Circuit Timing

Circuit Delay is a Consequence of the Physics of Transistors and Interconnections.

- As a designer, you need to understand these physics enough to make appropriate design decisions.
- Fortunately for us, CMOS can be accurately modeled in most cases as simple resistive/capacitive circuits:

- Circuit timing is part of the larger hierarchy of design decisions regarding performance.
# Performance Design Decisions

<table>
<thead>
<tr>
<th>Abstraction Layer</th>
<th>Example Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>functional specification</td>
<td>algorithm or ISA</td>
</tr>
<tr>
<td>microarchitecture</td>
<td>function unit multiplexing, pipelining</td>
</tr>
<tr>
<td>RTL</td>
<td>logic organization (factoring)</td>
</tr>
<tr>
<td>transistor circuits</td>
<td>transistor sizing, signal buffering</td>
</tr>
<tr>
<td>layout</td>
<td>wire lengths, layer assignment</td>
</tr>
<tr>
<td>device &amp; wire engineering</td>
<td>materials, processing</td>
</tr>
</tbody>
</table>

- Ultimate goal is to meet performance / cost (area) / power target for the functional specification.
- Subgoal is meet a upper bound on clock period.

## Synchronous Design Clock Constraint

1. Delay in Combinational Logic

\[ T \geq \tau_{\text{clk-Q}} + \tau_{\text{CL}} + \tau_{\text{setup}} + \tau_{\text{clk-skew}} \]

2. Delay in State Elements

3. Delay in wires (grouped with CL or State)

4. Clock Skew

For all paths.

Today: focus on transistor circuit and layout level. Micro-architecture/RTL later, device/wire engineering out of our control.
Parasitics

1. Resistance - the wires are not perfect conductors.
   Some more perfect than others.

2. Capacitance -

Circuit speed is dictated by "RC" delay =>
we need to understand the inherent resistance & capacitance of components,
(4 transistors) to predict circuit performance & power consumption. => design guidelines

Resistance

\[ R = \frac{1}{2 \mu L} \]

Assume T constant over wafer =>

\[ R = \frac{\rho}{W'} \]

aspect ratio

Note: all square geometries have same resistance (independent of area):

If is given \( \frac{1}{2}/\square \) ohms per square

Example:

\[ 4 \text{ ohms} \]

Example:

Changing layers (contacts) - additional R
45nm process:
inner metal layers: ~0.09 Ohms/sq
outer metal layers: ~0.028 Ohm/sq
vias: ~0.9 Ohm

Remember effective transistor "resistance":

n+let ~ 10kΩ/sq
p+let ~ 30kΩ/sq

Transistor Resistor Approximation

n-type:

\[ I_{ds} = \frac{1}{2} \times 150 \mu A = \frac{150 \mu A}{2} = 75 \mu A \]

\[ R = \frac{V}{I} = \frac{5}{75 \mu A} = \frac{5}{7.5 \times 10^{-6}} = 6.67 \times 10^5 \Omega \]

p-type:

\[ I_{ds} = \frac{1}{2} \times 10 \mu A = \frac{10 \mu A}{2} = 5 \mu A \]

\[ R = \frac{V}{I} = \frac{5}{5 \mu A} = \frac{5}{5 \times 10^{-6}} = 10^5 \Omega \]

\[ R = \frac{5V}{152.5 \mu A} = \frac{5}{152.5 \times 10^{-6}} = 32k\Omega \text{ for p-type} \]
Resistive Effects

Parasitic Capacitance

Metal 1  Metal 2  PolySi

\[ C_{\text{poly}} = \frac{A}{2} \]

We would like \( C_{\text{poly}} \) to be \( \frac{1}{2} \mu \f m \)

\[ C_{\text{poly}} = A \times \text{PolySi} + C_{\text{oxide}} \times P \]

The values are published by \textit{Moses}.

\[ C_{\text{poly}} = \frac{C_{\text{oxide}}}{2} \times \frac{A}{2} = \left( 6.12 \times 10^{-8} \text{ F/m} \right) \times \left( 0.55 \times 10^{3} \mu \text{m} \right)^{-1} \]

\[ = 0.06 \text{ F/\mu m}^{2} \]
Capacitance between layers

Capacitance of Diffusion Regions

- Terms a non-linear capacitor.
- Again we have an area contribution.
- Parametric contribution.
- Can be $C_n + C_p = C_n + C_p = \text{total}$.
- For performance estimates, we assume linear.

Transistor Gate Capacitance

- Approximately $200 - 225 \AA$.

Simple Approximation for Capacitive RC delays:
Assume the area of a plate capacitor with the channel $C_{ox} \frac{Cy}{2}$.
Transistor source/drain regions

Typical Capacitance Values

<table>
<thead>
<tr>
<th>Poly</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>Substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>poly</td>
<td></td>
<td></td>
<td></td>
<td>97:</td>
</tr>
<tr>
<td>m1</td>
<td>63:30</td>
<td></td>
<td></td>
<td>38:23</td>
</tr>
<tr>
<td>m2</td>
<td>17:42</td>
<td>51:52</td>
<td></td>
<td>19:60</td>
</tr>
<tr>
<td>m3</td>
<td>10:30</td>
<td>15:36</td>
<td>38:49</td>
<td>13:56</td>
</tr>
</tbody>
</table>

plate: fringe

C2510 UC Berkeley Fall 09
In a 45nm process, a unit inverter delay is 8-10ps.

**Gate Capacitance Calculation**

\[
C_{\text{gate}} = 14 \times (1.25)^2 \times 6 = 1.5
\]

\[
C_{\text{gate}} = 8 \times (1.25)^2 \times 6 = 0.75
\]

\[
R_C = \frac{6 \times 10^{-12}}{\frac{1}{42} \text{ps}} = 42 \text{ps}
\]

**Wire Coupling Capacitance**

- As wires scale down, capacitance goes up.
- Measured by B. Kingsbury:
  - W = 4, space = 3: 0.032 \text{ F/m edge}
  - W = 4: 0.015
  - W = 4, space = 1: 0.030
  - W = 3: 0.010

- This can have a significant effect on dynamic nodes & bus wires; because of capacitor shunt effects.

\[
\frac{V_0}{C_1} = \frac{V}{C_2} + \frac{C_2}{C_1} V_0
\]

- By capacitive charge: \( V = V_0 + \frac{C_0}{C_1 + C_2} V_0 \) (\( C_0 \neq C_1 \))
Node Coupling Effect

Combining R & C

Same effect applies to series connections of transistors.

- In the limit as each section gets shorter, behavior governed by diffusion (heat conduction) equation:
  - For step input at \( V \), response at \( V' \)
  - At a distance \( x = \frac{1}{2} \) from \( V \), is delayed by:
    \[ t = \frac{R C}{2} \]
  - The delay grows as the square of the distance (length)
  - RSI linearly as with C alone.
  - How important is this effect?
0.25 um process, minimal width wires

Combined R & C

$R \cdot C \cdot RC \cdot L$

$r2 \quad 0.21 \quad 127 \quad 0.027 \quad 3.7K$

$m1 \quad 0.16 \quad 60 \quad 0.0096 \quad 10K \sim 1\text{mm}$

$poly \quad 16 \quad 0.18 \quad 3.4 \quad 29$

$rt \quad 13 \quad 1584 \quad 21 \quad 4.8$

$pt \quad 9 \quad 1400 \quad 13 \quad 17.7$

$\frac{\text{N/um-length}}{}$

$\frac{\text{A/um-length}}{}$

length of wire in um for 50ps delay

Driving RC lines

Output voltage
$0-65\%$

Time elapsed
$\frac{\text{time elapsed (distributed RC)}}{}$

$\frac{\text{time elapsed (uniform RC)}}{RC}$

$E_{\text{GS2}} = \frac{1}{2} R_{\text{int}} C_{\text{int}} + R_{\text{d}} C_{\text{int}} + R_{\text{d}} C_{\text{f}} + R_{\text{d}} C_{\text{f}}$
Rise & Falls
times and propagation delay

In a 45nm process, $\tau$ is 4-5ps.

Timing Optimization

Logic Delay Review:

\[ \text{Delay} = \frac{V_o}{V_T} + \text{Gate} \]

\[ B = K_p, \quad \gamma = K_p \cdot Y = K_p \]

$C_{\text{min}} = W \cdot L$ (gate area), usually $L_{\text{min}}$

$\tau = C_{\text{gate}} \cdot (W \cdot Y)$

"C model"

Ignoring $C_{\text{min}}$: If Gate = 1, $V_o = \text{delay} = \tau$

In a 45nm process, $\tau$ is 4-5ps.
What happens to circuit speed as we make all transistors larger?

\[ \text{delay}(t) = \frac{V_{dd}}{4} \cdot \frac{1}{t} + \frac{1}{t} = 2t \]

1. Make all transistors bigger (wider) to offset delay.

- Switching energy \( \pm C \cdot V^2 \) \( \propto \) Power \( \propto Y \)
- More area \( \Rightarrow \) more Gain, less yield
- Is Gain independent of \( Y \)?  
  only independent of wires.

2. Add buffers to drive large loads.

Assume \( I_{off} = 3I \Rightarrow \frac{V_{dd}}{4} = 3.75 \)

\[ T = \frac{500}{7} + 500 \]

\[ T = \frac{20}{7} + \frac{25}{25} + \frac{325}{25} + \frac{325}{25} \]

\[ = 20 + 20 + 20 = 60 \text{ns}! \]

What is optimal scale factor?
Driving Large Capacitive Loads

For example in paddrooms, bus drivers, clock drivers, memory & lines...

How do we drive $C_L$ in minimum time starting from a minimum size network and $C_g$?

$$F = \frac{C_L}{C_g} \text{ (effective total fanout)}$$

if $F \gg 1$ delay $\approx C_L$

Strategy

- $N$ stages $\rightarrow 1/f^N = F$

- $f$ is scale up in inverter strength (transistor size) from stage to stage, which yields delay of $Ft$.

- Total delay $\approx NFt$

- If we use a large factor $f$ - get few stages each with a long delay

- If we use a smaller $f$ - whole delay per stage, but more stages

What value for $N$ minimizes overall delay?

- Total delay $\approx NFt$

- From before $1/f^N F$, $\Delta(t) = N \Delta(t)$

$$N \Delta(t) = \Delta(t)[\frac{1}{f(t)}]$$

fixed by total scaling

- Delay always $\approx \Delta(t)$ (because of step growth of internal stages)

- Relative time penalty

Factors of 3 to 10 normally used
How do we use this result?

What is the ratio of transistor sizes from stage to stage for an effective fastest (delay) of $f$?

$$\frac{C_2}{C_1} = \frac{R(C + C_2)}{R(C + C_1)}$$

$$\frac{C_2}{C_1} = \frac{f_2}{f_1} = \frac{R(C + C_2)}{R(C + C_1)}$$

$$f_1 = f_2 (1 - C_1 \cdot 2)$$

$$f_0 = f_1 (1 - C_1)$$

$C_0$ is input & tech dependent

for $v_1 = \sim 3/2$

$$f = 3 + (3-1)3/2 = 6$$

Each factor of stage to stage is min delay

---

### 3. Balance Delays

Example:

![Circuit Diagram]

- a) $C_p = 3$, $V_n = 1$  \[ T = 48 + 64 \cdot 2 = 68 \cdot 2 \]
- b) $C_p = 6$, $V_n = 2$  \[ T = 8 + 32 = 40 \cdot 2 \]
- c) $C_p = 12$, $V_n = 4$  \[ T = 16 + 16 = 32 \cdot 2 \]
- d) $C_p = 24$, $V_n = 8$  \[ T = 32 + 8 = 40 \cdot 2 \]
"Rebuffer" Long Wires

Wire Delay =
\[
\frac{1}{2} \cdot R_{\text{total}} \cdot C_{\text{total}} = \frac{1}{2} \cdot R \cdot C \cdot L^2
\]

Buffer adds some delay.
With too many splits, buffer delay dominates.
Modern designs rules don't allow.
Timing Closure: Searching for and beating down the critical path

Must consider all connected register pairs, paths from input to register, register to output. Don’t forget the controller.

- Design tools help in the search.
- Synthesis tools work to meet clock constraint, report delays on paths,
- Special static timing analyzers accept a design netlist and report path delays,
- and, of course, simulators can be used to determine timing performance.

Tools that are expected to do something about the timing behavior (such as synthesizers), also include provisions for specifying input arrival times (relative to the clock), and output requirements (set-up times of next stage).
Timing Analysis, real example

The critical path

Most paths have hundreds of picoseconds to spare.


Timing Analysis Tools

- Static Timing Analysis: Tools use delay models for gates and interconnect. Traces through circuit paths.
- Delay models capture
  - For each input/output pair: internal delay (output load independent)
  - output dependent delay
- Standalone tools (PrimeTime) and part of logic synthesis.
- Back-annotation takes information from results of place and route to improve accuracy of timing analysis.
- PC in “topographical mode” uses preliminary layout information to model interconnect parasitics.
- Prior versions used a simple fan-out model of gate loading.
Hold-time Violations

- Some state elements have positive hold time requirements.
- How can this be?
- Fast paths from one state element to the next can create a violation. (Think about shift registers!)
- CAD tools do their best to fix violations by inserting delay (buffers).
  - Of course, if the path is delayed too much, then cycle time suffers.
  - Difficult because buffer insertion changes layout, which changes path delay.

Conclusion

- Timing Optimization: You start with a target on clock period. What control do you have?
- Biggest effect is RTL manipulation.
  - i.e., how much logic to put in each pipeline stage.
- In most cases, the tools will do a good job at logic/circuit level:
  - Logic level manipulation
  - Transistor sizing
  - Buffer insertion
- But some cases may be difficult and you may need to help
  - Hand instantiate cells
Simple propagation delay model

In a 45nm process, \( \tau \) is 4-5ps.