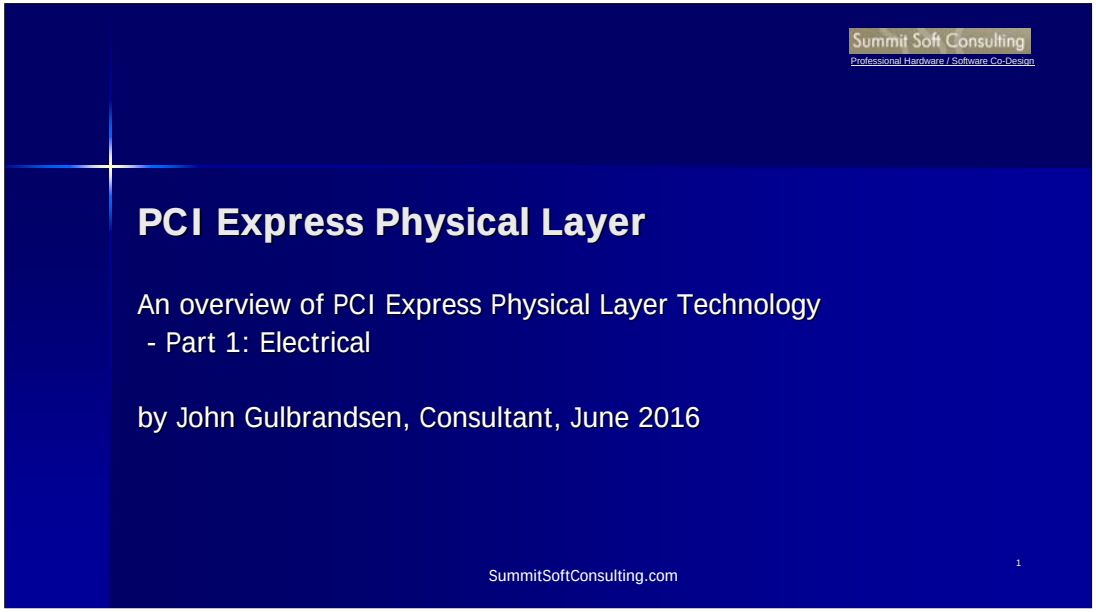
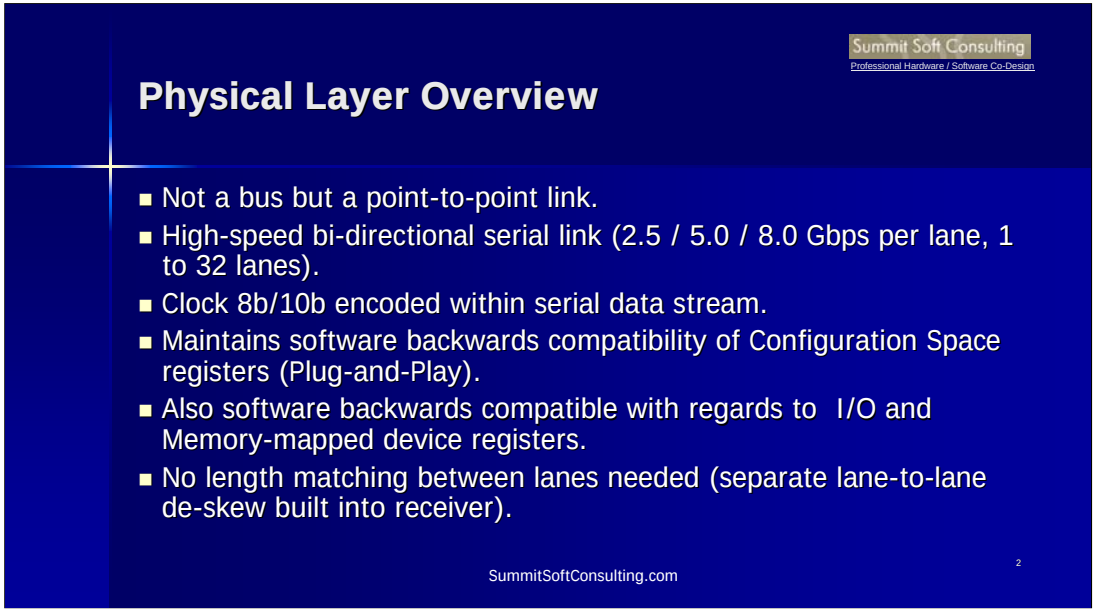


- I will in this presentation explain:
  - -The main problems with the legacy PCI bus that prompted the development of the new PCI Express architecture.
  - -An overview of the physical layer (hardware aspects) of the PCI Express bus and how it differs from PCI.
  - -Future presentations will cover higher protocol layers and the associated software features.



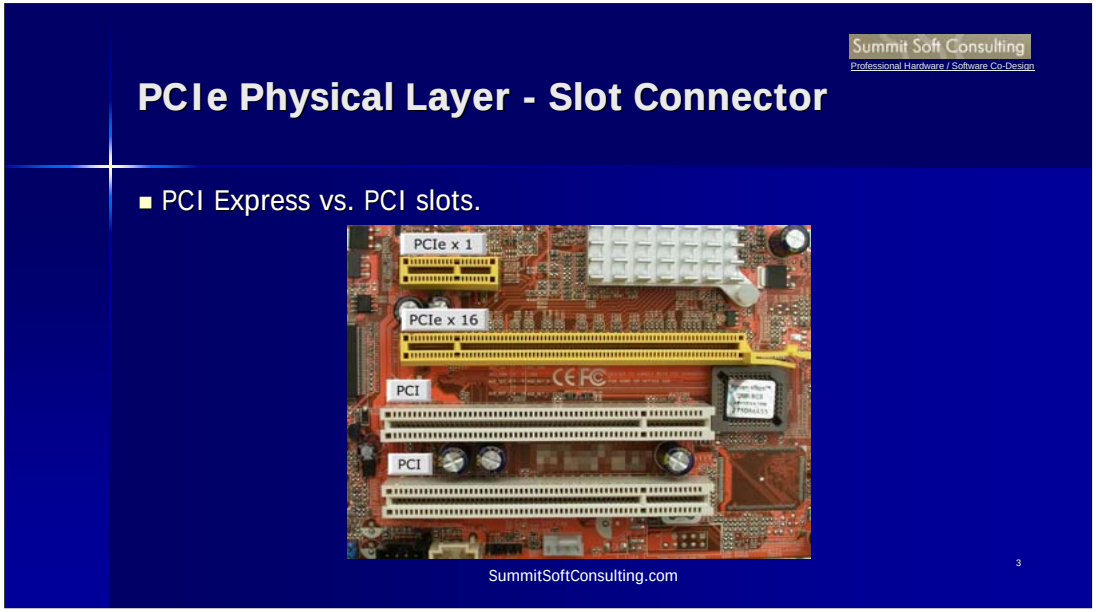
•After having used parallel buses over 21 years, PCI Express brought high-speed serial buses to the PC platform.

•Parallel data is serialized via shift registers and sent in serial form at gigabit per second link frequencies.

•8b/10b encoding was patented by IBM in 1984 and used in fibre channel and other serial link protocols.

-8b/10b encoding allows data to be encoded in such a way that the clock signal can be embedded within the serial data sent.

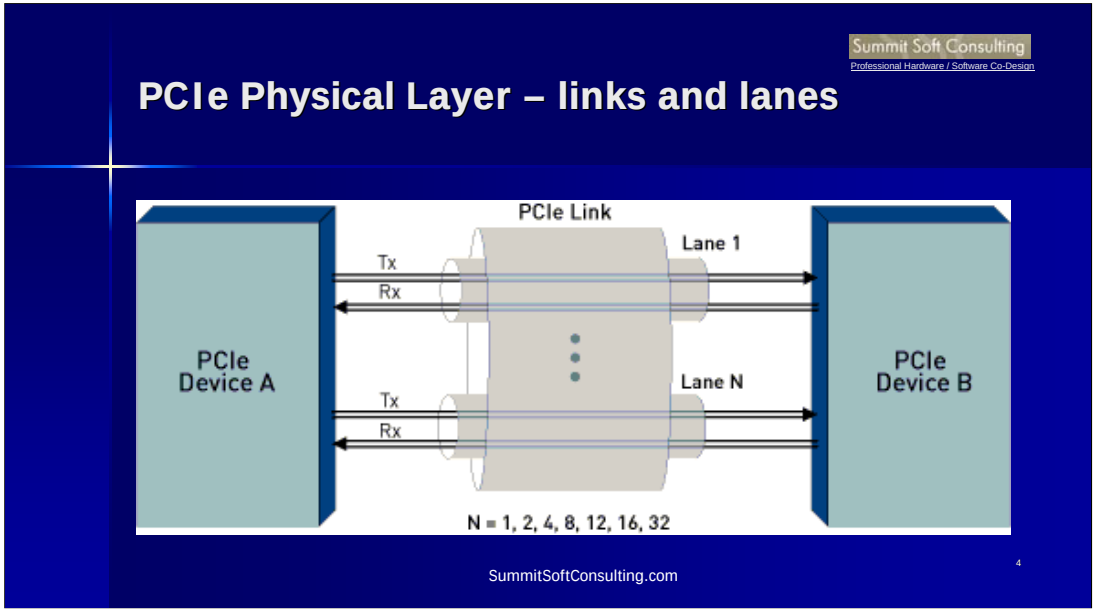
-This allows the receiver to synchronize its local receiver clock to the same clock the data was sent with, important at gigabit frequencies.



•The PCIe connectors are very similar to, but not compatible with, the PCI connectors (different pitch and number of pins).

•The PCI Express connectors use smaller pitch and are turned so not backwards compatible with any older connector type.

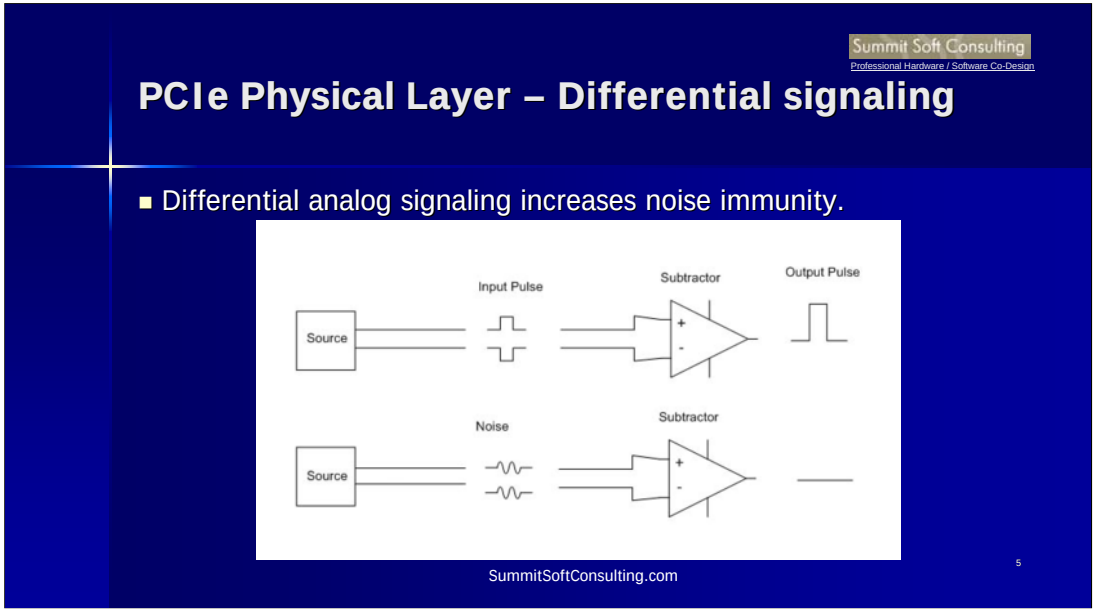
•The yellow PCIe x16 slot is normally used for graphics card but can be used by any PCIe card.



•A PCI Express link between the PC (root complex) and device (ASIC or FPGA) is comprised of 1 to 32 lanes

•Each lane is comprised of two uni-directional, differential signals, forming a bi-directional communication path.

•Each device uses a SERDES hardware block to SERialize and DESerialize between internal (parallel) to external (serial) data.



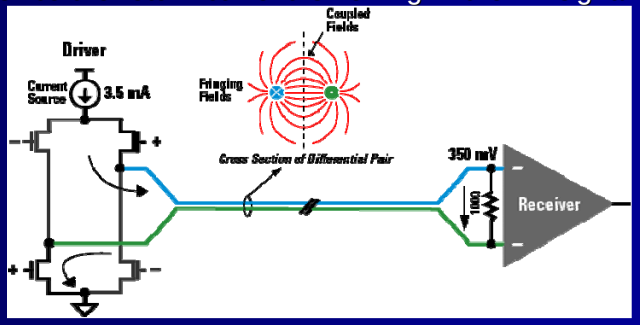
•The transmitter sends the inverted signal (-) along with the non-inverted signal (+).

•The + and – traces are routed together so any externally induced noise will affect both and cancel out at the receiver.

PCIe Physical Layer – Differential signaling

Summit Soft Consulting
Professional Hardware / Software Co-Design

- Differential analog signaling decreases electromagnetic interference (EMI) since the return current is flowing in the 2nd signal trace.

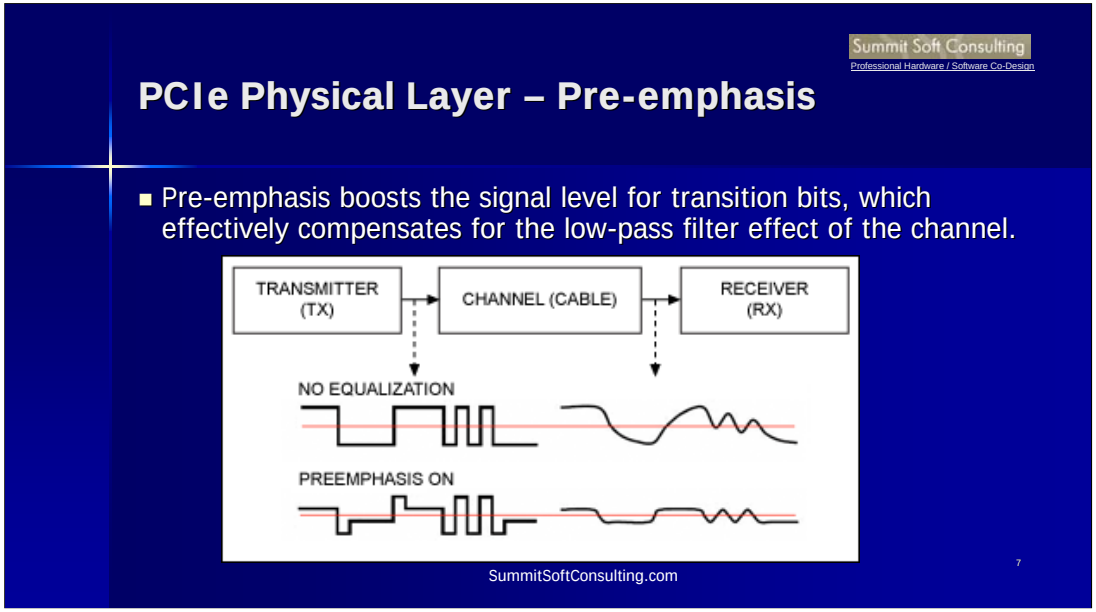


SummitSoftConsulting.com

6

•The driver constantly drives 3.5mA either on the + or – signal, which results in a 350mV differential voltage across the receiver termination (note: PCI Express uses Current Mode Logic, CML, which typically uses 4mA for a 400mV receiver voltage).

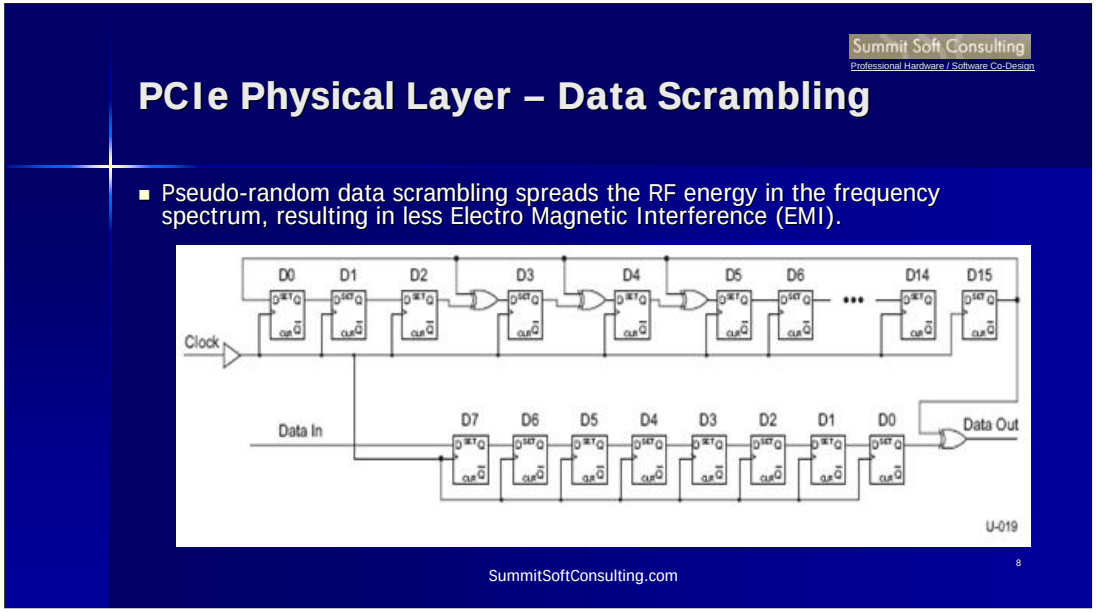
•Also, since the 3.5mA current is always driven (only the direction is alternated for a ‘1’ or ‘0’ bit) the risk of Ground Bounce due to SSO (Simultaneously Switching Outputs) is eliminated (issue for wide PCI buses).



•Without pre-emphasis, the high-frequency content of the signal will be attenuated by losses in long PCB or cable runs, which results in slower edge transitions at the receiver.

•By increasing the voltage of transition bits (more high-frequency energy goes in) we will be able to recover the correct waveform at the channel output (more high-frequency content comes out).

•Pre-emphasis helps combat Inter-Symbol Interference (ISI), which is caused by slow signal transitions between two bits, causing a bit to be incorrectly interpreted by the receiver.



•In order to randomize long sequences of ‘0’ or ‘1’ sent on the link, the above Linear Feedback Shift Register (LFSR) is used.

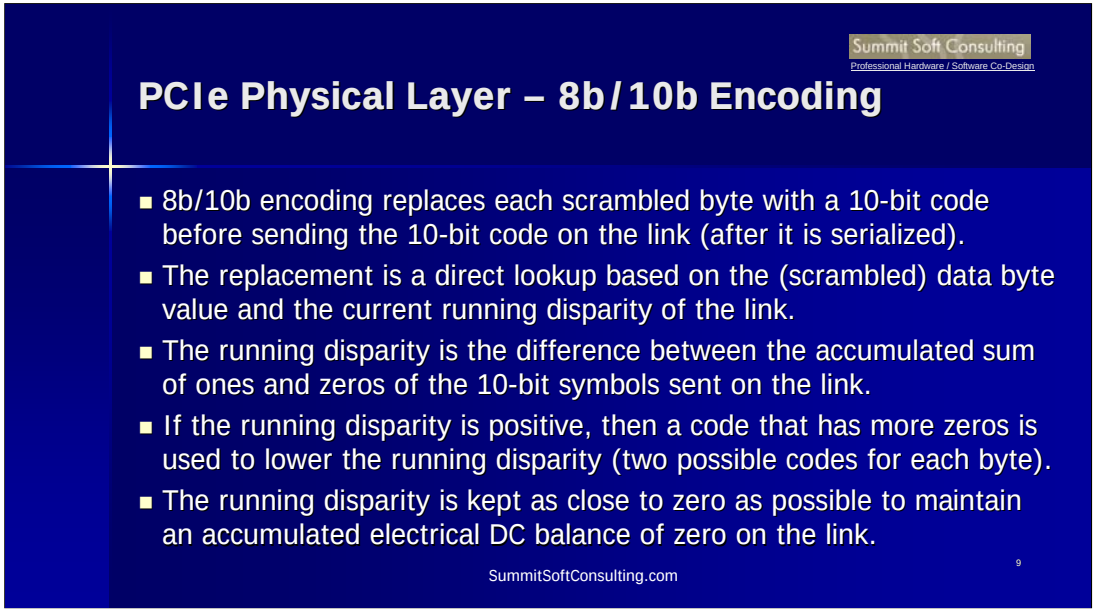
•A LFSR maintains an internal state machine, whose states are fed back into itself, resulting in a pseudo-random, predictable sequence of states.

•The pseudo-random LFSR state is then serially XOR’ed with the data bit stream, resulting in apparent random data on the link.

•The receiver uses an identical LFSR, which is reset at the same point in the data stream as the transmitter.

•By doing the same XOR operation on the data at the receiver, the original (unscrambled) bit data is recovered.

•Note: The COMMA 10b symbol is used to reset the LFSR on both the TX and RX LFSRs (8b/10b encoding explained later).

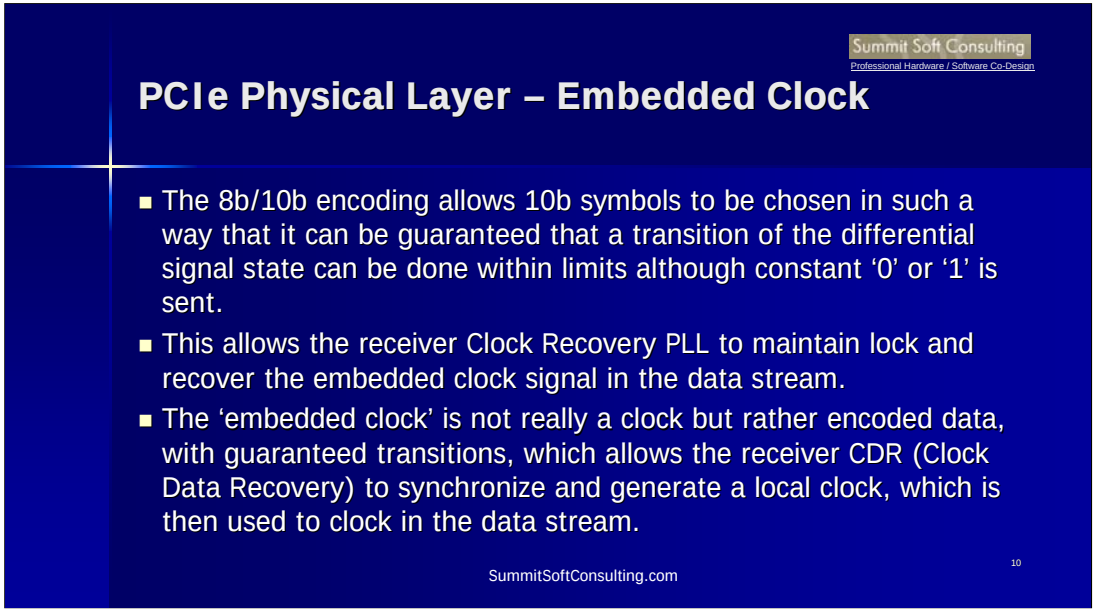


•8b/10b encoding was invented by IBM and initially used for Fibre Channel links.

•It is now used for PCI Express, USB 3.0 and many other high-speed SERDES technologies.

•The DC balance should be kept to zero on the link to ensure that the voltage at the receiver input is within the limits allowed.

•For example, if more ones than zeroes are sent on the link, the DC balance will increase over time and result in a total DC voltage level that is outside the (finite) limit allowed by the receiver, causing data capture errors.

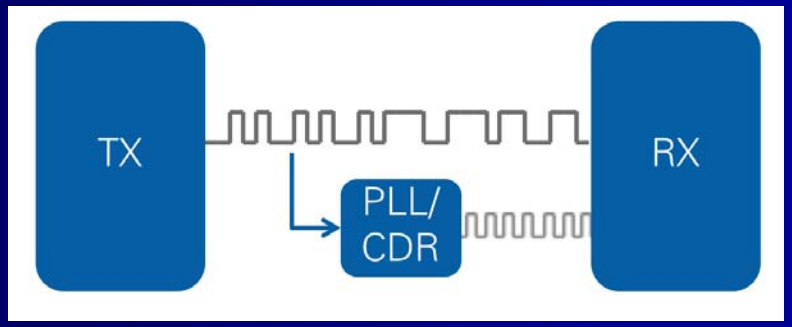


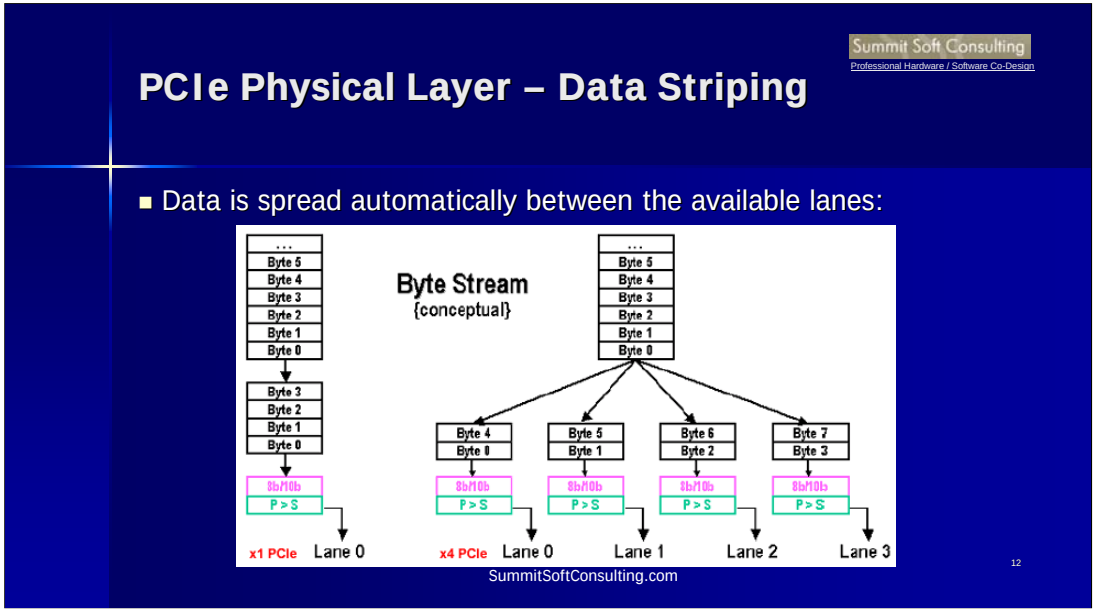
* See next page for example.

Summit Soft Consulting
Professional Hardware / Software Co-Design

PCIe Physical Layer – Clock Recovery

- The PLL/CDR recovers the clock, which is used to clock in data.



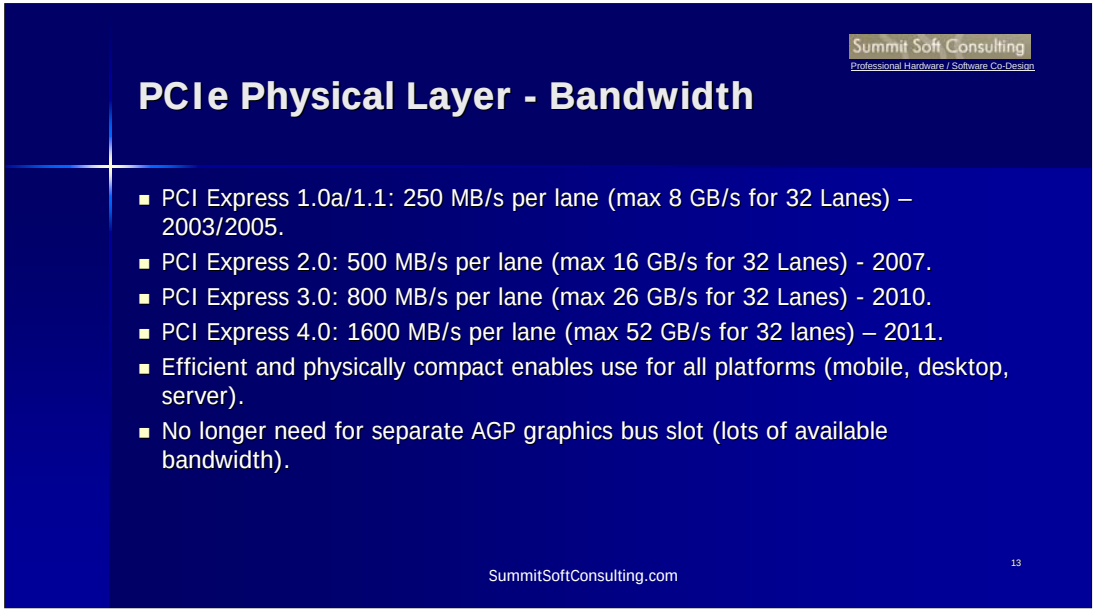


•When a link with multiple lanes (x4 shown), the PCIe hardware blocks will automatically stripe data across the available lanes, starting in lane 0.

•This will allow the bandwidth to be multiplied simply by adding lanes to the link.

•Regular I/O devices usually have x1 to x4 link width while graphics adapters typically use x16 link width.

•FPGAs typically use a built-in x4 SERDES hardware, allowing 1 GByte/s or 2 GByte/s depending on if Gen1 or Gen2 is used.

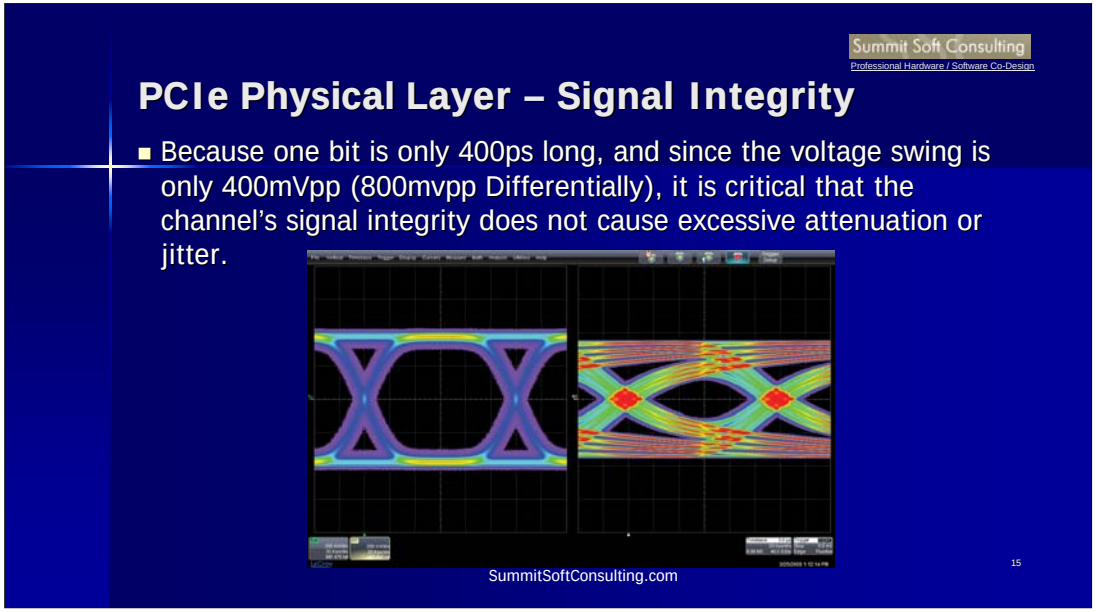


•This slide shows the performance per lane as well as maximum performance for a x32 lane link.

Summit Soft Consulting
Professional Hardware / Software Co-Design

PCIe Physical Layer – Link Training

- When a link comes up, both peers negotiate:
  - Data rate: 2.5 Gbps by default, can be negotiated up if supported by both peers.
  - Polarity inversion: Differential +/- signals can be swapped electrically, if physically crossed.
  - Lane/Lane deskew: Delay between all lanes in link are calibrated away automatically.
  - Lane and lane number reversal: 3..0 can be changed to 0..3, if crossed physically.
- This link training is negotiated via special 'Training Set' Ordered Sets, which each lane transmits after the link come up and has data locked.
  - Training Sets are explained in a future PCIe protocol presentation.

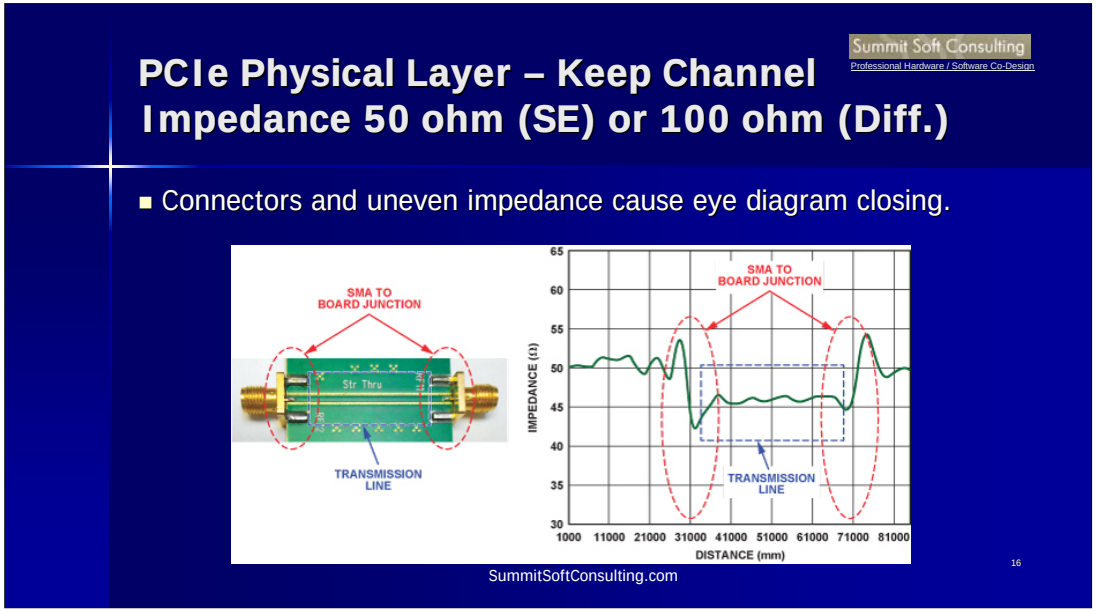


•Attenuation (reduction in receiver amplitude) is caused by low-pass filtering effects of the cable or PCB.

•Jitter (deviation of transition from the ideal position) is caused by electrical noise in the transmitter (power supply noise, random and thermal noise etc) as well as channel effects such as ISI, crosstalk, impedance discontinuities etc.

•Pre-emphasis is used to compensate for channel losses (attenuation).

•Excessive jitter and attenuation may cause trouble for the receivers PLL/CDR to lock onto the data stream, resulting in data reception errors.



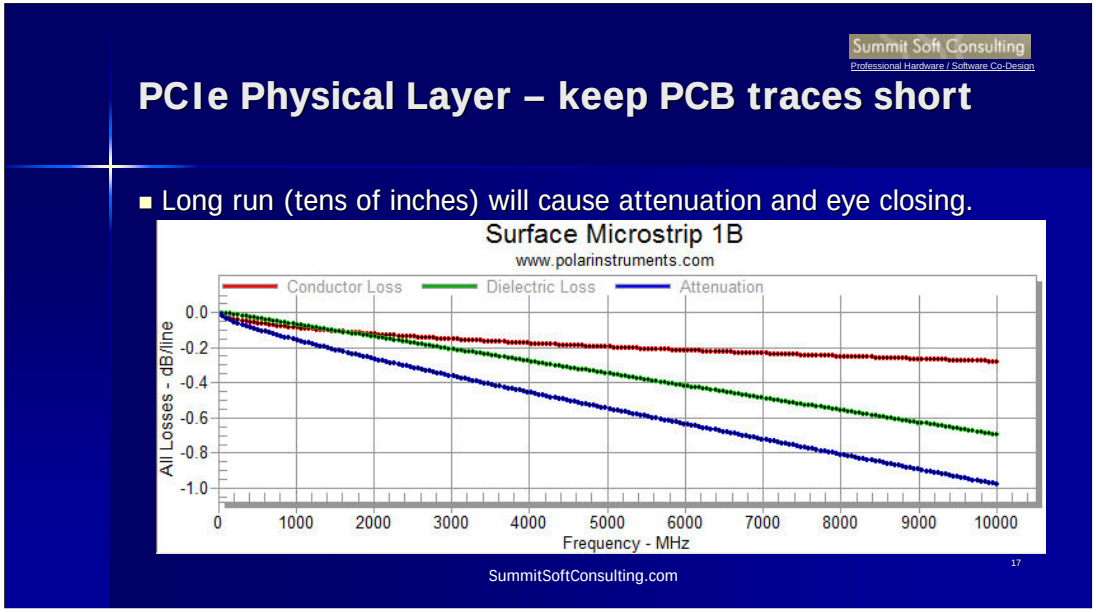
•Characteristic Impedance must be 50 ohm (single-ended) or 100 ohm (differential) across traces and connectors.

•Any deviation in impedance will cause reflections, jitter and eye closure.

•Wider trace copper results in higher capacitance so lower impedance; above, lower impedance.

•Connectors typically have less capacitance so overall inductive; above, higher impedance.

•Z0 = SQRT(L/C)



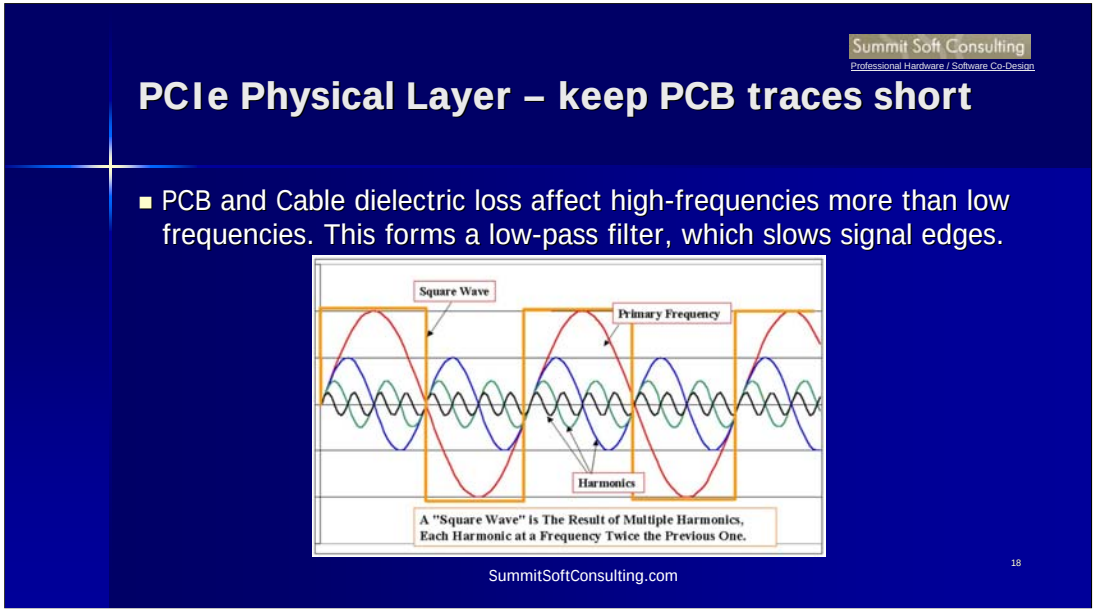
•At 5 gbps (PCIe Gen 2), we get 0.6dB loss per inch of PCB trace. 10 inches results in 6dB, which reduces the receiver eye diagram height to half of the transmitter eye. Higher speeds and longer channels may therefore cause trouble.

•It is mainly dielectric losses (reorientation and heating of the molecules in the PCB material) that cause attenuation.

•The conductor loss is not linear with the trace length but rather most significant at frequencies up to 1 GHz, mostly due to skin-effect.

•Skin-effect is the tendency of high-frequency currents to be pushed to the surface of the conductors, causing increased resistance.

•Re-drivers will restore signal integrity amplitude and clean up jittery signals without ill side-effects.

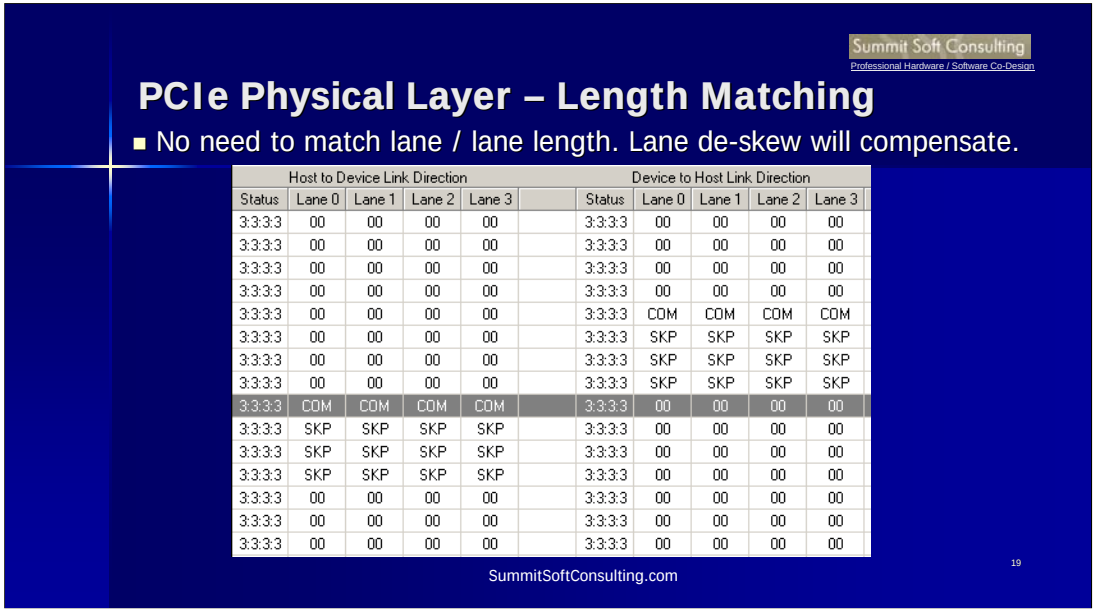


•Per Fourier analysis, a square wave is made up of sine waves of multiple frequencies. The more high-frequency content of the square-wave is removed, the more the resulting waveform looks like a sine wave.

•This is why low-pass filtering in long PCB trace runs result in slower rise and fall times.

•If not kept in check, signals will be too slow to change state between bits and ISI (Inter Symbol Interference) will result.

•In general, tens of inches cause no issues, several feet needs careful review of signal integrity.

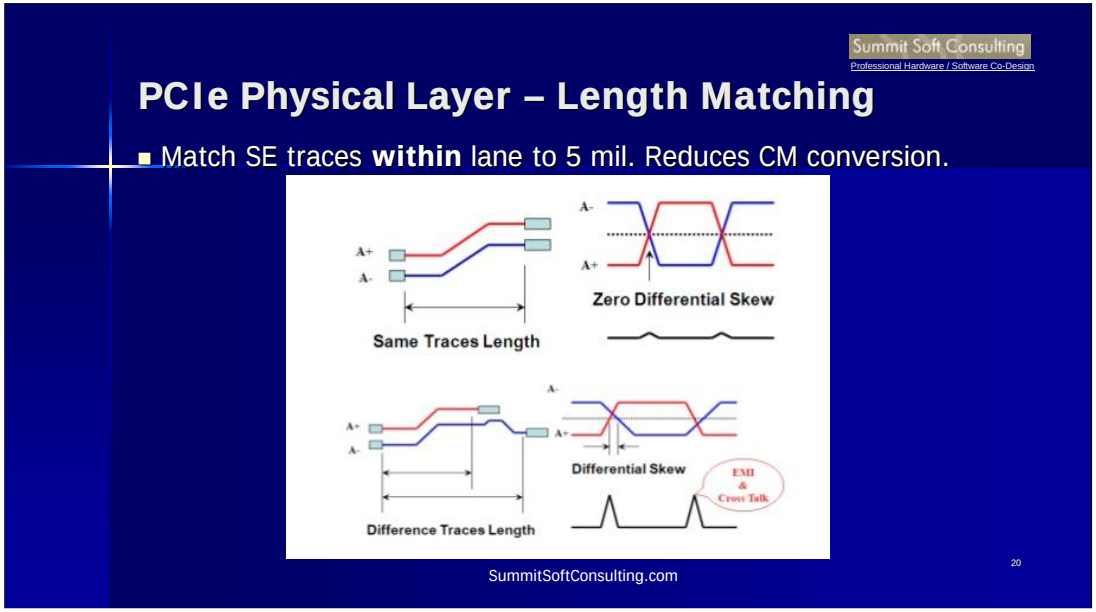


•During link training, the transmitters will send “Training Ordered Sets” (TS1 and TS2) in parallel across all lanes.

•A receiver will be able to use FIFOs to align the TS1/TS2 in time. A TS arriving early is delayed more than one arriving late, causing them all to be aligned in time.

•A maximum of five symbol times are allowed in skew across all lanes (5 x 4ns = 20ns). Assuming 6 inch propagation per ns, the physical lane trace length difference can be up to 120 inches or about 10 feet.

•Therefore, matching lane length to other lanes is a non-issue in PCI Express (even for higher Gen 2/3 speeds).

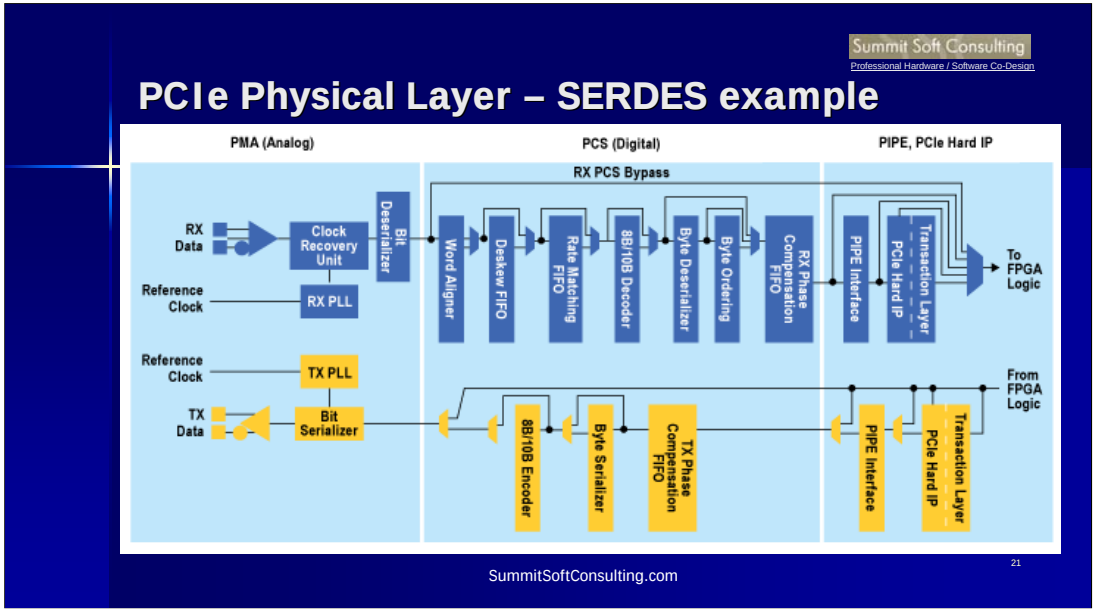


•Make sure to match the individual trace lengths WITHIN a differential pair to 5 mil.

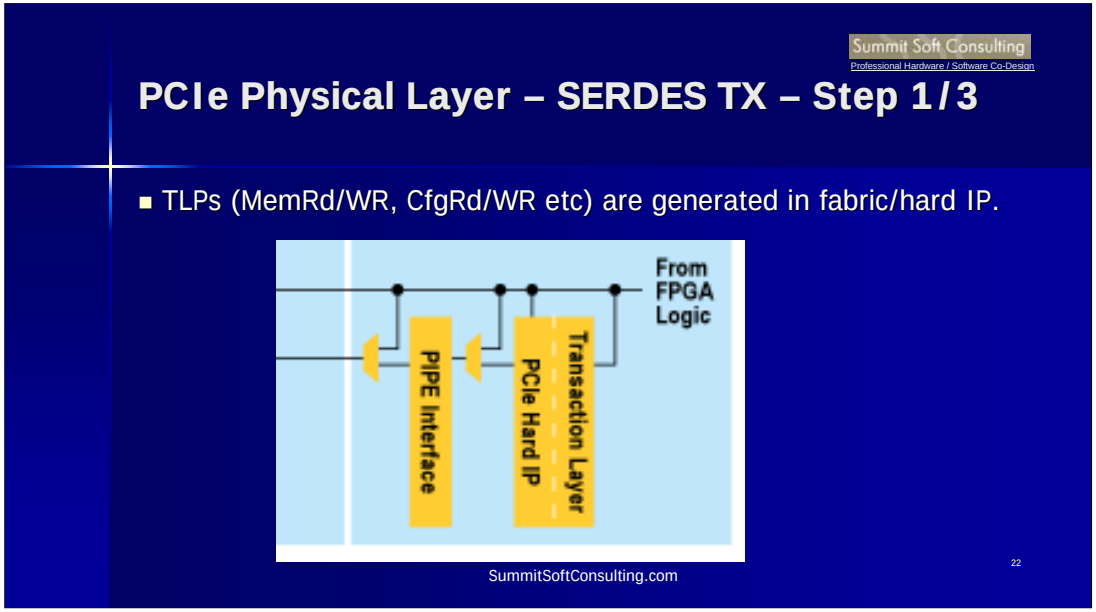
•Unmatched lines cause differential to common mode voltage conversion, where you'll get distortion of the signal.

•Also, you'll start to get common-mode voltage across the receiver inputs, with result that return currents start flowing in the PCB ground plane rather than in the differential return trace.

•This, in turn, could result in increased EMI since the loop area increases, causing a larger "antenna" effect.

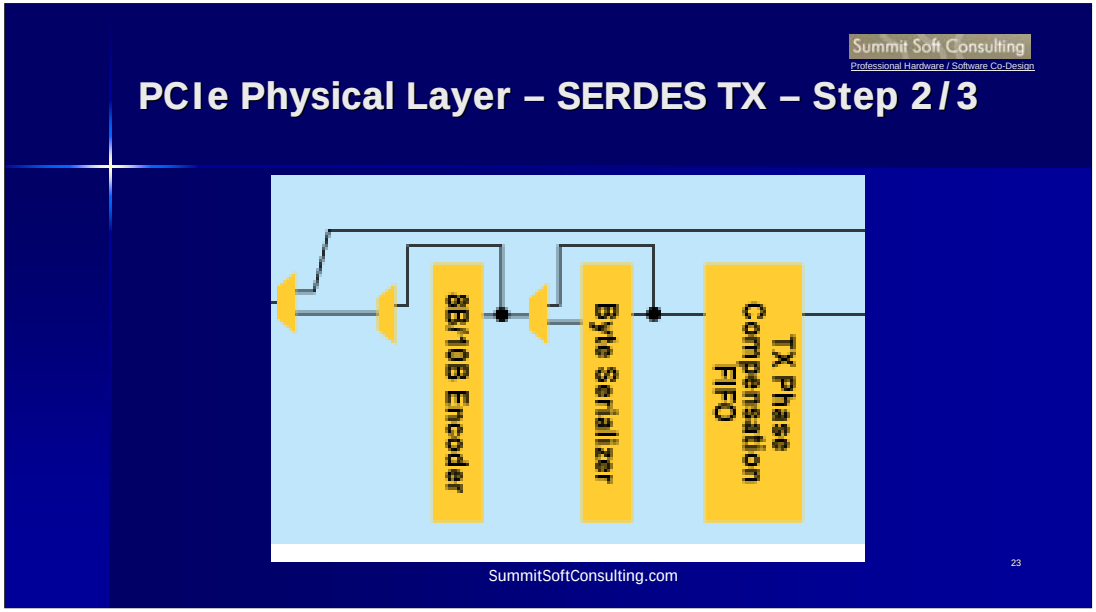


•TX Transaction Layer (Hard IP): Generates Transaction Layer Packets (TLPs), which are passed via the PIPE Interface.
•TX Phase Compensation FIFO: Simply a DC FIFO that allows differences in internal (FPGA) and external (PCIe) clocks.
•TX Byte serializer: Converts from 16-bit internal format to 8-bit format for 8b/10b encoder.
•TX 8b/10b encoder: Encodes 8-bit byte into 10-bit symbol, taking running disparity into consideration. Also encodes K/D flag.
•TX Bit Serializer: 8 bits to a 1-bit bit stream (shift register). Bits are clocked out with 2.5 Gbps PCIe clock.
•RX Clock Recovery: Locks on to the data pattern, recovers the 2.5 Gbps clock, identifies start of symbol location in the bit stream.
•RX Bit Deserializer: Converts the serial data into 10-bit encoded symbols.
•RX Word Aligner: (Not actually used when in FPGA mode)
•RX Deskew FIFO: Delays ordered sets if earlier than other lanes. Done during link training to align TS1/TS2/FTS OSs.
•RX Rate Matching FIFO: Adds/deletes SKP ordered sets to compensate for differing sender/receiver clock accuracy.
•RX 8b/10b decoder: Restores original 8-bit byte (as well as K/D symbol indicator).
•RX Byte Deserializer: Packs two bytes into 16-bit word and lowers clock from 250 MHz to 125 MHz.
•RX Byte Ordering: Swaps bytes such that COM symbol always starts in byte #0.
•RX Phase Compensation FIFO: DCFIFO to cross into the FPGA fabric's clock domain.
•Next pages will explain each block in more detail.



•The multi-DWORD Transaction Layer Packets are generated in a FIFO and passed on via the PIPE Interface as multiple 16-bit words.

•The PIPE interface is standardized in an Intel spec (Google for it).



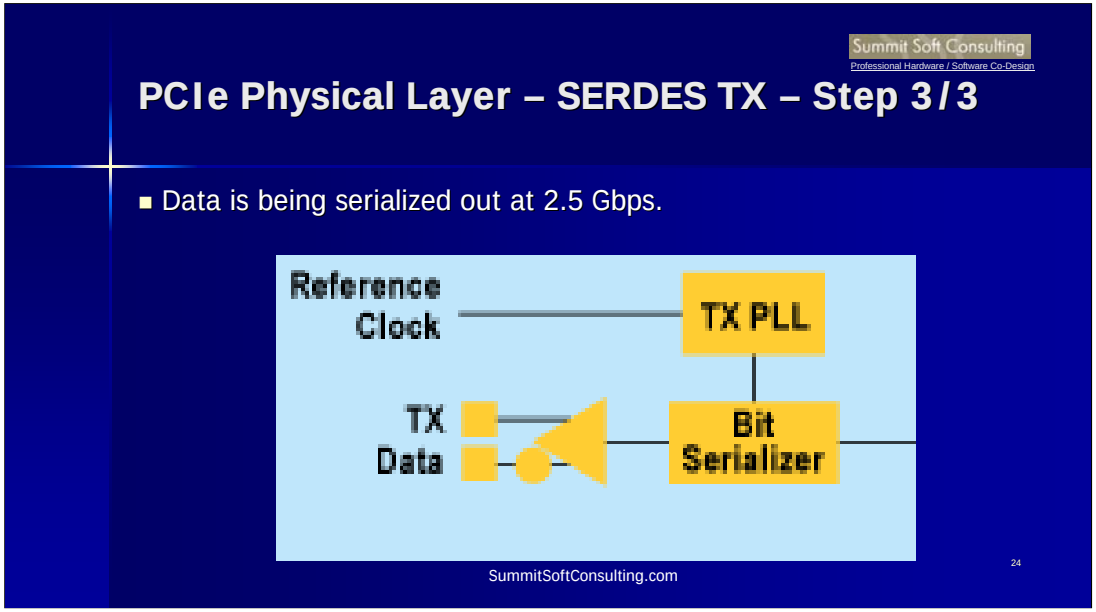
•Because the internal FPGA parallel PIPE clock may not be exactly synchronized to the external PCI Express clock, we need to cross clock domains. This is done via a 2-port FIFO, with dual clocks, here called “Phase Compensation FIFO”.

•Because the FPGA has limitations on clock frequency, it can not work with an internal 1 byte/250 MHz data path but rather uses a 2 byte wide data path clocked at 125 MHz. The “Byte Serializer” converts to a 1 byte/250 MHz data path.

•Note that the PCS (Physical Coding Sub layer) is implemented in silicon so does not have the same clock frequency limitation as the general FPGA fabric.

•The 8b/10b encoder creates the 10b codes needed to maintain DC bias on the link. It is also needed to later recover the embedded clock on the receiver side. See IBM patent for implementation details (http://www.google.com/patents/US4486739).

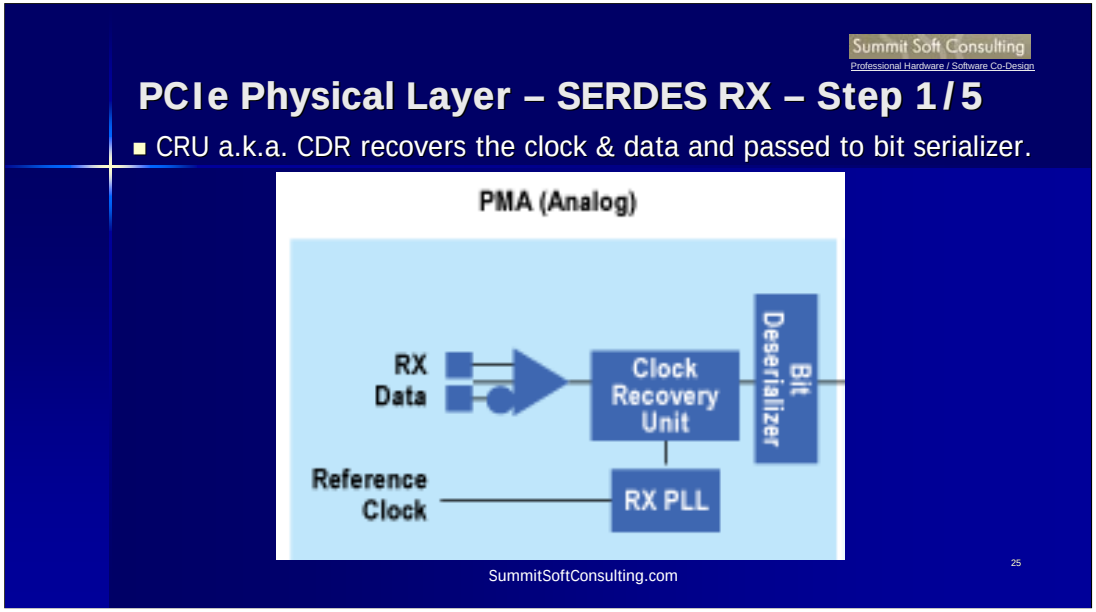
•Note that the scrambling block is missing from this diagram. There should be a LFSR after the 8b/10b encoder.



•Note that PCI Express uses a 100 MHz, out-of-band reference clock to drive the TX PLL. This ensures that the TX and RX SERDES modules are always synchronized with regards to clock frequency.

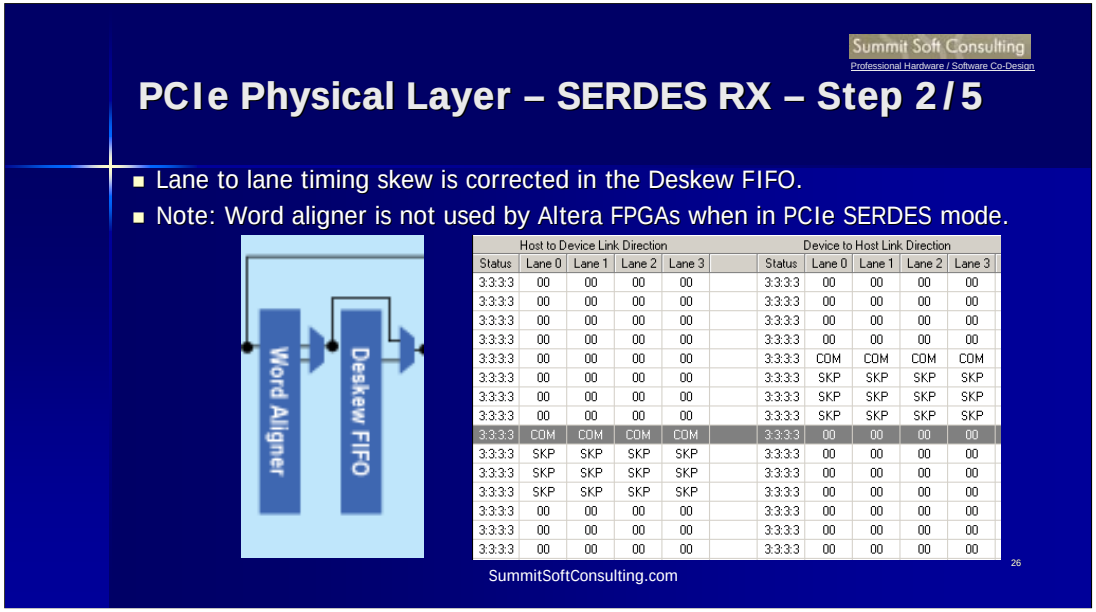
•Note that other high-speed serial technologies like USB 3, SAS, SATA do NOT use the same OOB reference clock but completely relies on the clock embedded in the data signal.

•The TX PLL multiplies the reference clock to 2.5 GHz, which is then used to clock out the TX data onto the wire.



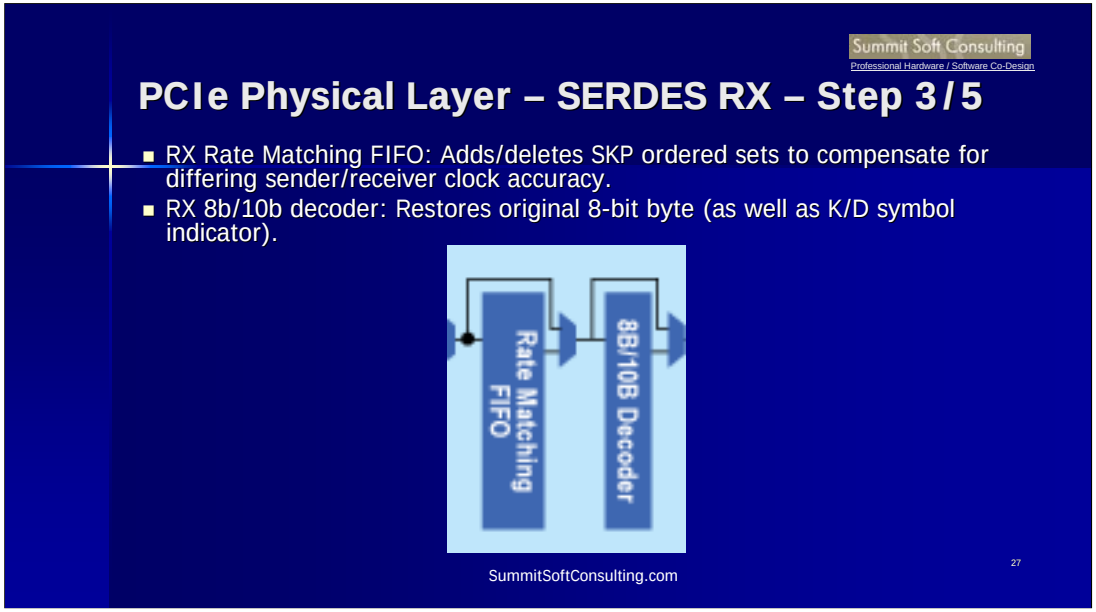
•The Clock Recovery Unit (a.k.a. Clock/Data Recovery) uses an internal PLL to lock onto the transitions in the 2.5 Gbps data being received. This forms a 2.5 GHz clock that then is used to clock in the actual data bits.

•The Bit Serializer detects the beginning of a special 10b symbol (COMMA/K28.1), which is unique, and then converts the serial data to parallel form.



•The Deskew FIFO on one lane communicates internally with the other lane deskew FIFOs to ensure aligned ordered sets during link training.

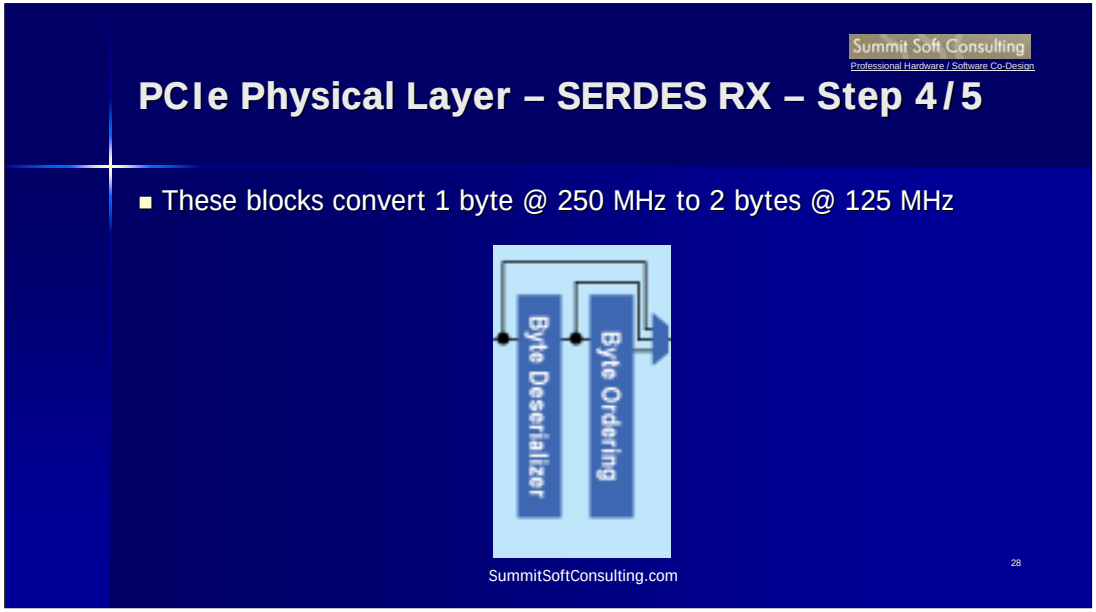
•Note: The Protocol Analyzer view to the right shows the lane timing AFTER internal deskew in the Protocol Analyzer. The actual skew between the lanes could otherwise be a couple of clock ticks off from lane to lane.



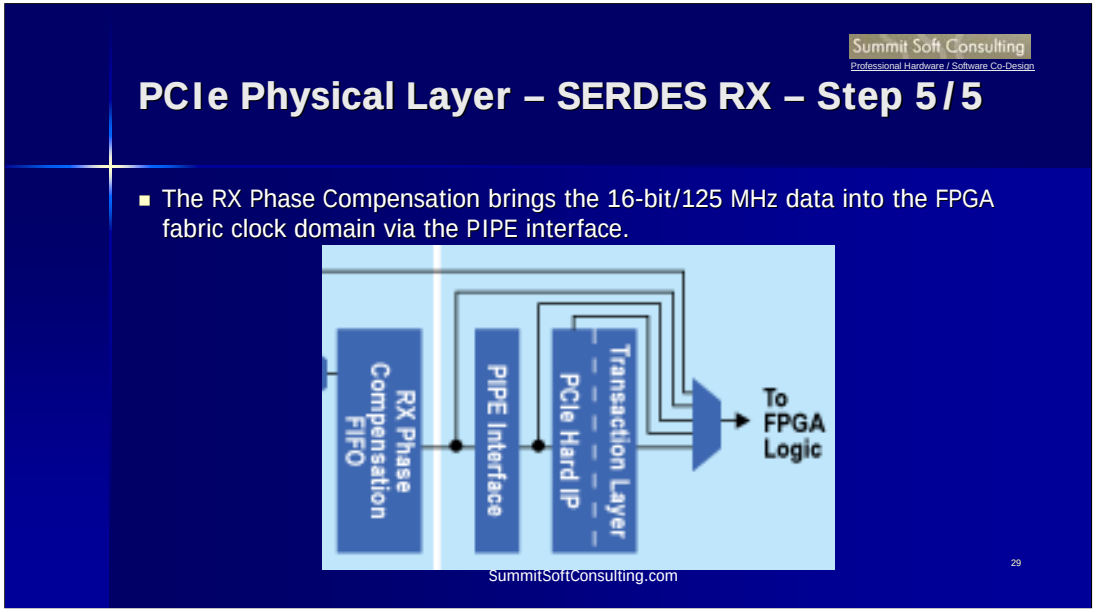
•Because the sending SERDES will not have an exactly matched clock with the RX side, the Rate Matching FIFO has the capability to remove or add SKP OS from the received data stream to avoid over or under-flow of its FIFO.

•The 8b/10b decoder is simply the reverse of what was done on the sending side. It restores the 8-bit data and K indicator.

•The K symbols are approximately ten special control symbols used for packet framing, symbol alignment, clock frequency compensation etc.



•The FPGA fabric is rather slow performing with regards to clock frequency so the incoming data needs to be converted to a wider path running at lower clock frequency before being passed to the general-purpose FPGA fabric.

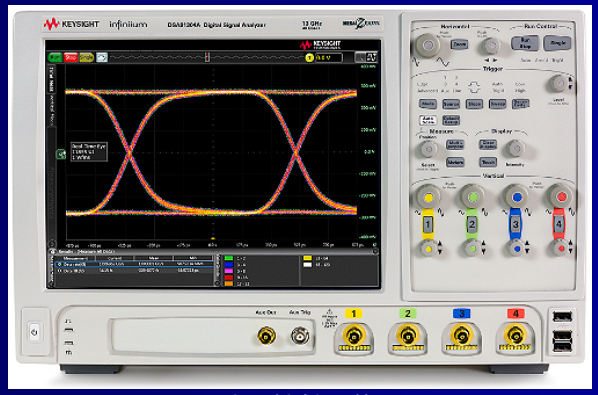


Note: The receiving-side FPGA logic will parse the received TLPs and act on it by executing internal state machine actions such as reading or writing from/to internal device memory.

Summit Soft Consulting
Professional Hardware / Software Co-Design

PCIe Physical Layer – Analog Lab Instruments

- Use a high-speed oscilloscope to validate eye opening on the RX side.



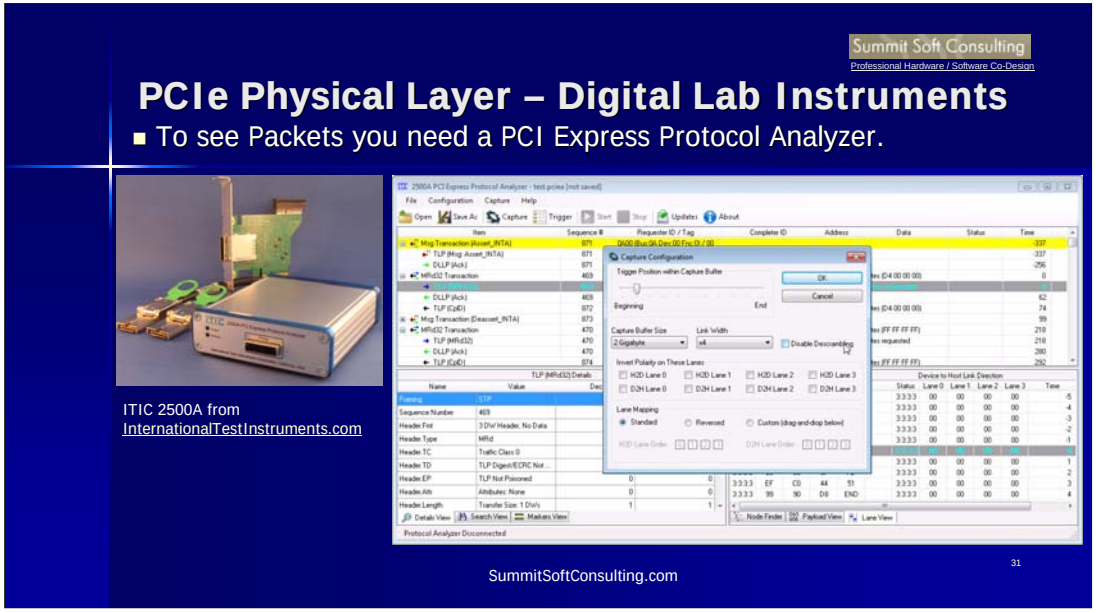
SummitSoftConsulting.com

30

•You'll need a bandwidth approximately 5x the bandwidth of the highest fundamental frequency in the data signal.

•For a Gen1 8b/10b encoded signal, which has a highest fundamental frequency of 1.25 GHz (one bit High, one bit Low per cycle), you'll need a 5 x 1.25 or approximately a 6 GHz oscilloscope.

•The same math, says 12 GHz for Gen 2 (5 Gbps), 20 GHz for Gen3 (8 Gbps) etc.



•Because the data is scrambled, 10b encoded and striped across lanes, you will need a multi-lane PCI Express Protocol Analyzer to view the data on the link under test.

•The ITIC 2500A does de-skew, descrambling, high-level packet display as well as analog lane inversion and lane swapping allowing you to, in detail, analyze packet traffic between your motherboard and custom PCIe plug-in card.

•See InternationalTestInstruments.com for details.

Summit Soft Consulting
Professional Hardware / Software Co-Design

PCIe Physical Layer – Books

I recommend reading these books for more in-depth information:

- PCI Express Electrical Interconnect Design
  - http://tinyurl.com/qwdkyju
- The Complete PCI Express Reference
  - http://tinyurl.com/jkuemwq
- PCI Express Technology 3.0
  - http://tinyurl.com/hwxayy6

About Summit Soft Consulting & John Gulbrandsen, Consultant

Summit Soft Consulting
Professional Hardware / Software Co-Design

Company History
Summit Soft Consulting was founded to offer expert consulting services in device driver and electronics peripheral device design. Over the 20 years we have been in the electronics and software engineering fields, we have had extensive experience with microcontrollers, digital and analog electronics, Windows x86/x64 software and device driver implementation, high-speed board design, signal and power integrity, advanced FPGA digital designs, USB and PCI Express Protocol Analyzer designs and much more.

Summit Soft Consulting is comprised of a team of consultants, each with complementary or overlapping skills related to Windows Systems Programming and Advanced Electronics Design. In addition, we work with a growing network of individual consultants across United States, which allows us to provide specialty niche expertise that, perhaps, may not be readily available in-house. The end benefit for you, our client, is that you will work with our competent and experienced team that safely will bring your project to completion within set cost and time budgets.

We are located in Aliso Viejo, Orange County, Southern California, mid-way between Los Angeles and San Diego.