Is it Memory or Logic? Blurring the Gap

Prof. Vijaykrishnan Narayanan
The Pennsylvania State University

In Collaboration with

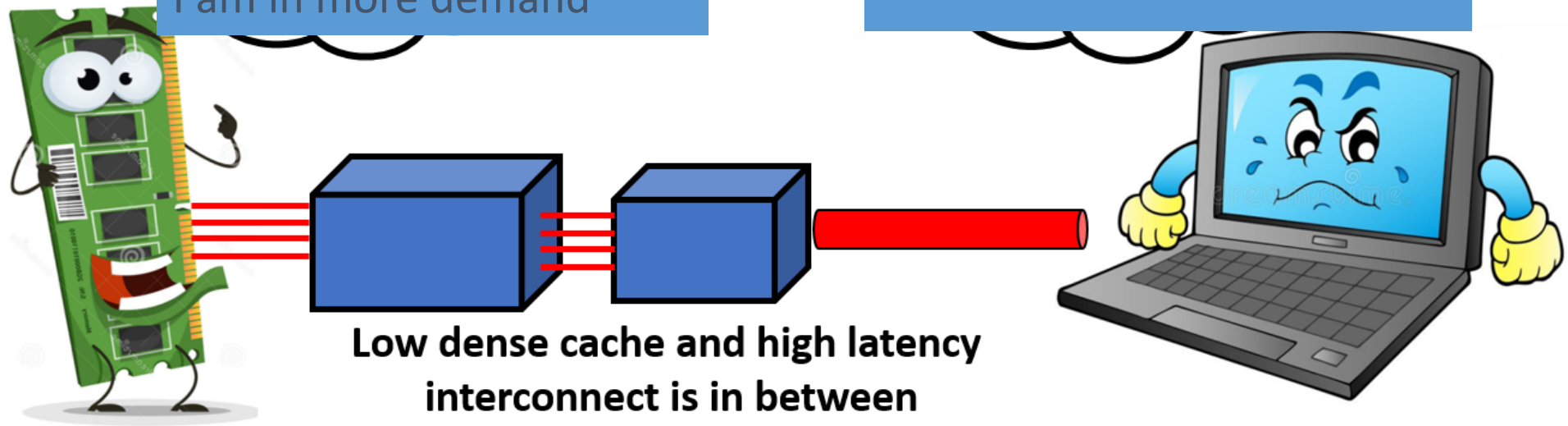
Professors Suman Datta, Sayeef Salahuddin, Sumeet Gupta, Sharon Hu, Michael Niemier, (Marvin) Mei-Fan Chang

Supported in part by NSF expeditions-in-computing, DARPA/SRC LEAST Center, NSF ASSIST ERC

September 2017

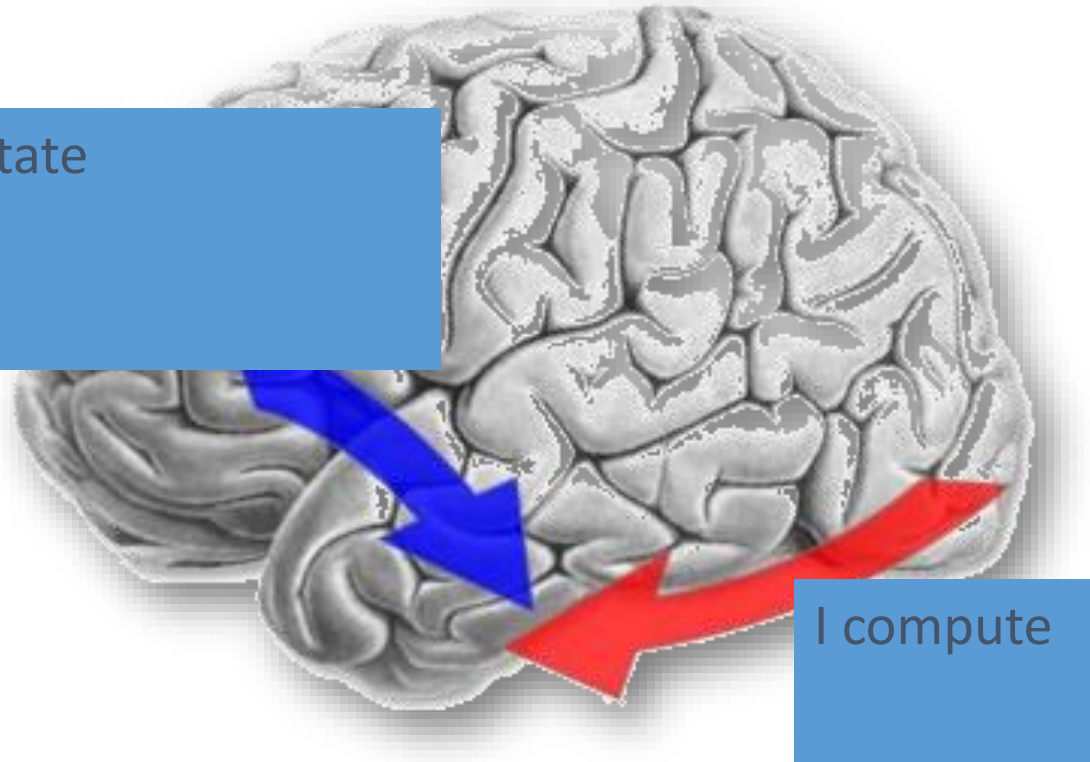
You will loose your mind without me !
I am in more demand

You are just a leaky bucket
I make most EDA revenue



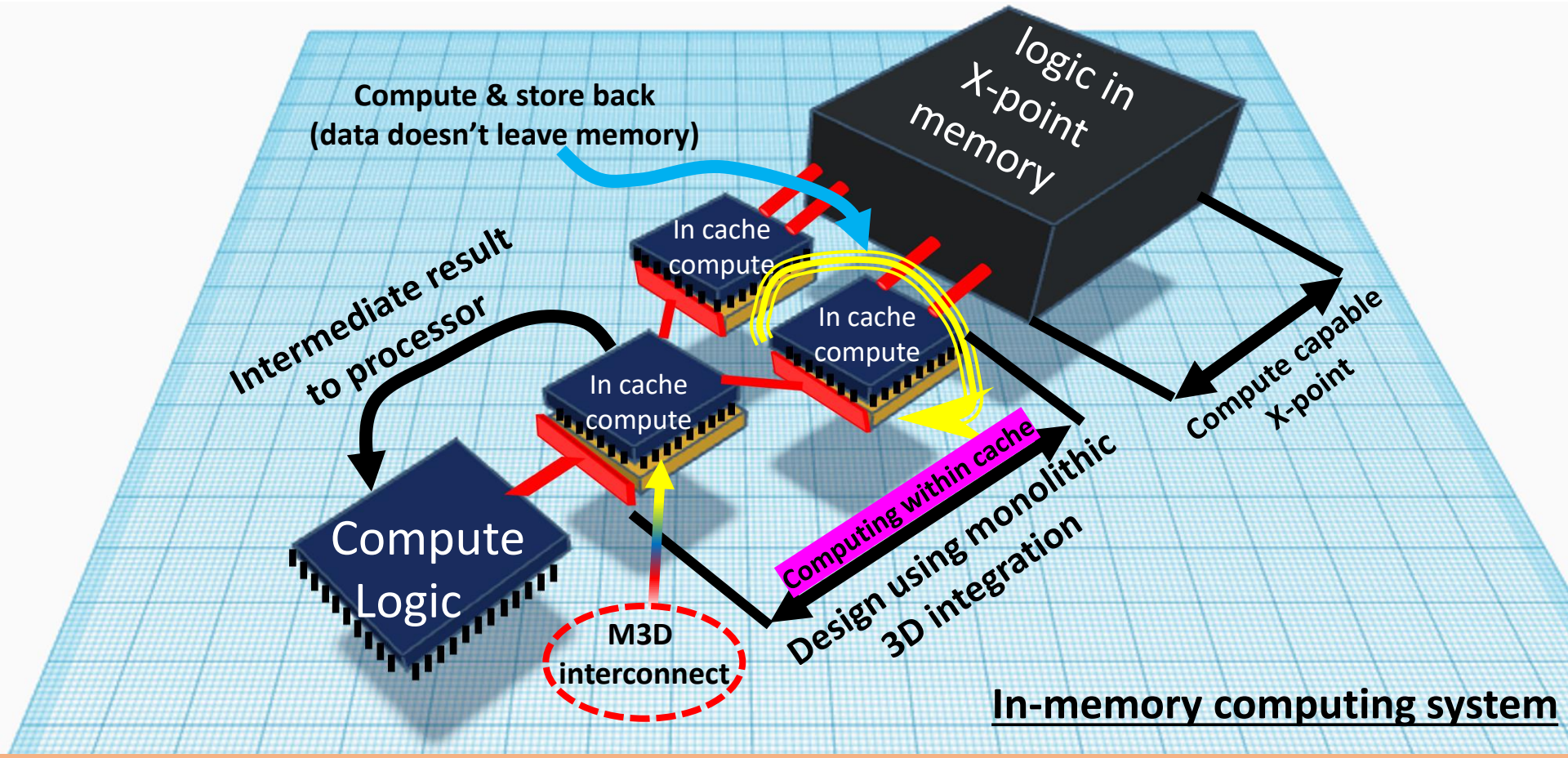
Low dense cache and high latency interconnect is in between

I store state



I compute

Why can’t you get smarter as well

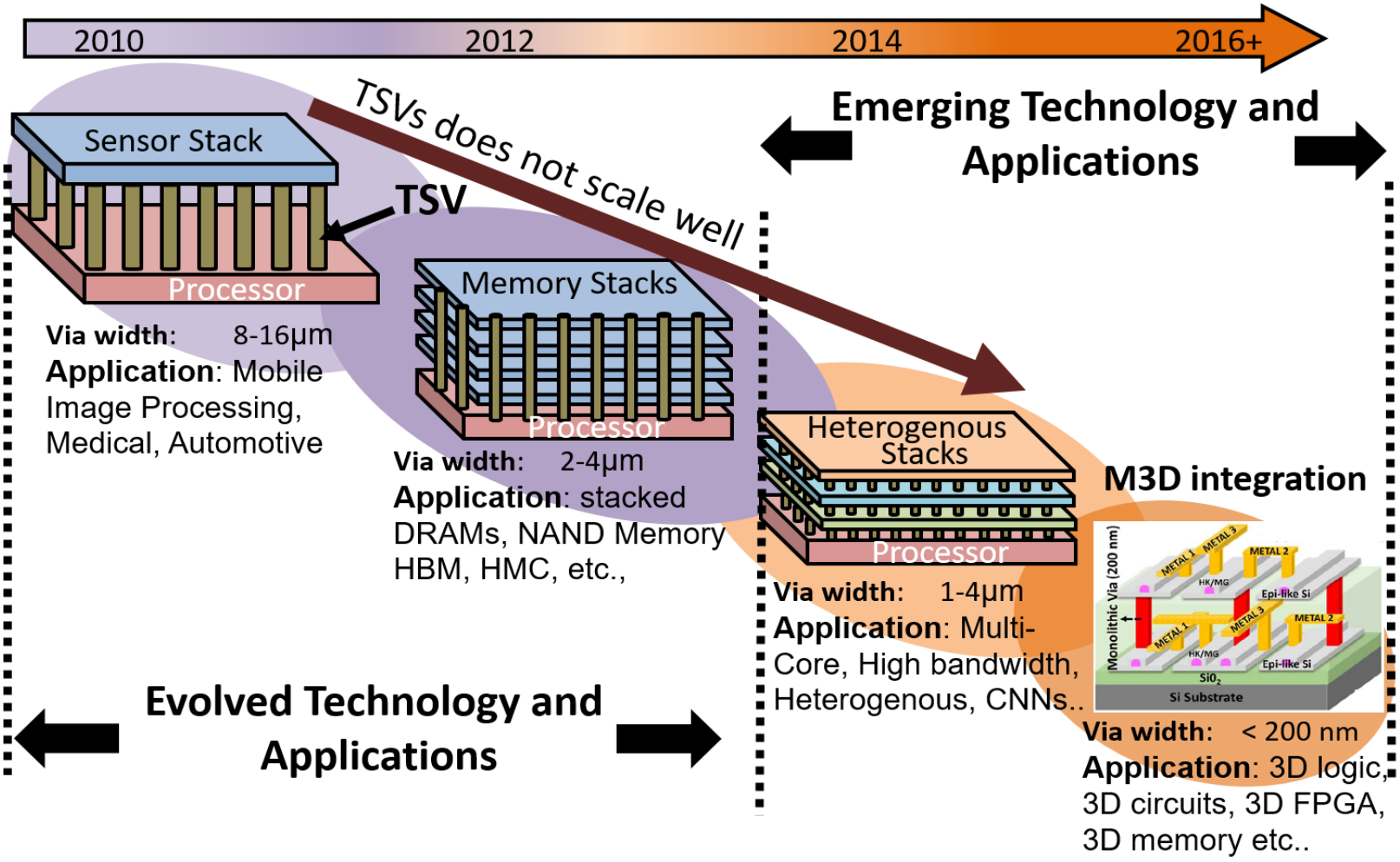


Offloading computations to memory helps in reducing data movement from farther memory to the processor

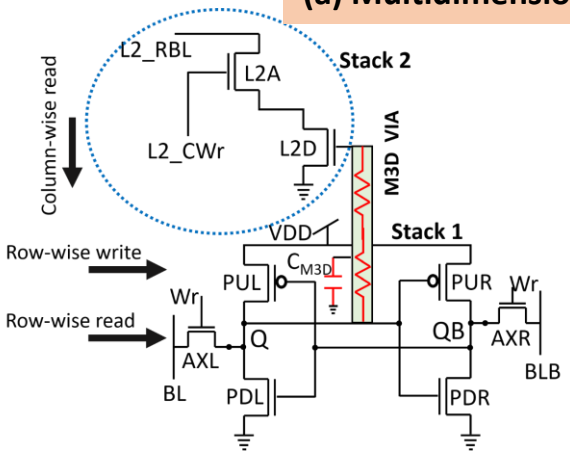
Overview

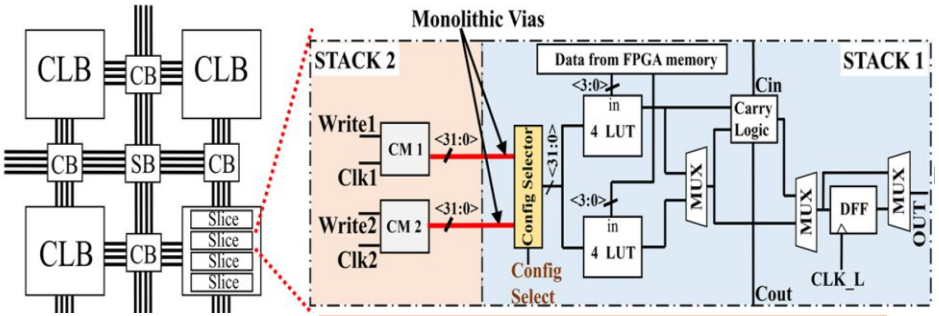
- Monolithic 3D Integration
- Configurable Memory-Logic Device
- Cross point arrays

3D Integration: Technology and Evolution



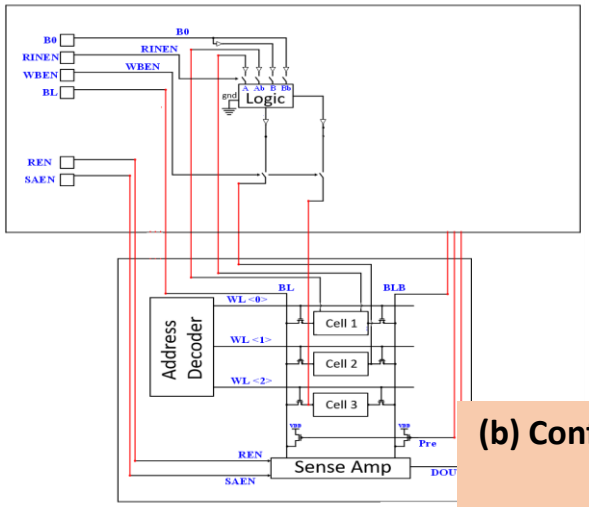
Compute-Oriented Caches



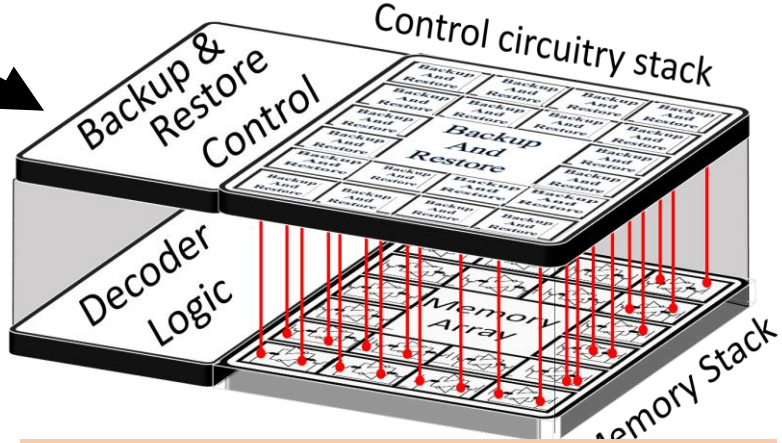


(c) Rapid reconfigurable M3D FPGA

Compute capable SRAM

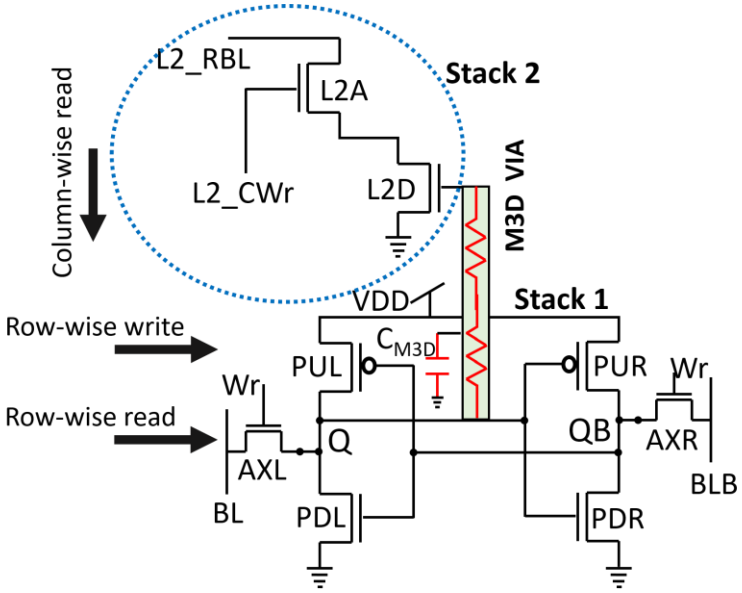


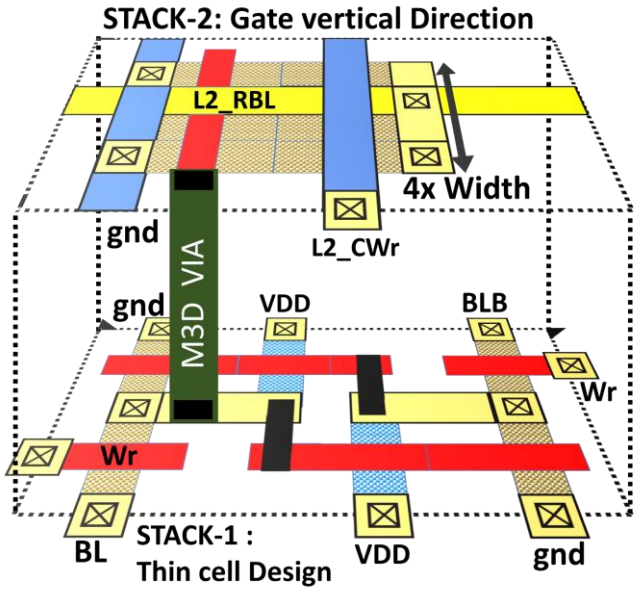
(b) Configurable Boolean logic near memory



(d) Feature addition: making SRAM non-volatile

Concurrent row and column accessible 3D memory



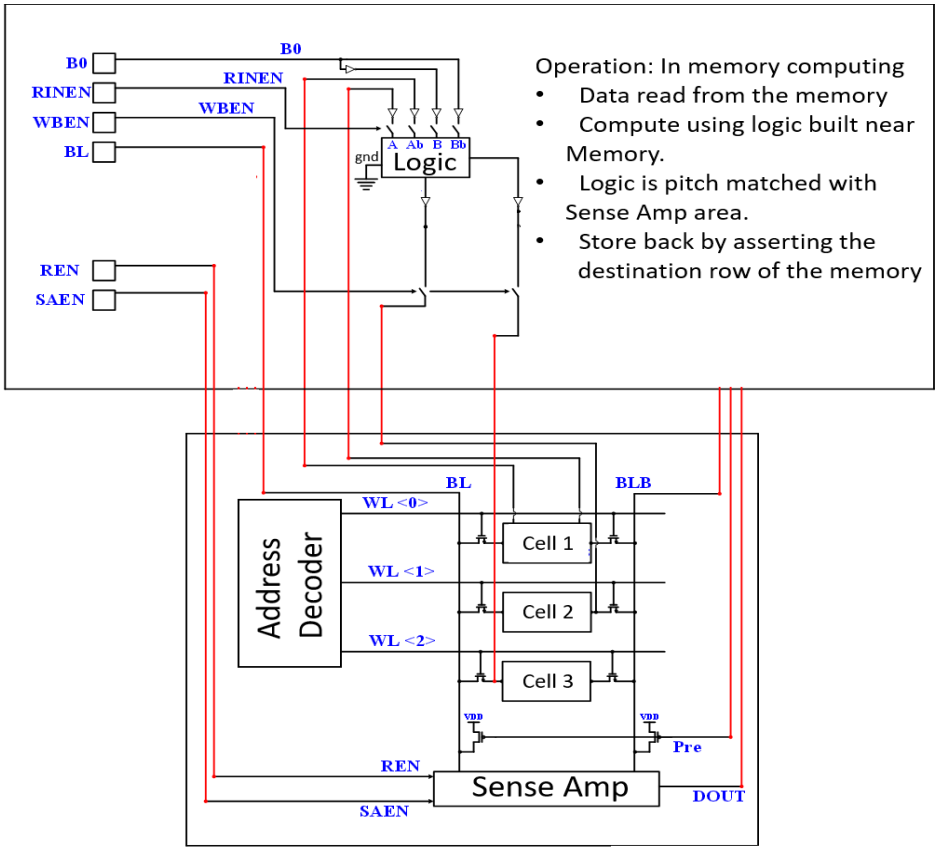


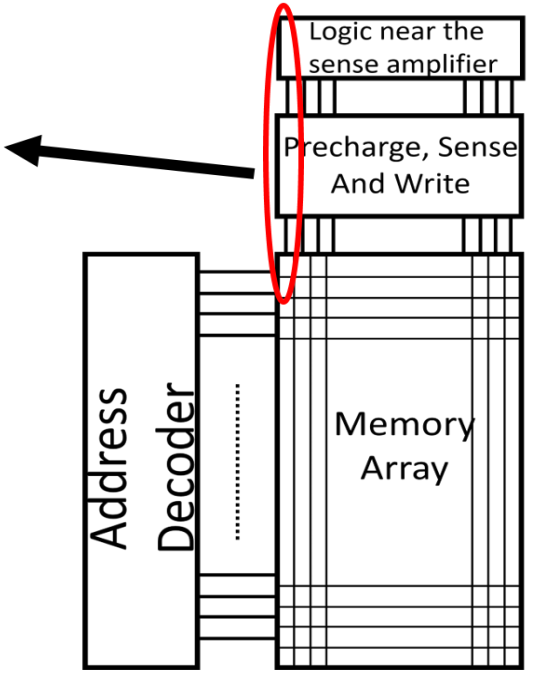
2.15x access time savings

This memory design can cater to applications requiring multi-dimensional data access for enhancing system performance.

Transforms traditional compiler optimizations

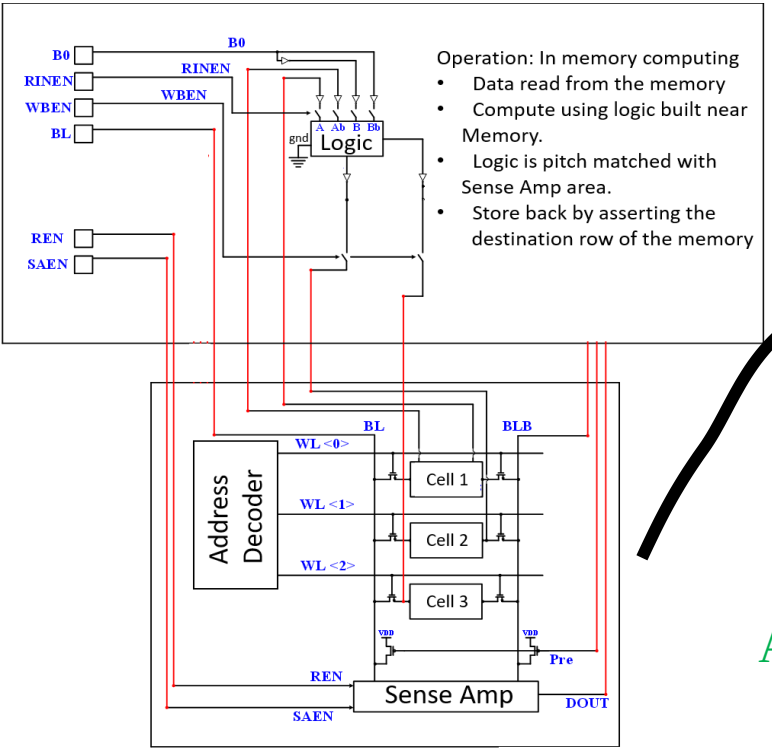
LUT based in-memory computing.



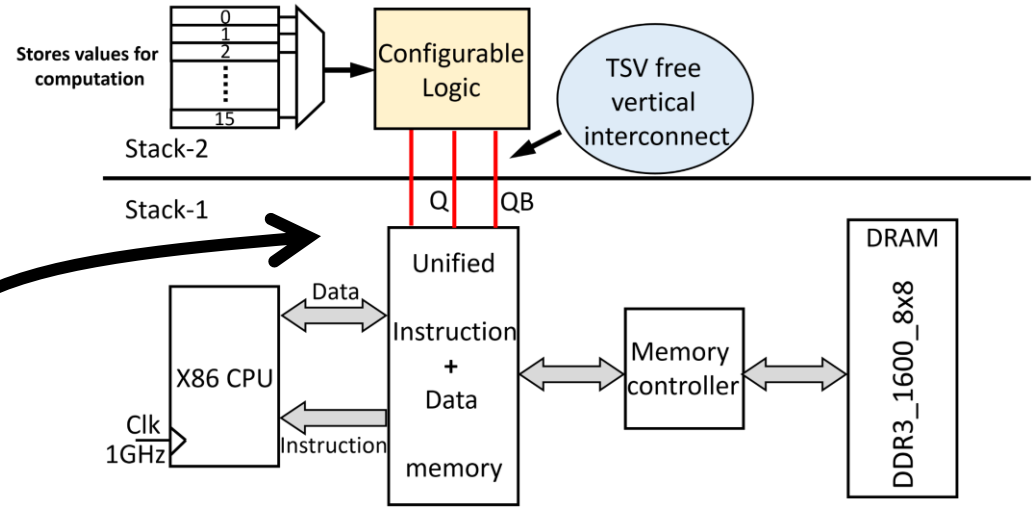


Boolean and arithmetic computations in memory along with store back feature

Code Compilation for in-memory compute



In memory compute capable SRAM



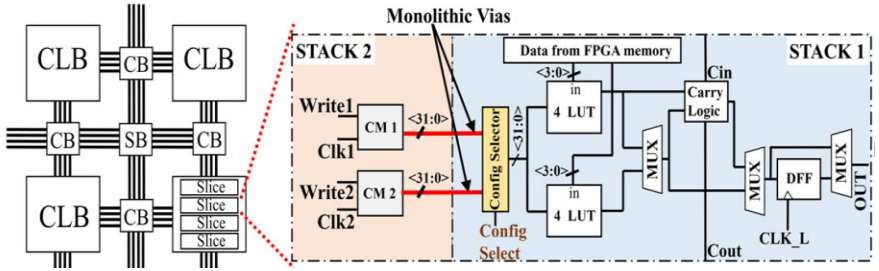
In memory compute SRAM as part of the evaluation system

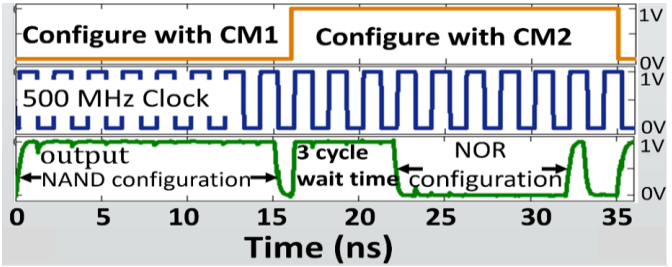
A workload of 1 million instructions with frequent Boolean operations were executed in this system

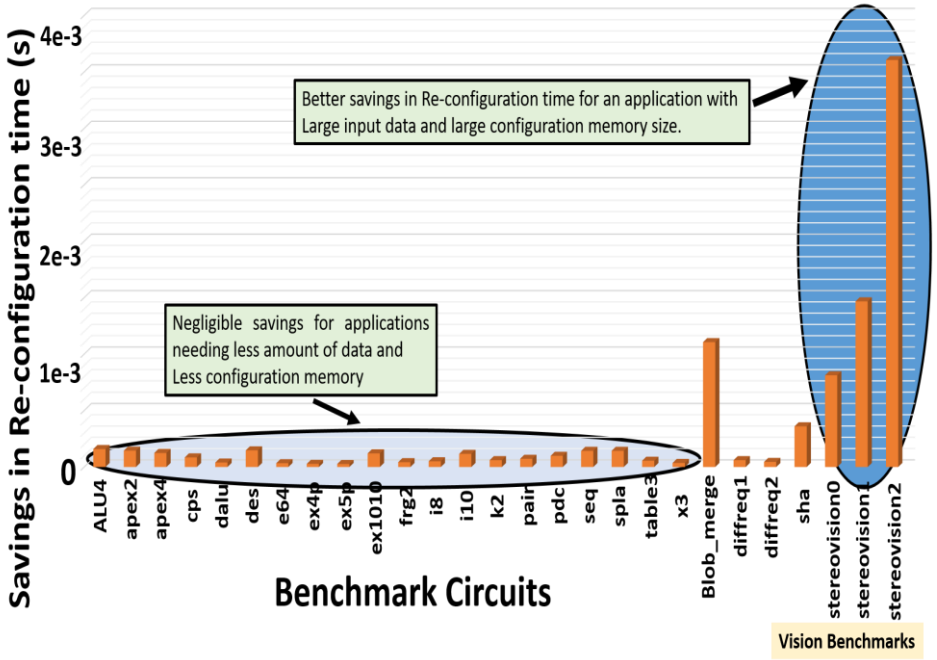
39% of the executions were done in memory

Where to do the compute – memory or logic ?

Rapid reconfigurable monolithic 3D FPGA (what FPGA)







Rapid reconfiguration enables
large applications to load in ‘ns’

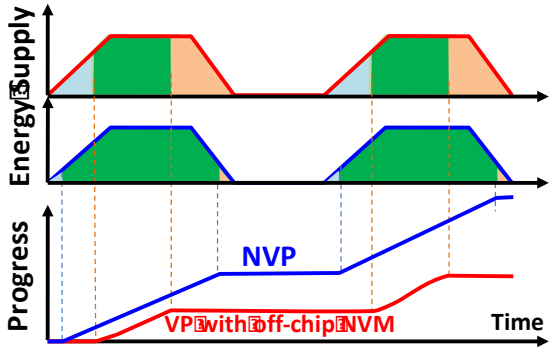
Overview

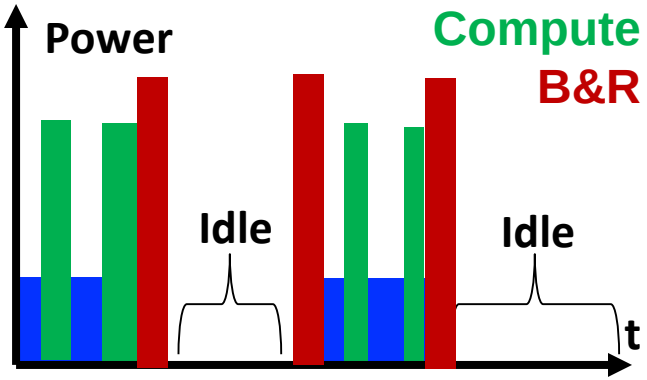
- Monolithic 3D Integration
- Configurable Memory-Logic Device
- Cross point arrays

New Pathway: Integration of NVM and Processor

Power Source Driven -- Frequent backup and restores (B&R)

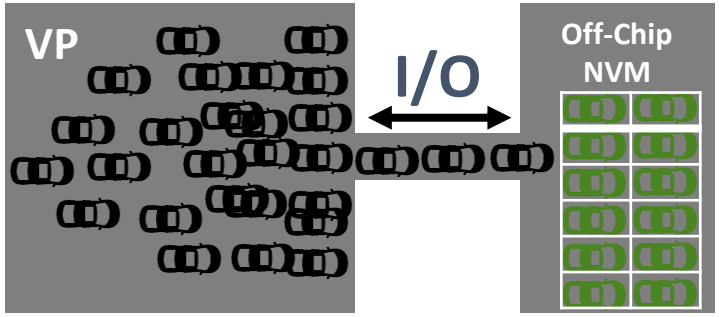
- e.g. Energy-harvesting computing systems → intermittent power supply
- e.g. Low stand-by power systems → complete fine-grain power-gating
- Key: reducing energy overhead!

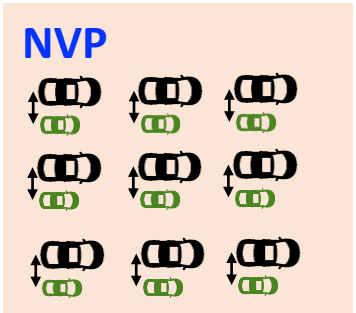


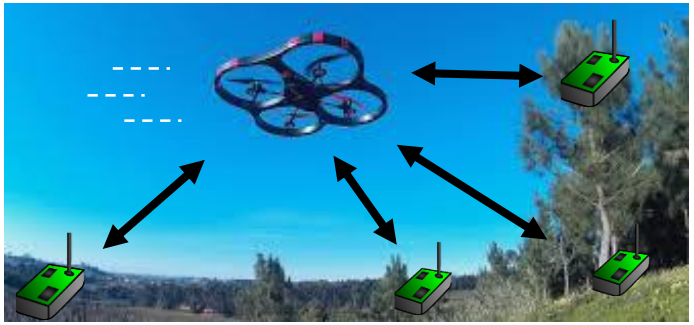


Event Driven -- Quick response in normally-off Applications

- Key: low-latency restore



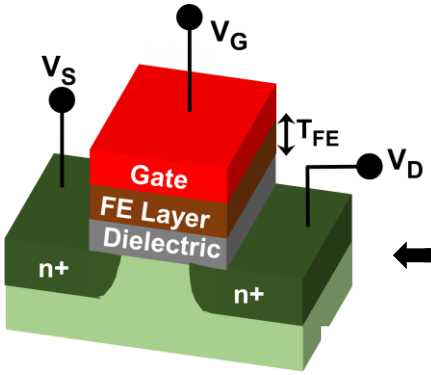




NCFET Modeling and Evaluation

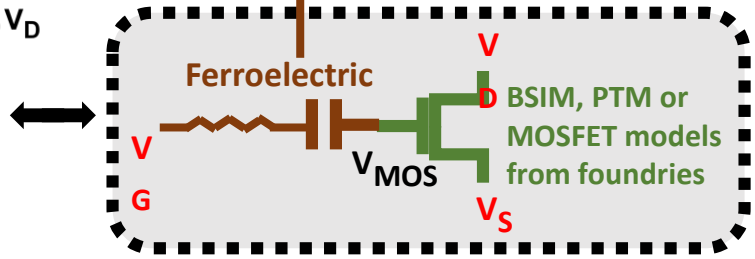
From Physics to Devices

- **Physics-based**: Employing time dependent LK equations solved self-consistently with MOSFET equations
- **Circuit compatible**: SPICE based model enables seamless integration of the model with circuit simulators
- Enables efficient device-circuit co-design and analysis from materials through circuits.



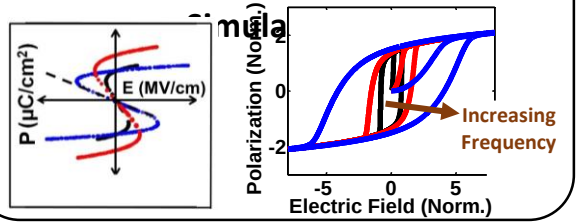
$$E - \rho\frac{dP}{dt} = 2\alpha P + 4\beta P^3 + 6\gamma P^5$$

$$\Delta V_{DEPOLARIZATION} = \frac{\lambda_{EFF}}{\varepsilon_{CON}\ A_{FE}} Q_{FE}$$

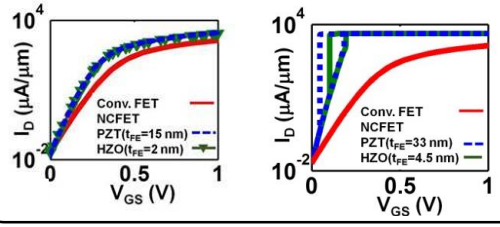


- Time dependent LK equations solved using circuit elements (Comparators, VCVS etc.)
- Depolarization field is accounted for to model non-ideal contacts
- Kinetic co-efficient ρ captures the finite polarization switching time
- On-going efforts to model multi-domain FE including minor loops and AFE.

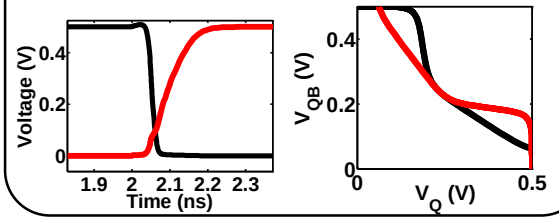
Ferroelectric P-E Simula



Device Simulations

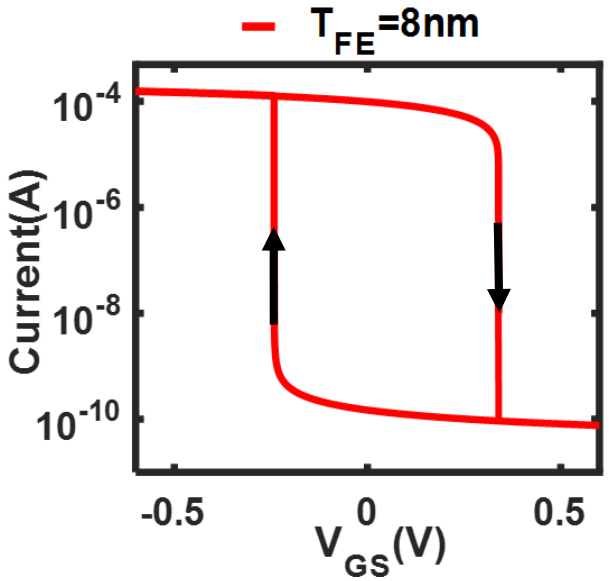


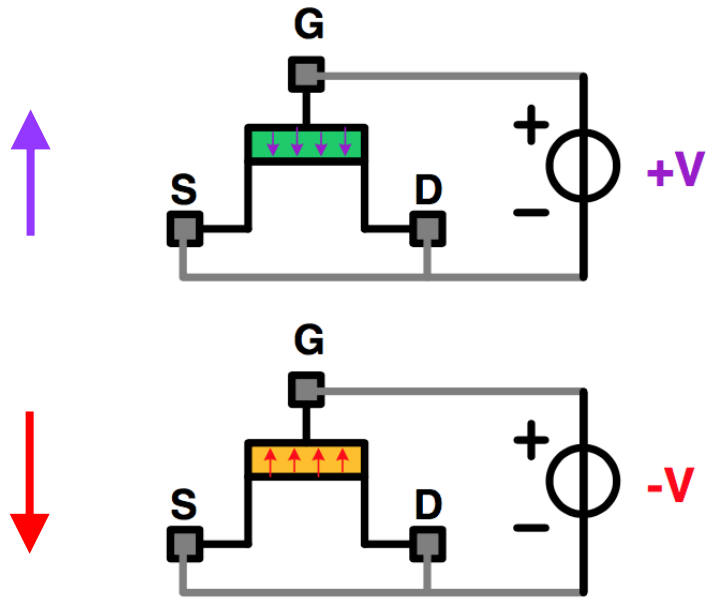
Circuit Simulations



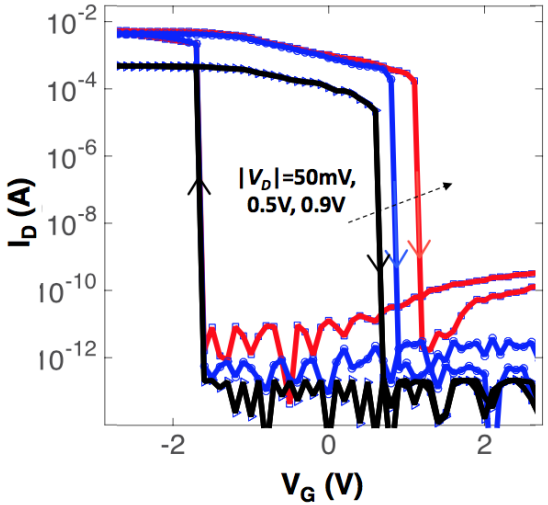
Focus: Memory-Logic Integration

Theory





Experiments

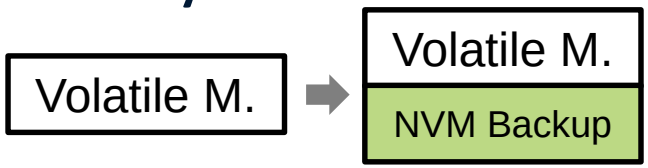


New Opportunities:
1. Memory-Logic Integration
2. NVM: scalable, low energy

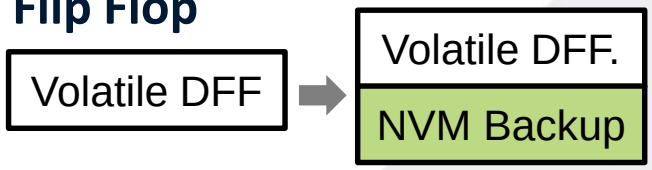
Experimental Results from Prof. Salahuddin

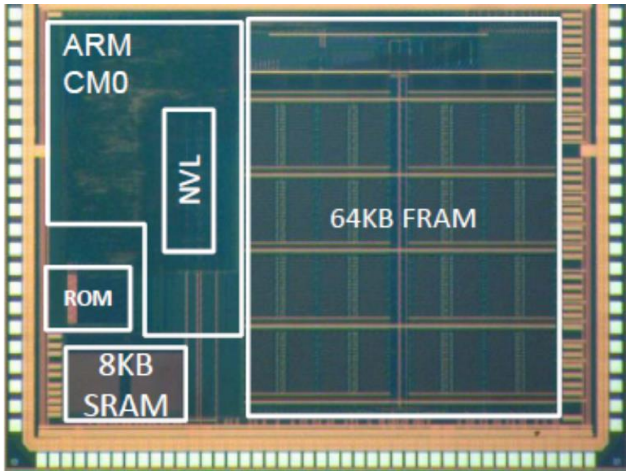
Xueqing Li et al, IEEE TED, vol. 64, no. 8, Aug. 2017; Patent filed;

Memory

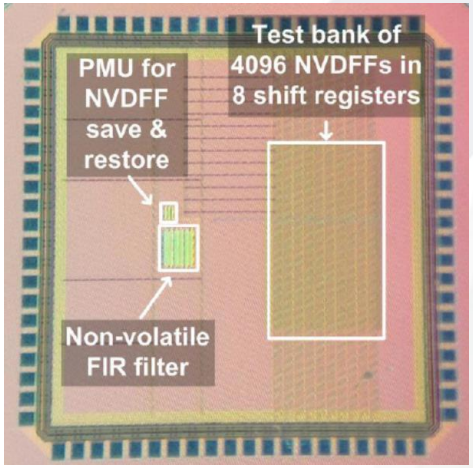


Flip Flop

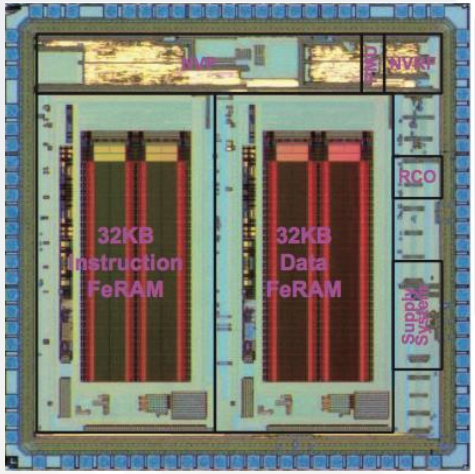




ISSCC'14, TI



ISSCC'13, MIT, TI



VLSI'17, Tsinghua

Our Contributions

FeFET NVP SoC

- nvSoftware
- nvMem
- nvLogic
- nvAccelerator
- nvIO
- nvCache

- nvDFF: intrinsic non-volatility
- NVM: intrinsic non-volatility

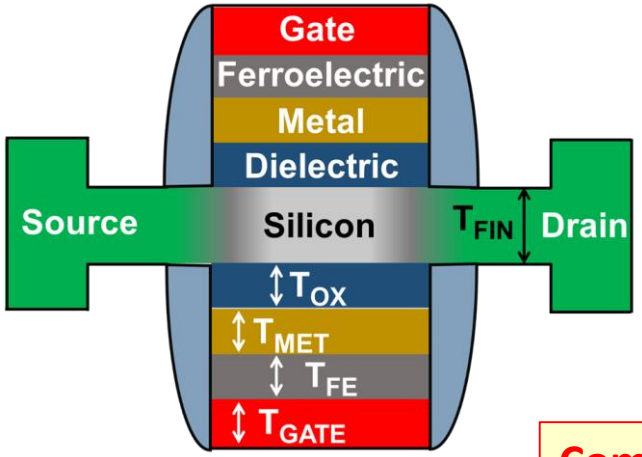
Compact Model

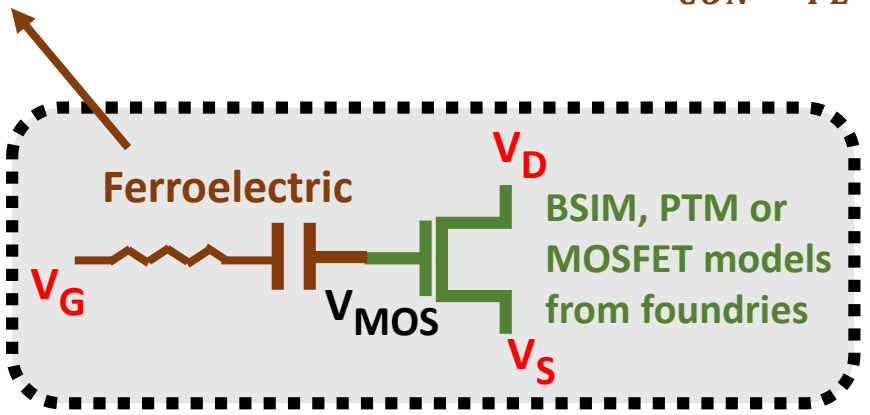
Key Features

- **Physics-based**: Employing time dependent LK equations solved self-consistently with MOSFET equations
- **Circuit compatible**: SPICE based model enables seamless integration of the model with testbenches. Compatible with existing commercial circuit simulation tools
- Enables efficient device-circuit co-design and analysis from materials through circuits.

$$E - \rho \frac{dP}{dt} = \alpha P + \beta P^3 + \gamma P^5$$

$$\Delta V_{DEPOLARIZATION} = \frac{\lambda_{EFF}}{\varepsilon_{CON}\ A_{FE}} Q_{FE}$$



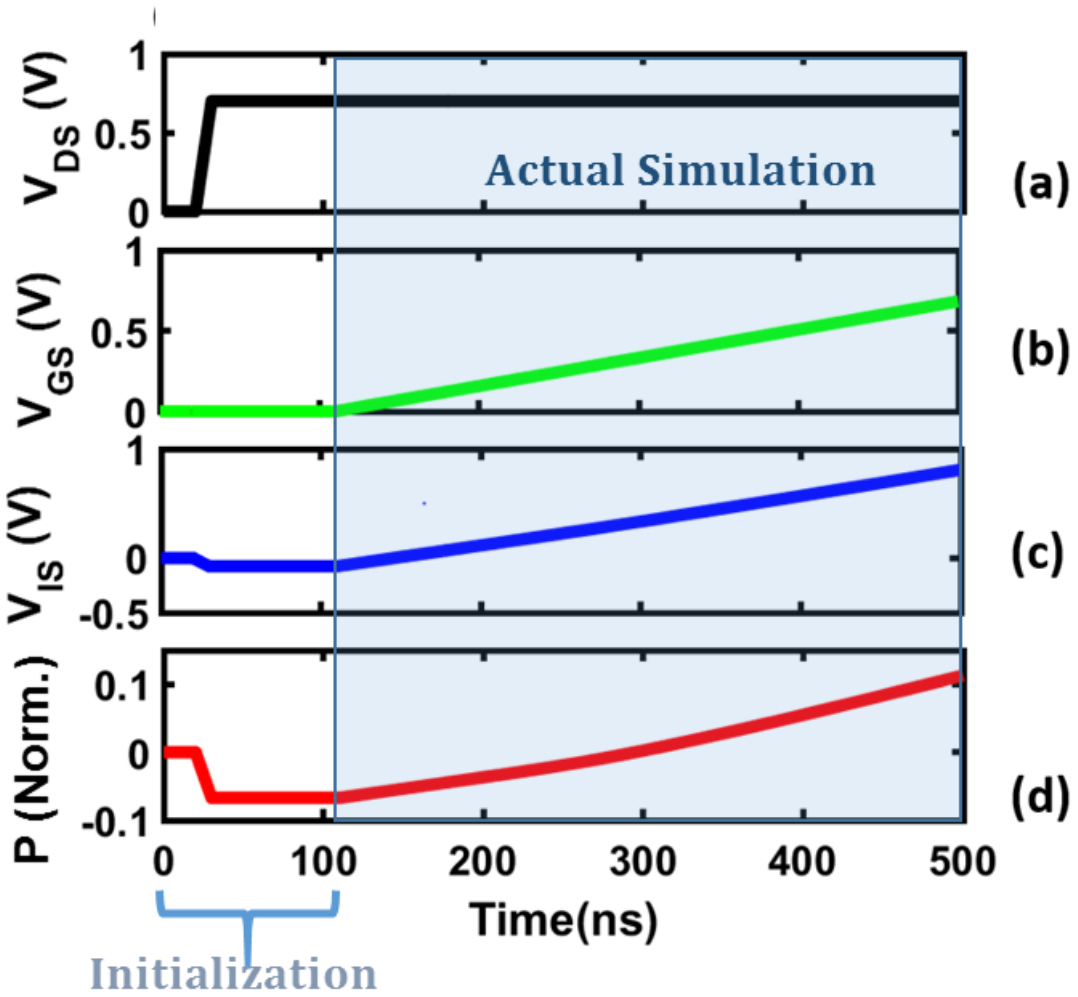


Aziz et al, EDL 2016

Compact model developed and is being refined

- Includes effects of finite polarization switching time
- Accounts for depolarization fields
- On-going effort to include other physical effects

Compact Model: Initialization

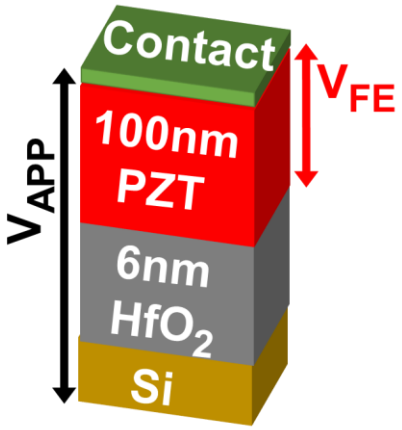


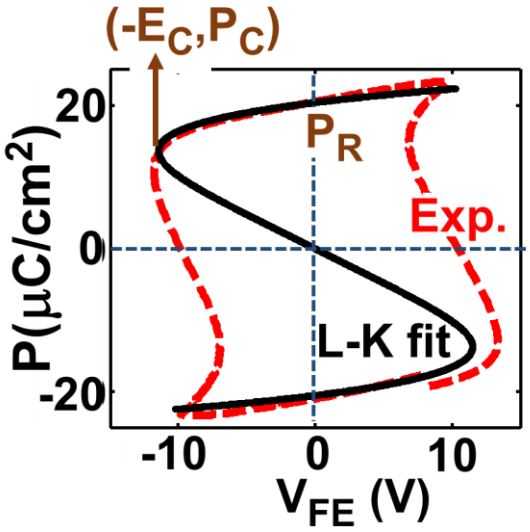
- Initialization is required to correctly capture the effect of gate and drain voltages on FE polarization.
- Device simulations are started with all voltages (V_{GS}, V_{DS}) at 0. The voltages are then ramped to desired values to capture the polarization trajectory.
- Circuit simulation are started by ramping supply rail from 0 to V_{DD} and then applying the inputs voltages.

S. Gupta, et al, IEEE TED, 2017

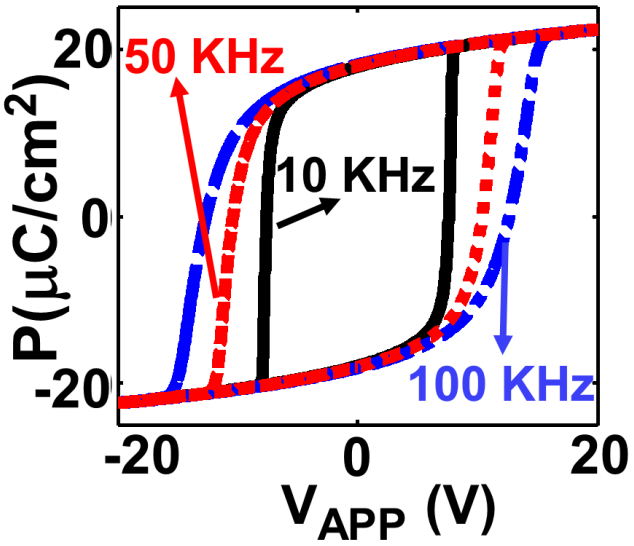
Model Calibration

$$E - \rho\frac{dP}{dt} = \alpha P + \beta P^3 + \gamma P^5$$





Simulations



α= -1.05x10^9 m/F

β=1x10^7 m^5/F/C^2

γ=6x10^{11} m^9/F/C^4

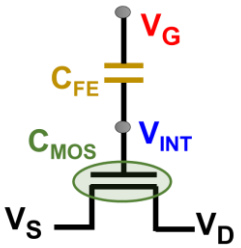
Static coefficients extracted from calibration against experiments

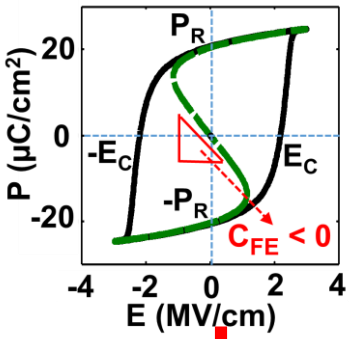
Behavior of the model with respect to varying input frequencies consistent with other works (Kobayashi et al, *VLSI Tech* 2015)

Aziz et al, EDL 2016

Compact Model: Capturing Different Modes of Operation

LANDAU KHALATNIKOV (LK) Model





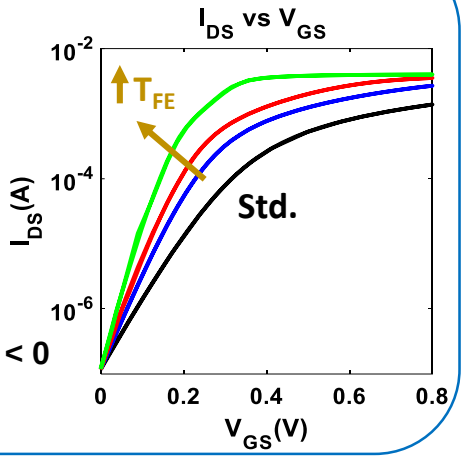
α= -1.05x10^9 m/F
β= 1x10^7 m^5/F/C^2
γ= 6x10^{11} m^9/F/C^4

STEEP-SWITCHING FEFET

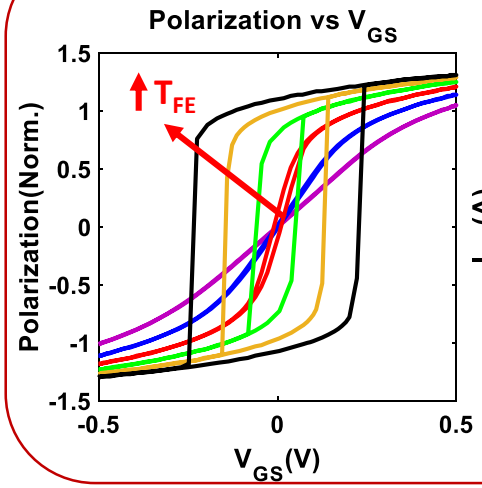
$$\frac{V_{MOS}}{V_G} = \frac{|C_{FE}|}{|C_{FE}| - C_{MOS}}$$

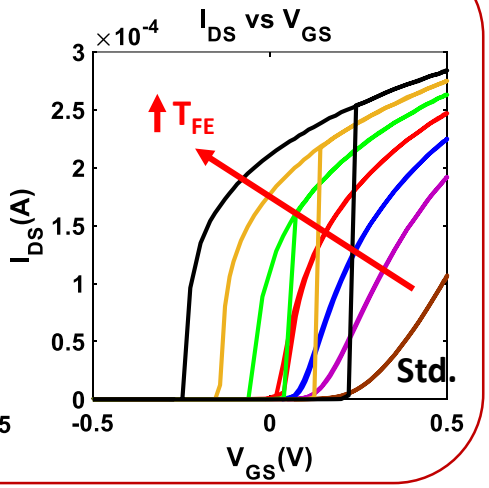
$$\frac{V_{MOS}}{V_G} > 1, \text{ for } C_{FE} < 0$$

SS<60mV/decade, for $C_{FE} < 0$

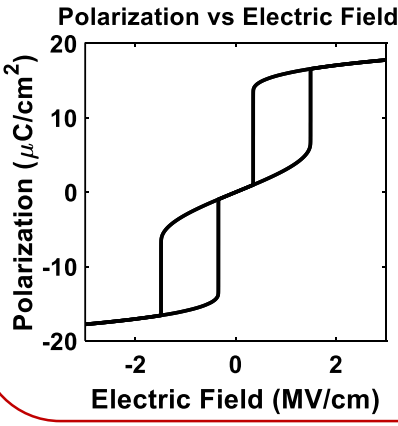


NON-VOLATILE FEFET





ANTI-FERROELECTRICITY (AFE)

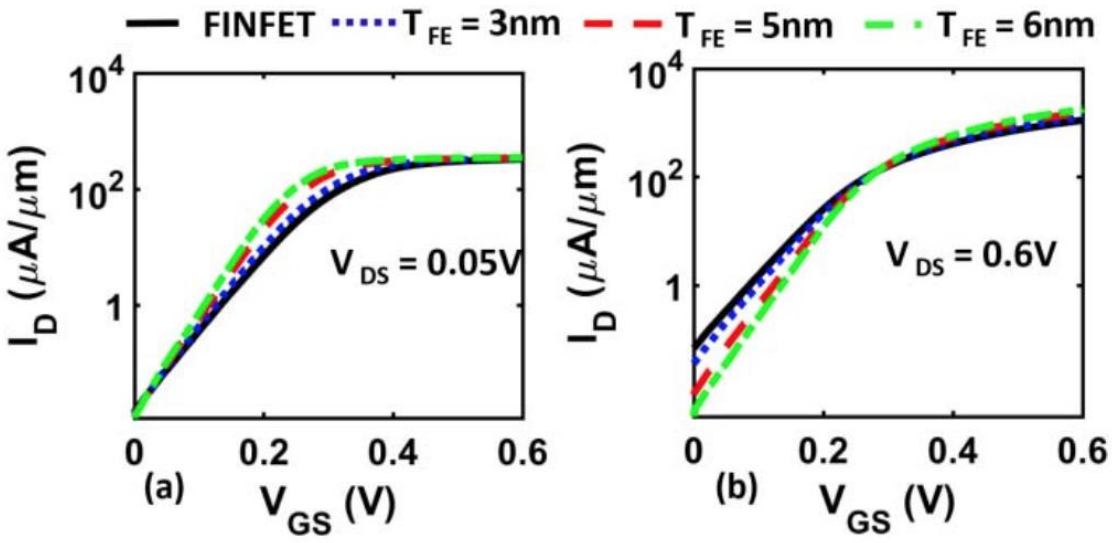


α= 1.55x10^9 m/F
β= -7.2x10^9 m^5/F/C^2
γ= 9x10^9 m^9/F/C^4

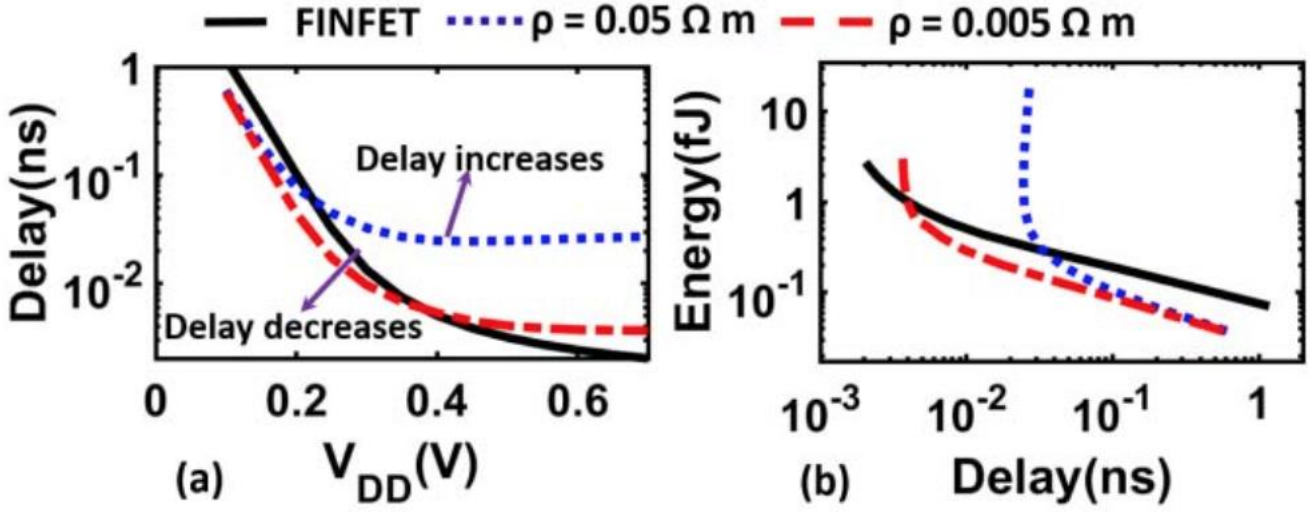
Device analysis with AFE ongoing

By varying the parameters (LK parameters, FE thickness (T_{FE}), oxide metrics of the underlying transistor), steep switching, non-volatile and anti-ferroelectric behavior can be obtained

FeFET Logic Design Benchmarking

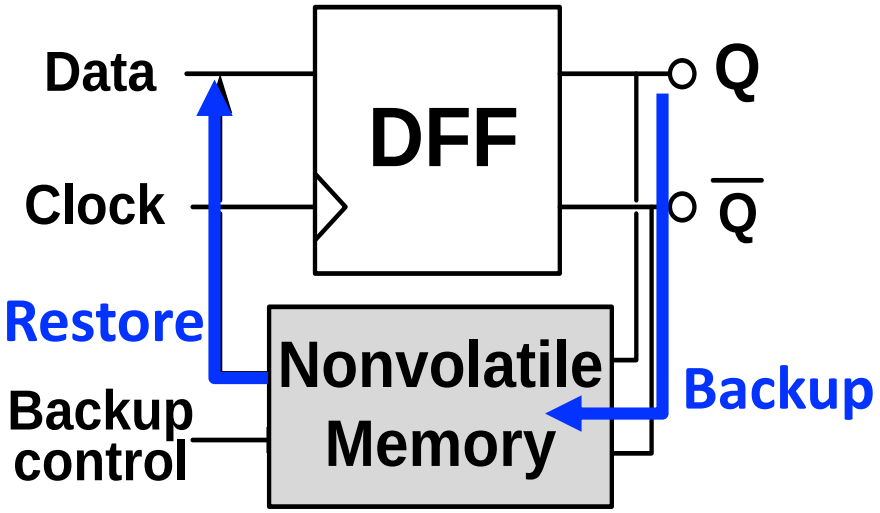


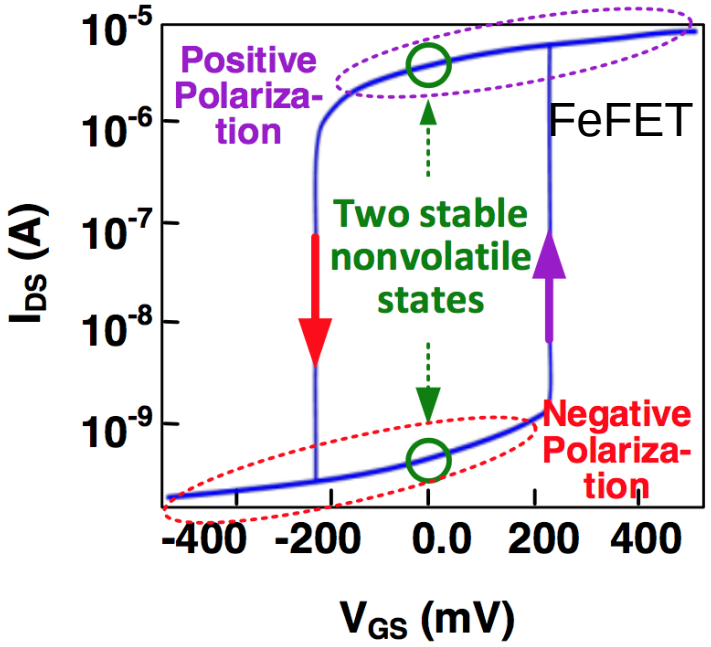
FeFET Id-Vgs comparison with MOSFET



For an inverter in seven-stage RO
$C_W = 0, T_{FE} = 5$ nm

nvDFF: Concept and Significance

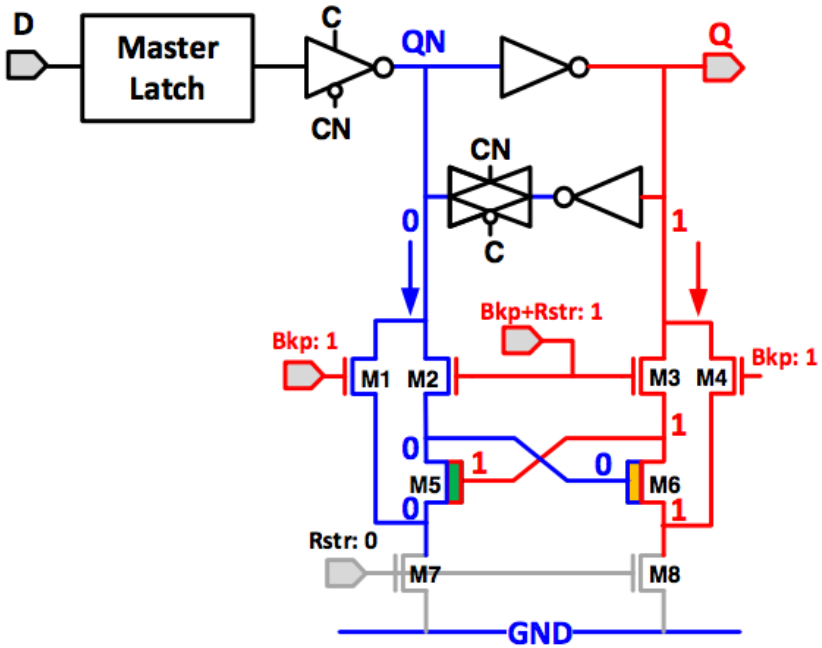




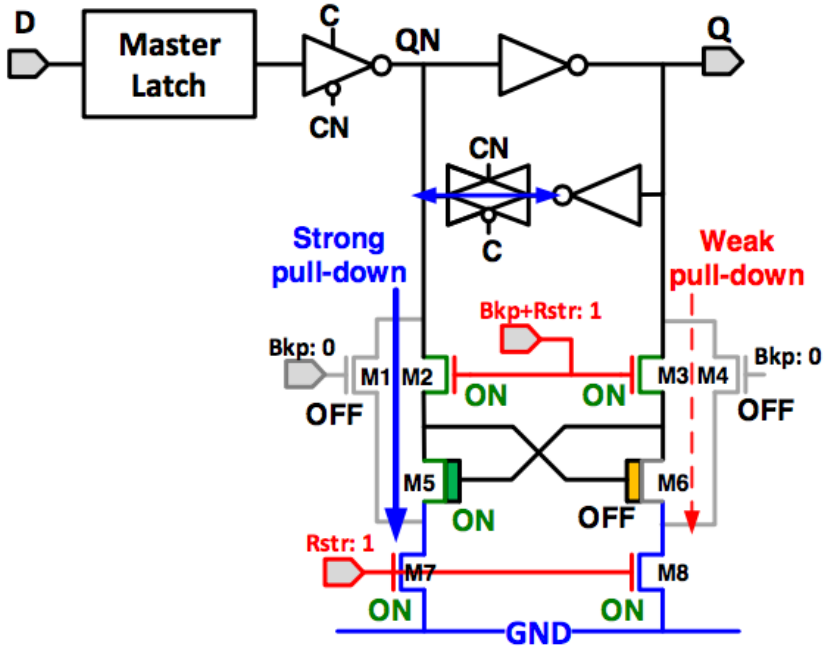
- **Low backup and restore (B&R) energy and latency;**
- On-demand or automatic B&R
- Other concerns: area, control complexity, retention time, circuit interface, process compatibility, etc.

nvDFF1: Edge Computing with FeFET – Operation Theory

Backup Operation



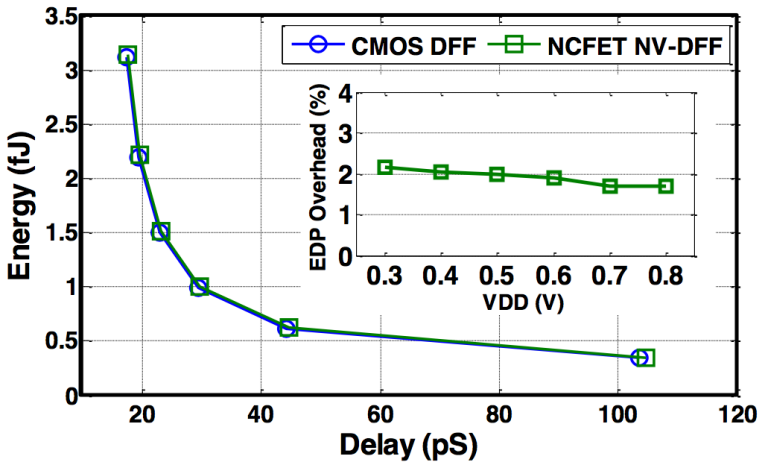
Restore Operation



Xueqing Li et al, IEEE TED, Aug. 2017; Patent filed;

nvDFF1: Edge Computing with FeFET – Benchmarking

Normal operation overhead



Performance benchmarking

	[10] Measured	[9] Simulated	[11] Simulated&	This Work Simulated*		
Tech. size	130nm	70nm	180nm	10nm		
Voltage	1.5V	1.0V	1.8V	0.3V-0.8V		
Material	PZT Cap	MTJ	ReRAM	6nm HfO_2, PZT		
				ρ=0.04	ρ=0.10	ρ=0.25
$T_{Backup+Restore}$	2.67µS	>10uS	1.3µS	277pS	583pS	1.29nS
$E_{Backup+Restore}$	**2.40pJ**	**382fJ**	**735fJ**	**1.38fJ**		
Break-Even Time	/	**0.83µS@25°C**	**1.47mS**	**55.9nS**		

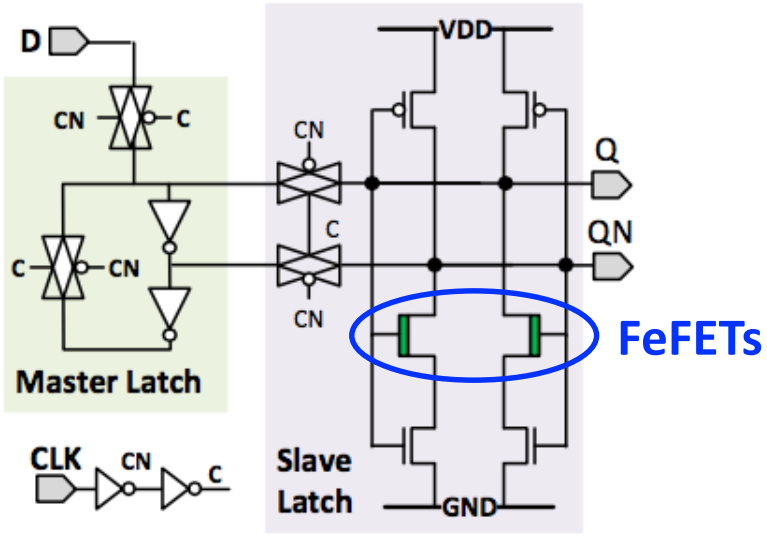
&: The results are for the topology of NVFF-I in [11] operating at 0.8 V supply (rise to 2.4V for ReRAM write) for the shortest break-even time.

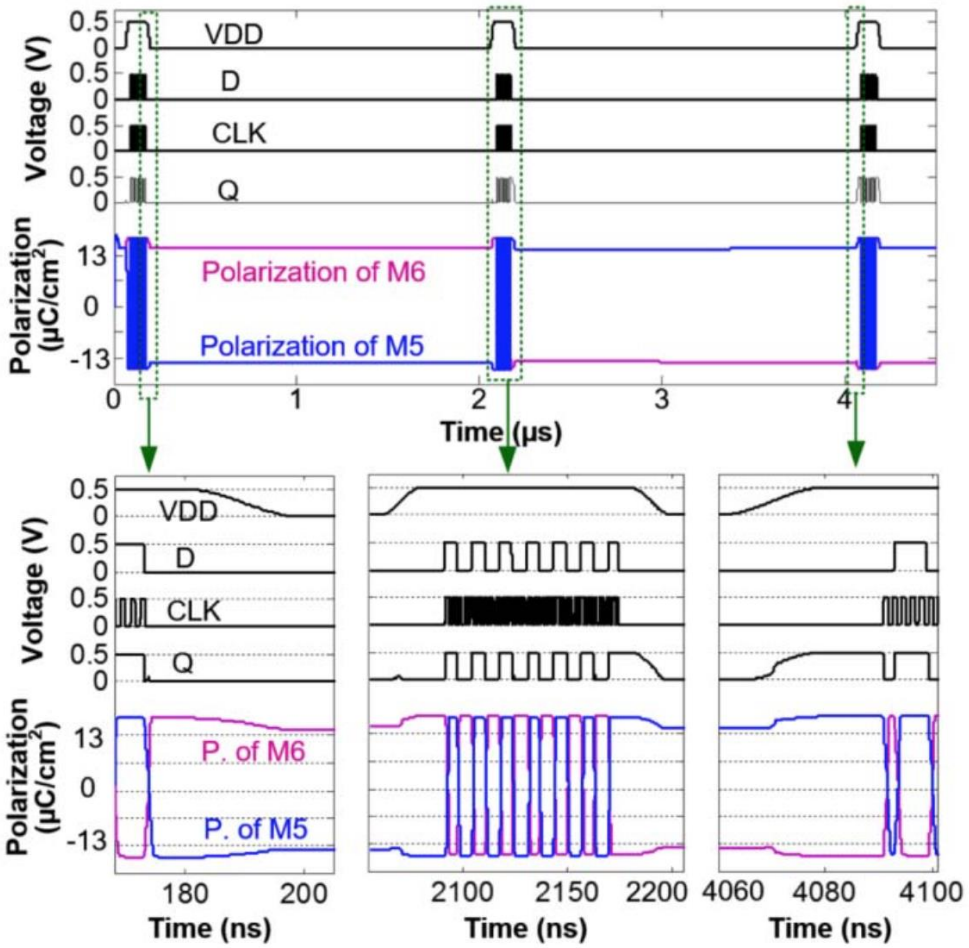
*: Backup and restore performance in this table is simulated at 0.5V supply.

- ✓ **Ultra-Low $E_{B\&R}$**
- ✓ **Ultra-Low latency**
- ✓ **Low normal-operation overhead**

Xueqing Li et al, IEEE TED, Aug. 2017; Patent filed;

nvDFF2: Intrinsically Nonvolatile DFF



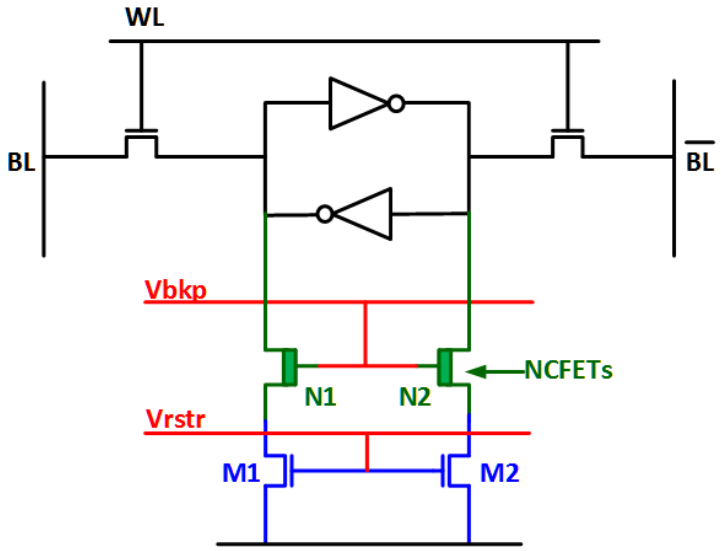


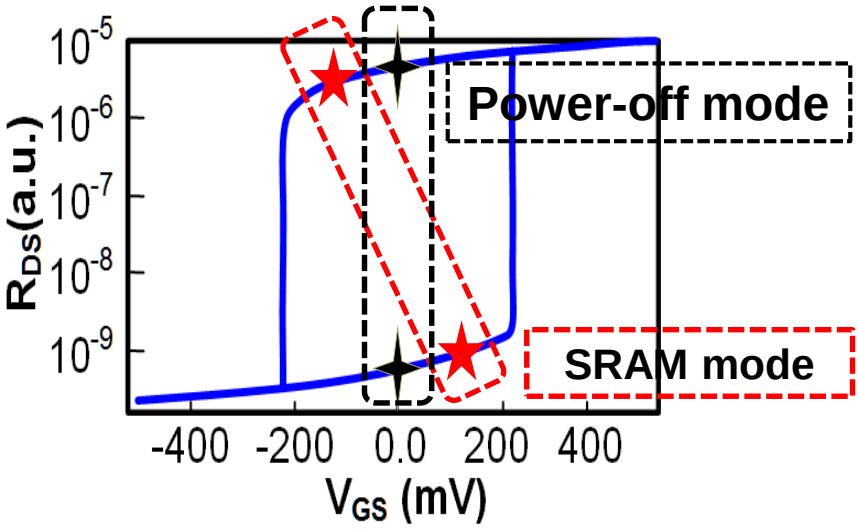
1. **Backup/restore features**
   - ✓ **Fast: done in sub-nS (1000x);**
   - ✓ **Low-energy: < 2.4fJ@0.8V (1000x);**
   - ✓ **Autonomous**
     - ➢ **no ext. control needed**

   Compared with FeCAP solution
2. **Normal operation features**
   - ✓ **Fast: GHz operation;**
   - ✓ **EDP overhead: ~35% (x4 fan-out)**
3. **Dense (only 2 added transistors), low-voltage, scalable, CMOS compatible**

Xueqing Li et al, IEEE Trans. CAS-I, vol.PP, no.99, pp.1-13; Patent filed;

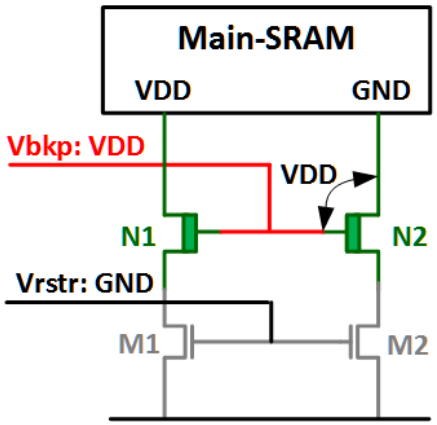
nvSRAM: Enabling Nonvolatile Computing with FeFET NVM



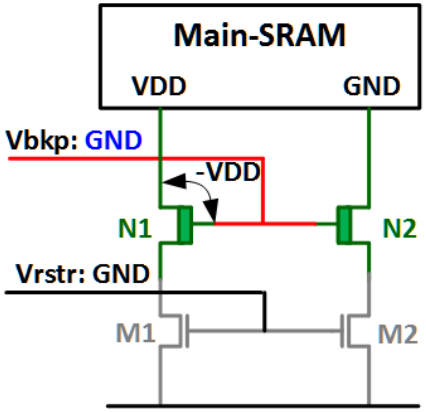


Restore theory: similar to nvDFF (based on pull-down difference)

Backup theory:



Backup step-1:
N2 → Pos. state

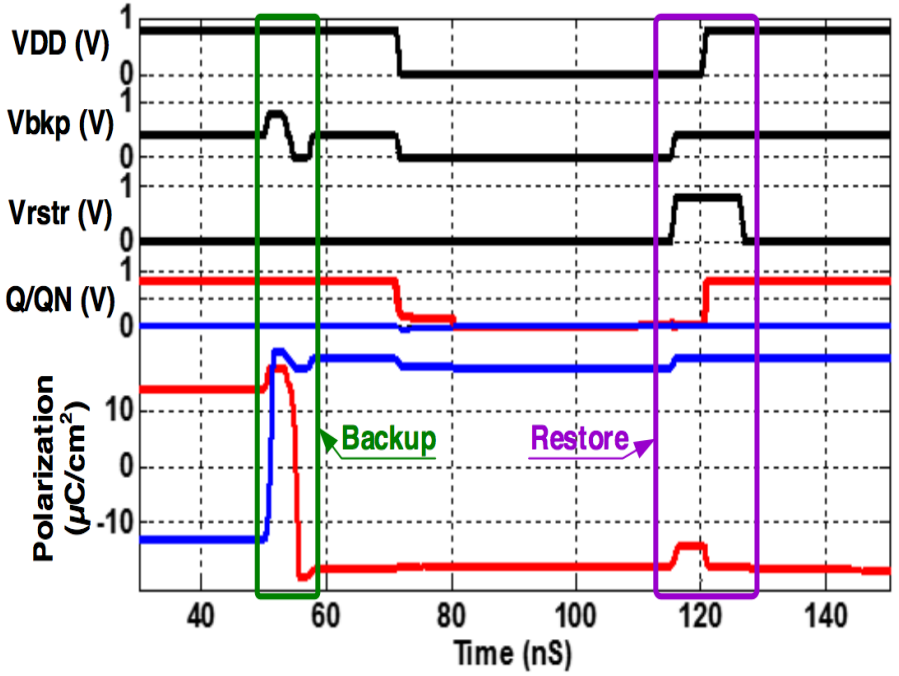


Backup step-2:
N1 → Neg. state

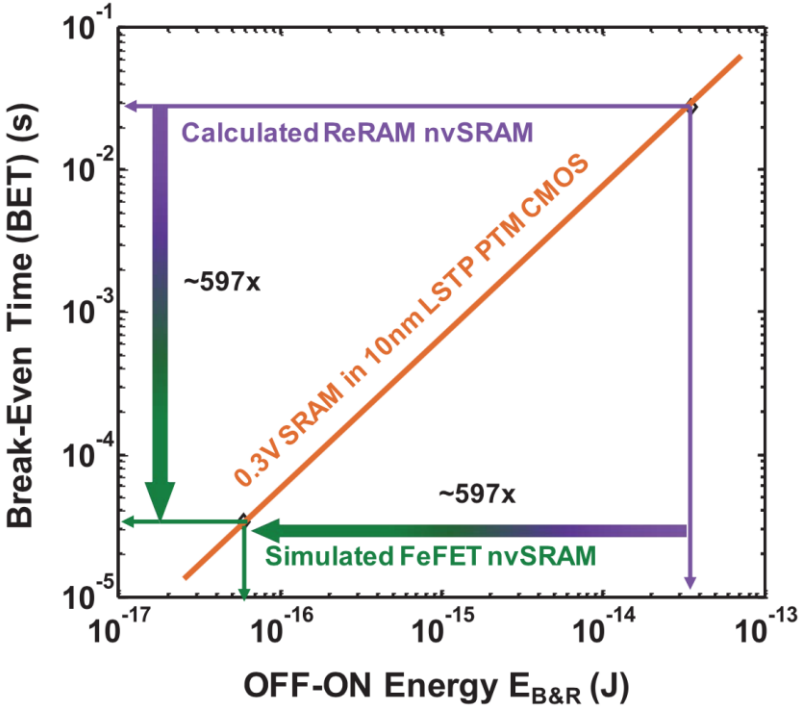
Xueqing Li *et al*, IEEE TED, July 2017; Patent filed

nvSRAM: Enabling Nonvolatile Computing with FeFET NVM

Transient Waveforms



Performance Evaluation



NO-DC-Current operation → ~600x $E_{B\&R}$ and Break-even time savings!
Enabling finer-grain power-gating with significantly lowered BET!

Break-even time (BET) is the maximum standby time of a volatile SRAM if an equal amount $E_{B\&R}$ is provided to sustain leakage.

Xueqing Li *et al*, IEEE TED, July 2017; Patent filed

Symmetric FeFET NVM: Flexible Data Analytics

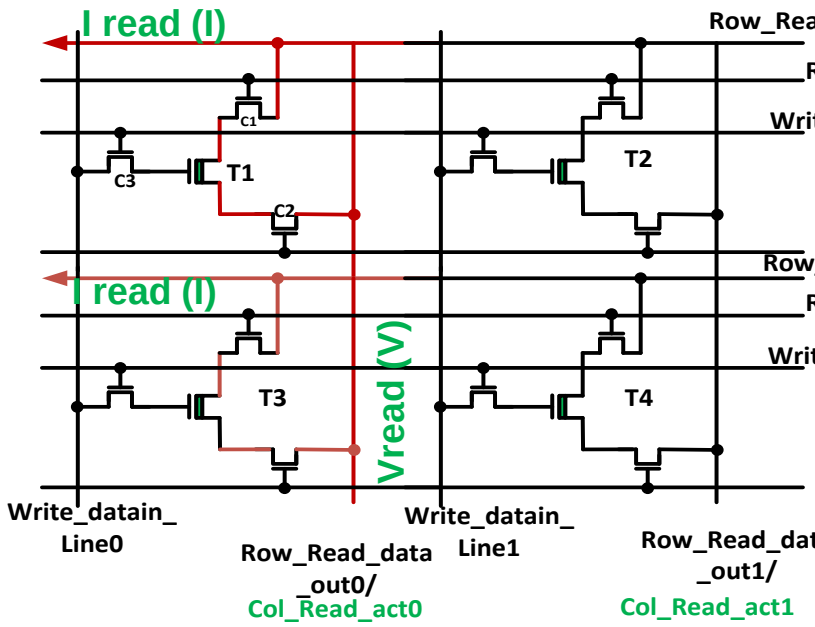


Fig.(a) Column Data read : T1 and T3

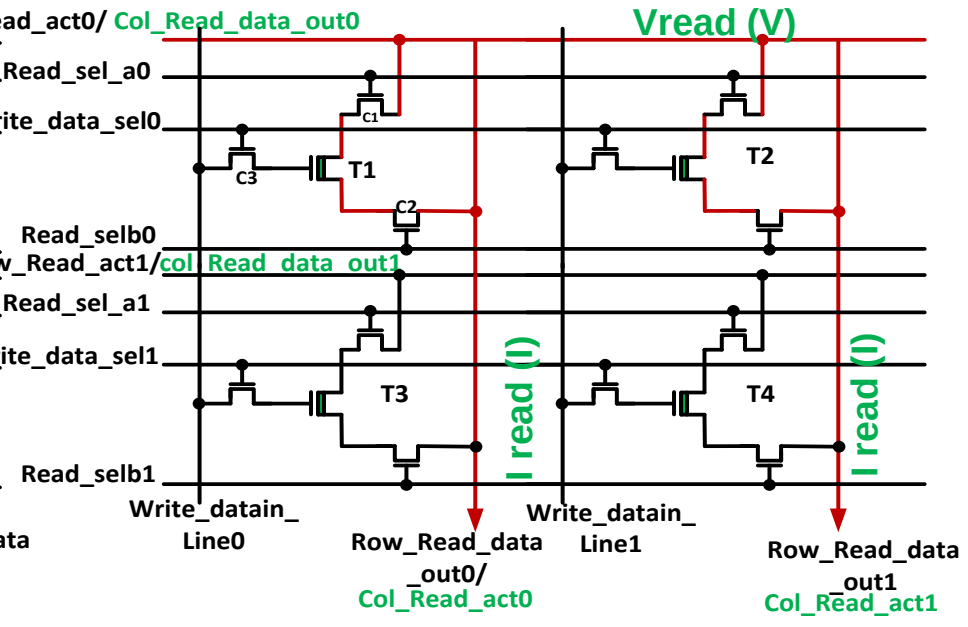
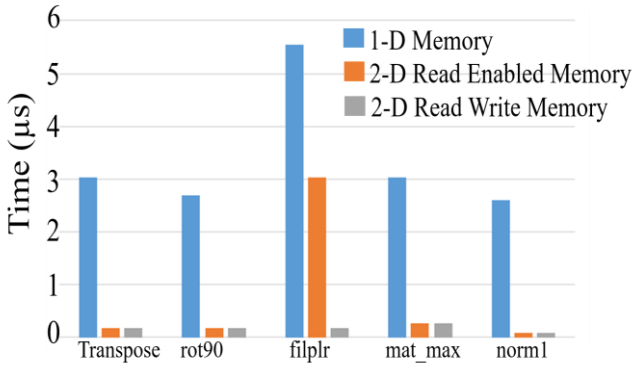
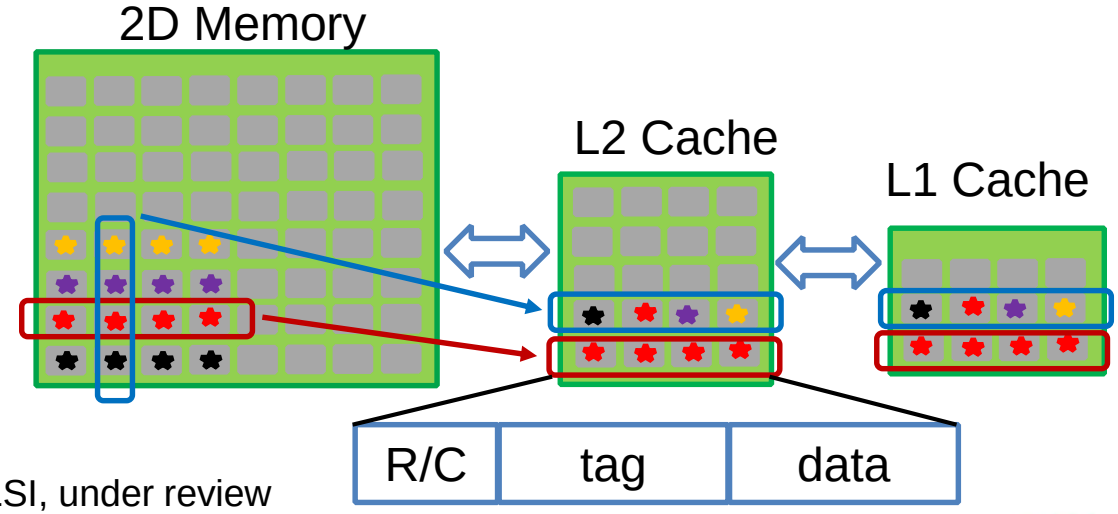


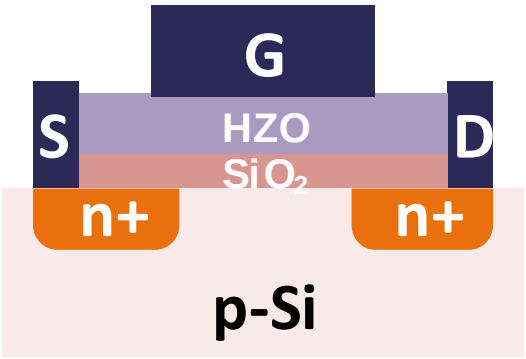
Fig.(b) Row Data read: T1 and T2

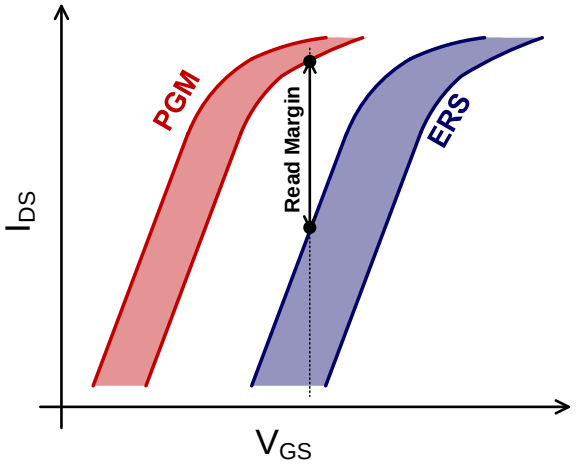


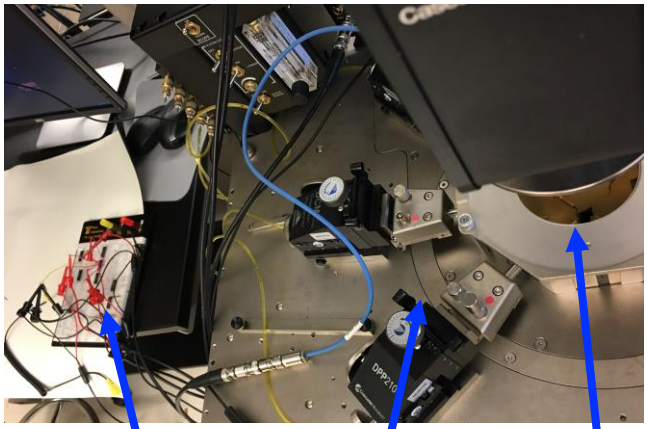


Sumitha George, Xueqing Li *et al*, IEEE TVLSI, under review

Recent Progress: Experimental FeFET NVM Circuit

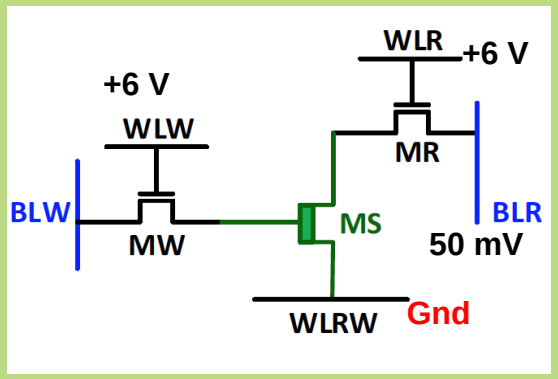




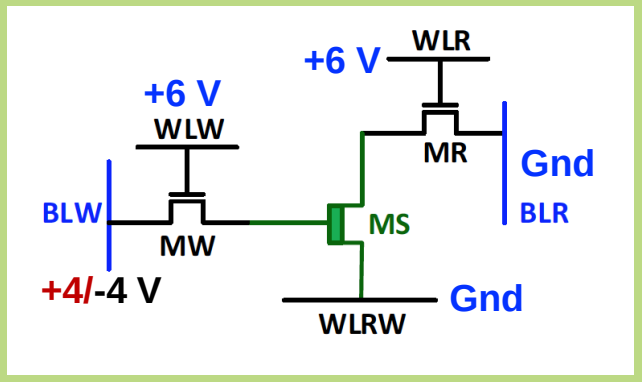


- 10 nm HZO/0.8 nm SiO_2/p-Si gate stack
- 40 μm / 2 μm with 2 μm gate overlap with S/D

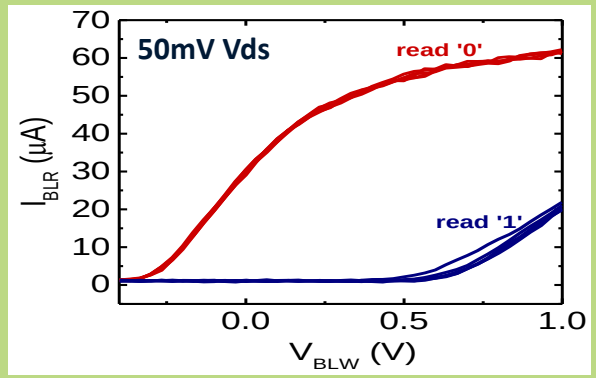
Read '0'/'1'



Write '0'/'1'



Read Results



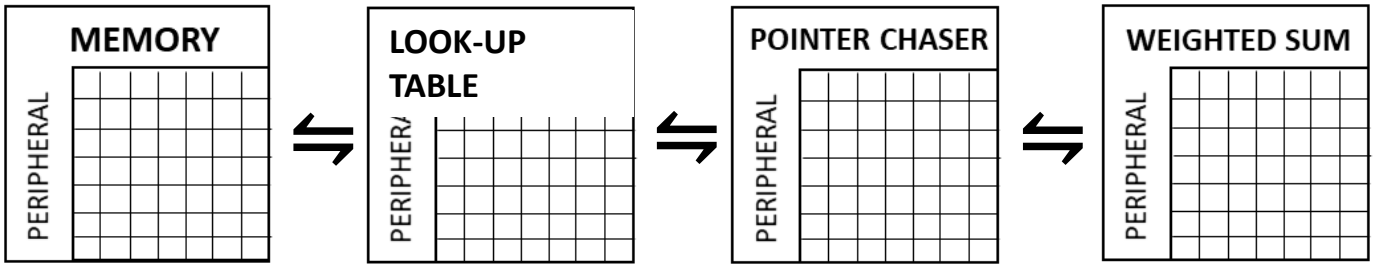
Ongoing collaboration work with Prof. Datta

Overview

- Monolithic 3D Integration
- Configurable Memory-Logic Device
- Cross point arrays

Cross-Point Peripheral Reconfigurability: Circuit-Architecture co-design

Each X-point array Reconfigurable as:



(MEM) (LUT) (PC) (WS)

Multiple Arrays To Form Programmable Unified System

MEM	WS	LUT
LUT	MEM	LUT
PC	LUT	MEM

To Accelerate

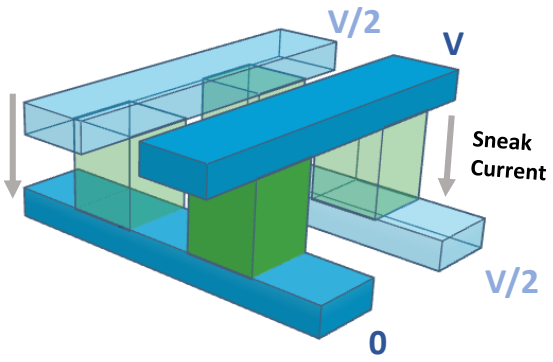
DATA-INTENSIVE APPLICATIONS

Security | Bioinformatics | Graph Analytics | Machine Learning

T. Leslie, "Micron's Automata Processor" in *Beyond CMOS HPC Workshop* (2016)

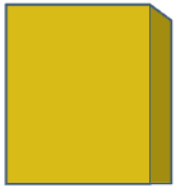
Study of selector for X-point: cross-point memories require selectors to eliminate current sneak path

Memory (only)



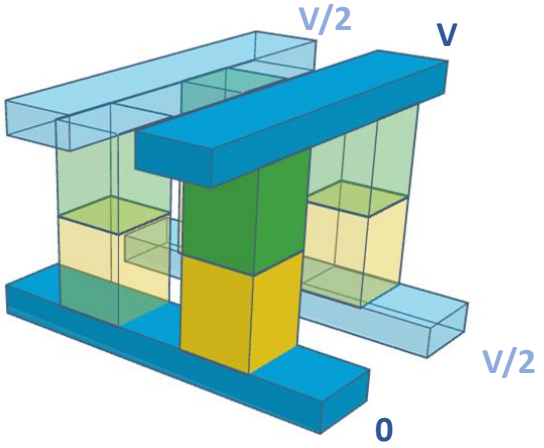
+

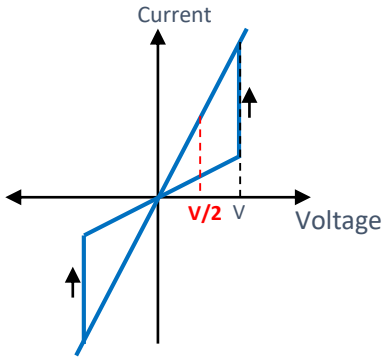
Selector



=

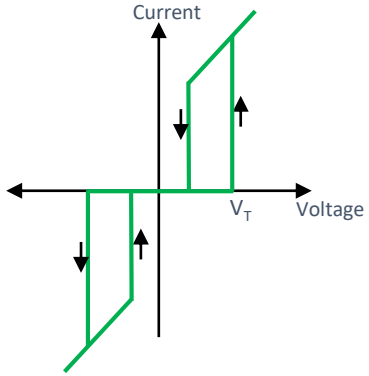
Memory + Selector





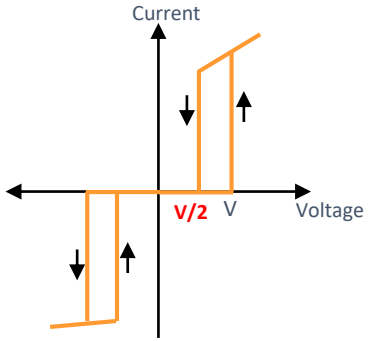
Significant sneak current through half accessed cells

+



Reduce sneak current exploiting non-linearity of selector

=



Summary

- New generation of design automation tools that drive the innovation in devices
- New computational models will create additional search dimensions for exploration tools
- Models to support circuit-architecture explorations
- Software design stack and protocols co-designed with device fabrics compound benefits
- Leaves room for several interesting questions – quest in many other countries in large funded projects
- Endurance, feature/voltage scaling, additional features: integrated sensing-compute structures, relative progress of other competing memory technologies.