ECE260B – CSE241A
Winter 2004

Interconnects

Website: http://vlsicad.ucsd.edu/courses/ece260b-w04
Outline

- Interconnects
- Resistance
- Capacitance and Inductance
- Delay
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**
  - ILD Deposition
  - Oxide Trench Etch
  - Metal Fill
  - Metal CMP

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

ECE 260B – CSE 241A Interconnects 5

Andrew B. Kahng, UCSD
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
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
ILD = interlevel dielectric

\[ C_{\text{int}} = e_{\text{ox}} \times \left( W \times L / t_{\text{ox}} \right) \]
Insulator Permittivities

- 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

<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>$\approx 1.5$</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>
Capacitance Values for Different Configurations

- Parallel-plate model substantially underestimates capacitance as line width drops below order of ILD height
  - Why?
Line Dimensions and Fringing Capacitance

- Line dimensions: $W$, $S$, $T$, $H$

- Sometimes $H$ is called $T$ in the literature, which can be confusing
Interwire (Coupling) Capacitance

- 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 applied to multilayer cross-sections
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

- $V_A$ and $V_V$ have opposite transitions (assume $R_A = R_V$ and $C_A = C_V \rightarrow dV_A / dt = -dV_V / dt$)

- Current through $C_C$ is given by $C_C \times (dV_A / dt + dV_V / dt)$. This should be the same as current through equivalent capacitor, which is $C_{CV} dV_V / dt$

- $\Rightarrow C_{CV} = C_C \times (1 + (dV_A / dt) / (dV_V / dt))$

- Equal rise/fall times (or both step inputs) $\Rightarrow C_{CV} = 2C_C$.

Figure 3: Equivalent circuit for two coupled lines
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?

![Aggressor and Victim](attachment:image.png)

*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.

![Diagram showing the relationship between an aggressor and a victim in terms of crosstalk effects.](Slide courtesy of Paul Rodman, ReShape)
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, …
Delay Uncertainty

- 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.

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**

\[ \Phi_{11} = \int_{S_1} B_1 \cdot ds_1 \]

\[
\text{Self Inductance} = \frac{\Phi_{11}}{I}
\]

**Mutual Inductance**

\[ \Phi_{12} = \int_{S_2} B_1 \cdot ds_2 \]

\[
\text{Mutual Inductance} = \frac{\Phi_{12}}{I}
\]
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)

\[
(R + j \omega L)
\]

- 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

\[
R + j \omega L
\]
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?
Outline

- Interconnects
- Capacitance and Inductance
- Resistance
- Delay
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>$\rho$ ((\Omega)-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** → wires get smaller and slower, devices get smaller and faster

- At higher frequencies, current *crowds* to edges of conductor (thickness of conduction = *skin depth*) → increased $R$
Conductor resistivity increases expected to appear around 100 nm linewidth - will impact intermediate wiring first - ~ 2006

Courtesy of SEMATECH

Copper Resistivity: The Real Story
Outline

- Interconnects
- Capacitance and Inductance
- Resistance
- Delay
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.
    - $t_{PLH}$ – 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 model 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.

Slide courtesy of Ken Yang, UCLA
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, $V/I$
- $I = \text{average drive current}$

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
RC Step Response - Propagating Wavefront

Step response of a distributed RC wire as function of location along wire and time
**RC Line Models and Step Response**

<table>
<thead>
<tr>
<th>Voltage Range</th>
<th>Lumped RC-network</th>
<th>Distributed RC-network</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0\rightarrow50%$ ($t_p$)</td>
<td>0.69 RC</td>
<td>0.38 RC</td>
</tr>
<tr>
<td>$0\rightarrow63%$ ($\tau$)</td>
<td>RC</td>
<td>0.5 RC</td>
</tr>
<tr>
<td>$10%\rightarrow90%$ ($t_r$)</td>
<td>2.2 RC</td>
<td>0.9 RC</td>
</tr>
</tbody>
</table>

Step Response of Lumped and Distributed RC Networks:
Points of Interest.

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

- Defined by Elmore (1948) as first moment of impulse response
- $H(t) = \text{step input response}$
- $h(t) = \text{impulse response} = \text{rate of change of step response}$
- $T_{50\%} = \text{median of } h(t)$
- $T_{ED} = \text{approximation of median of } h(t) \text{ by mean of } h(t)$
  - 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

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

- \( \mathcal{MP}(s_0, s_k) \) 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 \( e_{e_i} \), respectively denote the lumped resistance and capacitance of the edge \( e_i \).
Driving Large Capacitances

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

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 $t_p = U * t_{p0} + X/U * t_{p0}$
  - $U_{opt} = \sqrt{X} \quad ; \quad t_{p,\text{opt}} = 2 t_{p0} * \sqrt{X}$

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

- Cascade of buffers with increasing sizes ($U = \text{tapering factor}$) 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 $R_{\text{int}} < R_{\text{tr}}$.
Total line delay as function of driver size, load capacitance

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

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

- 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 $h = \text{optimal buffer size (represent driver as an intermediate buffer)}$

- Then $R_{tr} = R_0/h$ and $C_L = hC_0$

- Propagation Delay: $T_{pd} = \frac{0.4RC}{N} + 0.7 \left( \frac{R_0C}{h} + R_0C_0h + R_0C_0N \right)$

- 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