Introduction to CMOS VLSI Design

Delay Part B

Lecture by Jay Brockman
University of Notre Dame Fall 2008
Modified by Peter Kogge Fall 2010, 2011, 2015, 2018
Based on lecture slides by David Harris, Harvey Mudd College
http://www.cmosvlsi.com/coursematerials.html

Outline

- Delay Part A
  - Capacitance
  - Effective Resistance
  - RC Delay
- Delay Part B
  - Review
  - Inverter Delay
  - The Elmore Model
  - Effect of Load Capacitance
Capacitance

- Conductors separated by insulator have capacitance
- **Gate to channel capacitor** is very important
  - Creates channel charge necessary for operation
- Source and drain have capacitance to body
  - Across reverse-biased diodes
  - Called **diffusion capacitance** because it is associated with source/drain diffusion

Gate Capacitance

- Approximate channel as “connected to source”
- \( C_{gs} = \varepsilon_{ox} WL/t_{ox} = C_{ox} WL = C_{permicron} W \)
  - proportional to “width” ONLY
- \( C_{permicron} = \varepsilon_{ox} (L/t_{ox}) \) typically ~2 fF/\( \mu \)m of gate width
  - \( L \) and \( t_{ox} \) both scale with process
- Often just \( C_g \)
**Diffusion Capacitance**

- \( C_{sb}, C_{db} = \) “source/drain to bulk”
- Undesirable, called *parasitic* capacitance
- Capacitance depends on area and perimeter
  - Comparable to \( C_g \) for **contacted diffusion**
  - \( \frac{1}{2} C_g \) for **uncontacted**
  - Varies with process
  - Often just \( C_d \)
- “Contacted diffusion” occurs when there is a metal contact “touching” the diffusion
  - i.e. there’s a “wire” on the source or drain
- Appears on both source & drain
  - But ignore when connected to \( V_{dd} \) or Gnd
    - Capacitance is “shorted out”

---

**Contacted vs. Uncontacted Capacitance**

- **Transistor Layout**
  - Contacted Source/Drain – Physically Larger => more cap
  - Uncontacted Source/Drain – Smaller
- **Example nmos transistors in NAND**
Definitions

- **Waveform**
  - *Rise time* \( t_r \) = time to rise from 20% of \( V_{dd} \) to crossing 80% of \( V_{dd} \)
  - *Fall time* \( t_f \) = time fall from 80% of \( V_{dd} \) to crossing 20% of \( V_{dd} \)
  - *Edge Rate* \( t_{ref} \) = \( (t_r + t_f)/2 \)

- **Logic gate input to output**
  - *Propagation delay* \( t_{pd} \) = max time from input crossing 50% of \( V_{dd} \) to output crossing 50% of \( V_{dd} \)
    - \( t_{pdr} \) = delay when input is rising
    - \( t_{phl} \) = delay when output goes from High to Low
    - \( t_{pdf} \) = delay when input is falling
    - \( t_{plh} \) = delay when output goes from Low to High
    - *Delay* \( t_p \) = \( (t_{phl} + t_{plh})/2 \)
  - *Contamination delay* \( t_{pd} \) = min time from input crossing 50% of \( V_{dd} \) to output crossing 50% of \( V_{dd} \)
    - i.e with no load on output, in either direction

Effective Resistance

- **Shockley models have limited value**
  - Not accurate enough for modern transistors
  - Too complicated for much hand analysis

- **Simplification: treat transistor in on state as resistor**
  - Replace \( I_{ds}(V_{ds}, V_{gs}) \) with **Effective Resistance** \( R \)
    - \( I_{ds} \sim V_{ds}/R \)
  - \( R \) averaged across switching of digital gate

- **Too inaccurate to predict current at any given time**
  - *But good enough to predict RC delay*
Approximating Output Resistance (4.3.7)

- Want Effective Resistance that has ~ same current as transistor when $V_{DS} = V_{dd}/2$
- No single $R$ value matches transistor through switching event
- Approach:
  - Look at IV trajectory as input $V_{gs}$ rises high enough to start transition
    - While still being in saturation
  - Compute average $V/I$ on this trajectory

$$R = \ln(2) \times \frac{3}{4} \times \frac{V_{dd}}{I_{dsat}} \approx \frac{1}{2} \left( \frac{V_{dd}}{I_{dsat}} \right)$$

Approximating $R_{ON}$

- $C = 50 \text{ ff}$
- $\Delta t = 0.27 \text{ ns}$
- $R = \Delta t / (0.69 \times C) = 7.8 \ \text{K}\Omega$
Drawing Equivalent Circuits for Delay

- Model rise and fall as separate circuits
- Look at only the output of logic gate being modeled
  - But consider capacitance from the gate it is driving
- Model transistors that turn “OFF” by an open
- Ignore all capacitors with both ends to a rail
- Model transistors that turn “ON” by their resistance
  - If source is $V_{dd}$ or $V_{gnd}$, model voltage on that terminal as a “STEP Voltage”
- Ignore all capacitors with one end “floating”
- Lump ALL capacitors tied to same wire together
  - Ignore which rail other side goes to
- Draw as equivalent RC “Ladder” driven by step voltage

Inverter Delay Estimate

- Estimate the delay of a fanout-of-1 inverter
Inverter Delay Estimate

- Assume input going Low to High

Inverter Delay Estimate

- Estimate the delay of a fanout-of-1 inverter
Inverter Delay Estimate

- Estimate the delay of a fanout-of-1 inverter

\[ d \approx 6RC \]

What is Aggregate C?

\[ d \approx 6RC \approx 6 (7 \text{ K}\Omega)(2 \text{ fF}) \approx 84 \text{ ps} \]
\[ \approx \text{0.6 } \mu\text{m process} \]

\[ \approx 6 (2 \text{ K}\Omega)(2 \text{ fF}) \approx 24 \text{ ps} \]
\[ \approx \text{90 nm process} \]
Compare three delay cases

1

2

3

Is Case 1, Case 2 or Case 3 Faster?

Delay Case 1
Delay Case 1

\[ d = 84 \, RC \]

Delay Case 2
Note the geometric progression in size!
- 3X per stage.

The delay for each stage is the same

\[
d1 = 12 \text{ RC} \\
d2 = 36/3 \text{ RC} = 12 \text{ RC} \\
d2 = 108/9 \text{ RC} = 12 \text{ RC} \\
\]

\[
d = d1 + d2 + d3 = 36 \text{ RC} << 84 \text{ RC}
\]
Delay Case 3

\[
\begin{align*}
d_1 &= 9 \text{ RC} \\
d_2 &= 9 \text{ RC} \\
d_3 &= 9 \text{ RC} \\
d_4 &= 9 \text{ RC} \\
d_5 &= 129/16 \text{ RC} = 8.1 \text{ RC}
\end{align*}
\]

\[d = d_1 + d_2 + d_3 + d_4 + d_5 = 44.1 \text{ RC}\]

- Case 2 = 36 RC < Case 3 = 44 RC < Case 1 = 84 RC
- You can have too much of a good thing!

Some Review

- \(k\): transistor's width is \(k\) times unit width
- On resistance is \(1/k\) times unit transistor
- \(C_g\) is capacitance of gate to body
  - \(C_{g-k}\) is \(k\) times \(C_{g-unit}\) of unit transistor
- \(C_{diff}\) is capacitance to body from a contacted source or drain
  - Approximately = \(C_g\)
- Diffusion capacitance of uncontacted source/drain connection is less but approx as same
3 Input NAND

FIGURE 4.7 Equivalent circuits for a 3-input NAND gate

Elmore Delay Model

FIG 4.3 RC ladder for Elmore delay

\[ \text{Delay} = \sum_{i=1}^{N} R_{n_i} C_i \sim \sum_{i=1}^{N} C_i \sum_{j=1}^{n_i} R_j \]
What If Output Not At End?

\[ V_{\text{out}} = R_1 \sum C_i \]

Ignore \( R_2 \) thru \( R_n \), & *just sum C’s*

\[ \text{Delay} = R_1 \sum C_i \]

In reality, \( C_2, \ldots C_N \) “shielded” by R’s
(don’t have to charge all way to voltage at Y)
Thus a “conservative” estimate

Elmore Delay of 3 NAND

- Assume not driving any other gates
  - This is called the **Parasitic Delay**
- Fall Delay = \((R/3)3C + (2R/3)3C + (3R/3)9C = 12RC\)
- Rise time = \(R \times (9C + 3C + 3C)\)
  - Why is this worst case?
What’s the Capacitance We’re Driving?

\[ C_{\text{in}} = \text{capacitance a logic gate imposes on whoever drives it} \]

Assuming:
- gate width is indicated as \# in each transistor
- gate capacitance of unit transistor \( 1 \, C_g = C \)

**FIG 4.9** Logic gates sized for unit resistance

Simple Example: Effects of Drive

- Estimate worst-case rising and falling delay of 2-input NAND driving \( h \) identical gates.
Summary Delay Estimation

Assume driving $h$ identical 2-input NANDs

Only 1 pmos on for slowest case

Assume int node charged

Size 2 because in series

Elmore delays:
Worst Case Rise = $R(6+4h)C = (6+4h)RC$
Worst Case Fall = $(R/2)(2C) + R*(6+4h)C = (7+4h)RC$

Let's Do the Book’s 3NAND

• What is the load on Y?
• What is the fall time?
• What is the rise time?
Contamination Delay

- Contamination = min possible
- On fall, best if both bottom NMOS Ts on
  - Diffusion cap already drained
  - Only R effective left
  - Delay =
- On rise, best when ALL 3 PMOS turn on
  - Resistances in parallel
  - Delay =

Layout Dependent Capacitance

FIG 4.1 3-input NAND gate with unit rise and fall resistance

FIG 4.2 3-input NAND gate annotated with capacitances

Assuming all diffusion nodes contacted
Folding Wide Transistors

24/12 $\lambda$ Inverter

4$\lambda$ unit transistor has diff cap C

Total cap = 9C

Total cap = 4.5C

FIGURE 4.18 Layout styles: (a) conventional, (b) folded