Loss estimates for System Applications with Large Scale Switch – AUI Types

Upen Reddy, Kareti – Cisco Systems Inc.

Overview

- Scope
- Objectives
- Systems considerations
- 100Gb/s and 200Gb/s Conditions and Assumptions
- Review of Losses for suitable AUI types
- Summary /Observations

Scope

- Scope is limited to C2M AUI interfaces for all Ethernet rates and signaling rates
- Backplane and Copper cable applications are not covered.

Table from

https://www.ieee802.org/3/B400G/public/21_1028/B400G_overview_c_211028.pdf

Adopted Physical Layer Objectives

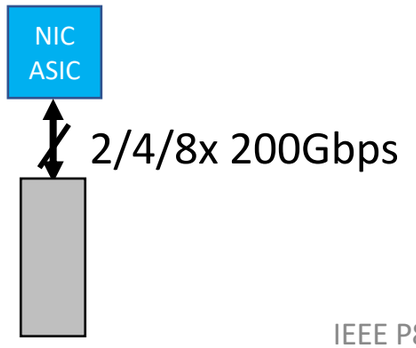
Ethernet Rate	Assumed Signaling Rate	AUI	BP	Cu Cable	MMF 50m	MMF 100m	SMF 500m	SMF 2km	SMF 10km	SMF 40km
200 Gb/s	200 Gb/s	Over 1 lane		Over 1 pair			Over 1 Pair	Over 1 Pair		
400 Gb/s	200 Gb/s	Over 2 lanes		Over 2 pairs			Over 2 Pair			
800 Gb/s	100 Gb/s	Over 8 lanes	Over 8 lanes	Over 8 pairs	Over 8 pairs	Over 8 pairs	Over 8 pairs	Over 8 pairs		
	200 Gb/s	Over 4 lanes		Over 4 pairs			Over 4 pairs	1) Over 4 pairs 2) Over 4 λ's		
	TBD								Over single SMF in each direction	Over single SMF in each direction
1.6 Tb/s	100 Gb/s	Over 16 lanes								
	200 Gb/s	Over 8 lanes		Over 8 pairs			Over 8 pairs	Over 8 pairs		

Objectives

- A look into Large scale Switch applications in systems and estimate losses for different AUI types that are suitable
  - Signaling rates covering 100Gb/s and 200Gb/s
  - 200Gb/s estimates include for PAM4 modulations
  - All estimates assumed RS-FEC (544,514,10) unless noted otherwise
  - Study impact of stronger RS-FEC (576,514,10) on losses due to increase of overheads

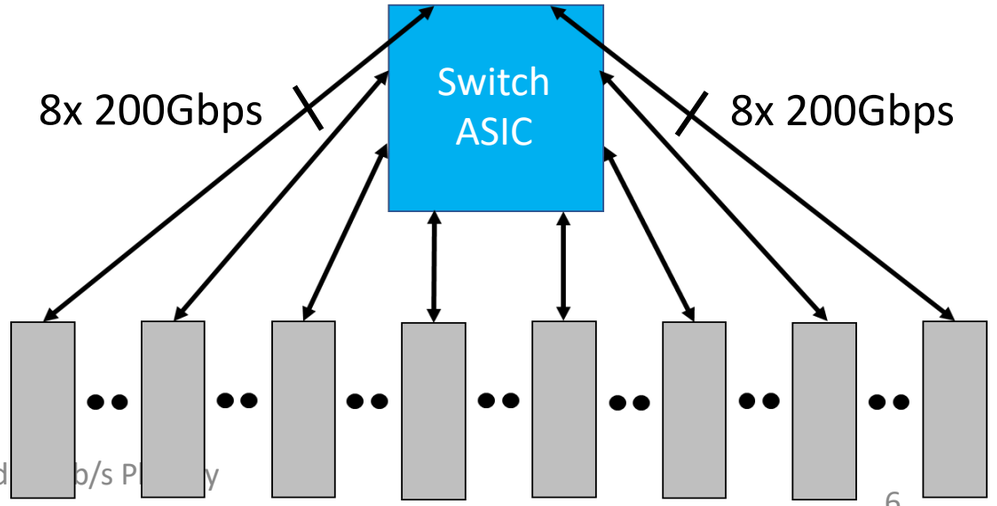
Why The Large Switch Use Case Matters

- NIC Use Case: (most previous contributions)
  - Low Radix (e.g. 8x200G)
    - = SERDES on <u>one</u> side of pkg
    - = <u>Small</u> Package
    - = Tight ball size & pitch is OK
  - 1 or 2 optics on one side (PCIe faceplate limited)
    - = very short channel length



1600G cage (e.g. QSFP-DD1600)

- Large Switch Use Case: (topic of this PPT)
  - High Bandwidth (102.4T), High Radix (e.g. 512x200G)
    - = SERDES on <u>all</u> sides of pkg
    - = <u>Large</u> Package (>16x the area)
    - = Wider ball size & pitch (solderability reasons)
  - 64 optics on <u>one</u> side (“pizza box” font panel)
    - = much longer channel lengths



64x1600G cages (e.g. QSFP-DD1600)

Continuing the trend for 100G Generation

- 802.3ck project for 100G generation adopted higher host loss for AUI interfaces

  - enabled retimer less switch hosts in 100G generation.

- Critical to
  - consider the switch not just the NIC
  - Support the retimer less switch hosts for 200G generation
  - Support the Switch integration trend 25.6T =>51.2T=>102.4T=>204.8T?

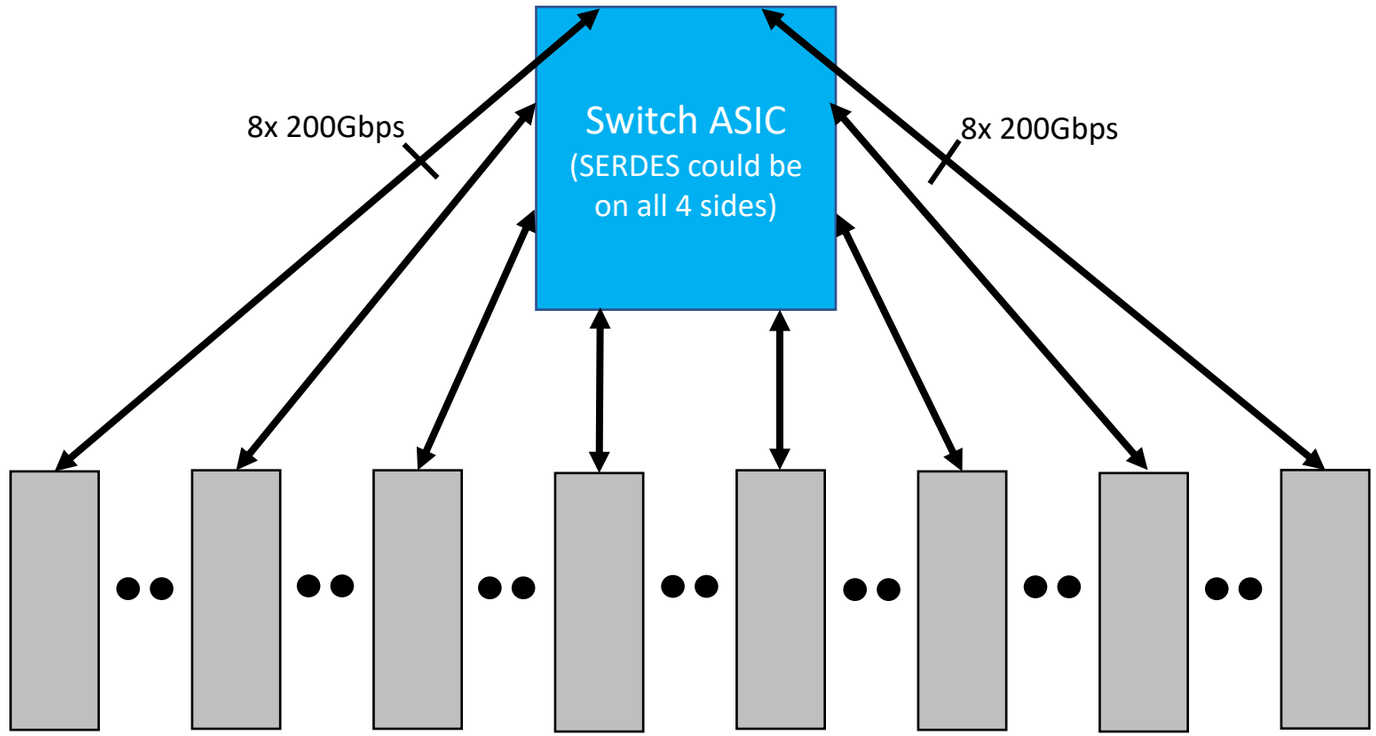
System considerations

- Scaling Switch from 51.2T Switch to 102.4T Switch poses significant challenges
  - Electrical, Mechanical, Thermal and power solutions must be significantly better than 2X of 51.2T systems
  - End-user Overall Power consumption limitations per rack and per system
    - Facility limitations
    - Green initiatives
  - Doubling the data rate limits the reaches using reasonably low power serdes architecture
  - Thermal aspect of the design is becoming very difficult with respect to system cooling solutions and is also a burden on overall system power consumption
  - Using more Re-timer/re-drivers to solve the reach issues places burden on
    - overall system power consumption
    - power delivery to the components
    - Thermal solutions
    - Pushes system Mechanical envelope larger

100Gb/s and 200Gb/s Conditions and Assumptions

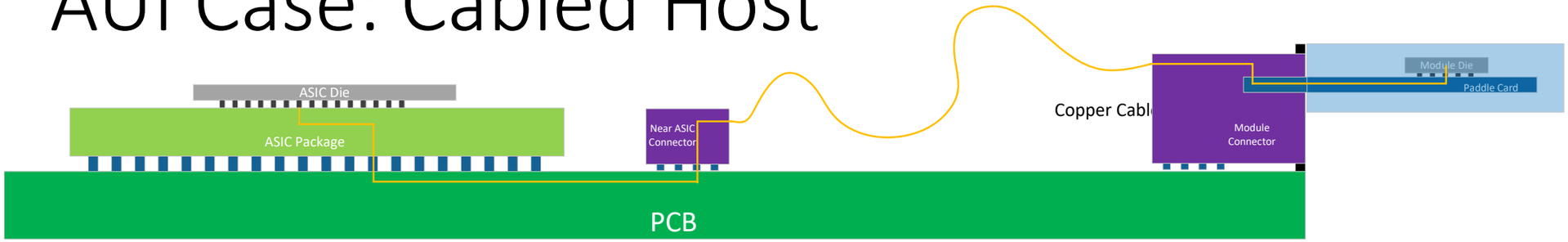
- 100G evaluations included here compliant to either AUI types considered in C2M standards of P802.3 ck or AUI types that are being considered in OIF like XSR+ for 100Gb/s
- Package and PCB loss improvement options like skipped layers technique are considers for all data rates – 100Gb/s and 200 Gb/s
  - These type of approaches has coverage limitation for large scale switch(s) and system(s)
  - Usage of these approaches are optimized to reduce overall max loss contribution
- 200G/s evaluations assume
  - PCB, cable, package substrate material improvements that are possible in next 2-4 years based on preliminary data
  - Connector losses are estimated based on available preliminary data
  - BGA and Package Core via designs and loss contributions are based on Preliminary simulations

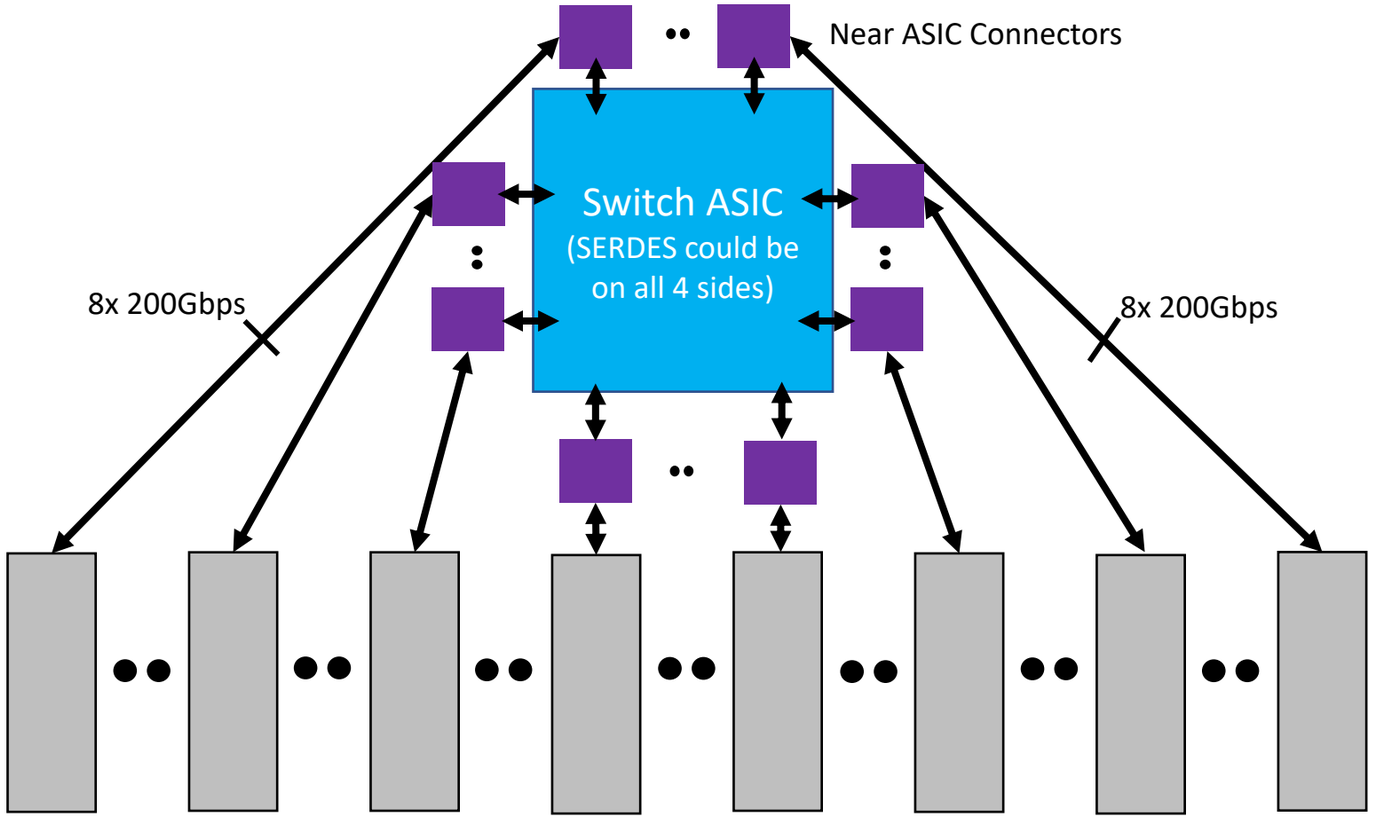
AUI Case: PCB Host



64x1600G cages (e.g. QSFP-DD1600)

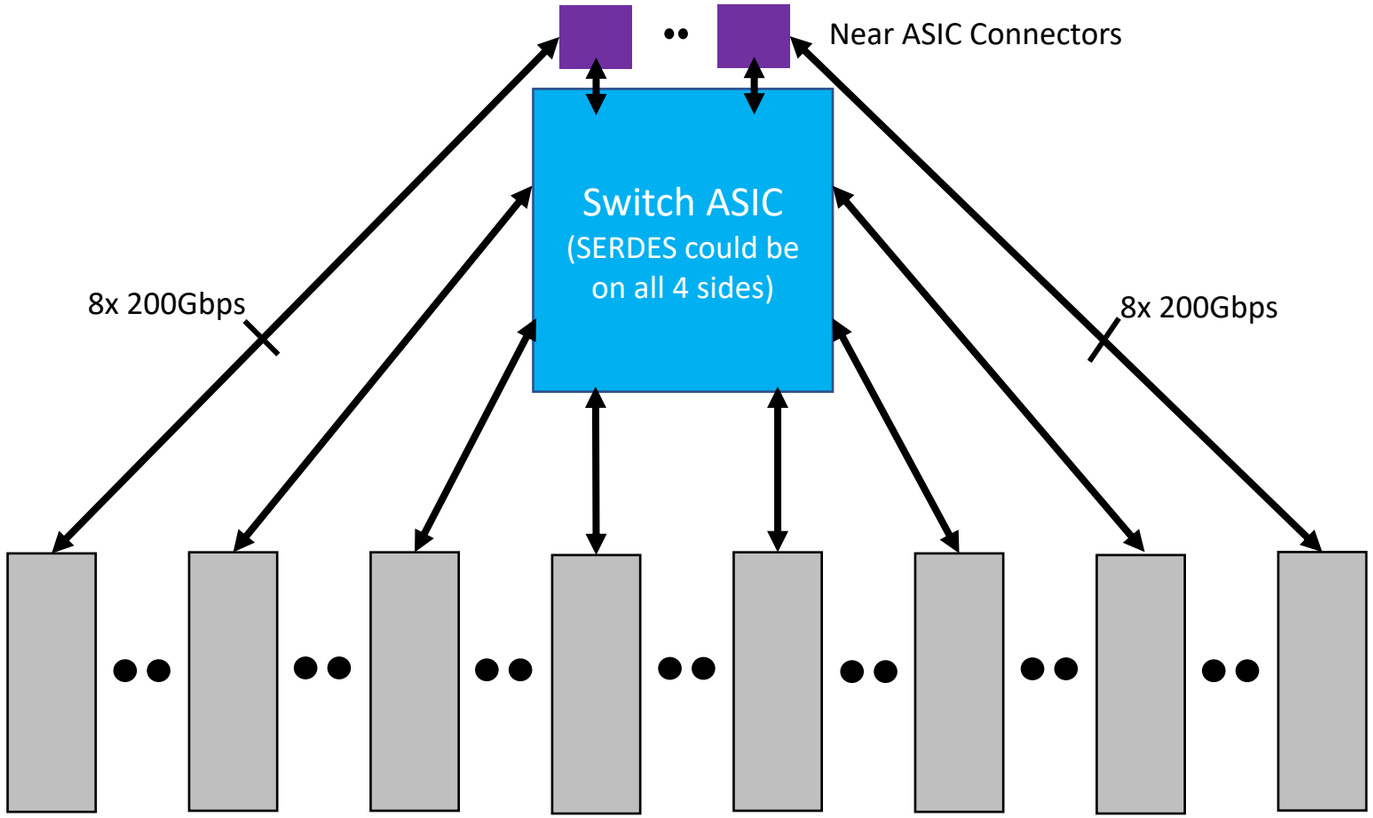
AUI Case: Cabled Host





64x1600G cages (e.g. QSFP-DD1600)

AUI Case: Hybrid (PCB, Cable) Host



64x1600G cages (e.g. QSFP-DD1600)

Loss estimates for All AUI types considered

	End to End Loss (Bump _ Bump) in dB			
AUI Type	**100G PAM4**	**200GPAM4** *(RS544 FEC)*	**200G PAM4** *(RS576 FEC)*	**Loss Increase with RS576 FEC**
Cabled host	*20.88*	*36.26*	*38.52*	*2.26*
PCB host	*25.70*	*43.51*	*45.30*	*1.79*
Hybrid host	*22.63*	*39.17*	*40.56*	*1.39*

Benefits offset by increased bit rate
(see next slide)

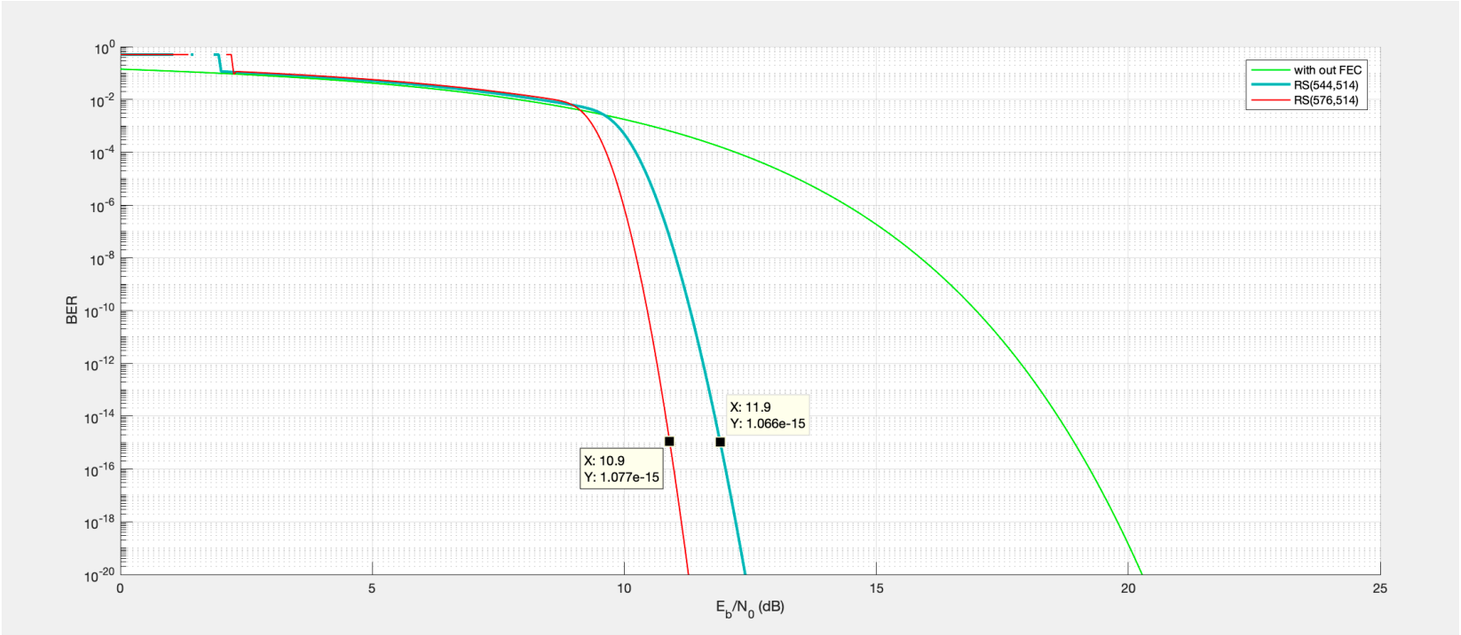
Other AUI types evaluated :

- ✓ Near Package Optics (NPO)
- ✓ Near Package Copper (NPC)
- ✓ Co-Packaged Optics (CPO)
- ✓ Co-Packaged Copper(CPC)

General trends were that the xPC had similar C2M channel losses (36-38 dB) and xPO had reduced channel losses

Observations

- FEC
  RS(544,514) vs RS(576,514) Coding Gain for PAM4
- Link Training



- Stronger FEC may not be helpful- as the coding gain from stronger FEC may be lost due to speed up required in data rate to handle additional overheads - resulting in higher channel losses
- For 36-38 dB loss range Coding Gain from the RS FEC (544,514) may only be sufficient to cover one segment of the channel . Under these conditions **Segmented FEC** is a better approach to meet overall channel performance
- Enabling the option of Link Training and Precoding for C2M interfaces would be of a great value

Summary and Observations

- The 200Gb/s generation needs to consider both NIC and Switch
- Reviewed Loss estimates for different AUI types that are suitable for all signal rates and Ethernet data rates
  - Part of 802.3 df adopted physical layer Objectives
- Analyzed implementations with anticipated chip packages consistent with large fixed box switch products
  - SerDes need to be able to work with a minimum of 36-38dB for Cabled Host
  - SerDes need to be able to work with a minimum of 43-45dB with PCB Host
- Call to Action : Power is the fundamental limit
  - We need to work to reduce the channels to save SerDes Power
    - System vendors to optimize system design to minimize loss
    - Substrate, PCB, Connector vendors must drive aggressively to improve performance
  - PCB Host created unusable channels without significant improvements
  - We need to consider serdes architecture that can address bump to bump loss at least up to 36-38 dB
    - Ideally improvements in materials can help shrink this, but it is too early to rely on furthermore improvements than already considered for 200G
  - Consider Enabling the option of Link Training (LT) and Precoding for C2M interfaces
  - Stronger FEC may not result in an improvement
  - Segmented FEC scheme should be strongly considered based on the review of the losses.