Transistor Sizing: How to Control the Speed and Energy Consumption of a Circuit

Jo Ebergen, Jonathan Gainsley, Paul Cunningham
Async Design Group
Sun Labs
Transistor Sizing: How to Control the Speed and Energy Consumption of a Circuit

Jo Ebergen, Jonathan Gainsley, Paul Cunningham
Asynchronous Design Group
Sun Labs
Introduction

- Transistor sizes (widths) determine
  - Speed of circuit
  - Energy consumption
  - Total area of circuit
  - Satisfaction of delay constraints

- Success or failure
How Do I Pick Transistor Widths?

- To optimize for speed?
- To optimize for energy?
- Automatically and quickly?
- Does a circuit have a speed limit?
- Is there a trade-off between speed and energy?
- How do I compare circuits built for different technologies?
An Example

• Given desired gate delays $s_0$, $s_1$, and $s_2$, fixed latch load $L$ and fixed wire load $W$
  • How do I find the sizes $x_0$, $x_1$, and $x_2$?
• Cycle time = $s_0 + s_1 + s_2$. What is minimum?
The Delay Model

• Defines relationship between gate sizes and delays
  • Capacitance driven by gate in time $s$
    = sum of all capacitances on node
      – $s = \text{gate delay}$
      – $x = \text{drive strength} [\text{capacitance/time}]$

\[
s_0 \times x_0 = C_{\text{dif}0} + C_L + C_{\text{in}1}
\]

• Input and diffusion capacitances are linear functions of drive strength [Idea of Logical Effort]
Units

• Unit of capacitance
  – $\kappa = \text{Input cap of min. sized inverter}$
  – All fixed loads must be converted

• Unit of delay $s$ (for stepup or slope)
  – $\tau = \text{Delay of ideal inverter, with no diffusion capacitance, driving copy of itself.}$
  – $\text{FO4} = 5\tau$

• Unit of drive strength $x$ (as in 4X, 8X)
  – $\kappa/\tau = \text{Capacitance per time unit}$
Technology Independence

- Units $\kappa$ and $\tau$ depend on technology
- Ex: TSMC 180nm, $\tau = 17$ps (FO4 = 85ps)
- Equations are independent of technology
- Allows comparisons of circuits in different technologies
- Warning: wire loads do not scale linearly
Logical Efforts

• Let me find gate capacitances as function of size $x$
• Input and diffusion capacitances are proportional to drive strength
• Logical Effort of input (LEin) = input capacitance per unit of drive strength
• Logical Effort of output (LEout) = diffusion capacitance per unit of drive strength

$C_{in0} = LE_{in0} \times x$

$C_{in1} = LE_{in1} \times x$

$C_{dif} = LE_{out} \times x$
Properties of Logical Effort

• Logical Effort = “effort” to compute logic function
• Logical effort is a time constant \([\kappa/(\kappa/\tau)]=\tau]\)
• \(\text{LE}_{\text{out}}\) = time to load diffusion capacitance of gate
  • parasitic delay
• \(\text{LE}_{\text{in}}\) = time to load input capacitance of gate (with the same drive strength)
• Logical efforts can be found from transistor diagram or empirically
Examples

- Some gates and their logical efforts
Back To Example

- Use delay model, now with Logical Efforts

\[
\begin{align*}
 s0 \cdot x0 &= 2x0 + 1x1 \\
 s1 \cdot x1 &= 1x1 + 1x2 + L \\
 s2 \cdot x2 &= \frac{4}{3}x0 + 1x2 + W \\
\end{align*}
\]

In Matrix form: \( S \cdot x = T \cdot x + C \)
In General

• $S^*x = T^*x + C$
• $S$ = diagonal matrix of gate delays
• $x$ = vector of drive strengths
• $T$ = logical effort matrix
  • $T_{ij} = 0$ no connection from gate $i$ to gate $j$
  • $T_{ij} \neq 0$ connection from gate $i$ to gate $j$
  • Describes topology of circuit
• $C$ = vector of fixed loads
• Equations can be extracted from netlist
Now What?

- Does a solution $x$ exist for given $S$, $T$, and $C$?
- How do I compute solution for $x$ efficiently?
- How to choose delays $S$?
- Special case: equal gate delays
  - Simpler model: $s^*x = T^*x + C$
  - Path delay = ($\#$ gates)$^*s$
  - Easy for satisfying delay constraints
  - More accurate
How To Compute Drive Strengths?

- Many ways to solve $s*x = T*x + C$
- Easiest is iteration
- Let $f(x) = (T*x + C)/s$
- $x(0) = 0$
- Repeat $x(i+1) = f(x(i))$ until convergence
- Converges quickly, if you choose $s$ well
An Example

\[ x_2 = \frac{(x_2 + 10 + \frac{4}{3}x_0)}{3} \]
\[ x_1 = \frac{(x_1 + 10 + x_2)}{3} \]
\[ x_0 = \frac{(2x_0 + x_1)}{3} \]

\[
\begin{align*}
x_2 &: 0 \quad 3.66 \quad 4.69 \quad 4.74 \quad .. \\
x_1 &: 0 \quad 4.89 \quad 5.23 \quad 5.25 \quad .. \\
x_0 &: 0 \quad 2.3 \quad 2.41 \quad 2.41 \quad ..
\end{align*}
\]
Critical Delay

- Only for circuits with cycles
  - Almost all async control ckts have cycles!
- Equal gate delay:  $s^x = T^x + C$
- Critical delay (cs) of circuit
  = largest real eigenvalue of $T$
- Feasible solution exists iff $s$ is larger than critical delay ($s > cs$)
- Critical delay is independent of fixed loads $C$
- Sizing algorithm converges if $s > cs$
Total size and critical delay

- Total size ($\sum x$’s) as function of gate delay
- Size grows as $C / (s - cs)$

![Diagram of logic gates with annotations](image)

![Graph showing Sum(x) versus Gate Delay](image)
Critical Delay and Limits

- Critical delay defines lower bound for gate delay, assuming equal gate delays
- Critical delay ($cs$) and # gate delays ($n$) in cycle define speed limit for throughput $= 1/(n*cs)$
Energy Estimation

• Dynamic Energy
  • Due to (dis)charging capacitance C
  • Proportional to $C \cdot V^2$

• Short-Circuit Energy
  • Due to crossover current
  • If input and output slope are equal, short-circuit energy
    $\approx \alpha \cdot$ dynamic energy consumption [Veendrick84]

• Static Energy
  • Due to leakage currents
  • Ignore for now
Units

• Unit of energy
  - $\varepsilon = \text{energy spent by ideal minimum inverter, excl. diffusion capacitance, driving a minimum inverter}$

• “Energy spent” = energy lost in resistors

• Unit $\varepsilon$ depends on technology, but equations do not

• Can be determined empirically
  - $\varepsilon = 2.9\text{fJ in TSMC 180nm, 1.8V.}$
More On Energy Estimation

- In equal-gate-delay model
- \( s^*x = T^*x + C \)
- Energy spent by gate \( \propto \) total output cap
- Energy spent by gate \( i \) \( \propto (T^*x+C)_i = s^*x_i \)
- Let \( p_i \) be activity index of gate \( i \) in an execution
- Total energy spent in an execution = \( \sum_i p_i s^*x_i \)
An Example: Magic Clock

- Every inverter has the same delay
- Control must charge and discharge load L
- \( E = 1 + \frac{2}{(s-2)} \) per unit load
- For equal delays, the best you can do
Comparing Circuits

• Independent of process technology
• Example: asynchronous controls of ripple FIFO
• How do different implementations compare in terms of energy versus performance?
The Implementations

• Chain of Rendezvous (COR):

• asP*:

• GasP:
More Implementations

- Berkel’s single-track handshake
- Singh & Nowick’s High-Capacity Pipeline
- IPCMOS by Schuster et al
- ....
- Magic clock
  - the lower bound and the ideal “synchronous” implementation.
## Critical Times

<table>
<thead>
<tr>
<th>Ckt</th>
<th>Critical Gate Delay</th>
<th># Gates in Cycle</th>
<th>Critical Cycle Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magic Clock</td>
<td>2.0</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>GasP</td>
<td>2.42</td>
<td>6</td>
<td>14.52</td>
</tr>
<tr>
<td>IBM IPCMOS 2(^1)</td>
<td>2.51</td>
<td>14</td>
<td>35.14</td>
</tr>
<tr>
<td>Singh-Nowick</td>
<td>3.38</td>
<td>8</td>
<td>27.04</td>
</tr>
<tr>
<td>asP*</td>
<td>3.95</td>
<td>8</td>
<td>31.6</td>
</tr>
<tr>
<td>Berkel’s single-track</td>
<td>4.21</td>
<td>6</td>
<td>25.26</td>
</tr>
<tr>
<td>Chain of C-elements</td>
<td>4.56</td>
<td>3(^3)</td>
<td>13.68</td>
</tr>
</tbody>
</table>
A Price/Performance Comparison

Normalized Energy Versus Gate Delay (Latch Loads Only)

- Magic Control
- Chain of C–elements
- 4–2 GasP
- asP*

Energy Per Cycle [Ω]
Gate Delay [τ]
A Price/Performance Comparison

Normalized Energy Versus Cycle Time (Latch Loads Only)

- GasP
- asP*
- Magic Control
- Chain of C–elements
It’s (Like) The Economy, Stupid

• Moving charges in circuit
  = Moving capital goods in economy
• Open input-output model
  • “Gate” = economic sector
  • “Capacitance” = demand for capital goods
  • “Drive strength” = supply of capital goods per time unit
  • “Energy” = total cost of capital goods
• Wassily Leontief (1906-1999)
• Has many applications
• Abundant literature
• A chip is like an economy
Summary

• Simple model for calculating transistor sizes
• Simple and efficient algorithms
• Good results obtained so far with equal gate delays
• **Critical delay** gives a price/performance characteristic for a control circuit
• Gives insight into speed-energy trade-offs
• Allows comparisons of circuits independent of process technology