100G Link Training Proposal

Kent Lusted, Intel

Ilango Ganga, Intel

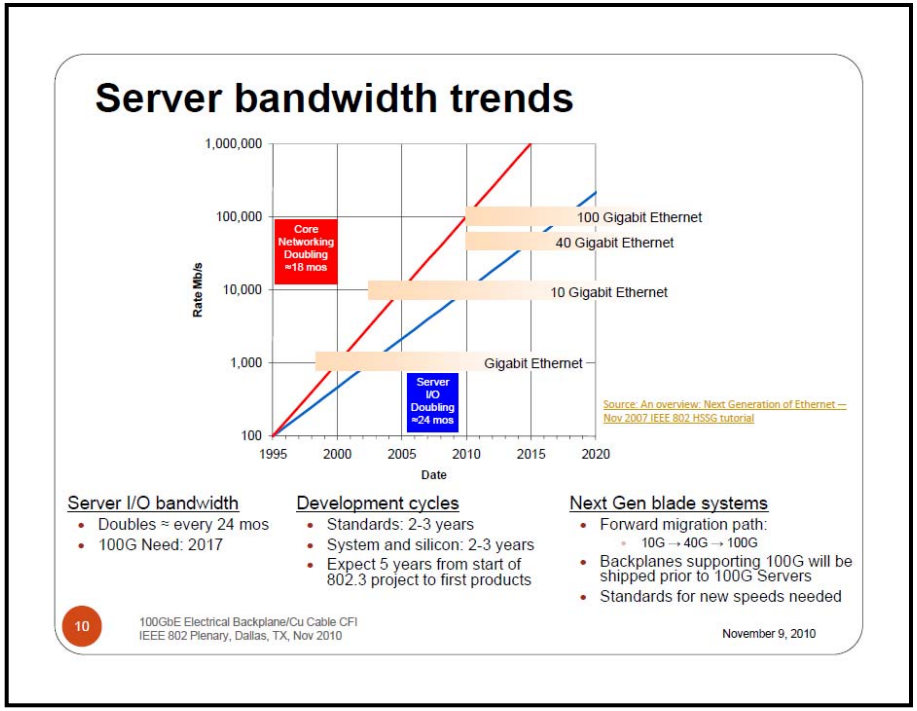
Changes from lusted_01_0512.pdf are shown in this color

Supporters

- Vasu Parthasarathy, Broadcom
- Scott Kipp, Brocade
- Arthur Marris, Cadence
- Bhavesh Patel, Dell
- Brad Booth, Dell
- Adee Ran, Intel
- Dave Chalupsky, Intel
- Rich Mellitz, Intel
- Andre Szczepanek, Inphi
- Oren Sela, Mellanox

100G Adoption

- First adopters of new PHY PMDs typically are customers seeking:
  - Higher bandwidth
  - Lower latency
  - Higher port count
- Latency degrades high performance applications
- FEC adds latency
- **To minimize latency & power penalty, vendors may not want to enable FEC unless it is required for the channel**



Source: http://www.ieee802.org/3/100GCU/public/nov10/CFI_01_1110.pdf

FEC Enablement Today

- Currently Cl 74 FEC is enabled during Auto-Neg if both link partners advertise FEC ability and >=1 request FEC
  - Receiver may not know if FEC is needed for the channel until after AN completes
  - Receive adaptation process is not defined in KR
- 100G CR4 channel adopted baseline needs FEC to meet 35dB loss objective
  - http://www.ieee802.org/3/bj/public/mar12/diminico_01a_0312.pdf
  - Some QSFP cables have EEPROM with optional loss fields in the memory
  - http://www.ieee802.org/3/bj/public/mar12/dudek_02a_0312.pdf
- 100GBASE-KR4 NRZ backplane channel adopted baseline needs FEC for the higher loss channel (35 dB)
  - No standardized way for PHY to know the channel
  - http://www.ieee802.org/3/bj/public/mar12/dudek_03_0312.pdf

FEC Enablement Proposal

- Defer decision of FEC enable to the link training process
  - Move FEC decision to PMD sublayer instead of AN
  - Receiver could implement link quality check in the receive adaptation process
    - How an RX determines and defines link quality is out of scope
  - Based on the results of the check, the local receiver tells link partner TX to enable or disable FEC
    - FEC enablement means enabling transcoding and FEC encoding data instead of leaving data in 64B/66B coded form
    - If *any* of the 4 physical lanes RX request FEC from the link partner TX, then *all* 4 physical lanes would get FEC encoded data from the TX.
  - Decision is communicated to link partner via control channel of link training protocol
- Add option to force FEC on or off
- **Use FEC in the path only if the receiver needs it for <u>that</u> channel!**
  - Permits symmetric and asymmetric FEC operation for reduced round trip latency ☺

Impact to AN Baseline

- Redefine 1 bit (D45/A24) in Link Codeword Base Page for FEC_Defer (F2)
  - If NRZ FEC is optional to implement:
    - If 0, then local device does not support NRZ FEC
    - If 1, then local device does support NRZ FEC
  - If NRZ FEC is mandatory to implement: then bit is not needed
- Redefine 1 bit (D44/A23) in Link Codeword Base Page for FEC_Force (F3)
- Leave current FEC Ability (F0) and FEC Request (F1) as is
  - Will apply only to Clause 74 FEC used with 10GBASE-KR, 40GBASE-KR4, 40GBASE-CR4, and 100GBASE-CR10
  - If PMD selected is 100GBASE-CR4, 100GBASE-KR4 (NRZ) then bits are don't care
- No bit for FEC clause 94 (PAM4) because FEC is mandatory for that PMD
  - PMD should treat F0, F1, F2 bits as don't care

* From IEEE P802.3REV D3.0

New Link Codeword Base Page

D 0	D 1	D 2	D 3	D 4	D 5	D 6	D 7	D 8	D 9	D 10	D 11	D 12	D 13	D 14	D 15
S 0	S 1	S 2	S 3	S 4	E 0	E 1	E 2	E 3	E 4	C 0	C 1	C 2	RF	Ack	NP

D 16	D 17	D 18	D 19	D 20	D 21	D 22	D 23	D 24	D 25	D 26	D 27	D 28	D 45	D 46	D 47
T 0	T 1	T 2	T 3	T 4	A 0	A 1	A 2	A 3	A 4	A 5	A 6	A 7	A 2	F 0	F 1

Figure 73–6—Link codeword Base Page

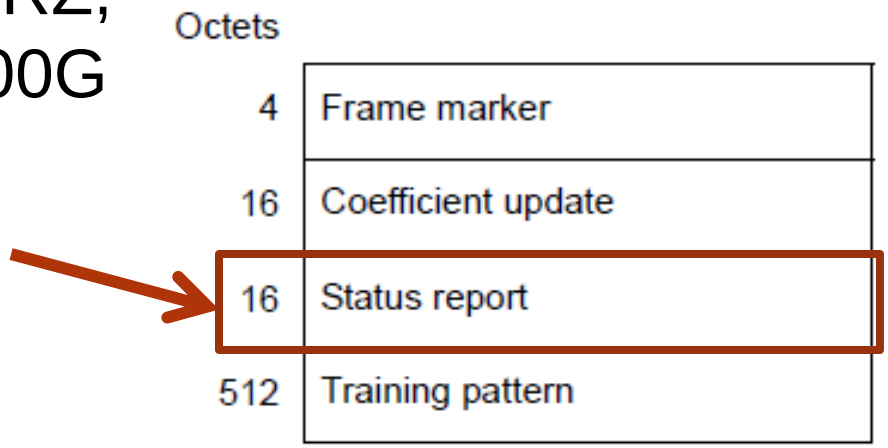
D45 = F2
D44 = F3

If NRZ FEC is mandatory to implement, then bit is not needed

* From IEEE P802.3REV D3.0

Basic Link Training Format

- Adopt basic format from 72.6.10 for 100G KR4 NRZ, 100G KR4 PAM4 and 100G CR4
- Add new 1 bit field to communicate local RX desired FEC Mode



* From IEEE P802.3REV D3.0

No Change to Coeff Update Field

Table 72–4—Coefficient update field

Cell(s)	Name	Description
15:14	Reserved	Transmitted as 0, ignored on reception.
13	Preset	1 = Preset coefficients 0 = Normal operation
12	Initialize	1 = Initialize coefficients 0 = Normal operation
11:6	Reserved	Transmitted as 0, ignored on reception.
5:4	Coefficient (+1) update	5 4 1 1 = reserved 0 1 = increment 1 0 = decrement 0 0 = hold
3:2	Coefficient (0) update	3 2 1 1 = reserved 0 1 = increment 1 0 = decrement 0 0 = hold
1:0	Coefficient (–1) update	1 0 1 1 = reserved 0 1 = increment 1 0 = decrement 0 0 = hold

* From IEEE P802.3REV D3.0

Update Status Report Field

- Add FEC Mode (Cell 14)
  - 0 = local RX determined that FEC is not required
  - 1 = local RX to use FEC (Enable FEC on link partner TX)

13:6

Table 72–5—Status report field

Cell(s)	Name	Description
15	Receiver ready	1 = The local receiver has determined that training is complete and is prepared to receive data. 0 = The local receiver is requesting that training continue.
~~14:6~~	Reserved	Transmitted as 0, ignored on reception.
5:4	Coefficient (+1) status	5 4 1 1 = maximum 1 0 = minimum 0 1 = updated 0 0 = not_updated
3:2	Coefficient (0) status	3 2 1 1 = maximum 1 0 = minimum 0 1 = updated 0 0 = not_updated
1:0	Coefficient (–1) status	1 0 1 1 = maximum 1 0 = minimum 0 1 = updated 0 0 = not_updated

* From IEEE P802.3REV D3.0

Example Implementation

Update Training Diagram

Define 4 new variables in state machine:
- local_rx_fec
- remote_rx_fec
- local_rx_fec_requested
- remote_rx_fec_requested

Change SEND_TRAINING:
- Add “local_rx_fec <= false”
- Add “remote_rx_fec <= false”

Change TRAIN_REMOTE:
- Add “local_rx_fec <= local_rx_fec_requested”

Change LINK_READY:
- Add “remote_rx_fec <= remote_rx_fec_requested”

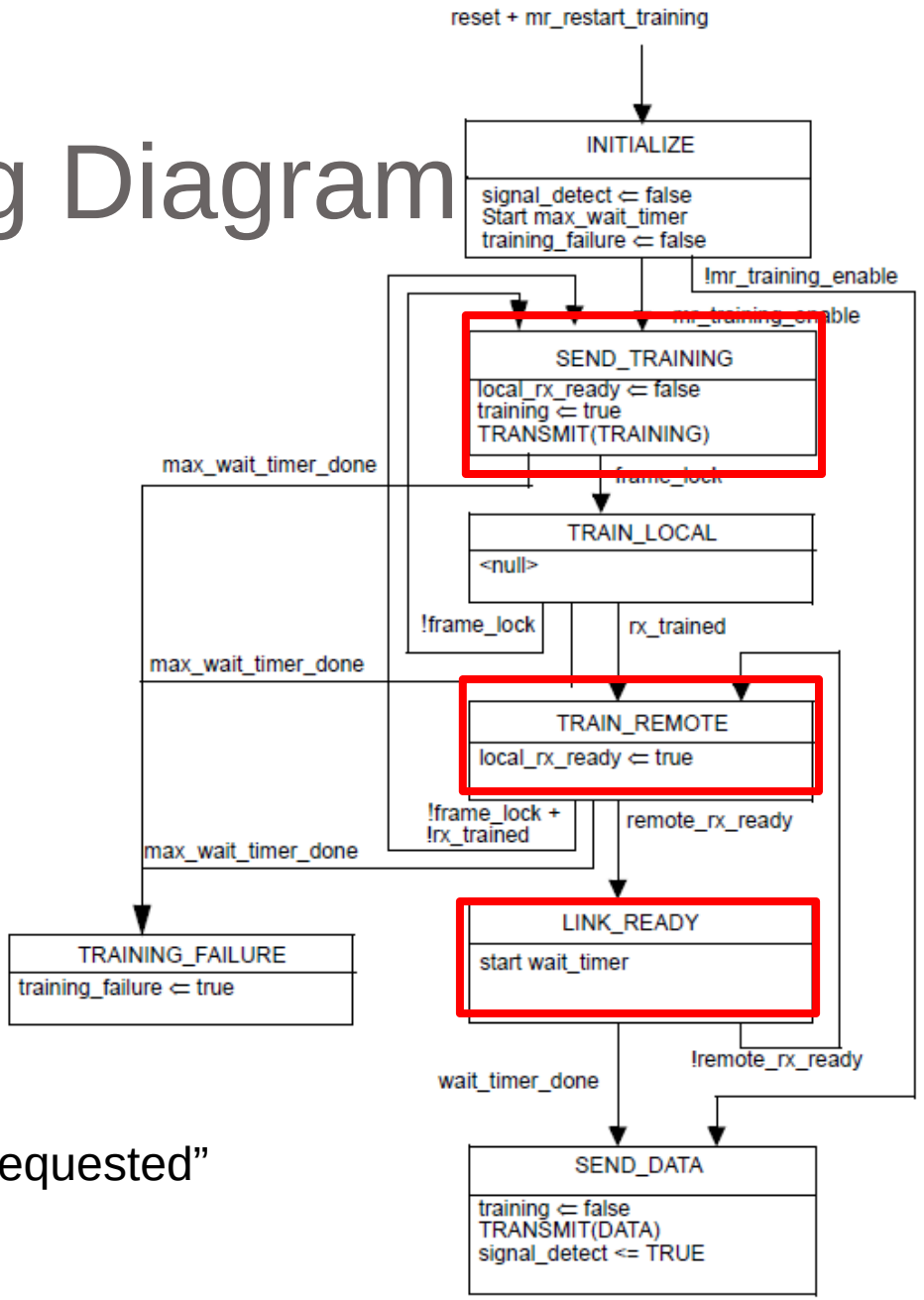


Figure 72–5—Training state diagram

* From IEEE P802.3REV D3.0

Example Implementation

Add Variables in Cl 72.6.10.3.1

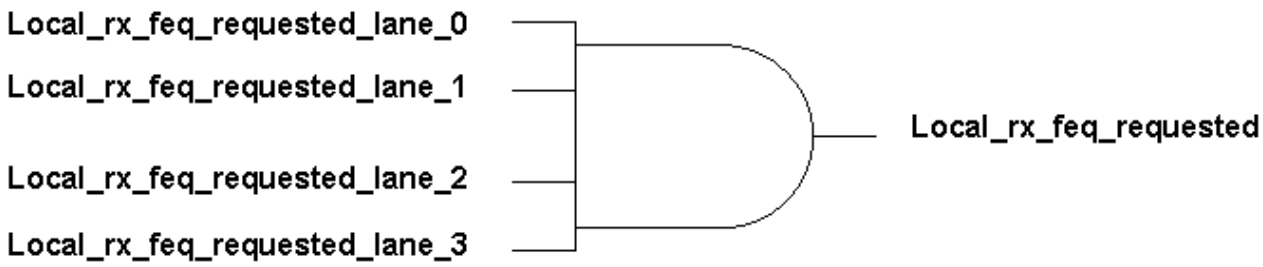
- local_rx_fec
  - Boolean variable that is set to local_rx_fec_requested by the training state diagram when rx_trained is asserted and is set to FALSE otherwise.
- remote_rx_fec
  - Boolean variable that is set to remote_rx_fec_requested by the training state diagram when remote_rx_ready is asserted and is set to FALSE otherwise.
- local_rx_fec_requested
  - Boolean variable that is set to TRUE if FEC is requested by the local receiver and is set to FALSE otherwise. The value is transmitted as the FEC mode bit on all outgoing training frames.
- remote_rx_fec_requested
  - Boolean variable that is set to TRUE if FEC is requested by the remote receiver and is set to FALSE otherwise. This value is received as the FEC mode bit on all incoming training frames.

* From IEEE P802.3REV D3.0

Example Implementation

Potential Changes:

- Service primitives to be defined in PMD Sublayer Interface to indicate the signaling to the FEC request/enable to the FEC sublayer
  - Add the PMD and what it is.
- FEC to be signaled 'enable' through service primitive if *any* of the lanes request to turn on FEC.
  - Individual lanes RX may indicate different FEC enable/disable indication but FEC is either enabled on all 4 lanes or none at all.
  - EX:



Example Implementation

Clause 30 Changes

- 30.5.1.1.15 aFECAbility
  - Change syntax to read as “unknown initializing or deferred, true state is not known
- 30.5.1.1.16 aFECmode
- Change syntax to read as “unknown initializing or deferred, true state is not known”
- 30.6.1.1.5 aAutoNegLocalTechnologyAbility
  - Insert FEC Defer after FEC Requested

* From IEEE P802.3REV D3.0

Example Implementation

Clause 45 Changes

- Update BASE-R LP status report, lane X register bit definitions and BASE-R LD status report, lane X register bit definitions
  - Registers 1.153, 1.155, 1.1201, 1.1202, 1.1203, 1.401, 1.402, 1.1403
  - In each table, add entry for FEC Mode (cell 14):
    - 0 = local RX determined that FEC is not required
    - 1 = local RX to use FEC (Enable FEC on link partner TX)
  - In each table, change reserved values to be range 1.x.13:6
- Create the appropriate subclause entries for FEC Mode
  - FEC Mode 1.X.14
    - The function and values for the FEC Mode bit is defined in Clause xx.x.x.x

* From IEEE P802.3REV D3.0

Example Implementation

Clause 45 FEC Control Register

- Update Register 1.171BASE-R FEC control register as follows

Bit(s)	Name	Description	R/W
1.171.15:4	Reserved	Value always zero, writes ignored	RO
1.171.3	Enable TX FEC	A write of 1 to this bit enables FEC in the transmitter A write of 0 to this bit disables FEC in the transmitter	R/W
1.171.2	Enable RX FEC	A write of 1 to this bit enables FEC in the receiver A write of 0 to this bit disables FEC in the receiver	R/W
1.171.1	FEC enable error indicati	A write of 1 to this bit configures the FEC decoder to indicate errors to the PCS layer	R/W
1.171.0	FEC enable	A write of 1 to this bit enables FEC A write of 0 to this bit disables FEC	R/W

- Add appropriate subclause entry for register 1.171.3 and 1.171.2

* From IEEE P802.3REV D3.0

Example Implementation

More Clause 45 Changes (2)

- Create a new register BASE-R PMD Status Register 4 (Register 1.158)
  - “The BASE-R PMD status 4 register is used for 100GBASE-CR4, 100GBASE-KR4 (NRZ) and other PHY types using the PMDs described in Clause xx supporting deferred FEC selection. The assignment of bits in the BASE-R PMD status 4 register is shown in Table xx”
  - Create a table with bits as shown on next page with the appropriate subclause entries by lane number.

* From IEEE P802.3REV D3.0

Example Implementation

New BASE-R PMD Status Register 4 (Register 1.158)

Bit(s)	Name	Description	R/W
1.158.8:15	Reserved	Value always zero, writes ignored	RO
1.158.7	Remote RX FEC Status 3	1 = remote RX determined that FEC is required for Lane 3 0 = remote RX determined that FEC is not required for Lane 3	RO
1.158.6	Local RX FEC Status 3	1 = local RX determined that FEC is required for Lane 3 0 = local RX determined that FEC is not required for Lane 3	RO
1.158.5	Remote RX FEC Status 2	1 = remote RX determined that FEC is required for Lane 2 0 = remote RX determined that FEC is not required for Lane 2	RO
1.158.4	Local RX FEC Status 2	1 = local RX determined that FEC is required for Lane 2 0 = local RX determined that FEC is not required for Lane 2	RO
1.158.3	Remote RX FEC Status 1	1 = remote RX determined that FEC is required for Lane 1 0 = remote RX determined that FEC is not required for Lane 1	RO
1.158.2	Local RX FEC Status 1	1 = local RX determined that FEC is required for Lane 1 0 = local RX determined that FEC is not required for Lane 1	RO
1.158.1	Remote RX FEC Status 0	1 = remote RX determined that FEC is required for Lane 0 0 = remote RX determined that FEC is not required for Lane 0	RO
1.158.0	Local RX FEC Status 0	1 = local RX determined that FEC is required for Lane 0 0 = local RX determined that FEC is not required for Lane 0	RO

* From IEEE P802.3REV D3.0

Example Implementation

BASE-R PMD Status Register 4 (Register 1.158) Entries

- Create subsection entries for each lane [0:3] after the new BASE-R PMD Status Register 4 (Register 1.158)
  - Local RX FEC Status [0:3]
    - When the PMD status register indicates that the receiver is trained, this bit maps to the state variable local_rx_fec for lane [0:3] as defined in xx.x.x.x.x
  - Remote RX FEC Status [0:3]
    - When the PMD status register indicates that the receiver is trained, this bit maps to the state variable remote_rx_fec for lane [0:3] as defined in xx.x.x.x.x

* From IEEE P802.3REV D3.0

Example Implementation

Clause 91 (NRZ FEC)

- Update appropriate section of clause 91 to support asynchronous operation

Summary

- To minimize latency & power penalties, vendors may not want to enable FEC unless it is required for the channel
  - Both 100GBASE-CR4 and 100GBASE-KR4 (NRZ) support 2 different channel limits based on FEC status
- Propose to move FEC decision for 100GBASE-CR4 and 100GBASE-KR4 (NRZ) to PMD sublayer instead of AN
  - Permits symmetric and asymmetric FEC operation
- Propose option to force FEC decision to on or off