15 January 2020

وَمَا أُوتِيتُمْ مِنَ الْعِلْمِ إِلَّا قَلِيلًا

20 جمادى الأولى 1441

Digital IC Design

Lecture 12
Power

Dr. Hesham A. Omran

Integrated Circuits Laboratory (ICL)
Electronics and Communications Eng. Dept.
Faculty of Engineering
Ain Shams University

This lecture is mainly based on “CMOS VLSI Design”, 4th edition, by N. Weste and D. Harris and its accompanying lecture notes

Why Low-Power?

- Battery powered devices
  - We want to add more features
  - But we want to reduce the size (size is limited by the battery)
  - And we want to extend the lifetime (also limited by the battery)
- Mains powered devices (with power cord)
  - We want to save cost
    - Cost of electricity
      - US data centers consuming 140 billion kW-hr annually by 2020
      - costing $13 billion annually in electricity bills only
    - Cost of cooling
      - Beyond 150W/chip liquid cooling or special heat sinks are required
  - We want to save the environment
    - Greenhouse emissions, climate change, etc.

Power and Energy

- ❑ Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.

- ❑ Instantaneous Power:

$$P(t) = I(t)V(t)$$

- ❑ Energy:

$$E = \int_0^T P(t)dt$$

- ❑ Average Power:

$$P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T P(t)dt$$

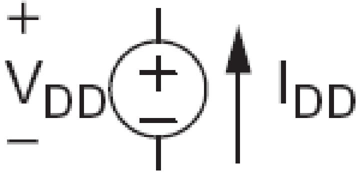
Energy

$$E = \int_0^T P(t)dt$$

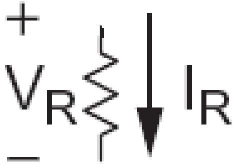
- ❑ Energy in circuits is usually expressed in Joules (J)
  - ▪ 1 W = 1 J/s
- ❑ Energy in batteries is often given in W-hr
  - ▪ 1 W-hr = (1 J/s)(3600 s/hr)(1 hr) = 3600 J
- ❑ At the end of the day, the battery has “Energy” not “Power”
- ❑ You can operate at low power
  - ▪ But if you need multiple cycles to do a given operation
    - You may end up consuming more energy
    - So your battery will die faster
- ❑ The focus should be **“Energy-Efficiency”** rather than “Low-Power”
  - ▪ Always think “Energy-wise”

Power in Circuit Elements

$$P_{VDD}(t) = I_{DD}(t)V_{DD}$$

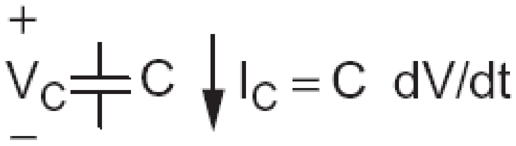


$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t)R$$



$$E_C = \int_0^\infty I(t)V(t)\,dt = \int_0^\infty C\frac{dV}{dt}V(t)\,dt$$

$$= C\int_0^{V_C} V(t)\,dV = \tfrac{1}{2}CV_C^2$$



Charging a Capacitor

- When the gate output rises
  - Energy stored in capacitor is

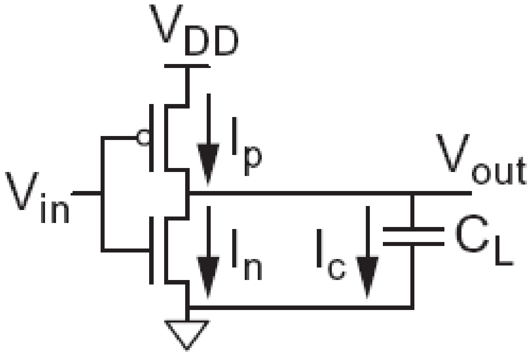
$$E_C = \tfrac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

$$E_{VDD} = \int_0^\infty I(t) V_{DD} dt = \int_0^\infty C_L \frac{dV}{dt} V_{DD} dt$$

$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

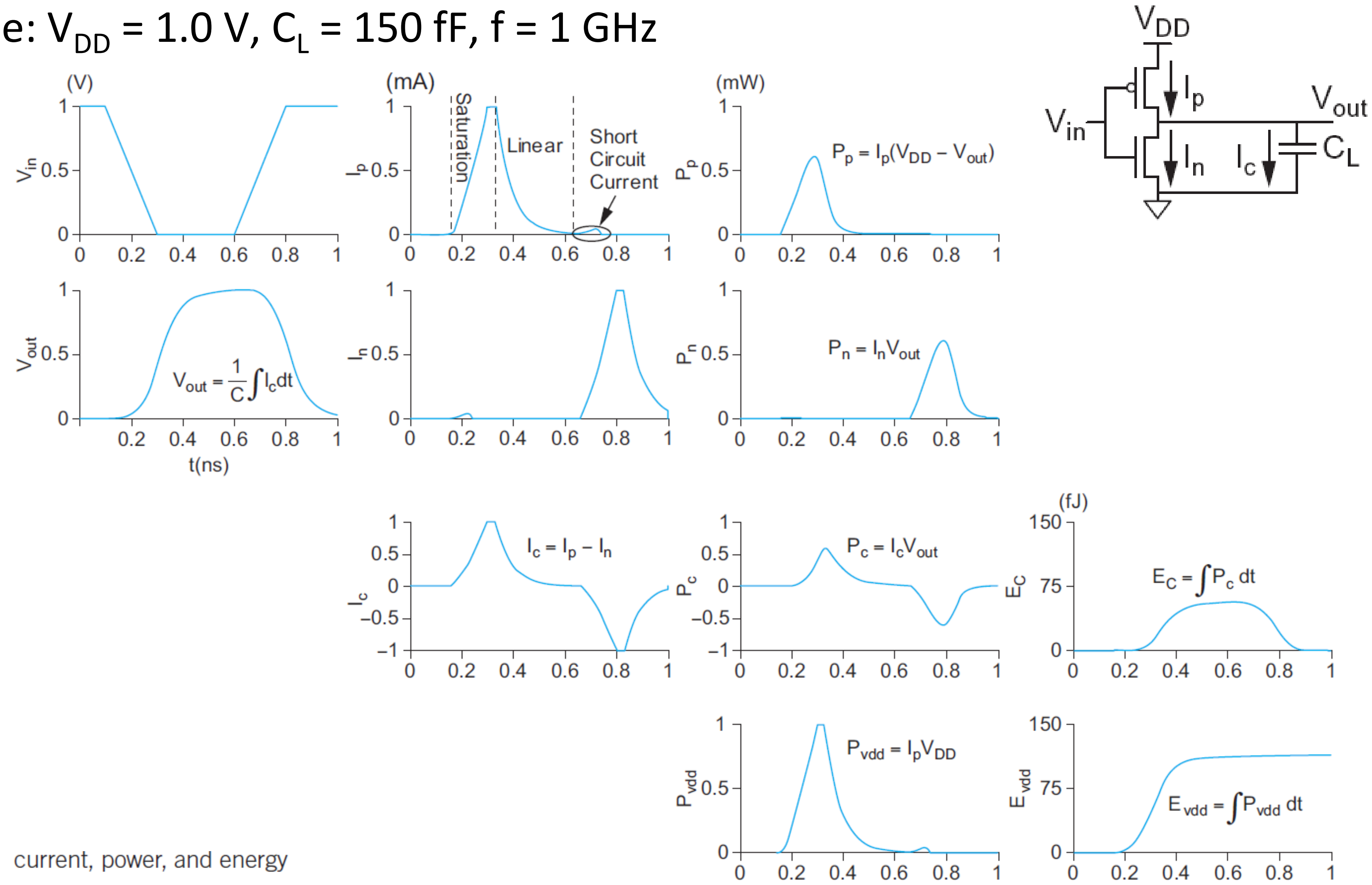
  - Half the energy from V_{DD} is dissipated in the PMOS transistor as heat, other half stored in capacitor



- When the gate output falls
  - Energy in capacitor is dumped to GND
  - Dissipated as heat in the NMOS transistor

Switching Waveforms

- Example: V_{DD} = 1.0 V, C_L = 150 fF, f = 1 GHz



current, power, and energy

Switching Power

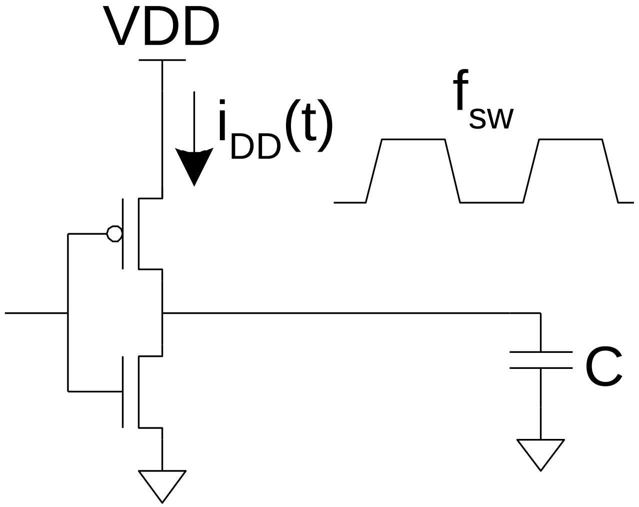
- Energy consumed in one switching cycle ($T_{sw} = 1/f_{sw}$):

$$E_{sw} = C_L V_{DD}^2$$

- Average switching power (a.k.a. dynamic power):

$$P_{sw} = \frac{E_{sw}}{T_{sw}} = C_L V_{DD}^2 f_{sw}$$

  - Quadratic dependence on V_{DD}
  - Linear dependence on f_{sw}



Activity Factor

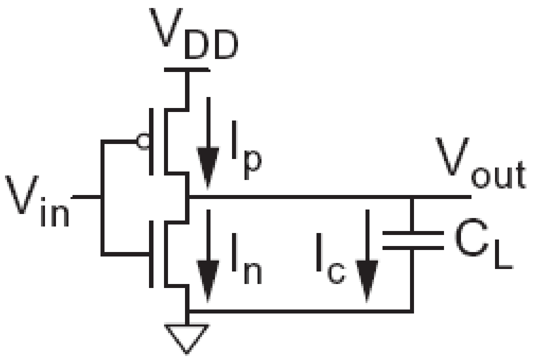
- ❑ Suppose the system clock frequency = f_{clk}
- ❑ Let $f_{sw} = \alpha f_{clk}$, where α = activity factor
  - The activity factor is the probability that the circuit node transitions from 0 to 1
  - If the signal is a clock, α = 1
  - If the signal switches once per cycle, α = 0.5

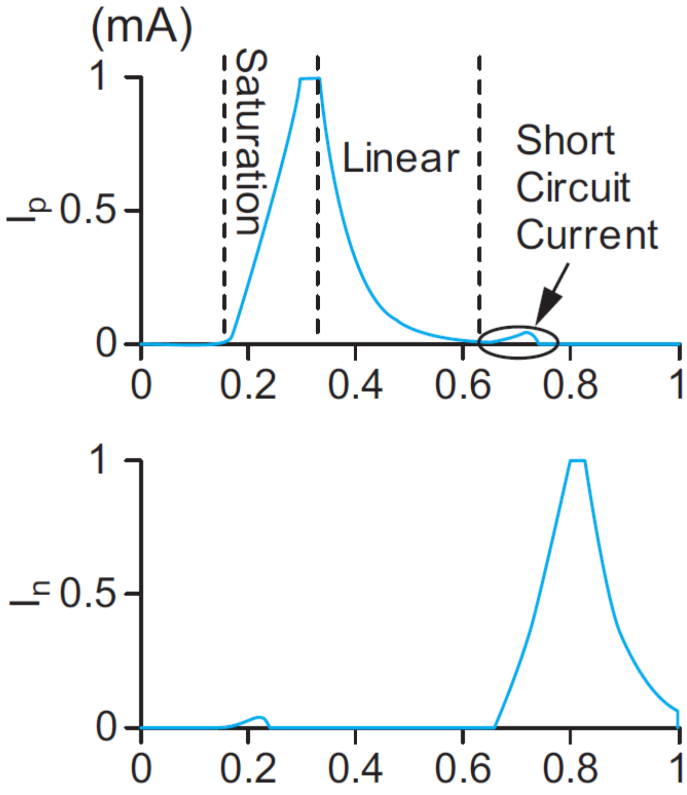
- ❑ Dynamic power:

$$P_{sw} = C_L V_{DD}^2 f_{sw} = \alpha C_L V_{DD}^2 f_{clk} = C_{eff} V_{DD}^2 f_{clk}$$

Short Circuit Current

- When transistors switch, both NMOS and PMOS networks may be momentarily ON at once
- Leads to a blip of “short circuit” current.
- Controlled by $\frac{V_t}{V_{DD}}$ ratio
  - Also depends on rise/fall times
- For $\frac{V_t}{V_{DD}} \approx 0.3$:
  - 2-10% of dynamic power
- Less important for nanometer technologies
- We will generally ignore this component



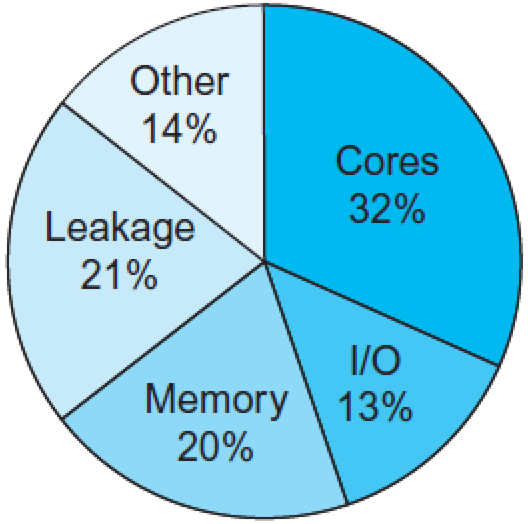


Power Dissipation Sources

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit current
- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current (in ratioed logic, to be discussed later)

Power Consumption Distribution

- In old technologies:
  - Roughly one-third of microprocessor power is spent on the clock
  - Another third on memories
  - The remaining third on logic and wires
- In nanometer technologies:
  - Nearly one-third of the power is leakage
  - High-speed I/O contributes a growing component too
- Example:
  - Active power consumption of Sun's 8-core 84 W Niagra2 processor
  - The cores and other components collectively account for clock, logic, and wires



Dynamic Power Example

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 λ
    - Activity factor = 0.02 (only necessary bank is activated)
  - 1.0 V 65 nm process
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate dynamic power consumption @ 1 GHz.
  - Neglect wire capacitance and short-circuit current.

Dynamic Power Example

- 50M logic transistors
  - Average width: 12 λ
  - Activity factor = 0.1
- 950M memory transistors
  - Average width: 4 λ
  - Activity factor = 0.02 (only necessary bank is activated)
- 1.0 V 65 nm process
- C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)

$$C_{\text{logic}} = \left(50\times10^{6}\right)\left(12\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\, fF/\mu m\right) = 27\text{ nF}$$

$$C_{\text{mem}} = \left(950\times10^{6}\right)\left(4\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\, fF/\mu m\right) = 171\text{ nF}$$

$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.02C_{\text{mem}}\right]\left(1.0\right)^{2}\left(1.0\text{ GHz}\right) = 6.1\text{ W}$$

Dynamic Power Reduction

$$P_{sw} = \alpha C_L V_{DD}^2 f_{clk}$$

❑ Try to minimize:

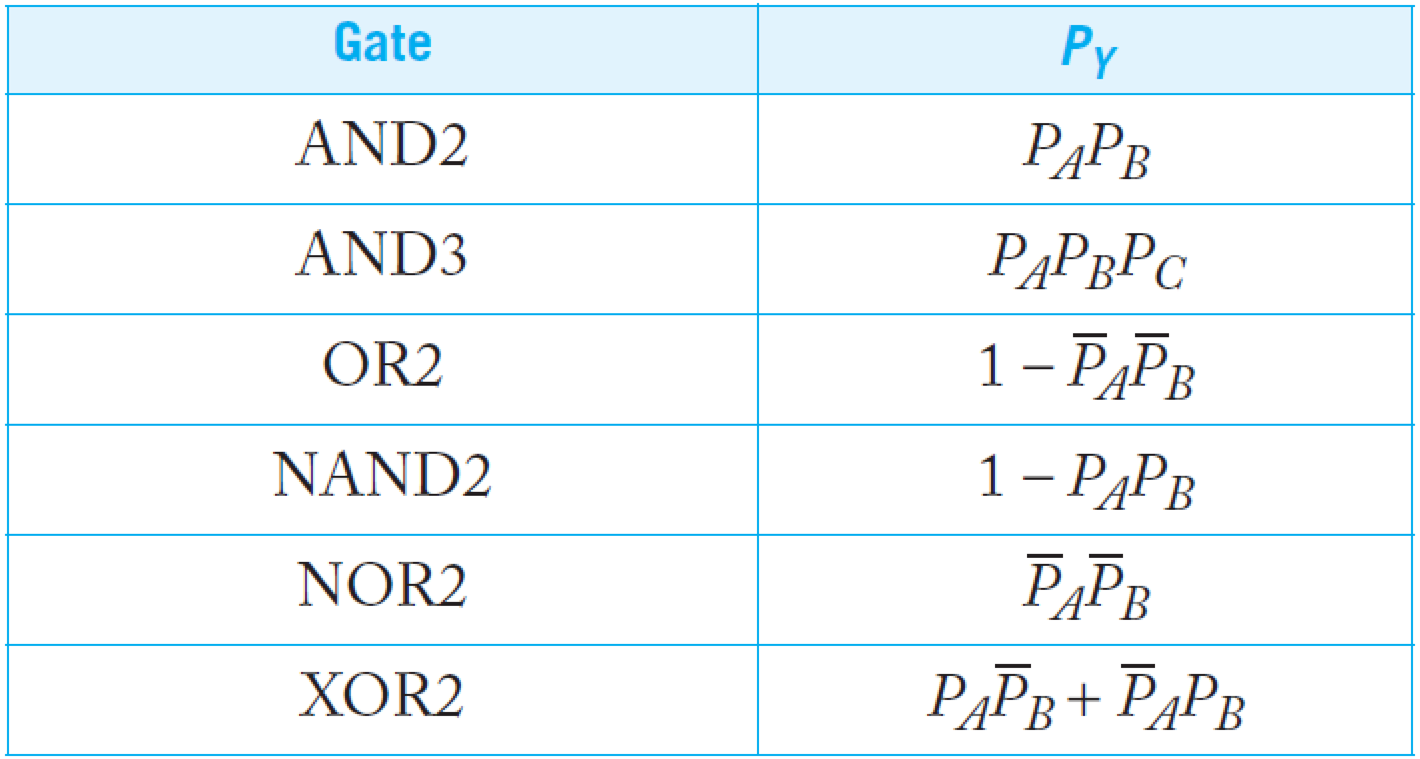
1. Activity factor
2. Capacitance
3. Supply voltage
4. Frequency

Activity Factor Estimation

- ❑ Let P_i = Probability(node i = 1)
  - $\overline{P_i}$= 1-P_i
- ❑ α_i = P_i * $\overline{P_i}$ = $P_i(1 - P_i)$
- ❑ Completely random data has P = 0.5 and α = 0.25
- ❑ Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0 ☺
- ❑ Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

Activity Factor Estimation

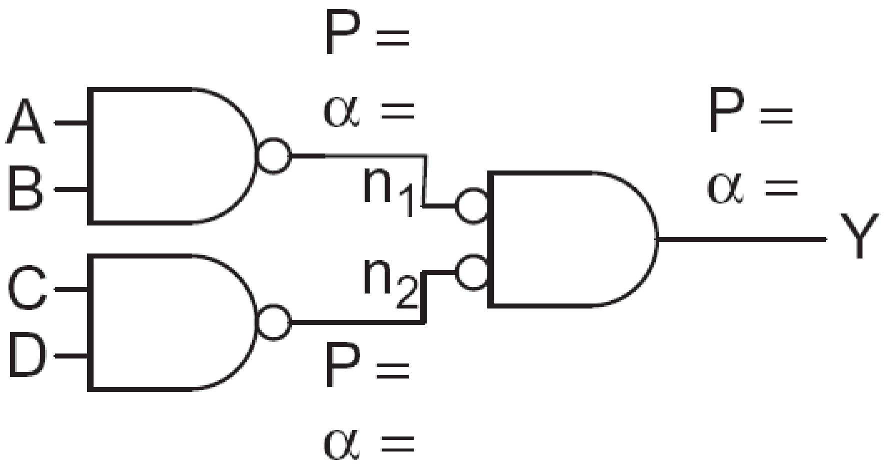
- ❑ Let P_i = Probability(node i = 1)
  - ▪ $\overline{P}_i$= 1-P_i
- ❑ α_i = P_i * $\overline{P}_i$ = $P_i(1-P_i)$



Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1-\overline{P}_A\overline{P}_B$
NAND2	$1-P_A P_B$
NOR2	$\overline{P}_A\overline{P}_B$
XOR2	$P_A\overline{P}_B+\overline{P}_A P_B$

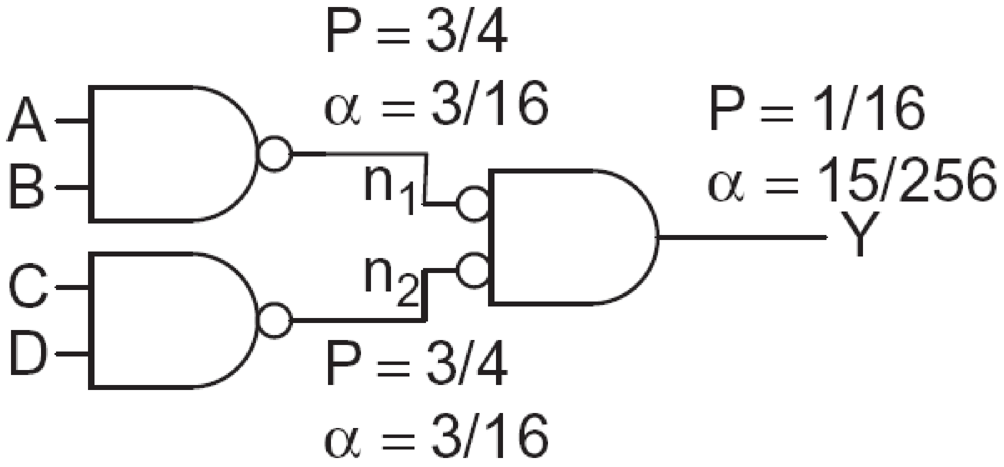
Example

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



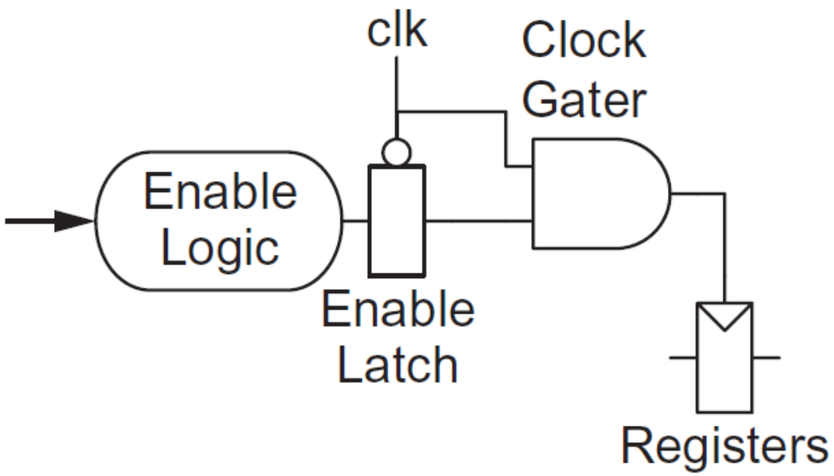
Example

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



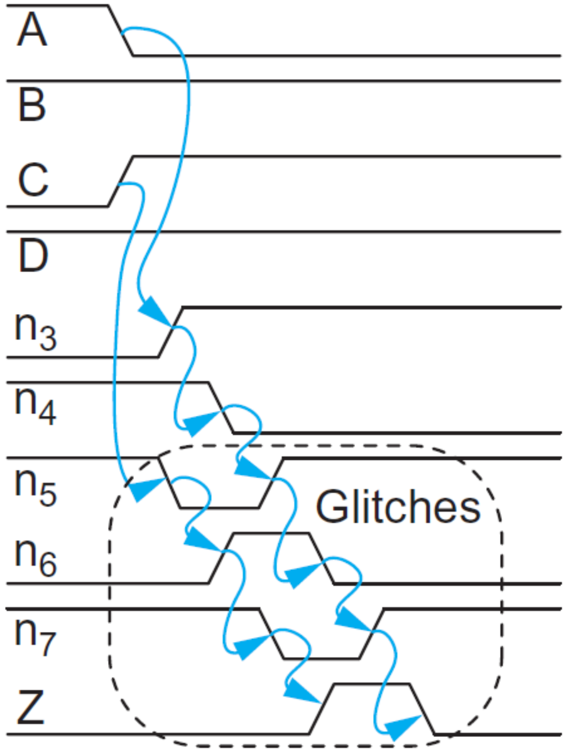
Clock Gating

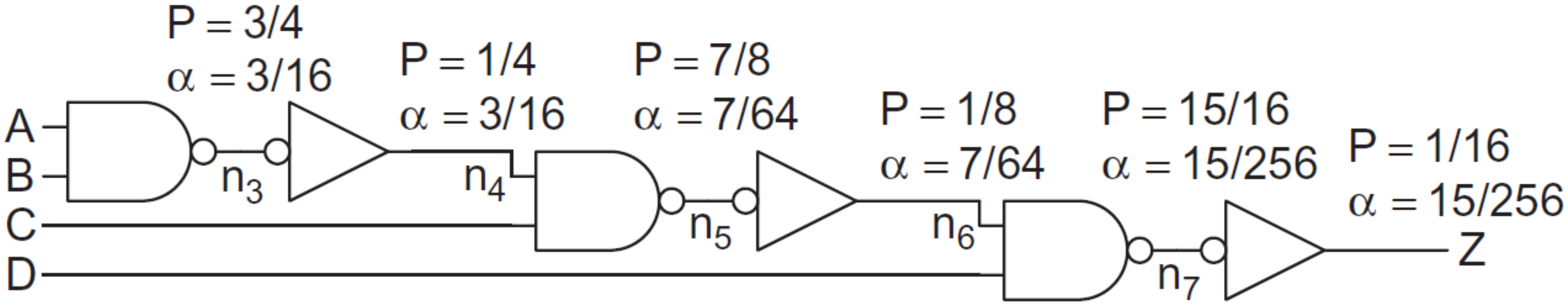
- The best way to reduce the activity is to turn off the clock to registers in unused blocks
  - Saves clock activity ($\alpha = 1$)
  - Eliminates all switching activity in the block
- Why directly ANDing the clock is a bad idea?



Glitches

- ❑ Gates will sometimes make spurious transitions called glitches.
- ❑ Glitches: When a single input change causes an output to change multiple times
- ❑ Example: ABCD from 1101 to 0111
- ❑ Glitches cause extra power dissipation (increase in activity factor) especially in chains of gate
- ❑ Causes majority of power in ripple carry adders and array multipliers





Capacitance

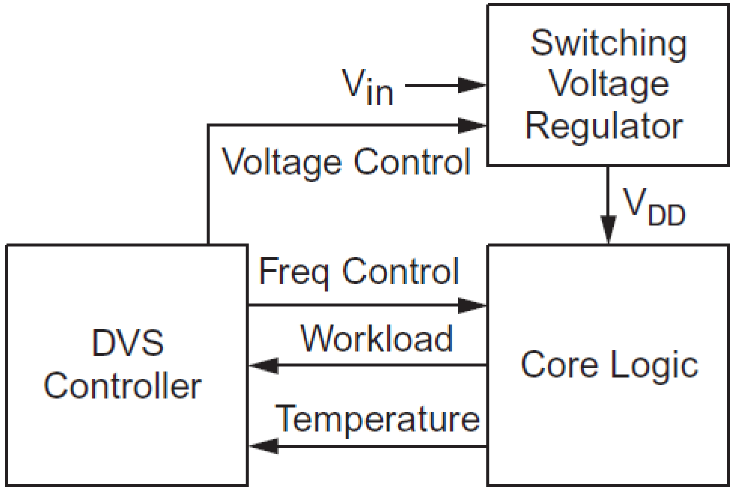
- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
  - To reduce power use $f > \hat{f} = 4$
  - For driving IO pads use $f = 8 \rightarrow 12$
  - Example: In a 64-bit adder, relaxing delay requirement by 10% can save 55% of energy!
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Voltage / Frequency

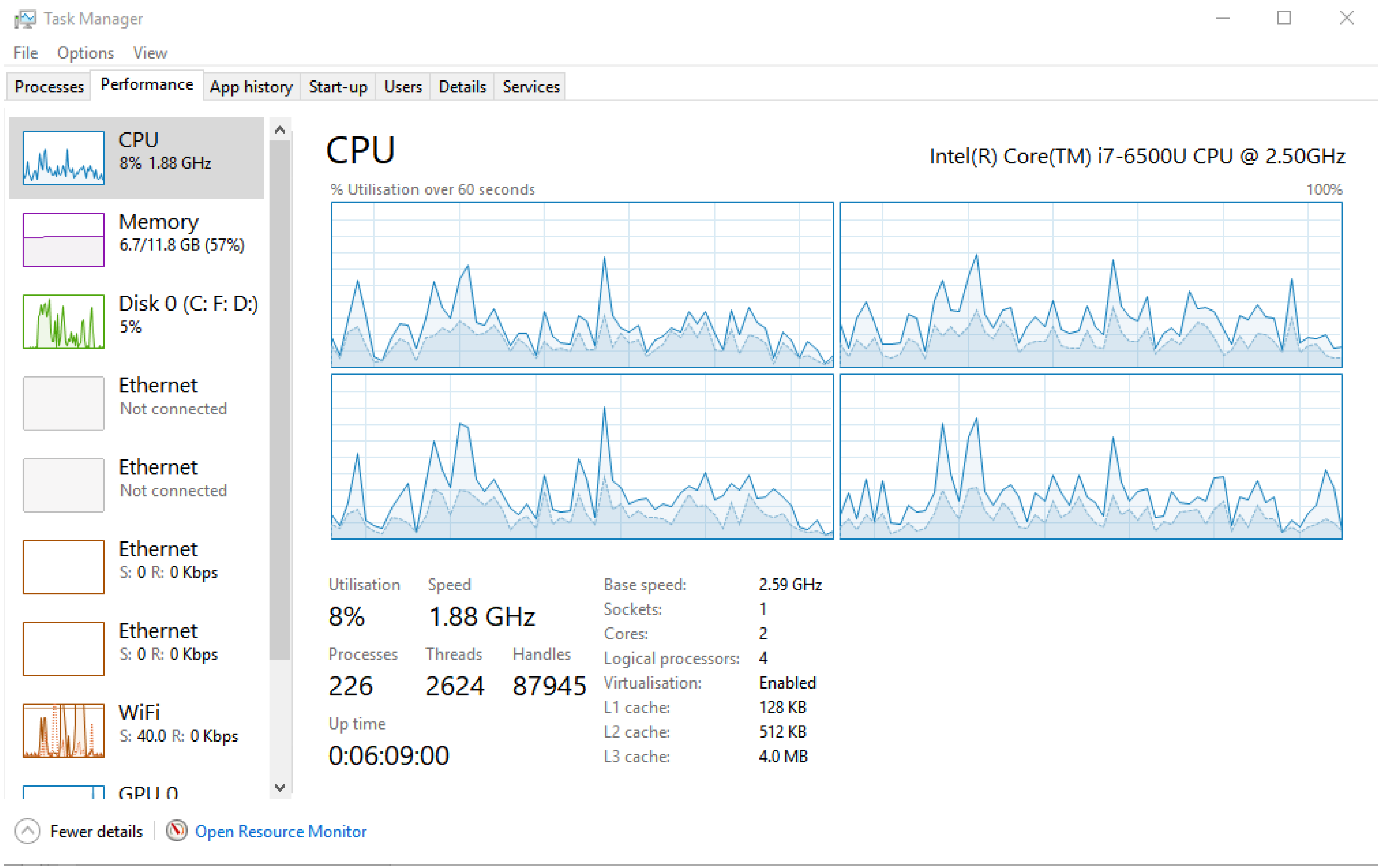
- ❑ If the frequency and voltage scale down in proportion, a cubic reduction in power is achieved.
- ❑ Run each block at the lowest possible voltage and frequency that meets performance requirements

- ❑ Voltage Domains
  - Provide separate supplies to different blocks
  - Level converters (shifters) required when crossing from low to high V_{DD} domains

- ❑ Dynamic Voltage/Frequency Scaling (DVFS)
  - Adjust V_{DD} and f_{clk} according to workload



Dynamic Voltage/Frequency Scaling (DVFS)



CPU
8% 1.88 GHz

Memory
6.7/11.8 GB (57%)

Disk 0 (C: F: D:)
5%

Ethernet
Not connected

Ethernet
Not connected

Ethernet
S: 0 R: 0 Kbps

Ethernet
S: 0 R: 0 Kbps

WiFi
S: 40.0 R: 0 Kbps

GPU 0

CPU

Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz

% Utilisation over 60 seconds

100%

Utilisation
8%

Speed
1.88 GHz

Processes
226

Threads
2624

Handles
87945

Up time
0:06:09:00

Base speed:	2.59 GHz
Sockets:	1
Cores:	2
Logical processors:	4
Virtualisation:	Enabled
L1 cache:	128 KB
L2 cache:	512 KB
L3 cache:	4.0 MB

Static Power

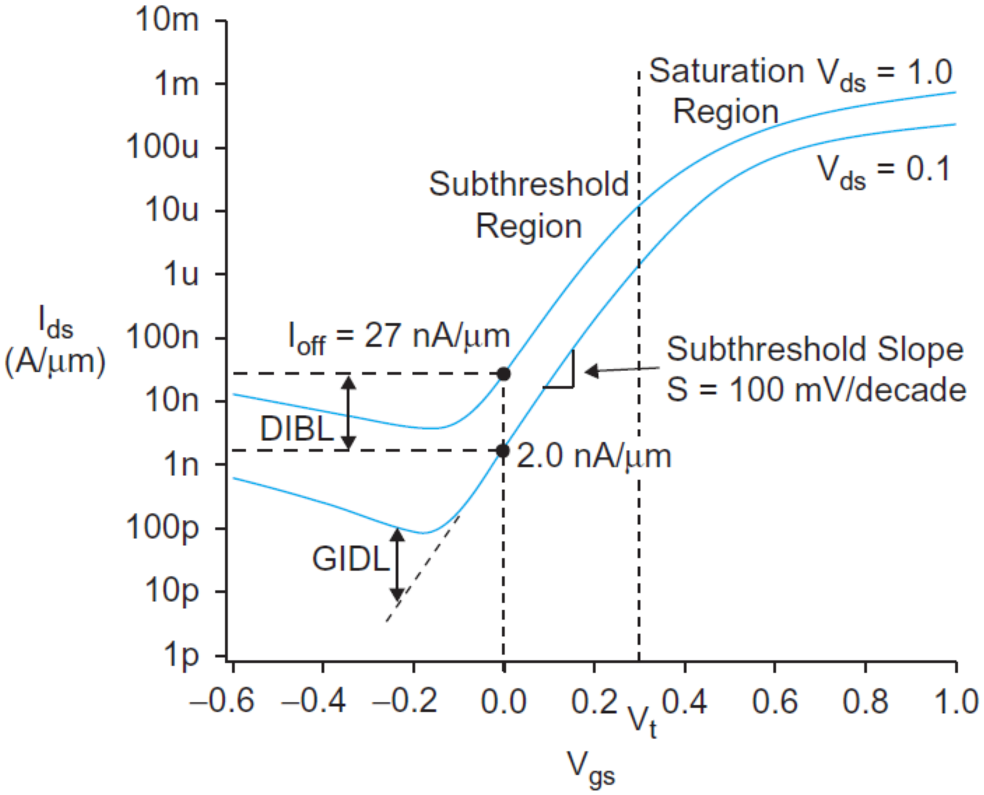
- Static power is consumed even when chip is idle.
  - Leakage draws power from nominally OFF devices
- Prior to the 90 nm node
  - Leakage power was negligible compared to dynamic power
  - It was of concern primarily during sleep mode
- In nanometer processes
  - Low threshold voltages and thin gate oxides
  - Leakage can account for as much as a third of total active power

Subthreshold Leakage

- ❑ For $V_{ds} > 2v_T = 2\frac{kT}{q} \approx 50mV$:

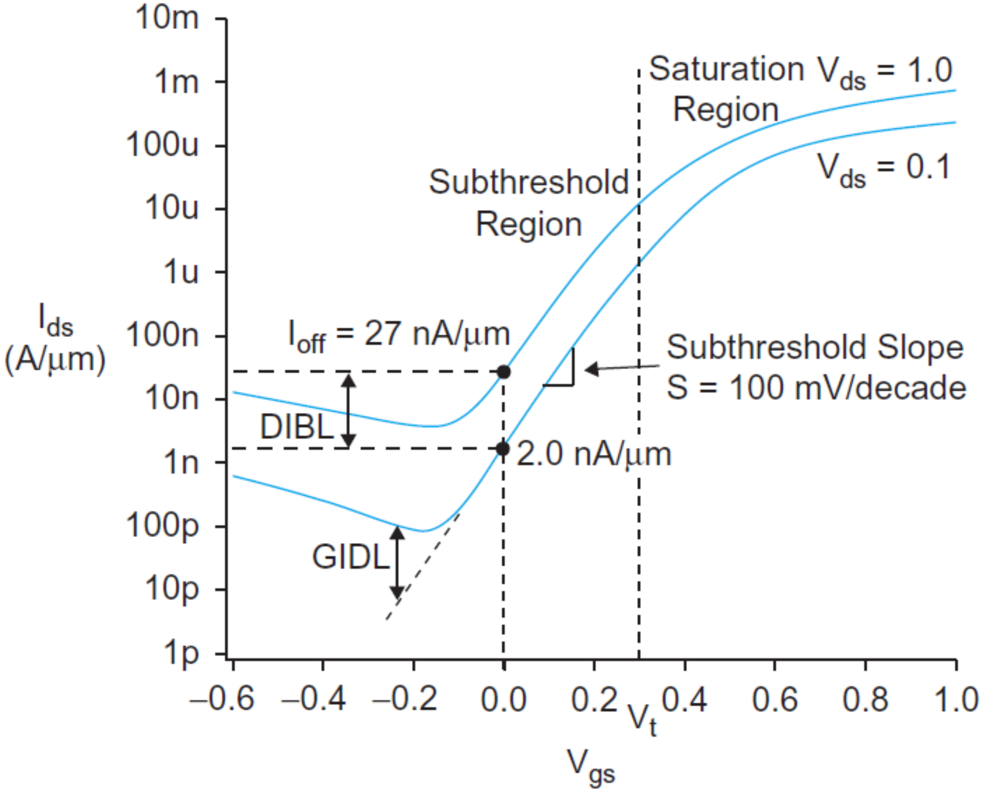
$$I_{\text{sub}} = I_{\text{off}} 10^{\frac{V_{gs} + \eta\left(V_{ds} - V_{DD}\right) - k_\gamma V_{sb}}{S}}$$

- ❑ I_{off}: Subthreshold leakage at $V_{gs} = 0$ and $V_{ds} = V_{dd}$
  - $I_{off} \propto e^{\frac{-V_t}{nv_T}}$
  - Specified at 25^oC, increases exponentially with temperature
  - $S = nv_T \ln 10 = nv_T / \log e = 2.3nv_T$: Subthreshold slope ~ 100mV/decade
- ❑ η: DIBL coefficient ~ 0.1V/V
- ❑ k_γ: Body effect coefficient ~ 0.1V/V



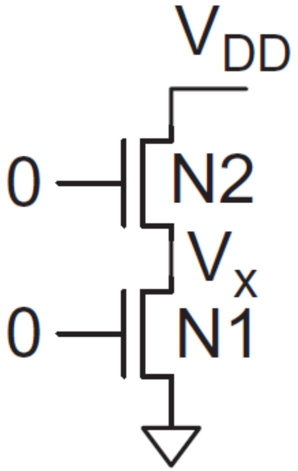
Multi-Threshold CMOS (Multi-$\boldsymbol{V_t}$)

- ❑ V_t can be tuned by adjusting the doping of the channel (beneath the gate oxide)
  - Adds additional photolithography and ion implantation steps
- ❑ $I_{off} \propto e^{\frac{-V_t}{n v_T}}$: Subthreshold leakage current at $V_{gs} = 0$ and $V_{ds} = V_{dd}$
  - Example (65nm technology):
    - Low V_t: $I_{off} \approx 100nA/\mu m$
    - Normal V_t: $I_{off} \approx 10nA/\mu m$
    - High V_t: $I_{off} \approx 1nA/\mu m$
- ❑ Use High V_t as default
  - Use Normal/Low V_t in critical paths only (to fix timing violations)



Stack Effect

- Series OFF transistors have less leakage
  - V_x is small
    - N1 will see lower DIBL
    - N1 will leak less
  - $V_x > 0$
    - N2 has negative V_{gs}
    - N2 will leak less
  - The smaller of N1 and N2 leakage will control the series path
  - Leakage through 2-stack reduces ~10x
  - Leakage through 3-stack reduces further
- Power gating makes use of this effect (also has higher V_t)



Static Power Calculation

- ❑ Estimate the total width of transistors that are leaking
- ❑ Multiply by the leakage current per unit width
- ❑ Assume half transistors OFF → Subthreshold leakage
- ❑ Assume half transistors ON → Gate leakage

Static Power Example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - Subthreshold leakage
    - Normal V_t: 100 nA/μm
    - High V_t: 10 nA/μm
    - High V_t used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible

Static Power Example

- 50M logic transistors: Average width: 12 λ
- 950M memory transistors: Average width: 4 λ
- Subthreshold leakage
  - Normal V_t: 100 nA/μm
  - High V_t: 10 nA/μm
  - High V_t used in all memories and in 95% of logic gates
- Gate leakage 5 nA/μm

$$W_{\text{normal-V}_t} = \left(50\times10^6\right)\left(12\lambda\right)\left(0.025\mu\text{m}/\lambda\right)\left(0.05\right) = 0.75\times10^6\ \mu\text{m}$$

$$W_{\text{high-V}_t} = \left[\left(50\times10^6\right)\left(12\lambda\right)\left(0.95\right)+\left(950\times10^6\right)\left(4\lambda\right)\right]\left(0.025\mu\text{m}/\lambda\right) = 109.25\times10^6\ \mu\text{m}$$

$$I_{sub} = \left[W_{\text{normal-V}_t}\times100\ \text{nA}/\mu\text{m}+W_{\text{high-V}_t}\times10\ \text{nA}/\mu\text{m}\right]/2 = 584\ \text{mA}$$

$$I_{gate} = \left[\left(W_{\text{normal-V}_t}+W_{\text{high-V}_t}\right)\times5\ \text{nA}/\mu\text{m}\right]/2 = 275\ \text{mA}$$

$$\text{P}_{static} = \left(584\ \text{mA}+275\ \text{mA}\right)\left(1.0\ \text{V}\right) = 859\ \text{mW}$$

Thank you!