Using FPGAs to Simulate Novel Datacenter Network Architectures at Scale

Zhangxi Tan, Krste Asanovic, David Patterson

Datacenter Network Infrastructure

- ❑ Network infrastructure is the "SUV of datacenter"
  - ▪Large Cisco switches/routers are expensive and unreliable
  - ▪Important for many optimizations
    - Improving server utilization (power consumption)
    - Supporting data intensive map-reduce jobs
- ❑ Many network architectures proposed recently
  - ▪VL2, Portland , Dcell, Thacker's container switch
- ❑ Different observations lead to many distinct design features
  - ▪Switch designs
  - ▪Network designs
  - ▪Application and protocols

Problems of Existing Evaluations

- ❑Scale is way smaller than real datacenter network
  - • << 100 nodes vs. O(10,000) nodes
- ❑Synthetic programs and benchmarks
  - •Datacenter Programs: Web search, email, map/reduce
- ❑Off-the-shelf switches architectural details are NDA
  - •Limited architectural design space configurations: E.g. change link delays, buffer size and etc.

Our Approach

- Build a "wind tunnel" for datacenter network using FPGAs
  - □Simulate O(10,000) nodes: each is capable of running real software
  - □Simulate O(1,000) datacenter switches (all levels) with detail and accurate timing
  - □Runtime configurable architectural parameters (link speed/latency, host speed)
  - □**Build on top of RAMP Gold**: A full-system FPGA simulator for manycore systems
  - □Prototyping with a rack of BEE3 boards

Node Software

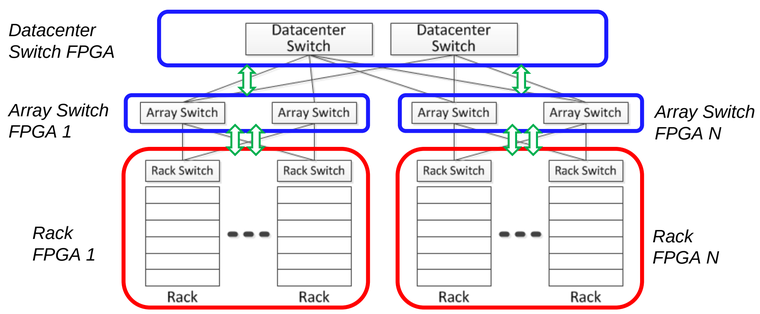
- LAMP + map/reduce
- Web 2.0 benchmarks
- Some research code and production code
  - E.g. Twitter memcached

Implementation

- Single FPGA Implementation (current)
  - □$750 Xilinx XUP V5 board
  - □64 cores (single pipeline), 2GB DDR2, FP, processor timing model, ~1M target cycles/second, 260x faster than SW
  - □Boot Linux 2.6.21 and Research OS

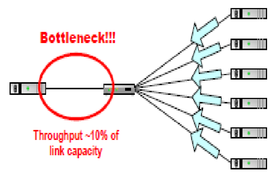
- Multi-FPGA Implementation for datacenter simulation (pending)
  - □BEE3 : 4 Xilinx Virtex 5 LX155T
  - □~512K cores +, 64GB DDR2, FP, timing model

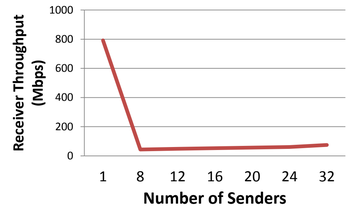
Simulator Architecture



- Modularized single-FPGA designs: two types of FPGAs
- Connecting multiple FPGAs using multi-gigabit transceivers according to physical topology

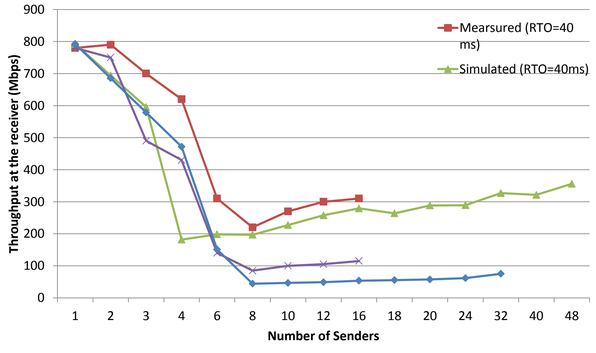
Case Study: Reproduce the TCP Incast Problem



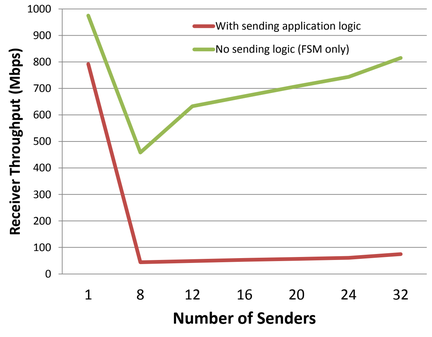


A TCP throughput collapse that occurs as the number of servers sending data to a client increases past the ability of an Ethernet switch to buffer packets.

Simulation vs. Measurement

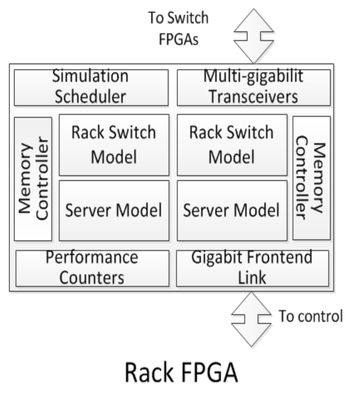


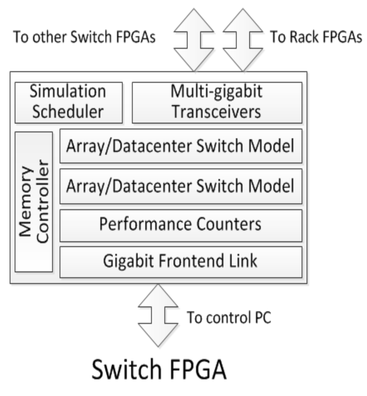
Importance of Node Software



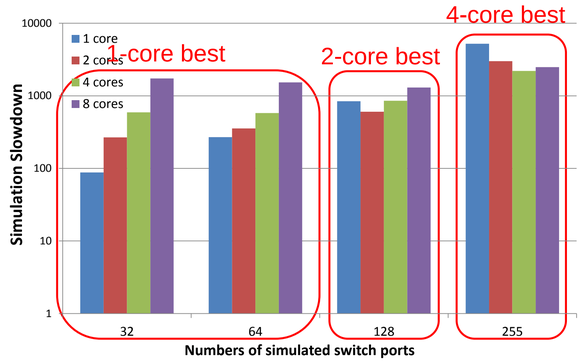
Simulator Scaling

- On 2007 FPGAs (Xilinx Virtex 5 LX155T)
  - □2 RAMP Gold pipelines, 128 servers on one FPGA, 512 on one BEE3 board;
  - □256 MB memory per simulated server
  - □88 FPGAs with 22 BEE3 boards for a 10,000 system
- On 2011 FPGAs (Xilinx Virtex 7)
  - □16 RAMP Gold pipelines, 1024 servers on one FPGA
  - □128 MB memory per simulated server
  - □11~12 single-FPGA boards for a 10,000 system
  - □Simulator cost: ~$120K
    - Board cost: $5,000 * 12 = $60,000
    - DRAM cost: $600 * 8 * 12 = $57,000
  - □O(10,000) real datacenter cost:
    - $36M in CAPEX, $800K in OPEX/mo.





FPGA vs. Parallel Software Simulator



- □Multi-core never better than 2x single-core
- □Top two software simulation overheads
  - □flit-by-flit synchronizations
  - □network microarchitecture details

Conclusion & Contribution

- Simulate node hardware with software at the scale of O(10,000)
- Node software significantly affects the simulation result
- RAMP Gold is promising for container-level experiments
- Production release: Q4, 2011