CS250
VLSI Systems Design

Fall 2020

John Wawrzynek

with

Arya Reais-Parsi
FPGA Overview

- Basic idea: two-dimensional array of logic blocks and flip-flops with a means for the user to configure (program):
  1. the interconnection between the logic blocks,
  2. the function of each block.

Simplified version of FPGA internal architecture (not scalable)
Reconfigurable Fabric Architecture: Degrees of Freedom

1. Logic Blocks
   - Capacity and internal structure of combination logic circuits and state element(s),
   - Clustering and internal interconnect

2. Interconnection Network Architecture
   - Circuit-switched not packet-switched,
   - Topology of network

3. Configuration Architecture
   - How is programming information loaded and distributed,
   - Configuration “depth”

4. Hard blocks: RAM, ALUs, Processor Cores, ...
   - Function(s), count, and how integrated into the fabric
Colors represent different types of resources:

- Logic
- Block RAM
- DSP (ALUs)
- Clocking
- I/O
- Serial I/O + PCI

A routing fabric runs throughout the chip to wire everything together.
Configurable Logic Blocks (CLBs)

*Slices define regular connections to the switching fabric, and to slices in CLBs above and below it on the die.*
Primitive: 5-input Look Up Tables (LUTs)

Computes any 5-input logic function.

Timing is independent of function.

Latches set during configuration.
Virtex 6-LUTs: Composition of 5-LUTs

May be used as one 6-input LUT (D6 out) ...

... or as two 5-input LUTS (D6 and D5)
The simplest view of a slice

- Four 6-LUTs
- Four Flip-Flops
- Switching fabric may see combinational and registered outputs.

An actual Virtex slice adds many small features to this simplified diagram. We show them one by one ...
Two 7-LUTs per slice ...

Extra multiplexers (F7AMUX, F7BMUX)

Extra inputs (AX and CX)
Or one 8-LUTs per slice ...

Third multiplexer (F8MUX)

Third input (BX)
ExtramuxestochoseLUToption...

From eight 5-LUTs
...to one 8-LUT.

Combinational
or registered outs.

Flip-flops unused by
LUTs can be used
standalone.
Virtex “Vertical” Logic

We can map ripple-carry addition onto carry-chain block.

The carry-chain block also useful for speeding up other adder structures and counters.
Putting it all together ... a SLICEL.

The previous slides explain all SLICEL features.

About 50% of the are SLICELs.

The other slices are SLICEMs, and have extra features.
Recall: 5-LUT architecture ...

32 Latches. Configured to 1 or 0.

Some parts of a logic design need many state elements.

SLICEMs replace normal 5-LUTs with circuits that can act like 5-LUTs, but can alternatively use the 32 latches as RAM, ROM, shift registers.
A SLICEM 6-LUT ...

- Memory data input
- Normal 6-LUT inputs.
- Memory write address
- Normal 5/6-LUT outputs.
- Memory data input.
- Control output for chaining LUTs to make larger memories.
- Synchronous write / asynchronous read
SLICEL vs SLICEM ...

SLICEL

SLICEM

SLICEM adds memory features to LUTs, + muxes.
Distributed RAM Primitives

All are built from a single slice or less.

Remember, though, that the SLICEM LUT is naturally only 1 read and 1 write port.
Configurable Interconnect

- Design Challenges (topology):
  - traversing long wires incurs delay and energy
  - switches (transistors) add significant delay
  - Mapping time

Switch matrix could be more richly populated

“connection block”
Xilinx FPGAs (tile interconnect detail)
Other Topologies

- Traditional:
  - From flexlogic, Inc.

- Clos Network
  - "uses about half the area of the traditional interconnect and uses only 5-7 metal routing layers"
Fat-Tree Based Interconnect

- Use “Rent’s rule” for proper thickness

**HSRA:**
High-Speed, Hierarchical Synchronous Reconfigurable Array

William Tsu, Kip Macy, Atul Joshi, Randy Huang
Norman Walker, Tony Tung, Omid Rowhani, Varghese George
John Wawrzynek, and André DeHon
Berkeley Reconfigurable, Architectures, Software, and Systems
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720-1776
Embedded Hard Blocks

- Many important functions are not efficient when implemented in the reconfigurable fabric:
  - multiplication, large memory, processor cores, ...

- Dedicated blocks take relatively little area and therefore could go unused.
Xilinx Virtex-5

Colors represent different types of resources:

- Logic
- Block RAM
- DSP (ALUs)
- Clocking
- I/O
- Serial I/O + PCI

A routing fabric runs throughout the chip to wire everything together.
Virtex DSP48E Slice

Efficient implementation of multiply, add, bit-wise logical.
Block RAM Overview

- 36K bits of data total, can be configured as:
  - 2 independent 18Kb RAMs, or one 36Kb RAM.

- Each 36Kb block RAM can be configured as:
  - 64Kx1 (when cascaded with an adjacent 36Kb block RAM), 32Kx1, 16Kx2, 8Kx4, 4Kx9, 2Kx18, or 1Kx36 memory.

- Each 18Kb block RAM can be configured as:
  - 16Kx1, 8Kx2, 4Kx4, 2Kx9, or 1Kx18 memory.

- Write and Read are synchronous operations.

- The two ports are symmetrical and totally independent (can have different clocks), sharing only the stored data.

- Each port can be configured in one of the available widths, independent of the other port. The read port width can be different from the write port width for each port.

- The memory content can be initialized or cleared by the configuration bitstream.
# Ultra-RAM Blocks

<table>
<thead>
<tr>
<th>Feature</th>
<th>Block RAM</th>
<th>UltraRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clocking</td>
<td>Two clocks</td>
<td>Single clock</td>
</tr>
<tr>
<td>Built-in FIFO</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Data width</td>
<td>Configurable (1, 2, 4, 9, 18, 36, 72)</td>
<td>Fixed (72-bits)</td>
</tr>
<tr>
<td>Modes</td>
<td>SDP and TDP</td>
<td>Two ports, each can independently read or write (a superset of SDP)</td>
</tr>
<tr>
<td>ECC</td>
<td>64-bit SECDED</td>
<td>64-bit SECDED&lt;br&gt;Supported in 64-bit SDP only (one ECC decoder for port A and one ECC encoder for port B)</td>
</tr>
<tr>
<td>Cascade</td>
<td>• Cascade output only (input cascade implemented via logic resources) &lt;br&gt;• Cascade within a single clock region</td>
<td>• Cascade both input and output (with global address decoding) &lt;br&gt;• Cascade across clock regions in a column &lt;br&gt;• Cascade across several columns with minimal logic resources</td>
</tr>
<tr>
<td>Power savings</td>
<td>One mode via manual signal assertion</td>
<td>One mode via manual signal assertion</td>
</tr>
</tbody>
</table>

**Figure 2-1:** UltraRAM URAM288_BASE Primitive
# State-of-the-Art - Xilinx FPGAs

<table>
<thead>
<tr>
<th>Device Name</th>
<th>45nm</th>
<th>28nm</th>
<th>20nm</th>
<th>16nm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPARTAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System Logic Cells (K)</td>
<td>862</td>
<td>1,314</td>
<td>1,724</td>
<td>2,586</td>
</tr>
<tr>
<td>CLB Flip-Flops (K)</td>
<td>788</td>
<td>1,201</td>
<td>1,576</td>
<td>2,364</td>
</tr>
<tr>
<td>CLB LUTs (K)</td>
<td>394</td>
<td>601</td>
<td>788</td>
<td>1,182</td>
</tr>
<tr>
<td>Max. Dist. RAM (Mb)</td>
<td>12.0</td>
<td>18.3</td>
<td>24.1</td>
<td>36.1</td>
</tr>
<tr>
<td>Total Block RAM (Mb)</td>
<td>25.3</td>
<td>36.0</td>
<td>50.6</td>
<td>75.9</td>
</tr>
<tr>
<td>UltraRAM (Mb)</td>
<td>90.0</td>
<td>132.2</td>
<td>180.0</td>
<td>270.0</td>
</tr>
<tr>
<td>HBM DRAM (GB)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HBM AXI Interfaces</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Clock Mgmt Tiles (CMTs)</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>DSP Slices</td>
<td>2,280</td>
<td>3,474</td>
<td>4,560</td>
<td>6,840</td>
</tr>
<tr>
<td>Peak INT8 DSP (TOP/s)</td>
<td>7.1</td>
<td>10.8</td>
<td>14.2</td>
<td>21.3</td>
</tr>
<tr>
<td>PCIe® Gen3 x16</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>PCIe Gen3 x16/Gen4 x8 / CCIX(1)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>150G Interlaken</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>100G Ethernet w/ KR4 RS-FEC</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Max. Single-Ended HP I/Os</td>
<td>520</td>
<td>832</td>
<td>832</td>
<td>832</td>
</tr>
<tr>
<td>GTY 32.75Gb/s Transceivers</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>120</td>
</tr>
<tr>
<td>GTM 58Gb/s PAM4 Transceivers</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>100G / 50G KP4 FEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) For Virtex Ultra-scale
(2) Industrial

---

**Virtex Ultra-scale**
Configuration Architecture

- How are the programming bits loaded and distributed?
- Configuration depth (number of stored on-chip configurations)
- Same interface often can provide read-back to save state / debug
- Design Challenge:
  - Configurations are very large (100’s of Mbits)
  - Moving many bits over chip interface requires time and energy

Many commercial FPGAs also have an internal reconfiguration controller that allows dynamic self reconfiguration.

![Diagram of 7 Series FPGA Master BPI Configuration Interface - Asynchronous Read Example](image-url)
Internal Reconfiguration

- Traditionally, long shift chains:
  - slow, relatively energy efficient
  - “Random access” structures have been tried.
  - permits fine-grain partial reconfiguration

Connections to logic blocks, programmable interconnection points, ...
Xilinx Configuration Layout

- “frame” is unit of reconfiguration
- serially loaded into chip

![Diagram of Xilinx Configuration Layout]
Multi-context FPGAs

A Time-Multiplexed FPGA

Steve Trimberger, Dean Carberry, Anders Johnson, Jennifer Wong

Xilinx, Inc.
2100 Logic Drive
San Jose, CA 95124
408-559-7778

steve.trimberger@xilinx.com

Figure 1. Time-Multiplexed FPGA Configuration Model

Garp: a MIPS processor with a reconfigurable coprocessor

Published 1997
Proceedings. The 5th Annual IEEE Symposium on Field-Programmable Custom Computing Machines

Rapid dynamic reconfiguration possible.

What’s the execution and programming model?
End of Lecture 4