



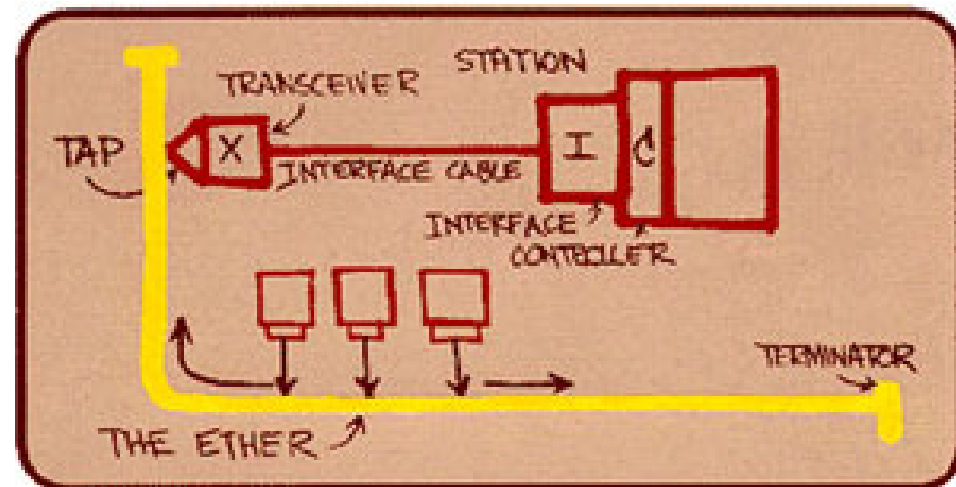
BROCADE

Chris Bull
Service Provider
Systems Engineer
ANZ



100 GIGABIT ETHERNET

So we've got 802.3ba,
now what do we do with it?





Standards

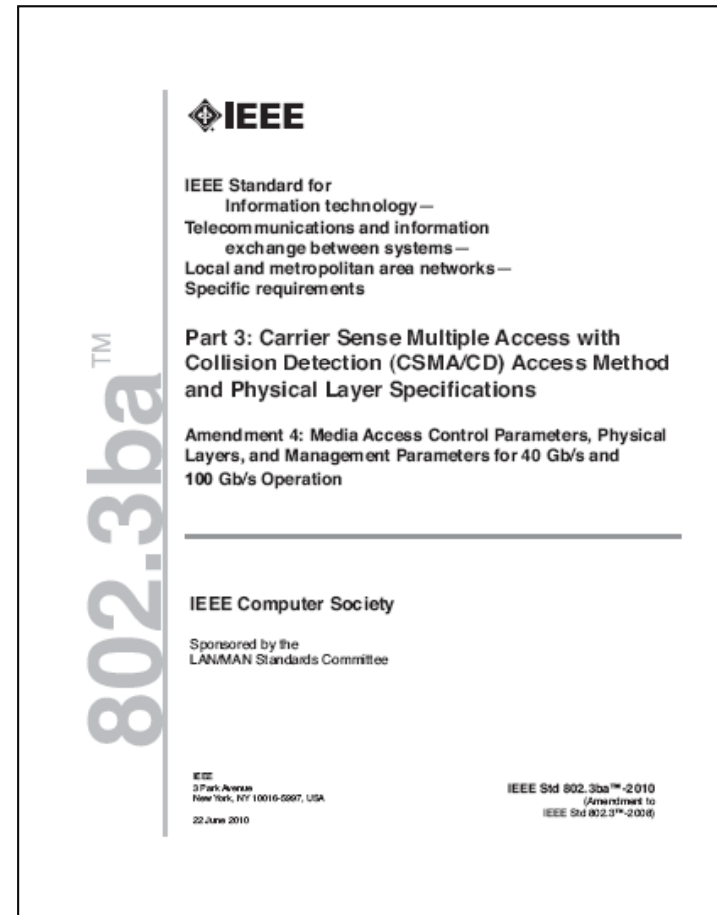
How we got to 802.3ba



Summary of Recent 40 Gigabit and 100 Gigabit Ethernet Developments

Completing the Standard for 40 and 100GE

- *We're done with 802.3ba!!!*
- Proceeded to state of “STOP MAKING CHANGES” in March 2010
- Final Draft 3.2 submitted for approval on April 29, 2010
- IEEE 802.3ba standard approved June 17, 2010
 - 457 pages will be added to IEEE 802.3-2008



Summary of Reach Objectives and Physical Layer Specifications – Updated July 2009

Physical Layer Reach	1 m Backplane	7 m Copper Cable	100 m OM3, 125 m OM4 MMF	10 km SMF	40 km SMF
----------------------	---------------	------------------	--------------------------	-----------	-----------

40 Gigabit Ethernet: Target Applications – Servers, Data Center, Campus, Metro, Backbone

Name	40GBASE-KR4	40GBASE-CR4	40GBASE-SR4	40GBASE-LR4	
Signaling	4 x 10 Gbps	4 x 10 Gbps	4 x 10 Gbps	4 x 10 Gbps	
Media	Copper Backplane	Twinax Cable	MPO MMF	Duplex SMF	
Module/Connector		QSFP Module, CX4 Interface	QSFP Module	QSFP Module, CFP Module	
Availability	No Known Development	2010	2010	QSFP 2011-2012 CFP 2010	

100 Gigabit Ethernet: Target Applications – Data Center, Campus, Metro, Backbone, WAN

Name		100GBASE-CR10	100GBASE-SR10	100GBASE-LR4	100GBASE-ER4
Signaling		10 x 10 Gbps	10 x 10 Gbps	4 x 25 Gbps	4 x 25 Gbps
Media		Twinax Cable	MPO MMF	Duplex SMF	Duplex SMF
Module/Connector		CXP Module	CXP Module, CFP Module	CFP Module	CFP Module
Availability		2010	2010	2010	2011-2012

Recent 100 GE Developments

- Shipping 1st generation media, test equipment, router interfaces, and optical transport gear in 2010/2011
- 2nd generation projects based on 4 x 25 Gbps electrical signaling have started
- New IEEE Copper Study Group approved in November 2010
 - 100GBASE-KR4 – 4 x 25 Gbps over backplane
 - 100GBASE-CR4 – 4 x 25 Gbps over copper cable
 - <http://www.ieee802.org/3/100GCU/index.html>
- IEEE is expected to start work in July, 2011 to define several interfaces
 - 100GBASE-SR4 – 4 x 25 Gbps over OM3 MMF
 - 100GBASE-FR4 – 4 x 25 Gbps over SMF for 2 km
 - CAUI-4 – electrical signaling to the CFP2
 - CPPI-4 – electrical signaling to the QSFP2/CFP4





Interfaces



A look at the optics and transport

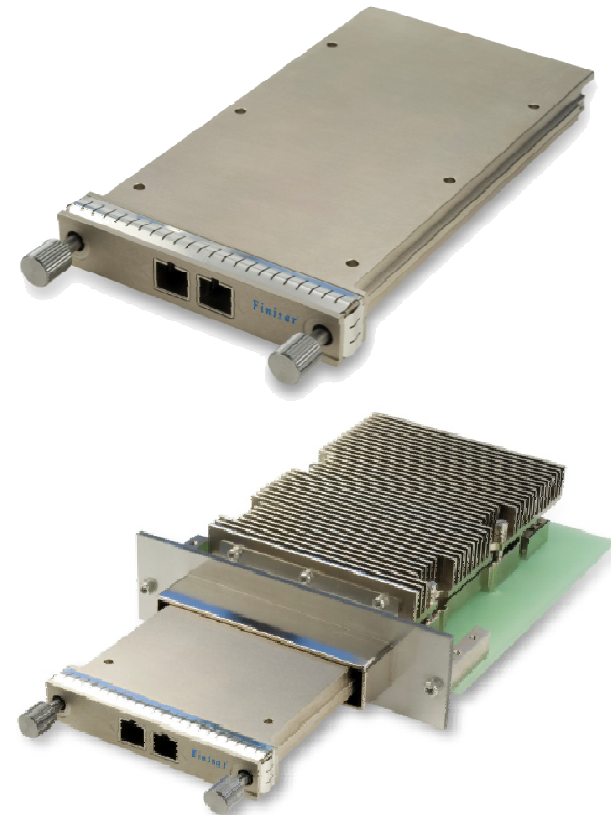
Cable problems

Mapping to carrier DWDM

40 and 100 GE CFP Modules

C (100) Form-factor Pluggable

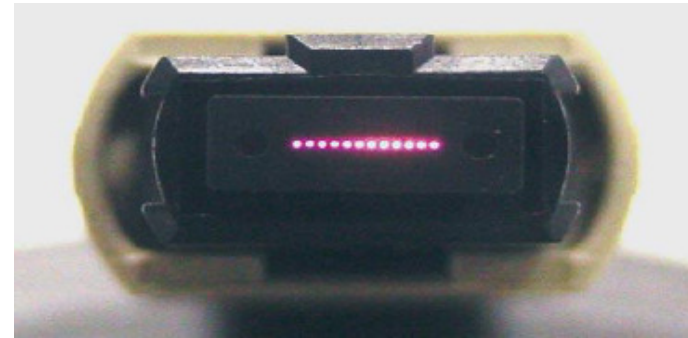
- New module optimized for longer reach applications
- Used for 40GBASE-SR4, 40GBASE-LR4, 100GBASE-SR10, 100GBASE-LR4 and 100GBASE-ER4
 - Dense electrical connector enables a variety of interfaces (10 x 10Gbps electrical lanes)
 - Integrated heat sink allows efficient cooling
 - 100 GE modules are complicated because of 10 Gbps electrical to 25 Gbps optical conversion
- About twice as wide as a XENPAK (14 mm high x 82 mm wide x 145 mm long)
- 100GBASE-ER4 not expected until 2011-2012



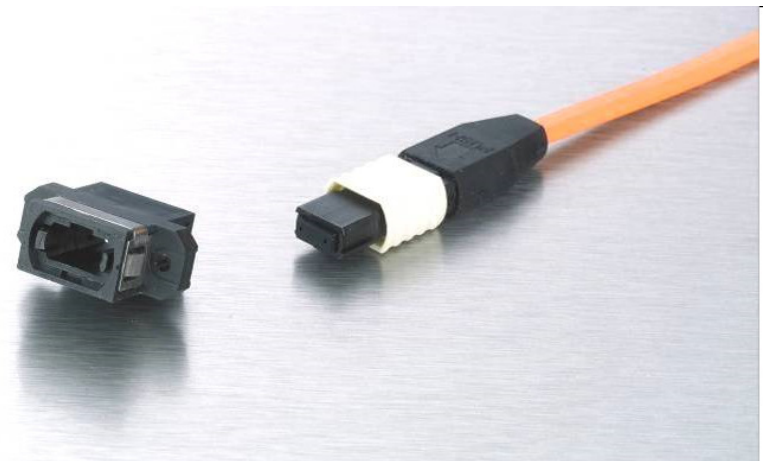
Example CFP and Cage Assembly

MPO/MTP Cable Assemblies

- MPO = “Multi-fibre Push On” assembly
 - Also called MTP by Corning
- Wide variety of high density cabling options
 - MPO to MPO
 - MPO cassette for patch panels
 - MPO breakout into SC, LC, etc
- 40GBASE-SR4
 - 12 fibre MPO cable, uses 8 fibres
- 100GBASE-SR10
 - 24 fibre MPO cable, uses 20 fibres
- May require new ribbon fibre infrastructure



12 Fibre MPO Cable



Fibernet MTP/MPO Cable

100 GE CXP Modules

C (100) X (10) Pluggable

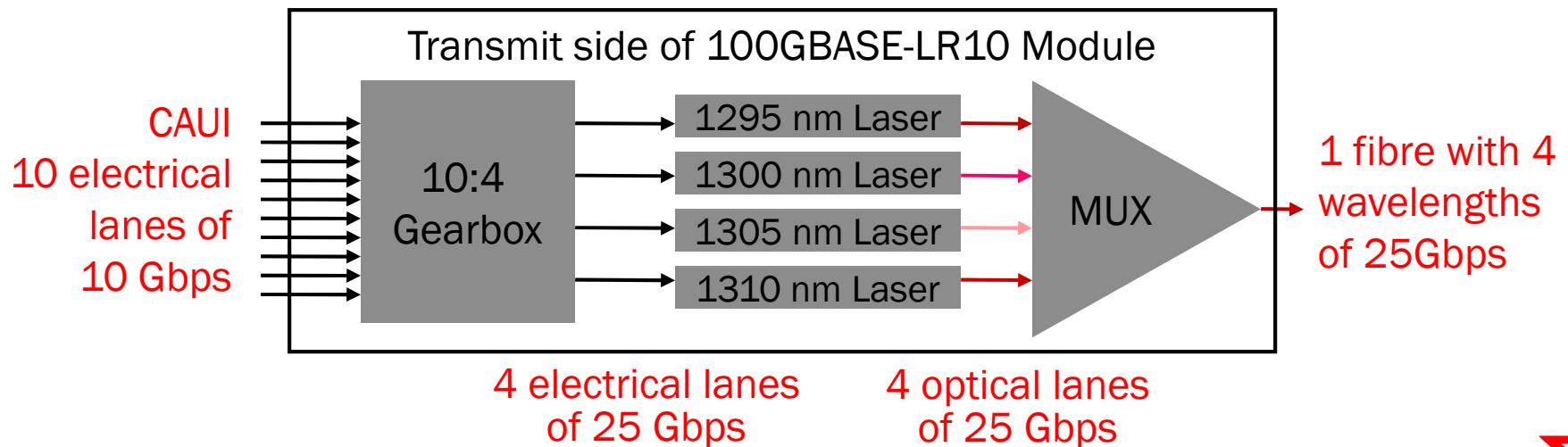
- Created for high density short reach interfaces
 - Targeted for data center applications
 - Small compact form factor enables high density but limits distance
- Used for 100GBASE-CR10, 100GBASE-SR10 and InfiniBand 12X QDR
 - 12 bidirectional channels/24 fibres
 - 100 GE uses 10 of the 12 channels
- Slightly wider than an XFP (27 mm wide x 45 mm long)



Finisar C.wire MMF CXP

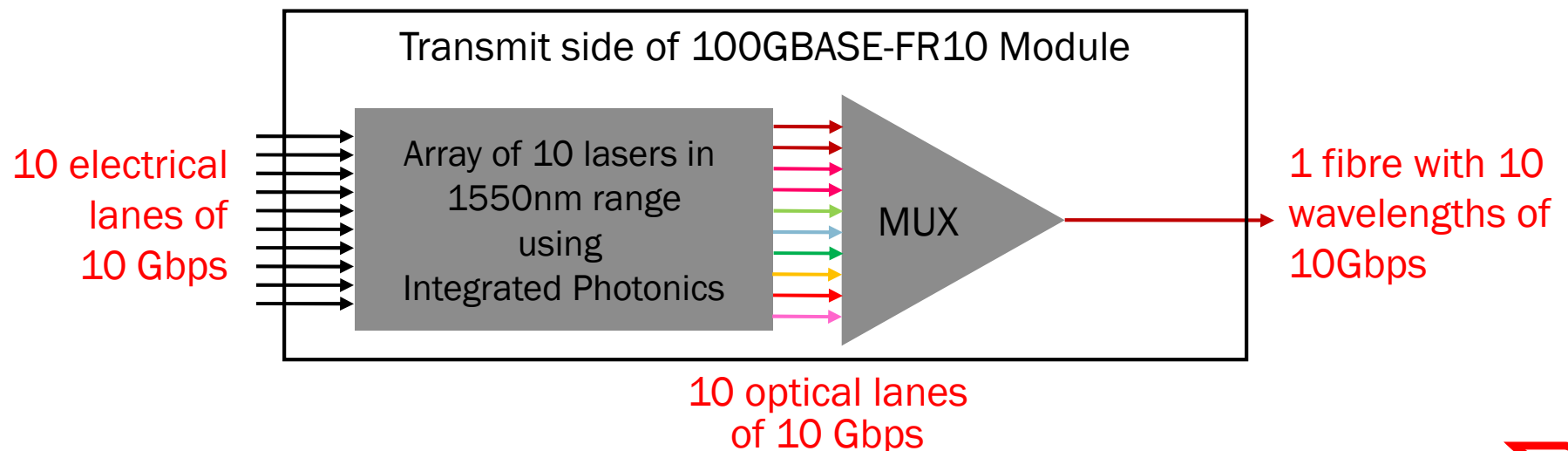
100GBASE-LR4 for 10km

- 10GBASE-LR4 solutions consume over 20 Watts because of a gearbox that converts the 10x10Gbps electrical lanes into 4 lanes of 25Gbps
- 10GBASE-LR4 is expensive because of gearbox and 25Gbps lasers



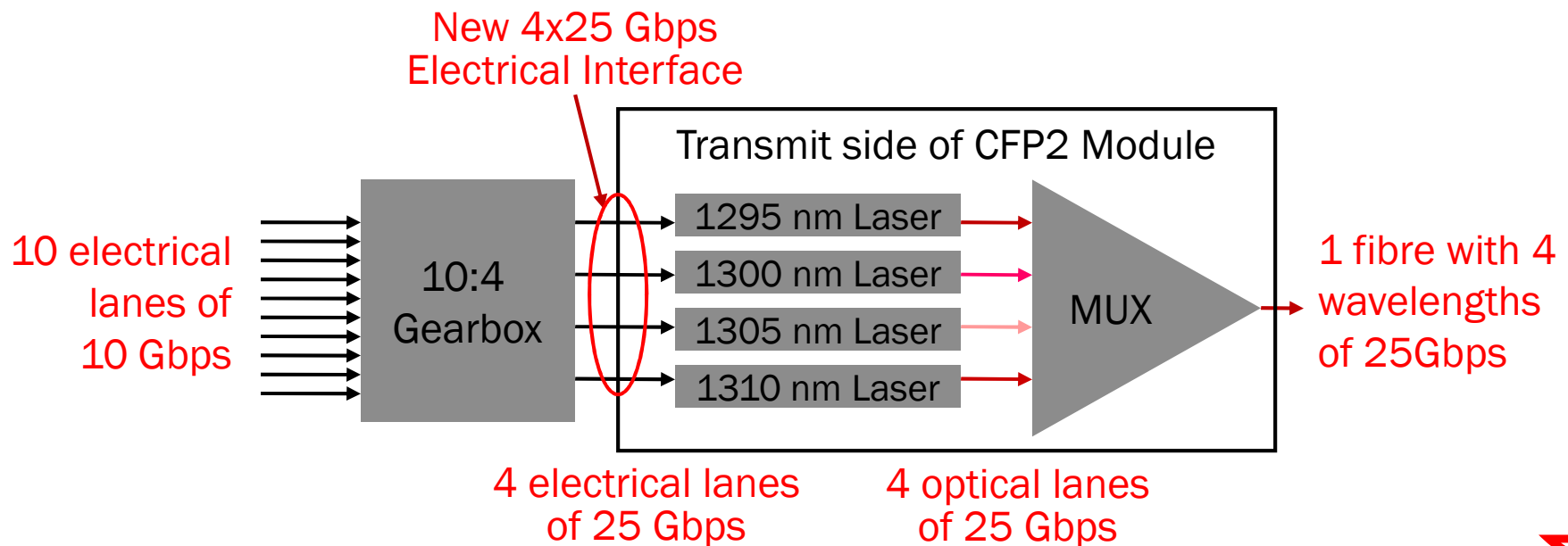
100GBASE-FR10 for 2 km

- 100GBASE-FR10 doesn't have the gearbox and consumes about 14W
- Solutions go either 2km or 4km at considerably less cost than 100GBASE-LR4 modules – possibly 10km
- Brocade is chairing the 10x10 Multi-Sourcing Agreement between Santur and JDSU to standardize the solution



The CFP2 Module

- The CFP2 module doesn't have the gearbox so will dissipate less power in second generation modules
 - Most 10 km solutions will consume less than 12 Watts
 - Probably available in the 2013 timeframe
- The gearbox will be in the switch ASIC or in a separate PHY chip



100 GE Technology Summary

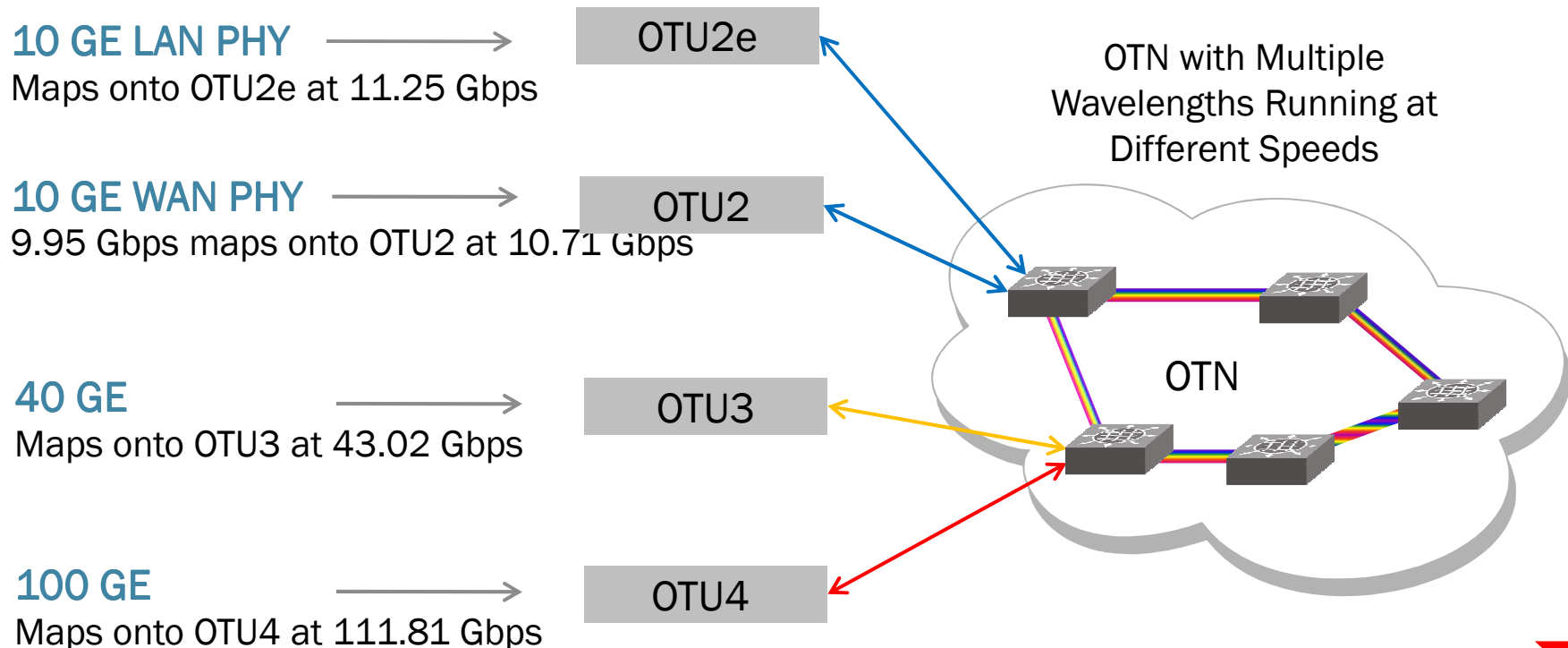
1st and 2nd Generation, MSA

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	100GBASE-FR10	100GBASE-FR4	LR10-10km	100GBASE-LR4	100GBASE-ER4
Standard Status	Future IEEE	Future IEEE	2010 IEEE	Future IEEE	2010 IEEE	2011 10x10 MSA	Future IEEE	Exceeds 10x10 MSA	2010 IEEE	2010 IEEE
Generation	2 nd	2 nd	1 st	2 nd	1 st	1 st	2 nd	1 st	1 st	1 st
Electrical Signaling	4 x 25 Gbps	4 x 25 Gbps	10 x 10 Gbps	4 x 25 Gbps	10 x 10 Gbps	10 x 10 Gbps	4 x 25 Gbps	10 x 10 Gbps	10 x 10 Gbps	10 x 10 Gbps
Media Signaling	4 x 25 Gbps	4 x 25 Gbps	10 x 10 Gbps	4 x 25 Gbps	10 x 10 Gbps	10 x 10 Gbps	4 x 25 Gbps	10 x 10 Gbps	4 x 25 Gbps	4 x 25 Gbps
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)

Optical Transport Network (OTN) Support

No WAN PHY needed at 40GE and 100GE

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





Implementation and Deployment



IX,

SP – low cost, best effort transport option

Cost of $n \times 10$ vs 100

Monitoring – a case for sFlow?

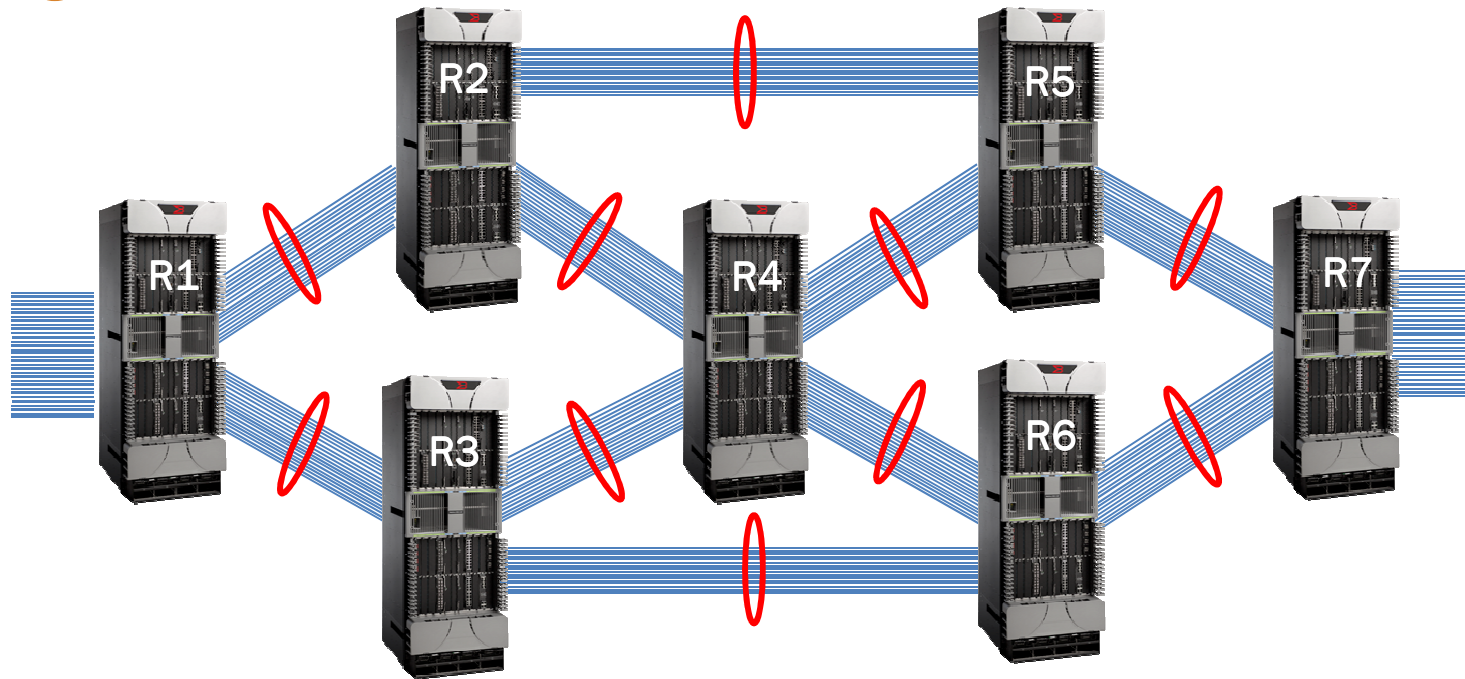
Applications for Early 100GE Deployments

- IP transit networks
- Internet Exchange Points
- High-end R&D networks
- Web 2.0, Content Data Centers
- High-capacity networks where transport equipment and router are collocated
- Networks with potential for suboptimal LAG member utilization
 - Flows greater than 10Gbps
 - Multicast over LAG



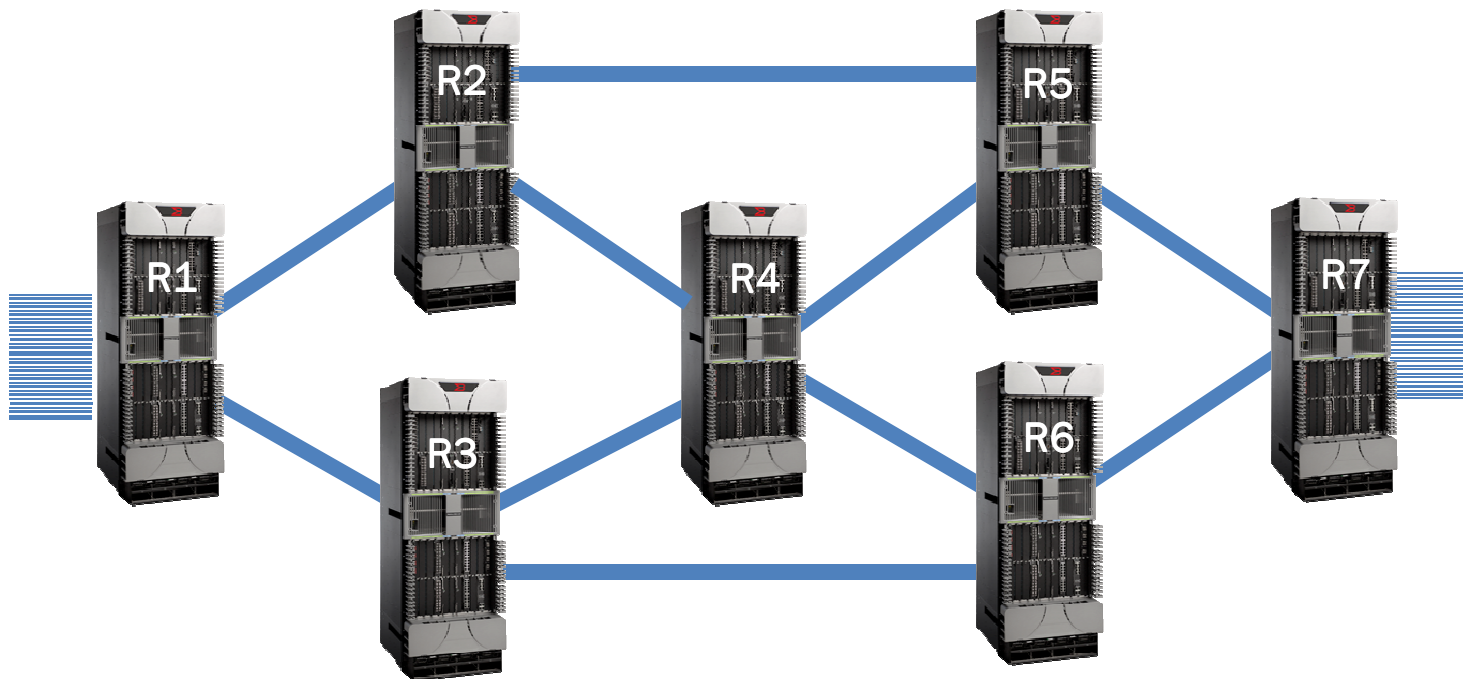
10Gbps Carrier Trunks

Keeping up with demand



- SP traffic demand goes up
- Inter-node links become congested
- Add incremental 10GE ports into LAG
- Demands on vendors to provide ever increasing LAG sizes

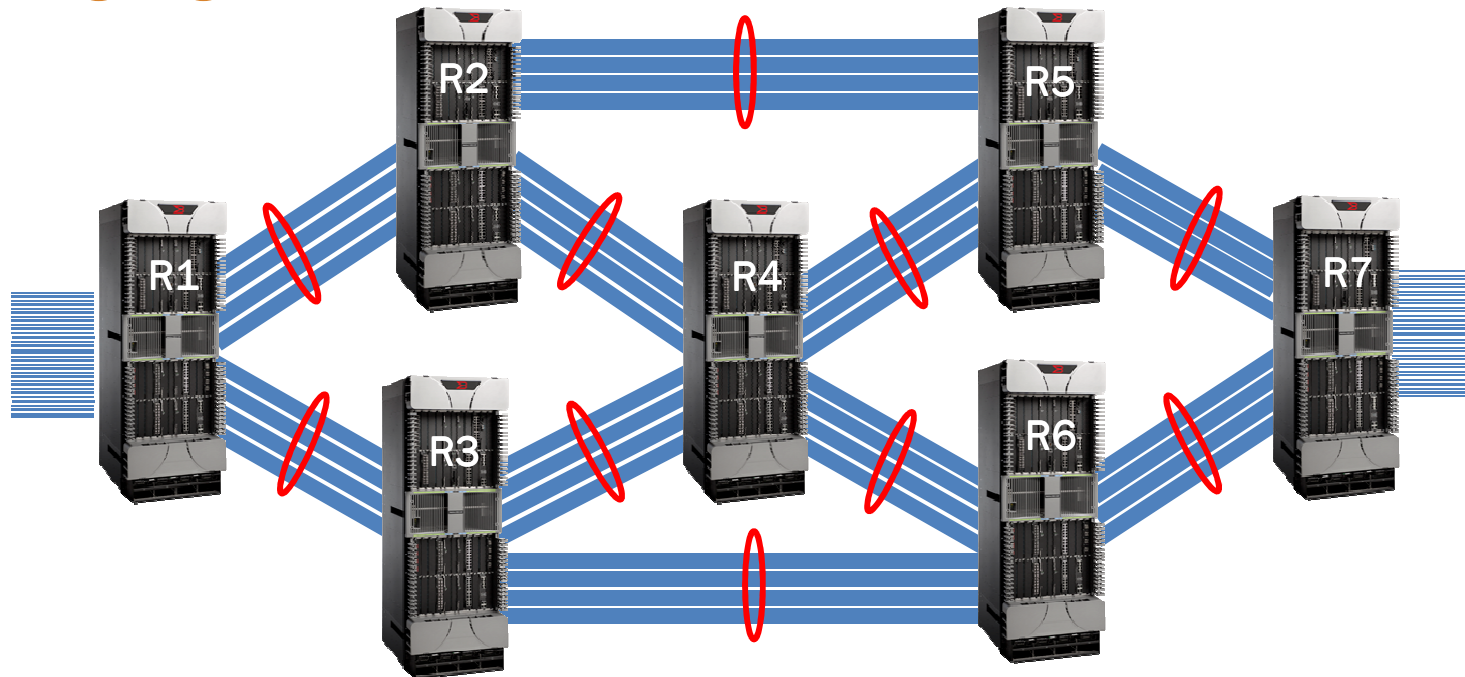
100Gbps in lieu of 10x10Gbps Carrier Trunks



- Fewer links, simpler management, greater spectral efficiency
- Better bandwidth utilisation per slot

400Gbps Inter-Node Carrier Trunks

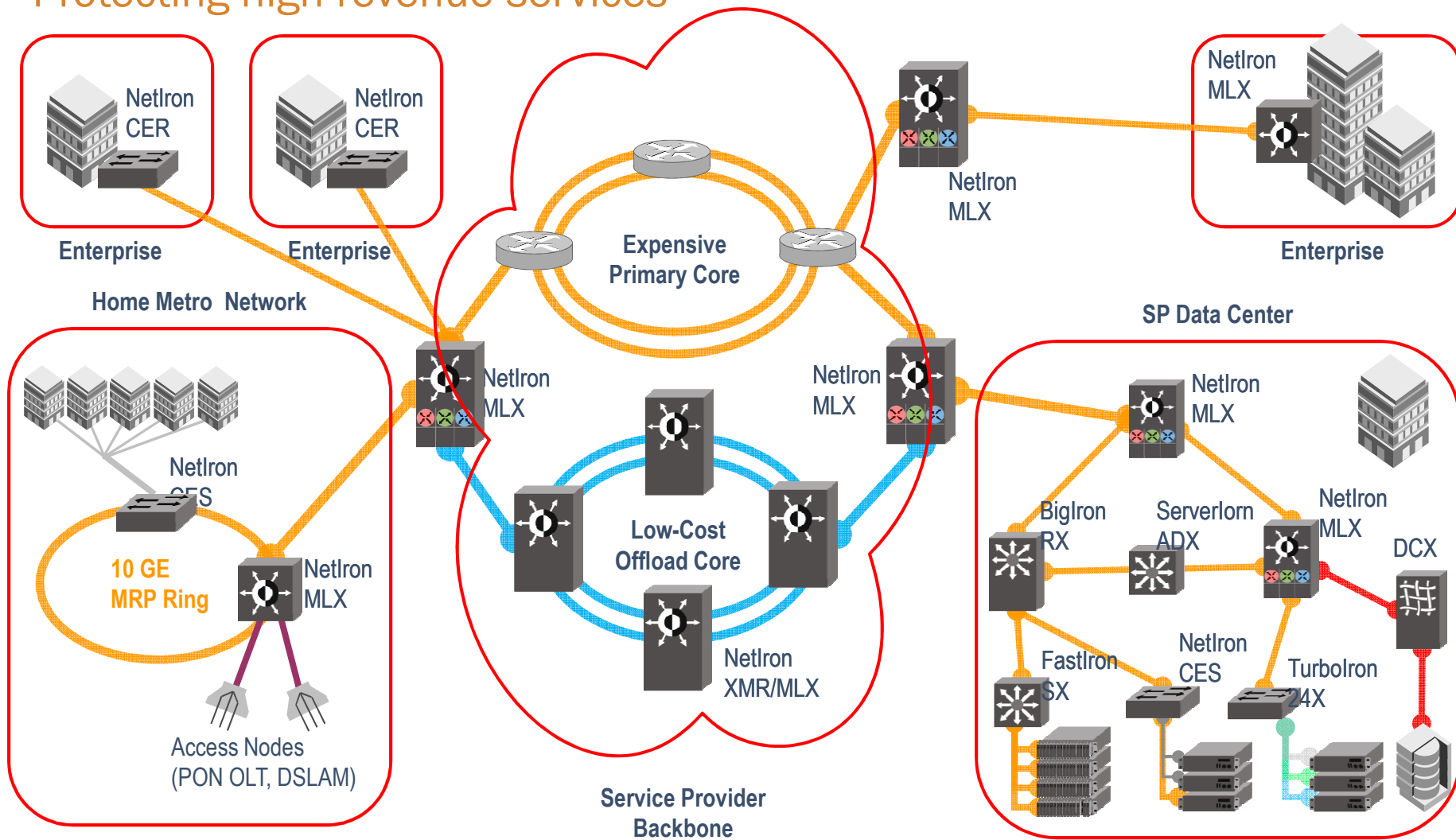
Here we go again...



- The demands for bandwidth are non-linear
- ...so don't get rid of those spare fibres just yet!

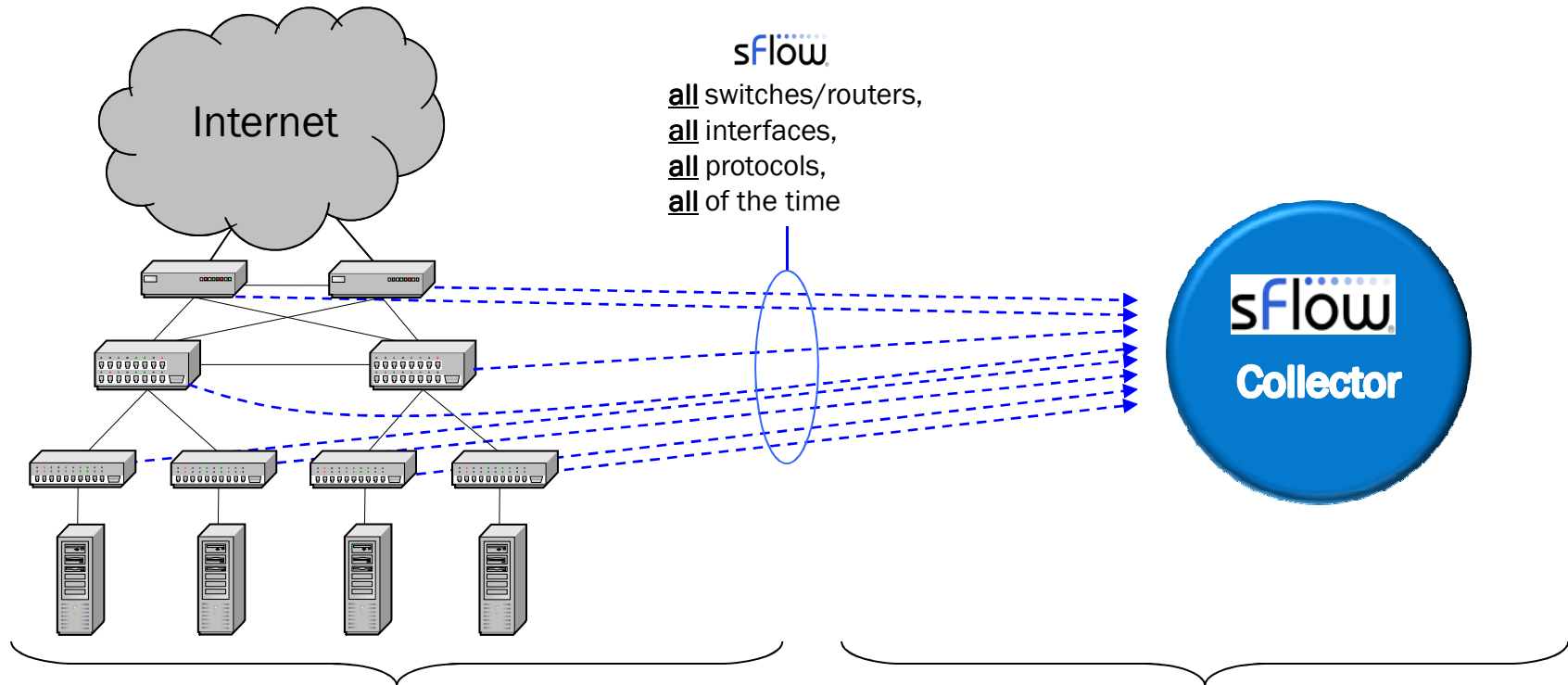
Best-effort Offload

Protecting high revenue services



sFlow Architecture

The network is the probe



Simple Agents

- Easy to implement
- Embedded, wire-speed
- Numerous (every device, every port)
- Packet sampling + counter-push

Smart Collector

- Collects sFlow from all network devices
- Scales to monitor the entire network (>60K interfaces)
- Performs complex analysis
- Alerts on abnormal traffic
- Captures and presents network state and history

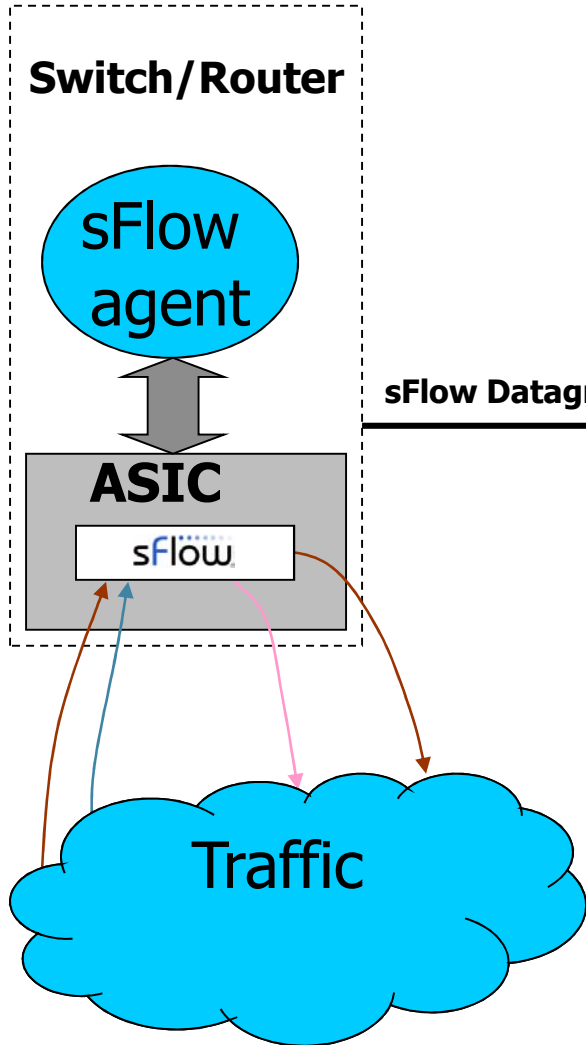


Details

Getting all the information



Detailed L2-L7 data in real-time



- Packet header (e.g. MAC,MPLS, IPv4,IPv6,TCP,UDP, ICMP, FCoE, ARP, STP)
- Sample process parameters (rate, pool etc.)
- Input/output ports
- Priority (802.1p and TOS)
- VLAN (802.1Q)
- Source/destination prefix
- Next hop address
- Source AS, Source Peer AS
- Destination AS Path
- Communities, local preference
- User IDs (TACACS/RADIUS) for source/destination
- URL associated with source/destination
- Interface counters





But is it fast enough?

Standard process is slowing down

Start now to get 400GE by 2016

- Why not just use 32 x 10GE?



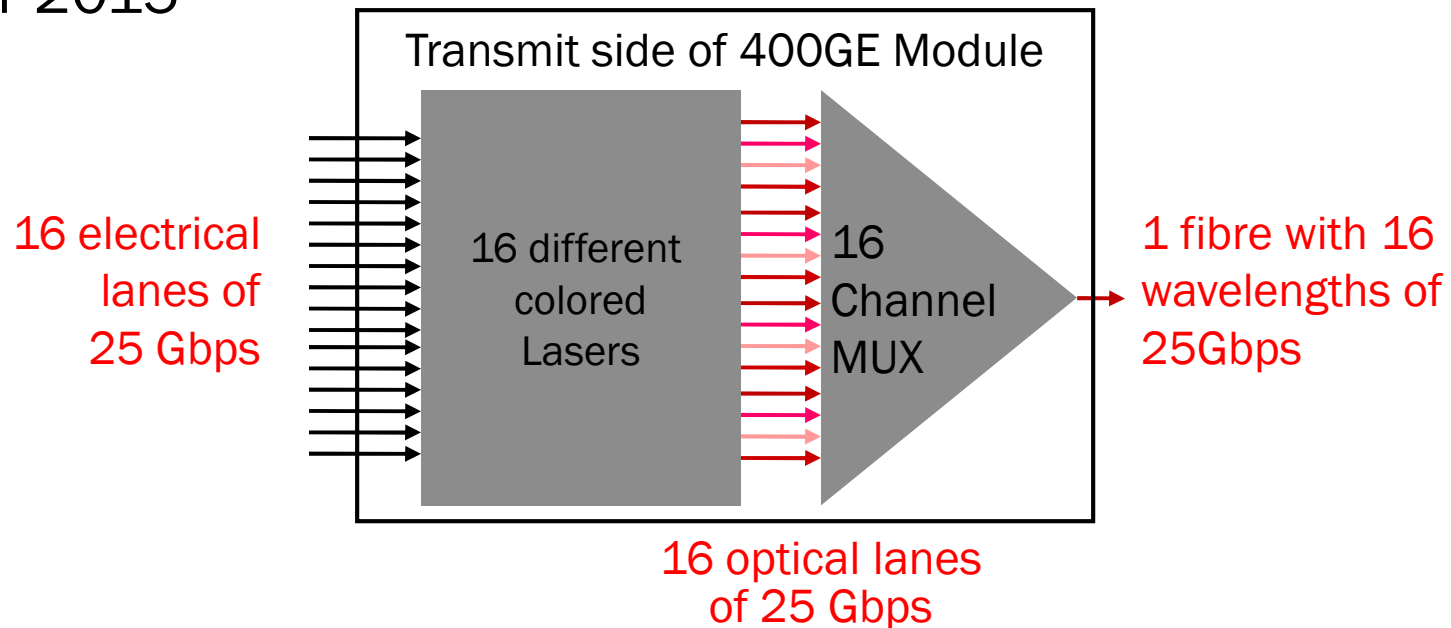
Beyond 100GE

- IEEE 802.3 is not standardizing anything beyond 100GE yet
- IEEE is expecting several project proposals to improve 100GE
 - 100GE over the Backplane
 - 100GE Twinax (This is likely to fail since there isn't a 10GE Twinax standard)
 - Energy Efficient Ethernet for 10/40/100GE Optical
 - 300 meter multimode 100GE fibre solutions
 - 80km single-mode fibre solutions (note there isn't a 10GE 80 km standard)
 - 4x25Gbps Multimode Solutions (only one vendor making 25GE VCSELs now)
- These 100GE projects will keep 802.3 busy for another couple of years
 - IEEE can only work on several projects at a time
- Ethernet will probably begin defining the next speed in 2012 or later
 - The next Ethernet speed will not be standardized for at least 5-6 years



The 400GE Module

- The 400GE module will be 16 channels wide and be larger than the current CFP
 - The module would probably dissipate over 40 Watts
- 40 Gbps serial lanes aren't going to be economical until after 2015



WDM Solutions Available Now

- 10GE CWDM SFP+ solutions are available soon
 - 16 channels for 40 km
 - 8 channels for 70 km
 - 20 nm spacing between wavelengths (1470-1610nm)
- 10GE DWDM SFP+ solutions are available next year
 - 32 channels for 160 km (1530-1560nm)
 - 64 channels for 160 km in 2012
 - ITU 100GHz spacing or about 0.8 nm between wavelengths
- That's 320 Gbps in 2011 and 640 Gbps in 2012!
 - Or wait until 2016+ for 400GE





Summary

100GE is here but not completely ready

Be mindful of the options – it has its place but is not the only game in town



So where does that leave us with 802.3ba

What have we got and what should we do with it?

- The standard is written, but there is still a way to go to implement this in a cost effective, practicable and scalable manner
- There are some clear use cases for 100GE
 - IX and IP transit for reducing large LAG groups
 - Lower cost alternative for best effort ISP traffic in a dual core
 - Provided we can get below 6 x 10GE
- Ethernet is continuing to evolve but it will be a while before the next speed jump
 - Work out what you can do with your current infrastructure
 - Mapping n x 10GE onto your DWDM transport
- Consider the whole cost of a link when comparing prices of n x 10 GE LAG vs 40 or 100 GE
 - Router line cards, ports and optics
 - Optical and transport gear
- You have options – tailor them to meet your needs!





Questions?





Backup

Additional reference information



802.3ba Nomenclature Suffix Summary

Speed	Medium		Coding Scheme	Lanes
	Copper	fibre		
40G = 40 Gbps 100G = 100 Gbps	K = Backplane C = Copper	S = Short Reach (100 m) L = Long Reach (10 km) E = Extended Long Reach (40 km)	R = scRambled 64/66B Encoding	n = Number of Lanes or Wavelengths N = 1 is not required as serial is implied.

Example: 100GBASE-ER4



40 GE QSFP Modules

Quad Small Form-factor Pluggable

- Created for high density short reach interfaces
 - Targeted for data center applications
 - Small compact form factor enables high density but limits distance
- Preferred optical module for 40 GE because of small form factor and cost
- Used for a variety of Ethernet and InfiniBand applications including 40GBASE-CR4 and 40GBASE-SR4
 - 4 bidirectional channels
 - Low power consumption
- Specifications defined to support 40GBASE-LR4 but QSFP not expected until 2011-2012
- Same faceplate size as an XFP but slightly shorter (8.5 mm high x 18.35 mm wide x 52.4 mm long)



Finisar quadwire MMF QSFP

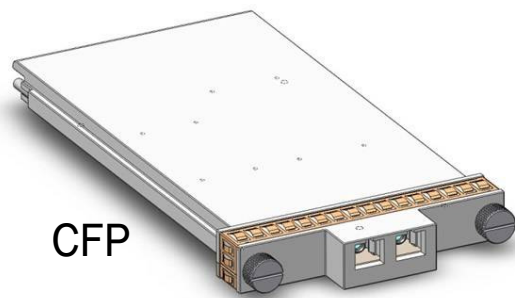


Mellanox Twinax Copper QSFP

100GE Module Design


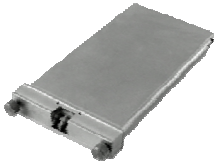






Two generations of 100GE are expected to take 5 years

	1 st Generation	2 nd Generation	3 rd Generation
Optical Module	CFP	QSFP2, CFP2	?
Electrical / Optical Interface	CAUI	Based on OIF's CEI-28G-VSR	?
Electrical Interface	10 lanes at 10 Gbps	4 lanes at 25Gbps	?
Release Date	2009	2014	?



100 GE 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



Part Codes and Availability

Product	Description
BR-MLX-100Gx2-X	2-port 100GE for MLX and XMR platforms. Requires MLX-e chassis.
BR-MLX-100Gx1-X	1-port 100GE for MLX and XMR platforms
BR-MLX-100G-Port-License	100GE 2 nd port License

- Ports on Demand (PoD) license:
 - Facilitates lower cost of adoption
 - Ability to enable second port dynamically as demand for capacity grows
- Availability:
 - General availability in calendar Q2 2011
 - Beta and early access starting end of calendar Q4 2010



Supported Reach

Brocade's 2-port 100GE Module

Physical Layer Reach	100 m OM3, 125 m OM4 MMF	10 km SMF	40 km SMF
100 Gigabit Ethernet (per IEEE 802.3ba)			
Name	100GBASE-SR10	100GBASE-LR4	100GBASE-ER4
Signaling	10 x 10 Gbps	4 x 25 Gbps	4 x 25 Gbps
Media	MPO MMF	Duplex SMF	Duplex SMF
Module/Connector	CFP Module	CFP Module	CFP Module
Availability	At GA	At GA	Post-GA
Wavelengths used (nm)	850 nm (10 pairs of fibre)	1294.53 to 1296.59 1299.02 to 1301.09 1303.54 to 1305.63 1308.09 to 1310.19	1294.53 to 1296.59 1299.02 to 1301.09 1303.54 to 1305.63 1308.09 to 1310.19



sFlow replaces counter polling

- sFlow agent automatically pushes full set of SNMP ifTable counters
- Compared to SNMP polling, counter push results in 10-20x fewer packets on network, reduces CPU load on switch and on network management software
 - XDR* is easier to encode/decode than ASN.1 used by SNMP
 - Counter push is not synchronised between devices
- Single sFlow collector can easily monitor 200,000 switch ports with 1 minute granularity. SNMP polling with 5 minute granularity requires 5-10 collectors.

*XDR (RFC 1832) is a standard for describing and encoding data transferred between systems with different architectures



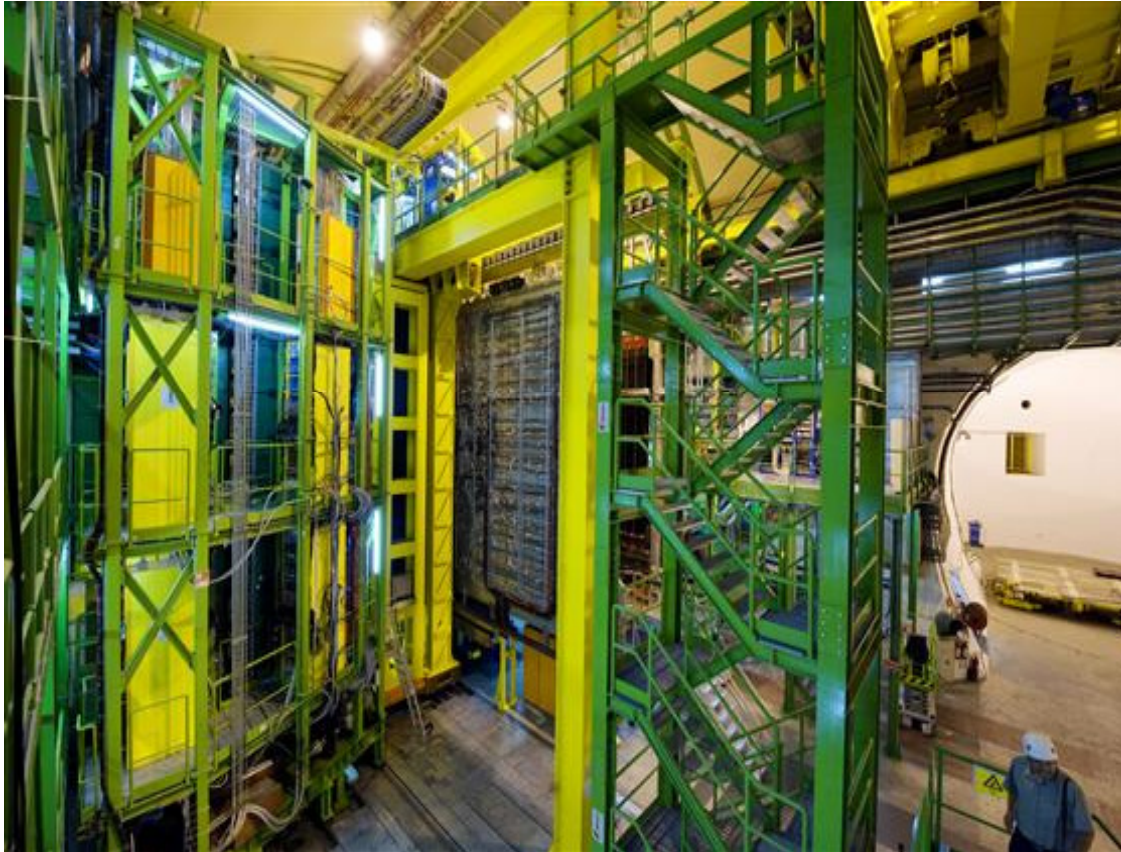
Two types of measurement that are scalable with known accuracy

- Periodic sampling of counters
 - Counting is fast, hardware supports counting, most systems count events, transactions, errors etc.
- Statistical sampling of packets
 - A variant on packet counting, count down to zero, capture the packet, reset the counter with a new random number
 - Why are these mechanisms scalable?
 1. They require minimal, fixed size state (just a block of counters per node). Total state space grows linearly with number of nodes.
 2. Very few operations required, easy to implement in hardware, very small impact when implemented in software
 3. Asynchronous, easily implemented without synchronization or locking mechanisms on: multi-port, multi-module, multi-thread, multi-core devices etc
 - Accuracy
 1. Not 100% accurate but sufficiently accurate for many applications including billing
 2. Sampling accuracy determined by number of samples, not total population (<http://blog.sflow.com/2009/05/scalability-and-accuracy-of-packet.html>)



Real world example: CERN

Large Hadron Collider



- High speed switched network used to collect measurements from the experiment and control the experiment
- Sophisticated monitoring of the network is essential for successful operation of the experiments
- CERN uses sFlow because of its scalability

"Because there are so many ports in the core switches, the SNMP query of interface counters takes a long time and occupies a lot CPU and memory resource."

CERN Investigation of Network Behaviour and Anomaly Detection (CINBAD)

Real world example

- "CERN's campus network has more than 50,000 active user devices interconnected by 10,000 km of cables and fibres, with more than 2500 switches and routers. The potential 4.8 Tbps throughput within the network core and 140 Gbps connectivity to external networks offers countless possibilities to different network applications."
- "Even in CERN 'academic' environment, we can not afford network downtimes, especially when LHC starts to produce peta bytes of data."
- "To acquire knowledge about the network status and behaviour, CINBAD collects and analyses data from numerous sources. A naive approach might be to look at all of the packets flying over the CERN network. However, if we did this we would need to analyse even more data than the LHC could generate. The LHC data are only a subset of the total data crossing via these links."
- "CINBAD overcomes this issue by applying statistical analysis and using sFlow, a technology for monitoring high-speed switched networks that provides randomly sampled packets from the network traffic."

