System Evolution with 100G Serial IO

Ali Ghiasi
Ghiasi Quantum LLC

100 Gb/s/Lane NEA Meeting
New Orleans

May 24th, 2017

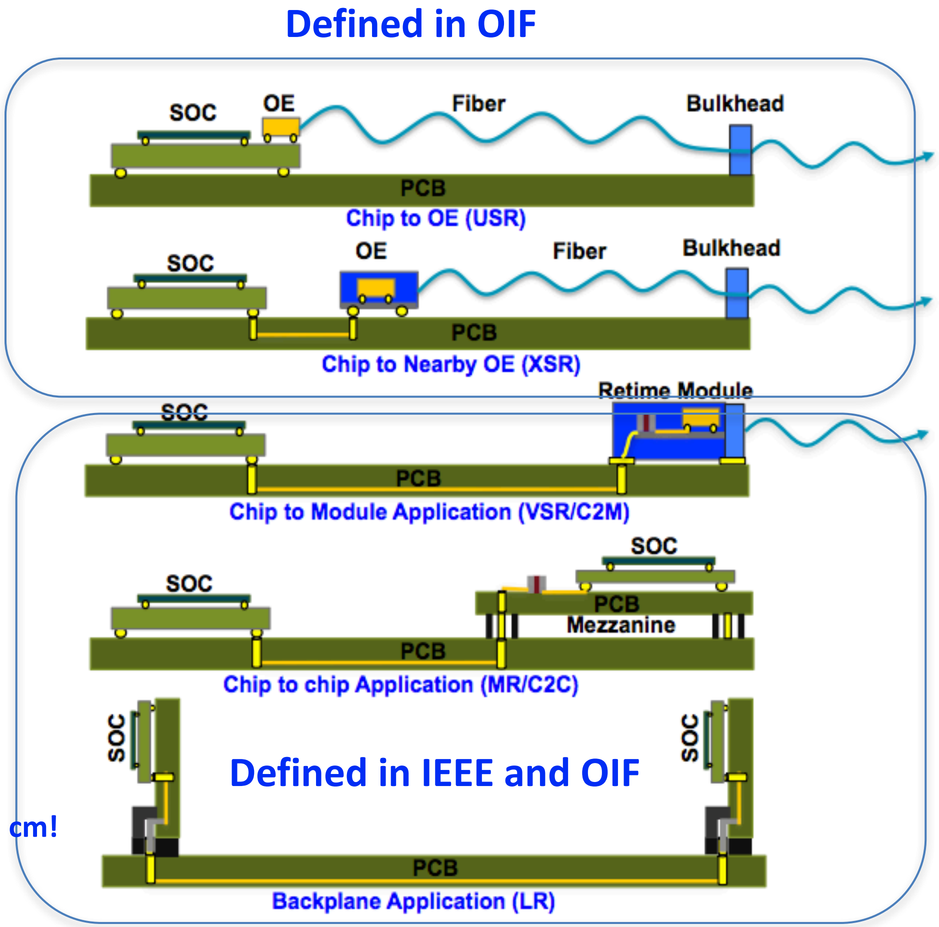
Overview

- **Since 10GBASE-KR superset ASIC SerDes have supported C2M, C2M, and backplane applications**
  - Adding KR/CR capability provided a solution to support Cu DAC and backplane small power penalty
  - The superset KR/CR SerDes supported C2M pluggable optics
- **At 112G need to reconsider our historical architecture to make sure the system is cost and energy efficient**
- **Expect 112G signaling to be based on PAM4 for following reasons:**
  - Higher order modulation such as PAM8, PAM12, PAM16 require stronger FEC with higher latency and eco-canceller due to discontinuity in the channels
  - More complex FEC and eco-canceller can’t be integrated into large ASICs
  - Any chip-to-module signaling other than PAM4 require a convertor chip for 100GBASE-DR and 400GBASE-DR4
  - Any FEC other than RS (544,514) require FEC termination and initiation in the module for 100GBASE-DR and 400GBASE-DR4 with significant latency impact
- **Considering eco-system requirement this contribution only considers PAM4 with KP4 FEC for 112G applications!**

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

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



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

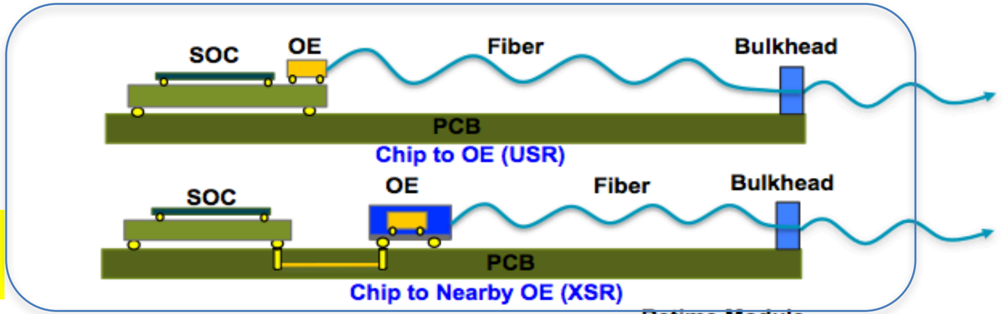
- **With estimated loss of 18 dB C2M specification is inline with our definition of C2C**
  - Bump to bump loss calculated by assuming ASIC package with 6 dB loss and small CDR package having 2 dB loss
  - 6 dB ASIC package assumes 30 mm trace and requires material better than GZ41
  - PCB reaches below assumes Tachyon 100/Megtron 7
  - C2M with 18 dB loss is more inline with current C2C SerDes
  - Should we consider defining OBO and/or MCM applications?

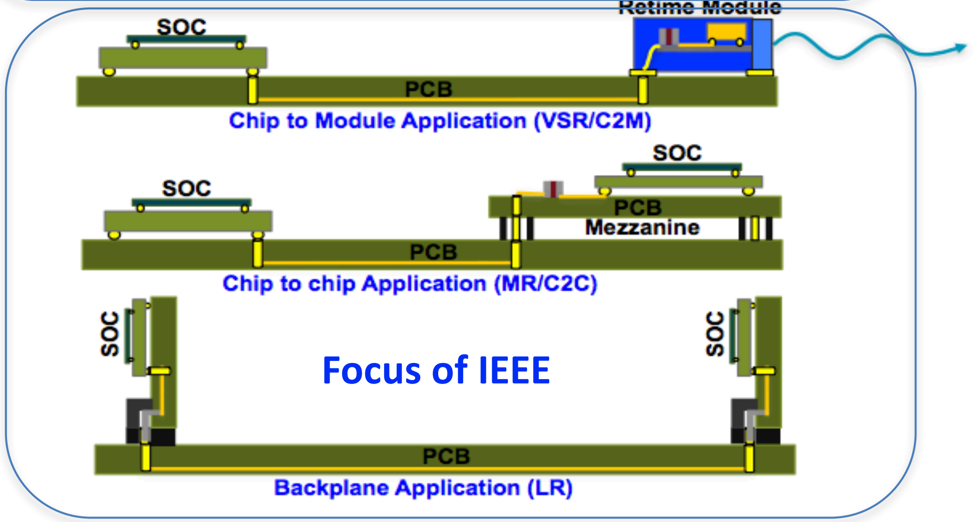
Application	Standard	Modulation	Reach	Ball-Ball Loss	Bump-Bump Loss
Chip-to-OE (MCM)	TBD	PAM4	< 1 cm	NA	2 dB
Chip-to-nearby OE (no connector)	TBD	PAM4	<10 cm*	5 dB	12 dB
Chip-to-module (one connector)	OIF-112G-VSR	PAM4	< 25 cm	18 dB	26 dB
Chip-to-chip (one connector)	TBD	PAM4	< 38 cm	20 dB	32 dB
Cabled Backplane (two connectors)	TBD	PAM4	<50 cm	24 dB	36 dB

Possibly C2C can be met with 24 dB SerDes

* Practical OBO implementation requires 10 cm!

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

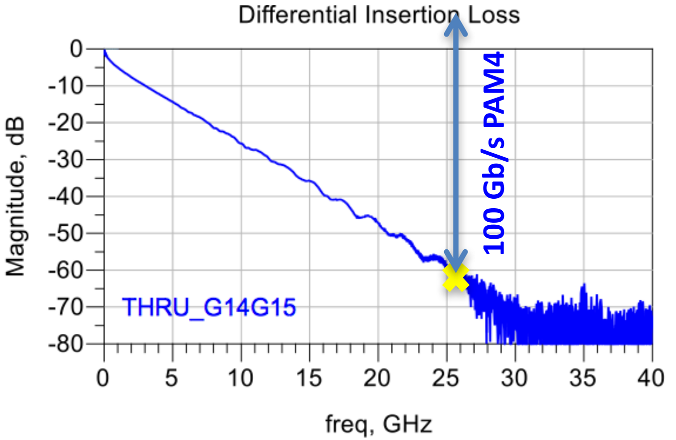



Retime Module
SOC
PCB
Chip to Module Application (VSR/C2M)
SOC
SOC
PCB
Mezzanine
PCB
Chip to chip Application (MR/C2C)
SOC
SOC
Focus of IEEE
PCB
Backplane Application (LR)

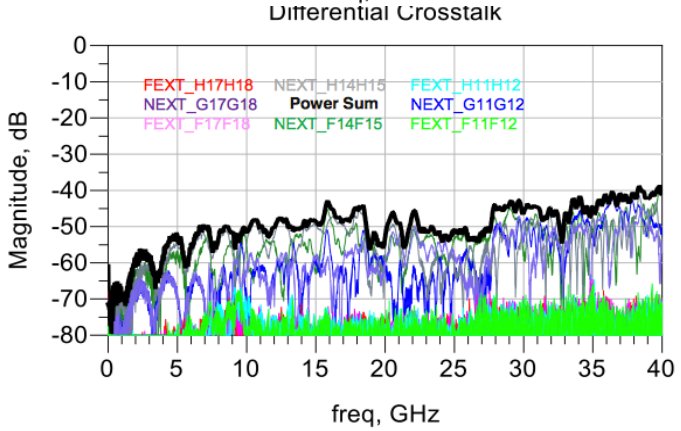
Conventional Backplane no Longer Feasible at 100 Gb/s!

- ❑ **TE Whisper 40" conventional backplane at 100 Gb/s PAM4 Nyquist has a loss of ~65 dB ***
- ❑ **1 m cabled backplane is viable with short daughter-card, in effect every lane needs a retimers!**

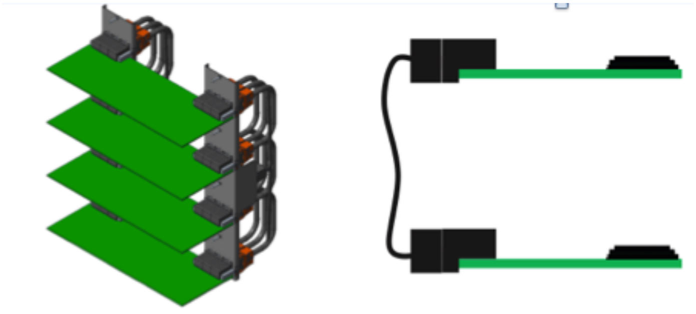
TE Whisper Conventional Backplane 40" with Meg 6 HVLP *

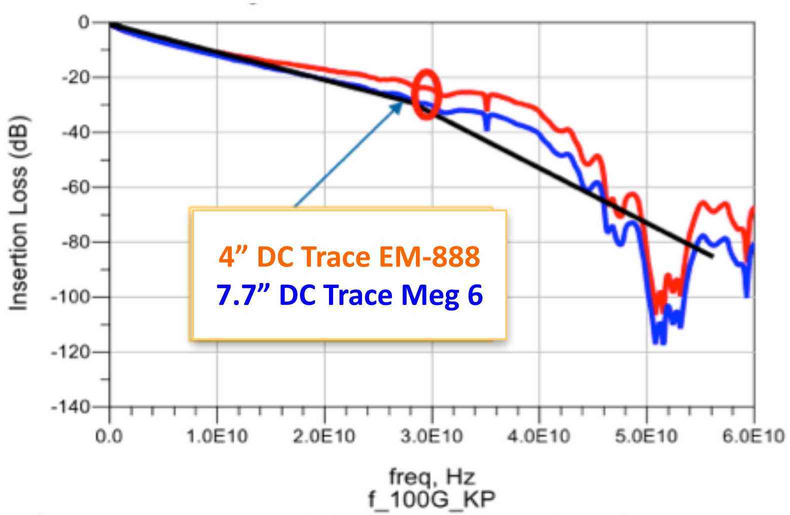


Even High Power 10GBase-T EQ Can equalize the Channel



TE Whisper 1 m Cabled Backplane **





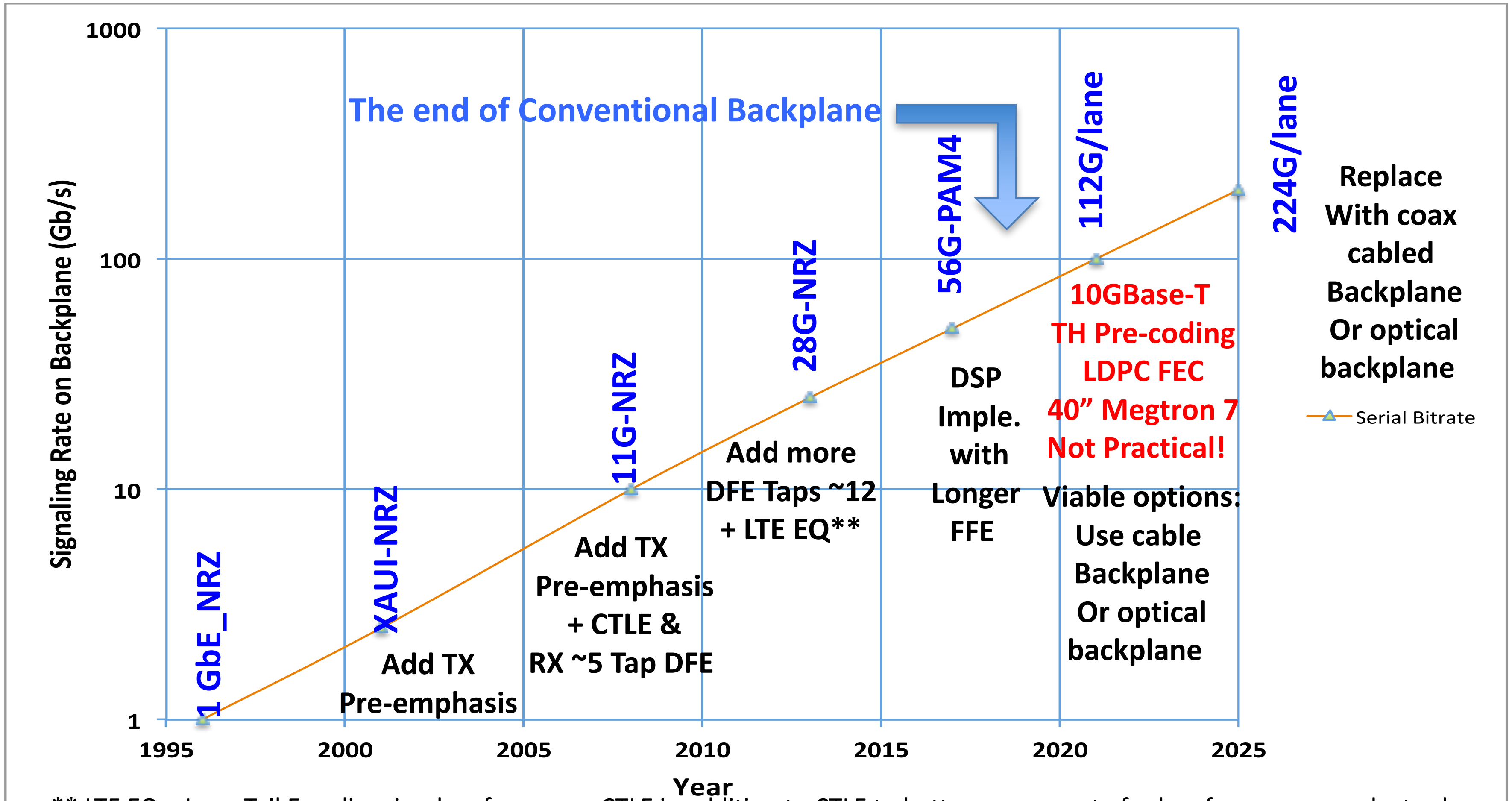
4" DC Trace EM-888
7.7" DC Trace Meg 6

* TE Whisper channel, http://www.ieee802.org/3/cd/public/channel/Reference_document_for_TE_Connectivity_Backplane_S-Parameter_Channels_07_28_16.pdf

** Achieving 100 Gb/s Channels, David Hester TE Connectivity, OIF 2016 100 Gb/s Workshop.

When do we need 100G signaling?

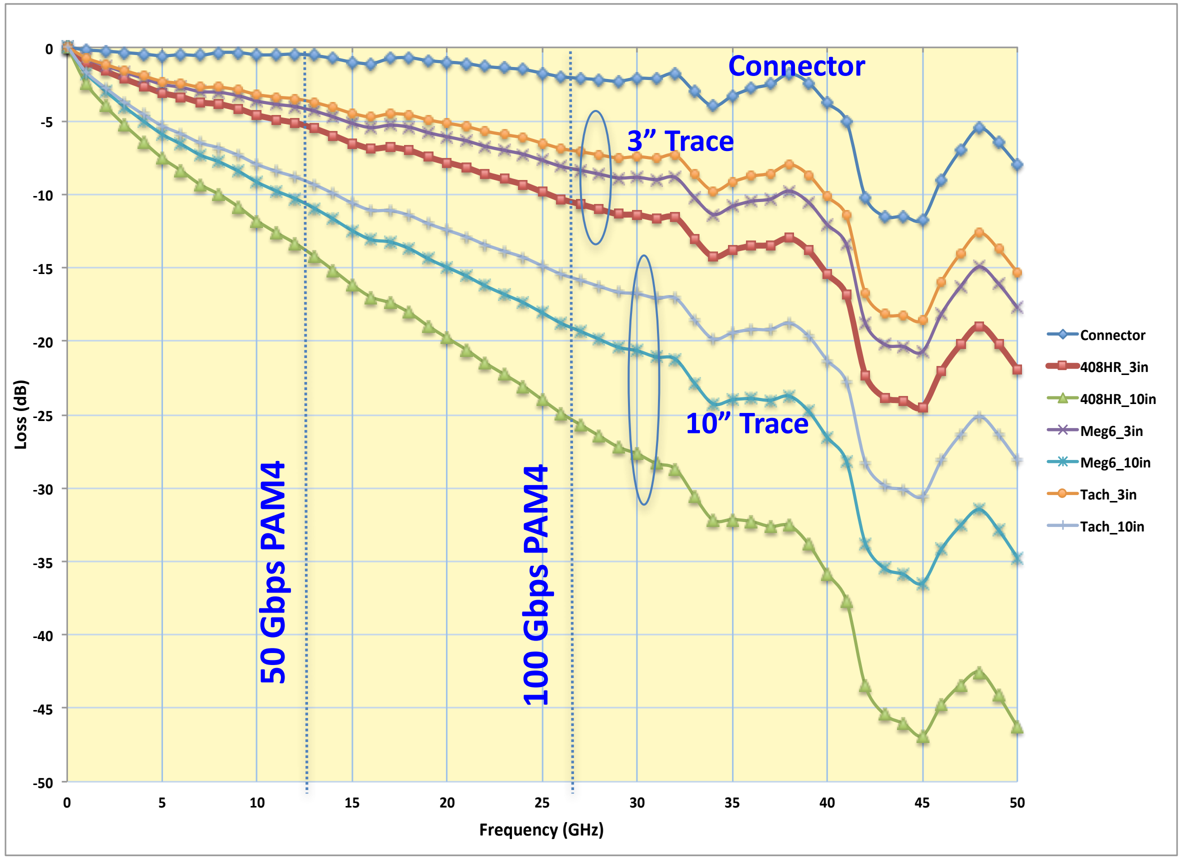
- Product based on 112G/lane are expected to be deployed by 2021



** LTE EQ = Long Tail Equalizer is a low frequency CTLE in addition to CTLE to better compensate for low frequency conductor loss.

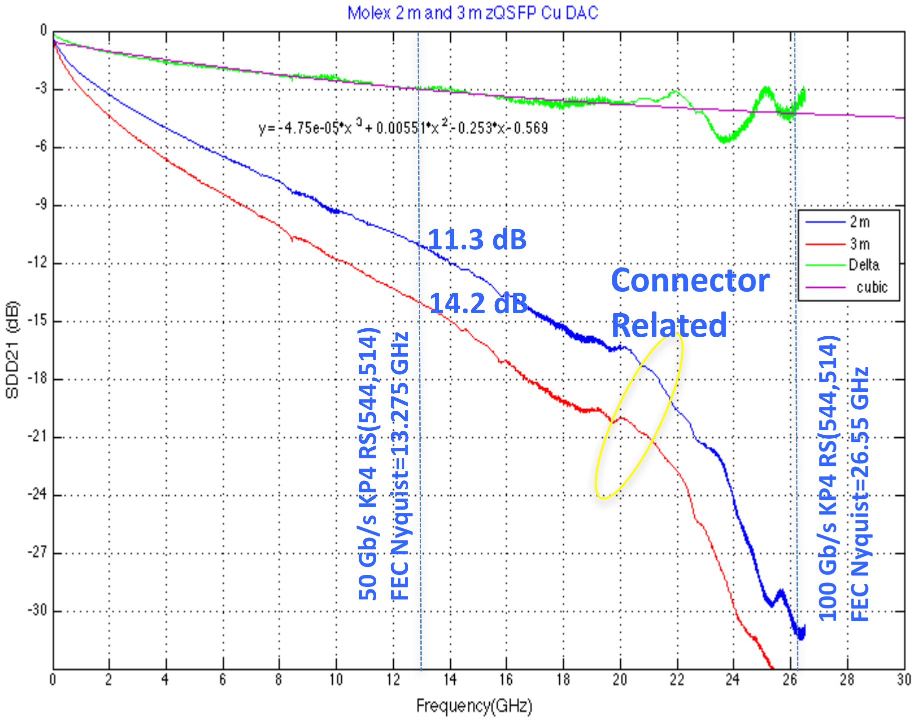
112G C2M Channels

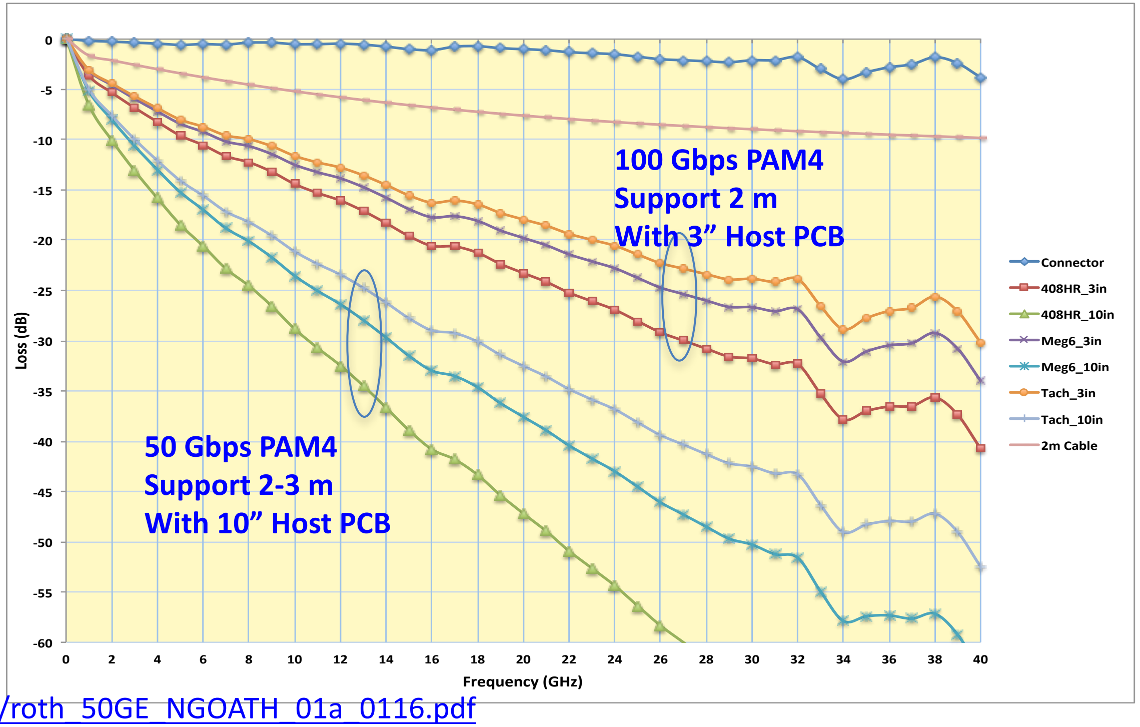
- **Connector assumed is Yamachi CFP2 which is capable of 53 GBd operation other connectors potentially could be improve**
  - VSR channel loss investigated with following material 408HR, Megtron 6 HVLP, Tachyon HVLP for 5.5 mil ½ oz stripline
  - To stay with 56G-VSR loss limit of 10.5 dB the host PCB trace will be <75 mm and even with ultra low loss material the end to end loss will be ~19.5 dB (7 dB for host ASIC and 2.0 dB CDR)!
  - CTLE receiver is no longer an option
  - Better to use C2C receiver and go little longer for PHYless design
  - With ~18 dB loss 125-250 mm of host PCB can be supported with end-end loss of 27 dB
    - Inline with 50G C2C definition.



Extending Cu DAC Operation from 50 to 100 Gbps

- **Construction of the hypothetical 100 Gb/s Cu DAC**
  - De-embed Molex zQSFP cable response then build a hypothetical DAC with Yamaichi CFP2 connector
  - Hypothetical 2 m Cu DAC with 10" trace has end-end loss of ~54 dB (assuming 2x~7 dB ASIC package)
  - Instead a 3" host Tachyon 100 with 2 m cable has end-end loss of ~ 37 dB (assuming 2x~7 dB ASIC package)
  - A high end DSP retimer could provide a passive Cu DAC solution for 2 m with <3" host but will be costly and high power
  - A better solution is to go with <10" PCB (PHY-less) and instead replace passive DAC with active DAC or AOC.



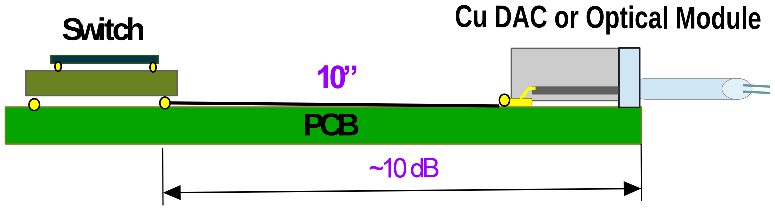


*zQSFP cable data, http://www.ieee802.org/3/50G/public/Jan16/roth_50GE_NGOATH_01a_0116.pdf

**CFP2 connector, http://www.ieee802.org/3/400GSG/public/13_05/nishimura_400_01a_0513.pdf

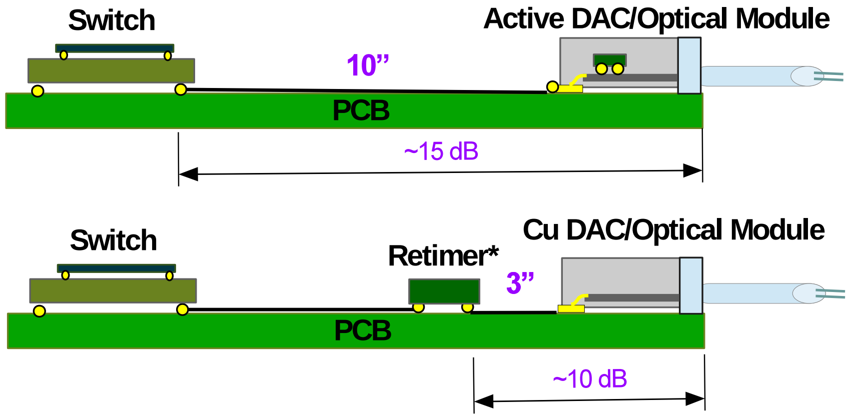
Evolution of Front Panel Ports

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

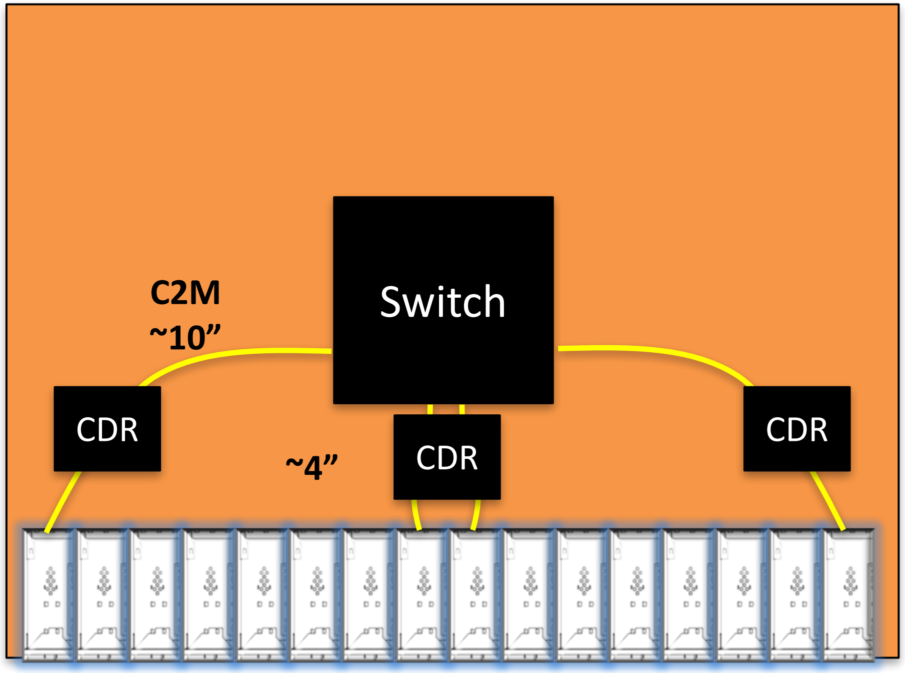


- ❑ Option I – PHY less design
  - Doesn’t support passive Cu DAC
  - Switch directly drives pluggable module, active Cu DAC, or AOC
  - Support 10” of Megtron 7/Tachyon PCB
  - Offers improve power and cost
  - Better overall choice as industry transition toward fiber centric
- ❑ Option II – Require PHY close to every module
  - Supports passive Cu DAC, active DAC, and AOC Support 3” of Megtron 7/Tachyon PCB
    - Flyover cable can extend the PHY to module distance but adds cost and manufacturability issues
  - Supports Active Cu DAC and optical modules
  - Retimer adds significant cost and power.

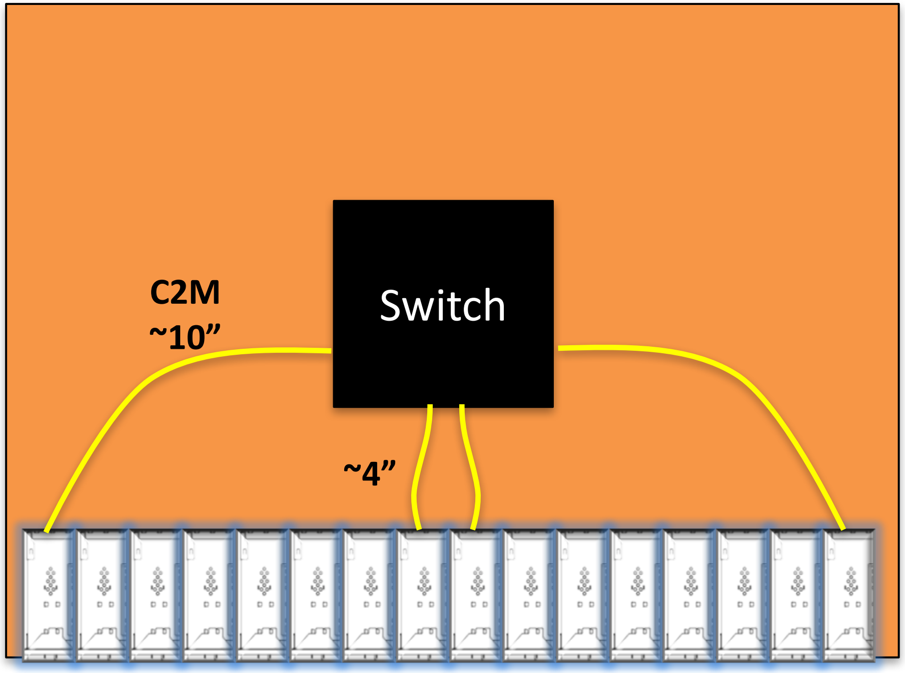
1RU/TOR Implementation

- Given that optical PMDs/AOC use retimer adding 2nd retimer/CDR on the host port add unnecessary power

Not Preferred!

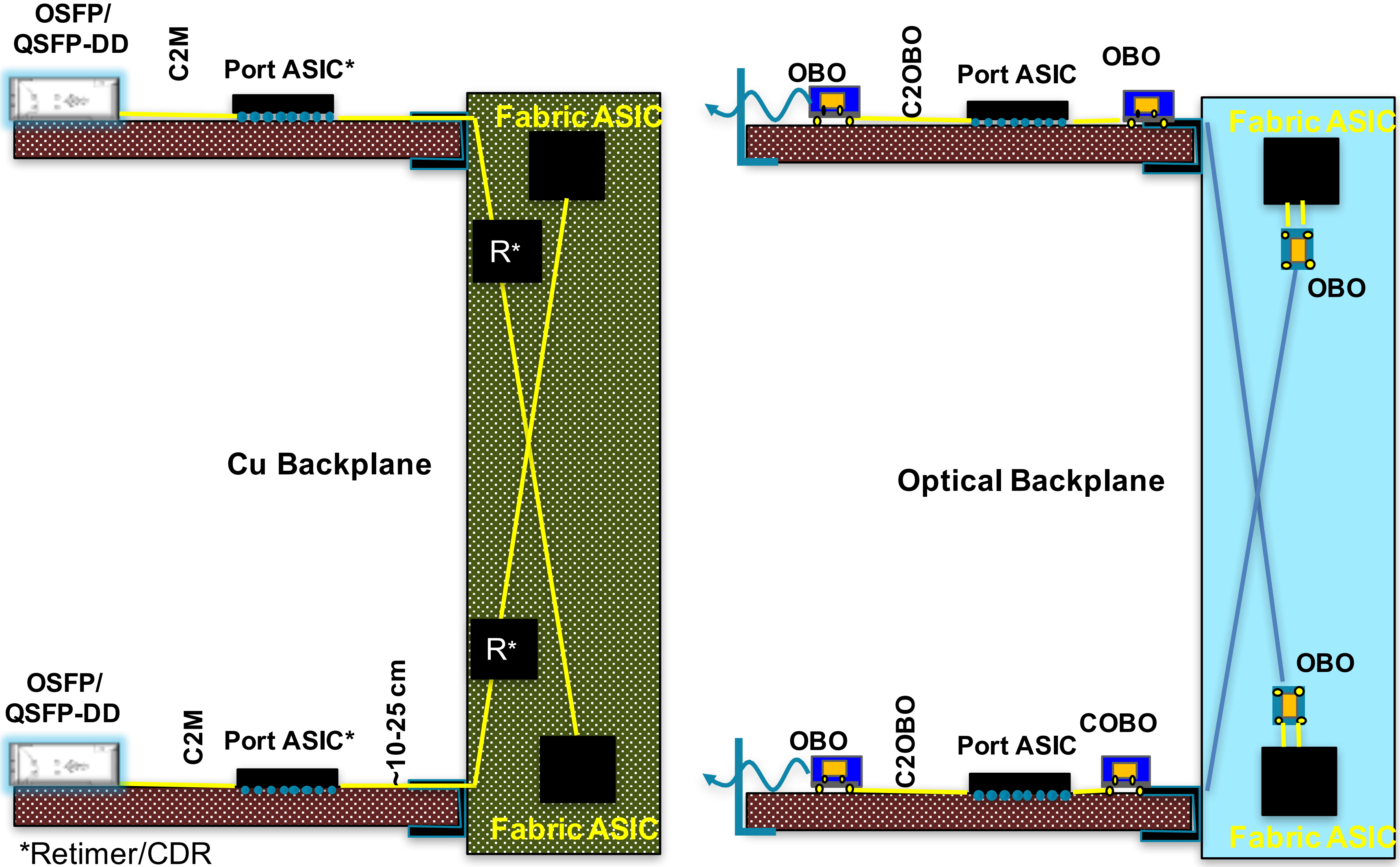


Preferred!



Chassis Implementation

- To support a practical size chassis most link would require a retimer/CDR
- In the time frame of consideration we should not rule out OBO and optical backplanes!



*Retimer/CDR

Summary

- **The 100G/lane will offer more efficient ASIC interface by doubling the switch BW**
  - OSFP/QSFP-dd or QSFP112 with 100 Gb/s/lane signaling could deliver 14.4-25.6 Tb/s front panel BW
  - The downside of 100G/lane IO are lack of 10 km PMD and 850 nm MMF PMDs support as these PMDs may require operation at 50 Gb/s/lane with inverse Mux
- **Given that at 100 Gb/s/lane supporting conventional 1 m backplane or 3 m passive cable no longer feasible one must first consider the architectural impact**
  - Conventional backplane likely will be replaced with cabled backplane, use Megtron 7/Tachyon 100 on a short backplane <50 cm linecard to fabric, add extra retimer to extend the reach, or use optical backplane
  - We need to focus on an energy efficient, cost effective, synergistic solution – PLEASE NO 100GBASET!
  - Instead of trying define a heroic passive Cu DAC solution, it would be simpler and more economical to use active Cu DAC or AOC
- **Given that 100GBASE-DR and 400GBASE-DR4 are based on PAM4 with KP4 FEC any other signaling and/or FEC would require PHY layer adding complexity and latency**
  - Potentially active Cu DAC may use internally other signaling
- **The transition to serial 100G/lane will not be smooth like 50G/lane transition**
  - Even with material like Megtron 7 or Tachyon 100 C2M loss will be ~18 dB requiring a C2C like equalizer
  - We can’t roll rule out OBO or co-package at 100 Gb/s/lane
  - Should we consider defining C2OBO interface
- **What has worked at 25G/50G may not be the optimum system/ASIC partition at 100G/lane!**