"Flash and Chips"
The FireBox Warehouse-Scale Computer

Krste Asanović, David Patterson, Randy Katz,]
+ FireBox group and ASPIRE Lab
krste@berkeley.edu

Fujitsu Labs, America Technology Symposium
June 24, 2015

Three WSC Generations

1. ~2000: Commercial Off-The-Shelf (COTS) computers, switches, & racks

2. ~2010: Custom computers, switches, & racks but build from COTS chips, drive new board-level standards (“Open Compute”)

3. ~2020: Custom computers, switches, & racks using custom WSC chips, drive new chip-scale open standards (“RISC-V”)

ASPIRE
UC Berkeley

Moore's Law: Old vs. New Conventional Wisdom

- Old CW: Moore's Law, each 18-month technology generation, transistor performance/energy improves, cost/transistor decreases
- New CW: generations slowing to 3 year -> 5+ year, transistor performance/energy slight improvement, cost/transistor *increases*!

2015: Moore's Law has ended for logic, SRAM, & DRAM (Maybe 3D Flash & new NVM continues?)

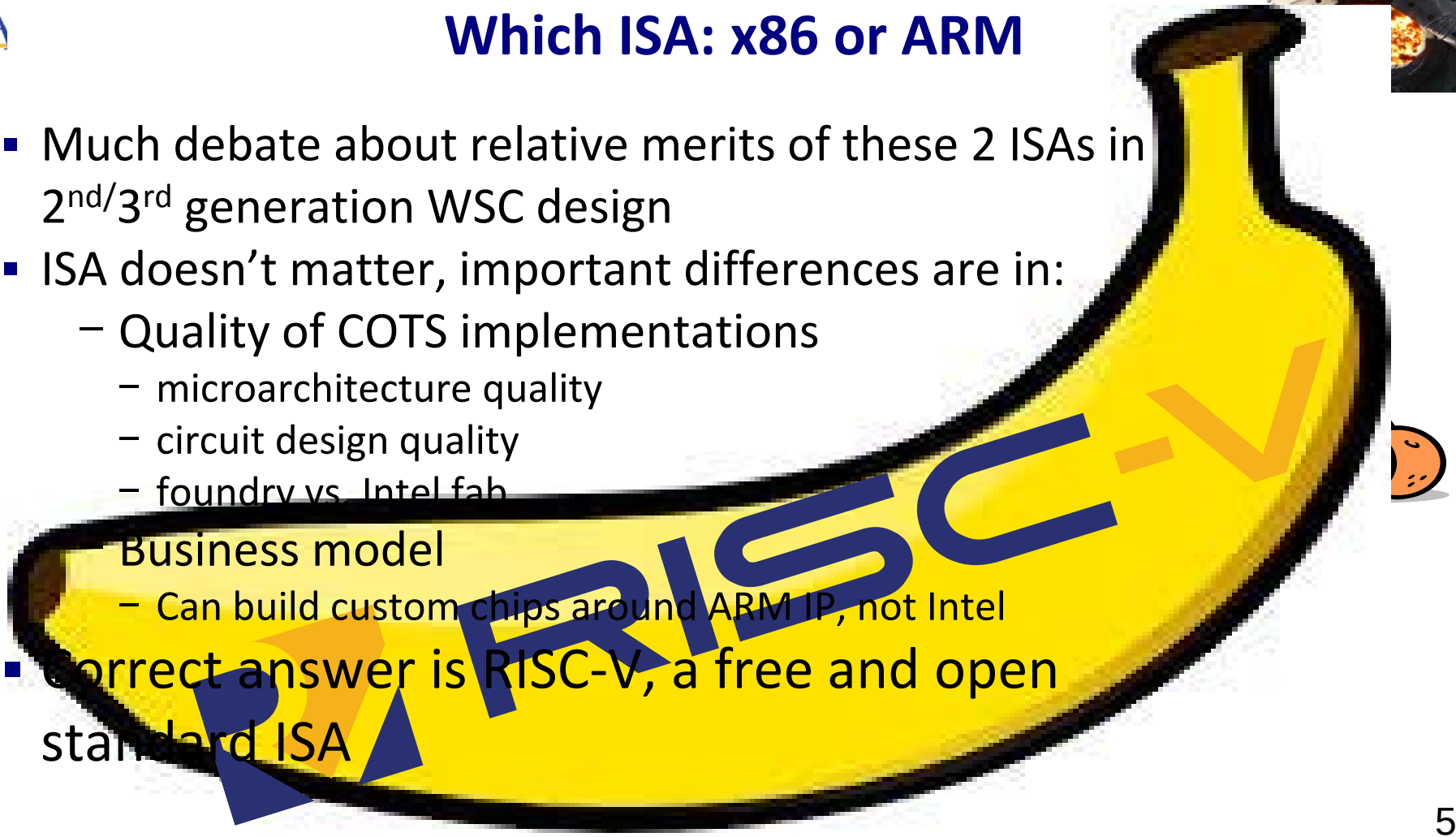
ASPIRE
UC Berkeley

Why WSC Custom Chips in 2020?

- Without transistor scaling, improvements in system capability have to come above transistor-level
  - More specialized hardware
- Good news: when scaling stops, custom chip costs drop
  - Amortize investments in capital equipment, CAD tools, libraries, training, … over decade vs. 18 months
- WSCs proliferate @ $100M/WSC
  - Economically sound to divert some $ if yield more cost-performance-energy effective chips
- New HW description languages supporting parameterized generators improve productivity and reduce design cost
  - E.g., Stanford Genesis2; Berkeley's Chisel, based on Scala

Which ISA: x86 or ARM

- Much debate about relative merits of these 2 ISAs in 2nd/3rd generation WSC design
- ISA doesn't matter, important differences are in:
  - Quality of COTS implementations
    - microarchitecture quality
    - circuit design quality
    - foundry vs. Intel fab
  - Business model
    - Can build custom chips around ARM IP, not Intel
- Correct answer is RISC-V, a free and open standard ISA



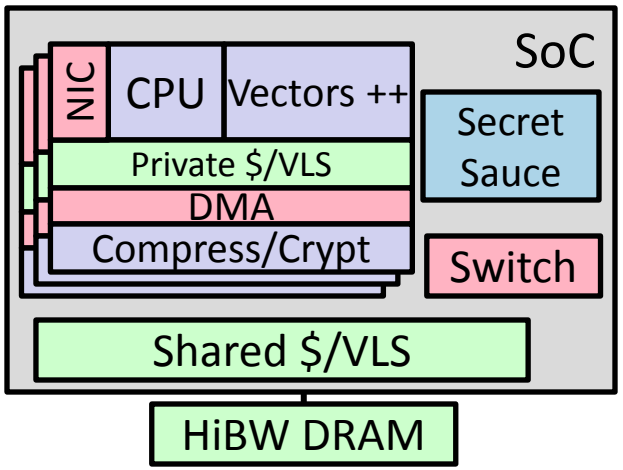
ASPIRE
UC Berkeley

RISC-V

www.riscv.org

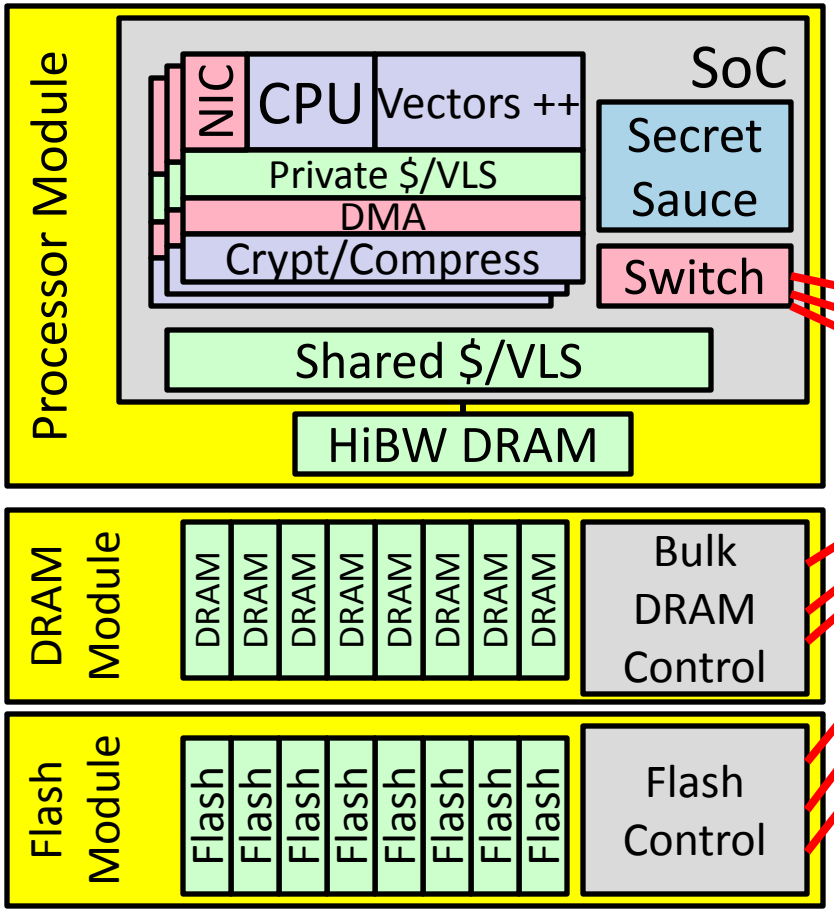
- New completely free, open ISA => shared open cores
  - Already runs GCC, Linux, glibc, LLVM, QEMU, …
  - RV32/RV64/RV128 variants for 32b/64b/128b address
- Designed for SoCs: allow per-SoC customization yet maintain software compatibility
  - Base ISA that will never change with only 40 integer instructions, and supports compiler, linker, OS, …
  - Optional extensions provide full general-purpose ISA, including IEEE-754/2008 floating-point
  - “Secret sauce” accelerator unique per SoC
- 11 64-bit chips at Berkeley so far (45nm, 28nm)
- Multiple external groups building commercial parts to ship this year
- RISC-V non-profit foundation to launch this summer

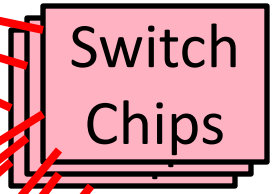
FireBox "Strawman" Processor Module



- ~64 identical vector/SIMD cores / SoC
  - Simplify resource management, software model
  - Provide redundancy for failure tolerance
  - Enhanced vector unit with specialized engines
- 8-32 DRAM chips on interposer for high BW
  - 32Gb chips give 32-128GB DRAM capacity/node
  - 500GB/s+ DRAM bandwidth
- Cache coherent on-chip, simplify OS for SoC
- Fast messaging NIC/DMA near each core
- Cache & BW partitioning for QoS support
- Compress/Encrypt engine so reduce size yet always encrypted outside node
- "Secret sauce" for per-SoC accelerators
- Via parameterized Chisel WSC-SoC generator
  - Easy to add custom application accelerators, tune architectural parameters

FireBox Rack



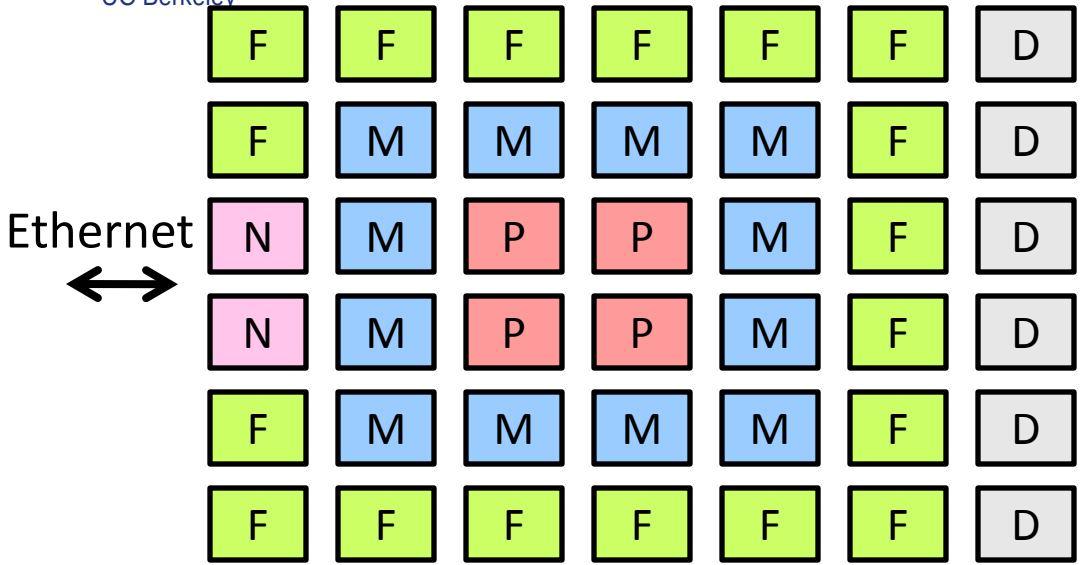


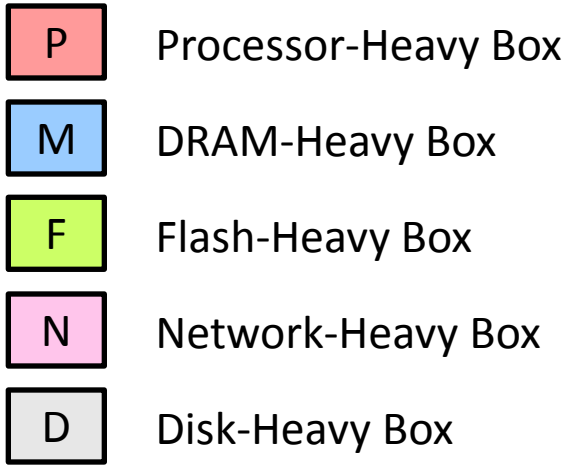
Up to ~1000 modules of any kind: Processor, DRAM, Flash

Many Pb/s photonic network

Massive Redundancy for Dependability

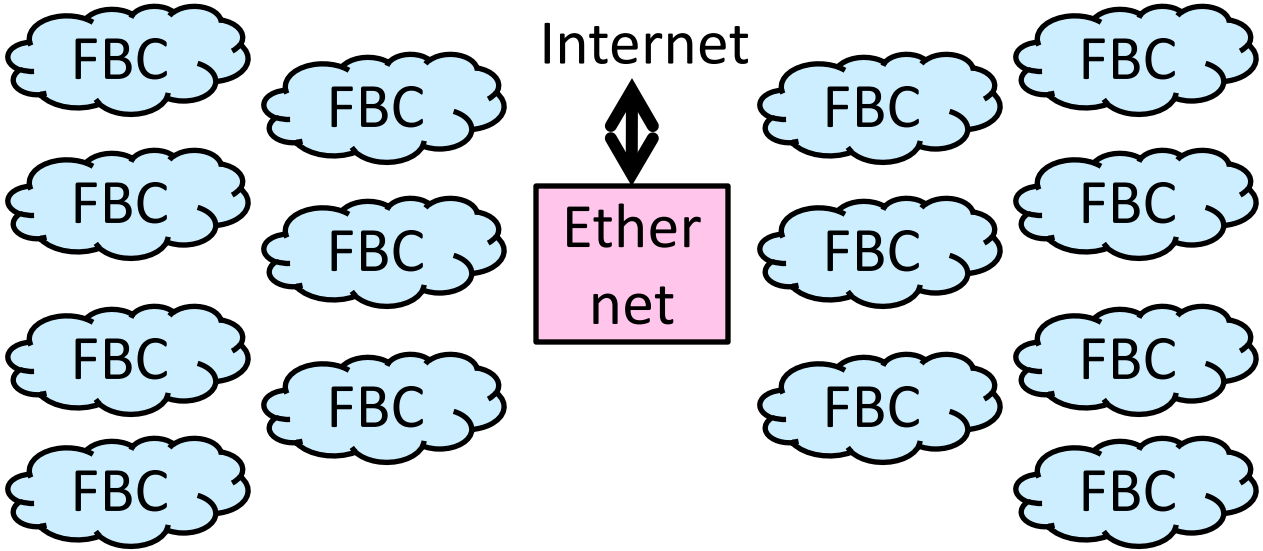
FireBox Cluster





- Configure P/M/F/N/D mix for workload & budget
- Each FireBox has up to ~1000 sockets
  - Each box can contain any mix of P/M/F/N/D modules, not just one kind
  - Switches located inside boxes
- 20 - 40 FireBox Racks / Cluster, Cluster is app target
- Network box connects WSC to standard networks (e.g. 400GbE)
- Disk boxes hold local disk arrays

Warehouse of FireBox Clusters (FBC)



- ≈50 KW/FireBox => ≈1 - 2 MW/Cluster
  => 10 - 20 Clusters/WSC
- Ethernet datacenter router to connect FBCs to each other and to Internet

ASPIRE
UC Berkeley

Acknowledgements

Help from ASPIRE insiders:

- Elad Alon
- Scott Beamer
- Joao Carriera
- Pi-Feng Chiu
- Adam Izraelevitz
- Kurt Keutzer
- Forrest Iandola
- Yunsup Lee
- Eric Love
- Martin Maas
- Bora Nikolic
- Frank Nothaft
- David Patterson
- Nathan Pemberton
- Colin Schmidt
- Vladimir Stojanovic
- Chen Sun
- Andrew Waterman

Help from ASPIRE outsiders:

- Michaela Blott, Xilinx
- Bob Brennan, Samsung
- Eric Brewer, UCB/Google
- Zhangxi Tan, UCB/Pure
- Amin Vahdat, Google
- Kees Vissers, Xilinx

ASPIRE Sponsors:

- Founding: DARPA, SRC C-FAR, Intel
- Affiliates: Google, HP, Huawei, LGE, NVIDIA, Oracle, Samsung

ASPIRE
UC Berkeley

Summary

- Solid-state datacenter is already here, now need to architect chips/systems to take advantage
- Custom chips are coming back
  - Vertical integration makes sense at warehouse-scale
  - Open-source software developed for open hardware
  - 99.9% of software identical, only change the secret sauce
  - Chip NRE costs will drop (or we're all doomed)
- Specialized coprocessors, not specialized cores
  - Ease of software development
  - Avoid latency & energy costs to move data out of caches
  - WSC management prefers arrays of homogeneous cores