Dedicated Inter-FPGA Networks for Scalable Reconfigurable Computing

Kentaro Sano
RIKEN Center for Computational Science (R-CCS)
Introduce Myself: Kentaro Sano

RIKEN Center for Computational Science
- Develop and operate Supercomputer Fugaku
- Facilitate leading edge infrastructures for research based on supercomputers
- Conduct cutting-edge research on HPC

Leader, Processor Research Team
- Exploration of future HPC architectures
- Advanced use of present HPC systems

Joint Laboratory at Tohoku University
- Visiting Professor
  "Advanced Computing Systems Lab"

Hiring researchers: R-CCS2105 or R-CCS2022

Supercomputer Fugaku
Kobe city
Riken Center for Computational Science

NANDA Workshop
Sep 5, 2022
Goal and Roadmap of Processor Research Team

Establish HPC architectures suitable for Post-Moore Era

Advancement of Fugaku
- Functional extension with FPGAs and eco-system
- System software and apps of task-flow computing

Exploration of New HPC Architectures
- Novel accelerators based on data-flow model (CGRA)
- System architectures

Near-sensor / Near-storage Processing
- FPGA-based processing for X-ray imaging detector

Exploration for Novel Computing Principle
- Specialized hardware design for quantum error correction

This talk
Outline

- Introduction
- ESSPER: FPGA Cluster Prototype
- Inter-FPGA network
- Implementation and evaluation
- Conclusions
Introduction

- **Accelerators for higher power-efficiency**
  - System power is the most critical issue. (Fugaku: 20+MW for operation, 30MW as max.)
  - Standard CPUs are not sufficient. Accelerators (Accs) for higher performance per power.

- **Reconfigurable computing with FPGAs**
  - GPUs are popular as gen-purpose Accs.
  - More specialized, higher efficiency. But we also need flexibility. **FPGAs!**

- **Prototype FPGA Cluster “ESSPER”**
  - Proof-of-concept system to evaluate FPGA-based extension of Fugaku.
  - Challenges:
    - How to scale with multiple FPGAs

---

What’s next? CPU, GPU, or Acc?
Motivation and Objective

Motivation

✓ What kind of inter-FPGA network are appropriate?
✓ Assumption: Many FPGAs in a system.
   Each of users uses them partially.

Objective

Find inter-FPGA network appropriate for a large-scale system with multiple users

✓ Investigate requirements
✓ Compare Direct and Indirect networks
✓ Propose Virtual circuit-switching network (VCSN)I
✓ Design, implement, and evaluate
ESSPER: FPGA Cluster Prototype
Elastic and Scalable System for High-Performance Reconfigurable Computing

Experimental prototype for research on functional extension with FPGAs

Supercomputer Fugaku

Connected w/ 100m cables
Modern Supercomputers are based on Many-core CPUs (& GPUs).

System Configuration of Fugaku

Photos & figs by Fujitsu
Elastic and Scalable System for High-Performance Reconfigurable Computing

Experimental prototype for research on functional extension with FPGAs

Connected w/ 100m cables
Elastic and Scalable System for High-Performance Reconfigurable Computing
Goal

✓ Technical investigation for functional extension of Fugaku.

**Architecture of ESSPER**

- **System network**
  - Login node
  - Fugaku node
  - Fugaku node
  - Fugaku node
  - Fugaku node

- **Computing network (Tofu-D)**
  - Fugaku node

- **Bridging network**
  - Manage node
  - FPGA node
  - FPGA node
  - FPGA node
  - FPGA node

- **Extension system**
  - Inter-FPGA network

- **100m cables**
Hardware Organization of ESSPER

Various servers
- CAD servers
- Storage server
- ARM servers

CPU - FPGA network
- 100G Infiniband
- Software-bridged driver (R-OPAE)

FPGA cluster
- FPGA host servers (x86)
- FPGA boards
- Inter-FPGA network

FPGA SoC
- AFU Shell design
- FPGA object class as HAL
- Programming by HLS, DSL
Inter-FPGA Network
Assumption and Requirements

- A lot of FPGA resources in a system
  - 100~, 1000~, or more?

- FPGAs are (globally or partially) connected by their dedicated networks
  - Hardware programmed on multiple FPGAs operates by communicating and synchronizing with each other.

- Each of multiple users acquires a part of FPGAs and execute tasks on them.
  - User A : 16 FPGAs with 2D torus network.
  - User B : 64 FPGAs with a tree network.
## Two Types of Networks

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Direct network</th>
<th>Indirect network</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>p2p-connection without switches</strong>, typical: torus network</td>
<td><strong>connection with switches</strong>, typical: Ethernet</td>
<td></td>
</tr>
<tr>
<td><strong>Switching</strong></td>
<td><strong>circuit</strong> or packet (w/ on-chip router)</td>
<td><strong>packet</strong></td>
</tr>
</tbody>
</table>

### Pros
- Direct network: **low latency**, easy to use with simple HW
- Indirect network: **flexibility**, small diameter, easy adoption of cutting-edge

### Cons
- Direct network: large diameter, inflexibility in resource allocation
- Indirect network: higher latency due to packet processing, complex and difficult to use
### Related Work: Networks for FPGAs in HPC/DC

<table>
<thead>
<tr>
<th>Type</th>
<th>Direct network</th>
<th>Indirect network</th>
<th>Indirect circuit-switching nw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics</td>
<td>p2p-connection without switches, typical: torus</td>
<td>connection with switches, typical: Ethernet packet</td>
<td>connection with optical switch (MEMS)</td>
</tr>
<tr>
<td>Switching</td>
<td>circuit or packet (w/ router)</td>
<td>packet packet</td>
<td>circuit or packet (w/ router)</td>
</tr>
<tr>
<td>Pros</td>
<td>low latency</td>
<td>flexibility, small diameter</td>
<td>low latency, flexibility</td>
</tr>
<tr>
<td>Cons</td>
<td>inflexibility, large diameter</td>
<td>higher latency, complex</td>
<td>expensive, signal attenuation</td>
</tr>
<tr>
<td>Representative systems</td>
<td>Cygnus @ U of Tsukuba</td>
<td>Catapult @ Microsoft</td>
<td>Noctua @ Paderborn U</td>
</tr>
<tr>
<td></td>
<td>Novo-G# @ U of Florida</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What Network is Appropriate for Multi-User System?

- **Inflexibility of direct network**
  - Cannot provide requested topology for partial use of FPGAs
  - Full torus cannot be provided.
    Only (n-1)-D torus or mesh available

- **Flexibility of indirect network**
  - Can provide *any topology for any part* of the FPGA nodes
  - Appropriate for operation of a large system with multiple users
  - However, *complicated to use* due to packet generation and destination control

*Example of 2D torus network.*
Partial usage is limited to 1D torus or 2D mesh.
**Proposal: Virtual Circuit Switching Network (VCSN)**

Provide arbitrary topology with virtual links over Ethernet

- Easy to use by simply sending data through a virtual topology.
  No complex control required for user logic.
Proposal: Virtual Circuit Switching Network (VCSN)

Provide arbitrary topology with virtual links over Ethernet
✓ Easy to use by simply sending data through a virtual topology.
   No complex control required for user logic.
Mechanism of Virtual Circuit-Switching

**Mux & Demux**
- Multiple virtual ports for User modules
- TDM of multi streams
- Intel Avalon-ST (stream)

**Frame encoder/decoder**
- Encode stream into Frames, or decode Frames into stream
- Destination MAC address is put on each Frame.

**User modules**
- Send / receive Avalon-ST data streams with multiple ports

**100G Ethernet**
- Ethernet Frame (L1)
- Intel’s Ether MAC IP

**User Hardware Modules**

- DC FIFO 512 to 256
- DC FIFO 512 to 256
- DC FIFO 256 to 512
- DC FIFO 256 to 512

**Virtual ports**

- Mux
- Demux

**Frame Encoder (Destination Table)**

- 100G Ethernet MAC IP
- QSFP28

**100G Ethernet Switch**
Payload Efficiency of VCSN

- Theoretical max efficiency: 99.54% due to Jumbo Frame

**Ethernet FCS : Frame Check Sequence**

NANDA Workshop
Sep 5, 2022
Implementation and Evaluation
FPGA Shells for Direct and Indirect Networks

Direct connection network (DCN)

Indirect network (VCSN)

VCSN Setup with 100Gbps Ethernet Switches

Intel PAC D5005 FPGA cluster with VCSN

- Two 16-port 100G Ethernet switches
- Two ports of FPGA are connected to a different switch (Dual Plane).

VCSN Configuration Examples

6-FPGA networks
• 1-D torus
• 2-D torus (2x3)

Virtual topology of bi-dir 2D Torus

Area and Latency

- **Area**

<table>
<thead>
<tr>
<th>DCN subsystem</th>
<th>ALM</th>
<th>Registers</th>
<th>M20k</th>
<th>DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL3</td>
<td>10474.2</td>
<td>10356</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Width converter</td>
<td>1275.4</td>
<td>3626</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Flow controller</td>
<td>1175.1</td>
<td>1452.4</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>Network subsystem</td>
<td>12924.7</td>
<td>15434.4</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Percentage</td>
<td>0.47%</td>
<td>0.14%</td>
<td>0.84%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

  | Stratix 10 SX280 | 2753000 | 11012000 | 11721 | 5760 |

- **Latency (minimum)**

<table>
<thead>
<tr>
<th>Network</th>
<th>path</th>
<th>latency [ns]</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCSN</td>
<td>virtual ports ↔ virtual ports</td>
<td>851.093</td>
</tr>
<tr>
<td>DCN</td>
<td>Cable link</td>
<td>243.137</td>
</tr>
<tr>
<td></td>
<td>Crossbar ↔ crossbar</td>
<td>490.942</td>
</tr>
</tbody>
</table>

**Area and latency:** DCN << VCSN

Throughput (point-to-point)

Throughput of VCSN rises slowly due to higher latency.

- P2P latency of VCSN: 851 ns
- P2P latency of DCN: 490 ns

VCSN has higher Max throughput.

- 100Gbps = 12.5 GB/s
- Jumbo frame of Ethernet is more efficient: 96% of the peak

Latency-tolerant stream computing should work well.
Comparison of Stream-Computing Performance

- **2D Fluid dynamics simulation**
  - Lattice Boltzmann method
  - 48 PEs / FPGA, 155 MHz
  - Streaming 2 GB test data

**FLOPS by DCN ≒ FLOPS by VCSN**

When bandwidth determines computational performance, **VCSN is equivalent to DCN for a large data.**
## Summary

<table>
<thead>
<tr>
<th><strong>Objective</strong></th>
<th>Find inter-FPGA network appropriate for a large-scale system with multiple users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our proposal</strong></td>
<td><strong>Indirect network with VCSN</strong>&lt;br&gt;Virtualized circuit-switching network over Ethernet frame for higher flexibility</td>
</tr>
<tr>
<td><strong>Comparison</strong></td>
<td><strong>x2 latency with slightly higher throughput</strong>&lt;br&gt;compared to DCN (direct connection network)</td>
</tr>
<tr>
<td><strong>Future work</strong></td>
<td>✓ System software to configure VCSN (almost implemented)&lt;br&gt;✓ Evaluation with application cases</td>
</tr>
</tbody>
</table>

**Hiring researchers:**
- R-CCS2105 or R-CCS2022