BROWN

TOPICS IN COMPUTING WITH EMERGING TECHNOLOGIES

FALL 2020

PROF. IRIS BAHAR

SEPTEMBER 21, 2020

LECTURE 4: TRANSISTOR DESIGN (CURRENT & POWER)

1

LECTURES AND OFFICE HOURS FOR WEEK #3

- Office hours this week
  - I will be holding office hours outside of MacMillian today from 4:30-5:30pm
  - Tomorrow I will have office hours over zoom from 10-11am (use the lecture zoom link)
  - Email me if you would like to meet outside of these times.
- Monday and Wednesday we will continue with our overview of memory design and conventional transistor design
  - How does a MOSFET transistor work?
  - What are the main sources of power dissipation in a MOSFET transistor?
  - How do emerging technologies operate?

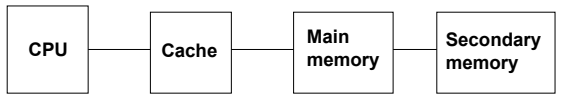
2

COMING UP FOR WEEK #4

- Paper discussions will start week #4
- Review the following papers:
  - *Emerging NVM: A Survey on Architectural Integration and Research Challenges*
  - *Memory that never forgets: emerging nonvolatile memory and the implication for architecture design*
- I will post papers some time next week
  - I will assign teams for reviewing the papers
  - Expect different team assignments weekly
- I will also assign discussion leaders for the week
  - If you want to volunteer, let me know
  - We will rotate discussion leaders throughout the semester. Expect to lead 2-3 times
- Starting the 6th or 7th week of class, students will be able to choose papers to review

3

MEMORY HIERARCHY REVIEW



Registers

Levels L1, L2, … (hardware implementation, SRAMs)

Virtual memory

(software implementation)

Hierarchical organization makes memory look large and fast

Memory access is checked in fast caches first before resorting to slow memory at lower levels.

4

LOCALITY

- Locality is a principle that makes having a memory hierarchy a good idea
- If an item is referenced,
  - temporal locality: it will tend to be referenced again soon
  - spatial locality: nearby items will tend to be referenced soon.
- Why does code have locality?
  - loops
  - instructions accessed sequentially
  - arrays, records

5

DIRECT MAPPED CACHE

- Simple approach: Direct mapped
  - block size is one word
  - every main memory location can be mapped to exactly one cache location
  - lots of words in the main memory share a single location in the cache
- How is the address composed for the cache?
  - cache address is identical with lower bits in the main memory address
  - tag (higher address bits) differentiates between competing main memory words
- We are taking advantage of temporal locality.

6

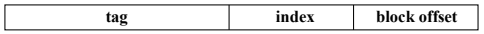
FLEXIBLE PLACEMENT OF BLOCKS

- Direct mapped cache
  - a memory block can go exactly in one place in the cache
  - use the tag to identify the referenced word
  - easy to implement, but rigid placement can cause high miss rate
- Fully associative cache
  - a memory block can be placed in any location in the cache
  - search all entries in the cache in parallel
  - requires a comparator associated with each cache entry
- Set-associative cache
  - a memory block can be placed in a fixed number of locations
  - n locations: n-way set-associative cache
  - a block is mapped to any of n locations in a set
  - Requires searching all locations of the set

7

LOCATING A BLOCK

- Address portions



  - Index selects the set.
  - Tag chooses the block by comparison.
  - Block offset is the address of the data within the block.
- The costs of an associative cache
  - comparators and multiplexers
  - time for comparison and selection
  - More tag bits to store in cache

8

TYPES OF CACHE MISSES

- Compulsory misses: happens the first time a memory word is accessed
  - the misses for an infinite cache
- Conflict misses: happens because two words map to the same location in the cache
  - Most prevalent in direct mapped cache, absent from fully-associative caches
- Capacity misses: happens because the program touched many other words before re-touching the same word
  - the misses for a fully-associative cache

9

ISSUES FOR SET-ASSOCIATIVE CACHES

- Set-associative caches have a significant HW overhead
- Tag lookup is more complicated
- The CPU would like the data as soon as possible
  - For direct mapped caches, there is only one choice of which data to send
  - What about a set-associative cache?
- Can you send the data to the CPU before the tag has been checked?
- ***What about power concerns?***

10

AVERAGE ACCESS TIME

- Hit time is also important for performance
- Average memory access time (AMAT)
  - AMAT = Hit time + Miss rate × Miss penalty
- Example
  - CPU with 1ns clock, hit time = 1 cycle, miss penalty = 20 cycles, I-cache miss rate = 5%
  - AMAT = 1 + 0.05 × 20 = 2ns
    - 2 cycles per instruction

11

INTERACTIONS WITH ADVANCED CPUS

- Out-of-order CPUs can execute instructions during cache miss
  - Pending store stays in load/store unit
  - Dependent instructions wait in reservation stations
    - Independent instructions continue
- Effect of miss depends on program data flow
  - Much harder to analyse
  - Use system simulation

12

"CONVENTIONAL" TRANSISTOR DESIGN

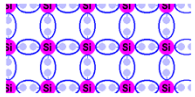
13

IMPACT OF DOPING ON SILICON RESISTIVITY

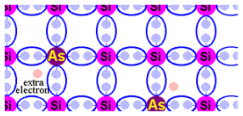
silicon

4.995×10^{22} atoms in cm^3
Resistivity 3.2×10^5 Ωcm

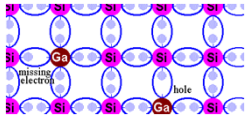
dope with phosphorous or arsenic → n-type



dope with boron (or gallium)→ p-type



1 atom in billion → 88.6 Ωcm
1 atom in million → 0.114 Ωcm
1 atom in thousand → 0.00174 Ωcm

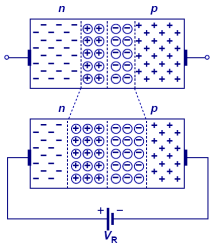


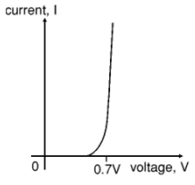
1 atom in billion → 266.14 Ωcm
1 atom in million → 0.344 Ωcm
1 atom in thousand → 0.00233 Ωcm

➔ Electrons are more mobile/faster than holes

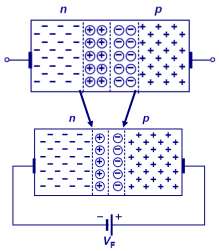
14

P-N JUNCTION REGIONS OF OPERATION





Characteristic of a Silicon Diode



In reverse bias, the width of the depletion region increases. The diode acts as a voltage-controlled capacitor.

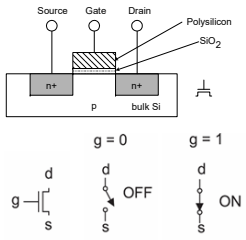
A forward bias decreases the potential drop across the junction. As a result, the magnitude of the electric field decreases, and the width of the depletion region narrows.

15

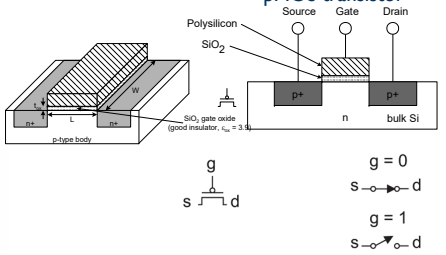
NMOS AND PMOS TRANSISTORS

Each transistor consists of a stack of a conducting gate, an insulating layer of silicon dioxide and a semiconductor substrate (body or bulk)

nMOS transistor



pMOS transistor



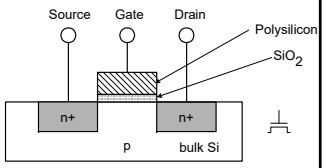
Body is typically grounded

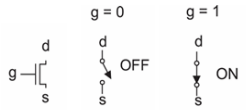
Body is typically at supply voltage

17

NMOS TRANSISTOR

- g=0: gate is at low voltage ($V_{gs}<V_{tn}$)
  - p-type body is at low voltage
  - source and drain junction diodes are OFF
  - transistor is OFF (no current flows)
- g=1: gate is at a high voltage ($V_{gs} \geq V_{tn}$)
  - negative charge attracted to body
  - inverts a channel under gate to n-type
  - transistor is ON (current flows)
  - transistor acts as resistor

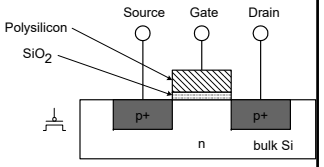


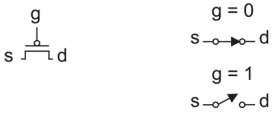


18

PMOS TRANSISTOR

- g=0: gate is at low voltage ($V_{gs}<V_{DD} - |V_{tp}|$)
  - positive charge attracted to body
  - inverts channel under gate to p-type
  - transistor is ON (current flows)
- g=1: gate is at a high voltage ($V_{gs} \geq V_{DD} - |V_{tp}|$)
  - n-type body is at high voltage
  - Source and drain junctions are OFF
  - transistor is OFF (no current flows)

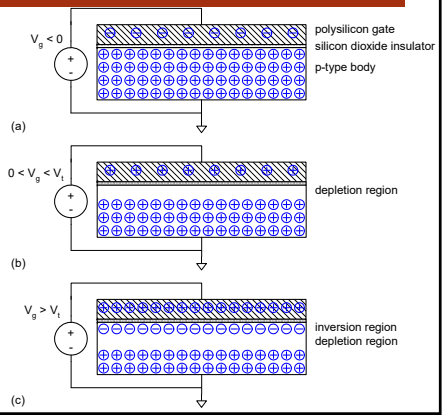




19

MOS CAPACITOR

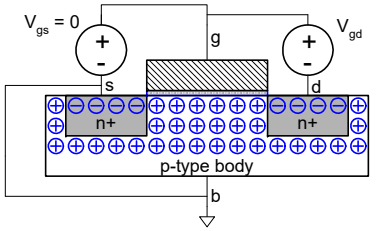
- Gate and body form MOS capacitor
- Operating modes
  - Accumulation
  - Depletion
  - Inversion



21

NMOS CUTOFF

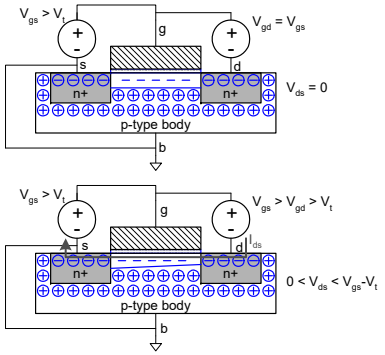
- No channel
- $I_{ds} = 0$



22

NMOS LINEAR

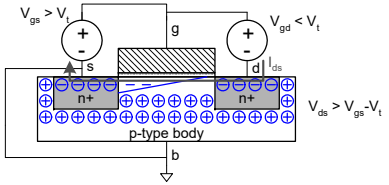
- Channel forms
- Current flows from d to s
  - e^- from s to d
- I_{ds} increases with V_{ds}
- Similar to linear resistor



23

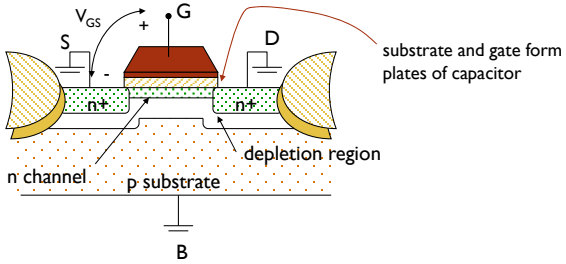
NMOS SATURATION

- Channel pinches off
- I_{ds} independent of V_{ds}
- We say current saturates
- Similar to current source



24

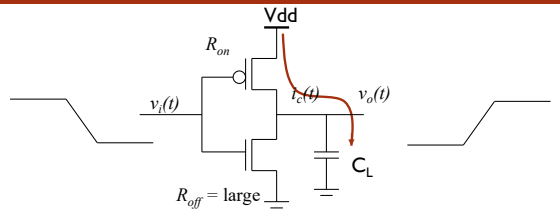
THRESHOLD VOLTAGE CONCEPT



- Depletion region: area devoid of mobile carriers (holes)
- Inversion layer: n-channel region under oxide
- The value of V_{GS} where strong inversion occurs is called the threshold voltage, V_T

25

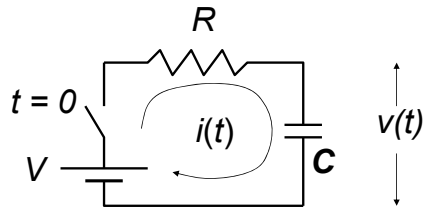
POWER OF A TRANSITION: P_{TRAN}



What goes into charging/discharging the capacitance C_L?

28

CHARGING OF A CAPACITOR



Charge on capacitor, q(t) = C v(t)

Current, i(t) = dq(t)/dt = C dv(t)/dt

29

$$i(t) = C\frac{dv(t)}{dt} = \frac{[V - v(t)]}{R}$$

$$\frac{dv(t)}{dt} = \frac{V - v(t)}{RC}$$

$$\int \frac{dv(t)}{V - v(t)} = \int \frac{dt}{RC}$$

$$\ln[V - v(t)] = \frac{-t}{RC} + A$$

Initial condition, $t = 0$, $v(t) = 0 \rightarrow A = \ln V$

$$v(t) = V[1 - e^{\frac{-t}{RC}}]$$

30

$$v(t) = V[1 - e^{\frac{-t}{RC}}]$$

$$i(t) = C\frac{dv(t)}{dt} = \frac{V}{R}e^{\frac{-t}{RC}}$$

31

TOTAL ENERGY

- Total energy per charging transition from the power supply V_{dd}

$$E_{trans} = \int_0^\infty Vi(t)dt = \int_0^\infty \frac{V^2}{R}e^{\left(\frac{-t}{RC}\right)}dt$$

$$E_{trans} = CV^2$$

32

ENERGY CONSUMED PER TRANSITION IN RESISTANCE

$$E = R\int_0^\infty i^2(t)dt = R\frac{V^2}{R^2}\int_0^\infty e^{\frac{-2t}{RC}}dt$$

$$E = \tfrac{1}{2}CV^2$$

33

ENERGY STORED IN CHARGED CAPACITOR

$$E = \int_0^\infty v(t)i(t)dt = \int_0^\infty V\left[1 - e^{\frac{-t}{RC}}\right]\frac{V}{R}e^{\frac{-t}{RC}}dt$$

$$E = \tfrac{1}{2}CV^2$$

34

TRANSITION POWER

- Gate output rising transition
  - Energy dissipated in pMOS transistor = ½ CV^2
  - Energy stored in capacitor = ½ CV^2
- Gate output falling transition
  - Energy dissipated in nMOS transistor = ½ CV^2
- Energy dissipated per transition = ½ CV^2
- Power dissipation:

$$P_{trans} = E_{trans}\,\alpha\, f_{ck} = \alpha\, f_{ck}\, \tfrac{1}{2}\, CV^2$$

α = activity factor

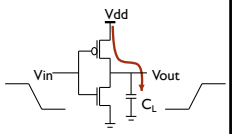
35

LOWERING DYNAMIC POWER

Capacitance:
Function of fan-out, wire length, transistor sizes

Supply Voltage:
Has been dropping with successive generations

$P_{dyn} = C_L V_{DD}^2 \alpha f_{ck}$



Activity factor:
How often, on average, do wires switch?

Clock frequency:
Recent scaling back…

37