Design and implementation of an Economical Approach to Test Screening on LT-spice Embedded SRAMs for Low-Power MCUs

Dr.A.Ranganayakulu [(1)] , Mr.D.Satyanarayana [(2)] , Vudumula Anil Kumar Reddy [(3),] Dudekula Nagula Meera Vali [(4),] Bandaru Gurubrahmam [(5),] Ballani Chandu [(6),]
[1,2] *Krishna Chaithanya Institute Of Technology & Sciences, Ece Department, Markapur, Andhra Pradesh.*
[3,4,5,6,] *Krishna Chaithanya Institute Of Technology & Sciences, UG Student-ECE, Markapur, Andhra Pradesh.*

Abstract. In this paper, For low-power microcontroller units (MCUs), embedded static random-access memory (SRAMs) with reasonably priced test screening circuitry are demonstrated. In the package test, where a sample is tested at room temperature (RT), pseudo-LT (PLT) conditions are imitated to avoid the probing test step at the low temperature (LT) of −40 ◦C. A good minimum operating voltage (Vmin) correlation between LT and PLT circumstances is confirmed by Monte Carlo simulation that takes into account both local Vt fluctuations and contact soft open failure (high resistance). Utilizing low-power 40-nm CMOS technology, test chips with two varieties of 4-Mbit single-port SRAM macros and 1-Mbit dual-port SRAM macros are designed and manufactured. The suggested test approach replicates LT conditions and screens out LT failures with less over screening, according to measurement of the data. This project can be implemented by using the LT-spice analysis.

Keywords: Embeded SRAMs,LT spice FPGA, Verilog HDL Static Random Access Memory(SRAM).

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, input-output (IO) interfaces, analog blocks, and memory blocks is crucially important to produce cost-competitive products [4]. Especially, because recent MCUs typically include one or more high-density (HD) single-port (SP) static random-access memory (SRAM) blocks for high-speed cache access, reducing testing times is important not only for embedded nonvolatile 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 built-in self test (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 is indispensable to screen dies, which exhibit temperature-dependent failures [10]. Additionally, package testing after die sawing and assembly is generally performed at room temperature (RT). Therefore, dies should be tested under at least three temperature conditions: high temperatures (HTs) of 125 ◦C, RT of 25 ◦C, and low temperature (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].

1.1 OVERSCREENED ISSUES OF VOLTAGE GB TECHNIQUES

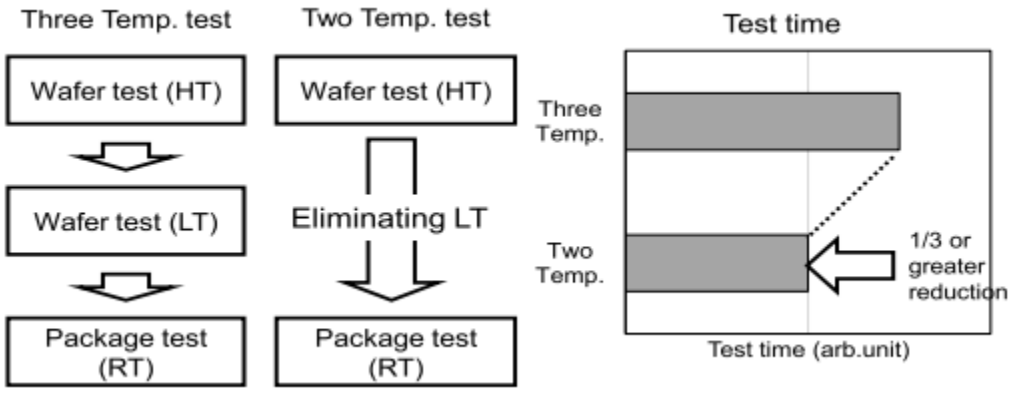


FIG.1 Test screening flows with three temperature steps (typical) and two temperature steps (eliminating LT).

In testing flow with either three-temperature steps or two-temperature steps as shown in Fig. 1, 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 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 (Vmin) at read operation (read-Vmin) becomes worse at HT because of SNM temperature dependence, whereas the SRAM Vmin at write operation (write-Vmin) becomes worse at LT. Here, we discuss how to screen out the write-Vmin 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 logic [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, the voltage GB technique sometimes induces overscreening, leading to undesirable yield loss [17]. Especially, the failure SRAM bits at LT has no constant write-Vmin offsets between LT and RT conditions because variations of write-Vmin 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.

Static random access memory (SRAM), type of random access memory that gives fast access to data but is physically relatively large. Random access memory (RAM) is computer main memory in which specific contents can be accessed (read or written) directly by the central processing unit (CPU) in a very short time regardless of the sequence (and hence location) in which they were recorded. SRAM consists of flip-flops, bistable circuits composed of four to six transistors. Once a flip-flop stores a bit, it keeps that value until the opposite value is stored in it. SRAM is used primarily for small amounts of memory called registers in a computer's CPU and for fast "cache" memory.

In contrast, dynamic RAM (DRAM) stores each bit in an electrical capacitor rather than in a flip-flop, using a transistor as a switch to charge or discharge the capacitor. Because it has fewer electrical components, a DRAM storage cell is smaller than SRAM. However, access to its value is slower, and, because capacitors gradually leak charges, stored values must be recharged approximately 50 times per second. In a SRAM chip each memory cell stores a binary digit (1 or 0) for as long as power is supplied. In a DRAM chip the charge on individual memory cells must be refreshed periodically to retain data. DRAM is generally used for main memory because the same size chip can hold several times as much DRAM as SRAM.

RAM, computer main memory in which specific contents can be accessed (read or written) directly by the central processing unit in a very short time regardless of the sequence (and hence location) in which they were recorded. Two types of memory are possible with random-access circuits: static RAM (SRAM) and dynamic RAM (DRAM). A single memory chip is made up of several million memory cells. In a SRAM chip each memory cell stores a binary digit (1 or 0) for as long as power is supplied. In a DRAM chip the charge on individual memory cells must be refreshed periodically to retain data. Because it has fewer components, DRAM requires less chip area than SRAM. Hence, a DRAM chip can hold more memory, though its access time is slower.

II. PROPOSED METHOD AND ITS METHODOLOGY

1.EXISTING SYSTEM

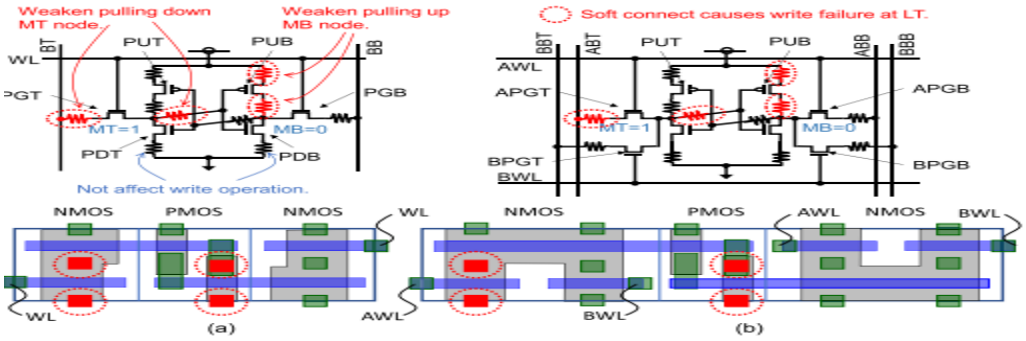


Fig2.1. (a) 6T SP-HD bitcell. (b) 8TDP bitcell

Fig. 2(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 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 [pass gate true (PGT)] or pull-up pMOS [pull up bar (PUB)], which have high resistance of more than several kilo-ohms, then the write-Vmin of the bitcell is much worse than that of RT conditions. Neither pull down true (PDT) nor pull down bar (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-Vmin with PGT = 0 and 10 kOHM for a 6T SP bitcell. The Vmin offset between RT and LT with PGT = 0 V, is 60 mV, whereas that with PGT = 10 kOHM is 240 Mv.

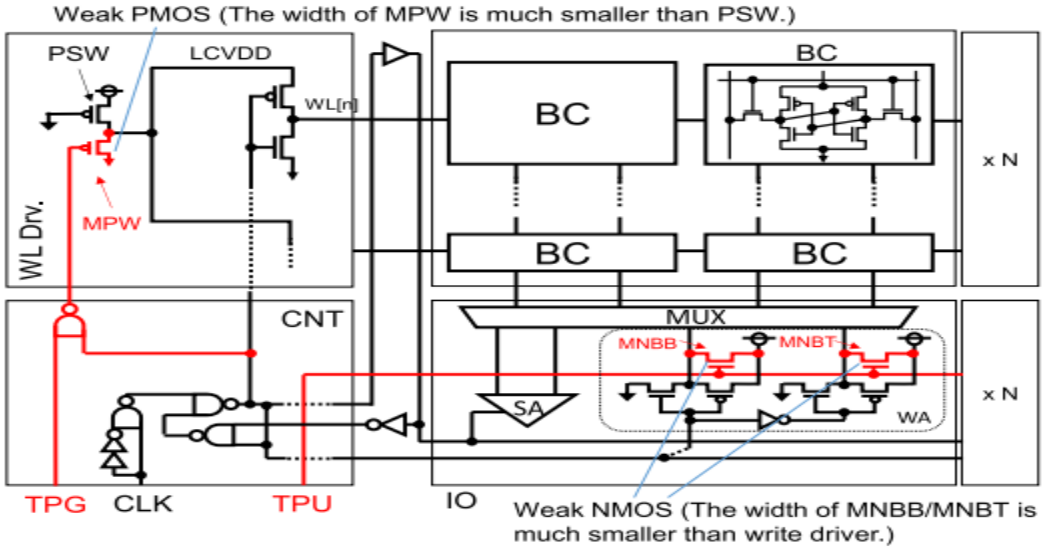
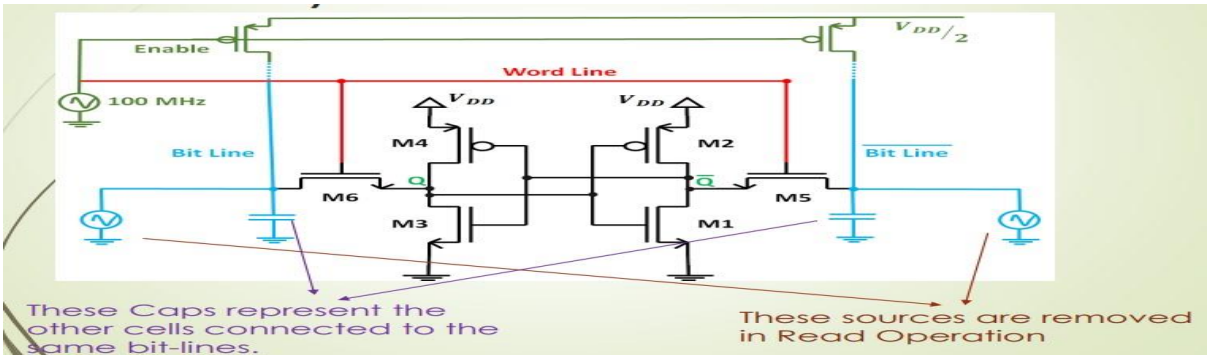


Fig2.2. Schematic of SP SRAM macro with PLT circuitry.

2.PROPOSED SYSTEM



PROPOSED 6T SRAM READ AND WRITE OPERATED CIRCUITS.

Static Random Access Memory (SRAM) is a type of volatile memory that is used to store and provide fast access to data in electronic devices. It is called "static" because it doesn't require

constant refreshing like Dynamic Random Access Memory (DRAM). SRAM retains data as long as power is applied to the system. Here's an explanation of key aspects of SRAM:

1. Cell Structure:

The basic storage unit in SRAM is a flip-flop circuit. Each memory cell typically consists of four to six transistors arranged in a cross-coupled latch configuration. This structure allows the cell to store a binary bit (0 or 1) in a stable state.

2. Operation:

SRAM stores data in a binary form (0s and 1s) using the on/off states of transistors in its flip-flop cells. Writing to SRAM involves setting the states of the transistors to represent the desired data. Reading involves sensing the states of the transistors to retrieve the stored information.

3. Volatility:

SRAM is a volatile memory, meaning that it loses its stored information when power is turned off. Unlike non-volatile memory types (such as Flash or hard drives), SRAM requires a continuous power supply to maintain the stored data.

4. Speed and Access Time:

SRAM is known for its fast access times and low latency. It allows for quick read and write operations compared to other types of memory like DRAM. This makes SRAM well-suited for cache memory in processors, where fast access to frequently used data is crucial.

5. Applications:

SRAM is commonly used in processor cache memories, as well as in other high-speed and high-performance applications where fast and frequent access to data is essential. It is used in various electronic devices, such as computers, embedded systems, and networking equipment.

6. Power Consumption:

While SRAM provides fast access times, it typically consumes more power per bit compared to DRAM. This is because the flip-flop cells require more transistors and a constant power supply to maintain the stored information.

7. Density and Cost:

SRAM tends to have lower storage density compared to DRAM. As a result, it is more expensive to manufacture and is often used in smaller capacities due to cost considerations.

8. Write Endurance:

SRAM cells generally have high write endurance, meaning they can withstand many read and write cycles. This characteristic is advantageous in applications where frequent data updates occur.

SP-SRAM

Single-Port (SP) SRAM (Static Random Access Memory) refers to a type of SRAM configuration where each memory cell has a single access port, allowing either read or write access at any given time. This contrasts with dual-port SRAM, where each memory cell has two access ports, enabling simultaneous read and write operations. Here's an explanation of key aspects of single-port SRAM:

1. Basic Structure:

In a single-port SRAM, each memory cell typically consists of six transistors arranged in a flip-flop configuration. This flip-flop configuration provides stable storage of a single binary bit (0 or 1).

2. Access Port:

The term "single port" indicates that each memory cell has a single access port for data operations. This means that at any given time, the cell can either be read from or written to, but not both simultaneously.

3. Read Operation:

During a read operation, the stored data in the selected memory cell is read out through the access port. The read operation is non-destructive, meaning it does not alter the stored data.

4. Write Operation:

During a write operation, new data can be written into the selected memory cell through the access port. The write operation may involve changing the state of the flip-flop configuration to represent the desired binary value.

5. Volatility:

Like other types of SRAM, single-port SRAM is volatile, meaning it loses its stored data when the power is turned off. Continuous power is required to maintain the stored information.

6. Applications:

Single-port SRAM is commonly used in various electronic devices and applications where fast and frequent access to data is required. It is found in processor cache memories, register files, and other memory structures within computing systems.

7. Speed and Latency:

Single-port SRAM provides fast access times and low latency, making it suitable for applications where quick access to data is critical. This characteristic is particularly important in high-performance computing environments.

8. Density and Cost:

The density of single-port SRAM is generally lower than that of dual-port SRAM. This means that, in terms of storage capacity, single-port SRAM may be more cost-effective to produce but is typically used in smaller capacities compared to other types of memory.

9. Write Endurance:

Single-port SRAM cells generally have high write endurance, allowing them to withstand many read and write cycles. This characteristic is advantageous in applications where frequent data updates occur.

III. RESULTS AND ANALYSIS DISCUSSION

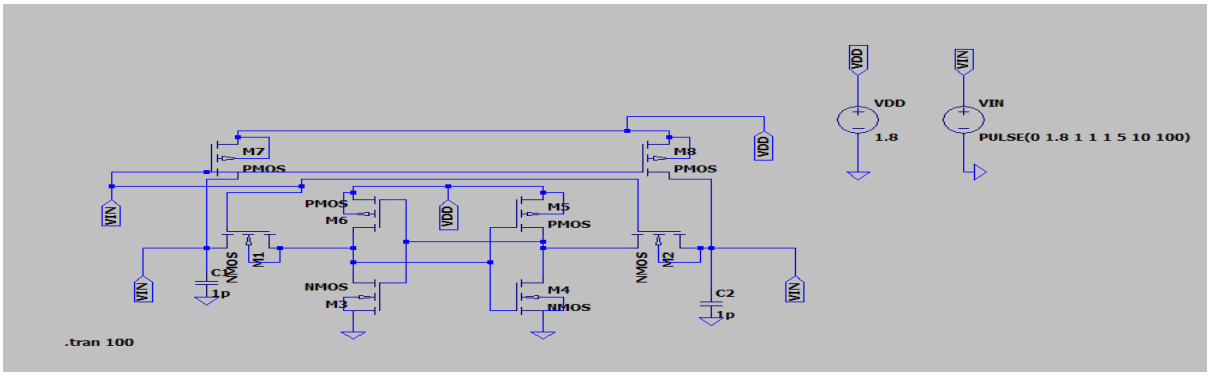


FIG.4.1 6T SRAM WRITE OPERATION CIRCUIT

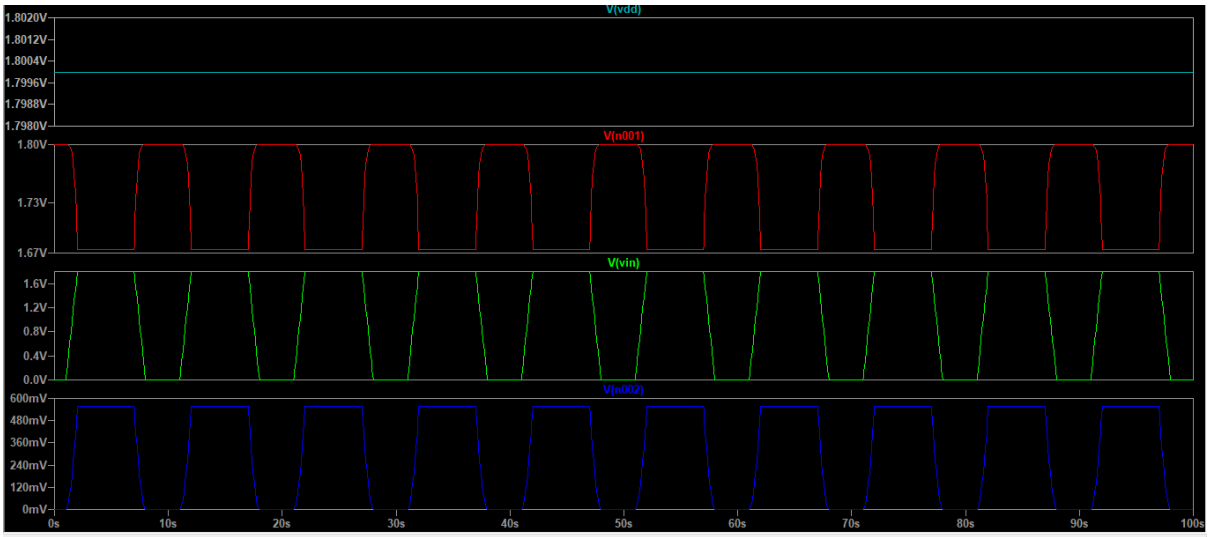


FIG2.6T SRAM WRITE OUTPUT

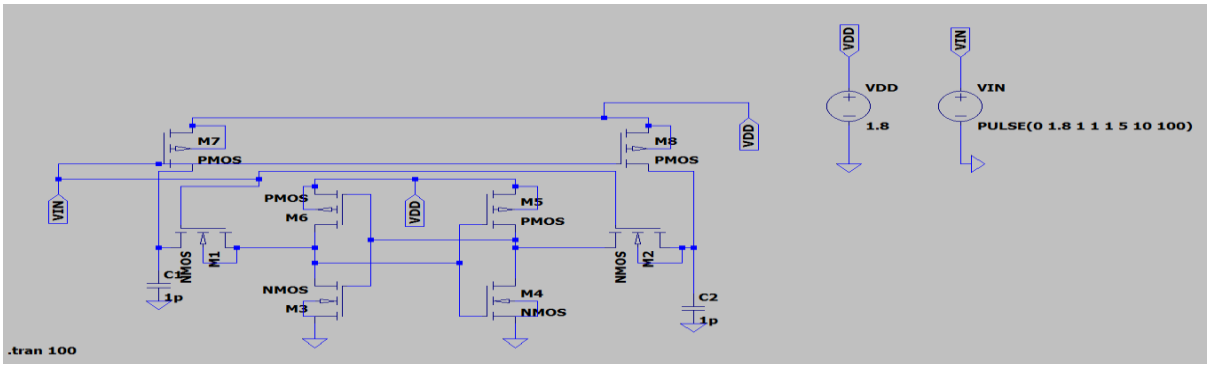


FIG.3. 6T SRAM READ OPERATION CIRCUIT

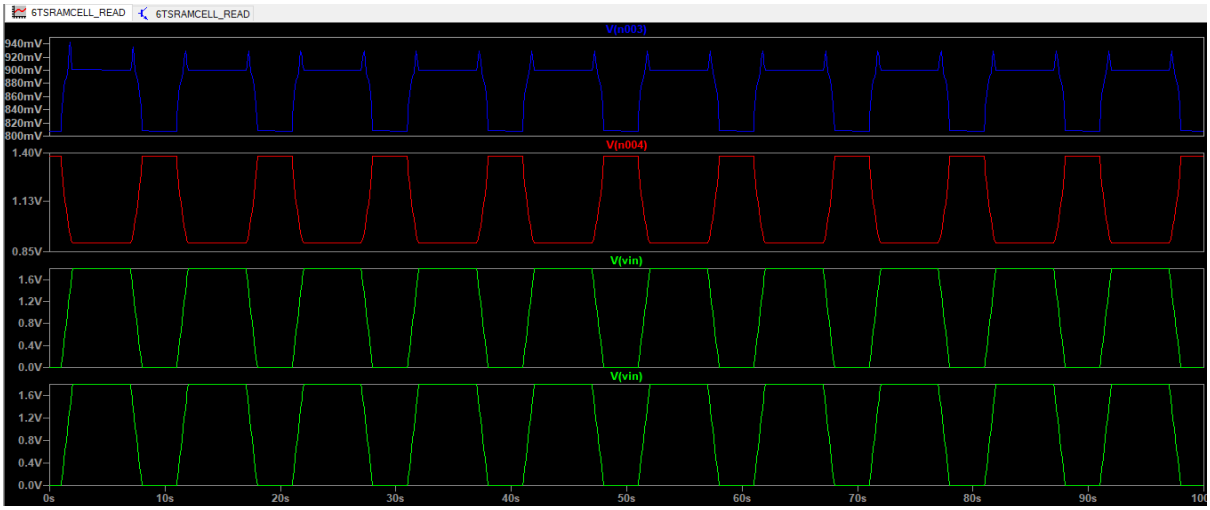


FIG.4 6T SRAM READ OUTPUT

CONCLUSION

This paper, We proposed the 6T SRAM test method for LT SPICE SP SRAM and DP SRAM to reduce testing times by eliminating LT condition tests. The proposed test circuit reproduced 6T conditions at RT with good correlation. We designed and fabricated SRAM macros with bitcells of three types on LT SPICE low-power CMOS technology and confirmed that the proposed test circuitry screens out LT 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 over screened dies compared to the conventional voltage GB test method (e.g., the proposed PLT test method reduces the number of over screened 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.

REFERENCES

[1] X. Ma, Y. Lu, R. P. Martins, and Q. Li, “A 0.4 V 430 nA quiescent current NMOS digital LDO with NAND-based analog-assisted loop in
28nm CMOS,” in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig.Tech. Papers, Feb. 2018, pp. 306–308.
[2] S. Kundu, M. Liu, S.-J. Wen, R. Wong, and C. H. Kim, “A fully integrated digital LDO with built-in adaptive sampling and active voltage positioning using a beat-frequency quantizer,” IEEE J. SolidState Circuits, vol. 54, no. 1, pp. 109–120, Jan. 2019.
[3] J. Oh, J.-E. Park, Y.-H. Hwang, and D.-K. Jeong, “25.2 A 480 Ma output-capacitor-free synthesizable digital LDO using CMP-triggered oscillator and droop detector with 99.99% current efficiency, 1.3 ns response time, and 9.8A/mm2 current density,” in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2020, pp. 382–384.
[4] R. J. Milliken, J. Silva-Martinez, and E. Sanchez-Sinencio, “Full on-chip CMOS low-dropout voltage regulator,” IEEE Trans. Circuits Syst. I, Reg.Papers, vol. 54, no. 9, pp. 1879–1890, Sep. 2007.
[5] J. Guo and K. N. Leung, “A 6-μW chip-area-efficient outputcapacitorless LDO in 90-nm CMOS technology,” IEEE J. Solid-State Circuits, vol. 45, no. 9, pp. 1896–1905, Sep. 2010.
[6] S. Bu, J. Guo, and K. N. Leung, “A 200-ps-response-time output-capacitorless low-dropout regulator with unity-gain bandwidth >100 MHz in 130-nm CMOS,” IEEE Trans. Power Electron., vol. 33,no. 4, pp. 3232–3246, Apr. 2018.
[7] X. Ma, Y. Lu, and Q. Li, “A fully integrated LDO with 50-mV dropout for power efficiency optimization,” IEEE Trans. Circuits Syst. II, Exp.Briefs, vol. 67, no. 4, pp. 725–729, Apr. 2020.
[8] K. N. Leung and P. K. T. Mok, “A capacitor-free CMOS low-dropout regulator with damping-factor-control frequency compensation,” IEEE J. Solid-State Circuits, vol. 38, no. 10, pp. 1691–1702, Oct. 2003.
[9] M. Kim, “Two-port feedback analysis on miller-compensated amplifiers,” 2022, arXiv:2207.10983.
[10] M. Kim and S. Cho, “A single BJT bandgap reference with frequency compensation exploiting mirror pole,” IEEE J. Solid-State Circuits,vol. 56, no. 10, pp. 2902–2912, Oct. 2021.
[11] A. S. Sedra and K. C. Smith, Microelectronic Circuits, 5th ed. Oxford,U.K.: Oxford Univ. Press, 2004.
[12] P. R. Gray, Analysis and Design of Analog Integrated Circuits, 5th ed.Hoboken, NJ, USA: Wiley, 2009.

[13] A. D. Grasso, G. Palumbo, and S. Pennisi, "High-performance four-stage CMOS OTA suitable for large capacitive loads," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 62, no. 10, pp. 2476–2484, Oct. 2015.
[14] J. Roberge, Operational Amplifiers: Theory and Practice (Radiation Laboratory). Hoboken, NJ, USA: Wiley, 1975. [Online]. Available:
[15] G. F. Franklin, D. J. Powell, and A. Emami-Naeini, Feedback Control of Dynamic Systems, 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[16] G. A. Rincon-Mora, "Active capacitor multiplier in miller-compensated circuits," IEEE J. Solid-State Circuits, vol. 35, no. 1, pp. 26–32,Jan. 2000.
[17] J.-E. Park, J. Hwang, J. Oh, and D.-K. Jeong, "32.4 A 0.4-to1.2 V 0.0057 mm2 55 fs-transient-FoM ring-amplifier-based low-dropout regulator with replica-based PSR enhancement," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2020, pp. 492–494.