Optical Interconnect Networks for Datacom and Computercom

Casimer DeCusatis, Ph.D.
IBM Distinguished Engineer & CTO,
System Networking Strategic Alliances
decusat@us.ibm.com
Twitter @Dr_Casimer

Tutorial Paper
OFC 2013
Anaheim, CA
Agenda

▪ What’s in the Data Center Network?
  – **Ethernet**: Not just for the LAN anymore
  – Fibre Channel, InfiniBand, and the WAN
  – Trends in CFP, QSFP, POI, WDM, VCSELs, Si Photonics, and more
  – 5 Mega-Trends for the Data Center

▪ Modern Data Systems Network Architecture
  – **Automated**: Virtual & Overlay Networks
  – **Integrated**: Software Defined Networks
  – **Optimized**: Flat, Converged, Scalable Fabrics
Start of the Data Center Land Rush

IT Solutions will be roughly a $65 B market in 2013, increasing to over $100 B within the next 3 years.

From 2013 to 2020, 90% of IT industry growth will be driven by “3rd Platform” technologies (mobile computing, cloud, & big data analytics) which today represent just 22% of spending.

Spending on data center infrastructure management will increase more than 20% in 2013, with annual sales reaching $690 M by 2016.

* From IDC Predictions 2013: The new data center dynamic R. Villars, J. Koppy, and K. Quinn (December 2012)
What makes up the data center network?

- **Ethernet**
  - Not just for the LAN anymore; architectures are changing
  - Server Pods, iSCSI, NAS, RoCE, FCoE,…
  - 10GE → 40 GE → 100 GE

- **Fibre Channel**
  - Continues to dominate the SAN
  - Alternatives like FCoE growing slowly

- **InfiniBand**
  - Server to server clustering

- **WAN has unique requirements**
  - Early adoption of 100G along with the network core

Plus vendor proprietary variants…
Protocol, performance and latency: One size does not fit all

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Operating System / Application</th>
<th>Small Computer System Interface (SCSI)</th>
<th>Performance</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCIP</td>
<td>FCP</td>
<td>TCP</td>
<td>lossless</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FCIP</td>
<td>IP</td>
<td>lossless</td>
<td></td>
</tr>
<tr>
<td>iSCSI</td>
<td>iSCSI</td>
<td>Ethernet</td>
<td>lossy</td>
<td></td>
</tr>
<tr>
<td>FCoE</td>
<td>FCP</td>
<td>Ethernet with DCB</td>
<td>lossless</td>
<td></td>
</tr>
<tr>
<td>Fibre Channel</td>
<td>FCP</td>
<td>FCP</td>
<td>lossless</td>
<td></td>
</tr>
<tr>
<td>Infiniband</td>
<td>SRP</td>
<td>SRP</td>
<td>lossless</td>
<td></td>
</tr>
</tbody>
</table>

**Operating System / Application**

Small Computer System Interface (SCSI)

- **FCP**: Fibre Channel Protocol
- **iSCSI**: Internet Small Computer System Interface
- **FCoE**: Fibre Channel over Ethernet
- **SRP**: Shortest Path Root Protocol

**Performance**

- **L4**: Layer 4
- **L3**: Layer 3
- **L2**: Layer 2

**Latency**

© 2012 IBM Corporation
Infiniband Roadmap

Are FDR (14G) and EDR (26G) expected to be supported over a ROADM solution?

Source - InfiniBand Trade Association

http://www.infinibandta.org/content/pages.php?pg=technology_overview
Fibre Channel shipped more than 84 Petabits/second of optical bandwidth in 2011 – more than any other networking protocol
- Fibre Channel also has a huge install base

Ethernet bandwidth was just under 73 Petabits/second.

In 2011, Fibre Channel sold over 11.7 million (M) SFP+ with 7.7M of those supporting 8 Gbps Fibre Channel (8GFC)
- 16G FC shipped only 156,000 units but is growing rapidly

Ethernet was the next leading protocol that shipped 17.8M transceivers with 11.8M supporting Gigabit Ethernet.

*LightCounting* is a seven year old market research company focused on the in-depth study of high speed interconnects for the datacom, telecom, and consumer communications markets. Their unique research includes coverage of the number and type of fiber optic transceivers sold throughout the world.
Ethernet performance growth is causing disruptions in DC fabrics:

- **10/40 GE & CEE** → Disrupts storage market (Fibre Channel SAN) and enables denser virtual environment
- **40 GE & CEE** → Will further disrupt cluster market (InfiniBand)
- **100/400 GE & CEE** → Will disrupt server IO market & structure in 4-6 years.
Mega-Trend 1: Big Data Centers

Worldwide High-End Datacenter Space

Scale vs TCO

- Cost/bit should go down with additional scale, not up
  - Consider analogies with compute and storage

- However, cost/bit doesn't naturally decrease with size
  - Complexity in pairwise interactions and any-to-any communication requires more advanced forecasting and control mechanisms
  - Lack of control and determinism in distributed protocols necessitates worst case over-provisioning
  - Complexity of automated configuration to deal with non-standard vendor configuration APIs
  - existing routing mechanisms do not allow for
    - scheduling
    - optimization of explicit objectives

From IDC Predictions 2013: The new data center dynamic
R. Villars, J. Koppy, and K. Quinn (December 2012)

E. Crabbe & V. Valancius, “SDN at Google”, IETF 841 IRTF meeting (August 1, 2012) and
U. Holzle keynote, ONS 2012
Mega-Trend 2: Big Data

- Worldwide information volume growing by 59% annually (Gartner Group, June 2011)
  - 2.7 Zettabytes in 2012, growing to 8 Zettabytes in 2015
  - Top 5 driving applications are financial transactions, email, images, web logs, and Internet traffic
  - 50% organizations managing 500 TB or more, 20% have over 10 PB
  - IT assets have large revenue generating potential, including forecasting (predictive analysis and “what if” scenarios), competitive positioning, customer retention/satisfaction

- Characteristics of big data (Information Week, April 2012)
  - Size: 30 TB or more
  - Type: multiple formats of unstructured, semistructured, and structured data (although 90% of new data is unstructured)
  - Latency: changing rapidly, requires accelerated analysis
  - Complexity: large files with incomplete or inconsistent data
  - Data bases may be hosted in the cloud (analytics as a service)

- Network Traffic Implications
  - Bulk transfer; High bandwidth, mostly east-west between servers, can overwhelm traditional networks
  - Data aggregation/partitioning; High bandwidth to exchange data between a large number of servers
  - Aggregation & shuffling patterns can become bottlenecks in an oversubscribed network (requires application tuning)
  - Control messages; Latency sensitive, low data rate

- Mega-Trends 1 & 2 drive new network architectures in the data center
Network Architecture - The Beauty of Trees

- In the beginning, Ethernet was used to interconnect stations (e.g. dumb terminals), initially through repeater & hub topologies...

And... eventually through switched topologies.

- Ethernet campus evolved into a tree structure
  - Typically: core, services, aggregation & access planes.
  - Traffic is mostly North-South (directed outside campus).
  - To avoid spanning tree problems, campus networks typically are divided at access.
The Beauty of Trees

- In the beginning, Ethernet was used to interconnect stations (e.g. dumb terminals), initially through repeater & hub topologies...

And... eventually through switched topologies.

- Ethernet campus evolved into a tree structure
  - Typically: core, services, aggregation & access planes.
  - Traffic is mostly North-South (directed outside campus).
  - To avoid spanning tree problems, campus networks typically are divided at access.

The industry liked the tree structure & applied it to DC
The Repotted Campus Tree

Soo... Campus Ethernet tree was repotted to the **Enterprise Data Center**. *which...*
Soo... Campus Ethernet tree was repotted to the Enterprise Data Center. **which**...

- Has different traffic patterns:
  - 50-75% East-West* in DC
  - 95% North-South in Campus

- Has different fabric performance needs
  - Lossless traffic for storage
  - Low latency & high bandwidth for clusters

Evolved into a virtual compute model, with different demands:

- From static workloads
  - to dynamic workloads
  - to multi-tenant, dynamic workloads

...which today results in complex and/or inefficient service plane (e.g. to protect East-West traffic)
The industry is looking for a smarter Data Center Infrastructure that solves these issues.
Data Center Infrastructure Requirements

**Integrated**
- Simple, consolidated management
- Network Service agility
- Software Defined Network platform

**Automated**
- Workload Aware Networking
- Dynamic provisioning

**Optimized**
- Converged (FCoE, iSCSI, NAS, RDMA)
- Single, flat (1 managed switch cluster)
- Traditional or OpenFlow based fabric
Open Datacenter Interoperable Network (ODIN)

Endorsed by Juniper, Brocade, Huawei, NEC, BigSwitch, Adva, Ciena, Alcatel-Lucent, Extreme Networks, Marist College

Layer2

WAN

MPLS/VPLS Enabled

Core Layer
Pooled, Virtual Appliances

Layer2

40 – 100 Gbps links

TOR/Access Layer
w/ TRILL, stacked switches & lossless Ethernet

Layer2

Embedded Blade Switches & Blade Server Clusters
w/embedded virtual switches

OpenFlow controller

SAN

FCoE
Gateway

10 Gbps links

8 Gbps or higher
FC links
10 Gbps
ISCSI / NAS links

FCoE Storage

* Gartner Group, “Debunking the myth of the single-vendor network”, 17 November 2010

http://www-03.ibm.com/systems/networking/solutions/odin.html
Mega-Trend 3: Pods
(will represent 1/3 of all data center cloud deployments by 2016)

Building Blocks

Chassis
14 half-wide bays for nodes

Compute Nodes

Storage Node
V7000
Expansion inside or outside chassis

Management Appliance
Optional

Networking
10/40GbE, FCoE, IB
8/16Gb FC

PureFlex
Pre-integrated compute, storage, networking, virtualization & management

Pods interconnect into warehouse-scale data centers
100,000+ servers with 10G links requires 1 Petabyte bandwidth

• 4 to 8 PureFlex Racks, with up to 42 servers each; total 336 servers, 672 10 GE ports (3.8:1 oversubscription); Multi-hop FCoE for Flex
• Optional LAG into stacked 8264s (simpler configuration)
• Fiber optic components play a key role in realizing the potential of data center networks

From Laser Focus World:
Optical technologies scale the datacenter,
by Hong Liu and Ryohei Urata, Google 12/01/2012
The **Quad Small Form-factor Pluggable** (often abbreviated as **QSFP** or **QSFP+**) is a compact, hot-pluggable transceiver used for data communications applications. QSFP+ transceivers are designed to support Serial Attached SCSI, 40G Ethernet, 20G/40G Infiniband, and other communications standards.

IEEE support sending Ethernet frames at 40 and 100 gigabits per second over multiple 10 Gbit/s or 25 Gbit/s lanes. Previously, the fastest published Ethernet standard was 10 Gigabit Ethernet.

QSFP+ pluggable interface, hosting a range of electrical and optical I/O solutions. Courtesy: Molex.
C form-factor pluggable (CFP) is a multi-source agreement to produce a common form-factor for the transmission of 100G and above digital signals. The c stands for the Latin letter C used to express the number 100 (centum).

Electrical interface: 10 x 10G in each direction
Optical interface variants: 10x10Gbit/s and 4x25Gbit/s variants of 100Gbit/s interconnects (typically referred to as 100GBASE-LR10 and 100GBASE-LR4 in 10 km reach, and 100GBASE-ER10 and 100GBASE-ER4 in 40 km reach respectively.)
CFP transceivers can support a single 100 Gbit/s signal like 100GbE or OTU4 or one or more 40 Gbit/s signals like 40GbE, OTU3 or STM-256/OC-768

References http://www.cfp-msa.org/
Enabling Optical Technologies

- **Parallel optical interconnects (POI, or Space division multiplexing)**
- Extensions of the switch backplane / optical bus over multimode fiber (individual cables, ribbons, or fanout cables)
- Pluggable transceivers or active optical cables
- Active optical cables have potential to reach cost/gigabit targets sufficient to displace copper for short distances (tens of meters) at data rates of 10-40G and higher
  - BER lower than copper (10e-15 vs 10e-12)
  - No Xtalk within cable, better EMI immunity than copper
  - Low power consumption (350 mW to 1 W per end)
  - SFP+, QSFP+, CXP form factors

- **Wavelength-division multiplexing (WDM)**
- Traditionally used for interconnecting multiple data centers over extended distances using single-mode fiber; gracefully scales telecom bandwidth over many generations
- Potential use for connections to and from the pods to core switches, replacing traditional parallel optical transceivers
- Integrated WDM transceivers (40G, 100G, 400G, and beyond) can aggregate signals from multiple channels with a common destination over a single strand of fiber (addressing cabling complexity issues)

100G for Telecom = Optimized for Distance, Capacity
100G for Datacom = Optimized for Cost, Power, Space
Lasers & Silicon Photonics

- **High-speed VCSELs, DFBs**
  - Low-power, inexpensive VCSELs) and multimode fiber already play a critical role for 10 Gbit/s communication within the datacenter.
  - Significant progress has been made in higher-speed VCSELs using alternate materials, overcoming the reliability and yield hurdles to scale VCSELs significantly beyond 20 Gbit/s link speeds has thus far proven difficult.
  - Furthermore, traditional VCSELs coupled with MMF have a limited distance-bandwidth product due to modal dispersion.

- **High Power DFB lasers**
  - Higher-power, more expensive DFB lasers and single-mode fiber are often used in the datacenter to cover the reach beyond 300 m at 10 Gbit/s. As we scale from 10G to 25G per lane, DFB lasers can use more novel quaternary materials such as indium gallium aluminum arsenide/indium phosphide (InGaAlAs/InP) that exhibit better high-temperature performance and higher speeds.
  - Novel DFB laser structures such as short-cavity and lens-integrated surface-emitting DFB lasers have also been demonstrated. These approaches provide higher device bandwidth and a narrower spectrum in comparison to their VCSEL-based counterparts to increase interconnect bandwidth and reach while maintaining low power consumption and cost.

- **Silicon Photonics**
  - IBM has recently transferred silicon nanophotonics into a commercial manufacturing foundry
  - A variety of silicon nanophotonics components, such as wavelength division multiplexers (WDM), modulators and detectors up to 25 Gbit/channel, are integrated side-by-side with a CMOS electrical circuitry in a 90 nm CMOS fab
Optical component technologies play a vital role in reducing cost, power, and cabling complexity.

But...what about software and network management issues? What’s the best way to deliver new network fabric services?

The availability of network services has been gated by vendor’s business priorities. Networking does not have a Linux development community equivalent.
Mega-Trend 4: SDN Makes Networks Smarter, More Open

VMware to buy Nicira for $1.26B in a strategic leap of faith

OpenFlow developed at Stanford / Berkeley

Open Networking Summit 2012 Plays to a Packed House

HP hops on the OpenFlow train with 16 new switches

IBM launches beefy OpenFlow switch for data centers, cloud

Juniper Embraces OpenFlow

Interop 2011 could have been called The OpenFlow Show.

10 Coolest Startups Of 2011

7. Big Switch Networks

Tech Titans Back OpenFlow Networking Standard

Nicira lands $9M for software to virtualize networks

Where is OpenFlow?

Source: Gartner technology hype cycle, adapted from Wikipedia

See SDN: a Theory of Everything
www.wired.com/insights/2012/12
Each network element has its own control and management plane
Optimized

OpenFlow Network

OS

Mgt Plane
Telnet, SSH, SNMP, NTP, SYSLOG, HTTP, FTP/TFTP

Control Plane
Network topology, ACLs, Forwarding & Routing, QoS, Link Management

Data Plane
Link, Switching, Forwarding, Routing

Services run as Apps

Software Defined Network Stack

Mgt Plane
Telnet, SSH, SNMP, NTP, SYSLOG, HTTP, FTP/TFTP

Apps
Multipath, Security, FCF,…

Control Plane
Network topology, ACLs, Forwarding & Routing, QoS, Link Management

Control plane is extracted from the network
Distributed or Centralized

- Ethernet topologies were built distributed
  Scalable but hard to monitor

- Openflow topologies (today) are centralized
  Control-data separation forces this model

Strengths of one approach are weaknesses of the other
Centralized is better suited to modern cloud applications
**Today’s Network**

- Sub-optimal traffic flow, limited scale
  - Traffic is not optimized across groups (cross-subnet VM-VM traffic must go North-South), even for virtual network services
  - Limited scaling (lacks multi-tenancy)
  - Static workload, Static network state

**Optimized Network**

- Optimized traffic flow, multi-tenant scale, disjoint multipathing
  + Optimizes traffic within and across groups (cross-subnet VM-VM traffic stays in server) – Application Aware Network
  + Multi-tenant scaling for CSPs
  + Dynamic workload, Network state moves with VM
Mega-Trend 5: Network Virtualization

- Number of VMs per socket is rapidly growing (10x every 10 years).
  - Increases amount of VM-VM traffic, increases network complexity associated with creating/migrating layer-2 (VLANs, ACLs...) & layer-3 (e.g. Firewall, IPS) attributes.
- Network virtualization makes many devices look like one device (the opposite of server virtualization)
- Server virtualization (VMWare on x86) drove MORE server volumes, not less. The same thing will happen with networking
- Virtual switches like IBM DVS 5000v (supporting IEEE 802.1Qbg) coordinates migration of the network state with the VM.
Isolate a few (long) flows for preferential treatment by applications

Federation of controllers with each controller handling (smaller) integrated system (pod)
Networking Technology Value Proposition

- No customer pays for the re-invention of the wheel
- Customers do pay for a smoother ride
We are entering an era of Network Affluence

With affluence comes a demand for Quality of Life:

• Can you ease my provisioning headache?
• Can you hide all complexity of the physical infrastructure?
• Can my applications talk to my network?
• Can you simplify how I monitor my network?

• Can “this particular communication” be of “Platinum” service
  • Can you guarantee certain latency characteristics? End to End?
  • Can you guarantee certain bandwidth? End to End?
Conclusions

- Traditional telecom applications spend more at the link end points to maximize spectral efficiency of long-distance fiber links – see for example recent R&D on coherent transmission devices.
- For datacom, spectral efficiency tends to be traded for lower power, cheaper transceiver cost and a network fabric with abundant path diversity.
- Within the datacenter, fiber resources are much more abundant, cheap, and easy to deploy. Thus, the transceiver cost must be aggressively reduced.

- As we enter an age of network affluence, the focus shifts to quality of service & application awareness.
- Five Mega-Trends affect optical network design in the data center:
  - Big Data Centers
  - Big Data
  - PODs (and new network architectures)
  - Network becomes Smarter & More Open
  - Network Virtualization

- Virtualization of the network is the next big frontier
  - Software-defined Networking (OpenFlow + DOVE overlays) are a disruptive technology in this landscape
  - Application velocity and deployment velocity become equally important

- The most cost effective path forward requires open standards and interoperable networks.
- Business process improvement provides more value add for IT than capex/opex reduction.