Operational Considerations for Deploying 100 Gigabit Ethernet

NANOG51

Brent van Dussen, Limelight Networks
<bvd@llnw.com>

Greg Hankins, Brocade
<ghankins@brocade.com>

Agenda

- What problems does 100 GbE solve?
- 100 GbE Technology Update
- Operational Considerations

What problems does 100 GbE solve?

- Higher capacity interfaces beyond 10 GbE
  - Core, edge, metro, HPC and data center applications
- General and vendor-specific LAG and ECMP issues
  - Scalability
  - Manageability
  - Hashing
  - Large flow distribution
- Side effects
  - Lower cost and higher density 10 GbE
  - Higher bandwidth enables new applications

LAG and ECMP Issues

- Often a great solution but doesn’t solve every bandwidth capacity problem
    - Scalability issues apply to any link speed
- Limitations on number of LAGs and number of links in a LAG
    - CPU resources are needed to run LACP which limits the number of LAGs per router
    - Extremely complex hashing algorithms are needed to scale number of links in a LAG

LAG and ECMP Issues

- Hashing is usually decoupled from link capacity
  - Links have no way to signal that they are full
  - Huge flows could exceed the capacity of the link in a LAG (rarely seen today with 10 GbE)
  - Lots of large flows could exceed the capacity of the link in a LAG if hashing breaks
- Hashing algorithm problems
  - Odd links
  - Too simple
  - Unable to hash on fields deep in the packet (MPLS VPNs)
- Even a good hashing algorithm hashes badly without header field diversity

Agenda

- What problems does 100 GbE solve?
- 100 GbE Technology Update
- Operational Considerations

Recent 100 GbE Developments

- Shipping 1st generation media, test equipment, router interfaces, and optical transport gear in 2010/2011
- 2nd generation projects based on 4 x 25 Gb/s electrical signaling have started
- New IEEE Copper Study Group approved in November 2010
  - 100GBASE-KR4 – 4 x 25 GB/s over backplane
  - 100GBASE-CR4 – 4 x 25 Gb/s over copper cable
  - http://www.ieee802.org/3/100GCU/index.html

Recent 100 GbE Developments

- MSA formed in December, 2010 to develop a 100 GbE CFP standard using 10 x 10 Gb/s signaling over 2 km SMF
  - Much lower cost than 4 x 25 Gb/s 100GBASE-LR4 CFPs
  - Draft standard is finished, final standard expected in March, 2011
  - 2 km, 4 km and 10 km media available today
  - http://10x10msa.org/
- IEEE is expected to start work in July, 2011 to define several interfaces
  - 100GBASE-SR4 – 4 x 25 Gb/s over OM3 MMF
  - 100GBASE-FR4 – 4 x 25 Gb/s over SMF for 2 km
  - CAUI-4 – electrical signaling to the CFP2
  - CPPI-4 – electrical signaling to the QSFP2/CFP4

100 GbE Technology Summary

1st and 2nd Generation, MSA

Physical Layer Reach	*1?* m Backplane	*3 - 5?* m Copper Cable	7 m Copper Cable	*100?* m OM3 MMF	100 m OM3, 150 m OM4 MMF	2 km SMF		10 km SMF		40 km SMF
Name	100GBASE-KR4	100GBASE-CR4	100GBASE-CR10	100GBASE-SR4	100GBASE-SR10	10x10	100GBASE-FR4	LR10-10km	100GBASE-LR4	100GBASE-ER4
Standard Status	Future IEEE	Future IEEE	2010 IEEE 802.3ba	Future IEEE	2010 IEEE 802.3ba	2011 10x10 MSA	Future IEEE	Exceeds 10x10 MSA	2010 IEEE 802.3ba	2010 IEEE 802.3ba
Generation	2nd	2nd	1st	2nd	1st	1st	2nd	1st	1st	1st
Electrical Signaling	4 x 25 Gb/s	4 x 25 Gb/s	10 x 10 Gb/s	4 x 25 Gb/s	10 x 10 Gb/s	10 x 10 Gb/s	4 x 25 Gb/s	10 x 10 Gb/s	10 x 10 Gb/s	10 x 10 Gb/s
Media Signaling	4 x 25 Gb/s	4 x 25 Gb/s	10 x 10 Gb/s	4 x 25 Gb/s	10 x 10 Gb/s	10 x 10 Gb/s	4 x 25 Gb/s	10 x 10 Gb/s	4 x 25 Gb/s	4 x 25 Gb/s
Media Type	Backplane	Twinax	Twinax	MPO MMF	MPO MMF	Duplex SMF	Duplex SMF	Duplex SMF	Duplex SMF	Duplex SMF
Media Module	Backplane	QSFP2	CXP	QSPF2	CXP, CFP	CFP	CFP2	CFP	CFP	CFP
Availability	2013	2013	2010	2013	2010	Q1 2011	2013	Q1 2011	2010 (CFP2 in 2013)	2012+ (CFP2 in 2013)

100 GbE Market Overview

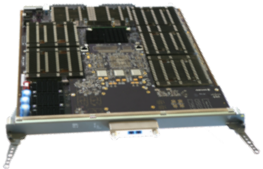
CFP Optics

Physical Layer Reach	100 m OM3, 150 m OM4 MMF	2 km(*) SMF	10 km SMF	
Media Module	 100GBASE-SR10	 LR10-4km	 LR10-10km	 100GBASE-LR4
Media Type	 MPO MMF	 Duplex SMF	 Duplex SMF	 Duplex SMF
Power (W)	6	14	15	20
Availability	Now	Now	Now	Now
Sample Relative List Price	$	5.3 x $	8.3 x $	11.6 x $

(*) 2 km MSA standard, some vendors support longer distances

100 GbE Market Overview

Router Interfaces and Media

Vendor	Feature Set	Product Line	CFP Media
Alcatel-Lucent	L2, IP, MPLS	 7450 ESS, 7750 SR	LR10-10km, 100GBASE-LR4
Brocade	L2, IP, MPLS	 MLX/XMR Series	100GBASE-SR10, LR10-4km, LR10-10km, 100GBASE-LR4
Cisco	IP, MPLS	 CRS-3	100GBASE-LR4
Juniper	IP, MPLS	 T1600, TX Matrix Plus	100GBASE-LR4

Optical Transport Network (OTN) Support

- IEEE has worked closely with the ITU-T SG15 to define interoperable Ethernet and optical transport standards
- Transport for 40 and 100 GbE is defined in ITU-T G.709 (Amendment 3, October 2009)

10 GbE LAN PHY — OTU2e

Maps on to OTU2e at 11.25 Gb/s

10 GbE WAN PHY — OTU2

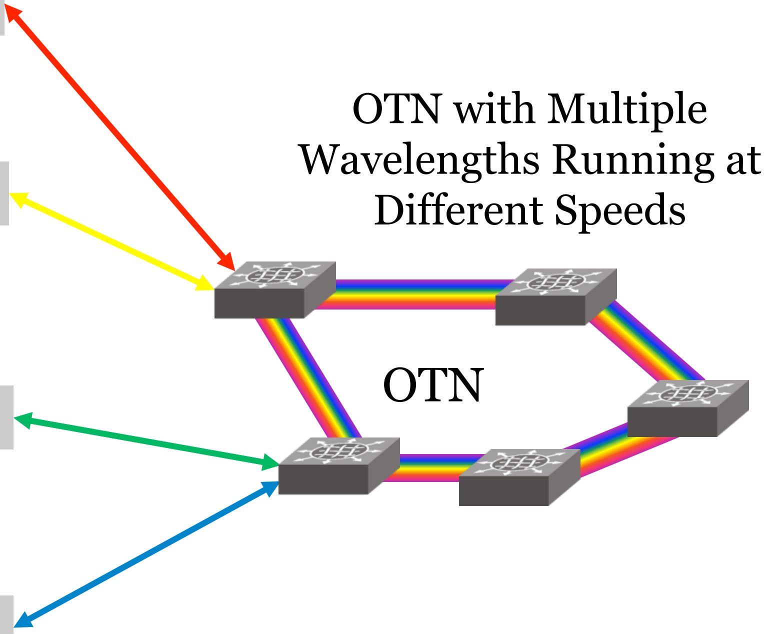
9.95 Gb/s maps on to OTU2 at 10.71 Gb/s

40 GbE — OTU3

Maps on to OTU3 at 43.02 Gb/s

100 GbE — OTU4

Maps on to OTU4 at 111.81 Gb/s



OTN with Multiple Wavelengths Running at Different Speeds

Agenda

- What problems does 100 GbE solve?
- 100 GbE Technology Update
- Operational Considerations

Motivational Drivers

- General rule of thumb that's surfaced
- Time to upgrade when…
  - Edge hosts come online with interface speeds equal to the highest interface speed on the network
  - Bandwidth over aggregated links 2-3x greater than max bandwidth of new interface technology
- On one hand we have huge hosts/flows that spike individual 10G LAG members
- With huge LAGs we end up spending more and more time installing and maintaining individual member interfaces

Technical Considerations

- Using SR10 is going to be limited to wherever MPO MMF is available
- Limits use to runs between routers in the same cage or from router to optical gear in the same cage
- LR4 vs. LR10

Benefits

- Significant man hours saved installing and supporting the same amount of bandwidth
- Cross-connect costs reduced
- Router density increased immediately ~20%, extending platform life
- No longer pushing up against vendor max number of LAG member limitations

Where Do We Need 100 GbE?

- Connectivity between various facilities in a city using optical gear and dark fiber
- Connectivity between routers in a particular facility to maintain core capacity
- Peering exchanges and PNIs with larger networks
- Backbone between city pairs

100 GbE Over Optical Gear

- Client and line facing cards available since early 2010
- 100GBASE-SR10 and 100GBASE-LR4 available for client ports
- Still uses standard 50 GHz spacing
- 88 channel filters makes for 8800 Gb/s over existing dark fiber pairs
- Shelf density of 100 G not as good as current 10 G
- Next iteration of 100 G will match or surpass 10 G shelf density

Thanks

Questions?