PCIe® 3.0 PHY Logical Layer Requirements

Debendra Das Sharma

Member, PWG and EWG

Agenda

- Problem Statement
- Existing Usage of K-Codes
- Metrics used for evaluation
- Current Direction on Encoding
- Summary & Call to Action

Disclaimer: Information contained here is **Preliminary/ Work In Progress**.
We are working towards Rev 0.3 draft.

Problem Statement

- PCIe® 3.0 data rate decision: 8 GT/s
  - ✓ Worst case HVM (High Volume Manufacturing) channel for client/ servers
    - – Same channels and length – backwards compatibility
  - ✓ Low power and ease of design
    - – Avoid using complicated receiver equalization, etc.
- Requirement: Double Bandwidth from PCIe 2.0
  - ✓ PCIe 1.0a data rate: 2.5 GT/s
  - ✓ PCIe 2.0 data rate: 5 GT/s
    - – Doubled the bandwidth from Gen 1 to Gen 2 by doubling the data rate
  - ✓ Data rate gives us a 60% boost in bandwidth
- Rest will come from Encoding
  - ✓ Replace 8b/10b encoding with a scrambling-only encoding scheme when operating at Gen3 data rate

Problem Statement

- Scrambling-only yields 25% BW improvement over 8b/10b coding
  - ✓ Allows 8G to have 2x BW of 5G with 8b/10b
    - Data Rate boost: 1.6
    - Encoding Efficiency Improvement: 1.25
    - Total B/W improvement: 1.6 x 1.25 = 2.0
- Phy layer impact
  - ✓ Transition density, run length, and DC balance due to lack of 8b/10b
- Logical Phy layer impact
  - ✓ Lack of distinct escape codes to delineate inter, intra packet boundaries

Agenda

- Problem Statement
- Existing Usage of K-Codes
- Metrics used for evaluation
- Current Direction on Encoding
- Summary & Call to Action

Existing Usage of K-Codes

- Two flavors for K-code use
  - ✓ Packet Stream (independent of link width)
  - ✓ Lane Stream (per-lane)
- Packet Stream relates to Packet Framing (Link-Wide)
  - ✓ STP - Start of TLP
  - ✓ END - End (Good) of TLP
  - ✓ EDB - End Bad of TLP
  - ✓ SDP - Start of DLLP
- Lane Stream relates to Ordered Sets:
  - ✓ Training Set #1 & #2
    - – Link training and negotiation
  - ✓ SKP Ordered Sets
    - – Periodic link clock compensation
    - – Recovery from bit slip/add
  - ✓ Electrical Idle Start/ Exit sequence
    - – Power management
- New encoding scheme needs to accommodate these existing usages

Agenda

- Problem Statement
- Existing Usage of K-Codes
- Metrics used for evaluation
- Current Direction on Encoding
- Summary & Call to Action

Error Detection Ability

- Robustness against bit errors considered
  - ✓ Bit flip, bit slip/add
- **Basic Fault Model:**
  - ✓ Guaranteed error detection against random bit flips in any TLP or DLLP or IDL or Ordered Set
    - – Must not alias to a TLP or a DLLP
      - • Can cause data corruption or flow-control problems
- Consider burst errors (e.g., receivers with DFE)
  - ✓ A bit flip can impact up to a certain number of consecutive bits on the same lane
- No guaranteed detection of error with bit slip/add
  - ✓ Same as 2.0 ability
- Eventual guaranteed recovery in the presence of multiple errors above including bit slip/add

Error Detection and Recovery

- Error Detection Ability of existing CRC schemes
  - ✓ TLP CRC guarantees 3-bit (flip) error detection
  - ✓ DLLP CRC guarantees 4-bit (flip) error detection
- Preference would be to guarantee error detection in the presence of random three bit flips in any TLP/ DLLP/ IDL/ Ordered Set
  - ✓ Should not alias to a valid TLP or DLLP
- Recovery from errors
  - ✓ No need to do self-healing
    - Can just go to Recovery for any error
    - Loss of self-healing ability from 2.0
    - Considered reasonable given that a 10^{-12} BER translates to an entry to Recovery every hour (typically 2-4 usec)
- Need to handle killer packets
  - ✓ Send a different bit stream on retry of a packet

Other Metrics

- Bandwidth Inefficiency must be low enough
  - ✓ 8b/10b had a 20% inefficiency
  - ✓ New scheme must be in the 1-2% range for inefficiency
- Time Overhead through Recovery as well as L0s/L1 exit must be minimal
  - ✓ Enables better power management without performance penalty
- Bytes must be the unit of transmission
  - ✓ Enables single-wide/double-wide type of parallel implementation
    - E.g., no end TLP in bit 3 and a new TLP starts in bit 4 within a byte
  - ✓ Prefer preserving as much framing rules as feasible
- Link transmitting idle must have scrambled Bytes on all lanes irrespective of width
- Should switch to new encoding after speed change from electrical idle
  - ✓ Avoids any extra synchronization
- Minimal changes beyond PHY layer
  - ✓ Ease of implementation

Approaches Considered

- Multiple proposals evaluated against metrics
- Choice of Scrambler:
  - ✓ Self synchronizing vs Additive
  - ✓ Link-wide scrambler vs per-lane scrambler
  - ✓ Error Multiplication in self synchronizing evaluated against the CRC for all possible link widths
- Phy layer packetization vs Substiution
  - ✓ Packetization for TLP/ DLLP/ IDL/ OS
  - ✓ Substitution uses defined bit patterns for K-codes and adopts an escape mechanism when data aliases to K-codes
- Lane level encoding for Phy layer packetization
  - ✓ 64/66 vs 128/130
- Current direction chosen based on the evaluation criteria

Agenda

- Problem Statement
- Existing Usage of K-Codes
- Metrics used for evaluation
- Current Direction on Encoding
- Summary & Call to Action

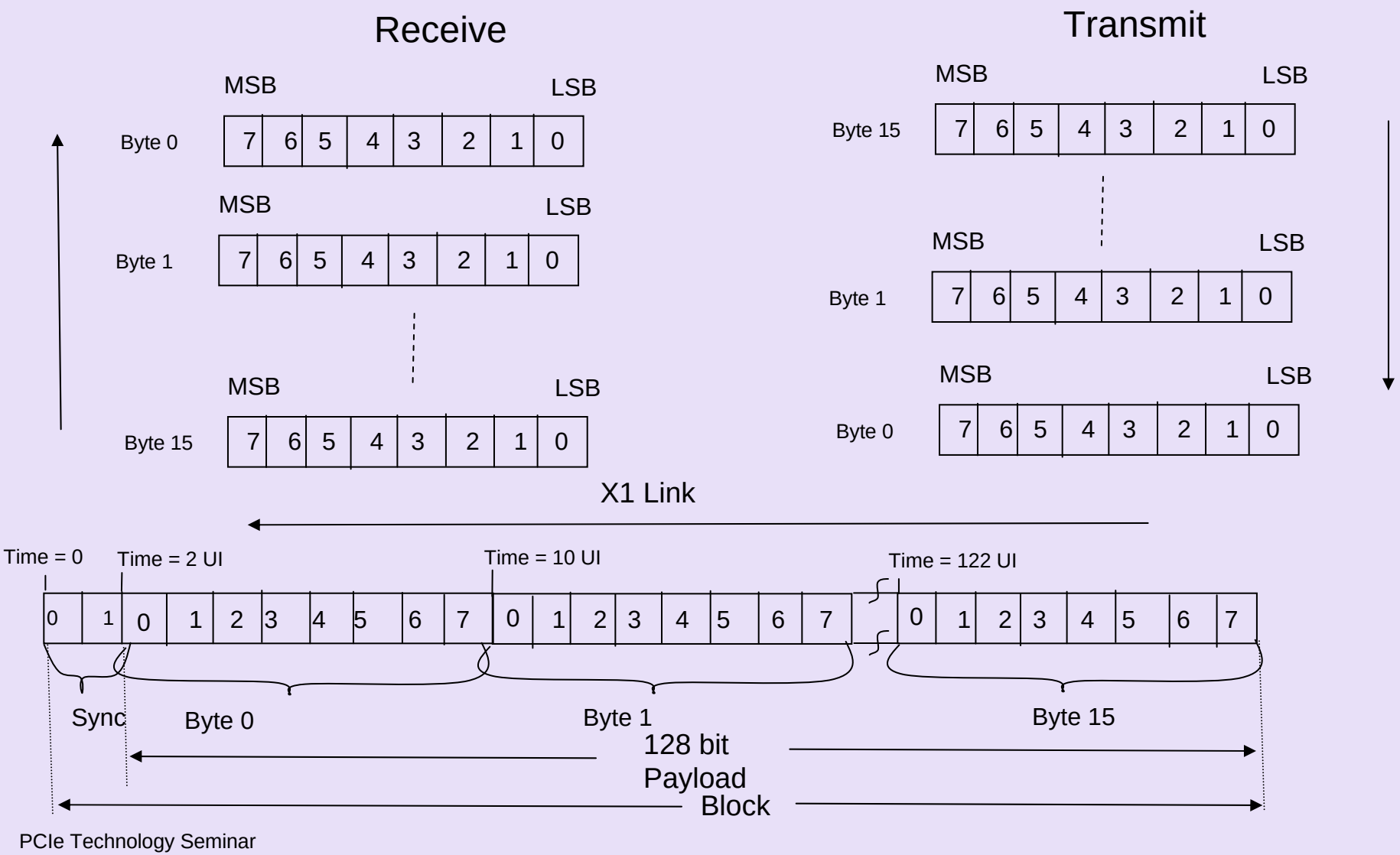
Overall Scheme

- Current direction is to use two levels of encapsulation
  - ✓ A 128/130 code on individual lanes
  - ✓ Physical layer packetization to identify “packet” boundaries
- Lane Level 128/130 Code
  - ✓ 130 bit code called a Block
  - ✓ Used for block lock
    - – Substitutes COM used for Byte lock in 8b/10b
  - ✓ Differentiate certain packet types
- Physical Layer packetization identifies packet boundaries. Packet types:
  - ✓ Link Level (TLP or DLLP)
  - ✓ Lane Level (Ordered Sets or IDL)
- Scrambling only (no 8b/10b) to provide edge density
  - ✓ Scrambling done on a per-lane basis
  - ✓ Additive scrambler

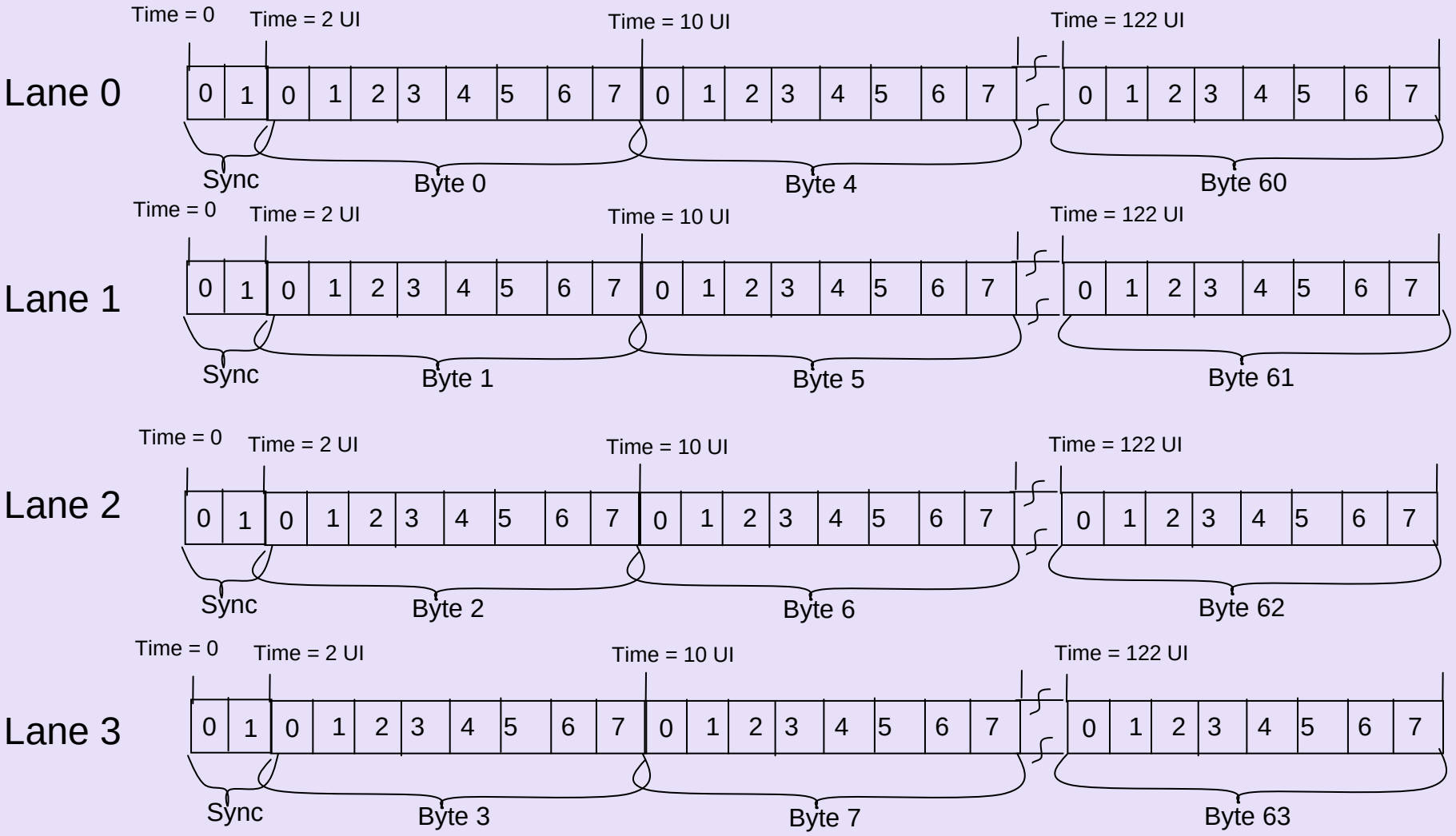
Lane Level Encoding

- Lane Level Encoding with a 128/130 code
  - ✓ 2-bit sync character followed by 128 bit payload
    - – 128 bit payload may be any combination of
      - Part of one or more Link level packets (TLP, DLLP)
      - One or more lane level packets (IDL, Ordered Sets)
    - – 2-bit sync character not scrambled
    - – 128 bit payload may or may not be scrambled
  - ✓ 10_ sync character used for certain Ordered Sets
    - – EIEOS not scrambled
      - 10_ [<00000000>_<11111111> ..8 times] represents EIEOS
      - Used for the dual purpose of low frequency patterns (in Recovery and Configuration) as well as establishing block lock
  - ✓ 01_ sync character is used when 128 bit information contains IDL, TLP, or DLLP
    - – Each bit in the 128 bit payload is scrambled
  - ✓ Alternatives for other Ordered Set encodings considered

Mapping of bits on a x1 Link



Mapping of bits on a x4 Link



Scrambler

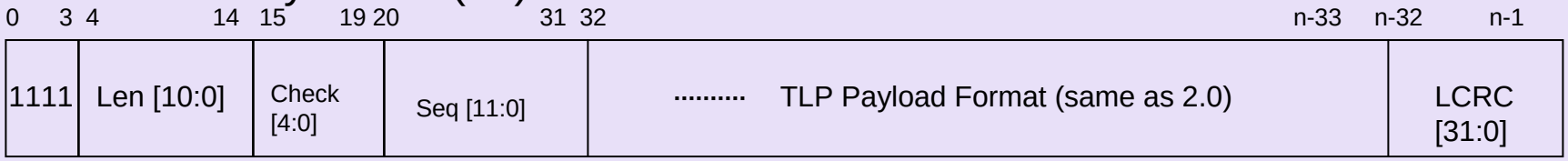
- Scrambler (Descrambler)
  - ✓ Lane level
  - ✓ Additive
    - LFSR not impacted by the incoming (outgoing) bit value
  - ✓ Degree 23 primitive polynomial (e.g., x23 + x5 + 1)
  - ✓ Reset on transmission (or receipt) of EIEOS
    - In Recovery/ Config
  - ✓ Advances with every bit in the 128-bit payload of the 128/130 code
    - Does not advance for the 2-bit sync character

Physical Layer Encapsulation

- First Byte (scrambled) indicates packet type:
  - ✓ 00000000 is Logical IDL
    - All subsequent lanes in same Byte-time should be IDL
    - Receivers check for all 0s (after descrambling) in IDL
    - PAD is identical to IDL
  - ✓ 1111xxxx is STP
    - Subsequent 11 bits (link wide) define the length
  - ✓ 00001111 is SDP
    - 2nd Byte also gets a fixed encoding
  - ✓ 00000011 is EDB
    - EDB packet is 4 bytes; each with the same value 00000011

P-Layer Encapsulation: TLP

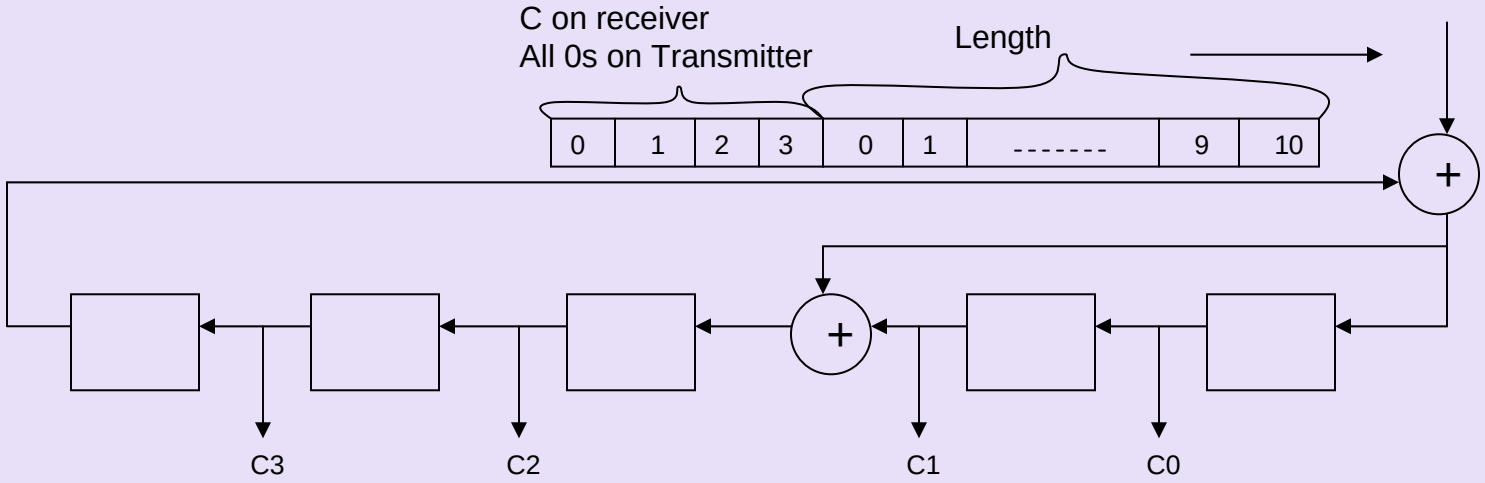
- Length known from the first 3 bytes
  - ✓ First 4 bits are 1111 (bit[0:3] = 4'b1111)
  - ✓ Bits 4:14 has the length of the TLP (valid values: 5 to 1031)
  - ✓ Bits 15:19 is check bits to cover the TLP Length field
    - – Primitive Polynomial (X4 + X + 1) protects 15 bit field
      - • Provides double bit flip detection guarantee (length 11 bits + CRC 4 bits)
    - – Odd parity covers the 15 bits (length 11 bits + CRC 4 bits)
      - • Guaranteed detection of triple bit errors (over 16 bits)
  - ✓ Sequence Number occupies bits 20:31
  - ✓ TLP payload is from the 4th byte position (same as 2.0)
  - ✓ No explicit END. Need to check first Byte after TLP for implicit END vs an explicit EDB => Ensures triple bit flip detection
  - ✓ All bytes are (de)scrambled



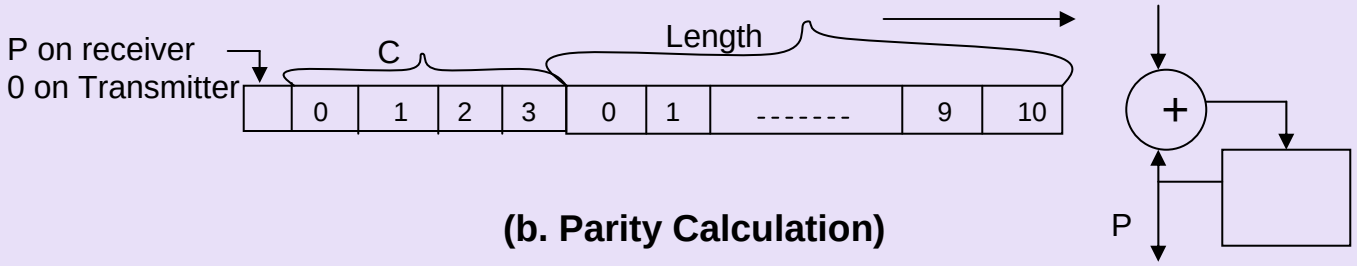
(TLP Layout)

[Len[10:0]: length of the TLP in DWs, Check[4:0]: Check Bits, 18:4, No END]

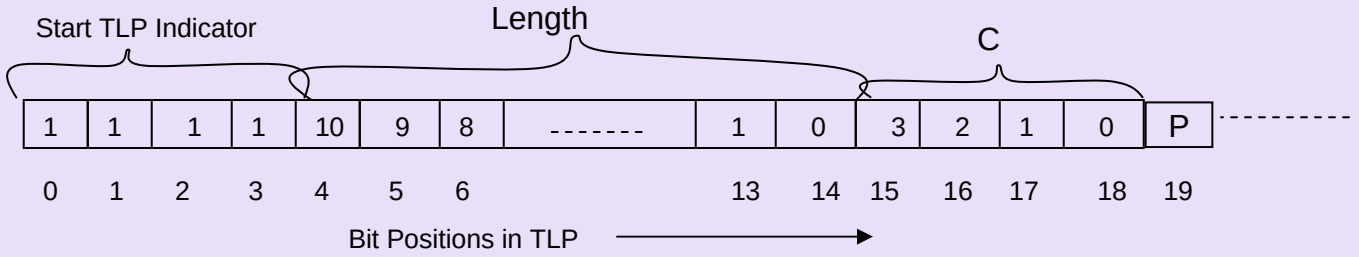
TLP Length Protection



(a. CRC Check Bit Calculation with x4+x+1)



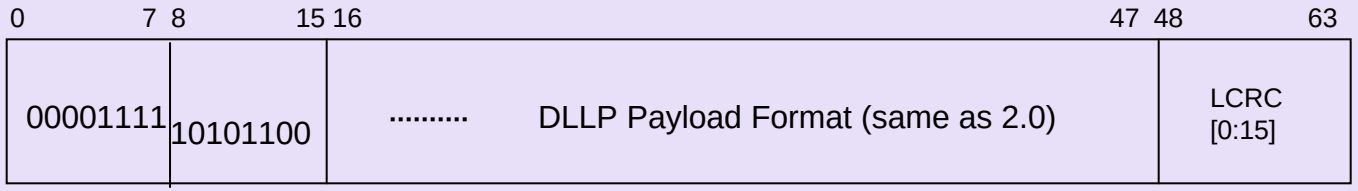
(b. Parity Calculation)



(c. Layout of bits in Packet)

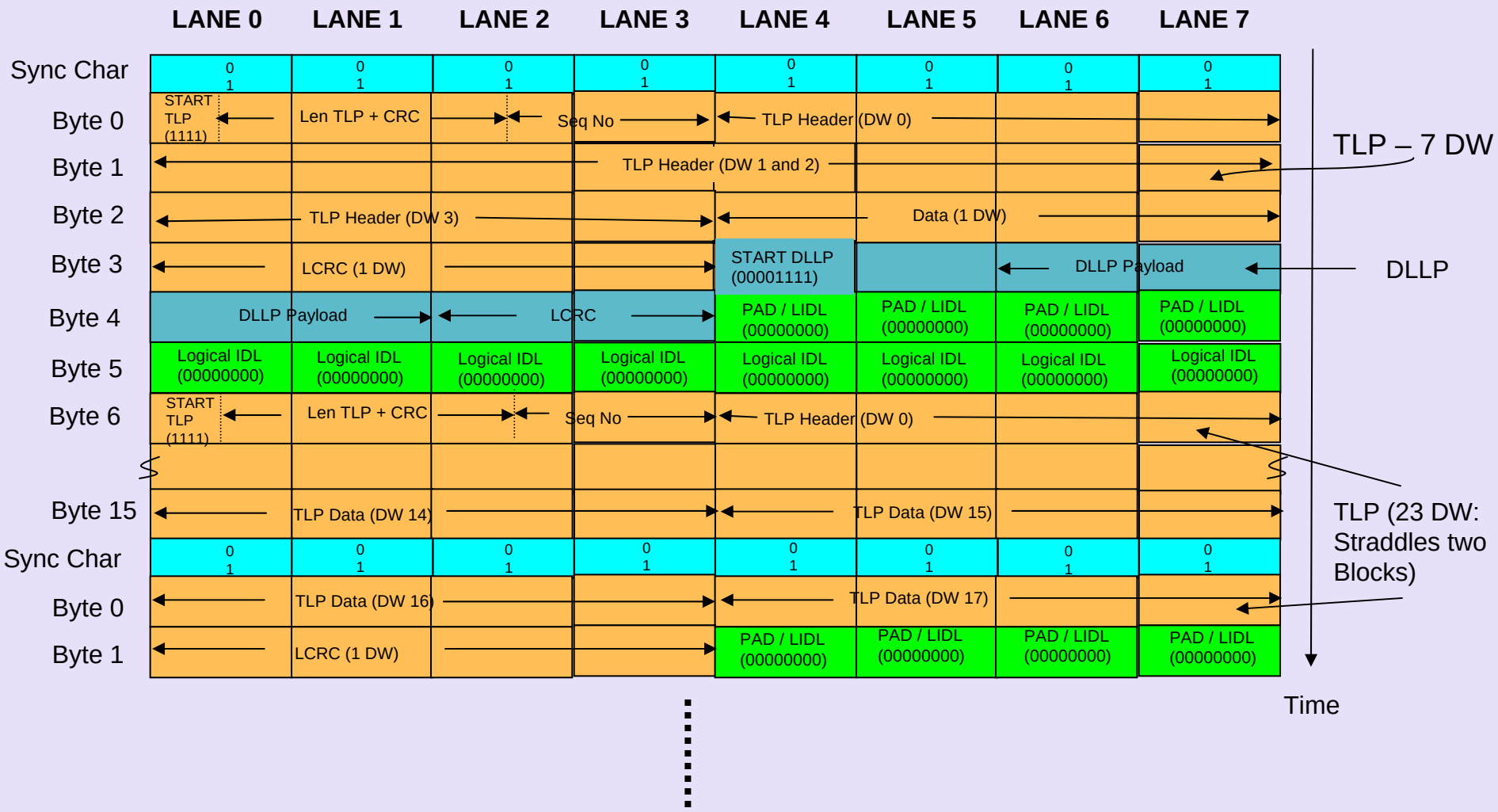
P-Layer Encapsulation: DLLP

- Preserve DLLP layout of 2.0 spec
- First byte is 0000_1111
- Second byte another unique value
  - ✓ Will allow to share encoding with some Ordered Sets if needed
- Next 4 bytes (2 through 5) are the DLLP layout
- Next 2 bytes (6 and 7): LCRC (identical to 2.0)
- No explicit END
- All bytes are (de)scrambled



(DLLP Layout)

Ex: TLP/ DLLP/ IDLs in x8



Error Detection and Recovery

- Framing error
  - ✓ The first byte is not one of the three allowed sets
  - ✓ After block lock, if the sync character is not 01 or 10
- Any error requires directing LTSSM to Recovery
  - ✓ CRC error or framing error
  - ✓ Stop processing any received TLP/ DLLP after error until we get through Recovery
  - ✓ Block lock acquired with EIEOS
  - ✓ Scrambler reset with each EIEOS
- Error Detection Guarantees
  - ✓ Triple bit flip detection within each TLP/ DLLP/ IDL/ OS
- Killer Packets: In Recovery.Idle, mandate a variable number of IDL bytes so that the same TLP retransmitted immediately after Recovery does not come out with the same bit pattern

Lane to Lane Deskew

- No explicit SKP OS need to be sent for clock compensation
  - ✓ SKP Ordered Sets for retiming repeaters under discussion
- Lane to Lane Deskew:
  - ✓ Lane to lane deskew in L0s done using special deskew patterns
  - ✓ In Recovery or Polling/ Config this can be done using TS ordered sets or EIEOS
- Clock Compensation:
  - ✓ Can be done prior to elastic buffer by stripping out the 2 bits in the (130, 128) code
  - ✓ 1.5% bandwidth loss does pay off to some extent (about 0.4%) in savings of the SKP OS (4 symbols every 1150 symbols)
  - ✓ Designs expected to be able to deal with the output of elastic buffer not having a Byte ready every cycle
- Loopback Solution under discussion

Agenda

- Problem Statement
- Existing Usage of K-Codes
- Metrics used for evaluation
- Current Direction on Encoding
- Summary & Call to Action

Summary & Call to Action

- Encoding scheme decided and development in progress
- Offers advantage of 25% bandwidth for 8GT/s (and above) data rate
- Rev 0.3 spec development in progress
- Track the spec development and plan for products accordingly

Thank you for attending the
PCIe Technology Seminar

For more information please go to
www.pcisig.com


PCI
SIG®