EE141-Spring 2010
Digital Integrated Circuits

Lecture 20
Technology Scaling

Administrativia

- Midterm 2 on We April 7 6:30-8pm in 277 Cory
  - Open book.
  - Covers everything until Lecture 17 (DOMINO LOGIC). Registers and timing NOT included.
- Review Session: Tu April 6 6:30-8pm in 289 Cory
- DO NOT FORGET THE PROJECT!
  - Extra office hours of TAs on Mo and Tu (during lab hours) in 353 Cory
- Project Phase 1 has been graded. Results will be posted early next week.
Class Material

- Last lecture
  - Timing
- Today’s lecture
  - Technology Scaling
- Reading (Ch 5)
**Timing Constraints**

- **Cycle time (max):** \( T_{\text{Clk}} > t_{\text{clk-q}} + t_{\text{logic}} + t_{\text{setup}} \)
- **Race margin (min):** \( t_{\text{hold}} < t_{\text{clk-q,min}} + t_{\text{logic,min}} \)

**Clock Constraints in Edge-Triggered Systems**

If launching edge is late and receiving edge is early, the data will not be too late if:

\[
t_{\text{clk-q}} + t_{\text{logic}} + t_{\text{setup}} < T_{\text{CLK}} - t_{JS,1} - t_{JS,2} + \delta
\]

Minimum cycle time is determined by the maximum delays through the logic

\[
t_{\text{clk-q}} + t_{\text{logic}} + t_{\text{setup}} - \delta + 2t_{JS} < T_{\text{CLK}}
\]

Skew can be either positive or negative
Clock Constraints in Edge-Triggered Systems

If launching edge is early and receiving edge is late:

\[ t_{clk-q, min} + t_{logic, min} - t_{JS, 1} > t_{hold} + t_{JS, 2} + \delta \]

Minimum logic delay

\[ t_{clk-q, min} + t_{logic, min} > t_{hold} + 2t_{JS} + \delta \]

(This assumes jitter at launching and receiving clocks are independent – which usually is not true)

Reference

<table>
<thead>
<tr>
<th>Clock Period</th>
<th>Adder</th>
<th>Absolute Value</th>
<th>Logarithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( a_1 + b_1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( a_2 + b_2 )</td>
<td>(</td>
<td>a_1 + b_1</td>
</tr>
<tr>
<td>3</td>
<td>( a_3 + b_3 )</td>
<td>(</td>
<td>a_2 + b_2</td>
</tr>
<tr>
<td>4</td>
<td>( a_4 + b_4 )</td>
<td>(</td>
<td>a_3 + b_3</td>
</tr>
<tr>
<td>5</td>
<td>( a_5 + b_5 )</td>
<td>(</td>
<td>a_4 + b_4</td>
</tr>
</tbody>
</table>

Pipelining
Loop Unrolling

Latch-Based Clocking

(Domino logic almost always uses latch-based clocking)
Latch vs. Flip-flop

- **In a flip-flop based system:**
  - Data launches on one rising edge
    - And must arrive before next rising edge
  - If data arrives late, system fails
    - If it arrives early, wasting time
  - Flip-flops have hard edges

- **In a latch-based system:**
  - Data can pass through latch while it is transparent
  - Long cycle of logic can borrow time into next cycle
    - As long as each loop finished in one cycle

Time Borrowing Example

![Diagram showing time borrowing example]
**Latch vs. Flip-flop Summary**

- Flip-flops generally easier to use
  - Most digital ASICs designed with register-based timing
- But, latches (both pulsed and level-sensitive) allow more flexibility
  - And hence can potentially achieve higher performance
  - Latches can also be made more tolerant of clock un-certainty
  - More in EE241

**CMOS Transistor Scaling**
### Goals of Technology Scaling

- **Make things cheaper:**
  - Want to sell more functions (transistors) per chip for the same money
  - Or build same products cheaper
  - Price of a transistor has to be reduced
- **But also want to be faster, smaller, lower power…**

### Technology Scaling

- **Benefits of 30% “Dennard” scaling (1974):**
  - Double transistor density
  - Reduce gate delay by 30% (increase operating frequency by 43%)
  - Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency)
- **Die size used to increase by 14% per generation (not any more)**
- **Technology generation spans 2-3 years**
Moore was not always accurate

Projected 2000 Wafer, circa 1975

Technology Scaling (1)

2X reduction every ~5 years

Minimum Feature Size
Technology Scaling (2)

Number of components per chip

Technology Scaling (3)

Propagation Delay
**Technology Scaling (4)**

![Graphs](https://i.imgur.com/3Q5Q5Q5.png)

(a) Power dissipation vs. year.

(b) Power density vs. scaling factor.

From Kuroda

EECS141

Lecture #20

---

**Technology Scaling Models**

- **Full Scaling (Constant Electrical Field)**
  - ideal model — dimensions and voltages scale together by the same factor $S$

- **Fixed Voltage Scaling**
  - most common model until 1990’s
  - only dimensions scale, voltages remain constant

- **General Scaling**
  - most realistic for today’s situation —
  voltages and dimensions scale with different factors
Scaling

- $W, L, t_{ox}: 1/S$
- $V_{DD}, V_T: 1/S$
- Area: $WL$
- $C_{ox}: 1/tox$
- $C_L: C_{ox}WL$
- $I_D: C_{ox}(W/L)(V_{DD}-V_T)^2$
- $R_{eq}: V_{DD}/I_{DSAT}$

Full Scaling (Dennard, Long-Channel)
Full Scaling (Dennard, Long-Channel)

- $W, L, t_{ox}: 1/S$
- $V_{DD}, V_T: 1/S$

- $t_p: R_{eq} C_L$
- $P_{avg}: C_L V_{DD}^2 / t_p$
- $P_{avg} / A: C_{ox} V_{DD}^2 / t_p$

Scaling Relationships for Long Channel Devices

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>Full Scaling</th>
<th>General Scaling</th>
<th>Fixed Voltage Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W, L, t_{ox}$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td></td>
</tr>
<tr>
<td>$V_{DD}, V_T$</td>
<td>$1/S$</td>
<td>$1/U$</td>
<td>$1$</td>
<td></td>
</tr>
<tr>
<td>$N_{SUB}$</td>
<td>$V/W_{dep}^2$</td>
<td>$S$</td>
<td>$S^2/U$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Area/Device</td>
<td>$WL$</td>
<td>$1/S^2$</td>
<td>$1/S^2$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>$C_{ax}$</td>
<td>$1/t_{ox}$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$C_L$</td>
<td>$C_{ox} WL$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td>$1/S$</td>
</tr>
<tr>
<td>$k_n, k_p$</td>
<td>$C_{ox} W/L$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$I_{av}$</td>
<td>$k_{np} V^2$</td>
<td>$1/S$</td>
<td>$S/U^2$</td>
<td>$S$</td>
</tr>
<tr>
<td>$t_p$ (intrinsic)</td>
<td>$C_L V / I_{av}$</td>
<td>$1/S$</td>
<td>$U/S^2$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>$P_{av}$</td>
<td>$C_L V^2 / t_p$</td>
<td>$1/S^3$</td>
<td>$S/U^3$</td>
<td>$S$</td>
</tr>
<tr>
<td>PDP</td>
<td>$C_L V^2$</td>
<td>$1/S^3$</td>
<td>$1/SU^2$</td>
<td>$1/S$</td>
</tr>
</tbody>
</table>
**Full Scaling (Dennard, Short-Channel)**

- $W, L, t_{ox}: 1/S$
- $V_{DD}, V_T: 1/S$

- Area: $WL$
- $C_{ox}: 1/tox$
- $C_L: C_{ox}WL$
- $I_D: WC_{ox}v_{sat}(V_{DD}-V_T-V_{SAT}/2)$
- $R_{eq}: V_{DD}/I_{DSAT}$

---

**Full Scaling (Dennard, Short-Channel)**

- $W, L, t_{ox}: 1/S$
- $V_{DD}, V_T: 1/S$

- $t_p: R_{eq}C_L$
- $P_{avg}: C_LV_{DD}^2/t_p$
- $P_{avg}/A: C_{ox}V_{DD}^2/t_p$
Transistor Scaling  
(Velocity-Saturated Devices)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>Full Scaling</th>
<th>General Scaling</th>
<th>Fixed-Voltage Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W, L, t_{ox}$</td>
<td>1/$S$</td>
<td>1/$S$</td>
<td>1/$S$</td>
<td></td>
</tr>
<tr>
<td>$V_{th}, V_T$</td>
<td>1/$S$</td>
<td>1/$U$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$N_{IDD}$</td>
<td>$V/W^2_{min}$</td>
<td>$S$</td>
<td>$S^2/U$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Area/Device</td>
<td>$W^2$</td>
<td>1/$S^3$</td>
<td>1/$S^3$</td>
<td>1/$S^3$</td>
</tr>
<tr>
<td>$C_{inv}$</td>
<td>1/$L_{min}$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$C_{p, WL}$</td>
<td>$C_{inv, WL}$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$I_{sat}$</td>
<td>$C_{inv, WY}$</td>
<td>1/$S$</td>
<td>1/$U$</td>
<td>1</td>
</tr>
<tr>
<td>Current Density</td>
<td>$I_{sat}/Area$</td>
<td>$S$</td>
<td>$S^2/U$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>$R_{on}$</td>
<td>$V_{th}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Intrinsic Delay</td>
<td>$R_{on}C_{inv}$</td>
<td>1/$S$</td>
<td>1/$S$</td>
<td>1/$S$</td>
</tr>
<tr>
<td>$P$</td>
<td>$I_{sat}V$</td>
<td>1/$S^3$</td>
<td>1/$U^2$</td>
<td>1</td>
</tr>
<tr>
<td>Power Density</td>
<td>$P/Area$</td>
<td>1</td>
<td>$S^2/U$</td>
<td>$S^2$</td>
</tr>
</tbody>
</table>

An interesting question

What will did cause this model to break?

- Leakage set by $kT/q$
  - Temp. does not scale
  - $V_T$ set to minimize power
- Power actually increased
  - Leakage increased drastically
  - $f$ increased faster than device speed
  - Hit cooling limit
- Process Variation
  - Hard to build very small things accurately (less averaging)