An exploration of the technical feasibility of the major technology options for 400GE backplanes

Brian Holden

Kandou Bus, S.A.

brian@kandou.com

IEEE 802.3 400GE Study Group

July 16, 2013

Geneva, Switzerland

KANDOU BUS

IP Disclosure

- As mentioned in the initial presentation of related material at the May 2013 Victoria 802.3 meeting, this presentation includes technology that may be the subject of multiple patent applications, applications in process, and patents by Kandou Bus, S.A.

- Assuming that this Study Group results in a PAR, that Kandou Bus, S.A. is committed to filing an Letter of Assurance against that PAR. (IEEE Patcom will not accept a LOA against a study group.)

- That LOA will guarantee that licenses to patents derived from these applications will be available on a Reasonable And Non-Discriminatory basis, should Kandou's technology be adopted into the 400GE specification

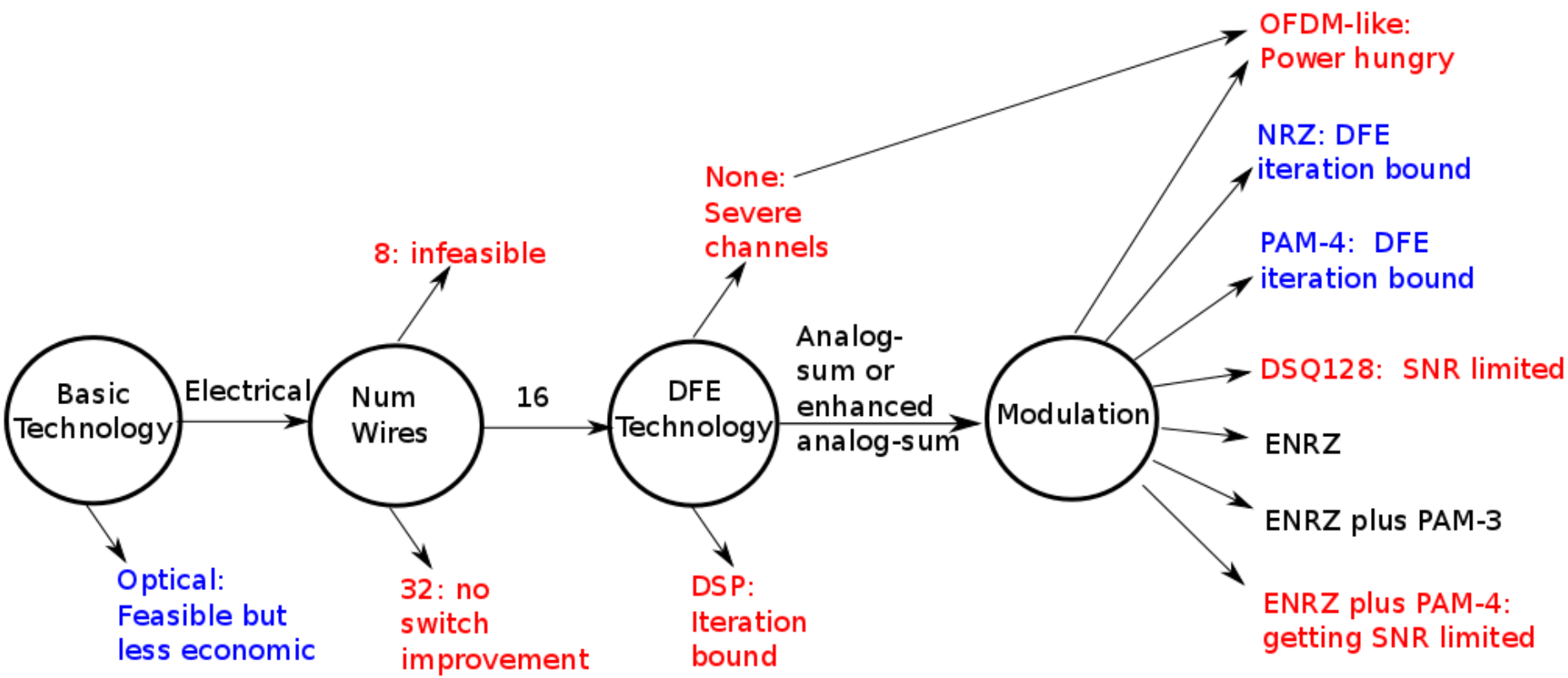
This presentation compares the major technology options for 400GE backplanes

- This presentation will compare:
  - Electrical vs. Optical
  - 8-wire vs. 16 wire vs. 32 wire
  - Modulation techniques
    - NRZ
    - PAM-4
    - DSQ128
    - OFDM-like
    - ENRZ
    - ENRZ plus PAM-3
    - ENRZ plus PAM-4
  - Time allowing, we will also discuss the relationship between ENRZ and PCS per questions asked during the Victoria meeting

KANDOU BUS

Decision tree for the options for 400GE backplanes

- **Below is a decision tree for 400 GE backplanes that this presentation will review in detail:**



Electrical vs Optical for 400GE backplane I/F

- In the 100GE generation, there was no compelling evidence and little discussion that an optical backplane was at all desirable or economic
  - Optical backplanes are clearly technically feasible
    - They are still in the high cost end of the spectrum
  - MCM and interposer technology has brought integrated optics along

- That said, having an electrical backplane for 400GE is still a highly desirable objective
  - If it can be accomplished, it will almost certainly be less expensive than having an optical backplane

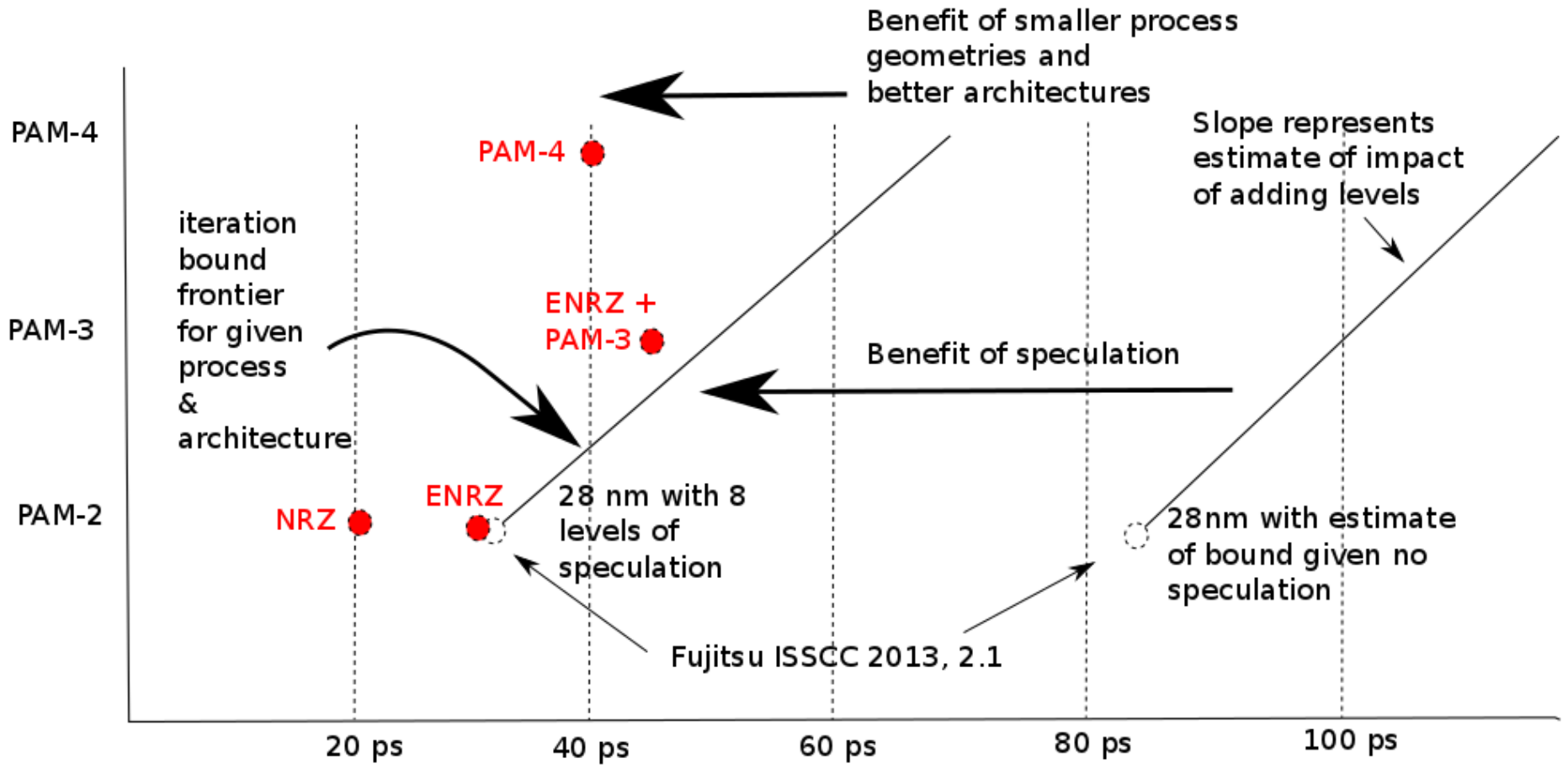
8 vs. 16 vs. 32 wires for 400GE backplane electrical I/F

- 8 wire
  - 10GBASE-KX4, 40GBASE-KR4, 100GBASE-KR4 and 100GBASE-KP4 all use an 8 wire electrical interface
    - This looks infeasible on its face – 50 Gb/s per wire (100 Gb/s per pair) is needed
    - No channels have been presented that support anything near this
- 16 wire
  - This looks like the sweet-spot – 25 Gb/s per wire is needed
    - Less dense than the 100GE generation, but possibly feasible
    - Extra pins does limit the switch bandwidth improvement
- 32 wire
  - Since switch cards are sometimes connector-pin limited, this would significantly limit the bandwidth gains available to 400GE generation switches
    - This would probably be a bad choice

DFE technology is key to the modulation choice for 400GE backplane electrical I/F

- The choice of the modulation scheme is partially driven by DFE technology
  - DFEs are needed to handle the challenging channels
  - The first tap is the most important tap
  - Virtually any DSP DFEs will be iteration-bound and not useful at these high rates
- One or more unit intervals will need speculation
  - An example state-of-the-art DFE is from 2013 ISSCC paper (Fujitsu, paper 2.1) used three UI's of speculation over two cross-linked phases (8 total guesses) to run at 32 GBaud at 28nm
  - There is on-going research on enhanced architecture DFEs
- Speculative DFEs for PAM-4 suffer a 4^N exponential speculative state increase
  - DFEs for PAM-3 suffer a 3^N exponential speculative state increase
  - Get diminishing returns from speculation because of mux delays
- The speed of DFEs are limited by their iteration bounds

Cartoon of the DFE iteration bounds for 400GE backplane electrical I/F



KANDOU BUS

Modulation Choices for 16-wire 400GE backplane electrical I/F

- **NRZ**
  - The typical choice, but the channels look grim at 50 GBaud
    - The best backplane-sized channels seen to-date dive off badly
    - It is unclear if this can really be made to work
    - The required large DFEs look hard to implement at 50 GBaud – may need many taps of speculation in order to close the loop
- **PAM-4**
  - PAM-4 is not as robust as NRZ
    - Requires a more ideal channel – harder as the rate increases
    - With a 1 Volt swing, the transmit level difference is only around 330mv
  - The required DFEs look hard to implement at 25 GBaud
    - For each speculative sample needed in order to close the loop, 4 states are needed. If more than one sample is needed, a 4 to the N state explosion happens in the inner loop

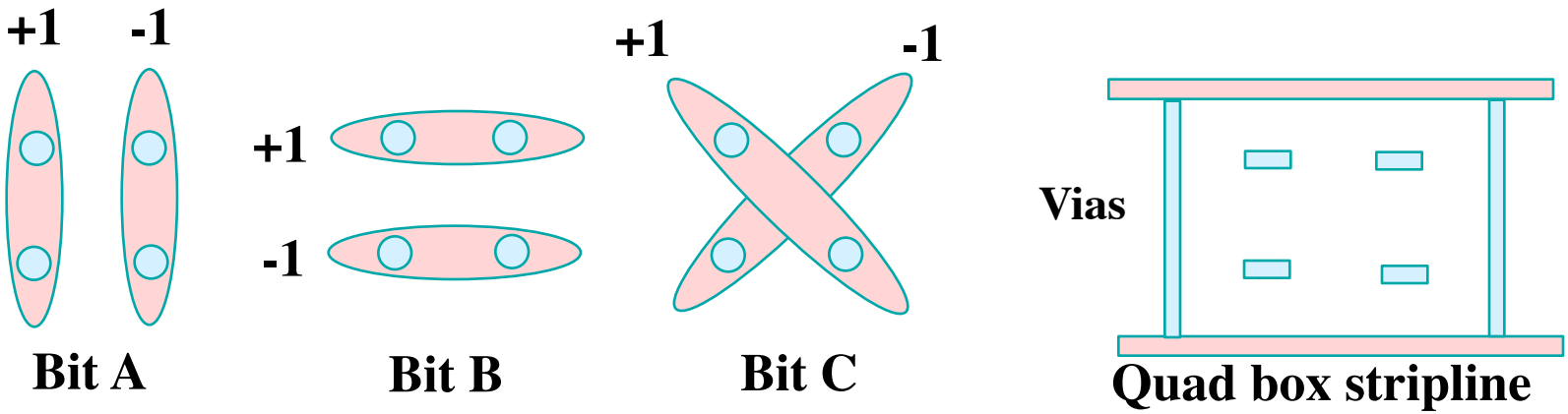
Modulation Choices cont. (16-wire 400GE)

- DSQ128 (from 10GBase-T)
  - This is a PAM-16 variant that gets 3.5 bits per symbol
    - (3.125 bits per symbol are delivered in the specification after some other coding is added)
    - The supported loss will be severely limited
    - The signal levels of adjacent symbols would only be 88 mv or so apart at the transmitter (1 volt * root2 / 16)
  - The required DFEs are still hard to implement even at 14.3 GBaud
    - If even one speculative sample is needed in order to close the loop, 128 speculative values would need to be calculated

- OFDM-like solutions
  - These can have have high spectral efficiency – can get 6 or more bits per Hz
    - DSP-like hardware is needed at both ends
    - The power consumption of these can get out of hand quickly
    - The complexity of these can get out of hand quickly

Review of Ensemble NRZ

- Ensemble NRZ coding delivers 3 bits over an ensemble of 4 wires
  - The symbol rate is 2/3rds of what is required for differential NRZ for the same throughput
  - The line power is 1/3 of the three differential NRZ channels that are otherwise needed at the same rate and throughput
- The 4 wires in the ensemble must have low intra-ensemble skew, on the same order as differential's intra-pair skew
  - Ensembles are terminated jointly to an AC ground at the receiver
- The transmit codewords are:
  - the 4 permutations of (1, -1/3, -1/3, -1/3) plus the 4 permutations of (-1, 1/3, 1/3, 1/3)



Modulation Choices cont. (16-wire 400GE)

- Ensemble NRZ (ENRZ)
  - Ensemble NRZ gets 3 bits per symbol over 4 wires
    - The symbol rate would be 33 GBaud, the per-wire rate is 25 Gb/s per wire (equivalent to 2 x 50 GBaud differential NRZ pairs)
    - Ensemble NRZ receivers are reference-less, like NRZ
    - For a 1 Volt total swing, the effective sub-channel swing for all three sub-channels is 500mv (measured at the output of a gain-of-1 inverse H4 transform with a zero-loss channel)
  - With a new specification that constrains the imbalance of the reflections amongst the four wires, binary DFEs can be used
    - Imbalances cause crosstalk between the three sub-channels
    - 33 GBaud binary DFEs will be much easier to implement than 50 GBaud binary DFEs
  - Increases the throughput without changing the DFE iteration bound frontier by delivering an effectively better channel
    - Four correlated wires are better than two

Modulation Choices cont. (16-wire 400GE)

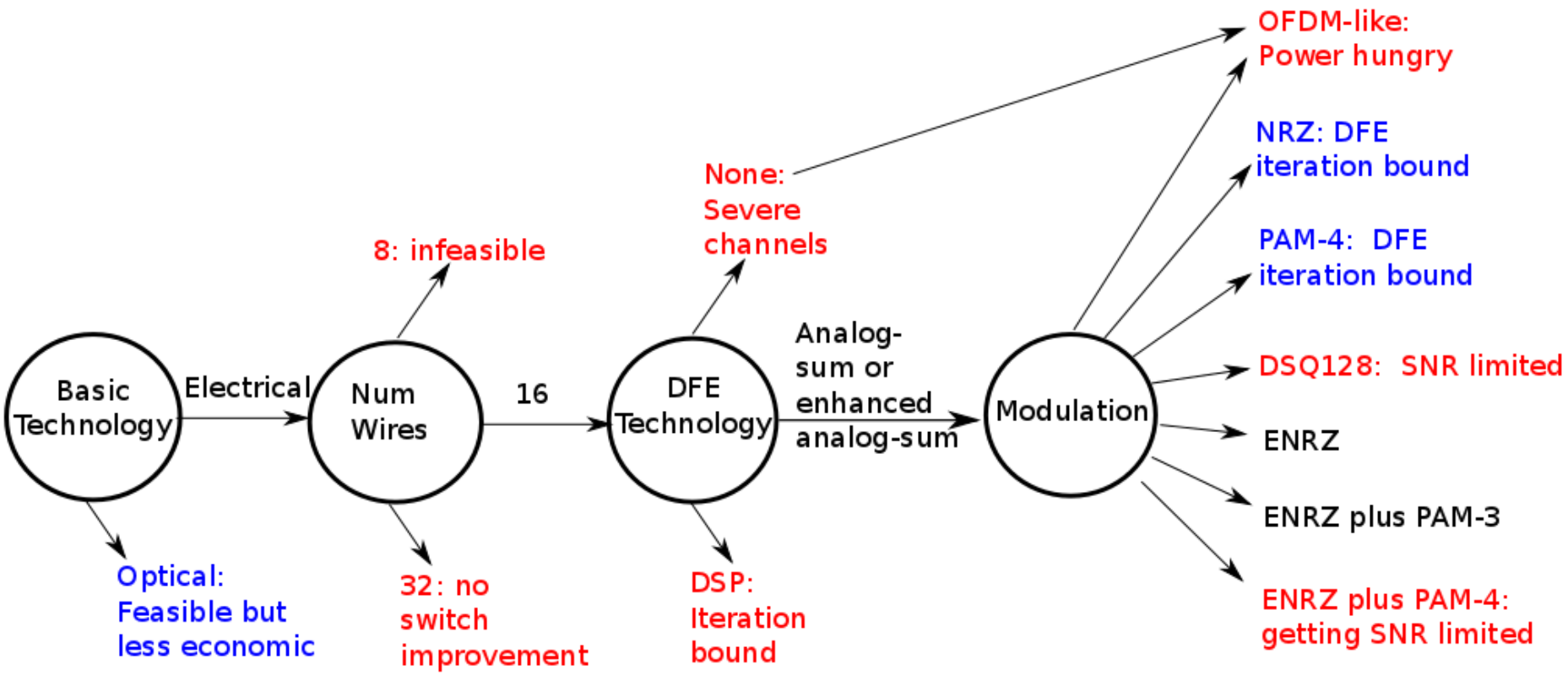
- ENRZ plus PAM-3
  - Ensemble NRZ can be combined with PAM-3
    - The Hadamard transform is fully linear
    - Combination would run at 22 Gbaud
    - With a 1 volt total swing, the effective sub-channel swing would be 250mv
  - The required DFEs are easier than those needed for PAM-4
    - For each speculative sample needed in order to close the loop, 3 states are needed. If more than one sample is needed, a 3 to the N state explosion happens in the inner loop

Modulation Choices cont. (16-wire 400GE)

- ENRZ plus PAM-4
  - Ensemble NRZ can be combined with PAM-4
    - The Hadamard transform is fully linear
    - Combination would run at 16.7 Gbaud
    - With a 1 volt total swing, the effective sub-channel swing would be 250mv
  - The required DFEs still look challenging
    - For each speculative sample needed in order to close the loop, 4 states are needed. If more than one sample is needed, a 4 to the N state explosion happens in the inner loop
    - DSP DFEs are probably still iteration-bound at 16.7 GBaud

Review of decision tree for the options for 400GE backplanes

- **Below is a decision tree for 400 GE backplanes that this presentation has reviewed in detail – the two ENRZ options are competitive**



PCS layer requirements For the use of ENRZ

- Since this proposed use of Ensemble NRZ delivers three phase-associated 33 Gb/s channels, the PCS requirements for a sixteen wire 400GE backplane are that of four 100 Gb/s serial lanes.
  - There is no need to have an additional PCS multiplexing layer to de-skew and align the three 33 Gb/s channels
  - The linear H4 transform that ENRZ is based on delivers the data with no time ambiguity between the channels
- The three channels act as if they were a single 100 Gb/s serial link for alignment purposes

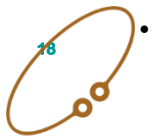
PCS interworking with 25 Gb/s optics

- It seems likely that 25 or 50 Gb/s optics will be used to support 400 GE channels
  - For these two cases, a 16-lane or 8-lane PCS layer would be used (16 x 25 or 8 x 50)
- Since a 3x33 Gb/s ENRZ link acts as if it were a single 100 Gb/s link, an additional PCS framing layer would be required after the ENRZ link to find the correct 25 Gb/s lanes for each of the optical links
  - Multiplexing layers are also needed in many other cases like this including using 50 Gb/s optics with a 16-lane PCS layer

KANDOU BUS

Conclusions

- ENRZ encoding may be a good choice to achieve technical feasibility for the possible objective of 16-wire 400GE backplane links
  - ENRZ reduces the excessive link rate needed by NRZ links, allowing their construction with existing and well-understood implementation techniques
  - ENRZ employs a relatively simple circuit
  - ENRZ relies on a low-skew four wire ensemble
  - ENRZ's linear nature allows implementation flexibility
  - ENRZ can be combined with PAM-3 or PAM-4
  - No ENRZ-specific PCS layer is required.
- The use of Ensemble NRZ saves power over NRZ
- Ensemble NRZ supports channels with frequency impairments similar to what NRZ requires
  - Ensemble NRZ is much more like NRZ than PAM solutions are
- Ensemble NRZ allows binary DFEs to be used
  - Ensemble NRZ can use ordinary binary DFE circuits if the differences in the reflections are constrained through a specification



KANDOU BUS