Cost-effective Test Screening Circuits for High-reliable Embedded SRAMs

Koji Nii,
Y. Yokoyama, K. Kobayashi

Kyoto Institute of Technology, Japan
Self-Introduction

Koji Nii received the B.E. and M.E. degrees from Tokushima Univ., and the Ph.D. degree from Kobe Univ. in 2008. In 1990, he joined Mitsubishi Electric Corp., and he was transferred to Renesas Technology Corp. in 2003, Renesas Electronics Corp. in 2010, where he has been worked on the research and development of embedded SRAMs, TCAMs, ROMs on 28nm to 0.8um bulk/SOI CMOSs and advanced 7-16nm FinFETs. In 2018, he joined Floadia Corp., which is an embedded Flash IP company in Tokyo.

He is now with TSMC Design Technology Japan, Inc., in charge of a head of memory design team for developing advanced FinFET SRAM compilers, and custom cache SRAMs, Register files and computing-in-memory IPs. His current responsibility is Director, Japan Memory Design Program, Memory Solution Division. Dr. Nii holds over 100 US patents and over 150 papers/presentations at major int’l journals/conferences. He served Technical Program Committees of the IEEE CICC and IEDM, and an Associated Editor of the IEEE Trans. on VLSI Systems. He was also a Visiting Professor of Graduate School of Natural Science and Technology, Kanazawa Univ. and now is a Senior Fellow of Green Innovation Lab., Kyoto Institute of Technology (2019-).
1. Introduction
2. Cost-effective pseudo low temperature (PLT) testing
3. Test screening for low-standby power SRAMs
4. Test screening for multi-port SRAMs
5. Summary
MCUs are implemented in various applications. MCUs certainly have CPU and memories.
Embedded SRAM

- High speed
- High cost (large area)
- Volatile
- Most major memory

Memory hierarchy for processor

"Classification of Semiconductor Memories and Computer Memories", Sidhartha, July 22, 2015


Embedded SRAM area ratio is increasing.
Single-port 6T SRAM bitcell schematic

@Noop (data hold)

<table>
<thead>
<tr>
<th>Node</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL</td>
<td>Low</td>
</tr>
<tr>
<td>PC</td>
<td>Low</td>
</tr>
<tr>
<td>BT</td>
<td>High</td>
</tr>
<tr>
<td>BB</td>
<td>High</td>
</tr>
<tr>
<td>MT</td>
<td>Low</td>
</tr>
<tr>
<td>MB</td>
<td>High</td>
</tr>
</tbody>
</table>

On Transistor inside bitcell

PC: Pre-charge circuit
Read/Write operations

**Read operation**

**Write operation**
Dummy Read

A bitcell must remain the bitcell data during dummy read.  

-> Static Noise margin.

Column MUX is often used to reduce bitline capacitance and achieve good aspect ratio.

Opposite characteristics to write margin
SRAM macros features

**Performance**
- Delay speed, constraint speed

**Power**
- Active power, Standby power

**Area**
- Capacity, Cost

Optimized for each applications with high reliability

Ensure the reliability by screening test
Outline

1. Introduction
2. Cost-effective pseudo low temperature (PLT) testing
3. Test screening for low-standby power SRAMs
4. Test screening for multi-port SRAMs
5. Summary
Background – Test cost and SRAM margin

- Test cost will increase because of complicated test.
- Important for low-end MCU to reduce test cost.

SRAM should be tested at both low temperature (LT) and high temperature (HT).
→ Costly multi-temperature test

* SNM: Static Noise Margin
# Motivation

<table>
<thead>
<tr>
<th>3-temp. test (Conv.)</th>
<th>High temp. (HT)</th>
<th>Low temp. (LT)</th>
<th>Room temp. (RT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Leakage failure</td>
<td>Write failure</td>
<td>Bonding failure</td>
</tr>
<tr>
<td></td>
<td>SNM failure etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-temp. test (Prop.)</td>
<td>Leakage failure</td>
<td></td>
<td>Bonding failure</td>
</tr>
<tr>
<td></td>
<td>SNM failure etc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Skipping LT test by screening out write failures at RT.
2. Reducing overscreened dies induced by skipping LT test.
Low Temperature Write Failure

In the NMOS-Slow and PMOS-Fast condition, the current difference (write ability) is decreased at LT.

Write ability: \( I_d(PG) - I_d(Pu) \)

\[
I_D = \frac{W}{2L} C_{ox} \mu_0 (V_{GS} - V_{th})^2
\]
Proposed Pseudo Low Temperature (PLT) Techniques

Reproducing properly the write ability of LT at RT.

\[ I_{DS} = \frac{W}{2L} C_{ox} \mu_0 (V_{GS} - V_{th} - V_{1/2})^2 \]
Issue of conv. Voltage Guardband technique

- Cause of write failure at LT.
- High resistance contacts increase the guardband voltage.
Overscreening Issue

Reducing over-screened dies while screening out LT failures is important for skipping LT test.
Vmin simulation results

Bitcell write Vmin simulation: Accelerated sigma Monte Carlo sim (N=4000)

Conv. Voltage Guardband technique

- Simulation Data @SF corner
- 140 mV GB
- Worst Cell
- Overscreened (3.0% of 4000 samples)

Proposed Pseudo low temperature technique

- Simulation Data @SF corner
- 10 mV GB
- Overscreened (0.7% of 4000 samples)

- Guardband voltage to screen out write failure is improved 140mV -> 10mV by PLT.
- PLT reproduces the low temperature write failures.
- The number of overscreened reduced 3.0% -> 0.7% by PLT.
Test circuit implementation in SRAM macros

Area overhead of PLT (Pseudo Low Temperature) test circuitry is only 0.01% in the 128-kbit SRAM macro.
SPICE sim. waveforms

PG test (SF -40°C)

WL is lowered in PG test.

ΔV₁=30 mV

WL

MT

25°C

-40°C

MB

PG test (PLT) 25°C

PU test (SS -40°C)

BT is raised in PU test.

ΔV₂=20 mV

BT

MT

25°C

-40°C

MB

25°C

-40°C

PU test (PLT) 25°C

PG test and PU test reproduce low temperature condition accurately.
Test chip and measurement results

- 40-nm technology.
- Overscreened dies are reduced by PLT. (29→1) in the case of using the voltage guardband calculated by simulation.
Outline

1. Introduction
2. Cost-effective pseudo low temperature (PLT) test
3. Test screening for low-standby power SRAMs
4. Test screening for multi-port SRAMs
5. Summary
Power scaling

Device scaling

<table>
<thead>
<tr>
<th>Device or circuit parameter</th>
<th>Scaling factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimension tox, L, W</td>
<td>1/κ</td>
</tr>
<tr>
<td>Doping concentration No</td>
<td>κ</td>
</tr>
<tr>
<td>Voltage V</td>
<td>1/κ</td>
</tr>
<tr>
<td>Current I</td>
<td>1/κ</td>
</tr>
<tr>
<td>Capacitance eA/t</td>
<td>1/κ</td>
</tr>
<tr>
<td>Delay time/circuit VC/I</td>
<td>1/κ</td>
</tr>
<tr>
<td>Power dissipation/circuit VI</td>
<td>1/κ2</td>
</tr>
<tr>
<td>Power density VI/A</td>
<td>1</td>
</tr>
</tbody>
</table>

Breakdown of power

\[ P_{total} = \frac{1}{2} CV^2 \cdot Freq + V \cdot I_{standby} \]

\[ P_{active} \quad P_{standby} \]

Increasing of standby power is significant.

R. Dennard et. al., JSSC 1974,

G.E. Moore, 2003 ISSCC.
Leakage current

The leakage current is increasing by the device scaling.

$V_\theta$ : Thermal voltage number
$W$ : Gate width
$n$, $a$ : Empirical numbers
$T_{ox}$ : Gate thickness
$E$ : Electric field

$\text{Gate off}$

$\text{Gate on}$

$I_{sub} \propto \frac{W}{L} \cdot \exp\left(\frac{-V_{th}}{n \cdot V_\theta}\right) \cdot \left[1 - \exp\left(\frac{-V}{V_\theta}\right)\right]$

$\Rightarrow$ Increased by threshold voltage ($V_{th}$) scaling

$I_{gate} \propto W \cdot \left(\frac{V}{T_{ox}}\right)^2 \cdot \exp\left(\frac{-a \cdot T_{ox}}{V}\right)$

$\Rightarrow$ Increased by gate thickness ($T_{ox}$) scaling

$\text{GIDL} \propto E_{DG}$

R. Inagaki et al., 2009 Information and media tech.
Leakage reduction (INV)

Normal body bias

\[ V_{GS} = V_{DD} \]

ON

\[ \text{OUT} = V_{DD} \]

OFF

\[ V_{SD} = 0 \text{V} \]

\[ V_{GD} = V_{DD} \]

Reverse body bias

\[ V_{GS} = V_{DD} - V_{2} \]

ON

\[ \text{OUT} = V_{DD} - V_{2} \]

OFF

\[ V_{SD} = V_{1} \]

\[ V_{GD} = V_{DD} - V_{2} \]

- I_sub can be reduced by increasing \( V_{SB} \)
- GIDL can be reduced by decreasing \( V_{GD} \)
- I_gate can be reduced by decreasing \( V_{GS} \)
Resume standby (6T SRAM bitcell)

Leakage current of SRAM bitcell can be reduced by these 3 biasing with holding the stored data.
Embedded SRAM design examples

- Motivation of reducing standby power and resume standby technique are explained.
- Present 2 examples: 40-nm automotive MCU, 110-nm consumer MCU.
- Each test chips were fabricated and measured as below table.
- New circuits were proposed to prevent the side effects of resume standby as below table.

<table>
<thead>
<tr>
<th>Proposed test circuit</th>
<th>(1) 40-nm Automotive MCU</th>
<th>(2) 110-nm Consumer/Industrial MCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reducing standby power</td>
<td>75% Cut @ 170 °C</td>
<td>75~79% Cut @ 25-150 °C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low cost retention test</td>
</tr>
</tbody>
</table>
Case 1: Automotive MCU

Requirements for Automotive MCUs

- Stable operation at 170 °C.
- Low power at high temperature of 170 °C.
- High reliability screening test for human life.

Conv. SRAM for automotive doesn’t have low leakage techniques because,
- Mature process has low-standby power
- Retention failure risks for safety.

Power train
- Engine
- Transmission

Infotainment
- Audio
- Dashboard
- GPS

ADAS
- Camera
- Sensor

Network
- CAN
- LIN
- Flex-ray

Safety
- Airbag
- Parking assistance
- back monitor

Chassis
- ABS
- Suspension
- Chassis

Body
- Power door
- Power window
- Climate control
- Seat control

K. Nii @ IEEE WRTLT2021, Nov. 25th, 2021
Bitcell optimization

- Optimize Vth
- Gate length $\rightarrow 1.5x$
- Tox $\rightarrow 1.2x$
- Area $\rightarrow 1.1x$
- Leak@TT 25 °C $\rightarrow 0.1x$

- Leakage power is less than 1/10.
- Enough SNM at 170 °C / WM at -40 °C
Bias controlling

1). Floating bit-line
2). Raising VSS source
3). Lowering VDD

Resume standby technique reduces 78 % bitcell leakage power compared with conventional normal standby.

Simulation Data

Leakage power (a.u.)

Conv. Normal
Prop. Resume standby

@FF 1.35 V 170 °C

78%
Proposed retention test circuit

Floating Bitline might induce retention failure.

It is difficult to screen out this faults by conventional test method.

Discharging bitlines to 0V realizing the worst condition of the retention test.
40nm Test Chip and Evaluation Data

Retention SRAM
160 kbit

Median value (mW/Mbit)

<table>
<thead>
<tr>
<th>Mode</th>
<th>25 °C (N=248)</th>
<th>150 °C (N=248)</th>
<th>170 °C (N=4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>17.30</td>
<td>1406</td>
<td>2518</td>
</tr>
<tr>
<td>Standby</td>
<td>1.86</td>
<td>389</td>
<td>643</td>
</tr>
<tr>
<td>Reduction</td>
<td>89%</td>
<td>72%</td>
<td>74%</td>
</tr>
</tbody>
</table>

74 % leakage power is reduced by resume standby technique.
Case 2: Consumer/Industrial MCU

Requirements for Consumer/Industrial MCUs

- Low-cost manufacturing.
- Low standby power at 25 °C and 150 °C because fan or heat sink less for low cost.
- Reducing test cost.

Advanced process

- Large leakage
- Costly

Mature process

- Low leakage
- Low cost
- Heat sink or fan less solution by low leakage at 150 °C
- Low standby mode
- Reduce test cost

SRAM test time in mature process
This bias controller can reduce leakage power with high data retention capability at all temperature range.

**Bias controlling**

The worst for data retention at -40 °C

Resume standby technique reduces 75% and 79% bitcell leakage power at 25 °C and 150 °C, respectively.
Retention test time issue

- The worst condition of retention is low temp -40°C.
- ARVDD is discharged very slowly only by ultra low leakage current at low temp.
- ARVDD has large capacitance.

About 10 sec pause time is necessary to saturate ARVDD. It significantly increases test time and test cost.
Proposed retention test for low cost

Proposed test modes can immediately discharge ARVDD and reduce test time.

The area overhead of test circuit is less than 0.03%.

<table>
<thead>
<tr>
<th></th>
<th>Conv.</th>
<th>Strict</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over-screen</td>
<td>None</td>
<td>Some</td>
<td>Few</td>
</tr>
<tr>
<td>Pause time</td>
<td>Long</td>
<td>Very</td>
<td>Short</td>
</tr>
</tbody>
</table>

Proposed Test circuit

Truth table of modes

<table>
<thead>
<tr>
<th>RS</th>
<th>TRS</th>
<th>TSE</th>
<th>mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Write/Read</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>-</td>
<td>Resume standby</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Strict test</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Medium test</td>
</tr>
</tbody>
</table>
Proposed retention test flow

- The proposed circuit can be controlled by BIST circuit.
- The combination of proposed tests are applied only for low temp.
- Medium test after strict test can reduce testing time and over-screened samples.
Test chip and measurement

110nm Test chip

- Strict test and medium test can screen out failure samples correctly.
- Proposed test flow can reduce test time to 1/50.

| Features          | 110-nm process technology | 2.5-Mbit (4096 word x 40 bit x 16) | 612 μm x 486 μm | 0.302 mm²@160-kbit | 0.518 Mbit/mm² | 4.7 ns | 6.8 ns (147 MHz) | Read: 90 μW/MHz | Write: 105 μW/MHz | 114.3 μW @150°C | 0.73 μW @25°C |

Retention test result at -40 °C

- True V-min +330 mV

Test time (s)

- Conv.(10s)
- Prop. Test flow

Strict & medium
Outline

1. Introduction
2. Cost-effective pseudo low temperature (PLT) test
3. Test screening for low-standby power SRAMs
4. Test screening for multiport SRAMs
5. Summary
Multiport SRAM

- SRAM margin is degraded due to Vth variation.
- Inherent problem of DP-SRAM “Write/Read-disturb issues”.

30Mb SRAM IP in media processor [1]

Dual port (DP) SRAM in multi-core processor [2]

Multiport (DP,2P) SRAMs are used for buffer memory for in-vehicle networks.
• Also, for video processor with high speed and low-power.
• High reliable screening test for human safety.
DP(2RW), 2P(1R1W) SRAMs

8T bitcell can be accessed by 2 ports at the same time.

8T bitcell can be used as DP(2RW) or 2P(1R1W) SRAM.
Disturbance issue occurs during the AWL and BWL are activated at the same time (Same row access). This issue reduces write/read operation margin.
Read Disturbance Issue of 8T bitcell

Read margin is reduced by the disturbance issue.
Write Disturbance Issue of 8T bitcell

Write margin is also reduced by the disturbance issue.
Challenges of DP-SRAM testing

- Need to consider write-read disturbs with clock skew dependence

**Synchronous clock test**

**Asynchronous clock test**

Not the worst case

Test time; enormous
Clock-skew dependences of $V_{\text{min}}$

**Write margin**

- **Zero skew**: WLA, WLB, MT, MB
- **Positive skew**: WLA, WLB, MT, MB
- **Negative skew**: WLA, WLB, BLA/BLA, MT, MB
- **Worst skew**: Ramping up, at ground level

**Read margin**

- **Zero skew**: WLA, WLB, MB
- **Worst skew**: Ramping up

$V_{\text{min}}$ vs. clock skew (ns)
Disturbance test mode

Expanding a “Aggressive” WL pulse to fully cover a “Victim” WL pulse.

Victim

Aggr.

Read worst: Negative skew

Write worst: Positive skew

Screening write and read Disturbance issue of DP SRAM

Proposed screening circuitry of DP-SRAM

Clock Generator

Complex gate

Pre Decoder

S.A. & W.D.

Port-B

TME /TAE

CLKB

Port-A

MUX

Mux
Proposed screening circuit

Clock Generator

- CLKA
- CLKB
- WLA
- WLB

- TME
- TAE
- Delay1
- Delay2
- Pre Decoder

Port-A (Test side)
TME=“H” TAE=“L”

Port-B (Disturb side)
TME=“H” TAE=“H”

synchronous clock phase

CLKA
(CLKA)

CLKB
(CLKB)

WLA
(WLA)

WLB
(WLB)

(Delay1)

(Delay1)

(Delay1)+(Delay2)
Test chip on 28nm

<table>
<thead>
<tr>
<th>Features</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>28-nm CMOS process</td>
</tr>
<tr>
<td>Macro configuration</td>
<td>32-kbit (512-word x 64-bit/word)</td>
</tr>
<tr>
<td>Macro size</td>
<td>16376 µm²</td>
</tr>
<tr>
<td>Area of proposed circuit</td>
<td>22.4 µm²</td>
</tr>
<tr>
<td>Total capacity</td>
<td>512-kbit(32-kbit x 16)</td>
</tr>
<tr>
<td>Bit density</td>
<td>2.1-Mbit/mm²</td>
</tr>
<tr>
<td>Area overhead of proposed circuit</td>
<td>0.14%</td>
</tr>
</tbody>
</table>
Measured shmoo plots

(a) Write operation

(b) Read operation

K. Nii @ IEEE WRTLT2021, Nov. 25th, 2021
Measured Vmin comparisons

(a) Write operation

(b) Read operation

Vmin w/ proposed test is always worse than or equal to the worst case Vmin.
Outline

1. Introduction
2. Cost-effective pseudo low temperature (PLT) test
3. Test screening for low-standby power SRAMs
4. Test screening for multi-port SRAMs
5. Summary
Summary

- Presented pseudo low temperature (PLT) test circuit. It can reproduce LT conditions at RT w/ good correlation. Confirmed reducing overscreened samples on 40nm testchips.

- Introduced test-screening circuits of embedded SRAMs with low-standby modes for low-cost/high-reliable MCUs.

- Also shown multi-port SRAM test screening circuits against inherent read/write disturbances.
Thank you!