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...

Greater speedups for highly parallelizable benchmarks...

"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, reconfigurability, [and/or] high computation density.”
What do ferroelectric devices offer?

Steep subthreshold swings

Analog synaptic behavior

Memory functionality

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.


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


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

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)
Time-dependent Landau Khalatnikov (LK) model

- LK model is SPICE compatible

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

\( \alpha, \beta, \gamma \) calibrated to hafnium zirconium oxide (HZO)

| \(\alpha\) | \(-7 \times 10^9\) m/F |
| \(\beta\) | \(3.3 \times 10^{10}\) m^5/F/coul^2 |
| \(\gamma\) | \(7 \times 10^9\) m/F |
| \(\rho\) | 0.25 |
| \(t_{FE}\) | 5.7 nm |

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

Multi-Domain Preisach Model

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

Multi-Domain Preisach Model

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

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

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**

Search data (e.g., 192.168.1.1)

Choose 01, instead of 10 as 01 more complete

**TCAM array architecture**

https://www.pagiantzis.com/cam/caminro
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 $P$ as well as representation of the projected space $\Omega_P$ 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 locality sensitive hashing (LSH) to define $P(\cdot)$.

**TCAM-supported hashing again an important compute kernel**

Look at TCAMs based on ASCENT technologies to support these models, other applications – consider FeFETs here...
4T, 2FeFET TCAMs (w/negative supply, LK model)

Comparison transistors

Storage elements

D = 0
S = 0
W = VDD

D = 1
S = 1
W = 0

VDD = -0.4V, VGS = negative, Write 0
VDD = 0.8V, VGS = positive, Write 1

VDD = 0.8V, VDD = 0.4V

Comparison

<table>
<thead>
<tr>
<th>Mode</th>
<th>D / D</th>
<th>Y / Y</th>
<th>WL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td>S = 1</td>
<td>0</td>
<td>VDD / -V'DD</td>
</tr>
<tr>
<td></td>
<td>S = 0</td>
<td>0</td>
<td>-V'DD / VDD</td>
</tr>
</tbody>
</table>

Search

<table>
<thead>
<tr>
<th>Mode</th>
<th>D / D</th>
<th>Y / Y</th>
<th>WL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Prior work considered TCAM based on LK model with negative supply

<table>
<thead>
<tr>
<th>Step</th>
<th>WL0/WL1</th>
<th>BL/BL</th>
<th>SL/SL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write 0</td>
<td>$V_{\text{write}}/0$</td>
<td>0/$V_{\text{DD}}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 0/$V_{\text{write}}$</td>
<td>$V_{\text{DD}}$/0</td>
<td></td>
</tr>
<tr>
<td>Write 1</td>
<td>$V_{\text{write}}/0$</td>
<td>$V_{\text{DD}}$/0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 0/$V_{\text{write}}$</td>
<td>0/$V_{\text{DD}}$</td>
<td></td>
</tr>
<tr>
<td>Don’t care 1</td>
<td>$V_{\text{write}}/0$</td>
<td>0/$V_{\text{DD}}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 0/$V_{\text{write}}$</td>
<td>$V_{\text{DD}}$/0</td>
<td></td>
</tr>
<tr>
<td>search</td>
<td>0/0</td>
<td>0/0</td>
<td>data</td>
</tr>
</tbody>
</table>

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


<table>
<thead>
<tr>
<th>Step</th>
<th>WL0/WL1</th>
<th>BL/BL</th>
<th>SL/SL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write 0</td>
<td>1 $V_{write}/0$</td>
<td>0/$V_{DD}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 0/$V_{write}$</td>
<td>$V_{DD}/0$</td>
<td></td>
</tr>
<tr>
<td>Write 1</td>
<td>1 $V_{write}/0$</td>
<td>$V_{DD}/0$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 0/$V_{write}$</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>Don’t care</td>
<td>1 $V_{write}/0$</td>
<td>0/$V_{DD}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 0/$V_{write}$</td>
<td>$V_{DD}/0$</td>
<td></td>
</tr>
<tr>
<td>search</td>
<td>0/0</td>
<td>0/0</td>
<td>data</td>
</tr>
</tbody>
</table>
**2T FeFET design (Preisach model)**

Comparison transistors

Storage elements

<table>
<thead>
<tr>
<th>Mode</th>
<th>BL/SL</th>
<th>BL/SL</th>
<th>SsL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td>S = 1</td>
<td>Step 1</td>
<td>V_write</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 2</td>
<td>V_write</td>
</tr>
<tr>
<td>S = 0</td>
<td>Step 1</td>
<td>0</td>
<td>V_write</td>
</tr>
<tr>
<td></td>
<td>Step 2</td>
<td>0</td>
<td>V_write</td>
</tr>
</tbody>
</table>

\[ \text{BL/SL}= 1 \text{V}, \quad V_{\text{GS}} = \text{positive, Write 1} \]

\[ \text{BL/SL}= 4 \text{V}, \quad V_{\text{GS}} = \text{positive, Write 1} \]

\[ \text{SsL} = V_{\text{write}}; \quad M_1 V_{\text{GS}} = 0, \text{unchanged}; \quad M_2 V_{\text{GS}} = -4 \text{V}, \text{Write 0} \]

Benchmarking (area comparisons)

Current projections suggest competitive density

---


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
- Delay assumes worst case
  - (i.e., 1-bit mismatch...)

![Diagram of TCAM array]
# Benchmarking (other figures of merit)

<table>
<thead>
<tr>
<th></th>
<th>64X64 size</th>
<th>16T CMOS</th>
<th>2T2R</th>
<th>4T-2FeFET</th>
<th>2FeFET</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell structure</strong></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Technology node</strong></td>
<td>45nm</td>
<td>45nm</td>
<td>45nm</td>
<td>45nm</td>
<td>45nm</td>
</tr>
<tr>
<td><strong>Cell area (µm²)</strong></td>
<td>1.12 (7.5x)</td>
<td>0.41<a href="2.7x">1</a></td>
<td>0.65 (4.3x)</td>
<td>0.15 (1x)</td>
<td></td>
</tr>
<tr>
<td><strong>ON/OFF ratio</strong></td>
<td>~10⁶</td>
<td>~100⁰²</td>
<td>~10⁴</td>
<td>~10⁴</td>
<td></td>
</tr>
<tr>
<td><strong>Search voltage</strong></td>
<td>1V</td>
<td>1V</td>
<td>1V</td>
<td>1V</td>
<td></td>
</tr>
<tr>
<td><strong>Search delay (ps)</strong></td>
<td>582 (1.7x)</td>
<td>350 (1.03x)</td>
<td>1013 (3.0x)</td>
<td>341 (1x)</td>
<td></td>
</tr>
<tr>
<td><strong>Search energy (fJ/bit/search)</strong></td>
<td>1.0 (2.4x)</td>
<td>1.2 (2.7x)</td>
<td>0.5 (1.3x)</td>
<td>0.4 (1x)</td>
<td></td>
</tr>
<tr>
<td><strong>Normalized EDP</strong></td>
<td>4.1x</td>
<td>2.8x</td>
<td>3.8x</td>
<td>1x</td>
<td></td>
</tr>
<tr>
<td><strong>Write scheme</strong></td>
<td>Voltage driven dynamic switching</td>
<td>Current driven</td>
<td>Voltage driven</td>
<td>Voltage driven</td>
<td></td>
</tr>
<tr>
<td><strong>Write voltage</strong></td>
<td>1V</td>
<td>Set 1.8V³</td>
<td>±4V</td>
<td>±4V</td>
<td></td>
</tr>
<tr>
<td><strong>Write time</strong></td>
<td>&lt; 2ns</td>
<td>~ 10 ns</td>
<td>10 ns</td>
<td>10 ns</td>
<td></td>
</tr>
<tr>
<td><strong>Write energy (fJ)row</strong></td>
<td>309 (3.5x)</td>
<td>288000 (3225x)³</td>
<td>512 (5.7x)</td>
<td>89 (1x)</td>
<td></td>
</tr>
</tbody>
</table>

---

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 VLSI, 2013.
## Benchmarking (other figures of merit)

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

![Benchmarking Diagram](image-url)
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

Either A=1 or B=1 will pull the bitline down

Read access transistors (ON)

2T+1FeFET memory cell

Write access transistors (OFF)

FeFET (storage device)

Memory array...

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.
Neuromorphic applications
Inference & training

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

Vector-matrix multiplication with crossbars

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

Analog synaptic device characteristics

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

Analog synapses

- Filamentary O-RRAM
  - High density
  - Electro-thermal switching
  - <100ns pulse widths
  - Low $G_{\text{max}}/G_{\text{min}}$ ratios demonstrated thus far

- Non-Filamentary RRAM
  - High density
  - Electro-thermal switching
  - <100ns pulse widths to be demonstrated
  - Low $G_{\text{max}}/G_{\text{min}}$ ratios demonstrated thus far

- Ferroelectric FET
  - 2T design proposed
  - Electric-field switching
  - 75ns pulse widths
  - Large and tunable $G_{\text{max}}/G_{\text{min}}$ ratios demonstrated thus far

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

Ferroelectric FET analog synapse

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

Effect of Pulse Scheme on $P_r$

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

Simulated G vs Pulse Number: Scheme 3

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

FeFET Analog Synapse: Scheme 3

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

Analog Synapse Benchmarking

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

FeFET-based binary crossbars: circuits

Basic cells consists of 2 FeFETs that store (complementary) weight bits. Complementary (input) bits applied to HL and \( \overline{HL} \).

1. Basic cells consists of 2 access transistors.
2. WL set to \( V_{DD} \).
3. BL and \( \overline{BL} \) set to read voltage (\( V_R \)).
4. Cell performs XNOR operation between input, weight.
5. Output read at VL by current or voltage.

FeFET-based crossbars: benchmarking

Benchmarking assumptions for 64x64 crossbar array
- **FeFET**: 10nm FinFET, $T_{fe}=10.5\text{nm}$, $V_{WL}=0.6$, $V_{W}=0.6$, $V_{R}=-0.55$
- **RRAM**: $R_{on}=10\text{K}\Omega$, $R_{off}=1\text{M}\Omega$, $V_{W}=2$
- **For both**: $V_{HL}=0.3$
- **Average case**: half input bits and half weights are 1

For both:
- **RRAM**: $W_{WL}$, $R_{off}$, $T_{h}$, $V_{R}$
- **FinFET** $W_{h}$


![Write power (for 1 column)](image)

For CMOS programming, can also set HIs to 0, SRAM write power low
- RRAM requires >> write voltage, higher power

![Read power](image)

- CMOS lowest due to I0s

![Read delay](image)

![Execution time](image)

![Register area](image)

![RC parasitic parameters used in simulations](image)

<table>
<thead>
<tr>
<th></th>
<th>Cell area ($F^2$)</th>
<th>$R_{wire}$ ($\Omega$)</th>
<th>$C_{wire}$ ($F$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMOS</td>
<td>150</td>
<td>0.245</td>
<td>0.059</td>
</tr>
<tr>
<td>FeFET</td>
<td>60</td>
<td>0.155</td>
<td>0.037</td>
</tr>
<tr>
<td>RRAM (1R*2)</td>
<td>4 ($\times$2) [30]</td>
<td>0.04</td>
<td>0.0096</td>
</tr>
<tr>
<td>RRAM (1T1R*2)</td>
<td>20 ($\times$2) [30]</td>
<td>0.09</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Takeaways ... promising metrics!