100Gb/s Electrical Links System View

Mark Gustlin – Xilinx
Beth Kochuparambil – Cisco
Kent Lusted – Intel
Gary Nicholl - Cisco
David Ofelt – Juniper
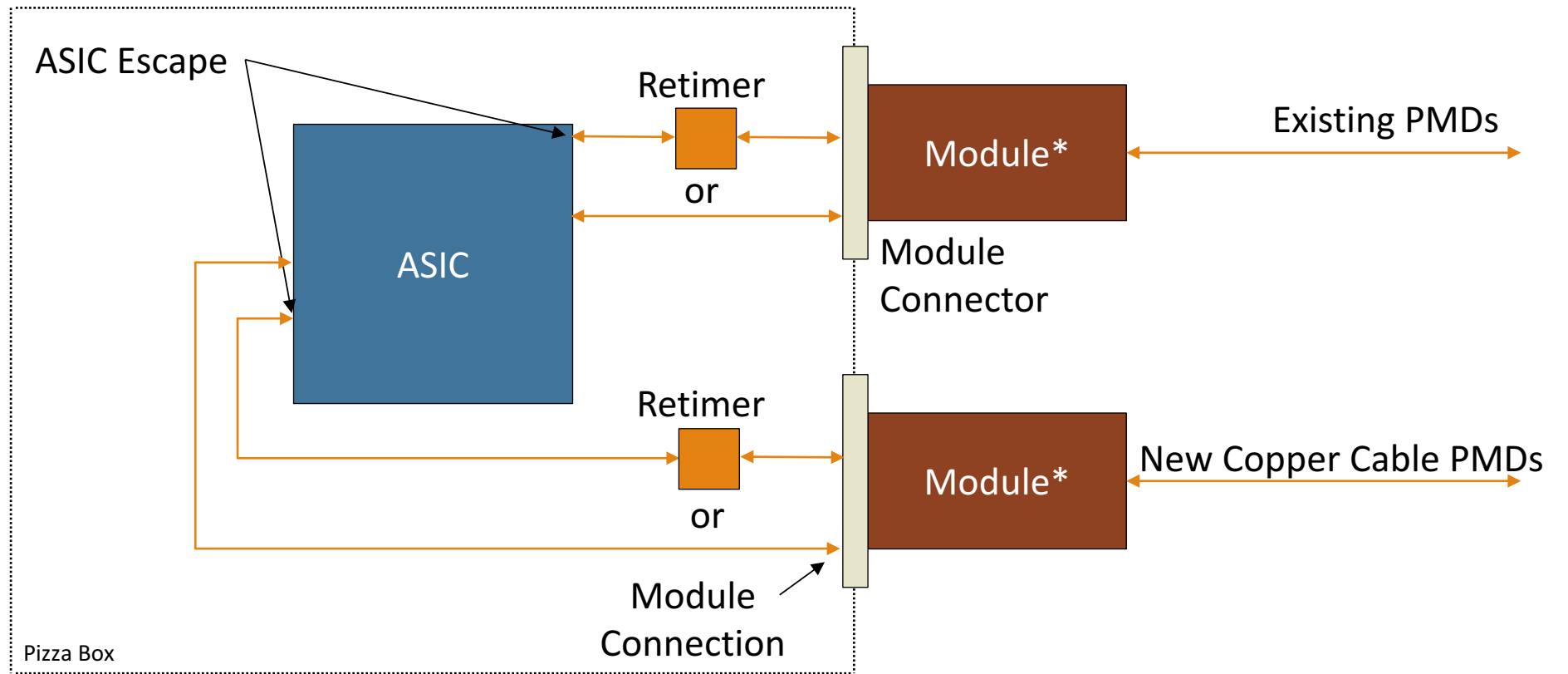Rob Stone - Broadcom

Ground Rules

This presentation is about motivating use cases and the issues around each

- Expect other presentations in response- providing:
  - objectives
  - technical feasibility (TF)
  - economic feasibility (EF)
  - broad market potential (BMP)

Some traditional objectives (ex: chip to module) may have different issues at each end of the link
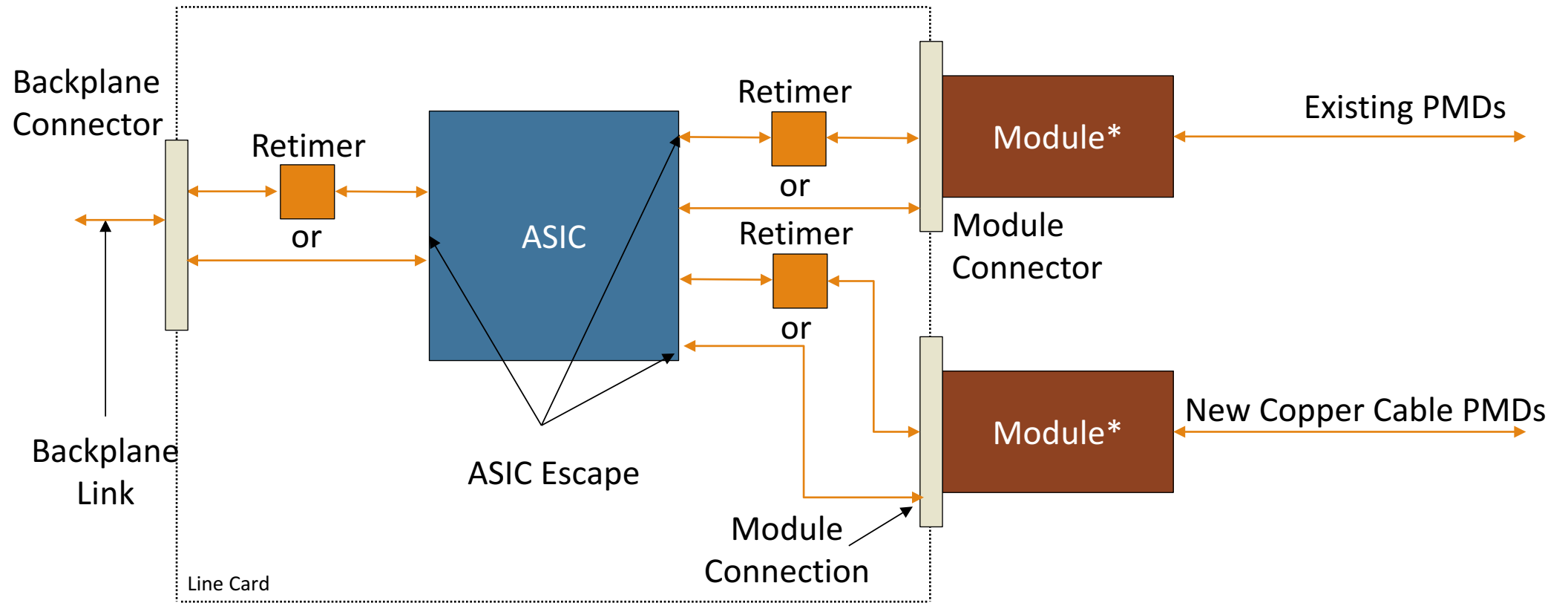
System Overview

System Overview – Pizza Box

ASIC Escape

Retimer

or

Module*

Existing PMDs

Module
Connector

ASIC

Retimer

or

Module*

New Copper Cable PMDs

Module
Connection

Pizza Box

* : Module can be traditional front-panel module or OBO

System Overview – Line Card

* : Module can be traditional front-panel module or OBO

General System Observations

For line-card based systems- a faster line card is less useful:
- If there isn't backplane bandwidth to support it
- If there isn't enough faceplate density (or board surface area) to get the bandwidth out of the box

For standalone systems – a faster ASIC can still be useful
- No increase in faceplate density, but will increase total bandwidth and crossbar radix

OIF 100Gb/s SERDES work:
- Are the current channels and objectives appropriate for the work here?

Boards are big – getting from the center to modules in the front corners can be tricky
- OBO don't necessarily help that much, since they can't all be packed right next to the ASIC

Copper Cable

Copper Cable

What are achievable reaches for:

◦ Passive copper cable?

◦ Active copper cable?

What are system implications to achieve these reaches?

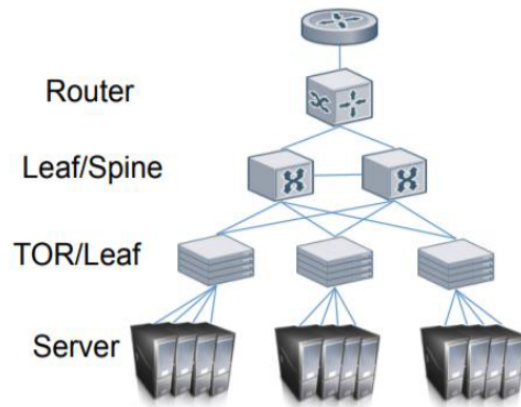What are the end-user implications of these reaches

◦ What common data center architectures work/don't' work?

What is the cost delta for active –vs- passive cable

◦ Impacts broad market potential for Cu cable if the cost approaches optics

Data Center Architecture Influence

Why Copper Cable??

Router

Leaf/Spine

TOR/Leaf

Server

Interconnection Volume

- Four sections per colo & multiple colos (≥ 4) per data center
- Volumes below are per section (except DCR to Metro)

A End	Z End	Volume	Reach (max)	Medium	Cost Sensitivity	Market Space
Server ‡	TOR	10k – 100k	3 m	Copper	Extreme	
TOR	LEAF	1k – 10k	20 m	Fiber (AOC)	High	LAN
LEAF	SPINE	1k – 10k	400 m	SMF	High	
SPINE	DCR	100 – 1000	1,000 m	SMF	Medium	Campus
DCR	Metro	100 – 300	10 - 80 km	SMF	Low	WAN

‡ Server-TOR links may be served by breakout cables

9   IEEE 802.3 400G Study Group - November 2013

Need to study highly cost sensitive and very short reach market.

Source: Brad Booth, Microsoft http://www.ieee802.org/3/400GSG/public/13_11/booth_400_01a_1113.pdf
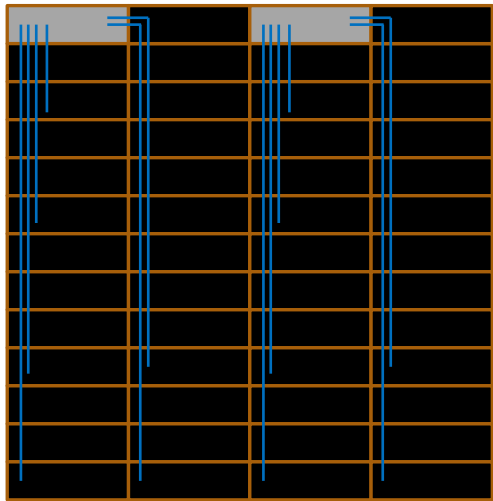
*Note that data is from 2013, however data center architecture hasn't drastically changed in recent years

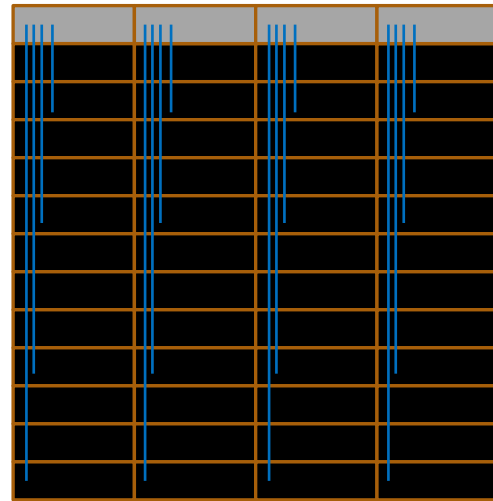IEEE 802.3 November 7, 2017, Consensus Building

9

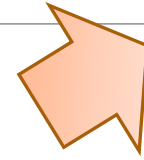Source: http://www.ieee802.org/3/cfi/1117_3/CFI_03_1117.pdf
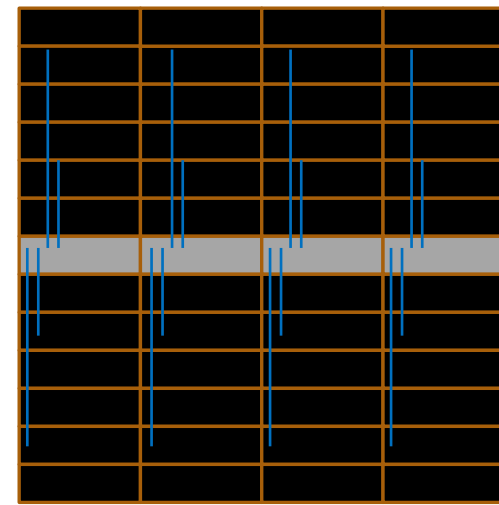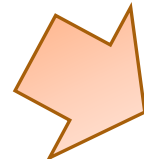
Copper Cables in Data Center Racks

"Inter-rack Switch"
5m DAC reach – 802.3by
(25GBASE-CR)

→

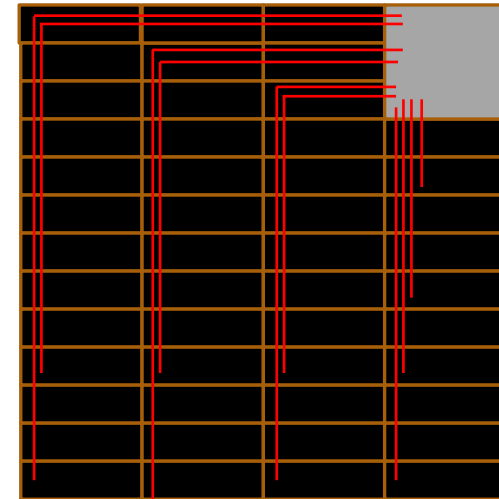"Intra-rack Switch"
3m DAC reach – 802.3cd
3m DAC reach – 802.3by
(25GBASE-CR-S)

?

"Middle of Rack Switch"
1-2m? DAC reach

What length is too short?

Are FEC or host loss changes needed?

"End of Row Switch"
No DAC reach

What is the impact to industry?

▬▬▬ = Cu cable
▬▬▬ = Fiber cable

⬜ = switch   ⬛ = server

Module Connection

Module Connection (AUI)

Backwards compatibility issues (see following slides)

What are the system implications to make 100Gb/s AUIs work?
◦ retimer proximity/universality, PCB material, fly-over cables required, etc, etc?
◦ Do these requirements affect EF or BMP?

What are the economic tradeoffs between redoing PMD budgets and spending significantly more on the system?

How different is the module TX/RX SERDES feature set than the ASIC TX/RX SERDES feature set?

Systems will need to support modules that run many rates of Ethernet
◦ Hard to have loss budgets that are different per rate when the same module can operate at different rates using breakout.

Backwards Compatibility Issues

Will we be able to or want to support existing (or soon to be existing) PMDs with a new AUI?
- Assumption is we will want to support 8023.bs/cd PMDs, especially the DR versions since those are 100G per lane

Issues include:
- FEC partitioning budgets
- FEC choice
- Latency
- Backwards compatibility

Solution Space to investigate:
1. Make 100Gb/s chip-to-module links work with 802.3bs and 802.3cd electrical FEC budgets
2. Change end-to-end link budget of 802.3bs and 802.3cd PHYs to allocate more errors to the electrical links
3. Terminate FEC in module and regenerate FEC for the wire (segment-by-segment FEC)
4. Add a (hopefully lightweight) wrapper FEC to protect 100Gb/s electrical links

200/400GbE "Legacy" Module Based PHYs

The majority of these PHYs are based on 50Gb/s per lane technology

- Supporting them with 100G electrical interfaces requires a reverse mux in the module
- Bit muxing is supported for changing lane widths

Only 400GBASE-DR4 uses 100Gb/s lane technology

FEC is end-to-end, RS(544,514)

PMD portion of the BER end-to-end budget is always 2.4×10^{-4}

200/400GE PHYs	Technology/Reach	FEC	FEC Coverage
200GBASE-DR4	4 lanes SMF, 500 m reach	RS(544,514)	End to End
200GBASE-FR4	4 WDM SMF, 2 km reach	RS(544,514)	End to End
200GBASE-LR4	4 WDM SMF, 10 km reach	RS(544,514)	End to End
400GBASE-SR16	16 lanes MMF, 100 m reach	RS(544,514)	End to End
400GBASE-DR4	4 lanes SMF, 500 m reach	RS(544,514)	End to End
400GBASE-FR8	8 WDM SMF, 2 km reach	RS(544,514)	End to End
400GBASE-LR8	8 WDM SMF, 10 km reach	RS(544,514)	End to End

100GbE "Legacy" Module Based PHYs

These PHYs are based on 10, 25, 50 and 100 Gb/s per lane technology
- ◦ Supporting them (with the exception of DR) with 100G electrical interfaces requires a reverse mux in the module

Only 100GBASE-DR uses 100Gb/s lane technology

FEC is end-to-end for some (SR10, LR4 and ER4 don't require FEC at all)

There are many derivative PMDs in the industry specified by MSAs etc.
- ◦ They reuse the IEEE PCS/FEC

PMD portion of the BER end to end budget is $2.4x10^{-4}$ for the DR PHY

| 200/400GE PHYs	Technology/Reach	FEC	FEC Coverage
100GBASE-SR10	10 lanes MMF, 100 m reach	No FEC	N/A
100GBASE-SR4	4 lanes MMF, 100 m reach	RS(528,514)	PMD only
100GBASE-LR4	4 WDM SMF, 10 km reach	No FEC	N/A
100GBASE-ER4	4 WDM SMF, 40 km reach	No FEC	N/A
100GBASE-SR2	2 lanes MMF, 100 m reach	RS(544,514)	End to End
100GBASE-DR	1 lane SMF, 500 m reach	RS(544,514)	End to End

Opt 1: Legacy BER Partitioning for 802.3cd/bs PHYs

For a complete PHY the error partitioning below gives us a BER of 1×10^{-13} (or equivalent frame loss ratio)

If we keep the C2M segment at a BER of 1×10^{-5} or better for 100G per lane interfaces, then we can reuse the cd/bs PMDs as is

- Assuming we were to keep using the RS(544,514) FEC
- Allows for both module and host backwards compatibility (see right side)
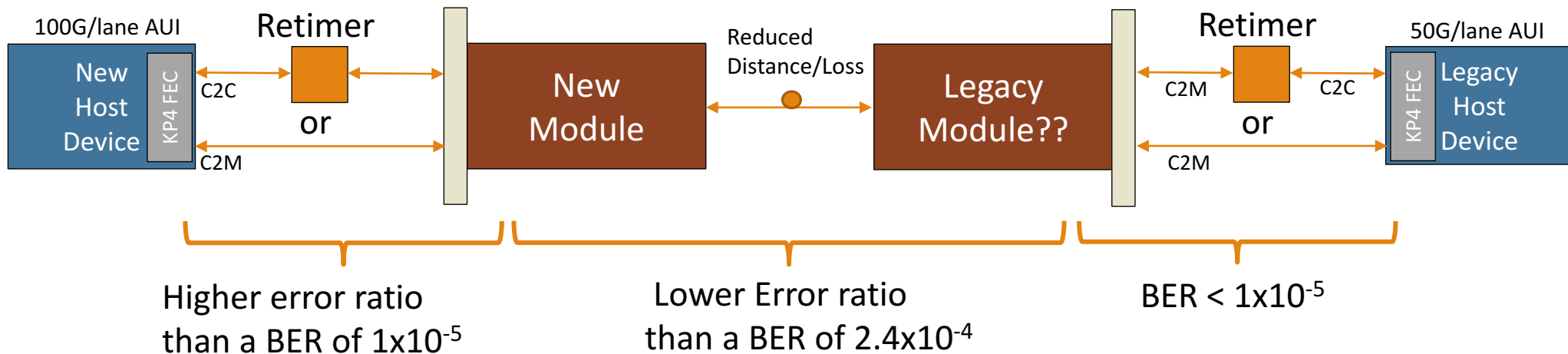
Otherwise we can't directly support these existing PMDs
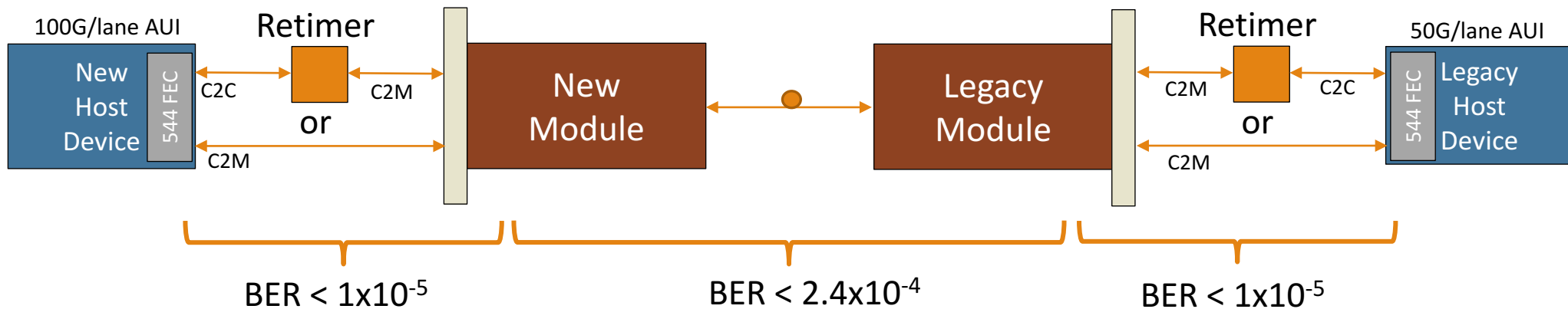
Opt 2: New BER Partitioning for 802.3cd/bs PMDs

Would need to redefine the PMD BER requirements (for example for 100GBASE-DR)

This would create new PMDs definitions, not compatible with the 802.3cd/bs PMD
- Might be possible to allow interop at reduced loss for legacy modules (might add confusion to the market?)
  - But might not be able to depending on the error floor?
- Allows for host backwards compatibility (see right side)

100G/lane AUI | Retimer | Reduced Distance/Loss | Retimer | 50G/lane AUI

New Host Device — KP4 FEC — C2C / C2M — or — New Module — Legacy Module?? — C2M / C2C — or — KP4 FEC — Legacy Host Device

Higher error ratio than a BER of 1×10^{-5}

Lower Error ratio than a BER of 2.4×10^{-4}

BER < 1×10^{-5}

Opt 3: Segment by Segment FEC

The 100G per electrical lane FEC will have to be terminated in the module, so no dependencies exist

- Same (RS(544,514)) or different FEC could be used for electrical interface
- Adds latency to the span and complexity/power to the modules
- Allows for both module and host backwards compatibility (see right side)

BER < open   BER < 2.4×10^{-4} BER < 1×10^{-5}
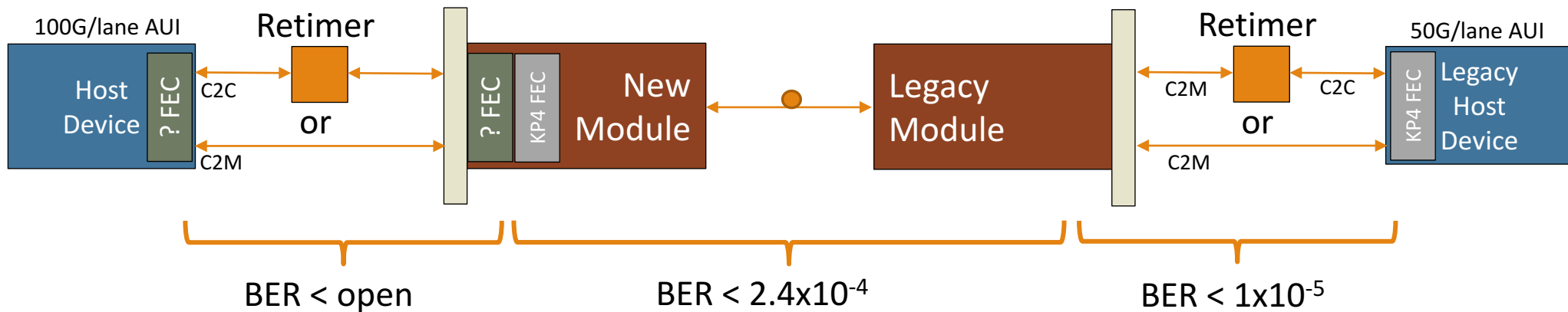
Opt 3: BER Partitioning for 802.3ba/bm PHYs

If there is interest in supporting these older PMDs, the FEC must be segment by segment

The 100G per lane FEC will have to be terminated in the module, so no dependencies exist
- Allows for both module and host backwards compatibility (see right side)

BER < open   No FEC or post FEC BER < 1×10^{-12} BER < 1×10^{-12}
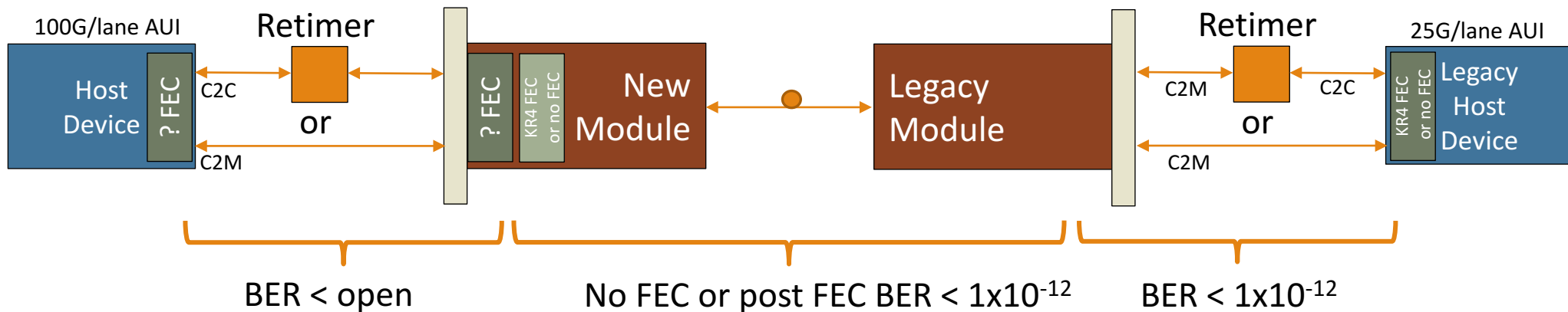
Opt 4: New FEC Wrapper

This option adds some additional FEC (wrapper) to the data

- ◦ Allows for both module and host backwards compatibility (see right side)
- ◦ This additional FEC is point to point across the AUI interface and terminated in the module
- ◦ Will add additional latency
- ◦ Will increase the data rate
- ◦ Add complexity/power to the module (and host device)
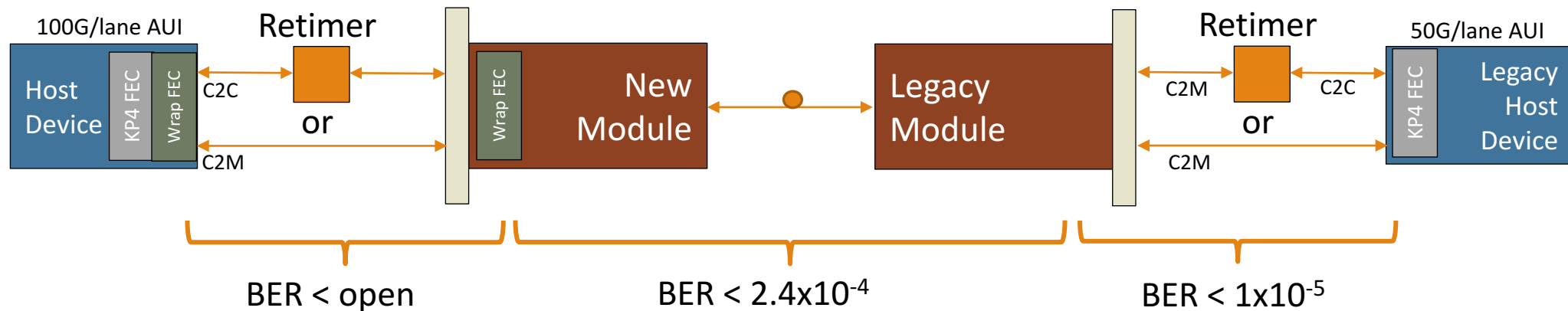
ASIC Escape

ASIC Escape

Next-generation ASICs want to (at least) double their bandwidth
- 50Gb/s generation have likely maxed out SERDES count per die/package
  - No next-generation product for some architectures without 100Gb/s SERDES

Are the ASIC SERDES universal (CR/KR/C2M/C2M) or are they just (say) C2C and retimers provide the other connectivity?

100Gb/s ASICs can use retimers to downshift to 50Gb/s AUIs (etc)
- No face plate density increase, but allows for ASIC bandwidth growth.

Switch Die Area Increasingly Consumed by I/O

Commercial Example: Ethernet Switch
- 128 x 25G Lanes
- 30% Chip Area is IO (shaded blue), 70% for other mission functions
- Grows to >> 30% in 50G IO generation!

Key requirement for 100G generation IO: preserve silicon area for value added mission functions
- Minimize Power
- Minimize Area (cost and device feasibility)
- PCS & FEC logic commonality (reuse where possible)
- "Balanced" serdes choice – area & power vs reach / capability
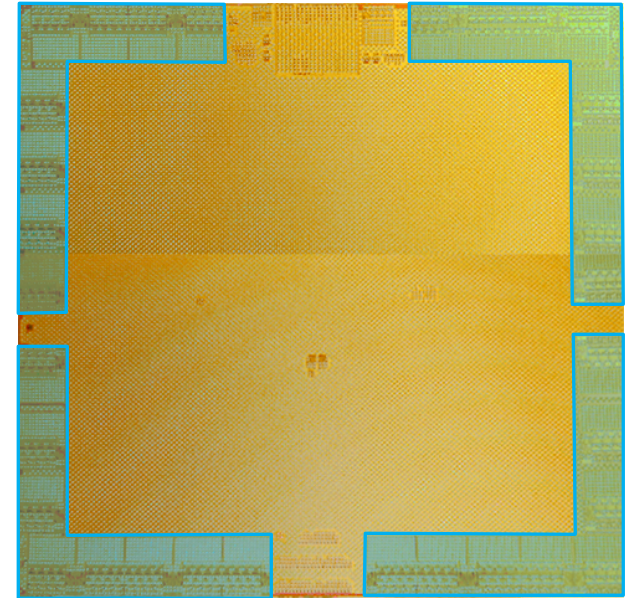
Package Design
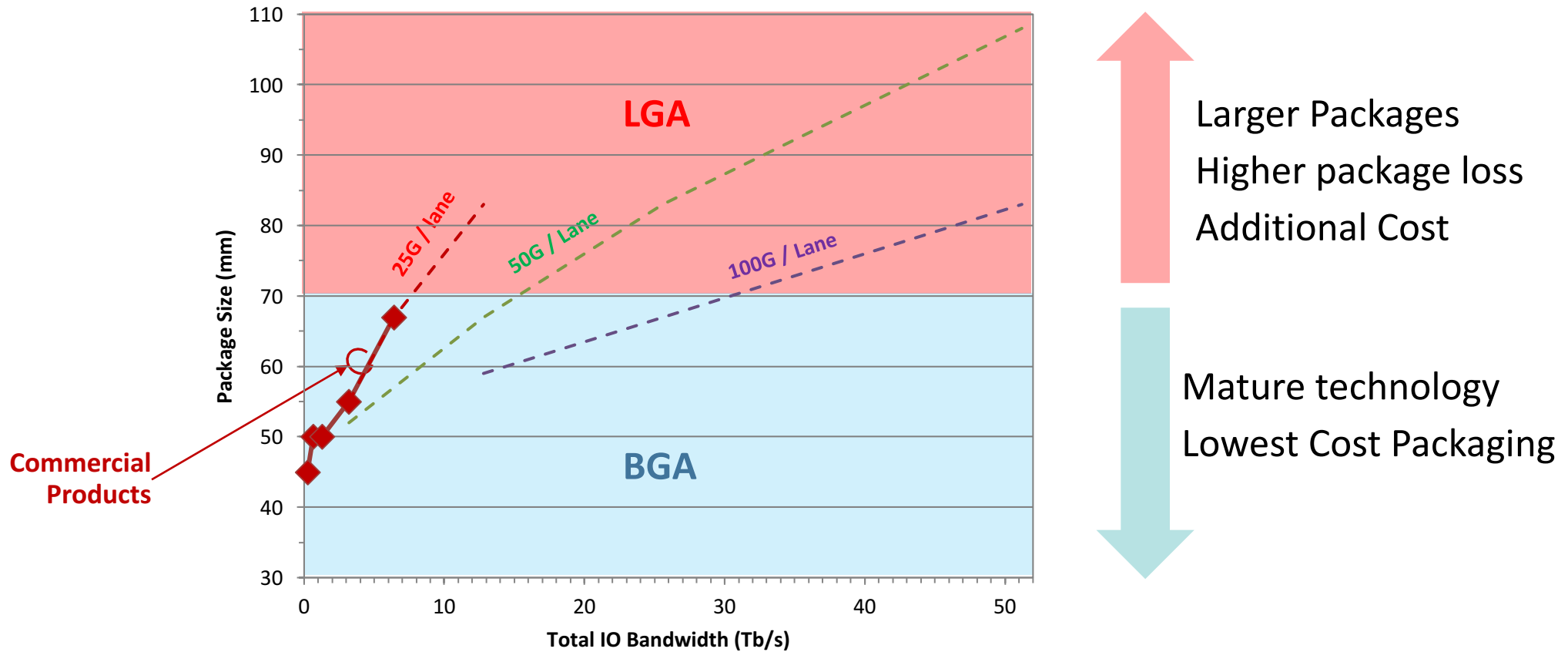
When is 100G electrical I/O required?

Driven primarily by switch package escape

Practical BGA limit is ~ 256 lanes in a 70 mm package, 1 mm ball pitch

Products announced already with 50G IO with this form factor (12.8T)

Switch processing capacity doubling every ~ 24 mo

100G IO required in products by ~ 2019 to continue the growth trend at a constant rate!

Leading Switch Capacity vs Year
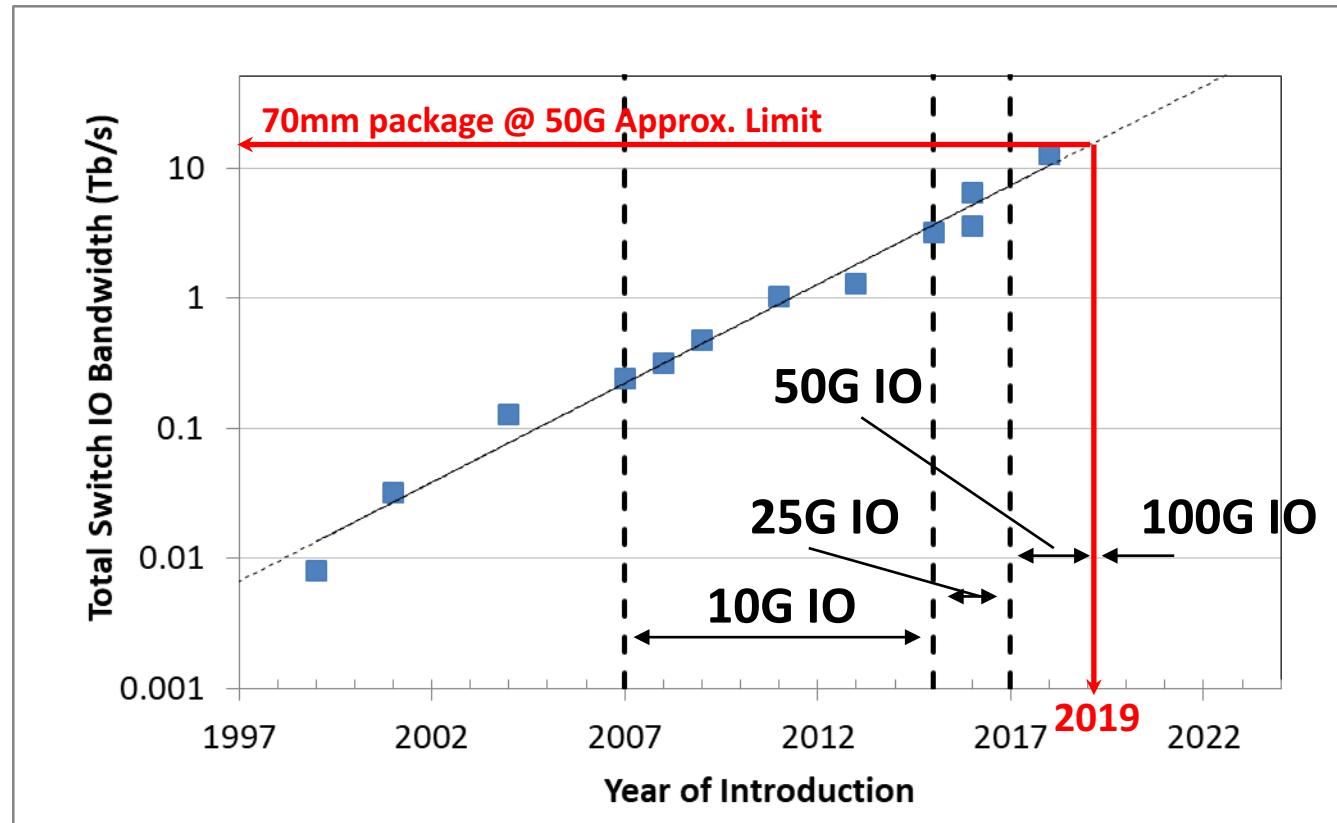
> 2019 ASIC requirements are expected to exceed
BW delivered by a conventional BGA with 50G IO

Non-retimed Feasibility Estimates

Interface	Architecture	Approximate Channel Loss (ball – ball, dB)	Feasibility Rank
Chip to Module (assumes 1RU system)	Conventional (10"PCB + Front panel module)	20^1	High
	Internally Cabled Host + Front-panel module	15^2 ?	High
	Mid-board Optical Module	$10 - 15^2$	High
Chip to Chip	Conventional (PCB + Mezz Connector)	40^1	Low
	Internally Cabled + Mezz Connector	~ $20 - 30$	Med
Backplane (KR)	1m Conventional PCB	60^1	Very Low
	Orthogonal	$20 - 40^2$	Med
	Cabled Backplane	$20 - 35^2$	Med
Copper Cable (CR)	Conventional DAC + PCB Host	60^1	Very Low
	Internally Cabled Host + DAC	~ $20 - 30$	Med

- C2M look achievable

- Chip to Chip, Backplane, Copper Cables require architecture change

1 – Projection - Scaled 2 x from 802.3cd existing channels
2 - Source: Nathan Tracy, TE Connectivity, DesignCon 2017 - CEI-112G: Considering Electrical Channels

Right sizing the IO standardization for switching applications

Not surprisingly, tall poles appear to be associated with the longer channels

More data and study required on both channels, as well as serdes capability

Initial 100G / lane electrical applications are likely to be associated with switch package escape, where a high power, larger area serdes will be prohibitive, and time to standardization is critical

Project scope implications?

Backplane

Backplane Considerations
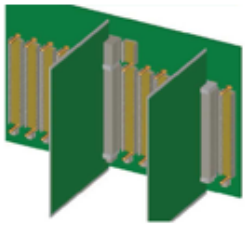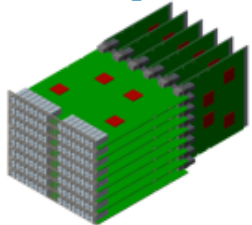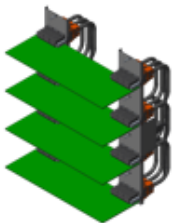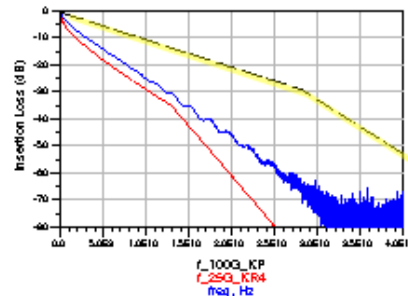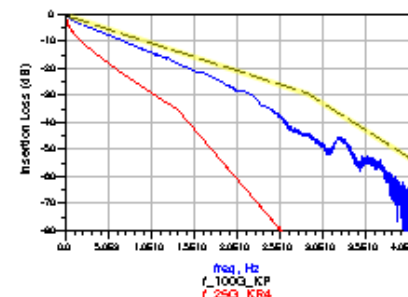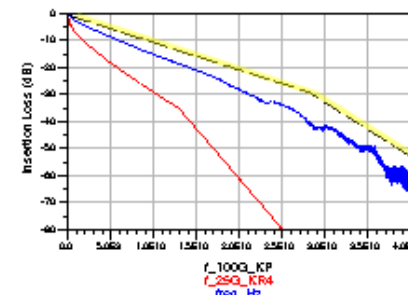
Line in the sand at 30dB at 28GHz
- Past project targets
- Need a starting point to discuss system routing needs

Traditional lengths struggle at this speed step.

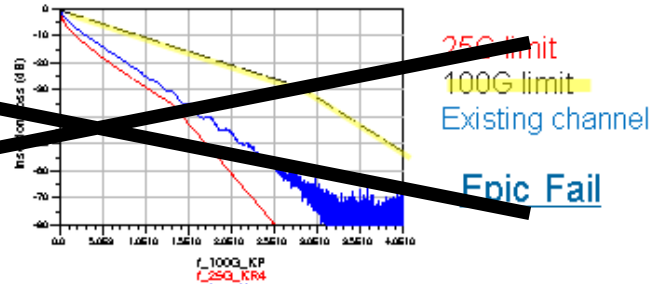PCB industry has made further progress in the last 12 months.

Slide/data Used with permission from Nathan Tracy, TE Connectivity. Data presented at DesignCon2017.

Backplane Considerations
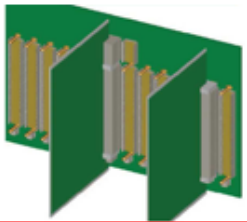
Traditional backplane:
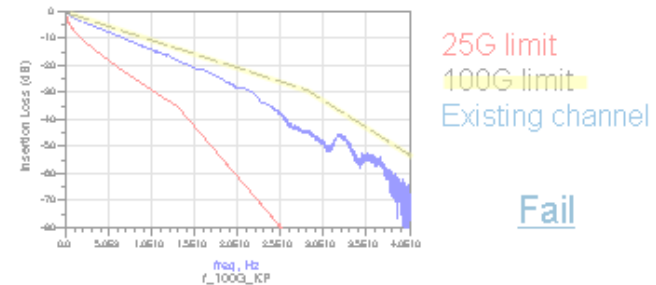- Need to shorten to ~14-17" of Meg6 to meet 30dB

Given routing required for on and off LCs, Routable distance on backplane itself is too limited.

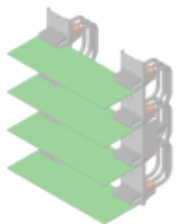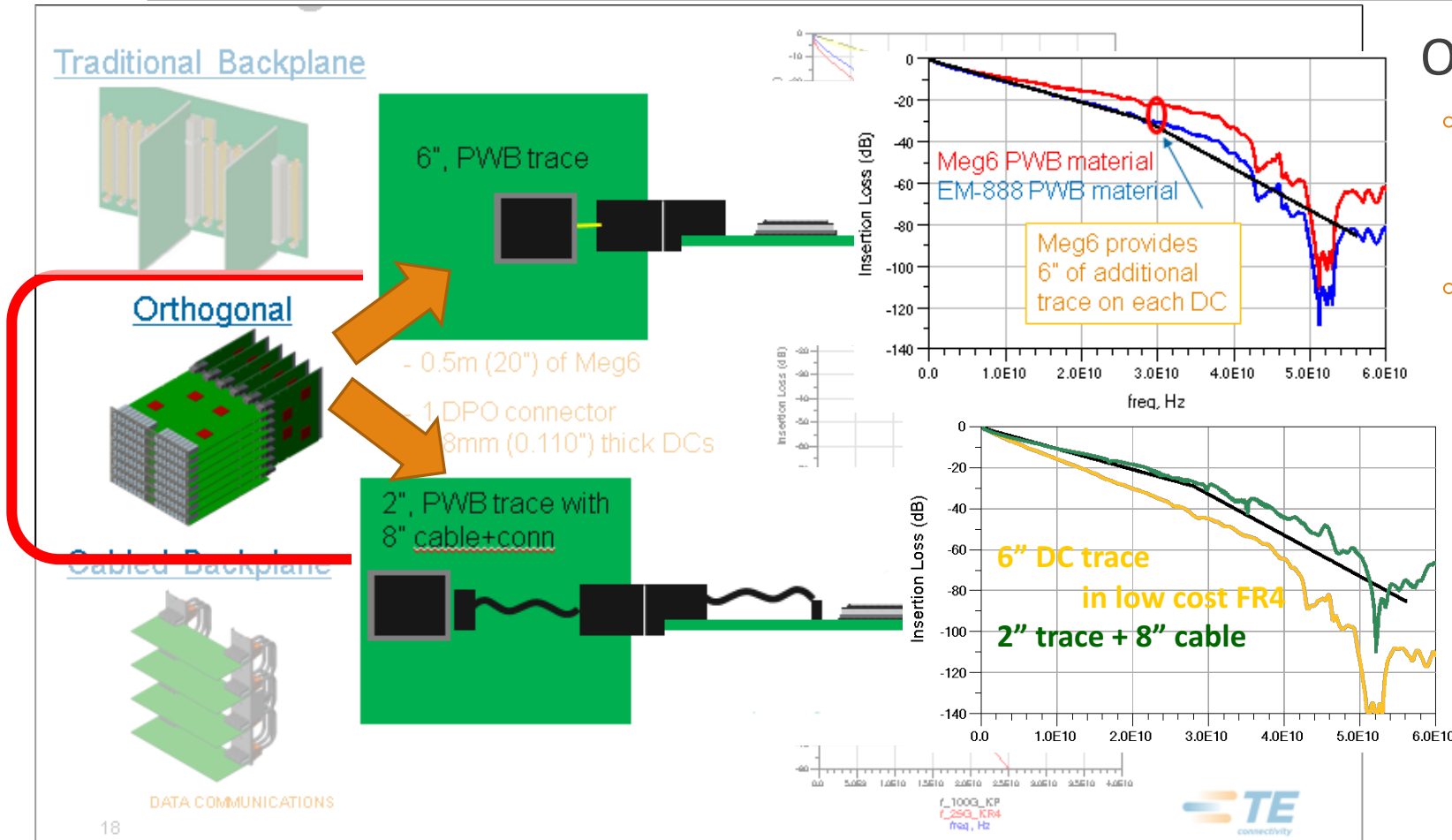Traditional backplane architecture isn't very realistic at this speed.
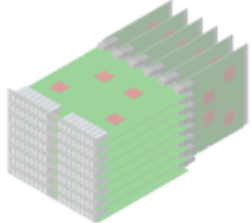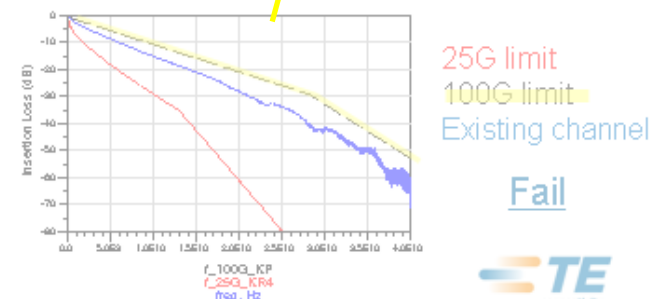
Approximate calculations:
- Holding 4dB for conn + vias
- 1.89dB/in → 13.75"
- 1.485dB/in → 17.5"
- Loss #s from Goergen_nea_01_0517
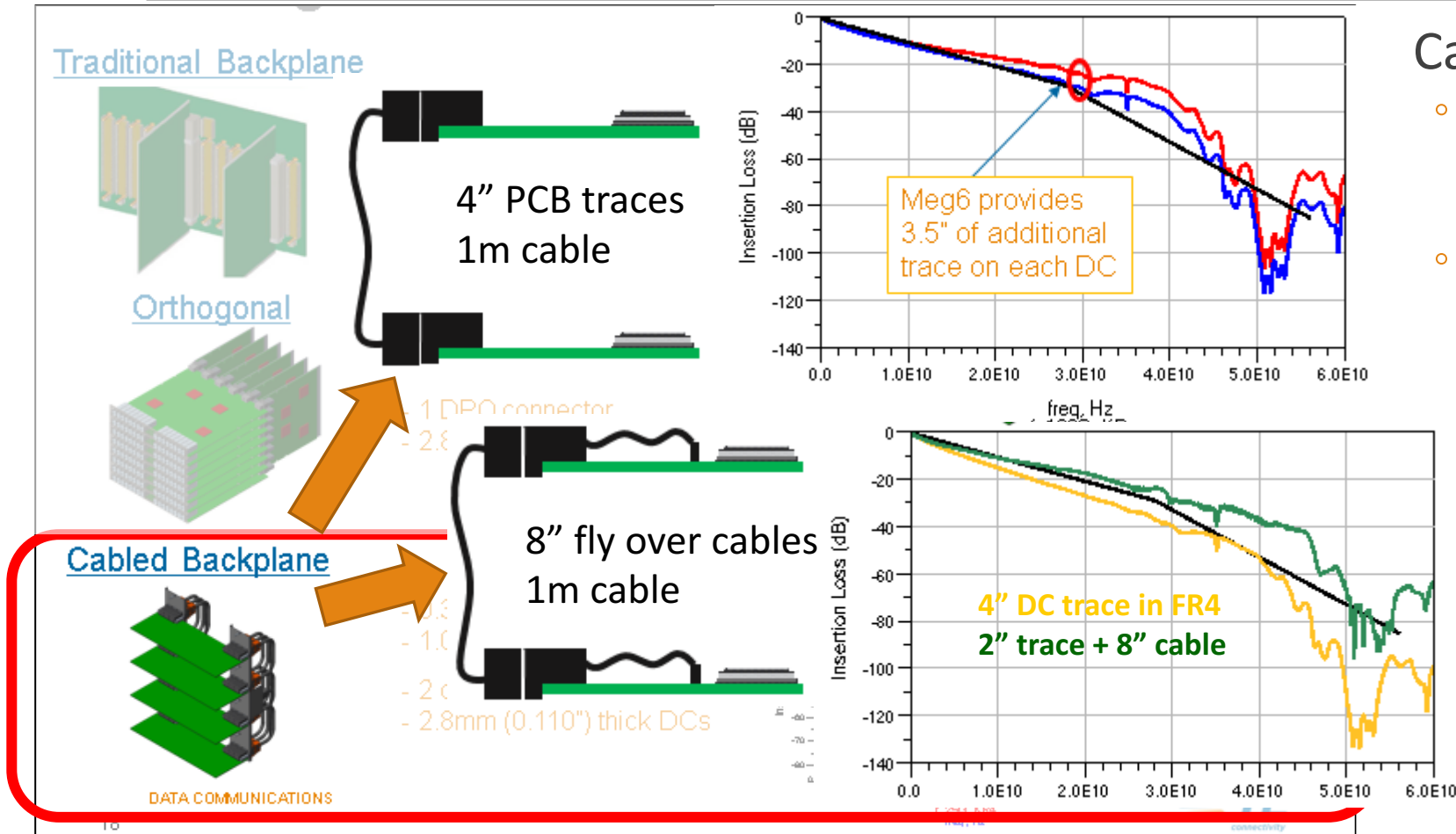
Backplane Considerations

Orthogonal Connector
- Shortening DC traces (on moderate materials) to closer to 6-8" fits in a 30dB budget
- Using 8" fly over cable fits in a 30dB budget

There are paths forward here.

Simulation results used with permission by Nathan Tracy, TE Connectivity.

Backplane Considerations

Traditional Backplane

4" PCB traces
1m cable

Orthogonal

Cabled Backplane

8" fly over cables
1m cable

DATA COMMUNICATIONS

Meg6 provides 3.5" of additional trace on each DC

4" DC trace in FR4
2" trace + 8" cable

Cabled Backplane
◦ Shortening DC traces (on moderate materials) to closer to 4-6" fits in a 30dB budget
◦ Using 8" fly over cable fits in a 30dB budget

There are paths forward here.

Simulation results used with permission by Nathan Tracy, TE Connectivity.

Backplane Summary

Design choices will factor in more than ever before:
- Medium used (PCB vs. cable)
- Surface roughness
- Vias & stubs
- Ground layer thickness
- Radiated emissions shielding

Pure-backplane links do not have the burden of needing to support optics and passive Cu cable
◦ Universal SERDES on ASICs are desirable (see nicholl_3bs_01c_0115*), but may make things too difficult
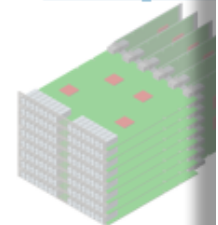
Backplane options that remain at 30dB are still useful for systems!

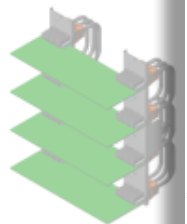* http://www.ieee802.org/3/bs/public/15_01/nicholl_3bs_01c_0115.pdf

Traditional Back

Orthogona

Cabled Backpl

DATA COMMUNIC

18
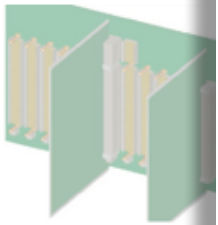
Retimers

Retimers

Are a useful tool to make connections function
- Can increase TF for trickier links

Increase cost per bit, power, and take up board surface area
- These have BMP and EF implications

Need everywhere or just sparingly where needed?
- If we use everywhere, then ASIC SERDES can be simpler

Discussion

We should keep in mind how long each objective is expected to take to standardize
- A subset of the objectives (e.g. C2M) may be economically worth splitting off to finish faster

Need to get consensus on what use cases should be supported

Thanks!