Extended Abstracts of the 2000 International Conference on Solid State Devices and Materials, Sendai, 2000, pp. 276-277

C-5-1 (Invited)

Technology Considerations for High-Speed High-Density Embedded Flash

Ko-Min Chang

Non-Volatile Memory Technology Center, Transportation Systems Group, Semiconductor Products Sector, Motorola
6501 William Cannon Drive West, MD: OE341, Austin, Texas 78735, USA
Phone: +512-895-8352 Fax: +512-895-2722 e-mail: ko-min.chang@motorola.com

1. Introduction

Consumers want the best. They want to live in comfortable homes, communicate with handy gadgets, commute in intelligent vehicles, and work in efficient offices. The explosive productivity of the electronics industry has fueled the demand of the consumers. PC-driven growth belongs to the last century. Wireless communications and consumer electronics now drive the demand of electronic components. System-on-a-chip (SoC) is a logical response by the IC industry to the market demand [1]. The inclusion of high-speed, high density embedded flash memory allows in-system reprogrammability along with better performance, lower power, less interference and reduced system cost [2].

2. Matching Technology with Market Requirements

Wireless communications and automotive electronics are two of the major markets Motorola serves. With the wireless phone industry progressing from 2G through 2.5G to 3G (with wireless multimedia capability) and the automotive industry preparing to transform from hydraulic-driven to electrical motor-driven systems, the push for high-speed, high density embedded flash is unmistakable. Currently, an embedded flash IP block is being prepared which delivers up to 2.0GB/s throughput to feed an on-chip processor. Traditionally, MCUs with embedded flash that run on 10-20 MHz clock have been the volume driver for a variety of applications. In our opinion, the technology and the choice of flash cell for these markets don't need to be the same as those that require high speed and high density [3]. In fact, we have chosen to outsource a good fraction of the medium-speed volume to foundries.

The first generation embedded flash Motorola introduced was based on a 1.5T source-coupled split-gate (SCSG) cell that delivered medium speed and density. Subsequently a 2T flash with a source-side select gate was introduced to address the speed issue at low supply voltages [4]. The first sub-40ns flash was introduced to feed a 32-bit RISC MCU with density up to 512K bytes [2]. Recently a cell based on an AMD 0.25um stand-alone flash was successfully integrated to a high-performance logic platform with shallow trench isolation (STI), cobalt salicide, and tungsten local interconnect [5]. Products with sub-25ns and density up to 1M bytes are being sampled to key customers. Cu-based interconnect system will appear on sub-0.18um embedded flash technology and there doesn't appear to be any show stopper [6].

3. Cell Selection for High Speed and High Density

The technology requirements for high-speed and high-density are, in fact, conflicting. Moreover, most of SoC's with embedded flash also require low active and standby power. The key to high-speed design is to have sufficient cell current under the worst-case operating conditions, e.g., low voltage or high ambient temperature. Given the fact that the tunnel oxide thickness will remain largely constant for the foreseeable future, due to the anomalous SILC in tunnel oxide, the intrinsic transconductance of a scaled flash bitcell will not be much better than that of previous generation. Increasing the cell width works against the small size necessary for high density. The biggest factor that controls the read current, then, is in the gate drive. The worst-bit gate drive in an array is defined by the difference between the word line voltage (Vwl) and the highest Vt (Vt,hi) of the bit population. As illustrated in Fig.1, for n-ch flash cells, the choice will be influenced by the granularity of the discharge block:

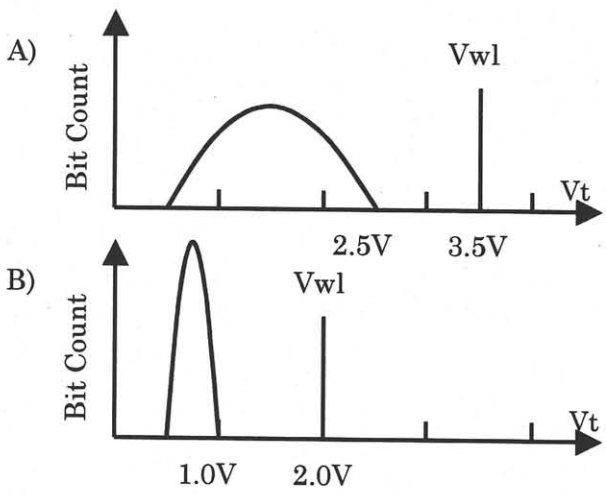

Fig. 1 Placement of word line voltage, Vwl, to obtain 25uA read current: A) ETOX-like cell, B) DINOR-like cell.

ETOX-like (large discharge block) Cells

(see the other side)

Page 1, Chang (Page number and author's last name with a light pencil)

Since hot-electron injection is used for programming, the relatively slow Fowler-Nordheim erase is typically performed in 64K-byte blocks to achieve the erase throughput. Even with compaction techniques, a typical Vt distribution is around 2V, which places Vt,hi at 2.5V. To have 25uA read current necessary for high-speed access, the word line voltage Vwl needs to be above 3.5V, which is well above the supply voltage of 1.2-1.8V used in sub-0.25um logic designs.

DINOR-like (small discharge block) Cells

Sector erase to higher Vt is achieved with positive control gate voltage. Programming towards lower Vt can be performed on 64 bytes or less at a time. This allows the bit-by-bit verification technique be used to tighten up the programmed Vt distribution to less than 0.5V. To have the same 25uA read current, the word line voltage Vwl needs only be above 2.0V, which is much closer to the supply voltage range.

For 2.5V-3.3V single-supply applications, the DINOR-like cells don't need to have boosted or pumped word lines, a clear advantage for low power. However, if an additional supply is available in the system, low standby power can also be had for the ETOX-like cells. This goes to illustrate that the final cell selection is highly dependent on the application space and system definition. There is not a universal set of criteria to guide the decision making process and it underscores the importance of having clear communications among the stakeholders.

4. Cost Pressure

Embedding flash IP adds complexity to the silicon fabrication process. The resultant system chip is bigger and will produce less number of good die per wafer. To reduce the die cost, aggressively scaling the area occupied by the flash IP block(s) is a necessary step. However, due to the inability to scale the tunnel oxide, the area of the flash IP block does not scale easily from one generation to the other. Array Efficiency (AE), defined as the percentage of total bitcell area to the total IP block area, is a convenient figure of merit for IP block size estimation. It is not uncommon to see the range of AE be 30% or lower for high-speed embedded flash IP blocks, which is much lower than that for a high-density stand-alone flash (AE>60%). Moreover, a high-speed, high-density flash IP block seldom occupies greater than 50% of the chip area. This has an interesting but important implication on cost: the chip size is relatively insensitive to flash cell size for high-speed, high-density embedded flash.

5. A Different Cost Reduction Path – Vpp Scaling

Instead of following the aggressive shrink path the stand-alone flash vendors set out for the flash cell size, which may have an adverse effect on wafer cost, the high-speed, high-density embedded flash IP providers must find a different cost reduction path. In our opinion, continual cost reduction can be achieved through Vpp scaling. Figure 2 shows the trend of various device oxides as embedded flash technology scales:

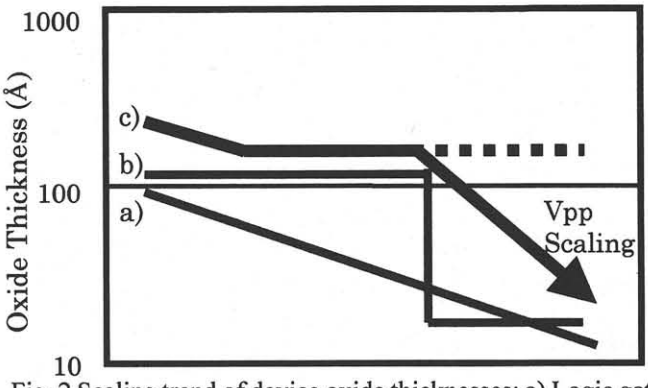

Fig. 2 Scaling trend of device oxide thicknesses: a) Logic gate oxide, b) Tunnel oxide, c) Vpp gate oxide. The arrow indicates a break from the traditional cell scaling path.

Vpp scaling can be realized in a number of ways. The more immediate implementation can be based on storage devices with non-conducting storage media, such as silicon nitride [7] and nanocrystaline silicon [8]. These devices are less sensitive to the anomalous SILC so the tunnel oxides can be aggressively scaled. With split-biasing, the magnitude of Vpp can be reduced to 4V. Over the horizon, magnetoresistive RAM (MRAM) [9] will push Vpp to below 2V with excellent performance.

Acknowledgments

The author would like to thank the NVM technology community within Motorola for stimulating discussions and the product groups for challenging business objectives. Special thanks to Paul Ingersoll for shaping the final version of the manuscript.

References

[1] S. Kohyama, *IEDM Tech. Digest, 1999*, p.8-13.
[2] C. Kuo et al., *Int. NVM Tech. Conf., 1998*, p. 28-33.
[3] K. Yoshikawa, *VLSI-TSA,1999*, p. 183-186.
[4] W.-H. Liu et al., *15th NVSM Workshop, 1997*, p.4.1.1
[5] D. Burnett et al., *NVSM Workshop, 2000*, p 59-61.
[6] C.-L. Chang et al., *NVSM Workshop, 2000*, p.62-64.
[7] I. Fujiwara et al., *NVSM Workshop, 1998*, p. 98-100.
[8] S. Tiwari et al., *Appl.Phys.Lett. 68, Mar/96*, p. 1377.
[9] M. Durlam et al., *ISSCC Tech. Dig.., 2000*, p.130