

pFLASH® Architecture Advantages

Technology Development, PMC

*“The Only pFLASH® Memory Product on Market Today with 40
Million Parts Shipped in Just Three Years” --- June, 2003*

Contents:

- Abstract
- Introduction
- 2T to Boost Flash Cell Conductivity
- No Program and Read Disturb
- No Sector Erase Disturb
- Fast Program Speed with Low Power
- Quality is Built-in with Designed
- Test by Performance®
- 100K Endurance is Guaranteed
- Data Retention is Intrinsicly Better
- Ideal for Code Flash and Embedded Flash
- Conclusion

Abstract

Two transistors (2T) PMOS cell is the quickest way to realize a reliable, fast programming, low power and low cost reprogrammable nonvolatile memory product. Only with the combination of 2T and P-Channel MOS device will result in disturbance free re-programmable nonvolatile memory. The absence of disturbance at both program and read cycles eliminate majority of the subtle reliability challenge for all of the other Flash technology in the industry today. This unique combination also enable the *Band-To-Band Tunneling* program which is the most efficient methodology for the least amount of reliability hazard related to endurance and data retention performance. The combination of *BTBT* program and *Channel Fowler-Nordheim (CFN)* erase not only result in the minimum power consumption but also the most scalable NOR Flash technology.

Introduction

Historically semiconductor devices with N- type of carriers are being used most due to their legacy reasons such as de-ionized water not widely available in the 1970s. The success of N-type EEPROM in the 1970s also led to its descending NMOS Flash technology which still prevailing in the 21st century. By the time when industry has its success on the CMOS technology, the advantage of PMOS Flash was uncovered in the mid-1980s. The fundamental physical mechanism of PMOS and NMOS Flash are almost the same, except the above mentioned reliability advantage of PMOS Flash. As can be seen in Figure 1, only electrons are being pushed in and out of the floating gate during program and erase operations.

2T to Boost Flash Cell Conductivity

In deep submicron technology, the deteriorating cell current became an overwhelming issue for all Flash technology immersing in high voltage complexity mixed with disturbance and programming strength. The ultimate solution of this issue could be resolved completely by 2T architecture. pFLASH, therefore, has extraordinary capability to maintain the cell conductivity by programming into deep depletion to increase the gate drive on the storage node. The select transistor will get around all the high voltage complexity so that the conductivity of the storage node is intact in deep submicron technology providing plenty of margins for the cell performance enhancement. The 2T scheme widens the program window tremendously in two ways.

First, so long as it is in depletion there are almost $8.4V = 10.5V - 2.1V$, $V_{program} - V_{read}$, of program window before running into the control gate leakage problem shown in Figure 2a. The deep depletion flexibility will convey cell conductivity in deep submicron technology. Second, after electrons start to accumulate on the floating gate it will reduce the floating gate voltage, therefore the injection rate of the hot electrons. Deeper depletion will result in tighter distribution for programmed cells due to this self limiting mechanism.

No Program and Read Disturb

The operation conditions of program and erase for this 2T PMOS cells are shown in Figure 2. The select gate will shield all the un-selected cells on the same bit line from the high programming voltage. So the program disturb issue becomes simple device isolation issue. The pFLASH architecture is designed to have floating gate programmed into deep depletion with perfect isolation so that the read disturb is avoided by the inherent design. Very small read bias is implemented to provide a reasonable read current at typical $12\mu A$ (see figure 2a). Figure 4 shows the lifetime for program disturb. The criteria for the lifetime is to assume a $0.1V$ shift in V_{cg} after the disturb operation. For the unselected cells along the same bit line the program disturb life time is about $1.2E14$ years at $120^{\circ}C$ after 10K cycling, while the read disturb life time for selected cells is about $6.9E7$ years.

No Erase Disturb

Since the *Channel Fowler-Nordheim* (CFN) erase is used by almost all major suppliers of EEPROM products, its advantage and reliability are well documented for the NMOS cells. As for the pFLASH cells, the erase disturb can be even lower simply because of the biasing design. As shown in Figure 2c, during erase, conducting channels formed underneath floating gates of programmed bits of all other unselected sectors connected to V_{cc} through the source junction. The deep depletion cell design will shield the floating gate from high N-well biasing at V_{pp} , such that the stress voltage be much lower than the conventional NMOS Flash. Therefore there will not be any erase disturb either.

Fast Program Speed with Low Power

As programming mechanism described above, fast programming, low power consumption, and low oxide damage are three distinctive advantages associated with the pFLASH cell architecture. Using super efficient *band-to-band-tunneling* as programming mechanism the injection efficiency, ratio of gate to drain current, is around 0.2% to 0.9% as shown in Figure 5. It is at least six orders higher than that of any conventional Flash cells using *channel hot electrons* for programming. Since the drain junction is biased negatively, electrons are being expelled toward the floating gate while the holes are attracted to the drain junction area. High tunneling efficiency implies low drain current needed for programming, which results even less hot holes through impact ionization. This mechanism sometimes is characterized as “hot holes free programming” as can be seen from Figure 6 that the program V_t monotonically decreased with cycling while no hot holes to fill up the hole traps generated by erase causing the erase window to be widened. In summary, the elegant biasing scheme of pFLASH® architecture offers the finest quality of Flash technology that eliminates disturbance, repels hot hole trapping, and achieves the most efficient Flash programming and erase capability.

Quality is Built-in with Design

All PMC products were designed with full production flow in mind, for example, many test and stress modes are designed in so the product and test engineers can screen out bad parts due to infant mortality with minimum test time. The circuit simulations are done with 20% of V_{cc} variation instead of 10% specified for V_{cc} range. The temperature range is -10°C to 110°C , though we only guarantee for 0° to 85°C . For data read operation, our devices will function up to 1.8V, though only specified for 3V. For ESD protection, we perform all three test modes (HBM, MM, and CMD) with typical results twice the minimum requirements. Our typical latch-up resistant level is above 200mA much higher than the required 110mA. These highly disciplined design targets lay the foundation for achieving superior product quality.

Test by Performance®

Once characterization data is available, we will decide a number of critical parameters to be monitored not only by product specs, but also by its normal distribution, whichever is tighter. Abnormal distribution parameters sometimes are the early

warnings of sorts. Either process parameters are drifting away such as CD or film thickness, or some severe degradation process is more than what our normal test flows can handle such as data retention or infant mortality. Once the alarm is triggered, the material will be either held for more thorough tests or scrapped if no effective test methods can be found. For either case, failure analysis will be performed to find out the root cause and subsequently appropriate actions will be taken to prevent any of the existing damage (out of distribution parameters) from becoming a major disaster. With this methodology we are able to guarantee our ship quality level to be well below 200 ppm for the past one and half years.

100K Endurance is Guaranteed

Figure 6 is the single cell cycling data. Figure 7a is a typical program Vt margin distribution for one 4Mb die before and after 50K cycling. Both indicate that the erase state read margin improves with cycling. It is an indication that the hole trapping generated during erase is negligible as that of EEPROM. Because program efficiency of pFLASH architecture is millions of times better compared to other convention Flash technologies, the pace of endurance degradation is much less. As shown in Figure 8, the worst program Vt margin of a 1Mb chip typically drops around 1.7V after 100K cycling without program verify, while 1.4V with program verify. The fact that all bits have 1.2V margin or larger once program verify is turned on is also an indication that the degradation is relatively minor such that the effects of oxide traps be overcome with longer program time. If another 0.3V Vt drop is added after high temperature bake at 150°C for 1000 hours as shown in Figure 9, the product will have at least 0.9V margin for read operation. With some due diligence we will be able to report endurance assurance greater than 100K in the near future.

Data Retention is Intrinsicly Better

The energy barrier of pFLASH cell is 4.14eV, while 3.04eV for NMOS cell. In other words, pFLASH cells are intrinsicly better than NMOS as far as data retention is concerned. However, in the real world, most retention failures are due to the extrinsic defects in the tunnel oxide area which are normally related to the wafer cleaning and defect control on the wafer surface during gate oxide growth and subsequent poly deposition process. Together with our intrinsic characteristic monitoring of every wafer

lot and test by performance(methodology, so far, we have not found any of failed returned parts to be associated with data retention issue literally, and we conservatively guarantee our parts for 20 years at 85°C.

Ideal for Code Flash and Embedded Flash

Another feature with this cell is that both BTBT program and CFN erase do not depend on the physical width of the stacked gate device. As the device dimensions shrink smaller the cell will function properly so long as the 2T devices are punch through free. The operation bias conditions need not be changed much due to the absence of various disturb mechanisms. The cell size reduction will rely mainly on the technology and no new program or erase mechanisms to be introduced. We will be able to shrink the cell once more advanced technology is available and we are committed to bringing out products with more advanced technology in a timely fashion to stay competitive. Furthermore, the advantages in this pFLASH architecture can be easily realized in the area of low density memories (such as code flashes), or where the ease of design and reliability are the main concerns (such as embedded flash, or known good die business etc).

Conclusion

Contrary to conventional wisdom, PMOS devices should be the natural choice for flash memories from reliable erase, fast program, and superior data retention perspectives. Additionally, the stable 2T architecture virtually eliminates subtle and complex physics issues such as over-erase, program disturbance, and read disturbance of 1T or 1.5T Flash. This superior architecture combined with highly disciplined design and test methodologies enable us to bring out a family of the finest quality pFLASH products with consistent reliability to the market.

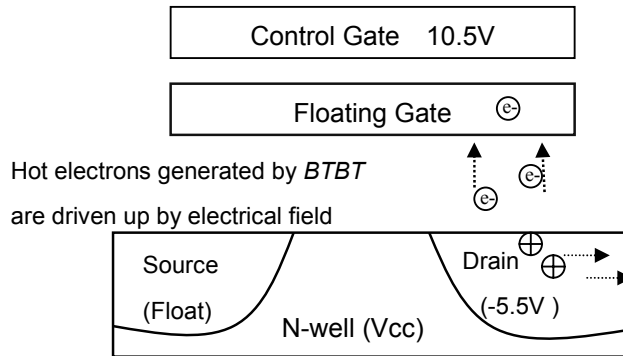


Figure 1a. Program biasing conditions. The electrons are generated by high bias in the drain junction underneath the gate. Hot electrons generated by *Band-To-Band Tunneling* are being extracted toward the floating gate while the holes toward the drain junction area. Please also note the device is not even conducting during programming, so very little power is consumed.

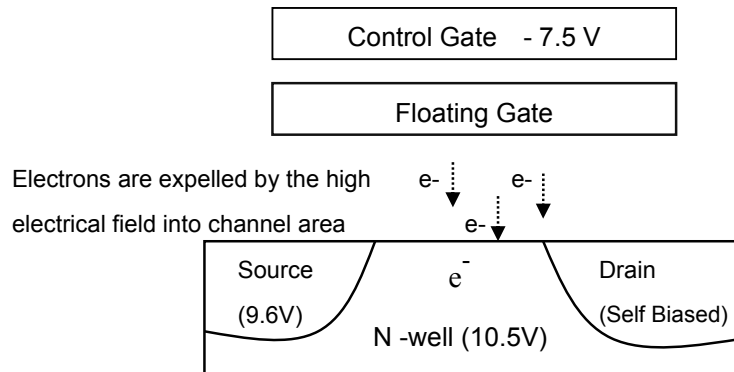


Figure 1b. The biasing conditions for erase. The electrons are extracted out of the floating gate into the channel area by *Channel Fowler-Nordheim (CFN)* tunneling. This scheme is being used by most of other flash memory vendors also.

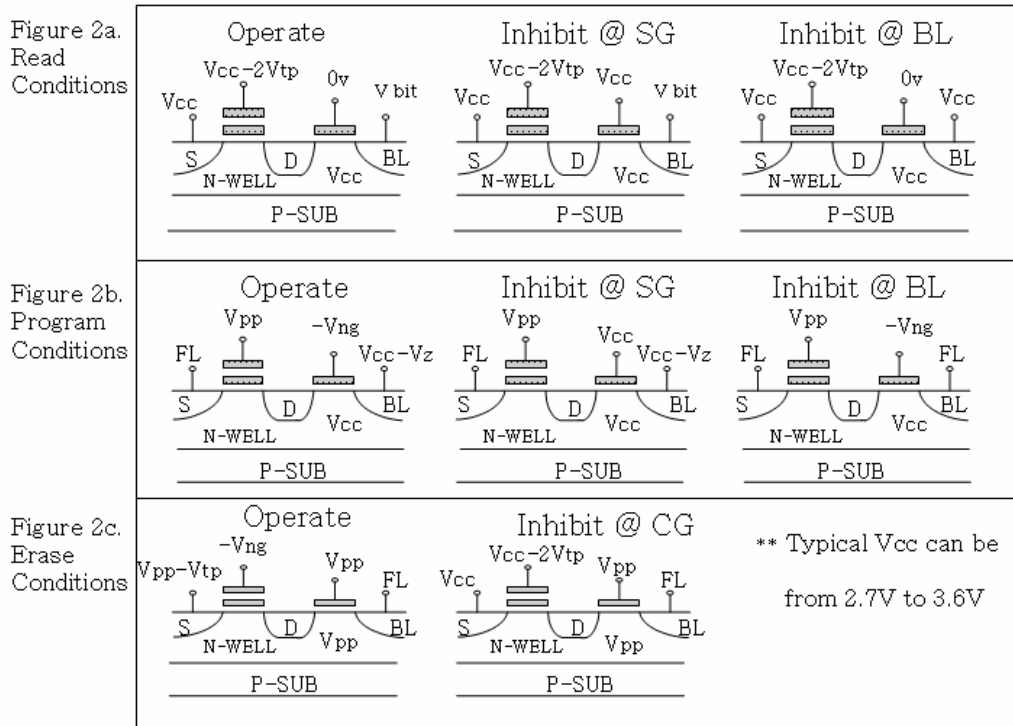


Figure 2. The Operation Conditions for the 2T PMOS Cells

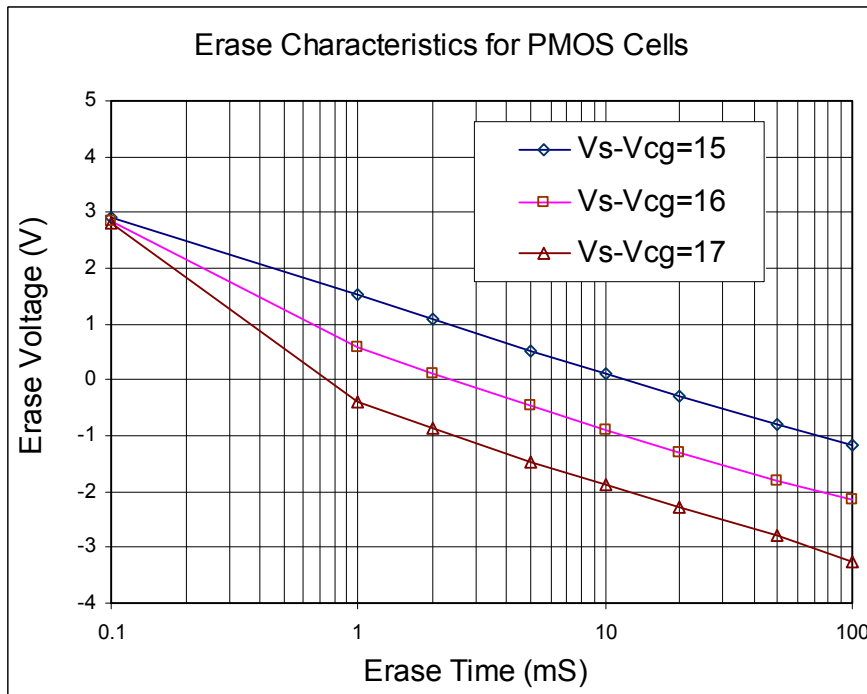
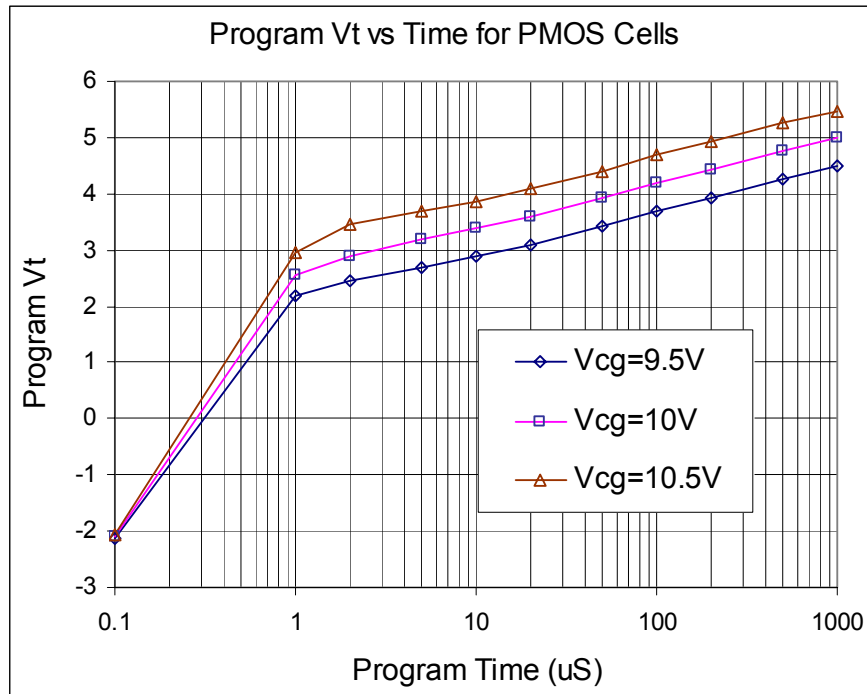
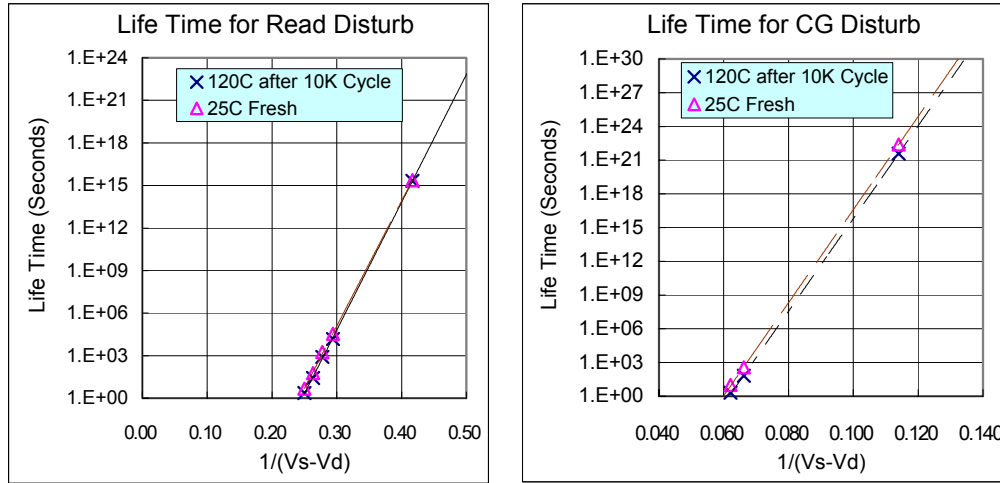


Figure 3. The Program and Erase Characteristics for the 2T PMOS Cells



Vcc	1/(Vs-Vd)	120C after 10K Cycle	25C Fresh	Vcg	1/(Vs-Vd)	120C after 10K Cycle	25C Fresh
5.2	0.250	2.3	4.8	19.3	0.055	2.00E-03	1.00E-02
5	0.263	27	60	18.3	0.058	5.00E-02	0.3
4.8	0.278	800	1800	17.3	0.062	2	9.8
4.6	0.294	15000	34000	16.3	0.066	65	370
3.6	0.417	2.2E+15	2.4E+15	10	0.114	3.8E+21	2.5E+22

Figure 4. The life time for read and program disturb for fresh dies at 25°C are 7.6E7 and 7.9E14 years respectively, while for 10K cycled dies at 120C are 6.9E7 and 1.2E14 years respectively. In another words, there will not be any read or program disturb associated with this 2T PMOS EEPROM cells.

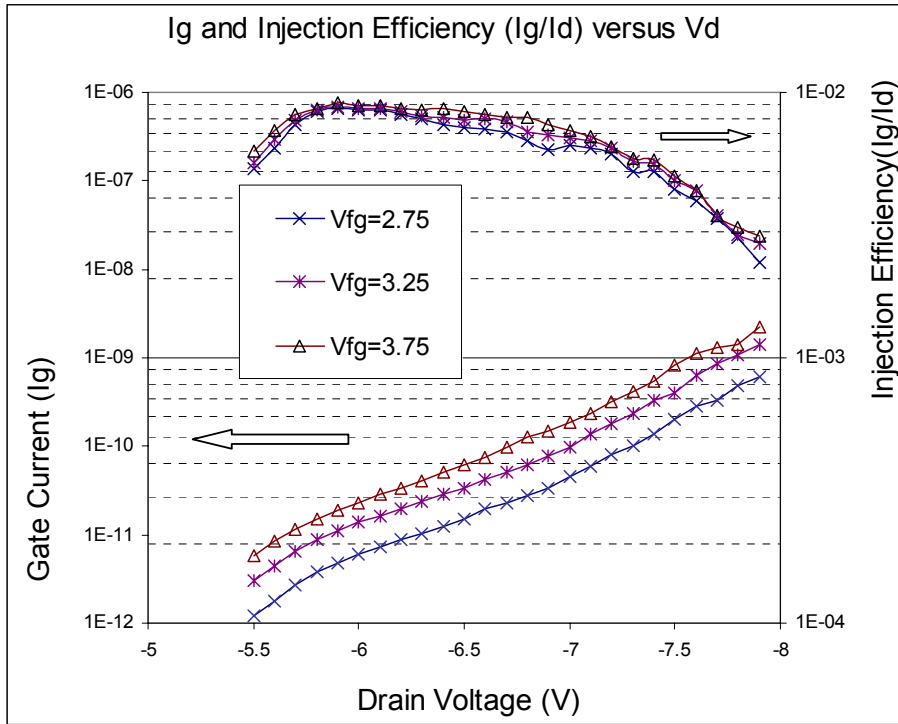


Figure 5. The drain current and program injection efficiency (defined as the ratio of gate to drain current) versus drain voltages.

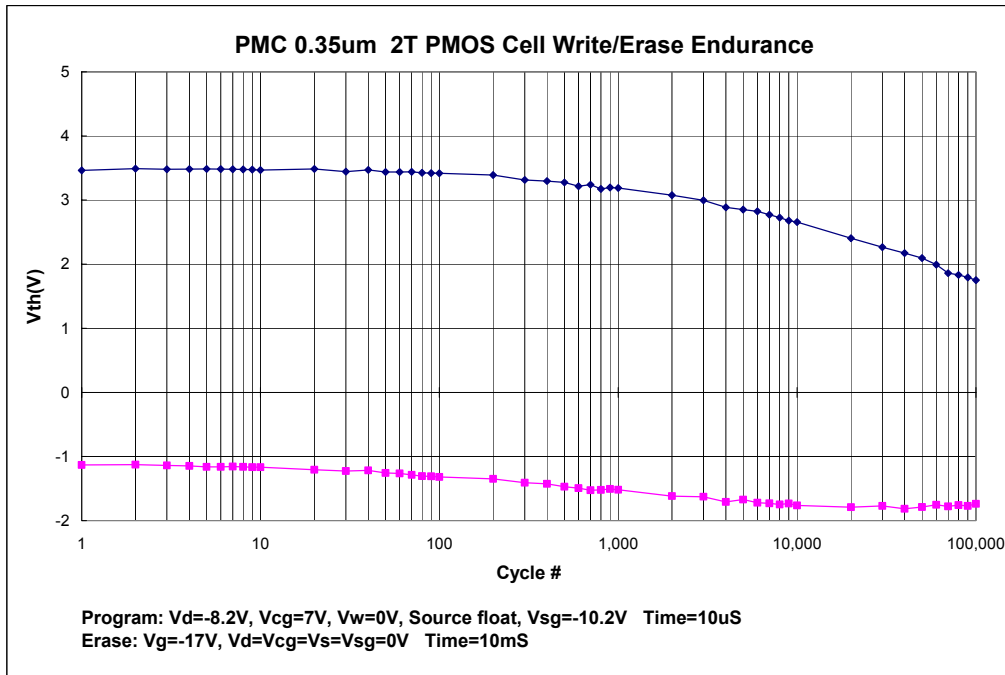


Figure 6. The endurance characteristics of the 2T PMOS Cells. It is interesting to see that the erase window improves through cycling. It is a clear demonstration of hot hole free programming, while only minor hole traps generated during erase and they will not be filled before 100K cycling.

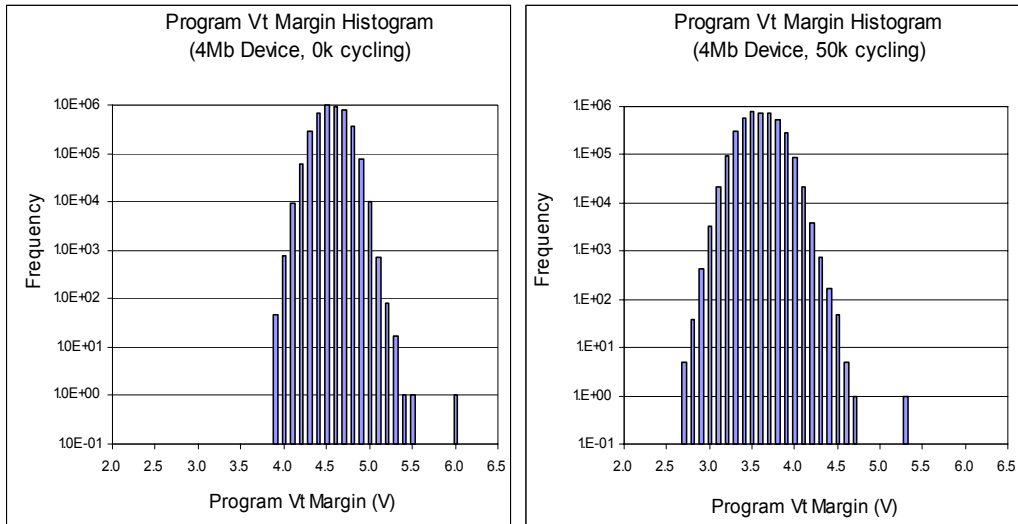


Figure 7a. Program Vt margin distribution for a 4Mb product before and after 50K endurance cycles. If $V_t < 0$, then it will report as read failure.

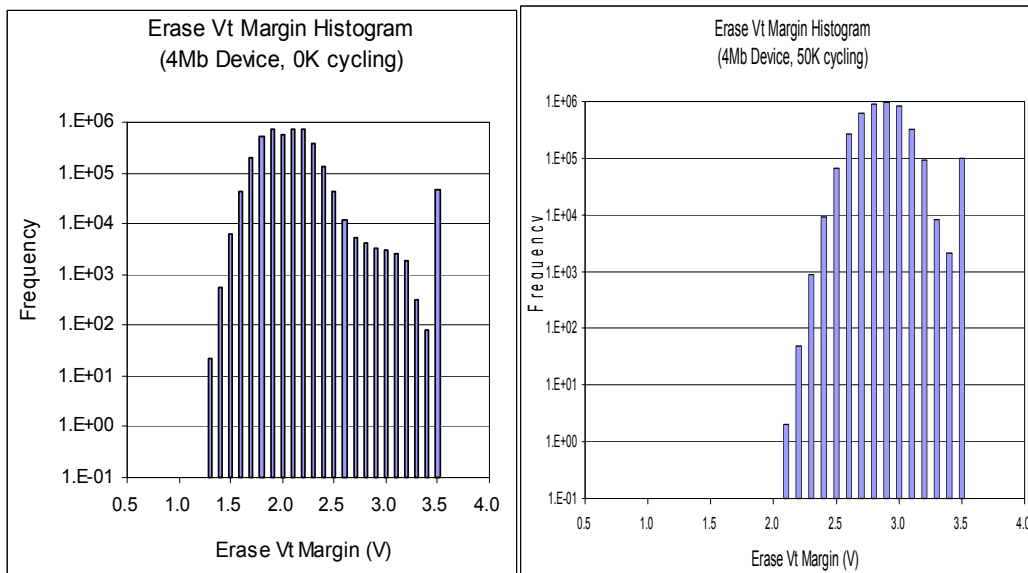


Figure 7b. Erase Vt margin distribution for a 4Mb product. V_t is defined as if $V_t < 0$, it will report as an erase failure. The peak at $V_t = 3.5V$ is due to test circuit cut-off. The erase margin actually improves with cycling, which also confirms the single cell data from previous figure.

Program Vt Margin	AVG	3 Sigma	Erase Vt Margin	AVG	3 Sigma
Fresh	4.56	0.463	Fresh	2.06	0.76
After 50K cycling	3.59	0.575	After 50K cycling	2.88	0.54
Change	0.97	Worse	Change	0.82	Better

Figure 7c. Summary table for both program and erase Vt margins for a 4Mb product before and after 50K endurance cycling.

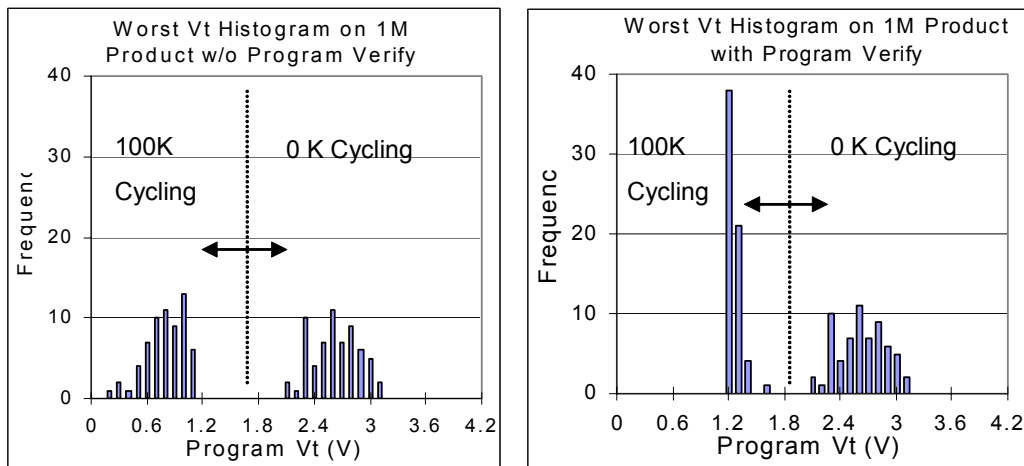


Figure 8a. Worst Bit Vt distribution for a 1Mb product after 100K cycling. Vt is defined as the voltage where the sense amplifier fails to read “0” if it is less than 0. In this product program verify is triggered at Vt=1.2V The program verified was performed after the regular 100K cycling.

Worst Vt on 1Mb Product With and Without Program Verify After 100K Cycling								
	0K	50K	100K	100K with PV	Shift (0-50K)	Shift (0-100K)	Shift (0-100K w/PV)	Shift (50K-100K)
Max	3.10	1.60	1.10	1.60	1.80	2.70	1.80	1.20
Min	2.10	0.50	0.10	1.20	1.20	1.50	0.90	0.20
AVG	2.62	1.14	0.79	1.25	1.48	1.83	1.37	0.35

Figure 8b. 100K Vt data.

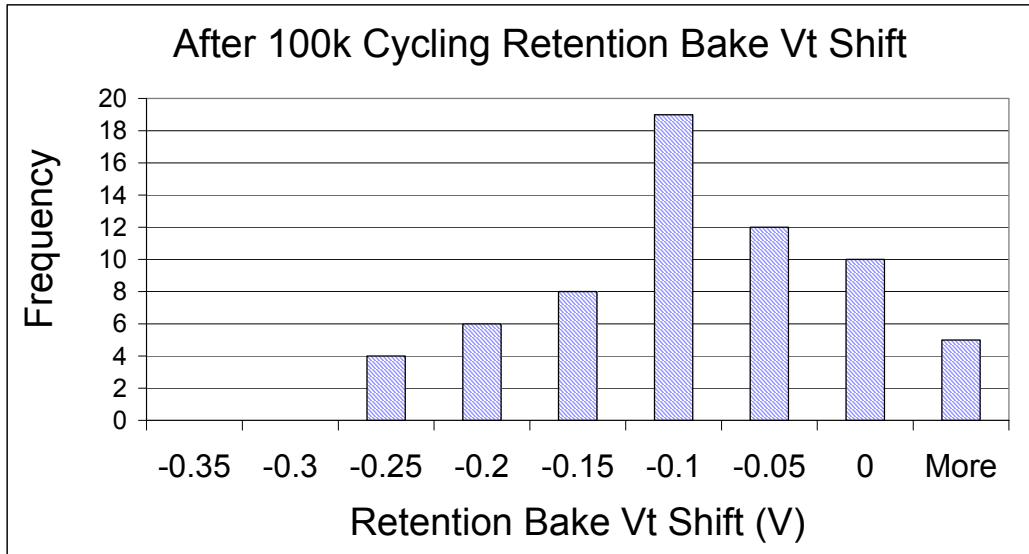


Figure 9. Worst Bit Program Vt Shift after bake retention test. All the samples were after 100K cycles endurance stress then put into the bake oven for 1,000 hours at 150°C bake oven for 1,000 hours at 150°C.