CSUN | CALIFORNIA STATE UNIVERSITY NORTHRIDGE

COMP122

COMP 122

DR JEFF SOFTWARE
INDIE APP DEVELOPER

Prof. David Patterson (UCB)

SPEAKER
Dave Patterson
Distinguished Engineer, Google; ACM A.M. Turing Award Laureate

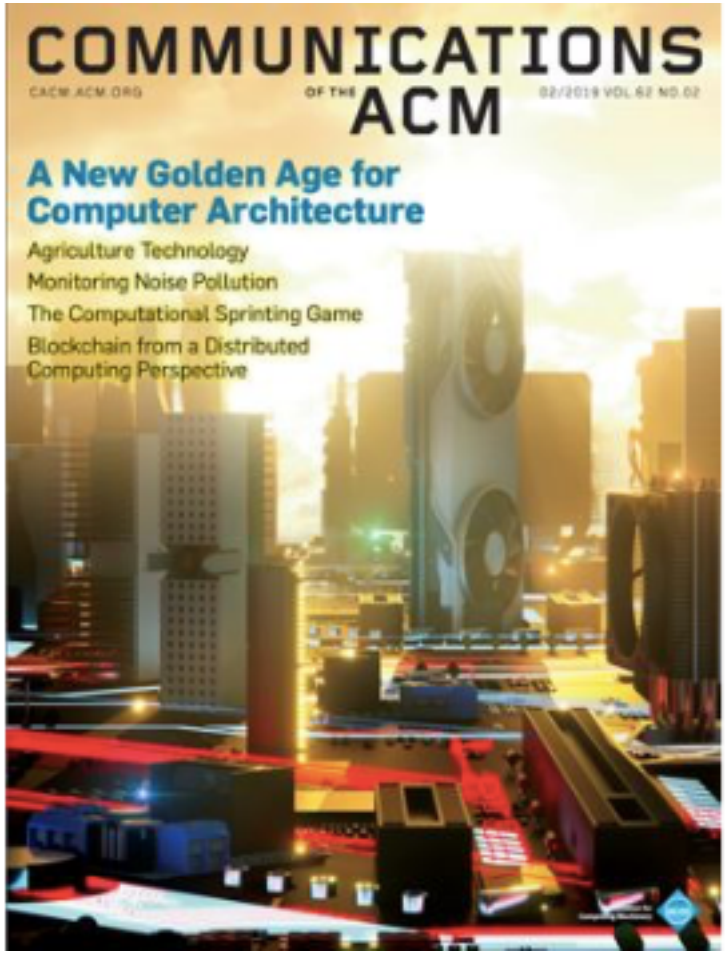
Webinar

Compiled by
Dr Jeff Drobman

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP122

ACM Tech Talk

New Golden Age? — David Patterson



The Learning Continues...

TechTalk Discourse: https://on.acm.org
TechTalk Inquiries: learning@acm.org
Learning Center & TechTalk Archives: https://learning.acm.org
Professional Ethics: https://ethics.acm.org
Queue Magazine: https://queue.acm.org

Lessons Learned

David Patterson

Lessons of last 50 years of Computer Architecture

1. *Software advances can inspire architecture innovations*
   - Microprogramming - control as SW
   - RISC, x86 ISA - (Hardware) translator vs interpreter
   - Open Architectures & Implementations
   - Agile Hardware Development
2. *Raising the HW/SW interface enables arch. opportunities*
   - Assembly to HLL ⇒ RISC
   - HLL to Domain Specific Language⇒DSA
3. *Ultimately the marketplace settles architecture debates*
   - Losers: 432
   - Winners: IBM S/360, 8086 (PC Era), RISC (Post PC Era)
   - Open vs Proprietary ISA (RISC-V vs ARM): Too soon to tell
   - ML DSA (SIMD vs GPU vs TPU vs FPGA vs startups): Too soon to tell

History

David Patterson

IBM Compatibility Problem in Early 1960s

By early 1960's, *IBM had 4 incompatible lines of computers!*

701	➡	7094
650	➡	7074
702	➡	7080
1401	➡	7010

Each system had its own:

- Instruction set architecture (ISA)
- I/O system and Secondary Storage: magnetic tapes, drums and disks
- Assemblers, compilers, libraries,...
- Market niche: business, scientific, real time, ...

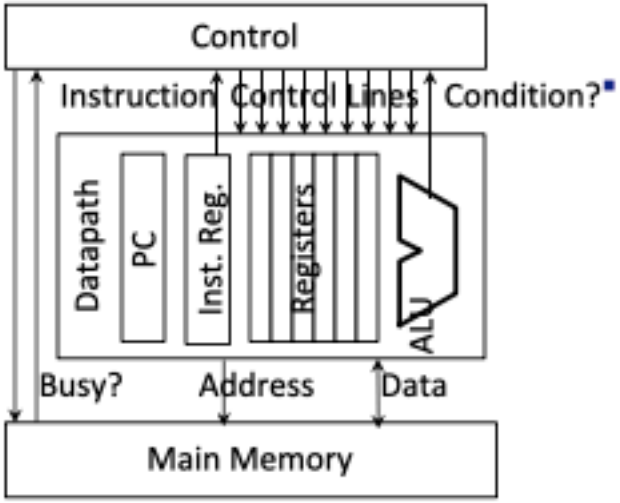
IBM System/360 – one ISA to rule them all

CPU Control

David Patterson

Control versus Datapath

- Processor designs split between *datapath*, where numbers are stored and arithmetic operations computed, and *control*, which sequences operations on datapath
- Biggest challenge for computer designers was getting control correct



- **Maurice Wilkes** invented the idea of *microprogramming* to design the control unit of a processor*

  - Logic expensive vs. ROM or RAM
  - ROM cheaper and faster than RAM
  - *Control design now programming*

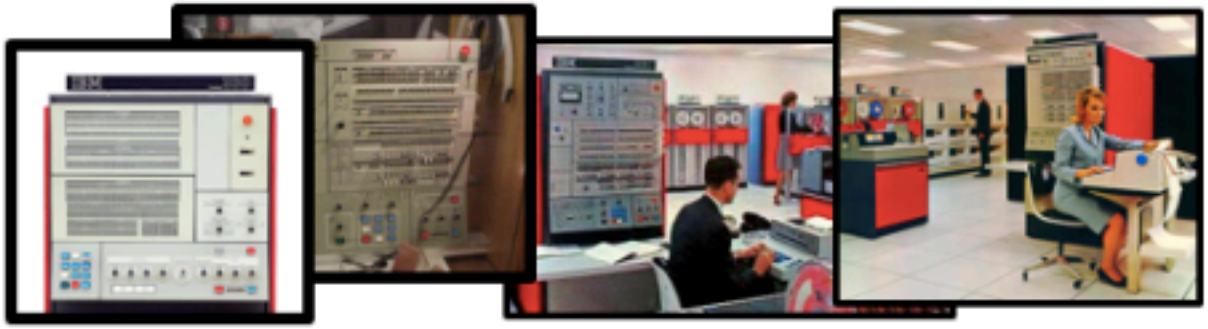
* "Micro-programming and the design of the control circuits in an electronic digital computer," M. Wilkes, and J. Stringer. *Mathematical Proc. of the Cambridge Philosophical Society*, Vol. 49, 1953.

RISC vs CISC

David Patterson

Microprogramming in IBM 360

Model	M30	M40	M50	M65
Datapath width	8 bits	16 bits	32 bits	64 bits
Microcode size	4k x 50	4k x 52	2.75k x 85	2.75k x 87
Clock cycle time (ROM)	750 ns	625 ns	500 ns	200 ns
Main memory cycle time	1500 ns	2500 ns	2000 ns	750 ns
Price (1964 $)	$192,000	$216,000	$460,000	$1,080,000
Price (2018 $)	$1,560,000	$1,760,000	$3,720,000	$8,720,000

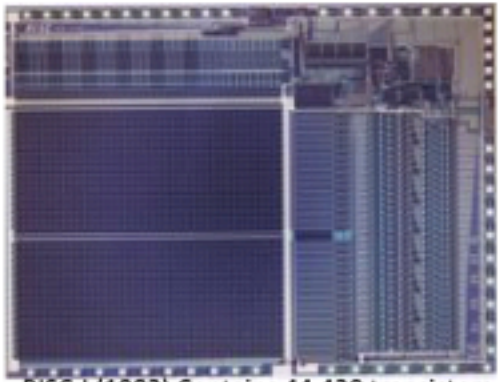


RISC History

MIPS & SPARC

David Patterson

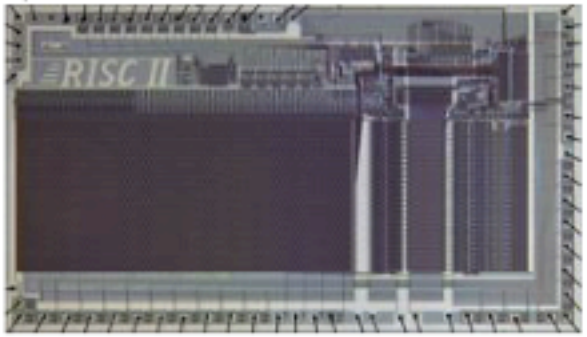
Berkeley and Stanford RISC Chips



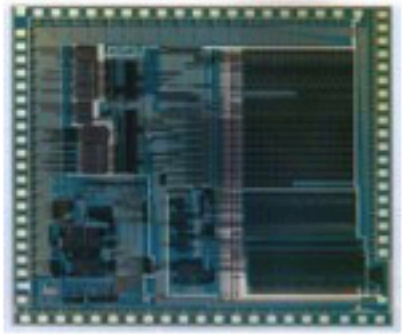
Fitzpatrick, Daniel, John Foderaro, Manolis Katevenis, Howard Landman, David Patterson, James Peek, Zvi Peshkess, Carlo Séquin, Robert Sherburne, and Korbin Van Dyke. "A RISCy approach to VLSI." *ACM SIGARCH Computer Architecture News* 10, no. 1 (1982)

RISC-I (1982) Contains 44,420 transistors, fabbed in 5 µm NMOS, with a die area of 77 mm², ran at 1 MHz

Hennessy, John, Norman Jouppi, Steven Przybylski, Christopher Rowen, Thomas Gross, Forest Baskett, and John Gill. "MIPS: A microprocessor architecture." In *ACM SIGMICRO Newsletter*, vol. 13, no. 4, (1982).



RISC-II (1983) contains 40,760 transistors, was fabbed in 3 µm NMOS, ran at 3 MHz, and the size is 60 mm²



Stanford MIPS (1983) contains 25,000 transistors, was fabbed in 3 µm & 4 µm NMOS, ran at 4 MHz (3 µm), and size is 50 mm² (4 µm) (Microprocessor without Interlocked Pipeline Stages)

14

RISC vs CISC

David Patterson

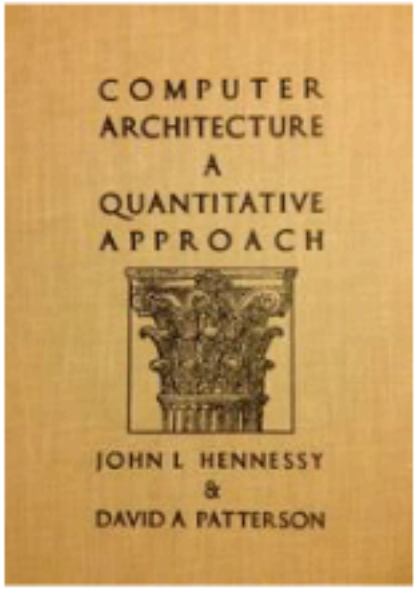
"Iron Law" of Processor Performance: How RISC can win

$$\frac{\text{Time}}{\text{Program}} = \frac{\text{Instructions}}{\text{Program}} * \frac{\text{Clock cycles}}{\text{Instruction}} * \frac{\text{Time}}{\text{Clock cycle}}$$

- CISC executes fewer instructions / program (≈ 3/4X instructions) but many more clock cycles per instruction (≈ 6X CPI)

  ⇒ RISC ≈ 4X faster than CISC

"Performance from architecture: comparing a RISC and a CISC with similar hardware organization," Dileep Bhandarkar and Douglas Clark, *Proc. Symposium, ASPLOS*, 1991.



RISC

David Patterson

CISC vs. RISC Today

PC Era

- Hardware translates x86 instructions into internal RISC instructions *(Compiler vs Interpreter)*
- Then use any RISC technique inside MPU
- > 350M / year !
- x86 ISA eventually dominates servers as well as desktops

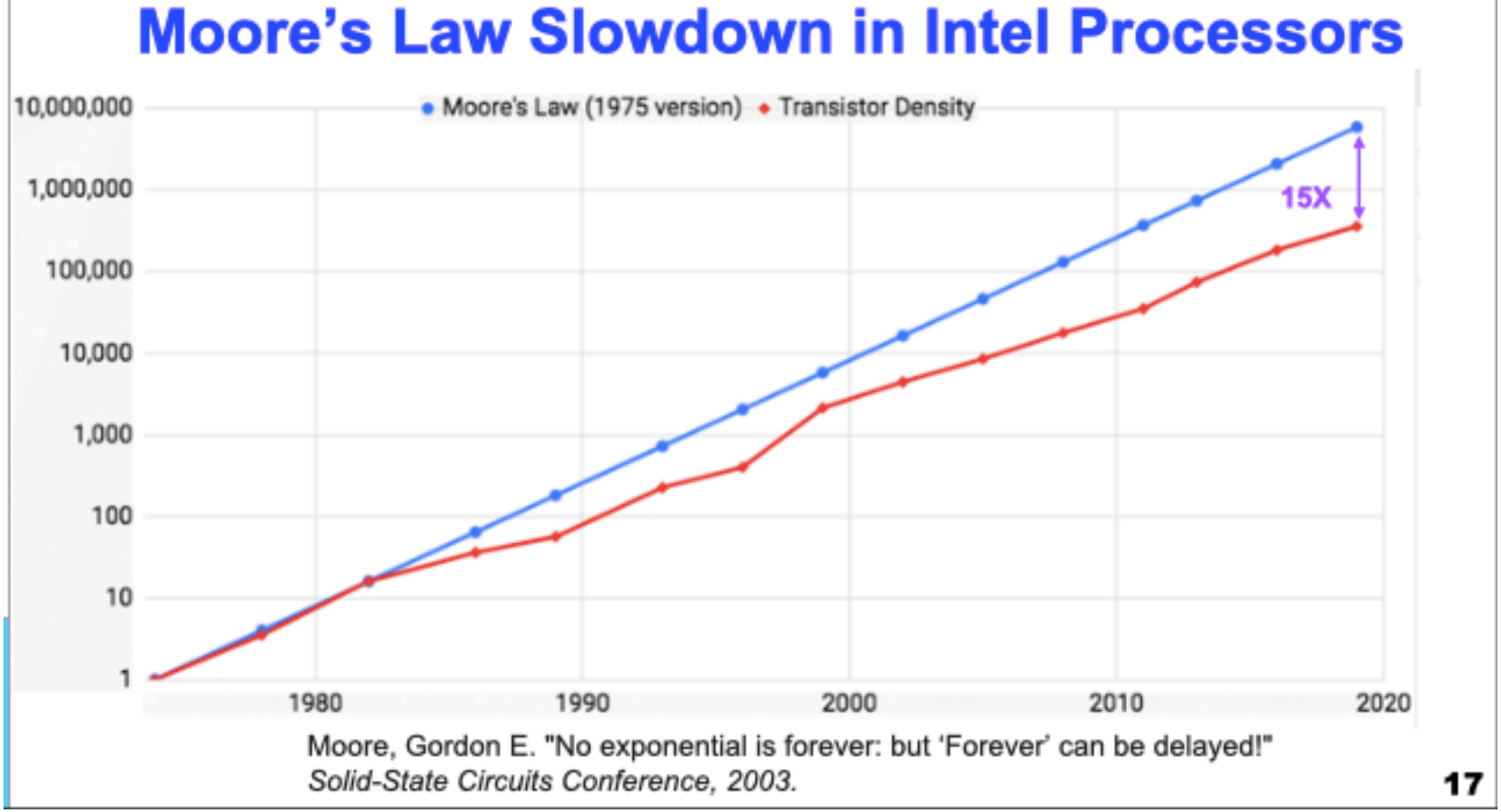
PostPC Era: Client/Cloud

- IP in SoC vs. MPU
- Value die area, energy as much as performance
- > 20B total / year in 2017
- 99% Processors today are RISC
- *Marketplace settles debate*

*"A Decade of Mobile Computing", Vijay Reddi, 7/21/17, *Computer Architecture Today*

End of Moore's Law

David Patterson

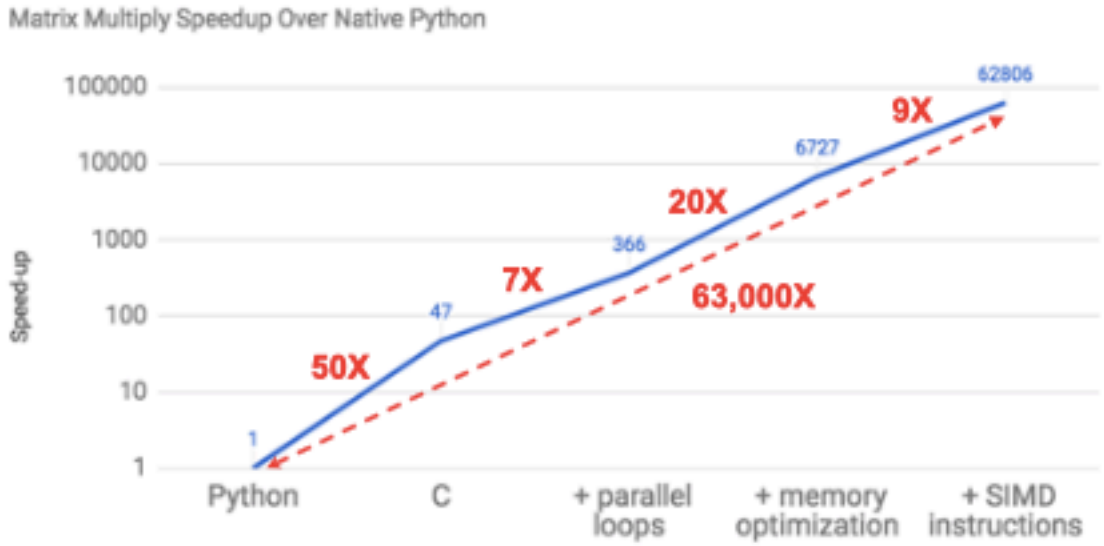


Multi-Core Perfromance

David Patterson

What's the Opportunity?

Matrix Multiply: relative speedup to a Python version (on 18 core Intel CPU)

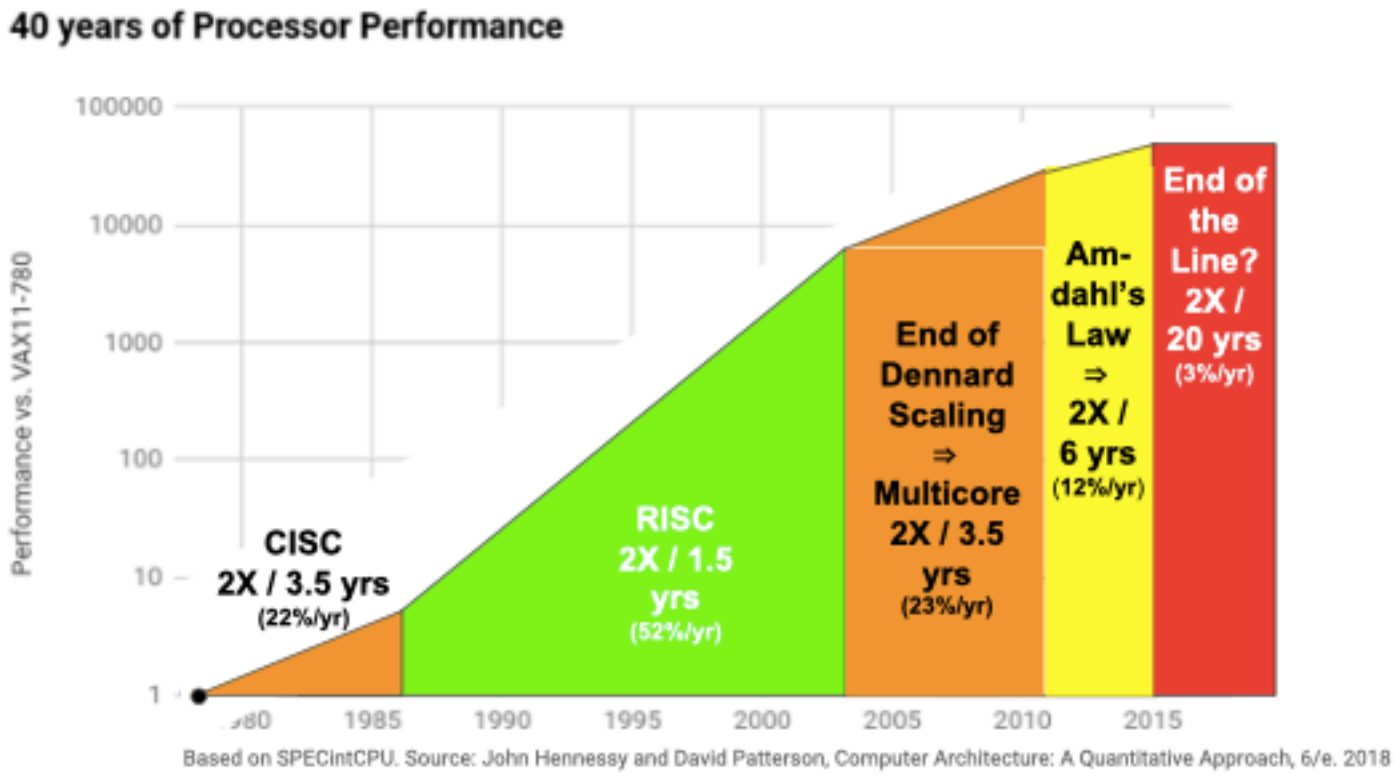


from: "There's Plenty of Room at the Top," Leiserson, et. al., *Science, to appear.*

Multi-Core Perfromance

David Patterson

End of Growth of Single Program Speed?



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

RISC-V Origin

David Patterson

RISC-V Origin Story

- UC Berkeley Research using x86 & ARM?
  - Impossible – too complex *and* IP issues
- 2010 started "3-month project" to develop own clean-slate ISA
  - Krste Asanovic, Andrew Waterman, Yunsup Lee, Dave Patterson
- 4 years later, released frozen base user spec

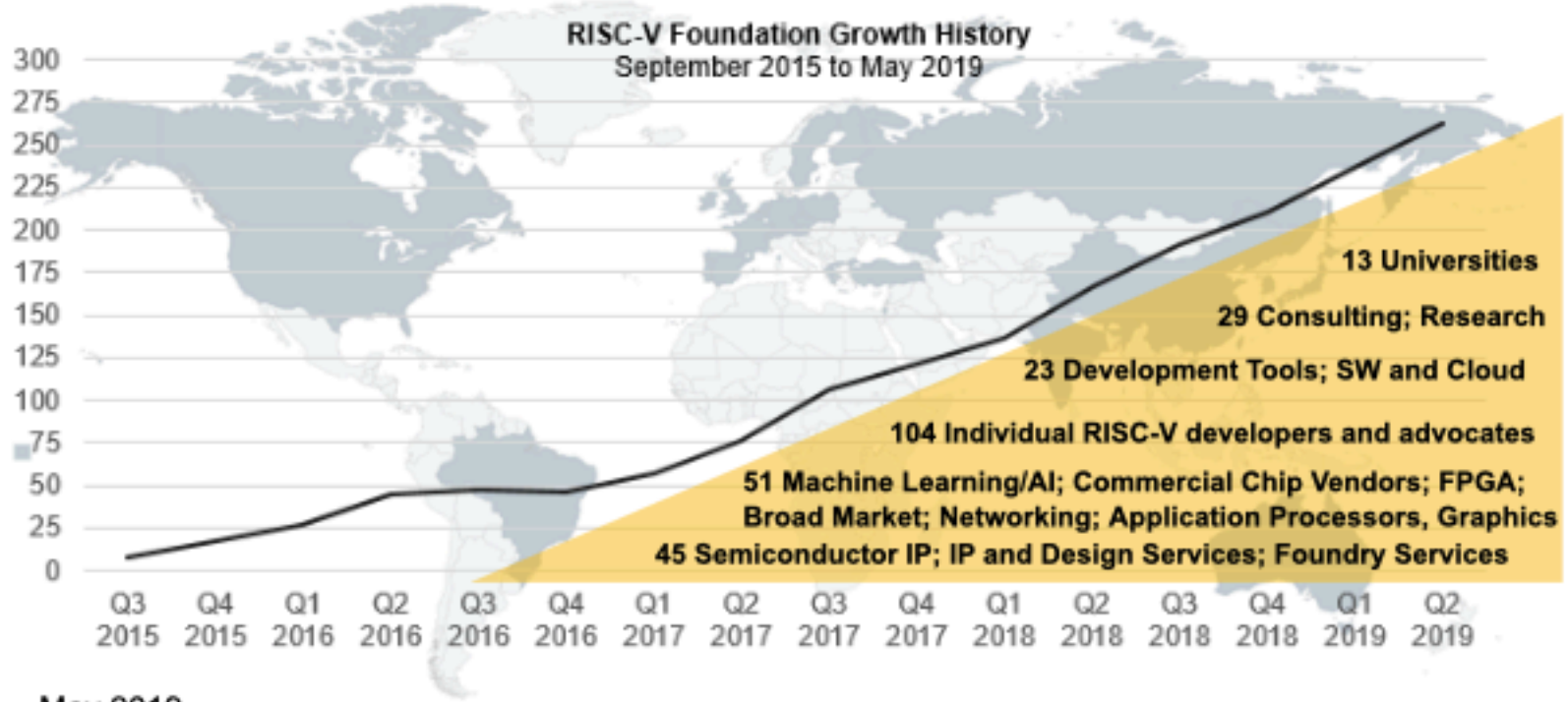
Why are outsiders complaining about changes of RISC-V in Berkeley classes?

37

RISC-V ISA

David Patterson

More than 300 RISC-V Members in 28 Countries Around the World



May 2019

RISC-V Affiliates

David Patterson

What's Next

David Patterson

What Opportunities Left? (Part I)

- SW-centric
  - Modern scripting languages are interpreted, dynamically-typed and encourage reuse
  - Efficient for programmers but not for execution
- HW-centric
  - Only path left is *Domain Specific Architectures*
  - Just do a few tasks, but extremely well
- Combination:
  - Domain Specific Languages & Architectures
  - Raises level of HW/SW Interface

DSA

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

Google TPU's

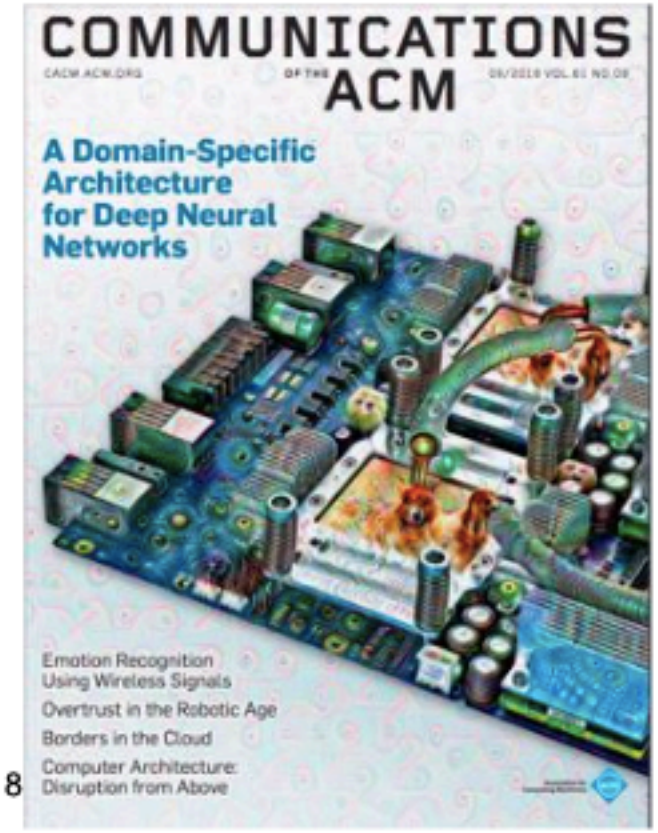
David Patterson

Tensor Processing Unit v1 (Announced May 2016)

Google-designed chip for neural net **inference**

In production use for 4 years: used by billions on search queries, for neural machine translation, for AlphaGo match, …

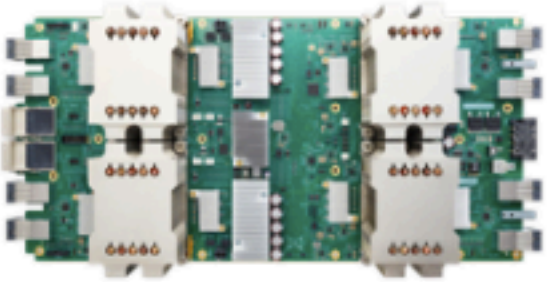
A Domain-Specific Architecture for Deep Neural Networks, Jouppi, Young, Patil, Patterson, *Communications of the ACM*, September 2018

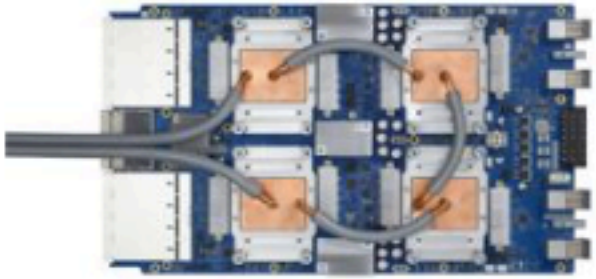


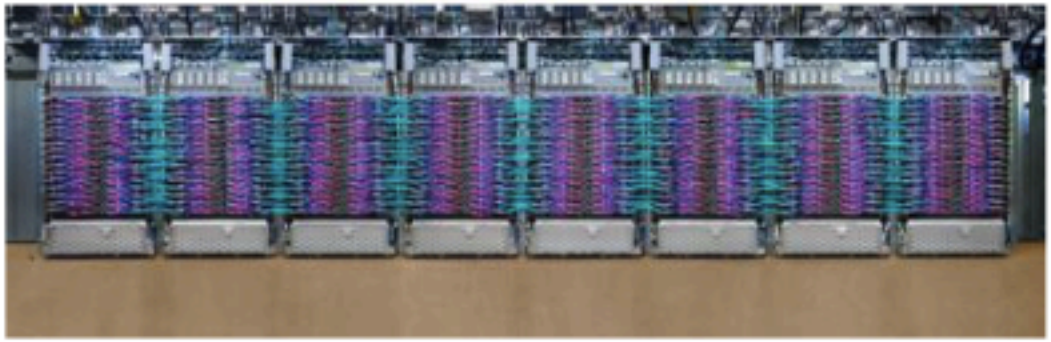
Google TPU's

David Patterson

Training: TPUv2 (5/2017), TPUv3 (5/2018)







Peak: 11.5 PetaFLOP/s

Peak: >100 PetaFLOP/s

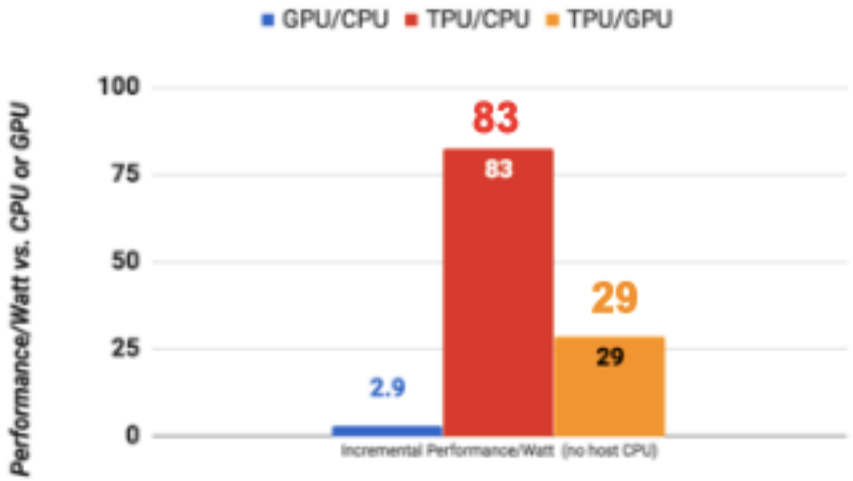
32

TPU Performance

David Patterson

Perf/Watt TPU vs CPU & GPU

Using production applications vs contemporary CPU and GPU



Measure performance of Machine Learning?

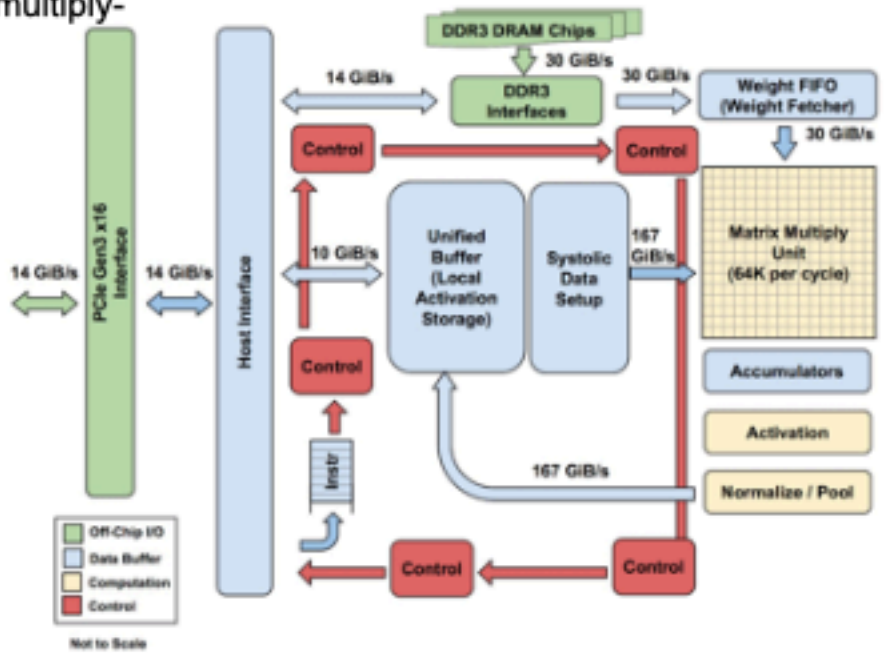
See MLPerf.org ("SPEC for ML")

- Benchmark suite being developed by 23 companies and 7 universities
- 1st Results Public 12/12/18

David Patterson

TPU: High-level Chip Architecture

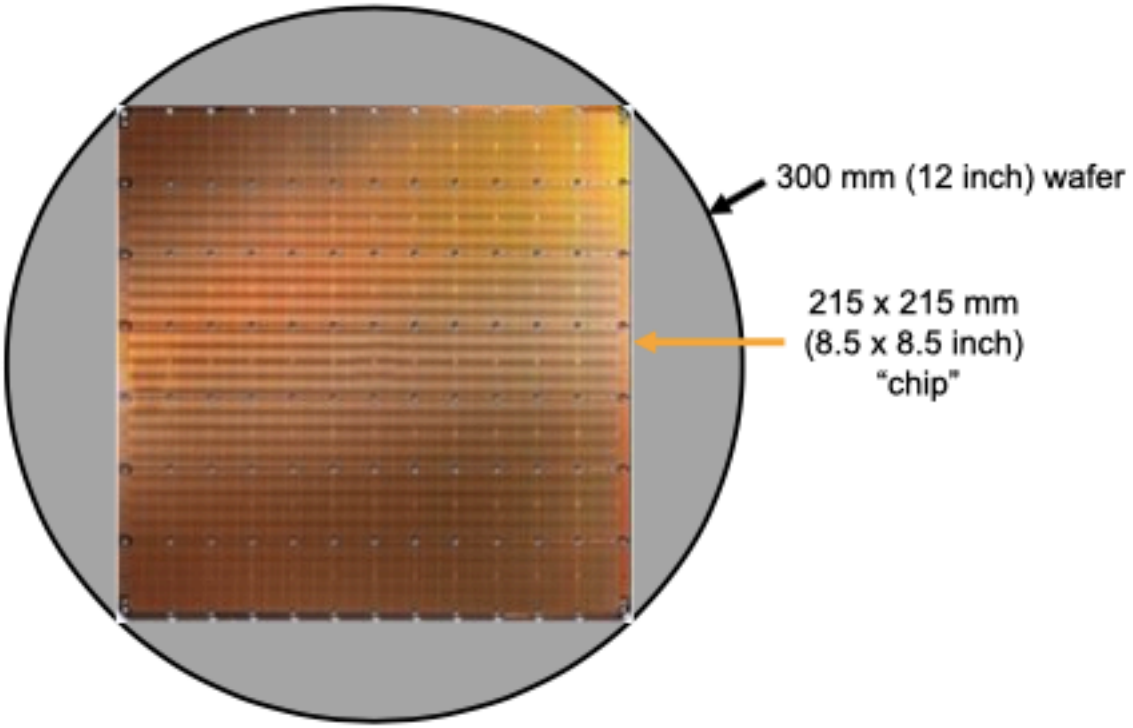
- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
  - 65,536 * 2 * 700M
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU
- 4 MiB of on-chip Accumulator memory + 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much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



Wafer-Scale ML

David Patterson

Cerebus announces ML Training "Chip" 8/19/19



Neural Network Arch

David Patterson

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 (20MB registers), small multipliers, caches, scatter/gather & coalescing HW
- 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
- > 100 startups with their own architecture bets
- *#3. Ultimately the marketplace settles architecture debates*

Security Challenge

David Patterson

Current Security Challenge

- Spectre: speculation ⇒ timing attacks that leak ≥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 ⇒ 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

Security

David Patterson

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

Agile Hardware?

David Patterson

What Opportunities Left? (Part III)

- *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**

43

No Quantum Computers

David Patterson -

Quantum Computing to the Rescue?

- *Quantum Computing - Progress and Prospects**
  - 12/2018 consensus study from National Academies
- "*Significant technical and financial issues remain towards building a large, fault-tolerant quantum computer and one is unlikely to be built within the coming decade.*"

  Gwynne, Peter. (2019). "Practical quantum computers still at least a decade away." *Physics World*. 32. 9-9. 10.1088/2058-7058/32/1/14.

•Mark Horowitz (Chair, NAE, Stanford, EE), Alán Aspuru-Guzik (U. Toronto, Chemistry), David Awschalom (NAE & NAS, U. Chicago, Physics), Robert Blakley (Citigroup), Dan Boneh (NAE, Stanford, CS), Susan Coppersmith (NAS, U. Wisconsin, Physics), Jungsang Kim (Duke, Physics & CS), John Martinis (UCSB & Google), Margaret Martonosi (Princeton, CS), Michele Mosca (U. Waterloo, Math & Physics), William Oliver (MIT, Physics), Krysta Svore (Microsoft), Umesh Vazirani (NAE, Berkeley, CS), National Academies, Washington D.C.
https://www.nap.edu/catalog/25196/quantum-computing-progress-and-prospects