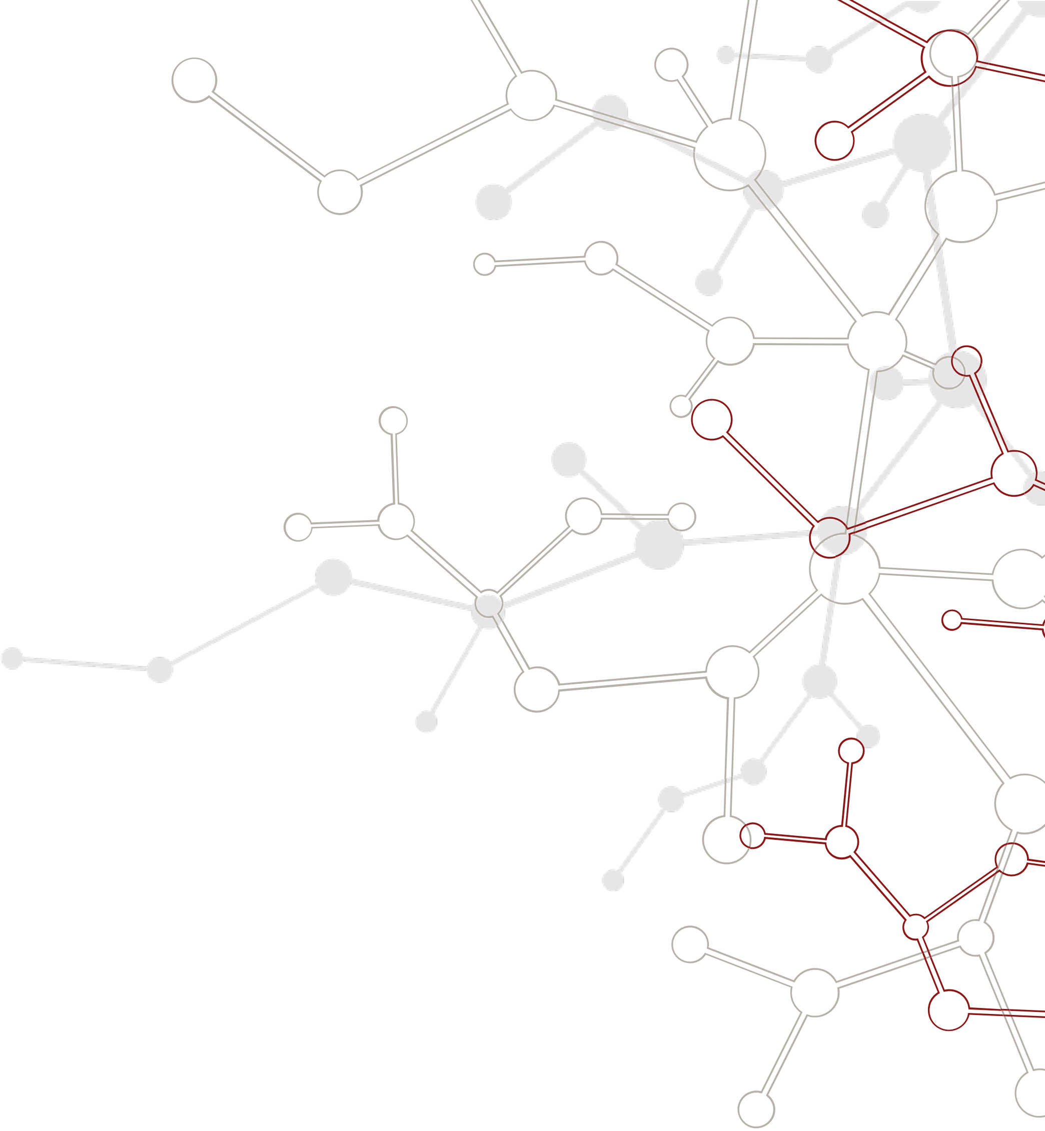
# Materials and Devices Working Group Highlight: Ferroelectrics



May 16, 2023

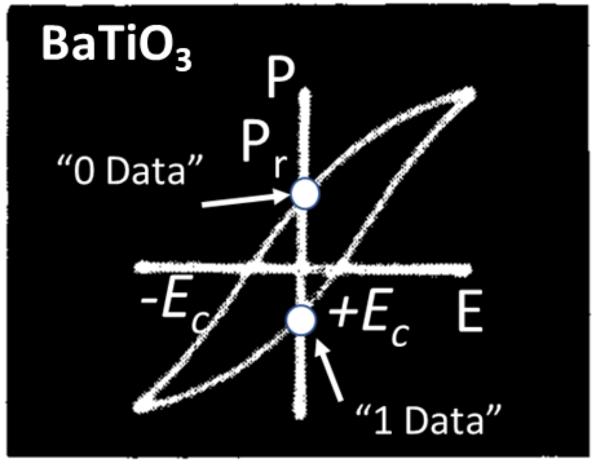
John D. Baniecki

SLAC National Accelerator Laboratory

# Outline

- □ Perovskite vs fluorites (each with challenges)
- □ Ferroelectric memories: FRAM, FeFET, FTJs
- □ Comparison of memory technologies
- □ Thermal processing for BEOL integration
- □ Compute-in-memory (CIM) accelerators

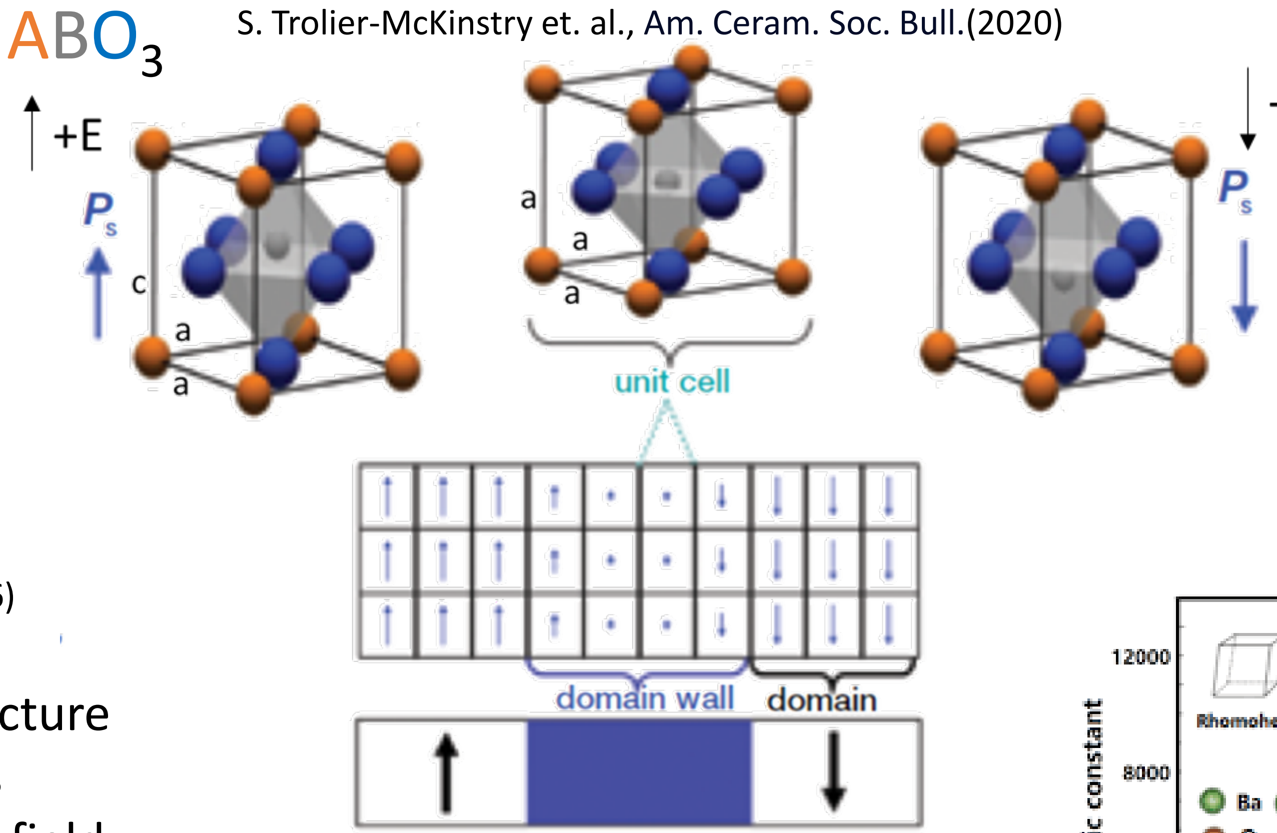
# Traditional: Perovskite-based (status quo for ~ 60 years)

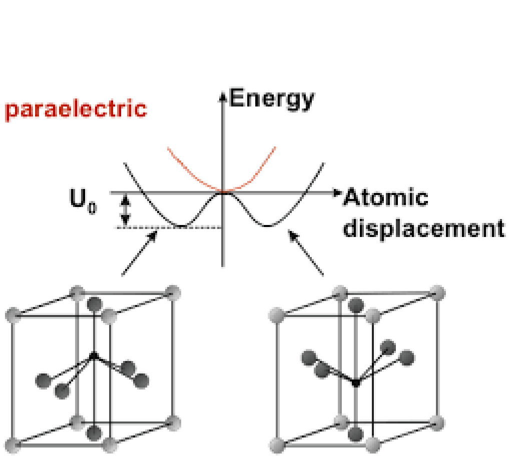


Oscillographic trace

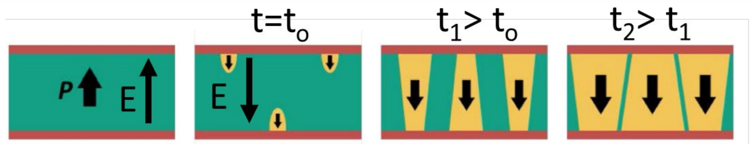
A. Von Hippel et. al., Eng. Chem. (1946)

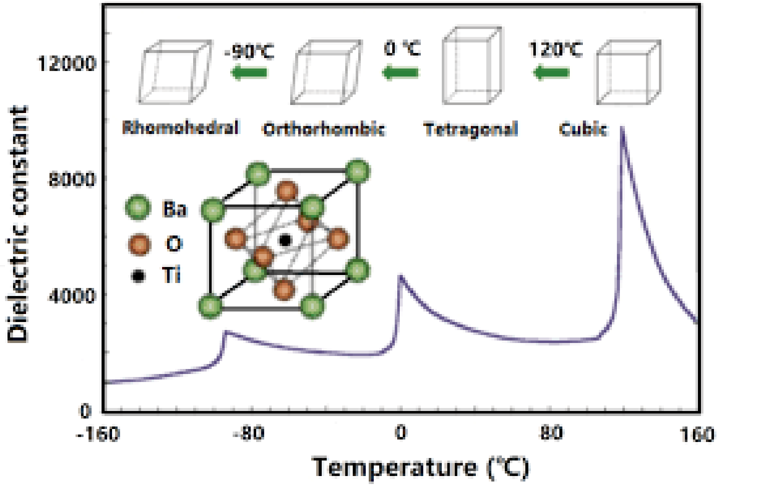
S. Trolier-McKinstry et. al., Am. Ceram. Soc. Bull.(2020)





- Non-centrosymmetric structure
- 2 stable polarization states switchable by an applied E field
- Difference in remnant polarization ($P_r$) through domain states





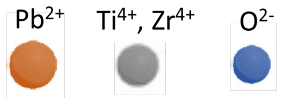
J. Su et. al., J Mater Sci: Mater Electron, (2019)

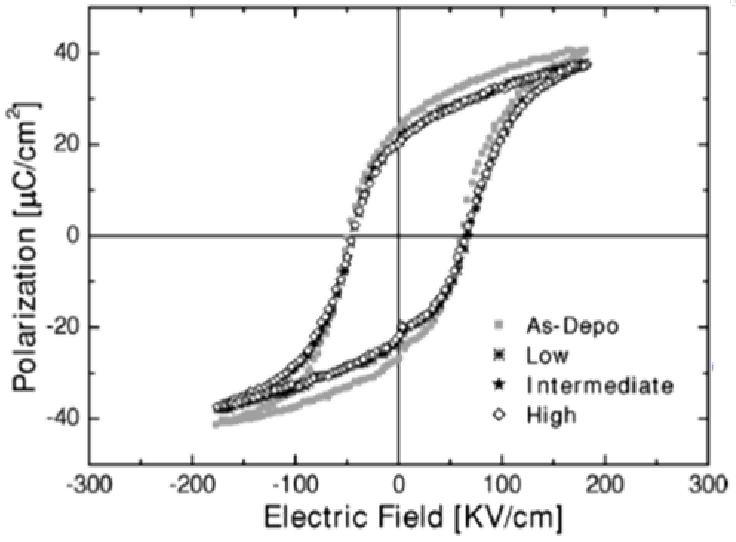
# Traditional: Perovskite-based (status quo for ~ 60 years)

$PbTiO_3$-$PbZrO_3$(PZT)

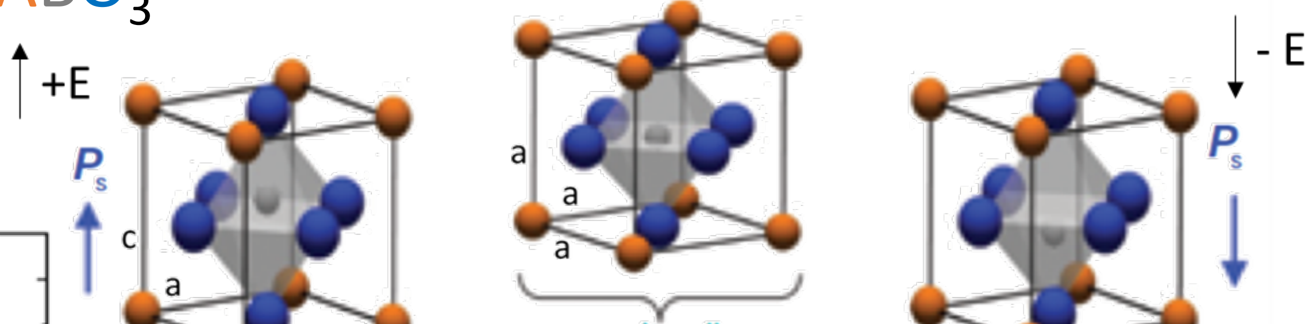
$ABO_3$

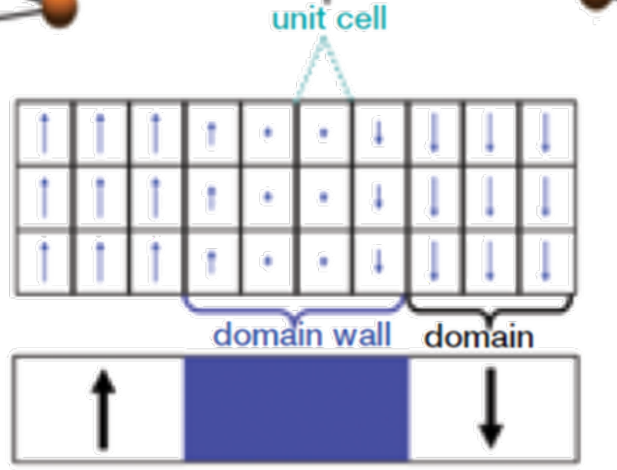
S. Trolier-McKinstry et. al., Am. Ceram. Soc. Bull.(2020)

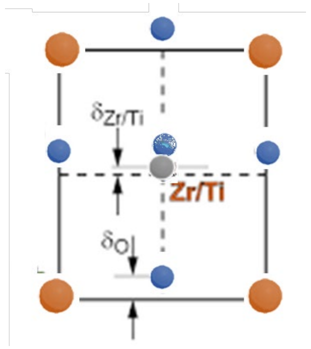




S.H. Choi et. al., Integr. Ferroelectr. (2006)







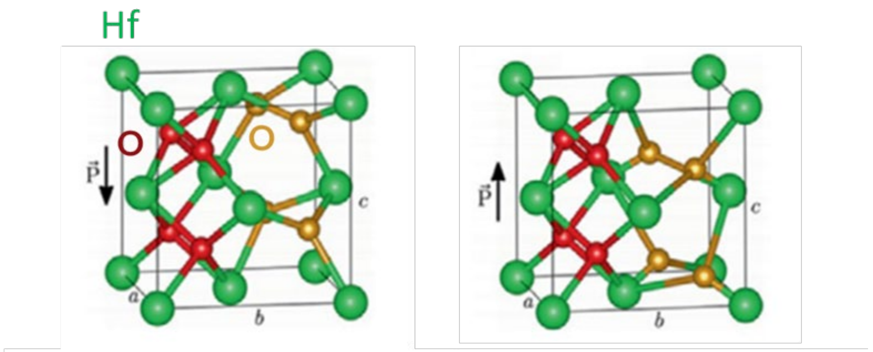
- Thick (> 70nm)
- Limited scaling (130 nm node)
- Low coercive field ($E_c$) ~100kV/cm
- Limited retention
- Not CMOS compatible

❑ PZT-based 1T1C memory in production for > 20 years FeRAM (Fujitsu), FRAM (TI), F-RAM (Infineon)

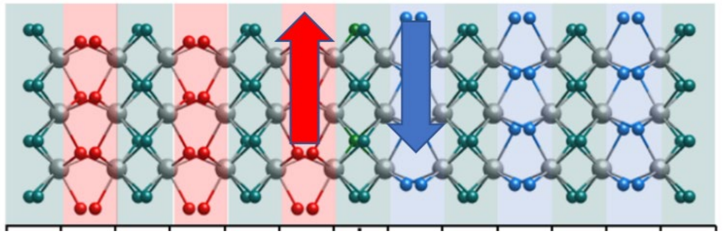
❑ Low density (4KB- 128 MB) niche applications ( IC card, robotic, automotive applications, ...)

# Fluorite-structured ferroelectrics (~2011)

*$HfO_2$-based*



Ferroelectric orthorhombic-III (*$Pca2_1$*)
(metastable phase)

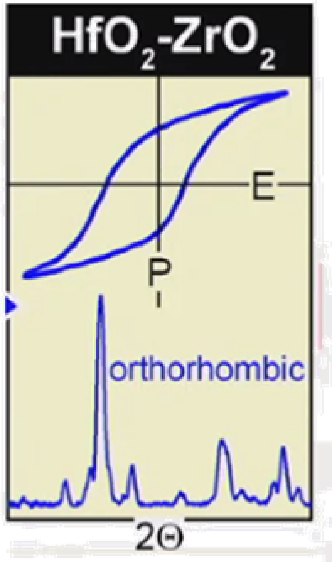


H.-J. Lee et al., Science (2020)

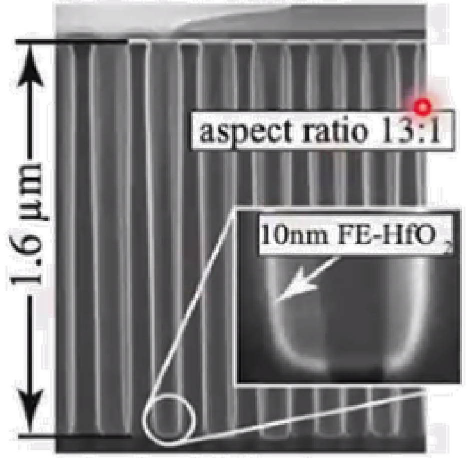
$HfO_2$ discovered to be ferroelectric in 2006 (Tim Böscke at Qimonda, formerly Infineon), published results in 2011

Böscke, T. S. et al., Appl. Phys. Lett., (2011)

*$HfO_2$-$ZrO_2$ Binary Alloys (HZO)*



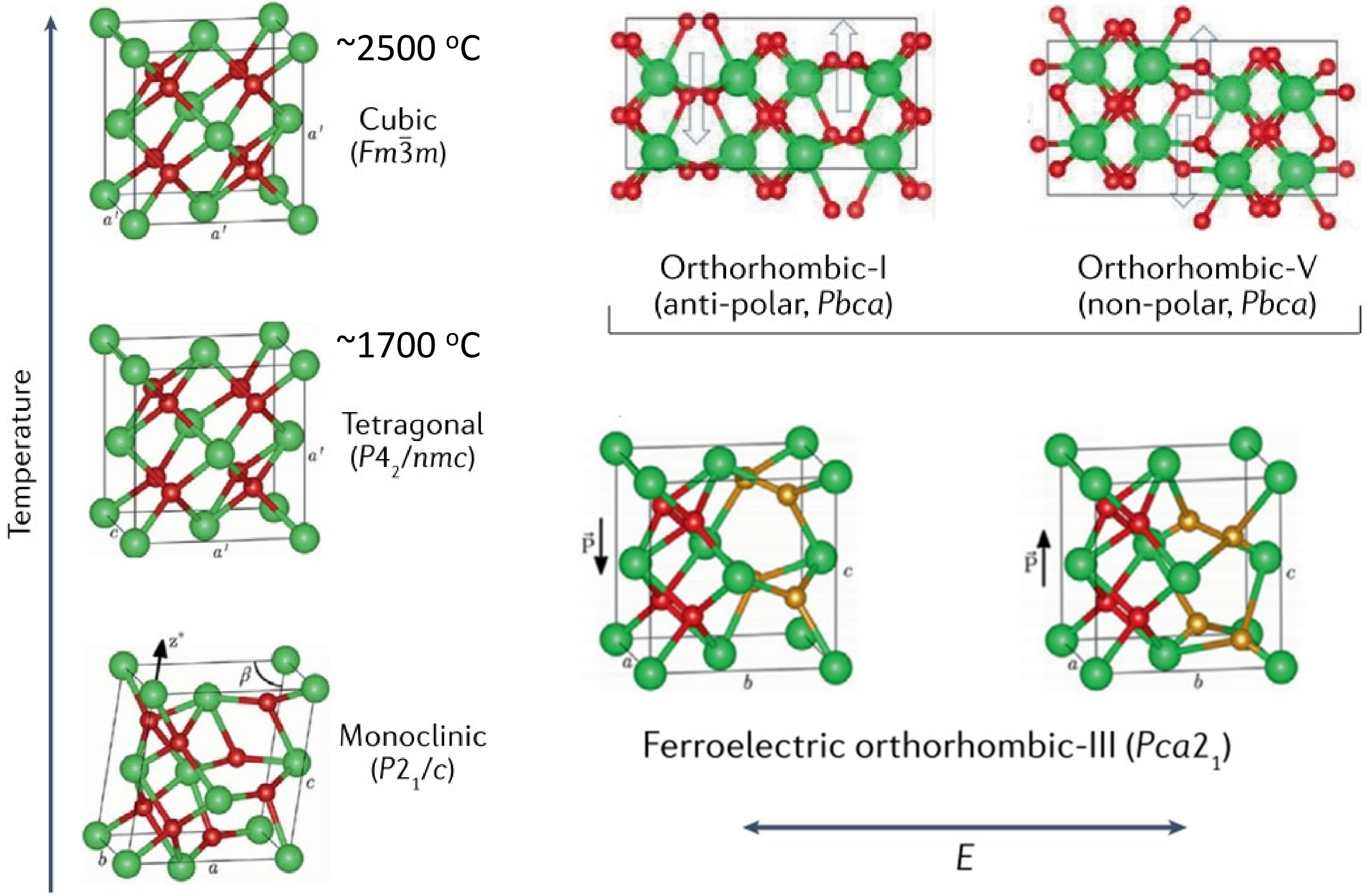
*Atomic-Layer-Deposition (ALD)*



- Thin (sub 3 nm)
- Fast switching (sub-ns)
- Scaling (22 nm & beyond)
- High $E_c$ (~1M V/cm)

- Good retention
- CMOS compatible
- ALD growth

# Plethora of phases present in $HfO_2$ system



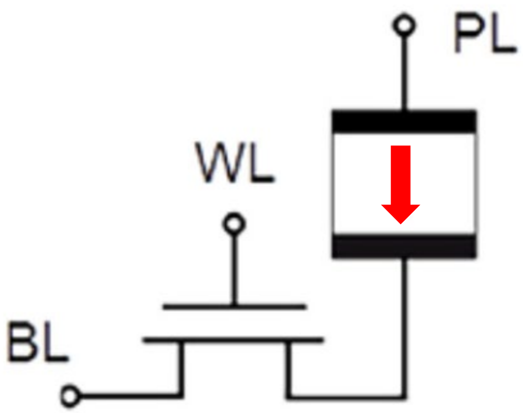
Schroeder, U. *et al.*, *Nat Rev Mater*. (2022).

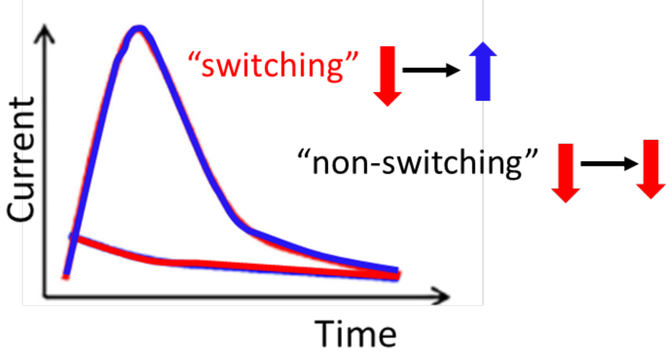
- Monoclinic phase is the room-temperature bulk stable phase.
- Phases separated by energies of 10s of meV
- Rapid heating and cooling with capping layers (e.g. TiN) stabilizes ferroelectric phase
- Dopants ( Zr, Al, Gd, La, Si, Sr, and Y )
- Field induced transformations
- Reliability issues

# Ferroelectric memories: Operation principles

## FeRAM (1T1C)

Ferroelectric Random-Access Memory
“DRAM-like”

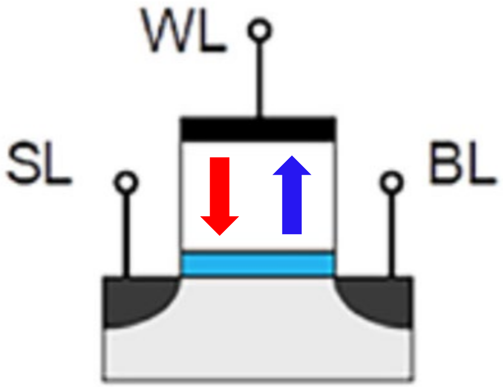


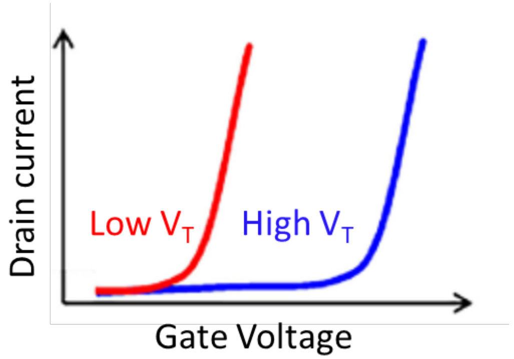


- Commercialized for > 20 years
- Destructive read
- High endurance > $10^{15}$

## FeFET(1T)

Ferroelectric Field Effect Transistor
“FLASH-like”

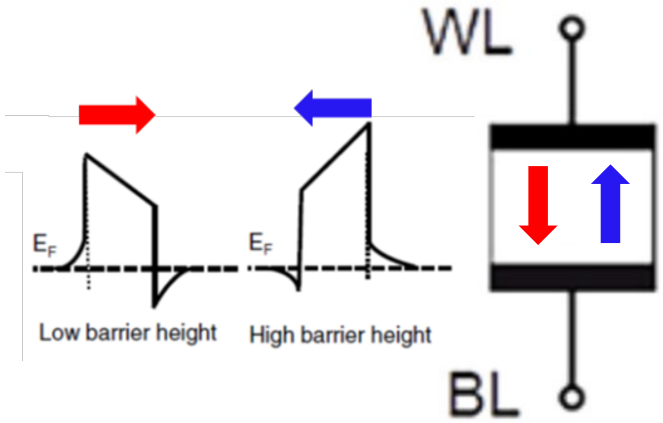


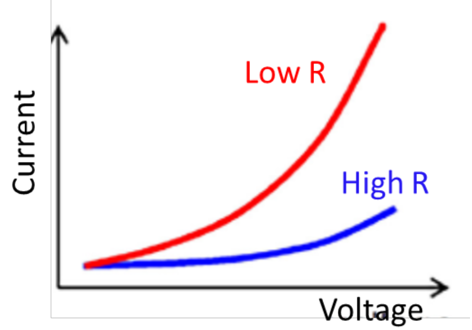


- Intensive R&D by Semiconductor Industry
- Nondestructive read, multiple bits
- High endurance  challenging

## FTJ(1R)

Ferroelectric Tunnel Junction
“Diode-like”

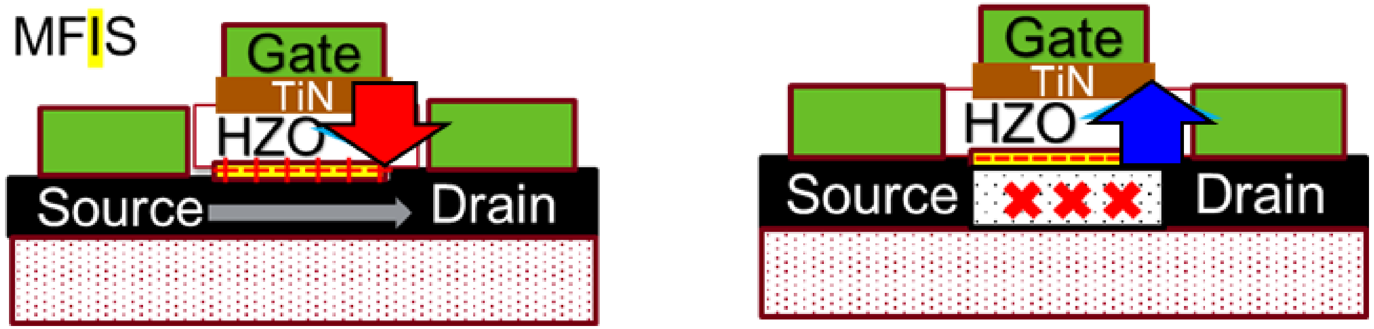




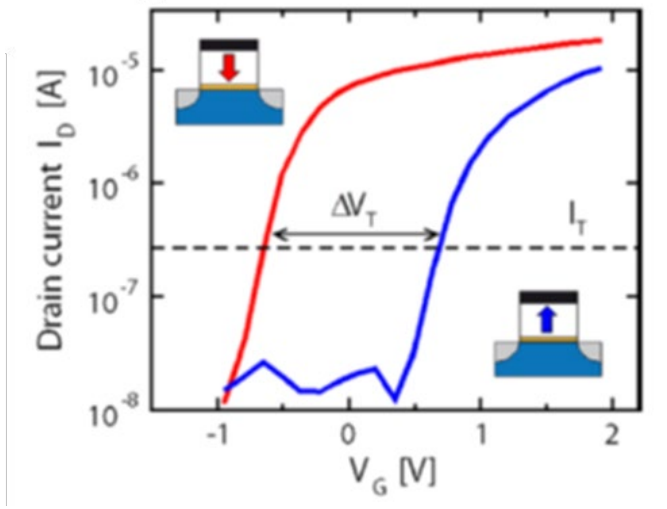
- Academia R&D
- Asymmetric free carrier screening lengths

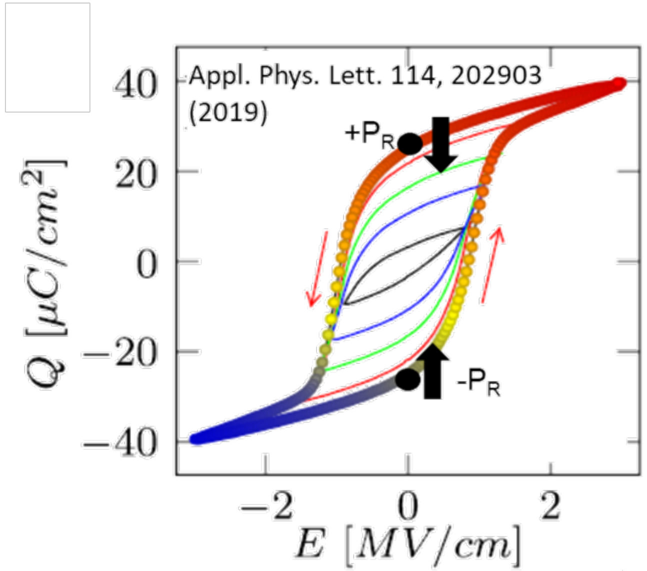
# Multi-state $HfO_2$-based FeFET memories

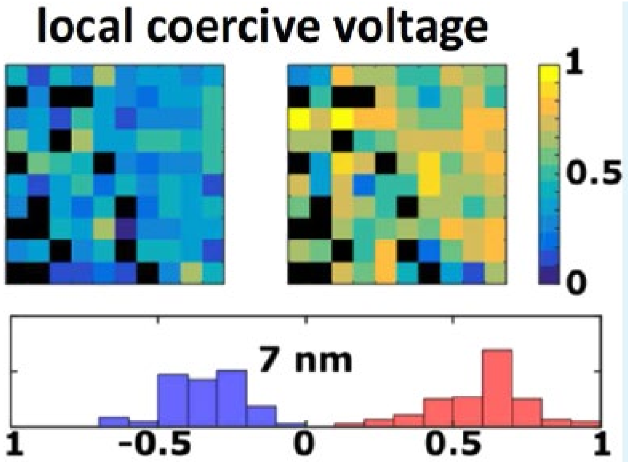
**$Hf_{0.5}Zr_{0.5}O_2$(HZO)** Ferroelectric Field Effect Transistor (FeFET)

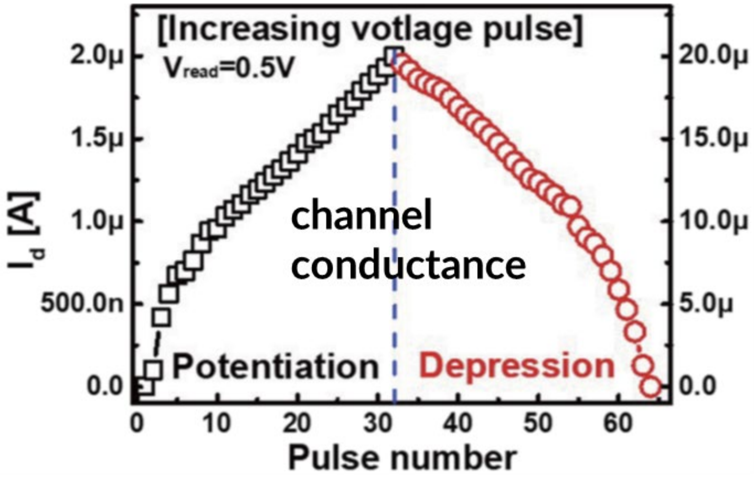


ACS Appl. Mater. Interfaces 2017, 9, 3792−3798




Appl. Phys. Lett. 114, 202903 (2019)

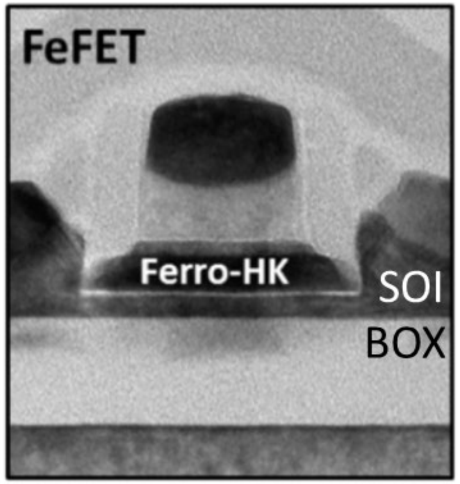


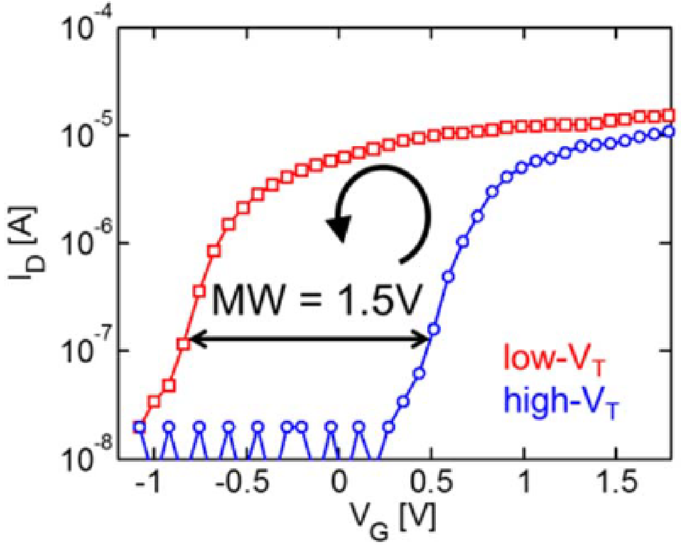


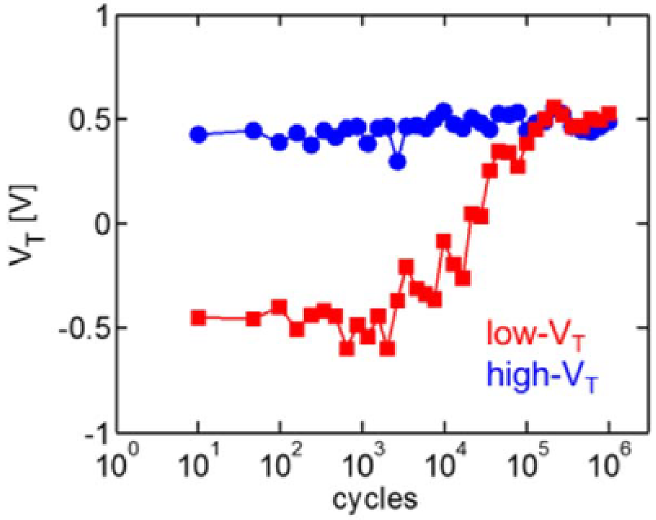
I Stolichnov et. al., ACS Appl. Mater. Interfaces (2018)

# Scaled $HfO_2$-based FeFETs

S. Dünkel , IEDM 2017, 22 nm FDSOI CMOS



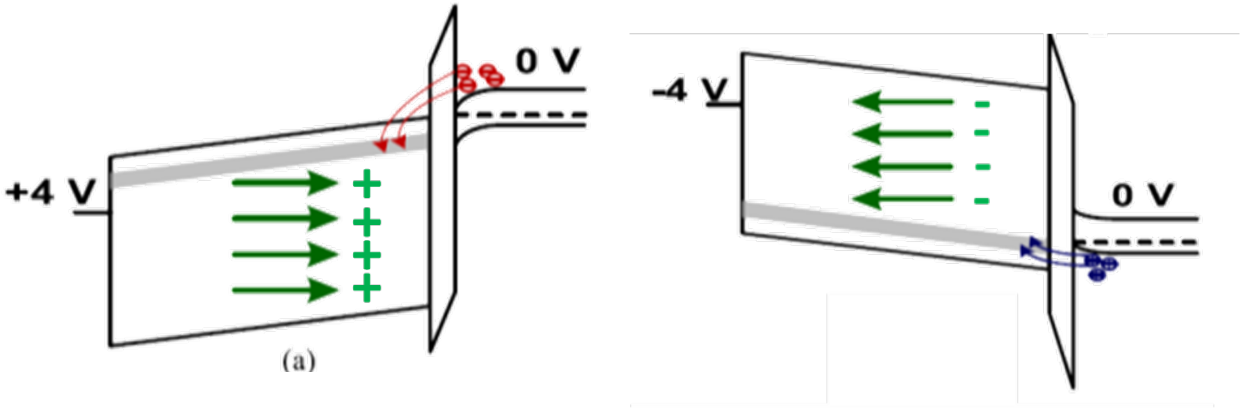




❑ Demonstrated at the 22/28 nm node

- MW of 1.5 V
- Scaled FeFET cells 0.025 µm²
- Endurance cycles up to $10^5$

❑ 3D NAND, GAA structures



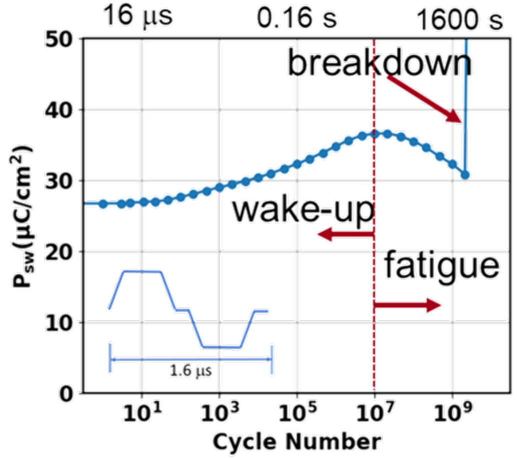
K. Ni et. Al., T-ED, 2018

K. Florent et. al., VLSI 2017, S-Y Lee et. al., JEDS 2021

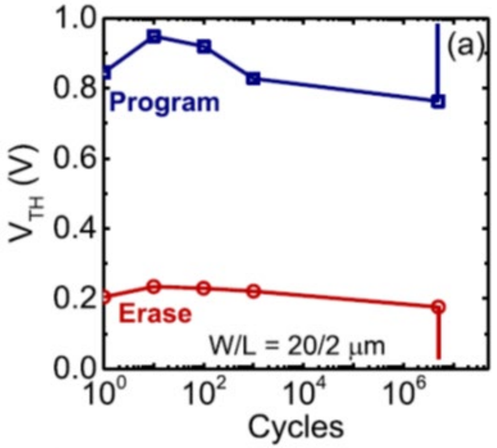
# Challenges of ferroelectric memories

## 1. Polarization variation/Endurance

MFM (> $10^{12}$ cycles)

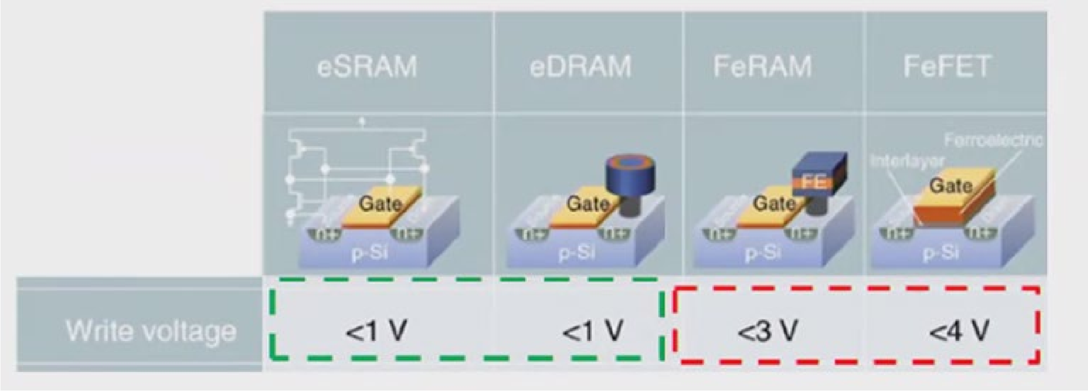


MFIS (<$10^9$ cycles)



## 2. High Write Voltage

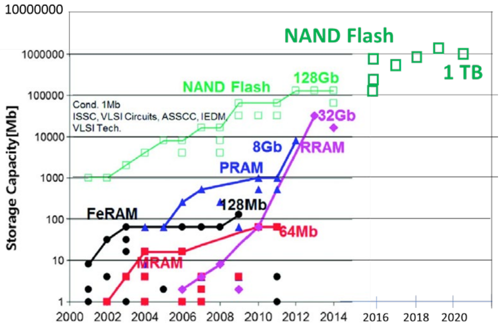
- Reduce write voltage to logic compatible level



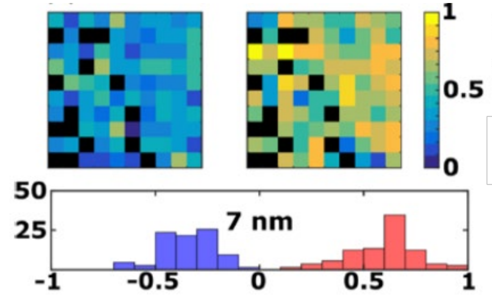
|  | eSRAM | eDRAM | FeRAM | FeFET |
|---|---|---|---|---|
| Write voltage | <1 V | <1 V | <3 V | <4 V |

## 3. Scaling and Density

Storage capacity

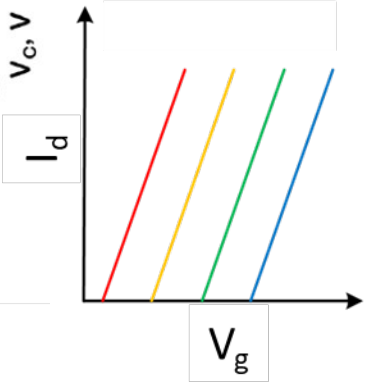


Stochasticity



TiN/HZO/TiN, 1x1 μm²

Multi-bit per cell

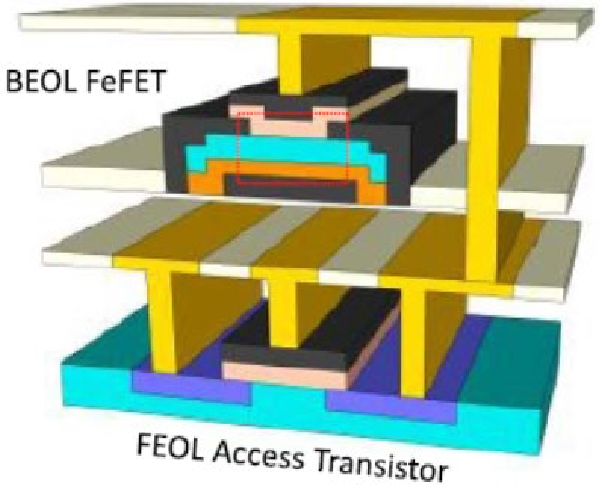


ACS Appl. Mater. Interfaces 2018, H. Mulaosmanovic et al, EDL 2018

## 4. 3D Integration

- Reduced latency, energy consumption
- BEOL process temperature < 400 °C

*BEOL: Back-end-of-line*
*FEOL: Font-end-of-line*

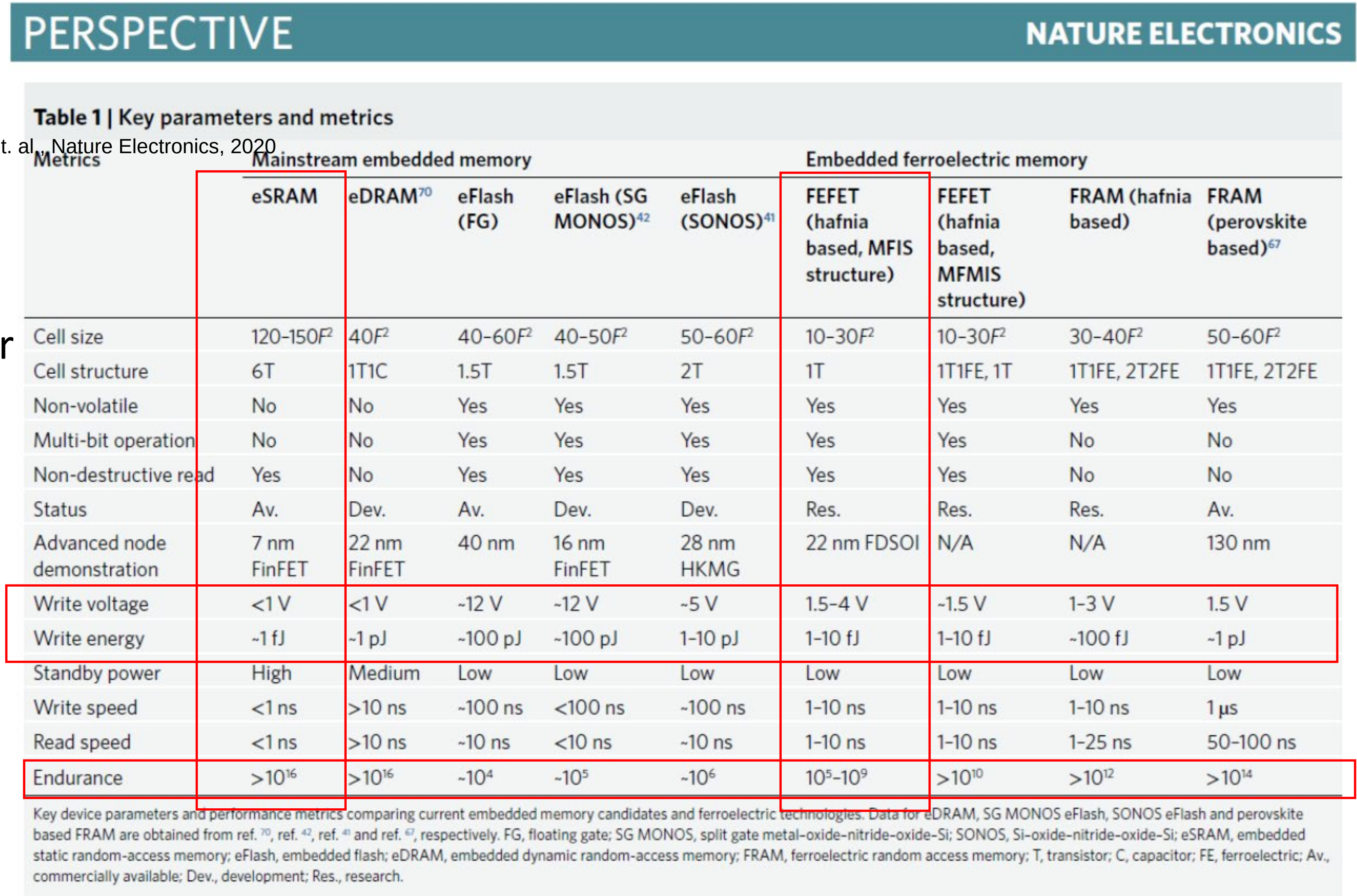


# Comparison of memory technologies

A. I. Khan et. al., Nature Electronics, 2020

- ❑ Impact of tech node on energy:
  2Pr ~ 60 $\mu$C/cm $^2$ , 1.5 V
  - 60F$^2$, 130 nm (FRAM perov.) ~ 913 fJ
  - 30F$^2$, 22 nm FeFET ~ 13 fJ
  - Hypothetical 30F$^2$, 5 nm ~ 4 fJ
- ❑ Pathways to attojoule switching not clear
- ❑ Advantages over eSRAM in energy efficiency will depend on technology node, compute to standby ratio (application specific)
- ❑ Need to reduce voltage, improve endurance

A. I. Khan et. al., Nature Electronics, 2020



PERSPECTIVE

NATURE ELECTRONICS

**Table 1 | Key parameters and metrics**

| Metrics | Mainstream embedded memory | | | | | Embedded ferroelectric memory | | | |
|---|---|---|---|---|---|---|---|---|---|
| | eSRAM | eDRAM[70] | eFlash (FG) | eFlash (SG MONOS)[42] | eFlash (SONOS)[41] | FEFET (hafnia based, MFIS structure) | FEFET (hafnia based, MFMIS structure) | FRAM (hafnia based) | FRAM (perovskite based)[67] |
| Cell size | 120–150$F^2$ | 40$F^2$ | 40–60$F^2$ | 40–50$F^2$ | 50–60$F^2$ | 10–30$F^2$ | 10–30$F^2$ | 30–40$F^2$ | 50–60$F^2$ |
| Cell structure | 6T | 1T1C | 1.5T | 1.5T | 2T | 1T | 1T1FE, 1T | 1T1FE, 2T2FE | 1T1FE, 2T2FE |
| Non-volatile | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Multi-bit operation | No | No | Yes | Yes | Yes | Yes | Yes | No | No |
| Non-destructive read | Yes | No | Yes | Yes | Yes | Yes | Yes | No | No |
| Status | Av. | Dev. | Av. | Dev. | Dev. | Res. | Res. | Res. | Av. |
| Advanced node demonstration | 7 nm FinFET | 22 nm FinFET | 40 nm | 16 nm FinFET | 28 nm HKMG | 22 nm FDSOI | N/A | N/A | 130 nm |
| Write voltage | <1 V | <1 V | ~12 V | ~12 V | ~5 V | 1.5–4 V | ~1.5 V | 1–3 V | 1.5 V |
| Write energy | ~1 fJ | ~1 pJ | ~100 pJ | ~100 pJ | 1–10 pJ | 1–10 fJ | 1–10 fJ | ~100 fJ | ~1 pJ |
| Standby power | High | Medium | Low | Low | Low | Low | Low | Low | Low |
| Write speed | <1 ns | >10 ns | ~100 ns | <100 ns | ~100 ns | 1–10 ns | 1–10 ns | 1–10 ns | 1 μs |
| Read speed | <1 ns | >10 ns | ~10 ns | <10 ns | ~10 ns | 1–10 ns | 1–10 ns | 1–25 ns | 50–100 ns |
| Endurance | >$10^{16}$ | >$10^{16}$ | ~$10^4$ | ~$10^5$ | ~$10^6$ | $10^5$–$10^9$ | >$10^{10}$ | >$10^{12}$ | >$10^{14}$ |

Key device parameters and performance metrics comparing current embedded memory candidates and ferroelectric technologies. Data for eDRAM, SG MONOS eFlash, SONOS eFlash and perovskite based FRAM are obtained from ref. [70], ref. [42], ref. [41] and ref. [67], respectively. FG, floating gate; SG MONOS, split gate metal-oxide-nitride-oxide-Si; SONOS, Si-oxide-nitride-oxide-Si; eSRAM, embedded static random-access memory; eFlash, embedded flash; eDRAM, embedded dynamic random-access memory; FRAM, ferroelectric random access memory; T, transistor; C, capacitor; FE, ferroelectric; Av., commercially available; Dev., development; Res., research.

# Back-end-of-line (BEOL) flash thermal processing

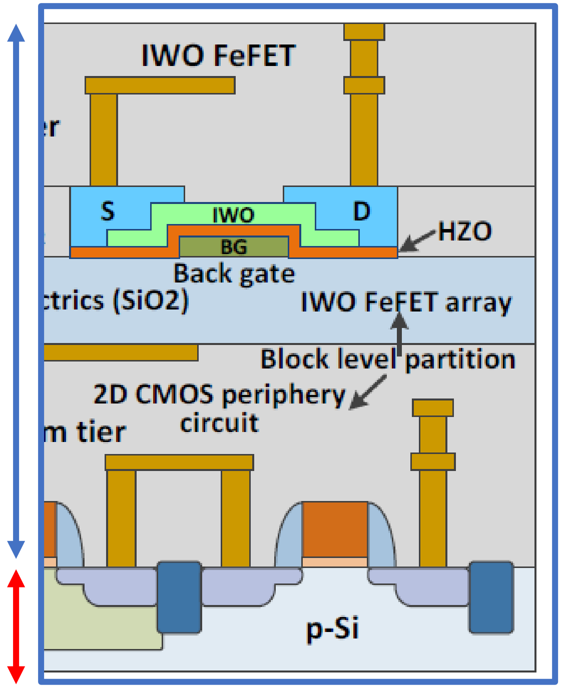
BEOL memory nodes

< 400 °C

Inter-tier dielectrics $SiO_2$

FEOL logic nodes

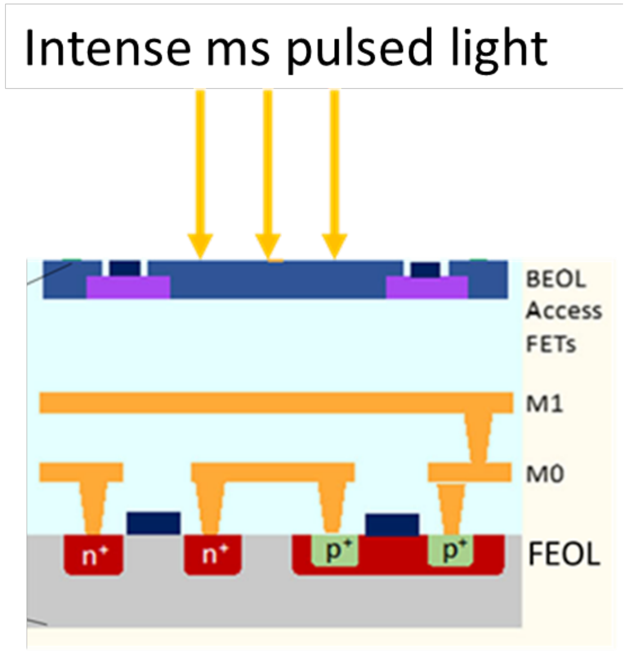
~ 1000 °C

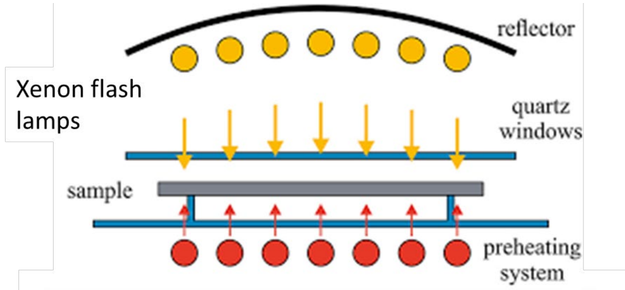


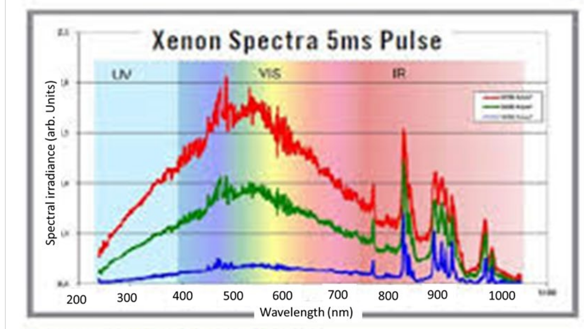
- Embedded memory in BEOL stack
  - Can bring memory and logic closer by stacking memory on top of processing nodes
  - Decrease energy consumption and latency

Dutta, S. *et al., IEEE Electron Device Lett.* **43**, 382–385 (2022).

How not to damage BEOL components?



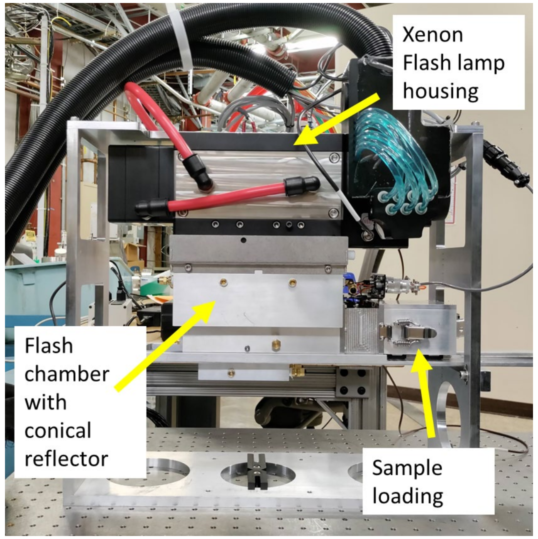


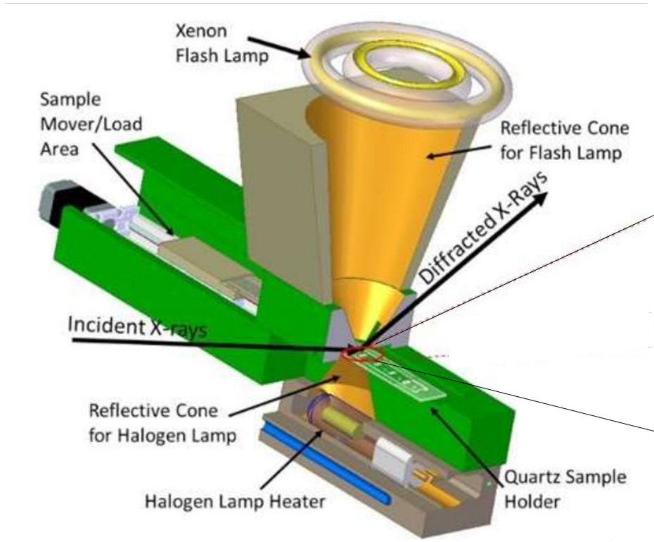


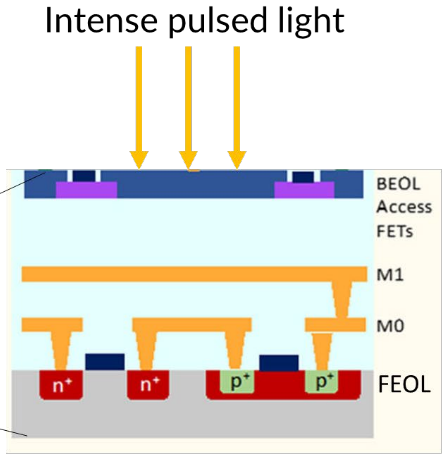
Xenon output spectrum

- ❑ Confine thermal transients to upper layers
- ❑ Ultra-fast thermal treatments

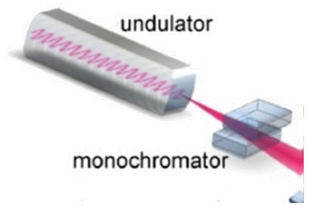
# Back-end-of-line (BEOL) flash thermal processing



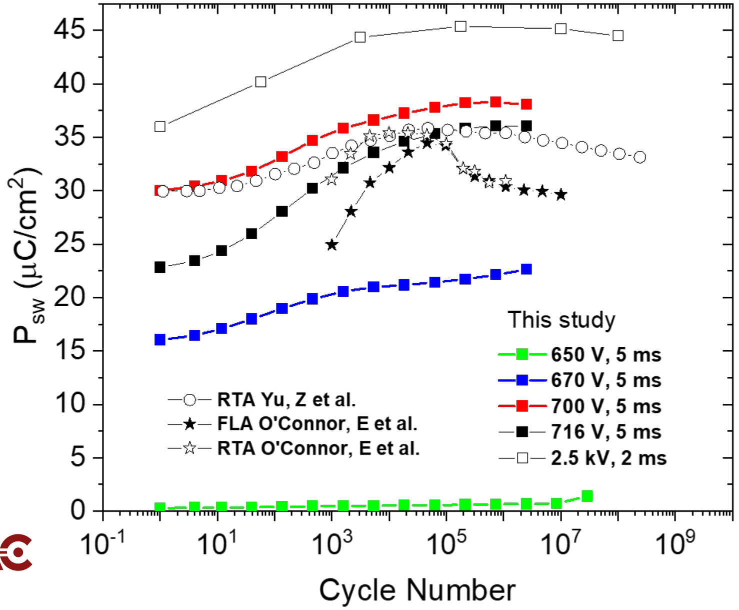


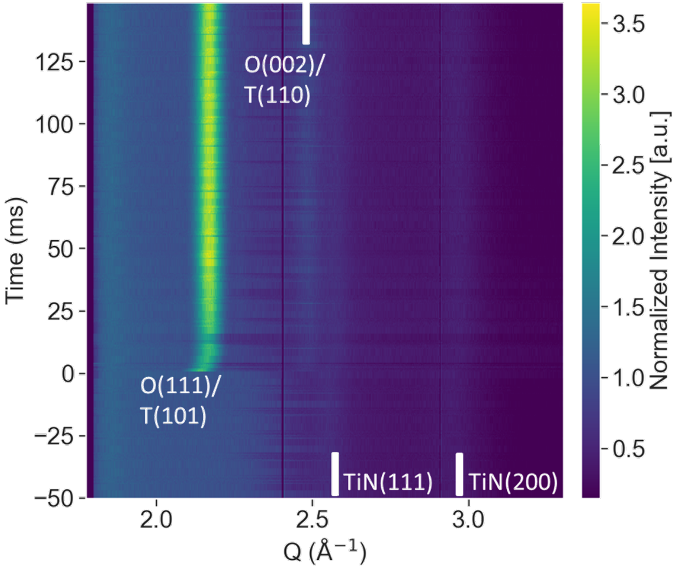


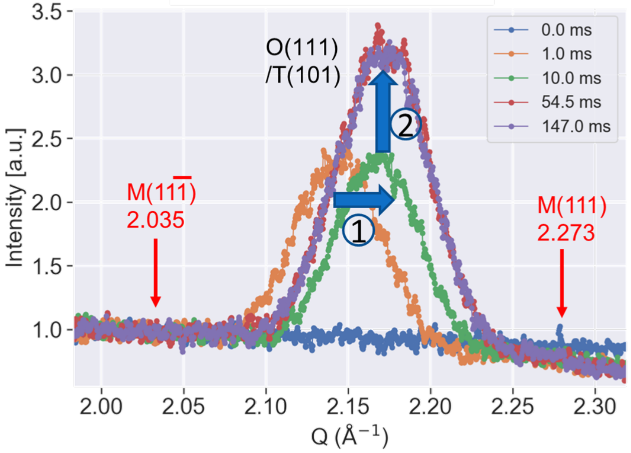
High Brightness In-Vacuum Undulator Beamline



- Static & time resolved studies
- Collect thousands of diffraction patterns during a flash time-temperature sequence
- See transformation in real-time





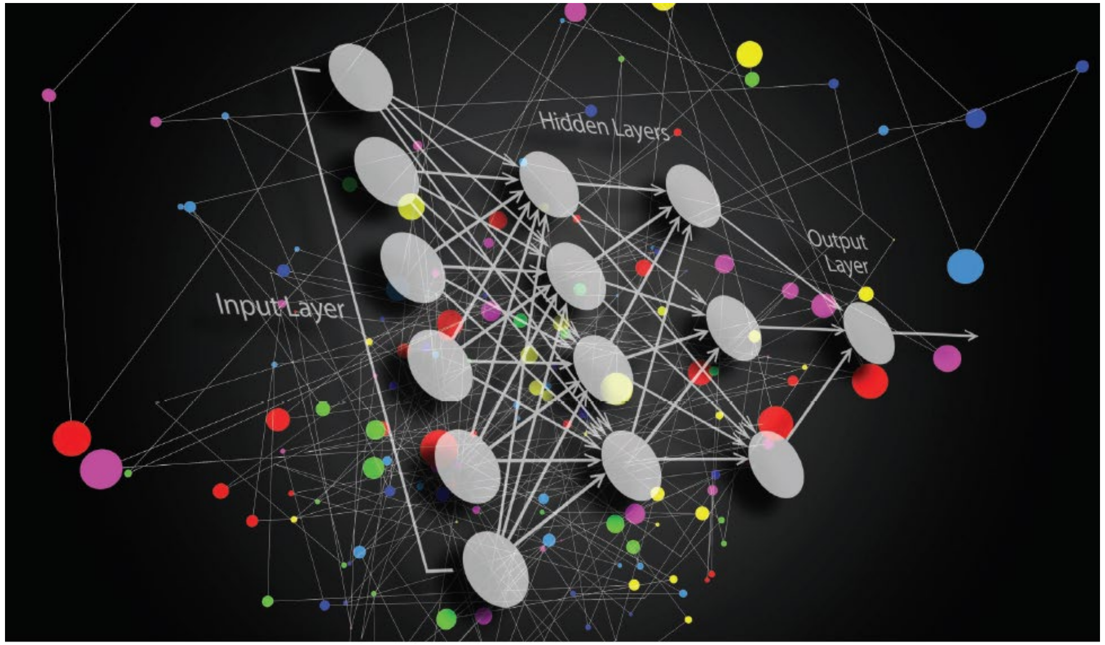


SLAC

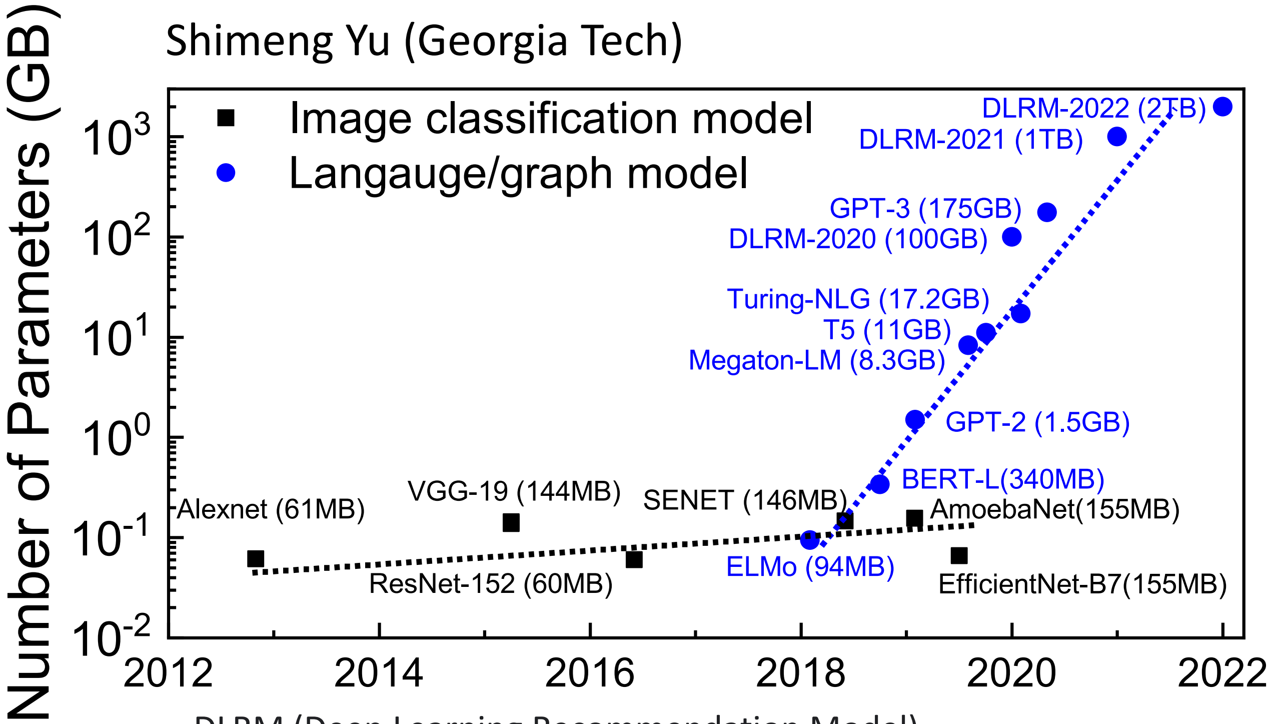
U.S. DEPARTMENT OF ENERGY | Energy Efficiency & Renewable Energy
ADVANCED MANUFACTURING OFFICE

# Compute-in-memory accelerators

## The Memory Demand of Modern AI Models



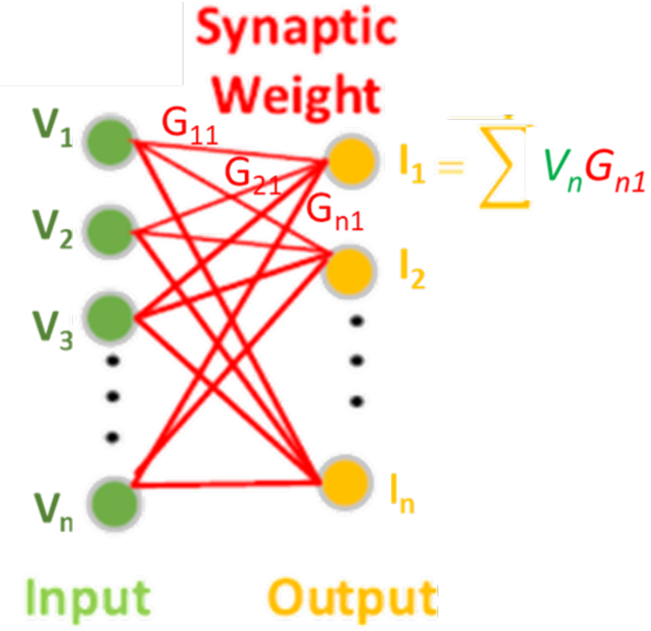
Shimeng Yu (Georgia Tech)



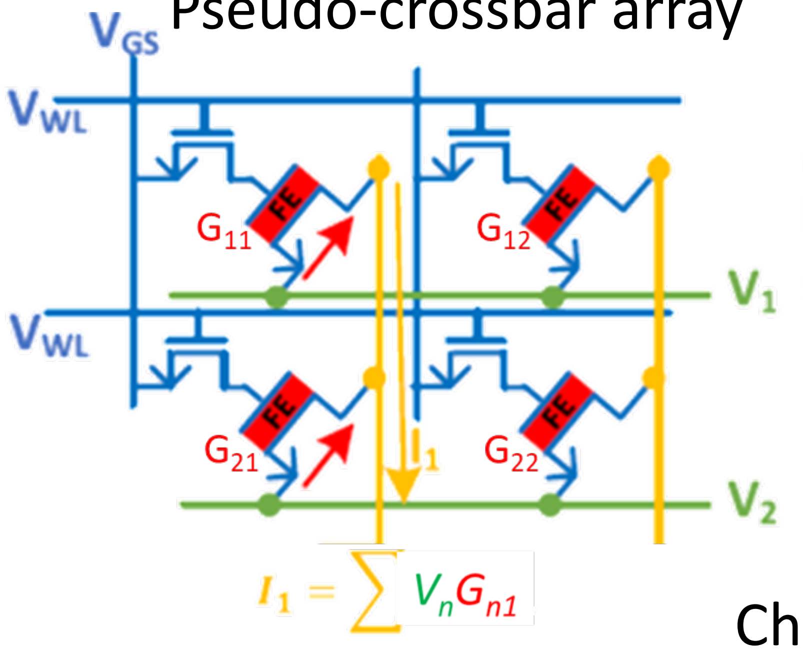
DLRM (Deep Learning Recommendation Model)

- ❑ Modern AI models can have more than a TB of parameters
- ❑ With on-chip memory limited by SRAM size, there is an extraordinary volume of data traffic between processor and off-chip memory that adds to energy consumption  and latency.
- ❑ Compute-in-memory (CIM) is a promising approach to overcome memory bottleneck where compute is moved closer to the data residing in the memory

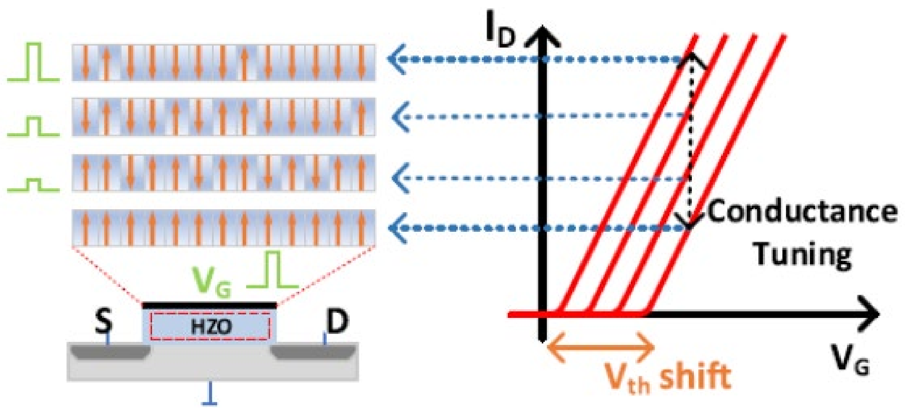
# FeFETS for compute-in-memory accelerators



Pseudo-crossbar array



HfO$_2$-based FeFETs



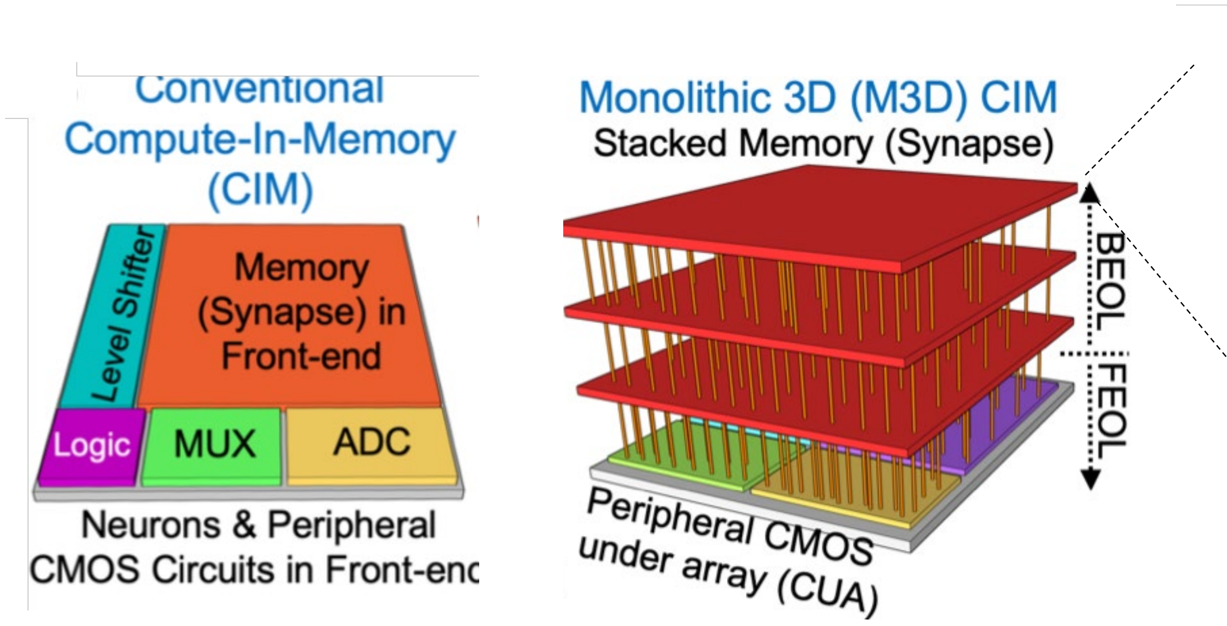
## Advantages

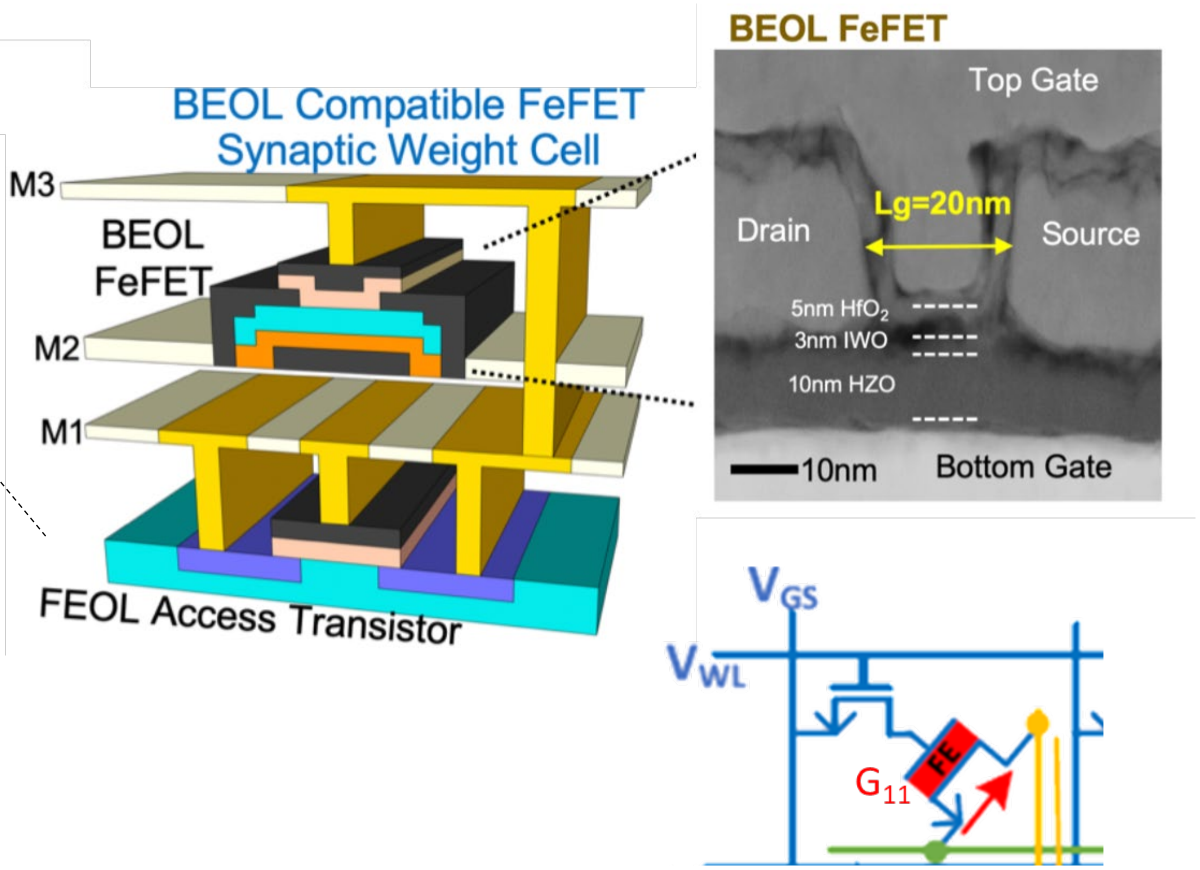
- Nonvolatile conductance tuning
- Low switching energy
- Fast read/write
- ALD w/deeply scaled CMOS nodes

## Challenges

- Endurance (for training)
- High write voltage
- Linearity (for training), stochasticity of conductance tuning
- Density (Legacy nodes, pseudo-crossbar array)
- Area-hungry peripheral circuits (e.g., level shifters (e.g., 45%), high-precision ADC, shift-&-add and buffers)
- How to leverage multi-bit density with peripheral logic scaling (device to system co-optimization)

# Monolithic 3D compute-in-memory





S. Dutta, (IEDM) 2020.

❑ Memory arrays in the BEOL on top of FEOL CMOS, peripheral circuits

❑ Significant advantage in terms of area, energy and latency

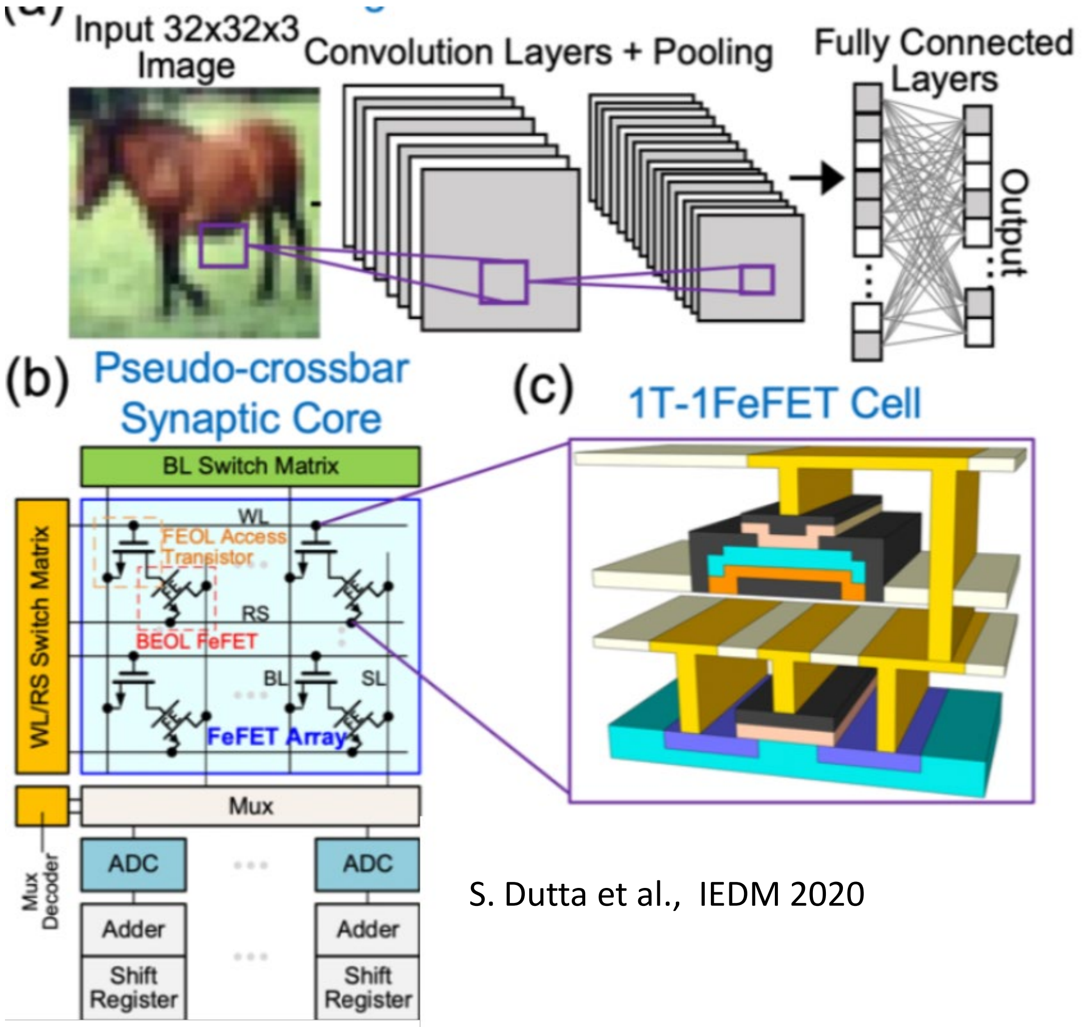
# NeuroSim Framework: Benchmarking IMC performance

**Shimeng Yu (Georgia Tech)** https://github.com/neurosim

❑ Hierarchical simulation framework that covers the device to algorithms to investigate design trade-offs

- Open-source simulator for “in-memory compute” interfaced with PyTorch
- Wide technology choices: SRAM, emerging NVM (RRAM, MRAM, FeFET, etc.) Periphery and interconnect accounted
- Widely used in academia worldwide (>300 registered users)
- Used by industry researchers from SRC/DARPA JUMP sponsors (Intel, TSMC, Samsung, SK Hynix, etc.)
- Validation with IMC prototype with TSMC 40nm RRAM (<2% error)
- VGG-8 model on CIFAR10 dataset (60,000 32x32 color images), with 8-bit weight and 8-bit activation precision.

X. Peng et al., IEDM 2019 and 2020 W. Li, et al. CICC 2020

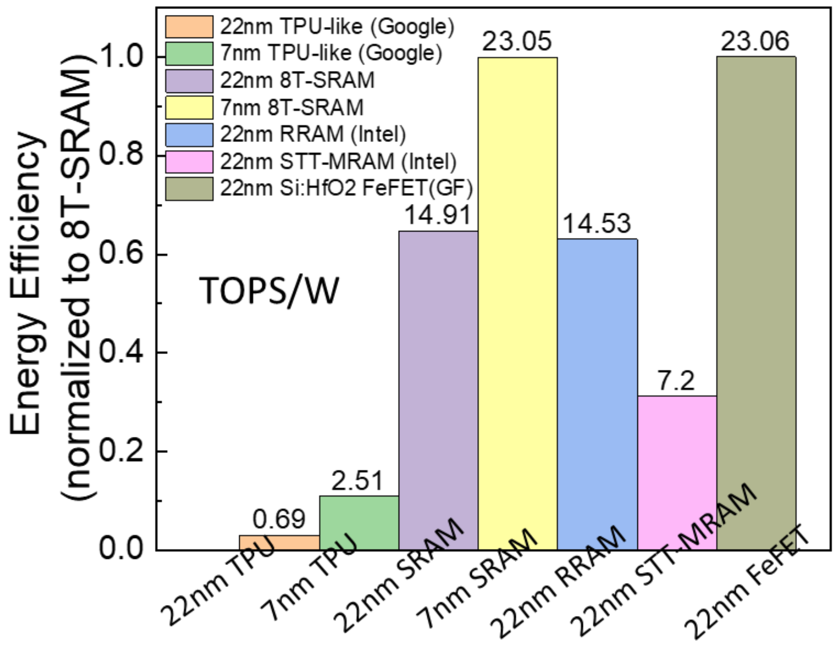


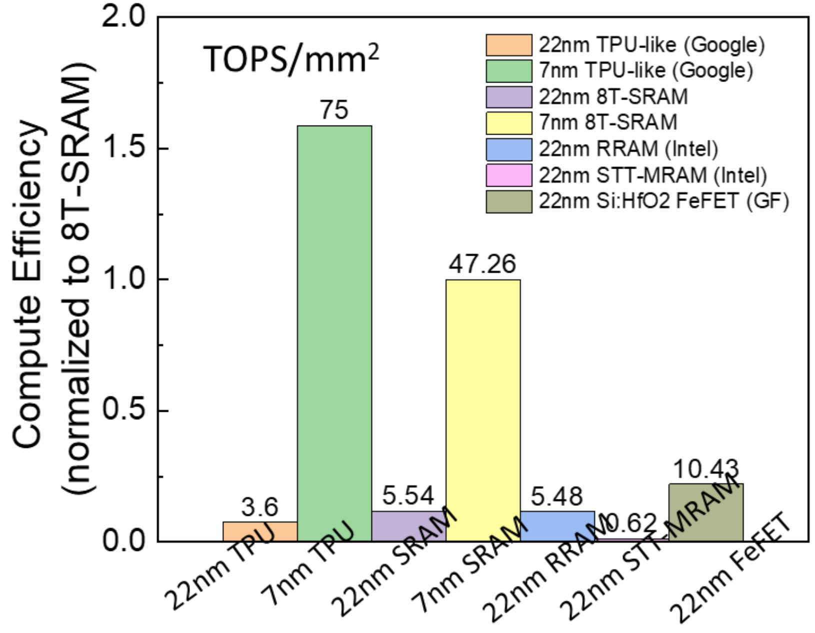
S. Dutta et al., IEDM 2020

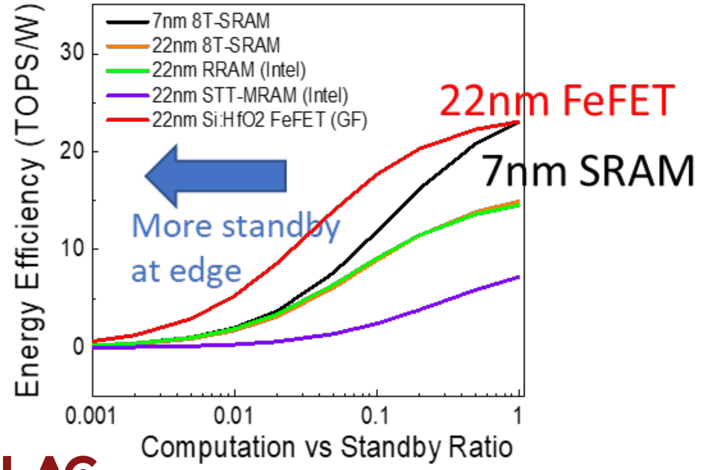
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited). VGG =Visual Geometry Group (CNN)

# Inference H/W Benchmark Results – TOPS/W & TOPS/mm$^2$

**Side courtesy Shimeng Yu (Georgia Tech)**







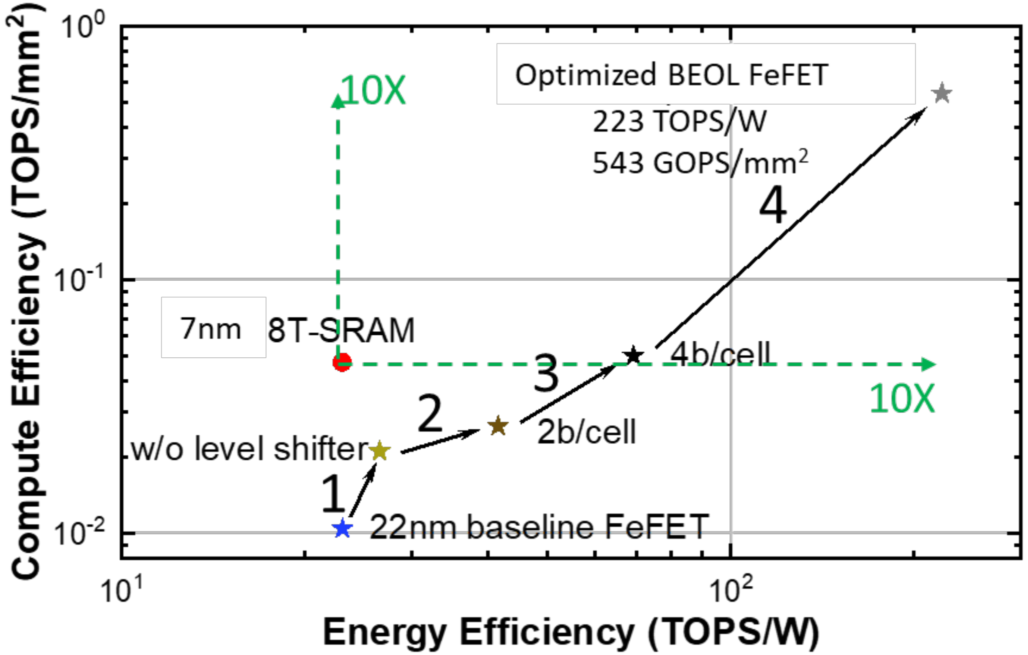
- Limited gain by technology scaling (e.g. TPU), need new approaches: in-memory computing.

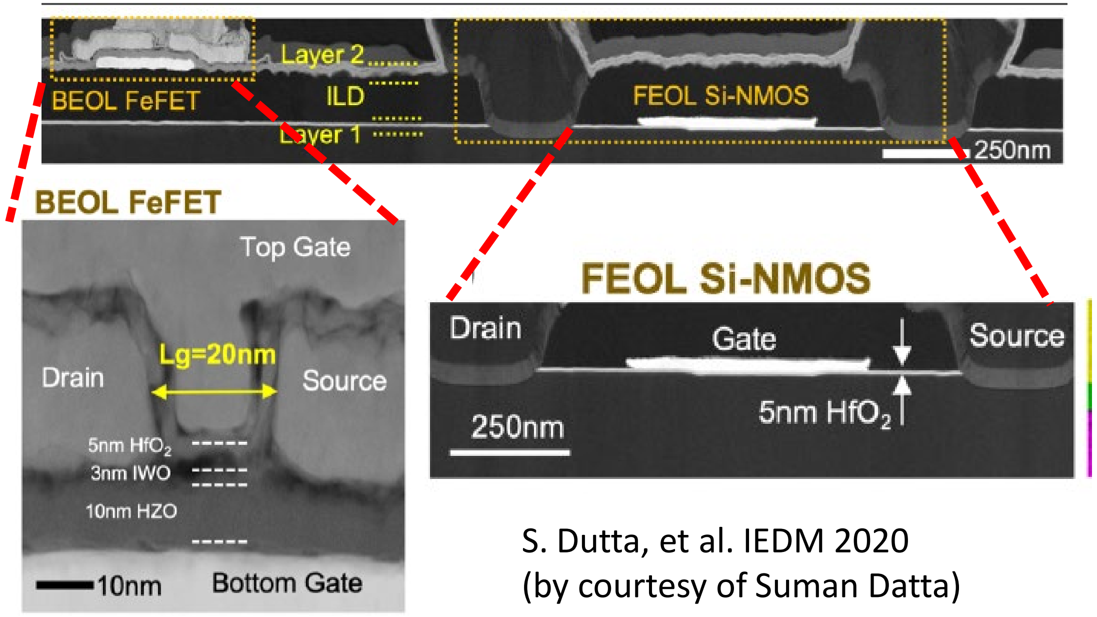
- 7nm SRAM TOPS/W is high, but suffers from leakage when the standby is frequent at edge.
- 22nm FeFET design shows superior TOPS/W, thanks to its high resistance (Ron)

# Roadmap of FeFET Improvements

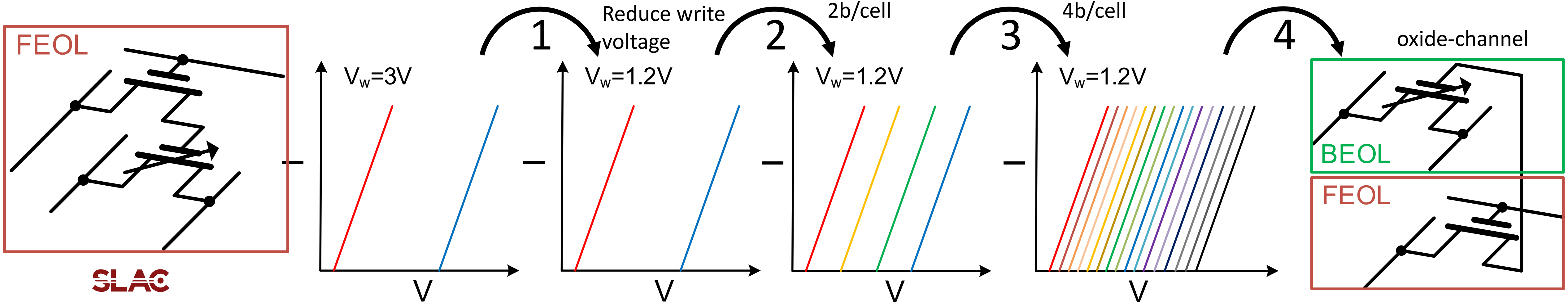
**Side courtesy Shimeng Yu (Georgia Tech)**

**End goal:** 10X improvement in TOPS/W and GOPS/mm$^2$ over 7nm SRAM





S. Dutta, et al. IEDM 2020
(by courtesy of Suman Datta)



# Summary

❑ $HfO_2$-based ferroelectrics offer potential for deeply scaled (22 nm and beyond), low switching energy (~1fJ/bit), non-volatile, fast (sub-ns), multi-bit CMOS compatible memories

❑ FeFET needs to reduce write voltage to logic compatible level, increase cycling endurance, further increase multi-bit per cell, and manage its variability/reliability, particularly in deeply scaled structures

❑ Monolithic 3D integration of BEOL memory and transistors has to overcome thermal processing challenges while maintaining high performance of devices

## Acknowledgement

Prof. Shimeng Yu (Georgia Tech)