Extended Abstracts of the 1998 International Conference on Solid State Devices and Materials, Hiroshima, 1998, pp. 136-137

Invited

Reliability Issues of Floating Gate Flash Memory

Chi Chang
AMD, Inc. One AMD Place, P.O. Box 3453, MS: 177, Sunnyvale, CA 94088, USA
Phone: 408-749-3890, Fax: 408-749-3718, Email: Chi.Chang@amd.com

1. Introduction

The vast majority of Flash memory devices being shipped today are based on floating gate technology, with the exception of MNOS (or SONOS) devices which are limited in unit quantity as well as chip density, and the emerging FeRAM device still in its infancy. Among the different floating gate Flash devices, more than 90% of them use tunnel oxide as the dielectric, separating the floating gate and the 'channel' of the MOS transistor. Over the last few years, approximately two billion or more units have been shipped to customers. Major Flash suppliers have claimed device reliability equal to or better than 100K endurance cycles, with 10 year data retention. This impressive spec signifies the high reliability of tunnel oxide-based floating gate Flash memory. Understandably, this feat can only be achieved by a combination of good process technology, robust circuit design, a solid manufacturing environment and intelligent wafer sort and back-end reliability screening. However, the memory cell size is shrinking swiftly, such that the amount of stored charges distinguishing between a '1' state and a '0' state is rapidly decreasing. The industry's relentless push toward higher density, as well as multi level cell technology (MLC), further reduces the available sensing margin of the cell. As a result, it becomes ever more challenging to maintain such high levels of reliability. In this paper, we will review some of the reliability concerns with tunnel oxide Flash devices that are cycling induced, as well as how these concerns can be adequately addressed.

2. Background

Among the Flash memory devices using tunnel oxide as the gate dielectric, there are several types in existence employing different P/E operational methods. The leading method is the NOR Flash, utilizing channel hot electron (CHE) programming and source-edge Fowler-Nordheim (F-N) tunneling erase, with a tunnel oxide of 100Å or thicker. These devices are primarily used for code storage. The DiNOR, or AND Flash, belongs to another class of devices which employ F-N tunneling near the gate edge for programming and F-N tunneling through the channel region with an ~90Å thick tunnel oxide for erase. The main application is mass storage, but they may also be used for code. The NAND device, which is typically for mass storage applications (slow random access time) operates with channel F-N for both program and erase through a tunnel oxide of ~90Å. Generally, high field operation, along with thinner oxide dielectric is used for mass storage devices (high-density arrays), and requires EDC (error detection/correction) algorithms as well as 'bad sector' management to guarantee the reliability for cycling intensive applications.

Due to the various hot carrier effects(s) as the result of applying intense electric fields across the tunnel oxide during P/E cycling, charge trapping (as well as de-trapping), do occur inside the oxide and near the oxide-silicon or the oxide-polysilicon interface. Consequently, the electrical characteristics of the memory devices are altered in the course. If care is not taken, their reliability may be compromised. Reliability in this paper, most notably, is concerned with cycling endurance, charge retention and read disturb.

3. Channel Hot Electron Programming (CHE)

CHE is commonly used due to its good byte program speed and its seemingly benign impact on reliability, at the expense however of a relatively high current. Electron trapping at the drain side typically is not an issue, provided the drain engineering is properly performed and the channel hot electron injection is appropriately controlled. Over a million programming cycles can be achieved on high density NOR Flash, without seeing any noticeable write speed degradation due to charge trapping. Programmed Vt distribution, with the aid of intelligent circuit regulation, can be controlled very tightly. This is important, since systematic over-programming can produce sluggish erase bits [1] due to electric field over-stress of the tunnel oxide. Another example is that over-programming can degrade the channel mobility of the cell, due to surface state generation [2], causing bits to be slow to erase verify.

4. F-N Tunneling Erase

The 'discharge' of electrons from the floating gate edge (at the source junction overlap area) will be used as an example for this discussion. Hot carrier effects, such as hot holes generated by the tunneling electrons (in the bulk oxide or at the anode), or hot holes originated from the band-to-band tunneling mechanism, can all participate in the oxide charge trapping phenomena, not to mention that a steady electron trapping inside the oxide is also occurring and is often compounded by the trap-state generation event going on in parallel. These events, which are dynamically taking place in the oxide (i.e., electron current flow, hole generation and hole diffusion toward the cathode [3], and charge trapping/de-trapping) alter the electrical characteristics of the tunnel oxide, hence possibly causing issues related to endurance, read disturb, and charge retention. In addition, operating with a constant field to avoid high peak field together with a mild erase speed has been shown to enable 1 million P/E cycles with essentially no degradation in erase time [4]. A constant field operation presumably also benefits other aspects of Flash array reliability.

5. Erratic Bits

A supposedly normally behaved sub-population of bits in a high-density array can suddenly exhibit accelerated erase characteristics and can become over-erased [5.6]. If special provisions are not made, this can be detrimental to the cycling endurance of the one-transistor cell array typically used for NOR Flash. This phenomenon has been closely studied and it has been found there is actually no 'permanent' damage suffered by the tunnel oxide associated with the erratic bits. The fast erase behavior of the erratic bit often disappears during cycling or after the bit is artificially 'refreshed,' such as subjecting it to a high temperature bake (250°C or above). It is also often recovered after the bit is exposed to some UV light. A model based on certain cluster formation of trapped positive charges inside the oxide near the polysilicon-SiO_2 interface has shown (by computer simulation) to result in a strong locally enhanced electrical field, hence the erase current [5,6], as shown in Fig.1. A molecular model explaining the nature of the oxide trap centers capable of trapping positive charges has also been proposed [7].

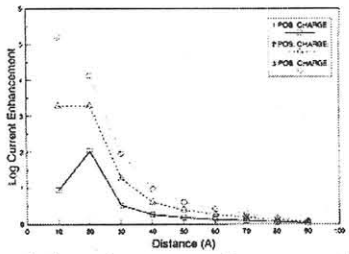

Fig. 1 Local current enhancement [6]

A commonly used method to 'repair' erratic bit over-erasure is through 'soft programming' [8] of the over-erased bits, and another is through a 'self-convergence' technique [9]. Generally speaking, the overhead time incurred in 'soft programming' to bring the over-erased bit(s) to a sufficiently positive Vt, is relatively short in comparison with the total array P/E time, provided that the erased array Vt distribution is sufficiently well behaved to begin with. If the 'self-convergence' technique were to be used in tightening the erase Vt distribution, it needs to be properly controlled, since co-injection of electrons and holes can often lead to permanent damage to the cell.

6. Read Disturb

Analogous to the erratic bits during erase, under a static read condition some erased bits in an array are observed to gain electron charges much faster after P/E cycling. This phenomenon has been investigated thoroughly [6,10,11] and again, it is shown to be related to positive charge trapping which induces a barrier lowering effect for electron injection. For the P/E induced read-disturb 'flier' bit, it also has been observed to often revert back to the typical bit characteristics after some additional cycling [11], thus bearing similar 'randomness' behavior as the aforementioned erratic bit. It is believed that the same model (local positive charge cluster formation) can also be applied to explain this phenomenon, and perhaps in this case the cluster of charges is located closer to the SiO_2-substrate interface instead. Various measures by process and design are taken to suppress its occurrence and the magnitude of it. Fabrication process, such as the wafer front end processing methodology [12], can be improved to suppress this enhanced read-disturb behavior. Circuit design, such as the optimization of P/E operation conditions, minimizing read disturb time, and reducing disturb field all need to be considered carefully to manage the read disturb issue.

7. Charge Retention

Similar to charge gain, a sub-population of the array bits exhibit enhanced charge loss rate under their own internal oxide field after P/E cycling. Analogous to the erratic bits and read-disturb bits, a number of 'flier' bits with enhanced charge loss rate increases with increasing P/E cycling [12]. The enhanced charge loss characteristics can disappear when the cell is given a high-temperature bake (125°C or above) [12]. They are cured presumably either by thermal de-trapping of the trapped positive charges, which are responsible for the enhanced oxide leakage via tunneling mechanism, or through some charge recombination mechanism. The 'instability' or the 'self-curing' effect of this local oxide leakage can even take place at room temperature [12]. Again, proper process optimization and good cell engineering, as well as the help of circuit design techniques, should be carefully considered to minimize this potential reliability issue.

Conclusion

The P/E cycling induced erratic behavior of bits can cause concerns for Flash reliability. Care should be taken to contain these potential problems. A classical thermionic emission process alone will no longer explain the mechanism(s) responsible for this erratic behavior, but the tunneling model (with a reduced barrier height as well as with a small thermal activation energy) associated with positive charge trapping, appears to give a better description. An accelerated high temperature bake study (for charge retention) can actually mistakenly lead to an incorrect interpretation of device reliability. This phenomenon also imposes a severe limit to tunnel oxide thickness scaling. As a result, it will be a real challenge to reduce the high voltage required for P/E for floating gate Flash memory.

Acknowledgements

I would like to express my thanks to Sameer Haddad and NVT/NVT Engineering, Fujitsu, and FASL for their contributions to this paper

References

1) J. Chen., et. al.: Proc.13th NVSMW(1994)
2) C. Chang: Flash Reliability (tutorial),Int.Rel.Phys.Symp. (1993)
3) I.C. Chen, et. al.: Appl. Phys. Lett. 49 (1986) 669
4) R. Bez et. al.: EDL-19 (1998) 37
5) T.C. Ong, et. al.:VLSI Tech. Symp. (1993) 83
6) C. Dunn, et. al.: Proc.Int.Rel.Phys.Symp.(1994) 299
7) C. Kaneta: Jpn. J. Appl. Phys. 35(1996) 1540
8) M. Van Buskirk, et.al.: U.S. Patent #5359558
9) S. Yamada, et. al.: IEDM Tech. Dig. (1991) 307
10) A. Brand, et. al.: Proc.Int.Rel.Phys.Symp.(1993) 127
11) S. Yamada, et.al.: Proc.Int.Rel.Phys.Symp.(1996) 108
12) F. Arai, et. al.: Proc.Int.Rel.Phys.Symp. (1998) 378