J Inf Process Syst, Vol.14, No.1, pp.270~279, February 2018
https://doi.org/10.3745/JIPS.01.0022

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Implementation of Multipurpose PCI Express Adapter Cards with On-Board Optical Module

Kyungmo Koo*, Junglok Yu*, Sangwan Kim*, Min Choi**, and Kwangho Cha*

Abstract

PCI Express (PCIe) bus, which was only used as an internal I/O bus of a computer system, has expanded its function to outside of a system, with progress of PCIe switching processor. In particular, advanced features of PCIe switching processor enable PCIe bus to serve as an interconnection network as well as connecting external devices. As PCIe switching processors more advanced, it is required to consider the different adapter card architecture. This study developed multipurpose adapter cards by applying an on-board optical module, a latest optical communications element, in order to improve transfer distance and utilization. The performance evaluation confirmed that the new adapter cards with long cable can provide the same bandwidth as that of the existing adapter cards with short copper cable.

Keywords

Device Network, Interconnection Network, On-Board Optical Module, PCI Express Bus

1. Introduction

PCI Express (PCIe) bus has been sincerely played a role of I/O bus to connect various devices in a system. It has also progressed by interoperating with CPU, chipset, and other units, in order to accommodate various devices, achieve high-performance bandwidth, and provide diversified services [1]. As PCIe switching chip, which was used for accommodating various end point devices, now provides communications between servers based on non-transparent bridging (NTB) [2] or Tunneled Window Connection (TWC) [3], PCIe bus can be used not only as a connector between devices but also as an interconnect network.

With the expanded use of PCIe bus, features of PCIe adapter cards required to improve. For example, transfer media of adapter cards need to change, considering the weak points of cooper wire cables like SAS cables in terms of transfer distance and cabling method.

In this paper, we propose and introduce prototypes of new PCIe adapter cards designed to extend the PCIe bus for use in the interconnect network for high-performance computing system. Because the cards are based on the PCIe bus technology, they can be used for connection with the remote devices, as well as for interconnection between servers.

Manuscript received August 18, 2017; first revision October 11, 2017; accepted November 6, 2017.
Corresponding Author: Kwangho Cha (khocha@kisti.re.kr)
* Dept. of Supercomputer System Development, Korea Institute of Science and Technology Information, Daejeon, Korea ({kookm, junglok.yu, sangwan, khocha}@kisti.re.kr)
** Dept. of Information and Communication Engineering, Chungbuk National University, Cheongju, Korea (mchoi@cbnu.ac.kr)

Until now, the optical PCIe adapter card was usually based on PCIe Gen2 and legacy optical modules such as QSFP. Unlike legacy optical modules, the on-board optical module has small size form factor and guarantees high speed. This is the one of the reason why the on-board optical module is regarded as earlier model of silicon-photonics prototype. In this study we used the newest on-board optical module, the vertical-cavity surface-emitting laser (VCSEL) type on-board optical module, to check its usability in PCIe Gen3 adapter cards.

In addition to using the latest optical modules, we designed the adapter cards can provide multi-purpose functionality. There are different adapter card types according to the location and the usage of them. Our adapter cards, however, can be used as multipurpose by changing their settings.

This paper is organized as follows: Section 2 discusses key technologies for implementing our PCIe adapter cards. In Section 3, we describe the basic structure of the prototype adapter cards. In Section 4, we analyze the performance of our proposed implementation. Finally, the conclusions are presented in Section 5.

2. Background

2.1 PCI Express Switching

The PCIe bus is the typical I/O bus widely used in recent computer systems to physically connect various devices in a single system. It adopts the PCIe bridge chip in order to overcome the limit of distance of electric signals and to allow connection with many devices. The PCIe switching chip is an advanced type of PCIe bridge chip, which not only extends the electric signal, but also provides special functions, such as isolation of devices and conversion of the address system [2,3]. Our prototype of new PCIe adapter cards can provide multiple functions by using the PCIe switching chip.

It is also suggested that PCIe can be used as data center network because of the advantage of PCIe switching processor [4]. In order to prove their proposed scheme, they used NTB for connecting servers. Because NTB uses specific system functions, it is required to develop and improve more general communication library for NTB. It was reported RDMA, one of the most powerful transfer method, could be implement on NTB [5]. PEACH2 [6] is an intra-node communication system based on PCIe and it is used for connecting accelerators in a node. PEACH2 also shows it was possible to use FPGA as a PCIe switching processor instead of commercial one. PCIe also can be used for connections such as FPGA-to-Host or FPGA-to-FPGA [7].

2.2 On-Board Optical Module

Recently, optical modules are developed considering high-speed transmission and compactness, influenced by the silicon photonics technology. In particular, the latest on-board optical modules are surface-mounted directly on PCB. Foxconn's MiniPOD [8] and MicroPOD [9] and Samtec's PCUO (PCIe over FireFly optical cable system [10]) are included in the type.

The optical module used in the adapter card is the PCUO module of Samtec's FireFly optical module family. The PCUO optical module has a heatsink mounted on it, and as in the case of PCUO ×8 optical assembly, two modules are connected with an MTP connector in a Y-shape. The PCUO optical module has 4 channels with the transmission speed of 14 Gbps for transmission/receiving, respectively.

3. Implementation of PCIe Adapter Card

This section provides the composition and features of PCIe adapter cards using an on-board optical module. In the following subsections, we will give the detailed explanation of our proposed scheme.

3.1 Proposed Scheme Overview

Fig. 1 shows the conceptual structure of the adapter card prototype proposed by this study. The two different types of adapter cards use Broadcom's PEX8733 and PEX8749, respectively, as their switching chip. Both supporting PCIe Gen3, the two chips offer NTB and accommodate 32 lanes and 48 lanes, respectively. The switching chips send and receive data via Samtec's PCUO module to transfer them with a PCIe bus method.

Samtec's PCUO module plays a role of converting data delivered from PCIe switching chip into optical signals, or converting optical signals delivered through an optical cable into electrical data to deliver PCIe switching chip. At this time, the optical signals handled with an on-board optical module include some control signals like RESET in addition to data signals.

Two important system signals processed internally in a PCIe adapter card are RESET and CLOCK. In order to manage these two signals, we designed the special hardware modules and they will be explained in the following subsection.

The operating method of the adapter card varies under card setting through configuration module. There are two methods of setting an adapter card. Hardware-type method changes the operation of the adapter card by changing the settings of DIP switches in the adapter card. In addition, a PCIe switching chip management program recorded set values on EEPROM of the adapter card and then the card operated by referring to the record.

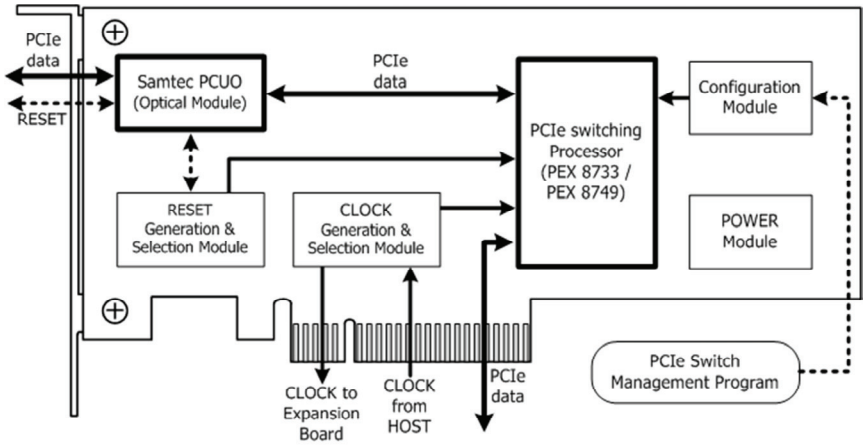


Fig. 1. Conceptual structure of PCIe adapter card prototype.

While changing the size of a PCIe switching chip and a card, two types of adapter card were created like Figs. 2 and 3. First is a PCIe ×8 adapter card using a PEX8749 chip with size of MD2 low-profile (167.64 mm × 64.41 mm). Second card is larger (167.64 mm × 79.50 mm) with ×16 speed and PEX8733 chip used.

Fig. 4 shows a PCUO module mounted on a prototype PCIe adapter card. As it has configuration where two heatsink integral optical modules with speed of PCIe Gen3 ×4 are connected to a single MPO connector, one PCUO provides Gen3 ×8. In other words, one PCUO module was used for prototyped ×8 adapter card, while two PCUO modules were used for ×16 adapter card.

Fig. 2. Prototype of PCIe ×8 adapter card.

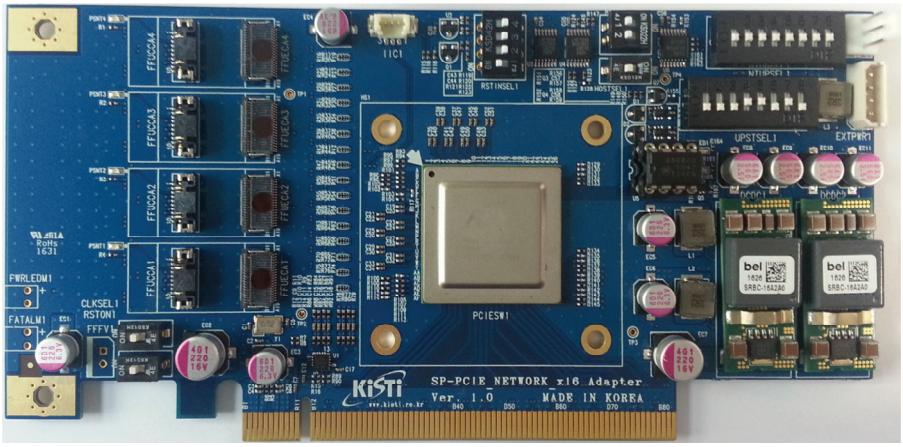


Fig. 3. Prototype of PCIe ×16 adapter card.

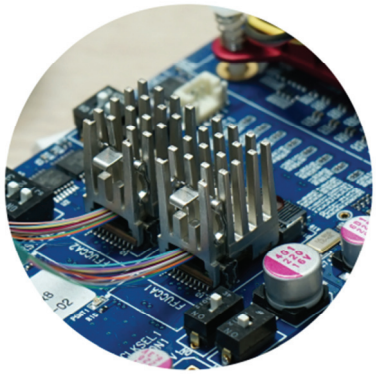


Fig. 4. PCUO module installed at ×8 adapter card.

3.2 Signal Selection Module

Depending on installation location and usage of the adapter card, each signal shall be internally created, or the system signal to be delivered from outside of the adapter card shall be selected. Like Fig. 5, we implemented the hardware module for signals selection in order to manage two special signals, RESET and CLOCK.

We analyzed the various usage and configuration options of adapter card and concluded that the proposed scheme should be able to manage three different types of sources and destinations of the signal. Signals can be delivered from a Host-side mainboard via slot or from a remote Host via optical module. In addition to that, it should be generated by the adapter cards themselves in some cases. Among these signals, the proper one is selected and passed to the final output via the signal buffer. The signal is delivered to a PCIe switching processor of the adapter card, optical module for a remote Host or a slot for Host.

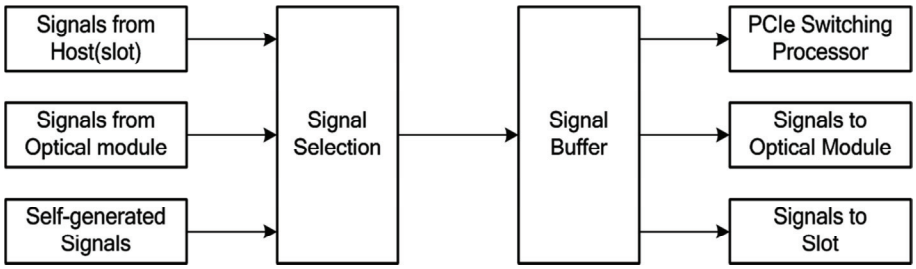


Fig. 5. Signal selection module of our proposed scheme.

4. Performance Evaluation

This section describes the test environment for verifying performance of PCIe adapter cards proposed in Section 3 and provides analysis of the performance verification results. Section 4.1 introduces a conceptual diagram on PCIe expansion and hardware and software to build a test environment. Section 4.2 offers test results and analysis thereof.

4.1 Experiment Environment

Fig. 6 is a conceptual diagram on the test environment, which shows the expansion of PCIe GPU card using PCIe interconnect adapter card. In Fig. 6, "A" is ×16 PCIe slot, at which PCIe adapter card is installed to connect the Host server and expansion box. To evaluate performance, the developed PCIe adapter card and Broadcom's PEX 8732, the existing one, were respectively installed and tested.

An expansion board with two slots (×16) was internally manufactured to expand GPU to the external and optical cable or copper wire-based mini-SAS was used as an expansion link between Host and GPU.

For adapter cards developed for expanded PCIe connection, optical cables were used, while Broadcom PEX 8732 adapter cards used mini-SAS cables. Connection through optical cable required one optical cable per ×8, but connection through mini-SAS cable required two cables per ×8. Table 1

lists specifications of hardware and software used to build the test environment. Host-side CPU is Intel Xeon E5-2600 v4 which has 20 lanes so the system fully supports x16 PCIe, and for GPU, NVIDIA Geforce GTX 1060 and GTX 1050 were used in ×16 and ×8, respectively. Experiment details are as follow:

- ×16 8733 prototype card vs. ×16 system local bus
- ×16 8733 prototype card vs. ×16 PLX PEX8732 adapter card
- ×8 8749 prototype card vs. ×8 PLX PEX8732 adapter card

The bandwidth of each experiment item is measured about CPU to GPU, GPU to CPU using CUDA Bandwidth test utility.

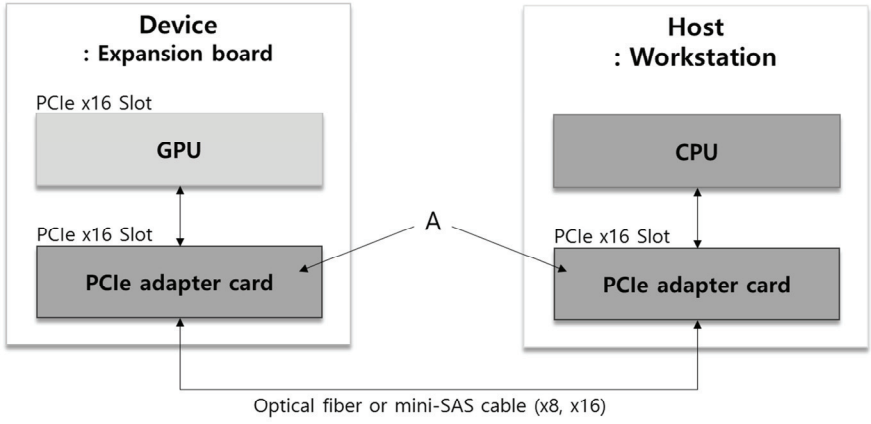


Fig. 6. Conceptual diagram of experiment environment.

Table 1. Experimental equipment specification

Type	Specification
CPU (Host)	Intel Xeon E5-2600 v4
Memory	8 GB DDR4 2133 MHz
GPU	NVIDIA GeForce GTX 1060 (6 GB) / 1050 (2 GB)
Expansion hardware & cable	[Proposed adapter card + optical cable 2ea] vs. [Broadcom adapter card + mini-SAS cable 4ea + SATA cable 1ea]
Expansion board	2 Slots (×16) Self-made type
OS	CentOs 7.3
Management software	PLX SDK & PDE v7.25
Bandwidth test SW	NVIDIA CUDA 8 Toolkit bandwidth test utility

4.2 Experiment Result

Figs. 7 and 8 show the results of measuring bandwidths while increasing the message size up to 1 MB from CPU (Host) to GPU and GPU to CPU (Host). Fig. 7 illustrates the results of measuring bandwidths when GPU was installed on PCIe GEN3 ×16 slot (local bus) in the server (workstation) and when GPU was installed on an external board using the adapter card proposed in this study (optical

link). It was clearly revealed that when the message size is 400 kB or less, use of local bus is better choice for performance, but as the message size increased, performance difference between two cases sharply decreased. However, there was still a certain performance difference in the 0.9–1 MB section with a relatively large message size, and when using the external expansion board, the performance was 98.5% of when local bus was used. We expect that the performance degradation is due to an overhead occurring in the process of using PCIe switching chip, which does not exist in the local method.

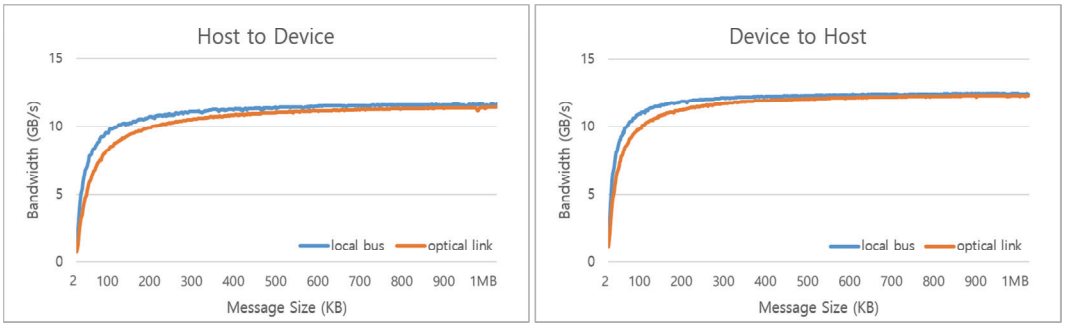


Fig. 7. Proposed adapter card with on-board optical module vs. system local bus (×16).

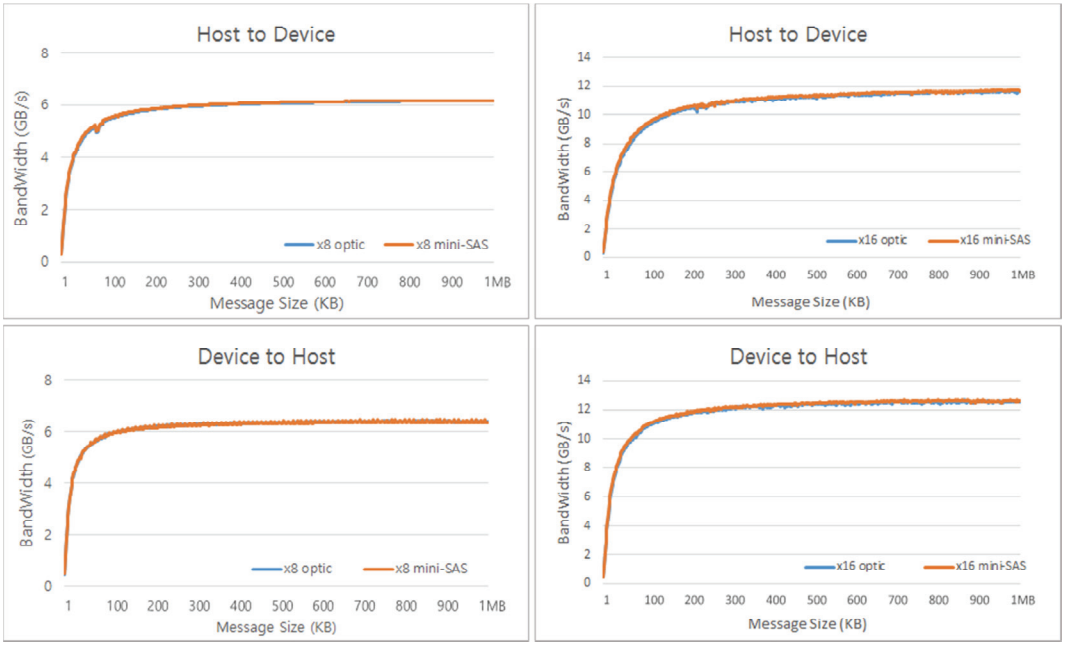


Fig. 8. Proposed adapter card with on-board optical module vs. Broadcom adapter card with mini-SAS cable.

Figs. 8 and 9 show the results of measuring bandwidths of the existing commercial product Broadcom's PEX 8732 mini-SAS-based adapter card and optical module-based adapter card. In this performance evaluation, GPU was commonly installed in an external board and it was connected to a server via adapter card. All test sections displayed almost same results.

Fig. 9 especially displays the saturated bandwidth. In other words, it shows the average bandwidth

when the message size are in the range of 0.9 MB to 1 MB. Overall, performance of Broadcom's PEX 8732 adapter card and that of the proposed adapter card are almost the same and it is hard to find out the serious performance degradation of our proposed scheme. To expand PCIe ×16 to the external, however, 16 copper cables and 1 SATA cable are needed in the case of mini-SAS, and the copper wire has a limitation in terms of distance (as for copper wire-type SAS cable, valid transmission distance is 10 m or less [11]). On the other hand, the proposed adapter card requires 1 or 2 optical cables and enables connection up to 100 m [10].

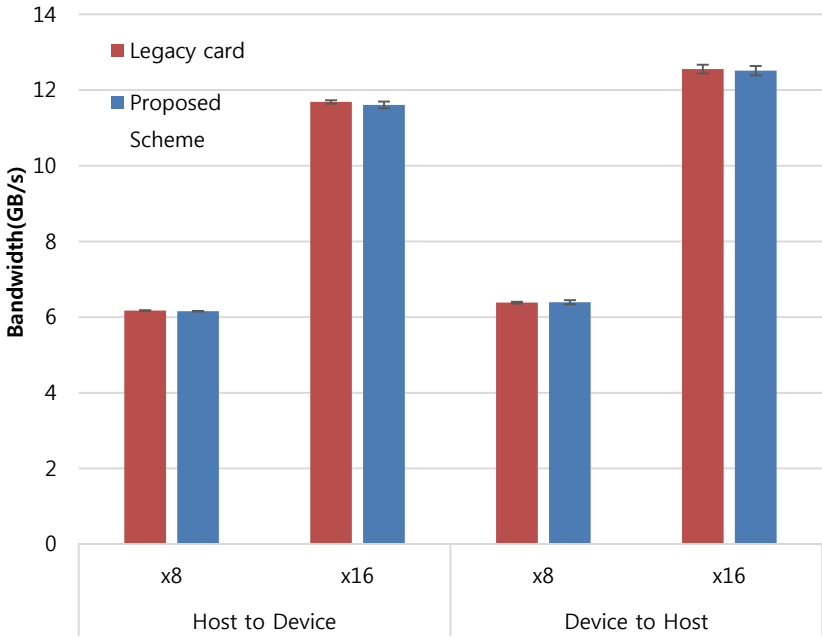


Fig. 9. Performance evaluation results (average bandwidth).

5. Conclusions

The optical module-based on-board PCIe interconnect adapter card proposed by this study is an interface card developed by using Broadcom's PEX 8733 and 8749 switching chips and Samtec's PCUO. With help of our proposed, scheme Gen3 PCIe bus can be expanded to the external through optical cables. According to performance comparison between the existing system mini-SAS-based PCIe interconnect adapter card and the optical module-based PCIe adapter card, almost no performance difference was found. However, copper wire-type mini-SAS cable has a limitation in terms of transmission distance and difficult cable work, while the optical cable which is required by the proposed adapter card expands a transmission distance and involves easy cabling. For these reasons, we believe that the proposed adapter card technology is expected to play a critical role to establish various networks by expanding PCIe bus.

We, however, mainly focused on PCIe devices expansion using on-board optical modules in this study. We also plan to improve our proposed scheme to include an interconnect functionality for host

to host connection. Since the dedicated software is required for the interconnecting, we are developing the software libraries for our proposed hardware. We expect that we can check the host to host performance based on our proposed hardware and software within near future.

Acknowledgement

This research was supported by Korea Institute of Science and Technology Information (KISTI).

References

[1] A. V. Bhatt, "Creating a PCI Express interconnect," PCI-SIG Whitepaper, 2002 [Online]. Available: http://derdoc.info/portfolio/NMSU/CS473/site/pciexpress_whitepaper.pdf.

[2] J. Regula, "Using non-transparent bridging in PCI Express systems," PLX Technology Inc., Sunnyvale, CA, 2004.

[3] Avago Technologies, "ExpressFabric platform PEX9797," 2015 [Online]. Available: https://www.broadcom.com/products/pcie-switches-bridges/expressfabric/pex9797.

[4] R. Hou, T. Jiang, L. Zhang, P. Qi, J. Dong, H. Wang, X. Gu, and S. Zhang, "Cost effective data center servers," in *Proceedings of 2013 IEEE 19th International Symposium on High Performance Computer Architecture*, Shenzhen, China, 2013, pp. 179-187.

[5] M. Choi and J. H. Park, "Feasibility and performance analysis of RDMA transfer through PCI Express," *Journal of Information Processing Systems*, vol. 13, no. 1, pp. 95-103, 2017.

[6] T. Hanawa, Y. Kodama, T. Boku, and M. Sato, "Interconnection network for tightly coupled accelerators architecture," in *Proceedings of 2013 IEEE 21st Annual Symposium on High-Performance Interconnects,* San Jose, CA, 2013, pp. 79-82.

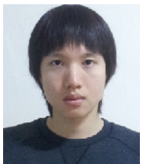
[7] M. Vesper, D. Koch, K. Vipin, and S. A. Fahmy, "JetStream: an open-source high-performance PCI Express 3 streaming library for FPGA-to-Host and FPGA-to-FPGA communication," in *Proceedings of 2016 26th International Conference on Field Programmable Logic and Applications*, Lausanne, Switzerland, 2016, pp. 1-9.

[8] Foxconn, "MniPOD specification," [Online]. Available: http://www.fit-foxconn.com/Images/Products/Spec/AFBR-822VxyZ_20160510175057424.pdf.

[9] Foxconn, "MicroPOD specification," [Online]. Available: http://www.fit-foxconn.com/Images/Products/Spec/AFBR-77D1SZ_20160510175052121.pdf.

[10] Samtec, "PCIe MID-BOARD OPTICS," 2016 [Online]. Available: http://suddendocs.samtec.com/ebrochures/samtec_pcuo_ebrochure.pdf.

[11] B. Hansen, "Extending SAS connectivity in the data center," presented in *the Storage Developer Conference*, Santa Clara, CA, 2013 [Online]. Available: http://www.snia.org/sites/default/files/files2/files2/SDC2013/presentations/BlockStorage/BobHansen_Extending_SAS_Connectivity_Data_Center-v1.pdf.



Kyungmo Koo https://orcid.org/0000-0002-9783-7223

He received M.S. degrees in School of Electrical Engineering and Computer Science from Gwangju Institute of Science and Technology in 2015. Since 2017, he has been with Department of Supercomputer System Development of Korea Institute of Science and Technology Information, where he is currently a research staff.

Junglok Yu

He received his Ph.D. degrees from KAIST, Daejeon, Korea in 2007. He was a senior engineer in Samsung Electronics from 2007 to 2010. He is currently a principal researcher with Supercomputing center, Korea Institute of Science and Technology Information. His research interests include parallel processing, cluster computing, and cloud computing.

Sangwan Kim https://orcid.org/0000-0003-1173-8759

He received M.S. degree in Computer Science and Engineering from POSTECH in 2001. His current research interests include reconfigurable computing and hardware acceleration.

Min Choi https://orcid.org/0000-0002-8031-1022

He received the B.S. degree in Computer Science from Kwangwoon University, Korea, in 2001, and the M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 2003 and 2009, respectively. From 2008 to 2010, he worked for Samsung Electronics as a Senior Engineer. Since 2011 he has been a faculty member of Department of Information and Communication of Chungbuk National University. His current research interests include high performance computing, cloud computing, interconnection network, and embedded computing.

Kwangho Cha https://orcid.org/0000-0003-3299-4575

He received the B.S. degree in computer science from Soongsil University, Korea in 1999 and the M.E. degree in information and computer engineering from Information and Communications University (ICU), Korea in 2002 and the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST) in 2012. Since 2002, he has been with Division of Supercomputing of Korea Institute of Science and Technology Information, where he is currently a senior research staff. His research interests include parallel and cluster computing, parallel file system, and high-performance computer systems.