ECE 5990

Note 6
Embedded Nonvolatile Memory (eNVM)

Edwin C. Kan
School of Electrical and Computer Engineering
Cornell University
Fall 2014

Outline

- How CMOS and EEPROM works?
- Critical aspects for embedded EEPROM in RFID
  - Isolation and stabilization for high voltages
  - Manufacturing process constraints
  - Area and power
- Other embedded nonvolatile memories
  - Embedded Flash with technology modification
  - FeRAM
  - Fuse, anti-fuse and one-time programmable (OTP)
  - ReRAM

Quotable Quotes

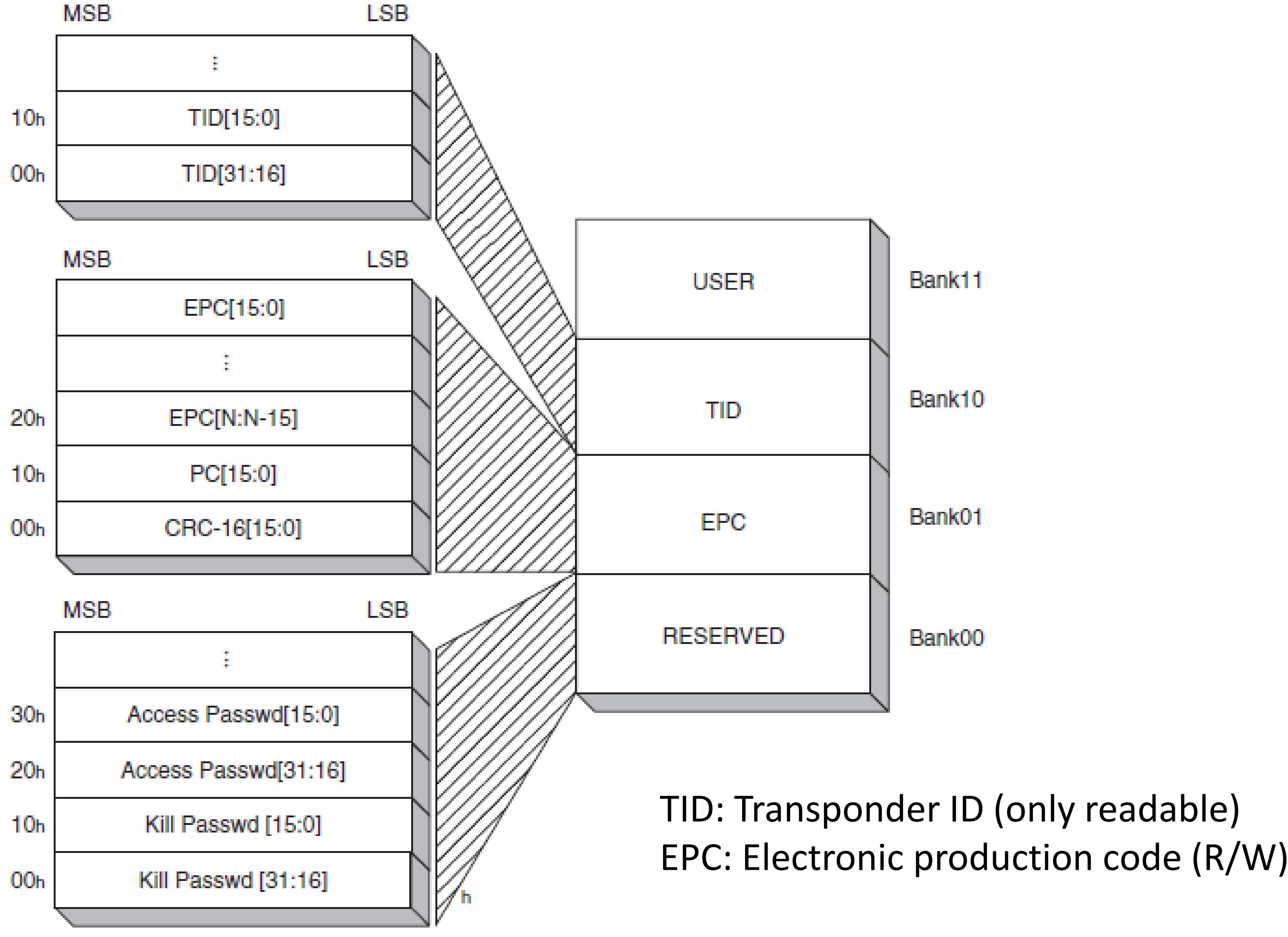
"The electronics and the semiconductor fields in particular are so dynamic and fast changing that today's concept may be obsolete tomorrow."

— Simon Min Sze (1935 - present)

Nonvolatile Storage on RFID Tags

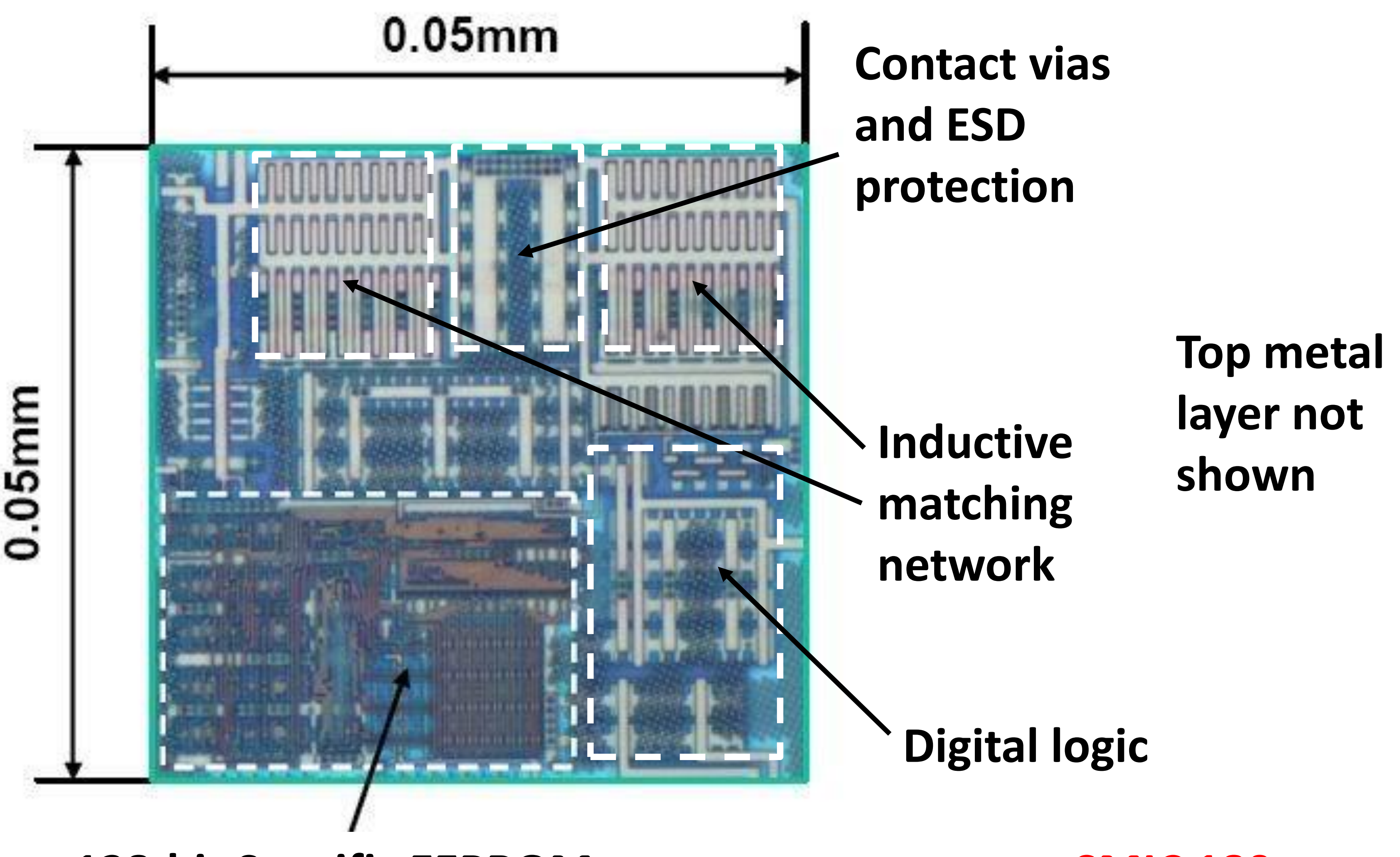
- RFID tag eNVM for:
  - Electronic production code (EPC, unique to item level)
  - Expiration date
  - Security password
  - Manufacturing data
  - Shipping information (especially cold train)
- 96 bits to several kilobits
- Peripheral circuits (decoders, high-voltage generation, sense amp, etc.) still significant layout areas

EPC in ISO 18000-6C



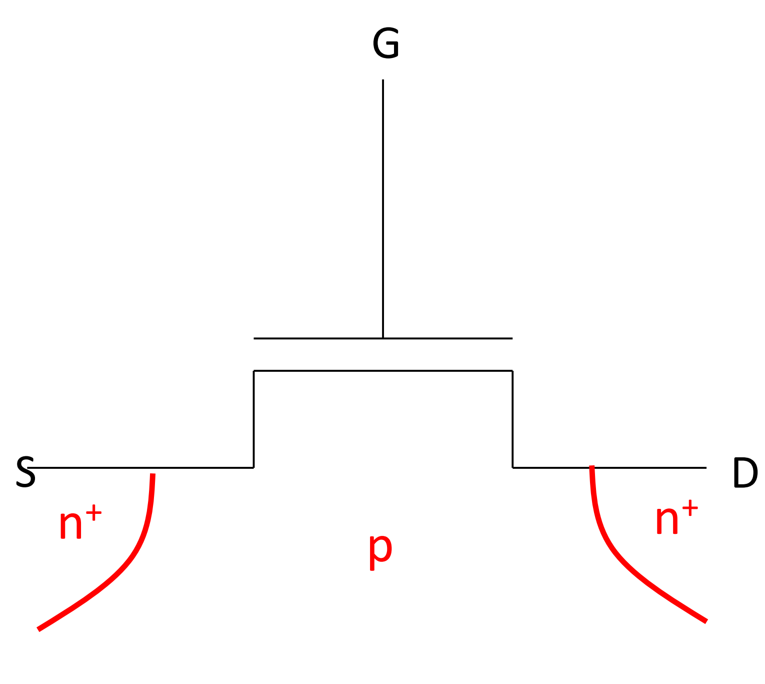
TID: Transponder ID (only readable)
EPC: Electronic production code (R/W)

Layout Design of eNVM on RFID Chips

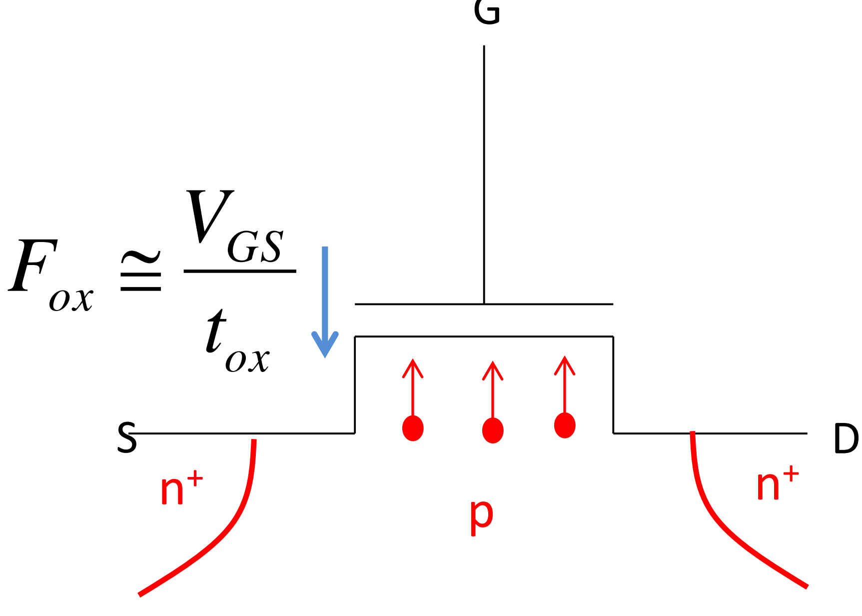


SMIC 180nm CMOS

MOSFET 101



$V_{GS} = 0; V_{DS} > 0$



$$F_{ox} \cong \frac{V_{GS}}{t_{ox}}$$

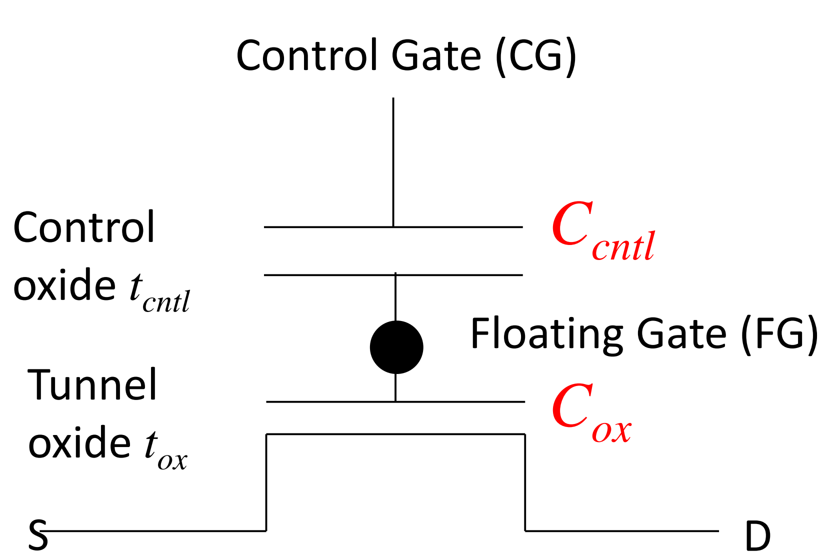
$V_{GS} > V_{th}; V_{DS} > 0$

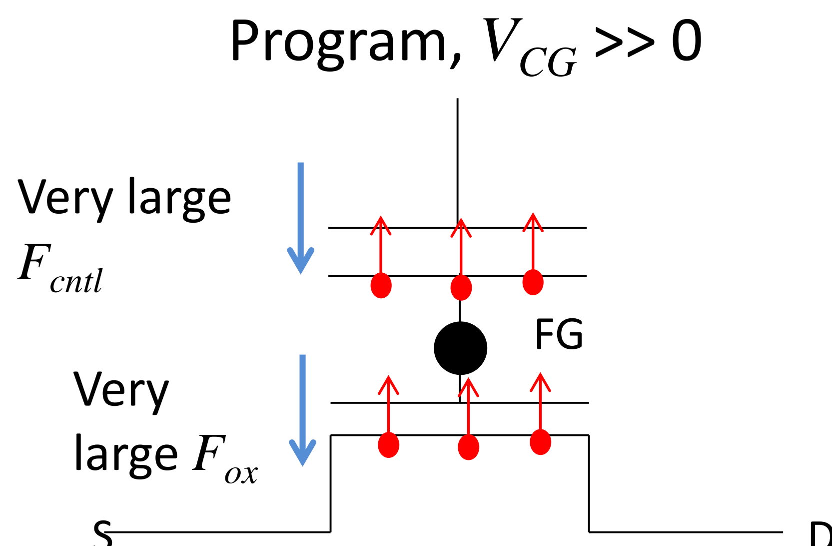
- F_{ox} : the gate oxide field
- t_{ox}: the gate oxide thickness
- ε_{ox} : oxide permittivity
- ε_{si} : silicon permittivity
- F_{si}: vertical field in channel

F_{ox} and t_{ox}

- When the gate oxide field F_{ox} = 0, the transistor channel has a n-p-n configuration (back-to-back diodes), and has minimal I_D, i.e., in the OFF state.
- A positive F_{ox} will attract electrons to the channel by the field effect with $\varepsilon_{ox} \cdot F_{ox} = \varepsilon_{si} \cdot F_{si}$.
- When V_{GS} and F_{ox} are sufficiently large to draw enough electrons, the transistor is ON.
- When V_{GS} and F_{ox} are so large, electrons can tunnel through the gate oxide to cause gate current.

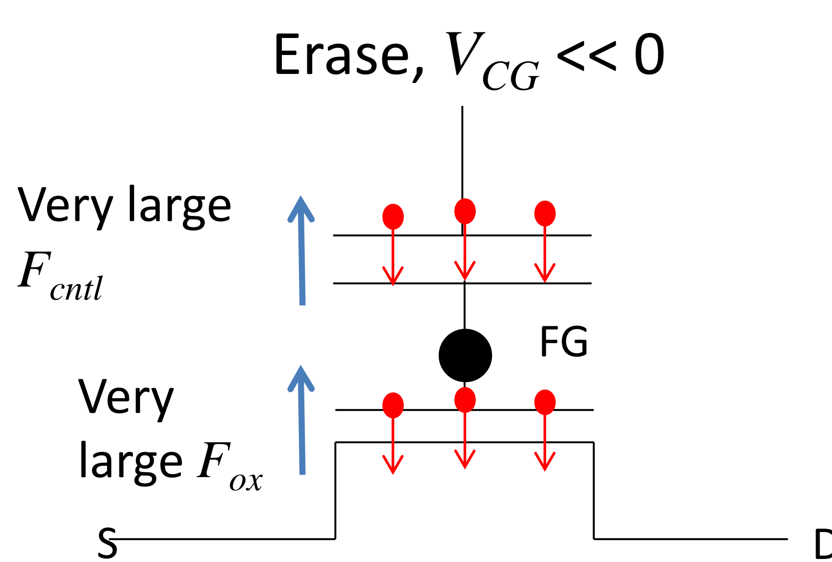
EEPROM 101





When Q_{FG} = 0, capacitive divider gives:

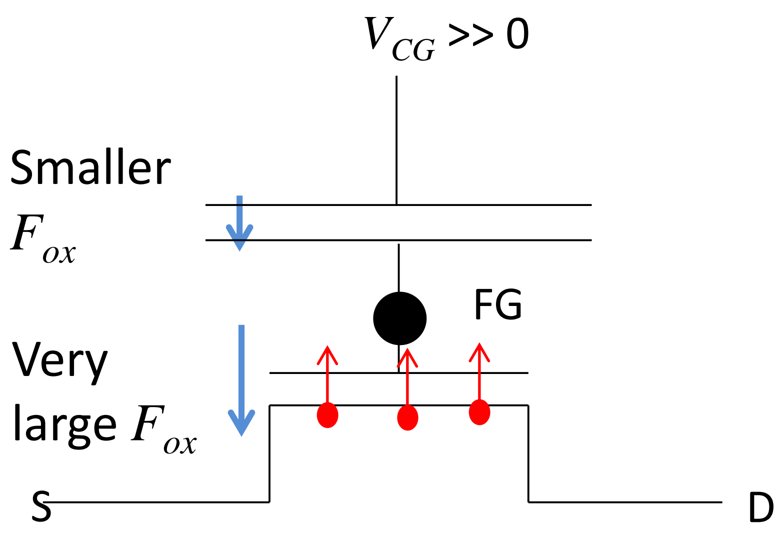
$$C_{cntl} \cdot (V_{CG} - V_{FG}) \cong C_{ox} \cdot (V_{FG} - V_S)$$



If the control and tunnel oxides have the same area and material, regardless of t_{cntl} or t_{ox}, $F_{cntl} = F_{ox}$

EEPROM Coupling Ratio

Now if $C_{cntl} > C_{ox}$ by permittivity or area, $F_{cntl} < F_{ox}$, and net charge can be injected into FG with $Q_{FG} \neq 0$.



The coupling ratio:

$$\eta = \frac{C_{cntl}}{C_{cntl} + C_{ox}} > 0.5$$

- If $Q_{FG} < 0$ (electron injection): $V_{th} \uparrow$
- If $Q_{FG} > 0$ (hole injection): $V_{th} \downarrow$

EEPROM Operations

- Both t_{ox} and t_{cntl} > 8nm SiO_2 to have charge retention time > 10 years.
- F_{ox} > 0.9V/nm to have reasonably fast charge injection for both program and erase

$$C_{cntl} \cdot (V_{CG} - V_{FG}) + Q_{FG} \cong C_{ox} \cdot (V_{FG} - V_S)$$

- Injection will stop when (self limiting):

$$\frac{V_{CG} - V_{FG}(Q_{FG})}{t_{cntl}} \cong \frac{V_{FG}(Q_{FG}) - V_S}{t_{ox}}$$

EEPROM Examples

$$V_{CG} - V_S = V_{cntl} + V_{ox} = V_{ox}\left(1 + \frac{V_{cntl}}{V_{ox}}\right) = V_{ox}\left(1 + \frac{C_{ox}}{C_{cntl}}\right) = \frac{V_{ox}}{\eta}$$

Low-voltage eNVM

- t_{ox} = 8nm and η = 0.72 (i.e., $C_{cntl} = 2.6\ C_{ox}$).
- $V_{prog} > 0.9V/nm \times 8nm \times 1/\eta = 10V;\ V_{erase} = -10V$

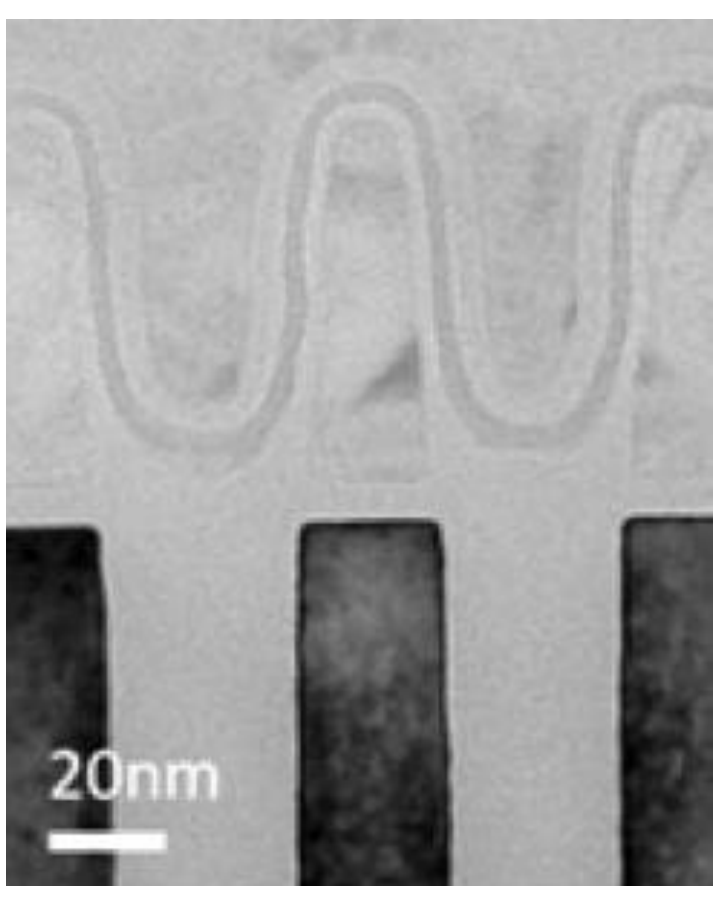
NAND Flash

- t_{ox} = 9nm and η = 0.5 (i.e., $C_{cntl} = C_{ox}$).
- $V_{prog} > 0.9V/nm \times 9nm \times 1/\eta = 16.2V;\ V_{erase} = -16.2V$

Possible Implementations: Data Flash

Dense Array Flash

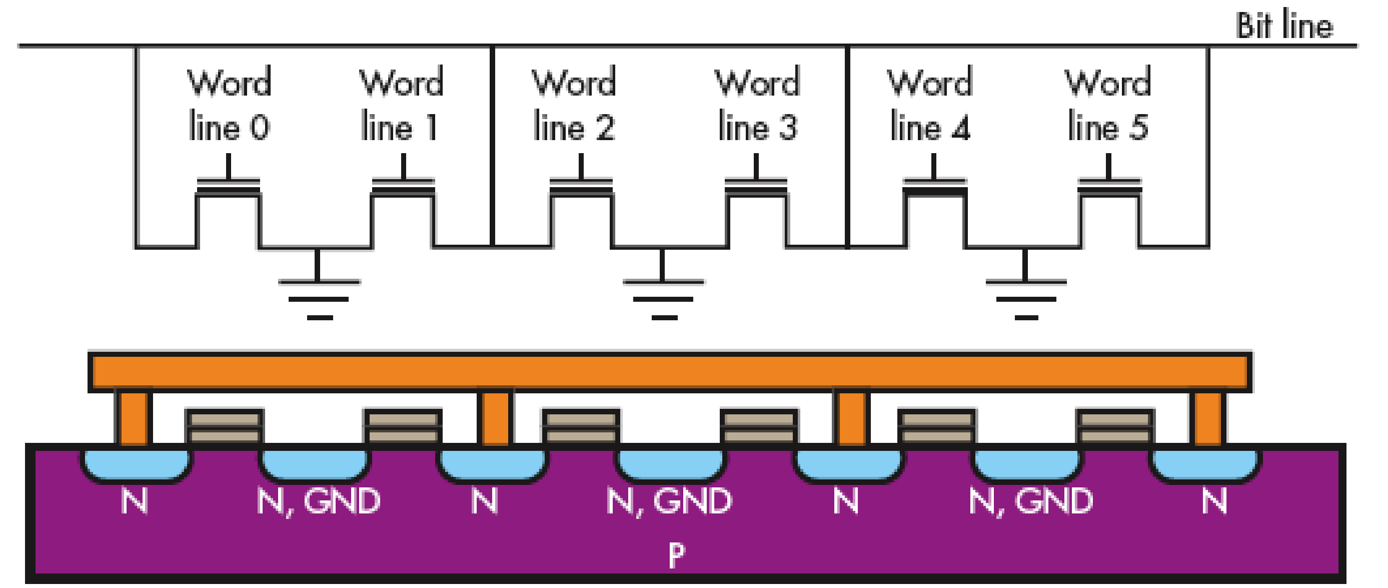
Cell area: $WL_{pitch} \times BL_{pitch}$



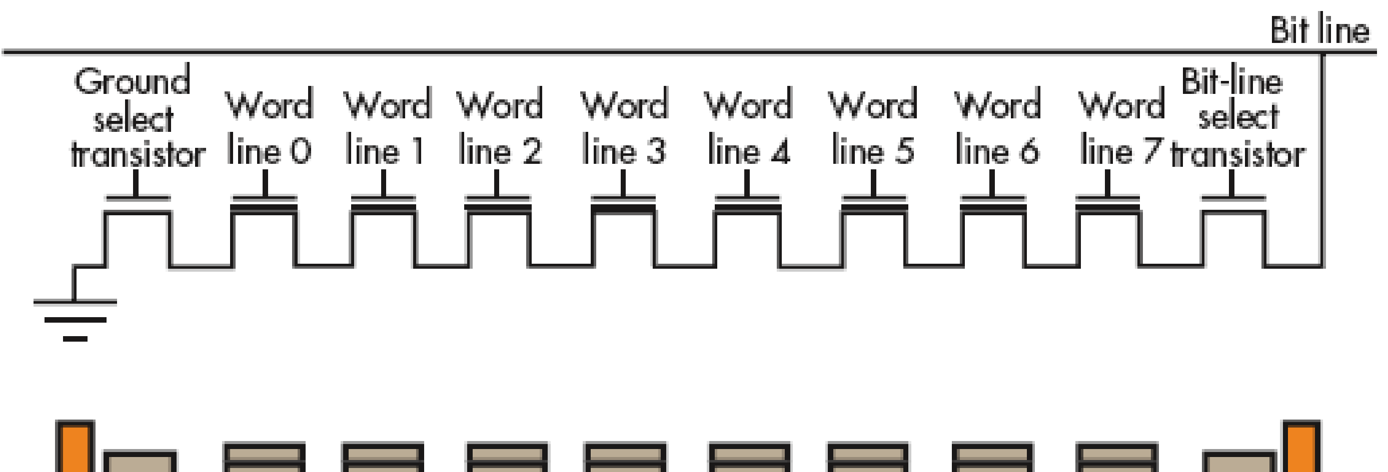
Tall floating gate with wrap-around control gate:

- Good coupling ratio
- Good isolation of adjacent floating gate
- Problematic high aspect ratio
- Larger cell pitch due to additional space to allow control gate wrapping

NOR and NAND Flash



NOR Flash: larger cells with faster access: code or cache



NAND Flash: smallest area with possible multiple-level cells (MLC): Data

Critical EEPROM Cell Parameters

- Cell layout area
- Retention time
- Program/erase (P/E) voltages
- Program/erase times
- Read access time
- Cycling endurance
- Bit error rate
- Cell capacity (eNVM is often single-bit per cell)

Group Exercise

Where is Flash memory applied? Why?

- Camera
- iPod
- Solid state drive (SSD)
-
-
-
-

Outline

- How CMOS and EEPROM works?
- Critical aspects for embedded EEPROM in RFID
  - Isolation and stabilization for high voltages
  - Manufacturing process constraints
  - Area and power
- Other embedded nonvolatile memories
  - Embedded Flash with technology modification
  - FeRAM
  - Fuse, anti-fuse and one-time programmable (OTP)
  - ReRAM

Possible Implementations: eNVM

eNVM

Control Gate (CG)

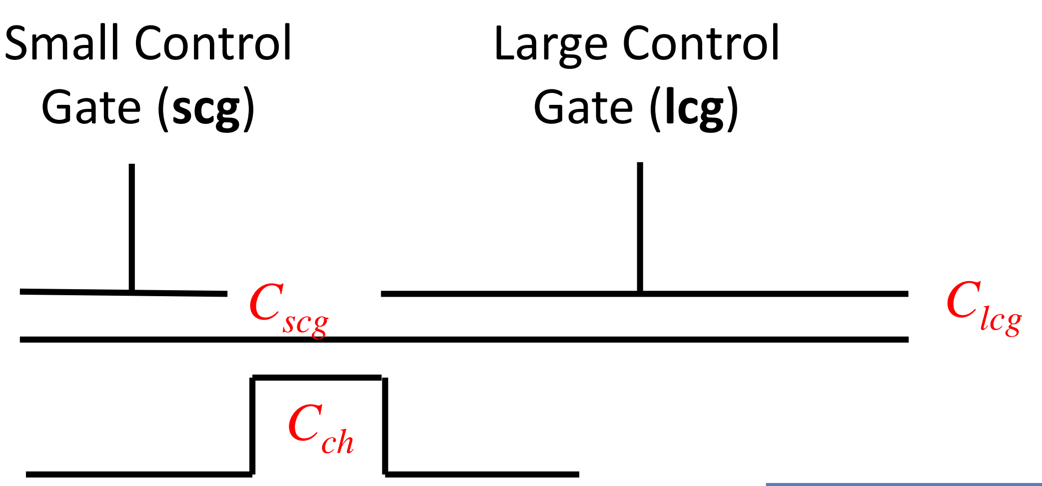
Large extended floating gate and control gate:

- Coupling ratio is a trade-off of cell area
- Positive voltage for program and negative voltage for erase
- Often dual-poly process is needed

As the cell area is much larger, NOR or even more sophisticated array structures (separating the read/write/erase WL and BL; additional control transistors) are all possible.

Unipolar Program/Erase Operations

Flash with Low Unipolar Program/Erase: $C_{ch}, C_{scg} << C_{lcg}$
Use of virtual floating ground



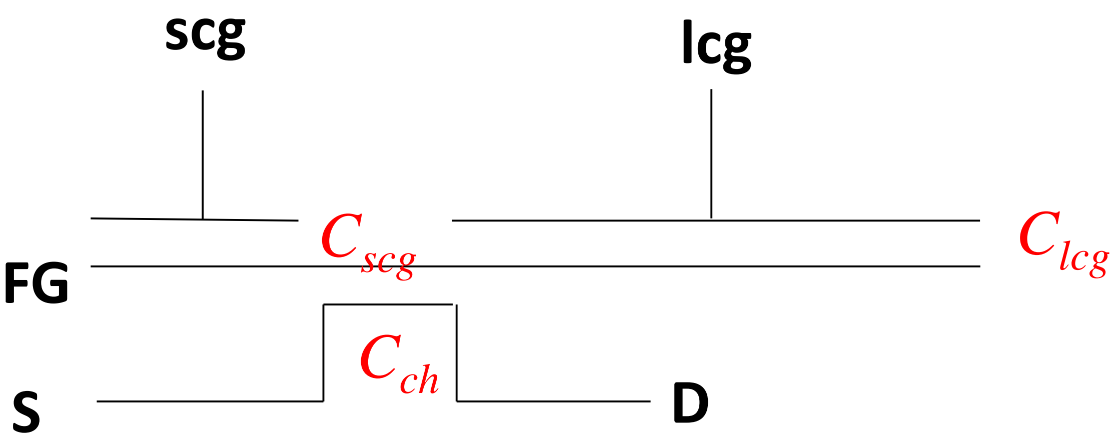
Sample Voltages

	V_{scg}	V_{lcg}	V_{ds}
Retention	0	0	0
Read	1.5V	1.5V	0.5V
Program	10V	10V	0
Erase	10V	0	0

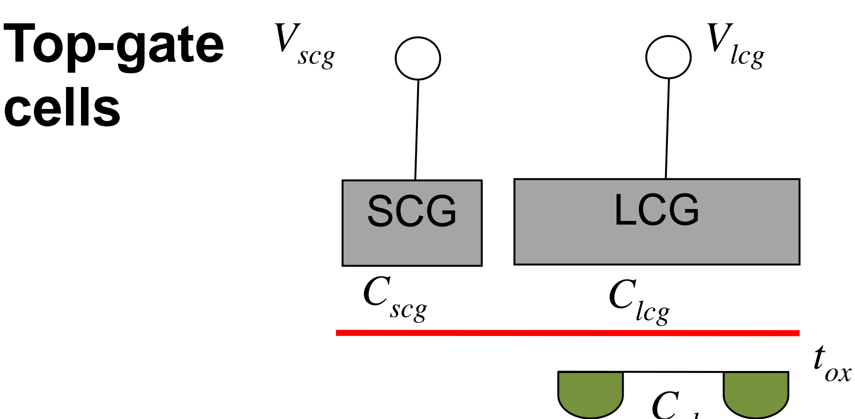
Electrostatic Equation for Neuromorphic EEPROM Cell

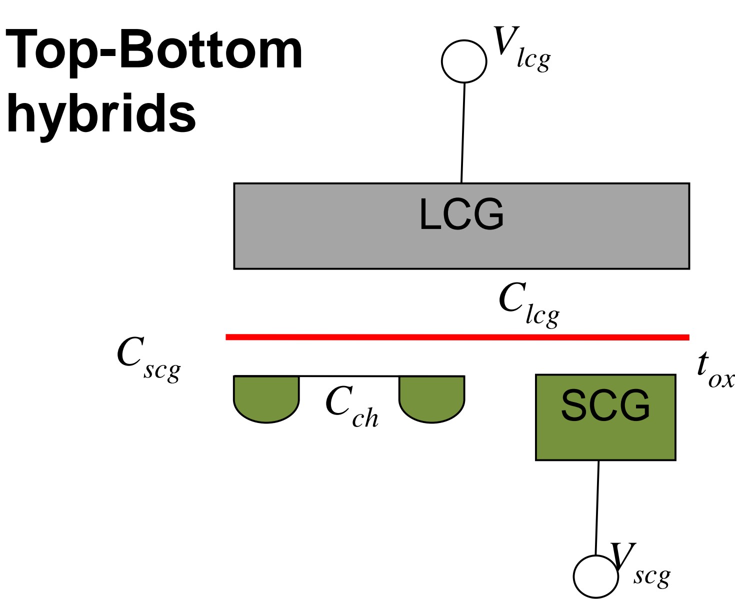
$$V_{FG} = \frac{1}{C_T}\left(Q + C_{scg}V_{scg} + C_{lcg}V_{lcg} + C_{ch}\frac{V_{ds}}{2}\right)$$

$$C_T = C_{scg} + C_{lcg} + C_{ch}$$

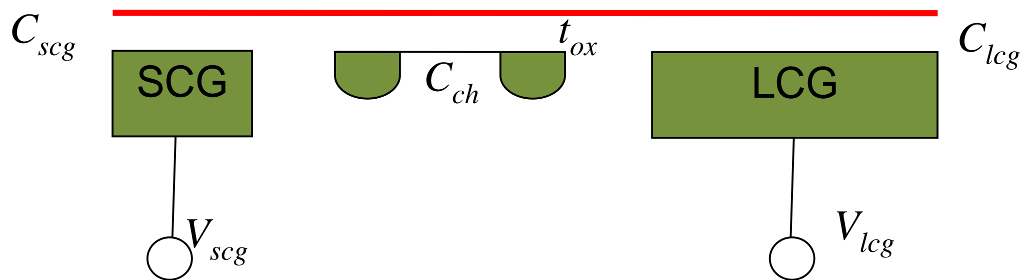


Unipolar Program/Erase Implementations



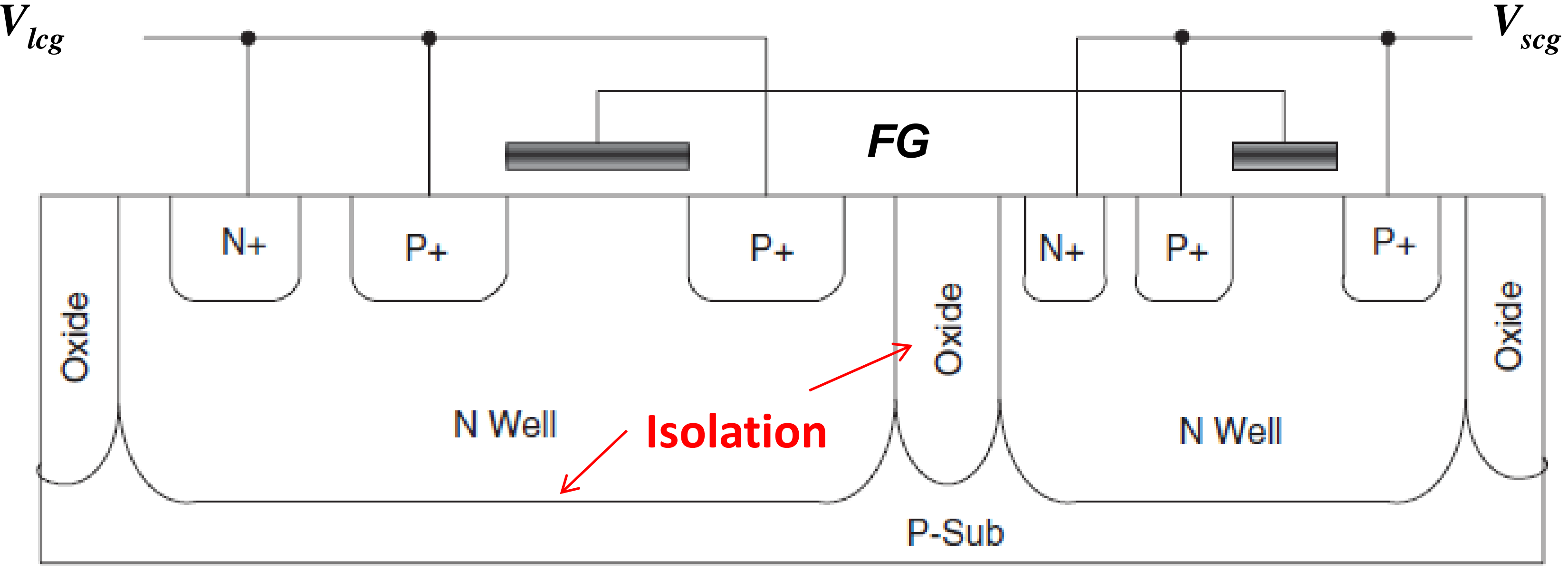


Inverted cells

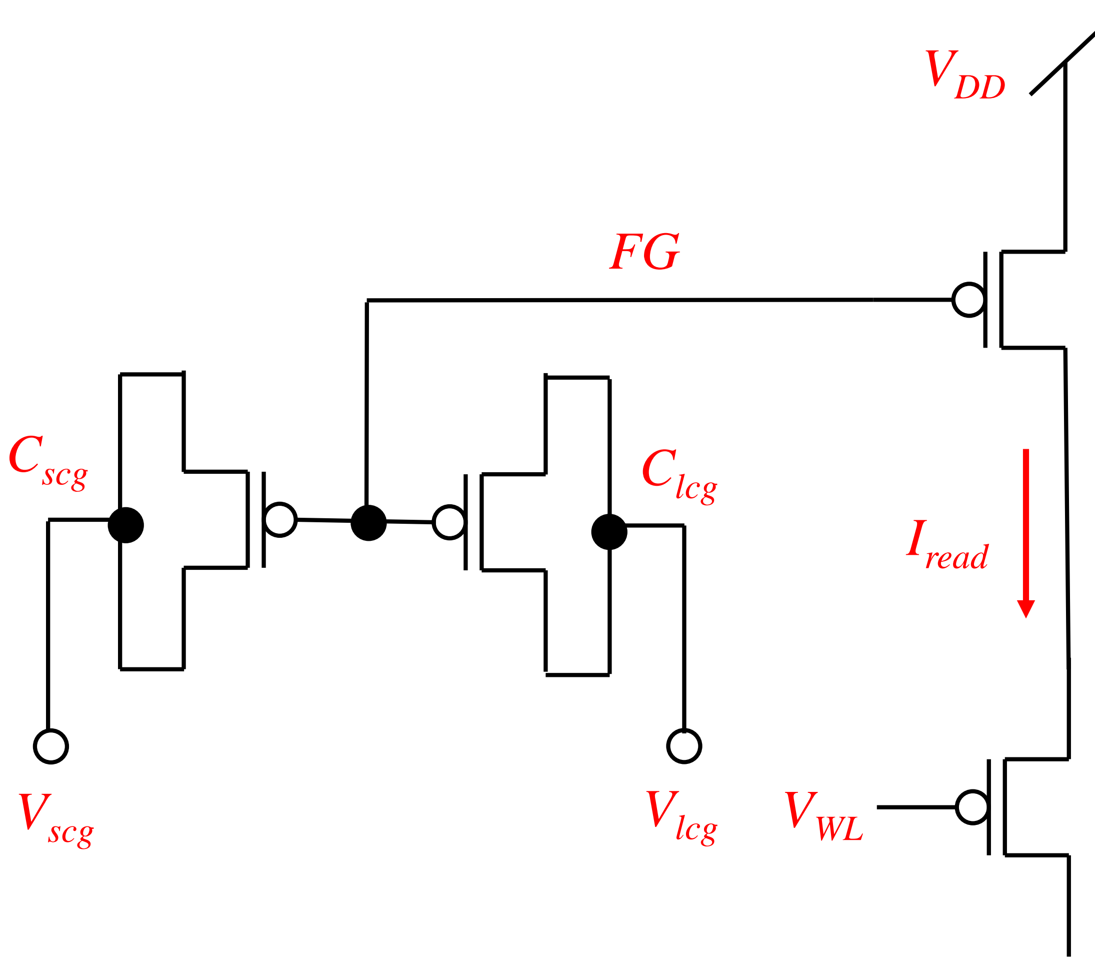


Electrical Isolation of Embedded EEPROM

- Floating gate should be entirely in poly1 (only the oxide around poly1 has high enough quality for retention)
- The control gate (can be poly2, active, M1, etc.) will see the highest P/E voltages
- High voltage needs good isolation (large design rules)
- High voltage is difficult to be on low impedance nodes

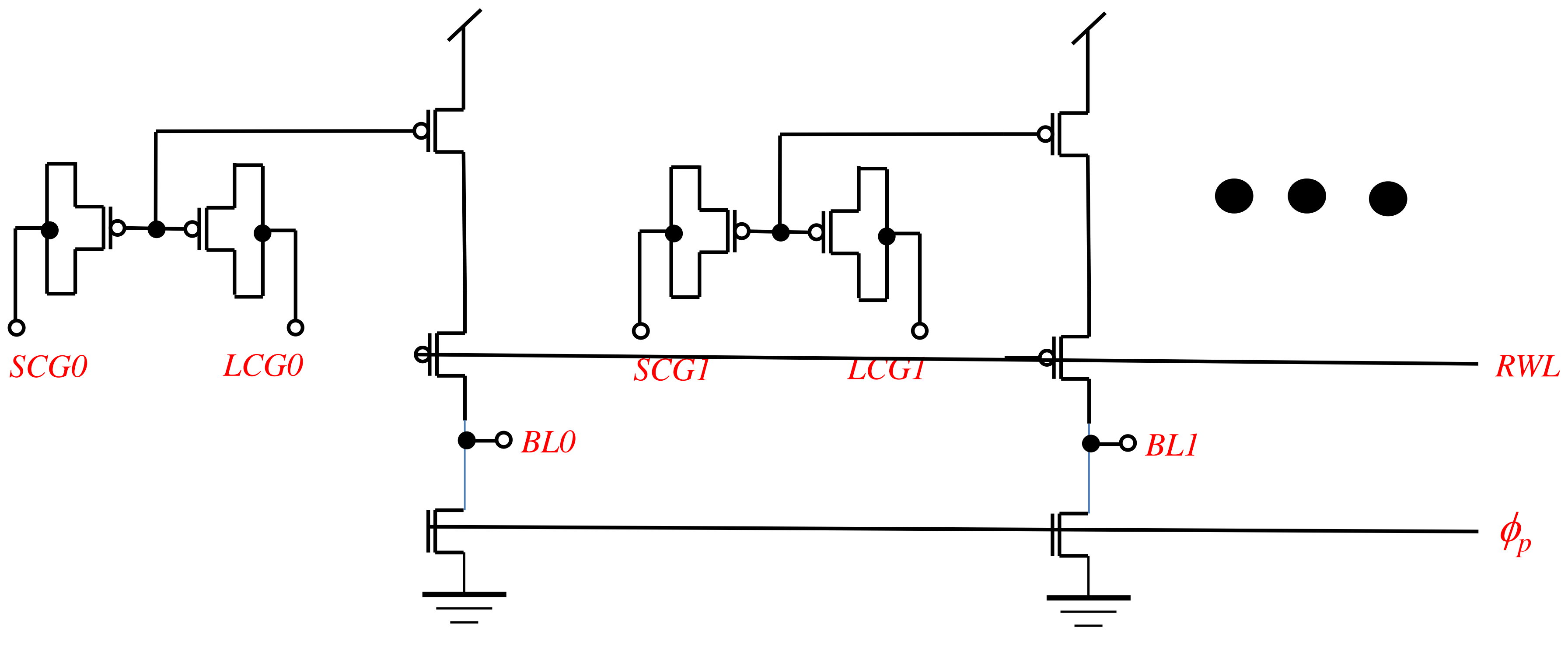


Single Poly Unipolar PMOS Cell



- Only gate nodes will see high voltages
- PMOS often gives better retention time as a trade-off for longer-time and higher-voltage P/E

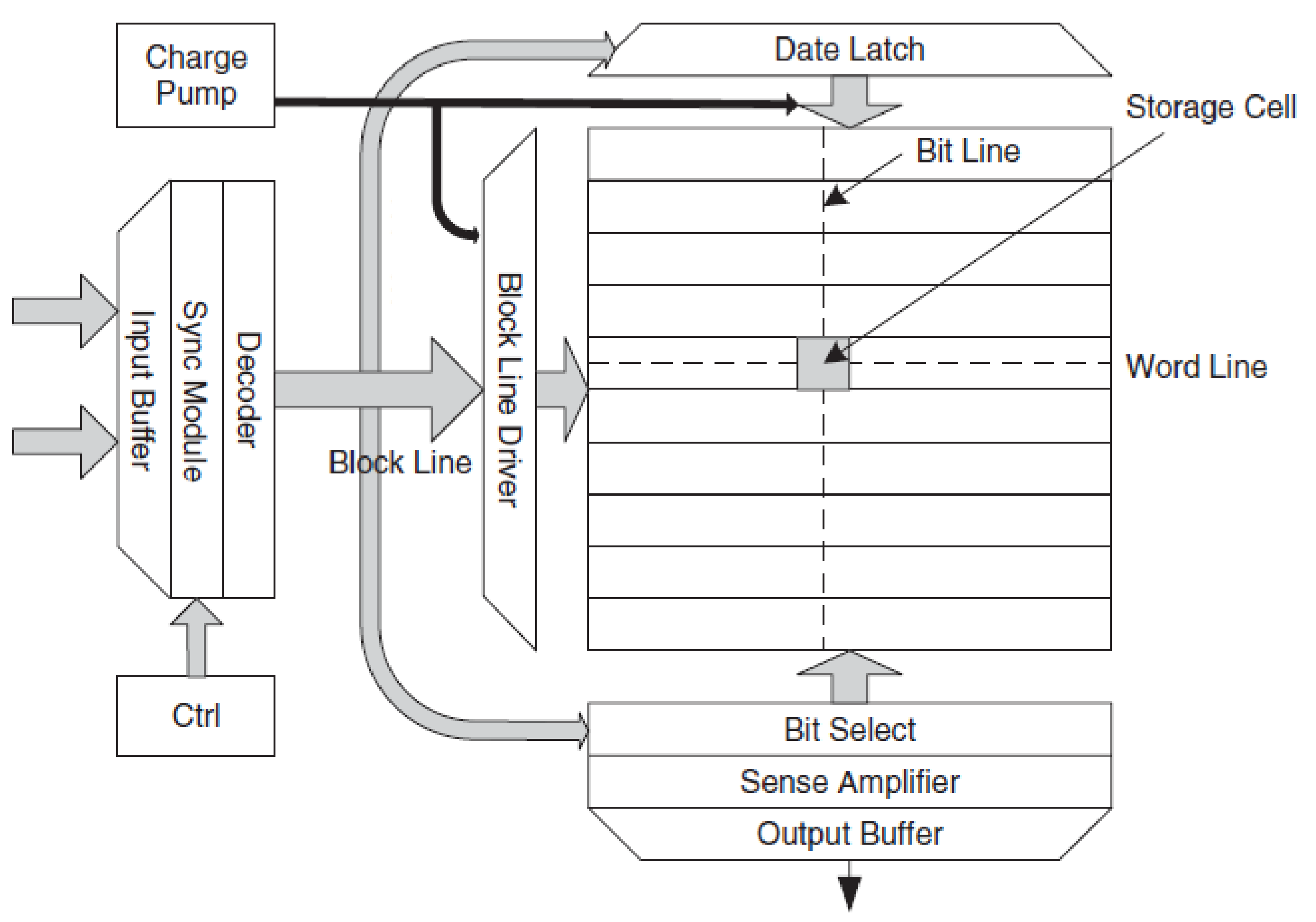
Single Poly Unipolar PMOS Array Circuits



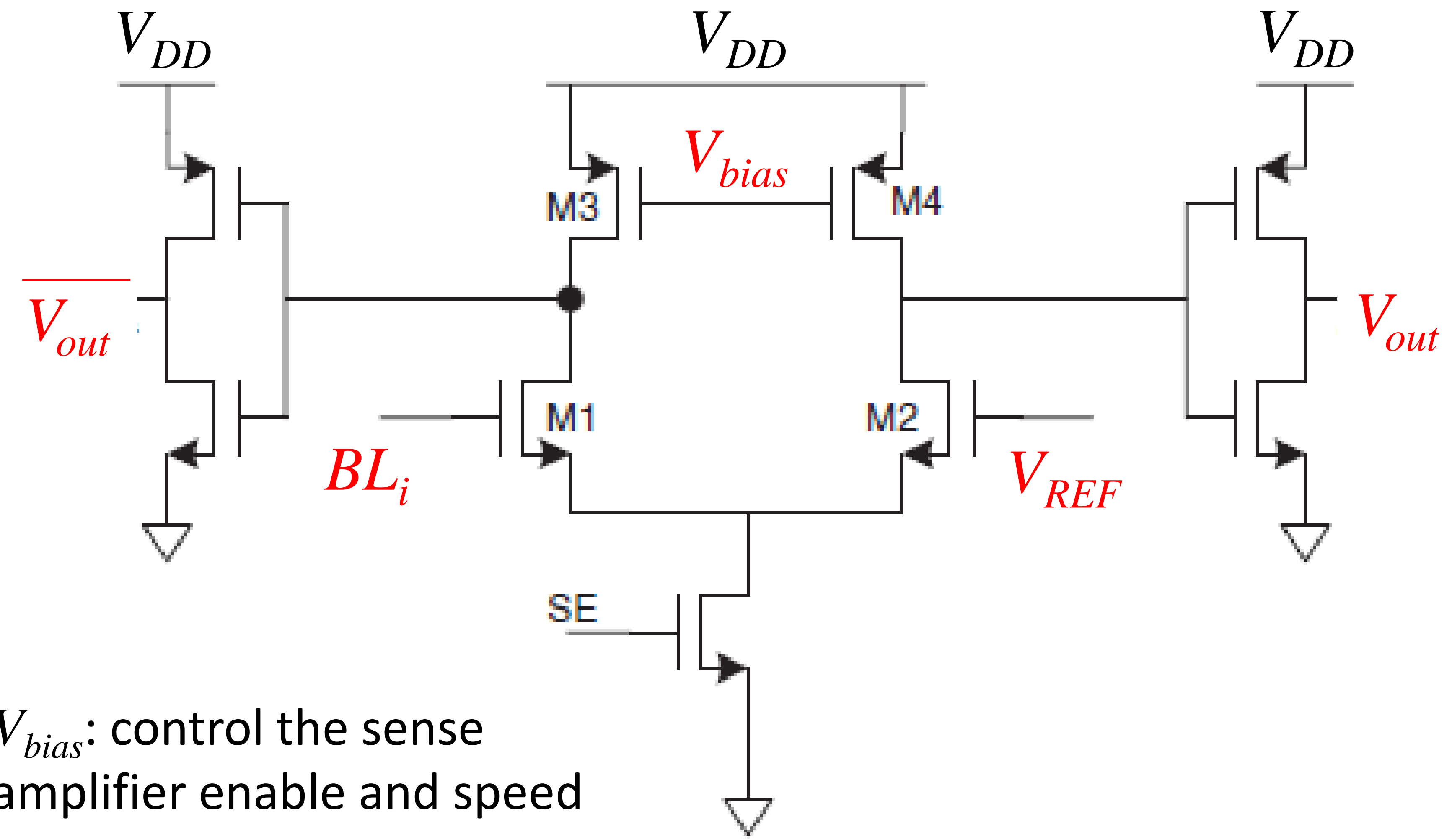
The array circuit here is similar to a "footless" ϕ_p NAND branch with RWL = 0 at the selected cell (RWL: Read Word Line).

SCG_i and LCG_i are logic functions of WWL and $DATA_i$ with high-voltage transmission gate (WWL: Write Word Line).

eNVM Array Architecture



Single-Ended Sense Amplifier



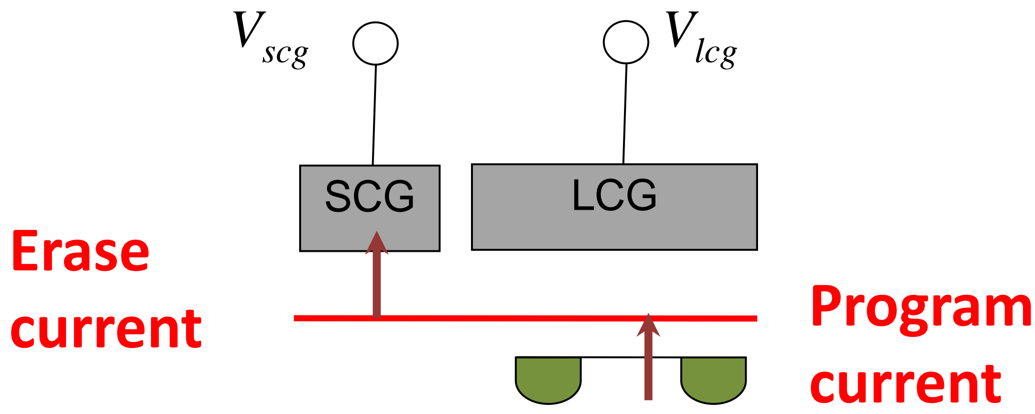
- V_{bias}: control the sense amplifier enable and speed
- V_{REF}: provide most sensitive differentiation between data levels of 0 and 1 in eNVM.

eNVM Architecture Variations

- As RWL (read word line, regular V_{DD} voltage) and WWL (write word line, high voltage feeding into V_{scg} and V_{lcg}) are separated, to reduce the power consumption in read, RWL can use a simple shifter to accomplish the N×1 array to a serial output.
- Separate access for read and P/E is especially popular for EPC Class 1 tags (96 – 512 bits) where the read distance can be further than P/E distance.
- Generation of accurate analog voltage for control of memory cell, sense amp, etc., is important and difficult, as timing of the tag chips still needs to be guaranteed.

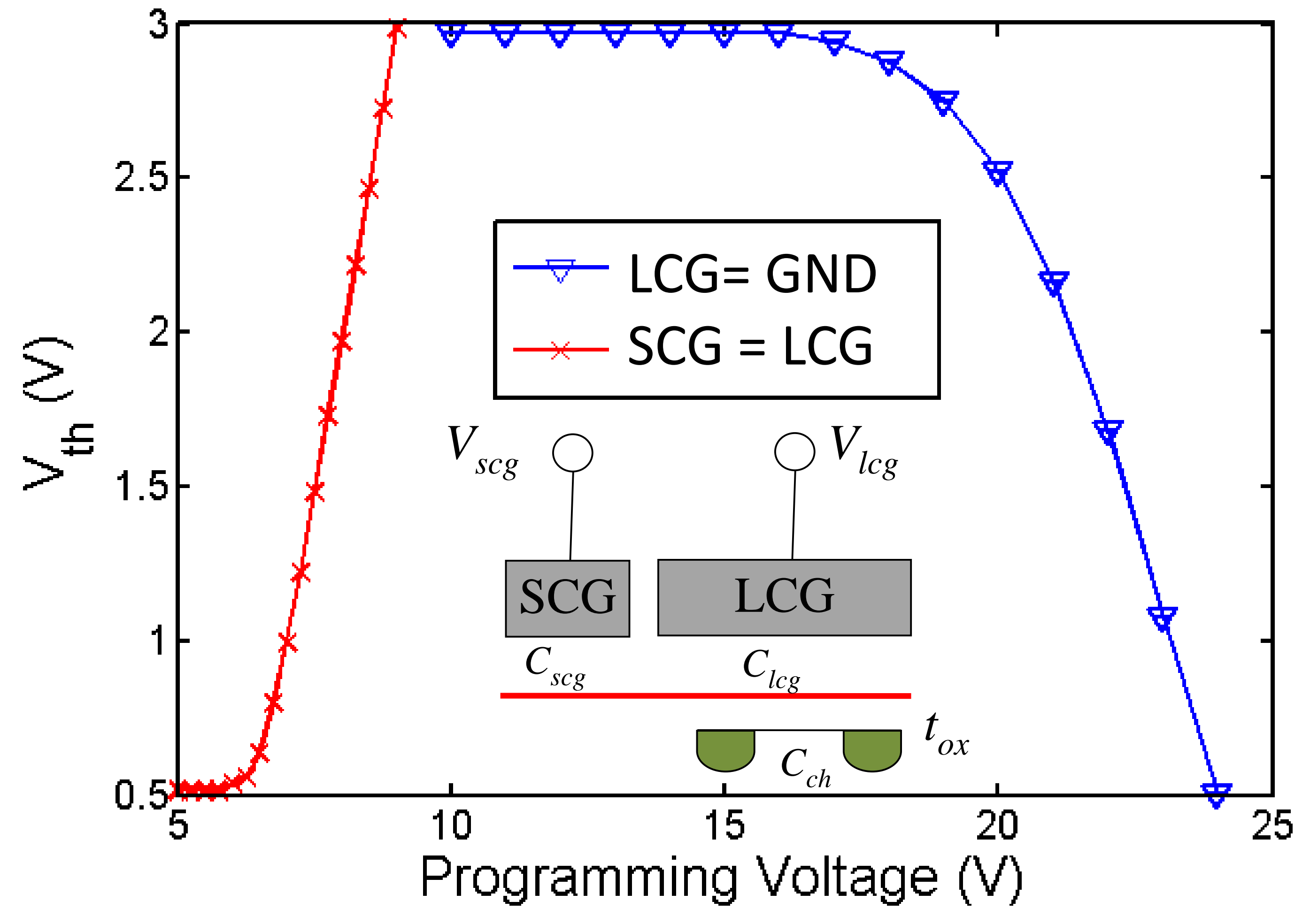
Cycling Endurance and Oxide Breakdown

- P/E pass some current through oxide and presently limit the EEPROM cycling endurance
- **Soft breakdown**: Amount of charge passing through oxide to generate interface states (shift V_{th}) or oxide traps (leaky path): stress induced leakage current (SILC)
- **Hard breakdown**: High voltage or current through oxide to cause contact material migration and cause a nanobridge path



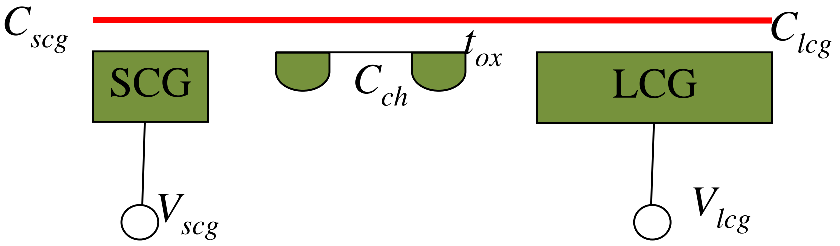
Top-Gate Unipolar EEPROM

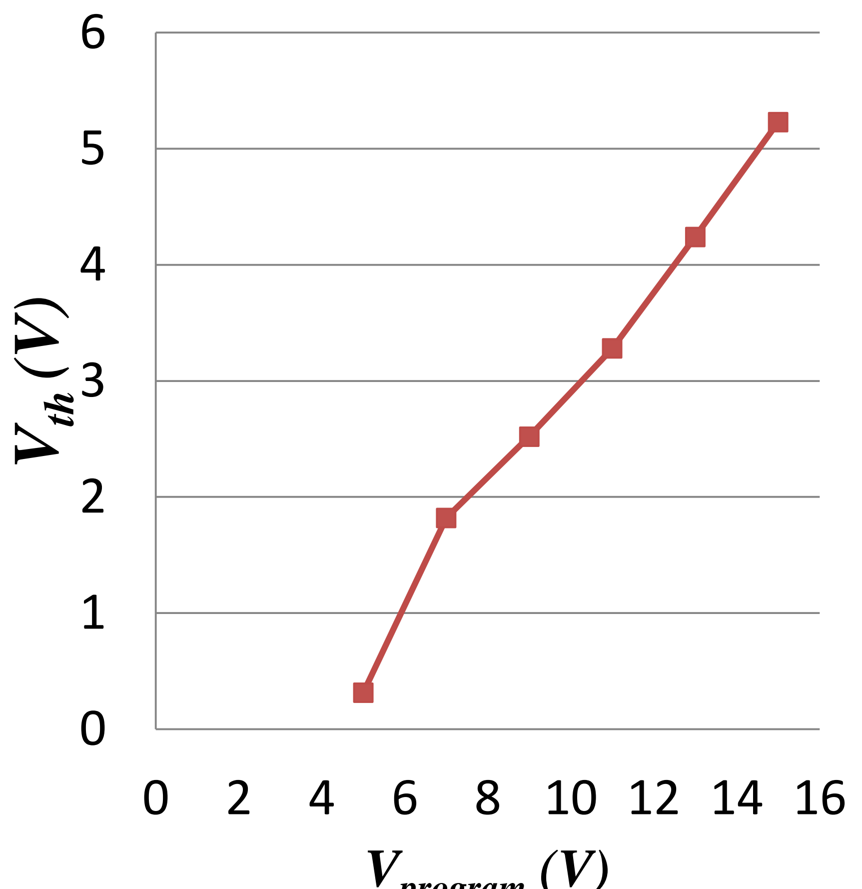
Transistor	Small Gate	Large Gate	Area Ratio
0.13 pF	0.62 pF	58 pF	92

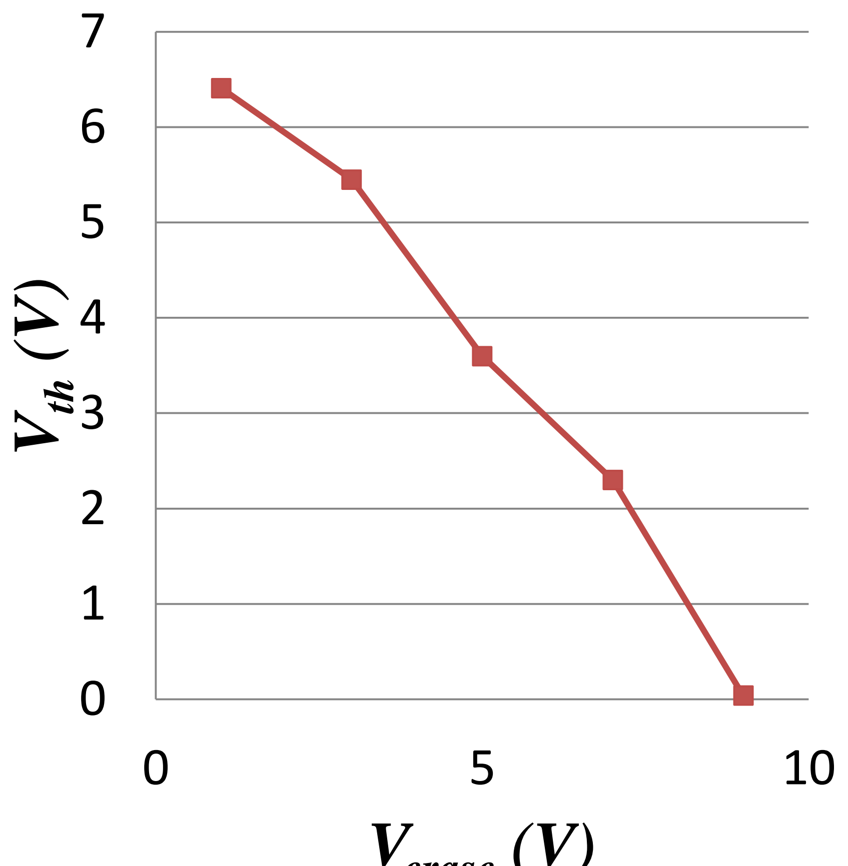


Single-Poly Inverted Cell Unipolar EEPROM

Transistor	Small Gate	Large Gate	Area Ratio
1 µm²	16 µm²	159 µm²	9.9







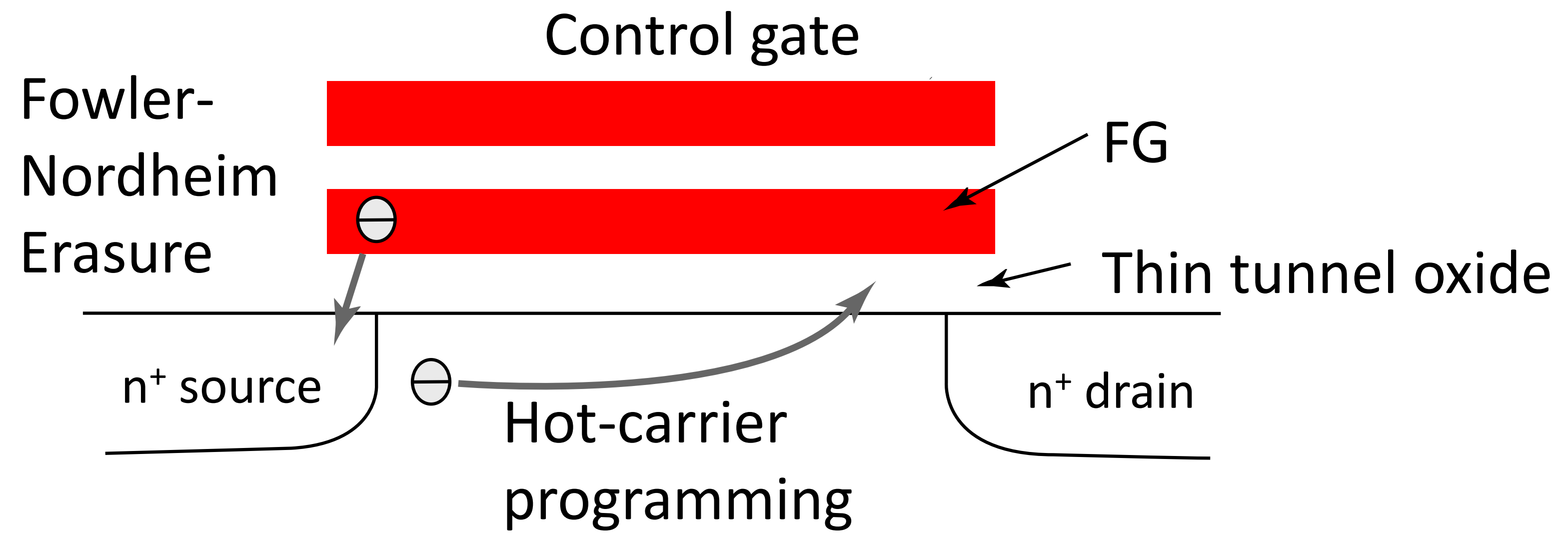
Outline

- How CMOS and EEPROM works?
- Critical aspects for embedded EEPROM in RFID
  - Isolation and stabilization for high voltages
  - Manufacturing process constraints
  - Area and power
- Other embedded nonvolatile memories
  - Embedded Flash with technology modification
  - FeRAM
  - Fuse, anti-fuse and one-time programmable (OTP)
  - ReRAM

Embedded Flash (eFlash) with Technology Modification and Better Charge Pumps

- The layout area of the previous embedded EEPROM is often too large (> 300 λ^2), not feasible for large number of bits due to overall cost (> 1kbits).
- When *the production can be customized* AND when *the charge pump can support pumping a low impedance node to above 5V,* there is another option of embedded EEPROM based on channel hot electron (CHE) injection.
- In comparison with storage-class Flash, this embedded flash memory often focuses more on reliability and performance than large storage space and bit cost, such as in automobiles or in precision control (Renesas is a big player)
- Often used in semi-passive RFID tags and in readers as power and cost are not so much constrained.

Channel Hot-Electron (CHE) Injection

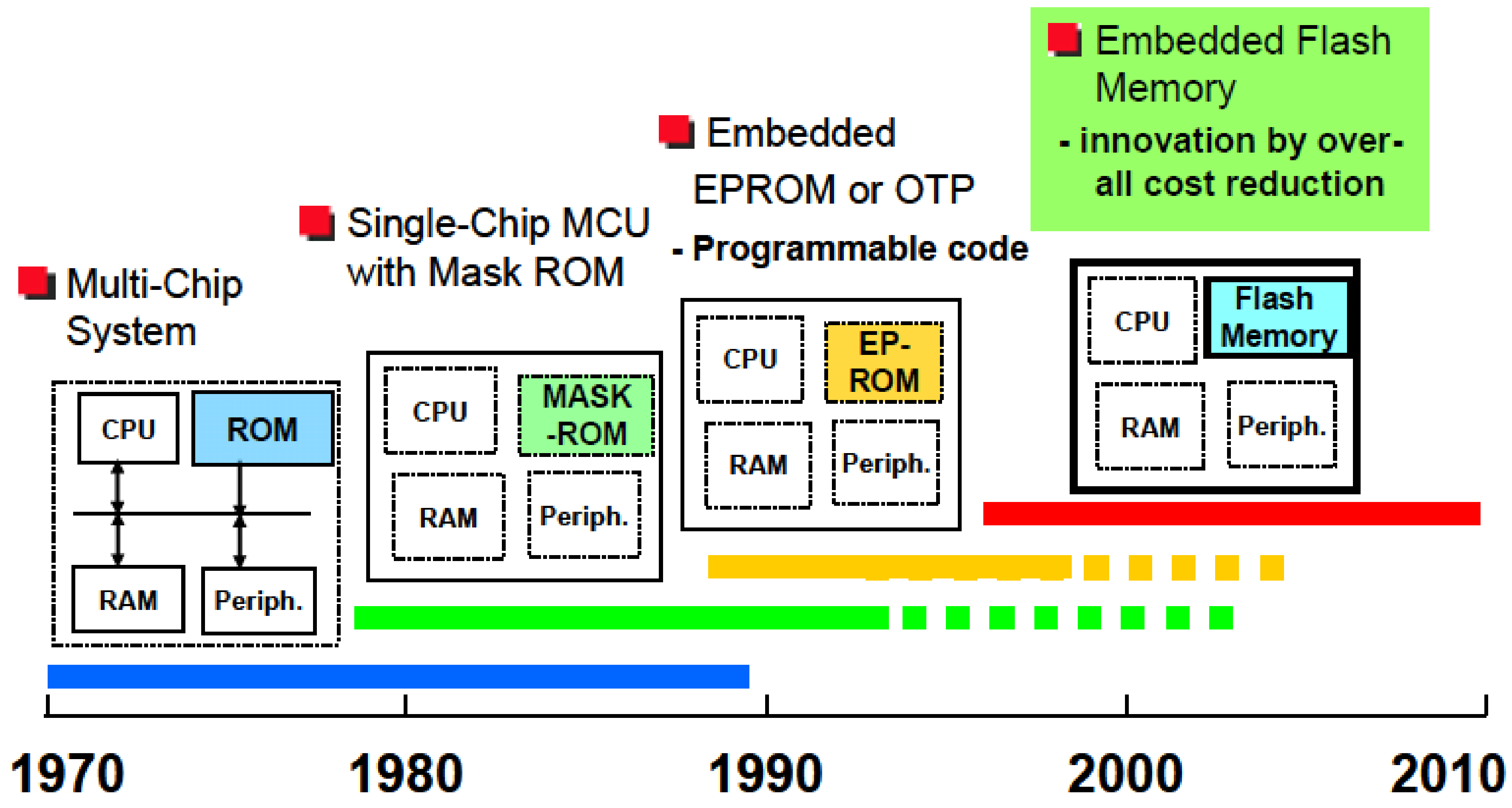


- Program by channel hot electon injection: ex., V_{GS} = 6V, V_{DS} = 5V.
- Erase by FN Tunneling: ex., V_{GS} = - 19V, V_{DS} = 0V.
- Substrate-assisted HCI can reduce gate and drain voltages to 4V.
- It is much harder to provide a high drain voltage (low impedance)
- High erase voltage is needed anyway (difficult to provide both hot electrons and hot holes in the same cell)

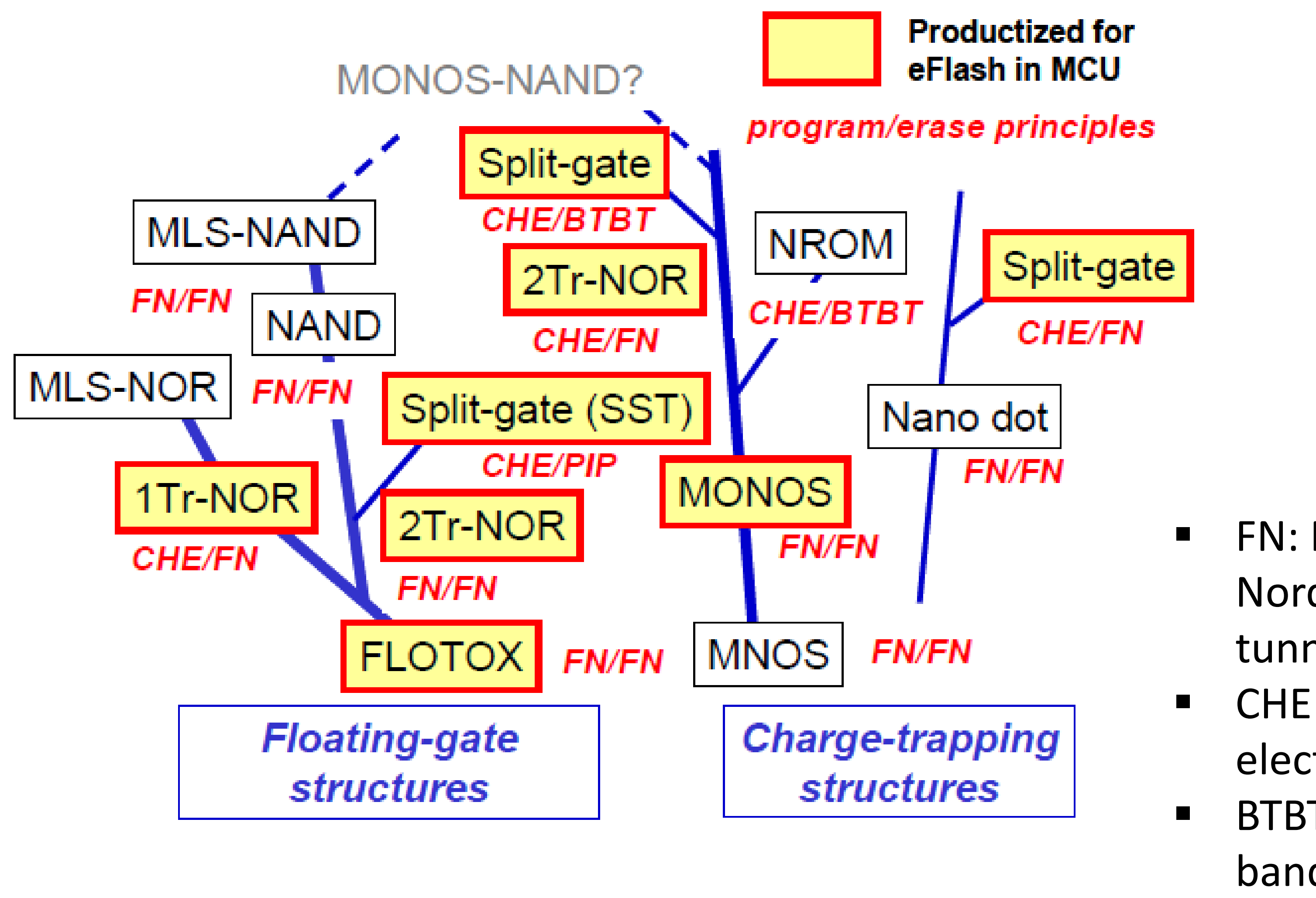
Custom vs. Fully-CMOS Compatible

SMIC 180nm Mixed-Signal HV Process	Custom EEPROM (FN or Hot Carrier)	HV CMOS Fully Compatible
Cell area	3.96 μm^2	110 μm^2
Number of masks	**28** (16 frontend + 12 backend)	**19** (7 frontend + 12 backend)
Memory area (512 bits) on 0.5mm^2 tag chip	0.41%	11.3%
Cost increase from area	0	10.89%
Cost increase from masks (10^7 tag chips)	30%	0
Total cost increase	19.1%	0

eFlash in Microcontroller Unit (MCU)



Possible P/E Mechanisms in NV Memory



- FN: Fowler-Nordheim tunneling
- CHE: channel hot electron
- BTBT: Band-to-band tunneling
- PIP: poly-interpoly-poly

eFlash Structure Variations: 1

Type	1T NOR	1.5T SuperFlash	2T NOR
Program	CHE	CHE/SSI	FN
Erase	FN (poly-sub)	FN (poly-poly)	FN (poly-sub)
Device Structure	Control gate Floating gate (N+-poly) N+ N+ S D P-sub	Floating gate Word-line N+ N+ S D P-sub	Control gate Access gate Floating gate N+ N+ N+ S D P-sub
Main Advantage	High density	Fast program	Low-power P/E

- SSI: Secondary substrate injection

eFlash Structure Variations: 2

Type	1T SONOS (NROM)	1.5T Nanodots	2T SONOS (PMOS)
Program	CHE	CHE/SSI	CHE
Erase	HH	FN	FN
Device Structure	O-N-O stack Control Gate Bit#1 Bit#2 N+ N+ P-substrate	Control Gate Access Gate Nano dots N+ N+ S D P-substrate	Access Gate Control Gate P+ P+ P+ S N-well D O-N-O stack P-substrate
Main Advantage	2 bits/cell	Fast, low-power P/E	Low-power P/E

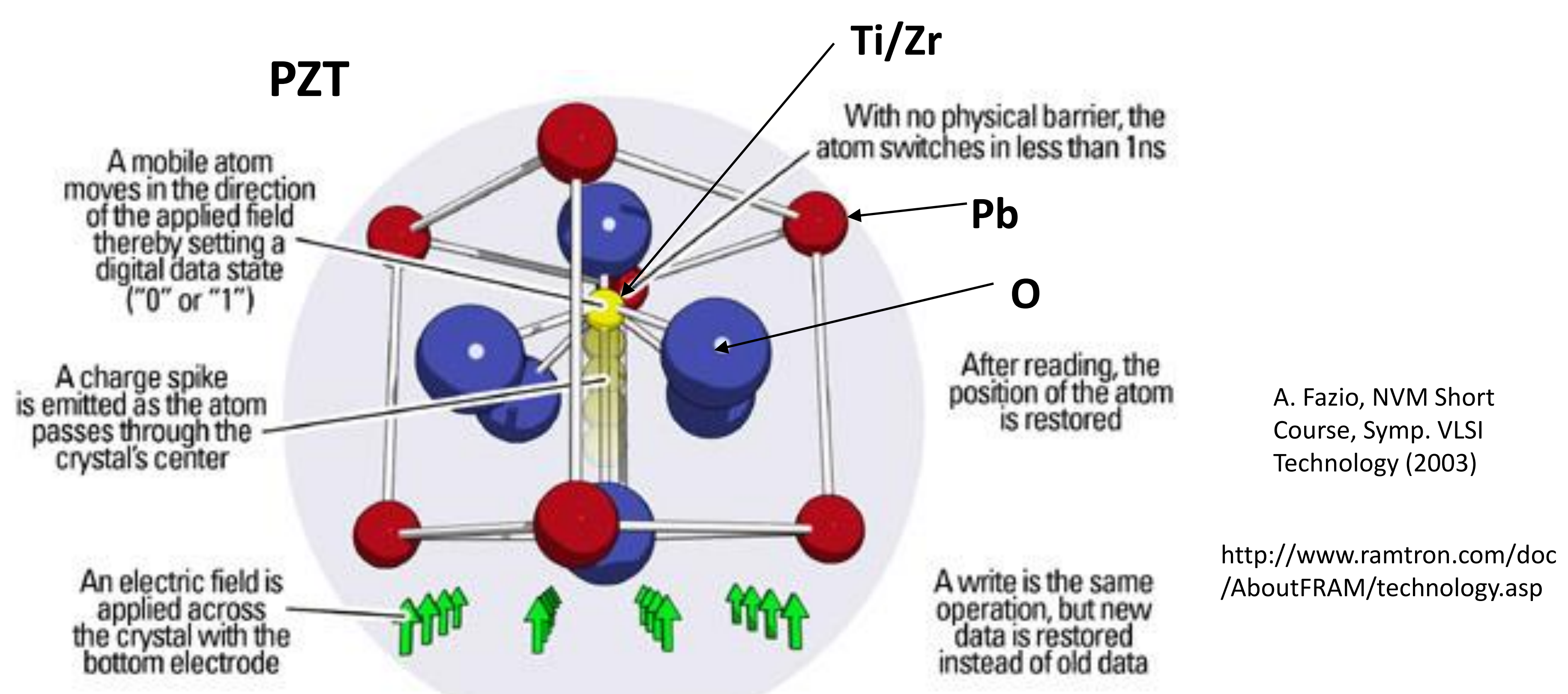
- HH: Hot hole (channel accumulated hole)

Outline

- How CMOS and EEPROM works?
- Critical aspects for embedded EEPROM in RFID
  - Isolation and stabilization for high voltages
  - Manufacturing process constraints
  - Area and power
- Other embedded nonvolatile memories
  - Embedded Flash with technology modification
  - FeRAM
  - Fuse, anti-fuse and one-time programmable (OTP)
  - ReRAM

Ferroelectric Atomic Structure

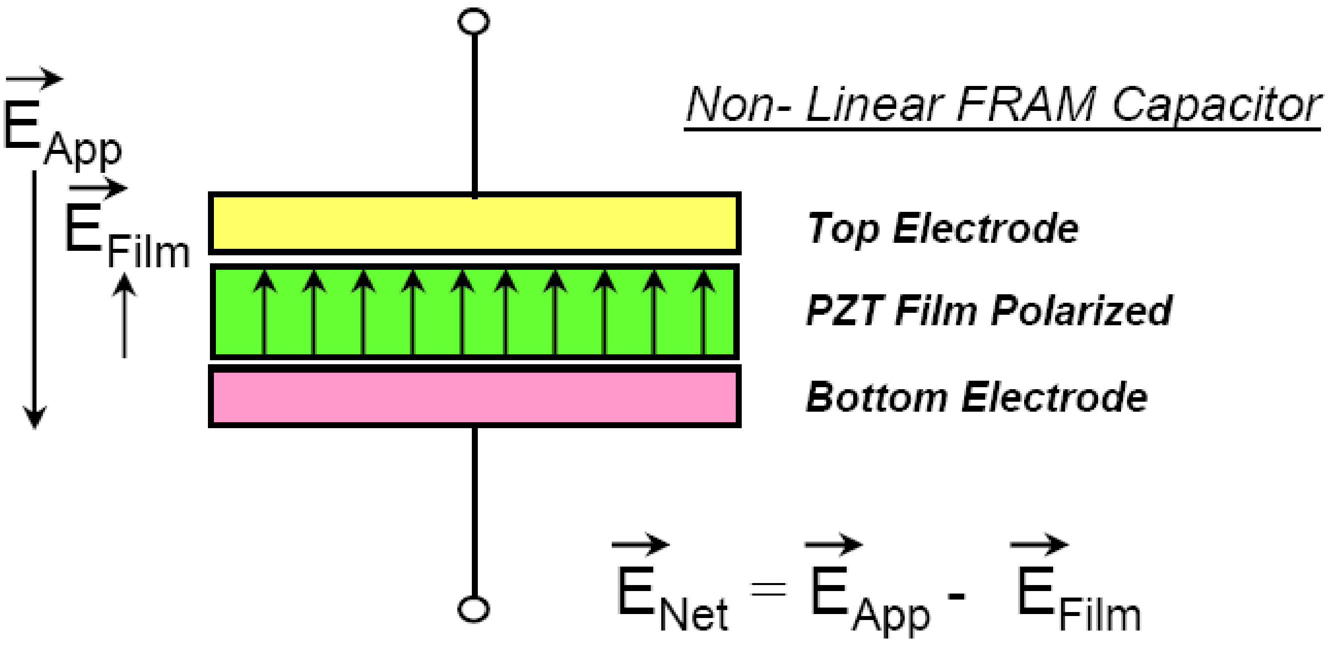
- Stores information as position of the atom (which gives a different polarization) – atom has bi-stable position
- Cyclic electric fields give a hysteresis, similar to ferromagnetics
- Read by sensing displacement current when voltage is applied

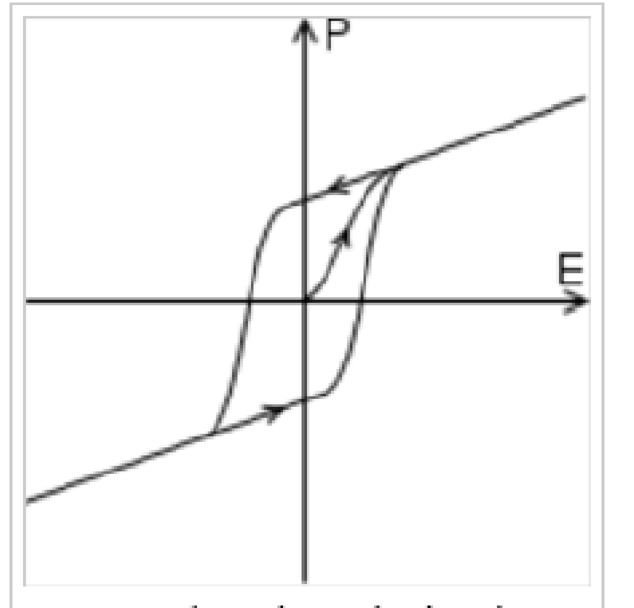


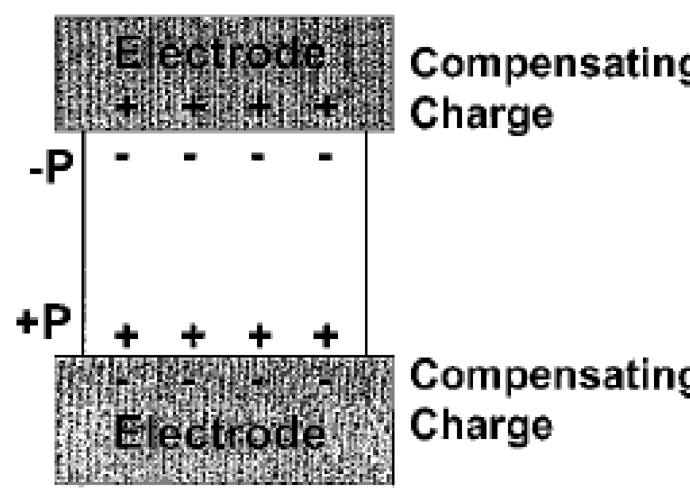
A. Fazio, NVM Short Course, Symp. VLSI Technology (2003)

http://www.ramtron.com/doc/AboutFRAM/technology.asp

Ferroelectric MIM Capacitor





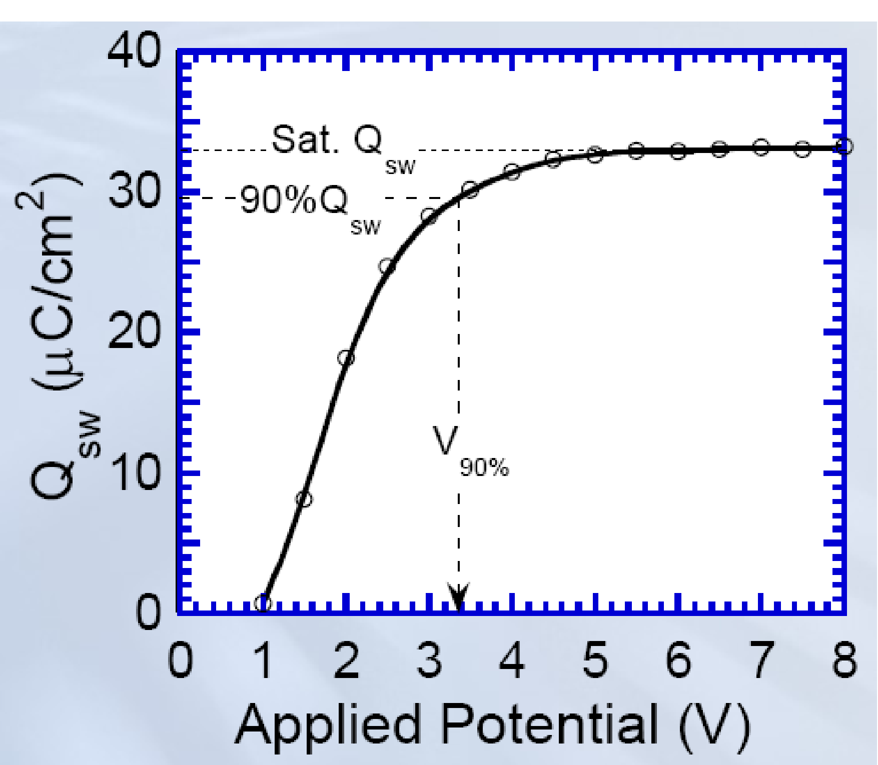


No depolarization field when $E_{App} = 0$ due to image charge in top and bottom electrodes.

T. P. Ma, et al, *IEEE EDL* 23, 386(2002)

Ferroelectric Capacitor Characteristics

- Data stored by electrically switching ions between stable positions in crystal structure



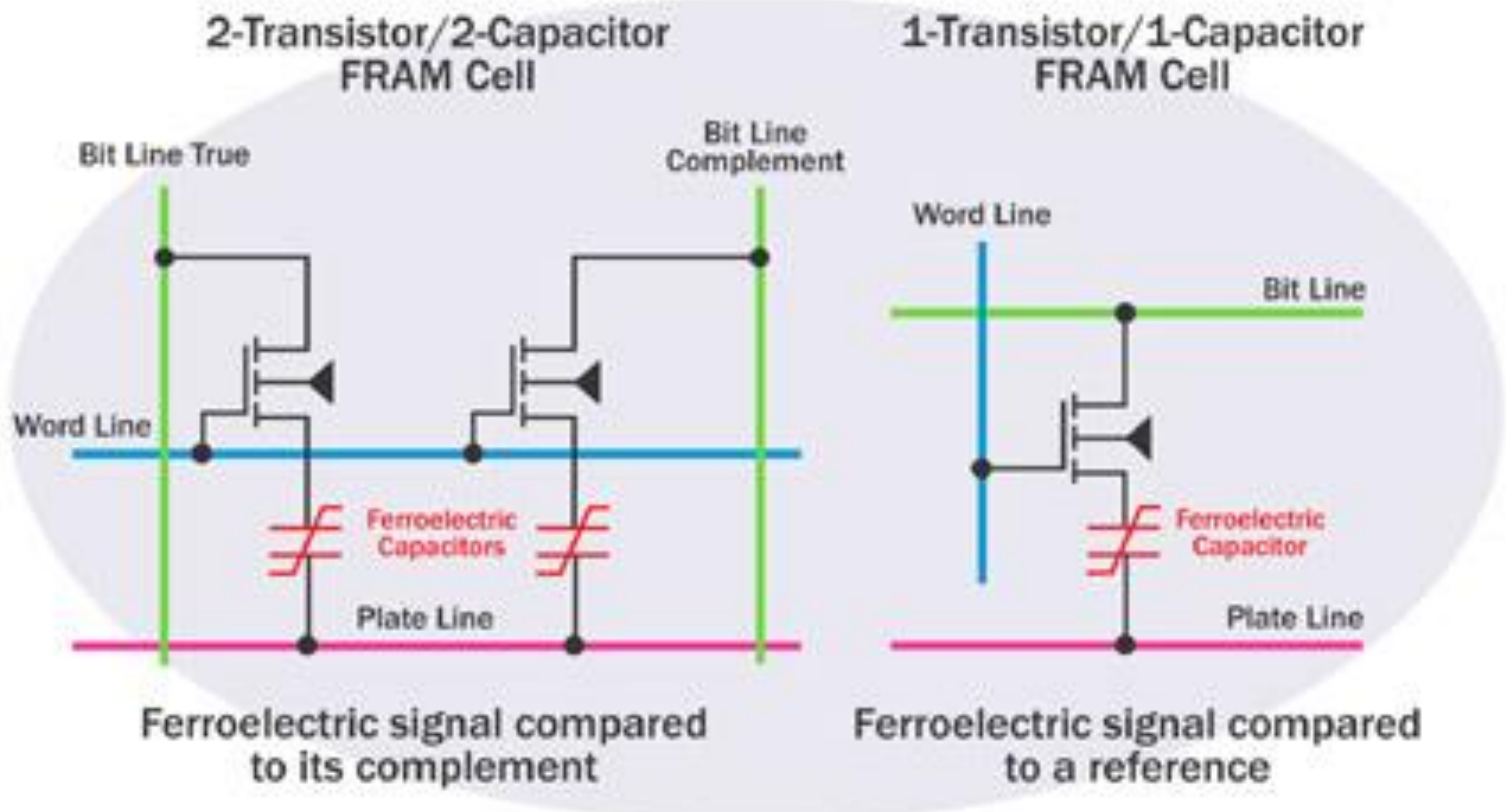
Polarization
Electrical Field
$+P_r$
$-P_r$
$-E_c$
$+E_c$
Q_{sw}

Millivolts
Time (usec)

G. Fox et al., SRC – NVM Forum, Stanford University, Nov. 18, 2004

FeRAM Circuits

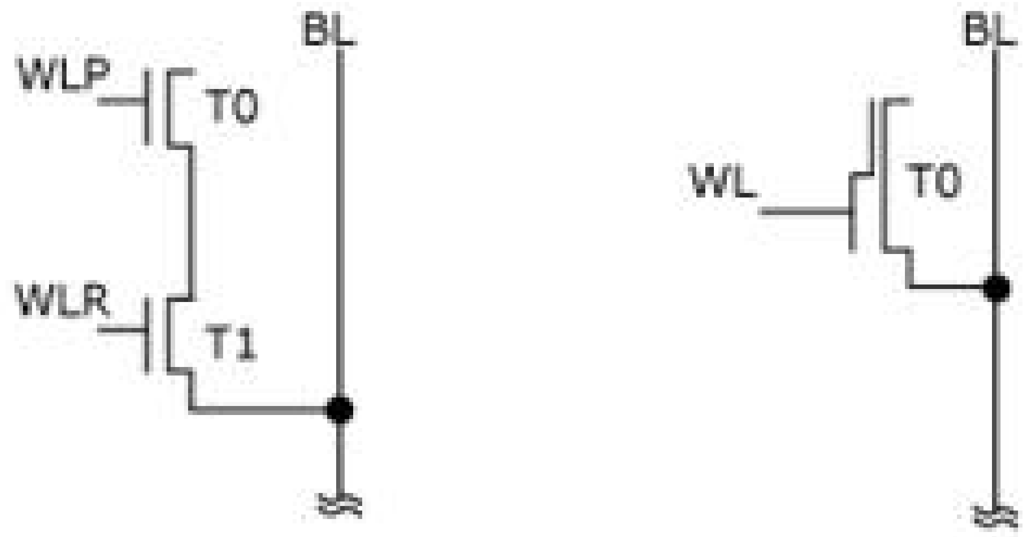
- 2T2C → 1T-1C
- Planar capacitor → on plug → 3D



http://www.ramtron.com/doc/AboutFRAM/technology.asp

Fuse-Based eNVM

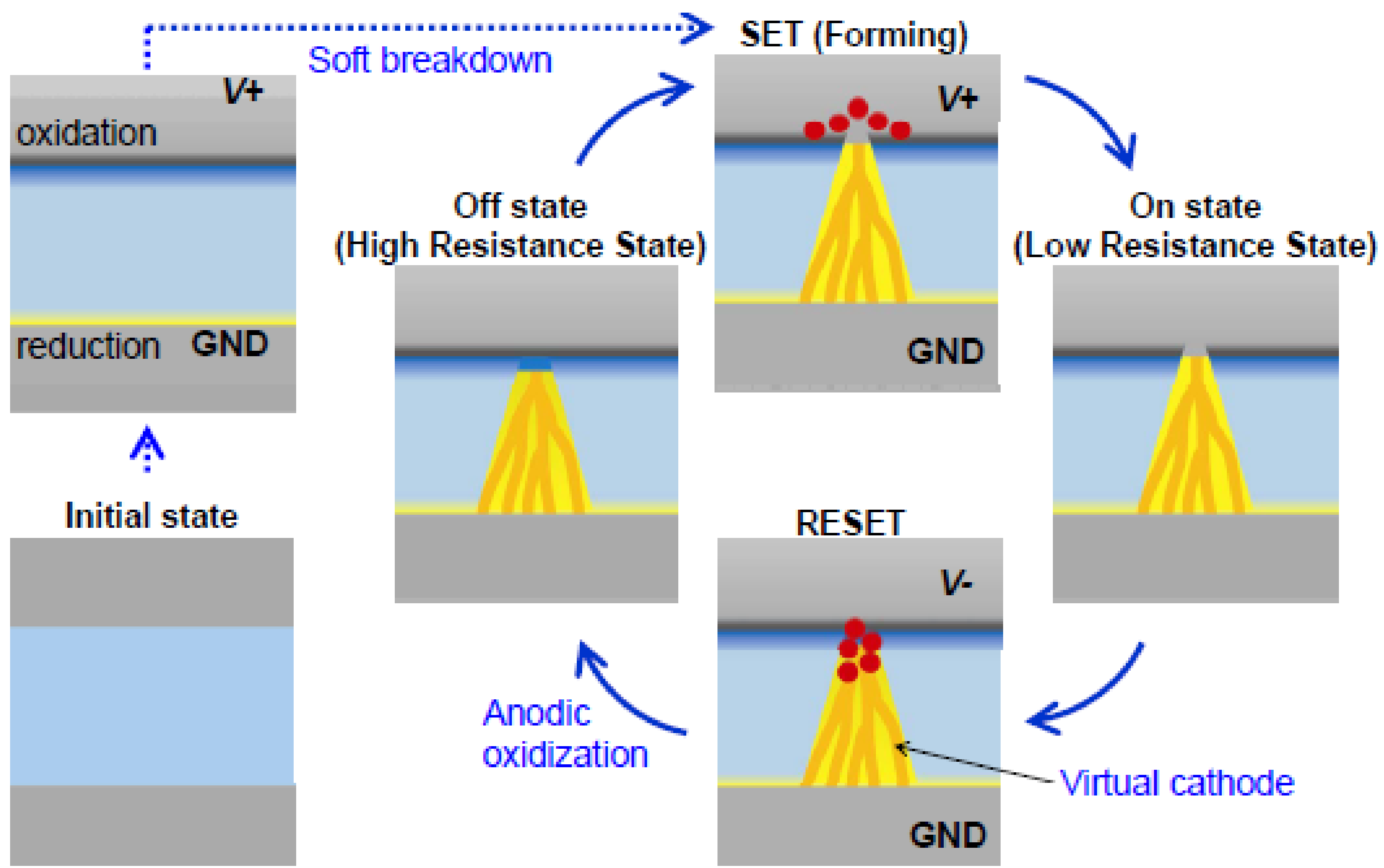
- Programmable read-only memory (PROM) or one-time programmable non-volatile memory (OTP NVM) sets each bit by a fuse or antifuse, most often by oxide breakdown (invented by Wen Tsing Chow in 1956 for storing target in missile control).
- The key difference from a strict ROM is that the programming is applied after the device is constructed.
- Cell area is much smaller than other eNVM.



Metal Oxide ReRAM

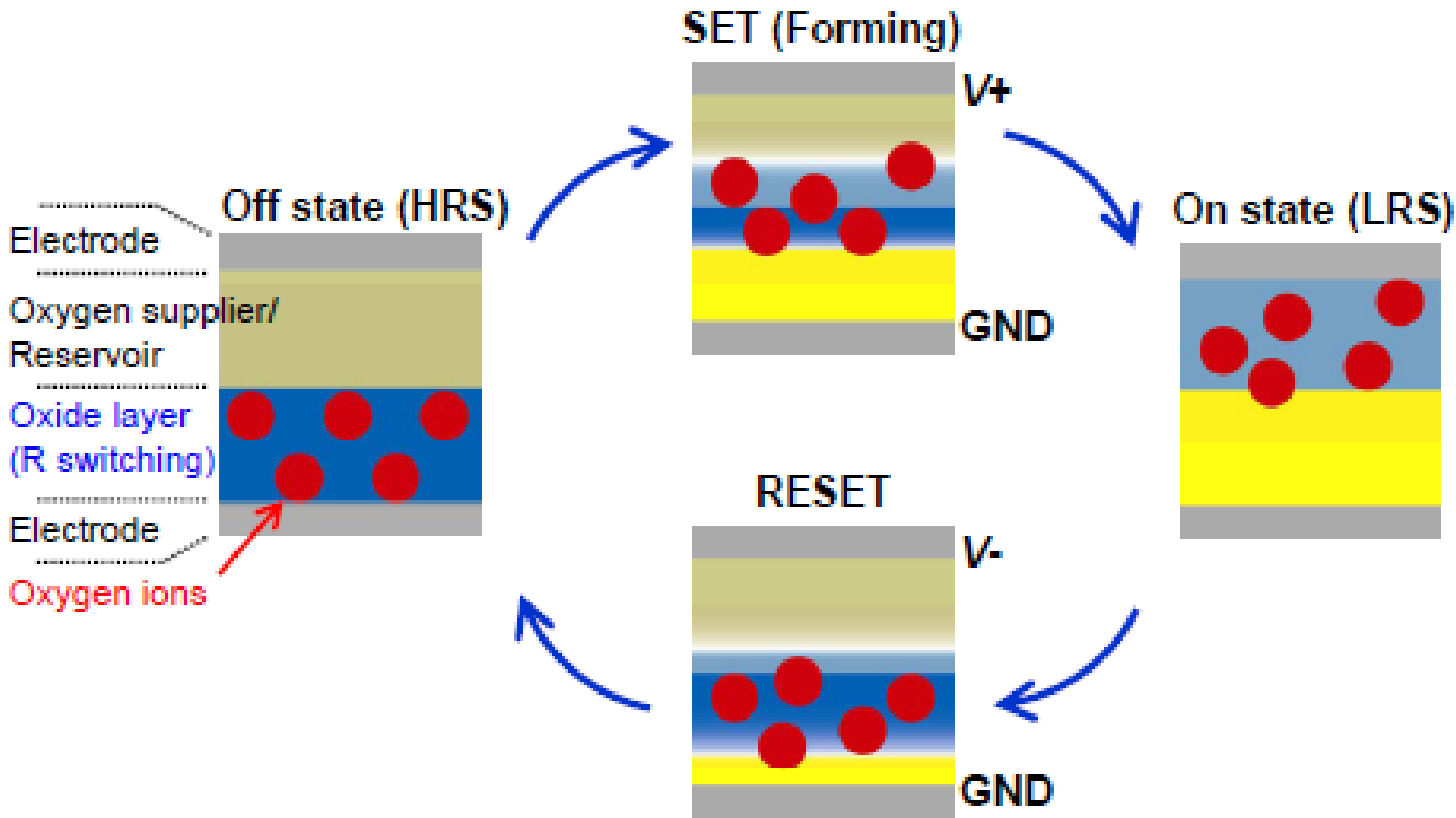
- Memory switching observed in many metal oxides
  - NiO_x, TiO_x, Nb_2O_5, Al_2O_3, Ta_2O_5, CuO_x, WO_x
  - Cr doped perovskite oxides: $SrZrO_3$, $(Ba,Sr)TiO_3$, $SrTiO_3$
  - Cu doped MoO_x, Al_2O_3, ZrO_2, Al doped ZnO
  - $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO)
  - Ta/CoO/Pt
- Top and bottom electrodes play major roles
- Physical mechanisms are often in debate
  - Conductive filaments between the metal electrodes
  - Interface-charge-modulated metal-semiconductor barrier height
  - Oxygen drift
- Most often, on-state resistance and set current not correlated to device area
- Most often, off-state resistance and reset current area dependent

Nano-Filament Model



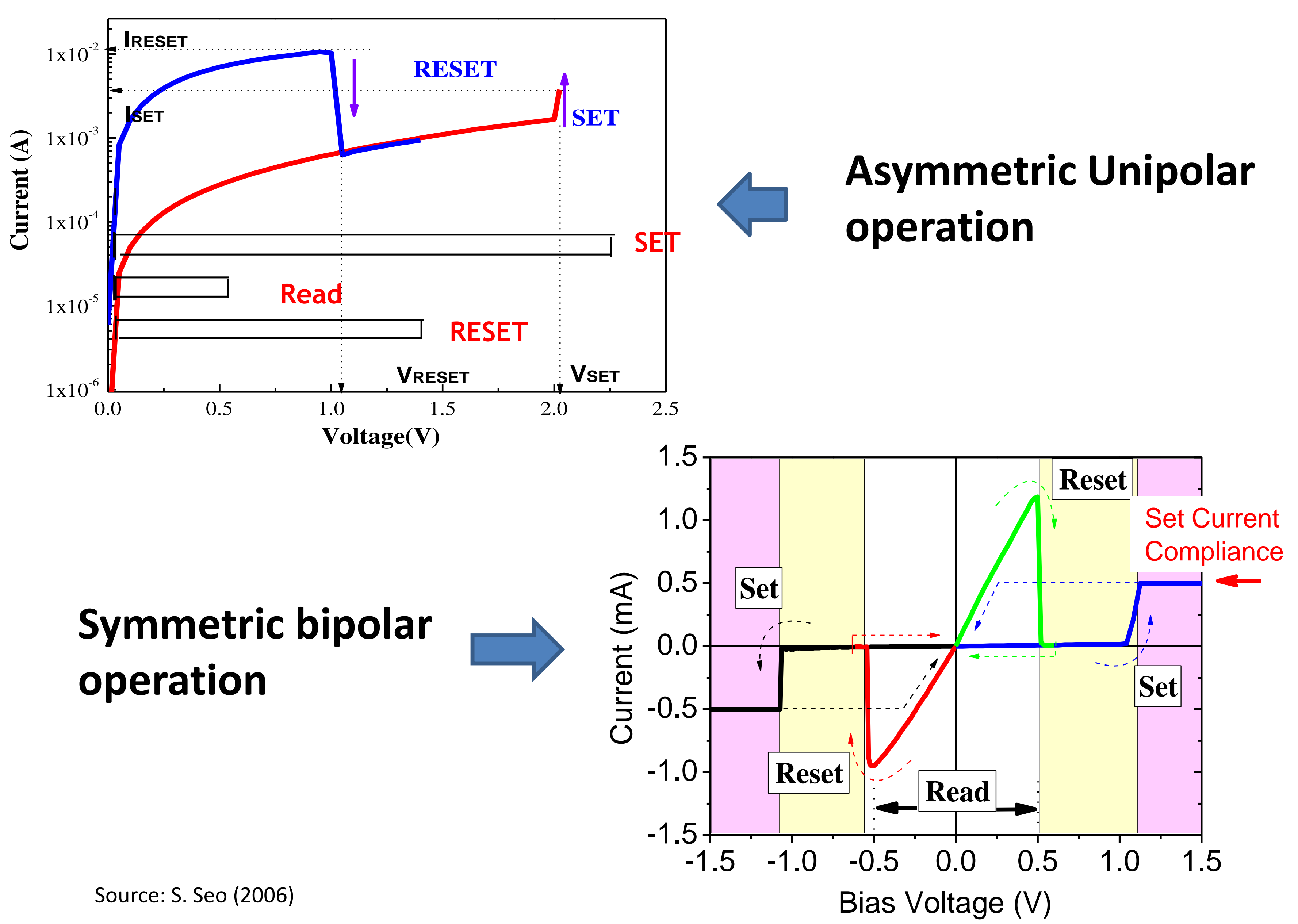
H. Akinaga and H. Shima, AIST, Japan, Nonvolatile Memory Workshop, 2009.

Oxygen Drift Model



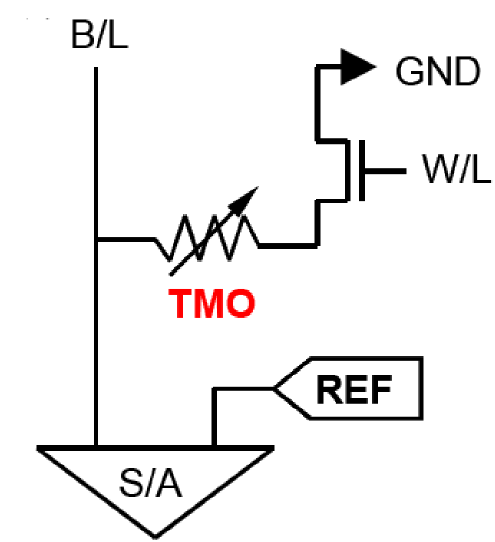
H. Akinaga and H. Shima, AIST, Japan, Nonvolatile Memory Workshop, 2009.

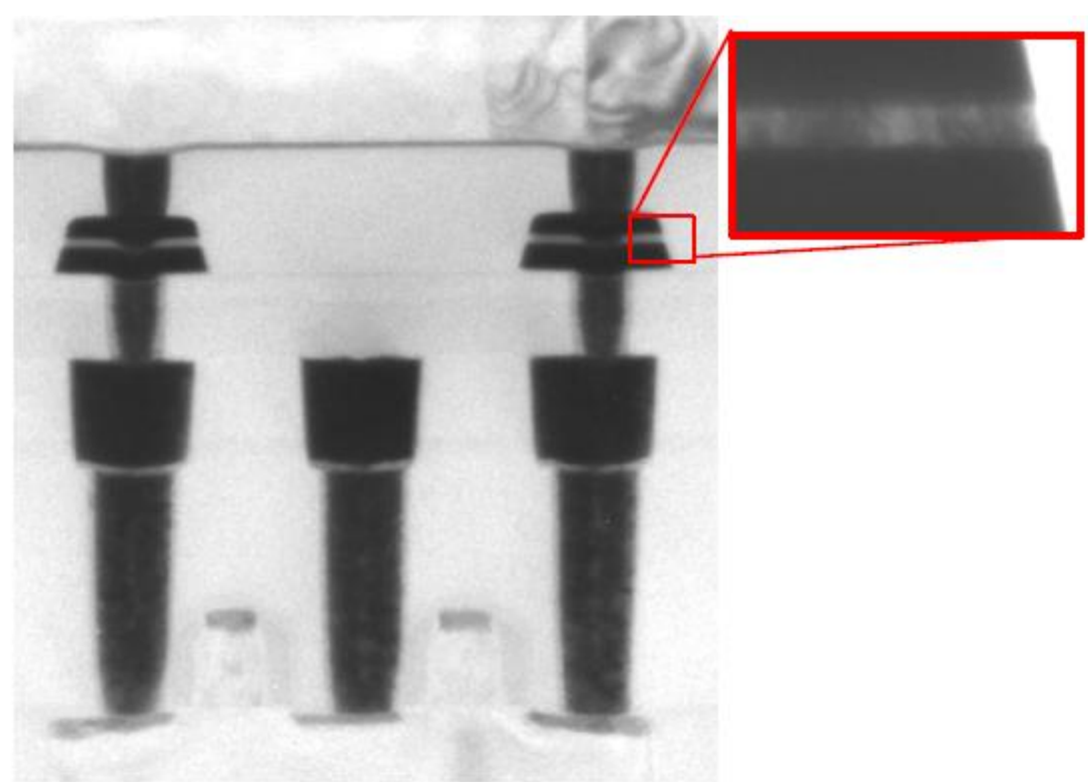
IV Polarity and Symmetry

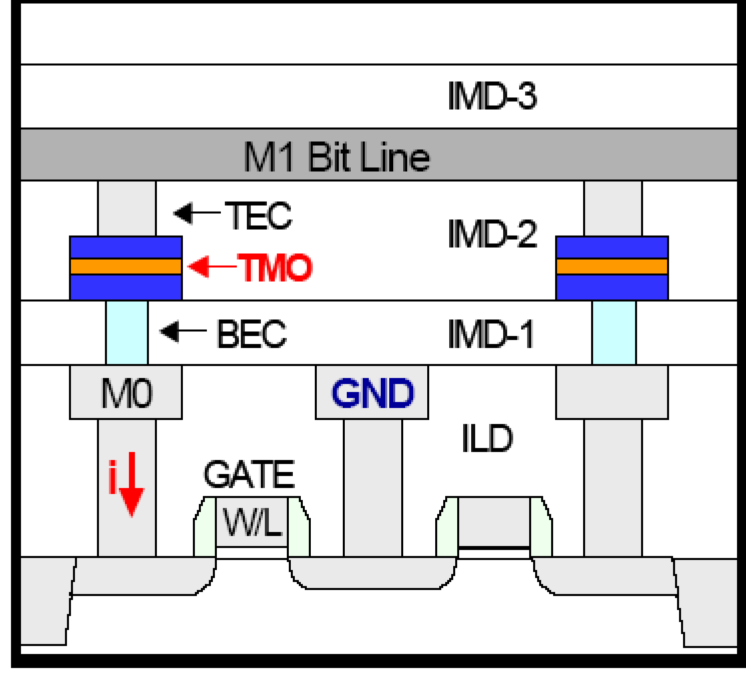


NiO Memory

- 3V set / 1.4V reset
- Direct overwrite during reset
- 10^6 cycles
- 2mA programming current







I. G. Baek, et al, "Highly scalable non-volatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses," *IEDM Tech. Dig.*, pp. 587 - 590, December 2004.

Commercial ReRAM

Maker	Samsung	Sharp	IBM (Zürich)
Material	**Binary Oxide (NiO, TiO_2, HfO_2, ZrO_2...)**	**CMR Perovskite ($Pr_{0.7}Ca_{0.3}MnO_3$)**	**TM-doped Perovskite ($SrZrO_3$, $SrTiO_3$)**
Proposed Mechanism	**trap charging /discharging (SCLC) + filamentary current path**	**trap charging /discharging (SCLC) + Schottky barrier**	**trap charging /discharging (SCLC)**
Electrode	**Noble metal, W, poly Si**	**Noble metal, YBCO, $LaAlO_3$**	**Noble metal, $SrRuO_3$**
Operation Voltage	**0 ~ 3V**	**- 5 ~ 5V**	**- 1.1 ~ 1.1V (SZO)**
Operation Current	**< 2mA**	**~ 200μA**	**~ 50μA (SZO)**
Pros.	**CMOS compatible, Uni-polar program**	**Multibit possible, Fast program(~20ns)**	**Multibit possible, Low current**
Cons.	**High operation current**	**CMOS compatibility, Seed layer, High operation voltage**	**CMOS compatibility, Seed layer, retention time**

Source: S. Seo (2006)

SCLC = space charge limited current

What Do You Learn

- Mechanisms and structures for embedded EEPROM
- Tradeoffs in operation (high voltage, unipolar, etc.) and manufacturing constraints (single poly, isolation, etc.)
- Other possible mechanisms for building eNVM