Memory

Can be categorized into:

- **Read Write Memory** (RWM)
  - Random Access Memory (RAM): static SRAM (faster) verses dynamic DRAM (smaller) structures possible. Access time independent of physical location of data.
  - Non-RAM: Serial Access Memory (FIFO, LIFO, Shift register) and Content Access Memory (CAM). Non-uniform access time.
- **Non-volatile Read Write Memory** (NVRWM): write time much larger than read time.
  - EPROM, E^2PROM, FLASH
- **Read Only Memory** (ROM)

A second classification for RAMs and ROMs:

- Static-load: no clock required.
- Synchronous: require a clock edge to enable memory operation.
- Asynchronous: recognize address changes and output new data. More difficult to build.

Memory: Architecture

Typical memory chip architecture:

Only one active line

2^{m+k} bits

2^{n-k} bits

A_k	Row decoder
A_k+1	Row decoder
	Row decoder
A_{n-1}	Row decoder
A_0 … A_{k-1}	Column decoder

column mux, sense amp, write buffers

n-bit address: $[A_{n-1}..A_k][A_{k-1}..A_0]$

For example: Let N = 1,048,576 and M = 8 bits for a 1 million byte memory.

$n = \log_2 N = 20$, $k = 8$ and $m = \log_2 M = 3$.

Then there are 2^{n-k} rows $= 2^{12} = 4096$ and

2^{k+m} columns$/2^3$ bits per word $= 2^8 = 256$ words.

Memory: Architecture

Typically, the single module version shown previously is good up to about 256Kbits.

For larger memories, a multiple module (P block) architecture is used.

Block 0   Block i   Block P-1

Row Address

Column Address

Block Address

Global Data bus

Global amplifier/driver

I/O

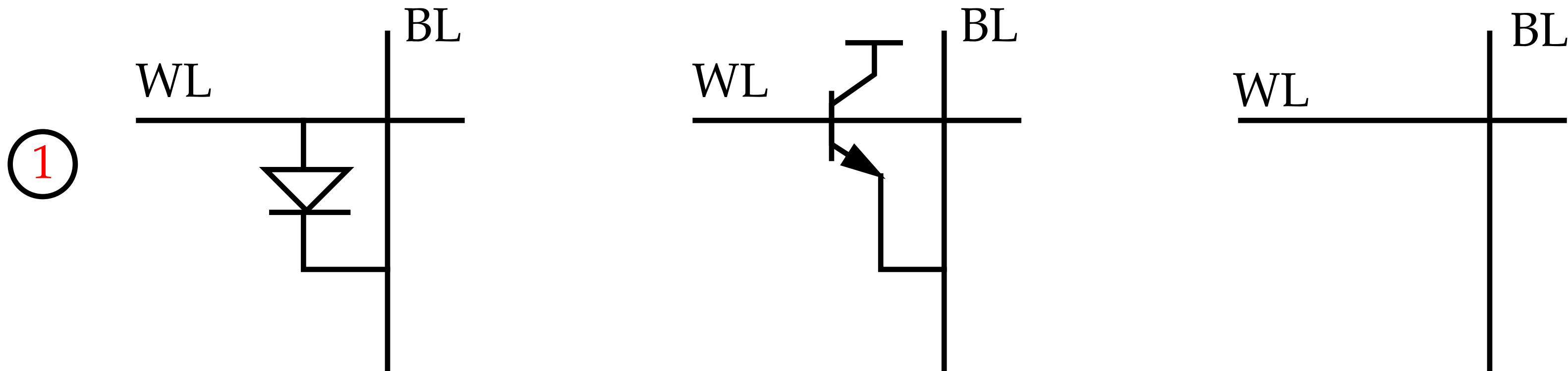
Address: [Row][Block][Col]

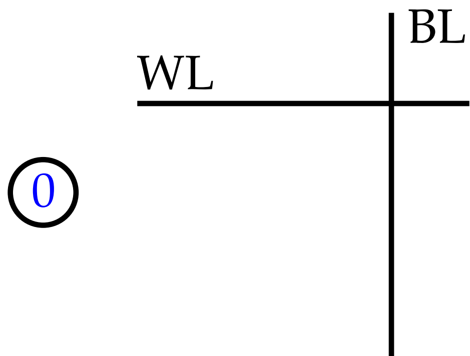
4 Mbit: P = 32 blocks with 128Kbits/block.

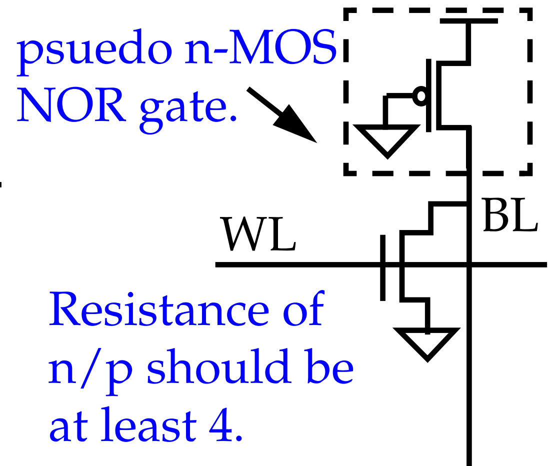
128Kbit block: 1024 rows and 128 columns.

Memory: ROM

ROM cells are permanently fixed: Several possibilities:







Diode supplies current to raise BL (bitline) for all cells on the row.

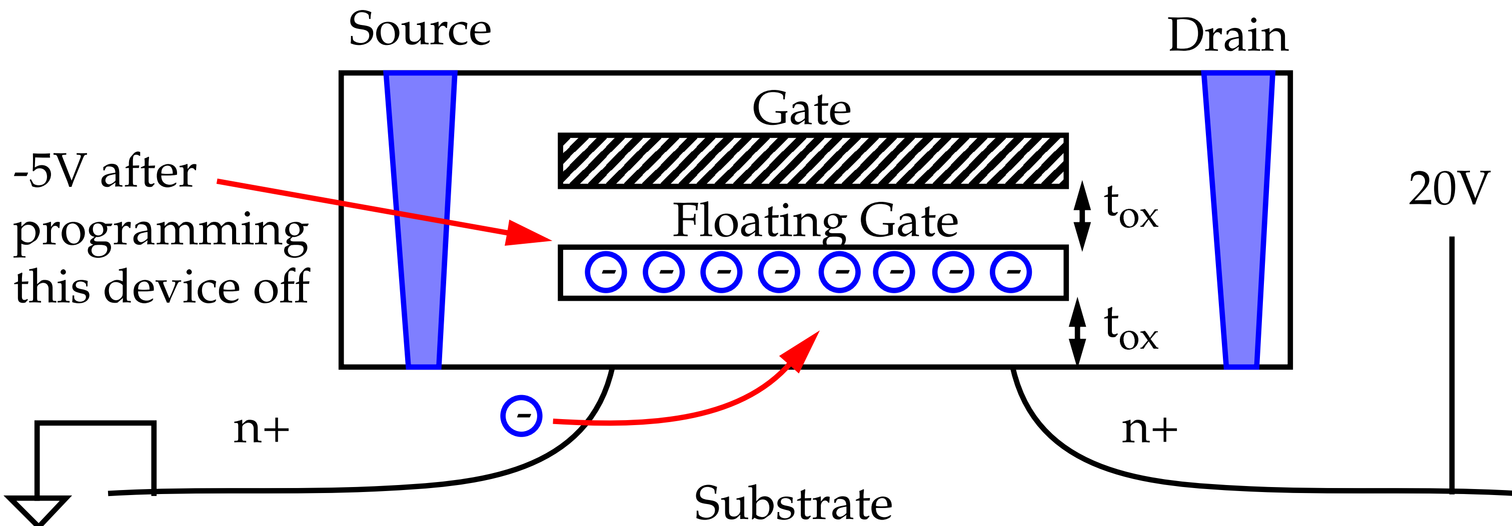
BJT supplies current to raise BL for each cell on the row. Requires V_{DD} to be routed.

p-MOS used to hold BL high. n-MOS provides pull-down path.

Memory: Non-volatile Read-Write Memories

Virtually identical in structure to ROMs.

Selective enabling/disabling of transistors is accomplished through modifications to threshold voltage. This is accomplished through a floating gate.



Applying a high voltage (15 to 20 V) between source and drain create high electric field and causes avalanche injection to occur.

Hot electrons traverse first oxide and get trapped on floating gate, leaving it negatively charged.

This increases the threshold voltage to ~7V. Applying 5V to the gate does not permit the device to turn on.

Memory: Non-volatile Read-Write Memories

The method of erasing is the main differentiating factor between the various classes of reprogrammable nonvolatile memories.

- **EPROM**:

  UV light renders oxide slightly conductive.

  Erase is slow (seconds to several minutes).

  Programming is slow (5-10 microsecs per word).

  Limited number of programming cycles - about 1000.

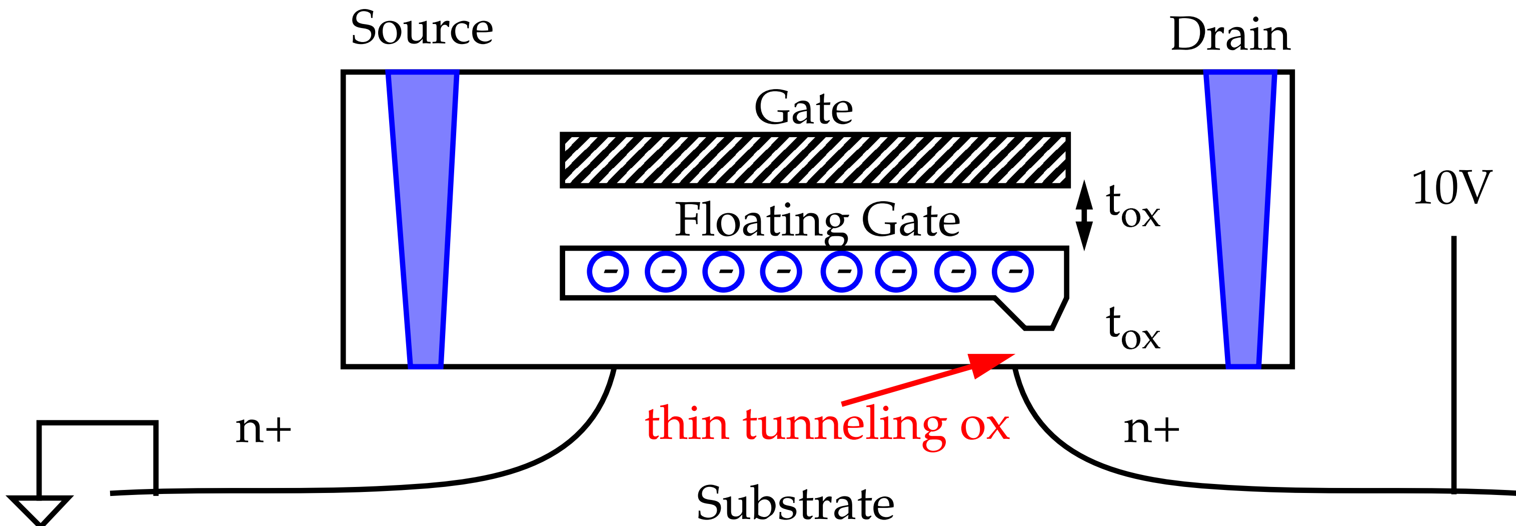
  Very dense - single transistor functions as both the programming and access device.

Memory: Non-volatile Read-Write Memories

- **EEPROM** or **E^{2}PROM**:

  Very thin oxide allows electrons to flow to and from the gate via Fowler-Nordheim tunneling.

  Erasure is achieved by reversing the voltage applied during writing.



Threshold control becomes a problem:

Removing too much charge results in a depletion device.

Remedy: Add an access transistor.

Memory: Non-volatile Read-Write Memories

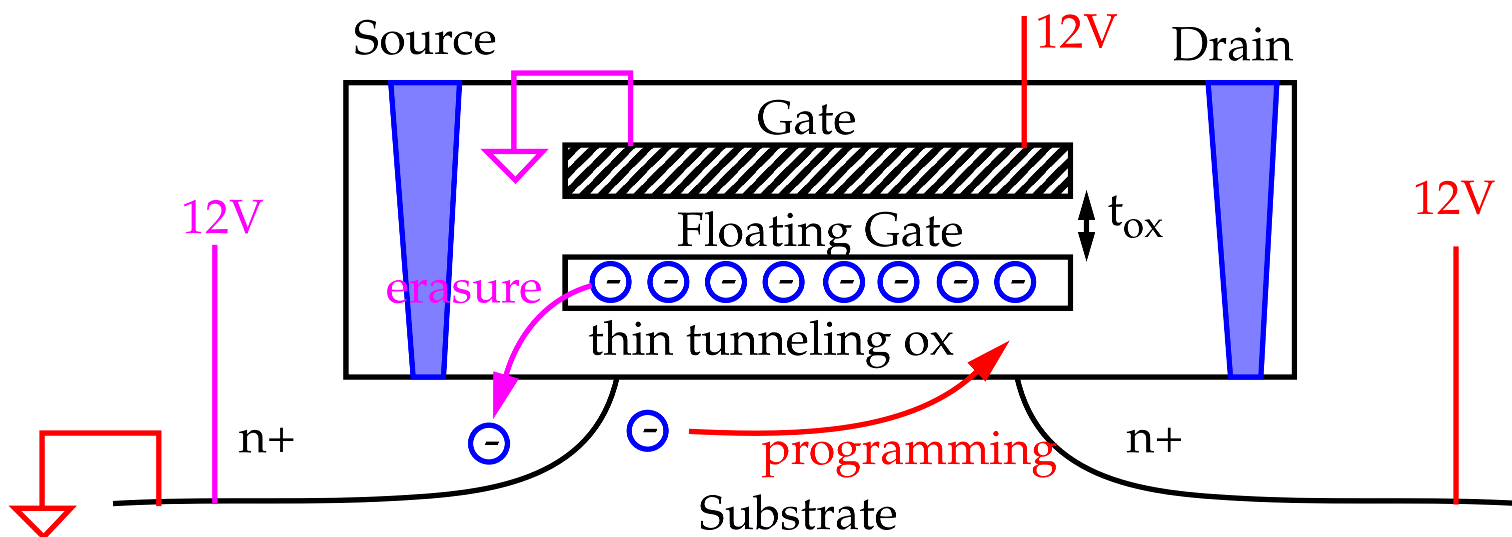
- **Flash EEPROM**:

  Combines density adv. of EPROM with versatility of EEPROM.

  Uses avalanche hot-electron-injection approach to program.

  Erasure performed using Fowler-Nordheim tunneling.

  Monitoring control hardware checks the value of the threshold during erasure - making sure the unprogrammed transistor remains an enhancement device.

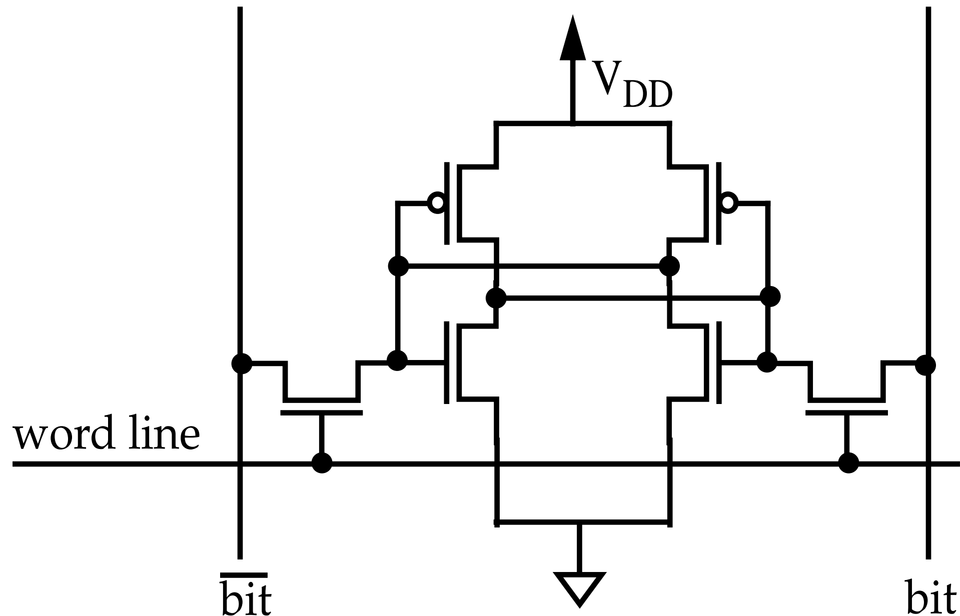


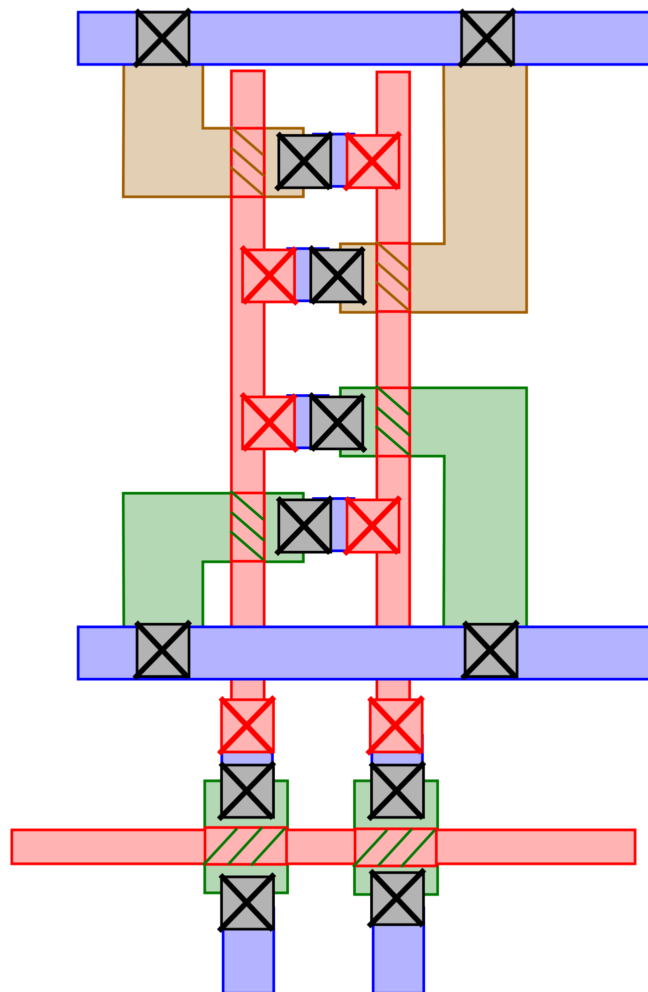
Programming performed by applying 12V to gate and drain.

Erasure performed with gate grounded and source at 12V.

Memory: Read-Write Memories (RAM)

SRAM:





Memory: Read-Write Memories (RAM)

Generic RAM circuit:

SRAM Read Operation:

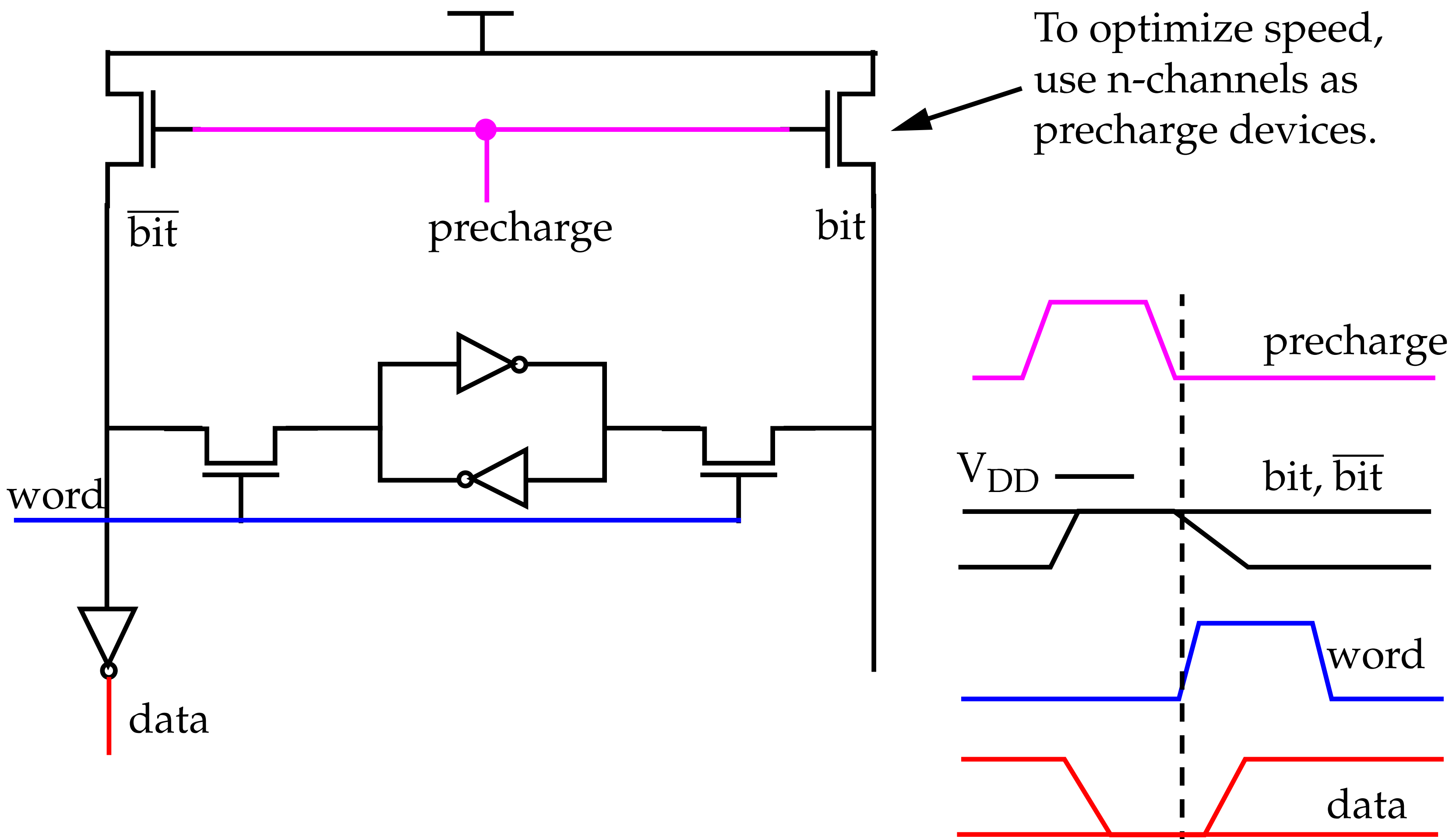
Assume bit and $\overrightarrow{\text{bit}}$ are at some value when the word line is asserted.

Since p channels are small, design focuses on pulling the bit line from high to low through the n-channel access transistor and the n-channel pull-down.

Memory: Read-Write Memories (RAM)

SRAM: Read Operation

Precharging bit and bit_bar to 5V before enabling the word line would help.



Memory: Read-Write Memories (RAM)

SRAM: Read Operation:

Key aspect of precharge RAM read cycle: Timing of RAM addresses, the precharge pulse and the enabling of the row decoder.

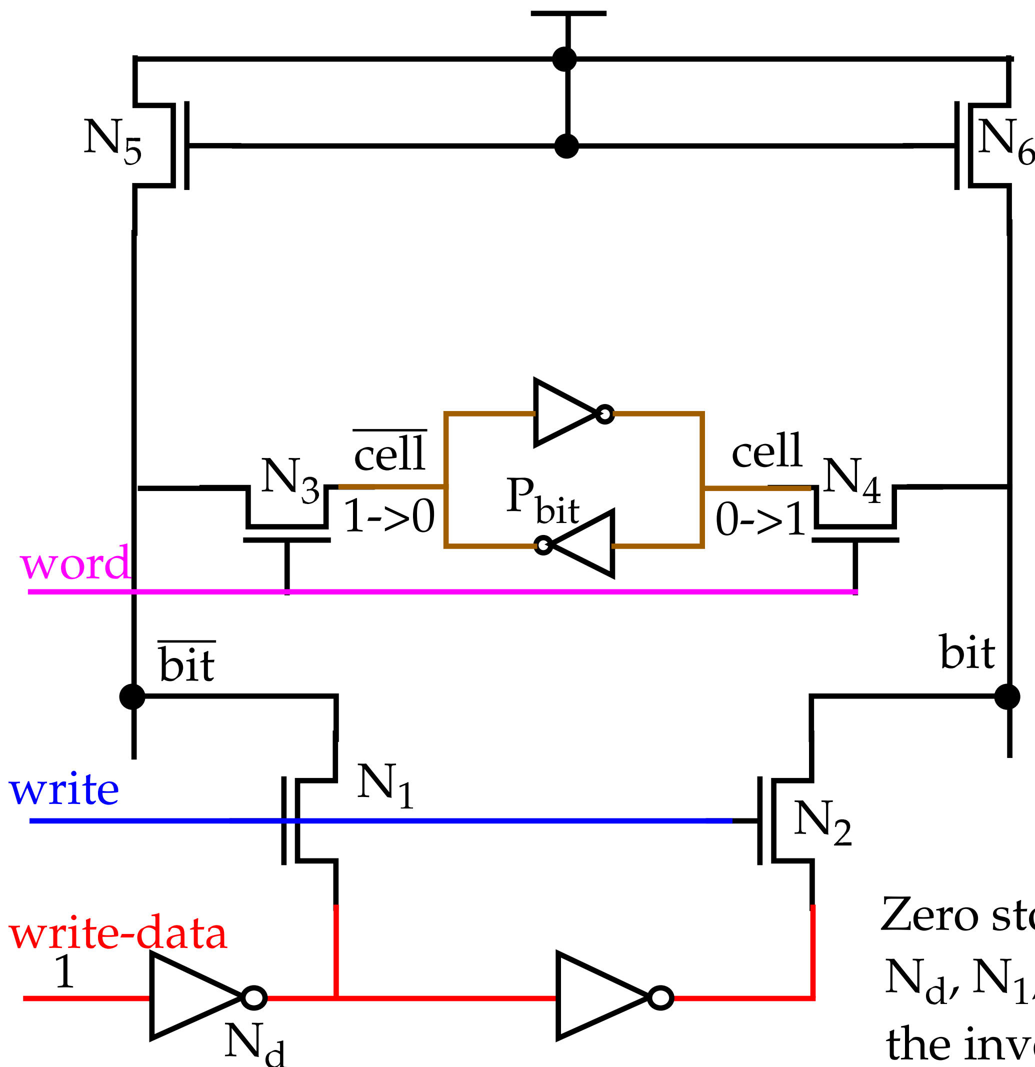
- If word assertion goes high **before** end of precharge, RAM cells on active wordline will see both lines ACTIVELY pulled high - may flip state.
- If address changes **after** the precharge cycle has finished, more than one word line will be accessed and more than one RAM cell will have the chance to pull the bit lines down -> erroneous READ data.

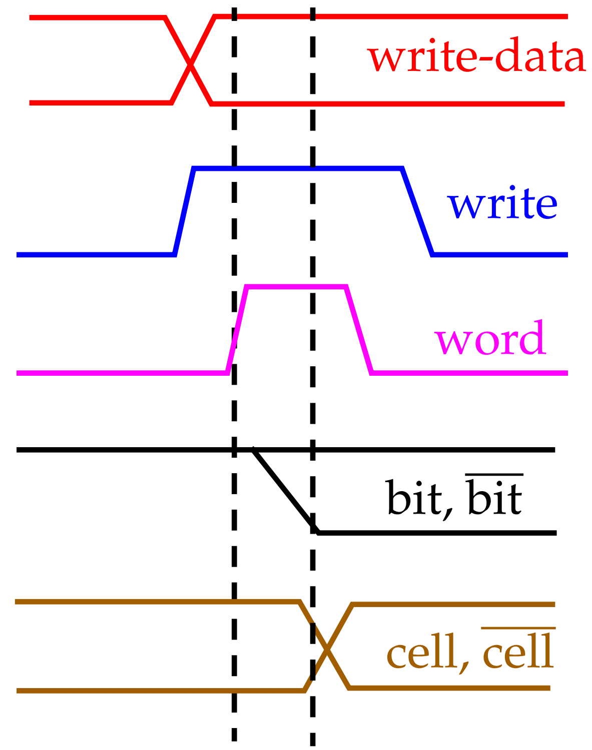
See Weste and Eshraghian for a configuration that eliminates precharging but requires transistor ratioing and a sense amplifier.

Also, row and column decoder layouts and options discussed.

Memory: Read-Write Memories (RAM)

SRAM: Write Operation:





Zero stored in cell originally.

N_d, N_1, and N_3 have to pull P_{bit} below the inverter threshold.

Memory: Read-Write Memories (RAM)

Register files:

Memory: Read-Write Memories (RAM)

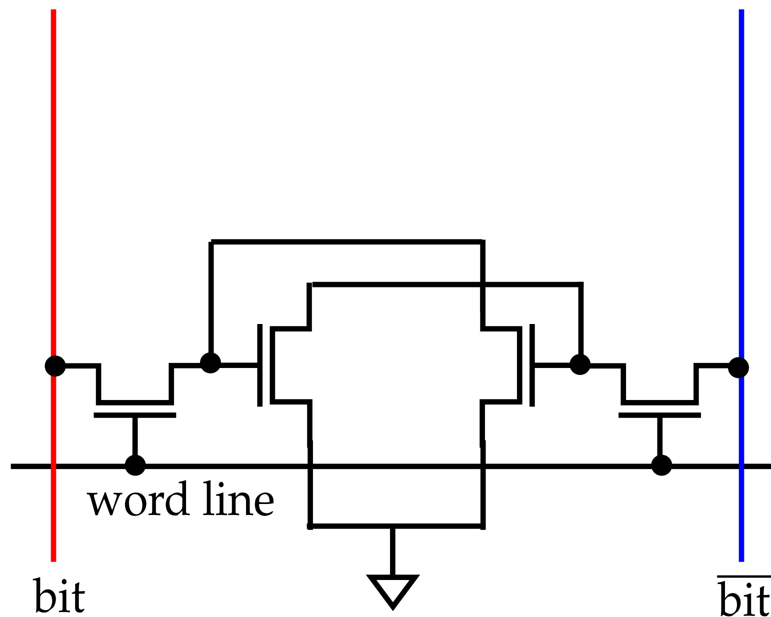
DRAM:

Refresh: Compensate for charge loss by periodically rewriting the cell contents.

Read followed by a write operation.

Typical refresh cycles occur every 1 to 4 milliseconds.

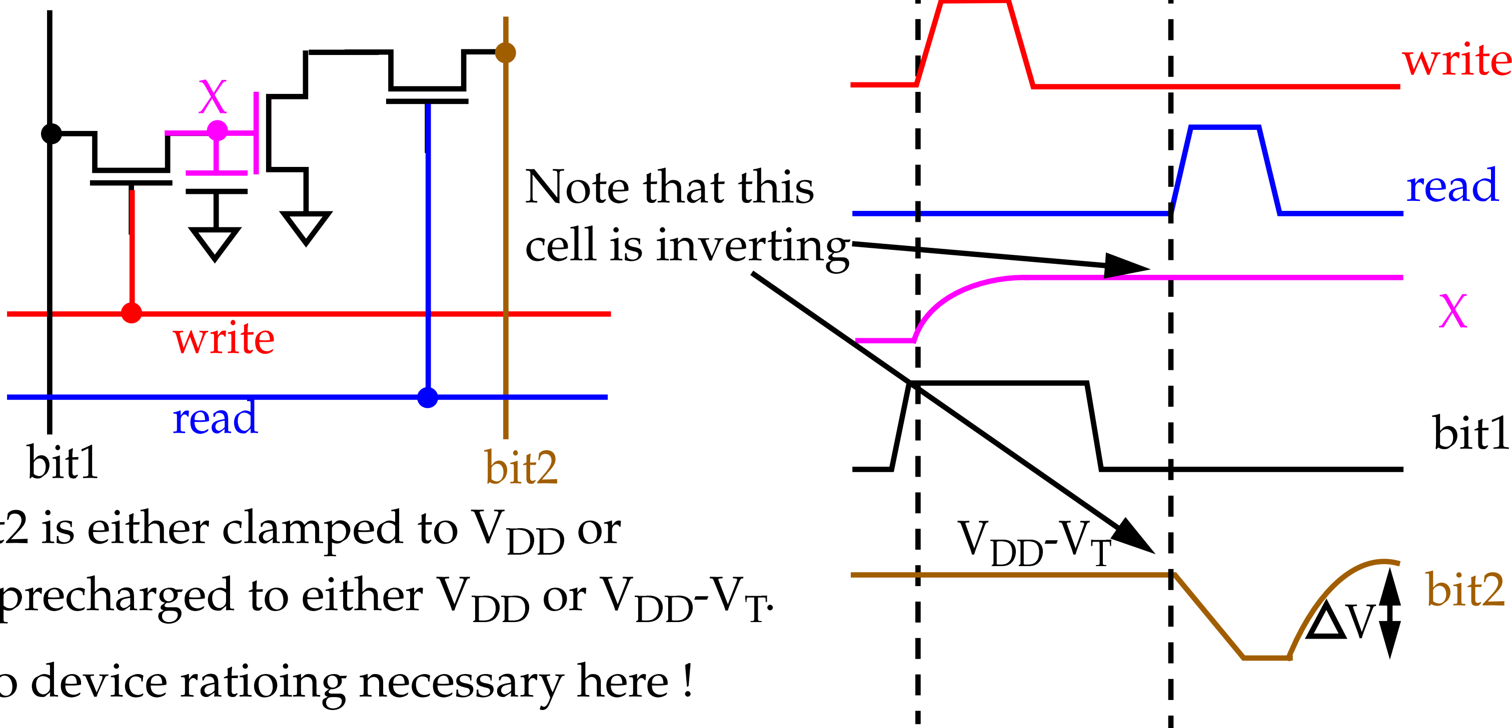
4 transistor DRAM created by simply eliminating the p tree in an SRAM cell.



Logic 1 values are, of course, a threshold below V_{DD}.

Memory: Read-Write Memories (RAM)

3T DRAM:



bit2 is either clamped to V_{DD} or is precharged to either V_{DD} or V_{DD}-V_T.

No device ratioing necessary here !

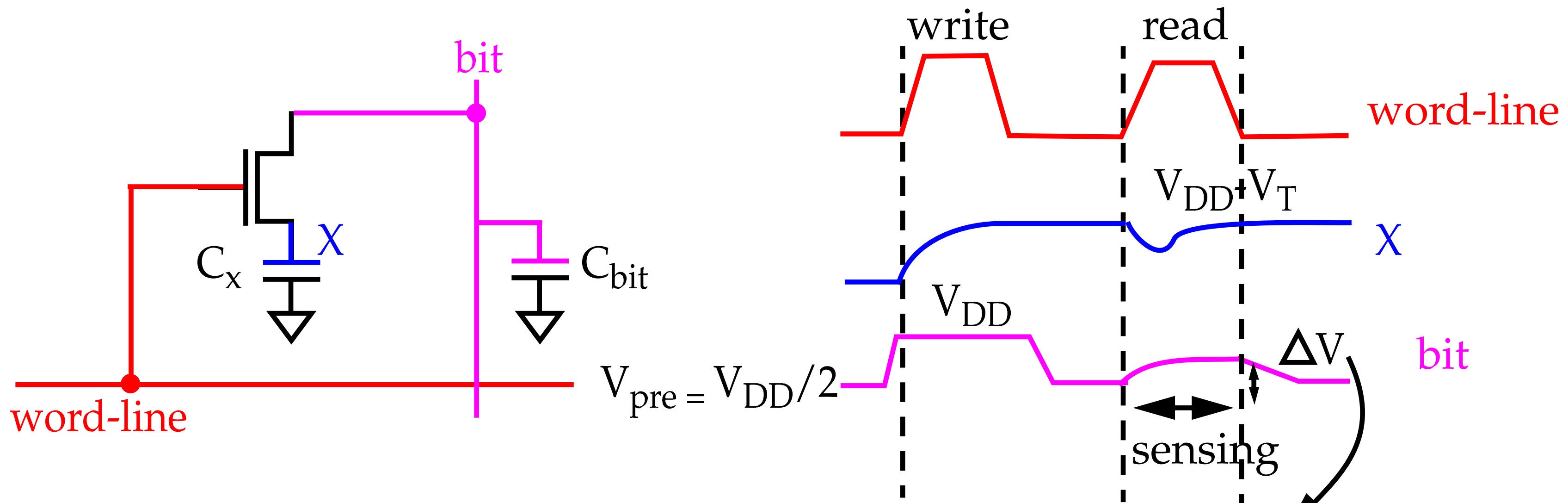
Most common method of refresh is to read *bit2*, place its inverse on *bit1* and assert *write*.

Precharge method of 'setting' *bit2* is preferred (no steady-state current).

Memory structure of choice in ASICs because of its relative simplicity in both design and operation.

Memory: Read-Write Memories (RAM)

1T DRAM



$$\Delta V = V_{bit} - V_{pre} = (V_x - V_{pre})\frac{C_x}{(C_x + C_{bit})}$$

During read operation, charge redistribution occurs between node X and node bit.

C_x is typically 1 or 2 orders of magnitude smaller than C_{bit} so the delta-V value is typically 250 mV.

Most pervasive DRAM cell in commercial memory design.

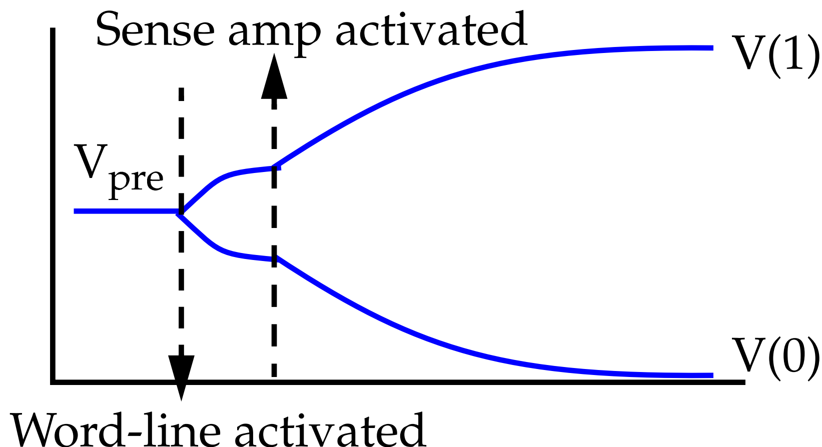
Memory: Read-Write Memories (RAM)

1T DRAM observations:

- Amplification of delta-V (through a sense amplifier) is necessary in order for the cell to be functional.

  Other cell designs used sense amps only to speed up the read operation.

- The read-out operation is destructive ! Output of sense amp is imposed onto the bit line with word-line high during read-out.



- 1T transistor requires an explicit capacitor (3T used gate capacitance).

  Capacitance must be large (~30fF) but area small - key challenge in design.

- Bootstrapping word-line to a value larger than V_{DD} circumvents V_T loss on storage capacitor.

Memory: Read-Write Memories (RAM)

Content Access Memory (CAM):

Determines if a match exists between a data word with a stored word.

Used in Translation-look-aside buffers.

Match is 0 if ANY SRAM cell has $bit/\overline{cell}$ or $\overline{bit}/cell$ equal to 1.

Each bit of the word is tied to the match line.

Dynamic or Pseudo n-MOS implementations possible.