Lecture #29

ANNOUNCEMENTS

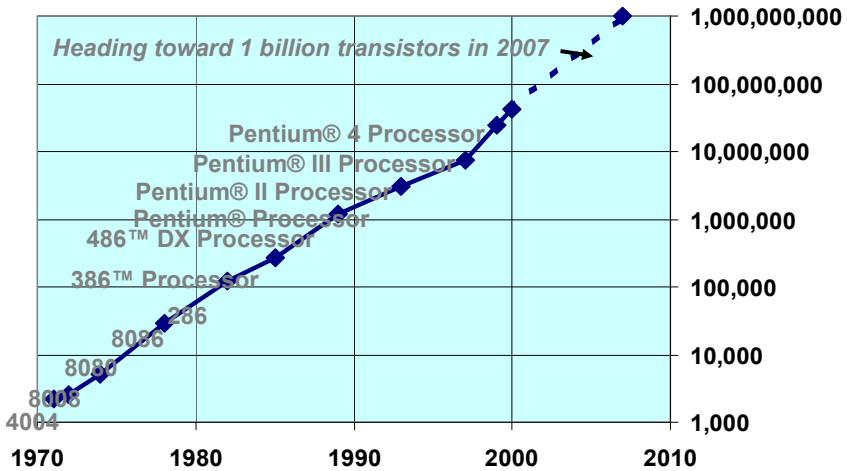
- HW#15 will be for extra credit
- Quiz #6 (Thursday 5/8) will include MOSFET *C-V*
- No late Projects will be accepted after Thursday 5/8
- The last Coffee Hour will be held this Thursday at 5/8
- Prof. King & TAs will hold office hours through 5/22

OUTLINE

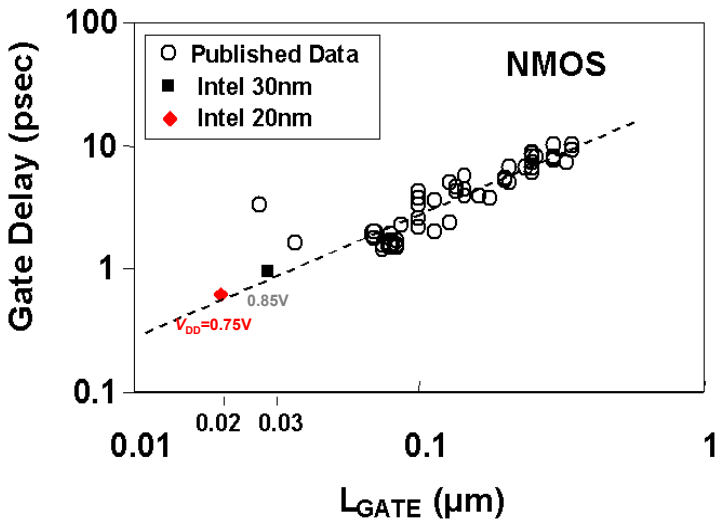
- MOSFET scaling (reprise)
- SOI technology
- MOS memory devices

Moore's Law

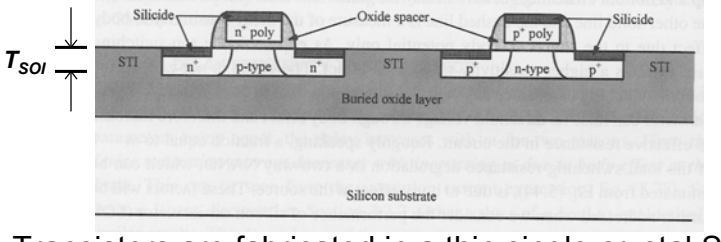
 # transistors/chip doubles every 1.5 to 2 years



Intrinsic Gate Delay ($C_{gate}V_{DD}$ / I_{Dsat})



Silicon on Insulator (SOI) Technology



- Transistors are fabricated in a thin single-crystal Si layer on top of an electrically insulating layer of SiO_2
  - ✓ Simpler device isolation → savings in circuit layout area
  - ✓ Low junction capacitances → faster circuit operation
  - ✓ Better soft-error immunity
  - ✓ No body effect
  - × Higher cost

Partially Depleted SOI (PD-SOI)

$$T_{SOI} > W_{dm}, \text{ where } W_{dm} = \sqrt{\frac{2\varepsilon_s(2\psi_B)}{qN_{body}}}$$

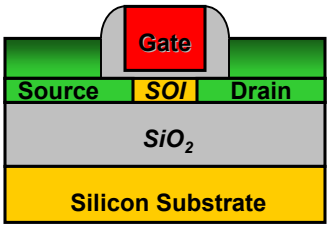
Floating body effect (history dependent):

1. When a PD-SOI NMOSFET is in the ON state, at moderate-to-high V_{DS}, holes are generated via impact ionization near the drain
2. Holes are swept into the neutral body, collecting at the source junction
3. The body-source pn junction is forward biased
4. → V_T is lowered → I_{Dsat} increases
   → "kink" in output I_D *vs.* V_{DS} curve

Fully Depleted SOI (FD-SOI)

$$T_{SOI} < W_{dm}, \text{ where } W_{dm} = \sqrt{\frac{2\varepsilon_s(2\psi_B)}{qN_{body}}}$$

- No floating body effect!
- V_T is sensitive to SOI film thickness
- Poorer control of short-channel effects due to fringing electric field from drain
- Elevated S/D contact structure needed to reduce R_S, R_D

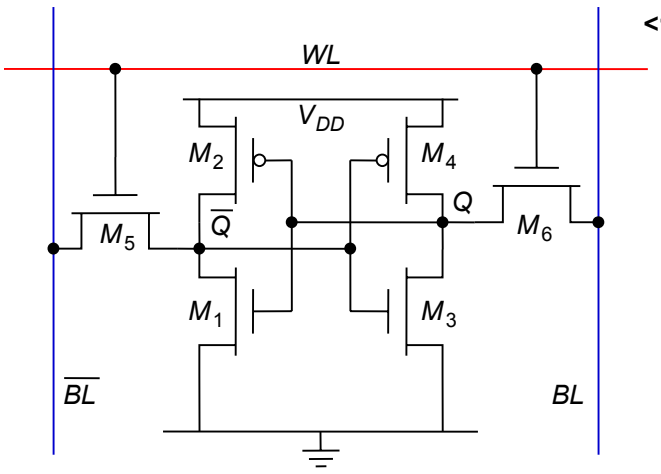


Semiconductor Memory

- **Volatile**
  - Static random access memory (SRAM)
  - Dynamic random access memory (DRAM)
- **Non-Volatile**
  - Mask programmed ROM
  - Programmable Read-Only Memory (PROM)
  - Electrically programmable ROM (EPROM)
  - Electrically erasable PROM (E^2PROM)
  - Flash EPROM

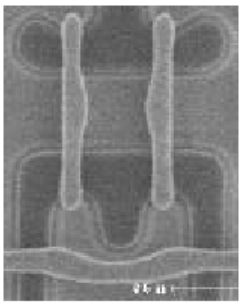
6-Transistor CMOS SRAM Cell

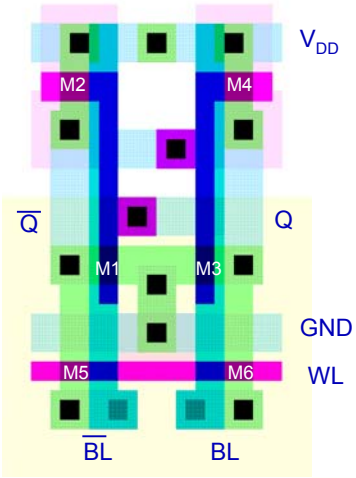
~1 ns read time
<10 ns write time



6T-SRAM: Layout

- Modern processes can fit a 6T SRAM cell in ~1.0μm^2



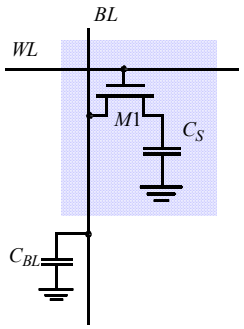


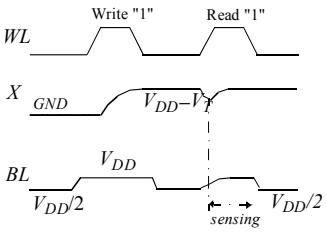
SRAM Scaling Challenges

- **Low standby power**
  - → low OFF current (*e.g.* 1 pA/cell)
    - → large V_T is required
- **Soft error immunity**

1-Transistor DRAM Cell

~10 ns read time
~100 ns write time





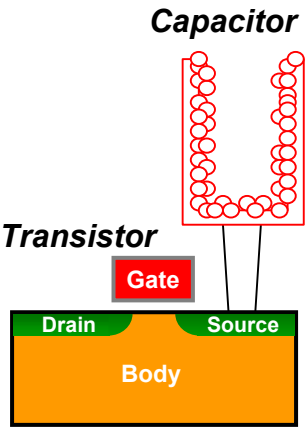
Write: C_S is charged or discharged by asserting WL and BL.
Read: Charge redistribution takes places between bit line and storage capacitance

$$\Delta V = V_{BL} - V_{PRE} = (V_{BIT} - V_{PRE})\frac{C_S}{C_S + C_{BL}}$$

Voltage swing is small; typically around 250 mV.

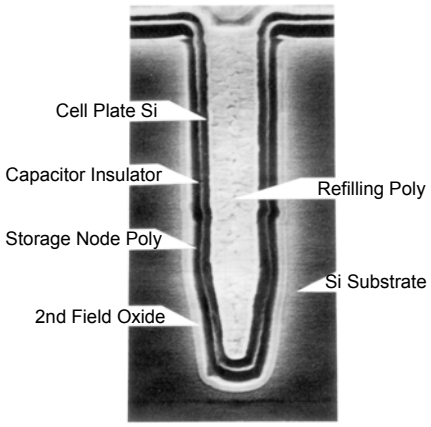
DRAM Cell Structure

- Desired characteristics:
  - ✓ low power consumption
  - ✓ long retention time
  - ✓ "fast" access time
  - ✓ soft error immunity
- ≥25fF/cell is required for sensing signal margin and retention time

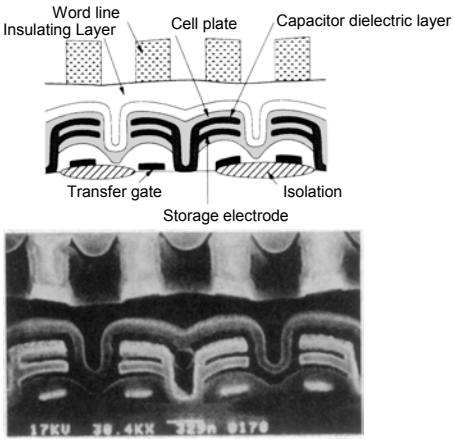


Advanced DRAM Capacitor Structures

Trench Capacitor



Stacked Capacitor

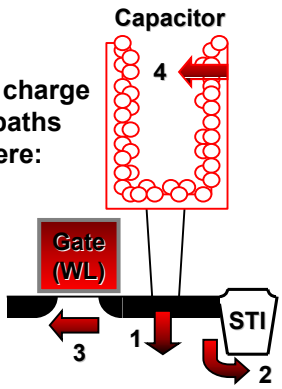


DRAM Scaling Challenge

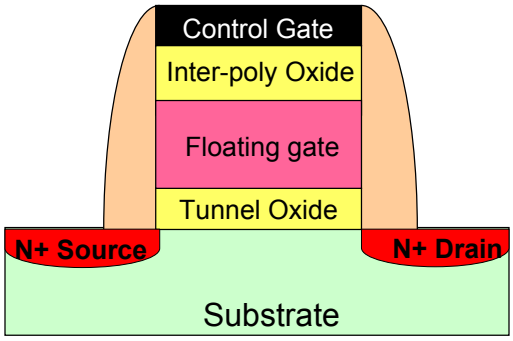
- Long retention time → low OFF current (~1 fA)
  - **large V_T is required**
- Fast access time → high ON current (~100 μA)
  - **large (V_{GS}-V_T) is required**

=> V_{DD} cannot be scaled down aggressively, for low power consumption

Possible charge leakage paths shown here:



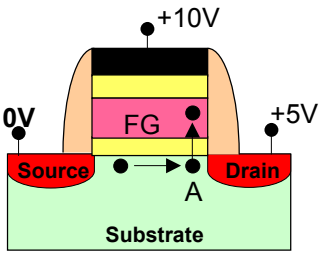
Flash EPROM Cell Structure

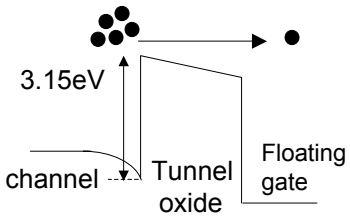


- To program this device, electrons are injected from the channel inversion layer into the floating gate through the tunnel oxide.
- The inter-poly oxide is thick, to prevent electrons from tunneling through it.

- Tunnel oxide: 8 nm thermal oxide
- Floating gate: 100 nm N+ poly-Si
- Inter-poly oxide: 16 nm CVD oxide or Oxide/Nitride/Oxide stack

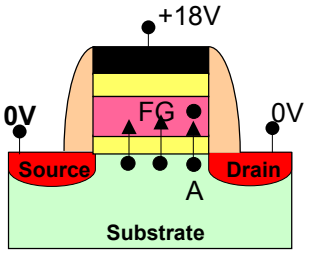
Program by Hot Electron Injection

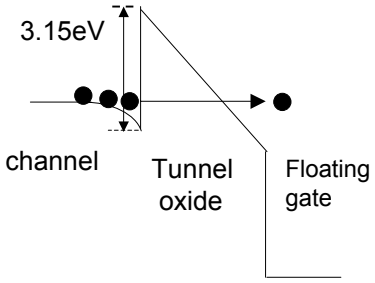




- Electrons are accelerated by the lateral E-field and gain enough kinetic energy at point A (near the drain) to surmount the potential barrier.
- Because of the control-gate bias, electrons are injected into the floating gate.

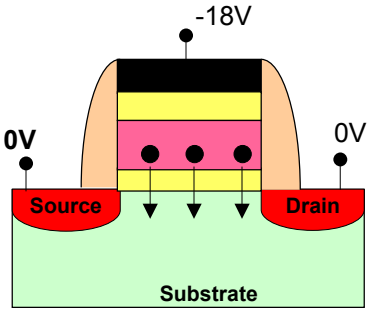
Program by Fowler-Nordheim Tunneling

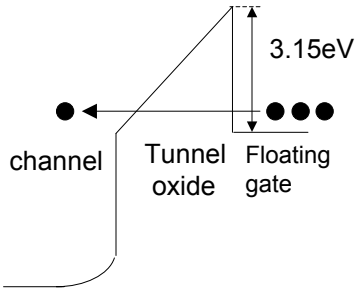




• For a sufficiently high control-gate bias, electrons can tunnel from the channel inversion layer into the floating gate.

Erase Operation

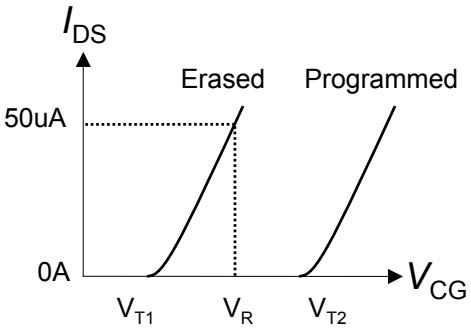




• Under a large negative control-gate bias, electrons tunnel out of the floating gate into the substrate.

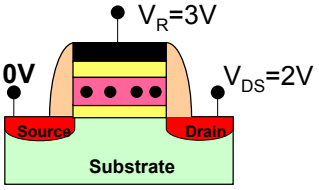
Sensing the Stored Data

Two V_T states:



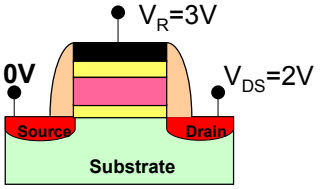
(1) Programmed state

$V_T = V_{T2} = 5V$, $I_{DS} = 0$

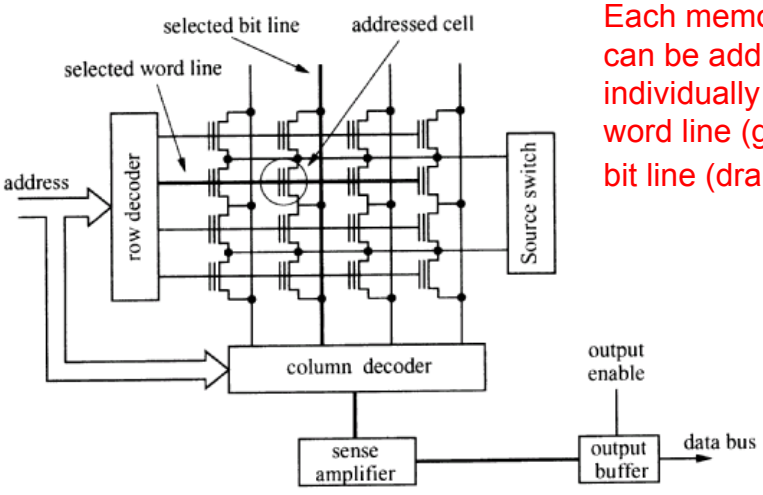


(2) Erased state

$V_T = V_{T1} = 1V$, $I_{DS} = 50$ uA

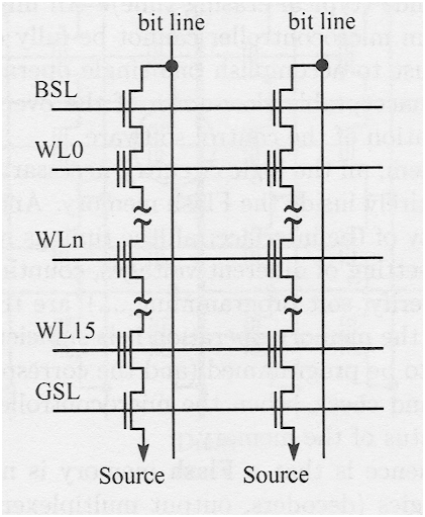


NOR Flash Memory Architecture



Each memory cell can be addressed individually by its word line (gate) and bit line (drain)

NAND Flash Memory Architecture

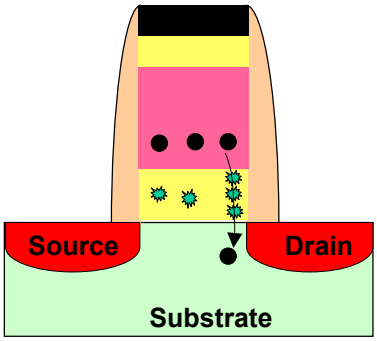


- For each bit line, 16 or 32 cells are connected, with one select transistor at each end of the bit line.
- Programmed $V_T > 0$ V
  Erased $V_T < 0$ V
- The source/drain region between each two adjacent cells are shared → high density

NOR *vs.* NAND Architecture

	NOR	NAND
Chip Density	Medium (64MB)	Very high (2GB)
Programming mechanism	Hot electron injection	F-N tunneling
Programming speed	1us ~10us	1ms
Erasing speed	ms byte/block erase	ms block erase
Random access	Yes	No
Application	Code storage	Data storage
Vendor	Intel, AMD	SanDisk, Toshiba, Samsung

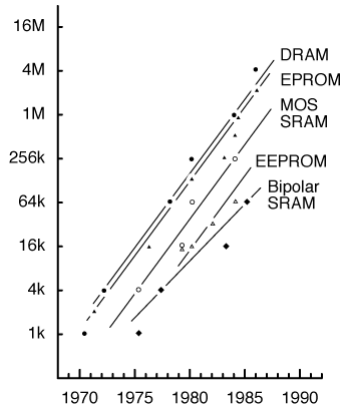
Flash E²PROM Scaling Challenges



- To achieve fast programming speed and low voltage operation, the tunnel oxide thickness must be scaled down.
- Defects in the tunnel oxide reduce the retention time and thereby limit the tunnel oxide scaling, however.
- Today >8nm tunnel oxide is used in commercial flash products.

Semiconductor Memory Trends

Capacity increases 4X every 3-4 years



Today:

- 1 Gb DRAM
- 512 MB SRAM
  (2MB on-chip cache SRAM)
- 1 Gb flash E²PROM