National Aeronautics and Space Administration

Flash Memory Reliability—Read, Program, and Erase Latency Versus Endurance Cycling

Jason Heidecker
Jet Propulsion Laboratory
Pasadena, California

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

JPL Publication 10-19 11/10

National Aeronautics and Space Administration

Flash Memory Reliability— Read, Program, and Erase Latency Versus Endurance Cycling

NASA Electronic Parts and Packaging (NEPP) Program
Office of Safety and Mission Assurance

Jason Heidecker
Jet Propulsion Laboratory
Pasadena, California

NASA WBS: 724297.40.49
JPL Project Number: 103928
Task Number: 03.02.03

Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109

http://nepp.nasa.gov

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the National Aeronautics and Space Administration Electronic Parts and Packaging (NEPP) Program.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Copyright 2010. California Institute of Technology. Government sponsorship acknowledged.

TABLE OF CONTENTS

Abstract 1
1.0 Introduction 2
2.0 Test Setup and Procedure 3
  2.1 Test Equipment 3
  2.2 Test Samples 3
  2.3 Erase, Program, and Read Timing Diagrams 4
  2.4 Test Procedure 6
3.0 Results 7
  3.1 Micron 128 Gb, MLC, MT29F128G08CJAAA 7
  3.2 Micron 8 Gb, SLC, MT29F8G08AAA 9
  3.3 Samsung 8 Gb, MLC, K9G8G08UOM 10
  3.4 Samsung 8 Gb, SLC, K9F8G08UOM 12
4.0 Summary 14
5.0 Recommendation to Space Missions 15
6.0 References 16

ABSTRACT

This report documents the efforts and results of the fiscal year (FY) 2010 NASA Electronic Parts and Packaging Program (NEPP) task for nonvolatile memory (NVM) reliability. This year's focus was to measure latency (read, program, and erase) of NAND Flash memories and determine how these parameters drift with erase/program/read endurance cycling.

1.0 INTRODUCTION

As the scope and requirements of space missions continue to increase, so does the need for local memory on board spacecraft to store science data before beaming it back to Earth. NAND Flash, although not radiation-hardened, is attractive for its large densities and nonvolatility. A single NAND Flash chip in a standard thin small outline package (TSOP) can offer up to 256 Gb (multiple die) of memory, whereas other nonvolatile alternatives such as CRAM, MRAM, FRAM, and EEPROM are limited to about 16 Mb—a factor of 1000× less. Unfortunately, Flash, like most NVM memories, is significantly slower than its volatile counterparts like SRAM and DRAM; so slow that erasing, programming, and reading a 128 Gb device takes 80 minutes at maximum clock frequency with typical erase/program/read latencies. However, if these latencies drift to maximum datasheet values, one erase/program/read cycle of the device can increase to three hours. This is a big concern for NASA space missions that utilize large amounts of NAND Flash and need to cycle it frequently (such as an Earth-observing mission completing orbits every 90 minutes). These types of missions need to understand how latencies can drift over the life of the part (i.e., endurance cycling).

"Latency" refers to the amount of time spent waiting on a particular operation to be completed. When a block erase command is sent to the NAND device, the host system must wait for the device to complete the operation before sending the next command. This time is known as "block erase latency," or t_{BERS}, and is on the order of a few milliseconds. "Read latency," or t_R, is the time (25–50 μs) it takes the device to prepare a page of data to be read out after the host system has sent the read command for a particular page address. Finally, "program latency," or t_{PROG}, is the period of time that must be observed by the host system (~200–2000 μs) while the NAND writes a page of data that has just been latched-in by the host.

This report describes erase, program, and read latencies measured at different levels of endurance cycling—0×, 0.5×, 1.0× and 2.0× datasheet endurance specification—for various NAND Flash devices. Latency distributions at each level were compared to measure change in latency values with cycling.

2.0 TEST SETUP AND PROCEDURE

2.1 Test Equipment

Testing was performed at the Jet Propulsion Laboratory (JPL) using the JD Instruments Automated Test Vector (ATV) test system (Figure 2.1-1). The JDI ATV system is comprised of a general-purpose test head, control unit, and PC. Custom load boards were made for mating the NAND Flash devices under test (DUTs) to the general-purpose test head (Figure 2.1-2). In order to obtain maximum test system reliability, parts were operated at 4 MHz, although these parts are capable of 50 MHz operation. This does not affect latency measurements.

Latencies were measured by monitoring the ready/busy (R/B#) pin on the DUT. This pin goes low during a given erase, program, or read operation and returns high once the operation has been completed. At 4 MHz operation, the test system can measure latencies with a resolution of 4.25 μs. For measuring program and erase latencies that can be on the order of hundreds and thousands of microseconds, this is not an issue. However, when measuring read latencies of about 25 μs, resolution is not that good. Read latency measurements of 25.5 μs (6× 4.25 μs) could mean an actual value between 21.75 and 25.5 μs. That is nearly 25% error in that case. It is recommended that future read latency testing be done with a test system that can reliably operate at higher frequencies. A 2× increase in operation frequency corresponds to a 2× improvement in resolution. At 40 MHz operation, resolution in latency measurements would improve by a factor of 10× to 0.425 μs. It is believed the limiting factor in the current test system is cabling. Improvements in the cabling between the ATV control unit and test head could provide the necessary 40 MHz operation in future testing.

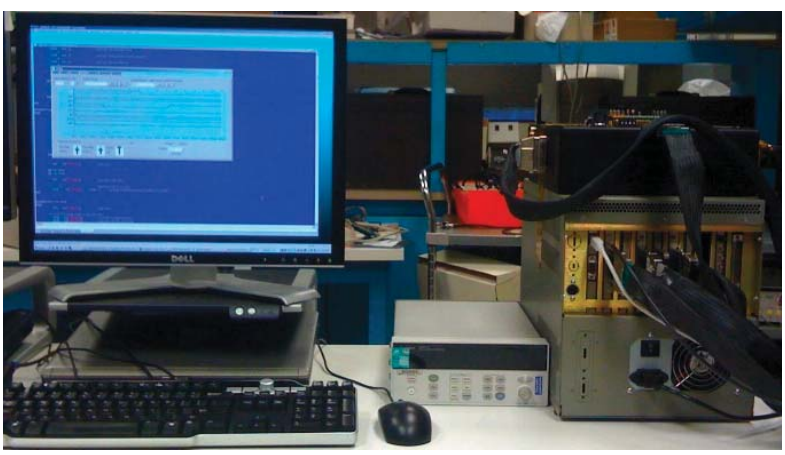

Figure 2.1-1. JDI ATV. Test head in upper right sitting atop the ATV control unit.

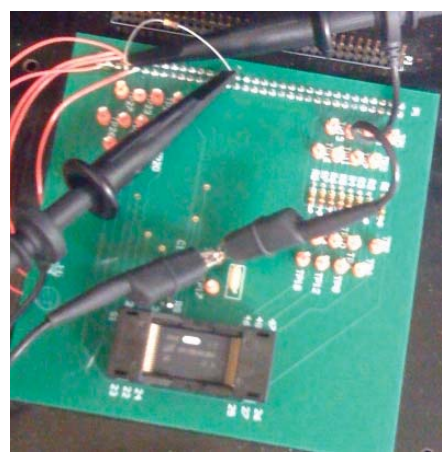

Figure 2.1-2. Custom DUT load board.

2.2 Test Samples

Both single-level-cell (SLC) and multi-level-cell (MLC) technologies were tested (Table 2.2-1). These specific part numbers were chosen as they have already been shown to have favorable performance under irradiation and therefore may be suitable for NASA space missions [1, 2]. All devices are single die with the exception of the 128 Gb device from Micron. It is packaged in the same 48-pin TSOP as the other devices; however, there are four vertically stacked 32 Gb die inside.

Table 2.2-1. Devices under Test

Part Number	Manufacturer	Density	SLC/MLC	Endurance Spec
MT29F128G08CJAAA	Micron	128 Gb	MLC	5,000
MT29F8G08AAA	Micron	8 Gb	SLC	100,000
K9G8G08UOM	Samsung	8 Gb	MLC	5,000
K9F8G08UOM	Samsung	8 Gb	SLC	100,000

2.3 Erase, Program, and Read Timing Diagrams

All four devices tested operate according to the open NAND Flash Interface (ONFI) standard. This means all four devices follow the same operational timing diagrams; the difference being that each device has its own specifications for the key latency parameters. The timing diagrams for the erase, program, and read operations are given in Figures 2.3-1 to 2.3-3 [3].

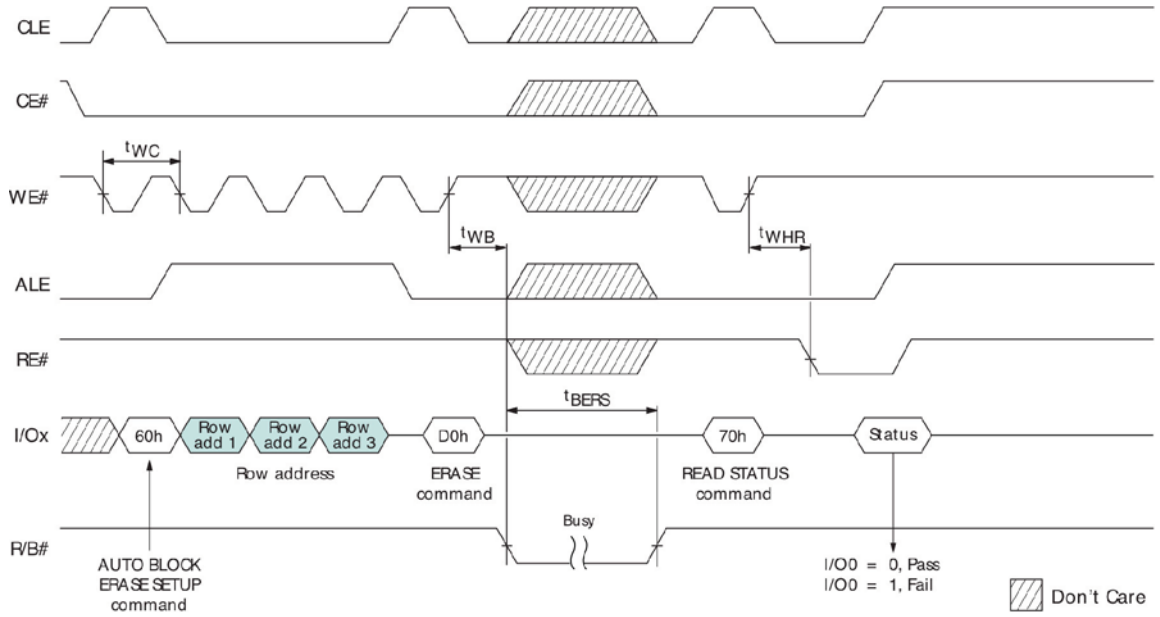

Figure 2.3-1. Block Erase. First, a "60h" command is sent to let the device know an address for a block erase is coming. This is followed by 3 cycles of addresses, which together define a single block. This is followed by the "D0h" command. Then the device will spend an amount of time doing the actual erasing, which is known as "erase latency," or t_{BERS}. R/B# will go low while the device is busy and return high once the device is ready.

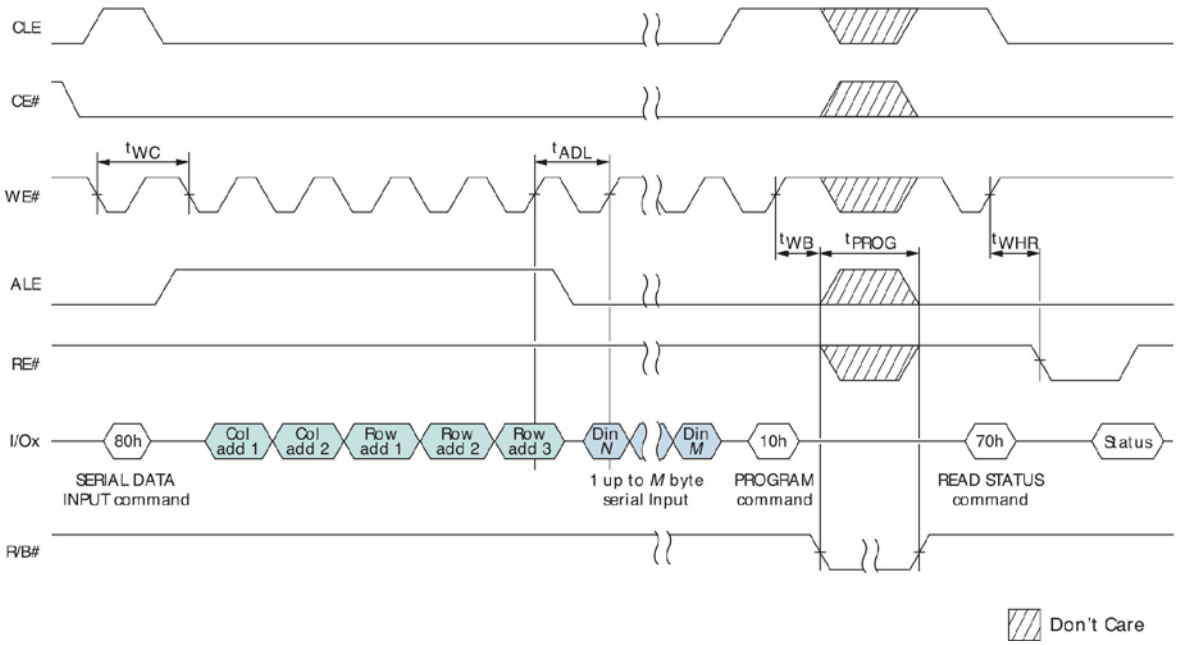

Figure 2.3-2. Page Program. First, a "80h" command is sent to let the device know an address for a page program is coming. This is followed by 5 cycles of addresses, which together define a single page. This is followed by serial input of the entire page of data. Then the device will spend an amount of time doing the actual programming, which is known as "program latency," or t_{PROG}. R/B# will go low while the device is busy and return high once the device is ready.

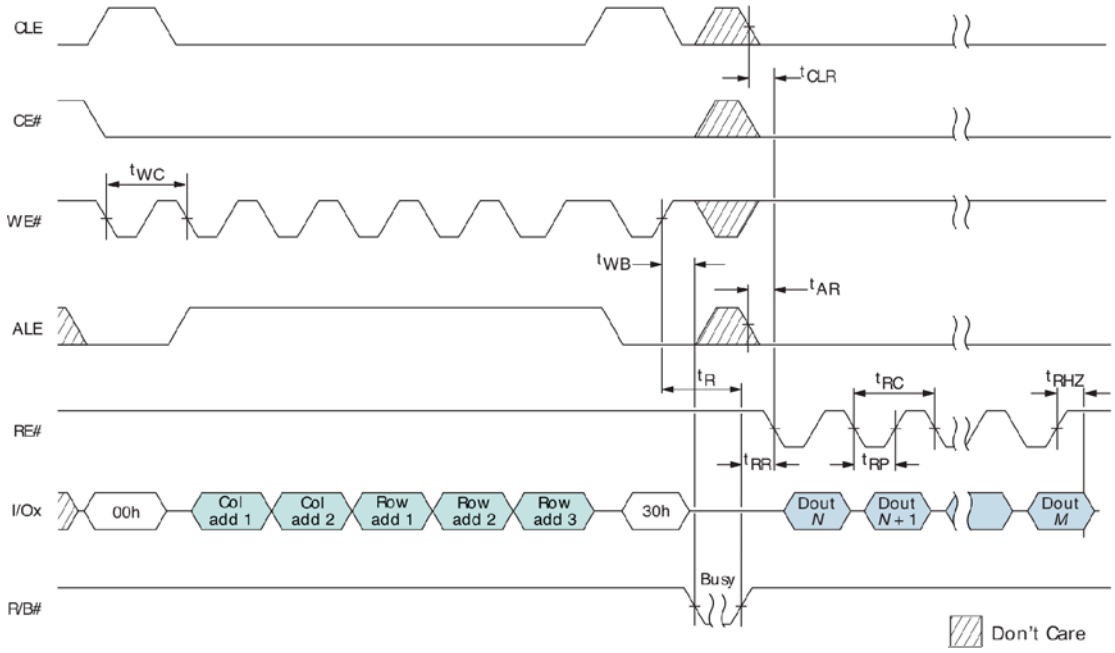

Figure 2.3-3. Page Read. First, a "00h" command is sent to let the device know an address for a page program is coming. This is followed by 5 cycles of addresses, which together define a single page. This is followed by the "30h" command. Then R/B# will go low while the device is busy (this is read latency, t_R) preparing the data to be read out by the host system.

2.4 Test Procedure

Testing was done on a single block at a time. 12 blocks from 4 samples/die of each NAND device in Table 2.2-1 were tested. This is a total 48 blocks per device type. The testing done on each block was as follows:

1. Erase Block, measure t_{BERS}.
2. Program Block All Zeros.
3. Repeat steps through 1–2 ten times (ten values for t_{BERS}).
4. Erase Block.
5. Program each page with checkerboard pattern. Record t_{PROG} for each page.
6. Read each page. Record t_R for each page.
7. Perform erase/program/read cycles.
8. Repeat steps 1–6 (further to be known as the "latency characterization") after 0.5×, 1.0×, and 2.0× endurance cycling specification for device.

For example, the testing for a block from the Micron 128 Gb MLC device would proceed as follows:

1. Latency characterization
2. 2,500 cycles
3. Latency characterization
4. 2,500 cycles
5. Latency characterization
6. 5,000 cycles
7. Latency characterization

All testing was done at nominal V_{CC} (3.3 V) and room temperature.

Cycling involved erasing the block, programming it, and reading it. Alternating checkerboard and inverse checkerboard patterns were used. This means for one cycle the data pattern would be "01010101" for all addresses and then it would be "10101010" for all addresses on the next cycle. Although certainly seen during testing (in great numbers in the MLC devices), bit errors were not recorded.

3.0 RESULTS

The following boxplots depict the change in erase, program, and read latency with cycling. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Table 3-1 gives the datasheet specifications for these latency parameters.

Table 3-1. Datasheet specifications for latency parameters.

Part Number	t_{BERS} (ms)		t_{PROG} (µs)		t_R (µs)	
	Typ	Max	Typ	Max	Typ	Max
MT29F128G08CJAAA	3	10	900	2200	–	50
MT29F8G08AAA	0.7	3	250	700	–	25
K9G8G08UOM	1.5	10	800	3000	–	60
K9F8G08UOM	1.5	2	200	700	–	25

3.1 Micron 128 Gb, MLC, MT29F128G08CJAAA

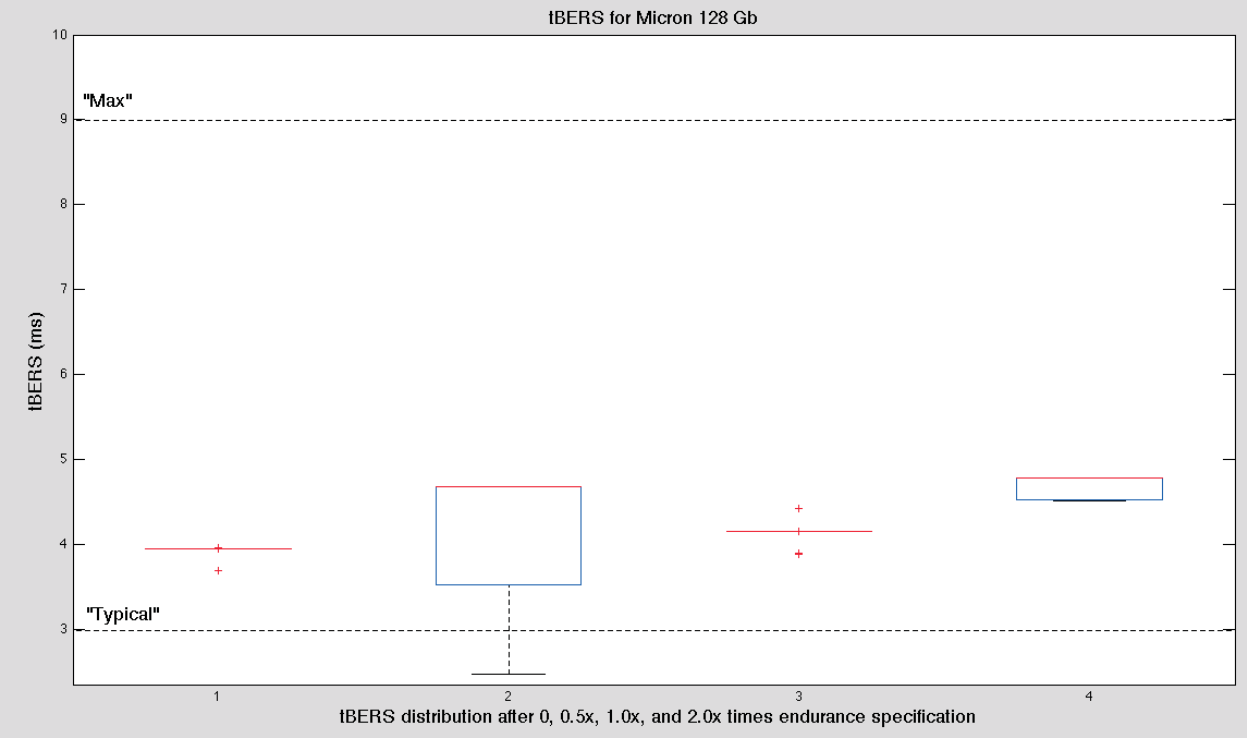

Figure 3.1-1. t_{BERS} for Micron 128 Gb after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

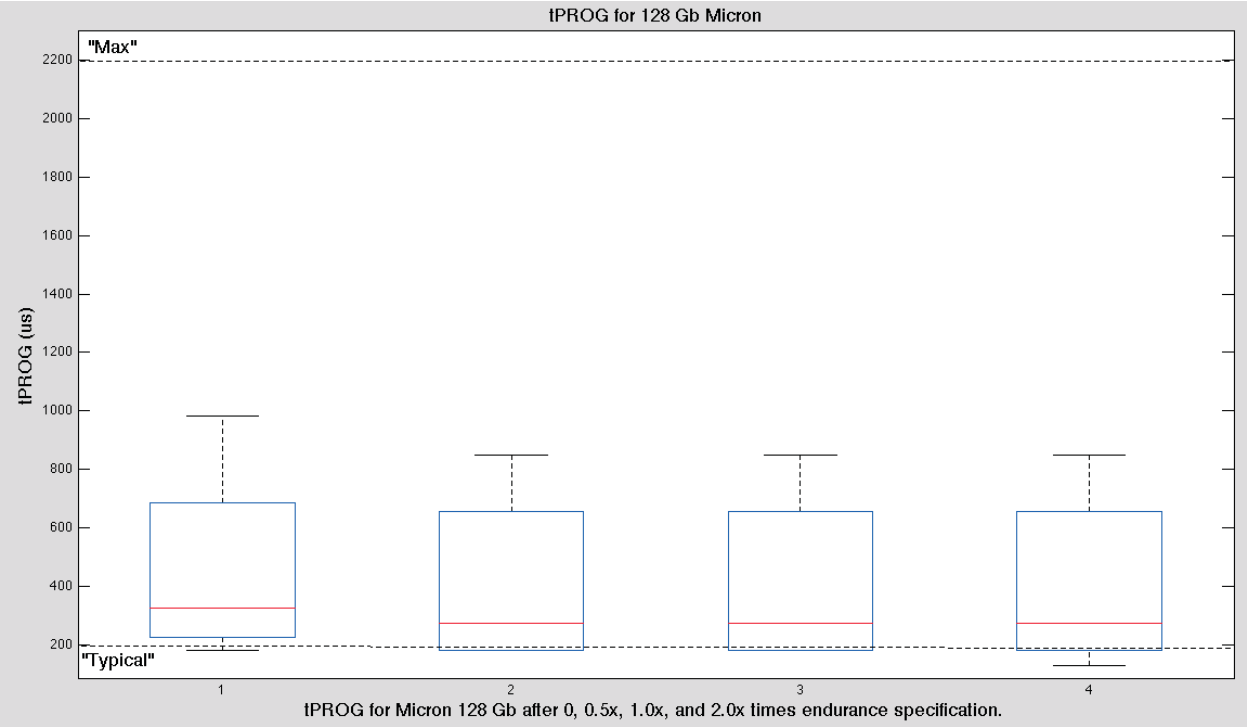

Figure 3.1-2. t_{PROG} for Micron 128 Gb after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

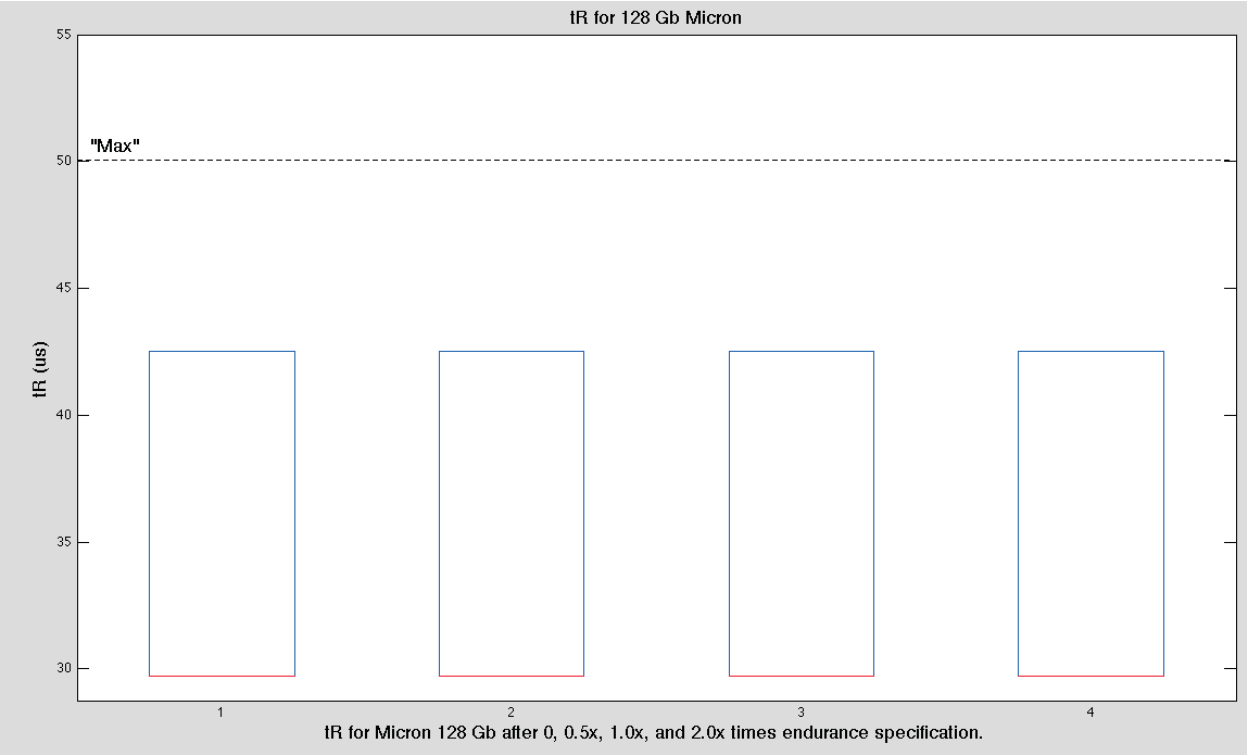

Figure 3.1-3. t_R for Micron 128 Gb after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

3.2 Micron 8 Gb, SLC, MT29F8G08AAA

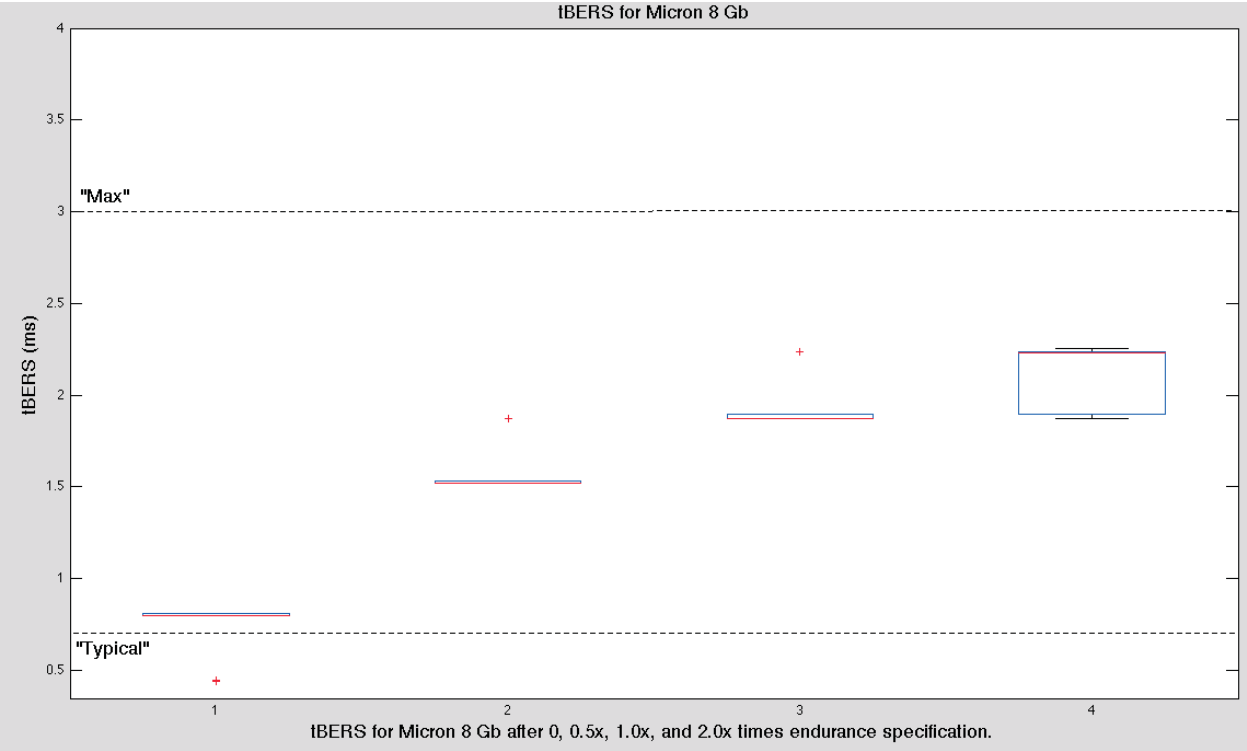

Figure 3.2-1. t_{BERS} for Micron 8 Gb after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

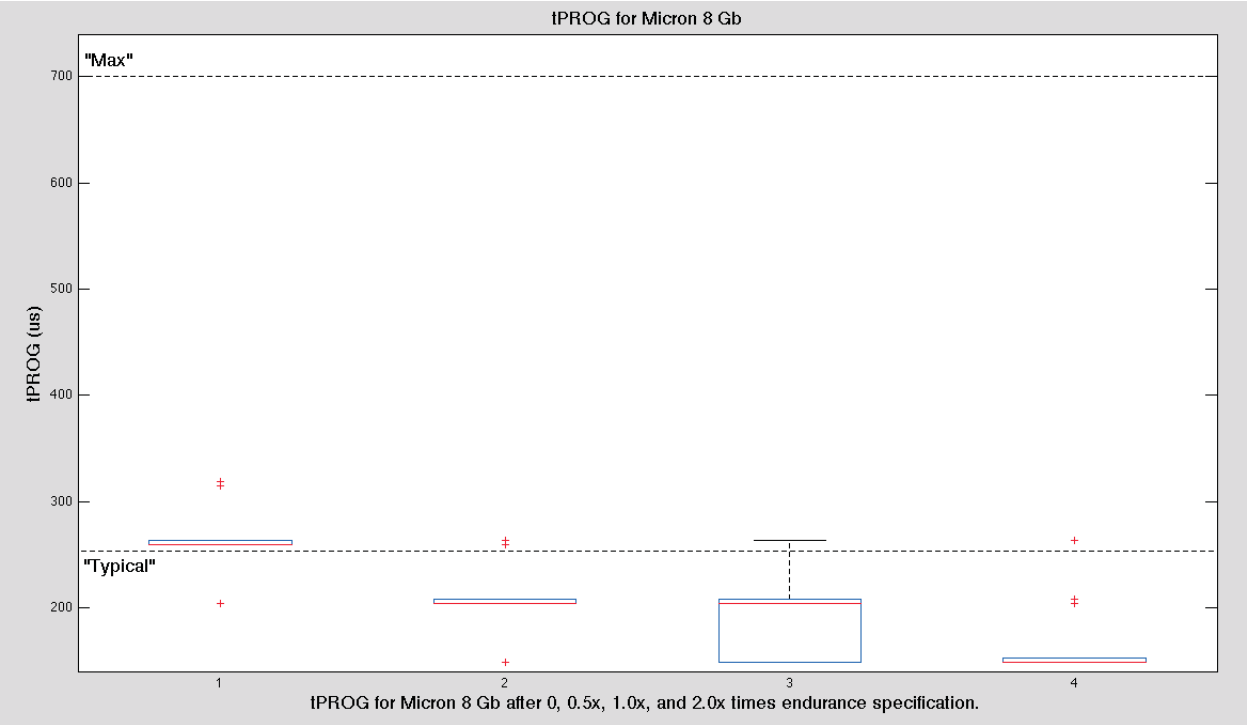

Figure 3.2-2. t_{PROG} for Micron 8 Gb after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

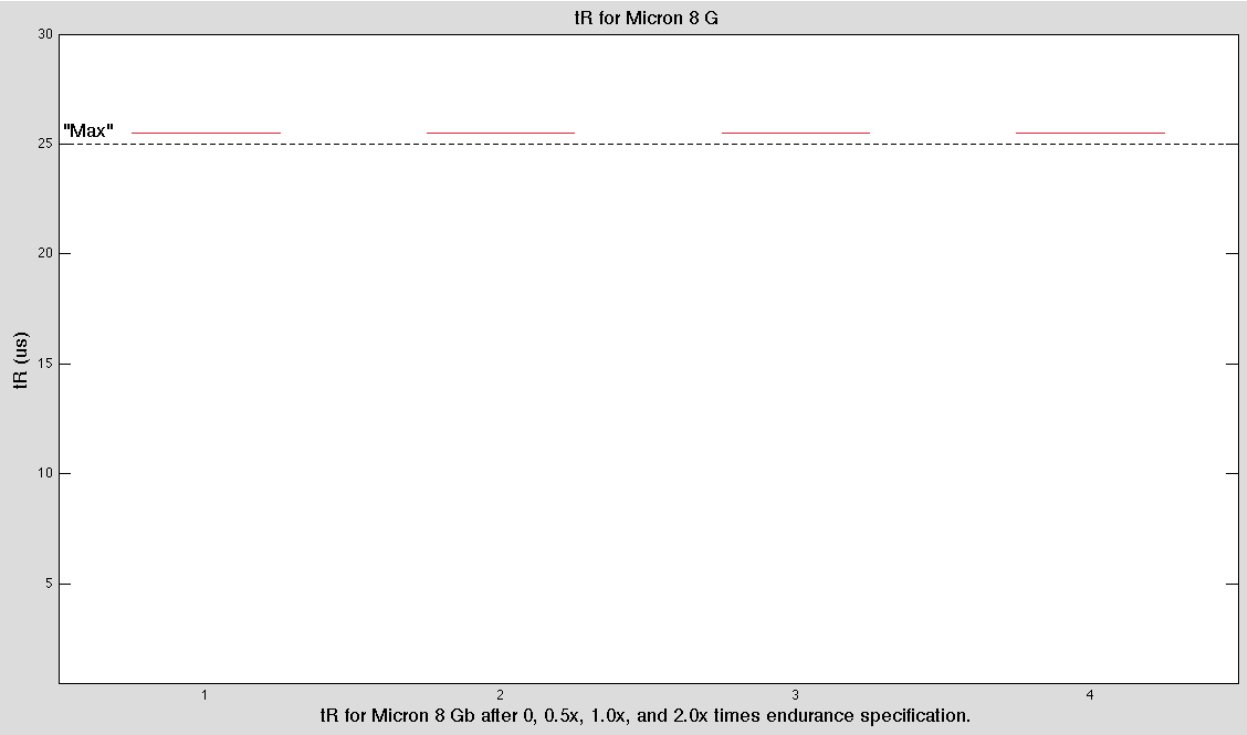

Figure 3.2-3. t_R for Micron 8 Gb after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

3.3 Samsung 8 Gb, MLC, K9G8G08UOM

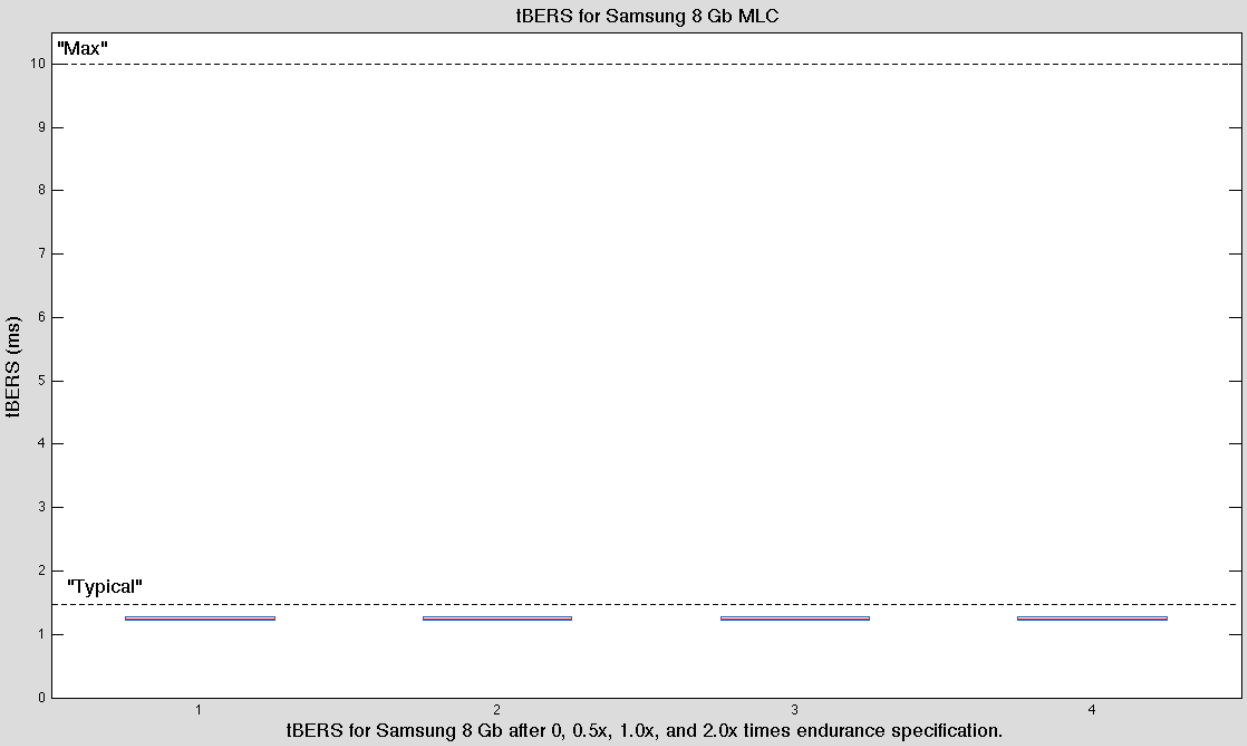

Figure 3.3-1. t_{BERS} for Samsung 8 Gb MLC after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

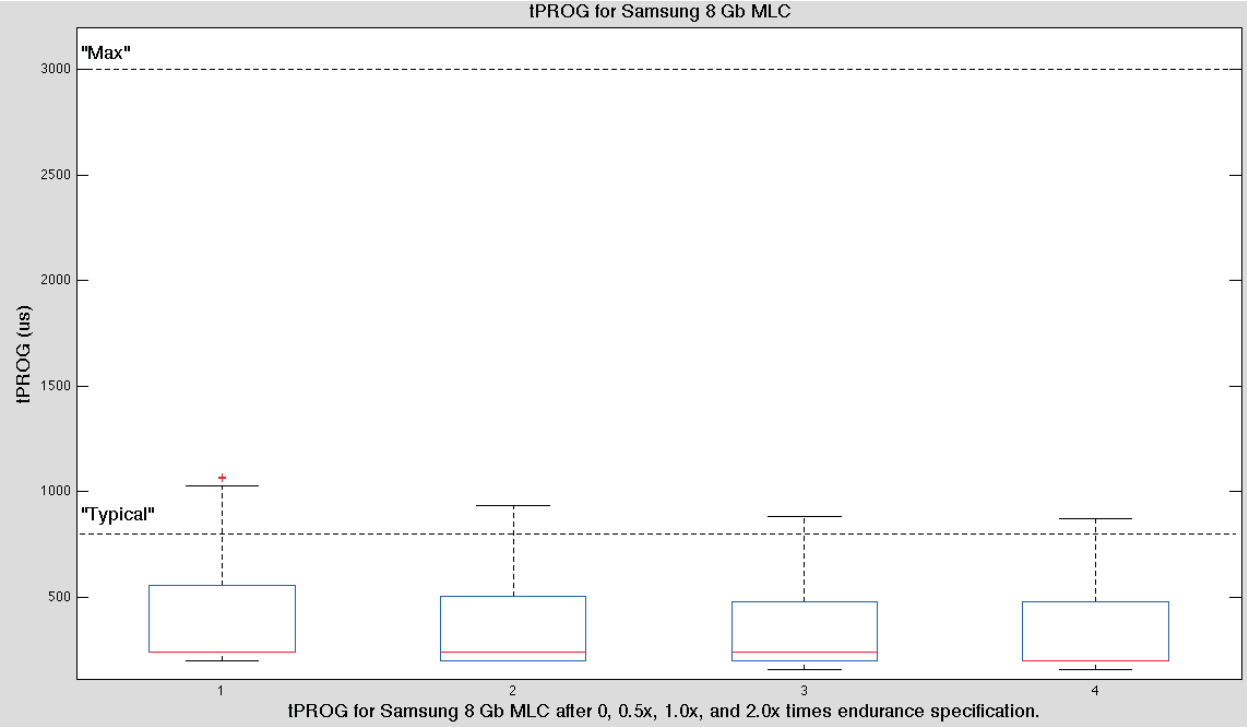

Figure 3.3-2. t_{PROG} for Samsung 8 Gb MLC after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

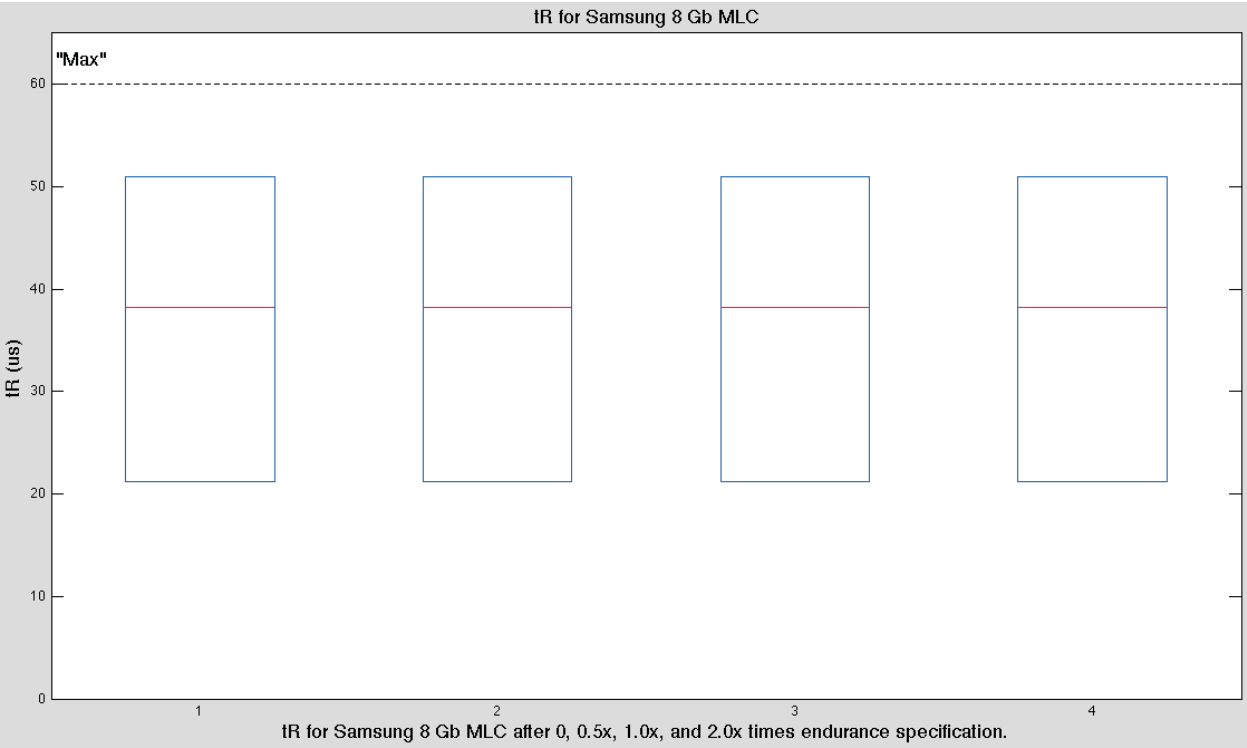

Figure 3.3-3. t_R for Samsung 8 Gb MLC after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

3.4 Samsung 8 Gb, SLC, K9F8G08UOM

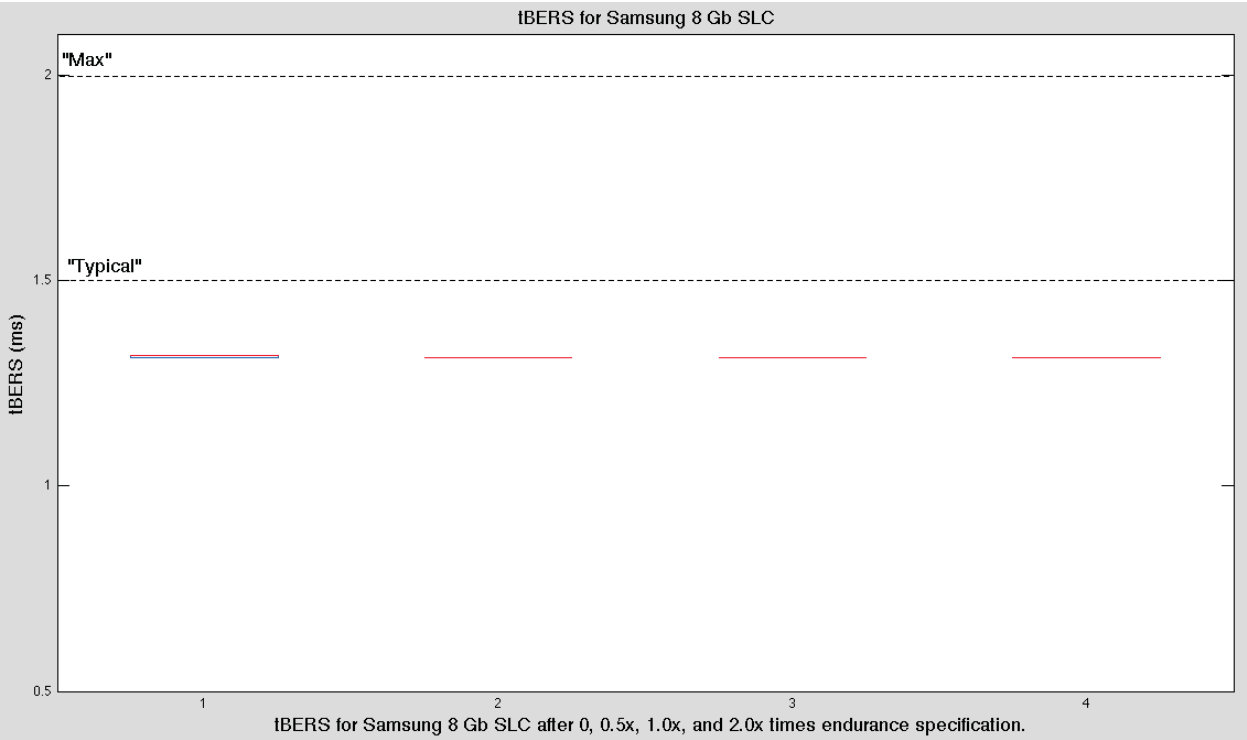

Figure 3.4-1. t_{BERS} for Samsung 8 Gb SLC after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

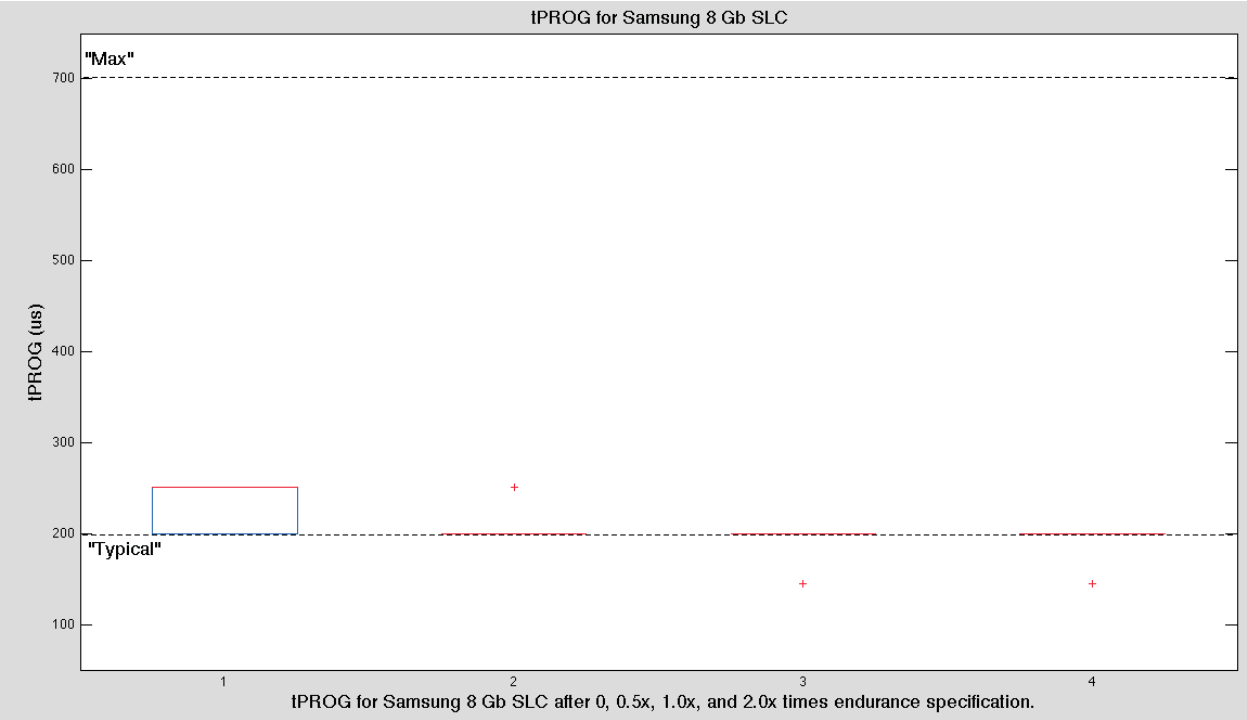

Figure 3.4-2. t_{PROG} for Samsung 8 Gb SLC after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

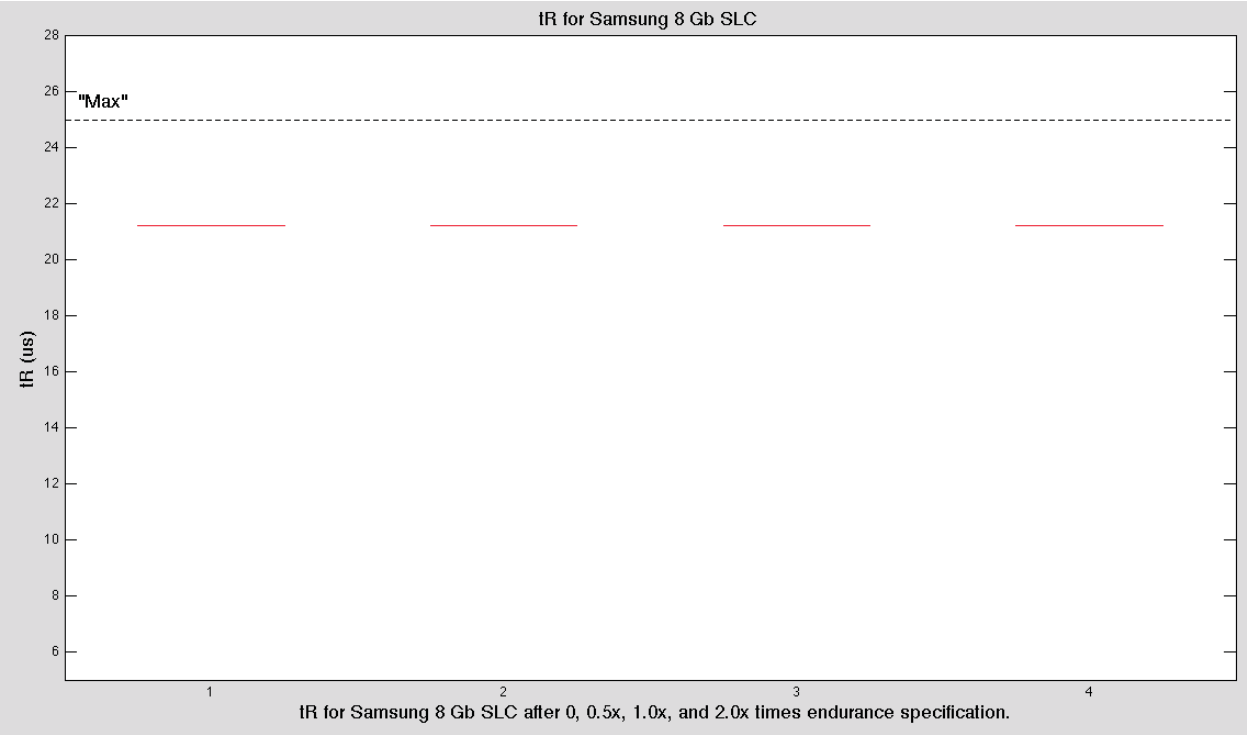

Figure 3.4-3. tR for Samsung 8 Gb SLC after 0, 0.5x, 1.0x, and 2.0x times endurance specification.

4.0 SUMMARY

Flash memory cells store their bits in floating gate memory cells. When these cells are neutral, they are said to be in a logical "1" state (erased). By adding charge (programming), the floating gates become negatively charged and are in the "0" state. Program and erase operations are done by Fowler-Nordheim tunneling, which utilizes high voltages to move charge on and off the floating gate, which is destructive to the gate oxide. Endurance cycling is meant to accelerate this failure mechanism. As the device consumes cycles, and defects are introduced to the gate oxide, it is expected that erasing and programming times would change. With cycling, bulk and interface traps are introduced in the gate oxide, which increases the threshold voltage of the cell [4]. It is therefore expected that erase and program times would change. However, because endurance cycling does not have an effect on how cells are read, read latencies should not change with cycling.

Table 4-1 summarizes the testing. Erase times increased, program times decreased, and read times stayed the same as expected. However, all three latencies seemed to not change with cycling in the Samsung devices. One could speculate that Samsung has introduced some kind of compensation circuitry to counteract the Fowler-Nordheim tunneling destruction in order to keep latency times consistent over the life of the product. Apparently, Micron has not done this. In the 32 Gb Micron device, erase times increased 50% and in the 8 Gb Micron device, erase times increased 300%. The 8 Gb Micron part also saw program times drop 80%, but t_{PROG} in the 32 Gb MLC did not change.

Table 4-1. Summary of Change in Latency Parameters with Cycling.

Part Number	t_{BERS}	t_{PROG}	t_R
MT29F128G08CJAAA	Increased	No change	No change
MT29F8G08AAA	Increased	Decreased	No change
K9G8G08UOM	No change	No change	No change
K9F8G08UOM	No change	No change	No change

5.0 RECOMMENDATION TO SPACE MISSIONS

All testing was done at room temperature and nominal V_{CC}. Because program and erase times can in fact change significantly, and are apparently device- and manufacturer-dependent, it is recommended that space missions with critical performance requirements (needing to use the devices with high duty cycle, i.e., programming and erasing without much time in between cycles) measure erase and program latency versus program-erase cycles at min/max V_{CC} as well as min/max temperature as part of the part qualification process.

6.0 REFERENCES

[1] Irom, F. et al., "Effects of Scaling in SEE and TID Response of High Density NAND Flash Memories," *Proceedings of IEEE Nuclear and Space Radiation Effects Conference* (2010).

[2] Irom, F. et al., "SEE Characterization and TID Response of Multi-Level 32Gb NAND Flash Memory," *Proceedings of IEEE Nuclear and Space Radiation Effects Conference* (2010).

[3] Micron datasheet, MT29F128G08CJAAA, 32gb_nand_mlc_l63a__2.fm - Rev. E 8/09 EN.

[4] Yang, H. et al. "Data Retention Characteristics of Sub-100 nm NAND Flash Memory Cells," *IEEE Electron Device Letters* (2003).