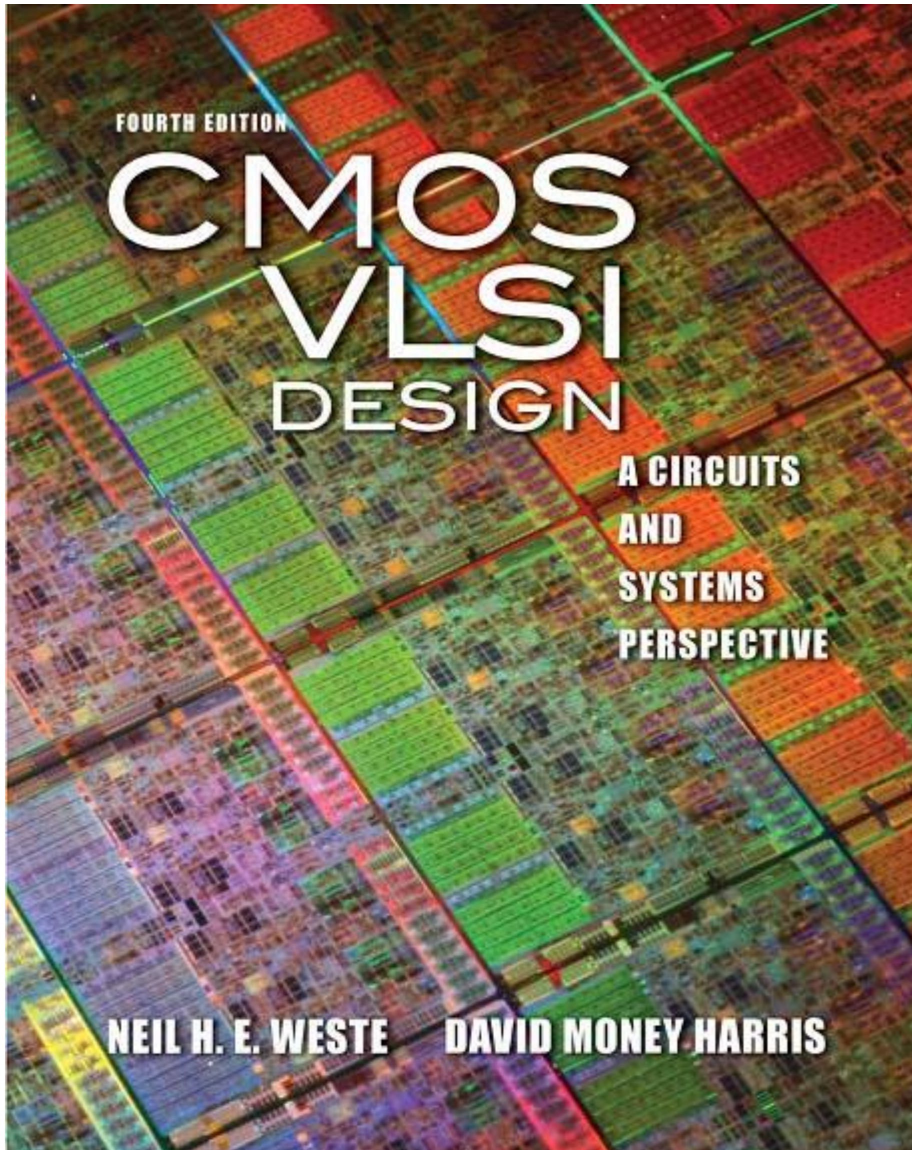


Lecture 7: Power

Outline

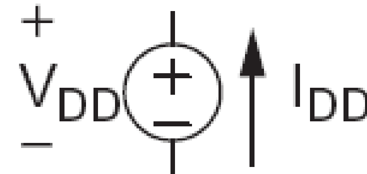
- Power and Energy
- Dynamic Power
- Static Power

Power and Energy

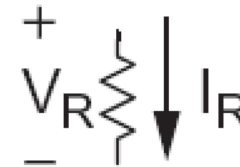
- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.
- Instantaneous Power: $P(t) =$
- Energy: $E =$
- Average Power: $P_{\text{avg}} =$

Power in Circuit Elements

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

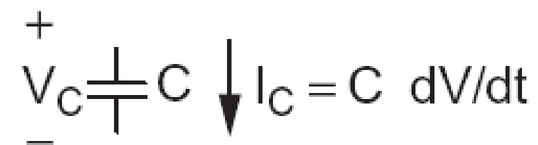


$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t)R$$



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

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



Charging a Capacitor

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

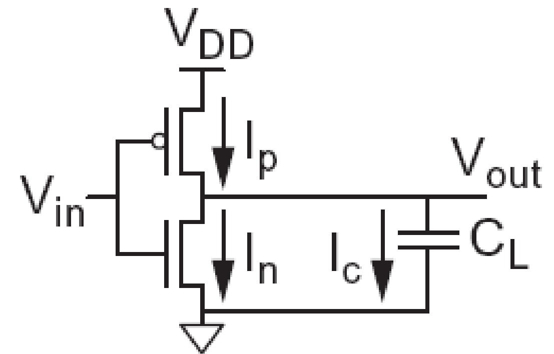
$$E_C = \tfrac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

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

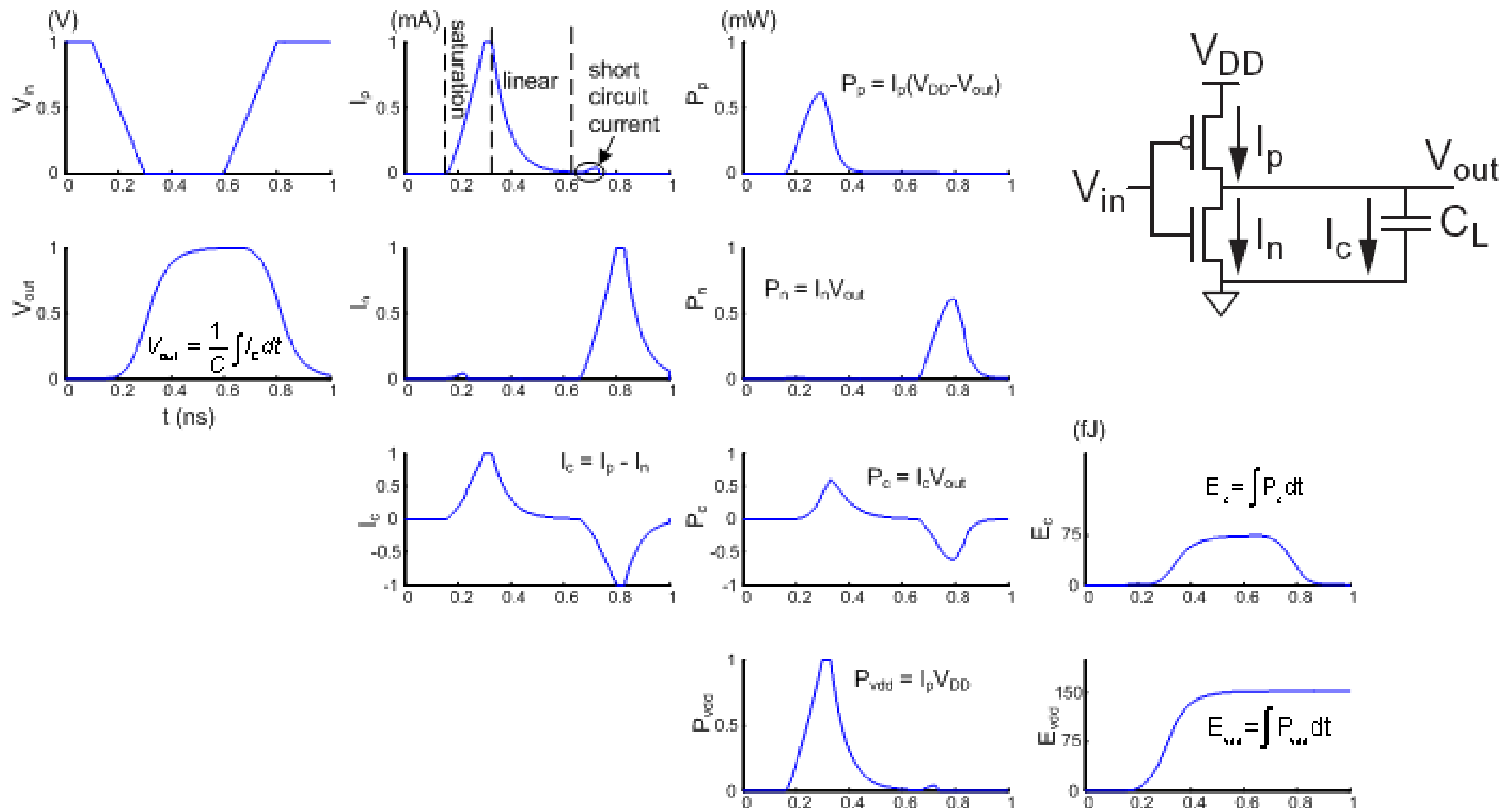
$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

  - 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



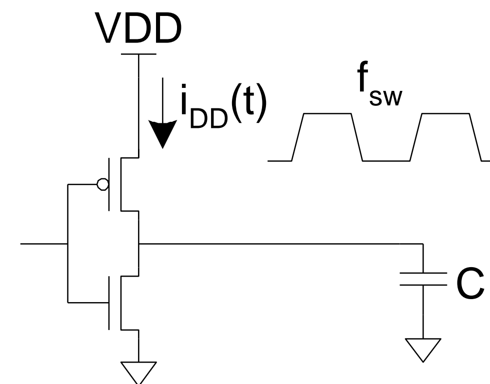
Switching Power

$$P_{\text{switching}} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}\,dt$$

$$= \frac{V_{DD}}{T}\int_0^T i_{DD}(t)\,dt$$

$$= \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right]$$

$$= CV_{DD}^{\ 2}f_{\text{sw}}$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$
- Dynamic power:

$$P_{\text{switching}} = \alpha C V_{DD}^2 f$$

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.
- < 10% of dynamic power if rise/fall times are comparable for input and output
- 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

Dynamic Power

- Consists of mainly switching power, short circuit power is neglected.
- To calculate dynamic power given V_{DD} and f, consider the capacitance of each node of the circuit including gate, diffusion, and wire capacitances.
- The effective capacitance is the true capacitance multiplied by the node activity factor.
- The switching power depends on the sum of the effective capacitances of all nodes.
- Activity factor is task-dependent.
- Low-power → minimize the power equation terms

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
  - 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.

Solution

$$C_{\text{logic}} = \left(50 \times 10^6\right)\left(12\lambda\right)\left(0.025\mu m / \lambda\right)\left(1.8 fF / \mu m\right) = 27 \text{ nF}$$

$$C_{\text{mem}} = \left(950 \times 10^6\right)\left(4\lambda\right)\left(0.025\mu m / \lambda\right)\left(1.8 fF / \mu m\right) = 171 \text{ nF}$$

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

Dynamic Power Reduction

- $P_{\text{switching}} = \alpha C V_{DD}^{\;2} f$
- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

Activity Factor Estimation

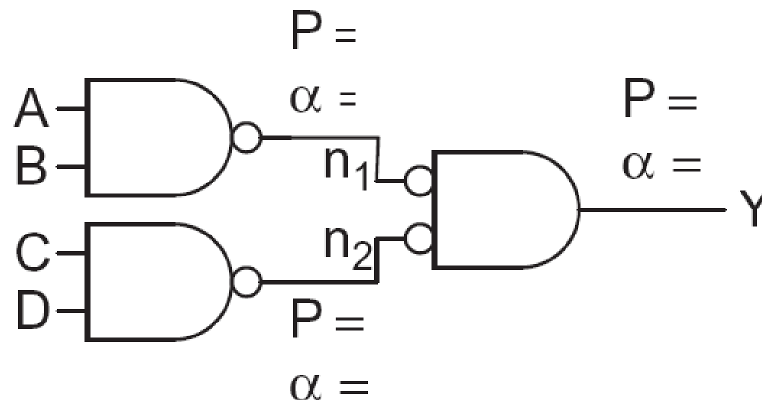
- Let P_i = Prob(node i = 1)
  - $\overline{P}_i = 1 - P_i$
- $\alpha_i = \overline{P}_i * P_i$
- Completely random data has P = 0.5 and α = 0.25
- Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0
- Data propagating through ANDs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

Switching Probability

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	$P_A \overline{P}_B + \overline{P}_A P_B$

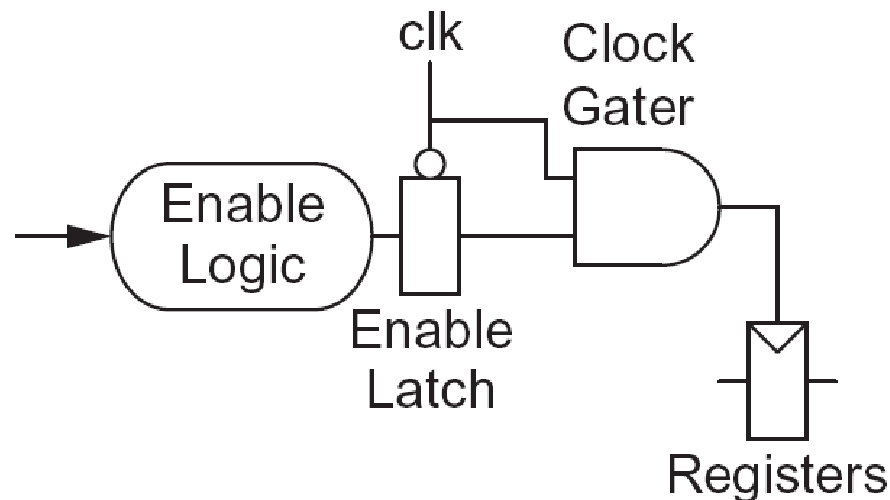
Example

- 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
- Construct the truth table and calculate the probabilities



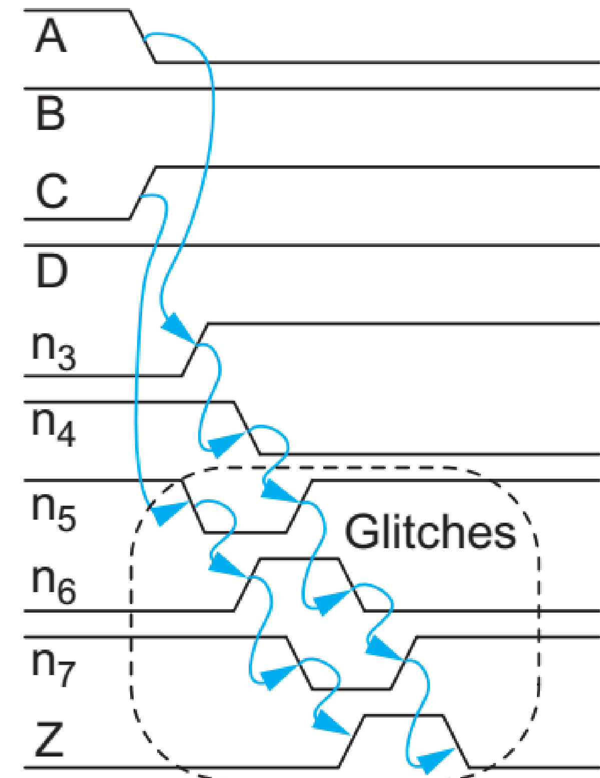
Clock Gating

- 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
  - Requires determining if block will be used



Glitches

- gates sometimes make spurious transitions called glitches when inputs do not arrive simultaneously
- The glitches cause extra power dissipation
- Chains of gates are particularly prone to this problem
- Glitching can raise the activity factor of a gate above 1



Capacitance

- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
  - Large gates with higher activity factors can be downsized to reduce power (at the expense of increasing logical effort and delay)
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Gate Sizing Under a Delay Constraint

- To compute energy in a circuit, consider:
  - a unit inverter has gate capacitance 3*C*,
  - a gate with logical effort *g*, parasitic delay *p*, and drive *x* has *gx* times as much gate capacitance and *px* times as much diffusion capacitance.
  - The energy of the entire circuit is the sum of the energies of each gate:

$$\text{Energy} = 3CV_{DD}^2 \sum_{i \in \text{nodes}} \alpha_i \left(\frac{C_{\text{wire}_i}}{3C} + p_i x_i + \sum_{j \in \text{fanout}(i)} g_j x_j \right)$$

Gate Sizing Under a Delay Constraint (2)

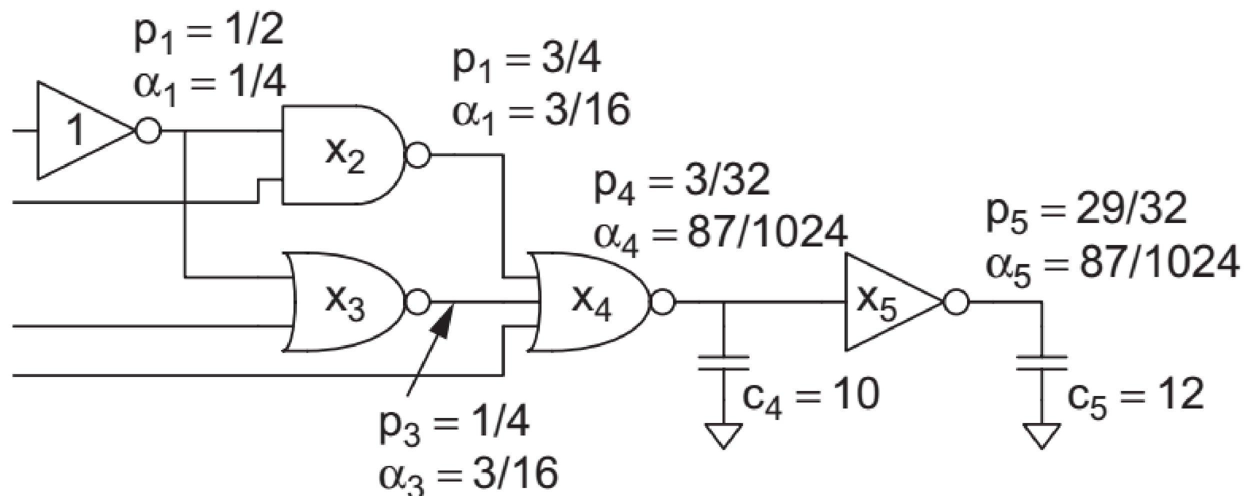
- By normalizing the equation:

$$E = \sum_{i \in \text{nodes}} \alpha_i \left(c_i + p_i x_i + \sum_{j \in \text{fanout}(i)} g_j x_j \right) = \sum_{i \in \text{nodes}} \alpha_i x_i d_i$$

- The problem is formulated as an optimization problem to minimize *E* such that the worst-case arrival time is less than some delay *D*.
- The problem is still a posynomial and has a unique solution that can be found quickly by a good optimizer.

Example

- Generate an energy-delay trade-off curve for the following circuit as delay varies from the minimum possible ($D_{\min}$ = 23.44 τ to 50 τ). Assume that the input probabilities are 0.5.

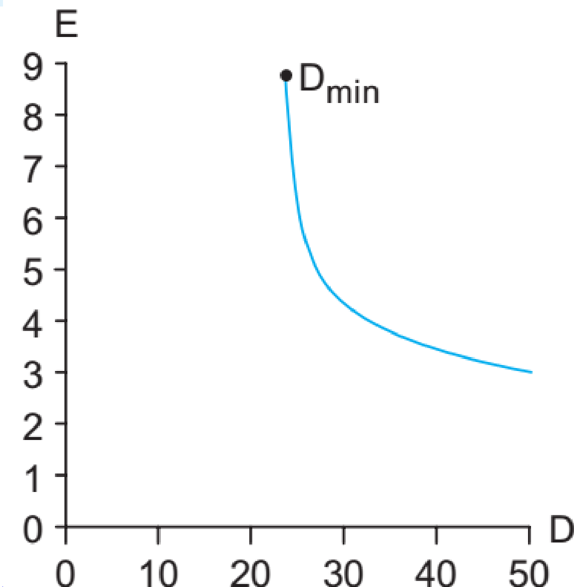


Solution

- The Energy of the circuit is:

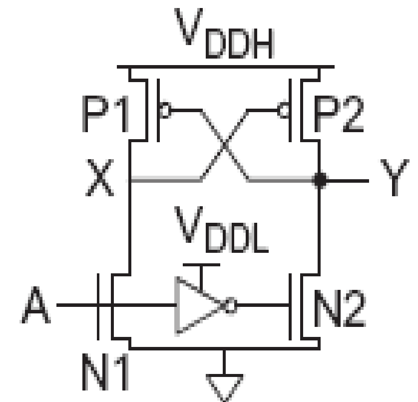
$$E = \frac{1}{4}\left(1 + \frac{4}{3}x_2 + \frac{5}{3}x_3\right) + \frac{3}{16}\left(2x_2 + \frac{7}{3}x_4\right) + \frac{3}{16}\left(2x_3 + \frac{7}{3}x_4\right)$$
$$+ \frac{87}{1024}\left(10 + 3x_4 + x_5\right) + \frac{87}{1024}\left(12 + x_5\right)$$

- The energy-delay trade-off curve obtained by an automatic Solver is depicted
- The delay cannot be minimized unless the input inverter size is increased



Voltage Domains

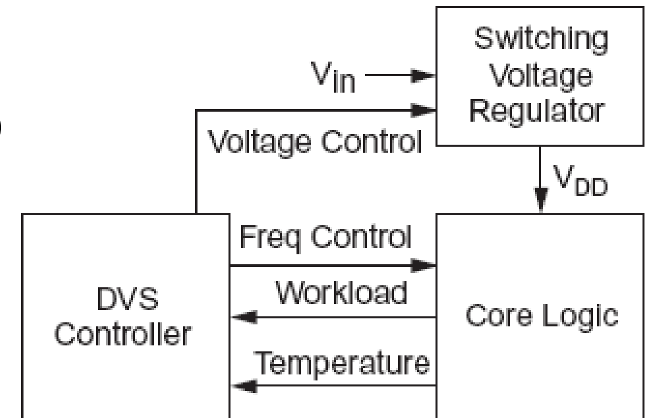
- Run each block at the lowest possible voltage and frequency that meets performance requirements
- Voltage Domains
  - Provide separate supplies to different blocks
  - Level converters required when crossing from low to high V_{DD} domains
  - Voltage domains are associated with a large area of the floorplan

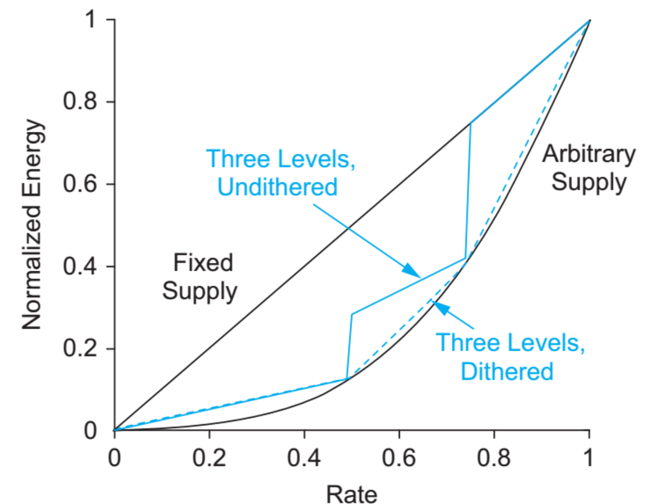


- Clustered Voltage Scaling (CVS) is an alternative approach to use two supply voltages in the same block with some constraints

Dynamic Voltage Scaling

- Dynamic Voltage Scaling (DVS)
  - Adjust V_{DD} and f according to workload
- DVFS
  - reducing the clock frequency to the minimum per task
  - reducing the supply voltage to the minimum necessary to operate at that frequency





Short-Circuit Current

- While the input switches, both pullup and pulldown networks are partially ON causing short-circuit current.
- It increases as the input edge rates become slower because both networks are ON for more time, and decreases as load capacitance increases.
- short-circuit current is a small fraction (< 10%) of current to the load and can be ignored for sharp input edges.
- Short-circuit power is strongly sensitive to the ratio $v = V_t / V_{DD}$, for $v=0.5$ short circuit current is zero.

Static Power

- Static CMOS gates have no contention current
- Static power is consumed even when chip is quiescent.
  - Leakage draws power from nominally OFF devices
  - Ratioed circuits burn power in flight between ON transistors

Subthreshold Leakage

- For $V_{ds} > 50$ mV

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

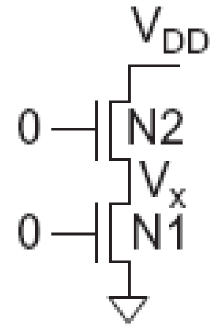
- I_{off} = leakage at $V_{gs} = 0$, $V_{ds} = V_{DD}$
  - η: the DIBL coefficient
  - K_γ: The body effect coefficient
  - S: Subthreshold slope

Typical values in 65 nm
I_{off} = 100 nA/μm @ V_t = 0.3 V
I_{off} = 10 nA/μm @ V_t = 0.4 V
I_{off} = 1 nA/μm @ V_t = 0.5 V
η = 0.1
k_γ = 0.1
S = 100 mV/decade

- I_{off} is usually specified at 25 °C and increases exponentially with temperature

Stack Effect

- Series OFF transistors have less leakage
  - V_x small, N_1 has low DIBL and small leak
  - $V_x > 0$, so N2 has negative V_{gs}



$$I_{sub} = \underbrace{I_{off} 10^{\frac{\eta(V_x - V_{DD})}{S}}}_{N1} = \underbrace{I_{off} 10^{\frac{-V_x + \eta((V_{DD} - V_x) - V_{DD}) - k_\gamma V_x}{S}}}_{N2}$$

$$V_x = \frac{\eta V_{DD}}{1 + 2\eta + k_\gamma}$$

$$I_{sub} = I_{off} 10^{\frac{-\eta V_{DD}\left(\frac{1+\eta+k_\gamma}{1+2\eta+k_\gamma}\right)}{S}} \approx I_{off} 10^{\frac{-\eta V_{DD}}{S}}$$

  - Leakage through 2-stack reduces ~10x
  - Leakage through 3-stack reduces further

Leakage Control

- Leakage and delay trade off
  - Aim for low leakage in sleep and low delay in active mode
- To reduce leakage:
  - Increase V_t: *multiple V_t*
    - Use low V_t only in critical circuits
  - Increase V_s: *stack effect*
    - *Input vector control* in sleep
  - Decrease V_b
    - *Reverse body bias* in sleep
    - Or forward body bias in active mode

Leakage Control (2)

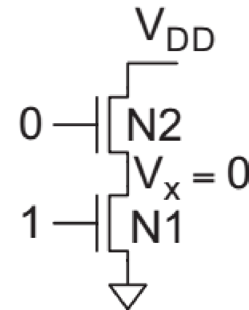
- Other forms of leakage must be considered to reduce Subthreshold leakage.
- Raising the doping level to raise V_t by controlling DIBL and short-channel effects increases BTBT.
- Applying a reverse body bias to increase V_t also causes BTBT to increase.
- Applying a negative gate voltage to turn the transistor OFF more strongly increases GIDL.
- Silicon on Insulator (SOI) circuits are attractive for low-leakage designs because they have a sharper subthreshold current roll-off.

Gate Leakage

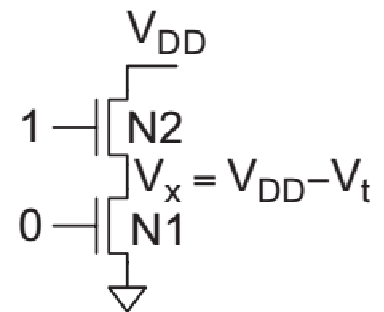
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using $t_{ox} > 10.5$ Å
  - High-k gate dielectrics help
  - Some processes provide multiple t_{ox}
    - e.g. thicker oxide for 3.3 V I/O transistors
- Control leakage in circuits by limiting V_{DD}

Gate Leakage (2)

- Gate leakage also depends on the voltage across the gate
- For the example in the figure
  - If $N1$ is ON and $N2$ is OFF, $N1$ has $V_{gs} = V_{DD}$ and has full gate leakage.
  - On the other hand, if $N1$ is OFF and $N2$ is on, $N2$ has $V_{gs} = V_t$ and experiences negligible gate leakage
  - In both cases, the OFF transistor has no gate leakage.
  - Thus, gate leakage can be alleviated by stacking transistors such that the OFF transistor is closer to the rail



(a)



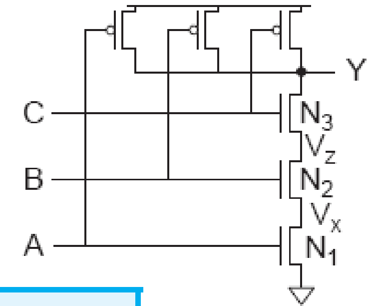
(b)

NAND3 Leakage Example

- 100 nm process

  $I_{gn} = 6.3$ nA $\quad I_{gp} = 0$

  $I_{offn} = 5.63$ nA $\quad I_{offp} = 9.3$ nA



Input State (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	stack effect	stack effect
001	0.7	0	0.7	stack effect	$V_{DD} - V_t$
010	0.7	1.3	2.0	intermediate	intermediate
011	3.8	0	3.8	$V_{DD} - V_t$	$V_{DD} - V_t$
100	0.7	6.3	7.0	0	stack effect
101	3.8	6.3	10.1	0	$V_{DD} - V_t$
110	5.6	12.6	18.2	0	0
111	28	18.9	46.9	0	0

Data from [Lee03]

Junction Leakage

- From reverse-biased p-n junctions
  - Between diffusion and substrate or well
- Ordinary diode leakage is negligible
- Band-to-band tunneling (BTBT) can be significant
  - Especially in high-V_t transistors where other leakage is small
  - Worst at $V_{db} = V_{DD}$
- Gate-induced drain leakage (GIDL) exacerbates
  - Worst for $V_{gd} = -V_{DD}$ (or more negative)

Static Power Estimation

- Static CMOS circuits have no contention current.
- Some other families inherently draw current even while quiescent. (e.g. pseudo nMOS logic)
- Static current is estimated by:
  - Estimate total width of transistors that are leaking,
  - multiplying by the leakage current per width,
  - and multiplying by the fraction of transistors that are in their leaky state (usually one half).
  - Add the contention current if applicable.
  - The static power is the supply voltage times the static current.

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 Vt used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible

Solution

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

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

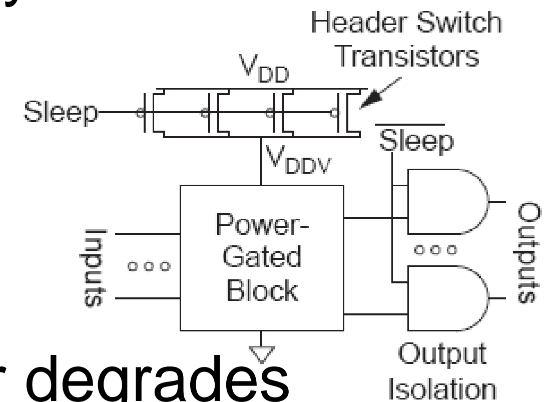
$$I_{sub} = \left[W_{\text{normal-V}_t}\times100\ \text{nA}/\mu\text{m}+W_{\text{high-V}_t}\times10\ \text{nA}/\mu\text{m}\right]/2 = 584\ \text{mA}$$

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

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

Power Gating

- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual V_{DD} (V_{DDV})
  - Gate outputs to prevent invalid logic levels to next block
- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize delay and voltage drop
  - Also, it should have low leakage during sleep
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough



Power Gating Design

- It can be done externally with a disable input to a voltage regulator or internally with high-V_t header or footer switches
- On-chip power gating can use pMOS header switch transistors or nMOS footer switch transistors
- *Fine-grained power gating* can be applied to individual logic gates, but placing a switch in every cell has enormous area overhead
- Practical designs use *coarse-grained power gating* where the switch is shared across an entire block
- The switch is commonly sized to keep this delay to 5–10%

Multiple Threshold Voltages

- Multiple threshold voltages can keep performance on critical paths with low-V_t transistors while reducing leakage on others with high-V_t transistors.
- Good design practice starts with high-V_t devices everywhere and selectively introduces low-V_t devices where necessary.
- Using multiple thresholds requires additional implant masks that add to the cost of a CMOS process.
- Alternatively, designers can increase the channel length, which tends to raise the threshold voltage via the short channel effect.

Variable Threshold Voltage

- V_{sb} controls the threshold voltage via the body effect
- In *variable threshold CMOS* (VTCMOS), a body bias is applied to achieve high I_{on} and low I_{off}
- For example, low-*Vt* devices can be used and a *reverse body bias* (RBB) can be applied during sleep mode to reduce leakage
- Alternatively, higher-*Vt* devices can be used, and then a *forward body bias* (FBB) can be applied during active mode to increase performance
- Improper body biasing can increase leakage via BTBT and junction leakage