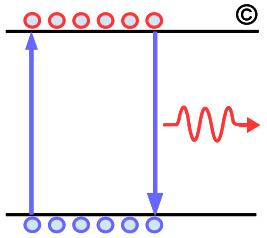


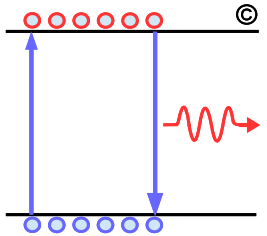
Architectural Consideration for 100 Gb/s/lane Systems

Ali Ghiasi
Ghiasi Quantum LLC

IEEE Meeting
Geneva

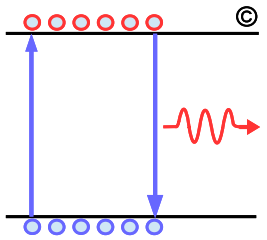
January 25, 2018

Overview

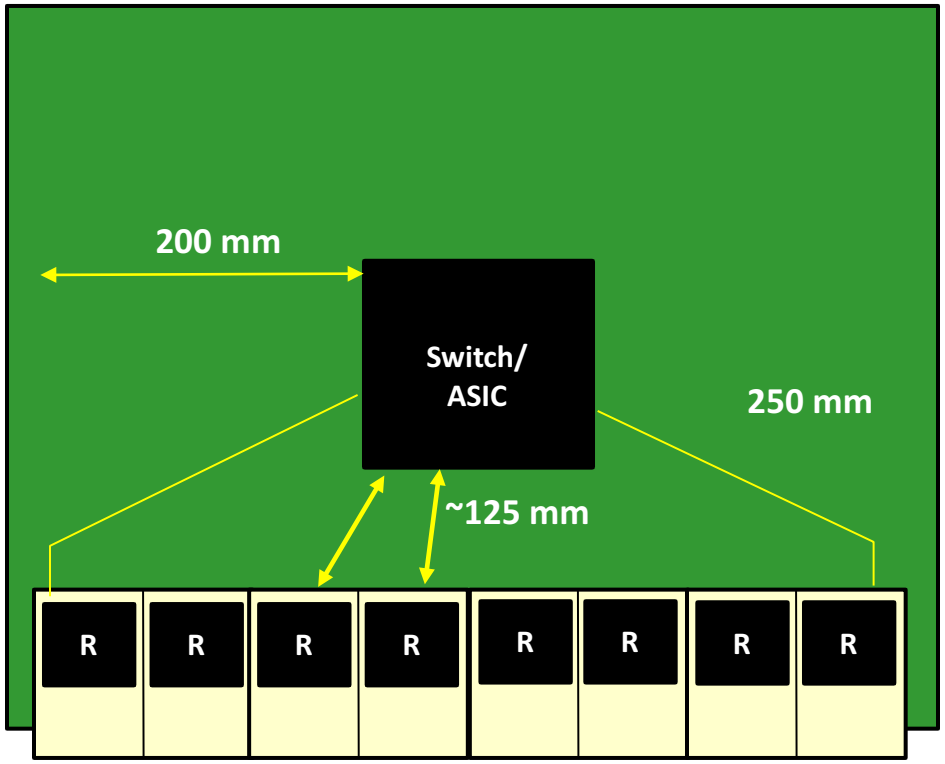


- **Since 10GBASE-KR superset ASIC SerDes have supported C2M, C2C, Cu Cable, and backplanes**
  - With power are area premium small at 10/25/50G ASIC SerDes build at Swiss army knife
- **With switch radix increasing to 256 and at 112G we should not assume every ASIC will implement KR/CR capability due to power and area penalty**
- **Expect 112G signaling USR, XSR, VSR, and MR to be based on PAM4 to maintain compatibility with 100GBASE-DR and 400GBASE-DR4 based on RS (544,514) FEC**
  - Higher gain – higher latency FEC may not meet intra-system latency requirements
  - Considering eco-system requirement this contribution only considers PAM4 with KP4 FEC for 112G applications.
- **Some have voiced support to preserve a very short passive Cu even as short as 1 m**
  - Supporting passive Cu cable require higher power host SerDes capable 35+ dB bump-bump
  - Supporting passive Cu cable may also require placing host ASIC close to the cage and use retimers, and/or use Flyover cable – just to support 1 m Cu cable may not justify on large switch ASIC
  - With switch radix increasing to 256 passive Cu DAC not longer meets server to 1st switch distance requirement
  - A host with 250 mm PCB and a loss of ~ 16 dB can support C2M applications but not Cu passive cable
- **A high radix 256 port switch even 2 m is too short with significant power added SerDes power, instead should consider**
  - Define host Type I - C2M loss is 16 dB so practical PCB can be constructed without extra retimers
  - Define host Type II - C2M loss limited to 10 dB so ~2 m Cu cable is supported but may require retimers
  - Both host types support AOC/Optics but only host type II supports Cu Cables
- **One concern raised is the added PD in the module CDR equalizing 16 dB channel, but given at 112G loss is our friend, a less reflective 16 dB channel may not be more challenging than a 10 dB more reflective channel**
  - At 112G reflectance and ILD are the greatest challenge for C2M and C2C applications
  - Need to use COM analysis for C2M to trade-off loss, return loss, and ILD.

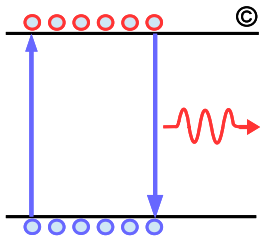
C2M Applications



- **Numerous study in IEEE and OIF have shown typical line card require about 250 mm host traces**
  - CAUI-4 loss budget is 10.2 dB supporting ~125 mm on mid-grade PCB material like Isola 408HR
  - Most line card implementation prefer not to use retimer to save power and instead use Megtron 6 like material to extend CAUI-4 PCB reach to ~250 mm
  - A C2M channel supporting ~125 mm by assuming best PCB material like Megtron 7 or Tacyhon 100 would not meet C2M applications
  - C2M applications need to support at least 200 m on PCB.



C2M Channel Reach



- **PCB loss estimate assumptions and tools for calculation**
  - Rogers Corp impedance calculator (free download but require registration) https://www.rogerscorp.com/acm/technology/index.aspx
  - The IEEE tool if updated could be another option to estimate channel reach http://www.ieee802.org/3/bj/public/tools/Reference__DkDf_AlegbraicModel_v2.04.pdf
  - Stripline ~ 50 Ω, trace width is 5.5 mils, and with ½ oz Cu
  - Isola 408HR DK=3.65, DF=0.0095, RO=2.5 um, Meg-6 DK=3.4, DF=0.005, RO 1.2 µm, Tachyon100 DK=3.02, DF=0.0021, RO=1.2 µm
  - To support equivalent PCB traces for C2M need at least 16 dB end-end channel loss.

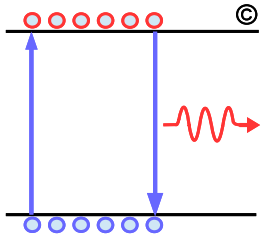
Host Trace Length (in)	Total Loss (dB)	Host Loss(dB)	Isola 408HR	Megtron 6	Tachyhon100
Nominal PCB Loss/in at 5.15 GHz	N/A	N/A	0.65	0.52	0.46
Nominal PCB Loss/in at 13 GHz	N/A	N/A	1.27	0.98	0.83
Nominal PCB Loss/in at 27 GHz	N/A	N/A	2.18	1.60	1.28
28G-VSR with one connector & HCB*	10.5	6.81	5.4	6.9	8.2
Current 112G-VSR draft+one connector & HCB**	13.5	8.5	3.9	5.3	6.6
112G-VSR with one connector & HCB**	16	11	5.0	6.9	8.6

Reach Inches Too Short

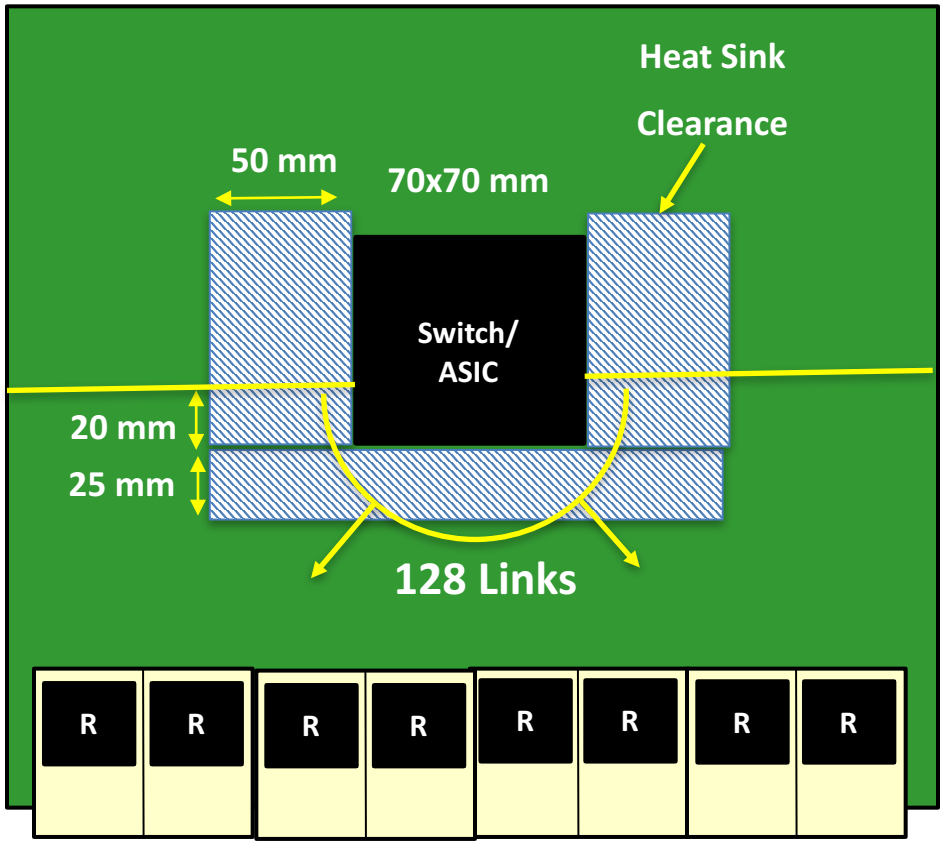
* Assumes connector loss is 1.69 dB and HCB loss is 2.0 dB at 12.89 GHz

** Assumes connector loss is 2.5 dB and HCB loss also 2.5 dB at 27 GHz.

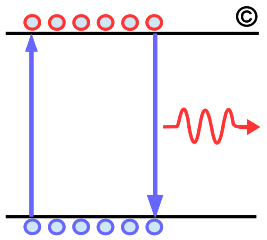
Thought on Flyover Cable



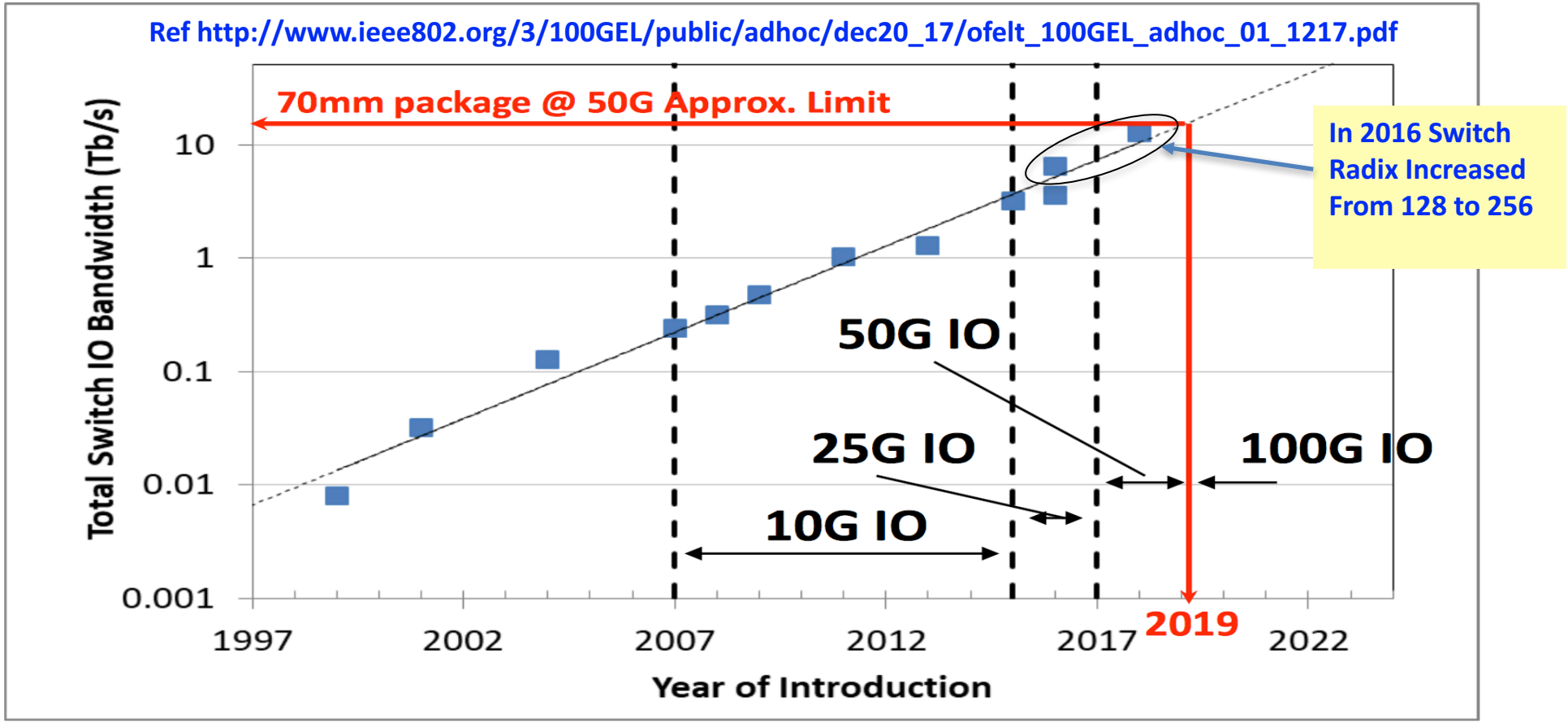
- **Interesting!**
  - May work OK for server or low-medium density applications
  - May add additional discontinuity due to Flyover connector
  - Flyover connectors may have higher ILD and worse RL due package-connector cascaded discontinuity and low loss
  - High density application may require upward of 100 mm PCB trace to break out to mount Flyover connectors and clear the heat sink
  - A reasonable question to ask: what would be be the maximum PCB trace to route 128 links (512 twin-ax cables) for the application shown?



Switch Evolution Trend

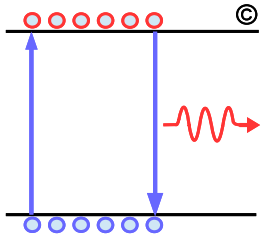


- **Since 2106 several Ethernet switches with radix of 256 have been introduced**
  - 256x50G recently announced and expect 256x100G in ~2 years
  - Single Ethernet switch ASIC is too large for one rack of servers.



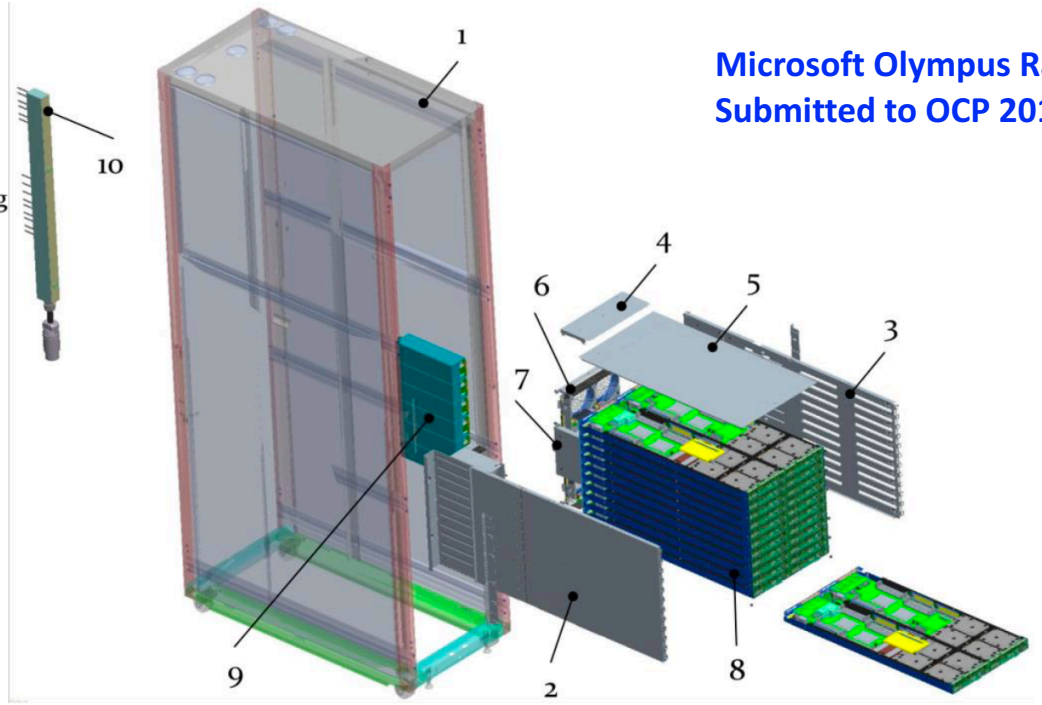
> 2019 ASIC requirements are expected to exceed BW delivered by a conventional BGA with 50G IO

Example Server Rack and TOR

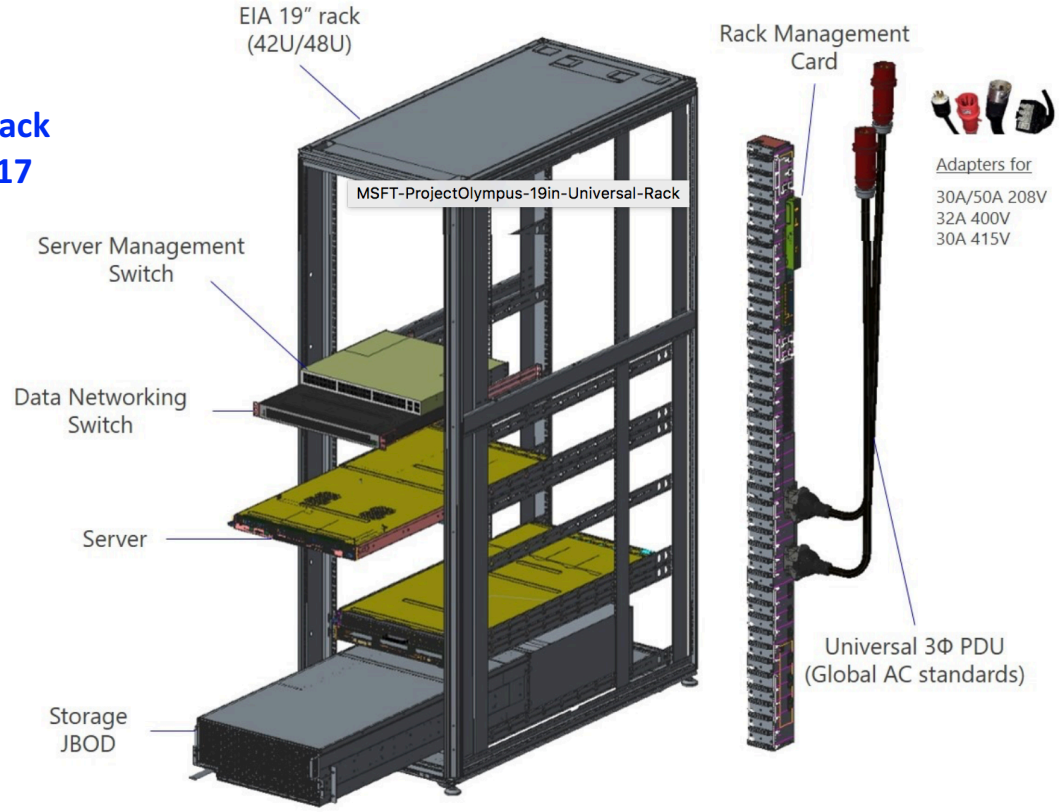


- **A decade ago half-width servers with 96 servers in a rack were common**
- **Today common server rack implementation only have 24-48 servers as result of**
  - Larger CPUs with more cores/memory and racks having JBOD, JBOF, and GPU.

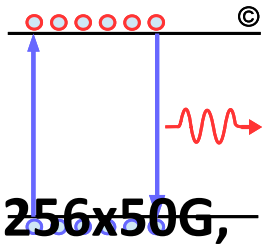
1. Rack
2. Left Side Wall (including Power Dist Board)
3. Right Side Wall
4. Rear Covers
5. Top Cover
6. Fan Door
7. Chassis Manager
8. Trays
9. Power Supplies
10. Power Distribution Unit



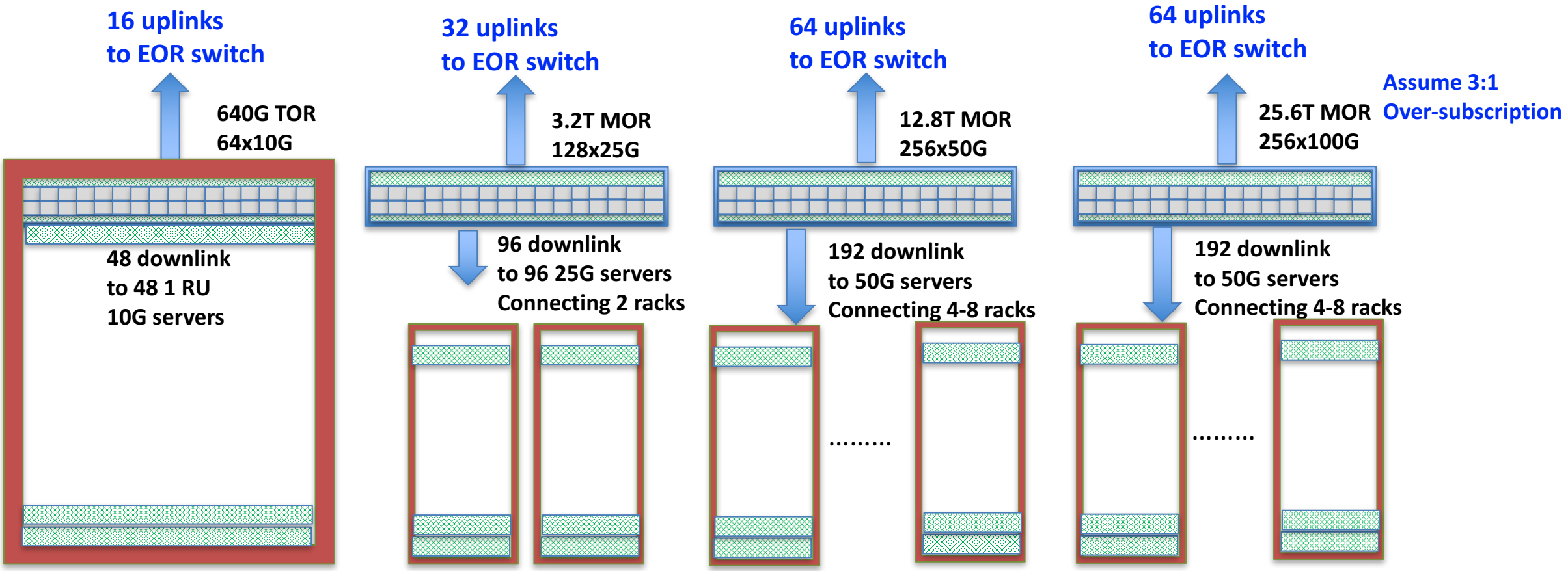
Microsoft Olympus Rack
Submitted to OCP 2017



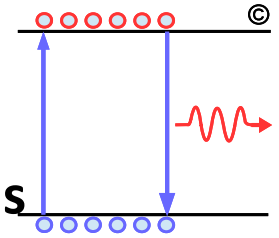
Datacenter Trends



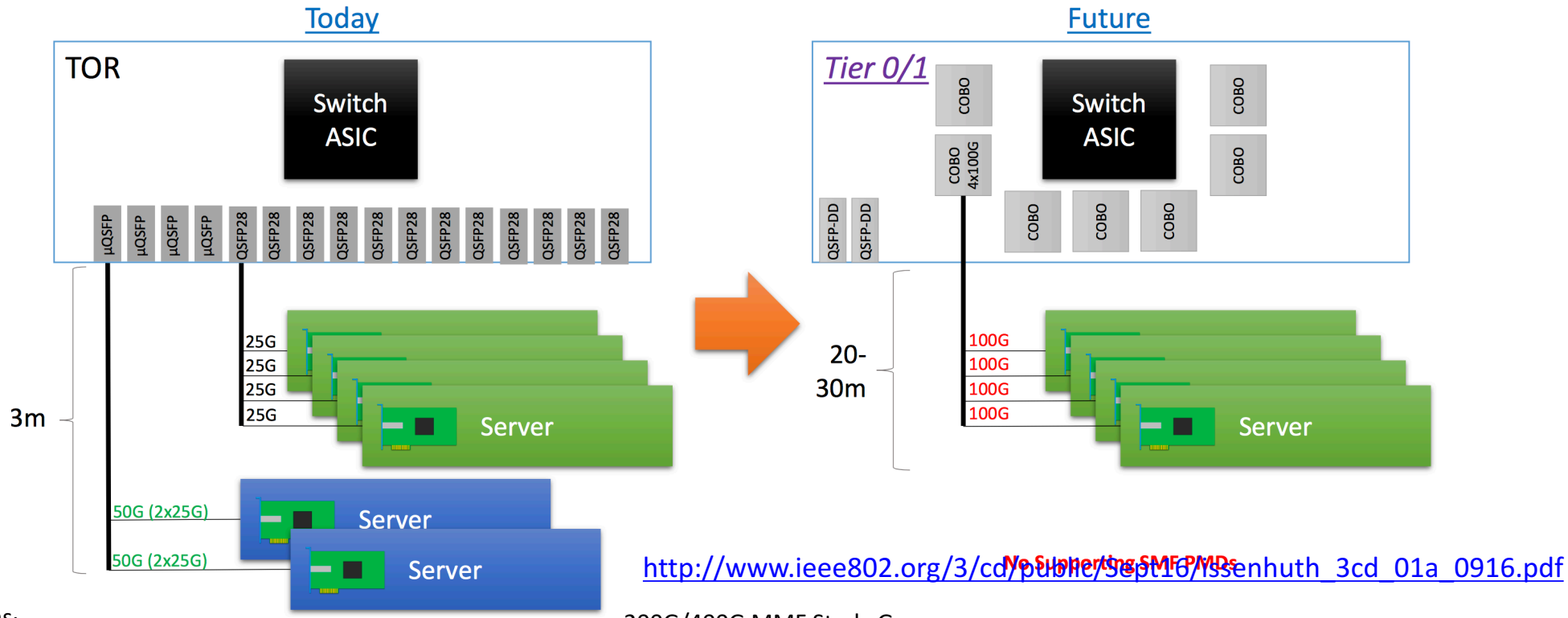
- **Switch radix over the last 9 years has increased from 64x10G, 128x25G, now to 256x50G, and likely to 256x100G by 2019/2020**
  - To mitigate full rack failure dual MOR switches may connect to each rack.



Emerging Trend: Server Connecting to MOR Switch



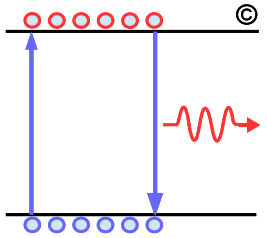
- **Microsoft evolution showing server directly connecting to MOR/Tier 0/1 switches as result of switch radix increase from 128 to 256 and fewer servers in a rack**
  - Passive Cu cable with reach limited to 1 m or even 2 m at 100 Gb/s/lane not very useful
  - We need to be responsive to emerging trend and not burden the system with Cu cable when the attach rate expected to be low and forces an impractical low loss host PCB!



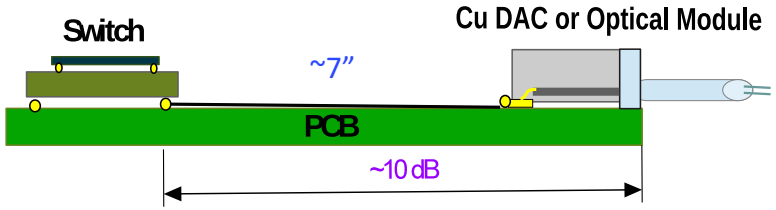
http://www.ieee802.org/3/cd/public/Sept16/issenhuth_3cd_01a_0916.pdf

No Supporting SMF PMDs

Evolution of Front Panel Ports

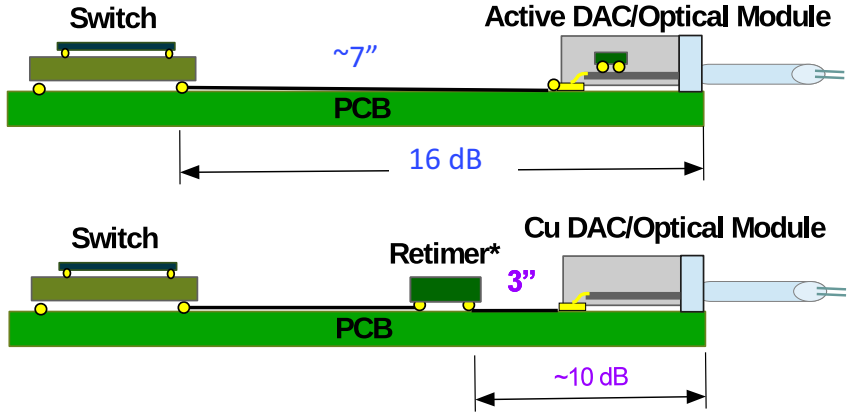


Pluggable at 25 Gb/s and 50 Gb/s



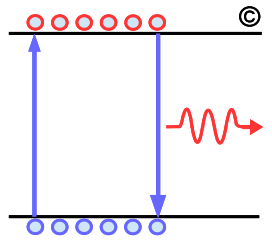
- PHY less design – what we are used to
  - Supports passive Cu DAC
  - Switch directly drives optical modules
  - Switch directly drives 3 m of Cu DAC
  - Offers optimum power and cost.

Pluggable at 100 Gb/s

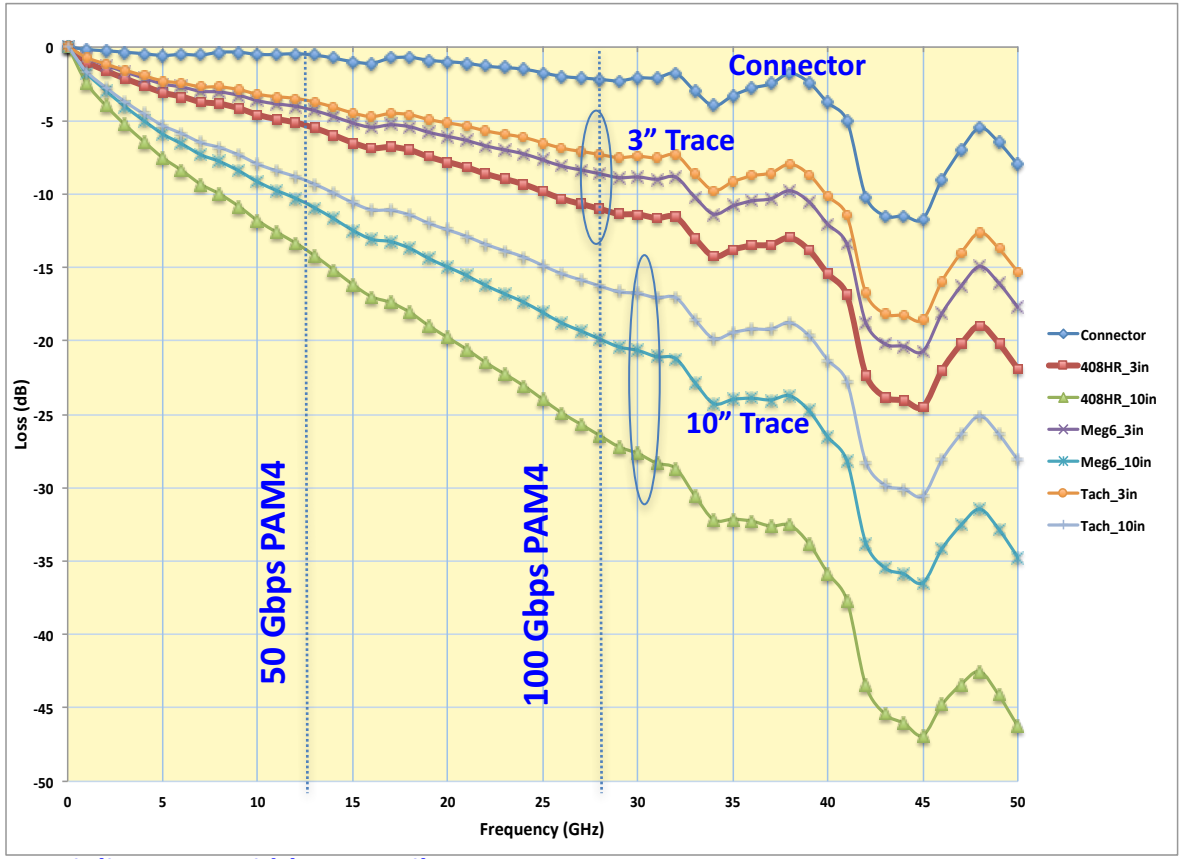


- Option I – PHYless Design – Channel loss 16 dB
  - Supports AOC, Active DAC, and Optics
  - Doesn’t support passive Cu DAC
  - 16 dB loss up to 250 mm PCB traces on premium material such as Megtron 7/Tachyon PCB
  - Offers improve power and cost
  - Better choice for MOR/Spine switches
- Option II – Require PHY – Channel loss 10 dB
  - Supports passive Cu DAC, active DAC, AOC, and optics
  - 10 dB loss supports 100 mm PCB traces on premium material such as Megtron 7/Tachyon
    - 10 dB loss can be allocated to use Flyover to extend the reachs
  - Retimer adds cost and power
  - Viable option for low radix switch's/TOR

112G VSR Channels

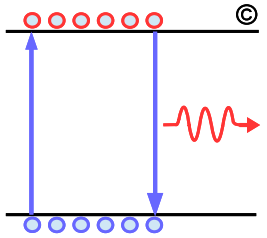


- **Connector assumed is Yamachi CFP2 which is capable of 53 GBd operation other connectors potentially may as well**
  - VSR channel loss investigated with following material 408HR, Megtron 6 HVLP, Tachyon HVLP for 5.5 mil ½ oz stripline
  - End-end channels constructed from from 3” or 10” host PCB traces + CFP2 connector + 1” 408HR (plug)
  - A CTLE receiver is even questionable if it can offer PVT margin at 50G PAM4
    - http://www.ieee802.org/3/bs/public/17_09/lim_3bs_01b_0917.pdf
  - Given that the 112G-VSR receiver will have few FFEs and/or 1-2 tap DFEs need to investigate range of 10-16 dB channels
  - At 112G with PAM4 in many instances few extra dB of loss can dampen resonance effect and ILD
  - It is time we move away from simple channel loss and RL to COM like tool to allow trading-off loss, return loss, and ILD.



Stripline trace width = 5.5 mils
Meg6 DK=3.4, DF=0.005, Ro = 1.2 µm
Tachyon 100 DK=3.02, DF=0.0021, Ro= 1.2 µm
Isola 408HR DK=3.65, DF=0.0095, Ro = 2.5 µm

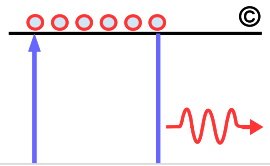
Common Package Build-up Substrate Material



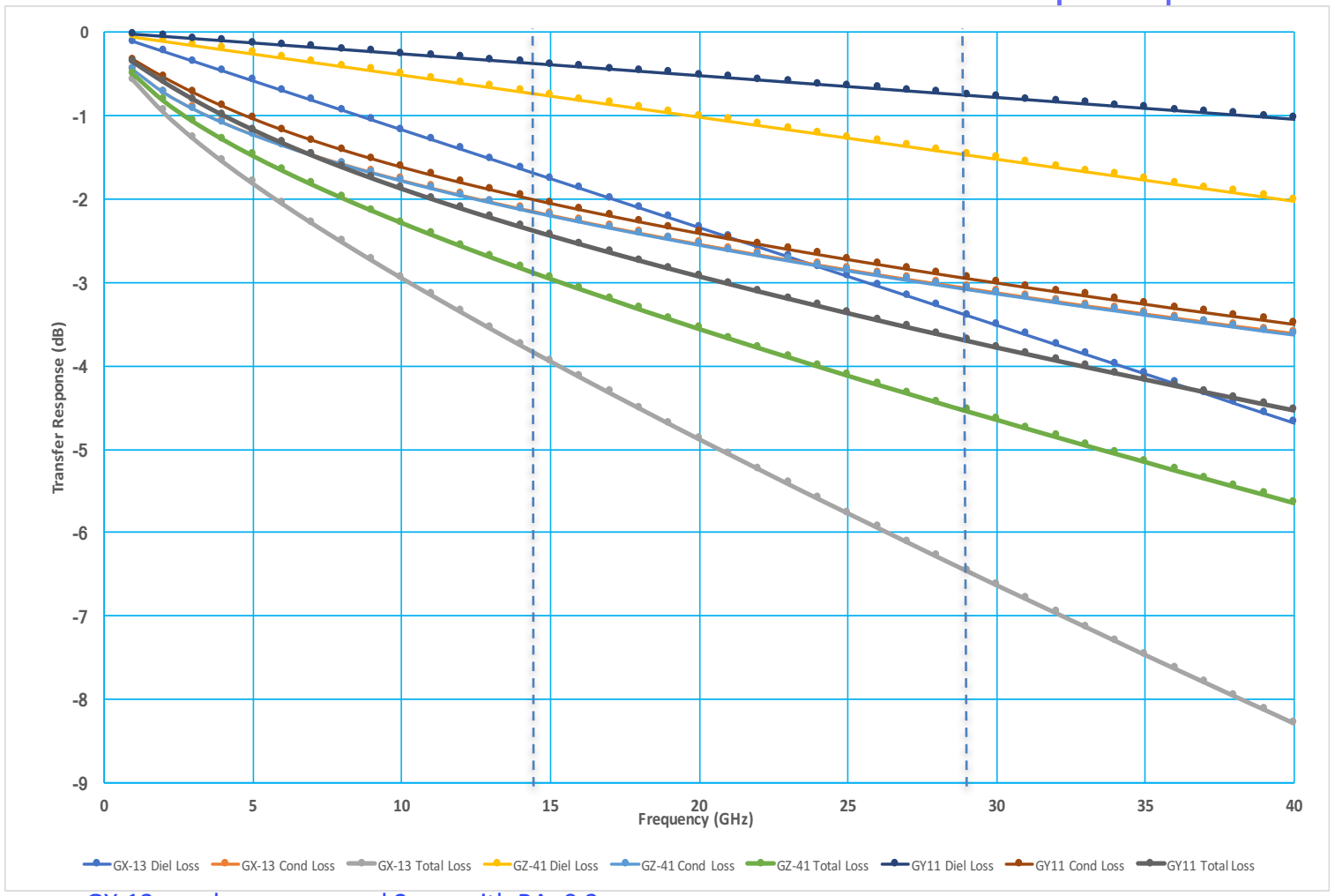
- **Low Df Build-up Material for High Frequency Signal Transmission of Substrates, Hirohisa Narahashi, ECTC 2013.**

ABF	Temp./degC	-10	5	10	25 (r.t.)	40	60	80	100
GX13	Dk	3.14	3.16	3.17	3.21	3.22	3.23	3.21	3.25
	Df	0.0096	0.0158	0.0166	0.0190	0.0204	0.0217	0.0230	0.0238
GZ41	Dk	3.22	3.23	3.24	3.26	3.26	3.26	3.26	3.28
	Df	0.0059	0.0069	0.0072	0.0076	0.0083	0.0090	0.0095	0.0102
GX92	Dk	3.11	3.13	3.13	3.14	3.15	3.15	3.16	3.17
	Df	0.0129	0.0150	0.0155	0.0173	0.0187	0.0194	0.0204	0.0216
GX-T31	Dk	3.17	3.17	3.18	3.24	3.25	3.25	3.26	3.27
	Df	0.0083	0.0094	0.0097	0.0135	0.0141	0.0151	0.0159	0.0169
GY11	Dk	3.06	3.07	3.07	3.10	3.07	3.08	3.10	3.10
	Df	0.0032	0.0035	0.0036	0.0039	0.0040	0.0045	0.0048	0.0054

Package Loss for 100 Gb/s PAM4

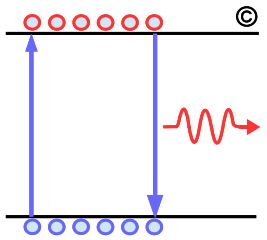


- **Current package loss for large ASIC with 30 mm trace assumed in IEEE COM is 3.0 dB**
  - The IEEE COM package trace likely was based on GX-13 material which will have excessive loss at 53 GBd PAM4
- **Estimated loss shown for 30 mm package trace having a moderate size 36x24 µm stripline trace**
  - Estimated loss for GX-13 is 6.5 dB
  - Estimated loss for GZ-41 is 4.5 dB
  - Estimated loss for GY-11 is 3.5 dB
- **Assuming next generation package substrates material it's reasonable to assume 4 dB substrate loss at 28 GHz for 30 mm trace**
  - ILD effects due to Cp, Cd, and via may add 1-2 dB of ripple!



GX-13 roughness assumed 2 µm with RA=0.3 µm
GZ-41 roughness assumed 2 µm with RA=0.3 µm
GY-11 roughness assumed 1 µm with RA=0.1 µm

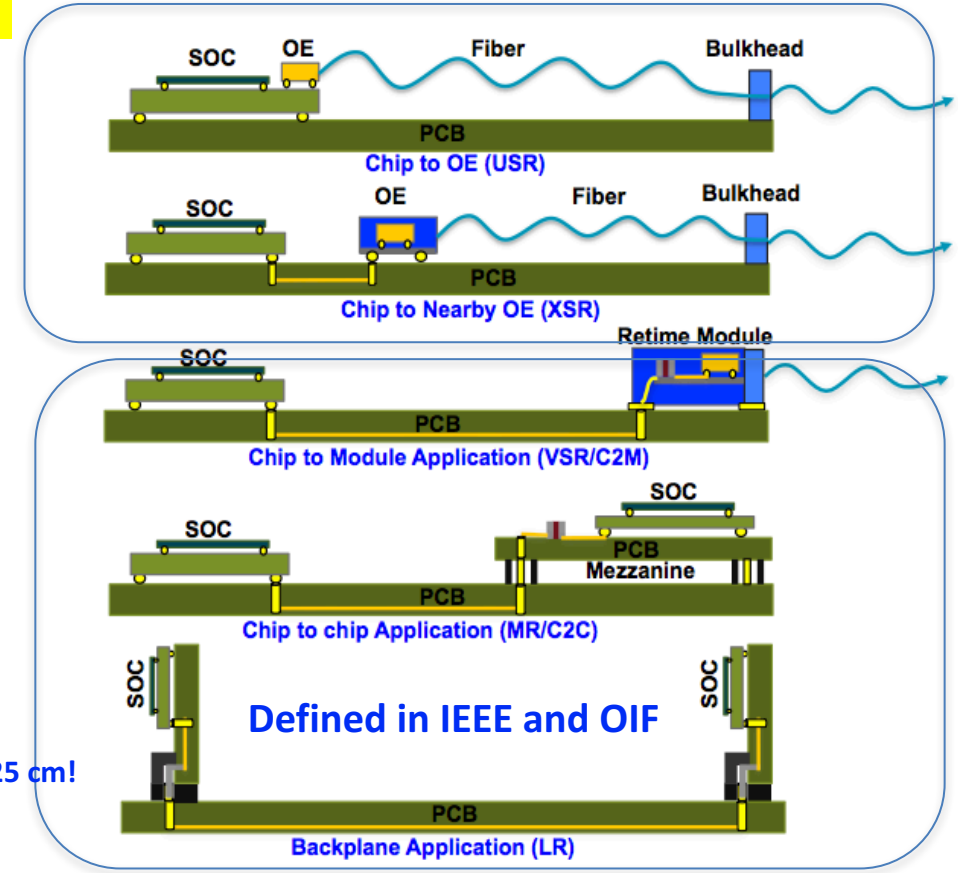
The 50G/lane Interconnect Ecosystems



- **OIF has defined both NRZ and PAM4 for MR, VSR, XSR, and USR**
- **IEEE P802.3bs and P802.3cd are defining PAM4 signaling for 50G/lane Chip-to-chip, chip-to-module, Cu DAC, and backplane**
  - An LR SerDes operating at 29 GBd may have 37 dB of loss from bump to bump!

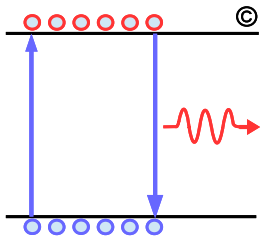
Application	Standard	Modulation	Reach	Loss Ball-ball	Loss Bump-bump
Chip-to-OE (MCM)	OIF-56G-USR	NRZ	< 1cm	2 dB@28 GHz	NA
Chip-to-nearby OE (no connector)	OIF-56G-XSR	NRZ/ PAM4	<7.5 cm[1]	8 dB@28 GHz 4.2 dB@14 GHz	12.2 dB@14 GHz 4.2 dB@14 GHz
Chip-to-module (one connector)	OIF-56G-VSR IEEE CDAUI-8	NRZ/PAM4 PAM4	< 10 cm[2] <20 cm	18 dB@28 GHz 10 dB@13.3 GHz	26 dB@28 GHz 14 dB@13.3 GHz
Chip-to-chip (one connector)	OIF-56G-MR IEEE CDAUI-8	NRZ/PAM4 PAM4	< 50 cm < 50 cm	35.8 dB@28 GHz 20 dB@13.3 GHz	47.8 dB@28 GHz[3] 26 dB@13.3 GHz
Backplane (two connectors)	OIF-56-LR IEEE 200G-KR4	PAM4 PAM4	<100 cm <100 cm	30dB@14.5 GHz 30dB@13.3 GHz	~37dB@14.5 GHz[4] 36dB@13.3 GHz

Defined in OIF



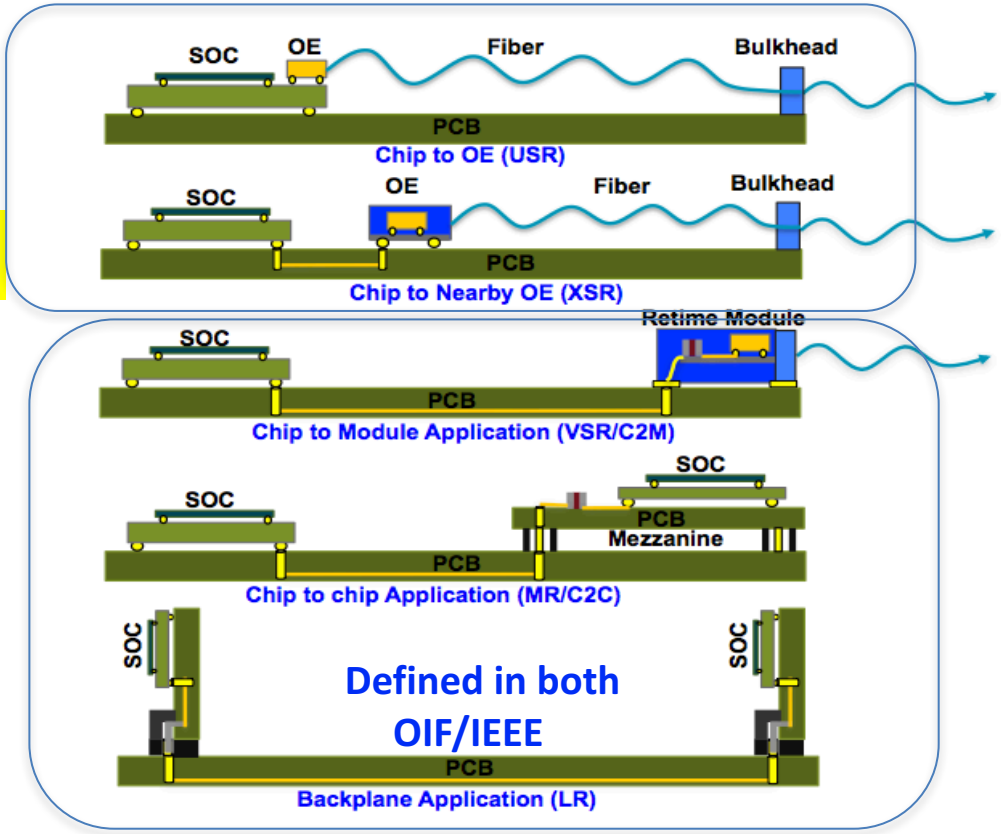
1. OIF XSR definition likely too short for any practical OBO implementation!
2. OIF VSR 10 cm reach assumes 10 cm mid-grade PCB but typical implementation uses Meg6/ Tachyon 100 with ~25 cm!
3. Include 2x6 dB for package loss but 47.8 dB seem beyond equalization capability
4. Include 2x3.5 dB for package loss.

The 100G/lane Eco-System will be follow 50G Eco-system



- **With estimated loss of 18 dB VSR specification is inline with our definition of MR**
  - Bump to bump loss calculated by assuming ASIC package with 4 dB loss and small CDR package having 1.5 dB loss
  - 4 dB ASIC package assumes 30 mm trace and requires material better than GZ41
  - PCB reaches below assumes Tachyon 100/Megtron 7.

OIF has defined USR/XSR but with little traction so far!

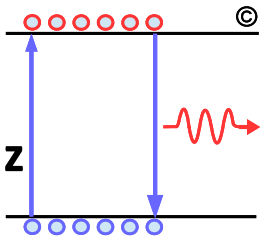


Application	Standard	Modulation	Reach	Ball-Ball Loss	Bump-Bump Loss
Chip-to-OE (MCM)	TBD	PAM4	< 1 cm	NA	2 dB
Chip-to-nearby OE (no connector)	TBD	PAM4	<10 cm*	5 dB	12 dB
Chip-to-module (one connector)	OIF-112G-VSR	PAM4	< 21 cm**	16 dB	21 dB
Chip-to-chip (one connector)	TBD	PAM4	< 39 cm	20 dB	28 dB
Cabled Backplane (two connectors)	TBD	PAM4	<55 cm	28 dB	36 dB

* OBO connector + package assumed having 3 dB loss

** VSR host packaged assumed 4 dB loss and the CDR packaged assumed to be 1 dB.

Summary



- **The 112G PAM4 is uncharted territory need quality measured S-parameters to at least 58 GHz**
  - With a representative connector compatible with SFP56, QSFP56, QSFP-dd, or OSFP
- **Need to consider next generation package material to limit 30 mm trace loss < 4 dB**
- **C2M likely will be the most important application and require ~16 dB loss on Megtron7/Tachyon 100**
  - Biggest challenge for C2M will be resonance effect and ILD in mid-band (5-15 GHz)
  - In many cases having few extra dBs of loss will help C2M pulses response
  - COM like tool will allow to trade-off loss, return loss, and ILD to allow higher loss – less reflective channels
  - Compliance methodology and MCB/HCB could end up to be Achilles' heel
  - We need to look outside the box including considering transmitter training at start up
  - Use of transmitter training and COM will allow to support 16 dB channel with negligible CDR power preimum
- **Key emerging trend in the data center are introduction of 256 radix switches and fewer servers per rack**
  - This trend impacts passive Cu cables broad market potential with 1-2 m reach
  - An MOR/1st layer switch potentially 30 m away Cu cable may not paly a role
- **Lets not sacrifice C2M application by cutting host PCB loss for sake of supporting an impractical 1m reach Cu cable**
  - Assuming Cu cable still has broad market potential it would be better to define 2nd host type with ~ 10 dB loss where the port can support 2 m Cu cable as well as optical PMD/AOC.