Continuing Work on Inner FEC Bypass: Path Forward

Brian Welch (Cisco)

Zvi Rechtman (Nvidia)

Piers Dawe (Nvidia)

Supporters

- Chris Cole (Quintessent)
- Drew Guckenburger (Maxlinear)
- Frank Chang (Source Photonics)
- Karl Muth (Broadcom)
- Vasu Parthasarathy (Broadcom)
- Adee Ran (Cisco)
- Phil Sun (Credo)
- Lenin Patra (Marvell)
- Ed Ulrichs (Intel)
- Sridhar Ramesh (Maxlinear)
- Roberto Rodes (Coherent)
- Mark Kimber (Semtech)
- Vipul Bhatt (Coherent)
- Mike Li (Intel)
- Jamie Gaudette (Microsoft)
- Scott Schube (Intel)
- Wilson Zhang (Innolight)
- Tom Palkert (Samtech/Macom)
- Karen Liu (Nubis Communications)
- Jamal Riani (Marvell)

Overview

- In May task force meeting, considerable support was shown for inner code bypass (welch_3dj_03c_2305.pdf)
  - Explored the motivations (latency and power) and mechanisms for bypass
  - Polling (shown on next slide) showed a high level of interest
- In the optical ad-hoc June meeting, dudek_3dj_optx_01_230629.pdf suggested an alternative method to bypass inner-FEC
  - The discussion showed a support for bypassing option as well
- In both May Interim and June ad-hoc, questions/discussion focused on:
  - Method of specification: Common or unique PMD specifications
  - Method of determination: Auto-negotiation, auto-detection or Management interface only

https://www.ieee802.org/3/dj/public/23_05/motions_3cwdfdj_2305.pdf

Straw Poll Results from May Interim

Straw Poll #13

I am interested in working towards enabling an inner code FEC bypass approach for 200 G/lambda IMDD optics

A. all single wavelength
B. multi-wavelength 2km
C. none
D. NMI
E. abstain

(chicago rules)

results: A: 76 , B: 61, C: 19, D: 22, E: 11

Path Forward - Discussion

- **Should we enable inner_FEC bypass**?
- **How should optical compliance in bypass mode be specified?**
- **How should the bypass mode be initiated?**
- **Intent of this presentation is to explore each of these topics**

Value Proposition

Latency and Power

Value Proposition - Latency

- Many AI/ML fabrics uses “Lossless” traffic as a key performance factor to obtain better network utilization
  - “Lossless” – usage of Flow Control (FC)/802.3 Annex 31B or Priority Flow Control (PFC)/802.1Q
  - Link Layer assign buffers to absorb links RTT (at least) for FC/PFC
- Consider 102.4T switch – 256 ports of 400GE
  - Every 1nsec latency increment result with ~25KB buffer size increment per Priority.
- Inner FEC has ~150nsec latency for 400GE port*

Number of Priorities per port	Additional buffers needed in the Switch due to RTT increment	Comments
1	~4MB	About Few % of total Switch buffers in a typical switch.
2	~8MB	
4	~15MB	
6	~23MB	

*Assuming Convolutional Interleaving is used

Value Proposition - Power

- Inner FEC expected to increase module power by 5-10%
  - Additional logic of inner FEC
  - Additional power for unique line slide clock domain
  - Increased analog power when running at higher overhead

- Effects on high density systems can be dramatic
  - Net power due to optics for 102T host (including overhead and cooling of optics) ~ 2.2 kW (assuming 26W for a 1.6T optical module)
  - System power savings of 180W(5%) - 340W(10%) possible with FEC bypass

Method of Specification

Optical Compliance Point Specifications

- **No Specification**: During link startup make determination of link operating margin, and decide operating mode based on that.
  - **Potential Advantages**: May allow for more links to operate in FEC_bypass given that typical link margin generally exceeds worst case link margin.
  - **Potential Disadvantages**: No guaranteed compliance point performance in bypass mode. Accommodations of link degradation due to environmental corners and aging unclear. Method (and location) of determining link health unclear.

- **Integrated Specification**: Common spec defining both FEC enable and bypass operation.
  - **Potential Advantages**: Guaranteed compliance point specs, ensuring margin for environmental degradation and aging.
  - **Potential Disadvantages**: Optics required to meet on worst case part, may allow for fewer bypassed links than prior option.

- **Discrete Specification**: Separate specs for FEC enable and bypass operation.
  - **Potential Advantages**: May be simpler to document than an integrated specification.
  - **Potential Disadvantages**: May be procedurally more complicated with current set of objectives.

Integrated Specification

Template

Proposed Transmitter Specifications

Description	800GBASE-xR4 *RS(544,514)* | *+Hamming(128,120)*		Unit
Signaling rate, each lane (range)	106.25 ± 50 ppm	113.4375 ± 50 ppm	GBd
Modulation Format	PAM4		
Lane wavelengths (range)	Value(s)		nm
Side-mode suppression ratio (SMSR), (min)	Value		dB
Average launch power, each lane (max)	Value		dBm
Average launch power, each lane (min)	Value		dBm
Outer Optical Modulation Amplitude (OMA_{outer}), each lane(max)	Value		dBm
Outer Optical Modulation Amplitude (OMA_{outer}), each lane(min) for TDECQ < 1.4 dB for 1.4 dB ≤ TDECQ ≤ TDECQ (max)	 Value Value		 dBm dBm
Transmitter and dispersion eye closure (TDECQ), each lane (max)	Value	Value	dB
TECQ (max)	Value	Value	dB
| TDECQ - TECQ | (max)	Value	Value	dB
Average launch power of OFF transmitter, each lane (max)	Value		dBm
Extinction ratio, each lane, (min)	Value		dB
Transmitter transition time (max)	Value		ps
Transmitter over/under-shoot (max)	Value		%
RIN_xOMA (max)	Value		dB/Hz
Optical return loss tolerance (max)	Value		dB
Transmitter reflectance (max)	Value		dB

Different rate & TDECQ/TECQ options to reflect different classes of transmitters

Proposed Receiver Specifications

Description	800GBASE-FR4 *RS(544,514)* | *+Hamming(128,120)*		Unit
Signaling rate, each lane (range)	106.25 ± 50 ppm	113.4375 ± 50 ppm	GBd
Modulation Format	PAM4		
Lane wavelengths (range)	Value(s)		nm
Damage threshold, each lane	Value		dBm
Average receive power, each lane (max)	Value		dBm
Average receive power, each lane (min)	Value		dBm
Receive power, each lane (OMA_{outer}) (max)	Value		dBm
Receiver reflectance (max)	Value		dB
Receiver sensitivity (OMA_{outer}), each lane (max)			
for TECQ < 1.4 dB	Value		dBm
for 1.4 dB ≤ TECQ ≤ SECQ	Value		dBm
Stressed receiver sensitivity (OMA_{outer}), each lane (max)	Value	Value	dBm
Conditions of stressed receiver sensitivity test:			
SECQ	Value	Value	dB
OMA_{outer} of each aggressor lane	Value		dBm

Different rate & SECQ/SRS requirements to accommodate the range of optical transmitters

Proposed Link Budget

Description	800GBASE-FR4		Unit
	RS(544,514)	*+Hamming(128,120)*	
Power budget (for max TDECQ)	Value	Value	dB
Operating distance	Value		m
Channel insertion loss	Value		dB
Maximum discrete reflectance	Value		dB
Allocation for penalties (for max TDECQ)	Value	Value	dB
Additional insertion loss allowed	Value		dB

Different power budget and penalties to reflect the different classes of transmitters

Method of Determination

Bypassing Methods

- Management Interface
  - In case of discrete Specification
    - Not much to do, PMD “capabilities”/“selection” already defined in MDIO and CMIS’s applications.
  - In case of Integrated Specification
    - New Bypass capability/selection controls
    - “Simple” 1 control bit per RX and TX.
  - Auto-Negotiation (via padding or other methods) may define additional Management status/controls similar to Clause 73 MDIO registers.
- Probably Management Interface is less controversial topic

Bypassing Methods

- Plug&Play Interoperability
  - Two main approaches – Auto detect and Backchannel/Auto-negotiation
  - Auto-detect
    - Tx can send either Inner-FEC enabled or bypassed, based on its performance
    - RX should be able to auto-detect both possible modes
  - Auto-Neg
    - Backchanneling and “agreement” between both ends on the most appropriate mode of operation
    - Several alternatives:
      - Use inner FEC padding bits
      - Between the PCSs
      - More alternatives..
- Based on ad-hoc discussion, Plug&Play method seems to require additional consensuses building work and future contributions.

Summary

Three separate discussions being had

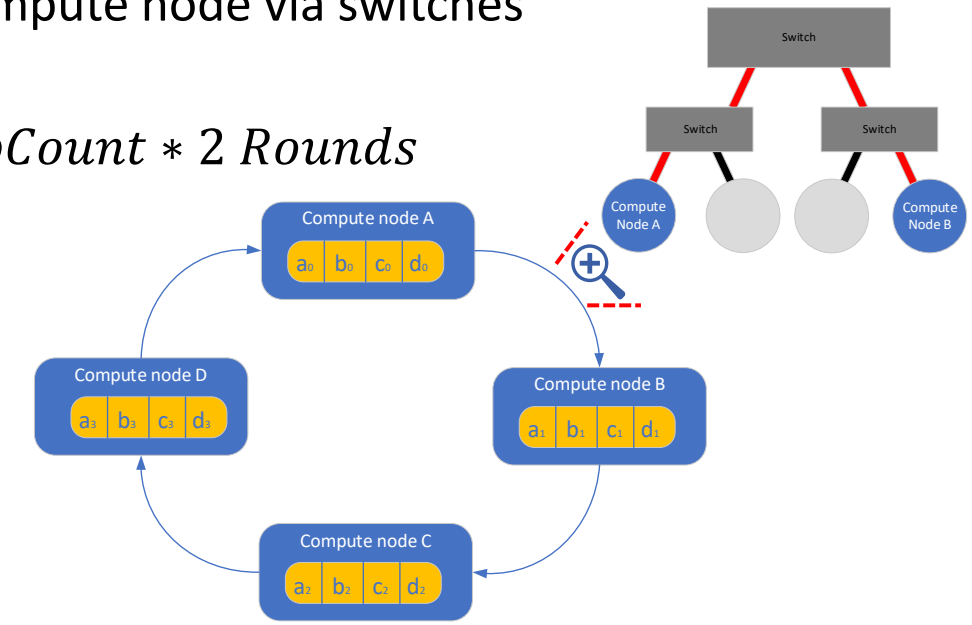
- **Should we enable inner_FEC bypass**: Seems to be strong consensus that we should
- **How should the bypass mode be specified**: Three options discussed thus far (No Spec, Integrated Spec, Discrete Specs)
- **How should the bypass mode be initiated**: Auto-Negotiation, Auto-Detection, or Management control interface only.
  - Many variants of Auto-Negotiation possible.
  - Host override via control interface expected for any case.

Backup

Ring AllReduce / AllGather

- Common traffic pattern in distributed AI training
- Used to communicate the model gradients from each compute node (e.g. GPU) to all others.
- Each compute node in the ring is connected to the other compute node via switches
  - Can go up to 5 Links between compute nodes in some cases
- Ring accumulate latency: $HopLatency * NumNodes * HopCount * 2\ Rounds$

Number of compute nodes	Additional accumulated latency
1200	~2msec
128	~150usec



- Some AI workloads like Convolutional Neural Network (CNN) might have computational time on the order of 10's of msec
  - Additional 2 msec of networking latency are tangible in such usecase.

Ring AllReduce /AllGather simulation

- Simulation of NCCL AllReduce of 128 compute nodes
- Bus BW = effective BW = Number of computed bytes / total completion time
- Latency impact is depended on the message size:
  - For short messages – The network latency is dominate.
  - For Large messages – The limitation on based on network and computational BW.
- Typical message sizes are between 50MB (~2^26) (ResNET/Mlperf) to few GB in Large Language Model (LLM)
  - For 64MB message size – the effective BW degradation is about 20%
- The amount of latency addition has direct impact on BW slowdown.
  - Which translates to “right shift” in the BW (S-curve)