Improving the CR loss budget

Piers Dawe, Nvidia

Supporters

- Rick Rabinovich   Keysight
- Nathan Tracy   TE Connectivity
- Karl Bois   TE Connectivity
- Bruce Champion   TE Connectivity

Problem statement

- The end-to-end loss budgets for CR and C2M are stable now

- After uncertainty, it now looks like CR would work, but:

- The allocation of losses for CR is a poor fit to the primary application, server-switch links
  - However, some designs do use it
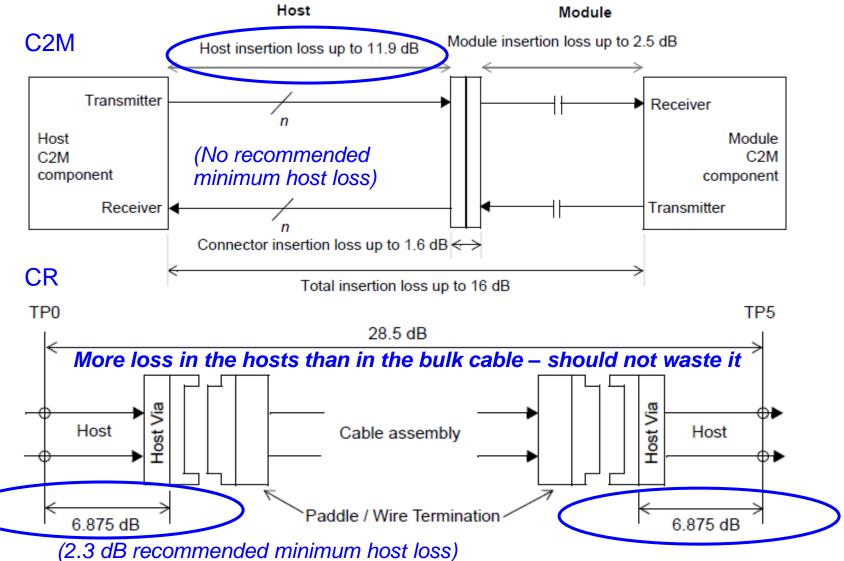
Last time (minutes for 19 May 2021)

- **Presentation #9:**
- "Improving the CR loss budget", Piers Dawe
- See: https://www.ieee802.org/3/ck/public/adhoc/apr28_21/dawe_3ck_adhoc_01_042821.pdf
- **Straw Poll #6**
- I would support a new pair of CR port types with reduced host insertion loss limits on one end (e.g., NIC) and increased host loss limit on the other end (e.g., switch) similar to slide 7 of dawe_3ck_adhoc_01_042821.pdf. (chicago rules)
- A: Yes
- B: No
- C: Need more information
- D: Abstain
- Results: A: 27, B: 13, C: 29, D: 7   (see comment #166)
- **Straw poll #7**
- I would support a new pair of CR port types with reduced host insertion loss limits on one end (e.g., NIC) and increased host loss limit on the other end (e.g., switch) similar to slide 7 of dawe_3ck_adhoc_01_042821.pdf. (Choose one)
- A: Yes
- B: No 19
- C: Need more information
- D: Abstain
- Results: A: 22, B: 11, C: 11, D: 6   (see comment #166)

5 dB less host trace loss in CR than C2M

C2M

CR

Host

Module

Host insertion loss up to 11.9 dB

Module insertion loss up to 2.5 dB

Transmitter

Receiver

Host
C2M
component

(No recommended minimum host loss)

Module
C2M
component

Receiver

Transmitter

Connector insertion loss up to 1.6 dB

Total insertion loss up to 16 dB

TP0

TP5

28.5 dB

More loss in the hosts than in the bulk cable – should not waste it

Host Via

Host

Cable assembly

Host Via

Host

Paddle / Wire Termination

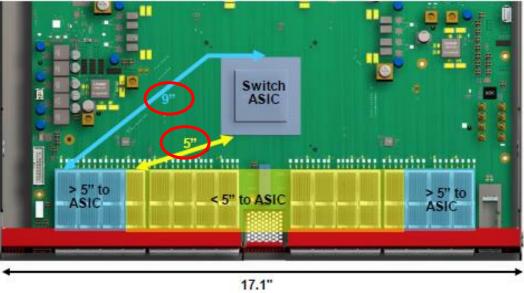6.875 dB

6.875 dB

(2.3 dB recommended minimum host loss)

Part 1, allocate the host loss wisely

•

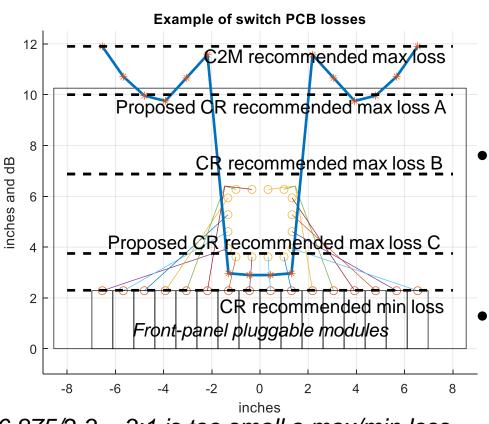# Architectural changes to ToRs due to reduced physical VSR reach

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

• Slide 4 from [1]
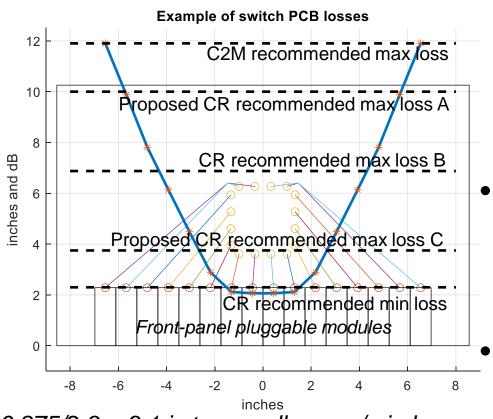
Example of switch PCB losses

Example of switch PCB losses

- C2M recommended max loss
- Proposed CR recommended max loss A
- CR recommended max loss B
- Proposed CR recommended max loss C
- CR recommended min loss
- *Front-panel pluggable modules*

inches and dB (y-axis) — inches (x-axis)

6.875/2.3 = 3:1 is too small a max/min loss ratio anyway

At 25G/lane and 50 G/lane we had 5.8:1

Want at least about 4:1. This example 4.1:1

- In this example, all paths are below the C2M recommended max loss (good)

- But only 8 out of 32 are within the CR recommended max loss (bad)

- Other distributions are possible (see next slide) but the issue remains

- There are fixes (see e.g. [2] but with costs
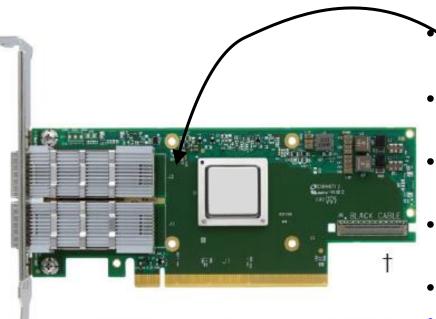
Another example of switch PCB losses

Example of switch PCB losses

12 — C2M recommended max loss

10 — Proposed CR recommended max loss A

8 — CR recommended max loss B

— Proposed CR recommended max loss C

— CR recommended min loss

Front-panel pluggable modules

inches and dB / inches

6.875/2.3 = 3:1 is too small a max/min loss ratio anyway

At 25G/lane and 50 G/lane we had 5.8:1

Want at least about 4:1. This example 5.8:1

- In this example, all paths are below the C2M recommended max loss but it needs lower loss/in than previous example

- But only 10 out of 32 are within the CR recommended max loss (bad)

- 4 paths are below the C2M recommended minimum

- There are fixes (see e.g. [2] but with costs
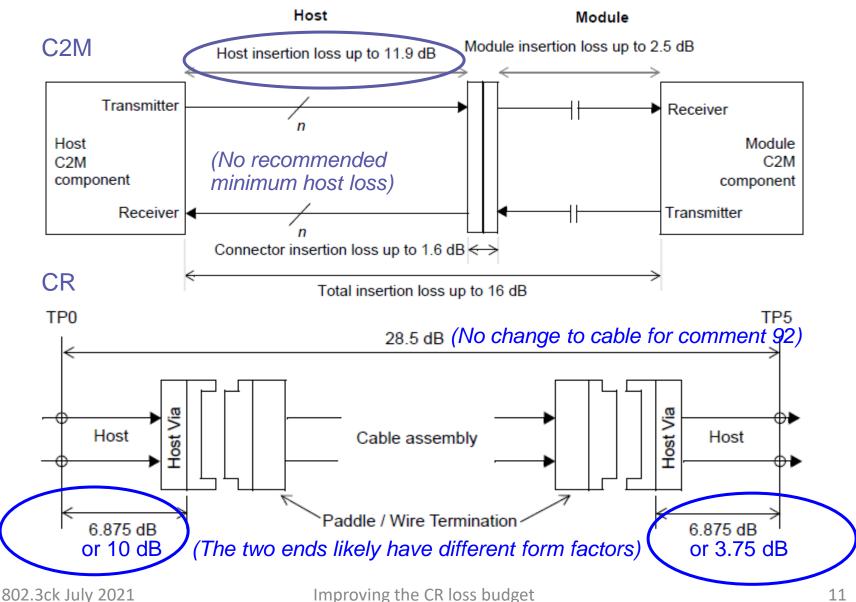
NIC losses

See comments 92 and 93 (reproduced later in this slide pack)

In the future, some LAN-on-motherboard (LOM) servers (different to NICs) could find the 6.875 dB allocation convenient

- IC to module length is typically short
  - More in reference [3]
- PCIe card is much smaller than switch card, but there are many more of them
- For both reasons, trace loss in dB/in is higher for NIC than switch
- But length wins: less loss is needed than in switch
- 3.75 dB is enough for the NIC
- >3 dB spare to give to the switch
  - 3.125 dB re-allocated in this proposal, a little more might be possible
- There are very few ports in a NIC and the host trace losses can be similar to each other, unlike in a switch
- Losses < CR recommended minimum would be natural

Re-allocating 3.125 dB from one host to another in CR

Host

Module

C2M

Host insertion loss up to 11.9 dB

Module insertion loss up to 2.5 dB

Transmitter → Receiver

Host C2M component

(No recommended minimum host loss)

Module C2M component

Receiver ← Transmitter

Connector insertion loss up to 1.6 dB

Total insertion loss up to 16 dB

CR

TP0

TP5

28.5 dB *(No change to cable for comment 92)*

Host | Host Via | Cable assembly | Host Via | Host

Paddle / Wire Termination

6.875 dB or 10 dB   *(The two ends likely have different form factors)*   6.875 dB or 3.75 dB

Result

- Twice as many switch ports (16 out of 32) are CR-capable now in this example
  - The last few ports might be uplinks, using C2M
  - C2M already has "short" and "long" hosts, so grades of host port isn't an alien concept
- Optimised for NIC server-switch links
  - All the NIC-based servers are "short" hosts
  - Some switch ports are "long" or "medium" hosts, all are as capable as "long" hosts or better
  - No significant extra complexity, very attractive for cost, power
- Also can be used to make a cluster switch from multiple pizza boxes
  - A mix of "short", "medium" and "long" ports. A cluster like this would have pre-planned connections, so cost and power savings outweigh the complexity

How is this managed?

- Host advertises its ability to other host via Clause 73 Auto-Negotiation
  - Which it is using anyway
  - So each host knows its and the other host's ability
- 3 ability classes of CR ports, with nominal max host loss allocations of:
- A: 10 dB, B: 6.875 dB as in D2.1, C: 3.75 dB
  - Uses 2 bits in the Link codeword Base Page
- A is typically near the core of the network (like USB type A)
  - Use A B C names rather than long medium short, to avoid confusion with C2M long and short, and with cable lengths
- Clause 73 Priority Resolution function lets A connect to C, B to B or C, and C to A, B or C
- Do we want MDIO registers to report local and remote host ability class?

Changes to Link codeword Base Page

73.6 Link codeword encoding

Replace Figure 73–6 with the following figure to make D43 indicate F4 rather than A22:
and D42 and D43 to indicate host loss ability

D 0	D 1	D 2	D 3	D 4	D 5	D 6	D 7	D 8	D 9	D 10	D 11	D 12	D 13	D 14	D 15
S 0	S 1	S 2	S 3	S 4	E 0	E 1	E 2	E 3	E 4	C 0	C 1	C 2	RF	Ack	NP

D 16	D 17	D 18	D 19	D 20	D 21	D 22	D 23	D 24	D41 D42	D 43	D 44	D 45	D 46	D 47
T 0	T 1	T 2	T 3	T 4	A 0	A 1	A 2	A 3	L1 L2	F 4	F 2	F 3	F 0	F 1

Figure 73–6—Link codeword Base Page

L1	L2	Host loss ability
0	0	B
1	0	A
0	1	C
0	0	Reserved

Change the first sentence of 73.6.4 as follows:

Technology Ability Field (A[~~22~~19:0]) is a ~~23~~20-bit wide field

Other changes to complete

- In Table 162-10, add limits A and C for linear fit pulse peak ratio (min), as in next slide
  - Change text in 162.9.3.1.2 to refer to the table
- In Table 162-15, add columns for Test 2 (high loss), A and C, with test channel insertion loss:
  - A: 6.875-3.75 = 3.125 dB lower (20.5 dB to 21.5 dB), and
  - C: 10-6.875 = 3.125 dB higher (26.75 dB to 27.75 dB), as in slide after next
  - No change needed for Test 1 (low loss)
- In 162.11.7.1.2, instead of using $S^{(HOSPR)}$ per 162.11.7.1.1 (z_p = 110.3 mm for 4.33 dB), use z_p = 30.7 mm for 1.205 dB *Added 14 Jul*
- In 162A.4, add equations for IL_PCBmax and ILHostMax A and C, show them in Fig 162A-1 and 2
- In 162A.5, add Value columns A, C in Table 162A-1 (ILChmin and ILMaxHost differ)
- Adjust figures 162A-3 and 4.

Table 162–10—Summary of transmitter specifications at TP2

Parameter	Subclause reference	Value	Units
Linear fit pulse peak ratio (min)	162.9.3.1.2	~~0.397~~	—

Type A (smaller number)
Type B 0.397
Type C (larger number)

-

Table 162–15—Interference tolerance test parameters

Parameter	Test 1 (low loss)		Test 2 (high loss)				Units
	Min	Max	Min			Max	
			A	B	C	A	B	C	
Test pattern	Scrambled idle encoded by FEC							
FEC symbol error ratio required[a]	$< 10^{-3}$							
Test channel insertion loss at 26.56 GHz[b]	10.5	11.5	20.5	23.625	26.75	21.5	24.625	27.75	dB
Cable assembly insertion loss at 26.56 GHz	10.5	11.5	17.75			19.75			dB
COM[c]	3		3						dB

[a]See 162.9.4.3.5 for definition of FEC symbol error ratio.
[b]Insertion loss between the two test references (see Figure 110–3b).
[c]The COM value is the target value for the SNR_{TX} calibration defined in 162.9.4.3.3 item f. The SNR_{TX} value measured at the Tx test reference should be as close as practical to the value needed to produce the target COM. If lower SNR_{TX} values are used, this would demonstrate margin to the specification but this is not required for compliance.

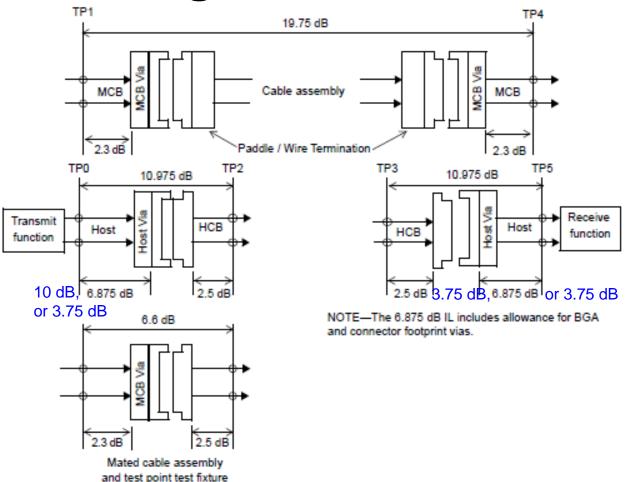- No change needed for Test 1 (low loss)

Equations 162A-1 and 162A-3

-

A $IL_{PCB}(f) \leq IL_{PCBmax}(f) = 1.4268(0.417\sqrt{f} + 0.1194f + 0.002f^2)$ (162A–1)

B $IL_{PCB}(f) \leq IL_{PCBmax}(f) = 0.9809(0.417\sqrt{f} + 0.1194f + 0.002f^2)$ (162A–1)

C $IL_{PCB}(f) \leq IL_{PCBmax}(f) = 0.535(0.417\sqrt{f} + 0.1194f + 0.002f^2)$ (162A–1)

A $IL_{Host}(f) \leq IL_{HostMax}(f) = 2.2775(0.417\sqrt{f} + 0.1194f + 0.002f^2)$ (162A–3)

B $IL_{Host}(f) \leq IL_{HostMax}(f) = 1.5658(0.417\sqrt{f} + 0.1194f + 0.002f^2)$ (162A–3)

C $IL_{Host}(f) \leq IL_{HostMax}(f) = 0.8541(0.417\sqrt{f} + 0.1194f + 0.002f^2)$ (162A–3)

Table 162A-1

- $IL_{Chmin}(f)$ is the channel insertion loss in dB between TP0 and TP5 representative of a minimum insertion loss cable assembly and a maximum loss host channel

Table 162A–1—Insertion loss budget values at 26.56 GHz

Parameter	A	B	C	Units
IL_{Chmax}		28.5		dB
IL_{Camax}		19.75		dB
IL_{Chmin}	19.75	19.75	13.5	dB
IL_{Camin}		11.0		dB
$IL_{MaxHost}$	14.1	10.975	7.85	dB
$IL_{MatedTF}$		6.6		dB

Figure 162A-3

Figure 162A–3—Cable assembly, host, and test fixture insertion loss at 26.56 GHz

Figure 162A-4

●

Figure 162A–4—Channel insertion loss at 26.56 GHz

Part 2, enable longer cables

- As shown, trace lengths and losses in a switch layout cover a range
  - Some low loss, some high loss, in every switch
- Some servers are further from the switch than others
- The max cable loss doesn't deliver passive cables as long as we would like
- People will put two and two together, and connect nearby servers to low-loss switch ports using increased-loss cables
  - Save 3.125 dB at each end, add 6.25 dB to cable
  - 3 m cable for about 1 in 4 of the switch ports is enough of an improvement to be worthwhile
  - Include this in the standard to head off a free-for-all of multiple out-of-spec specials
- A B C hosts could enable a "super short" category too, 19.75-6.25 = 13.5 dB max for A-A links, and intermediate categories: assume that these are not interesting

Enable longer cables: implementation

- 2 classes of cable, "short" or regular (19.75 dB, as today) and "long", $19.75+2*(6.875-3.75) = 19.75+6.25 = 26$ dB max
  - Achievable cable length 3 m
  - A B C hosts could enable a "super short" category too, $19.75-6.25 = 13.5$ dB max for A-A links, and intermediate categories: assume that these are not interesting

- Long cables connect port types C at both ends, short cables connect a valid (see earlier) combination of A, B, C
  - Each host knows its loss ability, learns remote loss ability from AN, learns cable loss class from its memory map

- In 162.11.2, cable assembly insertion loss, change text to refer to Table 162-17

Corrected 14 Jul

- In 162.11.7.1.1, add zp = 30.7 mm for the "long" cable

- In Table 162A-1, add a column for the A-short-A scenario (ILCamax differs) as on next slide

- Figure 162A-3 stays as slide 19, Figure 162A-4 as on slide 24

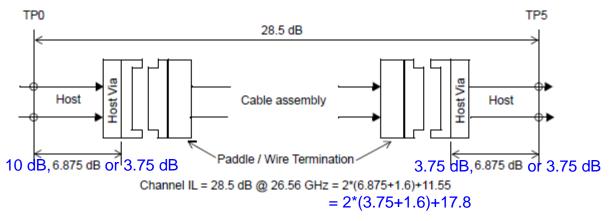# Table 162A-1 enabling long cable

- $IL_{Chmin}(f)$ is the channel insertion loss in dB between TP0 and TP5 representative of a minimum insertion loss cable assembly and a maximum loss host channel

Table 162A–1—Insertion loss budget values at 26.56 GHz

Parameter	Cable	A	B	C	Units
IL_{Chmax}			28.5		dB
IL_{Camax}	Regular		19.75		dB
IL_{Chmin}		19.75	19.75	13.5	dB
IL_{Camin}			11.0		dB
$IL_{MaxHost}$		14.1	10.975	7.85	dB
$IL_{MatedTF}$			6.6		dB
IL_{Camax}	Long		26		dB

Figure 162A-4 enabling long cable

●

TP0 .... 28.5 dB TP5

Host | Host Via | Cable assembly | Host Via | Host

Paddle / Wire Termination

10 dB, 6.875 dB or 3.75 dB   3.75 dB, 6.875 dB or 3.75 dB

Channel IL = 28.5 dB @ 26.56 GHz = 2*(6.875+1.6)+11.55

= 2*(3.75+1.6)+17.8

NOTE—Channel IL derived from cable assembly host, and mated test fixture

Figure 162A–4—Channel insertion loss at 26.56 GHz

Comment 92, improve the CR loss allocations

- *Subclause* **162.9.3** *P* **163** *L* **18** *#* 92   *Type* **TR**

- The draft CR loss budget wastes over 3 dB in nearly every case. The relative range of host losses, 6.875/2.3 = 3:1, is too small for switch layout yet not needed for NICs.

- The recommendation for the host traces plus BGA footprint and host connector footprint, 6.875 dB, compares very poorly with C2M's host insertion loss up to 11.9 dB, making passive copper to this draft expensive and unattractive for a switch, yet a full range of NICs can be made with only 3.75 dB. Server-switch links are asymmetric in form factor (e.g. QSFP-DD to 2 x QSFP) and will get made with an asymmetric loss budget, so it would be better for the standard to regularise what will happen anyway. C2M already has short and long ports.

- This change would also benefit CR switch-switch links because the shortest ports would get credit for their low loss.

- The symmetric budget is used for some designs under way and may be useful in future for LOM, so it is kept here, and the better way added.

Comment 92, *Suggested Remedy*

- 3 classes of CR ports, host loss allocations of A 10, B 6.875, C 3.75 dB. B is as D2.1.

- A connects to C, B to B or C, C to A, B or C.

- Use 2 bits in Clause 73 Auto-Negotiation Link codeword Base Page to advertise A, B or C to the other end. In the Priority Resolution function, an A port ignores a 100G/lane Technology Ability Field bit from an A or B port, a B port ignores a 100G/lane Technology Ability Field bit from an A port.

- In Table 162-10, add limits A and C for linear fit pulse peak ratio (min). Change text in 162.9.3.1.2 to refer to the table.

- In Table 162-14, add columns for Test 2 (high loss), A and C, with test channel insertion loss: A: 6.875-3.75 = 3.125 dB lower (20.5 dB to 21.5 dB), and C: 10-6.875 = 3.125 dB higher (26.75 dB to 27.75 dB). No change needed for Test 1.

- In 162A.4, add equations for IL_PCBmax and ILHostMax A and B and show them in Fig 162A-1 and 2. In 162A.5, add Value columns A, C in Table 162A-1 (ILChmin and ILMaxHost differ). Adjust figures 162A-3 and 4.

Comment 93, enable longer cables

- *Subclause* **162.11** *P* **177** *L* **29** *Type* **T**

- The poor max cable loss makes CR unattractive, while all NICs and some ports on any switch have host loss going to waste. Enabling longer cables on a minority of links is needed.

- In the remedy, each host knows the other host's loss class through AN and the cable's loss class from its I2C compliance code, so the situation is just like any other CR scenario, no extra management features needed in the spec for the long cable class.

Comment 93, *Suggested Remedy*

- 2 classes of cable, which could be called "short" (19.75 dB, as today) and "long", 19.75+2*(6.875-3.75) = 19.75+6.25 = 26 dB max (achievable cable length 3 m). Long cables connect port types C (see another comment) at both ends, short cables connect a valid combination of A, B, C.

- In 162.11.2, cable assembly insertion loss, change text to refer to Table 162-17.

- In 162.11.7.1.1, add zp = 30.7 mm for the "short" cable.

- In Table 162A-1, add a column for the A-short-A scenario (ILCamax differs).

- Illustrate in Figure 162A-4.

Summary

- Cost-effective CR is promising but needs asymmetric loss budget
- Three kinds of CR ports, A B C with max 10 dB, 6.875 dB as D2.1, 3.75 dB host loss. A can connect to C, B to A or B, A to any, with same cable as D2.1
- Add entries in Clause 73 Auto-Negotiation to advertise host loss ability class to the other end
- Define a "long" cable
    - Enables ~ 3 m, for class C ports at each end

References

1. Short Host Channel System Implications, Rob Stone
   https://ieee802.org/3/ck/public/18_05/stone_3ck_01a_0518.pdf

2. Thoughts on CR loss budget
   https://ieee802.org/3/ck/public/adhoc/apr10_19/dawe_3ck_adhoc_01b_041019.pdf

3. Server NIC Trace Lengths
   https://ieee802.org/3/ck/public/18_07/lusted_3ck_01a_0718.pdf

4. C2M AUI and Cu MDI Options
   https://ieee802.org/3/ck/public/18_05/ghiasi_3ck_01a_0518.pdf

5. Improving the CR loss budget
   https://ieee802.org/3/ck/public/adhoc/apr28_21/dawe_3ck_adhoc_01_042821.pdf

Thanks!

Backup follows

Backup

- Similar comments against D2.0

Improving the CR loss budget

D2.0 Comment 166, improve the CR loss allocations

- *Subclause* **162.9.3** *Page* **154** *Line* **21**   *Type* **TR**

- The draft loss budget wastes over 3 dB in nearly every case.

- The recommended maximum insertion loss allocation for the host traces plus BGA footprint and host connector footprint, of 6.875 dB, compares very poorly with C2M's host insertion loss up to 11.9 dB, making passive copper expensive and unattractive for a switch, while a full range of NICs can be made within only 3.75 dB. Server-switch links will get made with an asymmetric loss budget, so it would be better for the standard to regularise what will happen anyway. By the way, many server-switch links will be asymmetric anyway (different form factors at server and switch ends), and that's already allowed in this draft.

- This change would also benefit CR switch-switch links because the shortest ports would get credit for their low loss.

D2.0 Comment 166: *Suggested Remedy*

- As we have done for C2M, create two kinds of CR ports. Host loss allocations of 3.75 dB and 10 dB. Short can connect to short or long with same cable as today; long to long is not supported. Add entries in Clause 73 Auto-Negotiation to advertise short and long to the other end.

- In Table 162-10, provide separate limits for Linear fit pulse peak (min).

- In Table 162-14, provide separate rows for Test channel insertion loss: for testing the short host input the values for Test 2 are 10-6.875 = 3.125 dB higher (26.75 dB and 27.75 dB), while for the long host input the values for Test 2 are 6.875-3.75 = 3.125 dB lower (20.5 dB and 21.5 dB). No change needed for Test 1.

- In 162A.4, provide two equations for each of IL_PCBmax and for ILHostMax and show them in Fig 162A-1 and 2. In 162A.5, provide two Value columns in Table 162A-1. Adjust figures 162A-3 and 4.

- For discussion: should a "long" cable, 19.75+2*(6.875-3.75) = 19.75+6.25 = 26 dB max (maybe 3 m) be defined? A CR link could have no more than one of the three host, cable, and host being "long".

- We could choose other names than "short" and "long" for the ports, possibly "short" and "medium" (as a C2M host can be "longer"), or A and B, somewhat like USB.

- In 162.11.7.1.1, zp, representing the extra loss a host has above an MCB, could be made asymmetric but I believe that would not bring an improvement in accuracy.

- There could be a third kind of CR port with 6.875 dB but this would not be useful for server-switch links, would be useful for only a subset of switch-switch links, for which passive copper is a subset anyway, so it doesn't seem worthwhile.

D2.0 Comment 182, recommended minimum insertion loss

- *Subclause* **162A.4** *P* **260** *L* **40**   *Type* **T**

- This section, for CR, says "the recommended minimum insertion loss allocation for the transmitter or receiver differential controlled impedance PCBs is 2.3 dB at 26.56 GHz".

- This is the same as the 2.3 dB MCB PCB IL (but why?), and (ignoring connector via loss) 1/3 of the maximum host trace loss (6.875 dB). 92A.4 and 136A.4 use a ratio of 0.086/0.5 or 1/5.8 which allows more flexibility in host layout than 1/3 does. 120G has Host insertion loss up to 11.9 dB, and I didn't find a minimum host loss, although very low loss could be more of a concern in C2M than CR.

- *Suggested Remedy*

- Reduce the recommended minimum insertion loss allocation for the CR transmitter or receiver differential controlled impedance PCBs to whatever is justified. If the reasonable limit is a strong function of host package reflection, state whether the recommendation is for a "nominal worst" package, or what. Add a recommended minimum insertion loss for C2M host traces as appropriate.