Thoughts on CR loss budget

Piers Dawe
Mellanox

Introduction

- We would like to create a standard for 2 m passive copper links with no more than 28 dB loss ball-to-ball
- Proposed CR baseline [1] allocates 2 × 7 dB for hosts
- Presentations by Tracy [2] and Palkert [3] say that these things are not compatible
  - Shortfall of about 2 dB or 0.4 m, with today's connector and package performance assumptions
  - Depends on connector type
- Assuming RS(544,514) ("KR4") FEC

What could change?

1. Reduced host loss?
   - Both ends or one end?
2. Reduced cable length?
3. Thicker cable?
4. Stronger FEC?
5. Higher loss budget?
6. Improve the cable?
7. Lower loss connectors?
8. Anything else?

Reduced host loss?

- Proposed headline host loss for CR is 7 dB (each host)
- Proposed equivalent for C2M [4] is (16-2.5-2) = 11.5 dB TBC
- ~1.3 dB of each goes on vias and ASIC escape
- 5.7 vs 10.2 dB for trace loss – barely better than half the loss or distance
  - 7 dB is not enough for the usual "pizza box" TOR switch
  - Would need in-the-box cables, retimers on PCB, or don't support passive copper on a large proportion of ports in the TOR switch. See [5]
  - Burdens all ports, even those with active links connected, with additional cost
- 7 dB for switches should be increased not decreased
- Conclusion: No

Reduced host loss, both ends or one end?

- The large majority of few-metre links will be server-switch
- NICs in servers are to PCIe add-in card size
- Traces in NICs are significantly shorter than longest trace in switches, but there are many more NICs than switches so PCB material must be cheaper
- Net: maybe 1 dB can be taken from the NIC loss, but it should be given to the switch loss
- An asymmetric budget like this can be written (compare C2M which is asymmetric), but this is not enough to fix the problem by itself

Asymmetric host loss, switch-switch?

- If there were an asymmetric budget as on previous slide, a switch could have two kinds of copper-supporting ports
    1. Capable of connecting to a NIC with a max-loss cable (or a module or active cable)
    2. Connects to type 1 above (or a module or active cable)
    - Similar to the long ports/ short ports split (C2M / C2M and CR) which is already being proposed
- What is needed to interconnect a rack of pizza-box switches?

Reduced cable length?

- At 2 m, links are within one rack
  - Not connecting 3 racks to 1 TOR with ~2 m 100G/lane passive copper anyway
- If TOR is placed half way up the rack, 2 m links can reach any part of the rack
- So can e.g. 1.75 m
  - May imply constraints on layout of the rack cabling
- See [6] for examples of cable deployments – cases 2 and 4 use >~1.75 m, cases 1,3,5 would need >2.4 m so we have given up on them already
  - See detail in [6]. Can we improve on this?
- Unlike some of the other options, there is a gradual trade-off here:
  - Shorter reach loses a small proportion of possible links (pushing them to active cables), but doesn't break the paradigm or lose the large primary market for passive copper
- Worth further investigation

Thicker cable?

- Assumption is 26 AWG
- 24 AWG would be too heavy, too stiff, would not fit in QSFP-DD
- Conclusion: no

Stronger FEC?

- Would make 100GEL CR different to all other 50G/lane or 100G/lane Ethernet
  - Except coherent optics where the different FEC is in the modules not the host
  - Would increase the FEC overhead and therefore the signalling rate, reducing the net benefit of a stronger FEC
- Conclusion: this would probably work, but too costly and disruptive for 2 dB or 0.4 m.
- Not worth doing

Higher loss budget?

- Not all impairments such as host vias have been factored into signal quality yet
- Have we allowed what we need for real-world host connectors (e.g. worse reflections than MCB connectors)?
- COM doesn't understand quantisation noise, and thermal noise limit is coming into view at 100G/lane
- IC experts I spoke to say: don't do this
- Conclusion: can't agree to do this

Improve the cable?

- For octal-octal cables, don't expect much improvement in cable loss
- Server-switch links are likely to be SFP-SFP, or octal-SFP breakouts
  - Maybe several tenths of a dB lower loss for the same length than octal-octal
  - For which cable widths is what length important?
- Worth investigating, but may not be enough without other changes

Lower loss connectors?

- Lower loss connectors would be part of the host not the cable
  - Any loss reduction identified could be given to host or to cable
- At most a few tenths of a dB might be found for QSFP-DD or OSFP
- Other connector types with fewer lanes may have lower loss
  - Cables with them could be slightly longer for the same cable spec loss, or could allow longer host traces for the same end-to-end loss
  - But crosstalk may be worse
- Worth investigating, but may not be enough without other changes

What could change? revisited

1. ~~Reduced host loss?~~
   - Move loss from one end to the other (asymmetric loss)?
2. Reduced cable length?
3. ~~Thicker cable?~~
4. ~~Stronger FEC?~~
5. ~~Higher loss budget?~~
6. Improve the cable?
   - Be aware of different loss of different connector types
7. Lower loss connectors?
8. Anything else?

Thanks!

References

1. Baseline proposal for copper twinaxial cable specifications, Chris DiMinico
   http://ieee802.org/3/ck/public/19_03/diminico_3ck_01_0319.pdf
2. 100G OSFP Cable Assemblies, Nathan Tracy
   http://ieee802.org/3/ck/public/19_03/tracy_3ck_01a_0319.pdf
3. QSFP-DD 2m Cable Channels, Tom Palkert
   http://ieee802.org/3/ck/public/19_03/palkert_3ck_01_0319.pdf
4. Baseline Proposal for "100 Gb/s, 200 Gb/s, and 400 Gb/s Chip-to-Module Attachment Unit Interface", Mike Peng Li
   http://ieee802.org/3/ck/public/19_03/li_3ck_02b_0319.pdf
5. Short Host Channel System Implications, Rob Stone
   http://ieee802.org/3/ck/public/18_05/stone_3ck_01a_0518.pdf
6. Criteria for 100Gbps Copper Cable Solution, Joel Goergen
   http://ieee802.org/3/100GEL/public/18_03/goergen_100GEL_01_0318.pdf