Low-latency Data Flow Management for Heterogeneous Systems

Yusuke Fujii, Shunsuke Kurumatani, Hiroyuki Takahashi, Toshio Hitaka
fujii.yusuke@lab.ntt.co.jp, kurumatani.shunsuke@lab.ntt.co.jp, takahashi.hiroyuki@lab.ntt.co.jp, hitaka.toshio@lab.ntt.co.jp
NTT Software Innovation Center, NTT Corp.

1. Background

Accelerators[Fig.1] are effective devices for compute-intensive applications, and that are now widely used for several domains of applications.

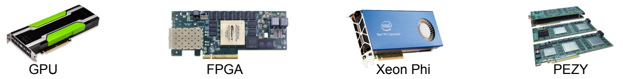

Fig.1: Several types of accelerators.

We expect to be able to accelerate compute/data-intensive applications by a suitable combination of devices which have different characteristics. For example, GPU is suitable for data-parallel computing like SIMD, and FPGA is suitable for task-parallel computing. Therefore, we expect to get more performance with these characteristics.

A typical heterogeneous system[Fig. 2] has some nodes. Each node is equipped with various types of accelerators. A node is connected to another node by network for receiving/sending data to/from external resources and communication between nodes. Some accelerator devices are interconnected by a PCIe network.

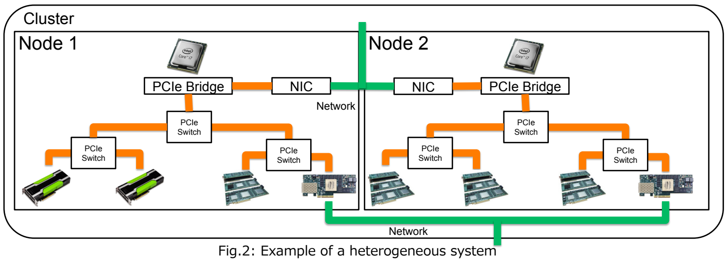

Fig.2: Example of a heterogeneous system

In the case of using various types of accelerators, performance of computing may be decreased due to I/O latency. Therefore, latency reduction is an important challenge to improve performance of computing in heterogeneous systems.

Our target is to integrate accelerators effectively into data centers for increasing data or processings; thus we want to use a model which is receiving data from network because of using distributed-computing.

2. Related work and Motivation

Some studies were focused on low-latency device communication:

- Direct communication GPU and FPGA:
  - PEACH2: Tsukuba University
  - [Springer '13]: Microsoft
- Packet processing in an FPGA and direct communicating to a GPU:
  - APEnet+: INFN
  - EXTOLL NIC: Deep project

※PEACH2 targets GPU to GPU direct communication, however it uses FPGA devices.

These studies are supportive of a homogeneous system that has cluster structured with the same nodes. On the other hand, the heterogeneous systems have clusters with different nodes; therefore, these studies did not take into account various types of device switching functions.

We will develop a new intelligent low-latency communication method for the heterogeneous system.

Our method's functions:

- **Direct devices communicating**
  - Device-to-device communication without host memory copy
  - Direct write payload data to another device from a network
- **Device switching**
  - Decision of the destination device based on a port number in packet headers
- **Resource management**
  - Managing data flow for load-balancing
  - Decision of the best device based on load of device

The methods use an FPGA equipped with four 10Gb Ethernet modules. The above functions will be implemented on the FPGA.

This research is currently in the beginning stages, and this poster describes the concept of the methods to investigate an effective processing infrastructure with heterogeneous systems.

3. Data flows of Applications via Network

We show examples[Fig.3] of three general data flows among an external resource, GPU, and FPGA.

Pattern 1:
Copy count = 6

Pattern 2
Copy count = 5

Pattern3:
Copy count = 4

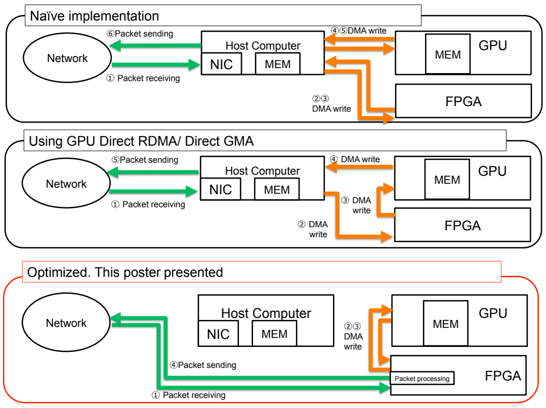

Fig.3: Example of data flows of applications via network.

4. Our Communication Method Design

We have shown first system design[Fig.4]. FPGA (PCIe Endpoint) receives UDP packet from network. Payload data are extracted from UDP packet on the FPGA's UDP/IP Stack. Payload data are processed application that are written to a device memory by PCIe DMAC.

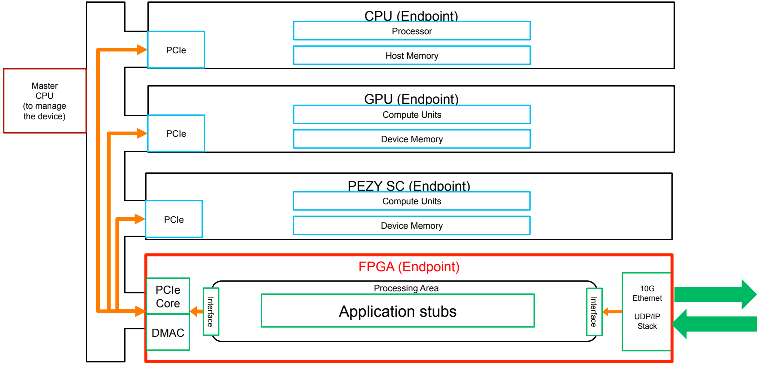

Fig.4: Block diagram of our communication method

5. Control Flows of GPU Processing Invoking

After the FPGA writes data to another device memory, some applications must invoke a kernel. We show the control flow on how to pull a trigger of kernel launch (example of GPU)[Fig.5].

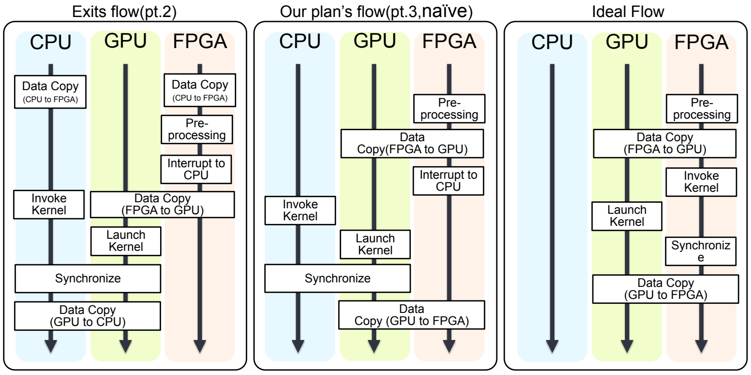

Fig.5: Control flows of GPU processing invoking

6. Future Work

We will investigate an effective processing infrastructure with heterogeneous systems according to develop the intelligent low-latency communication method. This poster presented a data flow manager based on low-latency communication techniques. We need to ensure the effectiveness of this concept. The first step will be to verify the concept by implementing PoC. We will then measure the effectiveness by using real-applications such as image classifications based on deep learning algorithm.

However, we should address the following issues:

- Make an abstraction layer for hiding hardware layer
- Manage interconnect networks automatically
- Make an easier to use the heterogeneous systems

Additionally, we should examine another device and network architectures:

- A network connection between nodes
  - Infiniband
- A network connection between devices
  - NVLink
  - CAPI

2nd Annual Meeting on Advanced Computing System and Infrastructure (ACSI2016) Jan, 18-20, 2016