Challenges and Opportunities in Agile and Open Computer Architecture

David Patterson
UC Berkeley and Google
June 23, 2019
Moore’s Law Slowdown in Intel Processes

We’re now in the Post Moore’s Law Era

Moore, Gordon E. "No exponential is forever: but ‘Forever’ can be delayed!"
Energy scaling for fixed task is better, since more and faster transistors

Power consumption based on models in "Dark Silicon and the End of Multicore Scaling," Hadi Esmaelizadeh, ISCA, 2011
End of Growth of Single Program Speed?

End of the Line?

2X / 20 yrs (3%/yr)

Am- dahl’s Law ⇒ 2X / 6 yrs (12%/yr)

End of Dennard Scaling ⇒ Multicore 2X / 3.5 yrs (23%/yr)

RISC 2X / 1.5 yrs (52%/yr)

CISC 2X / 3.5 yrs (22%/yr)

Based on SPECintCPU. Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018
Current Security Challenge

● Spectre: speculation $\Rightarrow$ timing attacks that leak $\geq 10$ kb/s
● More microarchitecture attacks on the way*
● Spectre is bug in computer architecture definition vs chip
● Need Computer Architecture 2.0 to prevent timing leaks**
● Software not yet secure $\Rightarrow$ how can hardware help?

* “A Survey of Microarchitectural Timing Attacks and Countermeasures on Contemporary Hardware,” Qian Ge, Yuval Yarom, David Cock, and Gernot Heiser, Journal of Cryptographic Engineering, April, 2018
** “A Primer on the Meltdown & Spectre Hardware Security Design Flaws and their Important Implications”, Mark Hill, 2/15/18, Computer Architecture Today
What Opportunities Left? (Part I)

- *Software advances can inspire architecture innovations*
- Why open source compilers and operating systems but not ISAs?
What’s Different About RISC-V? ("RISC Five", fifth UC Berkeley RISC)

- **Free and Open**
  - Anyone can use
  - More competition ⇒ More innovation

- **For Cloud & Edge**
  - From large to tiny computers

- **Secure/Trustworthy**
  - Design own secure core
  - Open cores ⇒ no secrets

- **Simple, Modular**
  - 25 years later, learn from mistakes of 1st generation RISCs
  - Far simpler than ARM and x86
  - Can add custom instructions

- **Community evolves and provides stability**
  - RISC-V Foundation owns RISC-V ISA
RISC-V Foundation Growth

September 2015 – February 2019

Platinum Gold Silver Auditor Individual

RISC-V Foundation Summit
International Members of RISC-V

RISC-V Members in 27 Countries Around the World!

Representing ~57% of the global population!!
Calista Redmond, New RISC-V Foundation CEO

Previously, Vice-President of IBM Z Ecosystem division; President of OpenPOWER Foundation.
RISC-V in Education

RISC-V spreading quickly throughout curricula of top schools
International Interest in RISC-V

English ($20)
riscbook.com/

Japanese (3,240 ¥)
www.amazon.co.jp/dp/4822292819/

Spanish (free)
riscbook.com/spanish/

Portuguese (free)
riscbook.com/portuguese/

Chinese (free)
riscbook.com/chinese/
NVDLA: An Open DSA and Implementation

- NVDLA: NVIDIA Deep Learning Accelerator for DNN Inference
- Free & Open: All SW, HW, and documentation on GitHub
- Scalable, configurable design
  - Each block operates independently or in pipeline to bypass memory
  - Data type configurable: int8, int16, fp16,
  - 2D MAC array configurable: 8 to 64 x 4 to 64
  - Size scales 6X (0.5 - 3mm²), power scales 15X (20 - 300 mW)
- RISC-V core as host (optional)
Free & Open Instruction Set (ISA) vs Free & Open Source Hardware?

- **Specifications**
  - Instruction Set Architecture (for example, RISC-V)
  - Similar to Portable Operating System Interface (POSIX) standard in software

- **Designs (“source code”)**
  - RISC-V Rocket
  - Similar to Linux in software

- **Products**
  - OURS Pygmy chip
  - Similar to RedHat 7.5 in software

---

3 Types of Specifications or Designs

1. **Free & Open**
   - No fee, anyone can use
   - Can design it yourself, share with others, get from others

2. **Licensable**
   - Company owns, pay fee to use
   - Can’t share with or get from others

3. **Closed**
   - Company owns, others cannot use
Need Free & Open Specification
To Have Free & Open Designs

<table>
<thead>
<tr>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free &amp; Open Spec</td>
</tr>
<tr>
<td>Licensable Spec</td>
</tr>
<tr>
<td>Closed Spec</td>
</tr>
</tbody>
</table>
Need Free & Open Specification To Have Free & Open Designs

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Free &amp; Open Designs</th>
<th>Licensable Designs</th>
<th>Closed Designs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Free &amp; Open Spec</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Licensable Spec</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Closed Spec</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Need Free & Open Specification To Have Free & Open Designs

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Free &amp; Open Designs</th>
<th>Licensable Designs</th>
<th>Closed Designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free &amp; Open Spec</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Licensable Spec</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closed Spec</td>
<td></td>
<td></td>
<td>Based on Closed Designs</td>
</tr>
</tbody>
</table>

Based on the table above, it is clear that to have free and open designs, there is a need for free and open specifications. This ensures that the designs can be freely used and modified, leading to a wider range of products and innovations.
Need Free & Open Specification To Have Free & Open Designs

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Designs (&quot;Source&quot;)</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free &amp; Open Spec</td>
<td>Free &amp; Open Designs</td>
<td>Based on Licensed or Closed Designs</td>
</tr>
<tr>
<td></td>
<td>Licensable Designs</td>
<td>$5M + 4%</td>
</tr>
<tr>
<td></td>
<td>Closed Designs</td>
<td>$25M</td>
</tr>
<tr>
<td>Closed Spec</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on Closed Designs
### Need Free & Open Specification To Have Free & Open Designs

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Free &amp; Open Designs</th>
<th>Licensable Designs</th>
<th>Closed Designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free &amp; Open Spec</td>
<td>“Open Source”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Licensable Spec</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closed Spec</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Based on Free & Open, Licensed, or Closed Designs
- Based on Licensable or Closed Designs
- Based on Closed Designs
OURS Pygmy microprocessor
28nm HPC+ TSMC @ 600 MHz
From scratch to tapeout ~7 months (Thanks to the RISC-V infrastructure)
Full RISC-V based heterogenous multicore architecture
64-bit control processor (RV64g)
~ 10mW active
12 energy-efficient AI engines based on custom RV vector extensions
INT8 : ~4 TOPS/watt
FP16 : ~0.35 TOPS/watt
1MB SRAM, LPDDR4 support
Retail price < ¥20 ($3)
Security and Open Architecture

- Security community likes simple, verifiable (no trap doors), alterable, free and open architecture and implementations
- Equally important is number of people and organizations performing architecture experiments
  - Want all the best minds to work on security
- Plasticity of FPGAs + open source RISC-V implementations and SW ⇒ novel architectures can be deployed online, subjected to real attacks, evaluated & iterated in weeks vs years (even 100 MHz OK)
- RISC-V may become security exemplar via HW/SW codesign by architects and security experts
What Opportunities Left? (Part II)

- **Software advances can inspire innovations**
- Agile: small teams do short development between working but incomplete prototypes and get customer feedback per step
- Scrum team organization
  - 5 - 10 person team size
  - 2 - 4 week sprints for next prototype iteration
- New CAD enables SW Dev techniques to make small teams productive via abstraction & reuse

=> **Agile Hardware Development**
Small chip tape-out 100 chips 1x1mm @ 28nm is affordable at $14,000!

AWS FPGA F1 instance ⇒ develop new prototypes using cloud (nothing to buy)

What Opportunities Left? (Part III)

▪ Only performance path left is **Domain Specific Architectures (DSAs)**
  - Just do a few tasks, but extremely well
▪ Achieve higher efficiency by tailoring the architecture to characteristics of the domain
▪ Not one application, but a domain of applications
▪ Different from strict ASIC since still runs software
Why DSAs Can Win (no magic)
Tailor the Architecture to the Domain

• More effective parallelism for a specific domain:
  • SIMD vs. MIMD
  • VLIW vs. Speculative, out-of-order
• More effective use of memory bandwidth
  • User controlled versus caches
• Eliminate unneeded accuracy
  • IEEE replaced by lower precision FP
  • 32-64 bit integers to 8-16 bit integers
• Domain specific programming language provides path for software
Tensor Processing Unit v1 (Announced May 2016)

Google-designed chip for neural net inference

In production use for 3 years: used by billions on search queries, for neural machine translation, for AlphaGo victory over Lee Sedol, …

*A Domain-Specific Architecture for Deep Neural Networks*, Jouppi, Young, Patil, Patterson, *Communications of the ACM*, September 2018
TPU: High-level Chip Architecture

- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
  - 65,536 * 2 * 700M
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU
- 4 MiB of on-chip Accumulator memory + 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much on-chip memory vs GPU
- 8 GiB of off-chip weight DRAM memory
Perf/Watt TPU vs CPU & GPU

Using production applications vs contemporary CPU and GPU

Measure performance of Machine Learning?

See MLPerf.org ("SPEC for ML")
- Benchmark suite being developed by 23 companies and 7 universities
- 1st Results November 2018
Current Neural Network Architecture Debate

- Google TPU: 1 core per chip, large 2D multiplier, software controlled memory (instead of caches)
- Nvidia GPU: 80+ cores, many threads, many registers
- Microsoft FPGA: customize “hardware” to application
- Intel CPU: 30+ cores, 3 levels of caches, SIMD instructions
  - Also bought Altera that supplies Microsoft’s FPGAs
  - Also bought Nervana, Movidius, MobilEye to offer custom chip DSA
- > 45 startups with their own architecture bets
Conclusion

- End of Moore’s Law, Dennard Scaling, General Purpose Performance, Security Challenges require Innovation in Computer Architecture
- Open Instruction Set Architecture vs. Licensable or Closed Instruction Set Architecture
- Agile Hardware Development vs. Large Scale Hardware Development
- Domain Specific Architecture vs. General Purpose Architectures