Auto-Negotiation for 25G-CR Cables

Venu Balasubramonian, Marvell

Matt Brown, APM

Objective

- Proposal for an Auto-Negotiation scheme and PMD designation for 25GbE passive copper cables that is
  1. In alignment with the approved objectives for the Task Force
  2. Consistent with use of AN in 802.3

Objectives for Passive copper: A recap

- Define a single-lane 25 Gb/s PHY for operation over links consistent with copper twin axial cables, with lengths up to at least 3m
- Define a single-lane 25 Gb/s PHY for operation over links consistent with copper twin axial cables, with lengths up to at least 5m

- Two PMD types:
  1. Type 1: Capable of driving at least 3m
  2. Type 2: Capable of driving at least 5m

Drivers behind a separate 3m objective

- 3m reach is sufficient for server to switch links in Cloud scale datacenter environments
  - Fully engineered links
  - Power and latency are premium
- Additional latency and power introduced by the CL91 FEC (or an FEC of equivalent gain) an unnecessary overhead

Assumptions for the 3m CR PMD

- If possible, specify without FEC to minimize latency
- If CL74 FEC is required, this can be negotiated using FEC ability bit as is used for various PHYs
- Host loss budget could potentially be lower than the 802.3bj host budget

Drivers behind a separate 5m objective

- Compatibility with 100G equipment with 100GBASE-CR4 capability and 4x25G modular interconnect
- Provides reach beyond a single rack for efficient use of high-port-density switches

Assumptions for the 5m CR PMD

- All specifications are similar to those for 100GBASE-CR4
- CL91 FEC is always required in the same way as for 100GBASE-CR4.
  - Transmitter always encodes FEC
  - Receiver has the option to correct errors or not to save some latency when possible

Auto-negotiation in 802.3

- Technology ability advertised by the two ends of a link
  - Highest common ability picked based on predefined priority resolution
- Advertised technology ability controlled by the higher layer control function (system user, pervasive management)
  - Based on the specific requirements of an application, user may choose not to advertise an ability even if it's supported by the underlying hardware

AN proposal for 25G-CR cables

- Add two new technology-dependent PHY types (placeholder nomenclature)
  1. 25G-CR-S (higher priority)
  2. 25G-CR-L
- 25G-CR-S: A PMD capable of supporting a total link budget consistent with the 3m objective
- 25G-CR-L: A PMD capable supporting a total link budget consistent with the 5m objective

Update technical abilities

Table 73-4 – Technology Ability Field encoding

Bit	Technology
A0	1000BASE-KX
A1	10GBASE-KX4
A2	10GBASE-KR
A3	40GBASE-KR4
A4	40GBASE-CR4
A5	100GBASE-CR10
A6	100GBASE-KP4
A7	100GBASE-KR4
A8	100GBASE-CR4
A9	25GBASE-KR
A10	25GBASE-CR-S
A11	25GBASE-CR-L
~~A9~~A12 through A24	Reserved for future technology.

Update priority resolution

Table 73-7 – Priority Resolution

Priority	Technology	Capability
1	100GBASE-CR4	100GBASE-CR4 100 Gb/s 4 lane, highest priority
2	100GBASE-KR4	100 Gb/s 4 lane
3	100GBASE-KP4	100 Gb/s 4 lane
4	100GBASE-CR10	100 Gb/s 10 lane
5	40GBASE-CR4	40 Gb/s 4 lane
6	40GBASE-KR4	40 Gb/s 4 lane
7	25GBASE-CR-S	25 Gb/s 1 lane
8	25GBASE-CR-L	25 Gb/s 1 lane
~~7~~9	10GBASE-KR	10 Gb/s 1 lane
~~8~~10	10GBASE-KX4	10 Gb/s 4 lane
~~9~~11	1000BASE-KX	1 Gb/s 1 lane, lowest priority

Usage scenario 1: Server to switch links on Cloud-scale Datacenters

- Server and switch advertise only CR-S capability
  - Engineered links with cables and host budgets guaranteed to meet the 3m objective
  - Inability to link up indicates link issues
    - Bad connector/cable etc.
    - Thus using the CR-L to try and link up under this scenario not a logical option

Usage scenario 2: Server to switch links in an enterprise data center

- Both switch and server support only the CR-S PMD type
  - Identical to the cloud scale datacenter usage scenario
- Only one of the end points supports CR-L PMD type
  - End user allowed to use only a CR-S (<3m) cable
  - The CR-S only end point advertises only CR-S, and the link Auto-negotiates to CR-S mode of operation
- Both end points support CR-S and CR-L PMD types (cable type plugged in unknown)
  - Case 1: “Aggressive” configuration
    - Advertise both CR-S and CR-L capabilities and try to link up in CR-S mode
    - Re-negotiate with CR-L only if link up fails
  - Case 2: “Conservative” configuration
    - Start with CR-L and switch to CR-S only if link partner ability limited to CR-S

Usage scenario 3: Inter-rack links

- Require both switch and server to be CR-L capable
  - End user allowed to plug in a CR-L (>3m) cable only if supported on both ends of the link
  - Case 1: Both CR-L and CR-S advertised
    - Link up attempted with the highest priority mode (CR-S)
    - Falls back to CR-L on failure to link up in CR-S mode
  - Case 2: Only CR-L advertised
    - Link up attempted and established with CR-L mode

Usage scenario 4: Systems with asymmetric host losses

- Asymmetric host losses can be supported only on engineered links
  - Need to guarantee that the total budget is within the CR-L spec limit
  - Try to link up with CR-L mode
    - Use link quality monitor (FEC statistics) to decide if link up successful
    - Report to higher layers if link up unsuccessful after multiple ANs

Summary

- Auto-negotiation based on two distinct PMD types covers all major usage scenarios
- Consistent with objectives, and traditional use of Auto-negotiation
- Consistent with what is physically possible based on the capabilities of port types targeted to meet the approved task force objectives
- Optimizes and simplifies PMD for end use cases

Thank You