PCI Express Optical Repeater for Server I/O Expansion

Yongseok Choi, Cheol-Hoon Lee

Abstract*—We suggested a new PCI Express Optical Repeater Device which can locate I/O Devices externally for I/O expansion, not locate all I/O Devices in Server that has limit in Volume and Power in this paper. This Device doesn't convert PCI Express signal to Optical Signal directly, but in PCI Express Switch Architecture, uses the layered architecture to make Switch's internal bus or interconnection the optical interface. Thus there is no disadvantage of wasting bandwidth. This solves signal attenuation and over-power dissipation for the sake of Optical Interface Characteristics and can meet distances requirement for Server Environment. We expect that the methods and research results in this paper can be used in Small Server which will want I/O Expansion

***Index Terms*—PCI Express, Optical Repeater, Server, I/O, Interconnection**

I. INTRODUCTION

In contemporary server architecture, the I/O interface lead to be problematic acting bottleneck point as the processing power and the I/O complexity are increased due to required data amount from the use of multi-core CPU and large amount of memory and required bandwidth of peripheral I/O. For solving this problem, the external interfaces of the server device such as Ethernet, IEEE1394, USB, etc. have already been transformed from conventional parallel interfaces to serial interface of the system, and the interface of the high-performance device inside the server has been solved to some extent by PCI Express. Since other interfaces than the PCI Express interface, require a relatively low bandwidth, those do not have a significant constraint on the distance and the bandwidth even with an conventional electrical signal, but the PCI Express interface that is used inside the server requires the high bandwidth (up to 2.5 Gbps or higher) and the plurality of transceiver for multi-lane transmit and receive operation. Because the server's main board size is limited, it has the limitation of the size of the peripheral device that can be mounted, and the device's power consumption, based on the power which can be supplied from the main board, also has the limitation, so for the devices requiring high-power, solving this problem using a separate source from the power supply, but this has also been a factor that limits the weight and size of the server. Our goal is to present the PCI Express optical repeater for Server I/O expansion for solving previous problems.Organization of paper as follows: Section 1 is introduction, section 2 is related work, section 3 is proposed repeater architecture, section 4 gives experimental result, and section 5 is conclusion.

Revised Version Manuscript Received on November 03, 2015.

Yongseok Choi, Senior Engineer, SW Contents Research Laboratory, Electronic and Telecommunication Research Institute, Daejeon, Republic of Korea.

Prof. Cheol-hoon Lee, Professor, Department of Computer Science & Engineering, Chungnam University, Daejeon, Republic of Korea.

II. RELATED WORKS

To connect a relatively large volume, a high power, or a plurality of PCI Express devices to the server, I/O expansion is required. so for this, the PCI SIG was published External Cabling Specification [1], but it continues to present a problem from the thickness and the size of the connector and the cable and also has many problems such as the skew between the signal, the constraints of reach, and the restriction for power consumption.

To solve the problem of the size and weight of the cable connector, Intel and Apple have announced new standards called Thunderbolt, but also has continued restriction on the distance due to electrical signal attenuation and signal skew problem.

Canadian company Adnaco, using two or more switch chipset, made an optical connection in place of the signal between the switch chips [2], but in this case, it has the disadvantage that the buffering steps are increased from using multiple switches and cannot use full bandwidth of the optical connection, instead, only use the signal bandwidth of the PCI Express.

Avago also demonstrate the optical connection between PCI Express 3.0 compliant PLX Switch chips, but also has the disadvantages that cannot use full bandwidth of the optical transceiver as Adnaco's product. Because the contemporary signal bandwidth of PCI Express is 8Gbps for PCI Express Gen 3, the bandwidth wasted in the 10Gbps optical connection is not much large, but a waste problem in the future 25Gbps ~ 50Gbps optical connection and beyond 16Gbps PCI Express signaling environment, will be expected to become more severe.

In this paper, we present a method of making an optical connection for extending PCI Express, without wasting bandwidth, including repeater structure that minimize the buffering delay due to the number of switches used and can increase the scalability, and give demonstration using FPGA to verify its feasibility and show the results for the experimental data. The details of this study are expected to be valuable for small servers, etc. that require the expansion for PCI Express I/O.

III. PROPOSED PCIE OPTICAL REPEATER

A. Server System including PCIe Optical Repeater

Server systems, including PCI Express optical repeater, is shown in Figure 1.

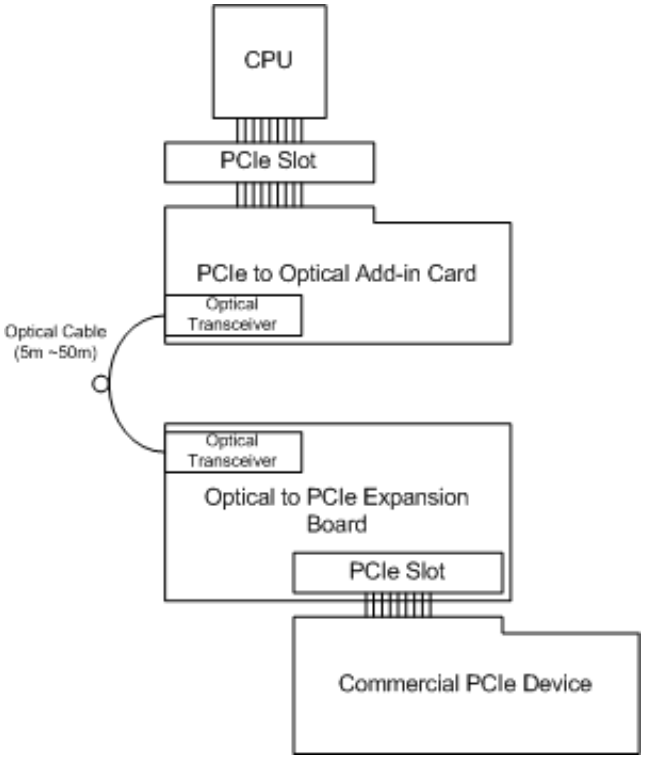

Figure 1 Server system including PCIe Optical Repeater

Server system, since it is mounting commercial PCIe devices directly to the PCIe slot supplied from the CPU, limits mountable device on the limited server area, and limits the device power when using the power supplied from the server.

With PCI Express optical repeater device consisting of PCIe to Optical add-in card and Optical to PCIe expansion board in figure 1, the small add-in card is mounted in the server, optical connection to expansion board is made using relatively small form factor connector, and commercial PCIe device can be mounted on expansion board, so PCIe expansion board has freedom for attaching any PCIe device regardless of the size of the device. In addition, because it can have a power supply unit for supplying the extension board, separately, the use of commercial PCIe devices with high power is also possible.

When accommodating conventional PCIe cable, it uses the add-in card for connecting the signal to the connector just after the impedance match with extracting the direct signal from the PCIe slot, so it is limited to the length and type of cable against the degradation of the PCIe signal, but when using the method proposed in this paper, it is possible to send and receive signals in ranges from some metres to tens of km in accordance with the type of the optical transceiver and the cable. Figure 1 shows a length of the cable from multi-mode fiber transceiver required for the actual server environment and it can be configured relatively inexpensive VCSEL.

Optical repeater were constructed by applying the general PCI Express switch device and the following Figure 2 is a diagram showing the internal structure of a typical PCI Express switch. [4]

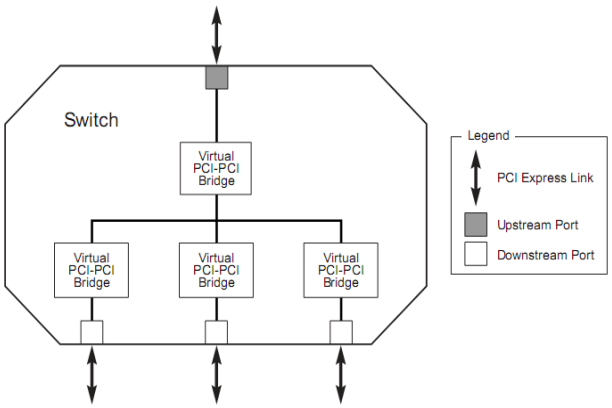

Figure 2 PCIe Switch Typical Internal Representation

PCI Express switch is configured as a virtual PCI-PCI bridge connected to the upstream port and a virtual PCI-PCI bridge connected to a downstream port and these are connected to each other using an internal bus. Given PCI Express optical repeater constitutes a virtual PCI-PCI bridge on the upstream port in add-in card form part and a virtual PCI-PCI bridge on the downstream port in an expansion board part form and changes internal bus to optical interface and the each part information will be described in detail in the following sections.

B. PCIe-to-Optical Add-in Card

The configuration of PCIe-to-Optical add-in card that is attached to server motherboards is shown in figure 3.

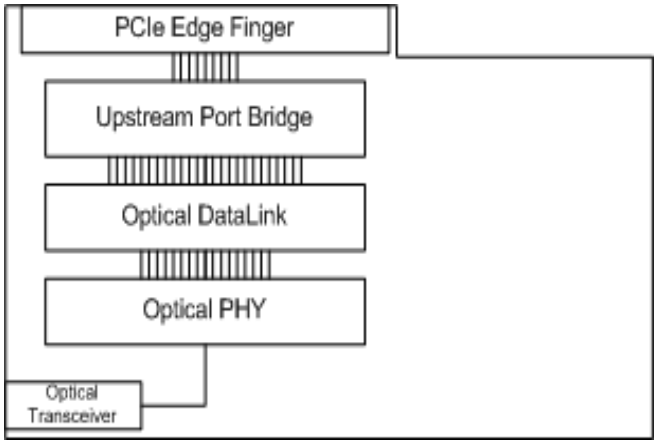

Figure 3 PCIe-to-Optical Add-in Card

PCIe edge finger for mounting the PCIe slot are connected to the upstream port bridge, corresponding to the virtual PCI-PCI bridge directly connected to the upstream port as figure 2. In figure 2, it is not clear the connection between the upstream port and the virtual PCI-PCI bridge, but upstream port bridge is made from the structure of PCIe layering that is composed of physical layer, data link layer, and transaction layer to perform the virtual PCI-PCI bridge function. PCIe-to-Optical add-in card can pass packets between the PCI Express interface and Optical interface using separate optical data link layer and physical layer, rather than to covert a PCI Express signal directly to the optical signal,

so it can utilize full bandwidth of optical interface for packet transmission and reception and the waste of bandwidth that results from the difference between the bandwidth does not exist.

C. Optical to PCIe Expansion Board

Configuration of Optical to PCIe Expansion Board for PCIE expansion slots are shown in the following figure 4.

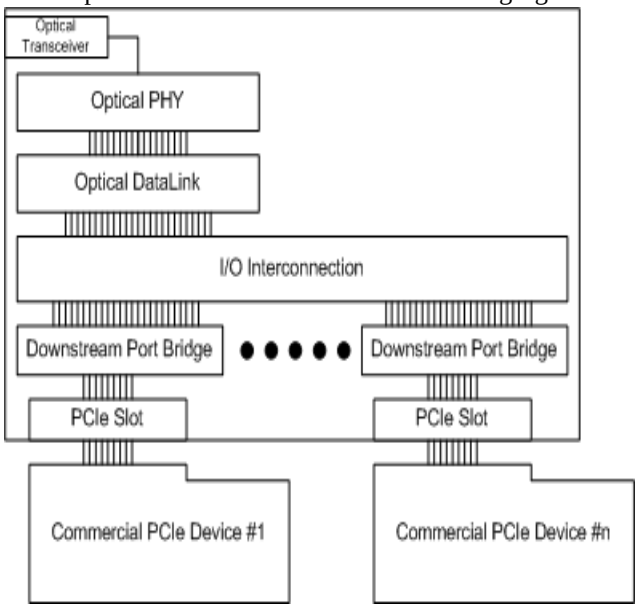

Figure 4 Optical-to-PCIe Expansion Board

Packets received by the optical transceiver is transmitted by an optical data link layer through the optical physical layer, which is transmitted to an appropriate downstream port bridge via I/O interconnection. I/O interconnection can forward the packet to an appropriate downstream port bridge after determining the right downstream port bridge, or to reduce the complexity it can forward the packet to all downstream port bridge and the downstream bridges determines whether the packet is corresponding to it to discard or process the packet.

In Figure 4, since it has the multiple downstream port bridge, I/O interconnection is required, but when there is only one downstream port bridge, optical data link layer is connected directly to the downstream port bridge without any I/O interconnection.

Downstream bridge has PCI Express hierarchy of transaction layer, data link layer, and physical layer to perform all of the work of the PCIe protocol and the transaction layer is configured to perform a virtual PCI-PCI bridge function shown in Fig 2.

As the operation of the PCIe-to-Optical add-in card shown in Figure 3, Packet received at the optical transceiver is transferred to downstream bridge via optical physical layer and optical data link layer and packet received at PCIe slot is also transferred to optical link through optical data link layer and optical physical layer to transmit and receive packet, and also the bandwidth wasted bandwidth resulting from the difference does not exist.

IV. EXPERIMENTAL RESULT

A. Optical Interconnection test between Upstream bridge and Downstream Bridge

The logic for the present study was implemented using Altera Stratix II GX FPGA development board with two optical transceivers and PCIe edge fingers, which is shown in the following figure 5.

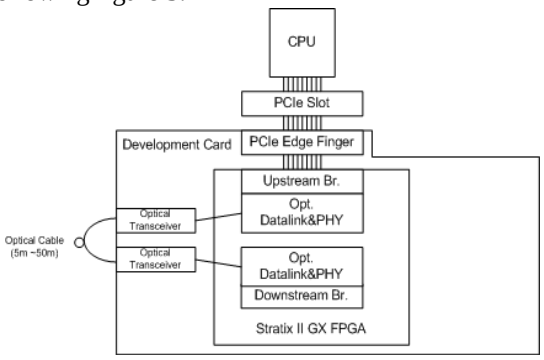

Figure 5 Test environment for Optical Interconnection

Since Stratix II FPGA development board with PCIe slots for the experimental environment has only one interface, we confirm the operation of the downstream port bridge optically connected to the upstream bridge.

To check for the downstream port operation, we confirmed that the Configuration packets received over the PCIe slot is transferred to the downstream bridge via optical transceivers and ultimately the downstream bridge is recognized as the device. Figure 6 and 7 shows its actual implementation feature and experimental result.

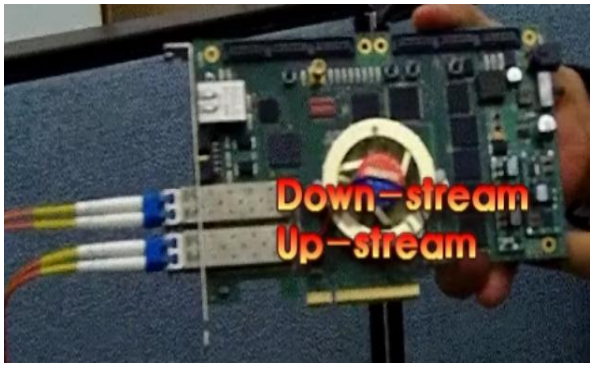

Figure 6 Implementation feature

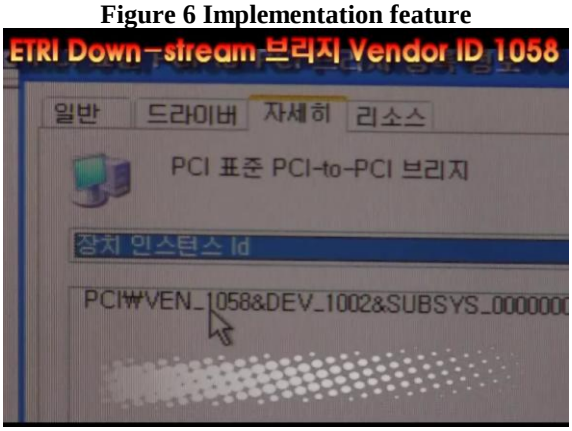

Figure 7 Experimental result

Figure 7 shows that the recognized downstream bridge has ETRI(Electronic Telecommunication Research Institute)'s vendor ID(1058h) and device ID(1002h) assigned by us.

Now the downstream bridge should be checked to operate as PCI Express upstream device. Since Altera Stratix II GX FPGA development board has only PCI Express edge finger but doesn't have PCI Express Slot as described above, we need some method to connect PCI Express optical repeater via adjusting the current configuration and this system will be described in the following subsection.

B. PCI Express downstream bridge test using separate PCIe Extension system

To check the operation as a PCI Express upstream device's PCI Express downstream bridge, we changed the optical interconnection between the upstream bridge and the downstream bridge to direct connection from inside the FPGA(removed the optical transceivers and made direct connection between the logical part of the physical layer) and assigned one optical transceiver to the electrical part of PCI Express physical layer of the downstream bridge and connected the optical fiber cable attached to the downstream bridge to PCI Express expansion system with optical transceiver interface. This system is shown in the following figure 8.

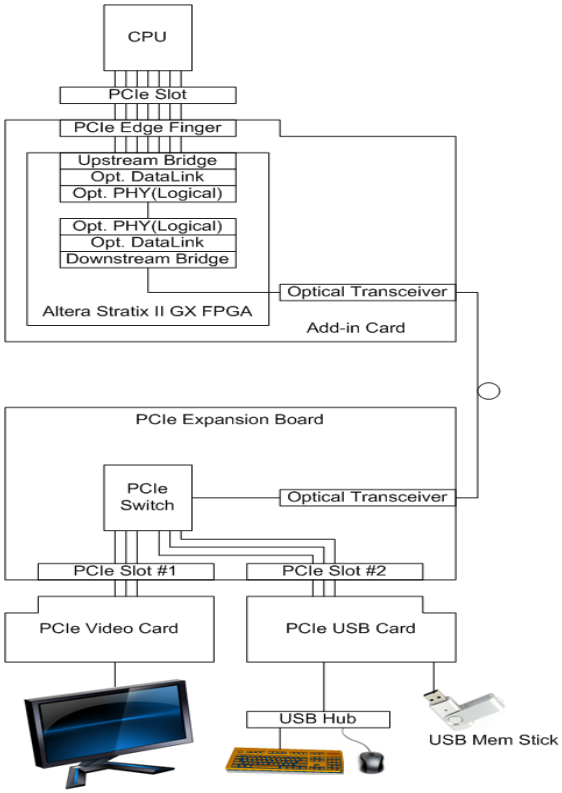

Figure 8 Downstream Bridge Test System

PCIe Switch present in the downstream bridge serves to distribute the PCIe signal received through the optical transceiver to the at least two PCI Express slot and PCIe video card and PCIe USB card were installed on each PCI Express slot, respectively. The stored video file in USB Memory Stick is delivered through PCIe USB Card to the host with Add-in card and the host processes the stored video file and delivers the video frame data via optical connection to the PCIe Expansion Board in which PCIe Video Card displays it. In addition, to access host CPU, PCIe USB Card has Keyboard and Mouse via USB Hub on it.

The actual operating system is shown in Figure 9.

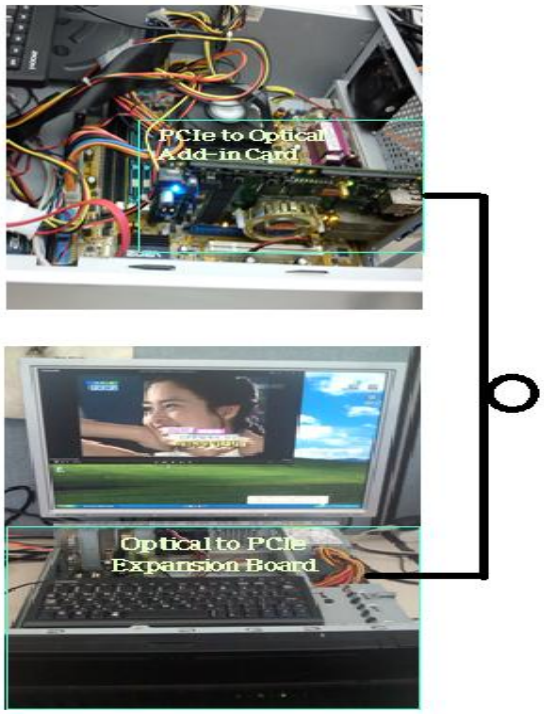

Figure 9 Actual System expanding PCIe via Optical Interconnection

As a result, the videos are stored on the USB Memory Stick made smooth playback on the video card.

V. CONCLUSION

We provided a way to use PCI Express switch internal bus as optical interconnection and, to verify feasibilities, using a FPGA development board, we give the result that Downstream bridge is fully recognized as the bridge device in PCI Express. To demonstrate downstream bridge's operation, we use separate PCIe Expansion board to make video application and we get the good result in displaying good quality video representation. The research results of the paper is expected to be utilized for having a value sufficient for the I/O server which has a limited extension in terms of the size and power.

ACKNOWLEDGEMENT

This work was supported by the ICT R&D program of MSIP/IITP, Korea. [10038764, Silicon Nano Photonics Based Next Generation Computer Interface Platform Technology]

REFERENCES

1. PCISIG, "PCI Express External Cabling Specification Revision 1.0", pp53-134, 2007
2. http://www.adnaco.com
3. Avago, "A Demonstration of PCI Express Generation 3 over a Fiber Optical Link", 2012
4. PCISIG, "PCI Express Base Specification Revision 2.1", pp43, 2009