



Enabling 100 Gigabit Ethernet

Implementing PCS Lanes





Contents

Introduction	2
The IEEE 802.3 Protocol Stack	3
PCS Layer Functions	4
Striping	4
The Need for Virtual Lanes	5
Multiplexing of Virtual Lanes	6
Testing Challenges.....	8
Receiver Testing	8
Transmit-Side Testing	10
Summary.....	11
Biographies	11



Introduction

Advances in data center technology and Internet usage have pushed enterprise data centers from Gigabit Ethernet links to 10 Gigabit Ethernet (GbE) links and search engines/carriers/ISPs from single 10 GbE links to multiple links. Projected growth indicates the need for higher-speed Internet connections¹. The IEEE 802.3ba task force has been established with the objective of standardizing 100 GbE and 40 GbE over the Internet and within the data center.

If higher-speed Ethernet is to be useful in the near term, implementations must take advantage of existing copper and fiber cables, both in the data center and over the Internet. This poses an interesting problem since no technology currently exists to transport 100 Gbps or 40 Gbps as a single stream over either media. In order to transport 100 GbE over single-mode fibers, for example, 4 different wavelengths will be required using LAN wavelength-division multiplexing (LAN WDM). Similarly, the device interfaces found in routers, switches and servers that drive these higher-speed links cannot currently handle single 100 Gbps or 40 Gbps data streams. They will be forced to resort to parallel electrical “lanes” to handle the flow of data; for example, 10 lanes of 10 Gbps.

As technology improves, the bandwidth of fiber wavelengths and electrical lanes will increase at independent rates. Efficient 100 and 40 GbE links will need to handle a number of combinations. A technique is needed to allow efficient implementation that allows for changing numbers and bandwidths of wavelengths and electrical lanes. Within the 802.3 Ethernet specification, the protocol stack layer that performs this function is called the physical coding sublayer, or PCS. This white paper describes the baseline proposal for the physical coding sublayer (PCS) for the 100-Gbps and 40-Gbps Ethernet interfaces currently under standardization within the IEEE 802.3ba task force.

¹ Schneider, David. *An Overview of Next-Generation 100 and 40 Gigabit Ethernet Technologies*. Available at http://www.ixiacom.com/library/white_papers.

The IEEE 802.3 Protocol Stack

Figure 1 shows a simplified version of the standard IEEE 802.3 protocol stack. The subject of this paper relates the PCS layer and the unique requirements for 100 Gbps Ethernet.

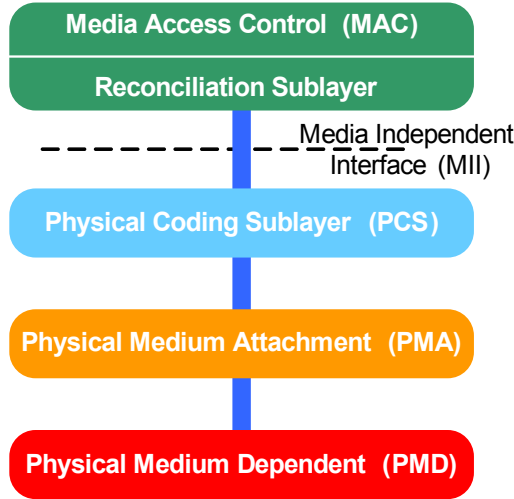


Figure 1. Simplified IEEE 802.3 protocol stack

Figure 2 shows the architectural partition for a 100 Gbps Ethernet link. The partitioning assumes:

- A highly integrated 100 GbE MAC/PCS chip that includes the packet interface and MAC, PCS, and PMA functions.
- An optical module that includes both PMA and PMD functions.
- A high-speed parallel electrical interface comprising n lanes between the MAC/PCS chip and the optical module.
- A parallel optical interface comprising m lanes, where lanes can be wavelengths or fibers.

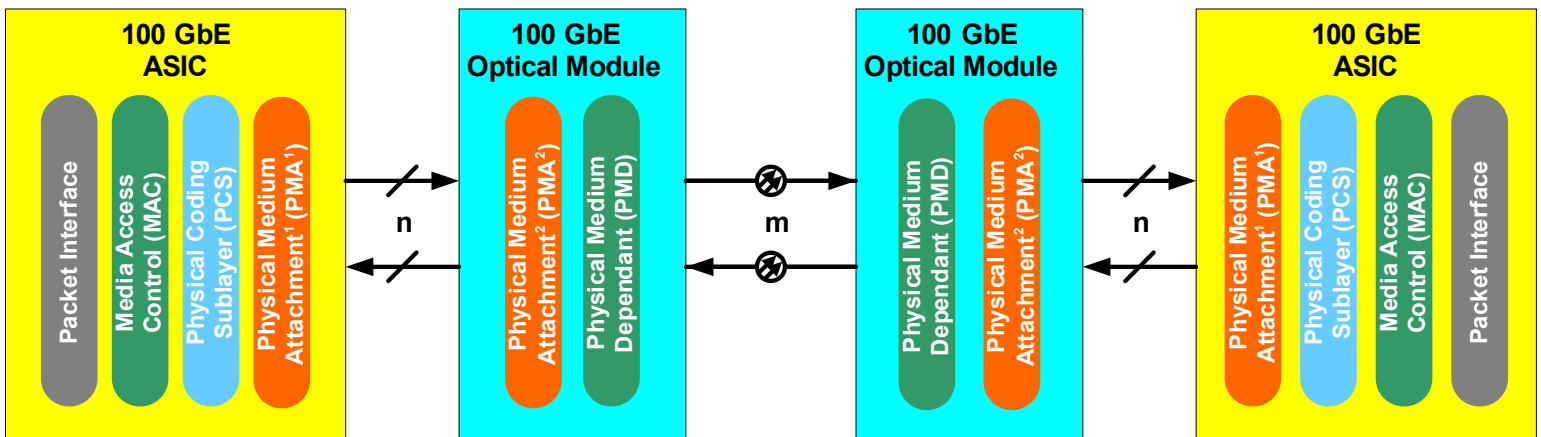


Figure 2. Architectural partition for a 100 GbE link

PCS Layer Functions

The requirements for the PCS layer include:

- Provide frame delineation
- Control signal transport
- Provide the clock transitions needed by SerDes-style electrical and optical interface
- Bond multiple lanes together through a striping or fragmentation methodology

Since the data must travel on multiple optical lanes as well as on multiple electrical lanes to the optical module, the striping mechanism should support:

- Low frame overhead that is independent of frame size
- Formatting that enables receiver-only lane deskew
- Evolving media and media interface widths
- Simple optical module implementations

Striping

Where the number of electrical and optical lanes is the same, a simple striping of 66-bit blocks across the lanes is sufficient, as shown in Figure 3 for four electrical lanes.

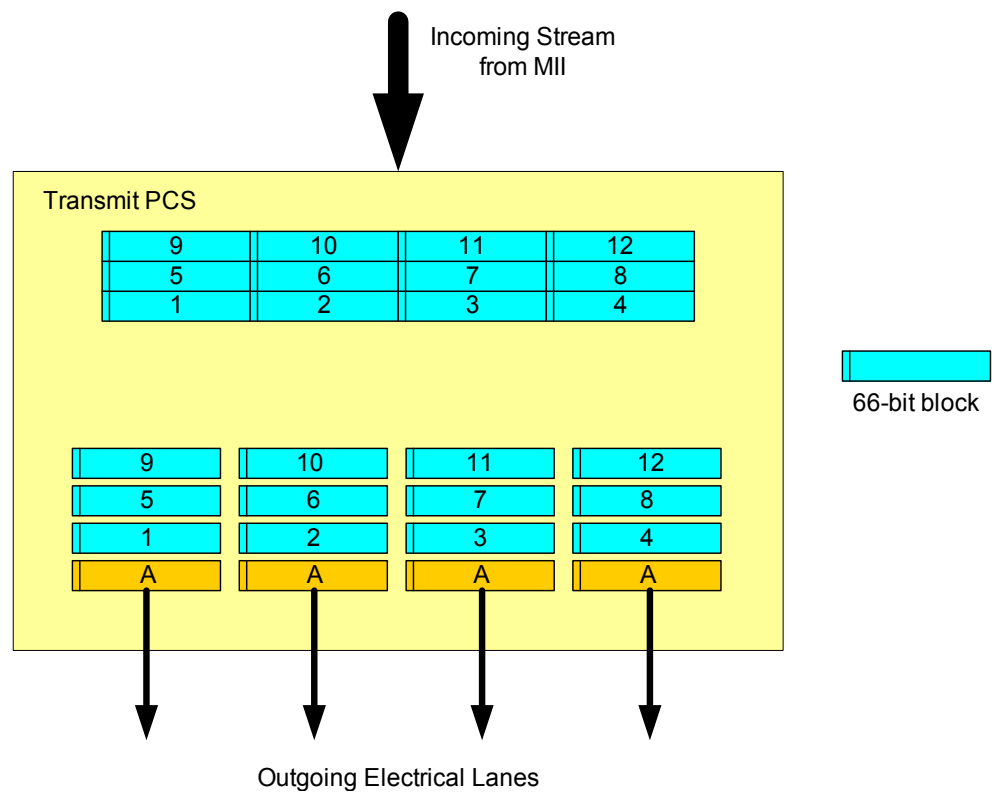


Figure 3. Simple striping of data



The transmitting PCS first performs 64/66b encoding on the incoming stream. The 66-bit blocks are distributed across the lanes on a round-robin basis. An alignment block is periodically added as a marker that allows the receive side to deskew the lanes. This allows the receiver to remove skew between the individual media lanes. Bandwidth for the alignment blocks is created by periodically deleting inter-packet gaps.

Line coding is implemented once at the aggregate level in the transmitting PCS block using a single scrambler in such a way as to ensure adequate transition density on any of the individual electrical and media lanes along the path. The deskew function is also only implemented once in the receiving PCS block, and again in such a way as to compensate for skew introduced over any of the electrical or optical interfaces along the path.

The advantages of this approach are obvious; it moves all of the scrambling and deskew features into a CMOS chip, where they are easy and cheap to implement, and therefore significantly simplifies the design of the optical module.

The Need for PCS Lanes

As discussed earlier, advancing electrical and optical technologies will require the ability to handle differing and changing numbers of electrical interface lanes versus optical lanes. To handle the general case, the PCS baseline proposal calls for data to be distributed on a per-66-bit block basis to a number of PCS lanes v , where the number of PCS lanes equals the least common multiple of the number of electrical lanes n and optical lanes m . Using the virtual lane concept, the optical module can be a very simple multiplexer, which merely bit multiplexes the data from n electrical lanes down to m media lanes. The receiver's PCS block can simply demultiplex the data back into the PCS lanes and then realign the skewed data.

A virtual lane is a continuous stream of 64/66b blocks. PCS lanes are created through a simple round-robin function which distributes 66-bit blocks, in order, to each virtual lane. In the case of an interface running with 20 PCS lanes, a single virtual lane would contain every 20th 66-bit block from the aggregate signal, as illustrated in Figure 4 below.

In order to allow the receiver to identify and deskew the individual PCS lanes, a unique alignment block is added to each virtual lane on a periodic basis. The alignment block is a special 66-bit control signal block that is unique to each virtual lane and cannot be duplicated in the data. The current proposal under consideration calls for an alignment block once every 16,384 blocks on each virtual lane.

The number of PCS lanes required depends on the electrical and optical lane combinations that need to be supported. The number of PCS lanes is the least common multiple of the number of electrical lanes and the number of optical lanes, regardless of whether the lanes are based on

numbers of wavelengths, numbers of fibers, or other considerations. This constraint ensures that the total number of PCS lanes can be mapped evenly over both the number of electrical lanes and the number of optical lanes; therefore, the data from any particular virtual lane will always reside on the same electrical and media lane across the link. This guarantees that no skew can be introduced between the bits within a virtual lane, which would be impossible to remove at the receiver.

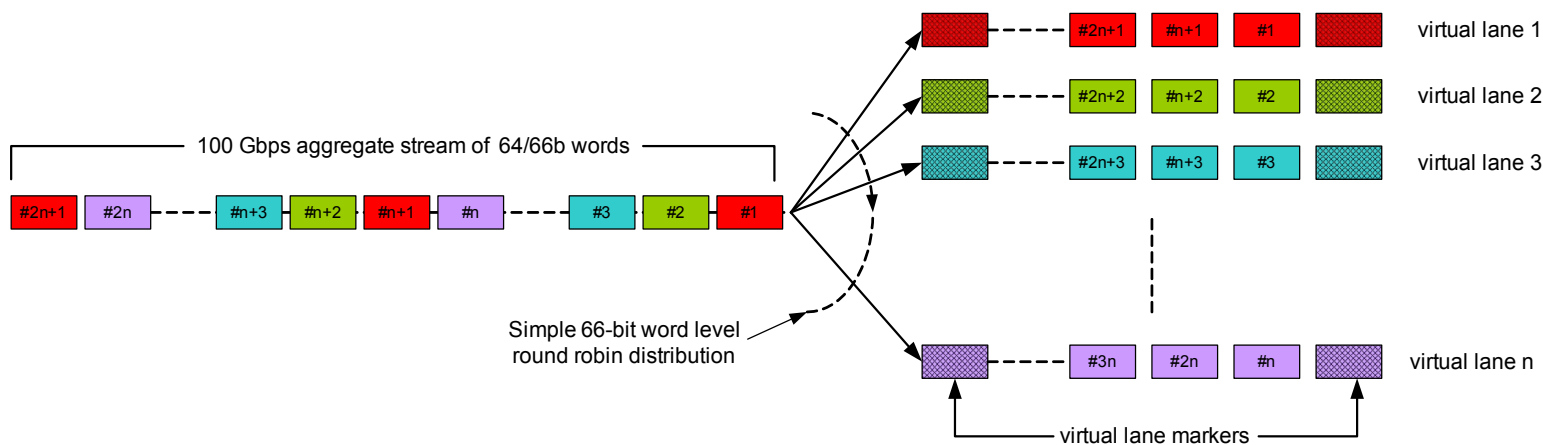


Figure 4. Virtual lane data distribution

Multiplexing of PCS Lanes

Virtual lane muxing is straightforward, consisting of bit-interleaving of all of the PCS lanes assigned to a particular electrical or optical lane and is performed in the PMA layer. Figure 6 illustrates a case with:

- 20 PCS lanes
- 10 electrical lanes
- 4 media lanes

In Figure 5, VL stands for virtual lane and the nomenclature n.m indicates virtual lane n, bit m. The PCS layer created the 20 PCS lanes from the data source flow, as covered above. The PCS layer bit-multiplexes each pair of PCS lanes into a single lane and sends 10 electrical lanes to the PMA in the optical module. The optical device's PMA then bit-multiplexes the 10 electrical lanes into the 4 optical lanes.

Due to skew on the electrical interface, the random startup state of the bit multiplexers and demultiplexers is not predictable where PCS lanes will end up in the receive PCS. The unique alignment words, however, provide the means by which the receive PCS layer identifies and reorders the individual PCS lanes.

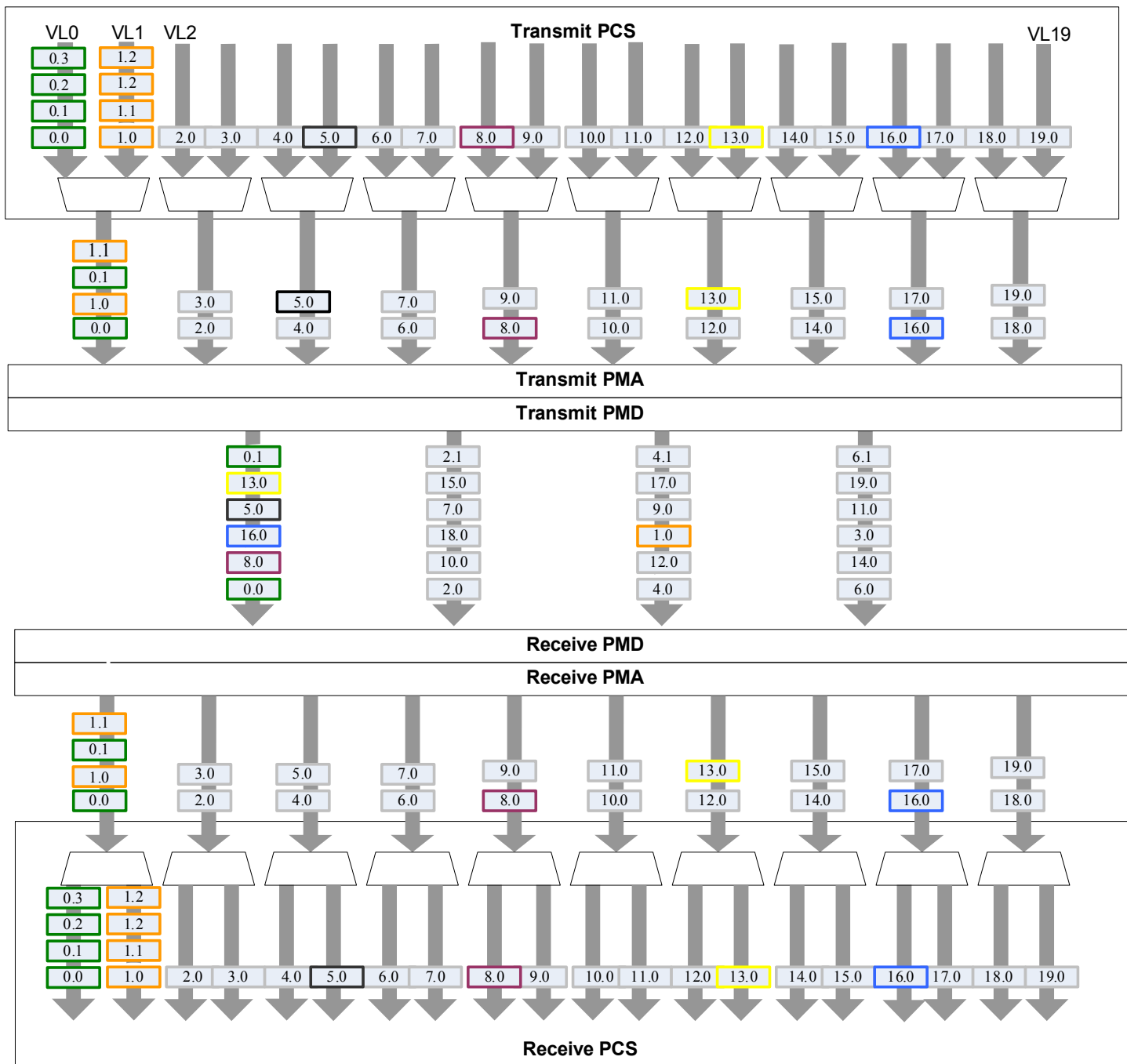


Figure 5. Multiplexing of PCS lanes to optical media



Testing Challenges

As with any new technology, 100 GbE testing will be a significant challenge. Although this paper largely concentrates on PCS layer testing issues, it's important to note that layer 2 through 7 testing is as important as ever. The 100 GbE and 40 GbE links will be incorporated in switches, routers, and servers almost immediately. It's important that real-world traffic be used to properly exercise the network device/server, to determine if it can properly sustain traffic over multiple high speed links. The 100 GbE interfaces must operate in an integrated manner with lower-speed interfaces.

As with every new standard, multiple companies will develop products that will conform to pre-standards and then to published standards. Proper test tools are needed to allow each vendor to independently assess their conformance to standards that will lead to high assurance of interoperability. The alternative approaches are back-to-back with one's own equipment or limited interoperability testing with other vendor's equipment. The former is self-deceptive, allowing a vendor to make the same mistakes in both transmit and receive functions. The latter is expensive and endlessly frustrating, due to multiple vendors' updates.

Receiver Testing

Concentrating on just the receiver side of the PCS layer, major test considerations include:

- 64/66b word sync lock
- Virtual lane alignment marker lock
- Skew tolerance and compensation
- Arbitrary mapping of received to transmitted PCS lanes

64/66b Word Sync Lock

Each received virtual lane must properly identify 64/66b blocks, based on the control bits. The receive logic must properly lock and come out of lock under conditions covered in the specification. In order to measure conformance, various bit sequences at the limits of the specification must be transmitted by the test equipment. The receiver of the interface under test must then be subjected to intentionally errored sync bits and out-of-sync detection must be verified. Additionally, a variety of intentionally errored patterns must be applied to also ensure that the receiver does not improperly sync.



PCS lane alignment marker lock

Each virtual lane must correctly lock onto alignment markers. The techniques used here are analogous to those used in 64/66b word sync lock testing. Boundary conditions need to be verified in addition to basic locking operations. For example, if the specification calls for virtual lane lock after n consecutive non-errored alignment markers, then testing must ensure that lock is not declared with any fewer alignment markers. Conversely, if virtual lane lock should be dropped with m consecutive errored alignment markers, then testing must ensure that lock is maintained with fewer errored alignment markers.

Skew tolerance and compensation

Skew at various electrical and optical interfaces causes PCS lanes to arrive with different delays relative to other PCS lanes. The virtual lane alignment markers are designed to make it possible for received PCS lanes to identify and realign themselves properly. In order to test skew tolerance and compensation algorithms, test equipment must introduce known, specific amounts of skew amongst the PCS lanes. The receiver's PCS layer should be able compensate for the skew mandated in the standard. Test equipment must also be used to measure the actual skew tolerance when it exceeds the specifications.

Arbitrary mapping of transmitted to received PCS lanes

As shown in Figure 6, a particular virtual lane may be transmitted on any of a number of optical lanes and then received in any virtual lane position. The red lines trace a set of the possible paths for virtual lane 0. The virtual lane concept, in fact, does not require that the number of physical lanes on the transmit and receive paths match. To make matters worse, there is no requirement to enforce where specific PCS lanes are mapped along the transmit or receive path. The only requirement is for all transmitted bits on a given virtual lane to end up together on the same physical receive lane. When there is an incongruity, which will occur as the various vendors' transmit and receive ASICs and optical modules proliferate, transmitted PCS lanes can appear on any received virtual lane.

Receivers must be able to reorder and reassemble any mapping of PCS lanes into a single 100 Gbps aggregated stream. In order to exercise this reordering, test equipment must be able to transmit PCS lanes in any order to emulate differing scenarios.

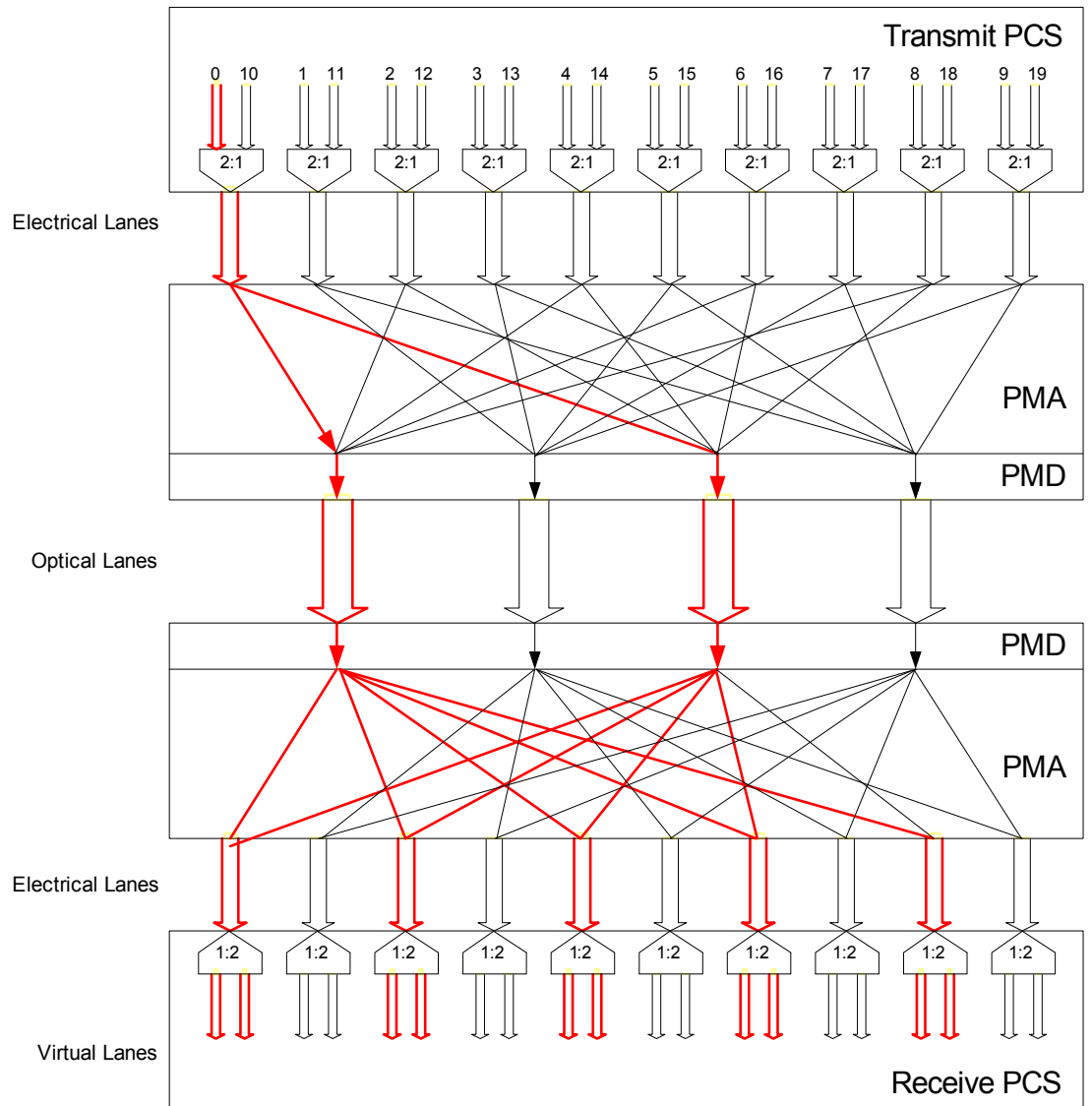


Figure 6. Lane ordering requires lane identification

Transmit-Side Testing

Transmit-side testing consists of verification and monitoring of various signal attributes:

- 64/66b sync bits. Proper transmission of sync bits must be determined and then monitored. Counts of erroneous sync bits must be available in real time during testing.
- Virtual lane alignment markers. Proper transmission of alignment markers must be verified and monitored, with error counts available.
- Lane skew. Actual lane skew must be measured and reported.
- Virtual lane mapping. The actual mapping of transmitted to received PCS lanes must be identified and reported.



Ixia and PCS

Ixia has been actively working with the IEEE 802.3ba task force. On June 17th, 2008 Ixia demonstrated the first PCS hardware implementation at NXTcomm08. The demonstration used Ixia's proof of concept 100 GbE implementation to generate and receive line-rate traffic, which was sent round-trip from Las Vegas to Los Angeles, and to analyze the results. At Interop 2009, Ixia showed the world's first 100G IP test module with a CFP interface. The 100 GbE interface card works in conjunction and in the same manner as all other Ixia interface cards and test applications, making it possible to test all types of devices and networks from layer 2 through 7.

Summary

100 GbE and 40 GbE technologies are rapidly approaching standardization and deployment. A key factor in their success will be ability to utilize existing fiber and copper media in an environment of advancing technologies. The physical coding sublayer (PCS) of the 802.3 architecture is in a perfect position to facilitate this flexibility. The current baseline proposal for PCS implementation uses a unique virtual lane concept that provides the mechanism to handle differing electrical and optical paths. Testing of this technology, however, will require refined and sophisticated test methodologies and equipment.

About the Author

Jerry Pepper is a Distinguished Engineer at Ixia, a manufacturer of network test hardware and software. Jerry has over 25 years of experience in FPGA/ASIC design. He has been an active participant in the IEEE HSSG / 802.3ba task force since January 2007, and is currently the hardware architect on Ixia's 100 GbE project, within the Ixia Labs. Jerry has been instrumental in development of Ixia's next generation load modules since 2000.



26601 Agoura Rd.
Calabasas, CA 91302
(Toll Free North America)
1.877.367.4942
(Outside North America)
+1.818.871.1800
(Fax) 818.871.1805
www.ixiacom.com

Info: info@ixiacom.com
Investors: ir@ixiacom.com
Public Relations: pr@ixiacom.com
Renewals: renewals@ixiacom.com
Sales: sales@ixiacom.com
Support: support@ixiacom.com
Training: training@ixiacom.com