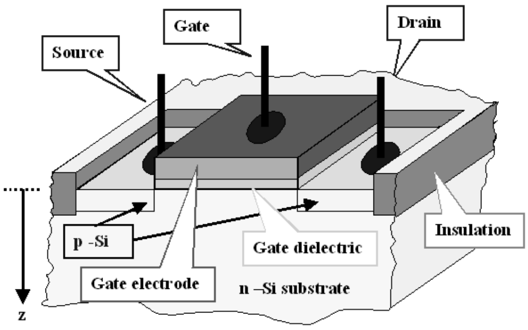
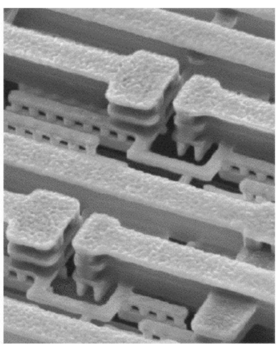
Energy-Aware Computing

Lecture 2: CMOS technology

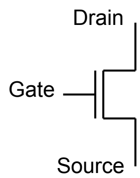
Basic components

- Transistors
  - Two types: NMOS, PMOS
- Wires (interconnect)

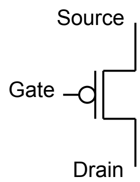




Transistors as switches



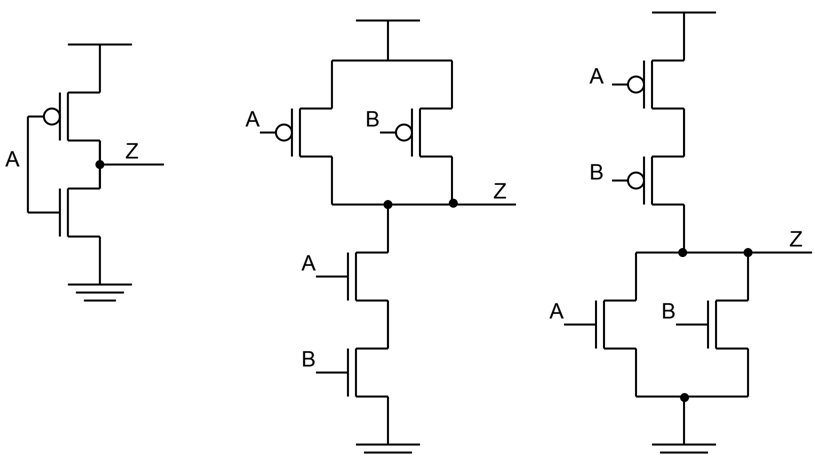
NMOS:
When G is @ logic 1 (actually over a tech dependent threshold voltage), the switch is on
S, D are connected, electric current flows



PMOS:
When G is @ logic 0 (under a threshold), the switch is on
S, D are connected, electric current flows

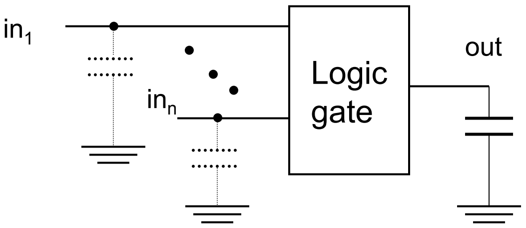
UoE/Informatics Energy-aware computing

Logic gates (static)



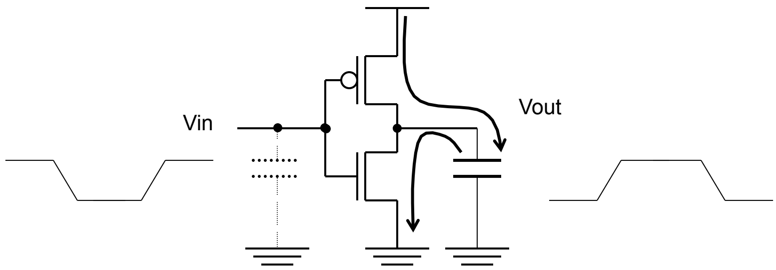
UoE/Informatics Energy-aware computing

Capacitances



- Input capacitance is due to:
  - Transistor gate capacitance (NMOS and PMOS)
- Output capacitance is due to:
  - Fanout; number of other gates driven by this one
  - Interconnect (wires)
  - Diffusion capacitance (transistor drain terminals)

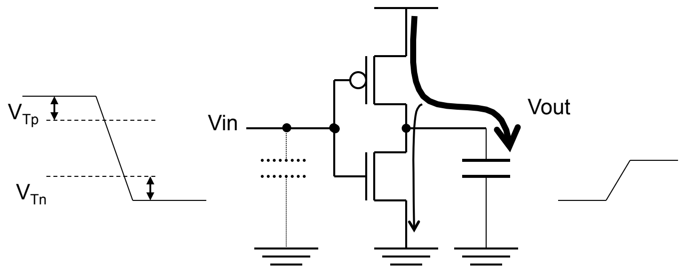
Switching current



Energy in electronic circuits

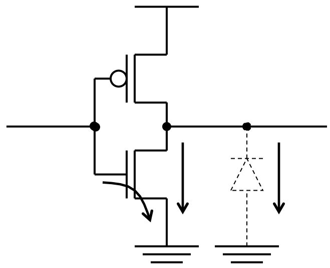
- Power: P = I * V
  - I - current drawn from supply
  - V - the supply voltage
- Energy: E = P * t
- Energy is expended when current flows
- CMOS circuits only draw current when they switch
  - If no input changes, only *leakage current* flows; more later

Short-circuit current



- Signal transition slopes are finite
  - Both NMOS and PMOS conduct for a while
  - Short-circuit!
- About 15% of total *dynamic power*
  - Not much to do except design for steep slopes

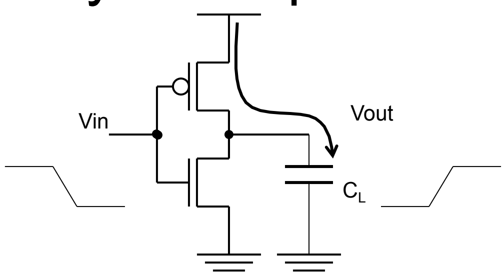
Leakage currents



- The transistor is not a perfect switch
  - Gate leakage
  - Sub-threshold current
  - Drain junction leakage

More on leakage in a future lecture

Dynamic power



Energy/output transition $C_L \cdot V_{DD}^2 \cdot P_{0 \to 1}$

Power (rate of energy consumption)

$$C_L \cdot V_{DD}^2 \cdot P_{0 \to 1} \cdot f$$

Dynamic power

Capacitance
Purely circuit design territory

Clock frequency
Upper limit set by circuit
Affects performance

$$P = C_L \cdot V_{DD}^2 \cdot P_{0 \to 1} \cdot f$$

Supply voltage
Upper/lower limits set by tech
Affects circuit switching time

Activity factor (<=1)
Circuit style dependent
Data dependent
“Algorithm” dependent

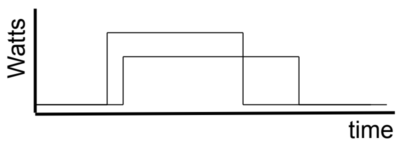
Energy

- Unit: Joules
- Measured over time
  - E.g. run-time of a benchmark
- Time is important!
- Often referred to as the power-delay product
- Energy is a good metric for
  - Battery life
  - Energy bills
- Proportional to total CV^2

Power

- Power: work done per time unit
- Units: Watts (= Joules / sec)
- Average and peak are of interest
- Good for
  - Predicting heat dissipation (avg power)
  - Setting the specs for the power delivery system (peak power)

Power vs Energy



- Which method (blue/red) is better?
- Compare both energy (area under the curve) and run-time
- Need metrics that combine time with energy and/or power

Energy - delay product

- $EDP = E * t = P_{avg} * t^2$
  - t is run-time
- Lower is better
- For systems expending equal energy, the fastest one has better EDP
  - Similar for systems with equal run-times
- There is a catch though!

Et^2

- EDP is misleading when circuits allow voltage scaling
- Assume systems A, B with
  - $E_A = 2E_B$, $t_A = t_B/2$
- If supply voltage of A can drop by half:
  - $E_A' = E_A / 4$, $t_A' = 2\ t_A$
- Therefore, A is better:
  - $E_A' = E_B / 2$, $t_A' = t_B$
  - Assuming B cannot have voltage scaling

More metrics

- What does MIPS/W represent?
  - The **reciprocal** of joules/instruction
  - Larger is better!
  - Essentially an energy metric
- What would be the EDP equivalent?
  - $MIPS^2/W$
  - Reciprocal! Larger is better

Technology progress

- Currently chips using 45nm, 32nm are being produced
- For comparison (src Intel fact sheet):
  - Rhinovirus = 20nm
  - Silicon atom = 0.24nm

Future trends/problems

- Variability
  - Very hard to control certain key parameters such threshold voltage
  - In the same chip, neighbouring transistors of the same size and orientation will operate differently
- Reliability
  - Some transistors will fail or will be so slow that appear faulty

Future trends/problems

- Leakage power is increasing
- Long interconnect is slower
  - Delay due to interconnect used to be negligible
  - Now most of the delay comes from interconnect and this is set to continue

Summary

- Transistors, wires
- Capacitances and what they depend on
- Transistor threshold voltage
- Dynamic, static power/energy
- Dynamic energy expended when switching occurs
- $P = C_L \cdot V_{DD}^2 \cdot P_{0\to 1} \cdot f$
- Metrics
- Future trends:
  - Leakage, variability, interconnect, (un)reliability