Cost-Effective Test Screening Method on 40-nm Embedded SRAMs for Low-power MCUs

Yoshisato Yokoyama, Yuichiro Ishii, Koji Nii, *Senior Member, IEEE*, and Kazutoshi Kobayashi *Senior Member, IEEE*

***Abstract*—Embedded SRAMs with cost-effective test screening circuitry are demonstrated for low-power microcontroller units (MCUs). The probing test step at the low temperature (LT) of -40°C is obviated by imitating pseudo-LT (PLT) conditions in the package test, where a sample is measured at room temperature (RT). Monte Carlo simulation is carried out considering local *Vt* variations as well as contact soft open failure (high resistance), confirming good minimum operating voltage (V_{min}) correlation between LT and PLT conditions. Test chips with two types of 4-Mbit single-port SRAM macros and 1-Mbit dual-port SRAM macro are designed and fabricated using low-power 40-nm CMOS technology. Measurement results demonstrate that the proposed test method reproduces LT conditions and screens out LT failures with less overscreening. The proposed test method eliminates 1/3 or more of the test costs.**

***Index Terms*—40 nm, MCU, Single-port SRAM, Dual-port SRAM, embedded memory, screening, testability, test cost, V_{min}**

I. Introduction

With process technology scaling, on-die embedded memory densities increase annually as do logic gates [1]. Consequently, the testing cost per transistor increases year by year [2]. Furthermore, screening tests have become more complicated, requiring much longer testing time of embedded memories [3]. In microcontroller unit (MCU) markets, reducing the necessary testing times of digital logic blocks, IO interfaces, analog blocks, and memory blocks is crucially important for production of cost-competitive products [4]. 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 [5],[6]. 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 [7],[8].

A low failure rate must be ensured after test screening and shipment. Device characteristics show temperature dependence [9] such that testing at low and high temperatures are indispensable to screen dies, which exhibit temperature-dependent failures [10]. Additionally, package testing after die sawing and assembly are generally performed at 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) [11]–[13]. 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. 1 presents testing flows with three temperature steps and two temperature steps as described herein. In this work, we propose cost-effective test screening with only two temperature conditions (HT and RT), skipping the LT test by introducing pseudo-LT (PLT) testing at RT [14].

The organization of this paper is the following. Section II gives a description of a conventional guard-band (GB) technique with tightening supply voltages for skipping the LT test flow; then its overscreening issue is discussed. The proposed PLT test is described in Section III. Section IV presents some results of our experimental evaluation of the circuit techniques. A brief conclusion is presented in Section V.

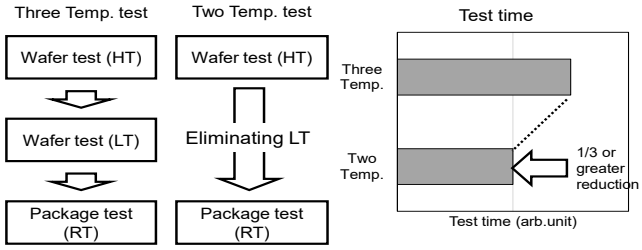


Fig. 1. Test screening flows with three temperature steps (typical) and two temperature steps (eliminating LT).

II. Overscreened Issues of Voltage Guard band Techniques

In testing flow with either three temperature steps or two-temperature steps as shown in Fig. 1, pin contact testing is

Manuscript received Jan. 10, 2021; revised March 23, 2021, and April 11, 2021; accepted May 02, 2021. Date of publication XX XX, 2021; date of current version April 11, 2021. (Corresponding author: Yoshisato Yokoyama.)

Yoshisato Yokoyama is with Renesas Electronics Corporation, Tokyo 187-8588, Japan, and also with Kyoto Institute of Technology, Kyoto 606-8585, Japan (e-mail: yokoesh@gmail.com).

Yuichiro Ishii and Koji Nii were with Renesas Electronics Corporation, Tokyo 187-8588, Japan.

Koji Nii and Kazutoshi Kobayashi are with Kyoto Institute of Technology, Kyoto 606-8585, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier XXX

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 standby mode. Solid failures caused by defects, which are not dependent on the temperature change, are also screened out by the first wafer probing test. Next, in the three temperature steps, the wafer probing test at 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 margin 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 write operation (write-V_{min}) becomes worse at LT. Here, we discuss 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 margin in the CMOS logics [15],[16]. The lowered voltage is also effective for embedded SRAMs. Fewer write margin bits can be screened out at RT testing with additional appropriate GB voltage. However, voltage GB technique sometimes induces overscreening, leading to undesirable yield loss [17]. 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.

Figures 2(a) and 2(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. 2(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. 2. Fig. 3 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. 4(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. 4(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. 4(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. 4(b). These dies are overscreened if we introduce RT testing with GB voltage instead of LT testing at typical voltages.

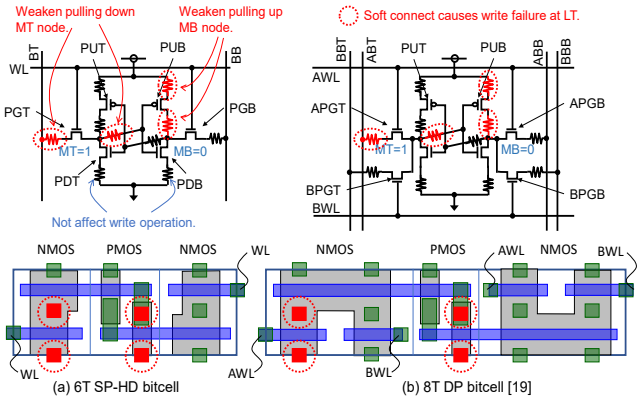


Fig. 2. Write failure models with abnormal contact-diffusion high resistance. Red rectangles are contacts with high resistance.

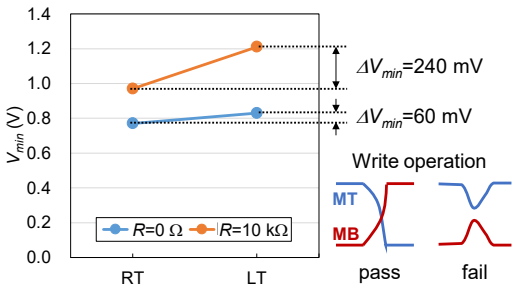


Fig. 3. Temperature dependence of write-V_{min} with normal resistance contact (R=0 Ω) and abnormal high-resistance contact (R=10 kΩ).

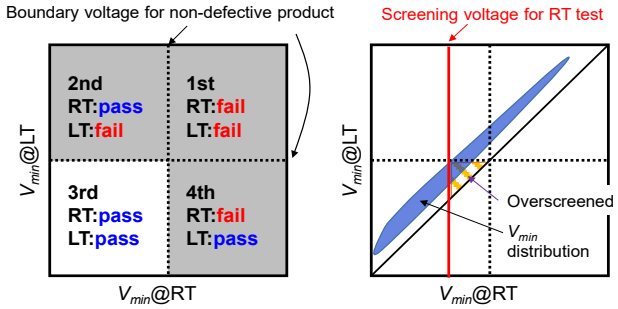


Fig. 4. Pass and fail categories by V_{min} at RT vs. LT.

III. Proposed Pseudo Low Temperature Test

A. Pseudo Low Temperature (PLT) Test

In this section, we propose a PLT test method to eliminate the LT test flow. Fig. 5 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. 2. Fig. 5(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. 5(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" [18],[19]. 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.

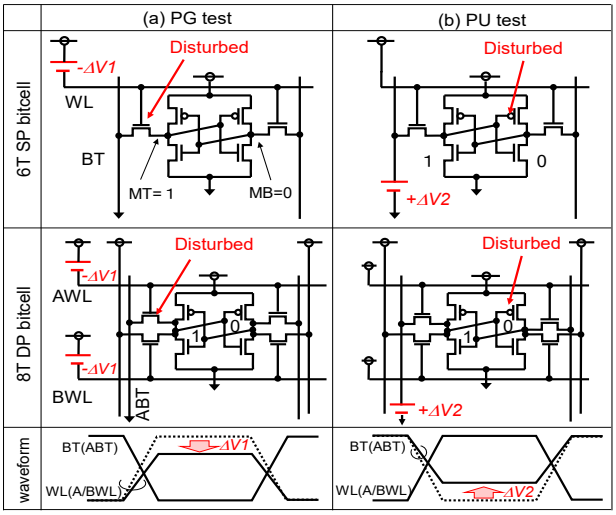


Fig. 5. Concept of reproducing the LT test conditions.

Fig. 6(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. 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. 6(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. 6(a) portray samples that pass the RT test and which fail the LT test if voltage GB technique and PLT would be applied.

Fig. 6(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. 5. $\Delta V1$ and $\Delta V2$ are calculable by the difference of Vt between low temperature and RT with SPICE simulation. They depend only on bitcell transistors. Compared to the original V_{min} distributions of RT vs. LT (Fig. 6(a)), the GB voltage between LT vs. PLT is smaller, narrowing the distributions. By finding the optimal disturbance conditions of $\Delta V1$ and $\Delta V2$, 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. 7 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" [19]. Slightly remained over-screened bits in the proposed scheme of Fig. 6 and 7 are derived from some variations of Vt shift from RT to LT.

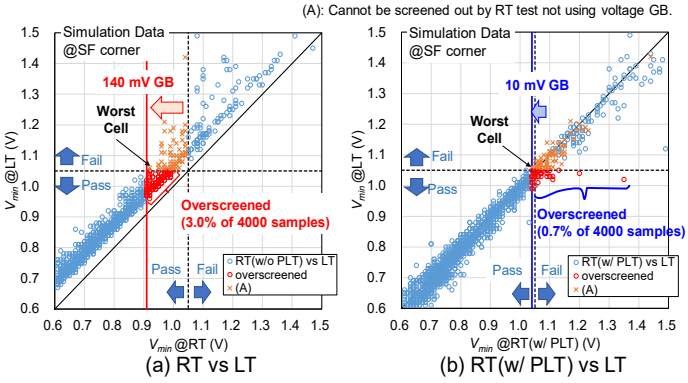


Fig. 6. Simulated write-V_{min} distribution of 6T SP-HD bitcell: (a) RT vs. LT and (b) RT with PLT vs. LT.

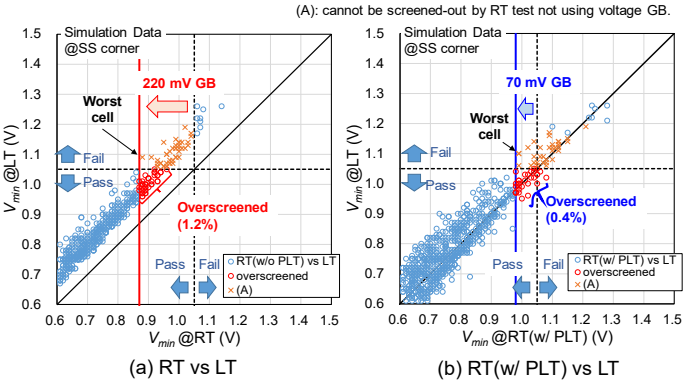


Fig. 7. Simulated write-V_{min} distribution of 8T DP bitcell: (a) RT vs. LT and (b) RT with PLT vs. LT.

B. Circuit Design

Fig. 8 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"), each pull-up NMOS "MNBB" and "MNBT" in Fig. 8 connected to each write amplifiers (WA in Fig. 8) 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. 9 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 I presents $\Delta V1$ and $\Delta V2$ settings shown in Fig. 5. 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 DP SteRAM, as shown in Fig. 7 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 [18],[19]. The calculated GB voltages for the conventional test flow and the proposed test flow are, respectively, 220 mV and 70 mV.

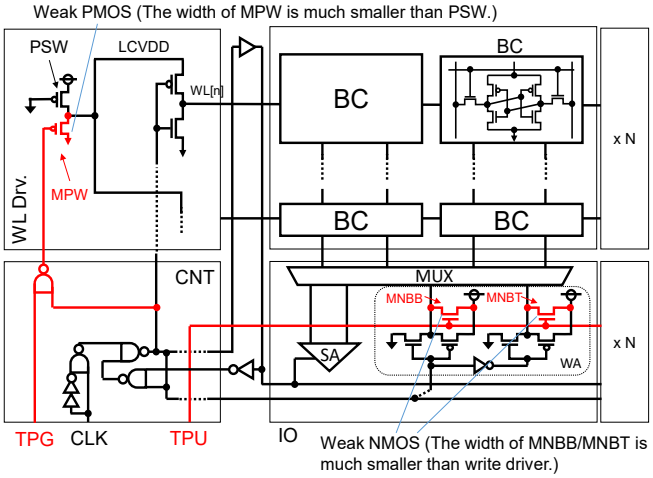


Fig. 8. Schematic diagram of SP SRAM macro with PLT circuitry.

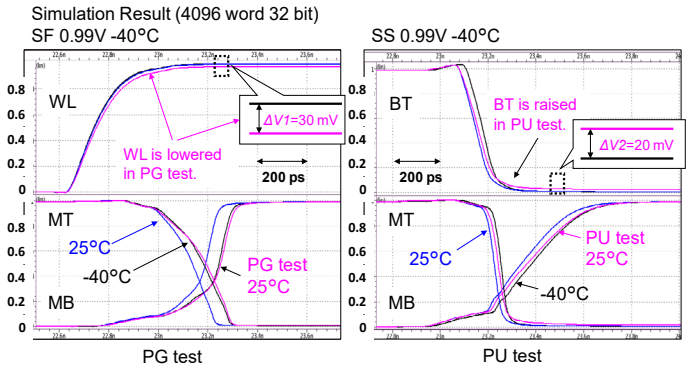


Fig. 9. Simulated waveform of write operation in PLT of 6T SP-HD bitcell.

TABLE I
BIAS SETTING FOR EACH BITCELL IN THE PLT TEST MODE

	$\Delta V1$	$\Delta V2$
6T SP-HD bitcell	30 mV	20 mV
6T SP-HC bitcell		
8T DP bitcell	33 mV	30 mV

IV. Design and Evaluation of 40 nm Test Chip

Fig. 10 portrays a photograph of the test chip die used for the proposed SRAM macros. For targeting low-power MCUs intended for IoT or wearable markets, we use 40 nm low-power CMOS technology. TABLE II 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.

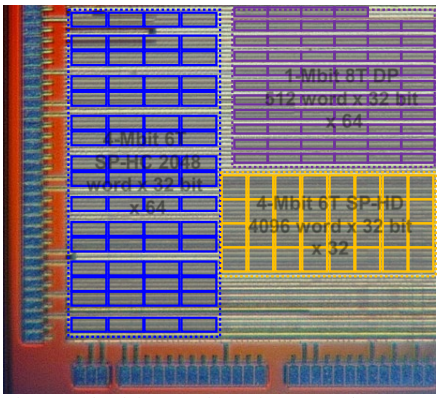


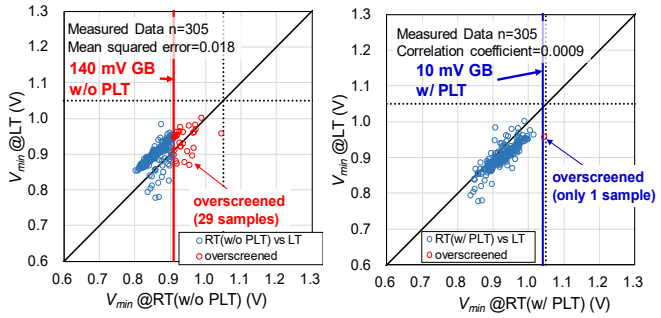
Fig. 10. Microphotograph of the test chip using 40 nm low-power CMOS.

TABLE II
TEST CHIP FEATURES

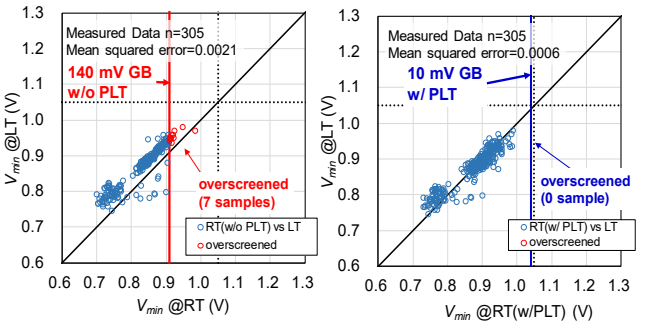
	Bitcell type		
	6T SP-HD	6T SP-HC	8T DP
Technology	40 nm low-power CMOS process		
Macro configuration	128-kbit (4096 word x 32 bit)	64-kbit (2048 word x 32 bit)	16-kbit (512 word x 32 bit)
Macro size of instances	207.3 µm x 183.9 µm (38121 µm^2)	291.1 µm x 111.5 µm (32468 µm^2)	274.3 µm x 67.1 µm (18398 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. 11(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. 6, 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 Figs. 16(b) and 16(c). TABLE III 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.

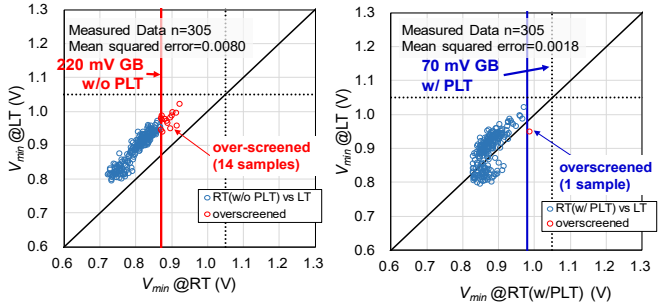
To adopt the PLT method for a leading product in early production stages, many measured data are collected in the LT and PLT tests, improving the GB voltage accuracy for the PLT test. During the mass production stage, one can eliminate the LT tests using the most appropriate GB voltage for the PLT test. 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. 11. 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 III
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%

V. CONCLUSION

We proposed the pseudo-low-temperature test (PLT) method for 40 nm single-port SRAM and dual-port SRAM to reduce testing times by eliminating low-temperature condition tests. The proposed test circuit reproduced low-temperature conditions at RT with good correlation. We designed and fabricated SRAM macros with bitcells of three types on 40 nm low-power CMOS technology and 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.

ACKNOWLEDGMENTS

The authors are grateful to members of the embedded memory development team for their supports and helpful discussions. We also thank Atsushi Miyanishi, Toshihiro Inada, Koji Tanaka, and Miki Tanaka for their encouragement and support.

REFERENCES

[1] 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.

[2] ITRS2001, “https://www.semiconductors.org/”.

[3] 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.

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

[5] 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.

[6] K. Nii et al., “A 65 nm Ultra-High-Density Dual-Port SRAM with 0.71μm^2 8T-Cell for SoC,” *VLSI Cir. Symp. Dig.,* 2006, pp. 130–131.

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.