Data Center Switch Architecture in the Age of Merchant Silicon

Nathan Farrington
Erik Rubow
Amin Vahdat
The Network is a Bottleneck

• HTTP request amplification
  – Web search (e.g. Google)
  – Small object retrieval (e.g. Facebook)
  – Web services (e.g. Amazon.com)

• MapReduce-style parallel computation
  – Inverted search index
  – Data analytics

• Need high-performance interconnects
The Network is Expensive

8xGbE

\[ \text{Rack 1} \quad \text{Rack 2} \quad \text{Rack 3} \quad \ldots \quad \text{Rack N} \]

\[ \ldots 48xGbE \text{ TOR Switch} \ldots \]

\[ \ldots 40x1U \text{ Servers} \ldots \]
What we really need: One Big Switch

- Commodity
- Plug-and-play
- Potentially no oversubscription
Why not just use a fat tree of commodity TOR switches?

10 Tons of Cable

- 55,296 Cat-6 cables
- 1,128 separate cable bundles

The “Yellow Wall”
Merchant Silicon gives us Commodity Switches

<table>
<thead>
<tr>
<th>Maker</th>
<th>Broadcom</th>
<th>Fulcrum</th>
<th>Fujitsu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>BCM56820</td>
<td>FM4224</td>
<td>MB86C69RBC</td>
</tr>
<tr>
<td>Ports</td>
<td>24</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Cost</td>
<td>NDA</td>
<td>NDA</td>
<td>$410</td>
</tr>
<tr>
<td>Power</td>
<td>NDA</td>
<td>20 W</td>
<td>22 W</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt; 1 μs</td>
<td>300 ns</td>
<td>300 ns</td>
</tr>
<tr>
<td>Area</td>
<td>NDA</td>
<td>40 x 40 mm</td>
<td>35 x 35 mm</td>
</tr>
<tr>
<td>SRAM</td>
<td>NDA</td>
<td>2 MB</td>
<td>2.9 MB</td>
</tr>
<tr>
<td>Process</td>
<td>65 nm</td>
<td>130 nm</td>
<td>90 nm</td>
</tr>
</tbody>
</table>
Eliminate Redundancy

• Networks of packet switches contain many redundant components
  – chassis, power conditioning circuits, cooling
  – CPUs, DRAM
• Repackage these discrete switches to lower the cost and power consumption
Our Architecture, in a Nutshell

• Fat tree of merchant silicon switch ASICs
• Hiding cabling complexity with PCB traces and optics
• Partition into multiple pod switches + single core switch array
• Custom EEP ASIC to further reduce cost and power
• Scales to 65,536 ports when 64-port ASICs become available, late 2009
3 Different Designs

- 24-ary 3-tree
- 720 switch ASICs
- 3,456 ports of 10GbE
- No oversubscription
Network 1: No Engineering Required

- 720 discrete packet switches, connected with optical fiber

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost of Parts</td>
<td>$4.88M</td>
</tr>
<tr>
<td>Power</td>
<td>52.7 kW</td>
</tr>
<tr>
<td>Cabling Complexity</td>
<td>3,456</td>
</tr>
<tr>
<td>Footprint</td>
<td>720 RU</td>
</tr>
<tr>
<td>NRE</td>
<td>$0</td>
</tr>
</tbody>
</table>

*Cabling complexity (noun): the number of long cables in a data center network.*
Network 2: Custom Boards and Chassis

- 24 “pod” switches, one core switch array, 96 cables

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost of Parts</strong></td>
<td>$3.07M</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>41.0 kW</td>
</tr>
<tr>
<td><strong>Cabling Complexity</strong></td>
<td>96</td>
</tr>
<tr>
<td><strong>Footprint</strong></td>
<td>192 RU</td>
</tr>
<tr>
<td><strong>NRE</strong></td>
<td>$3M est</td>
</tr>
</tbody>
</table>

This design is shown in more detail later.
Switch at 10G, but Transmit at 40G

<table>
<thead>
<tr>
<th></th>
<th>SFP</th>
<th>SFP+</th>
<th>QSFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>1 Gb/s</td>
<td>10 Gb/s</td>
<td>40 Gb/s</td>
</tr>
<tr>
<td>Cost/Gb/s</td>
<td>$35*</td>
<td>$25*</td>
<td>$15*</td>
</tr>
<tr>
<td>Power/Gb/s</td>
<td>500mW</td>
<td>150mW</td>
<td>60mW</td>
</tr>
</tbody>
</table>

* 2008-2009 Prices
Network 3: Network 2 + Custom ASIC

- Uses 40GbE between pod switches and core switch array; everything else is same as Network 2.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost of Parts</td>
<td>$2.33M</td>
</tr>
<tr>
<td>Power</td>
<td>36.4 kW</td>
</tr>
<tr>
<td>Cabling Complexity</td>
<td>96</td>
</tr>
<tr>
<td>Footprint</td>
<td>114 RU</td>
</tr>
<tr>
<td>NRE</td>
<td>$8M est</td>
</tr>
</tbody>
</table>

EEP

This simple ASIC provides tremendous cost and power savings.
Cabling Complexity

Network 1: 3,456
Network 2: 96
Network 3: 96
Footprint

Footprint (in rack units)

Network 1: 720
Network 2: 192
Network 3: 114
Partially Deployed Switch

Cable tray

72-fiber cable
(4 per pod switch)

Pod switch

10GbE cable

144 servers

Core switch array

Nathan Farrington
farrington@cs.ucsd.edu
Fully Deployed Switch

- 72-fiber cable (4 per pod switch)
- Pod switch
- 10GbE cable
- 24 pods; 144 servers each
- Core switch array
Pod Switch

144 Ports 10GbE Towards Compute Nodes

Line Card

Midplane

36 Ports 40GbE Towards Core Switch Array

Uplink Card
Logical Topology

Pod Switch

Line Cards

Midplane

Uplink Cards

Core Switch Array Cards

Core Switch Array

Optical Fiber

Copper Traces
Pod Switch Line Card

To Uplink Card #1
To Uplink Card #2
To Uplink Card #3
To Uplink Card #4
To Uplink Card #5
To Uplink Card #6

CPU
Switch ASIC (x4)

PHY (x48) (XAU1/10GBASE-KR)

1-8 SFP+ Cage (x8)
9-16
17-24
25-32
33-40
41-48

400 mm
432 mm
Pod Switch Uplink Card

To Core Cards
#1-3 or 4-6 or 7-9

QSFP Cage
(x6)

EEP ASIC
(x6)

Switch ASIC
(x2)

PHY (x24)
(XAUI/10GBASE-KR)

To Line Card #1
To Line Card #2
To Line Card #3

245 mm

150 mm
Why an Ethernet Extension Protocol?

• Optical transceivers are 80% of the cost
• EEP allows the use of fewer and faster optical transceivers
How does EEP work?

• Ethernet frames are split up into EEP frames
• Most EEP frames are 65 bytes
  – Header is 1 byte; payload is 64 bytes
• Header encodes ingress/egress port
How does EEP work?

- Round-robin arbiter
- EEP frames are transmitted as one large Ethernet frame
- 40GbE overclocked by 1.6%
Ethernet Frames
EEP Frame Format

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1B</td>
<td>SOF</td>
<td>EOF</td>
<td>LEN</td>
<td>Virtual Link ID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td></td>
<td></td>
<td>Payload Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65B</td>
<td></td>
<td></td>
<td>Payload</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SOF: Start of Ethernet Frame
EOF: End of Ethernet Frame
LEN: Set if EEP Frame contains less than 64B of payload
Virtual Link ID: Corresponds to port number (0-15)
Payload Length: (0-63B)
Why not use VLANs?

• Because it adds latency and requires more SRAM

• FPGA Implementation
  – VLAN tagging
  – EEP
Latency Measurements

![Graph showing round-trip latency vs. Ethernet frame size (bytes) for IEEE 802.1ad and EEP.]

- IEEE 802.1ad: Red line, increasing with frame size.
- EEP: Blue line, relatively flat with some variance.

---

Hot Interconnects
August 27, 2009

Nathan Farrington
farrington@cs.ucsd.edu
Related Work

  - Fat trees of commodity switches, Layer 3 routing, flow scheduling

  - Layer 2 routing, plug-and-play configuration, fault tolerance, switch software modifications

  - Layer 2 routing, end-host modifications
Conclusion

• General architecture
  – Fat tree of merchant silicon switch ASICs
  – Hiding cabling complexity
  – Pods + Core
  – Custom EEP ASIC
  – Scales to 65,536 ports with 64-port ASICs
• Design of a 3,456-port 10GbE switch
• Design of the EEP ASIC