

# Dedicated Inter-FPGA Networks for Scalable Reconfigurable Computing

Kentaro Sano

RIKEN Center for Computational Science (R-CCS)

# Introduce Myself : Kentaro Sano

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

## 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



Supercomputer Fugaku

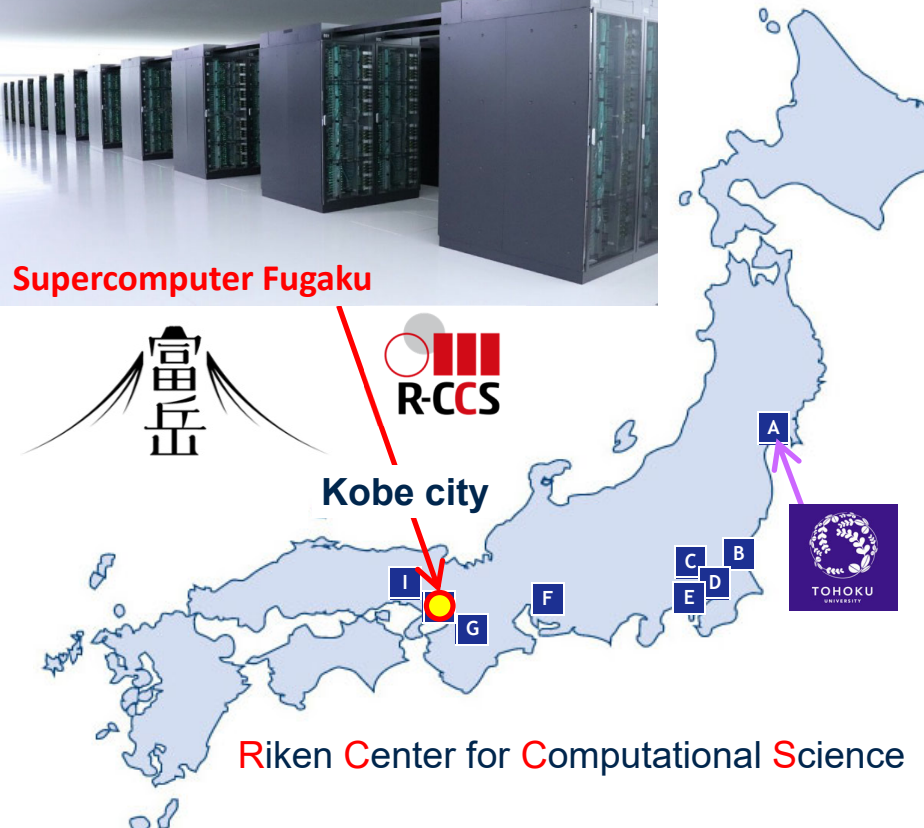
## Leader, Processor Research Team

- ✓ Exploration of future HPC architectures
- ✓ Advanced use of present HPC systems



## Joint Laboratory at Tohoku University

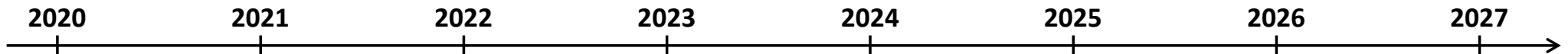
- ✓ Visiting Professor  
"Advanced Computing Systems Lab"



Riken Center for Computational Science

# Goal and Roadmap of Processor Research Team

## Establish HPC architectures suitable for Post-Moore Era



### Advancement of Fugaku

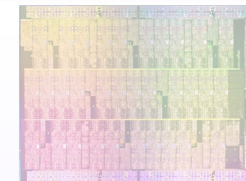
**This talk**

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



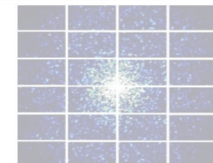
### Exploration of New HPC Architectures

- ✓ Novel accelerators based on data-flow model (CGRA)
- ✓ System architectures



### Near-sensor / Near-storage Processing

- ✓ FPGA-based processing for X-ray imaging detector

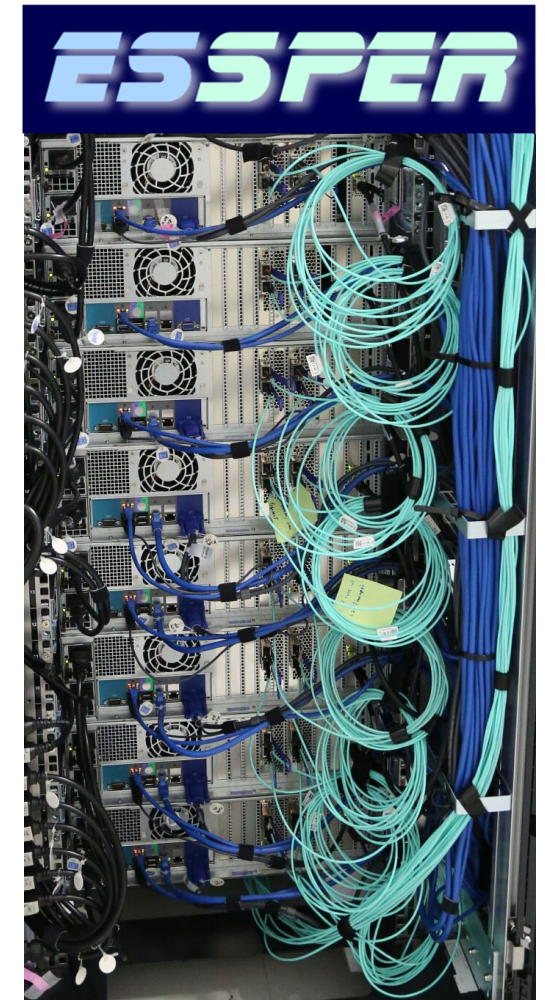


### Exploration for Novel Computing Principle

- ✓ Specialized hardware design for quantum error correction

# Outline

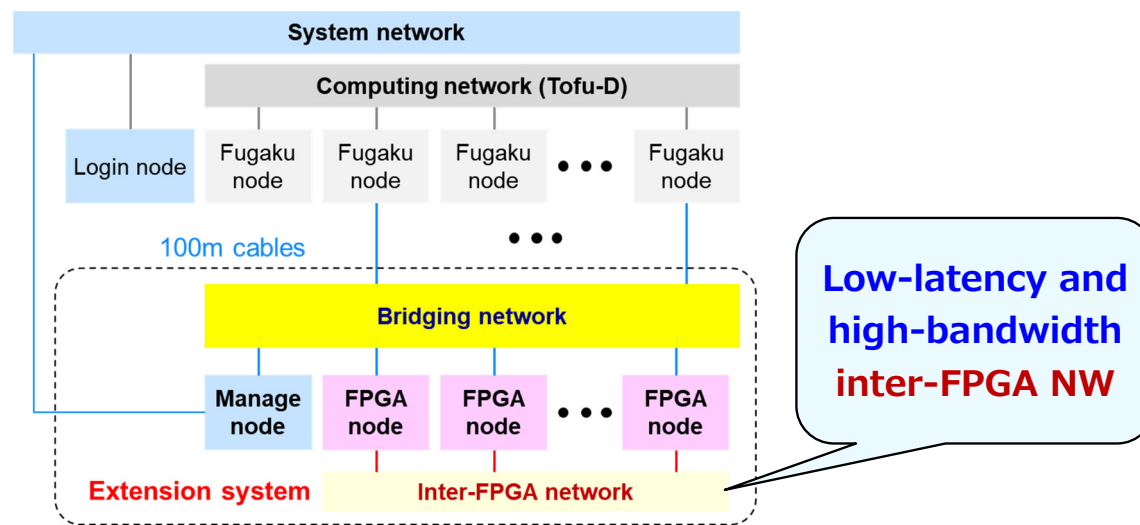
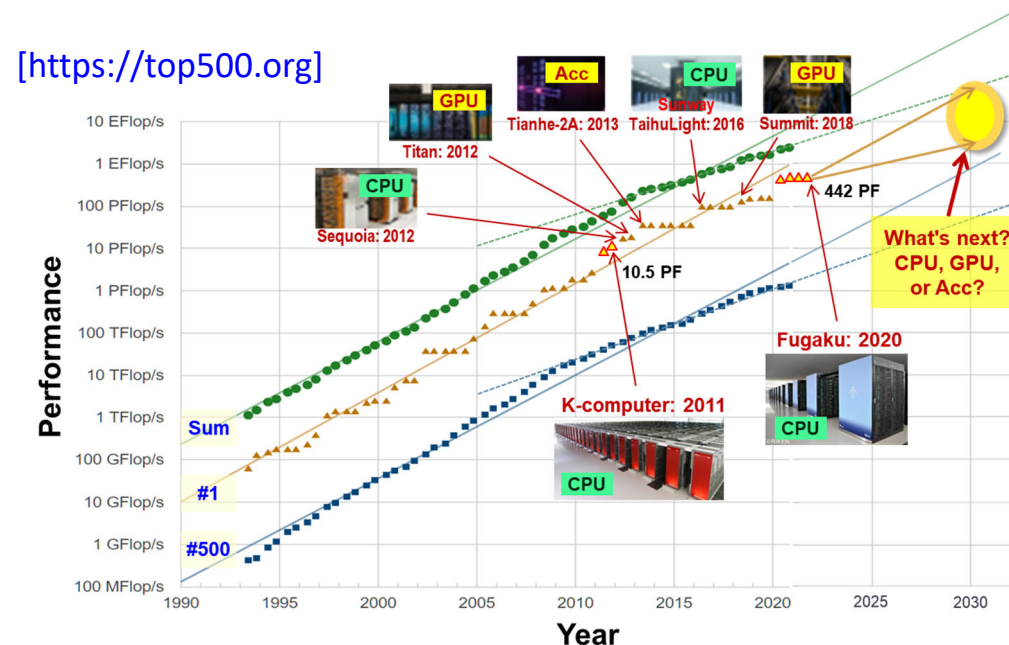
- Introduction
- **ESSPER** : FPGA Cluster Prototype
- Inter-FPGA network
- Implementation and evaluation
- Conclusions



ESSPER: FPGA Cluster System

# 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**



# Motivation and Objective

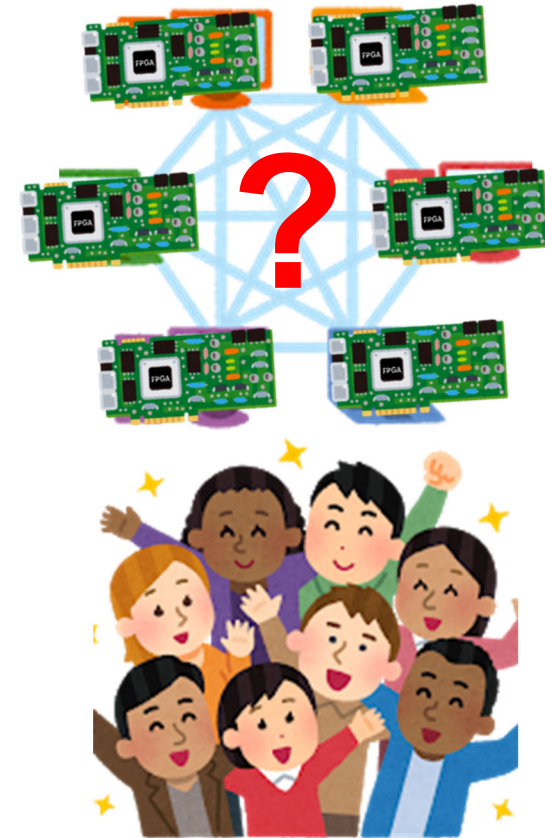
## • Motivation

- ✓ 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**

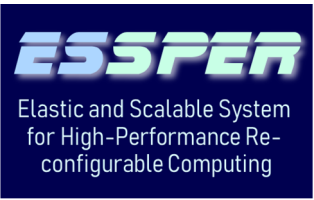
- ✓ Investigate requirements
- ✓ Compare **Direct and Indirect** networks
- ✓ Propose Virtual circuit-switching network (**VCSN**)I
- ✓ Design, implement, and evaluate



The logo for ESSPER, featuring the word "ESSPER" in a stylized, italicized font with a blue-to-green gradient and a white outline, set against a dark blue background.

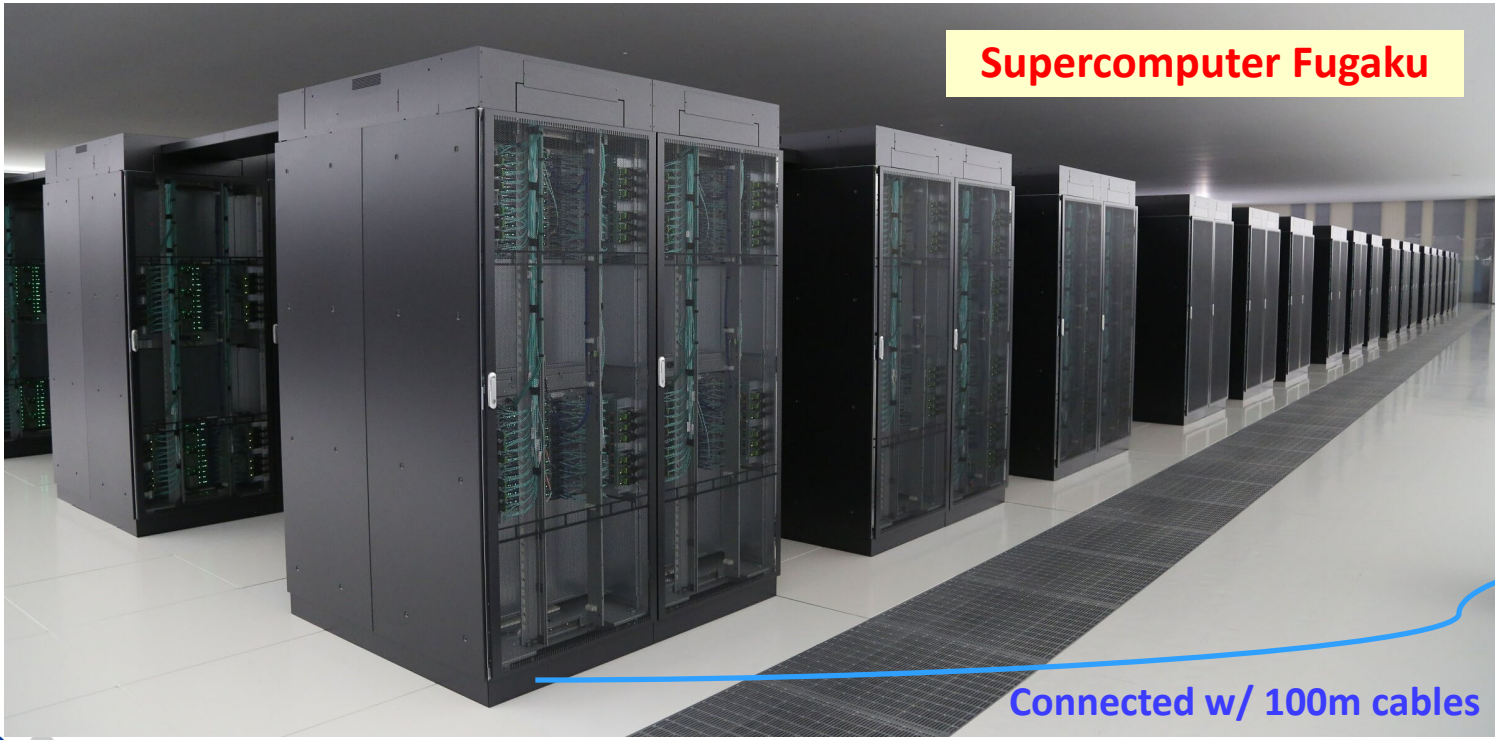
Elastic and Scalable System  
for High-Performance Re-  
configurable Computing

# ESSPER : FPGA Cluster Prototype



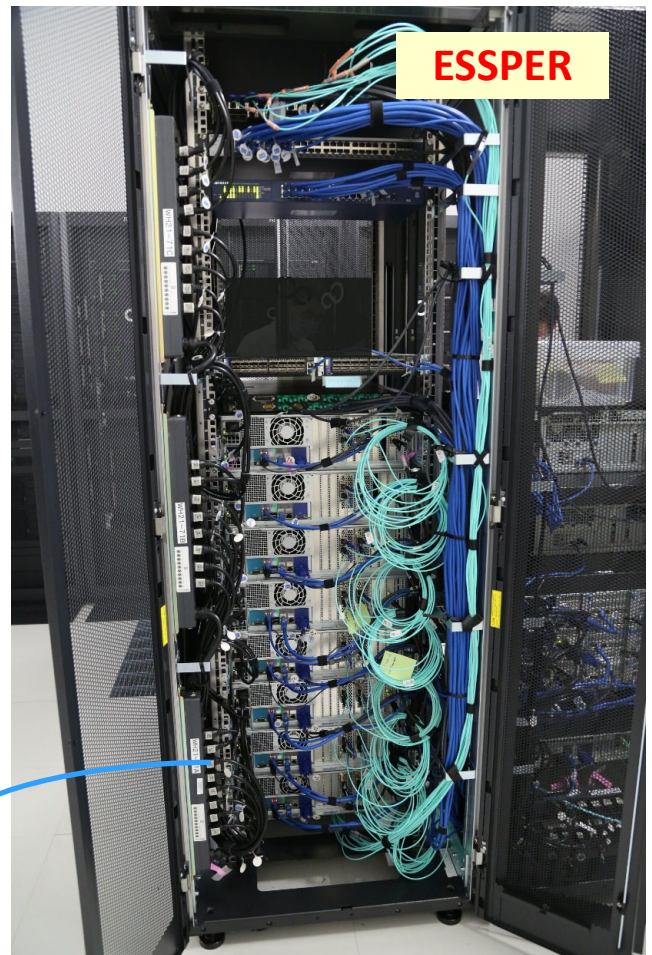
# Elastic and Scalable System for High-Performance Reconfigurable Computing

Experimental prototype for research on functional extension with FPGAs



Supercomputer Fugaku

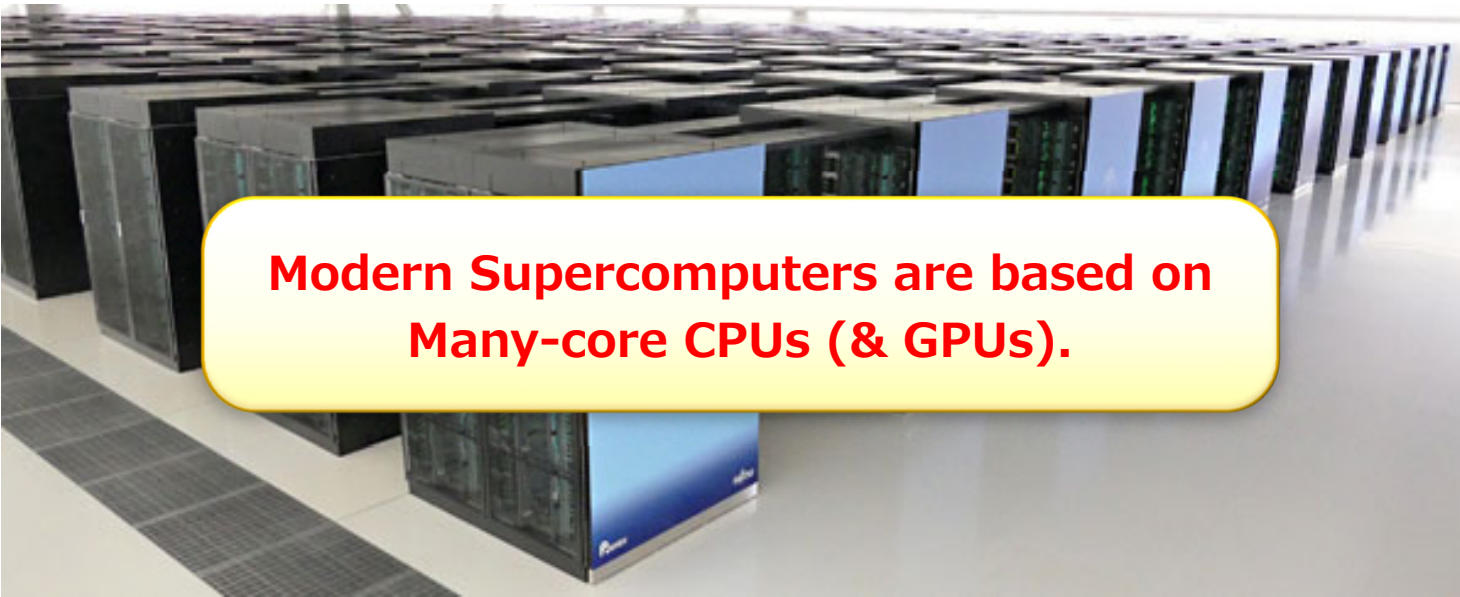
Connected w/ 100m cables



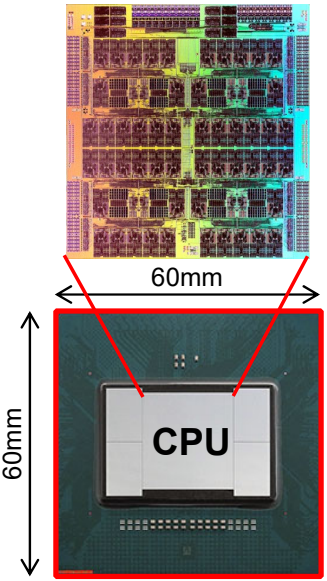
ESSPER



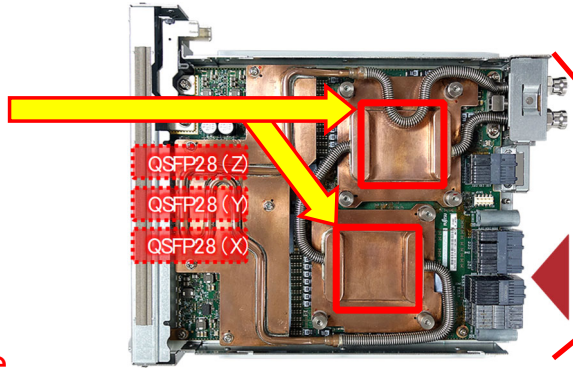
# System Configuration of Fugaku



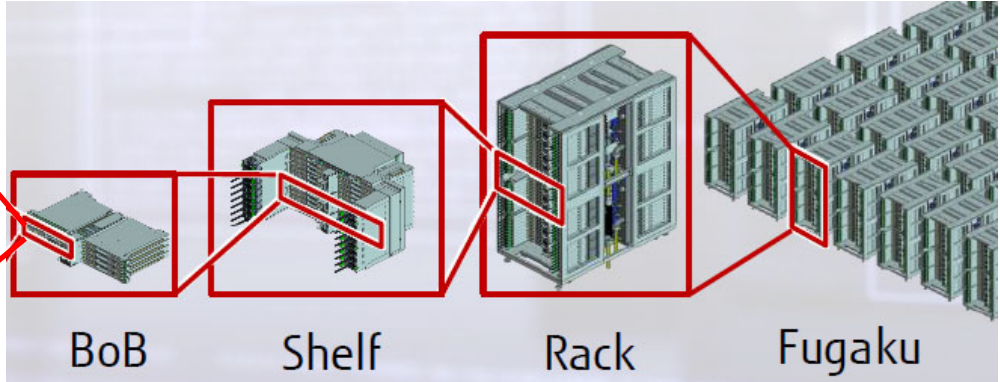
**Modern Supercomputers are based on Many-core CPUs (& GPUs).**



48+ cores / 1 node  
2.7+ TF

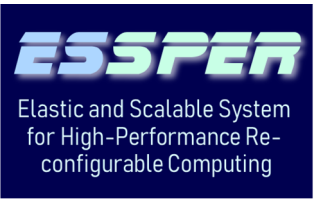


**CPU-Memory Unit (CMU)** 2 nodes  
5.4+ TF



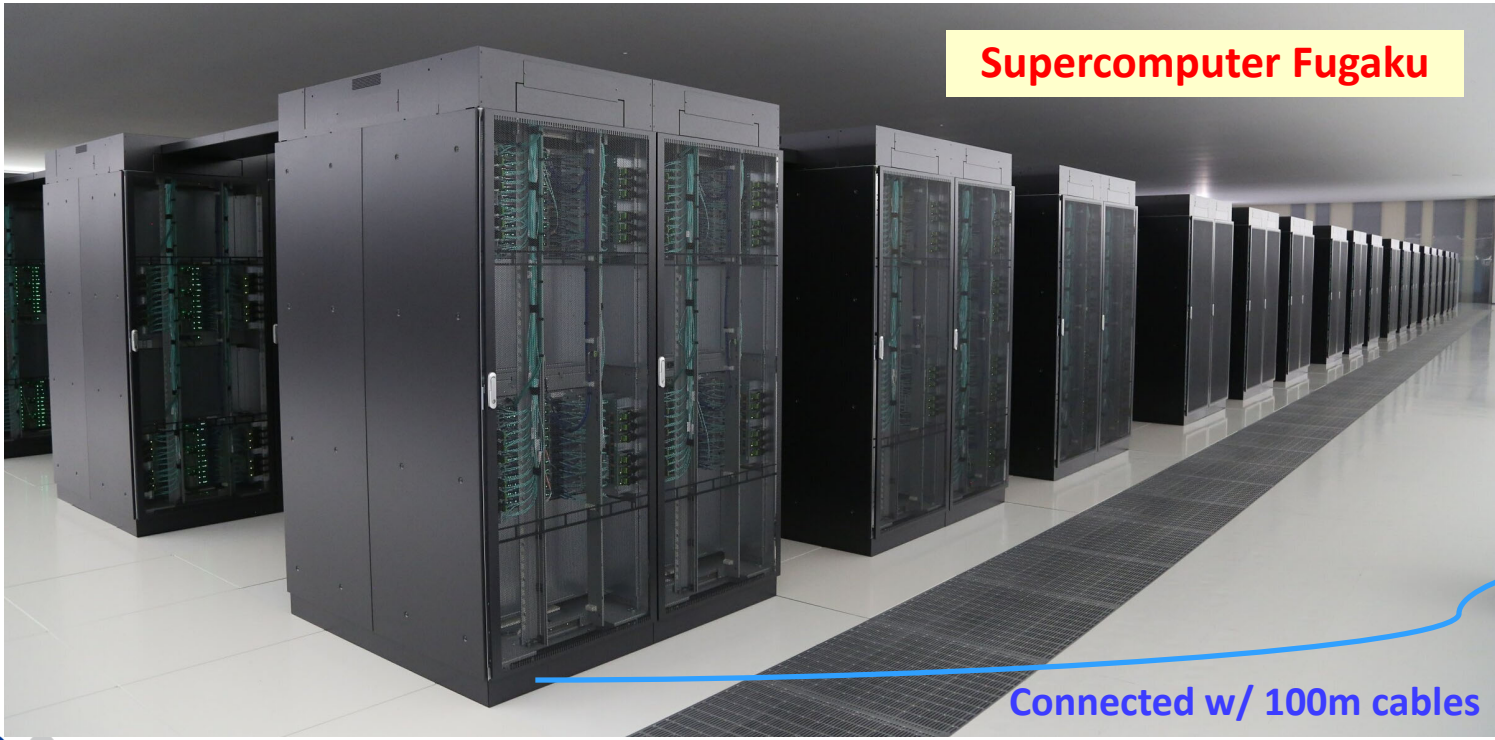
16 nodes	48 nodes	384 nodes	158,976 nodes
43+ TF	129+ TF	1+ PF	537 PF @ FP64 (414 racks)

Photos & figs by Fujitsu



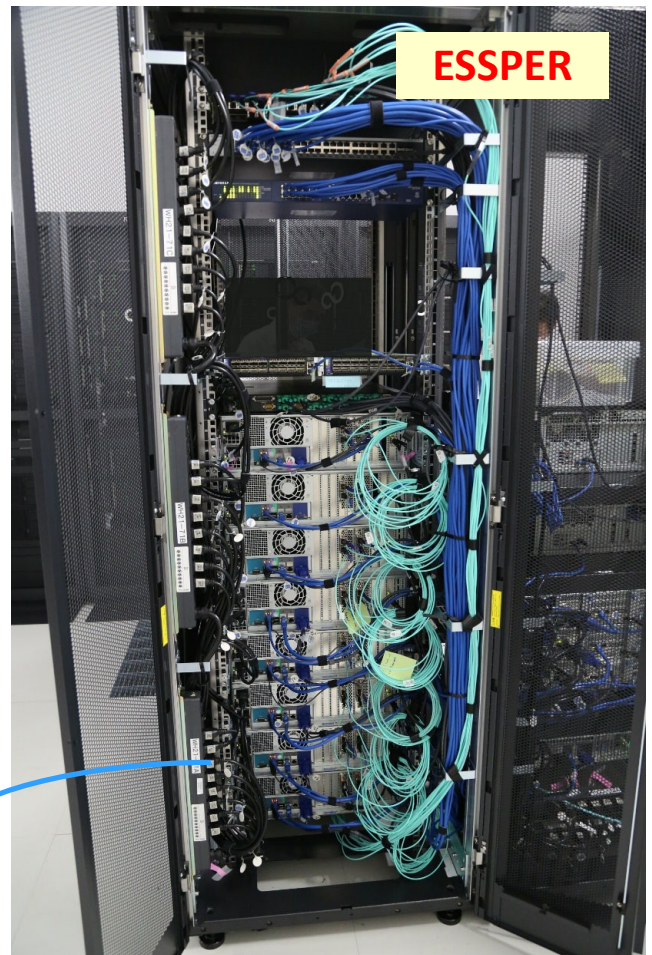
# Elastic and Scalable System for High-Performance Reconfigurable Computing

Experimental prototype for research on functional extension with FPGAs



Supercomputer Fugaku

Connected w/ 100m cables



ESSPER

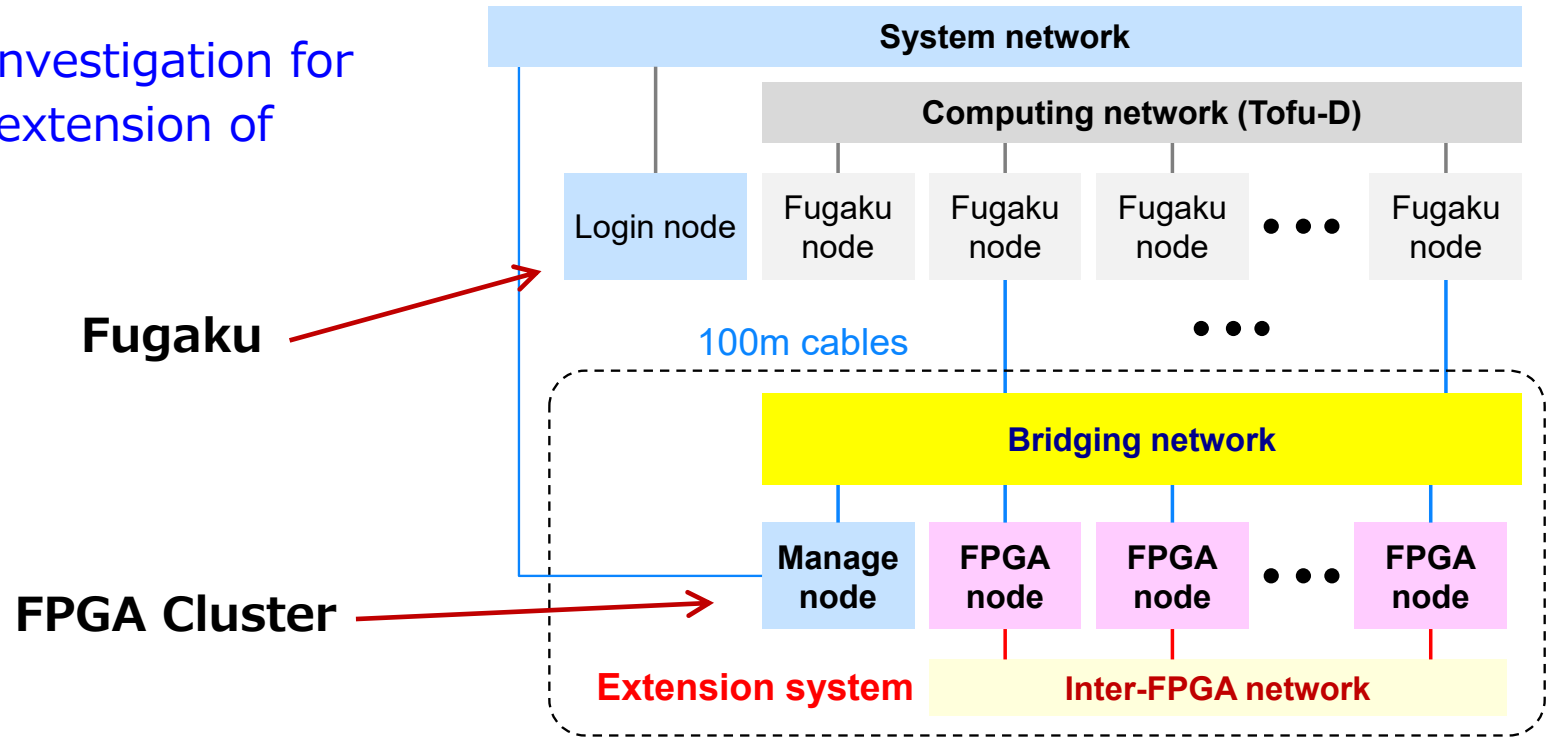
# Elastic and Scalable System for High-Performance Reconfigurable Computing



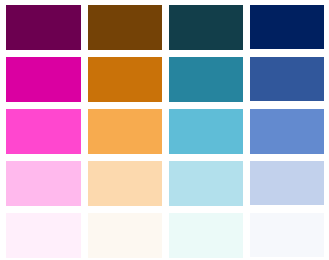
# Architecture of ESSPER

## Goal

- ✓ Technical investigation for functional extension of Fugaku.



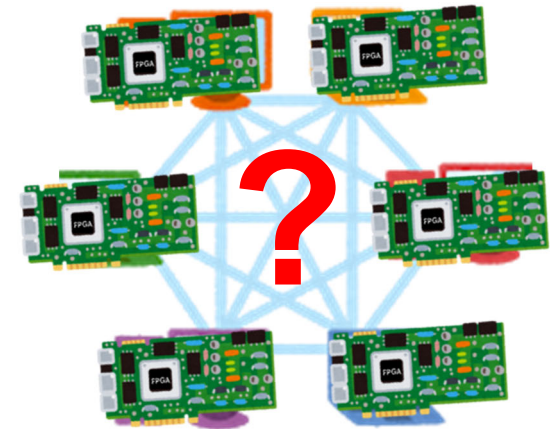




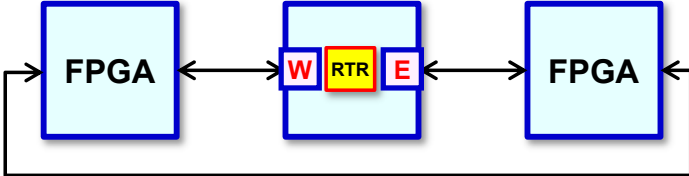
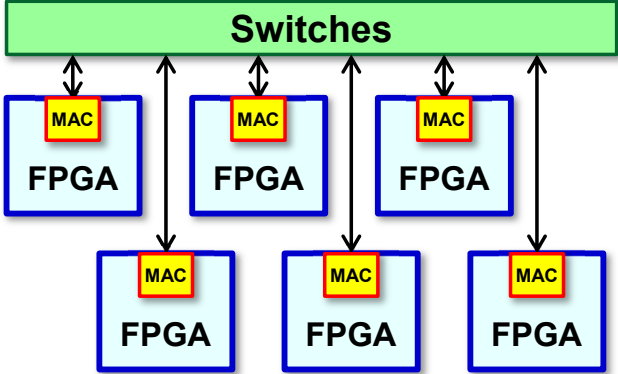
# Inter-FPGA Network

# 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

	Direct network	Indirect network
		
<b>Characteristics</b>	p2p-connection without switches, typical: torus network	connection with switches, typical: Ethernet
<b>Switching</b>	circuit or packet (w/ on-chip router)	packet
<b>Pros</b>	low latency, easy to use with simple HW	flexibility, small diameter, easy adoption of cutting-edge
<b>Cons</b>	large diameter, inflexibility in resource allocation	higher latency due to packet processing, complex and difficult to use



# Related Work : Networks for FPGAs in HPC/DC

Type	Direct network	Indirect network	Indirect circuit-switching nw
Characteristics	p2p-connection without switches, typical: torus	connection with switches, typical: Ethernet	connection with optical switch (MEMS)
Switching	circuit or packet (w/ router)	packet	circuit or packet (w/ router)
Pros	low latency	flexibility, small diameter	low latency, flexibility
Cons	inflexibility, large diameter	higher latency, complex	expensive, signal attenuation

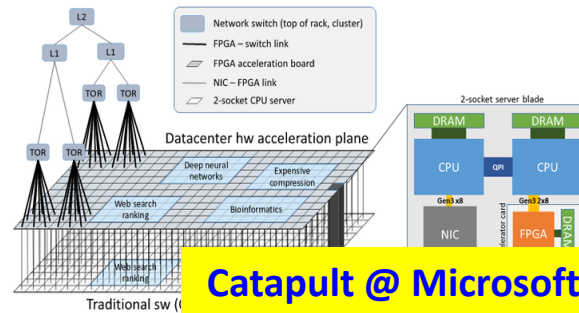
## Representative systems



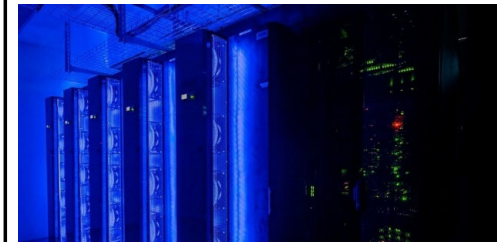
Cygnus @ U of Tsukuba



Novo-G# @ U of Florida



Catapult @ Microsoft



Noctua @ Paderborn U

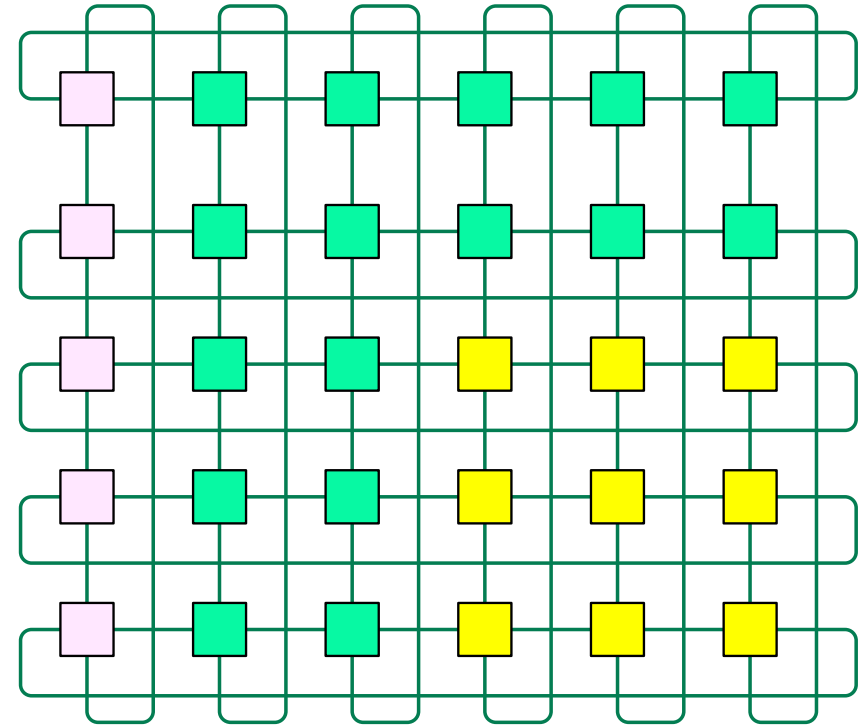
# What Network is Appropriate for Multi-User System?

- **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



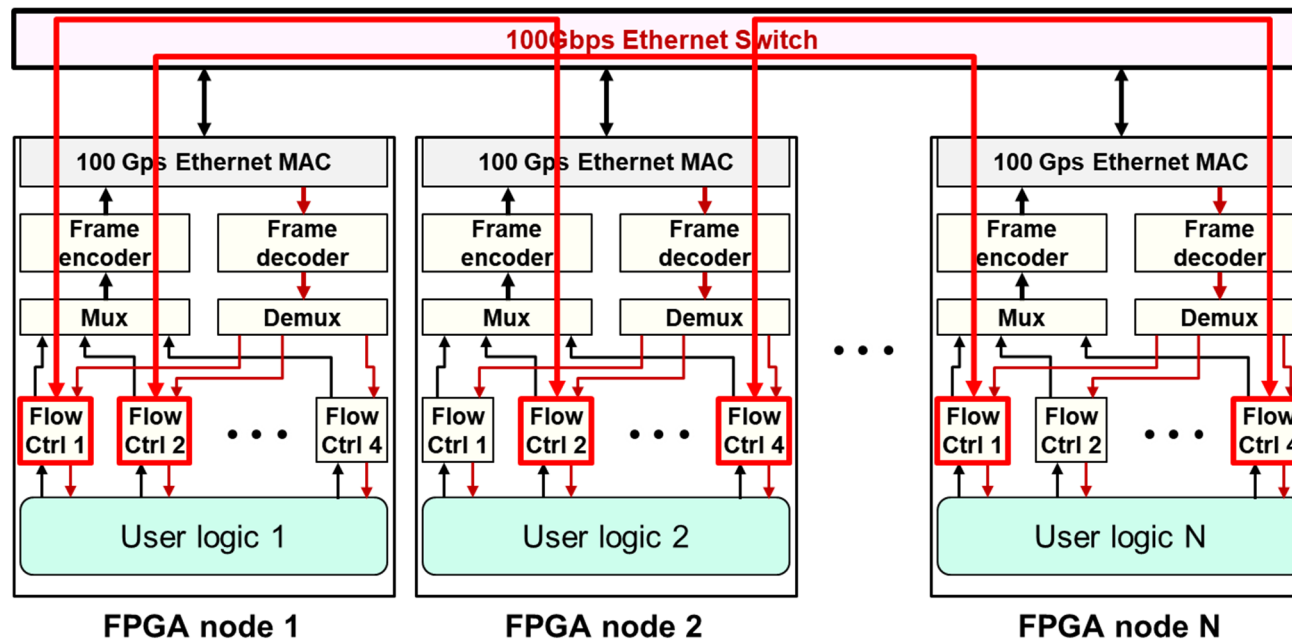
**Example of 2D torus network.**

Partial usage is limited to 1D torus or 2D mesh.

# Proposal: Virtual Circuit Switching Network (VCSN)

## Provide arbitrary topology with virtual links over Ethernet

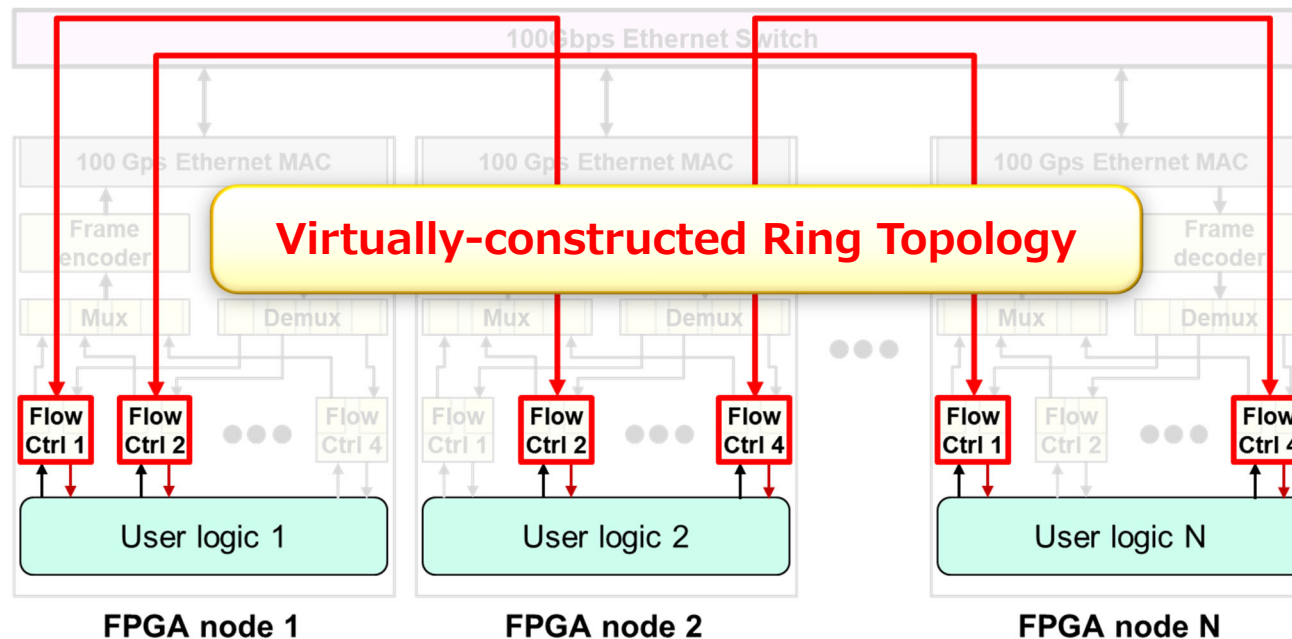
- ✓ Easy to use by simply sending data through a virtual topology.  
No complex control required for user logic.



# 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**

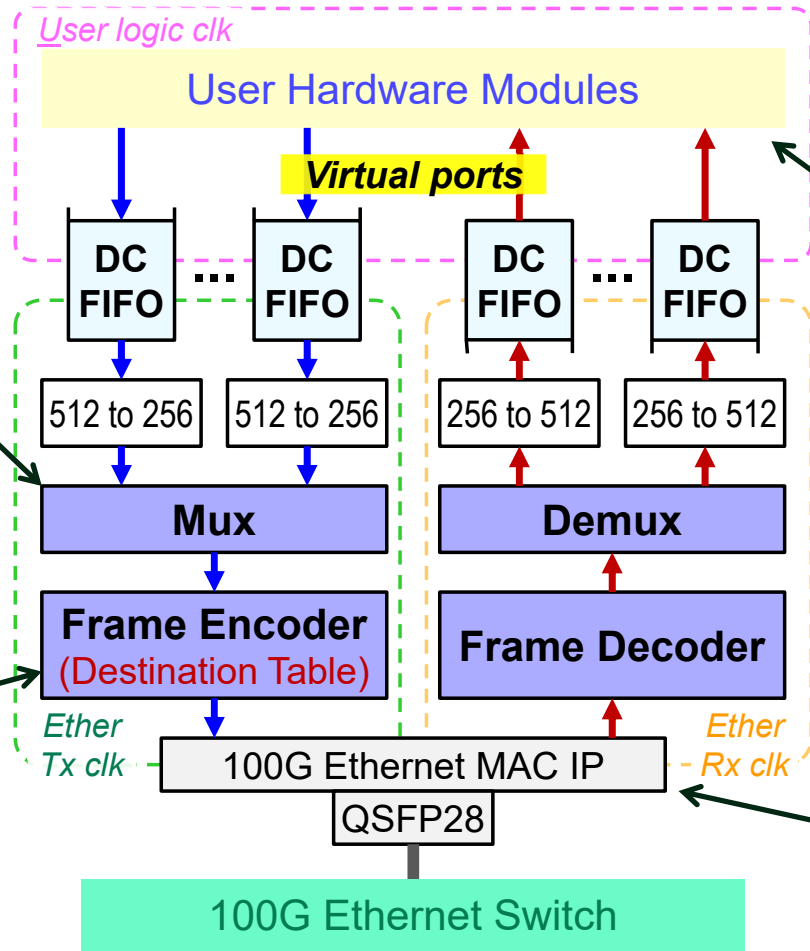
- Encode stream into Frames, or decode Frames into stream
- Destination MAC address is put on each Frame.

**User modules**

- Send / receive Avalon-ST data streams with multiple ports

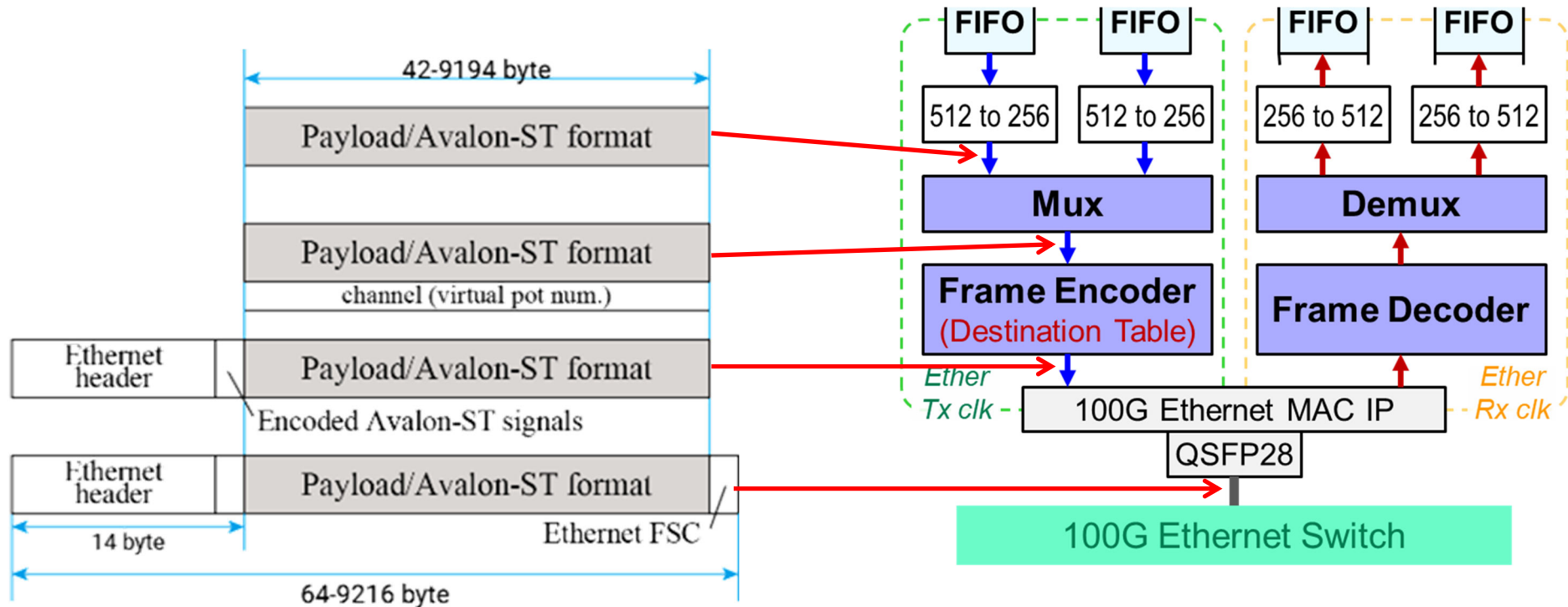
**100G Ethernet**

- Ethernet Frame (L1)
- Intel's Ether MAC IP



# Payload Efficiency of VCSN

- Theoretical max efficiency : **99.54%**  
due to Jumbo Frame



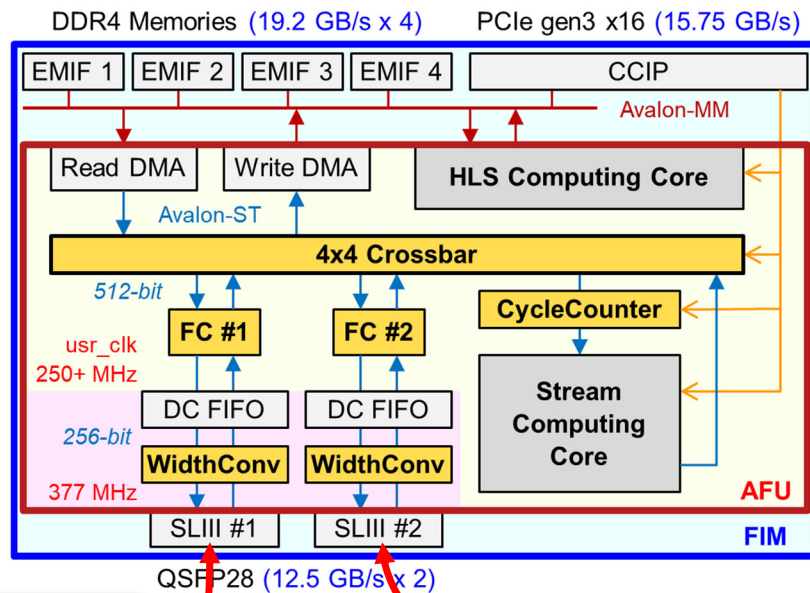
Ethernet FCS : Frame Check Sequence



# Implementation and Evaluation

# FPGA Shells for Direct and Indirect Networks

## Direct connection network (DCN)

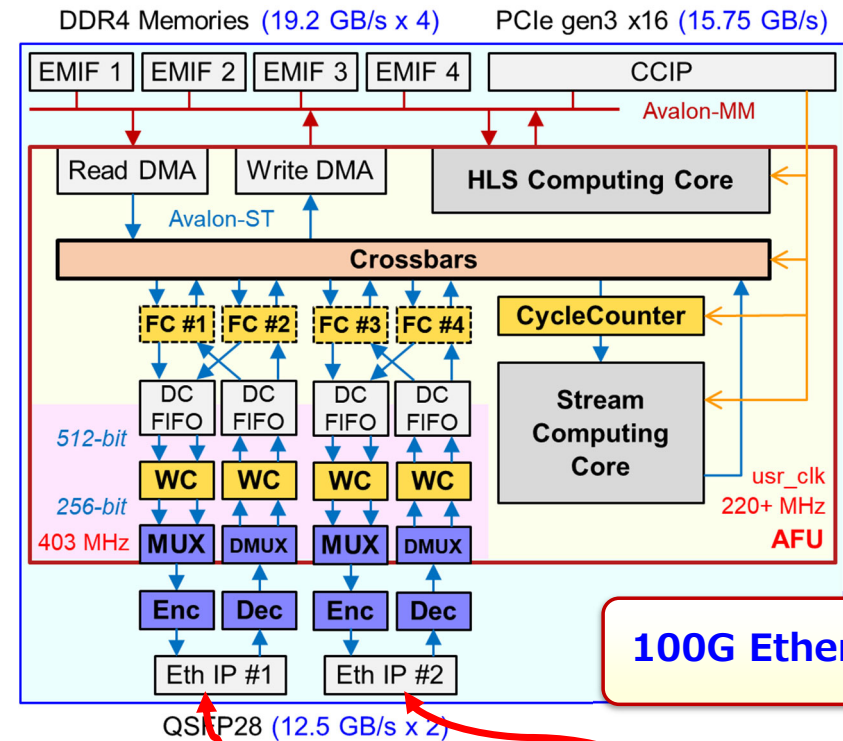


100G SerialLite3 with flow controller (FC)

Another FPGA board

Another FPGA board

## Indirect network (VCSN)



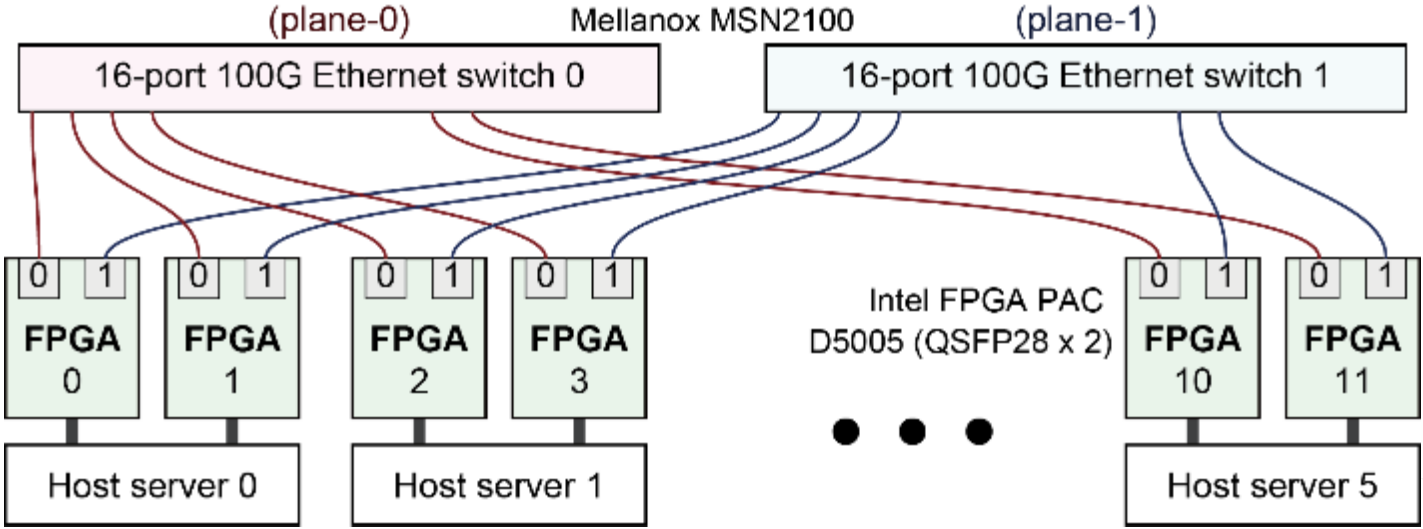
100G Ethernet

100G Ethernet SW

100G Ethernet SW



# VCSN Setup with 100Gbps Ethernet Switches



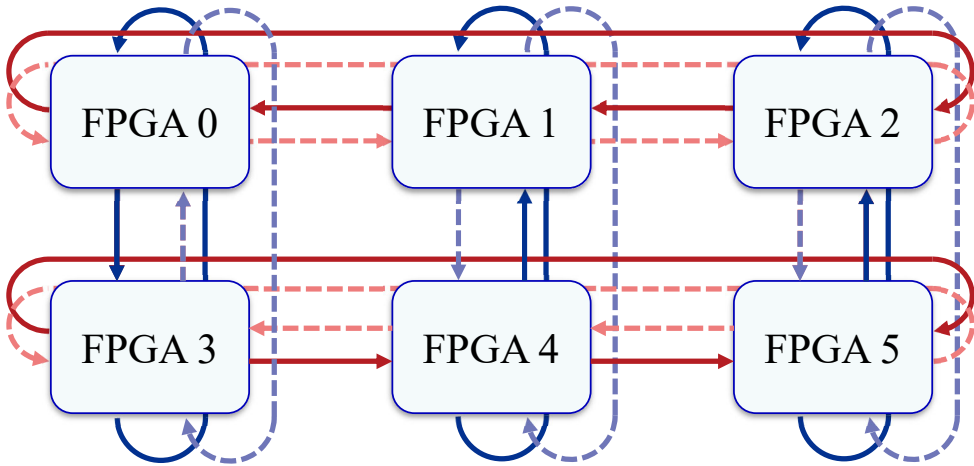
## Intel PAC D5005 FPGA cluster with VCSN

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

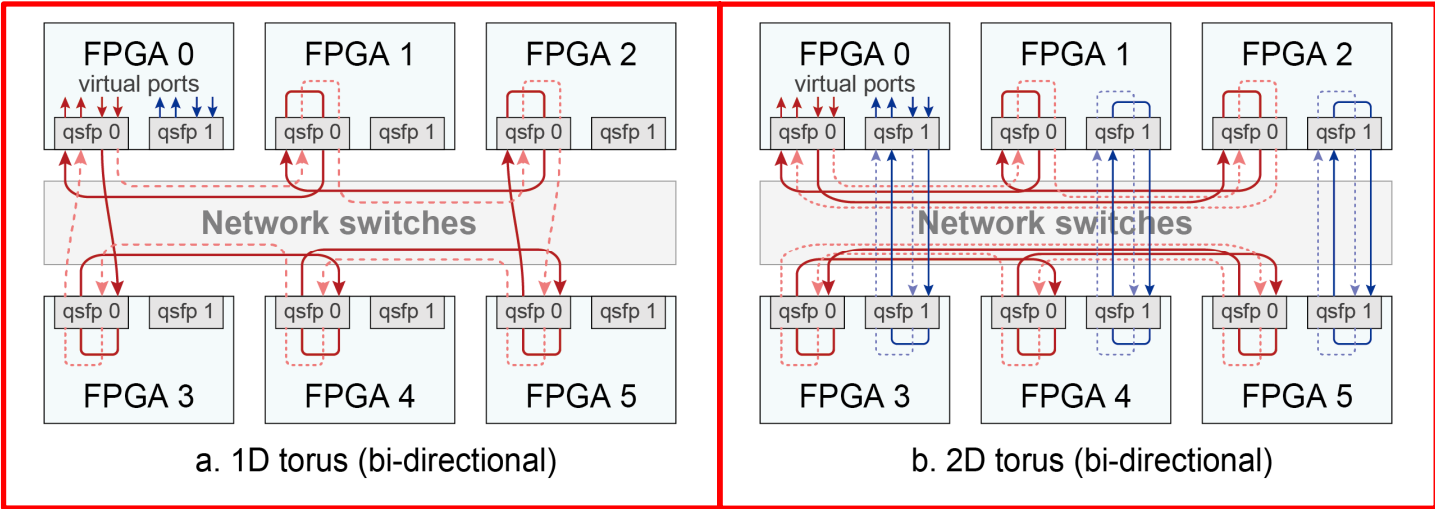
Tomohiro Ueno, Atsushi Koshiba, Kentaro Sano, "Virtual Circuit-Switching Network with Flexible Topology for High-Performance FPGA Cluster," Procs. of ASAP, pp.41-48, 2021.

# VCSN Configuration Examples

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



Virtual topology of bi-dir 2D Torus



Tomohiro Ueno, Atsushi Koshiba, Kentaro Sano, "Virtual Circuit-Switching Network with Flexible Topology for High-Performance FPGA Cluster," Procs. of ASAP, pp.41-48, 2021.

# Area and Latency

- Area

## DCN subsystem

	ALM	Registers	M20k	DSP
<b>SL3</b>	10474.2	10356	18	0
<b>Width converter</b>	1275.4	3626	0	0
<b>Flow controller</b>	1175.1	1452.4	80	0
<b>Network subsystem</b>	12924.7	15434.4	98	0
<b>Percentage</b>	<b>0.47%</b>	<b>0.14%</b>	<b>0.84%</b>	0.00%
<b>Stratix 10 SX280</b>	2753000	11012000	11721	5760

## VCSN subsystem

	ALM	Registers	M20k	DSP
<b>100G MAC</b>	22884.3	24922	0	0
<b>MUX/DEMUX</b>	9145.2	21458	28	0
<b>Encoder/decoder</b>	32818.6	43577	84	0
<b>FIFO</b>	2327.9	2336	832	0
<b>Width converter</b>	1275	3626	0	0
<b>Network subsystem</b>	68451.4	95919	944	0
<b>Percentage</b>	<b>2.49%</b>	<b>0.87%</b>	<b>8.05%</b>	0%
<b>Stratix 10 SX280</b>	2753000	11012000	11721	5760

- latency (minimum)

Network	path	latency [ns]
<b>VCSN</b>	virtual ports ⇔ virtual ports	<b>851.093</b>
<b>DCN</b>	Cable link	<b>243.137</b>
	Crossbar ⇔ crossbar	<b>490.942</b>

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

# Throughput (point-to-point)

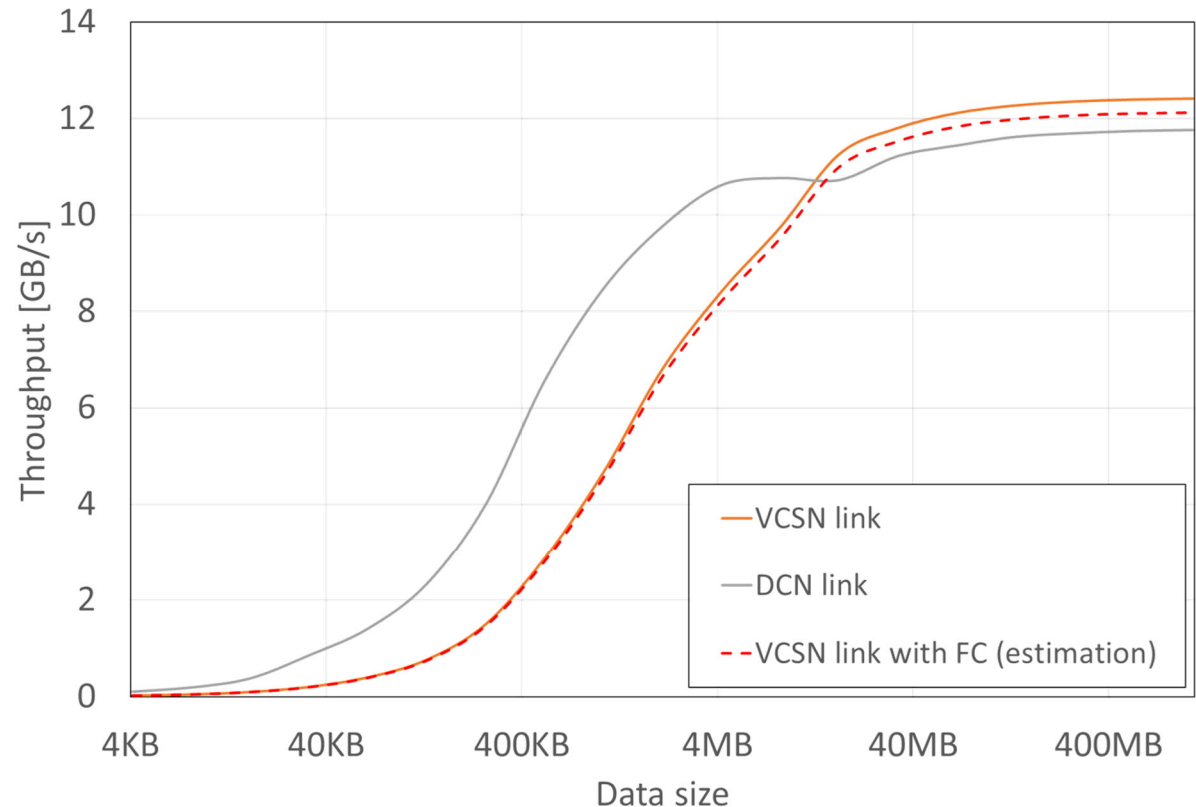
**Throughput of VCSN rises slowly due to higher latency.**

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

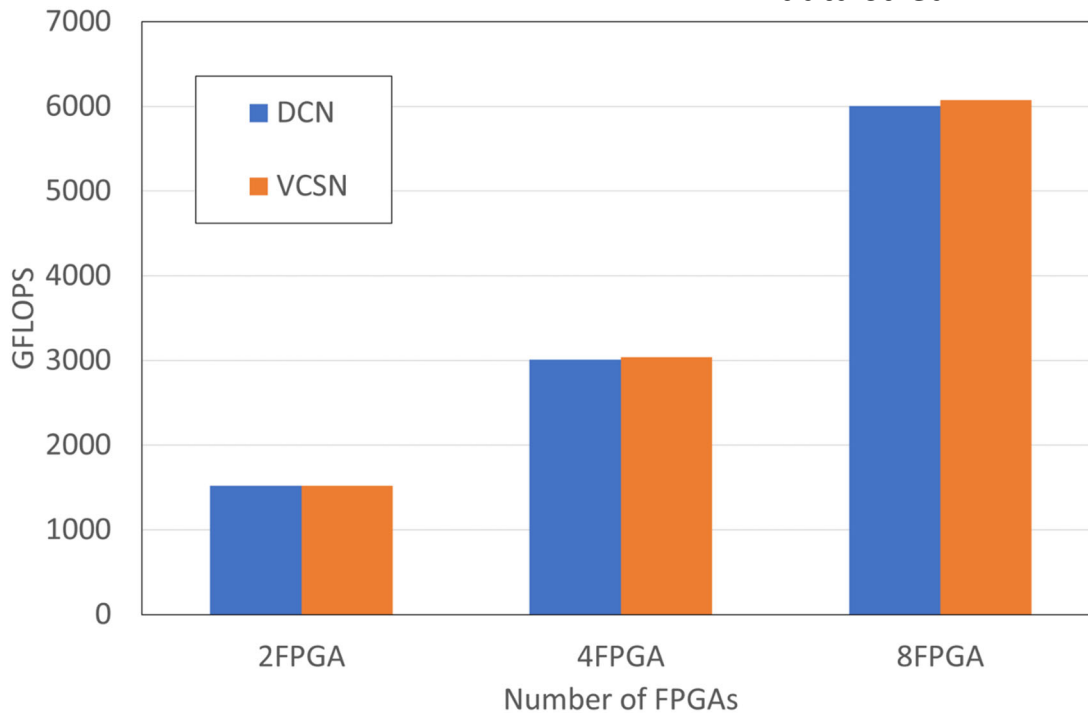
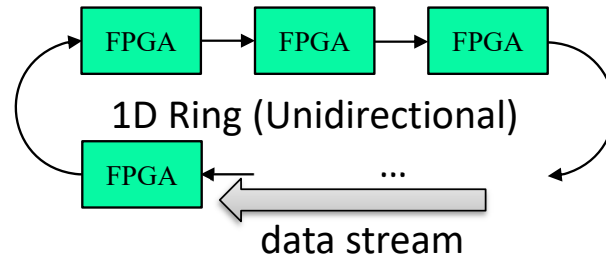
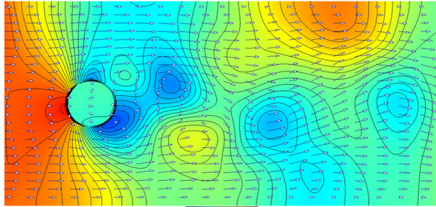
**VCSN has higher Max throughput.**

- ✓ 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  $\cong$  FLOPS by VCSN**

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

# Summary

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

**Our proposal** **Indirect network with VCSN**  
Virtualized circuit-switching network over Ethernet frame for higher flexibility

**Comparison** **x2 latency with slightly higher throughput**  
compared to DCN (direct connection network)

## Future work

- ✓ System software to configure VCSN (almost implemented)
- ✓ Evaluation with application cases

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

