# PCS error burst counting proposal

Adee Ran
Intel Corporation

# Supporters

- Richard Mellitz, Intel
- Arash Farhood, Cortina
- Rick Rabinovich, Alcatel-Lucent

# Problem statement

- CAUI-4 C2C receiver can include a DFE which can introduce error propagation.
- If CAUI-4 carries bit-muxed PCS lanes, error propagation can reduce MTTFPA.
- Assuming an adaptive DFE, error propagation is a system-level problem: the same receiver can either be totally safe or have severe error propagation, depending on channel conditions or transmitter transition time.
- No measurable result that correlates to MTTFPA is specified.
- Nothing in any of the CAUI-N specifications prevents using a DFE or addresses error bursts in any way.
- False packet acceptance is undetectable (by definition) and assumed to be very rare. Our unofficial objective (>AOU) is practically impossible to guarantee. We have no data on how real systems actually perform.

# Identifying bursts in the receiver

- Proposed below is a simple method of identifying error bursts and measuring their rate during normal receiver operation, **based on the existing BIP mechanism**: Multilane BIP Mismatch Counting (MBMC).
- Possible uses:
  - Reporting burst rates in stressed receiver tests.
  - Monitoring a full link (similar to BER estimation using BIP).

# How does it work?

- For the bit-muxing case, the CAUI-4 on the RX path interfaces PMA(4:20) attached to the RX lanes of the 100GBASE-R PCS.
- A burst of errors on one of the CAUI-4 lanes is thus striped across up to 5 PCS lanes (PCSLs).
  - For burst lengths of up to 5, the error bits will be mapped to one PCSL each.
  - For bursts longer than 5 bits, some PCSLs will get two or more adjacent errors.

# PMA demux from CAUI-4 to PCS

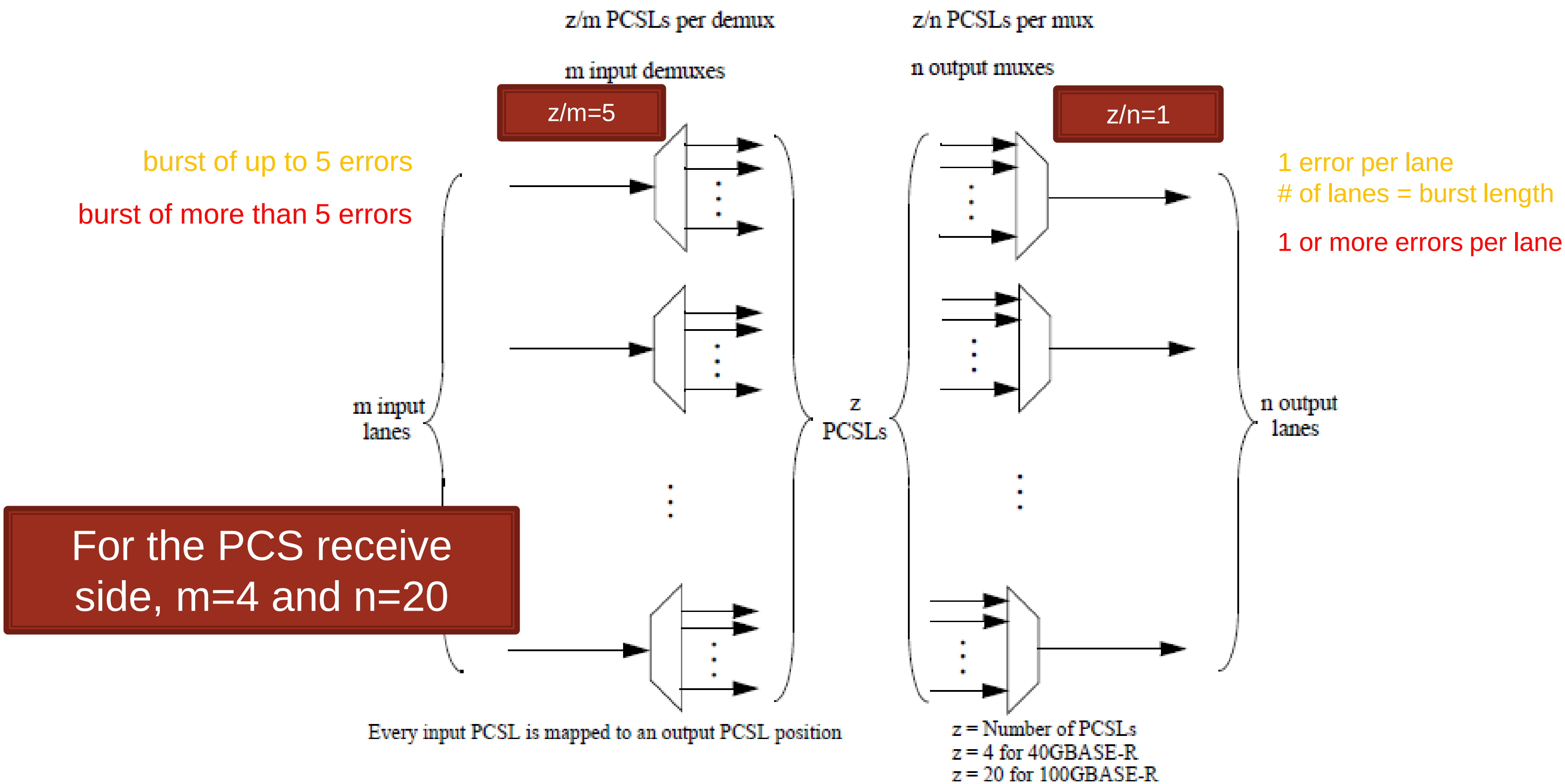


**Figure 83–4—PMA bit mux operation used in both Tx and Rx directions**

# Identifying bursts

- PCS detects errors on each PCSL separately using the BIP field in alignment markers (AMs).
  - Any event of up to 5 adjacent errors in the same PCSL will cause separate bit flips in the BIP field.
- After PCS lane alignment, AMs from all 20 lanes are available together as a group.
- After a burst of length L≤25 occurs, exactly L out of the 8*20 BIP bits in the next AM group will be flipped.

# Identifying bursts

- If the full link operates at BER=1e-12, errors are expected once per 10 seconds…
  - An isolated error will cause *one* of 20 the BIP counters to advance
  - If the error is propagated into a burst, *more than one* counter will advance
  - If one reads all 20 BIP counters 10 times per second (noting that they are clear-on-read) and sums the “1” bits then:
    - Getting 0 suggests no errors have occurred during this second
    - Getting 1 suggests a single error has occurred
    - Getting L suggests a single error burst of length L has occurred
    - “Suggests” assumes two or more independent bit errors within 0.1 second are unlikely; but in fact this is expected to happen once per 30 minutes .
- Under assumed BER levels, bursts are detectable and their lengths are measurable, but “false counts” may occur too often even with fast polling.

# Proposed improvement

- Monitoring can be made more accurate if Multilane BIP Mismatch Counting (MBMC) is implemented in the PCS:
  - Whenever a set of AMs is received, define L as the count of 1's in all BIP fields (= the burst length)
  - Define 4 new burst counters, one per value of L (1…4)
    - Whenever L>0, increment counter L (use counter 4 if L>4)
    - Make the counters clear-on-read
    - No need for more than 4, since even 4-error bursts should be very rare.
  - False counts occur only if two independent errors occur between two AMs.
    - Mean time to such event is >**28,000 years** for CAUI-4 (assuming BER=1e-15) or >**10 days** for a full 100GBASE-LR4 (assuming BER=1e-12).
    - As we shall see, 10 days is rare enough and doesn't create a problem.
- MBMC replaces polling the BIP counters and prevents false counts.

# Estimating MTTFPA based on MBMC

- If error propagation follows the Gilbert model [1] with parameter *a*, we can estimate *a* as the probability of 2-error bursts, p(EP), and calculate p(burst length≥4) as BER*p(EP)^3.
- If EP does not follow this model, errors can be more often:
  - e.g. two DFE taps with similar values can cause 3-error bursts with higher probability than expected: p(EP2)>>p(EP)^2.
  - More than two such taps can cause even more frequent 4-error bursts – but is less likely.
- If the test is performed on just a CAUI-4 link (no optical segment or negligible BER):
  - Measure the rates (events per second) of single errors $f_1$; 2-error bursts $f_2$; and optionally 3-error bursts $f_3$.
  - Estimate 4-lane BER as $p_1 = f_1 \cdot \frac{UI}{4}$, p(EP) as $p_{2|1} = f_2/f_1$, and optionally p(EP2) as $p_{3|2} = f_3/f_2$.
  - Estimate p(burst length≥4) for the whole CAUI-4 link as $p_1 \cdot p_{2|1}^3$ (optionally, $p_1 \cdot p_{2|1} \cdot p_{3|2}^2$).

[1] See cideciyan_02a_1111 in P802.3bj

# Estimating MTTFPA based on MBMC (cont.)

- If the test is performed on a full 100GBASE-LR4 link (which can have BER=1e-12 per lane), the rate of single errors $f_1$ can be dominated by the total link BER.
  - $f_1$ can be relatively large, but most of the errors are not on the CAUI-4 segment and thus do not propagate.
- Assuming 2-error events result only from error propagation on the CAUI-4 segment:
  - Measure the rates (number per second) of 2-error bursts $f_2$ and 3-error bursts $f_3$.
  - Estimate 4-lane 2-error burst probability as $p_2 = f_2 \cdot \frac{UI}{4}$, and p(EP) as $p_{3|2} = f_3 / f_2$.
  - Estimate p(burst length≥4) for the whole 100GBASE-LR4 link as $p_2 \cdot p_{3|2}^2$.

# Estimating MTTFPA – cont.

- Assume large MAC frames so approximately all error locations are “dangerous”
  - Shorter frames are safer (see backup).
- Assume any 4-error burst on the 4-lane link can create a CRC collision with p=$2^{-32}$.
- Estimated MTTFPA is

$$\frac{UI/4}{p(burst \geq 4) \cdot 2^{-32}}$$

$$\cong \frac{1.4 \cdot 10^{-9}}{p(burst \geq 4)}\ years$$

# Estimating MTTFPA – cont.

- Example:
  - If each of the four lanes has BER=1e-15 and measured burst rates yield **p(EP)=0.02** and **p(EP2)=0.1**, then

    $$p(burst \geq 4) = 10^{-15} \cdot 0.02 \cdot 0.1^2 = 2 \cdot 10^{-19}$$

    Resulting in MTTFPA≈7 billion years.
  - This is shorter than AOU, even though p(EP) is apparently small enough; suggests p(EP2) has to be used too.
- This estimate assumes max frame size, no idles, and all lanes are worst case; so it includes considerable guard band, and suggests the CAUI-4 segment is probably safe.

# How fast is MTTFPA estimation?

- Results presented in the ad-hoc meeting (see backup) show that a rough safe/unsafe decision can be made within a couple of days of operation.
  - Even if testing for sufficient time to detect 3-error bursts with good confidence.
- This may be considered too long for some uses; but we can consider running with increased stress to enable faster estimates (as will probably be required for BER testing as well).

# Is it needed if we adopt solution X?

- Specifying limits of DFE taps
  - How can anyone confirm this specification is met? ➔ Using MBMC!
- Differential encoding (precoding)
  - Can create multi-burst error propagation patterns such as 100001 (safe), 11011 (unsafe), 110011 (unsafe)…
  - These will be mapped to non-consecutive locations in the MAC frame and are not guaranteed to be detectable by CRC.
  - MBMC can detect this kind of bursts too – it actually measures burst *weight* rather than length.
- Block muxing/FEC: if adopted, probably no need for MBMC.

# How to treat the results?

- **Thresholds?**
  - MTTFPA should ensure good operation of a large network. But there is no reason to assume all links are worst-case simultaneously.
  - Even with very high p(EP), CAUI-4 BER of 1e-15 yields MTTFPA in millions of years.
  - If a *typical* links have MTTFPA of billions of years, and if bad links aren't common, the network can be assumed safe.
  - ➔ Suggest calculated MTTFPA > 1e9 years.
- **Normative or informative?**
  - PCS implementations already exist, some already deployed; can't rely on a new feature.
  - Good confidence requires ~90 hours of test time; testing every link this way is impractical.
  - ➔ Suggest an informative recommendation.

# Proposal

1. Add MBMC as a new optional PCS feature
   - Detailed draft changes discussed in CAUI-4 ad hoc. Updated version is available if adopted.
2. Add a *recommendation* that MBMC results based on a 90-hour measurement yield:

$$P(EP_1) \cdot P(EP_2)^2 < 3 \cdot 10^{-5}$$

3. Add a *recommendation* that MBMC results based on a 90-hour measurement yield:

   Estimated MTTFPA > 1e9 years

# Backup

# Effect of frame length

- Since the CRC does not span the IPG, the ratio of frame size to minimum IPG affects the MTTFPA: the shorter a frame is, the fewer positions it has for starting an "unsafe" burst.
- MTTFPA calculations should have a "safety factor" in p(FPA), dependent on frame size.
- For frame sizes below 2944 bits, CRC can always detect up to 5 errors [2]. Safety factor is 0.
- For frame size of 179*64=11456 bits (slightly below MTU limit):
  - Adding IPG and sync headers yields 11880 bits at the PCS.
  - There are only 616 initial locations for a CAUI-4 4-error burst which are "safe" (guaranteed to be detected): sync headers, last 3 blocks and IPG; "safety factor" is $\frac{11880-616}{11880} \cong 0.95$.
- We can approximate safety factor is 1 in the worst case.

[2] Koopman, P. "**32-bit cyclic redundancy codes for Internet applications**", Proc. DSN 2002. See table 1.

# Example

- Let's consider a CAUI-4 which operates at worst-case compliant conditions:
  - All four lanes have BER=1e-15
  - Gilbert model with p(EP)=0.03
  - ➔ MTTFPA ≈13e9 years (according to slide 12)
- Estimate how fast the counters advance for this system, and compare to cases when either its BER or its p(EP) are increased.

# Results

| Scenario | BER=1e-15; EPP=0.03 | BER=1e-14; EPP=0.03 | BER=1e-15; EPP=0.3 |
|---|---|---|---|
| Mean time to a single error (any BIP mismatch) | **2.7 hours** | **16 minutes** | **2.7 hours** |
| Mean time to burst with L=2 | **3.7 days** | **9 hours** | **9 hours** |
| Mean time to burst with L=3 | 125 days | 12 days | 30 hours |
| Mean time to burst with L=4 | 11 years | 59 weeks | 4 days |
| MTTFPA estimate | 13 billion years | 1.3 billion years | 13 million years |
| Mean time to false count of 2 uncorrelated errors | 28 thousand years | 284 years | 28 thousand years |