Low-Power Design for FPGAs

Tim Tuan
Xilinx Inc.
8/8/2005

Outline

• FPGA Power
  – Power Trends
  – Power Models
  – Power Analysis

• Architecture Innovations
  – Dedicated Cores
  – Low Leakage Programmable Interconnect
  – Power Gating
  – Heterogeneous Architecture
Traditional FPGA

Modern FPGAs Architectures

- 10-Gigabit Transceivers
- Digital Clock Manager
- CLBs
- PowerPC
- High Performance (Virtex-4 FX140)
- System Monitor, ADC
- Block RAMs, multipliers
- DSP48
- Low Cost (Spartan3e 3S1200E)
FPGA Power Trends

• Largest device of each generation shown

- Dynamic Power
- Static Power

• Max static power
• Dynamic power is design dependent; typical shown

Assuming Vdd, Vth, frequency, capacity all continue to scale at the same rate

- If past trends continue ...

ISLPED 2005 Tutorial
FPGA Power Trends

What does the ITRS Roadmap (2004) say?

Key limits:
- Fixed die size
- Slower product introduction
- Slower voltage scaling

Lower power is very desirable in FPGAs, but must be carefully traded off against performance and cost.

Lower power is very desirable in FPGAs, but must be carefully traded off against performance and cost.

Power Analysis Flow

Resource Characterization

Power Analysis
CLB Architecture Model

Configurable Logic Block (CLB)

Configuration SRAM - Largest Virtex-4 has >50Mb

Clock Tree Model

Input pin

ClnRoot

ClnCol

up to 8 rows
Power Breakdown

• Based on typical resource utilization, 12.5% switching activity, $f_{CLK}=100\text{MHz}$

<table>
<thead>
<tr>
<th>Block</th>
<th>% of Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic Slices</td>
<td>67%</td>
</tr>
<tr>
<td>LUT</td>
<td>47%</td>
</tr>
<tr>
<td>FF</td>
<td>26%</td>
</tr>
<tr>
<td>Carry</td>
<td>4%</td>
</tr>
<tr>
<td>I/O BAR</td>
<td>40%</td>
</tr>
<tr>
<td>Direct</td>
<td>27%</td>
</tr>
<tr>
<td>Double flip</td>
<td>15%</td>
</tr>
<tr>
<td>Hex</td>
<td>5%</td>
</tr>
<tr>
<td>Long</td>
<td>1%</td>
</tr>
<tr>
<td>CLK Rout</td>
<td>16%</td>
</tr>
<tr>
<td>CLK Mux</td>
<td>16%</td>
</tr>
<tr>
<td>CLK Select</td>
<td>4%</td>
</tr>
<tr>
<td>I/O BAR Ctrl</td>
<td>36%</td>
</tr>
</tbody>
</table>

Typical resource utilization

Dynamic Power Breakdown

• Based on typical resource utilization, 12.5% switching activity, $f_{CLK}=100\text{MHz}$

90nm Spartan-3 Logic Core

[Degalahal ASPDAC 2005]
Leakage Power Breakdown

- A few basic circuits consume the majority of total leakage power

![Leakage Power Breakdown Diagram]

Are FPGA Power-Inefficient?

- Depends on what you compare to...
- Case study: JPEG2000 encoder

![JPEG2000 Encoder Diagram]

<table>
<thead>
<tr>
<th></th>
<th>Power Dissipation (in Watts)</th>
<th>Relative MOPS/mW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device</td>
<td>Memory</td>
<td>Total</td>
</tr>
<tr>
<td>FPGA</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>DSP</td>
<td>1.7</td>
<td>2.4</td>
</tr>
<tr>
<td>ASIC</td>
<td>2.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Dedicated Blocks

- Customs logic is more energy-efficient than programmable logic
- Introduce custom logic cores for common functions
  - Virtex-4 optimizes for DSP applications with DSP48 custom cores

![DSP48 blocks](image)

### DSP48 vs. MULT18x18+Logic Cells

<table>
<thead>
<tr>
<th>Frequency (MHz)</th>
<th>Power (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
</tr>
<tr>
<td>300</td>
<td>80</td>
</tr>
<tr>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>500</td>
<td>120</td>
</tr>
<tr>
<td>600</td>
<td>140</td>
</tr>
</tbody>
</table>

- Virtex-II ~47 mW/100MHz
- Virtex-II Pro ~40 mW/100MHz
- Simulated Virtex-4 ~5.7 mW/100MHz

Low Leakage Interconnect

**Triple-Oxide**

- Use low-speed, low-power configuration SRAM
  - Not timing critical
  - Also used in interconnect passgates with higher gate boosting
- High-Vt is not enough in 90nm due to substantial gate leakage
  - Virtex-4 uses a mid-oxide device
  - 40% leakage reduction, no leakage increase from 130nm (Virtex2-Pro)
- Reduces much of the overhead of reconfigurability with no performance penalty
Power Gating

- In MPUs, use a high-Vt device to cut off power to inactive blocks
  - 20-2000X leakage reduction during sleep for <10% delay and area penalty
- In FPGAs, applicable to inactive and *unused* blocks
  - Each PG block may be one or more CLBs, or sub-CLB resources (e.g. LUT)

Design Issues

- Inputs, outputs often require interfacing logic
  - Relatively less overhead for larger blocks

*Coarser-grain power gating requires less overhead*
**Design Issues**

- **Power gates must be sized for peak current**
  - Larger blocks have less power gate overhead, as they typically have lower peak current relative to block size
  - E.g.: Block B has roughly the same power gate requirement as Block A

  ![Block A](image1)

  ![Block B](image2)

  **Coarser-grain power gating requires less overhead**

**CAD for Power Gating**

- Power saving may be limited by coarse-grain power gating
- Region-constrained placement (RCP) circumvents that limitation
  - Use placement constraints to group “used logic”

  ![Diagram of CAD flow](image3)

  ![Group logic with same temporal profile for dynamic power gating](image4)

  *Gayasen, FPGA ’04*
CAD for Power Gating

- RCP enables high power savings in coarse grain power gating architectures

![Graph showing leakage power savings with and without RCP]

Heterogeneous Architectures

- Heterogeneity addresses the over-design inherent in homogeneous FPGAs
  - ASICs use slack timing: Use fast cells for critical paths
  - In FPGAs, critical paths are unknown when designing silicon, hence everything is designed for high performance

- Heterogeneous FPGA
  - Mix fast and slow (but low power) resources. Use CAD tools to map timing critical logic to fast resources
Heterogeneous Interconnect

- Mixed high-speed and low-power interconnect resources
- Use router to perform slack timing

![Heterogeneous Interconnect Diagram]

- Also can be applied to logic blocks

CAD for Heterogeneous Arch

- Resize some interconnect muxes to consume less power
- Use standard timing-driven place and route tools
  - Critical paths should naturally be mapped to fast resources
  - 40% leakage reduction in the interconnect

<table>
<thead>
<tr>
<th>Architecture: Low power resources</th>
<th>Power</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>all imux, longs, 25% double, hex, 50% omux</td>
<td>-30.1%</td>
<td>-4.0%</td>
</tr>
<tr>
<td>all long, 75% imux, 25% double, hex, 50% omux</td>
<td>-26.0%</td>
<td>-3.5%</td>
</tr>
<tr>
<td>all imux, long, 75% double, hex, 50% omux</td>
<td>-41.5%</td>
<td>-6.8%</td>
</tr>
<tr>
<td>all imux, long, hex, 50% double, omux</td>
<td>-40.8%</td>
<td>-11.6%</td>
</tr>
</tbody>
</table>
CAD for Heterogeneous Arch

- Potential CAD improvements
  - Post-routing optimization: After PAR, re-assign non-critical nets to low power resources, if possible
  - Modify cost functions to include timing and power

Summary

- FPGAs trend towards higher power, but power reduction must be carefully weighed against performance and cost
- In 90nm Spartan-3 FPGA, dynamic power is still dominant (70%-84% of total core power)
- Dedicated blocks improve power efficiency at the cost of flexibility
- Triple-oxide process offers an effective, one-time solutions
- Power gating is effective for reducing static and standby power, especially with CAD support
- Heterogeneous architectures allow slack timing for power reduction with proper CAD support