# Thoughts on the BER Objective

**IEEE 802.3 Beyond 400 Gb/s Ethernet Study Group**

Mark Gustlin – Cisco

# Supporters

Cedric Lam - Google

Rob Stone – Facebook

# Past History

- My recollection of the BER debate for 802.3bs (400/200GE) was:
  - Some system vendors in the past were held to: any bit errors were a bad thing in the system
  - They were able to do that since we had margin in our interfaces and systems (things were easier at slower lane rates)
  - This group wanted a better BER (1E-15 or better)

  - Many component focused participants liked the 1E-12 for: shorter test time, better yield etc.
  - A compromise was stuck at 1E-13, with some justification that it kept the errors/sec similar to the past (thanks for Pete Anslow's presentation)

  - This moved the bar somewhat without a radical departure from the past

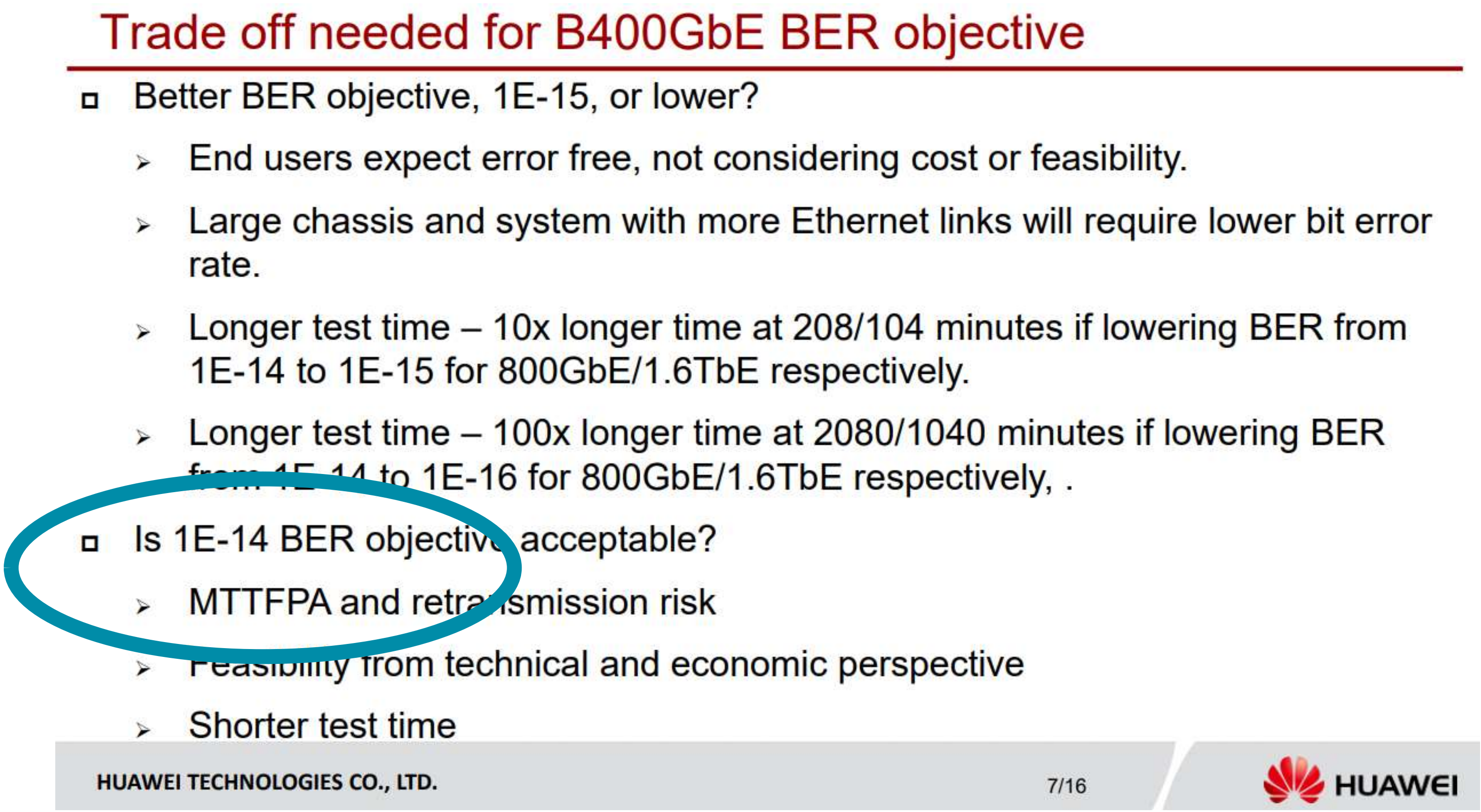
# Considerations for Beyond 400GE BER?

- We should consider:
  - Mean Time To False Packet Acceptance (MTTFPA)
  - Application needs
  - Cost in terms of power and gates for implementations

- We should not consider:
  - Errors/second (not too important if the application needs are met)

# Other Work Referenced:

- This works cites MTTFPA as one criteria for deciding on the BER objective (or the equivalent FLR)
- If we have a strong FEC, such as RS(544,514), I don't think that is a concern?



## Trade off needed for B400GbE BER objective

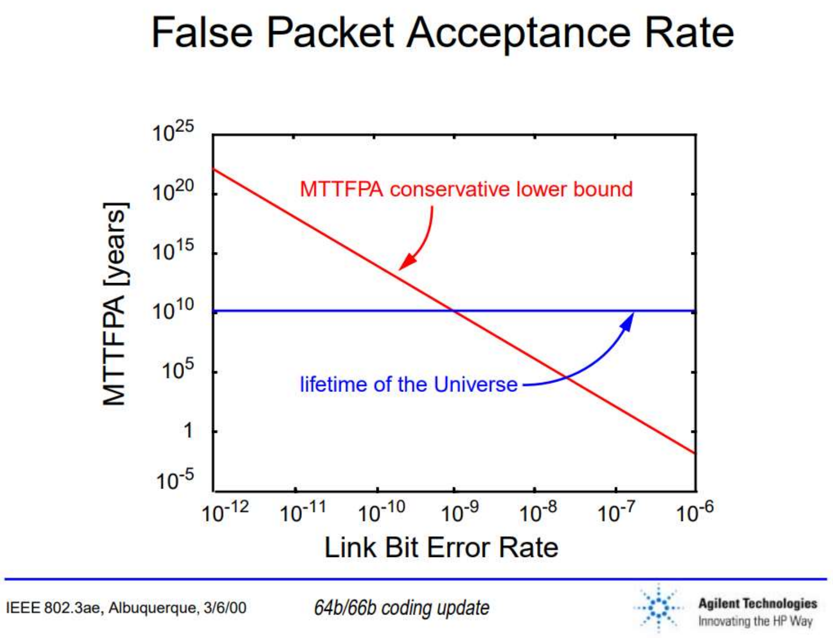
- Better BER objective, 1E-15, or lower?
  - End users expect error free, not considering cost or feasibility.
  - Large chassis and system with more Ethernet links will require lower bit error rate.
  - Longer test time – 10x longer time at 208/104 minutes if lowering BER from 1E-14 to 1E-15 for 800GbE/1.6TbE respectively.
  - Longer test time – 100x longer time at 2080/1040 minutes if lowering BER from 1E-14 to 1E-16 for 800GbE/1.6TbE respectively, .
- Is 1E-14 BER objective acceptable?
  - MTTFPA and retransmission risk
  - Feasibility from technical and economic perspective
  - Shorter test time

HUAWEI TECHNOLOGIES CO., LTD. 7/16 HUAWEI

From: wang_b400g_01_210315.pdf

# MTTFPA Importance

- The MTTFPA has been a metric used in Ethernet since at least 10GbE
  - Probably earlier, but I have not researched earlier
- It tells us how often the Ethernet link at the minimum BER is likely to pass a packet without flagging it as a bad packet
  - In other words, how often is a corrupted packet silently passed
- Calculating MTTPA considers the following:
  - Error rates and error models
  - Detection provided by the encoding (64B/66B etc.)
  - Detection provided by any FEC
  - Detection provided by the CRC32
- Rick Walker and company set the bar at:
  - The lifetime/age of the universe in the 10GE days



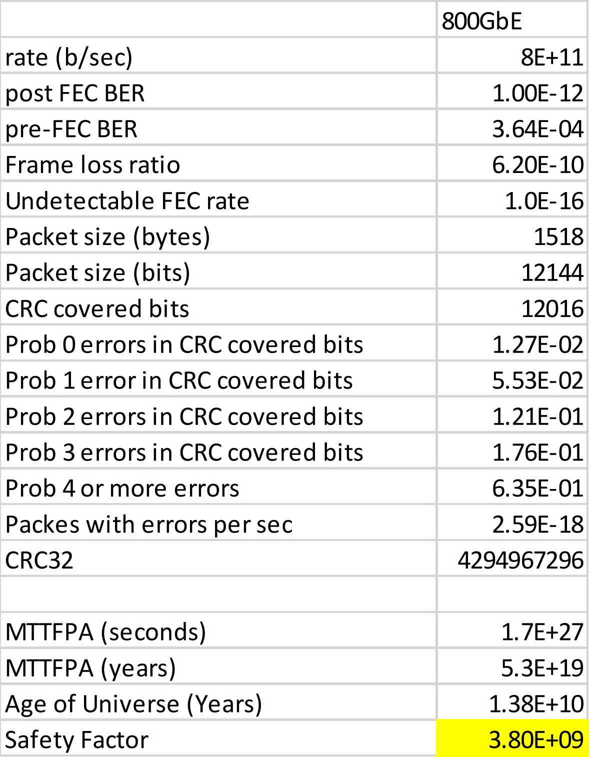
False Packet Acceptance Rate

IEEE 802.3ae, Albuquerque, 3/6/00 *64b/66b coding update* Agilent Technologies Innovating the HP Way

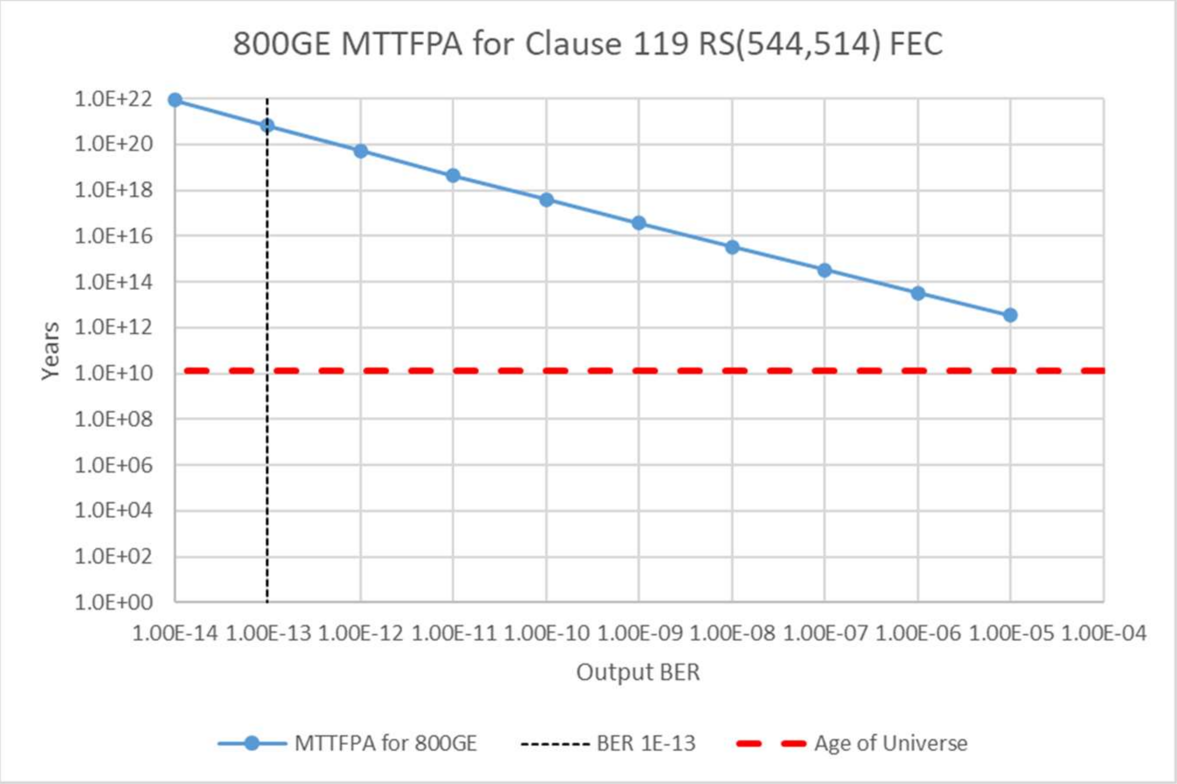
http://www.omnisterra.com/walker/pdfs.talks/albuquerque.pdf

# MTTFPA Calculations

- MTTFPA calculations show absolutely no concerns if we were to use the RS(544,514) FEC at these higher speeds and stayed at 1E-13 BER (or the equivalent FLR)
  - As long as error marking is on
  - All bets are off if you were to disable this…but why would you if you are concerned about passing bad packets
  - Thanks to Pete Anslow for his help on this!



| | 800GbE |
|---|---|
| rate (b/sec) | 8E+11 |
| post FEC BER | 1.00E-12 |
| pre-FEC BER | 3.64E-04 |
| Frame loss ratio | 6.20E-10 |
| Undetectable FEC rate | 1.0E-16 |
| Packet size (bytes) | 1518 |
| Packet size (bits) | 12144 |
| CRC covered bits | 12016 |
| Prob 0 errors in CRC covered bits | 1.27E-02 |
| Prob 1 error in CRC covered bits | 5.53E-02 |
| Prob 2 errors in CRC covered bits | 1.21E-01 |
| Prob 3 errors in CRC covered bits | 1.76E-01 |
| Prob 4 or more errors | 6.35E-01 |
| Packes with errors per sec | 2.59E-18 |
| CRC32 | 4294967296 |
| | |
| MTTFPA (seconds) | 1.7E+27 |
| MTTFPA (years) | 5.3E+19 |
| Age of Universe (Years) | 1.38E+10 |
| Safety Factor | 3.80E+09 |



# Application Needs?

- I think this really boils down to: what do the applications that are common at these speeds require from a BER/FLR perspective?

- DC applications do care about post FEC BER, which cause packet drops
  - These cause retries which can be slow to occur, slowing down overall throughput
  - Critical need is tail latency for ML applications, one drop can cause an increase in ML latency
  - Applications can add in extra overhead/redundancy to withstand some drops

- Typical BER is often more important than worst case
  - Typical BER is normally several orders of magnitude better than worst case
  - This is what on average impact applications

- Assuming you meet a BER of at least 1E-13, saving power is more important than further improving of the BER
  - Power is the most critical limiting factor in today’s data centers

- Overall: Balancing error rate with reduced power, complexity, latency and cost is critical to datacenter scaling

# Summary

- In study group: Adopt 1E-13 or better as the objective and decide in task force if we want a better BER for some or all of the PHYs (or the equivalent FLR)
  - Once we make progress on FEC structures, overhead, cost of an improved BER etc.

# Thanks!