Admin

- Project proposals due by Mo 5pm (by e-mail to jan@eecs.berkeley.edu)
  - Title
  - Short abstract of ½ page describing the problem you are trying to tackle
- No class nor office hours on Monday
ISSCC 2006

Keynotes (Monday Morning)
- *Where CMOS Is Going: Trendy Hype versus Real Technology*, T.C. Chen, VP of Science and Technology, IBM
- *ICs for Mobile Multimedia Applications*, Herman Eul, Infineon Technologies, Germany
- *The Future of Computing for Real-Time Entertainment*, Ken Kutagari, President and CEO of Sony Entertainment

Interesting Short Courses (on Su)
- F2: Embedded-SRAM Design
- F3: Circuit Design in Emerging Technologies

Tutorials (on Su)
- T3: Introduction to Statistical Validation and Techniques for Design Optimization
- T5: Multi-Level Cell Design for Flash Memory
- T7: 3D Integration

Interesting Sessions

Mo
- S5: Processors
- S7: Non-Volatile Memories

Tu
- S8: DRAM and TCAM
- S18: Clock and Data Recovery
- S21: Advanced Clocking, Logic and Signaling Techniques

We
- S22: Low-Power Multimedia
- S23: Technical and Architectural Directions
- S24: High-Performance Digital Circuits
- S29: Power Management and Distribution
- S30: Silicon for Biology
- S32: PLLs, VCOs and Dividers
- S34: SRAM
Some other stuff

- **Panels:**
  - Mo: Is the Digital Circuit Designer Dead?
  - Mo: Emerging and Disrupting Memory Technologies
  - Tu: Present and Future Classical Circuits with less than 25 transistors

- **Th Circuit Design Forum**
  - Multi-core architectures, designs and implementation challenges

---

Today’s lecture

- Using the models we have created so far to do create an environment for optimization

- **Reading:**
  - ICCAD paper by Stojanovic et al.
  - Chapters 2 and 3 in the text by K. Bernstein (High Speed CMOS Design Styles)
  - Background material from Rabaey, 2nd ed, Chapters 5, 6.
Static Timing Analysis

- Computing critical (longest) path delay
  - Longest path algorithm on DAG [Kirkpatrick, IBM Jo. R&D, 1966]
- Used in most ASIC designs today
- Limitations
  - False paths
  - Simultaneous arrival times

False Paths

Inside Carry Bypass Adder - 1

- Longest graphical/topological path runs along carry chain from stage to stage
- Longest path analysis would identify red path as critical
Signal Arrival Times

- NAND gate:

![NAND gate diagram]

Signal Arrival Times

- NAND gate:

![NAND gate diagram]
Simultaneous Arrival Times

- NAND gate:

Impact of Arrival Times

- Up to 25%
- A arrives early
- B arrives early

Delay

$t_{A} - t_{B}$
Optimization for Performance

- Performance critical blocks
- Start with a synthesized design
  - Easier to explore architectures
  - Easy to verify
  - Provides some level of performance optimization
- Understand the limits of synthesized designs

Performance Optimization

Increasing the performance increases power!
How to Increase Performance?

- Scale technology
- Circuit level:
  - Transistor sizing, buffering
  - Wire optimization, repeaters
  - Supply and Threshold voltage
  - Logic styles
  - Timing, latches
- Microarchitecture
  - Block topologies (adders, multipliers)
  - Pipelining
  - Parallelism

Sizing Logic Paths for Speed

- Frequently, input capacitance of a logic path is constrained
- Logic has to drive some capacitance
- Example: ALU load in an Intel’s microprocessor is > 0.5pF
- How do we size the ALU datapath to achieve maximum speed?
- Review the method of logical effort
Inverter Chain

If \( C_L \) and \( C_{in} \) are given:
- How many stages are needed to minimize the delay?
- How to size the inverters?

May need some additional constraints.

Delay Formula

\[ \text{Delay} \sim R_W (C_{int} + C_L) \]

\[ t_p = kR_W C_{int} (1 + C_L / C_{int}) = t_{p0} (1 + f / \gamma) \]

\( C_{int} = \gamma C_{gin} \) with \( \gamma \approx 1 \)
\( f = C_L/C_{gin} \) - effective fanout
\( R = R_{unit}/W ; C_{int} = WC_{unit} \)
\( t_{p0} = 0.7 R_{unit} C_{unit} \)
Apply to Inverter Chain

\[ t_p = t_{p1} + t_{p2} + \ldots + t_{pN} \]

\[ t_{pj} \sim R_{\text{unit}}C_{\text{unit}} \left( 1 + \frac{C_{\text{gin},j+1}}{\gamma C_{\text{gin},j}} \right) \]

\[ t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0} \sum_{i=1}^{N} \left( 1 + \frac{C_{\text{gin},j+1}}{\gamma C_{\text{gin},j}} \right), \quad C_{\text{gin},N+1} = C_L \]
Optimal Tapering for Given $N$

Delay equation has $N - 1$ unknowns, $C_{\text{gin},2} - C_{\text{gin},N}$

Minimize the delay, find $N - 1$ partial derivatives

Result: $C_{\text{gin},j+1}/C_{\text{gin},j} = C_{\text{gin},j}/C_{\text{gin},j-1}$

Size of each stage is the geometric mean of two neighbors

$$C_{\text{gin},j} = \sqrt{C_{\text{gin},j-1}C_{\text{gin},j+1}}$$

- each stage has the same effective fanout ($C_{\text{out}}/C_{\text{in}}$)
- each stage has the same delay

Optimum Delay and Number of Stages

When each stage is sized by $f$ and has same effective fanout $f$:

$$f^N = F = C_L / C_{\text{gin},1}$$

Effective fanout of each stage:

$$f = \sqrt[N]{F}$$

Minimum path delay

$$t_p = N t_{p0} \left( 1 + \sqrt[N]{F} / \gamma \right)$$
Example

\[ C_L/C_1 \text{ has to be evenly distributed across } N = 3 \text{ stages:} \]
\[ f = \sqrt[3]{8} = 2 \]

Optimum Number of Stages

For a given load, \( C_L \) and given input capacitance \( C_{in} \)
Find optimal number of stages, \( N \), and optimal sizing, \( f \)

\[ C_L = F \cdot C_{in} = f^N C_{in} \quad \text{with} \quad N = \frac{\ln F}{\ln f} \]

\[ t_p = N t_{p0} \left( F^{1/N} / \gamma + 1 \right) = t_{p0} \ln F \left( \frac{f}{\ln f} + \frac{\gamma}{\ln f} \right) \]

\[ \frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{\gamma} \cdot \ln f - 1 - \frac{\gamma}{f} = 0 \]

For \( \gamma = 0 \), \( f = e \), \( N = \ln F \)

\[ f = e^{1+\gamma/f} \]
Optimum Effective Fanout $f$

Optimum $f$ for given process defined by $\gamma$

$$f = e^{(1+\gamma/f)}$$

$f_{opt} = 3.6$
for $\gamma=1$

Impact of Loading on $tp$

With self-loading $\gamma=1$
Extending the Model

For given $N$: \( \frac{C_{i+1}}{C_i} = \frac{C_i}{C_{i-1}} \)

To find $N$: \( \frac{C_{i+1}}{C_i} \approx 4 \)

Method of logical effort generalizes this to any logic path

\[
\text{Delay} = \sum_{i=1}^{N} (p_i + g_i \cdot f_i) \quad \text{(in units of } \tau_{\text{inv}})\]

Logical Effort

\[
\text{Delay} = k \cdot R_{\text{unit}} C_{\text{unit}} \left( 1 + \frac{C_L}{\gamma C_{\text{in}}} \right) = \tau(p + g \cdot f) \]

- $p$ – intrinsic delay - gate parameter $\neq f(W)$
- $g$ – logical effort – gate parameter $\neq f(W)$
- $f$ – electrical effort (fanout)

Normalize everything to an inverter:
\( g_{\text{inv}} = 1, \quad \gamma_{\text{inv}} = 1 \)

Divide everything by $\tau_{\text{inv}}$
(everything is measured in unit delays $\tau_{\text{inv}}$)
Assume $\gamma = 1$. 
Delay in a Logic Gate

Gate delay:
\[ d = h + p \]

Effort delay \quad Intrinsic delay

Effort delay:
\[ \text{logical effort} \quad \text{effective fanout} = \frac{C_{out}}{C_{in}} \]

Logical effort is a function of topology, independent of sizing
Effective fanout (electrical effort) is a function of load/gate size

Logical Effort

- Inverter has the smallest logical effort and intrinsic delay of all static CMOS gates
- Logical effort of a gate presents the ratio of its input capacitance to the inverter capacitance when sized to deliver the same current
- Logical effort increases with the gate complexity
### Logical Effort

Logical effort is the ratio of input capacitance of a gate to the input capacitance of an inverter with the same output current.

Size factor: 1.5

Size factor: 1.8

![Logical Effort Diagrams]

### Logical Effort of Gates

<table>
<thead>
<tr>
<th>Fan-out (f)</th>
<th>Normalized delay (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(Fan-in)</td>
<td>t_{pNAND}</td>
</tr>
<tr>
<td>F(Fan-in)</td>
<td>t_{pINV}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>g</th>
<th>p</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Logical Effort of Gates

- Fan-out ($f$)
- Normalized delay ($d$)

For $g=1$, $p=1$, $d=f+1$

For $g=3.5/3$, $p=5.5/3$, $d=(3.5/3)f+1.8$

Add Branching Effort

Branching effort:

$$b = \frac{C_{on-path} + C_{off-path}}{C_{on-path}}$$
Multistage Networks

\[ \text{Delay} = \sum_{i=1}^{N} (p_i + g_i \cdot f_i) \]

Stage effort: \( h_i = g_i f_i \)
Path electrical effort: \( F = \frac{C_{out}}{C_{in}} \)
Path logical effort: \( G = g_1 g_2 \ldots g_N \)
Branching effort: \( B = b_1 b_2 \ldots b_N \)
Path effort: \( H = GFB \)
Path delay \( D = \sum d_i = \sum p_i + \sum h_i \)

Optimum Effort per Stage

When each stage bears the same effort:
\[ h^N = H \]
\[ h = \frac{H}{N} \]

Stage efforts: \( g_1 f_1 = g_2 f_2 = \ldots = g_N f_N \)
Effective fanout of each stage: \( f_i = h/g_i \)
Minimum path delay
\[ \hat{D} = \sum (g_i f_i + p_i) = NH^{1/N} + P \]
Optimal Number of Stages

For a given load, and given input capacitance of the first gate, find optimal number of stages and optimal sizing

\[ D = NH^{1/N} + P \]

\[ \frac{\partial D}{\partial N} = -\frac{H^{1/N}}{N} \ln(H) + H^{1/N} + \left( \frac{\partial P}{\partial N} \right) = 0 \]

Substitute ‘best stage effort’ \( h = H^{1/N} \)

Logical Effort Optimization Methodology

- For smaller problems, easy to translate into set of analytical expressions
- Feed them into Matlab optimizer
- With some careful manipulations, can be turned into a convex optimization problem (Stojanovic)
- Easily extended to add power/energy
Optimization for Performance

Options
- Technology choice
  - CMOS, bipolar, BiCMOS, GaAs, Superconducting
- Logic level optimizations
  - logic depth, network topology, fan-out, gate complexity
- Circuit optimizations
  - logic style, transistor sizing
- Physical optimization
  - implementation choice, layout strategy
- Wires are the key

Logic Level Optimizations

Logic Depth

Techniques: Restructuring, pipelining, retiming, technology mapping

Well covered by today's logic and sequential synthesis
Logic Optimizations (2)

Fanout

$T_p = O(FO)$
also effects wiring capacitance

Technique: Removal of common sub-expression
Start from tree structure/output

Logic Optimizations (3)

Fanin

$T_p = O(FI^2)$
Observation: only true if FI translates in series devices - otherwise linear
e.g. NAND pull-down
NOR pull-up

AVOID LARGE FAN-IN GATES! (Typically not more than FI < 4)
Logic Optimizations (4)

All the gates have the same drive current

Technology Mapping for Performance

Alternative coverings
Use low F1 modules on critical path(s)

Library composition?
Sizing/Supply Voltage

- Joint optimization of parameters essential – Trading off performance for energy

**Energy-delay sensitivity**

\[
\text{Sens}(V_{dd}) = \left. \frac{\partial E}{\partial V_{dd}} \right|_{V_{dd} = V_{dd}^*} - \left. \frac{\partial D}{\partial V_{dd}} \right|_{V_{dd} = V_{dd}^*}
\]

- Proposed by Zyban at ISLPED02
- \( \Delta E = \text{Sens}(A) \cdot (\Delta D) + \text{Sens}(B) \cdot \Delta D \)

At the optimal point, all sensitivities should be the same

---

**Alpha-power based delay model**

\[
t_p = \frac{K_d \cdot V_{dd}}{(V_{dd} - V_{on})^\alpha d} \left( \frac{W_{out}}{W_{in}} + \frac{W_{par}}{W_{in}} \right)
\]

- Fitting parameters
  - \( V_{on} \), \( \alpha_d \), \( K_d \)
Energy model

♦ Switching energy

\[ E_{Sw} = \alpha_0 \cdot \left( C(W_{out}) + C(W_{par}) \right) \cdot V_{dd}^2 \]

♦ Leakage energy

\[ E_{Lk} = W_{in} \cdot I_0(S_{in}) \cdot e^{-\frac{V_{th}-\gamma V_{dd}}{V_0}} \cdot V_{dd} \cdot D \]

Sensitivity to sizing and supply

♦ Gate sizing \(W_g\)

\[ -\frac{\partial E_{Sw}}{\partial W_i} = \frac{ec_i}{\tau_{nom} \cdot (h_{eff,i} - h_{eff,i-1})} \]

∞ for equal \(h_{eff}\) (\(D_{min}\))

♦ Supply voltage \(V_{dd}\)

\[ -\frac{\partial E_{Sw}}{\partial V_{dd}} = \frac{E_{Sw}}{D} \cdot 2 \cdot \frac{1-x_v}{\alpha_d - 1 + x_v} \]

\[ x_v = (V_{ou} + \Delta V_{th}) / V_{dd} \]
Sensitivity to $V_{th}$

- Threshold voltage ($V_{th}$)

\[
- \frac{\partial E}{\partial D} = P_{Lk} \cdot \left( \frac{V_{dd} - V_{on} - \Delta V_{th}}{\alpha_d \cdot V_0} - 1 \right)
\]

Low initial leakage
⇒ speedup comes for “free”