IEEE 802.3 25G Ethernet SG – Arch Ad Hoc Layering and Gaps

Eric Baden (ericb at broadcom com), presenting
Cedrik Begin (cbegin at cisco com),
Yong Kim (ybkim at broadcom com),
Gary Nicholl (gnicholl at cisco com)

Supporters

Brad Booth, Microsoft
John D’Ambrosia, Dell
Joel Goergen, Cisco
Peter Jones, Cisco
Rob Stone, Broadcom
Mike Dudek, Qlogic
Kapil Shrikhande, Dell
Mike Andrewartha, Microsoft
Ali Ghiasi, Ghiasi Quantum LLC
JC Castillo, CoMira Solutions
Don Cober, CoMira Solutions
Oded Wertheim, Mellanox
Scott Kipp, Brocade
Tom Ambrose, Emulex
Ryan Latchman, Macom

<name>, Affiliation

25G PCS Thoughts - recap

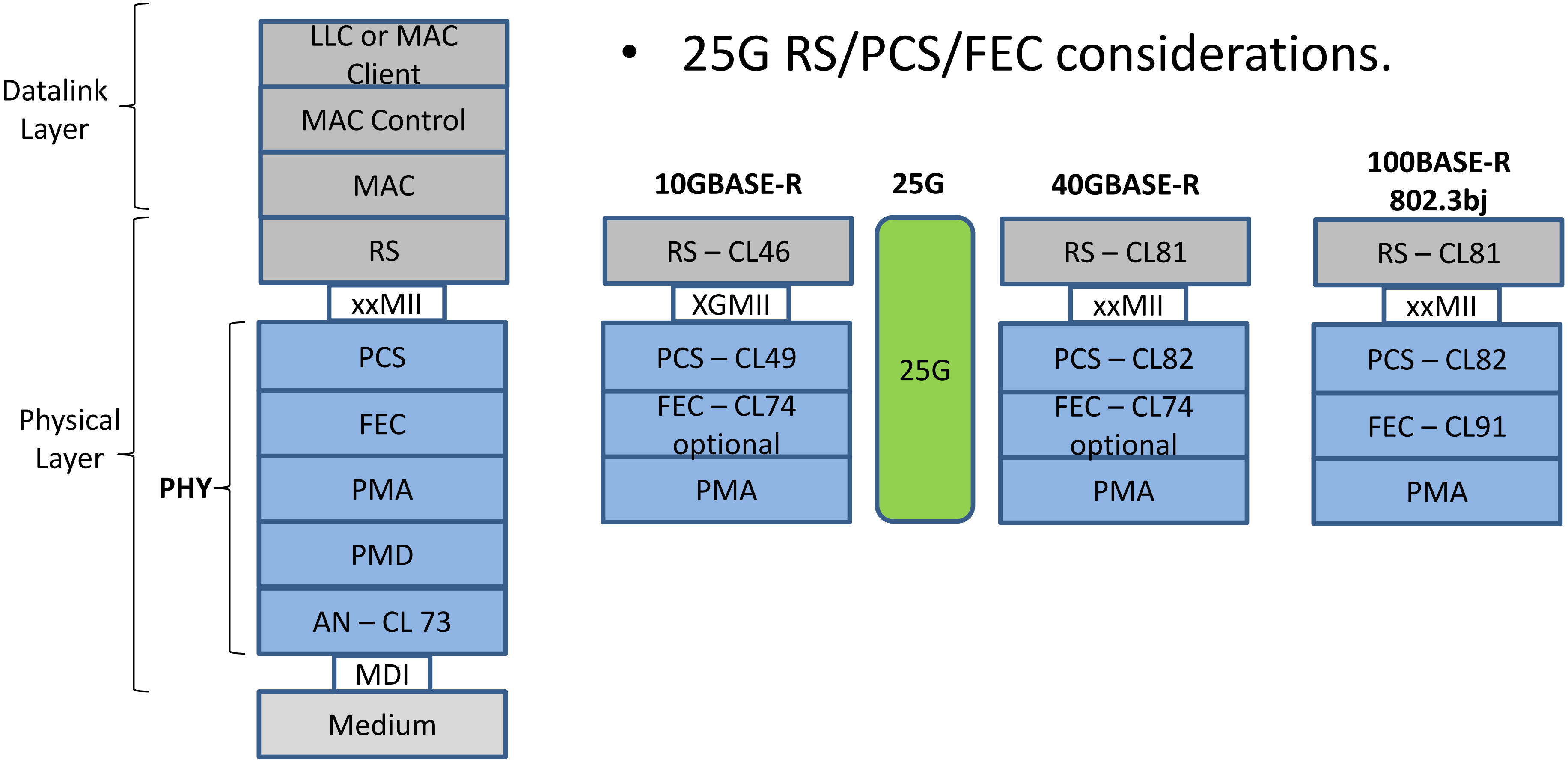
- Recap from Sept Interim (not to revisit)
  - Both 3m and 5m reach adopted as objectives (implicit ToR and InterR)
  - FEC/no FEC (implicit sub-set objectives of latency, cost, compatibilities
- Views
  - 10G speed up
  - 100G (.3 bj) quarter lane use
- Desires
  - NICs – implementations for 10G/25G and 40G
  - Switches – implementations for 100G/40G/25G and 10G

General and Common Ideas - Recap

- 64/66B.
- Lane rate of 25.78125G
- Alignment Marker eases the use of FEC (not FEC capability).
  - BIP has benefits. Bug-fix category or nice to have?
- Optional Auto-negotiation determines use of FEC and training, among other things.

[Sub-]Layering

- 25G RS/PCS/FEC considerations.

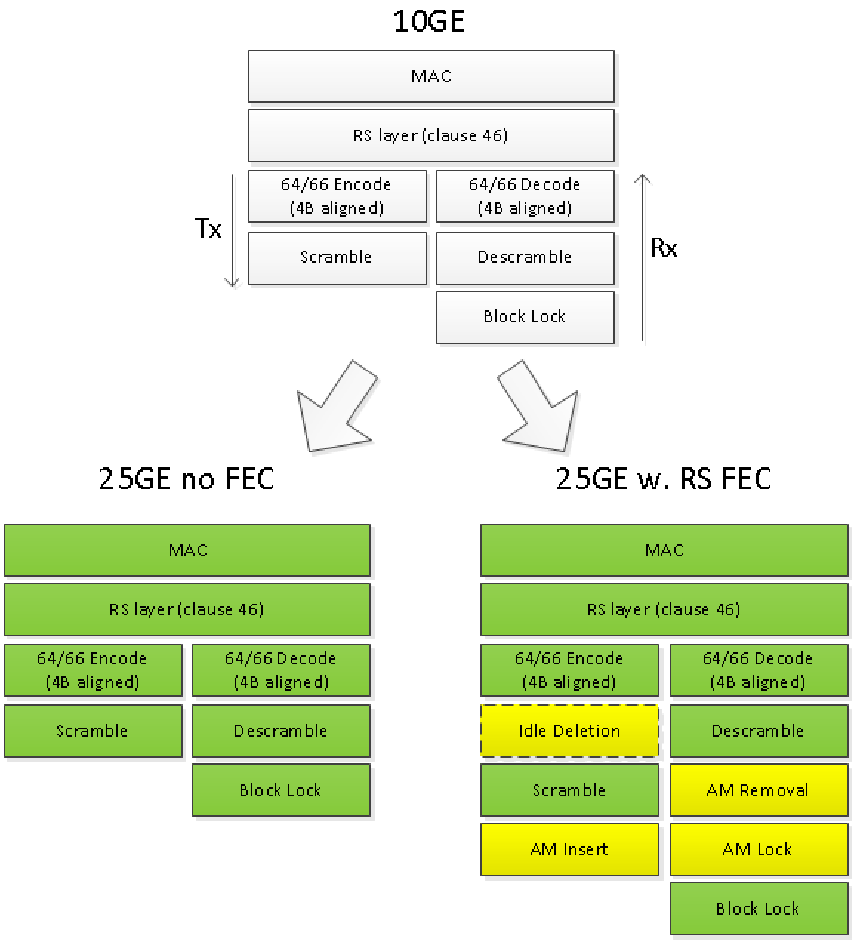


[Sub-]Layer Elements

- Closer look at the data path elements of 10GBASE-R, 40G/100G BASE-R, and recent .3bj work.
  - Examine RS/PCS/FEC datapath elements adopted for 25G Ethernet use, individual clause basis and also together.
  - Evaluate the choices for relevancy, technical merits, and ease of implementation.

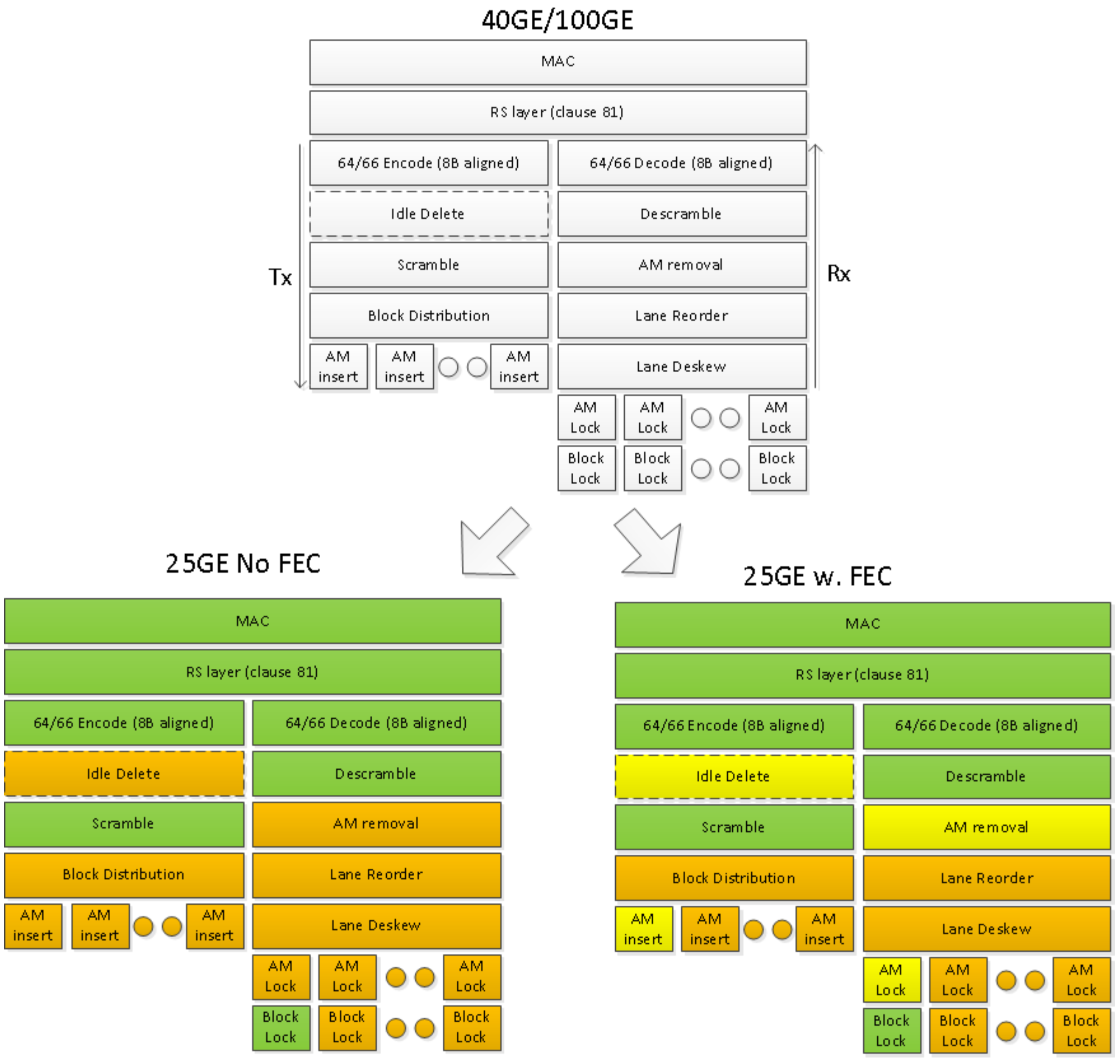
Details of 25G **Sub**-Sub-Layering considerations

25GE PCS using 10GE (CL49) building blocks



- 4 byte MII (CL46)
- For a 25GE without RS FEC, can use 10GE function as is, i.e. complete reuse (simply run 2.5x faster).
- To aid RS FEC, would add alignment marker insertion and removal in the 25GE PCS. (yellow blocks)

25GE PCS using 40/100GE (CL82) building blocks



40GE/100GE

MAC

RS layer (clause 81)

Tx

64/66 Encode (8B aligned)

Idle Delete

Scramble

Block Distribution

AM insert

AM insert

AM insert

Rx

64/66 Decode (8B aligned)

Descramble

AM removal

Lane Reorder

Lane Deskew

AM Lock

AM Lock

AM Lock

Block Lock

Block Lock

Block Lock

25GE No FEC

MAC

RS layer (clause 81)

64/66 Encode (8B aligned)

Idle Delete

Scramble

Block Distribution

AM insert

AM insert

AM insert

64/66 Decode (8B aligned)

Descramble

AM removal

Lane Reorder

Lane Deskew

AM Lock

AM Lock

AM Lock

Block Lock

Block Lock

Block Lock

25GE w. FEC

MAC

RS layer (clause 81)

64/66 Encode (8B aligned)

Idle Delete

Scramble

Block Distribution

AM insert

AM insert

AM insert

64/66 Decode (8B aligned)

Descramble

AM removal

Lane Reorder

Lane Deskew

AM Lock

AM Lock

AM Lock

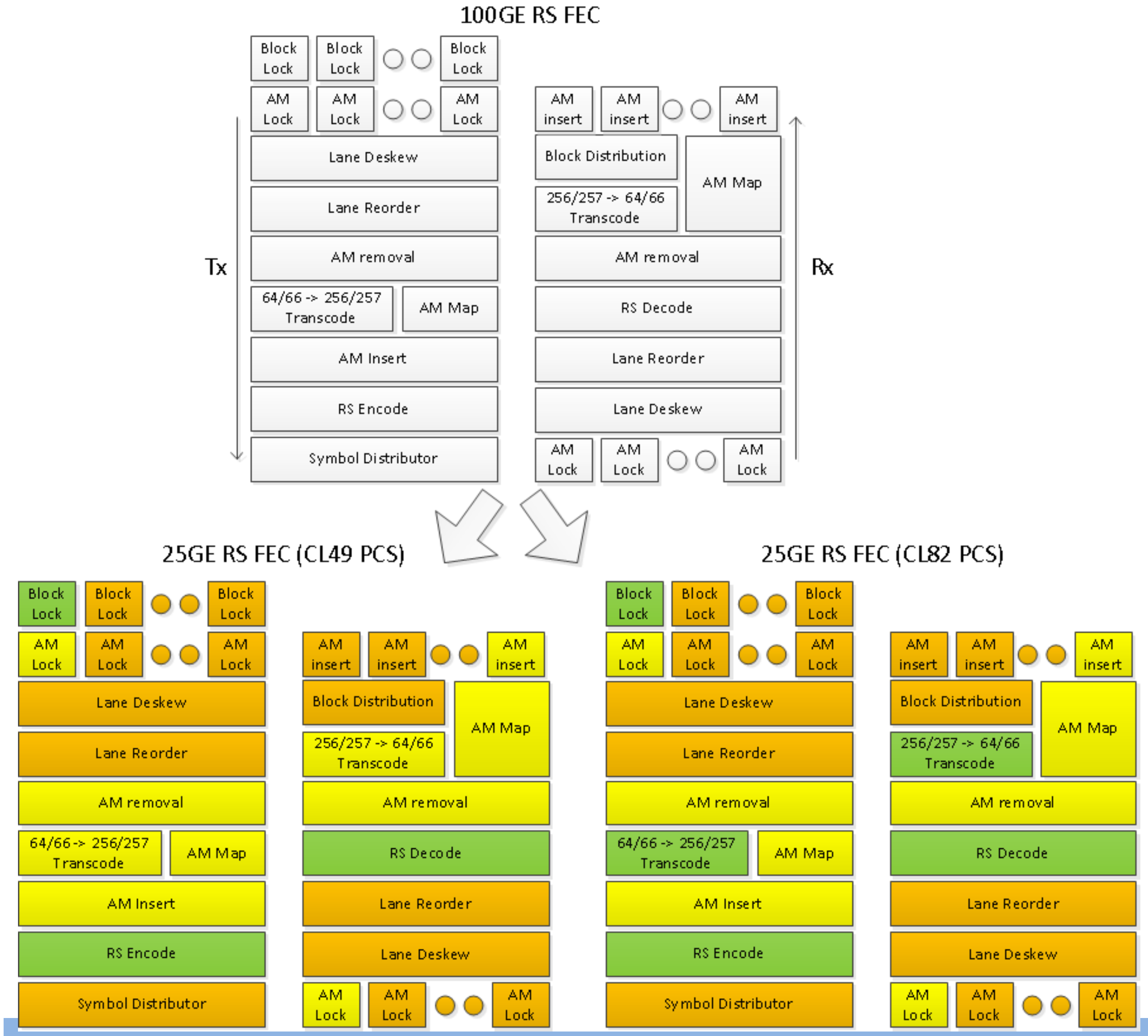
Block Lock

Block Lock

Block Lock

- 8 byte MII (CL81).
- Some function reuse, however would remove (orange blocks):
  - multiple per lane logic
  - block distribution and reorder/deskew.
- AM insertion/removal logic would need to change (yellow blocks) in order to reflect different rates of AM insertion/removal

Changes to RS FEC (CL91) for 25GE (8B vs. 4B)



100GE RS FEC

Tx

Block Lock | Block Lock | Block Lock

AM Lock | AM Lock | AM Lock

Lane Deskew

Lane Reorder

AM removal

64/66 -> 256/257 Transcode | AM Map

AM Insert

RS Encode

Symbol Distributor

Rx

AM insert | AM insert | AM insert

Block Distribution | AM Map

256/257 -> 64/66 Transcode

AM removal

RS Decode

Lane Reorder

Lane Deskew

AM Lock | AM Lock | AM Lock

25GE RS FEC (CL49 PCS)

Block Lock | Block Lock | Block Lock

AM Lock | AM Lock | AM Lock

Lane Deskew

Lane Reorder

AM removal

64/66-> 256/257 Transcode | AM Map

AM Insert

RS Encode

Symbol Distributor

AM insert | AM insert | AM insert

Block Distribution | AM Map

256/257 -> 64/66 Transcode

AM removal

RS Decode

Lane Reorder

Lane Deskew

AM Lock | AM Lock | AM Lock

25GE RS FEC (CL82 PCS)

Block Lock | Block Lock | Block Lock

AM Lock | AM Lock | AM Lock

Lane Deskew

Lane Reorder

AM removal

64/66 -> 256/257 Transcode | AM Map

AM Insert

RS Encode

Symbol Distributor

AM insert | AM insert | AM insert

Block Distribution | AM Map

256/257 -> 64/66 Transcode

AM removal

RS Decode

Lane Reorder

Lane Deskew

AM Lock | AM Lock | AM Lock

- For both options would remove (orange):
  - Per lane logic
  - Block distribution and deskew logic.
- For both options would need to change AM related logic to reflect difference in number of AMs and periodicity (yellow).
- Only difference between the two options is that the clause 49 based option would need the transcoders to not restrict the transcoding of its additional block codes.

Summary

- Clause 49 is the better starting point for a 25GE PCS.
  - Even in the case where an alignment marker is inserted to aid the RS FEC
- Changes are required to clause 91 FEC, whether or not the 25GE PCS is based on clause 49 or clause 82
  - Magnitude of changes are equivalent.

25G directions with optional FEC

PCS/FEC	10G	25G without any FEC	25G with CL74 FEC	25G with CL91 RS FEC	40G	100G
Block Coding	64/66B					
Lanes	1	1	1	1	4	4
RS	CL46 (4B)	CL46 (4B)	CL46 (4B)	CL46 (4B)	XLGMII (8B)	CGMII (8B)
PCS	CL49	CL49	CL49	CL49	CL82	CL82
Alignment Markers	N	N	N	Y	Y	Y
Trans Code	N/A	N/A	N/A	256/257B	N/A	256/257B
Reach		3+ m		5+ m		
Latency		Low	Medium	High		

ALIGNMENT MARKERS (AMS) - REVIEW

- Used by MLD PCS to De-skew across lanes
  - Inserted into data stream in groups, based on the number of PCS lanes.
  - IDLEs are deleted to offset bandwidth increase.
- One AM per PCS Lane
  - Four PCS lanes in 40G. Twenty PCS lanes in100G.
  - AMs in 40G are different from AMs in 100G.
- DC Balanced (same number of 1's as 0's)
  - 'Many' transitions for CDR maintenance.
- Spaced 16383 * Number of PCS lanes apart.
  - The 'space' is the number of 66 bit blocks between the end of one group of AMs and the beginning of the next group of AMs.
  - 40G **PCS**: AMs are inserted every 16383Blocks*66bits/Block***4**PCS Lanes/(4*10.3125G) =~ 105us
  - 100G **PCS**: AMs are inserted every 16383Blocks*66bits/Block***20**PCS Lanes/(10*10.3125G) =~ 210us
- Used with **CL91** FEC to determine Code Word (CW) boundaries
  - A CW is 5280 bits. Equivalent to 80 – 66 bit blocks.
  - 100G: 16384*66*20/5280 = 4096.
    - For 100G with CL91, AMs appear every 4096 CWs
- BIPs provide some link quality checking on per PCS Lane basis.
  - Parity doesn't always work in the presence of multiple bit errors.

ALIGNMENT MARKERS (AMS) – 25G PROPOSAL

- **Only when CL91 is enabled, periodically insert 4 AMs**
  - **AMs are Required** for use with CL91 FEC to determine Code Word (CW) boundaries
  - **Four consecutive AMs are Required** for use with CL91 FEC transcoding
  - *Simplifies implementations not requiring CL91 FEC*
  - Delete IDLEs to offset bandwidth increase
- **Space AMs to match 100G spacing, and meet CL91 needs**
  - 25G: 16384***5***66/5280 = 1024.
    - AMs appear every 1024 CWs
    - 16384Blocks*66bits/Block*5/(2.5*10.3125G) =~ 210us
- **Re-use** AM0, AM1, AM2, AM3 from **40G** CL82 PCS
  - *Known, simple, good properties (see previous slide)*
  - *Different from **100G** AMs (avoids any ambiguity)*
- BIPs not needed with CL91
  - Replace with fixed values?

THANK YOU!