CWMs IN IEEE 25GE (802.3by)

Eric Baden
Howard Frazier
Yong Kim
Rob Stone

Supporters

Mike Andrewartha, Microsoft
Vittal Balasubramanian, Dell
Andy Bechtolsheim, Arista Networks
Brad Booth, Microsoft
Jacky Chang, HP
Don Cober, CoMIRA Solutions, Inc
Bob Felderman, Google
Adil Haque, Altera
Jeff Hirschmann, Arista Networks
Tom Issenhuth, Microsoft
Alan Judge, Amazon
David Koenen, HP
Tom Kolze, Broadcom
Martin Langhammer, Altera
Mark Laubach, Broadcom,
Jon Lewis, Dell
Mike Li, Altera
Ron Muir, JAE
Bhavesh Patel, Dell
Rich Prodan, Broadcom,
Anshul Sadana, Arista Network
Rochan Sankar, Broadcom
Kapil Shrikhande, Dell
Tongtong Wang, Huawei
Xinyuan Wang, Huawei
Oded Wertheim, Mellanox
Joe Yohannan, Arista Networks

BACKROUND

- August, 2014, AMs first proposed as an option.
  - http://www.ieee802.org/3/25GSG/public/adhoc/architecture/gustlin_081214_25GE_adhoc.pdf
- October 2014, first specific proposal in SG
  - http://www.ieee802.org/3/25GSG/public/adhoc/architecture/baden_102214_25GE_adhoc.pdf
- Multiple other presentations in Ad hocs, Plenary, TF meetings, etc.
  - "PCS Thoughts and Considerations" - kim_100114_25GE_adhoc.pdf
  - "25G RS/PCS Considerations – A follow up" - kim_100814_25GE_adhoc.pdf
  - "Architectural Thoughts – 25G Interconnect" - booth_102914_25GE_adhoc.pdf
  - "25G Ethernet Layering and Gaps" - baden_25GE_01a_1114.pdf
  - "Architectural Thoughts – 25G Interconnect"- booth_25GE_01a_1114.pdf
  - …
- January Interim 2015 (baden_3by_01b_0115.pdf):
- **PCS and FEC Baseline Approved** (Y: 62 , N: 3 , A: 16)
  - CWMs are used for RS FEC locking
  - Constructed from concatenation of CL82 **40G** AM0, AM1, AM2, AM3

BACKROUND CONT'D

- March Plenary 2015 (*slavick_3by_01_0315.pdf*):
- **Amendment proposed and Failed** (Y: 31 , N: 11 , A: 35):
  1. Change to use Scramble and Test method for RS-FEC code-word delineation.
     1. Saves gate count (area and power) ~5%
     2. Matches method used by another standard [Fibre Channel] which already defines single lane RS-FEC solution (designs supporting both just rate scale)
     3. Mean Lock time is similar (CWM: 0.3ms v. SnT: 0.5ms)
     4. EEE wake lock method identical to Clause 74 – known data pattern during PCS scrambler bypass period1
     5. Re-use CL 74 PN-2112 scrambler (run over 5280b instead of 2112b)
- Other Considerations:
  7. OTN
  8. FlexE
  9. 1588

SUMMARY

- **There is no *compelling* reason to remove the CWMs.**
- **Such a change would not add value to the approved baseline.**
- **It would delay the project.**
- **The following slides address each of the items presented in the first proposal to remove the Code Word Markers:**

CONCERNs

WHAT PROBLEM ARE WE SOLVING WITH THE PROPOSED CHANGE?

- *WRT "Matches method used by another standard* [Fibre Channel] *which already defines single lane RS-FEC solution (designs supporting both just rate scale)"*
  1. This was never a part of our Broad Market Potential or Economic feasibility justification.
  2. IEEE AN mechanism and bit rates, training, data rates, and FEC selection are incompatible with FC.
     1. For instance, FC AN requires the SERDES to transmit at one rate and receive at a completely different rate.

- **CONCLUSION: Compatibility with Fibre Channel is not achieved and is not relevant.**

CONCERNs

WHAT PROBLEM ARE WE SOLVING WITH THE PROPOSED CHANGE?

- *WRT "Saves gate count (area and power) ~5% (**10K gates**)"*
  1. *See (slavick_3by_02_0315.pdf)*
     1. *CL108 FEC estimated to be 400K gates.*
  2. **Switches** support 100G already.
     1. All lanes lock to Alignment Markers (aka, CWMs). This is **FREE**.
     2. Switches support AM insertion and associated buffering in the PCS, for 40G, and 100G. **May require some flops per separate FEC PHY.**
     3. Switches support removal of AMs for 40G and 100G. This is **FREE**.
     4. Switches support Elastic Buffer for Clock Compensation. **May require 264 flops per 25G PHY**.
     5. Total estimate for Switch is **1K flops** (about 10K gates).
        1. A 32x100G (128x25G) Switch with 7 Billion transistors (7/4 = 1.75Billion Gates) adds 128*10K/1.7B **=** **.076%** total device gate count.
  3. **NICs** may support 100G already, maybe not.
     1. A 35Million Gate NIC might add 4 * 10000 = **.11428%** to the total device transistor count.

- **CONCLUSION. Negligible (if any) area difference.**

CONCERNs

WHAT PROBLEM ARE WE SOLVING WITH THE PROPOSED CHANGE?

- *WRT "Mean Lock time is similar (CWM: 0.3ms v. SnT: 0.5ms):"*
  1. The lock time is on the order of **2.5X** longer for the Test method vs. the CWM method.
     1. CWMs appear every **200us**
        1. **Two CWs required to 'lock' = 400us.**
     2. For SnT, 5280 bit slips takes about 1ms (**1.081344ms**)
  2. The CWM scheme tolerates errors in the CW (more robust). The SnT method does not.
     1. This can increase the lock duration in increments of 1ms.

- **CONCLUSION: CWM Lock Time faster and more robust than SnT approach.**

CONCERNs

WHAT PROBLEM ARE WE SOLVING WITH THE PROPOSED CHANGE?

- *WRT "EEE wake lock method identical to Clause 74 – known data pattern during PCS scrambler bypass period"*
    1. RCWMs can be shown to guarantee CW boundary lock within 400ns to 600ns, without scrambler bypass.
        1. RCWMs are identical in format to normal CWMs (as opposed to RAM)
        2. RCWMs do not require additional logic to identify unscrambled IDLEs and CW boundaries.
            1. http://www.ieee802.org/3/by/public/adhoc/architecture/cober_050615_25GE_adhoc.pdf
            2. http://www.ieee802.org/3/by/public/adhoc/architecture/wertheim_050615_25GE_adhoc.pdf
        3. No change to CL49 LPI FSMs.

- **CONCLUSION: RCWMs as good as or better than CL74 approach.**

CONCERNs

WHAT PROBLEM ARE WE SOLVING WITH THE PROPOSED CHANGE?

- *WRT "Re-use CL74 PN-2112 scrambler (run over 5280b instead of 2112b)"*
    1. There is no savings here.
    2. The scrambler exists for CL74 anyway.
    3. No demonstrable advantage for the SnT proposal.

- **CONCLUSION: No savings, and therefore no advantage.**

ADDITIONALLY EXPRESSED CONCERNs

- *WRT FlexE – 25G RS FEC Idle insertion/deletion may be inconsistent with FlexE*
  1. FlexE layers on top of IEEE Ethernet PHY types, and must accommodate those definitions.
     1. FlexE requires an Ethernet PHY to **disable** the 64/66b codecs. Therefore, **IEEE 25GE PHYs** are not compatible with FlexE.
        1. **All IEEE BASE-R type PHYs** are not compatible with FlexE.
     2. There may be other incompatibilities.
  2. FlexE proposal is embryonic, and therefore subject to definition.
  3. What is the relevance and the BMP for FlexE in 802.3by?
  4. FlexE considerations are out of scope for 802.3by.

- **CONCLUSION: No value in making a change either for FlexE.**

ADDITIONALLY EXPRESSED CONCERNs

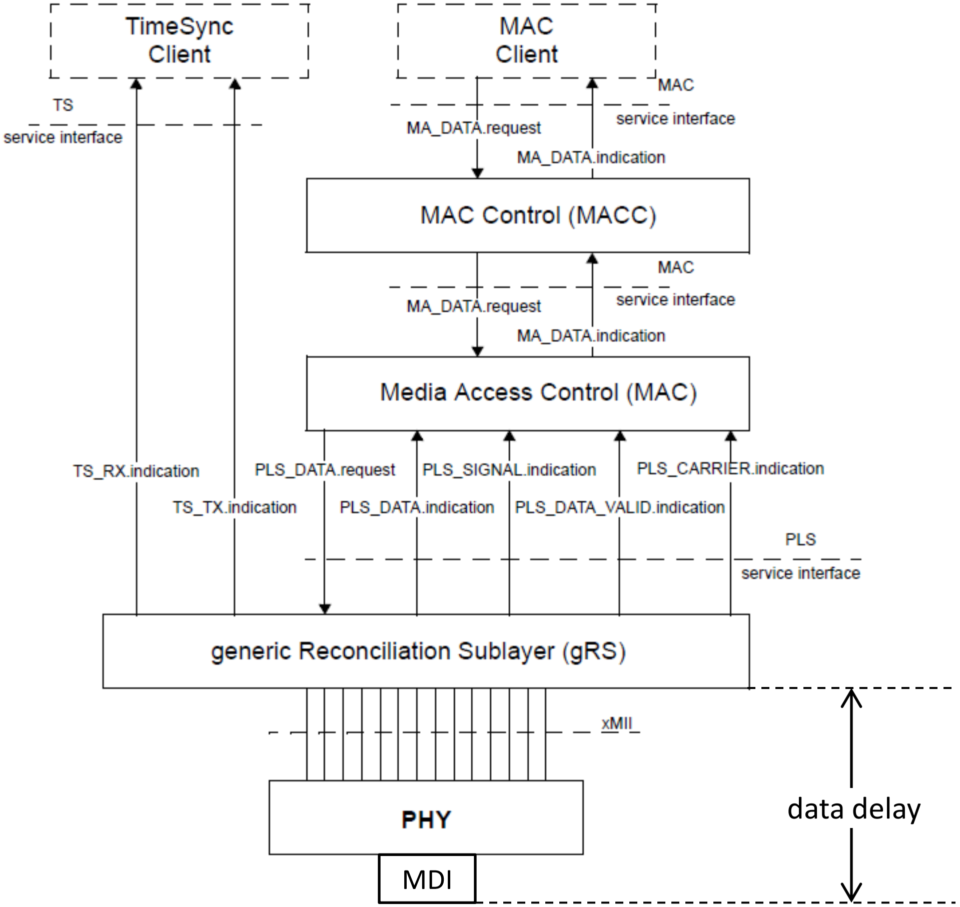
- *WRT OTN – OTN Transparency Perception*
  1. There is no known issue for OTN.
     1. Per trowbridge_3by_01_0115.pdf presentation, the FEC does not protect the OTN network.
     2. The effect of CWMs is no different from what happens in an MLD PCS.
     3. The CL108 FEC is bandwidth neutral.

- **CONCLUSION: No value in making a change either for OTN.**

ADDITIONALLY EXPRESSED CONCERNs

- *WRT 1588 Accuracy.*
  1. Integration of PCS and FEC allows for maximum 1588 accuracy.
  2. Separate PCS and FEC devices already handled via 802.3bf.
     1. 802.3bf provides a mechanism for a PHY to communicate its **range** of **delays**.
     2. See next slide.
  3. 1588 functions may be implemented in a PHY for max accuracy.
     1. Note: PHYs have variable delays.
     2. Allows for differentiation within the Market.
     3. Requires functions (PCS, RS) already provided by 802.3by
- **CONCLUSION: No value in making a change for 1588.**

IEEE Std 802.3bf support for time synchronization protocols (CL 90)



- TS_RX.indication and TS_TX.indication are generated in response to detection of a valid SFD at the xMII
- Data delay is represented by the quartet of values:
  - maximum transmit data delay
  - minimum transmit data delay
  - maximum receive data delay
  - minimum receive data delay
- Data delay values must be reported for each MMD

Motivation For CWMs in 25G RS-FEC Mode

- Initial motivation for including the CWMs for 25GE was consistency with 100G RS-FEC code-word locking implementation.
  - Allows sharing and/or re-use of RS FEC designs.
- 40GE AMs were chosen for 25GE single lane, to differentiate from 100G AMs at same lane rate.

SUMMARY

- ***Make all Ethernet PHYs supporting RS FEC consistent and maintain CWMs in 25G.***
  - **There is no *compelling* reason to change.**
  - **The CWMs have distinct advantages as outlined in this presentation.**
    - **Such as RCWMs for quick CW locking in EEE mode.**
  - **We should focus on making progress and adhere to the approved timeline, and maintain the CWMs.**

THANK YOU!

BACKUP

For 100G Ethernet:
The logic finds the first AM0, then the next AM0, and if they're good (9 or more good nibbles, allowing for errors in the CW), it sets amps_locked. That means the FEC CW boundaries have been recognized and locked.

Once all lanes are locked, all_locked is set. Deskew then happens, and then complete CWs start to be checked.
It's really once deskew is done that the fec_align_status is set, which is basically the FEC is locked bit.
The CWs don't really start to be checked until fec_align_status is true.

For FC, follow the following algorithm:

```
START: descramble the data, try the current bit, and decode it.
If ( S!=0) {  //  means no errors in the CW
    Change the bit pointer;
    GOTO START;
}
Else {
    Good = 1;
}
CONT: descramble the next CW at the same bit position, and decode it
If ( not correctable ) {
    Change the bit pointer;
    GOTO START;
}
Else {
    Good++;
    Bad =0;
    Codeword_sync = 1;
}
NOW IT TAKES THREE CONSECUTIVE CWs TO LOSE codeword_sync;
```