PCI Express* 3.0 Technology: PHY Implementation Considerations for Intel Platforms

Debendra Das Sharma

Principal Engineer

Digital Enterprise Group

Intel Corporation

TCIS007

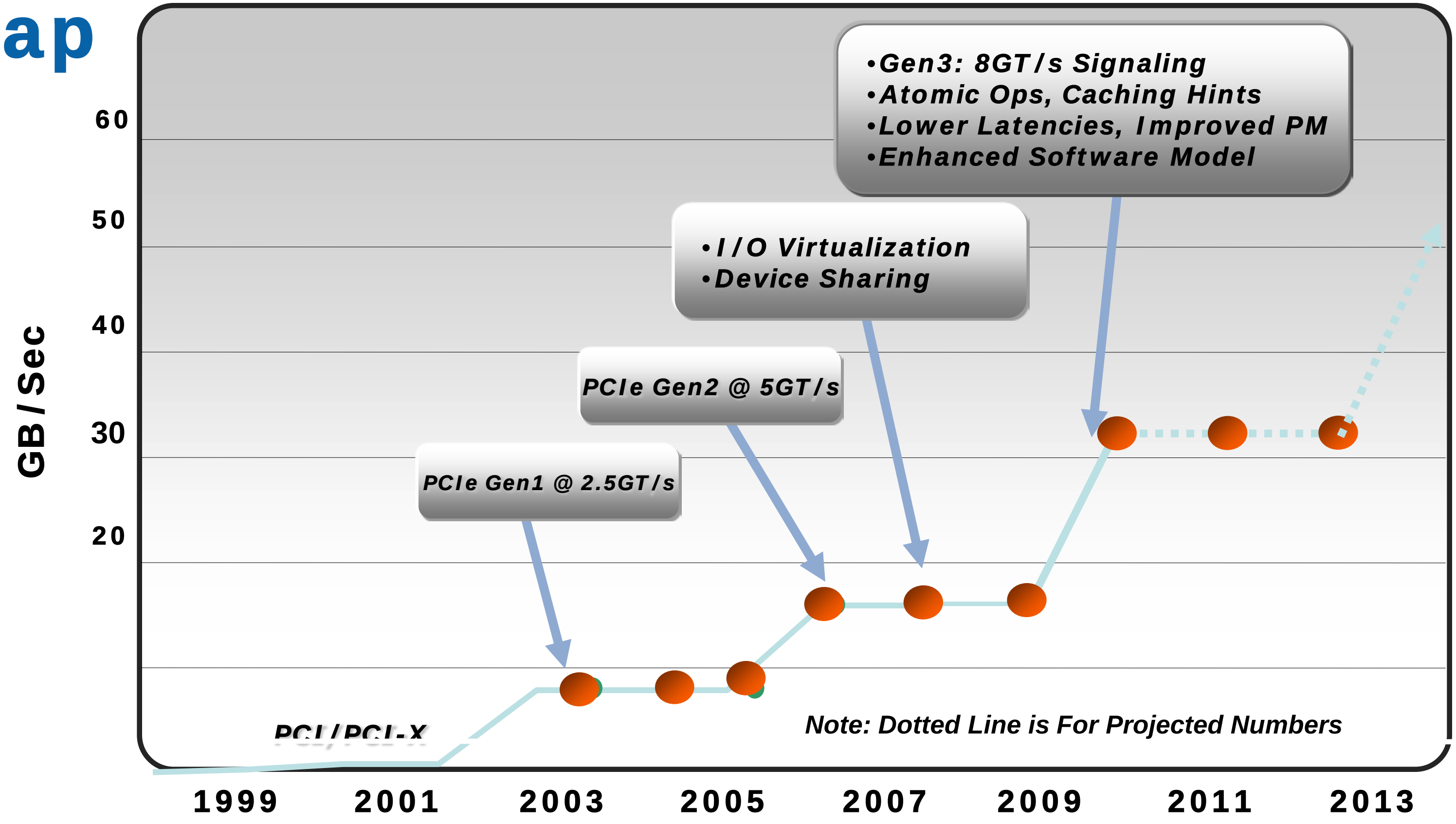
Sponsors of Tomorrow.™ intel

Agenda

- Problem Statement
- Existing Usage of K-Code in 8b/10b
- Encoding Scheme
- Transmitter Equalization and training
- Implementation Considerations
- Summary

Disclaimer: Information contained herein is derived from Intel technology path finding and is **Work In Progress** and is subject to change

PCI Express* (PCIe*) Technology Roadmap



- Gen3: 8GT/s Signaling
- Atomic Ops, Caching Hints
- Lower Latencies, Improved PM
- Enhanced Software Model

- I/O Virtualization
- Device Sharing

PCIe Gen2 @ 5GT/s

PCIe Gen1 @ 2.5GT/s

Note: Dotted Line is For Projected Numbers

	Raw Bit Rate	Link BW	BW/lane /way	BW x16
PCIe 1.x	2.5GT/s	2Gb/s	~250MB/s	~8GB/s
PCIe 2.0	5.0GT/s	4Gb/s	~500MB/s	~16GB/s
PCIe 3.0	8.0GT/s	8Gb/s	~1GB/s	~32GB/s

Based on x16 PCIe channel

Continuous Improvement: Doubling Bandwidth & Improving Capabilities Every 3-4 Years!

Problem Statement

- **PCI Express* (PCIe*) 3.0 data rate decision: 8 GT/s**
  - High Volume Manufacturing channel for client/ servers
    - Same channels and length for backwards compatibility
    - Low power and ease of design
      - Avoid using complicated receiver equalization, etc.

- **Requirement: Double Bandwidth from Gen 2**
  - PCIe 1.0a data rate: 2.5 GT/s
  - PCIe 2.0 data rate: 5 GT/s
    - Doubled the data rate/ bandwidth from Gen 1 to Gen 2
  - 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 PCIe 3.0 data rate

- **Double B/W: Encoding efficiency 1.25 X data rate 1.6 = 2X**

Challenge: New Encoding Scheme to cover 256 data plus 12 K-codes with 8 bits

Agenda

- Problem Statement
- **Existing Usage of K-Code in 8b/10b**
- Encoding Scheme
- Transmitter Equalization and training
- Implementation Considerations
- Summary

Review of K-Code Usage

- **Each K-codes is a unique 10-bit value**
  - Distinct from data and other 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 Transaction Layer Packet (TLP)
  - END - End (Good) of TLP
  - EDB - End Bad of TLP
  - SDP - Start of Data Link Layer packet (DLLP)
- **Lane Stream relates to Ordered Sets:**
  - Training Set #1 & #2: Training/ retraining
  - SKP Ordered Sets: clock compensation and byte realignment
  - Electrical Idle Start/ Exit sequence: Power Management
- **New encoding scheme accommodates these existing usages**

Functionality of K-Code needs to be preserved

Agenda

- Problem Statement
- Existing Usage of K-Code in 8b/10b
- **Encoding Scheme**
- Transmitter Equalization and training
- Implementation Considerations
- Summary

Requirements and Capabilities

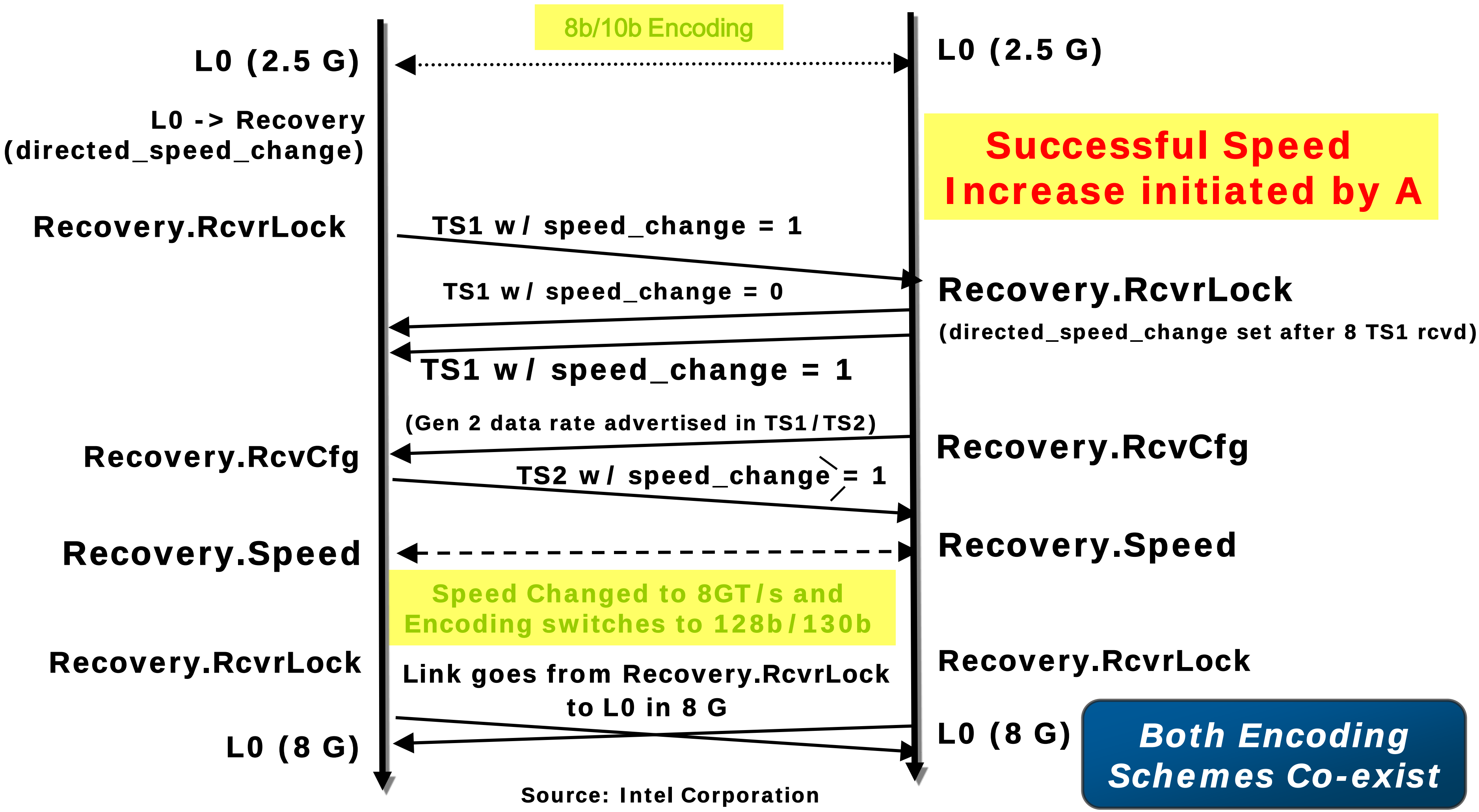
- **Basic Fault Model:**
  - Guaranteed error detection against random bit flips in any packet or Ordered Set
- **Eventual recovery from bit slip/add**
- **Handle killer packets**
  - Send a different bit stream on retry of a packet
- **Low bandwidth overhead (1-2%)**
- **Low L0s/L1 exit latency overhead**
  - Preserve aggressive power management with performance
- **Changes mostly limited to physical layer**
- **Protocol development concurrent with analysis / simulation done by Intel Pathfinding team**

New encoding scheme: Better Performance and Reliability than PCI Express* 2.0 Technology

LTSSM Speed Change: Example

LTSSM in Device A

LTSSM in Device B



8b/10b Encoding

L0 (2.5 G)

L0 (2.5 G)

L0 -> Recovery
(directed_speed_change)

Successful Speed Increase initiated by A

Recovery.RcvrLock

TS1 w/ speed_change = 1

TS1 w/ speed_change = 0

Recovery.RcvrLock

(directed_speed_change set after 8 TS1 rcvd)

TS1 w/ speed_change = 1

(Gen 2 data rate advertised in TS1/TS2)

Recovery.RcvCfg

Recovery.RcvCfg

TS2 w/ speed_change = 1

Recovery.Speed

Recovery.Speed

Speed Changed to 8GT/s and
Encoding switches to 128b/130b

Recovery.RcvrLock

Recovery.RcvrLock

Link goes from Recovery.RcvrLock
to L0 in 8 G

L0 (8 G)

L0 (8 G)

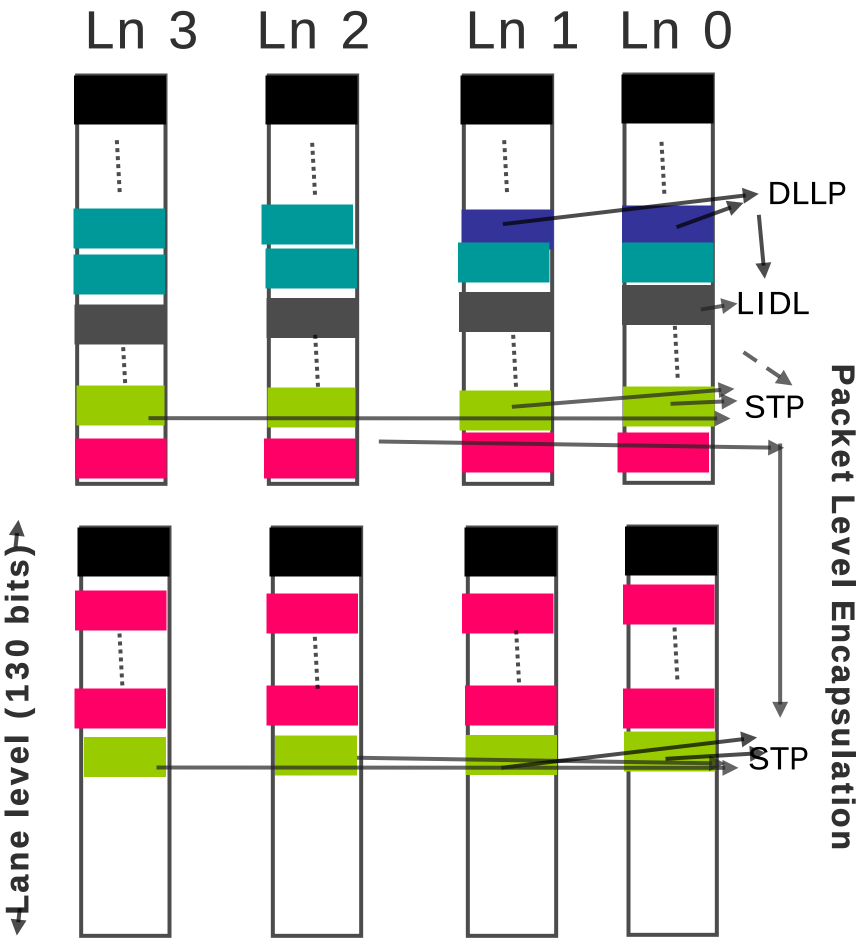
Source: Intel Corporation

Both Encoding Schemes Co-exist

128b/130b Encoding Scheme

Two levels of encapsulation

- Lane Level: Blocks
  - Data vs Ordered Sets
  - 2-bit Sync Header identifies Data Block vs OS (not scrambled)
  - 128-bit payload
  - Rationale:
    - Redundancy helps separate Data from OS
    - 128-bit payload chosen to match OS payload. 2-bit is low overhead for Sync header is low
- Data Block: Link wide with Framing preamble identifying packet boundary up-front
  - Multiple packets within a Data Block and one packet can straddle multiple Blocks
  - Framing preamble same overhead as in 8b/10b
  - Payload scrambled



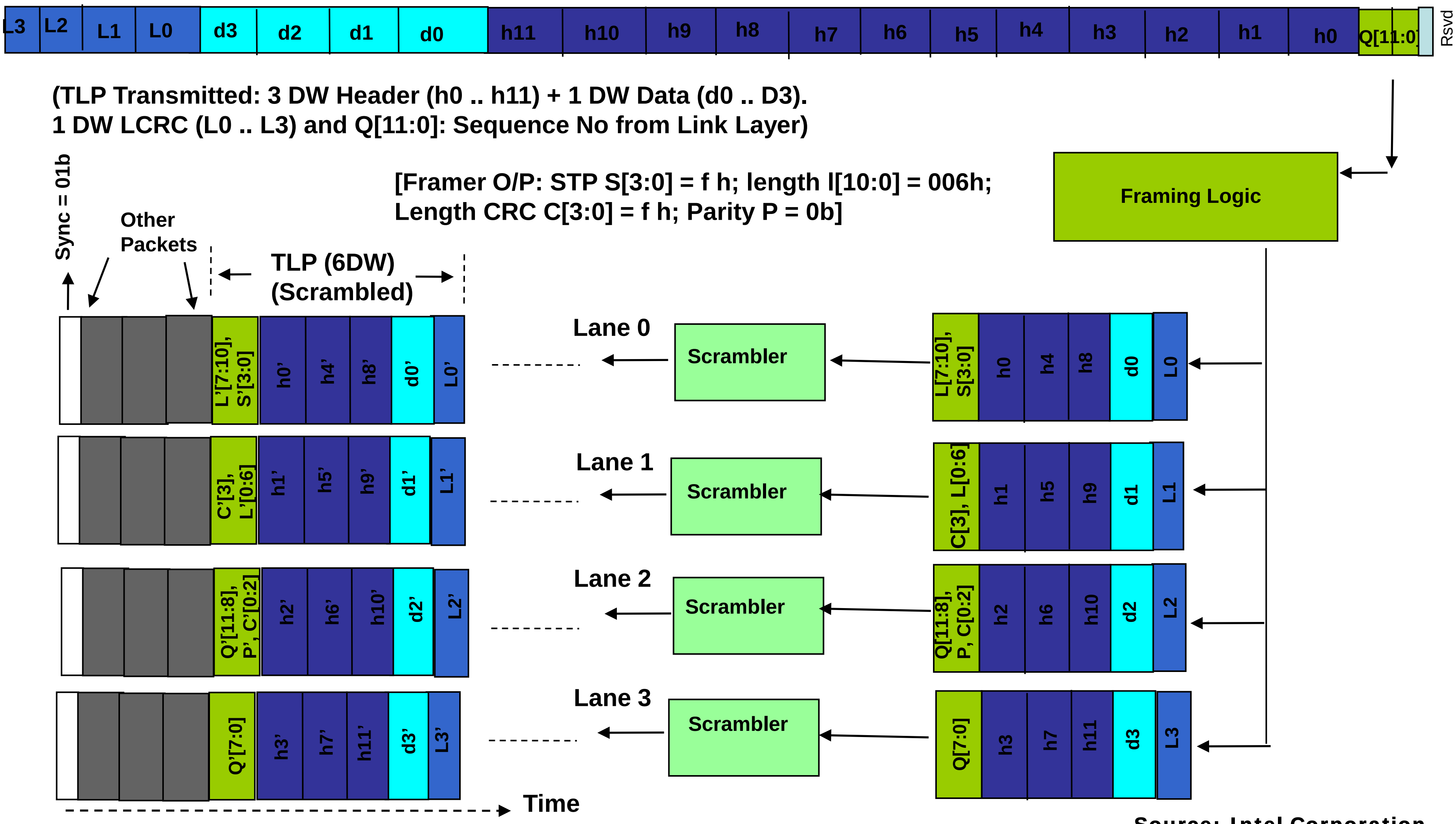
Source: Intel Corporation

Scrambling with two levels of encapsulation

Data Block

- Packets: Logical IDL (LIDL), DLLP, TLP, etc.
- Various sizes from 8b/10b time: 1 Symbol for LIDL, DLLP: 8 Symbols, TLP: Multiple of 4 Symbols
- Everything other than TLP is fixed size
- Need to ensure triple bit flip detection ability while keeping the sizes the same
  - New encoding: One Symbol becomes 8bit rather than 10bit in 8b/10b
- TLP and DLLP body is LCRC protected which provides triple bit flip detection ability
  - Framing preamble itself needs to ensure triple bit flip detection ability as it is used to determine packet boundary
- Challenge: Framing preamble itself has to be of variable length
- Solution: Use first Symbol encoding between different entities to be at a Hamming distance of 4 to ensure triple bit flip detection ability
  - Subsequent Symbols, if any, use some form of CRC protection itself
- Robustness features confirmed by analysis/ simulations in Intel path finding

Example of TLP Transmission in a X4



(TLP Transmitted: 3 DW Header (h0 .. h11) + 1 DW Data (d0 .. D3).
1 DW LCRC (L0 .. L3) and Q[11:0]: Sequence No from Link Layer)

[Framer O/P: STP S[3:0] = f h; length l[10:0] = 006h;
Length CRC C[3:0] = f h; Parity P = 0b]

Source: Intel Corporation

IDF2009
INTEL DEVELOPER FORUM

Ordered Sets

- Used for Link training, power management, clock compensation
- Additional usage: Block alignment
  - Can not scrambled
  - Must reset scrambler (so that both sides start at same point)
  - Can not do this during Data Blocks
    - Unlike 8b/10b where COM can never be aliased to between any two 10-bit Symbols, one can easily alias to any bit pattern when two Data Blocks are looked at consecutively for a fixed pattern
  - Must be done when all permutations are not possible (i.e., during Recovery)
  - Choice of Ordered Set encoding to be such that one can always correctly do block alignment
  - Another challenge is bit slip initially, based on past observation
  - Solution: Continuously do block alignment in Recovery while Ordered Sets are on
- Ordered Set for Clock compensation
  - Can not be scrambled (e.g., repeaters)
  - Need to carry information such as LFSR value to help trace tools
- Link training needs a spectrally rich pattern for better bit lock
- Solution: Most of TS1/TS2 are scrambled; Rest are not
- Problem scenarios and their solutions created and verified in analysis / simulation in Intel pathfinding

Sync Header Protection

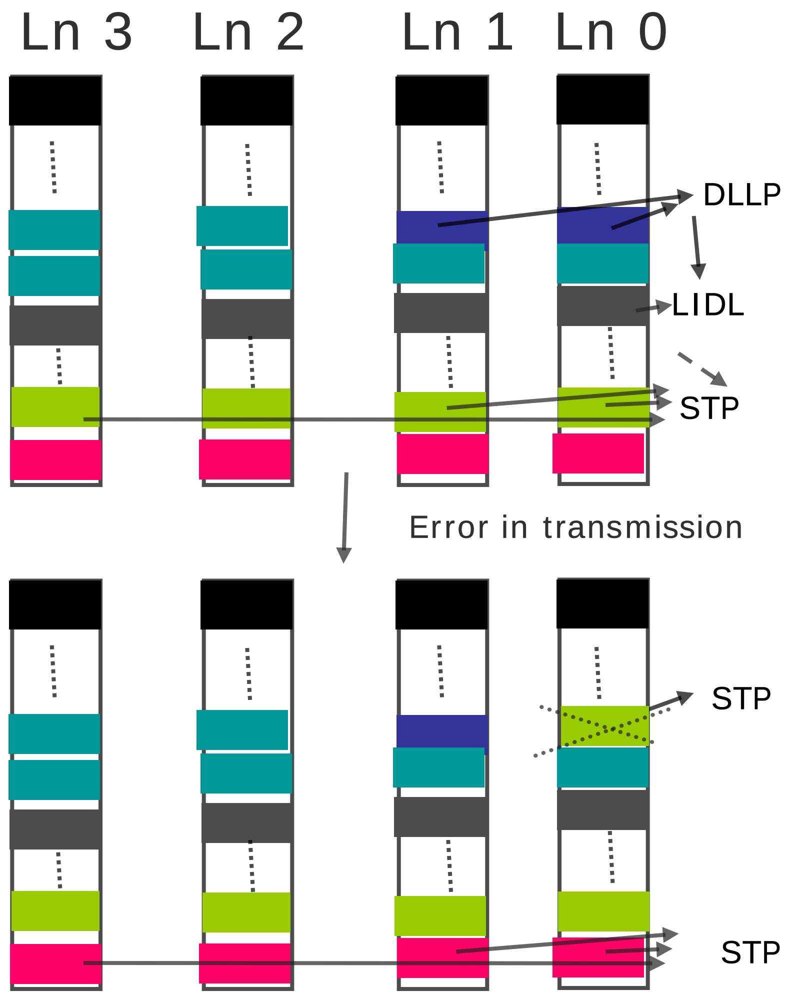
- Challenge: Sync Header is 2-bits: so even a two bit flip can potentially change a Data Block to an Ordered Set and vice-versa. Even worse the LFSR can be out of sync between the Tx and the Rx and introduce massive errors as a result and cause data corruption
- Solution: Define a “marker” of the appropriate type whenever there is a transition from Data Block to Ordered Set and vice-versa
  - Pre-notification of the Block type change – that itself is adequately protected
  - Data Block to OS: In last Data Block as a CRC-protected packet
  - OS to Data Block: As an additional 130-bit marker OS
  - Protects more than triple bit flip
  - Problem scenario as well as solution created and validated through analysis/ simulation (Intel path finding)

Error Recovery

- **Framing error detected by PHY**
  - Helps identify cases where the physical layer can either have its scrambler out of sync or fails to ascertain the next packet's framing preamble location

- **Any framing error directs LTSSM to Recovery**
  - Stop processing any received TLP/ DLLP after Recovery to avoid data corruption
    - The CRC within these packets become ineffective when the packet boundary is lost – random data can always alias to a good CRC
  - Block lock and scrambler reset happens through Recovery prior to packet being accepted
  - Link layer detected errors can be recovered through packet retry

- **Error Detection Guarantees maintained**
  - Triple bit flip detection within each TLP/ DLLP/ IDL/ OS



(Error causes a DLLP to alias to a TLP;

Can cause massive errors after that)

Source: Intel Corporation

Robust Error Detection and Recovery Mechanism required with PHY Framer

Agenda

- Problem Statement
- Existing Usage of K-Code in 8b/10b
- Encoding Scheme
- **Transmitter Equalization and training**
- Implementation Considerations
- Summary

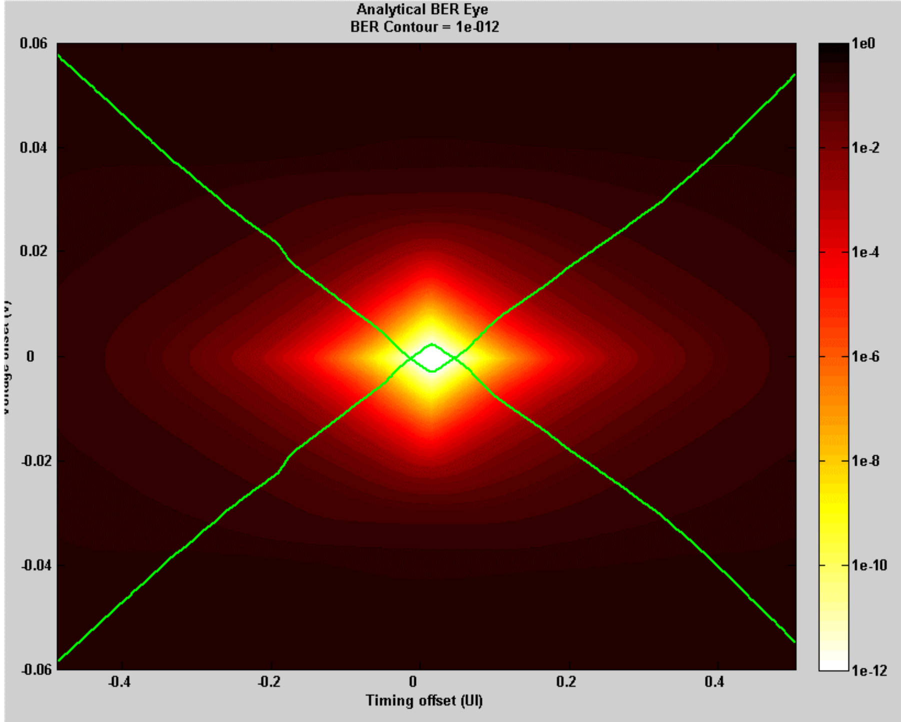
Transmitter Equalization

- 2.5 GT/s : Same de-emphasis for all
- 5 GT/s: Introduced platform dependent de-emphasis selection on a per-Link basis
  - -3.5 dB and -6 dB
- 8 GT/s: Our analysis shows that a static selection does not work for all channels due to variations in the receiver design, channel, as well as PVT
- Solution: Need to adjust each transmitter at the by its corresponding receiver in a fine-grain fashion (coefficients)
  - Need to do it once and store it for use on every entry to 8GT/s
  - Must start with some predefined value set by platform characteristics
  - Dynamic adjustment after that
  - Our analysis shows this approach results in working silicon

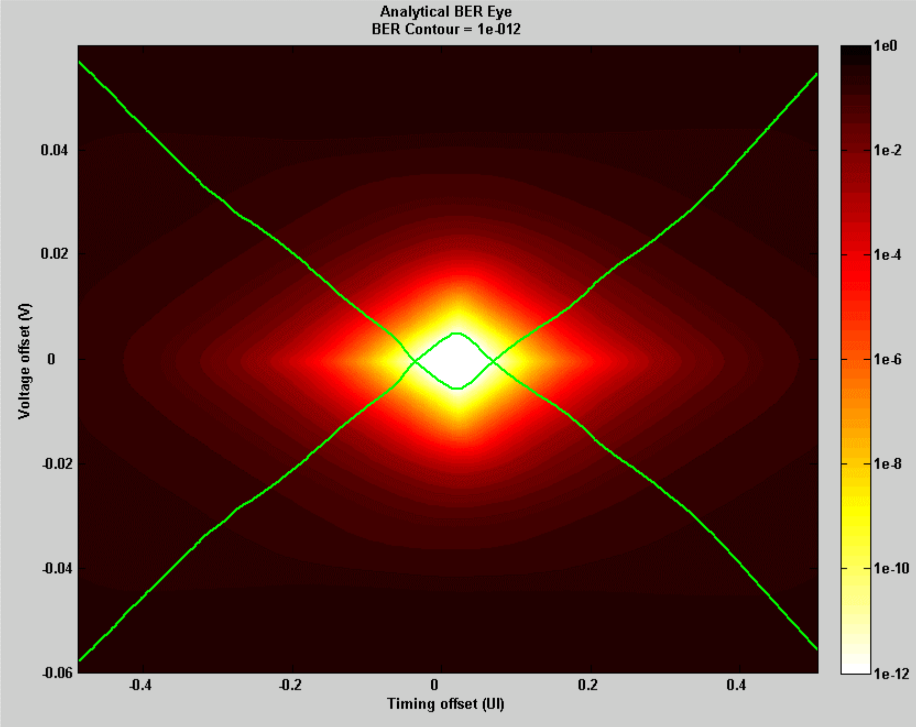
Tx EQ coefficient Optimization vs. Pre-set example

- The eye diagram on the left was the result of using the best pre-set Tx EQ values.
- The eye diagram on the right was same channel with optimized Tx EQ coefficients.
- The green contour shows the BER eye at 1e-12.
- Eye width opening increased from 7ps to 16ps (over 50% more Eye Width)
  - Both assumed a Tx EQ step size resolution of 1/32
  - Channel: 2 connector topology 18” pin-pin
  - Both used same Rx EQ that was re-optimized for each case.

BER Eye With
Best Pre-Set



BER Eye with
optimized Tx coef

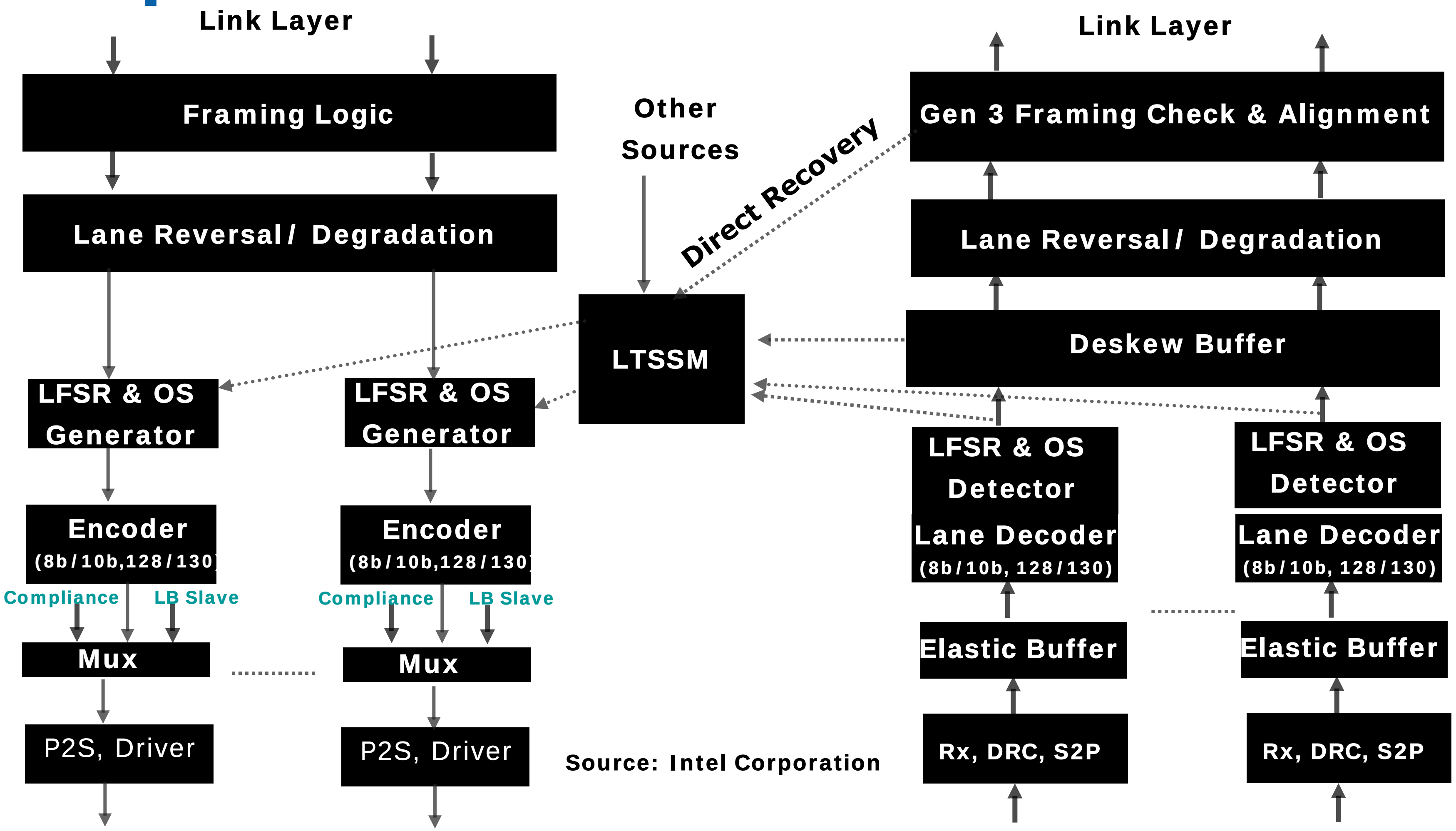


Source: Intel Corporation

Agenda

- Problem Statement
- Existing Usage of K-Code in 8b/10b
- Encoding Scheme
- Transmitter Equalization and training
- **Implementation Considerations**
- Summary

Sample Transmitter and Receiver



Source: Intel Corporation

Considerations: 1 byte offset to Link Layer for TLPs (EDB). Seq # aligned to LCRC

Agenda

- Problem Statement
- Existing Usage of K-Code in 8b/10b
- Encoding Scheme
- Transmitter Equalization and training
- Implementation Considerations
- **Summary**

Summary & Call to Action

- Overview of Logical PHY based on Intel analysis, simulations, and experience
- 128b/130b Encoding definition
- Equalization mechanism needed
- 25% bandwidth advantage with new encoding over 8b/10b encoding with enhanced reliability
- Track the PCI Express* (PCIe*) 3.0 Spec development in the PCI-SIG and at www.pcisig.com
- Track the PCIe PIPE Spec development at www.intel.com/technology/pciexpress/devnet
  - Plan for products accordingly

Additional Sources of Information on This Topic

- Other Sessions / Chalk Talks / Labs:
  - **TCIQ002** Q&A: PCI Express* 3.0 Technology
  - **TCIS006** PCI Express* 3.0 Technology: Device Architecture optimizations on Intel Platforms
  - **TCIS008** Electrical requirements for designing PCIe* 3.0 ASICs on Intel platforms
  - **USBS002** USB 3.0 Architecture and PHY Interface (PIPE) Specification Updates

- Demo/Booths:
  - PCI Express* Technology Community

- Additional Web-based Info:
  - www.pcisig.com
  - www.intel.com/technology/pciexpress/devnet

Legal Disclaimer

- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.
- Intel, Intel Inside, and the Intel logo are trademarks of Intel Corporation in the United States and other countries.
- *Other names and brands may be claimed as the property of others.
-

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the third quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Many factors could affect Intel's actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the corporation's expectations. Ongoing uncertainty in global economic conditions pose a risk to the overall economy as consumers and businesses may defer purchases in response to tighter credit and negative financial news, which could negatively affect product demand and other related matters. Consequently, demand could be different from Intel's expectations due to factors including changes in business and economic conditions, including conditions in the credit market that could affect consumer confidence; customer acceptance of Intel's and competitors' products; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Additionally, Intel is in the process of transitioning to its next generation of products on 32nm process technology, and there could be execution issues associated with these changes, including product defects and errata along with lower than anticipated manufacturing yields. Revenue and the gross margin percentage are affected by the timing of new Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel's response to such actions; and Intel's ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on changes in revenue levels; capacity utilization; start-up costs, including costs associated with the new 32nm process technology; variations in inventory valuation, including variations related to the timing of qualifying products for sale; excess or obsolete inventory; product mix and pricing; manufacturing yields; changes in unit costs; impairments of long-lived assets, including manufacturing, assembly/test and intangible assets; and the timing and execution of the manufacturing ramp and associated costs. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. The current financial stress affecting the banking system and financial markets and the going concern threats to investment banks and other financial institutions have resulted in a tightening in the credit markets, a reduced level of liquidity in many financial markets, and heightened volatility in fixed income, credit and equity markets. There could be a number of follow-on effects from the credit crisis on Intel's business, including insolvency of key suppliers resulting in product delays; inability of customers to obtain credit to finance purchases of our products and/or customer insolvencies; counterparty failures negatively impacting our treasury operations; increased expense or inability to obtain short-term financing of Intel's operations from the issuance of commercial paper; and increased impairments from the inability of investee companies to obtain financing. The majority of our non-marketable equity investment portfolio balance is concentrated in companies in the flash memory market segment, and declines in this market segment or changes in management's plans with respect to our investments in this market segment could result in significant impairment charges, impacting restructuring charges as well as gains/losses on equity investments and interest and other. Intel's results could be impacted by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. A detailed discussion of these and other risk factors that could affect Intel's results is included in Intel's SEC filings, including the report on Form 10-Q for the quarter ended June 27, 2009.

Rev. 7/27/09