*

Thoughts on CR loss budget

Piers Dawe, Mellanox

With input and support from

Rob Stone, Broadcom

Mike Dudek, Marvell

Rick Rabinovich, Keysight Technologies

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

*Slides with * are revised or added vs. [4]*

*

What could change?

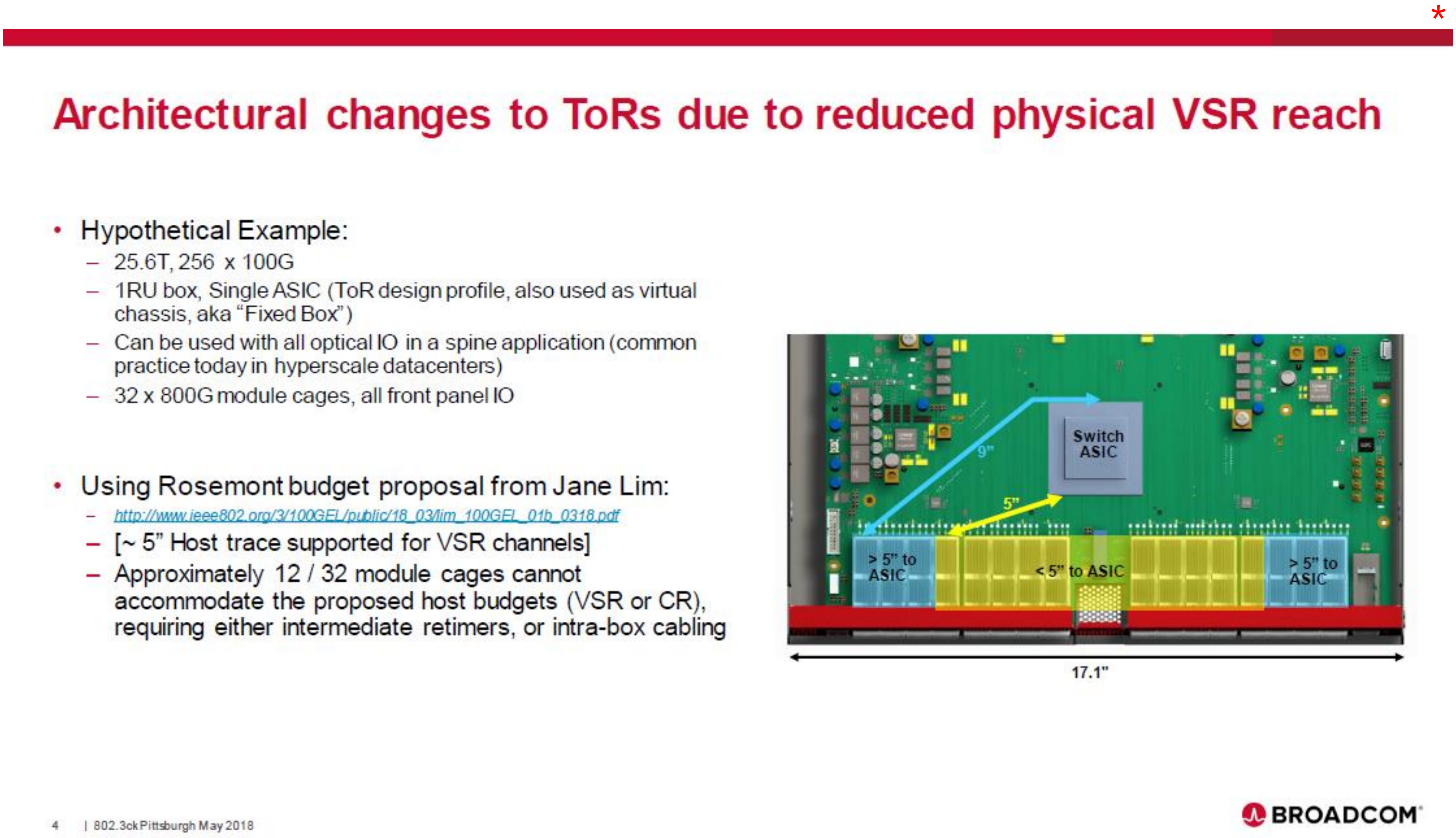
1. Reduced host loss?
   - Both ends or one end?
2. Reduced cable length?
3. Thicker cable?
4. Active cable?
5. Stronger FEC?
6. Higher loss budget?
7. Improve the cable?
8. Lower loss connectors?
9. Anything else?

Reduced host loss?

*

- Proposed headline host loss for CR is 7 dB (each host)
- Proposed equivalent for C2M [5] 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 full length passive copper - on a large proportion of ports in each TOR switch. See [6] slide 5, see [7]
  - Burdens all ports, even those with active links connected, with additional cost. *How much?*
- Possibly 7 dB could be slightly decreased, but with a larger proportion of ports with in-the-box cables or retimers or not for full length passive copper
- 7 dB for switches should be increased not decreased
  - Or, the extra dB in the C2M budget should be revisited
- Conclusion: Looks expensive, too different to C2M

*



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

- Slide 4 from [6]

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?

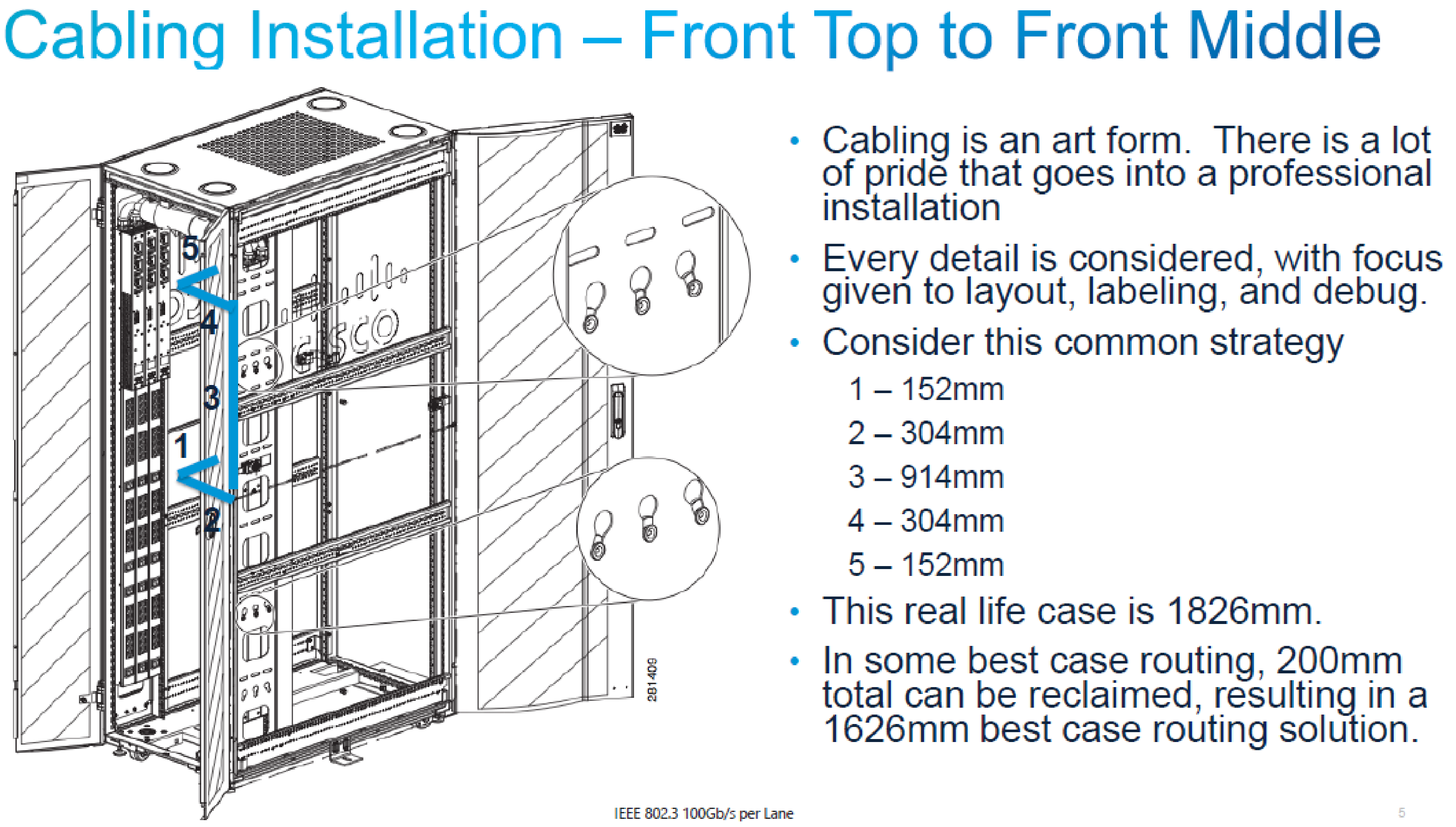
- If there were an asymmetric budget as on previous slide, a switch could have three kinds of ports:
  A. Highest loss, for optics modules and active cables only
  B. Low loss, connects to a NIC or other very low loss port with a max-loss cable, to a similar port with a shorter cable, or to any kind of port with a module or active cable
  C. Very low loss, including NICs, connects to similar or type B (above) with a max-loss cable, or to any kind of port with 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?
- How are switch clusters designed and installed, logistically? Are they pre-planned? Can different port types be managed? Do such clusters need 2 m?
- Worth considering for a planned, data centre or supercomputing environment

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 [8] *(next slides)* for examples of cable deployments – cases 2 and 4 use >~1.8 m, cases 1,3,5 would need >2.4 m so they will need some active cables
  - See detail in [8]. *How much slack is needed? 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

*



5
4
3
1
2
cisco
281409

Cabling Installation – Front Top to Front Middle

- Cabling is an art form. There is a lot of pride that goes into a professional installation
- Every detail is considered, with focus given to layout, labeling, and debug.
- Consider this common strategy

  1 – 152mm

  2 – 304mm

  3 – 914mm

  4 – 304mm

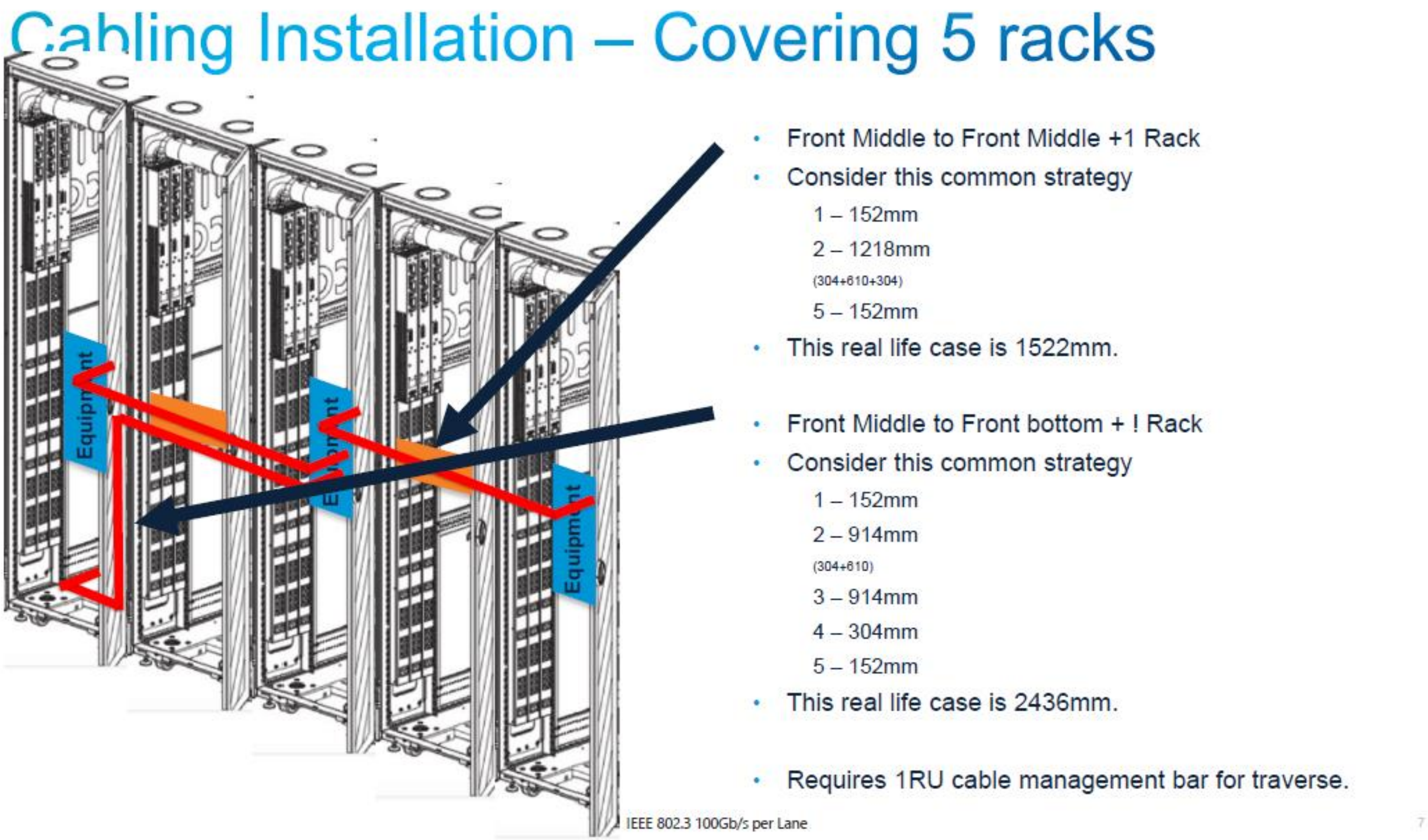
  5 – 152mm
- This real life case is 1826mm.
- In some best case routing, 200mm total can be reclaimed, resulting in a 1626mm best case routing solution.

IEEE 802.3 100Gb/s per Lane

5

- Case 2 from [8]

*



Cabling Installation – Covering 5 racks

- Front Middle to Front Middle +1 Rack
- Consider this common strategy
  - 1 – 152mm
  - 2 – 1218mm
  - (304+610+304)
  - 5 – 152mm
- This real life case is 1522mm.

- Front Middle to Front bottom + ! Rack
- Consider this common strategy
  - 1 – 152mm
  - 2 – 914mm
  - (304+610)
  - 3 – 914mm
  - 4 – 304mm
  - 5 – 152mm
- This real life case is 2436mm.

- Requires 1RU cable management bar for traverse.

IEEE 802.3 100Gb/s per Lane 7

- Case 4 from [8]

*

Thicker cable?

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

Active cable?

- Not explored in this slide set
- Active cables could look to the host like optical modules or like passive cables with different loss, noise and linearity
- Cost? Power?

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)?
- Investigations under way – let's see what they find
- Could we rely on COM yet go beyond 28 dB ball to ball?
  - Some loss limit is needed anyway to bound the range of signals the receiver has to cope with. Expected to be same receiver for CR as for KR. Limit should not be too far-fetched (no use having a high limit if cables would fail another spec anyway)
  - Package reflections are high and still being debated, COM doesn't understand quantisation noise, and receiver noise limit is coming into view at 100G/lane
  - Could the COM threshold be reduced enough to make a difference?
    - Present COM results seem to have a large "random" component related to reflection phase
- IC experts I spoke to say: don't go beyond the agreed 28 dB
- 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 (at the breakout end?) for the same end-to-end loss
  - But crosstalk (NEXT, SFP) may be worse
- Worth investigating, but may not be enough without other changes

What could change? revisited

*

1. ~~Reduced host loss?~~ *controversial*
   - Move loss from one end to the other (asymmetric loss)?
2. Reduced cable length?
3. ~~Thicker cable?~~
4. Active cable?
5. ~~Stronger FEC?~~
6. ~~Higher loss budget?~~
   - More reliance on COM? *We should improve COM's host assumptions anyway*
7. Improve the cable?
   - Be aware of different loss of different connector types
8. Lower loss connectors?
9. 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. Thoughts on CR loss budget
   http://ieee802.org/3/ck/public/19_03/dawe_3ck_01_0319.pdf
5. 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
6. Short Host Channel System Implications, Rob Stone
   http://ieee802.org/3/ck/public/18_05/stone_3ck_01a_0518.pdf
7. C2M AUI and Cu MDI Options
   http://ieee802.org/3/ck/public/18_05/ghiasi_3ck_01a_0518.pdf
8. Criteria for 100Gbps Copper Cable Solution, Joel Goergen
   http://ieee802.org/3/100GEL/public/18_03/goergen_100GEL_01_0318.pdf