# FEC-protected chip-to-module 25G-AUI specification

Piers Dawe

IEEE P802.3by, Mar 2015, Berlin

# Supporters

- Adee Ran — Intel
- Tom Palkert — Luxtera

# Introduction

- Chip-to-module 25G-AUI (Annex 109B) is used to connect a 25G module to a host
- Only one PMD type is defined for this: 25GBASE-SR, which uses FEC
- The 25G-AUI could be implemented as SFP (1 PMD in a module) or QSFP (4 PMDs in a module)
  - Other possible formats include CFP2 and CDFP
  - The same QSFP module could be used for 100GBASE-SR4 (using CAUI-4)
  - The same SFP socket in the host (e.g. a NIC) could be used for 25GBASE-CR (no other Ethernet at this speed)
  - The same QSFP socket in the host could also be used for 25GBASE-CR, CAUI-4 for 100GBASE-SR4 or 100GBASE-CR4
  - Also CAUI-4 for 100GBASE-LR4 (not so much in data centres where 25GBASE-SR 25G-AUI will mostly be used), and unofficial PMDs such as PSM4
- The FEC (if used) is in the host and so the chip-to-module 25G-AUI's errors are corrected by that FEC
  - If considering compatibility with 400G Ethernet, e.g. using CDFP: all 400G PMDs are expected to be FEC-protected
- It is desirable that the specifications for the alternatives above are compatible
  - The same hosts and modules could be used at other speeds – not discussed in detail
- **What exactly is needed for compatibility and what opportunity for cost reduction is possible for 25G-AUI? How can we reduce power and test time?**

# Compatibility

- Compatibility between copper and optical
  - e.g. 25GBASE-CR and CAUI-4
  - Some electrical specifications differ: silicon takes on a "personality" according to what is connected to it
  - Non-programmable characteristics such as host channel loss and return loss should be compatible
    - Items like voltage swing, coding, use of FEC can be different between copper and optical while keeping compatibility
    - Generally the host channel loss is limited by the copper requirements
      - Might not be true for all 25GBASE-CR variants?
      - So not proposing any change to host channel loss or return loss
- Compatibility between different optical AUIs / PMDs
  - Expect to be able to achieve this without changing "personality"
    - Except possibly the exact signalling rate for 400GBASE-SR16: not the subject of this presentation
  - BER requirements differ.  This is our opportunity
- Compatibility between different copper PMDs
  - Not the subject of this presentation
- Summary: want to cost-reduce 25G-AUI, keeping compatibility with CAUI-4 and future CDAUI-16

# Opportunities

- Easier (FEC-protected) BER can enable:
- Reduced test time of module and host input and output
  - Reduced cost. Reduced time of BER (input) test is obvious – for output test see next slide
- Reduced minimum voltage swing from host and module (relaxed output eye height spec)
  - Reduced power, hence cost. Reduced crosstalk, especially to a neighbouring copper port
- Relaxed jitter from host and module (relaxed output eye width spec)
  - Wider tuning range, can be used for reducing the power of adaptive tuning, relaxing any calibration for manual tuning, or providing a more robust interface
- Relaxed stressed eye requirements for host and module (relaxed eye height and width from the input's point of view)
  - Robustness, shorter test programs if there's more margin, might allow implementation-dependent power reductions
- A host that also supports CAUI-4 might choose not to take advantage of these
  - But it could reduce the voltage swing according to the module type
  - For a 25GBASE-SR module in SFP format, CAUI-4 doesn't apply, so there's no compatibility concern or downside
  - For a 25GBASE-SR/100GBASE-SR4 module in QSFP format, CAUI-4 may apply. The host could direct the module to e.g. reduce the voltage swing to save power when in 25G mode
- There will be very many hosts that don't support CAUI-4 (NICs, which don't need 4 lanes)

# Benefits: a lot or a little?

- Reduced test time of module and host output
  - The eye measurement procedure in CAUI-4, 25G-AUI takes a long time!
  - About 15 min per lane per condition, before any test acceleration tricks
  - The issue is the 4 million samples for an effective extrapolation to 1e-15
- Reduced minimum voltage swing from host and module (relaxed output eye height spec)
  - 128 or 144 lanes per switch card. I/O (including the module power) might consume very roughly 2 to 4 W per 100G interface (4 lanes), or very roughly 100 W per card
  - 800 mV pk-pk into 50+50 ohm load // 50+50 ohm matching R, so 16 mA in the output stage, plus another 30%? behind it. Say 1.5 V supply for host, 3.3 V for module. 0.016 * 1.3 * (1.5+3.3) = 100 mW/lane
    Multiply by 144 lanes, giving ~14 W per switch card. A significant fraction of this is wasted.
  - Note that the VCSEL can transmit over 100 m with less current than the 25G-AUI needs to transmit less than a foot!
  - So about 2% of the I/O power could be saved. Worth taking if it really is free, because so much of the other power is fixed
- Relaxed jitter from host and module (relaxed output eye width spec)
  - The power that adaptive tuning takes can be surprisingly high

# What to relax

- Address the very long test time and partly address the power requirement of the eye height spec by changing the spec from EH15 to EH6, and EW15 to EW6
- EH15 and EW15 represent a BER of 1e-15, EH6 and EW6 represent a BER of 1e-6, which is 1/50 of the 5e-5 limit for 25GBASE-SR
- There could be two 25G-AUI links in series with the optical link, so the hypothetical total BER is 5.2e-5. The FEC corrects 5.2e-5 random errors to 1e-12
  - Usually we don't need to add BERs in such a pessimistic way, but one should take care when FEC is involved, and this is affordable
- What about the EH6, EW6 limits?
  - If we make them small, we make the spec easier for the outputs (host and module)
  - If we make them large, we make the spec easier for the inputs (host and module)
  - If we change the limits for EH15 and EW15 to limits for EH6 and EW6 with the same values, we give the benefit to the outputs (mainly the high loss host ports)
    - The output can make a worse eye. The input must tolerate this worse eye but is allowed to make more errors
  - An output with a particular EH15, EW15 might have EH6, EW6 that are 5% to 40% larger? than EH15, EW15
    - The relation between EH15, EW15 and EH6, EW6 for stressed eyes used in testing should be consistent
  - An input that tolerates a particular EH15, EW15 at BER = 1e-15 might tolerate EH6, EW6 that are a little to 40% smaller? at BER = 1e-6

# Comment 145

- *Cl **109B SC 109B.1.1 P 214 L 22 # 145*** *Comment Type **TR***
- This bit error ratio spec goes with non-FEC PMDs that can't be connected to 25G-AUI. It adds a pointless burden of test cost and power - this is most obvious for a 25GBASE-SR module for which the PMD type is known.
- Also, any consideration of error correlation should take the FEC into account.
- The remedy below is intended to put no burden on the host and allow dual-use hosts or modules that are tested to CAUI-4 only.
- *Suggested Remedy* Change The bit error ratio (BER) shall be less than 10^-**15** with any errors sufficiently uncorrelated to ensure an acceptably high mean time to false packet acceptance (MTTFPA) assuming 64B/66B coding. to
- The bit error ratio (BER) shall be less than 10^-**6** with any errors sufficiently uncorrelated to ensure an acceptably high mean time to false packet acceptance (MTTFPA) assuming 64B/66B coding **and the RS-FEC of Clause 108**.
- In 109B.3.1, add exceptions:
- EW15 and EH15 do not apply.
- Limits for EW6 and EH6 A and B are 0.46 UI and 95, 80 mV.
- In 109B.3.2, add exceptions:
- EW15 and EH15 do not apply.
- Limits for EW6 and EH6 are 0.57 UI and 228 mV. VEC6 is defined as 20*log10(AV/EH6). Limit 4.5 dB.
- In 109B.3.3, add exceptions:
- Host implementer may comply to either the host stressed input test of 83E.3.3.2 (BER <= 1e-15) or to a test to BER<=1e-6 with the EW6, EH6 defined for the module output in 109B.3.2 with a VEC6 in the range of 3.5 dB to 4.5 dB with a target value of 4 dB.
- In 109B.3.4, add exceptions:
- Module implementer may comply to either the module stressed input test of 83E.3.4.1 (BER <= 1e-15) or to a test to BER<=1e-6 with the EW6, EH6 defined for the host output in 109B.3.1.
- *Proposed Response*
- PROPOSED ACCEPT IN PRINCIPLE.
- Pending task force review.

# Discussion

- It is likely that better-optimised specs such as this will be developed in the industry, e.g. company by company or in Fibre Channel and InfiniBand and MSAs
  - Fibre Channel has one, but it enables increased host loss, which is thought not compatible with 25GBASE-CR and 100GBASE-CR4 loss budgets
  - A single Ethernet standard spec will avoid multiple proprietary specs, helping the industry
- Do EH15, EW15 specs give interoperability at 1e-6 with another part tested to EH6, EW6?
  - i.e. when an output complies to an EH6, EW6 spec, and the input complies to 1e-15 at EH15, EW15
  - If the Gaussian tails of the worst output are steeper than (or similar to?) those of the worst input, it's OK
  - This is expected to be the case when the output is the module (short low loss traces)
  - In the other direction, the EH6, EW6 limits could be set a little greater than the present EH15, EW15 limits
    - No burden for a high loss host port (more Gaussian tails), might be a burden for a low loss host port, but would still be easier for the host IC than driving the high loss port
      However, a NIC would have no high loss ports
  - We could use say EH8, EW8 instead of EH6, EW6, reducing the difference between legacy 1e-15 method and FEC-protected method, but reducing the benefit of any change
  - We could use say EH8, EW8 in the host output eye spec, and EH6, EW6 in the stressed eye for module input test, but for the same numbers (mV and UI), testing to 1e-6. Provides outputs that must be "100x" better than inputs need without developing new limit values. Narrows any possible gap in the wrong direction between the EH*n* output and the EH15 input. Not as much test time improvement as EH6 output testing
  - This option is developed on the next page

# Comment 145 – revised proposal 1/2

- *Suggested Remedy* Change The bit error ratio (BER) shall be less than 10^-**15** with any errors sufficiently uncorrelated to ensure an acceptably high mean time to false packet acceptance (MTTFPA) assuming 64B/66B coding. to
- The bit error ratio (BER) shall be less than 10^-**6** with any errors sufficiently uncorrelated to ensure an acceptably high mean time to false packet acceptance (MTTFPA) assuming 64B/66B coding **and the RS-FEC of Clause 108**.

- (More on next page)

- In 109B.3.4 (25G-AUI C2M module input characteristics), add exceptions:
- Module implementer may comply to either the module stressed input test of 83E.3.4.1 (BER **<** 1e-15) or to a test to BER **<** 1e-6 with EW6, EH6 set at the limits for EW8, EH8 specified for the host output in 109B.3.1.
- Revise the PICS to follow the changes on this and the next slide

# Comment 145 – revised proposal 2/2

- In 109B.3.1 (25G-AUI C2M host output characteristics), add exceptions:
- A 25G-AUI C2M host output shall meet the specifications in 83E.3.1 with the following differences:
- EW15 and EH15 do not apply.
- The eye width and eye height measurement method is as 83E.4.2 except that the number of samples is equivalent to at least 400,000 bits to allow for construction of a normalized cumulative distribution function (CDF) to a probability of 1e-5 without extrapolation.
  - 83E.4.2 has 4 million, 1e-6
- The best linear fit is found over the range of probabilities of 1e-3 to 1e-5.
- Define EW8 = EW5 – 1.35 × (RJR + RJL), EH8 = EH5 – 1.35 × (RN0 + RN1)
  - 83E.4.2 has EH15 = EW6 – 3.19 × (RJR + RJL), EH8 = EH5 – 1.35 × (RN0 + RN1)
- Limits for EW8 and EH8 A and B are 0.46 UI and 95, 80 mV.
- Alternatively, 25G-AUI C2M host output may meet all specifications in 83E.3.1.
- In 109B.3.2 (25G-AUI module output characteristics), add exceptions, allowing CAUI-4 alternative similar to above:
- EW15 and EH15 do not apply.
- The eye width and eye height measurement method is as 83E.4.2 except that the number of samples is equivalent to at least 400,000 bits to allow for construction of a normalized cumulative distribution function (CDF) to a probability of 1e-5 without extrapolation.
- The best linear fit is found over the range of probabilities of 1e-3 to 1e-5.
- Define EW8 = EW5 – 1.35 × (RJR + RJL), EH8 = EH5 – 1.35 × (RN0 + RN1)
- Limits for EW8 and EH8 are 0.57 UI and 228 mV. VEC8 is defined as 20*log10(AV/EH8). Limit 5 dB.
- In 109B.3.3 (25G-AUI C2M host input characteristics), add exceptions, allowing CAUI-4 alternative similar to above:
- Host implementer may comply to either the host stressed input test of 83E.3.3.2 (BER <= 1e-15) or to a test to BER<=1e-6 with the EW6, EH6 set at the limits for EW8, EH8 specified defined for the module output in 109B.3.2, with a VEC6 in the range of 3.5 dB to 4.5 dB with a target value of 4 dB. VEC6 is defined as 20*log10(AV/EH6)

# Conclusions

- The non-FEC-aware chip-to-module 25G-AUI specification is unnecessarily expensive for use with 25GBASE-SR
  - Test costs, some wasted power
- A lower cost spec is needed
  - Implementers will derate the spec, standard or not
  - Want to allow hosts and modules qualified to (non-FEC-aware) chip-to-module CAUI-4 without retesting
  - Can avoid a fragmented market with unnecessary confusion by standardizing the lower cost spec
- Annex 109B:
  - Outputs specified to CAUI-4 spec or EH8 and EW8
  - Inputs tested to CAUI-4 spec or EH6, EW6 and 1e-6 BER
- Conservative spec, no added burden to CAUI-4 implementations yet allows cost reduction particularly for 25GBASE-SR

# Thank You