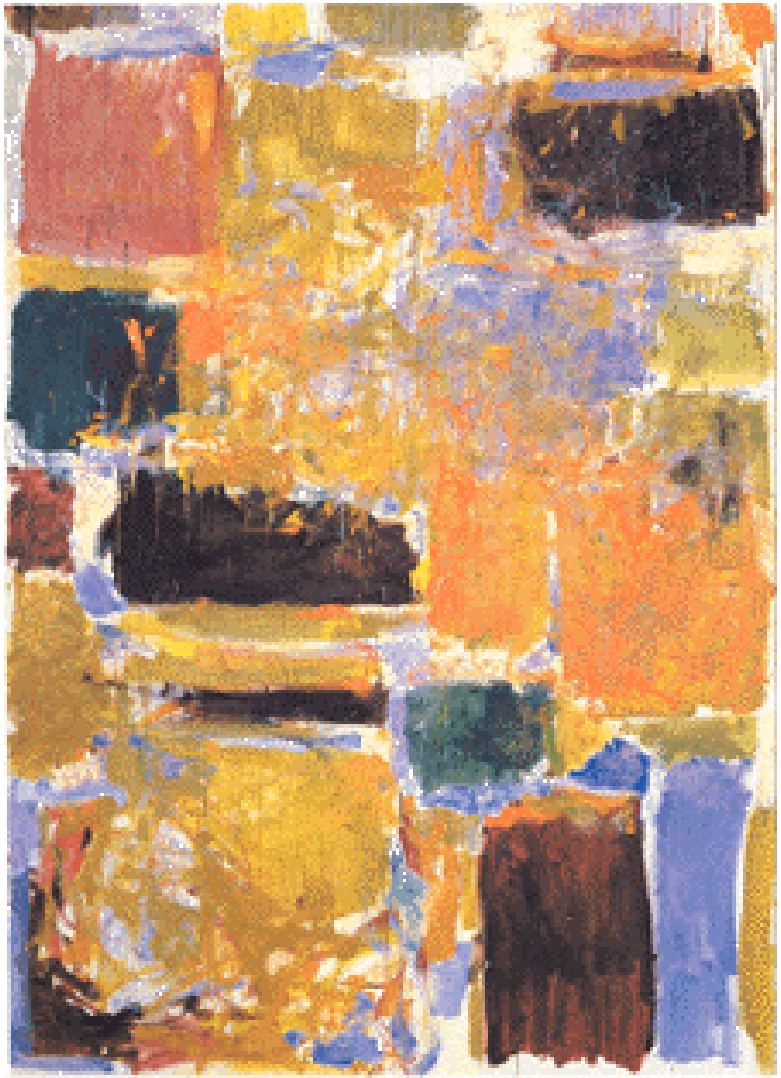


Power Dissipation

Where Does Power Go in CMOS?

- **Dynamic Power Consumption**

  Charging and Discharging Capacitors

- **Short Circuit Currents**

  Short Circuit Path between Supply Rails during Switching

- **Leakage**

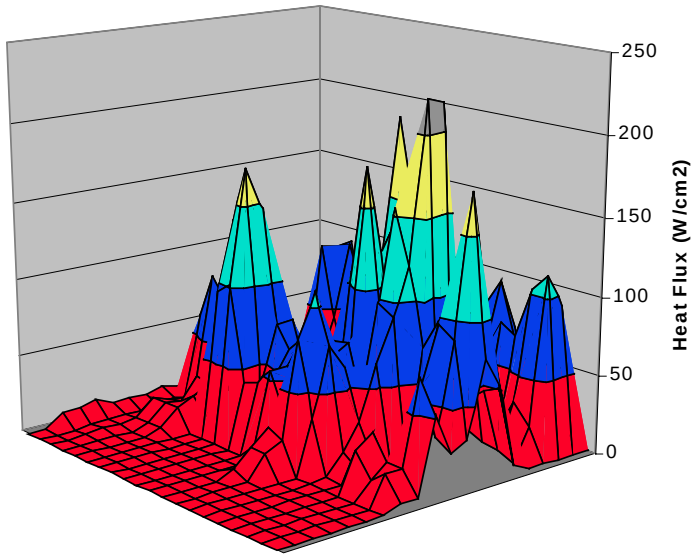
  Leaking diodes and transistors

Why Power Matters

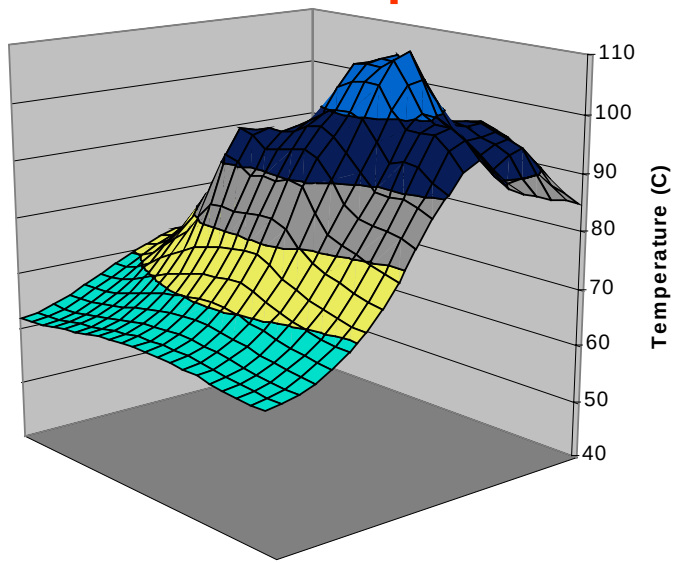
- Packaging costs
- Power supply rail design
- Chip and system cooling costs
- Noise immunity and system reliability
- Battery life (in portable systems)
- Environmental concerns
  - Office equipment accounted for 5% of total US commercial energy usage in 1993

Chip Power Density Distribution

Power Map

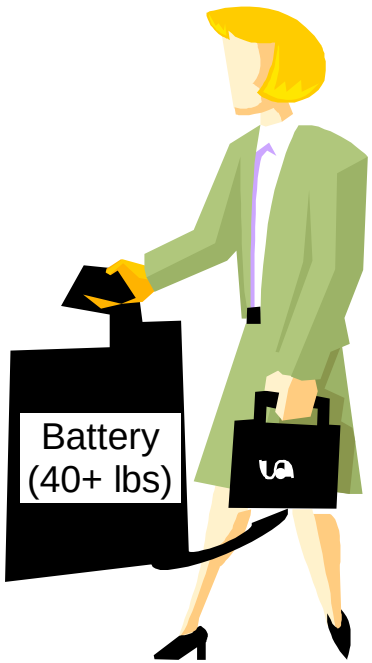


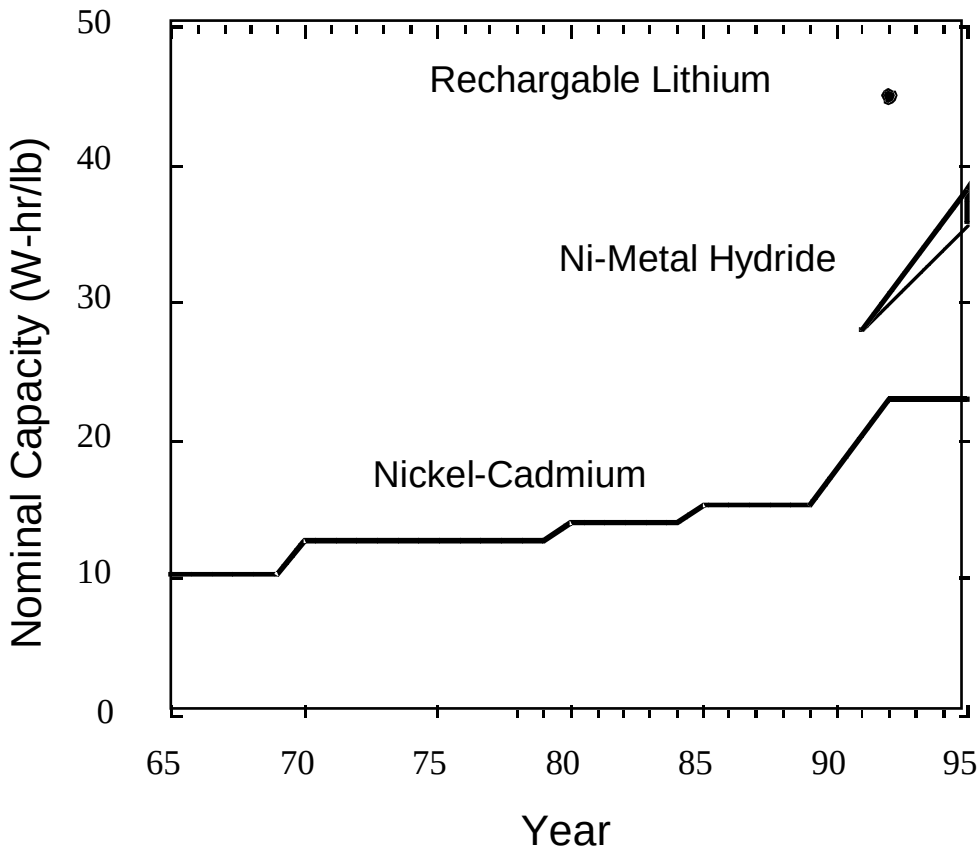
On-Die Temperature



- Power density is not uniformly distributed across the chip
- Silicon is not a good heat conductor
- Max junction temperature is determined by hot-spots
  - Impact on packaging, w.r.t. cooling

Why worry about power ? -- Battery Size/Weight





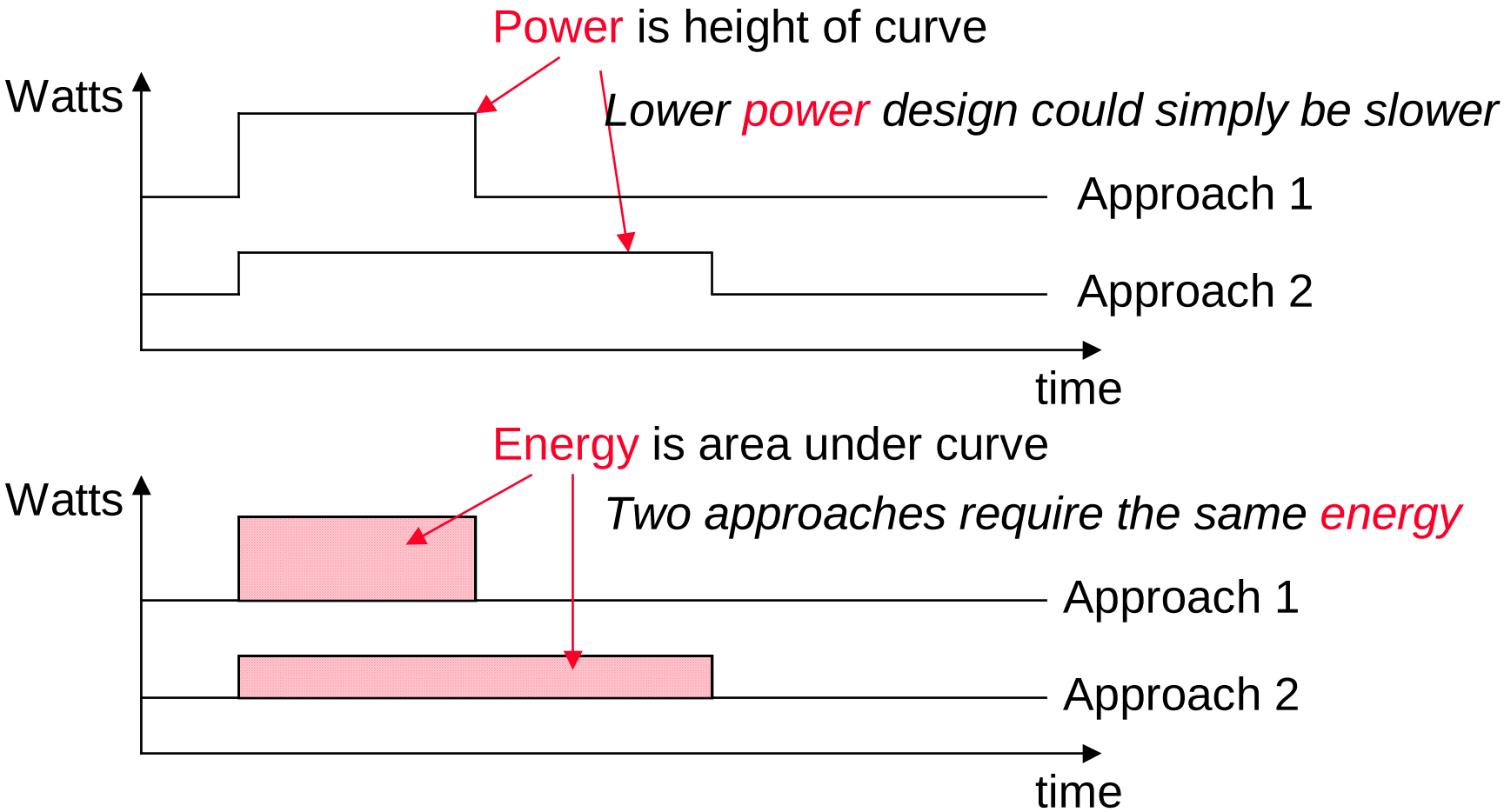
Expected battery lifetime increase over the next 5 years: 30 to 40%

From Rabaey, 1995

Power and Energy Figures of Merit

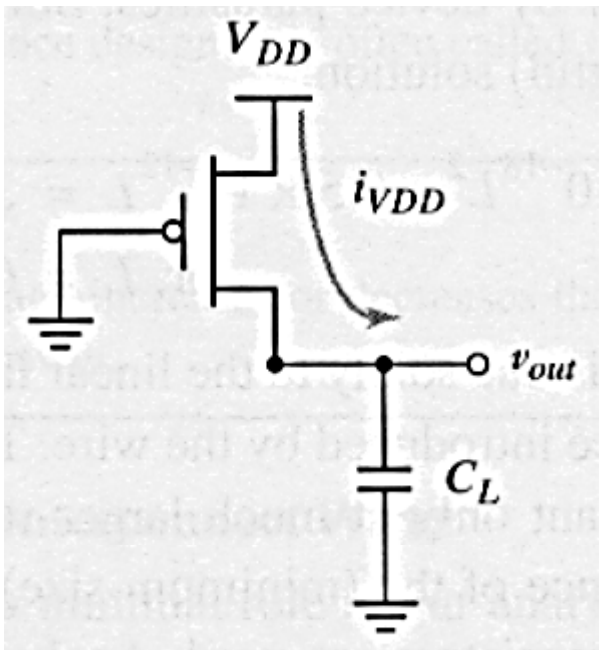
- Power consumption in Watts
  - determines battery life in hours
- Peak power
  - determines power ground wiring designs
  - sets packaging limits
  - impacts signal noise margin and reliability analysis
- Energy efficiency in Joules
  - rate at which power is consumed over time
- Energy = power * delay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

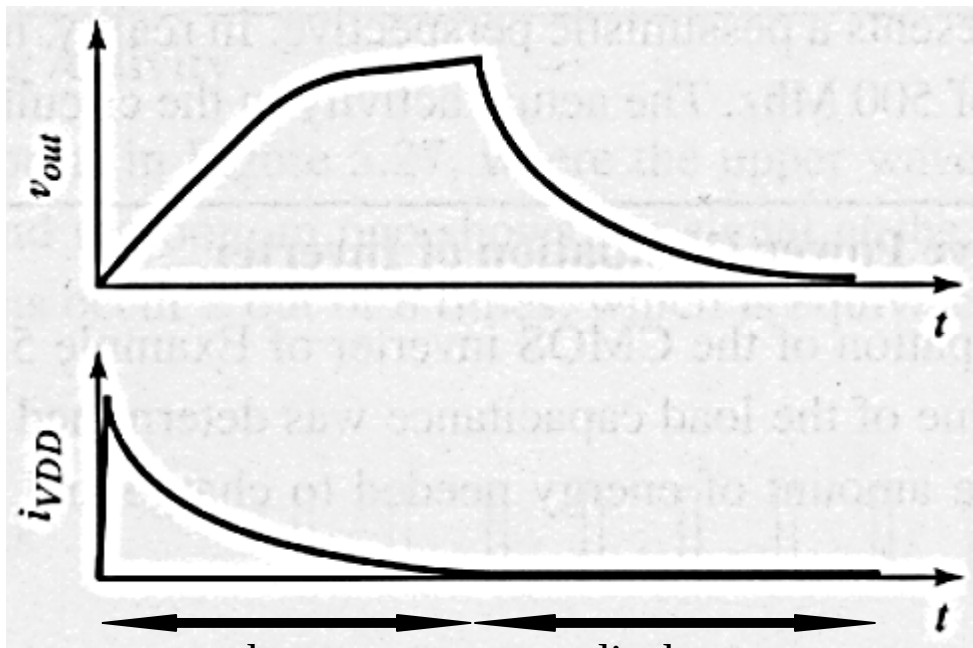
Power versus Energy



$$E_{V_{DD}} = \int_0^{\infty} i_{V_{DD}}(t) V_{DD}\, dt = V_{DD} \int_0^{\infty} C_L \frac{dv_{out}}{dt} dt = C_L V_{DD} \int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2$$

$$E_C = \int_0^{\infty} i_{V_{DD}}(t) v_{out}\, dt = \int_0^{\infty} C_L \frac{dv_{out}}{dt} v_{out}\, dt = C_L \int_0^{V_{DD}} v_{out}\, dv_{out} = \frac{C_L V_{DD}^2}{2}$$





$$P_{dyn} = C_L V_{DD}^2 f_{0\rightarrow 1}$$

$f_{0\rightarrow 1}$ represents the frequency of energy-consuming transitions ($0\rightarrow 1$ transitions for static CMOS)

Each switching activity (consisting of an $L\rightarrow H$ and $H\rightarrow L$ transition) takes a fixed amount of energy, equal to $C_L V^2_{DD}$, from the supply. Half of it is dissipated at the PMOS and the other half is stored in C_L. The stored energy is dissipated in NMOS transistor during discharge. The energy dissipation is independent of the size (and hence the resistance) of the transistors.

Consider: 0.25μm CMOS chip. 500MHz. Clock, average load capacitance of 15fF/gate (assuming a fan-out of 4) and V_{DD}=2.5V Power consumption per gate ≈ 50 μW. For a 1 million gate design, if a transition occurs at every clock edge, that would result a power consumption of **50 W**!

NOTE: Multiplying something by frequency is equivalent to dividing it by time (T = 1/f)

For the CMOS inverter of C_L = 6 fF, V_{DD} = 2.5 V, the energy needed to charge (and discharge) C_L:

$$E_{dyn} = C_L V^2_{DD} = 37.5 \text{ fJ}$$

The maximum possible (hypothetical) switching rate:

$$T = 1/f = t_{pLH} + t_{pHL} = 2\, t_p$$

For a t_p of 32.5ps, the dynamic power dissipation of the circuit:

$$P_{dyn} = E_{dyn}/2t_p = 580\mu W$$

An actual circuit can not be switched at this maximum rate, and even if it would be, the output does not swing from rail to rail. For a rate of 4 GHz. (T = 250ps), the dissipation reduces to 150 μW.

While the switching activity is easily computed for an inverter, it is far more complex in the case of more complex gates and circuits.

The switching activity of such a circuit is a function of the nature and the statistics of input signals. If the input signals remain unchanged, no switching happens and the dynamic power consumption is zero.

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} = C_L V_{DD}^2 P_{0\to1} f = C_{EFF} V_{DD}^2 f$$

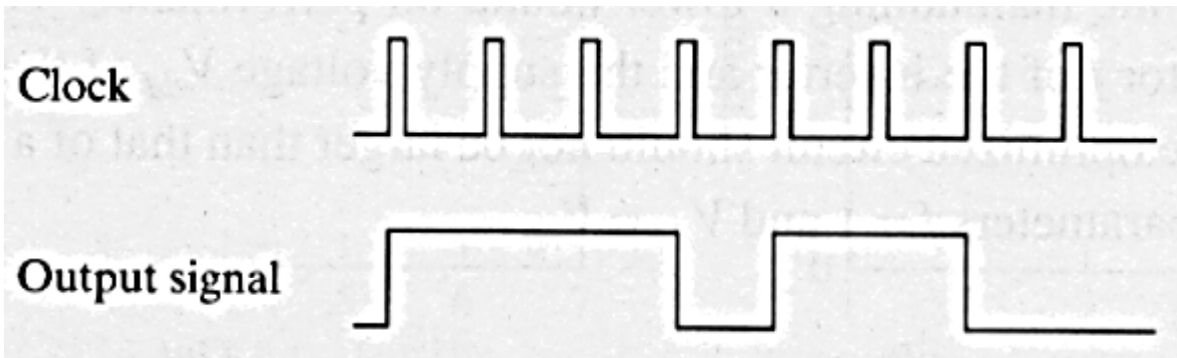
$C_{EFF} = P_{0\to1} C_L$ f = maximum event rate (clock rate)

$$E_{dyn} = C_L V_{DD}^2 P_{0\to1} = C_{EFF} V_{DD}^2$$

C_{EFF} = average capacitance switched every clock cycle

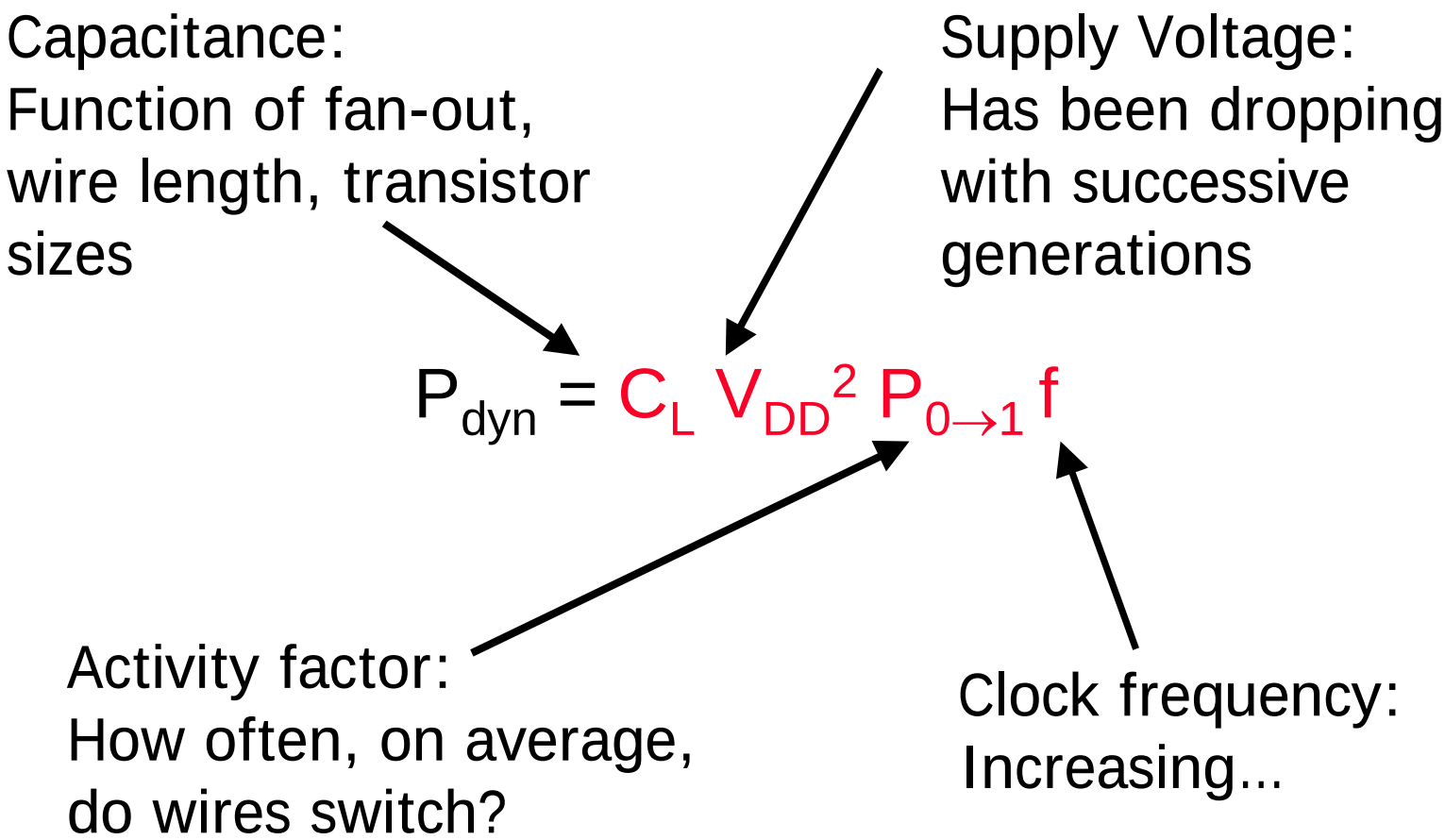
Reducing V_{DD} has a quadratic effect on P_{dyn}

$P_{0\to1}$ is the probability that a clock event results with a power consuming transition at the output



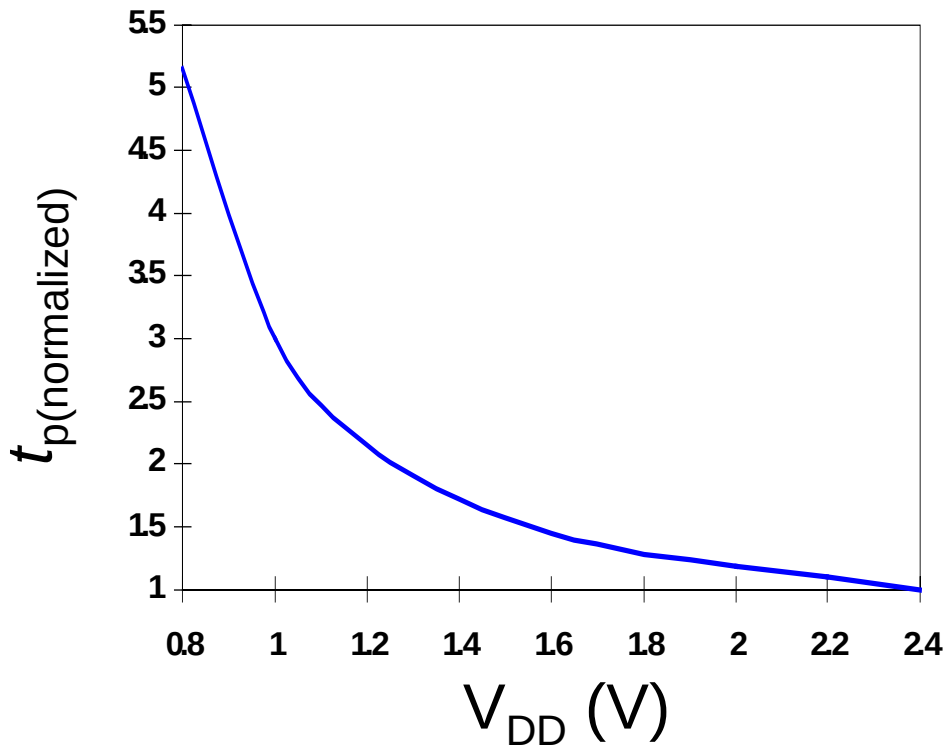
2 out of 8 times power consuming transitions $P_{0\to1}$=0.25

Lowering Dynamic Power



Dynamic Power as a Function of V_{DD}

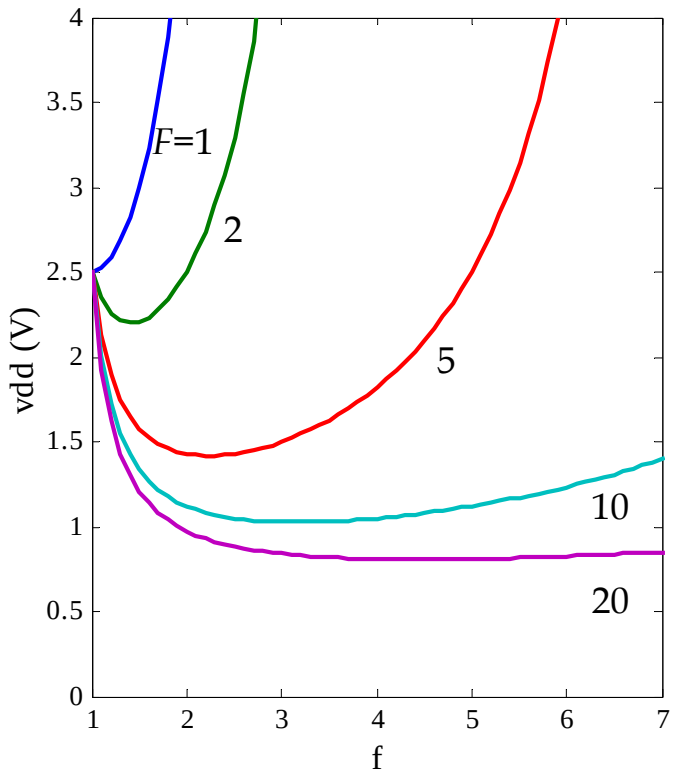
- Decreasing the V_{DD} decreases dynamic energy consumption (quadratically)
- But, increases gate delay (decreases performance)



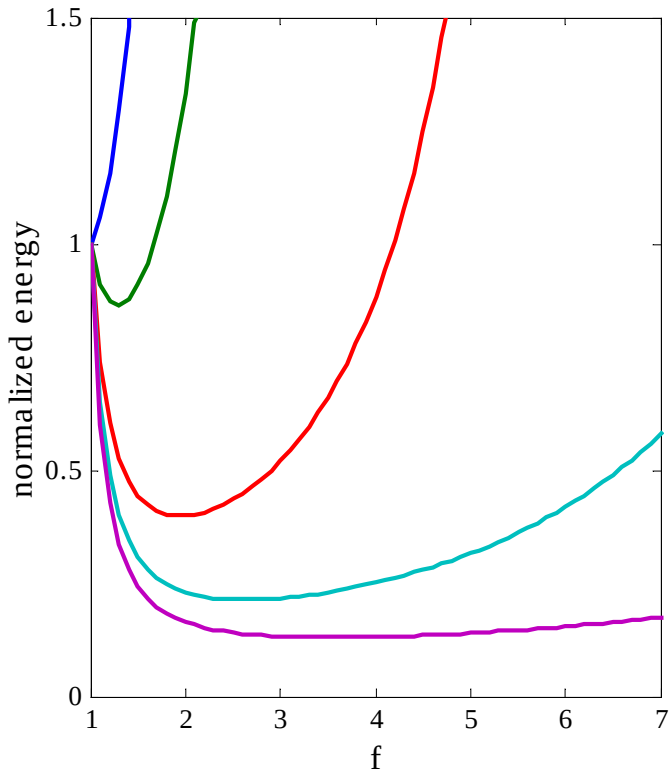
- Determine the critical path(s) at design time and use high V_{DD} for the transistors on those paths for speed. Use a lower V_{DD} on the other gates, especially those that drive large capacitances (as this yields the largest energy benefits).

Transistor Sizing

V_{DD}=f(f)

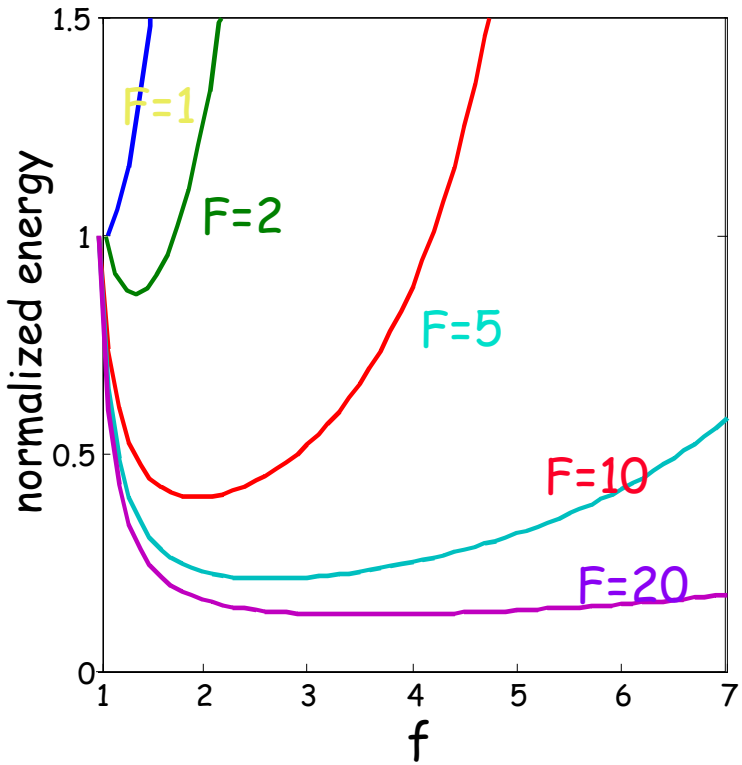


E/E_{ref}=f(f)



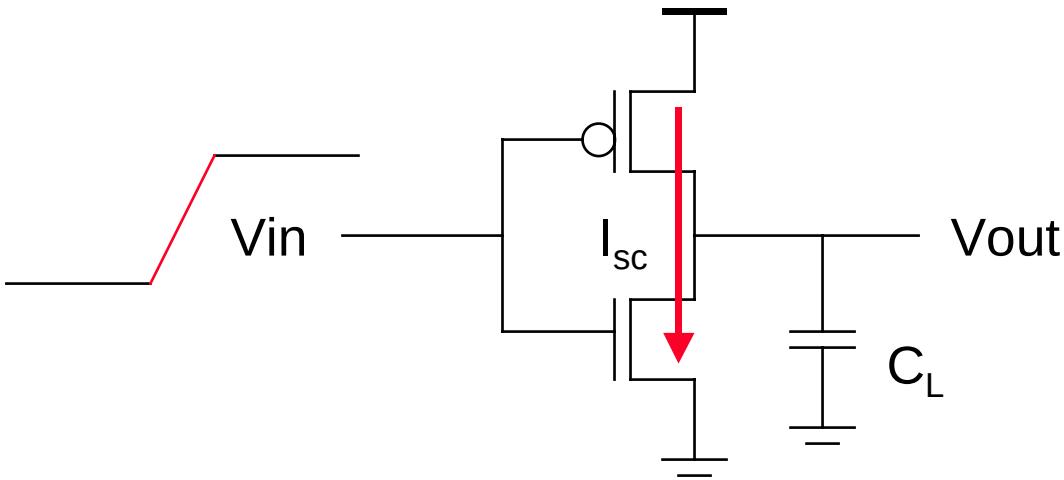
Dynamic Power as a Function of Device Size

- Device sizing combined with supply voltage reduction is very effective in dynamic energy consumption
  - gain is largest (A factor of ~10) for networks with large overall effective fan-outs ($F = C_L/C_{g,1}$)
- The optimal gate sizing factor (f) for dynamic energy is smaller than the one for performance, especially for large F's
  - e.g., for F=20, f_{opt}(energy) = 3.53 while f_{opt}(performance) = 4.47
- If energy is a concern avoid oversizing beyond the optimal

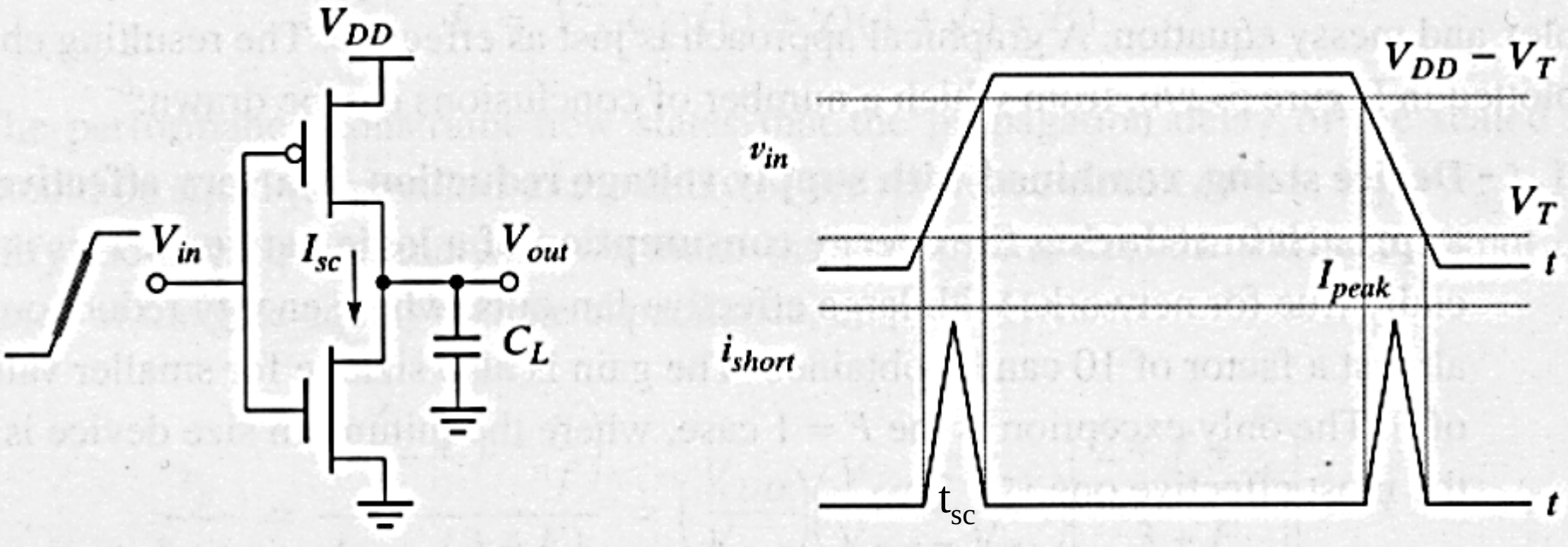


From Nikolic, UCB

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.



Under the assumption that the short circuit current spikes can be approximatted as triangles, and the inverter is symmetrical:

$$E_{dp} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak}$$

$$t_{sc} = \frac{V_{DD} - 2V_T}{V_{DD}}t_s \approx \frac{V_{DD} - 2V_T}{V_{DD}}\ \frac{t_{r(f)}}{0.8}$$

dp stands for “direct path”, t_s: 0-%100 transition time

Short Circuit Currents Determinates

$$E_{dp} = t_{sc} V_{DD} I_{peak} P_{0\rightarrow1}$$

$$P_{dp} = t_{sc} V_{DD} I_{peak} f_{0\rightarrow1}$$

- I_{peak} determined by
  - the 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
    - a function of C_L

$$P_{dp} = t_{sc}\, V_{DD}\, I_{peak}\, f_{0\to1} = t_{sc}\, I_{peak}\, V_{DD}\, P_{0\to1}\, f$$

Remembering: $P_{dyn} = C_L\, V^2_{DD}\, P_{0\to1}\, f = C_L\, V_{DD}\, V_{DD}\, P_{0\to1}\, f$

If we write $C_L V_{DD} = t_{sc}\, I_{peak}$ in P_{dyn}, we obtain P_{dp}

or we can write $C_L = t_{sc}\, I_{peak} / V_{DD}$ if we model the short circuit power dissipation by a capasitive load C_{sc}

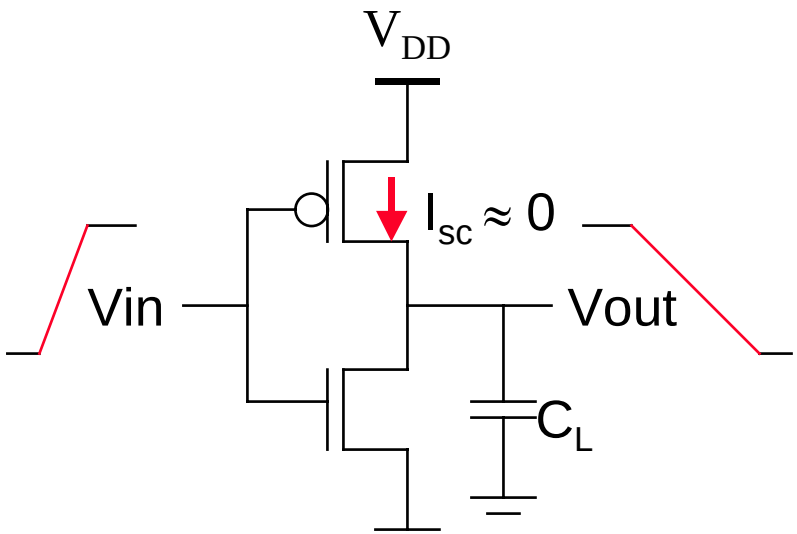
$C_{sc} = C_L\, P_{0\to1} = t_{sc}\, I_{peak}\, P_{0\to1} / V_{DD}$ we can rewrite P_{dp}

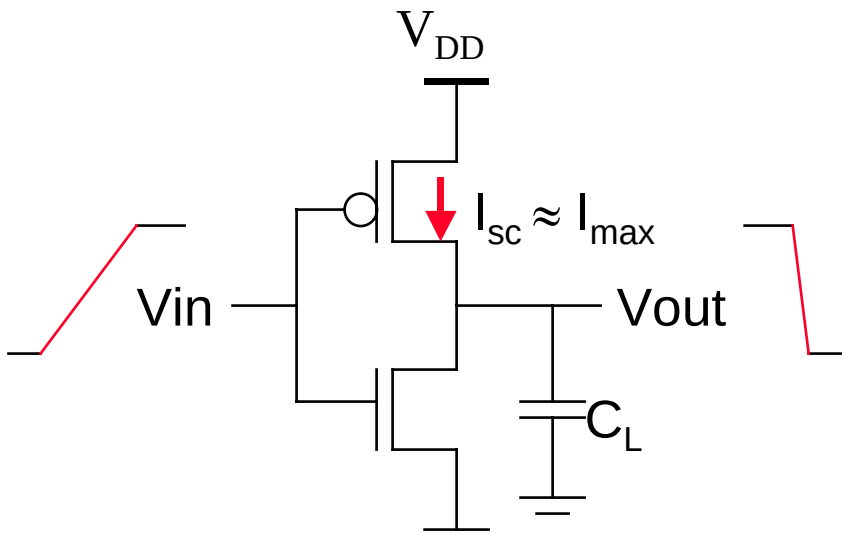
$$P_{dp} = C_{sc}\, V^2_{DD}\, f$$

Short circuit power dissipation can be modeled by adding a load capacitance $C_{sc} = t_{sc}\, I_{peak}\, P_{0\to1} / V_{DD}$ in parellel with C_L

The value of this short-circuit capacitance is a function of V_{DD}, the transistor sizes, and the input/output slope ratio.

Impact of C_L on P_{sc}





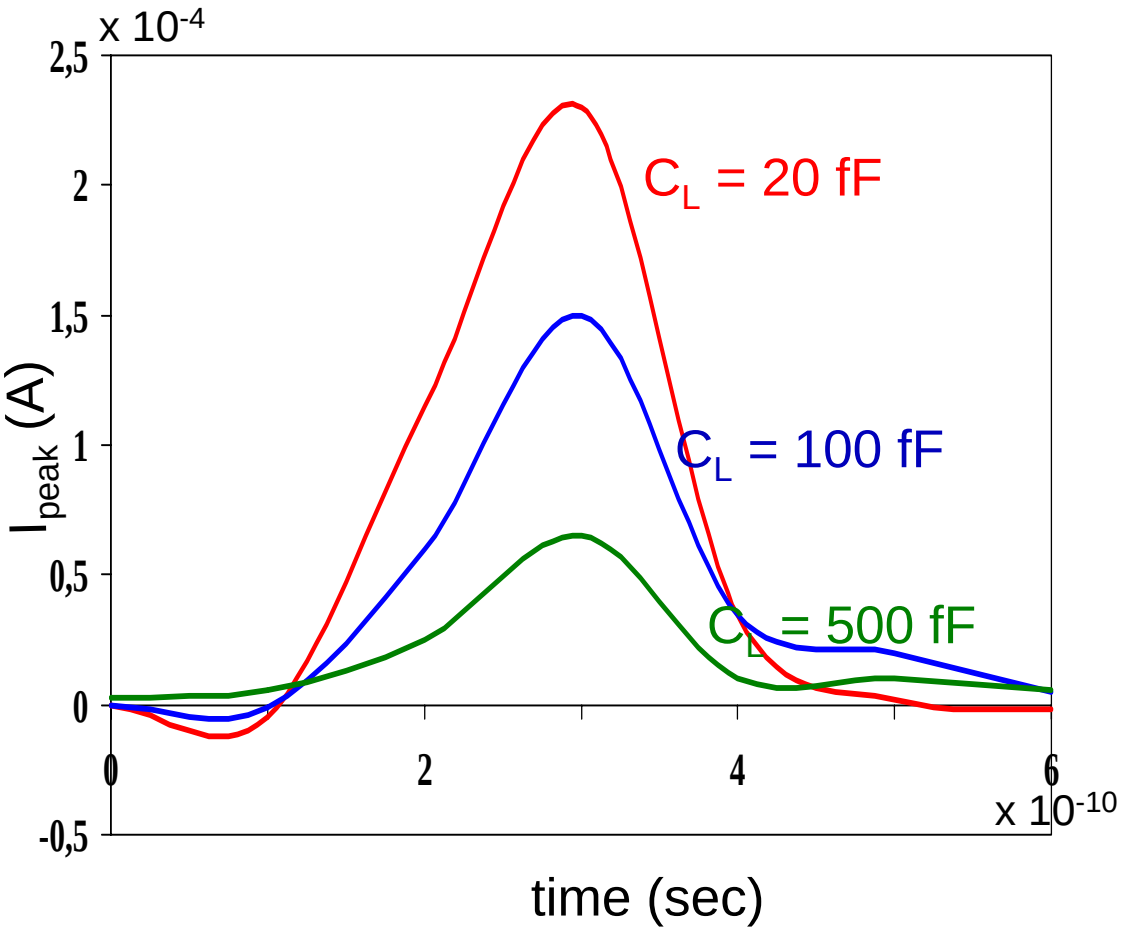
Large capacitive load

Small capacitive load

Output fall time significantly larger than input rise time.

Output fall time substantially smaller than the input rise time.

I_{peak} as a Function of C_L

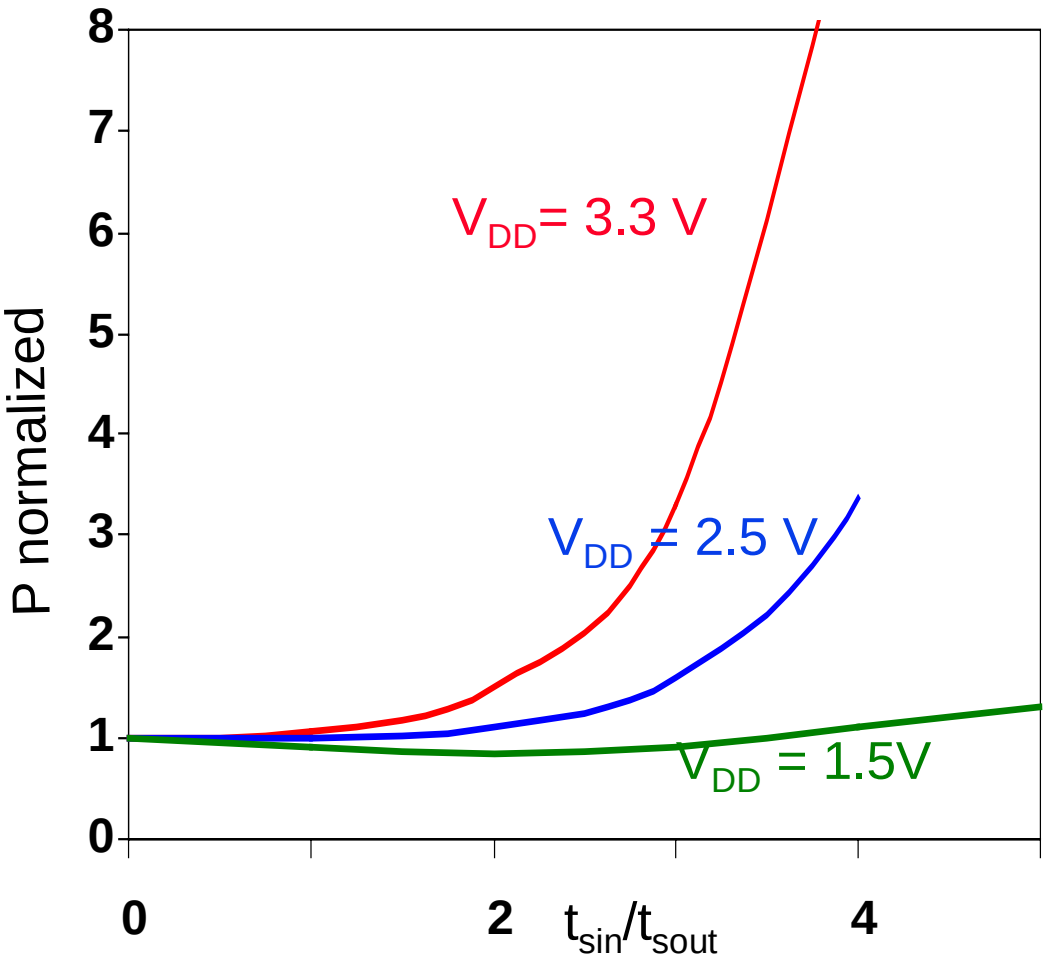


500 psec input slope

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_{dp} as a Function of Rise/Fall Times



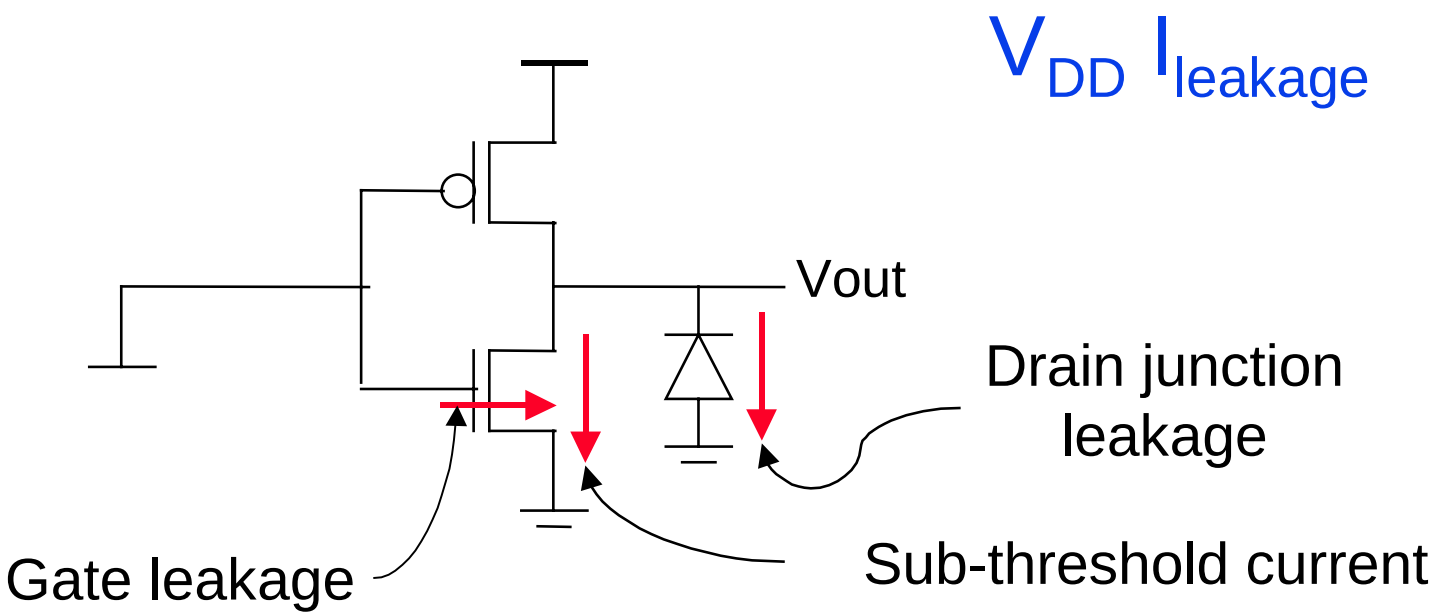
When load capacitance is small ($t_{sin}/t_{sout} > 2$ for $V_{DD} > 2V$) the power is dominated by P_{dp} (short circuit)

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

W/L_p = 1.125 μm/0.25 μm
W/L_n = 0.375 μm/0.25 μm
C_L = 30 fF

normalized wrt zero input rise-time dissipation

Leakage (Static) Power Consumption



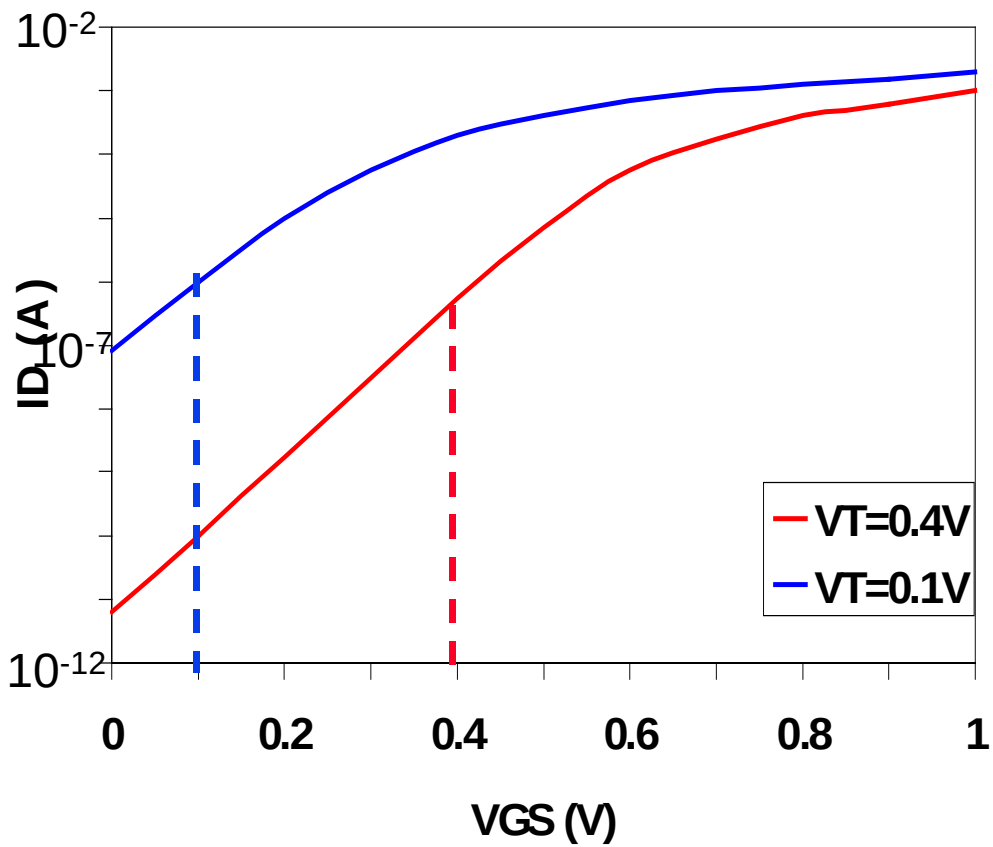
Sub-threshold current is the dominant factor.

All increase exponentially with temperature!

At 85ºC, the leakage increases by x 60

Leakage as a Function of V_T

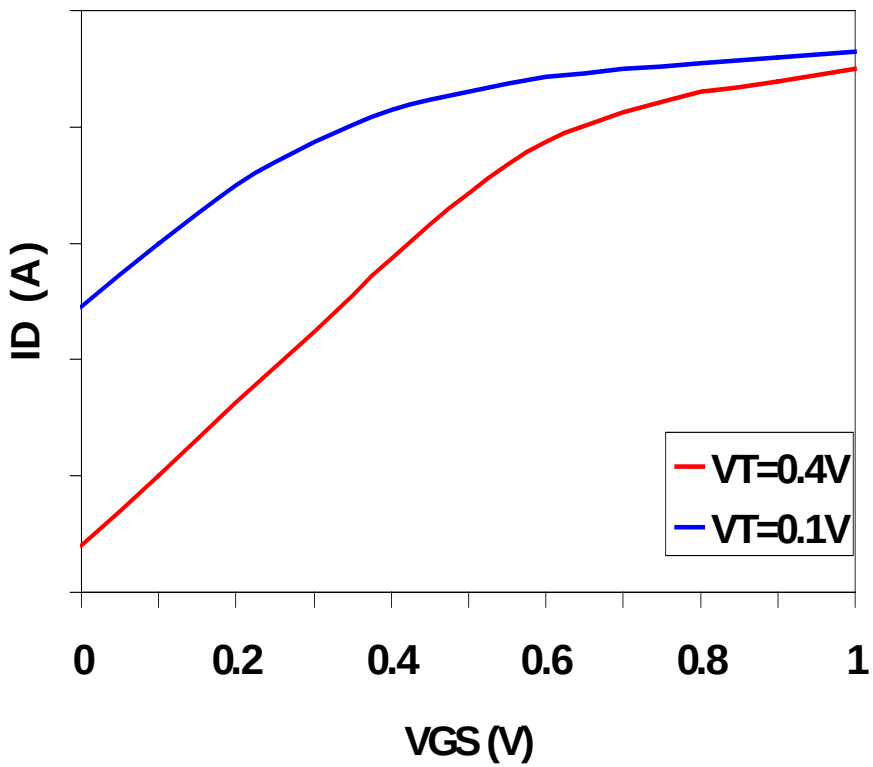
- Continued scaling of supply voltage and the subsequent scaling of threshold voltage will make subthreshold conduction a dominate component of power dissipation.



- An 200mV reduction in V_T (0.5-0.2=0.3V) multiplies the off current by 170. Assuming a million gate design, with V_{DD}=1.5V that results 2.6mW. If V_T is reduced to 100mV power increases to 0.5W!

Leakage as a Function of Design Time V_T

- Reducing the V_T increases the sub-threshold leakage current (exponentially)
  - 90mV reduction in V_T increases leakage by an order of magnitude
- But, reducing V_T decreases gate delay (increases performance)



- Determine the critical path(s) at design time and use low V_T devices on the transistors on those paths for speed. Use a high V_T on the other logic for leakage control.
  - A careful assignment of V_T's can reduce the leakage by as much as 80%

$\downarrow V_T \rightarrow \uparrow$ performance, $\uparrow$ static power

$\downarrow V_{DD} \rightarrow \downarrow$ performance, $\downarrow\downarrow$ dynamic power

For a 0.25μm CMOS process:

$V_{DD} = 3V, \quad V_T = 0.7V$

$V_{DD} = 0.45V, \; V_T = 0.1V$

} Same performance

$\times$45 times smaller dynamic power dissipation.

The optimal operation point depends upon the activity of the circuit. In the presence of a sizable static power dissipation, it is essential that the nonactive modules are powered down (disconnection from the supply rails or lowering the supply voltage)

PDP and EDP

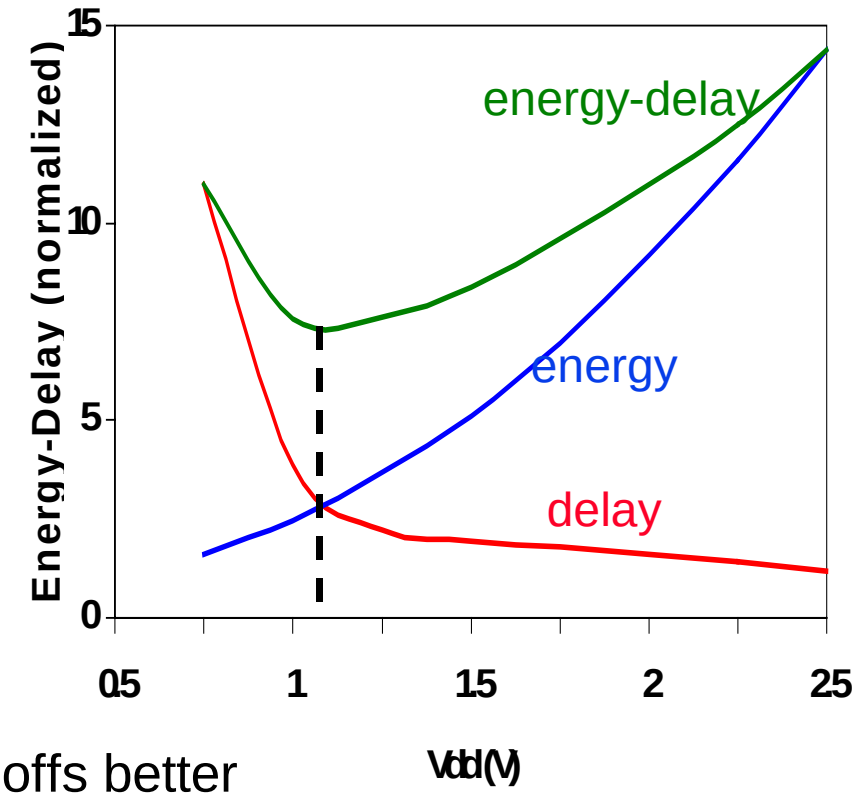
- Power-delay product (PDP) = $P_{av}\ t_p = C_L V_{DD}^2\ f_{max}\ t_p$

  Assuming max possible rate ($f_{max} = 1/2t_p$) PDP = $C_L V_{DD}^2/2$

  - PDP is the average energy consumed per switching event (Watts * sec = Joule)

- Energy-delay product (EDP) = PDP $t_p = P_{av}\ t_p^2$

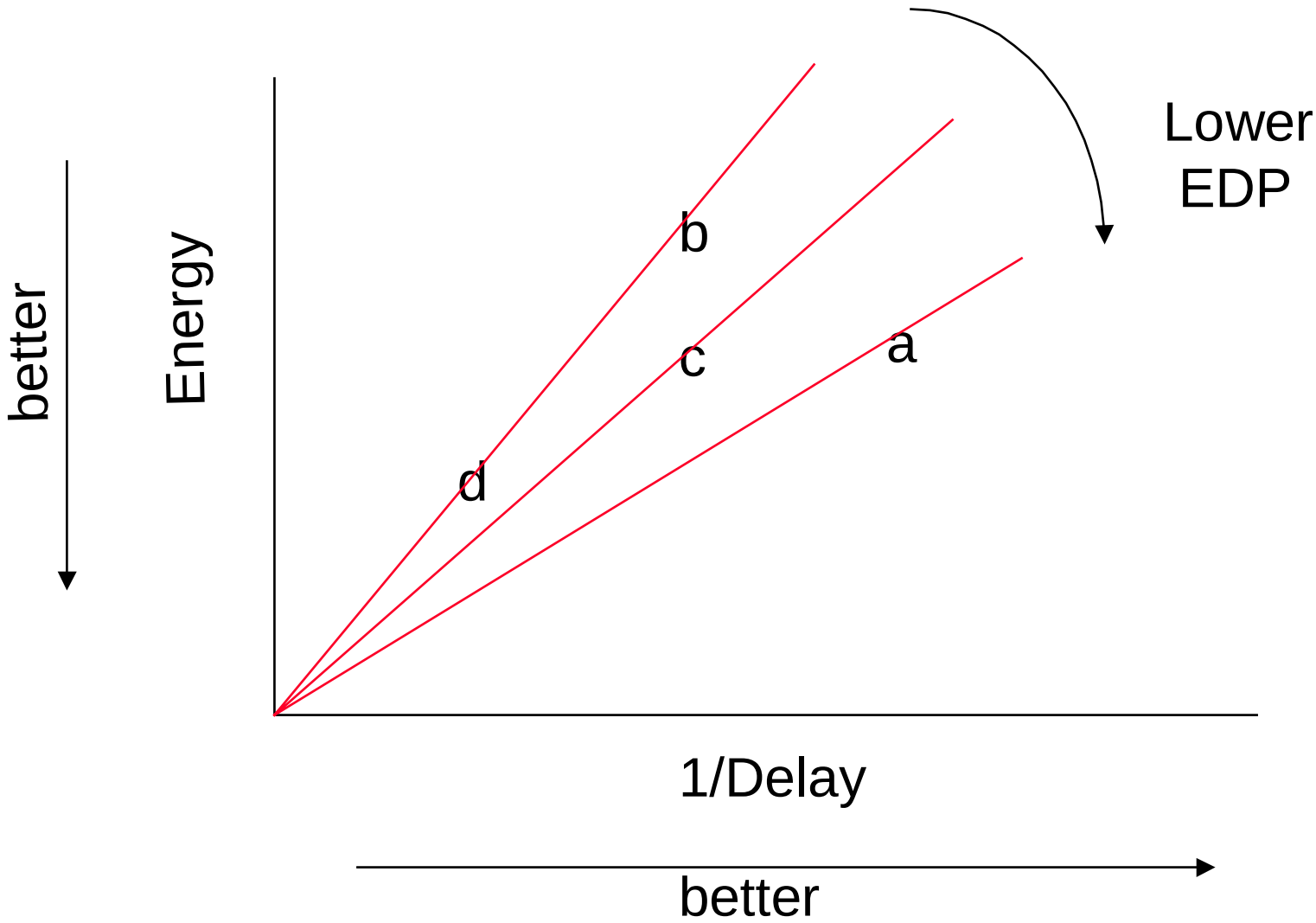
  - EDP is the average energy consumed multiplied by the computation time required
  - takes into account that one can trade increased delay for lower energy/operation (e.g., via supply voltage scaling that increases delay, but decreases energy consumption)



  - allows one to understand tradeoffs better

Understanding Tradeoffs

- Which design is the “best” (fastest, coolest, both) ?



Review: Energy & Power Equations

$$E = C_L V_{DD}^2 P_{0\to1} + t_{sc} V_{DD} I_{peak} P_{0\to1} + V_{DD} I_{leakage}$$

$$f_{0\to1} = P_{0\to1} * f_{clock}$$

$$P = C_L V_{DD}^2 f_{0\to1} + t_{sc} V_{DD} I_{peak} f_{0\to1} + V_{DD} I_{leakage}$$

Dynamic power (~90% today and decreasing relatively)

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

Leakage power (~2% today and increasing)

Principles for Power Reduction

- Prime choice: Reduce voltage!
  - Recent years have seen an acceleration in supply voltage reduction
  - Design at very low voltages still open question (0.6 … 0.9 V by 2010!)
- Reduce switching activity
- Reduce physical capacitance
  - Device Sizing: for F=20
    - f_{opt}(energy)=3.53, f_{opt}(performance)=4.47

Dynamic Power Consumption is Data Dependent

- Switching activity, $P_{0\rightarrow1}$, has two components
  - A static component – function of the logic topology
  - A dynamic component – function of the timing behavior (glitching)

2-input NOR Gate

A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

Static transition probability

$$P_{0\rightarrow1} = P_{out=0} \times P_{out=1}$$

$$= P_0 \times (1-P_0)$$

With input signal probabilities

$P_{A=1} = 1/2$

$P_{B=1} = 1/2$

NOR static transition probability

$= 3/4 \times 1/4 = 3/16$