EECS151 : Introduction to Digital Design and ICs

Lecture 14 – Wire & Energy

Bora Nikolić and Sophia Shao

AMD Athlon: First 1GHz CPU in 2000!

The Athlon's arrival signaled the opening salvos in what was coined 'The Gigahertz War'. The Pentium III had a lead role in the 'Gigahertz War' against AMD's Athlon processors between 1999 and 2000. Ultimately it was AMD who crossed the finish line first, shipping the 1GHz Athlon days before Intel could launch theirs.

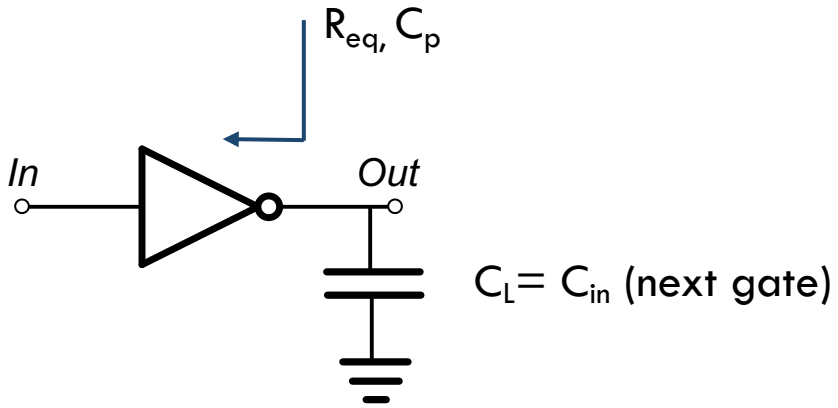
Delay Optimization

- Critical paths limit the operating speed of the system.
- Four main levels:
  - Architectural/Microarchitectural Level, e.g., # of pipeline stages
  - Logic Level, e.g., types of functional blocks
  - Circuit Level, e.g., transistor sizings
  - Layout Level, e.g., floorplanning
- Last lecture: using simple models that offer designers intuitions on logic and circuit optimizations.
  - RC delay model: transistor -> resistor + capacitor
  - Logical effort: further simplifies into a linear model

Generalizing to Arbitrary Gates

- Delay has two components: $d = gf + p$
- g: *logical effort (for same Req)*
  - Measures relative ability of gate to deliver current
  - $g = 1$ for inverter
- f: *electrical fanout* = C_{out} / C_{in}
  - Ratio of output to input capacitance
  - Sometimes called electrical effort
- p: parasitic/intrinsic delay *(for same Req)*
  - Represents delay of gate driving no load
  - Set by internal parasitic capacitance

Inverter RC Delay

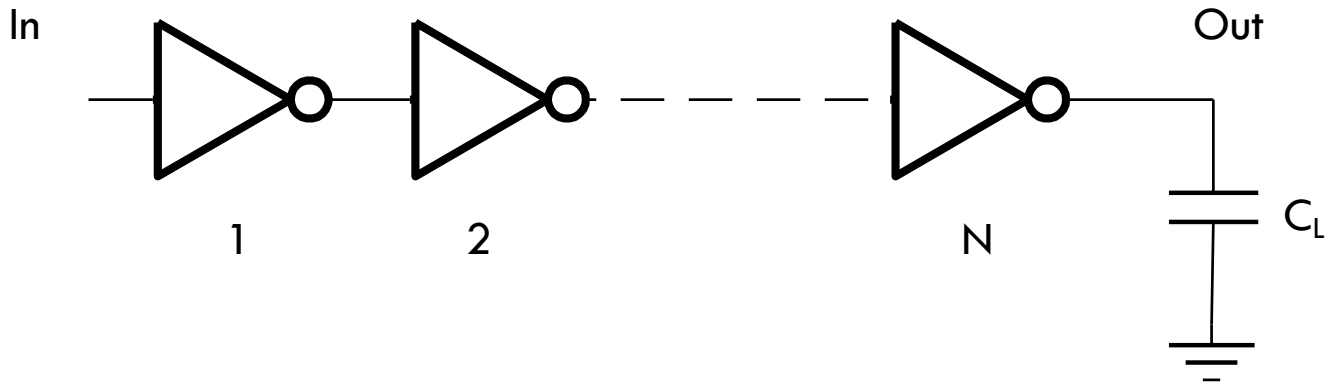


- $t_p = R_{eq}(C_p + C_L) = Req(Cin/\gamma + C_L)$
  - $\gamma = 1$ (closer to 1.2 in recent processes)
- $t_p = R_{eq}C_{in}(1+C_L/C_{in}) = \tau_{INV}(1+f)$
  - Propagation delay is proportional to fanout
- Normalized Delay = 1 + f

Fanout = $f = C_L/C_{in}$

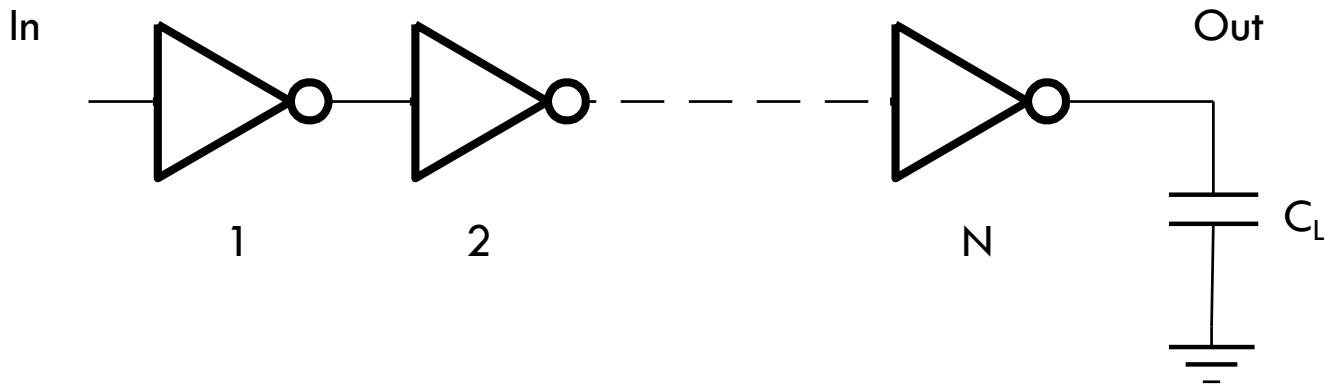
$t_p = \tau_{INV}(1+f)$

Multi-stage Logic Networks



- Logical effort generalizes to multistage networks
- *Path Logical Effort* $G = \prod g_i$
- *Path Electrical Effort* $H = \frac{C_{\text{out-path}}}{C_{\text{in-path}}}$
- *Path Effort* $F = \prod f_i = \prod g_i h_i$

Example: Minimizing the delay of an inverter chain



Delay = $t_{p1} + t_{p2} + \ldots + t_{pN}$

Delay = $(1+f_1) + (1+f_2) + \ldots + (1+f_N)$;

$f_1 = C_2/C_1$, $f2 = C_3/C_2$

f1* f2 * f3 … * fN = C_L/C_in

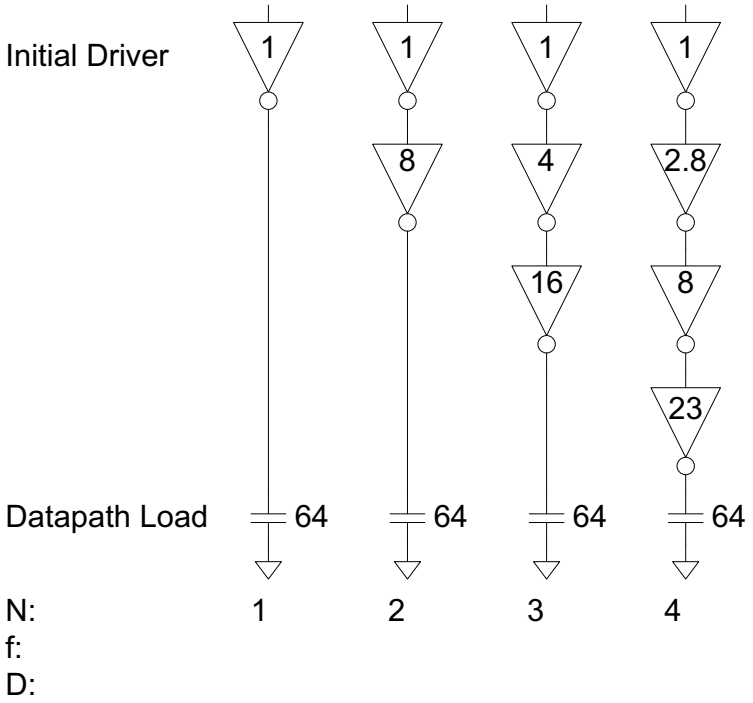
Minimum -> $f_1 = f_2 = \ldots f_N =$

Example: Best Number of Stages

- How many stages should a path use?
  - Minimizing number of stages is not always fastest
- Example: drive 64-bit datapath with unit inverter

$$D = NF^{1/N} + P$$

$$= N(64)^{1/N} + N$$

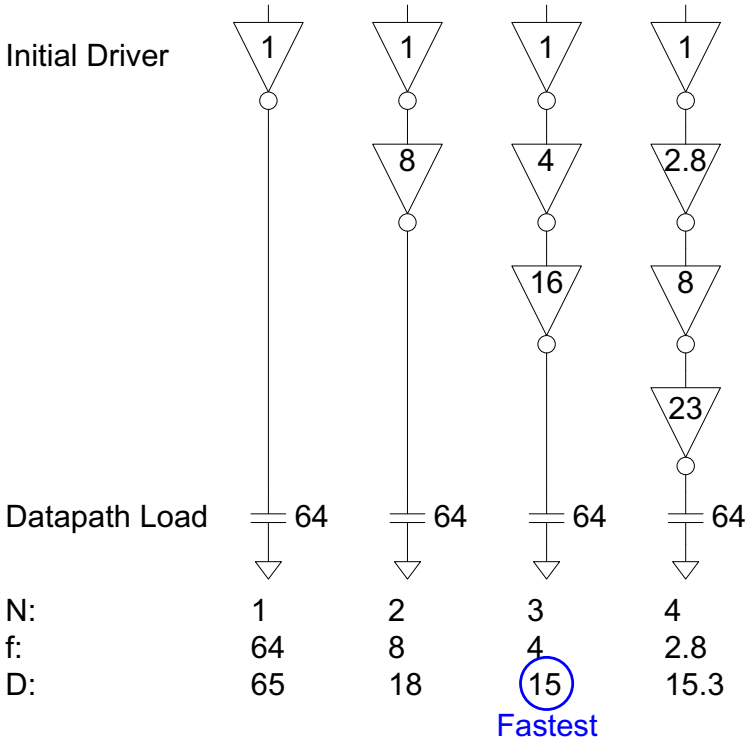


Example: Best Number of Stages

- How many stages should a path use?
  - Minimizing number of stages is not always fastest
- Example: drive 64-unit load with inverters

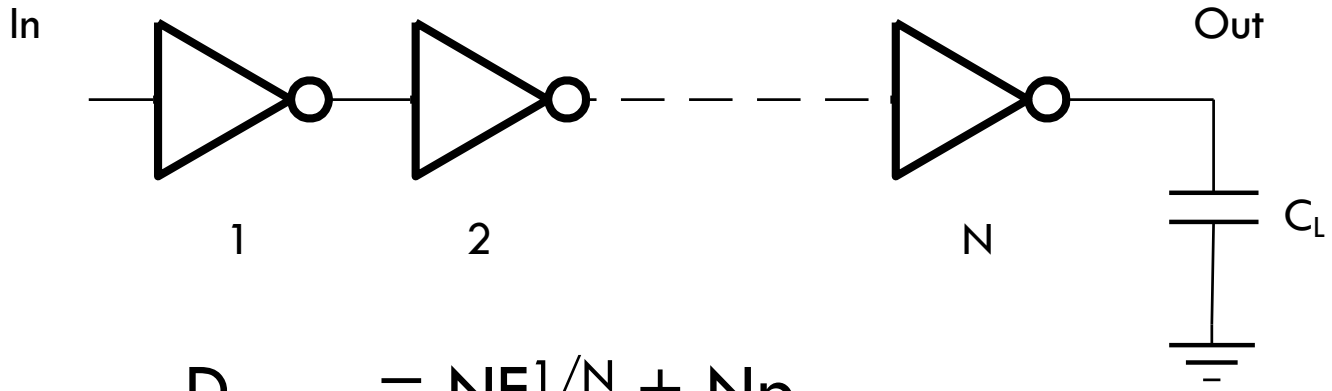
$$D = NF^{1/N} + P$$

$$= N(64)^{1/N} + N$$



Example: Best Number of Stages

- How many stages should a path use?
  - Minimizing number of stages is not always fastest



$$D = NF^{1/N} + Np_{inv}$$

- Define best stage effort $\rho = F^{\frac{1}{N}}$
- Neglecting parasitics ($p_{inv} = 0$), we find $\rho = 2.718$ (e)
- For $p_{inv} = 1$, solve numerically for $\rho = 3.59$

Logical Efforts Method

1) Compute path effort $F = GBH$

2) Estimate best number of stages $N = \log_4 F$

3) Sketch path with N stages $D = NF^{\frac{1}{N}} + P$

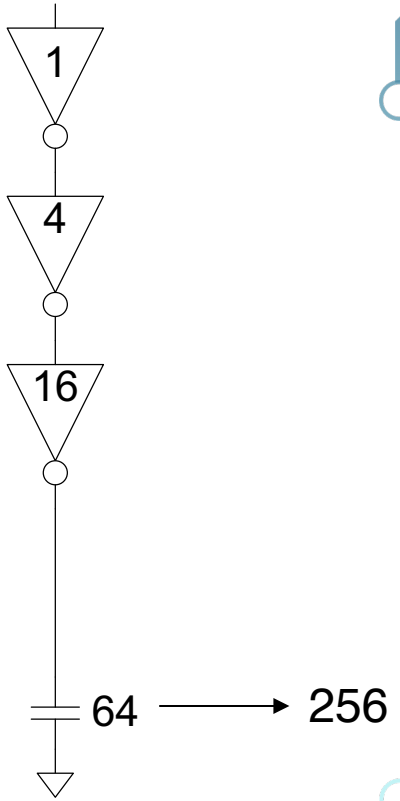
4) Estimate least delay $\hat{f} = F^{\frac{1}{N}}$

5) Determine best stage effort $C_{in_i} = \frac{g_i C_{out_i}}{\hat{f}}$

6) Find gate sizes

Quiz: Logical Effort

- If we increase the load from 64 to 256, which of the following choices will have the smallest delay?



- A. Remove the size 16 inverter.
- B. Add a size 64 inverter.
- C. Do nothing.

Summary

- Delay optimization is critical to improve the frequency of the circuit.
- The dimensions of a transistor affect its capacitance and resistance.
- We use RC delay model to describe the delay of a circuit.
- Two delay components:
  - Parasitic delay (p)
  - Effort delay (F)
    - Logical effort (g): intrinsic complexity of the gate
    - Electrical effort (f): load capacitance dependent

Wire Delay

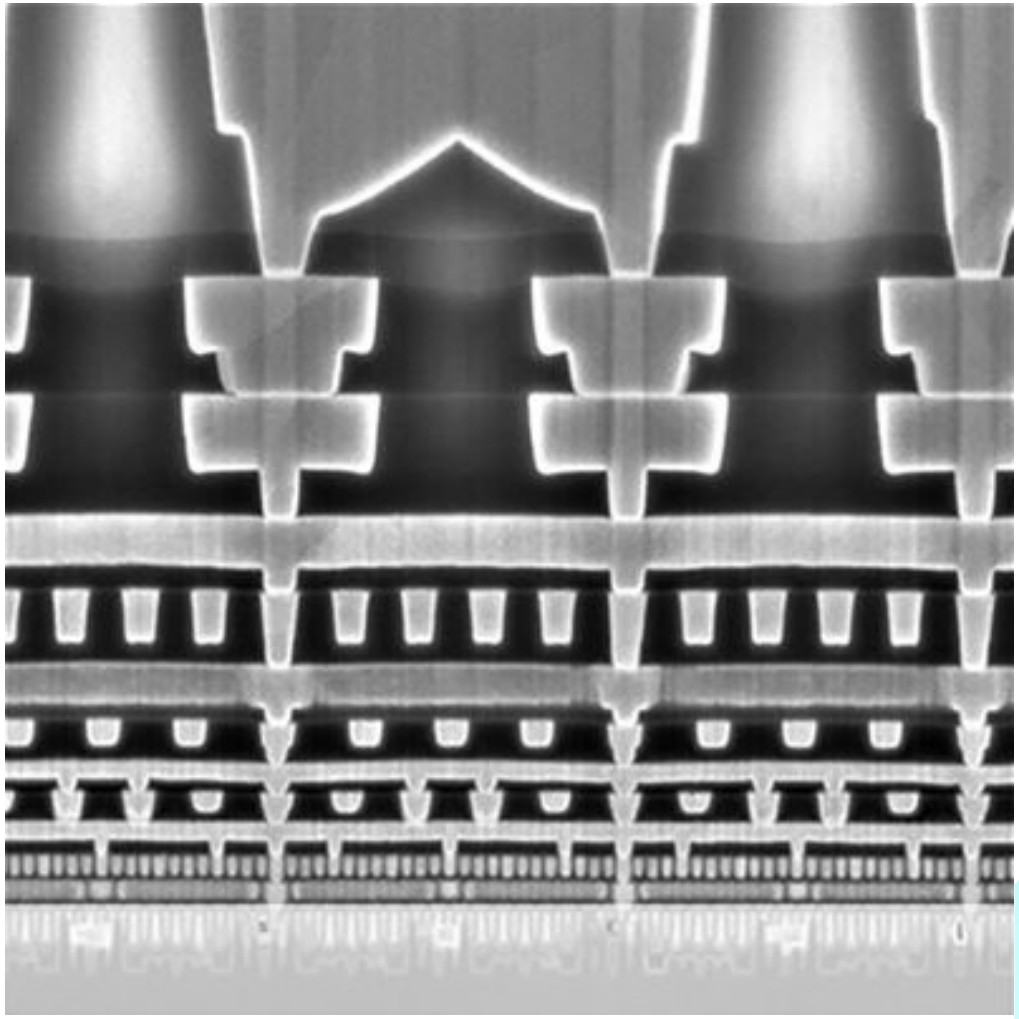
- Overview
- Wire RC Delay

Energy

- Overview
- Dynamic
- Static

A modern technology is mostly wires

- Transistors are little things under the wires
- Many layers of wires
- Wires are as important as transistors
  - Speed and power

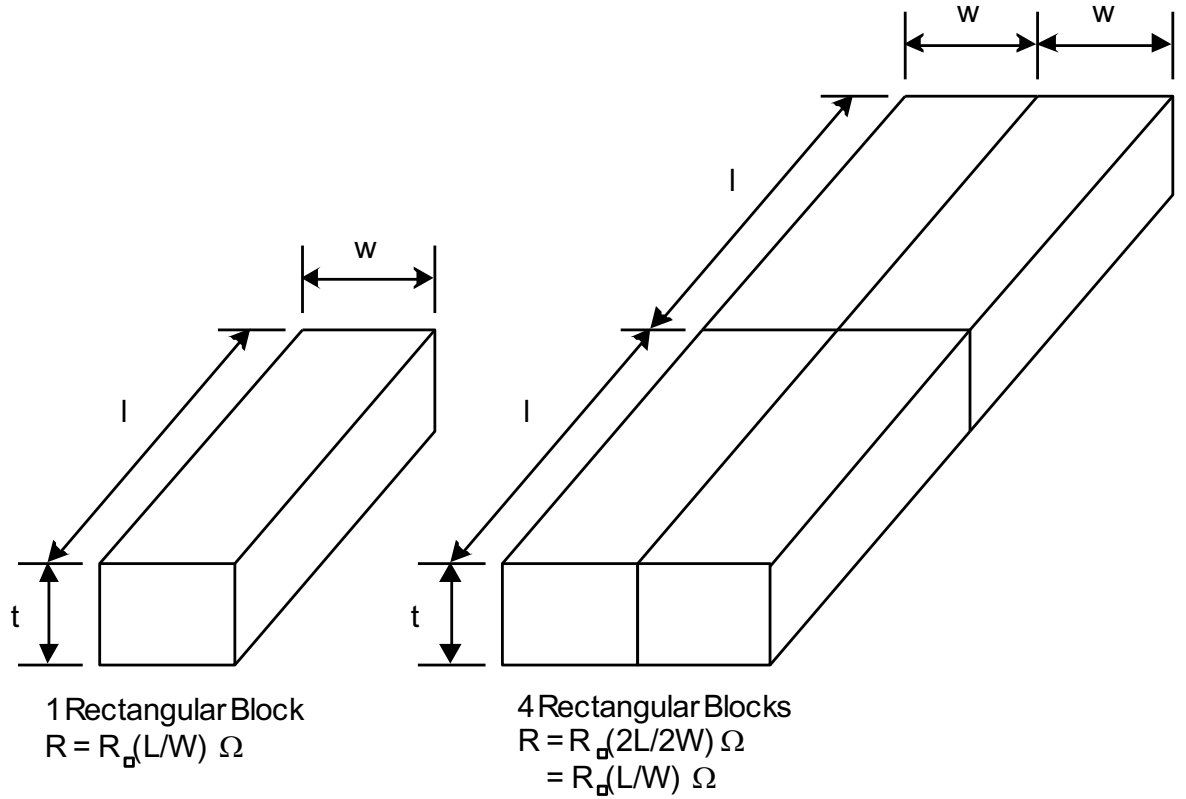


Wire Resistance

- ρ = *resistivity* (Ω*m)

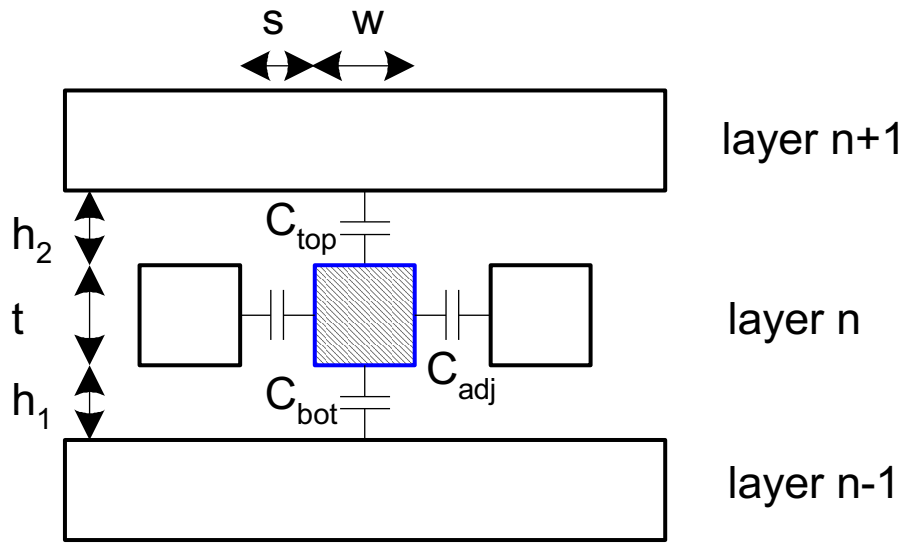
$$R = \frac{\rho}{t}\frac{l}{w} = R_{\square}\frac{l}{w}$$

- $R_{\square}$ = *sheet resistance* ($\Omega/\square$)
  - $\square$ is a dimensionless unit(!)
- Count number of squares
  - $R = R_{\square}$ * (# of squares)



Wire Capacitance

- Wire has capacitance per unit length
  - To neighbors
  - To layers above and below
- $C_{total} = C_{top} + C_{bot} + 2C_{adj}$



Wire Delay

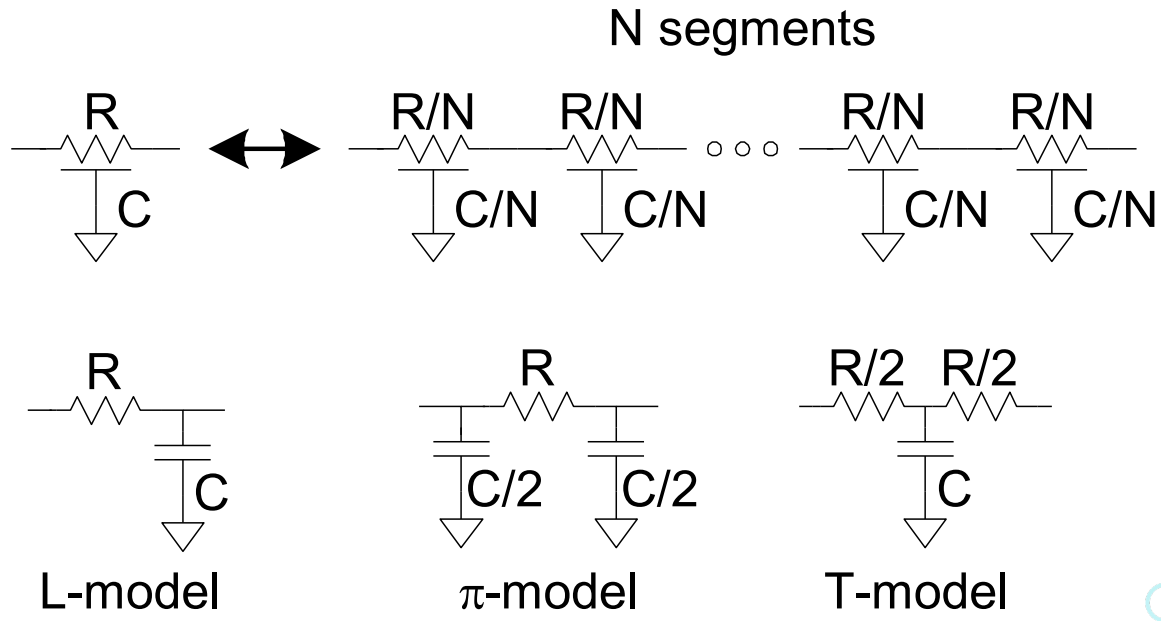
- Overview
- Wire RC Delay

Energy

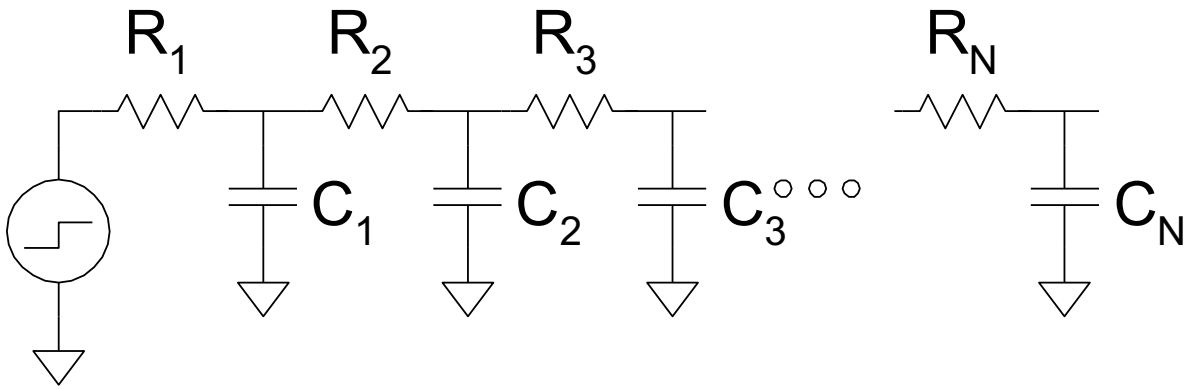
- Overview
- Dynamic
- Static

Wire RC Model

- Wires are a distributed system
  - Approximate with lumped element models
- 3-segment pi-model is accurate to 3% in simulation



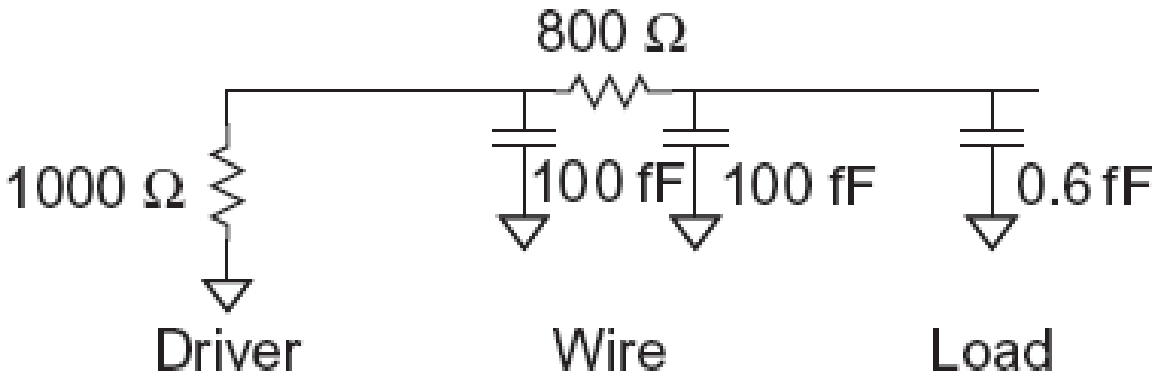
Elmore Delay for RC Tree



$$t_{pd} \approx \sum_{\text{nodes } i} R_{i-to-source} C_i$$

$$= R_1 C_1 + (R_1 + R_2) C_2 + \ldots + (R_1 + R_2 + \ldots + R_N) C_N$$

Example: RC Delay with Wire and Gate



Administrivia

- Guest lecture on verification
- Project starts this week.
- Midterm 2 in two weeks!

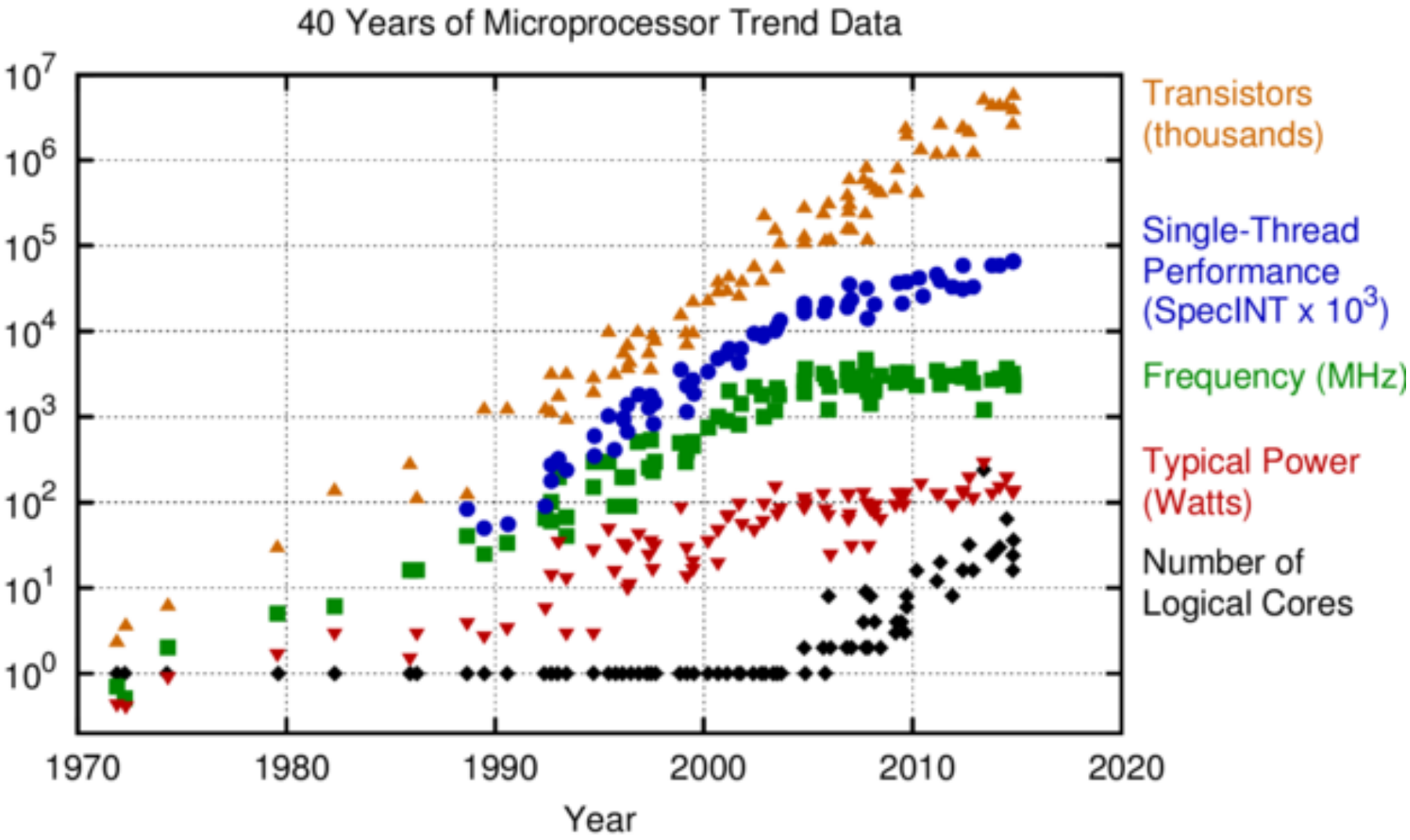
Wire Delay

- Overview
- Wire RC Delay

Energy

- Overview
- Dynamic
- Static

Processor Frequency Scaling



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

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:

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

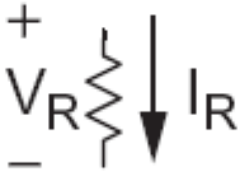
- Average Power:

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

Power in Circuit Element

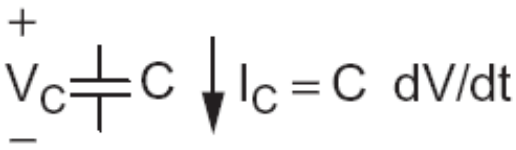
$$P_{VDD}(t) = I_{DD}(t) V_{DD}$$

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



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

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



Sources of Power Dissipation

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

Wire Delay

- Overview
- Wire RC Delay

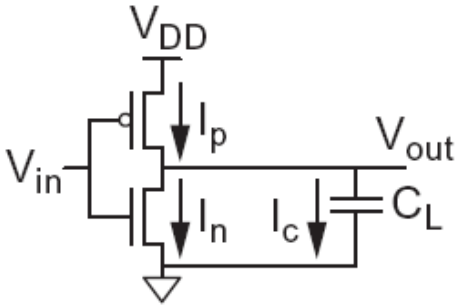
Energy

- Overview
- Dynamic
- Static

Charging and Discharging 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

Dynamic Power Reduction

How can we limit switching power?

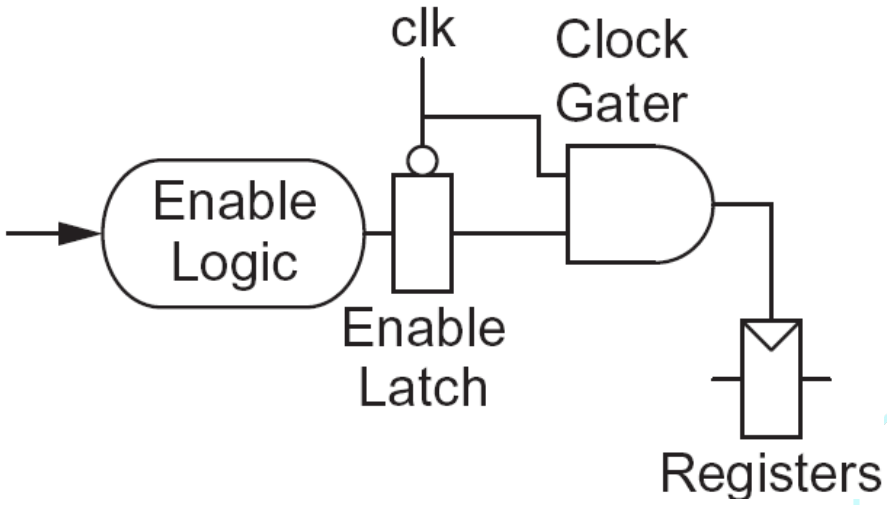
- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

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

Reduce Activity Factor

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

- Clock gating
- The best way to reduce the activity is to turn off the clock to registers in unused blocks
  - Saves clock activity (a = 1)
  - Eliminates all switching activity in the block
  - Requires determining if block will be used



Reduce Capacitance

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

- Gate capacitance
  - Fewer stages of logic
  - Smaller gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other

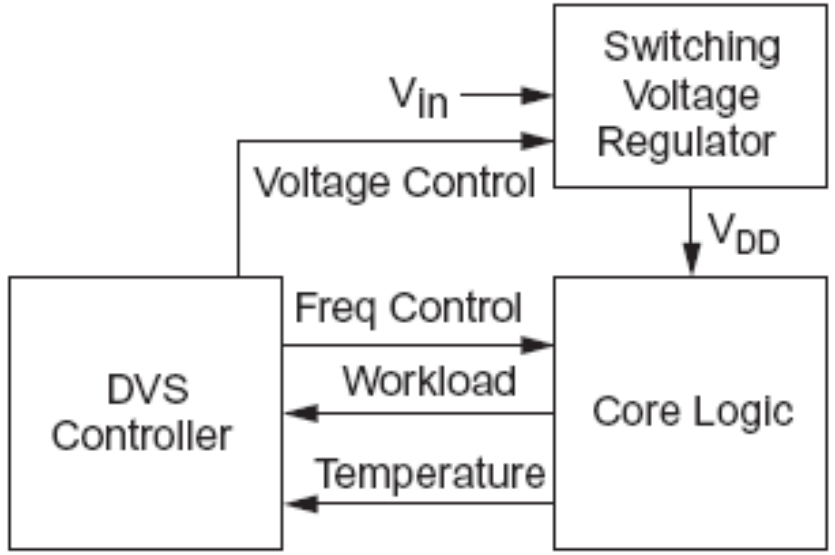
Reduce Voltage/Frequency

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

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

- Voltage Domains
  - Provide separate supplies to different blocks

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



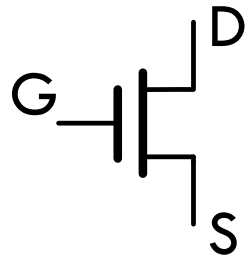
Wire Delay

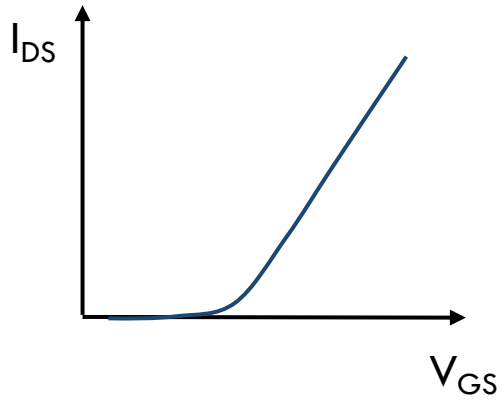
- Overview
- Wire RC Delay

Energy

- Overview
- Dynamic
- Static

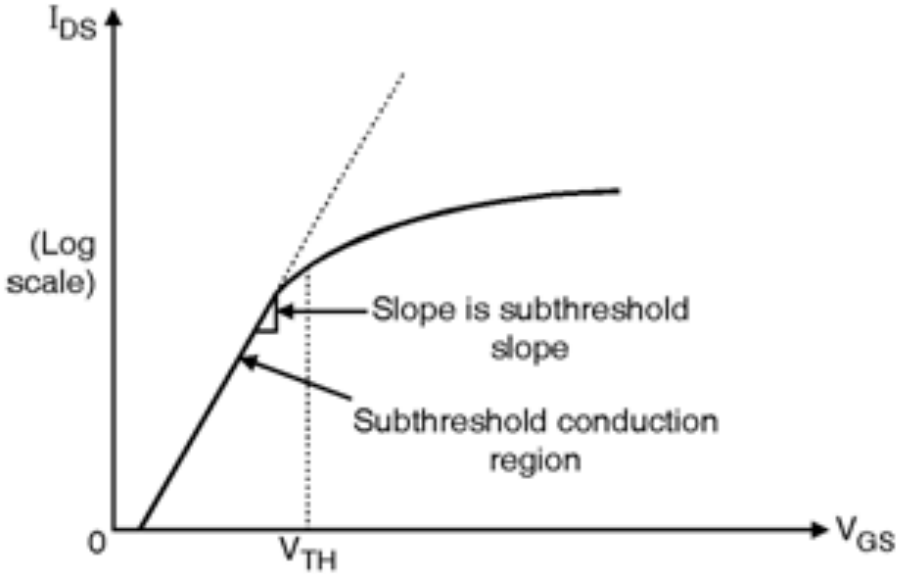
Subthreshold Leakage





Nearly linear

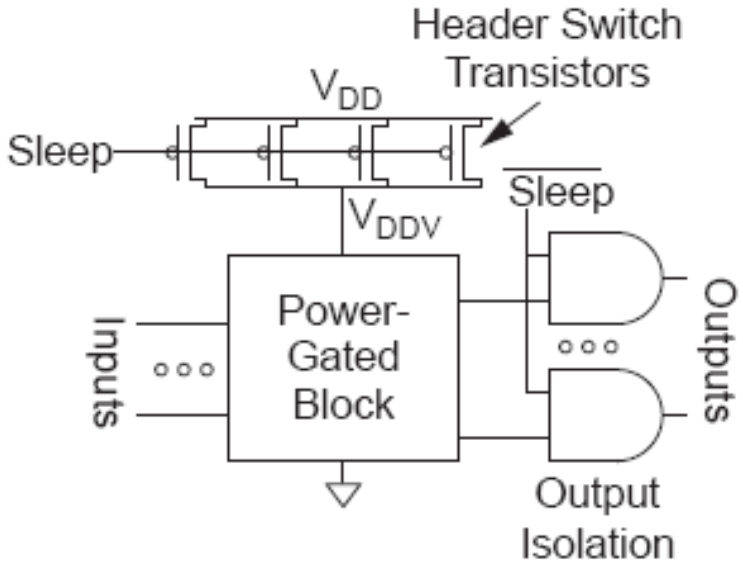
$I_{DS} \sim K(V_{GS}-V_{Th})$



I_{DS} Vs V_{GS} characteristics in log scale

Power Gating

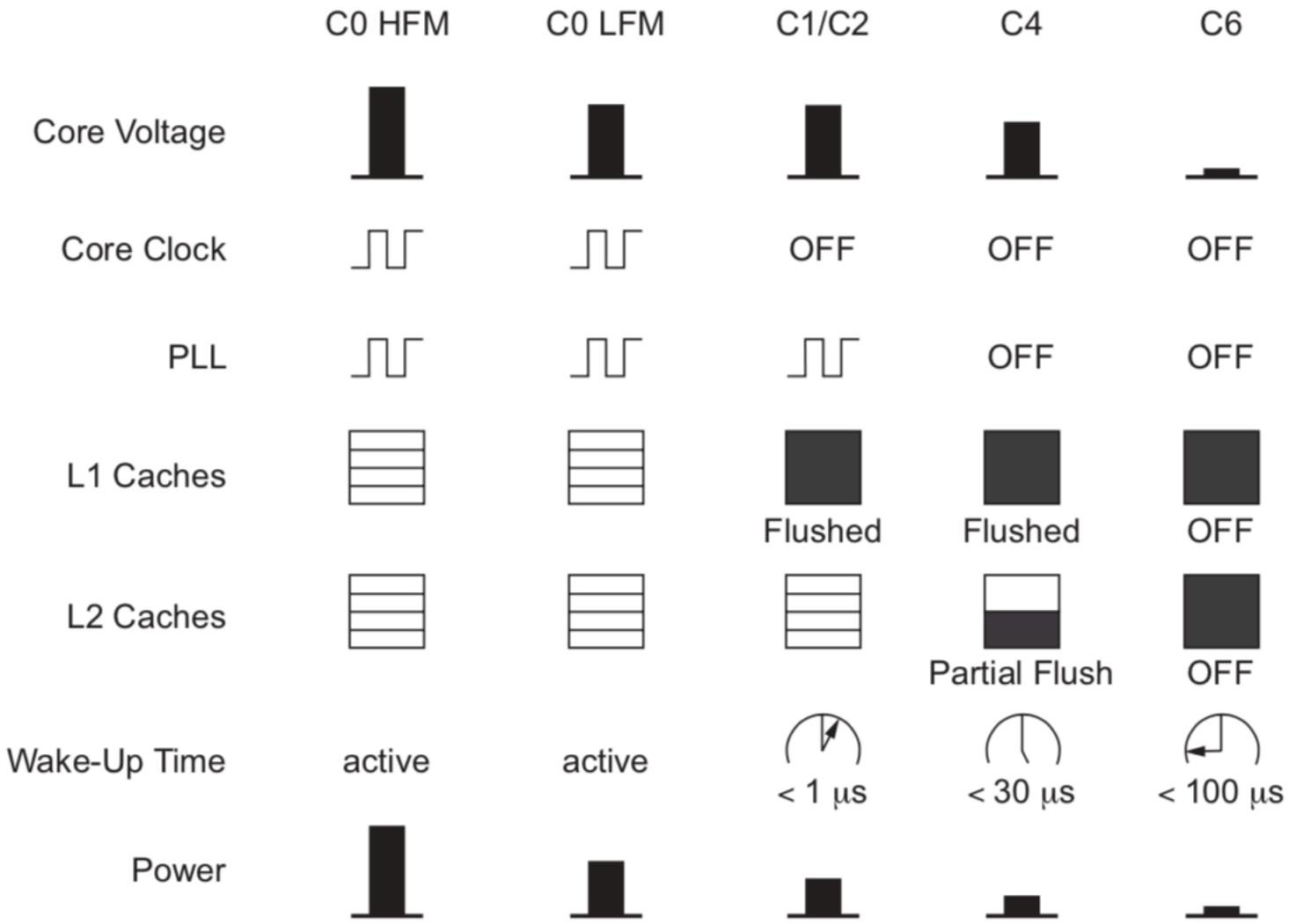
- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual V_{DD} (V_{DDV})
  - Gate outputs to prevent invalid logic levels to next block



- 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

Example: Power Management

- Power states



Summary

- Wire also contributes to delay, especially in modern technology.
- We can use RC model to capture wire delay as well.
- Energy becomes an increasingly important optimization goal.
  - Dynamic Energy
  - Static Energy