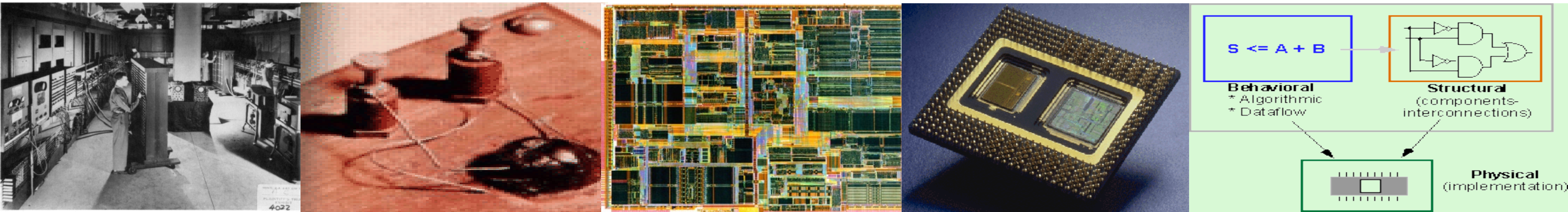
CpE 690: Introduction to VLSI Design

Lecture 12
Low Power Design

Bryan Ackland
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030

Adapted from Digital Integrated Circuits: A Design Perspective, Rabaey *et. al*., 2003
and Lecture Notes, David Mahoney Harris CMOS VLSI Design

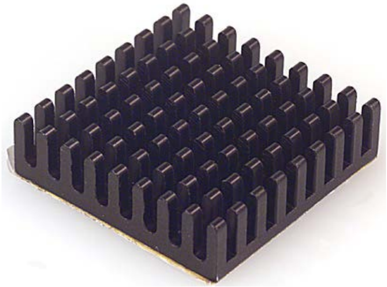


CMOS – a Low Power Technology

- CMOS developed in 1970’s as a low power technology
  - (almost) no DC current when gate is not switching
  - no static power dissipation
- CMOS replaces NMOS in 1980’s as dominant digital technology
  - NMOS designs dissipated about 200μW/gate
  - Power dissipation no longer an issue!
- CMOS process technology evolves to provide:
  - more transistors per chip (Moore’s Law)
  - faster switching speed (few MHz ⇨ hundreds of MHz)
- 1992 DEC announces Alpha 64-bit microprocessor
  - triumph of high speed CMOS digital design
  - first 200MHz processor, 1.7M transistors
  - **30W power dissipation !!**
  - Power dissipation is once again an issue!

Why Power Matters: Package & System Cooling

- Need to remove heat from high performance chips
  - max. operating temperature silicon transistors: 150 – 200 °C
  - above these temperatures, dopants diffuse

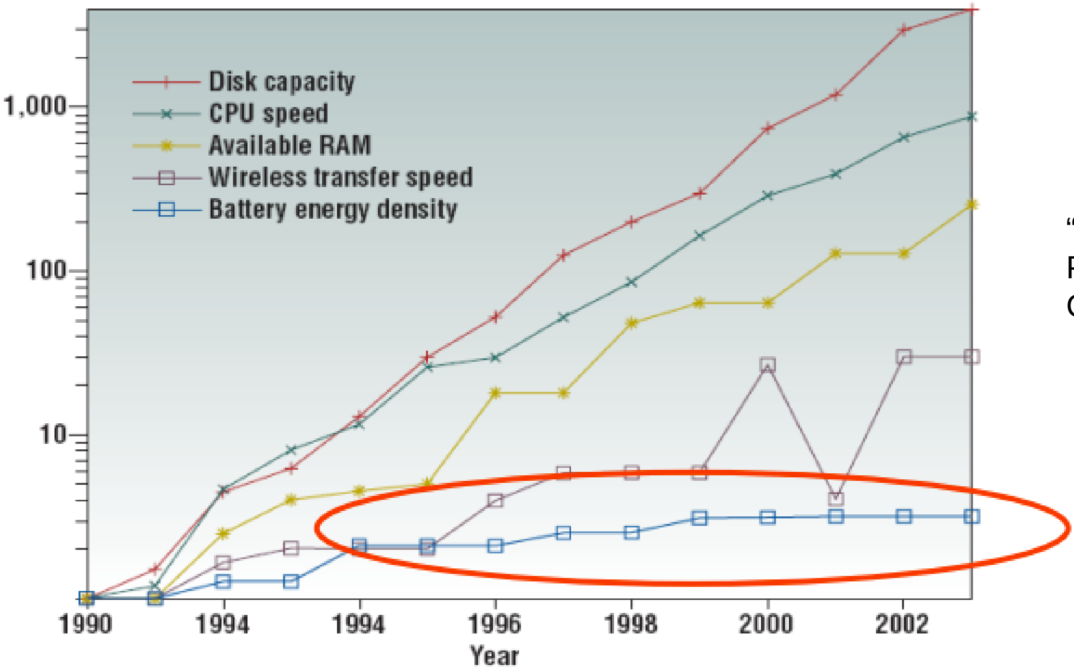


- Chip on PC board can dissipate 2-3 watts
- With suitable heatsink, maybe 10 watts
- With forced-air cooling (fans), up to 150W

- With sophisticated liquid cooling, maybe 1000W

Why Power Matters: Battery Size & Weight

- Today, we see more hand-held battery operated devices
- Unlike CMOS technology, battery technology has seen only modest improvements over last few decades



“Mobile Computing Environment”, Paradiso et. al. Pervasive Computing, IEEE 2005

- Expected battery lifetime increase over the next 5 years: 30 to 40%

Why Power Matters: Power Distribution

- Power Supply and Ground design
  - If VDD=1.0V, a 100W chip draws 100 amps!
  - Many package pins required
  - Virtex-6 1924-pin package:
    - 220 power and 484 GND pins
  - On-chip wiring distribute this current
  - Electro-migration issues

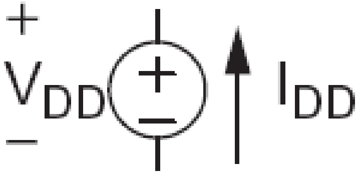
- On-chip noise and system reliability
  - Large currents switched through package and PCB inductance
- Environmental Concerns
  - Computers and consumer electronics account for 15% of residential energy consumption

Back to Basics: Power & Energy

- Power is drawn from a voltage source attached to the V_{DD} and GND pins of a chip.
- Instantaneous Power: $P(t) = I(t)V(t)$ (watts)
- Energy: $$E = \int_0^T P(t)\,dt$$ (joules)
- Average Power: $$P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T P(t)\,dt$$

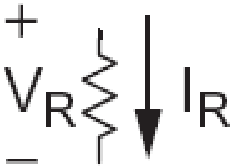
Back to Basics: Power in Circuit Elements

- Power Supply:



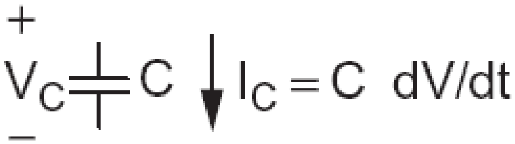
$$P_{VDD}(t) = I_{DD}(t)V_{DD}$$

- Resistor



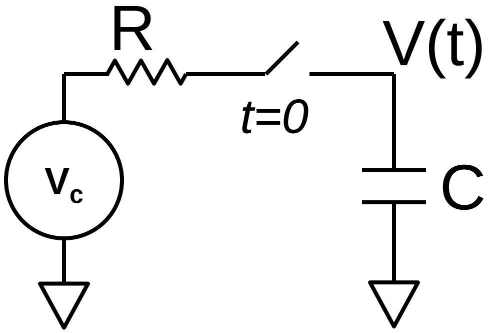
$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t)R$$

- Capacitor



Capacitors don't dissipate power!

  – but they do store energy:



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

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

Power Dissipation in CMOS

- $P_{total} = P_{dynamic} + P_{static}$

- 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: Charging a Capacitor

- When the gate output rises from GND to V_{DD}:
  - Energy stored in capacitor is

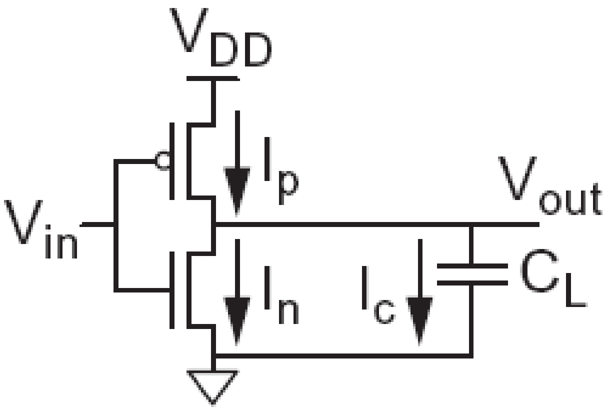
$$E_C = \tfrac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

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

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

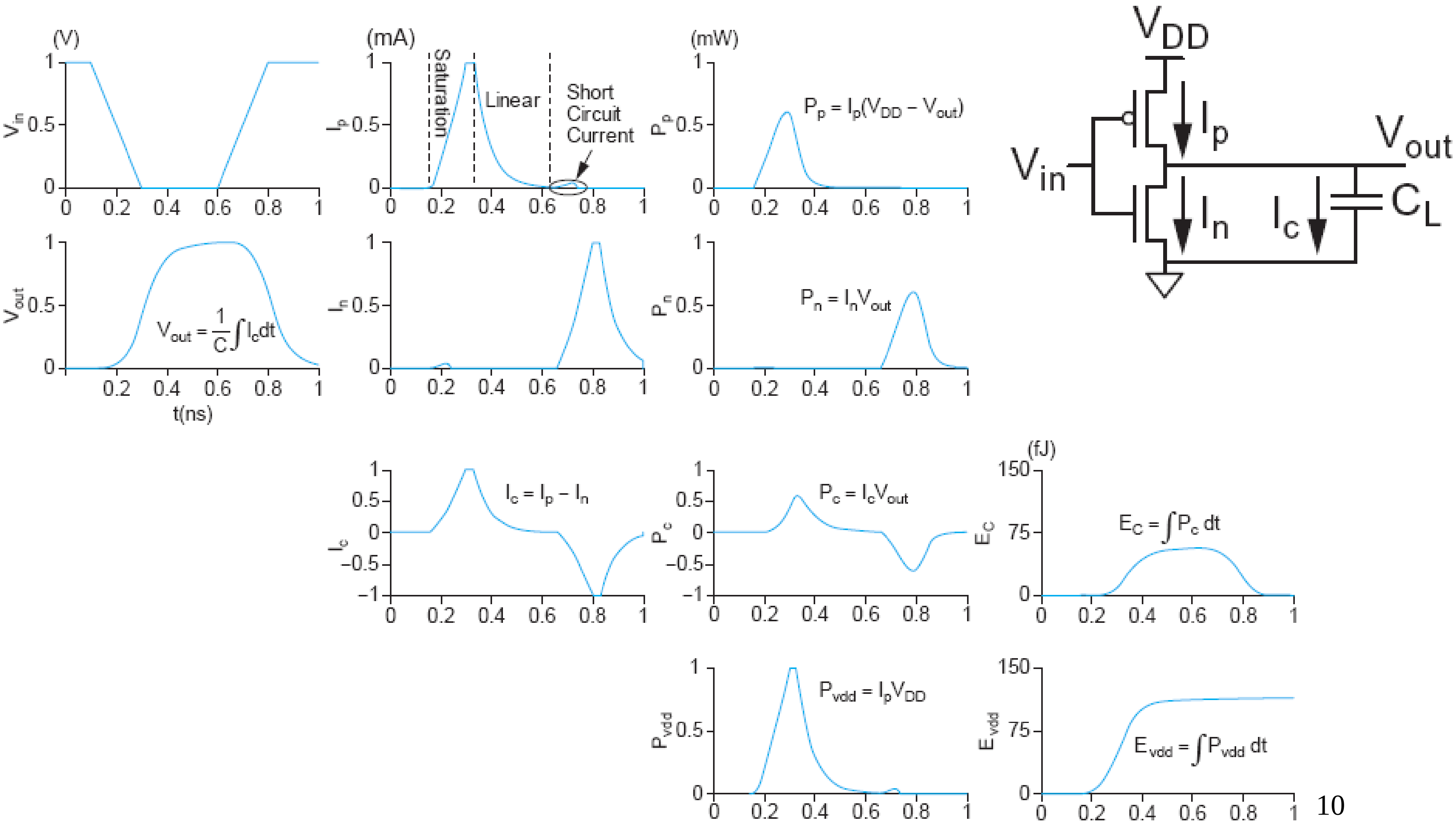
independent of size of transistors!



  - Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, other half stored in capacitor
- When the gate output falls from V_{DD} to GND
  - Stored 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 Waveforms

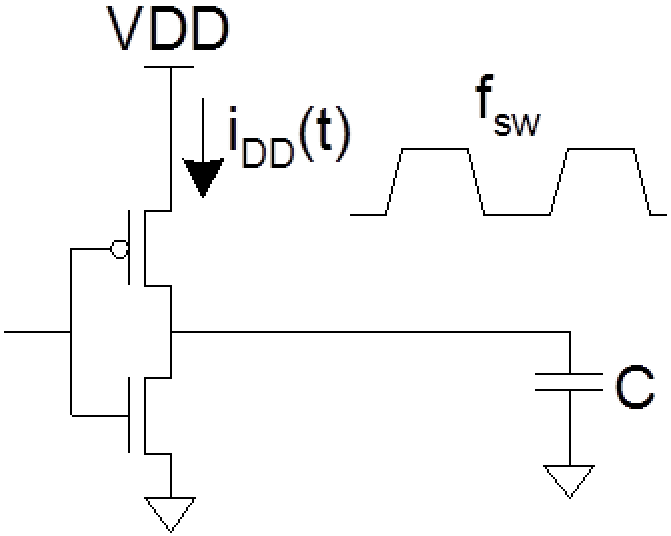
$$P_{switching} = \frac{1}{T}\int_0^T i_{DD}(t) V_{DD}\, dt$$

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

$$= \frac{V_{DD}}{T} \times \begin{bmatrix} total\ charge\ drawn \\ from\ power\ supply \\ in\ time\ T \end{bmatrix}$$

$$= \frac{V_{DD}}{T} \times [T f_{sw} C V_{DD}]$$

$$P_{switching} = C.{V_{DD}}^2.f_{sw}$$



Note: $P_{switching}$ is independent of drive strength of the nMOS and pMOS transistors

Activity Factor

- Suppose the system clock frequency = f
- Most gates do not switch every clock cycle
- Let $f_{sw} = \alpha f$, where α = activity factor
  - $\alpha = P_{0\rightarrow 1}$: probability that a signal switches from 0 to 1 in any clock cycle
  - If the signal is the system clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = 0.5$
  - If the signal is random (clocked) data, $\alpha = 0.25$
  - Static CMOS logic has (empirically) $\alpha \approx 0.1$

- Switching power of each node i is $P_i = \alpha_i . C_i . {V_{DD}}^2 . f$

$$\boxed{P_{switching} = \sum_i P_i = {V_{DD}}^2 . f . \sum_i \alpha_i . C_i}$$

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
  - 65 nm, 1.0V process (λ = 25nm)
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)

- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

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

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

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

Reducing Switching Power

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

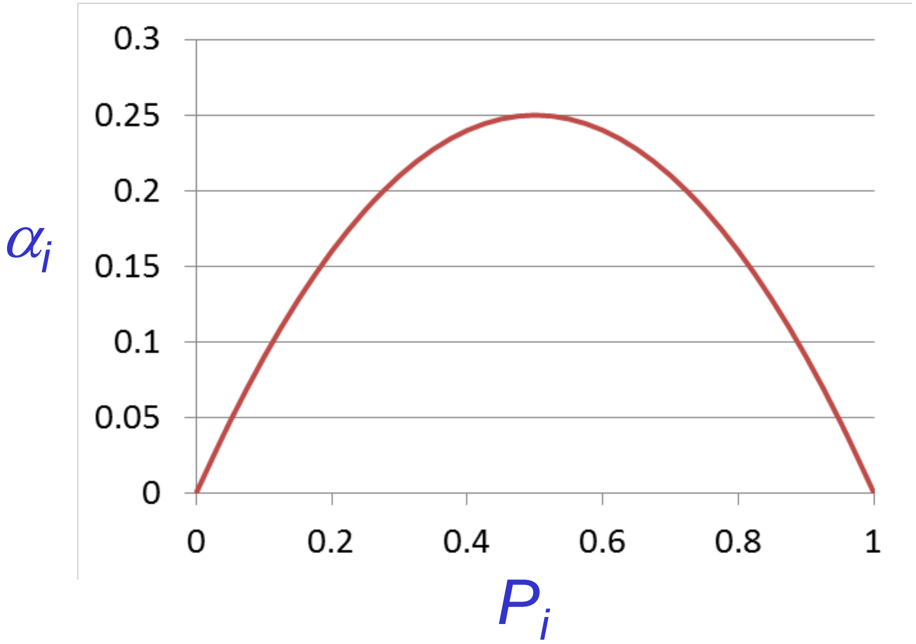
- So try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

Activity Factor Estimation

- Let P_i = probability (node *i* = 1)
  and $\overline{P}_i = (1 – P_i)$ = probability (node *i* = 0)

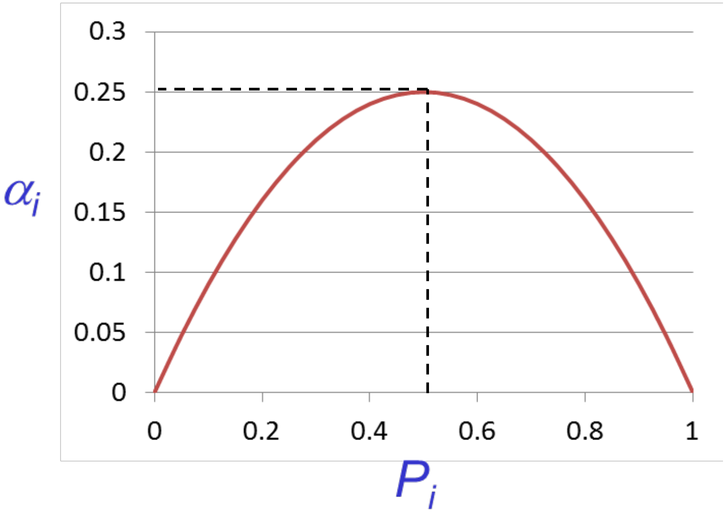
- α_i = prob. that node *i* makes a transition from 0 to 1, so

- $\alpha_i = \overline{P}_i \bullet P_i \ = (1 — P_i) \bullet P_i$



Activity Factor Estimation

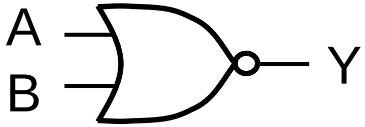
- For random data, $\alpha = 0.5 \cdot 0.5 = 0.25$



- 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

Example: Switching Probability of NOR2



- For NOR2, $P_Y = \overline{P}_A \bullet \overline{P}_B$

- $\overline{P}_Y = (1 - P_Y) = (1 - \overline{P}_A \bullet \overline{P}_B)$

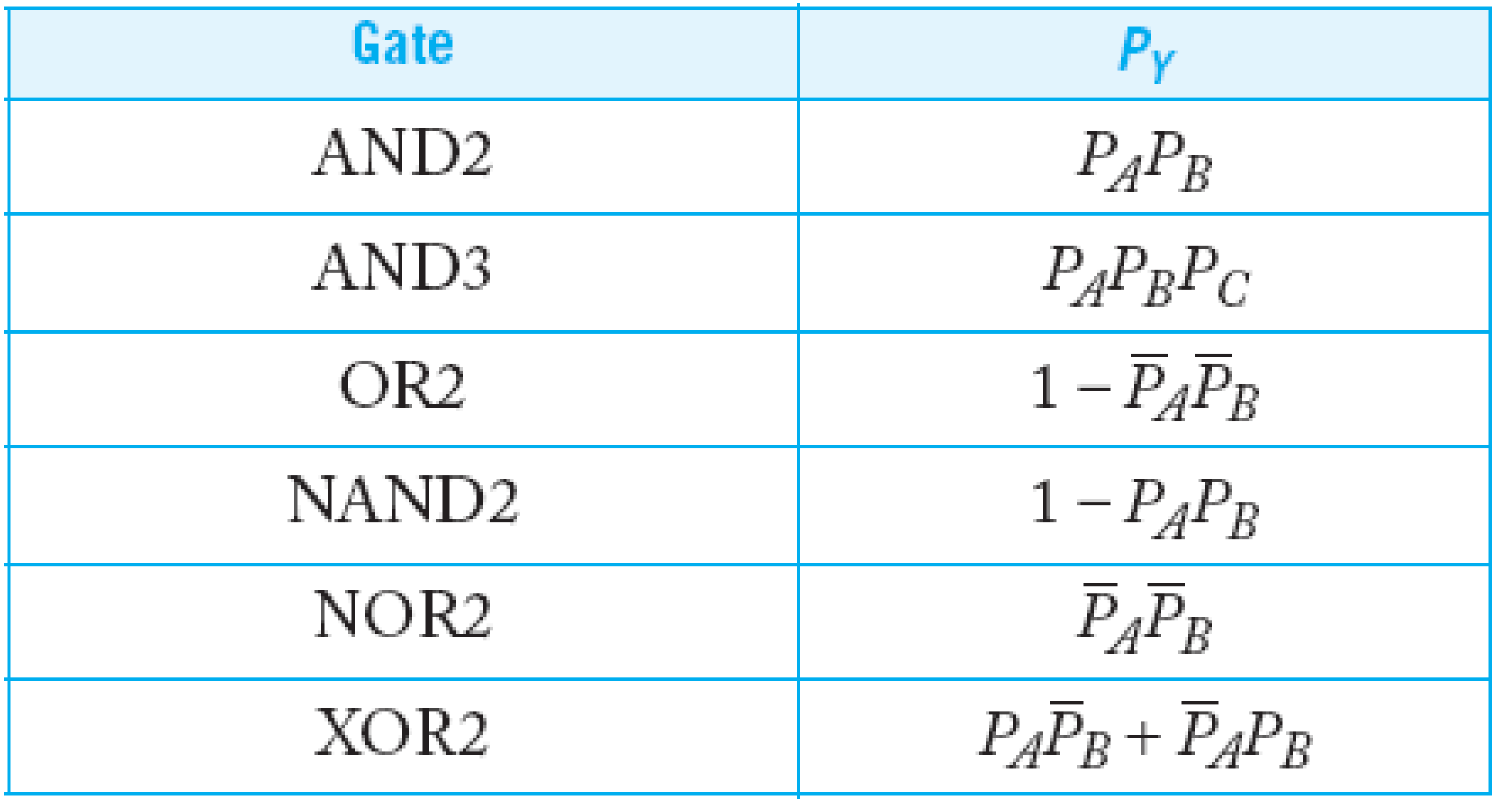
- $\alpha_Y = P_Y \bullet \overline{P}_Y$

  $= (\overline{P}_A \bullet \overline{P}_B) \bullet (1 - \overline{P}_A \bullet \overline{P}_B)$

A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0

- If $P_A = P_B = 0.5$, $P_Y = 0.25$, $\alpha_Y = 3/16 \approx 0.19$

Switching Probabilities (Static Gates)

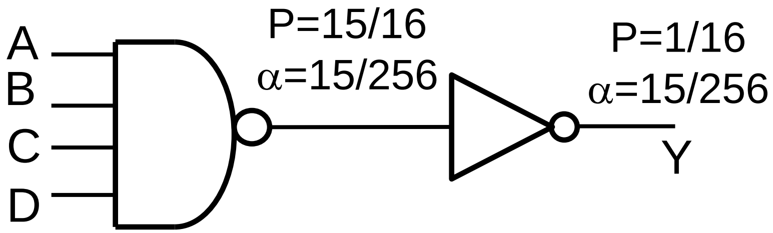


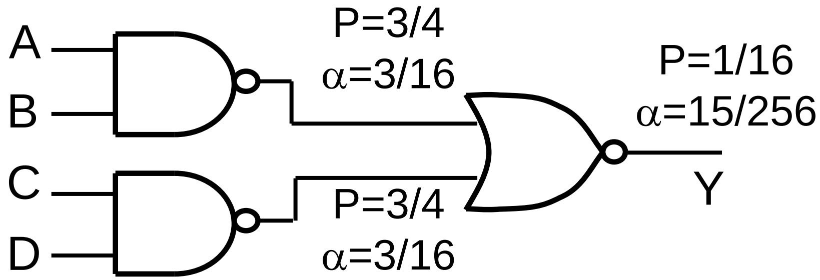
Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P}_A \overline{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P}_A \overline{P}_B$
XOR2	$P_A \overline{P}_B + \overline{P}_A P_B$

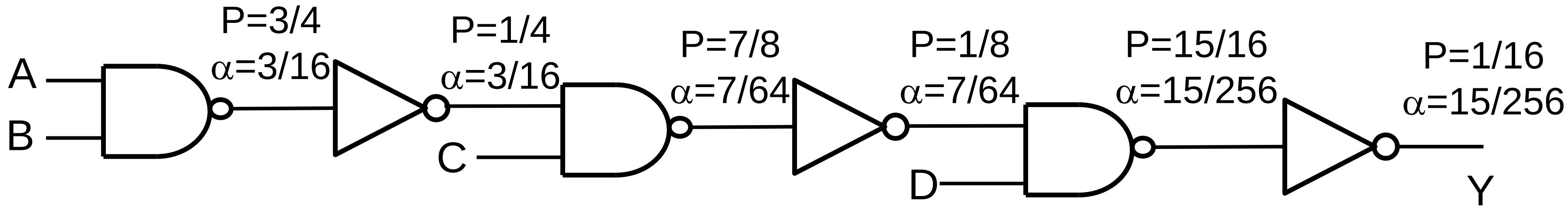
- Remember $\alpha_Y = \overline{P_Y} \bullet P_Y$

Example: 4-input AND gate

- Assume all inputs have P=0.5





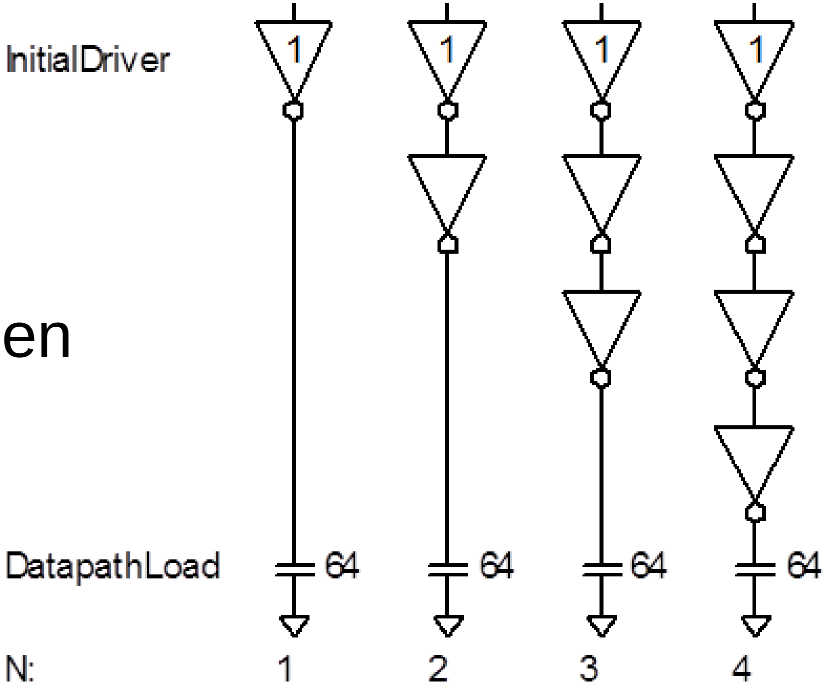


- Which has the lowest power?

Number of Stages vs. Power

- Power depends on activity and capacitance at each node
  - Generally fewer stages usually mean less power
- Compare this to delay
  - frequently add stages to improve delay (stage effort ≈ 4)

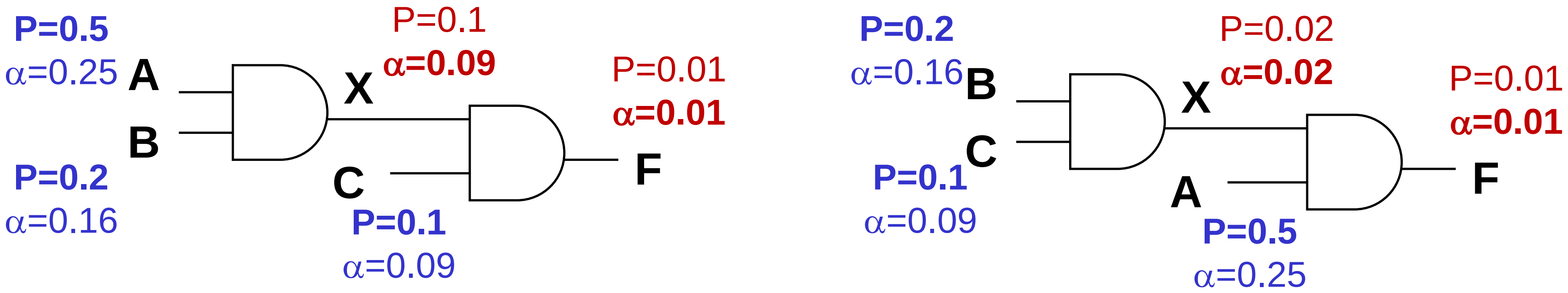
- Usually tradeoff between speed and power!



Relative Delay:	65	18	15	15.3
Relative Power:	66	82	106	134

Input Ordering

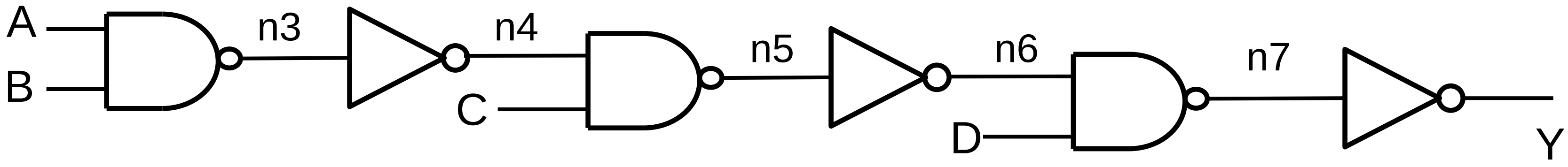
- What if inputs have different activity levels?

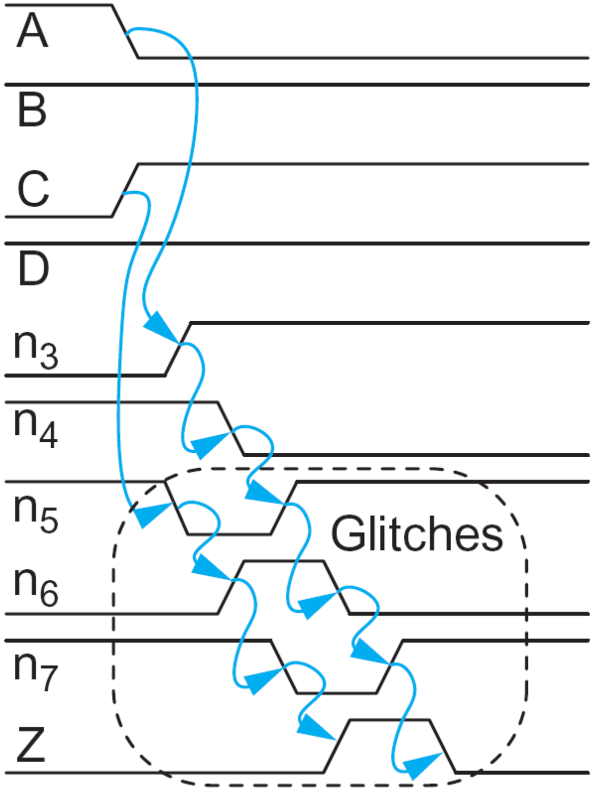


- Beneficial to postpone the introduction of signals with a high activity
  - i.e. signals with signal probability close to 0.5

Beware of Glitches!

- Extra transitions caused by finite propagation delay





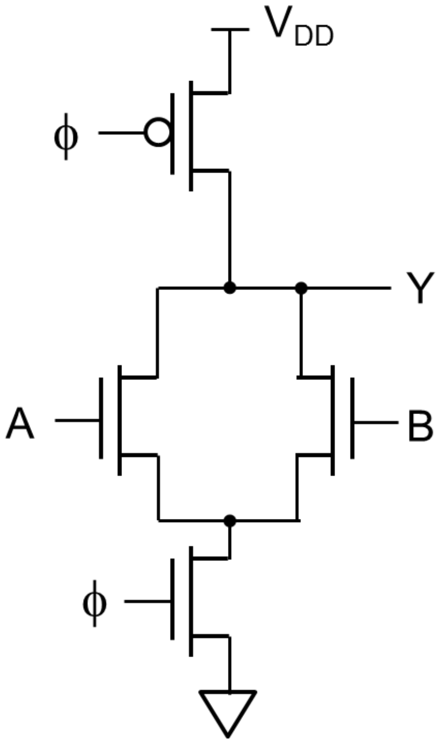
Suppose input changes from ABCD = “1101” to “0111” ?

Glitching occurs whenever a node makes more transitions than necessary to reach its final value

Glitching can raise the activity factor of a gate to greater than 1!

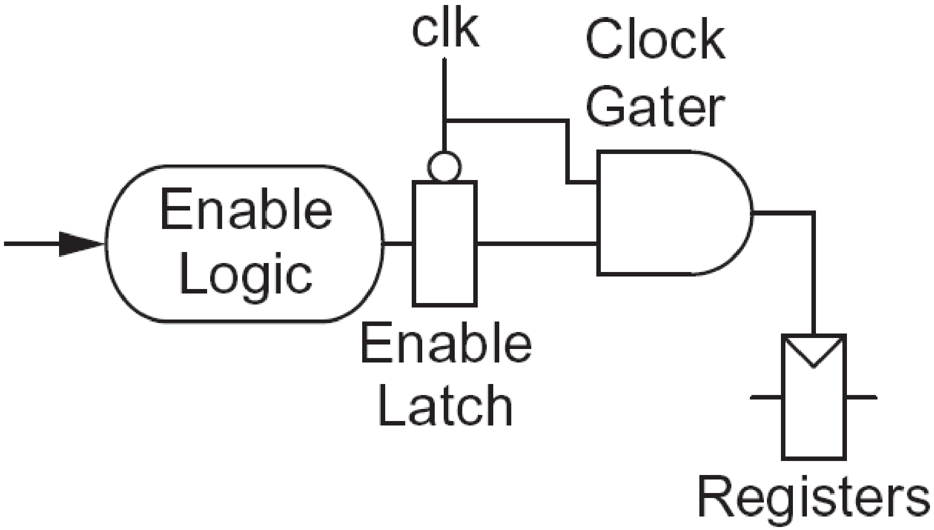
Activity of Dynamic Circuits

Exercise: What is activity factor of node Y (output of a dynamic NOR2) if P_A=P_B=0.5 ?



Clock Gating

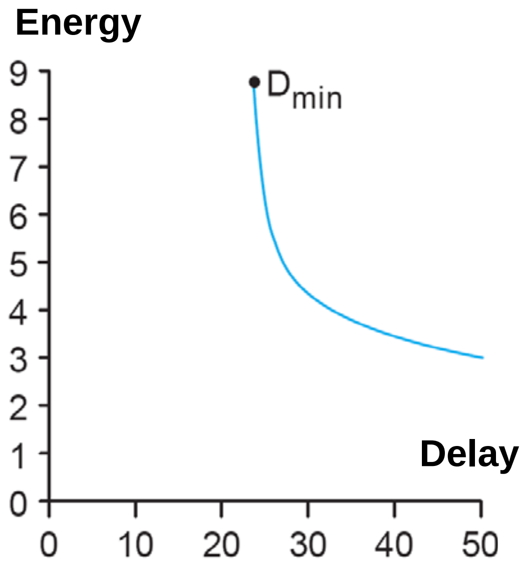
- Another way to reduce the activity is to turn off the clock to registers in unused blocks
  - – Saves clock activity ($\alpha = 1$)
  - – Eliminates all switching activity in the block
  - – Requires determining if block will be used



Capacitance

- Extra capacitance slows response and increases power
  - Always try to reduce parasitic and wiring capacitance
  - Good floorplanning to keep high activity communicating gates close to each other
  - Drive long wires with inverters or buffers rather than complex gates

- Gate sizing and number of stages
  - Designing network for minimum delay will usually result in a high-power network.
  - Small increase in delay (by reducing the # of stages or increasing the logical effort per stage) can give large reduction in power
  - There are no closed form solutions to determine gate sizes that minimize power under a delay constraint.
  - Can be solved numerically



Voltage

- Power dissipated in gate is $P_{av} = \alpha.f.C_L.V_{DD}^2$
- Energy per switching event* is $E_s = P_{av}/(2.\alpha.f) = (C_L.V_{DD}^2)/2$
  - Power & Energy can be significantly reduced by decreasing V_{DD}
- *But* delay of gate is $D = (C_L.\ \Delta V)/I$

  $$\approx (C_L.V_{DD})/[(\beta/2).(V_{DD}-V_t)^2]$$

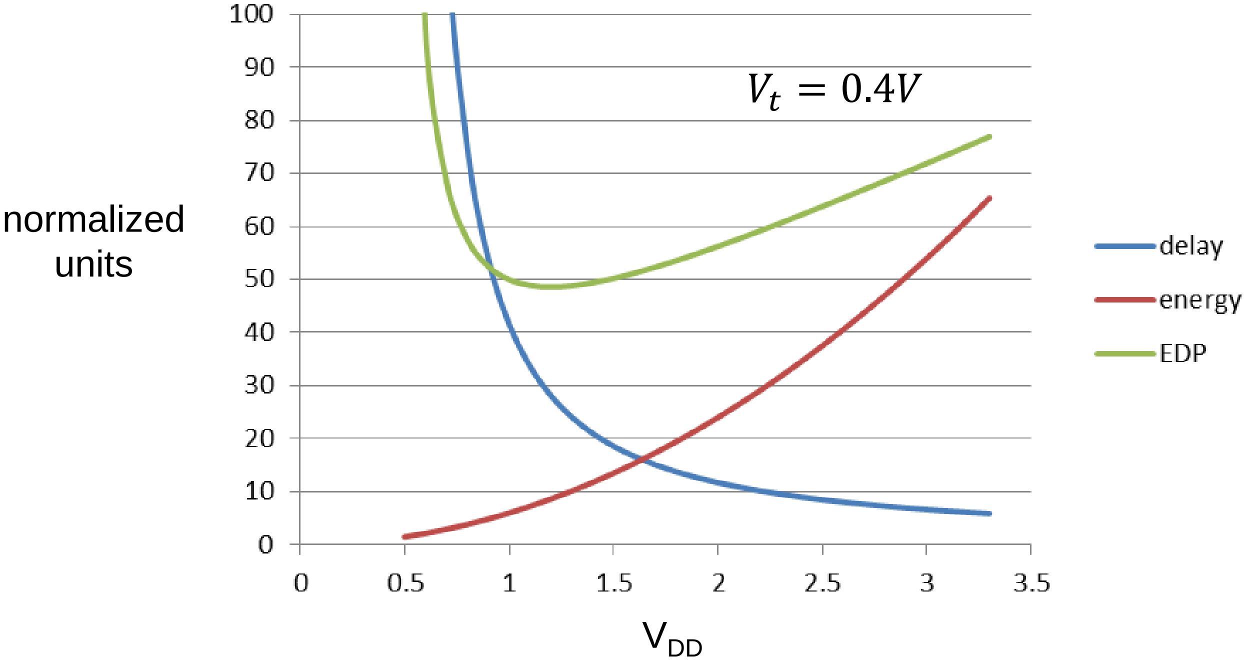
  - Decreasing V_{DD} increases delay
- Circuit can be made (almost) arbitrarily low power at the expense of performance – not very useful

** switching event is defined as a transition from 0→1 or 1→0*

Energy-Delay Product

- Introduce metric energy-delay product (EDP)

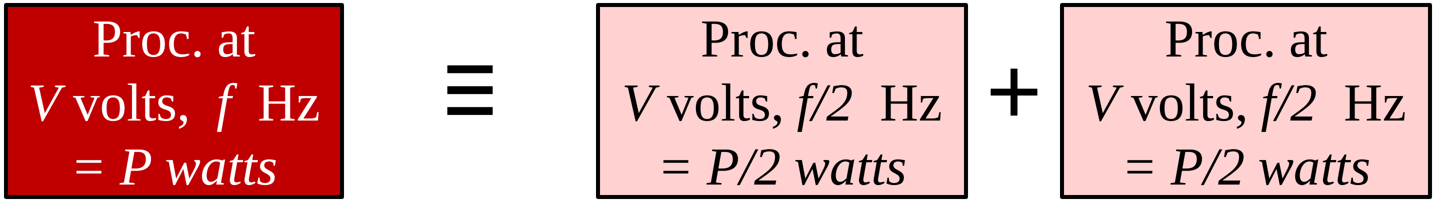
$$EDP = E_s . D = \frac{k . {C_L}^2 . {V_{DD}}^3}{(V_{DD} - V_t)^2}$$



- Minimum EDP at $V_{DD} = 3.V_t$ (for long channel process)

Frequency

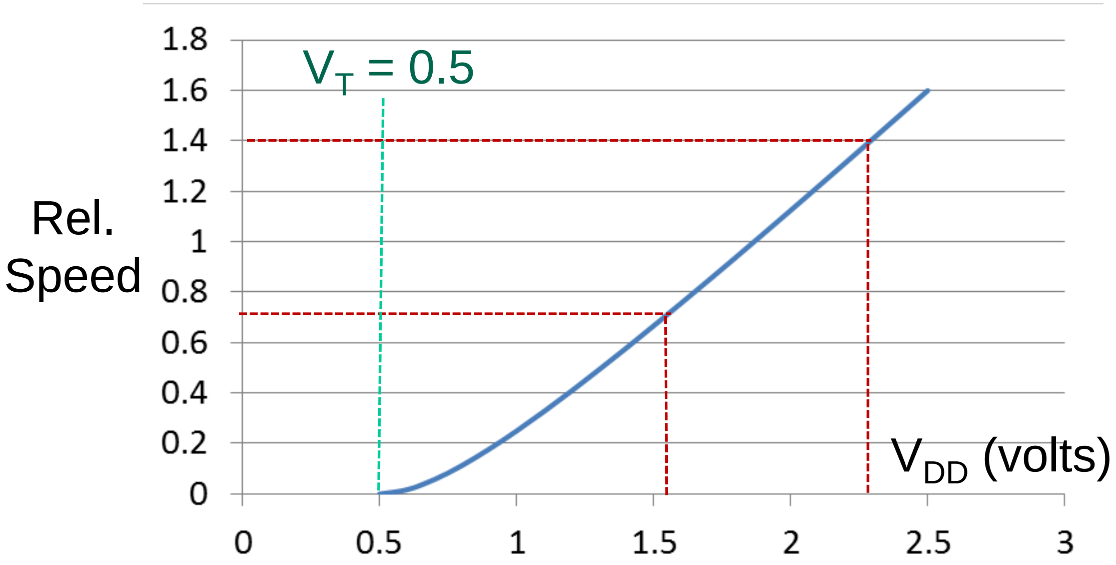
- Suppose we can do a task in T sec. on one processor
- Can we do it in T/2 sec. on two processors
  - if application has sufficient intrinsic parallelism
- How about doing it in T sec. on two processors running at half clock frequency?



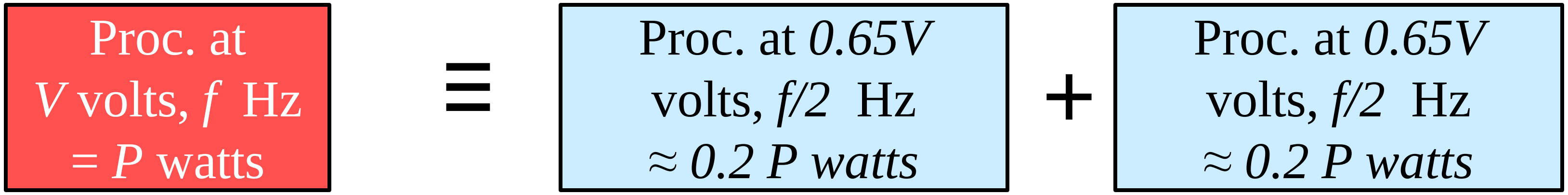
Proc. at V volts, f Hz $= P$ *watts* ≡ Proc. at V volts, $f/2$ Hz $= P/2$ *watts* + Proc. at V volts, $f/2$ Hz $= P/2$ *watts*

- This gives no net power savings.
- But $speed \propto (V_{DD} - V_T)^2 / V_{DD}$, so if we reduce clock frequency, we can also reduce V_{DD}:

Reduced Frequency & Voltage



In this example, reducing speed by factor of 50% allows voltage reduction of ~35%



Proc. at V volts, f Hz $= P$ watts

$\equiv$

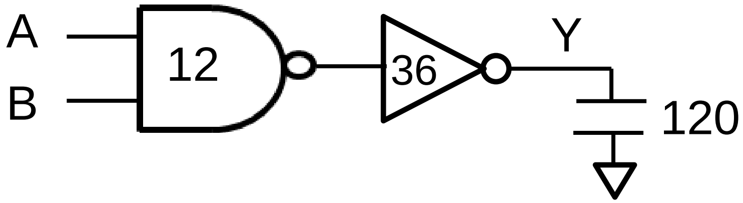
Proc. at $0.65V$ volts, $f/2$ Hz $\approx 0.2\,P$ *watts*

$+$

Proc. at $0.65V$ volts, $f/2$ Hz $\approx 0.2\,P$ *watts*

- Parallelism with reduced f and V_{DD} leads to lower power
  - diminishing returns as V_{DD} approaches V_T

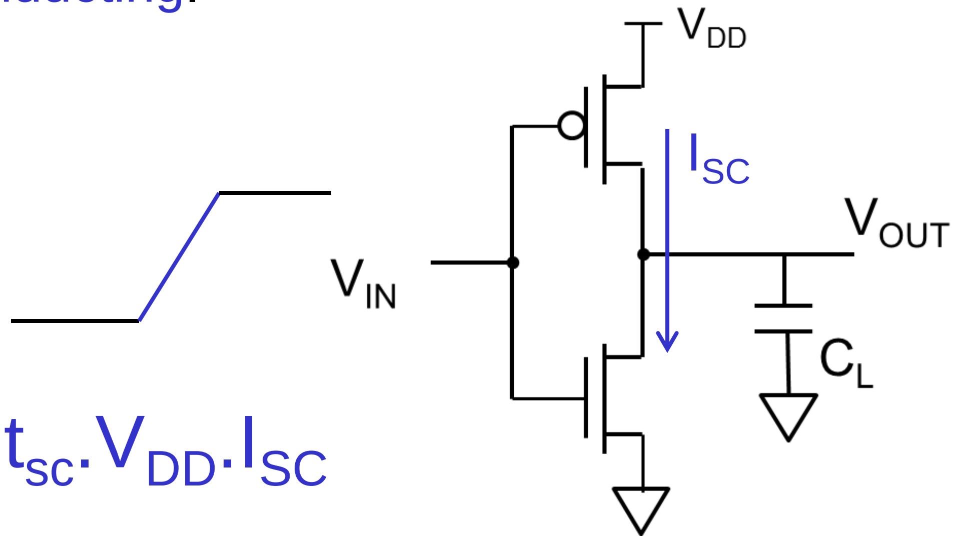
Dynamic Power Dissipation Example



- A NAND2 gate of size (input capacitance) 12C is driving an inverter of size 36C which in turn drives a load of 120C units of capacitance. Assume the inputs A, B are independent and uniformly distributed. What is the power dissipation of this circuit if the gate capacitance C of a unit sized transistor is 0.1fF, V_{DD} is 1.0V and the operating frequency is 1GHz?

Short-Circuit Power

- Finite slope of the input signal
  - sets up a direct current path between V_{DD} and GND for a short period during switching when both the NMOS and PMOS devices are conducting.

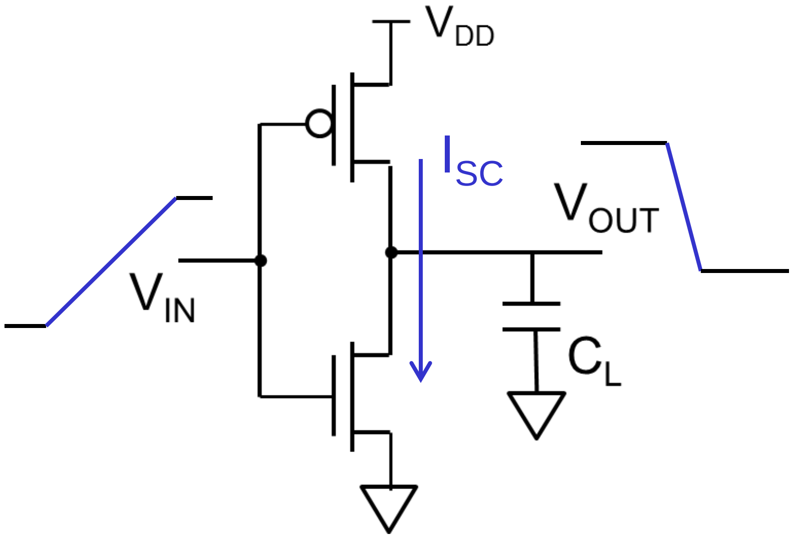


$$E_{sc} \approx t_{sc}.V_{DD}.I_{SC}$$

- Depends on duration and slope of the input signal, t_{sc}
- I_{SC} which is determined by
  - saturation current of the P and N transistors
    - depends on sizes, process technology, temperature, etc.
  - ratio between input and output slopes (a function of C_L)

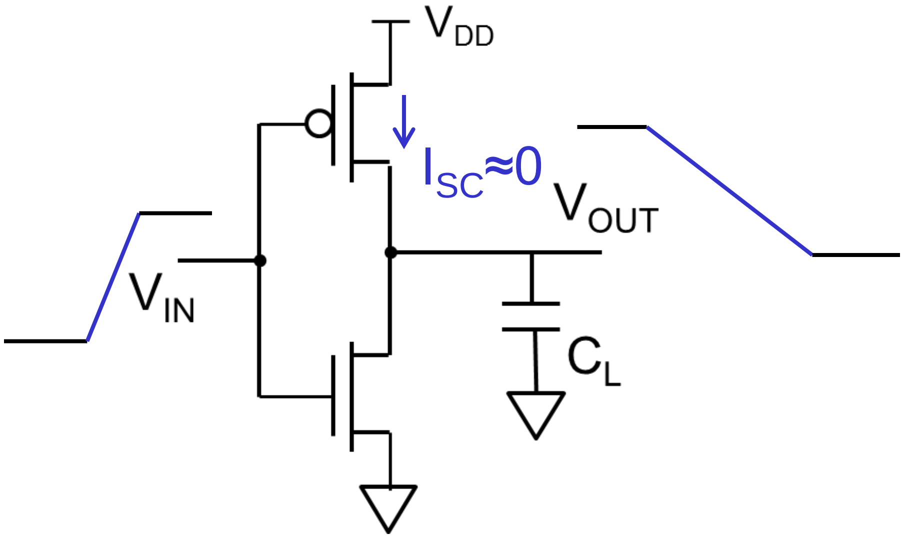
Slope Engineering

Small Capacitive Load



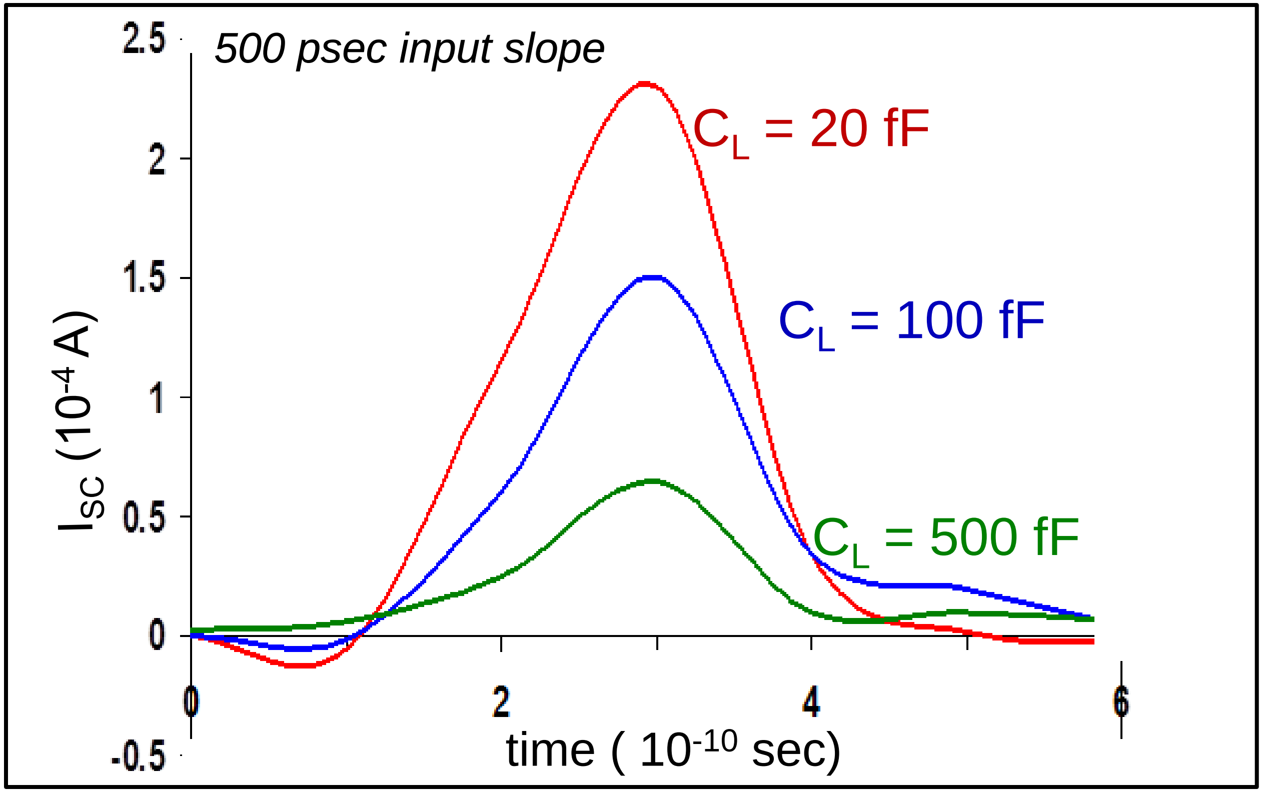
- Output fall time significantly shorter than input rise time
- Output "tracks" input as per DC transfer function
- Large I_{SC} when $V_{IN} = V_{SW} \approx V_{DD}/2$

Large Capacitive Load



- Output fall time significantly longer than input rise time
- Output transition lags input
- When $V_{IN} = V_{SW}$, V_{dsp} is still very small, so small I_{SC}

Impact of C_L on I_{SC}

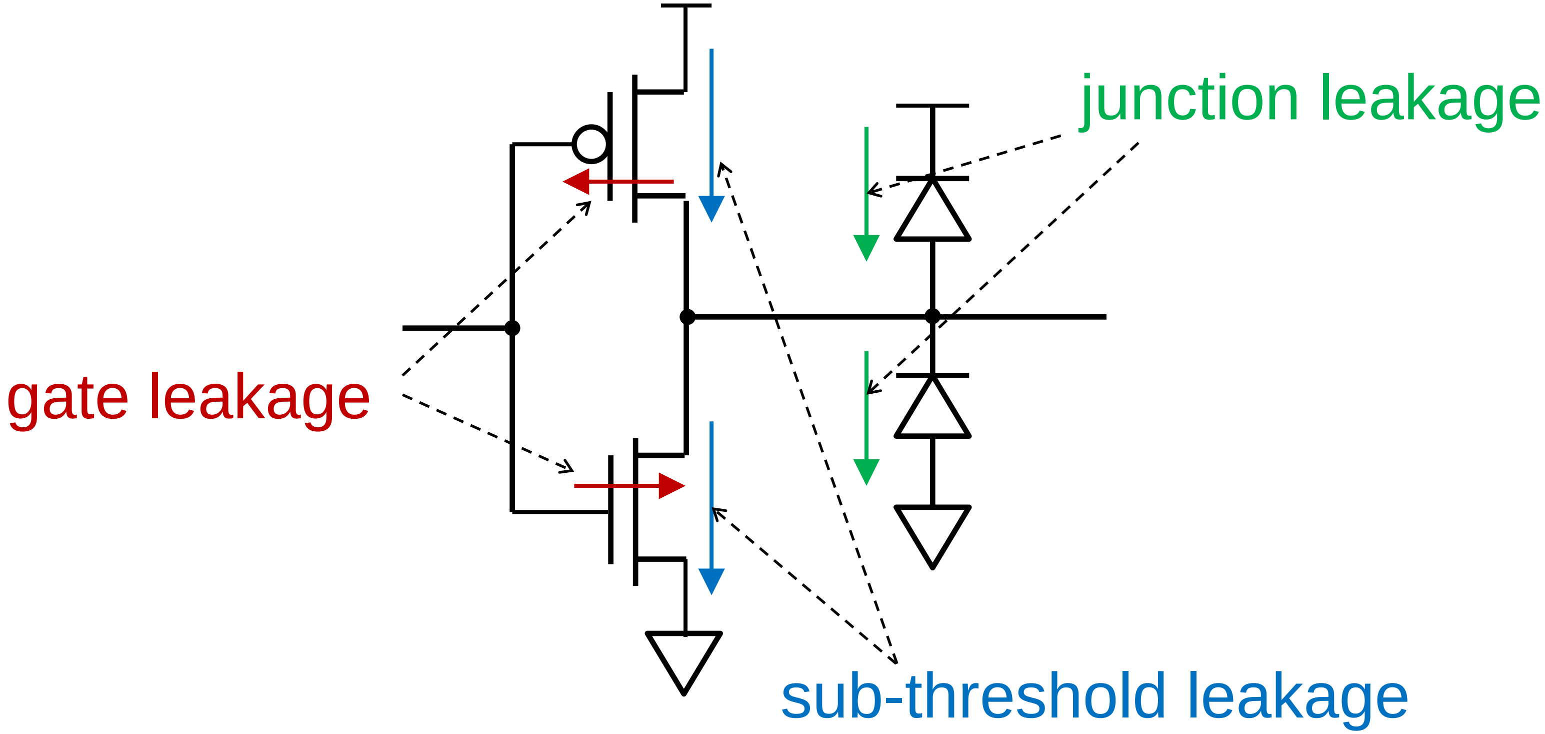


- When C_L is small, I_{SC} is large!
  - Short circuit dissipation is minimized by matching the rise/fall times of the input and output signals - slope engineering.
- Typically less than 10% of dynamic power if rise/fall times are comparable for input and output

Static Power Dissipation

- Static power is consumed even when chip is quiescent
  - i.e. powered up but not running
- Ratio’ed circuits (e.g. pseudo-NMOS) burn power in fight between ON transistors
  - known as contention current
- Leakage consumes power from current passing through normally off devices
  - sub-threshold current
  - gate leakage current
  - diode junction leakage current

Leakage Sources



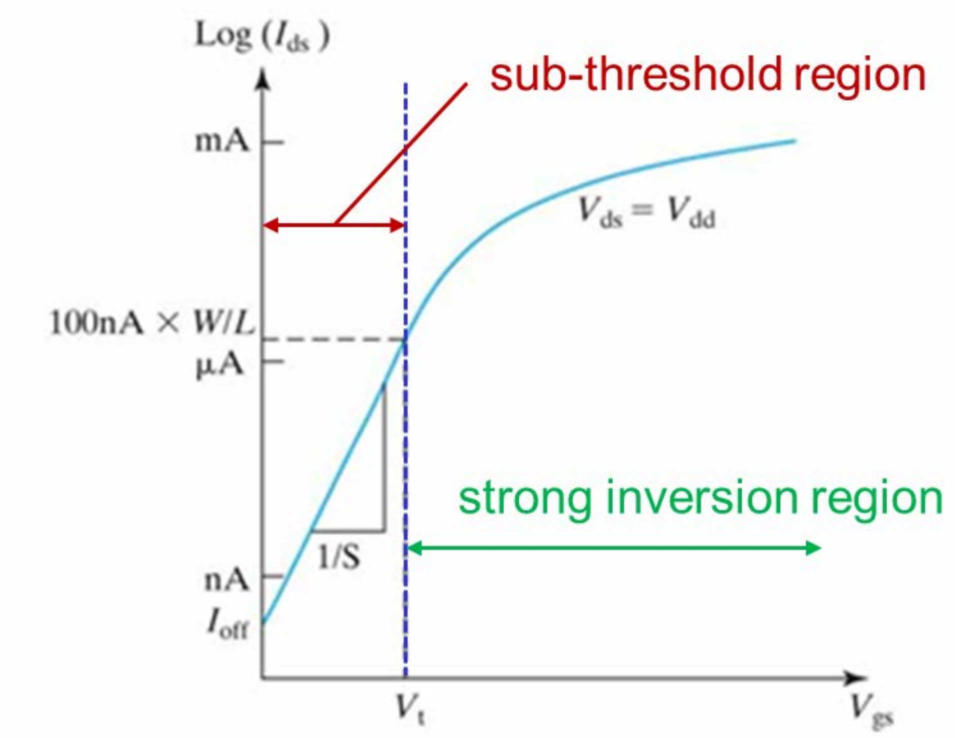
- Leakage currents are very small (per transistor basis)
  - prior to 130 nm, not usually an issue (except in sleep mode of battery operated devices)
  - but when multiplied by hundreds of millions of nanometer devices, can account for as much as 1/3 of active power
- All increase exponentially with temperature

Sub-threshold Leakage

- Shockley model assumes $I_d = 0$ when $V_{gs} \leq V_t$
- But in real transistors, $I_d \approx\ 100nA \times (W/L)$ when $V_{gs} = V_t$
- For $V_{gs} < V_t$, I_d decreases exponentially with V_{gs}

$$I_d = I_0 10^{\frac{(V_{gs}-V_t)}{S}}$$

where S is sub-threshold slope ≈ 100mV/decade



Sub-threshold Leakage

- In sub-threshold: $I_d = I_0 10^{\frac{(V_{gs} - V_t)}{S}}$

- In nanometer processes, as we reduce V_{DD}, we also reduce V_t to maintain good *on-current*

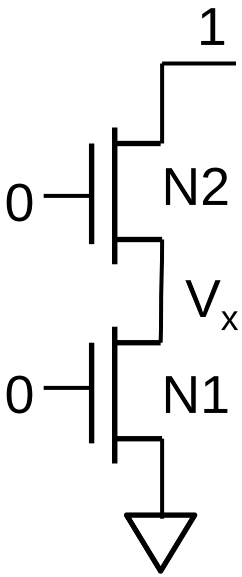
- But *off-current* $I_{off} = I_0 10^{\frac{(-V_t)}{S}}$ increases with smaller V_t

Typical values in 65 nm, V_{ds}=1.0V:

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

Stack Effect

- Series OFF transistors have less leakage
  - for N1 to have any leakage, $V_x > 0$
  - so N2 has negative V_{gs}
  - leakage through 2-stack reduces ~10x
  - leakage through 3-stack reduces further



- Leakage and delay trade off
  - Aim for low leakage in sleep and low delay in active mode

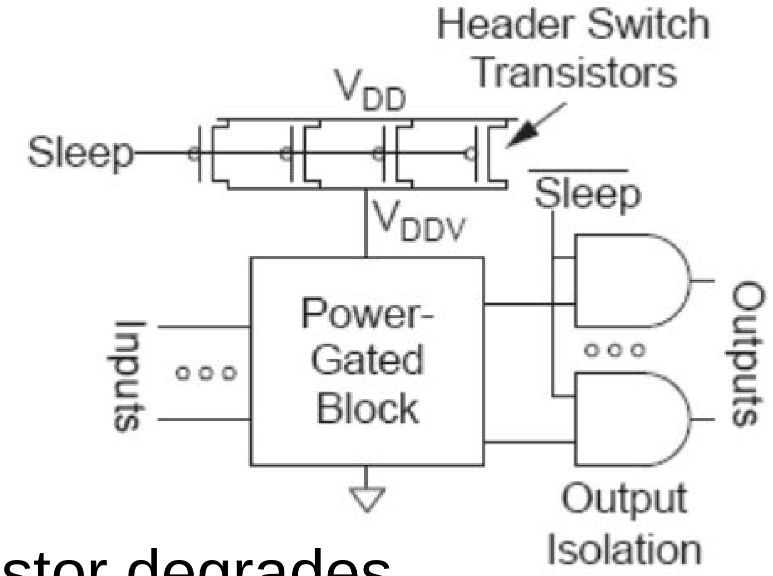
- To reduce leakage:
  - Increase V_t: *multiple V_t*
    - Use low V_t only in speed critical circuits
  - Increase V_s: *stack effect*
    - *Input vector control* in sleep

Gate & Junction Leakage

- Gate leakage extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches sub-threshold leakage at 65 nm
- An order of magnitude less for pMOS than nMOS
- Control gate leakage in the process using $t_{ox} > 10$ Å
  - High-k gate dielectrics help
  - Some processes provide multiple t_{ox}
    - e.g. thicker oxide for 3.3 V I/O transistors
- Junction leakage usually negligible
  - becoming little more significant in nanometer processes
- Control gate & junction leakage in circuits by limiting V_{DD}

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 impact
- Switching wide sleep transistor costs dynamic power
    - Only justified when circuit sleeps long enough

Static Power Example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - 65nm process, V_{DD}=1.0V, λ = 25nm
  - 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
  - Junction leakage negligible

Static Power Solution

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

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

$$I_{sub} = \left[W_{\text{normal-V}_\text{t}} \times 100 \text{ nA/}\mu\text{m} + W_{\text{high-V}_\text{t}} \times 10 \text{ nA/}\mu\text{m}\right] / 2 = 584 \text{ mA}$$

$$I_{gate} = \left[\left(W_{\text{normal-V}_\text{t}} + W_{\text{high-V}_\text{t}}\right)\times 5 \text{ nA/}\mu\text{m}\right] / 2 = 275 \text{ mA}$$

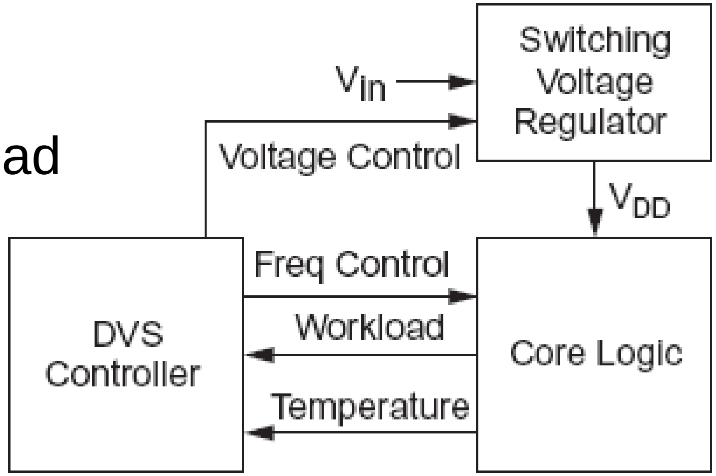
$$\text{P}_{static} = \left(584 \text{ mA} + 275 \text{ mA}\right)\left(1.0 \text{ V}\right) = 859 \text{ mW}$$

Voltage & Frequency Control

- Run each block at the lowest possible voltage and frequency that meets performance requirements

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

- Dynamic Voltage Scaling
  - Adjust V_{DD} and f according to workload


Switching Voltage Regulator
V_{in}
Voltage Control
V_{DD}
Freq Control
Workload
Temperature
DVS Controller
Core Logic