A Static Power Model for Architects

J. Adam Butts and Guri Sohi
University of Wisconsin-Madison
{butts,sohi}@cs.wisc.edu

33rd International Symposium on Microarchitecture
Monterey, California
December, 2000
Overview

The static power problem

- Leakage current
- Scaling trends

A static power model: $P_{\text{static}} = V_{\text{CC}} \cdot I_{\text{leak}} \cdot N \cdot k_{\text{design}}$

Attacking static power

- Power gating
- Using slower devices
- Applying speculation

Conclusion
A CMOS Gate

P transistor

V_{cc}

Input

Output

Load capacitance

N transistor
Sources of Power Consumption

Dynamic

C $\frac{dV}{dt}$ (charging of capacitive load)
Sources of Power Consumption

Dynamic

$I_{\text{short-circuit}}$ (both devices conducting)
Sources of Power Consumption

Static

\[ I_{\text{leakage}} \text{ (subthreshold, junction leakage)} \]
Technology Scaling

Dimensions reduced to increase performance and density

$V_{CC}$ decreases each generation...
- Limit dynamic power
- Limit electric fields

...requiring lower $V_T$
- Gate overdrive $= V_{CC} - V_T$

Leakage increases exponentially
- $P_{static} = V_{CC} I_{leak} \sim \exp (-V_T)$
Static Power Projections

Static power is an increasing fraction of total power

Today: Pentium III 1.13 GHz
- \( P_{\text{total}} \) (peak) = 41.4 watts
- \( P_{\text{static}} = 5.4 \) watts
- Static power is 13% of total
- Higher contribution on average

This is only getting worse
- \( P_{\text{static}} = P_{\text{dynamic}} \) in 3 generations
Important Characteristics of Static Power

1. Exponentially increasing due to $V_T$ scaling
   - Increasing faster than dynamic power

2. Adds to average power, not peak power
   - More expensive than dynamic power

3. Independent of transistor utilization
   - Transistors are not free
Model Derivation

Want an equivalent of $C \cdot V_{CC}^2 \cdot f$ for static power

Develop model from the bottom-up

- Lack of data precludes a top-down “data-driven” approach
- Start from BSIM3v3.2 transistor model

$$I_{D_{sub}} = I_{s0} \cdot \frac{W}{L} \cdot \left(1 - e^{\frac{-V_{ds}}{V_t}}\right) \cdot e^{\frac{V_{gs} - V_T - V_{off}}{n \cdot V_t}}$$

BSIM3 model eq.

Aspect ratio

$V_T$ dependence

A Static Power Model for Architects - J. Adam Butts and Guri Sohi
33rd International Symposium on Microarchitecture, December 2000
Apply BSIM to a single “off” (leaking) device
Model Derivation

\[
I_{D\text{sub}} = I_{s0}' \cdot \frac{W}{L} \cdot \left(1 - \frac{-V_{ds}}{V_t}\right) \cdot \frac{V_{gs} - V_T - V_{off}}{n \cdot v_t}
\]

Abstracted equation for a single device

\[
I_{D\text{sub}} = \hat{I}_{\text{leak}} \cdot \frac{W}{L}
\]

Group technology-dependent parameters together
Apply to large numbers of devices

Only devices which are off contribute to leakage

Average aspect ratio

Account for lower leakage of stacked devices

Model Derivation
Model Derivation

\[ I_{\text{leak}} = \hat{I}_{\text{leak}} \cdot N \cdot \left( \frac{W}{L} \right)_{\text{avg}} \cdot f_{\text{off}} \cdot f_{\text{stack}} \]

Group design-dependent parameters together
Static Power Model

Resulting power model has four parameters

- Technology-dependent (from scaling, process data)
- Design-dependent (from estimates, past designs)

\[ P_{\text{static}} = V_{CC} \cdot I_{\text{leak}} \cdot N \cdot k_{\text{design}} \]
The Design Constant

Represents an “average” device
- Aspect ratio (device size)
- Fraction of leaking devices
- Stacking factor

Depends on design style

Independent of technology → Allows for forward projection

![Graph showing the design constant vs. channel length](image)

- Datapath (adder)
- Associative (1 RW, 1 CAM port)
- SRAM (6T)
Attacking Static Power

Power reduction techniques address factors in the model equation:

\[ P_{\text{static}} = V_{\text{CC}} \cdot I_{\text{leak}} \cdot N \cdot k_{\text{design}} \]

Use power aware microarchitecture

- Use fewer devices
- Power gating

Employ slow devices

- Enables supply voltage reduction (voltage partitioning)
- Enables use of higher threshold voltage devices
Power Gating

Eliminate leakage by removing power to unused devices
- Analogous to clock gating
- Requires logic to determine power down/up conditions

Many power gating possibilities
- Floating point hardware
- Rare instruction decode logic
- Interrupt handling hardware

Power-up prediction problem
- Large decoupling capacitance
- Limited charging current & dI/dt
  → Several cycles of power-up latency
Speculative Power Gating

Power-up latency limits power gating potential

1. Do not gate power (no power savings)
2. Accept power-up latency (lower performance)
3. Build predictor for power-up condition

Adjustable misprediction penalties

- Power/performance bias

Sample Applications

- PC based prediction for special instruction needs
- PC based prediction for L1 miss handler (L1-L2 interface)
Using Slower Devices

Trade latency and area for power
- 2× devices at 0.5× frequency
  → Equivalent throughput with higher latency and lower total power

Reducing clock frequency helps only dynamic power
- Multiple threshold voltage technology (multiple frequency domains)
- Variable supply voltage (multiple supply voltage domains)

Architectural Issues
- Interdomain communication
- Latency tolerance
Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong

→ Average latency is decreased
Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong

→ Average latency is decreased

<table>
<thead>
<tr>
<th>Operation Latency</th>
<th>Relative Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without speculation</td>
<td>4</td>
</tr>
</tbody>
</table>
Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong

→ Average latency is decreased

<table>
<thead>
<tr>
<th></th>
<th>Operation Latency</th>
<th>Speculation Latency</th>
<th>Effective Latency</th>
<th>Relative Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without speculation</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance speculation</td>
<td>4</td>
<td>2</td>
<td>2.7</td>
<td>6</td>
</tr>
</tbody>
</table>

A Static Power Model for Architects - J. Adam Butts and Guri Sohi
33rd International Symposium on Microarchitecture, December 2000
Using Slower Devices with Speculation

Speculation is a latency tolerance technique
- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
  → Average latency is decreased

Use slower devices to save power and speculation to tolerate increased latency

<table>
<thead>
<tr>
<th></th>
<th>Operation Latency</th>
<th>Speculation Latency</th>
<th>Effective Latency</th>
<th>Relative Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without speculation</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance speculation</td>
<td>4</td>
<td>2</td>
<td>2.7</td>
<td>6</td>
</tr>
</tbody>
</table>

A Static Power Model for Architects - J. Adam Butts and Guri Sohi
33rd International Symposium on Microarchitecture, December 2000
Using Slower Devices with Speculation

Speculation is a latency tolerance technique
- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
→ Average latency is decreased

Use slower devices to save power and speculation to tolerate increased latency

<table>
<thead>
<tr>
<th></th>
<th>Operation Latency</th>
<th>Speculation Latency</th>
<th>Effective Latency</th>
<th>Relative Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without speculation</td>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Performance speculation</td>
<td>4</td>
<td>2</td>
<td>2.7</td>
<td>6</td>
</tr>
<tr>
<td>Power speculation</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

A Static Power Model for Architects - J. Adam Butts and Guri Sohi
33rd International Symposium on Microarchitecture, December 2000
Conclusions

Static power will become important ($V_T$ scaling)

A high-level model is available: $P_{\text{static}} = V_{CC} \cdot I_{\text{leak}} \cdot N \cdot k_{\text{design}}$

Reducing static power also reduces dynamic power

Speculation as a power savings technique
- Speculative power gating
- Allows use of slower devices with controlled performance penalty

What can architects do to impact static power dissipation?
- Latency/throughput tradeoffs
- Design partitioning (voltage/frequency domains)
- Identify idle resources, predict the need for them
- Identify opportunities for power speculation