Clause 73 exchange of DME frames not needed for front-panel ports

Piers Dawe Avago Technologies

Supporters

Scott Kipp	Brocade
David Cunningham	Avago Technologies
Tom Palkert	Luxtera

Relates to D2.0 comment 74

Why this presentation

- A bunch of backplane-oriented features and methods have been swept up in the CRn proposal
  - Just one aspect (CX4) in recent reflector thread
- Things designed for backplane are now in the draft for a front-side port
- Some consequences need deeper review: is this really what we want?
- Questions to answer
  - How to take a port designed for QSFP or CFP optical, remove the optical module, and plug in an electrical module with a CRn PMD
  - How to stop a port designed for QSFP or CFP CRn driving an optical module with 10G CDRs mad with its 156.25 Mb Ethernet signalling
  - How to auto-negotiate a mix of 4 x 10G and 40G links e.g. going through QSFP. Or don't do it?
- Also, smoothing the way to compatibility between Ethernet and Fibre Channel at the next speed
- This presentation doesn't answer all these questions but moves the discussion along

Background

- In addition to the normal MAC-to-MAC communication function, Backplane Ethernet has three additional functions, together called "Auto-negotiation"
  - Exchange of DME frames
  - Training
  - Parallel Detection
- Optical PMDs have never had these features
- **Do not wish to burden a host** that might support e.g.
  - 40GBASE-SR4 and 40GBASE-CR4 in the same QSFP socket
  - or 100GBASE-SR10 and 100GBASE-CR10 in the same CXP socket
  - or 40GBASE-SR4, 40GBASE-LR4, and 40GBASE-CR4 in the same (XLAUI) socket
  - or 100GBASE-SR10, 100GBASE-LR4, 100GBASE-ER4 and 100GBASE-CR10 in the same (CAUI) socket

Exchange of DME frames 1/1

- Clause 37 AN has a bad reputation so 802.3ap wrote a new Clause 73 AN
- Exchange of DME frames uses differential Manchester encoding at 3.2 ns (312.5 MBd, 156.25 Msymbols/s), a common factor of the 1000BASE-X, 10GBASE-X and 10GBASE-R (lane) signalling rates
- Quite heavy-duty signalling scheme with frame formats, state machines and so on
- Used to advertise "technology ability" (1000BASE-KX and/or 10GBASE-KX4 and/or 10GBASE-KR) and FEC and Pause ability
- The hierarchy of "technologies" is predefined
- Highest "technology" available at both ends is chosen

Exchange of DME frames 2/3

This heavy-duty protocol does just three things

1. Choosing signalling speed
   - Not relevant for most optical PMDs which have incompatible power levels
   - There is only one front-panel speed choice to be made: 10GBASE-CX4 vs. 40GBASE-CR4. CX4 doesn't know about Clause 73, so this choice is made by "Parallel detection", not DME signalling. 1000BASE-X is not relevant
   - Not necessary; Fibre Channel use a much simpler scheme, similar to "Parallel detection", to choose between e.g. 2GFC, 4GFC, 8GFC. See later
2. Choosing whether to use FEC
   - Interworking between FEC-enabled ports and ports without FEC can be done by "Parallel detection" or the Fibre Channel method
   - Should not involve PMD or PMA at all; this can be worked out by the PCS by observing incoming signal coding
3. Advertising Pause ability
   - Do not see the point of this; not a PHY feature at all; could be done after the link is up with Slow Protocol frames (same signalling method that Pause uses), or LLDP
   - Optical ports don't advertise Pause this way (or at all?)

Exchange of DME frames 3/3

- Requires a CDR that can work at 1/33 of usual signalling rate; ordinary CDRs won't; will lose lock, raise loss of lock alarm, possibly squelch the DME signal
- Requires a wire from PCS (in host) to wherever the AN function is (for CFP and optics-oriented QSFP, in module) to carry AN_LINK.indication.
  - Shown as below the PMD, may be combined with it in practice
  - Timing too fast to use management registers
  - CFP and QSFP don't have a pin for this
- **Exchange of DME frames is complicated and onerous**
  - **Unnecessary and should not be required on front-panel ports**
- **This is the 40 Gb/s and 100 Gb/s Ethernet Task Force – don't want to be developing 156.25 Megabit Ethernet!**

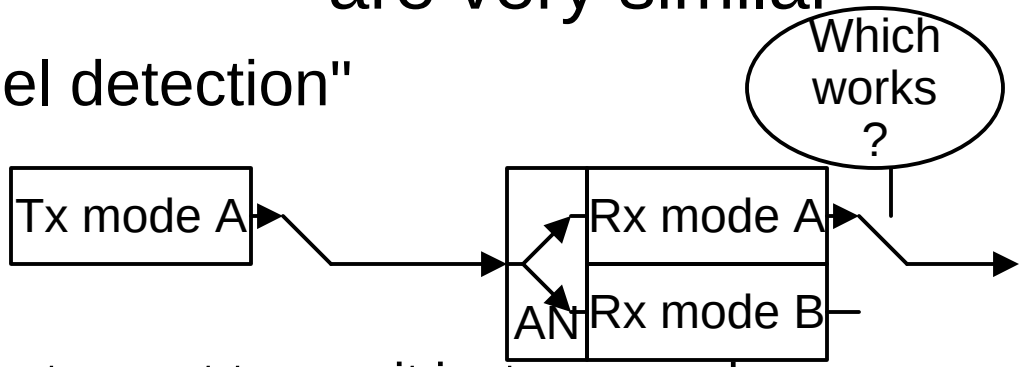
Parallel detection

- See 73.7.4.1
- Involves a receiver listening to a signal and trying to match it to the kinds of signal it can receive (we assume its transmitter can transmit the same kinds)
- This is the first step towards the Fibre Channel link startup method, where both transmitter and receiver try different kinds

- For description of Training, see backup slide

"Parallel detection" and "Link Speed Negotiation" are very similar

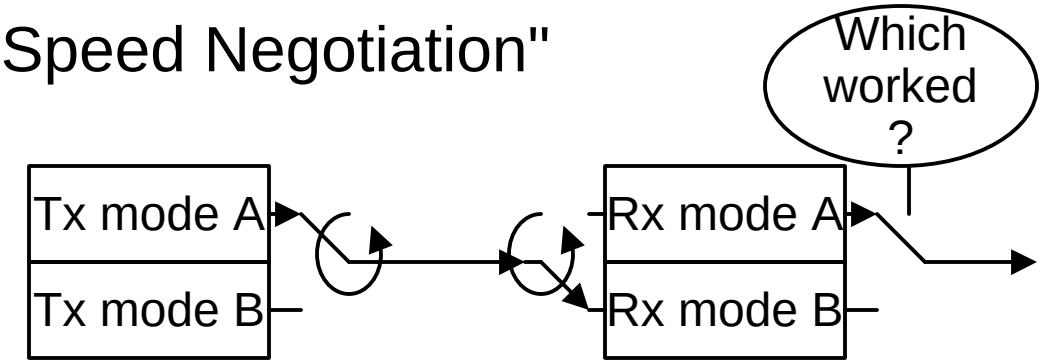
- "Parallel detection"



- One party must transmit just one mode
- Are the start-up timings defined?

See next slide for details

- "Link Speed Negotiation"



AAAABBBBAAAABBBBAAAABBBB ABABABABABABABABABABABAB

- Can negotiate coding, not just speed. Can be physically same Tx, Rx
- Transmitter cycles through its options slowly, receiver quickly
- Cycle rates (range) and total sequence time to be specified (FC has values for 8B/10B coding, is developing them for 64B/66B coding)

Fibre Channel way

- See FC-FS-2 Clause 26, Link Speed Negotiation*
- Transmitter cycles slowly through up to three or four transmit modes
  - In FC these would be signalling rates, e.g. 2G, 4G, 8G, 16G
- Receiver cycles more quickly through its receive modes
- Timings are defined so that at least four receive modes (if available) are tried against every transmit phase
- Transmitter cycles round three? times
- At the end, each side transmits in the highest-priority mode that it successfully received
- Can be used for choosing signalling speed
  - 10GBASE-CX4 vs. 40GBASE-CR4
  - As CX4 doesn't know about this (or Clause 73), it transmits and receives in the only language it knows. Works same as "Parallel Detection"
- Can be used for choosing whether to start with FEC on and off
  - CDR remains in lock all the time, receive PCS/FEC tries to parse the signal different ways e.g. by looking for lane markers
- **Provides good long-term structure for standard** for possible features in the future
  - e.g. distinguishing between a single 10GBASE-R lane and one in a 40GBASE-R4 or 100GBASE-R10 group
- **A port with only one mode doesn't need to do anything special**: e.g. 10GBASE-CX4; it transmits the only way it knows how and receives the only way it knows how
  - Only requirement is that its signal detect, lock detect and other behaviours do not go mad when receiving a cycling signal
  - Can recommend timing limits to make this simple
- * http://www.t11.org/ftp/t11/pub/fc/fs-2/06-085v3.pdf FC-FS-3 for 64B/66B 16GFC is in draft

Conclusion

For front panel ports (-SR*n*, LR*n,* ER4, CR*n*):

- Exchange of DME frames is a sledgehammer to crack a nut
  - CFP and QSFP sockets do not have a contact for AN_LINK.indication, can't use Clause 73 AN (with PCS in host and PMD in module) as it stands, even for CRn
  - Do not require exchange of DME frames on front panel ports
- Connection to optical module (nAUI or PPI) doesn't use Training
  - Does Training add value for CR*n*?
    - If it increases cable length by e.g. just 10%, don't use it
- "Parallel detection" is a simple form of Fibre Channel's Link Speed Negotiation
- Use optional Link Speed Negotiation method to allow interoperation with legacy PHY types
  - And if desired, distinguishing between 4 x 10GBASE-R and one, 40GBASE-R4
- Use Link Speed Negotiation method to allow interoperation between FEC and non-FEC ports

Non-FEC ports with only one speed don't need any form of AN

Backup: Training

- Provided to allow receiver equalizer to train itself on a signal so distorted that the receiver might not learn what to do on a regular scrambled signal
- Defined in 72.6.10.2
- Uses full-rate (10.3125 GBd) PRBS11 and differential Manchester encoded handshaking at 1/4 the Baud rate
- Allows a receiver to ask its peer transmitter to change its emphasis (and amplitude?) during the training phase, before link starts up
  - Receiver can receive at some SNR without training adjustments, but altering the transmitter spectrum might improve SNR
- 10GBASE-LRM does not use such a thing
- CR*n* channels (cables) expected to be more consistent than backplane channels; should be able to set emphasis for a long cable and it will be fine with a short cable
- Why not just transmit the best signal to start with?
  - If wished, could handshake using Slow Protocol frames to turn the transmitter down after link has started
- Would like to see evidence that training is really needed for CR*n*