400 GbE Discussion on skew variation and tolerance

Paul Brooks
Viavi solutions

This presentation discusses aspects of skew variation (dynamic skew) with regard to the emerging 400G 802.3 bs standard

Contributers

- Gary Nichol - Cisco
- Mark Gustlin - Xilinx
- Andre Szczepanek - Inphi
- Matt Brown - APM
- Chris Cole - Finisar
- Jonathan King - Finisar

Reference material

- IEEE 802.3 table 80-5
- http://www.ieee802.org/3/10GMMFSG/email/ppt00016.ppt
- http://www.ieee802.org/3/ba/public/tools/index.html
- 100G PSM4 Specification version 2.0 – table 10

What is skew variation?

- Two types of skew can exist in a parallel data transmission system.
- Static
  - A fixed UI offset (internal FIFOs, physical delay offsets etc)– typically managed at the PCS layer. Should not vary once link establish. If it changes you get a bit slip and hence an error condition
- Dynamic
  - Physical layer variability (fiber – relative delay with lambda), electrical buffer propagation delay with temperature
  - Can vary over time due to physical conditions (temperature, vibration, fiber plant movement)
  - Should not cause any error if varied within reasonable expected limits ➔ see discussion below

Skew variation tolerance at 100G

Table 80–5—Summary of Skew Variation constraints

Skew points	Maximum Skew Variation (ns)	Maximum Skew Variation for 10.3125 GBd PMD lane (UI)[a]	Maximum Skew Variation for 25.78125 GBd PMD lane (UI)[b]	Notes[c]
SP1	0.2	≈ 2	N/A	See 83.5.3.1
SP2	0.4	≈ 4	≈ 10	See 83.5.3.3 or 84.5 or 85.5 or 86.3.2 or 87.3.2 or 88.3.2.
SP3	0.6	≈ 6	≈ 15	See 84.5 or 85.5 or 86.3.2 or 87.3.2 or 88.3.2
SP4	3.4	≈ 35	≈ 88	See 84.5 or 85.5 or 86.3.2 or 87.3.2 or 88.3.2
SP5	3.6	≈ 37	≈ 93	See 83.5.3.4 or 84.5 or 85.5 or 86.3.2 or 87.3.2 or 88.3.2
SP6	3.8	≈ 39	N/A	See 83.5.3.5
At PCS receive	4	≈ 41	N/A	See 82.2.12

Comments on table 80-5

- Values accepted as generally good, practical and in line with real-life conditions
- Issues around what is the acceptable rate of change (skew variation v. time) that should be tolerated.
  - Typical guidance (from T&M vendors) that these are effects caused by slow changing events like thermal effects or fiber plant mechanical manipulation (vibration, movement)
  - Typical changes in mUI/sec (for thermal), vibration & mechanical stress figures are FFS.

- Suggest we base the skew tolerance numbers for 400G (802.3 bs) on the work done at 100G with the addition of upper limits for skew rate of change tolerance.

Which interfaces can be impacted

- Chip to module case
  - From host to module ➔ variation would be due to thermal effects (low rate of change mUI/sec)
  - From module to host (gearbox case) ➔ variation would be due to thermal effects (low rate of change mUI/sec)
  - From module to host (no gearbox) ➔ variation would be due to fiber plant + thermal effects (higher rate of change? uI/sec)
- Module (optical input)
  - variation would be due to fiber plant + thermal effects
  - If module has a mux then it may have to tolerate this skew.
    - It can also be passed onto the host i/f
  - No mux may be skew transparent
- Module (copper)
  - Host must absorb skew

PMDs

- SR16
  - Uses 28Gb NRZ signaling
    - Re-use SR4 numbers from .ba => FFS?
    - No mux – skew transparent when using a CDAUI-16?
- FR8
  - Uses 28Gb PAM-4 signaling
  - May have mux if CDAUI-16 based (terminate skew)
- LR8
  - Uses 28Gb PAM-4 signaling
  - May have mux if CDAUI-16 based (terminate skew)
- PSM4
  - Uses 58Gb PAM-4 signaling => faster = 2x ul's per unit time
  - Will need a mux (at least in 1st generation) (terminate skew)
  - 2.4 ns skew variation (from PSM4 specification)

- Question => does PAM-4 change anything?

Skew and MUX's

- MUXs will be a major component of 400GbE pluggables (especially in 1st generation with CDAUI-16 electrical i/f)

- CDAUI 16 => 25Gb PAM4 (TX from host to module)
  - 2 x 25G NRZ lanes into each PAM-4 lane
  - Must have skew variation tolerance between pairs of 25G NRZ lanes
  - Skew variation would arise from host SERDES changes
    - Slow => sub mUI/sec (thermal)?
    - Limit range => few uI
- 25Gb PAM4 => CDAUI-16 (from PMD to host)
  - If the MUX can support independent PAM-4 lane timing then the skew variation should be simply passed through the MUX as a skew variation in the respective pair of 25G NRZ lanes
  - Skew variation would arise from transmitter skew and physical medium (fiber)
    - Rate of change => FFS
    - Range => great => several 10's uI (FFS)

- Architectural implementation may require all input lanes to be on same clk + phase (this can impact skew tolerance)

Fiber effects

- Two main contributors:
  - dynamic skew due to parallel fibre cables being stretched/temperature changes, and wavelength changes in conjunction with chromatic dispersion
  - See => http://www.ieee802.org/3/ba/public/tools/index.html

- There are other effects which could be faster but they're probably very small in magnitude, if not negligible: Polarisation mode dispersion; microphony (ie changes in length due to noise and/or vibration, because these tend to be important only for short periods of fibre, the perturbation necessarily gets smaller with higher frequency); Laser wavelength changes due to patterning and laser self heating.

Work to be undertaken (including FFS)

- Validate if the values used at 100G (table 80-5) can be used for the 25Gbaud based interfaces (CDAUI-8, CDAUI-16, SR16, FR8, LR8)
- Determine what values should be used for 50 Gbaud (DR4) based PMD
- Determine rates of change of skew that should be tolerated across a reworked table 80-5 for 400G applications.

Conclusions

- The range (limits) of skew variation on the 25Gb signaling interfaces should reuse values from table 80-5
- Limits on the rate of change of skew show be given
  - The figures for rate of change of skew are topics FFS