NAND Device Physics

Akira Goda
Krishna Parat

Acknowledgement:

- **Intel and Micron Global NAND Team**

Micron/Intel Confidential

Outline

- Introduction – 2D & 3D Cell and Array Architecture
- Read – Cell I-V, Sensing, Interference, MLC, TLC
- Programming – Program Physics, Algorithms, PVS SSPC, Program Noise, Inhibit, Inhibit schemes
- Erase – Erase Physics, Algorithms
- Read Window Budget – Key components of RWB
- Cell Reliability Overview – Cycling, ICL, SBCL, Read Disturb, Floating Body effects, FPR, ECC

Micron/Intel Confidential

General Schedule

Day 1:

• Intro -	~30 slides	~1.5hr
• Read -	~30 slides	~1.5hr
• Single Cell Programming –	~15 slides	~1.0hr

Day 2:

• Array Program & Inhibit –	~30 slides	~1.5hr
• Erase –	~15 slides	~1.0hr
• RWB & Rel –	~25 slides	~1.5hr

➢ Feel Free to Interrupt and ask questions anytime during the class

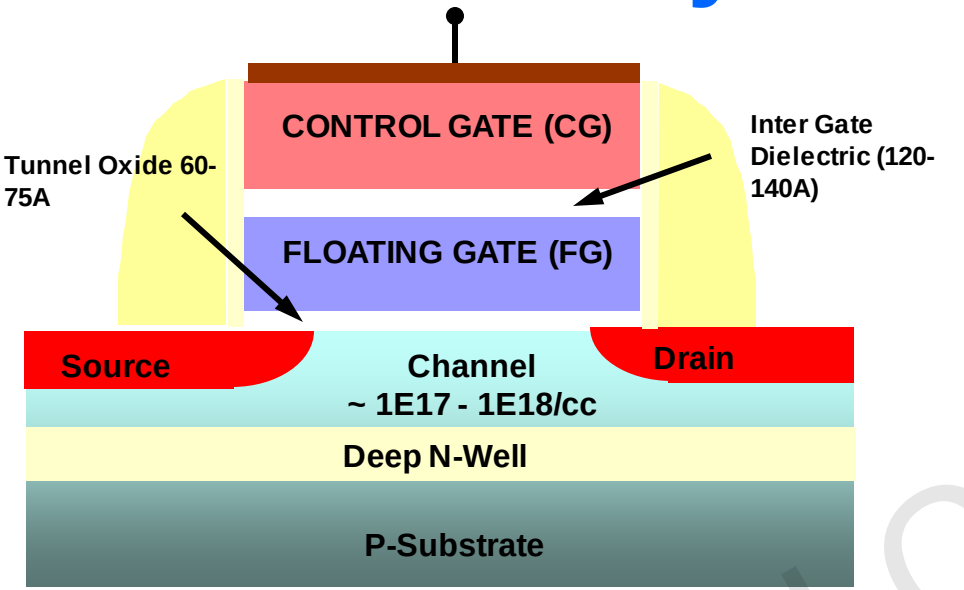
Micron/Intel Confidential

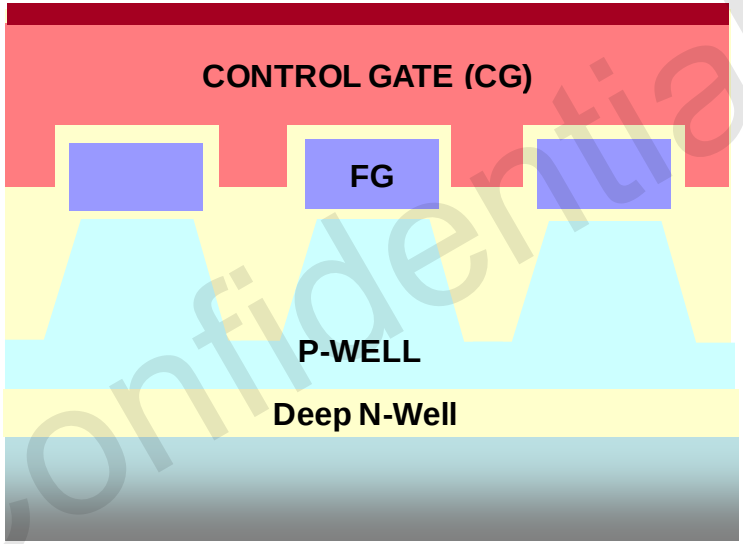
Introduction

- NAND Flash Memory Basics
- Device structure
- Device operation
- 2D Cell and String
- 2D Cell Structures
- 3D Cell
- 3D Cell Structures
- 2D Array
- 3D Array

Micron/Intel Confidential

Flash Memory Device (2D)

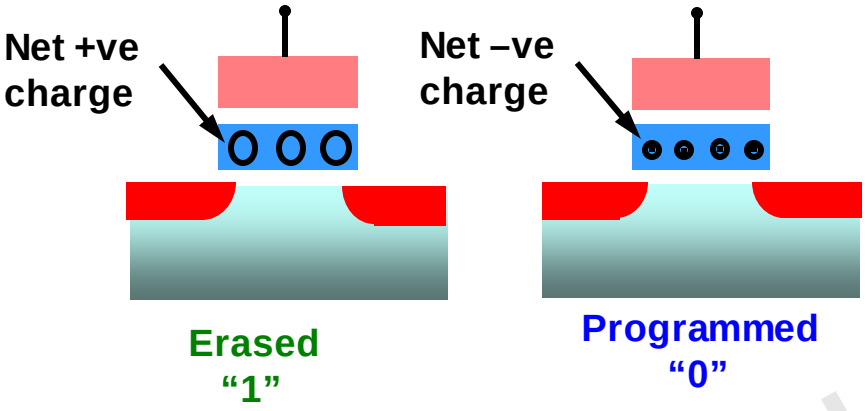


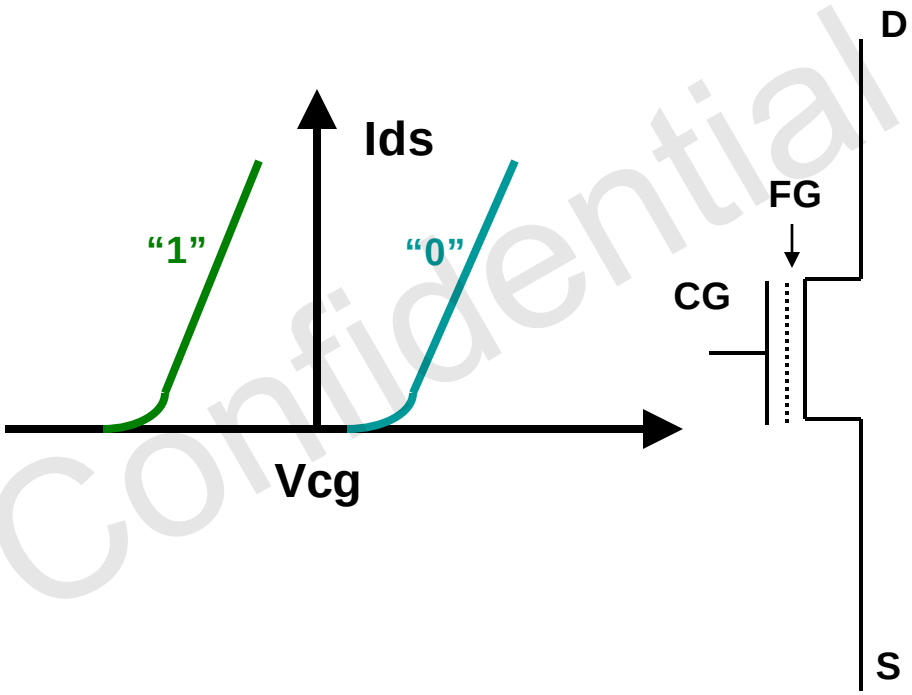


- Stacked Gate NMOS Transistor
  - Floating Gate: Charge storage node
  - Control Gate: Access/Control gate for the transistor
  - Tunnel-oxide: Gate oxide and dielectric through which charge is tunneled in and out of the floating gate (charge storage node)
  - Inter Gate Dielectric: Multi-layer/barrier dielectric with low leakage
  - Lightly doped Source/Drain Junctions optimized for leakage
  - Channel isolated from Si substrate to allow applying voltage to Channel

Micron/Intel Confidential

Flash Memory Device

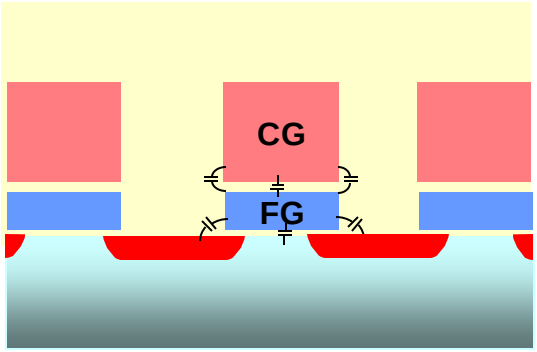


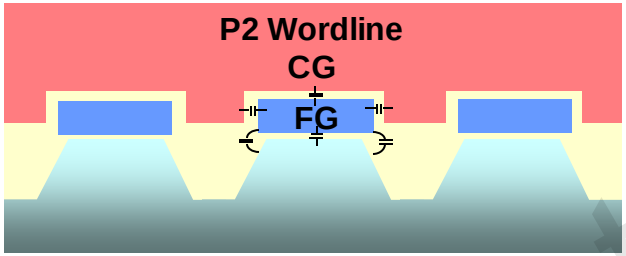


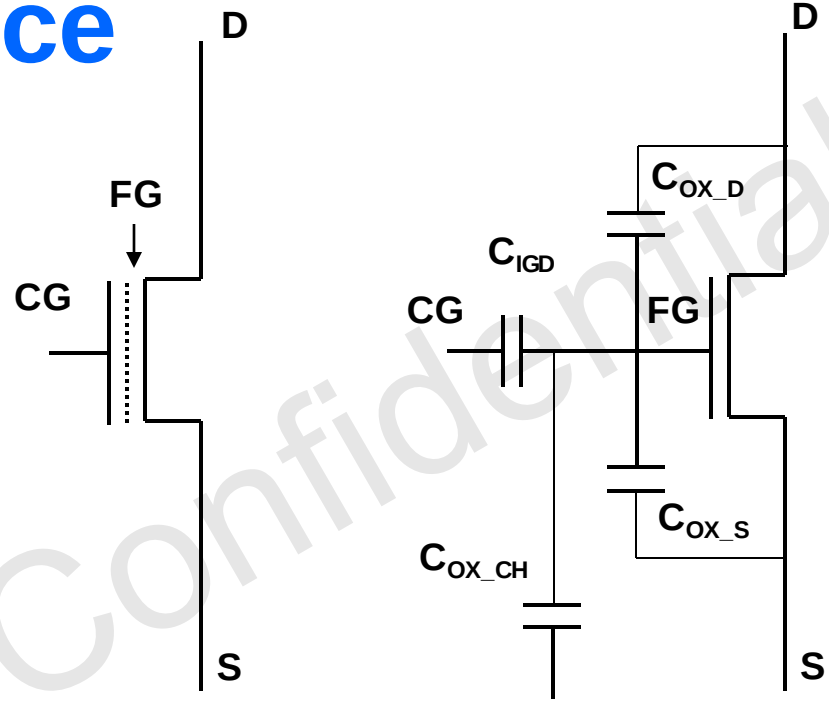
- Threshold Voltage shift = $-\Delta Q_{FG}/C_{IGD}$ { ΔQ_{FG} = Δ in charge on FG; C_{IGD} = CG to FG Capacitance }
- Program = Electrons Stored on the Floating Gate → High Vt
- Erase = Remove electrons from the Floating Gate → Low Vt
- Read = Look for current through the cell at given gate bias

Micron/Intel Confidential

Flash Memory Device





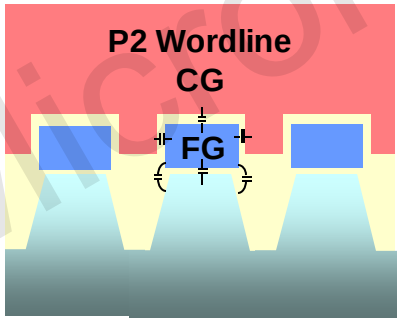


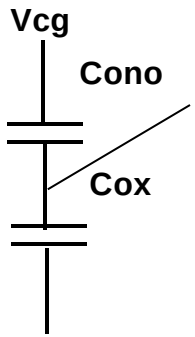
- Cell is a linear capacitor network with n-ch transistor attached
- Various cell capacitances can be calculated using TCAD
- FG Voltage, $V_{FG} = Q_{FG}/C_{TOT} + C_{IGD}/C_{TOT}*V_{CG} + C_{ox\_D}/C_{TOT}*V_D + \ldots$
  - Where, Q_{FG} is the net charge on FG, and $C_{TOT} = C_{IGD} + C_{OX\_D} + \ldots$
- Gate Coupling Ratio (GCR) = $dV_{FG}/dV_{CG} = C_{IGD}/C_{TOT}$
- High GCR is desirable for CG to have a good control of FG & Channel

Micron/Intel Confidential

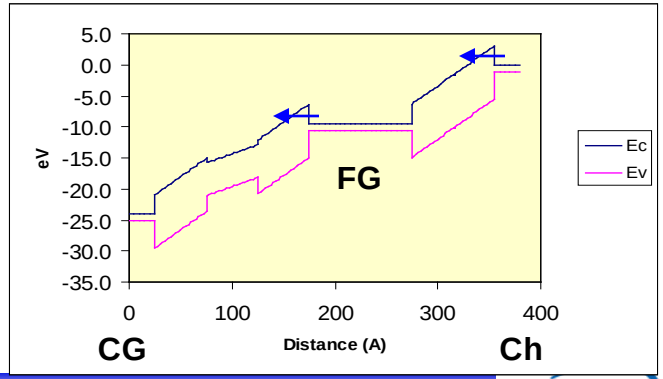
Flash Dielectric Constraints

- Data retention limits the Tunnel-oxide scaling to ~55-75A
  - Intrinsically a ~55A silicon dioxide with ~3eV barrier to N-Poly can provide the needed retention (<1E-16A/cm^2 of leakage).
- IGD thickness dictated by retention as well as the need for large Program/Erase window (high field leakage during Program/Erase)
  - ONO as IGD scaled down to ~110-120A Physical (EOT ~ 100A)
  - Hi-K IGD scales down to ~150-160A Physical (EOT ~ 80A)
  - High barrier important for low leakage
  - Low electrical thickness important for good gate control
  - Good Gate control can be achieved by increasing the coupling area between the CG and FG → CG wrap of the FG
- High doping of the FG and CG desired to minimize poly depletion



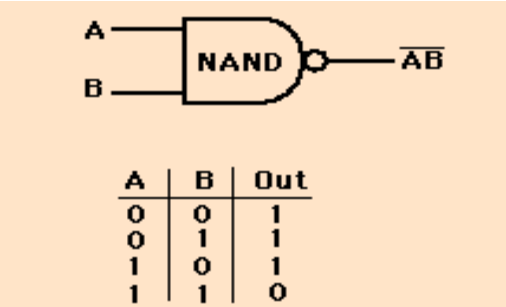


Vfg = Vcg* Cono/(Cox+Cono)

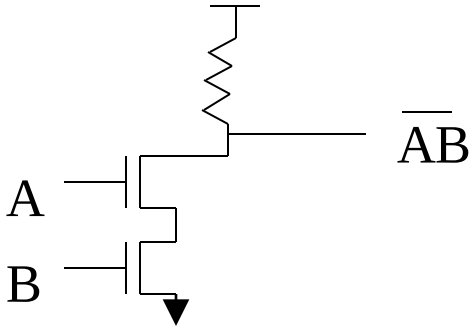


Micron/Intel Confidential

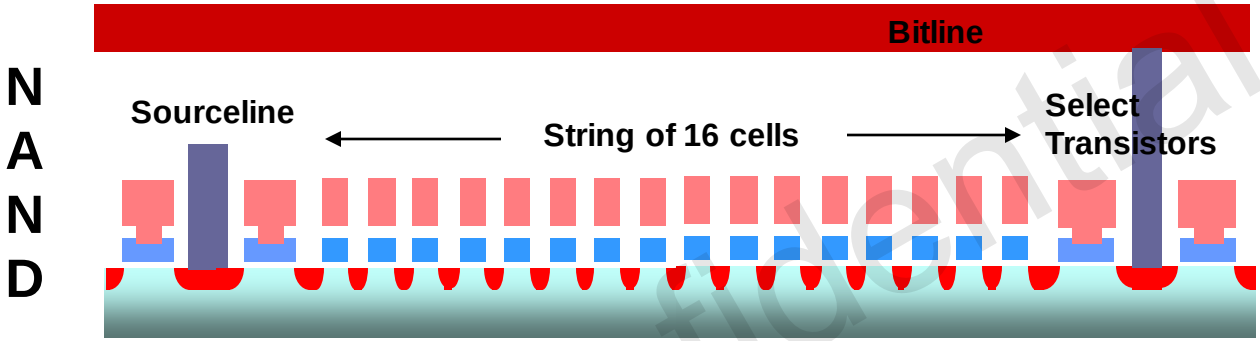
NAND Schematic Cross-sections



A	B	Out
0	0	1
0	1	1
1	0	1
1	1	0



1. Transistors connected thru each other to output & ground.
2. Current flows simultaneously through all transistors

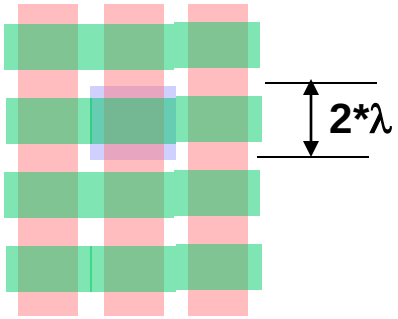


NAND string typically has 32-128 cells (+ Dummy cells). Schematic shows example with 16 cells.

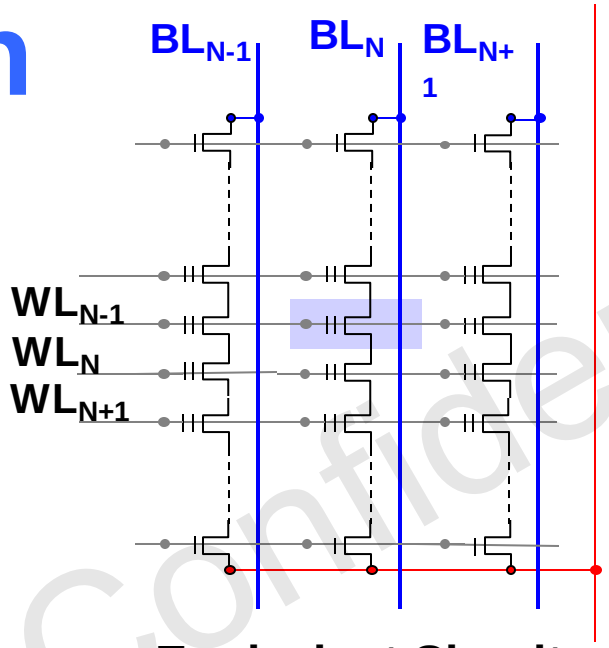
- Individual NAND cells do NOT require a contact → very compact cell layout (2D NAND - typically 4-5λ^2)
- One select transistor (SGD & SGS) at each end of the NAND string
- Select Gates, however, adds additional area overhead of about ~20%

Micron Intel Confidential

2D NAND Flash

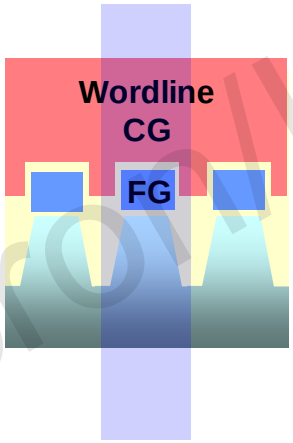


Layout

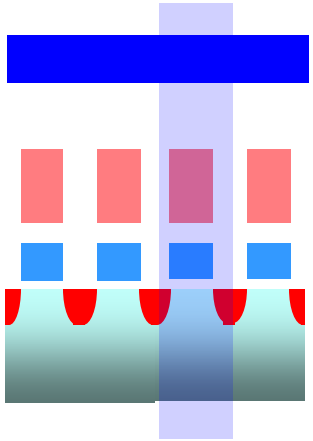


Equivalent Circuit

Select Transistors shown in the equiv ckt, but not shown in the layout or cross-sections



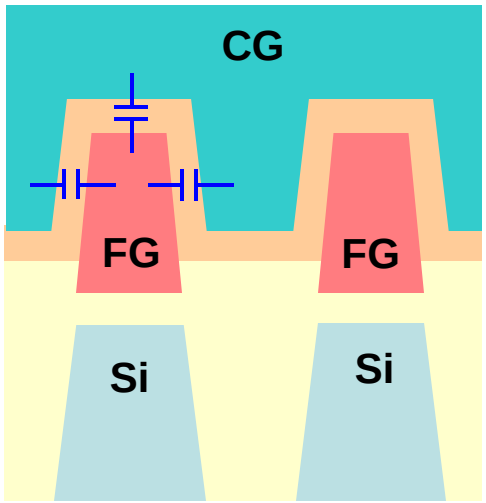
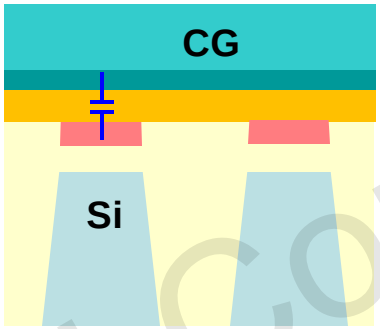
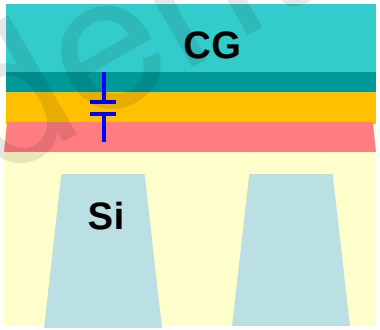
Y Cross-section



X Cross-section

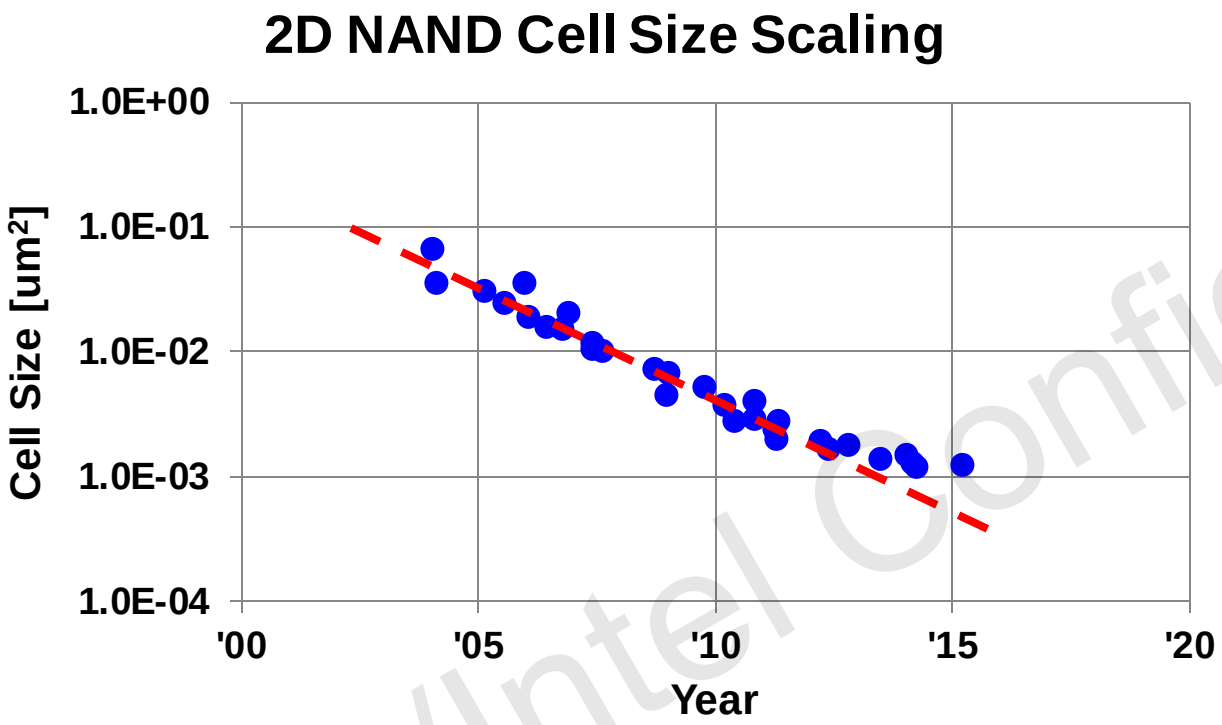
Micron/Intel Confidential

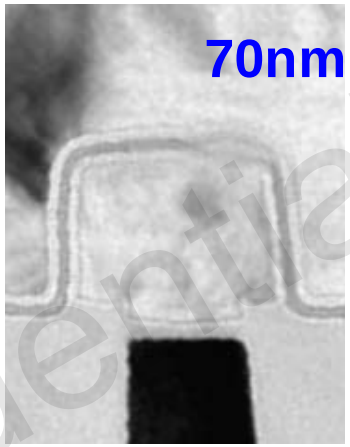
2D Flash Memory Cell Structures

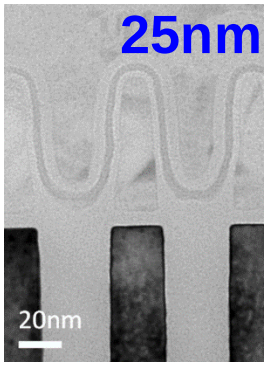
		
Wrap FG Cell	Planar Floating Gate Cell	Planar Charge Trap Cell
• NAND Technology up to 70s • Polysilicon Floating Gate • ONO IPD (IGD) • Polycide/Metal Control Gate	• 80s & 90s Technology • Hybrid Floating Gate • Multi-barrier dielectric Blocking dielectric (BD) • Metal Control Gate	• SONOS/TANOS • Nitride / Dielectric Charge storage • Oxide/High-K Blocking dielectric (BD) • Metal Control Gate

Micron/Intel Confidential

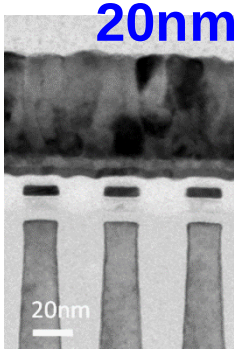
2D NAND Cell Scaling

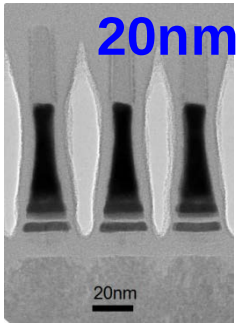






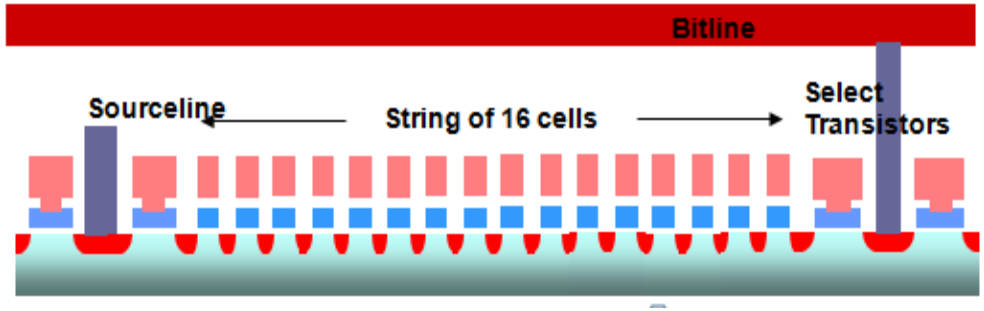
- 2D NAND scaling has slowed down:
  - Lithography Limitations
  - Small Cell Area Effects: Number Fluctuation
  - Proximity Effects: Cell to Cell interference
  - High Electric Field Effects





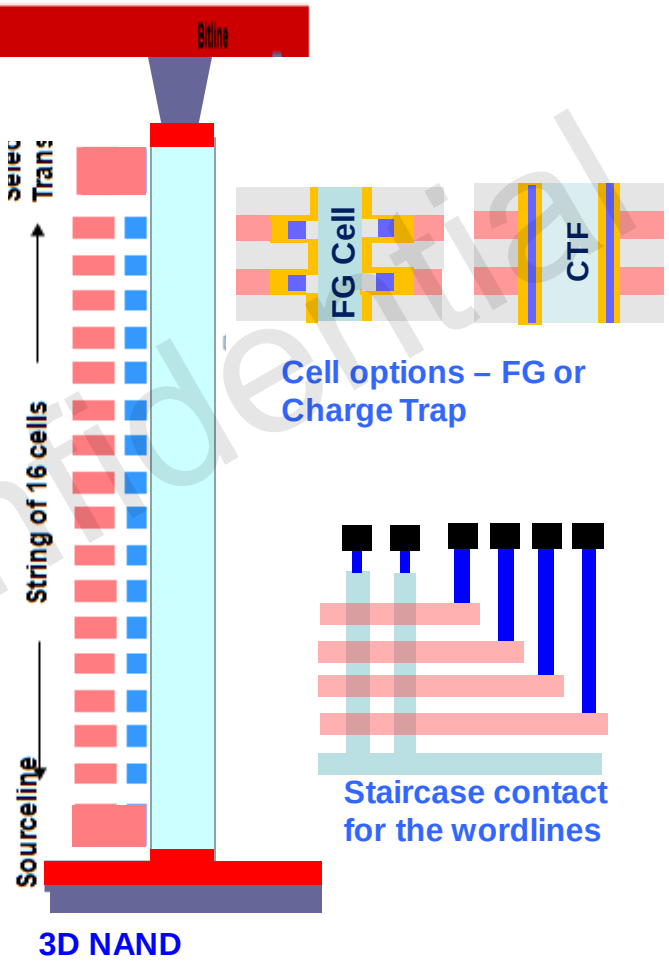
Micron/Intel Confidential

2D NAND → 3D NAND



2D NAND

Only 16 cells/string shown for clarity.
80s (L85) has 128 cell string.



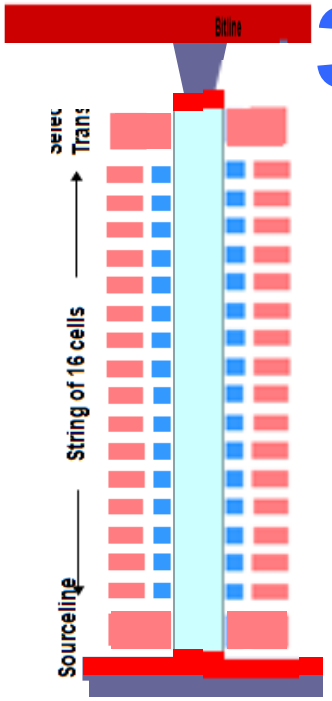
3D NAND

Only 16 cells/string shown for clarity.
Actual number of tiers is more.

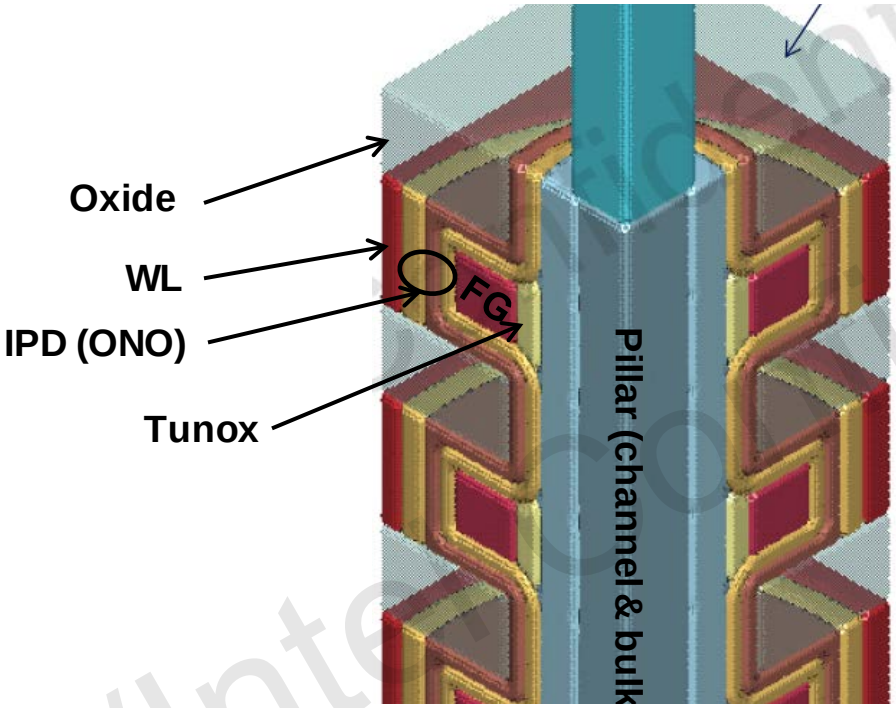
- Vertical NAND string
- Conductive or dielectric charge storage node
- Deposited poly-silicon channel
- New WL contacting scheme
- Large foot-print of the 3D Cell (~16X) will require quite a few layers to be stacked to achieve effective cell area scaling
- Increased demands on process technology – very high aspect ratio etches and fills

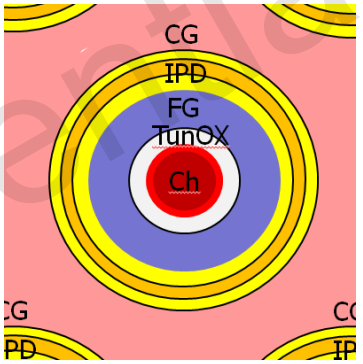
Micron/Intel Confidential

3D NAND Cell and String



3D NAND

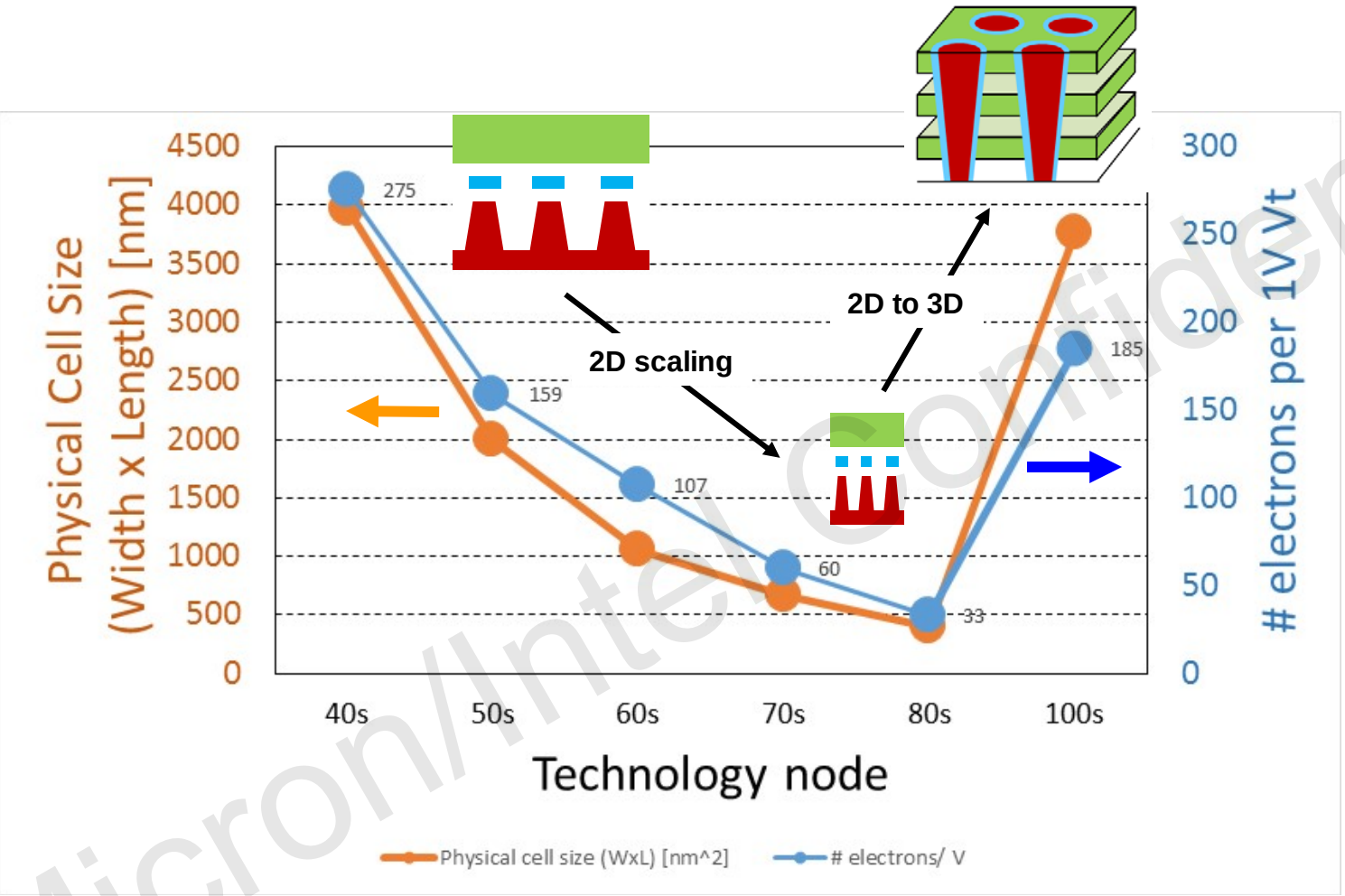




- Center of string is a cylinder of polysilicon: the “pillar”
- The pillar serves as the silicon bulk, and contains the channel
- The cell stack (Tunnel-oxide, Charge Storage Node, IGD, Control Gate) wraps around the pillar
- Adequate Gate coupling achieved through higher IGD/CG area

Micron/Intel Confidential

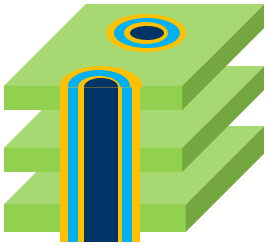
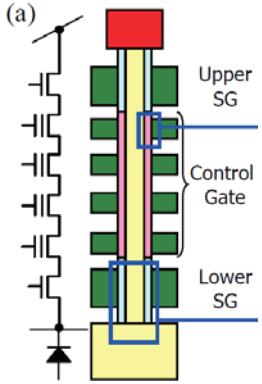
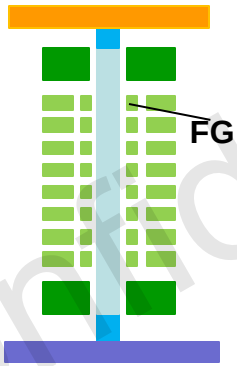
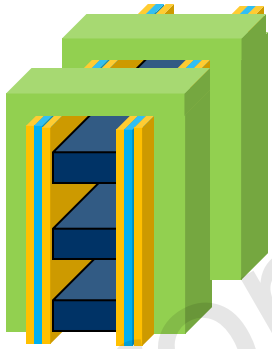
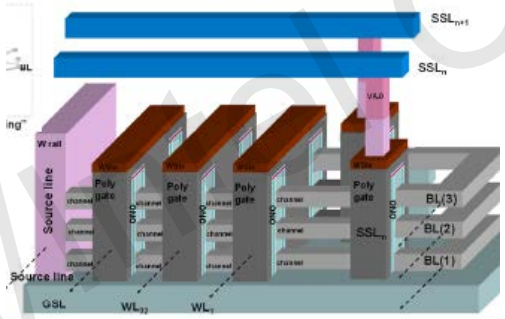
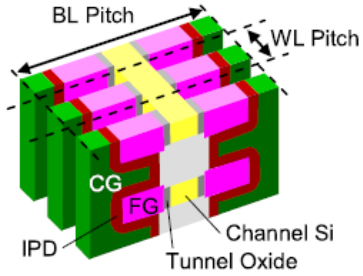
2D NAND to 3D NAND ~ Cell size~



3D NAND has a very large cell size owing to the gate all around architecture, realizing excellent electrical properties.

Micron/Intel Confidential

3D NAND Gallery: String and Cell

	SiN trap cell	Floating gate cell
Vertical string	  SiN trap H. Tanaka et al., VLSI 2007	  Intel-Micron
Horizontal string	  H.T. Lue et al., VLSI 2010	 K. Sakuma et al., EDL 2007

[Vertical String] Gate-All-Around, Cell properties advantage

[Horizontal String] Small cell size, Tier stack height/cell foot print advantage

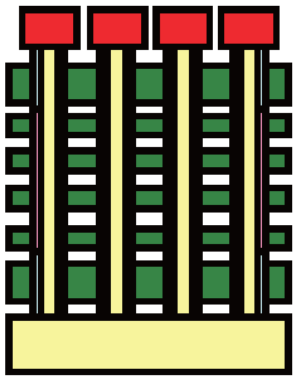
[SiN trap cell] Process integration & Low FGFG coupling advantage

[FG cell] Cell reliability advantage

Micron/Intel Confidential

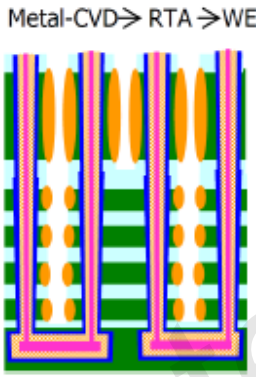
3D NAND Gallery: WL technology

Poly Si WL



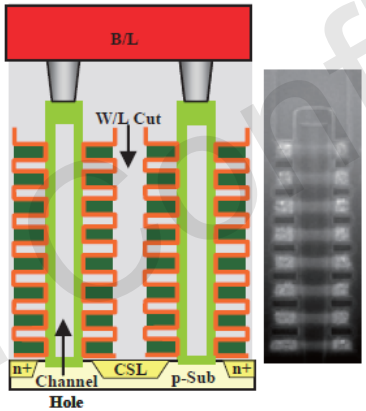
H. Tanaka et al.,
VLSI 2007

Salicided WL



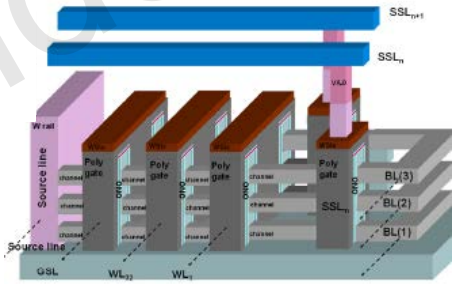
M. Ishiduki et al.,
IEDM 2009

Tungsten WL



J. Jang et al.,
VLSI 2009

WiSix WL

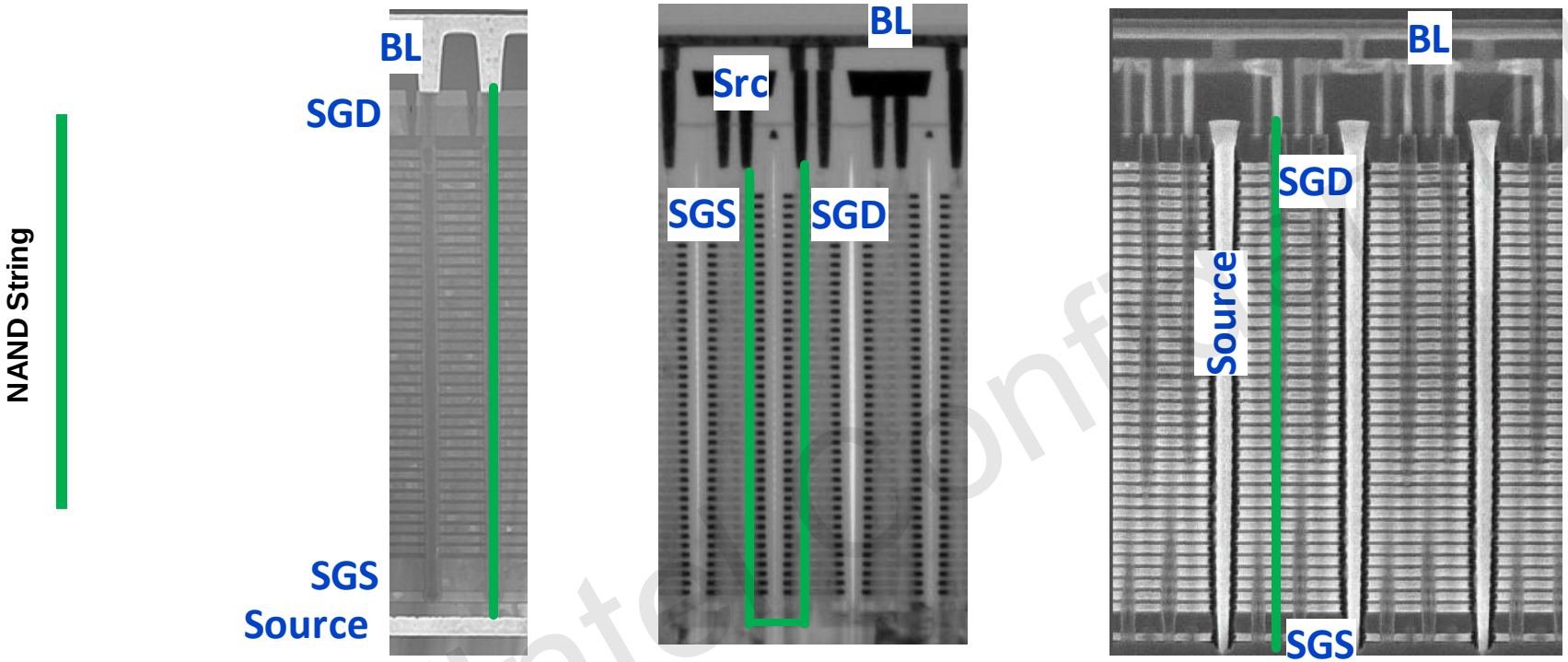


H.T. Lue et al.,
VLSI 2010

[Poly Si WL] Process integration and small foot print advantage.
[Low res. WL] Enabling long WL, string driver layout advantage.

Micron/Intel Confidential

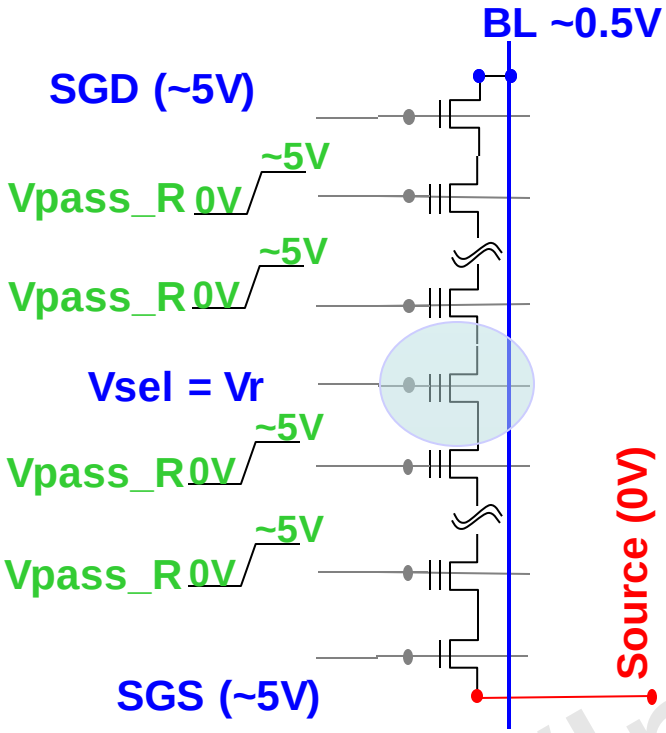
3D NAND Gallery ~ Production



	Intel-Micron 3D FG	Hynix SMArT	Samsung/Toshiba V-NAND
Cell	FG	SiN trap (BE-TANOS)	SiN trap (BE-TANOS)
WL	Poly	Replacement W	Replacement W
SGD	MOS	Cell	Cell
SGS	MOS	Cell	MOS
String	Vertical, Straight	Vertical, U-shape	Vertical, Straight
CMOS	Under Array	Outside Array	Outside Array

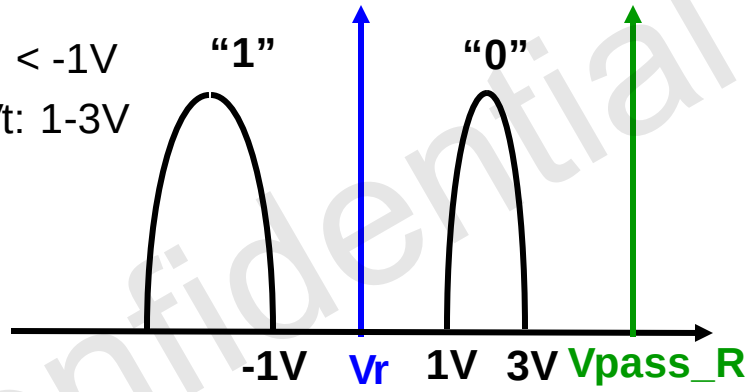
Micron/Intel Confidential

NAND Flash Read



Erased Cell Vt: < -1V

Programmed Cell Vt: 1-3V



- Bitline is biased to small voltage (~0.4-0.5V). SGD & SGS of the selected block (string) are turned on (Vg ~5V).
- Selected wordline is biased to Read Gate voltage (Vr). All the other wordlines in the string are biased to Vpass_R voltage (> the highest program Vt).
- If the selected cell Vt < Vr, the string will conduct and there will be current on the bitline. If the selected cell Vt > Vr, the string will not conduct and the current on the bitline will be zero. Sensing circuits detect this.

Micron/Intel Confidential

SLC, MLC, & TLC

SLC: 1 bit/cell: 2 states:

Level 0 = Erased

Level 1 = Programmed

MLC: 2 bits/cell: 4 states:

Level 0 = Erased

Level 1 = Pgm to Vt1

Level 2 = Pgm to Vt2

Level 3 = Pgm to Vt3

LSB (LP) Read with Vg = R2

MSB (UP) Read with Vg = R1 & R3

TLC: 3 bits/cell: 8 states:

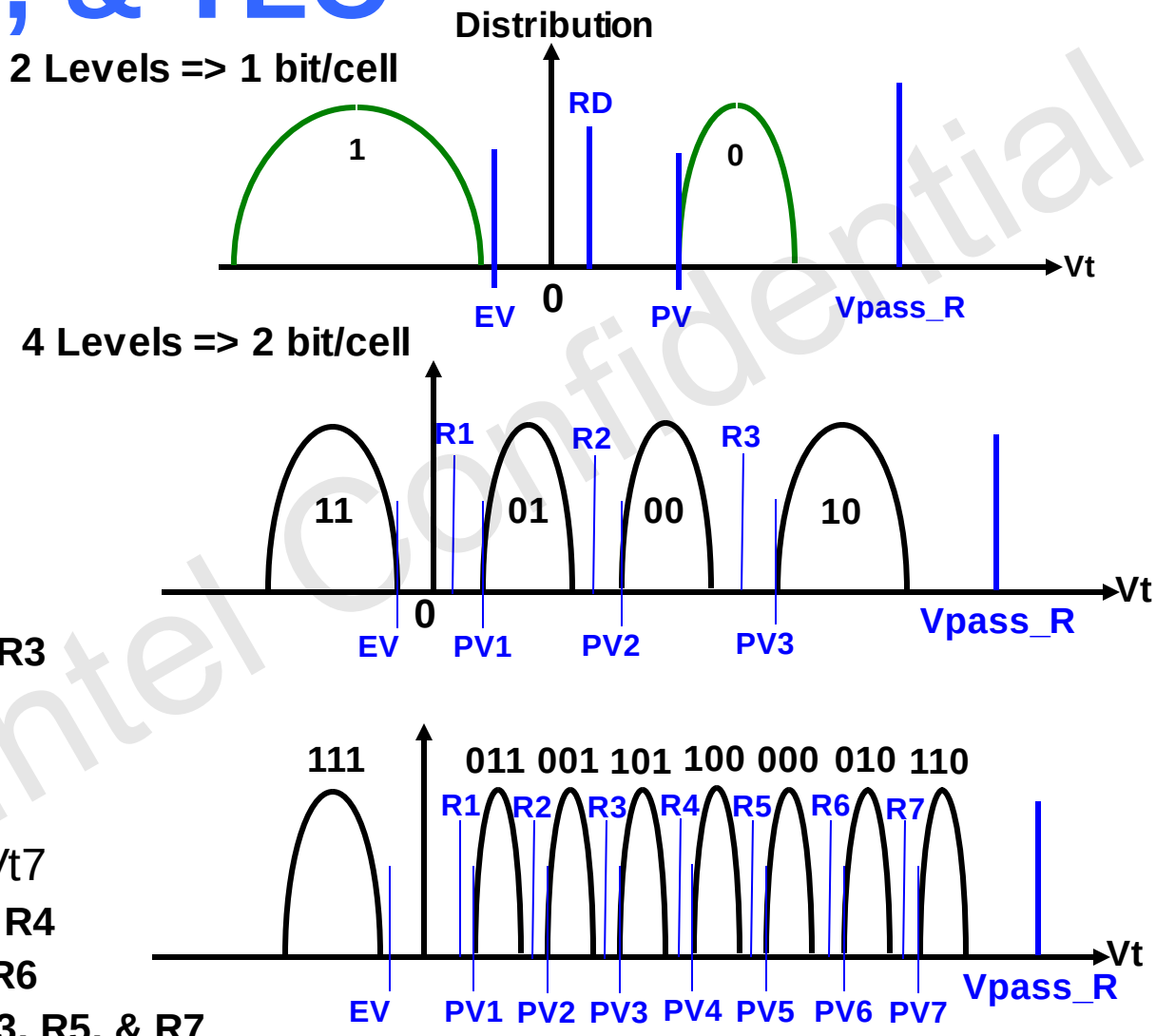
Level 0 = Erased

Level 1-7 = Pgm to Vt1 – Vt7

LSB (LP) Read: Read with Vg = R4

?SB (UP) Read with Vg = R2 & R6

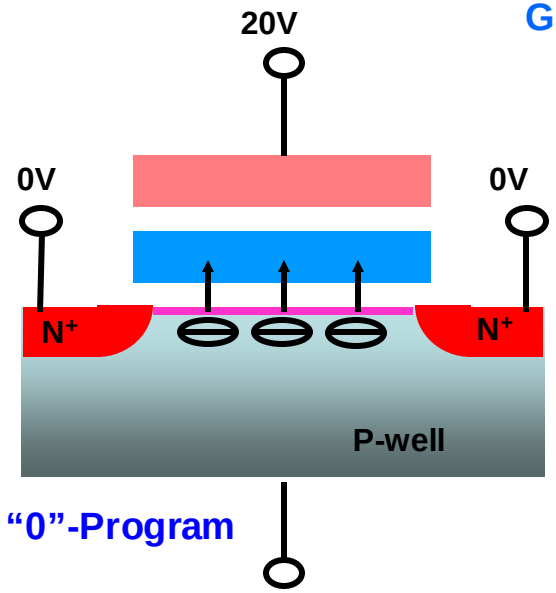
MSB (XP) Read with Vg = R1, R3, R5, & R7

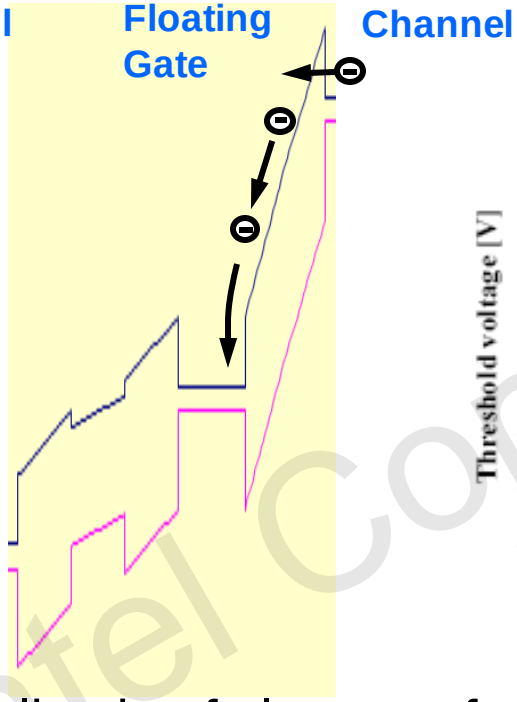


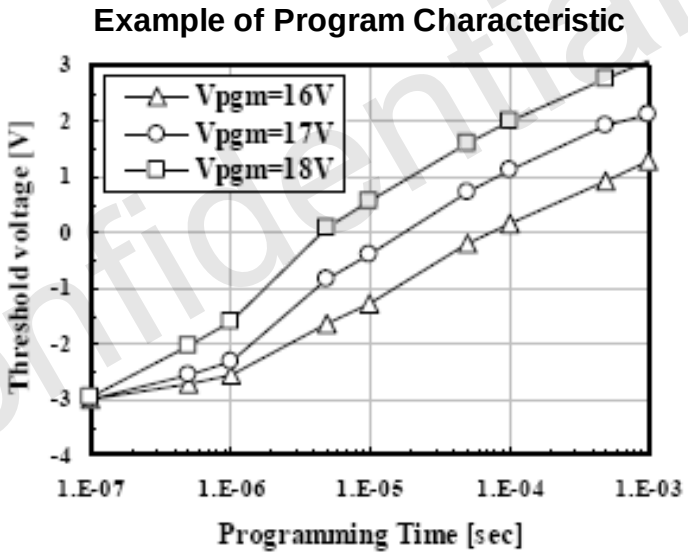
- **Having the threshold voltage of the cells defined to 2^n different levels allows for storing n bits per cells**

Micron/Intel Confidential

NAND Flash Programming - FN Tunneling



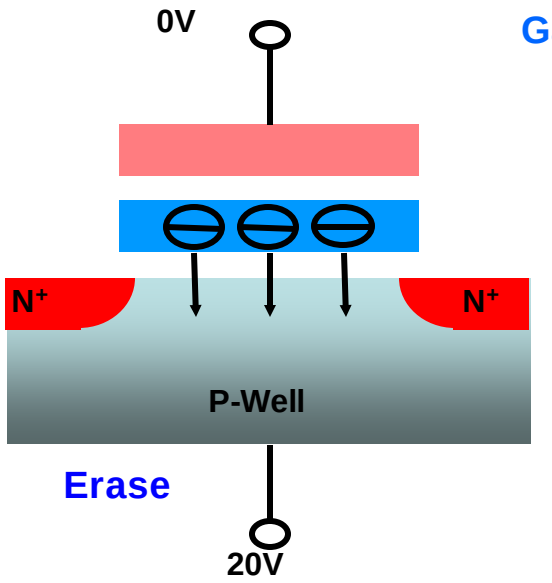


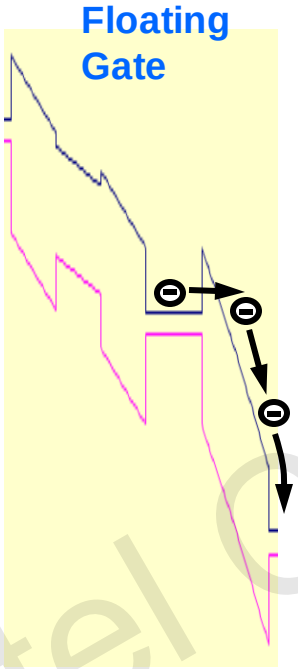


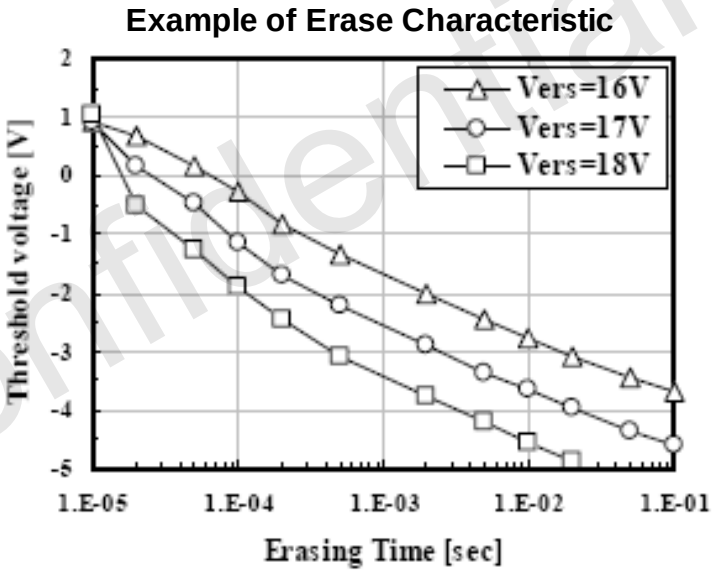
- Programming is by Tunneling in of electrons from the channel into the FG by biasing the Control Gate positive w.r.t the channel and S/D.
- "Effective" Program Time (cumulative pulse time ignoring set up and verify Overheads) ~100-200us.
- Tunnel-oxide E-field ~11-12MV/cm. Tunneling Current ~ 10mA/cm^2.
- Program page size ~ 16KB

Micron/Intel Confidential

NAND Flash Erase - FN Tunneling







- Erase is opposite of Programming by reversing the applied biases.
- Tunneling out of electrons from the FG into the channel by biasing the Well positive w.r.t the Control Gate. Erase time ~1-10ms.
- Tunnel-oxide E-field ~11-12MV/cm. Tunneling Current ~ 1-10mA/cm^2.
- Erase Block size ~ 24MB/96MB/36MB (B0KB/B1MA/B16)

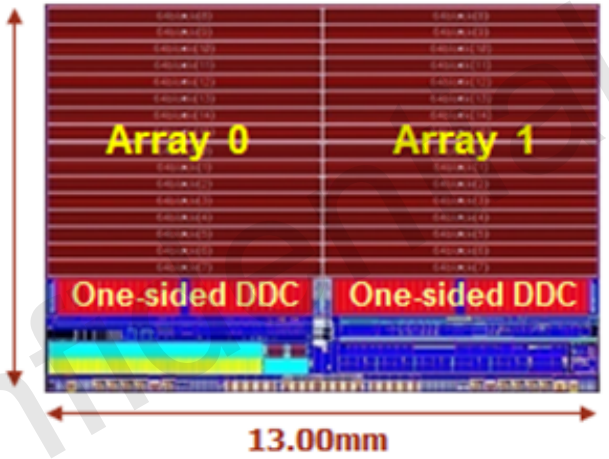
Micron/Intel Confidential

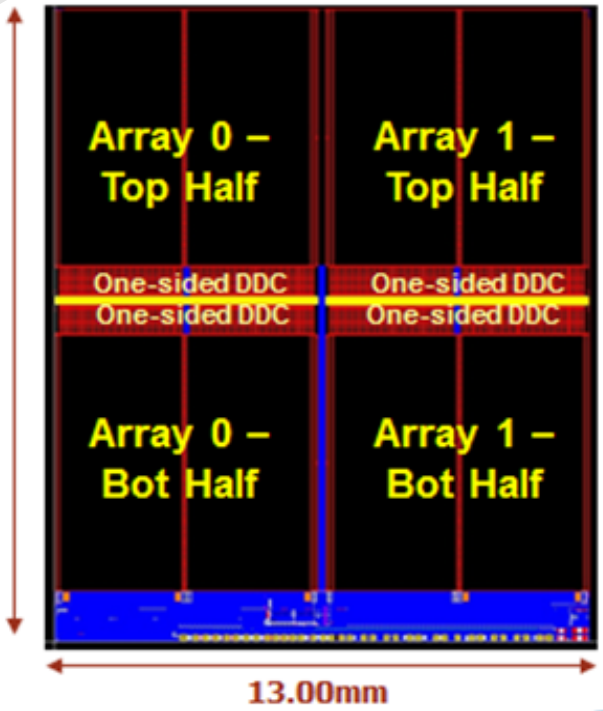
More on Read, Program and Erase Later

Micron/Intel Confidential

2D NAND Array – 80s

- **L84A**
  - Dual Plane
  - Bottom 1-sided 8KB DDC
  - 68 NAND String
  - Mixed String Driver **(Edge & Center)**
  - ECC: 1KB code word
  - Pad arrangement: 1-sided long
- **L85A**
  - Dual Plane – Folded Array
  - 1-sided 8KB DDC for half plane
  - 132 NAND String
  - Staggered Mixed String Driver
  - ECC: 2KB code word
  - Pad arrangement: 1-sided short

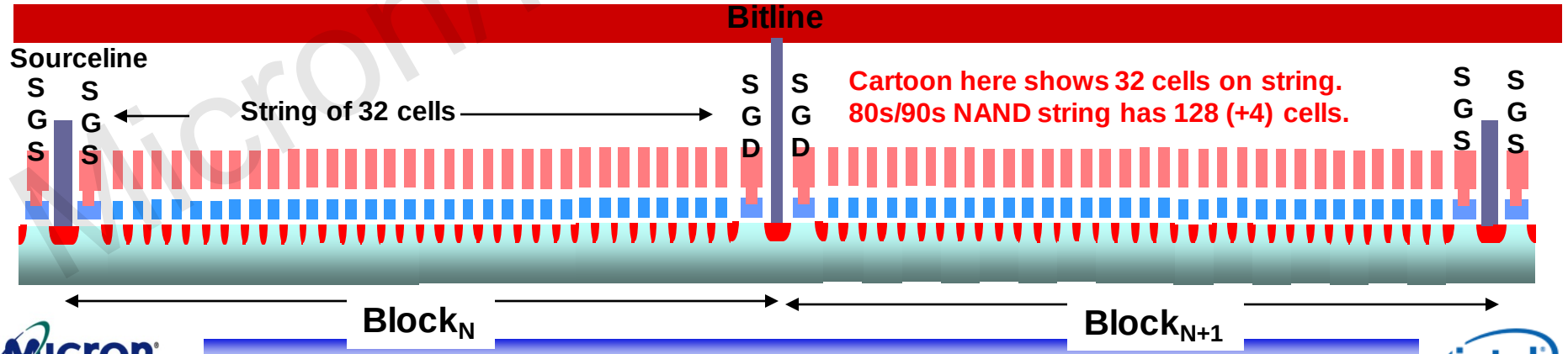




Micron/Intel Confidential

2D NAND Array

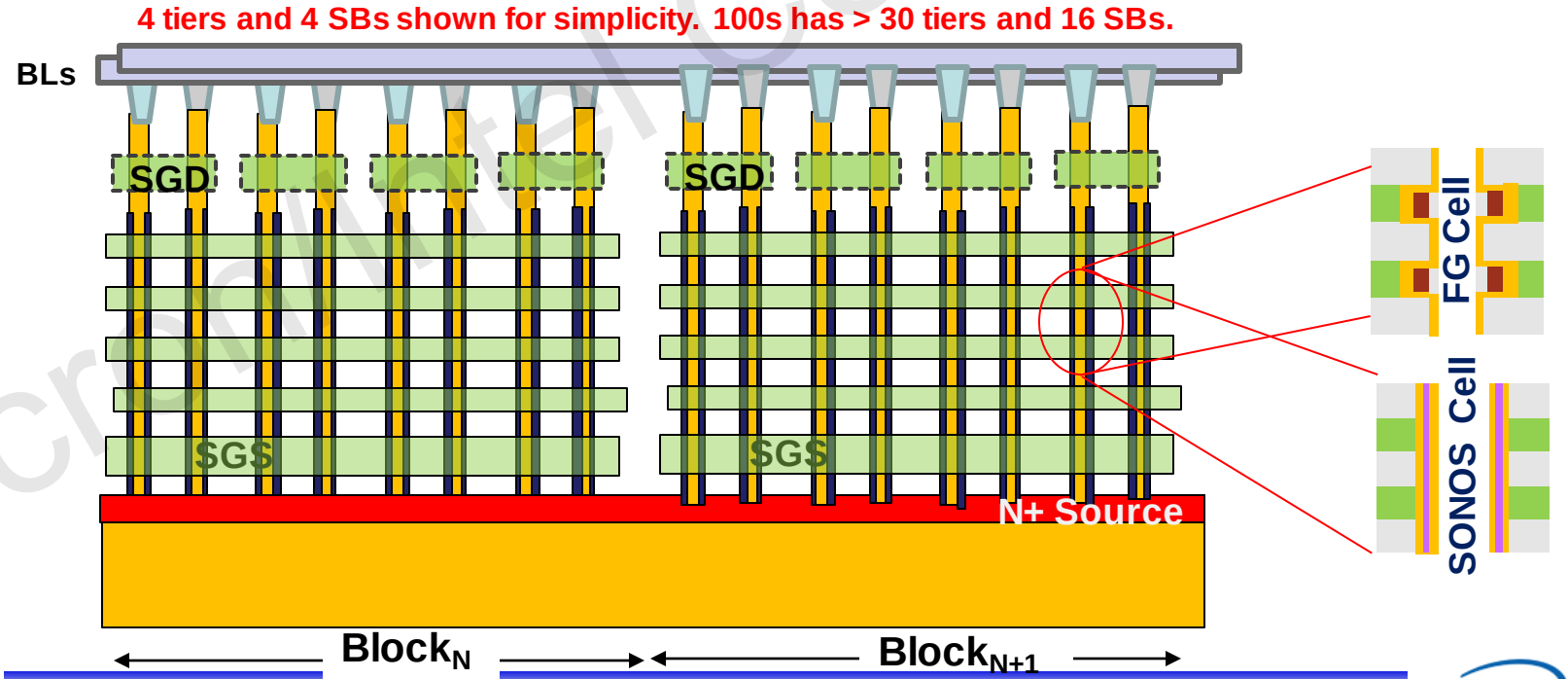
- Die is typically divided into 2 or 4 planes
- All blocks in a plane typically sit in a single P-Well in a single Deep N-Well (Referred to as ATUB)
- Bitlines (columns) typically span the full length (or half length) of the die
- Wordlines typically span the full width (or half the width) of the die
- Blocks are separated by the select transistors. A given string or a block is selected by selecting the appropriate Select Gates (SGS & SGD)
- Bitlines contact the drain side of the Select Gate Drain
- Source is common to the entire plane
- A block has 128 Rows (Wordlines). With 64-128 cells in the string, one Block = (64-128 WLs) * # of Cell on the WL (8-16KB)



Micron/Intel Confidential

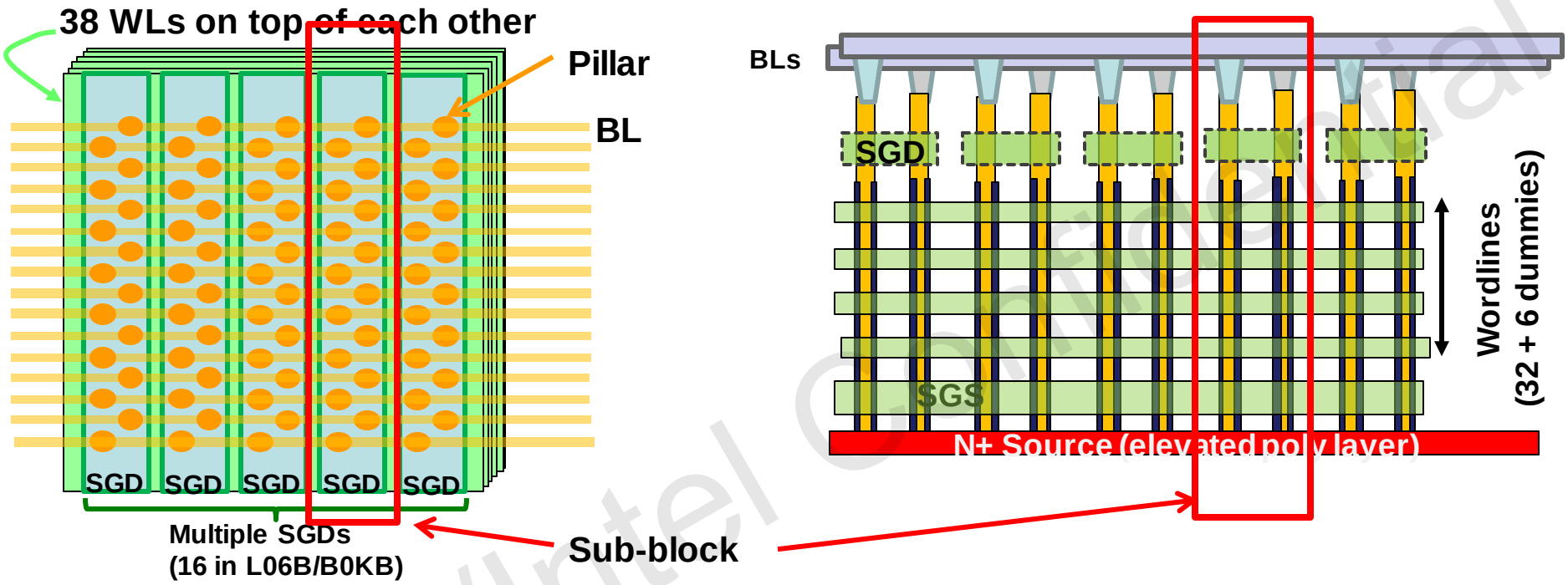
3D NAND Array (L06B/B0KB)

- 3D NAND introduces additional unit of array – sub-blocks. An erase block is divided into 16 sub-blocks (L06B/B0KB) – selected by the SGD.
- Wordlines are shared across the sub-blocks in a block.
- SGS is shared across the sub-blocks in a block.
- Pillar is the NAND string and the Atub.
- L06B/B0KB: With 32 tiers, 16 sub-blocks (SGDs/block), and 16KB page size, one Block = (32 WLs) * # of Cell on the WL (16KB) * 16 sub-blocks



Micron Confidential

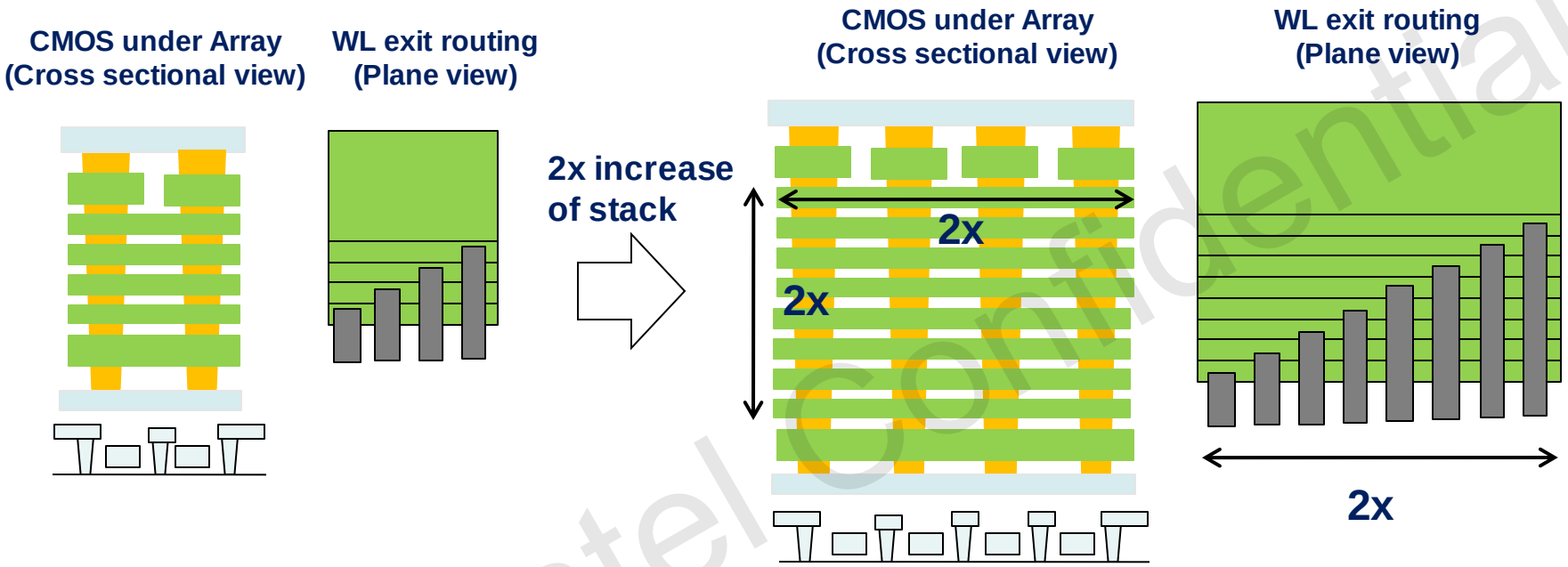
3D NAND Block (L06B/B0KB)



- Wordline: Many cells wide (16 in L06B/B0KB)
  - Ideally would have WL just one cell wide (as in 2D), but cannot fit string drivers/WL hook-ups for all the WLs into that small a pitch
    - ➢ So WL is wide to fit the string drivers and individual NAND strings are selected by SGDs – Multiples SGDs in a block → Many Sub-blocks / Block
- One page = intersection of one WL and SGD
- One cell = intersection of one WL, one BL, and one SGD

Micron/Intel Confidential

What Decides SGDs/Block



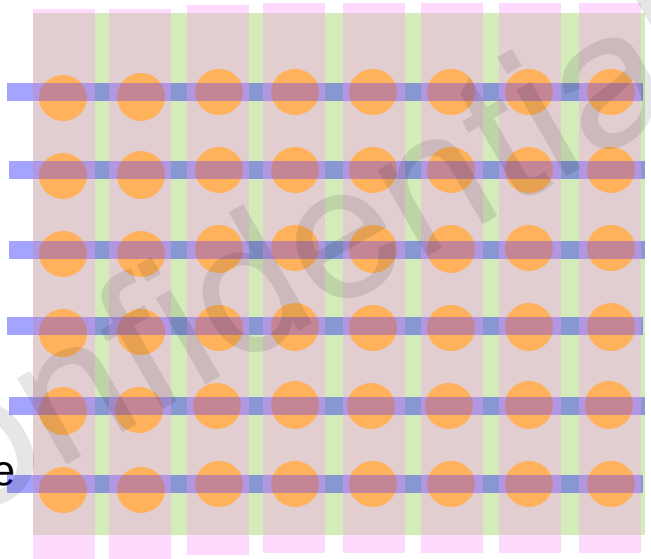
- Cell is 3D while CMOS circuit and routing are 2D.
- Cell 3D scaling (stacking) requires more 2D area for circuit and routing.
- This results in # sub-block(SGDs) increase → increased disturb stress and WL capacitance. WL RC remains unchanged.

Micron/Intel Confidential

3D NAND Pillar/SGD Layout

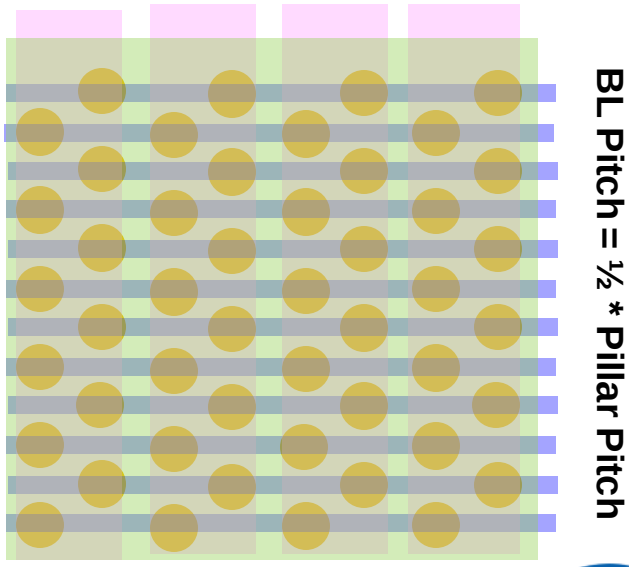
Simple Case

- Simple arrangement of pillars in square grid.
- BL Pitch = Pillar Pitch
- SGD Pitch = Pillar Pitch
- Pillars are not closely packed ➔ Poor array efficiency
- Smaller Page size ➔ Lower Program Bandwidth
- Block Size (WL width) decided by the String Driver layout needed to support WL and SG drivers for all the tiers and SGs



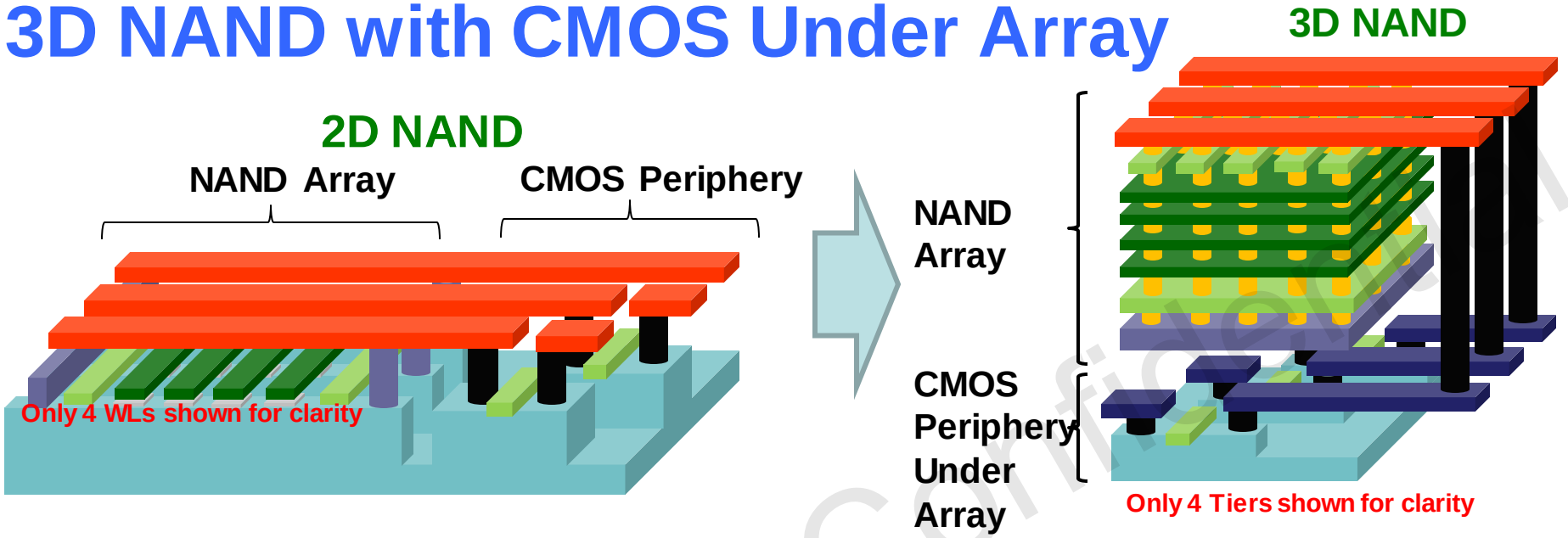
Stagger the Pillars (L06B/B0KB and L17A/B1MA):

- Hexagonal Close Packed Pillars ➔ Best array efficiency
- BL Pitch = ½ Pillar Pitch ➔ Capable of 2X page size ➔ Improves the Program bandwidth by 2X
- ~Doubles the SGD Pitch ➔ For a given WL width (dictated by String Driver layout) number of SGDs/sub-blocks is halved.

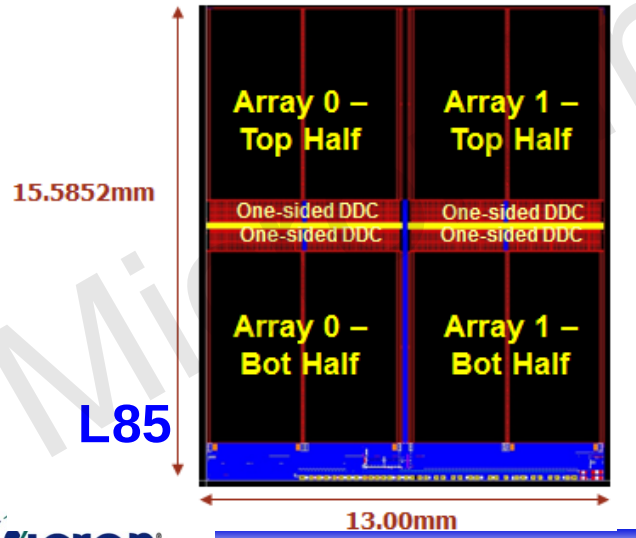


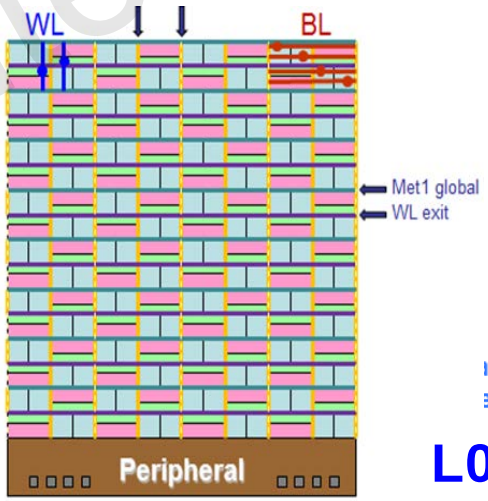
Micron/Intel Confidential

3D NAND with CMOS Under Array

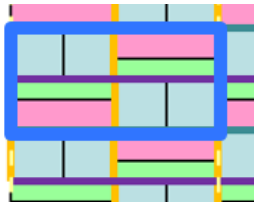


- 3D NAND is very compatible with CMOS/Periphery Under Array → Sense Amps, Array drivers placed under Array → Tile Architecture.



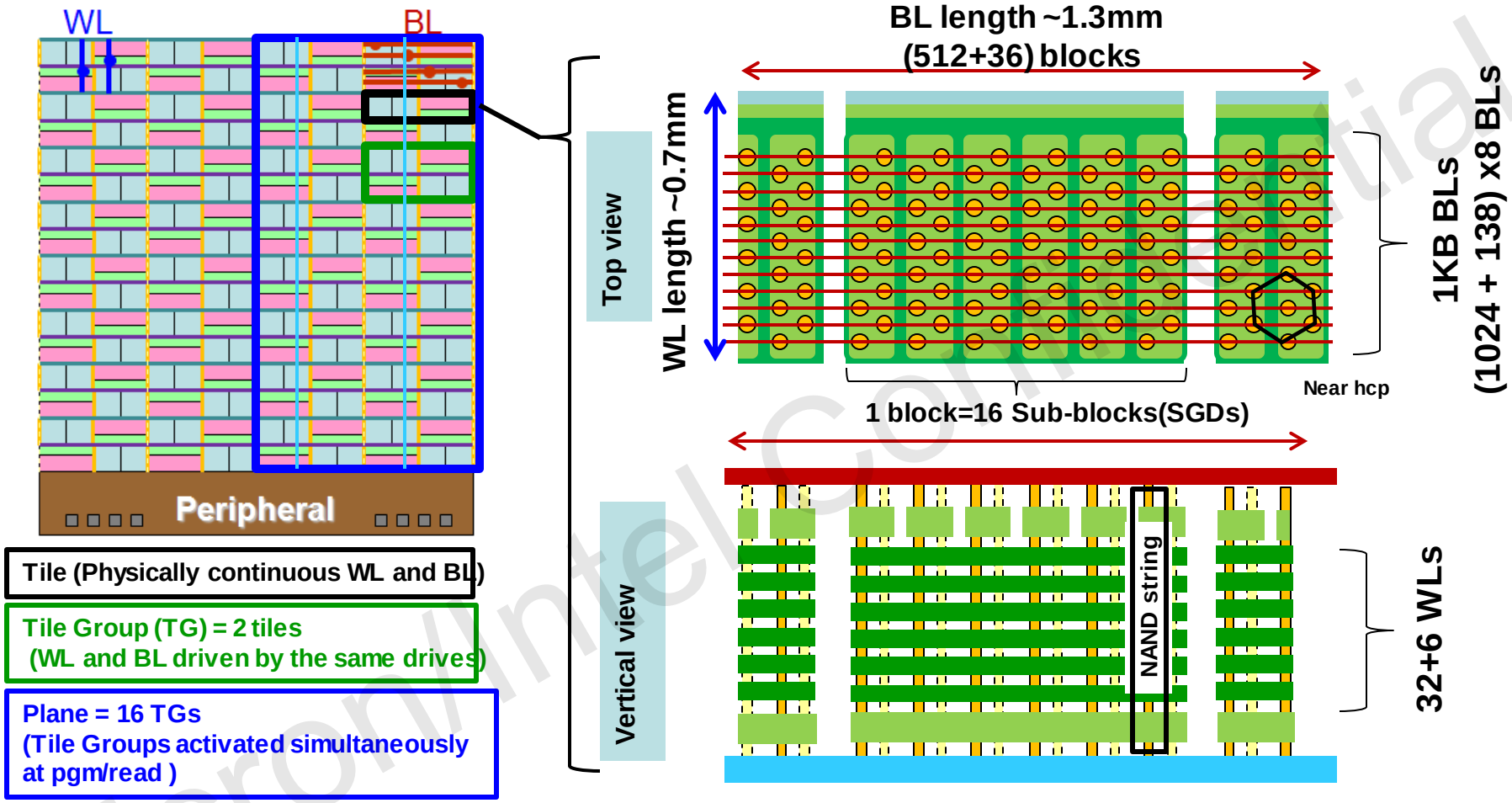


1 Tile Group (TG)
19008 BL x 548 block segments
Spans 4 SRC plates
ıre separated by Staircase/WL Exit
e are 32 tile groups per die (8 x 4)



Micron Intel Confidential

3D NAND Architecture



Page size: Number of cells program or read together
= number of BLs in a plane
= 32KB (1KB/tile x 2 tiles/TG x 16TGs/plane)

Pages per block: 1024 pages (=2pages/WL x 32WLs/sub-block x 16 sub-blocks/block)

Die density: 256Gb (=32KB/page x 1024pages/block x 512blocks/plane x 2planes/die)

Micron Intel Confidential

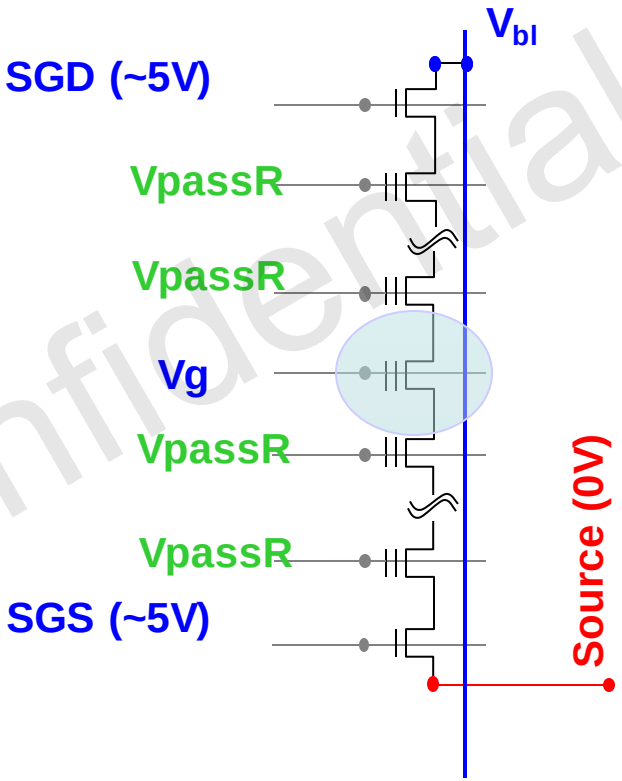
Read

- Cell/String I-V
- String Current Requirement
- 100s String Current limiters
- No-LDD issues
- Sensing – SBL & ABL
- Interference (FG-FG)
- 1-Pass and 2-Pass Programming
- Shifted Window

Micron/Intel Confidential

NAND Cell I-V

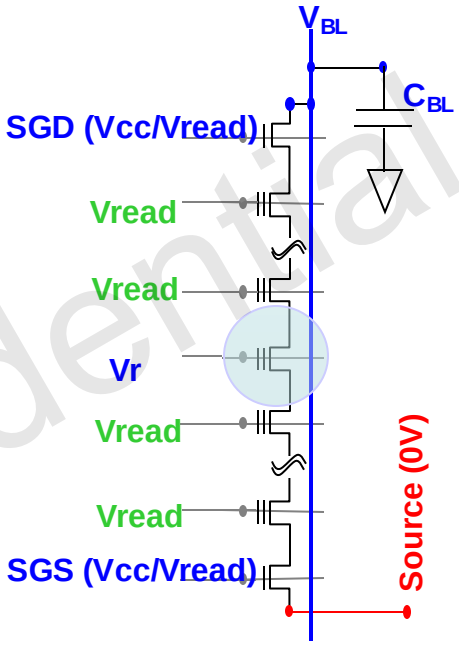
- A NAND String can comprise 32-128 NAND active cells plus two select transistors
- Like any transistor, there is gate bias (Vg) and Well/substrate bias (Vb) which can be directly applied for the selected cell
- However, there is no direct access to the source and drain of the selected cell. External bias can only be applied to the Bitline (Vbl)and the Sourceline (Vs)
- The select gates (SGS & SGD) need to be biased to a high Vg (~5V) to turn these on



- The deselected cells need to be turned on as well by making sure that these gates are biased well above the Vt of these cells
  - The deselected cells could have any Vt (Back-Pattern: BP) depending on whether they are in the Programmed or Erased state. The Cell Vts and the gate biases will dictate their channel resistance. This in turn will decide the actual source and drain bias the selected cell sees.

Micron/Intel Confidential

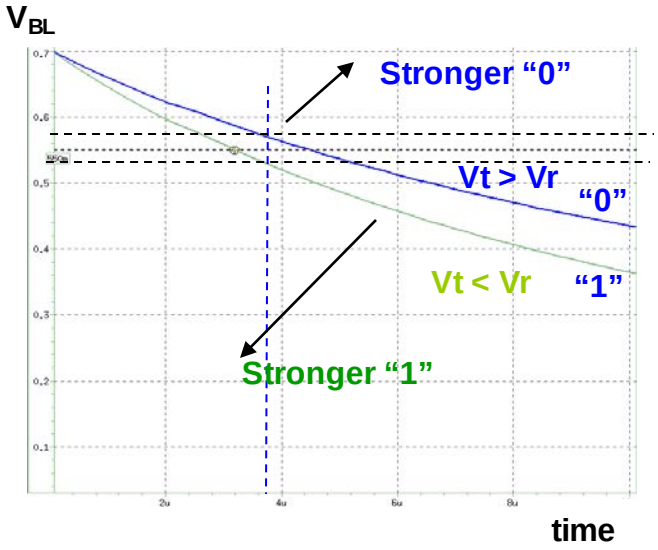
Basic Sensing Concept



Simple description of voltage-sensing concept (pre-100s)

- Selected bitlines are charged to a Vbl_init and then floated
- String to be read is selected:
  - Wordline to be read is biased to the Read Gate bias (Vr) and all the other wordlines in the string are biased to Vpass_r to make sure they are able to conduct
  - SGD and SGS are turned on.
- The cell current will start discharging the bitline
- After a certain amount of time (t_{SAD} ~ 5-10us) the Bitline voltage is sensed.
  - If the Vbl > Vbl_fin (not discharged) ➔ "0"
  - If the Vbl < Vbl_fin (discharged) ➔ "1"
- Let ΔVbl = Vbl_init - Vbl_fin
- Then, if Icell > $\Delta Vbl*C_{BL}/t_{SAD}$ ➔ "1"
  Icell < $\Delta Vbl*C_{BL}/t_{SAD}$ ➔ "0"
- For selected cell to be sensed, current should be dictated by selected cell and not rest of string.

➢ The string current >> sense current.



Micron Confidential

String Current Required

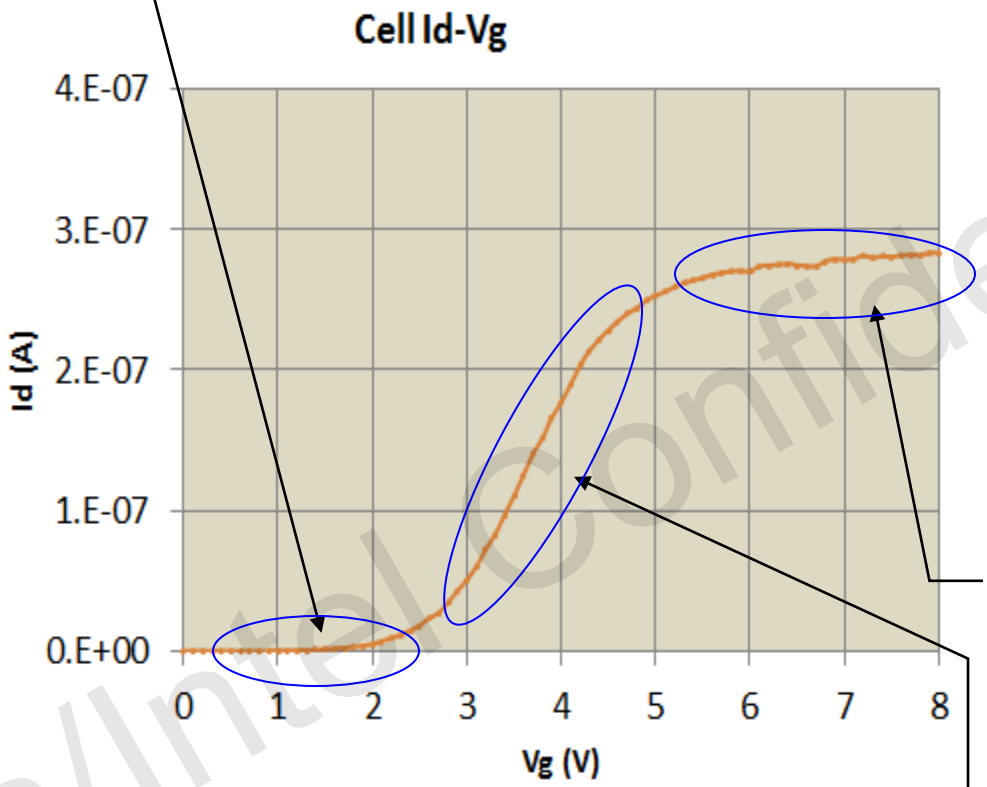
#	Item	Value	Consideration
1	Sense current	~10-15nA	Margin to leakage, Noise, Sense times and BL Stabilization (Performance).
2	Dual Strobe Margin	~1.5X-2.0X	PPV sense current is 1.5-2X higher
3	Sense current to String current Guardband	~3X	Current set by Cell Vt and not string, since we want to read cell. RWB.
4	Temp Guardband (90C to -10C)	~1.5X	Poly channel current degrades at low temperature (in single crystalline silicon current increases at low temp)
5	Cycling Guardband	~1.5X	Trap-up related current degradation
6	Erase string to Programmed String	~1.7X	Reduced Gate overdrive on Programmed string vs Erased string
7	90C Probe Min Erase string current Required	~120nA - 200nA	

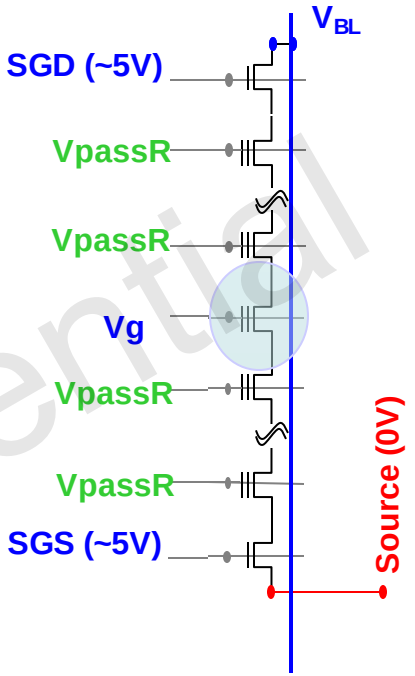
- Intrinsic Cell improvements (Temp/Cycling GB) and Process improvements (DHC interfaces, distributions, leakages) can improve many of these factors to manage the overall string current requirement.

Micron/Intel Confidential

NAND Cell I-V

Current limited by the cell. Full Bitline to Source voltage drop occurs across the selected cell. Assuming the Bitline voltage is Vbl and the source is at 0V, the selected cell will see Vds ~ Vbl.



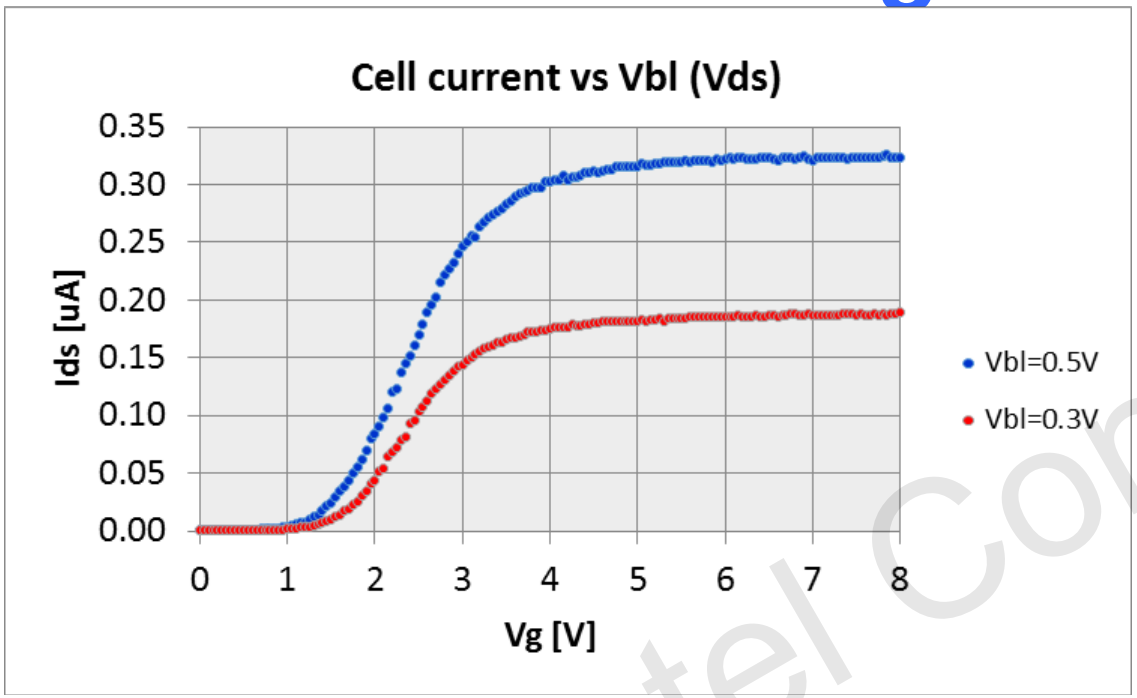


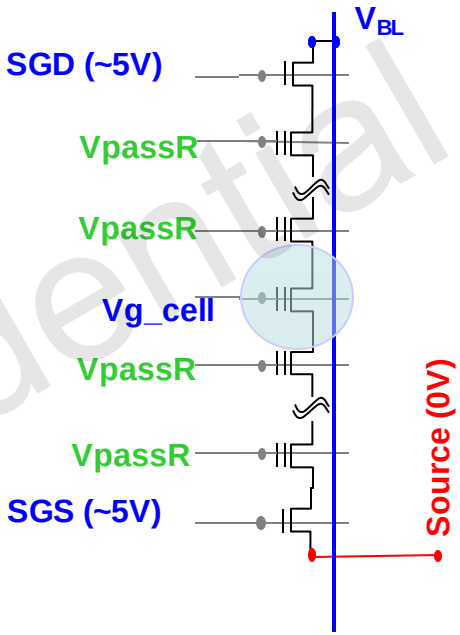
Saturated string current. Current limited by the total string resistance ~ LDD resistance as well as the net channel resistance of all the cells in the string and select gates. The selected Cell only sees a fraction of Vbl.

As the cell starts to conduct, there will be IR drop in the rest of the string causing the Vds across the selected cell to decrease. If the cell is on a lower wordline IR drop is mostly on the drain side causing the Vd to decrease. If the cell is on a higher wordline, the IR drop is mostly on the source side causing the Vs to increase. Since the Vs, Vds, Vgs for the cell are all varying the slope of the I-V does not really represent the cell Gm.

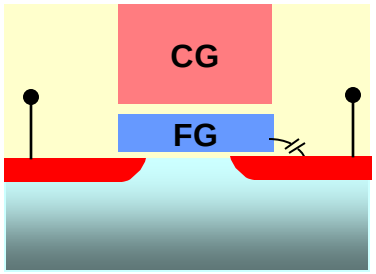
Micron/Intel Confidential

Cell I-V: BL Voltage





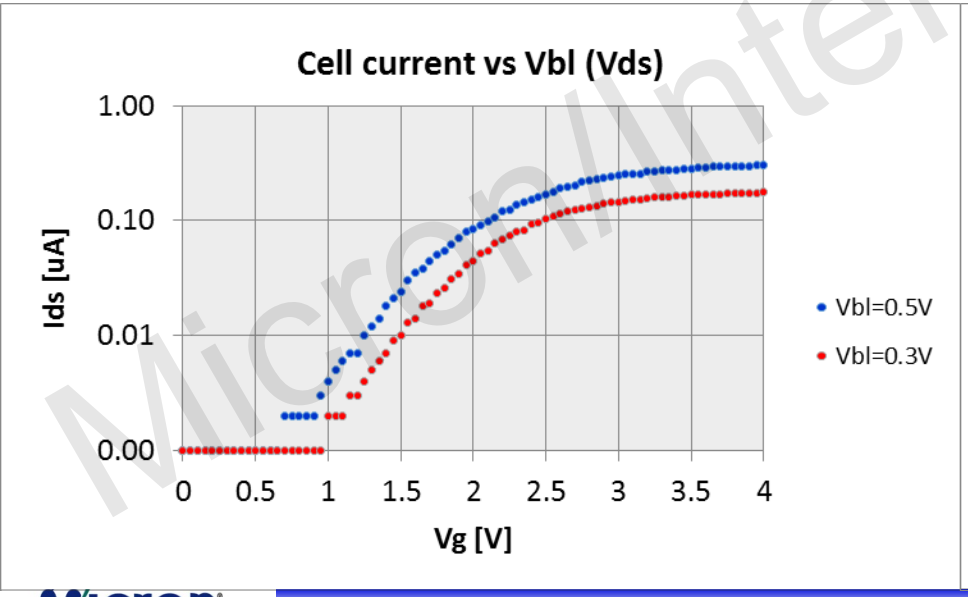
- Bitline voltage impacts the string saturation current as the string current is Bitline voltage/Net String resistance (Channel + S/D resistance of all the cells and SGs on the string).
- It also causes the Cell Vt to shift due to DIBL/DCR (Drain-induced Barrier Lowering / Drain Coupling Ratio).

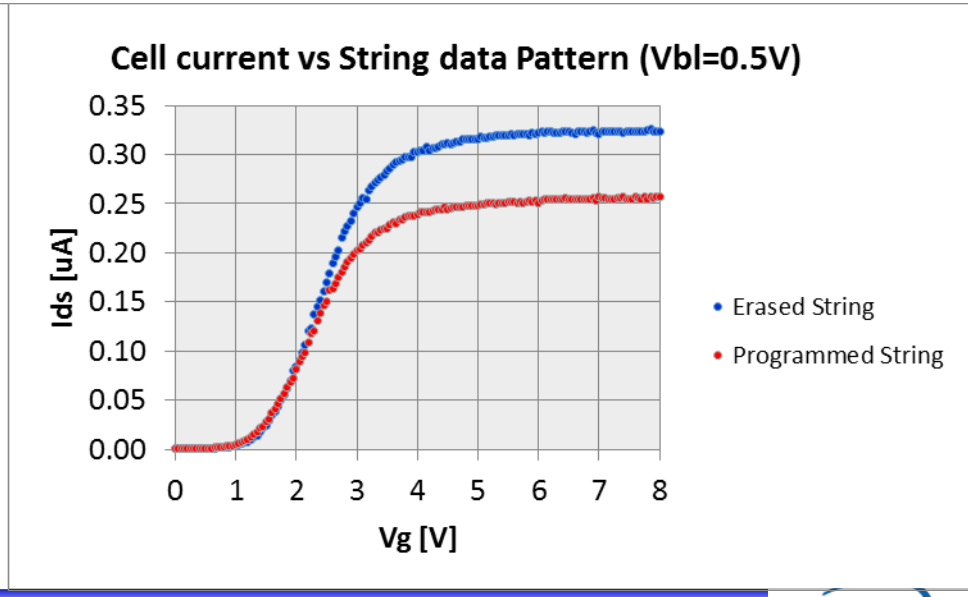


Micron/Intel Confidential

Key Cell I-V Parameters

- Some of the key cell parameters of interest are:
  - Sub-threshold slope – can be measured from the I-V
  - DCR/DIBL: Change in the cell Vt due to change in the Vds
  - Body effect: Change in the cell Vt due to change in the Vsb. On 3D NAND with floating body, there will be no body-effect. But change in Vs will impact Vgs-Vt.
  - Trans-conductance (Gm) – difficult to measure directly since the cell Vd & Vs in a string dynamically changes causing the Vt, Vgs-Vt to change as well
- Knowing the above for a cell allows us to construct the cell/string I-V



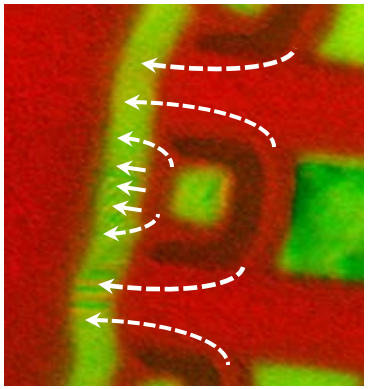


Micron/Intel Confidential

100s String Current

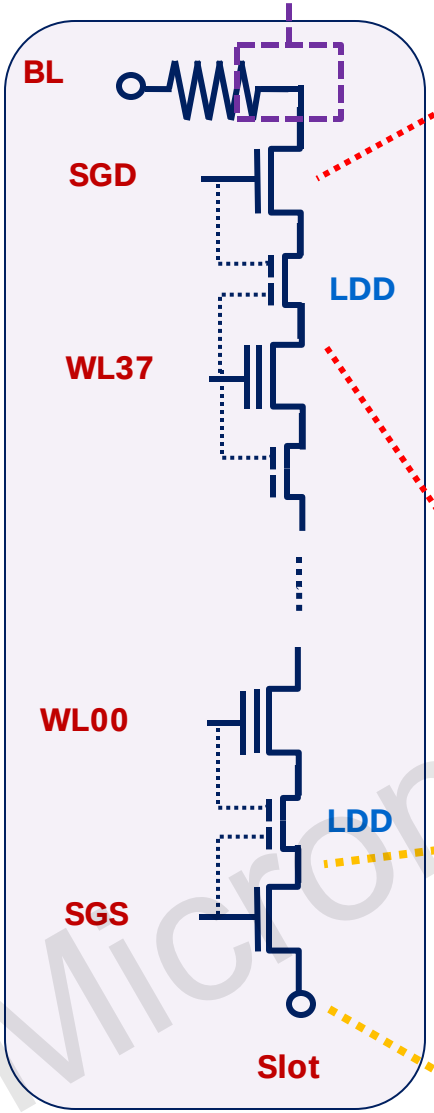
- String current health is one of the key parameters. There are three important factors to consider:

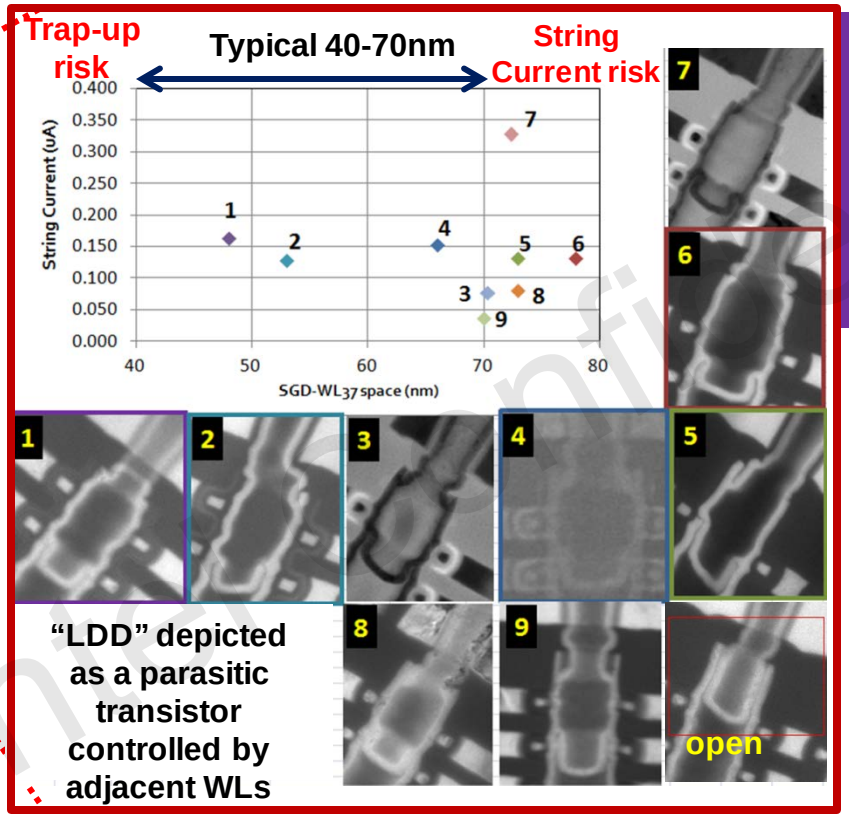
1. Channel material is made from poly-silicon. Inter-grain boundary and associated traps influence charge mobility significantly (~20-40x lower mobility compared to crystalline silicon). Moreover, the channel material is made of thin and hollow cylinder (implemented to improve cell Vt uniformity); charge transport is influenced by proximity of two oxide-poly boundaries
2. Cell transistors in the string are junction-less; the classic "LDD" region is not doped and relies on fringing field to operate (a less efficient coupling than the channel below cell)
3. Process integration such as 21L-24L, 24-41L, 21-45L are critical

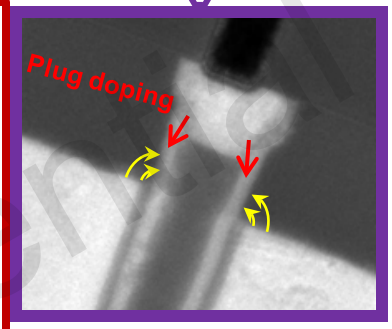


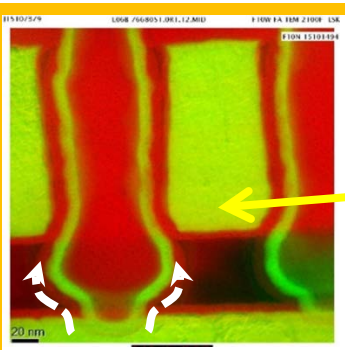
Micron/Intel Confidential

String Current: Process Integration

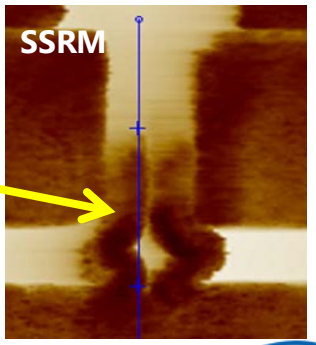






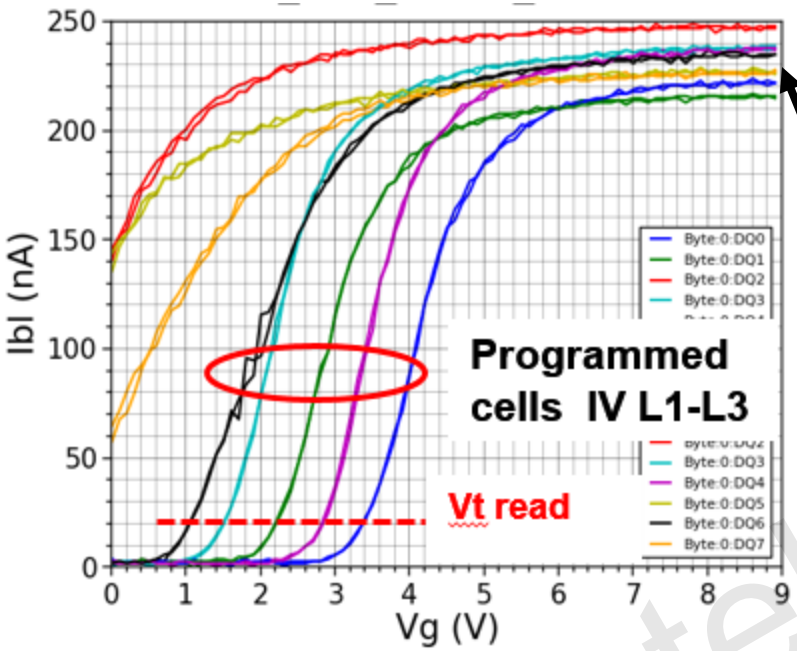


1. Dopant out diffusion from highly doped slot lowers LDD resistance. Too high doping lower SGS Vt (risk)
2. Main junction for erase GIDL generation
3. Electron injection point and key for string current performance



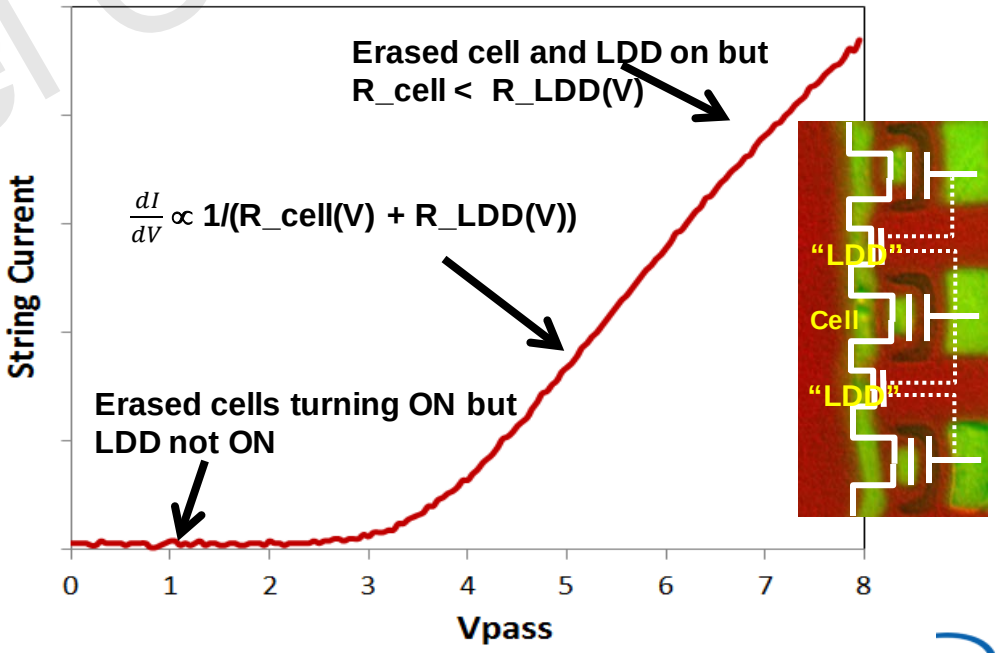
Micron/Intel confidential

Cell and String I-V



- Cell I-V (all WLs at Vpass_R ~8-9V, and sweep of selected WL) shows a steep turn-on of transistor (decided by Cell Vt) and saturation decided by string current.
- Limited by resistance "external" to the selected cell (Vts of other cells in the string, string to string variability, etc.)

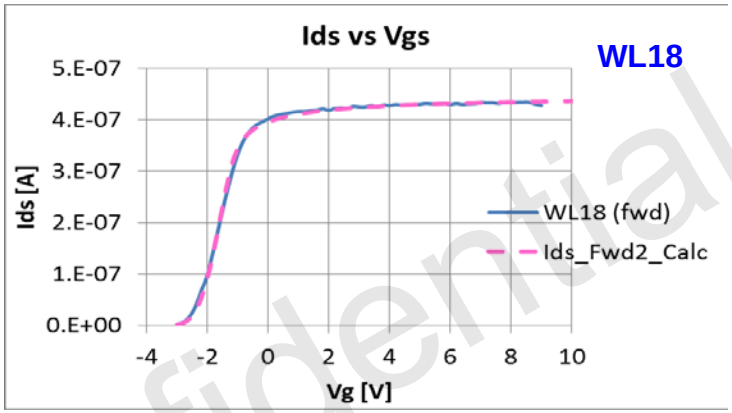
- String I-V (all WLs ganged and voltage swept) shows a gradual turn-on and continuously increasing current with WL voltage

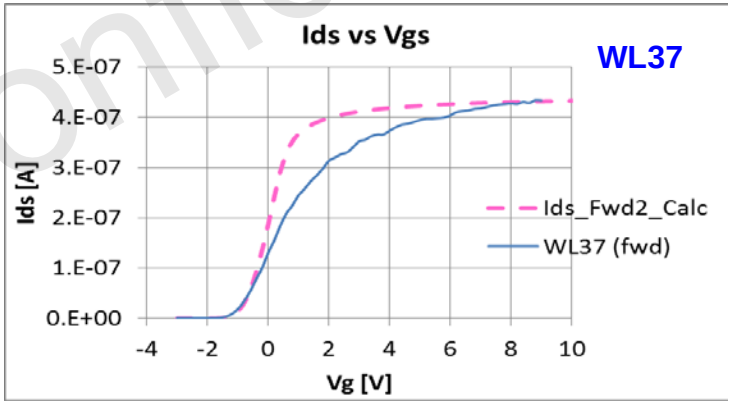


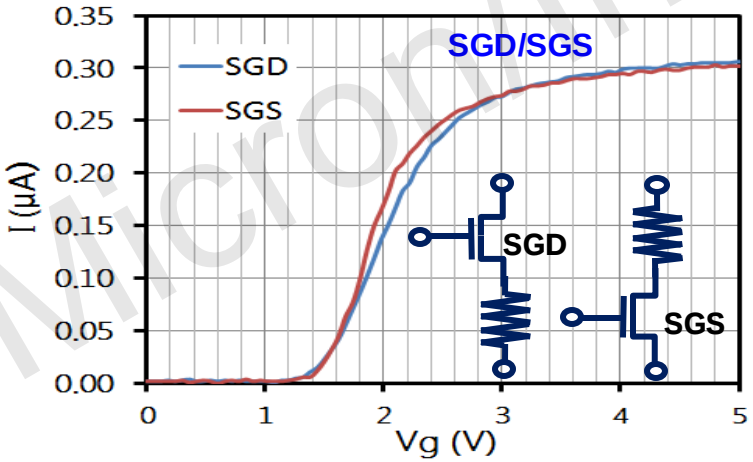
Micron/Intel Confidential

Cell and SG I-V

- Cell I-V can be modeled very well using simple transistor I-V equations (example WW18 Id-Vg shown here).
- Channel mobility is estimated to be ~5-10cm^2/V-sec (compared to 100-200cm^2/V-sec) for single crystalline Si (e.g. 80s).
- But the larger W/L ratio of 100s NAND string compensates for the reduction in mobility to provide adequate string current.
- WL37 example shows significant deviation of experimental I-V from expected I-V and → Softer saturation of current (significant voltage response) → indication that resistance between SGD-WL37 is still significant.





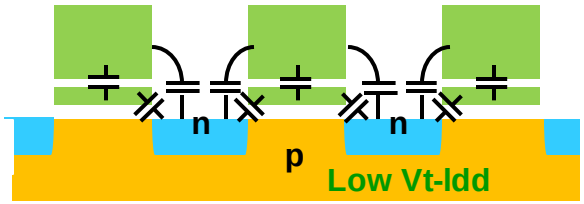


- Apparent "gm" of SGD transistor is lower than SGS as the main loading resistance is sitting on the source side of SGD transistor
- Loading resistance analysis can be helpful to segment location of string current limiters

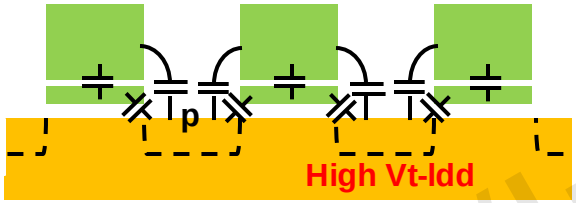
Micron Intel Confidential

No LDD - DHC

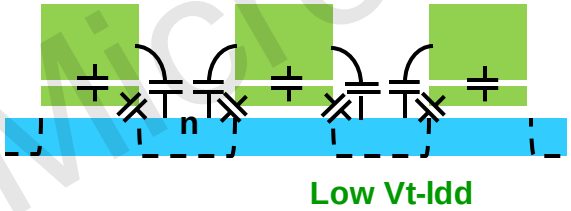
1. With n- LDD (2D)



2. No LDD, P-type channel



3. No LDD, N-type thin channel



- LDD region has Vt which shifts by Qfg and Vwl_adj.
- Vt_cell = max. (Vt_ch., Vt_ldd).
- We want Vt_cell = Vt_ch. Instead of Vt_ldd. Vt_ldd is more influenced by the neighboring cell and WL.
- Therefore, lower Vt_ldd is a requirement, leading to n-type doped thin channel.
  - N-type is to reduce Vt_ldd.
  - Thin channel is to eliminate bulk leakage path to realizing n-type channel.
  - Thin channel also makes the Vt less sensitive to the trap density in the channel (less volume → less charge)
  - Due to the poly Si trap (acceptor type trap), a net channel doping is ~un-doped with 1~3E18 n-type poly.

Micron/Intel Confidential

Vpass_R Required

$$\text{Vpass\_R Required} = \text{Average L7 Vt} + V_{BL} + \text{DCR/DIBL} * (2/3 * V_{BL}) + \gamma * \{\text{sqrt}(1 + V_{BL}) - \text{sqrt}(1 + 1/3 * V_{BL})\} + (\text{Vgs-Vt})$$

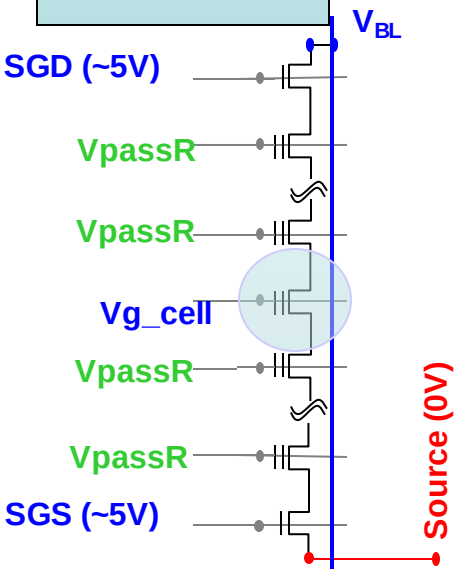
The Vt of the deselected cells. Including all sources, e.g. interference as well.

Since the sources of high WL#s is at ~Vbl, Vg needs to be higher by this much

Going from PV7 verify to Pass gate condition, the Vds across the high WL cells goes from 2/3*VBL to ~0. So, the Vt increases

Going from PV7 verify to Pass gate condition, the Vs for the high WL cells goes from ~0-1/3rd VBL to VBL. So, the Vt increases by the body effect

Gate overdrive to have acceptable string current

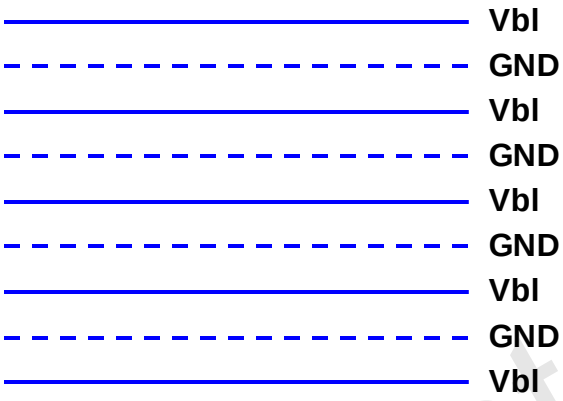


Higher Vpass_R will help increase the string current, but will result in higher Read Disturb. So, there is Read Disturb to RWB trade-off to consider. (We will discuss Read Disturb later)

Micron/Intel Confidential

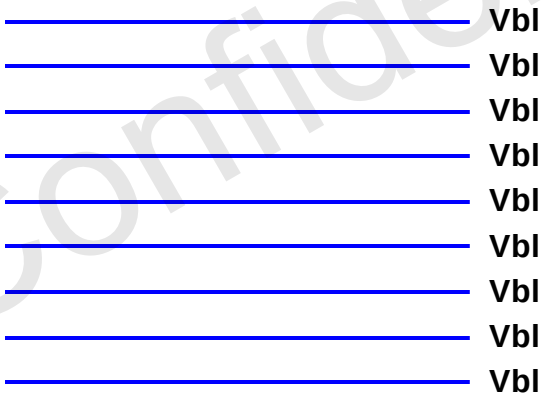
Shielded BL (SBL) and All BL (ABL)

SBL



- Every other BL is selected for pgm/read.
- Page size is a half of # physical BLs

ABL

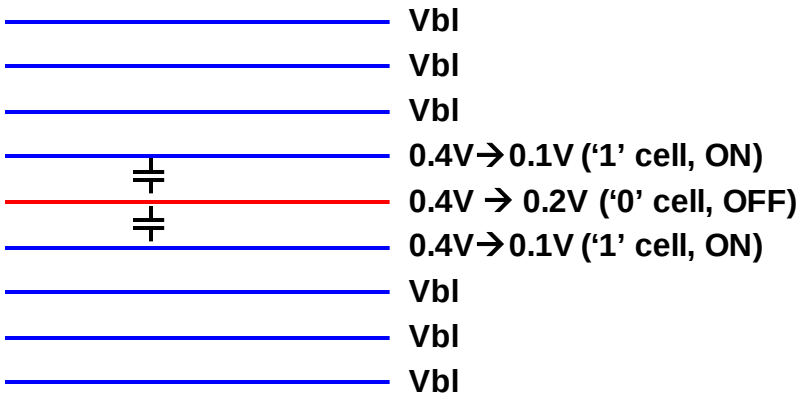


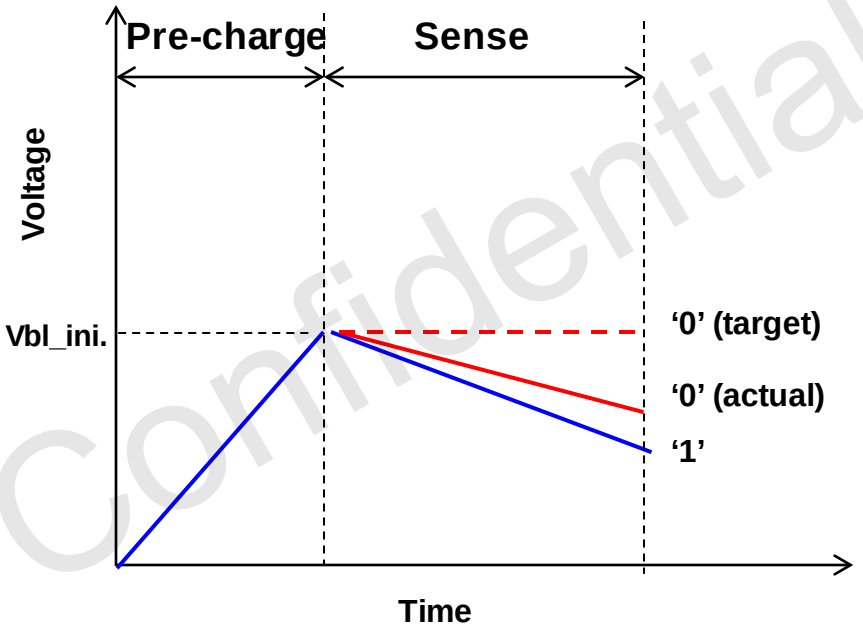
- All BLs are selected for pgm/read.
- Page size is the same as # physical BLs

ABL architecture is introduced to enhance pgm/read throughput by doubling a page size.

Micron/Intel Confidential

BL-BL Coupling Noise in ABL





- '1' cells will draw current and discharge the BLs connected to them
- BLs connected to '0' cells are expected to stay at the pre-charge voltage
- Due to the BL-BL capacitance (~0.4pF/side), '0' BLs adjacent to '1' BLs are pulled down as well and discharges the sensing node.
  - '0' Cells are read as '1' → Sensing error.
  - Voltage sensing can't be used in ABL.
  - BL voltage has to be held constant during sensing ➔ Active sensing

Micron/Intel Confidential

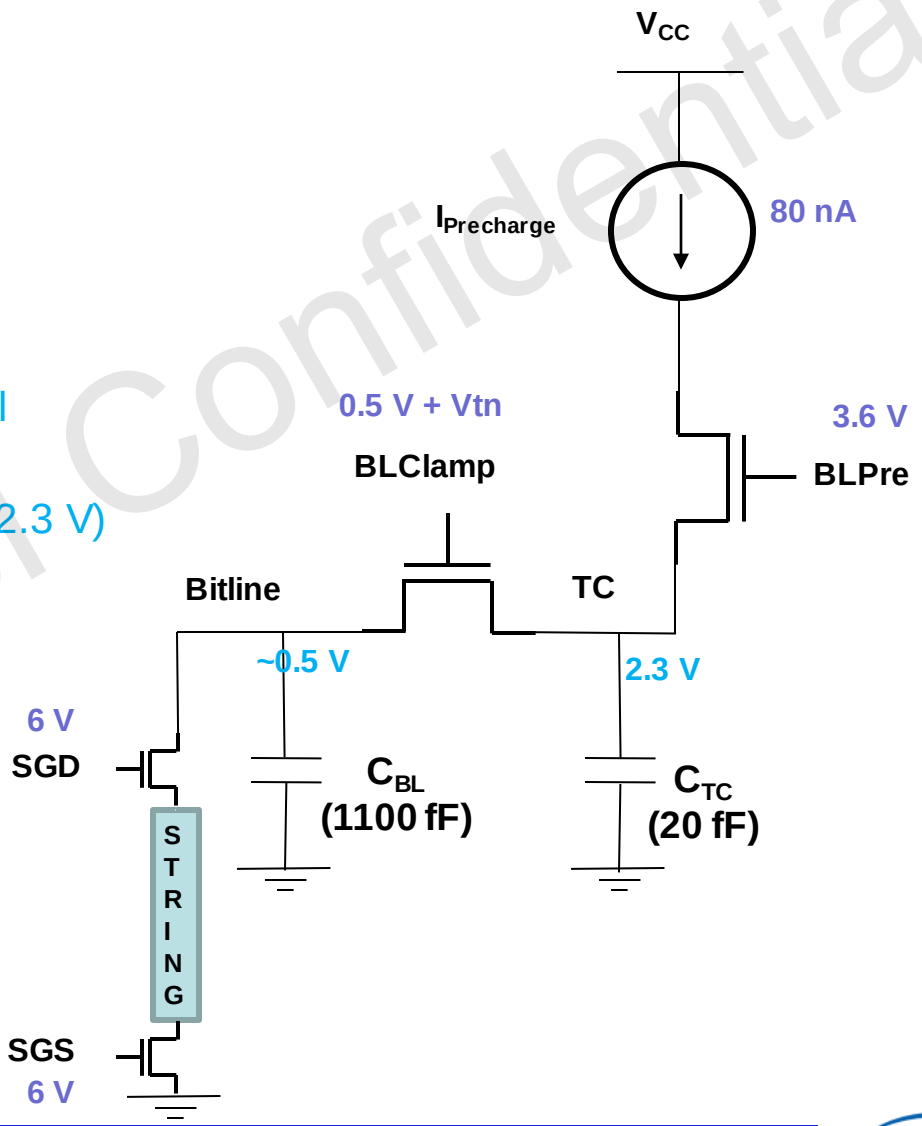
Active-Sensing – Step 1

Step 1: BL pre-charge/stabilization

- Supply precharge current (80 nA)
- BLPre strongly ON (3.6 V)
- BLClamp targeting Vbl (0.5 V)
- SGS/SGD ON (~6 V)

- For strings drawing < 80 nA,
  - BL charges to approx 0.5 V. Exact Vbl depends on $I_{BLClamp} = I_{string}$ condition.
  - $I_{precharge} - I_{BLClamp}$ charges TC to Vcc (2.3 V)

- For strings drawing > 80 nA,
  - Vbl will settle at < 0.5 V with exact Vbl depending on $I_{BLClamp} = I_{string} = I_{precharge}$
  - Since all the precharge current is being discharged by string, TC cannot charge up to V_{cc} and is ~ V_{bl}
  - TC is lower than sense-amp flip voltage (~1.5V), and these will be sensed as '1' regardless of rest of sensing sequence.



Micron/Intel Confidential

Active-Sensing – Step 2

Step 2: Turn OFF BLPre

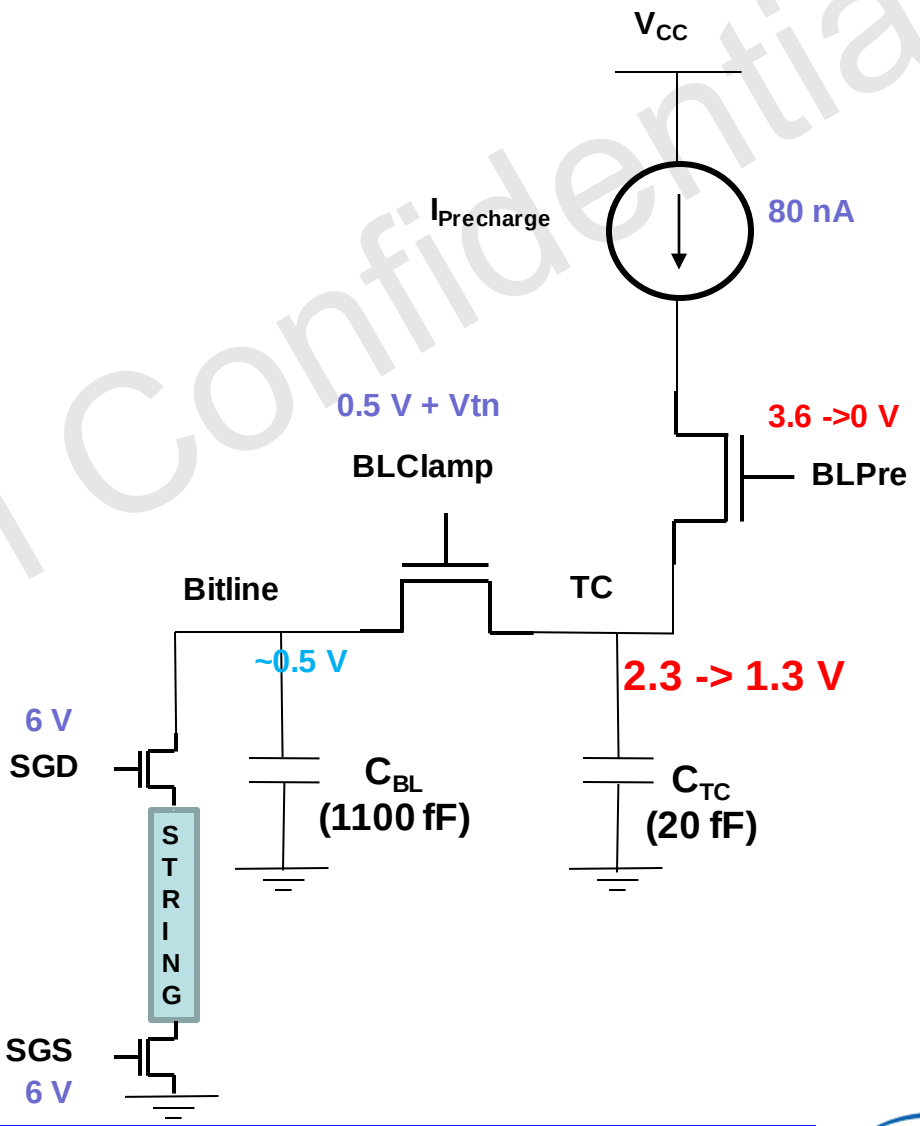
- BLPre OFF(0 V)
- Rest same.

- Since BLPre is OFF, there is no current feeding TC.
- TC will start getting discharged by ISTRING (= IBLClamp)
- If TC discharges below sense-amp flip voltage in a certain time, it is sensed as a '1' else it is sensed as a '0'.
- TC Develop time

  $= C_{TC} * (Vcc - VSA\_flip) / I\_sense$

  = 20 fF * 1 V / 20 nA ~1 us

- As long as TC > Vbl + ~0.2V, Vbl doesn't change through develop time.



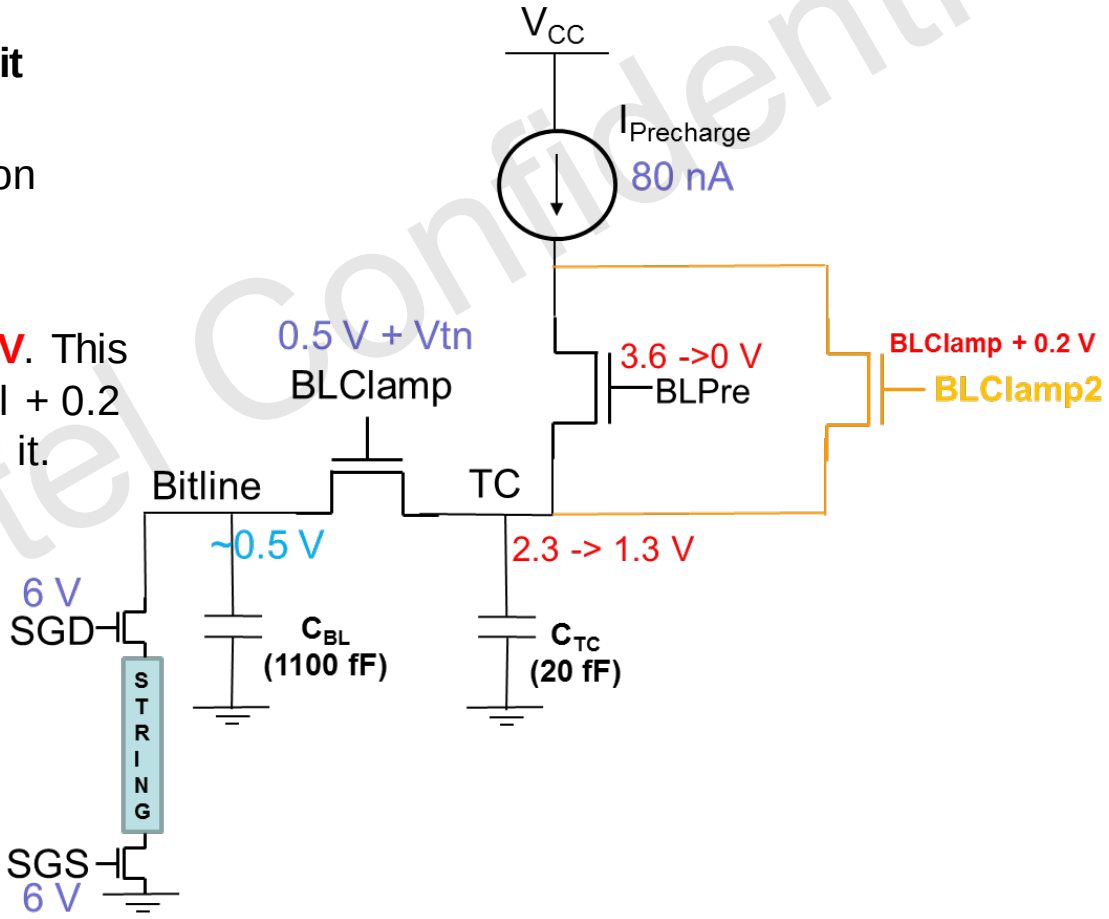
Micron/Intel Confidential

Active Sensing – Step 2

- **How to prevent Vbl movement if TC discharges too much?**

- **<u>Solution</u>: Feed current into TC if it discharges too low.**
  - Using BLClamp2 (new device on active-sensing)

- Trim BLClamp2 to **BLClamp + 0.2 V**. This way TC never discharges below Vbl + 0.2 V because $I_{precharge}$ will start feeding it.

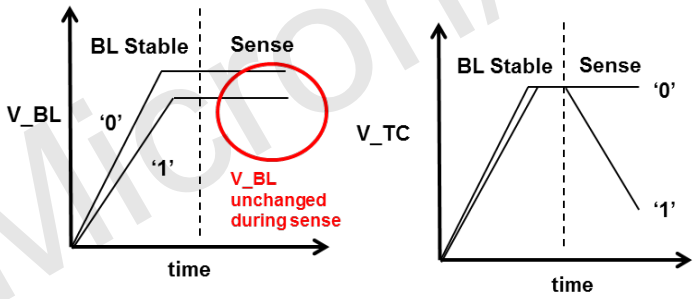


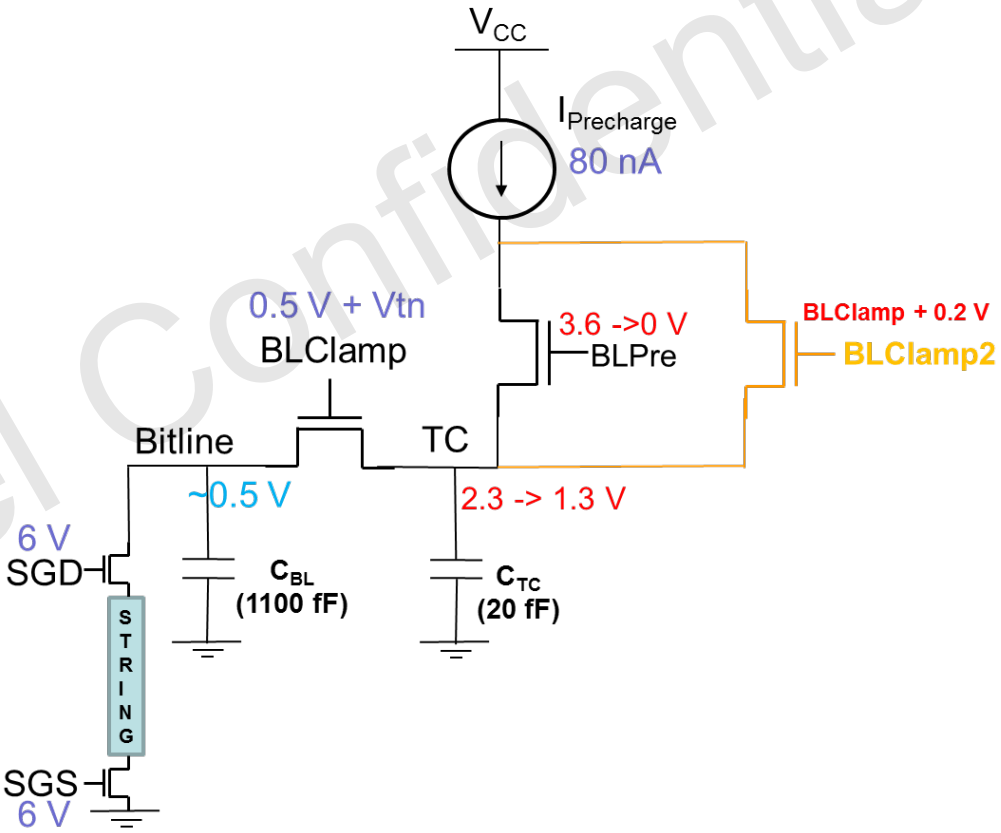
Micron/Intel Confidential

Active Sensing – Summary

To summarize:

- *Step1: Stabilize BL voltages (with SG ON) and precharge TC to Vcc.*
- *Step2: Turn off precharge current, and discharge TC. TC voltage is sensed after develop time*
  - *BL voltage doesn't change during TC develop time. This is accomplished by putting safeguard current into TC if it discharges too low, using BLClamp2.*

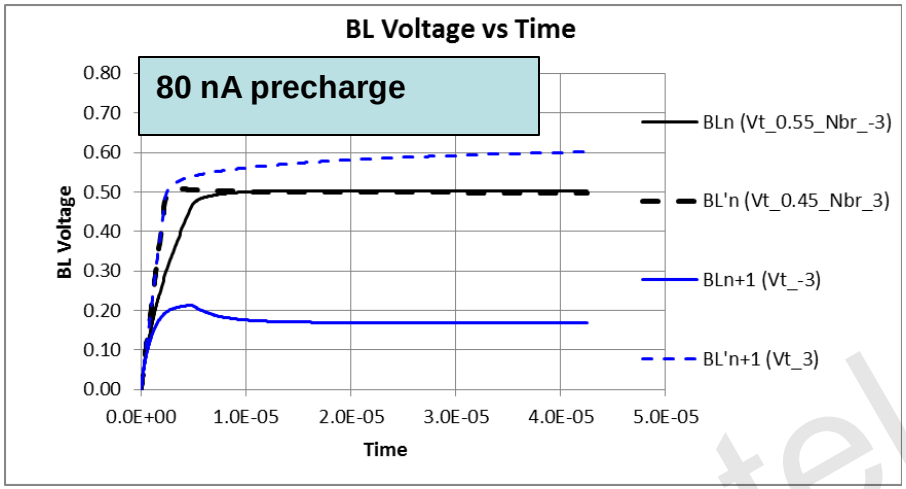


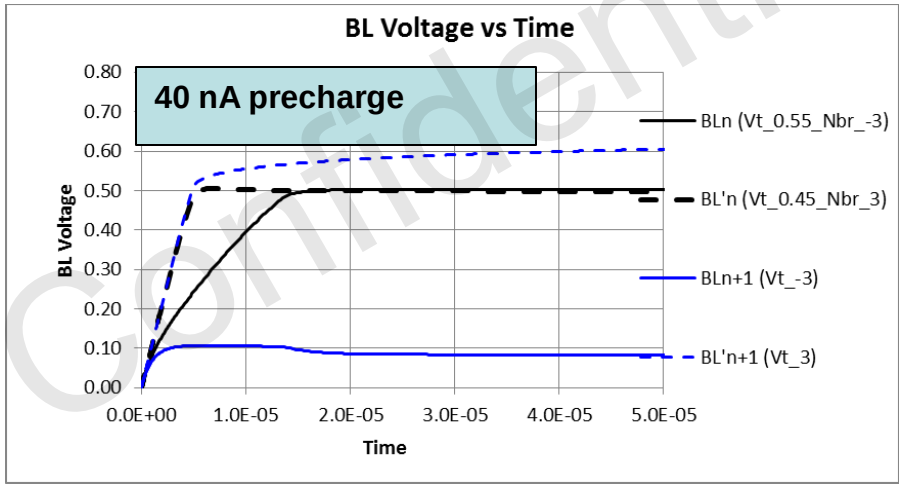


Micron/Intel Confidential

Need for knockout (KO)

- Step 1 (BL stabilization) time depends on Iprecharge
  - Larger the Iprecharge, faster the BL stabilization.





- Low Vt strings will be drawing full Iprecharge through entire sensing sequence.
  - Worse Icc / power
  - 'SRC bounce' leading to incorrect sensing.

- What is Source bounce?
  - Low-Vt cells draw large current constantly → IR drop in the SRC plate/line → Local SRC goes higher than global SRC (depending on location (R) and # low-Vt strings (I)).
  - If SRC is higher than intended, effective Vt looks higher.

Micron/Intel Confidential

Knockout (KO)

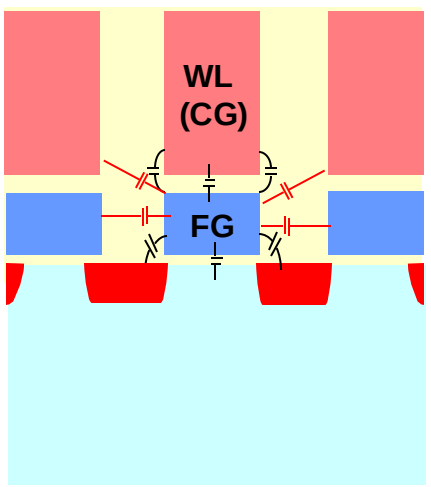
- **Knockout involves repeating the sense a second time, after knocking out the high current strings.**
  - Coarse sense: Do full regular sensing sequence. Inaccuracy due to source bounce will make some Vt look higher. Use this sense to identify really low-Vt strings. Develop time decides target current.
  - Knock the low-Vt strings out (i.e. make their precharge current zero), then wait for BL stabilization (with BLpre/BLClamp2 high) for neighbor BL/TC to recover.
  - Do a fine sense (take BLpre/BLClamp2 low and develop TC again)

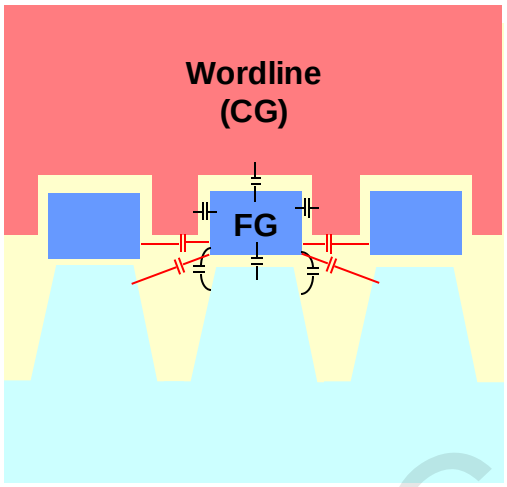
- Source bounce risk is higher for active-sensing because low-Vt strings are constantly drawing full precharge current whereas
  - Voltage sensing: No precharge current during develop.
  - Current sensing: Equal to sense current target ~ 20 nA, which is much smaller.

- Use of KO depends on tradeoff between
  - Using high Iprecharge to speed up BL stabilization, and taking penalty of KO/extra-sense time.
  - Using lower Iprecharge to avoid KO/extra-sense time but longer BL stabilization.

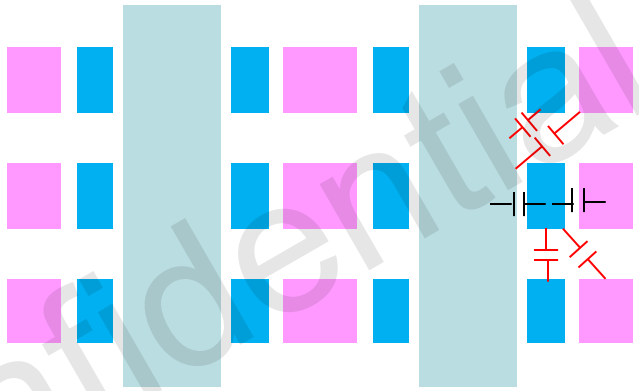
Micron/Intel Confidential

Cell to Cell Interference





2D



3D

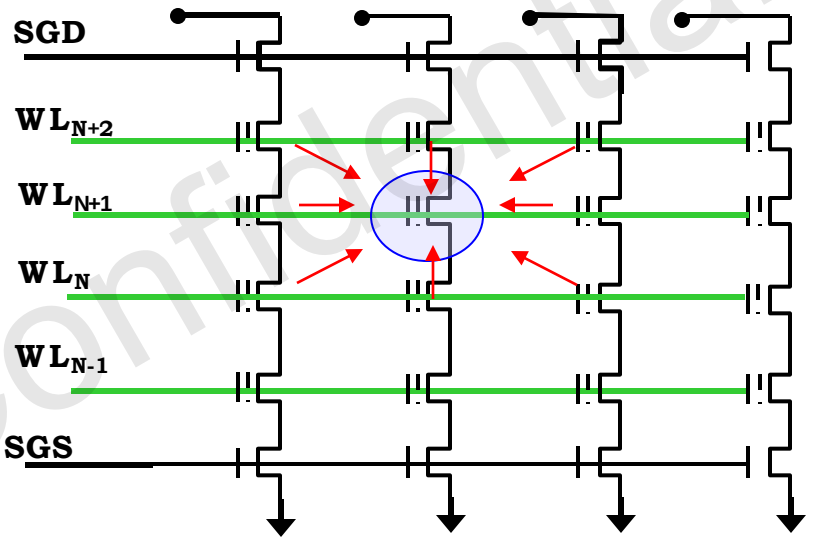
- The Cell Vt is primarily decided by the charge on its floating gate.
- But the charge and voltage on the neighboring nodes (neighboring floating gate, Wordlines, channel, etc.) also impact the cell Vt.
- Vt of the victim cell can change due to the influence of the aggressor node onto the victim cell's Channel/LDD or the floating gate.
- Two key terms are:
  1. "FG-FG": delta_Vt_victim / delta_Vt_aggressor

     Aggressor can be BL, WL, or Diagonal neighbor → WL-WL, BL-BL, Diag
  2. "WL-FG": delta_Vt_victim / delta_Vg_$WL_{n+/-1}$

Micron/Intel Confidential

FG-FG & $WL_{n\pm1}$–FG Coupling

- Impact of FG-FG interference:
  - During Programming all cells first start off in the Erased State.
  - Some cells are placed during a programming operation. Other cells (neighbors) are placed during a subsequent programming operation.
  - Effective Vt of the Cells Programmed first, moves higher as a result of the programming of the neighboring cells due to coupling from those cells.

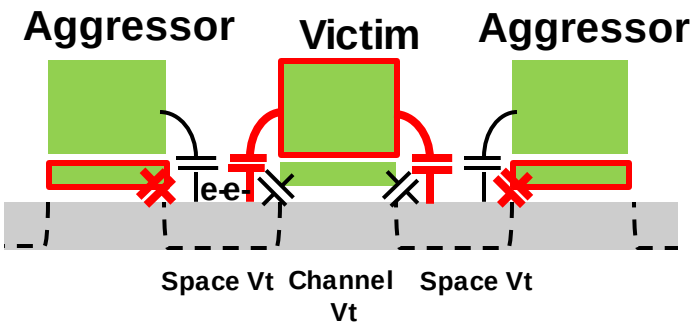


	80s	100s
FG-FG (WL-WL)	~4.5%	~5%
FG-FG (BL-BL)	~10%	~0%
FG-FG (Diag)	~1%	~0%
WL-FG	~10%	~13%

- Impact of WL-FG Coupling:
  - Coupling from the neighboring wordline make the Vt depend on the neighbor WL voltage.

Micron/Intel Confidential

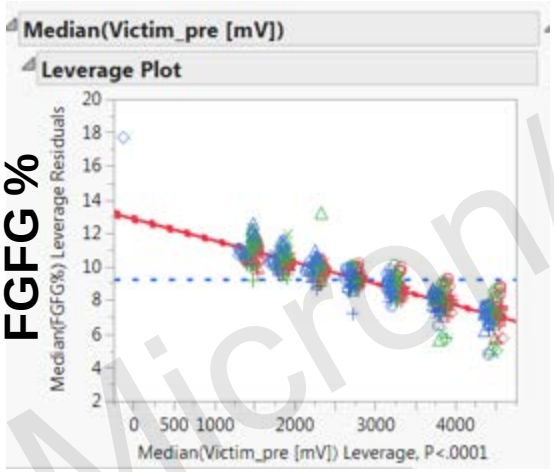
FGFG dependency on Vt level and cycling



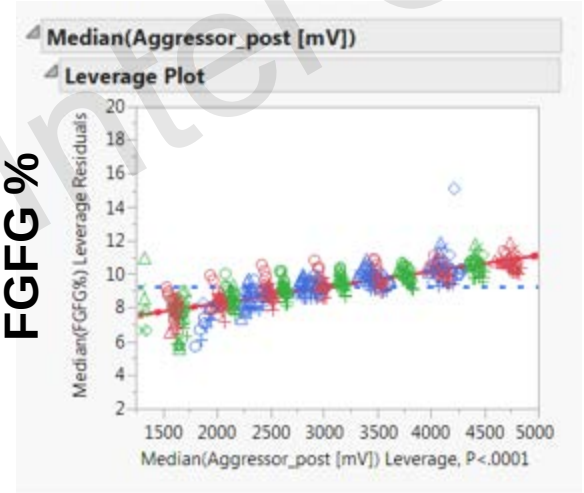
"FG-FG interference" is determined not only by FG-FG capacitance.

WL-Ch and FG-Ch play a role. too.
Space Vt has strong impact on 'FG-FG' Vt shift.

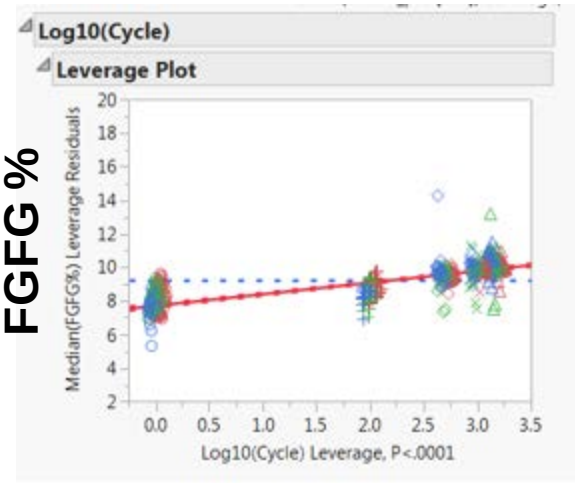
- Low Victim Vt (low WL voltage),
- High Aggressor Vt (more neg. FG potential),
- Cycling (trapping over space)
- Low temperature
degrade 'FGFG'
by making space Vt higher.



Victim Vt



Aggressor Vt



P/E cycling

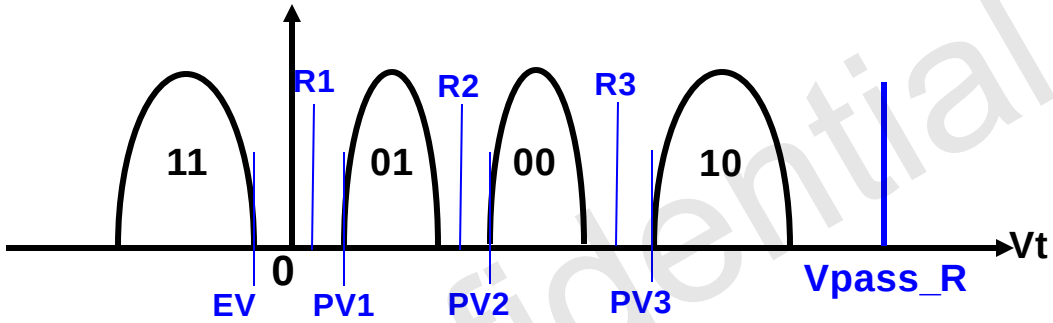
Micron/Intel Confidential

MLC

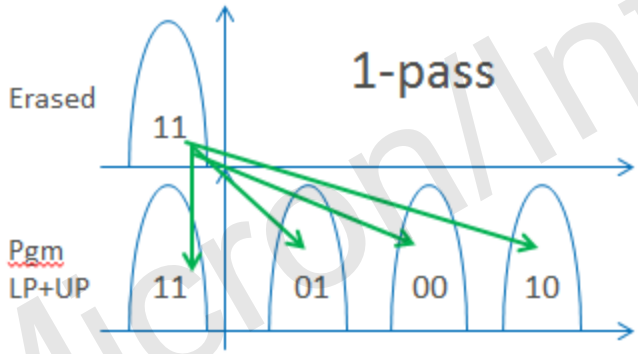
MLC: 2 bits/cell: 4 states:

Level 0 = Erased
Level 1 = Pgm to Vt1
Level 2 = Pgm to Vt2
Level 3 = Pgm to Vt3

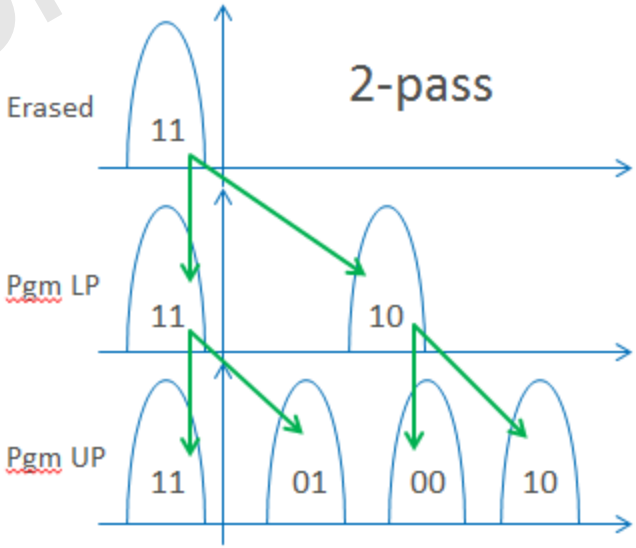
LSB (LP) Read with Vg = R2
MSB (UP) Read with Vg = R1 & R3



1 Pass vs. 2 Pass Programming



More susceptible to Cell-Cell interference.
More efficient for Program time

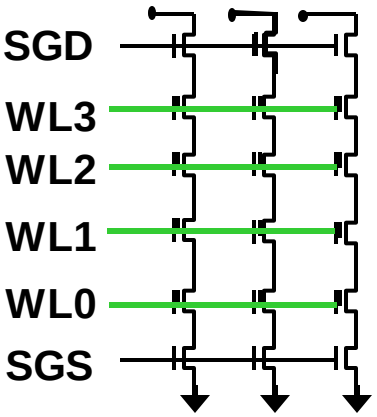


More immune to Cell-Cell interference.
Slightly inefficient for Program time

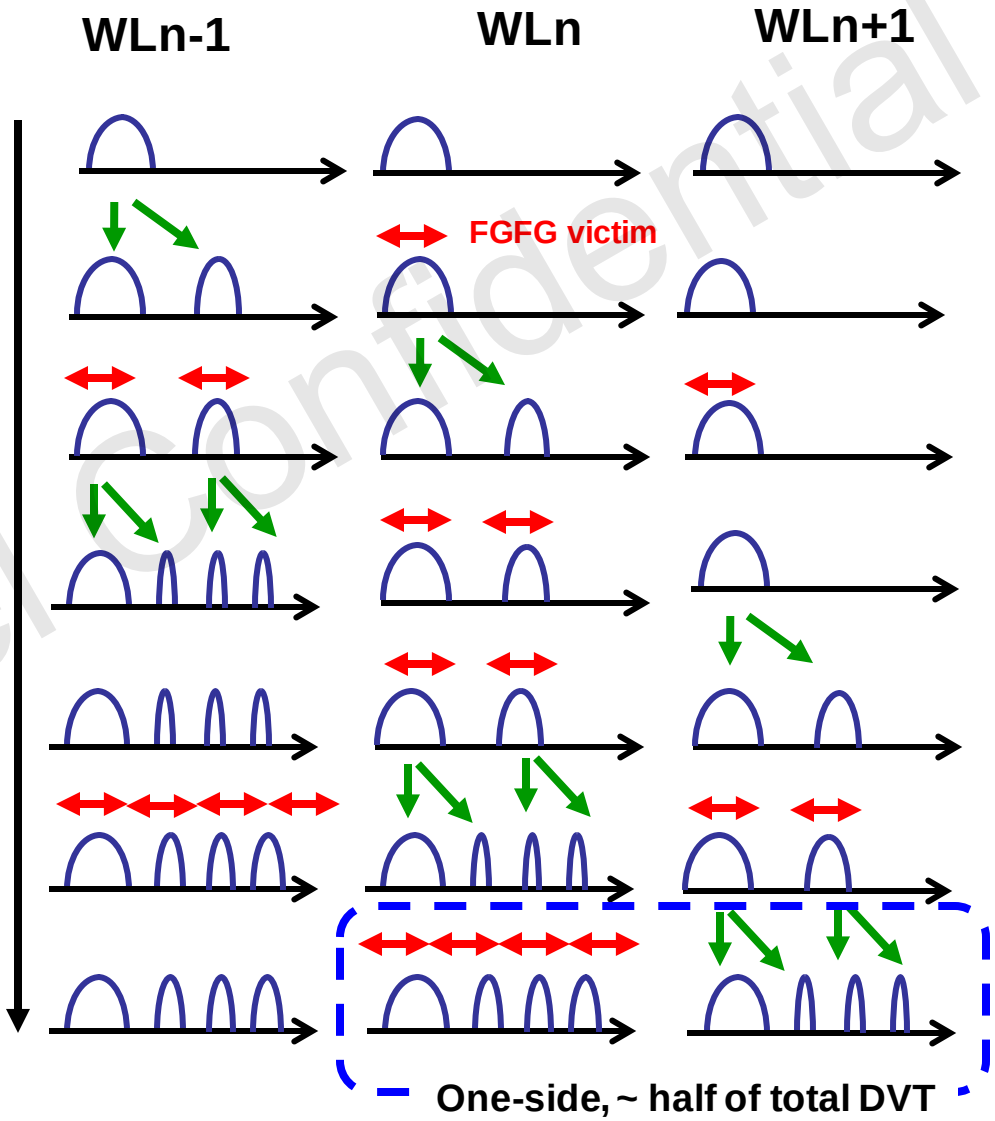
Micron/Intel Confidential

Back and Forth programming

Programming Order



	1st Pass (LP)	2nd Pass (UP)
WL3	5	7
WL2	3	6
WL1	1	4
WL0	0	2



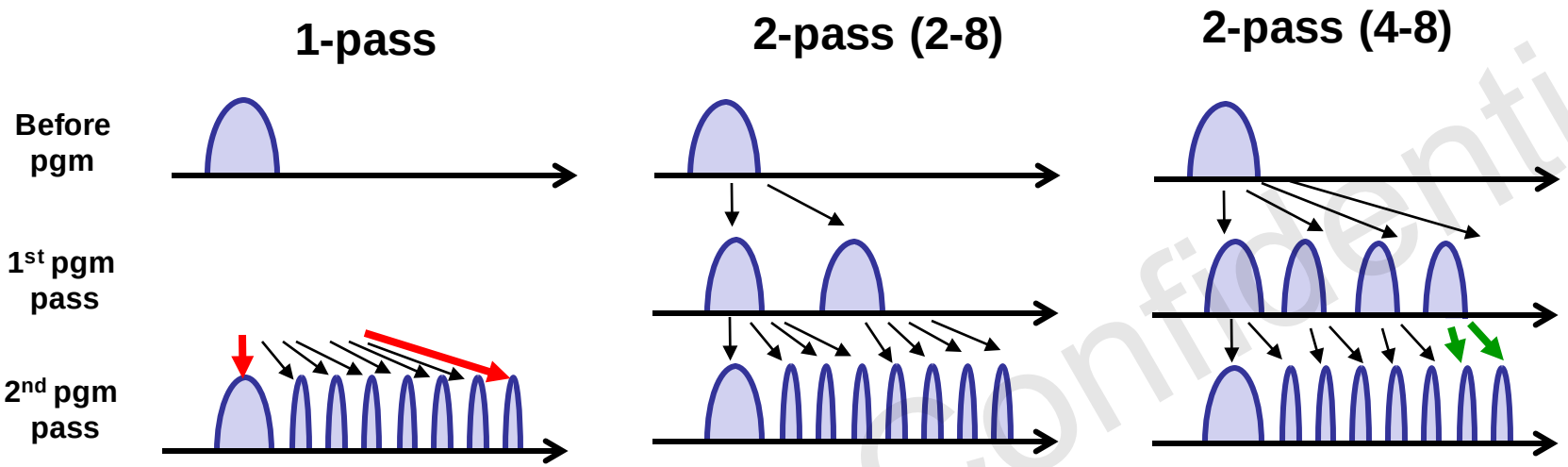
Programming is performed From the source side to the drain side.

With multi pass programming, Back and forth programming order.

This minimizes FGFG impact on the Final Vt states. (1-side of WL, ~half of total DVT.)

Micron/Intel Confidential

TLC programming sequence



# of pulses/verifies for the first pass

Zero ← → **More**

# of pulses/verifies for the second pass

More due to more FGFG aggression ← **Need smaller Vstep, more pulses**

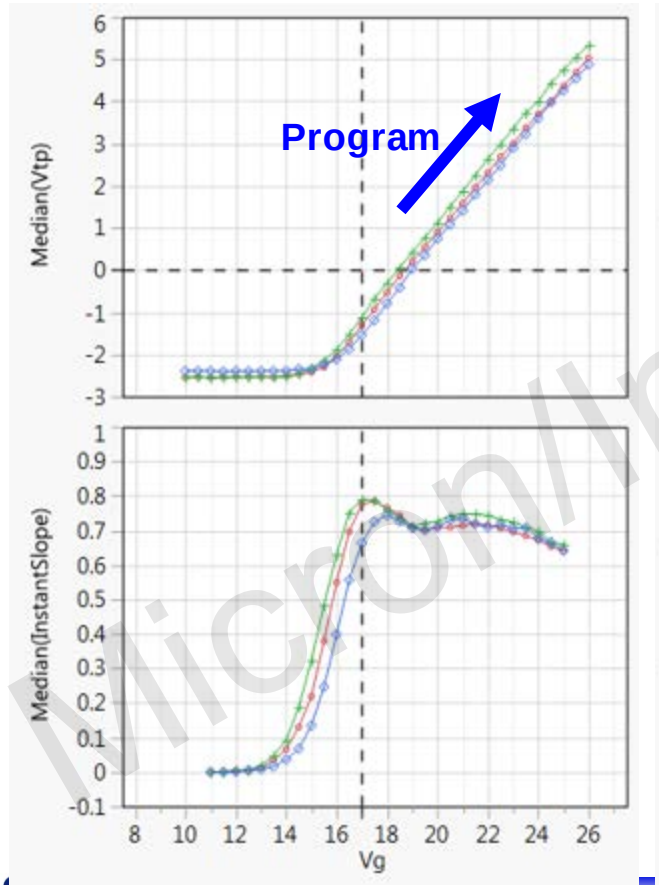
Need smaller Vstep, more pulses → **Less due to less FGFG aggression**

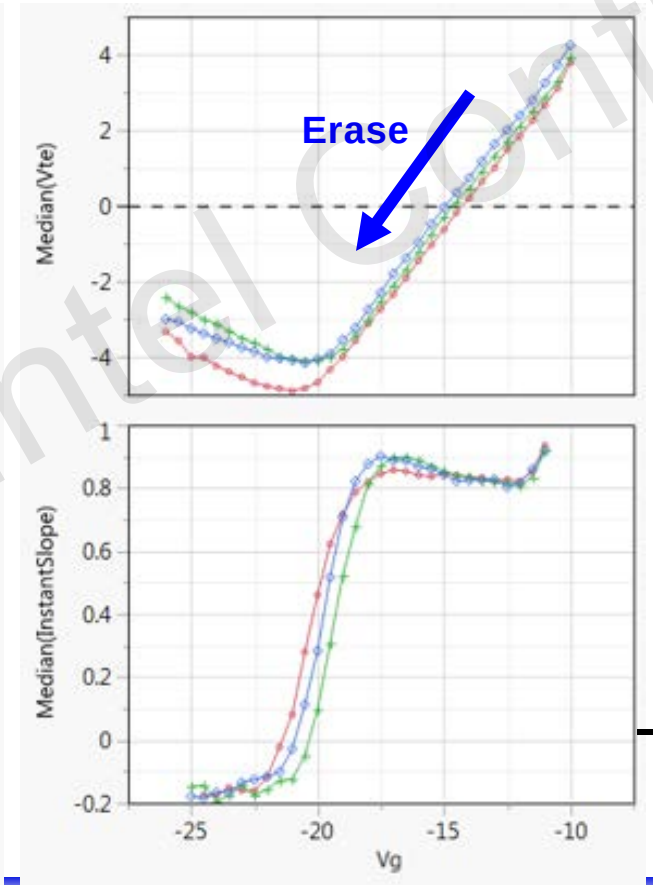
There are competing effect on tProg among programming algos.
The best algo. for tProg is FGFG dependent.
(FGFG<2%) 1-pass, (FGFG<4%) 2-8, (FGFG ~6%) 4-8.

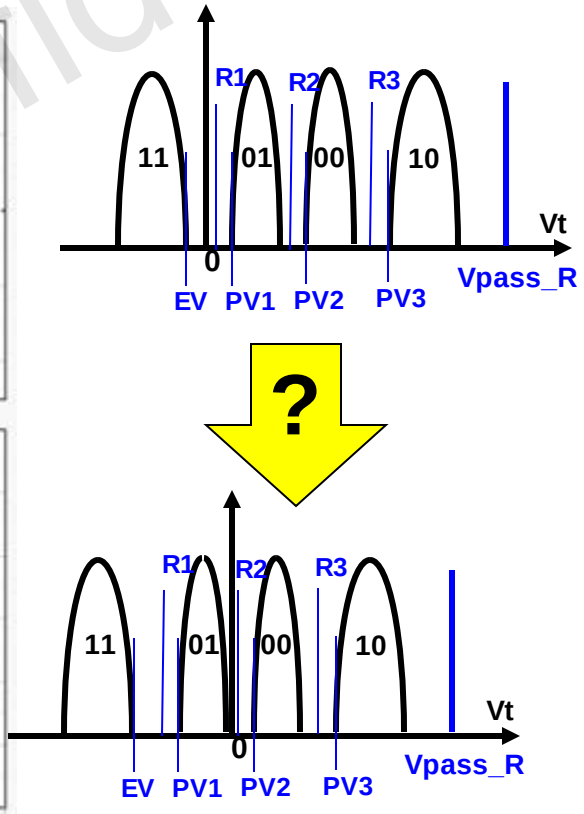
Micron/Intel Confidential

“Shifted” Read Window

- Since NAND has somewhat symmetric Program/Erase window (e.g. +5V/-5V), ideally we would like to set the Vt window to be straddling the 0V.
- However, this requires ability to read –ve Vt → -ve WL voltage is required
- ➔ Adds design complexity and die size

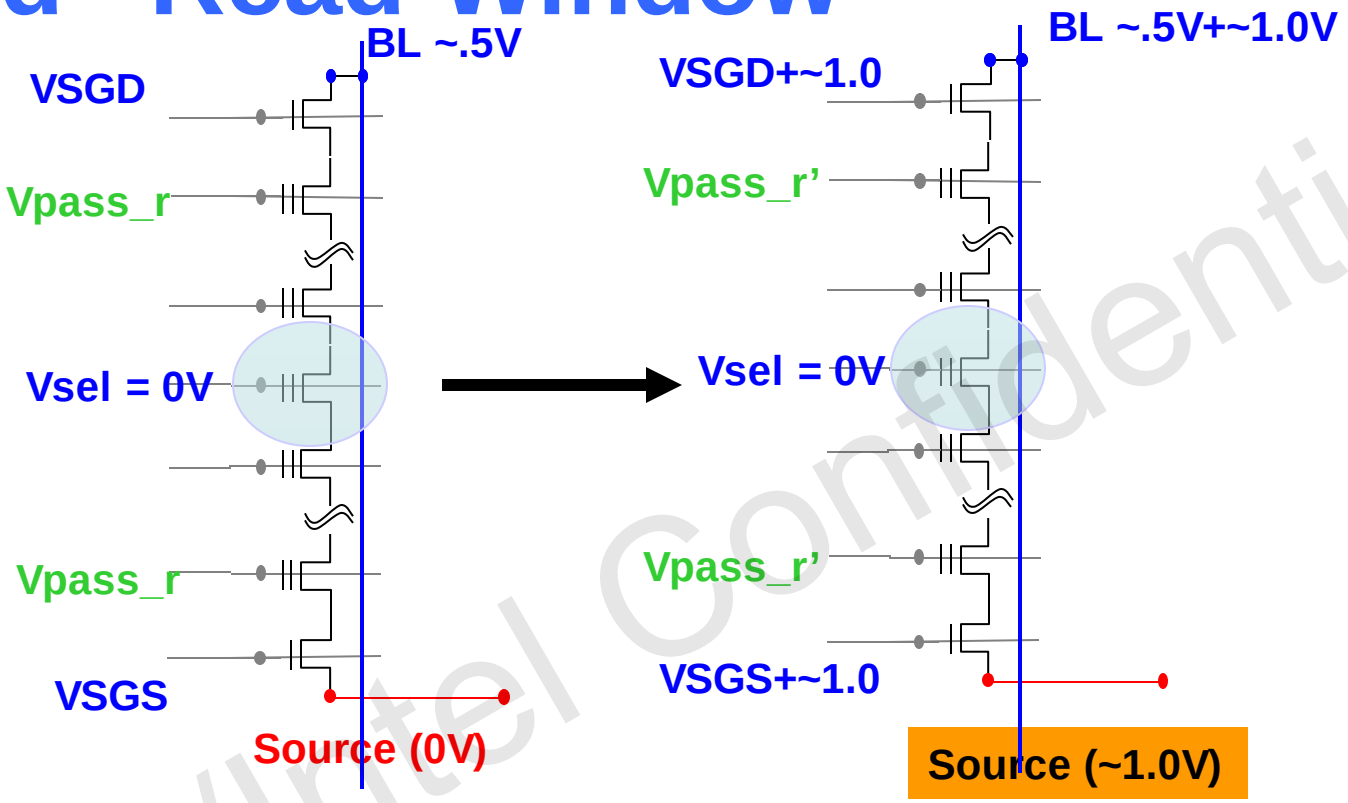






Micron/Intel Confidential

“Shifted” Read Window



- We can enable –ve Vt read capability by shifting the source voltage +ve
  - With Vs = 0V and Vg = 0V, select cell will not conduct if Vt > 0V (Vgs – Vt < 0)
  - With Vs = 1V and Vg = 0V, select cell will not conduct if Vt > -1V (Vgs – Vt < 0)
- +ve source bias allows –ve Vt read capability without needing –ve voltage on die. Ability to shift window –ve is limited to 1.0-1.5V due to other circuit limitations.

➢ Shifted window allows another flexibility to optimize MLC/TLC Vt centering

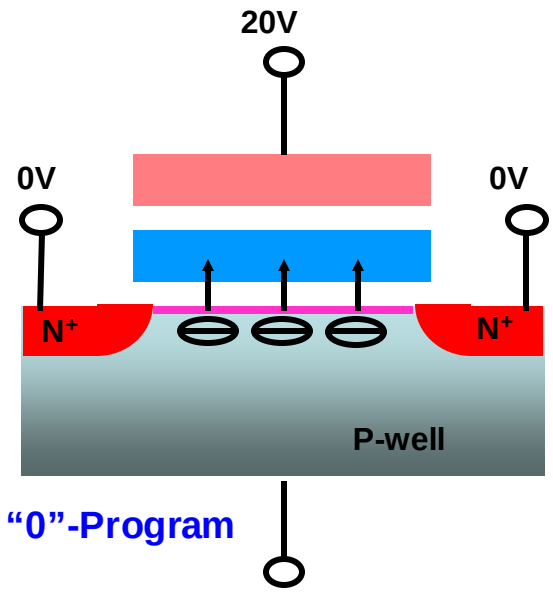
Micron/Intel Confidential

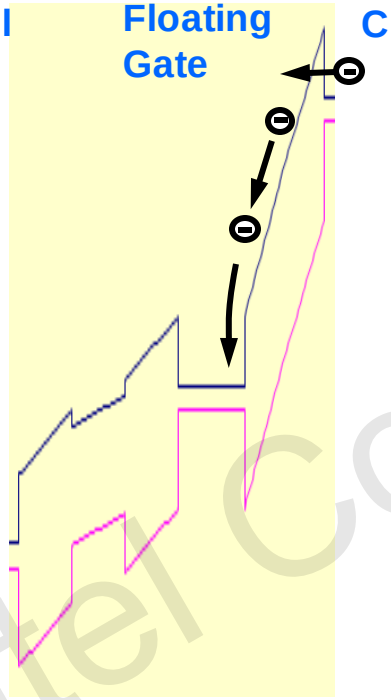
Programming

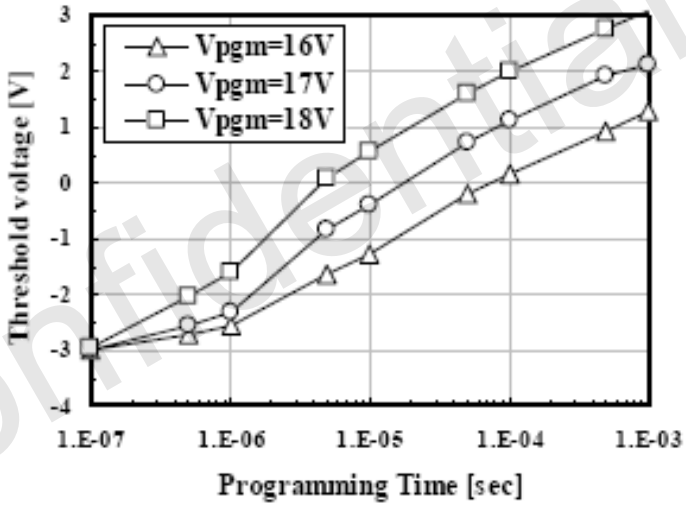
- FN Tunnel Programming
- Program/Erase Vt window requirement
- Program slope
- Program Saturation
- Curvature effect in 3D NAND
- Program Vt sigma
- Programming algorithm
- Program Noise
- SSPC / Proximity Effect
- Program Inhibit / Program Disturb
- Boost loss mechanisms
- Hot-e Disturb
- SG Leakage

Micron/Intel Confidential

NAND Flash Programming - FN Tunneling



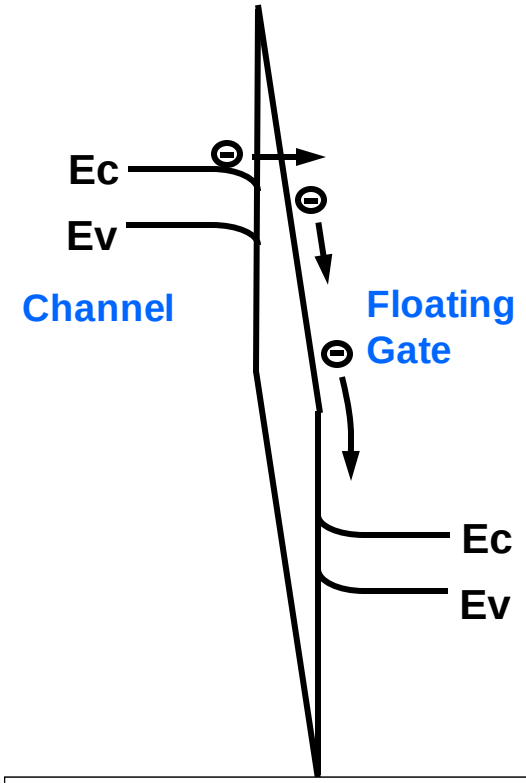




- Tunnel Programming from channel by biasing the Top Gate positive with respect to the channel and S/D.
- Actual Program Time (cumulative pulse width) ~100us. Program current ~ Displacement and Tunneling current. Low current allows large parallelism

Micron/Intel Confidential

FN Tunneling

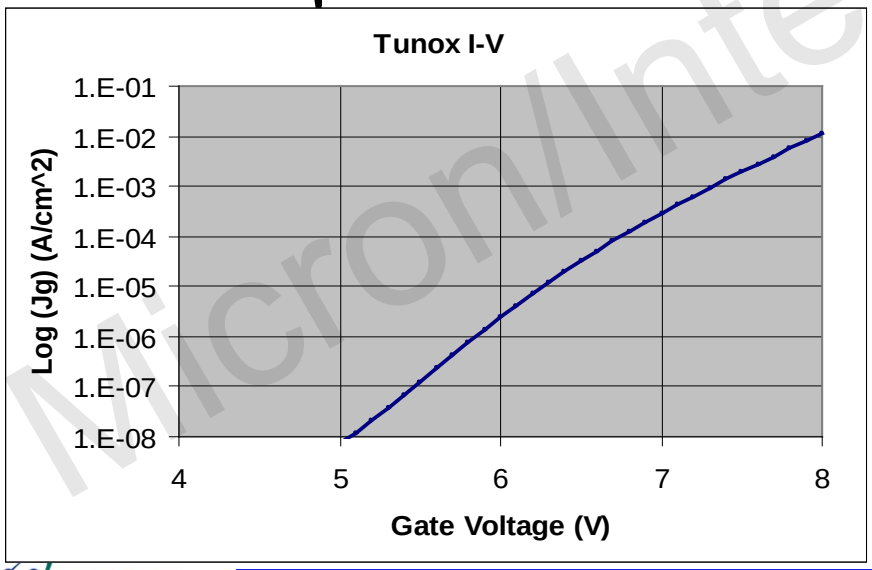


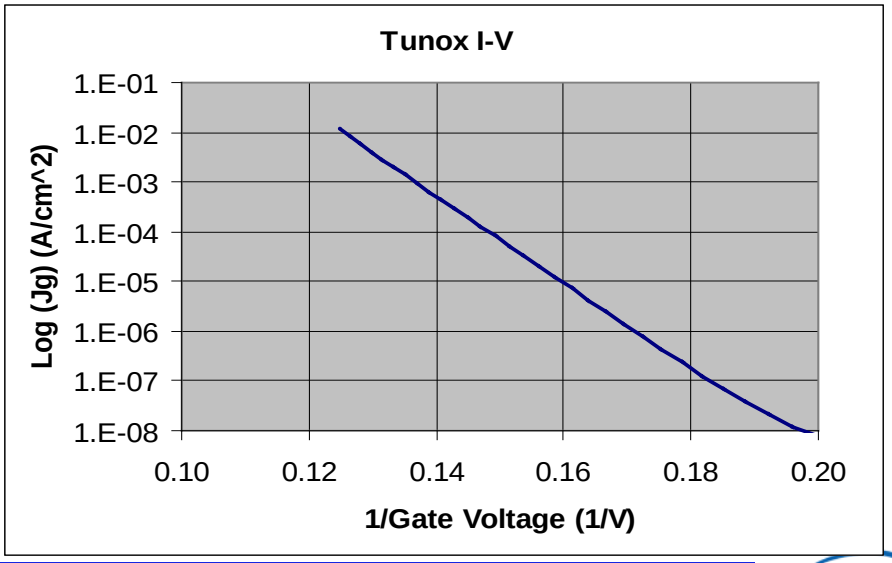
- Electrons tunnel through the first 25-30 Å of SiO_2 and then drift to the FG
- Fowler-Nordheim tunneling equation says that current varies with field as:

$$J = a E^{2} e^{\left(\frac{-b}{E}\right)}$$

Coefficients a and b are also commonly known as α and β

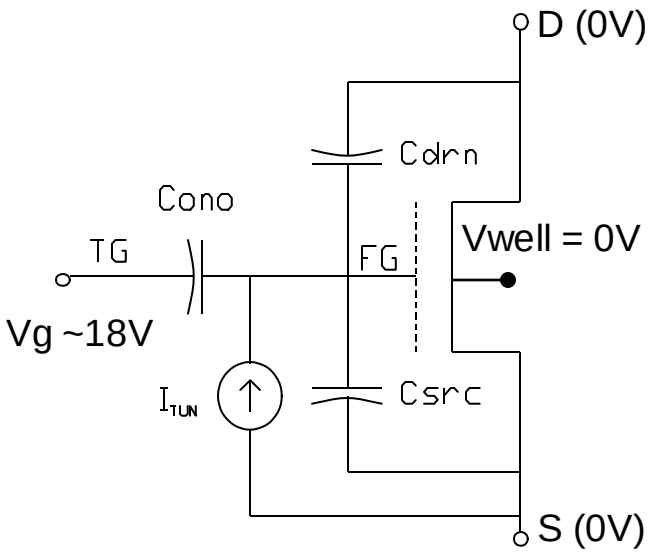
➢ The current changes by ~10X with ~0.8V change in the FG voltage





Micron/Intel Confidential

Program V_T vs. Time



- Equiv. ckt shows Floating Gate getting charged by I_{TUN} (actually –ve current as it is the electrons tunneling onto FG)

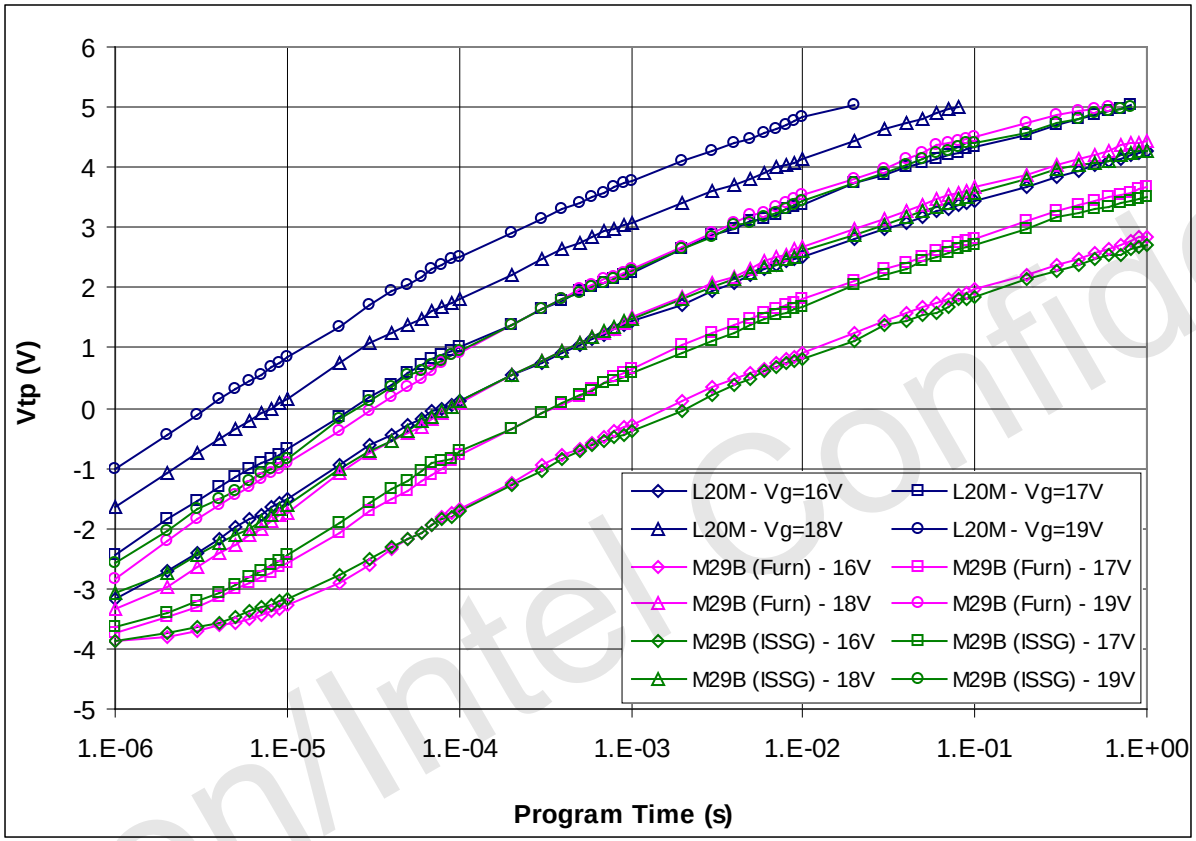
$$V_{FG} = V_{CG} \cdot GCR + \frac{Q_{FG}}{C_{TOT}} \qquad V_T = \frac{V_{TFG}}{GCR} - \frac{Q_{FG}}{C_{ONO}}$$

- At the beginning of program, V_{FG} is high and I_{TUN} is large, so high tunnel current and programming is fast

- As number of electrons on FG increases the FG voltage reduces (by ΔQfg/Ctot) decreasing I_{TUN} as given by the FN Equation
- Each ~0.8V decrease in V_{FG} drops I_{TUN} by ~10x
  - Programming rate drops by 10x, 100x, etc as the floating gate gets charged by .8V, 1.6V, ...
    - ➢ Makes Programming vary as a "log time"

Micron/Intel Confidential

Programming



Example Exhibit only. Exact Vt vs Vg dependence is very technology specifics (Tunox thickness, GCR, UV_Vt, ..)

- Log Time behavior of the program Vt can be seen above
- A rough rule of thumb is that the cell Vt changes by 1.3-1.5V/dec under programming conditions (~10-100us). The slope decreases to ~0.5-0.6V/dec as we go longer in time (~30-300s).

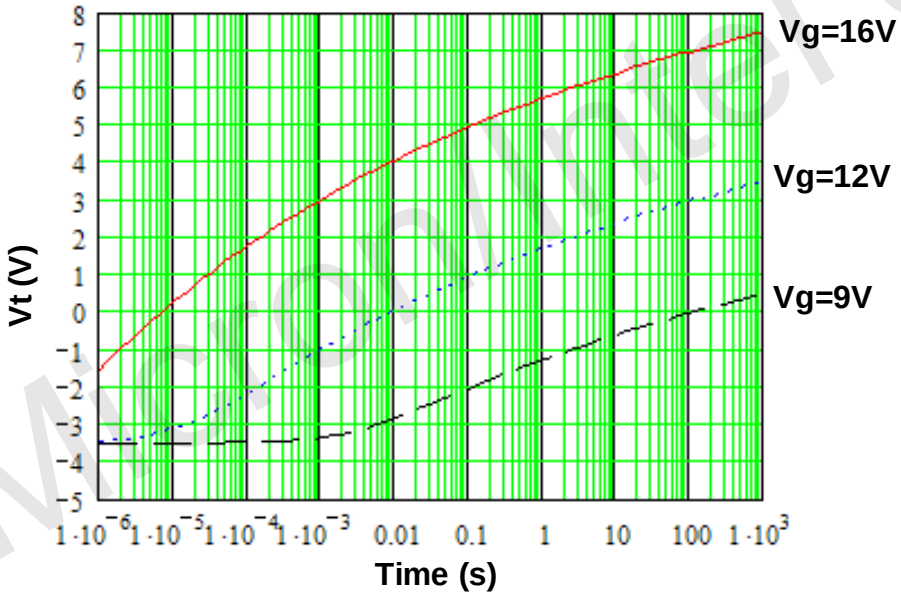
Micron/Intel Confidential

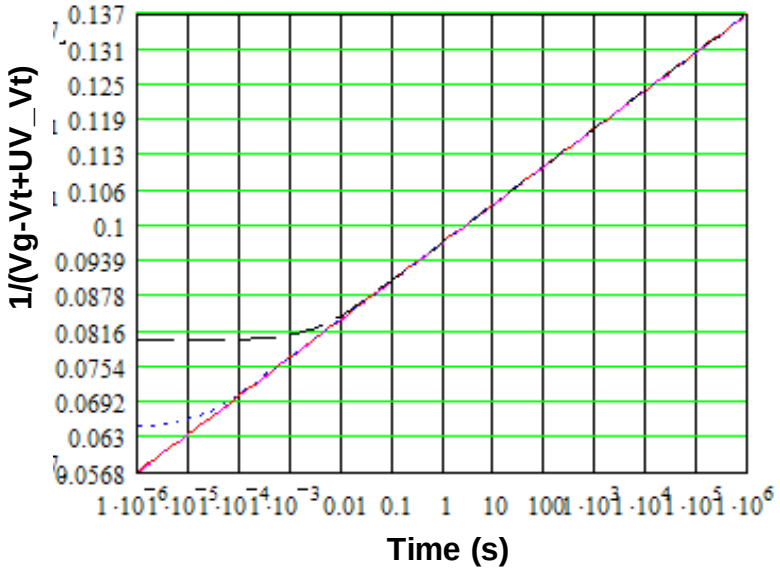
1/(Vg-Vt) Curves

Plotting 1/(Vg-Vt) vs. ln_t gives a straight line with an intercept and slope given by:

$$\left[\left(Vg - Vw + Vth \cdot Gcr^{-1}\right) - Vtx\right]^{-1} := \frac{Gcr}{\beta \cdot T_{ox}} \cdot \ln\left(\frac{Area \cdot \alpha \cdot \beta \cdot 1s}{Ctot \cdot T_{ox}}\right) + \frac{Gcr}{\beta \cdot T_{ox}} \cdot \ln\left(\frac{t}{1 \cdot s}\right)$$

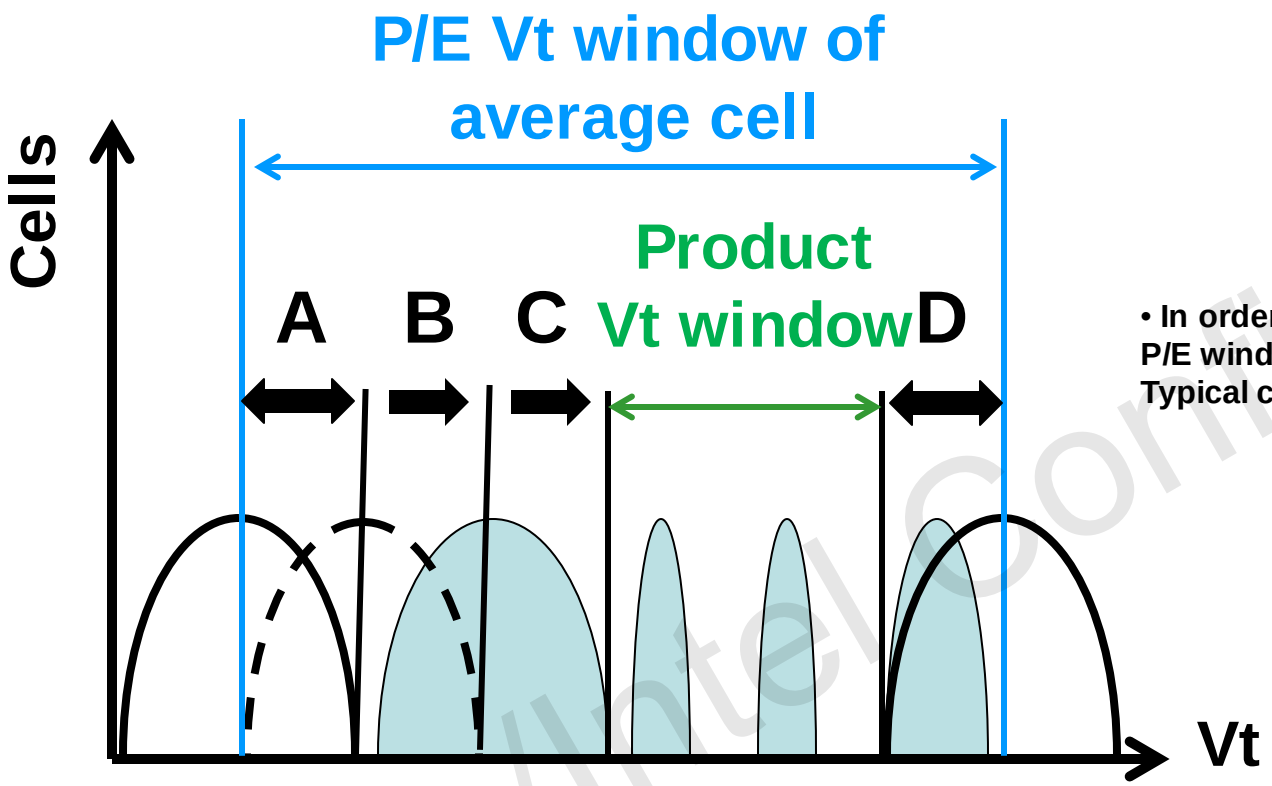
- Where Vg is the gate voltage, Vth/Gcr is the UV_Vt of the cell
- Vw is the Well or S/D voltage (depending on erase or programming)
- α and β are the FN tunneling coefficients. Areas is the tunneling area and Ctot is the Total FG capacitance. Tox is the Tunnel-oxide thickness





Micron Confidential

P/E window requirement

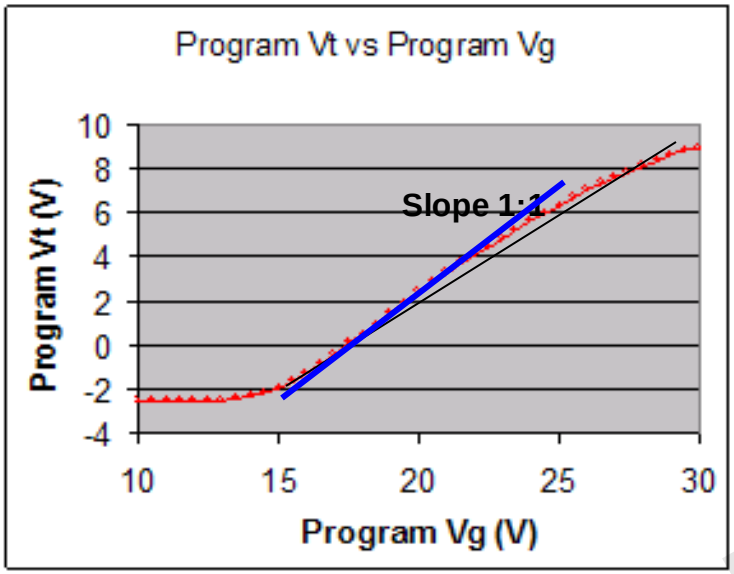


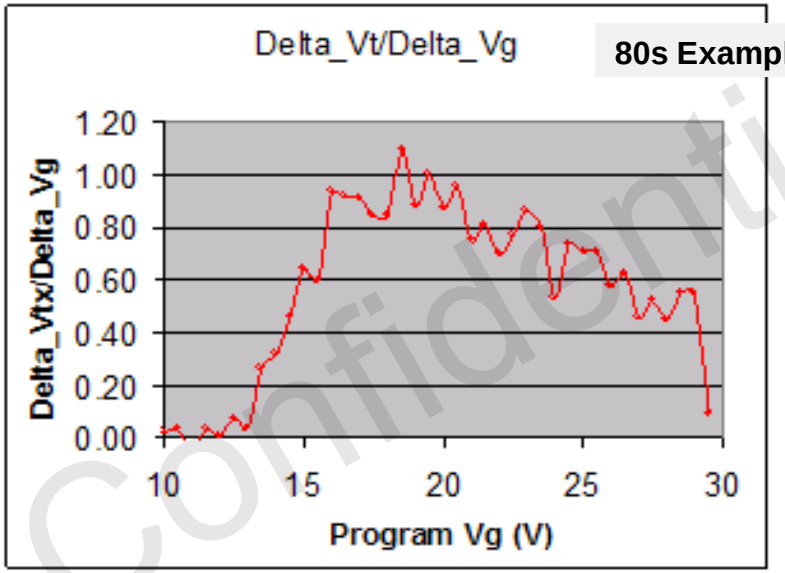
- In order to support ~4V P/E window in product, Typical cell needs to have >8V P/E window.

Product P/E window	PV3-R1	~4V
A. Cycling trap up margin	Vte_sat. degradation post cycling	~2V
B. FGFG margin	Erased Vt shifts up after neighbor cells programmed	~0.25V (=10% x 2.5V (average Vt shift with RRP data)
C. EVS margin	Cell to cell Vte variation margin	~2V (=0.25V x 4sigma (~+1 sigma from ECC requirement x2 (EVS and PVS))
D. PVS margin	Cell to cell Vtp variation margin	
Required Cell P/E window		**> 8.25V**

Micron/Intel Confidential

Pgm Vt vs Program Vg (80s)



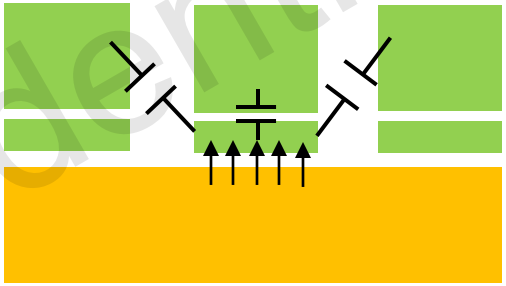


- Programming is done as ISPP (incremental step pulse programming)
- For a given programming pulse width, the program Vt tracks the program gate voltage. VgVt (= Vpgm – Vtx) is a good cell programming metric.
- VgVt is primarily determined by the Cell UV_Vt, FN Tunneling coefficients, Tunnel-oxide and ONO thicknesses, GCR, etc.
  - VgVt-UV_Vt is proportional to Cell EOT (= Tox/GCR)
- We would expect ΔVtx/ΔVg to be ~1.0. However, some deviation from 1.0 is seen due to depletion effects (GCR difference between Program and Read), and CG impacting the Vt directly (instead of through FG).

Micron/Intel Confidential

Programming Slope

- Typically the program slope should be = 1.0 (Assuming fixed Vinh).
- When you increase the Vcg (during Pgm) by ΔVcg, the Vfg moves up by GCR*ΔVcg and as a result we inject electrons into the FG and FG comes down by GCR*ΔVcg. Later when you read the cell Vt, you will need to apply a higher Vcg given by GCR*ΔVcg/GCR = ΔVcg to get the cell to turn on.

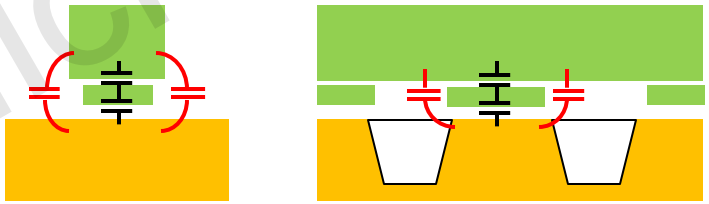


  - **ΔVt = ΔVcg during programming! ➔ Program slope = 1.0**
- If the Vinh on the adjacent WL was changed from pulse to pulse, then there will be coupling from the adjacent WL and which would have moved the Vfg by more than GCR*ΔVcg, so ΔVt will be more
  - **ΔVt = ΔVcg + 2 * WL-FG * ΔVinh**
- If the GCR during programming and Read were different (e.g. due to Poly depletion effects and different IPD fields during Programming and Read):
  - **Program slope = GCR_p/GCR_r** (Assuming fixed Vinh)

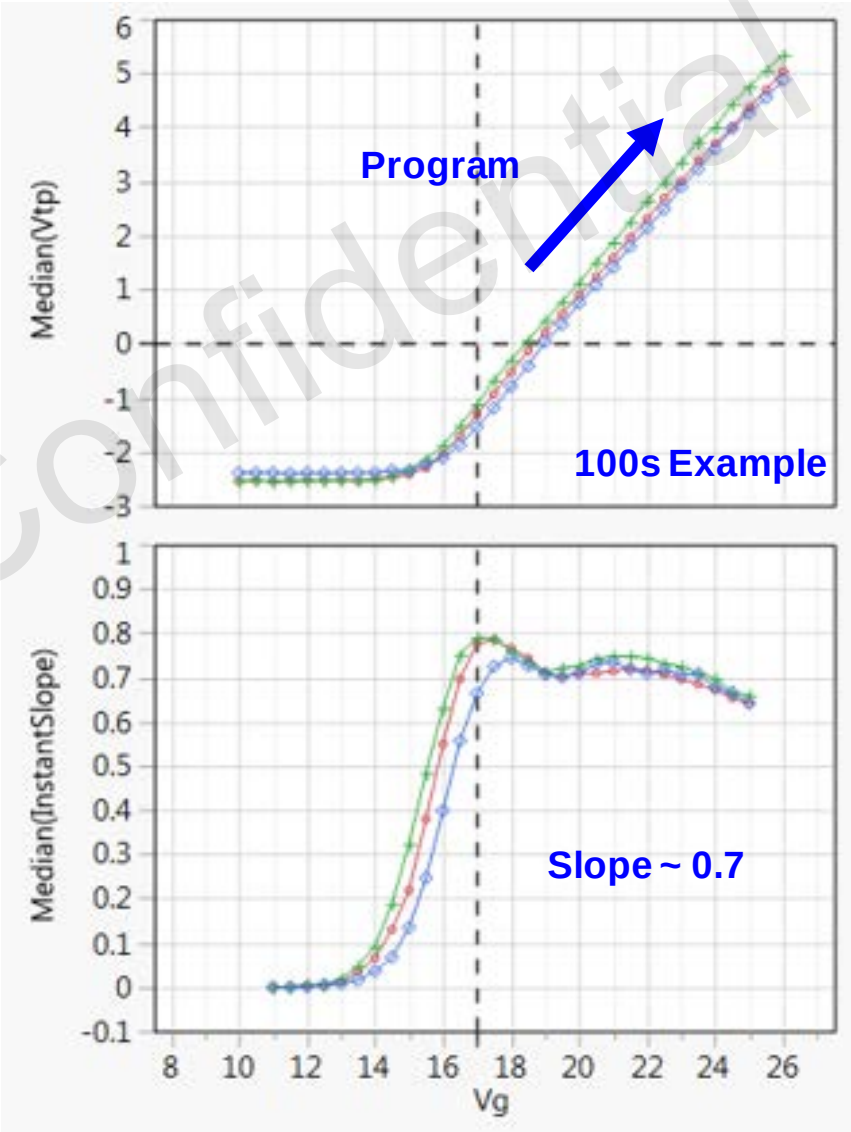
Micron/Intel Confidential

Programming Slope

- Additional factors impacting pgm. Slope:
  - As the cell programs, field across the IPD increases and there could be a competing effect causing increasing IPD leakage causing <1.0 Program slope.
  - In the limiting case, cell will stop programming (Programming saturation: Vtp_sat).
  - (Partial capture (e.g. 70%) of the electrons in the FG does NOT change the program slope. It can cause VgVt to increase slightly, but slope will not be affected as long as capture % doesn't change)
  - Direct CG-AA capacitance can reduce Program slope:
  - Program slope ~ Ceot/(Ceot+CG-Ch)



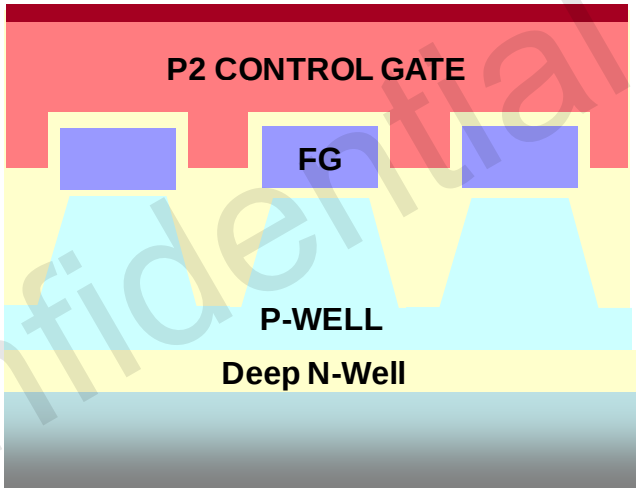
Direct CG-AA capacitance



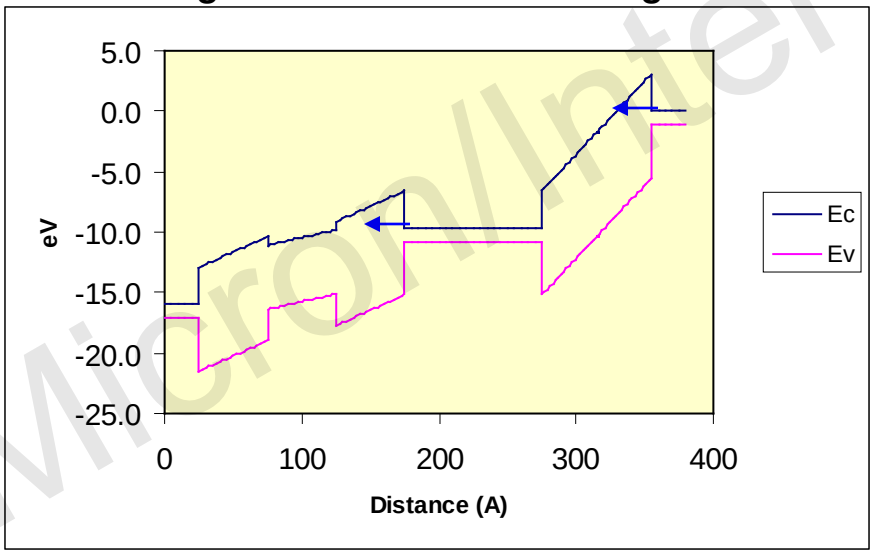
Micron/Intel Confidential

Program Vt Saturation

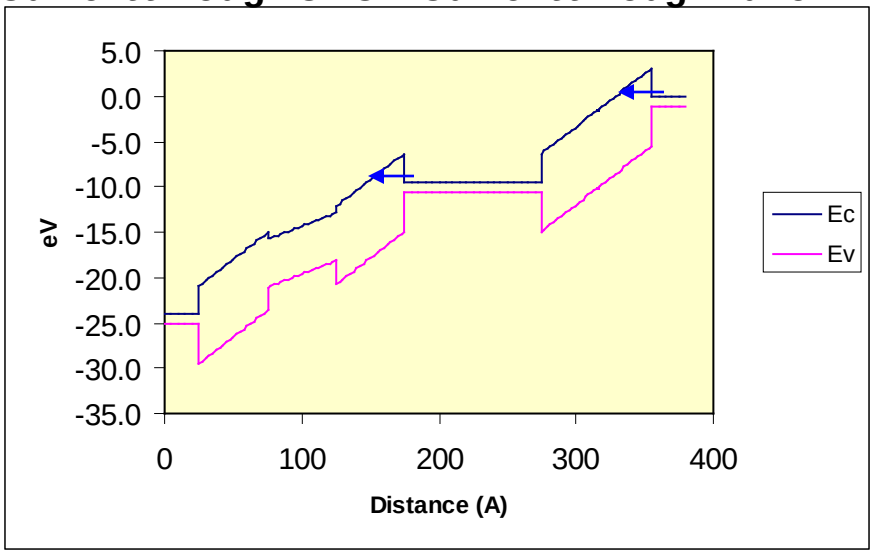
- Vt saturation at high Vt because of electrons tunneling out through the IPD.
- As the programming gate voltage increases and the cell gets programmed higher and higher, the field across the IGD increases and the leakage through the IGD starts becoming large and can become equal to the tunneling current



Low Program Vg and Vt (Vg = 16V, Vt =0V). Current through ONO << Current through Tunox



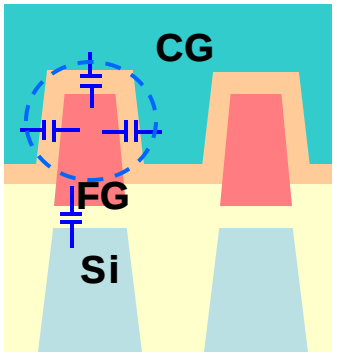
High Program Vg and Vt (Vg = 24V, Vt = 8 V). Current through ONO = Current through Tunox



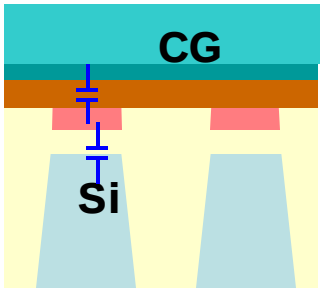
Tunox = 80A, ONO = 150A Physical / 125A Electrical, Wrap Ratio ~ 2.5

Micron/Intel Confidential

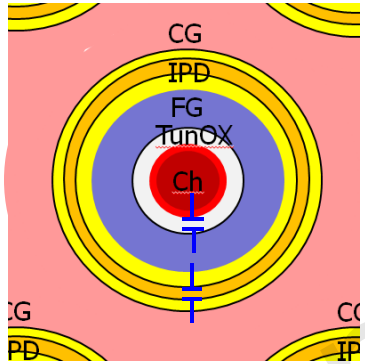
3D NAND Pgm/Ers Window



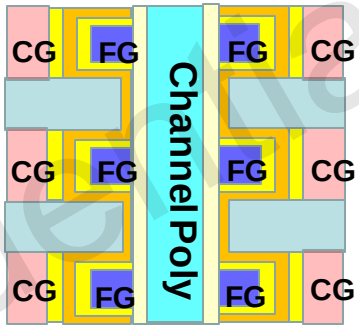
70s: Wrap



80s: High-K / MG



100s 3D cell

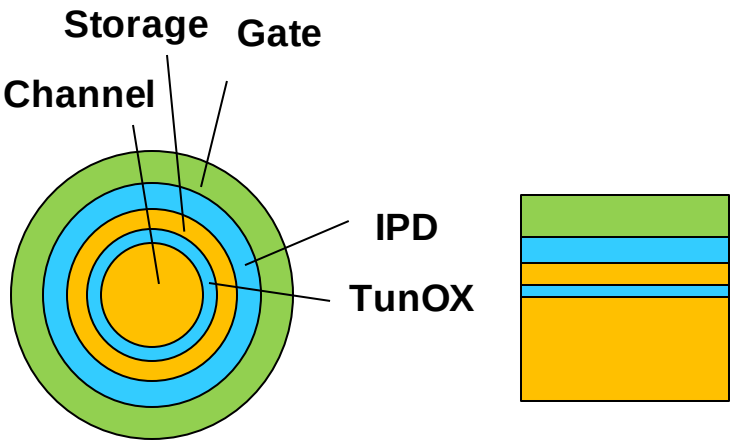


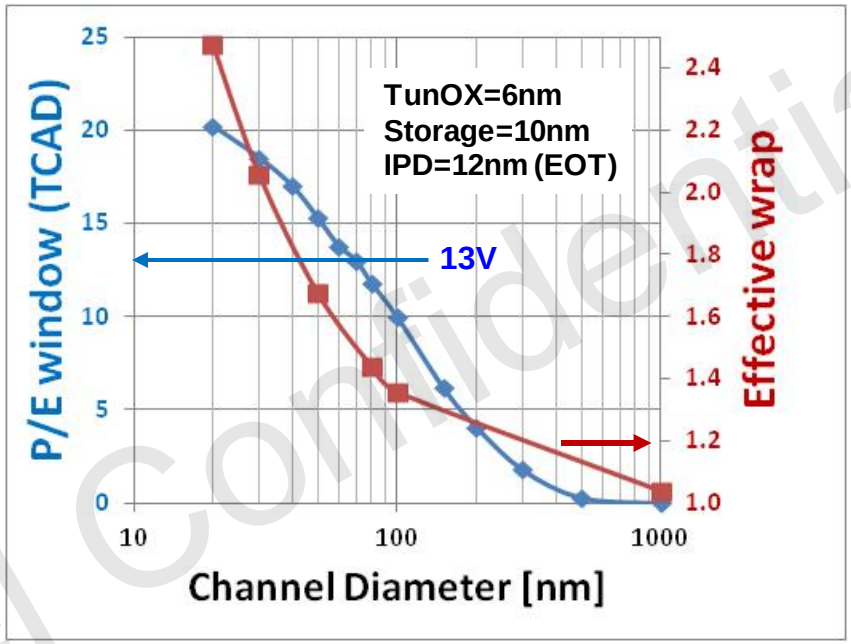
- Good Program/Erase window requires high Gate coupling Ratio (FG couples stronger to the Control Gate) → Higher Tunnel-oxide field and Lower IGD field
- 3D NAND (100s) cannot have conventional wrap (like 70s) or High-K/MG (like 80s). 3D NAND counts on the circular shape to provide higher coupling area between FG & CG relative to FG & Si.
- Alternatively it can be thought of as electrical field being higher closer to the center of the circle (e.g. Si/Tunox interface) than outside (FG/IGD interface) which allows FN current through Tunox but not IGD

Micron/Intel Confidential

GAA Wrap Effect

Gate-all-around (GAA) **Planar**

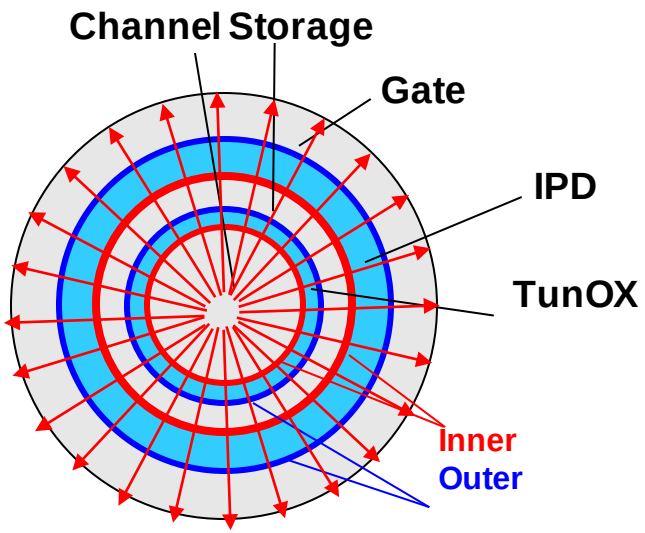




- Due to GAA architecture of 3D, IPD area is larger than TunOX area, providing effective wrap ratio > 1.0
- This increases ratio of E_field_tunox / E_field_ipd, enabling a wider progam/erase Vt window.
- <70nm channel diameter is required to achieve >13V P/E window.

Micron/Intel Confidential

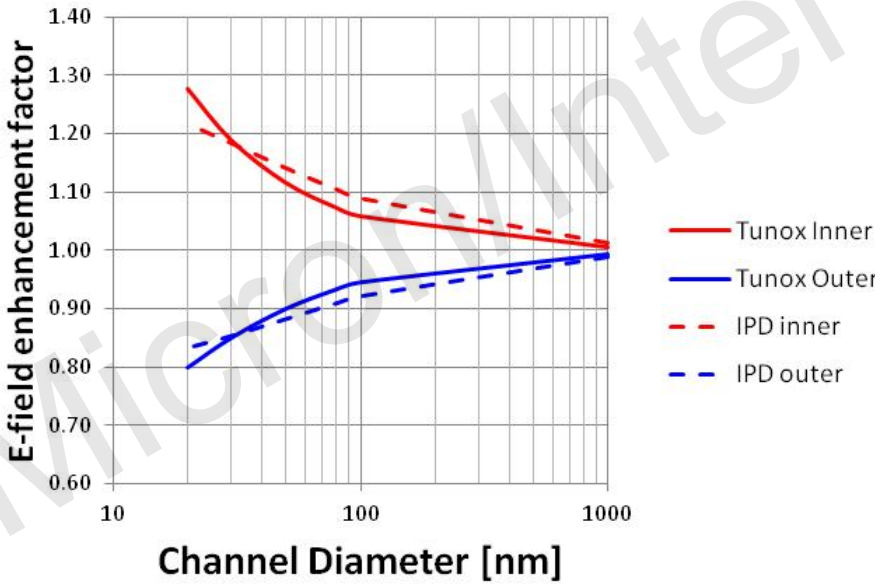
Curvature E-field effect

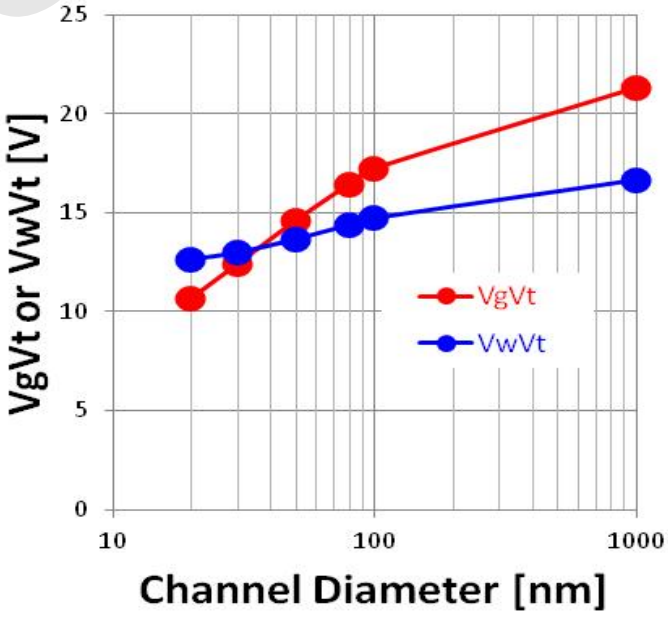


- E-field across TunOX and IPD are modulated due to the curvature effect.
- E-field at inner side of the interface increases and it decreases at the outer interface.

- Impact on P/E window is small.
Inner(tuox) vs. Inner (IPD) compete for Vtp_sat.
Outer(tuox) vs. outer (IPD) compete for Vte_sat.

- Impact on VgVt and VwVt is asymmetric.
VgVt has stronger diameter dependency. GCR and E-field effects are additive.

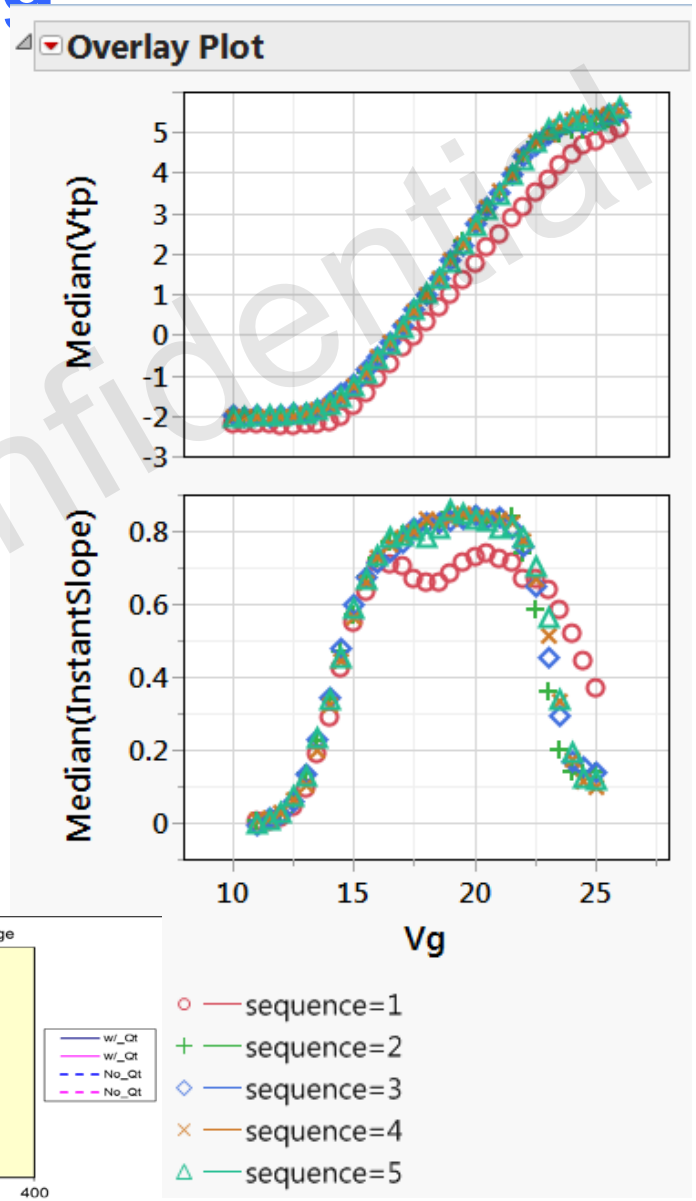


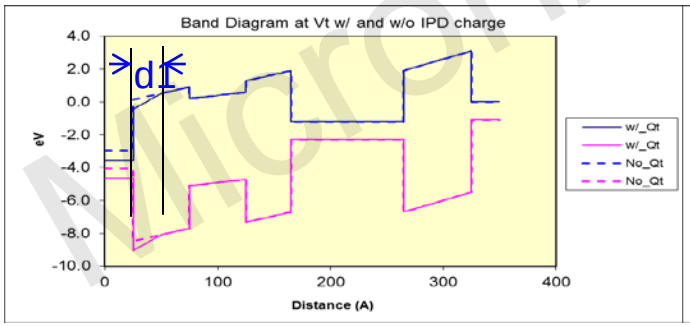


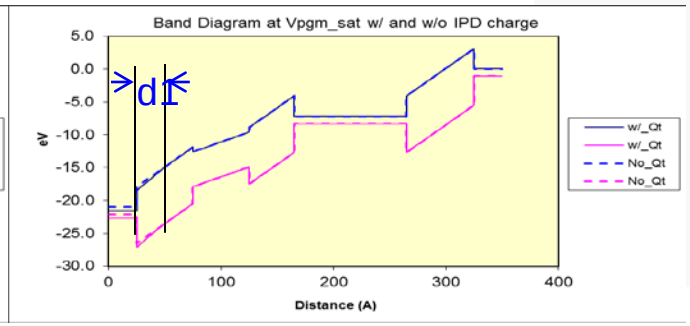
Micron/Intel Confidential

Programming Conditioning

- 3D NAND suffers from low GCR and hence low P/E window
- We can use all the help we can to improve P/E window
- Trapping some charge in the middle Nitride (by subjecting Cell to Program sat condition) of IPD result in reduced IPD leakage and thus improved Programming and Vtp_sat in subsequent programming operations
- Pre-conditionging/FCC: using several targeted program/erase cycles at probe to “condition” the cell for better programming capability.
- It is critical that conditioning doesn’t go away during assembly bake or reliability bakes.

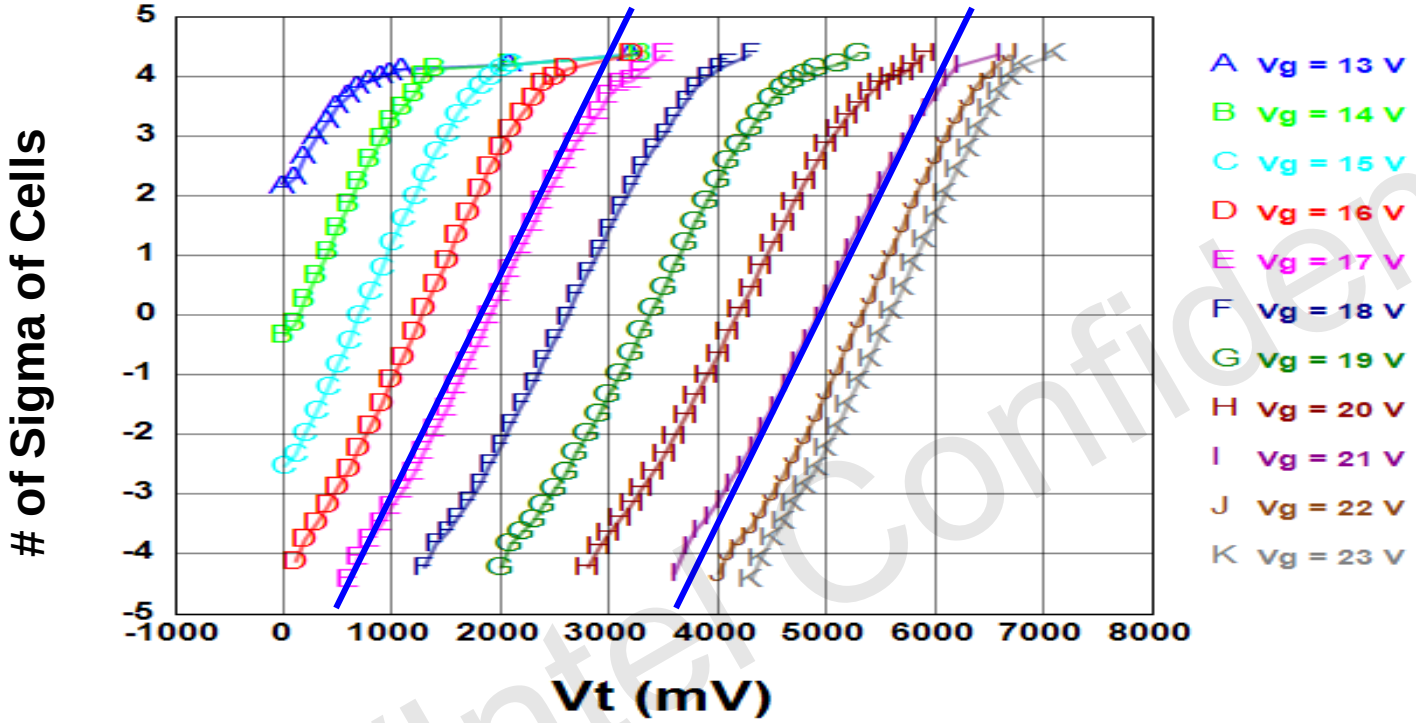






Micron/Intel Confidential

Array Programming Characteristics



- Cells in array program to different Vt's when programmed using a given gate voltage pulse
  - Fast cells are those that program to highest Vt, while slow cells are those that program to lowest Vt, for a given gate voltage
  - Slope of distribution is PVS (**P**rogram **V**t **S**igma)
- Fast cell and slow cell move by same amount for subsequent pulses, maintaining the difference in Vt's reached for a given gate voltage

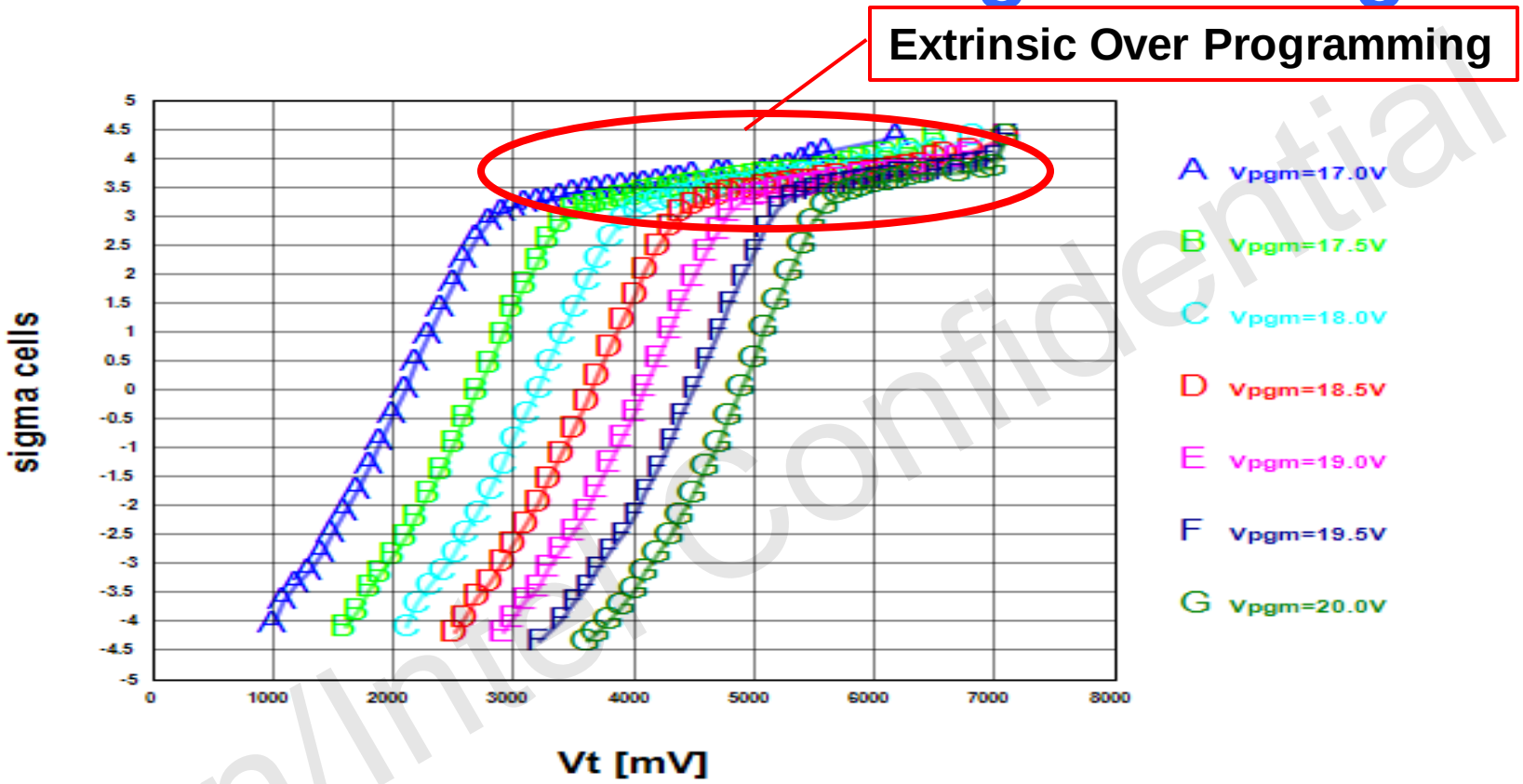
Micron/Intel Confidential

Program(Erase) Distributions

- PVS is one of the most important parameter for NAND cell. It sets the program time, Program disturb bit error rate, max voltage requirements, as well as dielectric scalability. It is very critical to minimize PVS
- Program/Erase Distribution are affected by several factors –
  1. Native Vt variations in the cell – random doping/trap variations
  2. Tunneling current variations from cell to cell due to - Tunnel barrier variations, Si/Poly1 asperities, etc.
  3. Cell Coupling Ratio variations – variation in W, L, Recess, Tox, Tono, etc.
  4. Others - Bird's beak, systematic effects from design
- Empirically we see that the PVS gets worse with cell scaling – almost as 1/sqrt(W*L). Here item #1 and #2 would be expected to scale as 1/sqrt (W*L) ➔ 1/k.
- Our experience suggests that, item #3 and #4 can be fixed/improved to the level that they are not fundamental limiters.

Micron/Intel Confidential

EOP – Extrinsic Over Programming



- In addition to the intrinsic PVS, 100s also showed extrinsic Over Programming tail
- Was attributed to the tunnel-oxide weakness (electrical) requiring optimization of TB dep/cut

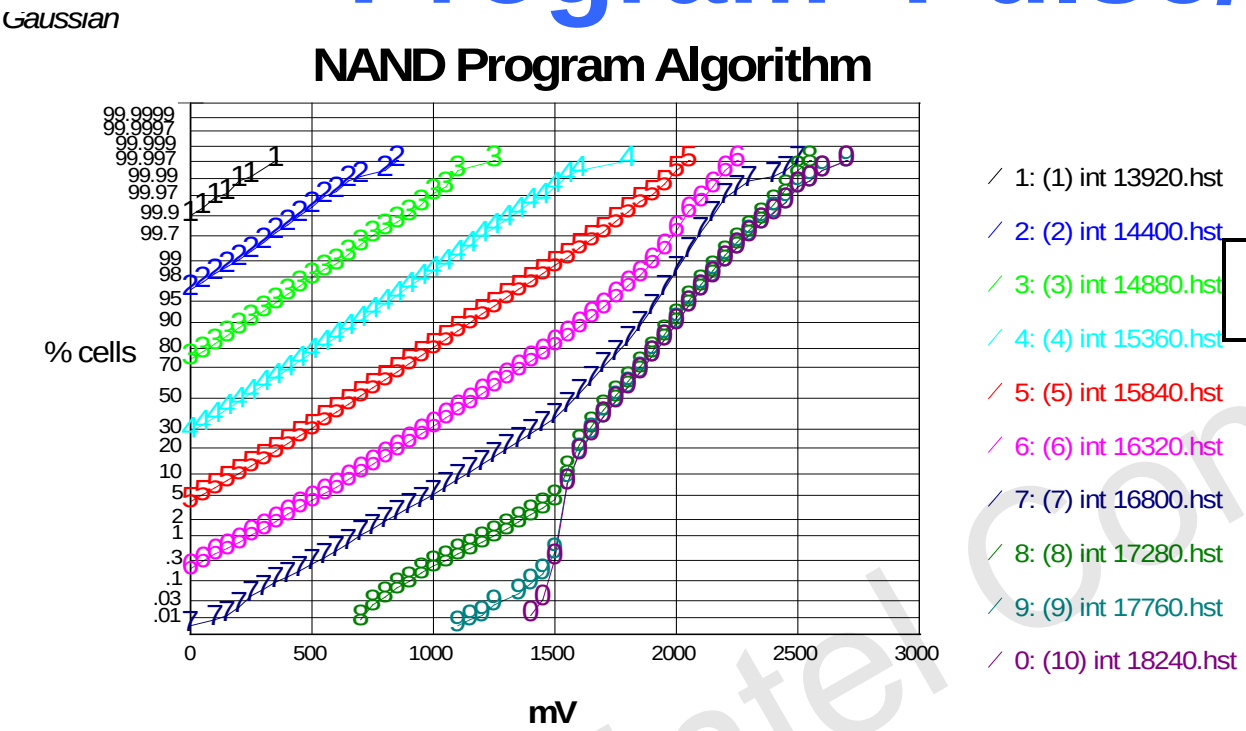
Micron/Intel Confidential

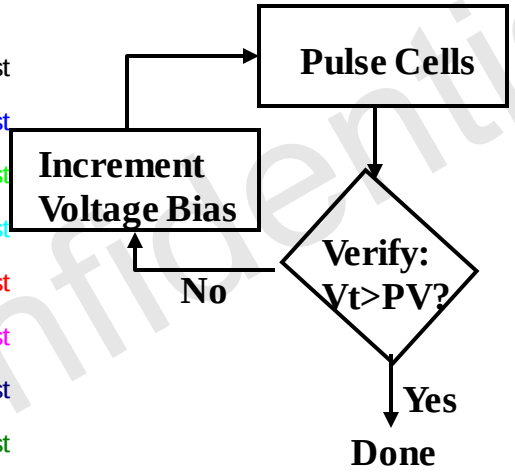
Programming Algorithm

- Since different cells in the array program to different Vt's when programmed using a given gate voltage pulse, we need to use some sort of algorithm to be able to program the cells to tight program distribution
- Programming (Placement) algorithm:
  1. Start with a low programming gate voltage
  2. Give a programming gate pulse (~20us)
  3. Check the cell Vt (Verify operation).
     - If cell Vt > Target Vt, end programming loop
     - If the cell Vt < Target Vt, increment the gate voltage (~0.5-1.0V).
     - Repeat steps 2-3.
- Starting gate voltage needs to be low enough to ensure the cell does not over program (dictated by lowest Vg-Vtx) in the first pulse
- Final gate voltage is dictated by the cell that has the highest Vg-Vtx
- Total number of programming pulses, and hence the programming speed is dictated by the spread in the Vg-Vtx (=> PVS)

Micron/Intel Confidential

Program- Pulse/Verify

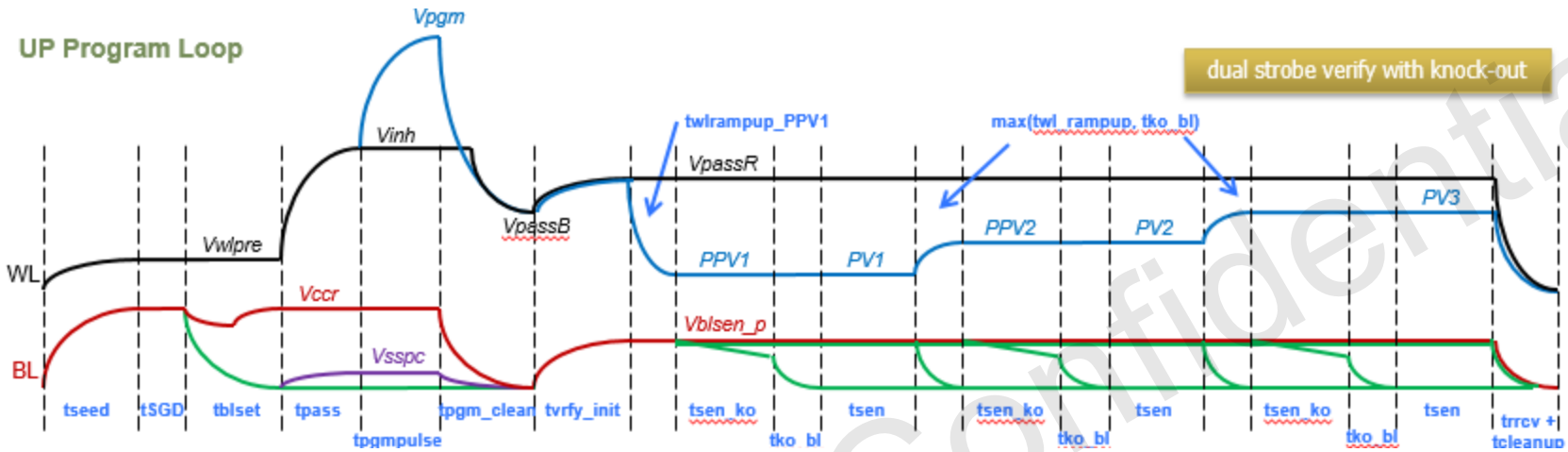




- Each cell in 16kB Page Buffer is twiddled off when cell successfully Verified. Example above shows cell being programmed using ~0.5V gate step starting with a gate voltage of 13.9V.
  - Fastest to program cell crosses program verify (~1.5V) at Vg=15.4V
  - Slowest to program cell crosses program verify (~1.5V) at Vg=18.2V
  - Ideally this should give a program Vt distribution width that is 0.5V wide. However, the actual distribution is >1V (due to Program Noise (see later))
  - Without a pulse/verify algorithm the distribution would have been 3-4V wide

Micron Intel Confidential

Simplified Programming waveform

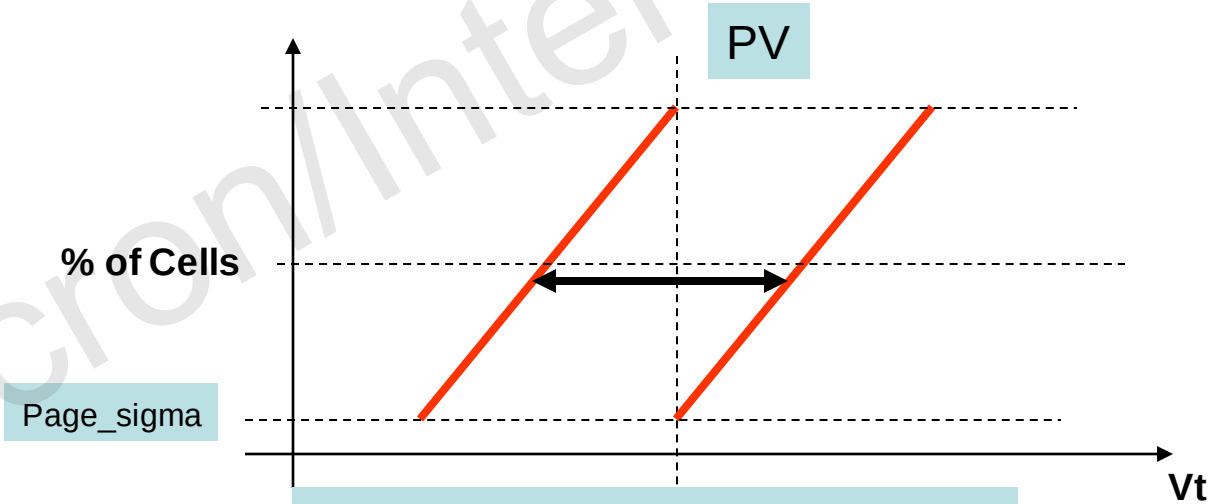


- Total Program time (t_{PROG}) is # of pulses * loop-time
- We discussed number of pulses earlier (dependent on PVS, gate-step, cfbyte)
- The Loop time is made up of the above components, and each is carefully optimized.

Micron/Intel Confidential

Program Speed: Page Pgm

- Program Time
  - The gate step will determine the program Vt distribution that can be achieved. Smaller the gate step tighter the Program Vt distribution
  - Increased PVS leads to a smaller VgVt_min, and larger VgVt_max
  - As a result, the Vg_start will have to be lower, and Vg_end higher leading to increased number of pulses and increased program time
    - Vg_end - Vg_start = 2*#_sigma*PVS + Guard-band (PV_max - ~PV_min)
    - # steps = ΔVg / Gate_Step



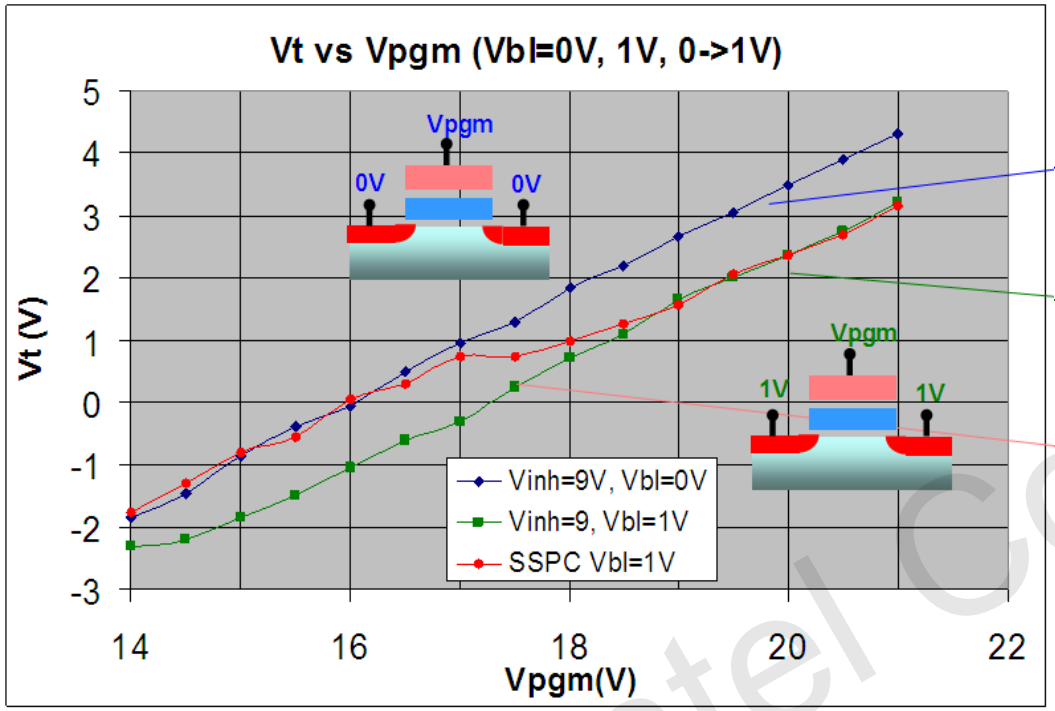
Micron/Intel Confidential

SSPC – Selective Slow Programming Convergence

- If we step up the program gate voltage by ΔVg and give a programming pulse, the cell Vt will also increase by the same amount (ΔVg)
- SSPC is a scheme that will cause the cell to move by less than ΔVg even though the Gate voltage increased by ΔVg.
- This is done by biasing the S/D/Channel (Bitline) of the cell of the cell to small voltage ~0.5V – 1.0V.
- Since the Tunneling current (Programming current) is actually a function of the FG to Channel field, even though the Vg is increased by ΔVg the actual channel to control gate voltage is increasing only by (ΔVg – ΔVbl) and as a result the increased Vg doesn't result in the corresponding increased current -> slow down of programming.
- Good way to achieve program distributions tighter than the gate step. Helps with programming speed as the gate step can be large without compromising the program placement width.

Micron/Intel Confidential

SSPC - Selective Slow Programming



Original Vt vs Vg for cell with Vbl = 0V

New Vt vs Vg for the cell with Vbl = 1V

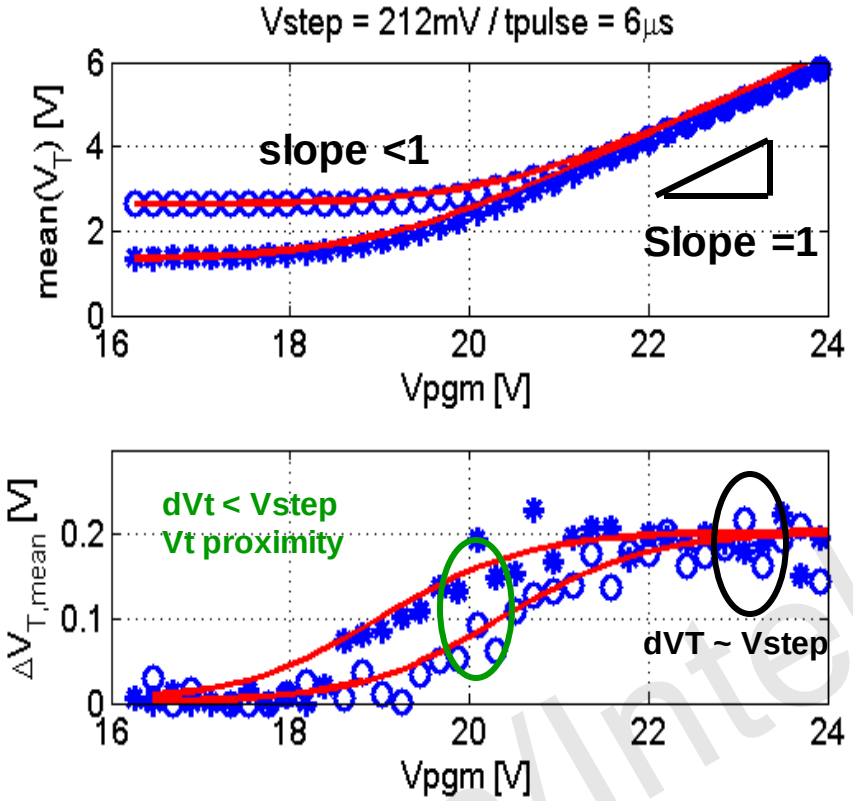
SSPC Regime. Vbl goes from 0V to 1V as Vg goes from 17.0V to 17.5V and then stays at 1V until it gets inhibited

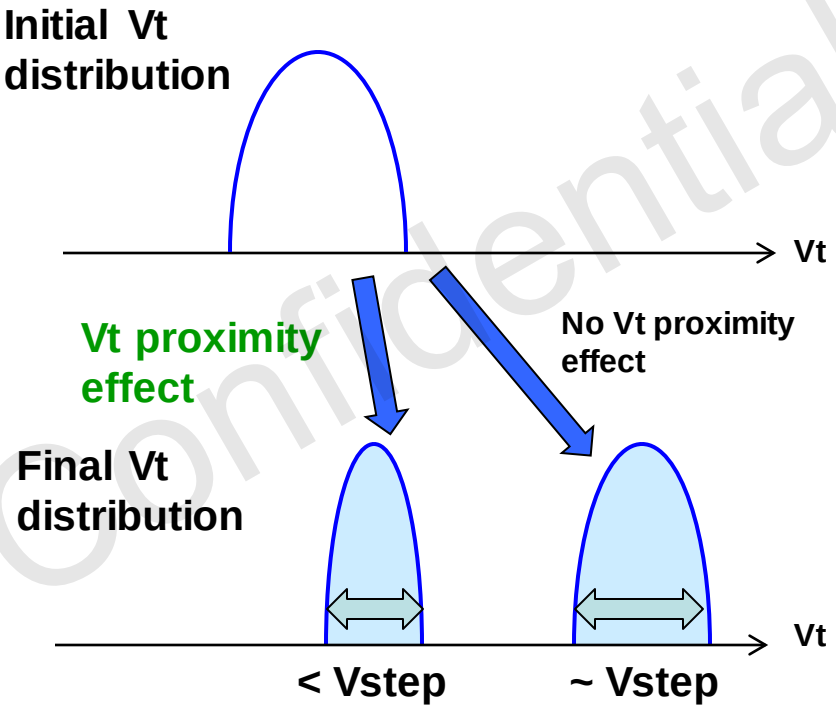
- As the cell reaches close to the target Vt, (as determined by passing the pre-verify), SSPC voltage is enabled: Vbl goes from 0V to the SSPC voltage (~0.5-1.0V).
- This slows down the programming of the cell for the next couple of pulses, allowing a finer placement of the cell Vt.

$\Delta Vt/\Delta Vbl = (1\text{-}CCR)/GCR \sim (1 + 2*WL\text{-}FG\ Coupling)$ -> "Magic Ratio"

Micron/Intel Confidential

Vt proximity effect

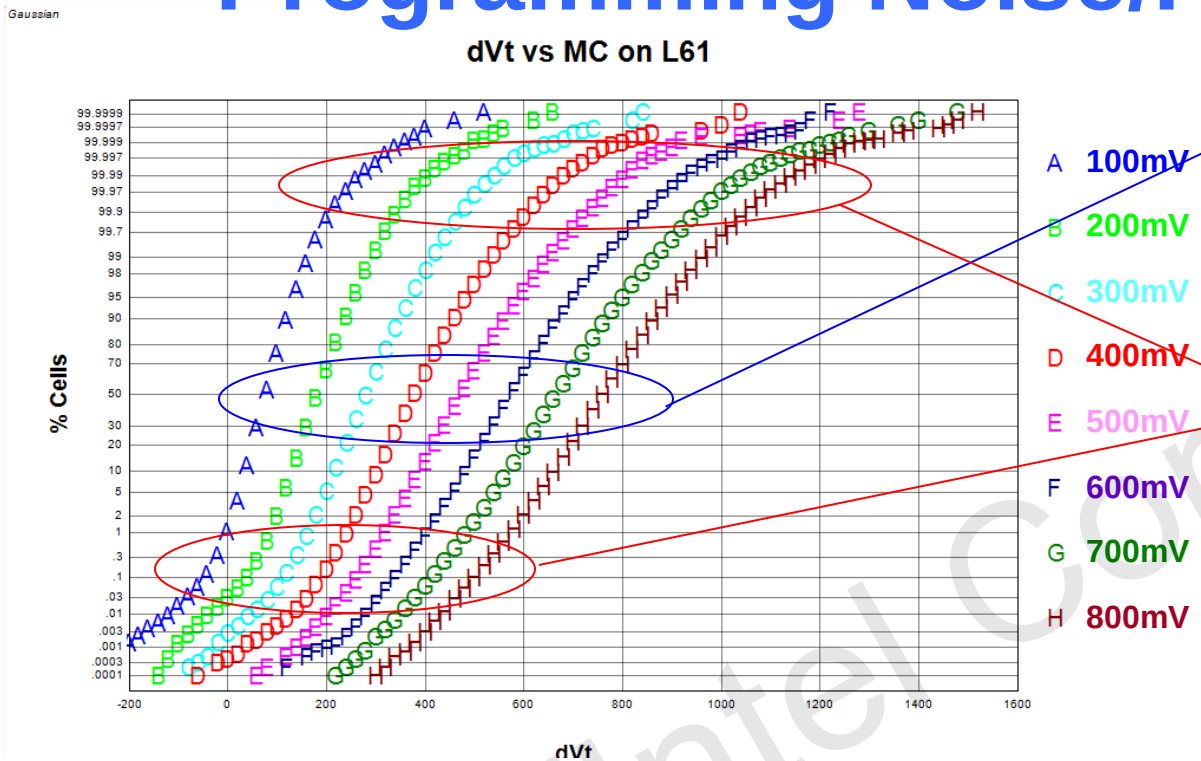




- If the program gate voltage starts low (Vg_start < Vt_init + VgVt) initially cell moves by ΔVt < ΔVg (Pgm slope < 1). Once Vg > Vt + VgVt, the cell Vt moves at the same rate as ΔVg (converges to Program Slope of ~1).
- If the final Vt is not far from initial Vt, a tighter state width is achieved.

Micron/Intel Confidential

Programming Noise/Fluctuation



Average cell Vt increases by the same amount as the Program gate step

However, large number of cells move by much larger than the gate step or much smaller than the gate step

- When programming a large number of cells using a staircase gate algorithm (Gate step = ΔVg), a typical cell Vt will move by ~ ΔVg from step to step. However, we find that all cells don't move by this amount and there is a distribution of Delta_Vts – A normal distribution followed by a tail
  - Normal Distribution comes from number fluctuation mostly dictated by cell capacitance that will dictate the number of electron/program step
  - Tail distribution comes from erratic over-programming (random events of enhanced tunneling). This part tends to increase with cycling as oxide degrades

Micron Intel Confidential

Programming Noise/Fluctuation

Tech Node	W	L	Tox	Tipd	Cipd	# Electrons/V	Δ_Vg	# Electron	Prog Noise
40s	55	72	75	135	4.39E-17	275	0.400	110	38
50s	40	50	75	130	2.54E-17	159	0.400	64	50
60s	30	35	75	120	1.71E-17	107	0.400	43	61
70s	28	24	75	110	9.67E-18	60	0.400	24	81
80s	20	20	60	85	5.31E-18	33	0.400	13	110
100s	251	15	70	130	2.96E-17	185	0.400	74	46

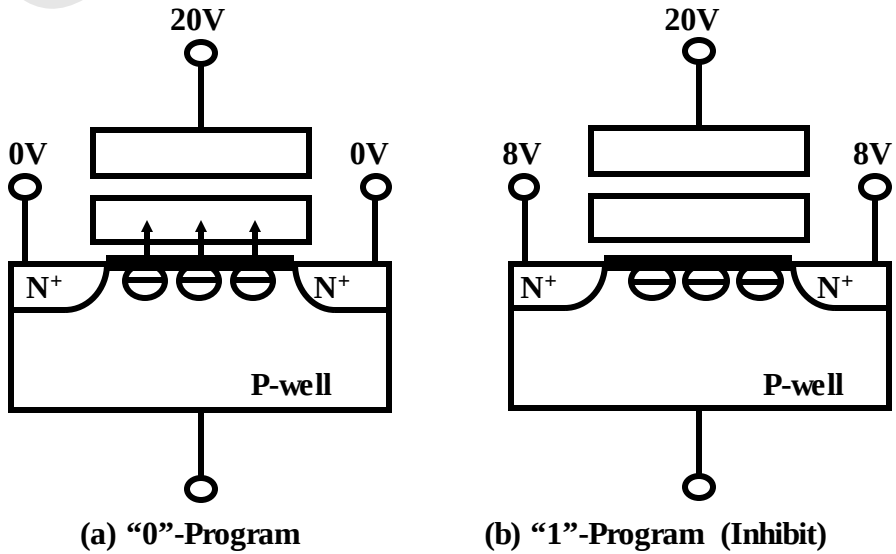
- A simple 1st order calculation can help us compute the Programming noise which is in reasonable agreement with data
  - During every programming step we put certain number of electrons on to the floating gate
  - Since $\Delta Vt = \Delta Vg$, and $\Delta Qfg = \Delta Vt * Cipd$, we can compute the amount of charge added.
  - Dividing this by charge of an electron we can calculate the number of electrons added to the FG on the average every pulse
  - The standard deviation is sqrt(mean) → ΔVt_sigma
- This is an approximate calculation. More exact calculations incorporating FN equation can be done as well.
- Higher cell capacitance of 100s cell leads to less Program noise

Micron/Intel Confidential

Inhibit

- We described the NAND Programming as FN tunneling from channel to the FG with the channel grounded and the WL high at 20V
- So, one question is how do we prevent ("inhibit") the other cells on the WL from programming up as well, since we would not want all cells on the WL to get programmed up
- The way we achieve inhibiting programming on all the other cells on the selected WL is called the "Program Inhibit"

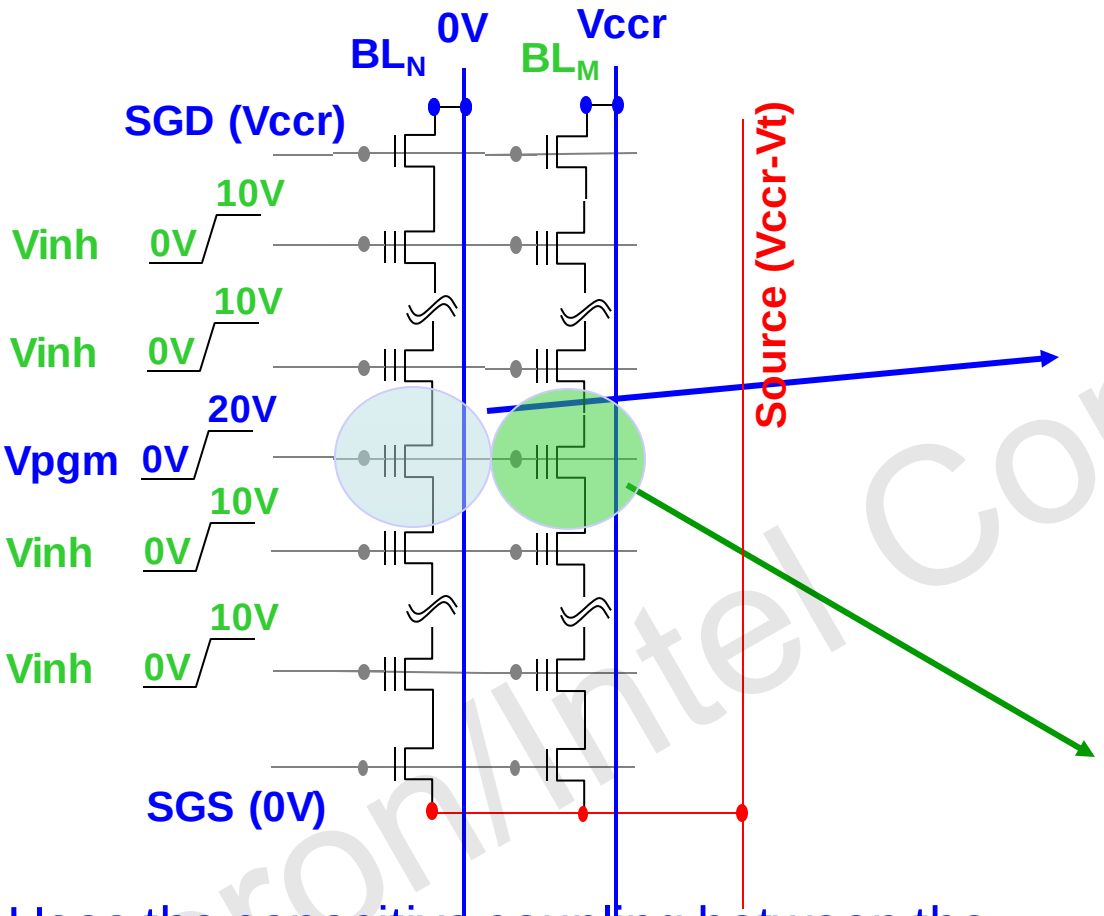
  - Program inhibited on other cells on the same wordline by floating those string and letting the channel and S/D "boost" up to ~6-8V, thus reducing the field across the Tunox

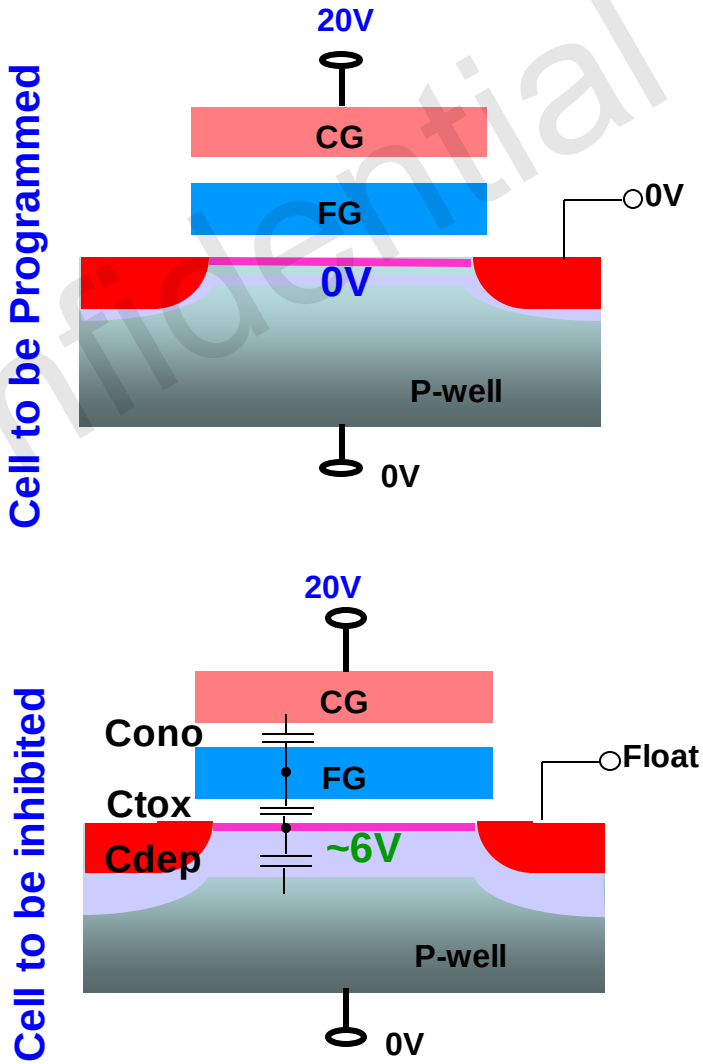


(a) "0"-Program (b) "1"-Program (Inhibit)

Micron/Intel Confidential

Program Inhibit (Self-Boosting)

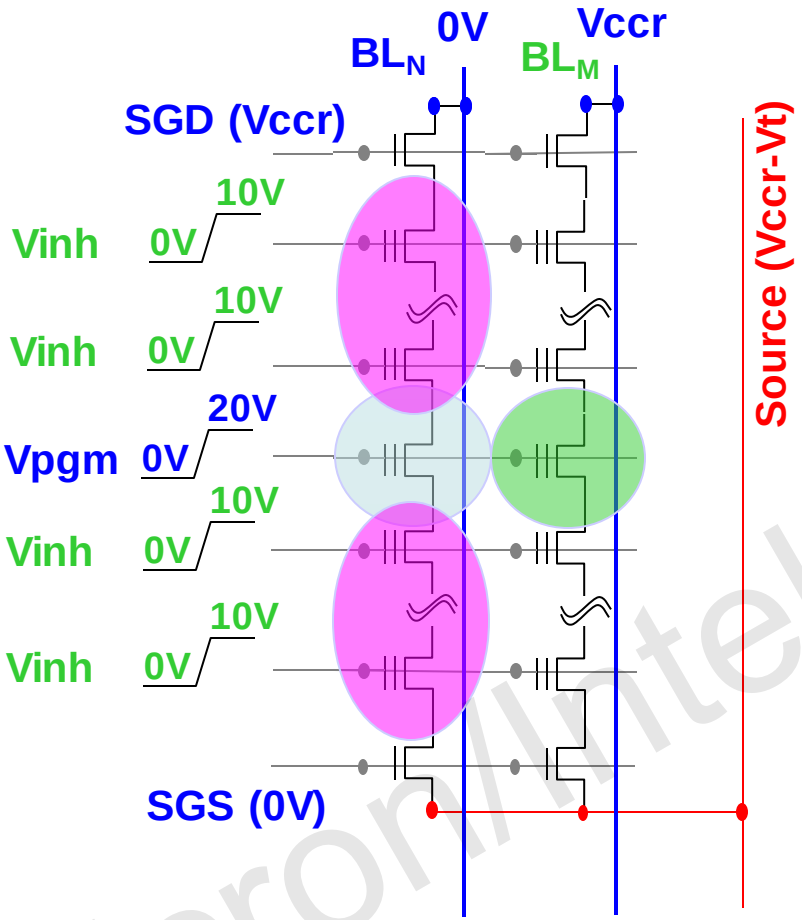




- Uses the capacitive coupling between the Control Gate and the Channel to boost the channel to high enough a voltage to inhibit programming

Micron/Intel Confidential

Program/Inhibit Disturb

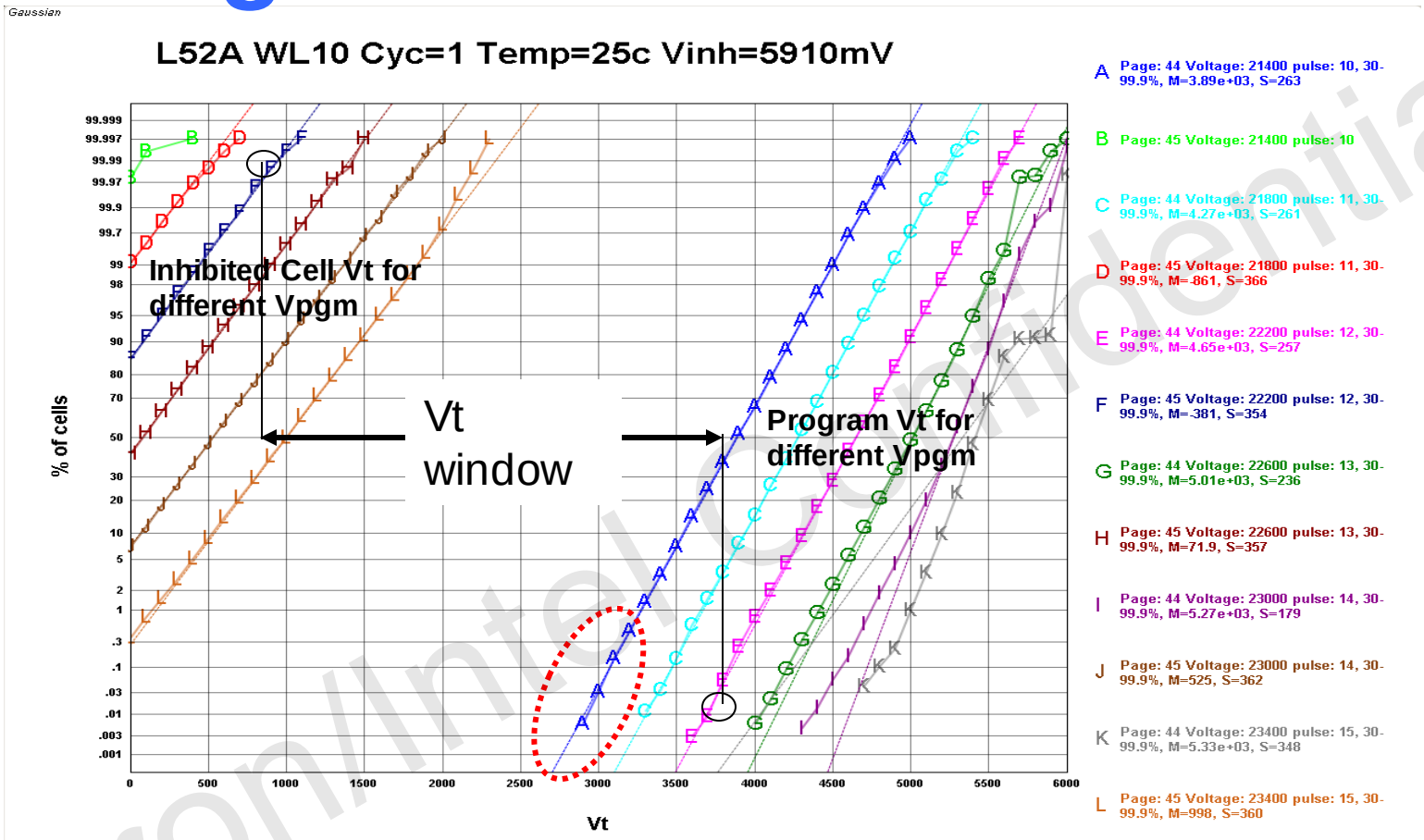


- There are two types of disturbs during programming:
  - Cells on the same WL see program disturb. If the Programming WL voltage is high and boosted channel voltage is not high enough these cells can get disturbed up
  - Cells on the selected bitline see Inhibit disturb. Here the channel is grounded and all the deselected WL are sitting at Inhibit gate voltage. If the Inhibit gate voltage is high, these cells can get disturbed up

- Program Disturb happens when the cells on that WL are programmed
- Inhibit disturb happens every time any other WL in that block is programmed

Micron/Intel Confidential

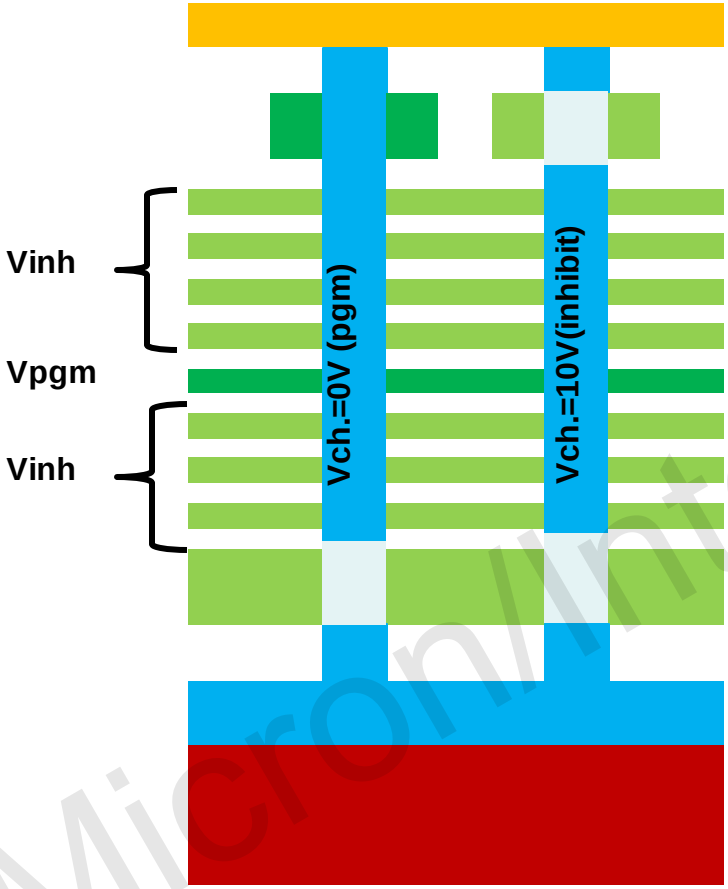
Program Disturb Window



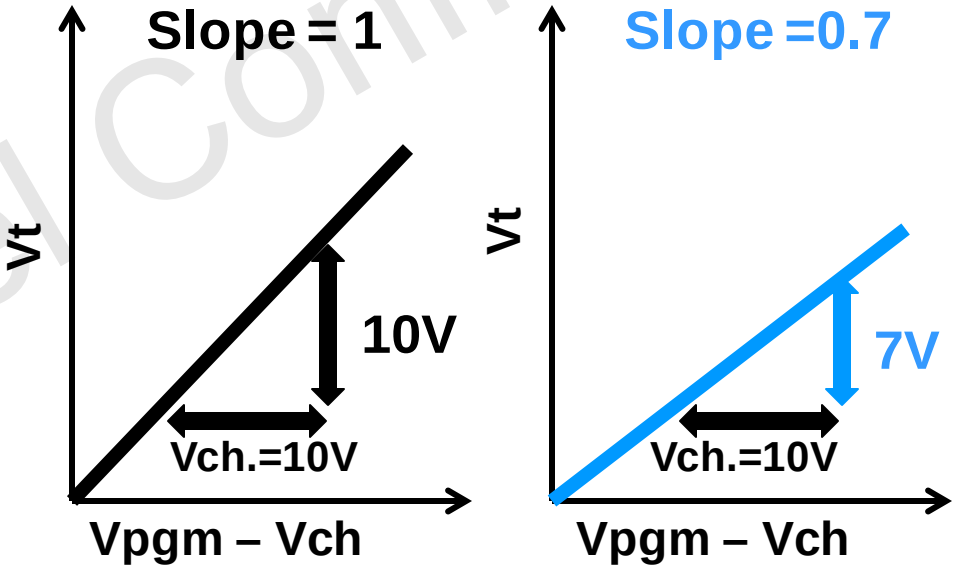
- To guarantee good Program Disturb Window, we need tail to tail window > PV7
  - Intrinsic Program Disturb window > PV7 – R1 + ~**7*PVS** + FG-FG GB

Micron/Intel Confidential

Programming slope requirement for program disturb window



Program disturb Vt window



Pgm slope <1 makes pgm disturb window more challenging.

Micron/Intel Confidential

Boosting Requirements

- If Vinh is too low, we do not get enough boosting and we have program disturb fail
- If Vinh is too high then we can have Inhibit disturb
  - The cells on the selected strings see the Inhibit disturb for the duration of 31 programming
  - For the high Vinh, in addition to the Inhibit disturb, we can also have high inhibit voltage induced disturbs – hot-e disturb, GIDL disturb, punchthrough disturb, etc.

➢ So, need to have adequate window between these two where we can avoid disturb

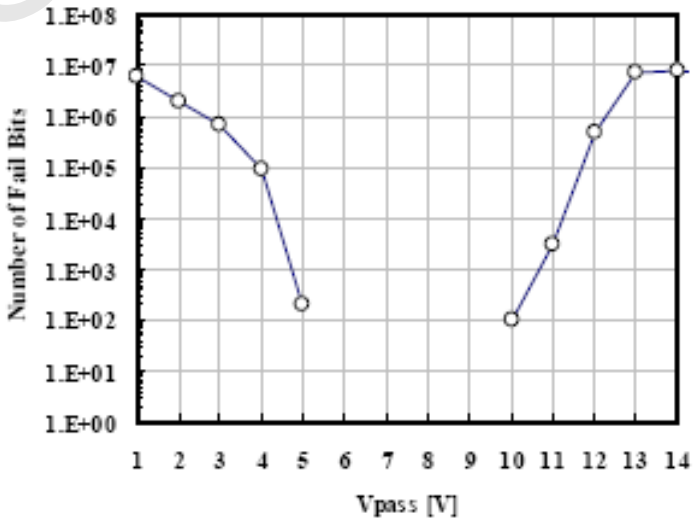
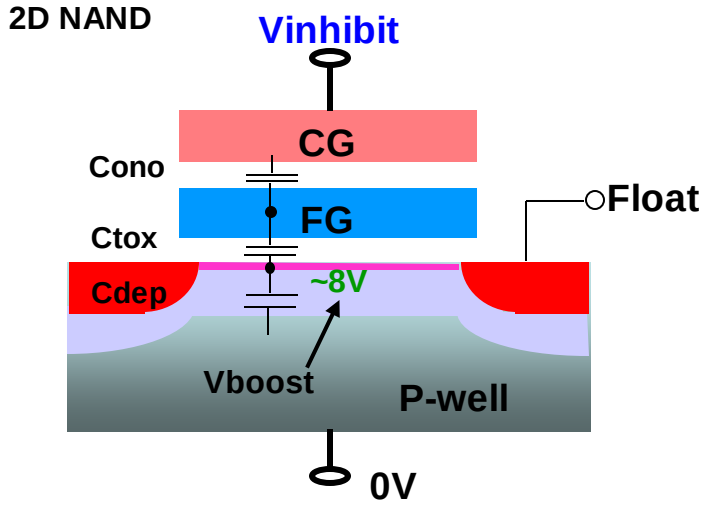
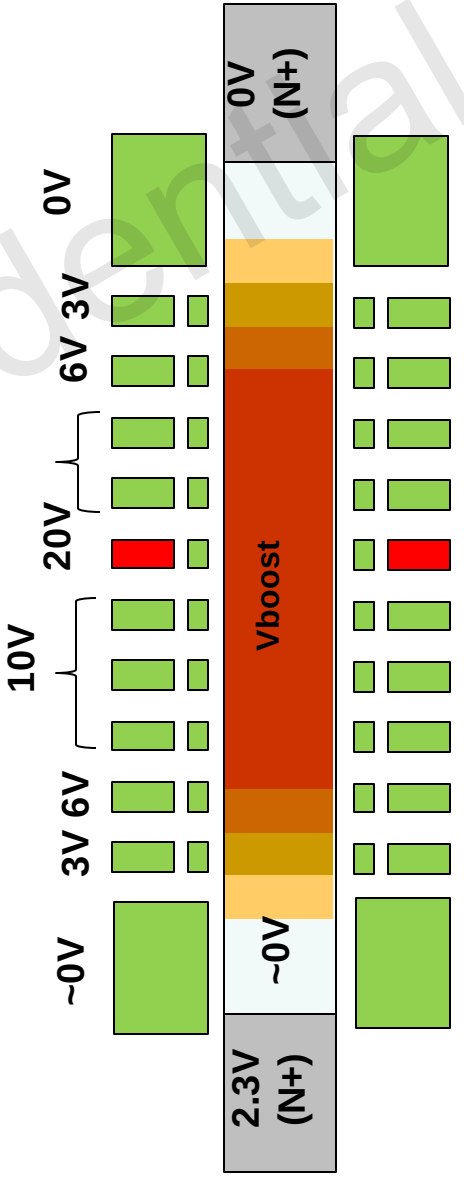


Fig. 10. Program disturbance of the NAND cell arrays with 70nm design rules as a function of the pass voltage.

Micron/Intel Confidential

2D vs 3D Program Disturb (PD)



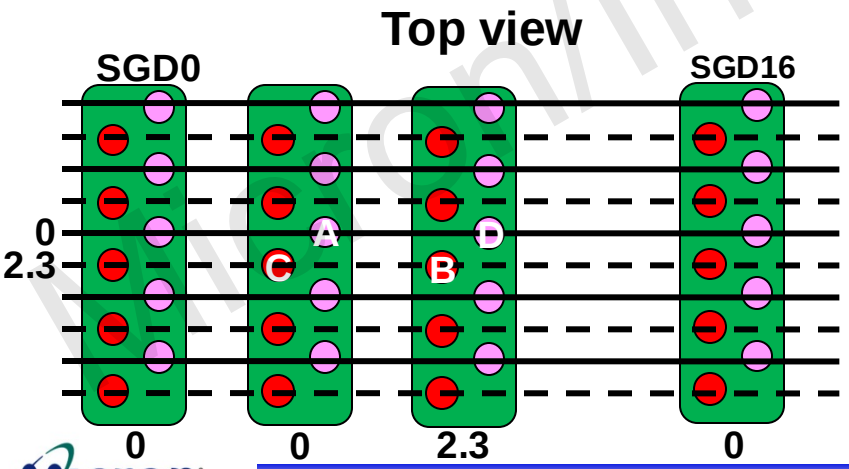


- Vboost created by capacitive divider – Vinh "fights" the depletion capacitance of the grounded P-well
- Coupling must also fight leakage mechanisms discharging Vboost
- 3D with Floating pillar has an advantage – no grounded P-well to fight → Boost ratio should be ~1
- But leakage still a risk, especially because of polysilicon pillar (grain boundaries and other defects)

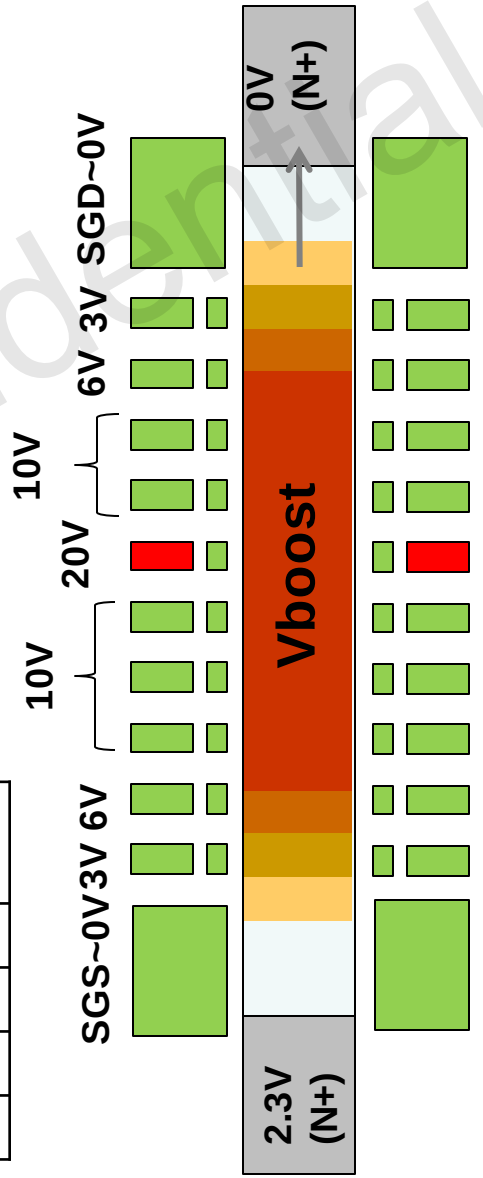
Micron/Intel Confidential

Inhibit and boost

- No grounded P-well to fight: higher boost coupling ratio
- Leakage through the SGS or SGD is a risk, so blocking schemes at block edge are important
- Each WL split among 16 sub-blocks (16 SGDs)
  - 1x conventional inhibit (high SGD, high BL)
  - 15x new inhibit (low SGD, low BL)
- "A" especially sensitive to SGD leakage, because BL=0
- More efficient boost can lead to hot-e issues: on inhibited subblocks, high electric fields can be present, leading to h+/e- pairs and hot-e injection
- Careful optimization of Vpass voltage and inhibit scheme needs to be achieved to prevent this issue

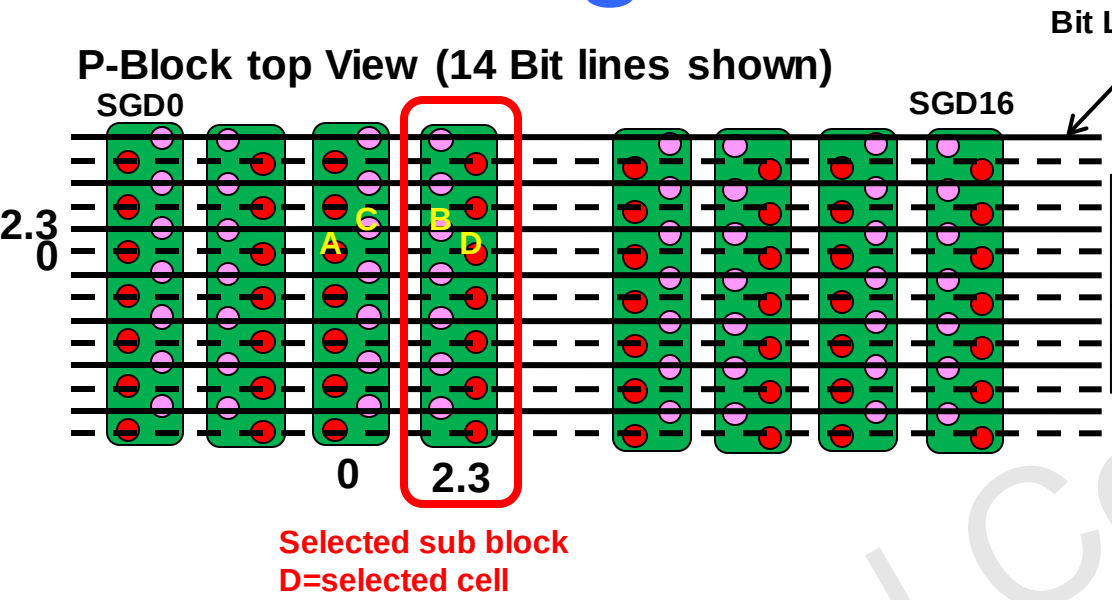


Pillar	Inhibit condition
A	New VSGD=0
B	Conventional
C	Stronger
D	Programmed



Micron Confidential

3D Program Disturb Biases



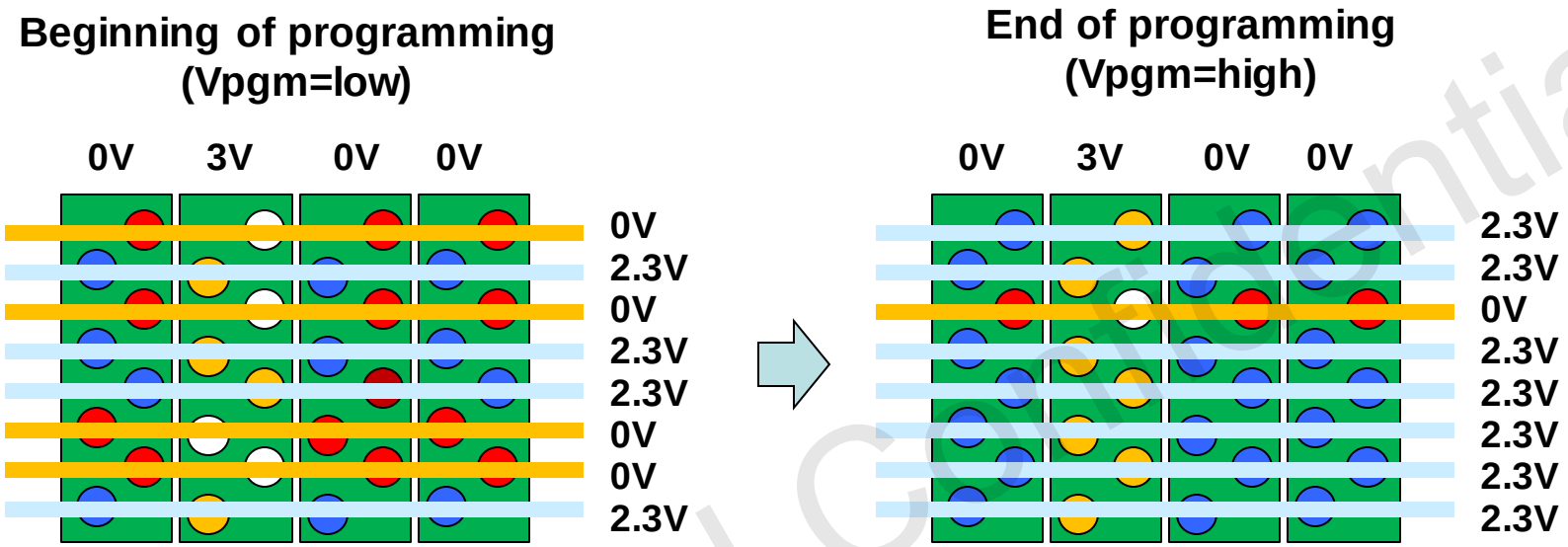
2D case

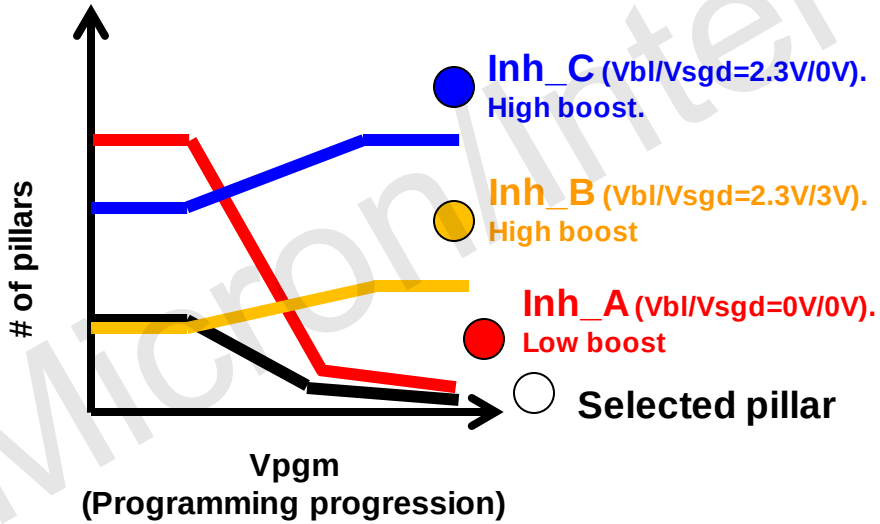
stress condition	WL	SGD	BL	pillar
A	vpgm	0	0	VboostA
B	vpgm	2.3	2.3	VboostB
C	vpgm	0	2.3	VboostC
D (programmed)	vpgm	2.3	0	0

- 3D advantage is good boost ratio (surround gate and floating pillar)
- But 3D must inhibit cells on selected WL and deselected SGD (top, cases "A" & "C"). "A" worse due to the low BL voltage.
- Condition "C" expected to have good boost (low SGD leakage), but has higher requirement because final programming pulses are higher voltage

Micron Intel Confidential

Boosting Pattern Progression

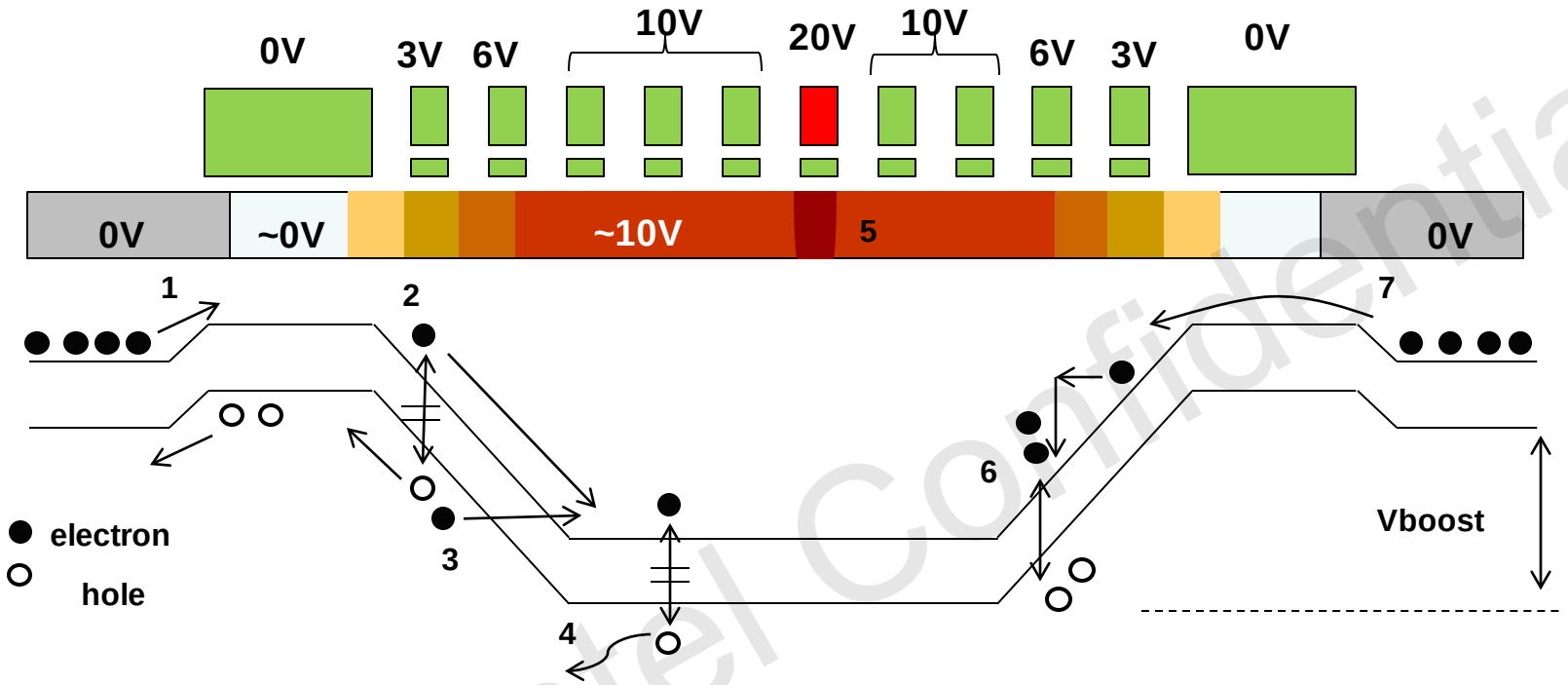




- As programming progresses, numbers of inh_A pillars decreases because more BLs are inhibited.
- Therefore boosting requirement is less for inh_A condition.
- Any of Inh_A,B and C can be a boosting limiter.

Micron/Intel Confidential

Boost loss mechanisms



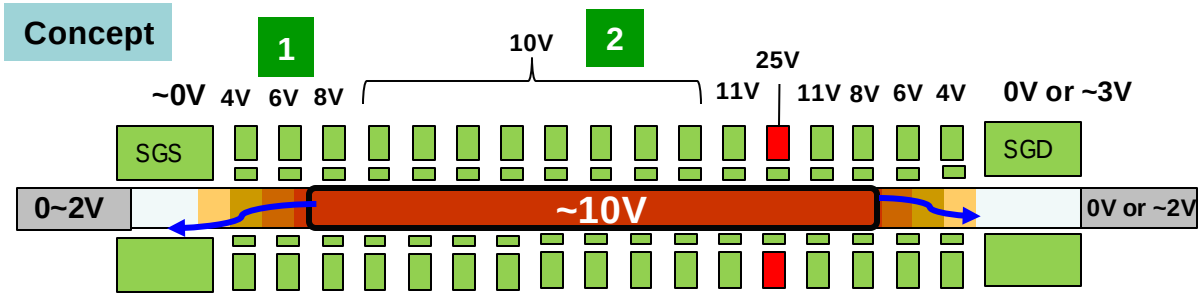
	Mechanism	Description
1	Bipolar gain	Holes under SG are injected into N+ then electrons are injected back.
2	e/h at edge	Thermally generated electrons are injected into channel due to E-field.
3	BTBT	Band-to-band tunneling at edge high E-field.
4	Hole diffusion	Thermally generated holes are diffused into edge before recombined with electrons..
5	Local boost hot e/h	Ch. Under Vpgm WL are locally boosted and creases hot e/h injection.
6	Impact ionization	Electrons are accelerated by edge E-field and create e/h pairs.
7	SG S/D leak	Source-drain leakage of SG transistor.

Electron injection into or hole loss from the boosting area cause boosting loss.

Micron/Intel Confidential

Self boost and Hot-e

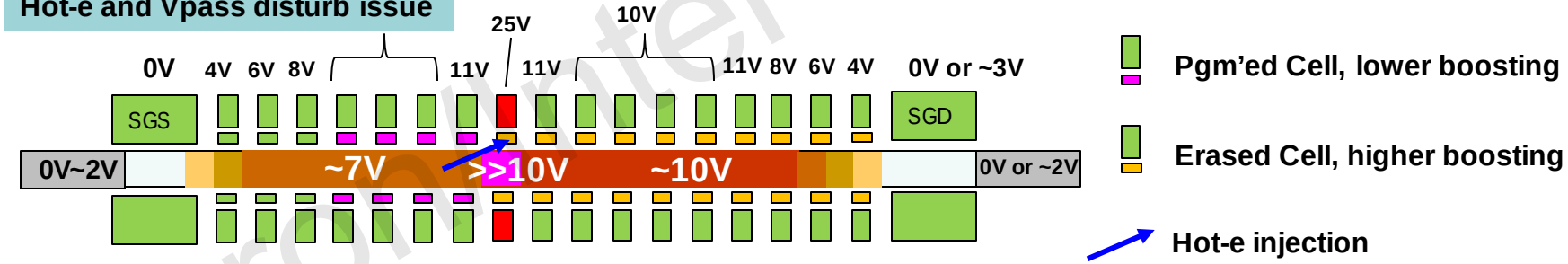
Concept



1. [Edge scheme] Edge WLs has gradual biasing to minimize the pillar E-field at the edge. This reduces leakage at the edge, pulling down the boosting.

2. [Self-boost scheme] ~ All WLs are biased to ~10V, making the boosting area (capacitance) large. This reduces the boosting loss with a given leakage at the edge. dV=dQ/C.

Hot-e and Vpass disturb issue



Hot-e injection disturb due to the lateral E-field.

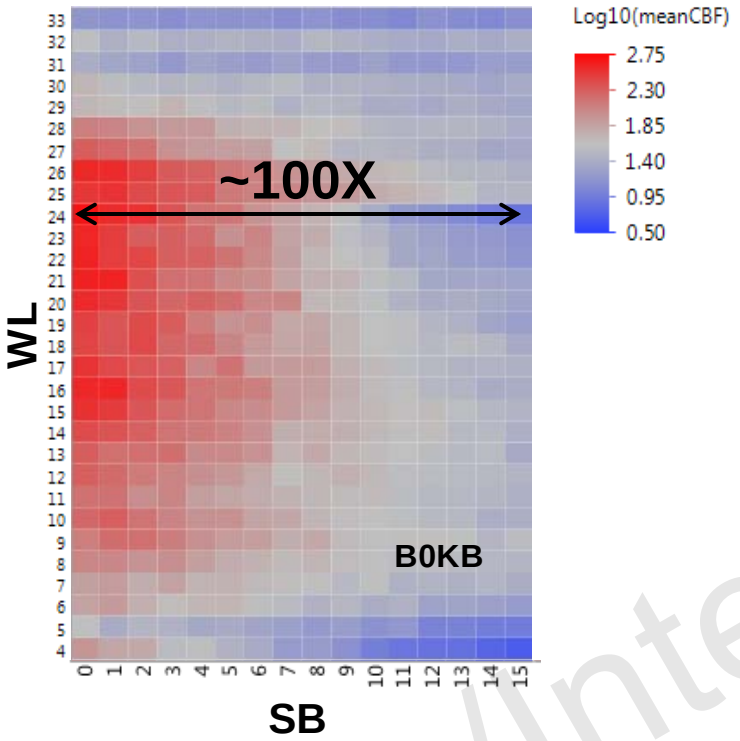
1. Data pattern is asymmetric, resulting in asymmetric boost potential.
2. Pillar under Vpgm is boosted higher due to WL-Pillar coupling.

→ The lateral E-field along the pillar create hot-e generation and injection to the cell.

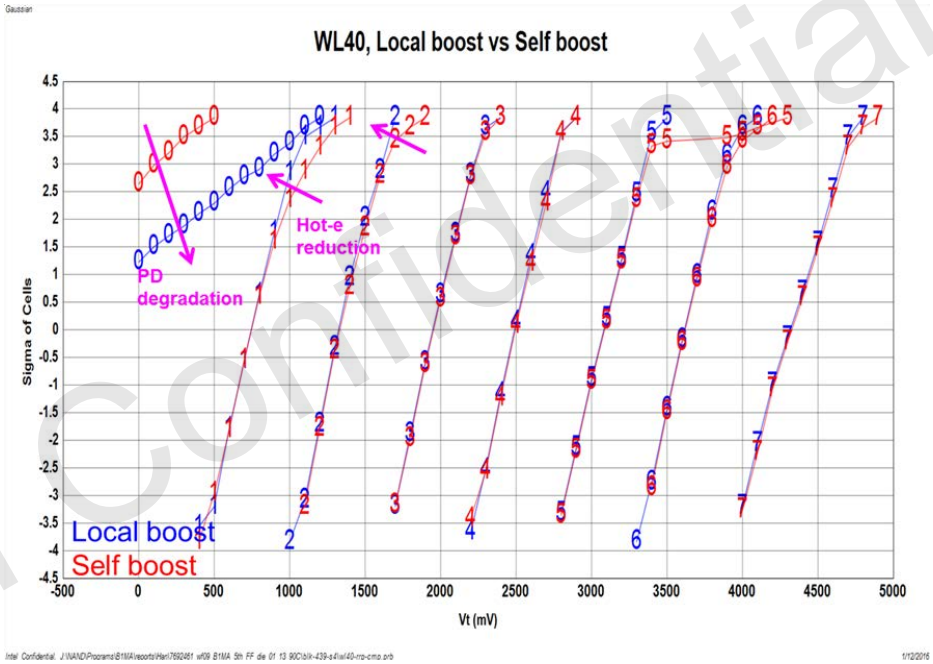
3. Also, a high Vpass is applied to all WLs, degrading Vpass disturb.

Micron Intel Confidential

PDOP: Hot-e impact on E2



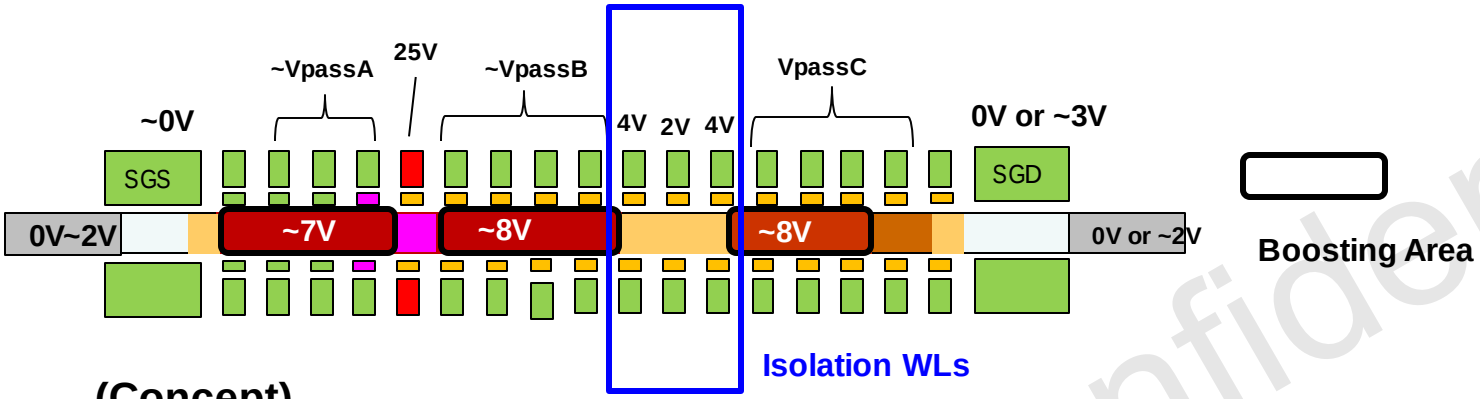
FN vs hot-e trade-off on B1MA



- E2 is worse on center WLs (~WL24), and on the first subblock
- A clear trend is present with subblock (SB 15 on WL24 is ~100x better than SB0): the mechanism is impacting the subblocks that are inhibited after the setup
- The issue is caused by high Vpgm needed to bring center WLs to L7, leading to hot-e generation on already programmed subblocks
- Worst at 30C, uncycled, and it's improved by either cycling or high temperature

Micron/Intel Confidential

Local Self-boost



(Concept)

Segmenting the boosting area by putting isolation WLs.
Source side boosting scheme is shown above.
Based on the location of the iso. WLs,
Other variations such has Drain side boosting (iso. WLs are at SGS side) and Local self-boost (iso. WLs are at both sides) exist.

(Benefit)

Boost potential can be optimized by each boosting area depending on the data pattern and other circumstances. Vpass can be lowered in general, Improving Vpass disturb.

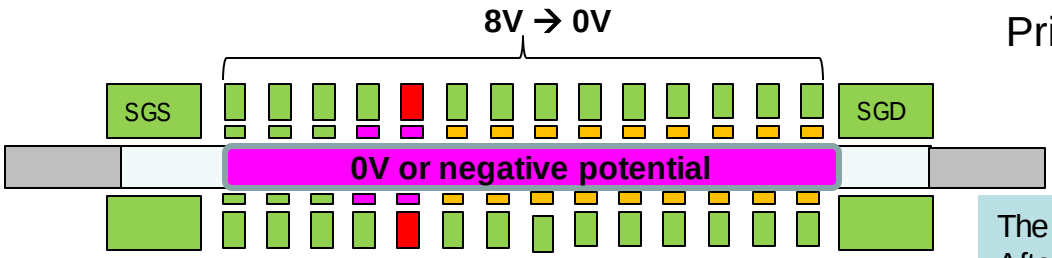
(Limitation)

Another lateral E-field can be generated around the iso. WLs.
The small boosting area (capacitance) can lose the potential easily.

Micron/Intel Confidential

Seeding

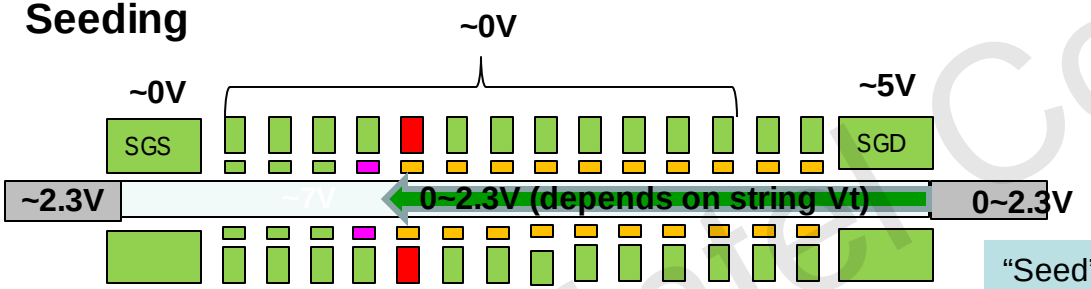
Before programming pulse
(after the previous pgm/verify loop)



"Seeding" sets the initial pillar potential Prior to boosting.

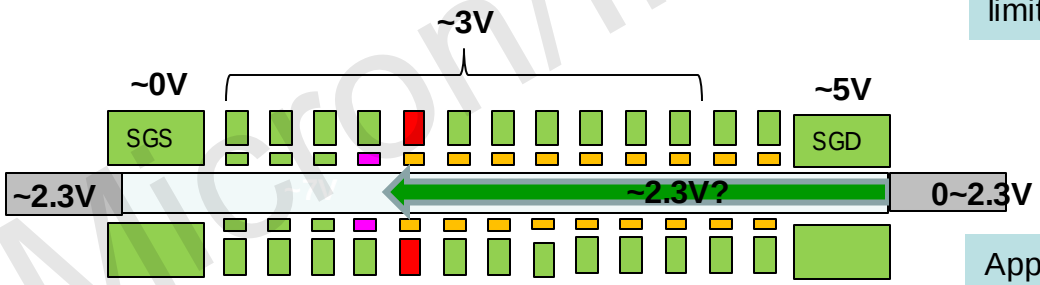
The pillar can be negative potential
After the previous pgm/PV, not desirable.

Seeding



"Seed" resets the pillar potential to 0V or
Higher depending on the string Vt.
Though all cells are the drain side is erased, the space Vt limits the seed capability. (Similar to All WL EV)

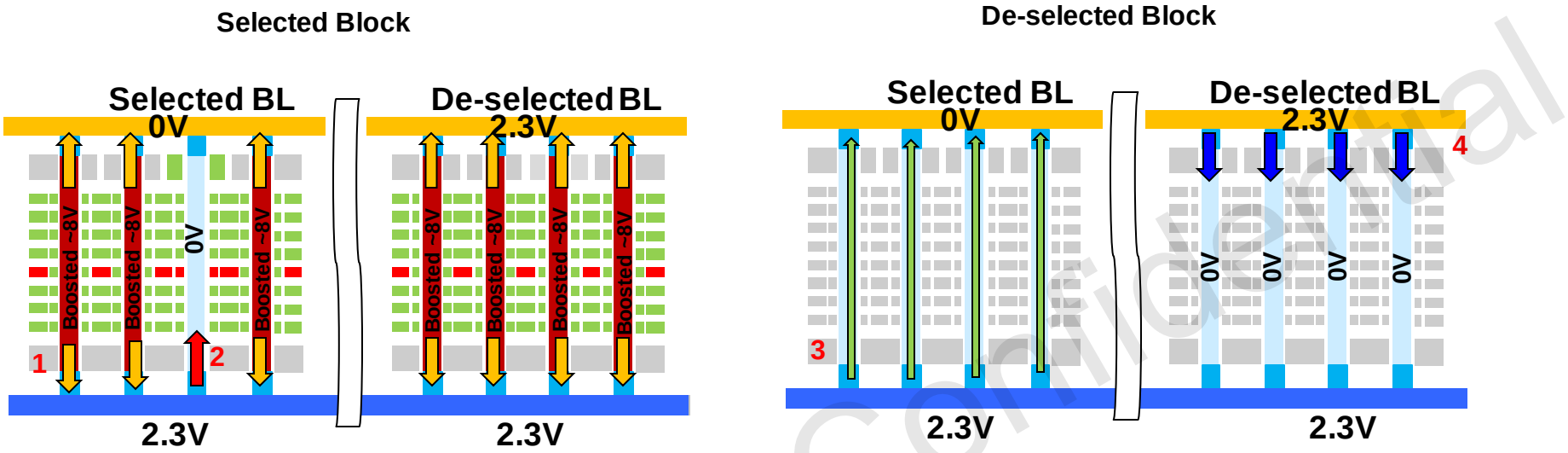
Seeding with high WL_seed



Applying positive WL bias for higher seeding often turns out the net loss. Because you lose the boosting later.
+1V WL at seed provides;
≤1V seed gain vs.
~1V boost loss (WL-Pillar coupling is strong in 3D NAND)

Micron/Intel Confidential

SG leakage during programming



1. SGS or SGD leakage pulling down boosting. Program disturb.
2. SGS leakage pulling up the channel. Slow to programming.
3. SGS leakage pulling up the channel then pulling up BL. Slow to programming.
4. SGD leakage pulling down BL. Program disturb.

(*) SG leakage includes both source/drain punch through and GIDL.

1,2 and 4 are common modes with 2D NAND.
3D NAND has increased risk due to more leakage caused by poly Si channel.

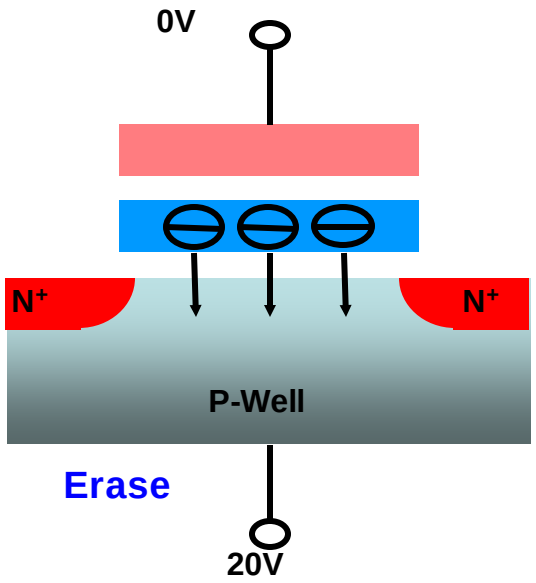
Micron/Intel Confidential

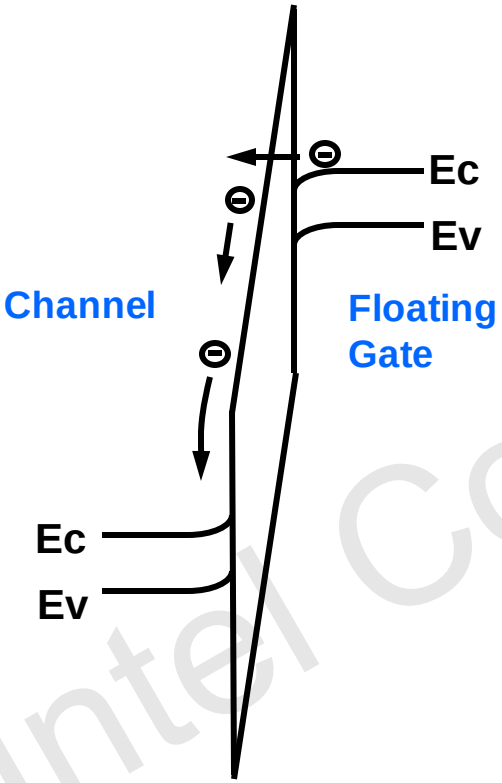
Erase

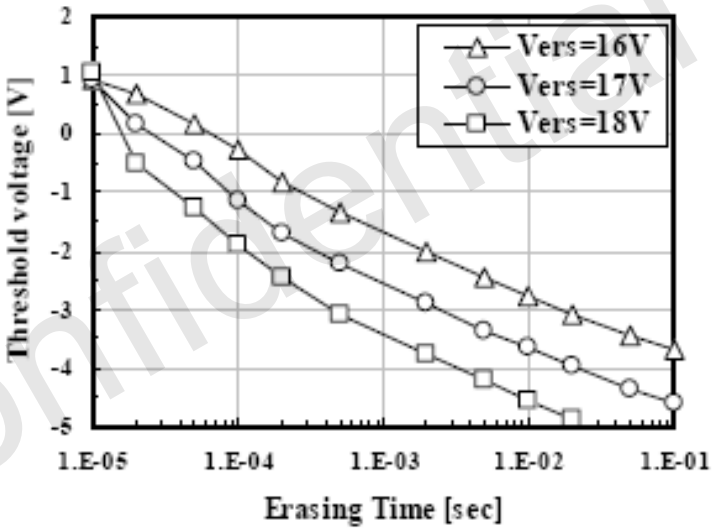
- FN Tunnel Erase
- Erase Inhibit
- Erase slope
- Erase Vt sigma
- 3D NAND Erase
- SG GIDL Requirements
- Erase Verify

Micron/Intel Confidential

NAND Flash Erase - FN Tunneling



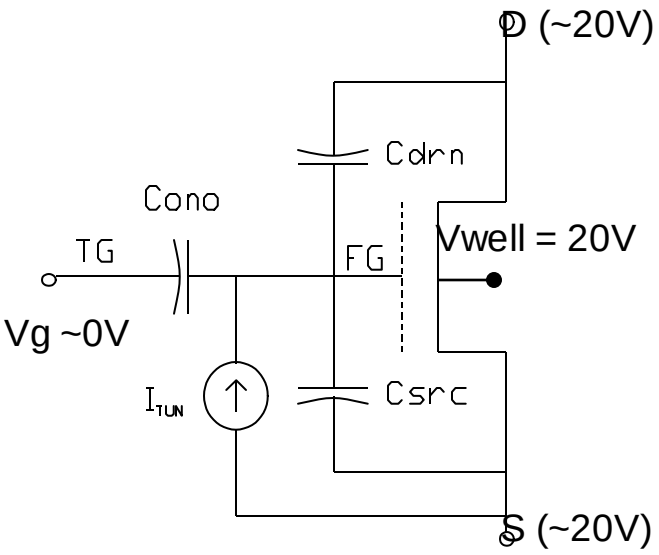




- Tunnel Erase using Channel Erase with high positive voltage applied to the Well with gate grounded.
- Erase time ~1ms
- Erase Physics is essentially same as Program Physics, except now we tunnel from FG into the P-Well

Micron/Intel Confidential

Erase V_T vs. Time



- Equiv. ckt shows electron tunneling off the Floating Gate by I_{TUN}

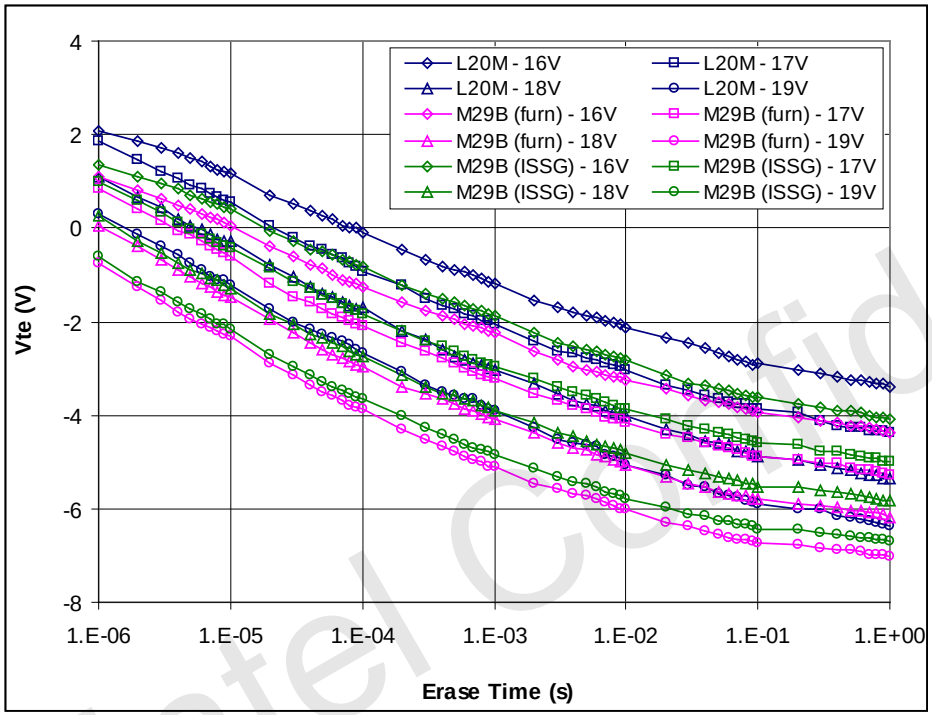
$$V_{FG} = V_{CG} \cdot GCR + \frac{Q_{FG}}{C_{TOT}} \qquad V_T = \frac{V_{TFG}}{GCR} - \frac{Q_{FG}}{C_{ONO}}$$

- At the beginning of Erase, V_{FG} is low and I_{TUN} is large, so high tunnel current and erasing is fast

- As number of electrons on FG decreases the FG voltage increases (by ΔQfg/Ctot) decreasing I_{TUN} as given by the FN Equation
- Each ~0.8V increase in V_{FG} drops I_{TUN} by ~10x
  - Erasing rate drops by 10x, 100x, etc as the floating gate gets charged by .8V, 1.6V, … From the control gate perspective, the erase rate drops by 10x, 100x, etc as the Cell Vt decreases by ~1.4V, 2.8V, …
    - ➢ Makes Erase Vt vary as a "log time"

Micron/Intel Confidential

Erase



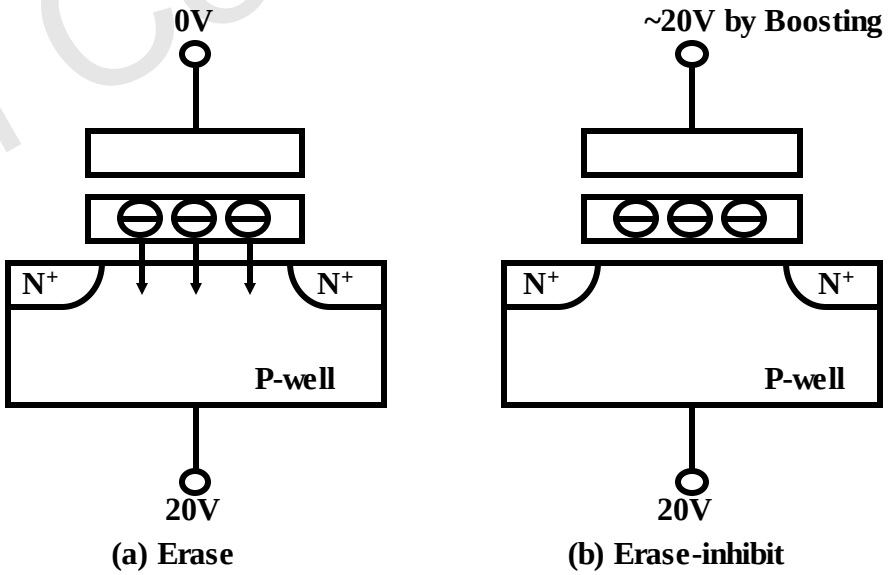
- Log Time behavior of the Erase Vt can be seen above
- A rough rule of thumb is that the cell Vt changes by 1.0V/dec under erasing conditions (~100us-1ms).
- Actual NAND Product uses a +ve Well voltage with grounded gate to erase the cell. However at Param (ETEST), often a –ve voltage is applied to the WL and Well grounded. Electrically both are equivalent

Micron/Intel Confidential

Erase Inhibit (2D NAND)

- We described the NAND Erase as FN tunneling from FG to the channel WL grounded and the P-Well high at 20V. But the P-Well is common to all the blocks
- So, the question is how do we prevent ("inhibit") all the other blocks from erasing as well, while erasing only the selected block.
- The way we achieve inhibiting Erasing on all the other blocks is called the "Erase Inhibit"

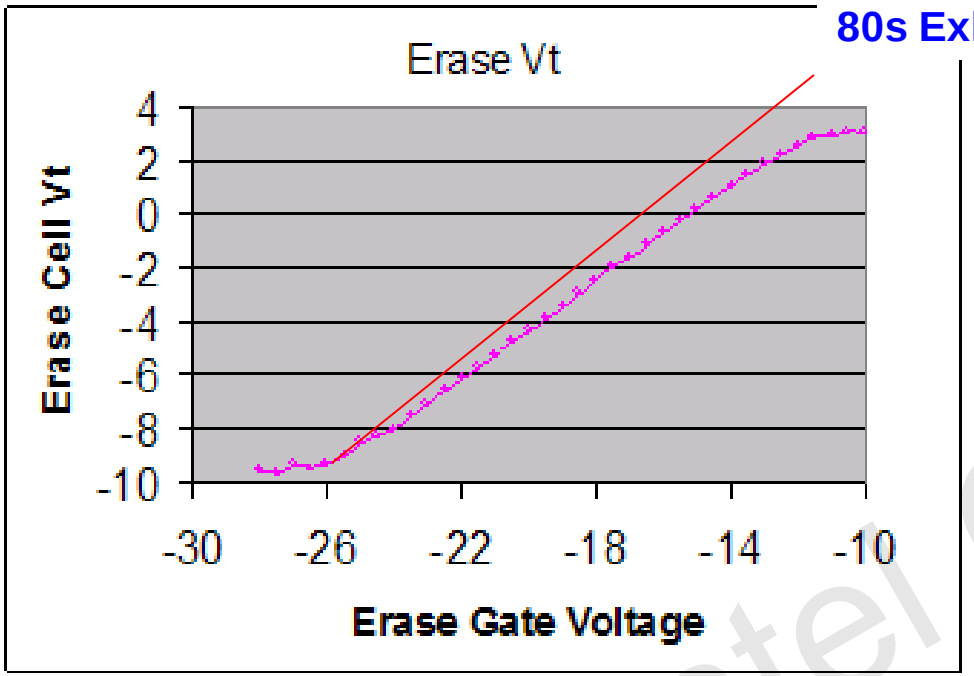
- P-Well goes high for all blocks
- WL on Selected blocks grounded, so these get erased
- Erase inhibited in deselected blocks by floating the WL & allowing them to get "boosted" up along with the Well

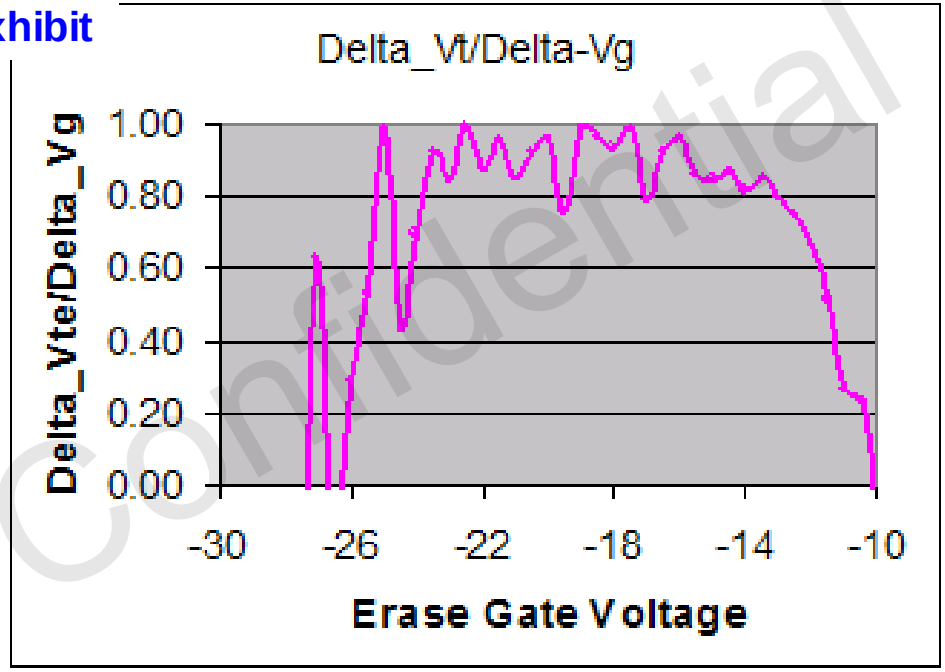


Micron/Intel Confidential

Erase Vt vs Vg (Vw)

80s Exhibit



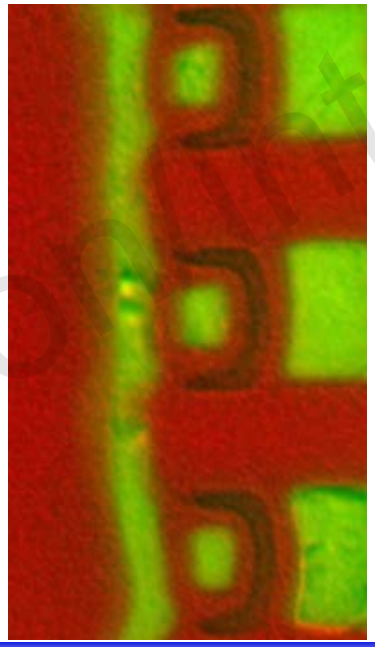


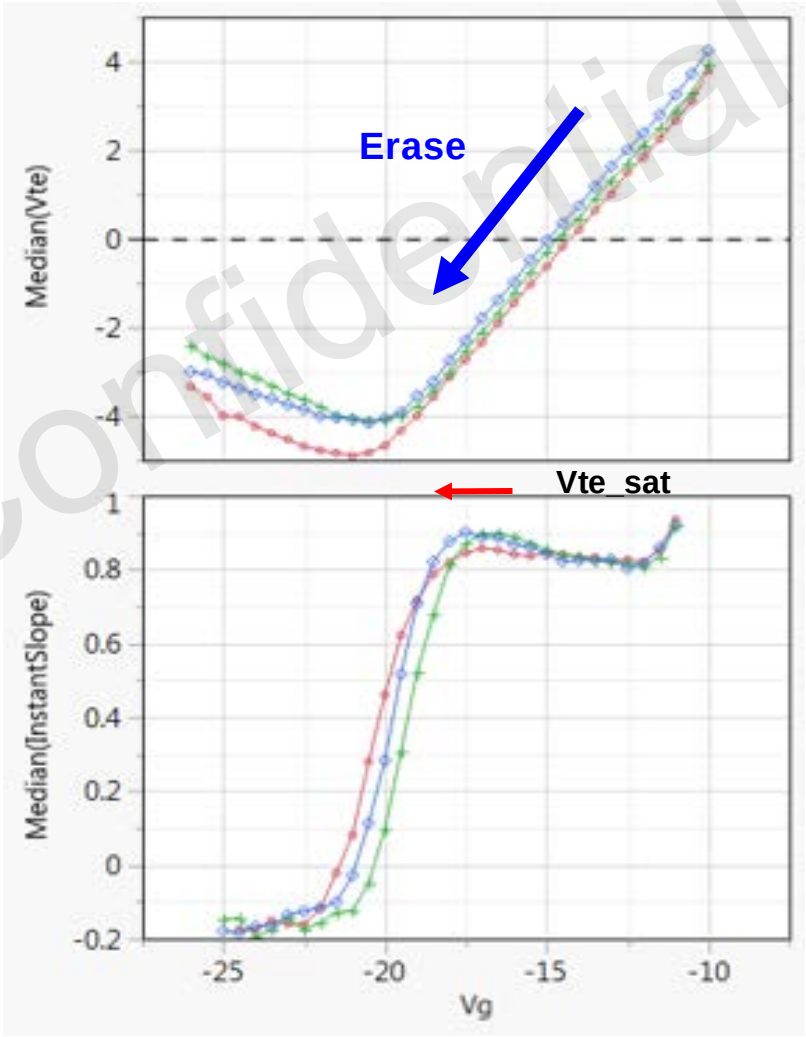
- For a given Erase time, the Erase Vt tracks the erase voltage. VwVt (= Vw + Vte) or VgVt (= Vg-Vt) is a good cell erase metric.
- VgVt or VwVt is primarily determined by the Cell UV_Vt, FN Tunneling coefficients, Tunnel-oxide and ONO thicknesses, GCR, etc.
- We expect ΔVte/ΔVg to be ~1.0. However, deviation from 1.0 is seen due to GCR difference between Erase and Read conditions as well as Poly depletion effects.
- At very low erase Vt the Vt saturates because of electrons tunneling from the CG through the ONO into the FG balances the electrons tunneling out.

Micron/Intel Confidential

Erase Vt vs Vg (Vw) (100s)

- 100s has shallower erase saturation Vt
- Some of the factors for this are:
  - Lower GCR
  - N+ doped Poly
  - Sharper corner for CG
  - Flank Nitride Shunting the electrical distance to channel



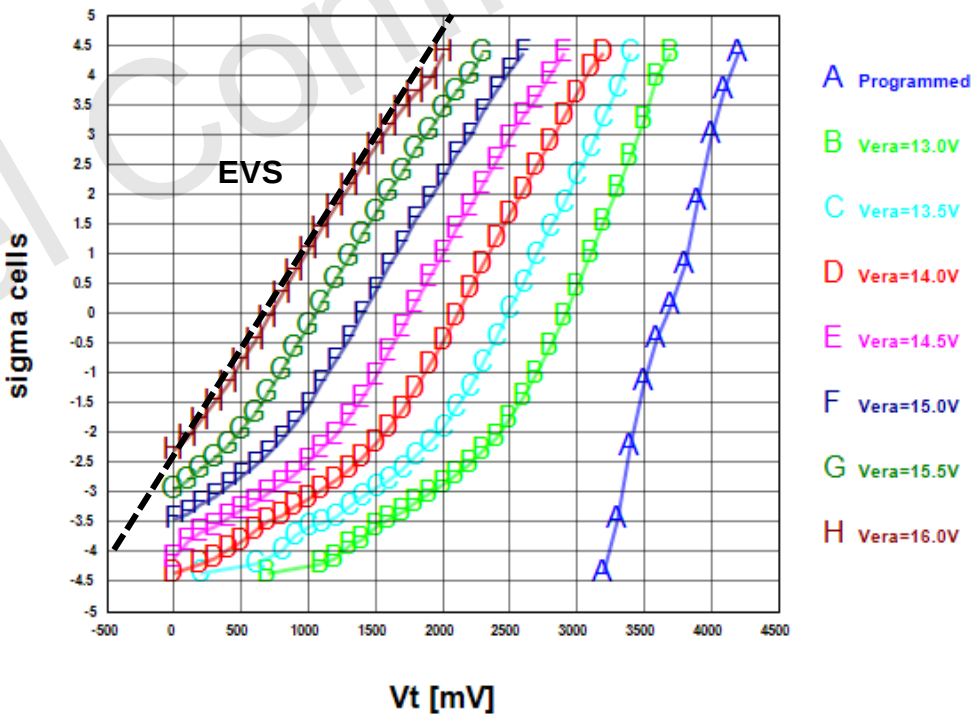


Micron Intel Confidential

Erase Vt Sigma

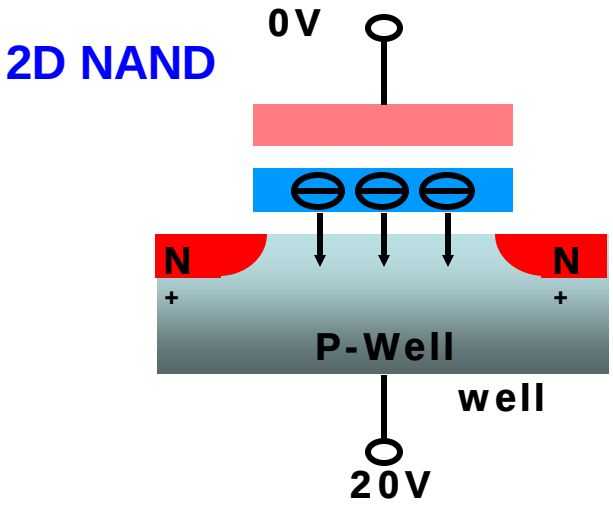
Similar to the Program Vt Distribution within the array, we have erase Vt distribution as well, since all the cells in the array do not erase to the same Vt due to cell to cell variations

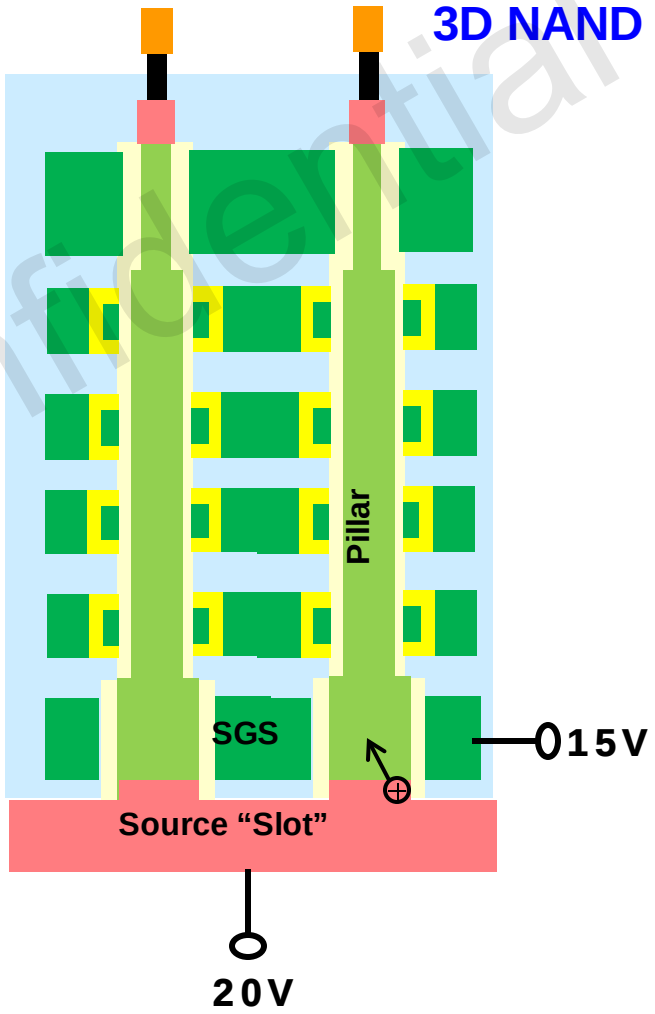
- The erase distribution comes about from differing rates of erase between different cells – (Coupling ratio variations, Z and L variations, UV_Vt variations, FN-Tunneling co-efficient variations, Pillar Voltage variation…)
- During erase we need to make sure that the slowest to erase cell is erased well below R1 Vt and it will stay below R1 despite the subsequent FG-FG coupling from neighboring cells



Micron/Intel Confidential

3D NAND Erase

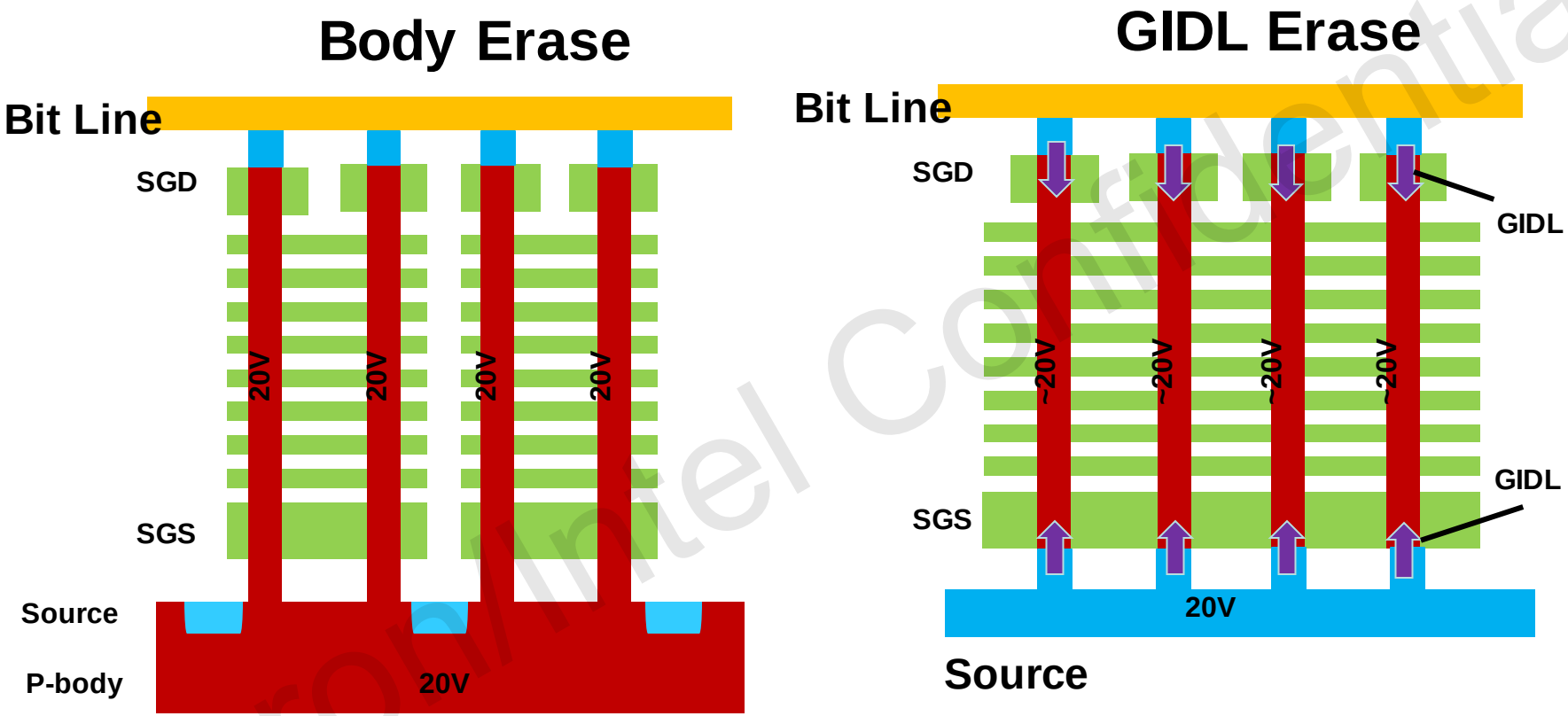




- 2D NAND is erased by applying ~20V to the P-Well/Channel
- 3D NAND requires the voltage being applied to the source slot and transferred to the pillar
- Requires injecting holes into pillar, but n^+ slot cannot do this.
- Solution: Use SGS GIDL leakage (right)
- SGS reliability from hot-carrier damage is a key concern for 100s 3D NAND.

Micron/Intel Confidential

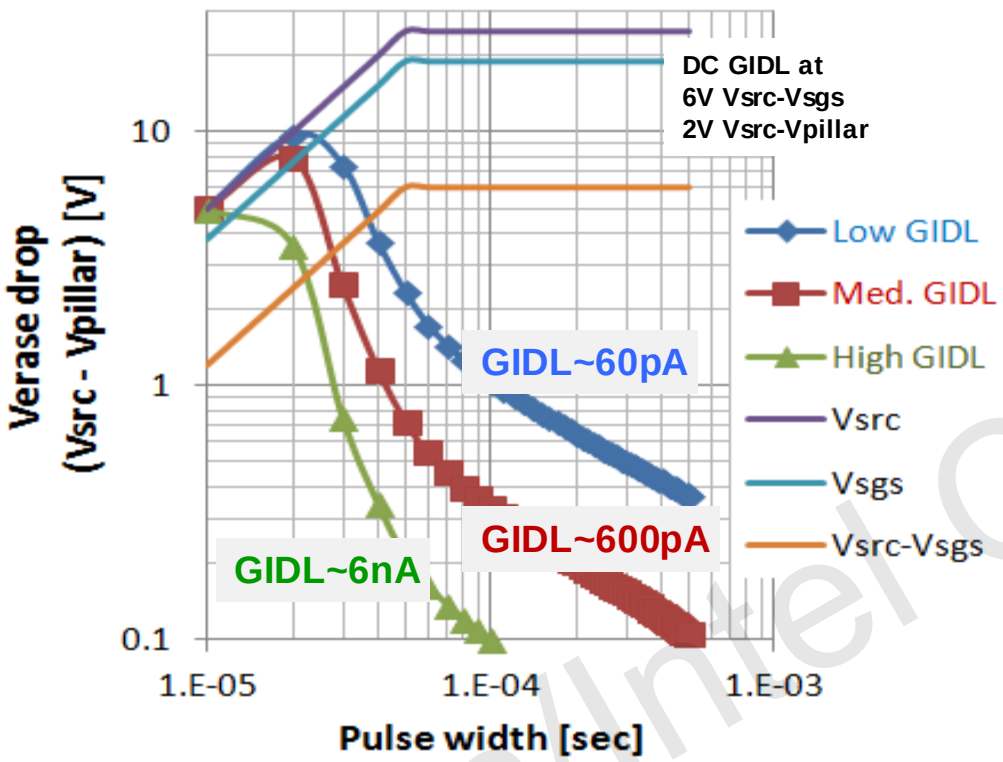
Body erase and GIDL erase

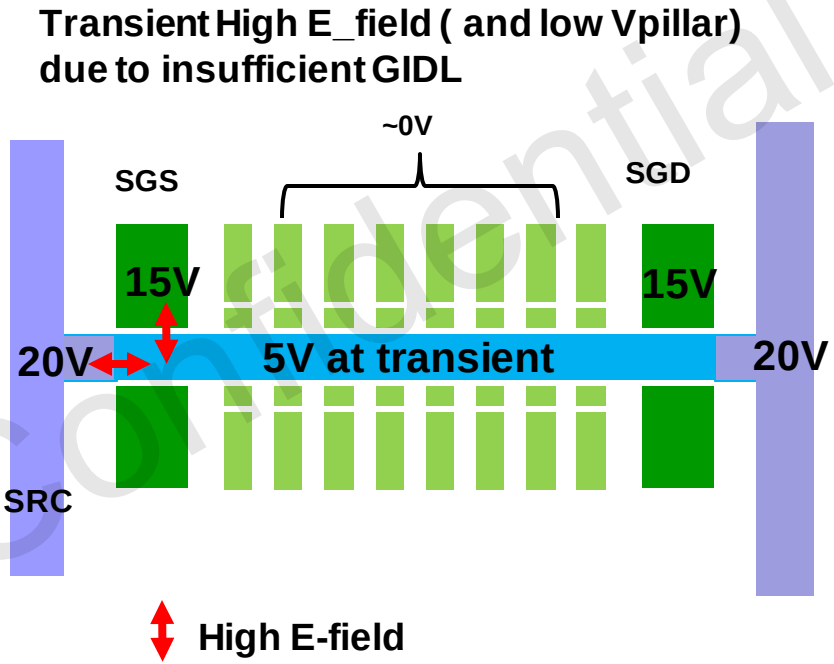


- Body Erase directly biases the pillars at erase voltage.
  Pillar to P-body interface resistance control is challenging
- GIDL Erase requires sufficient GIDL current to charge the pillars.
- 100s/110s do the GIDL erase.

Micron Intel Confidential

Erase GIDL current requirement

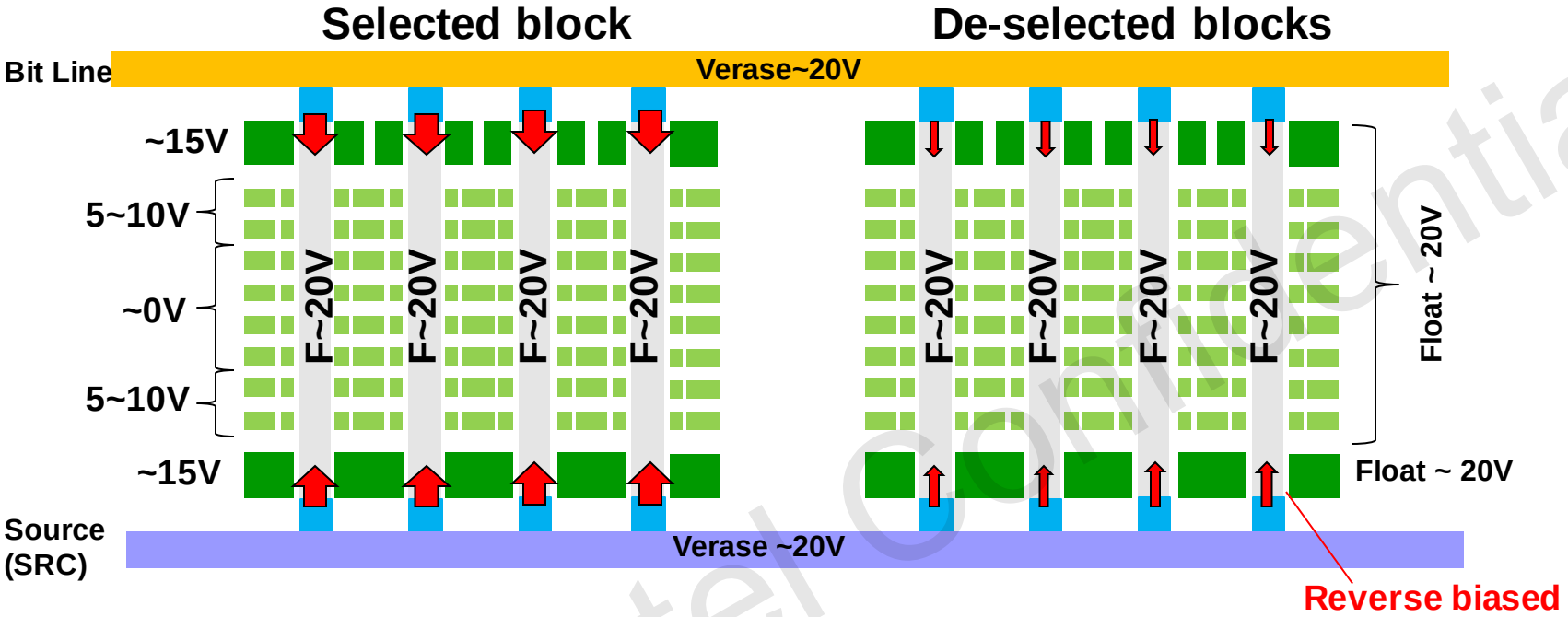




- SGS GIDL current has to be large enough to charge pillars within the erase ramp time.
- Insufficient GIDL current results in large voltage drop between Vsrc and Vpillar. Transient Hot-e and SG ox stress, High Verase and large erase property variation (EVS).

Micron/Intel Confidential

3D NAND Erase Bias Condition

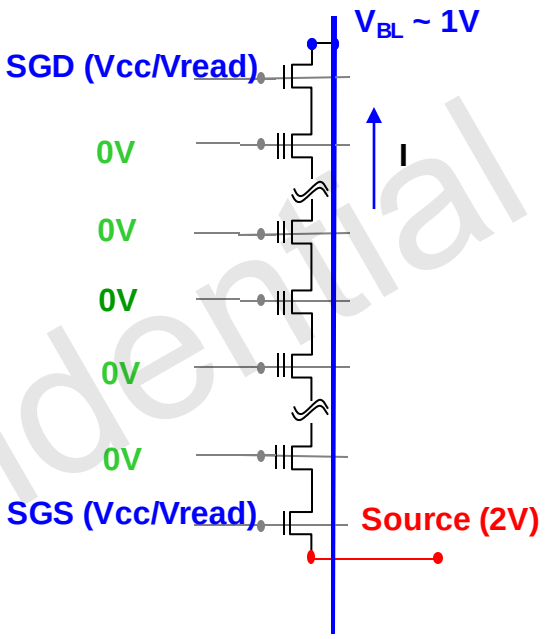


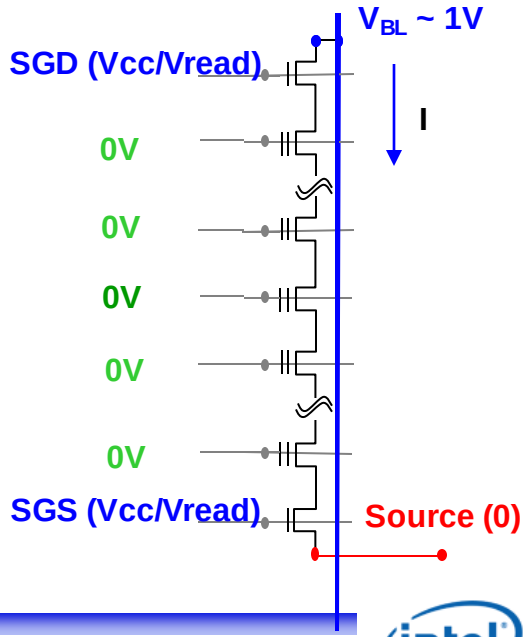
- SRC/BL is biased at erase voltage of ~20-25V
- SGS/SGD is biased ~5V below SRC (negative Vgs) on Selected Block to generate GIDL current to charge up the pillar.
- WLs of the Selected block are grounded to erase them
- SGS/SGD on Deselected block is floated
- WLs of the Deselected Blocks are floated which makes them go up to the same potential as the Deselected block Pillars → Inhibits erase

Micron/Intel Confidential

Erase Verify (EV) (2D)

- Historically (until 80s) Source side verify for erase verify (Assure cell Vt << 0V)
  - The source is biased to ~2V. WLs at ~0V.
  - The bitlines are floating. Due to the current from the source, the bitlines float up.
  - To pass the verify, the Vbl needs to go >1V. For the BL voltage to float up above 1V, all the cells in the string need to have a "Vt" < -1V.
  - Adjacent WL bias is different between EV (~0) and Read (VpassR~7V). This allows EV to verify negative Vt by the amount of (VpassR_n+/-1 x WL-FG coupling x 2sides).
  - The body bias on the cell provides EV to verify negative Vt by additional amount
- Conventional Drain side verify can be used as well and rely on WL-FG coupling to achieve Read Vt << 0V.

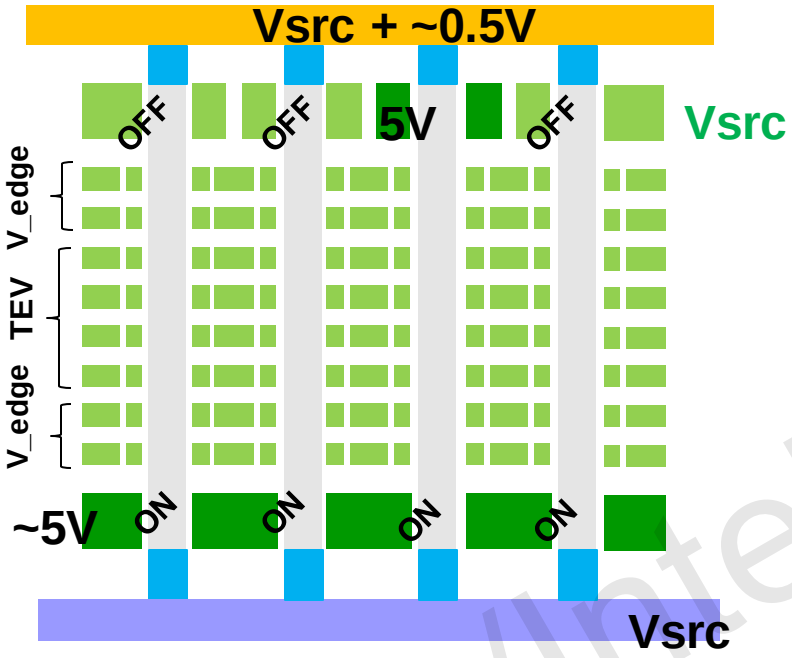


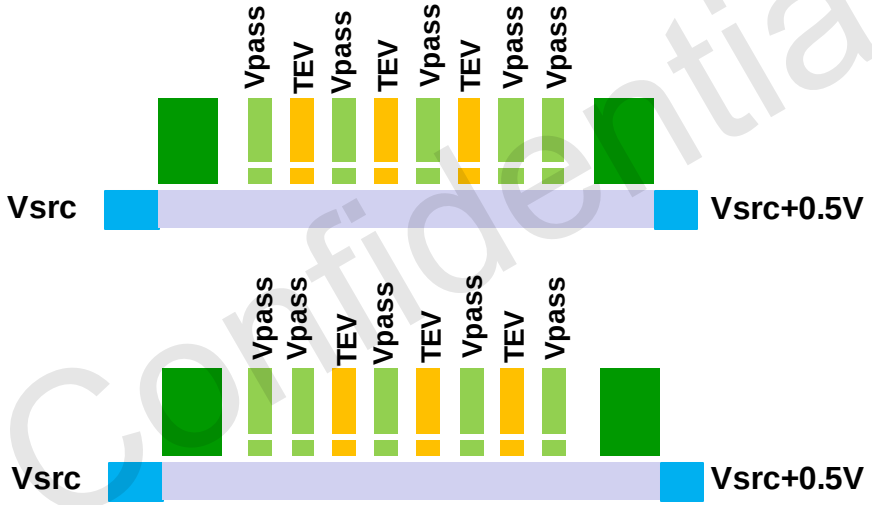


Micron/Intel Confidential

3D NAND Erase Verify

Erase Verify



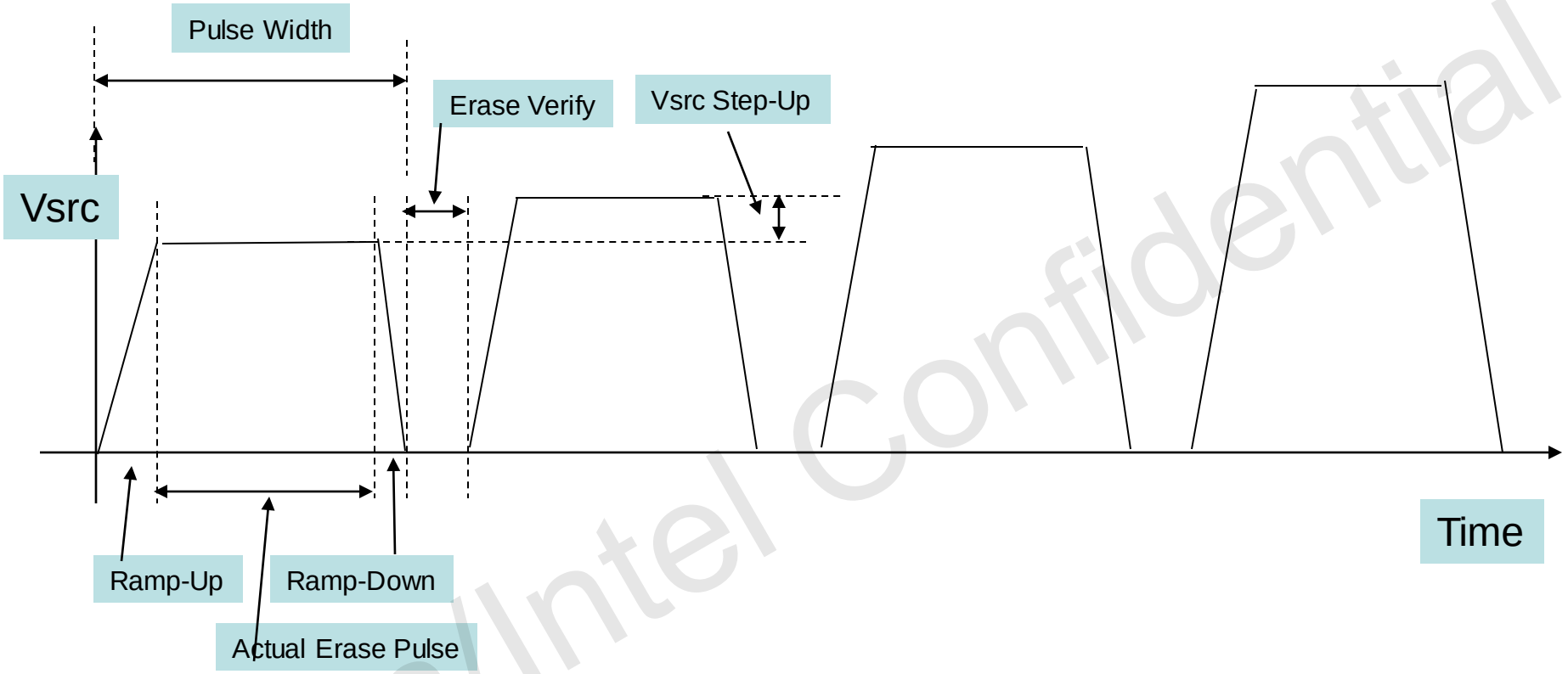


Alt-WL Erase Verify

- Each sub-block needs to be individually verified.
- Lack of LDD in the 3D makes the string shut-off at low Vg (no inversion charge over the LDD at low gate bias) → Simple string verify with TEV on all WLs do not work. → Two pass verify for each string with every other WLs at Vpass (~8-9V) to invert the LDD region.
- Continue with erase pulses until all sub-blocks pass verify (Cell Vt < Tgt)

Micron/Intel Confidential

Erase Voltage Pulse Shape



- Since Erase Vt is once again a logarithmic function of time, 1V of additional erase will require 10X longer erase time or 1V higher Erase voltage.
- By using an erase voltage that is stepped up after each verify, the exponential time impact can be turned into a linear time impact

Micron/Intel Confidential

Short Read Window Budget (RWB) Overview

Micron/Intel Confidential

Full (nearly) RWB Component List

Cell/TD/PE Parameters:

- RTS
- DRB/ICL
- FG Coupling
- VgVt Sigma
- Erratic Programming
- Back-Pattern
- Source Rise in Verify
- Poisson Pgm Noise
- Read Disturb
- Cell Tempco Sigma
- Pkg Stress
- Process Variation
- Trim Errors at Probe

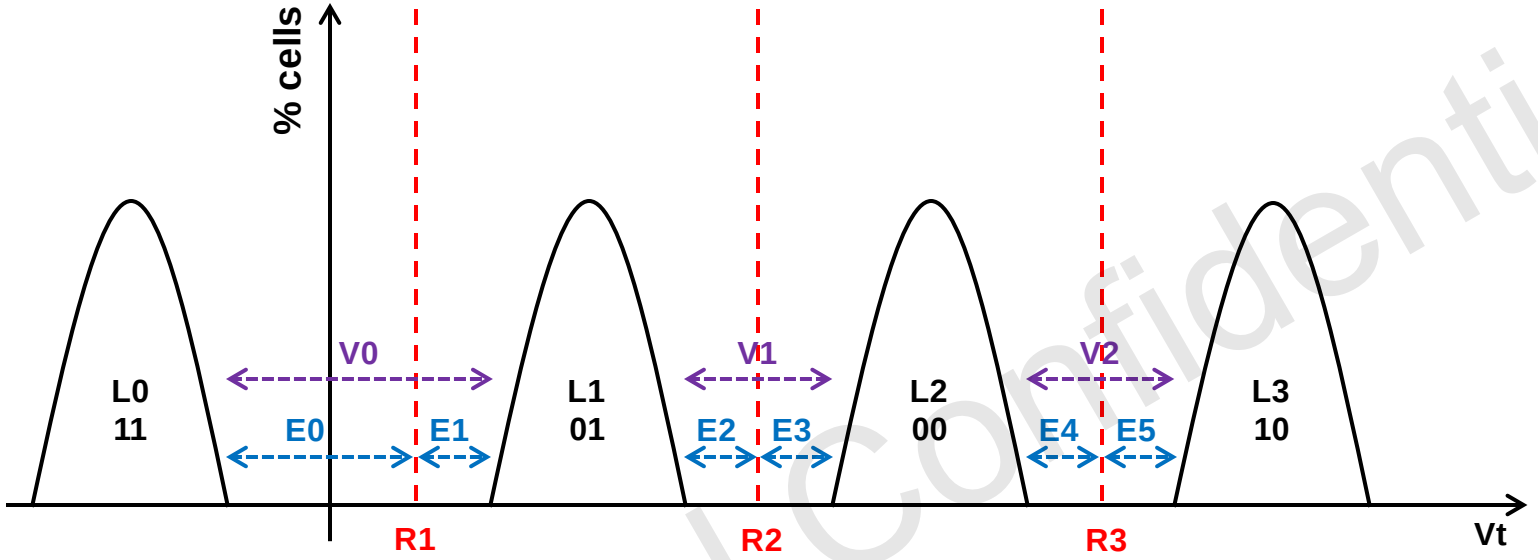
Alg/DE/Spec Parameters:

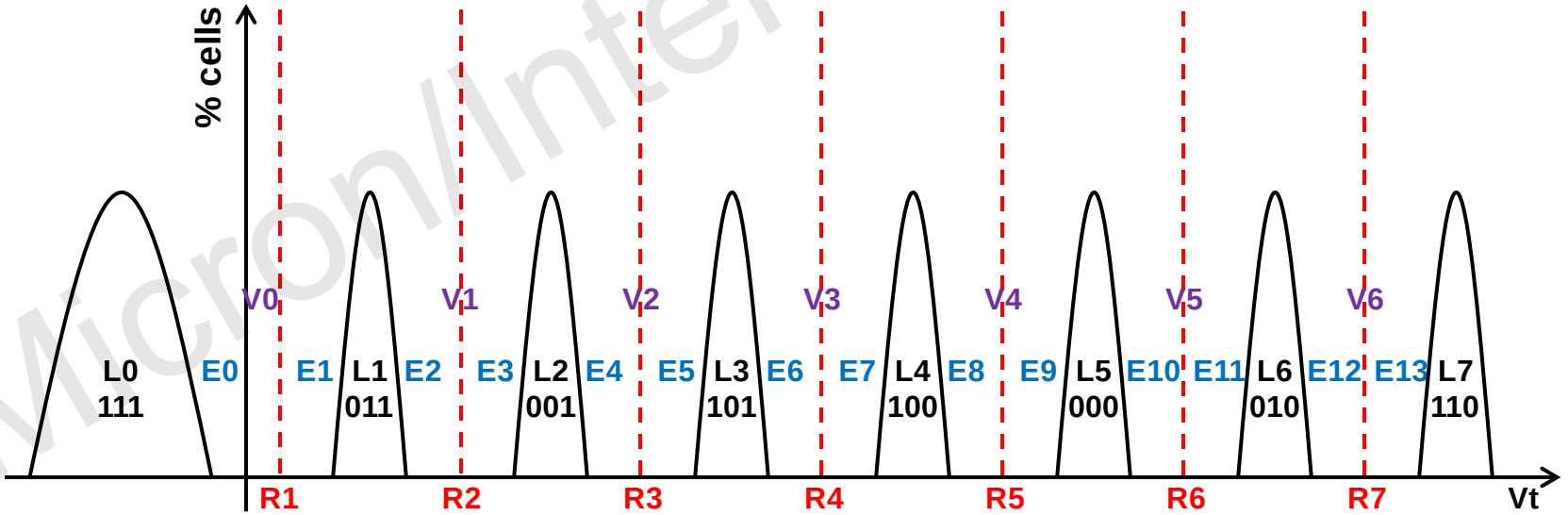
- ECC Capability
- Pgm Gate Step & SSPC
- Sensing Noise
- Data Pattern (PRP, etc)
- DAC Range & Step
- Temp Compensation Sigma
- Sense Amp Offset Sigma
- Circuit Response to Pkg Stress
- WL Voltage Regulation
- Random Unit Variation
- Density & Architecture
- Cycle Count
- Read Count

- Accounting for all these factors to arrive at the bottoms up the different state-widths and the margin requirement is the RWB methodology
- Post silicon we can directly measure the silicon and confirm the net window

Micron/Intel Confidential

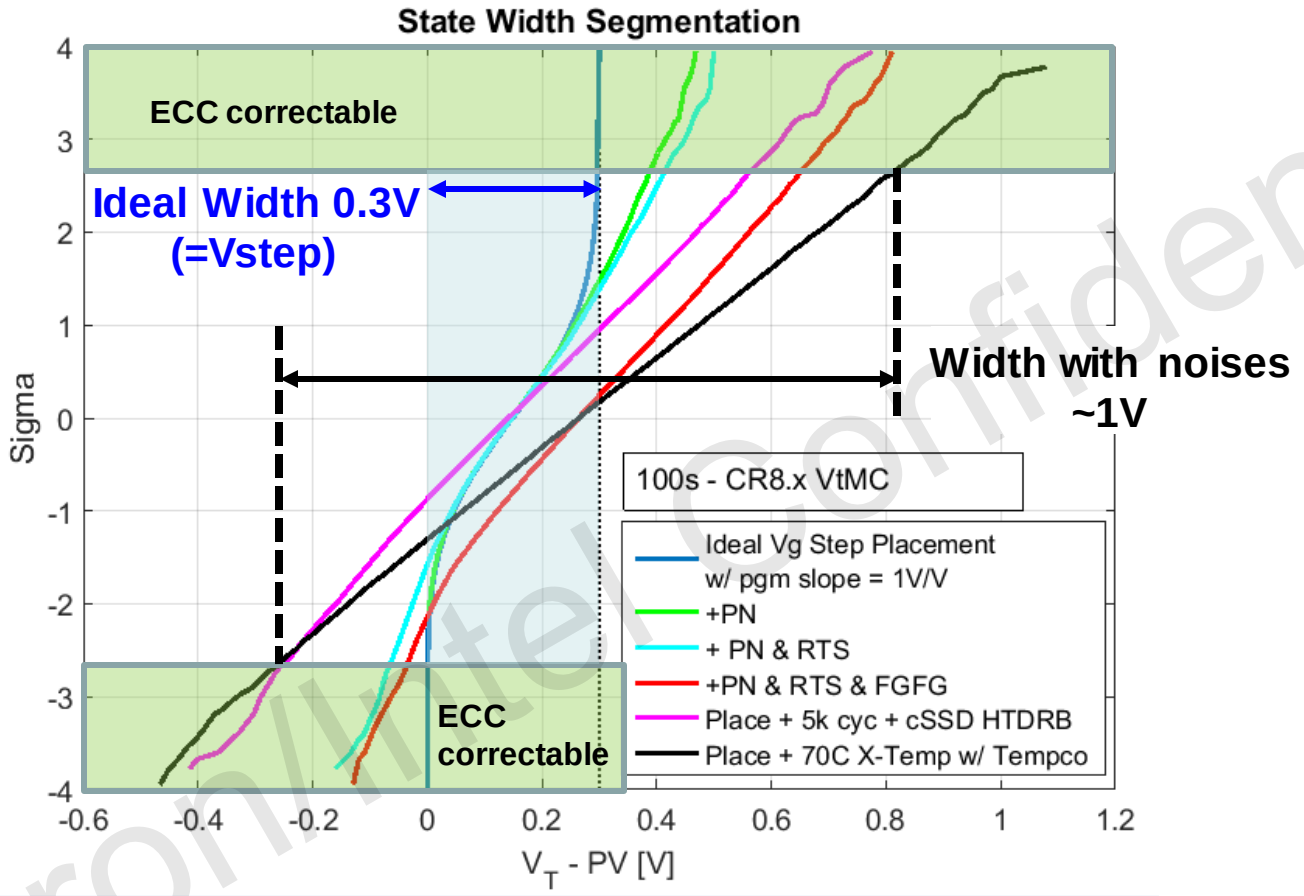
MLC/TLC distributions





Micron/Intel Confidential

State Creation 100s Example



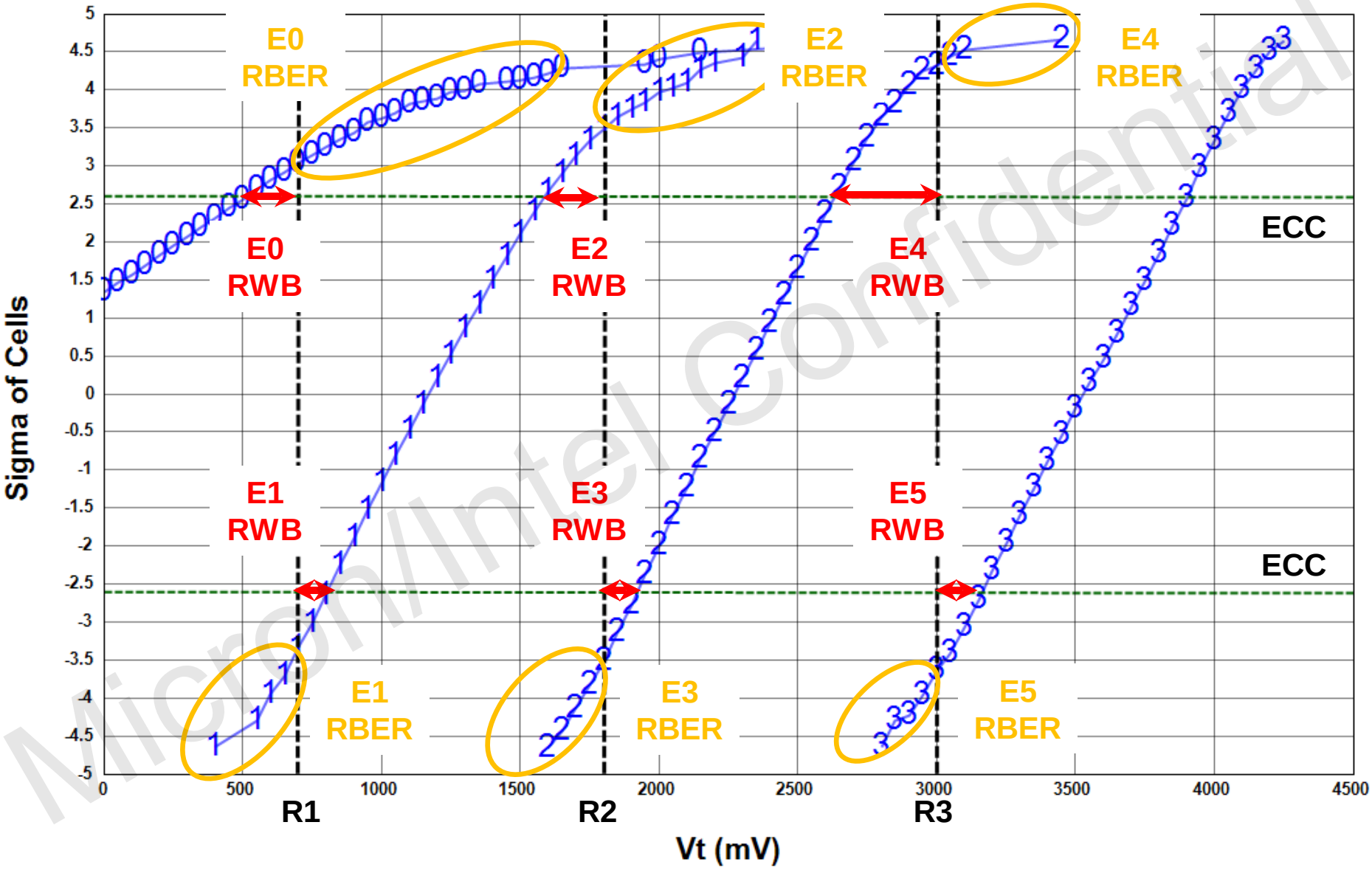
Assumptions: 300mV gate step, coupling from 1WL neighbor, MLC 1-pass algo.

Actual Vt width with noise components

is much larger than the ideal width dictated by Vgstep only.

Micron/Intel Confidential

RWB and RBER definition



Micron/Intel Confidential

Short Cell Reliability Overview

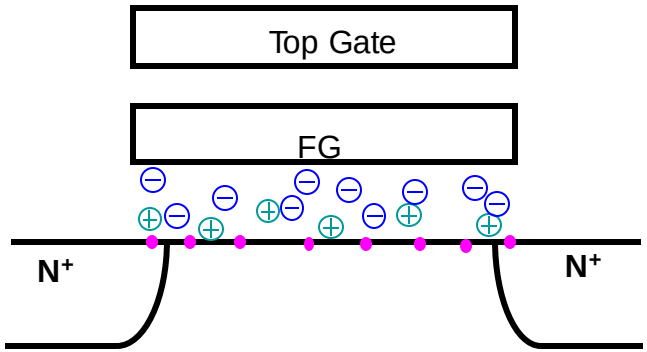
Micron/Intel Confidential

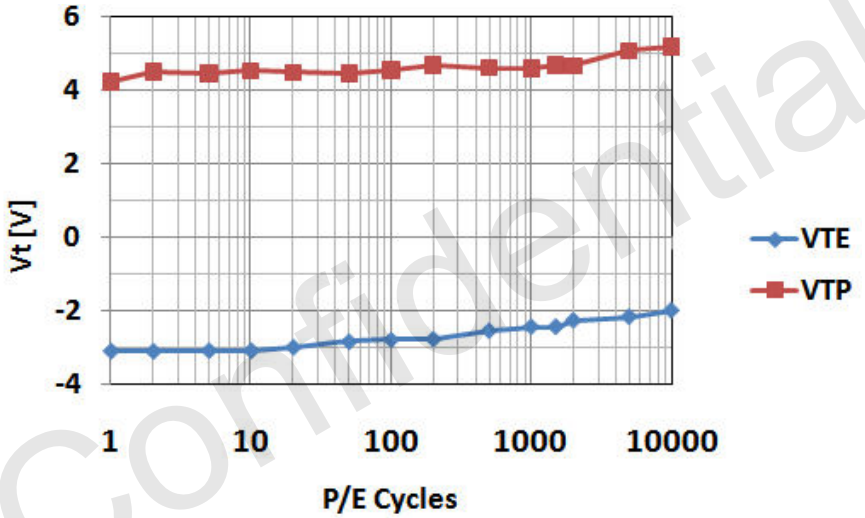
Reliability Overview

- Endurance & Trap-up
- Data Retention – SBCL, ICL
- RTS
- Read Disturb – FN & Hot-e
- Transient Vt and First Page Read effects
- Floating Body Effect
- ECC

Micron/Intel Confidential

Endurance





- Tunnel Program/Erase generates uniform electron traps, hole traps, and interface states
- Trapping of electrons in the Tunnel oxide lead to an increase in the threshold voltage of the Cell. Typically Program Vt and Erase Vt increases (at fixed P/E voltage/time). Erase Vt shift is tpically > Program Vt shift.
- Need to make sure that even after cycling we are completing erasing all the cells to Vt well below R1 level.

Micron/Intel Confidential

Trap-up impact on reliability

- Lower VgVt increases the sensitivity to the pass voltage during read, degrading read disturb
- Increased PVS (i.e. wider distribution) degrades PD because inhibited distribution gets wider
- Very fast to program cell can get misplaced after the first program pulse or due to hot-electron injection (over-programming)
- Channel interface degradation can trigger lower capability to hold boosting voltage further degrading PD (especially at high temperature)
- Trapped charge contributes to the cell Vt during program operation, but is released more easily over time, impacting RWB odd edges (E1/E3/E5) due to Intrinsic Charge Loss (ICL)
- Higher VwVt leads to lower endurance capability
- Also increased Tev-Vttop distance leads to lower endurance capability

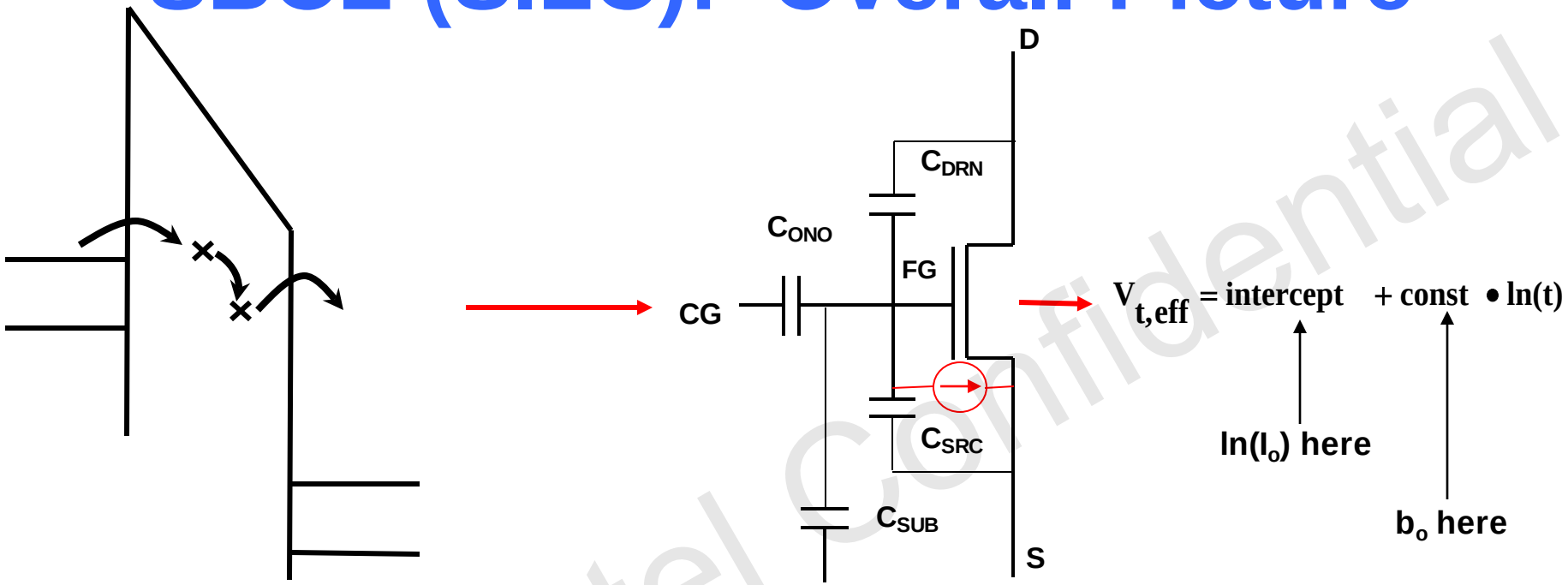
Micron/Intel Confidential

Data Retention

- Common Data Retention issues include:
  - Single bit charge loss or charge gain due to SILC related mechanisms
  - Cell Vt shift due to the de-trapping of trapped charges in the oxide
  - Charge loss or charge gain due to oxide defects, physical defects, contaminations, etc.

Micron/Intel Confidential

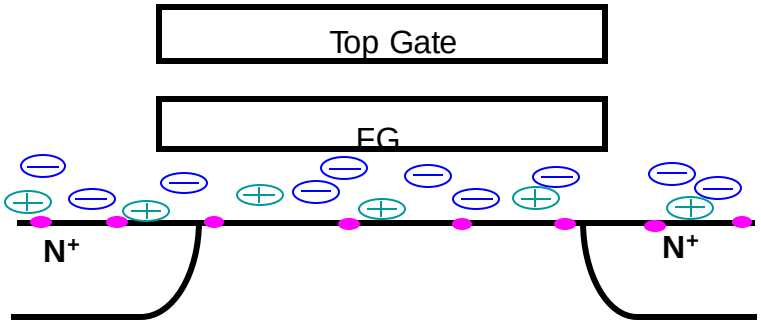
SBCL (SILC): Overall Picture

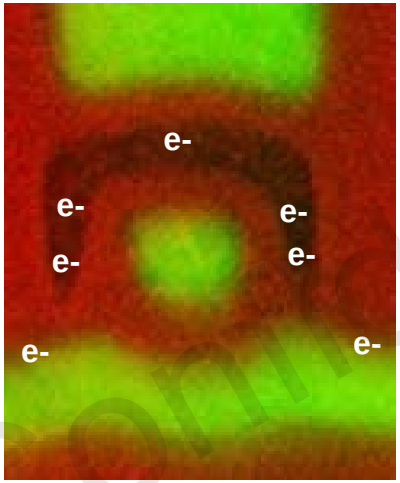


- Cycling leads to traps. Traps lead to TAT
- TAT has an exponential I-V, with a magnitude exponentially dependent on trap position
- Random trap positions lead to a wide defect tail which moves in parallel fashion with ln(t)
- Individual-cell SBCL is noisy because broken bonds are erratic – this is known from many oxide mechanisms
- The variation that we care about is in the intercept term, which stems from the I-V prefactor
  - Cell-to-cell variation is extreme value (in Vt), Weibull in time => based on our data and on STM's Monte-Carlo sims and consistent with percolation simulations of oxide TDDB
  - Cycling has ~ square root dependence on defect density, as do many other oxide mechanisms
  - Temperature dependence is weakly positive, consistent with other tunneling measurements in SiO_2
  - Defect density increases with hole injection, consistent with other oxide degradation mechanisms
  - Thicker oxide has exponential effect, ~consistent with 10x per 2 Angstroms in WKB (10(1/6) = 1.5x/Angstrom predicted)
  - Changes in UV Vt are predictable based on the equivalent circuit – simple Vt offset in asymptotic region

Micron/Intel Confidential

Intrinsic Charge Loss (ICL)



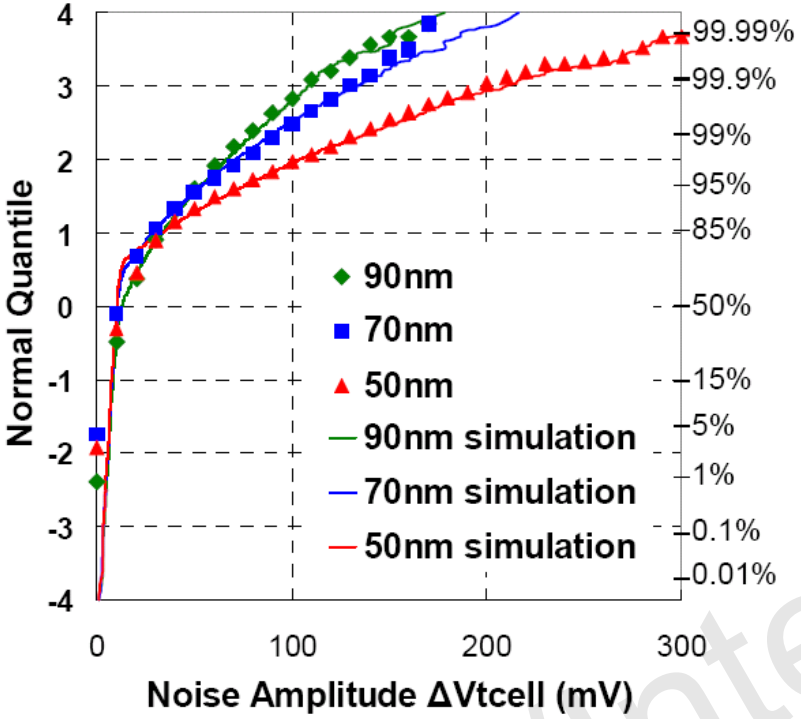


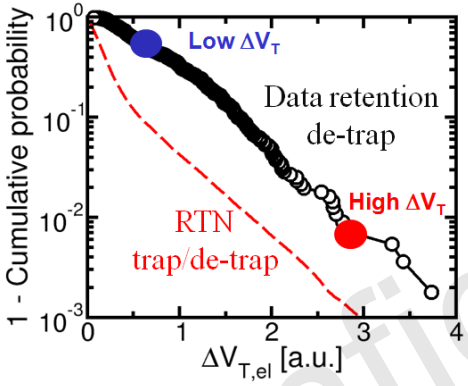
Parasitic trap/de-trapping Location in 3D NAND.

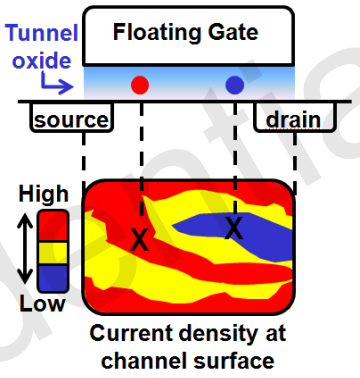
- Tunnel Program/Erase generates electron traps, hole traps, and interface states
- The trap-up occurs in the channel as well as the S/D as well as nearby dielectrics and all these trapped charges can effect the cell Vt. Detrapping of these will cause the Vt to shift
- Most Window closure and post cycling high temperature bake charge-loss (ICL) are due to bulk electron de-trapping
- ICL is also called HT-DRB due to the high activation energy causing the problem to be severe at high Temp

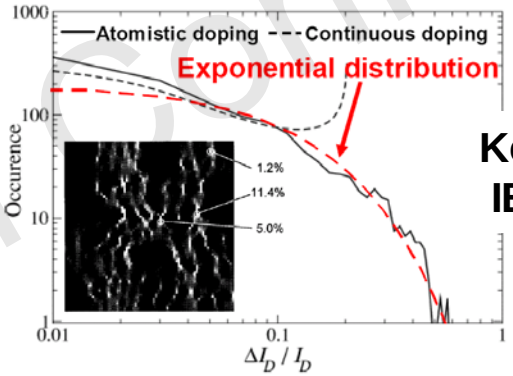
Micron/Intel Confidential

Random Telegraph Signal (RTS)







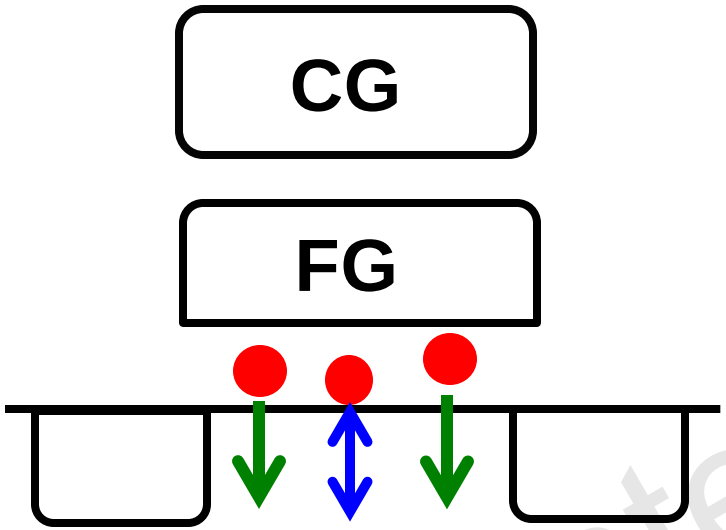


Koichi Fukuda, et al.
IEDM 2007

- Large-amplitude-noise tail portion of the noise distribution produced by the current percolation effect
- The noise amplitude is found to follow $1/(LeffWeff)^n$ making it worse as the cell is scaled
- In 3D w/ Poly channel, the additional grain-boundaries in the current path will impact RTS. However, large W of 3D reduces the impact.

Micron/Intel Confidential

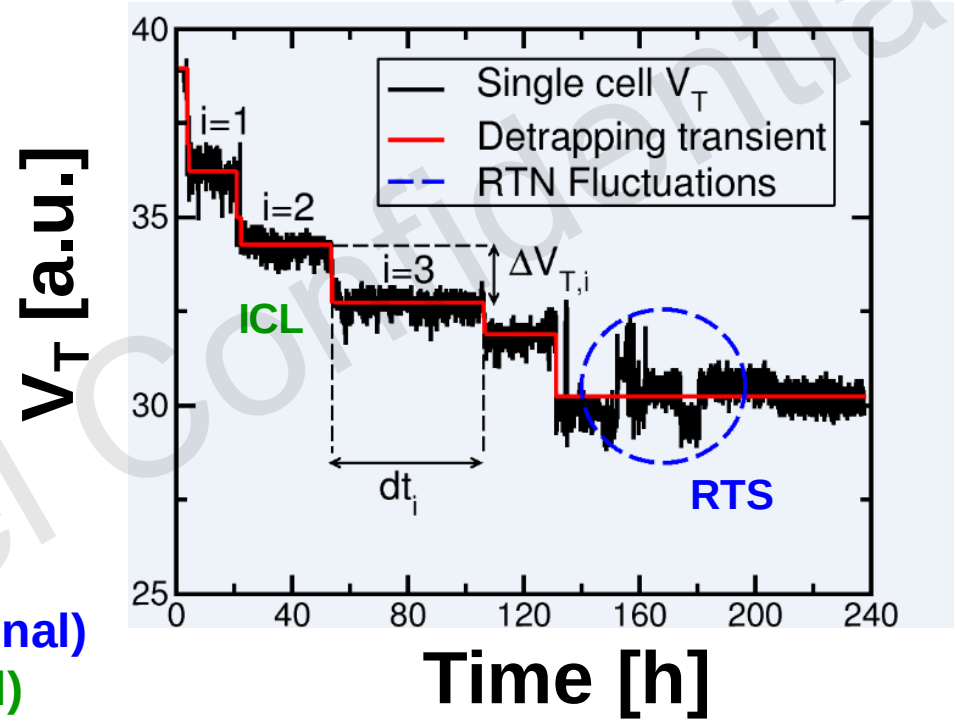
ICL and RTS



RTS (Capture/Emission, bi-directional)

ICL (De-trap, mainly uni-directional)

Time dependent V_T fluctuations



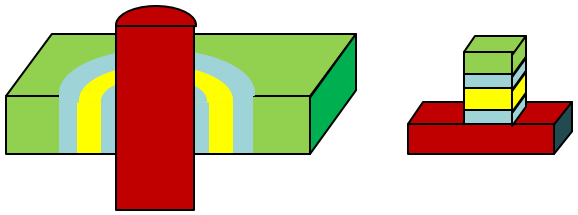
Both of ICL and RTS are Vt shift caused by charge trapping/de-trapping mainly at tunnel oxide. Vt shift characteristics and statistics can be described by

(1) # of traps and trap/de-trap time constant

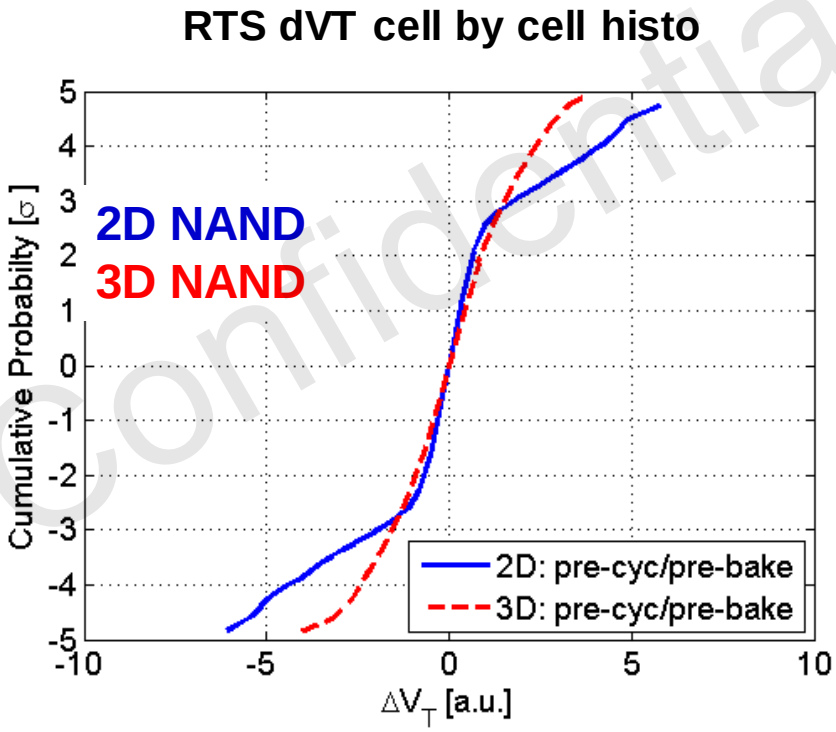
(2) ΔV_T shift per trap

Micron/Intel Confidential

3D NAND vs. 2D NAND



	3D	2D
ΔV_{T_ave} / e	Small	Large
#e fluctuation	Small %	Large %
ΔV_T/ e spread by percolative conduction	Large chan. width	Narrow width with STI edge
	Poly Si chan.	Dopant fluctuations



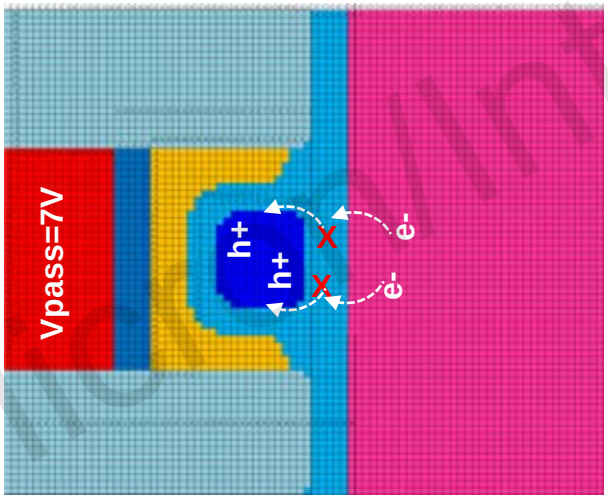
Key difference from 2D NAND is ΔV_T / electron statistics.
With a successful poly Si engineering, better RTS than 2D NAND, owing to the large channel width minimizing percolative current conduction,

Micron/Intel Confidential

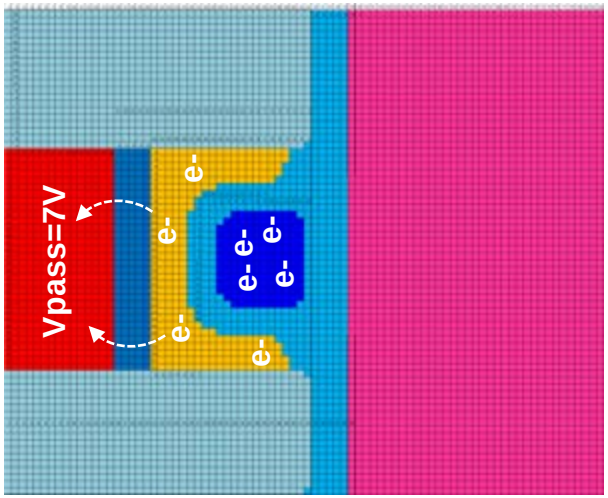
Read Disturb

- Unselected WLs in the string are biased at Vpass ~7-8V on 100s
- This leads to two issues:
  - Charge can be injected through the tunnel oxide, with a TAT mechanism that is the opposite of SBCL, mainly affecting L0 cells, due to higher electric field on the tunnel oxide: RD CG
  - Charge trapped in the IPD2 layer can be ejected through IPD3 to the CG, mainly affecting highest level (L3 in MLC, L7 in TLC), due to the highest charge trapped there and the highest electric field through IPD3: RD CL

RD CG: Read Disturb Charge Gain



RD CL: Read Disturb Charge Loss



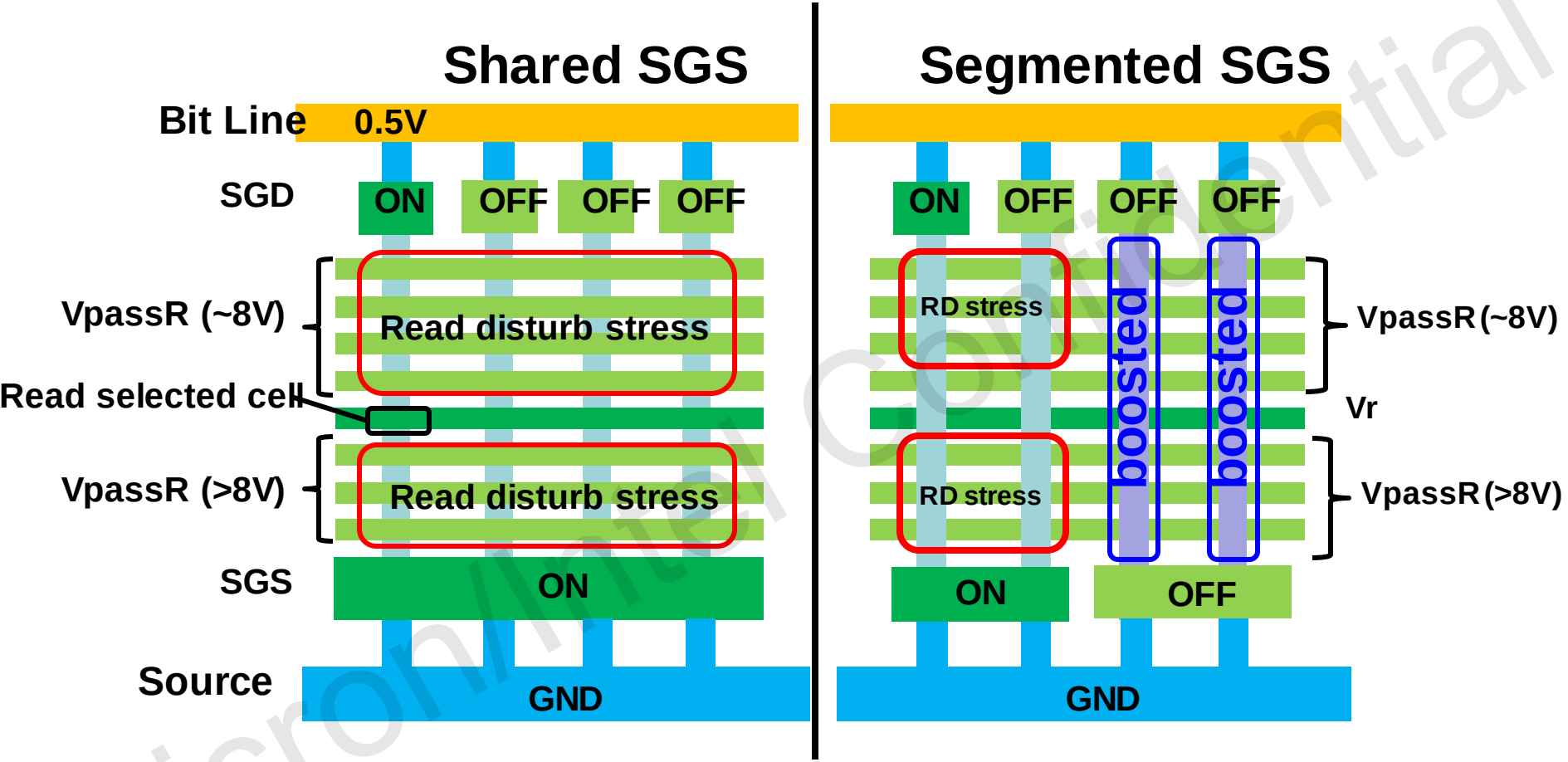
Micron Intel Confidential

Read Disturb

- Read disturb CG is one of the highest risk mechanisms on 100s
- Vpassr is higher than old technologies (~7-8V vs 5-6V) due to polysilicon channel and missing LDD region: need to create the channel not only under the cell, but also create inversion in the region between two cells
- Also the higher number of pages in the block is impacting, since the reliability goal is defined in terms of number of full block reads
- Read disturb CG gets worse with cycling, due to more traps present in the tunnel oxide (same as SBCL), and also due to higher UVVT of the cell (opposite of SBCL)
- Higher Vpassr makes also FN component present on 100s, together with TAT, creating a dependence on the VgVt of the cell
- Goal is defined at the highest cycle count for the product, e.g. 3k full block reads after 1.5k P/E cycles for TLC, or 10k full blocks reads after 3k P/E cycles for MLC

Micron Intel Confidential

3D NAND Read Disturb

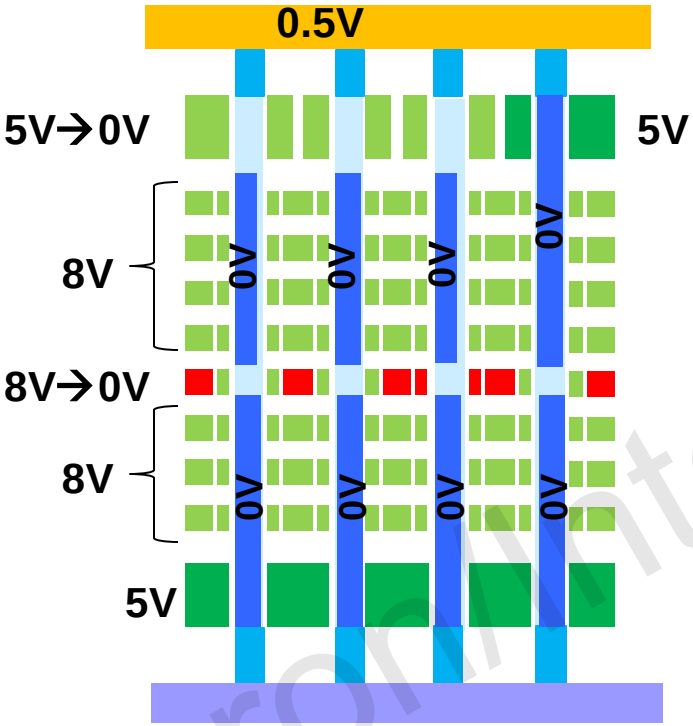


[Shared SGS] Pillars of the de-sel. sub-blocks are grounded, leasing extra read disturb stress (Rel. deg.) and WL-Pillar capacitance (Perf. and power loss).
[Segmented SGS] Some of the de-sel. pillars (75% for B1MA) are boosted During read, reducing read disturb stress and WL-Pillar capacitance.

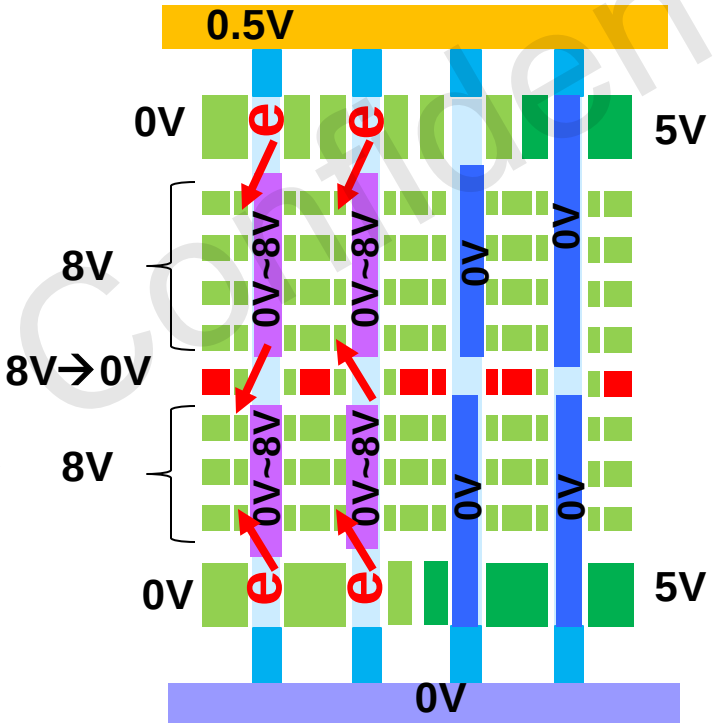
Micron/Intel Confidential

Hot-e RD

Shared SGS



Segmented SGS



Due to the High E_field at the edge of the boosted pillar, Hot-e injection disturb can occur in the segmented SGS, requiring bias and waveform optimization to minimize the hot-e disturb.

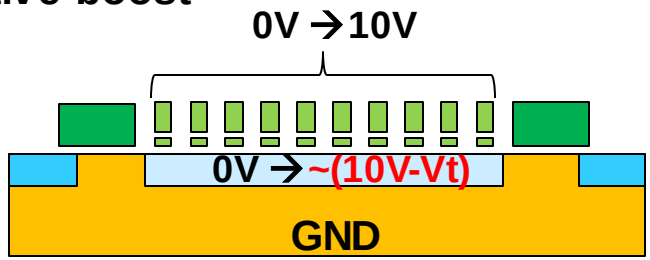
Micron Intel Confidential

Floating body effect

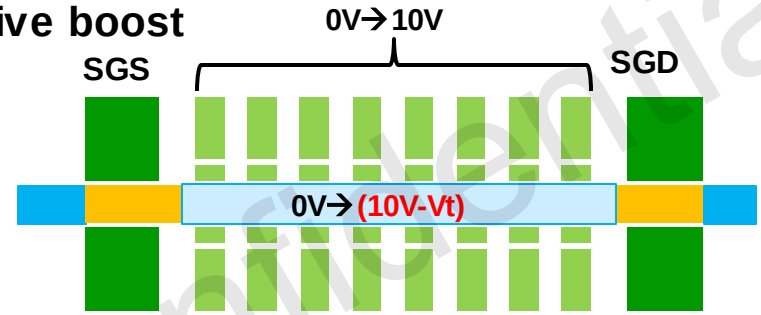
2D NAND

3D NAND

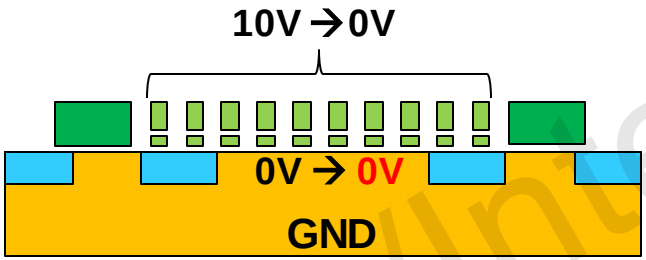
Positive boost



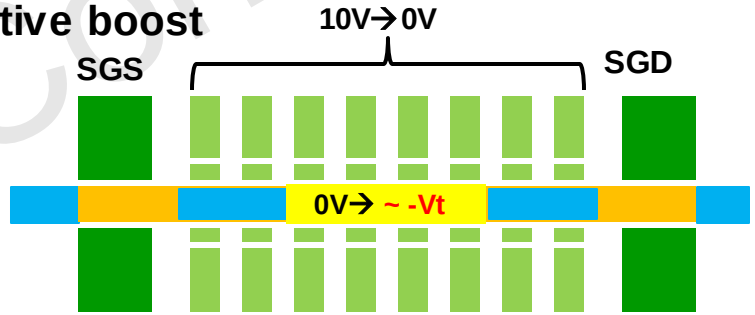
Positive boost



Negative boost doesn't occur.



Negative boost



- **Positive boost can occur in both 2D and 3D NAND.**
  This is known as boosting. Lack of electrons.
  The positive potential can be sustained by the reverse biased PN junction.
- **Negative boost can't occur in 2D NAND** because the accumulation layer is connected to pwell which supplies infinite amount of holes.
- **Negative boost can occur in 3D NAND** if the accumulation layer is sandwiched by the inversion layers. Lack of hole supply.
  The negative potential can be sustained by the reverse biased virtual PN junction.

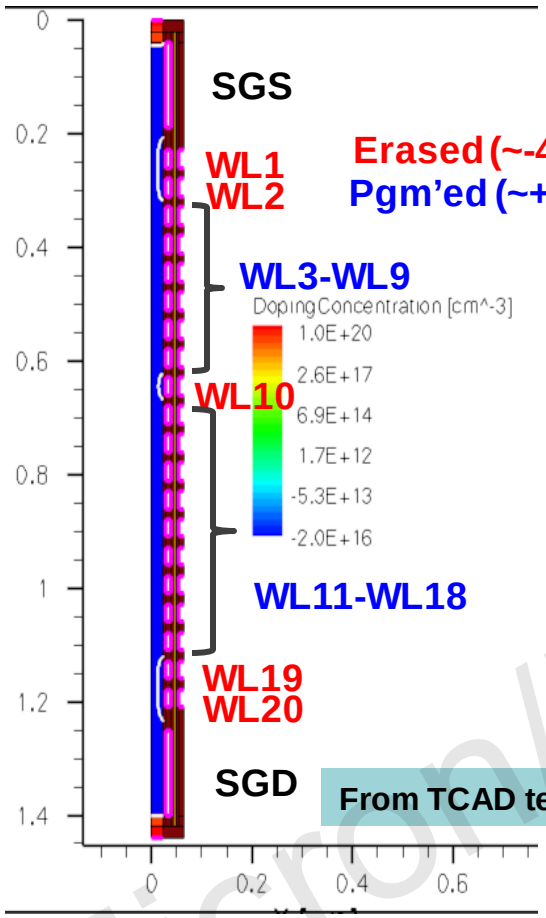
Assume WL to Ch coupling ratio ~1

Lack of electrons vs. thermal equilibrium

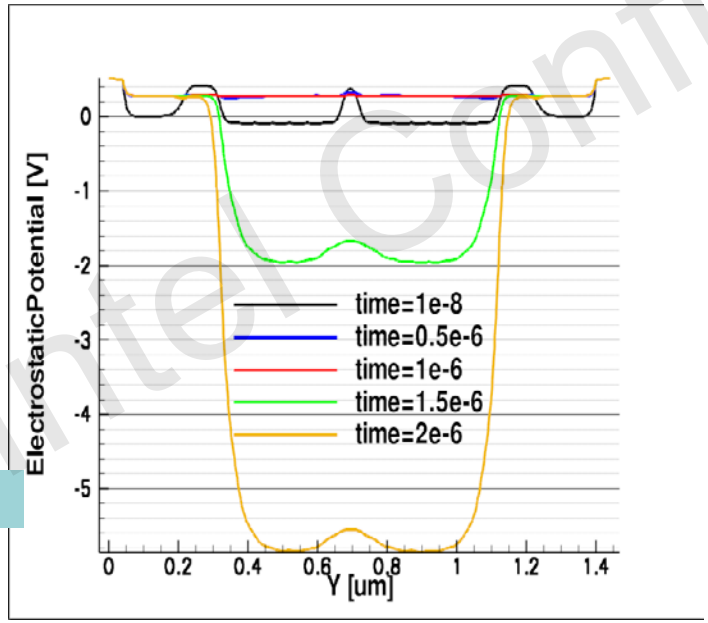
Lack of holes vs. thermal equilibrium

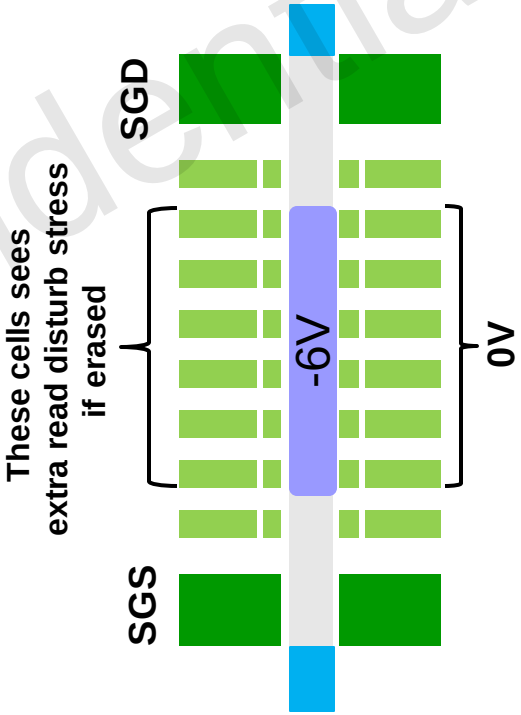
Micron/Intel Confidential

Extra read disturb stress



	Ramp up ~1us	Ramp down ~1us
WLs	0V to 8V	8V to 0V
N+ source	0V	0V
SG	3V	3V

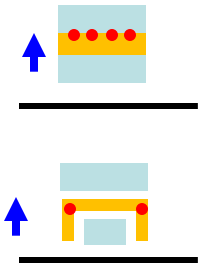
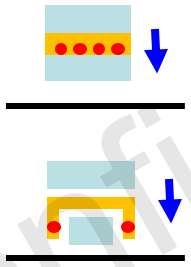
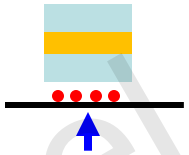
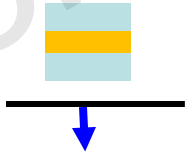
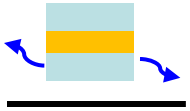
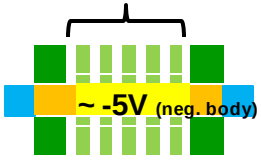
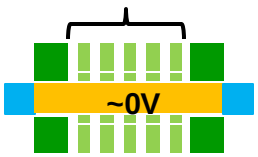




- At the end of read after WL ramping down from 8V to 0V, the body goes to negative potential. Sustained by the virtual PN junction.
- Erased cell (WL10 in this simulation) sees excess read disturb stress due to the negative body potential.

Micron/Intel Confidential

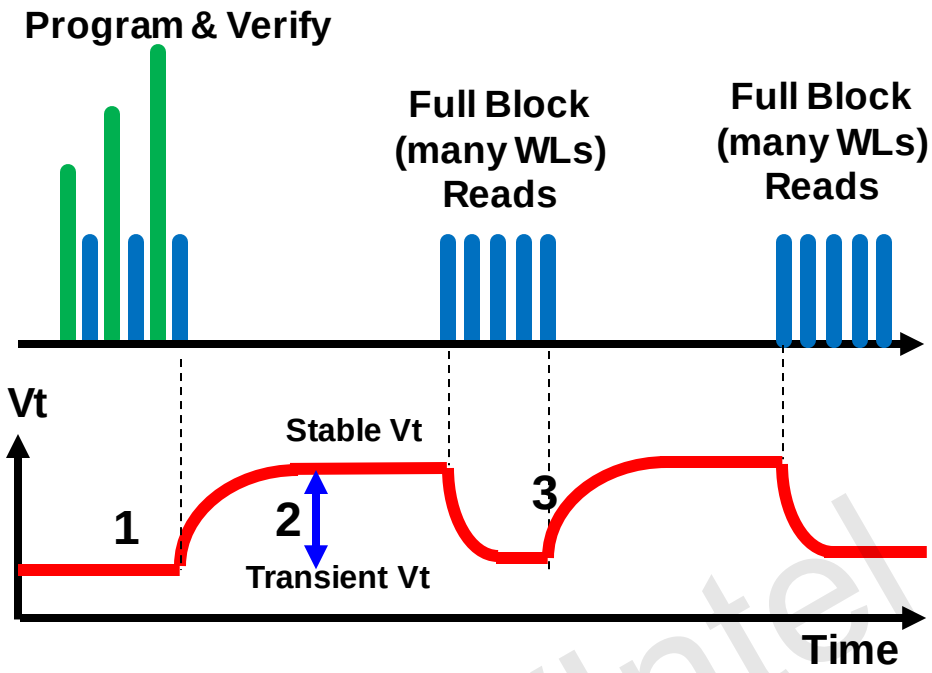
Transient Vt Mechanisms

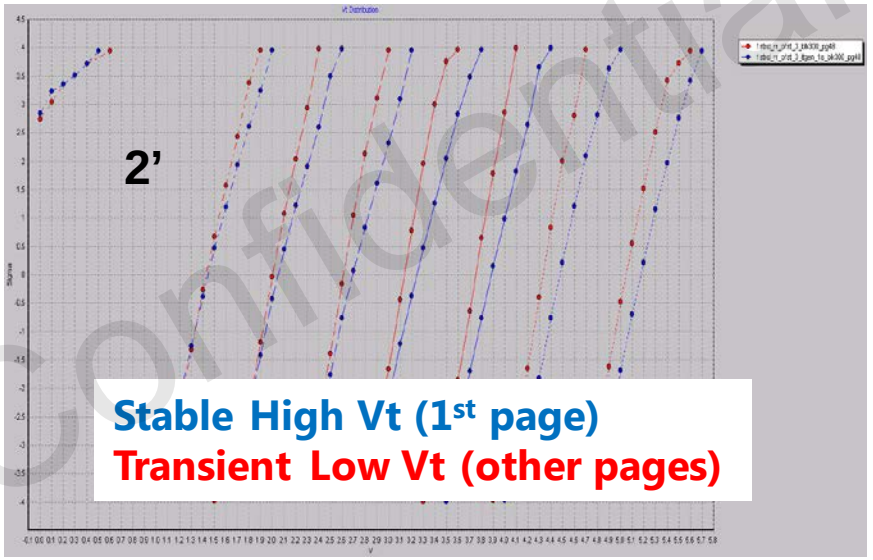
	Transient States Vg ~ +8V (Read/Pgm_inh)	Stable states Vg ~0V (idle time)	Transient Vt -> Stable Vt
Charge re-distribution or Polarization at IPD			Low VT to High Vt (QCG)
Charge trapping at Ox/Si interface			High VT to Low Vt (QCL)
Mobile charge diffusion			QCL when +Vg attract e. QCG when neg. body moves e away. (= cell with e- is stable)
Negative body potential	~+8V → 0V (ramp down) 	~0V 	Neg. body can trigger the transient. Or Neg. body can latch The transient states Created by +Vg.

A weak gate/body voltage creates a transient Vt state without changing the # charges at the storage (FG), leading quick charge loss or gain (QCL or QCG)

Micron/Intel Confidential

First Page Read (QCG example)





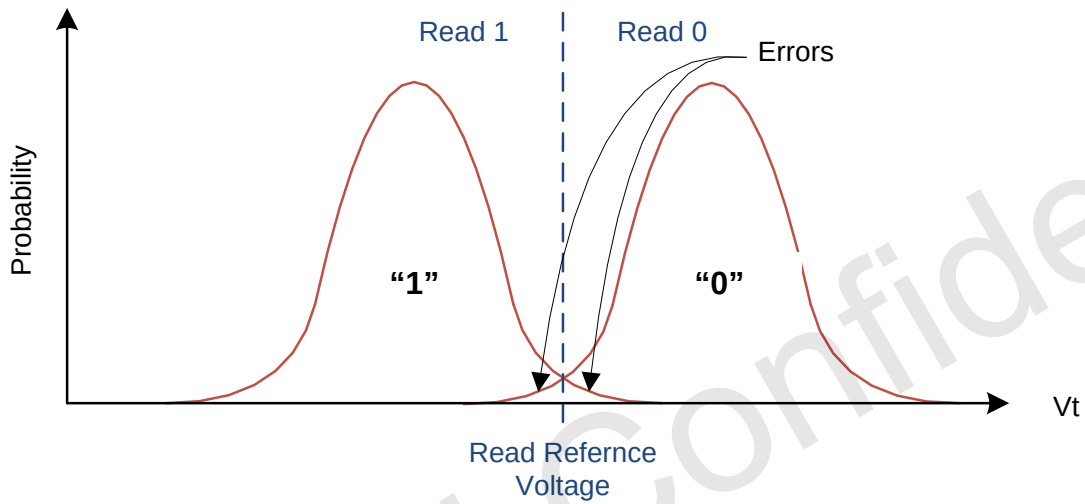
1. Pgm&Verify includes gate stress, hence reading the transient Vt.

2 & 2'. Vt transient to the stable Vt. Delta between stable and transient Vt varies among cells, therefore, Vt distribution shifts and widens.

3. 1st page read after the idle time reads the stable Vt. As read progresses, Vt moves to the transient Vt due to VpassR stress.

Micron/Intel Confidential

Source of Error in NAND Flash

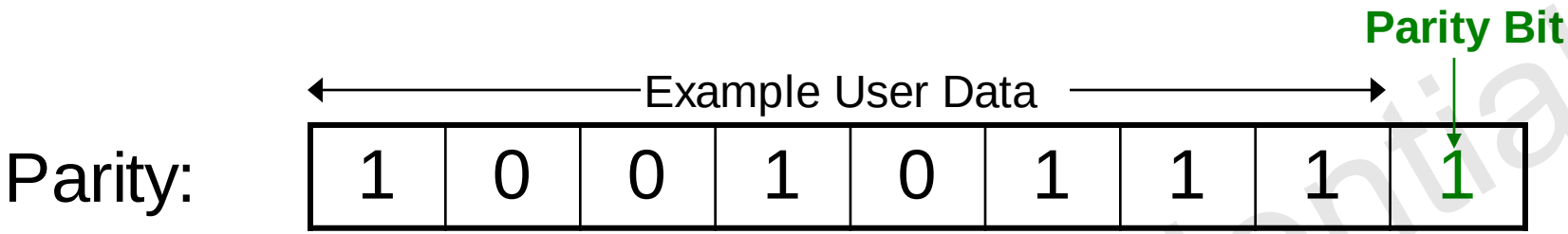


NAND distributions overlap

- Program noise
- Erratic over programming
- Program disturb
- Read disturb
- Charge Loss (ICL, SBCL)
- X-Temp
- …….

Micron/Intel Confidential

ECC: Example

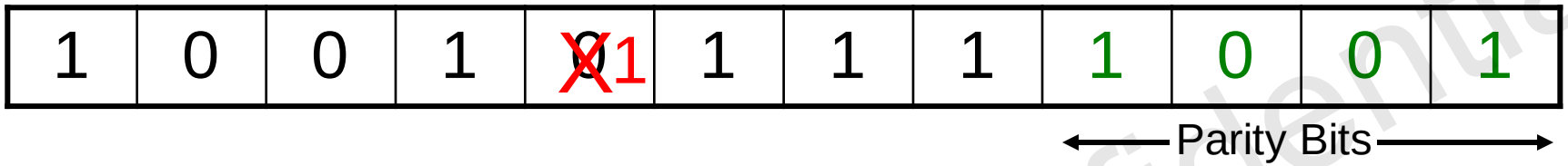


- Parity is a familiar scheme for detecting single-bit errors
  - 9^{th} bit is added to every byte and set to a value that turns the sum of all 9 bits into an even number
  - If any bit flips to the wrong state, the bits won't add up right, and the system will detect an error
  - Result: The scheme detects any single-bit error, can't correct any error, and fails if there are 2 errors
- Adding one parity bit to every byte is a scheme that localizes an error to within 9 bits

Micron/Intel Confidential

ECC: Simple Example

- By adding *four* parity bits to every byte, the error can be localized to the bad bit, and *corrected*



- Each ECC bit is a parity check on some subset of the 8 user bits:
  - 1st ECC bit = parity(bits 4,5,6,7)
  - 2nd ECC bit = parity(bits 1, 2, 3, 5, 6)
  - 3rd ECC bit = parity(bits 0, 2, 3, 5 6)
  - 4th ECC bit = parity(bits 0, 1, 3, 4, 6)
- Example: Suppose user bit 4 fails … how could the system figure that out?
  - Bit 4 is checked by the 1st and 4th ECC bits, which both flag the error
  - These two ECC bits share only bits 4 and 6, so the error must be one of those two
  - But bit 6 is checked by the 2nd and 3rd ECC bits, which are unaffected by the error, so bit 6 can't be it
  - Therefore, bit 4 must be the error
- The exercise will work on any of the bits (even the ECC bits themselves)
- This simple "Hamming" ECC detects and corrects a single error in a byte, but fails on >1 bit

Micron/Intel Confidential

How is Error Correction Done?

- For every *k* data bits, add *m* parity bits to generate a *n=k+m* size codeword

- Allow errors to occur, but have parity bits which can be used to correct the errors.
- Waste of bits due to addition of *m* parity bits per *k* data bits, but huge gains in UBER (uncorrected bit error rate), for a give RBER (raw bit error rate) so worth it.
- How many parity bits do we need? Depends on the RBER of the NAND component and the level of performance (UBER) expected from NAND.
- Large code-word sizes can improve the correction efficiency
- Once we specify the RBER and the UBER, and the code-word size, parity bits required can be calculated.

Micron/Intel Confidential

ECC Math

- Suppose one knows that a fraction P_{bit} of bits will fail in a device, called the "raw bit error rate"
- The probability that the CW will have "b" failing bits is given by the binomial probability formula:

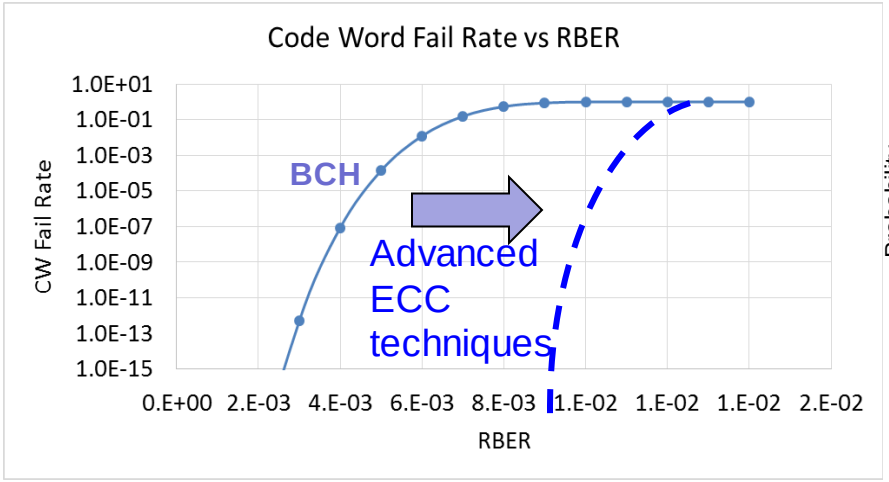
$$P_{CW} = \frac{E!}{(E-b)!b!} \cdot P_{bit}^{b} \cdot (1-P_{bit})^{E-b}$$

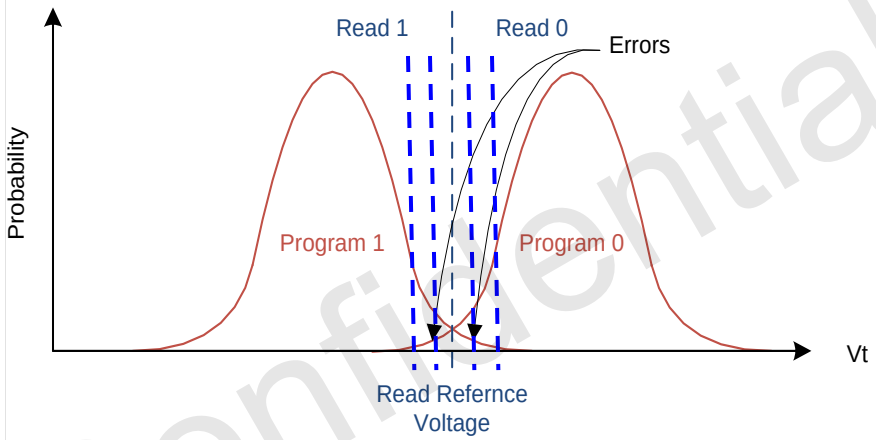
  - Here E is the # of bits in the ECC CW (1kB + 126 bytes = 9200 bits) and "!" means "factorial"

- If the ECC scheme can handle "ECC" failing bits, then the ECC fails if the # of failing bits is "ECC+1" or greater.

Micron/Intel Confidential

ECC Capability





- Based on spare bytes on our dies (16KB + 2208 ECC Bytes) we can correct up to 72 bits/1kByte correction capability.
- This doesn't mean we can tolerate 72bit/(1kB*8) of average error rate, since then every other codeword will have >72 bits failing, which will be uncorrectable.
- To guarantee good outgoing die quality the raw bit error rate (average error rate) needs to be much lower (~3.3E-3 ~27 bits/1kByte).
- Since the uncorrected error is very steep function(power of ~72), very important to make sure that worst page on the die stay below the RBER limit.
- More advanced ECC techniques that can use analog Vt information are used to improve correction capability and allow for even higher RBER (up to ~1E-2).

Micron/Intel Confidential

~Appendix~
Impact of Trapped Charge on Cell Program/Erase

Micron/Intel Confidential

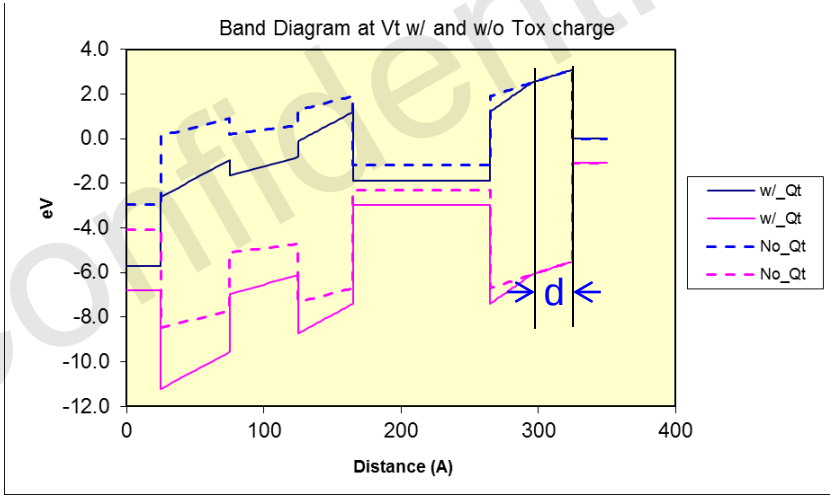
Charge Trapped in Tunox - UV_Vt, VgVt, Vtp_sat

- Consider Charge Trapped at a distance "d" from the Tunox-Si interface. (Assume d >25A to ignore barrier lowering and impact to tunneling)
- EOT is (Tox + IPD_eot/Wrap_Ratio) or Tox/GCR.

At Vg=Vt or Vg corresponding to onset of programming (Eox=Epgm (~12MV/cm)), the E_field in the tunnel-ox at Ox/Si interface is unchanged by trapped charge. But E-field above the trapped charge is higher by (Q/ε).

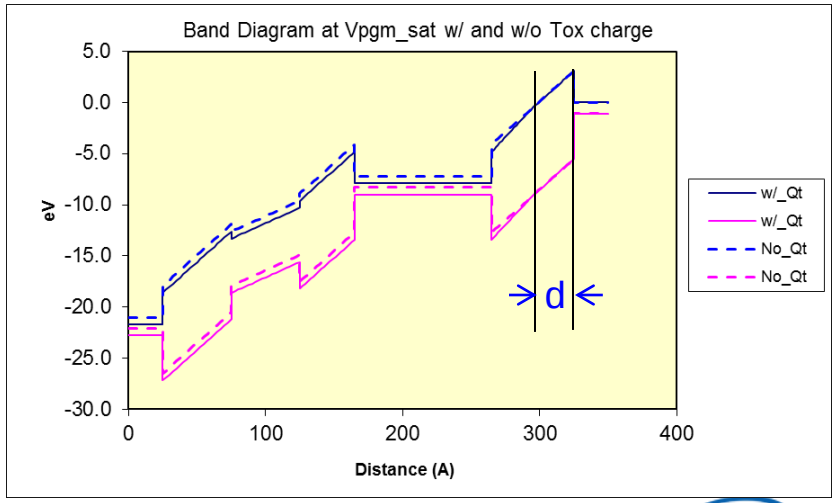
- Delta_UV_Vt = (-Q/ε)*(EOT-d)
- Delta_Vpgm = (-Q/ε)*(EOT-d)
- Delta_VgVt = 0

(If trapped charge is electrons, then –Q becomes a +ve quantity and hence UV_Vt shifts in the +ve direction)



Again at Vpgm_sat, the E-field in IPD is same w/ or w/o Tox trapped charge (~12MV/cm). Only region that has higher E-field is region between "d" and "FG"

- Delta_Vpgm_sat = (-Q/ε)*(Tox-d)
- Delta_Vtp_sat = (-Q/ε)*(Tox-d)



Micron/Intel Confidential

Charge Trapped in Tunox – VwVt, Vte_sat

- Consider Charge Trapped at a distance "d" from the Tunox-Si interface. (Assume d >25A to ignore barrier lowering and impact to tunneling)

For Erase, we need to maintain the same E-field (Eerase ~12MV/cm) at the FG/Tox interface. Which means same E-field in the IPD. The only region with higher E-field is the region between "d" and Si/Ox interface.

- Delta_Vera (Well voltage) = $(-Q/\varepsilon)*(d)$

(Vera here refers to the +ve Well voltage)

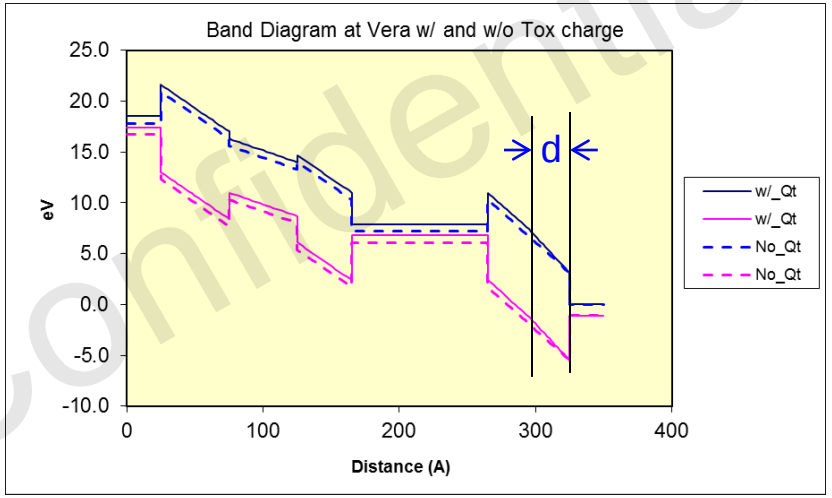
And we had:

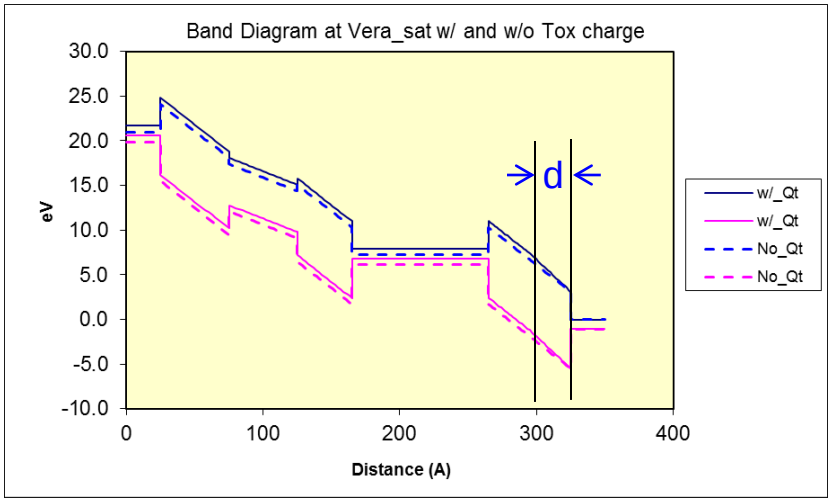
- Delta_UV_Vt = $(-Q/\varepsilon)*(EOT-d)$
- Delta_Vw_Vt = $(-Q/\varepsilon)*(EOT)$

- At Vera_sat, the E-field in IPD is same w/ or w/o Tox trapped charge (~12MV/cm). Only region that has higher E-field is region between "d" and Si/Ox interface.

- Delta_Vera_sat = $(-Q/\varepsilon)*(d)$
- Delta_Vte_sat = $(-Q/\varepsilon)*(EOT-d)$

(Vte_sat here is the actual Vt (NOT magnitude) = VwVt – Vera_sat)





Micron/Intel Confidential

Charge Trapped in IPD - UV_Vt, VgVt, Vtp_sat

- Consider Charge Trapped at a distance “d1” from the CG-IPD interface. (Assume d >25A to ignore barrier lowering and impact to tunneling).

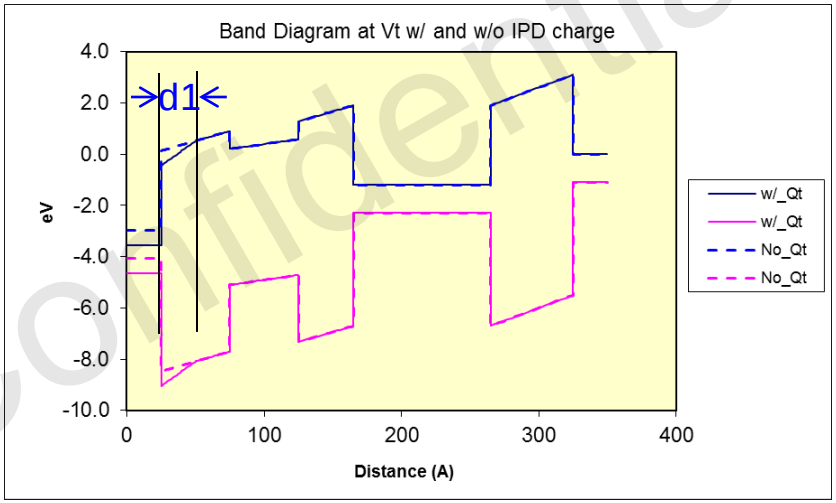
- Even though schematically the charge is shown trapped in the top IPD, the forumula is applicable charge trapped in Nitrde or bottom IPD as well.

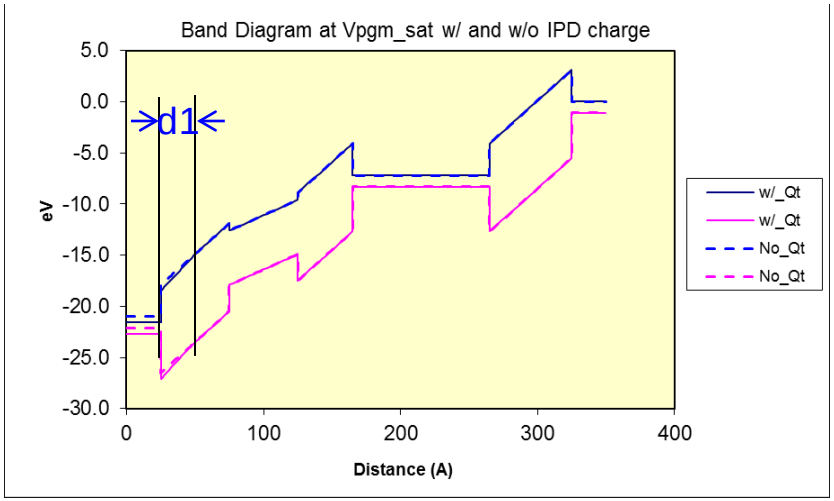
At Vg=Vt or Vg corresponding to onset of programming (Eox=Epgm (~12MV/cm)), the E_field in the tunnel-ox at Ox/Si interface is unchanged by trapped charge. But E-field above the trapped charge is higher by (Q/ε).

- Delta_UV_Vt = (-Q/ε)*(d1)
- Delta_Vpgm = (-Q/ε)*(d1)
- Delta_VgVt = 0

Again at Vpgm_sat, the E-field in IPD is same w/ or w/o trapped charge (~12MV/cm). Only region that has higher E-field is region between “d1” and “CG”

- Delta_Vpgm_sat = (-Q/ε)*(d1)
- Delta_Vtp_sat = (-Q/ε)*(d1)





Micron/Intel Confidential

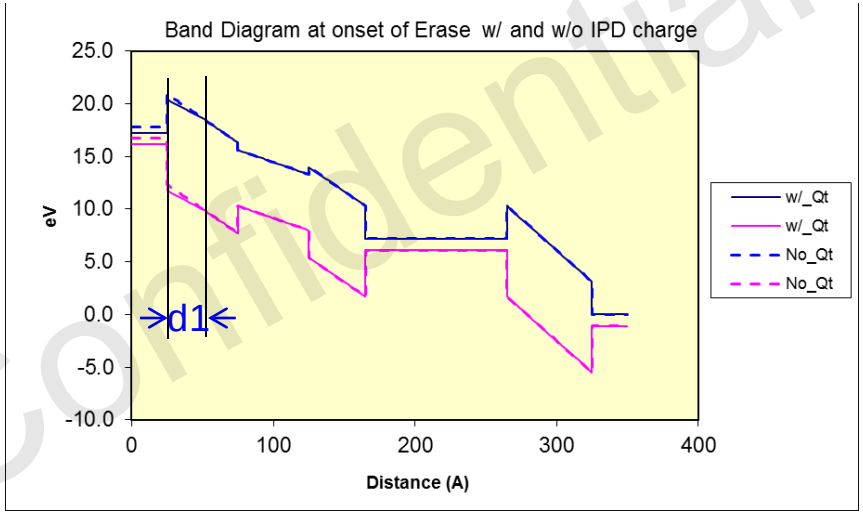
Charge Trapped in IPD – VwVt, Vte_sat

- Consider Charge Trapped at a distance "d1" from the CG-IPD interface. (Assume d >25A to ignore barrier lowering and impact to tunneling)

For Erase, we need to maintain the same E-field (Eerase ~12MV/cm) at the FG/Tox interface. Which means same E-field in the IPD. The only region with lower E-field is the region between "d1" and CG/IPD interface.

- Delta_Vera = (Q/ε)*(d1)

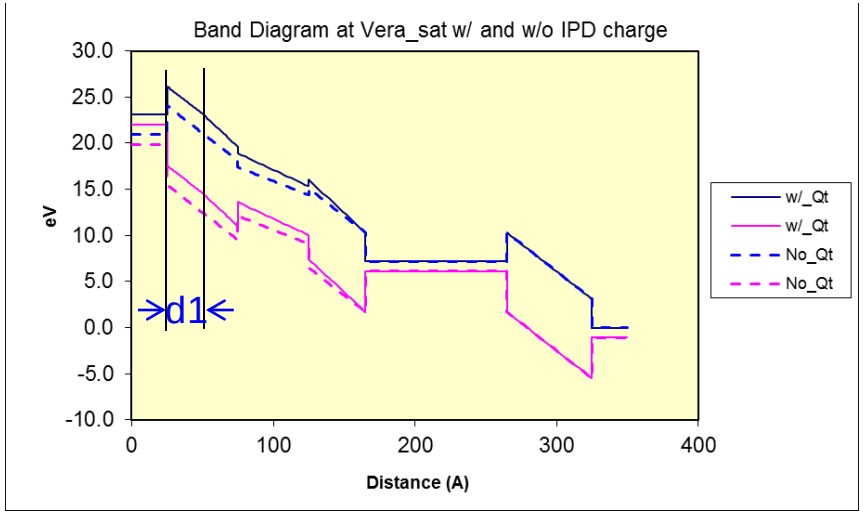
(Vera here refers to the +ve Well voltage)



And we had:

- Delta_UV_Vt = (-Q/ε)*(d1)
- Delta_Vw_Vt = 0
- At Vera_sat, the E-field in the top part of IPD is same w/ or w/o IPD trapped charge (~12MV/cm). But the region below d1 through FG will have higher E-field.
- Delta_Vera_sat = (-Q/ε)*(IPD_EOT-d1)
- Delta_Vte_sat = (Q/ε)*(IPD_EOT-d1)

(Vte_sat here is the actual Vt (NOT magnitude) = VwVt – Vera_sat)



Micron/Intel Confidential

Calculator

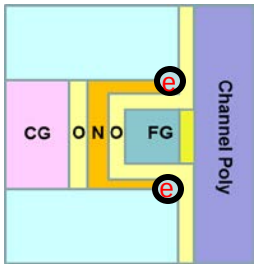
		Tox	IPD3	IPD1	Comment	
-Qt/ε	V/cm	2.32E+06	2.32E+06	2.32E+06	Charge density/cm^2	
Tox	A	60	60	60	Tox thickness	
IPD	A	115	115	115	IPD EOT	
Wrap Ratio	#	1.3	1.3	1.3	Plecholder wrap ratio	
EOT	A	148.5	148.5	148.5	Tox + (IPD EOT / Wrap Ratio)	
d	A	30			Tox Trap location from Channe linterface	
d1	A		95.0	25.0	IPD Trap location from CG interface	
					Tox Eq	**IPD3 / IPD1 Eq**
Δ_UV_Vt =	V	2.74	2.20	0.58	Delta_UV_Vt = (-Q/ε)*(EOT-d)	Delta_UV_Vt = (-Q/ε)*(d1)
Δ_VgVt =	V	0.00	0.00	0.00	Delta_VgVt = 0	Delta_VgVt = 0
Δ_VwVt =	V	3.44	0.00	0.00	Delta_Vw_Vt = (-Q/ε)*(EOT)	Delta_Vw_Vt = 0
Δ_Vpgm_sat =	V	0.70	2.20	0.58	Delta_Vpgm_sat = (-Q/ε)*(Tox-d)	Delta_Vpgm_sat = (-Q/ε)*(d1)
Δ_Vera_sat =	V	0.70	0.46	2.09	Delta_Vera_sat = (-Q/ε)*(d)	Delta_Vera_sat = (-Q/ε)*(IPD_EOT-d1)
Δ_Vtp_sat	V	0.70	2.20	0.58	Delta_Vtp_sat = (-Q/ε)*(Tox-d)	Delta_Vtp_sat = (-Q/ε)*(d1)
Δ_Vte_sat	V	2.74	-0.46	-2.09	Delta_Vte_sat = (-Q/ε)*(EOT-d)	Delta_Vte_sat = (Q/ε)*(IPD_EOT-d1)
Δ_P/E Window	V	-2.05	2.66	2.66		

- Charge trapping in Tunox reduces P/E window.
- Charge trapping in IPD can significantly improve the P/E_sat window.

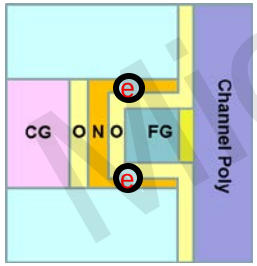
Micron Confidential

Other Charge Trapping Locations

- In addition to the Tox and IPD (directly in the path of channel Vt and/or Tunneling), we could have charge trapped around the cell (flank nitride, over LDD (beak region)).
- These charge can couple to the channel and/or FG. (But they themselves do not directly alter the Tunox or IPD E-field):
  - The coupling to channel (LDD, beak, etc.) can be thought of as UV_Vt change by channel doping change. It will impact the P/E Vts the way channel doping would have impacted P/E vts. So, VgVt, VwVt, Vtp_sat, Vte_sat, etc. will change by the same amount as Δ_UV_Vt.



    - $\Delta V_t > 0$ during the read operation;
    - $\Delta V_{pgm} = 0;\ \Delta V_g V_t = \Delta V_{pgm} - \Delta V_t = -\Delta Vt < 0;\ \Delta V_{tp\_sat} = \Delta Vt > 0;$
    - $\Delta V_{era} = 0;\ \Delta V_w V_t = \Delta V_{era} + \Delta V_t = \Delta Vt > 0;\ \Delta V_{te\_sat} = \Delta Vt > 0;$
    - $\Delta P/E_{window} = 0$
  - The coupling to FG is really going to be indistinguishable from those charges being on the FG itself



    - $\Delta V_t > 0$ during the read operation;
    - $\Delta V_{pgm} = \Delta V_t;\ \Delta V_g V_t = 0;\ \Delta V_{tp\_sat} = 0;$
    - $\Delta V_{era} = -\Delta V_t;\ \Delta V_w V_t = 0;\ \Delta V_{te\_sat} = 0;$

Micron/Intel Confidential

Charge Trapping Summary

- Even though we looked at the impact of trapped charge in each of the different locations in isolation…
- …In practice, we could have charge trapped in multiple locations simultaneously
- And also based on the location of the charge, it will couple to both the channel as well as the FG and may partially be in the path of the tunneling (Tunox or IPD) as well.
- So, the real impact to the cell P/E characteristic will be some superimposition of the various impacts described.

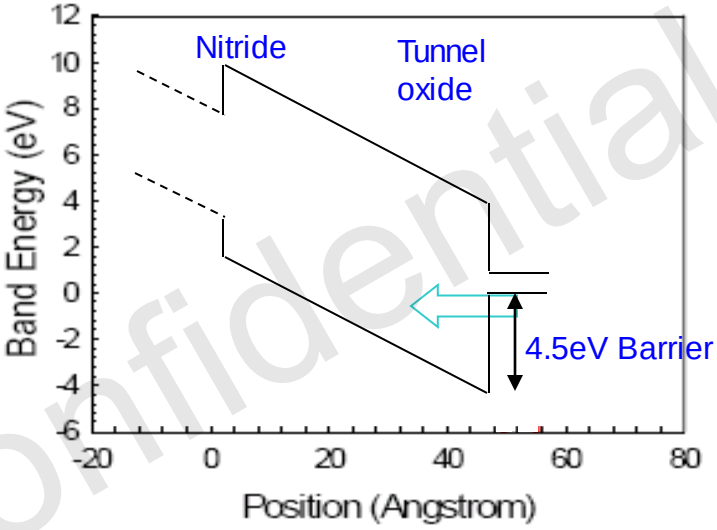
Micron/Intel Confidential

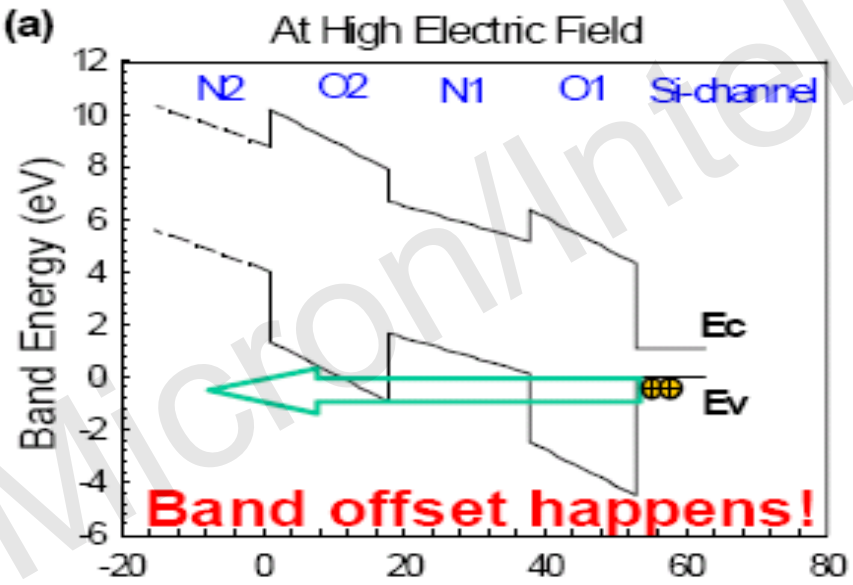
Back-Up

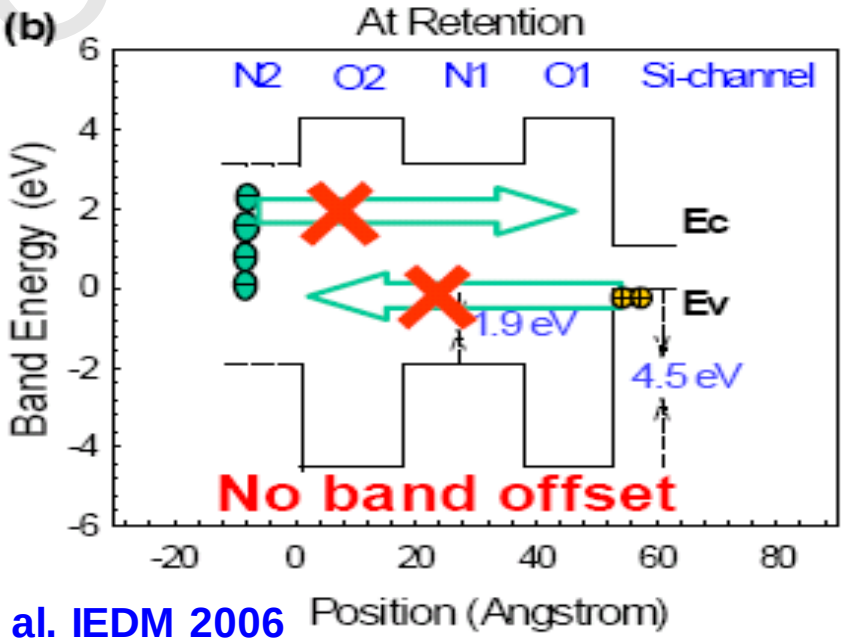
Micron/Intel Confidential

Nitride Storage - Erase

- Nitride Storage memories erase through hole injection which requires very high field.
- Engineered Tunnel barriers addresses this problem



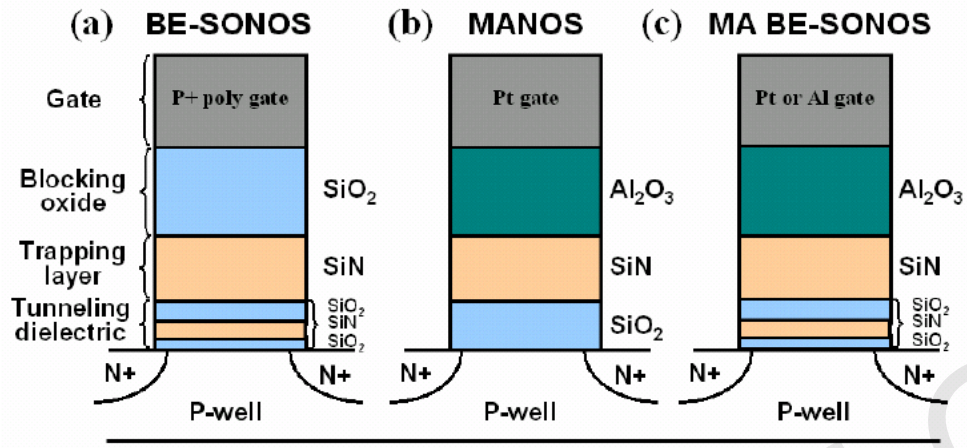




H. T. Lue, et al. IEDM 2006

Micron/Intel Confidential

Nitride Storage - Erase



S-C. Lai, et al. NVSMW 2007

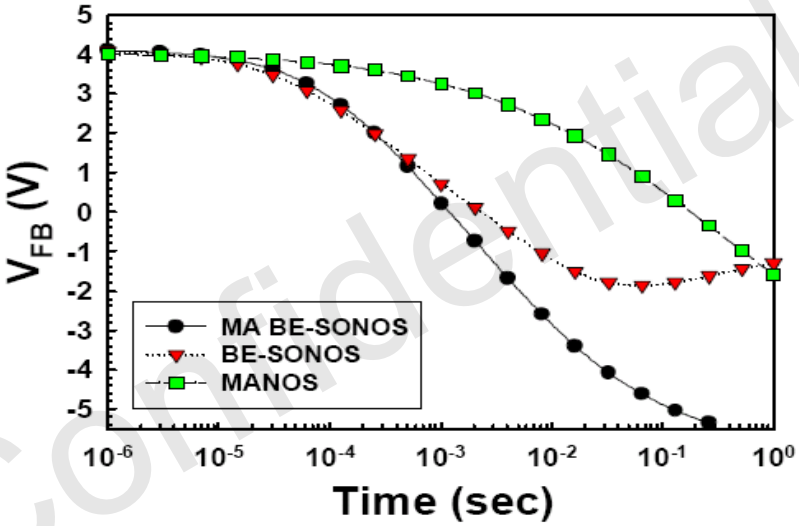


Fig. 2 Erase curves (V_{FB}-time) for the 3 devices at V_G= -18V, where the gate material for MA BE-SONOS and MANOS is platinum.

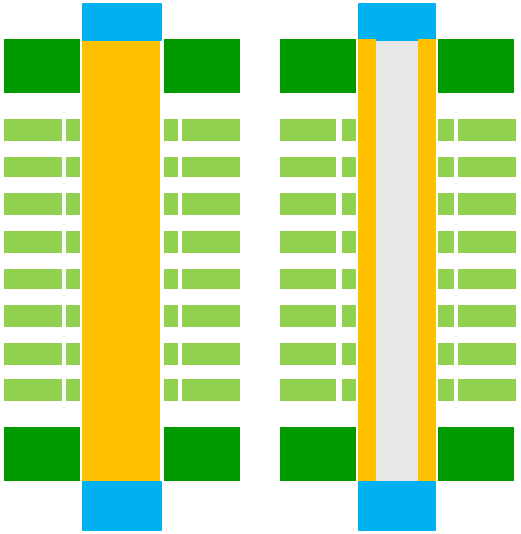
- Key requirements for good erase characteristics:
  - Engineered tunnel barrier
  - High work function Control Gate
  - High-K blocking dielectric

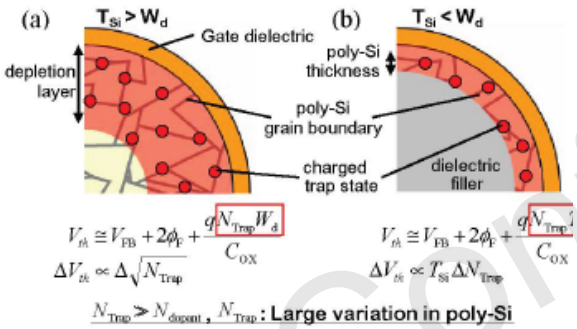
Micron/Intel Confidential

Hollow Channel

Solid channel

Hollow channel





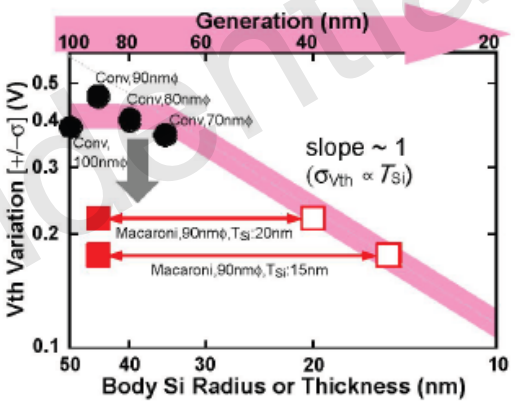


Figure 15: Extendability of 'Macaroni' body vertical FET. V_{th} variations in 300mm wafer are plotted as a function of body Si radius or thickness.

Fukuzumi *et al.* , IEDM2007

- **Hollow channel has three advantages due to the thin channel thickness { ~30nm (solid) to ~10nm (hollow) }**
  - **Vt sigma (PVS) improvement due to fewer # of traps involved in the depletion layer.**
  - **Boosting leakage reduction due to fewer G/R centers.**
  - **Sub-threshold slope and cut-off improvement due to better channel controllability, allowing n-type channel.**

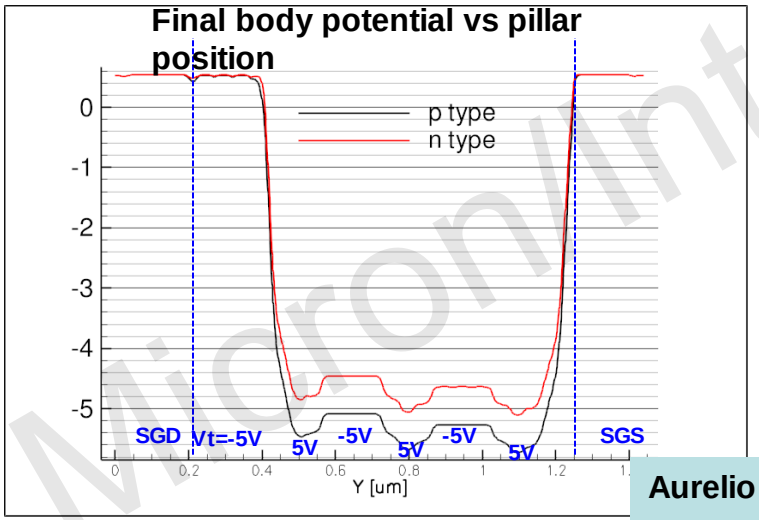
Micron/Intel Confidential

Floating Body Effect - Read

- At the end of the Read waveform, the pillar body can go negative (WL ramp down)
- The resulting negative voltage can linger for some time resulting in read disturb.
- Can happen in other waveforms and contexts as well.

Ramp up and down is 1us

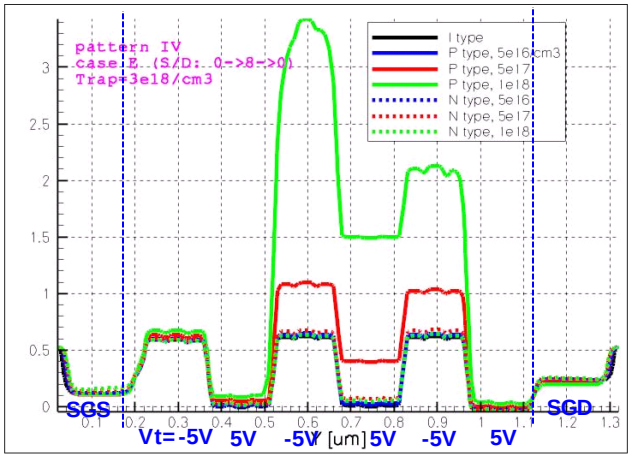
SG	WLs	N+
3V	0V to 8V to 0V	0V



Aurelio

Ramp up and down is 1us

SG	N+ S/D	WL
0V	0V to 8V to 0V	0V

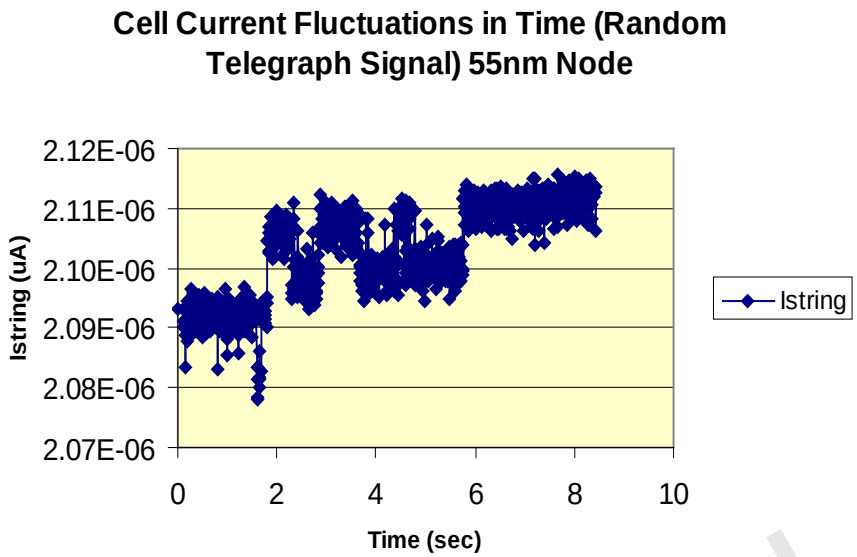


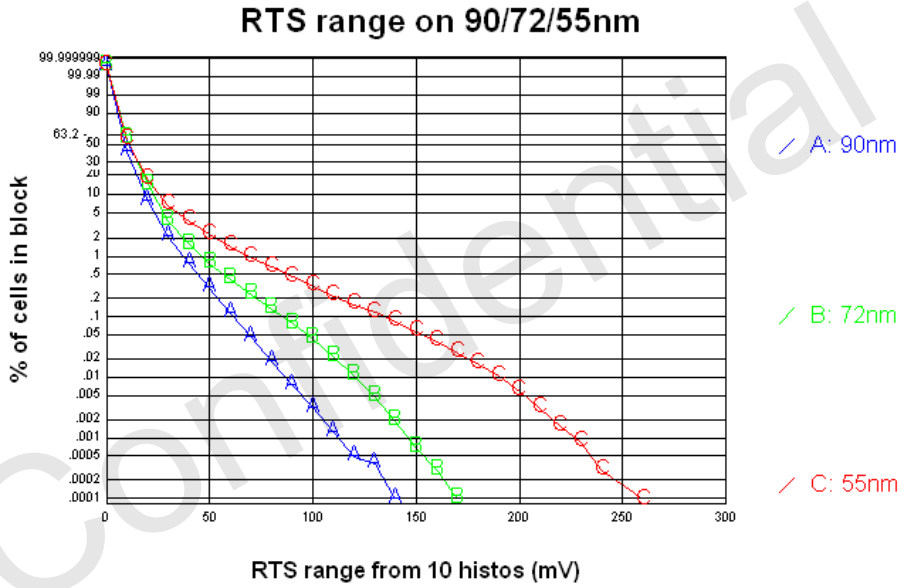
Haitao

Final body potential vs pillar position

MicronIntel Confidential

Random Telegraph Signal (RTS)





- Channel electron trapping and de-trapping during read causes current to fluctuate in discrete steps, like a random telegraph key being hit -> RTS noise
- The longer you look (many reads), the bigger the events you'll see...
- The more number of cells you look, the bigger the events you'll see

Micron/Intel Confidential