A-5-1 (Invited)

Nonvolatile Memory Challenges toward Gigabit and Nano-scale Era and a Nano-scale Flash Cell: PHINES

Sam Pan, C.C. Yeh, Rich Liu, and C.Y. Lu , Macronix International Co. Ltd.; email: sampan@mxic.com.tw

Abstract

Our analysis of all the existing Flash technologies indicates that none of them is ready for scaling into the deep nanometer regime. The floating gate NOR will hit the limit at 65nm, and NAND scaling probably cannot go beyond 45nm. As for nitride electron storage Flash, however, there seems no fundamental limiter although ONO quality and smarter cell operation scheme are still of significant technological challenges. Among many nitride Flash technologies, we propose a newly invented PHINES cell which is a promising Flash technology in nano-scale era.

Introduction

According to the ITRS Flash roadmap [1], to make 1Gb NOR Flash mass production worthy, 90nm node with 2bit MLC or MBC is required assuming 1cm^2 die size, and the timing is around 2005. Ironically, from the technological feasibility standpoint, the last technology node of floating gate NOR Flash was predicted to be 65nm, based on extrapolation of the difference between physical and electrical cell dimensions vs. generations, which drops to zero at 45nm [2]. On the other hand, lithography capability below 65nm will probably come sooner than later, driven by logic technology. How to take the advantage of lithographic advancement and push Flash technology beyond 65nm becomes an urgent issue. In this study, we will re-examine the fundamentals of Flash scaling and suggest a direction for future Flash research and development.

Floating gate Flash scaling considerations

There are two major difficulties in scaling floating gate Flash. The first is on the non-scalability of tunnel oxide and inter-poly ONO due to reliability concerns. Without the vertical scaling, the horizontal shrinkage is difficult. Consequently, cell size reduction in past several generations is mostly accomplished by the reduction of passive portion of cell area such as field isolation, bird's beak, wordline spacing and diffusion overlap. In the case of NOR Flash, Figure 1 shows the trend of cell gate length, cell size and product density since 1um. The corresponding process features are shown Table I. Note that the process complexity increases dramatically to realize 0.13um, and the trend probably continues. Table 2 shows the ITRS flash roadmap. It can be inferred that for technology nodes beyond 50nm, more than 50% of cell area is occupied by active transistor, which is almost twice the historical value of 25%. In other words, more process breakthroughs are needed to cut down the cell passive area. The dismal NOR scaling scenario is mainly caused by hot-e programming, which imposes the limit of the cell gate length, probably around 0.13um [3]. The constraints of NOR gate length scaling are drain turn-on leakage, band-to-band leakage, read current, and programming speed. Fig. 2 shows the process windows for 0.15-0.1um gate length. Based on the NOR cell gate length trend with a scaling factor of 0.7, 65nm node may need 0.11um gate length. This means at 65nm node, device operation window may already be very small, if exists at all. Another fundamental limitation is caused by the floating gate itself. Owing to the stray capacitance coupling, gate coupling ratio will drop dramatically when the wordline spacing is smaller than 40nm even under an ideal scenario of scalable floating gate thickness [4]. It is very likely that the practical limit of NAND flash scaling is at 45nm technology node.

NROM Flash scaling considerations

NROM Flash is to utilize localized electron trapping for physical two bits per cell nonvolatile memory [5]. Programming is accomplished by conventional channel hot electron trapped at drain edge, and erase by band-to-band hot hole. The novel reverse read with high enough drain bias depletion to eliminate the influence of trapped electrons on carrier conduction, gives rise to two bit operation. It is a neat Flash cell with simple process architecture. The device and reliability physics of NROM cell is, however, very complicated [6-9] with new phenomena still being understood. From the scaling perspective, although NROM is free from the floating gate issues, disturb problems of various sorts, especially, read disturb, are potential show stoppers. The general concern for two bit separation is manageable based on existing knowledge.

PHINES: a promising nano-scale Flash cell

To overcome the shortcomings encountered in floating gate flash and NROM, a novel Flash cell named PHINES (Programming by hot Hole Injection Nitride Electron Storage, pronounced as "finesse") has been proposed[10]. PHINES uses an MXVAND cell structure arranged in a buried diffusion virtual ground array similar to NROM as shown in Fig. 3. But as indicated in Table 3, PHINES operation is different from any SONOS type cells. First of all, fresh cell should be properly engineered with low Vt. The erase is performed to raise Vt by channel FN tunneling under positive wordline bias. Programming is done by lowering local Vt through bit-by-bit band-to-band hot hole injection, and thus, two bits per cell can be realized as shown in Fig. 4. In Fig. 5, programming of the adjacent cell sharing the same bitline (drain) and wordline (gate) is inhibited by properly biasing the unselected bitline (source), and likewise, the cells sharing the same bitlines but with a different wordline also have enough program disturb margin. The dramatic reduction of band-to-band current by biasing the source is crucial for preventing program disturb in the virtual ground PHINES array. Furthermore, it can be demonstrated that this source modulation effect is more pronounced for a shorter channel length, a good indication of scalability. PHINES read operation window improves with a shorter channel length as well, and consequently, drain bias during read can be reduced to improve read disturb. Since programming speed is very fast with low programming current, PHINES is suitable for parallel programming. From the performance standpoint, PHINES can serve both code and data Flashes. Fig. 6 shows good single cell cycling performance. As far as charge retention is concerned, PHINES should be better since hole induced damage is minimized due to the fact that only a small amount of hot hole is injected into ONO through a bit-by-bit verification scheme.

Conclusion

Flash technologies utilizing nitride electron trapping have the potential to dominate in future generations although there still exist technology bottlenecks such as better ONO and better cell operation design. And we first propose that the newly invented PHINES cell is one of the few promising candidates for nano-scale Flash.

References

[1]Process Integration, Device and Structures and Emerging Research Devices, ITRS, p. 14, 2001 edition,

[2]Stefan Lai, Future Trends of Nonvolatile Memory Technology, Intel press release, December 2001.

[3]T. H. Fan et al, to be published.

[4]J. Lee et al, NVSM workshop, p. 90, 2001.

[5]B. Eitan et al, SSDM, p.522, 1999.

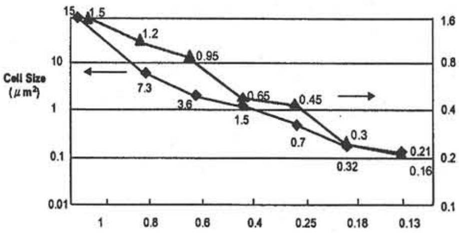
[6]W.J, Tsai et al, IEDM Tech. Digest, p.719, 2001

[7]B. Eitan et al, SSDM, p.534, 2001

[8]Y. Roizin et al, NVSM Workshop, p. 128, 2001.

[9]W.J. Tsai et al, IRPS, p. 34, 2002.

[10]C.C. Yeh et al, to be published; US patent pending.



Production Year	88	91	93	96	98	00	02	
Largest Product	4M	8M	16M	32M	64M	128M	256M	Die size 80-120mm^2
Product Sweet Spot	2M	4M	8M.	16M	32M	64M	128M	Die size 40-60mm2
Cell Size (F2)	15	10.9	10	9.4	11.2	9.8	9.5	

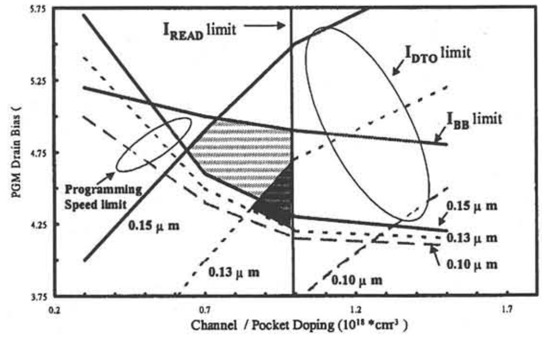
Fig 1. Demonstrated ETOX Code Flash Cell Size/Cell Gate Length

Process\Tech Node	1µm	0.8µm	0.6µm	0.4µm	0.25µm	0.18µm	0.13µm
Tunnl ox (A)	120-110	120-110	110-100	110-100	100-90	100-90	100-90
Interpoly dielectric (A)	300-250	250-200	200-180	180-160	160-140	160-140	160-140
Weff/Leff (µm)	0.6/0.4	0.5/0.35	0.45/0.32	0.35/0.25	0.29/0.2	0.19/0.14	0.13/0.1
GCR	0.6	0.55	0.55	0.6	0.6	0.6	0.6
Iread (uA)	100-75	75	60	50-60	50-60	50-60	50-60
Self-align source	NO	NO	Yes	Yes	Yes	Yes	Yes
STI/SA-STI	NO	NO	NO	NO	Yes	Yes	Yes
Dual/STI	NO	NO	NO	NO	NO	NO	Yes
Unlanded/SA Contact	NO	NO	NO	NO	NO	NO	Yes
Poly CMP	NO	NO	NO	NO	NO	NO	Yes
Local Interconnect	NO	NO	NO	NO	NO	NO	Yes
WxL/Cell size	0.2	0.21	0.23	0.28	0.29	0.28	0.17

Table 1. Device/Process Features Demonstrated In ETOX Code Flash

YEAR OF PRODUCTION	2002	2005	2007	2010	2013	2016
DRAM Pitch(nm)	115	80	65	45	32	22
MPU / ASIC Pitch (nm)	130	80	65	50	35	25
MPU Printed Gate Length (nm)	75	45	35	25	18	13
MPU Physical Gate Length (nm)	53	32	25	18	13	9
Flash technology node - F (nm)	130	90	70	50	35	25
Flash NOR cell size area factor a in multiples of F2	10-12	11-14	11-14	12-15	13-16	14-17
Flash NAND cell size area factor a in multiples of F2 SLC/MLC	5.5	4.5	4.5/2.3	4.5/2.3	4.5/2.3	4.5/2.3
Flash NOR typical cell size (µm2)	0.1986	0.101	0.061	0.034	0.018	0.01
Flash NOR Lg-stack (phsical - µm)	0.25-0.27	0.2-0.22	0.19-0.21	0.17-0.19	0.14-0.16	0.12-0.14
Flash NOR highest W/E voltage (V)	8-10	7-9	7-9	7-9	7-9	7-9
Flash NAND highest W/E voltage (V)	18-20	18-20	17-19	17-19	16-18	16-18
Flash NOR I_read (µA) [8]	35-43	31-39	29-37	27-33	25-31	22-28
Flash Coupling Ratio [9]	0.65-0.75	0.65-0.75	0.6-0.7	0.6-0.7	0.6-0.7	0.6-0.7
Flash NOR tunnel oxide thickness (nm)	9.5-10	8.5-9.5	8.5-9.5	8-9	8	8
Flash NAND tunnel oxide thickness (nm)	8.5-9	8-9	7.5-8	6-7	6-7	6-7
Flash NOR interpoly dielectric thickness (nm)	12-14	10-12	9-11	8-10	6-8	4-6
Flash NAND interpoly dielectric thickness (nm)	13-15	12-14	10-12	10-12	9-11	9-11
Flash endurance (erase/write cycles)	1E5	1E5	1E5	1E6	1E6	1E7
Flash nonvolatile data retention (years)	10-20	10-20	10-20	10-20	20	20
Flash maximum number of bits per cell (MLC)	2	4	4	8	8	8
NOR WxL/cell size (one bit per cell)	17%	20%	34%	42%	58%	78%
NOR product density sweet spot (one bit per cell)	128M	256M	512M	1G	2G	4G

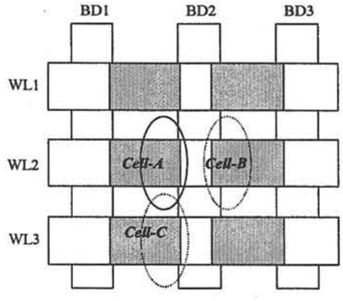
Table 2. ITRS Non-Volatile Memory Technology Roadmap (the last two rows are our inferences)



- L=0.13um is the scaling limit.
- I_{DTO} and I_{READ} are major limiting factors in scaling.

I_{DTO} : Drain turn on leakage
I_{BB}: Band to band leakage
I_{READ}: Cell read current

Fig.2. NOR Type FG Device Scaling Limitation



BD: Buried Diffusion
WL: Wordline

Fig 3. PHINES Virtual Ground Array.

	Program	Erase	Read
WL2	-V	Vh	Vwl
BD2	V	-V	Vs
BD1	0	-V	Vd
WL1,3	0	Vh	0
BD3	V/2	-V	Vs

Table 3. Program, Erase and Read Bias Conditions for Cell A

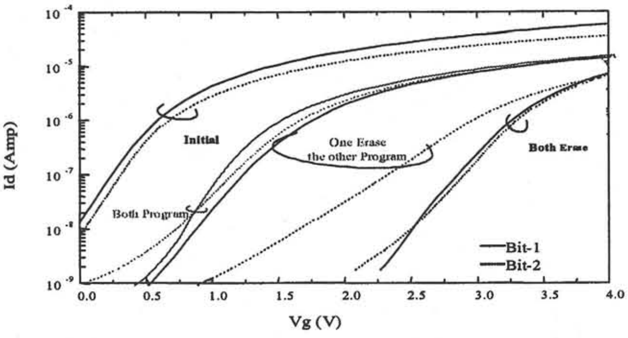


Fig.4 IV Characteristics of 4 Cell states: Initial, Both program, One Erase the Other Program and Both Erase

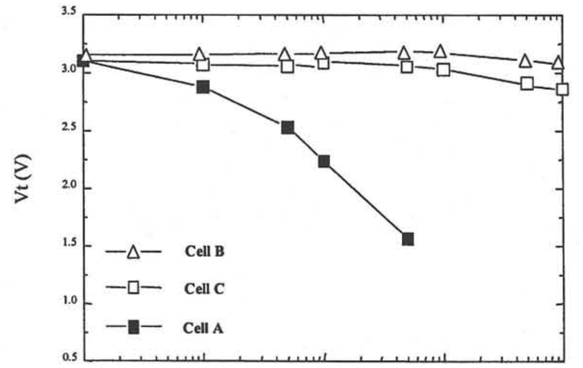


Fig.5 The programming Characteristic of Cell A and Disturb Behaviors of Cells B and C.

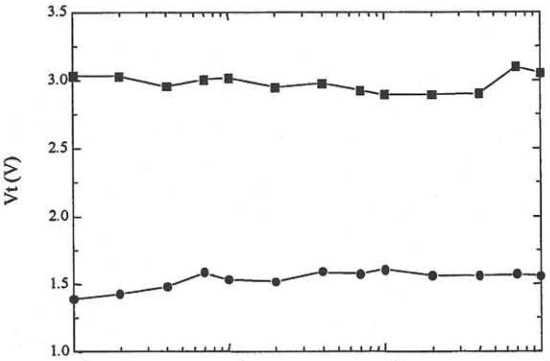


Fig.6. Cycle Endurance Characteristic of a Single PHINES Cell