Lecture 02: Performance and Power Topics
Logistics

Course logistics
- Recitation room: APM 2301 Wednesday noon – 12:50pm
- Datapaths, memories (Lecture 2) moved into Recitation 2
- More time for Lab 1++ (more Verilog exercises), and Verilog coding for performance moved to Recitation 3

Comments
- The material is self-contained (lecture + book). The prerequisites are (1) familiarity with logic design (UG level), (2) willingness to trace pointers, and (3) ability to identify some basic physical relationships (\( Q = CV \), \( V = IR \), etc.) in the material presented.
- This course serves several (CE) goals: replaces part of the ECE 260 sequence; gives “what you need to know about devices, interconnects, blocks, design” for CSE CE students; gives first exposure to ASIC design process.

Reading:
- Smith Chapter 1: Introduction to ASICs (types of ASICs, design flow, economics of ASICs, cell libraries)
- Smith Chapter 2: CMOS Logic (transistors, process, design rules, combinational logic cells, sequential logic cells, datapath logic cells, I/O cells)
- Smith Chapter 3.1, 3.2: Transistor parasitics, slew times
- Smith Chapter 11: Verilog
- Interconnect performance analysis (look for readings)
- References mentioned last time: Weste/Eshragian, Rabaey, Bakoglu
Outline

- Interconnects
- Resistance
- Capacitance and Inductance
- Delay
- Power
Circuit Performance Estimation

Critical Path Timing Analysis

Deep Sub-micron (DSM) MOSFET models

Accurate interconnect delay and noise models

Slide courtesy of Kevin Cao, Berkeley
What are some implications of reverse-scaled global interconnects?
Intel 130nm BEOL Stack

Intel 6LM 130nm process with vias shown (connecting layers)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Pitch (nm)</th>
<th>Thick (nm)</th>
<th>Aspect Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolation</td>
<td>364</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Poly-silicon</td>
<td>336</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>Metal 1</td>
<td>350</td>
<td>280</td>
<td>1.6</td>
</tr>
<tr>
<td>Metal 2, 3</td>
<td>448</td>
<td>360</td>
<td>1.6</td>
</tr>
<tr>
<td>Metal 4</td>
<td>756</td>
<td>570</td>
<td>1.5</td>
</tr>
<tr>
<td>Metal 5</td>
<td>1120</td>
<td>900</td>
<td>1.6</td>
</tr>
<tr>
<td>Metal 6</td>
<td>1204</td>
<td>1200</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Aspect ratio = thickness / minimum width
Damascene and Dual-Damascene Process

- Damascene process named after the ancient Middle Eastern technique for inlaying metal in ceramic or wood for decoration

- **Single Damascene**
  - IMD DEP
  - Oxide Trench Etch
  - Metal Fill
  - Metal CMP

- **Dual Damascene**
  - Oxide Trench / Via Etch
  - Metal Fill
  - Metal CMP
Cu Dual-Damascene Process

- Polishing pad touches both up and down area after step height
- Different polish rates on different materials
- Dishing and erosion arise from different polish rates for copper and oxide

Cu Damascene Process

- Bulk copper removal
- Barrier removal
- Oxide over-polish
- Oxide erosion
- Copper dishing
**Area Fill & Metal Slot for Copper CMP**

- **Dishing** can thin the wire or pad, causing higher-resistance wires or lower-reliability bond pads

- **Erosion** can also result in a sub-planar dip on the wafer surface, causing short-circuits between adjacent wires on next layer

- **Oxide erosion and copper dishing** can be controlled by area filling and metal slotting
Evolution of Interconnect Modeling Needs

- Before 1990, wires were thick and wide while devices were big and slow
  - Large wiring capacitances and device resistances
  - Wiring resistance << device resistance
  - Model wires as capacitances only

- In the 1990s, scaling (by scale factor $S$) led to smaller and faster devices and smaller, more resistive wires
  - *Reverse scaling* of properties of wires
  - RC models became necessary

- In the 2000s, frequencies are high enough that inductance has become a major component of total impedance
Global Interconnect Delay

![Graph showing the relationship between delay and minimum feature size for gate and interconnect delay. The graph indicates that gate delay decreases as the minimum feature size decreases, while interconnect delay increases.](image-url)
Interconnect Statistics

- What are some implications?

\[ S_{\text{Local}} = S_{\text{Technology}} \]
\[ S_{\text{Global}} = S_{\text{Die}} \]
Outline

- Interconnects
- Capacitance and Inductance
- Resistance
- Delay
- Power
ILD = interlevel dielectric

\[ C_{int} = \frac{\varepsilon_{ox} W}{t_{ox} L} \]

\[ S_{C,\text{wire}} = \frac{S \times S_L}{S} = S_L \]
Insulator Permittivities

<table>
<thead>
<tr>
<th>Material</th>
<th>$\varepsilon_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free space</td>
<td>1</td>
</tr>
<tr>
<td>Aerogels</td>
<td>~1.5</td>
</tr>
<tr>
<td>Polyimides (organic)</td>
<td>3-4</td>
</tr>
<tr>
<td>Silicon dioxide</td>
<td>3.9</td>
</tr>
<tr>
<td>Glass-epoxy (PC board)</td>
<td>5</td>
</tr>
<tr>
<td>Silicon Nitride ($\text{Si}_3\text{N}_4$)</td>
<td>7.5</td>
</tr>
<tr>
<td>Alumina (package)</td>
<td>9.5</td>
</tr>
<tr>
<td>Silicon</td>
<td>11.7</td>
</tr>
</tbody>
</table>

- Huge effort to develop low-k dielectrics ($\varepsilon_r < 4.0$) for metal
- Reduces capacitance → helps delay and power
- Materials have been identified, but process integration has been difficult at best
Line Dimensions and Fringing Capacitance

- Line dimensions: $W$, $S$, $T$, $H$
- Sometimes $H$ is called $T$ in the literature, which can be confusing
Capacitance Values for Different Configurations

- Parallel-plate model substantially underestimates capacitance as line width drops below order of ILD height

  Why?
Interwire (Coupling) Capacitance

- Leads to coupling effects among neighboring wires
Interwire Capacitance

<table>
<thead>
<tr>
<th>Layer</th>
<th>Poly</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacitance (aF/um) at minimum spacing</td>
<td>40</td>
<td>95</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>115</td>
</tr>
</tbody>
</table>

- **Example:** Two M3 lines run parallel to each other for 1mm. The capacitance between them is $85\text{aF/um} \times 1000\text{um} = 85000\text{aF} = 85\text{fF}$

- **Interwire capacitance today reaches ~80% of total wire capacitance**
Capacitance Estimation

- Empirical capacitance models are easiest and fastest
  - Handle limited configurations (e.g., range of T/H ratio)
  - Some limiting assumptions (e.g., no neighboring wires)

\[
C_{\text{wire}} = \varepsilon_{\text{ox}} \left[ \left( \frac{W}{H_{\text{ILD}}} \right) + 0.77 + 1.06 \left( \frac{W}{H_{\text{ILD}}} \right)^{0.25} + 1.06 \left( \frac{T_{\text{wire}}}{H_{\text{ILD}}} \right)^{0.5} \right]
\]

Capacitance per unit length

- Rules of thumb: e.g., 0.2 fF/um for most wire widths < 2um
  - Cf. MOSFET gate capacitance ~ 1 fF/um width

- Pattern-matching approaches
Capacitive Crosstalk Noise

- Two coupled lines

- Cross-section view

- Interwire capacitance allows neighboring wires to interact

- Charge injected across $C_c$ results in temporary (in static logic) glitch in voltage from the supply rail at the victim
Crosstalk From Capacitive Coupling

- Glitches caused by capacitive coupling between wires
  - An “aggressor” wire switches
  - A “victim” wire is charged or discharged by the coupling capacitance (cf. charge-sharing analysis)

- An otherwise quiet victim may look like it has temporarily switched

- This is bad if:
  - The victim is a clock or asynchronous reset
  - The victim is a signal whose value is being latched at that moment

- What are some fixes?

Slide courtesy of Paul Rodman, ReShape
Crosstalk: Timing Pull-In

- A switching victim is aided (sped up) by coupled charge
- This is bad if your path now violates hold time
- Fixes include adding delay elements to your path
Crosstalk: Timing Push-Out

- A switching victim is hindered (slowed down) by coupled charge
- This is bad if your path now violates setup time
- Fixes include spacing the wires, using strong drivers, …
Relatively greater coupling noise due to line dimension scaling

Tighter timing budgets to achieve fast circuit speed ("all paths critical")

→ Train wreck?

Timing analysis can be guardbanded by scaling the coupling capacitance by a "Miller Coupling Factor" to account for push-in or push-out. **Homework Q3:** (a) explain upper and lower bounds on the Miller Coupling Factor for a victim wire that is between two parallel aggressor wires, assuming step transitions; (b) give an estimate of the ratio (Delay Uncertainty / Nominal Delay) in the 90nm and 65nm technology nodes.

---

*Slide courtesy of Kevin Cao, Berkeley*
Inductance

- Inductance, $L$, is the flux induced by current variation.
- Measures ability to store energy in the form of a magnetic field.
- Consists of self-inductance and mutual inductance terms.
- At high frequencies, can be a significant portion of total impedance:
  $$Z = R + j\omega L \quad (\omega = 2\pi f = \text{angular freq})$$

**Self Inductance**

$$\Phi_{11} = \int_{s_1} B_1 \cdot ds_1$$

**Mutual Inductance**

$$\Phi_{12} = \int_{s_2} B_1 \cdot ds_2$$

**Self Inductance**

$$\text{Self Inductance} = \frac{\Phi_{11}}{I}$$

**Mutual Inductance**

$$\text{Mutual Inductance} = \frac{\Phi_{12}}{I}$$
Inductance

- When signal is coupled to a ground plane, the current loop has an inductance.
  - More apparent for upper layer metals and longer lines
  - Simple lumped model:

- Gives interconnect transmission-line qualities
  - Propagates signal energy, with delay; sharper rise times; ringing

- Magnetic flux couples to many signals \(\rightarrow\) computational challenge
  - Not just coupled to immediately adjacent signals (unlike capacitors)
  - Coupling over a larger distance
  - Bigger lumped model: matrix of coupling coefficients not sparse
Inductance is Important…

- If $\omega L \approx R$ where $\omega = 2\pi f = 2\pi \left( \frac{1}{\pi t_r} \right)$
- Copper interconnects $\rightarrow$ R is reduced
- Faster clock speeds
- Thick, low-resistance (reverse-scaled) global lines
- Chips are getting larger $\rightarrow$ long lines $\rightarrow$ large current loops
- Frequency of interest is determined by signal *rise time*, not clock frequency
On-Chip Inductance

- Inductance is a loop quantity
- Knowledge of return path is required, but hard to determine
- For example, the return path depends on the frequency
**Frequency-Dependent Return Path**

- At low frequency, \((R \gg \omega L)\) and current tries to
  - minimize impedance
  - minimize resistance
  - use as many returns as possible (parallel resistances)

\[
\begin{align*}
\text{Gnd} & & \text{Gnd} & & \text{Gnd} & & \text{Signal} & & \text{Gnd} & & \text{Gnd} & & \text{Gnd} & & \text{Gnd} \\
\end{align*}
\]

- At high frequency, \((R \ll \omega L)\) and current tries to
  - minimize impedance
  - minimize inductance
  - use smallest possible loop (closest return path) \(\Rightarrow L\) dominates, current returns “collapse”
  - Power and ground lines always available as low-impedance current returns

\[
\begin{align*}
\text{Gnd} & & \text{Gnd} & & \text{Gnd} & & \text{Signal} & & \text{Gnd} & & \text{Gnd} & & \text{Gnd} & & \text{Gnd} \\
\end{align*}
\]
Inductance Trends

- Inductance = weak (log) function of conductor dimensions

- Inductance = strong function of distance to current return path (e.g., power grid)
  - Want nearby ground line to provide a small current loop (cf. Alpha 21164)

- Inductance most significant in long, low-R, fast-switching nets
  - Clocks are most susceptible
Inductance vs. Capacitance

- **Capacitance**
  - Locality problem is easy: electric field lines “suck up” to nearest neighbor conductors
  - Local calculation is hard: all the effort is in “accuracy”

- **Inductance**
  - Locality problem is hard: magnetic field lines are not local; current returns can be complex
  - Local calculation is easy: no strong geometry dependence; analytic formulae work very well

- **Intuitions for design**
  - Seesaw effect between inductance and capacitance
  - Minimize *variations* in L and C rather than absolutes
    - E.g., would techniques used to minimize variation in capacitive coupling also benefit inductive coupling?

- **Homework Q4:** Conceive and describe as many ways as you can for managing (controlling) effects of both interconnect inductance as well as capacitance coupling. Some “hint” keywords: shield, split, space, slew, size, ...
Outline

- Interconnects
- Capacitance and Inductance
- Resistance
- Delay
- Power
Resistance & Sheet Resistance

- Resistance seen by current going from left to right is same in each block

\[ R = \frac{\rho L}{T W} \]

Sheet Resistance \( R_{\square} \)

\[ R_1 \equiv R_2 \]
Bulk Resistivity

- Aluminum dominant until ~2000
- Copper has taken over in past 4-5 years
- Copper as good as it gets

<table>
<thead>
<tr>
<th>Material</th>
<th>$\rho$ (Ω-m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver (Ag)</td>
<td>$1.6 \times 10^{-8}$</td>
</tr>
<tr>
<td>Copper (Cu)</td>
<td>$1.7 \times 10^{-8}$</td>
</tr>
<tr>
<td>Gold (Au)</td>
<td>$2.2 \times 10^{-8}$</td>
</tr>
<tr>
<td>Aluminum (Al)</td>
<td>$2.7 \times 10^{-8}$</td>
</tr>
<tr>
<td>Tungsten (W)</td>
<td>$5.5 \times 10^{-8}$</td>
</tr>
</tbody>
</table>
Interconnect Resistance

<table>
<thead>
<tr>
<th>Material</th>
<th>Sheet Resistance ($\Omega/\square$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n- or p-well diffusion</td>
<td>1000 – 1500</td>
</tr>
<tr>
<td>$n^+, p^+$ diffusion</td>
<td>50 – 150</td>
</tr>
<tr>
<td>$n^+, p^+$ diffusion with silicide</td>
<td>3 – 5</td>
</tr>
<tr>
<td>$n^+, p^+$ polysilicon</td>
<td>150 – 200</td>
</tr>
<tr>
<td>$n^+, p^+$ polysilicon with silicide</td>
<td>4 – 5</td>
</tr>
<tr>
<td>Aluminum</td>
<td>0.05 – 0.1</td>
</tr>
</tbody>
</table>

- Resistance scales badly
  - True scaling would reduce width and thickness by $S$ each node
  - $R \sim S^2$ for a fixed line length and material
- **Reverse scaling** $\rightarrow$ wires get smaller and slower, devices get smaller and faster
- At higher frequencies, current *crowds* to edges of conductor (thickness of conduction = *skin depth*) $\rightarrow$ increased $R$
Copper Resistivity: The Real Story

Conductor resistivity increases expected to appear around 100 nm linewidth - will impact intermediate wiring first - ~ 2006

Courtesy of SEMATECH
Outline

- Interconnects
- Capacitance and Inductance
- Resistance
- Delay
- Power
Gate Delay

- Gate delay is a measure of an input transition to an output transition.
  - May have different delays for different input to output paths.

- Different for an upward or downward transition.
  - $t_{p_{LH}}$ – propagation delay from LOW-to-HIGH (of the output)

- A transition is defined as the time at which a signal crosses a logical threshold voltage, $V_{THL}$.
  - Digital Abstraction for 1 and 0
  - Often use $V_{DD}/2$. 
Static CMOS Gate Delay

- Output of a gate drives the inputs to other gates (and wires).
  - Only pull-up or pull-down, not both.
  - Capacitive loads.

- Delay is due to the charging and discharging of a capacitor and the length of time it takes.

- The delay of EACH is treated as separately calculable

\[ t_{PD} = t_{PD1} + t_{PD2} \]
RC Model

- We can model a transistor with a resistor
  - (Take into account the different regions of operation?)
  - (Use a realistic transition time to model an input switching?)

- We can take the average capacitance of a transistor as well

- The easy model (one we will primarily use):
  - Delay = $R_{DRV}C_{LOAD}$ (the time constant)
  - $R$ proportional to $L/W$
    - Wider device (stronger drive)
    - Smaller $R_{DRV}$ shorter delay.
Another common expression for delay is $C\Delta V/I$.

- Based on the capacitance charging and discharging
- $\Delta V$ is the voltage to the transition ($V_{DD}/2$)

Very similar model except we are breaking $R$ into 2 components, $V/I$

- $I$ = average drive current

This helps understand what determines $R$

- $I$ is proportional to mobility and $W/L$
- $I$ is proportional to $V^2$ ($V$ is proportional to $V_{DD}$)
- For example, we can anticipate what might happen if $V_{DD}$ drops.
Interconnect: Distributing the Capacitance

- The resistance and capacitance of an interconnect is distributed.

- Model by using R and C.
  - Π Model is the best
  - Distributed model uses N segments.
    - More accurate but computationally expensive
    - Number of nodes blows up.
  - Lump model uses 1 segment of Π.
    - Sufficient for most nets (point to point)

Distributed using multiple lumps of Π model of a single wire
Step response of a distributed RC wire as function of location along wire and time
RC Line Models and Step Response

\[
T_{th} = \ln \left( \frac{1}{1 - Th} \right) \times T_{ED} \quad (\text{e.g., } T_{0.9} = 2.3 \times T_{ED}; \ T_{0.632} = T_{ED})
\]
Elmore Delay

- Defined by Elmore (1948) as first moment of impulse response
- $H(t) = \text{step input response}$
- $h(t) = \text{impulse response} = \text{rate of change of step response}$
- $T_{50\%} = \text{median of } h(t)$
- $T_{ED} = \text{approximation of median of } h(t) \text{ by mean of } h(t)$
  
  \[ T_{ED} = \int_0^\infty tv'_{out}(t)dt = \int_0^\infty th(t)dt \]
  
  - Works for monotonic waveforms
  - Is an overestimate of actual delay
  - Works well with symmetric impulse response (e.g., gate transition)
Elmore Delay for RC Network

\[ T_{ED}(k) = r_d C_{s_0} + \sum_{\forall i \in MP(s_0, s_k)} r_{e_i} \left( \frac{c_{e_i}}{2} + C_i \right) \]

- \(MP(s_0, s_k)\) is the main path between source \(s_0\) and sink \(s_k\)
- \(C_i\) is the capacitance of the (sub)tree rooted at node \(i\)
- \(e_i\) is the unique parent edge of node \(i\) when the tree is rooted at the source
- \(r_d\) is the driver on-resistance at the source \(s_0\)
- \(r_{e_i}\) and \(c_{e_i}\) respectively denote the lumped resistance and capacitance of the edge \(e_i\).

Homework Q5: (a) Write down the Elmore delay from node In to node O2 in Example A. (b) How efficiently can Elmore source-sink delay at all sinks in a given RC tree be evaluated? Explain the efficient (okay: linear-time) method of evaluation.
Driving Large Capacitances

$$t_{pHL} = \frac{C_L \cdot V_{swing}}{2 \cdot I_{av}}$$

Transistor Sizing
Driving Large Capacitances: Inverter As Buffer

- Total propagation delay = $t_p(\text{inv}) + t_p(\text{buffer})$

- $t_{p0}$ = delay of min-size inverter with single min-size inverter as fanout load

- Minimize $t_p = U \cdot t_{p0} + \frac{X}{U} \cdot t_{p0}$
  - $U_{opt} = \sqrt{X}$ ; $t_{p, opt} = 2 \cdot t_{p0} \cdot \sqrt{X}$

- Use only if combined delay is less than unbuffered case
Delay Reduction With Cascaded Buffers

- Cascade of buffers with increasing sizes (U = tapering factor) can reduce delay.
- If load is driven by a large transistor (which is driven by a smaller transistor) then its turn-on time dominates overall delay.
- Each buffer charges the input capacitance of the next buffer in the chain and speeds up charging, reducing total delay.
- Cascaded buffers are useful when $R_{\text{int}} < R_{\text{tr}}$. 

Slide courtesy of Mary Jane Irwin, PSU
Total line delay as function of driver size, load capacitance

Homework Q6: Derive the optimum (min-delay) value of U.
Reducing RC Delay With Repeaters

- RC delay is quadratic in length → must reduce length
  - $T_{50} = 0.4 \times R_{\text{int}} \times C_{\text{int}} + 0.7 \times (R_{\text{tr}} \times C_{\text{int}} + R_{\text{tr}} \times C_{L} + R_{\text{int}} \times C_{L})$

- Observation: $2^2 = 4$ and $1+1 = 2$ but $1^2 + 1^2 = 2$

- Repeater = strong driver (usually inverter or pair of inverters for non-inversion) that is placed along a long RC line to “break up” the line and reduce delay
Optimum Number and Size of Repeaters

- Let \( h = \) optimal buffer size (represent driver as an intermediate buffer)

- Then \( R_{tr} = R_0/h \) and \( C_L = hC_0 \)

- Propagation Delay: \( T_{pd} = \frac{0.4RC}{N} + 0.7\left(\frac{R_0C}{h} + RC_0h + R_0C_0N\right) \)

- Compute \( N_{opt} \) and \( h_{opt} \) by setting \( \frac{\partial T_{pd}}{\partial N} = 0 \) and \( \frac{\partial T_{pd}}{\partial h} = 0 \)

- Optimal buffer number: \( N_{opt} = \sqrt{\frac{(0.4RC)}{0.7R_0C_0}} \)

- Optimal buffer size: \( h_{opt} = \sqrt{\frac{R_0C}{RC_0}} \)
Repeaters vs. Cascaded Buffers

- Repeaters are used to drive long RC lines
  - Breaking up the quadratic dependence of delay on line length is the goal
  - Typically sized identically

- Cascaded buffers are used to drive large capacitive loads, where there is no parasitic resistance
  - We put all buffers at the beginning of the load
  - This would be pointless for a long RC wire since the wire RC delay would be unaffected and would dominate the total delay
Outline

- Interconnects
- Capacitance and Inductance
- Resistance
- Delay
- Power
Power Dissipation

Lead Microprocessor’s power continues to increase

Power delivery and dissipation will be prohibitive(?)

Power (Watts)

Year


Power delivery and dissipation will be prohibitive(?)

Courtesy, Intel
Power Density

Power density too high to keep junctions at low temp (?)

Courtesy, Intel
Power and Energy Figures of Merit

- **Power consumption in Watts**
  - Determines battery life in hours
  - Energy density \( \sim 120 \text{W-hrs/kg} \)

- **Peak power**
  - Determines power ground wiring designs
  - Sets packaging limits (\( 50 \text{W} / \text{cm}^2 \) ? \( 120 \text{W} \) total ?) \( ($1/\text{Watt}$ ?)
  - Impacts signal noise margin and reliability analysis (Why?)

- **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

Slide courtesy of Mary Jane Irwin, PSU
Power Versus Energy

Power is height of curve

Lower power design could simply be slower

Approach 1

Approach 2

Energy is area under curve

Two approaches require the same energy

Approach 1

Approach 2
Power dissipation in static CMOS gate: 3 components

- **Dynamic capacitive (switching, “useful”) power**
  - Still dominant component in current technology
  - Charging and discharging the capacitor

- **Crowbar current (short-circuit power)**
  - During a transition, current flows through both P and N transistors simultaneously for a SHORT period of time
  - Slow transitions worsen short-circuit power

- **Leakage (“useless power”) current**
  - Even when a device is nominally OFF ($V_{GS}=0$), a small amount of current is still flowing
  - With many devices, can add up to hundreds of mW
Reducing Dynamic Capacitive (Switching) Power

Capacitance: Function of fan-out, wire length, transistor sizes

Supply Voltage: Has been dropping with successive generations

\[ P_{\text{dyn}} = C_L V_{DD}^2 P_{0\rightarrow1} f \]

Activity factor: How often, on average, do wires switch?

Clock frequency: Increasing…

Slide courtesy of Mary Jane Irwin, PSU
Crowbar (Short-Circuit) Current

- 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.

- When $V_{TN} < V_{IN} < V_{DD} + V_{TP}$
  - Both transistors are ON
  - Current flowing directly from $V_{DD}$ to $V_{GND}$ is crowbar current

- Usually not a problem, e.g.,
  - P is ON strongly (LIN but with small $V_{DS}$ if at all)
  - N is barely ON

Slide courtesy of Ken Yang, UCLA
Leakage (Inactive, “Useless”) Power

- Three sources of leakage

- The dominant is the Source-to-Drain leakage current
  - Even when $V_{GS} = 0$, a small amount of charge is still present under the gate
  - Exponentially related to the gate (and S/D) voltage

$$I_D \propto \frac{W}{L} \exp\left(q(V_{GS} - V_T) / nkT\right)$$

- Source/Drain are junctions and some amount of reverse bias, $I_S$ is present
  - Typically much smaller than S/D leakage

- Gate tunneling leakage
  - When $t_{ox}$ is only 5-10 atoms, easy for tunneling current to flow
  - More of an issue sub 0.10-μm technology
2001 ITRS Projections of $1/\tau$ and $I_{sd,\text{leak}}$ for HP, LP Logic

![Graph showing projections of $1/\tau$ and $I_{sd,\text{leak}}$ over years from 2001 to 2015.](image)

- $1/\tau$ — High Perf.
- $I_{sd,\text{leak}}$ — High Perf.
- $1/\tau$ — Low Pwr
- $I_{sd,\text{leak}}$ — Low pwr

The graph illustrates the projected values of $1/\tau$ and $I_{sd,\text{leak}}$ for both high performance (HP) and low power (LP) logic technologies from 2001 to 2015.
Projections for Low Power Gate Leakage

- Need for high K driven by Low Power, not High Performance

Oxy-nitride no longer adequate: high K needed

Simulated $I_{\text{gate}}$, oxy-nitride

$I_{\text{gate}}$ spec. from ITRS

Year

$T_{\text{ox}}$ (normalized)

$J_{\text{gate}}$ (normalized)

Summary: Power and Energy Equations

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

\[ f_{0\to1} = P_{0\to1} \times f_{\text{clock}} \]

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

- Dynamic power (~90% today and decreasing relatively)
- Short-circuit power (~8% today and decreasing absolutely)
- Leakage power (~2% today and increasing relatively)

• Designers need to comprehend issues of memory and logic power, speed/power tradeoffs at the process (HiPerf vs. LowPower) level,
Assignments

- Do Verilog lab
- Homework questions 1, 2, 3 are due on Tuesday
- Read Sections 3.1-3.2, Chapter 11