PCS error burst counting proposal

Adee Ran
Intel Corporation

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.
- Nothing in any of the current or proposed 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.
- No measurable result that correlates to MTTFPA is specified.

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 for short).
- 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

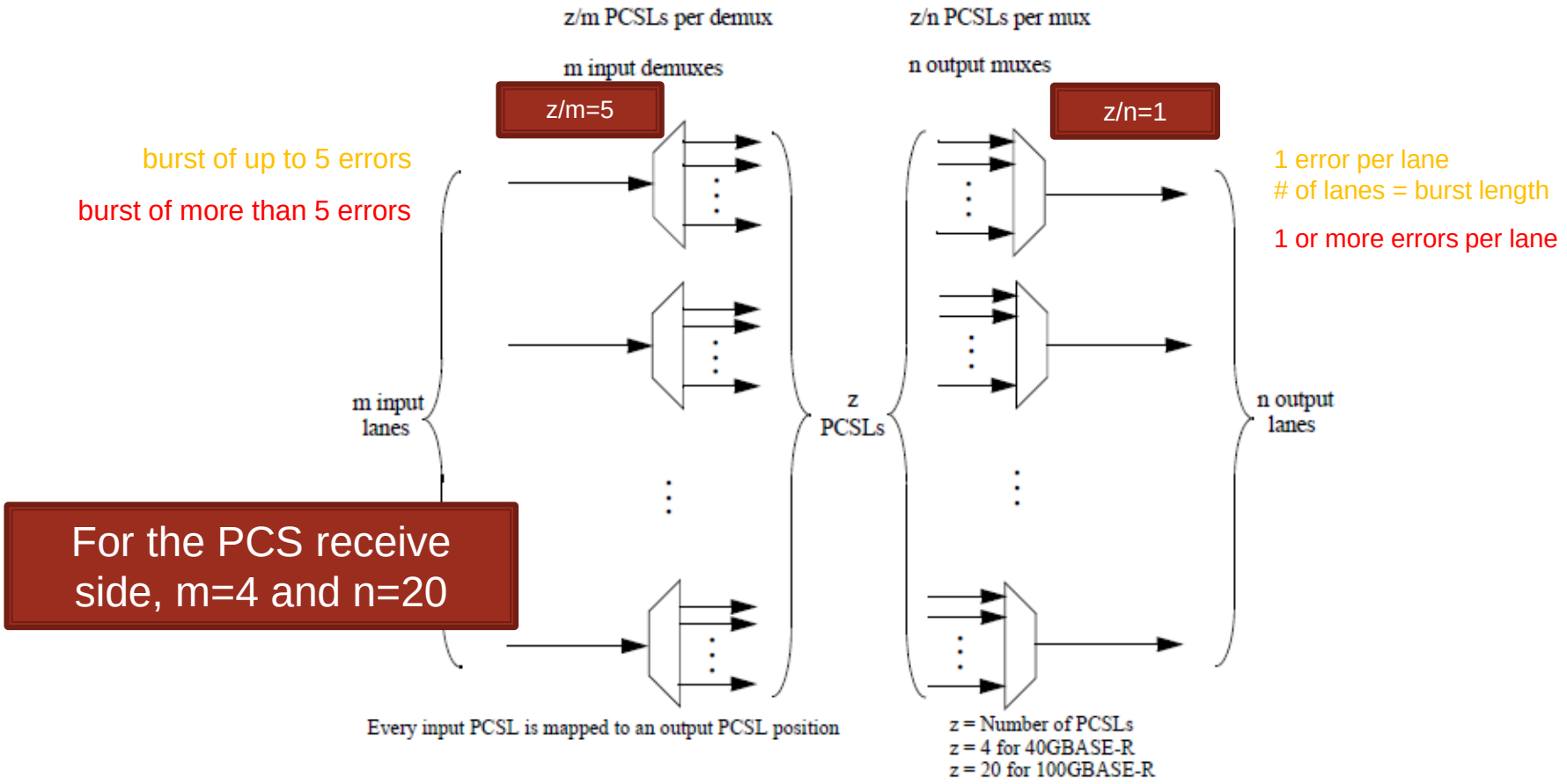

Figure 83–4—PMA bit mux operation used in both Tx and Rx directions

Identifying bursts

- PCS detects errors on each PCSL separately using the BIP field in alignment markers (AMs).
  - BIP can identify any event of up to 5 adjacent errors in the same PCSL with separate bit flips in the BIP field.
    - This means bursts up to 25 errors will be detectable with the accurate length.
  - Having more than one burst between adjacent AMs may flip some BIP bits twice; but assuming CAUI-4 has BER<1e15, this is extremely unlikely.
- After PCS lane alignment, AMs from all 20 lanes are available together as a group.
- After a burst of length L occurs, exactly L out of the 8*20 BIP bits in the next AM group will be set.

Identifying bursts

- If the full link operates at BER=1e-12 then 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 error 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 very unlikely; but in fact this is expected to happen once per 5-6 hours.
- Under assumed BER levels, bursts are detectable and their lengths are measurable, but "false counts" may occur if polling isn't fast enough.

Proposed improvement

- Monitoring can be made easier and 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
    - More than 4 can be used, but we assume even 4-error bursts should rarely occur.
  - False counts occur only if two independent errors occur between two AMs; this has negligible probability.
- MBMC replaces polling the BIP counters and prevents false counts.

Estimating MTTFPA based on MBMC

- Assumption: all four lanes have same BER and error propagation following the Gilbert model [1] with probability p(EP) ➔ same p(burst length≥4).
- Under this assumption:
  - Measure the rate of single errors f_1 over time; estimate 4-lane BER as $p_1 = f_1 \cdot UI$
  - Measure the rate of 2-error bursts f_2 over time; estimate p(EP) as $p_2 = f_2/f_1 \cdot UI$
  - Optionally: measure the rate of 3-error bursts f_3 over time; estimate p(EP^2) as $p_3 = f_3/f_2 \cdot UI$
  - Estimated p(burst length≥4) for the whole CAUI-4 link is $p_1 \cdot p_2^3$ (optionally, $p_1 \cdot p_2 \cdot p_3^2$)

[1] See cideciyan_02a_1111 in P802.3bj

Estimating MTTFPA – cont.

- Assume frames are 179*64=11456 bits long
  - Slightly below MTU limit
  - Shorter frames improve MTTFPA; and below 2944 bits, CRC can always detect up to 5 errors [2]
- Adding IPG and sync headers yields 11880 bits at the PCS.
- There are 11264 out of 11880 locations where a dangerous 4-error burst can be placed
  - Excluding all sync headers, last 3 blocks and IPG.
- Assume a 4-error burst starting on these locations can create a CRC collision with $p=2^{-32}$

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

Estimating MTTFPA – cont.

- Estimated MTTFPA is

$$\frac{11880/4\ UI}{p(burst \geq 4) \cdot 2^{-32} \cdot 11264} \cong \frac{1.4 \cdot 10^{-9}}{p(burst \geq 4)}\ years$$

- Example: if all four lanes have BER=1e-15 and p(EP)=0.03, we get MTTFPA ≈13 billion years.
  - This estimate assumes max frame size, no idles, and all lanes are worst case.
  - But it also assumes the Gilbert model; If EP does not follow this model, long bursts may occur more often than expected.
    - e.g. two DFE taps with similar values can cause 3-error bursts with almost the same probability as 2-error bursts.
    - More than two such taps can cause frequent 4-error bursts – seems unlikely.
- In practice, calculating p(EP)=0.03 means the CAUI-4 link is probably safe if it meets the BER requirement.

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 assuming bad links aren't common, the network is 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

- Add MBMC as a new optional PCS feature
  - Detailed draft changes discussed in CAUI-4 ad hoc. Updated version is available if adopted.
- Add a *recommendation* that calculated MTTFPA using MBMC based on a 90-hour measurement is above 1e9 years.

Backup

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 ≈15e9 years (according to slide 11)
- 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	380 years	38 years	14 days
MTTFPA estimate	13 billion years	1.3 billion years	13 million years
Mean time to false count of 2 uncorrelated errors	6,000 years	60 years	6,000 years