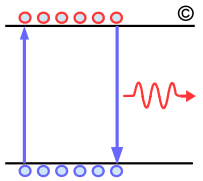


Architectural Consideration for 100 Gb/s/lane Systems

Ali Ghiasi
Ghiasi Quantum

Feng Hong
Huawei

Xinyuan Wang
Huawei

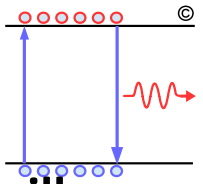
Yu Xu
Huawei

IEEE Meeting
Rosemount

February 7, 2018

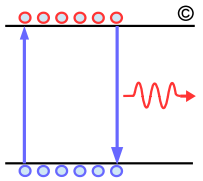
HUAWEI

Overview

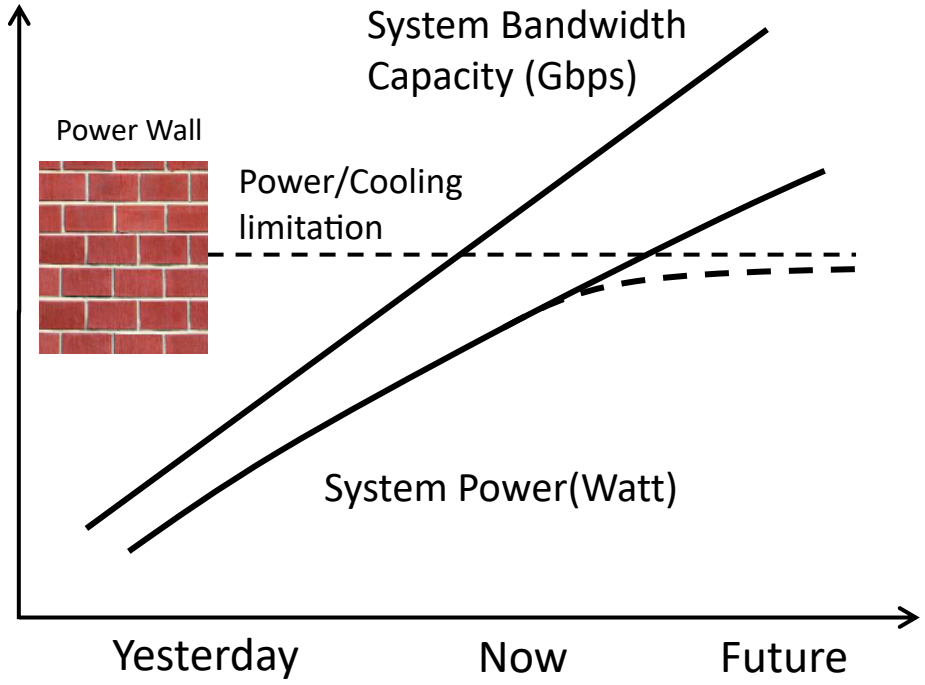


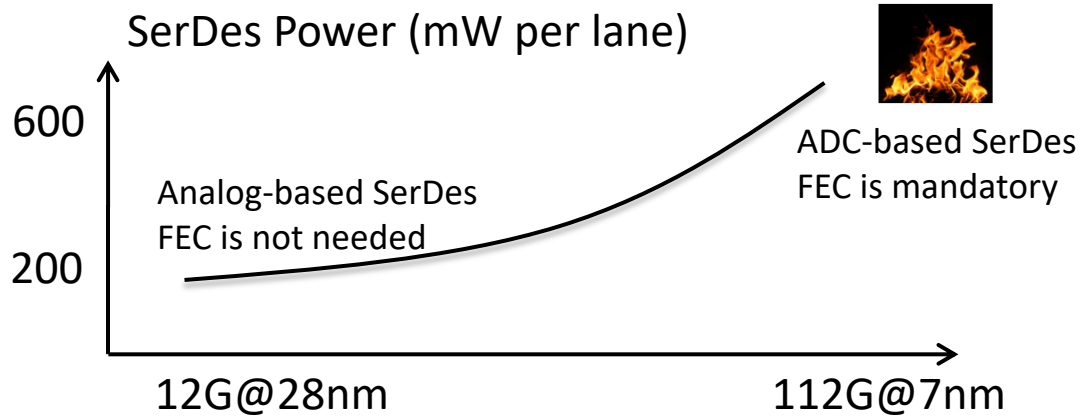
- **High capacity systems based on 112G/lane electrical will test conventional cooling limits and will come at cost premium**
  - 112G/lane electrical is necessary to enable next generation routers and high capacity data center switches
- **The cost benefit of 112G system may only be realized in large scale applications requiring highest capacity**
- **What is most important for initial 112G systems deployment are**
  - C2M supporting at least 200 mm PCB trace
  - C2C supporting at least 400 mm + 1 connector
  - Re-use of RS (544, 514)
- **Study group should also consider defining 0.5 m conventional or 1 m cabled backplane with 25 dB ball-ball or 35 dB bump to bump loss (assuming 5 dB package loss)**
  - Both RS (544, 514) as well as stronger FEC should be studied
- **Study group also may consider Cu cabling solution with following caveats**
  - Cu cabling should not compromise C2M PCB trace length
  - High radix 256 switches significantly reduces 1st switch to server use case given Cu cable reach is <3 m
  - Extra retimers and higher power LR SerDes on the host raises system max operating power
  - Active-Cu/AOC doesn't raise the max system operating power as the retimer in the active-Cu/AOC replaces a higher power SMF module
  - Given the level of support for 2 m Cu DAC one option to explore is asymmetrical link optimized for switch to server without compromising TOR switch PCB reach.

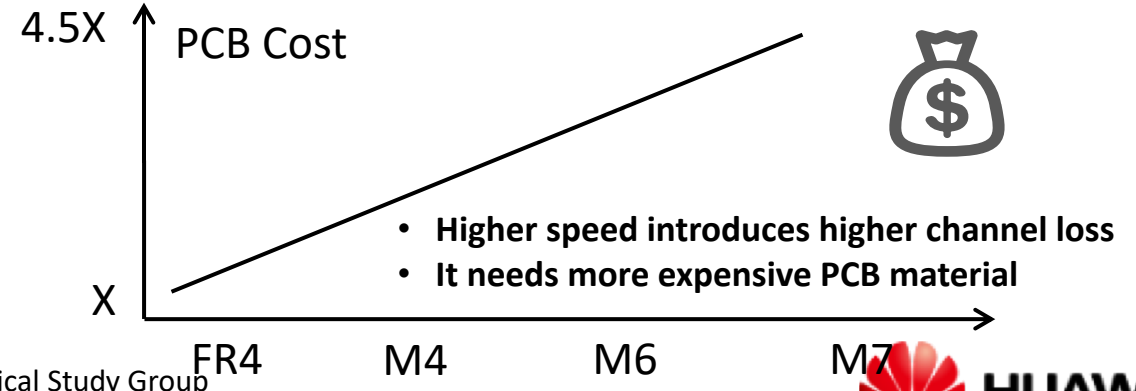
100G/lane System Concerns: Power and Cost Challenges



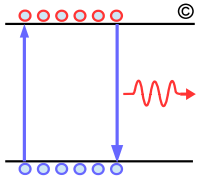
- **Cost/Gb and power/Gb increasing with migration from 25G to 50G and 100G**
  - CMOS technology scaling has slowed down
  - 100G/lane system power may exceed limits of conventional air-cooled
  - 100G/lane is required for next Gen routers and leading edge Hyper-scale but may not be the answer for every data center!



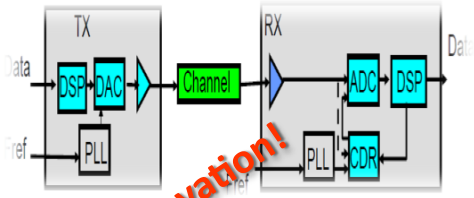




112G Electrical Backplane: Innovations are needed for Both Passive Channel and SerDes

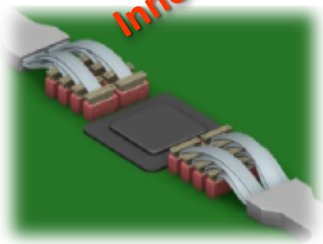


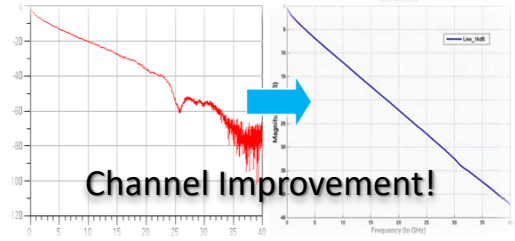
Is Electrical Still A Viable Solution?

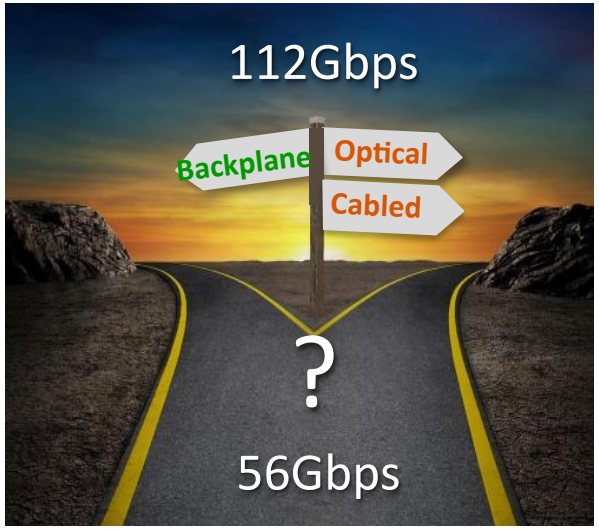


112G backplane

Electrical switch





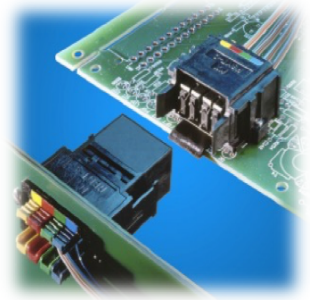


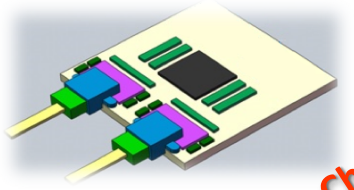
Will Optical Replace Electrical?

Optical Interconnect

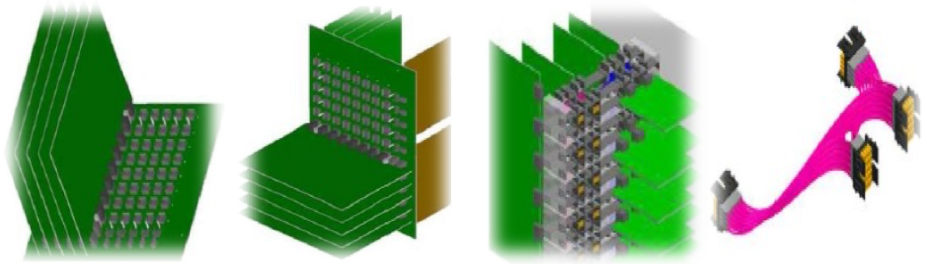
Optical IO / Electrical switch

Optical IO / Optical switch



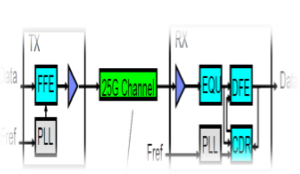


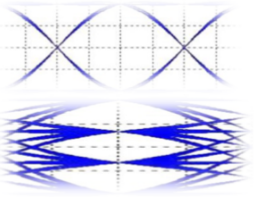
Challenge: COST!

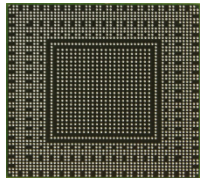


56G backplane

Electrical switch

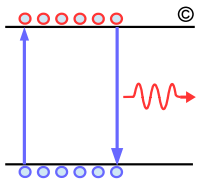




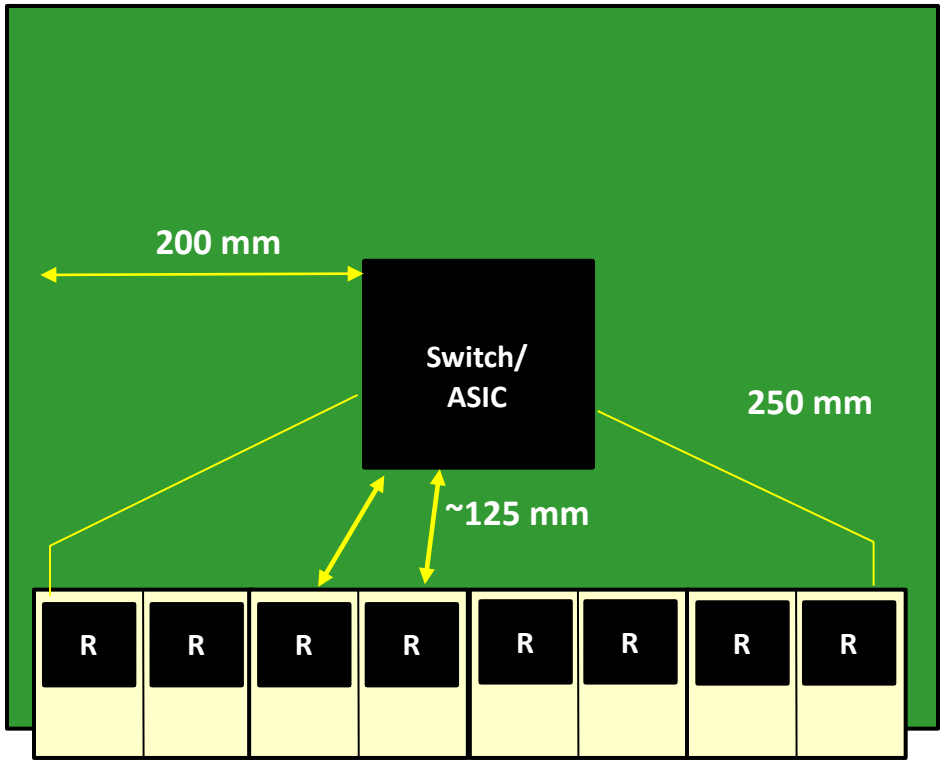


HUAWEI

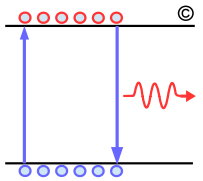
C2M Applications



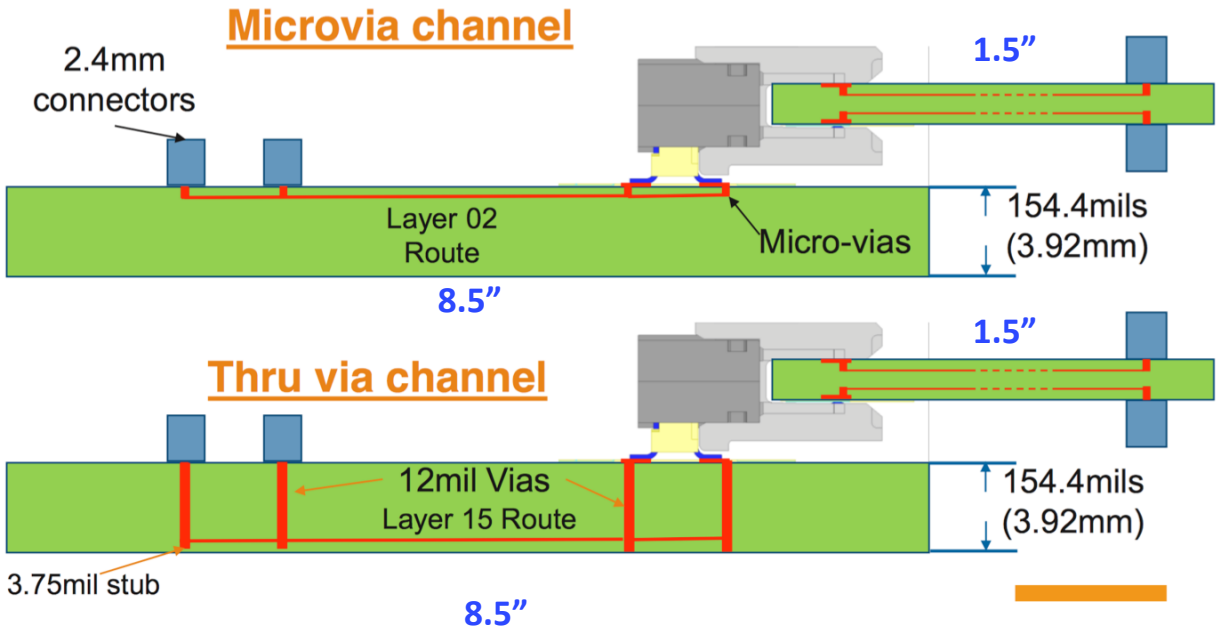
- **Numerous study in IEEE and OIF have shown typical line card require about 200-250 mm host traces**
  - CAUI-4 loss budget is 10.2 dB supporting ~125 mm on mid-grade PCB material like Isola 408HR
  - Most line card implementation prefer not to use retimer to save power and instead use Megtron 6 like 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 material such as Megtron 7/Tachyon.

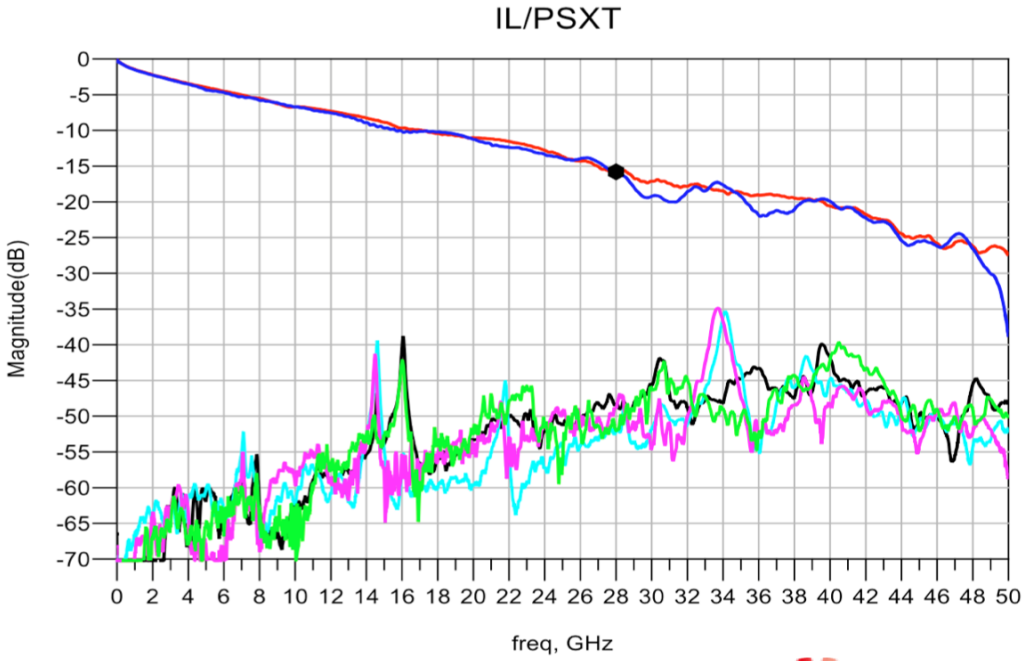


C2M Needs Practical PCB Trace Length and Construction

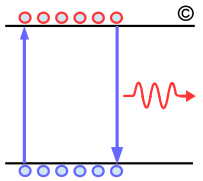


- **TE OSFP channel data is an example of a well built C2M channel**
  - http://www.ieee802.org/3/100GEL/public/18_01/tracy_100GEL_01a_0118.pdf
  - But the laser micro-via not feasible for complex board with several routing layers
  - 2X cal trace showed 1.36 dB/in loss @28 GHz (~1.3 dB/in @26.55 GHz)
  - 8.5" host channel on Megtron 7 HVLP+OSFP Connector+1.5" plug PCB has loss of ~15 dB@26.5 GHz





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 15 dB end-end channel loss consistent with tracy_100GEL_01a_0118

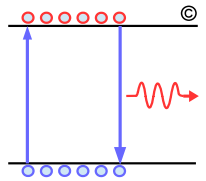
Host Trace Length (in)	Total Loss (dB)	Host Loss(dB)	Isola 408HR	Megtron 6	Tachyhon100
Nominal PCB Loss/in at 5.15 GHz	N/A	N/A	0.65	0.52	0.46
Nominal PCB Loss/in at 13 GHz	N/A	N/A	1.27	0.98	0.83
Nominal PCB Loss/in at 27 GHz	N/A	N/A	2.18	1.60	1.28
28G-VSR with one connector & HCB*	10.5	6.81	5.4	6.9	8.2
Iim_100GEL_adhoc_01_022618 Proposed	11.7	7.2	3.3	4.5	5.6
Current 112G-VSR draft+one connector & HCB**	13.5	9	4.1	5.6	7.0
100G C2M by Scaling 28G + connector & HCB**	15	10.5	4.8	6.6	8.2

Reach Inches Too Short

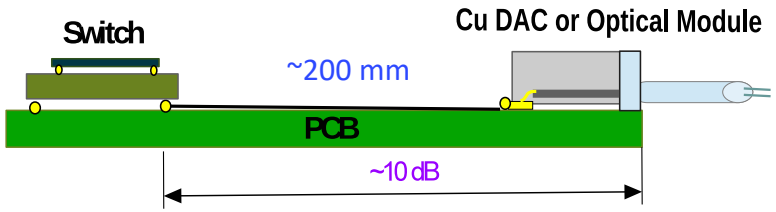
* Assumes connector loss is 1.69 dB and HCB loss is 2.0 dB at 12.89 GHz

** Assumes connector loss is 2.0 dB and HCB loss also 2.5 dB at 27 GHz.

Evolution of Front Panel Ports

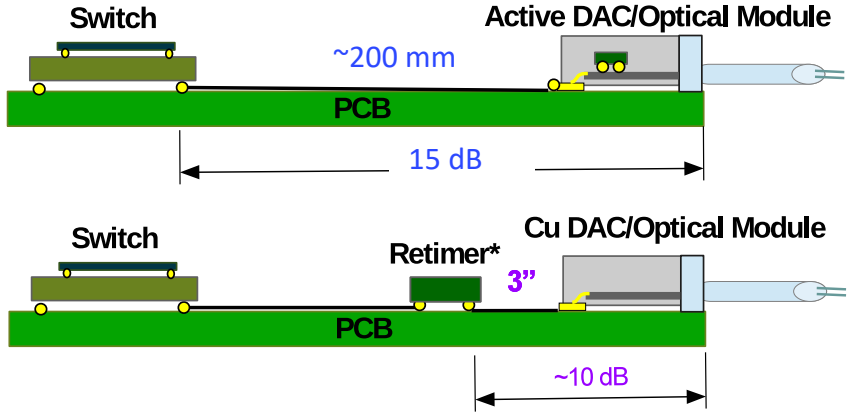


Pluggable at 25 Gb/s and 50 Gb/s



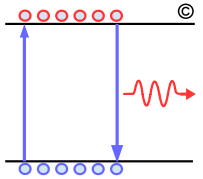
- PHY less design – what we are used to
  - Supports passive Cu DAC
  - Switch directly drives optical modules
  - Switch directly drives 3 m of Cu DAC
  - Offers optimum power and cost.

Pluggable at 100 Gb/s

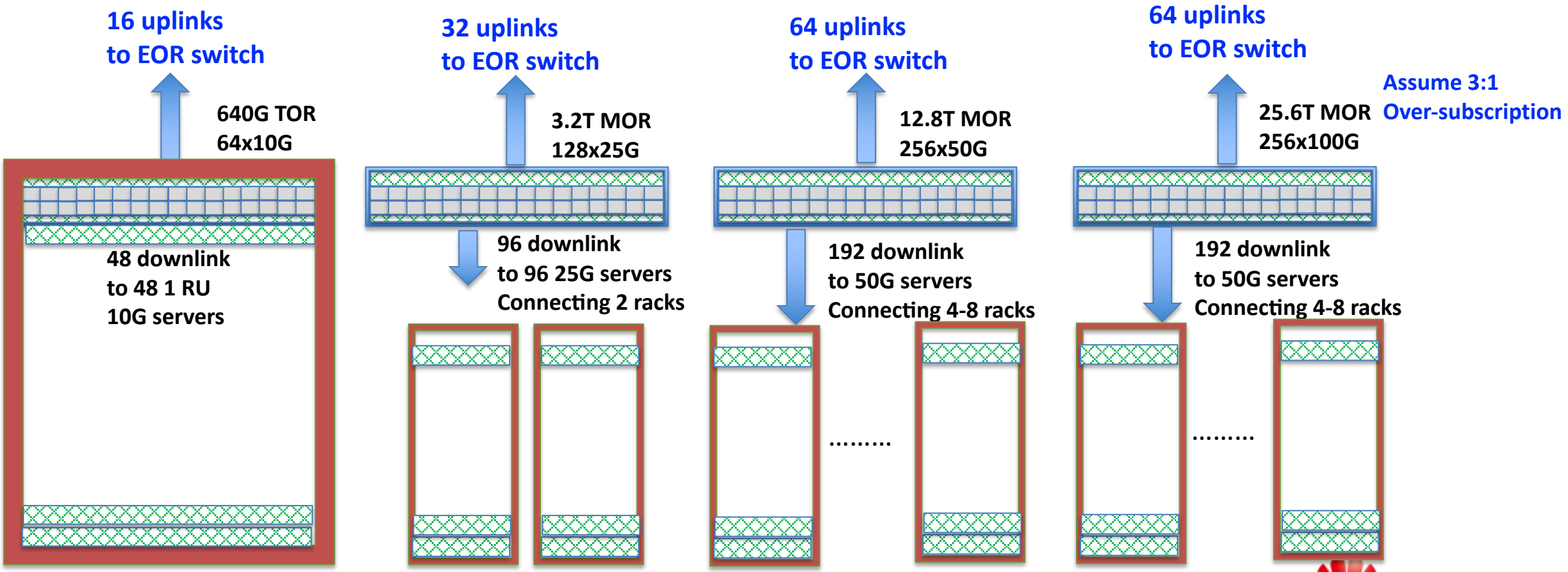


- Option I – PHYless Design – Channel loss 15 dB
  - Supports AOC, Active DAC, and Optics
  - Doesn’t support passive Cu DAC
  - 15 dB loss supports at least 200 mm PCB traces on premium material such as Megtron 7/Tachyon PCB
  - Offers improve power and cost
  - Better choice for MOR/Spine switches
- Option II – Require PHY – Channel loss 10 dB
  - Given that high radix switches if used as TOR require connecting servers on 4-6 racks passive DAC no longer feasible
  - Low capacity switches that can serve single server rack can just stay with 50G signaling
  - Adding 100G retimer assuming 1W/lane on a system having 16 line card with each line card based on 256by100G will add whopping 4 KW to the system power envelop!

Datacenter Trends

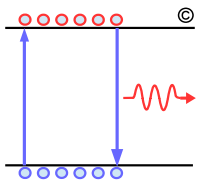


- **Switch radix over the last 9 years has increased from 64x10G, 128x25G, now to 256x50G, and likely to 256x100G by 2019/2020**
  - To mitigate full rack failure dual MOR switches may connect to each rack.



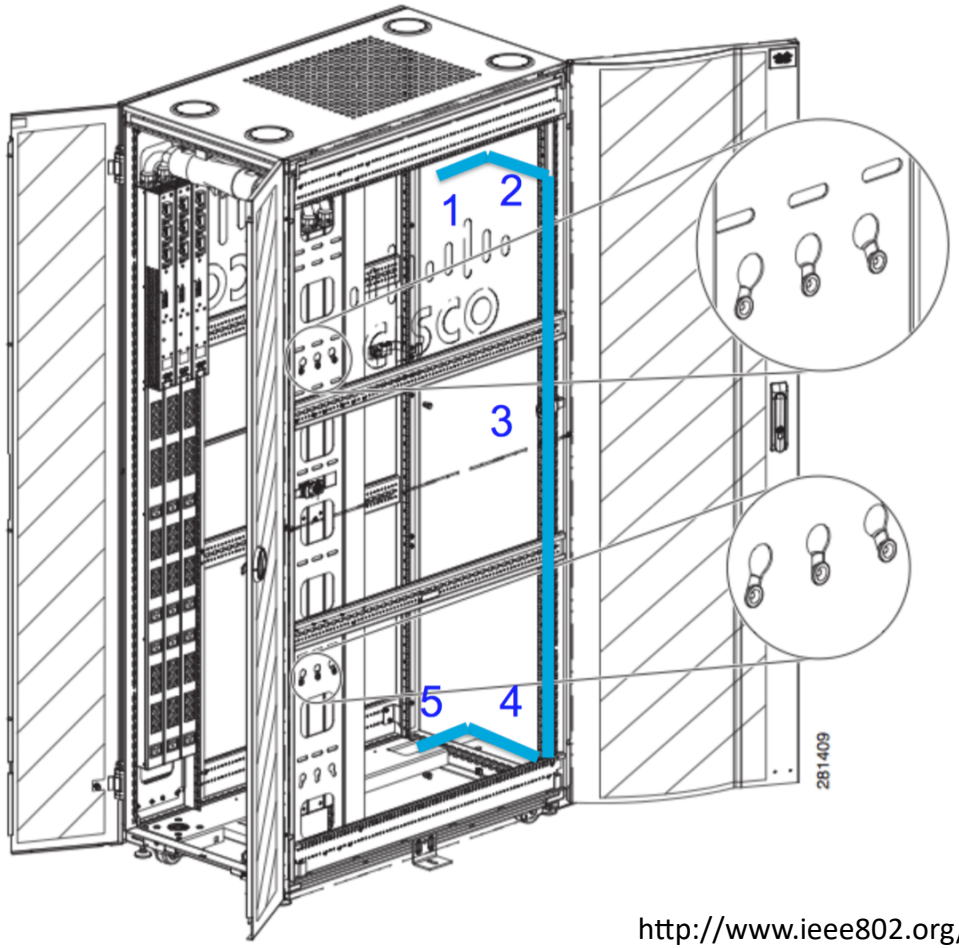
HUAWEI

Study Performed By Joel Goergen in 802.3by Indicate 3 m is necessary for Cu Cable!



- **Given that high radix switches can connect to 4-6 racks of server passive Cu cable no longer a viable option for 1st level switch-servers**
  - Potential use case for Cu cables at 112G will be switch to switch and one may not assume asymmetrical link
  - Application not driven by network performance may use an small TOR switch within the rack for simplicity and use 25G/50G Cu cabling!

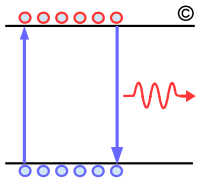
Cabling Installation – Top to Bottom



- Consider this common strategy
  - 1 – 152mm
  - 2 – 304mm
  - 3 – 1778mm
  - 4 – 304mm
  - 5 – 152mm
- This real life case is 2690mm.

http://www.ieee802.org/3/by/public/July15/goergen_3by_02a_0715.pdf

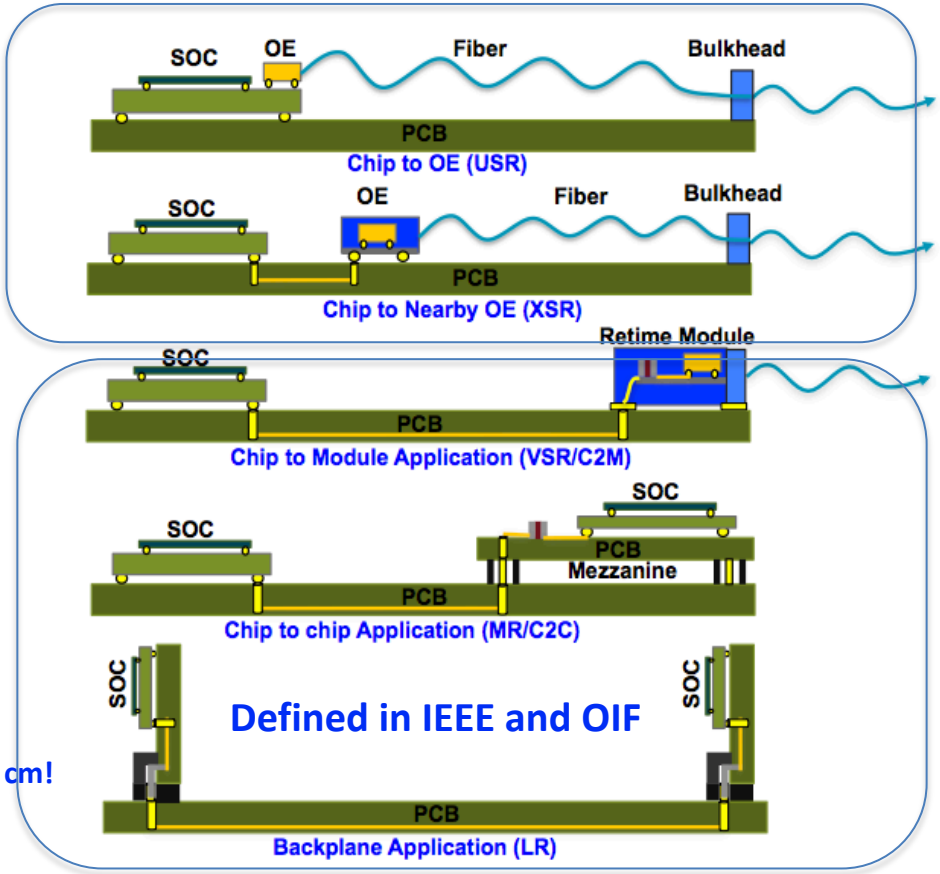
The 50G/lane Interconnect Ecosystems



- **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

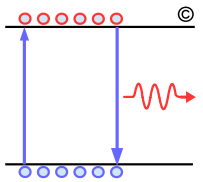
Defined in OIF



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.

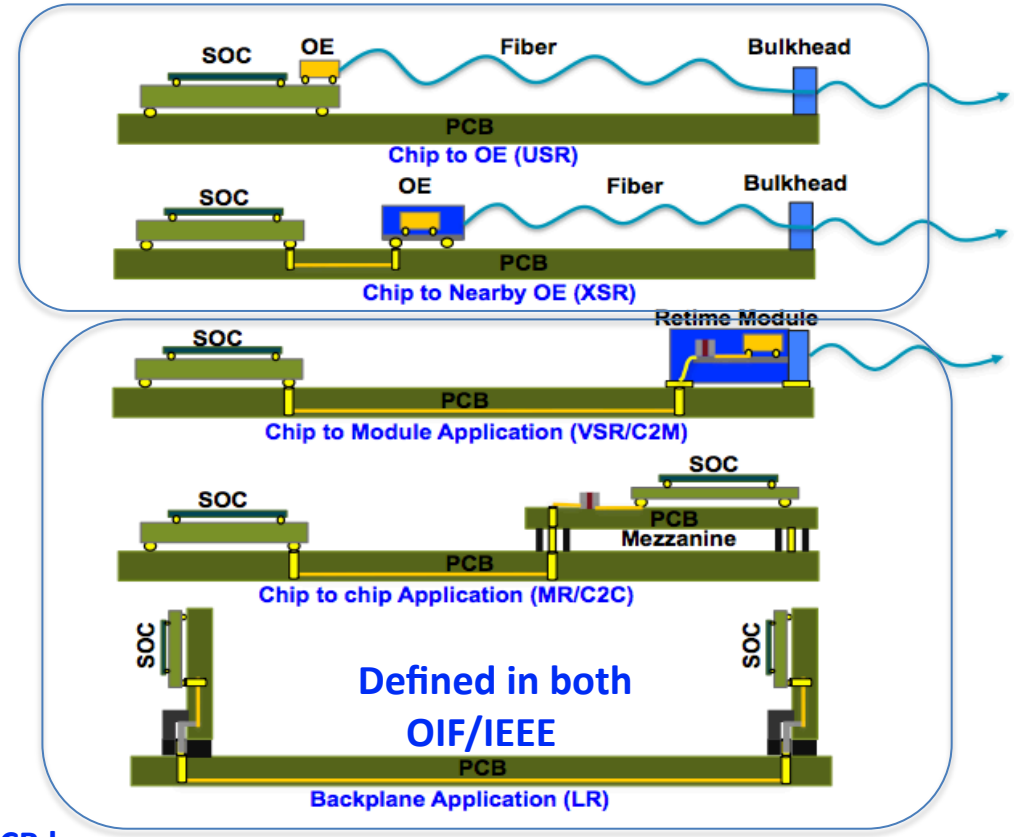
HUAWEI

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 5 dB loss
  - PCB reaches below assumes Tachyon 100/Megtron 7
  - Bump-bump loss for LR SerDes reduced by 1 dB from 50G PAM to account for additional impairment related to crosstalk, reflection, and ILD.

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)	C2M	PAM4	< 20 cm**	15 dB	20 dB
Chip-to-chip (one connector)	C2C	PAM4	< 40 cm	20 dB	30 dB
Cabled Backplane (two connectors)	LR	PAM4	<50 cm	25 dB	35 dB

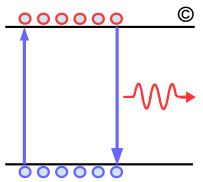


* OBO connector + CDR package assumed having 2 dB loss

** C2M host packaged assumed 5 dB loss and the CDR packaged assumed to reuse 2 dB HCB loss.

HUAWEI

A possible path forward is to optimize the 2 m Cu DAC for Switch to Server



- **Proposed host loss in Ghiasi_100GEL_01_0318 is 10.5 dB vs 8 dB in lim_100GEL_01_0318**
  - Given the primary application of 2 m Cu DAC is switch to server
  - Limit NIC PCB loss to 4 dB, allocate +2.5 dB to switch PCB, use 1.5 dB excess budget for more robust 2 m Cu
  - With 28.5 dB ball to ball budget one could support 4-5 dB loss for switch package loss and with 2-3 dB for NIC

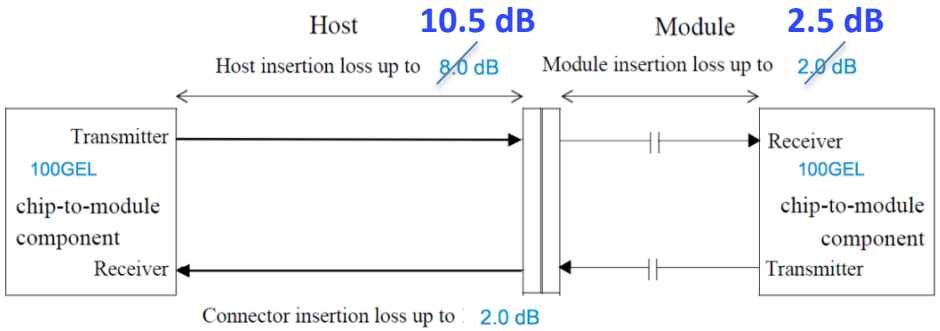


Figure 1: 100GEL C2M insertion loss budget at 26.56 GHz

lim_100GEL_01_0318

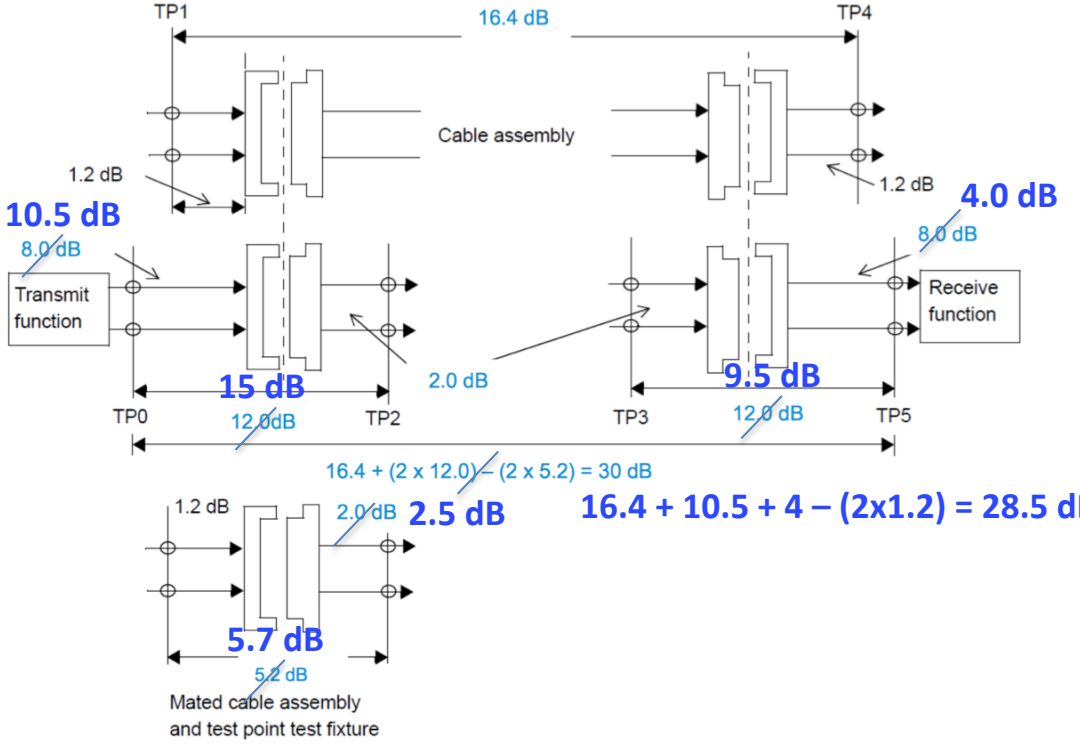
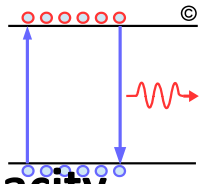


Figure 2: 100GEL CR 30dB insertion loss budget at 26.56 GHz

Summary



- **The primary applications that will benefit from 112G are the high capacity routers delivering capacity needed for 5G networks and high radix switches enabling next generation hyper scale data centers**
  - Managing power and cost will be key challenge for these type of systems
- **What is necessary to enable these next generation system based on 112G/lane electrical IO are**
  - C2M with at least 200 mm PCB (15 dB) support
  - C2C with at least 400 mm PCB (20 dB ball-ball)
  - Reuse of RS (544, 514) for C2M and C2C interfaces
- **Backplane applications based on 0.5 m conventional PCB or 1 m cabled backplane with 35 dB loss should also be considered as long as does not delay the C2M and C2C development**
  - For backplane application should consider both RS (544, 514) as well as stronger FEC
- **Cu cable since introduction of SFP+ CU DAC has been a huge success, but introduction of switches and QSFP-dd/OSFP supporting 256 lanes has diminished value of Cu DAC for TOR-Servers applications**
  - Cu cable should be considered in this project as long it does not scarifies C2M PCB reach
- **How to move forward not sacrificing C2M PCB reach and support 2 m Cu cable objective:**
  - Define optical MDI based on 15 dB loss and Cu MDI with 10 dB, a port with 10 dB loss can support Cu and optics
  - Given the primary application of 2 m DAC is switch to server an asymmetrical link budget as shown can support high density TOR as well as NIC without need to create superset ports.