A Demonstration of PCI Express Generation 3 over a Fiber Optical Link

PLX Technology and Avago Technologies

White Paper

Introduction

In 2007, the PCI SIG released an external cabling specification enabling interconnection of PCI Express systems at 2.5 Gb/s (Generation 1) and 5Gb/s (Generation 2), facilitating access to PCIe extension and expansion. Copper cabling solutions appeared in the market in a variety in channel widths to provide connectivity, but were limited in link distance and became bulky in size & weight at higher channel counts. These copper cabling solutions were able to meet the basic needs of emerging external expansion applications for Gen1 and Gen2.

With the recent adoption of PCIe's Gen3 specification, copper interconnects are finding it increasingly difficult to keep up with price, performance and size/weight needs for 8Gb/s interconnects.

Fiber Optic technology provides an alternate solution to high channel count PCIe Gen3 interconnects, with a value proposition of increased link distances, lower size/weight, higher performance and competitive pricing.

PLX Technology, an industry leader in PCIe IC solutions, and Avago Technologies, the industry leader in parallel fiber optic products, recently collaborated on a demonstration showing the first PCI Express Gen3 end-to-end fiber optic link, delivering a full 64Gbps (128Gbps bidirectional) performance for PCIe applications. This application note describes the components used, the demonstration configuration and a basic overview of functionality of fiber optic links.

PCIe and Optical Applications

The PCI Express (PCIe) bus serves as a high-speed serial IO designed to provide connections between peripherals (graphics cards, memory/disk drives, external IO cards) and the Central Processing Unit (CPU). PCIe is implemented as a point-to-point connection using a pair of differential electrical interconnects (Tx/Rx) connecting two end point devices. PCIe bandwidth can be scaled by adding multiple lanes. The overall bandwidth can be increased by combining channels into x4, x8, x16, and x32 lane links.

PCIe peripherals typically take the form of expansion cards and are connected to the motherboard interface via slot connectors. Thus, the most common usage for PCIe is to make connections between chips inside-the-box where both the central processor and peripherals are co-located. PCIe is the dominant interconnect for CPU-based applications within desktop PCs, workstations as well as large, high-end servers.

A number of applications exist where PCIe may be used to connect a central processor with devices outside the box. As it is, even at 2.5 Gb/s (Gen1) and 5Gb/s (Gen2), physical connections are limited to a few meters in length using the available copper cables. Physical link distances decrease at higher data rates, and as such, operation at 8.0Gb/s PCIe Gen3 will further reduce the usable distance of copper cables. Thus, users have expressed interest in optical solutions for PCIe applications requiring longer distances. Optical fiber based solutions allow connections over a much longer distance and are capable of providing better bit error rate performance, better immunity to electromagnetic interference and are thinner & lighter allowing for easier placement and routing.

Using optical solutions, nearly anything that is connected using PCIe today can now be connected remotely.   This allows users to leverage the ubiquity of PCIe for many applications such as memory/disk system interconnects, high-end audio/video applications, high performance computing and multi-chassis system interconnects.

Optical solutions can be a key enabler for network and computer architects who see value in using PCIe as an I/O technology for data center connectivity. Using PCIe to natively connect servers, switches & storage elements can lower overall system costs by reducing or eliminating the number of protocol conversion chips. In addition, this increased system simplicity provides advantages from a latency, power and dollars per gigabit perspective.

For the physical connection, PCIe 3.0 will need parallel solutions of at least eight lanes. The most natural fit is using parallel optical Transmit/Receive (Tx/Rx) modules using vertical cavity surface emitting laser arrays and provide up to 150 meters of connectivity. Twelve channel Parallel Optic devices capable of operating at 8Gb/s per lane are commonly available with multiple mechanical form-factors available today from Avago Technologies including MiniPOD™, Micro-POD™, CXP, PPOD and QSFP+.

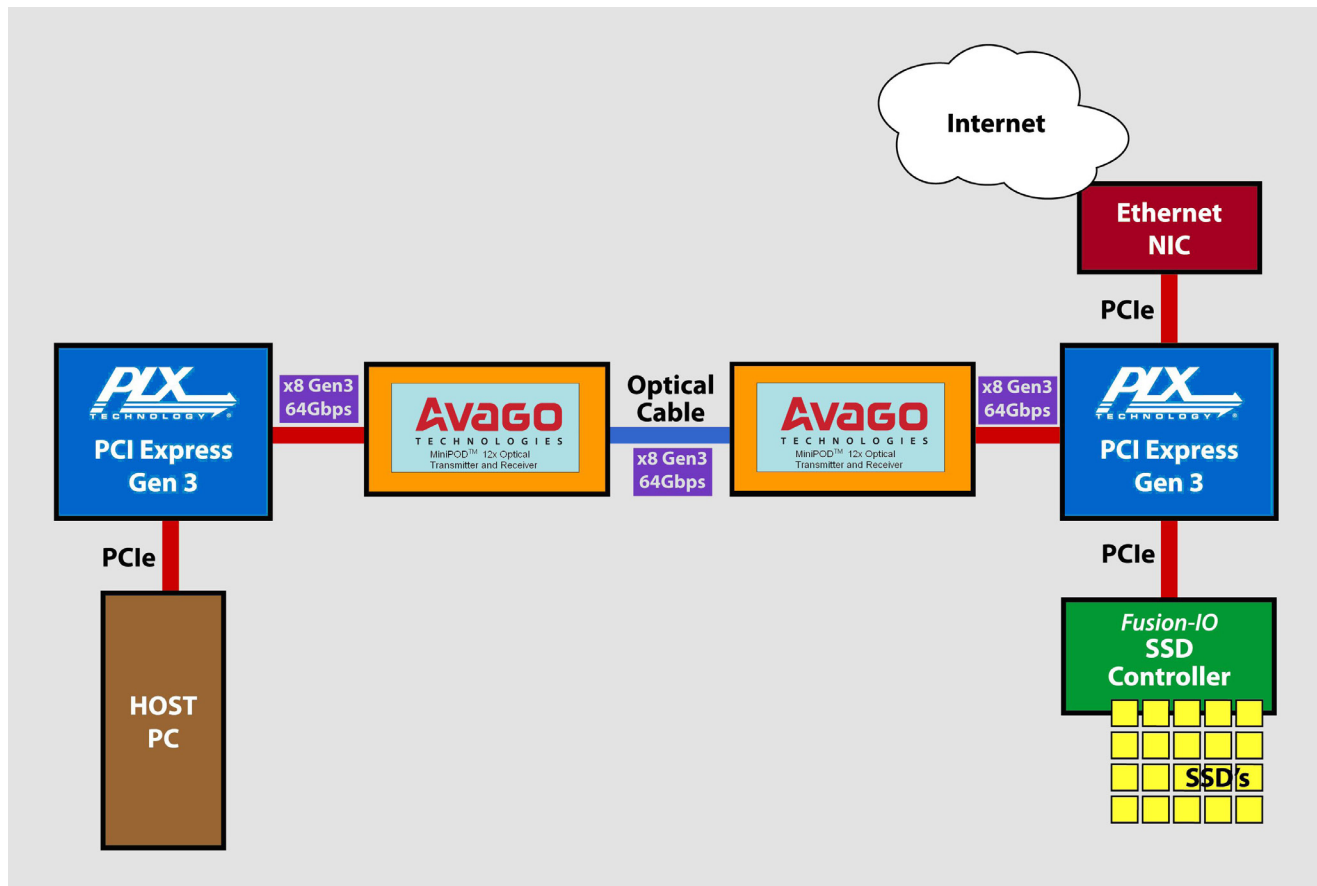## PLX Technology and Avago Technologies Demonstration

Figure 1. Block Diagram Optical System

The proof of concept demonstration consisted of a host PC housing a PLX designed adapter card, employing the PEX8748, 48 lane Gen3 switch. Shown in Figure 2, the card contains a daughter mounting assembly for which the AFBR-81/82 Series optical transmitter and receiver modules are mated to the PEX switch. At the opposite end of the optical link, a second switch card with another pair of AFBR-81/82 Series optical transmitter and receiver modules reside on a distribution board to provide fan-out and upstream data aggregation for express peripherals, such as SSD drives and Ethernet HBA cards.
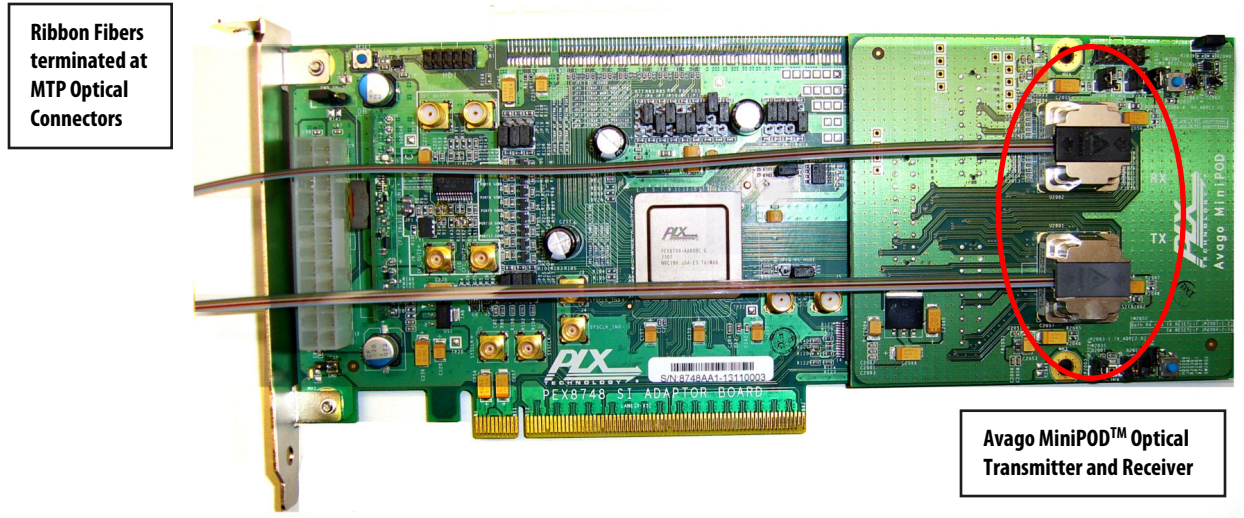

Ribbon Fibers terminated at MTP Optical Connectors

Avago MiniPOD™ Optical Transmitter and Receiver

Figure 2. PEX8748 SI card with Avago Technologies MiniPOD™ adapter

Demonstration Test Results

In this demonstration a PCIe Gen3 x8 link was successfully implemented over 30 meters of OM3 multi-mode optical fiber. Below are representative examples of eye quality plots, taken at the PLX receiver for cable lengths of 10 and 30 meters. These eye quality plots show good signal integrity and error free data is recovered, even after 30 meters of fiber.
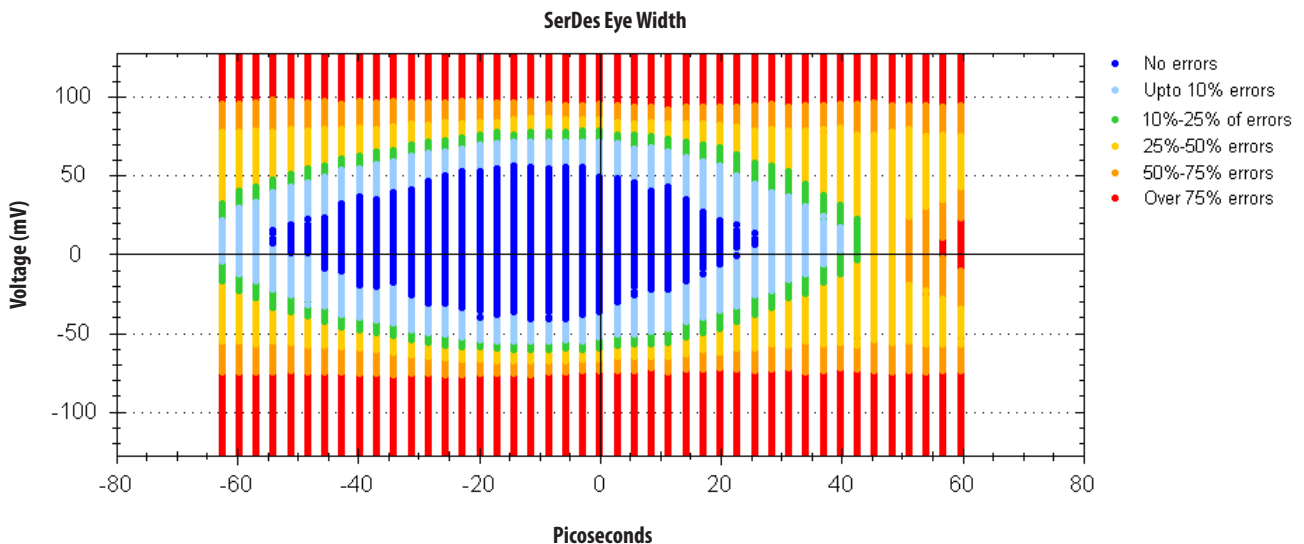

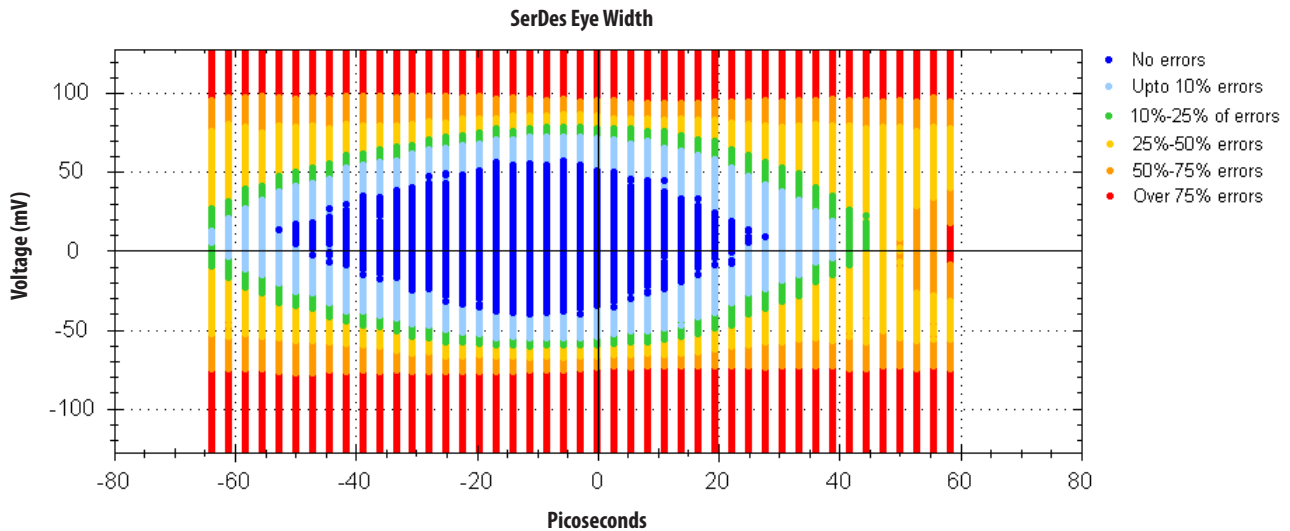Figure 3. Eye measurement at PLX switch (10 Meters)

3

SerDes Eye Width

Figure 4. Eye measurement at PLX switch (30 Meters)

PCIe Functionality supported with Optics:

The demonstration shows PCIe as an active, eight lane aggregation pipe. In many such applications, the lane remains active at all times and much of the PCIe ASPM (Active State Power Management) functions are not used. As implemented, the link supports the following PCIe functionality:

• Asynchronous operation (no native SSC, but SSC isolation provisions)

• L0 active state only (Link enable/disable functional under controlled operating system)

• PCIe normal link speed negotiation.

• Configurable for PCIe standard link width down training.

It does not support:

• PCIe Active State Power Management

• In-band synchronous resets (Out of band independent reset only)

How Does PCIe Work over Optics?

The standard PCIe protocol standard specifies two features that are typically difficult to implement when using re-drivers and optical links. Receiver Detection and Electrical Idle Control.

Receiver Detection

Receiver Detection is the mechanism by which a transmitting device can determine if there is proper loading of the transmission line. Where proper loading exists, the transmitter is triggered to operate in one of several modes based on what is detected at the device receiver. Optical links typically present a 50 ohm termination at all times. As such, optical links may not accurately indicate the presence of a valid PCIe receiver at the opposite end of the link. The PEX family of switches have the ability to mask receiver detection and instead use the decoding of incoming data as a means for link negotiation.

Electrical Idle Control

Electrical Idle is a second key parameter of concern with optical links. Electrical Idle indicates when a link has entered a temporarily inactive state (quiesed) and can herald entry/exit into low power states and link speed changes. To be PCIe compliant, the link transmitter must both be in a static state and also hold the transmission line to a fixed common mode. Line chatter or improper bias can lead to false EIDLE detection and/or exit from the EIDLE state. Within the PEX family of switches, provisions are made to ignore the standard electrical idle stimulus, while tracking specific data symbols so as to allow link speed negotiation. Links are able to enumerate, communicating proper link, lane and speed information and come to full operational bandwidth. However, the modified EDILE functionality does not allow for Active State Power Management entry and exit. The active port can be physically disabled and re-enabled and link retraining will occur. However, registry management, (i.e. resource reallocation, BAR programming, possible register re-programming of endpoints, message blocking etc) needs alternate handling routines. Several resources for message handling are provided within the PEX family of switches.

Typically, when employing optical fiber, both ends of the link will not reside in the same enclosure. This means they will not share the same Reset nor the same system clock. In the demonstration described above, there is no means to implement a synchronous reset or clock across the link. The PEX family of devices is capable of using asynchronous clock operation. In these configurations, system Spread Spectrum Clocking (SSC) cannot be active. However, because the interface is optical, there is a reduced need for EMI suppression of the link. In cases where system emissions are high, the PEX family of devices with SSC isolation feature allow the system to operate with lower emissions, yet keep the optical link in a constant frequency mode. Contact a PLX FAE for design notes on asynchronous operation.

The remote optical card can undergo a separate reset upon power up and is ready for link training once the host box becomes active. Typically, the remote box is powered ahead of the system box (Server/PC). Alternately, if the operating system (OS) is under full control of the user, the systems can be powered up in any order. Once both systems are confirmed to be powered (for example, the customer user software checks for link status) the OS can initiate standard system enumeration/programming methods. Contact PLX for reference design material.

While the high-speed 8Gbps signals are of primary importance, it is worth noting the pre-existing PCIe external cabling specification also defines extra signals that will not be carried in the AFBR-81/82 Series optical solution. For instance, CREFCLK, the 100MHz Cable Reference Clock is not needed since the clock is recovered from the datastream by the PCIe transceivers. In addition, the following electrical power pins are not applicable when using an optical cable.

a. SB_RTN   – the Signal return for single-ended Sideband Signals
b. CPRSNT# – the Cable Installed/Downstream Subsystem Powered-up
c. CPWRON – the Upstream Subsystem's Power Valid Notification
d. CWAKE#   – the Power Management Signal for Wakeup Events
e. CPERST   – the Cable PERST# Cable Platform Reset pins

Optics as the Physical Media

In the demonstration, the Avago Technologies MiniPOD™ AFBR-81/82 Series (12 lane, 10Gbps, Tx/Rx) are used to demonstrate the high-speed performance of PCI Express Gen3.0.

For this proof-of-concept demonstration, eight of the optical lanes are made active and four lanes are left unused.

Each end of the physical link is terminated using a PLX PCIe Gen 3 Switch IC. PLX PCIe switches include both clock/data recovery and TX/RX equalization for each high-speed port. Clock and Data Recovery (CDRs) are not required in the optical module and as a result, the latency advantage of PCIe is preserved. The MiniPOD™ optical modules operate at PCIe 3.0 8.0Gbps per lane for the demonstration but are capable of operation over a wide range of line rates from 1Gpbs to over 10.3125Gbps. As a result, these optical devices can operate at PCIe 2.x at 5.0Gbps and PCIe 1.x at 2.5Gbps operation without configuration changes and without any tradeoff in performance.

As shown below, the MiniPOD™ optical modules are coupled via a flat ribbon cable that is terminated using an industry standard MTP connector.
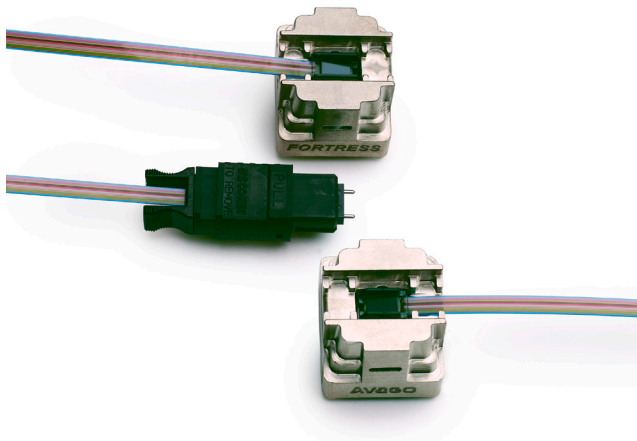
**Figure 5. MiniPOD™ 12-channel embedded parallel optics (Transmit/Receive)**

The MiniPOD™ devices use a Meg-Array connector allowing module detachability from the host/adapter card assembly. The PLX IC TX and RX signals directly interface to the Avago Technologies modules, with only AC-coupling required between the two elements. To maximize performance the electrical I/O ports of the MiniPOD™ module allow for fixed equalization/emphasis and amplitude adjustment. Adjustments can be made using a two-wire serial control on a per lane basis for each optical module.

There are a number of inherent advantages in using embedded parallel optics devices such as the MiniPOD™. Embedded parallel optics are defined as modules that are not mounted at the board edge but rather directly onto the PCB in order to gain advantages in size/density, signal integrity, and EMI immunity.

The MiniPOD™ has a footprint of 18.6 mm x 22 mm and a height of either 14.5 mm for a flat fiber cable housing or 15.6 mm for the round cable housing. Advantages of flat ribbon cables include lower profile, and dense tiling of modules, while round cables offer ruggedization and more flexibil¬ity in fiber routing. The MiniPOD™ uses a low-cost optical turn connector, called Prizm™ connector and the footprint uses the industry-standard 9x9 pin Meg-Array connector used for the electrical interface.

High-speed electrical signals at 8 Gbps progressively degrade over trace lengths in PCB materials due to capacitive skin effects. Designers utilize high performance PCB materials, transmitter de-emphasis, receiver equalization, and CDRs to address these signal integrity issues. These techniques add cost and power to the overall solution. Furthermore, if system designs use edge-mounted solutions, such as optical or electrical PCIe cables the PCB trace lengths can be very long, 12-20 inches or more. Designs using embedded parallel optics, such as the MiniPOD™ can minimize these long traces by being easily positioned within five inches of the high-speed IC output.

Furthermore, embedded optic cables can terminate to an MTP connector at the front panel, thus occupying must less front panel space than a comparable PCIe or CXP edge-mounted solution. By being a fiber optic solution, an PCIe Gen 3 signal can easily traverse over 100m of multi-mode fiber, while being unaffected by Electromagnetic Interference (EMI) and cross-talk which is a common affliction when using copper cables. Figure 6 shows the MiniPOD™ solution versus a PCIe copper cable. The embedded optical module solution is superior in size, bend radius, ease of installation, signal integrity and EMI immunity.
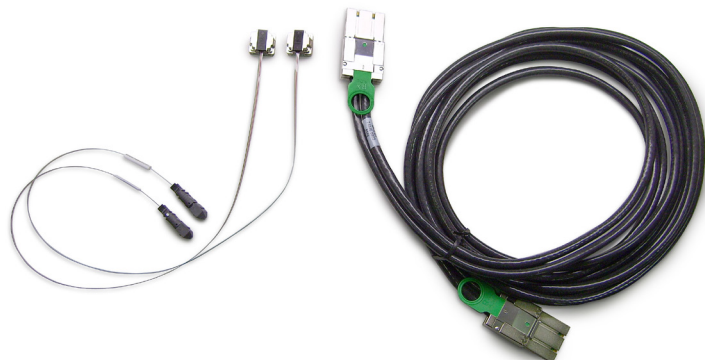
Figure 6. MiniPOD™ embedded parallel optics with cable size comparison versus PCIe copper cable
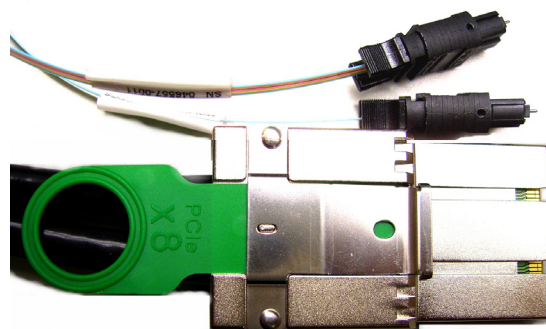
Figure 7. Size comparison of an MTP parallel fiber optic termination versus a PCIe copper cable

Below is a table of typical operating specs for the MiniPOD™ optical link. For detailed specifications consult Avago Technologies.
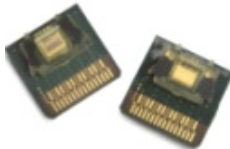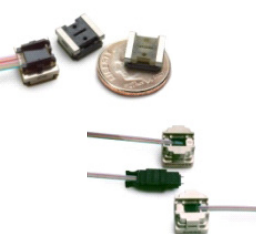
Parameter	Value	Units	Notes
Optical Wavelength	850	nm	VCSEL technology
Example Data rates per lane	10.3125 8.0 5.0 2.5	Gbps	100GbE PCIe 3.x PCIe 2.x PCIe 1.x
Number of operational lanes	12 10 8		Twelve lanes available 100GbE utilizes the middle ten lanes PCIe x8 can use the first eight lanes, or the middle eight lanes
Link Length	100 @ 10.3125Gbps	m	OM3, 2000 MHz.Km 50μm MMF Operation can be >100m @ 8.0Gbps
	150 @ 10.3125Gbps	m	OM4, 4700 MHz.Km 50μm MMF Operation can be >150m @ 8.0Gbps
Operating Temperature Range	0-70	°C	Case Temperature
Power Supply Voltage	3.3 & 2.5	V	
Management Interface	Two-Wire Serial		Compatible with industry standard Two-wire serial protocol scaled for 1.2 volt LVCMOS. Can also tolerate 3.3V LVTTL.
Output Optical Power: Average	+2.4	dBm	With optical connector and pigtail attached at the fiber output
Electrical Interface	Meg-Array		1.27mm pitch and 4mm contact mate height

Market Trends in Optics

Traditionally, optical transceivers have had cost as an implementation barrier. Avago Technologies, is a vertically integrated high-volume manufacturing and leader in the optical transceiver industry with over 30 years of experience and the highest quality levels in the industry. As such, Avago Technologies is cognizant of the long term trends and cost versus performance tradeoffs. As signaling speeds increase, copper solutions are increasingly challenged from a total system installation and cost perspective. At speed of 8 Gbps and above, optical solutions can be performance and cost competitive to active electrical cabling options.

Avago Technologies has a host of connectivity options. The AFBR-75x series products for example, address the trend towards smaller, shorter distance solutions as needed in many consumer applications. The MiniPOD™ solution addresses the need for longer distance links with higher lane counts suitable for many Data Center and high performance computing applications.

Avago Technologies' 10Gbps per Optical Channel Solutions

Platform	AFBR-75 Series	MiniPOD™/MicroPOD™	PPOD	SFP+/ QSFP+/CXP
Format	1 or 2 Channel	12 Channel Tx & Rx Modules	12 Channel Tx & Rx Modules	1 Channel Transceiver 4 Channel Transceiver 12 Channel Transceiver
Approx $/Gbps	< $1	<$5	<$8	<$12
Application/ Key Feature	Low cost Small Form Factor High density Limited Distance	High density, high speed (100G+) Small Form Factor At least 100m distance	High density, high speed (100G+) At least 100m distance	Pluggable / Pay-as-you-go, flexibility At least 100m distance
Datarate	10.3125Gbps	10.3125Gbps	10.3125Gbps	Varies
Power mW/Gbps	<15	<25	<25	<30

Embedded Solutions   **Pluggable Solutions**

Summary

An optical solution carrying high-speed PCIe 3.0 exists today using solutions from PLX Technology and Avago Technologies. A similar optical implementation of this kind will warrant some careful consideration by the design engineer. And while some features in the PCIe Gen3 standard are not directly compatible with fiber optic technologies, the features available in the PLX PCIe Gen3 switch products and the Avago Parallel Optics can be used to enable system architects to extend the reach of PCIe Gen3 outside-the-box.

The technology demonstration discussed here combines the advantages of PCIe (scalability, low latency, flexibility) and parallel optical interconnects (reach, bandwidth, distance, and low power).

For additional technical discussion please contact your field representative from Avago Technologies and PLX Technologies.

PLX Technology

PEX 8748 48-Lane, 12-Port PCI Express Gen 3 (8 GT/s) Switch, 27 x 27mm FCBGA

The ExpressLane™ PEX 8748 is a 48-lane, 12-port, PCIe Gen 3 switch device developed on developed on 40nm technology. PEX 8748 offers Multi-Host PCI Express switching capability that enables users to connect multiple hosts to their respective endpoints via scalable, high-bandwidth, non-blocking interconnection to a wide variety of applications including servers, storage, communications, and graphics platforms. The PEX 8748 is well suited for fan-out, aggregation, and peer-to-peer traffic patterns. Included is PLX's proprietary visionPAK debug software, which allows, for example, internal receive-eye observation after equalization and access to the devices' internal debug registers thus enabling faster time to market.

Avago Technologies

MiniPOD™ parallel optics, AFBR-82 Series & AFBR-81 Series

MiniPOD™ parallel optics are 12-channel embedded parallel optics transmitter and receiver modules support lane rates of up to 10 Gigabits per second (Gbps) for an aggregate bandwidth of up to 120 Gbps. The small-footprint modules feature a low-cost, removable fiber cable connection and a pluggable electrical connection that provide flexible cable management at installation, simplifying design and lowering cost for switching and supercomputing applications.

The new MiniPOD™ optical modules incorporate Avago Technologies 850-nm Vertical-Cavity Surface Emitting Laser (VCSEL) technology, Avago Technologies PIN array technology, and Avago Technologies integrated laser driver and receiver IC technology, which combine to provide robust electrical and optical performance at these high data rates. Using separate transmitter and receiver modules provides design flexibility and lowers the total solution cost for the optical interconnect. Incorporating programmable equalization and de-emphasis into the modules' highly compact 22- by 18.5-mm form factor allow system designers to optimize dense board layouts with superior signal integrity and system margin.

CXP optical modules, AFBR-83PDZ

The Avago Technologies AFBR-83PDZ is a twelve-channel, pluggable, parallel, fiber-optic CXP transceiver for high-speed, high-density optical interconnect applications. This device integrates twelve data lanes in each direction with greater than 120 Gbps aggregate bandwidth. Each lane can operate at 10.0 Gbps or 10.3125 Gbps up to 100 m using OM3 fiber. These modules are designed to operate over multimode fiber systems using a nominal wavelength of 850 nm.

AFBR-83PDZ is compliant to SFF-8642: MINI MULTILANE SERIES: SHIELDED INTEGRATED CONNECTOR. The electrical interface uses an 84-contact edge type connector. The optical interface uses a 24-fiber MTP (MPO) connector. Each of the twelve channels is compliant on a per lane electrical and optical specification basis with the 100 GbE 802.3ba (100GBASE-SR10 and CPPI) specification. Furthermore, each of the twelve channels is also compliant to the InfiniBand Architecture Release 1.2.1 QDR Specification.