A 45nm 1.3GHz 16.7 Double-Precision GFLOPS/W RISC-V Processor with Vector Accelerators

Yunsup Lee¹, Andrew Waterman¹, Rimas Avizienis¹, Henry Cook¹, Chen Sun¹,², Vladimir Stojanovic¹,², Krste Asanovic¹

¹University of California, Berkeley
²Massachusetts Institute of Technology
Upheaval in Computer Design

- Energy efficiency constrains everything
  - Incorporate specialized and heterogeneous accelerators into general-purpose processors
  - Write processor generators to express a design space, and do vertically-integrated design space exploration extensively

Source
[1] Why migration to 20nm bulk CMOS and 16/14nm FINFETS is not the best approach for semiconductor industry, IBS, Handel Jones, 2014.
3mm X 6mm Chip
Fabricated in 45nm SOI
75m+ transistors

Dual-Core RISC-V Processor with Vector Accelerators

1MB SRAM Memory Structure for Testing

Monolithically-Integrated Silicon Photonic Links
Transmitter: Wade OFC’14
Receiver: Georgas VLSI’14
Chip Architecture

Dual-Core RISC-V Vector Processor

1MB SRAM Array

Core Logic

VRF

L1D$

Core

L1I$

L1VI$

Rocket Scalar Core

Hwacha Vector Accelerator

16K L1I$

32K L1D$

8KB L1VI$

Arbiter

Rocket Scalar Core

Hwacha Vector Accelerator

16K L1I$

32K L1D$

8KB L1VI$

Arbiter

Coherence Hub

FPGA FSB/HTIF

1MB SRAM Array
• RISC-V is a new, open, and completely free general-purpose ISA
  – Developed at UC Berkeley
• RISC-V designed to be flexible and extensible
  – Better integrate accelerators with host cores
• RISC-V software ecosystem
  – binutils, GCC, Newlib, glibc, GDB, LLVM, Linux, QEMU
• External users contributing to ecosystem
Rocket Scalar Core

- 64-bit 6-stage single-issue in-order pipeline
- Design minimizes impact of long clock-to-output delays of compiler-generated RAMs
- 64-entry BTB, 256-entry BHT, 2-entry RAS
- MMU supports page-based virtual memory
- IEEE 754-2008-compliant FPU
  - Supports SP, DP FMA with hw support for subnormals
# ARM Cortex-A5 vs. RISC-V Rocket

<table>
<thead>
<tr>
<th>Category</th>
<th>ARM Cortex-A5</th>
<th>RISC-V Rocket</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISA</td>
<td>32-bit ARM v7</td>
<td>64-bit RISC-V v2</td>
</tr>
<tr>
<td>Architecture</td>
<td>Single-Issue In-Order</td>
<td>Single-Issue In-Order 6-stage</td>
</tr>
<tr>
<td>Performance</td>
<td>1.57 DMIPS/MHz</td>
<td>1.72 DMIPS/MHz</td>
</tr>
<tr>
<td>Process</td>
<td>TSMC 40GPLUS</td>
<td>TSMC 40GPLUS</td>
</tr>
<tr>
<td>Area w/o Caches</td>
<td>0.27 mm$^2$</td>
<td>0.14 mm$^2$</td>
</tr>
<tr>
<td>Area with 16K Caches</td>
<td>0.53 mm$^2$</td>
<td>0.39 mm$^2$</td>
</tr>
<tr>
<td>Area Efficiency</td>
<td>2.96 DMIPS/MHz/mm$^2$</td>
<td>4.41 DMIPS/MHz/mm$^2$</td>
</tr>
<tr>
<td>Frequency</td>
<td>&gt;1GHz</td>
<td>&gt;1GHz</td>
</tr>
<tr>
<td>Dynamic Power</td>
<td>&lt;0.08 mW/MHz</td>
<td>0.034 mW/MHz</td>
</tr>
</tbody>
</table>

- **PPA reporting conditions**
  - 85% utilization, use Dhrystone for benchmark, frequency/power at TT 0.9V 25C, all regular VT transistors
- 10% higher in DMIPS/MHz, 49% more area-efficient
Hwacha Vector Accelerator

Vector Issue Unit

Bank0 Ctrl

Bank0 1R1W SRAM

Bank1 Ctrl

Bank1 1R1W SRAM

Bank7 Ctrl

Bank7 1R1W SRAM

64-bit Integer Multiplier

SP/DP Floating-Point Units

Vector Memory Unit

from Rocket

L1 VI$

Read Ports

Write Ports

Shared L1 D$
– After a 2-cycle initial startup latency, the banked RF is effectively able to read out 2 operands/cycle.
Processor Generators

- Express hardware as highly parameterized generators
  - Helps tune the design under different performance, power, and area constraints
- Parameters include:
  - number of cores
  - cache sizes, associativity, number of TLB entries, cache-coherence protocol
  - number of floating-point pipeline stages
  - width of off-chip I/O, and more
Writing Generators with Chisel

• RTL generator written in Chisel
  – HDL embedded in Scala

• Full power of Scala for writing generators
  – object-oriented programming, functional programming
Physical Design Flow

Chisel Source Code

Chisel

RTL Code (Verilog)

Synthesis
Place-and-Route

Gate-level Netlist

Formality
Formal Verification

PrimeTime/StarRC
Static Timing Analysis

VCS Post-PNR
Gate-level Simulation

Signed-Off Design

The core is synthesized and place-and-routed independently, and instantiated twice
# Chip Results

<table>
<thead>
<tr>
<th>Chip Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Process</strong></td>
<td>45nm SOI CMOS, 11 metal layers</td>
</tr>
<tr>
<td><strong>Package</strong></td>
<td>C4 area I/O, flip-chip bonded to PCB</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td></td>
</tr>
<tr>
<td>Processor</td>
<td>2.8mm X 1.1mm</td>
</tr>
<tr>
<td>1 Core</td>
<td>1.37mm X 1.06mm</td>
</tr>
<tr>
<td>SRAM Array</td>
<td>1.1mm X 4mm</td>
</tr>
<tr>
<td><strong>Standard Cells</strong></td>
<td></td>
</tr>
<tr>
<td>Processor</td>
<td>425K (85K flip-flops)</td>
</tr>
<tr>
<td>1 Core</td>
<td>192K (36K flip-flops)</td>
</tr>
<tr>
<td><strong>SRAM Bits</strong></td>
<td></td>
</tr>
<tr>
<td>Processor</td>
<td>1246K</td>
</tr>
<tr>
<td>1 Core</td>
<td>621K</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>1GHz (Nominal), 250MHz-1.3GHz</td>
</tr>
<tr>
<td><strong>Voltage</strong></td>
<td>1V (Nominal), 0.65V-1.2V</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>300mW-430mW (Nominal), 40mW-960mW</td>
</tr>
</tbody>
</table>
Measurement Setup

45nm RISC-V Vector Processor

Virtex 6 FPGA Board

Laptop

512MB DRAM

only used in basic testing mode

FSB 150MHz

1Gbps Ethernet
## Shmoo Plot of DP GFLOPS/W

Running Double-Precision Matrix Multiplication on Vector Accelerator

<table>
<thead>
<tr>
<th>Frequency (MHz)</th>
<th>Vdd (V)</th>
<th>0.65</th>
<th>0.70</th>
<th>0.75</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
<th>1.00</th>
<th>1.05</th>
<th>1.10</th>
<th>1.15</th>
<th>1.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 MHz @ 0.65 V</td>
<td></td>
<td>15.8</td>
<td>12.8</td>
<td>10.6</td>
<td>8.7</td>
<td>7.1</td>
<td>5.7</td>
<td>4.6</td>
<td>3.7</td>
<td>3.1</td>
<td>2.6</td>
<td>2.1</td>
<td>1.6</td>
</tr>
<tr>
<td>250 MHz @ 0.7 V</td>
<td></td>
<td>16.7</td>
<td>14.0</td>
<td>11.6</td>
<td>9.6</td>
<td>7.9</td>
<td>6.4</td>
<td>5.3</td>
<td>4.3</td>
<td>3.6</td>
<td>3.0</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>250 MHz @ 0.75 V</td>
<td></td>
<td>14.9</td>
<td>12.4</td>
<td>10.3</td>
<td>8.6</td>
<td>7.0</td>
<td>5.8</td>
<td>4.8</td>
<td>4.0</td>
<td>3.3</td>
<td>2.8</td>
<td>2.2</td>
<td>300</td>
</tr>
<tr>
<td>250 MHz @ 0.8 V</td>
<td></td>
<td>15.6</td>
<td>13.0</td>
<td>10.9</td>
<td>9.1</td>
<td>7.5</td>
<td>6.2</td>
<td>5.2</td>
<td>4.3</td>
<td>3.7</td>
<td>3.0</td>
<td>2.4</td>
<td>350</td>
</tr>
<tr>
<td>250 MHz @ 0.85 V</td>
<td></td>
<td>13.6</td>
<td>11.3</td>
<td>9.6</td>
<td>7.9</td>
<td>6.6</td>
<td>5.5</td>
<td>4.7</td>
<td>3.9</td>
<td>3.3</td>
<td>2.6</td>
<td>350</td>
<td></td>
</tr>
<tr>
<td>250 MHz @ 0.9 V</td>
<td></td>
<td>14.1</td>
<td>11.8</td>
<td>9.9</td>
<td>8.3</td>
<td>6.9</td>
<td>5.8</td>
<td>4.9</td>
<td>4.2</td>
<td>3.5</td>
<td>2.8</td>
<td>350</td>
<td></td>
</tr>
<tr>
<td>250 MHz @ 0.95 V</td>
<td></td>
<td>12.2</td>
<td>10.3</td>
<td>8.6</td>
<td>7.2</td>
<td>6.1</td>
<td>5.2</td>
<td>4.4</td>
<td>3.7</td>
<td>3.0</td>
<td>2.5</td>
<td>350</td>
<td></td>
</tr>
<tr>
<td>250 MHz @ 1 V</td>
<td></td>
<td>12.5</td>
<td>10.5</td>
<td>8.8</td>
<td>7.5</td>
<td>6.3</td>
<td>5.4</td>
<td>4.5</td>
<td>3.8</td>
<td>3.1</td>
<td>2.5</td>
<td>350</td>
<td></td>
</tr>
<tr>
<td>1 GHz @ 1 V</td>
<td></td>
<td>10.8</td>
<td>9.1</td>
<td>7.7</td>
<td>6.5</td>
<td>5.6</td>
<td>4.7</td>
<td>4.0</td>
<td>3.2</td>
<td>2.4</td>
<td>2.2</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.2 GHz @ 1.2 V</td>
<td></td>
<td>11.0</td>
<td>9.5</td>
<td>7.9</td>
<td>6.7</td>
<td>5.8</td>
<td>4.9</td>
<td>4.2</td>
<td>3.5</td>
<td>2.7</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.3 V</td>
<td></td>
<td>11.2</td>
<td>9.4</td>
<td>8.3</td>
<td>6.9</td>
<td>5.9</td>
<td>5.0</td>
<td>4.3</td>
<td>3.5</td>
<td>2.7</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.4 V</td>
<td></td>
<td>11.4</td>
<td>9.7</td>
<td>8.3</td>
<td>7.1</td>
<td>6.1</td>
<td>5.2</td>
<td>4.4</td>
<td>3.6</td>
<td>2.8</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.5 V</td>
<td></td>
<td>9.8</td>
<td>9.5</td>
<td>8.3</td>
<td>7.1</td>
<td>6.1</td>
<td>5.2</td>
<td>4.4</td>
<td>3.6</td>
<td>2.8</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.6 V</td>
<td></td>
<td>8.6</td>
<td>8.7</td>
<td>8.3</td>
<td>7.2</td>
<td>6.3</td>
<td>5.3</td>
<td>4.5</td>
<td>3.6</td>
<td>2.8</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.7 V</td>
<td></td>
<td>8.7</td>
<td>8.7</td>
<td>8.3</td>
<td>7.2</td>
<td>6.3</td>
<td>5.3</td>
<td>4.5</td>
<td>3.6</td>
<td>2.8</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.8 V</td>
<td></td>
<td>8.8</td>
<td>8.8</td>
<td>8.3</td>
<td>7.2</td>
<td>6.3</td>
<td>5.3</td>
<td>4.5</td>
<td>3.6</td>
<td>2.8</td>
<td>2.0</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>1.3 GHz @ 1.9 V</td>
<td></td>
<td>8.8</td>
<td>7.6</td>
<td>7.1</td>
<td>6.2</td>
<td>5.5</td>
<td>4.7</td>
<td>3.8</td>
<td>2.9</td>
<td>2.1</td>
<td>1.5</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.8 V</td>
<td></td>
<td>7.3</td>
<td>6.6</td>
<td>5.7</td>
<td>4.8</td>
<td>4.0</td>
<td>3.2</td>
<td>2.4</td>
<td>1.6</td>
<td>1.0</td>
<td>0.6</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.7 V</td>
<td></td>
<td>5.8</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.6 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.5 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.4 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.3 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.2 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.1 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>550 MHz @ 0.0 V</td>
<td></td>
<td>5.7</td>
<td>5.0</td>
<td>4.1</td>
<td>3.4</td>
<td>2.7</td>
<td>1.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

- **More Efficient**: Nominal 1GHz@1V, 7.3 GFLOPS/W
- **Less Efficient**: Max Frequency 1.3GHz@1.2V, 4.2 GFLOPS/W
- **Most Efficient**: Most Efficient 250MHz@0.65V, 16.7 GFLOPS/W
- **VDD at 0.8V**: 550MHz@0.8V, 12.5 GFLOPS/W

Not Operational
Energy Efficiency Comparison

<table>
<thead>
<tr>
<th>@0.8V</th>
<th>Frequency (GHz)</th>
<th>64-bit GFLOPS</th>
<th>Power (W)</th>
<th>Efficiency (GFLOPS/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Gene/Q</td>
<td>1.60</td>
<td>204.8</td>
<td>29.7</td>
<td>6.9</td>
</tr>
<tr>
<td>IBM Cell</td>
<td>3.20</td>
<td>108.8</td>
<td>22.5</td>
<td>4.8</td>
</tr>
<tr>
<td>This Work</td>
<td>0.55</td>
<td>1.72</td>
<td>0.138</td>
<td>12.5</td>
</tr>
</tbody>
</table>

– BG/Q and IBM Cell fabricated in same 45nm SOI
– Conservatively assume BG/Q and Cell achieves peak GFLOPS, we achieve 78% of peak GFLOPS
– Power numbers only for the core with private caches
  • Blue Gene/Q: Cores dissipate 54% of total power
  • IBM Cell: Assume that cores dissipate 50% of total power
– Why better energy efficiency than others?
  • Simpler, but yet more energy-efficient microarchitecture
More on Comparison

• But BG/Q is clocked 3X faster and Cell is 6X faster?
  – If the end goal is to provide better energy efficiency then use simpler microarchitectures and rely on parallelism for performance.

• But BG/Q and Cell have big on-chip caches? What about I/O power?
  – We only count the power dissipated in the core and the private L1 caches.

• But BG/Q and Cell have 100X more total GFLOPS!
  – Sorry, we only had budget for a small test chip.
Conclusions

• Processor generators written in high-level languages can produce energy-efficient, high-performance hardware
  – Our dual-core RISC-V vector processor achieves 16.7 DP GFLOPS/W at 0.65 V and a maximum frequency of 1.3 GHz at 1.2 V
• Open-source RISC-V ISA can serve as a competitive base ISA for integrating specialized heterogeneous accelerators
• Rocket chip generator and software tools open-sourced at http://riscv.org
Acknowledgment

- DARPA award HR0011-11-C-0100
- DARPA award HR0011-12-2-0016
- Center for Future Architecture Research, a member of STARnet, a Semiconductor Research Corporation program sponsored by MARCO
- NVIDIA graduate fellowship
- ASPIRE Lab industrial sponsors and affiliates Intel, Google, Nokia, NVIDIA, Oracle, and Samsung
- All POEM team members at MIT, UC Berkeley, CU Boulder