From Silicon to System: How Hyper-scale Impacts Your Design

Brad Booth
Chair, Consortium for On-Board Optics
Secretary, 25 Gigabit Ethernet Consortium
Principal Architect, Microsoft Azure Networking
Outline

- Hyper-scale
- HSDN: Paths instead of routes
- COBO: Moving the optics inside
- FlexE: 1.6 Terabit Ethernet enabler
- Summary
Cloud Growth Driving Hyper Network Expansion

Network Demand & Capacity Growth

Santa Clara, CA
USA April 2015
Hyper-scale is the Enabler

- 3000+ ISPs peered with globally
- >3x Microsoft's fiber network in N. America could stretch to the moon and back
- +1.4 million Route miles
- 1.5 million + Requests processed per second
- 110m+ petabytes of data move through Microsoft's datacenters each year
- $15 billion+ Microsoft’s cloud investment
Microsoft’s network is one of the two largest in the world
Key Elements at Hyper-scale

Crop Rotation

Security

Power

Cost

Bandwidth
Power Example^  

- Retimer is ~800 mW per 100G
- Each link would require two retimers
  - One retimer for each end of the link
- 1M links would use 1.6 MW
- >500 homes (with 3 KW ADMD) just for retimers

---

Santa Clara, CA  
USA April 2015

^ Rough calculation not based on actual deployment
* ADMD = after diversity, maximum demand
Hierarchical SDN (HSDN)

Forwarding Plane

HIERARCHICAL UNDERLAY PARTITION (UP)

UP-0

UP-1

UP-1

UP-2

UP-2

UP-2

UP-2

Control Plane

HIERARCHICAL CONTROL

HSD N-C
UP-0

HSD N-C
UP-1

HSD N-C
UP-1

HSD N-C
UP-1

HSD N-C
UP-1

HSD N-C
UP-2

HSD N-C
UP-2

HSD N-C
UP-2

HSD N-C
UP-2

HSD N-C
UP-2

HSD N-C
UP-2

One Consistent Abstraction Paradigm → “Infinite” Horizontal Scaling

Santa Clara, CA
USA April 2015

IETF draft-fang-mpls-hsdn-for-hsdc-02
The Life of a Packet

HSDN Label

 Payload

IP Lookup, Push PL0, Push PL1, Push PL2, Push VL

UPBN = Underlay Partition Border Node

Santa Clara, CA
USA April 2015
HSDN Scaling Example

- Scales to tens of millions of network endpoints
- Small forwarding tables
  - Label forwarding information base (LFIB)

<table>
<thead>
<tr>
<th>Number of Server Endpoints</th>
<th>Maximum LFIB with ECMP only</th>
<th>Maximum LFIB with ECMP and TE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3M</td>
<td>~1K</td>
<td>&lt; 14K</td>
</tr>
<tr>
<td>10M</td>
<td>&lt; 2K</td>
<td>&lt; 24K</td>
</tr>
<tr>
<td>40M</td>
<td>&lt; 3K</td>
<td>&lt; 36K</td>
</tr>
</tbody>
</table>

Santa Clara, CA
USA April 2015
On-Board Optics

- Have been around for more than a decade
  - 300-pin MSA
- Today’s solutions are proprietary 😞
- Growing deployment of optics
  - Datacenter networks
  - Consumer applications
- Requires industry standards
  - “Goodness” of z-axis pluggability to the y-axis
Consortium for On-Board Optics

- Specifications for on-board optical modules
  - Electrical interfaces, pin-outs, connectors, thermals, etc.
  - Interchangeable and interoperable modules
- Better front-to-back airflow on equipment
- Increased switch faceplate bandwidth density
- Website: http://cobo.azurewebsites.net
Flexible Ethernet

- Nothing to do with bend radius
- OIF effort targeting
  - Bonding Ethernet PHYs into pipes carrying data at greater rate than a single PHY
  - Channelization to carry several smaller data rates in one or more bonded PHYs

* OIF – Optical Internetworking Forum (www.oiforum.com)
Transport Use Case

- For use in metro and long-haul applications
- Optical bandwidth is reach dependent

Santa Clara, CA
USA April 2015
Datacenter Use Case

- Create larger bandwidth pipes
- Getting to 1.6T
In Summary: Impact of Hyper-scale

- Hyper-scale wanted 25G Ethernet
- July 2014
  - Consortium formed
  - IEEE study group created
- 40G market forecast downgraded

![Graph showing the impact of 25G & 50G]

Dell’Oro January 2015
Thank You!