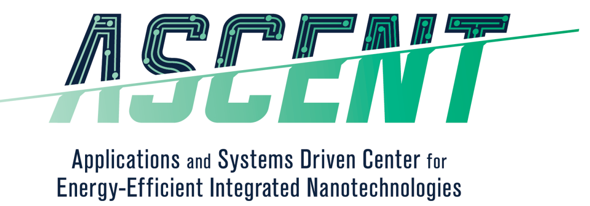
Beyond Logic Applications for Ferroelectric Field Effect Transistors

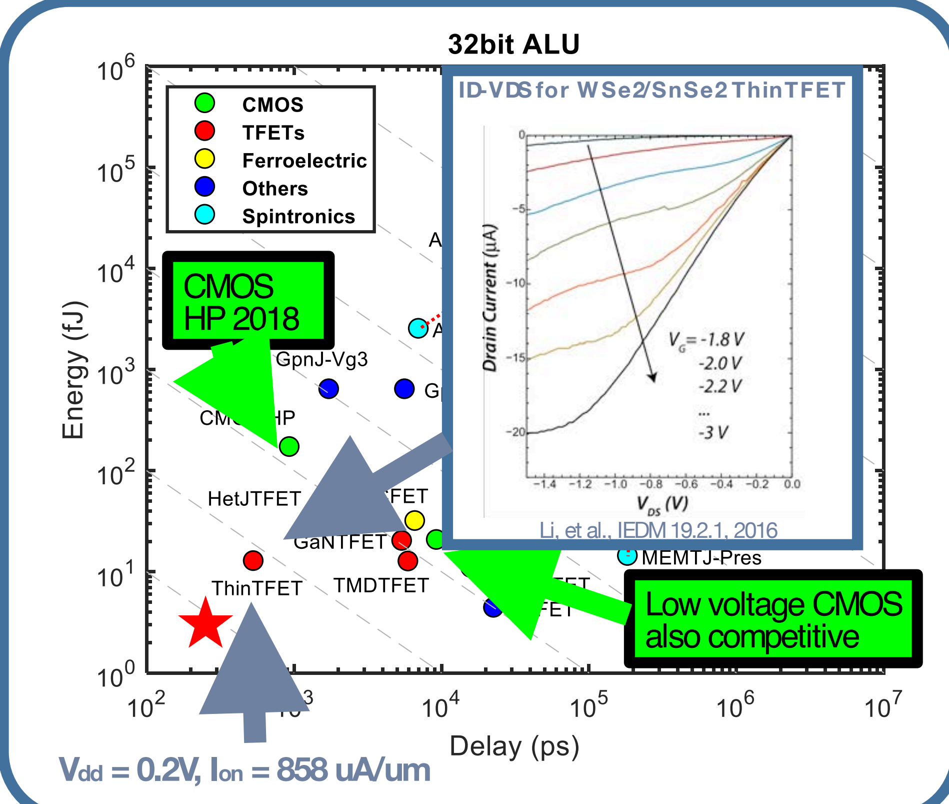
Michael Niemier

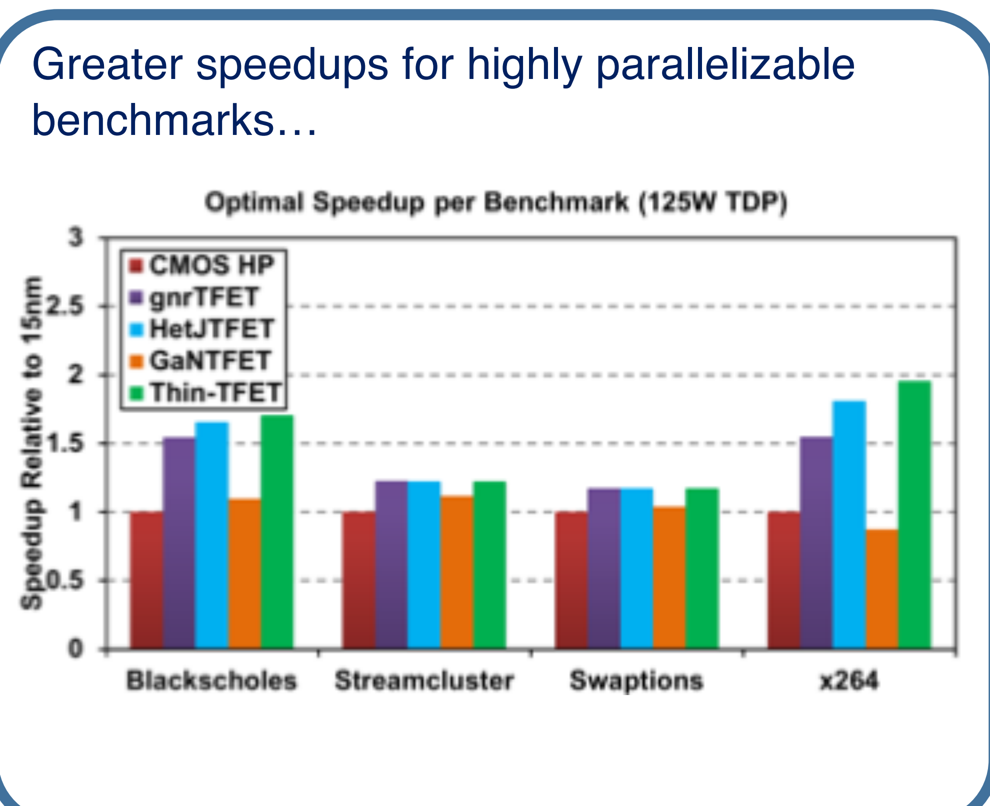
University of Notre Dame



This work was supported in part by ASCENT, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA

How does technology scaling impact m/c scaling?





Slightly better improvements at low TDP, but still only 2X to 3X…

Robert Perricone, X. Sharon Hu, Joseph Nahas, and Michael Niemier, "Can Beyond-CMOS Devices Illuminate Dark Silicon?" to appear in *Communications of the ACM*, 2018.

*"NRI research has explored a broad spectrum of beyond-CMOS devices for a 'new logic switch' to replace the current CMOS-based transistor ... a '***better switch***' has not been found. Comprehensive benchmarking of beyond-CMOS devices ... has revealed little or no advantage of these devices over CMOS for conventional Boolean logic and the von Neumann architecture."*

"some devices demonstrate unique characteristics suitable for novel architectures or computing paradigms, e.g., non-volatility in logic devices, reconfigurablity, [and/or] high computation density."

Semiconductor Research Corporation, "nanoelectronic COomputing REsearch (nCORE)," https://www.src.org/compete/ncore/ (2017)

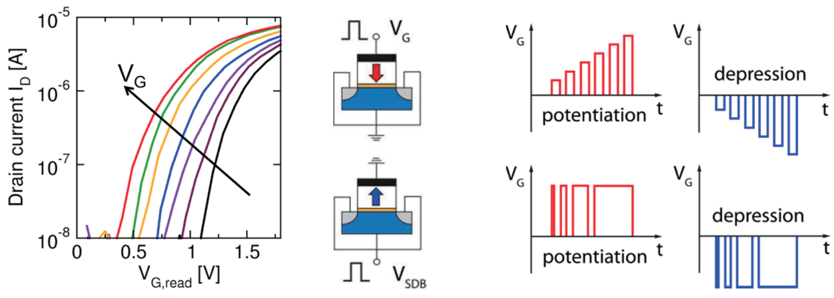
What do ferroelectric devices offer?

Steep subthreshold swings

Lee, et al., Prospects for Ferroelectric HfZrOx FETs with Experimentally CET=0.98nm, SSfor=42mV/dec,

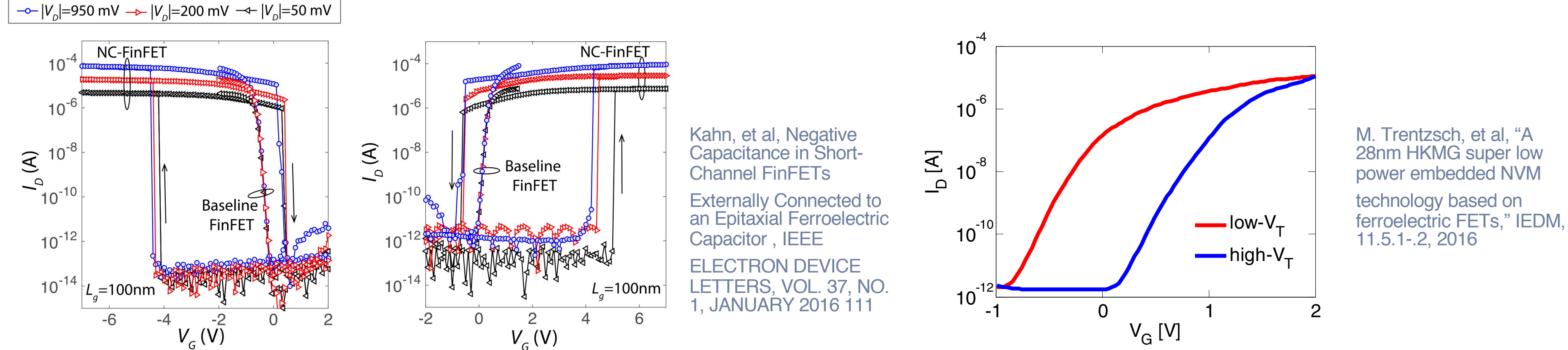
SSrev=28mV/dec, Switch-OFF <0.2V, and Hysteresis-Free Strategies, IEDM 2015.

Analog synaptic behavior



H. Mulaosmanovic, Novel ferroelectric FET based synapse for neuromorphic systems, VLSI Symposium, 2016.

Memory functionality



Kahn, et al, Negative Capacitance in Short-Channel FinFETs

Externally Connected to an Epitaxial Ferroelectric Capacitor , IEEE

ELECTRON DEVICE LETTERS, VOL. 37, NO. 1, JANUARY 2016 111

M. Trentzsch, et al, “A 28nm HKMG super low power embedded NVM

technology based on ferroelectric FETs,” IEDM, 11.5.1-.2, 2016

Devices with integrated ferroelectrics are well-positioned to address aforementioned space!

Talk outline

- FeFET device, models
- FeFETs for **logic-in-memory (LIM)**, **compute-in-memory (CIM)**
  - Emphasis on design/benchmarking of content addressable memories (LIM)
  - Briefly discuss FeFET-based CIM
- FeFETs for **neuromorphic applications**
  - FeFET-based analog synapse
  - FeFET-based (binary) convolutional neural networks (CNNs)
- Wrap-up

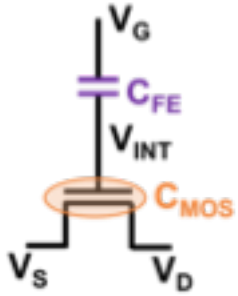
Background

FeFET device structure & operating modes

FeFET transistor structurally similar to bulk MOSFET or FinFET

- Ferroelectric (FE) layer integrated into gate stack

Ferroelectricity demoed in hafnium zirconium dioxide (highly compatible with CMOS)



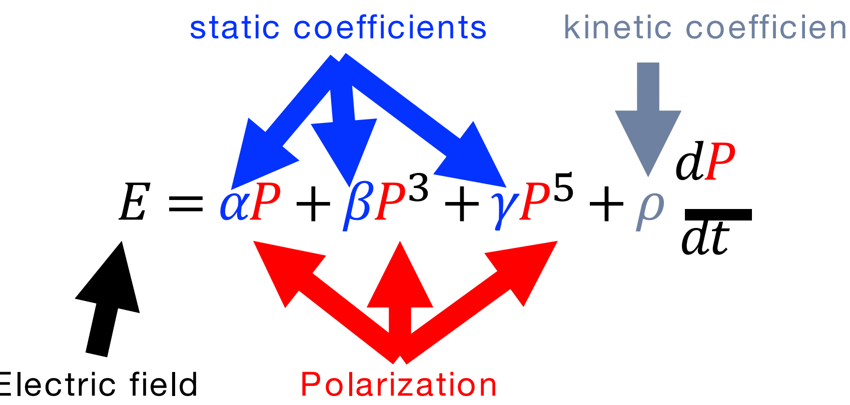
Interplay between FE material + underlying transistor capacitance results in different modes of operation:

- **Non-volatile mode** (device can maintain state)
- **Steep switching mode** (aimed at high performance)

Aziz, et al., “Computing with Ferroelectric FETs: Devices, Models, Systems, and Applications,” p. 1289-1298, DATE 2018.

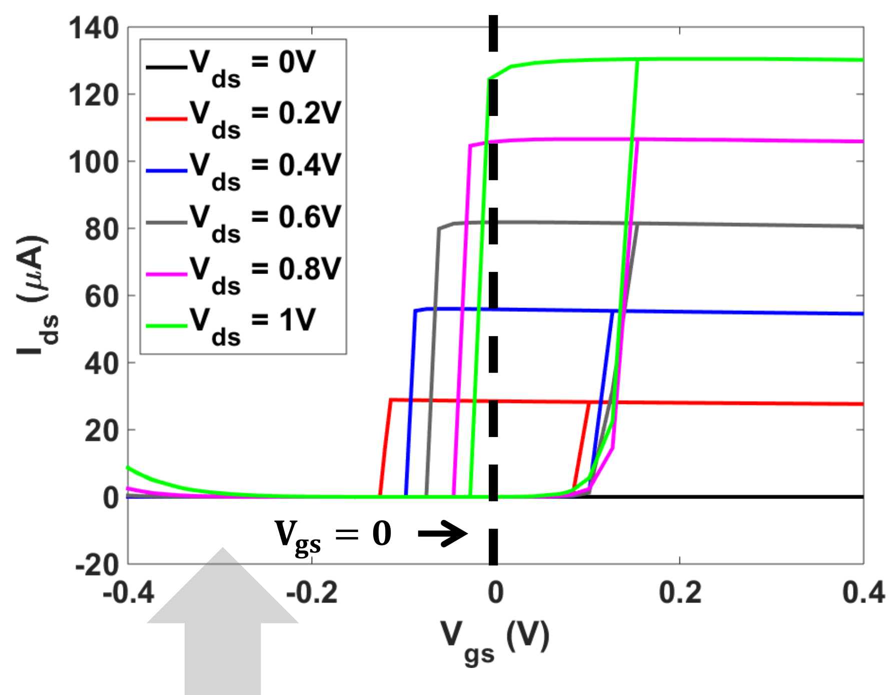
Time-dependent Landau Khalatnikov (LK) model

- LK model is SPICE compatible



α, β, γ calibrated to hafnium zirconium oxide (HZO)

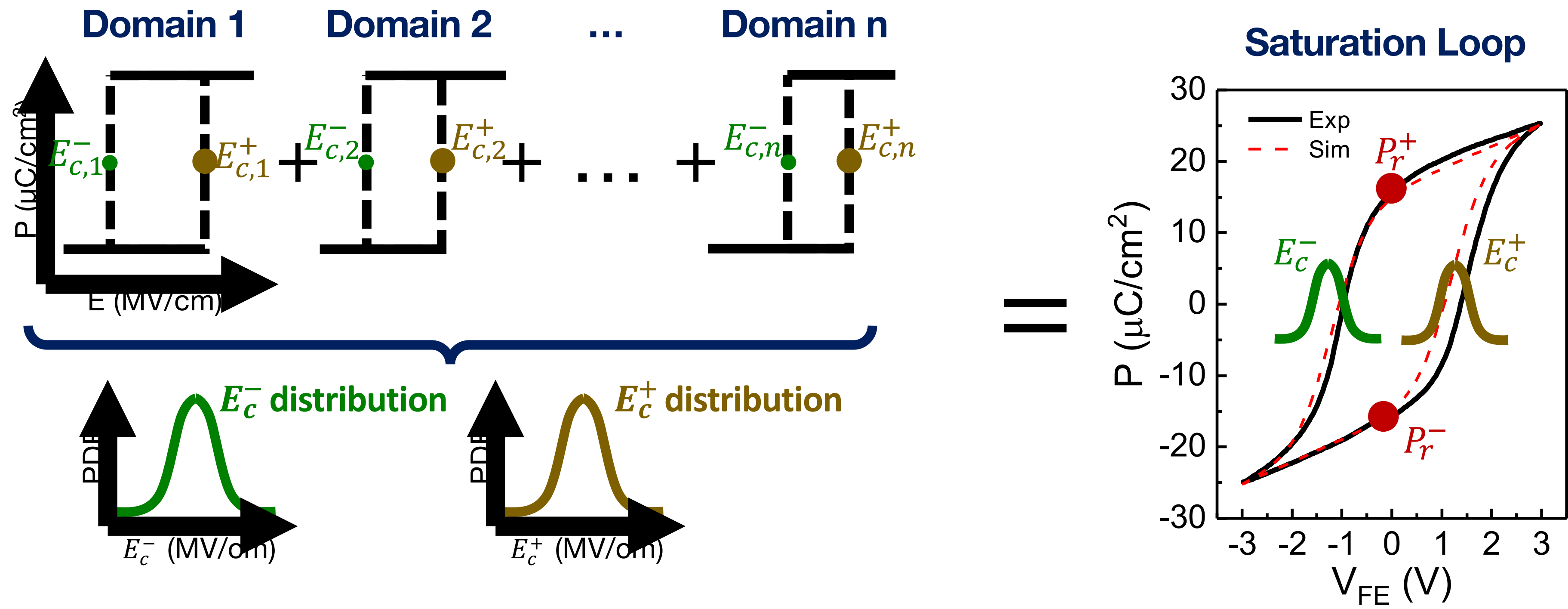
α	-7x10^9 m/F
β	3.3x10^{10} m^5/F/coul^2
γ	7x10^9 m/F
ρ	0.25
t_{FE}	5.7 nm



FeFET simulated by combining self-consistent LK equation with 45 nm PTM

A. Aziz, S. Ghosh, S. Datta, and S. K. Gupta, "Physics-based circuit-compatible SPICE model for ferroelectric transistors," IEEE EDL, vol. 37, no. 6, pp. 805–808, 2016

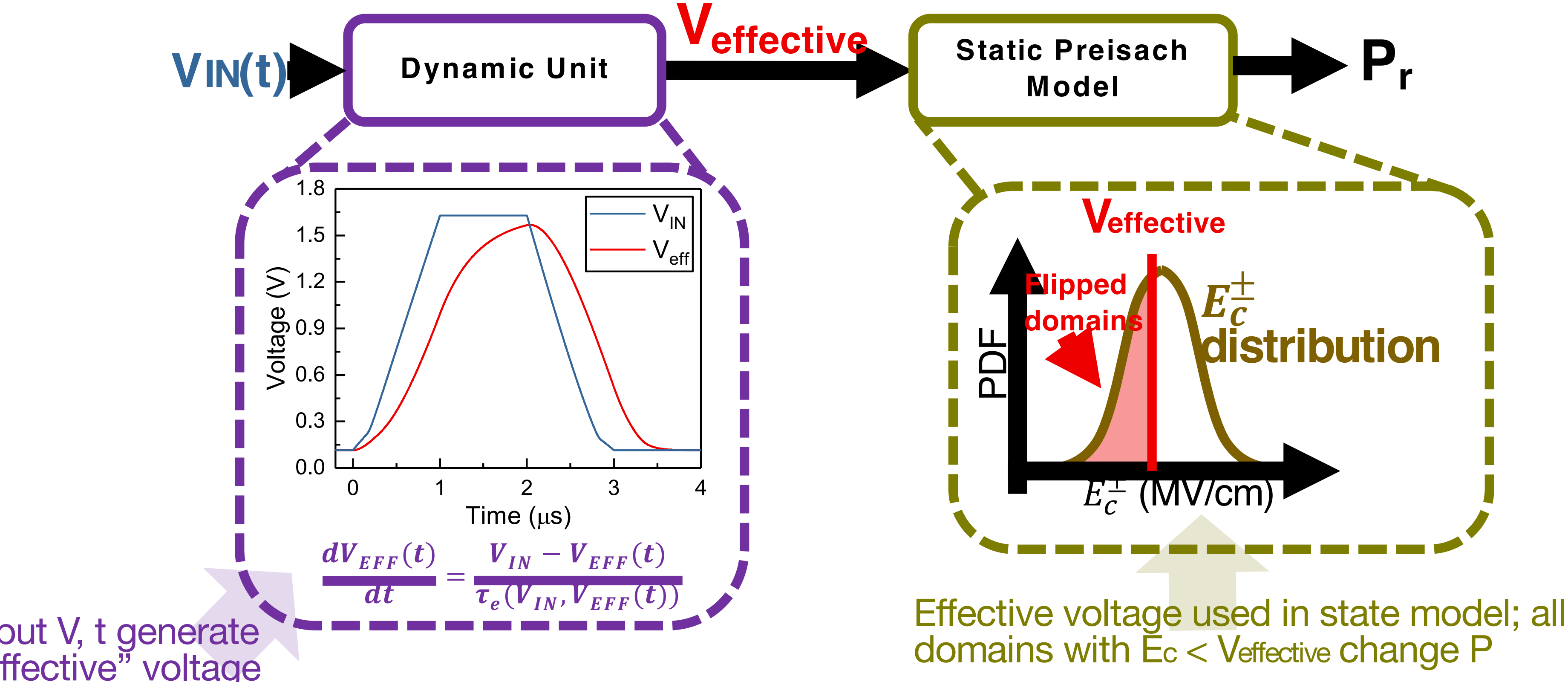
Multi-Domain Preisach Model



- The response of HZO film is described by the total contributions of many ideal ferroelectric domains of varying $E_c^{\pm}$.

K. Ni, M. Jerry, J.A. Smith, and S. Datta, "A Circuit Compatible Accuracy Compact Model for Ferroelectric FETs," in VLSI Symposium 2018.

Multi-Domain Preisach Model



Input V, t generate "effective" voltage used in standard DC model

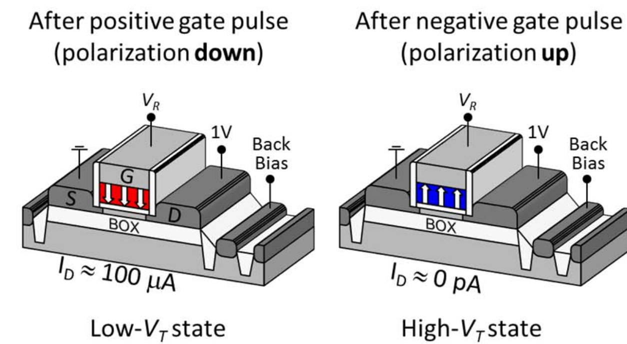
Effective voltage used in state model; all domains with Ec < Veffective change P

Calibration to the measured data allows model to accurately capture $P_r = f(V_{IN}, t)$

K. Ni, M. Jerry, J.A. Smith, and S. Datta, "A Circuit Compatible Accuracy Compact Model for Ferroelectric FETs," in VLSI Symposium 2018.

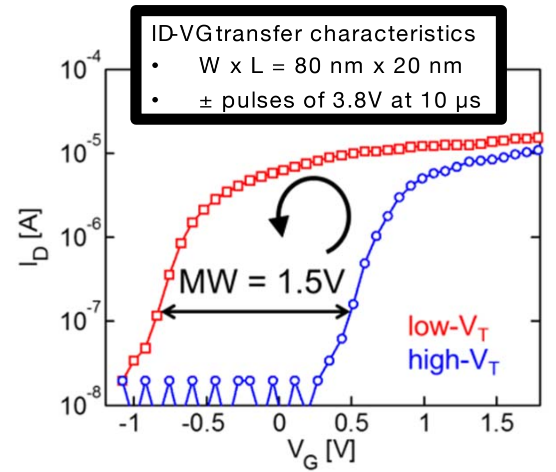
Operation

- Can switch FeFET polarization with:
  - Positive gate voltage pulse (program)
  - Negative gate voltage pulse (erase)



- Pulse causes stable, reversible V_t shift
  - Low V_t, high V_t depends on dipole's orientation

- 2 distinguishable states = memory window
  - Sense with readout of drain current



May tradeoff pulse duration, amplitude depending on application-level figures of merit

Dünkel, et al., "A FeFET based super-low-power ultra fast embedded NVM technology for 22 nm FDSOI and beyond," IEDM, 19.7.1, 2017.

Logic-in-memory & Compute-in-memory

Logic-in-memory: CAMs

Content Addressable Memory (CAM)

- Fast HW search O(1) for search intensive apps
- Often use **ternary CAM** (TCAM) – i.e., store 1, 0, or X (where X is “don’t care” (DC))
- TCAMs applicable to database apps, neural networks, *routers and switches*, etc.

Address lookup with TCAM, RAM

TCAM address					
00	1	0	1	X	X
01	0	1	1	0	X
10	0	1	1	X	X
11	1	0	0	1	1

01
Search result

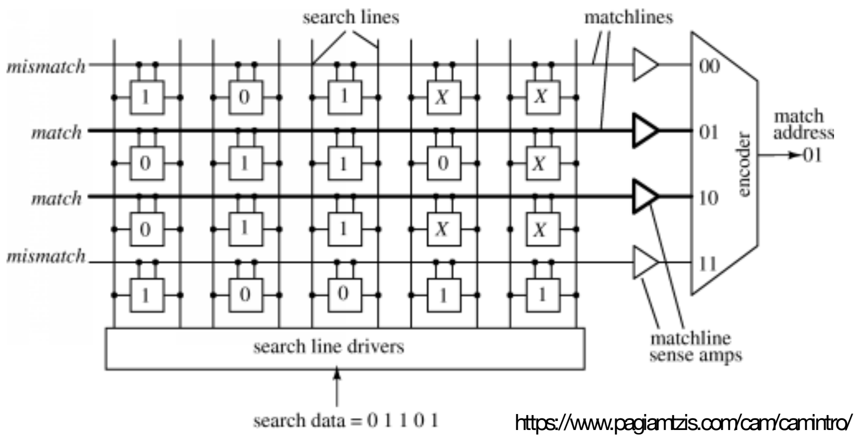
RAM address	
00	Port = A
01	Port = B
10	Port = C
11	Port = D

Output Port=B

Search data: 0 1 1 0 0 (e.g., 192.168.1.1)

Choose 01, instead of 10 as 01 more complete

TCAM array architecture

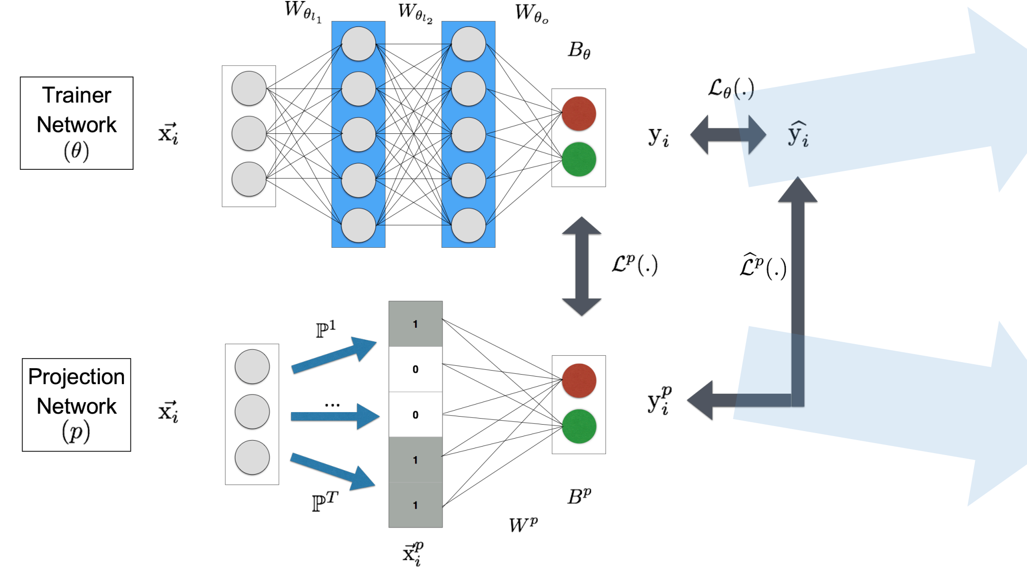


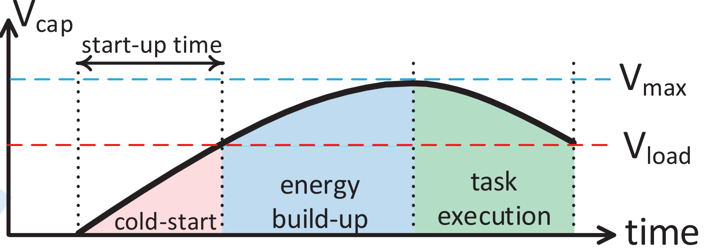
https://www.pagiamtzis.com/cam/camintro/

Emerging neuromorphic computing models

e.g., **Projection Networks**

Train neural network, lightweight network in lockstep





Goal: more accurate, powerful machine learning models in resource constrained environments

The choice of the type of projection matrix $\mathbb{P}$ as well as representation of the projected space $\Omega_{\mathbb{P}}$ in our setup has a direct effect on the computation cost and model size. We propose to leverage an efficient randomized projection method using a modified version of ocality sensitive hashing (LSH to define $\mathbb{P}(.)$.

TCAM-supported hashing again an important compute kernel

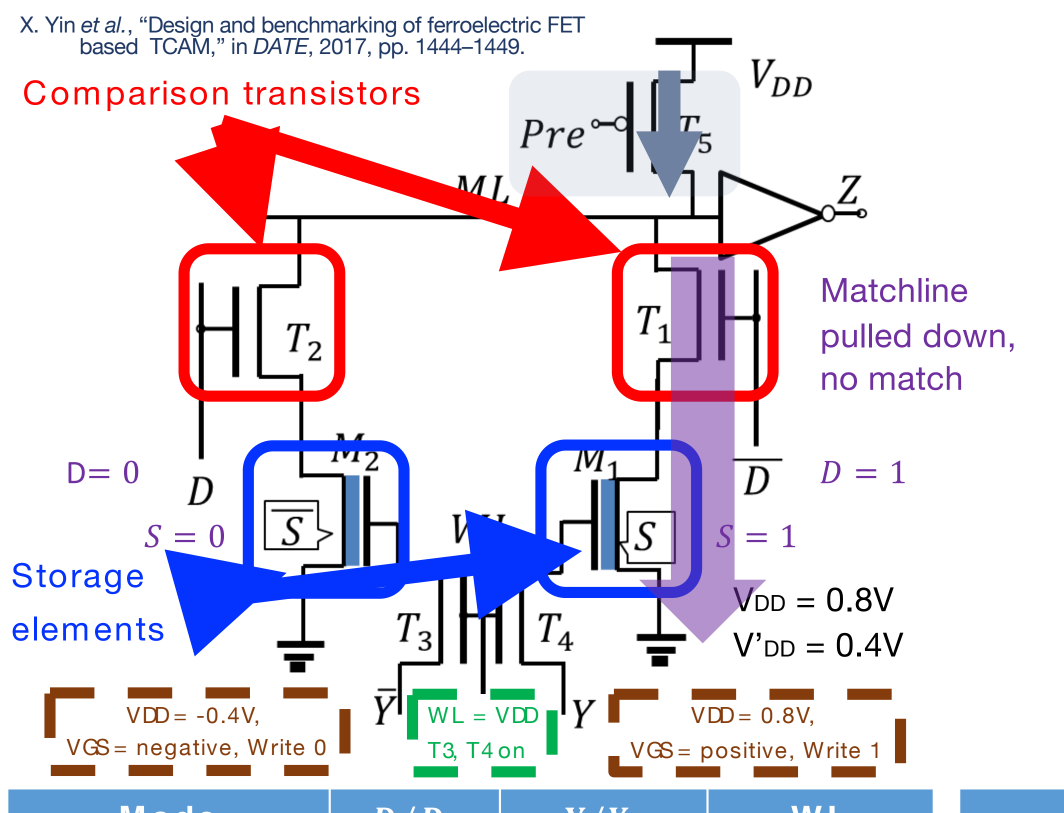
ProjectionNet: Learning Efficient On-Device Deep Networks Using Neural Projections

Sujith Ravi
Google Research, Mountain View, CA, USA
sravi@google.com

Look at TCAMs based on ASCENT technologies to support these models, other applications – consider FeFETs here…

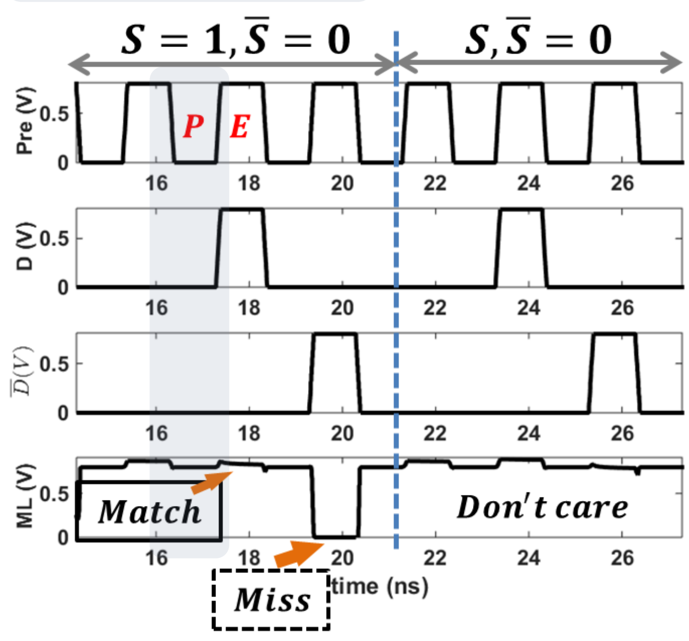
4T, 2FeFET TCAMs (w/negative supply, LK model)

X. Yin *et al.*, "Design and benchmarking of ferroelectric FET based TCAM," in *DATE*, 2017, pp. 1444–1449.



1st, pre-charge matchline

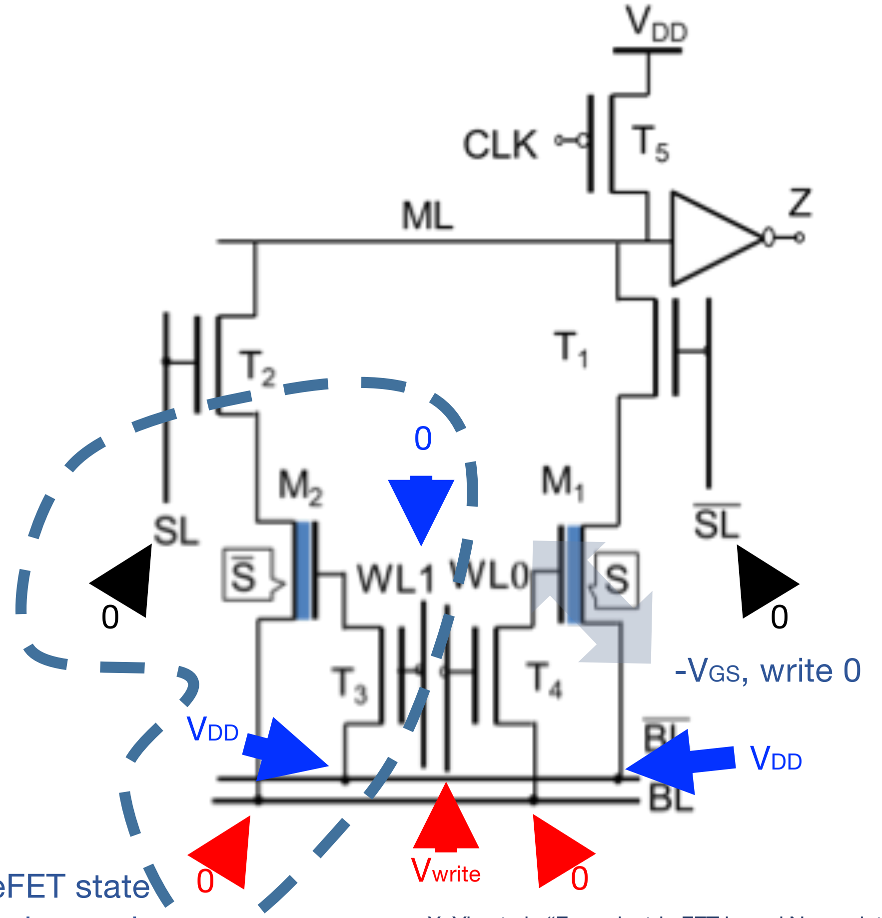
P = Pre-charge phase, **E** = Evaluate phase



Mode		$D/\overline{D}$	$Y/\overline{Y}$	WL
Write	S = 1	0	VDD / -V'DD	VDD
	S = 0	0	-V'DD / VDD	VDD

Mode		$D/\overline{D}$	$Y/\overline{Y}$	WL
Search	D = 1	VDD / 0	0	0
	D = 0	0 / VDD	0	0

4T, 2FeFET TCAMs (w/o negative supply, LK model)

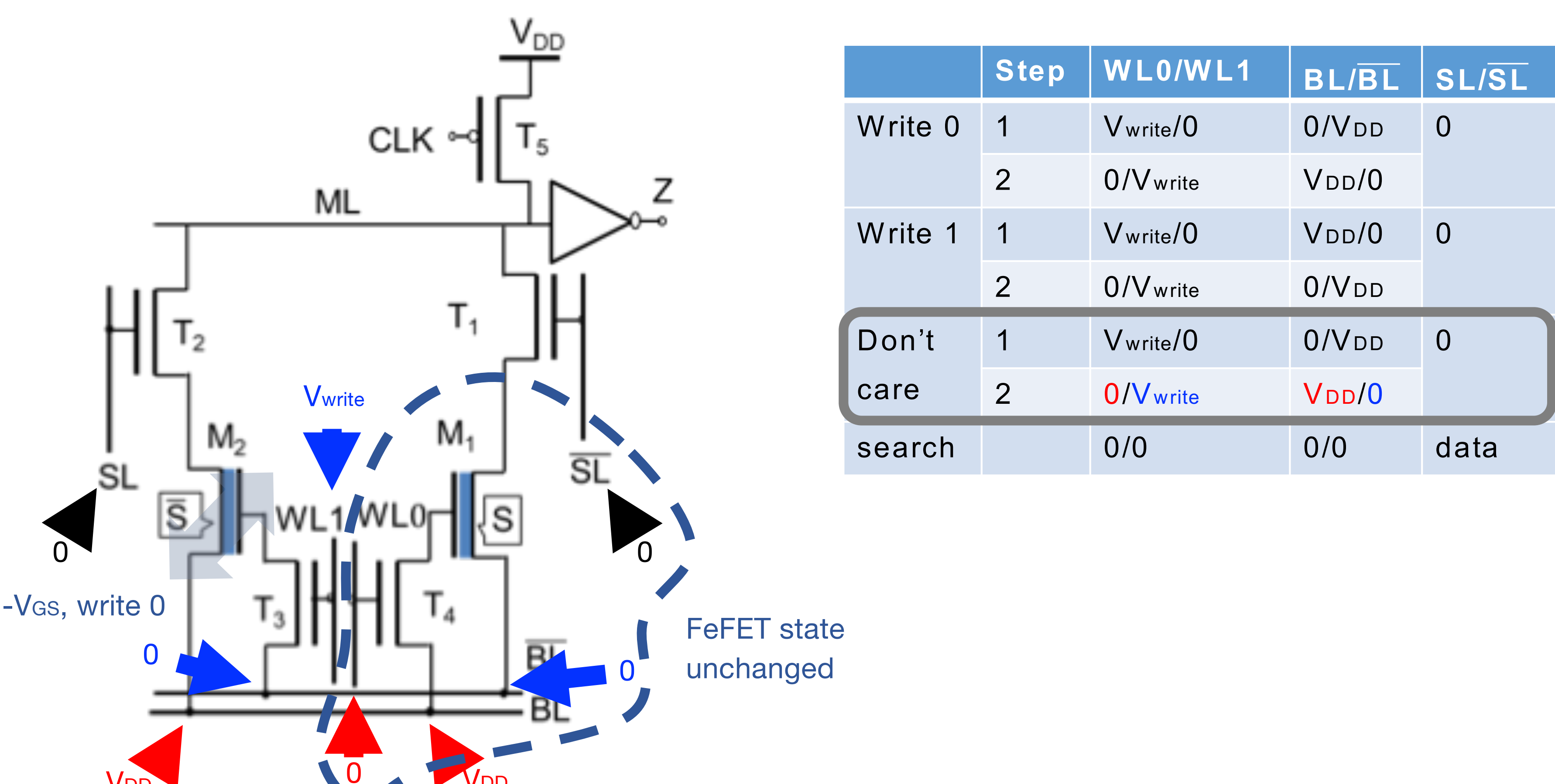


Prior work considered TCAM based on LK model with *negative* supply

	Step	WL0/WL1	BL/$\overline{BL}$	SL/$\overline{SL}$
Write 0	1	V_{write}/0	0/V_{DD}	0
	2	0/V_{write}	V_{DD}/0	
Write 1	1	V_{write}/0	V_{DD}/0	0
	2	0/V_{write}	0/V_{DD}	
Don't care	1	V_{write}/0	0/V_{DD}	0
	2	0/V_{write}	V_{DD}/0	
search		0/0	0/0	data

X. Yin et al., "Ferroelectric FET based Non-volatile Logic-in-Memory Circuits," IEEE TVLSI (in submission), 2018.

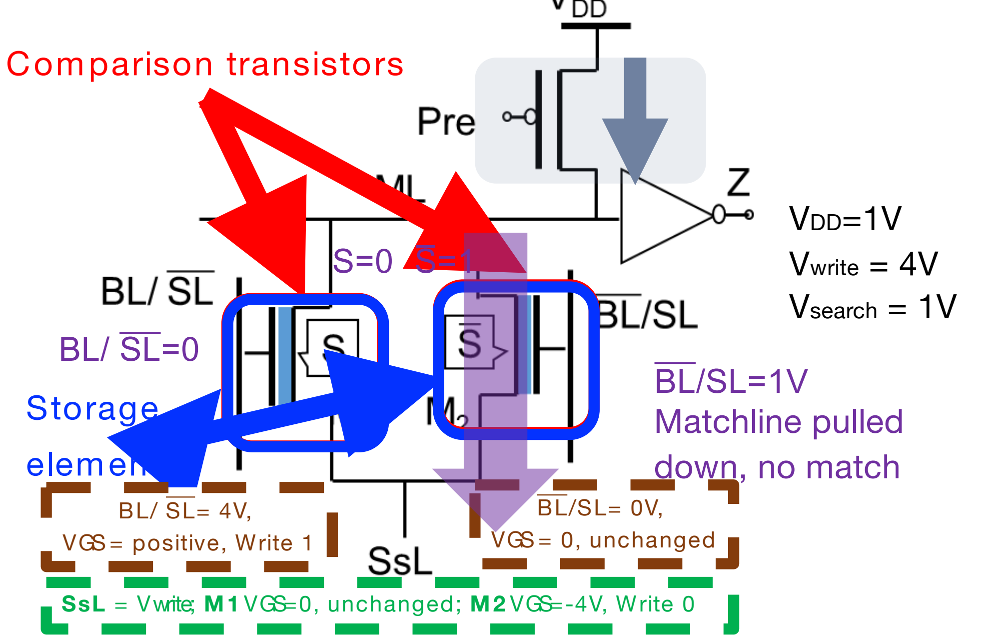
4T, 2FeFET TCAMs (w/o negative supply, LK model)

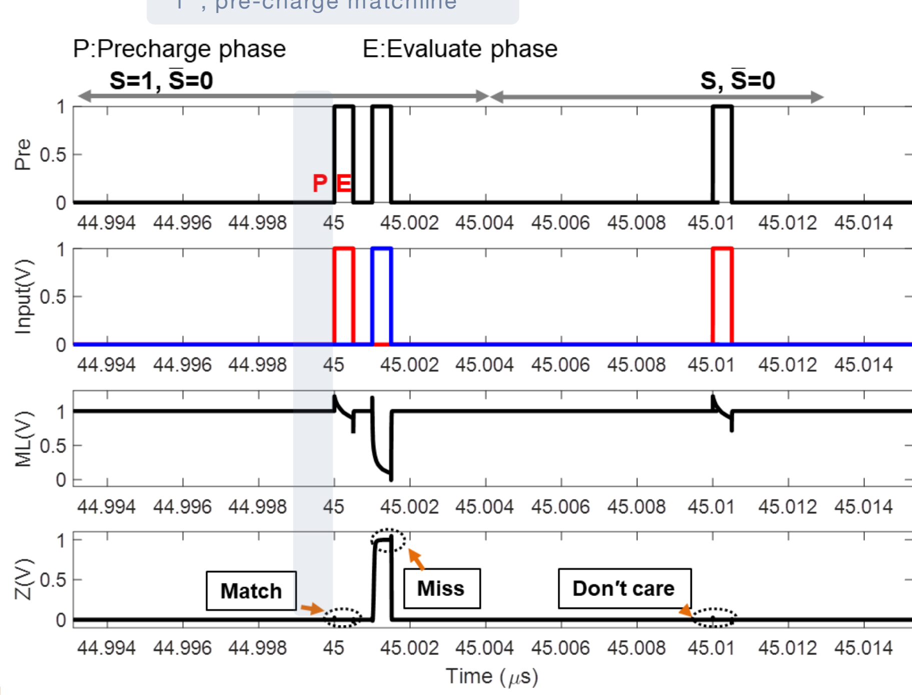


	Step	WL0/WL1	BL/$\overline{BL}$	SL/$\overline{SL}$
Write 0	1	V_{write}/0	0/V_{DD}	0
	2	0/V_{write}	V_{DD}/0	
Write 1	1	V_{write}/0	V_{DD}/0	0
	2	0/V_{write}	0/V_{DD}	
Don't care	1	V_{write}/0	0/V_{DD}	0
	2	0/V_{write}	V_{DD}/0	
search		0/0	0/0	data

X. Yin et al., "Ferroelectric FET based Non-volatile Logic-in-Memory Circuits," IEEE TVLSI (in submission), 2018.

2T FeFET design (Preisach model)



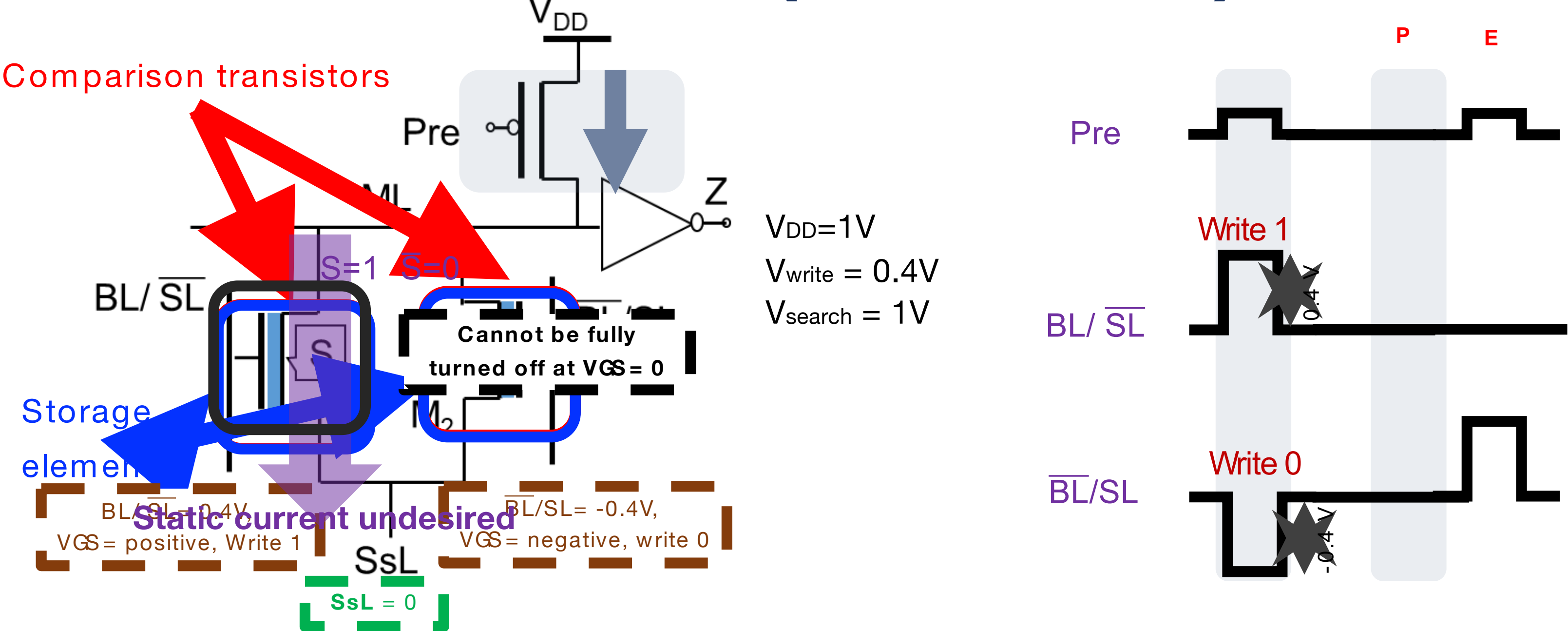


Mode			BL/ $\overline{SL}$	$\overline{BL}$/SL	SsL
Write	S = 1	Step 1	V_{write}	0	0
		Step 2	V_{write}	0	V_{write}
	S = 0	Step 1	0	V_{write}	V_{write}
		Step 2	0	V_{write}	0

Mode		BL/ $\overline{SL}$	$\overline{BL}$/SL	SsL
Search	1	0	V_{search}	0
	0	V_{search}	0	0

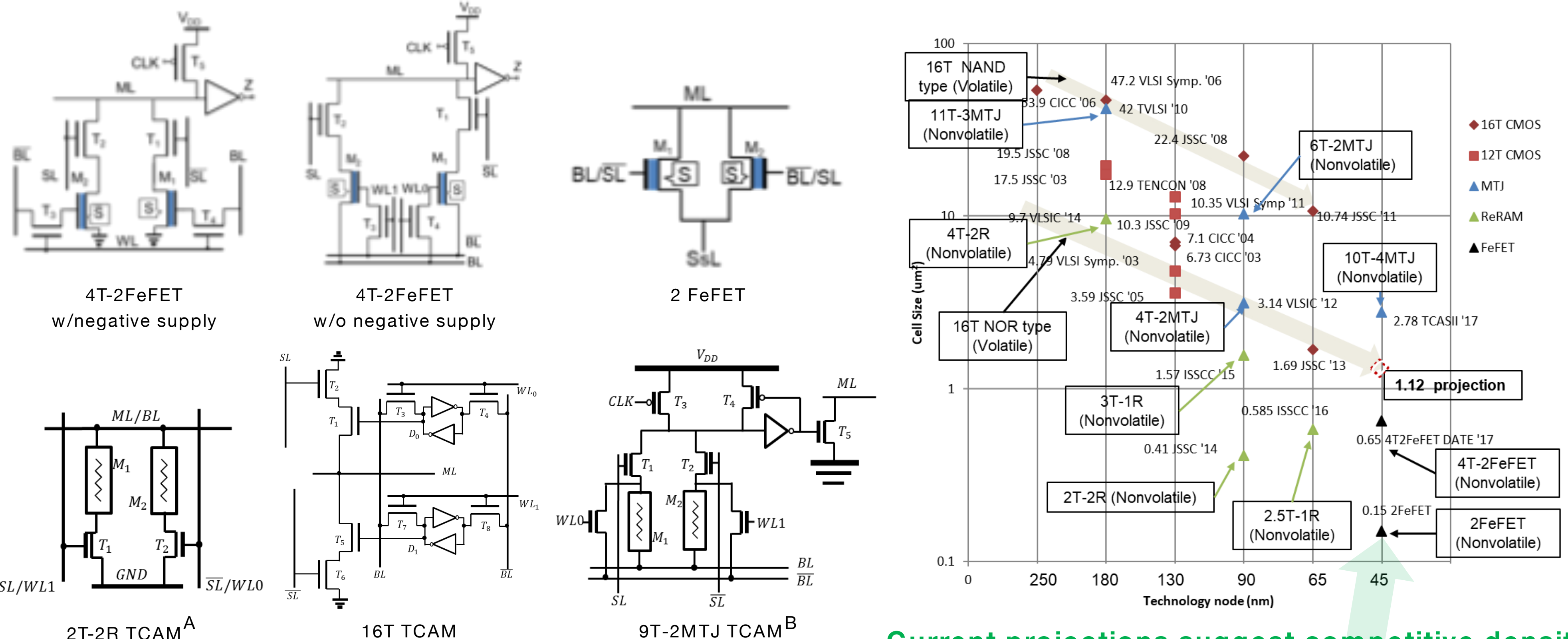
X. Yin et al., "Design and Benchmarking of 2-Ferroelectric FET TCAM," IEEE TCAS (in submission), 2018.

2T FeFET design (LK model)



X. Yin et al., "Design and Benchmarking of 2-Ferroelectric FET TCAM," IEEE TCAS (in submission), 2018.

Benchmarking (area comparisons)



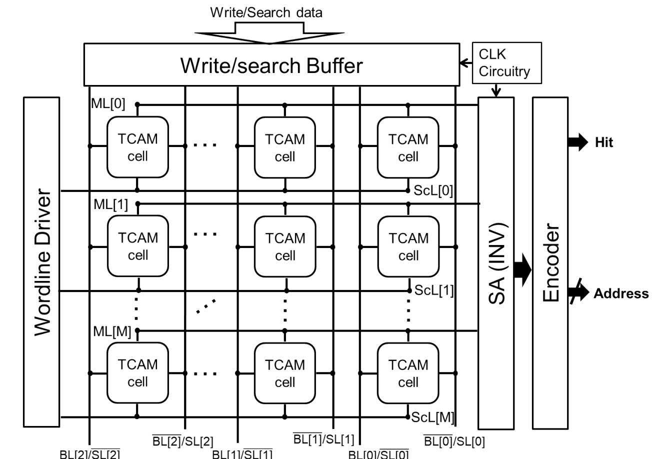
Current projections suggest competitive density

[A] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1 mb 0.41 µm2 2t-2r cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 896– 907, 2014.

[B] S. Matsunaga, A. Katsumata, M. Natsui, T. Endoh, H. Ohno, and T. Hanyu, "Design of a nine-transistor/two-magnetic-tunnel-junction- cell-based low-energy nonvolatile ternary content-addressable memory," *Japanese J. of Applied Physics*, vol. 51, no. 2S, p. 02BM06, 2012

Benchmarking methodology

- All designs evaluated in context of 64x64 array
- Assume
  - 45 nm PTM
  - Inverter-based SA
  - Minimum sized transistors for TCAM cell, SA
- Extract wiring parasitics from DESTINY
  - M. Poremba, et al., "Destiny: A tool for modeling emerging 3D NVM and EDRAM caches," in *DATE*, 2015, pp. 1543–6.
- Delay assumes worst case
  - (i.e., 1-bit mismatch…)



Benchmarking (other figures of merit)

64X64 size	16T CMOS	2T2R	4T-2FeFET	2FeFET
Cell structure				
Technology node	45nm	45nm	45nm	45nm
Cell area (μm^2)	1.12 (7.5x)	0.41[1] (2.7x)	0.65 (4.3x)	0.15 (1x)
ON/OFF ratio	~10^6	~100[2]	~10^4	~10^4
Search voltage	1V	1V	1V	1V
Search delay (ps)	582 (1.7x)	350 (1.03x)	1013 (3.0x)	341 (1x)
Search energy (fJ/bit/search)	1.0 (2.4x)	1.2 (2.7x)	0.5 (1.3x)	0.4 (1x)
Normalized EDP	4.1x	2.8x	3.8x	1x
Write scheme	Voltage driven dynamic switching	Current driven	Voltage driven	Voltage driven
Write voltage	1V	Set 1.8V[3] Reset 1.2V[3]	±4V	±4V
Write time	< 2ns	~ 10 ns	10 ns	10 ns
Write energy (fJ/row)	309 (3.5x)	288000 (3225x)[3]	512 (5.7x)	89 (1x)

1. Li, Jing, et al. "1Mb 0.41 µm 2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing." IEEE VLSIC, 2013.
2. Lastras-Montano, Miguel Angel, et al. "Architecting energy efficient crossbar-based memristive random-access memories." IEEE NANOARCH 2015
3. Li, Shuangchen, et al. "Nvsim-cam: a circuit-level simulator for emerging nonvolatile memory based content-addressable memory." ICCAD, 2016.

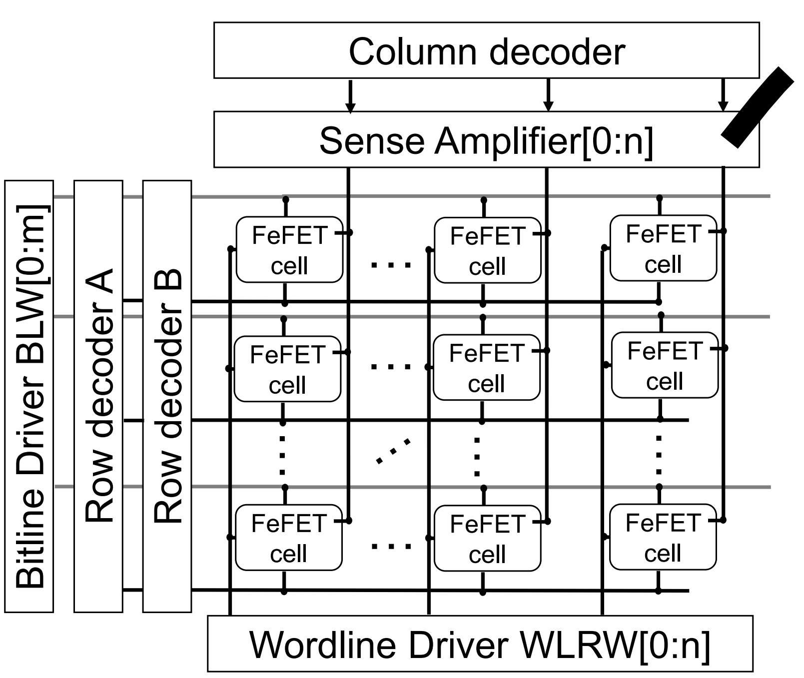
Benchmarking (other figures of merit)

64X64 size	16T CMOS	2T2R	4T-2FeFET	2FeFET
Cell structure				
Technology node				
Cell area (µm²)				
ON/OFF ratio				
Search voltage				
Search delay (ps)				
Search energy (fJ/bit/search)				
Normalized EDP				
Write scheme				
Write voltage				
Write time				
Write energy (fJ/row)				

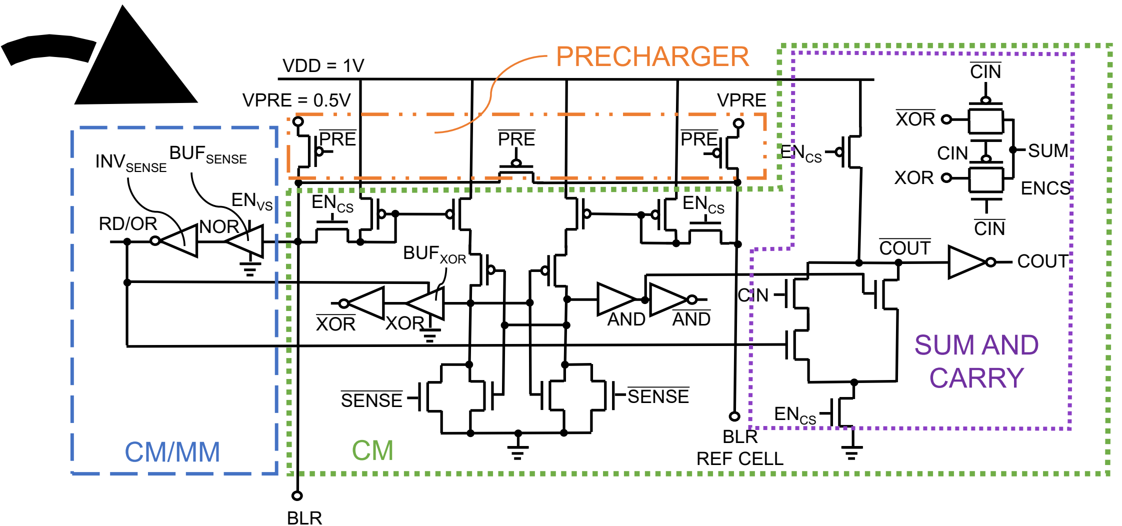
1. Li, Jing, et al. "1Mb 0.41 µm 2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing." IEEE VLSIC, 2013.
2. Lastras-Montano, Miguel Angel, et al. "Architecting energy efficient crossbar-based memristive random-access memories." IEEE NANOARCH 2015
3. Li, Shuangchen, et al. "Nvsim-cam: a circuit-level simulator for emerging nonvolatile memory based content-addressable memory." ICCAD, 2016.

FeFET-based CIM: architecture

FeFET-based CIM architecture



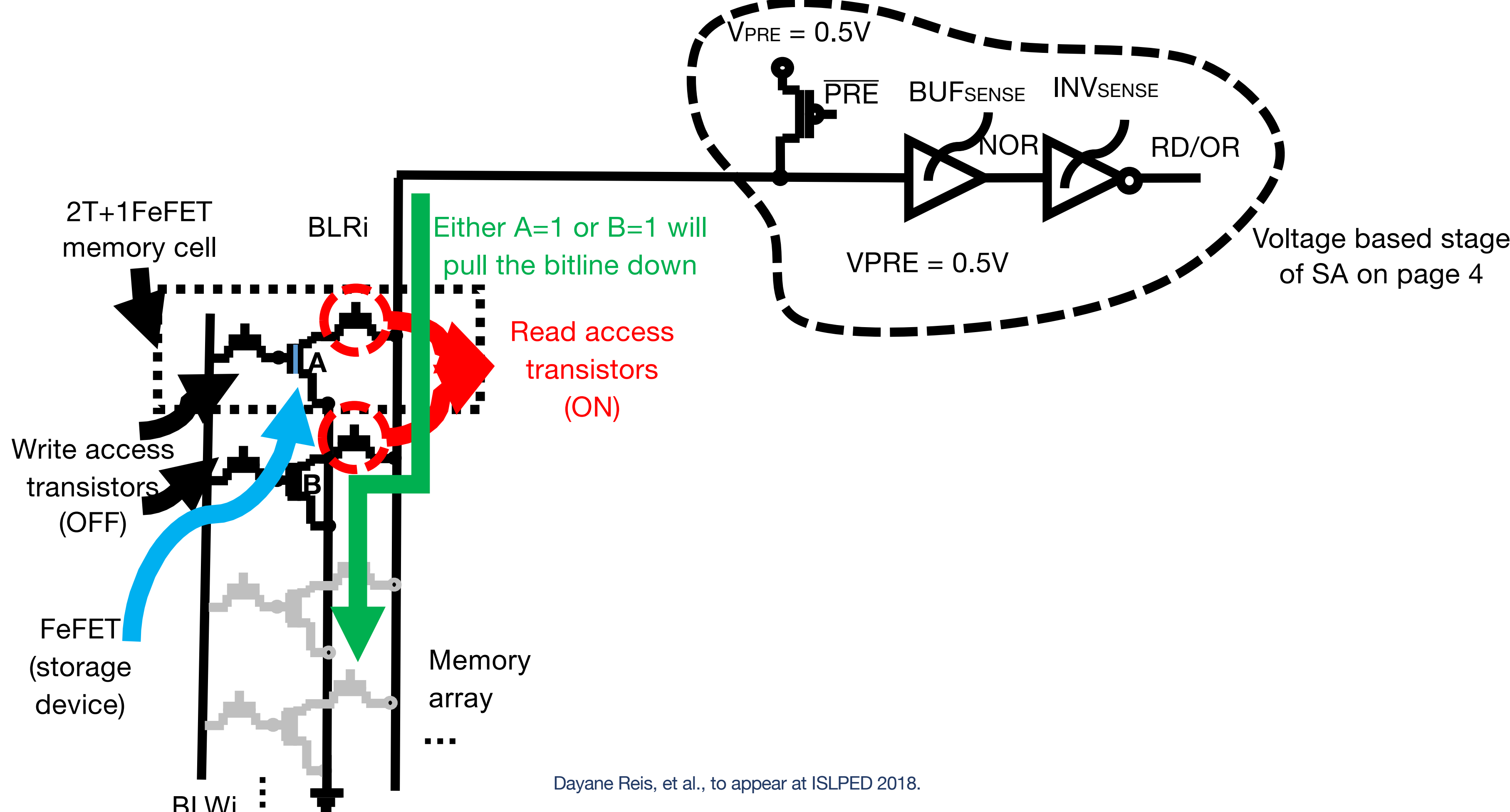
FeFET-CIM customized sense amp (SA)



- **CM/MM** is voltage-based sense scheme responsible for (N)OR logic and reads
- **CM** is current-based sense scheme used for Boolean (N)AND, X(N)OR, and ADD; also leverages voltage scheme
- **SUM** and **CARRY** is additional circuitry for carry and sum

Dayane Reis, et al., to appear at ISLPED 2018.

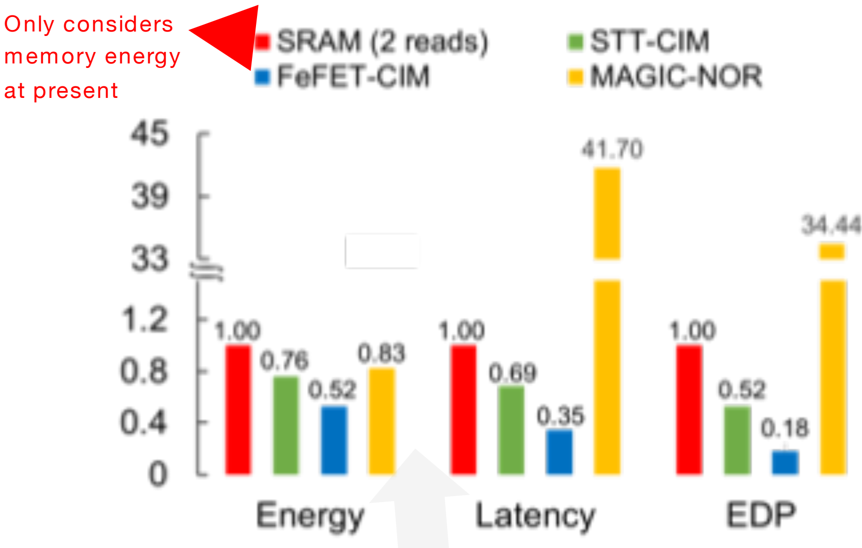
FeFET-based CIM: OR operations



Dayane Reis, et al., to appear at ISLPED 2018.

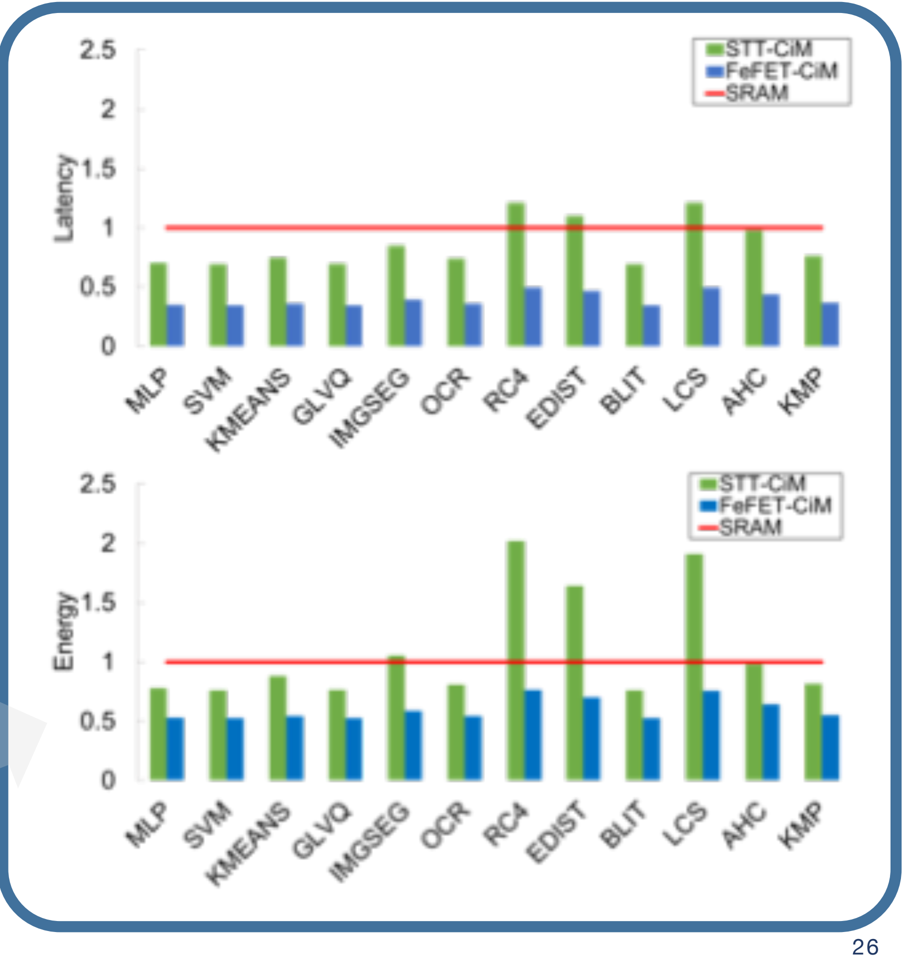
FeFET-based CIM: benchmarking

Only considers memory energy at present



FeFET-CIM has speed-ups (energy reductions) of ~119X (~1.6X) and ~1.97X (~1.5X) over ReRAM and STT-RAM CIM for in-memory addition of 32-bit words

FeFET-CIM approach offers an average speedup of ~2.5X and energy reduction of ~1.7X when compared to a conventional (not in-memory) approach.



STT-CIM: Computing in Memory with Spin-Transfer Torque Magnetic RAM

Shubham Jain, Ashish Ranjan, Kaushik Roy, Anand Raghunathan, School of Electrical and Computer Engineering, Purdue University, {jain130,aranjan,kaushik,raghunathan}@purdue.edu

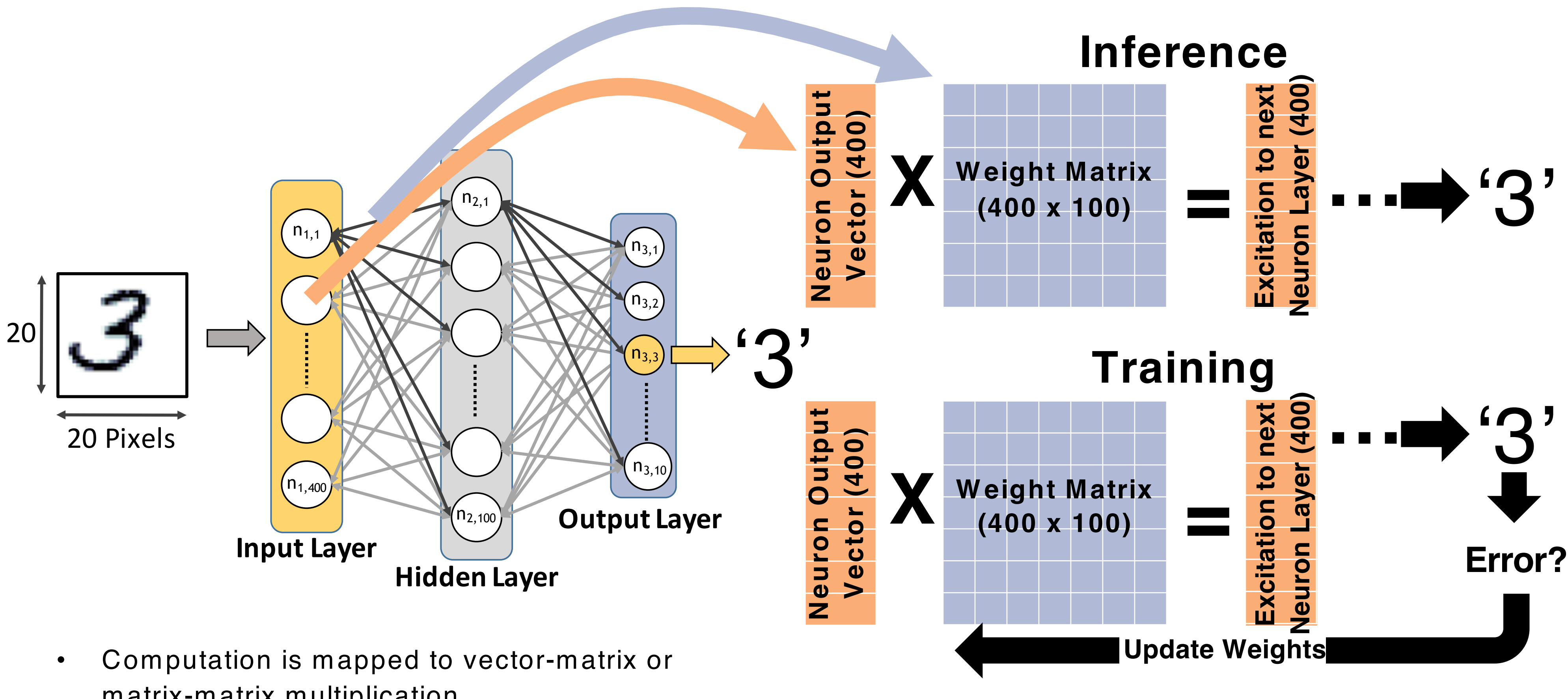
MAGIC-NOR: IEEE TRANSACTIONS ON NANOTECHNOLOGY, VOL. 15, NO. 4, JULY 2016 — Logic Design Within Memristive Memories Using Memristor-Aided loGIC (MAGIC)

Nishil Talati, Saransh Gupta, Pravin Mane, and Shahar Kvatinsky, *Member, IEEE*

Dayane Reis, et al., to appear at ISLPED 2018.

Neuromorphic applications

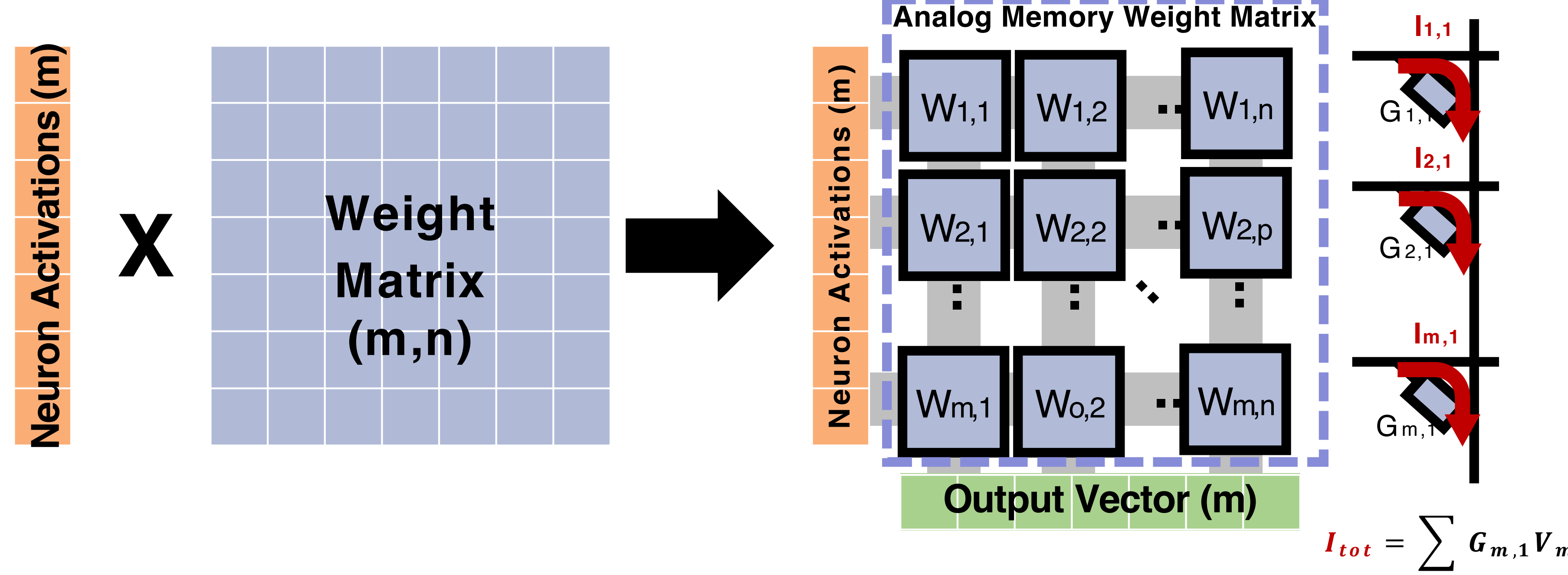
Inference & training



- Computation is mapped to vector-matrix or matrix-matrix multiplication

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

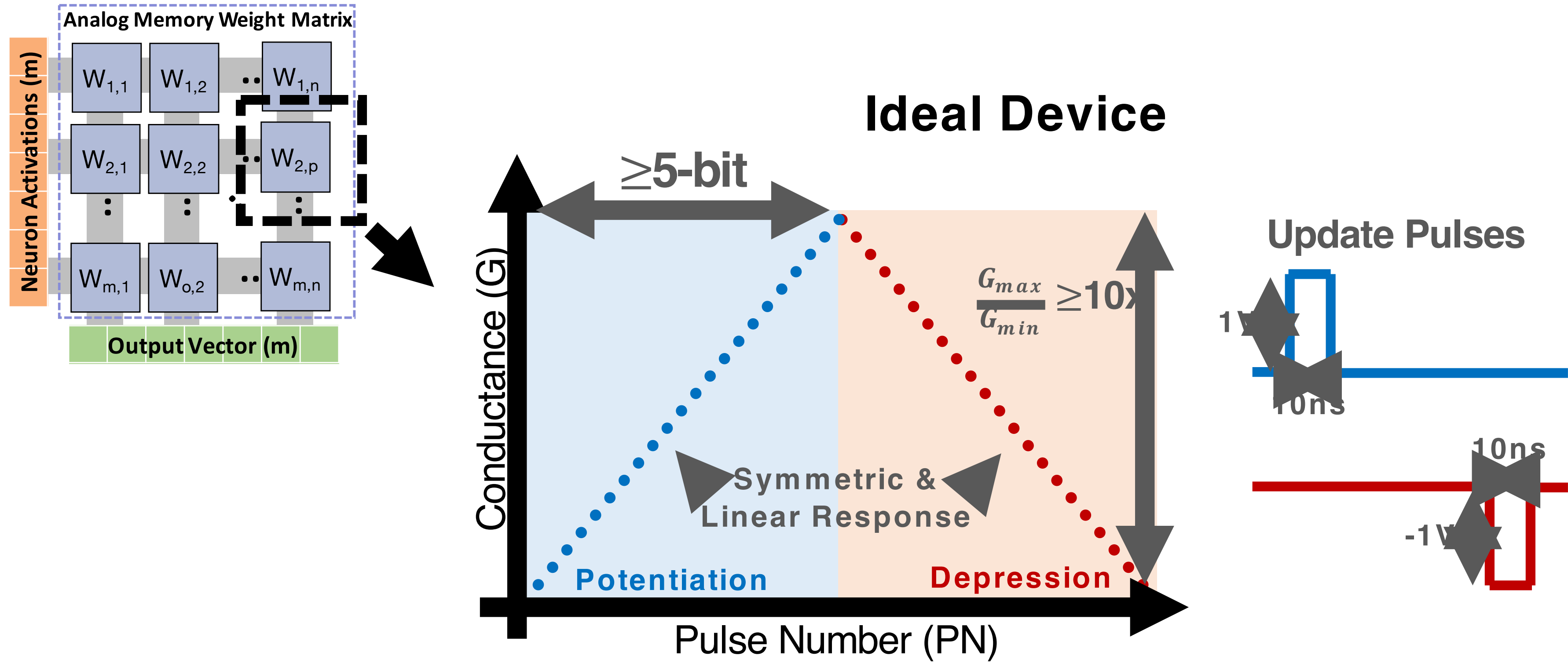
Vector-matrix multiplication with crossbars



- Dense analog synaptic memory arrays perform MACs and update at the location of the data

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

Analog synaptic device characteristics

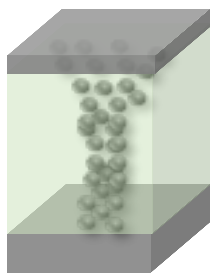


- Synaptic memory needs to be high density, low latency, energy efficient, and preserve high network accuracies.

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

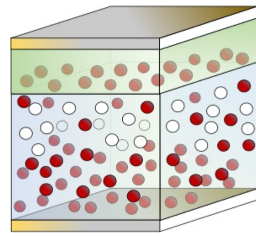
Analog synapses

Filamentary O-RRAM



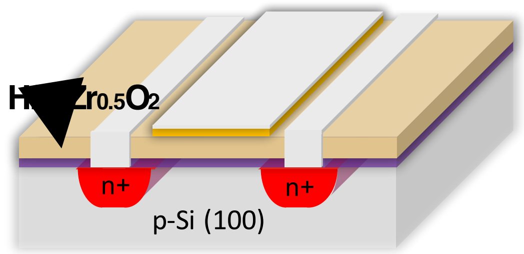
- ✓ **High density**
- Electro-thermal switching
- ✓ **<100ns pulse widths**
- Low G_{max}/G_{min} ratios demonstrated thus far

Non-Filamentary RRAM



- ✓ **High density**
- Electro-thermal switching
- <100ns pulse widths to be demonstrated
- Low G_{max}/G_{min} ratios demonstrated thus far

Ferroelectric FET

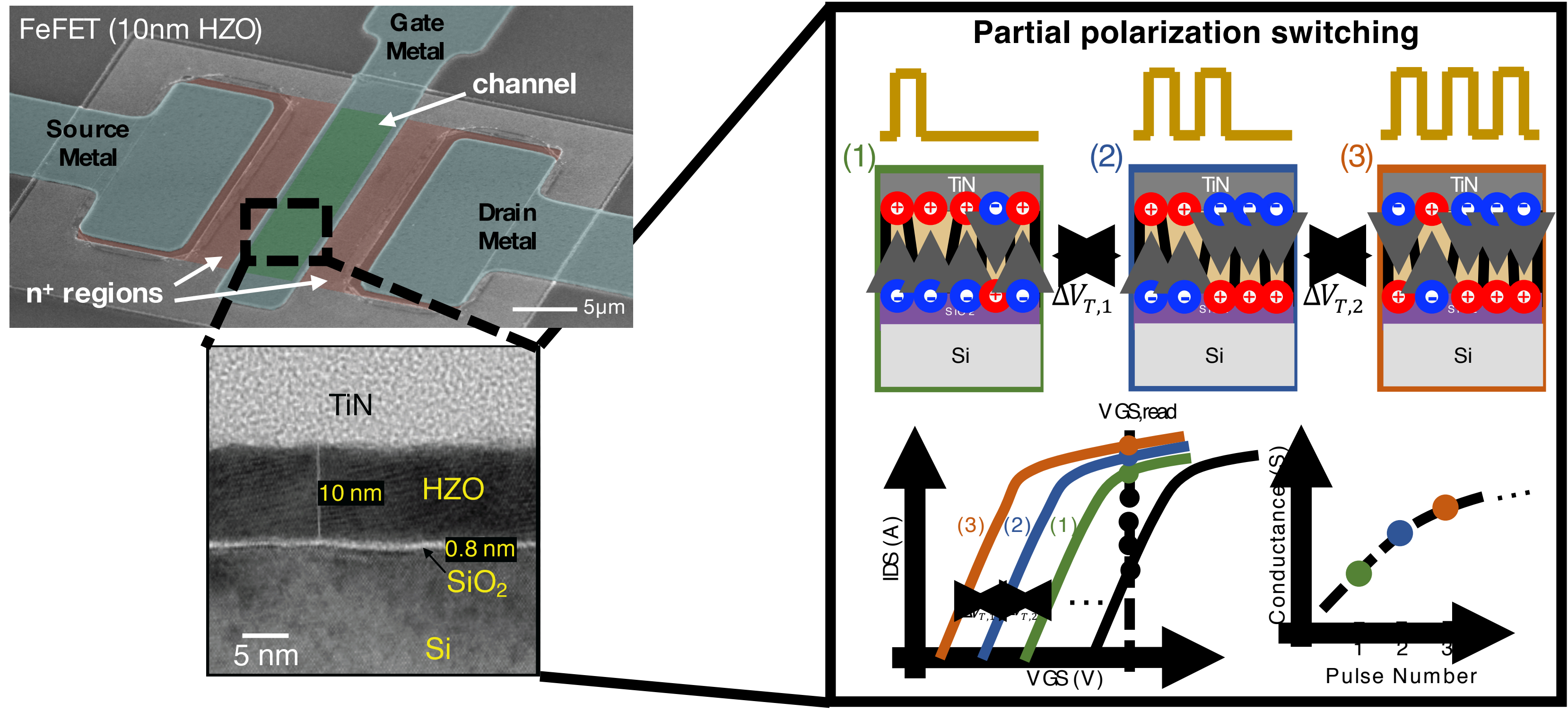


- 2T design proposed
- ✓ **Electric-field switching**
- ✓ **75ns pulse widths**
- ✓ **Large and tunable G_{max}/G_{min}**

- Ferroelectric FET is a promising candidate for an analog synaptic memory device.

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

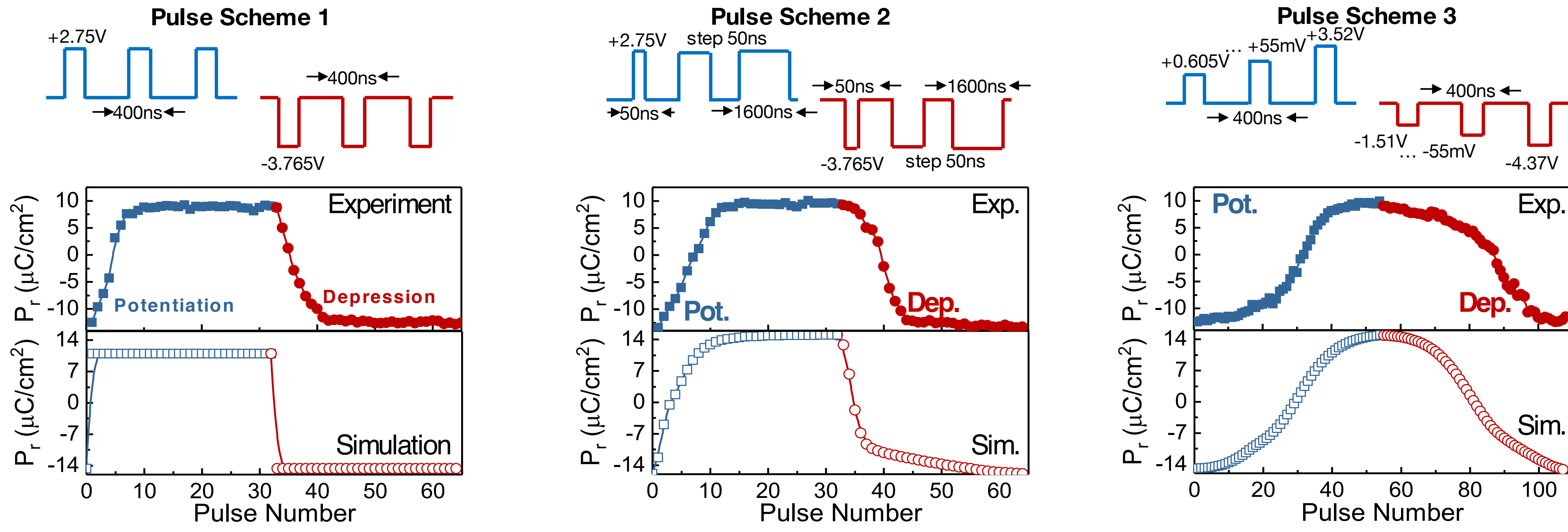
Ferroelectric FET analog synapse



- Electric-field controlled partial polarization switching in ferroelectrics FETs can be harnessed for synaptic memory with nanosecond updates.

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

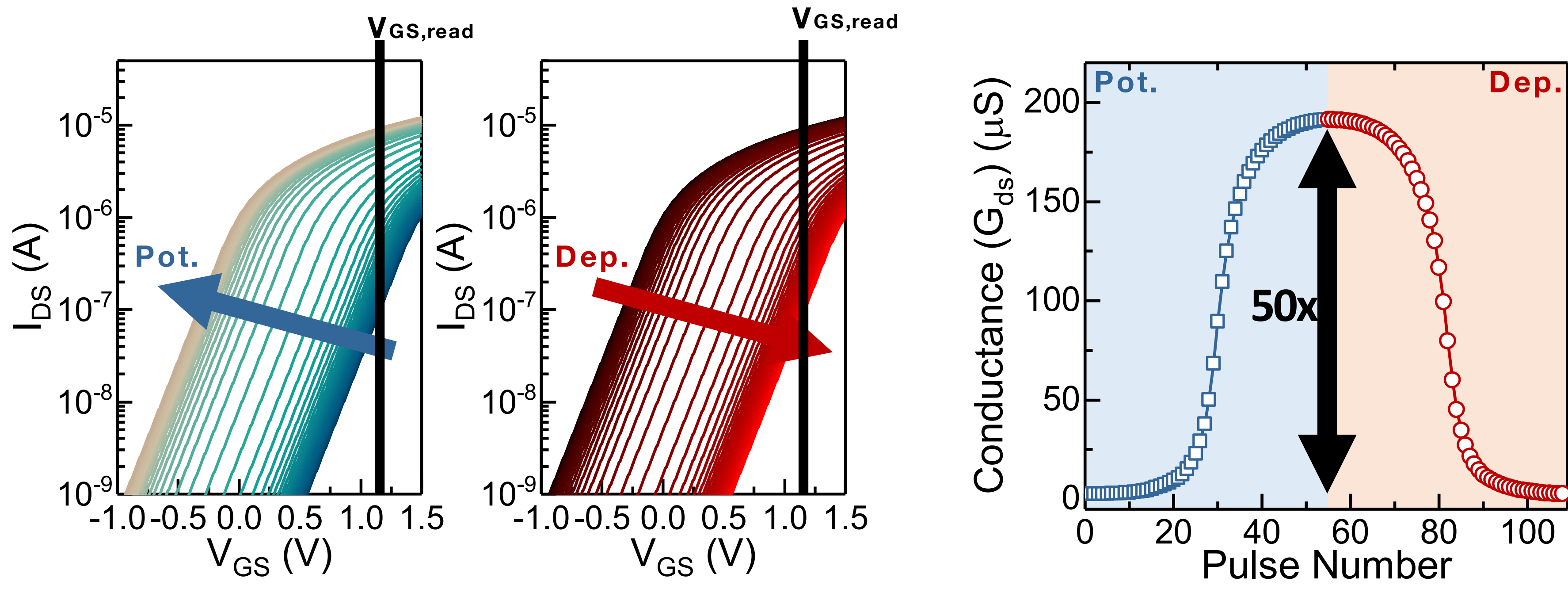
Effect of Pulse Scheme on P_r



- Multi-domain Preisach model accurately captures the the response of the remnant polarization

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

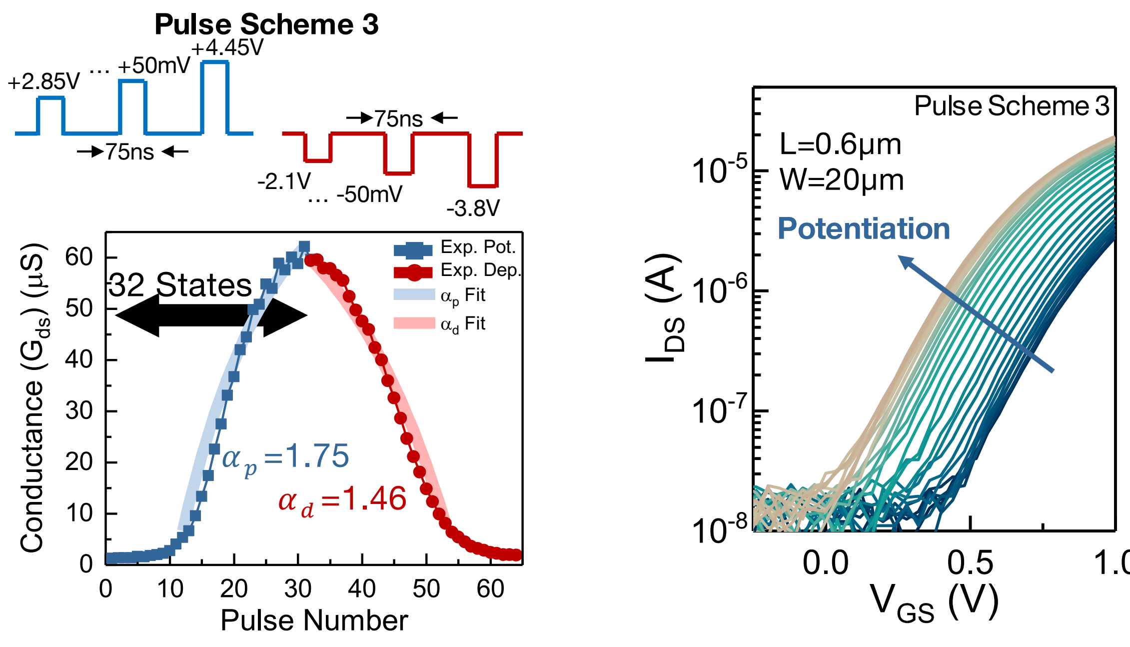
Simulated G vs Pulse Number: Scheme 3



- FeFET synapse response from simulated P_r in programming scheme 3.

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

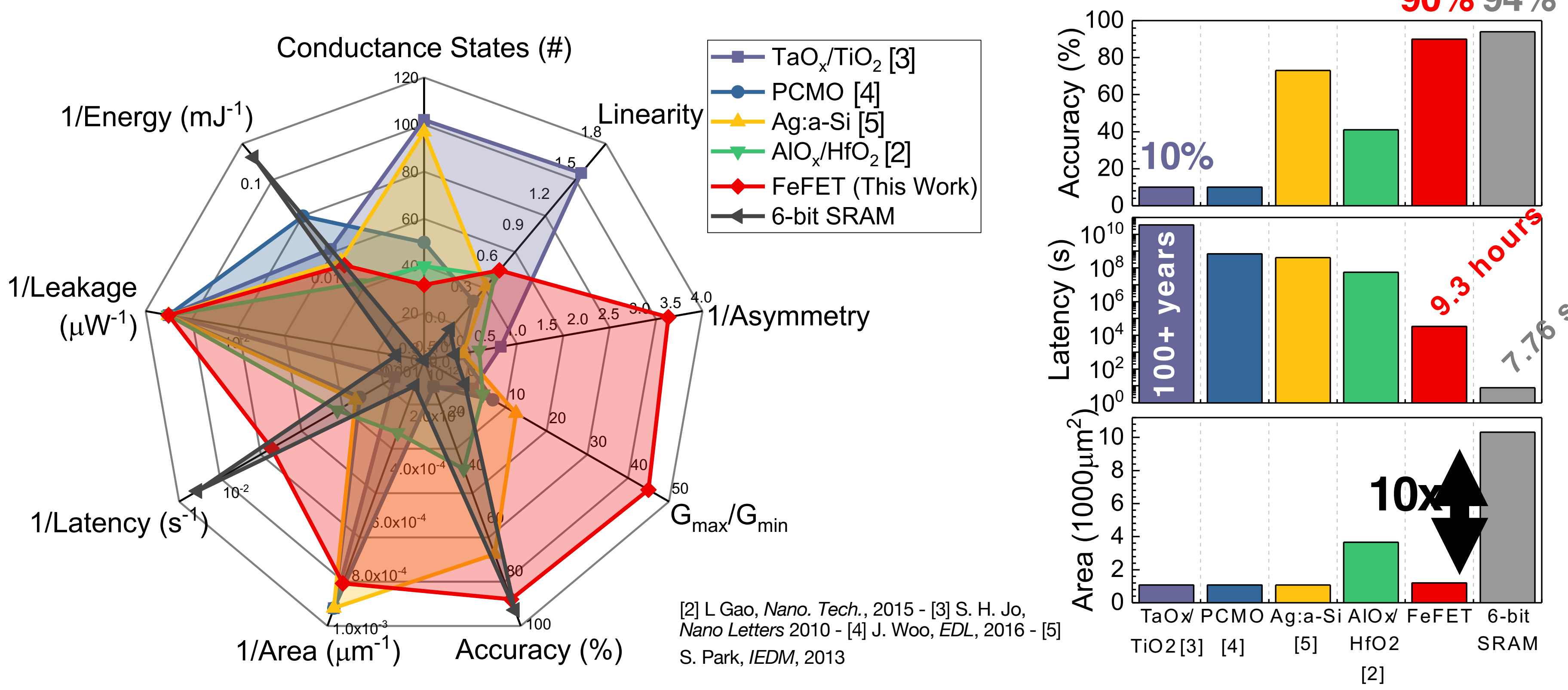
FeFET Analog Synapse: Scheme 3



- Partial polarization switching within the ferroelectric gate oxide results in a gradual decrease/increase (potentiation/depression) in V_T.

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

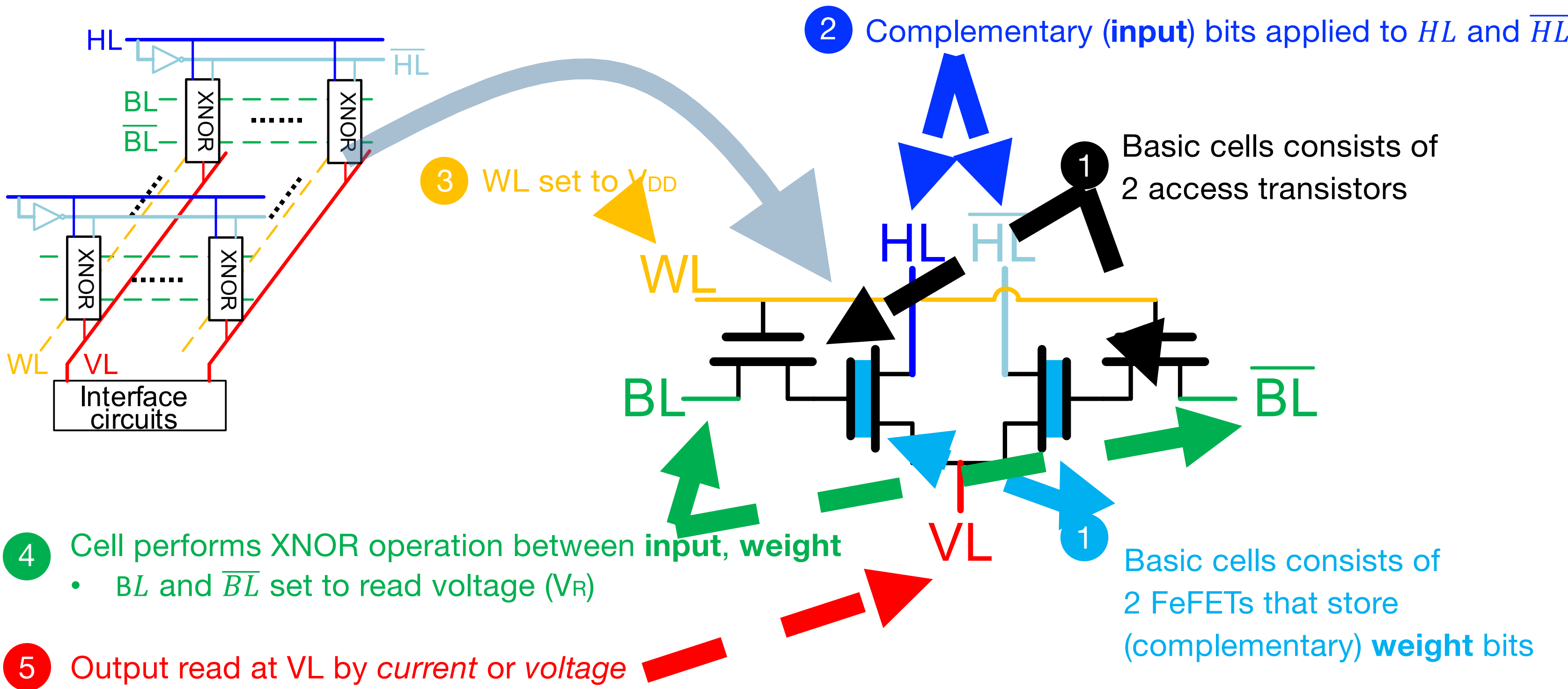
Analog Synapse Benchmarking



- FeFET under pulse scheme 3 exhibits the reduced footprint, high accuracy, and low latency

Jerry, Matthew, et al. "Ferroelectric FET analog synapse for acceleration of deep neural network training." in IEDM, p. 6-2, 2017.

FeFET-based binary crossbars: circuits



Xioaming Chen, Xunzhao Yin, Michael Niemier, and Xiaobo Sharon Hu, "Design and Optimization of FeFET-based Crossbars for Binary Convolutional Neural Networks," to appear in *Design, Automation, and Test in Europe (DATE)*, 2018.

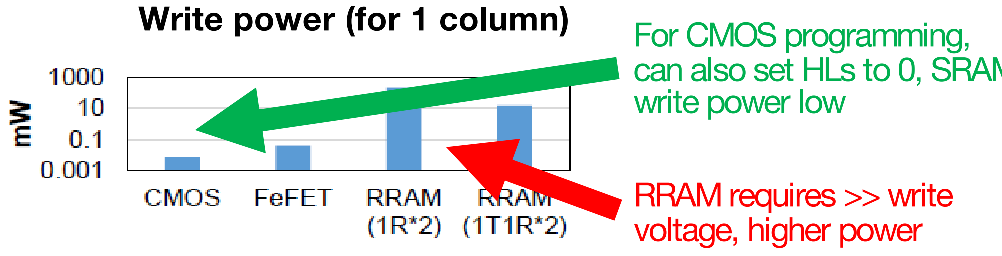
FeFET-based crossbars: benchmarking

Benchmarking assumptions for 64x64 crossbar array

- **FeFET:** 10nm FinFET, T_{fe}=10.5nm, V_{WL}=0.6, V_W=0.6, V_R=-0.55
- **RRAM:** R_{on}=10KΩ, Roff=1MΩ, V_W=2
- **For both:** V_{HL}=0.3
- **Average case:** half input bits and half weights are 1

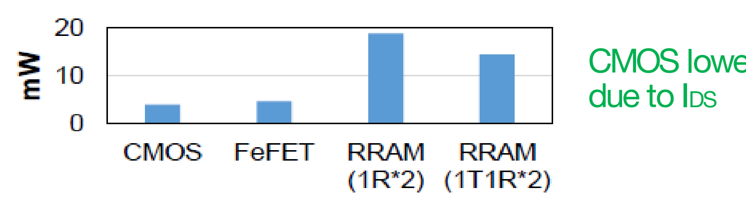
RC parasitic parameters used in simulations

	Cell area (F^2)	R_{wire} (Ω)	C_{wire} (fF)
CMOS	150	0.245	0.059
FeFET	60	0.155	0.037
RRAM (1R*2)	4 (×2) [30]	0.04	0.0096
RRAM (1T1R*2)	20 (×2) [30]	0.09	0.022

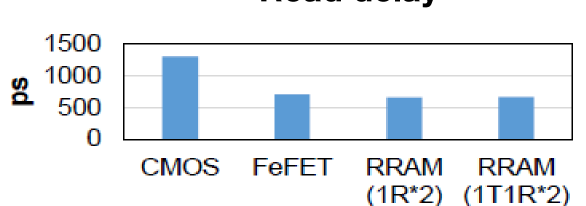


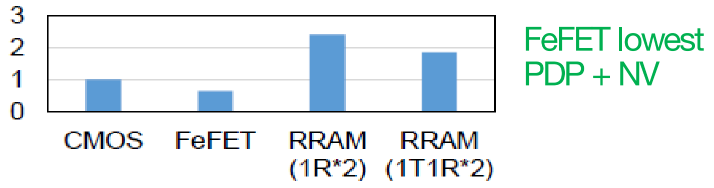
For CMOS programming, can also set HLs to 0, SRAM write power low

RRAM requires >> write voltage, higher power



CMOS lowest due to I_{DS}

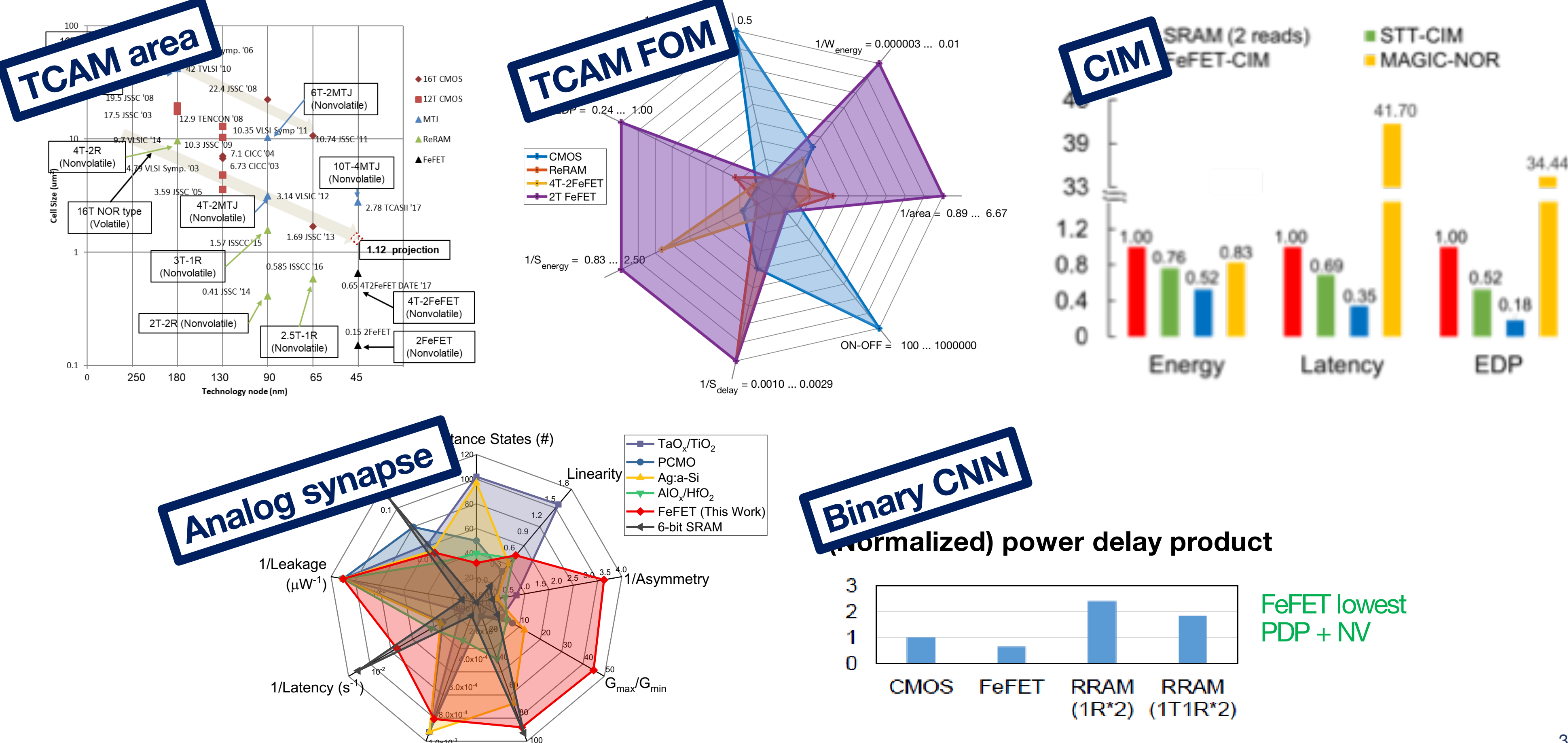




FeFET lowest PDP + NV

Xioaming Chen, Xunzhao Yin, Michael Niemier, and Xiaobo Sharon Hu, "Design and Optimization of FeFET-based Crossbars for Binary Convolutional Neural Networks," to appear in Design, Automation, and Test in Europe (DATE), 2018.

Takeaways … promising metrics!



Joint University Microelectronics Program

www.src.org/program/jump

Semiconductor Research Corporation

@srcJUMP