Challenges at Circuits Designs for Nonvolatile Memory and Logics in Dependable Systems

Dec. 6, 2013 @ JST DVLSI, Tokyo, Japan

Prof. Meng-Fan (Marvin) Chang

Memory Design Lab. (MDL)
Department of Electrical Engineering
National Tsing Hua University (NTHU), Taiwan
Outline

- Nonvolatile memory (NVM) and logics (nvLogics) in dependable systems
- Challenges at designing ReRAM
- Challenges at designing Flash
- Challenges at designing 3D NVM & nvLogics
- Summary
Volatile vs. Nonvolatile Memory

- **Volatile memory:**
  - Fast, low VDDmin
  - High endurance
  - Working memory
- **Non-volatile memory (NVM):**
  - Slow, high write-voltage
  - Limited endurance
  - Power-off storage
- **Two-macro structure in SoCs**
NVM in Dependable Systems

Typical Chips: SRAM + NVM + Logics

- NVM enables power-off operations
  - Provides power-off storage for program and data (RAM)
  - Provides states storage for selected logics (flip-flops)
  - Reduce standby power
  - Reduce thermal effect
  - Reduce voltage/thermal stress time
Systems Using NVM - Challenges

Today’s challenges

- Large store power + long store time
  => Limited power on/off frequency
  => Vulnerable to sudden power failure
- Slow restore (wake-up/read) time
- Lost local states/data for logics

Typical Chips: SRAM+ NVM + Logics

- Data stored to NVM (slow & large power)
- Data restored to SRAM (slow)

Idle period: Wasted Power & Voltage/Thermal stress
Using Nonvolatile Logics (nvLogic)

Two-Macro solution

- Complex interface
- Serial data transfer
  - Slow store/restore
- Large area penalty
- Lost local states

Nonvolatile SRAM + Flip-flop

- SRAM + NVM within a cell
  - Direct connect (nvSRAM)
- Flip-Flop + NVM (nvFF)
- Fast power on/off
  - parallel store/restore operations
Using Emerging NVM and nvLogics

- **Preferred NVM**
  - Low-power write
  - Low write-voltage
    - eliminate HV devices
  - Fast read and write
  - Low-voltage read

- **Using nvLogics**
  - Fast store/restore
  - Store local states
  => Enable frequent power interrupts

- **Low-voltage nvLogic**
  - Reduce V/T stress
Recent Researches in MDL, NTHU

**NVM & ReRAM**

- **ISSCC 2010**
- **ISSCC 2011**
- **ISSCC 2011**
- **ISSCC 2012**
- **ISSCC 2010**
- **ISSCC 2011**
- **ISSCC 2014**

**3D Memory**

- **VLSI 2010**
- **VLSI 2011**
- **VLSI 2013**

**Low Voltage SRAM**

- **VLSI 2009**
- **VLSI 2010**
- **VLSI 2011**
- **VLSI 2012**
- **VLSI 2013**
- **ISSCC 2014**
Challenges at ReRAM Designs
Examples:
High-Speed ReRAM
Area-Efficient ReRAM
Low-Voltage ReRAM
Recent ReRAM Devices

- Larger write current is required for:
  - High uniformity, long data retention,
  - Rapid write

$\Rightarrow$ Large-area switches
# Recent ReRAM Macros

<table>
<thead>
<tr>
<th>Year</th>
<th>ReRAM Macros</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>~2010</td>
<td>2Mb ReRAM (1T1R)</td>
<td>JSSC 2007</td>
</tr>
<tr>
<td></td>
<td>64Mb ReRAM (3D Cross-point)</td>
<td>ISSCC 2010 (Unity)</td>
</tr>
<tr>
<td></td>
<td>4Mb ReRAM (1T1R)</td>
<td>ISSCC (ITRI+NTHU)</td>
</tr>
<tr>
<td>2011</td>
<td>4Mb ReRAM (1T1R, 7.2ns-R/W)</td>
<td>ISSCC</td>
</tr>
<tr>
<td></td>
<td>8Mb ReRAM (Cross-point)</td>
<td>ISSCC</td>
</tr>
<tr>
<td>2012</td>
<td>4Mb ReRAM (1T1R, 0.5V-R)</td>
<td>ISSCC</td>
</tr>
<tr>
<td></td>
<td>1Mb BJT-ReRAM (0T1R, 4.2ns-Read)</td>
<td>VLSI Symp.</td>
</tr>
<tr>
<td>2013</td>
<td>32Gb ReRAM (Cross-point)</td>
<td>ISSCC</td>
</tr>
<tr>
<td></td>
<td>1Mb ReRAM (1T1R, 0.27V-R)</td>
<td>ISSCC</td>
</tr>
<tr>
<td>2014</td>
<td>16Gb ReRAM</td>
<td>ISSCC</td>
</tr>
</tbody>
</table>

**Embedded (1T1R)**
**Mass-storage (Cross-point)**
ReRAM Challenges: Disturb vs. Bias

- **Write operation**
  - Set: HRS (Hi-R) to LRS
  - Reset: LRS (Low-R) to HRS

- **Read operation**
  - Large $V_R$ cause read disturb

$\Rightarrow$ Requires low BL bias ($V_{BL-R}$)

<table>
<thead>
<tr>
<th>NMOS-RRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL&lt;0&gt;</td>
</tr>
<tr>
<td>SL&lt;0&gt;</td>
</tr>
<tr>
<td>WL&lt;1&gt;</td>
</tr>
<tr>
<td>SL&lt;1&gt;</td>
</tr>
<tr>
<td>$V_R$</td>
</tr>
<tr>
<td>$V_{BL-R}$</td>
</tr>
<tr>
<td>$V_{SET}$</td>
</tr>
<tr>
<td>$0$</td>
</tr>
<tr>
<td>$V_{RESET}$</td>
</tr>
<tr>
<td>$0$</td>
</tr>
<tr>
<td>LRS($R_L$)</td>
</tr>
<tr>
<td>HRS($R_H$)</td>
</tr>
<tr>
<td>$1^\circ$ / $0^\circ$</td>
</tr>
<tr>
<td>$I_{LRS}$</td>
</tr>
<tr>
<td>$I_{HRS}$</td>
</tr>
<tr>
<td>$I_{CELL}$</td>
</tr>
</tbody>
</table>

Lee, H. Y., VLSI-TSA 2010
Wide resistance distribution

- Large resistance (R) and $I_{LRS}$ variation
- Ultra-small-R reference cells cause large/tail $I_{REF}$
ReRAM Challenges: Bias & Speed

- **Bitline bias fluctuation**
  - BL-bias cannot exceed 0.3V
  - Conventional dynamic $V_{BL}$ generation
    - Sensitive to process and Temp. variation

- **Read access time**
  - Small $I_{CELL}$
  - MLC, low $V_{BL}$
  - Read vs. write speed
    - Slow read speed for long BL (large capacity)

![Graph showing $V_{BL}$ fluctuation and access time vs. cells per bit-line.](image)
A High-Speed ReRAM Device - ITRI

1T-1R configuration

<table>
<thead>
<tr>
<th></th>
<th>SET</th>
<th>RESET</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL</td>
<td>$V_{G\text{-}SET}$</td>
<td>$V_{DD}$</td>
<td>$V_{DD}$</td>
</tr>
<tr>
<td>BL</td>
<td>$V_{SET}$</td>
<td>0</td>
<td>$V_{BL}$</td>
</tr>
<tr>
<td>SL</td>
<td>0</td>
<td>$V_{RESET}$</td>
<td>0</td>
</tr>
<tr>
<td>State</td>
<td>$LRS(R_L)$</td>
<td>$HRS(R_H)$</td>
<td>$1^* / 0^*$</td>
</tr>
<tr>
<td>I</td>
<td>$I_{LRS}$</td>
<td>$I_{HRS}$</td>
<td>$I_{CELL}$</td>
</tr>
</tbody>
</table>

MLC

![MLC Chart]

- Level-1
- Level-2
- Level-3
- Level-4

# 15
Example: High-Speed ReRAM

- Parallel-Series Reference-Cell (PSRC)
  - Narrow reference current ($I_{\text{REF}}$) distribution

- Process-Temperature-Aware Dynamic BL-bias circuit (PTADB)
  - Stable BL bias to avoid read disturb
Example: High-Speed ReRAM

- 4Mb High-Speed ReRAM:
  - 7.2ns random read/write access time
  - Small reference variation
  - High-speed read circuit
    - Read disturb, R-variation

SS Sheu & MF Chang, ISSCC, 2011
Low-VDD Read Challenges

- Use RRCS for read
- Removal of BL clamper
- Body-Drain-Driven CSA (BDD-CSA)
  - Reduced SA headroom

- Use RRCS for read
- Removal of BL clamper

- CM/Diode (M1) Headroom
- Lower VDD
- RRCS

- BL Clamper (BLC) Headroom
- RRCS + BDD-CSA
- 0.4V (high yield)

- BL Bias ($V_{BL}$)
- Dynamic & Higher $V_{BL}(0.35V \sim 0.2V)$

- CM-CSA + BLC
- CM-CSA w/o BLC
- BDD-CSA
Example: Low-VDD Read Scheme

- **Standby mode**
  - SE=0, $V_{MAT} = V_{REF} = VDD$
  - BL=DBL= 0V

- **Active mode (Ymux on)**:
  - BL-$V_{MAT}$ charge sharing causes drop in $V_{MAT}$
  - M1/M2 precharge BL/DBL
Example: Low-VDD Read Scheme

- Faster read speed at low VDD
  - 2.9x faster than voltage-mode SA (VSA) at VDD=0.5V
  - 2.1x faster than conventional CSA (CM-CSA) at VDD=0.5V
Example: Low-VDD ReRAM Macro

Technology  | 65nm CMOS Logic Process
---|---
Capacity  | 4Mb (4 x 1Mb)
Sub-array (1Mb)  | 1024 BL x 512 WL
Macro Interface  | Asynchronous
ReRAM Device  | SET= ~2.5V (~25uA)  
             | RESET= ~1.5V (~50uA)
VDD  | Write: 0.48V~1V  
    | Read: 0.32V~1V
Testchip Size  | 4.74mm²
Examples: High-Density ReRAM Cells

- **Vertical Parasitic-BJT (VPBJT)**
  - Logic process, npn
  - Emitter: NLDD implant
  - Base: thin self-aligned P-pocket
  - Collector: N-Well (SL)
  - Min. 4\(F^2\)
VPBJT-ReRAM vs. NMOS-ReRAM

- Larger current density
  - >10x than NMOS
  - Enable smaller cell area + sufficient write current

- Smaller macro area
  - 4~7x reduction
  - Larger capacity, greater reduction

(Measured results)
Thermal-Aware Bitline Bias (TABB)

- Dynamic bitline (BL) bias voltage ($V_{BL-R}$)
  - Track $V_{BE}$ across temperatures (T)
  - Constant $V_R$ across T

=>$Larger I_{CELL}$

MF Chang, VLSI 2013
Examples: High-Density ReRAM Macros

- Cross-process scalability

<table>
<thead>
<tr>
<th>Technology</th>
<th>0.18um Logic</th>
<th>65nm Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>1Mb (8b-IO)</td>
<td>2Mb (16b-IO)</td>
</tr>
<tr>
<td>Sub-blocks</td>
<td>256Kb x 4</td>
<td>1Mb X 2</td>
</tr>
<tr>
<td>RRAM Cell</td>
<td>HfO₂ RRAM (NTHU+ITRI)</td>
<td>TION RRAM (NTHU+TSMC)</td>
</tr>
<tr>
<td>Read Power</td>
<td>6.3mA @ 100Mhz</td>
<td>2.8mA @ 100MHz</td>
</tr>
<tr>
<td>Read Speed</td>
<td>4.2ns</td>
<td>4.7ns</td>
</tr>
<tr>
<td>Write Speed</td>
<td>&lt; 5ns</td>
<td>&lt;10us</td>
</tr>
<tr>
<td>Testchip Size</td>
<td>3157um x 3907um</td>
<td>1900um x 2580um</td>
</tr>
<tr>
<td>Interface</td>
<td>Asyn. NOR</td>
<td>Asyn. NOR</td>
</tr>
<tr>
<td>Features</td>
<td>1. Fast Write</td>
<td>1. Pure logic process</td>
</tr>
<tr>
<td></td>
<td>2. BEOL RRAM</td>
<td>2. Contact layer</td>
</tr>
</tbody>
</table>

MF Chang, VLSI 2013
Challenges for Fast-Read NOR-Flash

Example: Calibration-based CSA
Current-Mode Sense Amplifier (CSA)

- Read-path input offsets
  - Variations in BL bias, SA device, $I_{\text{cell}}$ and $I_{\text{ref}}$

![Diagram showing current-mode sense amplifier with various input offsets and mismatches.](image-url)
Concept of High-Speed CSA - AVB

- Asymmetric-Voltage-Biased (AVB)

**CSA Offset Sources**

**CSA Operation**

**BL Precharge**
1. $V_{BL}, C_{BL}$ variations
2. $I_{REF}$ variations

$V_{BL}$, Bias $\rightarrow I_{CELL}$

$I_{REF}$ generation $\rightarrow I_{REF}$

**Conventional CM-CSA**
- Long $T_{PEE}$ to suppress (1)

**Proposed AVB-CSA**
- Asym. Voltage Bias (AVB) + Short $T_{PEE}$

**Sensing Operation**
3. (Input-stage $V_{TH}$ mismatch)

IV-Conversion (current-mirror, current-load, etc.)

$V_{CP}$ $\rightarrow V_{RP}$

$V_{RP}$ Comparator (VCMP) $\rightarrow$ Digital Out

**Use $\Delta V_{AP}$ to compensate (1)~(4)**

**V_{TH} Nulling for (4)**

**Summed Read-path offset ($I_{OS-SUM}$)**

$= (1) + (2) + I_{OS-SA}$

$= (1) + (2) + (3) + (4)$
Schematic of Proposed AVB-CSA

- Use inactive sub-array to provide dummy BLs for $I_{REF}$
- With $\Delta I_{AP-OS} = -I_{OS-SUM}$ to compensate offset
- $\Delta V_{AP}$ option unit (VOU) provides trimmed $\Delta V_{AP}$ to each AVB-CSA. ($\Delta V_{AP} = V_{AP-CP} - V_{AP-RP}$)
High-Speed CSA - Measured Results

- 1.15x improvement @ 512-cells/BL, VDD=0.9V
- 1.52x improvement @ 2048-cells/BL, VDD=0.9V

Calibration time (<1% of test-time)

MF Chang, A-SSCC 2013
Challenges at 3D NVM

Examples:
1. 3D Vertical-Gate (3DVG) NAND
2. 3D Sequential Layered NVM
3. 3D Nonvolatile Logics
# Published 3D NAND

<table>
<thead>
<tr>
<th>Year</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>~2006</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
<td>2010</td>
<td>2011</td>
<td>2012</td>
<td></td>
</tr>
<tr>
<td><strong>Samsung</strong></td>
<td><strong>Toshiba</strong></td>
<td><strong>Toshiba</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td></td>
</tr>
<tr>
<td>Stacked NAND</td>
<td>BiCS VLSI Symp.</td>
<td>P-BiCS VLSI Symp.</td>
<td>Island-gate SSL decoded 3D VG VLSI Symp.</td>
<td>PN diode decoded 3DVG VLSI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEDM 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-layer TFT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEDM 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mxic</strong></td>
<td><strong>Toshiba</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td><strong>Mxic</strong></td>
<td></td>
</tr>
<tr>
<td>VSAT VLSI Symp.</td>
<td>TCAT VLSI Symp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Univ. of Tokyo</strong></td>
<td><strong>Univ. of Tokyo</strong></td>
<td><strong>Univ. of Tokyo</strong></td>
<td><strong>Univ. of Tokyo</strong></td>
<td><strong>Univ. of Tokyo</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-SGT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEDM 2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Simply stacked</strong></td>
<td><strong>One etch Concept</strong></td>
<td><strong>Various 3D NAND innovations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3D Vertical-Gate (3DVG) NAND

- Etching is not perfectly vertical
- i.e. 500mV top-bottom Vth difference.

Source: Hung and Lue (MXIC), IEDM 2013
## Challenges of 3DVG NAND

**Cross-Layer Variation**

<table>
<thead>
<tr>
<th>Layer</th>
<th>Top</th>
<th>Bottom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Vth</td>
<td>Lower</td>
<td>Higher</td>
</tr>
<tr>
<td>Program speed</td>
<td>Slower</td>
<td>Faster</td>
</tr>
<tr>
<td>PGM&amp;RD Disturb</td>
<td>less</td>
<td>more</td>
</tr>
</tbody>
</table>

=> Require layer-aware scheme

- Higher failure rate than 2D NAND due to the process complexity
  - Need more ECC bits
  => Need faster fail-bit-detection scheme

---

### Forward Read Vt comparison of PL1 and PL2

VLSI 2013, MXIC+NTHU
Example: Layer-Aware-Program-Verify & Read

- **Conventional PV**
  - Same target threshold voltage \( V_{THP} \) across layers
  - Top layer (Layer\([k]\)) program to higher \( V_{THP} \) which causes endurance degradation

- **Proposed LA-PV & R**
  - Different \( V_{THP} \) across layers
  - Lower \( V_{THP} \) for Layer\([k]\) to reduce endurance degradation

---

**Bit Counts**

- **Layer\([1]\)**: “E”
- **Layer\([k]\)**: “E”
- **Layer\([1]\)**: “P”
- **Layer\([k]\)**: “P”

**VTH Distribution after Disturbance**

- SM2 > SM
- SM2’ = SM

SM: Sensing Margin
Example: Measurement Result of 3DVG-NAND

- MLC cell Vth distribution with LA-PV & R

<table>
<thead>
<tr>
<th>Technology</th>
<th>Macronix 3DVG (2-layer memory), 4-metal process</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL Half Pitch</td>
<td>37.5nm</td>
</tr>
<tr>
<td>BL Half Pitch</td>
<td>75nm</td>
</tr>
<tr>
<td>Device</td>
<td>Poly Silicon TFT BE-SONOS Charge-trapping NAND Device</td>
</tr>
<tr>
<td>Memory Density</td>
<td>256Mb (1bit/cell) or 512Mb (2bit/cell)</td>
</tr>
<tr>
<td>Page size</td>
<td>512 Byte/page</td>
</tr>
<tr>
<td>Power Supply</td>
<td>3V</td>
</tr>
</tbody>
</table>
3D Sequential Layered (3DSL) NVM

- A low-thermal process:
  - Less impact on gate dielectrics, S/D structures
  - Available in NDL, Taiwan

- Design & Test Challenges
  - Different cell performance across layers
  - Different thermal-effect across layers
  - In-process monitor/testing
    - Full function test?
    - At-speed test?
  - To be appear in 2013 IEDM (highlight paper)
Example: 3D nvSRAM/nvLatch Cell

Two 3D-stacked resistive device

6T SRAM w/o RFS w/ RFS

Write margin improves 1.64~2.4x
Trade WM for RSNM
  - RSNM is improved 1.42x at TT corner
=> improves VDDmin

Chou & Chang, NTHU/ITRI, Symp. VLSI 2010 / JSSC 2012
Example: 3D nvSRAM/nvLatch Cell

- **On/Off Energy:**
  - Store/re-store energy
  - Standby time vs. on/off frequency

![Graph showing standby energy and standby time for different technologies and process nodes.](image)

- RRAM-based 2-macro
- eFlash-based 2-macro
- 0.18um SRAM @0.4V
- 90nm SRAM @0.2V
- 65nm SRAM @0.2V

One-Macro Scheme (This work)
Example: 3D nvSRAM Macro

- A 16Kb 8T2R nvSRAM macro
  - ITRI’s RRAM + 0.18um CMOS
  - Low-VDDmin & Fast power-on/off speed
  - Enable Logic-in-Memory

16Kb Rnv8T macro

SRAM+ Flash

This work

12T-SONOS

Store time (normalized)

Store Energy (normalized)

Peripheral VDD

PASS

FAIL

VDDmin=0.45V

Cell VDD

Chou & Chang, NTHU/ITRI, Symp. VLSI 2010 / JSSC 2012
Nonvolatile memory is one of the enablers for DS
- Power interrupts to reduce voltage and temp. stress
- Against sudden power failure

Emerging memories
- X-RAM (STT, ReRAM, ..), 3D memory
- Low power and fast read/write operations
- Enable nonvolatile logics

Challenges for designing NVM
- Read disturbance, resistance variation, reference current generations, area/speed vs. write current … etc.

Silicon examples
- ReRAM: high-speed, low-voltage & area-efficient
- 3D-Memory: TSV-RAM, 3D-VG NAND, 3D-SL NVM
- Nonvolatile latch and SRAM

Collaboration of system, circuit and device is needed
Thank You for Your Attentions

Acknowledgements
NTHU, ITRI, NDL, TSMC and MXIC