Proposal for FEC in 40G/100G Ethernet

Piers Dawe, John Petrilla

Avago Technologies

Supporters

- Jonathan King — Finisar
- Petar Pepeljugoski — IBM
- Mike Dudek — JDSU

Contents

Why now?

- Why this project?
  - The need for and use of FEC increases as signalling speeds rise, and digital gates get cheaper and lower power
- Why this meeting?
  - It's overdue
  - Catching the "last proposal" deadline
  - Enabled by evolving and stabilising HSE architecture
    - Need and opportunity is "across the sublayers"
- **P802.3ba (HSE) should consider FEC more seriously than previous projects**

Everyone else uses FEC

- FEC (forward error correction) turns a mediocre to bad BER into a good BER. It mitigates noisy eyes. Many versions exist with different gains and costs
- Ethernet uses FEC in:
  - 1000BASE-PX (EPON)
  - 10GBASE-KR
  - 10GEPON
  - 10GBASE-T
  - DSL
- High end long haul telecoms use FEC
- **P802.3ba links have limited power budgets and SNR, for e.g. eye safety reasons. Parallel lanes have crosstalk. We can benefit from FEC too.**

Common PCS implies common FEC

- PCS for 40G and 100G be very similar, and based on 64B/66B
- Same PCS for all port types – medium independent
- Includes virtual lanes or similar, and lane striping
- "Logically" same for 25G lanes as for 10G lanes
- Best place for FEC is in an ASIC with other line-speed digital stuff – i.e. with the PCS
- Free
  - If you believe gates are free, FEC is free
  - Allows medium-independent link monitoring and protective maintenance
    - See backup
- Desirable for multiple PMD types
  - I expect FEC will be essential for 25G lanes and long reach 10G
    - See next slide
  - Already expected as at least an option for backplane
  - Will benefit the PMDs that can be made to work without FEC
- **A common, medium-independent PCS implies a common, medium-independent FEC**

Which PMDs need FEC?

- All would benefit from FEC
- Links with optical amplifiers will benefit 50% more
  - Similarly with APDs (avalanche photodiodes)
- Any future high-tech 100G modulation scheme e.g. DQPSK will probably need FEC
- Anything at 25G/lane
  - If optical: for constant laser SNR/Hz (RIN), the extra bandwidth means worse SNR on faster channels. We can spend the SNR on EDC
  - If electrical: 10GBASE-KR backplane Ethernet has already chosen to allow FEC, 25G crosstalk will be significantly worse
- Links significantly affected by near-Gaussian crosstalk or "random" jitter will benefit significantly
- Expect all 25G, all 1550 nm, links with SOA, and some others, will need FEC
- Don't know enough about 10G/lane 10 m electrical cable to comment
- Is that every PMD type?

Line rate

- CDRs will be used as glue within boxes
- There is no economic justification for redesigning these modules and ICs, especially the several-times-respun CDR chips within, for the low volumes of HSE in the first few years
  - Should share volume with 10GE
- Some link types will have multiple 10G modules at one end
- Therefore the 10G signals must be at the 10G line rate, and have the characteristics of a 10GBASE-R signal (statistically balanced etc)
  - Will pass correctly through a XFP or SFP+ module
- **Therefore the FEC must be in band**
  - e.g. KR FEC ~1dB gain, to 7% FEC for ~3dBo gain
  - Payload might be ~9.3 G*"bit"*/s/lane for 10G lanes
- Compare WAN PHY rate: 9.58464*64/66 ~9.2942 G"b"/s
  - What's a "*bit*": these payload numbers are somewhat squishy because there is non-information (idles, preamble, FCS,...) in an Ethernet stream, that could be removed if wished
- 25G line rate should be exactly 10 / 4 * 10.3125 GBd
  - Allowing simple 10:4 gearbox
- Use gapped clock concept to keep the MAC speed at 100G or 40G, as for lane markers
  - Some idles can be removed using an RS Deficit Idle Count

FEC vs. EDC

- Both take some silicon and power
- FEC copes well against random noise (poor SNR). EDC needs better-than-usual SNR to work
- EDC copes well with slow but linear "channels" (e.g. coax cable). FEC won't open a systematically closed eye but will receive a noisy one
- EDC has dynamic range issues. FEC doesn't
- FEC adds more latency than analog EDC, but digital EDC can be late too
- FEC and EDC can be used in combination
  - Design of each should be mindful of the other

FEC vs. CDR

- Both take some silicon and power
  - We are told <50 mW* / 10G lane (both directions) for FEC vs. 200 mW / 10G lane / direction / end for CDRs. 50 mW vs. 800 mW
- FEC copes well against random noise (poor SNR). CDR removes random jitter and noise. Have to repair signal before it's too late
- FEC can be integrated nicely into the PCS. CDRs should be placed in the right place; may need extra packages
- FEC adds some latency, CDRs only 1 or 2 UI
- FEC protection domain can span several CDR jitter domains
- FEC and CDR can be used in combination
  - * at 0.13 um

Block size and latency

- Large blocks are more effective but cost more latency (and power?)
  - FEC blocks can be shortened, at a cost in efficiency
- KR blocks are 32, 66-UI blocks long including 32 check bits = 204.8 ns ~ 41 m of cable (one way)
- RS(255, 239) blocks are 255 bytes including (255-239=16) check bytes = 2040 UI = 197.8 ns
- 10GEPON RS(255, 223) blocks also 255 bytes
- Latency is a minimum 2 and a bit blocks (?)
- For comparison,
  - LRM PMA and PMD round-trip delay limit is 921.6 ns
    - (Believe current implementations are way lower than this)
  - 10GBASE-T PHY round-trip delay limit is 2,560 ns
    - ("Too much")

"Systematic" code

- In FEC, "systematic" means that the information to be protected is transmitted as-is, with the protecting information in addition
- FEC encoding seems like CRC generation but on much shorter and fixed size blocks
  - Low power, low latency, low cost
- Decoding/error correction is more interesting
  - But still lower power than auxiliary CDRs!
  - We are told <50 mW / 10G lane (both directions) for FEC vs. 200 mW / 10G lane / direction / end for CDRs

Above or below the striping?

- Expect FEC to be below or in place of the 58-bit scrambler in 64B/66B code
  - Avoid having to correct 3 errors for every one on the line
- One FEC engine acting on aggregate stream or as many FEC engines as physical? virtual? lanes?
  - If per virtual lane, can FEC-correct or deskew, or both, at will and in either order
- More gates vs. faster gates
  - How practical is FEC correction at 100 Gb/s?
- Interleaving
  - Lane distributor provides a free interleaving function
    - Bursts of errors will be split up, e.g. seen as every 10th bit the other side of the distributor
    - If a 10-bit burst is rare, is that enough to be useful? (or is it the other way round: if consecutive errors get turned into errors 10 bits apart, KR burst correction won't work so well?)
- Latency
  - Reduced latency for same block size if above the striping

Auto-negotiation?

- If FEC everywhere, and OK for sync-up time, then not needed
- If "systematic" FEC, can transmit always: AN not needed
- If not, could start link without FEC, agree to turn it on
  - In concept, like 10GBASE-KR, but can be nicer
  - As the lanes are at 10.3125 GBd or 41.25 GBd anyway, no need to negotiate at a different signalling rate
    - 10GBASE-KR uses 312.5 GBd, the highest common denominator of 3.125 GBd and 10.3125 GBd. This very different rate is an unnecessary burden for CDRs
  - Any AN protocol needs to be robust to errors
    - Not a problem, just a design criterion
    - Can use lane markers or ordered sets
- Or start with FEC on and turn it off if the link partner doesn't show he understands it
  - Like Fibre Channel speed negotiation

Encoding, decoding, power and size

- Encoding, the straightforward part, takes little power
- Verifying a perfect block is similar to encoding
- Correcting an imperfect block takes more calculation
  - Believe that FEC power consumption depends on error rate: if no errors, few gates toggle, low power
  - If this is not true, and if "systematic" code, receiver can turn off its FEC, without auto-negotiation, if it determines that its error rate is low enough
    - And turn it on if it deteriorates
    - Doesn't matter if reverse direction is using FEC or not
- It is assumed that the encoder size of RS(255, 239) for 10Gbps is about 20Kgates; decoders are larger
  - See 3av_0707_daido_1 for more information
- 10GBASE-KR FEC estimated at 16 kgates + RAM (both directions)
  - See 10GBASE-KR_FEC_Tutorial_1407.pdf page 73

FEC and optical amplifiers

- FEC gain depends on the link type
  - e.g. for a traditional optical link like 10GBASE-L, 10GBASE-KR FEC has an SNR gain of ~2 dBe, translating into ~1 dBo power budget gain, e.g. lower required transmit power or higher (relaxed) sensitivity. RS(255, 239) check has about 3 times that gain.
  - for a RIN or jitter limited link the SNR gain is as above but the power budget gain can be as good as infinite: "removes an error floor"
    - This may apply to some crosstalk situations
  - For an optical link using an optical amplifier at the receiver, the noise depends on the signal such that the expected power budget gain is 50% more: 1.5 or 4.5 dB
    - If the optical amplifier is an SOA (semiconductor optical amplifier) there may be an additional issue with crosstalk and patterning near overload and additional benefit

Proposed staged decisions

- P802.3av, 10GEPON did this very well, spread over about a year (shown in backup):
- Interested in FEC?
- Optional or mandatory to transmit?
- Transmit whole information ("systematic")?
- 66-bit oriented block size?
- What block size / latency is acceptable?
- How much SNR gain needed?
- Choose FEC code to deliver above
- **P802.3ba, HSE should work through same or similar decision path**

How strong?

- Think of 5 categories:

1. Ostrich
   - "We didn't need it when I was young"
   - 0% lost throughput, 0 dB benefit
   - Not a likely outcome of the analysis
2. KR
   - 1.5% but scavenged => 0% lost throughput, ~1 dBo, 2 dBe benefit
3. ~4%
   - Intermediate
4. RS(255, 239) or similar strength
   - ~7% lost throughput, ~3 dBo (un-amplified), 6 dBe
5. RS(255, 223) e.g. 10GEPON
   - 12.9% lost throughput, ~7.2 dBe
6. Super FEC
   - Up to 33% lost throughput, substantial gain (e.g. 8dBo?) , long latency
   - Not a likely outcome of the analysis

Any of the middle options 2, 3, 4, 5 would be a better choice than sleepwalking

There are many code options: choose after deciding the characteristics we want

Implication

- Need to have believable estimates of SNR for all PMD types in the frame
  - "In the frame " means all in the first wave of standardization (P802.3ba), and if known, highly likely candidates in second wave
  - Might expect very high end PMDs to provide a secondary FEC function
  - But much cleaner if same FEC can be used for all link types
- If estimates not available in time, be conservative and pick a moderate FEC

Conclusions

- P802.3ba needs FEC more than previous projects
- Because PCS is to be medium-independent, expect FEC will be also
- FEC everywhere (or nowhere – but that is not an option), at least optionally
- Use step-by step decision making per slide 16
- Choose FEC strength based on thorough SNR estimates of all first-wave PMDs

References

- 10GBASE-KR FEC Tutorial
  http://ieee802.org/802_tutorials/july06/10GBASE-KR_FEC_Tutorial_1407.pdf
  See IEEE Std 802.3, 74
- 802.3ah EFM
- Reed-Solomon code (255, 239) operating on 8-bit symbols
  See IEEE Std 802.3, 65.2 and e.g.
  http://ieee802.org/3/efm/public/sep01/rennie_1_0901.pdf
  http://ieee802.org/3/efm/public/jul02/p2mp/khermosh_general_1_0702.pdf
  http://ieee802.org/3/efm/public/sep02/p2mp/rennie_p2mp_1_0902.pdf
- 802.3av
  RS(255, 223)
- See P802.3av 92.3 and e.g.
  http://ieee802.org/3/av/public/2007_07/3av_0707_daido_1.pdf
- http://ieee802.org/3/av/public/2007_01/3av_0701_effenberger_1.pdf
- http://ieee802.org/3/av/public/2007_03/3av_0703_kramer_1.pdf
- http://ieee802.org/3/av/public/2007_05/3av_0705_lynskey_1.pdf
- http://ieee802.org/3/av/public/2007_05/3av_0705_effenberger_4.pdf

Backup

KR FEC Summary

KR FEC Pros & Cons

- **Costs**
  - Latency: if FEC is done per lane,218 ns without error marking, 422 ns with error marking
    (Note 422 ns is the time of flight for ~84 m of fiber.)
    Reduced latency if FEC is done on aggregate stream
  - Chip Area: TBD
  - Power Dissipation: ~100 mW
- **Benefits**
  - Free - likely already available for backplanes and copper cables support
    Likely needed for 100G, 40 km
  - Keeps solution within host IC
  - Turn on when needed to correct errors due to longer fiber lengths or PCB traces, perhaps in combination with a CDR on Tx side.
  - Corrects raw 4.4E-8 BER into 1E-12 BER.
  - Enables continuous error monitoring without requiring traffic
  - No increase in overhead – no sacrifice of data rate
  - Handshaking for compatibility with a non-FEC partner already developed by KR
    - Would use simpler scheme with 10.3125 GBd signalling where appropriate

Comfort blanket

- FEC allows counting corrected errors
  - Can monitor the health of a link when it is perfectly well
  - Could do preventive maintenance if wished
  - As long as using "stream FEC" (protecting basically all the bits on the line) rather than "burst FEC" (protecting frames but not idles, as in GEPON), can monitor the health of the link even when there is no traffic
    - Don't need FEC to do this: could add idle PCS coding violation counter anyway

FEC vs. MAC

- Considering the transmission aspect:
- A CDR will clean up a signal (jitter bulkhead) but not guarantee it
- A FEC function will correct errors, will guarantee* that it's good and/or flag any errors
- A whole port with MAC and CRC function will guarantee* that it's good and/or flag any errors but WON'T correct them
- If a medium-heavy repeater is desired, it can use FEC without needing a whole MAC, at considerable saving in complexity
- * Not really a guarantee – everything in life is statistical

Staged decisions in 10GEPON

- P802.3av, 10GEPON did this very well, spread over about a year:
- Interested in FEC?
  - 2006? Sept 09, presentations on various FEC coding concerns, FEC gain requirement for 10GEPON downstream, redundancy, coding and latency
- Optional or mandatory to transmit?
  - Mar 07 10GEPON shall accommodate FEC's parity bandwidth by reducing the MAC's effective data rate (sub-rating)
- Transmit whole information ("systematic")?
  - Mar 07 Scheme in slides 3-7 in 3av_0703_mandin_2.pdf as baseline for upstream FEC framing and synchronization
- 66-bit oriented block size?
  - Mar 07 Baseline scheme the FEC codeword structure in illustration on slide 5 in 3av_0703_mandin_2.pdf for downstream (so that the FEC codeword structures on the upstream and downstream are identical). Slides 3-7 in 3av_0703_mandin_2.pdf as baseline for upstream FEC framing and synchronization.
- May 07 FEC algorithm's input Nx65bit payloads (2nd bit of sync header plus 64 bits of data) pre-pended with padding consisting of zeros to bring the input codeword to the required size; notwithstanding, both bits of the sync header shall be transmitted, while the padding shall not be transmitted, as in 3av_0705_effenberger_4.pdf. Baseline for FEC framing 3av_0701_effenberger1_1.pdf, 3av_0703_kramer_1.pdf, and 3av_0705_lynskey_1.pdf.
- What block size / latency is acceptable?
- How much SNR gain needed?
- Choose FEC code to deliver above
  - Nov 07 RS(255, 223)
- **P802.3ba, HSE should work through same or similar decision path**