An Asynchronous FPGA with Two-Phase Enable-Scaled Routing

Christopher LaFrieda, Ben Hill, and Rajit Manohar
Computer Systems Laboratory
Cornell University
Outline

- Asynchronous FPGA Architecture
- Two Phase Routing
  - Supporting circuitry.
- Enable Scaling
  - Supporting circuitry.
  - Loops and reconvergent paths.
  - Choosing Vddl.
- Results
- Conclusions
Asynchronous FPGA Architecture: Tile Fabric
FPGA Architecture: Configurable Logic Block (CLB)
FPGA Architecture: Switch Box (SB)
FPGA Architecture: 4:4 Switch

[Diagram showing the FPGA architecture with labeled components: L[0-3].f, L[0-3].e, L[0-3].t, C, PC4, WCHB Buffer, R[0-3].f, R[0-3].e, R[0-3].t.]
Interconnect Protocol

- Four-phase interconnect is power hungry.
- What about 1-of-4 codes? Bundled-data?
  - Hurts applications that do not have regular datapaths.
  - Makes software much more complex.
- Two-phase interconnect:
  - Half the switching of four-phase.
  - Requires minimal changes to implement.
Two-Phase Interconnect: CLB
Two-Phase Interconnect:
4:4 Switch

undersized to save
area
Enable Scaling Interconnect

- Voltage scaling enable signals:
  - Forward latency remains fixed.
  - Trade off: lower power for longer backward latency.

- Scaling enable voltage may not impact performance:
  - This is true when not operating at peak throughput.
  - FPGA user designs never operate near peak throughput.

- Circuits support enable scaling specific paths:
  - There is one global low Vdd (Vddl).
  - Switches can be configured to use Vdd or Vddl.
Enable Scaling: 4:4 Switch

can choose Vdd or Vddl
Enable Scaling: Pipeline Dynamics (two-phase)

\[ \tau = l_f + l_b \]
Enable Scaling: Pipeline Dynamics (two-phase)

\[ \tau = l_f + l_b \]

Token Tail \hspace{2cm} Token Head

\[ T = \frac{1}{\tau} \]
Enable Scaling: Loops

- Ideal Loop (2.5 GHz)
- Hole-Limited Loop (2.08 GHz)
- Token-Limited Loop (1.25 GHz)

Common in FPGA user designs.
Enable Scaling: Loop Throughput

Tokens vs. Throughput

Throughput (GHz)

$k$

$n - k$

$n_l_f$

$n_l_b$

kTokens
Enable Scaling: Loop Throughput

Tokens vs. Throughput

Throughput (GHz)

k

nl_f

n - k

nl_b

nl'_b

Nominal

Enable scaling

kTokens

0 1 2 3 4 5 6 7 8
Enable Scaling:
Enable Scaled Loops

Token-Limited Loop (1.25 GHz) → Enable-Scaled Loop (1.25 GHz)
Enable Scaling: Reconvergent Paths
Enable Scaling: Reconvergent Path Throughput
Enable Scaling: Reconvergent Path Throughput
Digression: Two-Phase Improves Performance
Enable Scaling:
Finding $L_b'$

Loops:

$$l_b' = \frac{l_f(n-k)}{k}$$

Reconvergent Paths:

$$l_b' = \frac{n_l l_f}{n_s} + l_b - l_f$$
Enable Scaling: Finding $V_{ddl}$
### Results: MCNC Benchmarks

<table>
<thead>
<tr>
<th>Name</th>
<th>Array Size</th>
<th>LUT Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigkey</td>
<td>36x36</td>
<td>1707</td>
</tr>
<tr>
<td>clma</td>
<td>47x47</td>
<td>8383</td>
</tr>
<tr>
<td>diffeq</td>
<td>20x20</td>
<td>1497</td>
</tr>
<tr>
<td>dsip</td>
<td>36x36</td>
<td>1370</td>
</tr>
<tr>
<td>elliptic</td>
<td>31x31</td>
<td>3604</td>
</tr>
<tr>
<td>frisc</td>
<td>30x30</td>
<td>3556</td>
</tr>
<tr>
<td>s38584.1</td>
<td>41x41</td>
<td>6447</td>
</tr>
<tr>
<td>tseng</td>
<td>17x17</td>
<td>1047</td>
</tr>
</tbody>
</table>
Results: Performance Gains
Results: Performance Gains

- Bigkey: +40%
- dma: +40%
- diffeq: +70%
- disp: +70%
Results: Performance Gains

![Bar chart showing frequency gains for different benchmarks across 4-Phase, 2-Phase, and 2-Phase ES phases.]
Results:
Power Savings

Normalized Power

-15%
-30%
+3%
Results: Power Savings

<table>
<thead>
<tr>
<th>Application</th>
<th>4-Phase</th>
<th>2-Phase</th>
<th>2-Phase ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigkey</td>
<td>-38%</td>
<td>-20%</td>
<td>-30%</td>
</tr>
<tr>
<td>dma</td>
<td>-57%</td>
<td>-25%</td>
<td>-35%</td>
</tr>
<tr>
<td>diffq</td>
<td>-56%</td>
<td>-22%</td>
<td>-32%</td>
</tr>
<tr>
<td>dsp</td>
<td>-43%</td>
<td>-18%</td>
<td>-30%</td>
</tr>
<tr>
<td>elliptic</td>
<td>-26%</td>
<td>-15%</td>
<td>-25%</td>
</tr>
<tr>
<td>frisc</td>
<td>-57%</td>
<td>-20%</td>
<td>-30%</td>
</tr>
<tr>
<td>s3858.1</td>
<td>-56%</td>
<td>-22%</td>
<td>-32%</td>
</tr>
<tr>
<td>tseng</td>
<td>-56%</td>
<td>-18%</td>
<td>-30%</td>
</tr>
</tbody>
</table>
Conclusions

- Two-phase interconnect:
  - Needs converters in the CLB and two-phase buffers in the SB.
  - Reduces power consumption by ~40%.
  - Improves performance in reconvergent path limited designs by 40-70%.

- Enable-scaled interconnect:
  - Needs virtual Vdds and voltage converters at the 4:4 switches.
  - Enable scaling reduces power consumption by ~30%.

- Methods complement each other. Together they increase area by only 12% and reduce power by up to 57%.
Thank You

And, don’t forget to vote for best paper!