SeaMicro SM10000-64 Server

Building Datacenter Servers Using Cell Phone Chips

Ashutosh Dhodapkar, Gary Lauterbach, Sean Lie, Dhiraj Mallick, Jim Bauman, Sundar Kanthadai, Toru Kuzuhara, Gene Shen, Min Xu, Chris Zhang

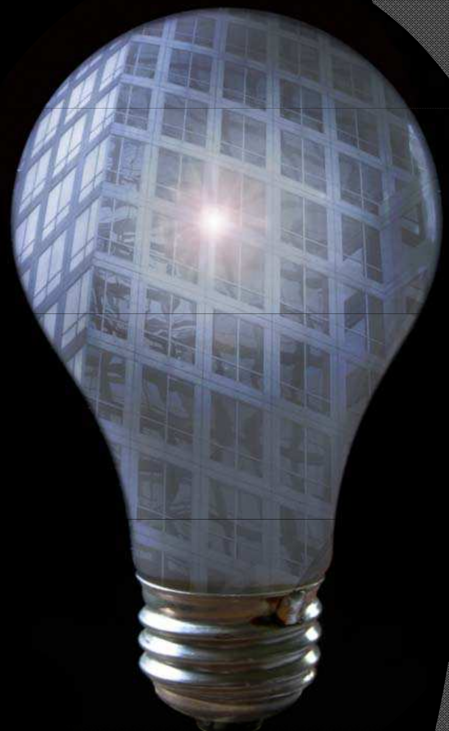
SeaMicro

Overview

- Power in the Datacenter
- Application Trends
- SeaMicro Architecture
  - CPU Selection
  - Interconnect Fabric
  - I/O Virtualization
  - Management Software
- Application Performance
- Summary

Power: The Issue in the Datacenter

- Power is the largest Op-Ex item for an Internet company; >30% of Op Ex
- Volume servers consume 1% of the electricity in the US – More than $3 Billion dollars per year *
- Datacenters reaching power limits
  - Reducing power will extend life of existing datacenters – saving 100's of millions of dollars in CapEx



* 2007 EPA Report to Congress on Server and Data Center Efficiency, Public Law 109-43
Power Provisioning for a Warehoused Sized Compute

Cloud's Killer Apps

- Compute moving to server side
  - Clients primarily for display: smart phones, tablets, etc.
- Free to users - cost amortized over large user base
  - Optimizing datacenter TCO is critical to profitability
  - Costs: bandwidth, power, servers, switching, storage, firewalls, load balancers, buildings, management
- Growing exponentially
  - Datacenters facing new challenges around logistics of "super-size"

- Killer apps
  - are collections of small, simple, and bursty workloads
  - have big data sets, high concurrency, lots of communication

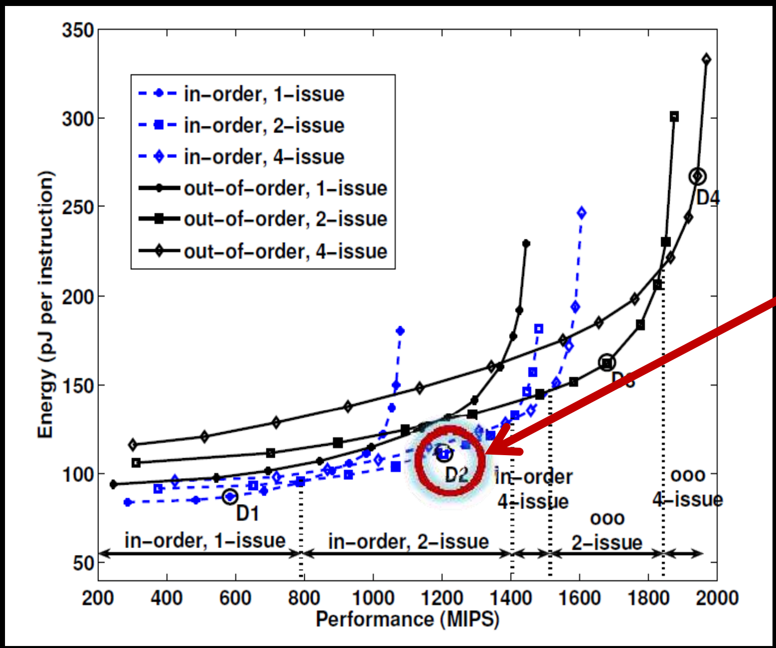
SeaMicro Architecture

- Cluster in a box: integrated compute, storage, and network
- Large number of inexpensive, energy-efficient "Cell Phone" CPUs interconnected using a low-latency, high-bandwidth fabric
- Unlike multi-core, provides natural scaling of all system resources:
  - O/S threads, networking stack
  - Memory bandwidth and capacity
  - NICs and network bandwidth
  - Disk controllers, disk bandwidth/capacity
- Purpose built for Cloud's Killer Apps: architecture maps well to internet workloads

SeaMicro Architecture: CPU

Purpose built for Cloud's Killer Apps

- Collection of small, simple, and bursty workloads
- Big data sets, high concurrency, lots of communication



CPU: Intel Atom™, a "Cell Phone" CPU

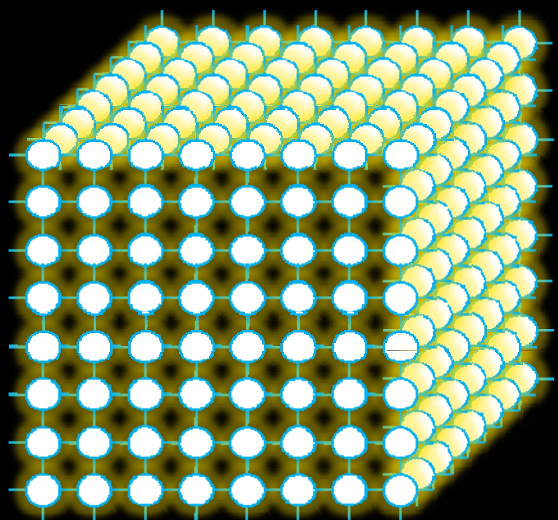
- Better match the workload
- Leverage efficiencies of Scale Down: operate at a more efficient point on the energy-performance curve*
- Derive cost advantage from smaller silicon area and higher volumes
- Superior Performance/Watt/$ compared to server class CPUs

* Adapted from Azizi et al. "Energy-Performance Tradeoffs in Processor Architecture and Circuit Design: A Marginal Cost Analysis," ISCA 2010

SeaMicro Architecture: Fabric

Purpose built for Cloud's Killer Apps

- Collection of small, simple, and bursty workloads
- Big data sets, high concurrency, lots of communication



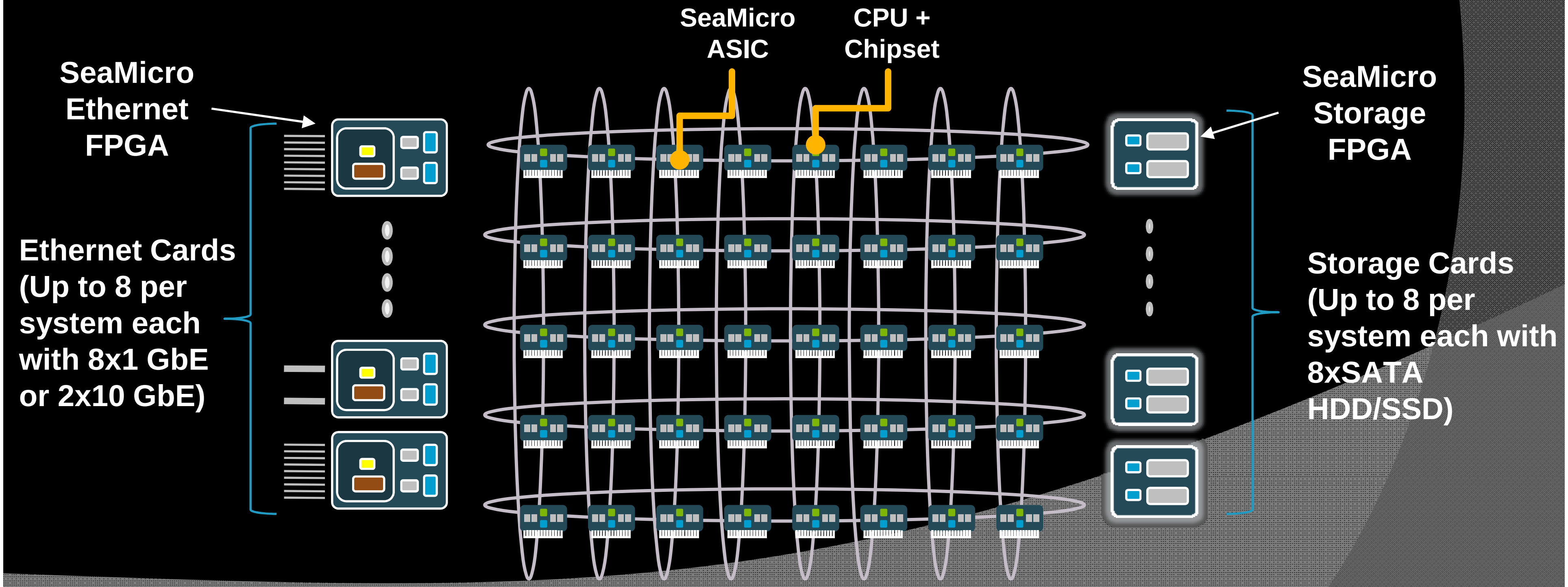
512 CPUs interconnected using a high bandwidth fabric in a 3D torus topology

- High scalability: distributed architecture based on low-power ASICs
- High bandwidth: 1.28Tbps
- Low-latency: < 5us between any two nodes
- High resiliency: multitude of paths between two nodes allowing easy routing around failures

SeaMicro Architecture: I/O Virtualization

Virtualized I/O devices

- Network/storage shared and amortized over large number of CPUs
- Improves utilization and enables optimizations, e.g. shared partitions
- Reduces components in the system, thus improving cost/power

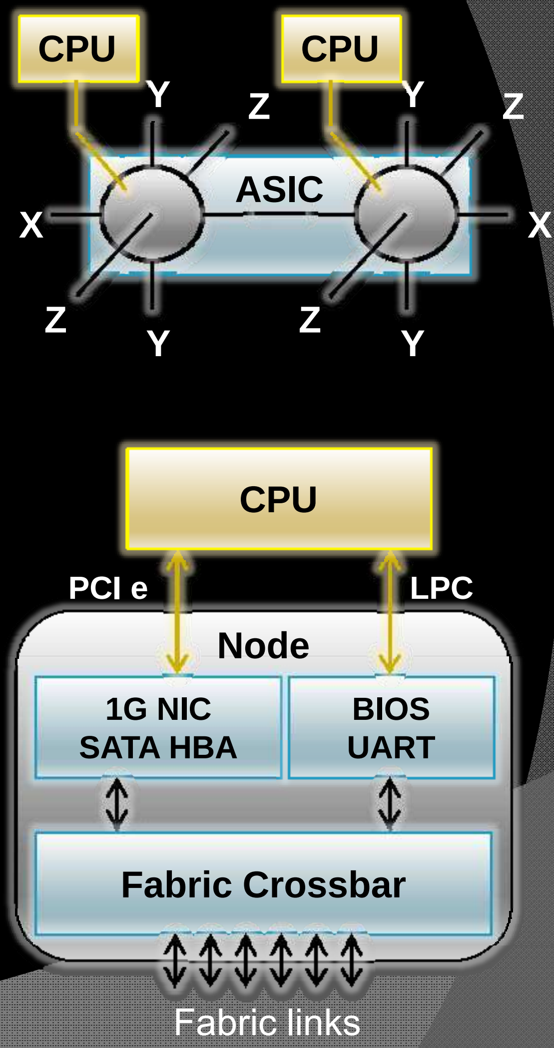


SeaMicro ASIC

- 90nm G TSMC technology
- 15 mm^2 per node
- 289 pin plastic BGA package
- Low power: <1W per node
- Key Features:
  - Fabric switch
  - I/O virtualization
  - Clock generation
  - Node management

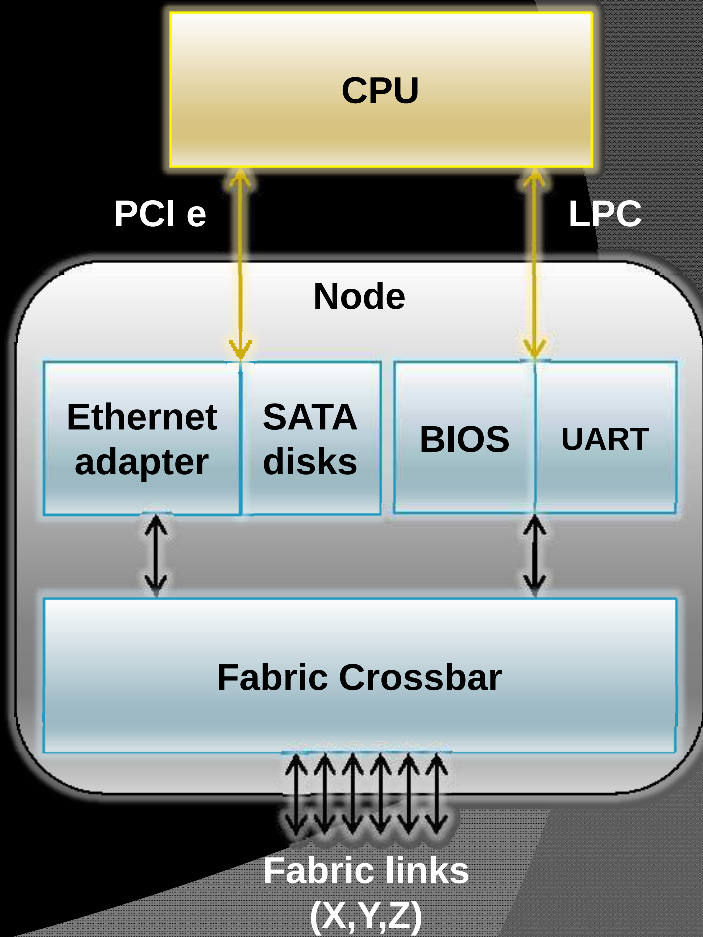
SeaMicro Fabric

- 512 logical fabric nodes
  - ASIC contains 2 fabric nodes
- Each node has
  - 6 fabric links (2.5Gb/s SERDES) to neighboring nodes (2X, 2Y, 2Z)
  - 1 PCIe link to a CPU
  - Crossbar to switch between 7 links
- Nodes connected in an 8x8x8 3D Torus
  - Each dimension is an 8-node loop
  - Total bandwidth is 1.28 Tbps
  - High path diversity for redundancy
- Fabric is cut-through, loss-less, deadlock free, has multiple QOS levels



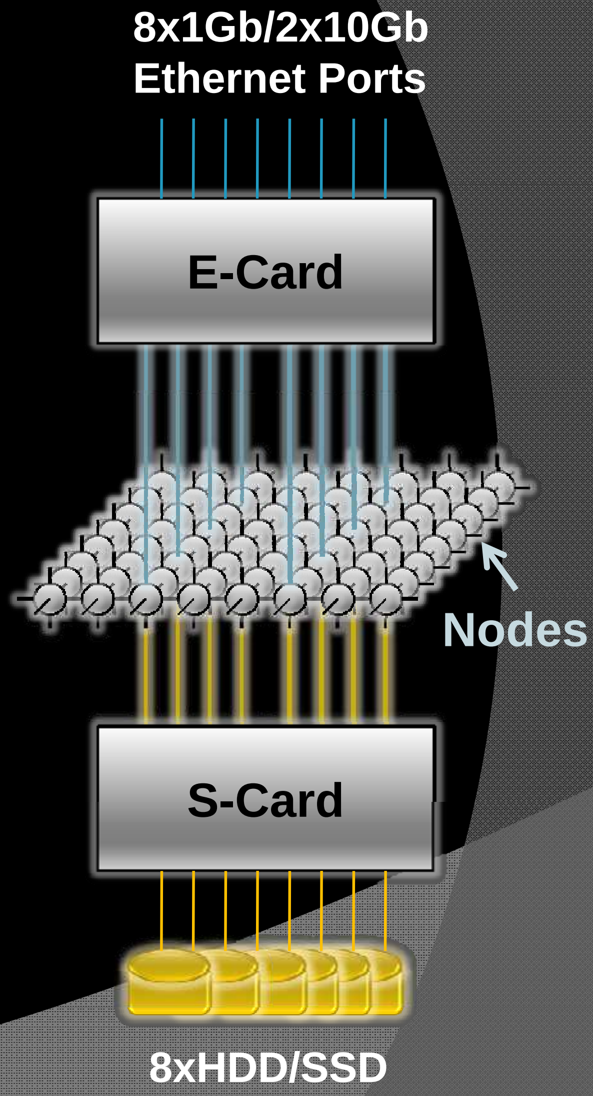
I/O Virtualization

- Node presents 4 virtual devices to CPU
  - PCIe: Ethernet adapter, 4 SATA disks
  - LPC (ISA): BIOS, UART
- Node packetizes data from CPU and routes it through the fabric to multiple shared I/O cards
  - Ethernet traffic is routed to other nodes or network I/O cards with uplink ports
    - Each node is a port on a distributed switch. Internal MAC addresses are hidden behind I/O card
    - Table at ingress port on I/O card provides fabric destination
    - Nodes keeps track of destination ports for external MACs. Fabric address encoded in Internal MAC
  - Disk, BIOS, and Console requests are routed to storage I/O cards which hold up to 8 SATA disks

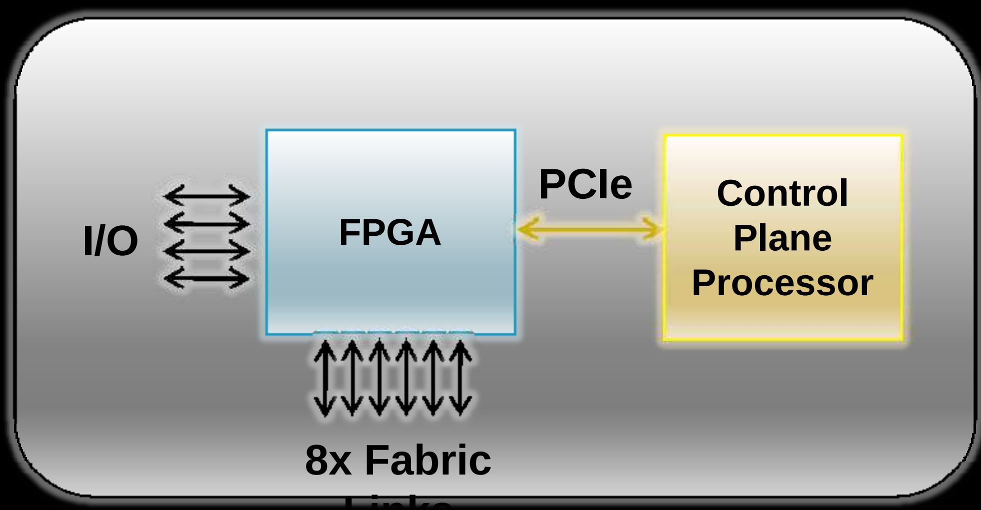


I/O Aggregation Cards

- Bridge between the fabric and I/O
  - Connected to the fabric in the Y-dimension
  - Terminate fabric protocol on one side and talk to I/O devices on the other
- Two types of I/O cards
  - E-Card: 1G/10G network connectivity
  - S-Card: SATA storage, BIOS, Console
- Any node can talk to any I/O card
  - 1 E-Card and 1 S-Card per Z-plane



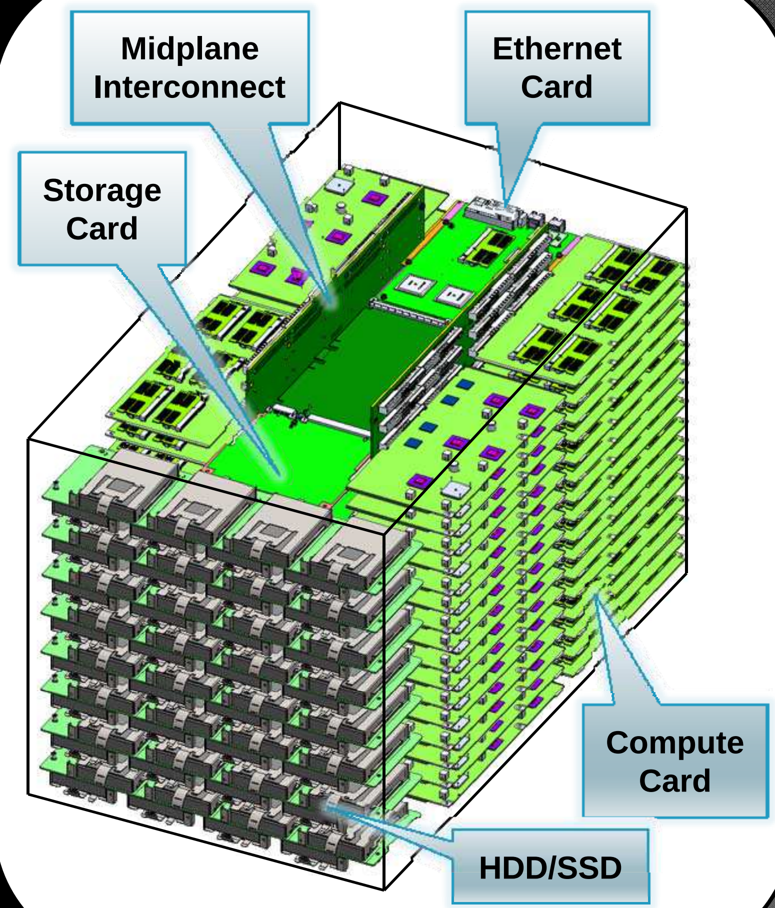
I/O Card Architecture



- Co-designed architecture
  - HW/SW boundary flexible: optimized for performance
  - High speed datapaths implemented in FPGA
  - Control plane implemented in microprocessor
  - FPGA enables rapid feature enhancement based on customer feedback. Power/cost amortized over 100's of nodes

SM10000-64 Server

- 10 RU chassis
- Dual mid-plane fabric interconnect
- 64 compute cards
  - 4 ASICs/card: 512 fabric nodes
  - 4-6 Dual-core CPUs/card: 512-768 cores
  - 4GB DRAM/CPU: 1-1.5TB DRAM
- 1-8 shared Ethernet cards:
  - Up to 160 Gb/s external connectivity
- 1-8 shared Storage cards:
  - Up to 64 SATA/SSD drives
- Shared infrastructure:
  - N+1 redundant Power supplies, fans, management Ethernet and console



Management Software

- Implements key real-time services
  - Fabric routing: fault isolation and failover
  - Ethernet control plane (MAC/IP learning, IGMP)
  - Layer4 load balancer management
  - Terminal server
  - Power supply and fan control
- Configuration, Management, and Monitoring
  - Integrated DHCP server
  - CLI and SNMP interfaces for network/storage/server management
  - Performance/Fault monitoring tools

Benchmark: Apache Bench

SM10000-64 consumes 1/4th the power, for equivalent performance

- Apache 2.2 with Apache Bench. CPUs running CentOS 5.4
- Retrieve 16KB files over 10 min.

	SeaMicro	1U Xeon Server
System Configuration	**SM10000-64 256 x Dual Core 1.66GHz Atom processors**	Industry Standard Dual Socket Quad Core Xeon L5630 2.13GHz
Systems Under Test	**1**	45
Apache Throughput/Sec	**1,005,056**	1,005,120
Apache Request File Size	**16KB**	16KB
System Power	**2,490W**	10,090W
Space consumed in Racks	**10 RU**	45 RU

Summary: SM10000-64

- Internet workloads are increasingly moving compute to the server side.
- Minimizing Power and thus TCO of datacenters is critical to internet businesses.
- The SM10000-64 is a major step forward to address the challenges of the datacenter
  - Provides a 4x reduction in power/space for equivalent performance, compared to traditional 1RU servers.

Shipping in volume

- 768 Intel Atom cores, 1.5 TB DRAM, 1.28Tbps fabric
- 64 SATA/SSD disks, 160Gbps uplink network bandwidth
- Integrated load balancer and management SW

Thank You!