Methods for True Power Minimization

Robert W. Brodersen¹, Mark A. Horowitz², Dejan Markovic¹, Borivoje Nikolic¹ and Vladimir Stojanovic²

¹Berkeley Wireless Research Center, UC Berkeley
²Stanford University
Power limited operation

Achieve the highest performance under the power cap
Power limited operation

Design optimization curves

Achieve the highest performance under the power cap
Power limited operation

Achieve the highest performance under the power cap
Power limited operation

How far away are we from the optimal solution?
Power limited operation

Design optimization curves

Global optimum – best performance
Power limited operation

Maximize throughput for given energy or
Minimize energy for given throughput
Design optimization

♦ There are many sets of parameters to adjust
♦ Tuning variables

- Circuit (sizing, supply, threshold)
- Logic style (domino, pass-gate, …)
- Block topology (adder: CLA, CSA, …)
- Micro-architecture (parallel, pipelined)
Design optimization

- There are many sets of parameters to adjust
- Tuning variables
  - Circuit (sizing, supply, threshold)
  - Logic style (domino, pass-gate, …)
  - Block topology (adder: CLA, CSA, …)
  - Micro-architecture (parallel, pipelined)

Globally optimal boundary curve: pieces of E-D curves for different topologies
Outline

♦ Circuit optimization
  ♦ Joint optimization
  ♦ Select the most promising sets of tuning variables

♦ Circuit & μArchitecture examples
  ♦ Adder
  ♦ Add-Compare

♦ Conclusions
Energy-delay sensitivity

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

- Proposed by Zyban at ISLPED02

\[ \Delta E = Sens(A) \cdot (\Delta D) + Sens(B) \cdot \Delta D \]

At the optimal point, all sensitivities should be the same
Alpha-power based delay model

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

♦ Fitting parameters
\[ V_{on}, \alpha_d, K_d \]
Alpha-power based delay model

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

- Fitting parameters
  - \( V_{on}, \alpha_d, K_d \)
- Effective fanout, \( h_{eff} \)
Energy model

♦ Switching energy

\[ E_{Sw} = \alpha_{0\rightarrow1} \cdot \left( C(W_{out}) + C(W_{par}) \right) \cdot V_{dd}^2 \]

♦ Leakage energy

\[ E_{Lk} = W_{in} \cdot I_0(S_{in}) \cdot e^{\frac{-\left(V_{th} - \gamma V_{dd}\right)}{V_0}} \cdot V_{dd} \cdot D \]
Sensitivity to sizing and supply

♦ Gate sizing ($W_i$)

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

♦ Supply voltage ($V_{dd}$)

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

\[ x_v = \frac{(V_{on} + \Delta V_{th})}{V_{dd}} \]

∞ for equal $h_{eff}$

($D_{min}$)
Sensitivity to Vth

- Threshold voltage ($V_{th}$)

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

Low initial leakage
⇒ speedup comes for “free”
Optimization setup

♦ Reference/nominal circuit
  • sized for $D_{min} @ V_{dd}^{nom}, V_{th}^{nom}$
  • known average activity

♦ Set delay constraint

♦ Minimize energy under delay constraint
  • gate sizing
  • $V_{dd}, V_{th}$ scaling
Kogge-Stone tree adder topology

- Off-path load (gates + wires)
- Reconvergence (inside ●-block)
Tree adder: Sizing optimization

Reference: all paths are critical

Nominal 
\((D_{\text{min}}, E_{\text{nom}})\)

Sizing opt. 
\((1.1D_{\text{min}}, 0.3E_{\text{nom}})\)

Internal energy peaks \(\Rightarrow\) Big savings for small delay penalty with resizing
Joint optimization: sizing and Vdd

\[ \Delta E = \text{Sens}(V_{dd}) \cdot (-\Delta D) + \text{Sens}(W) \cdot \Delta D \]
Results of joint optimization

Sensitivity table

<table>
<thead>
<tr>
<th>Sens</th>
<th>W</th>
<th>Vdd</th>
<th>Vth</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D_{nom}, E_{nom})</td>
<td>∞</td>
<td>1.5</td>
<td>0.2</td>
</tr>
<tr>
<td>(D_{nom}, E_{min})</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Energy efficient curve \( f(W, V_{dd}, V_{th}) \)

Nominal Design \( (D_{nom}, E_{nom}) \)

\( (D_{nom}, E_{min}) \)

80% of energy saved without delay penalty
Results of joint optimization

Sensitivity table

<table>
<thead>
<tr>
<th>Sens</th>
<th>W</th>
<th>Vdd</th>
<th>Vth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(D_{nom}, E_{nom})$</td>
<td>∞</td>
<td>1.5</td>
<td>0.2</td>
</tr>
<tr>
<td>$(D_{nom}, E_{min})$</td>
<td>1 (reference)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(D_{min}, E_{nom})$</td>
<td>22</td>
<td>16</td>
<td>22</td>
</tr>
</tbody>
</table>

Energy efficient curve $f(W,V_{dd},V_{th})$

Nominal Design $(D_{nom}, E_{nom})$

Delay $[D_{nom}]$

Energy $[E_{nom}]$

80% of energy saved without delay penalty

40% speedup for same energy
A look at tuning variables

Supply

Threshold

Limited range of tuning variables

\[ \text{Sens}(V_{dd}) = \text{Sens}(W) = \text{Sens}(V_{th}) = 1 \]
A look at tuning variables

Supply

Threshold

Limited range of tuning variables
Reducing the number of dimensions

Threshold and sizing nearly optimal around the nominal point
Scope of circuit optimization

Effective region +/-30% around nominal delay
Circuit & μArchitecture optimization

♦ Revisit the old argument for parallelism

(a) reference

$\frac{f}{2}$

(c) pipeline

(b) parallel

♦ What happens if we can choose optimal Vdd and Vth for each design?
Balance of leakage and switching energy

Optimal designs have high leakage current
Conclusions

♦ All design levels need to be optimized jointly
♦ Equal marginal costs ⇒ Energy-efficient design
♦ Peak performance is VERY power inefficient
♦ Today’s designs are not leaky enough to be truly power-optimal
♦ Pipelining starts to gain advantage over parallelism