

Portland State University

**PDXScholar**

---

Dissertations and Theses

Dissertations and Theses

---

1-1-2010

# The Development of Embedded DRAM Statistical Quality Models at Test and Use Conditions

Satoshi Suzuki

*Portland State University*

Follow this and additional works at: [https://pdxscholar.library.pdx.edu/open\\_access\\_etds](https://pdxscholar.library.pdx.edu/open_access_etds)

**Let us know how access to this document benefits you.**

---

## Recommended Citation

Suzuki, Satoshi, "The Development of Embedded DRAM Statistical Quality Models at Test and Use Conditions" (2010). *Dissertations and Theses*. Paper 341.

<https://doi.org/10.15760/etd.341>

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: [pdxscholar@pdx.edu](mailto:pdxscholar@pdx.edu).

The Development of Embedded DRAM Statistical Quality Models at Test and  
Use Conditions

by  
Satoshi Suzuki

A thesis submitted in partial fulfillment of the  
requirements for the degree of

Master of Science  
in  
Electrical and Computer Engineering

Thesis Committee:  
W. Robert Daasch, Chair  
Branimir Pejcinovic  
C. Glenn Shirley  
Christof Teuscher

Portland State University  
©2010

## **Abstract**

Today, the use of embedded Dynamic Random Access Memory (eDRAM) is increasing in our electronics that require large memories, such as gaming consoles and computer network routers. Unlike external DRAMs, eDRAMs are embedded inside ASICs for faster read and write operations. Until recently, eDRAMs required high manufacturing cost. Present process technology developments enabled the manufacturing of eDRAM at competitive costs.

Unlike SRAM, eDRAM exhibits retention time bit fails from defects and capacitor leakage current. This retention time fail causes memory bits to lose stored values before refresh. Also, a small portion of the memory bits are known to fail at a random retention time. At test conditions, more stringent than use conditions, if all possible retention time fail bits are detected and replaced, there will be no additional fail bits during use. However, detecting all the retention time fails requires long time and also rejects bits that do not fail at the use condition. This research seeks to maximize the detection of eDRAM fail bits during test by determining effective test conditions and model the failure rate of eDRAM retention time during use conditions.

## Acknowledgements

I would like to take this opportunity to express my gratitude to many wonderful people who have encouraged me during my graduate study at Portland State University. First of all, this research would not have been successful without the help and support of my academic advisor Dr. Robert Daasch. His outstanding knowledge and research involved in digital design and test helped me not only achieve the goal but also improve my critical thinking and problem solving skills. Also, I am very thankful to Glenn Shirley for providing me with industrial insights from his more than 20 years of Quality and Reliability modeling experience. He has helped me stay focused on the needs from customers for this research.

Also, I would like to express my gratitude to Professor Branimir Pejcinovic and Professor Christof Teuscher as members of my M.S. Defense Examination Committee. I would also like to thank the staff in the Electrical and Computer Engineering Department for their assistance throughout my graduate study.

I would like to thank LSI Corporation for giving me this research opportunity and necessary funds. Thanks to the ICDT lab colleagues Shavali, Ritesh, Kalyan, Biki, Muhammad, Saikiran, Saurabh, and Kapil for the support and encouragements. I personally would like to thank my parents and my sister who live in Japan for their constant support. I would like to thank my grandfather for unforgettable memories who died while I was in the U.S. Moreover, I am thankful to Marina for staying close together and being patient with me for many years. I would not be what I am today without her support and intelligence.

## Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgement</b>	<b>ii</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>vii</b>
<b>1 Introduction: Motivation, Contribution, and Implementation</b>	<b>1</b>
1.1 Motivation of this Research . . . . .	1
1.2 Contributions . . . . .	3
1.3 Implementation of this Research . . . . .	3
1.4 Organization of this Thesis . . . . .	4
<b>2 Backgrounds: RAM History and Quality Models</b>	<b>5</b>
2.1 History and Differences of Random Access Memories . . . . .	5
2.1.1 History of RAMs: SRAM, DRAM, and Embedded RAMs . .	6
2.1.2 Design Differences between SRAM and DRAM . . . . .	8
2.1.3 Differences between embedded DRAM and external DRAM	10
2.1.4 eDRAM Applications . . . . .	11
2.2 Operation and Retention Time Fails of eDRAM . . . . .	11
2.2.1 Operation of eDRAM: : 1T1C-DRAM Architecture . . . . .	12
2.2.2 Retention Time Fails of eDRAM . . . . .	14
2.2.3 Repair and Use Error Detection of DRAM . . . . .	16

2.2.4	eDRAM Technology . . . . .	17
2.3	Quality Figures-of-Merit . . . . .	17
2.3.1	Retention Time Characteristics and Statistical Model . . .	19
2.3.2	Weibull Distribution for Retention Time Characterization .	22
<b>3</b>	<b>Experiments: Test Approach</b>	<b>23</b>
3.1	Device Under Test: eDRAM Test Chip . . . . .	23
3.1.1	Description of eDRAM Test Chip . . . . .	23
3.1.2	Process Skews . . . . .	24
3.2	Test Environment . . . . .	26
3.2.1	IC Tester (ATE) . . . . .	27
3.2.2	eDRAM Test Program . . . . .	28
3.3	Data Collection Process . . . . .	32
3.4	Modeling eDRAM Bit Fails . . . . .	32
3.4.1	Model Extraction for Single Retention Time Bits . . . . .	34
3.4.2	Test and Use Condition Model . . . . .	38
3.4.3	Fail Bit Rate Prediction For Variable Retention Time Bits .	39
<b>4</b>	<b>Results: Parameter Extraction and Quality Models</b>	<b>45</b>
4.1	Quantitative Results to Derive Quality Models . . . . .	45
4.1.1	Test Conditions: Process, Block, Temperature, Voltage . . .	45
4.1.2	Observed SRT and VRT Results . . . . .	50
4.1.3	Summary of Parameter Extraction . . . . .	51
4.2	Methods of Quality Models . . . . .	54
4.2.1	Single Retention Time (SRT) Use and Test Model . . . . .	55
4.2.2	SRT and VRT Data Example in 2-D . . . . .	57

<b>5</b>	<b>Conclusion: Implications, Applications, and Future Work</b>	<b>59</b>
5.1	Applications . . . . .	59
5.2	Conclusion . . . . .	62
5.3	Future Work . . . . .	62
	<b>References</b>	<b>64</b>

## List of Tables

2.1	Comparison between SRAM and DRAM . . . . .	10
2.2	Comparison between External DRAM and Embedded DRAM . . .	11
2.3	1T1C-eDRAM Future Trend [1] . . . . .	17
3.1	eDRAM Test Chip Process Skews . . . . .	25
3.2	Test and Use Condition Experimental Parameters . . . . .	38
4.1	eDRAM Experimental Conditions: 90 distinct skew, voltage, and temperature conditions. . . . .	45
4.2	Single Retention Time Parameter Coefficients . . . . .	57
5.1	SRT and VRT Cumulative Counts: Original Conditions {Process Skew=w9 (fast), Temperature=125°C, $t_{test} = 1200\mu s$ , $t_{use} = 600\mu s$ , VDD=1.00V, VP=0.40V} Note that <i>Yield Loss</i> =146. . . . .	59
5.2	SRT and VRT Cumulative Counts: Modified Conditions 1 {Process Skew=w9 (fast), $t_{test} = 1200\mu s$ , $t_{use} = 600\mu s$ , Temperature=105°C, VDD=1.20V, VP=0.45V} Note that <i>Yield Loss</i> =69. . . . .	61
5.3	SRT and VRT Cumulative Counts: Modified Conditions 2 {Process Skew=w9 (fast), $t_{test} = 1500\mu s$ , $t_{use} = 800\mu s$ , Temperature=105°C, VDD=1.20V, VP=0.45V} Note that <i>Yield Loss</i> =145. . . . .	61

## List of Figures

2.1	Categories of Memory Arrays . . . . .	6
2.2	History of Random Access Memory . . . . .	7
2.3	6T–SRAM Cell Schematic . . . . .	8
2.4	One Bit DRAM Cell Schematic and Physical Layout: reformatted from <i>CMOS VLSI Design</i> , Weste and Harris, Pearson Education, Inc., 2005. [2] . . . . .	9
2.5	Change from External DRAM to Embedded DRAM . . . . .	10
2.6	1T1C–DRAM Cell Schematic . . . . .	12
2.7	DRAM Cell Read Operation Timing Diagram . . . . .	13
2.8	Categories of Single Event Error . . . . .	14
2.9	Fail Bit Characteristics with Time . . . . .	15
2.10	Sources of Retention Fails: reformatted from <i>CMOS VLSI Design</i> , Weste and Harris, Pearson Education, Inc., 2005. [2] . . . . .	16
2.11	eDRAM Product Test and Use Flow with Figures-of-Merit Mea- sured at Test and Use Conditions . . . . .	18
2.12	Quality of eDRAM Memory Bits at Use and Test Conditions . . . .	21
3.1	65nm eDRAM Test Chip: a) is the top view and b) is the back side view. Floorplan is not to scale. The die size is 4mm $\times$ 3mm. . . . .	24
3.2	65nm eDRAM Test Chip Floorplan . . . . .	25
3.3	Test Environment in the ICDT Lab . . . . .	26
3.4	Test Environment Block Diagram . . . . .	27
3.5	Credence Quartet One ATE . . . . .	28

3.6	Test Environment in ICDT Lab at Portland State University . . . .	29
3.7	Shmoo Data Log: Sweep retention time from $262\mu s$ to $2659\mu s$ (x-axis) and also sweep VDD voltage from 0.85V to 1.20V (Y-axis) . .	30
3.8	Shmoo Test Program Pseudo Code for eDRAM Test Chips . . . . .	31
3.9	Data Conversion Process: Note only the fail bits found during the test are recorded in data logs. . . . .	32
3.10	Pass and Fail Bit Characteristics in Time Domain . . . . .	33
3.11	Single Retention Time Modeling: Note that y-axis is the <i>Weibit</i> and x-axis is the $\ln(t_{ret})$ of Equation 3.3. . . . .	36
3.12	Two-Dimensional Time Domain ( $t_{test}$ and $t_{use}$ ) to analyze DPPM, Pass, Overkill, Yield Loss: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins. (See also Figure 2.17 for the eDRAM quality figures-of-merit.) . . . . .	41
3.13	VRT Two Retention Times Model for Row Counts: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins. . . . .	41
3.14	VRT Two Retention Times Model for Smoothen Counts: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins. . . . .	42
3.15	VRT Two Retention Times Model for Cumulative Counts: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins. . . . .	43

3.16	VRT Two Retention Times Model for Fraction Failure at each retention time stop: Note that y-axis is the maximum retention time bins and x-axis is the minimum retention time bins. . . . .	44
4.1	Fail Bit Counts per One Million Bits Sorted by Memory Block at 125°C, $VP = 0.45V$ , and $VDD = 1.20V$ . . . . .	46
4.2	Fail Bit Counts per One Million Bits Sorted by Process Skew at 125°C, $VP = 0.45V$ , and $VDD = 1.20V$ . . . . .	47
4.3	Fail Bit Counts per One Million Bits Sorted by Temperature at skew = w9, $VP = 0.45V$ , and $VDD = 1.20V$ . . . . .	49
4.4	Fail Bit Counts per One Million Bits Sorted by $VDD$ at skew = w9, $VP = 0.45V$ , $T = 125^\circ C$ . . . . .	49
4.5	Fail Bit Counts per One Million Bits Sorted by $VP$ at skew = w9, $VDD = 1.20V$ , $T = 125^\circ C$ . . . . .	50
4.6	w4 (Nominal) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions . . . . .	51
4.7	w8 (Slow) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions . . . . .	52
4.8	w9 (Fast) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions . . . . .	52
4.9	w10 (Slow) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions . . . . .	53
4.10	w12 (Slowest) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions . . . . .	53
4.11	Fail Bit Counts per One Million Bits (DPM) across All 5 skews up to 2.6ms Retention Time under Environmental Conditions . . . . .	54

4.12	SRT Conservative Modeling Plot of Nominal Process Chip at Test Condition: $VP = 0.45V$ and $VDD = 1.20V$ . See Section 3.4.1 for reference. . . . .	55
4.13	Weibull Fitted Shape Parameter over 90 Different Test Conditions .	56
4.14	SRT and VRT Quality Example in Two-Dimension . . . . .	58
5.1	Cumulative SRT and VRT Counts at High Temperature Condition (1) and High Voltage Condition (2) . . . . .	60

## **1 Introduction: Motivation, Contribution, and Implementation**

Today, embedded or integrated Dynamic Random Access Memory (eDRAM) is used in electronic applications, such as gaming consoles, to support read and write data operations. Unlike external and commodity DRAM, eDRAM is embedded inside the Application Specific Integrated Circuits (ASICs). Compared to standalone memory design, eDRAM provides more design flexibility, faster bandwidth, and lower power consumption because of the shorter electrical path from memory to logic system.

DRAM and Static Random Memory (SRAM) are the two major volatile digital memories. The eDRAM provides the advantages of DRAM over SRAM such as higher data density for a given area. However, the eDRAM has higher manufacturing cost because of its trench storage capacitor; hence embedded SRAM (eSRAM) has dominated the embedded RAM market. Recently, new developments in process technology have enabled eDRAM to compete with eSRAM. Therefore, for large-density applications, eDRAM has begun to replace eSRAM. In order to continue the replacement trend, the quality of eDRAM after test must be addressed [3] and is the primary focus of this research.

### **1.1 Motivation of this Research**

Unlike SRAM, DRAM retains its information in a storage capacitor for a limited amount of time, called the retention time. A memory cell's retention time is largely set by defects and intrinsic leakage currents. Hence, DRAM requires its memory cells to be refreshed every few hundred microseconds (i.e., refresh time). The retention time for some bits in a memory block is shorter than the other bits in

the same block. In some cases, the retention time of a particular bit varies between different read and write cycles [4]. During use, retention time fails occur when the retention time of the memory bits are shorter than the refresh time.

The retention time fails influence the DRAM quality. For example, the system refresh time is  $400\mu\text{sec}$ . Suppose for a particular memory cycle, the retention time of a specific bit is  $600\mu\text{sec}$ . The bit retention time is greater than the system refresh time, so the bit retains the data throughout the memory cycle. But, in the next memory cycle, the retention time is reduced  $300\mu\text{sec}$  which is shorter than the refresh time, and the bit would lose its data in this memory cycle. If the variation of retention time of  $300\mu\text{sec}$  and  $600\mu\text{sec}$  was fortunately found by testing the test retention time, then the bit can be replaced or removed with a new bit at test. During use, there will be no system problem from the bit's retention time. However, if the variable retention time bit was found to pass at test, the bit might sometimes fail during use. This variable retention time of a bit is a reduction of the DRAM quality in use.

As the DRAM begins to be embedded on ASICs and the leakage current and the possibility of defects increase due to smaller transistor size, the variable retention time fails will be a more significant factor in the observed eDRAM quality level.

Test conditions are more stringent than use conditions. If variable retention time bits and the dead on arrival (DOA) bits are detected and repaired at test, a limited number of additional bits will fail during use. However, detecting all possible variable retention time bits at test requires long test times and significantly increases test cost. The purpose and goal of this research is to maximize the detection of

variable retention time bits at test and limit the test cost. By determining the effectiveness of test conditions to detect variable retention time bits, the eDRAM quality level can be improved.

## **1.2 Contributions**

In summary, the following contributions emerge from this research:

- Identify the average retention time characteristics of eDRAM bits by repeating identical retention time tests with a wide range of sample chips.
- Model the trends of retention time fail bits based on experiments over a range of temperature, voltage, location of memory blocks, and process skew.
- Develop statistical quality models to evaluate eDRAM quality at use after screening the entire memory at test conditions.

## **1.3 Implementation of this Research**

In this eDRAM memory research, the eDRAM statistical quality model was extracted from the observed data. The quality model predicts the fraction failing memory bits in use and the fraction overkill after screening at a particular test condition. The model can be used by test engineers to identify test and use conditions to minimize the number of retention time fail bits in use in trade-offs of test time, yield, and overkill. In Chapter 2 and 3, the quality figures-of-merit are described in detail.

## 1.4 Organization of this Thesis

The thesis is divided into five chapters. Chapter 1 outlines the motivation and implementation of the eDRAM research. Chapter 2 discusses the background, history, and current trends of RAMs by illustrating the differences among DRAM, SRAM, and external and embedded DRAM. Chapter 3 explains the eDRAM experiments. The quality modeling strategy is also presented in Chapter 3. Chapter 4 presents the results of the retention time characterization and model parameter extractions. Chapter 5 concludes the thesis with data-driven applications and describes the lessons learned and possible future work.

## 2 Backgrounds: RAM History and Quality Models

In modern technology, the operations that logic systems perform are in the form of machine language with binary data. The machine language program needs to be stored in a memory so that the logic systems can read and write the instructions to store and modify the data. Random Access Memory (RAM) is the most common memory device to store instructions and data. Today, many applications demand RAMs with large density, high data rate, low power consumption, and small area. Recently, embedded DRAM (eDRAM) has become very attractive because of its large density and the lowering of its manufacturing cost [2, 5].

Chips need to be tested prior to shipping. The chips have to pass a series of required tests as simple as continuity tests and complex functional and structural tests. In particular, memory tests allow test engineers to check an array of memory bits and detect the addresses of fail bits. Because eDRAM performance is limited by retention time, the eDRAM quality is measured in two time scales set by use conditions and test conditions. In Chapter 2, the eDRAM use quality after test is described.

### 2.1 History and Differences of Random Access Memories

Figure 2.1 shows the basic categories of random access memories. There are two types of random access memories: *volatile* and *non-volatile*. When power is removed, volatile memories lose the stored contents. Two common volatile memory examples are Static Random Access Memories (SRAMs) and Dynamic Random Access Memories (DRAM). Non-volatile memories store data permanently after

power is removed, and the non-volatile memories can preserve the data. Examples of non-volatile memories are Read Only Memories (ROMs) such as Erasable Programmable ROM (EPROM), Electrically Erasable PROM (EEPROM), and CD-ROMs [6]. In Section 2.1, DRAMs, SRAMs, and embedded DRAMs are the primary focus to be discussed.

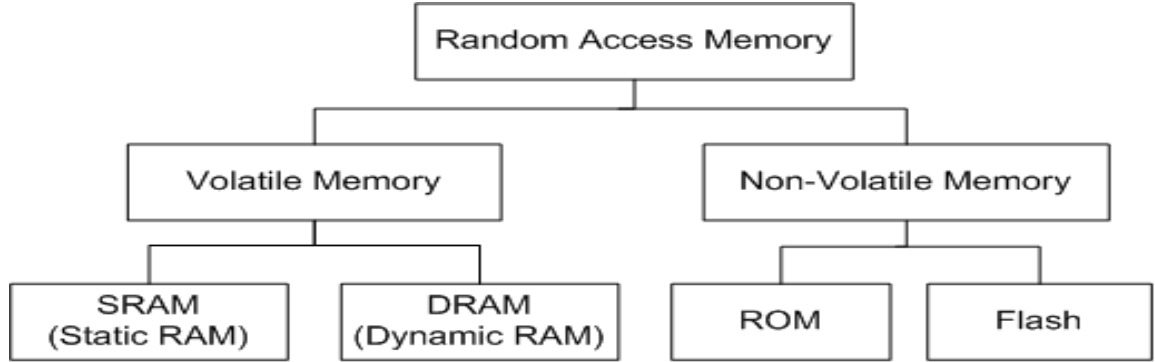


Figure 2.1: Categories of Memory Arrays

### 2.1.1 History of RAMs: SRAM, DRAM, and Embedded RAMs

Random Access Memory devices are connected to the processor by a memory bus which is a high speed interface that can transfer data at high rates. To increase the RAM performances, RAMs tend to be integrated on logic processors as on-chip memories [3]. Figure 2.2 shows the history of change in the use of the RAM.

The three major trends in Figure 2.2 are described in the chronological order.

#### 1. Logic System and RAMs with Extra I/O Bus

Since late 1960s, RAM design and technology have been developed separately from ASICs. To read and write data, an external I/O bus establishes a link between the logic system and RAMs.

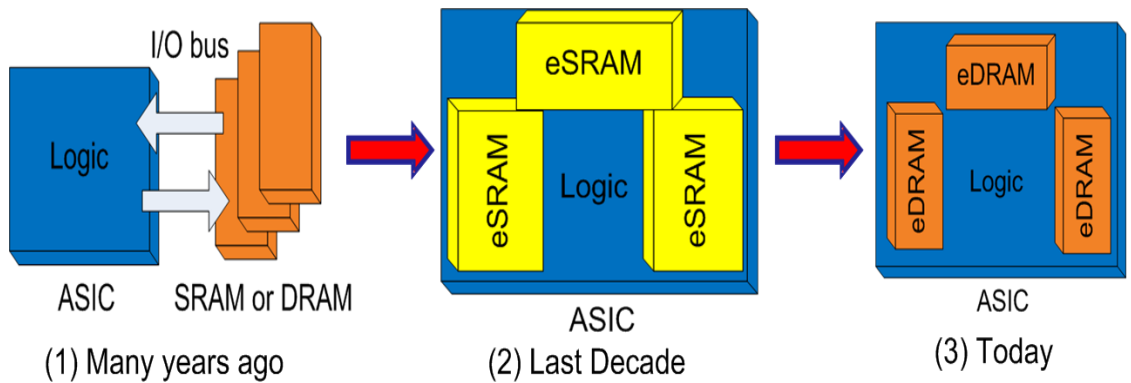


Figure 2.2: History of Random Access Memory

## 2. SRAM embedded on ASIC

To shrink the total board and chip area and improve performance, RAMs were embedded within the ASICs. With the RAMs embedded, ASIC designers have flexibility to design a compact I/O bus between the logic and RAMs. In 1990s, the first RAM to be integrated on ASICs was SRAM because of low manufacturing cost and high speed. At that time, integrated DRAM manufacturing cost was higher because DRAM cells required an expensive trench capacitor for data storage. For twenty years, SRAM dominated the on-chip memory design in ASIC technology.

## 3. DRAM embedded on ASIC

The embedded SRAM requires 50% or more of the ASIC die area. DRAM process manufacturing cost decreased in the 1990s, and large-density DRAM became very attractive. However, DRAM exhibits a retention time fail that the SRAM does not have. Since 1999, large density DRAMs have been integrated on ASIC (i.e., eDRAM). One of the early application examples is the Play Station 2 from Sony Entertainment.

In the next section, the design differences between SRAM and DRAM are specified in more detail.

### 2.1.2 Design Differences between SRAM and DRAM

In Section 2.1.2, the major choices of RAMs are presented in terms of memory architectures, area, power, and manufacturing cost.

#### Static Random Access Memory (SRAM)

The major SRAM architecture is a six-transistor SRAM cell (6T-SRAM). Figure 2.3 shows the six-transistor SRAM cell. The SRAM cell consists of a single wordline and two complementary bitlines. The cell design includes a pair of cross-coupled inverters and two access transistors, one for each bitline and a total of six transistors. The data is stored on the cross-coupled inverters and is accessed through the pair of the access transistors [6].

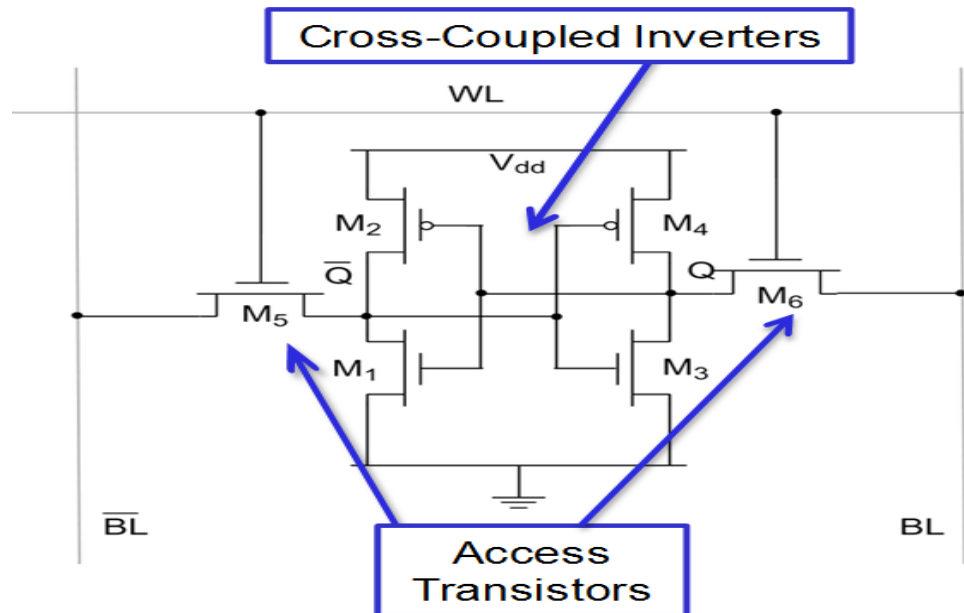


Figure 2.3: 6T-SRAM Cell Schematic

## Dynamic Random Access Memory (DRAM)

The DRAM uses a single access transistor for both read and write operations and stores the data on a static capacitor. The advantage of DRAM is the structural simplicity. In one memory cell (i.e., one bit), there is a single pass transistor and a capacitor compared to six transistors in the SRAM architecture. The structural simplicity enables very high density in a given die area. However, the storage trench capacitor must be grown vertically downward into the silicon substrate as shown in Figure 2.4 (2). For this reason, the DRAM manufacturing cost is considerably higher when compared to SRAM. A disadvantage of DRAM is data in DRAM must be continually refreshed to maintain the valid data from leakage current [2].

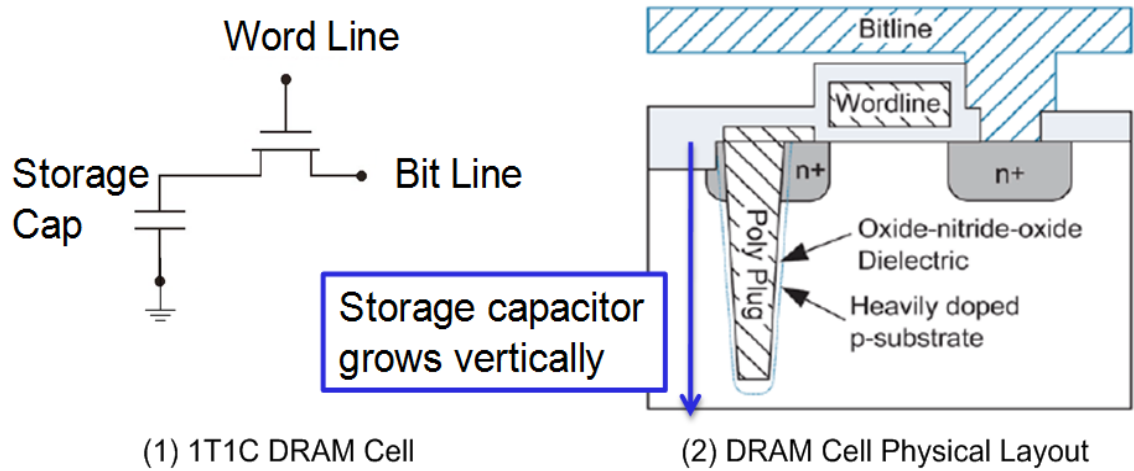


Figure 2.4: One Bit DRAM Cell Schematic and Physical Layout: reformatted from *CMOS VLSI Design*, Weste and Harris, Pearson Education, Inc., 2005. [2]

Table 2.1 summarizes the design comparison between the SRAM and DRAM.

Table 2.1: Comparison between SRAM and DRAM

Parameter	SRAM	DRAM
Data Rate	Fast	Nominal
Density	Nominal	Large
Cell Architecture	Complex	Simple
Cost	Nominal	High
Primary Disadvantage	Require large area	Refresh periodically

### 2.1.3 Differences between embedded DRAM and external DRAM

DRAM was first introduced in 1960s, and generally used as external memory devices. Embedded DRAM inside ASIC design has been difficult because DRAM physical layouts required additional and expensive physical layers for the trench capacitor within ASIC system layouts [2].

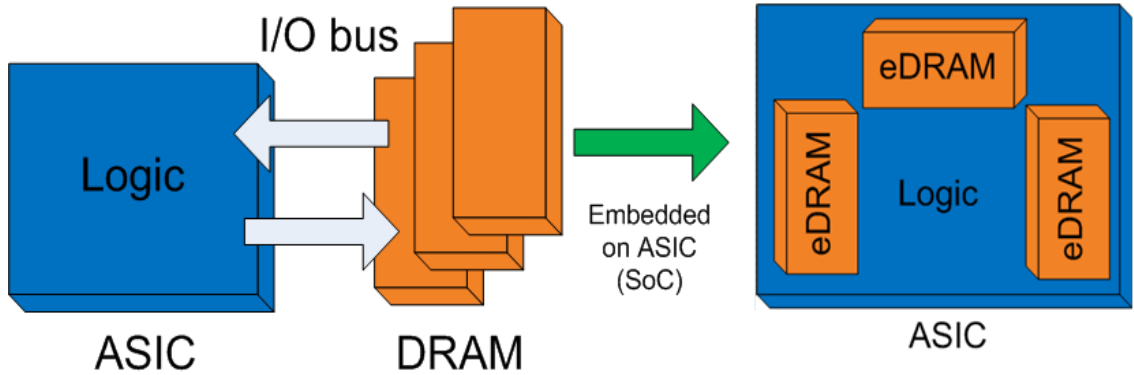


Figure 2.5: Change from External DRAM to Embedded DRAM

When DRAM is embedded, the external I/O bus to the memory can be eliminated, which reduces power consumption and noise (Figure 2.5). ASIC pin counts and Printed Circuit Board (PCB) layers can be reduced as well. Overall, embedding DRAM helps simplify and shrink the physical PCB size. The eDRAM increases ASIC design flexibility by merging logic system and memories by designing specific

memories for applications. Table 2.2 compares embedded DRAM with external DRAM.

Table 2.2: Comparison between External DRAM and Embedded DRAM

<b>Parameter</b>	<b>External DRAM</b>	<b>Embedded DRAM</b>
Data Rate	Nominal	Fast
Total Area	Nominal	Small
Noise	High	Low
Design Flexibility	Fixed	High
Cost	Nominal	High

#### **2.1.4 eDRAM Applications**

The largest eDRAM application is 3D graphics accelerator chips [3]. The graphic processor takes advantages of low power consumption and high data rate performance of eDRAMs for ASICs. Computer network routers are also a demanding application which requires large memory capacity for millions of entries and high data rate between logic and memory [4].

## **2.2 Operation and Retention Time Fails of eDRAM**

To analyze the retention time for DRAM cells, the operation of DRAM cells needs to be understood. This is because the test conditions are determined based on the sources of the retention time and test engineers need to detect memory bits whose retention time is shorter than use refresh time. In Section 2.2, the operation of the 1-Transistor 1-Capacitor (1T1C) DRAM cell and the sources of the retention time fails are described.

### 2.2.1 Operation of eDRAM: : 1T1C-DRAM Architecture

Figure 2.6 shows the schematic of 1T1C-DRAM architecture. DRAM cells are connected with a decoder and sense amplifier to conduct write and read operations. Similar to the SRAM cell, the data stored in the DRAM cell is accessed to write and read through an access transistor.

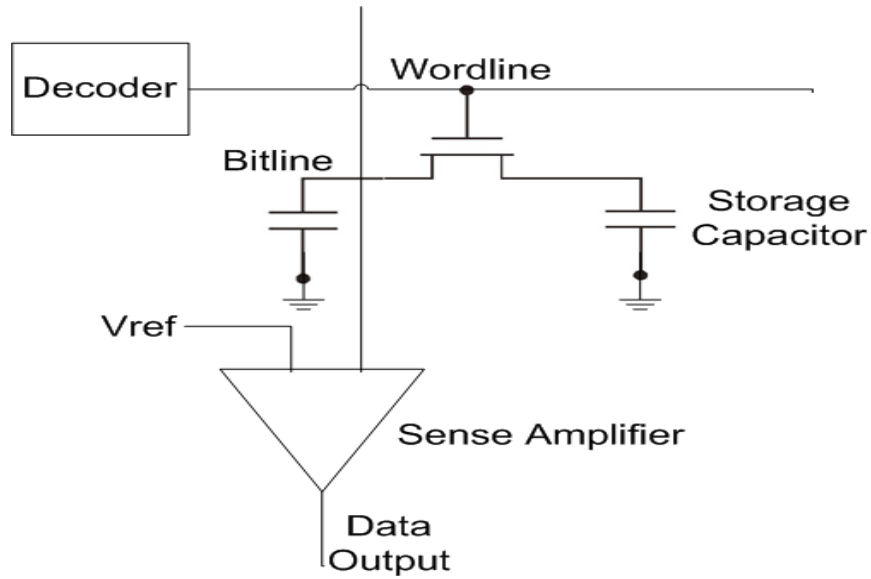


Figure 2.6: 1T1C-DRAM Cell Schematic

On a write operation, when the logic system needs to write a logic-high, the bitline is driven to VDD and the bitline charges the capacitor to VDD. Similarly, when the logic system wants to write a logic-low, the bitline voltage is forced to GND and the capacitor discharges through the access transistor until the cell retains a logic-low.

The read operation timing diagram is shown in Figure 2.7. First, the bitline is precharged at  $V_{DD}/2$ . Normally, digital data is stored as binary data. A logic-high is typically represented as the maximum voltage  $V_{DD}$ . A logic-low is typically represented as the minimum voltage  $GND$ . If the capacitor holds a logic-high, then the charge stored in the storage capacitor is distributed through the access transistor to the bitline. The bitline voltage settles down with a small voltage increase. On the other hand, if the capacitor holds a logic-low, the capacitor assumes charge from the bitline and the bitline voltage settles down with a small voltage decrease. To determine whether the storage capacitor has stored a logic-high or logic-low, sense amplifiers measure the bitline voltage change.

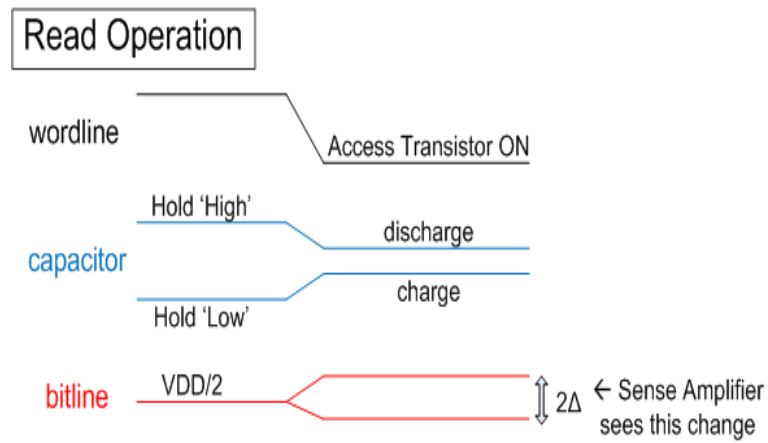


Figure 2.7: DRAM Cell Read Operation Timing Diagram

When the access transistor is off, the transistor does not transfer the read or write operation from the bit-line to the storage capacitor. The charge is isolated at the storage capacitor. As time goes by, the capacitor loses the charge due to the leakage current from insulating regions. The time until the 1T1C-DRAM cell loses the logic information is called *retention-time*. To avoid losing the logic information,

the memory cell is refreshed at every a few hundred microseconds. The time when the memory cells are refreshed is called *refresh-time*.

### 2.2.2 Retention Time Fails of eDRAM

The DRAM retention time characteristics are divided into multiple categories. In the same memory, DRAM bits have different retention time characteristics. DRAM bits with two or more possible retention times are called variable retention time (VRT) [7]. The DRAM cell retention time characteristics are classified as shown in Figure 2.8.

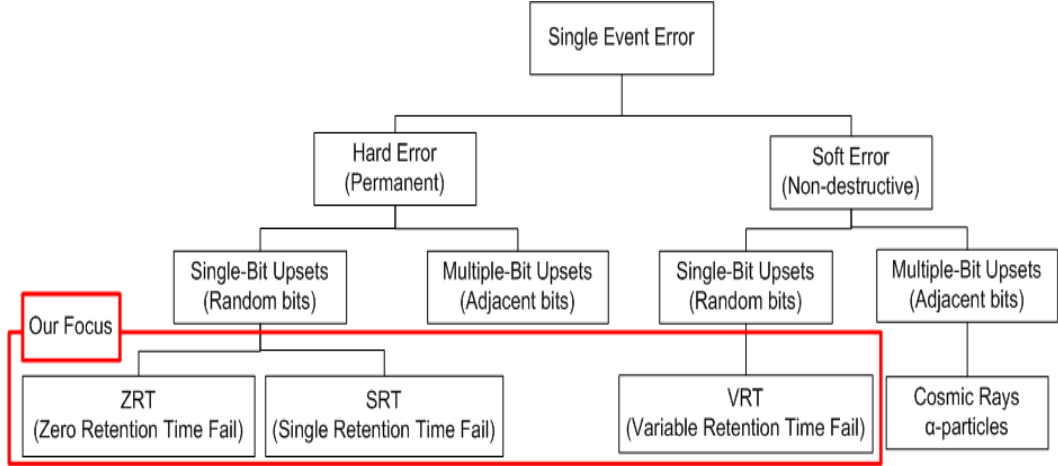


Figure 2.8: Categories of Single Event Error

DRAM errors can be categorized into *hard-errors* and *soft-errors*. Hard-errors corrupt bits permanently and the DRAM cells with hard-errors cannot be recovered (e.g., stuck-at faults). On the other hand, the DRAM cells with soft-errors randomly change the retention time at different memory refresh cycles. The randomness of the retention time is caused by defects [9]. The variable retention time could cause some bits fail before the refresh time.

Figure 2.9 classifies the three different types of retention time fail bits and for comparison, one pass bit is plotted. If a bit always fails at the initial test time, this bit is a hard-error and is called *zero-retention-time* (ZRT) fail bit because there is no retention time to hold any valid information (case (2) in Figure 2.9). The vast majority of memory cell retention times are constant from cycle to cycle. Such memory bits are named *single-retention-time* (SRT) bits. The SRT bit which has its retention time shorter than use refresh time are hard errors. Hard error SRT bits must be screened out at test (case (3) in Figure 2.9).

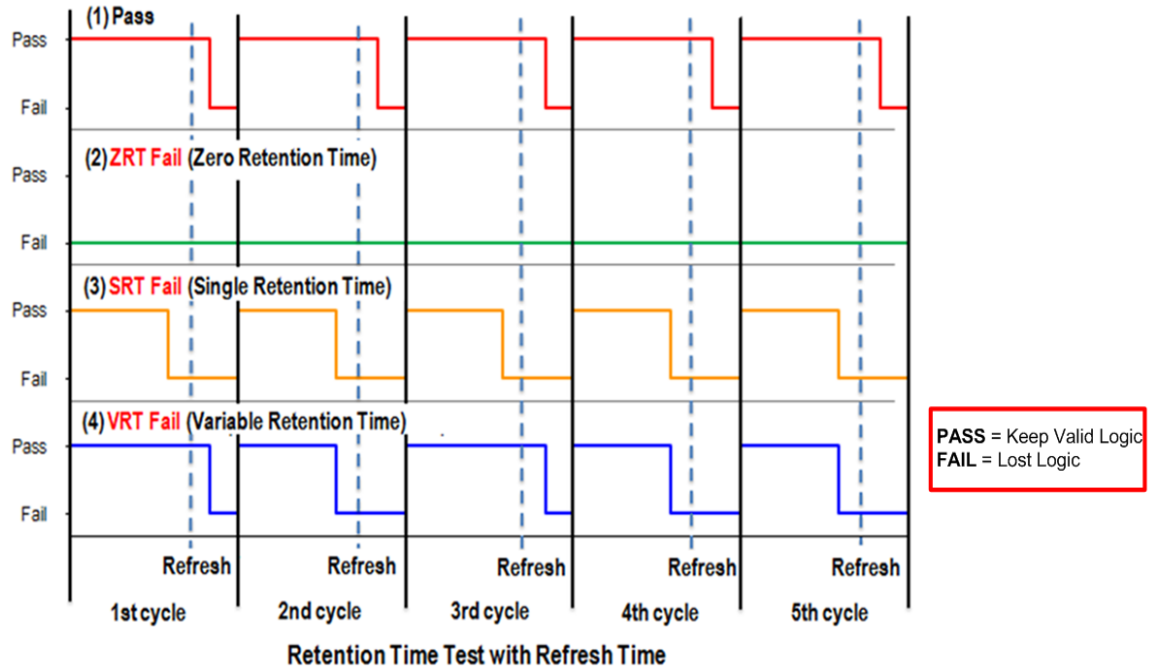


Figure 2.9: Fail Bit Characteristics with Time

In a very few cases, defects cause the retention time to vary from one refresh cycle to the next cycle. A bit with a variable retention time is called *variable-retention-time* (VRT) [7, 8]. As seen in Figure 2.9, bits in DRAM can display different retention time characteristics. The three main leakage sources that shorten retention time

are (1) reverse-bias  $pn$  junction leakage, (2) subthreshold leakage, and (3) direct tunneling current. Figure 2.10 describes where the leakage and defects occur in the DRAM cell. Defects in the DRAM cell can cause bits to have variable retention times [9].

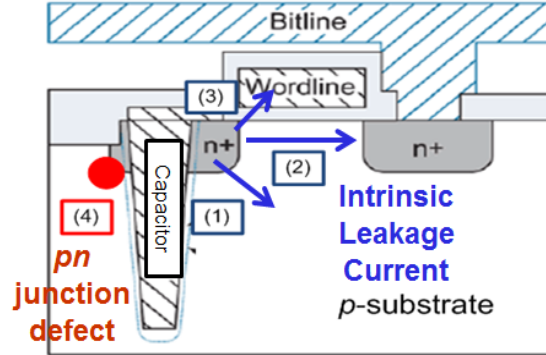


Figure 2.10: Sources of Retention Fails: reformatted from *CMOS VLSI Design*, Weste and Harris, Pearson Education, Inc., 2005. [2]

### 2.2.3 Repair and Use Error Detection of DRAM

After manufacturing and test are completed, good chips are sent to end-users. Typically, at test single retention time bits whose retention time is shorter than the use refresh time are replaced with new available bits or the entire chips are rejected before being shipped to end-users. However, some DRAM bits with variable retention time that passed at test might have shorter retention time in use than refresh time. Such bits need to be corrected with a software method like *Error Correcting Code* (ECC). ECC can correct many soft-errors by using information redundancy. Most memory systems in use today include error detection and correction codes. [10].

### 2.2.4 eDRAM Technology

For better performance, the ASIC technology follows Moore's Law to smaller feature size, lower power consumption, and faster data rate. Table 2.3 shows the current and future trends of 1T1C-eDRAM.

Table 2.3: 1T1C-eDRAM Future Trend [1]

Year of Production	2008	2009	2010	2013
eDRAM 1/2 pitch (nm)	90	65	65	45
1T1C bit cell size (fF)	12-30	12-30	12-30	12-30
Metal Mask Layer	3-5	3-5	3-5	3-6
Read Operating Voltage (V)	2	1.8	1.7	1.6

According to the International Technology Roadmap for Semiconductors [1], the 1T1C-DRAM trend is toward smaller cell size and lower read operating voltage. The smaller device size will influence the retention time for eDRAM because the smaller sized device will likely produce more intrinsic leakage current [1].

## 2.3 Quality Figures-of-Merit

Most semiconductor logic products are tested and shipped to end-users. In general, products must function under multiple temperature and voltage conditions at test and during use. Figure 2.11 describes the flow of product test and use after assembly for eDRAM product test including the quality figures-of-merit.

At test, chips are screened under test conditions. In this eDRAM case, test set points, such as test retention time and voltage and temperature conditions are the test conditions. All eDRAM bits are tested under test conditions. Generally, test conditions are more stringent than use conditions in order to screen out as

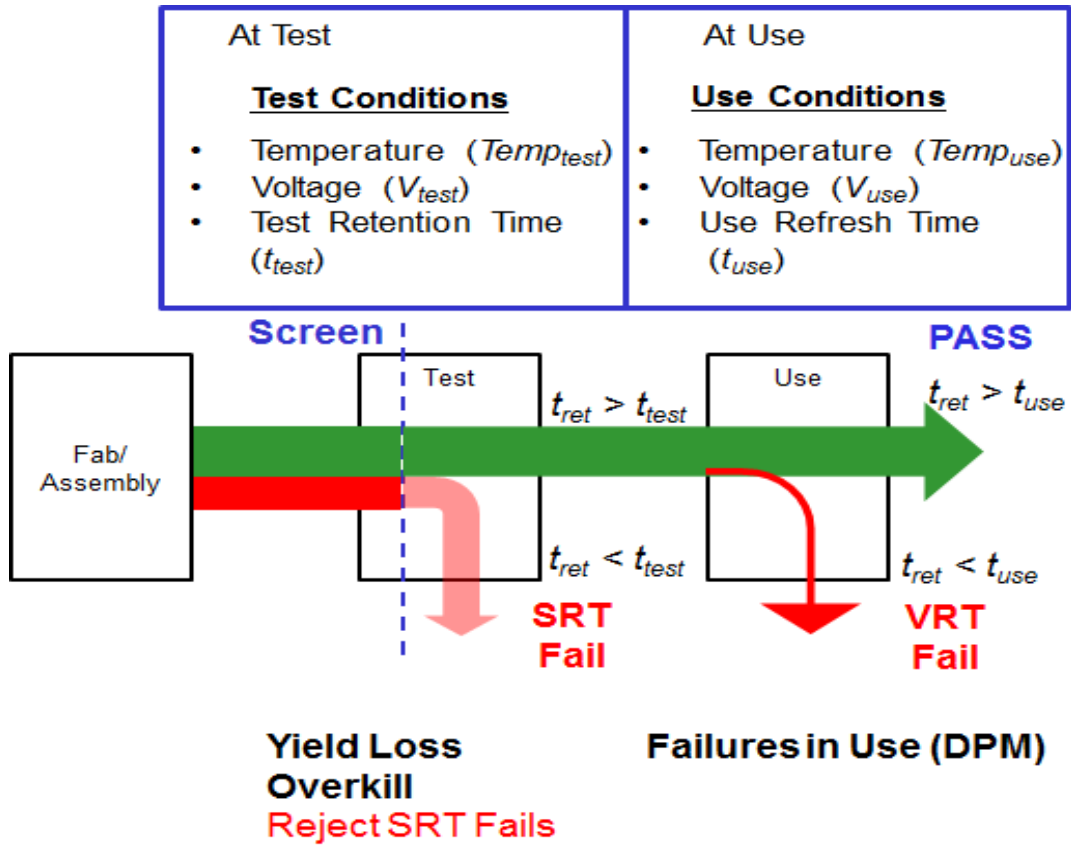


Figure 2.11: eDRAM Product Test and Use Flow with Figures-of-Merit Measured at Test and Use Conditions

many fail bits as possible. However, there are bits that would not fail under use conditions. The bits that would not retain the data during use but were rejected at more stringent test conditions are called *overkill*. At test, the memory yield is measured by the fail bits at test over a total number of tested bits.

For instance, memory bits with variable retention time are not easy to detect under one test condition because of the random retention time, so there is a chance of bits passing at test but failing in use. The bits that pass at test but fail in use are called test escape. The retention time fail bits found during use is a measure of the eDRAM quality. Especially, memory bits with variable retention time is the main focus of the eDRAM quality during use.

### **2.3.1 Retention Time Characteristics and Statistical Model**

In order to test the DRAM retention time as described in Section 2.2.2, the retention time of memory bits is measured for longer times than the use refresh time (e.g. reported on data specification sheets). The memory bits are tested at every few hundreds of microseconds of retention times to find out whether bits pass or fail, refresh bits, and test the bits at longer time again. Temperature and voltage conditions are known to be good accelerating factors for memory cell [11]. More stringent test conditions help observe retention time characteristics and determine effective test conditions for screening.

To display the results of the eDRAM quality based on test conditions and use conditions, two dimensional retention time scales are used to identify the quality level as shown in Figure 2.12. The horizontal axis is the use refresh time and the vertical axis is the test retention time. Red dots in the same figure indicate fail bits

with minimum retention time in the horizontal axis and maximum retention time in the vertical axis. The minimum and maximum retention times are applied to the use refresh time and test retention time, respectively because the worst case of the variable retention times of a bit is considered. The test retention time and use refresh time define the quality figures-of-merit. For example, if the minimum and maximum retention times are the same, the bits have only a single retention time and the bits are plotted on the 45-degree line between the horizontal and vertical axes. The other bits have variable retention times and the worst case of variable retention times (i.e., minimum and maximum retention time) are considered to plot the variable retention time bits in Figure 2.12.

The tested bits with shorter retention time than the test retention time are rejected and when possible the retention time fail bits are replaced with new available bits on the same chip. The rejected bits are located in Figure 2.12 labeled Rejected. The use refresh time defines the memory bits that pass in use after tested. For example, if the bits have a shorter retention time than use refresh time, the bits will fail during use. The fraction of fail bits during use is a measure of eDRAM quality. In Figure 2.12 labeled DPM at use. Since the test condition is more stringent than use condition, the test condition unnecessarily fails some bits that would meet the use condition in Figure 2.12 labeled "Overkill." It is important to minimize the overkill and preserve the yield while maintaining good quality level by selecting effective test conditions to minimize DPM at use.

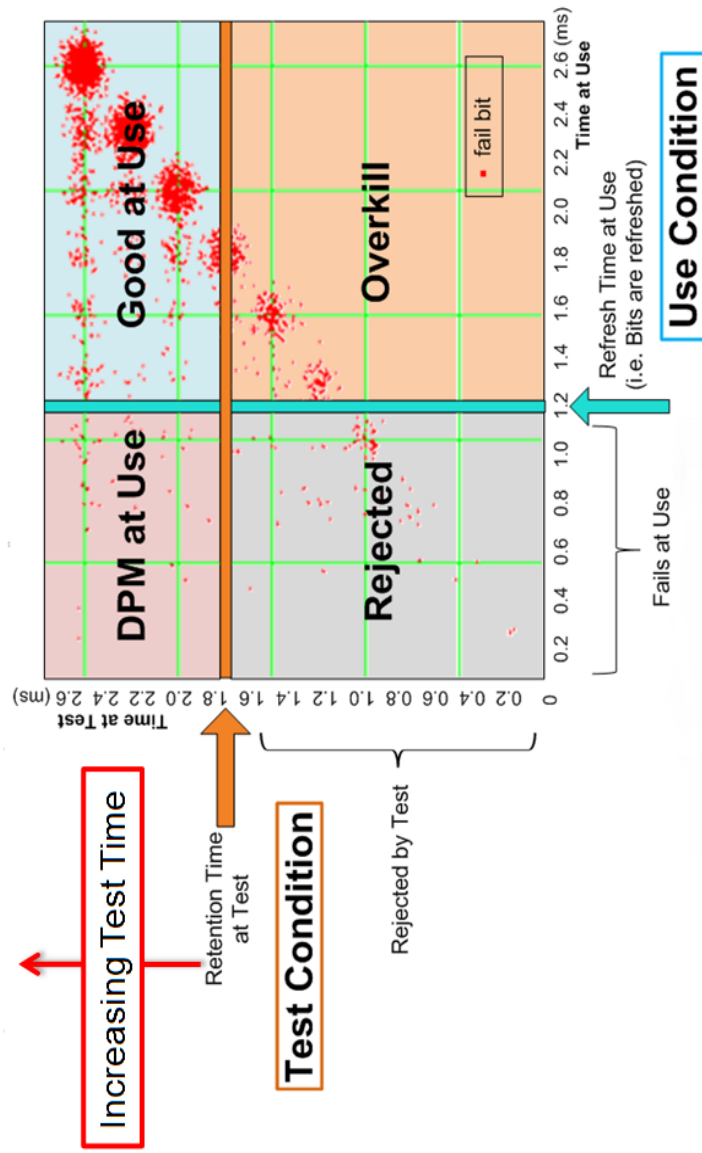


Figure 2.12: Quality of eDRAM Memory Bits at Use and Test Conditions

### 2.3.2 Weibull Distribution for Retention Time Characterization

To model the retention time for DRAM bits, a Weibull distribution is used. Lieneweg found that the Weibull distribution was a good way to characterize retention time distributions [14]. The strength of the Weibull distribution is its flexible shape as a model for many different kinds of data. By the adjustment of two distribution parameters (scaling parameter  $\alpha$  and shape parameters  $\beta$ ), the retention time model can be fitted to the data.

By modeling the retention time with the Weibull distribution, the fraction of remaining and surviving at time  $t$  is shown in Equation (2.4).

$$S(t) = \exp \left[ - \left( \frac{t}{\alpha} \right)^\beta \right] \quad (2.1)$$

In Chapter 3 and 4, the Weibull distribution applications are more described.

### 3 Experiments: Test Approach

The purpose of this research is to model the DRAM retention time for a specific technology. The eDRAM test chips need to be tested to characterize the retention times of eDRAM bits. Chapter 3 details the experimental methods and steps to obtain a statistical model of DRAM retention times.

#### 3.1 Device Under Test: eDRAM Test Chip

Device Under Tests (DUTs) were tested to observe the temperature and voltage characteristics of retention time fail bits in eDRAM. In subsections under Section 3.1, the test setup including the DUTs are described.

##### 3.1.1 Description of eDRAM Test Chip

Figure 3.1 shows a packaged eDRAM test chip. The eDRAM test chips from five distinct process skews were tested on a Credence Quartet One in Integrated Circuits Design and Test (ICDT) lab at Portland State University. The power supply voltage VDD and substrate bias VP were controlled by JTAG. Many testes were performed by an on-chip BIST. The test chip is packaged in a Fine-Pitch Ball Grid Array (fpBGA). The total pin count including IO, clock, power and unused pins is 896. The die size of the IC inside the package is 4mm  $\times$  3mm. In the test chips, *p*-channel is used as the access transistor. The advantage of using pMOS as the access transistor in the eDRAM cell is lower leakage.

There are four separate memory blocks inside each test chip shown in Figure 3.2. Figure 3.2 is the 65nm eDRAM test chip used during the experiments. Each memory block contains 1,218,750 DRAM test bits. Additional bits are available in

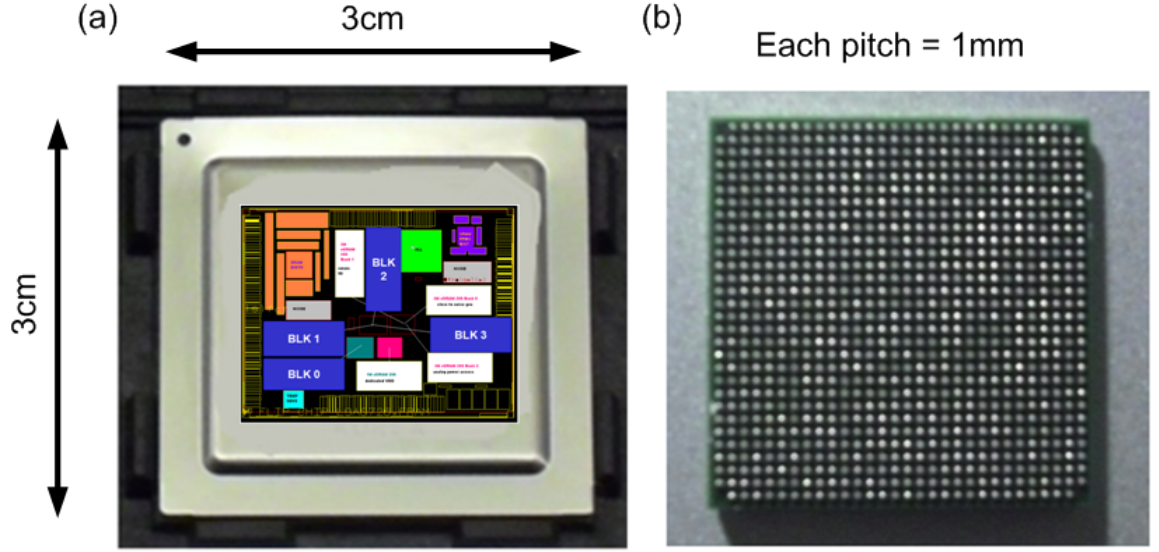


Figure 3.1: 65nm eDRAM Test Chip: a) is the top view and b) is the back side view. Floorplan is not to scale. The die size is  $4\text{mm} \times 3\text{mm}$ .

each memory block. An on-chip temperature sensor is located at the south west corner in Figure 3.2 to monitor its die temperature during the experiments.

### 3.1.2 Process Skews

To characterize retention times across the process variations such as oxide thickness and doping, five distinct process skews were provided. Table 3.1 compares the process differences and similarities among five different process skews. The fastest chips were not provided by a research sponsor because of too low yield. All chips are prescreened to screen out dead on arrival (DOA) bits before being sent to Portland State University.

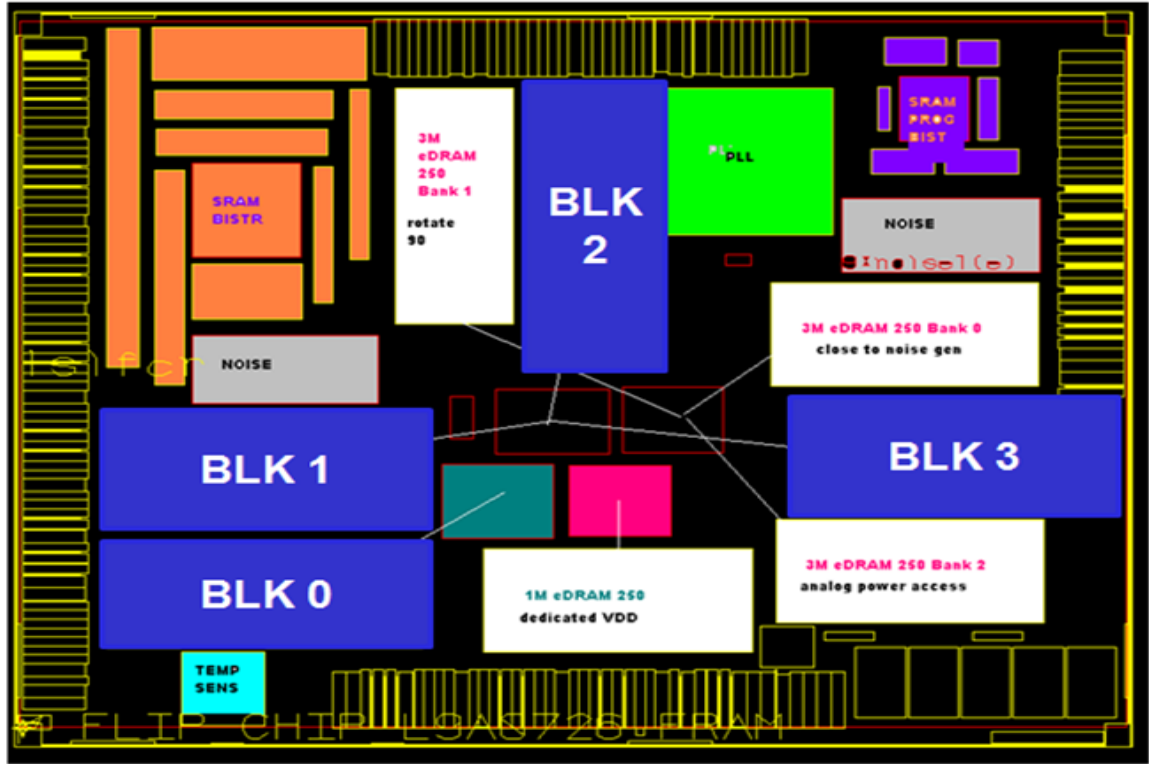


Figure 3.2: 65nm eDRAM Test Chip Floorplan

Table 3.1: eDRAM Test Chip Process Skews

Skew Name	Poly Dimension	N-type Doping	P-type Doping	$t_{ox}$
w4	typical	typical	typical	typical
w8	slow	slowest	slowest	low
w9	typical	fast	fast	typical
w10	typical	slow	fast	typical
w12	slow	slowest	slowest	low

### 3.2 Test Environment

All test and measurements were performed in Integrated Circuits Design and Test (ICDT) lab at Portland State University (Figure 3.3). The Credence Quartet One was utilized for the testing of the eDRAM chips. An external temperature controller keeps the package temperature of the eDRAM chips at constant temperature. The following subsections describe the test environment in more detail.



Figure 3.3: Test Environment in the ICDT Lab

Figure 3.4 is the simplified block diagram of the test environment during the experiments. The Quartet One can perform various types of required tests such as parametric and memory tests through software test programs controlled by the console which can also control the storage test data logs. For the voltage bias, the ATE forces minimum, nominal, and maximum voltages to the DUTs and takes parameter measurements. For temperature bias, the external temperature controller unit, LB300-i from Silicon Thermal controls the package temperature on

the DUTs, and the ATE can monitor the die temperature through the temperature sensor within the DUT and record the temperature results on log files for data analysis.

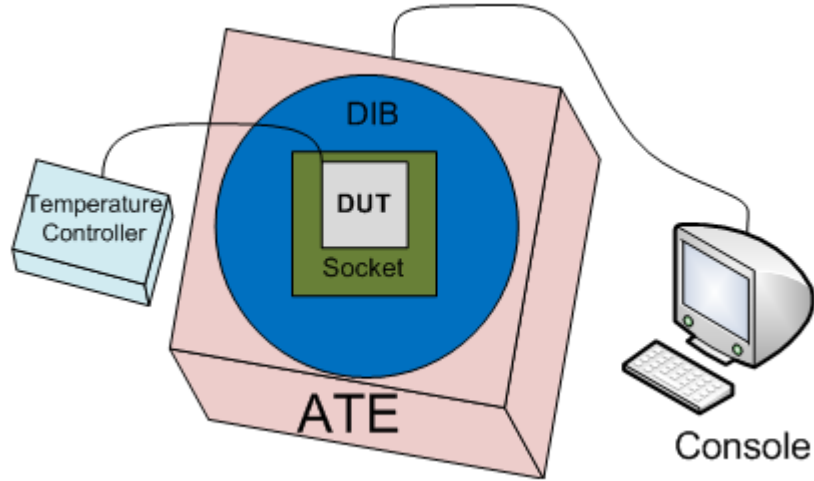


Figure 3.4: Test Environment Block Diagram

### 3.2.1 IC Tester (ATE)

Credence Quartet One in Figure 3.5 is a 200MHz and 512-I/O pin system. This ATE can perform digital and mixed-signal tests on ASICs, programmable logic, microprocessors, and etc. The driver and comparator voltage range is from -2.5V to +7.5V. The I/O bus speed is 200MHz. Each tester subsystem was calibrated before testing the DRAM test chips.

To make contact between the ATE and DUT, a Device Interface Board (DIB) is utilized as shown in Figure 3.6. All necessary pins on the DUT are mapped to the Tester {Pogo Pins} through the DIB. An IC socket is attached on the top of the DIB to accept one DUT package at a time. The external temperature controller

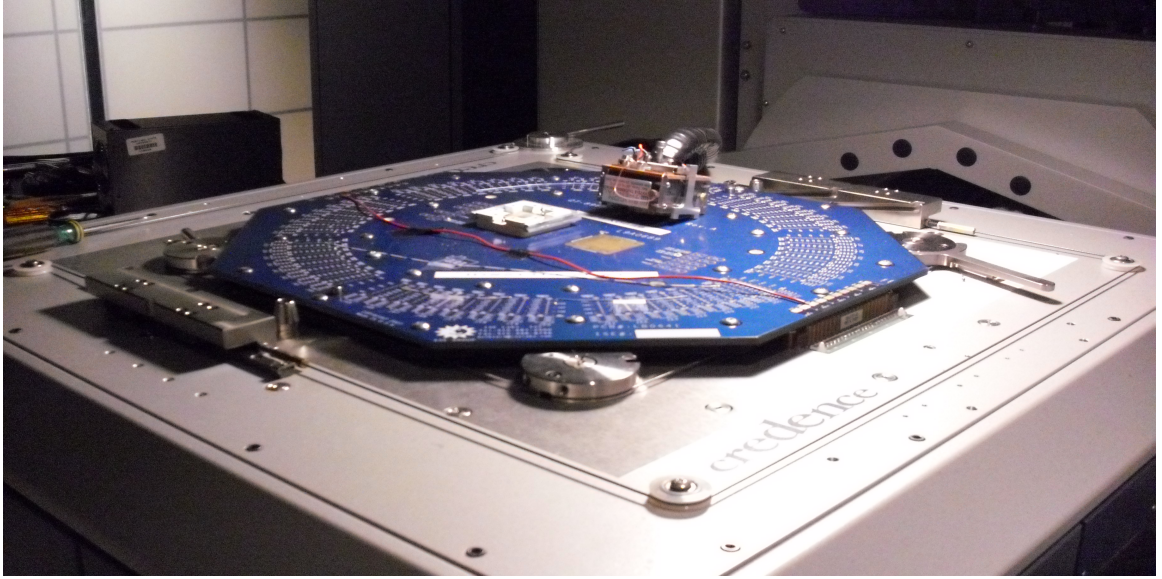


Figure 3.5: Credence Quartet One ATE

is used to maintain the package temperature of the DUT at a user-determined temperature during the experiments.

### 3.2.2 eDRAM Test Program

The test program is executed from the ATE console and the program logs the DUT response throughout the test execution. The log files are used to record where and when retention time fails and other fails occur. The data is extracted from the DUT.

During the testing of a DUT, all parametric tests, scan tests, and functional tests are performed first, and the die temperature is measured with the on-chip temperature sensor. The temperature value is recorded in binary by Analog-to-Digital Converter and converted to Celsius to monitor the constant temperature. The binary resolution is  $\pm 0.7^{\circ}\text{C}$ . Only when the DUT passes all the initial tests, the

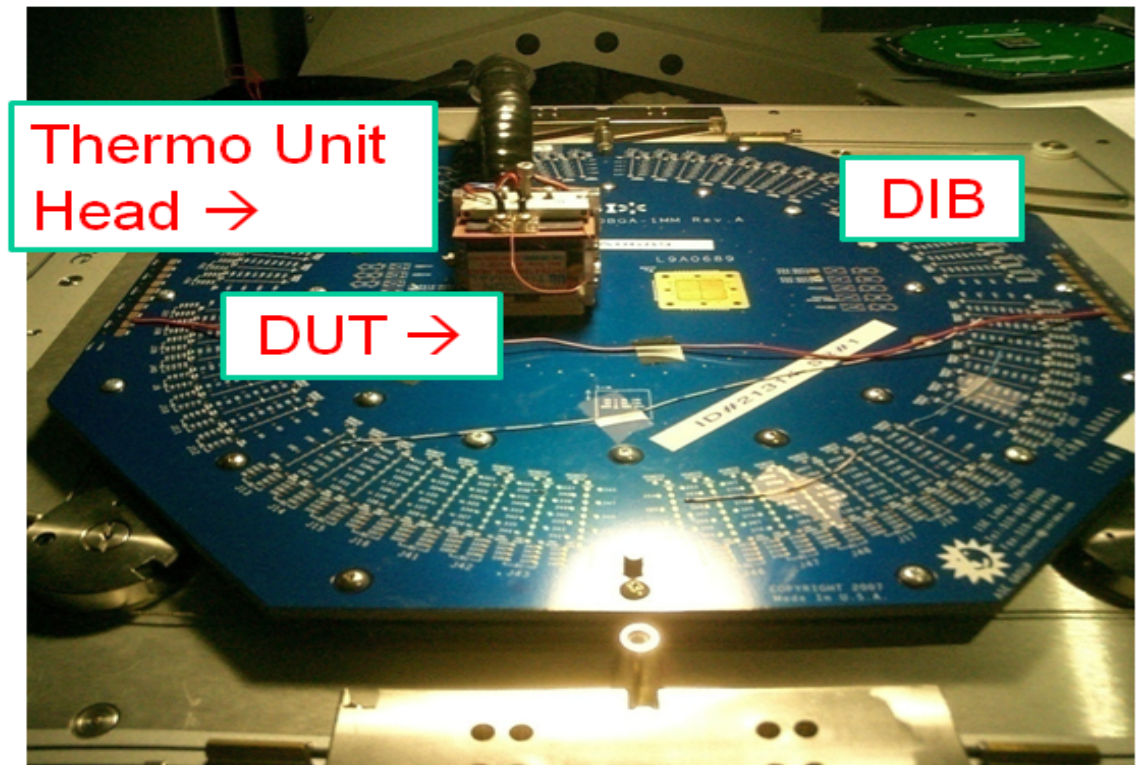


Figure 3.6: Test Environment in ICDT Lab at Portland State University

ATE begins to perform shmoo programs for retention time analysis.

A shmoo program helps identify fail characteristics by sweeping multiple test parameters. In the case of eDRAM testing, the shmoo creates plots to count the number of retention time fails with voltage and test retention time as the y-axis and x-axis, respectively. Figure 3.7 shows a two-dimension shmoo plot at one particular test condition. Note that the shmoo process for the DRAM test program sweeps the retention time from  $262\mu\text{s}$  to  $2659\mu\text{s}$  over three VDD voltage test conditions. During the shmoo, the VDD is also swept from 0.85V to 1.20V. The shmoo tables cannot give the logical address of the fail information. Note that the number of fail bit counts increases as the test retention time goes up.

		Block 1: VP = 0.40V, Test Count = 1									
		Fail Bit Counts per 1,218,750 bits									
VDD	1.20V	0	0	0	0	1	1	4	5	7	14
	1.00V	0	0	0	0	0	0	1	2	4	7
	0.85V	0	0	0	0	0	0	0	1	3	5
		262	480	916	1134	1352	1787	2005	2223	2441	2659
		Retention Time ( $\mu\text{s}$ )									

Figure 3.7: Shmoo Data Log: Sweep retention time from  $262\mu\text{s}$  to  $2659\mu\text{s}$  (x-axis) and also sweep VDD voltage from 0.85V to 1.20V (Y-axis)

Once the shmoo tests are completed, the test program searches on-chip registers for the logical addresses of retention time fails and it reports the fail bits with logical address as well as the voltages, memory location, and the cycle number (i.e., loop). The same test is conducted five times to distinguish single retention time (SRT) fails with variable retention time (VRT) fail bits. Figure 3.8 displays the shmoo test program pseudo code for eDRAM test chips.

```

Test Program Pseudo Code

Begin
Perform Parametric Tests
Measure Temperature Die
Create Shmoo Plots
Loop: Test Count (1 – 5)
    Loop: Block (0 – 3)
        Loop: VP (0.40V, 0.45V)
            Loop: VDD (0.85V, 1.00V, 1.20V)
                Loop: Ret Time (0.2μ - 2.6μs)
Measure Temperature Die
Save Data Log
End of Program

```

Figure 3.8: Shmoo Test Program Pseudo Code for eDRAM Test Chips

For each DUT, the test program is executed at three temperatures (105°C, 115°C, and 125°C), three VDD voltages (0.85V, 1.00V, and 1.20V), and two VP voltages (0.40V and 0.45V). Furthermore, there are ten sample chips from each five skews, a total of 50 chips. A total number of test runs for the experiments is 150. At 105°, the test time per chip varies from one hour to several hours. At 125°, the test time ranges between six hours and 12 hours depending on the selected DUT process skew because of larger retention time samples. Note that in each test, the temperature is measured at the beginning and end to insure the die temperature variance as small as possible ( $\pm 1$ bit difference).

### 3.3 Data Collection Process

During test, ASCII log files generated from the tester are primitive and very difficult to use. The log files need to be converted to the files that are easy to analyze. Figure 3.9 indicates the flow of the data conversion.

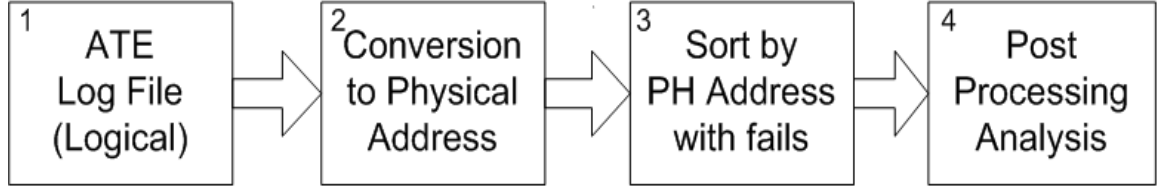


Figure 3.9: Data Conversion Process: Note only the fail bits found during the test are recorded in data logs.

The log files from the ATE only contain the information about logical addresses with fails. In step 2, each log file is parsed in order to map the logical addresses into the physical address (x, y). In step 3, the file with physical addresses are reorganized in test conditions and retention time stops. In step 4, the collected fail bits across all process skews are placed into a single file to analyze retention times of the bits all at once.

The characteristics of retention time are identified by the method described in Section 3.4, and the trends and results of retention time fail bits are discussed in Chapter 4.

### 3.4 Modeling eDRAM Bit Fails

The collected retention time fail bits are categorized into the three different types: Zero Retention Time (ZRT), Single Retention Time (SRT) and Variable Retention Time (VRT) bits by physical address, test condition, and retention time. Figure

3.10 describes the differences between pass and fail bits in binary representations along with time. The “0” means memory bits hold the valid logic and “1” means bits lose the data.

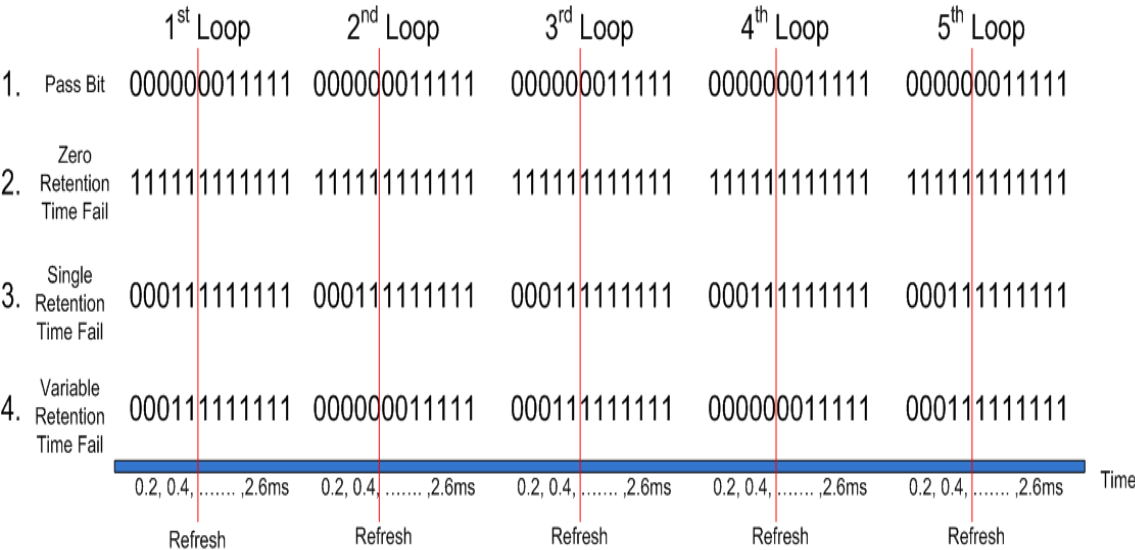


Figure 3.10: Pass and Fail Bit Characteristics in Time Domain

The pass bits (e.g., labeled 1 in Figure 3.10) can hold the valid logic information till the bits are refreshed. However, there are bits which fail before refresh time (e.g., labeled 2, 3, and 4 in Figure 3.10). Zero Retention Time (ZRT) bits fail at all times. Single Retention Time (SRT) fail bits pass for small amount of time and eventually fail at the same time in every cycle (loop) before the refresh time. Variable retention time fail bits fail one or more retention time tests. That is, the VRT bits pass in some cycles, but they fail in the other cycles. Collecting the retention time data for many samples and loops can help identify the trends of fail bits at different test conditions and statistically model the failure rate of memory bits.

### 3.4.1 Model Extraction for Single Retention Time Bits

The first step in modeling the DRAM retention time is to develop a model for a single bit retention time. For a single bit, the Weibull distribution is used to describe the probability of a DRAM bit has a retention time  $t_{ret}$  exceeding  $t$  is shown in Equation 3.1 (e.g.,  $S(t)$ ).  $S(t)$  is the fraction of a bit has a retention time larger than time  $t$ .

$$S(t) = 1 - F(t) = P(t_{ret} \geq t | S, T, V) = \exp \left[ - \left( \frac{t}{\alpha(S, T, V)} \right)^\beta \right] \quad (3.1)$$

$S$ ,  $T$ ,  $V$  in the Equation 3.1 denote the process and experimental conditions of the eDRAM test chips:  $S$  is a process variation (skew),  $T$  represents the DUT temperature,  $V$  stands for both the bias voltage and supply voltage (i.e., VP and VDD).  $t_{ret}$  is the retention time of a memory bit.  $\alpha$  and  $\beta$  are the Weibull scale and shape parameters, respectively.

To fit the retention time to the Weibull distribution, Equation 3.1 is converted Weibull scale. The model assumes that each fail bit has a single retention time.

$$Weibit = \ln(-\ln(1 - F)) = \beta \ln(t) - \beta \ln(\alpha(S, T, V)) \quad (3.2)$$

Weibull shape parameter  $\beta$  is the slope in Equation 3.2 and the y-intercept is  $-\beta \ln(\alpha)$ .  $F$  is the cumulative fraction. Figure 3.11 displays one example of the Weibull single retention time model along with the observed data. By adjusting the fit lines in Figure 3.11 to bound all the observed data, the single retention time

is modeled conservatively to slightly overestimate the Weibit. The term, “conservative” that means the fit is chosen so that the model bounds all the observed data and the conservative model slightly overestimates the observed data as opposed to best-fit model which averages the observed data. The model is fitted for all three temperatures in Figure 3.11. After the fit model is conservatively targeted for each process skew, the shape parameter is chosen.

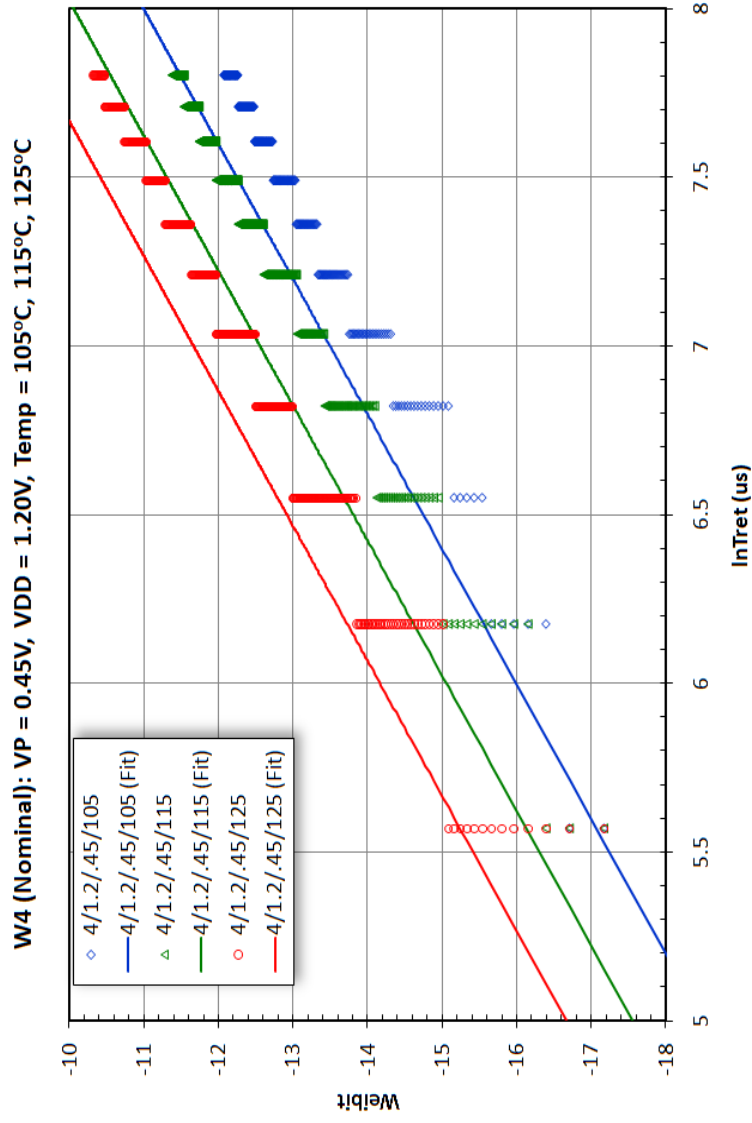


Figure 3.11: Single Retention Time Modeling: Note that y-axis is the *Weibit* and x-axis is the  $\ln(t_{ret})$  of Equation 3.3.

The shape parameter  $\beta$  is determined so that the Weibull model is conservative. The temperature and voltage dependence to the retention time fails is characterized. By fitting the *y-intercept* to the Weibull model, the dependence of the temperature, VP, and VDD on the single retention time fail bits is extracted. Equation 3.3 includes the coefficients of the experimental parameters with one offset to extract the experimental parameters and validating the conservative model.

$$\ln(\alpha) = a + b \cdot VP + c \cdot VDD + \frac{Q}{k_B \cdot T} \quad (3.3)$$

The  $a$ ,  $b$ ,  $c$ , and  $Q$  are the fitting coefficients for the voltage and temperature dependence to the experimental data. In fact, the coefficient,  $a$  is helpful to make adjustments by moving the fit model horizontally on the y-axis to cover the observed data. Figure 3.12 describes the experimental parameters dependence on the fail bit counts with a nominal process. From the parameter extraction,  $a$ ,  $b$ ,  $c$ , and  $Q$  are determined for each skew.

The single bit retention time fail model needs to be expanded to an array of memory bits to analyze the quality of memory blocks. For an array of size  $N$  bits (i.e., a single memory block), the probability that every independent bit in the array has retention time less than  $t$  can be written as a survival function.

$$S_{array}(t) = [S_{bit}(t)]^N = \exp \left[ -N \left( \frac{t}{\alpha(S, T, V)} \right)^\beta \right] \quad (3.4)$$

In summary, with the Weibull shape and scale parameters chosen, the single retention time failure rate is characterized and the model can be expanded to the user-determined number of memory bits in an array because each bit is independent. The quality for SRT fail bits is determined by identifying the figures-of-merit

such as DPM in use, overkill and yield loss at test in Section 3.4.2.

### 3.4.2 Test and Use Condition Model

Test conditions and use conditions are important to characterize the retention time of memory bits due to the sensitivities to the experimental conditions. Usually, the test condition is more stringent than the use condition because all possible fail bits are preferred to be rejected at the test condition. Table 3.2 lists the experimental test and use condition parameters such as the time, voltage, and temperature.

Table 3.2: Test and Use Condition Experimental Parameters

Test Condition	Use Condition
$t_{test}$	$t_{use}$
$VDD_{test}$	$VDD_{use}$
$VP_{test}$	$VP_{use}$
$T_{test}$	$T_{use}$

Note that the  $t_{use}$  means the refresh time in use. Yield loss is evaluated at test. The number of memory bits with shorter retention time than the test retention time are rejected and the fraction failing at test is the yield loss in Equation 3.5.

$$YieldLoss(t_{test}|S, T, V) = F(t_{ret} < t_{test}|S, T_{test}, V_{test}) \quad (3.5)$$

Overkill is estimated under the use condition after test. The number of bits whose retention time is shorter than test retention time but longer than the use retention time is overkill. Equation 3.6 shows the expression for the overkill.

$$Overkill(t_{use}, t_{test}|S, T, V) = F(t_{ret} < t_{test}|S, T_{test}, V_{test}) - F(t_{ret} < t_{use}|S, T_{use}, V_{use}) \quad (3.6)$$

To measure the quality at the use condition, fraction fail in use for a memory block is calculated as shown in Equation 3.7.  $V_{test}$  and  $V_{use}$  include  $VDD$  and  $VP$  voltages.

$$EndUse(t_{use}, t_{test}|S, T, V) = \frac{F(t_{ret} > t_{test}|S, T_{test}, V_{test}) - F(t_{ret} > t_{use}|S, T_{use}, V_{use})}{F(t_{ret} > t_{test}|S, T_{test}, V_{test})} \quad (3.7)$$

The environmental parameters such as temperature,  $VP$ , and  $VDD$  show dependence on the fail bit prediction model.

### 3.4.3 Fail Bit Rate Prediction For Variable Retention Time Bits

Unlike Single Retention Time (SRT) bits, Variable Retention Time (VRT) bits have at least two distinct retention times over multiple clock cycles. The VRT bits with short retention times can be rejected at the test condition; however some VRT bits might pass at test and might fail later at the use condition because of the randomness of the variable retention times. For the VRT modeling, assuming test retention time as maximum retention time and use refresh time as minimum retention time, a two-dimensional retention time analysis is utilized. During the shmoo data collections, each test is looped five times. The maximum retention time and minimum retention time of the VRT bits are recorded as the best and

worst case retention times. The objective to identify the VRT quality is to find the cumulative VRT fail bit counts and fraction failure at each recorded retention time set point during the experiment mentioned in Chapter 3. The steps to obtain the cumulative VRT fail bit counts and fraction failure are described in Section 3.4.3.

Figure 3.13 is the combined table of the use and test conditions. Other data under different conditions are located in the ICDT02/edram server in ICDT Lab. In this case, the test retention time is set at  $1400\mu\text{s}$  and the use refresh time is set at  $300\mu\text{s}$  assuming the test condition is more stringent than the use condition. By using the table, the overkill is the proportion of the VRT bits which are rejected at the test but are supposed to pass at the use condition. The fraction fail in use is estimated to be the proportion of bits that are not rejected at the test but fail at the use condition. The yield loss is the proportion of VRT bits that are rejected by the test condition over the number of tested bits. The limit for this VRT analysis is the figures-of-merit are only estimated by the observed data up to 2.6ms. However, the test retention time and use refresh time typically occur at times less than the 2.6ms. Hence, the test and use time-scale analysis gives an insight of VRT behaviors. Here, the retention time bin is  $218\mu\text{s}$  because of the tester time resolution.

Figure 3.14 is a example of a two-dimensional table that shows the VRT row count over 40M bit samples with the maximum (*y-axis*) and minimum (*x-axis*) retention times at one particular test condition. The data is obtained from the experiments in the ICDT lab at PSU.

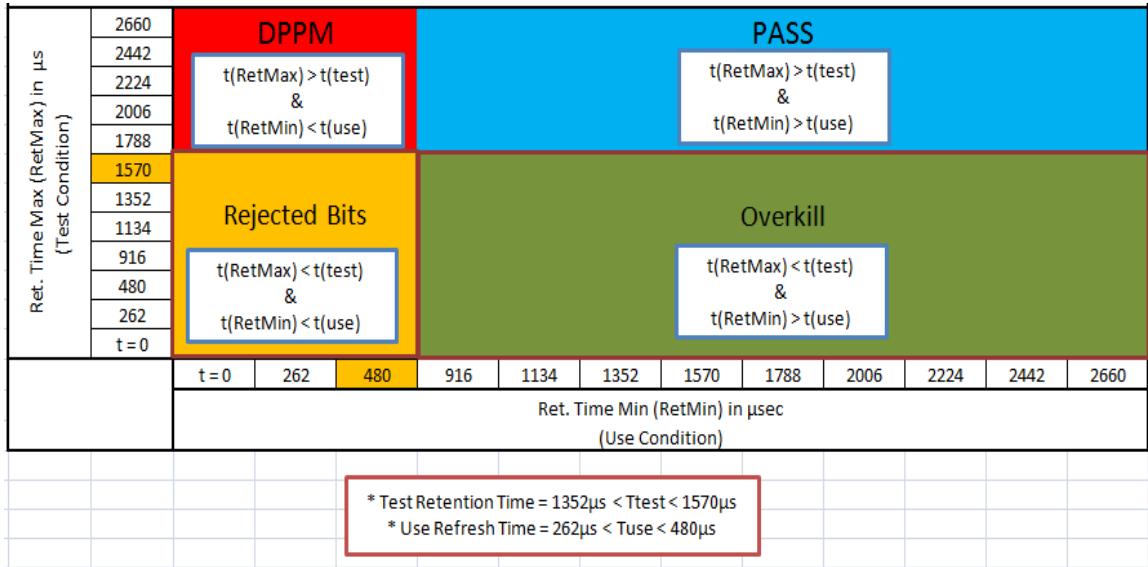


Figure 3.12: Two-Dimensional Time Domain ( $t_{test}$  and  $t_{use}$ ) to analyze DPPM, Pass, Overkill, Yield Loss: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins. (See also Figure 2.17 for the eDRAM quality figures-of-merit.)

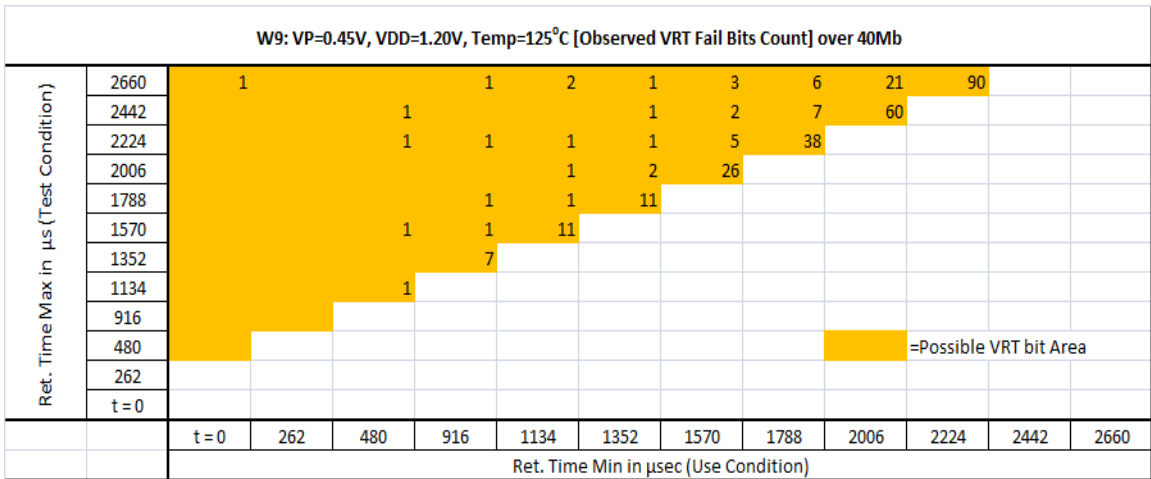


Figure 3.13: VRT Two Retention Times Model for Row Counts: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins.

Since the samples of the VRT bits are not large when compared to the SRT bits, the two-dimensional VRT table is smoothed with the estimated number of VRT bits in the empty cells in Figure 3.15. For example, if a cell is empty during the experiments, an estimate is added in the VRT table for the better modeling.

W9: VP=0.45V, VDD=1.20V, Temp=125C [Smoothed Count] over 40Mb													
Ret. Time Max in $\mu$ s (Test Condition)	2660	1	0	0	1	2	1	3	6	21	90		
	2442	0	0	1	1	1	1	2	7	60			
	2224	0	0	1	1	1	1	5	38				
	2006	0	0	1	1	1	2	26					
	1788	0	0	1	1	1	11						
	1570	0	0	1	1	11							
	1352	0	0	1	7								
	1134	0	0	1									
	916	0	0										
	480	0											
	262												
	t = 0												
<div>* Test Retention Time = 1352<math>\mu</math>s &lt; Ttest &lt; 1570<math>\mu</math>s * Use Refresh Time = 262<math>\mu</math>s &lt; Tuse &lt; 480<math>\mu</math>s</div>													
		t = 0	262	480	916	1134	1352	1570	1788	2006	2224	2442	2660
Ret. Time Min in $\mu$ sec (Use Condition)													

Figure 3.14: VRT Two Retention Times Model for Smoothen Counts: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins.

The cumulative count is summed by each test condition and use condition. Figure 3.16 is one example of cumulative counts of VRT fails enclosed by the test condition and use condition. Suppose the test retention time is  $1600\mu s$  and refresh time is  $500\mu s$ . The cumulative retention time fail count is six bits over 40mega bits.

Next, the cumulative count is normalized by the number of test bits (i.e., approximately 48Mbits) and it is expressed in the fraction of fails. At the same time, the same procedures are conducted at the use condition. The fraction at test provides the yield. Using Equation 3.8, the fraction of the variable retention fail bits in use is calculated given the yield at test. The denominator of Equation 3.8 is the yield

W9: VP=0.45V, VDD=1.20V, Temp=125C [Cumulative Count] over 40Mb													
Ret. Time Max in $\mu$ s (Test Condition)	2660	1	1	1	2	4	5	8	14	35	125		
	2442	1	1	2	4	7	9	14	27	108			
	2224	1	1	3	6	10	13	23	74				
	2006	1	1	4	8	13	18	54					
	1788	1	1	4	10	16	32						
	1570	1	1	6	12	29			= Cumulative Fails based on Test and Use Condition				
	1352	1	1	7	20								
	1134	1	1	8					=Possible VRT bit Area (cumulative)				
	916	1	1										
	480	1											
	262												
t = 0													
		t = 0	262	480	916	1134	1352	1570	1788	2006	2224	2442	2660
Ret. Time Min in $\mu$ sec (Use Condition)													

Figure 3.15: VRT Two Retention Times Model for Cumulative Counts: Note that the y-axis shows the maximum retention time bins and the x-axis shows the minimum retention time bins.

at test.

$$EndUse(t_{use}, t_{test} | S, T, V) = 1 - \frac{F(t_{RetMin} > t_{use} | S, T_{use}, V_{use})}{F(t_{RetMax} > t_{test} | S, T_{test}, V_{test})} \quad (3.8)$$

Figure 3.17 is an example of fraction failure at each retention time stop at test and use conditions. By using the two tables from test and use conditions, the fraction failure in use is estimated with Equation 3.8.

Now, the fraction of VRT fail bits in use given the test condition is estimated and the fraction failure set by the test retention time and use refresh time determines the eDRAM quality level.

W9: VP=0.45V, VDD=1.20V, Temp=125C (Test) -> VP=0.40V, VDD=1.00V, Temp=105C (Use) [Fraction]												
Ret. Time Max in $\mu$ s (Test Condition)	2660	2E-08	2E-08	4E-08	8E-08	1E-07	2E-07	3E-07	6E-07	2E-06	0E+00	
	2442	2E-08	2E-08	6E-08	1E-07	2E-07	3E-07	5E-07	2E-06	0E+00		
	2224	2E-08	2E-08	8E-08	2E-07	3E-07	4E-07	1E-06	0E+00			
	2006	2E-08	2E-08	8E-08	2E-07	3E-07	7E-07	0E+00				
	1788	2E-08	2E-08	1E-07	2E-07	6E-07	0E+00					
	1570	2E-08	2E-08	1E-07	4E-07	0E+00						
	1352	2E-08	2E-08	2E-07	0E+00							
	1134	2E-08	2E-08	0E+00								
	916	2E-08	0E+00									
	480	0E+00										
* Test Retention Time = 1352 $\mu$ s < Ttest < 1570 $\mu$ s * Use Refresh Time = 262 $\mu$ s < Tuse < 480 $\mu$ s												
	t = 0	262	480	916	1134	1352	1570	1788	2006	2224	2442	2660
Ret. Time Min in $\mu$ sec (Use Condition)												

Figure 3.16: VRT Two Retention Times Model for Fraction Failure at each retention time stop: Note that y-axis is the maximum retention time bins and x-axis is the minimum retention time bins.

## 4 Results: Parameter Extraction and Quality Models

### 4.1 Quantitative Results to Derive Quality Models

In Section 4.1, the trends of retention time fails based on the experimental parameters are analyzed. Table 4.1 summarizes the experimental and process parameters during the test denoted in Chapter 3. It is important to understand how experimental conditions influence the retention time of eDRAM bits. For example, the magnitude of temperature and voltage varies the retention time or not. Also, the location of the memory blocks inside a test chip influences the retention time fail bit counts or not. According to the retention time data shown in Section 4.1.1., the four memory blocks were all considered equivalent. The memory blocks were combined as one entity.

Table 4.1: eDRAM Experimental Conditions: 90 distinct skew, voltage, and temperature conditions.

Parameter	The Number of Values in Each Parameter
Process Skew	<b>5</b> skews (slow, nominal, fast, slowest, slowest & +20% $C_{store}$ )
Temperature	<b>3</b> temperatures (105°C, 115°C, 125°C)
Voltage (VDD)	<b>3</b> voltages (0.85V, 1.00V, 1.20V)
Voltage (VP)	<b>2</b> voltages (0.40V, 0.45V)

#### 4.1.1 Test Conditions: Process, Block, Temperature, Voltage

##### Memory Block Location

First of all, the four memory block locations in Figure 3.2 are analyzed. If there is no significant difference in retention time across the different memory locations, the four blocks can be grouped into one common set. The test condition used to

identify the relationship among the blocks is at the highest temperature and highest voltages (i.e.,  $T = 125^\circ$ ,  $VP = 0.45V$ , and  $VDD = 1.20V$ ) where the condition is known to produce the largest number of retention time fail bits. Figure 4.1 shows the number of retention time fail bit counts normalized by one million bits (i.e., DPM).

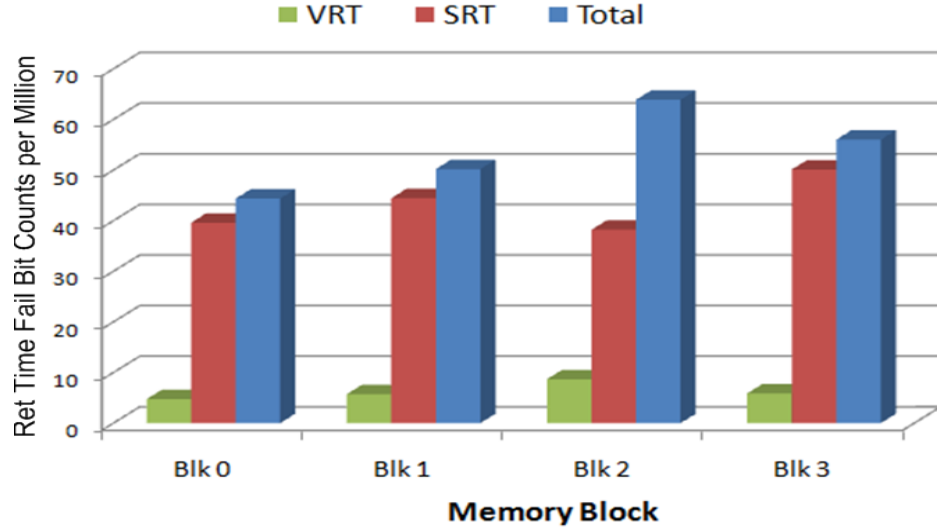


Figure 4.1: Fail Bit Counts per One Million Bits Sorted by Memory Block at  $125^\circ C$ ,  $VP = 0.45V$ , and  $VDD = 1.20V$

In Figure 4.1, VRT (green) is a variable retention time fail. SRT (red) is a single retention time fail. Total (blue) means the sum of the VRT and SRT. The data analysis shows that statistically there is no significant retention time difference across the four blocks. Hence, the four blocks are considered equivalent and grouped in one entity. In other words, the test chip has 4.8Mbits instead of separate 1.2Mbits from each block, and in each process skew, there are ten sample chips. In summary, for each process skew, 48Mbits are used for the parameter extraction and retention time statistical modeling.

### Process Skew

Figure 4.2 is the plot of the retention time fail bit counts normalized by one million bits against the five distinct process skews. The vertical axis is again retention time fail bits counts measured at 2.6ms similar to Figure 4.1. The horizontal axis is changed to the five available process skews.

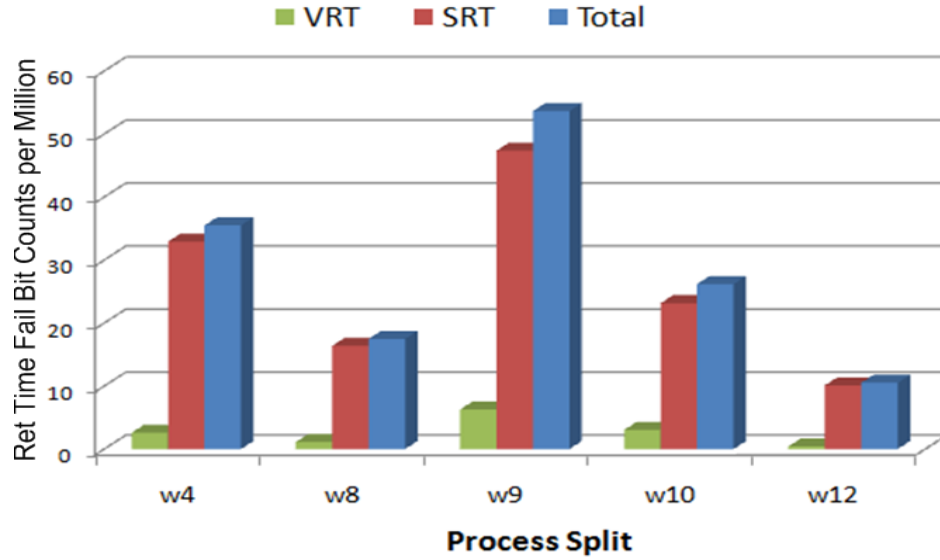


Figure 4.2: Fail Bit Counts per One Million Bits Sorted by Process Skew at 125°C, VP = 0.45V, and VDD = 1.20V

Note that w9 has the highest fail bit counts in SRT and VRT fails. The highest fail bit counts in w9 can be explained by the higher leakage current. The w9 is a fast access transistor and to achieve fast data rate, there would be more leakage current and this leakage current shortens the retention time and causes more fail bits. On the other hand, w8 and w12 are slow devices and especially w12 has more capacitance than any other process skew. The bigger capacitor can hold the logic information for longer time, so the bigger capacitor lowers the number of fail bit counts at any condition. The w4 (nominal) has less than 100 fail bit

counts per million bits at the higher temperature and voltages than the datasheet specification. In general, if the fail bit counts per one million bits is less than 100, it's considered as good products. Even at higher temperature and voltages with fast chips than nominal use condition (i.e.,  $T=105^{\circ}\text{C}$ ,  $VP = 0.40\text{V}$ , and  $VDD = 1.00\text{V}$ ), the fail bit counts per one million bits is at most 60. For this reason, the 65nm eDRAM test chips have an excellent quality.

### Temperature

Figure 4.3 is a plot of the retention time fail counts normalize by one million bits against the test temperature at  $105^{\circ}\text{C}$ ,  $115^{\circ}\text{C}$ , and  $125^{\circ}\text{C}$ . Higher temperature is known to generate more fail bits. Figure 4.3 indicates the retention time fail bit count increases exponentially. The data supports the temperature dependance on the retention time by the Arrhenius equation (i.e.,  $\exp(-\frac{Q}{kT})$ ). The  $Q$  in the Arrhenius equation is the activation energy (eV), the  $k$  is the Boltzmann's constant ( $8.617 \times 10^{-5}\text{eV/K}$ ), and the  $T$  is the die temperature. For example, as seen in Table 4.2, the  $Q$  for w4 skew is estimated to be 0.475 (eV). Hence, temperature is considered as one of the parameters that impacts the retention time fail bit counts.

### VDD (Bitline Voltage)

The supply and bias voltages are also known to be a parameter that influences the number of fails [11]. There are two DC voltages in each 1T1C DRAM memory cell: VDD and VP. Figure 4.4 is the VDD plot. As the VDD voltage increases, the fail bit counts per one million bits also increase. Both the VRT and SRT fail bit counts increase with increasing VDD.

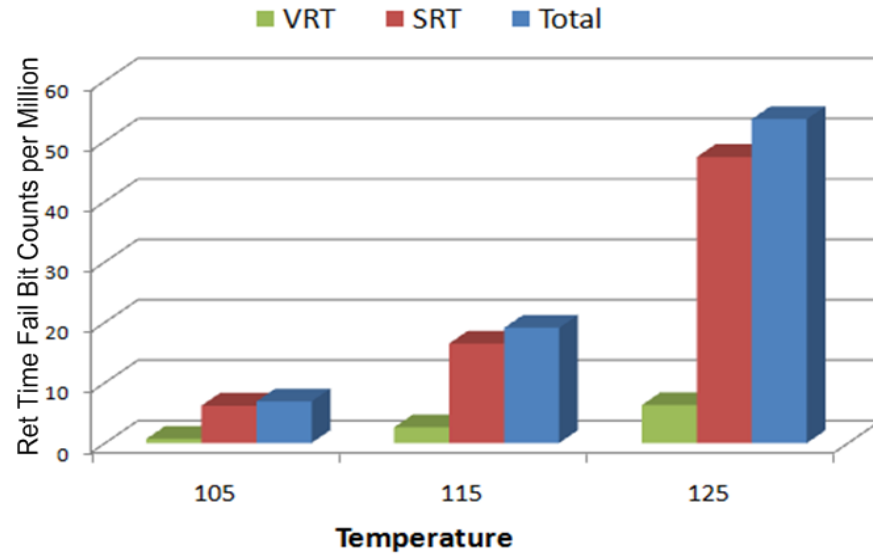


Figure 4.3: Fail Bit Counts per One Million Bits Sorted by Temperature at skew = w9,  $VP = 0.45V$ , and  $VDD = 1.20V$

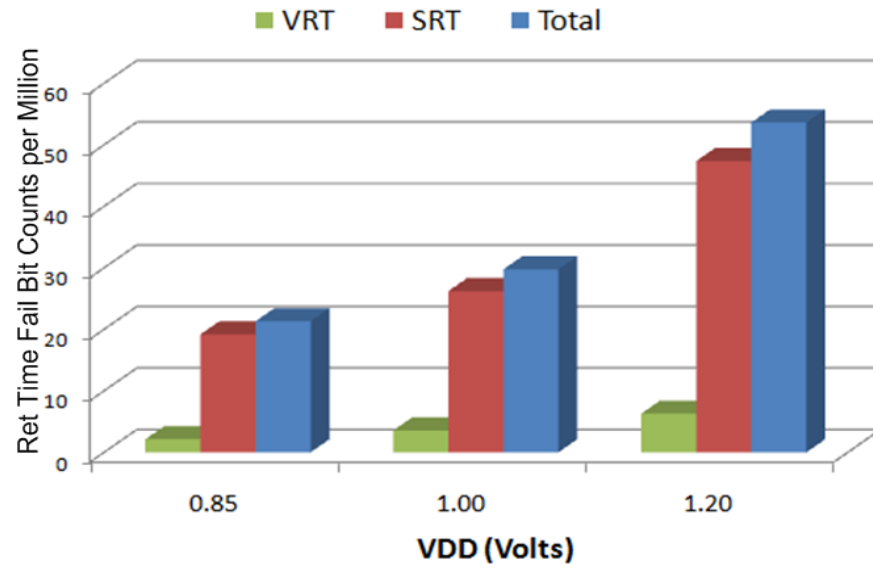


Figure 4.4: Fail Bit Counts per One Million Bits Sorted by  $VDD$  at skew = w9,  $VP = 0.45V$ ,  $T = 125^{\circ}C$

### VP (Body bias Voltage)

Figure 4.5 shows the trends of fail bit counts normalized by one million bits based on the VP voltage. Similar to VDD, the retention time fail counts increase as the VP voltage increases. Compared to VDD, the VP voltage change causes more fail bit counts by the smaller amount of increment from 0.40V to 0.45V. Since the magnitude of the two voltages influence the fail bit counts, the two voltages can't be grouped in one set. Hence, the voltage parameters are considered for the parametric extraction separately described in Section 4.2.

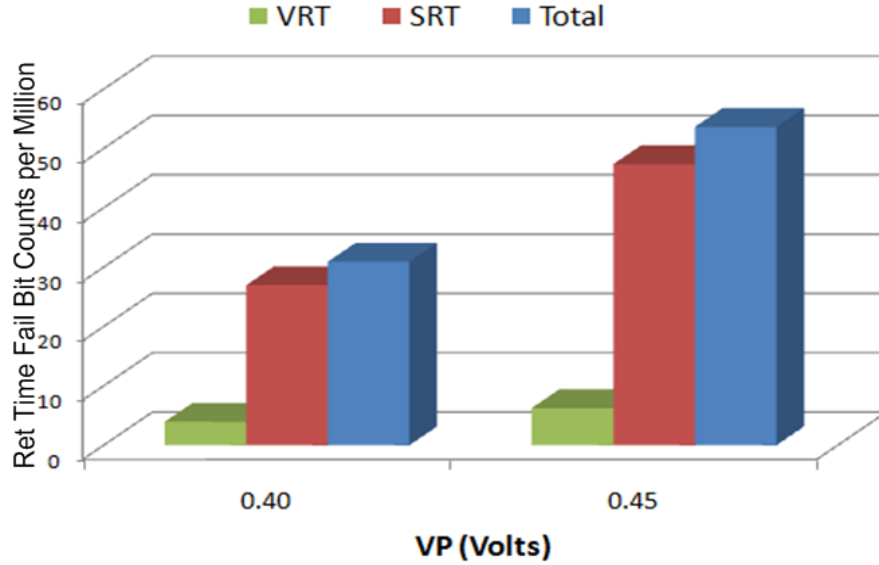


Figure 4.5: Fail Bit Counts per One Million Bits Sorted by  $VP$  at skew = w9,  $VDD = 1.20V$ ,  $T = 125^{\circ}C$

#### 4.1.2 Observed SRT and VRT Results

Figures 4.6-4.10 are the fail bit counts per one million bits based on the test conditions up to 2.6ms. Figure 4.11 shows the retention time fail counts normalized over one million bits for all the experimental test conditions. Even for the fast chips

(e.g., w9) at the most stringent test condition, the normalized fail bits counts are less than 100, which is considered as good quality. At the nominal use condition (i.e., Process Skew = w4, VP = 0.40V, VDD=1.00V, T=105°C), the normalized fail count is less than 5 up to 2.6ms. If the VRT fail bits in use are less than redundancy bits, these VRT fail bits can be corrected with the Error Correcting Code internally by software.

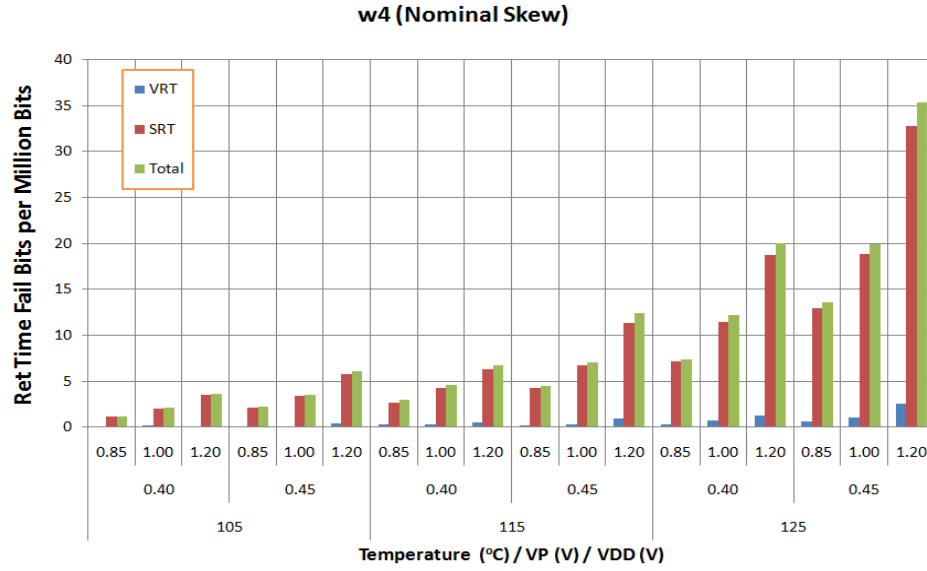


Figure 4.6: w4 (Nominal) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions

#### 4.1.3 Summary of Parameter Extraction

After analyzing the retention time fail characteristics across the four memory blocks, it is concluded that four separate memory blocks can be grouped and treated as one common set because there is no more than  $\pm 10\%$  of difference in the fail bit counts per one million bits counts between the blocks. The grouping of the memory groups eliminates the need for a block parameter in the retention time

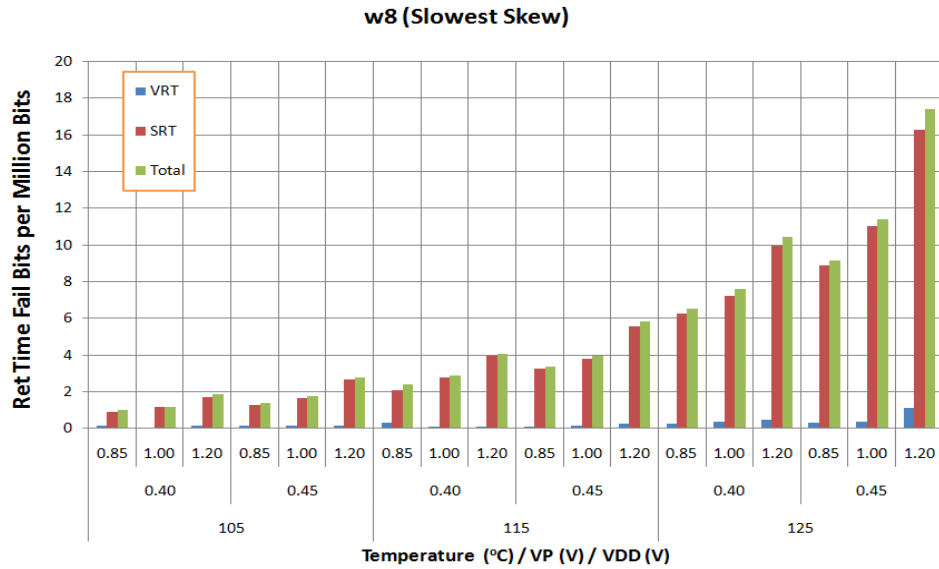


Figure 4.7: w8 (Slow) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions

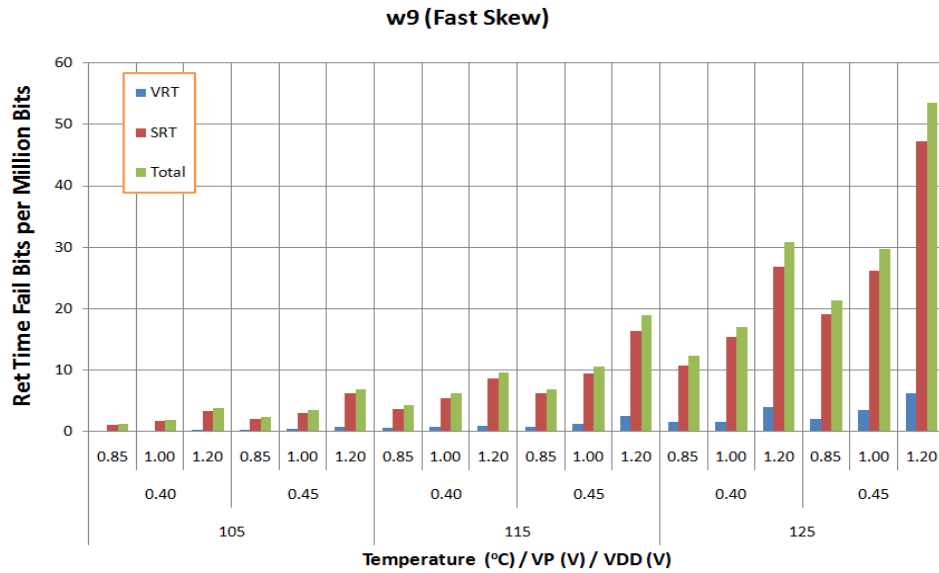


Figure 4.8: w9 (Fast) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions

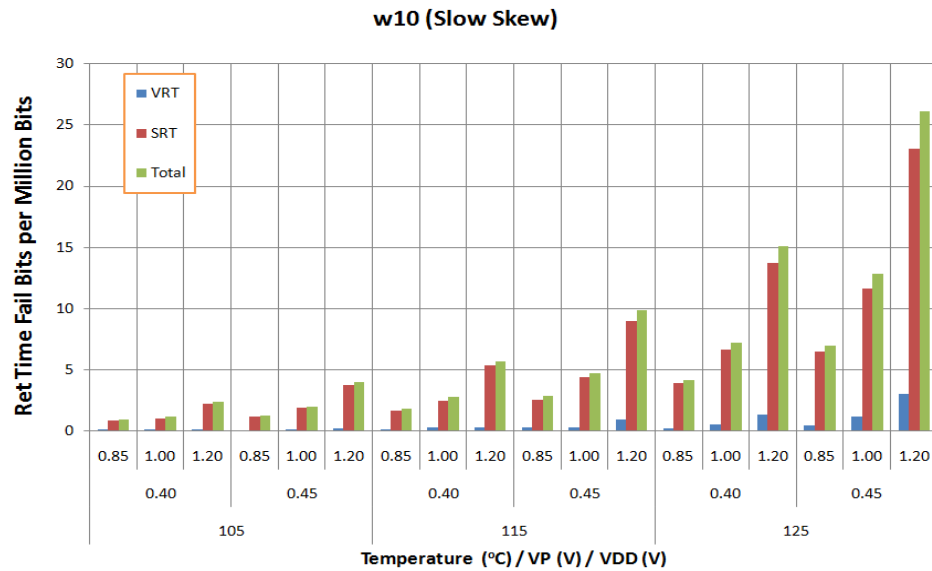


Figure 4.9: w10 (Slow) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions

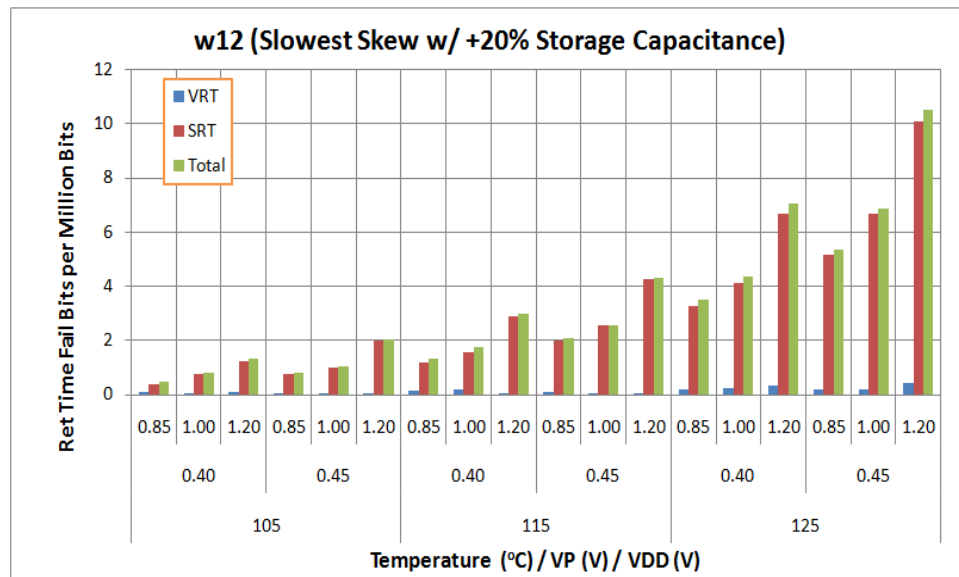


Figure 4.10: w12 (Slowest) Observed Fail Bit Counts per One Million Bits up to 2.6ms under Environmental Conditions

Temp (°C)	VP (V)	VDD (V)	Process Skew														
			w4 (nominal)			w8 (slowest)			w9 (fast)			w10 (slow)			w12 (slowest*)		
			VRT	SRT	Total	VRT	SRT	Total	VRT	SRT	Total	VRT	SRT	Total	VRT	SRT	Total
105	0.40	0.85	0	1	1	0	1	1	0	1	1	0	1	1	0	0	0
		1.00	0	2	2	0	1	1	0	2	2	0	1	1	0	1	1
		1.20	0	4	4	0	2	2	0	3	4	0	2	2	0	1	1
	0.45	0.85	0	2	2	0	1	1	0	2	2	0	1	1	0	1	1
		1.00	0	3	3	0	2	2	0	3	3	0	2	2	0	1	1
		1.20	0	6	6	0	3	3	1	6	7	0	4	4	0	2	2
115	0.40	0.85	0	3	3	0	2	2	1	4	4	0	2	2	0	1	1
		1.00	0	4	5	0	3	3	1	5	6	0	2	3	0	2	2
		1.20	0	6	7	0	4	4	1	9	10	0	5	6	0	3	3
	0.45	0.85	0	4	4	0	3	3	1	6	7	0	3	3	0	2	2
		1.00	0	7	7	0	4	4	1	9	11	0	4	5	0	3	3
		1.20	1	11	12	0	6	6	3	16	19	1	9	10	0	4	4
125	0.40	0.85	0	7	7	0	6	7	2	11	12	0	4	4	0	3	3
		1.00	1	11	12	0	7	8	2	15	17	1	7	7	0	4	4
		1.20	1	19	20	0	10	10	4	27	31	1	14	15	0	7	7
	0.45	0.85	1	13	14	0	9	9	2	19	21	0	7	7	0	5	5
		1.00	1	19	20	0	11	11	4	26	30	1	12	13	0	7	7
		1.20	3	33	35	1	16	17	6	47	53	3	23	26	0	10	11

Figure 4.11: Fail Bit Counts per One Million Bits (DPM) across All 5 skews up to 2.6ms Retention Time under Environmental Conditions

modeling. However, the two voltages (i.e., VP and VDD), temperature, and process skew are distinguishable, meaning there are significant differences when those parameters change (See Table 4.2 for the parameter extraction results). Therefore, the temperature and voltage parameters need to be included to extract when the failure rate of memory bits is modeled. Moreover, the proportion of VRT bit fails are roughly 10% of the SRT bits for each accelerated test condition. Section 4.2 describes the result of the quality models.

## 4.2 Methods of Quality Models

In the statistical quality modeling, the Single Retention Time (SRT) and Variable Retention Time (VRT) fails are modeled separately. As described in Chapter 3, there are two distinct retention times for VRT bits: the first one is minimum retention time (RetMin) and the second is maximum retention time (RetMax).

If a memory bit is a SRT fail bit, the RetMin and RetMax hold the same value and fail at the same retention time during multiple tests. While, the VRT fail bits hold different retention times filled as the RetMin and RetMax during the multiple tests. Section 4.2.1 describes the results of the SRT retention time analysis, and Section 4.2.2 indicates the results of the VRT retention time analysis.

#### 4.2.1 Single Retention Time (SRT) Use and Test Model

Using the Weibull distribution as described in Chapter 3, the Single Retention Time (SRT) fail bits are characterized as a memory block to predict the failure rate and conclude the quality level with DPM, yield loss, and overkill. In order to do so, the conservative modeling is performed for the SRT bits. Figure 4.12 is an example of the results of the Weibull SRT modeling at a particular test condition.

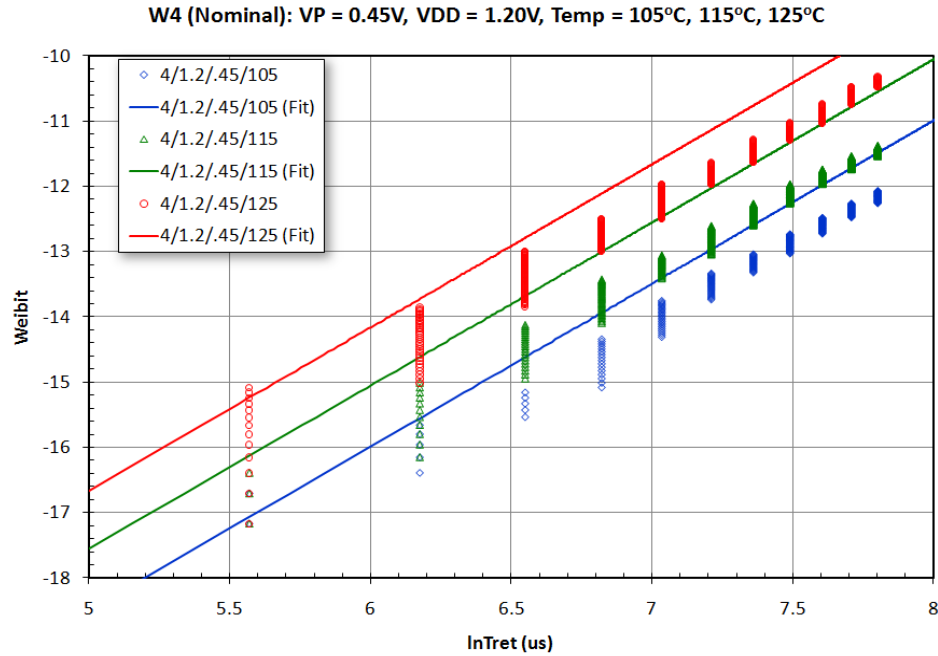


Figure 4.12: SRT Conservative Modeling Plot of Nominal Process Chip at Test Condition:  $VP = 0.45V$  and  $VDD = 1.20V$ . See Section 3.4.1 for reference.

By adjusting the  $a$  in Equation 3.6 to conservatively bound all the observed data, the beta (Weibull shape parameter) is extracted. For all test conditions conducted at PSU, the Weibull shape parameter  $\beta$  is found to be close to 2.5 in Figure 4.13. Due to small sample size (e.g., w8 slow chips), there are a few outliers in Figure 4.13. Since the conservative modeling is performed, the  $\beta$  is chosen to be 2.5 for all process skews. Next, the experimental parameters are also extracted given the shape parameter equal to 2.5.

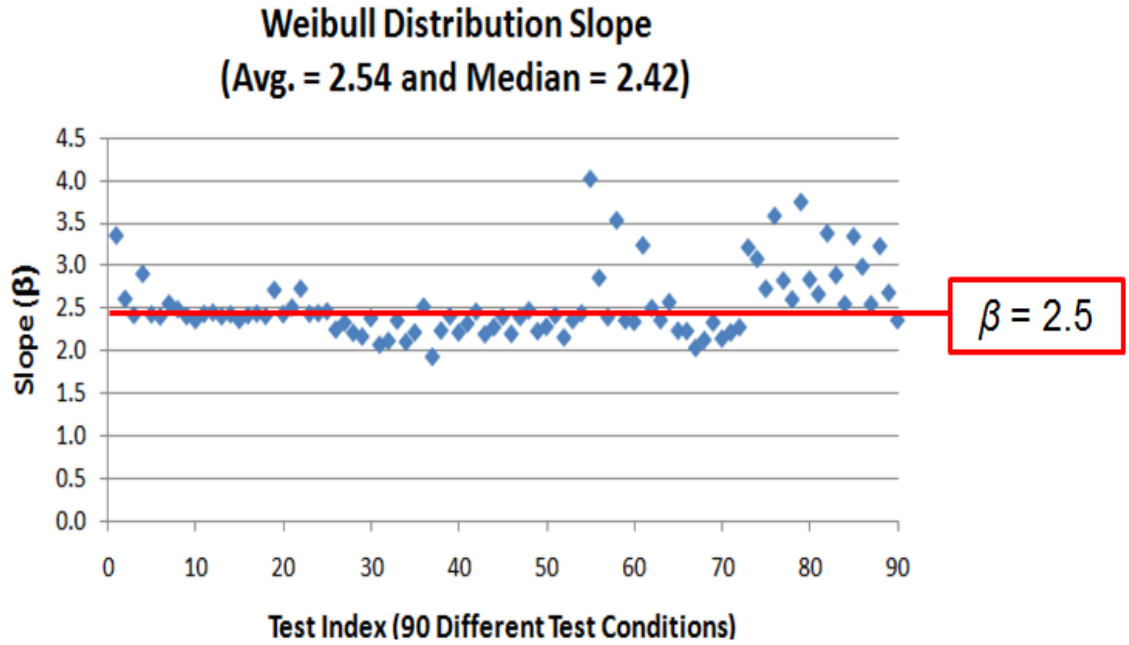


Figure 4.13: Weibull Fitted Shape Parameter over 90 Different Test Conditions

Table 4.2 shows the extracted SRT model parameter coefficients. The slopes  $\beta$  of the Weibull distribution for all the skews are set to be equal to 2.5 for ease of fitting the models to the observed data. The extracted  $\beta$  value is consistent with Lieneweg data (1998) in Figure 3 ( $\beta=3.2$ ) [14]. The y-intercepts of the Weibull distribution is adjusted to fit into the observed data by changing  $a$  in Equation

4.1. The y-intercepts include the coefficients of the temperature, VP, and VDD parameters along with one flexible coefficient which is independent of the test condition parameter as shown in Equation 4.1. The extracted activation energy  $Q$  is also consistent with Lieneweg data (1988) in Figure 6 [14].

$$\ln \alpha = a + b * VP + c * VDD + \frac{Q}{k_B T} \quad (4.1)$$

Table 4.2: Single Retention Time Parameter Coefficients

skew	4	8	9	10	12
<b>Beta</b>	2.5	2.5	2.5	2.5	2.5
<b>Q (eV)</b>	0.475	0.455	0.542	0.445	0.500
<b>a</b>	1.1	1.1	-0.9	2	0.4
<b>b</b>	-4.28	-3.23	-4.63	-4.18	-3.46
<b>c</b>	-1.16	-0.764	-1.16	-1.41	-1.02

#### 4.2.2 SRT and VRT Data Example in 2-D

Figure 4.14 shows raw bit counts of SRT and VRT under the same temperature and voltages conditions at test and use. In Figure 4.14, SRT bits are on the 45-degree line because SRT bits have only one retention time:  $t_{RetMin} = t_{RetMax}$ . VRT bits are located at the top-left corner because the VRT bits have multiple retention times:  $t_{RetMax} > t_{RetMin}$ . Test and use set points ( $t_{test}$  and  $t_{use}$ ) define the test and use fails. By setting the test and use points, the figures-of-merit, such as *DPM*, *Overkill*, *Reject*, and *Pass* are evaluated. In Chapter 5, one application using this method is described.

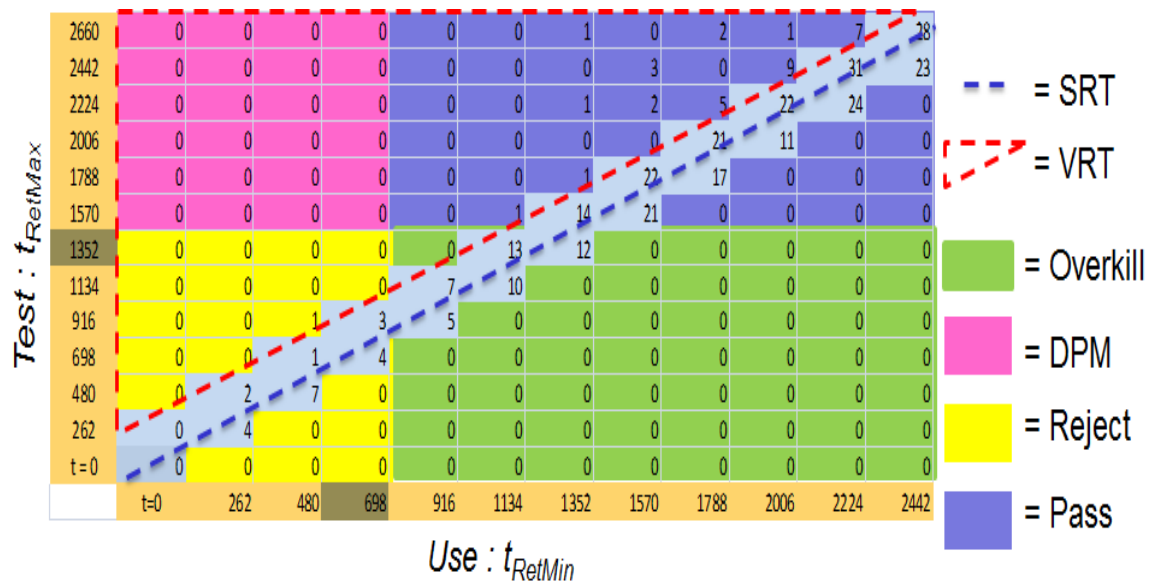


Figure 4.14: SRT and VRT Quality Example in Two-Dimension

## 5 Conclusion: Implications, Applications, and Future Work

### 5.1 Applications

In Section 5.1, one data-driven application example using the quality figures-of-merit is presented. In general, voltage is more convenient than temperature to adjust retention times because keeping temperature constant during test requires a temperature controller and thermo sensor. If high voltage conditions at a low temperature can provide the equivalent yield, test escapes, and overkill compared to a high temperature and nominal voltages, the high voltage tests would be less cost compared to the high temperature tests. Using the retention time data, similar SRT and VRT cumulative counts at Condition 1: a high temperature and low voltages and Condition 2: a low temperature and high voltages are examined (Figure 5.1).

Table 5.1 displays the cumulative retention time fails set by  $t_{test} = 1200\mu s$  and  $t_{use} = 600\mu s$  at a high temperature ( $125^\circ$ ) and nominal voltage conditions. The w9 (fast) skew is selected for this application example because of large sample size. The cumulative counts are classified as *DPM*, *Pass*, *Reject*, and *Overkill* (i.e., the quality figures-of-merit). Note that *Yield Loss* is the sum of *Reject* and *Overkill*.

Table 5.1: SRT and VRT Cumulative Counts: Original Conditions {Process Skew=w9 (fast), Temperature= $125^\circ C$ ,  $t_{test} = 1200\mu s$ ,  $t_{use} = 600\mu s$ , VDD=1.00V, VP=0.40V} Note that *Yield Loss*=146.

Test \ Use	Bad ( $t_{RetMin} < t_{use}$ )	Good ( $t_{RetMin} > t_{use}$ )
Pass ( $t_{RetMax} > t_{test}$ )	<i>DPM</i> =1	<i>Pass</i> =682+
Bad ( $t_{RetMax} < t_{test}$ )	<i>Reject</i> =40	<i>Overkill</i> =106

### Condition 1 (High Temperature)

- Process Skew = w9 (fast)
  - VDD = 1.00V
  - VP = 0.40V
  - Temp = 125°C
- Nominal Conditions

### Set conditions

Test Retention Time ( $t_{test}$ ) = 1200 $\mu$ s  
Use Refresh Time ( $t_{use}$ ) = 600 $\mu$ s

2660	1	11	25	41	84	128	198	301	407	527	720	829
2442	1	11	25	40	83	126	196	297	401	518	685	729
2224	1	11	25	40	82	125	195	295	397	504	590	550
2006	1	11	25	40	81	124	194	293	386	414	414	414
1788	1	11	25	40	81	124	193	285	312	312	312	312
1570	1	11	25	40	81	124	190	225	225	225	225	225
1352	1	11	25	40	81	122	146	146	146	146	146	146
1134	1	11	25	40	80	97	97	97	97	97	97	97
916	1	11	25	40	65	65	65	65	65	65	65	65
698	1	11	25	35	35	35	35	35	35	35	35	35
480	1	11	17	17	17	17	17	17	17	17	17	17
262	1	9	9	9	9	9	9	9	9	9	9	9
t=0	1	1	1	1	1	1	1	1	1	1	1	1
	t=0	262	480	698	916	1134	1352	1570	1788	2006	2224	2442

### Condition 2 (Raise Voltages)

- Process Skew = w9 (fast)
- VDD = 1.20V
- VP = 0.45V
- Temp = 105°C

### Retention Time Plot Cumulative Counts - sum fail counts in box

x-axis:  $t_{RetMin}$  (Use)  
y-axis:  $t_{RetMax}$  (Test)

2660	0	6	15	22	34	58	87	135	180	223	285	336
2442	0	6	15	22	34	58	86	134	177	219	274	297
2224	0	6	15	22	34	58	86	131	174	207	231	231
2006	0	6	15	22	34	58	85	128	166	177	177	177
1788	0	6	15	22	34	58	85	128	145	145	145	145
1570	0	6	15	22	34	58	84	105	105	105	105	105
1352	0	6	15	22	34	57	69	69	69	69	69	69
1134	0	6	15	22	34	44	44	44	44	44	44	44
916	0	6	15	22	27	27	27	27	27	27	27	27
698	0	6	14	18	18	18	18	18	18	18	18	18
480	0	6	13	13	13	13	13	13	13	13	13	13
262	0	4	4	4	4	4	4	4	4	4	4	4
t=0	0	0	0	0	0	0	0	0	0	0	0	0
	t=0	262	480	698	916	1134	1352	1570	1788	2006	2224	2442

Use:  $t_{RetMin}$

Figure 5.1: Cumulative SRT and VRT Counts at High Temperature Condition (1) and High Voltage Condition (2)

Next, the quality figures-of-merit at high voltages (VDD=1.20V, VP=0.45V) and the nominal temperature (105°) are evaluated.

Table 5.2: SRT and VRT Cumulative Counts: Modified Conditions 1 {Process Skew=w9 (fast),  $t_{test} = 1200\mu s$ ,  $t_{use} = 600\mu s$ , Temperature=105°C, VDD=1.20V, VP=0.45V} Note that *Yield Loss*=69.

Test \ Use	Bad ( $t_{RetMin} < t_{use}$ )	Good ( $t_{RetMin} > t_{use}$ )
Pass ( $t_{RetMax} > t_{test}$ )	<i>DPM</i> =0	<i>Pass</i> =265+
Bad ( $t_{RetMax} < t_{test}$ )	<i>Reject</i> =22	<i>Overkill</i> =47

With only the temperature and voltage adjustments, high voltages under a low temperature still do not provide equivalent figures-of-merit. Next, test and use set points  $t_{test}$  and  $t_{use}$  under the same voltages and temperature conditions are adjusted to obtain the equivalent figures-of-merit under the original conditions. Table 5.3 displays the figures-of-merit after the  $t_{test}$  and  $t_{use}$  are fitted to the original figures-of-merit.

Table 5.3: SRT and VRT Cumulative Counts: Modified Conditions 2 {Process Skew=w9 (fast),  $t_{test} = 1500\mu s$ ,  $t_{use} = 800\mu s$ , Temperature=105°C, VDD=1.20V, VP=0.45V} Note that *Yield Loss*=145.

Test \ Use	Bad ( $t_{RetMin} < t_{use}$ )	Good ( $t_{RetMin} > t_{use}$ )
Pass ( $t_{RetMax} > t_{test}$ )	<i>DPM</i> =0	<i>Pass</i> =190+
Bad ( $t_{RetMax} < t_{test}$ )	<i>Reject</i> =34	<i>Overkill</i> =111

At  $t_{test}=1500\mu s$  and  $t_{use}=800\mu s$ , the figures-of-merit are now equivalent with the original conditions. In summary, using the quality figures-of-merit, trade-offs of *DPM*, *Overkill*, and *Yield Loss* can be considered. Voltage and temperature conditions set DRAM bit retention time. The Weibull parameter  $\alpha$  described in Chapter

3 and 4 is used to adjust the retention time by independently changing voltages and temperature. Test and Use set points  $t_{test}$  and  $t_{use}$  reflect application requirements.

## 5.2 Conclusion

The eDRAM quality analysis utilizing the 65nm ASIC technology is performed and the 65nm eDRAM quality is excellent with less than 100 fail bit counts per one million at the desired use condition. During the retention time fail bit analysis, the effect of VRT fail bits to the eDRAM quality is approximately 10% of that of SRT fail bits. Also, the body bias voltage VP has more influences in retention time fail bits counts than the supply voltage VDD. For the overkill, since the VRT effect is very small, the test condition needs to be stringent enough to reject SRT fail bits (See Section 4.3.1 for reference).

## 5.3 Future Work

The eDRAM quality study initiates the utilization of use and test conditions to eDRAM retention time fails. The data collected for this research is sufficient to conduct the quality analysis of two different types of retention time fails in DRAM. There are further interesting opportunities that can be pursued to improve the study of eDRAM memory bits.

1. The variable retention time (VRT) bits are not fully modeled. The model presented in this thesis can provide fully modeled Single Retention Time bits based on the use and test conditions; however, the VRT bits failure rate is predicted based on the observed data.
2. The Error Correcting Code (ECC) with redundancy bits utilized at the use

condition need to be implemented.

3. The lifetime (reliability) analysis on eDRAM bits can be analyzed by monitoring the bits for a long run. The memory-bit reliability analysis will be beneficial as the technology aims for the shorter gate length.
4. The chances of state change from SRT bits to VRT bits are not analyzed. Over many refresh cycles, some portion of SRT bits might have variable retention times at any following cycles.

## References

- [1] International Technology Roadmap for Semiconductors 2007. [Online]. Available: <http://www.itrs.net/>
- [2] N. H. E. Weste and D. Harris, *CMOS VLSI Design*. Boston: Pearson Education, Inc., 2005, ch. 9, pp. 734–739.
- [3] D. Keitel-Schulz and N. Wehn, “Embedded dram development: Technology, physical design, and application issues,” *IEEE DESIGN TEST OF COMPUTERS*, vol. 18, no. 03, pp. 7–15, 2001.
- [4] —, “Issues in embedded dram development and applications,” *System Synthesis, International Symposium on*, vol. 0, p. 23, 1998.
- [5] M. L. Bushnell and V. D. Agrawal, *Essentials of Electronic Testing for Digital, Memory, and Mixed-Signal VLSI Circuits*. Boston: Kluwer Academic, 2000, vol. 17.
- [6] N. H. E. Weste and D. Harris, *CMOS VLSI Design*. Boston: Pearson Education, Inc., 2005, ch. 9, pp. 715–734.
- [7] Y. Mori, K. Ohyu, K. Okonogi, and R. Yamada, “The origin of variable retention time in dram,” in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, 5-5 2005, pp. 1034 –1037.
- [8] P. Restle, J. Park, and B. Lloyd, “Dram variable retention time,” in *Electron Devices Meeting, 1992. Technical Digest., International*, 13-16 1992, pp. 807–810.

- [9] D. Yaney, C. Lu, R. Kohler, M. Kelly, and J. Nelson, “A meta-stable leakage phenomenon in dram charge storage 8212;variable hold time,” in *Electron Devices Meeting, 1987 International*, vol. 33, 1987, pp. 336 – 339.
- [10] A. K. Sharma, *Semiconductor Memories: Technology, Testing, and Reliability*. New York: IEEE Press, 1999, ch. 2, pp. 40–75.
- [11] M. C.-T. Chao, H.-Y. Yang, R.-F. Huang, S.-C. Lin, and C.-Y. Chin, “Fault models for embedded-dram macros,” in *DAC '09: Proceedings of the 46th Annual Design Automation Conference*. New York, NY, USA: ACM, 2009, pp. 714–719.
- [12] W. Weibull, “A statistical distribution of wide applicability,” *Journal of Applied Mechanics*, pp. 293–297, 1951.
- [13] Applied Reliability ECE510 Lecture Notes. [Online]. Available: <http://web.cecs.pdx.edu/~cgshirl/>
- [14] U. Lieneweg, D. Nguyen, and B. Blaes, “Assessment of dram reliability from retention time measurements,” *Flight Readiness Technology Assessment NASA EEE Parts Program*, 1998.
- [15] P. A. Tobias and D. C. Trindade, *Applied Reliability*. New York: Van Nostrand Reinhold, 1994, ch. 4, pp. 81–104.