ASIC/FPGA Chip Design

Power Dissipation

Mahdi Shabany

Department of Electrical Engineering

Sharif University of technology

Outline

- Introduction
- Dynamic Power Dissipation
- Static Power Dissipation

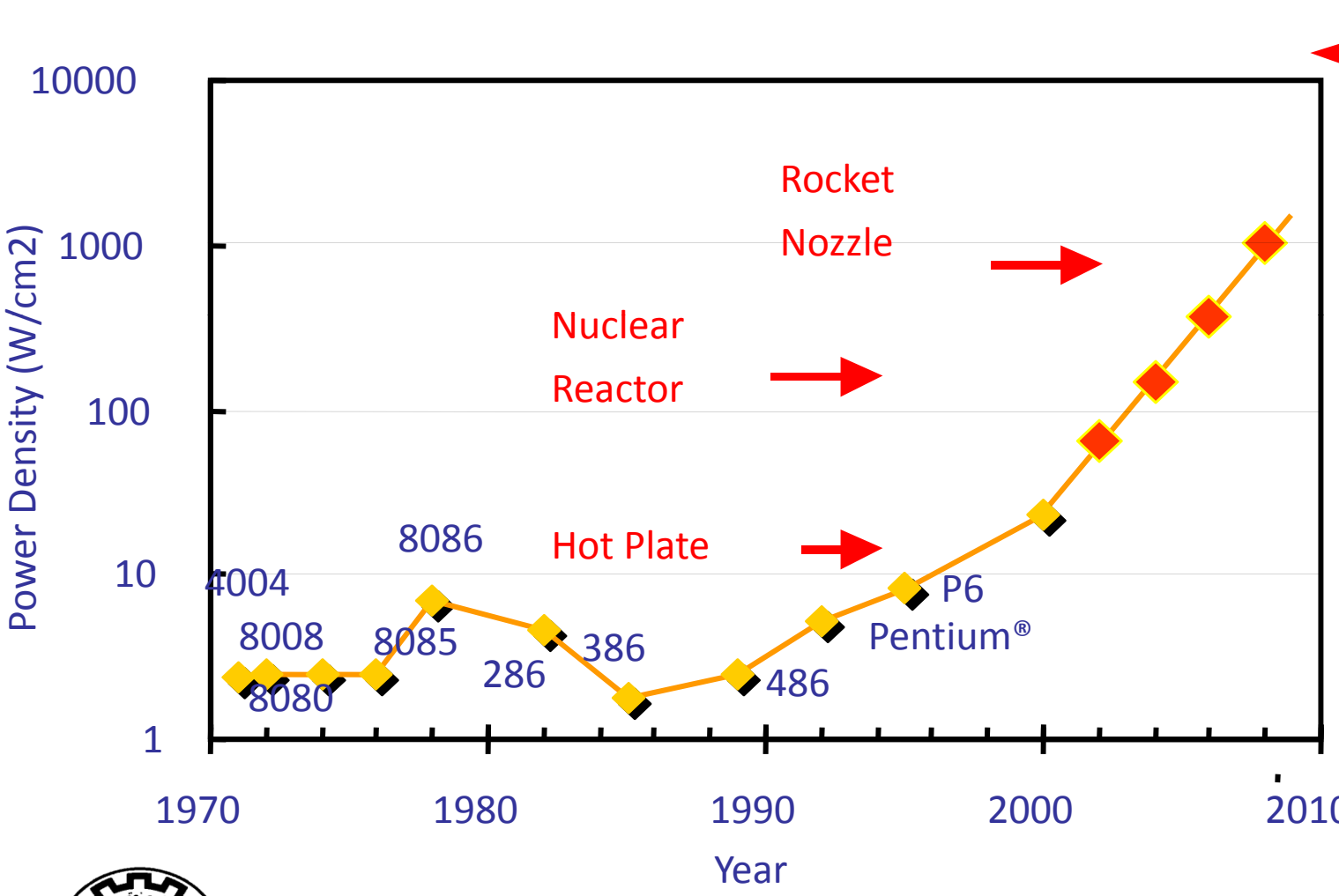
Outline

- Introduction
- Dynamic Power Dissipation
- Static Power Dissipation

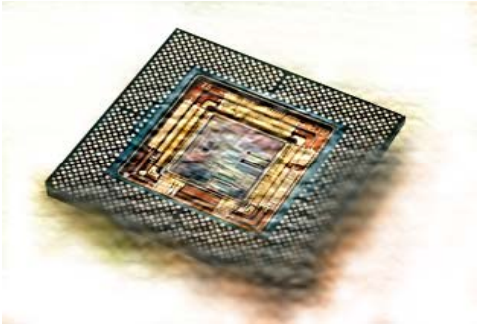
Why Power Matters?

- Packaging costs
- Power supply rail design
- Chip and system cooling costs
- Noise immunity and system reliability
- Battery life (in portable systems)
- Environmental concerns

Why worry about power? Chip Power Density



…chips might become hot…

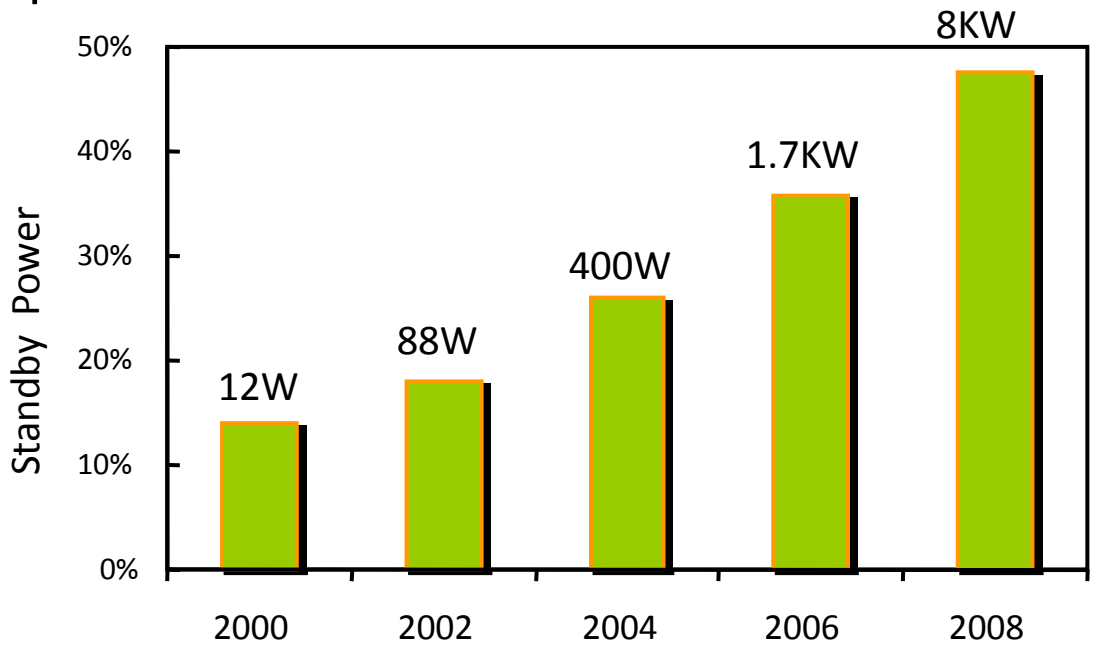


Source: Borkar, De Intel®

Why worry about power? Standby Power

Year	2002	2005	2008	2011	2014
Power supply V_{dd} (V)	1.5	1.2	0.9	0.7	0.6
Threshold V_T (V)	0.4	0.4	0.35	0.3	0.25

- Drain leakage will increase as V_T decreases to maintain noise margins and meet frequency demands, leading to excessive battery draining standby power consumption.

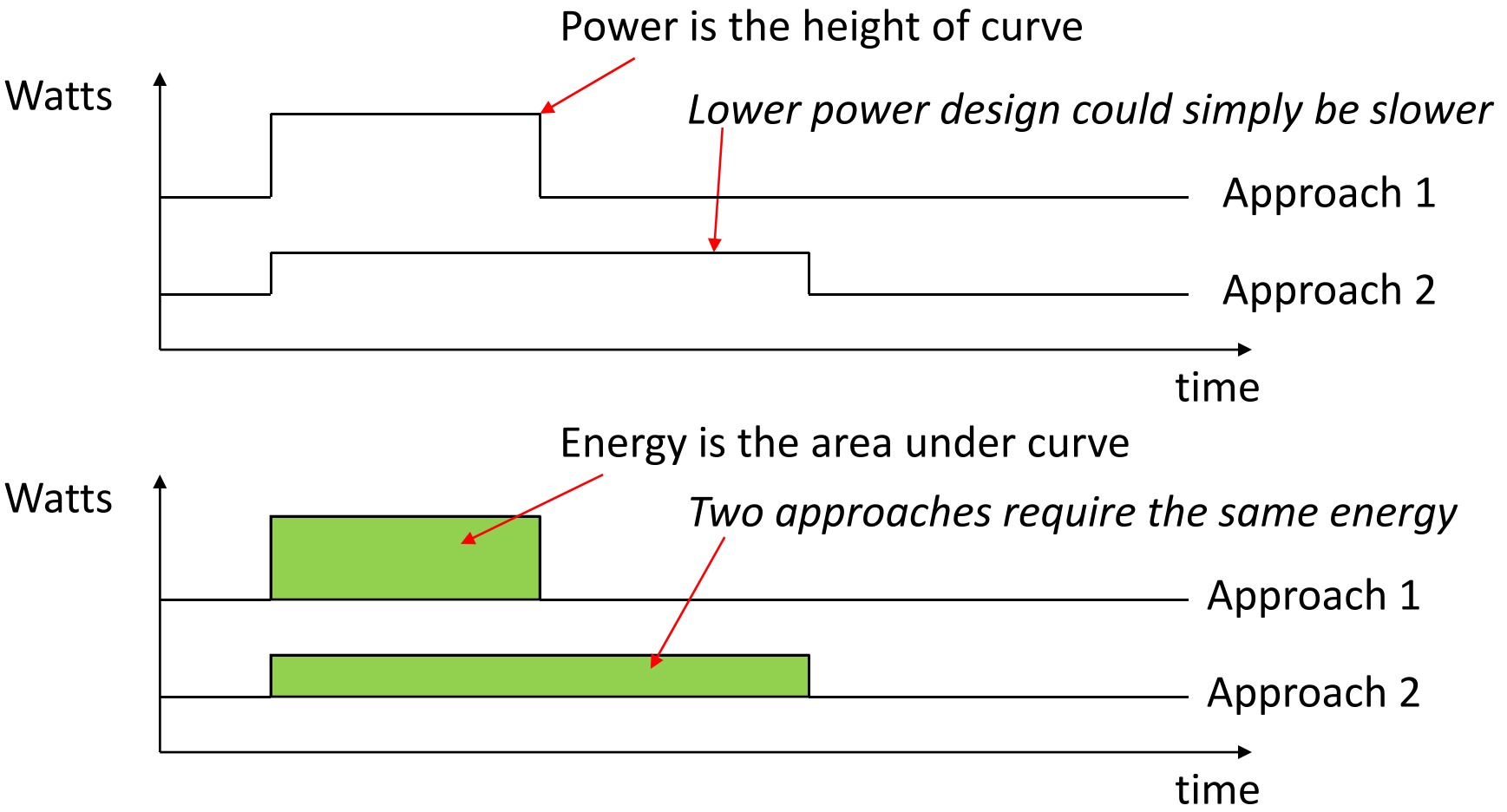


Power and Energy Figures of Merit

- Power consumption in Watts
  - Determines battery life in hours
- Peak power
  - Determines power ground wiring designs
  - Sets packaging limits
  - Impacts signal noise margin and reliability analysis
- Energy efficiency in Joules
  - Rate at which power is consumed over time
- Energy = power * delay
  - Joules = Watts * seconds
  - Lower energy number means less power to perform a computation at the same frequency

Power is the rate at which energy is delivered or exchanged; Power dissipation is the rate at which energy is taken from the source and converted into heat

Power vs. Energy



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

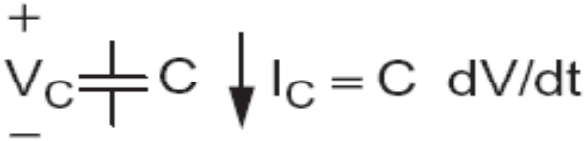
Power in Circuit Elements

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

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

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

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



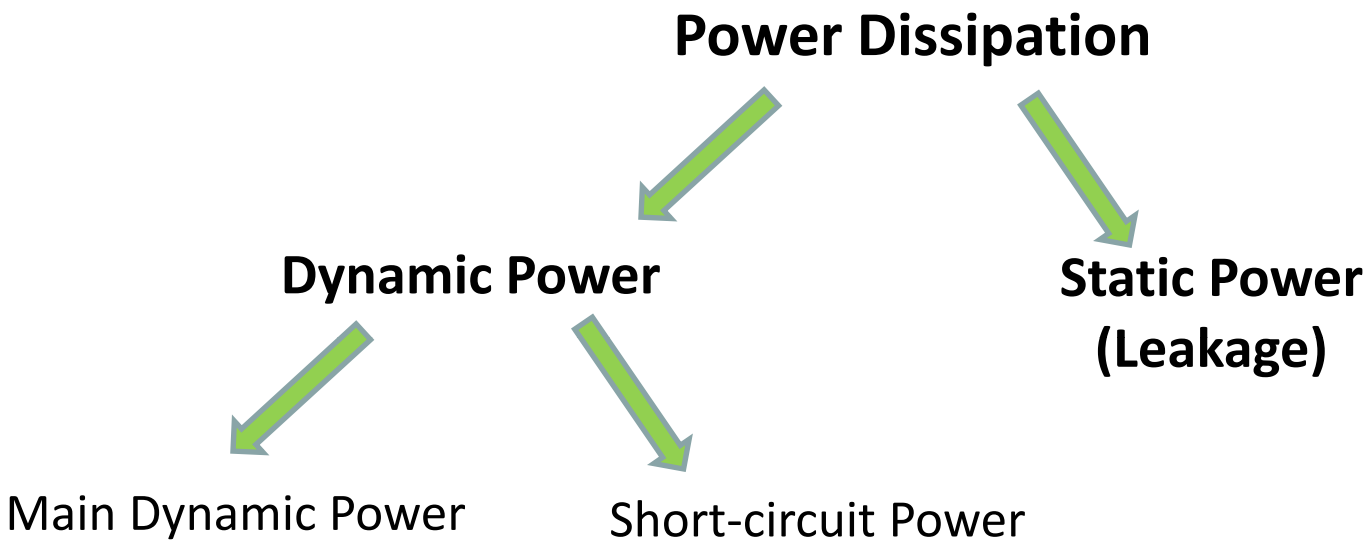
Power Dissipation

- **Power:** Due to the current flowing from supply to ground

$$P = I_D V_{DD}$$

- **Power Dissipation:**
  - **Dynamic Power :** Occurs only when the gate switches
    - Charging/discharging of load capacitances
    - Short-circuit power during switching (when both NMOS and PMOS are ON)
  - **Static Power**: Due to the presence of a path in the gate b/w the power supply & GND
    - In CMOS, when circuit is quiescent (no switching) one of the transistors is OFF thus ideally no current flows through an OFF transistor so no current b/w VDD and GND thus zero static power

Power Dissipation



CMOS Total Energy & Power Equations

$$E = C_L V_{DD}^2 \alpha_{0\to1} + \alpha_{sc} V_{DD}^2 C_L + V_{DD} I_{static}$$

$$f_{0\to1} = \alpha_{0\to1} * f_{Clk}$$

$$P = C_L V_{DD}^2 f_{0\to1} + \alpha_{sc} V_{DD}^2 C_L f_{Clk} + V_{DD} I_{static}$$

Dynamic Power (~90% today and decreasing relatively)

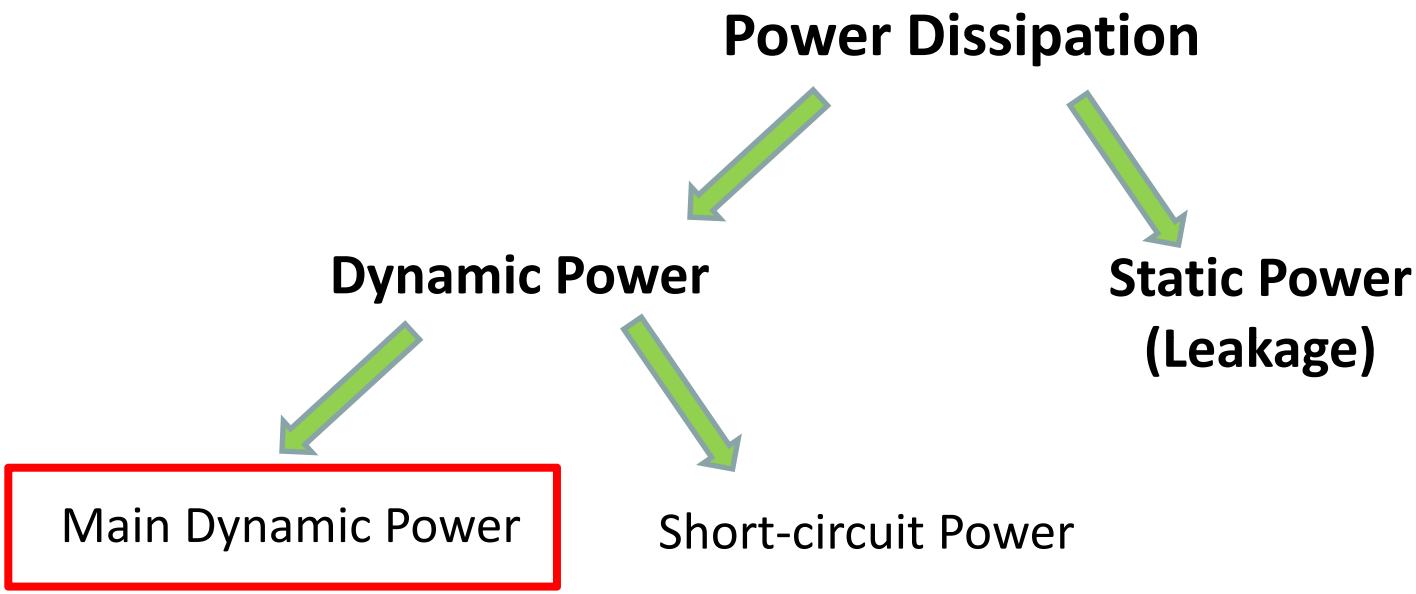
Short-circuit Power (~8% today and decreasing absolutely)

Static Power (~2% today and increasing)

Outline

- Introduction
- Dynamic Power Dissipation
- Static Power Dissipation

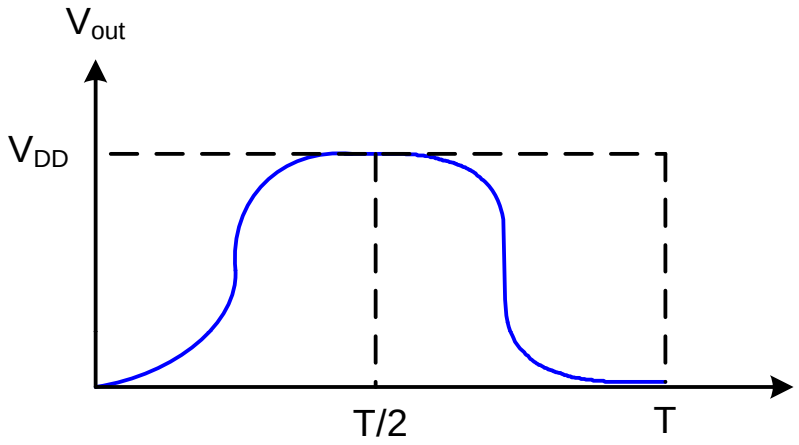
Power Dissipation



Power Dissipation: Main Dynamic Power

- Due to the charging/discharging the load capacitances

$$P_{avg} = \frac{1}{T}\int_0^T v(t)i(t)dt$$



$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2}(V_{out})C_L\frac{dV_{out}}{dt}dt + \int_{T/2}^{T}V_{out}(-C_L\frac{dV_{out}}{dt})dt\right]$$

$$\Rightarrow P_{avg} = \frac{1}{T}\left[\left(\frac{V_{out}^2}{2}C_L\right)_0^{T/2} - \left(\frac{V_{out}^2}{2}C_L\right)_{T/2}^{T}\right] = \frac{1}{T}V_{DD}^2C_L = C_LV_{DD}^2f_{Clk}$$

Power Dissipation: Dynamic Power

- Average Dynamic Power:
  - Linearly dependent to f_{Clk} (Clock frequency)
  - Independent of the transistor sizing

- Considering the utilization factor:

$$P_{avg} = \alpha C_L V_{DD}^2 f_{Clk} \qquad \alpha : \text{Activity Factor}$$

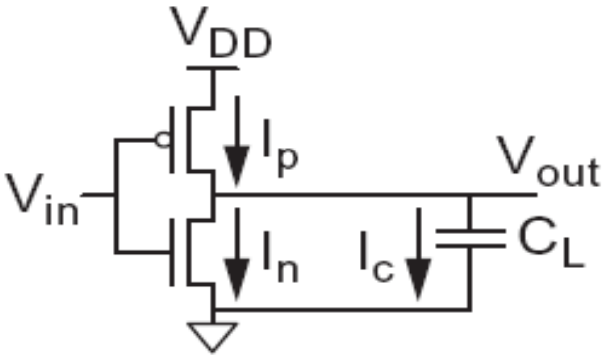
- In general, a chip with higher area burns more power unless its utilization factor is lower

- Power Delay Product: (dissipated as heat in transistors)

$$PDP = C_L V_{DD}^2$$

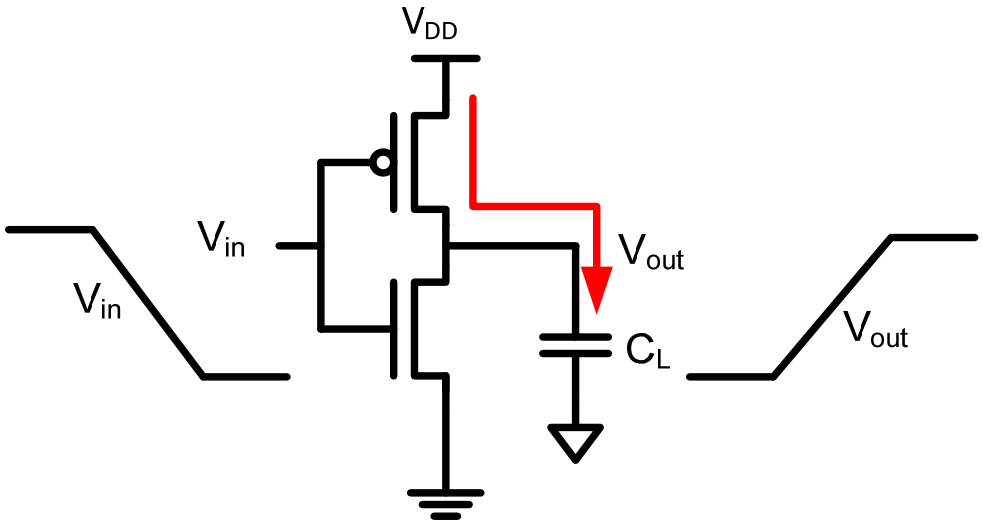
Charging a Capacitor

- 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 the capacitor



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

Dynamic Power Consumption



$$\text{Energy/transition} = C_L * V_{DD}^2 * \alpha_{0\to1}$$

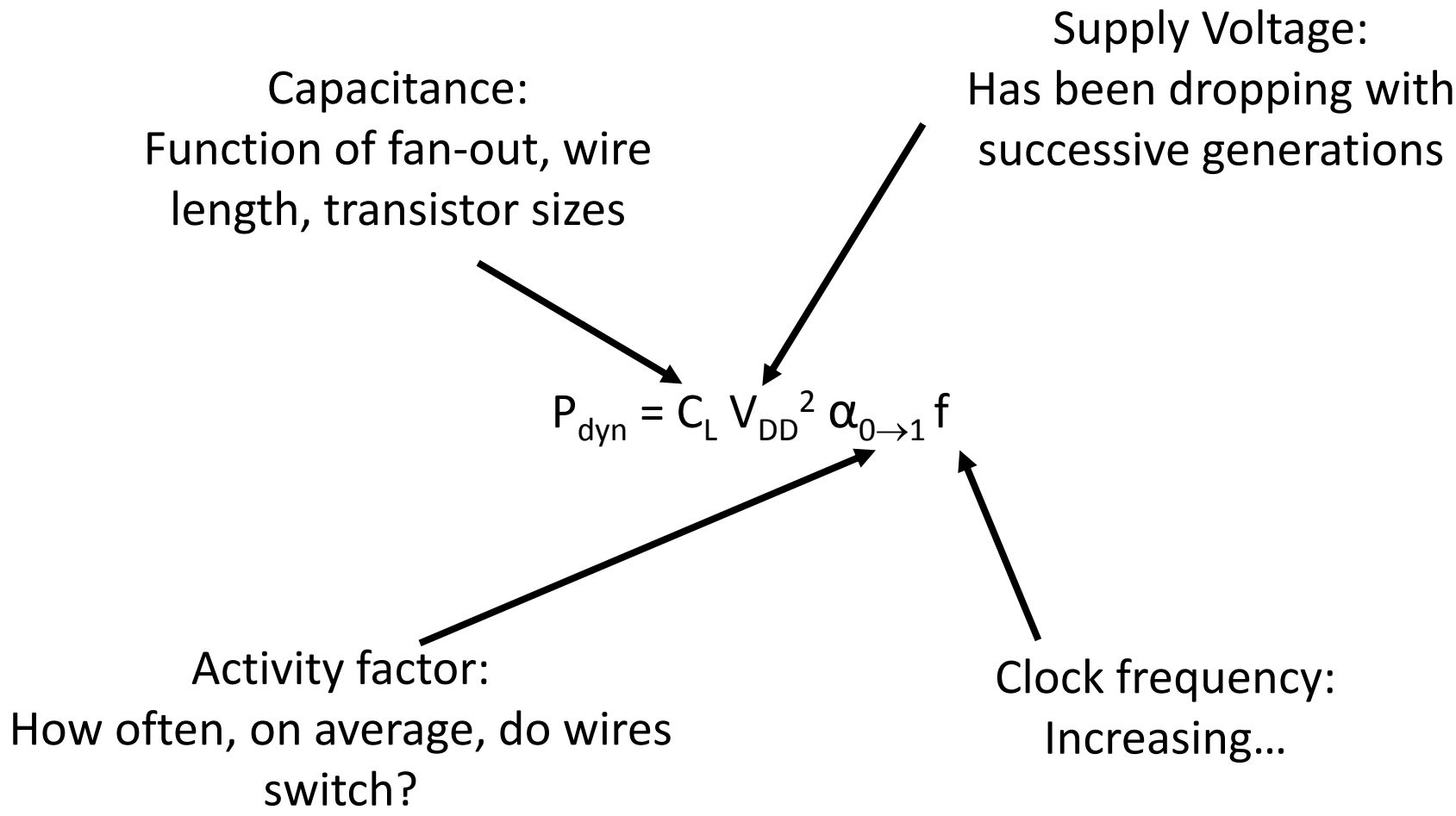
$$P_{dyn} = \text{Energy/transition} * f = C_L * V_{DD}^2 * \alpha_{0\to1} * f$$

$f_{0\to1}$

$$P_{dyn} = C_{eff} * V_{DD}^2 * f \quad \text{where } C_{eff} = \alpha_{0\to1} C_L$$

- Not a function of transistor sizes!
- Data dependent - a function of switching activity!

Lowering Dynamic Power



Lowering Dynamic Power

- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

$$P_{dyn} = C_L V_{DD}^2 \alpha_{0\rightarrow1} f$$

Lowering Dynamic Power – Activity Factor

❑ Probability that output is "zero" in one cycle and will be "one" in the next cycle

$$\alpha_{0\to1} = P_0 P_1 = \frac{N_0}{2^N}\frac{N_1}{2^N} = \frac{N_0(2^N - N_0)}{2^{2N}}$$

➢ where

N_0: Number of zero entries in the output column of the function truth table

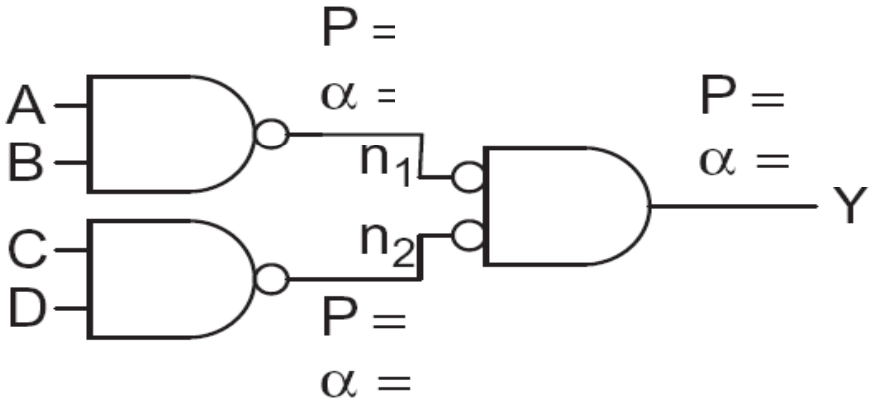
N_1: Number of one entries in the output column of the function truth table

❖ **Example**: A 2-input NOR

$$\alpha_{0\to1} = \frac{3(4-3)}{2^4} = \frac{3}{16} \Rightarrow P_{avg} = \frac{3}{16} C_L V_{DD}^2 f_{Clk}$$

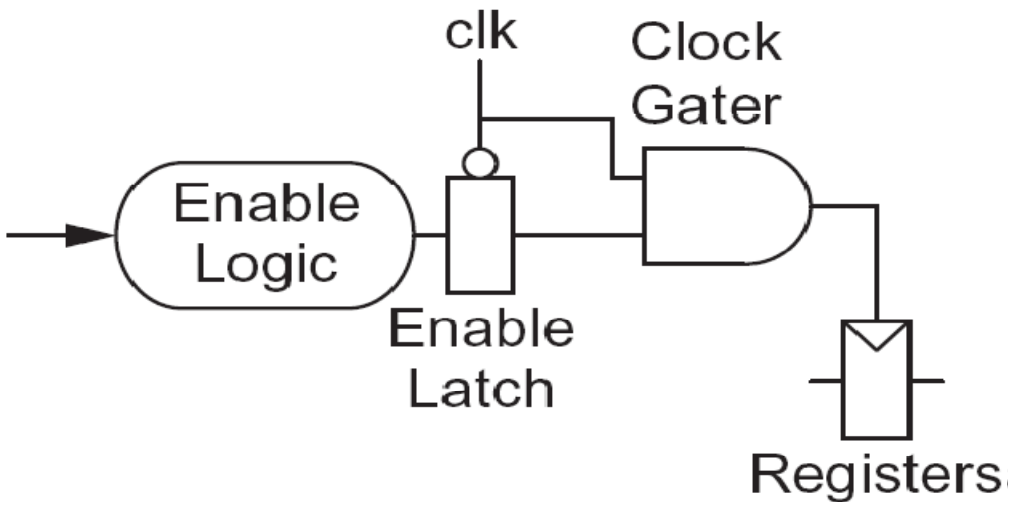
Example

- 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



Lowering Dynamic Power: 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

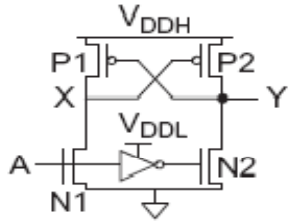


Lowering Dynamic Power: Capacitance

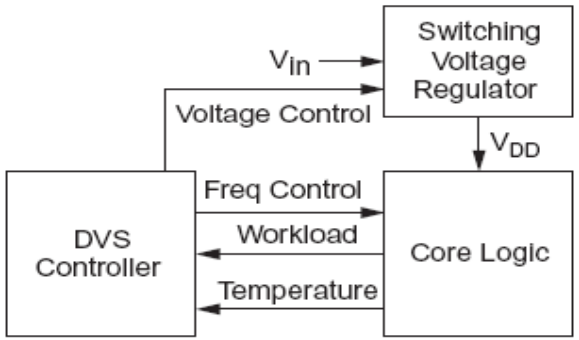
- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Lowering Dynamic Power: Voltage / Frequency

- 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



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



Power Dissipation: Dynamic Power

- In a digital CMOS circuit:

$$t_p \cdot I_{sat} = C\Delta V$$

$$t_p \cdot K(V_{DD} - V_t)^2 = C\Delta V \Rightarrow t_p \propto \frac{CV_{DD}}{(V_{DD} - V_t)^2} \Rightarrow f_{max} \propto \frac{(V_{DD} - V_t)^2}{V_{DD}} \propto V_{DD}$$

- Therefore, it can be shown that

$$P_{avg} \propto CV_{DD}^2 f \propto V_{DD}^3$$

$$\Rightarrow V_{DD} \uparrow \Rightarrow \text{delay} \downarrow \Rightarrow \text{Power} \uparrow$$

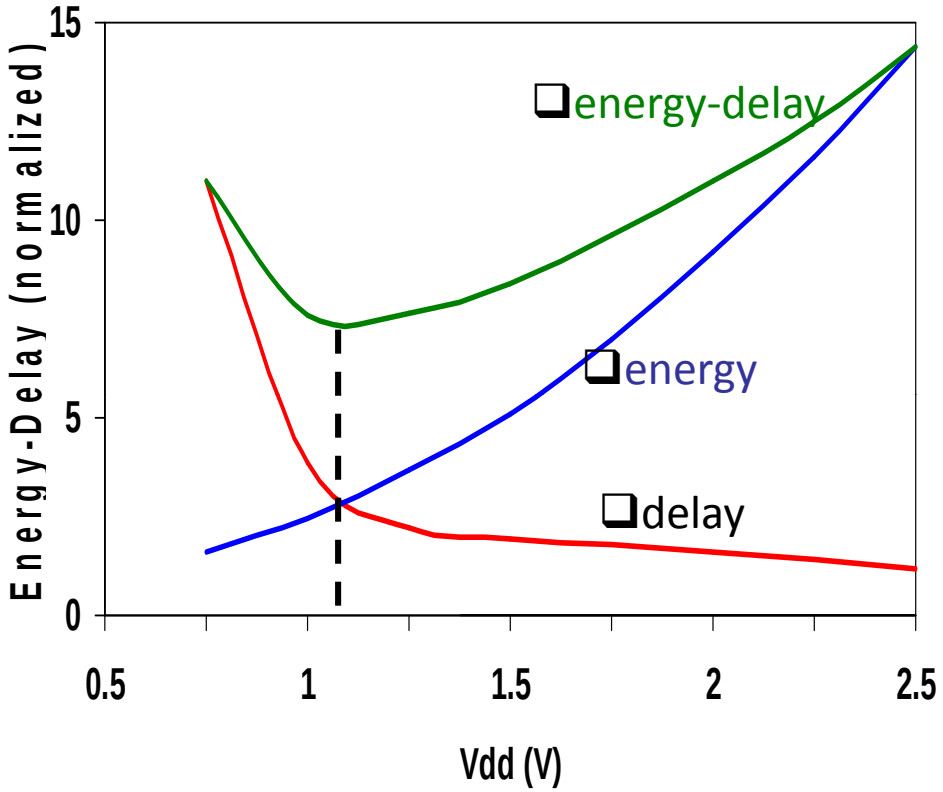
Throughput can be compromised for power

PDP and EDP

- **Power-delay product (PDP)** = $P_{av} * t_p = C_L V_{DD}^2$
  - PDP is the average energy consumed per switching event (Watts * sec = Joule)
  - Lower power design could simply be a slower design
  - For a given structure the PDP may be made arbitrarily low by reducing the supply voltage that comes at the expense of performance.

- **Energy-delay product (EDP)** = PDP * $t_p = P_{av} * t_p^2$
  - EDP is the average energy consumed multiplied by the computation time required
  - Takes into account that one can trade increased delay for lower energy/operation (e.g., via supply voltage scaling that increases delay, but decreases energy consumption)

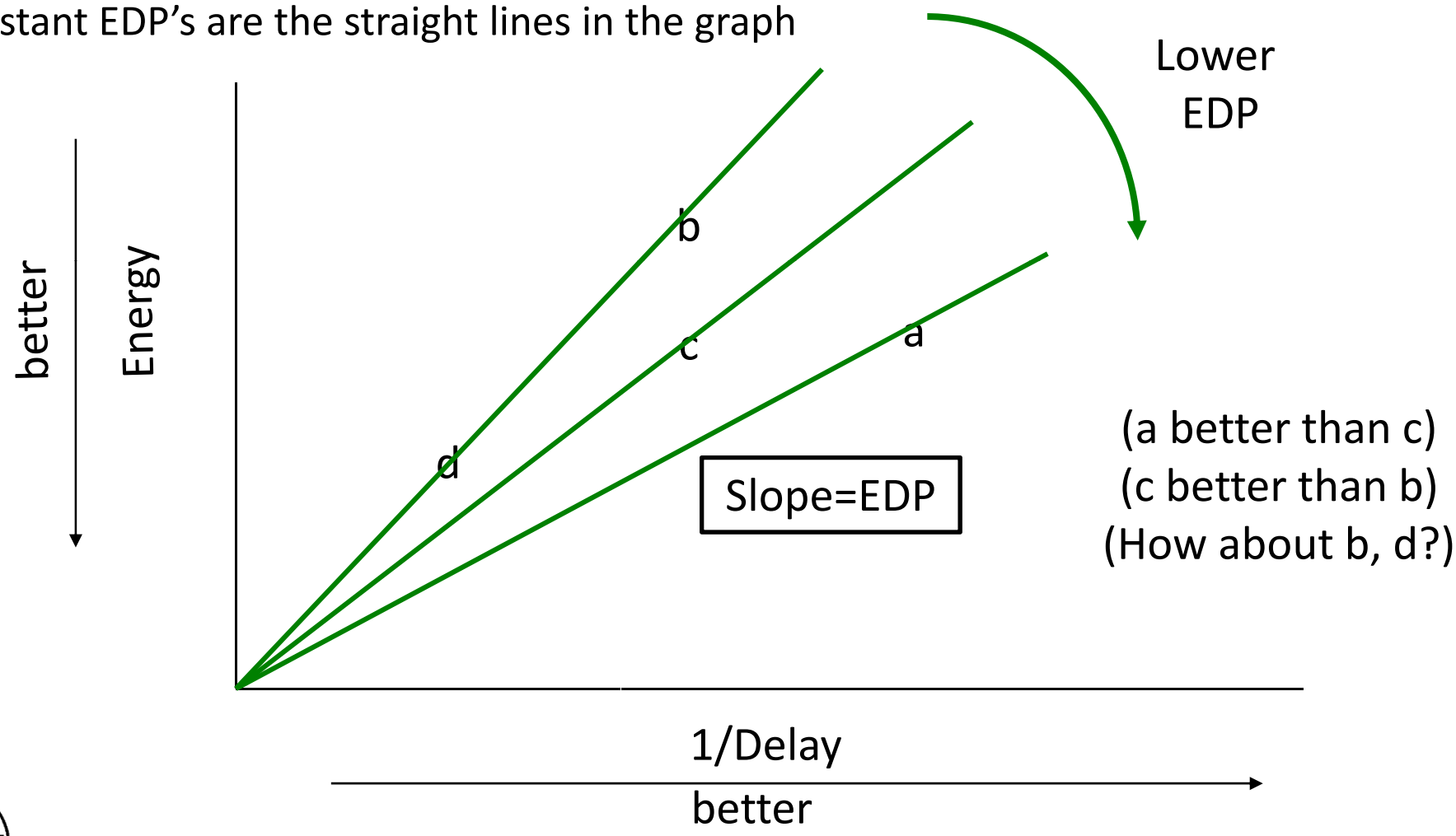
PDP and EDP



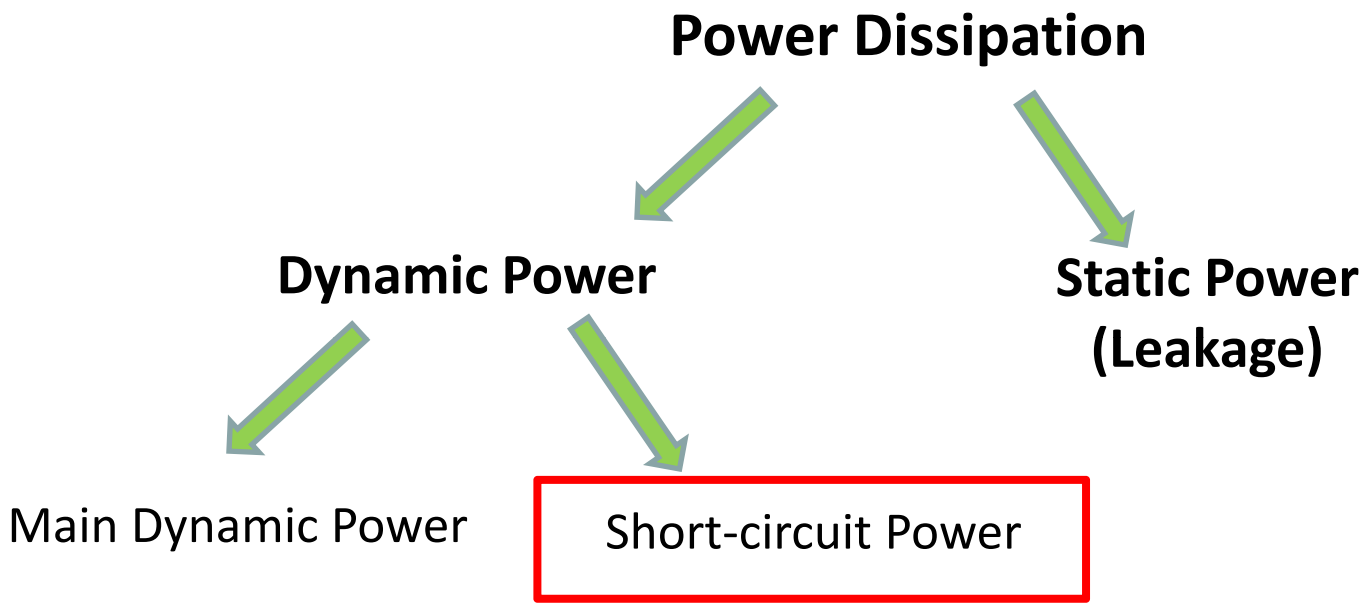
Rule-of-thumb: $V_{DD}^{Opt} = \frac{3}{2}\left(V_t + \frac{V_{DS}^{Sat}}{2}\right)$

Understanding Tradeoffs

- Which design is the “best” (fastest, coolest, both) ?
  - Constant EDP’s are the straight lines in the graph

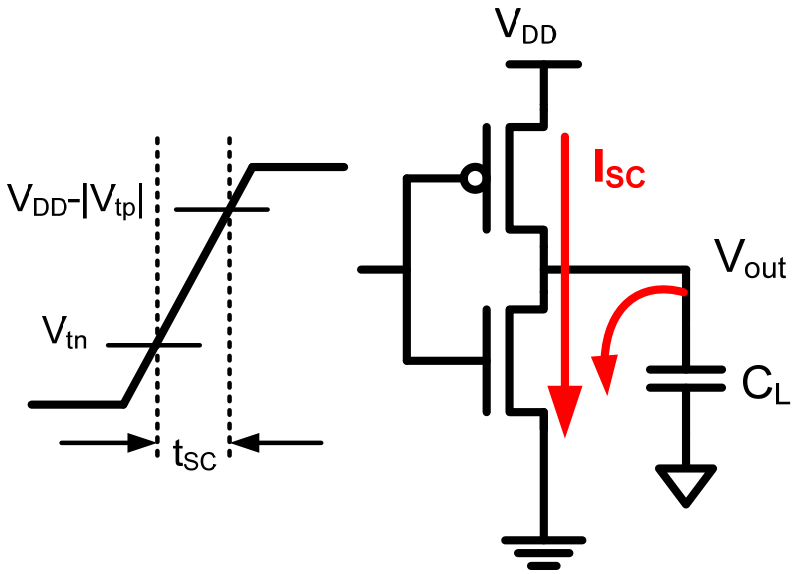


Power Dissipation

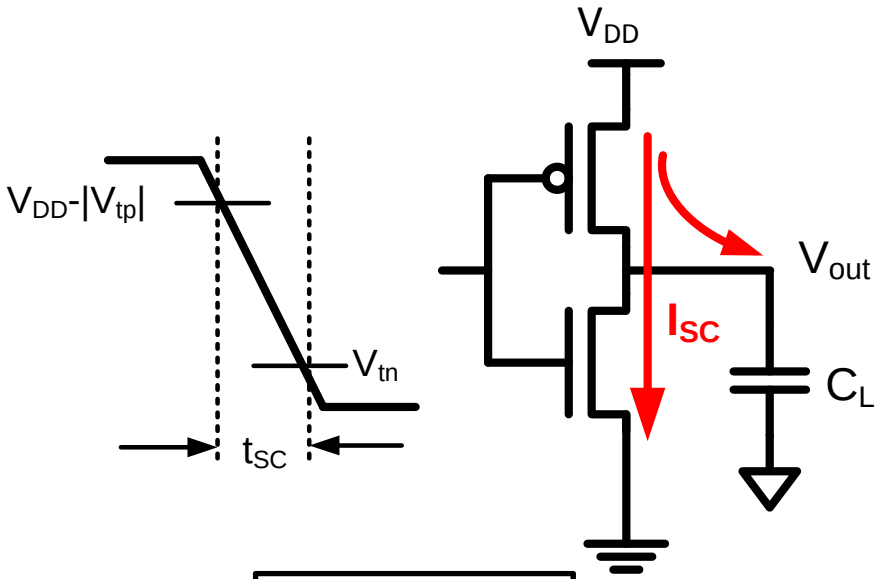


Short Circuit Power Consumption

❑ Finite slope of the input signal causes a direct current path between V_{DD} and GND for a short period of time during switching when both the NMOS and PMOS transistors are conducting.



Rise Time



Fall Time

Short Circuit Currents Determinates

$$t_{sc} = t_{sc}^{r} + t_{sc}^{f}$$

$$I = C\frac{dv}{dt} \Rightarrow t_{sc} I_{sc,avg} = C_{sc} V_{DD}$$

$$I_{sc} = \frac{t_{sc} \cdot I_{sc,avg}}{T}$$

Peak and duration of I_{sc} both increase as the input slope decreases

$$P_{sc} = I_{sc} V_{DD} = t_{sc} I_{sc,avg} V_{DD} f$$

$$\Rightarrow P_{sc} = C_{sc} V_{DD}^2 f = \alpha_{sc} C_L V_{DD}^2 f$$

$$\Rightarrow E_{sc} = C_{sc} V_{DD}^2 = \alpha_{sc} C_L V_{DD}^2$$

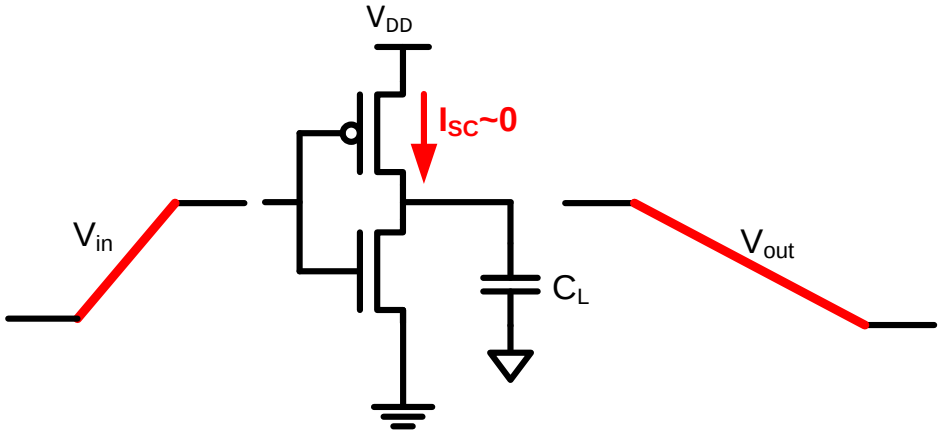
Short Circuit Currents Determinates

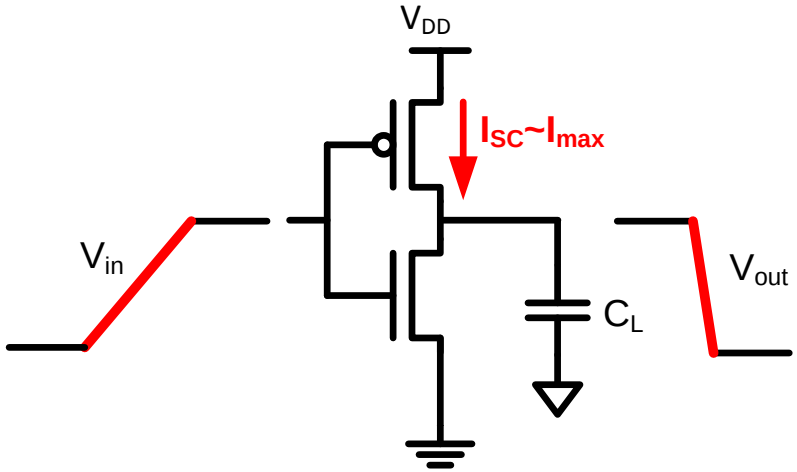
$$P_{sc} = \alpha_{sc} C_L V_{DD}^2 f$$

$$E_{sc} = \alpha_{sc} C_L V_{DD}^2$$

- I_{peak} determined by
  - Saturation current of the P and N transistors, which depend on their sizes, process technology, temperature, etc.
  - Strong function of the ratio between input and output slopes
    - Function of C_L

Impact of C_L on I_{sc}





Large capacitive load

⇩

Output fall time significantly larger than input rise time.

Small capacitive load

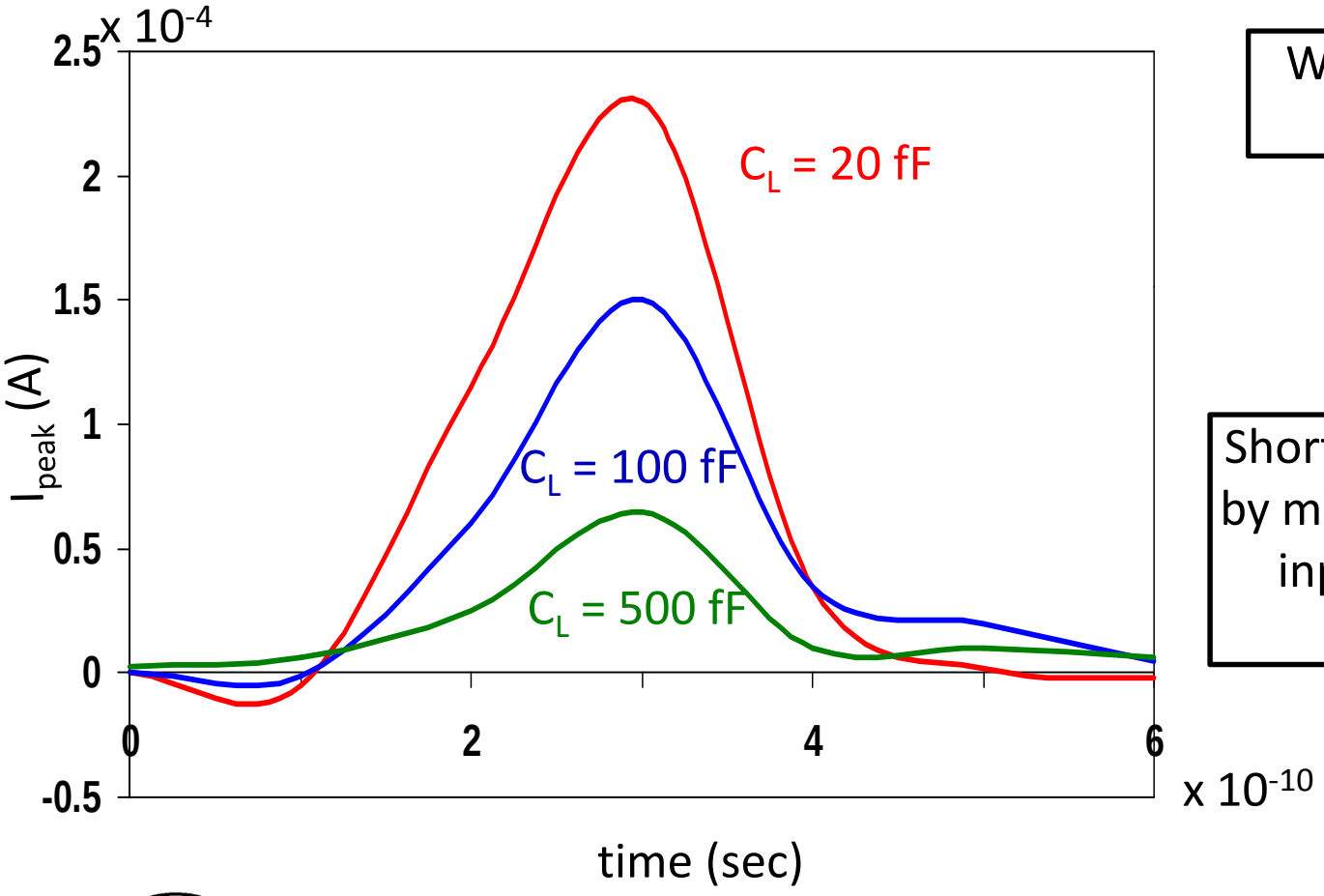
⇩

Output fall time substantially smaller than the input rise time.

❑ As the source-drain voltage of the PMOS is approximately 0 during transition, the device shuts off without ever delivering any current, so I_{sc} is close to zero.

❑ Drain-source voltage of PMOS equals VDD for most of the transition period, giving maximum I_{sc}

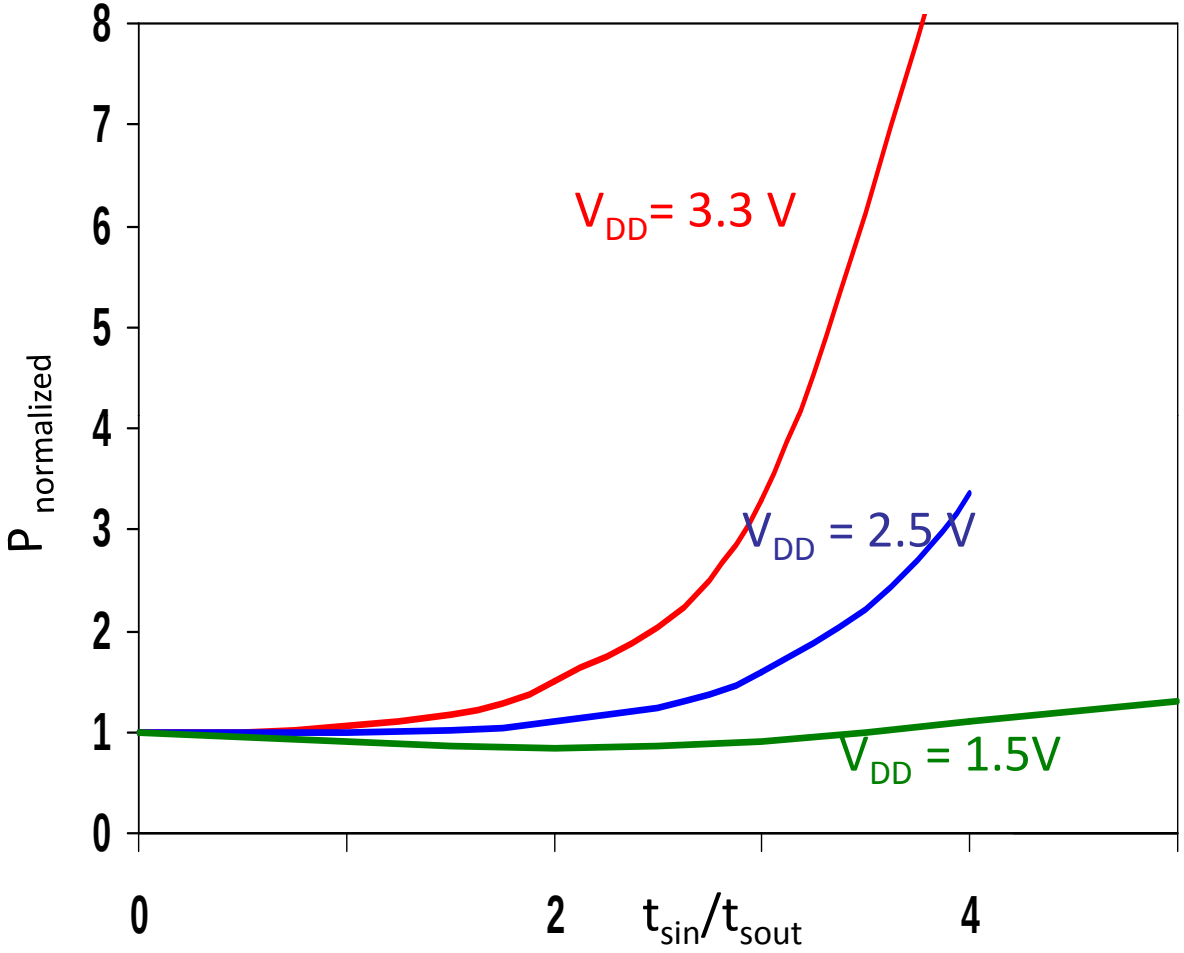
I_{peak} as a Function of C_L



When load capacitance is small, I_{peak} is large.

Short circuit dissipation is minimized by matching the rise/fall times of the input and output signals - slope engineering.

P_{sc} as a Function of Rise/Fall Times



When load capacitance is small ($t_{sin}/t_{sout} > 2$ for $V_{DD} > 2V$) the power is dominated by P_{sc}

If $V_{DD} < V_{Tn} + |V_{Tp}|$ then P_{sc} is eliminated since both devices are never ON at the same time.

- For large capacitance values, all the power dissipation is devoted to charging and discharging the load capacitance.
- When the rise/fall times of inputs and outputs are equalized, most power dissipation is associated with dynamic power and only a minor fraction (<10%) is devoted to P_{sc}.

Dynamic Power Example

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 L_{min}
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 L_{min}
    - Activity factor = 0.02
  - 1.0 V 25 nm process
  - C = 1 fF/mm (gate) + 0.8 fF/mm (diffusion)
- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Dynamic Power Example

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

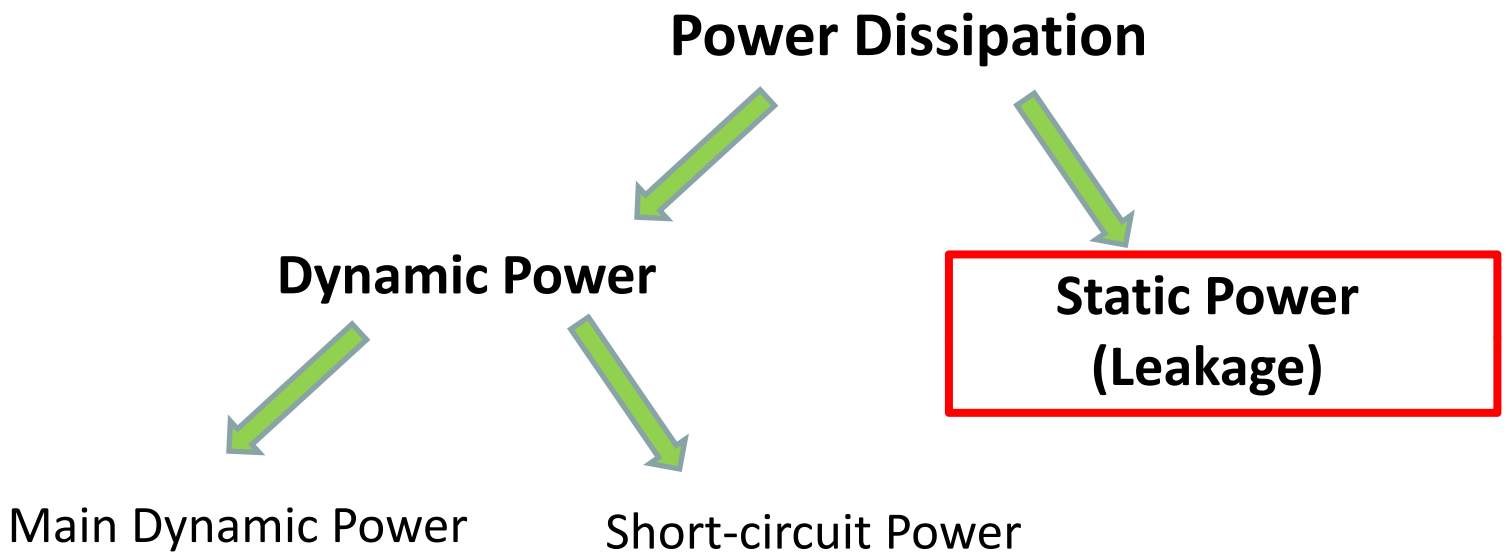
$$C_{\text{mem}} = \left(950 \times 10^6\right)\left(4\lambda\right)\left(0.025\mu m / \lambda\right)\left(1.8\, fF / \mu m\right) = 171 \text{ nF}$$

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

Outline

- Introduction
- Dynamic Power Dissipation
- Static Power Dissipation

Power Dissipation



Power Dissipation: Static Power

- Non-ideal Effects: small leakage current flows through the OFF transistor (I_{static})

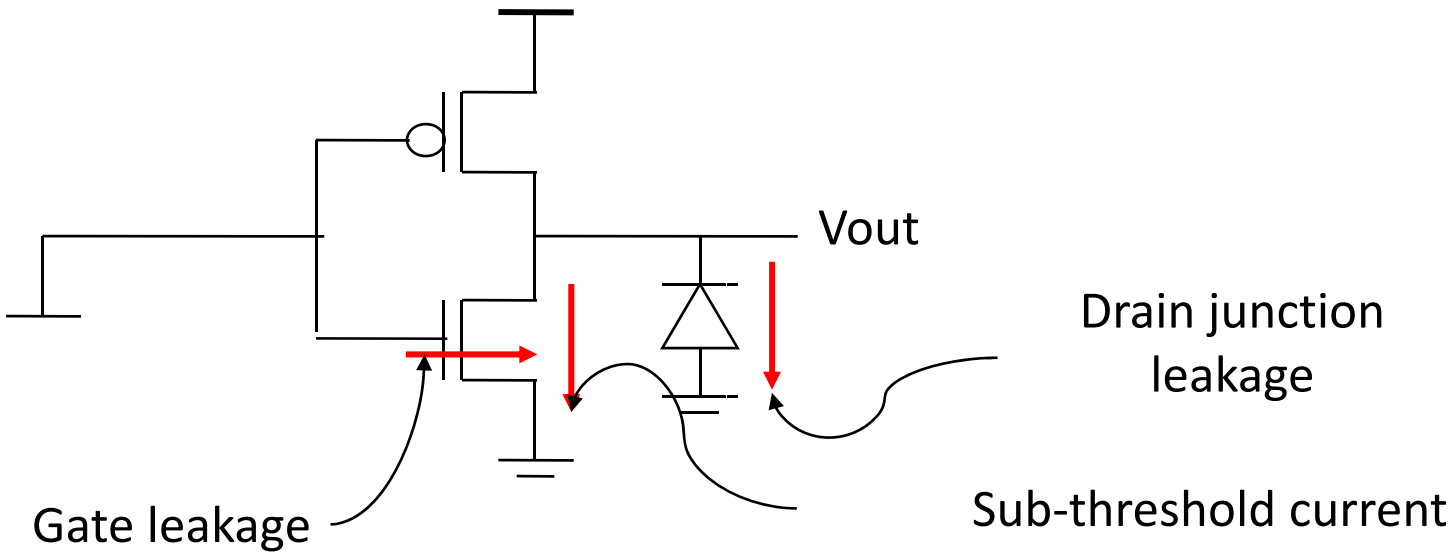
$$P_{static} = \frac{1}{T}\int_0^T i_{static} V_{DD} dt = I_{static} V_{DD}$$

- Sources of Leakage:
  - **Sub-threshold Conduction**: Exponentially increases as V_T scales down
  - **Tunneling through the gate oxide**: Exponentially increases as oxide thickness decreases
    - (Important for 130nm and smaller technologies)
  - **Leakage through reverse-biased diodes**

Static power dissipation an issue in deep sub-micron processes

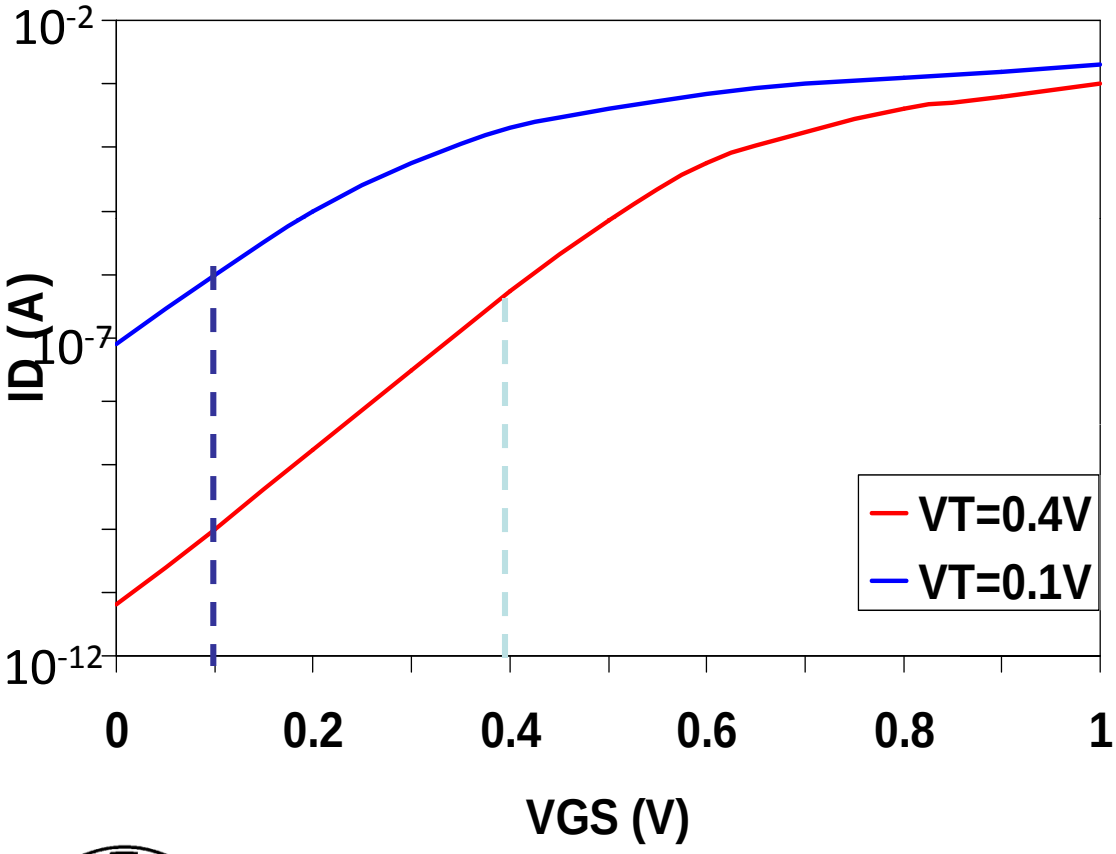
Power Dissipation: Static (Leakage) Power

- Sub-threshold current is the dominant factor.
- All increase exponentially with temperature!



Leakage as a Function of V_T

- Continued scaling of supply voltage and the subsequent scaling of threshold voltage will make sub-threshold conduction a dominate component of power dissipation.



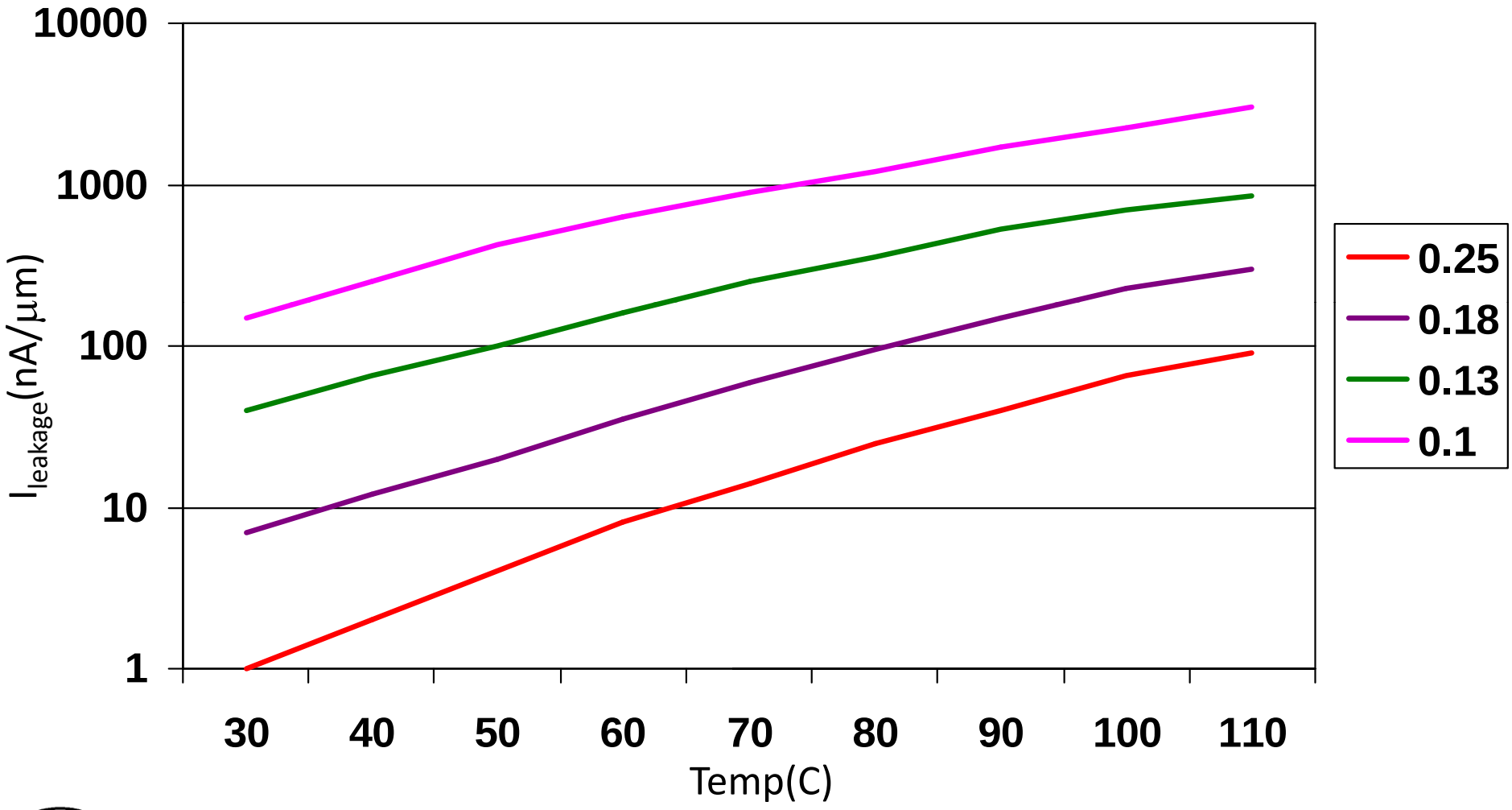
An 90mV/decade V_T roll-off - so each 255mV increase in V_T gives 3 orders of magnitude reduction in leakage (but adversely affects performance)

TSMC Processes Leakage and V_T

	CL018 G	CL018 LP	CL018 ULP	CL018 HS	CL015 HS	CL013 HS
V_{dd}	1.8 V	1.8 V	1.8 V	2 V	1.5 V	1.2 V
T_{ox} (effective)	42 Å	42 Å	42 Å	42 Å	29 Å	24 Å
L_{gate}	0.16 μm	0.16 μm	0.18 μm	0.13 μm	0.11 μm	0.08 μm
I_{DSat} (n/p) (μA/μm)	600/260	500/180	320/130	780/360	860/370	920/400
I_{off} (leakage) (ρA/μm)	20	1.60	0.15	300	1,800	13,000
V_{Tn}	0.42 V	0.63 V	0.73 V	0.40 V	0.29 V	0.25 V
FET Perf. (GHz)	30	22	14	43	52	80

(G: generic, LP: low power, ULP: ultra low power, HS: high speed)

Exponential Increase in Leakage Currents



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

Gate Leakage

- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches sub-threshold 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}

Static Power Example

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

Solution

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

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

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

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

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

Review: Designing Fast CMOS Gates

- Transistor sizing
- Progressive transistor sizing
  - MOS closest to the output is smallest of series MOS transistors
- Transistor ordering
  - put latest arriving signal closest to the output
- Logic structure reordering
  - replace large fan-in gates with smaller fan-in gate network
- Apply “logical effort”
- Buffer (inverter) insertion
  - separate large fan-in from large C_L with buffers
  - uses buffers so there are no more than four TGs in series