200G/lane Electrical interfaces – System implications

Adee Ran, Cisco

Liav Ben-Artsi, Marvell

Supporters

- Upen Reddy Kareti, Cisco
- Itamar Levin, Intel

This presentation is about…

These beasts

And that

Adopted Physical Layer Objectives & Nomenclature

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 200GAUI-1		Over 1 pair 200GBASE-CR1			Over 1 Pair TBD	Over 1 Pair TBD		
400 Gb/s	100 Gb/s							Over 4 Pair TBD		
	200 Gb/s	Over 2 lanes 400GAUI-2		Over 2 pairs 400GBASE-CR2			Over 2 Pair TBD			
800 Gb/s	100 Gb/s	Over 8 lanes 800GAUI-8	Over 8 lanes 800GBASE-KR8	Over 8 pairs 800GBASE-CR8	Over 8 pairs 800GBASE-VR8	Over 8 pairs 800GBASE-SR8	Over 8 pairs TBD	Over 8 pairs TBD		
	200 Gb/s	Over 4 lanes 800GAUI-4		Over 4 pairs 800GBASE-CR4			Over 4 pairs TBD	1) Over 4 pairs TBD 2) Over 4 λ's TBD		
	TBD								Over single SMF in each direction TBD	Over single SMF in each direction TBD
1.6 Tb/s	100 Gb/s	Over 16 lanes 1.6TAUI-16								
	200 Gb/s	Over 8 lanes 1.6TAUI-8		Over 8 pairs 1.6TBASE-CR8			Over 8 pairs TBD	Over 8 pairs TBD		

We need to approach all of this holistically

11 April 2022 IEEE P802.3df Task Force, Architecture and Logic Ad hoc Page 3

Source: dambrosia_3df_logic_220411a

Outline

- AUI C2M endpoints
- The ToR switch use case
- Loss budget
- Architecture implications
- Call for action

Evolution of C2M AUI endpoints

XLAUI, CAUI-10

- Interface between two simple devices ("CDRs")
- BER assumed negligible

CAUI-4, 25GAUI

- Loss isn't negligible, ISI became a problem
- Reference receiver has a CTLE
- BER specified as 1e-15

400GAUI-8, 50GAUI-1

- BER<1e-6 (protected by full-link FEC)
- Better CTLE in reference receiver

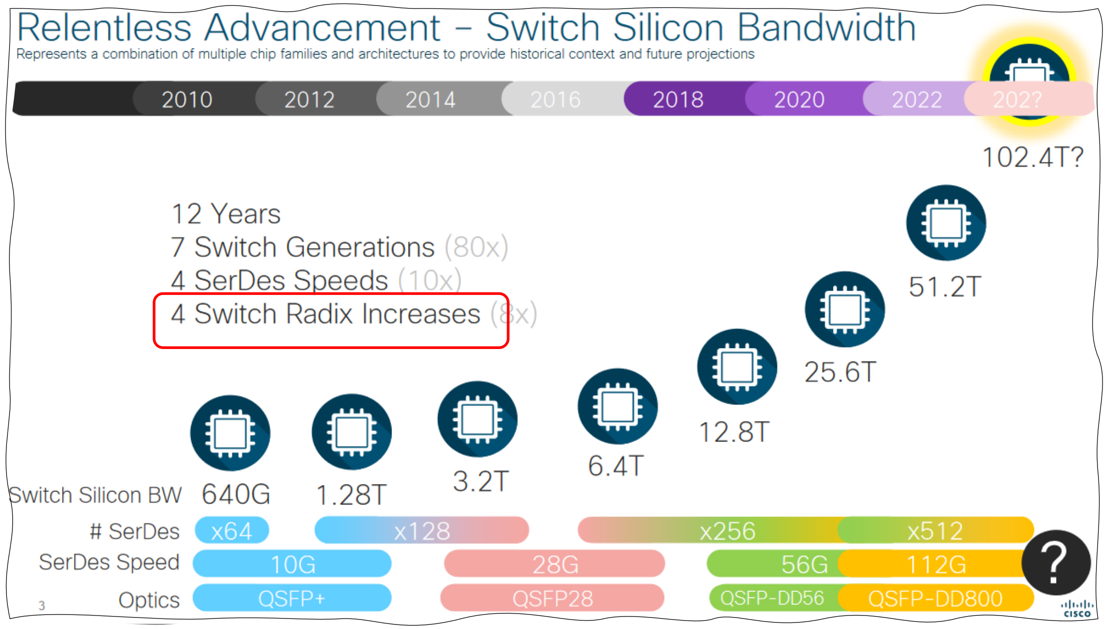
400GAUI-4, 100GAUI-1

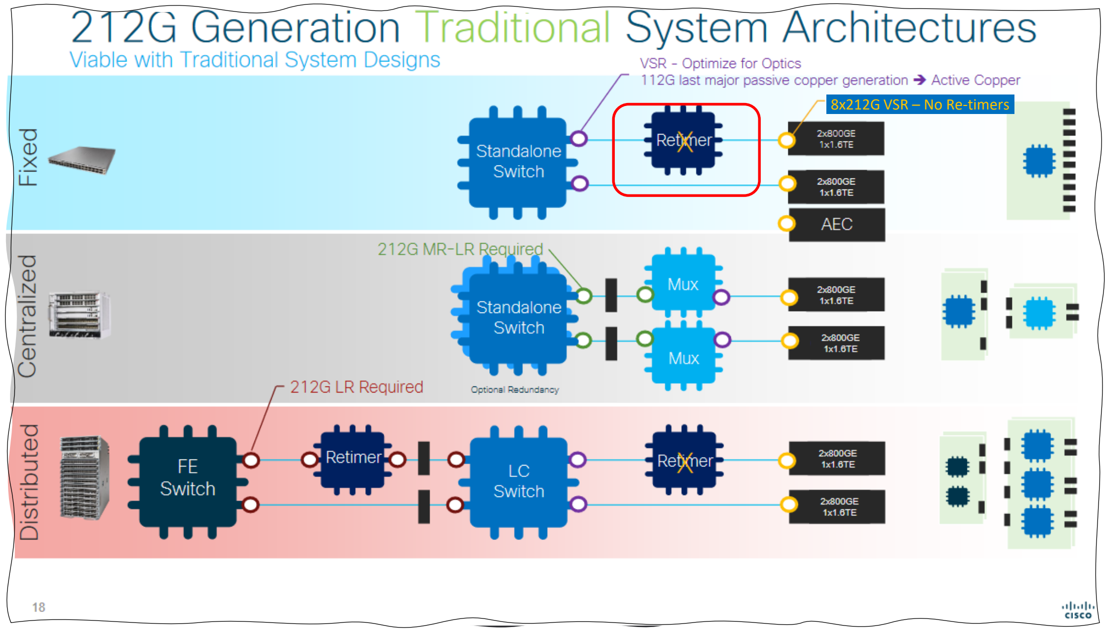
- CTLE is not sufficient – reference receiver includes a DFE
- Similar to an electrical PMA/PMD, but with a different specification methodology

1.6TAUI-8, 200GAUI-1

- **Question #1: what will it look like?**

Switch applications



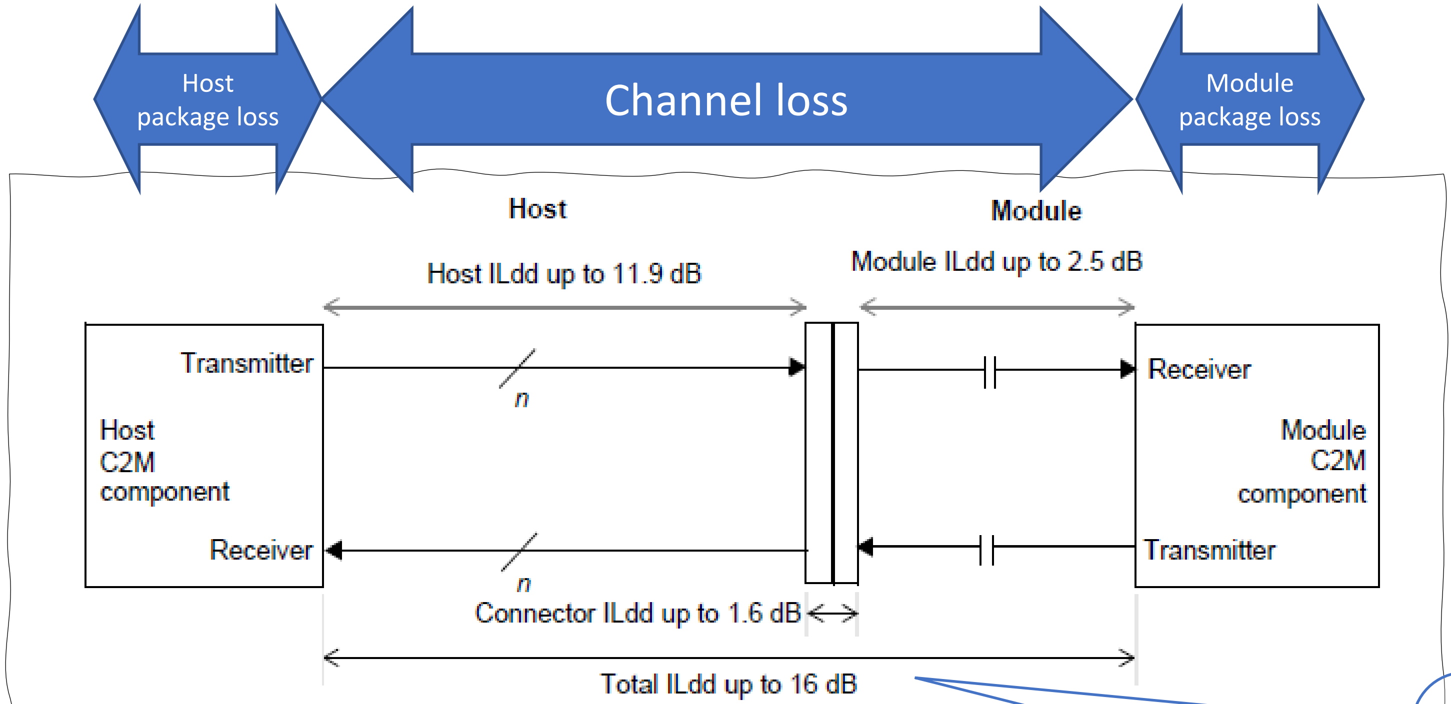


Source: chopra_b400g_01_210208 (slides 3 and 18)

Additional switch use cases

- There are other switch architectures
  - Co-packaged optics (CPO)
  - Near-package optics (NPO)
- **Not the scope of this presentation.**

C2M elements (from 802.3ck)



NOTE—The number of lanes *n* is equal to 1 for 100GAUI-1, 2 for 200GAUI-2, and 4 for 400GAUI-4.

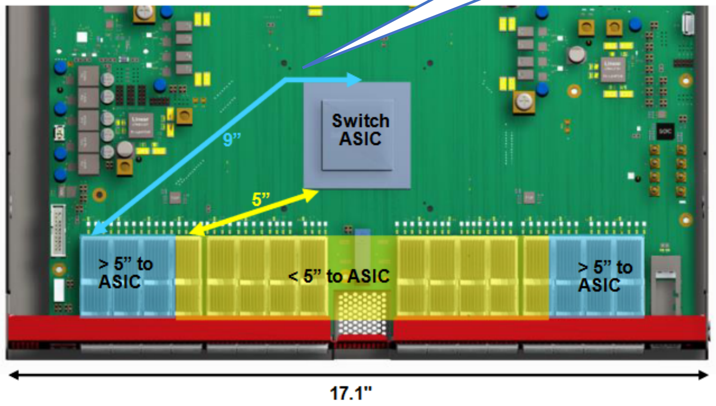
Figure 120G–2—100GAUI-1, 200GAUI-2, and 400GAUI-4 C2M insertion loss budget at 26.56 GHz

This number wasn't an easy decision

ToR switch geometry

Architectural changes to ToRs due to reduced physical VSR reach

- Hypothetical Example:
  - 25.6T, 256 x 100G
  - 1RU box, Single ASIC (ToR design profile, also used as virtual chassis, aka “Fixed Box”)
  - Can be used with all optical IO in a spine application (common practice today in hyperscale datacenters)
  - 32 x 800G module cages, all front panel IO
- Using Rosemont budget proposal from Jane Lim:
  - http://www.ieee802.org/3/100GEL/public/18_03/lim_100GEL_01b_0318.pdf
  - [~ 5” Host trace supported for VSR channels]
  - Approximately 12 / 32 module cages cannot accommodate the proposed host budgets (VSR or CR), requiring either intermediate retimers, or intra-box cabling



4 | 802.3ck Pittsburgh May 2018

BROADCOM

9” (lower bound)
from switch package to connector pads
+
Large switch ASIC package

This application influenced the Annex 120G specifications, which assume a ball-to-ball IL of **16 dB** @ 26.56 GHz.

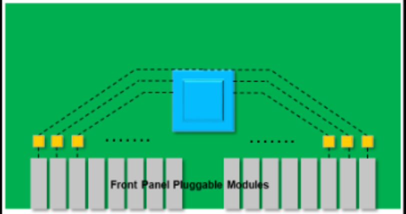
Assuming PAM4:
~32 dB @ 53 GHz?

Source: stone_3ck_01a_0518

Other options discussed in 802.3ck

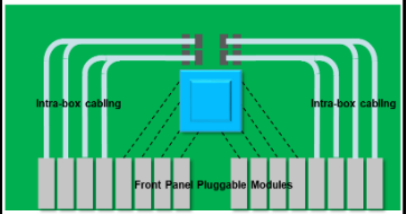
How Shorter Host Loss Maps to Possible "Universal Port" Solutions

Add retimers



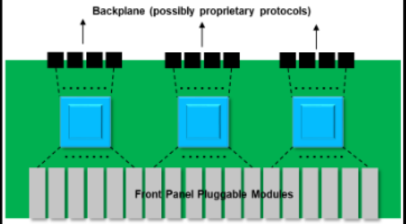
- Middle ports within proposed VSR budget do not require additional retimers
- Edge ports use additional retimers (shown in yellow) to enable longer overall host channels
- **Pros:** similar architecture to prior generation systems
- **Cons**: Cost and power of additional retimers

Intra-box cables



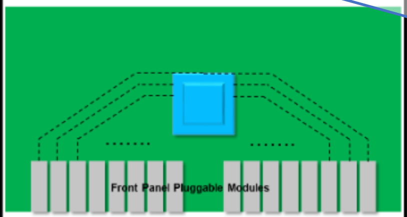
- Edge ports use intra-box cables to enable longer physical reach, but staying within proposed VSR budgets
- **Pros:** System does not incur cost or power of additional retimers, commonality with existing "retimerless" designs
- **Cons**: Increases mechanical complexity, may impact airflow, cost of cable and associated mechanicals

Multi-ASIC Linecards (Chassis Systems)



- Each ASIC can connect to fewer, closer module ports, which are supported within VSR proposed budget
- **Pros:** Similar "PHYless" design to current generation systems
- **Cons:** Does not address single ASIC "fixed box" designs forecast to be the dominant volume of the datacenter market

MR Capable Modules



- Enable modules with MR capability
- **Pros:** Similar "PHYless" design to current generation systems
- **Cons:** Requires MR support in modules, potentially increasing module power. Serdes may require training, and appropriate management support. Doesn't work on all ports with DAC – so not a universal port

5 I 802.3ckPittsburgh May 2018

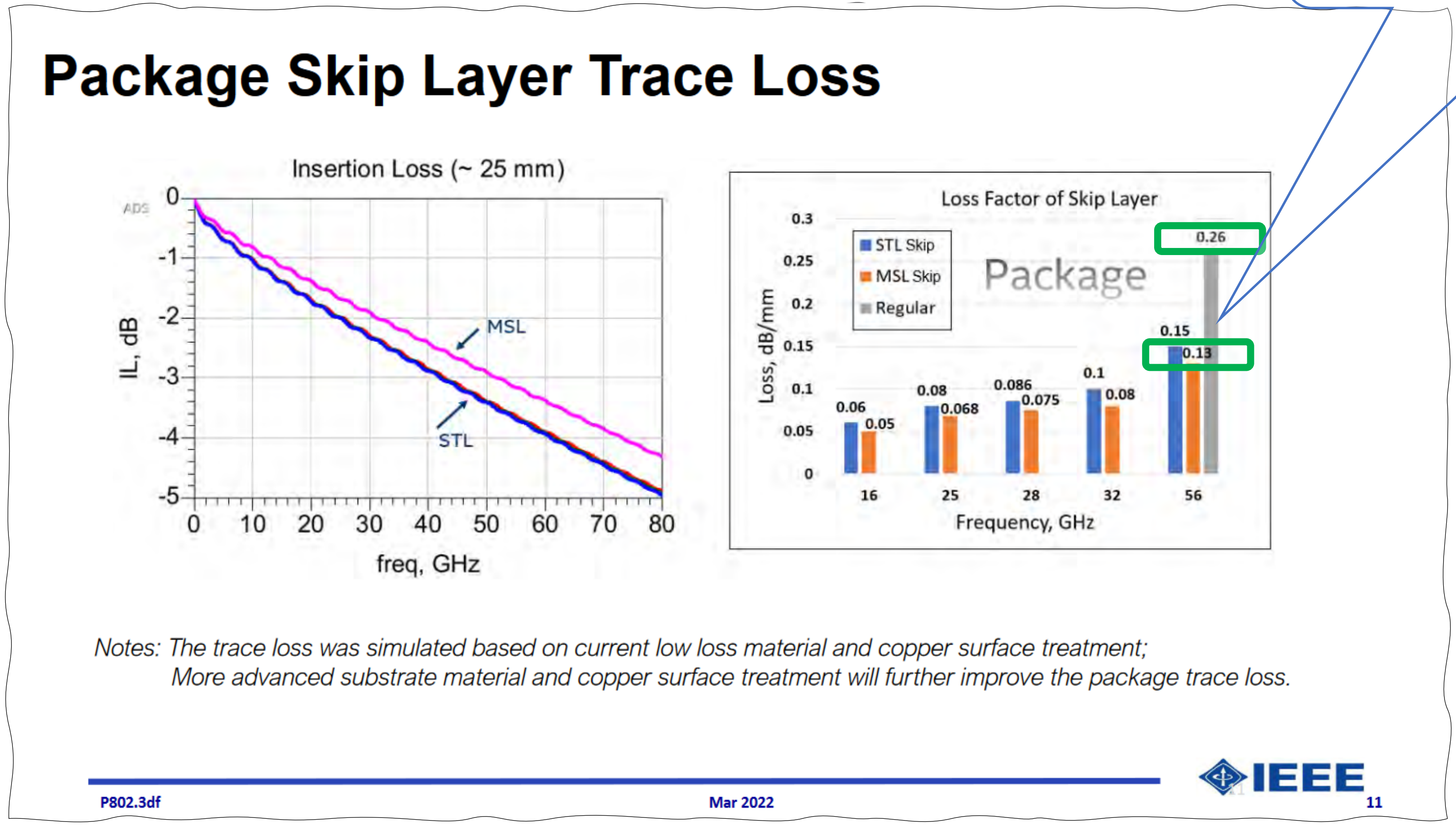
BROADCOM

This path was chosen (Annex 120G)

Source: stone_3ck_01a_0518

Package considerations

There are ways to reduce package trace loss to perhaps 0.13 dB/mm



Source: mli_3df_01_220316

But…

High-radix switch packages can't use skip layer and microstrips in all lanes.

These methods are typically used in the longer traces (e.g. 40 mm) to make them "look like" the reference package…

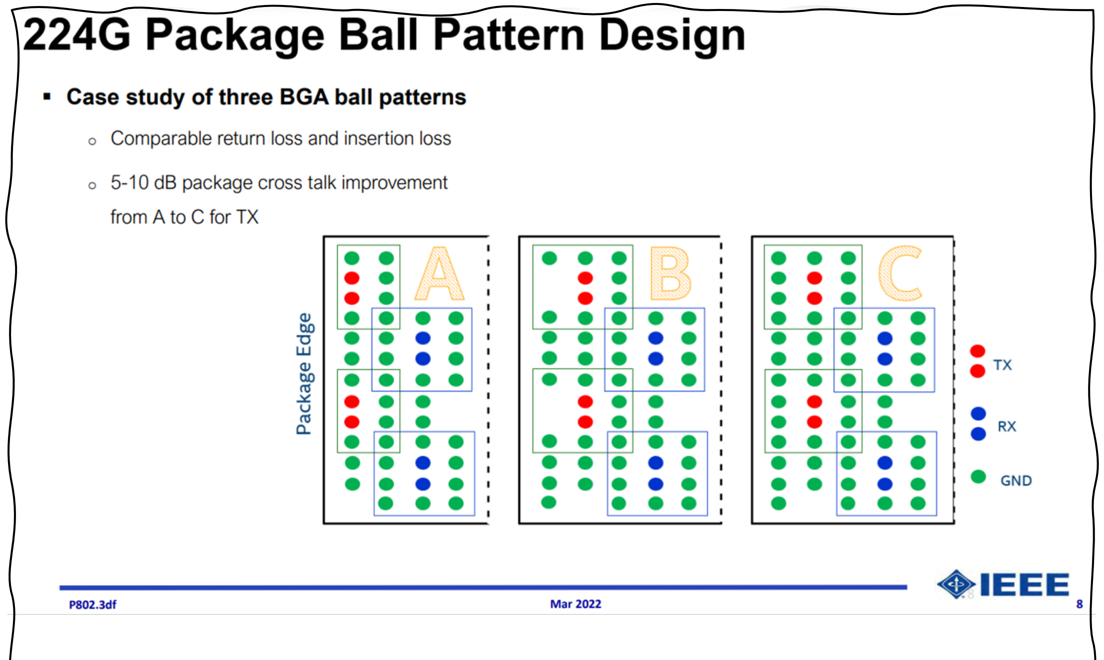
The 802.3ck COM reference package is based on "regular" trace of 31 mm

31*0.26 ≈8 dB @ 53 GHz
+ ~1 dB core via
⇒ 9 dB allocation for switch package?
1-3 dB for module package?

Ball pattern of a high-speed radix switch

Thought exercise:

Assume the minimum presented Tx/Rx separation, populate 256 lanes…

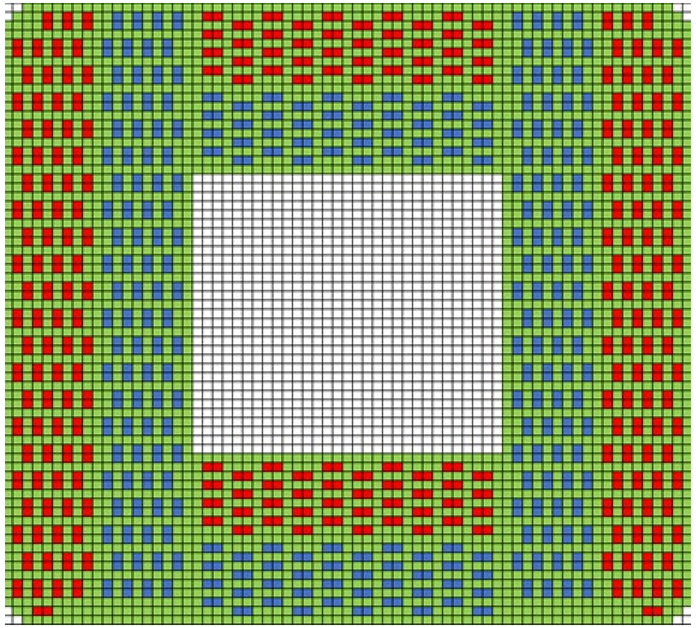


Source: [mli_3df_01_220316]

Just the AUI signals require a 69x69 grid (in practice, more are needed)

⇒ larger package than previously assumed (>75 mm square?)

⇒ longer traces



Host package and PCB recommendations

224G PAM4 Package Design Summary (slide 26)

- **Desired next generation package trace loss target for interpretation flexibility: 0.1 dB/mm at Nyquist frequency**
  - Skip-layer trace routing is required for mitigating the transmission loss
  - Low loss material and advanced copper surface treatment are required
- **0.8mm ball pitch is recommended (0.65mm or smaller preferred)**
- **Smaller ball size can further reduce discontinuities and package loss**
- **BGA ball pattern needs to be PCB breakout friendly and fully shielded**
- **Ground stitching via pitch < 1/10 wavelength along TX/RX traces and < 1/4 wavelength everywhere else in the vicinity of the 224G channel routing are required**

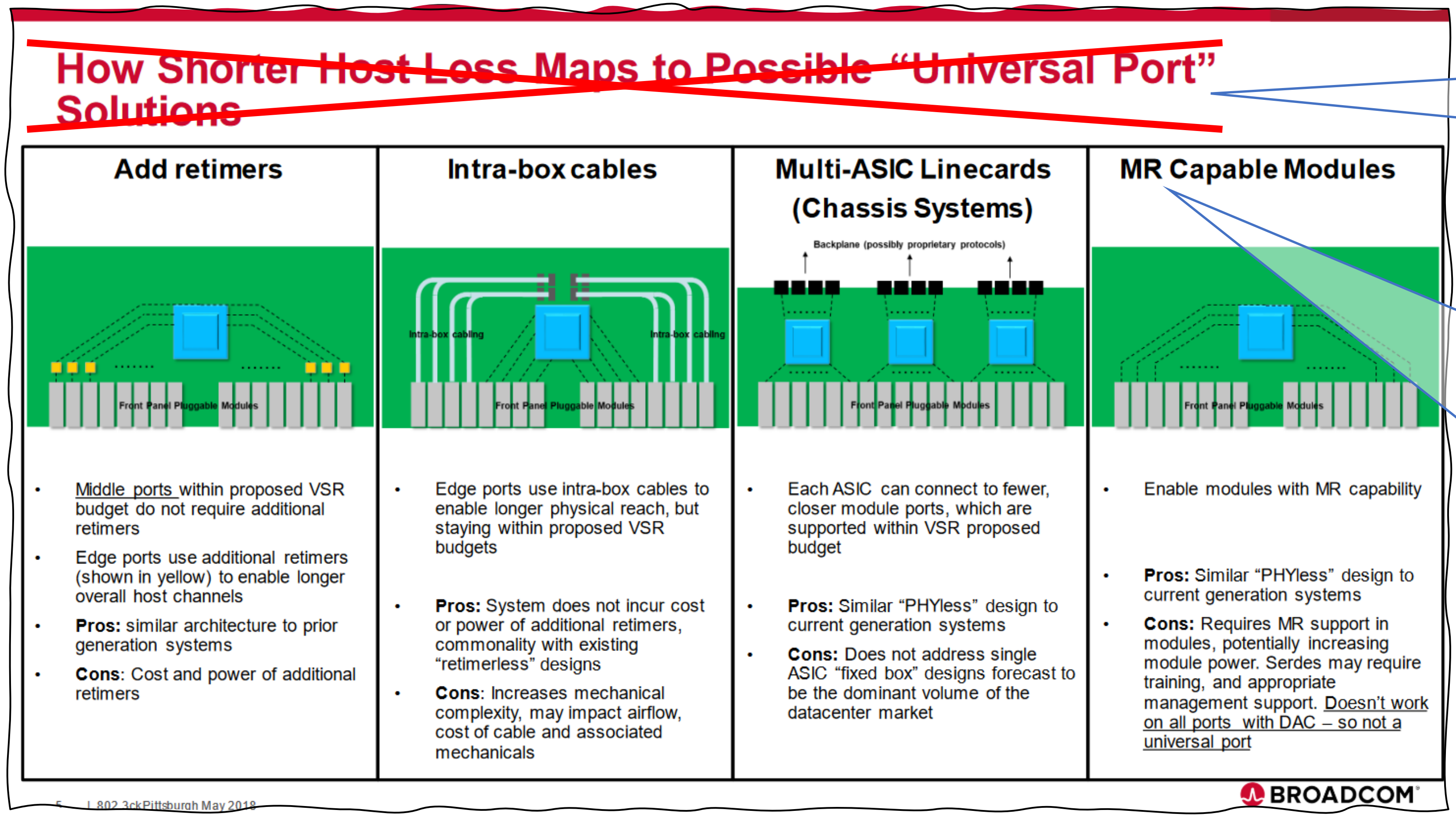
With the density required for high-radix switch ASIC package and PCBs, these design recommendations may not always be met

224G PAM4 PCB Design Summary (slide 27)

- **Desired next generation PCB trace loss target for interpretation flexibility: 1 dB/inch at Nyquist frequency**
  - Skip-layer trace routing is required
  - Ultra low loss material is required
  - HVLP copper surface treatment is required
- **PCB via stub length < 8mil is required**
- **Well controlled process variation of Dk, Df and dielectric thickness is required**

Source: mli_3df_01_220316

What about 802.3df?



Source: [stone_3ck_01a_0518]

At 200G/lane even an optical-only port is challenging.
"Universal port" may be possible with active electrical cables…

If we just sum the maximum numbers at 53 GHz:

32+9+3 > 40 dB – more than the traditional "Long Reach"…?

Other methods – retimers, cables – may be needed in some of the links

With PAM4, 30-35 dB end-to-end seems feasible

⇒ Question #2: What are the channel assumptions for 200G/lane C2M?

FEC architecture implications

- Achieving BER<<1e-4 with 200G/lane AUI isn't a safe assumption.
  - As mentioned in rabinovich_3df_01a_220224, even a relatively "easy" channel does not reach that goal. More so with switch AUI of 30-35 dB.
  - Assuming the RS(544,514) (KP FEC) for the AUI FEC, as suggested in gustlin_3df_logic_220411, its full correction capability will likely be required for one end of the link.
- As it seems:

End-to-end	❌
Encapsulated with "imperfect" outer (optical) FEC (inner end-to-end FEC corrects some "optical" errors)	❌
Encapsulated with "perfect" outer (optical) FEC, inner end-to-end FEC protects AUIs on both ends	🤞
Segmented	✔️

Architecture/Holistic approach

- As stated in dambrosia_3df_logic_220411a, we should also consider cases with more than one AUI on one or both sides.
- Our options are:

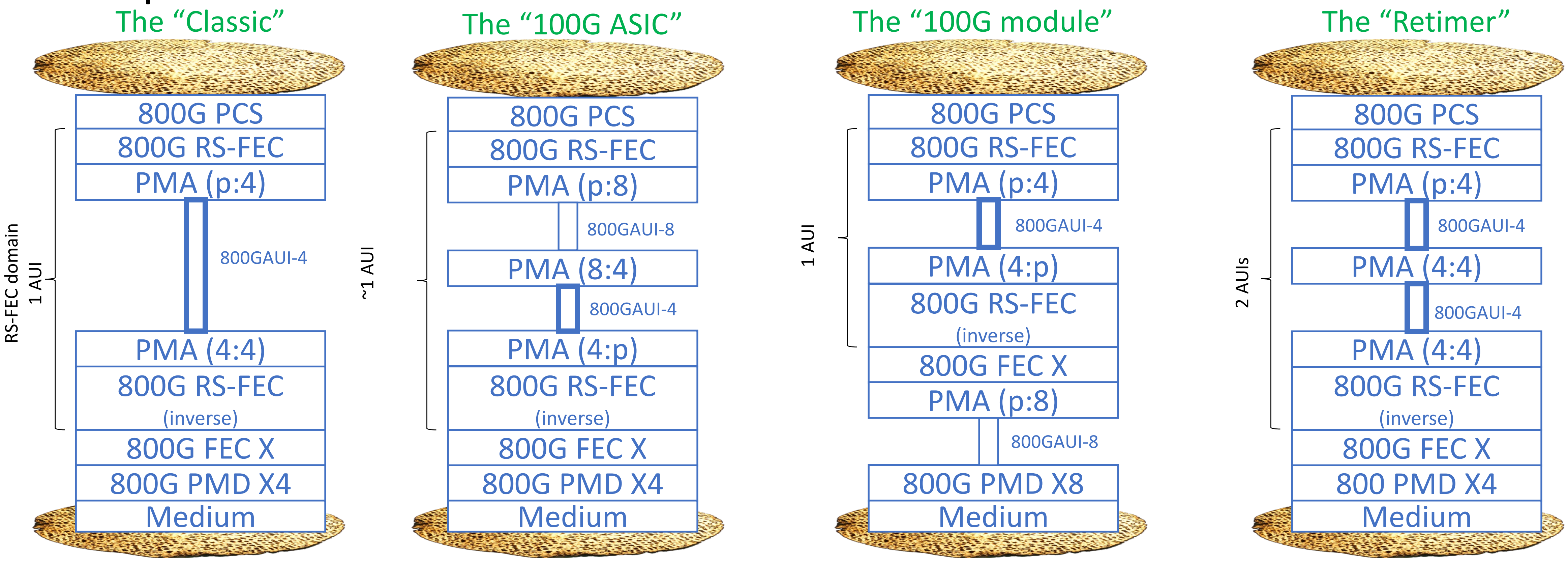


Image adapted from https://commons.wikimedia.org/wiki/File:Shmura_Matzo.jpg

Implication of segmented FEC

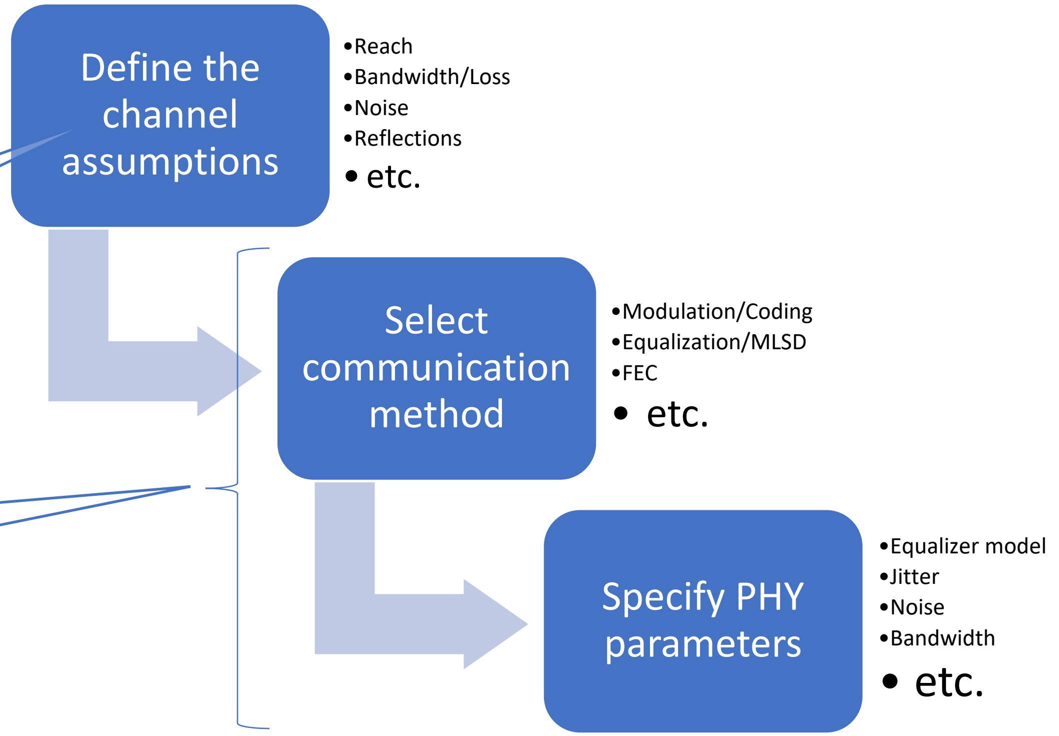
- Frame loss is a result of uncorrectable codewords on either of the FEC segments
  - These events are independent of each other, so easy to analyze and monitor
- Uncorrectable codeword ratio (UCR) of FEC-protected AUIs should be allocated from the total budget
  - From the UCR of the FEC-protected AUIs, we can calculate the maximum pre-FEC BER as we had in previous projects
  - More than one AUI can be in one FEC domain
  - **For now, assume the pre-FEC BER is 5e-5 to support two AUIs**
- Given maximum BER and channel assumptions, we can start analyzing reference Tx and Rx parameters…
  - So we need to define our channel assumptions!

Thoughts about our process

- How do we standardize an electrical interface?
- Tried-and-true way, in many projects

We need to agree on this part (question #2) first

Then answer question #1 (although some parts may already be assumed)



Partial answer to question #1

(200G/lane AUI endpoints)

- Authors' opinion:
  - At least as complex as reference Rx/Tx of 100G electrical PMDs
    - Including, e.g., a strong equalizer
  - BER similar or slightly better than 100G electrical PMDs
    - Assuming segmented FEC architecture with KP FEC
  - Every AUI segment must be protected by FEC; FEC domain (between encoding and decoding) spans at most two adjacent 200G/lane AUIs

Call for action

- We need a clear process of adopting a loss budget
  - Proposals in terms of lengths and IL (assuming PAM4); detailed results with S-parameters would help
  - Explain the targeted application (switch, NIC, other)
- Until we adopt a loss budget we can’t make any decisions on device electrical parameters (including reference Tx/Rx) or even modulation
  - Proposals in this area may be premature
- Let’s not intermix these steps (e.g. run COM analysis on channels before loss budget is adopted)

Questions? Comments?

Thank you