Doctoral Dissertation

Design Optimization for Low-power and Highly Reliable Embedded SRAMs on Advanced CMOS Platforms

2021

Kyoto Institute of Technology

Yoshisato Yokoyama

Abstract

This thesis reports a study of embedded static random access memories (SRAMs). The first and second chapters present the introduction and salient challenges. Challenges of this thesis are low power, testability, multi-port and high density and speeding-up in an advanced technology for embedded SRAMs. The third through sixth chapters introduce several examples for these SRAM challenges. The seventh chapter concludes this thesis.

Specifically, the third chapter presents a description of three examples of standby power reduction using resume standby technique. Resume standby can reduce standby power with retention of memory data. SRAMs embedded in automotive or consumer microcontroller units (MCUs) are implemented with resume standby. The standby power of 40-nm 1.1 V SRAM and a 110-nm 1.5 V SRAM are reduced by 1/4-1/5 compared with conventional nominal standby power. In fact, 40-nm 3.3 V SRAM is reduced by 1/50. Risks for resume standby for automotive MCUs and mature technology MCUs are described. The circuits proposed to address the issue are implemented and confirmed using test chips.

The fourth chapter presents a description of the disturbance issue for differential 8T SRAM bitcell. The disturbance issue occurs when two wordlines (WLs) in a bitcell are activated simultaneously. This chapter presents techniques for reducing active power or hastening by reducing disturbance issue effects. The evaluation results of test chips show that speeded-up SRAM has a good operation margin and that the active power can be reduced to 82%.

The fifth chapter presents a description of the testability of embedded SRAMs. This chapter presents a new technique to reproduce write failures under low-temperature conditions in an SRAM macro. This technique can screen out low-temperature write failures at room temperature and can thereby eliminate the low-temperature test for SRAM macro. Overscreened samples are reduced by 1/29 compared with the conventional technique. The test chip was fabricated using 40-nm technology. The effectiveness of this circuit has been confirmed.

The sixth chapter presents a description of how to achieve ultra-high-density SRAM in an advanced technology such as 7-nm. High-resistance WL and bitline (BL) make it difficult to ensure SRAM write/read operations. A new technique to drive WLs and BLs at both edges is proposed. These techniques are compatible with major write/read assist circuits. An SRAM with 29.2 Mbit/mm^2 ultra-high-density is achieved using these techniques. A good VDD-minimum has been confirmed from the test chips.

Contents

List of figures

List of tables

Chapter 1   Introduction

1.1. Background of the Research Area

Today, many semiconductor devices are used in electrical products of all kinds. Integrated circuits (ICs) using MOSFETs have achieved high integration. Furthermore, ICs with huge numbers of transistors have been designated as very large-scale integrated circuits (VLSI). Actually, VLSI are implemented not only in personal computers, TVs, game consoles, mobile phones and automotive devices, but also in home electronics, IoT devices, industrial equipment, medical equipment and artificial intelligence (AI) applications. A microcontroller unit (MCU) is an LSI has microprocessors and memory macros that store programs and data necessary for device control. Recently, the amount of data capacity handled by MCUs has increased rapidly. The scale of implemented memory capacity has been steadily increasing accordingly.

There are memories of various types. Fig. 1 shows a hierarchy for computing systems that are classified as embedded or external memories. The former is implemented inside the MCU. The latter is implemented outside the MCU. The MCU performance can be improved to implement memories. In addition, volatile and non-volatile memories exist. Flash memory is a typical non-volatile memory. A flash memory can store data even if a power supply is cut off. Furthermore, flash memory has very high memory density: one memory cell consists of one transistor (1T). Nevertheless, much time and much power must be used for read and write operations. Therefore, such memory is unsuitable for frequently accessed data.

Embedded SRAMs are used widely as the most basic memory for frequently accessed data. In embedded volatile SRAM, one memory cell consists of 6T. It has a higher memory density than logic gates such as flip-flops. Moreover, it is the most primitive memory and almost process lines have prepared SRAM. Currently, more than 1 billion transistors are implemented in an LSI. Furthermore, embedded SRAMs occupy more than 50% of the LSI area. Therefore, improving the embedded SRAM performance directly improves the LSI performance. Generally, intellectual property (IP) cores including SRAM macros require some features. They are typically speed (or performance), power, area (or cost), and yield. The purpose of this thesis is to present optimization of each feature of the embedded SRAMs according to the applications and to verify SRAM macro reliability using screening tests.
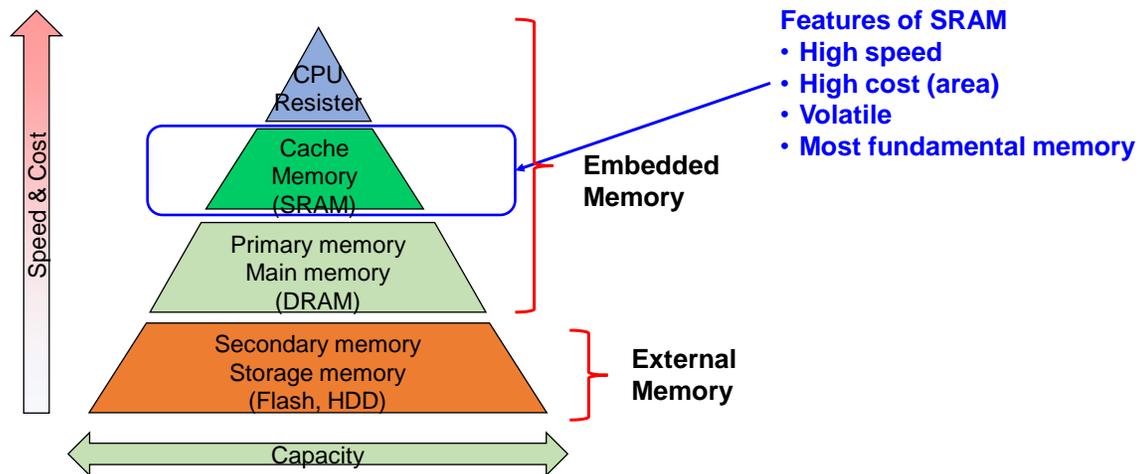
Fig. 1 Memory hierarchy in computing systems.

1.2. Overview of this Thesis

A simply visualized outline of this thesis is presented in Fig. 2. First, the background and objective of this study are described. Crucially important issues related to embedded SRAMs are presented in Chapter 2.

Chapter 3 explains the power reduction technique. Standby power reduction is especially important for volatile SRAM. The resume standby technique is demonstrated for standby power reduction and column selection by wordlines is demonstrated to reduce active power. Side effects using the resume standby technique and the screening test technique for them are also described respectively for automotive MCU and consumer MCUs.

Chapter 4 explains the importance of multi-port SRAMs. Disturbance issues occur especially with multi-port SRAMs. This issue degrades the SRAM performance. In this chapter, techniques to reduce effects of the disturbance issue and to improve performance are demonstrated. Chapter 4 also demonstrates the use of a screening test technique to address the disturbance issue.

Chapter 5 presents a demonstration of the technique to reduce test costs. Low-cost MCUs must have reduced test costs. The pseudo-low-temperature technique can reproduce low-temperature conditions at room temperature using the gate bias technique. The technique reduces testing costs by obviating low-temperature tests.

Chapter 6 describes a useful technique to improve the SRAM macro density in advanced technologies such as a 7 nm FinFET process. In advanced process technologies, wire resistance

disturbs the speeding up of SRAM macros. Chapter 6 demonstrates techniques to improve the performance of SRAM macros despite the large SRAM bitcell array with high-resistance wordlines and bitlines.
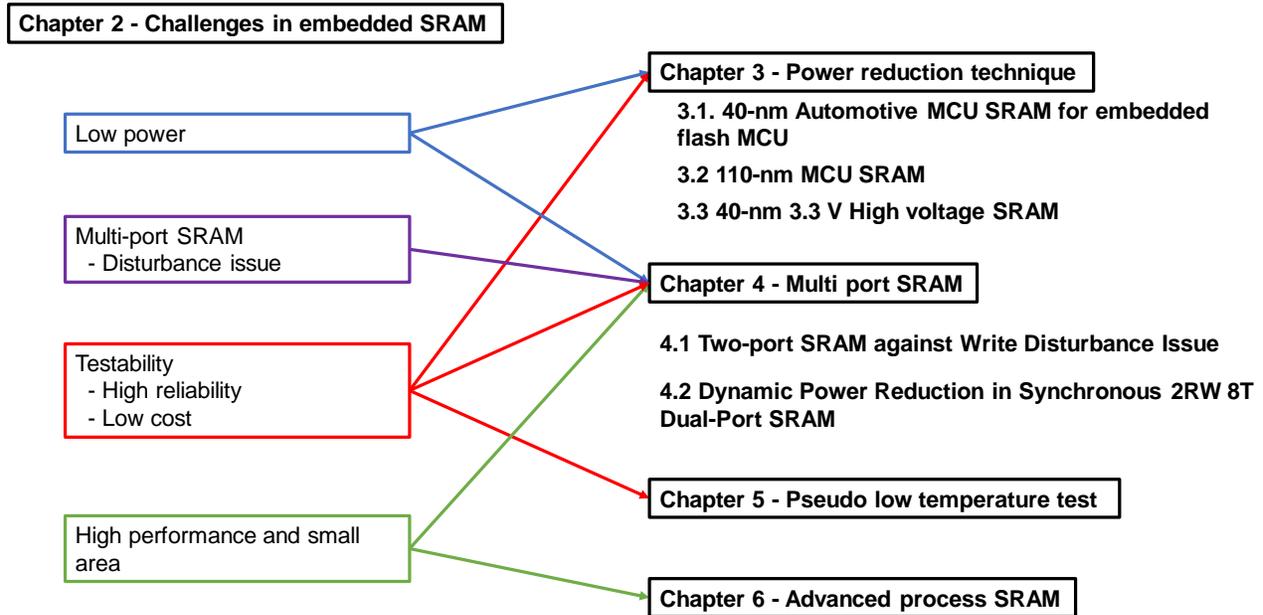
Fig. 2 Outline of this thesis.

Chapter 2　Challenges for Embedded SRAM

2.1. Introduction

2.1.1. Function of SRAM macro

First, basic operations of SRAM are explained. Fig. 3 shows a typical functional symbol of an embedded single-port (SP) SRAM macro along with the timing diagram. The symbols CLK, CEN, and WEN respectively denote the clock input, cell enable control, and write enable control. In the figure, the A[a-1:0], D[b-1:0], and Q[b-1:0] respectively denote the address input bus, data input bus, and data output bus, where variables a and b respectively represent integer numbers depending on the word depths and data bit width. For instance, for an SRAM macro with 4096 word depth and 32 bit width, a is 12 and m is 32. MCU devices in recent years require the use of many SRAM macros, which have various word depth and bit width configurations. The SRAM macro operates with synchronization of the system clock CLK. The other inputs CEN, WEN, A[a-1:0] and D[b-1:0] must satisfy the set-up time and hold-time by the positive clock edge every clock cycle. When the control input CEN is high (CEN= "H"), the SRAM macro is in a standby mode (no operation) despite of CLK asserting: it is not accessed for read or write operations. The SRAM operates for reading or writing depending on the WEN input level when the CEN is low (CEN= "L"). When WEN= "H", the SRAM reads out the stored data in memory cells addressed by A[a-1:0]. Conversely, when WEN= "L", the SRAM writes the data bits for memory cells addressed by A[a-1:0].
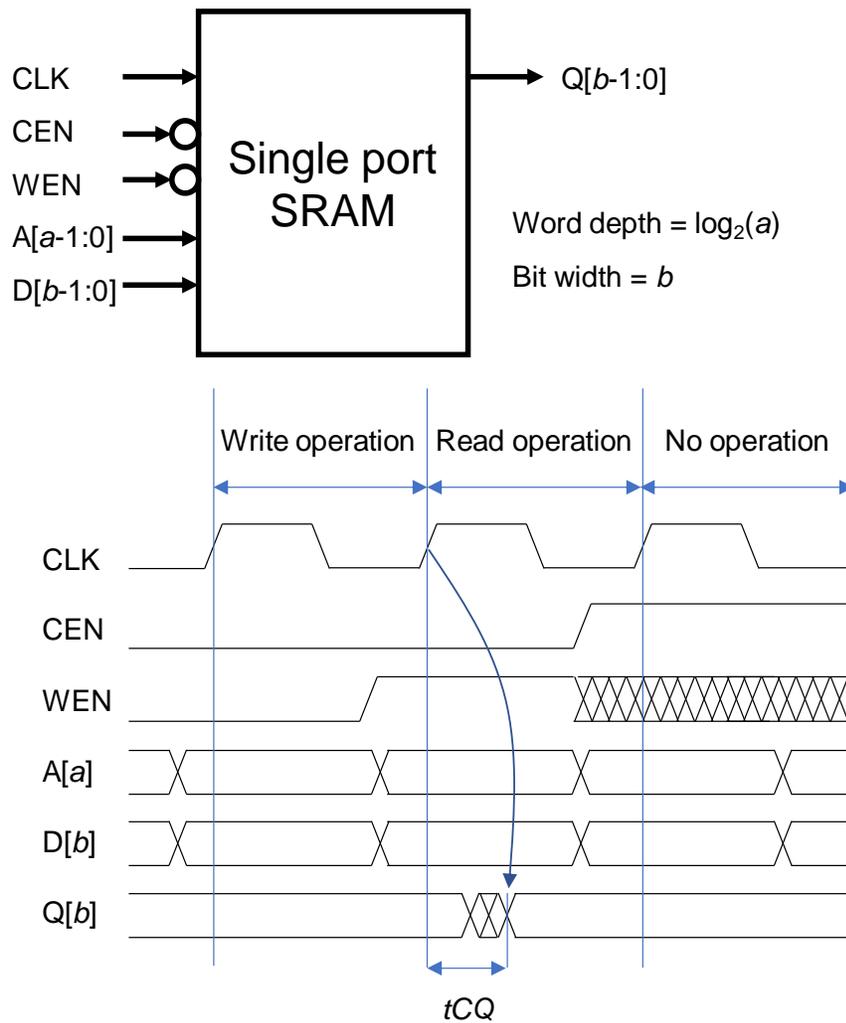
Fig. 3 Typical functional symbols and timing chart of a single-port SRAM macro.

2.1.2. 6T SRAM cell

Fig. 4 portrays a schematic of a 6T SRAM bitcell. It consists of the cross-coupled inverter pair and two pass-gate NMOS (access transistors). Each inverter has a pull-down NMOS (drive transistor) and a pull-up PMOS (load transistor). The wordline (WL) is connected to the gate terminal of both accessed NMOS. The bitline pairs bitline true (BT)/ bitline bar (BB) are connected to the source terminals of both access NMOS. These transistor sizes must be designed for the optimal size which achieves both a small area and sufficient stability. Empirically, each transistor size is designed such that the drain current of each transistor, the load PMOS, access NMOS, and drive NMOS, becomes approximately 1:2:4. Detailed discussions of the circuit design optimization for the 6T SRAM bitcell are omitted here because the optimization depends strongly on the process

technology.
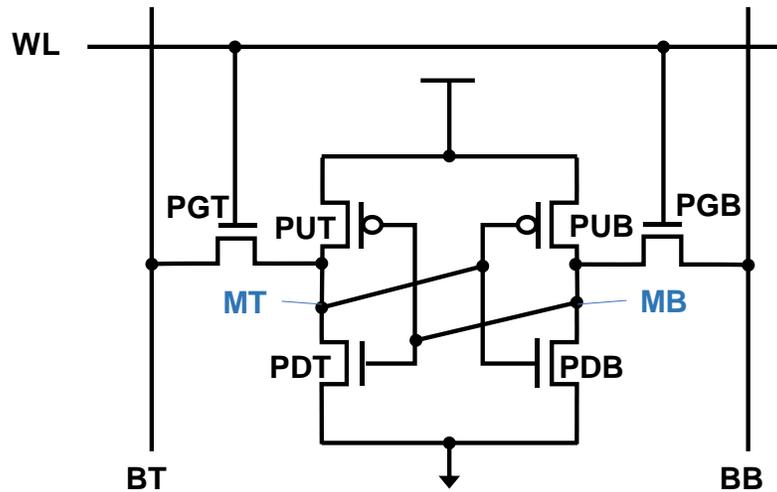
Fig. 4 Schematic showing a 6T single-port SRAM bitcell.

Fig. 5 shows an SRAM bitcell in standby mode. In standby mode, WL is low. Also, BT / BB is precharged to a high level by the pre-charge circuit. The 0/1 data are latched respectively to the memory node true (MT) / memory node bar (MB) by the cross-coupled inverters. Figure 6 shows the SRAM bit cell during the read operation. Read operations are done by pulling bitlines from a high level to a low level. The PC node in Fig. 6 changes from low to high. The pre-charge circuit of BT / BB turns off; then BT / BB goes to a high-impedance state (Hiz). Then WL changes from low to high to select a bitcell. Then, BT is discharged by PGT and PDT, which are on. The memory data can be read out to the outside bitcell. Actually, for speeding up, when the BT is discharged to a certain voltage, the sense amplifier (SA) is started to amplify the BT / BB potential difference and to read the data. Fig. 7 shows the write operation. Actually, write operations are performed by pulling down the memory cell high level node to a low level. In the case of Fig. 7, MB is pulled down from a high level to a low level. Nodes CTW/CBW in Fig. 7 are dependent on write data: only one of them becomes low; another remains high to precharge the bitline. In this case, CTW="H" and CBW="L". Similarly to the read operation, WL and PC become low. The write amplifier (WA) discharges BB. WA and PGB forcibly pull down MB, which is pulled up by PUB; then PDT turns off, PUT turns on, MT changes from a low level to a high level; then the write operation is completed.

Node	Level
WL	Low
PC	Low
BT	High
BB	High
MT	Low
MB	High

Fig. 5 Bitcell in standby (No-operation).

Fig. 6 Bitcell in read operation.

Fig. 7 Bitcell in write operation.

2.1.3. Configuration of SRAM macro

Fig. 8 shows a block diagram of a standard SRAM macro. It can be divided roughly into a control block, bitcell array, WL decoder and driver block, IO block, and timing generator block. The clock generator in the control block detects the rising edge of the clock signal (CLK) and generates an internal clock. The pulse width of the internal clock required for the write operation and the activation timing of the sense amplifier necessary for the read operation is generated by the timing generator. Here, the row address and the column address are described. The SRAM macro generally divides the address into a row address and a column address to achieve the requested word depth, bit width, X / Y ratio of the outer shape, power, and performance. The row address activates a certain wordline. The column address selects a certain bitline in one datain and dataout. Fig. 8 shows a case of column MUX2, either one of BT[0]/BB[0] or BT[1]/BB[1] is selected by the column address. Therefore, WLs and BLs can select a unique bitcell. Typically, the transfer NMOS is used for the multiplexor of the column MUX for a write circuit, transfer PMOS is used for read circuit. Here, the bitcell operation selected by the row address, but not selected by the column address is explained. Generally, unselected columns are pre-charged continuously because there are advantages rerated to IR-Drop to avoid the rush current, which pre-charges all bitlines when closing the operation. Fig. 9 shows the operation of the unselected column. Actually, WL is turned on even if the bitcells are unselected by the column address; BT / BB maintains a high level. This operation is designated as a "dummy read operation". During this operation, the MT voltage

rises slightly because of the on-resistance division between PGT and PDT.

Fig. 8 Basic block diagram of SRAM macro.

Fig. 9 Bitcell in dummy read operation.

2.2. Static Noise Margin and Write Margin

As shown in Fig. 9, the WL turns on while the BL is pre-charged during the dummy read operation. The PG / PD ratio is adjusted so that the raised MT node does not exceed the PUB / PDB logical threshold and so that the bitcell can retain cell data even during a dummy read operation. This threshold ratio is designated as the beta ratio. The stability of the cell data during the dummy read operation is designated as the static noise margin (SNM) [26]. However, the ease of inverting cell data during the write operation is called the write margin (WM). Actually, WM is determined by the PUB / PGB ratio. That is designated as the alpha ratio. SRAM bit cells must have two conflicting characteristics: stability and writability, i.e., SNM and WM.

The difficulty of achieving both SNM and WM is increased considerably if the variation of transistors must be considered. The variations of transistors are classifiable to global variation and local variation. Global variations represent different characteristics of silicon wafers or wafer lots. The local variations must be considered not only for different characteristics among chips in a wafer, but also for pair transistors in an SRAM bitcell. Here, the expression of the process corner condition is defined. The process corner condition is expressed by the threshold voltage of PMOS (V_{tp}) and NMOS (V_{tn}). For example, the process corner condition is "FS" for the case in which V_{tp} is low (Fast) and V_{tn} is high (Slow). Fig. 10 portrays the process corner conditions with x-axis V_{tn} and y-axis V_{tp}. Diamonds with a solid line represent global corners based on global variations. Red circles show local corners based on local variations from each global corner. The SNM has a smaller margin at the FS corner; the WM is smaller at the SF corner. The blue and red solid lines in Fig. 10 respectively represent the boundaries of the operation margin of SNM and WM. The local variation is increasing more and more. The pass range in Fig. 10 has been narrowed by recent device scaling.
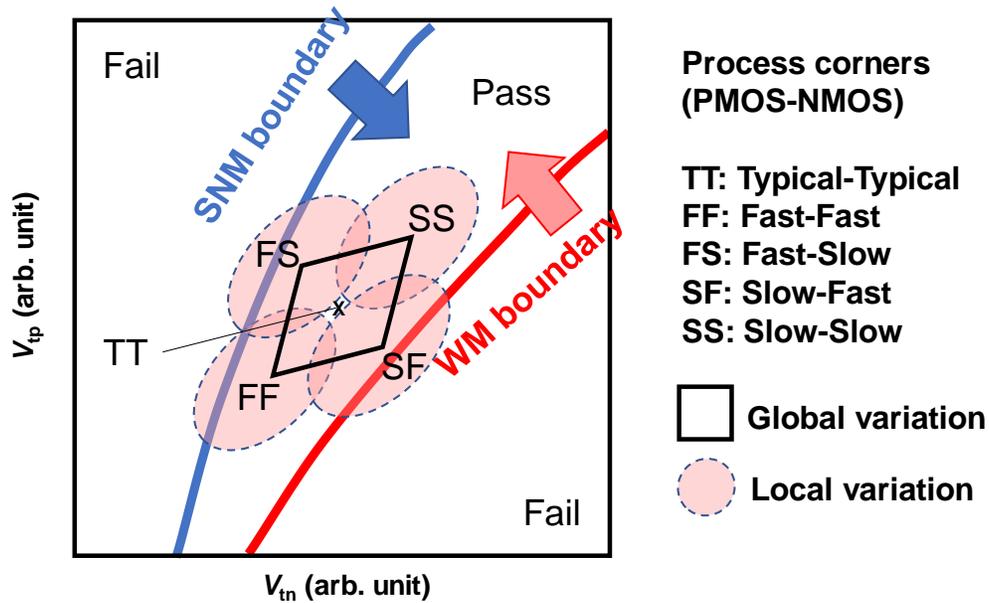
Fig. 10 Process corner definition and SNM/WM boundary.

2.3. Low Power for SRAM

The power consumption is classified for power of two kinds: active power and standby power. Active power is consumed when SRAM macros execute a write or read operation. The standby power is consumed not only during write and write operations but also when SRAM macros are in standby. Fig. 11 presents an illustration of the waveforms of power consumption by the active power and standby power. The active power is consumed by charging and discharging the capacitance. It is expressed by equation (i). Variable C represents the capacitance discharged and charged. Also, V represents the power supply voltage. *Freq* denotes the frequency of operations.

$$P_{active} = \frac{1}{2}CV^2 \cdot Freq \tag{i}$$

The standby power is expressed as equation (ii). $I_{standby}$ denotes the constant current in standby mode.

$$P_{standby} = V \cdot I_{standby} \tag{ii}$$

The amount of power consumption is expressed as equation (iii).

$$P_{total} = \frac{1}{2}CV^2 \cdot Freq + V \cdot I_{standby} \qquad \text{(iii)}$$
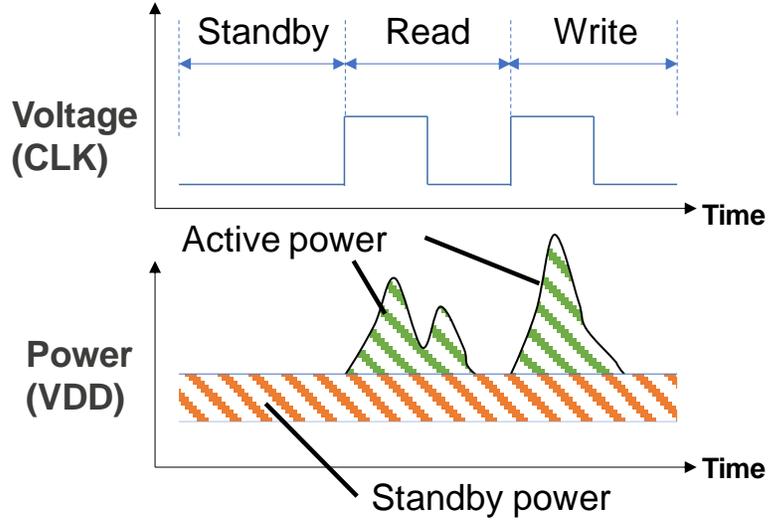
Fig. 11 Active power and standby power.

Along with Moore's law, transistor density has increased considerably over the years. As the gate oxide thickness has decreased, power supply voltages have decreased. Fig. 12 portrays a plot of the active power and the standby power of each generation [1]. Although the processor frequency increases, the active power is increased gradually by lowering the supply voltage. Furthermore, the increase of standby power is more severe. Techniques to reduce the standby power of the SRAM macro are crucially important. They occupy large areas of the SoCs and MCUs. Fig. 13 presents an illustration of three components of leakage current of transistors as an example of NMOS transistors. The first is sub-threshold leak current between the drain and source (I_{sub}) of off-gate. The second is a gate-induced drain-leakage current (GIDL) between the drain and body of the NMOS (GIDL) of the off-gate. The third is gate leakage between the gate and body of the PMOS (I_{gate}) of the on-gate. I_{sub} increases exponentially depending on the threshold voltage (V_{th}). Therefore, I_{sub} increases exponentially depending on the temperature, too. Actually, GIDL, which occurs between the bands by a high electric field between the gate overlapping region and the drain, increases depending on the gate-drain voltage. I_{gate} occurs by the gate-body (or gate-drain or gate-source) voltage when the gate is on. I_{gate} also increases depending on the voltage between the gate and other pins. As the channel length decreases, V_{th} decreases. Moreover, the gate oxide is thinned;

the electronic field is strengthened by device scaling. Therefore, all components of the standby power are increased rapidly by device scaling.

Furthermore, the standby power increases at high temperatures because of increased I_{sub}. MCUs for controlling automotive engines are exposed to high temperatures. Therefore, extremely low standby power is necessary at the high temperature of 170°C. In addition, industrial MCUs that are not equipped with a cooling fan to reduce costs necessitate low standby power at the high temperature of 150°C. Furthermore, MCUs for Internet of Things (IoTs) are under lower temperatures than the former, but must be battery-powered for as long as 10 years. Reducing power consumption considerably is an important task for development.
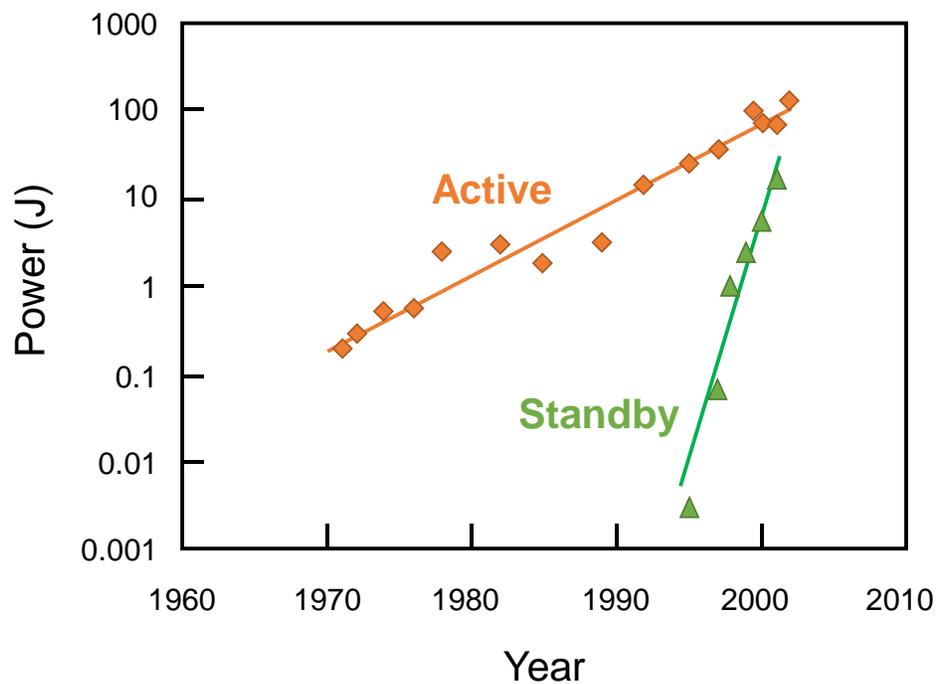
Fig. 12 Power consumption of processors [1].

Gate off   Gate on
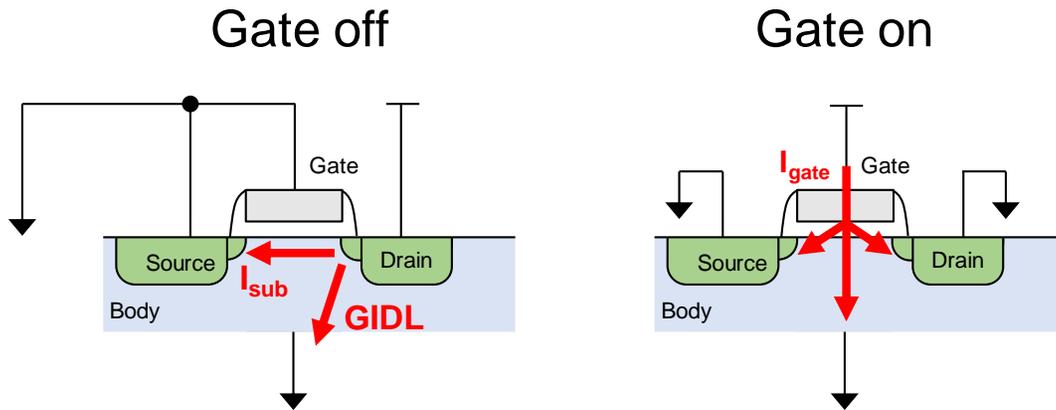
Fig. 13 Leakage current in a transistor.

2.4. Multi-port SRAM

As technology advances, the amount of the implemented memory tends to increase. This increase is attributable to increased data processing to improve the application performance. However, as described above, increasing the clock frequency increases the power. System architectures tend to increase parallel processing to improve the processing performance. As a result, demand for multi-port SRAMs has increased as opposed to single-port SRAMs, which have only one write/read port. The multi-port SRAMs can perform write/read operations in parallel. These SRAMs are mainly suitable for communications and image processing, thereby improving the processor performance.

Dual port (DP) SRAMs and Two port (2P) SRAMs are typical multi-port SRAMs. Fig. 14 presents functional symbols of DP SRAMs and 2P SRAMs. Actually, DP SRAMs have two ports that can do write/read operations. The ports can access different addresses simultaneously or asynchronously. DP SRAMs are often expressed as "2RW SRAM". Although 2P SRAMs also have two ports: one port can only do write operations; the other port can only do read operations. Fig. 15 presents a schematic of a differential 8T SRAM bitcell. Whereas the 6T bitcell has a WL and a pair of BLs, as shown in the Fig. 4, the 8T bitcell has two WLs, two pairs of BLs, and two pairs of pass-gate transistors for multiple access.

The differential 8T SRAM bitcell is extremely useful because it is useful for both DP SRAM and 2P SRAM. It is prepared for almost all process nodes. However, that bitcell has an important specific issue designated as a "disturbance issue". The disturbance issue occurs when two ports are

activated simultaneously for a bitcell. This issue decreases the write/read operation margin and worsens the minimum voltage of operation (V_{min}) [18]. Details of this issue are presented in Chapter 4 .

Word depth = $\log_2(a)$
Bit width = b
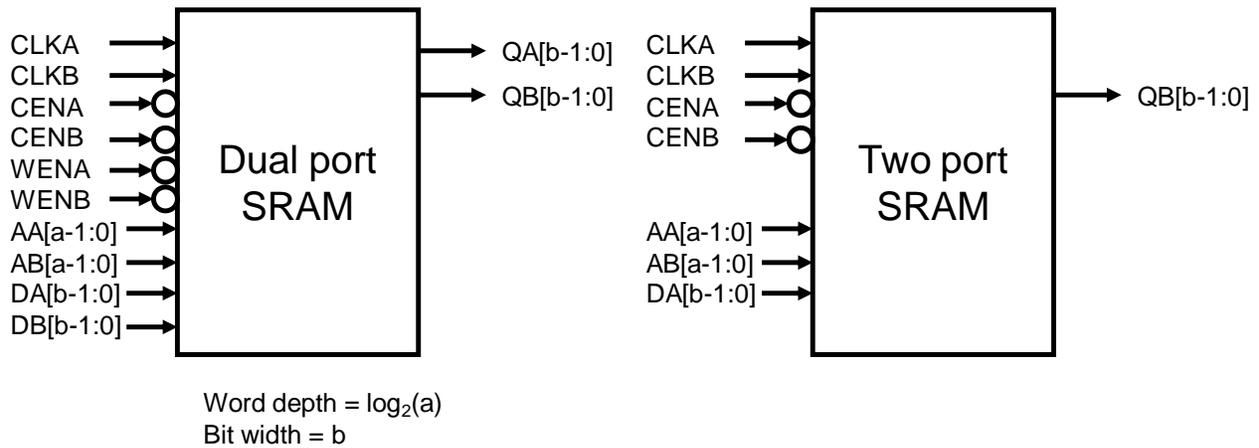
Fig. 14 Function symbols of DP SRAM and 2P SRAM.
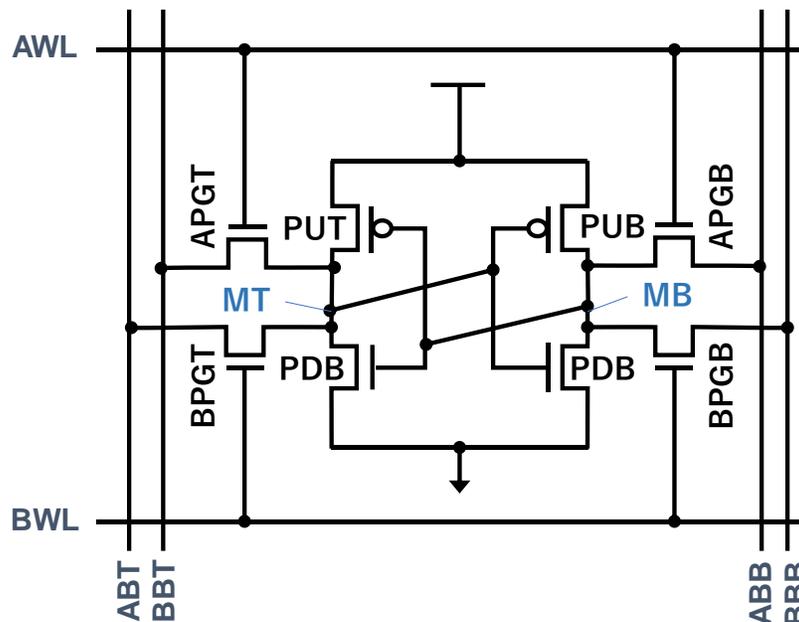
Fig. 15 Schematic of differential 8T SRAM bitcell for DP and 2P-SRAMs.

2.5. Testability

Implementation of transistors in MCUs has proceeded year-by-year by technology scaling. The transistor density has increased. The cost per transistor has decreased. However, the large number of transistors, demands large amounts of time to test one chip to screen out failures. The time to occupy the tester directly increases the test cost. Fig. 16 presents the tendency of manufacturing and test costs [15]. In addition, new failure modes are occurring because of the complexity of process technologies and IP or MCU functions. Some failure modes might occur under certain processes, and voltage and temperature (PVT) conditions. Moreover, a special screening test is necessary to screen out complicated failure modes. For example, the DP SRAMs presented above have special failure modes [40][64]. Huge amounts of time must be spent in screening out failures by the typical function test.

The testability of MCUs requires different features depending on the application of the MCUs. Because failures can be life-threatening for automotive MCUs, robust and high-quality tests are necessary to ensure screening-out of failures. Reducing test costs for low-priced general-purpose MCUs is important because reducing the price of MCUs and selling many dies is necessary.
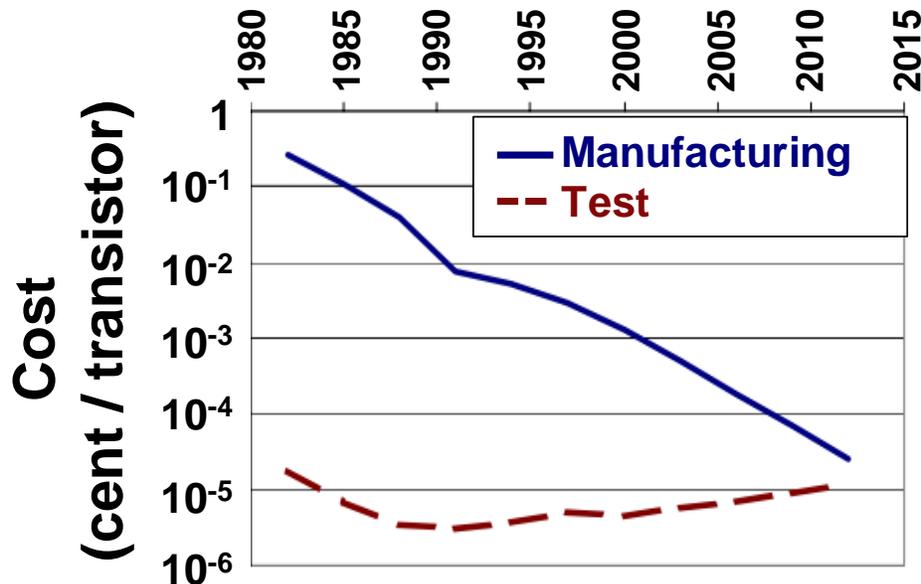
Fig. 16 Tendency of manufacturing and test cost [15].

Screening tests at multiple temperatures also increase test costs. Especially, low-temperature testing requires special equipment. It takes much time to test costs and to stabilize the temperature.

By contrast, Fig. 17 presents the temperature dependence of SNM and WM of certain SRAM bitcells. Actually, SNM decreases at high temperatures. Also, WM decreases at low temperatures. To screen out high-temperature and low-temperature failures in MCUs that have SRAM macros, it is necessary to conduct tests at each temperature.
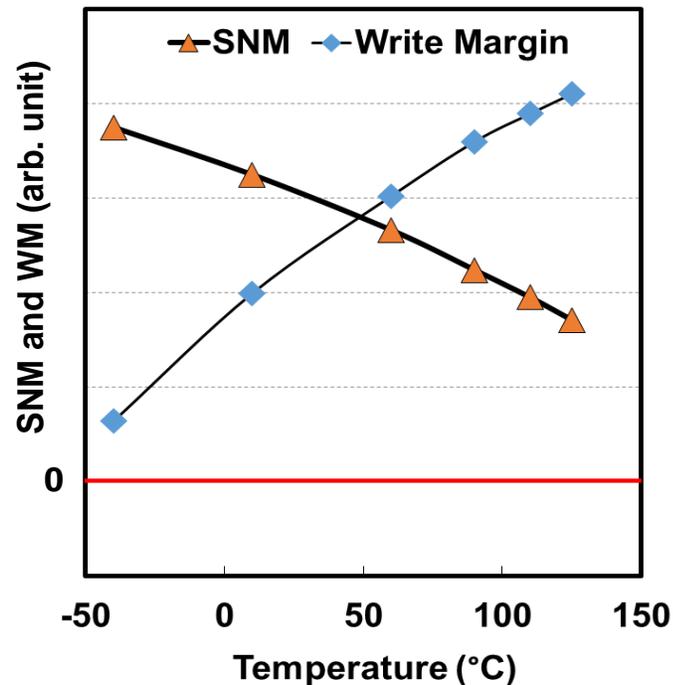
Fig. 17 Temperature dependence of SRAM bitcell SNM and WM.

2.6. High-density SRAMs in Advanced Technology

Because the Fin Field-Effect Transistor (FinFET), which has a three-dimensional (3D) structure, has been adopted under 16-nm, the device scaling has proceeded. Applications are as a general purpose processor, artificial intelligence (AI), graphics processing unit (GPU), and so on. For those applications performance, power and area (PPA) improvements are necessary.

The SRAM macro consists of the bitcell array and the peripheral circuits for bitcells, as presented in Fig. 8. Fig. 18 shows the tendency of SRAM bitcell area [71]. After 2017, the scaling of the bitcell looks slower than before. Therefore, it is more important to ascertain how to make a smaller peripheral circuit to improve PPA. Additionally, to ensure continuity of the SRAM cell layout and their characteristics, it is necessary to secure a certain clearance between the SRAM

bitcell array and the peripheral circuits in the layout plot. This clearance causes dead space in the layout design of SRAM macro, and worsens memory density. When improving the memory density, one must avoid dividing the SRAM bitcell array. Therefore, the WLs and BLs in a bitcell array lengthen. Furthermore, the wire resistances of WLs and BLs increase rapidly because of the scaling. Fig. 19 presents the tendency of wire resistance of tight interconnect [72]. The increased wire resistance causes an increase in the resistance capacitance delay (RC delay). For that reason, it worsens the SRAM macro performance.
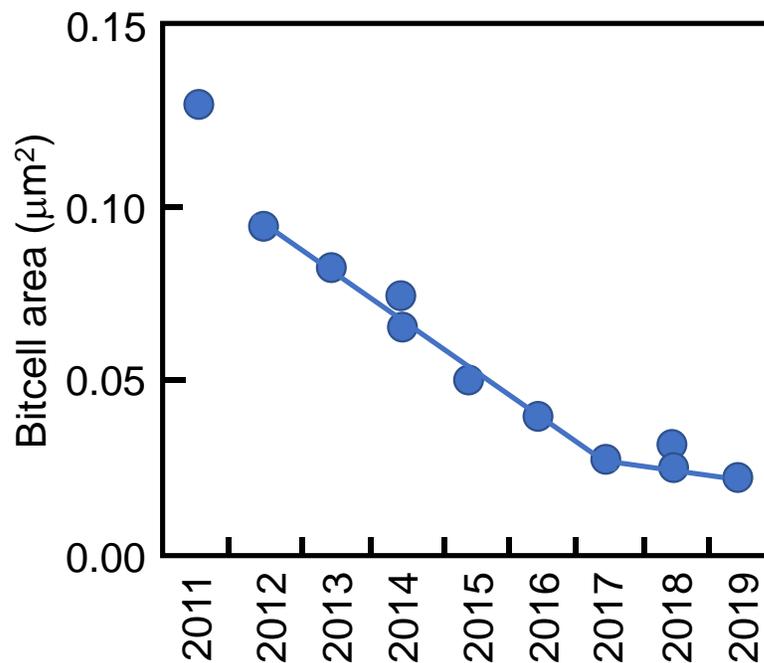
Fig. 18 SRAM bitcell area trend [71].
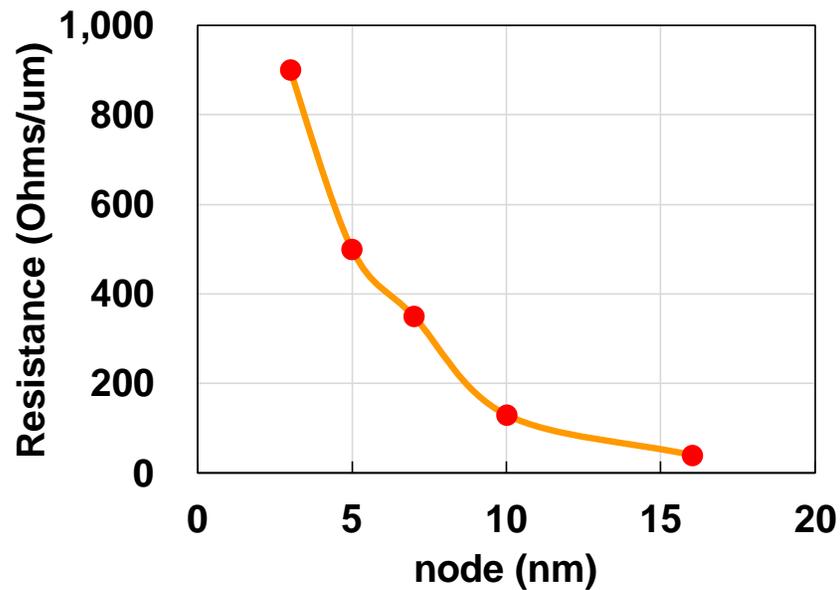
Fig. 19 Trend of tight interconnect wire resistance [72].

2.7. Summary

The basic SRAM macro operation principle was explained: two margins of SNM and WM, with their conflicting behaviors, are necessary for SRAM bitcell operating. Balancing these margins year-by-year is becoming difficult because of variations of transistor characteristics. Four issues of SRAM were explained. They are low power, multi-porting, testability, and high density. The respective challenges presented by these issues are explained in chapters 3–6.

Chapter 3   Power Reduction Technique for SRAM

In this chapter, techniques to reduce the power consumption of SRAM macros are introduced.

The power gate technique, which cuts off power supplies, is known as a technique to reduce the standby power of standard logic gates such as NAND and NOR. However, SRAM macros that must retain memory data cannot be cut off power supplies. The resume standby technique is reportedly extremely effective at reducing the standby power of SRAM macros while maintaining memory data [4].

Fig. 20 shows the leakage current of an inverter: the PMOS on and the NMOS is off. The leakage current has three components. The first is a sub-threshold leak between the drain and source of the NMOS (I_{sub}). The second is a GIDL current (GIDL) between the drain and body of the NMOS (GIDL). The third is the gate leakage between the gate and body of the PMOS (I_{gate}). V_{SB} is 0 V, V_{DB} is VDD, V_{GS} is VDD, and the voltage of "OUT", which is the output node of the inverter, is VDD. Next, the resume standby technique is applied for the inverter. The source nodes of the PMOS and NMOS are separated from each body, the source of the PMOS is lowered V_2. Also, the source of the NMOS is raised V_1. Then, V_{SB} becomes V_1; I_{sub} is reduced by the body effect. V_{DS} and V_{GD} become VDD-V_2, GIDL and I_{gate} are decreased respectively by weakening of the electric field. Furthermore, although the voltage of OUT becomes VDD-V_2, it is treated almost as a high level if V_2 is sufficiently lower than the logical threshold voltage.

Fig. 21 also presents the components of leakage current in 6T SRAM bitcell. The VDD source (ARVDD) and VSS source (ARVSS) of the bitcell are separated from each body. Similarly to the inverter in Fig. 20, the leakage current of the cross coupled inverters is reduced by biasing ARVDD and ARVSS. In addition, the sub-threshold leakage and GIDL of pass-gates are also reduced by lowering BT and BB. All bitcell leakage components are reduced while maintaining memory data. If the voltage difference between ARVDD and ARVSS (cell bias) is too small, then the bitcell might not retain the memory data. The appropriate cell bias differs for each bitcell. Designing the appropriate bias circuits is necessary for each bitcell in each process.
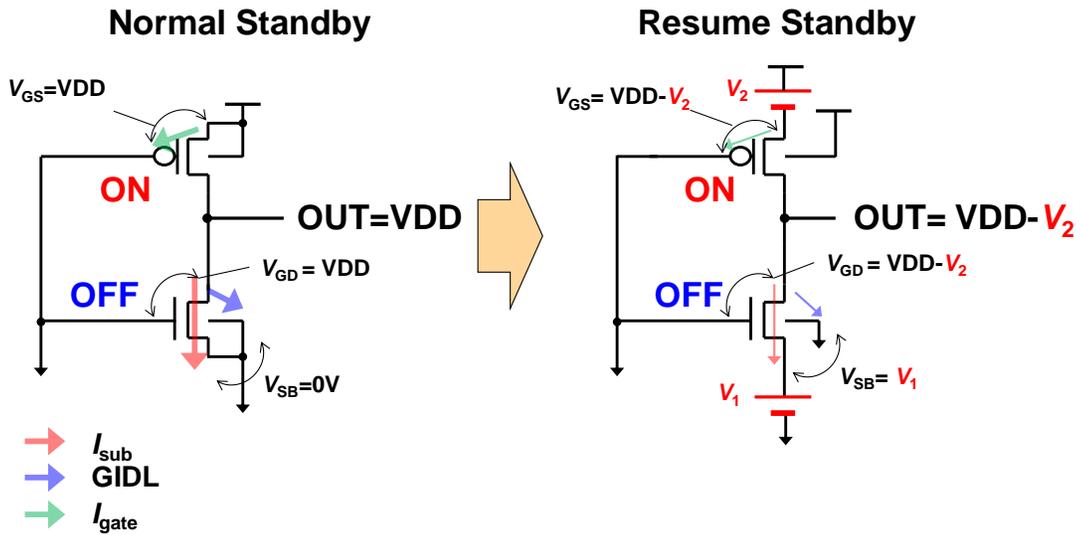
Fig. 20 Leakage reduction for an inverter by resume standby.
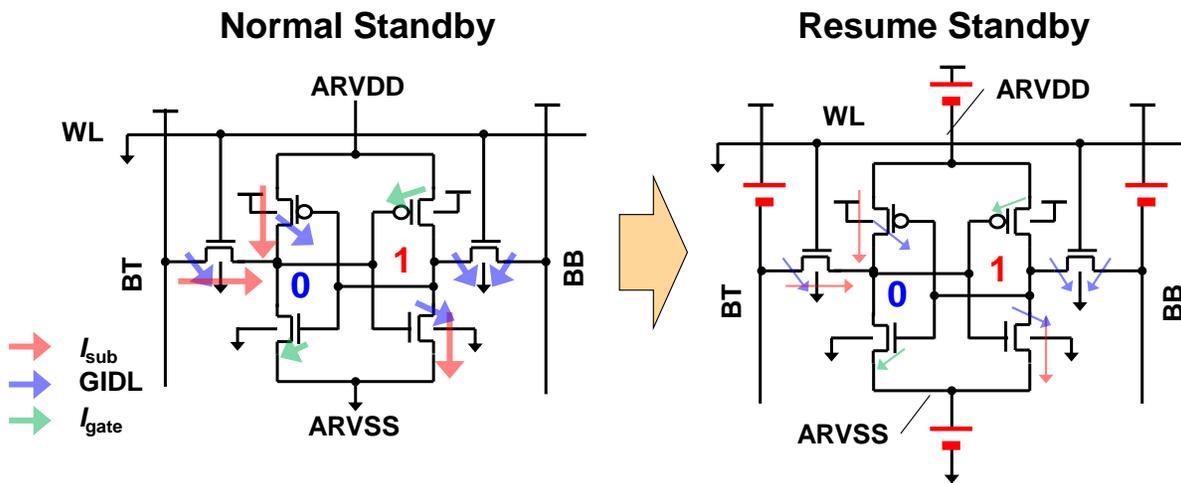
Fig. 21 Resume standby for 6T bitcell.

Furthermore, if the voltage of power supply is large, then not only the standby power but also active power should be reduced for reducing the power throughout an MCU. The major component of active power in SRAM macro is the charge–discharge current by an SRAM bitcell array. As described above, a unique bitcell is selected by the row address and the column address. Fig. 22

presents a circuit diagram of a MUX4 bitcell array. The column address selects 1 BL from 4 BLs. All bitcells selected by WL discharges BLs. The discharging current by unselected columns is excessive power consumption because they are unused.
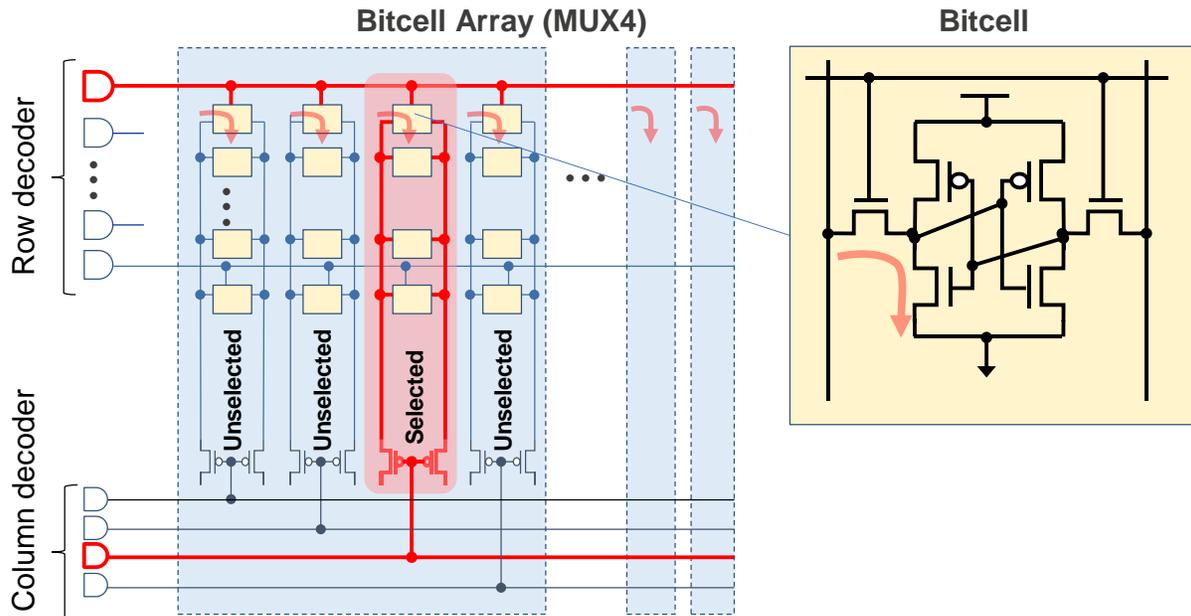
Fig. 22 Excessive current with unselected columns.

The following chapters present three actual examples that illustrate reduced standby power by application of the resume standby techniques.

The first is an SRAM macro for the automotive MCUs. 74% standby power is reduced at the high temperature of 170°C. In addition, a test screening circuit to prevent the retention failure which occurs as a side effect of the resume standby is demonstrated. The second is an SRAM macro for industrial MCUs. Actually, 75% standby power is reduced at 150°C. A test screening circuit is used to prevent an increase of the test time for the retention test at low temperatures. The third is an SRAM macro consisting of 3.3 V thick gate transistors for IoT MUCs. 98% standby power is reduced by the resume standby. In addition, 60% active power is reduced.

3.1. 40-nm Automotive MCU SRAM for embedded Flash MCU

3.1.1. Introduction

Embedded flash MCUs are widely used in industrial applications because of their flexible programmability at low power within an extremely small area [2]. For automotive applications, they are strongly required for high reliability (extremely low failure rate and long lifetime) under severe environments. Especially for an electronic (engine) control unit (ECU) in a power-train control module of a vehicle, MCUs have to operate safely at high temperatures of 150°C, whereas typical consumer products support up to 125°C. In high-temperature conditions, reducing standby power is strictly important to prevent thermal runaway. Furthermore, to ensure robust operation as a fail-safe if an emergency such as overheating occurs because of a broken fan belt, it is necessary to operate at 170°C for short periods. One candidate for embedded nonvolatile memory in advanced technology nodes is split-gate cell technology with charge-trapping structures of metal–SiO_2–$SiNx$–SiO_2–Si (MONOS) [3], which enables high-temperature operation and high reliability with zero standby power. However, the number of writing data in MONOS is limited because of its endurance performance. Typically, some SRAM macros working as cache/buffer memories are also embedded in a MCU combined with the MONOS macro.

Full-CMOS 6T SRAM bitcells are widely used not only for MCU but also as system-on-chips (SoCs) in advanced CMOS technologies. Many leakage reduction techniques have been reported [4][13][20], but almost all are targeting to reduction of the standby power at the waiting mode at room temperature or in conditions for mobile applications. A leakage reduction technique at the operating mode under high-temperature conditions for high-performance processors has been reported [5], but it is not supported for operation at high temperatures over 150°C. Degradation of the minimum operation voltage (V_{min}) of the embedded SRAM is another critical issue related to high-temperature operation. At temperatures over 125°C, the worst static-noise margin (SNM) in half-select bitcells during read/write operation worsens rapidly. Then it is necessary to enhance the SNM for automotive applications.

3.1.2. 6T SRAM bitcell optimization

The 6T SRAM bitcell is optimized by changing both the gate length and gate width for each pull-up (PU) PMOS, pull-down (PD) NMOS, and pass-gate (PG) NMOS to enhance the static-noise-margin (SNM) and write-margin (WM) and to reduce the standby leakage in high-temperature conditions. A process technology is also optimized by tuning the threshold voltage (V_t)

and gate oxide thickness (T_{ox}) with modification of MOS device structures for low-leakage MCU applications.

Fig. 23 shows a schematic of the typical 6T SRAM bitcell that is widely used in MCU and SoC. The normalized sizing of each MOS and normalized T_{ox}, bitcell area, and normalized leakage power at a typical process condition (TT: typical NMOS, typical PMOS) at 25°C are also shown in Fig. 23. Our target of proposed SRAM bitcell is reduction to 1/10 of the conventional typical 6T SRAM bitcell within 10% area overhead.

Fig. 24 presents temperature dependences of the leakage power, SNM, and WM for conventional and proposed 6T SRAM bitcells with consideration of the worst local variation (5.7 sigma standard deviation) in 40 nm technology. Therein, WM is given by the N-curve current metric [6]. The SNM of the conventional SRAM bitcell is less than 0 mV at temperatures over 150°C, where stored data would not be held anymore, although the proposed optimized 6T SRAM bitcell has sufficient SNM even at extremely high temperatures above 170°C. The WM is degraded by optimizing bitcell, but the proposed bitcell still has sufficient margin in the whole temperature range from -40°C to 170°C.

	Conv.	Prop.	Diff (Prop./Conv.)
PU (W/L)	0.5/1	0.6/1.5	+20%/+50%
PD (W/L)	1.5/1	1.5/1.5	+-0%/+50%
PG (W/L)	1/1	1.2/1.5	+20%/+50%
Tox	1	1.2	+20%
Area	1	1.1	+10%
Leakage Power @typical25 °C	1	0.1	-90%
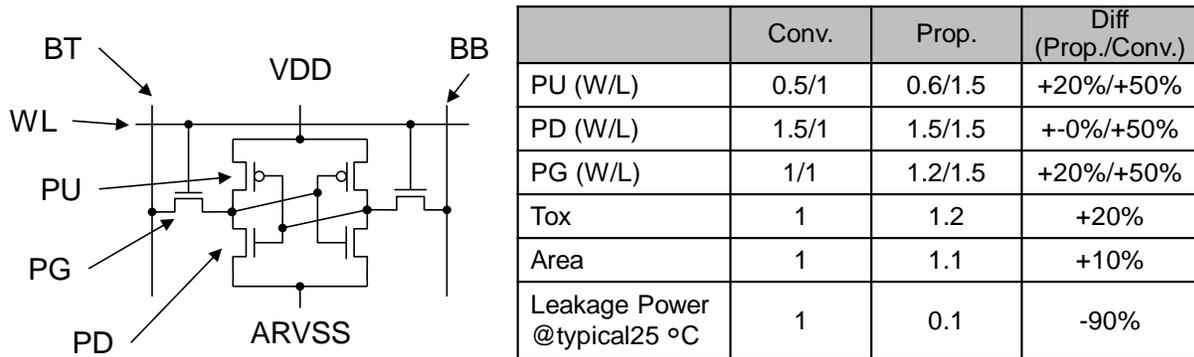
Fig. 23: Optimization of the 6T SRAM bitcell (normalized).
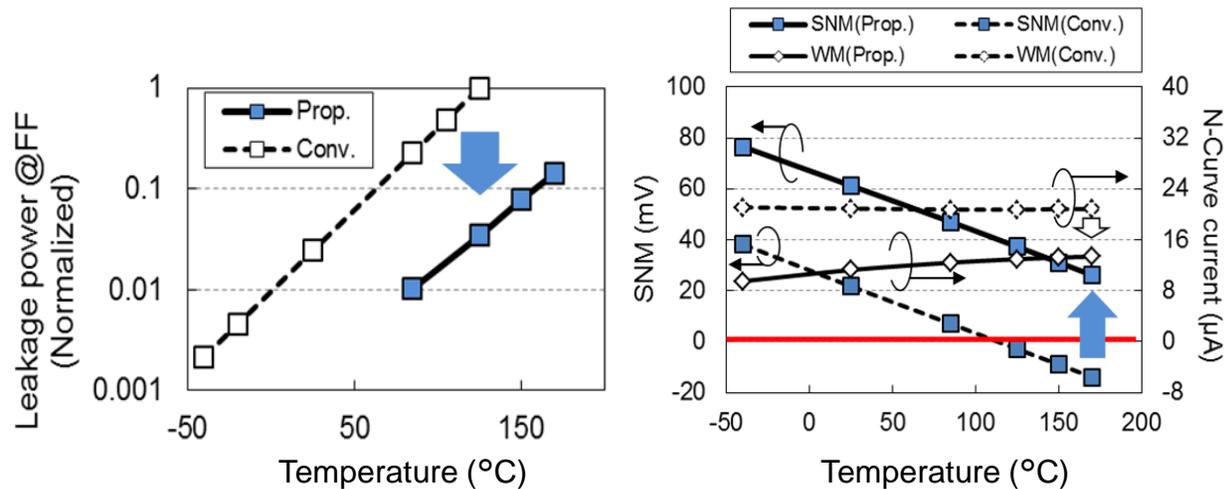
Fig. 24: Temperature dependence of the static-noise-margin (SNM) and write margin (WM) as defined by the N-curve current.

3.1.3. Leakage reduction circuits at high temperatures

The resume standby technique can reduce standby power of SRAM bitcell with keeping memory data as beginning of this chapter. Three biases, 1) floating bitline [9], 2) lowered VDD [10], and 3) raised VSS for cell-array (ARVSS) can be applied for SRAM bitcell [10]. Fig. 25 shows the estimated leakage power (normalized) by SPICE simulation of the optimized 6T SRAM bitcell at the worst process corner (FF), 1.35 V maximum supply voltage and 170°C. The leakage powers of 1)–3) can be reduced respectively by 33%, 51%, and 46%. Combined with these three reduction techniques, the leakage power can be reduced remarkably by 78%.
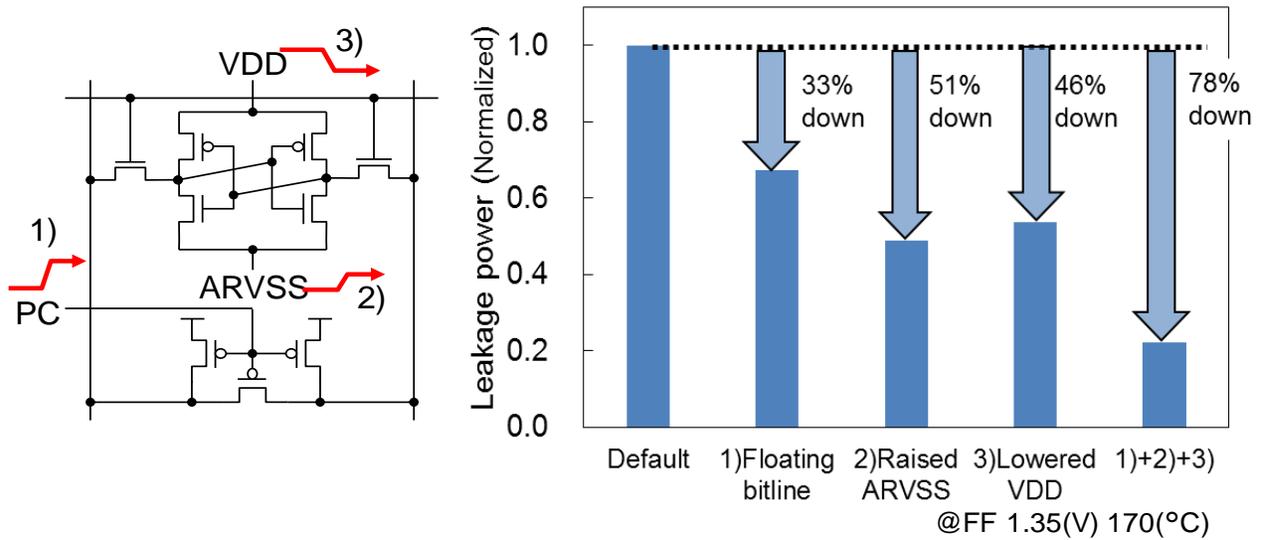
Fig. 25 Leakage reduction of the proposed 6T SRAM bitcell by SPICE simulation at process-worst (FF), 1.35 V maximum voltage, and 170°C.

Fig. 26 depicts the power saving modes of the SRAM macro. This study assesses two SRAM types: the retention SRAM macro with separated power supplies for the cell-array (RAMVDD) and peripheral (VDD) individually, and the standard SRAM macro with a common supply power (VDD) for both the cell-array and peripheral. The retention SRAM and standard SRAM have a header PMOS switch and a footer NMOS switch for peripherals, and a footer NMOS switch with a diode-connected NMOS for the cell array.

Each SRAM macro has three power-saving modes, the resume standby (RS) mode, the resume standby with low lowering supply voltage (RSLV) mode, and the deep standby (DS) mode.

In the RS mode, the header PMOSs and footer NMOSs for peripheral and footer NMOSs for cell-array are entirely cut off, but the power supplies of VDD and RAMVDD are still given by on-chip-regulators, that is Voltage Down Converters (VDCs). To reduce the leakage power further, the voltage level of VDD and RAMVDD are lowered by controlling the VDCs in the RSLV mode. In DS mode, the voltage level of VDD given by VDC is forcing to 0 V to cut off leakage perfectly. The standard SRAM macro cannot hold stored data in the DS mode, although the retention SRAM macro can hold the stored data because the RAMVDD is supplied by the individual VDC.
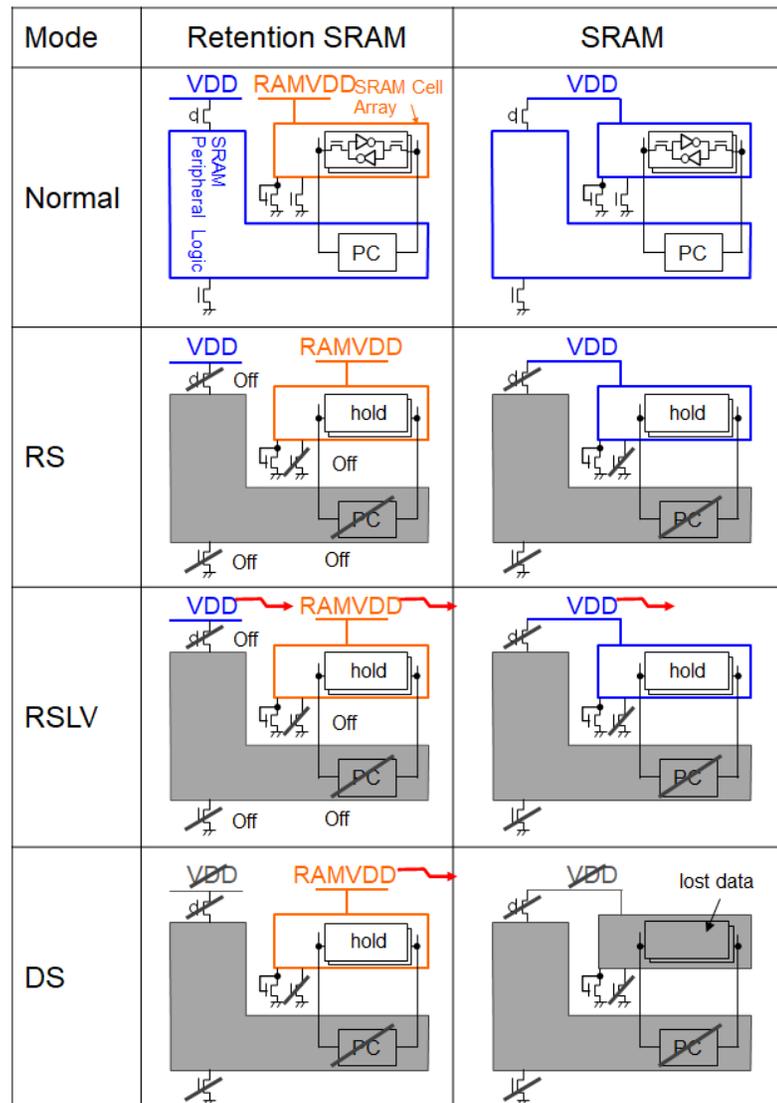
Fig. 26 Power saving modes of SRAM macros of two types.

3.1.4. Practical 160 kbit SRAM macro design with rush current suppressor and a test circuit for screening weak retention bitcells

Floating bitline can reduce the leakage current of bitcell. However, it has a demerit regarding to the rush current. The rush current might cause significant IR-drop, and have negative effects on memory data of other SRAM macros or analog circuits. Fig. 27 shows the circuit diagram of the proposed SRAM macro with a rush current suppressor (RCS) and a test circuit for screening weak

retention bitcells. At the transition from the normal operation mode to the RS mode, the bitline charger PMOSs are turned off immediately, the voltage levels of the bitlines are balanced by all stored data (0/1) of SRAM bitcells in a column, and the power switch footer NMOS for ARVSS is cut off. The voltage level of ARVSS is raised by leakage current of SRAM cells and balanced on-current of the diode-connected footer NMOS. However, at resumption from RS mode to the normal mode, first weak pre-charger PMOSs are turned on and the footer NMOS for ARVSS is turned on. Then after some time, full-charging with large PMOS is executed. Therefore, the rush current peak can be suppressed.

Fig. 28 shows simulated waveforms of the power supply current when the state of the SRAM macro transits from the RS mode to the normal mode. The first row shows how the bitlines are charged to the VDD level with and without the proposed RCS. Without the proposal, the bitlines are charged by the general pre-charge transistors, which have sufficient drivability for charging bitlines within the operating cycle. Therefore, the bitline slope is steep. However, with the proposal, the bitlines are charged by transistors dedicated to charge them at the resumption from the RS mode. Since these transistors are designed to be weak, so the slope is gradual. The second row shows the power current with and without the proposed modification. Using RCS, the peak current is reduced from 198 mA to 38 mA: an 80% reduction. The area overhead of RCS is 120.087 μm^2, which is 0.15% of the SRAM macro size. It is necessary to reduce the current peak by inserting a delay between signals to connect the RS control terminal of each SRAM instance if many SRAM instances are embedded in a chip.

Although the leakage power of an SRAM cell can be reduced by a floating bitline, it might cause retention failure. If the pass gate transistor of the SRAM cell has a leakage path to bitline as shown in Fig. 29, the floating bitline might cause retention failure that cannot be screened using ordinary tests, depending on the resistance of R1 in Fig. 29. Because the balanced voltage of floating bitlines differs depending on the temperature, source voltage, and data written, it is difficult to test under worst conditions. The potential of this failure should be terminated for the quality of the automotive MCUs. To screen out this retention failure, the SRAM macro has a circuit for the retention test mode. Using this mode (when the TEP signal shown in Fig. 27 is high), testing can be executed with the bitlines discharged to 0 V, which is almost the worst condition to keep MT=1 in Fig. 29.

Fig. 30 shows the effectiveness of the proposed retention test. If the resistance of R1 in Fig. 29 is 20–40 kΩ, then the data flip time as shown in Fig. 30 will be longer than 2.0 ns. The fault can be screened only by the proposed retention test. The failure can be screened by an ordinary function test if the R1 resistance is less than 20 kΩ. However, if the resistance of R1 is greater than 40 kΩ, then the SRAM macro will function correctly, therefore it is not failure.
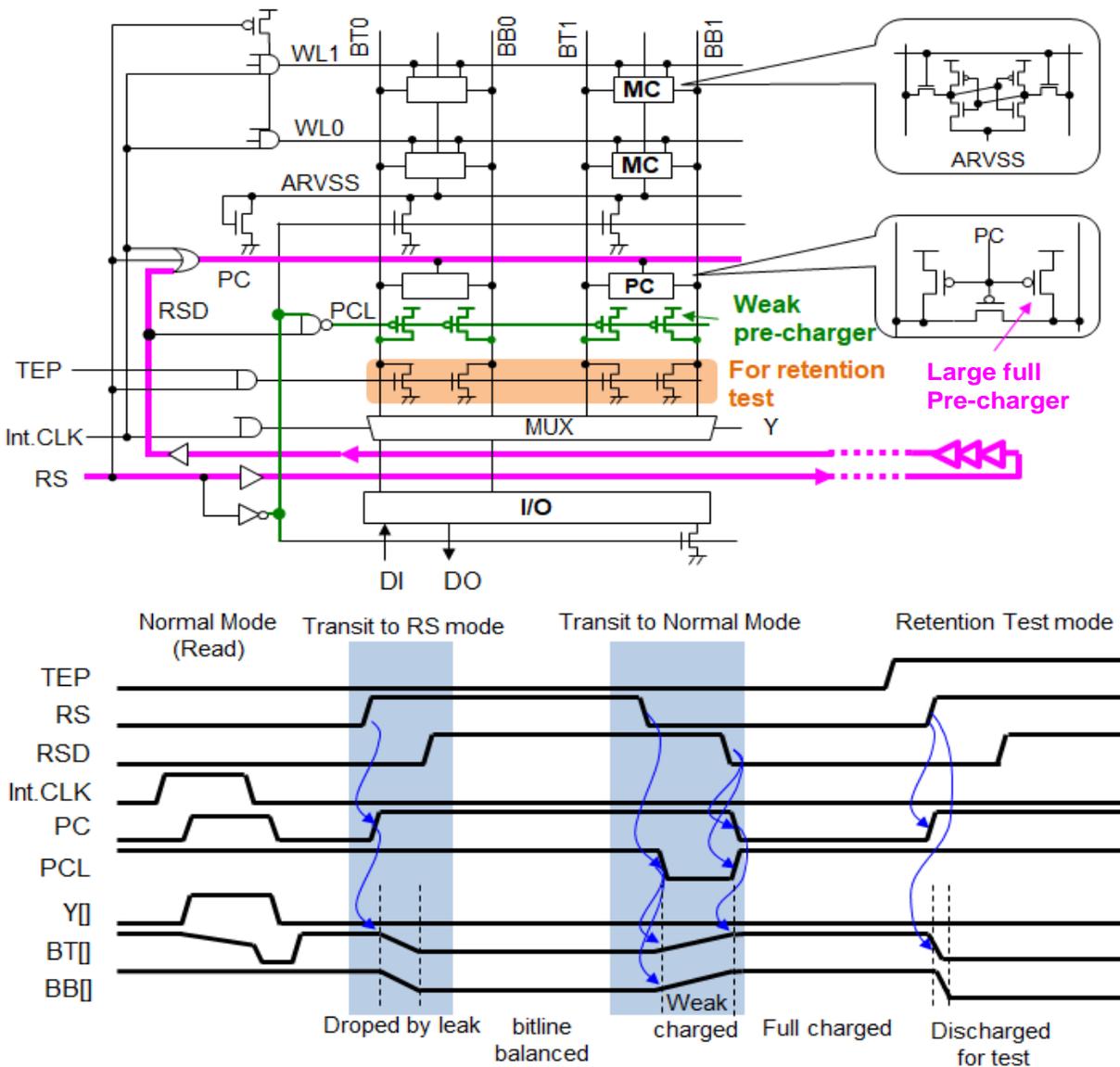
Fig. 27 Circuit diagram of proposed SRAM macro and transition timing chart.
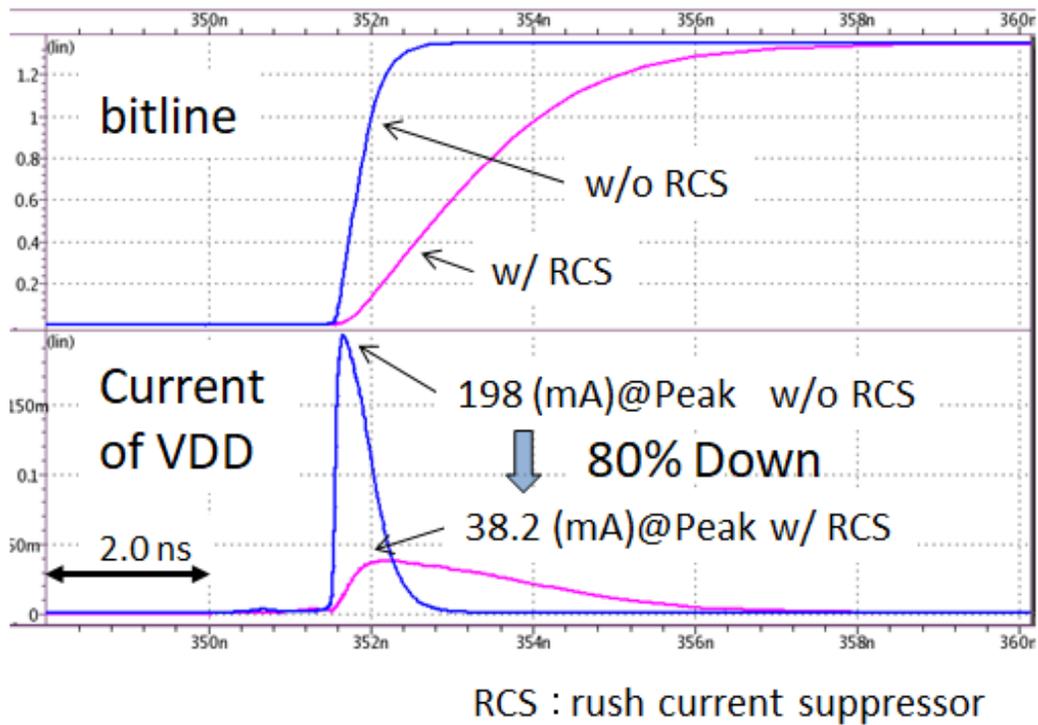
Fig. 28 Simulation waveforms of rush currents with and without a rush-current-suppressor (RCS) for 160 kbit SRAM macro at process-worst (FF), 1.35 V maximum supply voltage and 170°C.
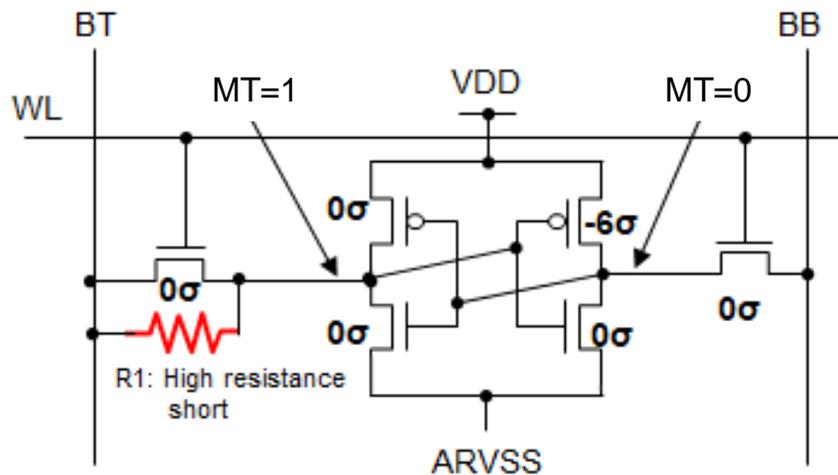
Fig. 29 Fault model to be detected by a weak-retention screening circuit.
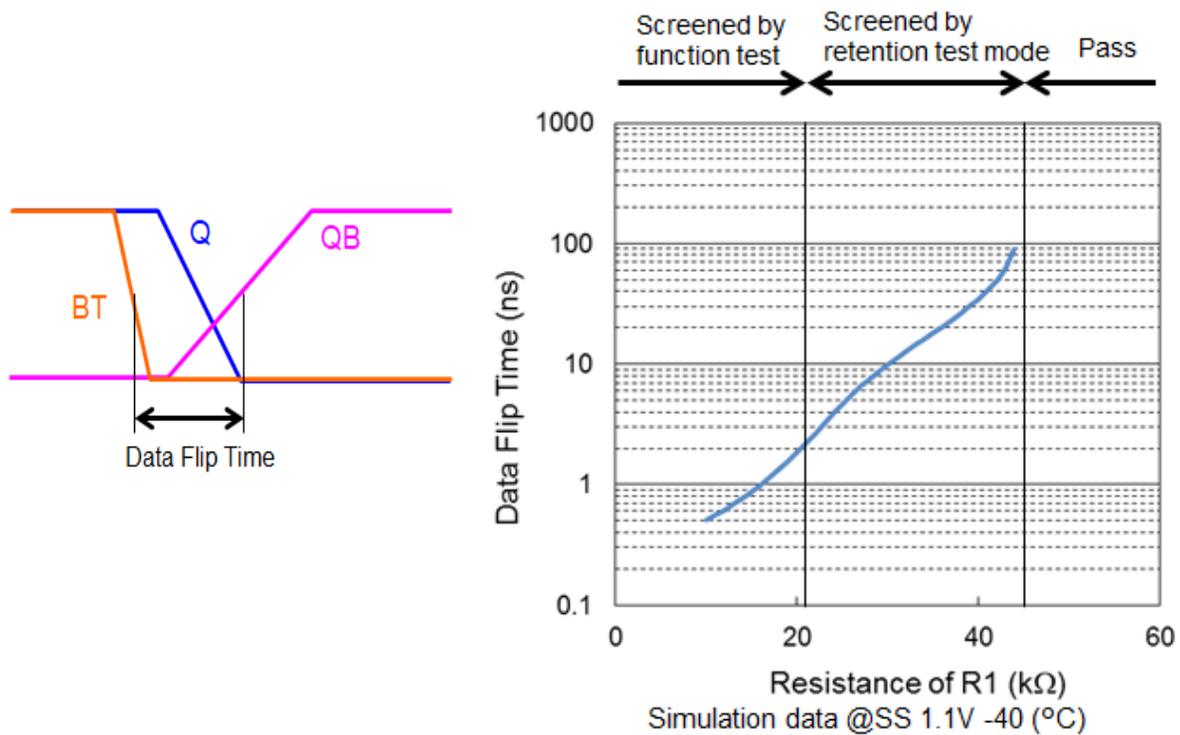
Fig. 30 Effectiveness of the retention test (TEP mode).

3.1.5. Test chip implementation and measurement results

Fig. 31 shows a die microphotograph of a test chip using a 40-nm embedded flash technology [3].

Table 1 presents test chip features. Eight 160-kbit retention SRAM macros are implemented in the test chip (completely 1.25-Mbit). The test chips were observed full read/write functions in the wide temperature range from -40°C to 170°C. Fig. 32 presents a typical Shmoo plot at 170°C, showing V_{min} vs. the access time. The access time of the measured 160-kbit SRAM macro is 2.14 ns at the typical supply voltage of 1.25 V. The read/write operation at 0.6 V with 16 ns access time was confirmed. Fig. 33 shows the measured V_{min} of the 1.25-Mbit SRAM macros in the test chip. The measured sample dies are 239/239/4 at 25°C/150°C/170°C, respectively, where good distributions and sufficient operating margin to the 1.25 V typical supply voltage were observed. All dies are fabricated in an almost typical process condition. Results show that the V_{min} at high temperature of 150°C or 170°C is lower than the V_{min} at 25°C, as expected, because the bitcell is designed to be better V_{min} of the SNM rather than that of WM in the typical process condition, even at a high temperature, to ensure aging degradations of the SNM.

226.6 (µm)

343.3 (µm)

Retention SRAM
160 kbit

1.25 Mbit(160 kbit x 8)
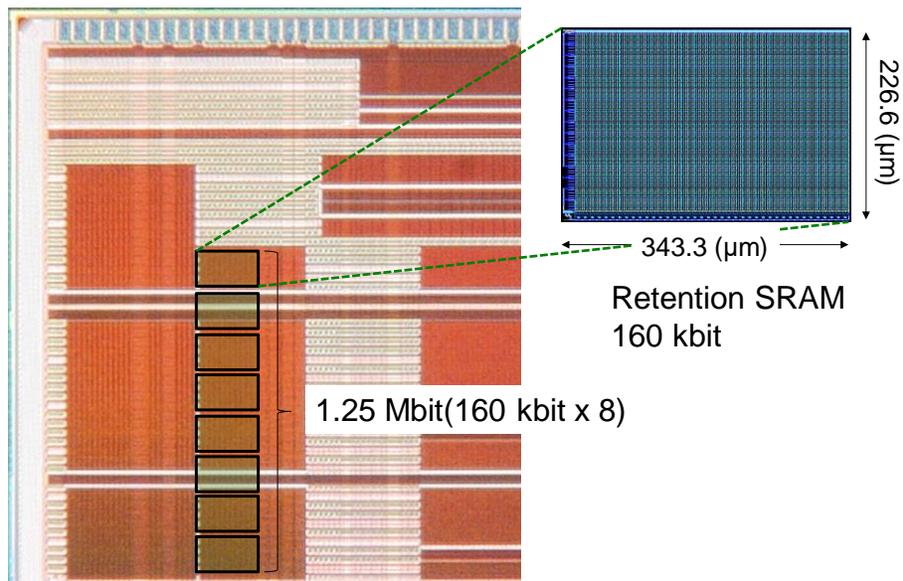
Fig. 31 Microphotograph of a test chip taken using 40-nm embedded flash CMOS technology.

Pass

Access time is **2.14 (ns)**
at 1.25 (V) (Typical voltage)

Fail

VDD (V)

Read access Time (ns)
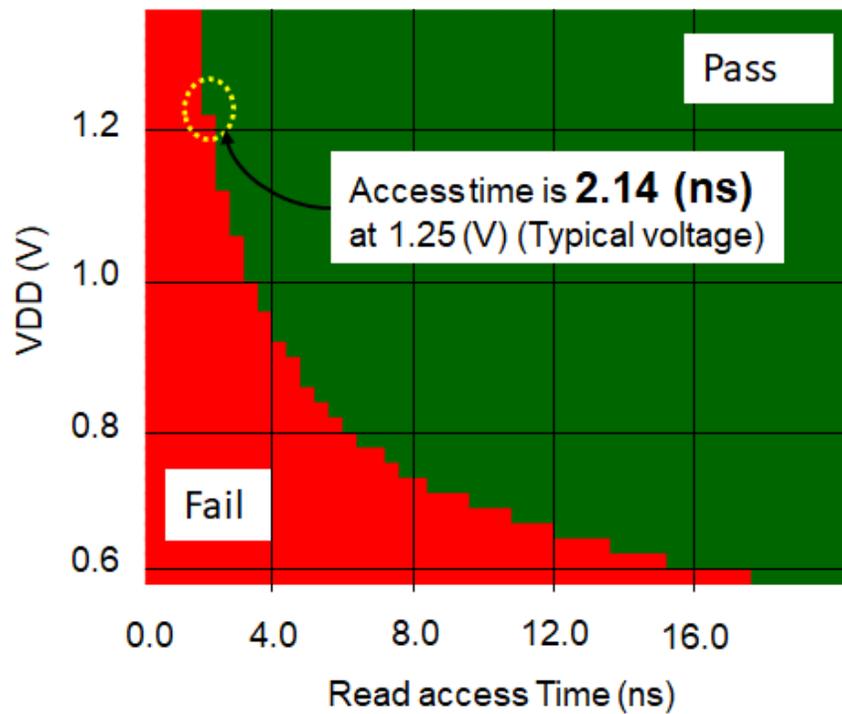
Fig. 32 Typical Shmoo plot of 160 kbit SRAM macro at 170°C.

Fig. 34 shows cumulative distribution plots of the normalized leakage power to 1-Mbit at 25/150/170°C. The median values of the measured leakage powers at 25°C (170°C) are 17.30 μW (2518 μW), 2.28 μW (682 μW) and 1.86 μW (643 μW) at the normal mode, the RSLV mode and the DS mode, respectively.

Table 1 Test chip features

	Features
Technology	40-nm embedded flash process [2]
Macro configuration	160 kbit (4-kw × 40-bit)
Macro size	343.3 μm × 226.6 μm (77801 μm^2) @160-kbit
Total capacity	1.25 Mbit (160-kbit × 8)
Bit density	2.106 Mbit/mm^2
Leakage Power	17.3/2.28/1.86 μW/Mbit (Normal/RSLV/DS) @typical 25°C
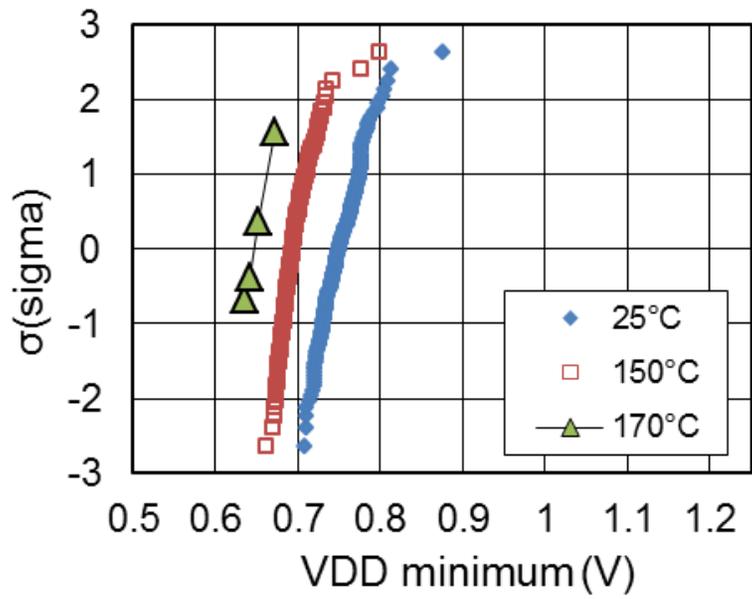
Median Value

170°C	0.647 (V) (N=4)
150°C	0.694 (V) (N=239)
25°C	0.750 (V) (N=239)

Fig. 33: Measured VDD minima of 1.25 Mbit (160 kbit × 8) SRAM.
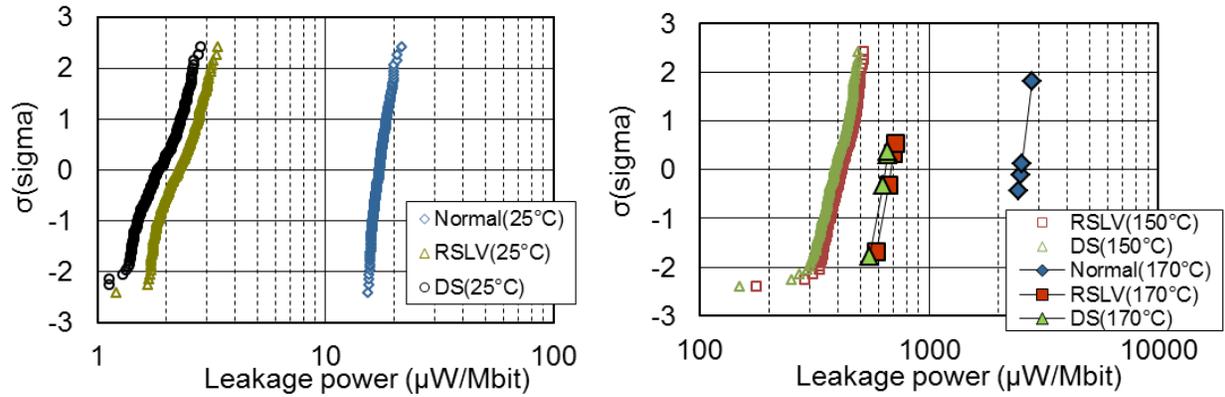
Median Value (µW/Mbit)			
Mode	25 °C(N=248)	150 °C(N=248)	170 °C(N=4)
Normal	17.30	1409	2518
RSLV	2.28	417	682
DS	1.86	389	643

Fig. 34: Measured leakage powers normalized to 1 Mbit.

3.1.6. Estimated SRAM leakage power for prototype MCU

Fig. 35 portrays a block diagram of the prototype chip, which is a Flash MCU designed for automotive applications. Retention SRAM demands a large capacity as cache for flash macro. The number of writing data of flash macros will reach the limit if flash macros are written every time the engine stops. On the other hand, SRAM macros without retention (total 4.3 Mbits) include high-speed SRAM as a CPU-cache memory. The three on-chip regulators (VDC) create three core-voltage powers, VDD, RAMVDD, and VDDsys. The power of the retention SRAM is supplied with RAMVDD, which is always-on. Therefore, the retention SRAM data are always held. Fig. 36 shows the simulated leakage power of the prototype chip, in the Normal/RSLV/DS mode, of the SRAM (total 11.1 Mbit required by the prototype MCU) and VDC embedded in the prototype chip. During the DS mode, it does not consume leakage power except SRAM and VDC. The leakage power can be reduced 70% at the RSLV mode, and 88% at the DS mode.
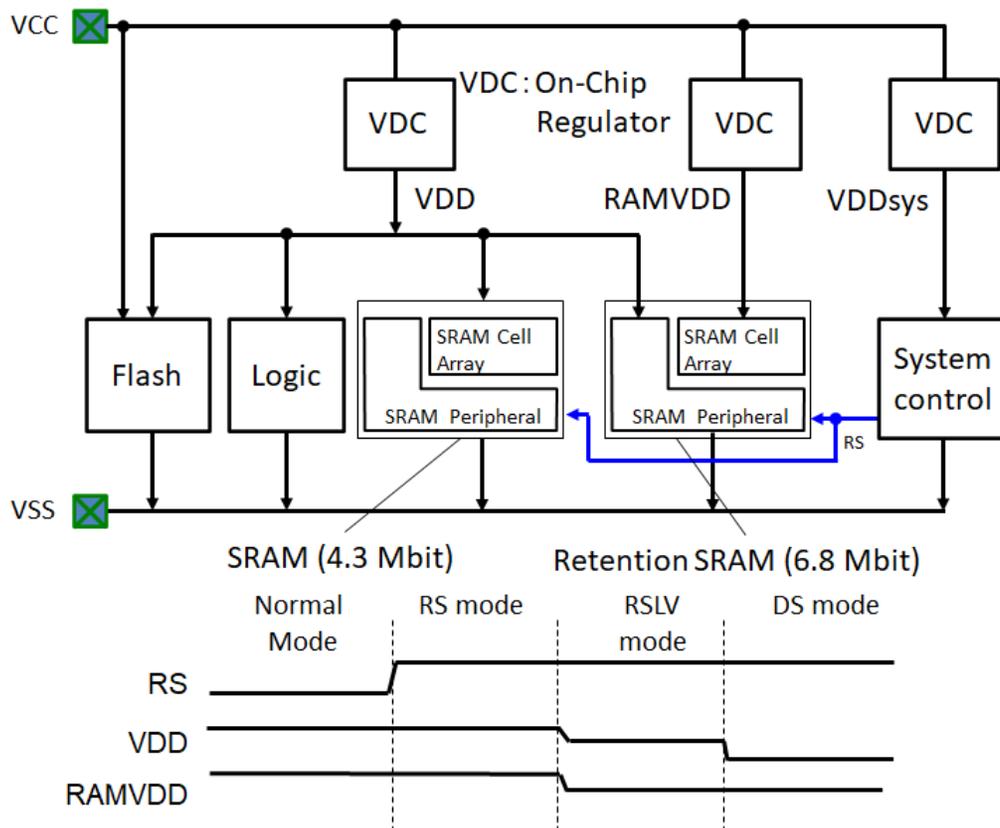
Fig. 35 Power structure of a prototype MCU chip.
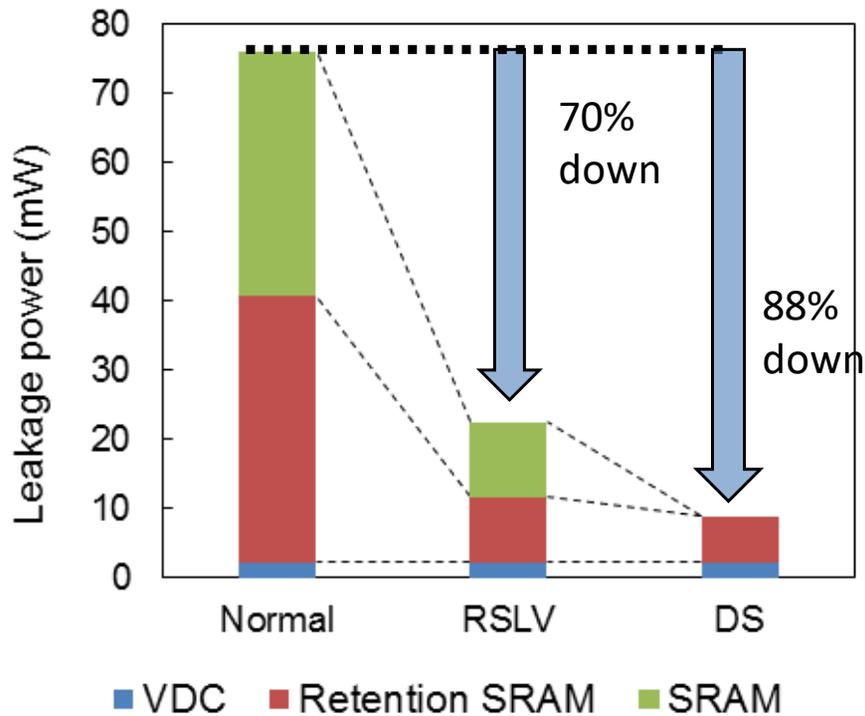
Fig. 36 Simulated leakage power of SRAM and VDC in prototype MCU chip at leakage-worst process (FF), 1.35 V, and 170°C.

3.1.7. Conclusion

Some leakage reduction techniques, newly proposed rush-current suppression circuit (RCS), and a retention test circuit were explained. Results show that raising the source of the pull-down MOS and floating bitlines engenders drastically reduces of memory-cell leakage. Separating the memory-cell power line from the peripheral one enables the completely cutoff the peripheral leakage. Moreover, the proposed RCS was demonstrated to restrain the peak current at the transition from the RS mode. A 160 kb SRAM macro able to operate at 170°C with the proposed circuits was designed and fabricated using a 40 nm embedded flash process. Results show that leakage power in the DS mode of 1.86 μW/Mbits at 25°C and 643 μW/Mbits at 170°C were achieved, respectively. At 170°C, 74% power reduction was achieved compared with the normal mode.

3.2. 110-nm Consumer MCU SRAM

3.2.1. Introduction

Internet-of-Things (IoT) devices are demanded both low active power and low standby power in the sleep mode with cost effective solutions. Especially, reducing a total Bill of Materials (BOM) cost is much important in low-end SoC/MCUs for consumer/industry IoT applications. Those devices are typically prepared on mature technology nodes around 90 nm to 180 nm for reducing manufacturing cost. It is because analog blocks are dominant, not needed huge core logics and embedded SRAMs. In addition, it tends to be required at 150 °C high-temperature operation and standby even for non-automotive products because of the cost reduction by heat sink/fan less solutions.

Meanwhile, the testing cost per transistor increases year by year [15]. Reducing the test cost is also an important factor in such low-end SoC/MCUs. Fig. 37 depicts a diagram of an SoC/MCU chip with built-in self-tests (BISTs) for logic and memory. On memory-BIST for SRAMs, techniques to screen out failures accurately and to shorten test time have already been reported [16] - [18]. However, there is no discussion on reducing the embedded SRAM retention test time for low leakage standby modes.

In this section, an ultra-low standby power embedded SRAM macro with resume standby circuits on a 110-nm SoC/MCU is introduced. Besides, a 2-stage test screening method for the resume standby SRAM is newly proposed. It is applied for only the low-temperature test without high-temperature test in the memory BIST flow to minimize the total testing time without any test coverage loss.
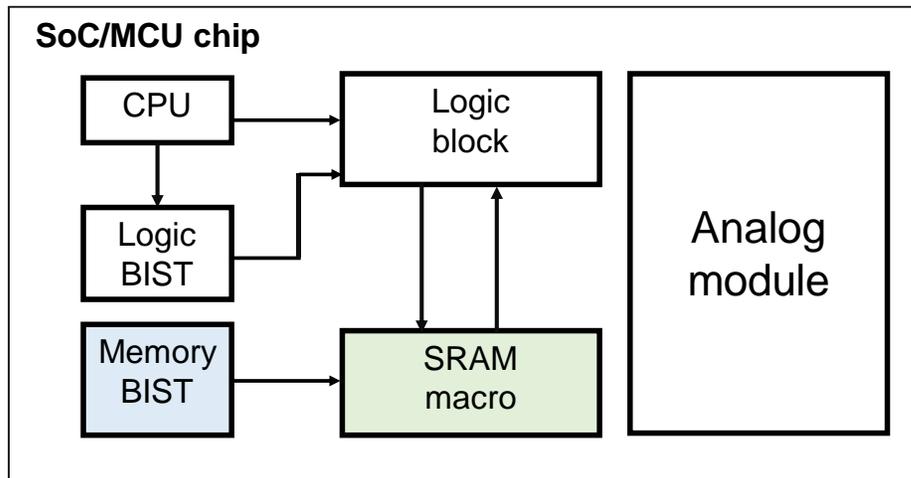
Fig. 37 Embedded SRAM with memory BIST on an SoC/MCU chip.

3.2.2. Resume Standby SRAM on 110-nm

SRAM source bias control techniques in the low-standby mode are effective for reducing standby power [19] - [25]. Fig. 38 (a) shows a 6T SRAM bitcell circuit with source bias control on 110-nm technology. To apply independent bias in the resume standby mode, each power source line (ARVDD) and ground source line (ARVSS) in bitcells is separated with VDD and VSS, respectively. Fig. 38 (b) shows effects of leakage reductions in each bias technique at 25°C and 150°C. Three bias techniques 1)Bitline lowering, 2)ARVSS rising and 3)ARVDD lowering, are applied to reduce much leakage power. In this case, enough cell bias at resume standby mode should be kept. Fig. 39 (a) plots the temperature dependencies of the read static noise margin (SNM) [26]and write margin (WM) defined by write-trip-point [27] by considering with local variations. The worst process condition for SNM is FS (NMOS: fast, PMOS: slow) corner. Whereas the worst process condition for WM is SF (NMOS: slow, PMOS: fast). The Monte Carlo simulation results show that the bitcell has less WM than SNM, but each μ/σ exceeds 5.5, which corresponds no failure bits up to around 500 Mbit. It is found that both are enough margins at typical 1.5 V +/- 10% operation. A mature process usually has no special process step for SRAM bitcells to reduce the manufacturing cost, so the pull-up/pull-down/pass-gate MOSs are same characteristics of core MOSs. The large on-current of the pull-up PMOS is the cause to decrease WM and to increase standby leakage current. That is the reason why ARVDD source bias is applied as well as bitline and ARVSS biases. Fig. 39 (b) plots the estimated minimum operating voltage (V-min) for the resume standby mode with source biases at each temperature of -40/25/150°C. It is found that the retention margin at low temperature (LT) become worse than the high temperature (HT). The data retention test should be done at LT as described later. Fig. 40 shows each test time of embedded SRAM with and without resume standby circuit. Memory BIST proceeds the IDDQ test, read/write functions test, DC stress screening test and data retention test. The functions test takes a short time for small SRAM capacity with typical test clock frequency. Neither DC screening nor IDDQ take significant testing times. Meanwhile, the testing time of data retention with the resume standby circuit is increased due to the long pause time to screening the retention failures. Because, the source bias level will be shifted slightly with long time constant by small SRAM leakage current in the 110nm mature process technology.

(a) Source bias techniques

(b) Leakage current w/ source bias
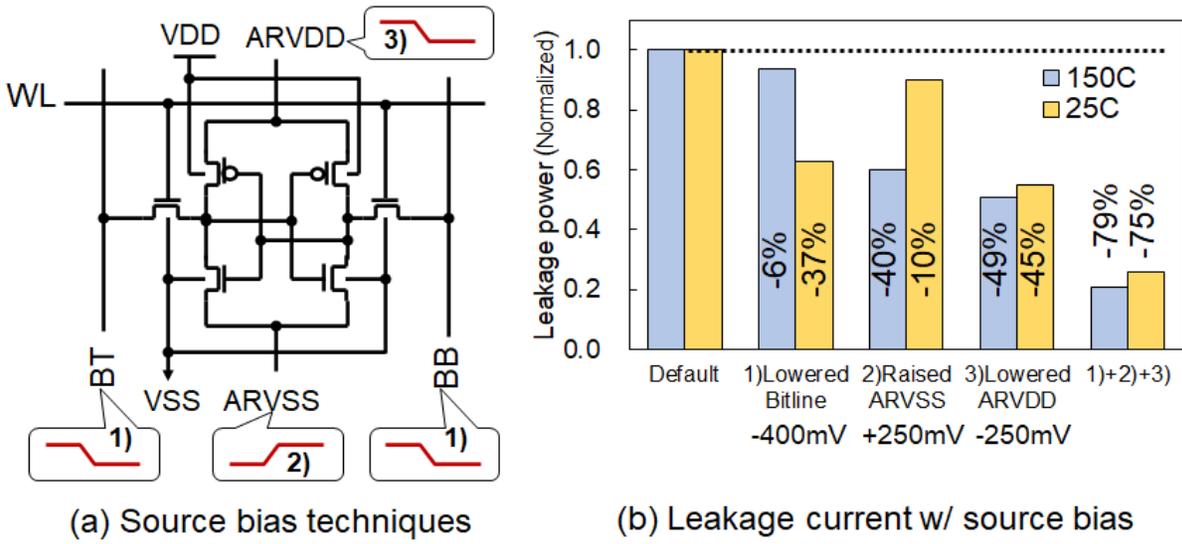
Fig. 38 110-nm 6T SRAM bitcell circuit with source biases and estimated leakage power reduction at 25°C and 150°C by SPICE simulation.

(a)SNM and Write Margin@1.35V

(b)Retention V-min with source bias
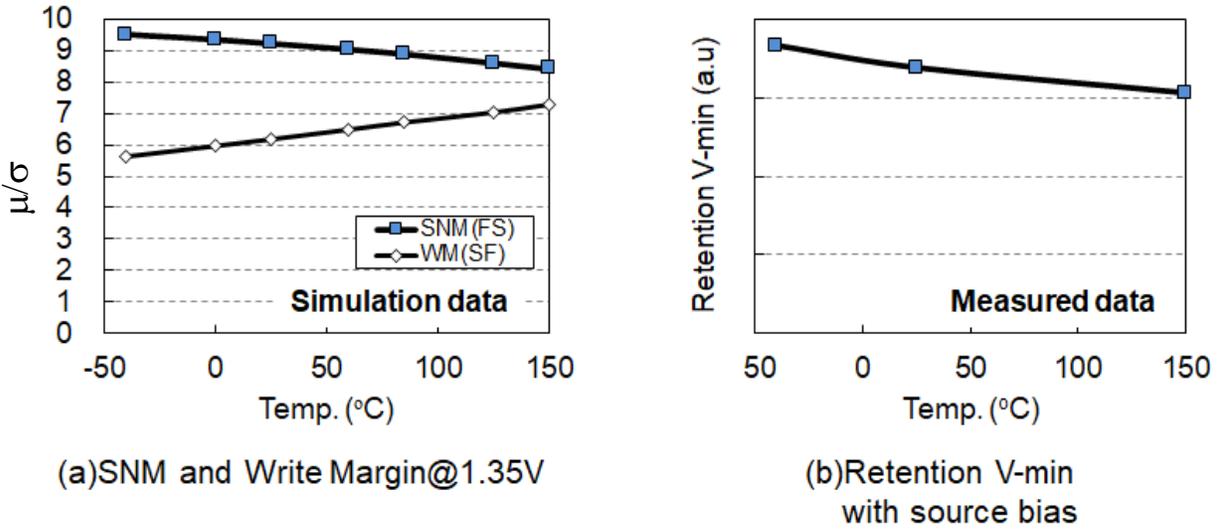
Fig. 39 a) Simulated temperature dependencies of the static-noise margin (SNM) and write margin (WM) at 1.35 V, and b) Estimated retention V-min based on measured data.
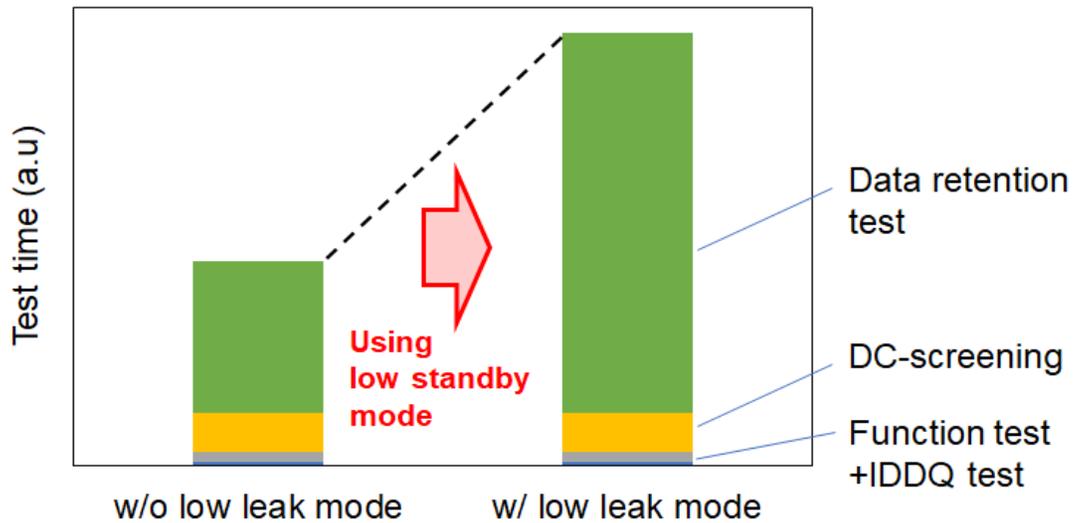
Fig. 40 Each test time of embedded SRAM w/ and w/o resume standby.

The source bias technique with the diode connected MOS shifts the source line voltage by its Vth. In precisely, the voltage of the source line is determined by the balances between the diode connected MOS current and leakage current of the bitcell array. The source bias techniques are applied to both ARVDD and ARVSS to keep the voltage difference between ARVDD and ARVSS (cell bias), which is proportional to the data retention margin. Fig. 41 (a) illustrates the proposed source bias control circuit with 2-stage data retention test mode. The SRAM macro transits to the standby mode with holding stored data if the signal RS becomes high. Both of ARVDD and ARVSS are biased by the circuit to reduce standby power without any on-die regulators. As descried in the previous paragraph, PMOS has large leakage current, therefore it is also important to bias ARVDD to reduce the pull-up PMOS leakage current. However, the bitcell has smaller data retention margin at LT than HT due to the unbalance of leakages as shown in Fig. 39 (b). The always-on NMOS (N_{AO} in Fig. 41 (a)) connected to ARVSS to the circuit in order to keep the cell bias [4]. N_{AO} has a weak drivability which is smaller than bitcell leakage at HT, and is larger than bitcell leakage at LT. Fig. 41 (b) shows the temperature dependency of voltages of ARVDD, ARVSS and cell bias carried out by SPICE simulation. N_{AO} discharges ARVSS to 0 V at LT, then the cell bias is increased with small standby power overhead at HT. Therefore, the proposed circuit can reduce the standby leakage current not only at HT but also at room temperature (RT) and LT with enough cell bias.

(a) Bias control circuit

(b) Voltages of ARVDD/ARVSS
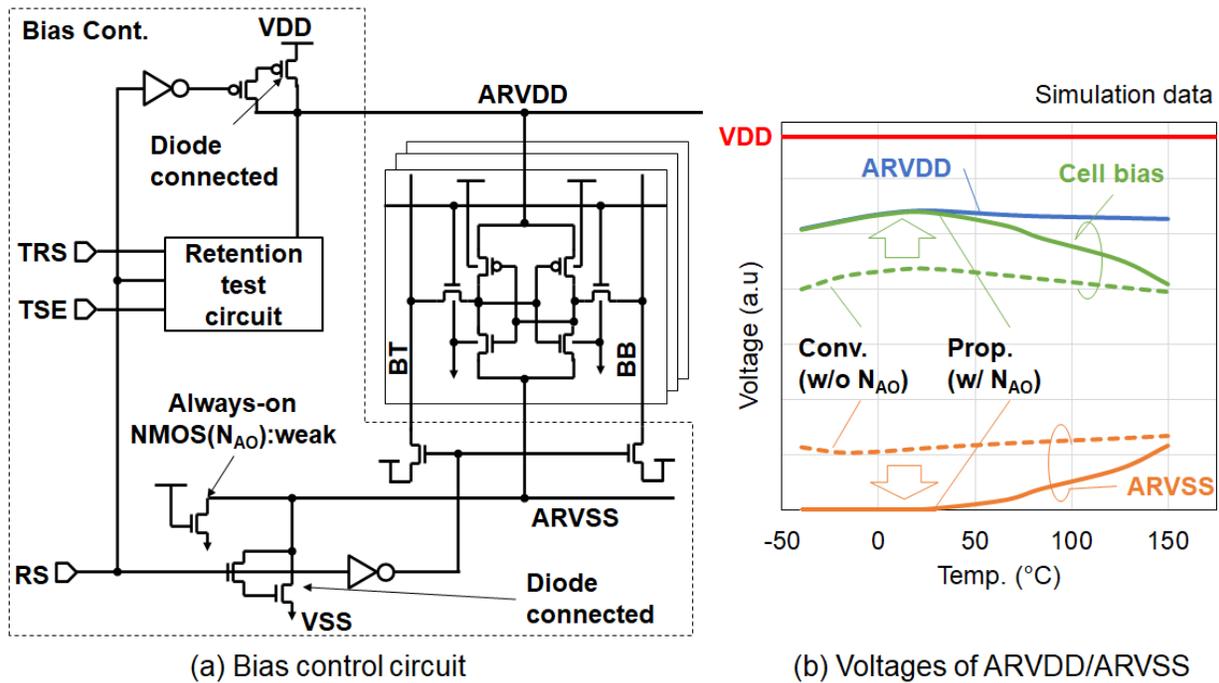
Fig. 41 Circuit diagrams of bias control circuit for resume standby and SPICE simulation results of the cell bias with ARVDD and ARVSS bias levels.

3.2.3. 2-stage Retention Test Circuit

A target SRAM leakage current is less than 1 pA/cell for low-power SoC/MCU applications in 110-nm technology nodes. Besides, due to the large parasitic capacitance of the matured process, discharging ARVDD takes long time at LT. It increases testing cost significantly. Fig. 42 shows the voltage dependency of failure bitcell counts in the retention mode with source bias in cases of 1 sec., 10 sec. and 30 sec. of pause time. The failure count of 10 sec. and 30 sec. are larger than that of 1 sec. It means 1 sec. pause time is not enough to stabilize the ARVDD bias level. If 10 sec. is needed for the retention test, the testing cost becomes over x10 higher.
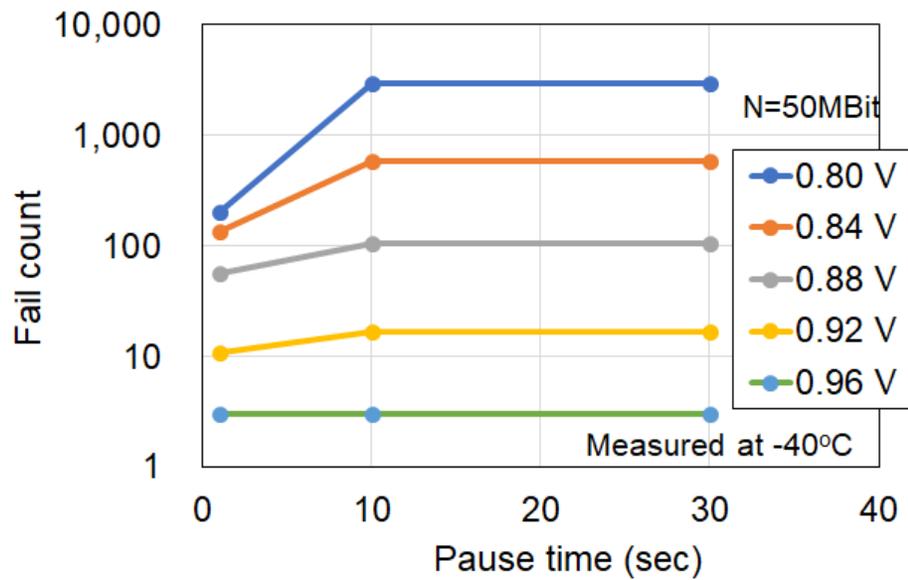
Fig. 42 Pause time dependency of retention test with source bias.

The 2-stage retention test method to reduce pause time is explained. Fig. 43 shows the proposed retention test circuit. The test circuit has a pull-down weak NMOS (N_{PD}). N_{PD} is connected to ARVDD, and is controlled by the additional control pins TRS and TSE. The retention test circuit operates in two modes. One is "strict test", the other is "medium test". Both test modes are controlled by a memory-BIST circuit. When RS is asserted, ARVDD is discharged by the leakage current of the bitcell array, and it takes long time to saturate ARVDD to a certain voltage V_{RET} as shown in Fig. 43. In the strict test mode, N_{PD} immediately discharges ARVDD to the voltage V_{STR} which is lower than V_{RET} and pulls down ARVDD continuously. The strict test might have some over-screened chips that are not failure samples, but it can certainly screen out failure chips with very short pause time. In the medium test mode, when the internal node RPLAVD is charged to Vth of INV_{DE} in Fig. 43, N_{PD} is turned off and stop discharging ARVDD, and then ARVDD bias level saturates to V_{RET}. It is expected that the results of the medium test can well reproduce the true V-min compared to the strict test and the medium test can avoid over-screening. Proposed test flow is shown in Fig. 44. The combination of the strict test and the medium test are applied for the LT test condition, and can reduce the average pause time with preventing over screening. The area overhead of the proposed retention test circuit is less than 0.03% of the SRAM macro.

Fig. 45 shows SPICE simulation waveforms of the proposed SRAM macro at the SS corner and -40°C, confirming the expected operation and bias levels. In this condition, ARVDD in the medium test reaches to V_{RET} earlier than the conventional test.
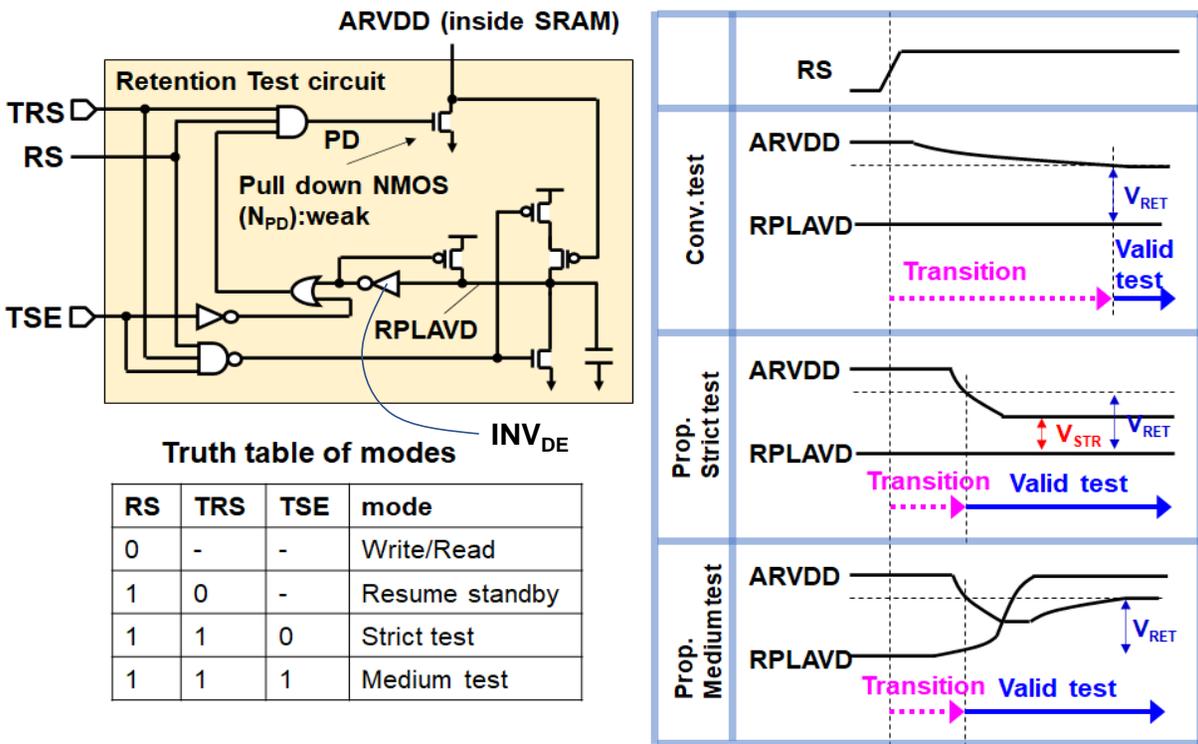
Fig. 43 Proposed retention test circuit, truth table and waveforms.
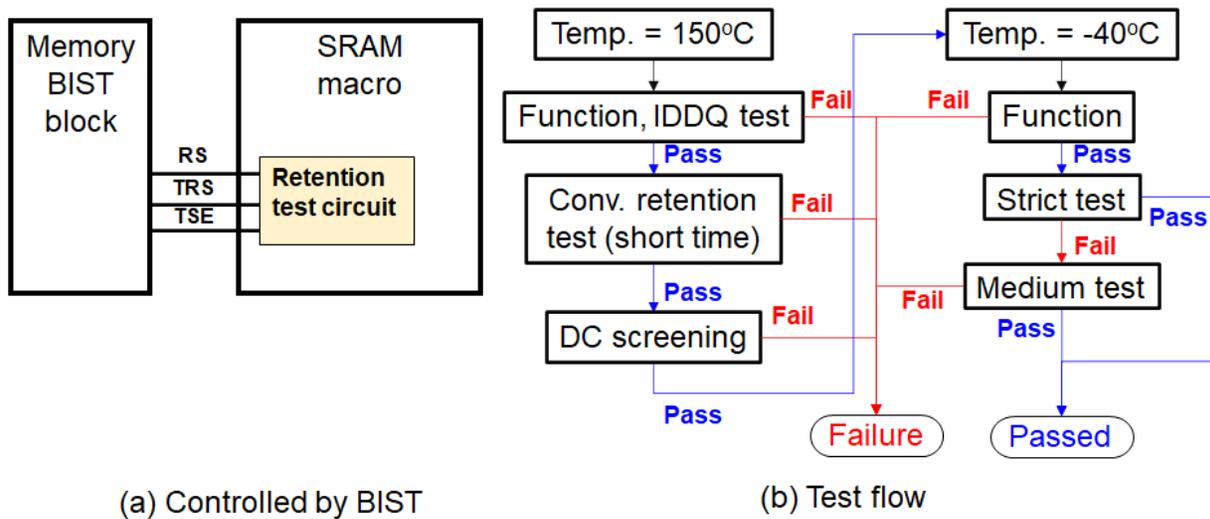
Fig. 44 Proposed 2-stage retention test flow with memory-BIST.
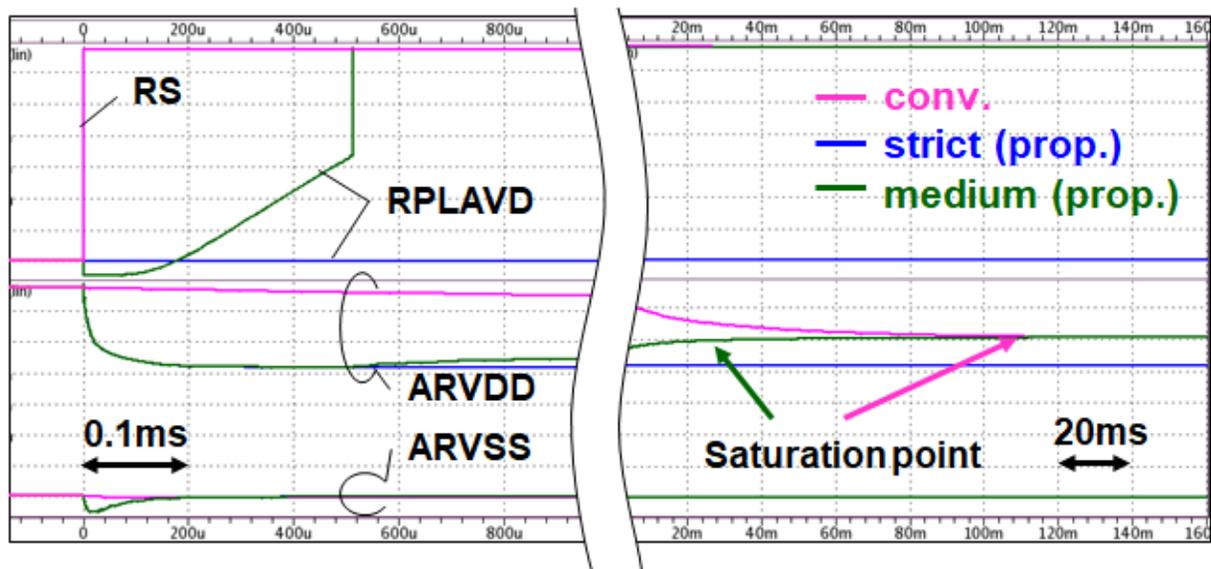
Fig. 45 SPICE simulation waveforms of retention test circuit of 4k-word x 40-bit SRAM at worst condition (1.35 V, process SS and -40°C)

3.2.4. DESIGN AND FABRICATION OF TEST CHIP

Fig. 46 portrays a die photograph of a test chip and the layout plot of a SRAM macro on the 110-nm platform. The 160-kbits SRAM macro with the proposed retention test circuit are implemented in the test chip with 16 instances. Table 2 summarizes the test chip features. Full read/write operation, and stable retention were observed at -40°C to 150°C. The dynamic powers in read (write) operation are 90 (105) μW/MHz at 1.5 V typical supply voltage and 25°C. Fig. 47 shows a typical Shmoo plot of the read access time vs V-min. The read access time is 4.74 ns at the typical condition of 1.5 V. Fig. 48 (a) shows cumulative distribution functions (CDFs) of V-min at -40°C and 150°C. The median values of V-min at process TT (NMOS: typical, PMOS: typical) and SF (NMOS: slow PMOS: fast) corners are 0.95 V and 1.14 V respectively, those values are enough margin for typical operating voltage of 1.5 V. Fig. 48 (b) shows CDFs of standby power without source bias and with source bias conditions at 25°C. The median values are 2.3 μW without source bias and 0.73 μW with source bias respectively. The standby power per bitcell is 0.28pW at 25°C, reduced by 70% by applying the proposed source bias technique.
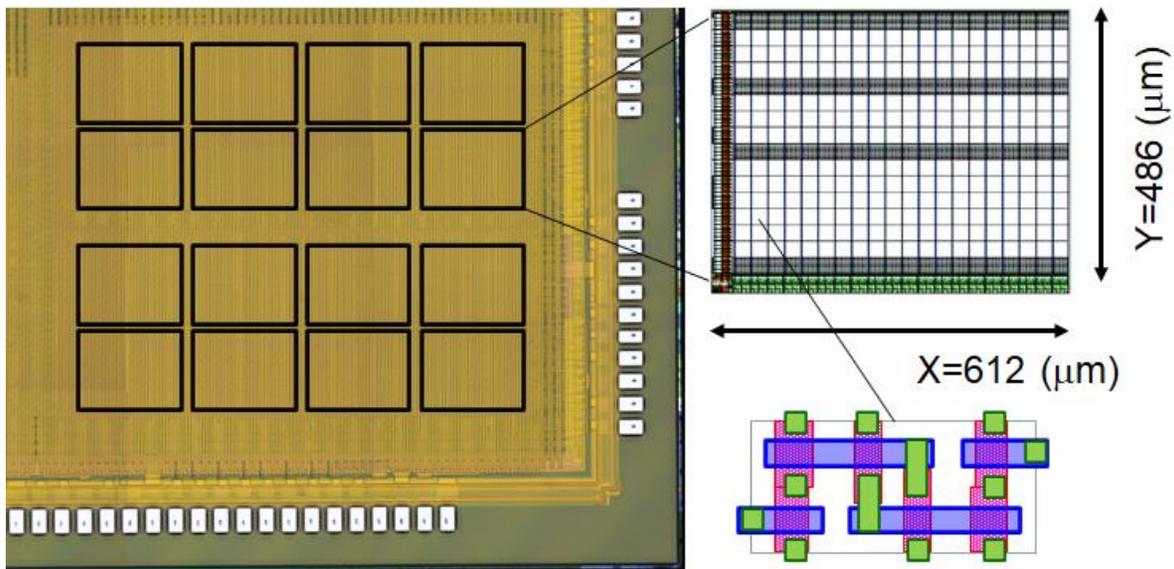
Fig. 46 Photograph of the test chip and layout plot of SRAM macro.

Table 2 Features of the test chip.

	Features
Technology	110-nm process technology
Macro configuration	2.5-Mbit (4096 word x 40 bit x 16)
Macro size	612 µm x 486 µm 0.302 mm^2@160-kbit
Bit density	0.518 Mbit/mm^2
Access time @typ	4.7 ns
Cycle time @typ	6.8 ns(147 MHz)
Dynamic power @typ	Read: 90 µW/MHz Write:105 µW/MHz
Standby power @2.5-Mbit	114.3 µW @150°C 0.73 µW @25°C
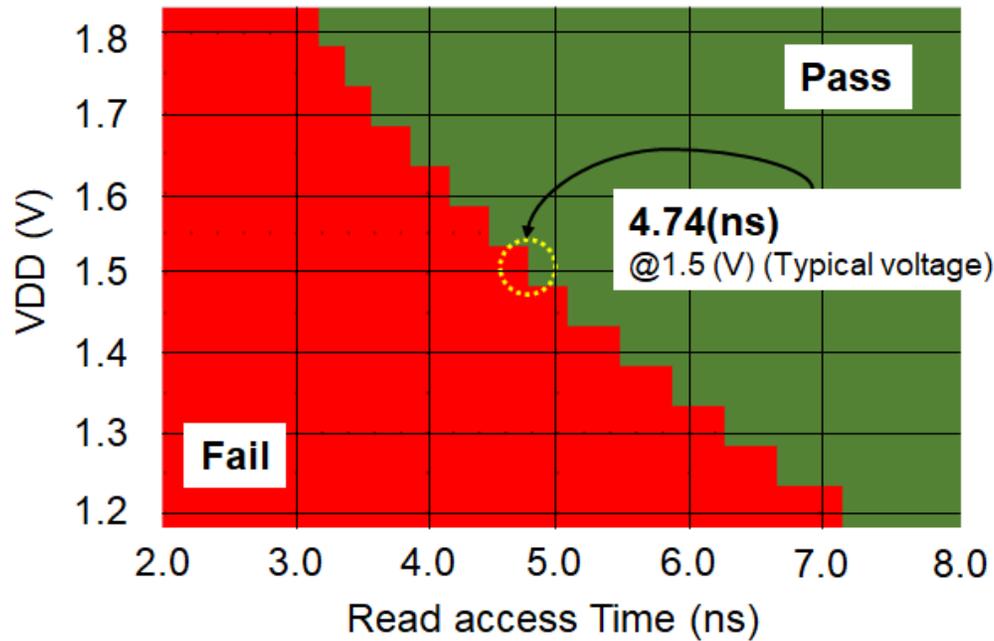
Fig. 47 Shmoo plot V-min vs. access time at 150°C

(a) Write/Read/Retention V-min

(b) Standby power @ 25°C
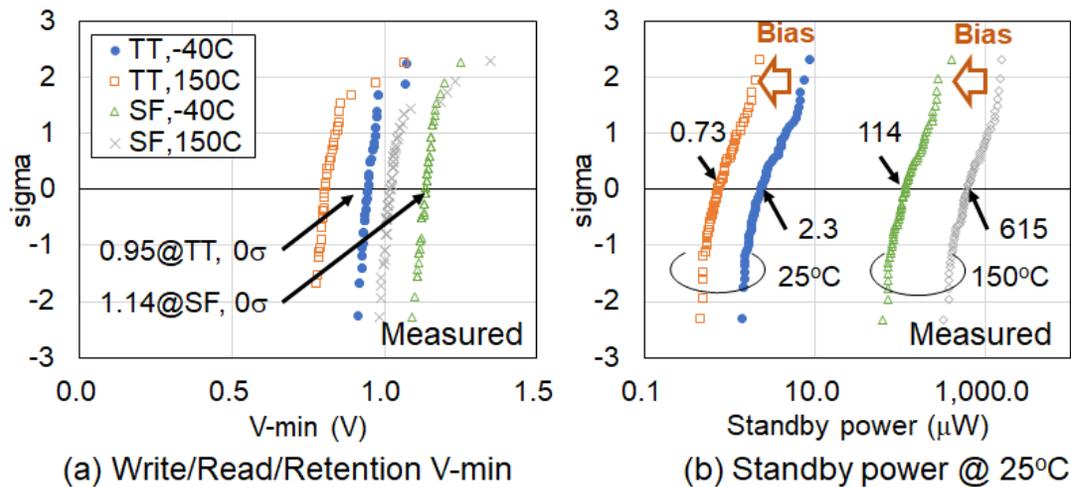
Fig. 48 Distribution of measured V-min at -40/150°C and leakage power at 25°C.

Fig. 49 (a) and (b) show CDFs of the retention V-min of 0.2 sec conventional test (conv. 0.2s), 10 sec conventional one (conv. 10s), the proposed strict test and the medium test at SS/-40°C and TT/-40°C respectively. Both the strict and medium tests are executed for 0.2 sec pause time. The conv. 10s can be assumed as the true V-min of each chip due to its enough long pause time. The

conv. 0.2s distributes smaller than the conv. 10s. The median values of conv. 0.2s is shifted by -330 mV with respect to conv. 10s at SS/-40°C. It goes out failure chips. The V-min of the strict test and medium test are distributed larger than the conv. 10s. Therefore, it can screen-out failure chips correctly. The strict test is distributed larger than the medium test, then the strict test might excessively screen out pass chips. Probabilistically, if there are 1/100 of over-screened chips at the 1st strict test and followed by the 2nd medium test, the 2-stage retention test time at LT is almost same as the 1st strict test time in average. As a result, the proposed test flow can reduce the data retention test time to 1/50 as shown in Fig. 50.
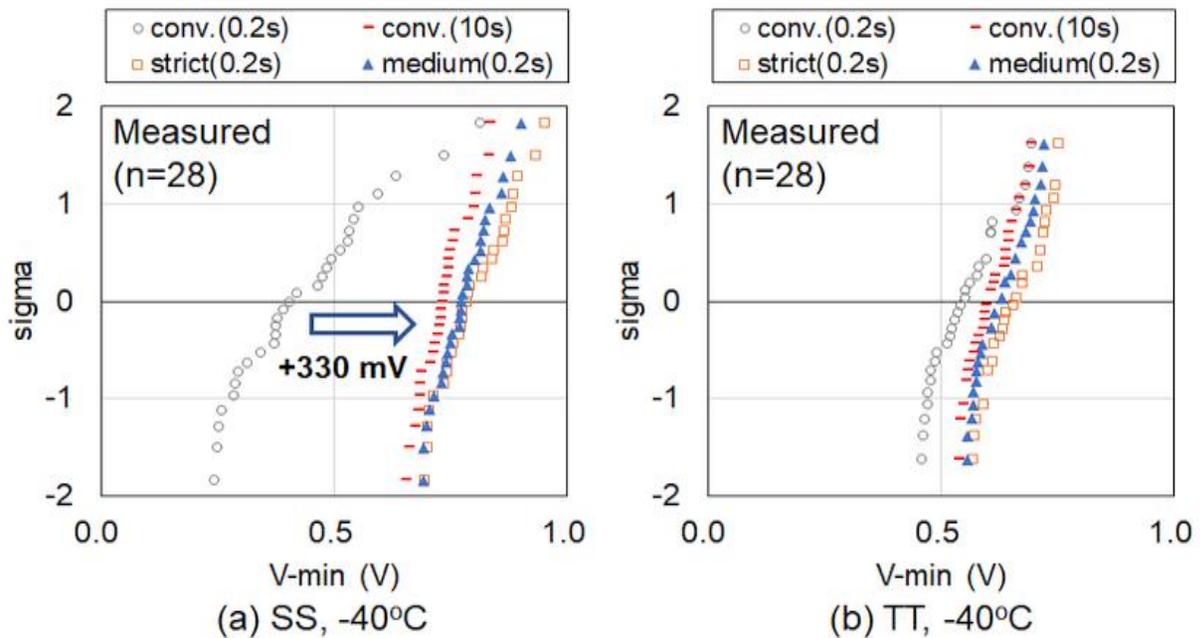
Fig. 49 Measured retention V-min at -40°C and effects of proposed retention test modes.
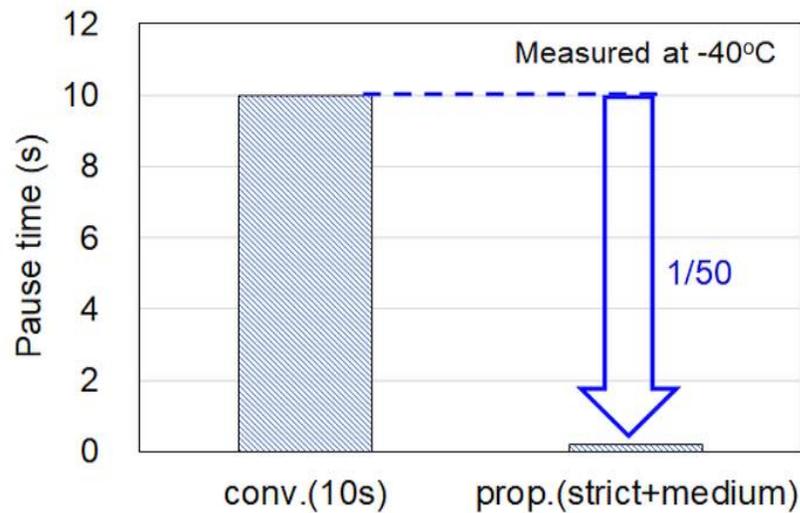
Fig. 50 Data retention time comparison with the conventional method.

3.2.5. Summary

An ultra-low standby power embedded SRAM macro with resume standby circuits were implemented on a 110-nm SoC/MCU. It can reduce standby power at HT and RT with enough cell bias. A 2-stage test screening method for resume standby mode was proposed. Test chips including 130kbit SRAMs on 110-nm were confirmed full reading and writing operations. Measured resume standby power at 1.5 V typical supply voltage and 25°C was reduced by 70% compared to conventional one, achieving 0.28 pW/cell. The test time was reduced to 1/50 with only 0.03% area overhead.

3.3. High Voltage MOS SRAM

3.3.1. Introduction

Recently, in the rapidly growing the IoT market, it is expected that all applications will be connected wirelessly. Thus, there are strong demands for extremely low-power operation for longer battery life. An edge equipment consists of some sensors and low-power micro-controller-unit (MCU) with embedded non-volatile-memories (NVM) and RF circuits. To reduce the total BOM cost, an edge computing system is required to remove the power module IC (PMIC), which is often

used in the mobile system for suppling multi-voltages to SoCs/MCUs. Fig. 51 shows a single power supply system without PMIC in an IoT application, whereas high performance mobile systems require the power management system with PMIC. To implement the single power supply system with a buttery, MCUs typically have voltage down converters (VDC) internally. On the intermitted operation in the edge equipment, the MCU is computing along with the obtained data from sensors in the active mode. Meanwhile almost systems are shut down in the standby/sleep mode. However, the latest data from sensors buffered in the embedded memory of MCU are needed to be retained before sleeping. The embedded NVM has a good data retention characteristic with less leakage power, however the endurance has limitations and writing/programing power and speed does not meet the target of such embedded buffer memories.

In the work, the buffer/backup (BB) SRAM using 3.3 V IO MOSs for PMIC less low-power MCU is proposed for IoT applications. Using volatile memories with low leakage current have been reported [6],[21],[31],[33][34], however, either an additional process steps or custom process to fabrication increasing the chip cost is required. The proposed BB SRAM macro can be implemented without any process customizations.
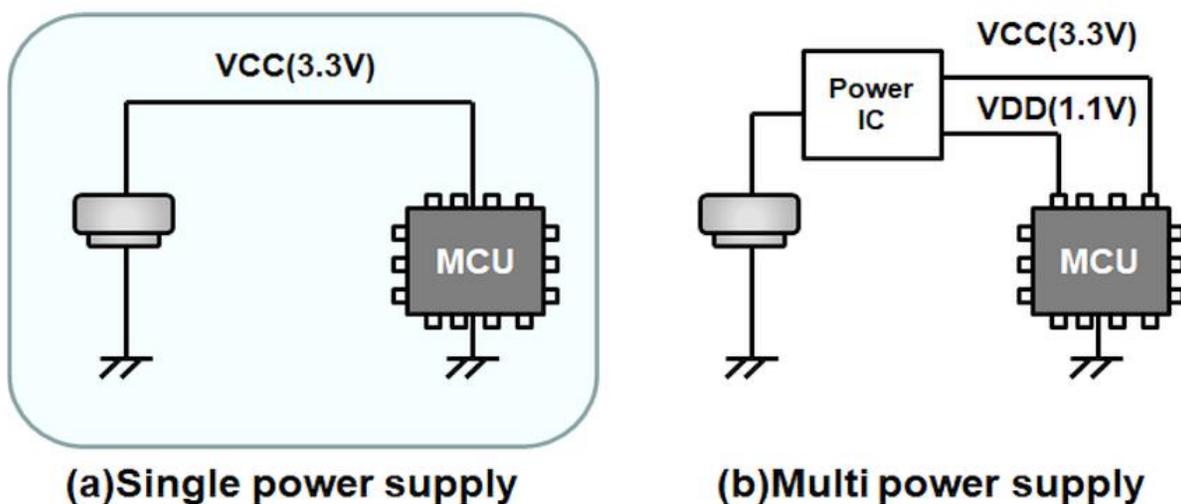
Fig. 51 Power-supply for MCU.

3.3.2. Buffer/Backup SRAM with 3.3 V Thick-Gate-Oxide IO MOS

Fig. 52 (a) and 2(b) show the conventional power management system and proposed one using on-die VDCs. In the conventional system, main blocks are shut-down by power gating and no standby power in the corresponding VDC, becoming almost zero standby power in these blocks.

However, in the other always power-on block, the VDC typically has several μA standby current [29] to generate the core voltage of 1.1 V in a 40-nm technology to retain the SRAM data and the wakeup/sleep system control logic. On the other hand, in the proposed power management system, always power-on blocks including the BB SRAM and logics are implemented in the 3.3 V IO MOS regions without any VDCs as shown in Fig. 52 (b). In this case, there is no leakage power of VDC, only consuming the leakage powers in the BB SRAM and the system logics.

   To reduce the total standby power of BB SRAM in the MCU, a 6T single-port SRAM bitcell using the 3.3 V thick-gate-oxide IO MOSs is proposed. Fig. 53 (a) illustrates the circuit of the 6T SRAM bitcell. It has separated sources which are shown in Fig. 53 (a) as VDD/ARVDD and VSS/ARVSS, respectively. The 6T SRAM bitcell is optimized by changing both gate length (L) and gate width (W) for each pull-up (PU) PMOS, pull-down (PD) NMOS, and pass-gate (PG) NMOS to reduce the leakage current keeping with acceptable area and access speed. Optimized L/W ratio of each PU, PD and PG is shown in Fig. 53 (b). Fig. 53 (c) illustrates the layout of the proposed 6T SRAM bitcell, which area is 2.888 μm^2. Source bias control techniques [13],[20],[28],[30] are effective for reducing standby leakage power in SRAM bitcells. In this work, the SRAM macro which is consist of 3.3 V supply voltage transistors including bitcell arrays is proposed. Then, both VDD source bias and VSS source bias techniques are introduced to effectively reduce the leakage power in the SRAM bitcell array. Fig. 54 shows the circuit diagram of the bias controller in the proposed BB SRAM macro. If input signal "RS" is set to "H", the SRAM macro transits to the resume standby mode that can hold the stored data with less leakage current than the normal standby mode. The signals "NMA[2:0]" are pins to adjust the bias levels of ARVDD and ARVSS in the resume standby mode.
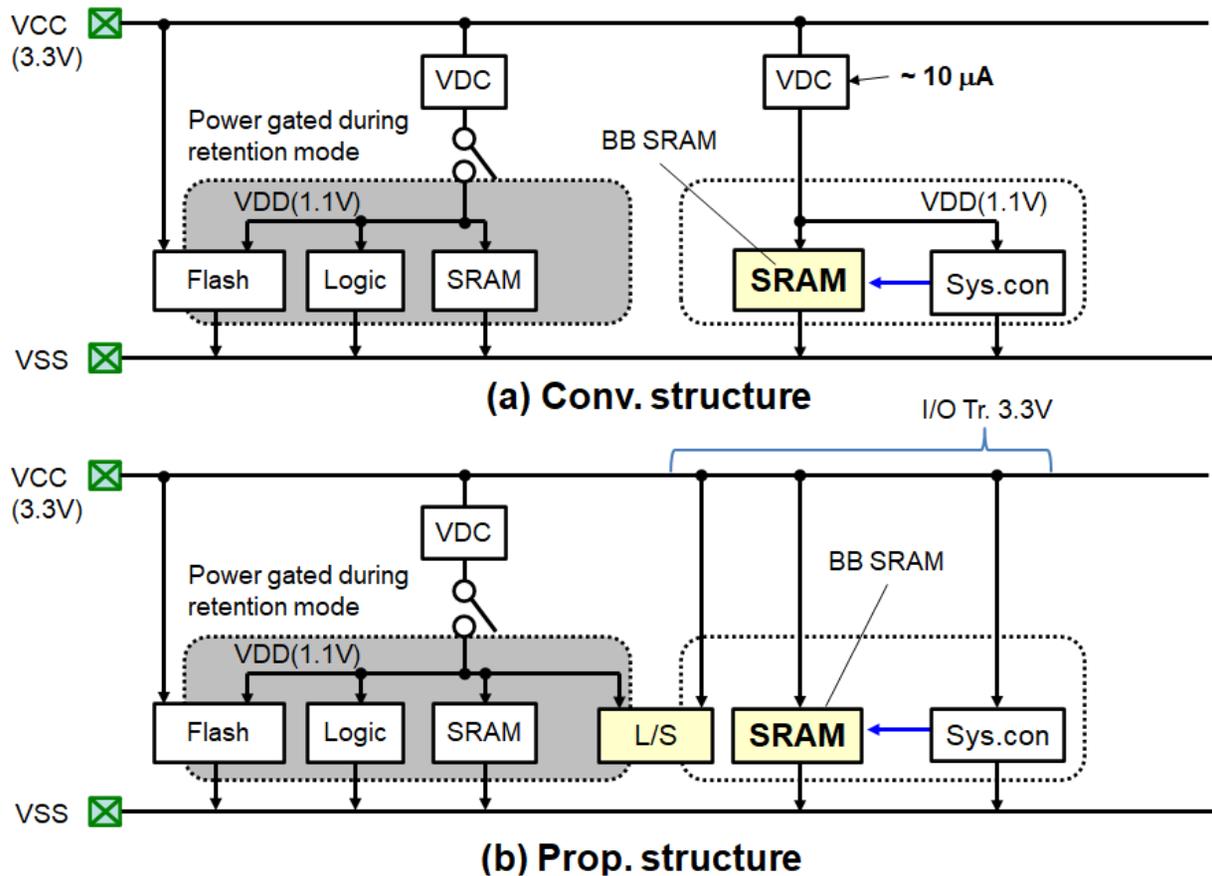
Fig. 52 Power management system using on-die voltage down converter (VDC).

Fig. 55 shows the simulation result of the leakage current reduction. The bias controller lowers the voltages of ARVDD and bitline (BL), and raises the voltage of ARVSS, and the leakage current is reduced by 98%. All simulations are executed at the worst condition of FF (NMOS:Fast and PMOS:Fast) process corner, 3.6 V supply voltage (+10%), and 65°C. Fig. 56 (a) shows the temperature dependencies of the read static noise margin (SNM) [26], write margin (WM) defined by write-trip-point [26] of proposed SRAM bitcell at -6.0 σ, 2.7 V and each worst condition. Fig. 56 (b) shows simulation results of the retention margin obtained by [26] in the both case of the normal standby (w/o source bias) and the resume standby (w/ source bias) conditions. The biased voltages generated by Fig. 54 are reproduced and reflected in the result of Fig. 56 (b). The simulation results show that the read SNM and WM have enough margin, and the retention margin becomes worst at the -40°C, but still has margin.

	Ratio
PU (W/L)	1/1.4
PD (W/L)	1/1.7
PG (W/L)	1/1.7

(a) Circuit

(b) W/L ratio

2.888 (um²/cell)

CONT
POLY
DIFF
M1

V1
M2
V2
M3

(c) Proposal layout
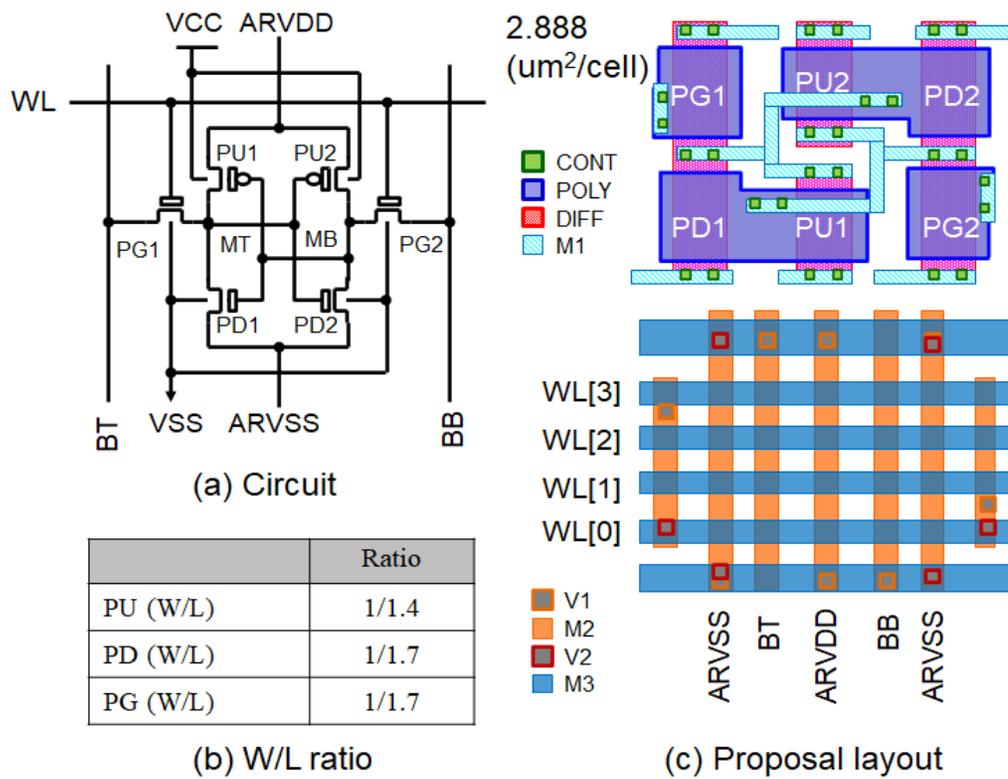
Fig. 53 6T SRAM bit cell using 3.3 V thick- gate-oxide IO MOSs
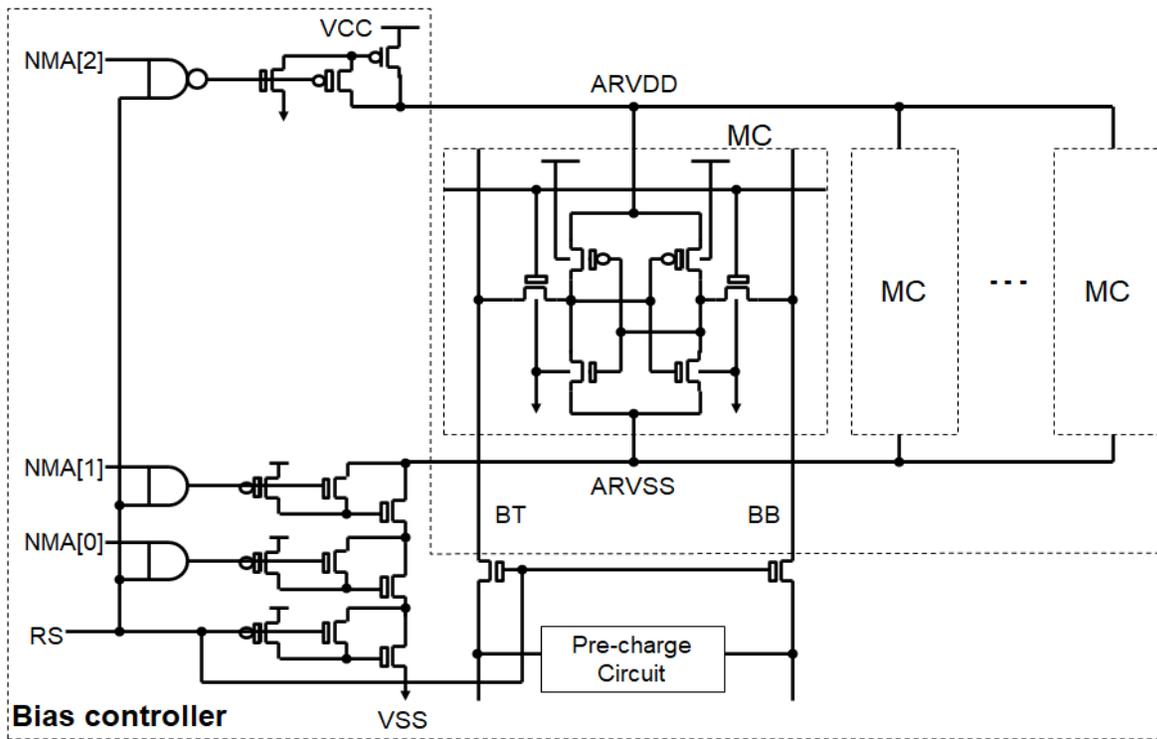
Fig. 54 Circuit diagram of bias controller in the proposed BB SRAM. macro.
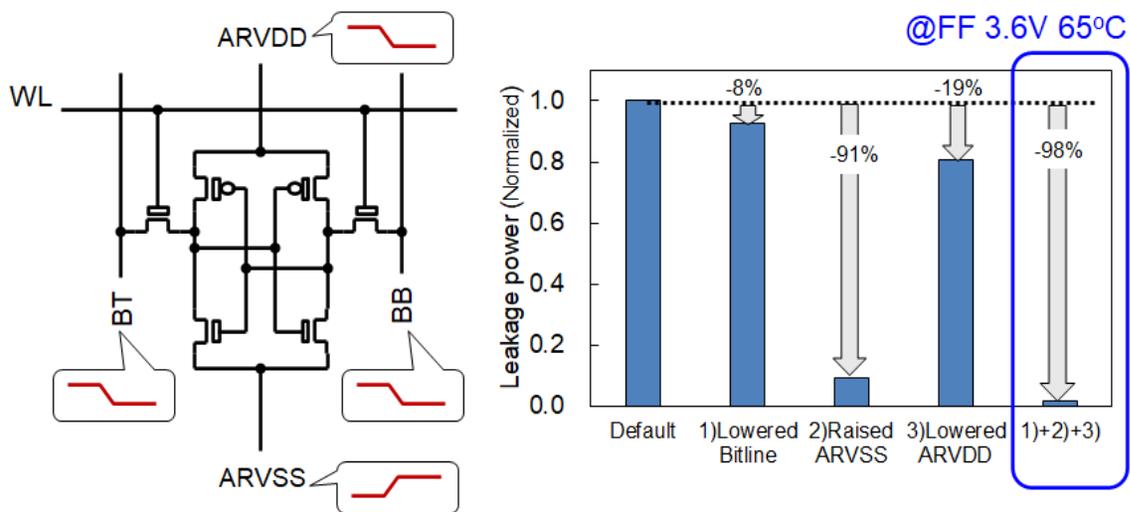
Fig. 55 Simulated leakage power reduction in the retention mode.
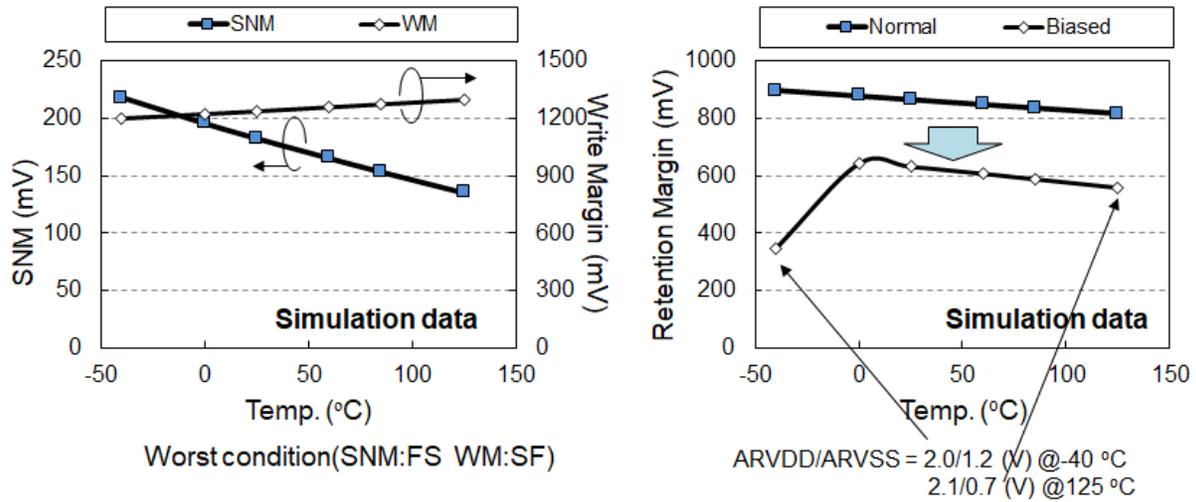
Fig. 56 Simulated temperature dependences of the static-noise margin (SNM), write margin (WM) and retention margin.

3.3.3. Dynamic Power Reduction

The proposed buffer/backup (BB) SRAM with 3.3 V IO MOSs can reduce the standby leakage as described before, however the dynamic power increases compared to typical SRAMs with 1.1 V core MOS. The techniques to reduce dynamic power are needed. The reported technique to select a column address with wordline (WL) using the advantage of wiring area is effective[19]. As shown in Fig. 53 (a), the proposed cell has four WLs in a cell, then not only a row address but also a column address can be selected by WLs. Fig. 57 indicates a simple block diagram of the designed SRAM macro. The signal of RA and CA in Fig. 57 are row addresses and column addresses respectively. The WL decoder can select a unique address specified by RA and CA. Furthermore, column MUXs in the I/O block can select a column address, as a result, it is possible to reduce the power consumption for unselected BLs. Fig. 58 (a) and Fig. 58 (b) show SPICE simulation waveforms at the worst process-voltage-temperature (PVT) conditions in the write and read operations, respectively. The signals of MT/MB in Fig. 58 (a) are internal complementary nodes in memory cell, and CTR/CBR in Fig. 58 (b) are differential inputs for a sense amplifier as shown in Fig. 57. The implemented 64-kbit BB SRAM macro has enough write/read operation margin considering with 6-sigma local variations. Fig. 58 (c) shows the SPICE simulation waveforms during transmission from the normal standby to the resume standby mode and from the resume standby mode to the normal standby mode. The waveforms with different colors indicate variations

of NMA[2:0]=000~111. The voltages of ARVDD and ARVSS are adjusted using NMA[2:0]. Then, power consumption to transit from the resume standby mode to normal mode is reduced by adjusting NMA[2:0]. That power consumption is 1083 µW/MHz if NMA[2:0]=111, and it becomes to 521 µW/MHz when NMA[2:0]=000. The setting of NMA[2:0]=000 is effective when switching the normal standby and resume standby modes frequently.
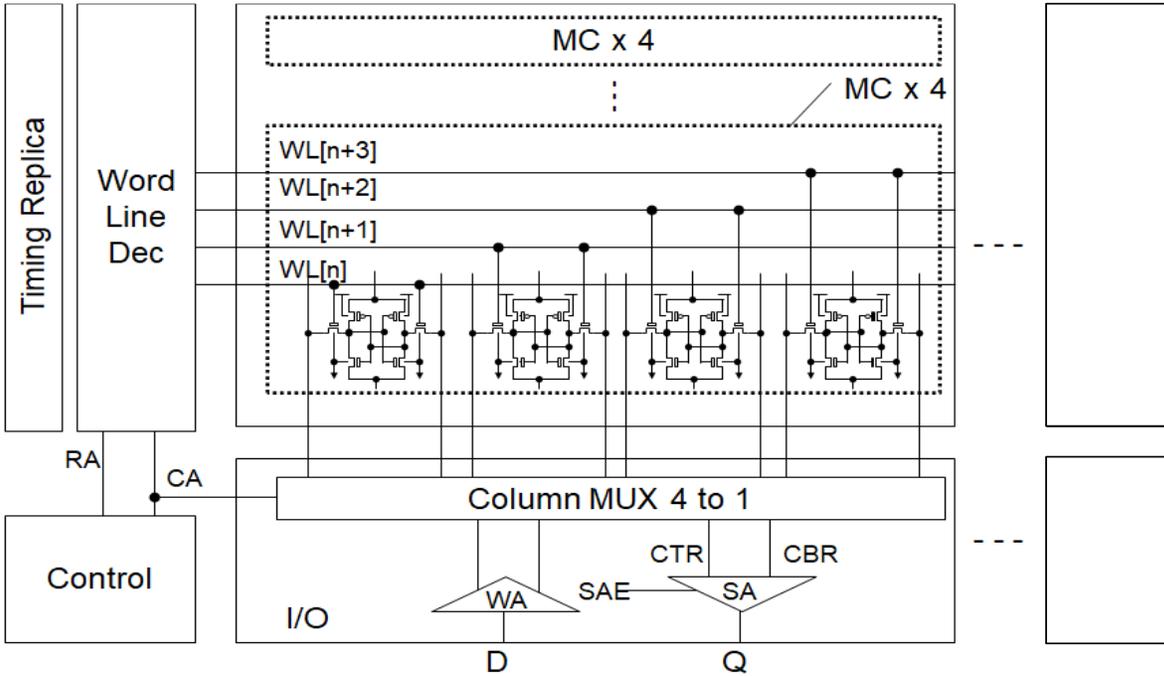
Fig. 57 Block diagram of proposed Retention SRAM.
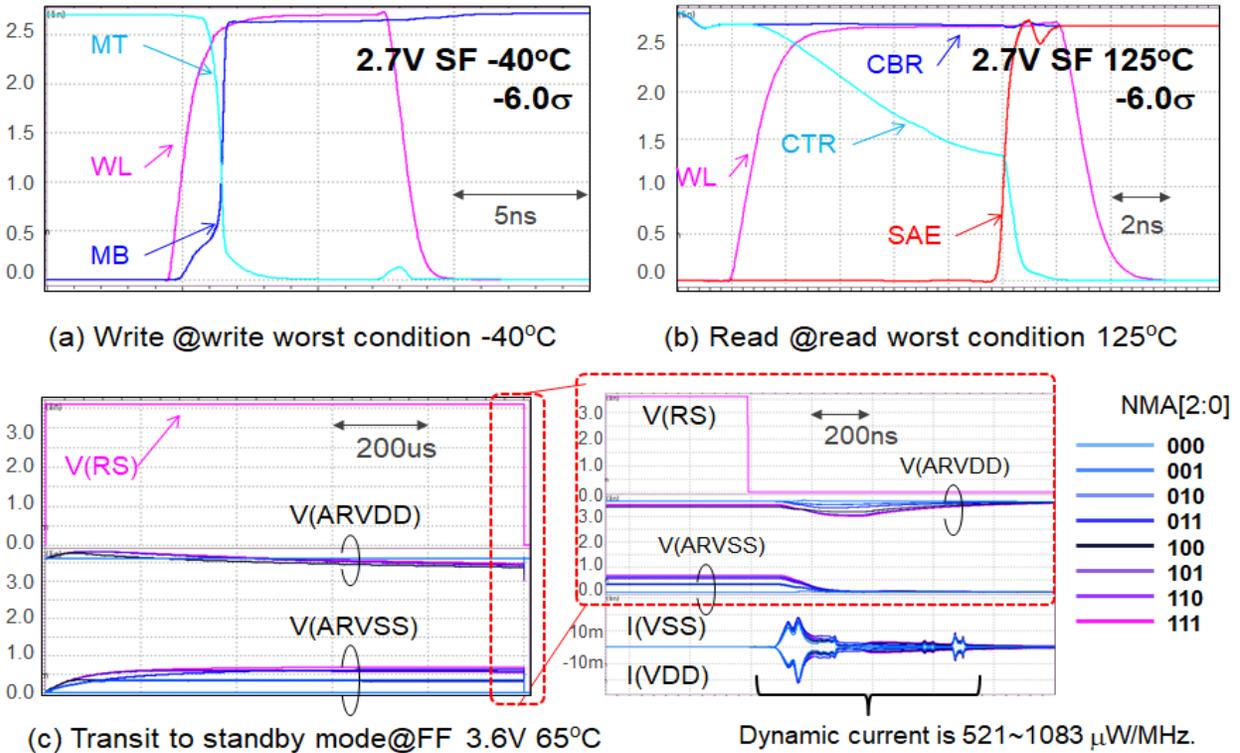
Fig. 58 Simulation waveforms of 64-kbit BB SRAM (2048-word x 32-bit).

3.3.4. Design and Evaluation of Test Chip

Fig. 59 portrays a die photograph of a test chip using the 40 nm CMOS technology [32]. The 64-kbit BB SRAM macro with the proposed circuit is implemented in the test chip. Table 3 presents a summary of the test chip features. Full read/write functions at temperatures of -40°C to 125°C were observed. Fig. 60 presents a typical Shmoo plot at 125°C, showing the minimum operating voltage (V_{min}) vs. the access time. The access time of the measured SRAM macro is 16.7 ns at the typical supply voltage of 3.3 V. It includes the delay of a level shifter. Fig. 61 (a) and Fig. 61 (b) show cumulative distribution functions (CDFs) of write/read V_{min} and retention V_{min} at the RS mode, respectively at temperatures of -40°C to 125°C. The total number of measured dies is 25. We can see that the median of write/read V_{min} for the typical process is 1.28 V, the median of retention V_{min} is 1.19 V and enough operation margin for typical supply voltage of 3.3 V. Fig. 62 (a) shows the CDF of standby leakage powers in the RS modes at 65°C and 125°C. The median of leakage power of 64-kbit SRAM is 6.0 μW at 125°C and 0.33 μW at 65°C, respectively. The leakage power at 25°C is lower than measurement accuracy of the tester. Therefore, it is extrapolated as 47 nW from

that at 65°C and simulated dependency of leakage current on temperature by SPICE simulation.
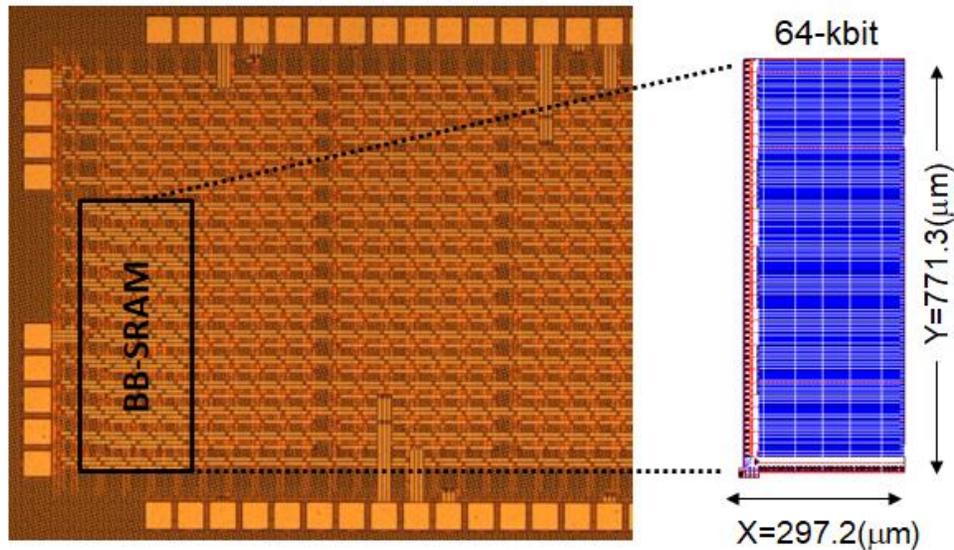
Fig. 59 Microphotograph of the Test chip and layout plots of proposed retention SRAM macro using 40-nm technology.

Table 3 Features of the test chip.

	Features
Technology	40-nm embedded flash process
Macro configuration	64-kbit (2048 word x 32 bit)
Macro size (included level shifter)	297.2 µm x 771.3 µm 225655 µm^2@64-kbit (Non-rectangular)
Bit density	0.277 Mbit/mm^2
Access time @typ	16.7 ns (included level shifter)
Cycle Time @typ	23.8 ns(42 MHz)
Dynamic power @typ	Read:174 µW/MHz Write:180 µW/MHz
Standby power @64-kbit	6.0 µW @typ 125°C 0.33 µW @typ 65°C

Fig. 62 (b) shows the measured dynamic power consumptions of write and read operations *vs* supply voltage. The measured dynamic power consumptions with the proposed circuitry are 174 µW/MHz for the read operation and 180 µW/MHz for the write operation at the typical process, 3.3 V and 125°C, respectively, obtained good correlations with simulation results. The dynamic

power with the multi-interleave WL scheme can reduce to 40% power consumption compared to without scheme, that read power is 426 µW/MHz. Table 4 shows a comparison with previous reports[19][28]. If the VDC has 2 µA of standby current at least, which is expected 1/3.5 of latest report by [29]. Proposed BB SRAM using 3.3 V IO MOS does not needed such VDC leakage power, having advantage of 1/140 total standby leakage power reduction, as shown in Table 4.
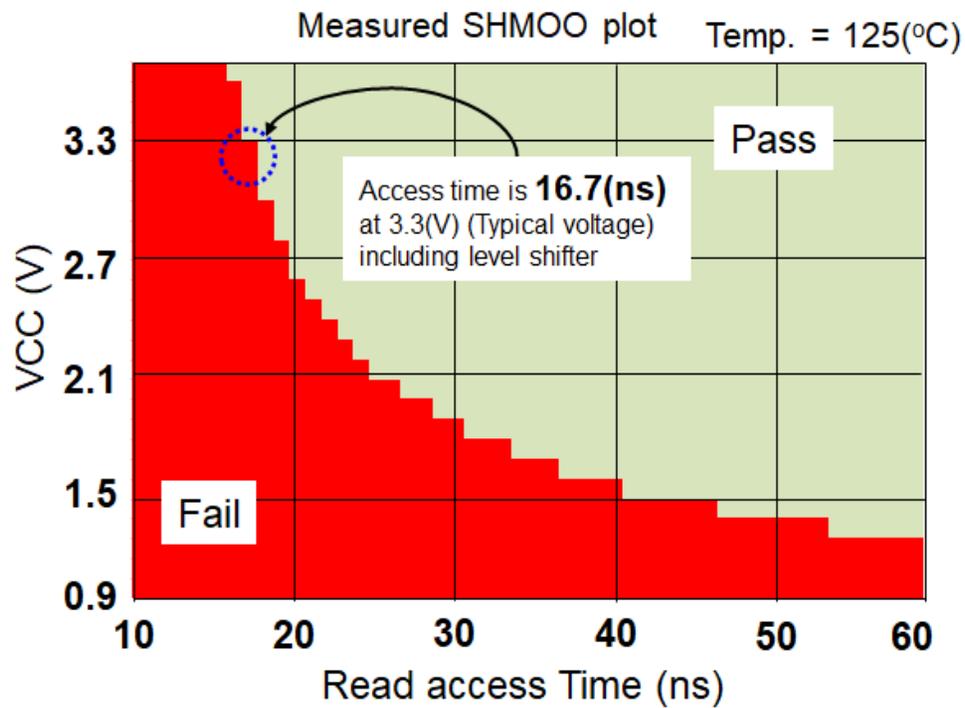
Fig. 60 SHMOO plot V_{min} vs. the access time at 125°C.
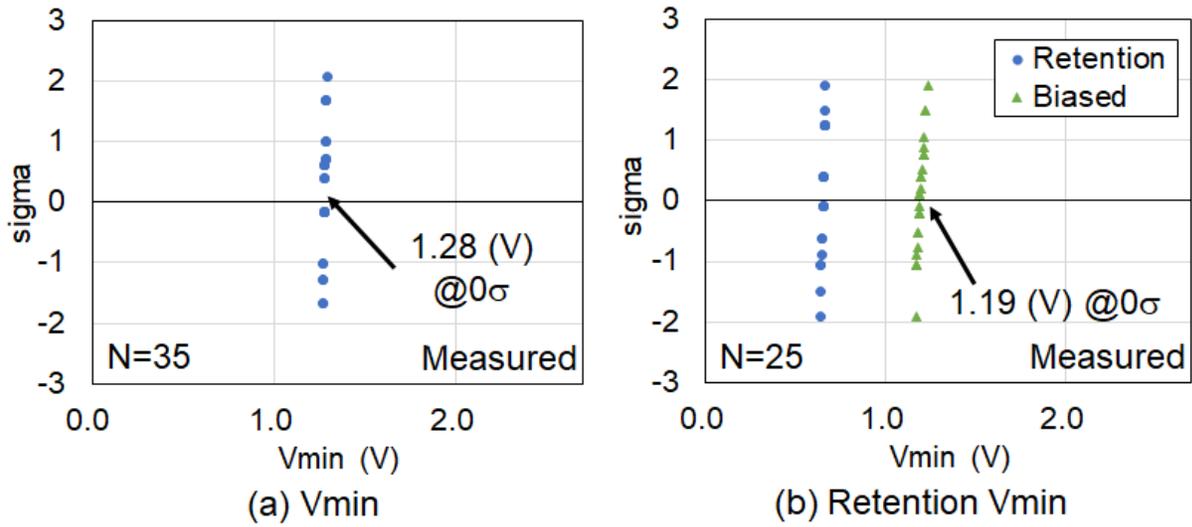
Fig. 61 Distribution of measured V_{min} at temperatures of -40°C to 125°C.
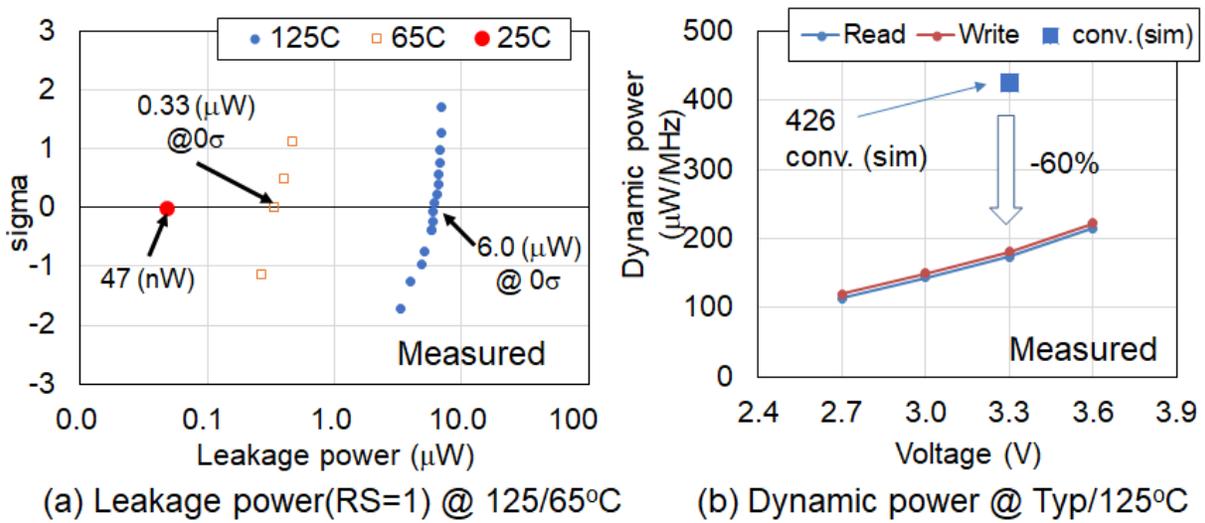
Fig. 62 Distribution of measured leakage power and voltage dependency of dynamic power.

Table 4 Comparison with previous reports.

		ISSCC'14[19]	VLC'17[28]	This work
Standby power	SRAM(*1)	2.6 nW @25°C	0.86 nW @25°C	330 nW@65°C 47 nW@25°C
	SRAM+VDC(*2)	6.60 μW @25°C	6.60 μW @25°C	
Techology node		65nm	65nm	40nm
Voltage		1.2V	0.75V	3.3V
Advanced or additional Process		Yes	Yes	No
Transistor		thick-gate	thin-gate	thick-gate
Cell Area		2.159 μm²	0.5408 μm²	2.888 μm²
Capacity		128 kbit	128 kbit	64 kbit
Dynamic power		25 μW/MHz	6.3 μW/MHz	174 μW/MHz

(*1):Normalized to 64-kbits.
(*2):Assumed VDC has 2 μA standby current at 3.3 V.

3.3.5. Conclusion

An effective standby power reduction of a buffer/backup SRAM in an MCU for a PMIC less edge system in IoT applications was proposed. It is implemented using 3.3 V thick-gate-oxide IO MOSs without any on-die VDC for effectively reducing the leakage power. Four multiples interleave WL circuitry was also introduced to reduce the dynamic power. A test chip with a 64-kbit SRAM macro was designed and fabricated using a 40-nm technology. From the measured data, it was obtained that the leakage power was 330 nW at 65°C (47 nW at 25°C), which is smaller 1/140 than other works. The read/write dynamic power is reduced by 60% by ther interleave WL circuitry.

Chapter 4   Multi-port SRAM

First, multi accessing of DP-SRAM/2P SRAM is explained.

Fig. 63 portrays simple block diagrams for a 1R1W 2P SRAM macro and a 2RW DP SRAM macro. A high-density 8T-bitcell with decoupled read bitlines for 1R1W 2P SRAM has been reported [38]. However, this type of bitcells must be designed carefully with a single-ended read-sense-amplifier to ensure a sufficient read margin against cell leakage current. The 8T bitcell with differential bitline pairs in Fig. 63 can be used for both DP and 2P SRAMs to maintain a sufficient operating margin with high-speed access time by a differential read sensing scheme. It has also been beneficial to reduce the bitcell qualification cost and turn-around-time by particularly addressing only one 8T bitcell layout type for both DP and 2P SRAMs.

Fig. 64 shows access modes of the 2P SRAM and DP SRAM. AWL and BWL are the wordline for A-port and the wordline for B-port, respectively. The 2P SRAM and DP SRAM have two access modes. One is "different-row access mode" when two ports access different rows each other (Fig. 64 (a)). Another is "same-row access mode" when two ports access the same row at the same time (Fig. 64 (b)). The 8T bitcell has a specific issue called as "Disturbance issue", which occurs only during the same-row access mode[69].
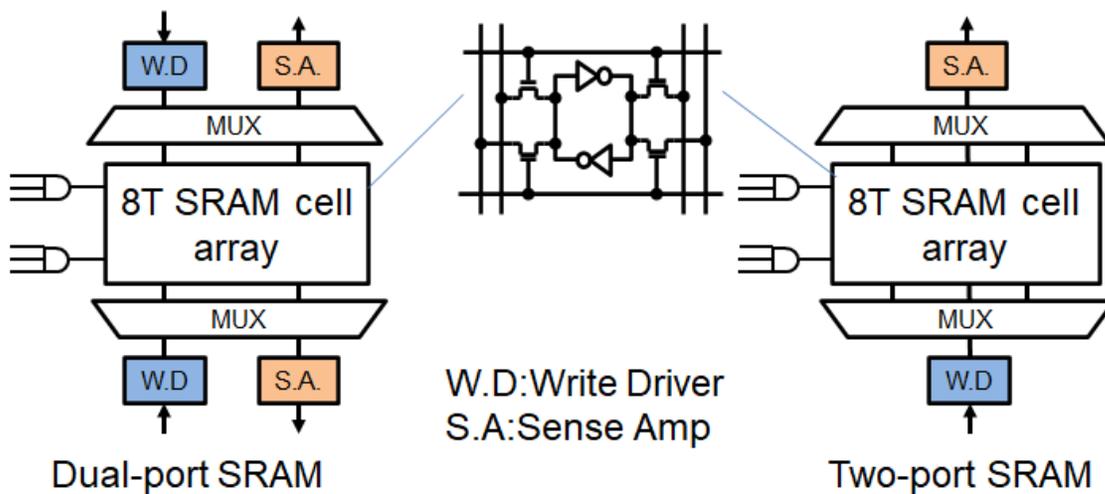
Fig. 63 Block diagrams for DP SRAM and 2P SRAM.
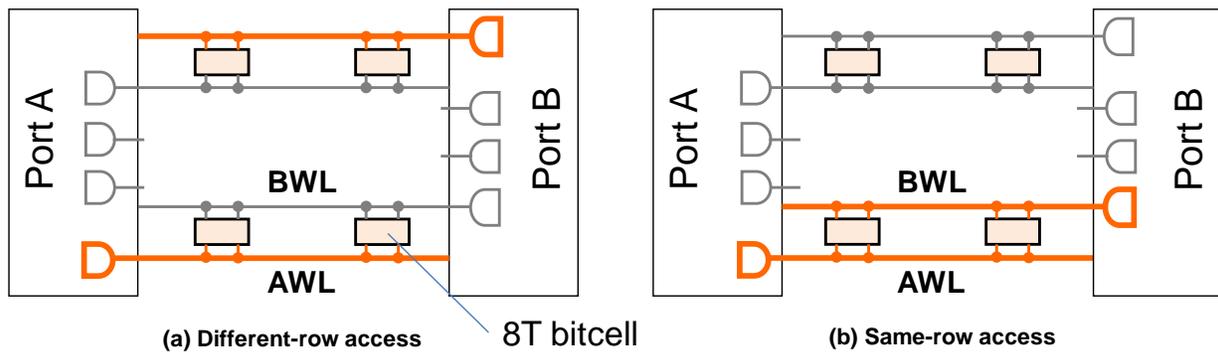
Fig. 64 Different-row access and same-row access.

The 8T SRAM has two disturbance modes: read disturbance and write disturbance. Fig. 65 presents an illustration of the read disturbance issue by activating the same-row address. The different row access mode works as a single port (SP) 6T SRAM: disturbance does not occur. Meanwhile undesirable cell current flows through the other bitline into the PD NMOS in the accessed bitcell in the same-row access mode. In that case, I_{read} of target the bitline through the PG and PD NMOSs decreases by additional cell current from other ports' bitline.
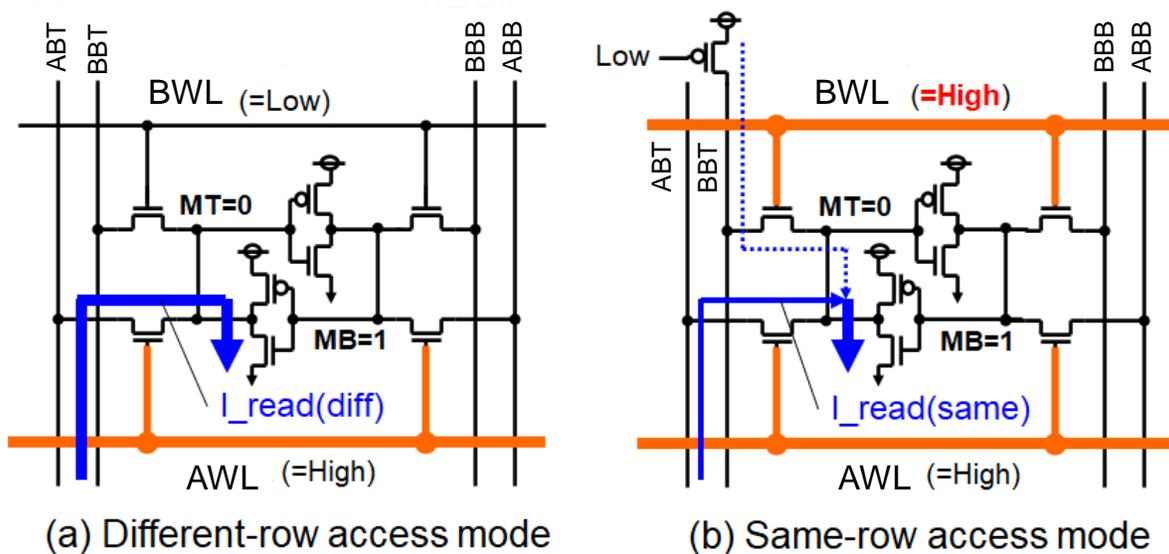
Fig. 65 Read disturbance issue.

Fig. 66 presents the measured distributions of I_{read} (n=20320) of the different row access mode and the same row access mode. Both distributions in the same and different row access modes are

observed on the normal distributions. The delta of I_{read} depends on the sigma value: it is 18% at 0 sigma (mean value) and 30% at -5 sigma as presented in Fig. 66 (b). Fig. 67 shows the other disturbance issue in the write operation. It is also disturbed by another BL current as well as a read-disturbance issue[40]. In the different row access mode, the write operation will be performed if the pulling down current (IPD) is larger than pulling up current (IPU) (Fig. 67 (a)). In the same row access mode, IPU is increased by the pre-charge circuit of BBT, disturbing the write ability.

(a) I_{read} distribution (same vs diff)

(b) Decrease of I_{read} (same vs diff)
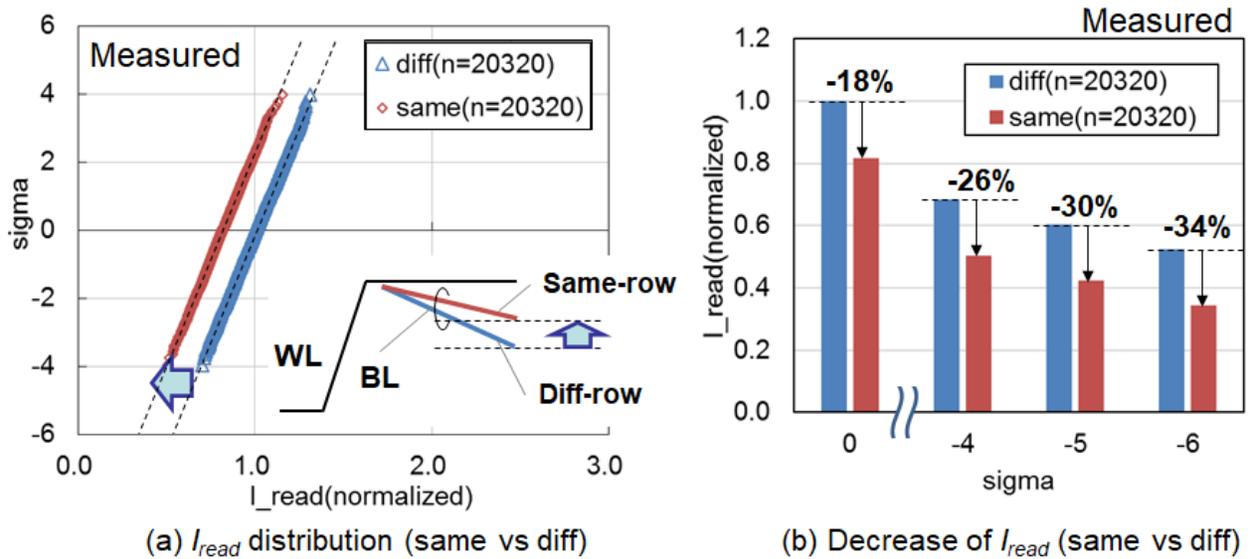
Fig. 66 Measured I_{read} distribution (different-row *vs* same-row access) of 40-nm CMOS process.
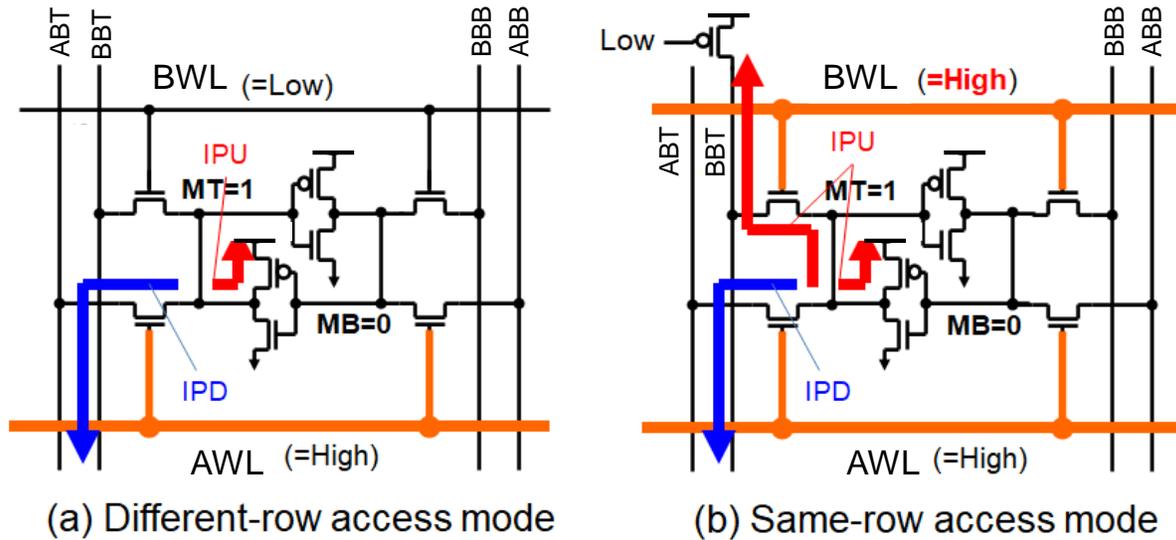
Fig. 67 Write disturbance issue.

Disturbance issue reduces read/write margin of 8T bitcell. The asymmetric Bitcell layout causes a certain unwanted mismatch offset in the capabilities of each A/B-port and true/bar. Some layout types of 8T bitcell have been reported [47]-[49],[51]. In mature processes technology, the 8T bitcell was a tendency to be designed for small area. In under 90-nm process technology, it tends to emphasize symmetry. Three types of 8T bitcell layout (a), (b) and (c) are presented in Fig. 68. The pair of pass-gate (APGT/APGB, BPGT/BPGB in Fig. 15) of (a) and (b) have some asymmetry. They might induce a mismatch (offset) of the read cell current (I_{read}) between true/bar or A-port/B-port bitlines. Fig. 68 (c) has little asymmetry in each pass-gates. The mismatch (offset) between APGT and APGB for port-A, BPGT and BPGB for port-B of (c) is expected to be smaller than (a) and (b) by the symmetric layout.
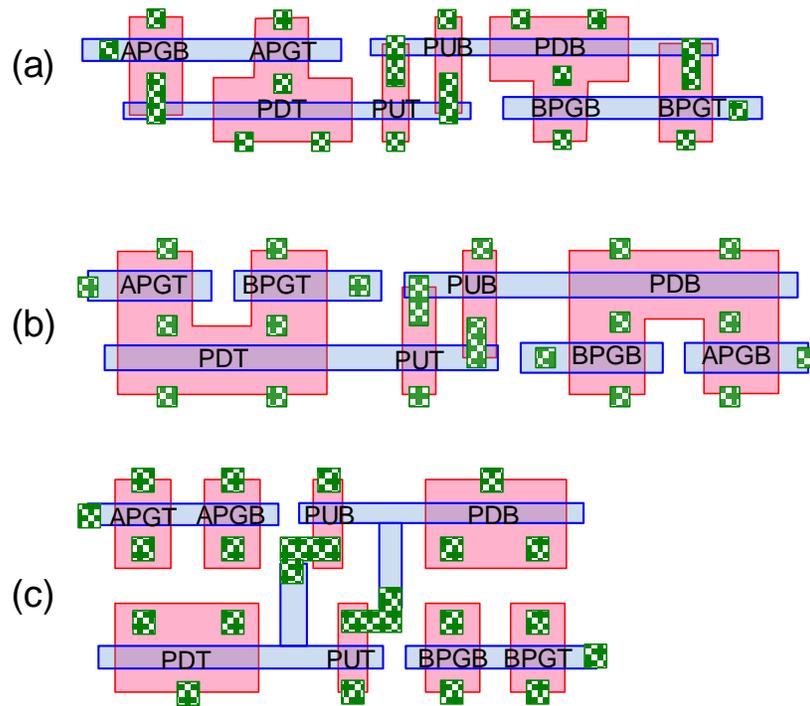
Fig. 68 8T SRAM bitcell layouts.

It is possible to confirm the symmetricity of each 8T bitcell layout by comparing measured cell current (I_{read}) distributions. Fig. 69 presents the measured distributions of I_{read} for A-true, A-bar, B-true and B-bar BLs implemented using a 40-nm technology. The (a)-(c) in Fig. 69 correspond to the (a)-(c) in Fig. 68, respectively. In the graph, all values of I_{read} are normalized by median and sigma. The (c) layout was observed to have the smallest differences among all I_{read}, whereas the layouts of (a) and (b) have offsets between true/bar or port-A/-B bitlines. From the measured I_{read} distributions, the write-margin (WM), read-margin (RM), and static-noise margin (SNM) are expected to be improved by virtue of the small offsets, resulting in a good minimum operating voltage (V_{min}). However, if the layout with some offsets should cause extra design margins, inducing the overhead of area, power and timing, and deterioration of V_{min}.
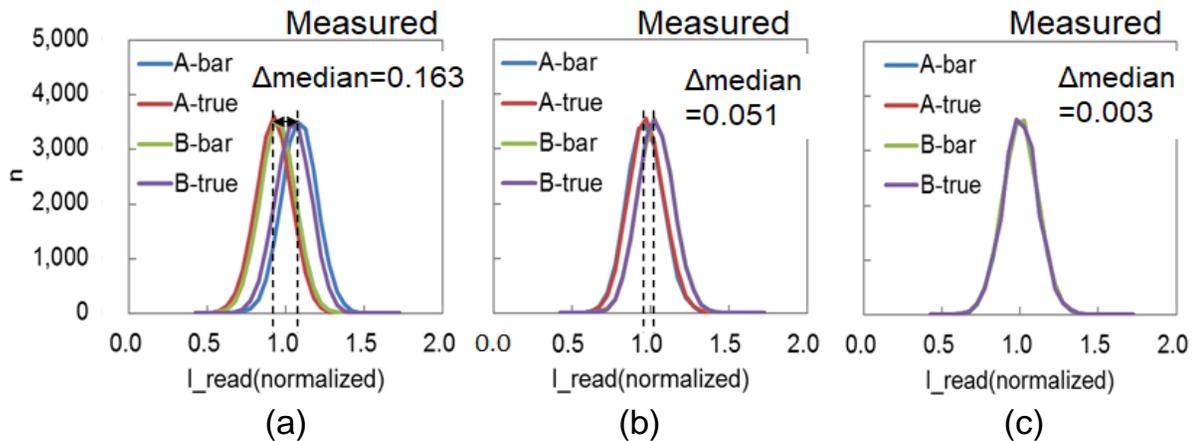
Fig. 69 Measured I_{read} distributions of A/B-true/bar BL (n=20320).

In this chapter, two techniques are introduced to prevent specs from getting worse by disturbance issue.

4.1. Two-port SRAM Against Write Disturbance Issue

4.1.1. Introduction

MCUs for the automotive device are combined with others using in-vehicle local area networks (LANs) such as the controller area networks (CANs) [36] and FlexRay [37]. For buffer memory of the in-vehicle LAN, 2-read/write (2RW) dual-port (DP) SRAMs and 1-read/1-write (1R1W) two-port (2P) SRAMs are used because of their high bandwidth and applicability to a different clock phase in each port, which facilitates high-speed communication processing. These features will become more important for future higher-speed in-vehicle LANs. Along with the SP SRAM, the DP and 2P SRAMs are also mandatory for automotive applications.

Previously reported studies have addressed only 6T single-port (SP) SRAM. This report describes the design and process optimizations of 8T bitcell for DP and 2P SRAMs in MCUs with a 40 nm embedded flash CMOS technology. A test screening circuit is also proposed to achieve sub-parts per million failures with more than 15-year reliability.

4.1.2. Write and Read Disturbance Issues

In this paragraph, the disturbance issues of the 8T bitcell is classified as DP and 2P SRAMs,

and screening test circuit is introduced for DP and 2P SRAMs.

Fig. 70 presents the disturbance mode of the 8T SRAM bitcell. The disturbance mode is categorized in four modes by the write/read mode of the victim wordline (WL) and aggressive WL. It is defined that the read-disturbance mode as read/read (victim/aggressive) or read/write access conditions. In the case of the write/write access condition, the mode is write-disturbance [18]. The write/read case is defined as half write-disturbance. Often, the pulse width of write WL becomes wider than the pulse of aggressive WL. In write-disturbance, the victim is disturbed during a long period. In the half write-disturbance, the aggressor port is read operation. Therefore, the disturbance period is shorter than write-disturbance, as shown in Fig. 70. There are only the read disturbance mode and half write-disturbance mode for the 2P SRAM. The write-disturbance mode never occurs in the 2P SRAM macro because there is only one write-port. Fig. 71 shows the relation between WL width and the worst timing for half the write-disturbance. The horizontal axis shows the difference between the pulse width of the victim write-WL and the aggressive read-WL. The vertical axis is the time from the falling edge of the victim WL to the falling edge of the aggressive WL. It negates after the victim WL in the positive skew condition, while it negates before the victim WL in the negative skew condition. In the zero skew condition, the victim and aggressive WLs negate simultaneously. The write margin becomes the minimum in the positive skew region if the pulse width of the aggressive WL is equal to the victim WL. The worst skew point for the write margin decreases in accordance with the decreasing pulse width of the aggressive WL.

Fig. 70 Read/Write disturbance modes of 8T SRAM bitcell.
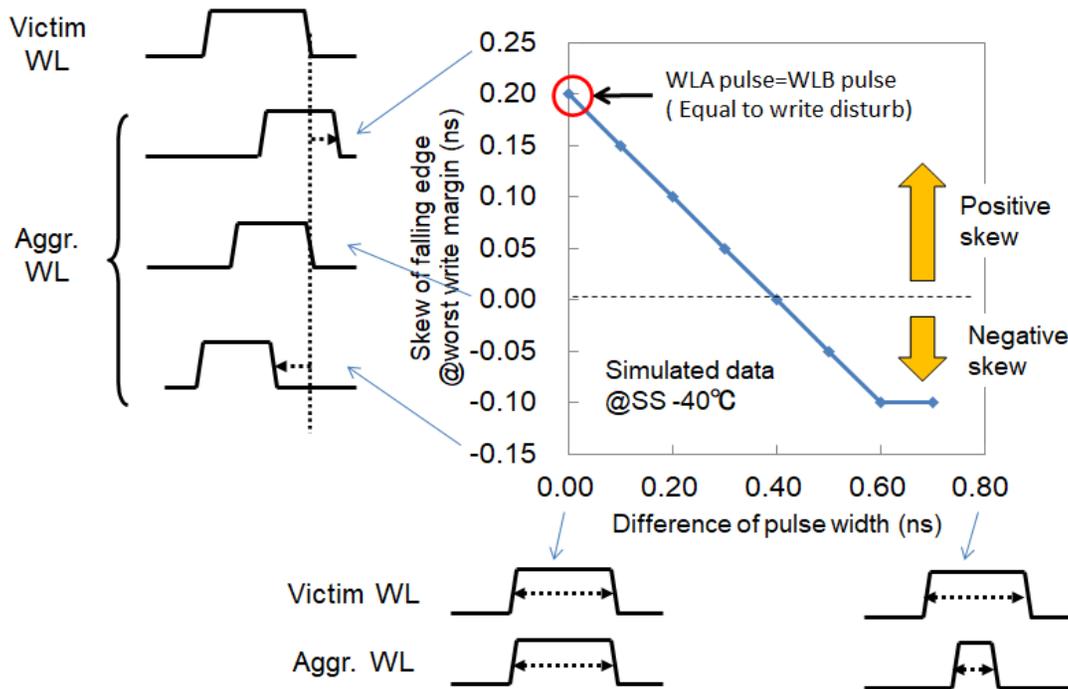
Fig. 71 The worst skew in the half write-disturbance.

Test circuits for screening disturbance failures are necessary for both DP and 2P SRAMs for screening the write-disturbance failure. The test circuit, which generates an aggressive WL pulse with fully covered the victim WL in the test mode, has been proposed for the DP SRAM [40]. The same test circuit with tuned delay timing is proposed for screening the disturbance failures of the 2P SRAM appropriately. Fig. 72 shows the proposed test circuit of the 2P SRAM. This circuitry delays the rising edge or falling edge of the WL pulse depending on the test mode. TME is a test signal to switch the operation into the test mode. When TAE is low, AWL is the aggressor and covers BWL. If TAE is high, then BWL becomes the aggressor and covers BWL [40]. It is possible to screen out disturbance failures of the 2P SRAM by tuning delay-1 and delay-2 to match the worst skew. This circuitry, which consists of complex gates and two delay elements, delays the rising edge or falling edge of the WL pulse depending on the test mode. Actually, TME is a test signal to switch the operation into the test mode, whereas TAE identifies which port produces the disturbing WL pulse with negative/positive clock skews [40].
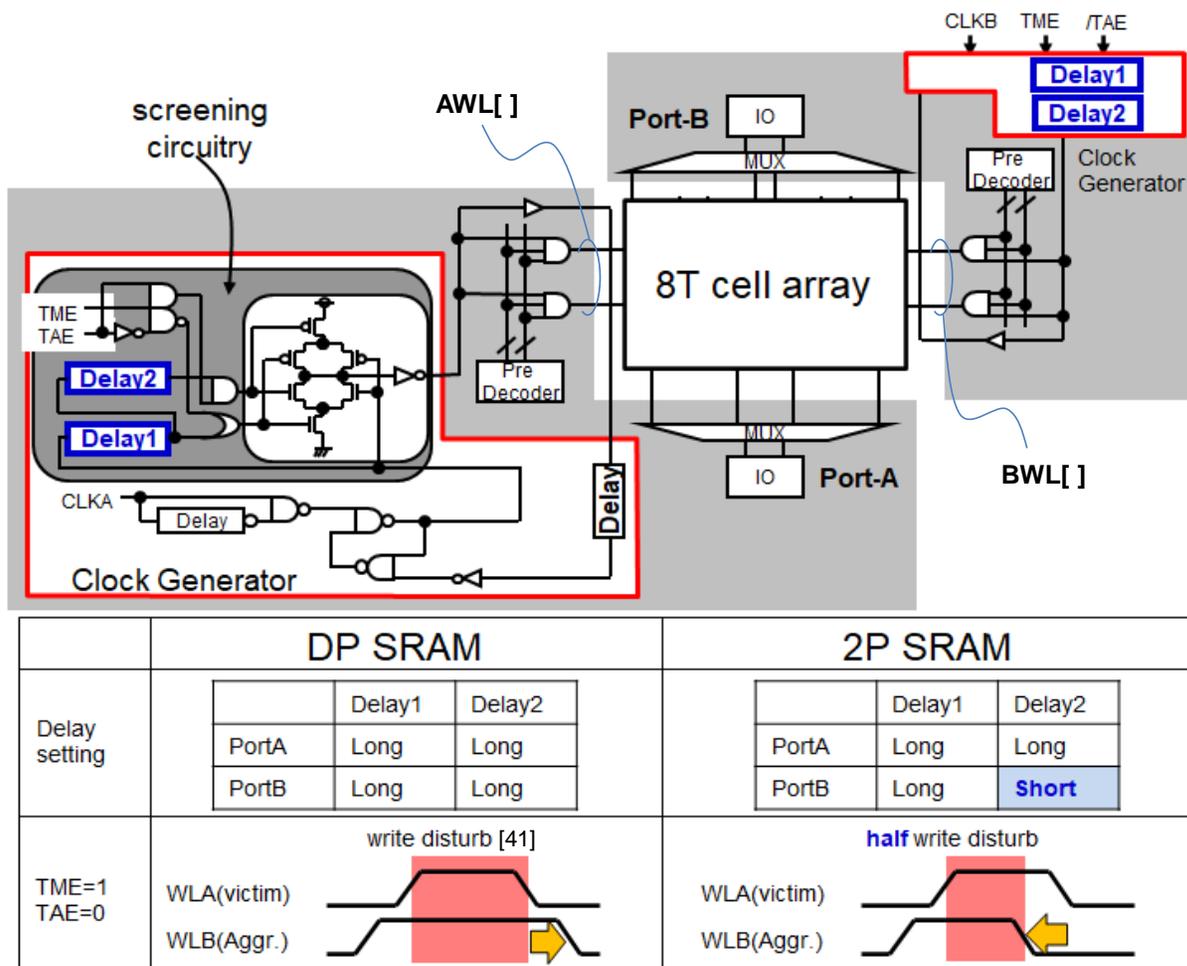
Fig. 72 Test screening circuits for DP and 2P SRAMs.

Fig. 73 presents simulated waveforms of the screening test circuit at the -40°C worst temperature condition. The aggressor WL fully covers the victim WL in the DP SRAM, whereas the aggressive WL is negated slightly earlier than the victim WL following Fig. 71 in the 2P SRAM. The write-margin in the test mode is slightly worse than that of the normal mode because the rising edge of aggressor WL covers that of the victim WL. In this way, the disturbance failures are prevented from outflowing. Fig. 74 presents V_{min} vs. the clock skew SHMOO of the designed 2P SRAM macro. The peak of V_{min} of normal mode is 0.9 V. V_{min} tuning for the write-disturbance is 1.1 V, drawn by the dashed red line. V_{min} in the case of tuning for the half write-disturbance is 0.9 V, drawn by the blue solid line. Write-disturbance tuning will overkill, but it will screen disturbance failures properly.
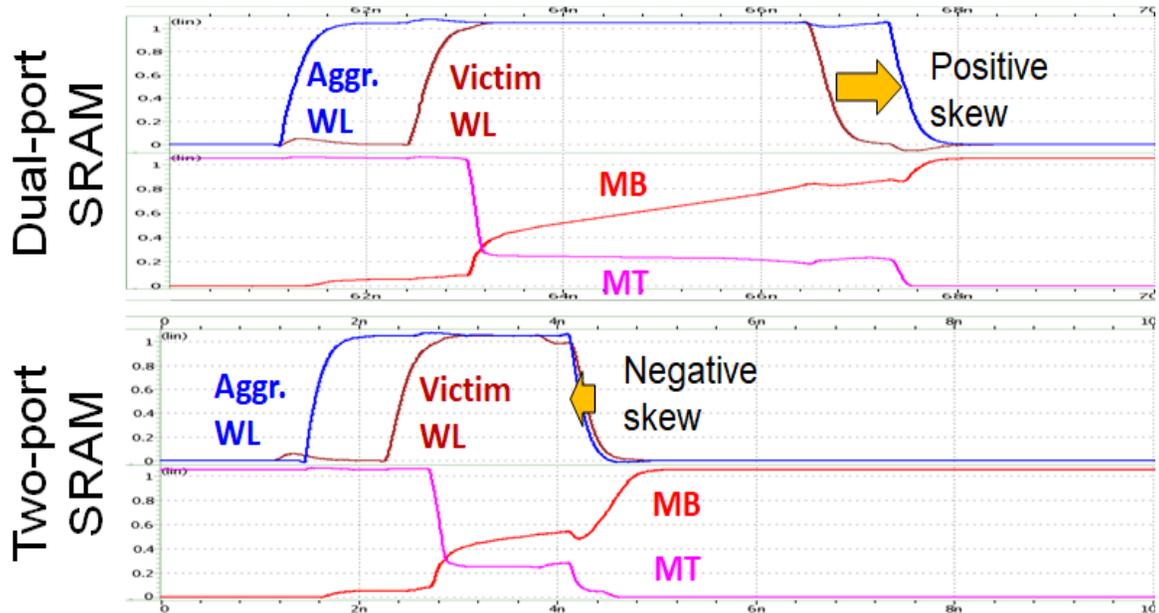
Fig. 73 Simulated waveform at screening test mode with 5.7 sigma local variation.
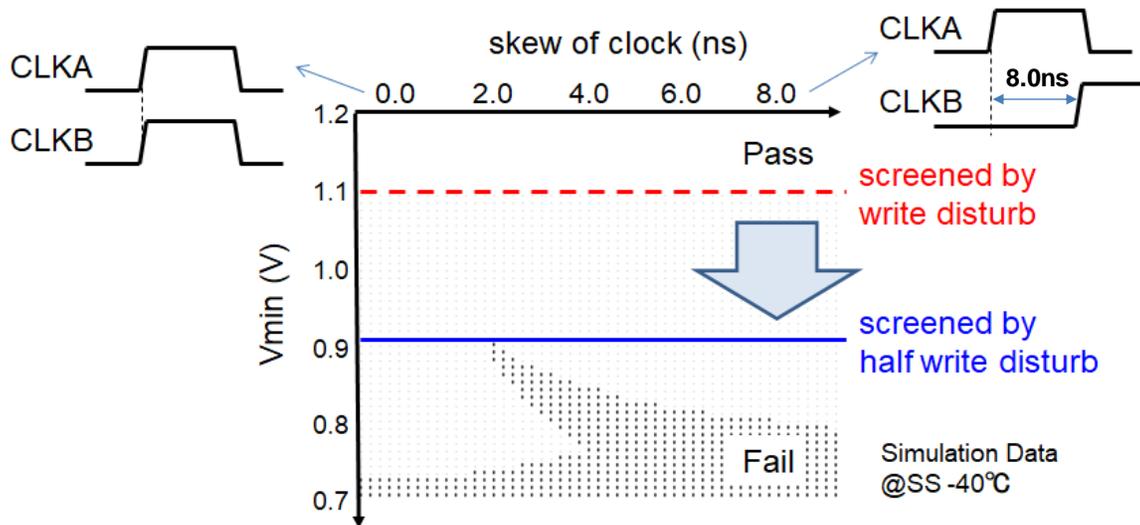
Fig. 74 Simulated V_{min} vs clock skew plot for 2P SRAM.

4.1.3. Design and Evaluation of Test Chip

Fig. 75 is a die photograph of the test chip using the 40 nm embedded flash CMOS technology [3]. Table 5 presents a summary of the test chip features. Sixteen 24-kbit DP SRAM macros (total 390-kbit) and a 6-kbit 2P-SRAM macros are implemented respectively in a test chip. Full

read/write functions were observed at wide temperatures of -40°C to 170°C. Fig. 76 presents a typical shmoo plot at 170°C, showing V_{min} vs. the access time. The access time of the measured 24-kbit DP SRAM macro is 1.67 ns, and a 6-kbit 2P SRAM macro is 1.15 ns at the typical supply voltage of 1.25 V.

Fig. 77 shows a cumulative distribution function (CDF) plot of V_{min}. The number of measured sample dies is 32/32/5 at -40/150/170°C, where we can see good distributions and sufficient operating margin to the 1.25 V typical supply voltage. All dies were fabricated in almost typical process conditions. Fig. 78 shows CDF plots of the normalized leakage power to 1-Mbit at 25/150/170°C. The leakage power of the DP and 2P SRAM are summarized. The median values of the measured leakage powers at 25/150/170°C are 36.9/3601/7059 μW/Mbit.

Fig. 79 presents shmoo plots of V_{min} vs. clock skew between port-A and port-B at 25°C. The worst V_{min} peak is observed at a specific clock skew condition for each DP and 2P SRAM in normal operation conditions. In test mode with the proposed screening test circuit, no V_{min} peak is found, with only slight worsening of the worst peak V_{min} of the normal mode. The proposed test screening is conducted appropriately without overkill screening.
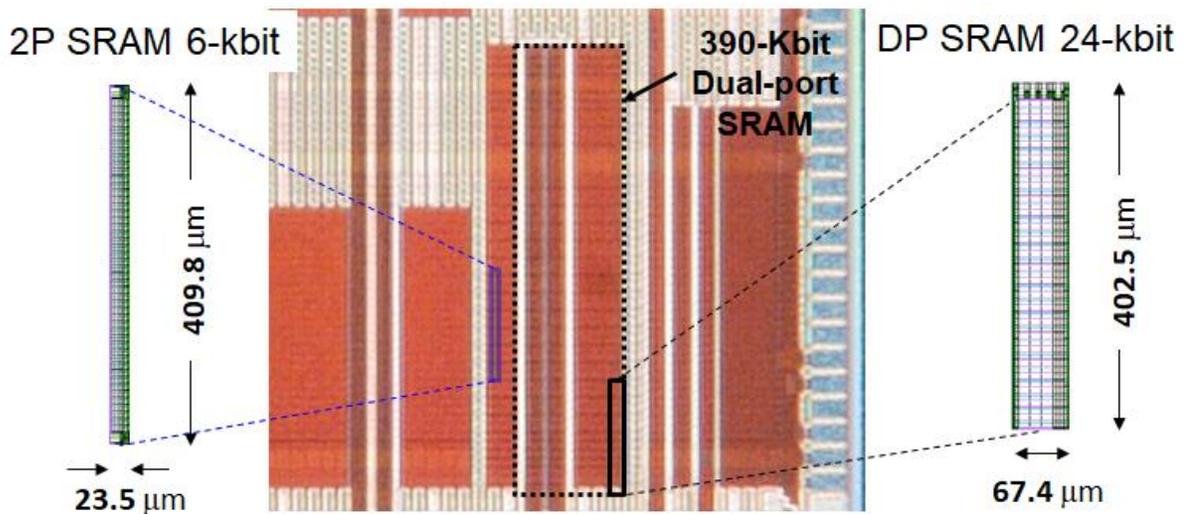
Fig. 75 Microphotograph and layout plots of the test chip.

Table 5 Features of the test chip

		Features
Technology		40 nm embedded flash process
Dual port SRAM	Macro configuration	24-kbits (512word x 48bit)
	Macro size	402.5 µm x 67.4 µm (27149 µm^2)@24-kbit
	Bit density	0.863 Mbit/mm^2
	Total capacity	390-Kbit (test chip)
Two port SRAM	Macro configuration	6-Kbits (128word x 48bit)
	Macro size	409.8 µm x 23.5 µm (9634 µm^2)@6-kbit
	Bit density	0.608 Mbit/mm^2
	Total capacity	6-kbit (test chip)
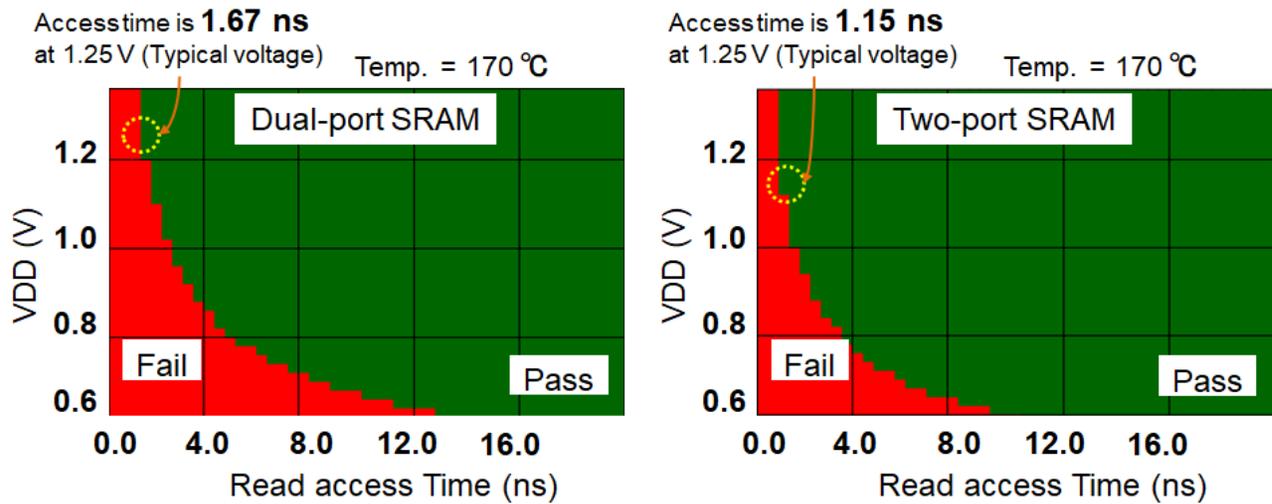
Fig. 76 SHMOO plots of V_{min} vs access time.
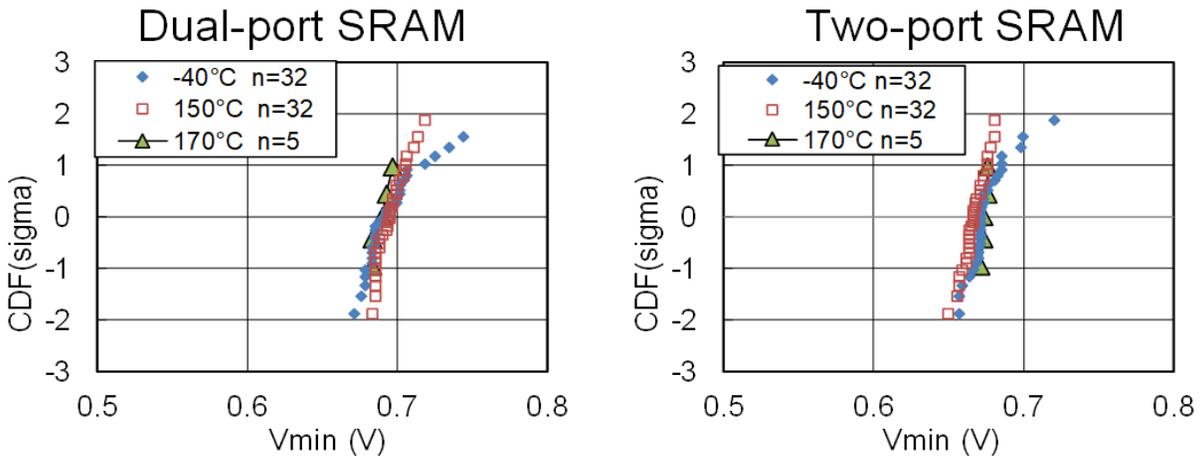
Fig. 77 Measured V_{min} distributions at -40/150/170°C.
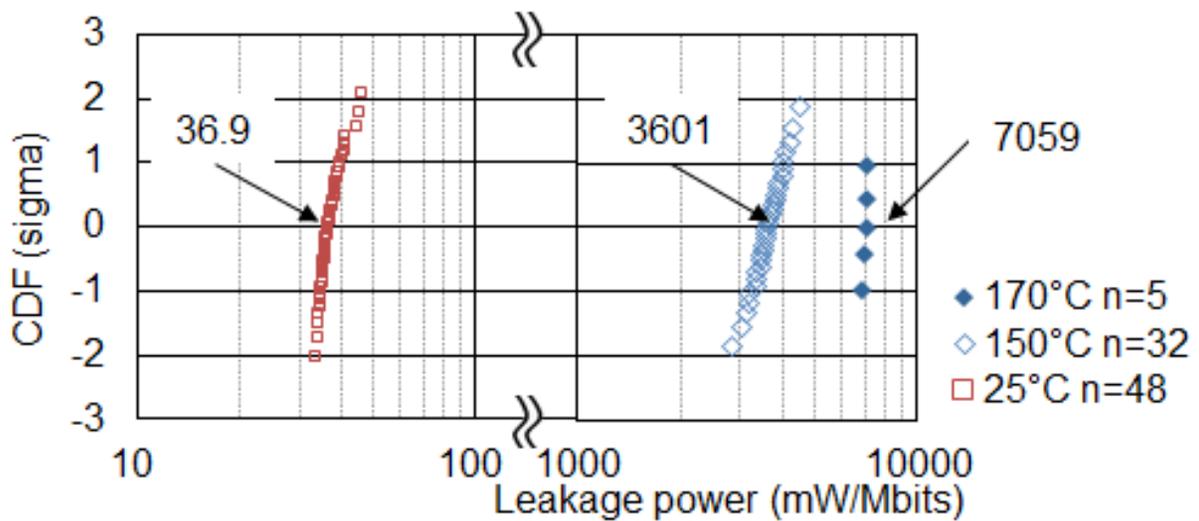
Fig. 78 Distributions of measured leakage power of 390-kbit DP SRAM and 6-kbit 2P SRAM at 25/150/170°C.
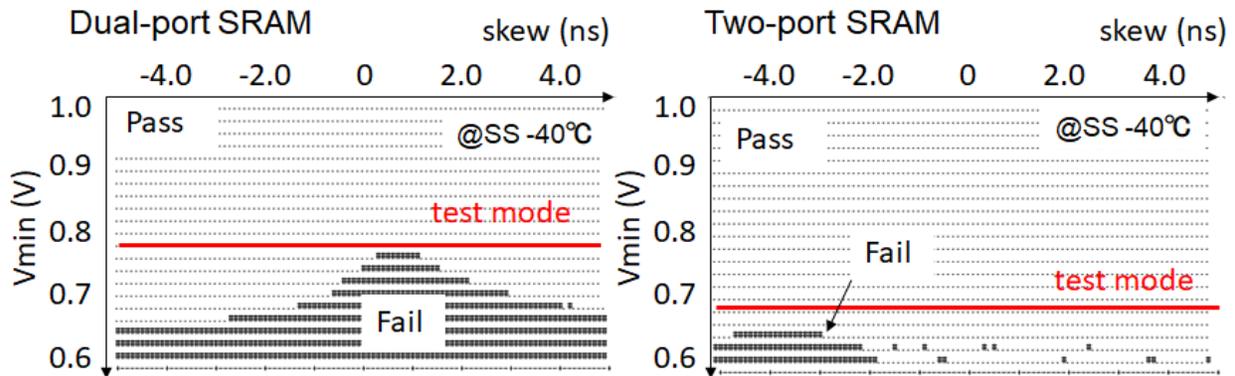
Fig. 79 Shmoo plots of V_{min} vs clock skew.

4.2. Dynamic Power Reduction in Synchronous 2RW 8T Dual-Port SRAM

4.2.1. Introduction

Along with the device scaling down, the multi-processing is more required than higher clock frequency f to achieve energy-efficient SoCs with high performance. Embedded SRAMs are key devices for reducing power consumption, especially being essential not only single-port but also multi-port access for parallel processing. As a data cache or shared cache memory, 2-read/write (2RW) dual-port (DP) SRAMs are used frequently for these multi-CPU and many-core architectures[41]-[43]. Furthermore, DP SRAMs also function in several applications such as FPGAs [44] or dynamically reconfigurable processors [45][46]. Many reports of the relevant literature describe studies related to the DP SRAM designs [47]-[53]. Enhancement of write/read margins for lower voltage operation, improvement of the density, and prevention or detection of disturbance issues between port accesses have been discussed mainly in earlier reports of the literature.

4.2.2. Detecting same-row and adjusting margin circuit

The probability of occurrence of the same row access is 1/row if the row address is accessed randomly. For example, it is 0.4% in the 256-row macro configuration. Although there is a small

probability of the same row access mode, it should be considered for maintaining the design margin. In other words, 99.6% of operations have excessive timing margins because the different row access has better timing margins for both reading and writing. This over-timing of margins consumes undesirable dynamic power by an extra discharge of BL capacitance. To eliminate this extra dynamic power consumption, an adjusting WL pulse timing control circuit is proposed. Fig. 80 presents a block diagram of the proposed circuit, where ARA is the row address of the A-port, and ARB is the row address of the B-port. If the input row addresses of port-A and port-B are same, and the output signal SR of the row address comparator (XNOR) becomes high. For that reason, the parasitic capacitance of the replica BL increases. As a result, the WL pulse width will be longer by increased replica delay. The triggered timing of the sense enable for the sense amplifier is also shifted in the read operation. In the write operation, the WL pulse width and enabling write driver period increase if the same row address is applied.
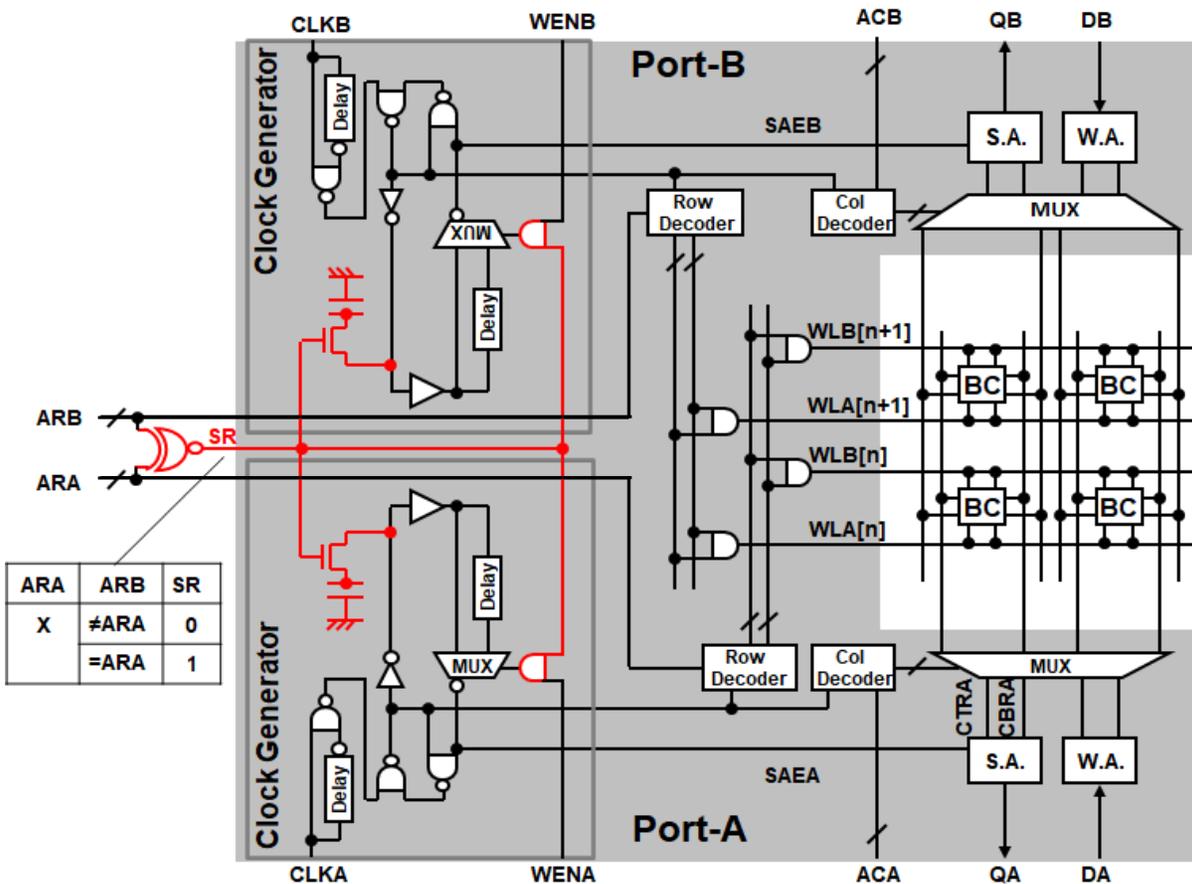
Fig. 80 Block diagram of proposed DP SRAM.
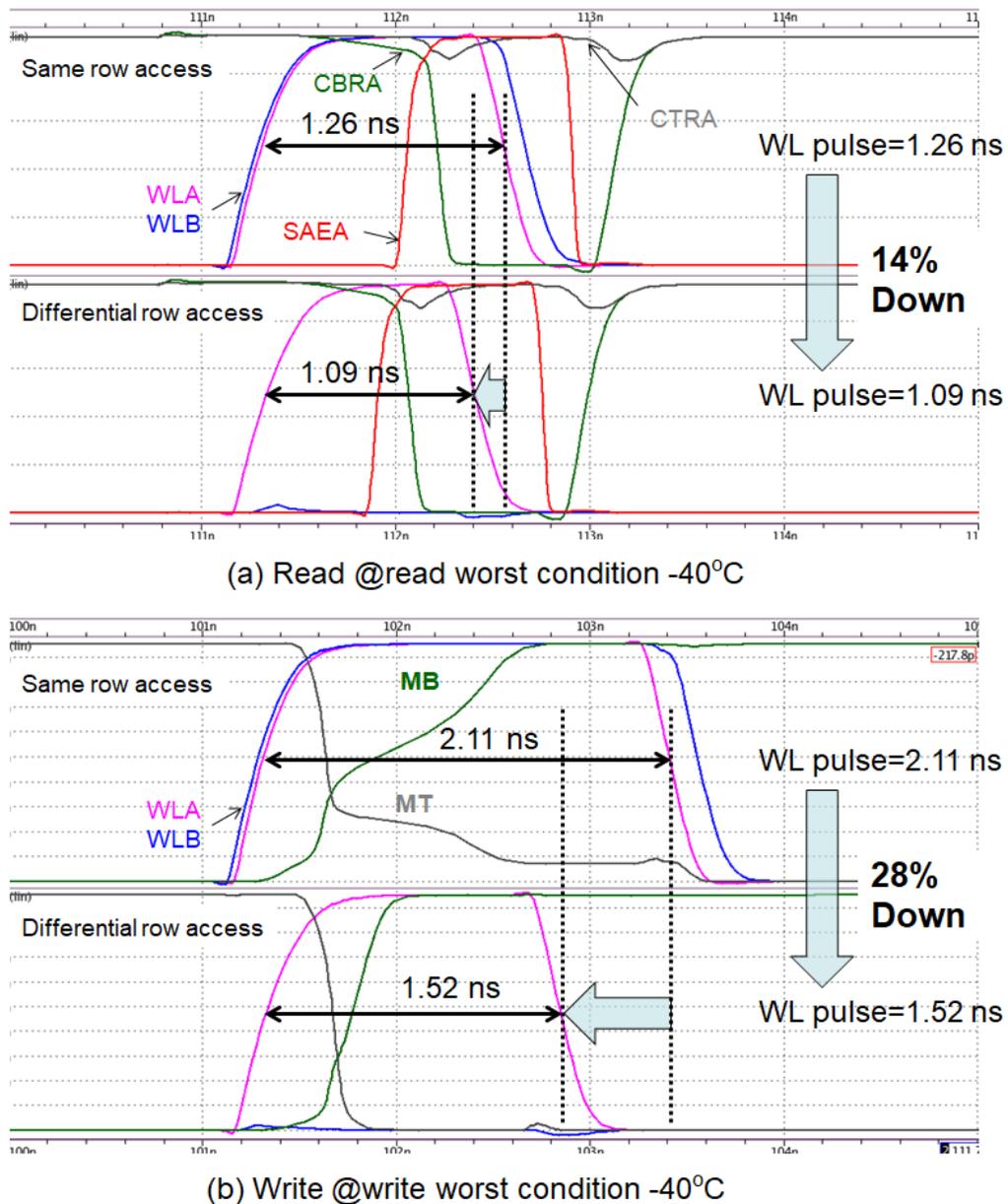
Fig. 81 Simulation waveforms of 38-kbit MUX8 (2048-word x 19-bit) with adjusting WL pulse timing control.

Fig. 81 (a) and Fig. 81 (b) respectively depict SPICE simulation waveforms at the worst PVT conditions in the read operation and write operation. The simulated WL pulse width for the read operation, as presented in Fig. 81 (a), can be shorten by 14% in the different row access mode compared to the same row access mode. Here, it is assumed the same BL swing: delta voltage

between the S.A. inputs of CBRA and CTRA in Fig. 81 (a) is almost identical to the different row access mode. However, the simulated WL pulse width for the write operation, as presented in Fig. 81 (b), can be shorten further by 28% in the different row access mode compared to the same row access mode. This simulation demonstrated that the write disturbance strongly affects the differences of WL width between the same and different row access modes. From the simulation results presented in Fig. 81 (a) and Fig. 81 (b), a dynamic power comparison is estimated in the two types of MUX configurations under the worst PVT condition. The MUX means the column address choose the bitlines as shown in the Fig. 22. Fig. 82 portrays the estimated results of expecting dynamic power reduction by the proposed circuits. In the read and write operations, the dynamic power is reduced by ~7% and ~18%, respectively.
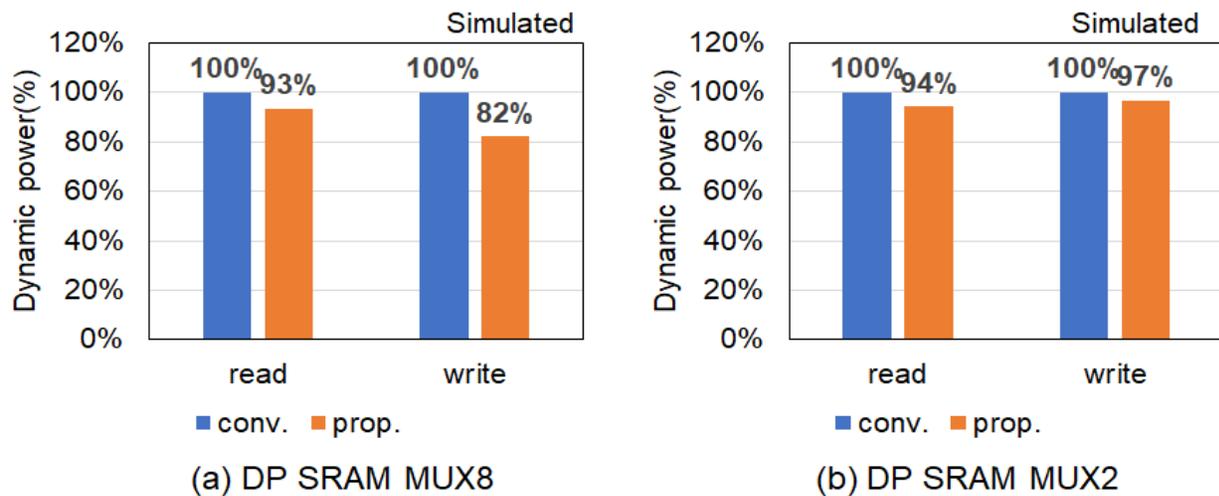
Fig. 82 Simulation result of dynamic power consumption at typical condition.

4.2.3. Detecting same-row and adjusting margin circuit

Fig. 83 is a die photograph of a test chip using the 40 nm CMOS technology. The DP SRAM macros of MUX2 and MUX8 are implemented in the test chip. Table 3 presents a summary of the test chip features. Full read/write functions at temperatures of -40°C to 125°C were observed. Fig. 84 (a) and Fig. 84 (b) show CDFs of V_{min} at -40°C for two macros. The total number of measured dies is 70. The median of V_{min} for TT process is 0.71 V for each macro. Fig. 85 shows the measured dynamic power consumptions of write and read operations vs. the supply voltage. The power consumption measurements obtained using the proposed circuitry are, 19.5 µW/MHz (read operation) and 19.6 µW/MHz (write operation), respectively, under the random address accessing (probability of the same row access is less than 4%), at the typical PVT condition (TT/1.1 V/25°C). The measured data show that the dynamic powers of proposed DP SRAM macros are reduced by 7% and 18%, respectively, for reading and writing operations compared to the conventional macros.

The proposed circuit in Fig. 80 has no disadvantage for access time or address setup time. Furthermore, the area overhead is less than 1%.
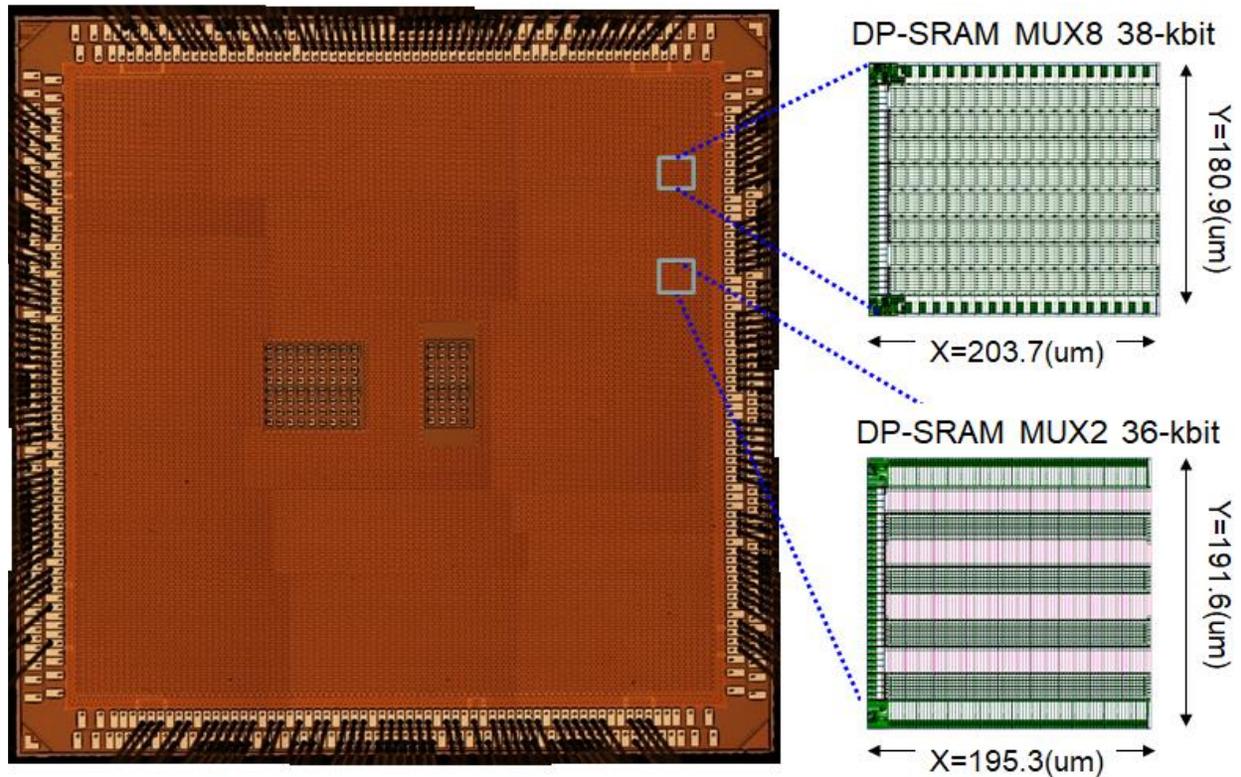
Fig. 83 Microphotograph of the Test chip and layout plots of proposed two DP SRAM macros using 40-nm technology.

Table 6 Features of the test chip.

		Features
Technology		40-nm CMOS process
Macro configuration	MUX2	36-kbit (512 word x 73 bit)
	MUX8	38-kbit (2048 word x 19 bit)
Macro size	MUX2	195.3 μm x 191.6 μm 37419 μm²@36-kbit
	MUX8	180.9 μm x 203.7 μm 36849 μm²@38-kbit
Bit density	MUX2	0.95 Mbit/mm²
	MUX8	1.01 Mbit/mm²
Access time @typical	MUX2	1.4 ns
	MUX8	1.6 ns
Cycle time @typical	MUX2	2.0 ns
	MUX8	2.3 ns
Dynamic power @typical	MUX8	Read: 19.5 μW/MHz Write: 19.6 μW/MHz

(a) Vmin @ -40°C (MUX2)   (b) Vmin @ -40°C (MUX8)
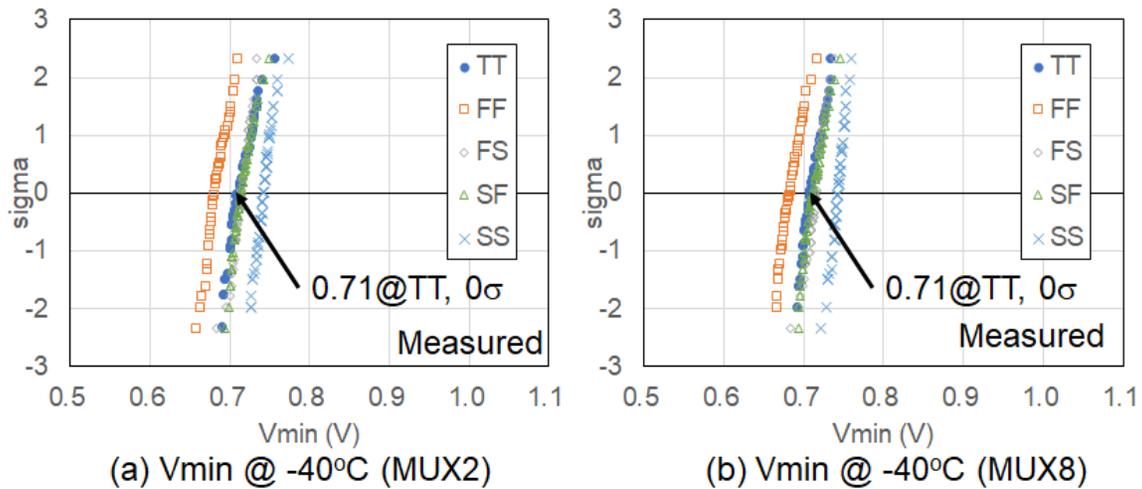
Fig. 84 Distribution of measured V_{min} at -40°C worst temperature.
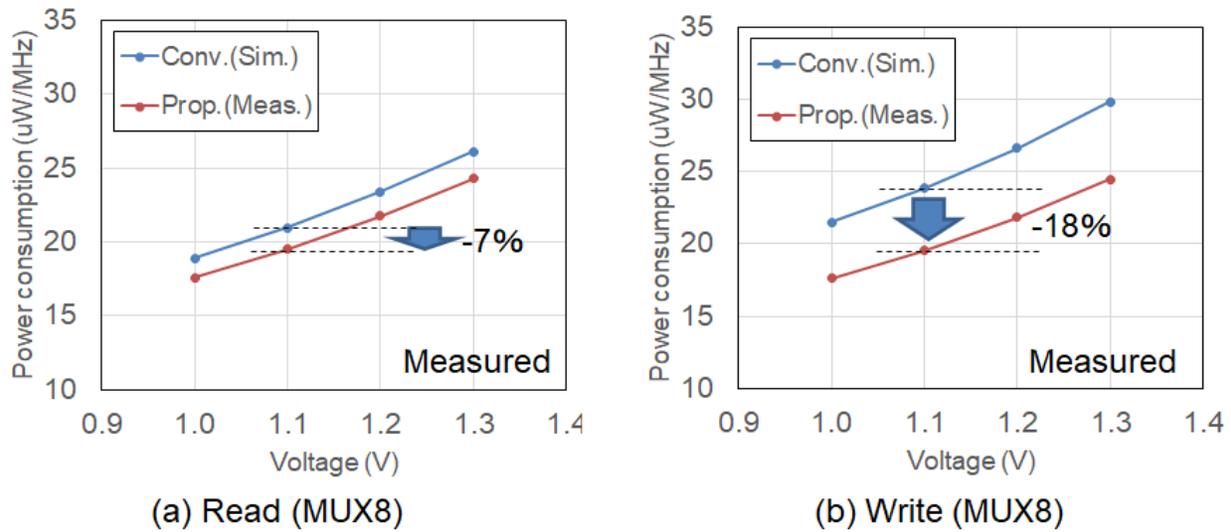
Fig. 85 Measured dynamic power vs supply voltage at 25°C, TT-process.

4.2.4. Conclusion

An adjusting WL pulse timing control circuit was proposed to reduce read and write dynamic power for a 2RW DP SRAM. Row addresses for port-A and port-B were compared, and the proposed circuit detected whether the row addresses are same or not, which is an inherent access mode of 2RW 8T DP SRAM design. Well-balanced 8T DP SRAM bitcell was demonstrated using 40-nm technology. Test chips including macros of two MUX types were designed and fabricated using 40-nm technology. The measured data show that read and write powers were reduced, respectively, by ~7% and ~18%.

Chapter 5 Pseudo Low Temperature Test

5.1. Introduction

With process technology scaling, on-die embedded memory densities increase annually as do logic gates [54]. Consequently, the testing cost per transistor increases year by year [15]. Furthermore, screening tests have become more complicated, requiring much longer testing time of embedded memories [55]. In microcontroller unit (MCU) markets, reducing the required testing times of digital logic blocks, IO interfaces, analog blocks, and memory blocks is crucially important for production of cost-competitive products [56]. Especially, because recent MCUs typically include one or more high-density single-port (SP) SRAM blocks for high-speed cache access, reducing testing times is important not only for embedded non-volatile memories such as flash, but also for embedded SRAMs. Dual-port (DP) SRAMs are used frequently as buffer memories in interface blocks and for image processing hardware accelerators [57],[58]. For these embedded SRAMs, various testing cost reduction methods have been proposed, such as using effective testing patterns, formulating test times for data retention, and using parallel BIST [59], [60].

A low failure rate must be ensured after test screening and shipment. Device characteristics show temperature dependence [61] such that testing at low and high temperatures are indispensable to screen dies, which exhibit temperature-dependent failures [62]. Additionally, package testing after die sawing and assembly are generally performed at room temperature (RT). Therefore, dies should be tested under at least three temperature conditions: high temperatures (HT) of 125°C, RT of 25°C, and LT of -40°C. Particularly, embedded SRAMs show different failure modes at HT and LT because of the different temperature dependencies of reading and writing margins and the static-noise margin (SNM) [26],[63],[64]. Therefore, the embedded SRAM should be tested at both LT and HT. However, changing the temperature during wafer probing test takes much time, thereby increasing costs directly. The LT test is particularly costly because of its testing environment. Fig. 86 presents testing flows with three temperature steps and two temperature steps as described herein. In this section, a cost-effective test screening with only two temperature conditions (HT and RT) is proposed. It can skip the LT test by introducing pseudo-LT (PLT) testing at RT [65].
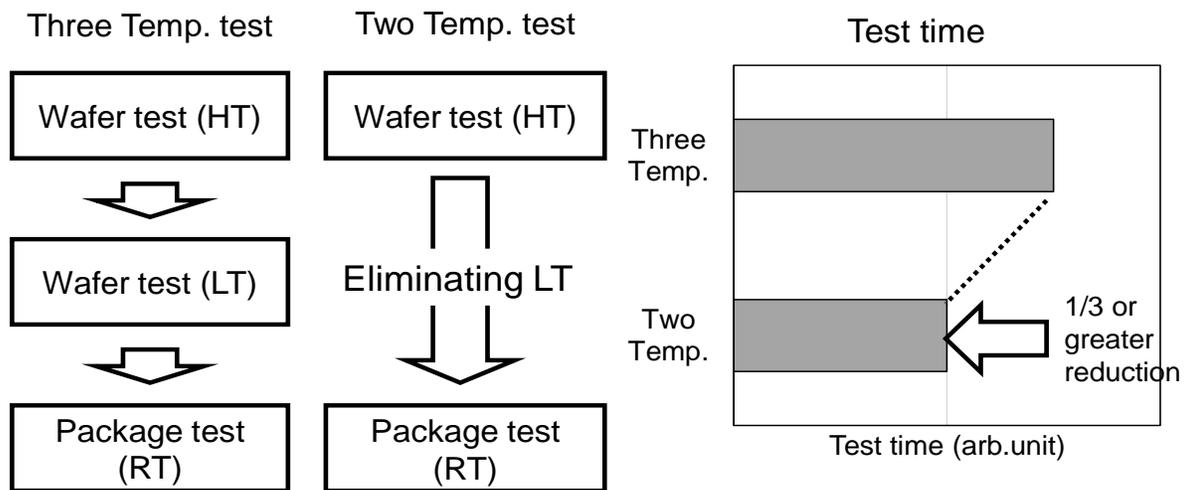
Fig. 86. Test screening flows with three temperature steps (typical) and two temperature steps (eliminating LT).

5.2. Overscreened Issues of Voltage Guard Band Techniques

In testing flows with either three temperature steps or two-temperature steps as shown in Fig. 86, pin contact testing is followed by wafer probing tests at HT first. It is typically executed to screen out SNM failures at the read operation in the embedded SRAM and to screen out leakage failures that occur during the standby mode. Solid failures caused by defects, which are not dependent on the temperature change, are also the screened out by the first wafer probing test. Next, in the three temperature steps, the wafer probing test at the LT is executed. Then the package test at RT is conducted after die sawing and package assembly. Elimination of second testing in LT conditions is effective at reducing testing time. The first HT testing is a necessary step for screening out the worst leakage conditions. The test after packaging cannot be skipped because it removes assembly failure dies.

Testing at LT is performed mainly to detect failure bits that have less write margins at the write operation in the embedded SRAM. As described above, the SRAM minimum operating voltage (V_{min}) at read operation (read-V_{min}) becomes worse at HT because of SNM temperature dependence, whereas the SRAM V_{min} at the write operation (write-V_{min}) becomes worse at LT. Here, this section introduces how to screen out the write-V_{min} failures at RT. By lowering the voltage in the test mode, one can perform more stringent tests and screen dies with poor operating margins in CMOS logics

[66],[67]. The lowered voltage is also effective for embedded SRAMs. Fewer write margin bits can be screened out at RT testing with additional appropriate guard band (GB) voltage. However, voltage GB technique sometimes induces overscreening, leading to undesirable yield loss [68]. Especially, the failure SRAM bits at LT has no constant write-V_{min} offsets between LT and RT conditions because variations of write-V_{min} offsets are caused not only by different temperature dependencies of MOS characteristics but also by abnormal contact-diffusion high resistances (soft open) in write operations.

Fig. 87 (a) and (b) respectively depict soft open failure models of contact-diffusion high resistance in the 6T SP and DP 8T SRAM bitcells during write operations. Here, there are SP SRAM bitcells of two types. One is the high-density (HD) type for a compact area. The other is high-current (HC) type for high-speed caches. Both bitcell layouts have identical topology, except for the values of transistor sizes. Some contact holes in a bitcell might take part in the write failure under LT conditions. As shown in Fig. 87 (a), if an abnormal contact hole is connected to either source or drain nodes of pass-gate NMOS: PGT or pull-up PMOS: PUB, which have high resistance of more than several kilo-ohms, then the write-V_{min} of the bitcell is much worse than that of RT conditions. Neither PDT nor PDB affects write operations. The same model is used for DP 8T SRAM bitcell, as shown in Fig. 87. Fig. 88 shows SPICE simulation results of temperature dependencies of write-V_{min} with PGT=0 Ω and 10 kΩ for a 6 T SP bitcell. The V_{min} offset between RT and LT with PGT=0 Ω, is 60 mV, whereas that with PGT=10 kΩ is 240 mV.

If V_{min} of each die is shown in a graph with V_{min} at RT on the x-axis and V_{min} at LT on the y-axis, then V_{min} is classified into four quadrants by boundary voltage for non-failure products, as shown in Fig. 89 (a). Dies in the third quadrant are non-failures that have passed both RT and LT testing. The dies in the first, second, and fourth quadrants are failed dies, which failed during either RT or LT testing. The blue regions in Fig. 89 (b) show images of V_{min} distributions. Actually, RT testing should be executed at the screening test with GB voltage (shown as the red line in Fig. 89 (b)) to detect all failure dies in the second quadrant. Otherwise, some failure dies in LT conditions cannot be detected in the RT testing with no GB voltage. Some dies are between GB voltage and typical voltage in the third quadrant, as shown in Fig. 89 (b). These dies are overscreened if RT testing with GB voltage would be applied instead of LT testing at typical voltages.
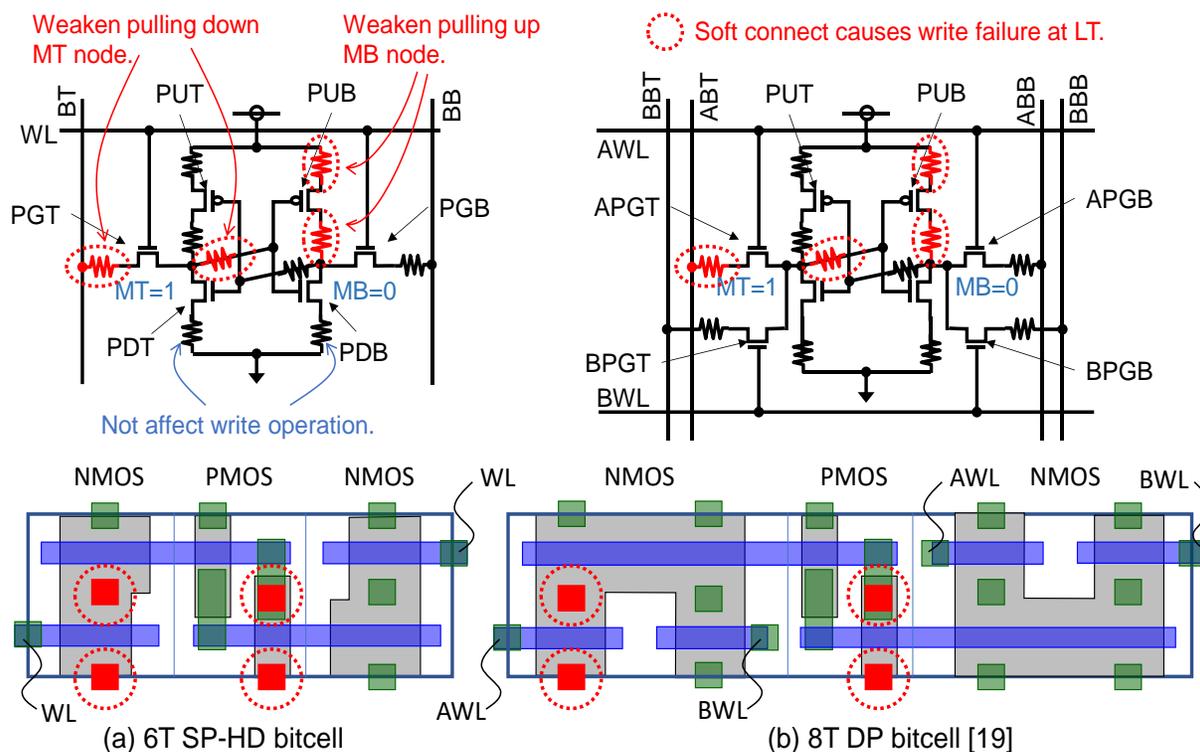
Fig. 87. Write failure models with abnormal contact-diffusion high resistance. Red rectangles are contacts with high resistance.
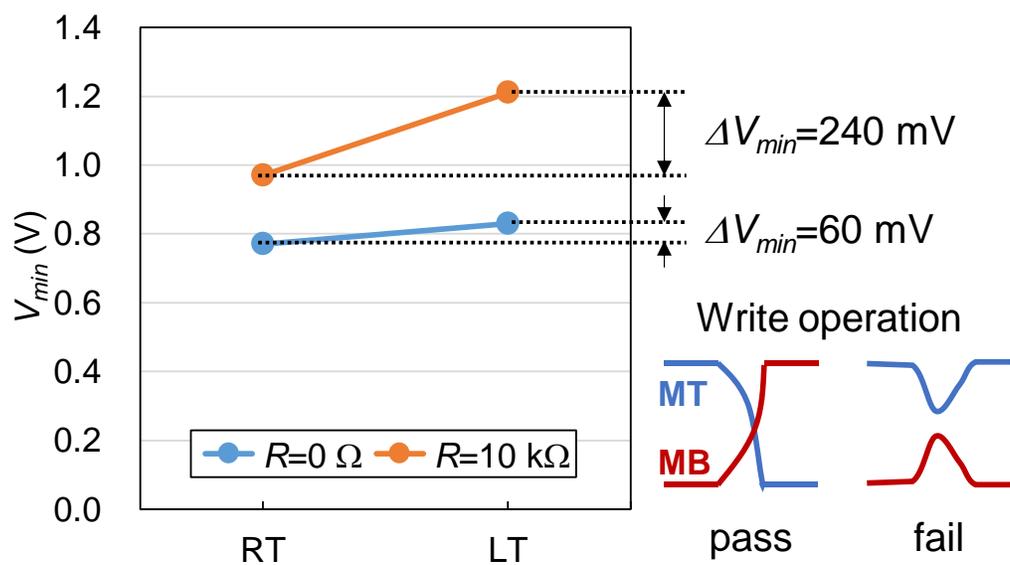
Fig. 88. Temperature dependence of write-V_{min} with normal resistance contact (R=0 Ω) and

abnormal high-resistance contact (R=10 kΩ).
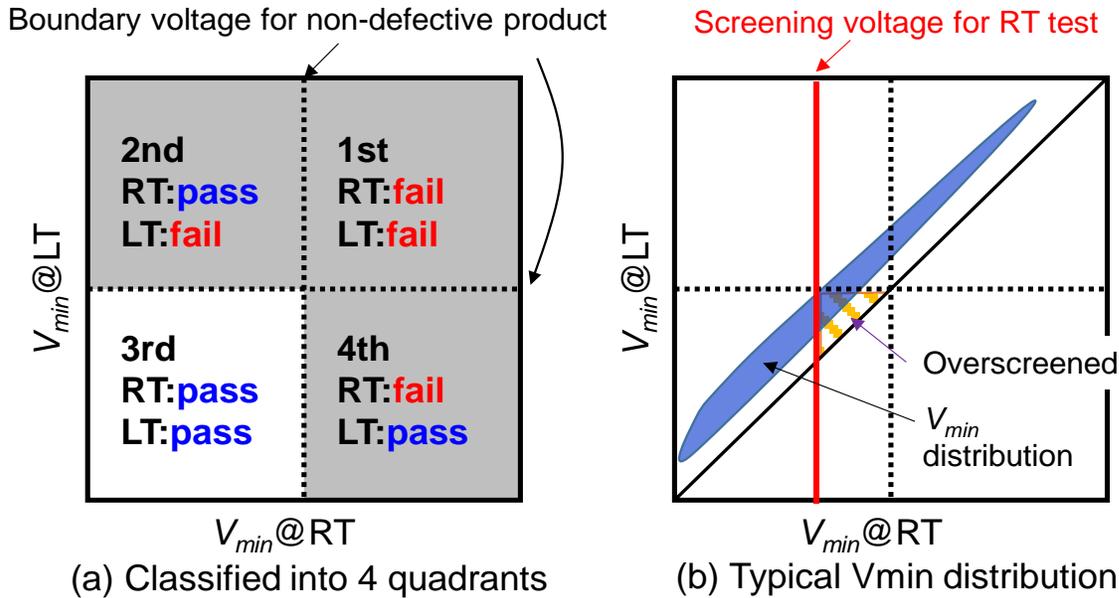
Fig. 89. Pass and fail categories by V_{min} at RT vs. LT.

5.3. Proposed Pseudo Low Temperature Test

In this section, a PLT test method is proposed to eliminate the LT test flow. Fig. 90 presents the concept of imitation of the write disturbance condition at LT. The bitcell construct and transistor names are the same as those shown in Fig. 87. Fig. 90 (a) depicts the proposed test method: "PG test". In general, Vt is defined as the drain–source voltage (V_{GS}) when the current drain–source (I_{DS}) reaches a certain value. Therefore, Vt can be pseudo-shifted by lowering the WL voltage. The wordline (WL) voltage is lowered to reproduce LT conditions in which the threshold voltage (Vt) of PGT and PGB is raised. Fig. 90 (b) depicts another proposed test method: "PU test". Similarly to the PG test, the bitline (BL) is raised to reproduce LT conditions of PUT and PUB for the write operation. The number of overscreened bitcells can be reduced using this technique. An 8T DP bitcell has two WLs and BLs to achieve read/write operations simultaneously and asynchronously. The 8T DP bitcell has a "disturbance issue" [40],[70]. The worst case of DP SRAM write operation arises when WLs for the A-port and B-port are activated simultaneously. The BLs of the port that does not execute the write operation is charged to the high level. Similarly to the 6T SP bitcell, the GB voltage can be estimated using accelerated Monte Carlo simulation.
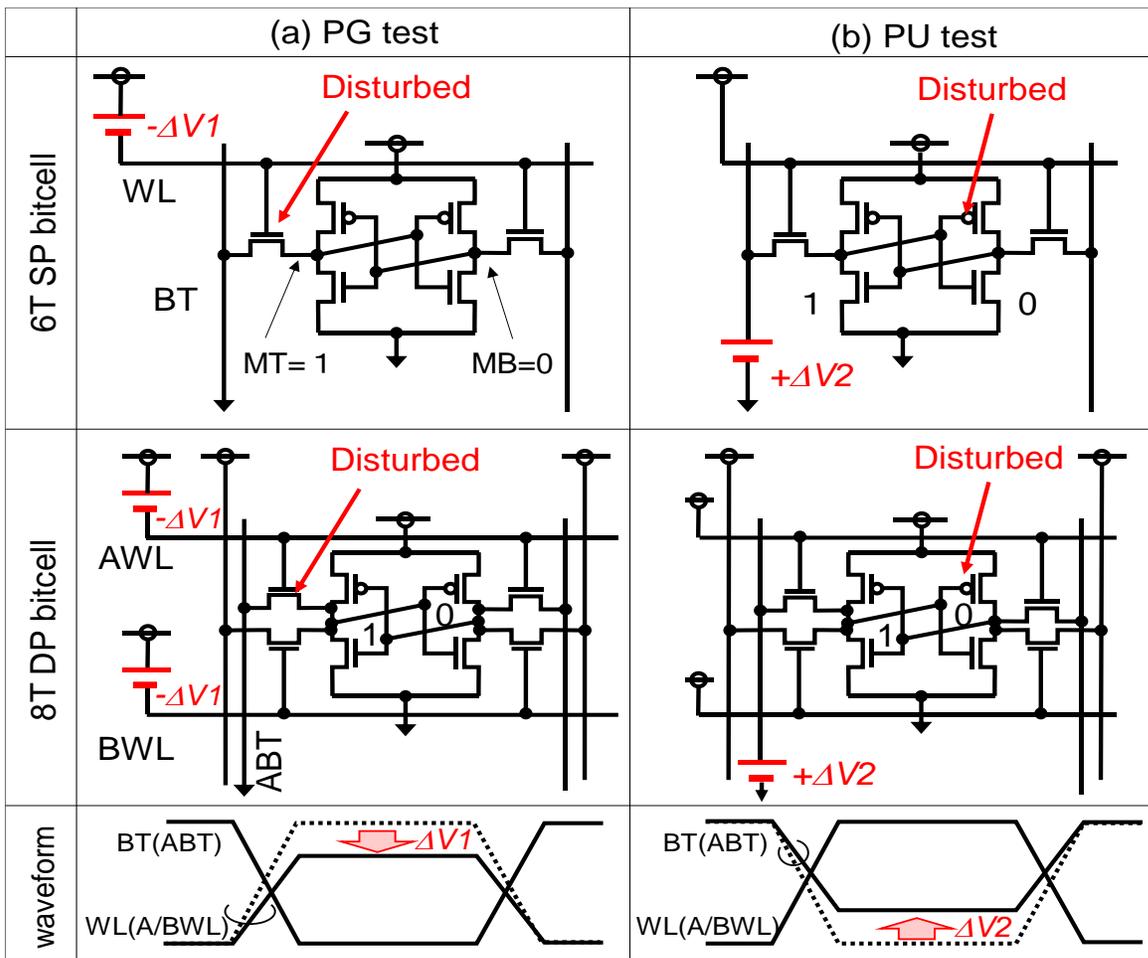
Fig. 90. Concept of reproducing the LT test conditions.

Fig. 91 (a) presents simulation results for RT vs. LT write-V_{min} of 40 nm 6T SP-HD bitcells with resistance at PGT or PUB in the worst process conditions (SF: slow NMOS and fast PMOS). A simulation with N=4000 iterations with three times accelerated sigma-Vt was carried out using Monte Carlo simulation. The N means a number of dies. In this case, N=4000 has variation that is equal to about 10 sigma. The horizontal axis shows V_{min} at RT. The vertical axis shows V_{min} at LT. If the boundary voltage for non-failure dies is 1.05 V, then the GB voltage at RT can be expected to be lower than 0.91 V to screen out defects appropriately. The "worst cell" in Fig. 91 (a) determines the GB voltage to screen out low-temperature failures at RT. The estimated percentage of overscreened dies is 3.0%. Cross plots in Fig. 91 (a) portray samples that pass the RT test and which fail the LT test if voltage GB technique and PLT would be applied.

Fig. 91 (b) presents simulation results of write-V_{min} distributions at LT vs. PLT, given by setting $\Delta V1$ = 30 mV and $\Delta V2$ = 20 mV, respectively, as shown in Fig. 90. $\Delta V1$ and $\Delta V2$ are calculated by

the difference of *Vt* between LT and RT with SPICE simulation. They depend on the bitcell transistors specification. Compared to the original V_{min} distributions of RT vs. LT (Fig. 91 (a)), the GB voltage between LT vs. PLT is smaller, narrowing the distributions. By finding the optimal disturbance conditions of *ΔV1* and *ΔV2*, the good V_{min} correlation between RT (PLT) and LT is obtainable. In this case, only 10 mV V_{min} offsets are observed, reducing the number of overscreened bitcells by 76.6% compared to the conventional method conducted using GB voltage only. Fig. 92 presents simulation results for RT vs. LT write-V_{min}. The worst process conditions for the 8T DP bitcell write operation are "SS: slow NMOS and slow PMOS" [70]. Slightly remained over-screened bits in the proposed scheme of Fig. 91 and Fig. 92 are derived from some variations of Vt shift from RT to LT.
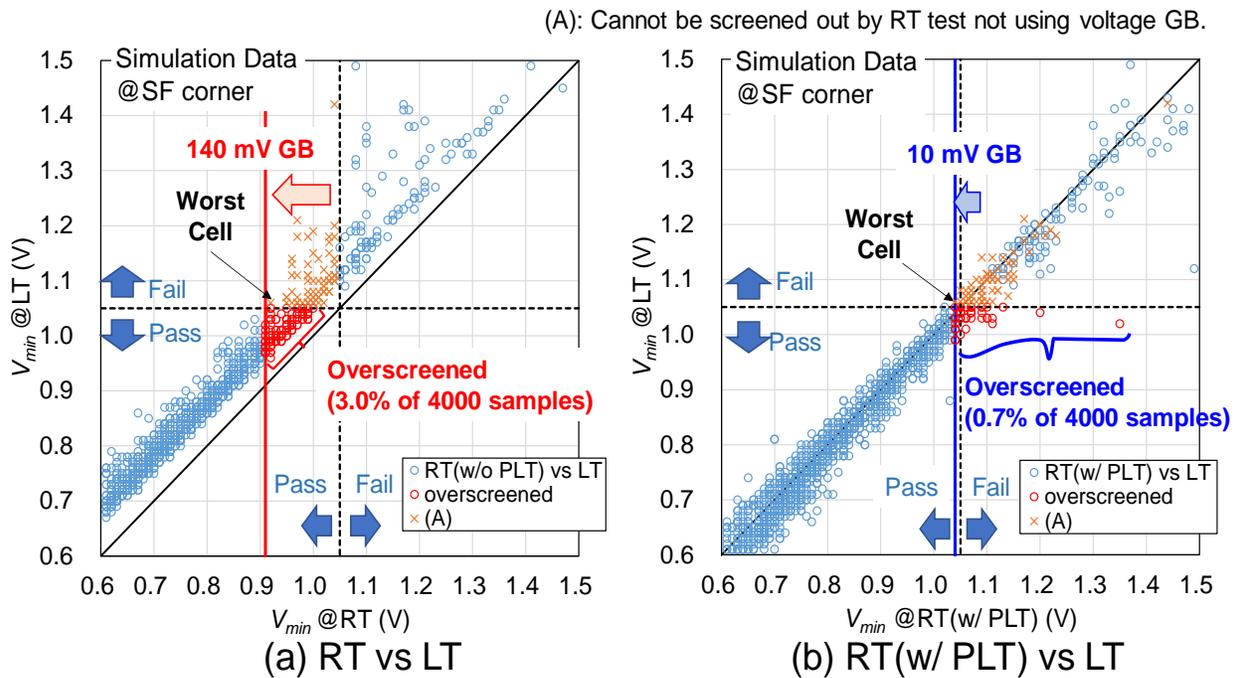
Fig. 91. Simulated write-V_{min} distribution of 6T SP-HD bitcell: (a) RT vs. LT and (b) RT with PLT vs. LT.
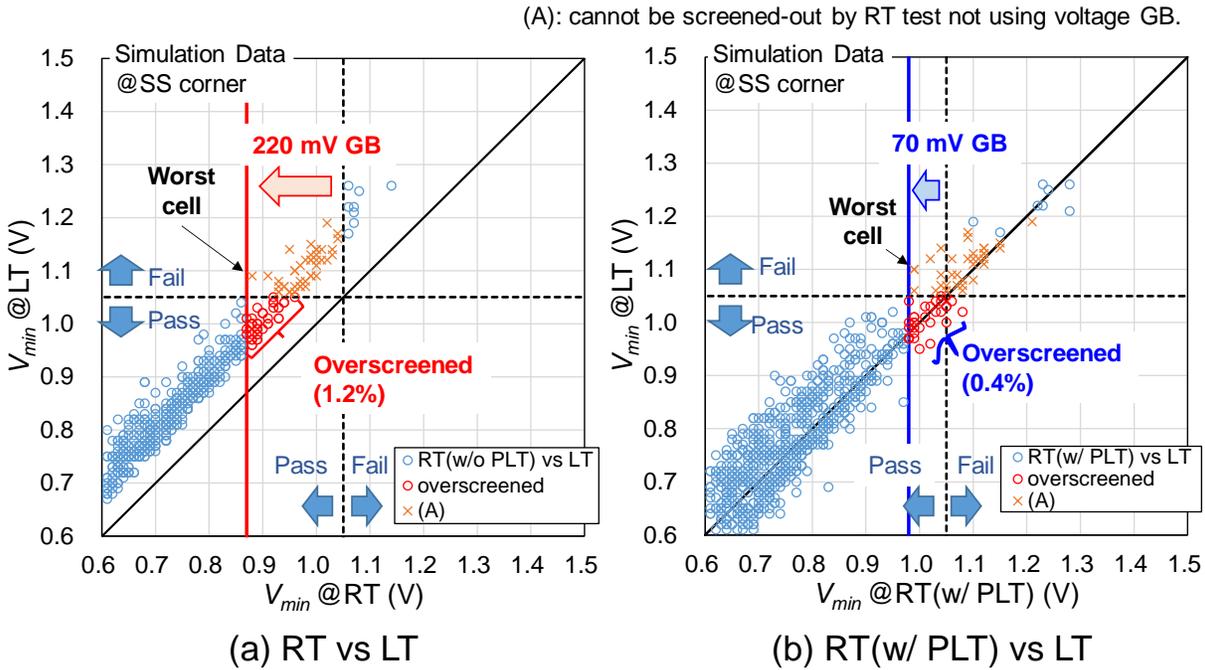
Fig. 92. Simulated write-V_{min} distribution of 8T DP bitcell: (a) RT vs. LT and (b) RT with PLT vs. LT.

Fig. 93 portrays a schematic diagram of SP SRAM macro with the proposed PLT test circuitry. To control the pseudo-test mode, two test input signals are added: TPU and TPG. When the TPU signal is enabled ("1"), pull-up NMOSs "MNBB" and "MNBT" in Fig. 93 connected to each write amplifiers (WA in Fig. 93) turns on. The corresponding BL voltage rises slightly from the VSS level during the write operation. Similarly, when the TPG signal is enabled ("1"), the weak pull-down PMOS MPW connected to LCVDD, which is the source node of each WL with pulled-up always-on PMOS PSW, turns on to slightly lower the WL voltage from the VDD in the write operation. The corresponding WL voltage decreases slightly from the VDD level in the write operation. This PLT test circuity can also be adapted for the 8T DP bitcell, similarly to the 6T SP bitcell. Fig. 94 presents circuit simulation waveforms of the 6T SP-HD bitcell under the PLT test condition at RT. The WL voltage (signal WL) is lowered by about 30 mV in the PG test mode. The BL voltage (signal BT) is raised by about 20 mV. Results show that the flipping time of the internal nodes (signals MT and MB) in the PLT test mode resembles that of the nodes in LT test conditions.

Table **7** presents $\Delta V1$ and $\Delta V2$ settings shown in Fig. 90. The same settings of $\Delta V1$ and $\Delta V2$ are applied for both 6T SP-HD and SP-HC bitcells because those transistor characteristics are very similar. Different settings of $\Delta V1 = 33$ mV and $\Delta V2 = 30$ mV are applied for 8T DP bitcell. This is because of the threshold voltage differences between 6T SP bitcell and 8T DP one and disturbance condition by both A-port and B-port activation in 8T DP one. The distributions of V_{min} of the DP SRAM, as shown in Fig. 92 are broader than that of 6T SP SRAMs in low voltage regions. It is

caused by the disturbance issue that is an inherent issue of the 8T DP bitcell [40],[70]. The calculated GB voltages for the conventional test flow and the proposed test flow are, respectively, 220 mV and 70 mV.
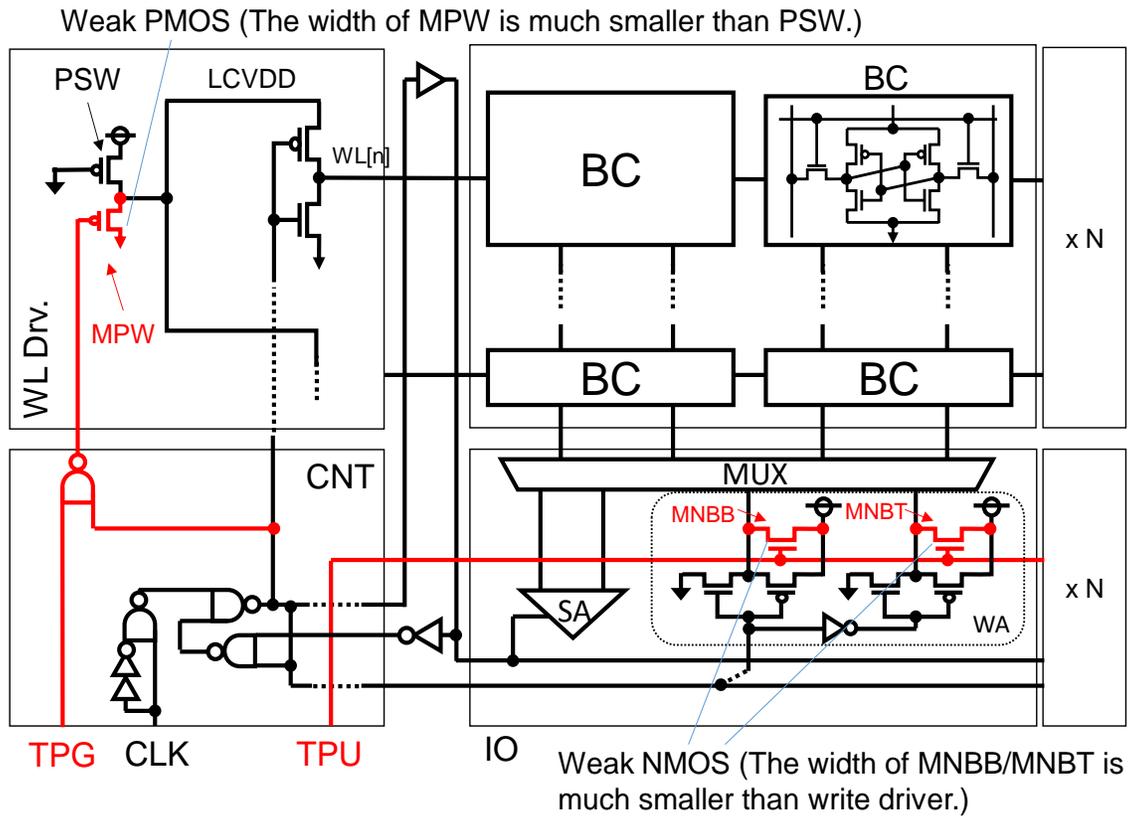
Fig. 93. Schematic diagram of SP SRAM macro with PLT circuitry.
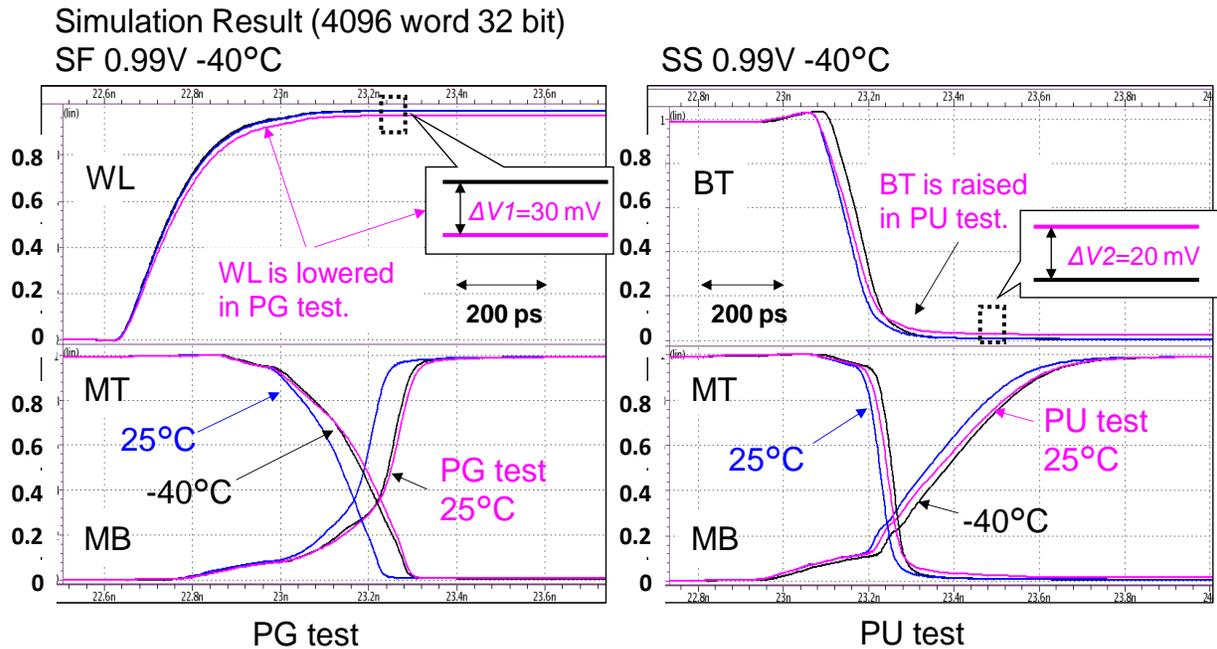
Fig. 94. Simulated waveform of write operation in PLT of 6T SP-HD bitcell.

Table 7 BIAS SETTING FOR EACH BITCELL IN THE PLT TEST MODE

	ΔV1	ΔV2
6T SP-HD bitcell	30 mV	20 mV
6T SP-HC bitcell		
8T DP bitcell	33 mV	30 mV

5.4. Design and Evaluation of 40 nm Test Chip

Fig. 95 portrays a photograph of the test chip used for the proposed SRAM macros. For targeting low-power MCUs intended for IoT or wearable markets, the test chip was fabricated using 40 nm low-power CMOS technology. Table **8** presents a summary of the test chip features. Thirty-two

128-kbit SRAM macros (total 4-Mbit) using 6T SP-HD bitcell, 64 64-kbit SRAM macros (total 4-Mbit) using 6T SP-HC bitcell, and 64 16-kbit SRAM macros using 8T DP bitcell are implemented in a test chip. The area overhead of PLT test circuitry is less than 0.01% in each SRAM macro. Power overhead of PLT is about 10%; PLT is used only for the screening test.
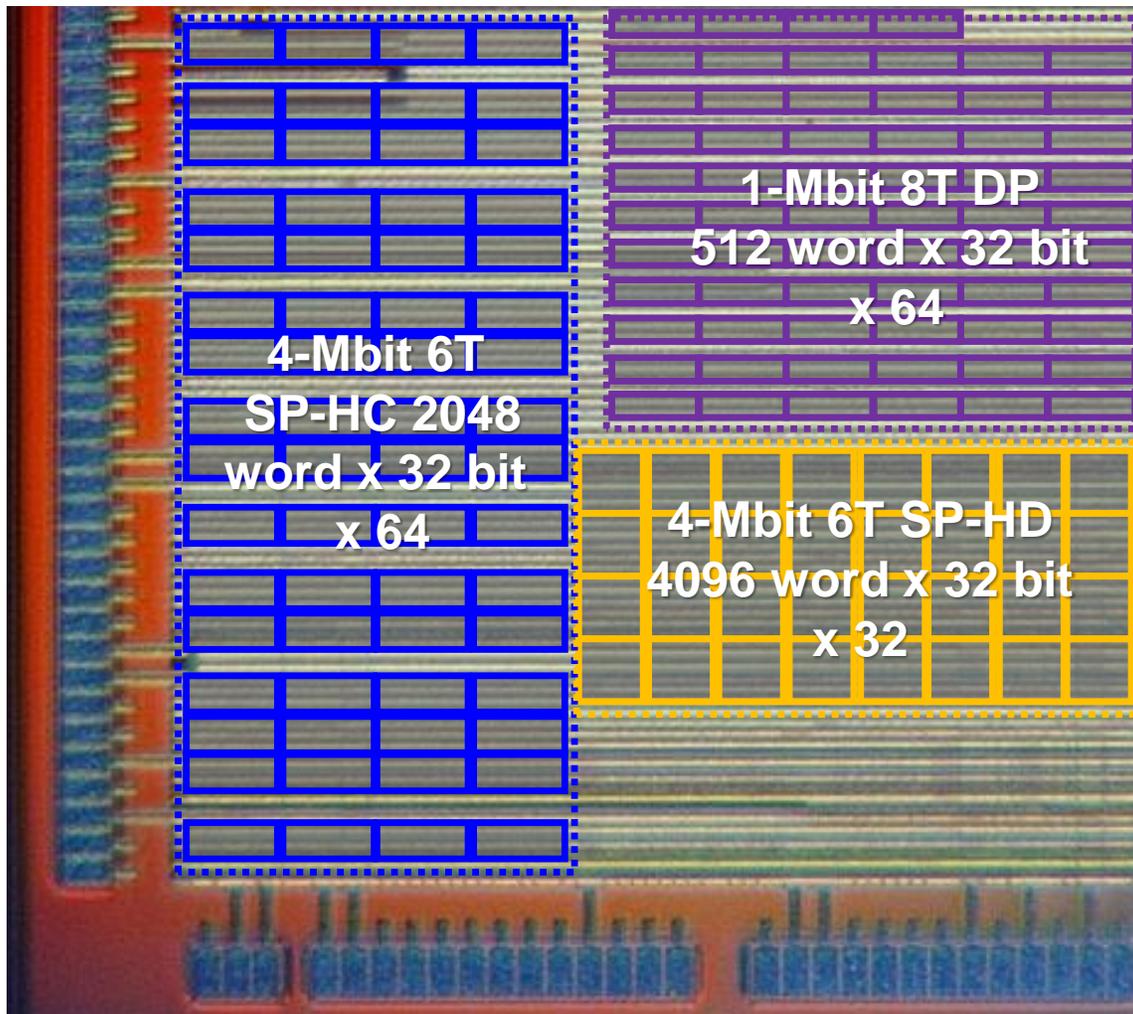
Fig. 95. Microphotograph of the test chip using 40 nm low-power CMOS.

Table 8 TEST CHIP FEATURES

	Bitcell type		
	6T SP-HD	6T SP-HC	8T DP
Technology	40 nm low-power CMOS process		
Macro configuration	128-kbit (4096 word x 32 bit)	64-kbit (2048 word x 32 bit)	16-kbit (512 word x 32 bit)
Macro size of instances	207.3 μm x 183.9 μm (38121 μm^2)	291.1 μm x 111.5 μm (32468 μm^2)	274.3 μm x 67.1 μm (18398 μm^2)
Bit density	3.28 Mbit/mm^2	1.92 Mbit/mm^2	0.85 Mbit/mm^2
Total capacity	4 Mbit	4 Mbit	1 Mbit
Access time @SS 1.04V -40°C	2.92 ns	0.98 ns	2.32 ns
Leakage power @FF 125°C	3742 μW/Mbit	8140 μW/Mbit	13356 μW/Mbit

Fig. 96 (a) shows measured V_{min} distributions of 6T SP-HD bitcell at RT and LT after HT testing for 140 mV GB w/o PLT and 10 mV GB w/ PLT. In all, 305 dies are measured, but HT failure dies are not shown in the graph. Similarly, to Fig. 91, it is apparent that the distribution of V_{min} is shifted by PLT on the LT=RT line. There are 29 overscreened dies used in the conventional method w/o PLT flow, although only 1 overscreened die is observed using the proposed method with PLT. Similarly to the 6T SP-HD bitcell, measured V_{min} distributions of 6T SP-HC bitcell and 8T DP bitcell are shown in Fig. 96(b) and (c). Table 9 presents the number of screened out dies for each testing step. We can see that test cost of PLT is 66% of that of the conventional 3 temperature test. The number of overscreened samples of PLT is less than the conventional voltage GB test. The same GB voltage of 6T SP-HD bitcell is applied to the 6T SP-HC bitcell because both schematics and layout topologies are equal. They have the same write failure modes.

To adopt the PLT method for a leading product in early production stages, many measured data are collected in the LT and PLT tests, improving the GB voltage accuracy for the PLT test. During the mass production stage, one can eliminate the LT tests using the most appropriate GB voltage for the PLT test. In actuality, some discrepancies might arise between the SPICE model of transistor and the silicon data. Therefore, the most appropriate GB voltage is updated by the large amount of evaluation data in mass production.
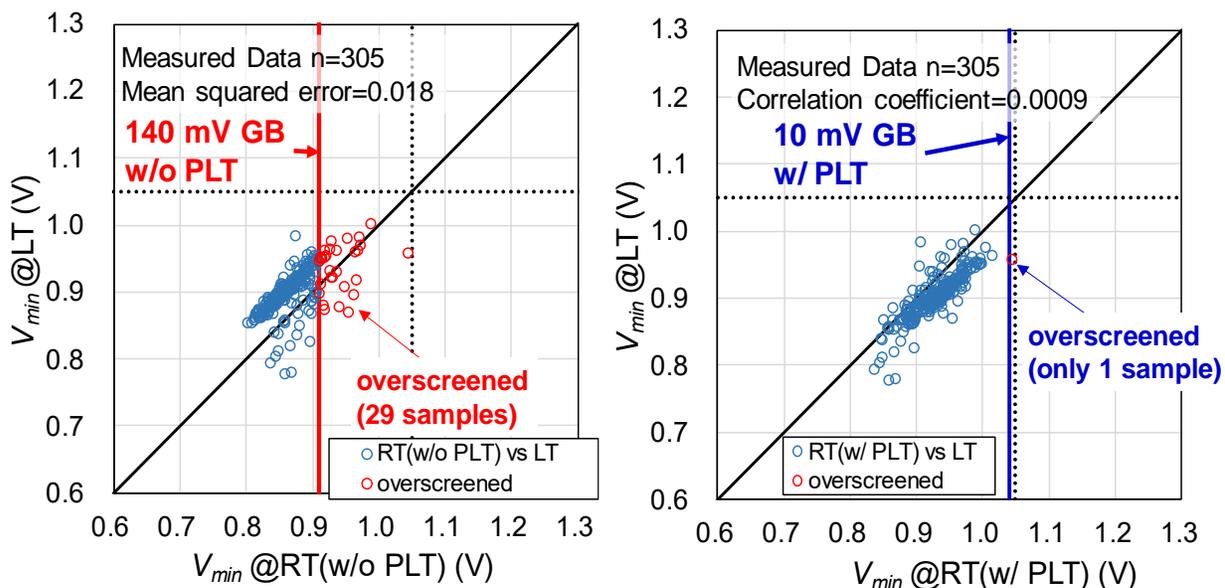
(a) 4-Mbit 6T SP-HD SRAM macro
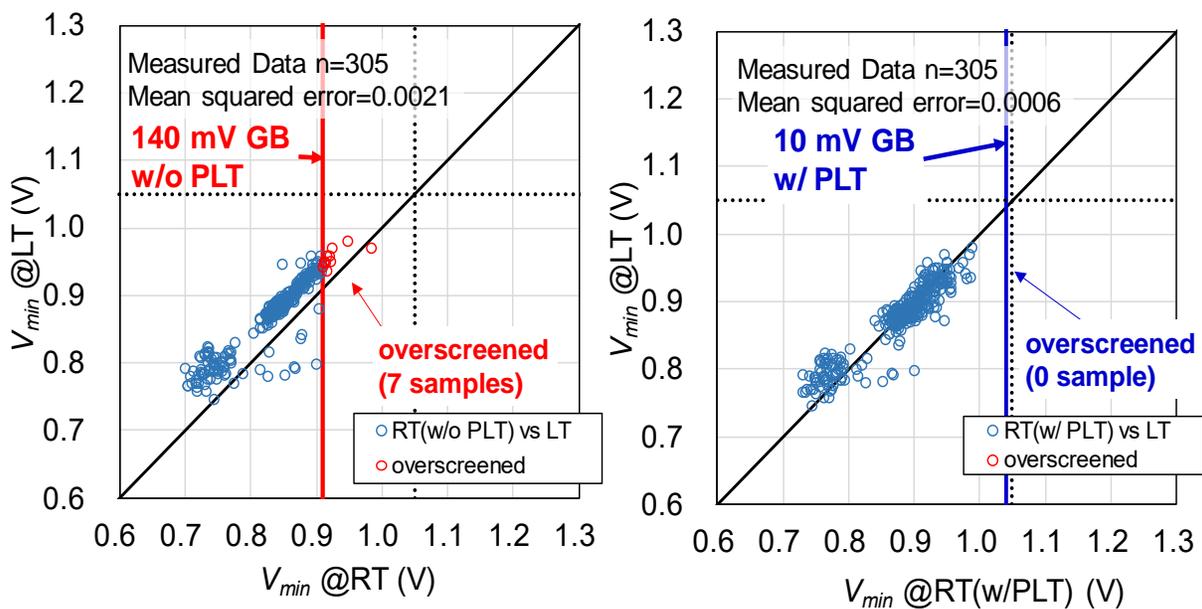
(b) 4-Mbit 6T SP-HC SRAM macro
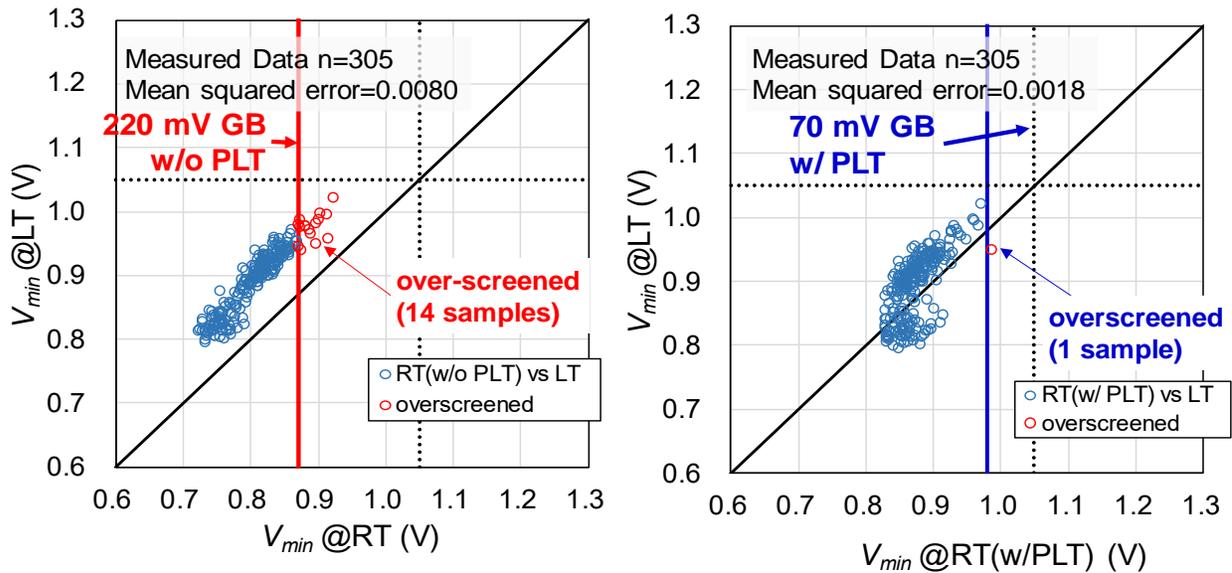
(c) 1-Mbit 8T DP SRAM macro

Fig. 96. Measured V_{min} distributions of SRAM macros with and without PLT test mode. (a) 4-Mbit 6T SP-HD macro, (b) 4-Mbit 6T SP-HC macro, and (c) 1-Mbit 8T DP macro.

Table 9 SUMMARIES OF SCREENED DIES

Bitcell	Test method	Number of dies (fail/test)			Test time	Rate of overscreened
		(1)HT test	(2)LT test	(3)RT test		
SP-HD	(conv.) 3-temp.	27/305	0/278	0/278	100%	0%
	(conv.) 2-temp. Voltage GB	27/305		29/278	66%	10.4%
	(prop.) PLT	27/305		1/278	66%	0.4%
SP-HC	(conv.) 3-temp.	6/305	0/299	0/299	100%	0%
	(conv.) 2-temp. Voltage GB	6/305		7/299	66%	2.3%
	(prop.) PLT	6/305		0/299	66%	0.0%
DP	(conv.) 3-temp.	3/305	0/302	0/302	100%	0%
	(conv.) 2-temp. Voltage GB	3/305		14/302	66%	4.6%
	(prop.) PLT	3/305		1/302	66%	0.3%

5.5. Conclusion

The pseudo-low-temperature test (PLT) method for 40 nm single-port SRAM and dual-port SRAMs were proposed to reduce testing times by eliminating low-temperature condition tests. The

proposed test circuit reproduced low-temperature conditions at RT with good correlation. SRAM macros with bitcells of three types on 40 nm low-power CMOS technology were designed and fabricated, and it was confirmed that the proposed test circuitry screens out low-temperature failures at RT. The proposed technique can reduce test costs by around 1/3 compared to the conventional three temperature test. Moreover, it can reduce the number of overscreened dies compared to the conventional voltage guard-band test method (e.g. the proposed PLT test method reduces the number of overscreened dies for 6T SP-HD bitcell SRAM from 29 to 1). The proposed testing method can be applied to other low-power CMOS platforms, with similar expected effects.

Chapter 6   Advanced Process SRAM

6.1. Introduction

Scaling in semiconductor process technologies enables a small SoC chip including SRAM with large capacity, which accelerates mobile-device market as well as artificial intelligence (AI), autonomous driving. Meanwhile, it is frequently pointed out also in area that the cost for 7nm and below jumps up enormously [73], thus every effort to reduce the cost in terms of design technology became more important than ever. SRAM which generally occupies 50% or more inside a single chip encounters a difficulty due to a higher metal-wire resistivity; Longer bitline (BL) and wordline (WL) prevent the SRAM macro sizes from being shrunk. To overcome this issue, various schemes such as using double WL [74], dual write-drive assist [75] and write/read-assist circuits [74], [75] have been proposed, but the maximum bit-cell array size does not exceed 256 row x 256 column class. In this section, 512 row x 512 col SRAM without dividing the macro by using dual-drive circuits for both BL and WL, which achieves the bit density of 29.2 Mb/mm^2 is introduced. Fig. 97 plots an estimated wordline (WL) delay in 7nm technology. Even if the fin number of the WL driver is added to increase its drivability, higher WL resistance prevents the WL delay from being improved. On the other hand, Fig. 97 also shows how the SRAM write margin [27] behaves against the bitline (BL) length at a high/low temperature. In the figure, write margin less than 0 means the write failure. The worst condition for the write margin generally appears at a low temperature. However, it is noticeable that the write margin gets worse in accordance with the BL length, and the degradation rate is quite severe at a high temperature rather than at a low temperature.
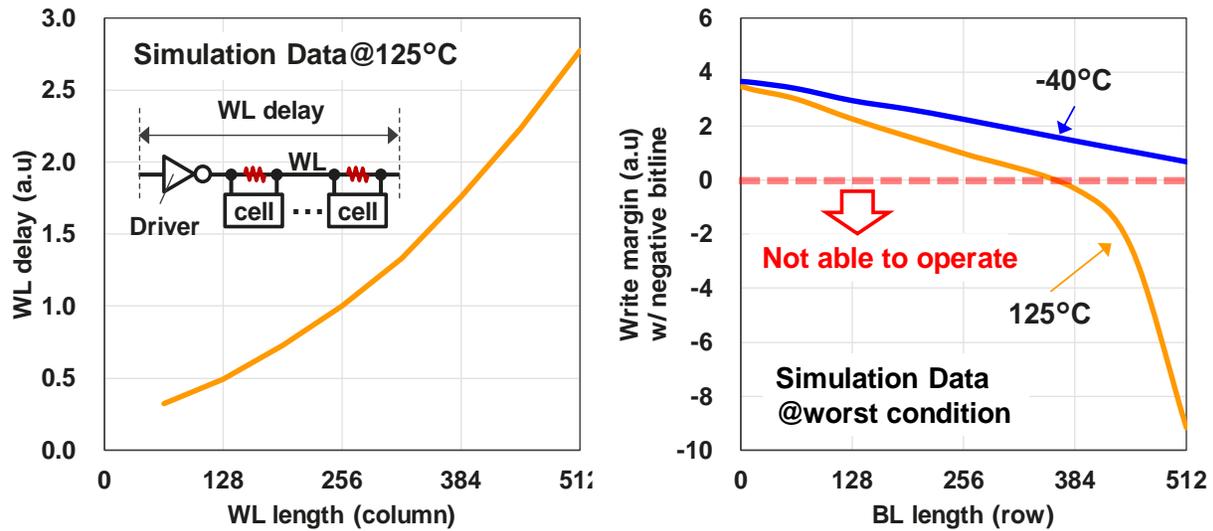
Fig. 97 Wordline (WL) delay vs. WL length and write margin dependence on bitline (BL) length in 7nm Fin-FET process.

6.2. Wordline Dual Drive and Bitline Dual Drive

Fig. 98 is the block diagram of the SRAM macro developed in this work. SRAM bit-cell array with 512 row x 512 col uses the 6T SRAM with the highest density that fab offers. This SRAM macro has new blocks: WL dual drive (WLDD) block, BL dual drive (BLDD) block and column replica. Note that these WLDD and BLDD circuits are allocated at the far-edge of the WL and BL, respectively, which enables the high-speed operation for such long wirings with 512 row/column, combined with the read/write assist circuits.

Fig. 99 illustrates the circuit diagram of WLDD. There are 2 drivers for a single row: WL drv. close to the Dec. and sub-WL drv. far from the Dec. The signal WLBF which drives sub-WL drv. is wired at the upper metal layer with much smaller parasitic capacitance than WL. Thus, the WLBF delay becomes very small, and WL is driven at both edges immediately. Read-assist by lowering WL voltage is applied to ensure static noise margin (SNM). WL voltage is pulled down by the PMOS directly attached to each WL, and the sub-WL drv.'s power is supplied by the level generator which is adjusted to be equal to WL voltage level. Fig. 99 compares the simulated waveforms with and without WLDD at the worst condition for read operation. The rising time of WL is 240 ps improved by WLDD at the slowest point compared to the conventional one. The area overhead of WLDD is about 1 %.

Fig. 100 illustrates the circuit diagram of proposed BLDD. BL is driven by 2 drivers: a standard

write amplifier (write amp.) and a "far-amplifier (far amp.)" attached a far end of BL. The sources of the both amps are connected to LWVSS, which supplies negative voltage for write assist. The coupling capacitance for obtaining negative BL is placed over the bit-cell array by using upper metal layer. It enables to generate a stable voltage level independent on the BL length and suitable for memory compiler design. Furthermore, it has no area overhead. When BT is discharged by the write amp. and comes to a certain voltage, the BL is also discharged at the far edge from a write amp. Thus, the far amp. is triggered which support BT/BB to discharge even though the BL suffers from its own higher wire resistance. The waveform in Fig. 100 shows simulation result at the worst condition. BL without BLDD cannot reach VSS level. Contrary to this, BL with BLDD is pulled-down well to the negative voltage, and MT/MB (see Fig. 98) flipping is successfully done. The area overhead of BLDD is less than 1 %.

Higher wire resistance causes timing delay not only to WL/BL but also to the internal signals to realize read/write operations. Fig. 101 shows the block diagram of the WL decoder with replica circuits to compensate timing delay. To reduce the rise delay from CLK to WL, the signal XA has repeaters allocated at well-tap row, enabling no area penalty. The path to activate the sense-amplifier (SA) is also shown in Fig. 101 using replica circuits [76]. Our self-timing replica circuit consists of row and column segments; in the row replica, logic replica cells which give RC delay to RBL1/2 path are introduced, whereas in the column replica, CWL1/2 path which matches with WL length (or the number of columns) is wired only by using metal wiring, which generates the timing delay that reflects the WL resistance.
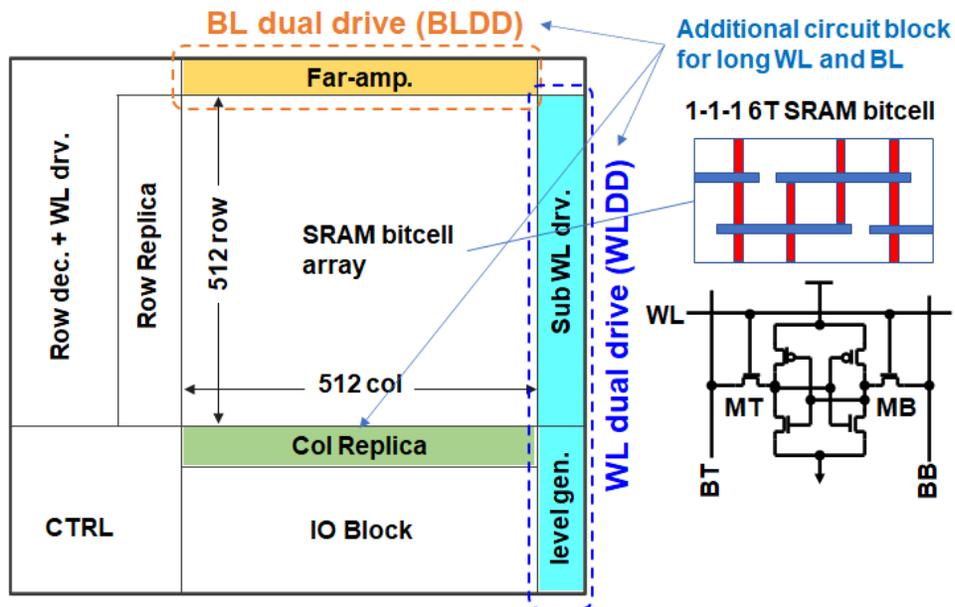
Fig. 98 Block diagram of proposed SRAM (512 x 512 array).
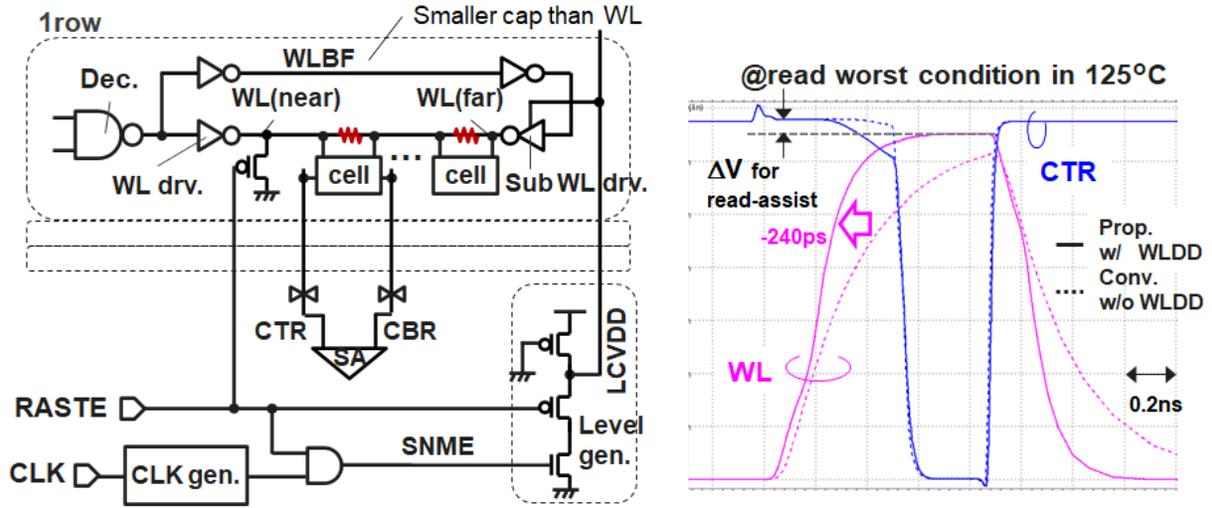
Fig. 99 Circuit diagram and SPICE simulation result (WLDD).
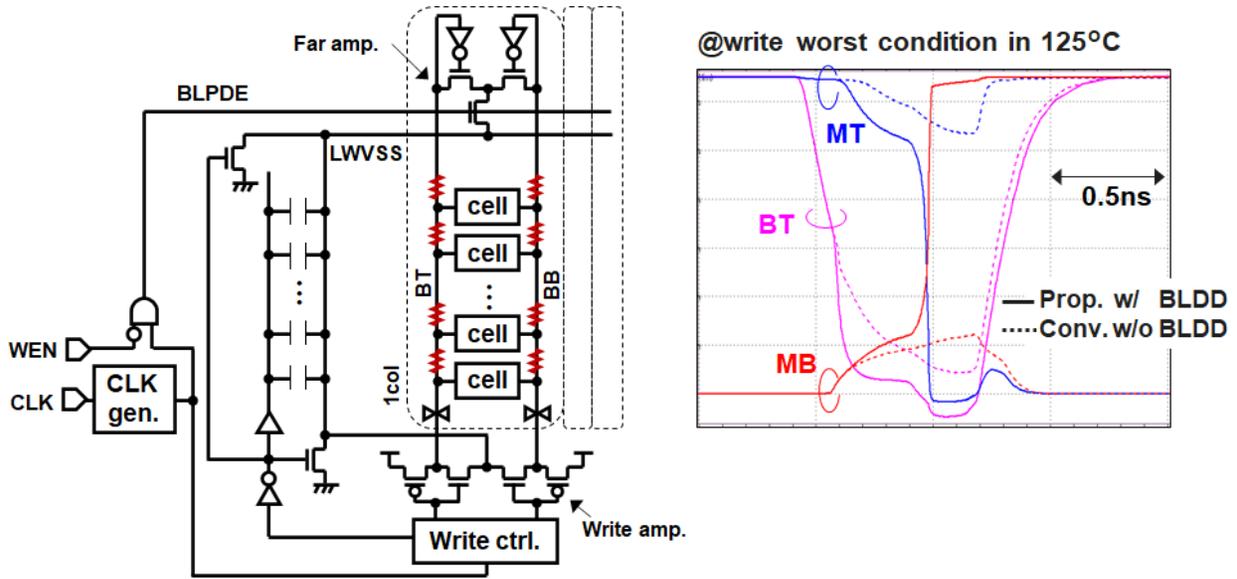
Fig. 100 Circuit diagram of the proposed BL dual drive (BLDD) with negative BL and SPICE simulation result.
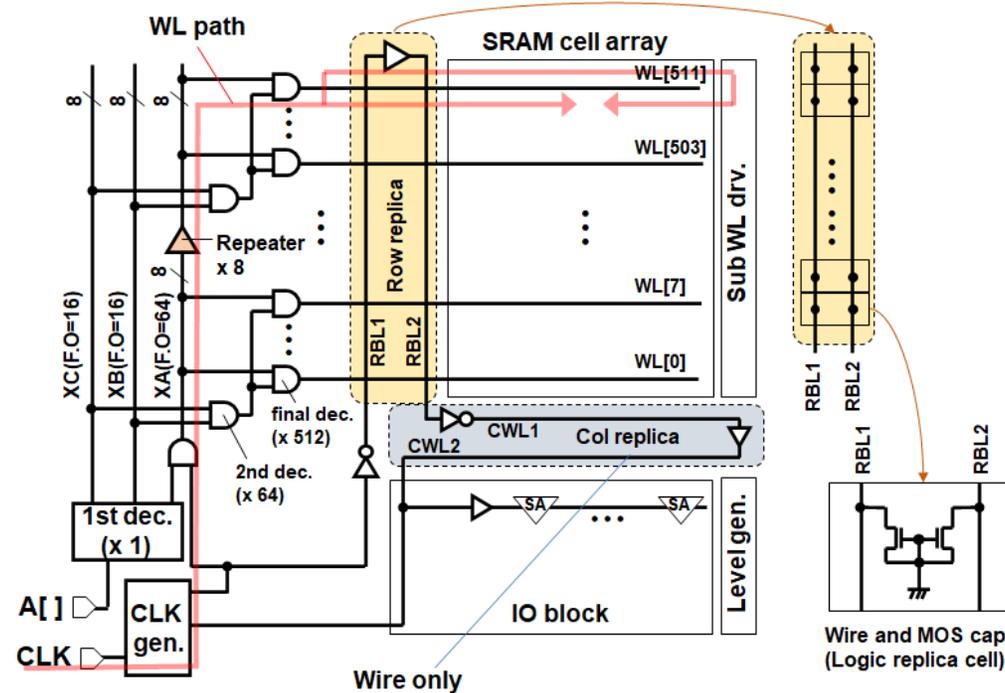
Fig. 101 Circuit path of WL driver and timing replica.

6.3. Measurement Results

Fig. 102 shows the evaluation results of a fabricated test chip. The left figure of Fig. 102 is a Shmoo plot of 256K byte. The access time of a conventional SRAM macro is adjusted to gain same differential voltage as using WLDD. The access time of the proposed SRAM macro is 0.82 ns that is improved 109 ps by WLDD in this condition. The minimum operation voltage (V_{min}) is improved by 120 mV compared to conventional macro. The right graph in Fig. 102 plots the relationship between operating voltage and the failure probability calculated from fail bit count of 8 Mbit. The proposed SRAM macro has better V_{min} that is about 0.65 V higher than the conventional SRAM macro without WLDD and BLDD. Fig. 103 shows the die microphotograph of the test chip and Table **10** shows features of the proposed SRAM macro.
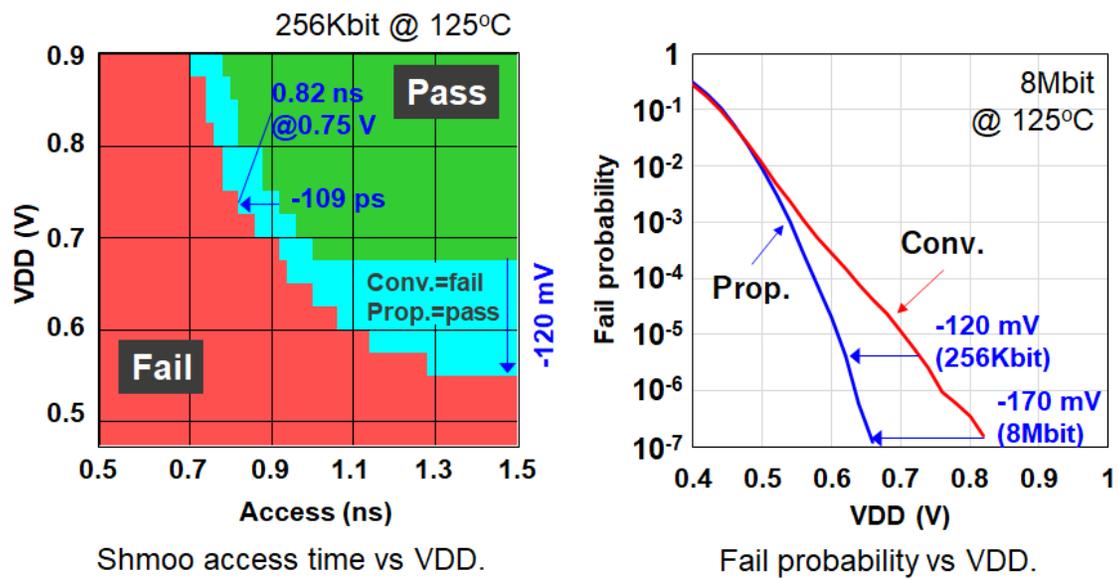
Fig. 102 Evaluation result of test chip.
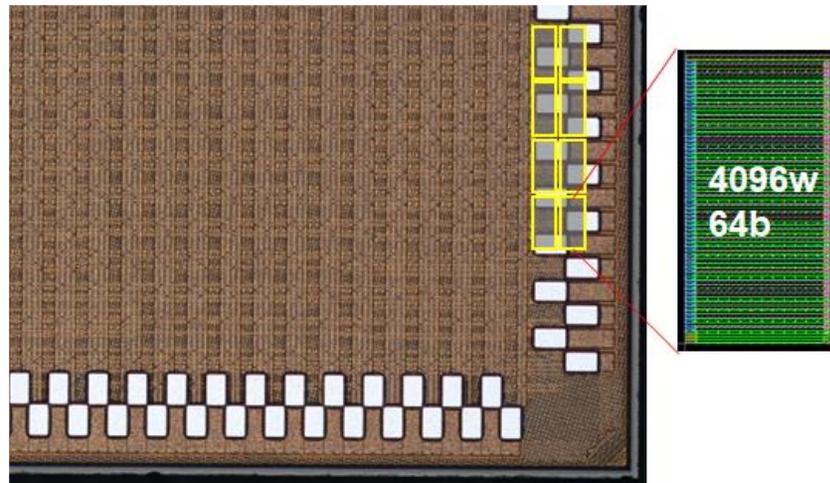
Fig. 103 Photograph of test chip.

Table 10 Features of the test chip

	Features
Technology	7-nm CMOS process
Macro configuration	4096w x 64b (512row x 512col)
Macro size	65.34 µm x 131.031 µm
Bit density	29.2 Mbit/mm^2
Access time	0.82 ns
Dynamic power	Read: 5.8 µW/MHz Write: 9.1 µW/MHz
Standby power in normal standby	16.9 µW @25°C

6.4. Conclusion

The dual drive techniques for WLs and BLs were proposed. The proposed techniques can drive long WLs and BLs in large bitcell array, and it is helpful for high density. It was confirmed that the proposed techniques can improve the V_{min} by 120 mV and the access time by 12 % using the test chips fabricated in 7-nm CMOS technology. The designed SRAM macro achieved the 29.2 Mb/mm^2 of memory density.

Chapter 7   Conclusion

This thesis described design optimization techniques for embedded SRAMs on advanced technologies. These study findings indicate measures to address the following four main points.

1. Low power techniques of SRAMs (Chapter 3)
2. Optimization of multi-port SRAMs (Chapter 4)
3. Test cost reduction techniques (Chapter 5)
4. High-density SRAM on an advanced process technology (Chapter 6)

These circuit techniques contribute to the resolution of SRAM critical issues in the advanced technologies.

Chapter 1 is a preface of the background and objective of this study. The structure of this thesis was explained. Before presenting practical design techniques to improve the four points above, undesirable problems for developing SRAMs are introduced in Chapter 2. The basic functions of the SRAM bitcell were explained. It was described that the bitcell requires two operation margins of the WM and SNM. Then, as basic knowledge for low-power techniques, the leakage current components of the MOS transistor and the reasons why the standby power increases as process scaling were explained. The sub-threshold leakage is increased by lowering the threshold voltage of the transistors, GIDL, and gate leakage are increased by strengthening the electric field with gate oxide scaling. The chapter additionally described that the active power increase is slowed by lowering the supply voltage, whereas the standby power has been increasing rapidly with process scaling. Reducing the standby power of LSI is extremely difficult. It was also explained that the screening test cost has been increasing along with the number of transistors. The screening tests of the MCUs are bipolarized. The tests for the automotive MCUs must screen out the failures strictly, whereas tests for low-cost consumer MCUs must reduce the test cost. It was also described that the SRAM performance is worsened by wire resistance, which increases rapidly in the advanced technologies.

In Chapter 3, the resume standby technique was explained. It can reduce the standby power of SRAMs while retaining memory data. Three examples for applying the resume standby technique were demonstrated. The standby power capacities of three designed SRAMs were reduced respectively by 78%, 74%, and 98% using the resume standby technique. Furthermore, tests to screen out retention failures, which are side effects of the resume standby techniques, were proposed. To ensure that the retention failures are screened out, the test to pulled down BL when the WL are low was proposed for automotive MCUs. Another screening test was proposed for low-cost consumer MCUs. The test can reduce the retention test time by reproducing the VDD source voltages of bitcell in the resume standby mode. It was also confirmed that the active power of 3.3 V thick gate SRAMs was reduced 60% by column selecting with WLs.

Chapter 4 explained the 8T bitcells for the DP and 2P SRAMs. 8T bitcells have the special issue

designated as a "disturbance issue", which occurs when two ports select the same row address simultaneously. The issue reduces the SRAM read and write margins. The DP and 2P SRAM performance was worsened to ensure the operation margins considering the disturbance issue. The examples presented in Chapter 4 improve the performance by decreasing effects of the disturbance issue. The first one proposed that the 2P SRAM, which has a read-only port, can be sped up because the period of the disturbance issue is shorter when the aggressor port executes a read operation. The other proposed that the active power can be reduced by detecting the same row access. The proposed SRAM widens the WL pulse width only during the same row access. The active power was reduced by 18%.

In Chapter 5, a method to reduce test costs was proposed. The screening test reproduces the low-temperature write failures at room temperature. The LT test is terminated. It is expected that the test cost becomes 2/3. It was also confirmed that the number of overscreened samples are reduced by 1/29 by the proposed technique, compared with the conventional voltage guard band technique.

In Chapter 6, high-speed and high-density techniques for advanced processes such as 7 nm were proposed. The RC delay is reduced and the WM is improved by the dual drive techniques, which drives WLs and BLs at both edge of WLs and BLs. The proposed techniques can drive long WLs and BLs quickly. Therefore, it avoids subdivision of the bitcell array. Furthermore, the proposed techniques have good correlation with the conventional read / write assist circuits. Using the proposed techniques, the V_{min} was improved by 120 mV. Moreover, the access time was improved by 12%. Memory density of 29.2 Mb/mm^2 was achieved.

References

[1] G.E. Moore, "No exponential is forever: but "Forever" can be delayed!," *2003 IEEE International Solid-State Circuits Conference*, 2003. Digest of Technical Papers.

[2] Y. Yano, "Take the Expressway to go Greener," *ISSCC Dig. Tech.* Papers, pp. 24-30, 2012.

[3] T. Kono, T. Ito, T. Tsuruda, T. Nishiyama, T. Nagasawa, T. Ogawa, Y. Kawashima, H. Hidaka, T. Yamauchi, "40nm embedded SG-MONOS flash macros for automotive with 160MHz random access for code and endurance over 10M cycles for data," *ISSCC* Dig. Tech. Papers, pp. 212-213, 2013.

[4] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa, K. Osada, "A 300-MHz 25-µA/Mb-leakage on-chip SRAM module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor," *IEEE Journal of Solid-State Circuits*, pp. 186-194, vol. 40, 2005.

[5] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, M. Bohr, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 895- 901, 2005.

[6] C. Wann, "SRAM cell design for stability methodology," in *Proc. IEEE VLSI-TSA,* pp. 21-22, 2005.

[7] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y. Ng, L. Wei, Y. Zhang, K. Zhang, M. Bohr, "A 4.0 GHz 291Mb voltage-scalable SRAM design in 32nm high-κ metal-gate CMOS with integrated power management," *ISSCC* Dig. Tech. Papers, pp. 456-457,457a, 2009.

[8] J. Piyush, J. Dasani, A. Kumar, "Sleep circuit for SRAM core with improved noise-margin," *ICICDT* Dig. Tech. Papers, pp. 139-142, 2008.

[9] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, S. Iwade, "A 90-nm low-power 32-kB embedded SRAM with gate leakage suppression circuit for mobile applications," *IEEE Journal of Solid-State Circuits*, pp. 684-693, vol. 39, 2004.

[10] Y. Takeyama, H. Otake, O. Hirabayashi, K. Kushida, N. Otsuka, "A low leakage SRAM macro with replica cell biasing scheme," *IEEE Journal of Solid-State Circuit*s, pp. 815-822, vol. 41, 2006.

[11] W. Yih, Hong Jo Ahn, U. Bhattacharya, C. Zhanping, T. Coan, F. Hamzaoglu, W. M. Hafez, J. Chia-Hong, P. Kolar, S.H. Kulkarni, L. Jie-Feng, N. Yong-Gee, I. Post, W. Liqiong, Z. Ying, K. Zhang, M. Bohr, "A 1.1 GHz 12 µA/Mb-Leakage SRAM Design in 65 nm Ultra-Low-Power CMOS Technology With Integrated Leakage Reduction for Mobile Applications,"

IEEE Journal of Solid-State Circuits, vol. 43, no. 1, pp. 172-179, Jan. 2008.

[12] K. Nii, M. Yabuuchi, H. Fujiwara, Y. Tsukamoto, Y. Ishii, T. Matsumura, Y. Matsuda, "A cost-effective 45nm 6T-SRAM reducing 50mV Vmin and 53% standby leakage with multi-Vt asymmetric halo MOS and write assist circuitry," in *2013 Symp. ISQED* Dig. Tech. Papers, 2013, pp. 438–441.

[13] N. Maeda, S. Komatsu, M. Morimoto, K. Tanaka, Y. Tsukamoto, K. Nii, Y. Shimazaki, "A 0.41 µA Standby Leakage 32 kb Embedded SRAM with Low-Voltage Resume-Standby Utilizing All Digital Current Comparator in 28 nm HKMG CMOS," *IEEE Journal of Solid-State Circuits, vol. 48, no. 4, pp. 917-923, Apr. 2013.*

[14] H. Fujiwara, M. Yabuuchi, M. Morimoto, K. Tanaka, M. Tanaka, N. Maeda, Y. Tsukamoto, K. Nii, "A 20nm 0.6V 2.1µW/MHz 128kb SRAM with No Half Select Issue by Interleave Wordline and Hierarchical Bitline Scheme," *Symp. VLSI Circuits* Dig. Tech. Papers, 2013, pp. C118-C119.

[15] ITRS2001, "https://www.semiconductors.org/".

[16] Q. Xu et. al, "Test scheduling for built-in self-tested embedded SRAMs with data retention faults," *IET Journals & Magazines*, vol. 1, no. 3, pp.256-264, 2007.

[17] Y. Yokoyama et. al, "A cost effective test screening method on 40-nm 4-Mb embedded SRAM for low-power MCU," *A-SSCC*, pp1-4, 2015

[18] Y. Ishii et.al, "A 28 nm Dual-Port SRAM Macro With Screening Circuitry Against Write-Read Disturb Failure Issues," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp.2535-2544

[19] Toshikazu Fukuda, K. Kohara, T. Dozaka, Y. Takeyama, T. Midorikawa, K. Hashimoto et al., "13.4 A 7ns-access-time 25µW/MHz 128kb SRAM for low-power fast wake-up MCU in 65nm CMOS with 27fA/b retention current," in *Proc. ISSCC*, pp. 236-237, Feb. 2014.

[20] T. Kamei et. al, "A resume-standby application processor for 3G cellular phones," *ISSCC*, pp.336-531, Vol.1, 2004

[21] N. Maeda et. al "A 0.41 µA Standby Leakage 32 kb Embedded SRAM with Low-Voltage Resume-Standby Utilizing All Digital Current Comparator in 28 nm HKMG CMOS," *IEEE Journal of Solid-State Circuits*, pp.917-923, Vol.4, 2013

[22] Y. Yokoyama et. al, "40nm Ultra-low leakage SRAM at 170 deg.C operation for embedded flash MCU," *ISQED*, pp.24-31, 2014.

[23] Y. Yokoyama et. al, "40-nm 64-kbit Buffer/Backup SRAM with 330 nW Standby Power at 65°C Using 3.3 V IO MOSs for PMIC less MCU in IoT Applications," *A-SSCC*, pp.9-12, 2018.

[24] K. Osada et. al, "16.7-fA/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray-induced multierrors," *IEEE Journal of Solid-State Circuits*, pp.1952-1957, Vol. 38, no.11,

2003.

[25] R. Islam et. al, "Low power SRAM techniques for handheld products," *ISLPED*, pp.198-202, 2005.

[26] E. Seevinck, F. J. List, and J. Lohstroh, "Static-Noise Margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits,* vol. 22, no. 5, pp. 748-754, Oct. 1987.

[27] R. Heald and P. Wang, "Variability in sub-100 nm SRAM designs," *ICCAD* Digest, pp. 347-352, 2004.

[28] Makoto Yabuuchi et. al, "A 65 nm 1.0 V 1.84 ns Silicon-on-Thin-Box (SOTB) embedded SRAM with 13.72 nW/Mbit standby power for smart IoT," *2017 Symposium on VLSI Circuits,* pp.220 - 221.

[29] Qianneng Zhou et. al, "Embedded DC-DC Voltage Down Converter for Low-Power VLSI Chip," *APCCAS*, pp.670 – 673, 2006.

[30] Patent US7087942, 2006

[31] Yoshisato Yokoyama et. al, "40nm Ultra-low leakage SRAM at 170 deg.C operation for embedded flash MCU," *ISQED*, pp.24-31, 2014.

[32] Yoshisato Yokoyama, et. al, "A cost effective test screening method on 40-nm 4-Mb embedded SRAM for low-power MCU," in *Proc. A-SSCC*, Industry2-3, Nov. 2015

[33] Koji Nii, et. al, "A dynamic/static SRAM power management scheme for DVFS and AVS in advanced automotive infotainment SoCs," *2016 Symposium on VLSI Circuits, pp.220 – 221*

[34] Takahiko Ishizu, et. al, "A 140 MHz 1 Mbit 2T1C gain-cell memory with 60-nm indium-gallium-zinc oxide transistor embedded into 65-nm CMOS logic process technology," *2017 Symposium on VLSI Circuits*, pp.C162 – C163

[35] Y. Yokoyam, Y. Ishii, H. Kojima, A. Miyanishi, Y. Tsujihashi, S. Asayama et al., "40nm Ultra-low leakage SRAM at 170 deg.C operation for embedded flash MCU," in *Proc. ISQED*, pp. 24-31, Mar. 2014.

[36] Int. Standardization Organisation, ISO 11898. "Road vehicles – Interchange of Digital Information - Controller Area Network (CAN) for High-Speed Communication," 1993.

[37] "The flexray communication system specification, version 3.0.1," http://www.flexray.com.

[38] L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. K. Montoye, et al., "Stable SRAM Cell Design for the 32 nm Node and Beyond," in *Symp. VLSI Circuits* Dig. Tech. Papers, June 2005.

[39] K. Nii, Y. Tsukamoto, S. Imaoka and H. Makino, "A 90 nm dual-port SRAM with 2.04 um^2 8T-thin cell using dynamically-controlled column bias scheme," in *ISSCC* Dig. Tech. Papers, pp. 508-509, 543, Feb. 2004.

[40] Y. Ishii, H. Fujiwara, S. Tanaka, Y. Tsukamoto, K. Nii, Y. Kihara and K. Yanagisawa "A 28

nm dual-port SRAM macro with screening circuitry against write-read disturb failure issues," *JSSC*, pp. 2535-2544, vol.46, issue 11, Nov. 2011.

[41] T. Shiota et al., "A 51.2 GOPS 1.0 GB/s-DMA single-chip multi-processor integrating quadruple 8-way VLIW processors," *ISSCC* Digest, pp. 194-195, 593, Feb. 2005.

[42] M. Nakajima, T. Yamamoto, M.Yamasaki, K. Kaneko, T. Hosoki, "Homogeneous Dual-Processor core with Shared L1 Cache for Mobile Multimedia SoC," in *VLSI Circuits*, Digest of Technical Papers, pp. 216-217, June 2007.

[43] Igor Loi, Luca Benini, "A multi banked - Multi ported - Non blocking shared L2 cache for MPSoC platforms," *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp.1-6, Mar. 2014.

[44] https://www.xilinx.com/support/documentation/data_sheets/ds099.pdf

[45] T.Toi, N.Nakamura, T.Fujii, T.Kitaoka, K.Togawa, K.Furuta and T.Awashima, "Optimizing Time and Space Multiplexed Computation in a Dynamically Reconfigurable Processor," *ICFPT 2013*, pp.106-111, Dec. 2013.

[46] T.Toi, N.Nakamura, Y.Kato, T.Awashima, K.Wakabayashi, "High-level Synthesis Challenges for Mapping a Complete Program on a Dynamically Reconfigurable Processor," *IPSJ Trans. on System LSI Design Methodology*, Vol. 3, pp.91-104 , Feb. 2010.

[47] Patent US6529401, 2003.

[48] K. Nii, Y. Masuda, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, M. Igarashi, K. Tomita, N. Tsuboi, H. Makino, K. Ishibashi, and H. Shinohara, "A 65 nm Ultra-High-Density Dual-Port SRAM with $0.71\mu m^2$ 8T-Cell for SoC," *VLSI Cir. Symp. Dig.*, pp. 130-131, 2006.

[49] K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, Y. Oda, K. Usui, T. Kawamura, N. Tsuboi, T. Iwasaki, K. Hashimoto, H. Makino, and H. Shinohara, "A 45-nm Single-port and Dual-port SRAM family with Robust Read/Write Stabilizing Circuitry under DVFS Environment," *VLSI Cir. Symp*. Dig., pp. 212-213, 2008.

[50] S.Tanaka, Y.Ishii, M.Yabuuchi, T.Sano, K.Tanaka, Y.Tsukamoto, K.Nii and H.Sato, "A 512-kb 1-GHz 28-nm Partially Write-Assisted Dual-Port SRAM with Self-Adjustable Negative Bias Bitline," *VLSI Cir. Symp. Dig., pp. 113-114,* 2014.

[51] Y.Yokoyama, Y.Ishii, K.Tanaka, T.Fukuda, Y.Tsujihashi, A.Miyanishi, S.Asayama, K.Maekawa, K.Shiba and K.Nii, "40 nm Dual-port and two-port SRAMs for automotive MCU applications under the wide temperature range of -40 to 170°C with test screening against write disturb issues," in *Proc. A-SSCC*, pp25-28, Nov. 2014.

[52] Yen-Huei Chen, Kao-Cheng Lin, Ching-Wei Wu, Wei-Min Chan, Jhon-Jhy Liaw, Hung-Jen Liao and Jonathan Chang, "A 16nm Dual-Port SRAM with Partial Suppressed Word-line, Dummy Read Recovery and Negative Bit-line Circuitries for Low VMIN Applications," *VLSI Cir. Symp. Dig.*, 2016.

[53] Y.Ishii, M.Yabuuchi, Y.Sawada, M.Morimoto, Y.Tsukamoto, Y.Yoshida, K.Shibata, T.Sano, S.Tanaka and K.Nii, "A 5.92-Mb/mm2 28-nm Pseudo 2-Read/Write Dualport SRAM using Double Pumping Circuitry," in *Proc. A-SSCC*, pp. 17-20, Nov. 2016.

[54] M. Fujigaya, N. Sakamoto, T. Koike, T. Irita et al., "A 28 nm High-κ Metal-Gate Single-Chip Communications Processor with 1.5 GHz Dual-Core Application Processor and LTE/HSPA+-Capable Baseband Processor," in *IEEE ISSCC* Dig. Tech. Papers, 2013, pp. 156–157.

[55] P. K. Veenstra, F. P. M. Beenker and J. J. M. Koomen, "Testing of random access memories: theory and practice," *IEEE Proc. G - Electronic Circuits and Systems*, vol. 135, no. 1, pp. 24–28, 1988.

[56] H. Deshayes, "Cost of test reduction," in *Proc. Int. Test Conf.*, 1998, pp. 265–271.

[57] K. Nii, Yabuuchi, Y. Tsukamoto et al., "A 45-nm Single-port and Dual-port SRAM family with Robust Read/Write Stabilizing Circuitry under DVFS Environment," *VLSI Cir. Symp. Dig.,* 2008, pp. 212–213.

[58] K. Nii et al., "A 65 nm Ultra-High-Density Dual-Port SRAM with $0.71\mu m^2$ 8T-Cell for SoC," *VLSI Cir. Symp. Dig.,* 2006, pp. 130–131.

[59] B. Wang, J. Yang, and A. Ivanov, "Reducing test time of embedded SRAMs," in *Proc. IEEE Int. Workshop on Memory Technology*, Design and Testing (MTDT), San Jose, July 2003, pp. 47–52.

[60] D. C. Huang and W. B. Jone, "A parallel transparent BIST method for embedded memory arrays by tolerating redundant operations," *IEEE Tran. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 5, pp. 617–628, 2002.

[61] S. M. Sze, "Physics of Semiconductor Devices," Wiley, New York, 1969.

[62] C. Tseng et al., "Cold Delay Defect Screening," in *Proc. IEEE VLSI Test Symposium*, Apr. 2000, pp. 183–188.

[63] Y. Yokoyama, Y. Ishii, H. Kojima, A. Miyanishi, Y. Tsujihashi, S. Asayama et al., "40 nm Ultra-low leakage SRAM at 170 deg.C operation for embedded flash MCU," in *Proc. ISQED*, Mar. 2014, pp. 24–31.

[64] K. Nii, Y. Tsukamoto, Y. Ishii, M. Yabuuchi, H. Fujiwara, and K. Okamoto,"A Test Screening Method for 28 nm HK/MG Single-Port and Dual-Port SRAMs Considering with Dynamic Stability and Read/Write Disturb Issues, " in *Test Symposium (ATS)*, Nov. 2012, pp. 246–251.

[65] Y. Yokoyama, Y. Ishii, T. Inada, K. Tanaka, M. Tanaka,Y. Tsujihashi et. al., "A cost-effective test screening method on 40-nm 4-Mb embedded SRAM for low-power MCU," in *Proc. A-SSCC*, Nov. 2015, pp. 1–4.

[66] H. Hao and E. J. McCluskey, "Very-low-voltage testing for weak CMOS logic ICs," in *Proc. Int. Test Conf.,* 1993, pp. 275–284.

[67] P. Engelke, I. Polian, M. Renovell, S. Kundu et al., "On Detection of Resistive Bridging Defects by Low-Temperature and Low-Voltage Testing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 2, pp. 327–338, 2008.

[68] P. Engelke, I. Polian, M. Renovell, B. Seshadri, and B. Becker, "The pros and cons of very-low-voltage testing: an analysis based on resistive bridging faults," *Proc. IEEE VLSI Test Symposium*, pp. 171–178, 2004.

[69] Y. Ishii, H. Fujiwara, S. Tanaka, Y. Tsukamoto, K. Nii et al., "A 28 nm dual-port SRAM macro with screening circuitry against write-read disturb failure issues," *IEEE J. Solid-State Circuits*, vol. 46, issue 11, Nov. pp. 2535–2544, 2011.

[70] Y. Yokoyama, Y. Ishii, K. Tanaka, T. Fukuda et. al., "40 nm Dual-port and two-port SRAMs for automotive MCU applications under the wide temperatures of -40 to 170°C with test screening against write disturb issues," in *Proc. A-SSCC*, Nov. 2014, pp. 25–28.

[71] Jonathan Chang, Yen-Huei Chen, Gary Chan, Hank Cheng, Po-Sheng Wang, Yangsyu Lin, Hidehiro Fujiwara, Robin Lee, Hung-Jen Liao, Ping-Wei Wang, Geoffrey Yeap and Quincy Li, "15.1 A 5nm 135Mb SRAM in EUV and High-Mobility-Channel FinFET Technology with Metal Coupling and Charge-Sharing Write-Assist Circuitry Schemes for High-Density and Low-VMIN Applications," *IEEE International Solid- State Circuits Conference* 2020.

[72] IRDS2015, 2017, https://irds.ieee.org/editions

[73] Lisa T. Su, Samuel Naffziger, Mark Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," pp. 1.1.1 - 1.1.8, *IEEE International Electron Devices Meeting* 2017.

[74] Jonathan Chang, Yen-Huei Chen, Wei-Min Chan, Sahil Preet Singh, Hank Cheng, Hidehiro Fujiwara, Jih-Yu Lin, Kao-Cheng Lin, John Hung, Robin Lee, Hung-Jen Liao, Jhon-Jhy Liaw, Quincy Li, Chih-Yung Lin, Mu-Chi Chiang, Shien-Yang Wu, "12.1 A 7nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," *IEEE International Solid-State Circuits Conference*, pp.206-207, 2017.

[75] Taejoong Song, Jonghoon Jung, Woojin Rim, Hoonki Kim, Yongho Kim, Changnam Park, Jeongho Do, Sunghyun Park, Sungwee Cho, Hyuntaek Jung, Bongjae Kwon, Hyun-Su Choi, JaeSeung Choi, Jong Shik Yoon, "A 7nm FinFET SRAM using EUV lithography with dual write-driver-assist circuitry for low-voltage applications," *IEEE International Solid - State Circuits Conference*, pp.198-200, 2018.

[76] B.S. Amrutur and M.A. Horowitz, "A replica technique for wordline and sense control in low-power SRAM's," *IEEE J. Solid-State Circuits*, pp.1208-1219, no. 8, Vol.33, 1998

[77] Zheng Guo, Daeyeon Kim, Satyanand Nalam, Jami Wiedemer, Xiaofei Wang and Eric Karl, "A 23.6Mb/mm2 SRAM in 10nm FinFET technology with pulsed PMOS TVC and stepped-WL for low-voltage applications," *IEEE International Solid State Circuits Conference*,

pp.196-198, 2018.

Publications

Journal Papers

1) <u>Yoshisato Yokoyama</u>, Yuichiro Ishii, Koji Nii and Kazutoshi Kobayashi, "Cost-Effective Test Screening Method on 40-nm Embedded SRAMs for Low-Power MCUs", IEEE Transactions on Very Large Scale Integration (TVLSI) Systems, vol.29, no. 7, pp. 1495-1499, July 2021.

Conference Proceedings

1) <u>Yoshisato Yokoyama</u>, Yuichiro Ishii,Hidemitsu Kojima,Atsushi Miyanishi,Yoshiki Tsujihashi,Shinobu Asayama,Kazutoshi Shiba,Koji Tanaka,Tatsuya Fukuda,Koji Nii,Kazumasa Yanagisawa, "40nm Ultra-low leakage SRAM at 170 deg.C operation for embedded flash MCU", submitted to Fifteenth International Symposium on Quality Electronic Design (ISQED), 2014

2) <u>Yoshisato Yokoyama</u>,Yuichiro Ishii,Koji Tanaka,Tatsuya Fukuda,Yoshiki Tsujihashi,Astushi Miyanishi,Shinobu Asayama,Keiichi Maekawa,Kazutoshi Shiba,Koji Nii, "40 nm Dual-port and two-port SRAMs for automotive MCU applications under the wide temperature range of −40 to 170°C with test screening against write disturb issues", submitted to IEEE Asian Solid-State Circuits Conference (A-SSCC), 2014.

3) Koji Nii, Makoto Yabuuchi, <u>Yoshisato Yokoyama</u>,Yuichiro Ishii, Takeshi Okagaki, Masao Morimoto, Yasumasa Tsukamoto, Koji Tanaka, Miki Tanaka, Shinji Tanaka, "2RW dual-port SRAM design challenges in advanced technology nodes", submitted to IEEE International Electron Devices Meeting (IEDM), 2015

4) <u>Yoshisato Yokoyama</u>,Yuichiro Ishii,Toshihiro Inada,Koji Tanaka,Miki Tanaka,Yoshiki Tsujihashi,Koji Nii, "A cost effective test screening method on 40-nm 4-Mb embedded SRAM for low-power MCU", submitted to IEEE Asian Solid-State Circuits Conference (A-SSCC), 2015

5) <u>Yoshisato Yokoyama</u>,Yuichiro Ishii,Haruyuki Okuda,Koji Nii, "A dynamic power reduction in synchronous 2RW 8T dual-port SRAM by adjusting wordline pulse timing with same/different row access mode", submitted to IEEE Asian Solid-State Circuits Conference (A-SSCC), 2017

6) <u>Yoshisato Yokoyama</u>,Tomohiro Miura,Yukari Ouchi,Daisuke Nakamura,Jiro Ishikawa,Shunya Nagata,Makoto Yabuuchi,Yuichiro Ishii,Koji Nii, "40-nm 64-kbit Buffer/Backup SRAM with 330 nW Standby Power at 65°C Using 3.3 V IO MOSs for PMIC less MCU in IoT Applications", submitted to IEEE Asian Solid-State Circuits Conference (A-SSCC), 2018

7) <u>Yoshisato Yokoyama</u>,Kenji Goto,Tomohiro Miura,Yukari Ouchi,Daisuke Nakamura,Jiro Ishikawa,Shunya Nagata,Yoshiki Tsujihashi,Yuichiro Ishii, "A Cost Effective Test Screening Circuit for embedded SRAM with Resume Standby on 110-nm SoC/MCU", submitted to IEEE Asian Solid-State Circuits Conference (A-SSCC), 2019

8) <u>Yoshisato Yokoyama</u>,Miki Tanaka,Koji Tanaka,Masao Morimoto,Makoto Yabuuchi,Yuichiro Ishii,Shinji Tanaka, "A 29.2 Mb/mm2 Ultra High Density SRAM Macro using 7nm FinFET Technology with Dual-Edge Driven Wordline/Bitline and Write/Read-Assist Circuit", submitted to IEEE Symposium on VLSI Circuits, 2020

Patents

United States Patents

1. 10,964,404   Semiconductor device
2. 10,897,248   Semiconductor device
3. 10,830,814   Semiconductor device
4. 10,811,405   Semiconductor device
5. 10,552,261   Semiconductor device and memory module
6. 10,490,262   Semiconductor device
7. 10,424,575   Semiconductor device
8. 10,373,675   Semiconductor storage device
9. 10,109,337   Memory macro and semiconductor integrated circuit device
10. 9,830,980   Semiconductor device, test program, and test method
11. 9,728,272   Semiconductor device, test program, and test method
12. 9,711,208   Semiconductor storage device with reduced current in standby mode
13. 7,782,654   Static random access memory device
14. 7,352,650   External clock synchronization semiconductor memory device and method for controlling same
15. 7,286,390   Memory cell and semiconductor integrated circuit device

Acknowledgments