IEEE P802.3by

Understanding the Tradeoffs for 25GBASE-CR

Dan Dove, DNS for Emulex

March Plenary Meeting, Berlin 2015

March 9, 2015

Supporters

- □ Venugopal Balasubramonian – Marvell
- □ Paul Kolesar – Commscope
- □ George Zimmerman – CME Consulting
- □ Matt Brown – Applied Micro
- □ Jonathan King - Finisar
- □ William Lo – Marvell
- □ Ali Ghiasi – Ghiasi Quantum LLC
- □ Jeff Slavick – Avago Technologies
- ❖ Kent Lusted – Intel
- ❖ Rick Rabinovich – Alcatel Lucent
- ❖ Jacky Chang – HP

❖ added after 03/08/2015 upload

Outline

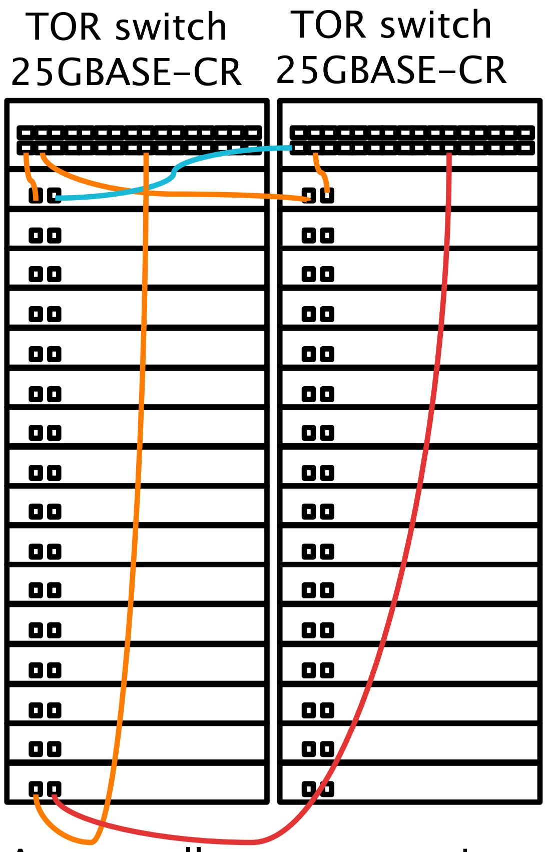
- □ Introduction & Problem Statement
- □ Decision Making Process
- □ Data for Analysis
- □ Conclusion(s)

Introduction

- □ IEEE P802.3by has two reach objectives on twin-ax cables (3m and 5m).
- □ To achieve 3m, there is ongoing debate about whether FEC is mandatory, but data suggests that for worst case 3m cables, BASE-R FEC will be required.
- □ For 5m applications, a higher gain FEC (RS-FEC) is required.
- □ Because the RS-FEC adds latency and die area, there is a desire to use the lower gain BASE-R FEC for shorter reach applications and only require RS-FEC for longer reach applications.
- □ Some end users may wish to avoid FEC altogether, even if engineered links are required, to minimize latency. If each approach were treated as a unique PMD, this would leave three different PMDs. (No FEC: CR-N, BASE-R FEC: CR-S, RS-FEC: CR-L)
- □ There is a question about how the market will respond to having multiple PMD instances, and how to auto-negotiate when priority may be different for low-latency applications, than for longer-reach applications.
- □ The key question remains "Do we have a single "One Size Fits All (OSFA) PMD, or do we have separate PMDs to achieve broad market potential?"

Problem Statement (Traditional Ethernet Configuration)

Traditional Ethernet Configuration
Plug-n-Play Ready
Redundant Switches



Assume all copper ports in servers are 25GBASE-CR as well. It all just works.

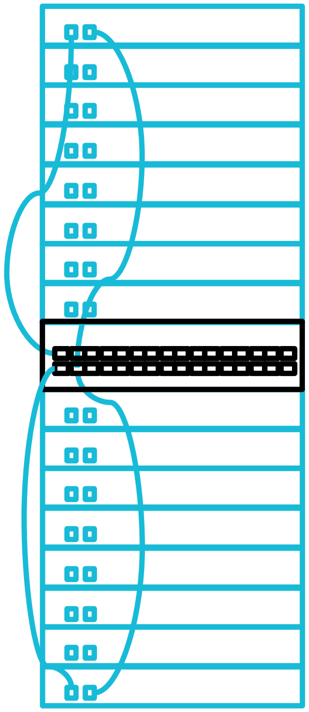
- Cables may be of normal grade, high grade, or ultra high grade to define performance.
- Regardless of performance grade, all ports link up.
- Ports negotiate to defined performance level combined with cable performance capability.
- Ex: Blue cables = ultra-high grade – May be more costly, but meet “no FEC” requirements.
- Ex: Gold cables = high-grade. May be up to 3m in reach, or shorter with lower cost materials.
- Ex: Red cables = normal-grade. May be up to 5m in reach, or shorter with even lower cost materials.

- In this example, customer uses different grade cables, receives different grade of performance but all ports link and perform at 25Gbps.
- Ethernet customers are used to this model and expect it!

Problem Statement (Proposed Engineered Configuration)

Engineered
Configuration

MOR switch
25GBASE-CR



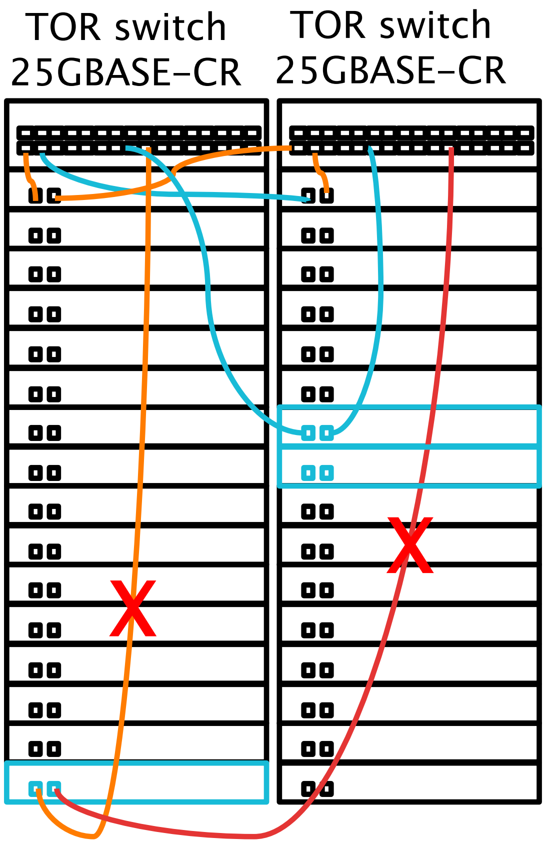
Assume all copper ports in
servers are 25GBASE-CR-S.
Configuration is constrained.

- •Cables MUST be of ultra high grade to define performance.
- •Ports negotiate to defined performance level combined with cable performance capability.
- •Ex: Blue cables = ultra-high grade – May be more costly, but meet "no FEC" requirements.
- •Ex: Server/NIC ports defined to optimum performance potentially using more expensive materials (blue boxes customized)

- •In this example, customer has engineered their physical structure, purchased custom server NICs, configured S/W to ensure FEC is disabled, and will not achieve link unless all elements are within very tight electrical specifications.

Problem Statement (Lack of interoperability)

Heterogeneous Ethernet Configuration
NOT Plug-n-Play Ready



Some copper ports are- L and some are -S. Customer expects plug and play!

•Cables may be of normal grade, high grade, or ultra high grade to define performance.

•Ports negotiate to defined performance level combined with cable performance capability.

•Ex: Blue cables = ultra-high grade – May be more costly, but meet “no FEC” requirements.

•Ex: Gold cables = high-grade. May be up to 3m in reach, or shorter with lower cost materials.

•Ex: Red cables = normal-grade. May be up to 5m in reach, or shorter with even lower cost materials.

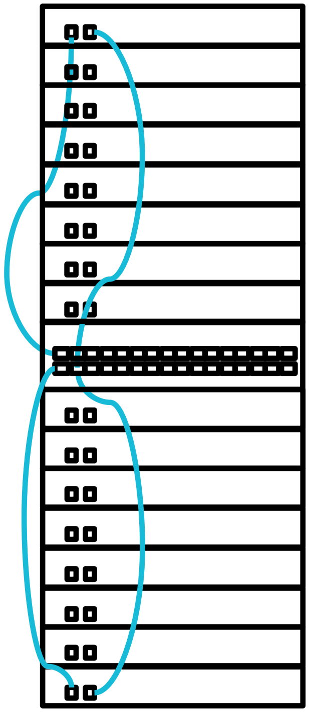
•In this example, customer assumes plug-n-play. Purchasing dept orders “25GBASE-CR” just like they ordered “10G” and expects it all to just work together. Doesn’t realize “-S” and “-L” impact interoperability because mixing works “sometimes”.

•In some cases, it works with slightly higher BER than expected, causing support calls.

Problem Statement (A Better Engineered Configuration)

BETTER
Engineered
Configuration

MOR switch
25GBASE-CR



Assume all copper ports in servers are 25GBASE-CR.

•Cables MUST be of ultra high grade to define performance.
•Ports negotiate to defined performance level combined with cable performance capability.
•Ex: Blue cables = ultra-high grade – May be more costly, but meet “no FEC” requirements.
•Ex: Server/NIC ports defined to optimum performance potentially using more expensive materials, customer defines port specs which are STILL compliant, just better than minimum spec.

•In this example, customer has engineered their physical structure, purchased custom server NICs, configured S/W to ensure FEC is disabled, and will not achieve link unless all elements are within very tight electrical specifications.

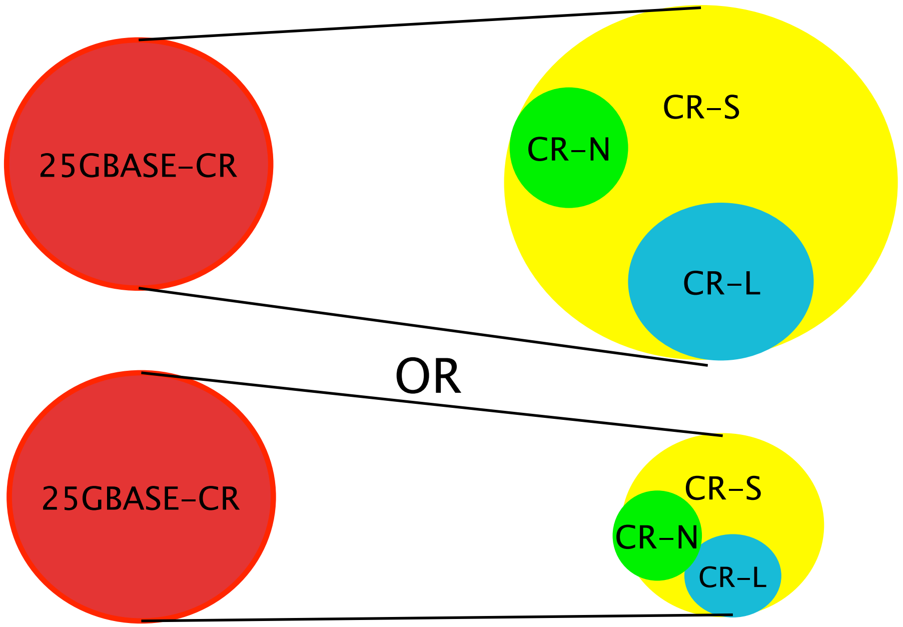
•But they do it with a SINGLE IEEE specification for 25G and don’t create confusion in the market.
•The only additional cost is in the cables which are designed to ultra-high grade.

Decision Making Process

- □ The author believes that a key benefit of Ethernet is providing plug-and-play interoperability to end users.
- □ There is concern that multiple PMDs that will not interoperate would potentially lead to customer confusion and ultimately reduce broad market potential.
- □ The key difference between PMDs under discussion is not in the area of Analog Front End (AFE) or PCS, but which form of FEC to use.
- □ We understand that different applications may be sensitive to different key parameters, primarily cost, power, die-area, so we seek to understand if there is sufficient differential cost/power/die-area to justify potential confusion of having multiple, non-interoperable copper PMDs at 25G.
- □ If no significant cost/power/die-area difference exists, we believe a single PMD with Auto-Negotiation to the alternative FEC option is best for the industry.
  - □ One PMD with Auto-Negotiation capable of disabling FEC, or setting FEC to the appropriate mode of operation to ensure link comes up.

Decision Making Process

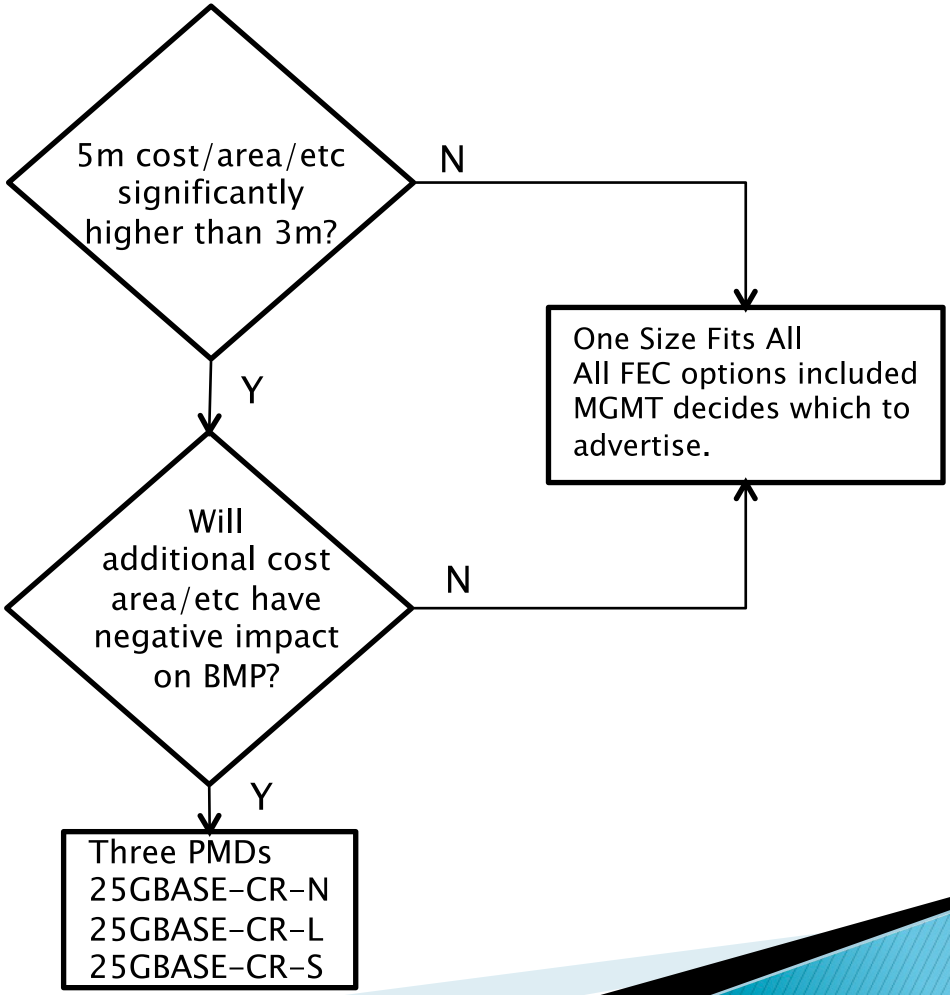
- Key Question
  - Will a multiple PMD solution increase or decrease Broad Market Potential?



  - The author believes that if only a small differential cost is possible, will not lead to increase in market volume, but interoperability issues will lead to reduction in BMP

Decision Making Process

- The author believes that the following process should apply to making the decision on one PMD for all reach objectives, vs multiple PMDs.


5m cost/area/etc significantly higher than 3m?
N
Y
One Size Fits All
All FEC options included
MGMT decides which to advertise.
Will additional cost area/etc have negative impact on BMP?
N
Y
Three PMDs
25GBASE-CR-N
25GBASE-CR-L
25GBASE-CR-S

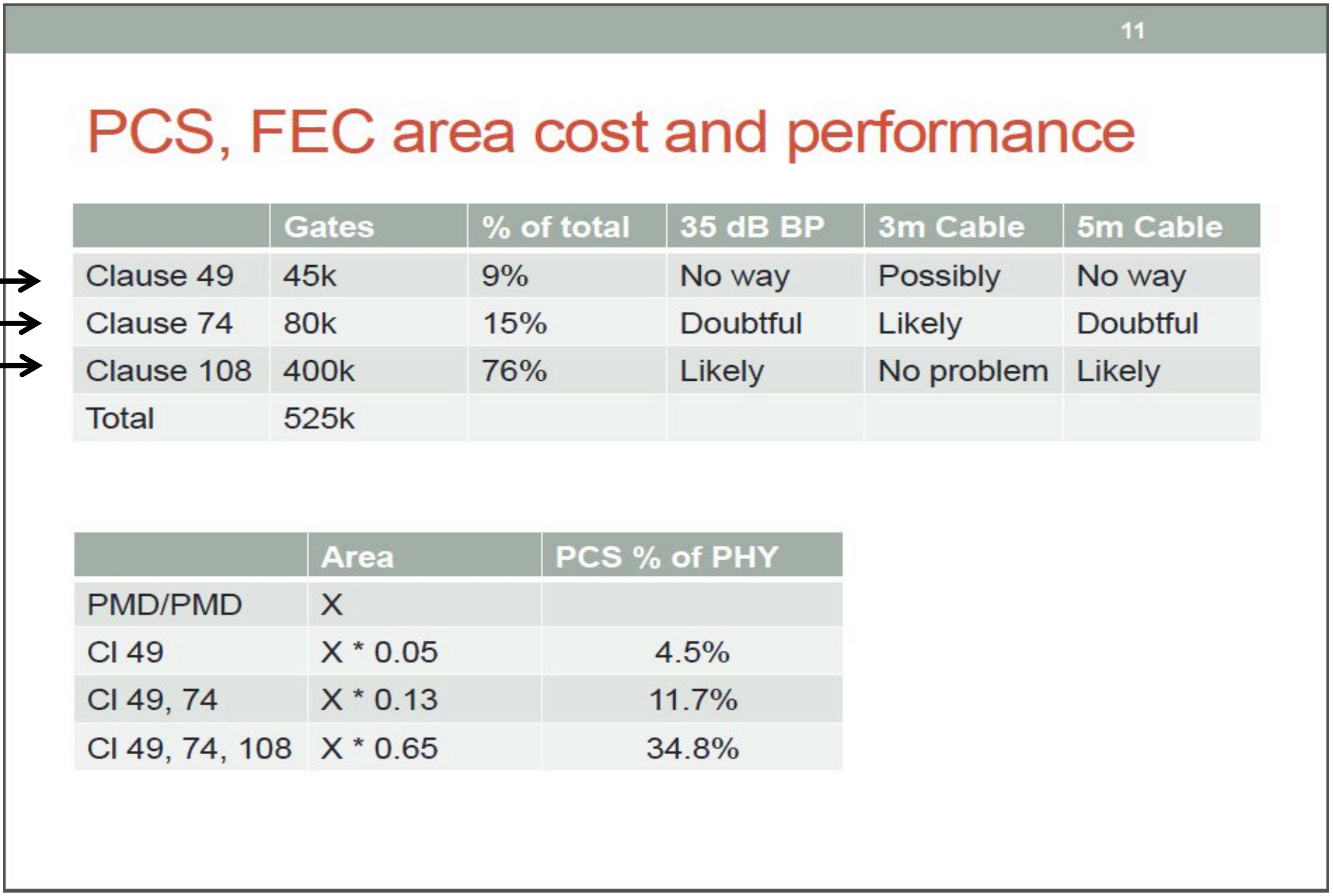
Data For Analysis

Clause Decoder

"No FEC"

"BASE-R FEC"

"RS-FEC"



PCS, FEC area cost and performance

	Gates	% of total	35 dB BP	3m Cable	5m Cable
Clause 49	45k	9%	No way	Possibly	No way
Clause 74	80k	15%	Doubtful	Likely	Doubtful
Clause 108	400k	76%	Likely	No problem	Likely
Total	525k				

	Area	PCS % of PHY
PMD/PMD	X	
Cl 49	X * 0.05	4.5%
Cl 49, 74	X * 0.13	11.7%
Cl 49, 74, 108	X * 0.65	34.8%

ref: slavick_022515_25GE_adhoc

Note: Area of die related to geometry of IC fab process. As geometry reduces, digital areas will tend to reduce much faster than required analog portions making inclusion of FEC even less of an impact in area, cost, power.

Data For Analysis

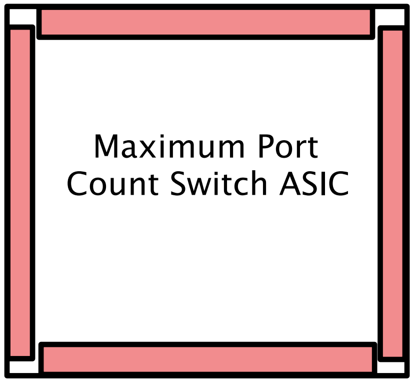
- **Die Area/Gate Count**
- A OSFA PHY requires ~ 525K gates to implement all FEC options per slavick_022515_25GE_adhoc.
- Die area potentially impacts cost but its not a direct cost factor.
  - If the design is ball-limited (ie: the size of the package and number of balls is greater than the die required for the job, die area itself is not a factor.
- RS-FEC requires the greatest amount of die-area as identified by # of gates.
  - According to slavick_022515, ~400K gates per instance of RS-FEC.
  - For 100G ports that may be used to create break-out to 25G, the RS-FEC option will most likely already be included in the port design. Thus, no incremental gate count.
  - For 25G ports that support MMF fiber, RS-FEC will be required, thus no incremental gate count.

Data For Analysis

- □ Arguably, 525K gates, an increment of ~22% of the PCS is significant, however, for most system ASIC applications, this becomes a relatively small contribution to the design.
  - □ A “State of the Art” NIC design > 35,000,000 gates and thus incremental gate count translates to ~3% of design.
  - □ A “State of the Art” switch design > has even more gates and thus incremental gate count translates to <1% of design.
    - □ Example: Broadcom announced a 128 port 25G switch chip with 7 Billion transistors. Assuming 4 transistors/gate, almost 2 BILLION gates.
    - □ 128 ports * 400K gates for RS-FEC yields < 3% impact (assuming no sharing).
    - □ http://www.enterprisetech.com/2014/09/24/broadcom-fights-ethernet-rivals-tomahawk-chips/
  - □ Its been observed that key portions of an RS-FEC implementation may be shared in a multi-port implementation. (ex: AM search)

Data For Analysis

- Some might argue that for very specific ASIC configurations, the 1-3% die area may lead to much larger IC areas.
- Ex: High port count switch ASIC bound by I/O
  - As each port grows, the overall size of the I/O ring grows and potentially leads to a larger IC



- Is this a “general case”, or a “very specific case”?
  - It assumes the I/O is currently the bounding limitation
  - It assumes additional logic must reside within the I/O ring
  - It ignores future geometry shrinks will generally reduce the problem
  - It ignores that for general purpose switching, 100G ports and 25G MMF ports will require RS-FEC anyway
- Do we create customer confusion in the market, and potentially fragment the market, to solve a “very specific case”?

Conclusion(s)

□ The perceived benefit of having multiple copper PMDs that target specific applications must be balanced against the cost of confusion in the market place.

□ Confusion, especially in a new market, can have a negative effect on market adoption, delay implementation and ultimately undermine long-term volumes.

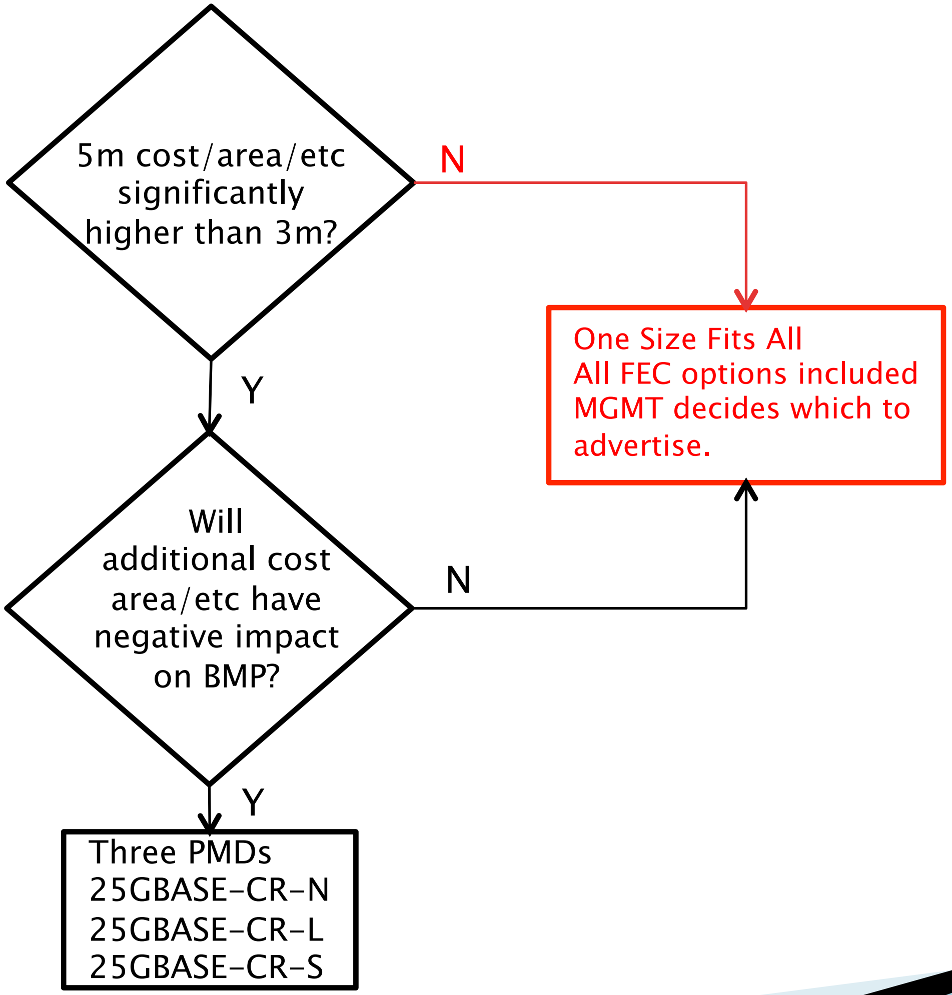
□ Confusion among the supplier base as to which PMD to design, which one will receive customer acceptance, and potential duplication of products can lead to inefficiencies which actually add cost.

□ The differential cost/area impact of having FEC in the PHY does not have a significant effect on key applications (Servers/Switches) **

** Generally true, may not apply for all specific implementations

Conclusion(s)

- The author believes the data, if corroborated by multiple suppliers, directs us to the following conclusion.



5m cost/area/etc significantly higher than 3m?

One Size Fits All
All FEC options included
MGMT decides which to advertise.

Will additional cost area/etc have negative impact on BMP?

Three PMDs
25GBASE-CR-N
25GBASE-CR-L
25GBASE-CR-S

Proposal

- Specify a single 25GBASE-CR PMD with mandatory implementation of RS-FEC, BASE-R FEC, and no-FEC. The FEC mode may be manually configured by the user or determined through auto-negotiation by advertising acceptable modes and resolving highest common capability.
- This may be the basis for a motion, or a straw poll position.

Q&A