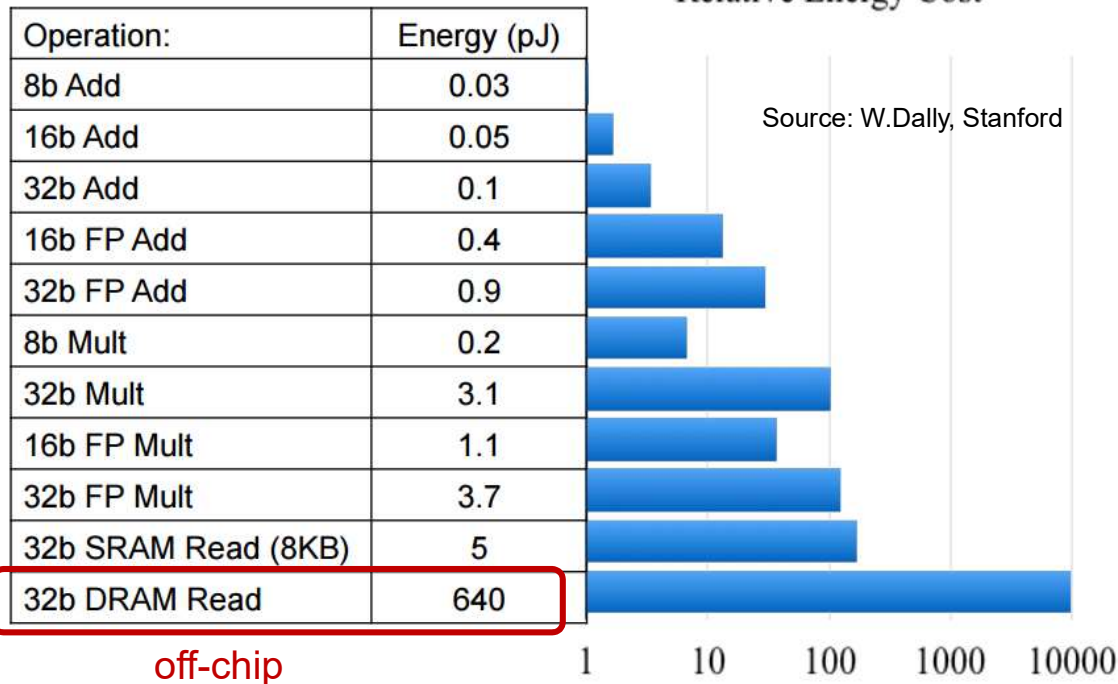
	Envision (2017)	UNPU (2018)	NVidia (2019)
Technology (nm)	28 UTBB FD-SOI	65	16
Peak performance [GOPS]	(1 to 4) × 102	1382 (4) / 345 (16)	4010 (8 bit)
Active area [mm ²]	1.87	16	3.1
Precision [b]	4-16	1-16	8
Power [mW] @ frame rate [fps]	44 @ 47*	297 @ ? *	?
Min/max energy efficiency [TOps/J]	0.26–10 (~2.5 @ 4-bit)	11.6(4 bit) / 3.08(16 bit)	~ 9.09 (8 bit)

* AlexNet convolutional layers only

- Saturating performance/energy efficiency, limited on-chip memory, expensive (\$300M design cost in 7 nm)
- Biology is many orders of magnitude more energy efficient

Problems with Digital Accelerators

Relative Energy Cost



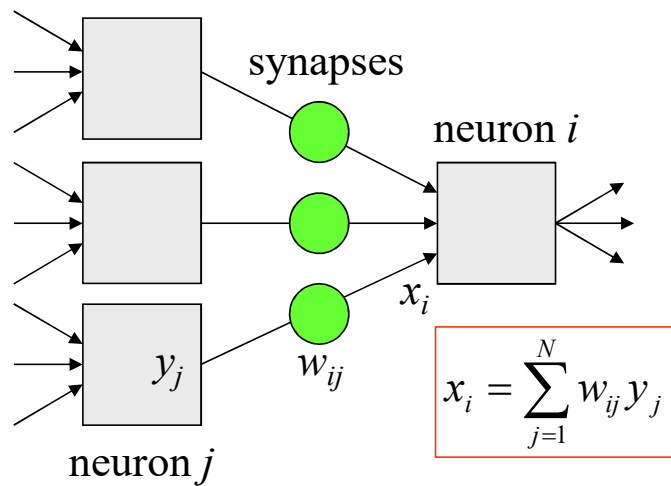
Large energy / latency overhead for moving data due to

- out of / near memory computing and bulky VMM circuitry
- off-chip weights

Part II.
**Neurocomputing with
Memory Devices**

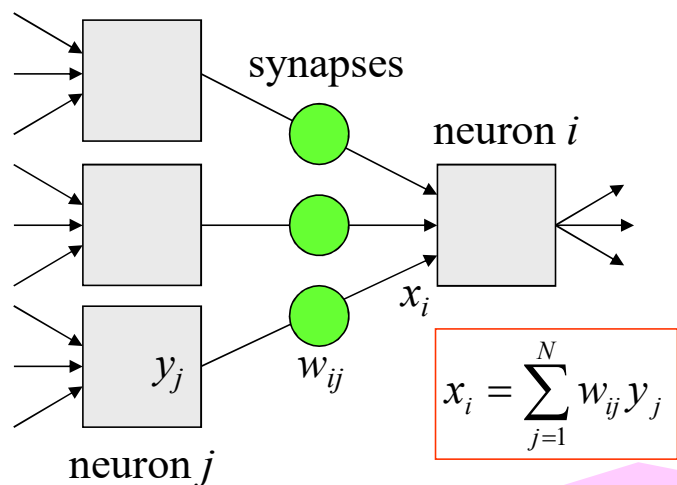
Implementing Basic Neuromorphic Operation

Dot-Product Computation ...

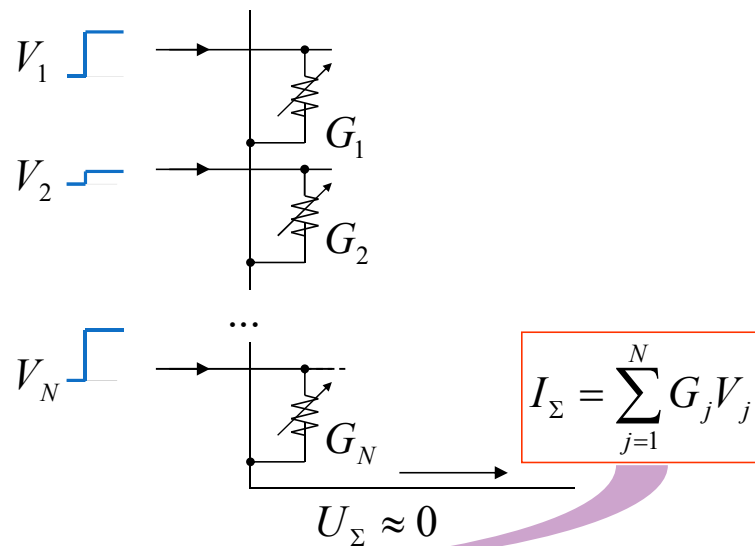


Implementing Basic Neuromorphic Operation

Dot-Product Computation ...



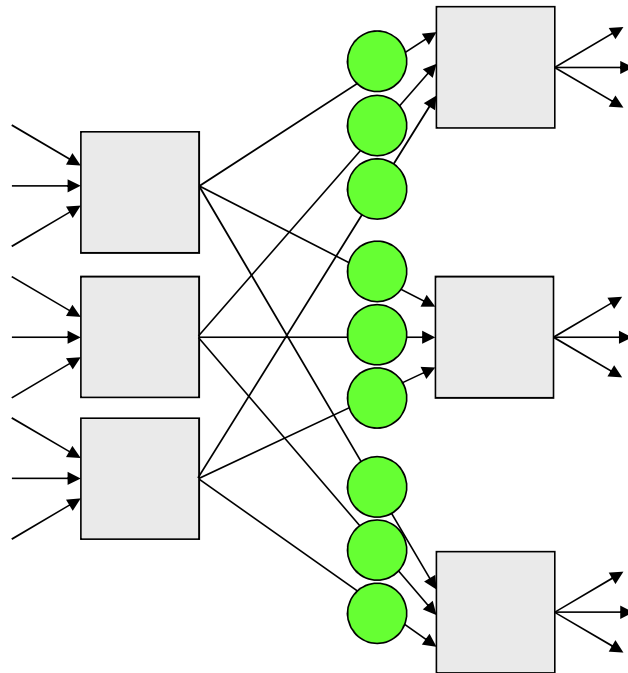
... by Analog Circuit



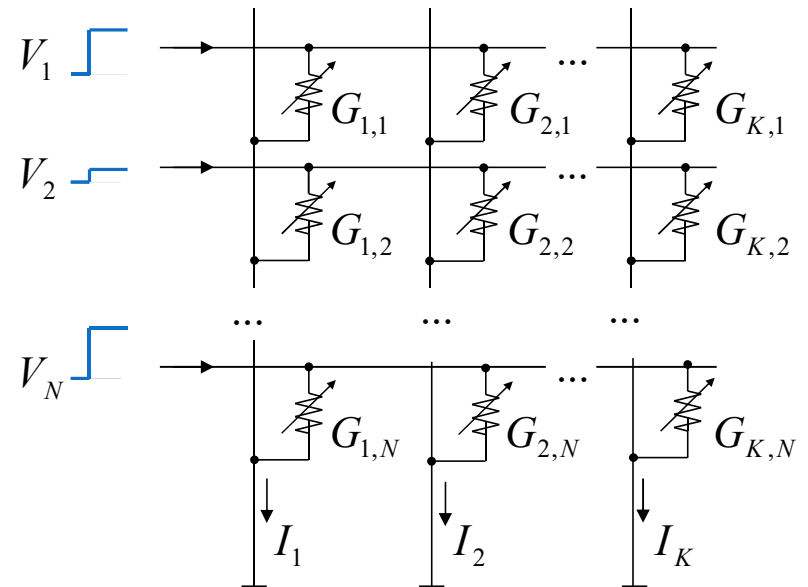
- Utilize fundamentally accurate Ohm's and Kirchhoff's laws
- In-memory computing (no need to move around synaptic weights)

Implementing Basic Neuromorphic Operation

Vector-by-Matrix Multiplication (VMM)



... by Analog Circuit

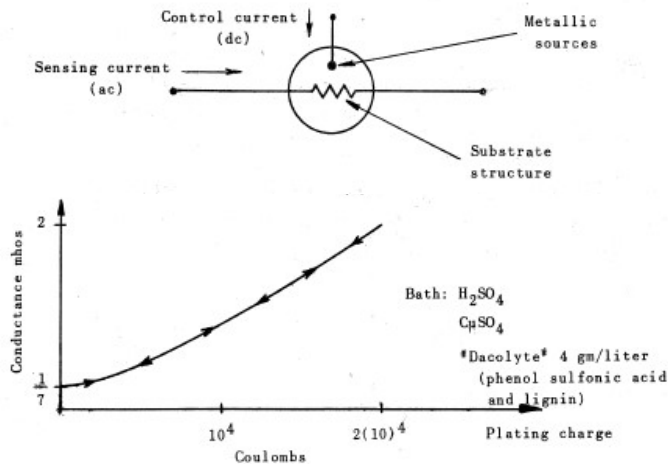
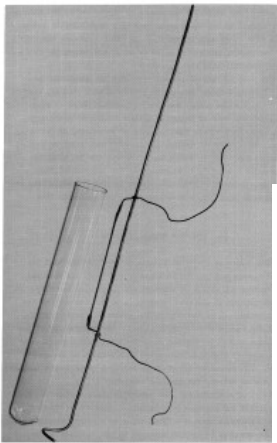


- No dense adjustable-conductance nonvolatile devices until recently

Earlier Work on Analog Neurocomputing

Widrow's "Memistor"

- AdaLiNe concept and its hardware implementation



B. Widrow and M.E. Hoff, Jr., *IRE WESCON Convention Record*, 4:96 1960

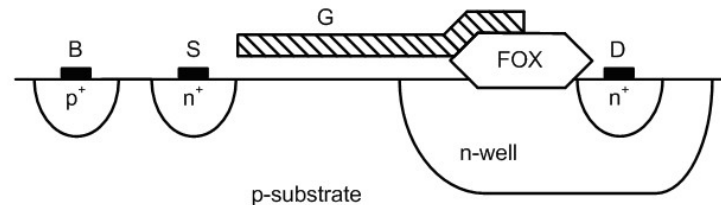
- Automated 4x4 b/w image classification in follow-up early 1970s work
- Coincided with the rise of digital processors (Intel 4004)



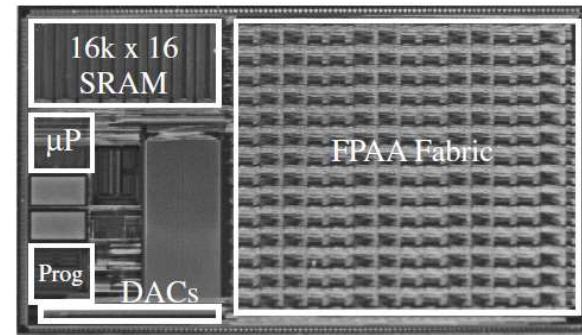
"Synaptic Transistor"

(from the late 1980s: C. Mead, C. Diorio, J. Hasler,...)

- "Extended drain" NMOS structure



- Hasler's recent FPAAs chip



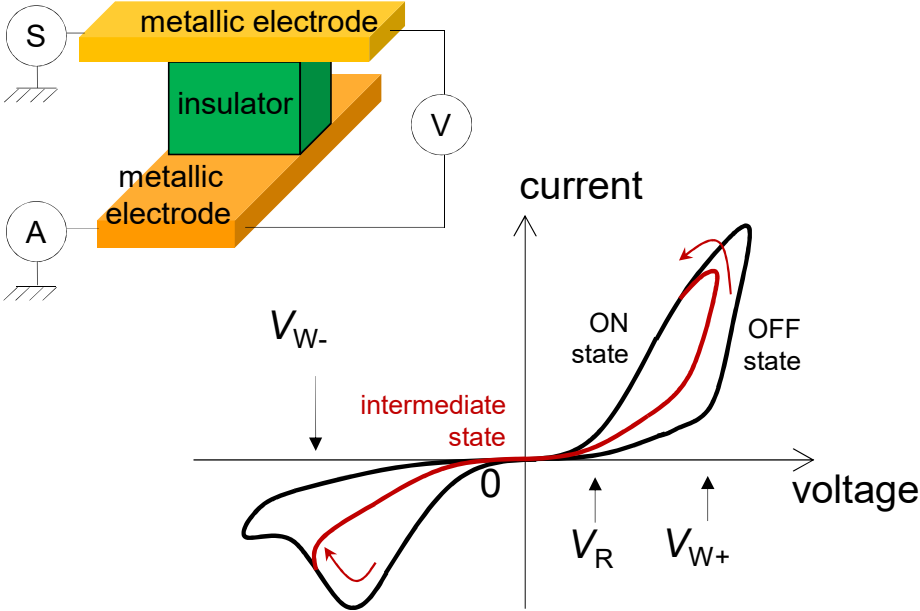
S. George et al. *IEEE Trans. VLSI*, 2016

>1000 F^2 per synaptic weight!

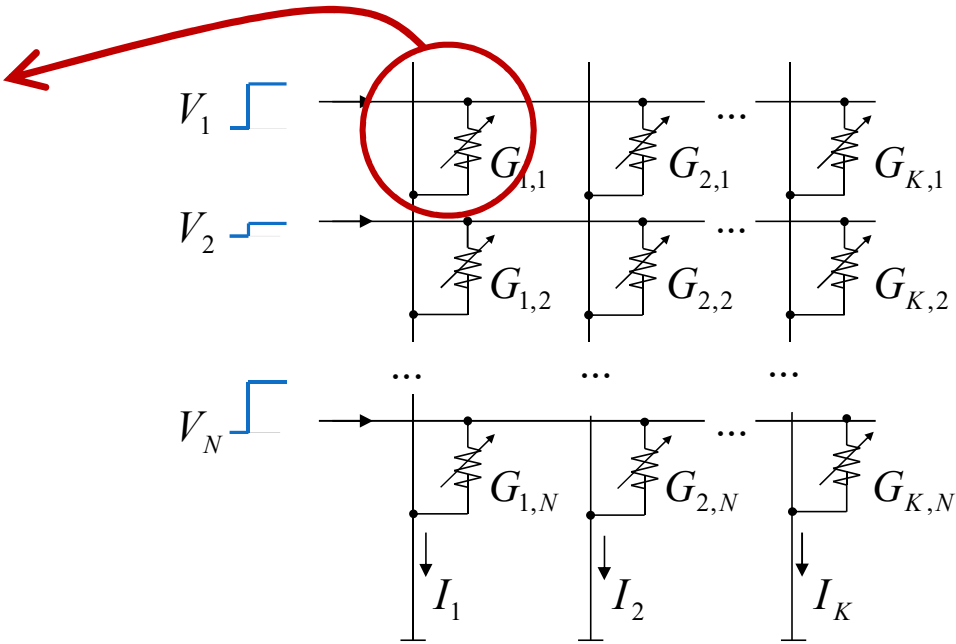
- PRO:** allow using common CMOS foundries
- CON:** large cells → low speed → low energy efficiency

New Life for Old Concept

Metal-Oxide Memristors



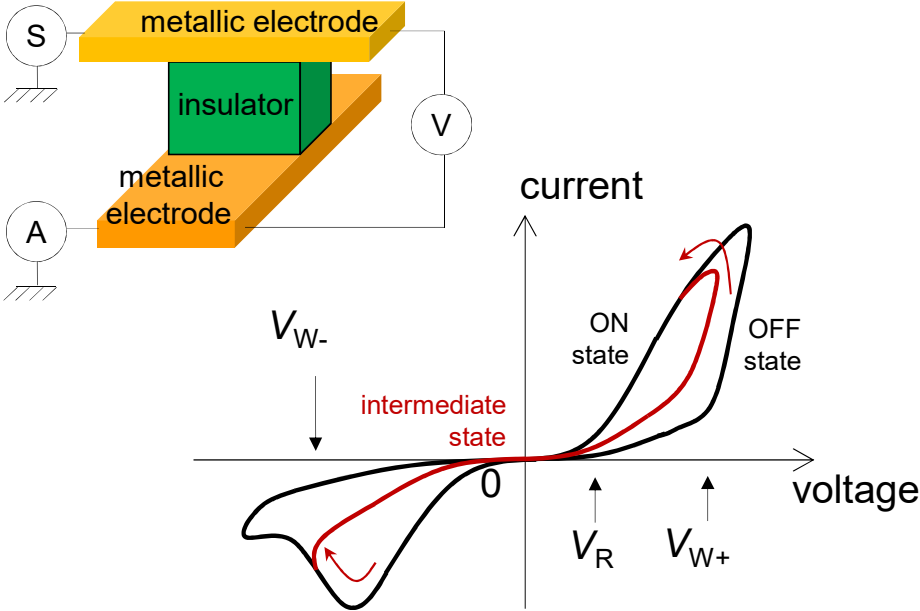
Analog VMM Circuit



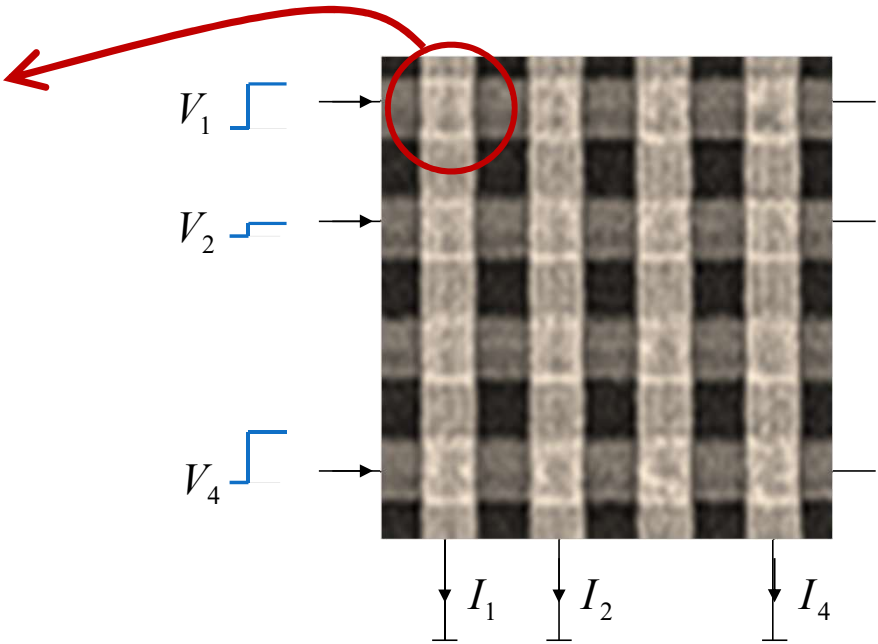
Applying $|V| \leq V_R$ does not disturb the memory state for nonvolatile memristors

New Life for Old Concept

Metal-Oxide Memristors



Analog VMM Circuit

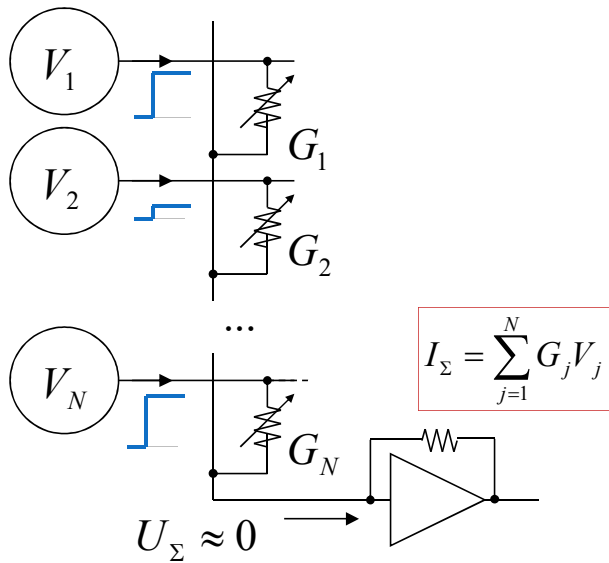


Tunable conductance, nonvolatile, extremely compact footprint when using passive memristors

Different Options for Analog / Mixed-Signal VMMs

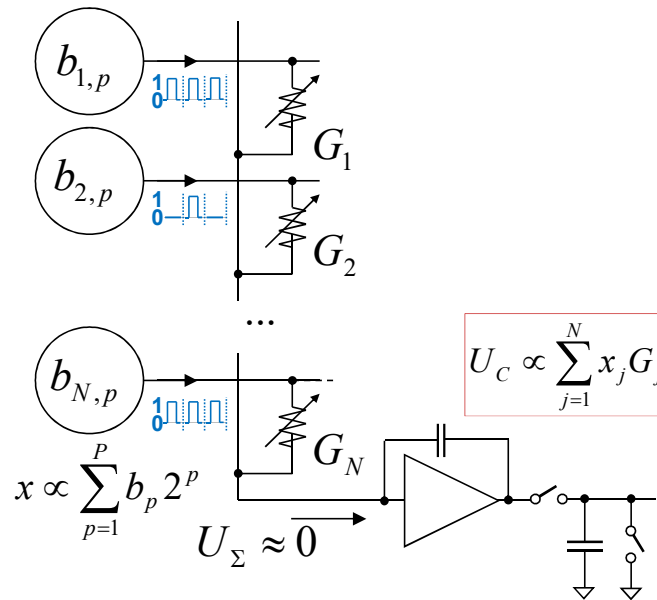
(classification based on how P -bit inputs are encoded)

Instant



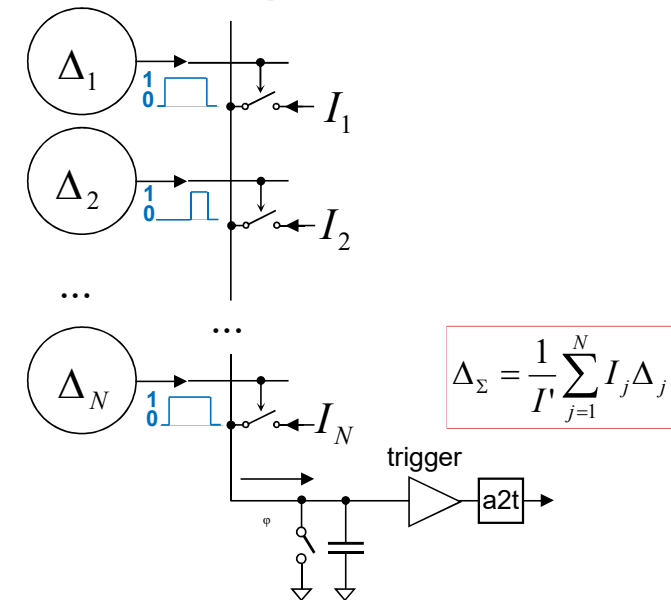
- Inputs are encoded in voltage amplitudes
- Dot-product is proportional to the output current

Linear



- Digital inputs are presented bit-by-bit
- Dot product is accumulated in P steps by successive integration and re-scaling
- Or accumulate in digital domain (HP's ISAAC)

Exponential



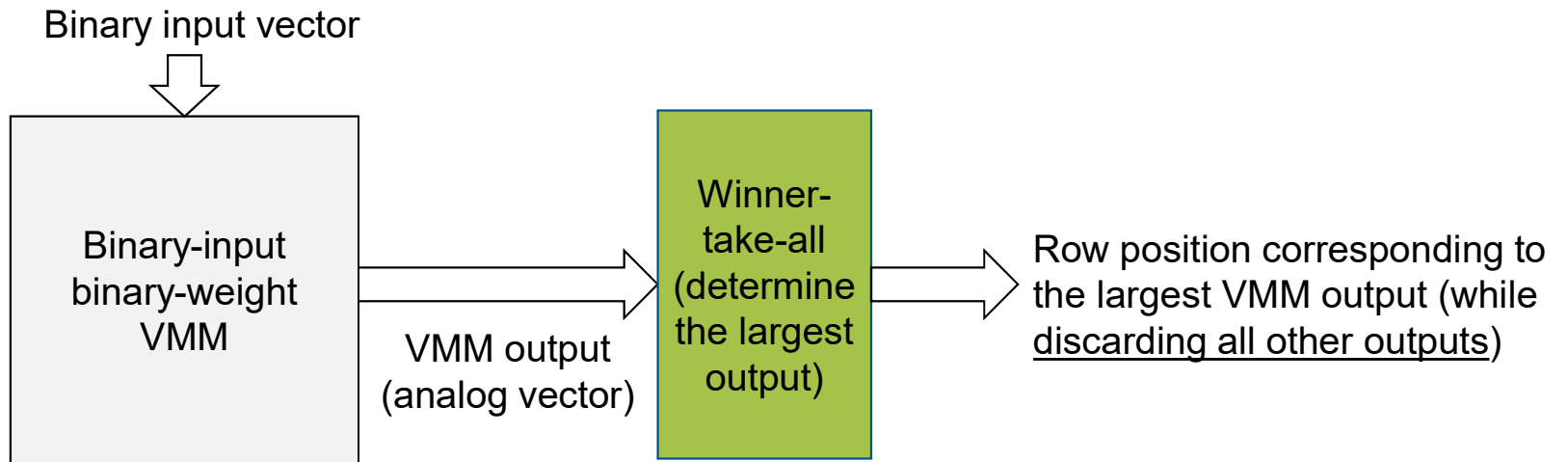
- Digital inputs are encoded in duration of fixed-amplitude pulses
- Dot product is proportional to the output pulse duration

Encoding in biology is closer to the fixed-amplitude / variable duration scheme

M. Bavandpour et al, IEEE S3S, 2019
M. Bavandpour, S. Sahay et al, IEEE TVLSI, 2019

Approximate Content Addressable Memory (CAM)

also known as “hyperdimensional” memory

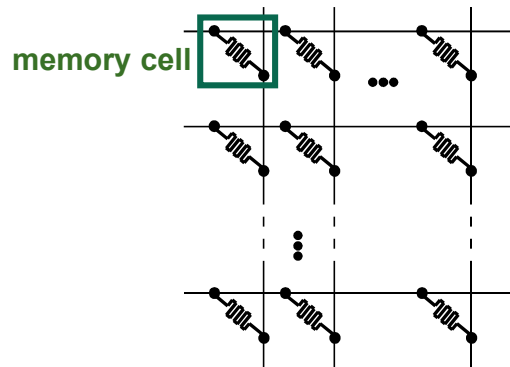


- CAM's essential operation is Hamming distance calculation
- Hamming distance between two vectors = dot-product of two vectors

Hyperdimensional computing is VMM with binary weights/inputs followed by winner-take-all circuit

Memory Array Options

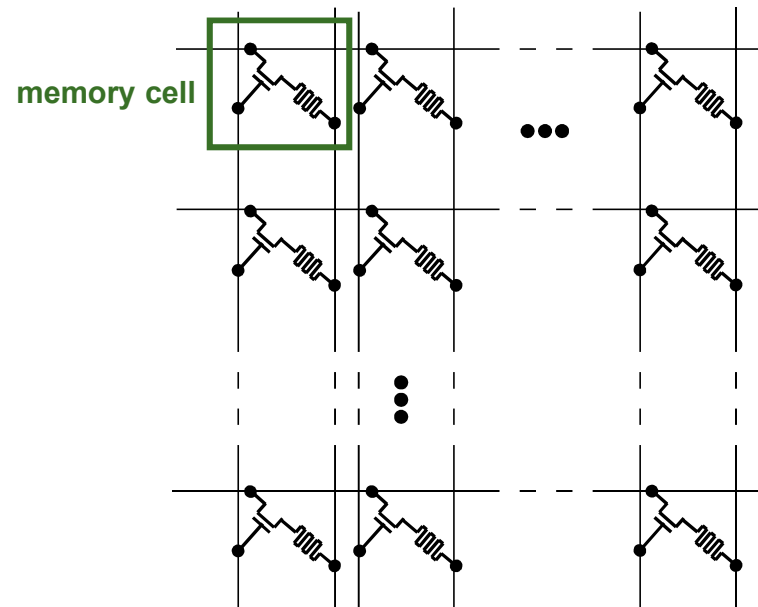
Passive (“0T1R”)



very dense

need very uniform devices

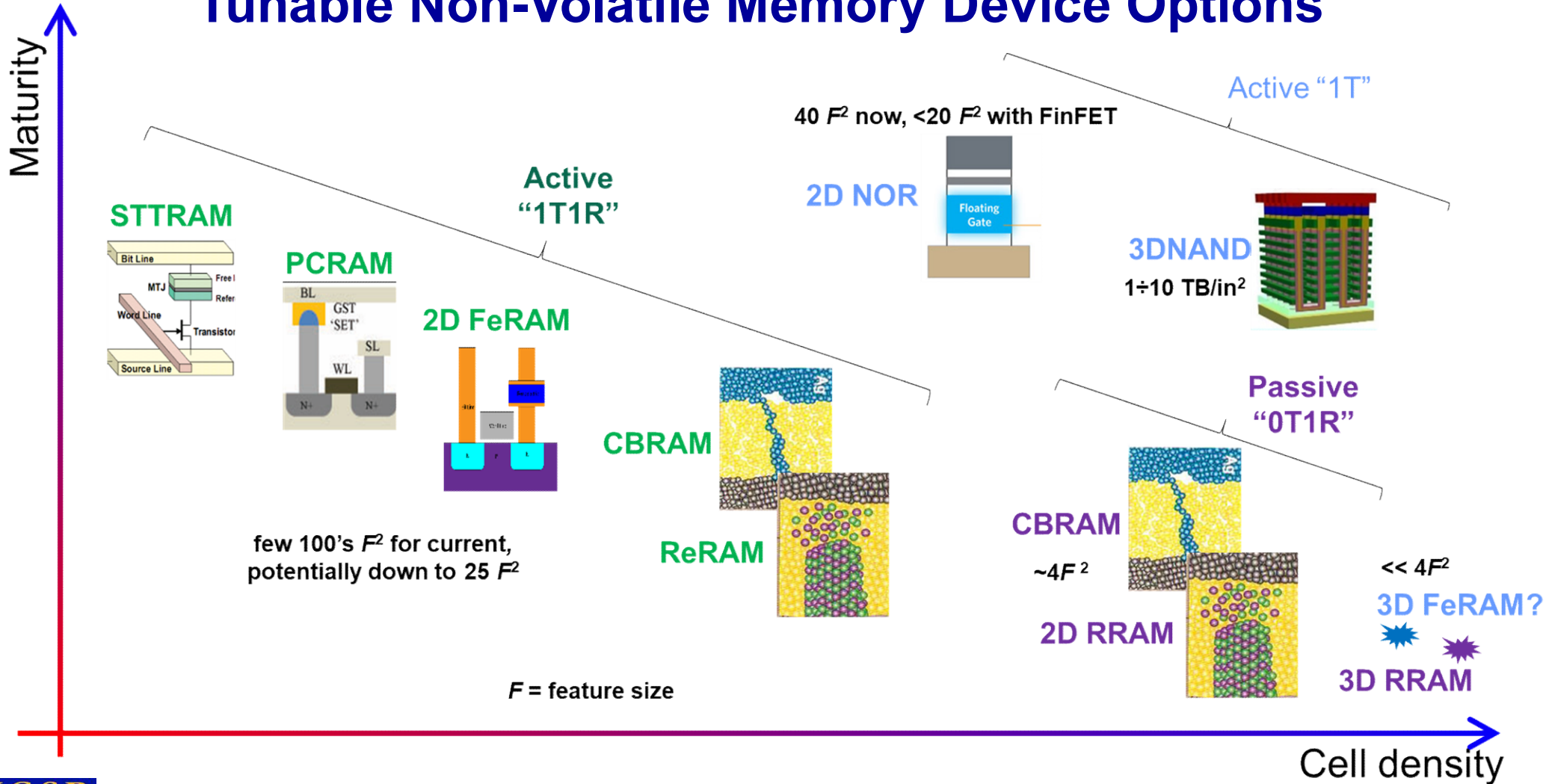
Active (“1T1R”)



less dense due to bulky transistor

relaxed d2d requirement

Tunable Non-Volatile Memory Device Options

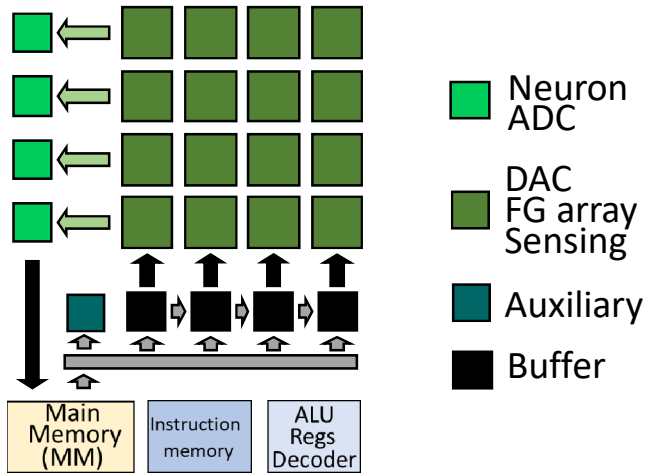


General Philosophy for System Level Demo

- Focus on ex-situ-trained neuromorphic inference
 - the simplest, yet very practical
- Has to be flexible enough
 - to run all recent practical NN models for image and language processing, e.g. those Google reported as workloads on their data centers
- Mixed-signal design
 - Most of the NN models rely on the weight reuse, which implies producing and storing some intermediate temporary data in the network. Temporary data are better to store in digital domain
 - Digital for more exotic (e.g. vector-by-vector outer product computations, different activation functions)
 - Analog circuits to implement dense VMMs, which dominate inference in all models

UC Santa Barbara's Inference Accelerator

aCortex architecture (one tile)



Details:

- Custom (energy-optimized) mixed-signal architecture, with estimates shown for NOR flash (projected from 100k-weights experimental chip)
- Analog neuron input bus
- Digital interfaces (DACs/ADCs)

System-level estimates for 55 nm NOR flash 4-bit and comparison with digital counterpart at the same compute precision / process node

	INC-V1	ResNet	GNMT
Network specifications			
# parameters	7.2e06	1.1e07	1.3e08
# operations	5.2e09	2.0e10	2.6e09
Area breakdown (%)			
MM	18.1	4.53	2.2
Sensing	15.5	23.3	25.1
FG arrays	24.2	36.5	39.3
P/E	26.3	14.7	11.3
Others	16	23	22.1
Energy breakdown (%)			
MM	38.8	23.9	8.3
Sensing	16.2	11.4	23.8
FG arrays	3.03	2.13	4.45
Buses	31.6	41.3	12.4
Leakage (buses)	4.4	17.4	46.7
Others	6.9	4	4.4
Performance summary			
Area (mm ²)	35.4	142	293
Power (mW)	14.9	19.8	16.1
Latency (ms)	3.1	8.75	0.59
EE (TOp/J)	114	120	283
Throughput (TOp/s)	1.69	2.37	4.54

	Chip area (mm ²)	Latency (ms)	EE (TOp/J)
Digital CMOS	257	749	0.14
Analog FG	293	0.59	283

2

- System-level efficiency for EE-optimized designs is mainly limited by memory density

2

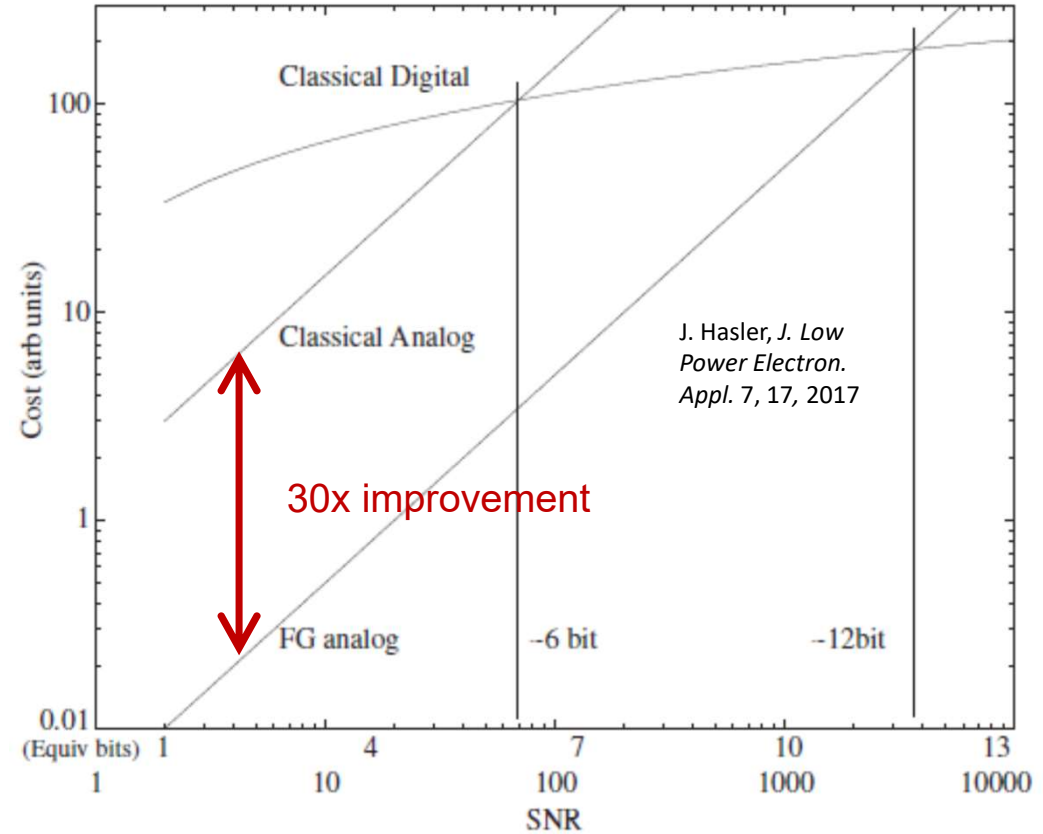
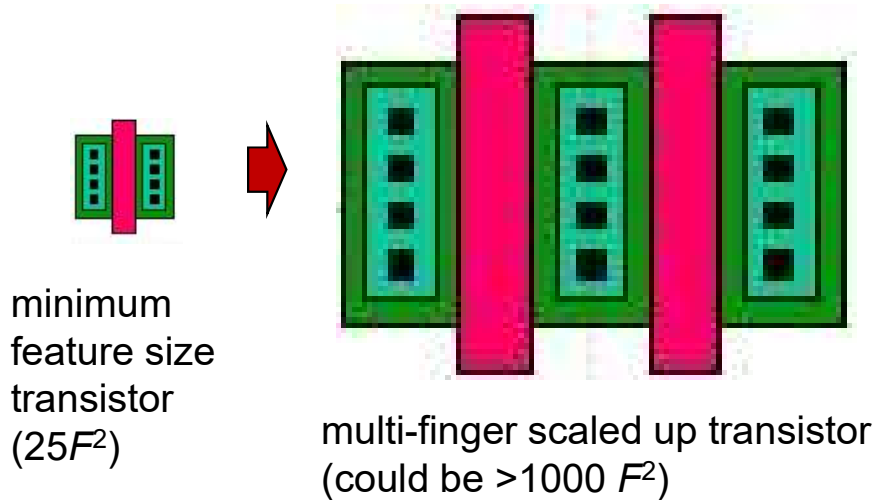
- Small networks (e.g. with $< \sim 10^4$ Op / pixel in 55 nm) will be dominated by I/O

3

M. Bavandpour et al, IEDM, 2018

Conductance Tunability at Rescue from Process Variations

Analog circuits' transistors are scaled up to properly operate with worst-case process variations

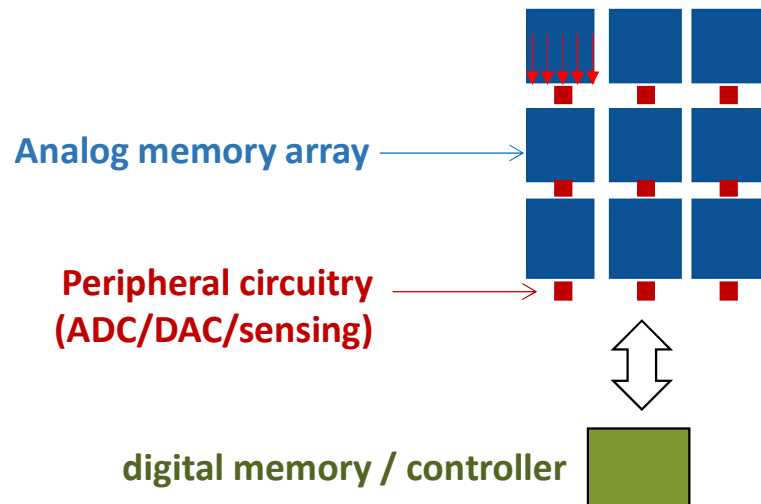


Memory state tunability → low overhead for dealing with process variations

Tradeoff between Energy-Efficiency and Throughput

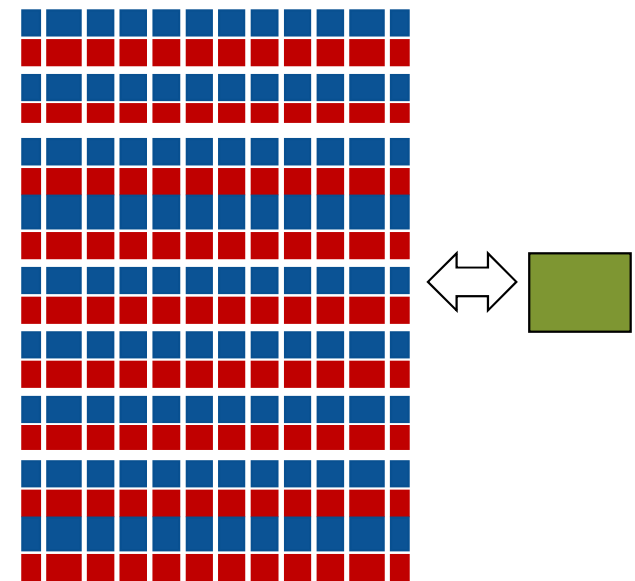
Energy-efficient design (for edge)

less periphery / more compact

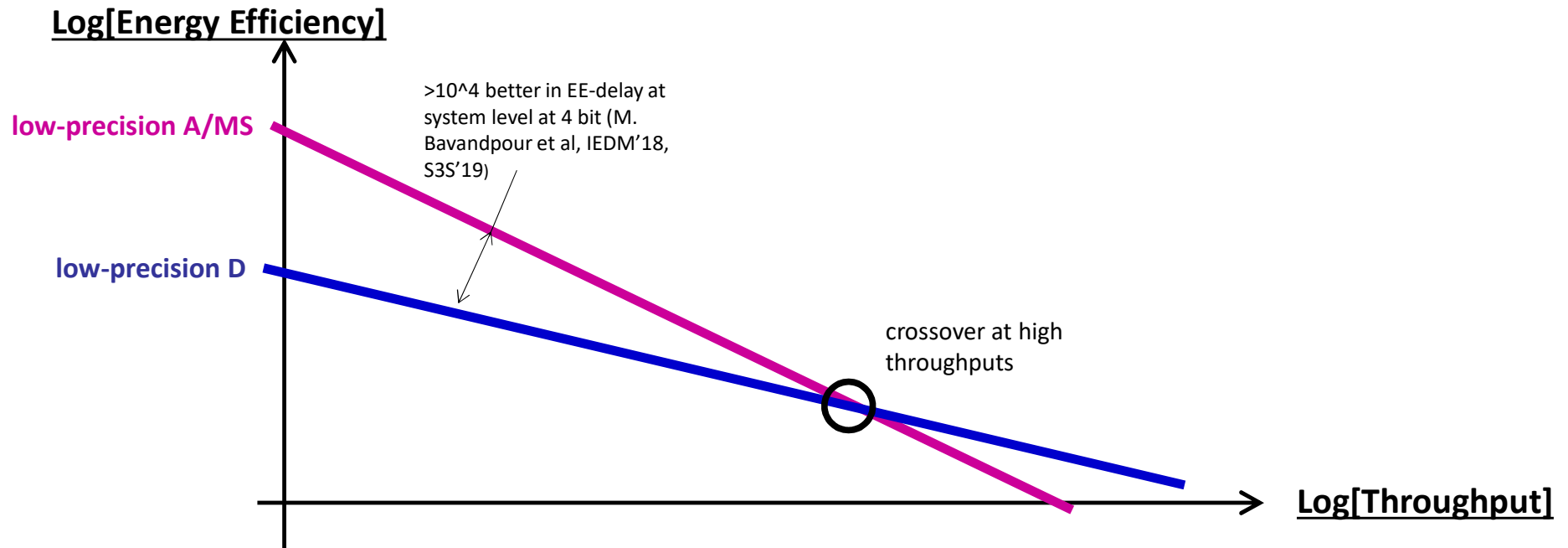


High-throughput design (for cloud)

more VMM computations in parallel

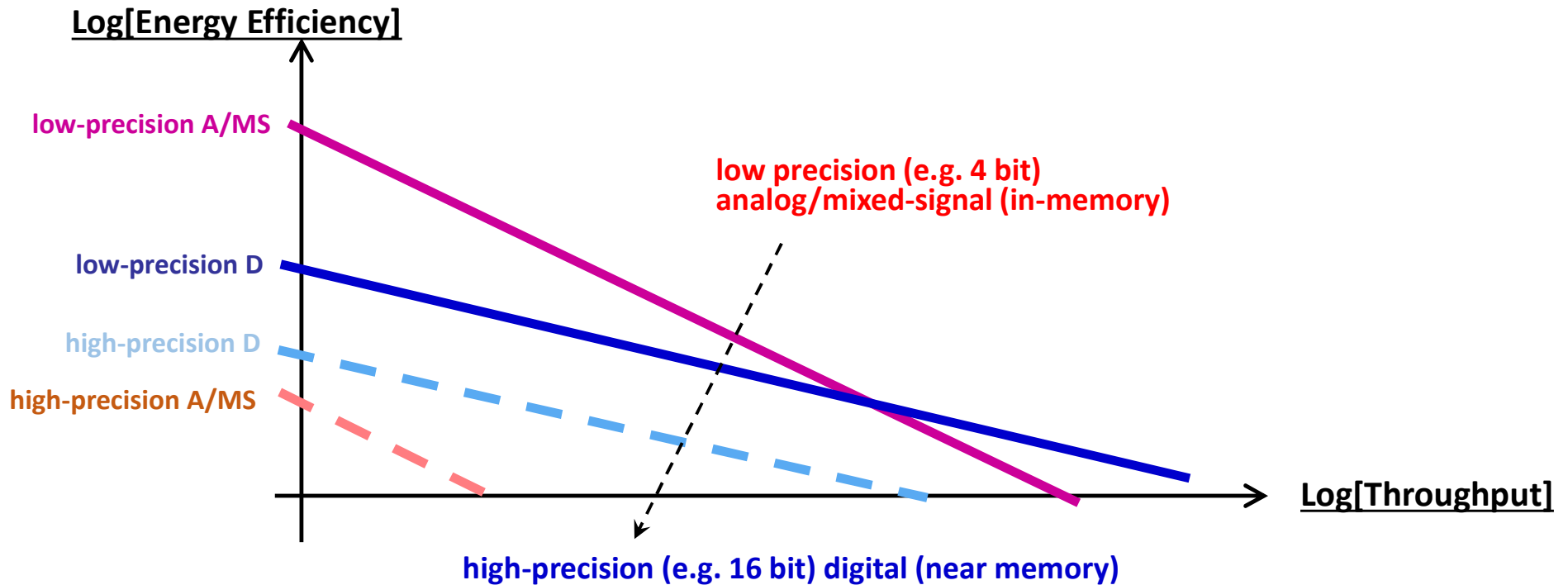


Tradeoff between Energy-Efficiency and Throughput



Natural tradeoff between energy efficiency and throughput (at device, circuit, architecture levels)

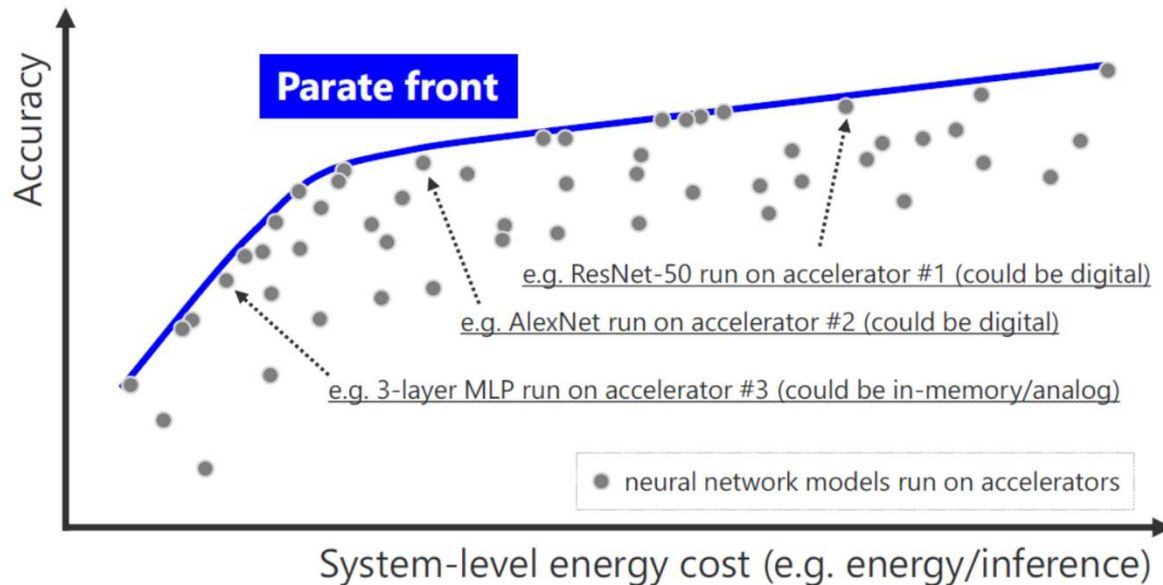
Tradeoff between Energy-Efficiency and Throughput



Mixed-signal for low-to-medium precision & EE optimal, digital for higher precision & throughput-optimal

Proper Way to Compare Performance Metrics

- Tradeoff between energy-efficiency and functional performance (image classification accuracy)



Source: J. Deguchi et al, IEDM'19

Report chip or system level metrics, e.g.

- Frames per seconds per area at certain accuracy
- Energy per inference at certain accuracy

Oblivious to type of network / computing precision / type of hardware

TOP/s/cm² or TOP/J as proxy could be very misleading

Useful applications for lower accuracy?

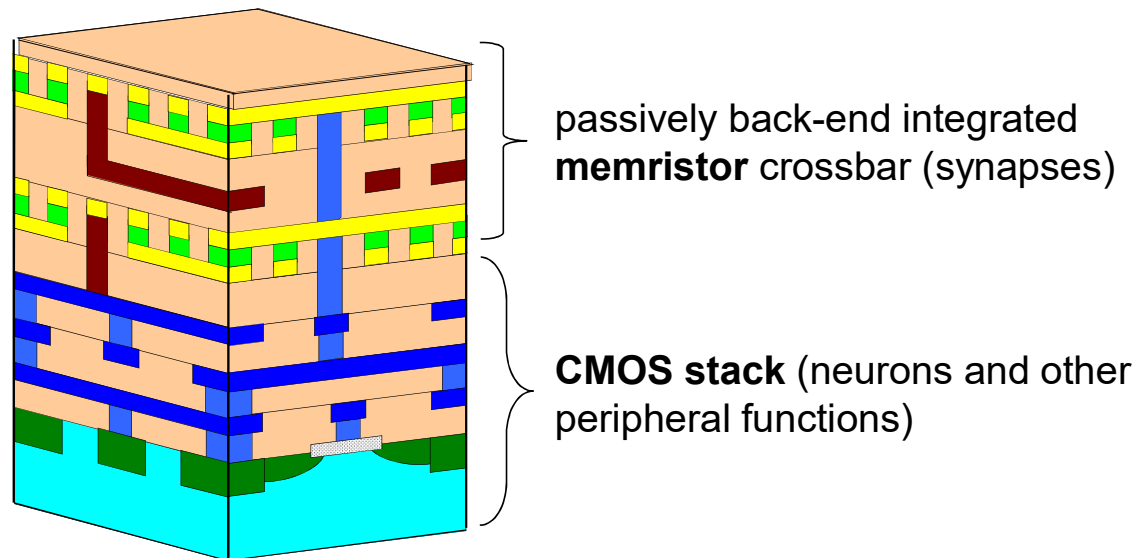
Part II: Key Takeaways

- Memristors are enabling element for VMMs
- >100x better chip-level energy-efficiency for inference accelerators due to
 - compact footprint & tunability of passively integrated memristors
 - “in-memory” computing in analog VMM circuits
- Digital conversion circuits, needed to take advantage of weight reuse, reduce processing throughput

Part III.
**Device Requirements and
Challenges**

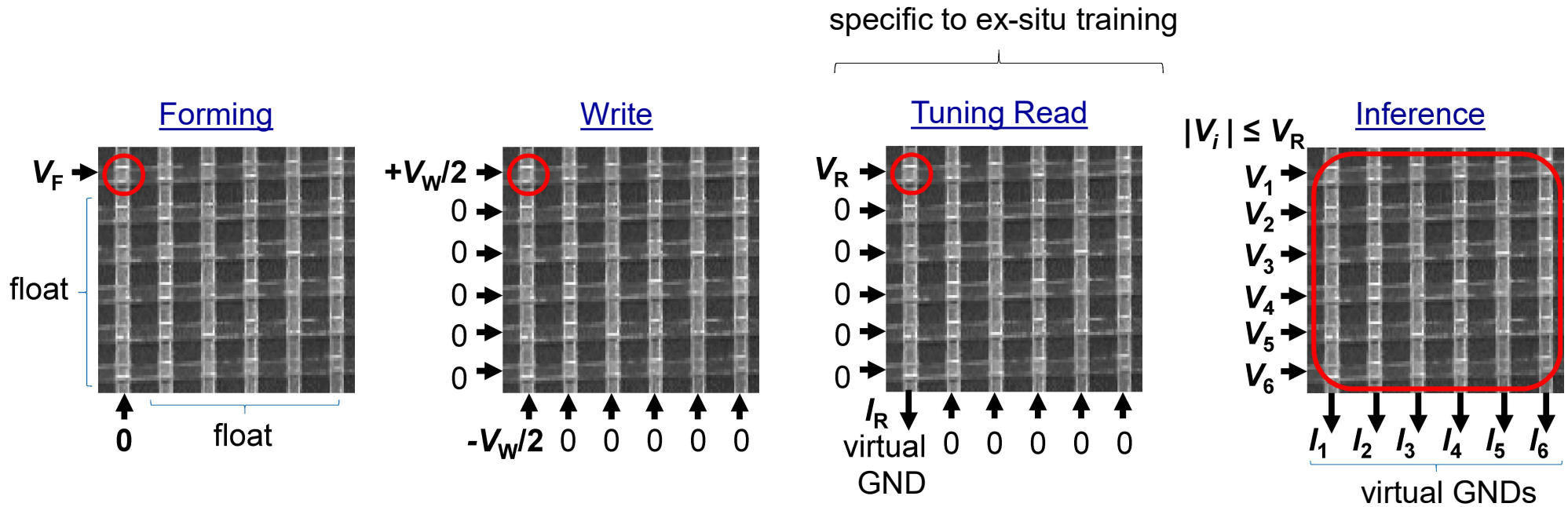
CMOS Compatibility

Hybrid monolithic 3D hybrid circuits using passively integrated devices as an ultimate goal



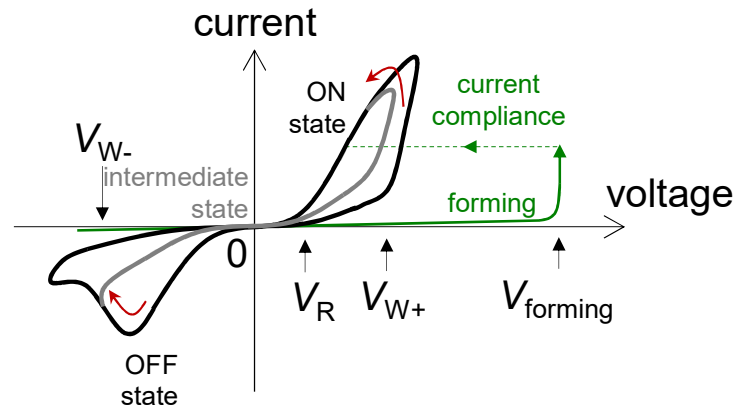
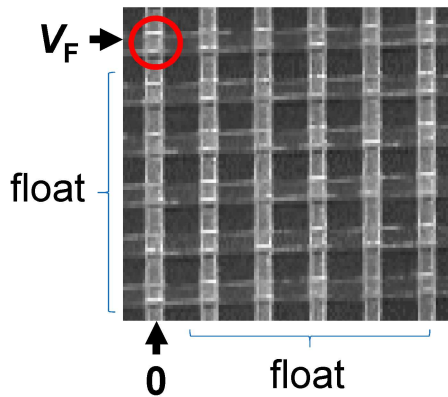
Device Requirement: Compatible with CMOS backend memristor integration process and memristor operation

Key Operations on Memristor Crossbar Circuits



Crossbar Compatible Forming

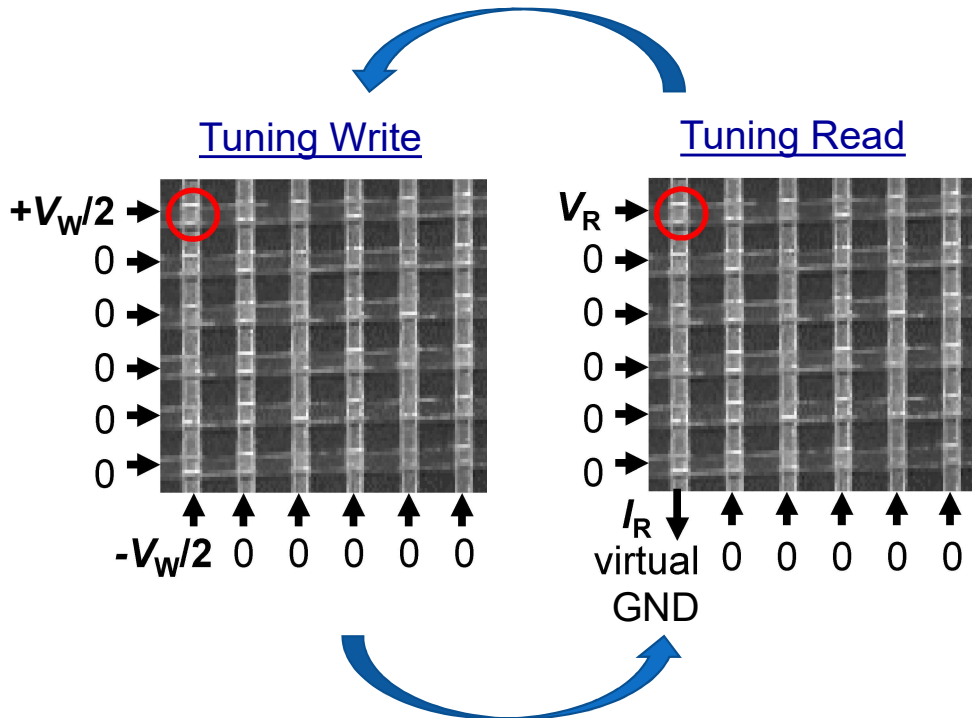
Forming



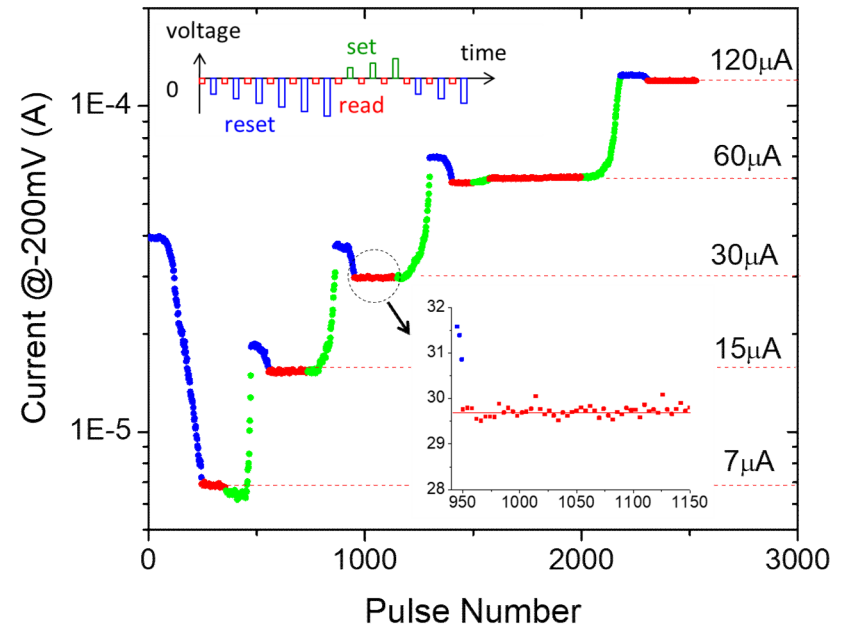
Large, potentially destructive currents through already formed crosspoint devices when $V_{forming} \gg V_w$

Device Requirement: Forming process compatible with passive crossbar circuits (relaxed for 1T1R circuits)

Endurance, Switching Speed/Energy



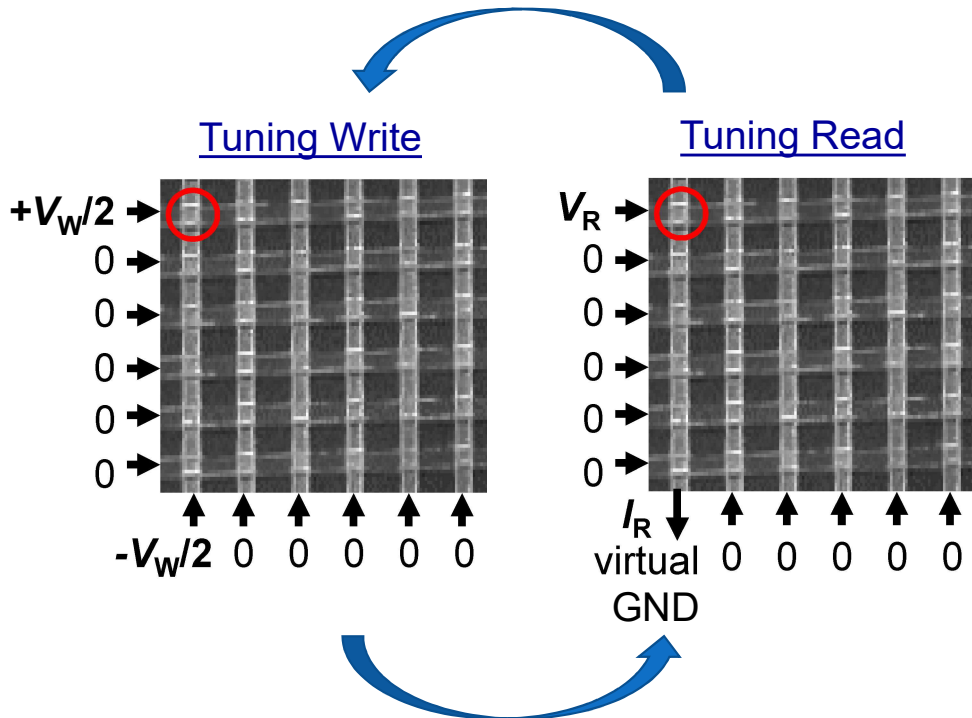
- Example of tuning device to different states



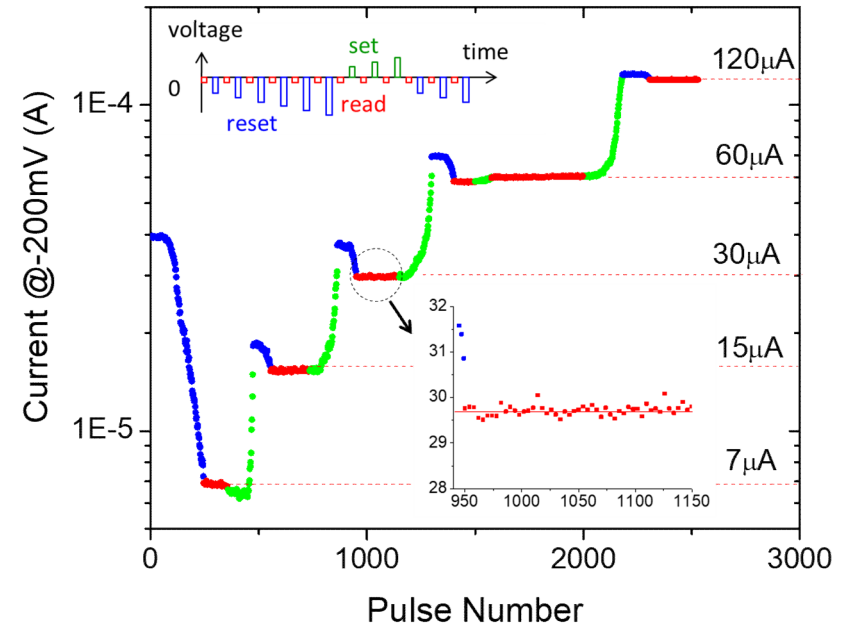
For ex-situ trained inference accelerators weights are programmed (tuned) infrequently

→ Acceptable to use slow / power hungry write-verify algorithm, which adapts to the variations in the device I-Vs

Endurance, Switching Speed/Energy



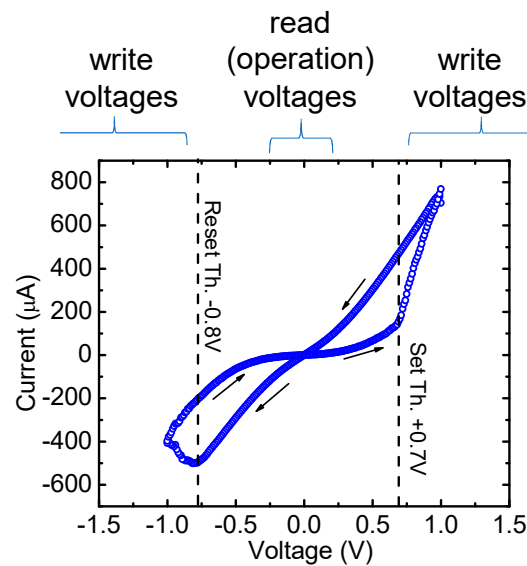
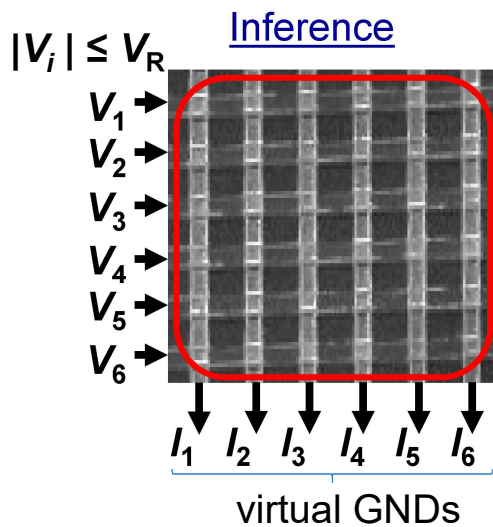
- Example of tuning device to different states



Write speed / energy, switching endurance, switching dynamics are not important for inference
 Less severe requirements for device variations for inference

Static I-V linearity

- Typical I-V for in-house devices

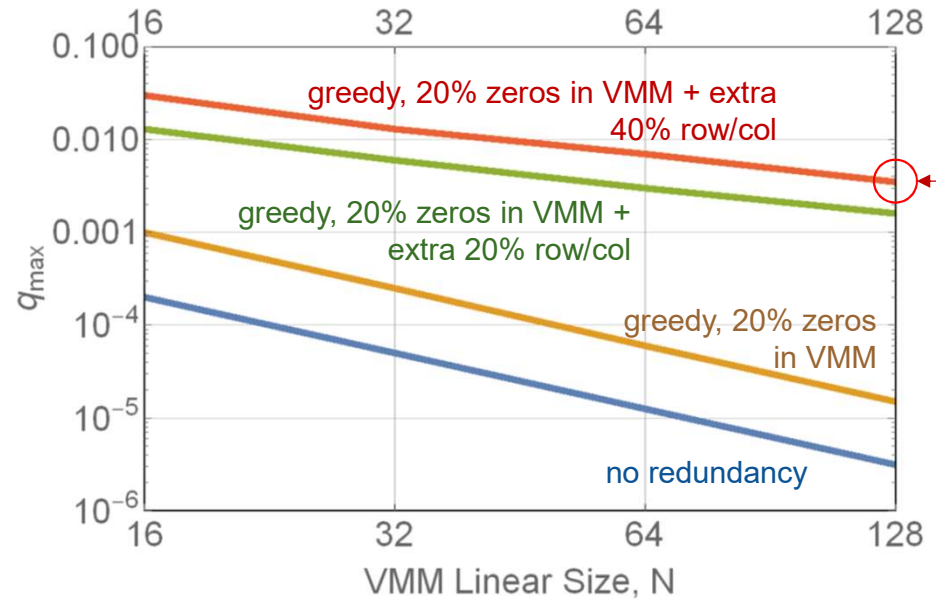
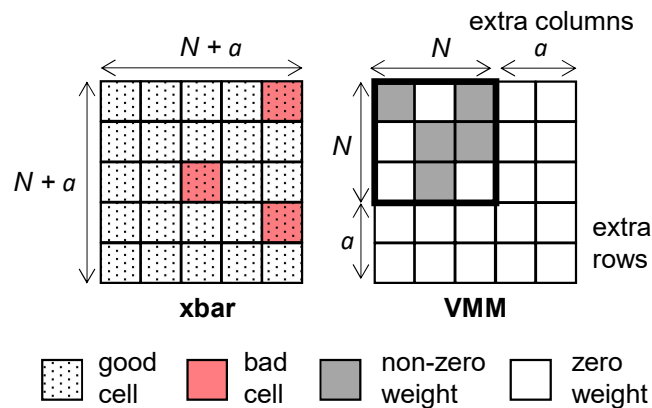


- Linear I-V is needed for precise multiplication, e.g. $I(V_R) / (2I(V_R/2)) > 0.95$ for 4-bit precision
- I-Vs are typically very linear at small (non-disturbing) voltages

Major Challenge #1: Poor Yield of Memory Devices

- **Impact of bad devices:** study of finding the largest fraction of bad devices for 95% successful mapping

- Defect tolerance by searching for good sub-array with greedy row/col permutations and utilizing redundancy (VMM sparsity and provisioned extra rows / columns)

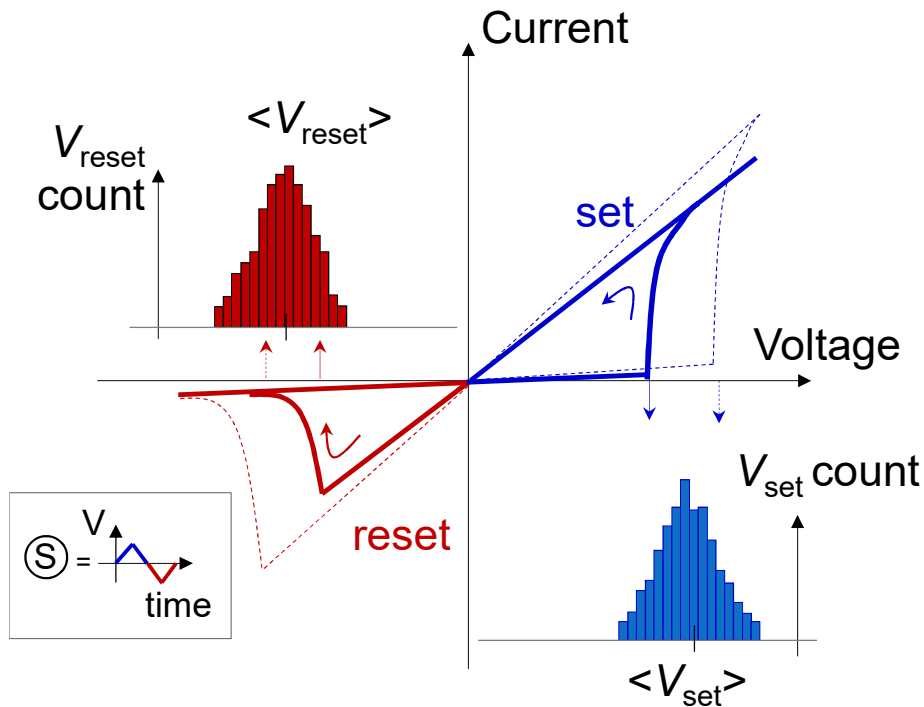


< 0.5% bad devices with 60% redundancy for 128x128 VMM

- Simplified model assuming stuck-on-off bad devices (failed to form) or out-of-spec (with poor retention, high noise)
- Some overhead for permuting block's input/output
- Higher tolerance with chip-in-the-loop / defect-aware ex-situ training (*F. Merrikh Bayat et al, ICCAD'19*) but less viable commercially

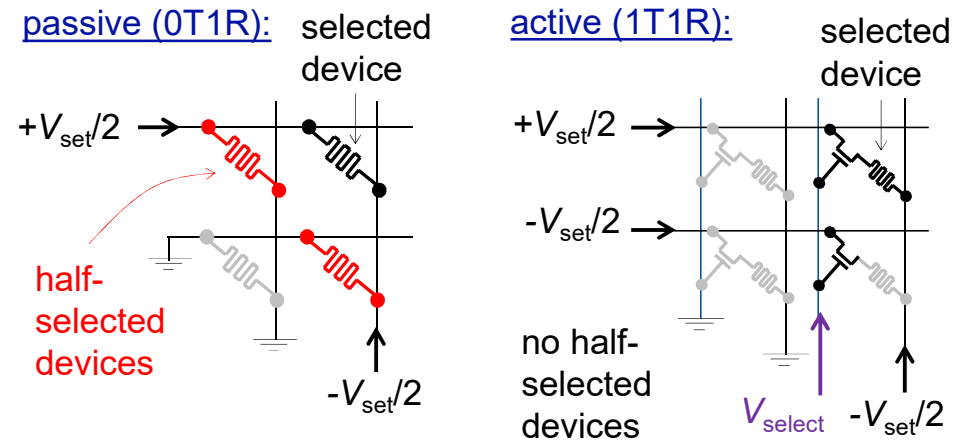
Major Challenge #2: Switching Threshold Variations

Switching voltage threshold variations



- Switching threshold = voltage at which current changes by $> 10\%$ when applying voltage ramp

Half-select disturbance

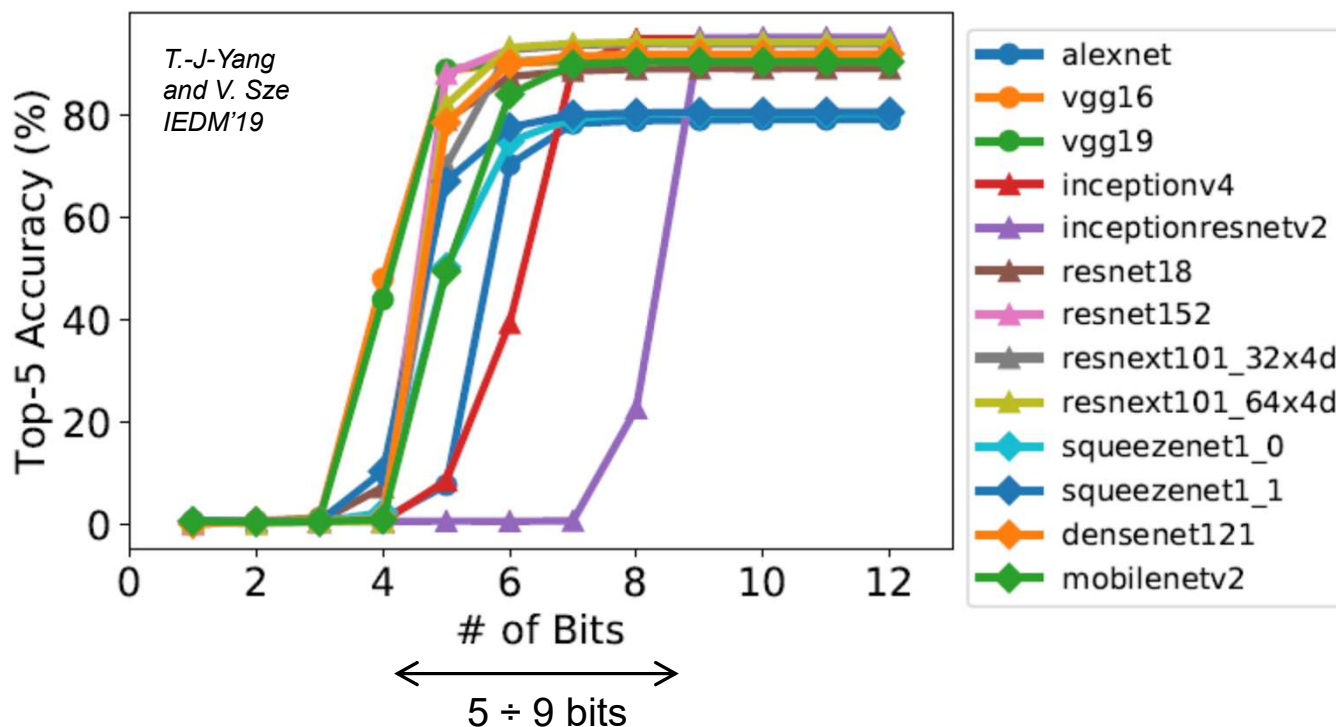


- Disturbance of half-selected devices when tuning devices with the largest voltage threshold
- at least c.v. < 0.3 for "V/2-biasing" scheme, with σ margins, and naïve algorithm to avoid disturbance
- No such problem for 1T1R memories (at the cost of 100-1000x larger cell so far)

D. Strukov, Nature Materials, 2019

Weight Tuning Requirements for Inference Applications

Impact of weight precision on ImageNet classification



Further improvements possible, e.g. perform some critical operations in digital domain

No loss in performance for ~ 3 % ÷ 0.2 % tuning precision in the dynamic range

$$\text{Tuning Precision} = 100\% \times \frac{|G_{\text{desired}} - G_{\text{actual}}|}{(G_{\text{max}} - G_{\text{min}})}$$

Major Challenge #3: High Cell Current

Why it is a problem?

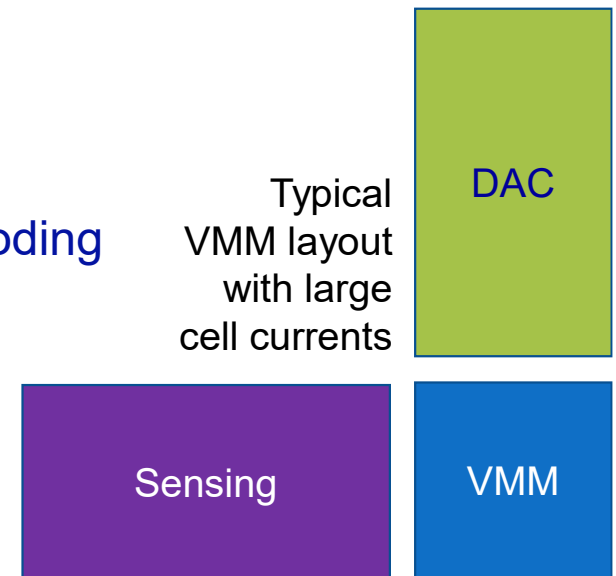
- larger operating (read) currents → smaller input/output array impedance → larger DAC/sensing overhead
- large write currents limits the size of the passive crossbar arrays due to IR drop or swells access transistor in active arrays

Pros of some active cells:

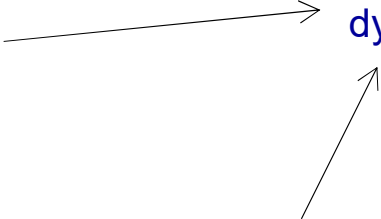
- high input/output array impedance for all 1T schemes
- high input impedance for 1T1R schemes with linear encoding
- no leakages/IR drops problem during forming/write

Optimal for operation: 100s pAs to 10s nAs

- too low SNR (compute precision) for smaller currents



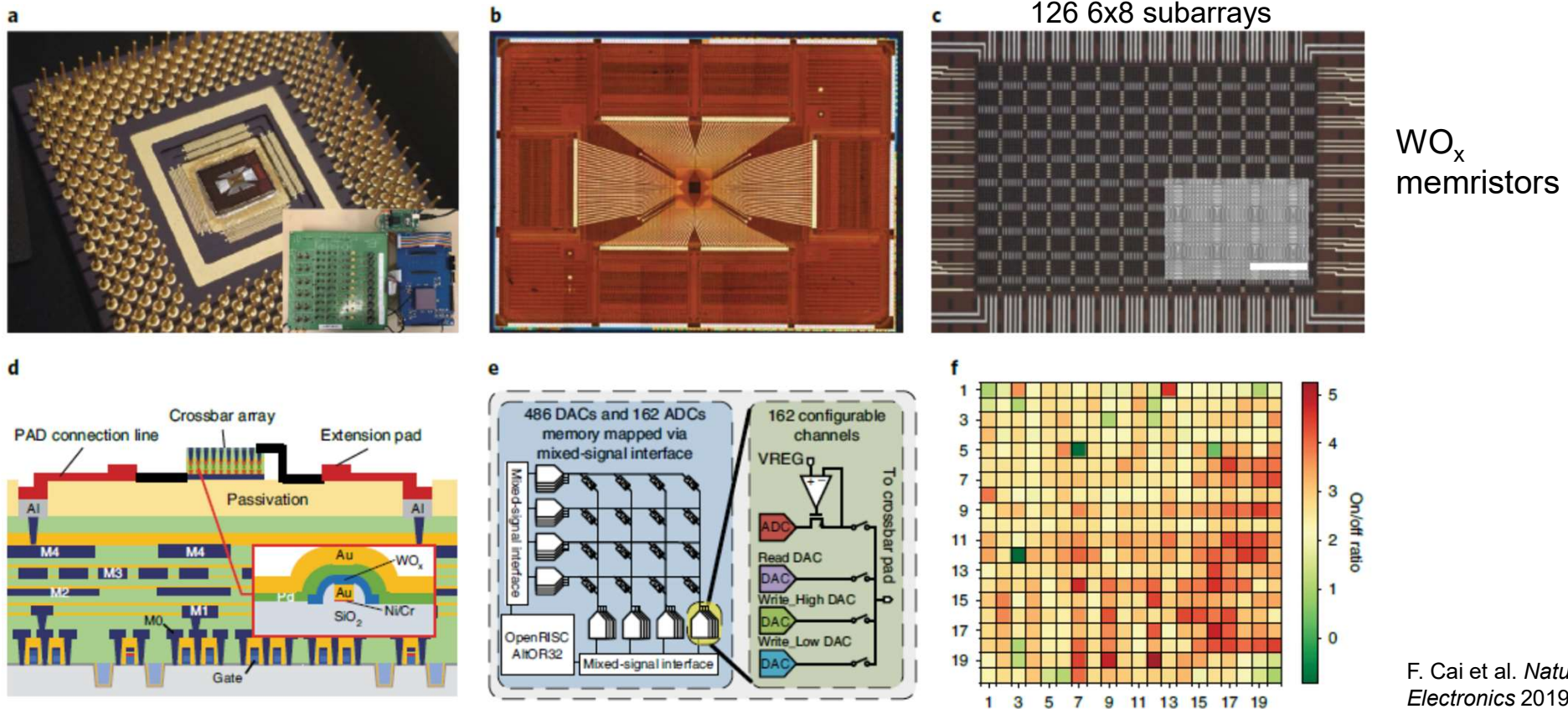
Part III: Key Takeaways

- For the simplest, yet most practical neuromorphic inference applications, the most important metrics:
 - CMOS and crossbar compatibility
 - density, especially for energy-efficiency optimal designs
 - multi-level analog memory (4 ÷ 8 bits or 32 ÷ 256 states)
 - high retention (~ months)
 - Key challenges are
 - poor device yield and I-V uniformity (need < 1% bad devices, $\sigma V_{sw} \ll V_{sw} / 2$)
 - high switching / operating currents (need 100s pAs to 10s nAs for operation)
 - Less important / desired specs ease to achieve: dynamic / static I-V linearity, write speed/energy, endurance, noise (RTN could be too high for some devices), ON/OFF ratio
 - Much more demanding device uniformity requirements for on-line/in-situ learning, e.g. SNN with STDP learning, though relaxed retention requirements for training accelerators
- < 3% tuning accuracy in the dynamic range
- 

Part IV.

Examples of Recent Mixed-Signal Neuromorphic Hardware Prototypes

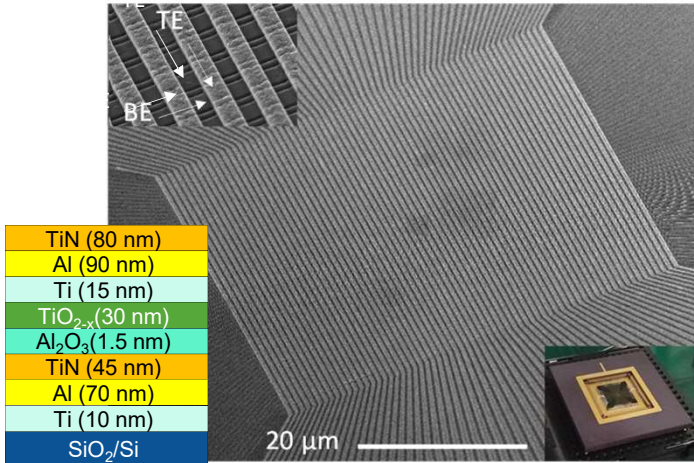
UMich's Chip with Backend-Integrated Metal-Oxide Memristors



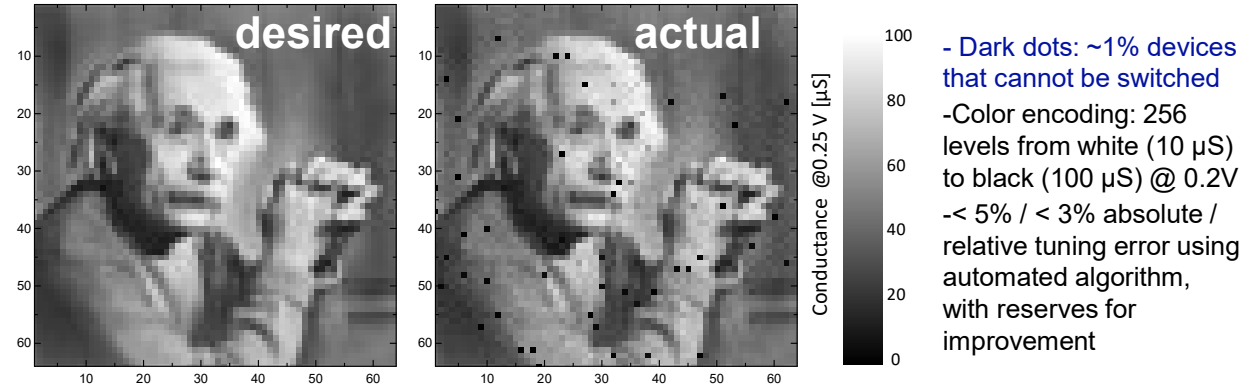
- The first fully-integrated CMOS / 0T1R memristor chip of its kind
- Poor \sim min-scale retention, only small fraction (16 \times 14) of memristors used at demo, very limited statistics, poor 600 μm^2 per cell density, small \sim 2.5 ON/OFF range

UCSB's Metal-Oxide Memristor Chip

Passive 64 × 64 xbar circuit



High precision conductance tuning

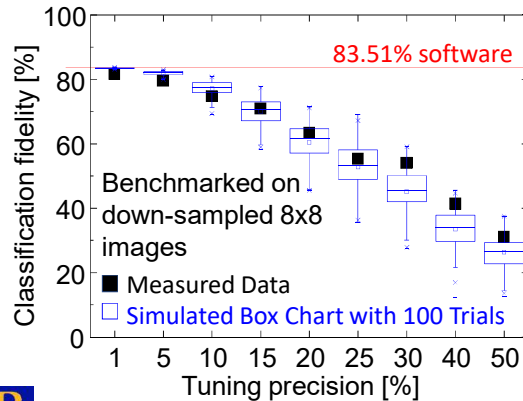


H. Kim *et al.* 2019 (unpublished)

Other work on classifier demos: M. Prezioso *Nature*'15, *IEDM*'15, F. Merrih Bayat *Nat. Comm.* 18

Recent work on 3D circuits: I. Kataeva *ISCAS*'19, B. Chakrabarti *Sci. Rep*'17, G. Adam *TED*'17

MNIST classification demo



Highest complexity analog-grade passive xbar demo

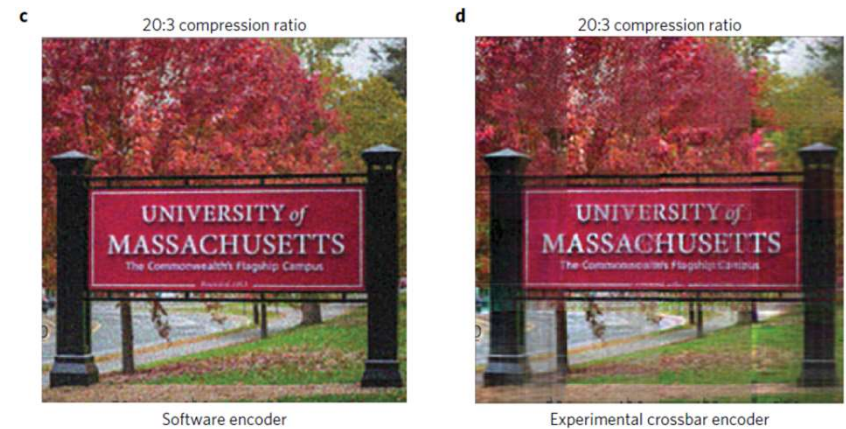
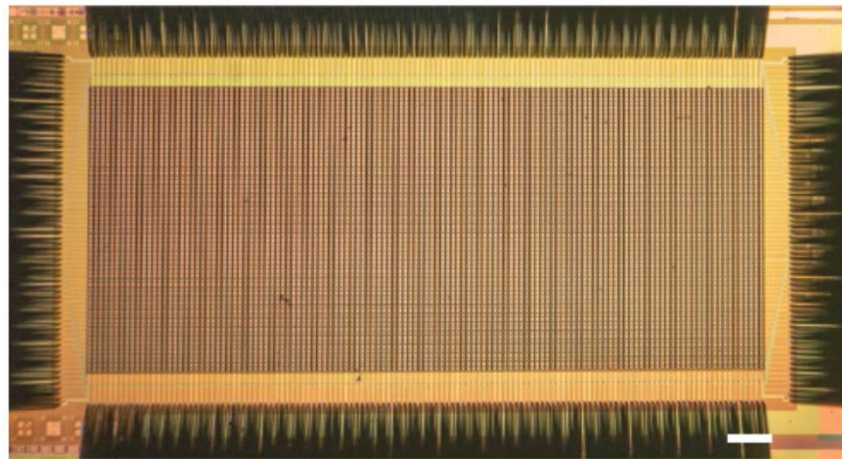
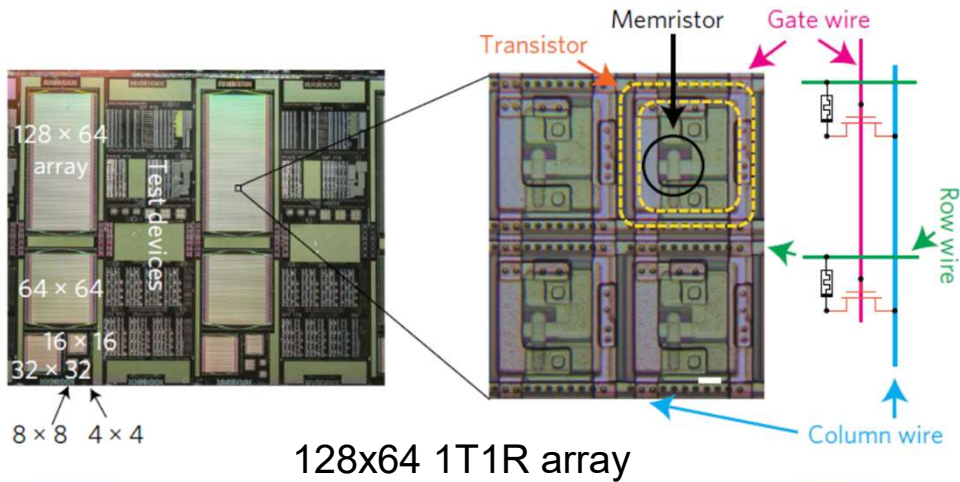
Details:

- Al₂O₃/TiO_{2-x} active bilayer by reactive sputtering
- ~250 nm wide lines, passive (0T1R) integration, 0.25 μm² per memory cell
- CMP/dry etching and TiN/Al electrodes for higher conductance
- Uniform I-Vs (~1% unswitchable devices, ~26% variation in threshold)

Important future steps:

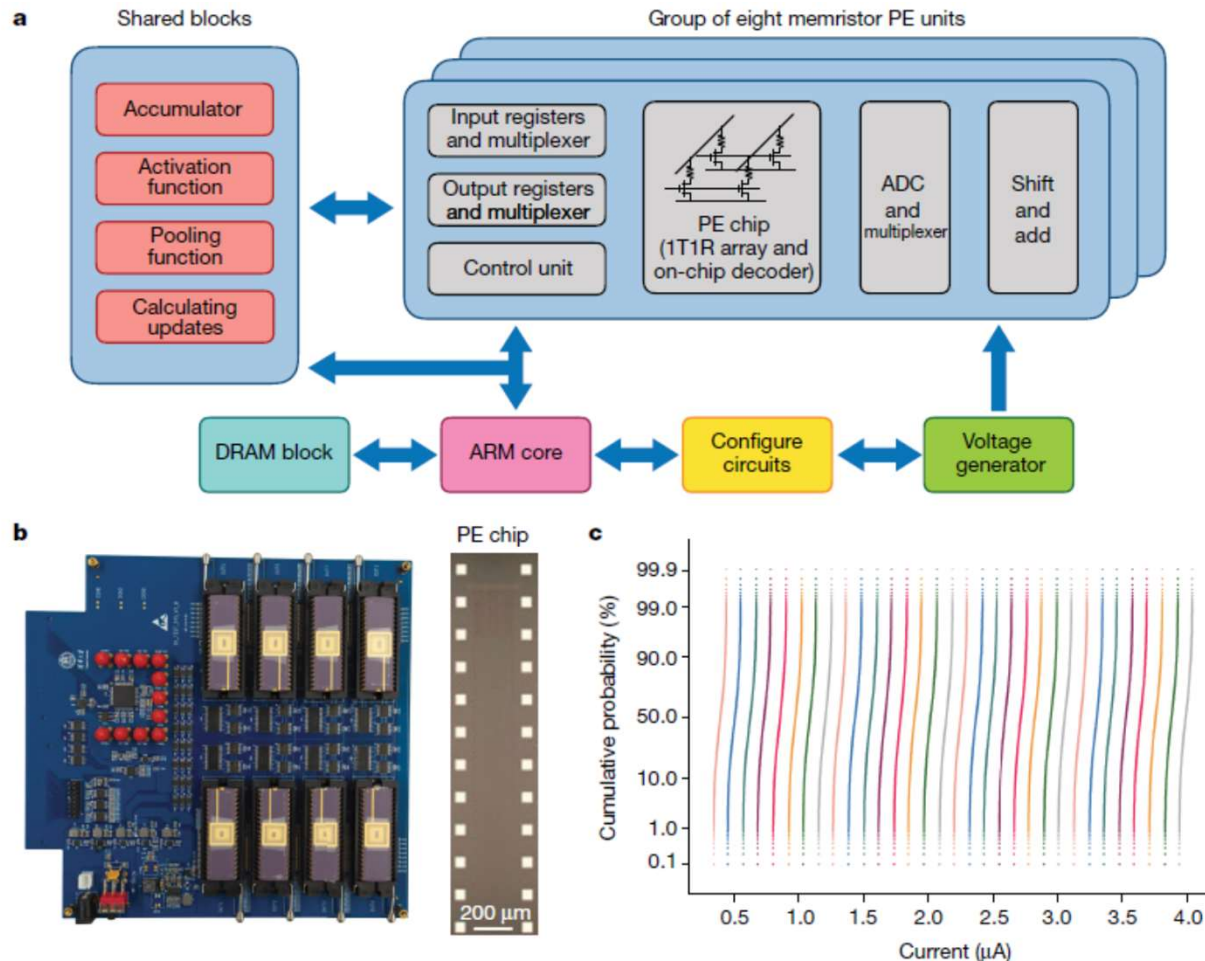
- 3D monolithic integration with CMOS (ongoing E2CDA work)
- Lowering device resistance (by feature size scaling)

HPL / UMass 1T1R Metal-Oxide Memristor Chip



- The first fully integrated CMOS / 1T1R HfO₂ memristor chip of its kind
- Very high >99% yield, linear I-V, excellent analog properties
- Used in many impressive demos (inference, training, reinforcement learning, unsupervised learning)
- Extremely bulky ~2,500 μm² cells, High (~ 1 mA for ON state) currents

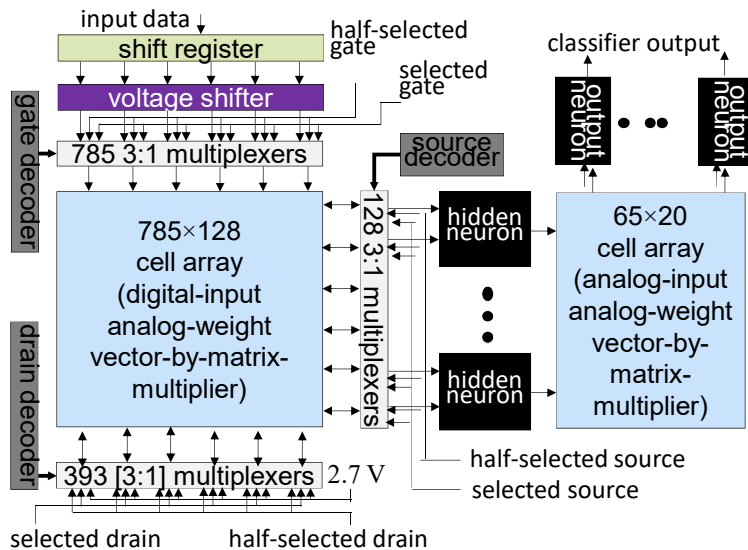
Tsinghua U's 1T1R Memristor Chip for Neurocomputing



- Largest scale demo of its kind
- Board-level integrated CMOS neural network circuitry with 8 1T1R memristor chipw
- Each chip is 128x16 1T1R array of TiN/TaO_x/HfO_x/TiN devices
- Used to demonstrate convolutional neural network
- Extermely bulky ~ 200 μm² cell

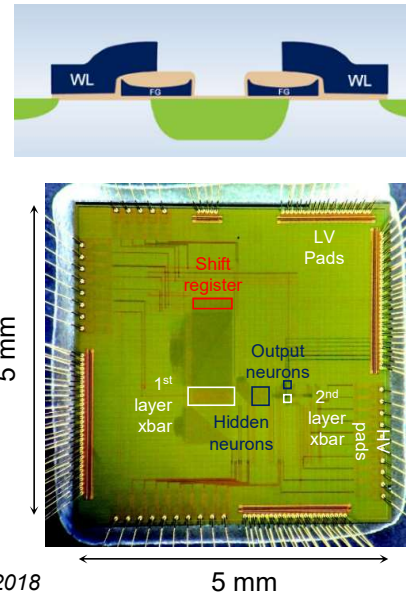
Neuromorphic Inference with 2D NOR flash

High-level architecture

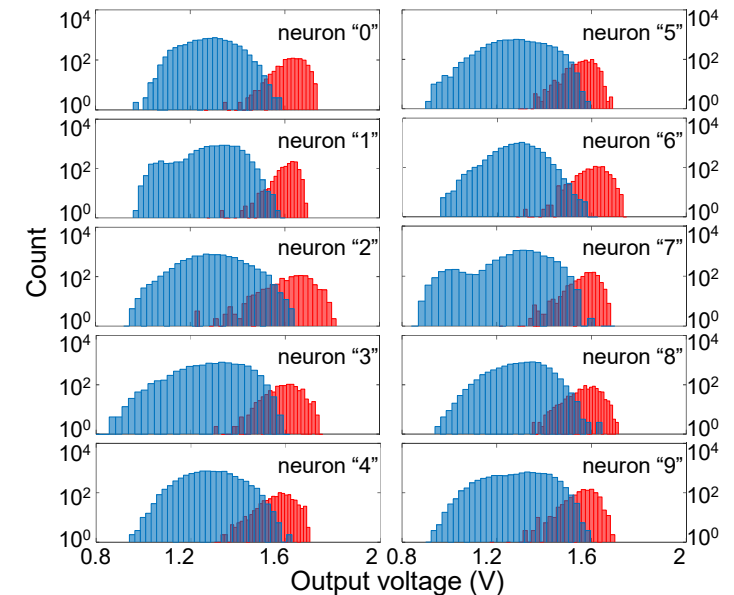


X. Guo. et al., IEDM, 2017; F. Merrih Bayat, IEEE TNLS, 2018

NOR eFlash chip



Measured results (10,000 MNIST test patterns)



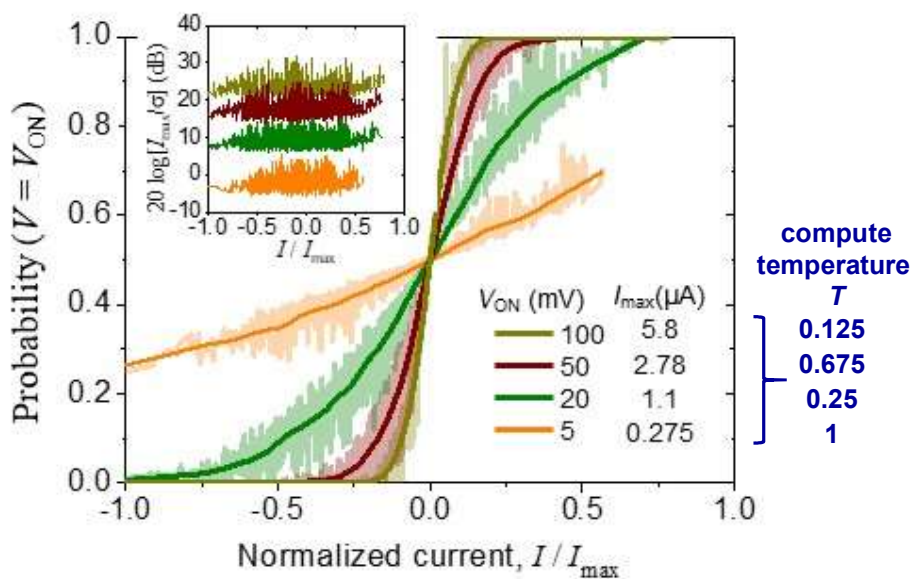
Summary:

- 28x28 B/W input, 10-class output, >100,000 NOR flash synapses, 64 hidden layer CMOS neurons, 180-nm process with eFlash
- 94.65% experimental fidelity (96.5% theoretical)
- < 1- μ s latency, < 20 nJ energy per pattern (reserves for improvement for both with better neuron design)
- Much better in speed and energy efficiency over digital circuits at comparable MNIST fidelity (10^6 better energy-delay than IBM TrueNorth)
- Reproducible, temperature insensitive, no change in performance after 7 months
- More recent work using 55-nm ESF3 NOR-flash technology (CICC'17, IEDM'18'19), scalable to 28 nm

Solving Optimization Problems with Hopfield Network

Stochastic dot-product circuit:

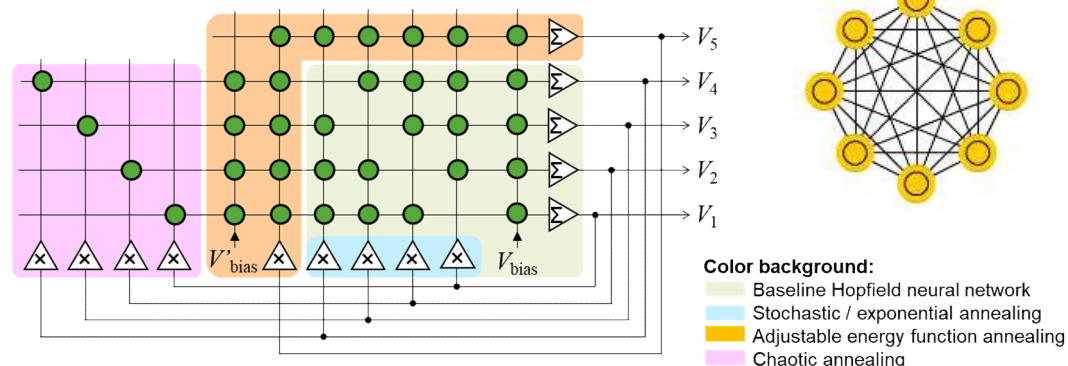
add intrinsic/extrinsic noise from memory array to dot-product current and feed it to comparator



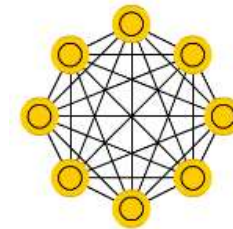
M.R. Mahmoodi et al. Nature Communications, 2019

- Sigmoid slope (i.e. SNR or compute temperature T) controlled dynamically by the applied voltage V_{ON}
- Estimated >10,000x improvement for energy-delay metric over competitive approaches

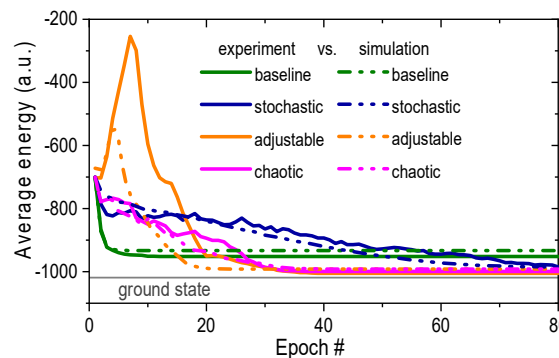
Basic idea of the demo:



Hopfield Network (HN)



Experimental results for weighted graph partitioning



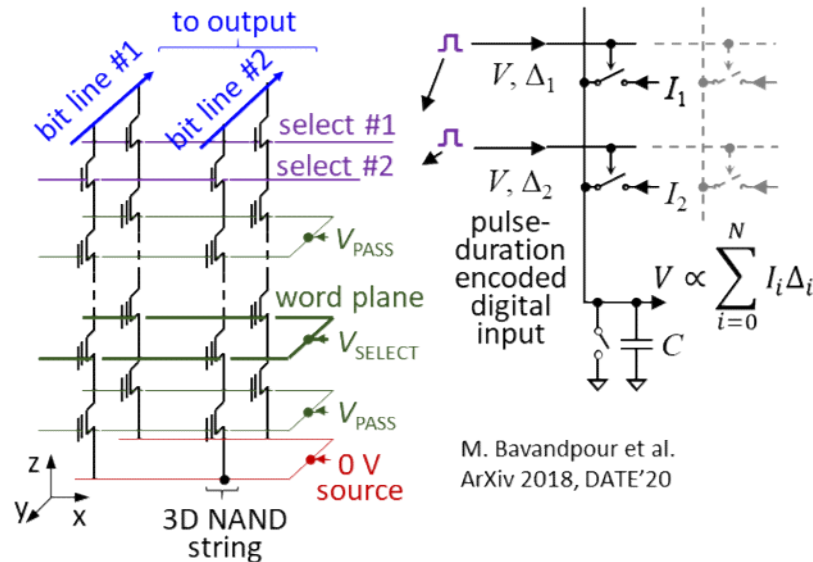
More results using 64x64 passive memristor xbar

- weighted max-clique
- weighted vertex cover
- independent set
- weighted graph partitioning

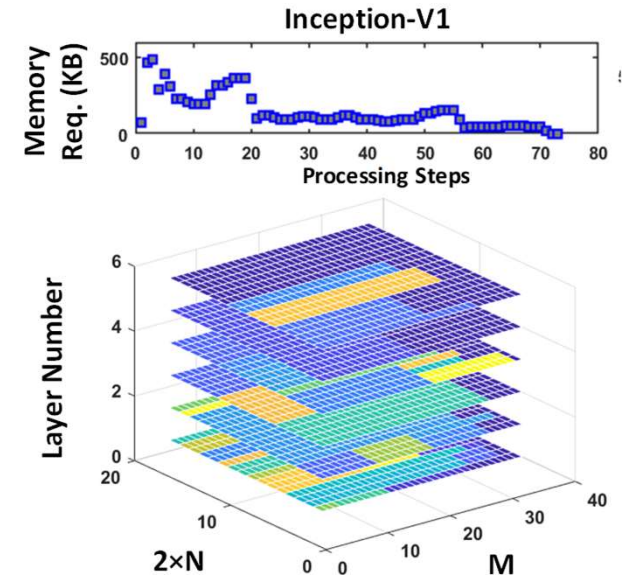
M.R. Mahmoodi et al. IEDM'19

3D NAND Inference Accelerator for Large-Scale Models

- Vector-by-matrix multiplier with native 3D NAND memory blocks



- Mapping common neural models



	TPU	3D NAND (64 layers)
Technology node (nm)	28	55
Precision (bit)	8	4
Area (mm ²)	330	18
Energy efficiency (TOp/J)	0.43	70
Throughput (TOp/s)	92	11

- Experimental results, excluding the off-chip weight transfer overhead. Google's Tensor Processing Unit (TPU) is optimized for throughput
- #Results of computer simulations of 3D aCortex architectures, with all weights stored on chip, optimized for energy efficiency

Summary of Recent Memristor Prototypes

based on filamentary (nonvolatile) memristors

Tech.	Switching Material	Refs.	Xbar Size	Yield (%)	Functional Demo	Cell Size (μm^2)	Average Forming ¹⁰ Current (μA) / Voltage (V)	Retention ¹ (@°C)	Endurance ¹¹ (cycles)	Tuning Precision	Set Switching Statistics (μ , σ)	$G_{\text{max}}/G_{\text{min}}$ ($\mu\text{S}/\mu\text{S}$)	Comment on the fabrication process
0T1R	Ta ₂ O _{5-x}	[49]	18×2	100	18×2	-	250 / ~1.1	-	-	-	-	1500/850	lift-off /Pd electrodes
		[38]	16×3	78	4×3	-	1000 / ~2	-	>100 k	-	1.25, 0.1	1800/1300	lift-off /Au and Pd electrodes
	WO _x	[47]	8×8 ⁸	-	-	0.01	-	-	-	~50 % ⁷	3.5, -	4/0.1	lift-off/ /Ag electrodes
		[41]	11×3	-	11×3	-	1000 / ~1.8	-	-	-	0.85, 0.05	-	lift-off /Pd electrodes
		[29,39,42]	32×32	-	25×20, 5×10, 4×4	~9	>170 / NA	-	-	~35 % ²	1.7, -	3/1 ¹⁵	lift-off /Au and Pd electrodes
	[30]	108×54 ⁶	-	16×14, 4×4	~600 ⁵	-	< 1 min	-	-	-	2.4/1	lift-off /Au and Pd electrodes	
	SiGe	[48]	14×1	-	1×1	100 ¹⁴	- / > 4 V	> 48 h @85°	>1 M	-	4, 0.15	40/0.16	High temperature epitaxial growth
	TiO _{2-x}	[20]	10×2	100	10×2	1	100 / 6	-	-	~2 %	1, 0.1	500/50	Wired XBAR/ lift-off/Au electrodes
		[5,43]	12×12	100	12×12	0.16	200 / 1.9	>10 years @30	>200 k	< 3%	0.9, 0.1	200/6	lift-off
		[12,15]	20×20	100	20×20	0.25	220 / 1.5	>20 h @120	>1 M	< 2.5%	1.0, 0.18	200/6	lift-off
[19,44]		2×10×10	100	2×10×10	0.49/2	100 / 2.5 ⁹	>25 h @100	>1 M	< 1%	1.1, 0.15	100/0.1	Ar IBE	
[28,17]	64×64	~99	64×64	0.25	100 / 3.2	>20 h @100 >10 years @RT	>1 M	< 5%	1.2, 0.13	200/6	Fully CMOS-compatible using etch-down planar process		
1T1R	HfO ₂	[13,36,37,46]	128×64	>99 ¹³	128×64	~2500 ¹²	-	>10 years @RT	-	<3.1 % ⁴	2, -	900/100	lift-off
		[35]	128×8	-	960	-	>150 / >3 V	-	-	< 35 %	-	40/5	lift-off
		[40]	1K	-	448	~25	-	~1 m @30	-	~20 %	3.5, -	0.01/100	-

¹ The retention data is the *minimum* reported number and does *not* necessarily determine the time-to-failure. ² Based on Fig. S3 in the supplementary information of the [29] and it is not clear if the data are obtained after tuning of all cells or recorded after programming each cell. ³For the top stack, average forming voltage/current for the bottom layer is 2 V/50 μA . ⁴ Based on Fig. S3 in supplementary information of the [46]. ⁵ Based on the cross-sectional SEM image in Figure S10. It is 256.8 μm^2 based on the data in supplementary note 10. ⁶ 126 subarrays of 6×8 is used to build an array of 108×54. ⁷ Based on Fig. 4d in [47] and conductances are measured after programming each subarray, not the whole crossbar. ⁸ A 40×40 array is developed by combining 25 subarrays of 8×8. ⁹ Data for the low-resistance state with sensible gradual retention loss at high resistance levels. ¹⁰ Maximum set voltage is used if forming statistics are not reported. ¹¹ Data might not be comparable among different works since the test conditions have been different. ¹² Based on the SEM figures provided in Fig. 1c of both [46] and [16]. ¹³ Based on Fig.1c in [36]. ¹⁴ Denser single devices are reported too, but most experimental results are for 5 μm × 5 μm devices. ¹⁵ This represents the average range of conductance values observed in the crossbar. There is a significant variation between different devices.

Part V.
Concluding Remarks

Drawing Inspiration from Human Brain for Future Neuromorphic Hardware Systems



- Is the brain structure a result of a fortunate fluke? Or it represents the best solution among different possibilities?
- Are all principles behind brain operation useful for algorithm / HW design?

Drawing Inspiration from Human Brain for Future Neuromorphic Hardware Systems



- Human brain is
 - not superfast
 - efficient only in tasks that help to survive (e.g., cannot do number crunching)
 - outperformed by latest algorithms in speech and image (by ~4%) recognition on well defined benchmarks

Drawing Inspiration from the Human Brain for Future Neuromorphic Hardware Systems



- Signal propagation in neuronal axons compared to integrated circuits is
 - much slower (<100 m/s cf. $\sim 10^7$ m/s)
 - much less energy efficient (> $1\mu\text{J}$ cf. <50 fJ per bit per mm)
- Spike encoding is the necessity rather than useful feature?

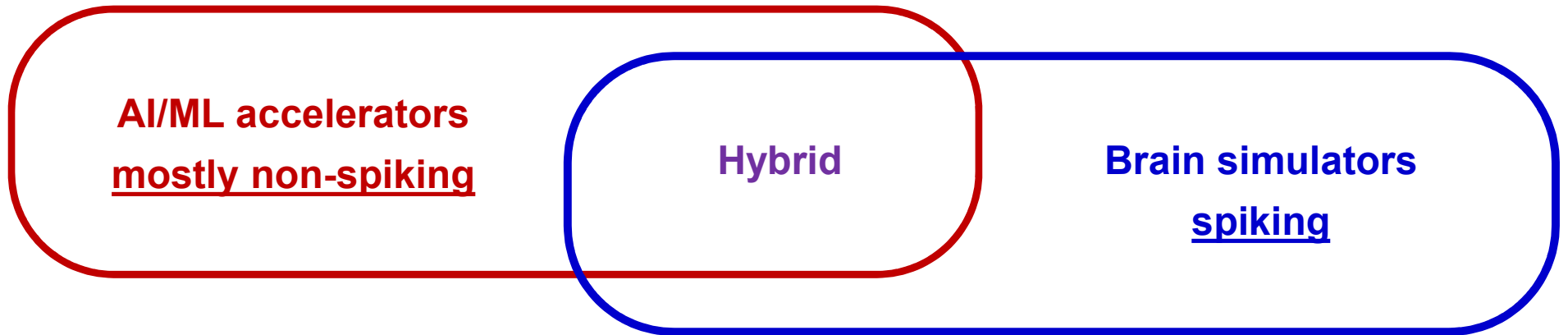
- Brain has huge variability (because it is the most energy-efficient way to reproduce?) which is tolerated by massive space/time redundancy with unique and slow “chip-in-the-loop retraining”
- Sparse encoding is the necessity rather than useful feature?

Drawing Inspiration from Human Brain for Future Neuromorphic Hardware Systems



- Is the brain structure a result of a fortunate fluke? Or it represents the best solution among different possibilities? (looks like a fluke)
- Are all principles behind brain operation useful for algorithm / HW design? (definitively no)
- Not all tricks the brain uses for computing are useful
- However, we are still missing the holy grail human-intelligence algorithm

Artificial Intelligence / Machine Learning Hardware



practical (boring) HW for today's ML/AI algorithms

e.g. dynamic vision sensors with spiking frontend and non-spiking backend

just like GPUs helped the revolution in ML, efficient spiking HW could lead to breakthrough in advanced AI algorithms

need efficient implementation of vector-by-matrix operation for all

Summary (I)

- Major memristor challenges:
 - poor yield
 - poor device uniformity
 - high cell currents
- Much more severe uniformity requirements and additional device challenges (endurance, write energy) for training accelerators and on-line/in-situ learning, e.g. SNN with STDP learning

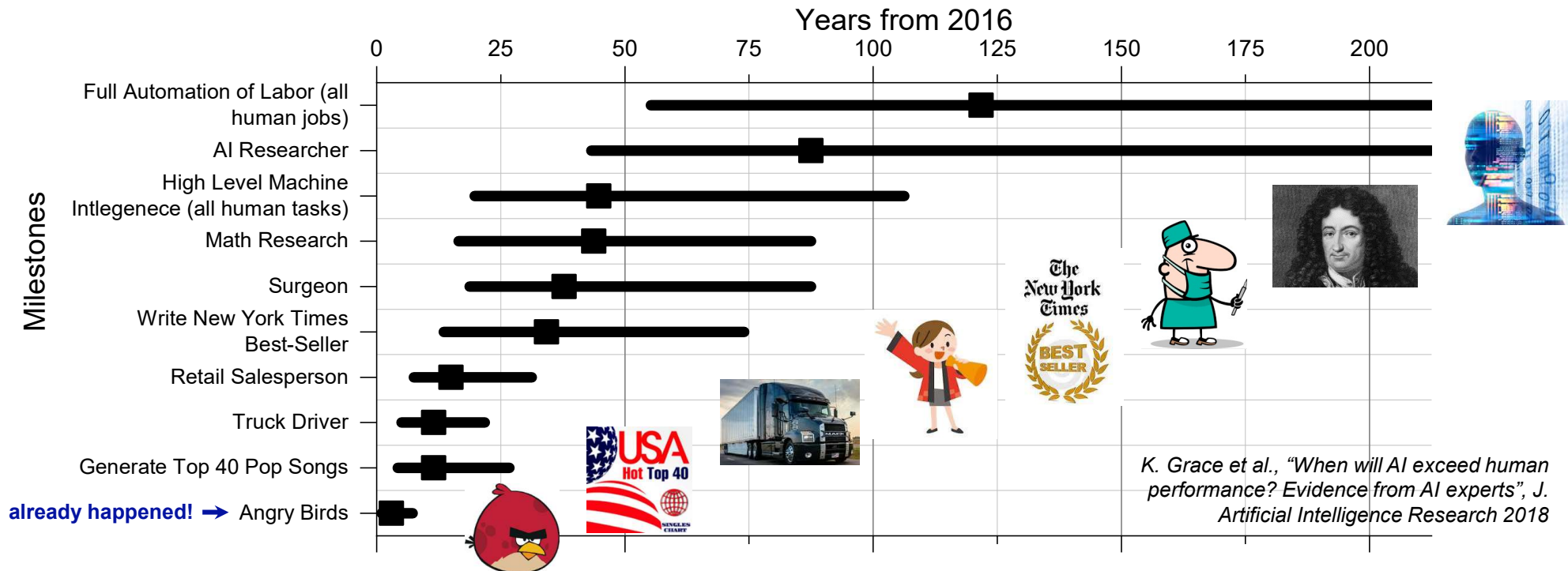


Summary (II)

- Neuromorphic inference with ex-situ training as natural entry-level application of mixed-signal neural networks
 - the simplest in terms of device requirements, yet very practical
 - at least 100x better in energy-delay over purely digital system according to the experimental results for small-scale system, and system-level projections to bigger systems
- Most promising memory technologies
 - Long term: passive 3D memristors, 3D NAND
 - Short term: embedded NOR flash, 1T1R memristors
- More advanced networks:
 - Straightforward extension to inference accelerators for more advanced approaches (stochastic neural networks, neurooptimization, spiking neural networks)
- Need biologically plausible neuromorphic hardware for brain simulations
- Novel applications driven by analog circuits?

When Will AI Take Over Humans?

(survey of 352 expert researchers published at NIPS'15 and ICML'15)



- 50% chance of automating all human jobs (better or more cheaply with AI) in 120 years
- Most optimistic AI progress predictions in Asia, least optimistic in North America

AI won't replace but rather empower human kinds! (hopefully ;)

Acknowledgments

Current members at my research group at UC Santa Barbara:



Hae Jin Kim



Tinish Bhattacharya



M. Reza Mahmoodi



Nikita Buzov



Mohammad Bavandpour*



Zahra Fahimi



Shabnam Larimian



Michael Klachko*



Subham Sahay*

Key collaborator: Konstantin Likharev (Stony Brook University)

* Graduating / leaving group at the end of the Spring 2020

Sponsors (past and present):

