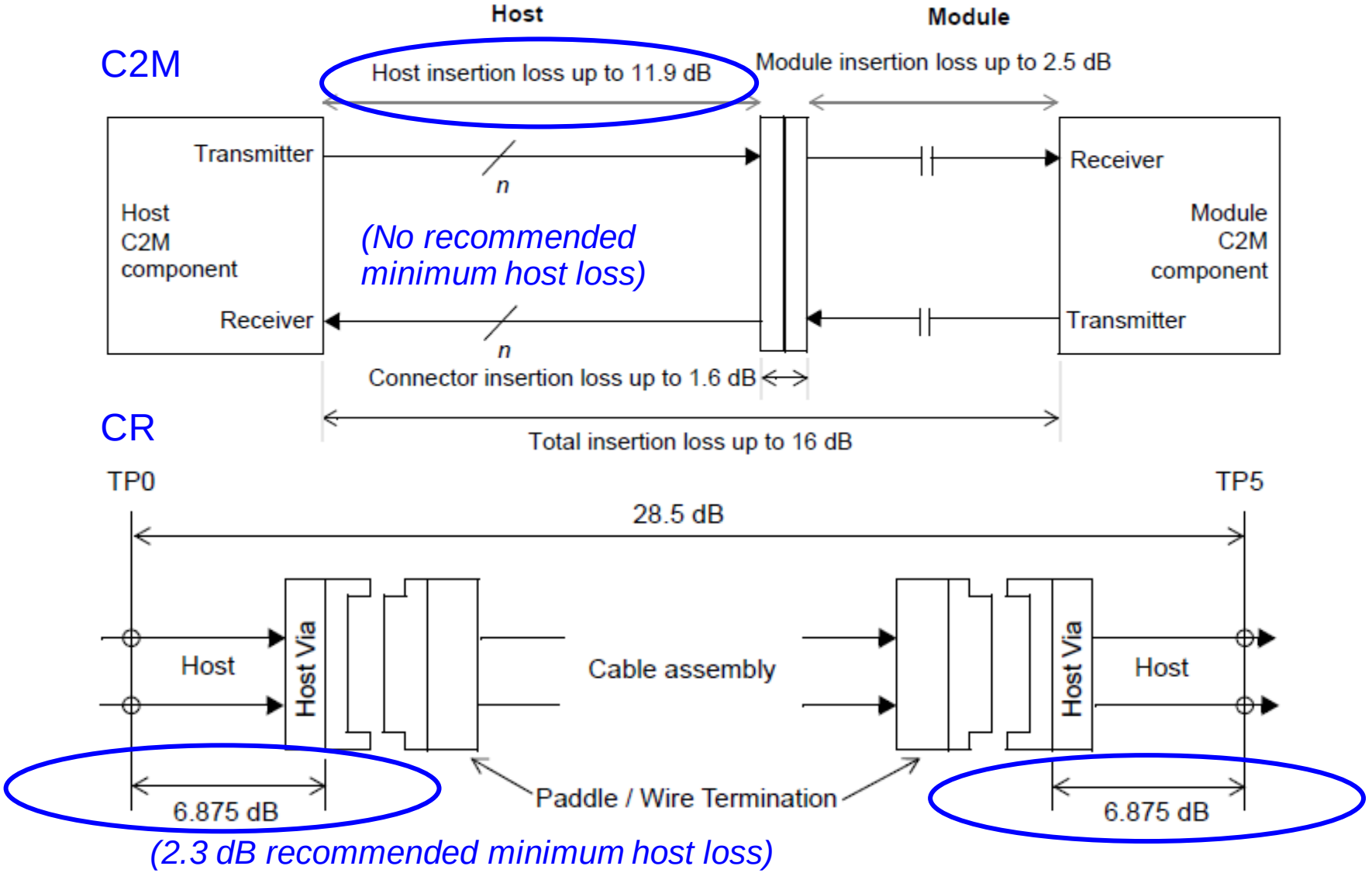
Improving the CR loss budget

Piers Dawe, Nvidia

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
- There are secondary issues with very low host trace losses

5 dB less host trace loss in CR



Architectural changes to ToRs due to reduced physical VSR reach

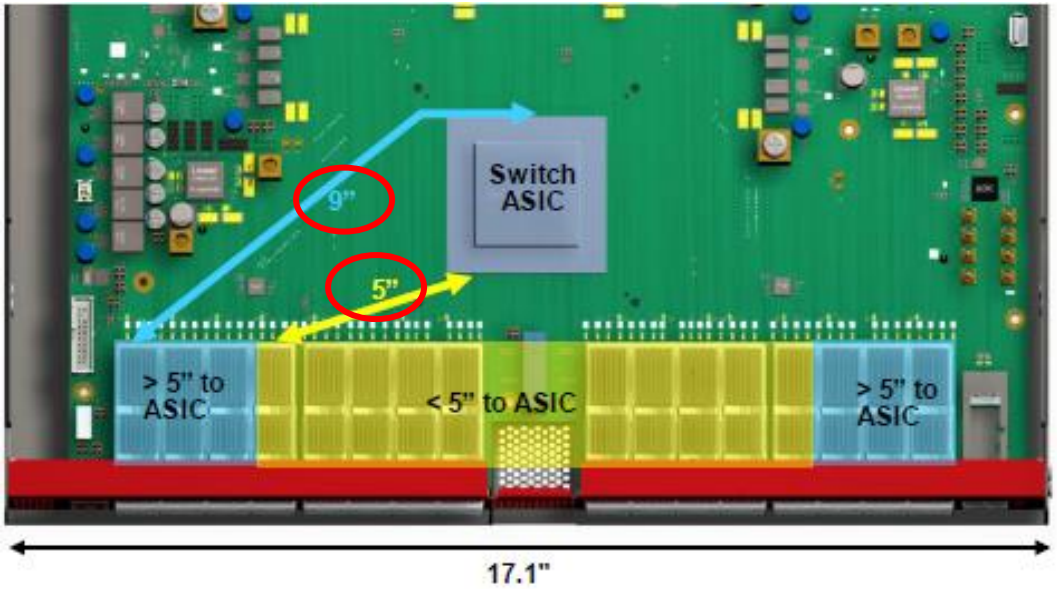
Hypothetical Example:

- 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

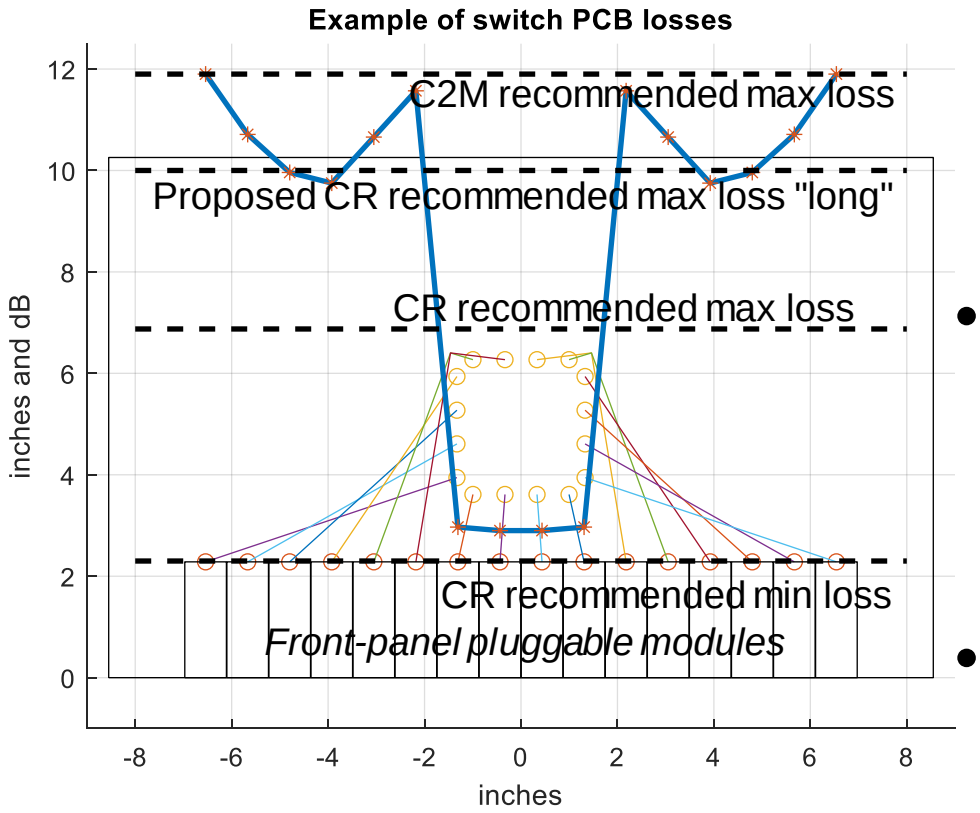
Using Rosemont budget proposal from Jane Lim:

- 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



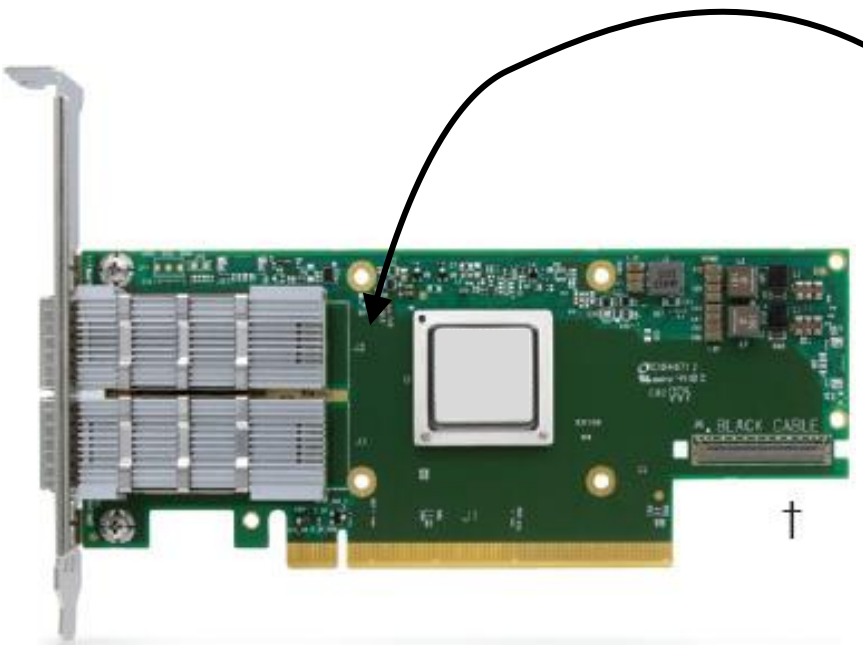
6.875/2.3 = 3:1 is too small a max/min ratio anyway

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

Want at least about 4:1

- 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 but the issue remains
- There are fixes (see e.g. [2] but with costs

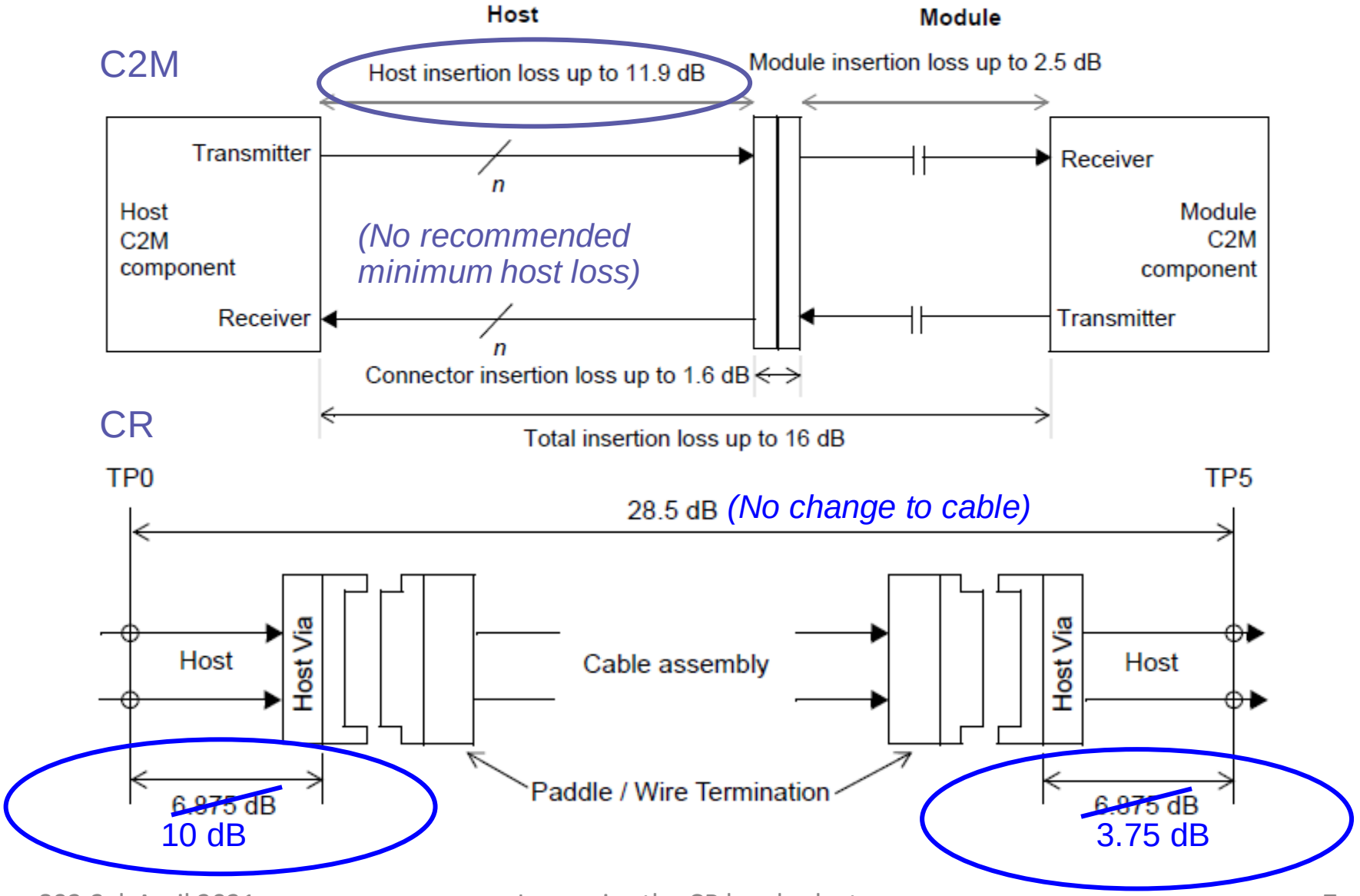
NIC losses



- 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 in this proposal, a little more might be possible – for discussion
- Would like to use losses < CR recommended minimum

See comments 166, 182 (reproduced later in this slide pack)

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



Result

- Twice as many switch ports (16 out of 32) are CR-capable now in this example
- Optimised for server-switch links
  - All servers are "short" hosts, switch ports are "long"
  - No significant extra complexity, very attractive
- Also can be used to make a cluster switch from multiple pizza boxes
  - A mix of "short" and long ports. Something like this would be planned, so cost and power savings outweigh the complexity

Opportunity

- Could connect two "short" ports with a higher-loss (longer) cable
  - Maybe 3 m? See comment 166 (slide 12)
  - Connecting two servers (NICs)
    - Not interesting?
  - Connecting a proportion of servers, further away than others, to switch
    - Interesting!
  - A minority of switch ports to minorities in other switches
    - ?

Related issues

- C2M as no recommended minimum host loss
  - Even though [ref] shows that very low loss host with worst-case package can be troublesome
- CR has a recommended minimum host loss
  - Although CR reference receiver would cope better than C2M reference receiver
  - Same value as MCB trace loss
- Seems to be wrong way round
  - Reduce (or remove?) the recommended minimum host loss

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.

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.

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.

Summary

- Cost-effective CR is promising but needs asymmetric loss budget
- Two kinds of CR ports. Short host loss 3.75 dB, long host loss 10 dB. Short can connect to short or long with same cable
- Add entries in Clause 73 Auto-Negotiation to advertise short and long to the other end.
- Supporting changes:
  - In Table 162-10 transmitter specifications at TP2, provide separate limits for Linear fit pulse peak (min).
  - Modify Test 2 of Table 162-14, Interference tolerance test parameters for short and long hosts. No change needed for Test 1.
  - Revise 162A to implement the intent
- Consider defining a "long" cable
  - A CR link could have no more than one of the three host, cable, and host being "long".

Thanks!

References

1. Short Host Channel System Implications, Rob Stone
   http://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
   http://ieee802.org/3/ck/public/18_05/ghiasi_3ck_01a_0518.pdf