FireSim: Productive, Scalable, FPGA-Accelerated Cycle-Accurate Hardware Simulation using Cloud FPGAs

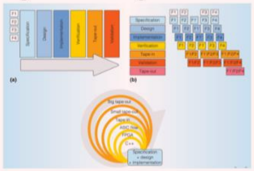
Sagar Karandikar, Howard Mao, Donggyu Kim, David Biancolin, Alon Amid, Borivoje Nikolic, Randy Katz, Jonathan Bachrach, Krste Asanovic.
Dept. of Electrical Engineering and Computer Sciences, UC Berkeley. Contact: sagark@eecs.berkeley.edu

XDF

Why build a scalable FPGA-accelerated HW simulator?

- Rapidly expanding ecosystem of open HW, but no fast, accessible simulators

Agile HW Design for ASICs



FPGAs in the Cloud

aws

Microsoft Azure

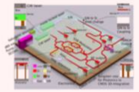
Alibaba Cloud

HUAWEI

- Next-gen datacenters won't be built from commodity components:

The end of Moore's Law

Custom Silicon in the Cloud

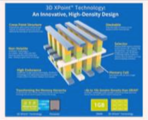


Fast networks

e.g. Silicon Photonics

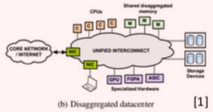
Deeper memory/storage hierarchies

e.g. 3DXPoint



New DC organizations

e.g. disaggregation



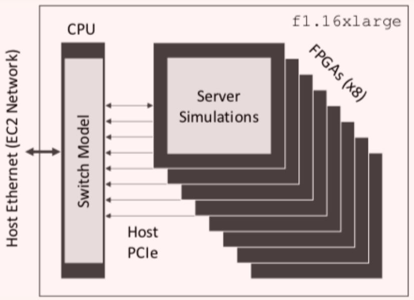
(b) Disaggregated datacenter [1]

FireSim Features

- **Model HW at scale, with high simulation rate (10s-100s MHz):**
  - CPUs down to microarchitecture (automatically transformed from Chisel RTL to cycle-exact FPGA Simulator)
  - Network links/switches (C++ software models via HW/SW co-simulation)
  - Novel accelerators/other hardware (since can transform arbitrary Chisel)
  - Validated abstract models for standardized components like DRAM (MIDAS Memory Model)
- **Run real software:**
  - Real OS, networking stack (Linux)
  - Real frameworks/applications (e.g. Memcached, Ray, Caffe)
- **Be highly-productive:**
  - Uses a commodity platform (EC2 F1)
  - Highly-automated: `firesim` command-line to manage/deploy sims
    - Similar to `docker` or `vagrant`, but for FPGA Simulators
    - Reproducible: Included scripts to reproduce results from ISCA'18 paper
  - Encourage collaboration between systems devs and architects

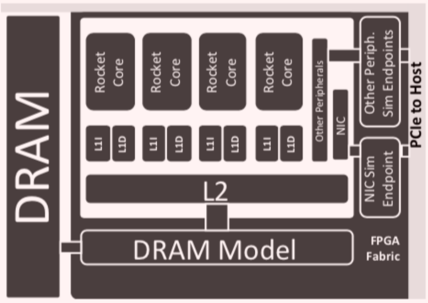
Mapping a simulation to EC2 F1

- Server Simulation
  - Highly-parallel if expressed as target RTL
  - We have the RTL: transform into a model
  - Put it on the FPGAs
- Network simulation
  - Little parallelism in switch models (e.g. a thread per port)
  - Need to coordinate all of our distributed server simulations
  - So use CPUs + host network



Example Datacenter Target Design

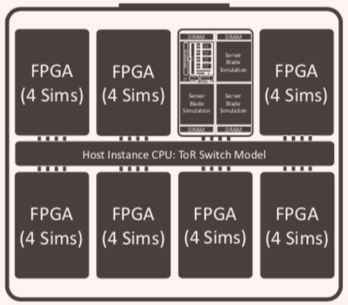
- Server blades (RTL), each with:
  - Quad-core RISC-V Rocket @ 3.2 GHz
  - 16 KiB I$, 16 KiB D$, 256 KiB L2
  - 16 GB DRAM
  - 200 Gbps Ethernet NIC
  - Optional Accelerators
  - Single node:
    - Runs at 150+ MHz (no net), 40 MHz (net)
    - Costs 40 cents/hour on EC2 spot market
- High-performance network (SW):
  - Parameterizable BW/link latency
    - e.g. 200 Gbps, 2μs
  - Easy to add your own link-layer
    - We provide Ethernet
  - Switches with configurable # of ports
  - Configurable topology



FireSim Server Blade Block Diagram

Rack-scale (32-node) simulation metrics

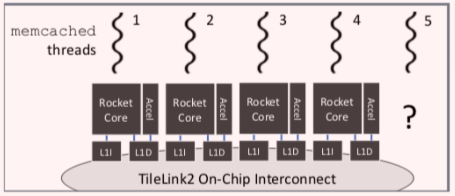
- Four quad-core server simulations per FPGA
  - = 32 server simulations per f1.16xlarge
  - = 128 simulated cores per f1.16xlarge
  - One simulation management thread per-FPGA
- 32-port, 200 Gbps per-port ToR switch model
  - One thread-per-port (f1.16xlarge has 64 vCPUs)
- Runs at ~10.7 MHz, ~1.4 billion insts/sec
- $13.20/hr on-demand, ~$2.60/hr spot



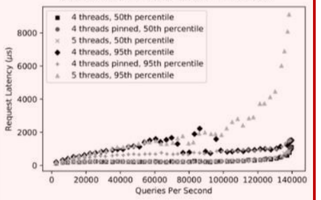
Reproducing end-to-end application latency effects from real clusters

- Leverich and Kozyrakis show effects of thread-imbalance in memcached in *EuroSys '14*: tail-latency is adversely affected, median latency is not
- We run `memcached` on one node and run the `mutilate` load generator on 7 nodes

Thread-imbalance: 5 threads on a 4 core system:



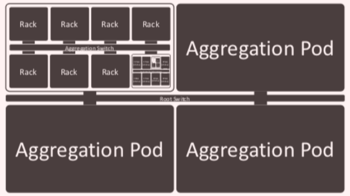
Reproducing thread-imbalance latency effects on an 8-node cluster in FireSim



1024-Node Datacenter Simulation

- Can scale to:
  - 1024 server blades (4096 cores)
  - 32 ToR Switches
  - 4 Aggregation, 1 Root switch
- Host resources:
  - 32 f1.16xlarges (256 FPGAs)
  - 5 m4.16xlarges (aggregation, root switches)
  - 5 m4.16xlarges
- Runs at 6.6 MHz (27 billion insts/s)
- Sample `memcached` run (512 servers, 512 load-generators):

	50th Percentile (μs)	95th Percentile (μs)	Aggregate Queries-Per-Second
Cross-ToR	79.3	128.2	4.7 million
Cross-aggregation	87.1	111.3	4.5 million
Cross-datacenter	93.8	119.5	4.1 million



1024 Server Simulation Topology

Latest Updates

- FireSim is now open-source! `https://fires.im`

Acknowledgements/References

The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000849, DARPA Award Number HR0011-12-2-0016, RISE Lab sponsor Amazon Web Services, and ADEPT/ASPIRE Lab industrial sponsors and affiliates Intel, HP, Huawei, NVIDIA, and SK Hynix. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or of the industrial sponsors.