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

Goal

*Increase work done per dollar (CapEx + OpEx)*
Hardware Evolution in Data Centers

Scale out vs. scale up
Many commodity servers rather than few expensive servers
Hardware Evolution in Data Centers

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

System-on-Chip (SoC)
CPU, IO controllers, NIC/fabric switch on the same die
Silicon Photonics
High-bandwidth / low-latency interconnect (resource disaggregation)
Hardware Evolution in Data Centers

- 2000: 1 rack unit (RU)
- 2004: 2 RU
- 2008: 4-10 RU
- 2010: Rack-scale
- 2012: Calxeda Server Node
- 2014: Fabric Integration
  - Rack Fabric
  - Optical Interconnects
  - Modular refresh

Towards Rack-scale Computing: Challenges and Opportunities
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
Rack-scale Computer in 2020?

<table>
<thead>
<tr>
<th></th>
<th>Today’s traditional rack</th>
<th>2020 Rack-scale Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Cores (#servers)</td>
<td>~100s (20-40)</td>
<td>~100,000s (1,000s)</td>
</tr>
<tr>
<td>Memory</td>
<td>~1 TB</td>
<td>~100s TB</td>
</tr>
<tr>
<td>Storage</td>
<td>~100 TB (flash + spinning disk)</td>
<td>~100s PB (NVM)</td>
</tr>
<tr>
<td>Bandwidth / server</td>
<td>10 Gbps</td>
<td>1 Tbps</td>
</tr>
</tbody>
</table>
How far are we from rack-scale computing?

<table>
<thead>
<tr>
<th></th>
<th>Today’s rack</th>
<th>2020 Rack-scale Computer</th>
<th>2014 Rack-scale Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Cores (#servers)</td>
<td>~100s (20-40)</td>
<td>~100,000s (1,000s)</td>
<td></td>
</tr>
<tr>
<td>Memory</td>
<td>~1 TB</td>
<td>~100s TB</td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td>~100 TB (SSD/HDD)</td>
<td>~100s PB (NVM)</td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>10 Gbps / server</td>
<td>1 Tbps / server</td>
<td></td>
</tr>
</tbody>
</table>
How far are we from rack-scale computing?

<table>
<thead>
<tr>
<th>Core count</th>
<th>AMD SeaMicro SM15000-64</th>
<th>HP Moonshot Redstone</th>
<th>Boston Viridis</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 512 cores (64 servers) in 10 RU</td>
<td>• 2,048 cores (256 servers) at rack scale</td>
<td>• 1,152 cores (288 servers) in 4U</td>
<td>• 192 cores (48 servers) in 2 RU</td>
</tr>
<tr>
<td>• 2,048 cores (256 servers) at rack scale</td>
<td></td>
<td>• 11,520 cores (3,200 servers) at rack scale</td>
<td>• 7,680 cores (1,920 servers) at rack scale</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Today’s rack</th>
<th>2020 Rack-scale Computer</th>
<th>2014 Rack-scale Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Cores (#servers)</td>
<td>~100s (20-40)</td>
<td>~100,000s (1,000s)</td>
</tr>
<tr>
<td>Memory</td>
<td>~1 TB</td>
<td>~100s TB</td>
</tr>
<tr>
<td>Storage</td>
<td>~100 TB (SSD/HDD)</td>
<td>~100s PB (NVM)</td>
</tr>
<tr>
<td>Network</td>
<td>10 Gbps / server</td>
<td>1 Tbps / server</td>
</tr>
</tbody>
</table>
## How far are we from rack-scale computing?

<table>
<thead>
<tr>
<th>Current</th>
<th>2020 Rack-scale Computer</th>
<th>2014 Rack-scale Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>#Cores (#servers)</strong></td>
<td>~100s (20-40)</td>
<td>~100,000s (1,000s)</td>
</tr>
<tr>
<td>Memory</td>
<td>~1 TB</td>
<td>~100s TB</td>
</tr>
<tr>
<td>Storage</td>
<td>~100 TB (SSD/HDD)</td>
<td>~100s PB (NVM)</td>
</tr>
<tr>
<td>Network</td>
<td>10 Gbps / server</td>
<td>1 Tbps / server</td>
</tr>
</tbody>
</table>

### 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?

<table>
<thead>
<tr>
<th></th>
<th>Today’s rack</th>
<th>2020 Rack-scale Computer</th>
<th>2014 Rack-scale Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Cores (#servers)</td>
<td>~100s (20-40)</td>
<td>~100,000s (1,000s)</td>
<td>~1,000s (100s-1000s)</td>
</tr>
<tr>
<td>Memory</td>
<td>~1 TB</td>
<td>~100s TB</td>
<td>~10 TBs</td>
</tr>
<tr>
<td>Storage</td>
<td>~100 TB (SSD/HDD)</td>
<td>~100s PB (NVM)</td>
<td>~1 PB (SSD/HDD)</td>
</tr>
<tr>
<td>Network</td>
<td>10 Gbps / server</td>
<td>1 Tbps / server</td>
<td>~10s -100s Gbps / server</td>
</tr>
</tbody>
</table>

Network

- AMD SeaMicro SM15000-XE
  - 1.28 Tbps fabric (20 Gbps / server)
- Mellanox ConnectX-3 Pro
  - 2x 40-Gbps NICs
- Intel M XC Connector (expected Q3’14)
  - Up to 32 fibers (25 Gbps / fiber)
  - Up to 800 Gbps / server
How far are we from rack-scale computing?

<table>
<thead>
<tr>
<th></th>
<th>Today’s rack</th>
<th>2020 Rack-scale Computer</th>
<th>2014 Rack-scale Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Cores (#servers)</td>
<td>~100s (20-40)</td>
<td>~100,000s (1,000s)</td>
<td>~1,000s (100s-1000s)</td>
</tr>
<tr>
<td>Memory</td>
<td>~1 TB</td>
<td>~100s TB</td>
<td>~10 TBs</td>
</tr>
<tr>
<td>Storage</td>
<td>~100 TB (SSD/HDD)</td>
<td>~100s PB (NVM)</td>
<td>~1 PB (SSD/HDD)</td>
</tr>
<tr>
<td>Network</td>
<td>10 Gbps / server</td>
<td>1 Tbps / server</td>
<td>~10s ~100s Gbps / server</td>
</tr>
</tbody>
</table>

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)

Paolo Costa
Towards Rack-scale Computing: Challenges and Opportunities
New Hardware, Old Software

Towards Rack-scale Computing: Challenges and Opportunities

New Hardware, Old Software

Towards Rack-scale Computing: Challenges and Opportunities

MapReduce
New Hardware, Old Software
AMD's Seamicro SM15000 server gets Red Hat Openstack certification

Important step to spur sales
by Lawrence Latif
Thu Jun 13 2013, 14:57

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 SM15000 server has been certified for CDH4, an Apache Hadoop distribution by Cloudera.

The company is pitching the server, with its up to 312 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.

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
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
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
Research Questions: Networking

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

- 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: Networking

• What is the correct topology?
  – Centralized vs. distributed switch
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

Paolo Costa
Towards Rack-scale Computing: Challenges and Opportunities
Research Questions: Networking

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

• Where to put memory/storage servers?
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

- Programming abstractions
- Storage
- Rack-scale design
- Network

http://research.microsoft.com/rackscale/
Rack-scale Computing @ MSR Cambridge

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

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

Aleksandar Dragojević, Dushyanth Narayanan, Orion Hodson, Miguel Castro
Rack-scale Computing @ MSR Cambridge

- Programming abstractions
- Storage
- Rack-scale design
- Network

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

http://research.microsoft.com/rackscale/
Rack-scale Computing @ MSR Cambridge

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

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

Programming abstractions
Storage
Network

Rack-scale design

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

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!