Charge-Trapping (CT) Flash and 3D NAND Flash

Hang-Ting Lue

Macronix International Co., Ltd.
Hsinchu, Taiwan
Email: htlue@mxic.com.tw

Outline

❑ Introduction

❑ 2D Charge-Trapping (CT) NAND

❑ 3D CT NAND

❑ Summary

Outline

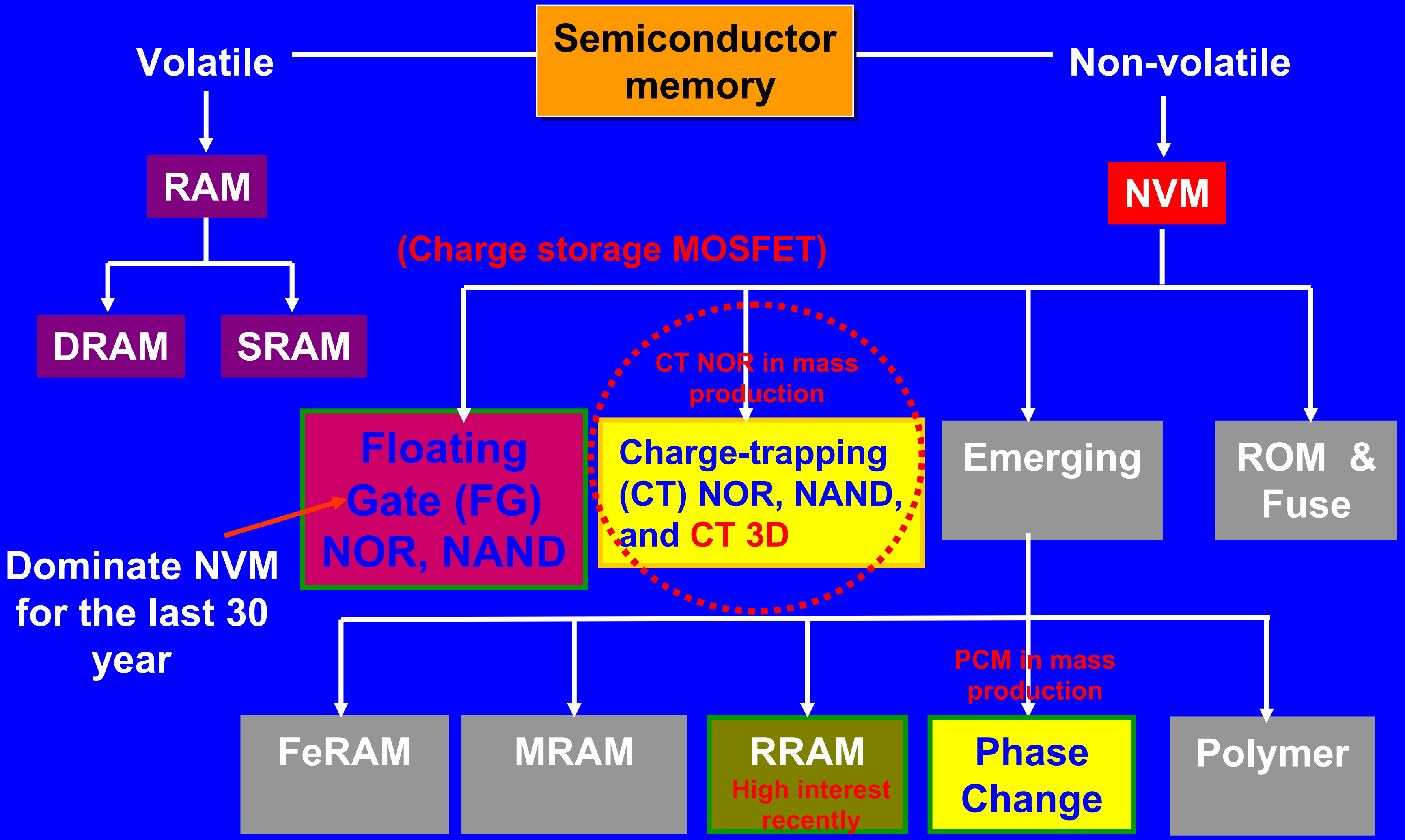
- ❑ Introduction

- ❑ 2D Charge-Trapping (CT) NAND

- ❑ 3D CT NAND

- ❑ Summary

Categories of Semiconductor Memory



Semiconductor memory

Volatile

Non-volatile

RAM

DRAM

SRAM

NVM

(Charge storage MOSFET)

Floating Gate (FG) NOR, NAND

Dominate NVM for the last 30 year

CT NOR in mass production

Charge-trapping (CT) NOR, NAND, and CT 3D

Emerging

ROM & Fuse

FeRAM

MRAM

RRAM

High interest recently

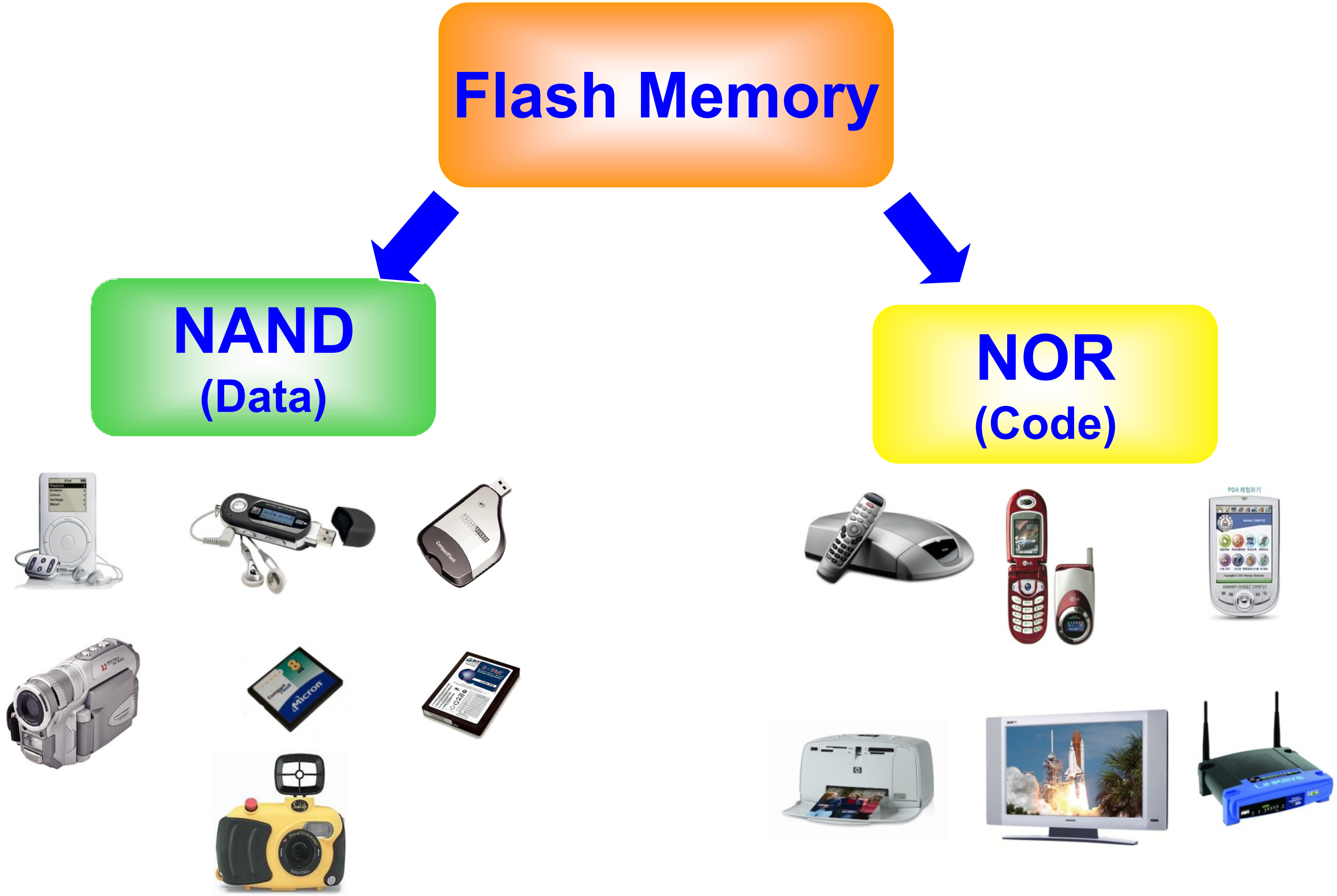
PCM in mass production

Phase Change

Polymer

□ CT NAND are discussed here.

Flash Memory Applications

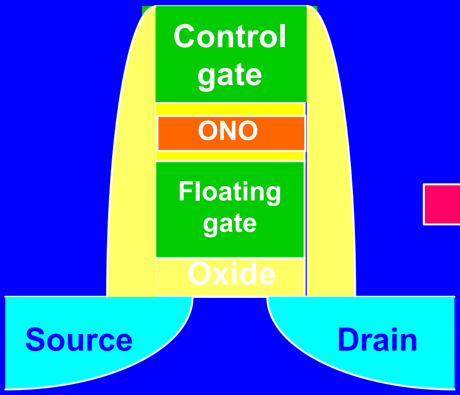


Flash Memory

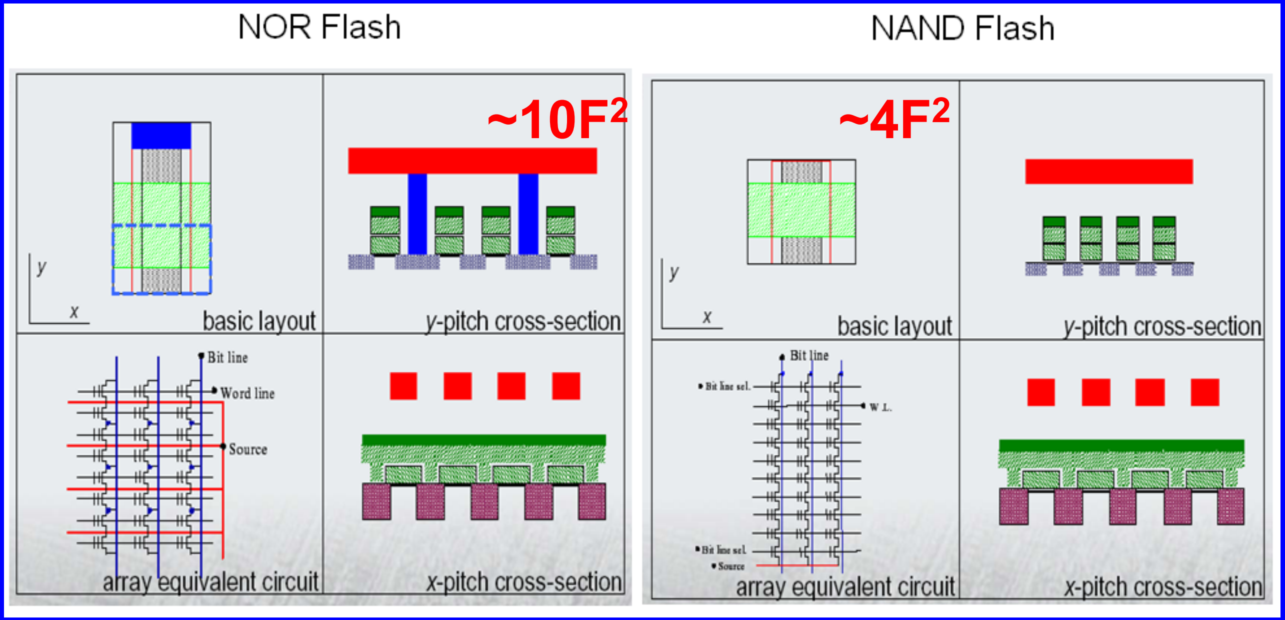
NAND
(Data)

NOR
(Code)

NOR and NAND Flash Memory



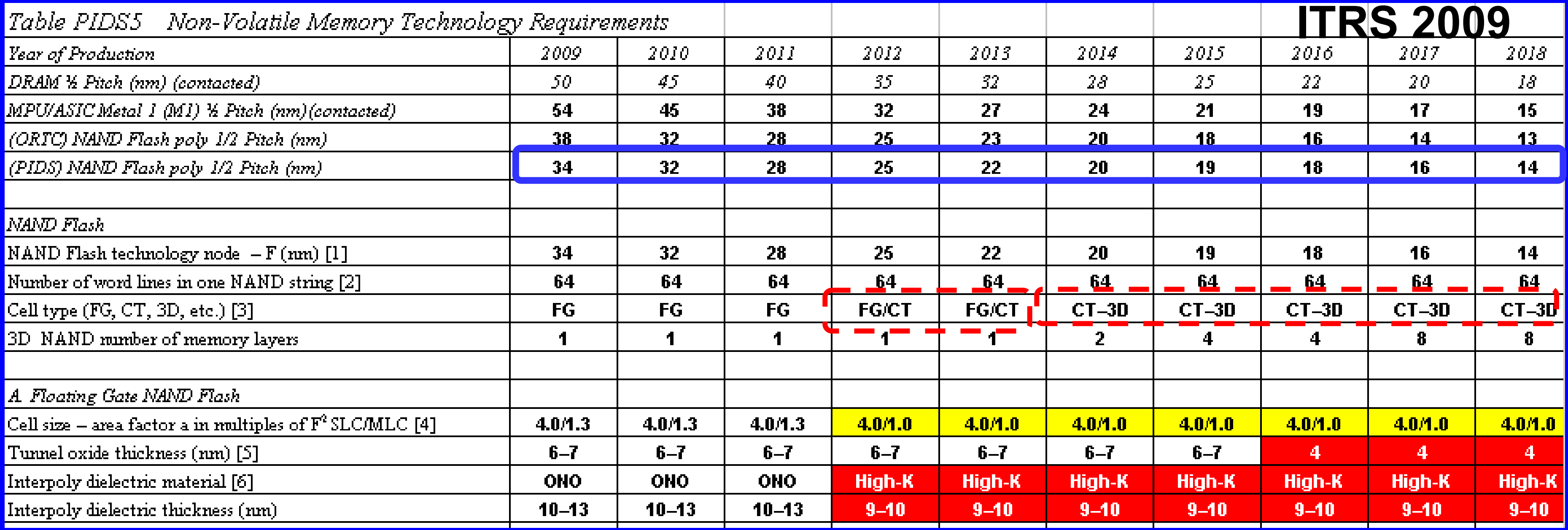
Single cell structure



□ Due to the excellent scalability and performances, NAND Flash has enjoyed the highest density

□ NOR Flash scaling is much slower than NAND so far

NAND Flash Scaling Roadmap – CT and 3D

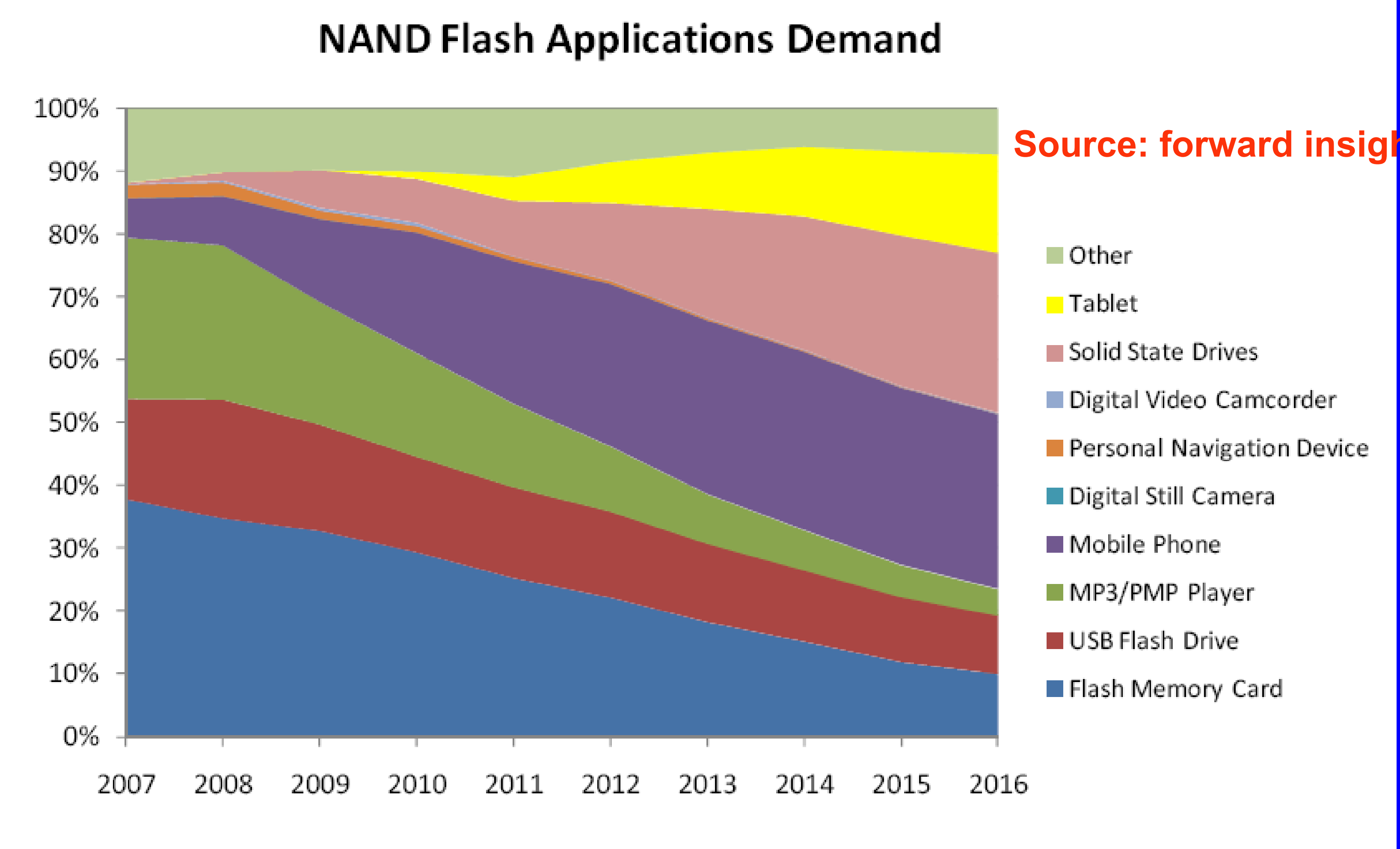


ITRS 2009

Table PIDS5 Non-Volatile Memory Technology Requirements										
Year of Production	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
DRAM ½ Pitch (nm) (contacted)	*50*	*45*	*40*	*35*	*32*	*28*	*25*	*22*	*20*	*18*
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)	54	45	38	32	27	24	21	19	17	15
(ORTC) NAND Flash poly 1/2 Pitch (nm)	38	32	28	25	23	20	18	16	14	13
(PIDS) NAND Flash poly 1/2 Pitch (nm)	34	32	28	25	22	20	19	18	16	14
NAND Flash										
NAND Flash technology node – F (nm) [1]	34	32	28	25	22	20	19	18	16	14
Number of word lines in one NAND string [2]	64	64	64	64	64	64	64	64	64	64
Cell type (FG, CT, 3D, etc.) [3]	FG	FG	FG	FG/CT	FG/CT	CT–3D	CT–3D	CT–3D	CT–3D	CT–3D
3D NAND number of memory layers	1	1	1	1	1	2	4	4	8	8
A Floating Gate NAND Flash										
Cell size – area factor a in multiples of F^2 SLC/MLC [4]	4.0/1.3	4.0/1.3	4.0/1.3	4.0/1.0	4.0/1.0	4.0/1.0	4.0/1.0	4.0/1.0	4.0/1.0	4.0/1.0
Tunnel oxide thickness (nm) [5]	6–7	6–7	6–7	6–7	6–7	6–7	6–7	4	4	4
Interpoly dielectric material [6]	ONO	ONO	ONO	High-K	High-K	High-K	High-K	High-K	High-K	High-K
Interpoly dielectric thickness (nm)	10–13	10–13	10–13	9–10	9–10	9–10	9–10	9–10	9–10	9–10

- □ NAND Flash has been scaled to 25nm (TLC, 3b/c) so far, even faster than ITRS prediction.
- □ 3D charge-trapping (CT) device is a possible solution to continue NAND Flash scaling below 1Xnm node.

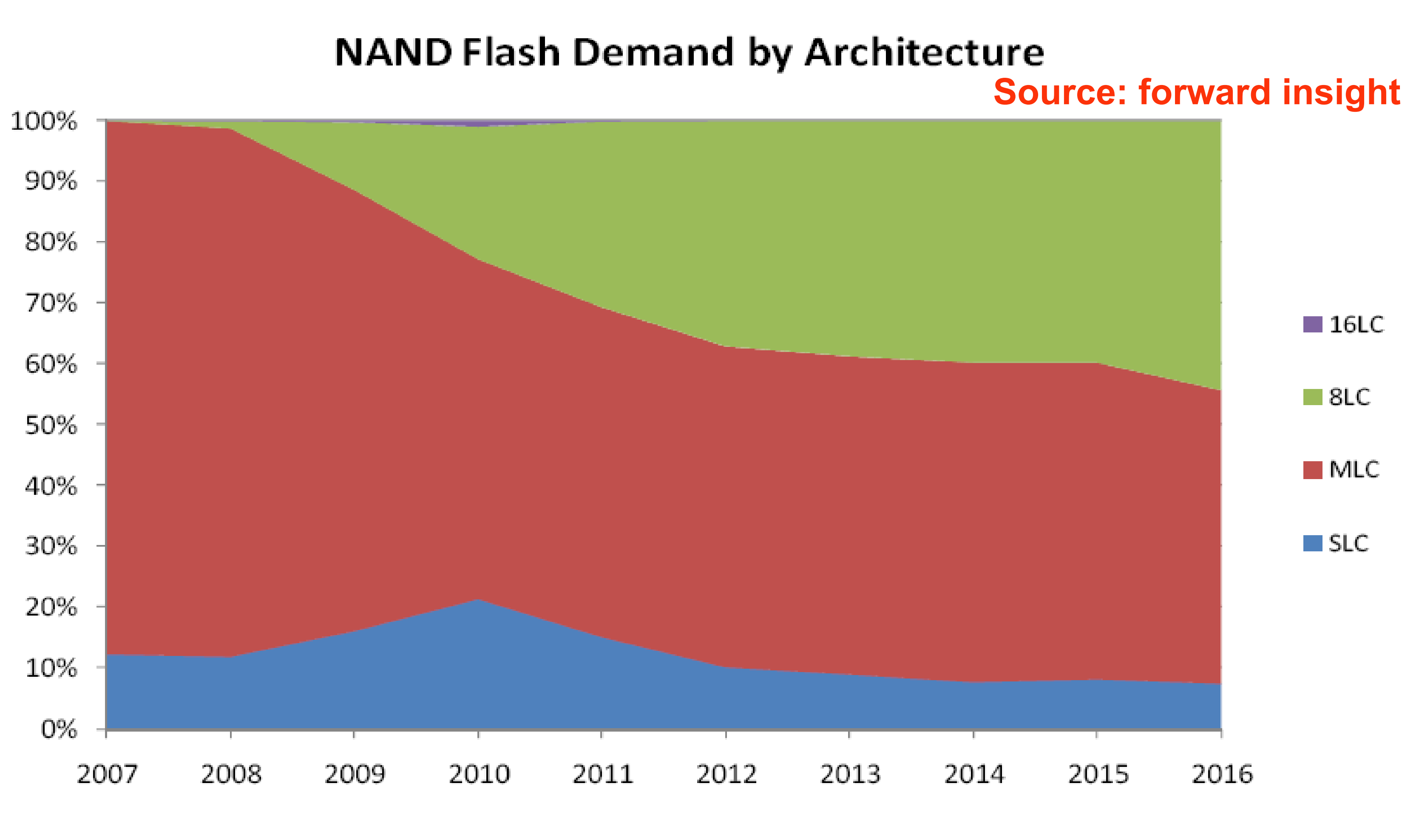
NAND Demand Forecast



Source: forward insight

NAND Flash enjoys a ~70% CAGR recently
Major driving force: Mobile application, Tablets, and SSD……

MLC/TLC/QLC



Source: forward insight

It is forecasted that the 16LC (4b/c) will only appear in a short period. TLC (3b/c) and MLC (2b/c) are the major products, while SLC keeps a small portion.

Endurance and Retention Forecast

Source: forward insight

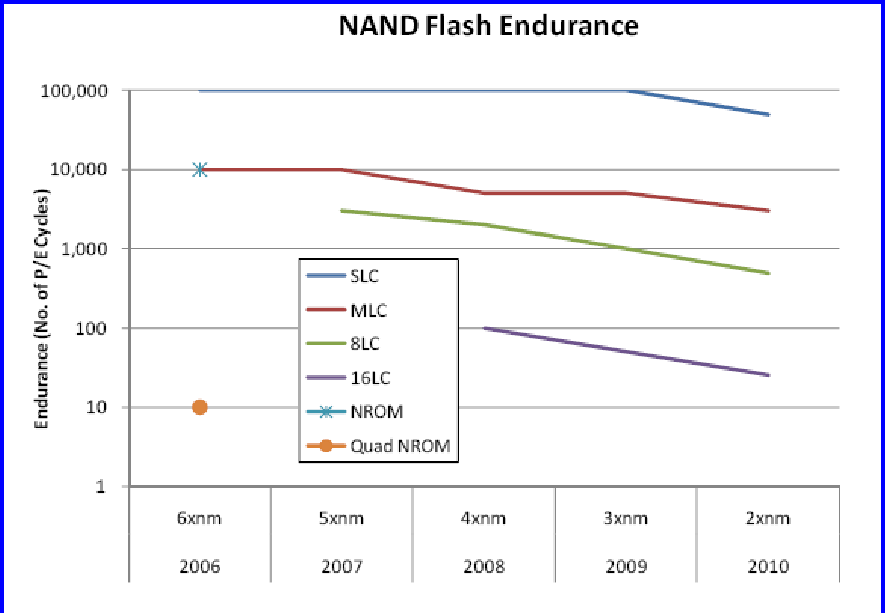
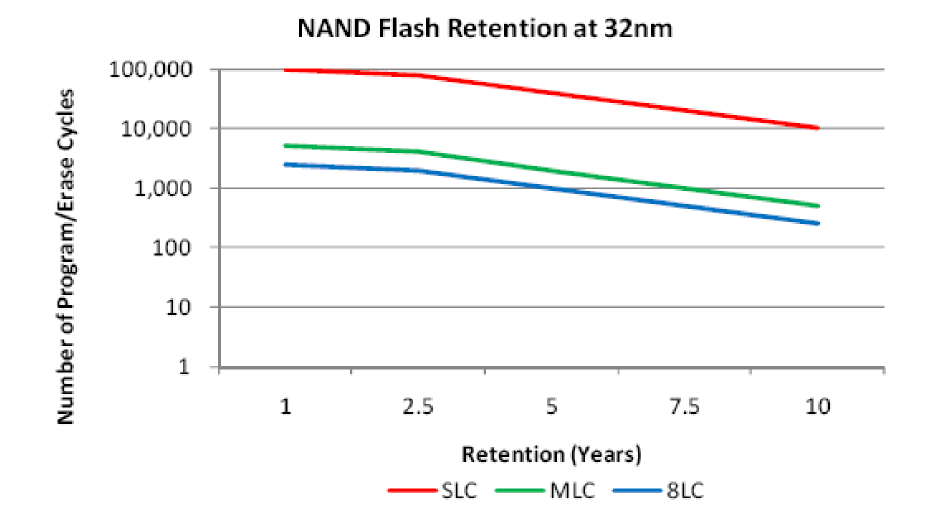
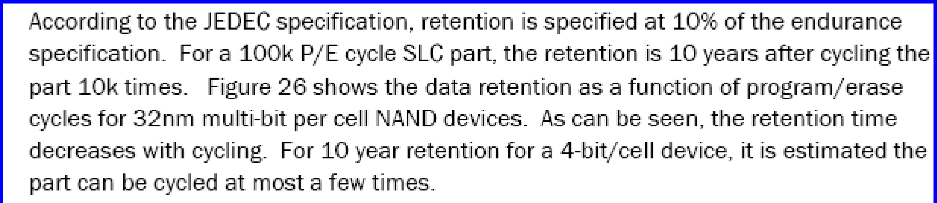


Figure 26. NAND Flash Endurance (32nm Process Technology)





According to the JEDEC specification, retention is specified at 10% of the endurance specification. For a 100k P/E cycle SLC part, the retention is 10 years after cycling the part 10k times. Figure 26 shows the data retention as a function of program/erase cycles for 32nm multi-bit per cell NAND devices. As can be seen, the retention time decreases with cycling. For 10 year retention for a 4-bit/cell device, it is estimated the part can be cycled at most a few times.

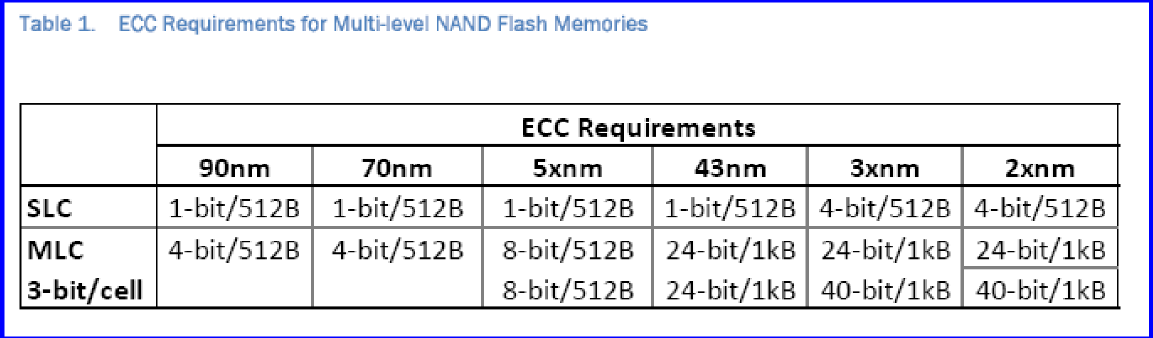
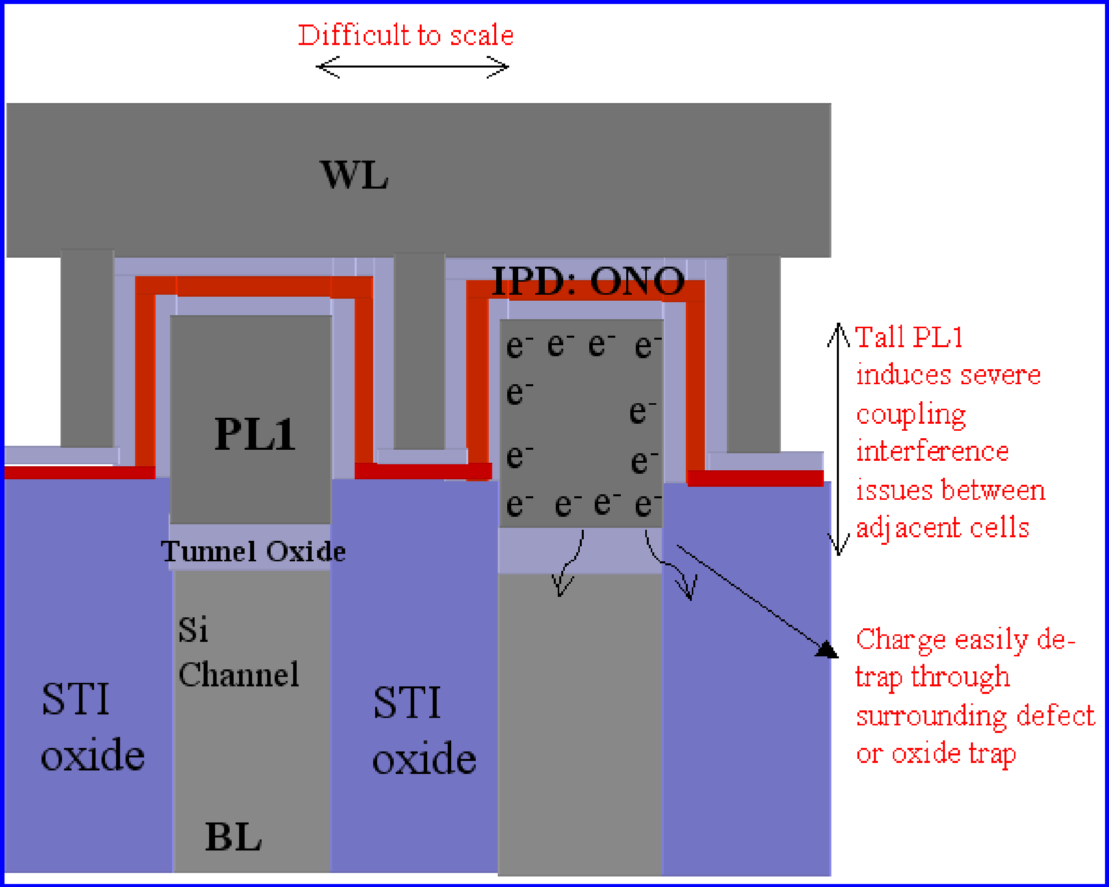


Table 1. ECC Requirements for Multi-level NAND Flash Memories

	ECC Requirements					
	90nm	70nm	5xnm	43nm	3xnm	2xnm
SLC	1-bit/512B	1-bit/512B	1-bit/512B	1-bit/512B	4-bit/512B	4-bit/512B
MLC	4-bit/512B	4-bit/512B	8-bit/512B	24-bit/1kB	24-bit/1kB	24-bit/1kB
3-bit/cell			8-bit/512B	24-bit/1kB	40-bit/1kB	40-bit/1kB

Endurance and retention continue to degrade. More than 40-bit ECC/page is necessary at 2X node.

Will NAND Flash Scale to 1X nm?

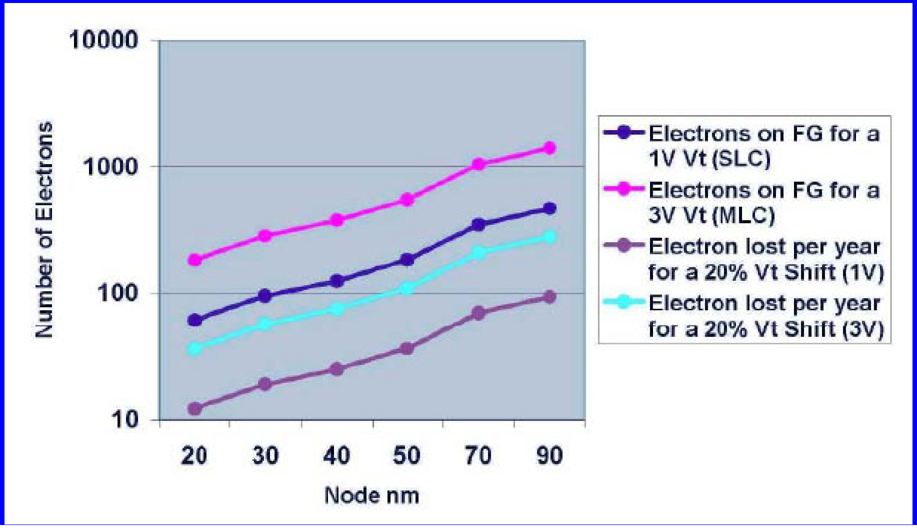


Tall PL1 induces severe coupling interference issues between adjacent cells

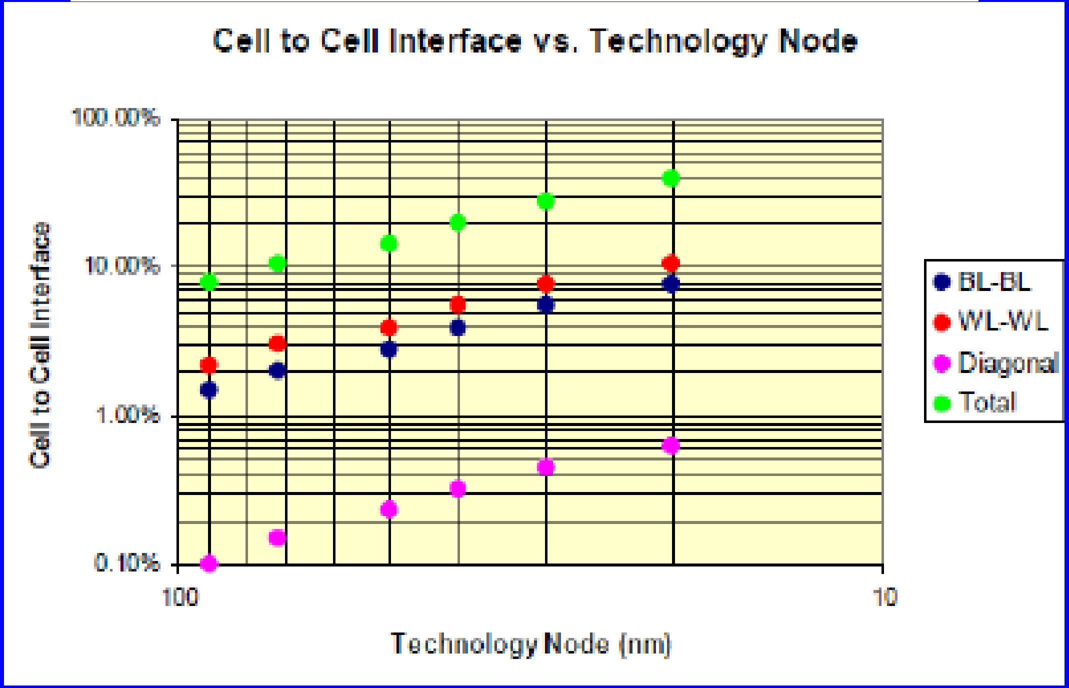
Charge easily de-trap through surrounding defect or oxide trap

IPD ONO thickness scales below 11nm
High-K IPD?
Tox scales below 7nm?
Thinner FG height and STI depth?

NAND Flash is going to run out of electrons!



Few electron number is the fundamental brick wall, especially for multi level cell

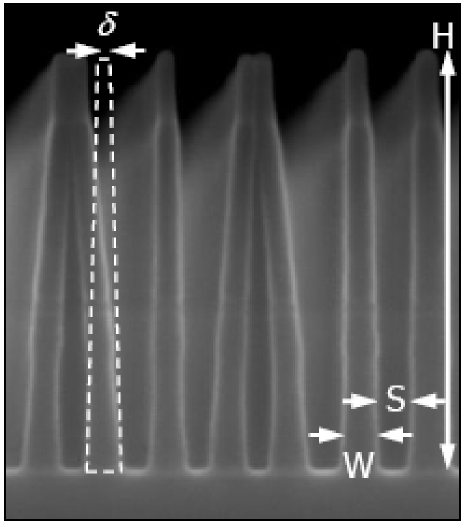


FG interference is huge (>40%) at 20nm node.

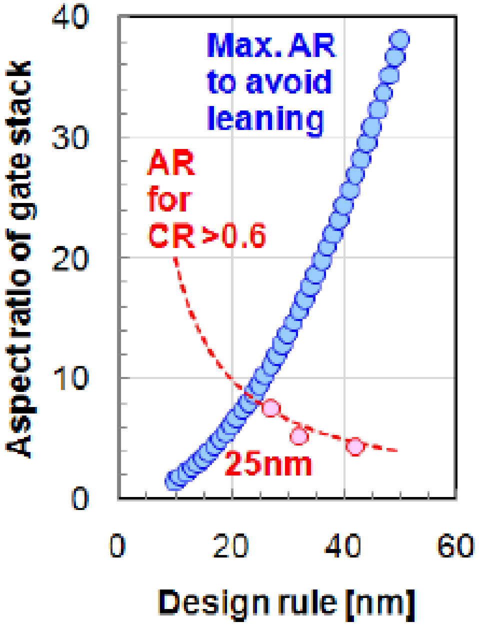
Scaling Issue - Physical Limitation

- ❑ **Gate stack leaning governed by Laplace pressure: DR >25nm**
  - **Even with low surface tension IPA (γ=21.7mN/m, just 1/3 of water)**
- ❑ **IPD leakage and CP void inhibit scaling of FG cell beyond 30nm**

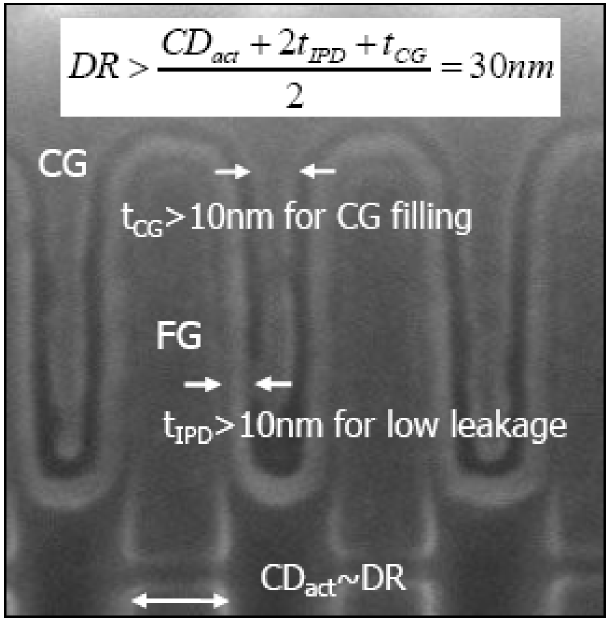
BL-direction



Deformation by Laplace pressure* $\delta = \dfrac{3\gamma H}{8WSE\rho} < \dfrac{1}{3}S$



WL-direction



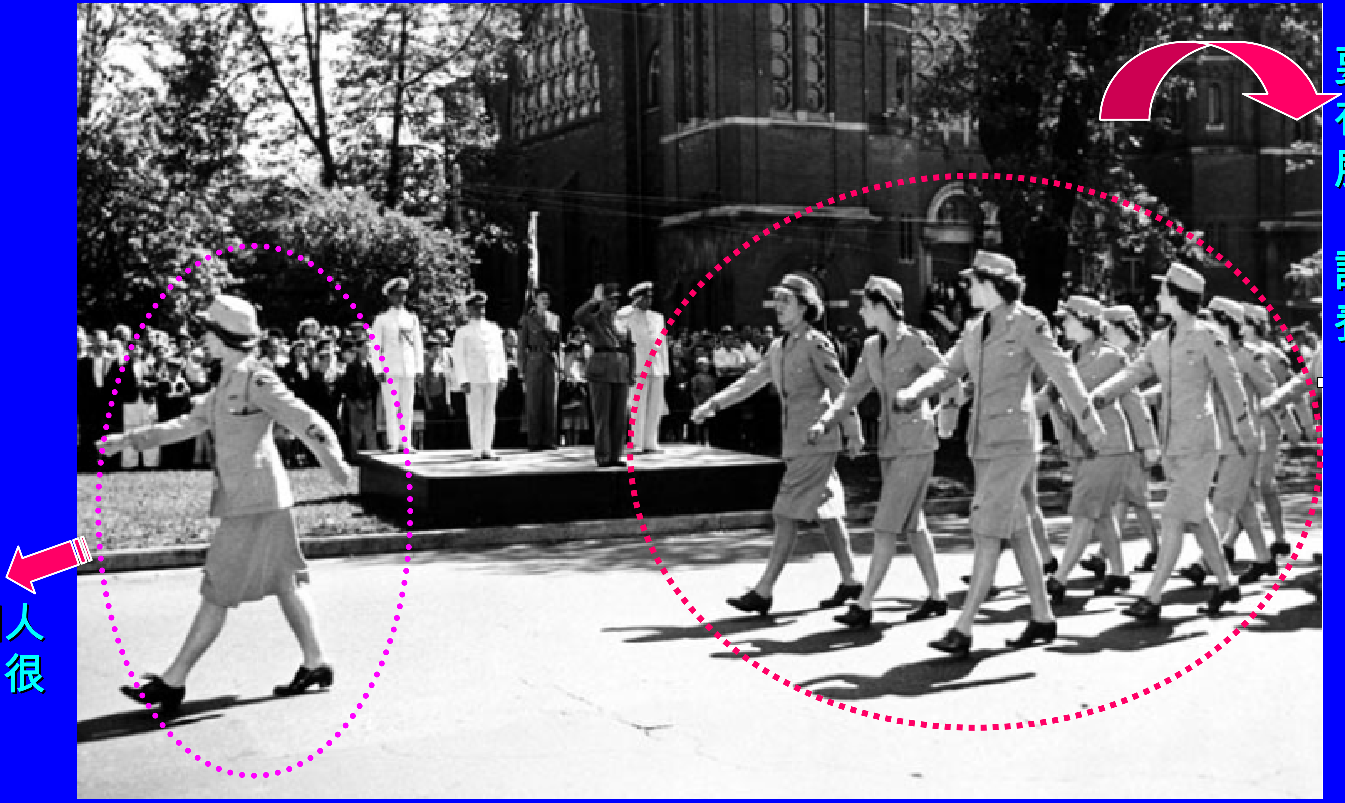
$$DR > \frac{CD_{act} + 2t_{IPD} + t_{CG}}{2} = 30nm$$

γ: surface tension of liquid, E: Young's modulus of poly-Si, ρ: density of poly-Si

*T. Abe, et. al., Journal of MEMS, vol. 4, no. 2, pp. 66-75, 1995

Source: J. Choi, IMW Short Course

Challenge of Cell Uniformity



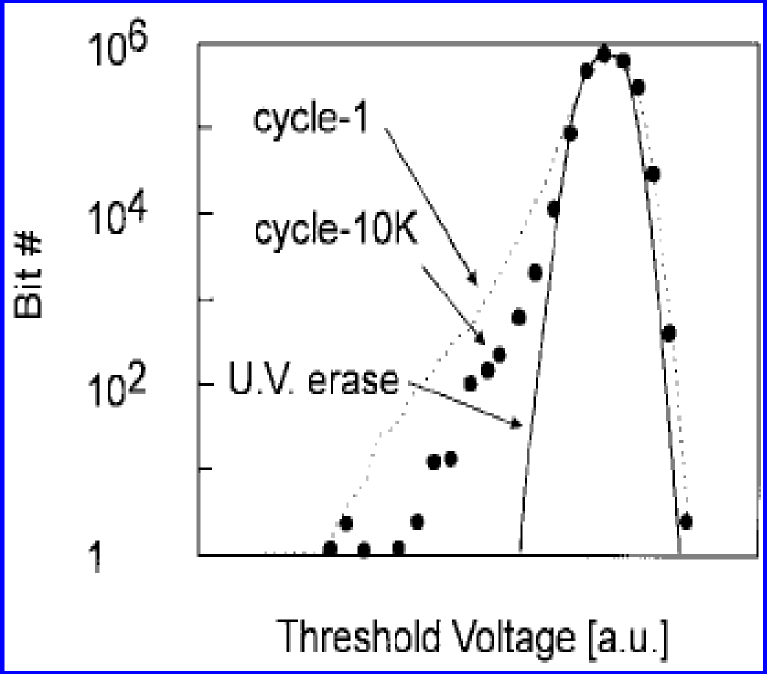
要讓每一個班兵在同一時間內，展現一致的動作，需要長時間的訓練與默契的培養

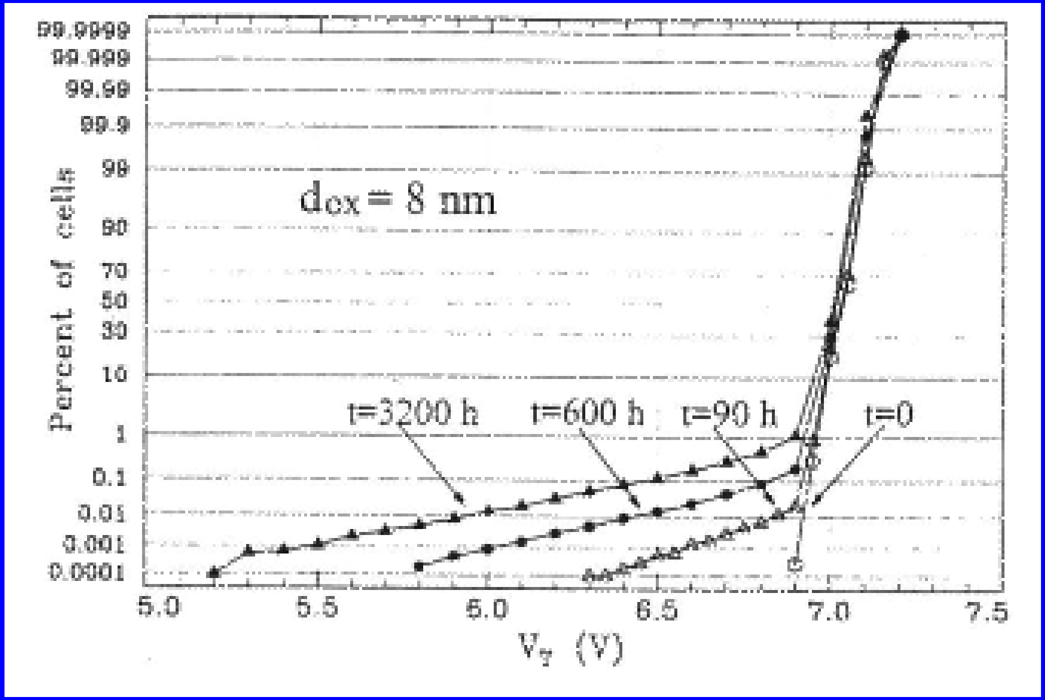
班長一個人踢正步，很簡單

□ Scaling generally makes uniformity very worse

□ Controlling cell uniformity is critical in overall performances

Challenge of Tail Bit





P. Cappelletti, *et al.*, IEDMTech. Dig., p.291, 1994.

- □ FG always has tail bits……(retention and P/E)
- □ More severe as Tox scales…..
- □ NOR don’t have tolerance
- □ NAND has more tolerance → ECC and many system-level design

Summary of FG Scaling Challenges

1. Geometry difficulty → gap filling and gate leaning
2. Reliability → Retention/endurance, noise….
3. Interference
4. High voltage and WL-WL breakdown..

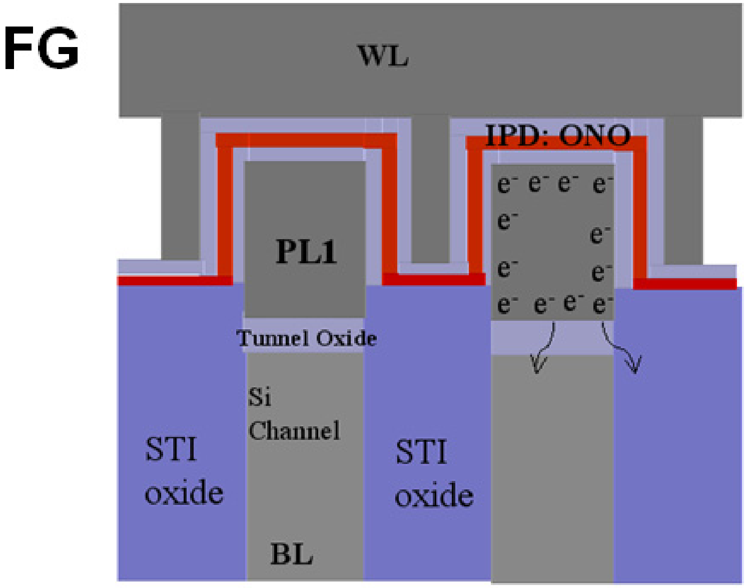
*5. Lithography limitations for 1Xnm……

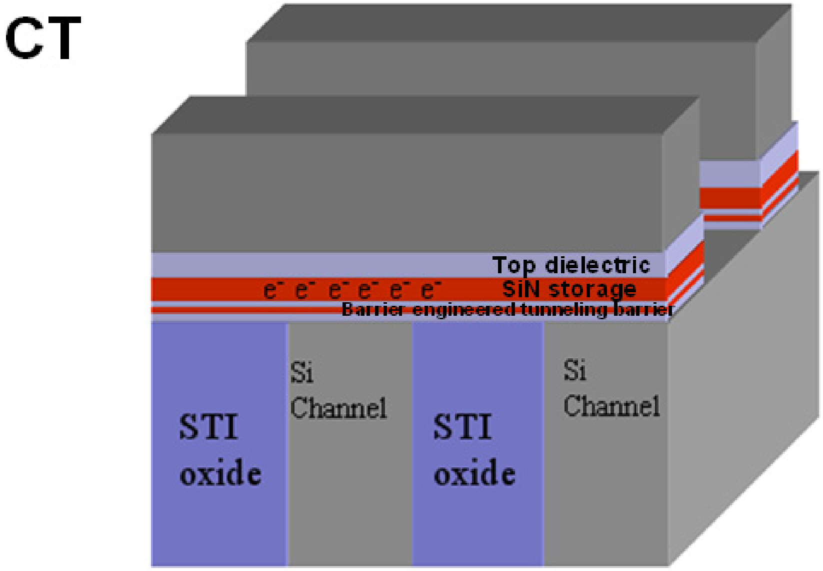
However, scaling efforts never stop……

Outline

- ❑ Introduction
- ❑ 2D Charge-Trapping (CT) NAND
- ❑ 3D CT NAND
- ❑ RRAM
- ❑ Summary

Brief Comparison of 2D FG and CT



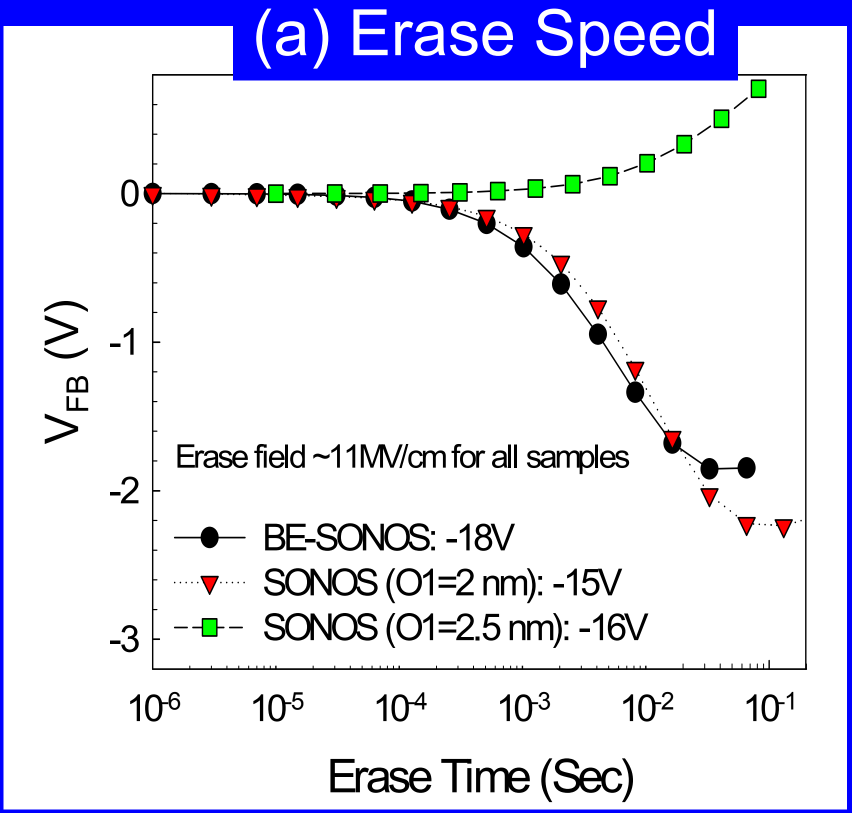


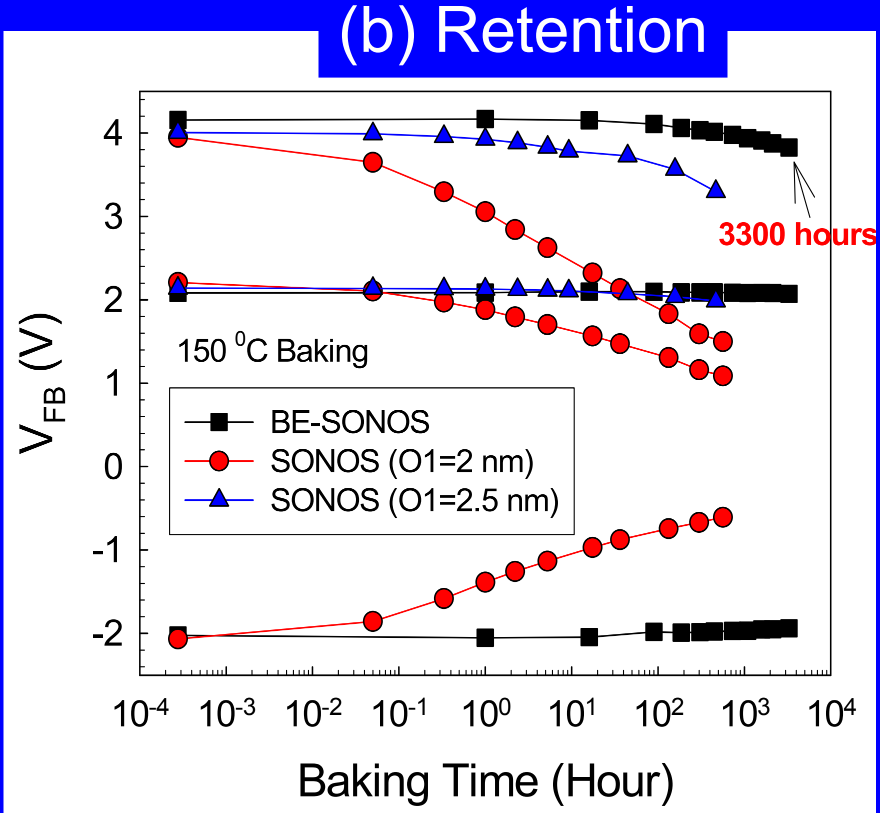
1. Gap filling difficulty and complicated topology → 1. Near-Planar structure
2. FG-FG interference and disturbs → 2. Discrete traps, no FG
3. Few-electron retention and sensitivity to oxide defect (SILC) → 3. Deep traps (High E_A). Immune to point defect in tunnel oxide

- □ **CT is simpler in topology**
- □ **More immunity to tunnel oxide defect and SILC**

Problem of Conventional SONOS

H. T. Lue (Macronix), et al, ICSICT, 2008.


(a) Erase Speed


(b) Retention

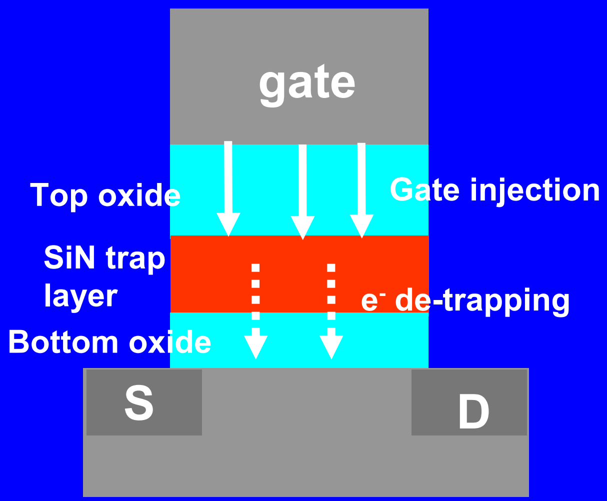
□ When O1> 25A, the erase becomes too slow (gate injection current is larger!)

□ When O1< 25A, data retention is too poor!

□ Erase and retention dilemma is the general issue

Charge-Trapping Devices Need BE Tox or High-K Top Dielectric

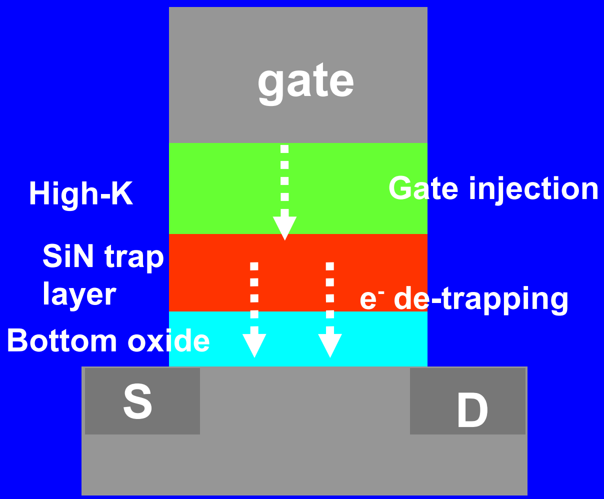
SONOS



E_{O2}=E_{O1}, and gate injection is larger than electron de-trapping → no memory window for erase

MANOS

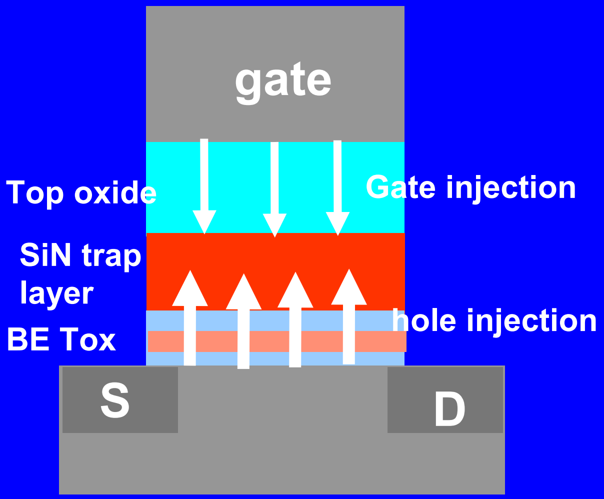
C. H. Lee, IEDM 2003.



By higher-K top oxide, E_{O2} is smaller, leading to smaller gate injection during erase.

BE-SONOS

H. T. Lue, IEDM 2005.



E_{O2}=E_{O1}, but BE Tox has larger hole injection than gate injection

□ Unlike FG, CT device is designed in a planar structure without GCR design.

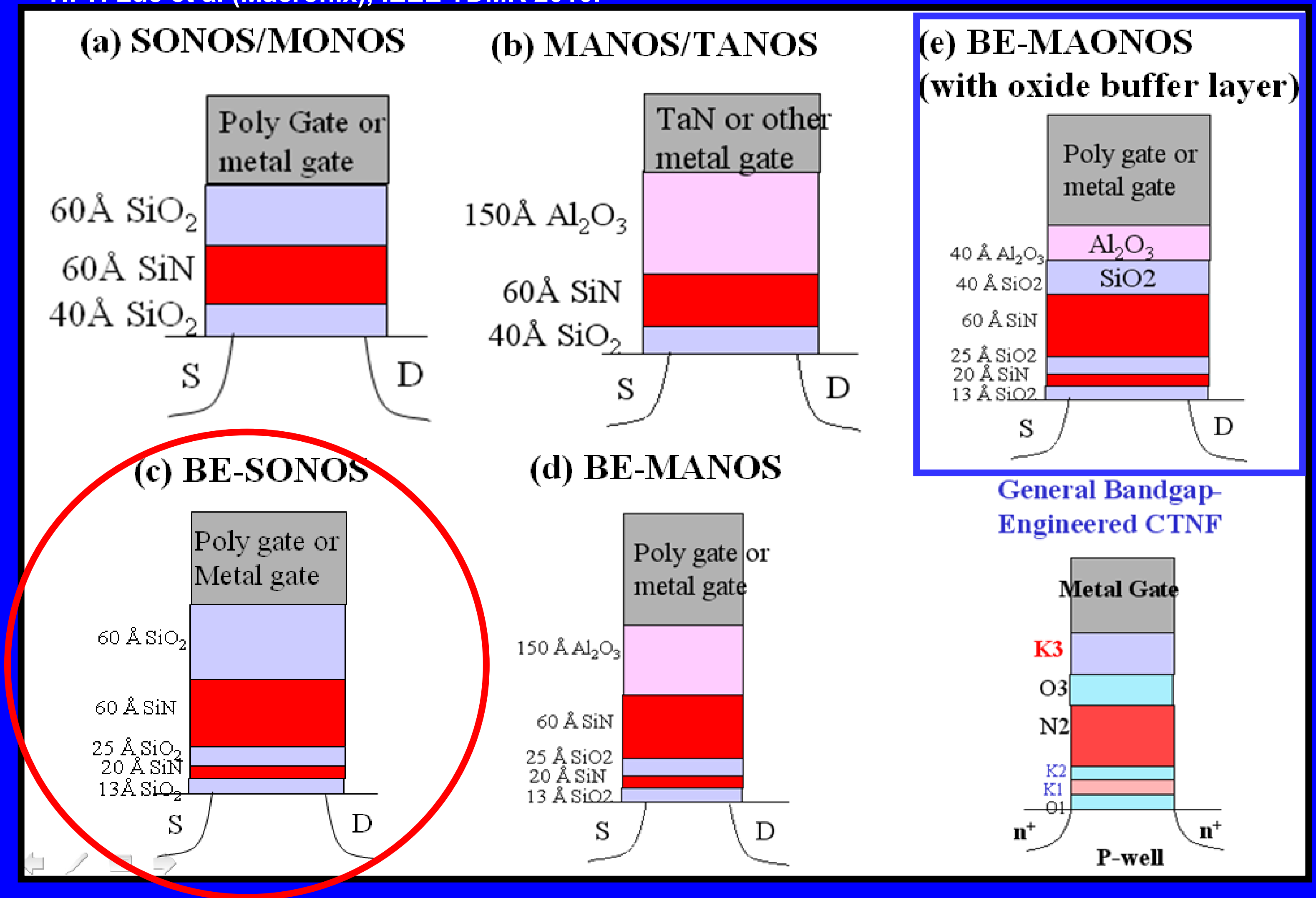
□ Bottom tunnel oxide (E_{O1}) has the same E field with top oxide (E_{O2}), leading to small memory window during the erase.

□ High-K top dielectric can reduce the gate injection

□ BE Tox can improve the hole injection for faster erase

Glance over various CT Devices

H. T. Lue et al (Macronix), IEEE TDMR 2010.



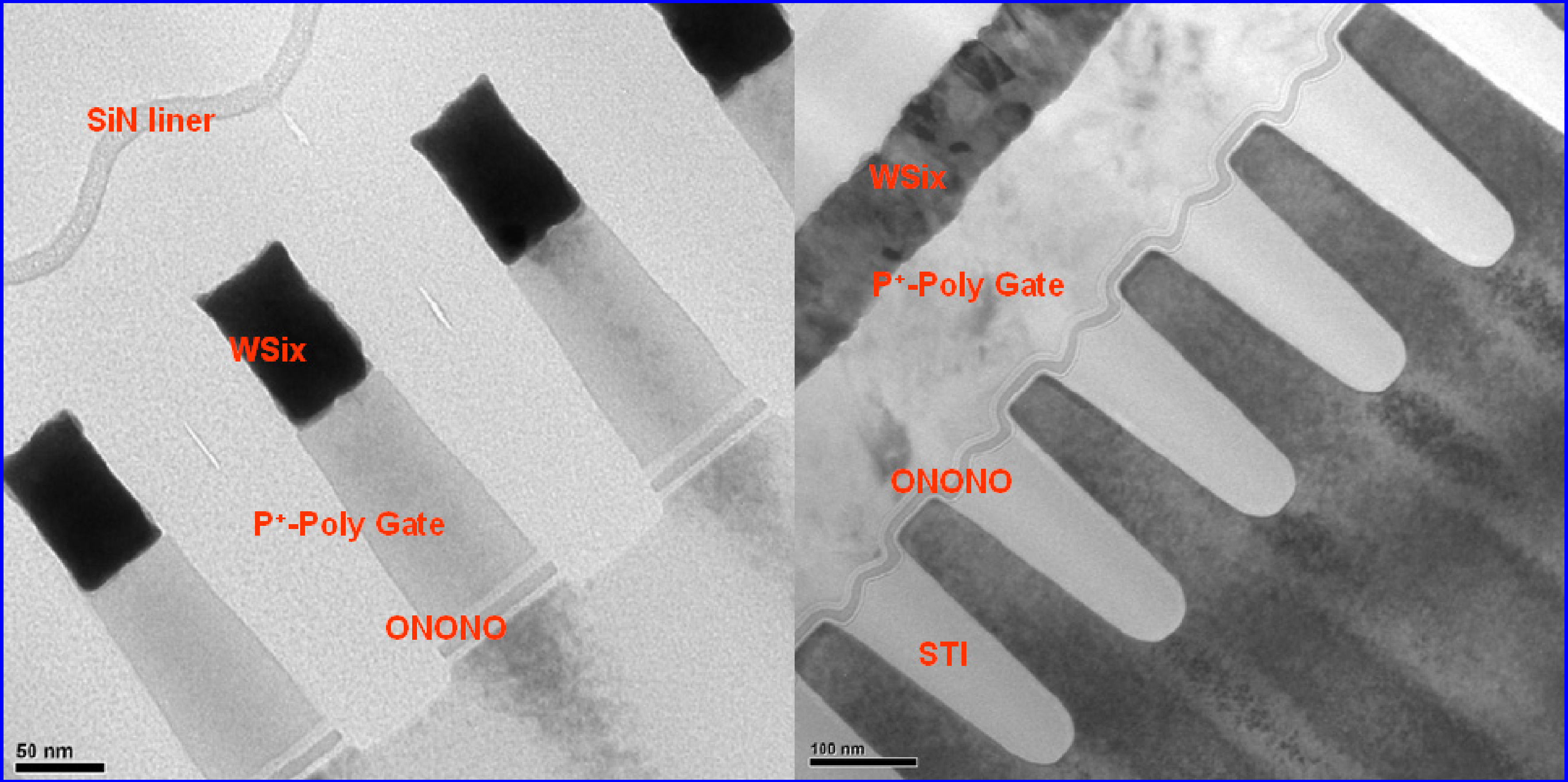
Theoretically the highest performance

Best reported reliability

→ No new materials, fast learning time

□ High-K CT devices (such TANOS) requires more learning time in reliability

BE-SONOS NAND Flash



H. T. Lue et al (Macronix), IMW, 2010.

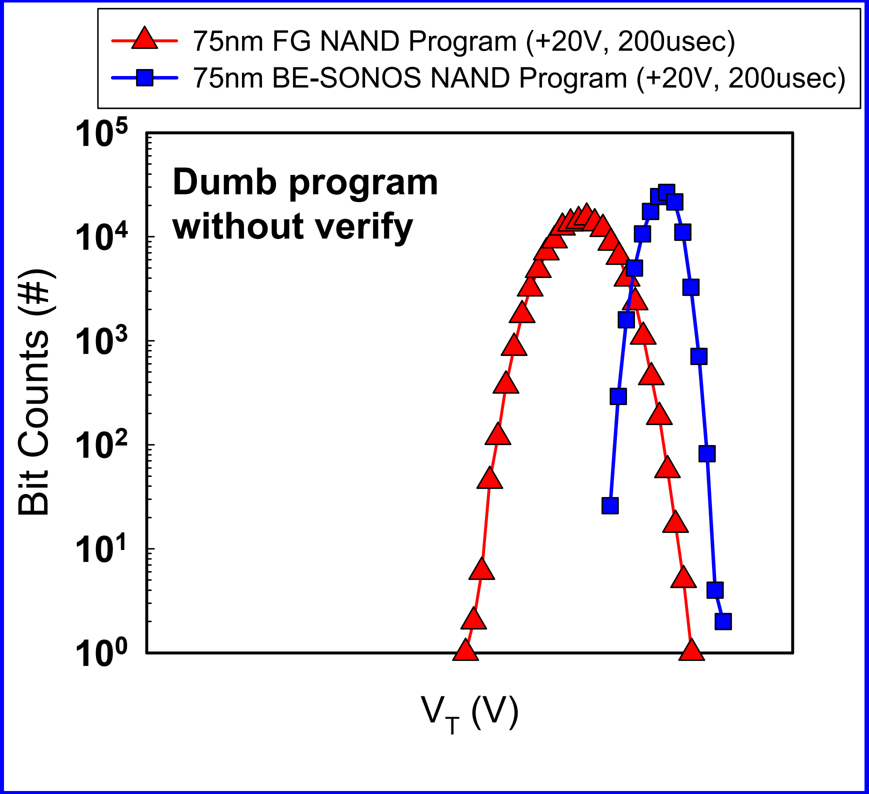
□ A 75nm BE-SONOS NAND Flash test chip has been demonstrated.

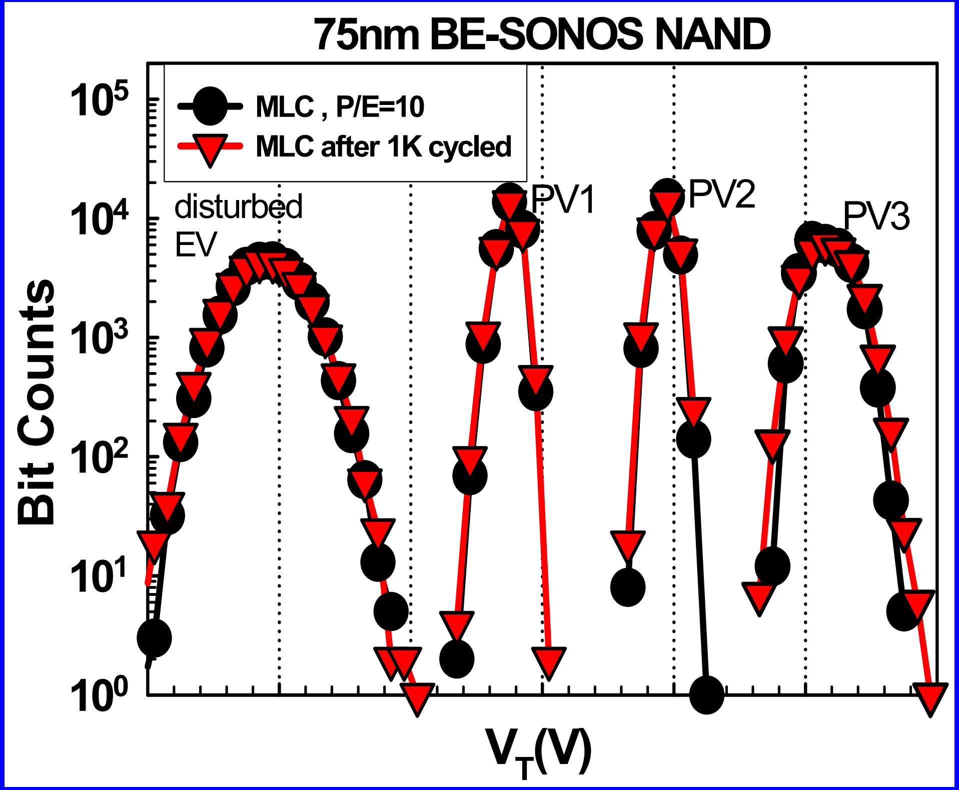
□ Near planar STI. Conventional materials (oxide, nitride, poly)

□ A highly reliable 38nm node BE-SONOS NAND will be published at IEDM 2010.

BE-SONOS NAND Performances

H. T. Lue, (Macronix), IMW short course





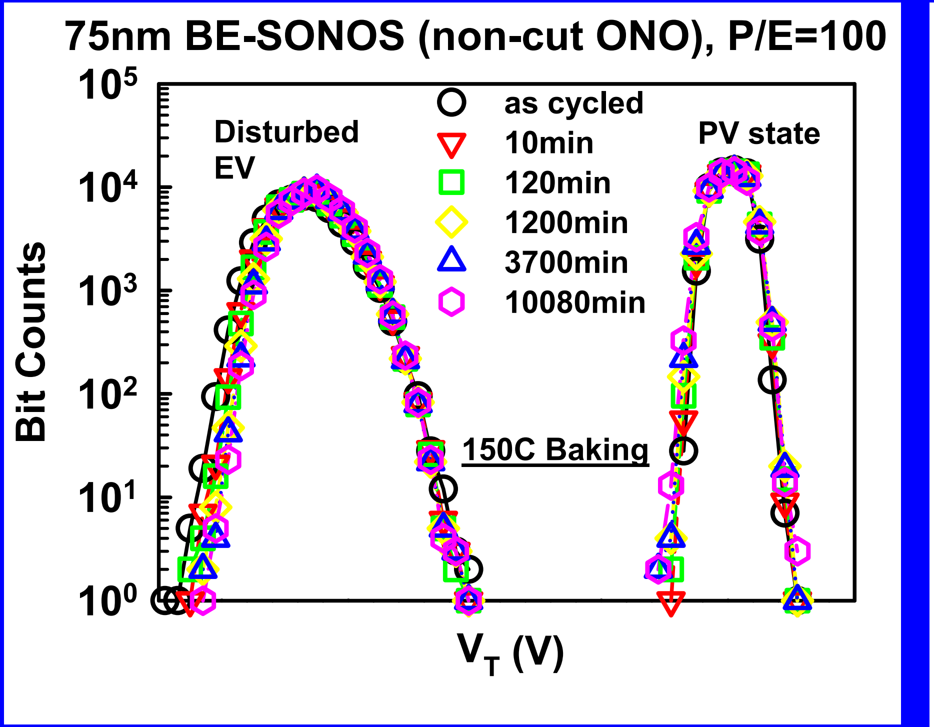
□ Our BE-SONOS NAND programming distribution can be tighter than FG due to simpler topology that minimizes the variation.

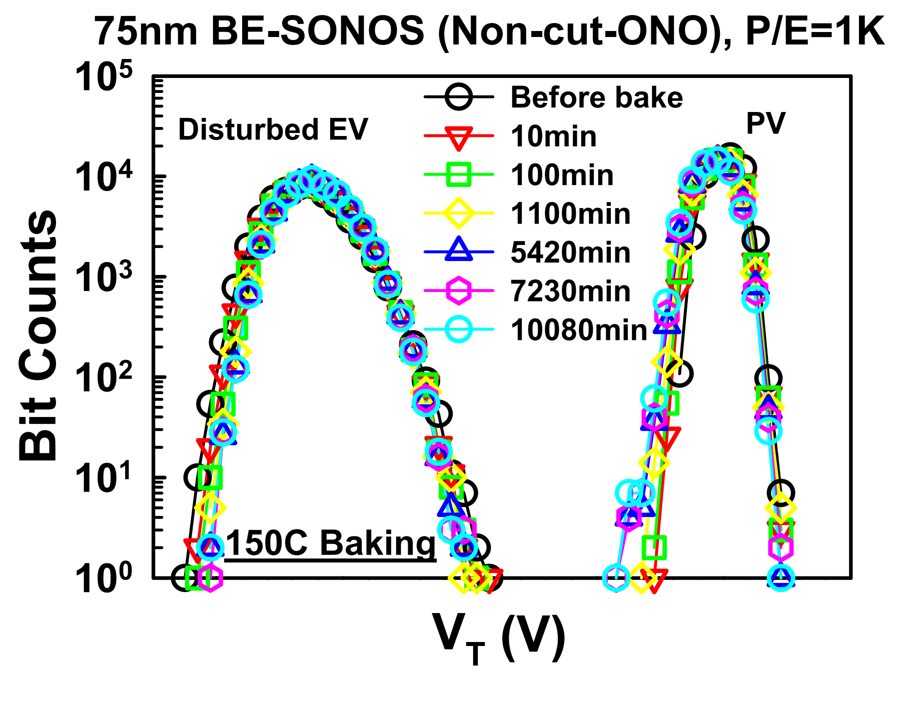
□ Good programming and read performances.

□ MLC operation of BE-SONOS NAND test chip is successful.

Retention of BE-SONOS NAND

C. C. Hsieh, (Macronix), IEDM 2010.





- □ Retention is excellent and no single tail bit found.
- □ The best reported CT reliability so far.
- □ No so called charge lateral migration issue (with our optimized SiN trapping layer and process integration)

We have developed a successful 2D CT BE-SONOS NAND with excellent reliability

However, current FG NAND has already scaled to ~25nm node with TLC

Therefore, CT NAND must look for further scaling below 1X nm node

Scaling Challenge of 2D CT NAND Below 20nm Node

□ Lithography difficulty below 1X nm
□ Few-electron storage and statistics
□ RTN (noise)
□ Interference of CT NAND is still observed
□ High voltage requirement is approximately the same with FG

→ 2D CT NAND probably has a similar (or a little more) scalability with FG NAND

Outline

- ❑ Introduction
- ❑ 2D Charge-Trapping (CT) NAND
- ❑ 3D CT NAND
- ❑ RRAM
- ❑ Summary

“Simply Stacked” 3D NAND Flash

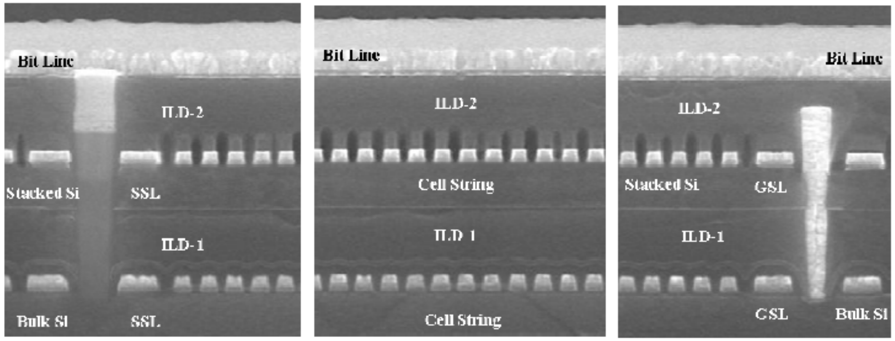


Fig. 4 Vertical SEM photograps of the fabricated 3D stacked NAND cell string. The 2nd active layer is SOI like perfect single crystal.

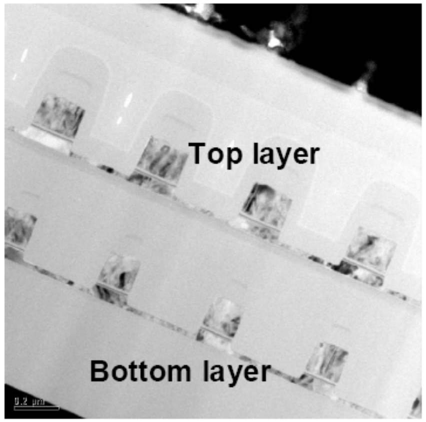


Fig. 3 Channel-length direction of double-layer TFT NAND devices.

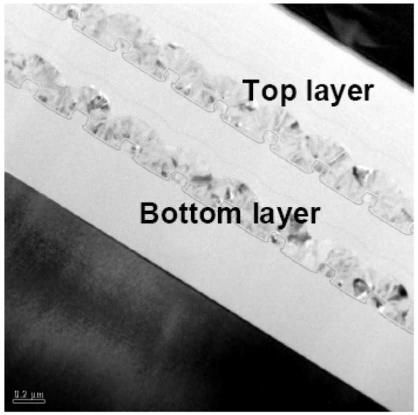


Fig. 4 Channel-width direction of double-layer TFT NAND devices.

3D TANOS devices
Samsung: IEDM 2006

3D TFT BE-SONOS devices
Macronix: IEDM 2006

□ 3D stackable NAND Flash using charge-trapping devices were firstly demonstrated in 2006.

□ Charge-trapping (CT) TANOS and BE-SONOS devices were used.

□ To stack many layers may linearly increase the cost → Not good when more than 4 layers are used.

□ However, for <4 layers the cost is reduced. The process seems doable in principle for 2X nm node….

Bit-cost scalable (BiCS) NAND Flash

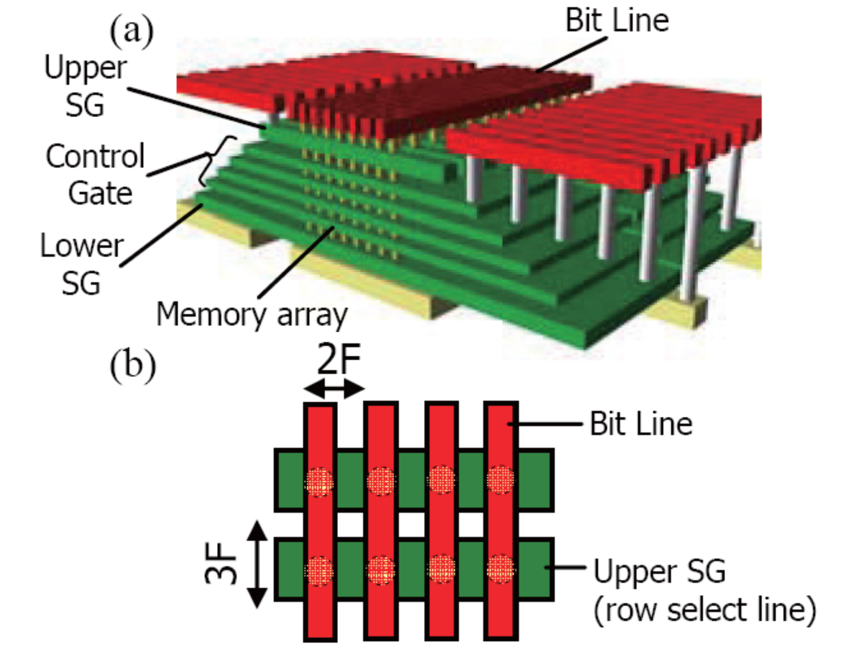
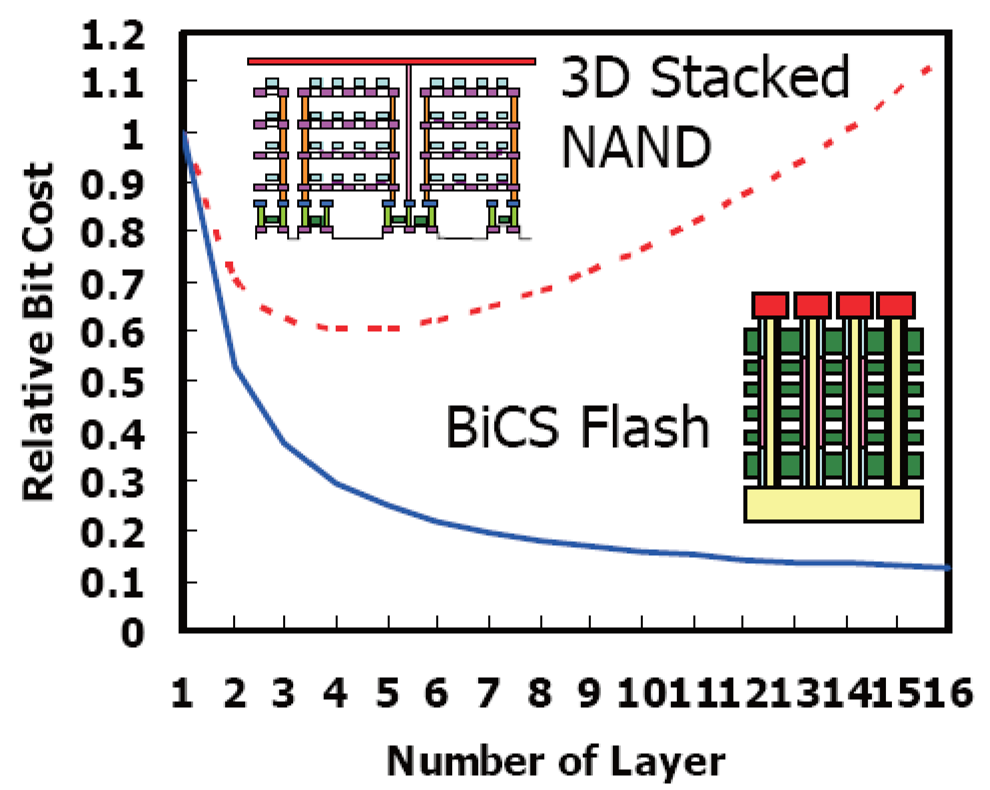


Fig. 3 (a) Birds-eye view of BiCS flash memory, (b) Top down view of BiCS flash memory array.



TOSHIBA: VLSI Symposia 2007

□ A break-through concept was proposed by TOSHIBA.

□ It uses a only one critical contact drill hole for many layers, thus the bit cost is scalable when more than 16 layers are used.

□ 3D NAND is a way to bypass the difficulty in lateral scaling

□ Also keep the electron number……

3D NAND Flash Architectures

P-BiCS

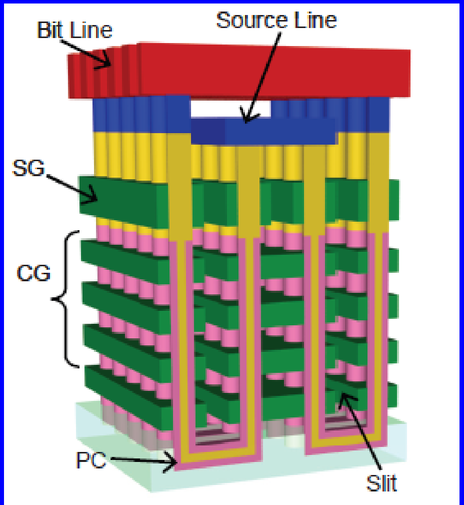
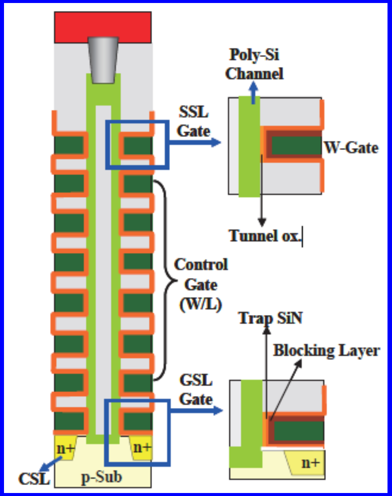


Fig. 1 Schematic of P-BiCS flash memory.

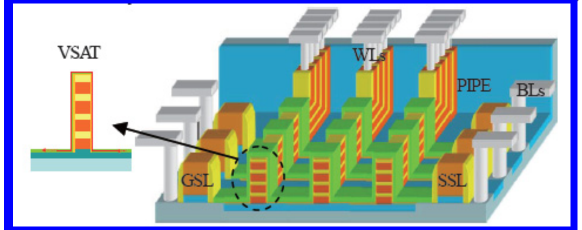
Ryota K., et. al. 2009 VLSI

TCAT



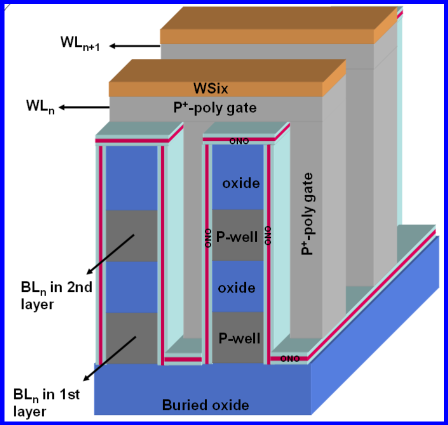
Jaehoon J., et. al. 2009 VLSI

VSAT



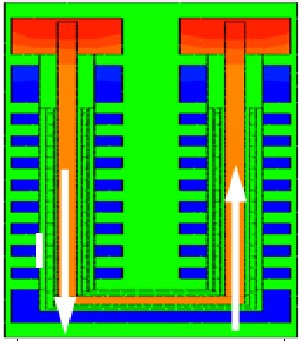
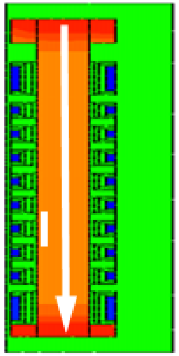
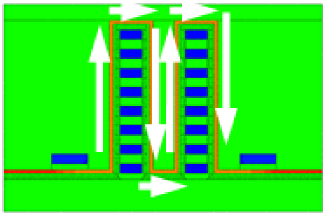
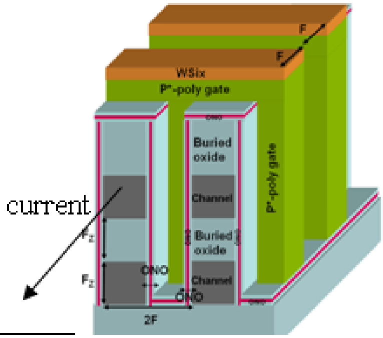
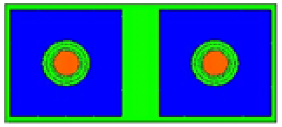
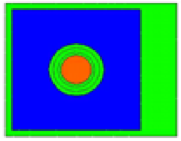
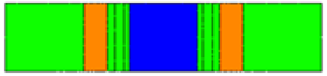
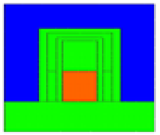
Jiyoung Kim, et. al. 2009 VLSI

VG



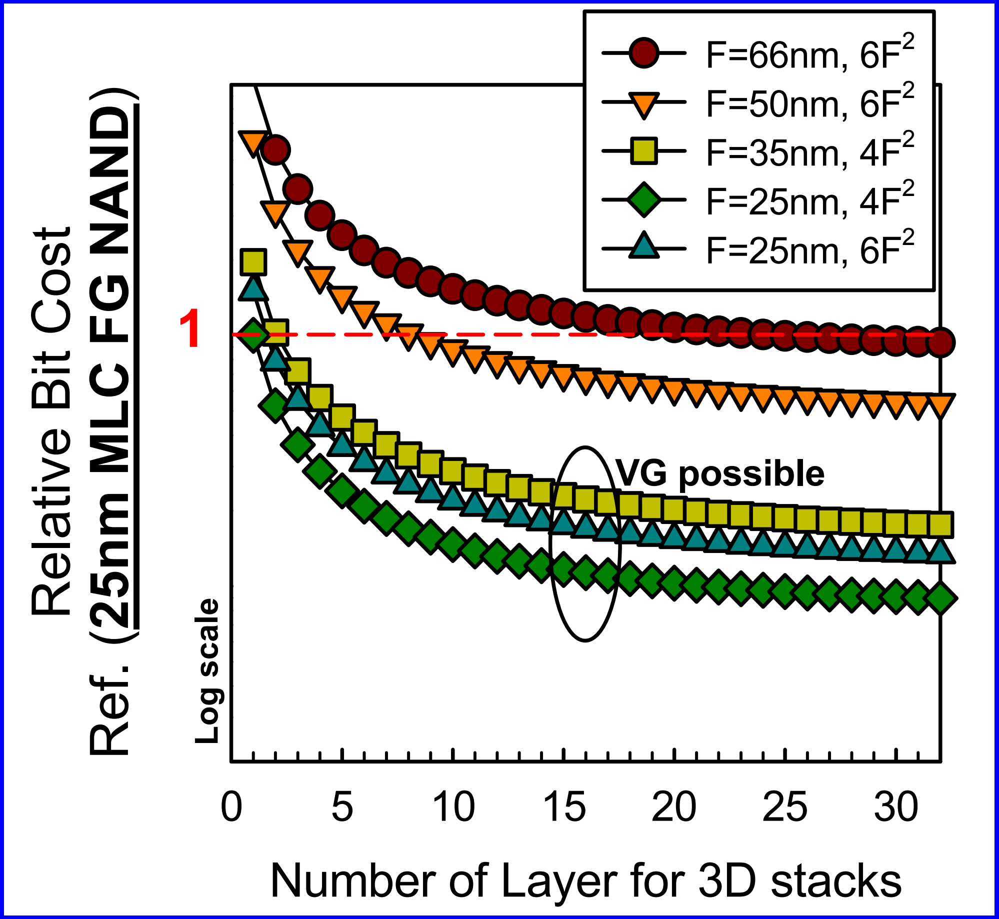
Wonjoo Kim, et. al. 2009 VLSI

3D NAND Flash Comparison

	P-BiCS	TCAT	VSAT	VG
String				 current F F WSix P+-poly gate ONO Buried oxide Channel Buried oxide Channel P+-poly gate F_Z F_Z ONO ONO 2F
Cell Shape				
Cell Size in X, Y direction	$6F^2$ (3F*2F)	$6F^2$ (3F*2F)	$6F^2$(3F*2F)	$4F^2$ (2F*2F)
Gate Process	Gate first	Gate last	Gate First	Gate Last
Current Flow direction	U-turn	Vertical	Multi-U-turn	Horizontal
Device Structure	GAA	GAA	Planar	Double Gate
Possible minimal F	~50 nm	~50 nm	~50 nm	~2X nm

[P-BiCS] R. Katsumata, et al, VLSI Symposia, pp. 136-137, 2009. [TCAT] J. Jang, et al, VLSI Symposia, pp. 192-193, 2009. [VSAT] J. Kim, et al, VLSI Symposia, pp. 186-187, 2009. [VG] W. Kim, et al, VLSI Symposia, pp. 188-189, 2009.

3D NAND Bit Cost Analysis – More realistic calculation

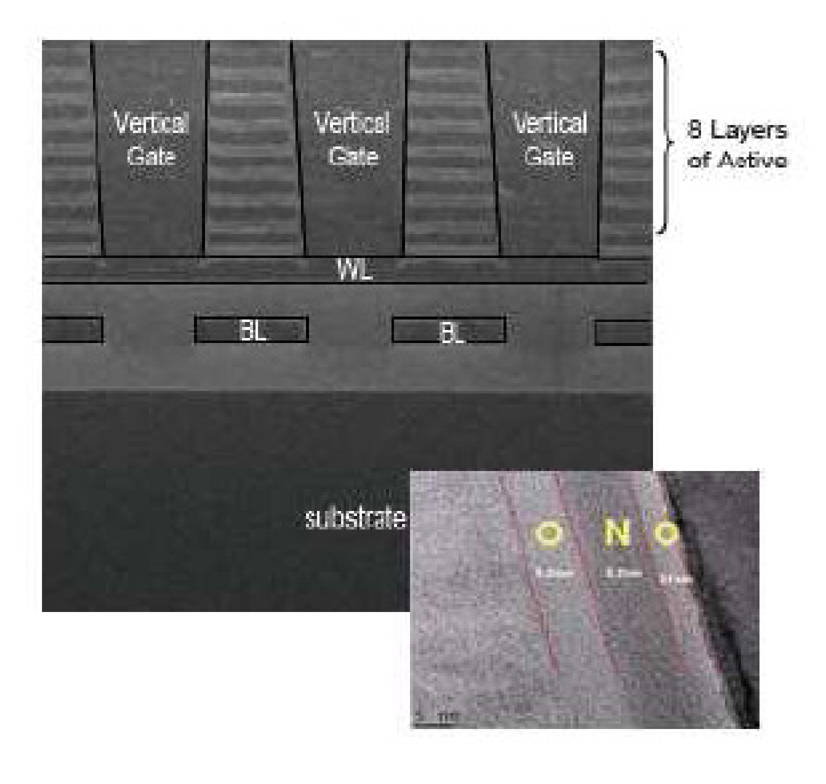


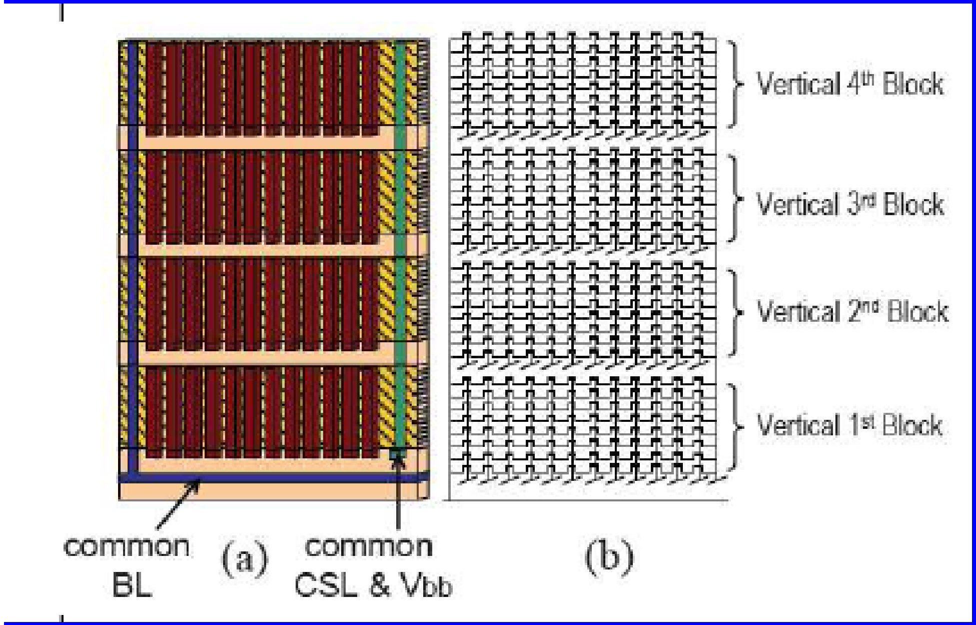
PS: Additional **processing cost** and **array efficiency loss** when adding one more memory layer are considered….

□ If 3D starts from >65nm 6F^2 cell size, it is hard to compete with current 25 nm FG NAND

□ 3D NAND is best to have cell size below 3X nm → VG is possible

Previous VG NAND Architecture





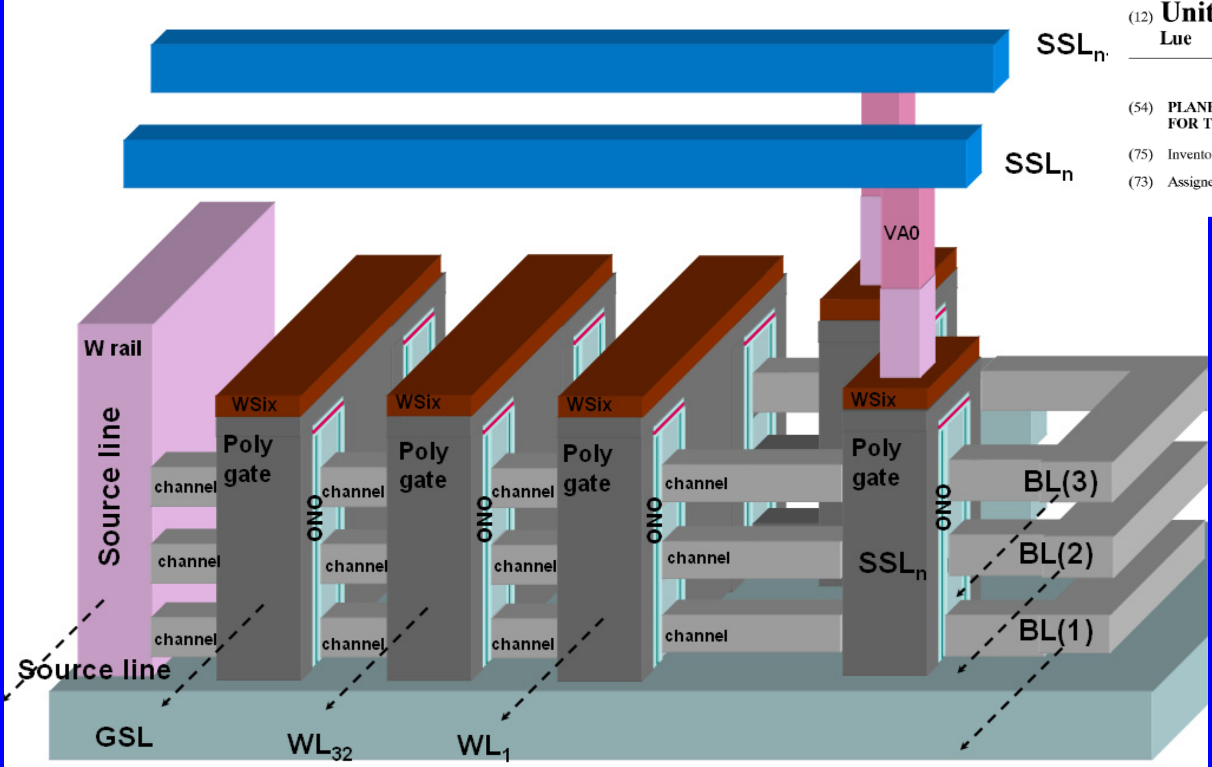
W. Kim, et al, VLSI Symposia, pp. 188-189, 2009.

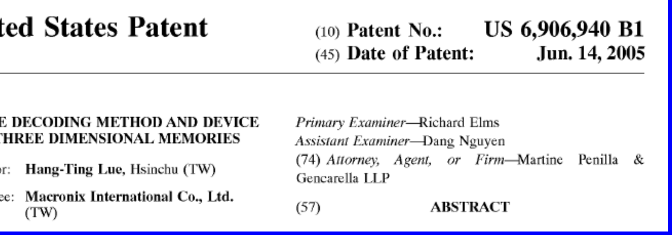
□ Relative large pitch. (>0.3um)

□ WL and BL located at the bottom.

□ The array decoding method (in-layer multiplex decoder) is very complicated, and wastes array efficiency

Modified VG NAND Architecture




(12) **United States Patent**
Lue

(10) **Patent No.:** **US 6,906,940 B1**
(45) **Date of Patent:** **Jun. 14, 2005**

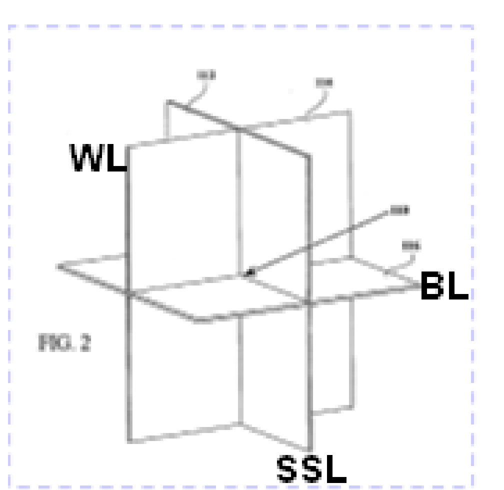
(54) **PLANE DECODING METHOD AND DEVICE FOR THREE DIMENSIONAL MEMORIES**

(75) Inventor: **Hang-Ting Lue**, Hsinchu (TW)

(73) Assignee: **Macronix International Co., Ltd.** (TW)

Primary Examiner—Richard Elms
Assistant Examiner—Dang Nguyen
(74) *Attorney, Agent, or Firm*—Martine Penilla & Gencarella LLP

(57) ABSTRACT



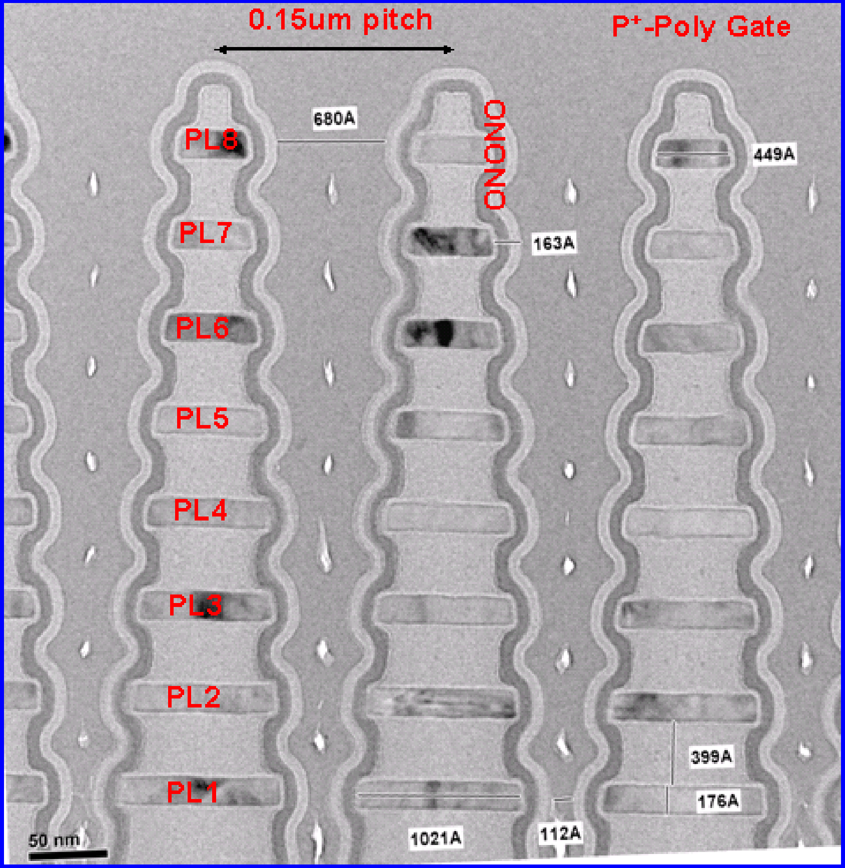
"Plane-decoding"

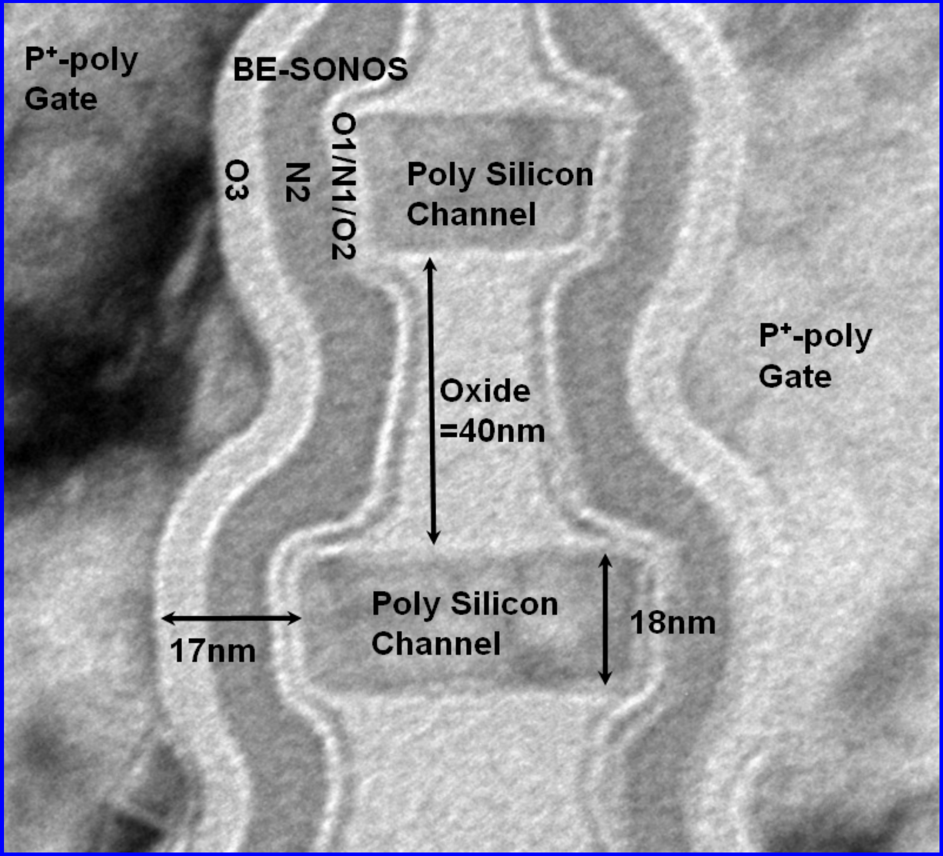
H. T. Lue, et al (Macronix), VLSI 2010.

- □ Conventional WL, BL are grouped into "planes".
- □ One additional SSL's device also grouped into "planes".
- □ Three planes select a memory cell.
- □ WL and BL can be at top.

Array X-Direction

H. T. Lue et al, VLSI 2010.





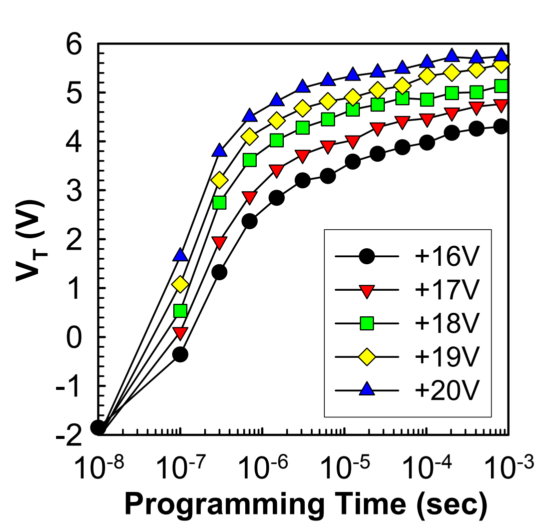
□ 75nm half-pitch, 8-layer device is fabricated

→ Equivalent cell size = **0.001406 um^2 (MLC)**

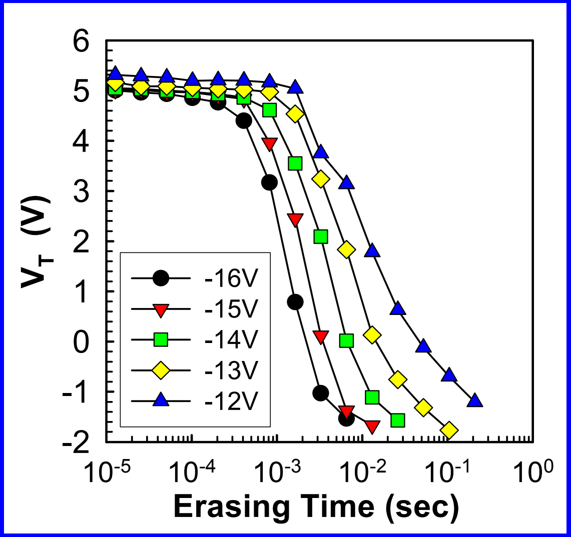
□ Each device is a double-gate TFT BE-SONOS device

Device characteristics of VG NAND

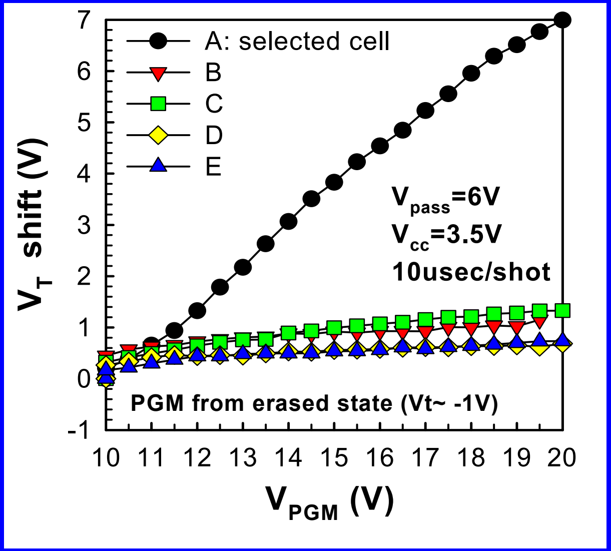
Programming



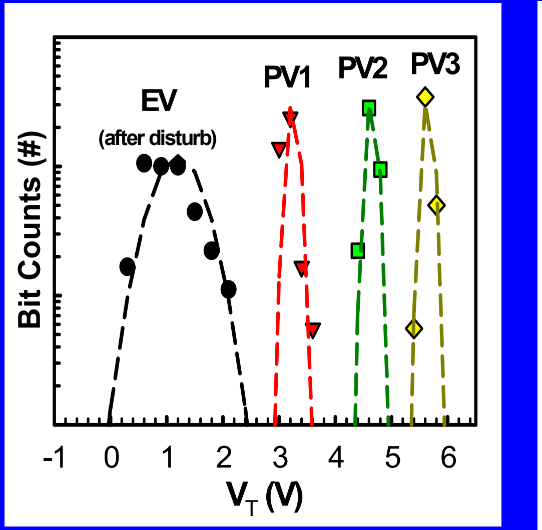
Erasing



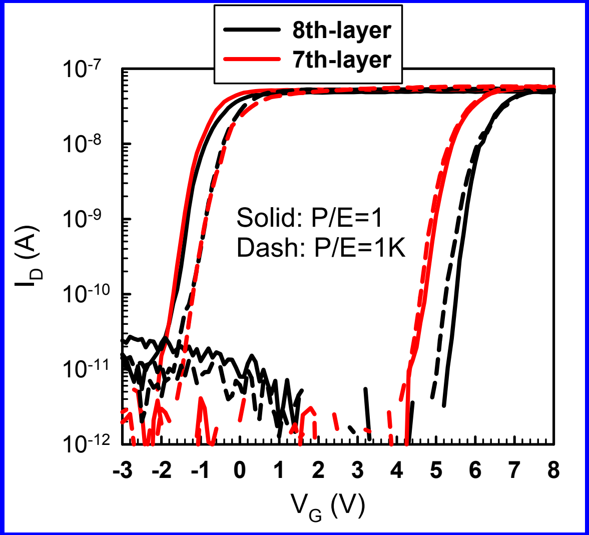
Program inhibit



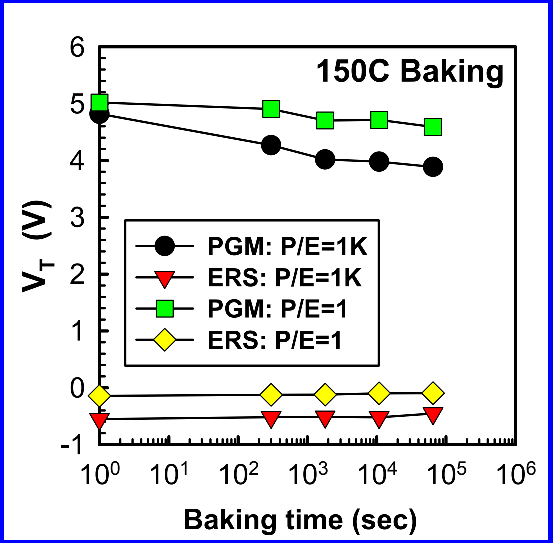
MLC capability



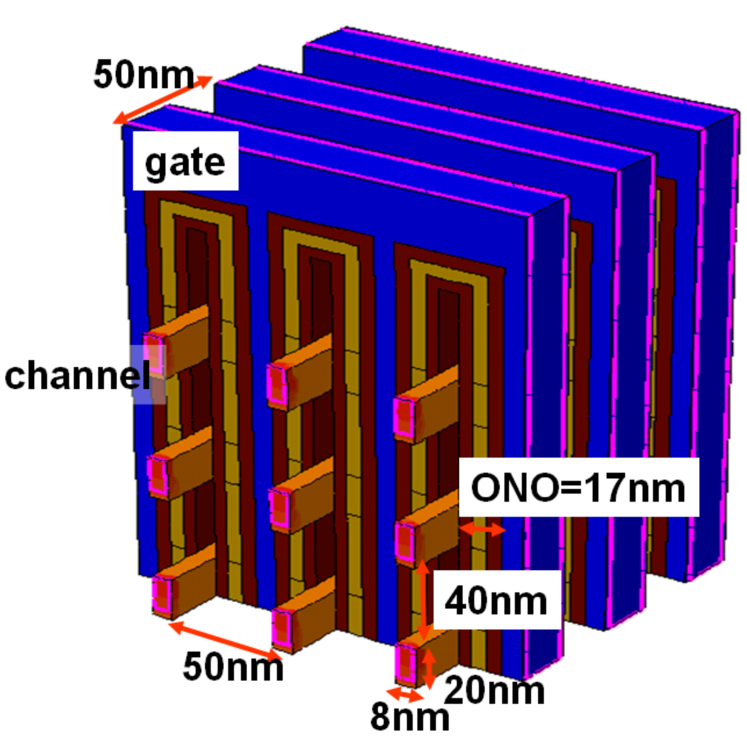
Endurance

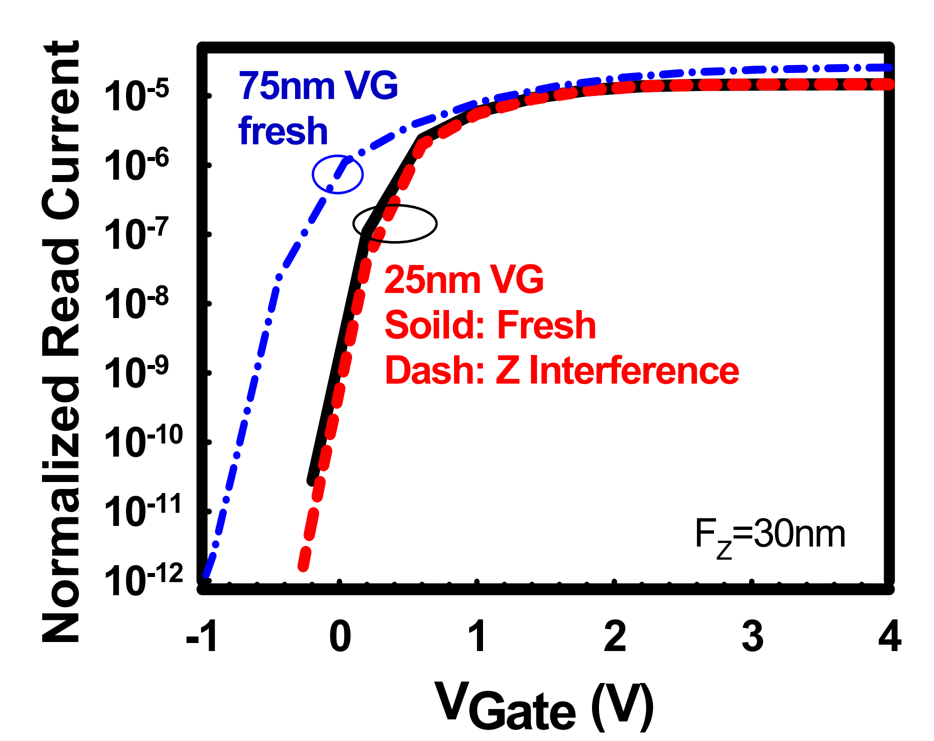


Retention



Scaling Simulation to 25nm





□ Scalability to 25nm is feasible based on the simulation.

Summary of 3D NAND

□ Many 3D memory architectures → Still hot topic

□ Scalability of cell size is important → Keep fewer memory stacks

□ Basic device physics is mostly known → No new materials except TFT

□ Decoding methods are key issues

□ Processing for the deep hole/trench is critical

□ Variability of TFT

Thank You for Your Attention!