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

Why build a scalable FPGA-accelerated HW simulator?
- Rapidly expanding ecosystem of open HW, but no fast, accessible simulators
- Agile HW Design for ASICs
- FPAGAs in the Cloud
- Next-gen datacenters won’t be built from commodity components:
  - Higher memory/storage hierarchies e.g. NVMe
  - New DC organizations e.g. disaggregations

FireSim Features
- Model HW at scale, with high simulation rate (10s-100s MHz):
  - CPUs down to microarchitecture (automatically transformed from Chisel RTL to cycle-accurate FPGA Simulator)
  - Network links/switches (C++ software models via HW/SW co-simulation)
  - Novel accelerators/other hardware (once 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 sim
  - 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 FPAGAs
- 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 @ 3.3 GHz
  - 16 KB I-cache, 16 KB D-cache, 256 KB L2
  - 16 GB DRAM
  - 200 Gbps Ethernet NIC
  - Optional Accelerators
  - Single node:
    - Runs at 350 MHz (no nic), 40 MHz (net)
    - Costs 40 cents/hour on EC2 spot market
- High-performance network (SW):
  - Parametrizable BW/link latency
    - e.g. 200 Gbps, 25ns
  - Easy to add your own link-layer
    - We provide Ethernet
  - Switches with configurable # of ports
  - Configurable topology

Rack-scale (32-node) simulation metrics
- Four quad-core server simulations per FPGA
  - 32 server simulations per f1.16large
  - 128 simulated cores per f1.16large
  - One simulation management thread per FPGA
- 32-port, 200 Gbps per-port ToR switch model
  - One thread-per-port (1.16large has 64 vCPUs)
  - Runs at ~107 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
- LeVeque 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 multilane load generator on 7 nodes

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 HR0011-12-2-0018, NSF’s 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.