Microsoft Research

Microsoft

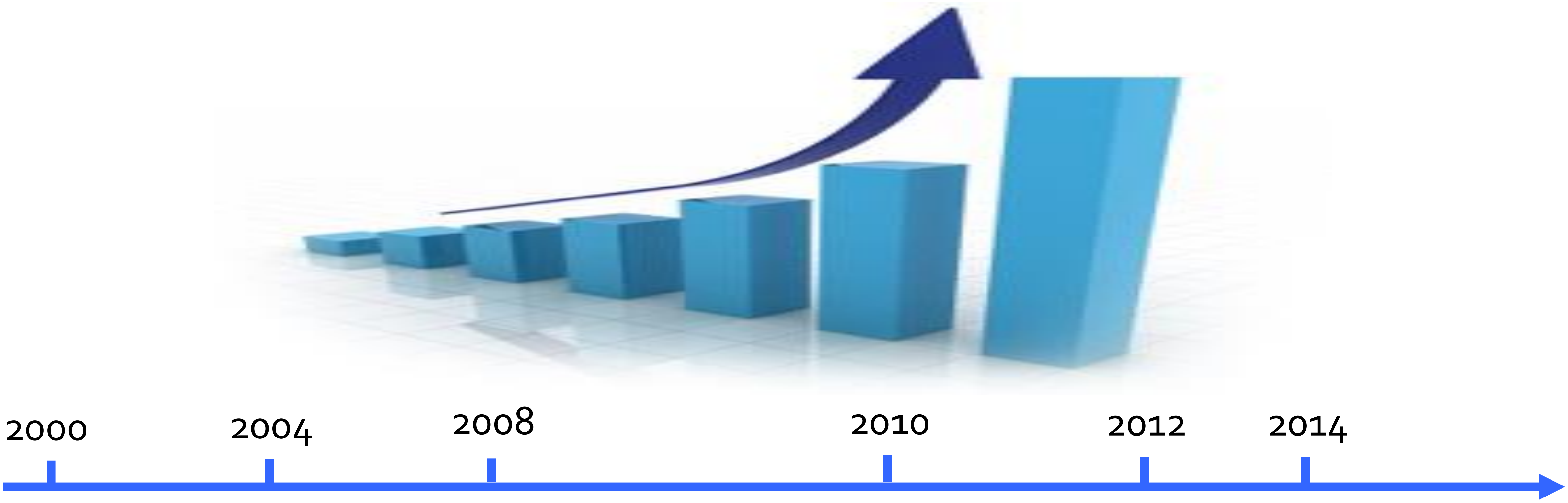
Towards Rack-scale Computing
Challenges and Opportunities

Paolo Costa

paolo.costa@microsoft.com

joint work with
Raja Appuswamy, Hitesh Ballani, Sergey Legtchenko, Dushyanth Narayanan, Ant Rowstron

Hardware Evolution in Data Centers



2000 2004 2008 2010 2012 2014

Goal

Increase work done per dollar (CapEx + OpEx)

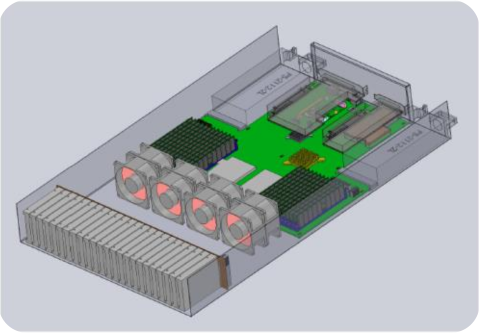
Hardware Evolution in Data Centers

2000
2004
2008
2010
2012
2014

Scale out vs. scale up

Many commodity servers rather than few expensive servers

Hardware Evolution in Data Centers



2000

2004

2008

2010

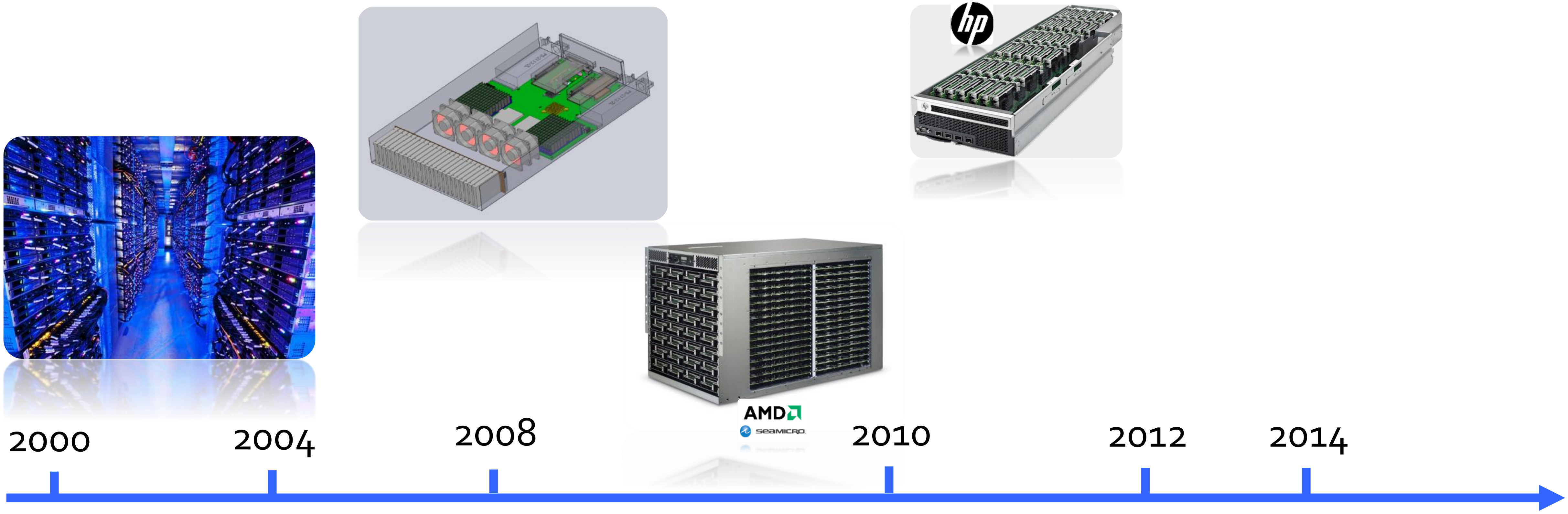
2012

2014

Custom layout

Remove unnecessary components (e.g., GPGPUs, USB ports)

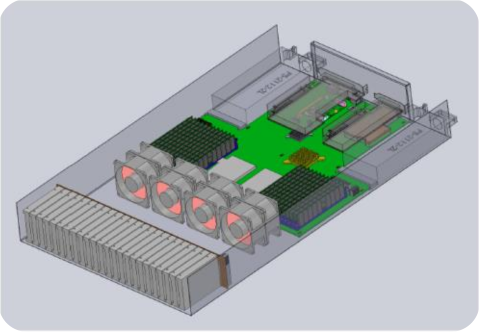
Hardware Evolution in Data Centers

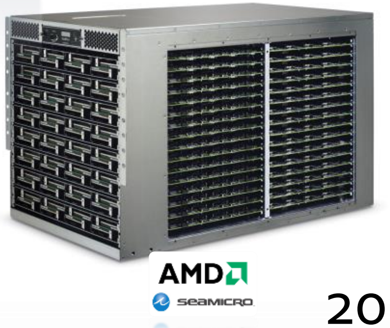


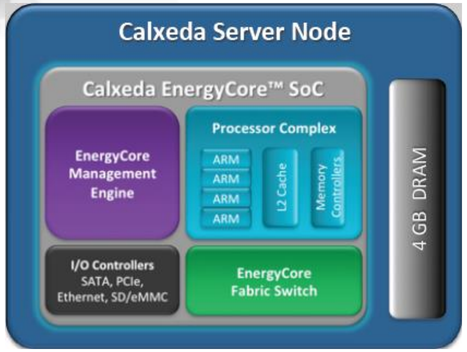
Integrated fabrics

Higher density and lower power consumption

Hardware Evolution in Data Centers





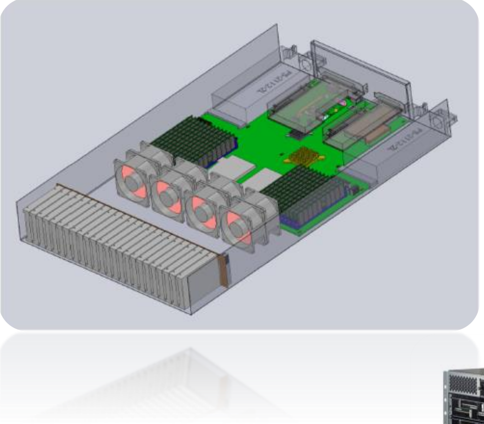


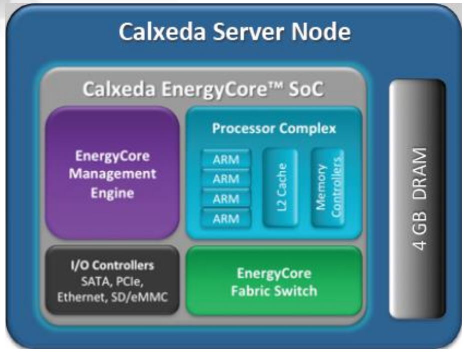
2000 2004 2008 2010 2012 2014

System-on-Chip (SoC)

CPU, IO controllers, NIC/fabric switch on the same die

Hardware Evolution in Data Centers





Calxeda Server Node

Calxeda EnergyCore™ SoC

EnergyCore Management Engine

Processor Complex

ARM ARM ARM ARM

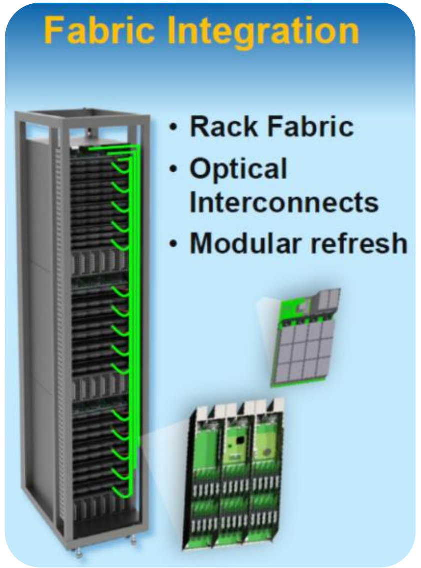
L2 Cache

Memory Controllers

4 GB DRAM

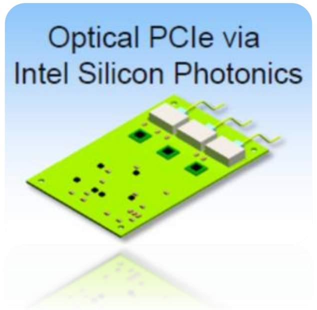
I/O Controllers SATA, PCIe, Ethernet, SD/eMMC

EnergyCore Fabric Switch



Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular refresh



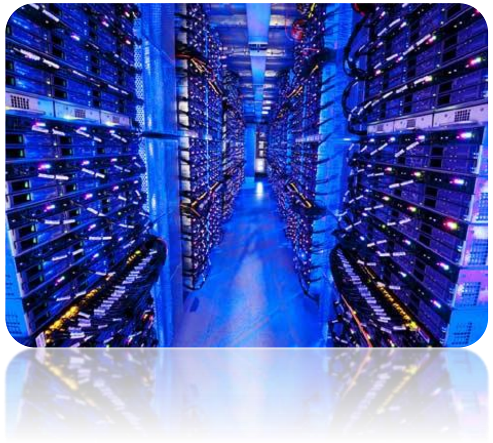
Optical PCIe via Intel Silicon Photonics

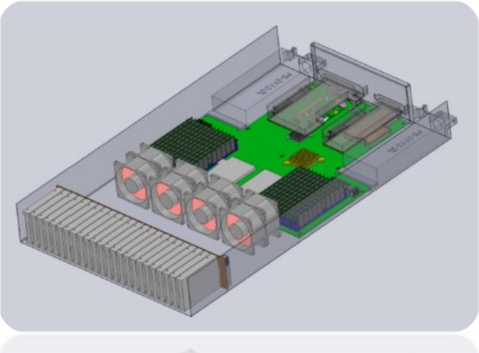
2000 2004 2008 2010 2012 2014

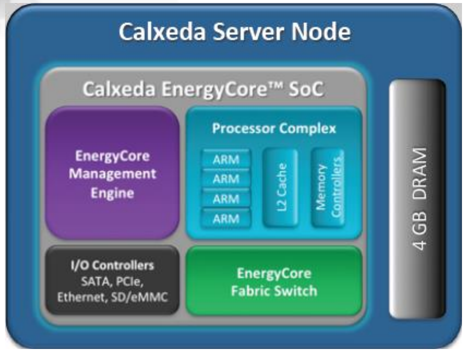
Silicon Photonics

High-bandwidth / low-latency interconnect (resource disaggregation)

Hardware Evolution in Data Centers







Calxeda Server Node

Calxeda EnergyCore™ SoC

EnergyCore Management Engine

Processor Complex

ARM

ARM

ARM

ARM

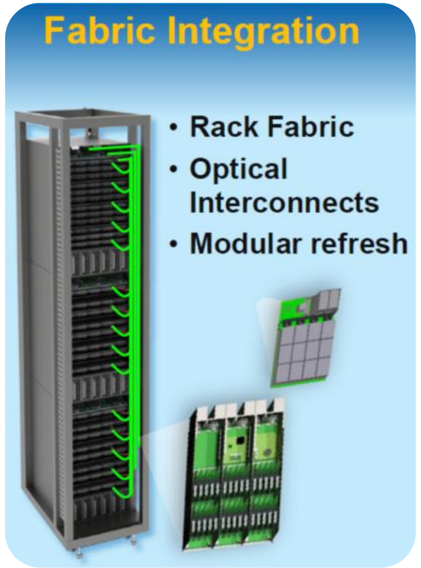
L2 Cache

Memory Controllers

4 GB DRAM

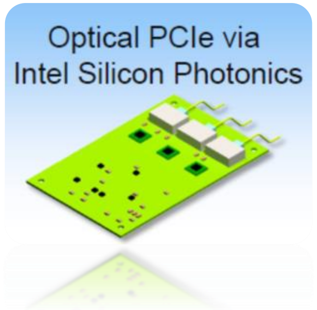
I/O Controllers SATA, PCIe, Ethernet, SD/eMMC

EnergyCore Fabric Switch



Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular refresh

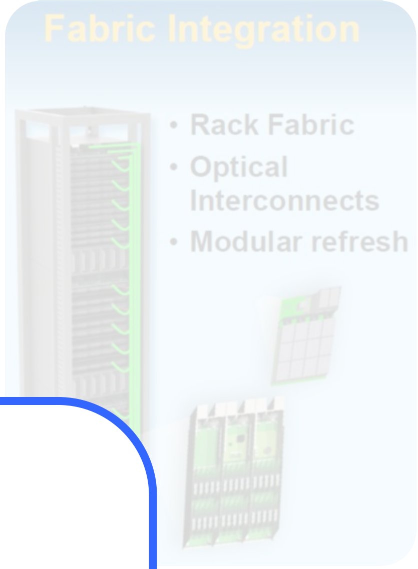


Optical PCIe via Intel Silicon Photonics

2000 2004 2008 2010 2012 2014

1 rack unit (RU) 2 Ru 4-10 Ru Rack-scale

Hardware Evolution in Data Centers



Rack-scale Computers

The rack is the new unit of deployment in data centers
Sweet spot between single-server and cluster deployments

2000

1 rack unit (RU) 2 Ru 4-10 Ru Rack-scale

Rack-scale Computer in 2020?

	Today’s traditional rack	**2020** **Rack-scale Computer**
#Cores (#servers)	~100s (20-40)	~100,000s (1,000s)
Memory	~1 TB	~100s TB
Storage	~100 TB (flash + spinning disk)	~100s PB (NVM)
Bandwidth / server	10 Gbps	1 Tbps

How far are we from rack-scale computing?

	Today's rack	2020 Rack-scale Computer	2014 Rack-scale Computer
#Cores (#servers)	~100s (20-40)	~100,000s (1,000s)	
Memory	~1 TB	~100s TB	
Storage	~100 TB (SSD/HDD)	~100s PB (NVM)	
Network	10 Gbps / server	1 Tbps / server	

How far are we from rack-scale computing?

	Today's rack	**2020** **Rack-scale Computer**	**2014** **Rack-scale Computer**
#Cores (#servers)	~100s (20-40)	~100,000s (1,000s)	*~1,000s* *(100s-1000s)*
Memory	~1 TB	~100s TB	
Storage	~100 TB (SSD/HDD)	~100s PB (NVM)	
Network	10 Gbps / server	1 Tbps / server	

Core count

- AMD SeaMicro SM15000-64
  - 512 cores (64 servers) in 10 RU
  - 2,048 cores (256 servers) at rack scale
- HP Moonshot Redstone
  - 1,152 cores (288 servers) in 4U
  - 11,520 cores (3,200 servers) at rack scale
- Boston Viridis
  - 192 cores (48 servers) in 2 RU
  - 7,680 cores (1,920 servers) at rack scale

How far are we from rack-scale computing?

	Today's rack	2020 Rack-scale Computer	2014 Rack-scale Computer
#Cores (#servers)	~100s (20-40)	~100,000s (1,000s)	*~1,000s* *(100s-1000s)*
Memory	~1 TB	~100s TB	*~10 TBs*
Storage	~100 TB (SSD/HDD)	~100s PB (NVM)	*~1 PB* *(SSD/HDD)*
Network	10 Gbps / server	1 Tbps / server	

Memory

- AMD SeaMicro SM15000-XE
  - 2 TB RAM (32 GB/server) in 10 RU
  - 8TB RAM at rack scale
- HP Moonshot Redstone
  - 1.12 TB (4 GB/server) in 4U
  - 11.25 TB at rack scale

Storage

- AMD SeaMicro FS-5084-L
  - 336 TB storage in 5 RU
  - 2.5 PB at rack scale

How far are we from rack-scale computing?

	Today's rack	2020 Rack-scale Computer	2014 Rack-scale Computer
#Cores (#servers)	~100s (20-40)	~100,000s (1,000s)	*~1,000s (100s-1000s)*
Memory	~1 TB	~100s TB	*~10 TBs*
Storage	~100 TB (SSD/HDD)	~100s PB (NVM)	*~1 PB (SSD/HDD)*
Network	10 Gbps / server	1 Tbps / server	*~10s -100s Gbps / server*

Network

- AMD SeaMicro SM15000-XE
  - 1.28 Tbps fabric (20 Gbps / server)
- Mellanox ConnectX-3 Pro
  - 2x 40-Gbps NICs
- Intel MXC Connector (expected Q3'14)
  - Up to 32 fibers (25 Gbps / fiber)
  - Up to 800 Gbps / server

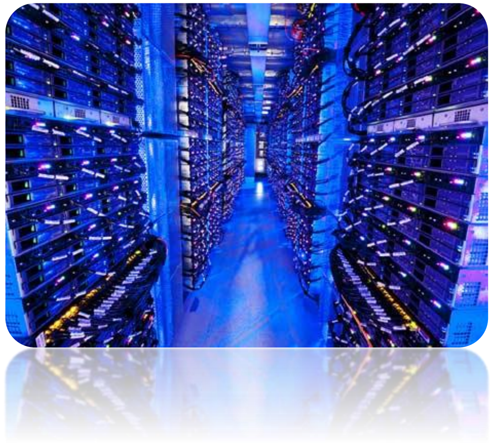
How far are we from rack-scale computing?

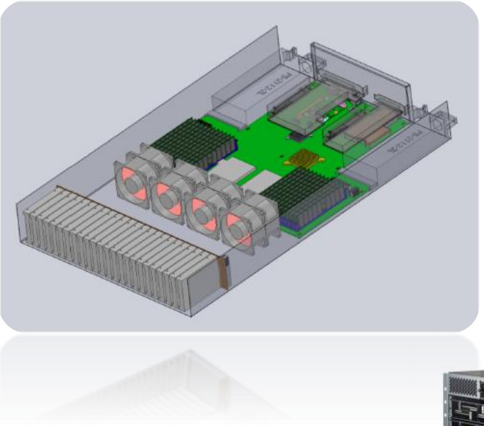
		2020	**2014**
	Today's rack	**Rack-scale Computer**	**Rack-scale Computer**
#Cores (#servers)	~100s (20-40)	~100,000s (1,000s)	*~1,000s* *(100s-1000s)*
Memory	~1 TB	~100s TB	*~10 TBs*
Storage	~100 TB (SSD/HDD)	~100s PB (NVM)	*~1 PB* *(SSD/HDD)*
Network	10 Gbps / server	1 Tbps / server	*~10s -100s Gbps / server*

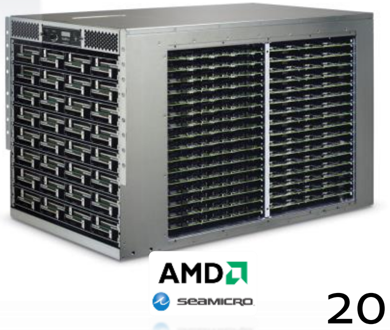
Not just quantity…

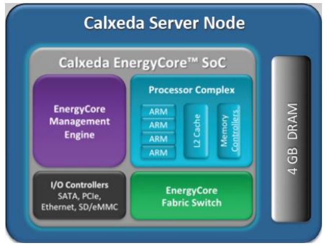
- 3D stacking
  - Cache-like performance for RAM?
- NVRAM
  - Fast byte-addressable storage
- Silicon photonics
  - Low latency (10s-100s ns at rack-scale)

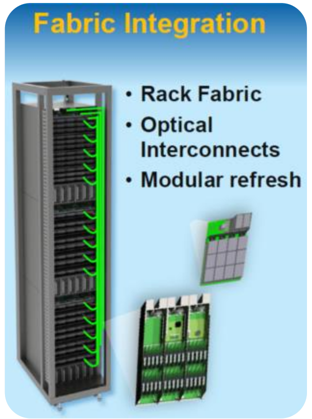
New Hardware, Old Software

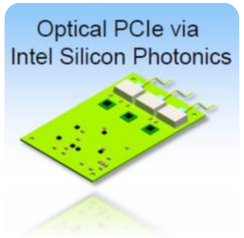






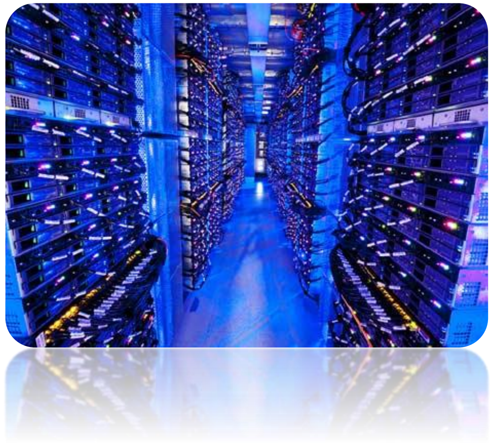


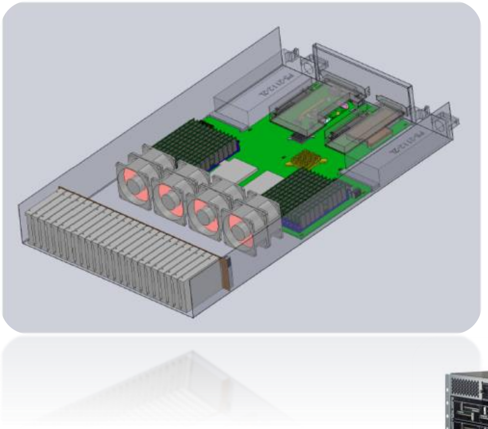


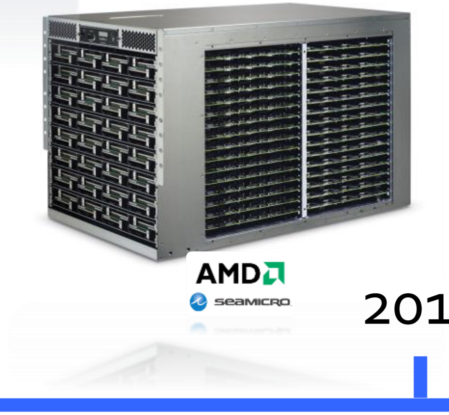


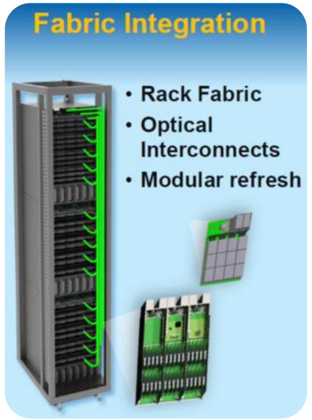
2000 2004 2008 2010 2012 2014

New Hardware, Old Software

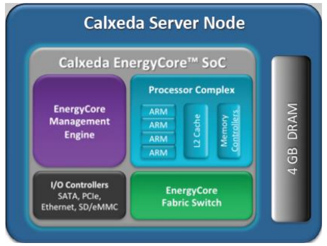


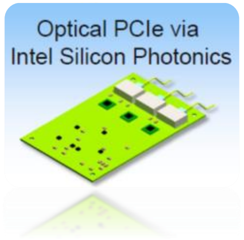




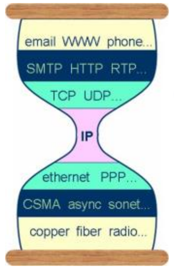

Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular refresh



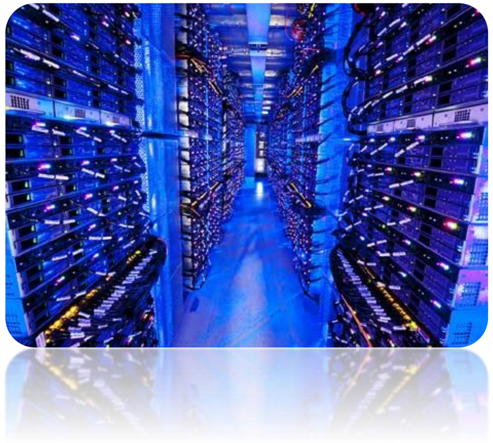

Optical PCIe via Intel Silicon Photonics

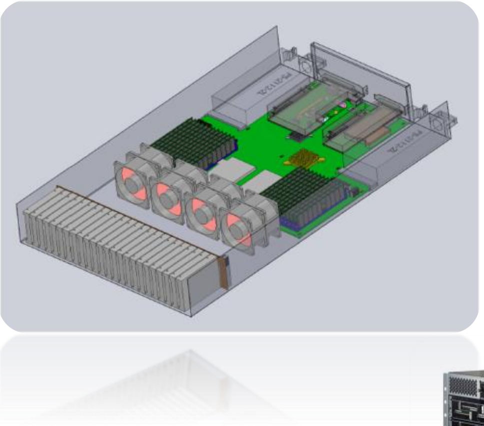
2000 2004 2008 2010 2012 2014

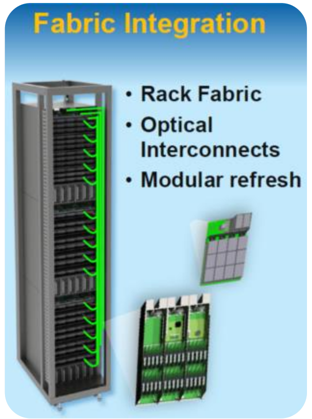


Google
MapReduce

New Hardware, Old Software

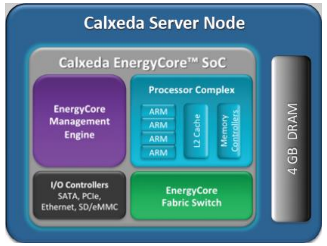


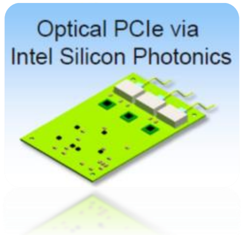




Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular refresh

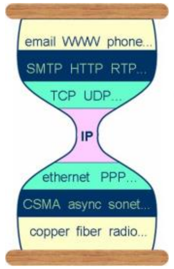




Optical PCIe via Intel Silicon Photonics

2000 2004 2008 2010 2012 2014

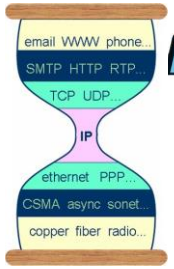
Windows 2000 Datacenter Server



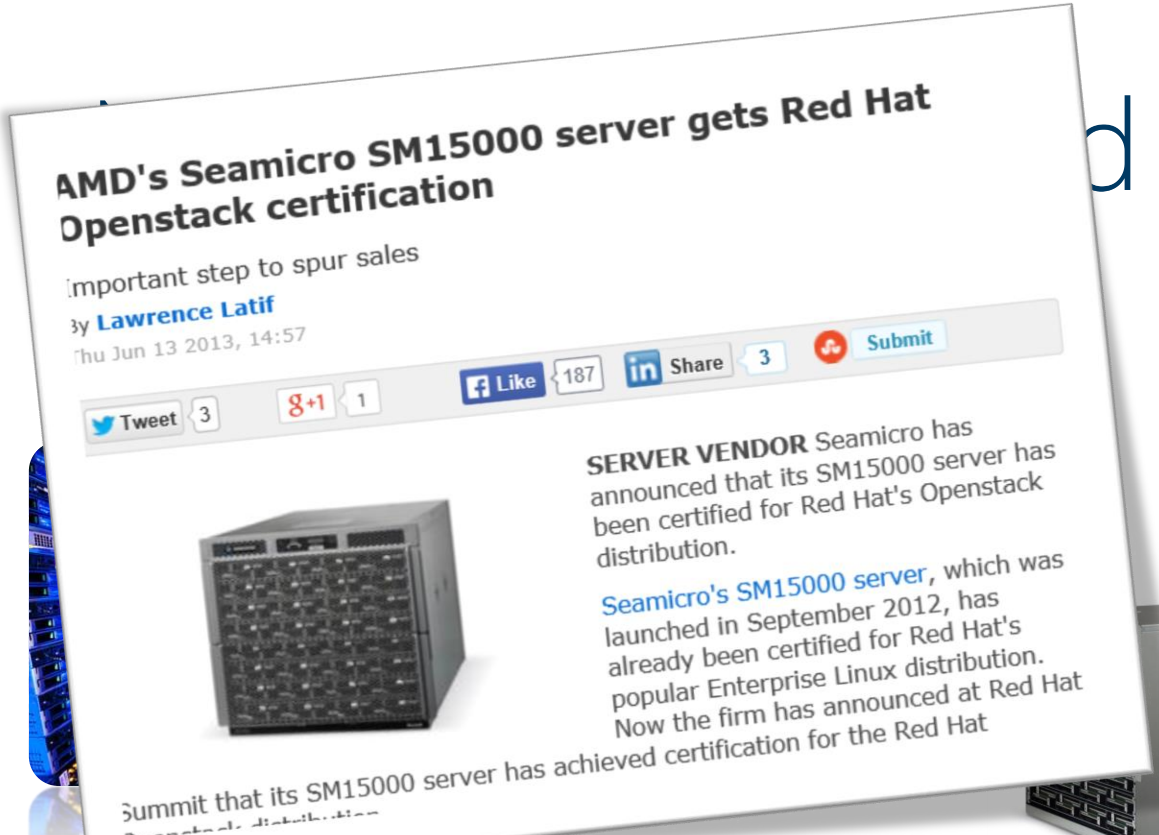
Google MapReduce

Windows Server 2012 Datacenter

Microsoft



hadoop



AMD's Seamicro SM15000 server gets Red Hat Openstack certification

Important step to spur sales

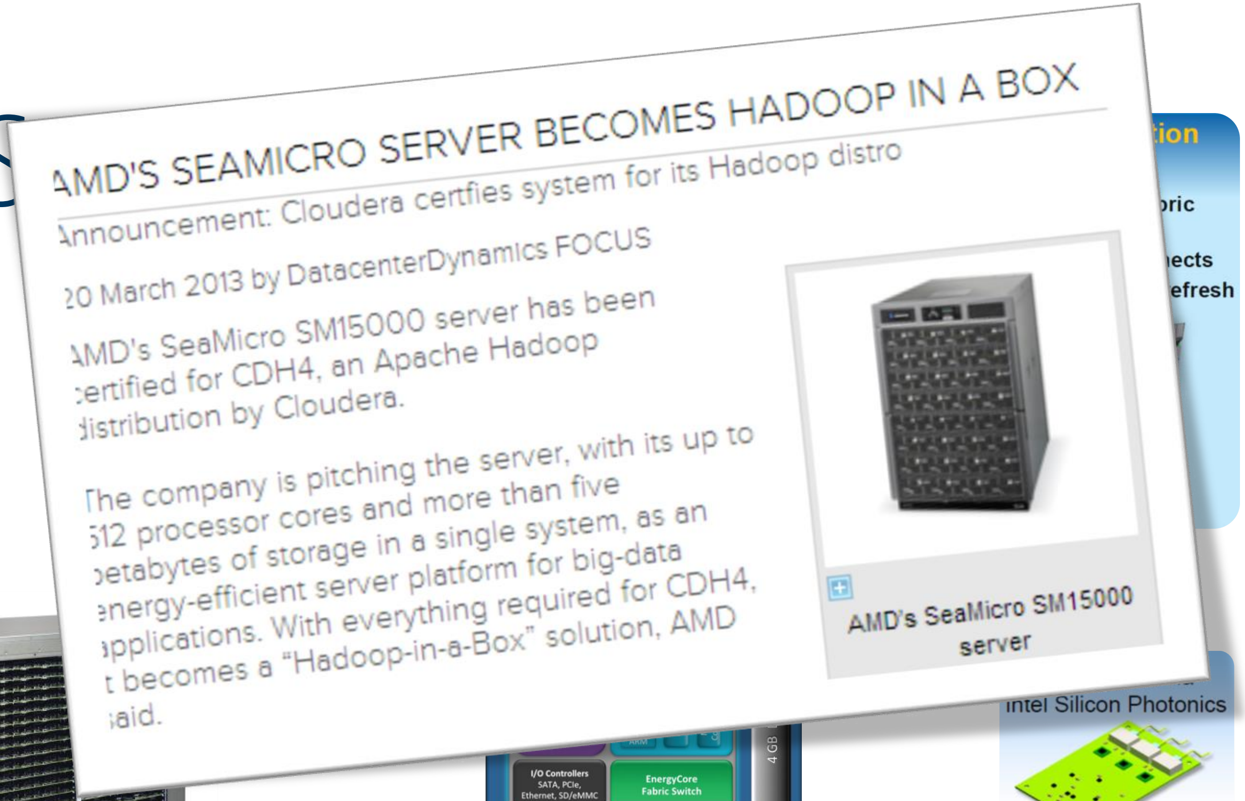
By **Lawrence Latif**

Thu Jun 13 2013, 14:57

Tweet 3 | g+1 1 | Like 187 | in Share 3 | Submit

SERVER VENDOR Seamicro has announced that its SM15000 server has been certified for Red Hat's Openstack distribution.

Seamicro's SM15000 server, which was launched in September 2012, has already been certified for Red Hat's popular Enterprise Linux distribution. Now the firm has announced at Red Hat Summit that its SM15000 server has achieved certification for the Red Hat Openstack distribution



AMD'S SEAMICRO SERVER BECOMES HADOOP IN A BOX

Announcement: Cloudera certfies system for its Hadoop distro

20 March 2013 by DatacenterDynamics FOCUS

AMD's SeaMicro SM15000 server has been certified for CDH4, an Apache Hadoop distribution by Cloudera.

The company is pitching the server, with its up to 512 processor cores and more than five petabytes of storage in a single system, as an energy-efficient server platform for big-data applications. With everything required for CDH4, it becomes a "Hadoop-in-a-Box" solution, AMD said.

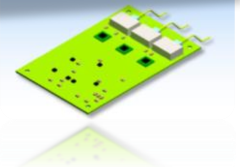
AMD's SeaMicro SM15000 server

...tion

...bric

...ects
...efresh

Intel Silicon Photonics



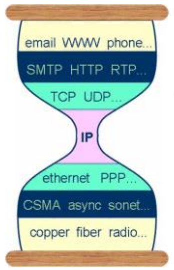
I/O Controllers SATA, PCIe, Ethernet, SD/eMMC

EnergyCore Fabric Switch

4 GB

AMD Seamicro

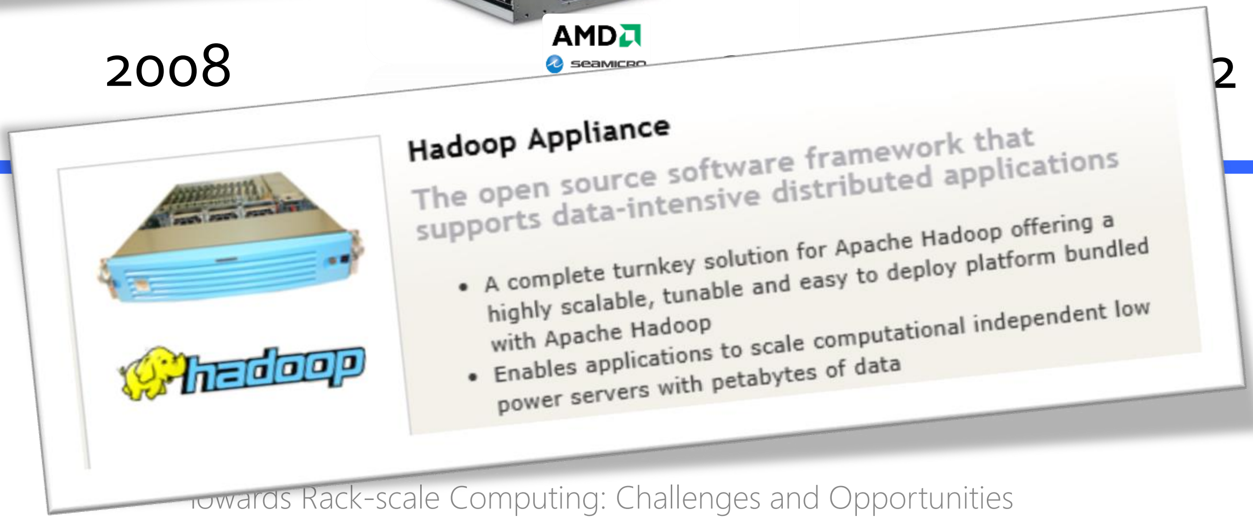
2000 | 2004 | 2008 | ...2 | 2014



Windows 2000 Datacenter Server

Google

MapReduce



Hadoop Appliance

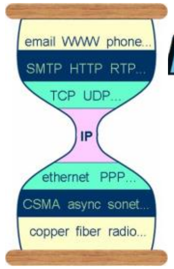
The open source software framework that supports data-intensive distributed applications

- A complete turnkey solution for Apache Hadoop offering a highly scalable, tunable and easy to deploy platform bundled with Apache Hadoop
- Enables applications to scale computational independent low power servers with petabytes of data

hadoop

Windows Server 2012 Datacenter

Microsoft

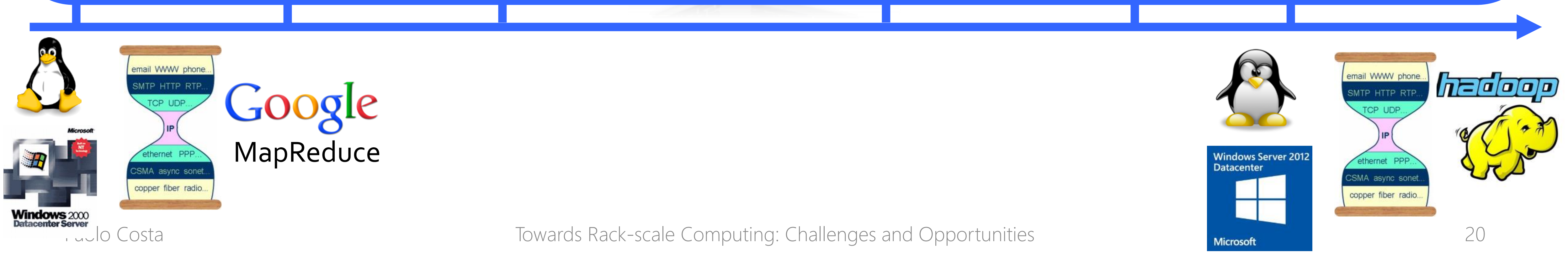


hadoop

New Hardware, Old Software

Achieving many of the benefits of Rack-scale Computers requires adequate software support

Great opportunity for system researchers to rethink the software stack and hw/sw co-design

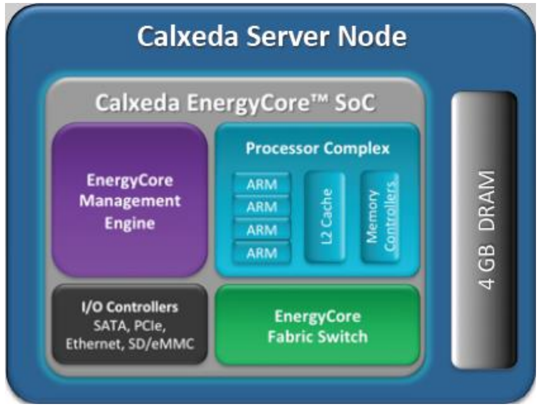


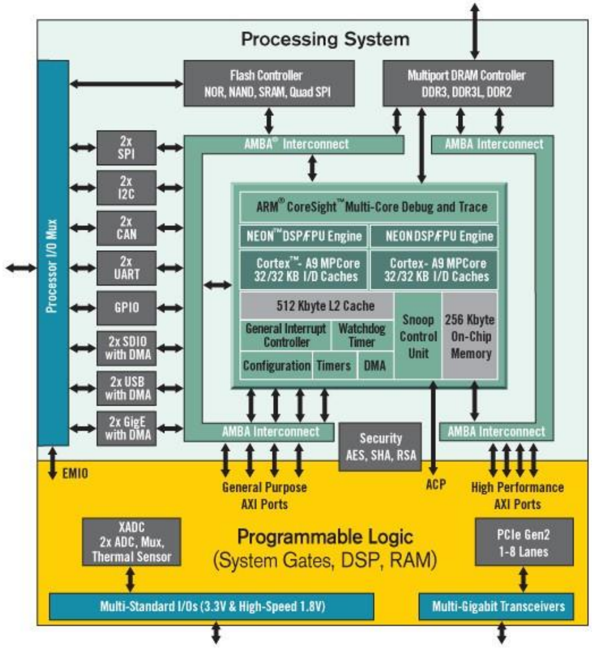
MapReduce

Research Questions: Architecture

What's the best usage of the silicon area?

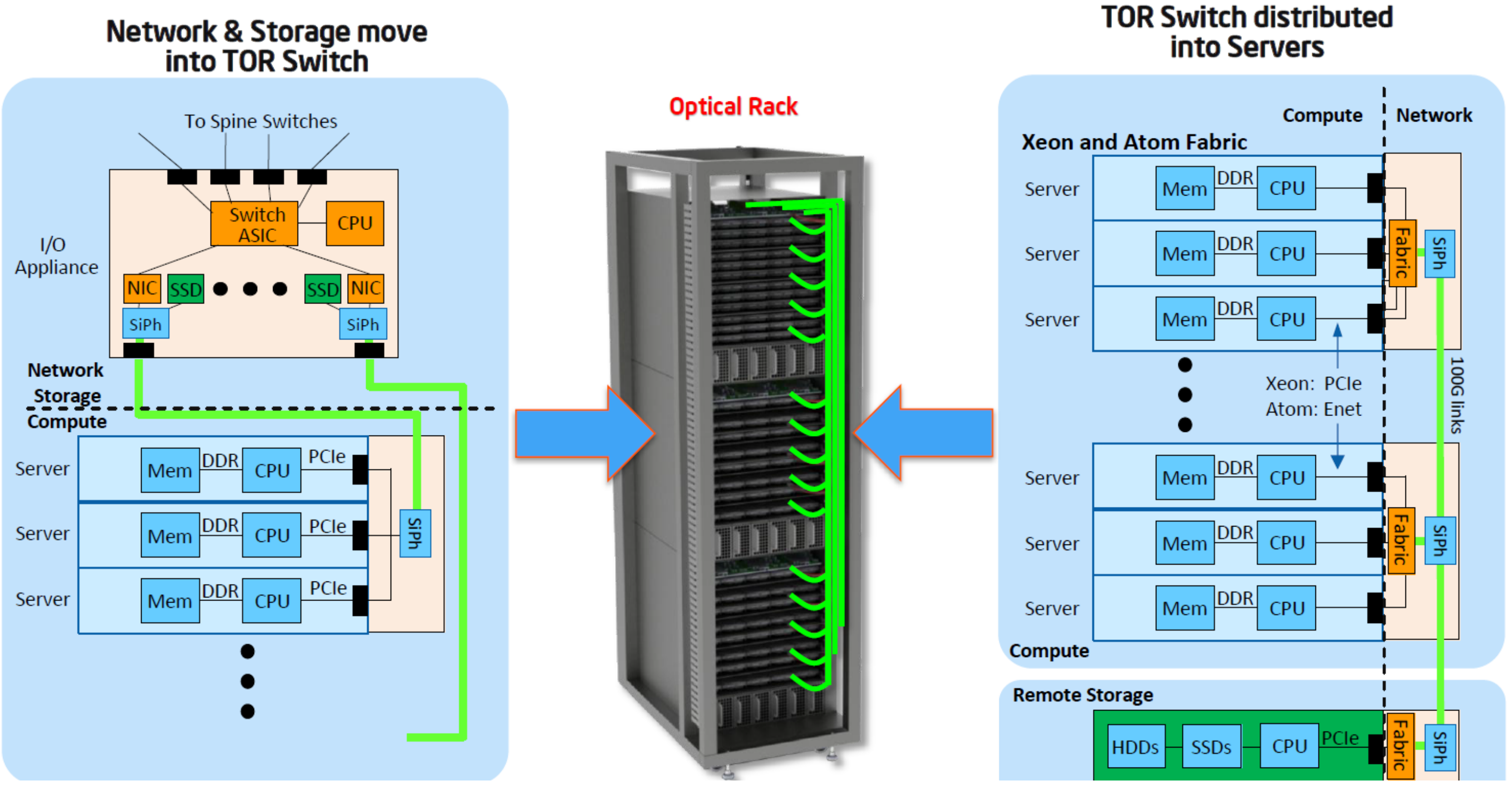
- Homogenous vs. heterogeneous cores
- General-purpose cores vs. accelerators
  - e.g., FPGAs, neural accelerators (NPUs)
- On-chip vs. off-chip functionality




Processing System
Flash Controller NOR, NAND, SRAM, Quad SPI
Multiport DRAM Controller DDR3, DDR3L, DDR2
Processor I/O Mux
2x SPI
2x I2C
2x CAN
2x UART
GPIO
2x SDIO with DMA
2x USB with DMA
2x GigE with DMA
AMBA® Interconnect
AMBA Interconnect
ARM® CoreSight™ Multi-Core Debug and Trace
NEON™ DSP/FPU Engine
NEON DSP/FPU Engine
Cortex™- A9 MPCore 32/32 KB I/D Caches
Cortex- A9 MPCore 32/32 KB I/D Caches
512 Kbyte L2 Cache
General Interrupt Controller
Watchdog Timer
Configuration
Timers
DMA
Snoop Control Unit
256 Kbyte On-Chip Memory
AMBA Interconnect
Security AES, SHA, RSA
AMBA Interconnect
EMIO
General Purpose AXI Ports
ACP
High Performance AXI Ports
XADC 2x ADC, Mux, Thermal Sensor
Programmable Logic (System Gates, DSP, RAM)
PCIe Gen2 1-8 Lanes
Multi-Standard I/Os (3.3V & High-Speed 1.8V)
Multi-Gigabit Transceivers

Research Questions: Networking

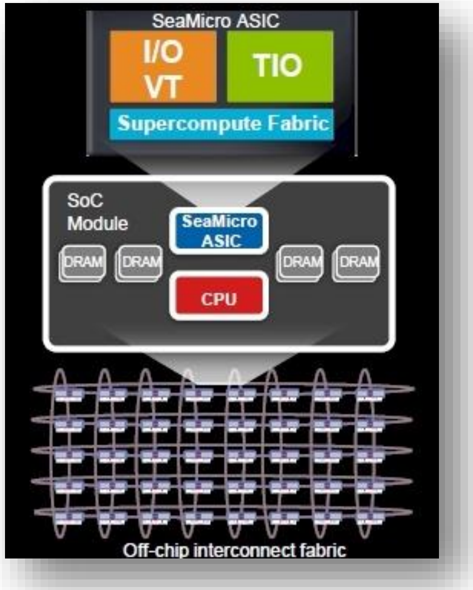
- What is the correct topology?
  - Centralized vs. distributed switch

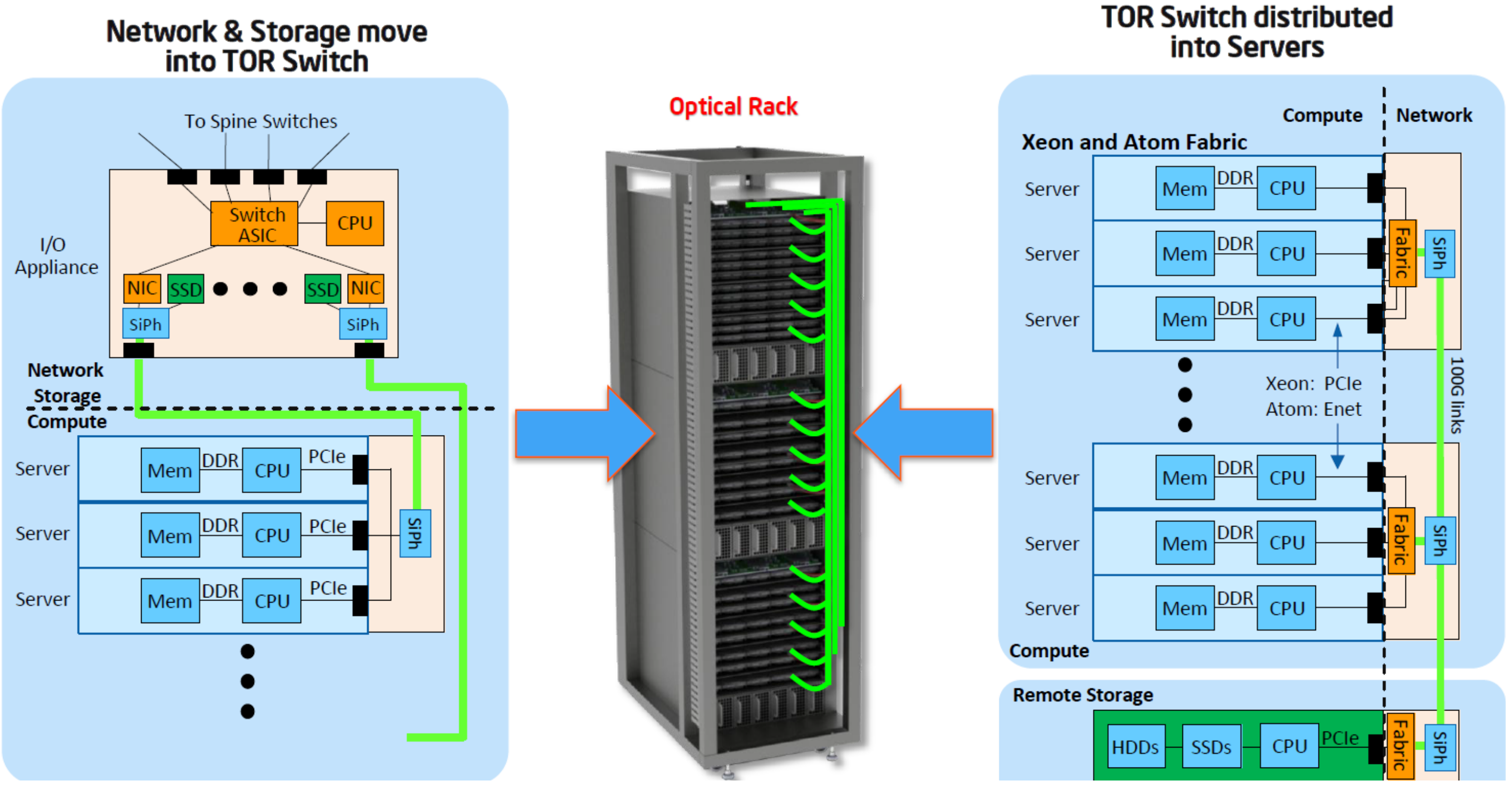


Research Questions: Networking

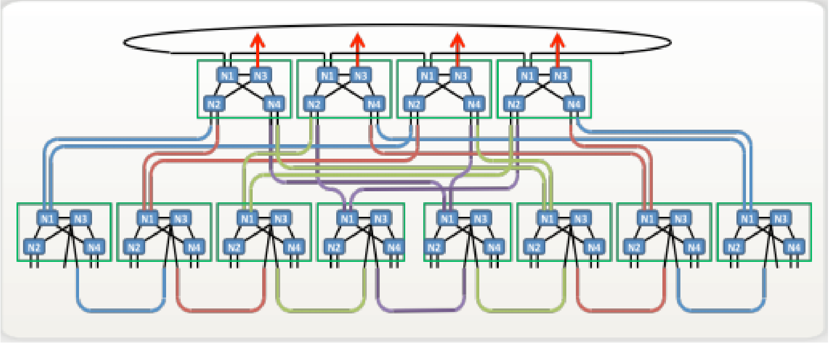
- What is the correct topology?
  - Centralized vs. distributed switch

AMD seamicro





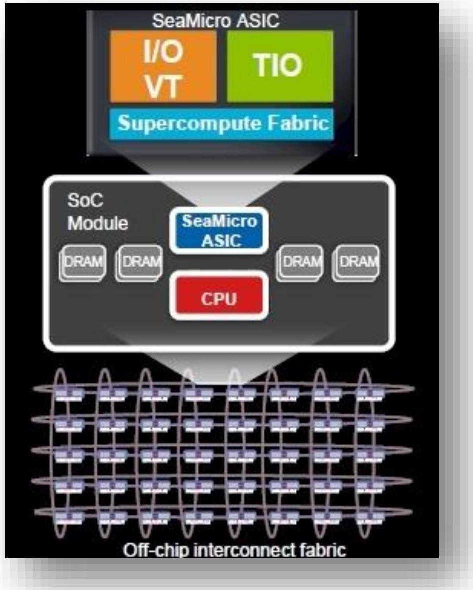
BOSTON
Servers · Storage · Solutions



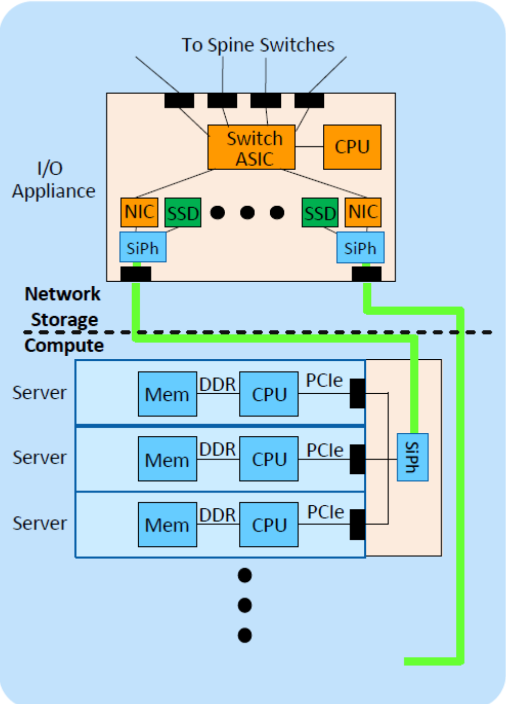
Research Questions: Networking

- What is the correct topology?
  - Centralized vs. distributed switch
  - Application-agnostic vs. application specific

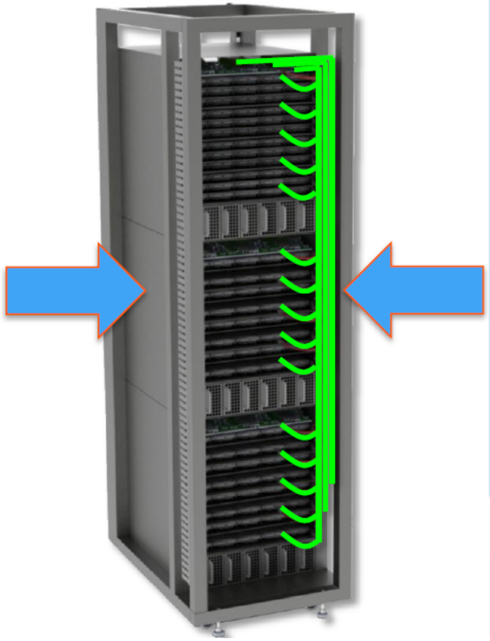
AMD seamicro



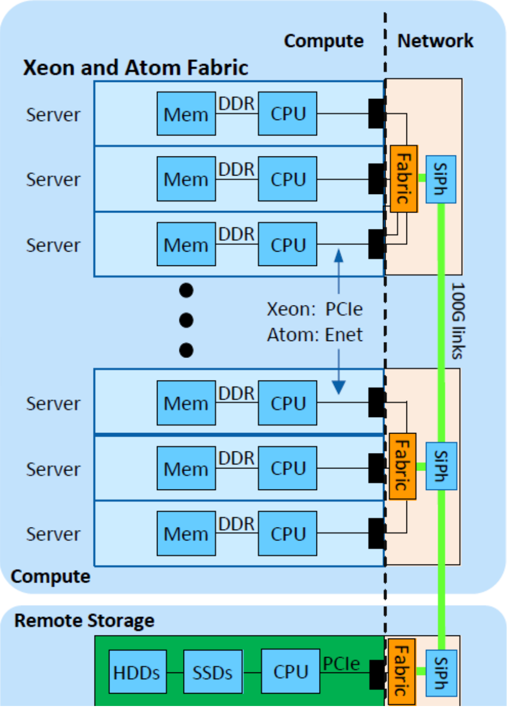
Network & Storage move into TOR Switch

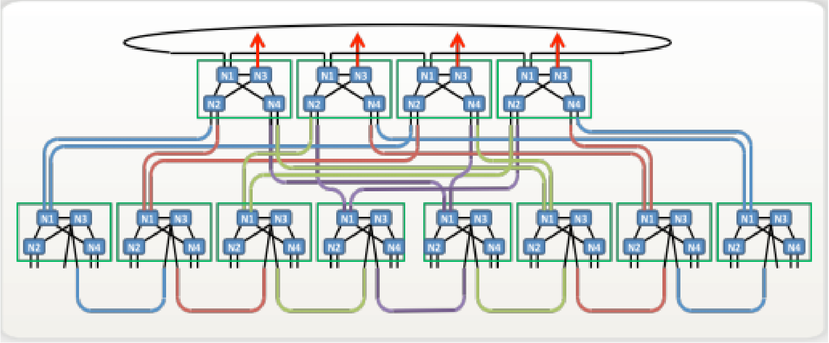


Optical Rack



TOR Switch distributed into Servers

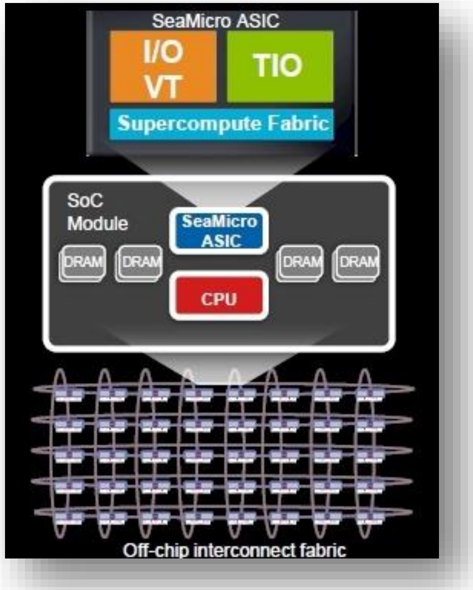




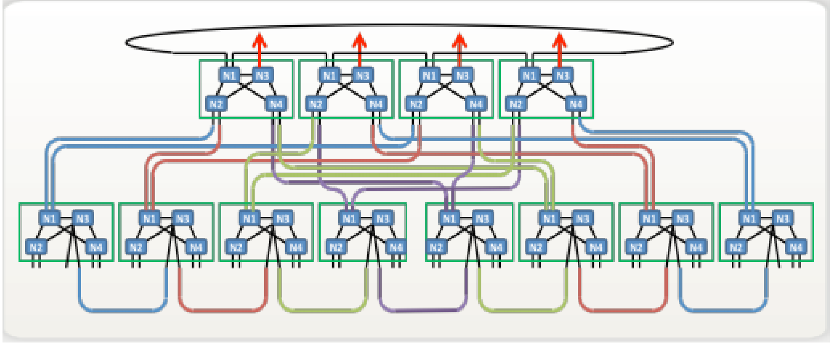
Research Questions: Networking

- What is the correct topology?
  - Centralized vs. distributed switch
  - Application-agnostic vs. application specific
- Where to put memory/storage servers?

AMD
seamicro

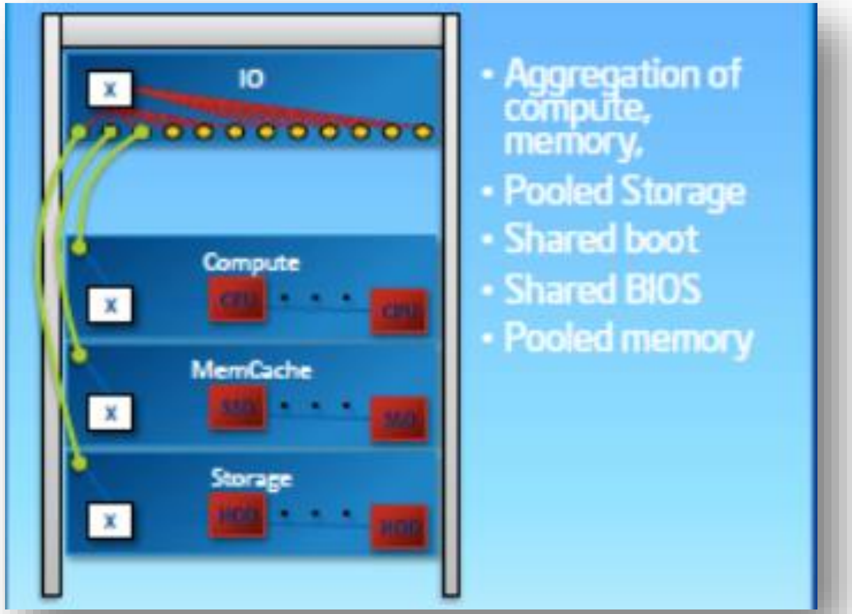


BOSTON
Servers · Storage · Solutions



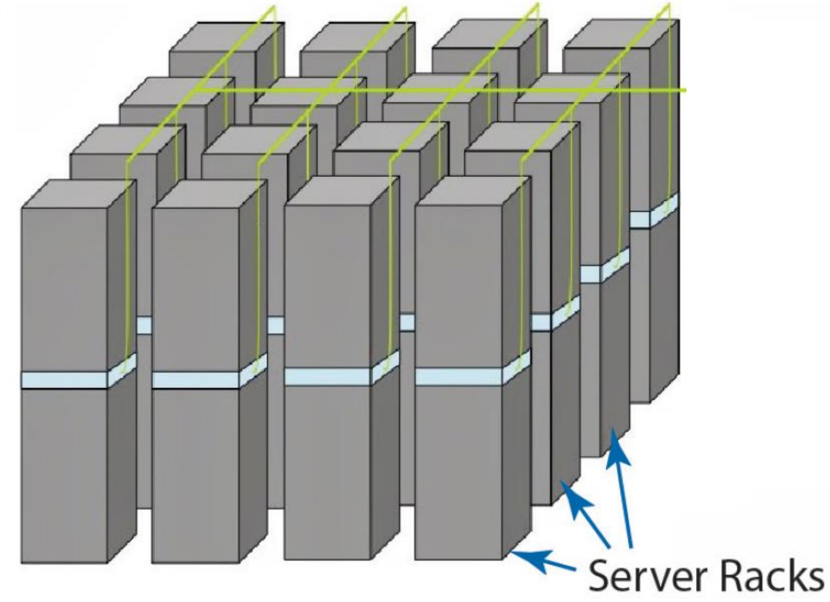
Research Questions: Networking

- What is the correct topology?
  - Centralized vs. distributed switch
  - Application-agnostic vs. application specific
- Where to put memory/storage servers?
- Converged fabric
  - How to handle memory, storage, IP traffic?



Research Questions: Networking

- What is the correct topology?
  - Centralized vs. distributed switch
  - Application-agnostic vs. application specific
- Where to put memory/storage servers?
- Converged fabric
  - How to handle memory, storage, IP traffic?
- Inter-rack connectivity
  - How to extend beyond rack-scale?
    - over-subscription and protocol bridging



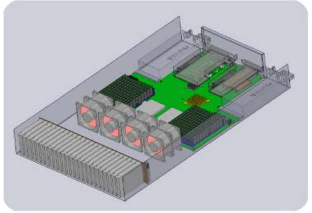
Research Questions: OS / Storage

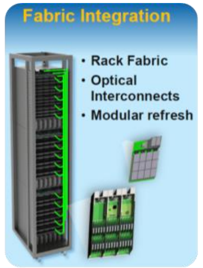
- Rethink the cache-hierarchy
  - High-performance (3D stacking) vs. high-capacity tier (NVRAM)
  - What's the correct ratio?
  - Are SSDs / HDDs to be used only for cold data?
- Impact on existing (and new!) applications?
  - Cache-like RAM and byte-addressable fast storage
- How to schedule application tasks?
  - Joint scheduling (CPU, memory, network, storage)

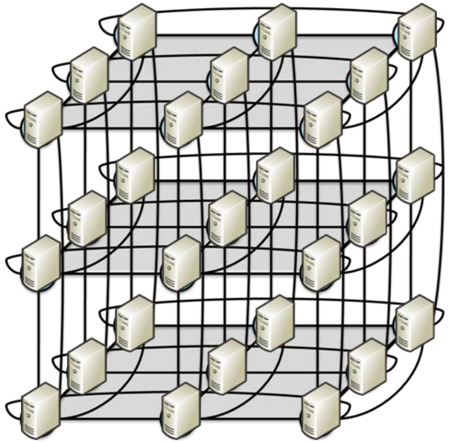
Research Questions: Distributed Systems

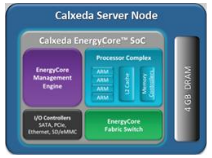
- RaSCs are different from many-core setups
  - Separate failure domains, no cache coherency
  - Rack-scale computers are distributed systems (albeit not traditional)
- How to handle remote resources?
  - Consistency and fault-tolerance
- What are the right programming abstractions?
  - Shared memory, message passing, MapReduce, …

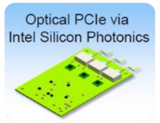
Rack-scale Computing @ MSR Cambridge

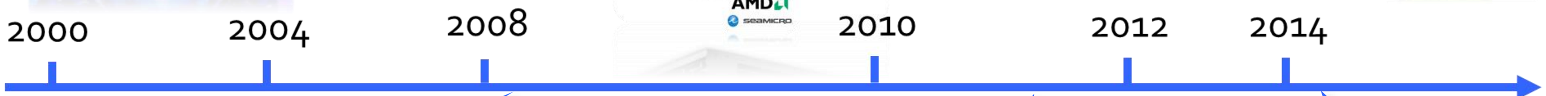


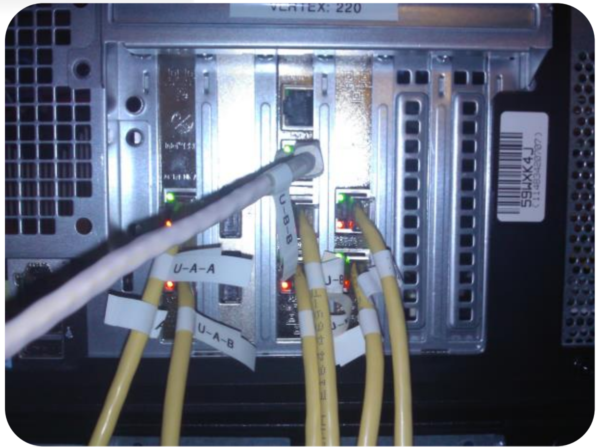












CamCube Vision [WREN'09]

Custom Routing [SIGCOMM'10]

In-network Aggregation [NSDI'12]

Network Abstractions [HPDC'13]

Rack-scale Computing @ MSR Cambridge

Programming abstractions

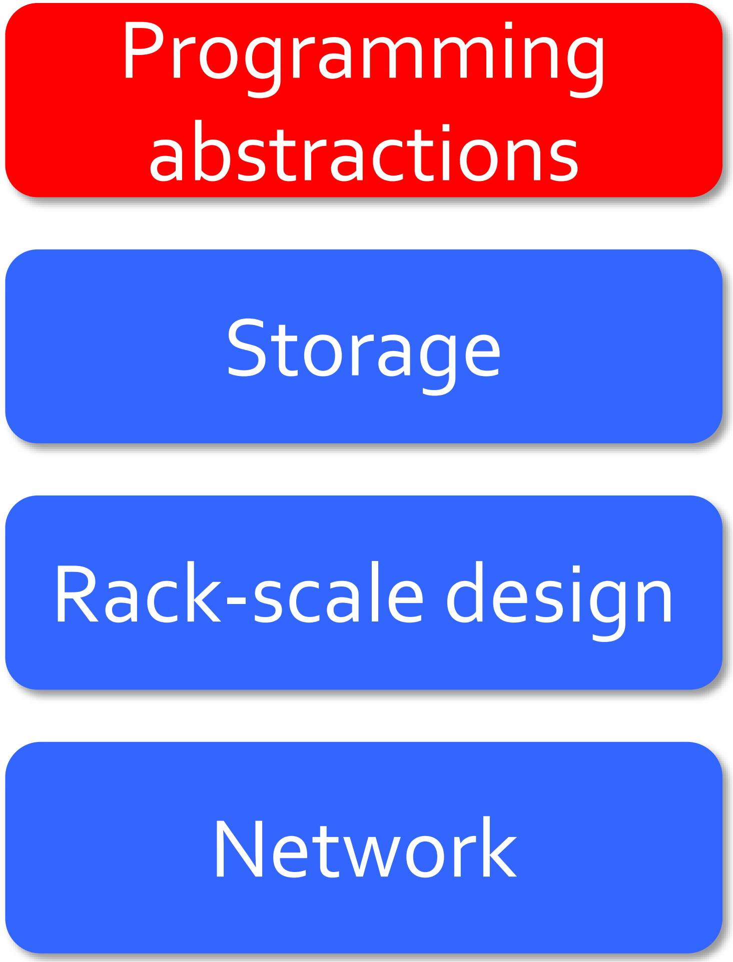
Storage

Rack-scale design

Network

http://research.microsoft.com/rackscale/

Rack-scale Computing @ MSR Cambridge


Programming abstractions

Storage

Rack-scale design

Network

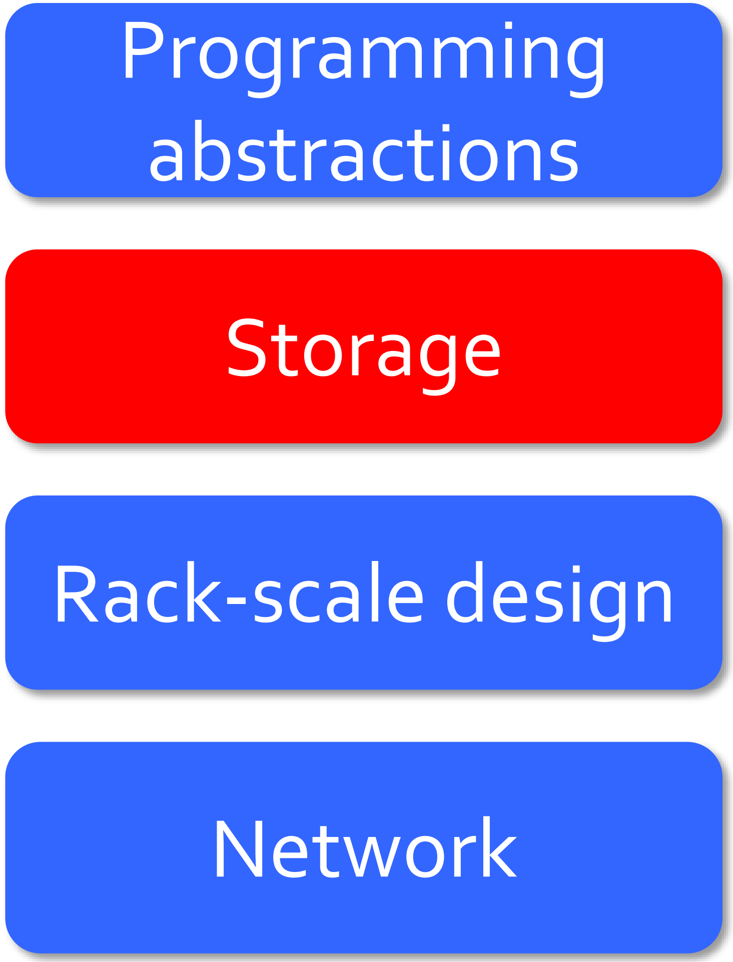
http://research.microsoft.com/rackscale/

FaRM [NSDI'14]

- RDMA-based distributed platform
  - Transaction support
  - Lock-free reads
  - Support for object colocation
- Hardware alone is not enough
  - Software stack customization is needed
- High performance
  - 167 M key lookups (31 us latency) on a 20-server testbed

Aleksandar Dragojević, Dushyanth Narayanan,
Orion Hodson, Miguel Castro

Rack-scale Computing @ MSR Cambridge


Programming abstractions

Storage

Rack-scale design

Network

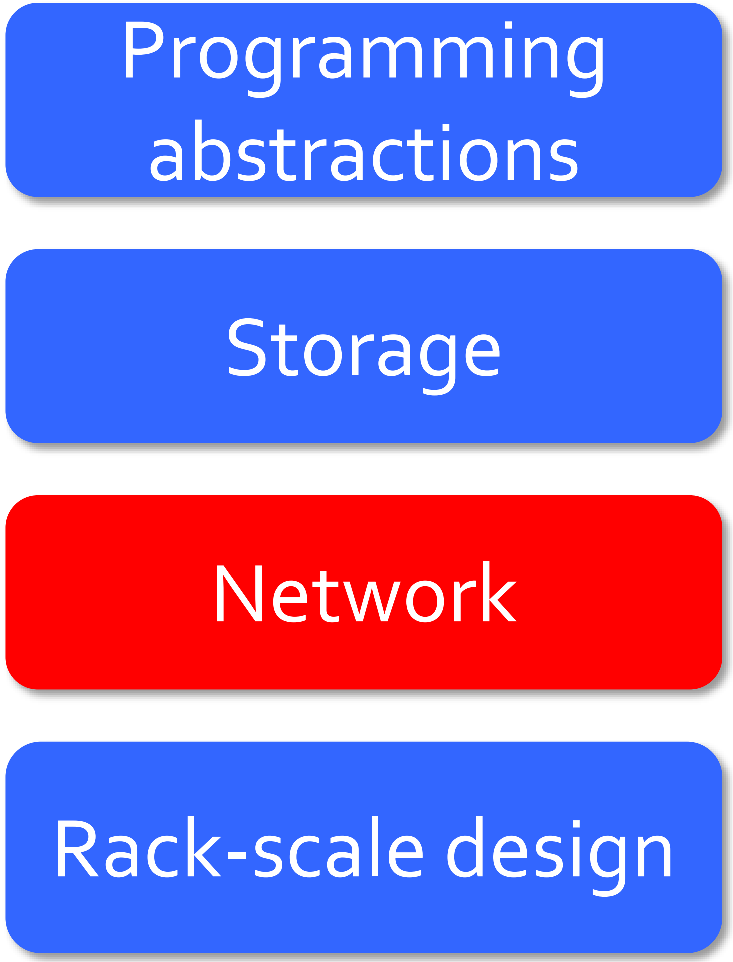
http://research.microsoft.com/rackscale/

Pelican

- Rack-scale storage appliance for “cold” data
- Hardware and software co-design
  - High storage density
  - Low cost
  - Low power consumption
  - Fault tolerant

Austin Donnelly, Richard Black, Sergey Legtchenko, Ant Rowstron, Dave Harper, Shobana Balakrishnan, Eric Peterson, Adam Glass

Rack-scale Computing @ MSR Cambridge



Programming abstractions

Storage

Network

Rack-scale design

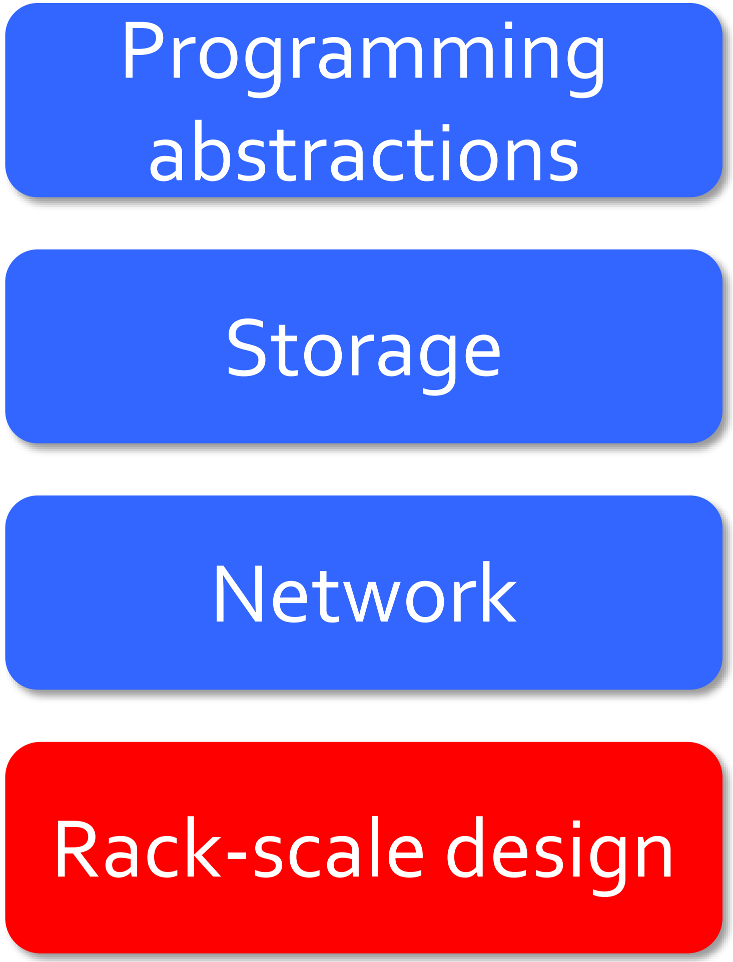
RaSC-Net [HotCloud'14]

- How to design a network stack for Rack-scale computers?
  - Routing and congestion control
- Support for:
  - Multiple paths
  - Low latency
  - Consolidated workloads

http://research.microsoft.com/rackscale/

Paolo Costa, Hitesh Ballani, Dushyanth Narayan

Rack-scale Computing @ MSR Cambridge



Programming abstractions

Storage

Network

Rack-scale design

http://research.microsoft.com/rackscale/

DRackAr

- How to master the design space?
  - Topology, resources provisioning, …
- Input:
  - Hardware components
  - Constraints (e.g., max power budget)
  - Target workload
  - Utility function
- Output: Rack configuration

Sergey Legtchenko, Ant Rowstron

Summary

- Rack-scale computing:
  - 1,000s of cores
  - TBs of RAM and PBs of storage
  - Intra-rack high bandwidth / low latency connectivity
- This can improve the performance of existing apps…
  - graph processing, machine learning jobs, in-memory DBs, …
  - …but also enable new ones!
- Call to action
  - Hardware has been changing a lot…
  - …now it's up to us to change the software too!