About Us : Cybermedia Center, Osaka University

Cybermedia Center, Osaka University, Japan

As a resource provider of knowledge and technology derived from advanced researches conducted in Osaka University, the Cybermedia Center (CMC) offers support in the areas of large-scale computation, information communication, multimedia content and education. The center also works closely with educational and research organizations within Osaka University, as well as with industries and institutes outside the University. By sharing its resources and encouraging local communities to use its facilities for public lectures and other events, CMC has helped to create a more internationally-oriented IT society for the region.

Research Divisions

Informedia Education Research Division is involved in constructing an advanced information education environment, providing information and information ethics education, and conducting research and education activities for faculty development of information education staff.

Multimedia Language Education Research Division seeks to create an ideal environment for language education by developing an innovative, user-friendly learning management system for all language teachers/learners, and self-learning software for foreign-language learners. It also supports the operation and maintenance of several computer-assisted language laboratories, and provides students with opportunities to optimize their learning of foreign languages.

Large-Scale Computational Science Research Division is involved in assisting in the operation of the supercomputer system, disseminating technologies for visualizing computational results, providing education on advanced technologies for using the supercomputer system, and conducting education and research activities for computational science and other related courses.

Applied Information Systems Research Division conducts education and research into system architecture and operating technology involving large-scale data, to assist in the operation of our supercomputer and cloud systems, and to support users. It also performs research and education in the visualization of large-scale data and the architecture of cyber-physical systems.

University-wide Information and Communications Infrastructure Services Promotion Division is involved in promoting and managing the smooth execution and enhancement of university-wide support services which the Cybermedia Center is implementing, such as the maintenance, operation, and user-support of information communication systems installed for education, research, and clerical work.

Kyoto
Tokyo
CYBERMEDIA CENTER
Osaka

Location

SX-ACE

PC cluster

Computer Assisted Science Research Division supports efficient computer applications and education (relevant also to supercomputers) aimed at identifying and solving scientific problems. It also conducts education and research activities in mathematical and computational modeling of scientific problems.

Cybercommunity Research Division is involved in the design of digital libraries, cyber communities, and social networks, building information modeling (BIM), development of risk management systems for urban areas, and evaluation of urban infrastructure, while providing computer-aided design and graphic science education.

Advanced Network Environment Research Division supports the operation and utilization of the Osaka Daigaku Information Network System (ODINS), which introduces novel networking technologies such as high-speed networks, and mobile networking environments, with lower energy consumption. It also conducts educational activities on networking technologies, security issues, information ethics, etc., for university students and staff. In addition, it conducts state-of-the-art research on network-related topics.

Visit us at www.cmc.osaka-u.ac.jp

Cybermedia Center
Osaka University

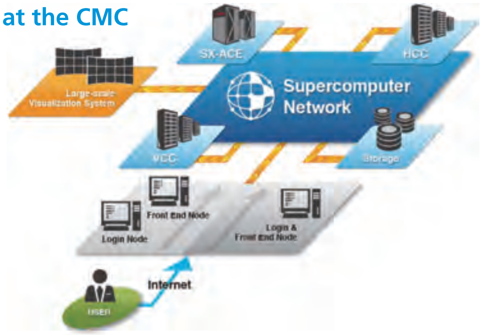
Large-scale Computing and Visualization Systems at the Cybermedia Center

Cybermedia Center, Osaka University, Japan

Large-scale Computing and Visualization Systems at the Cybermedia Center

Overview of High-performance Computing Environment at the CMC

Large-scale computing systems (SX-ACE, VCC and HCC), and large-scale visualization systems are deployed on CMC-Supercomputer network, a.k.a CMC-SCinet, a low-latency and wide-bandwidth network. This architectural design allows users to access to large-scale storage systems, perform large-scale high-performance computation and analysis on our large-scale computing systems, and then visualize its computation and analysis results without loosing any important information on our large-scale visualization system.



Large-scale Computing System

The large-scale computing systems at the CMC are classified into (1) Vector-typed Supercomputer and (2) Scalar-typed Supercomputer.

Type: Vector
OS: Super UX
# of nodes: 1536
# of cores : 6144
Total memory: 96 TB
Peak performance: 423 TFlops

SX-ACE

SX-ACE is a "clusterized" vector-typed supercomputer, composed of 3 clusters, each of which is composed of 512 nodes. Each node equips 4-core multi-core CPU and a 64 GB main memory. These 512 nodes are interconnected on a dedicated and specialized network switch, called IXS (Internode Crossbar Switch) and forms a cluster. Note that IXS interconnects 512 nodes with a single lane of 2-layer fat-tree structure and as a result exhibits 4 GB/s for each direction of input and output between nodes.

Library

MathKeisan (BLAS, LAPACK, etc)
ASL, ASLSTAT, ASLQUAD
MPI/SX
HPF/SX
XMP

Type: Scalar
OS: Linux
# of nodes: 62
# of cores : 1240
Total memory: 3.968 TB
Peak performance: 24.8 TFlops
Accelerator: NVIDIA Tesla K20 × 51

VCC (PC Cluster for large-scale visualization)

PC cluster for large-scale visualization (VCC) is a cluster system composed of 62 nodes. Each node has 2 Intel Xeon E5-2670v2 processors and a 64 GB main memory. These 62 nodes are interconnected on InfiniBand FDR and forms a cluster. Also, this system has introduced ExpEther, a system hardware virtualization technology. Each node can be connected with extension I/O nodes with which GPU resource, and SSD on 20 Gbps ExpEther network. A major characteristic is that this cluster system is reconfigured based on user's usage and purpose by changing the combination of node and extension I/O node.

HCC (General-Purpose PC Cluster)
Type: Scalar (VM)
OS: Linux
# of nodes: 575
# of cores : 1150
Total memory: 2.6 TB
Peak performance: 16.6 TFlops

IPC-C (Image Processing PC Cluster on Campus)
Type: Scalar
OS: Windows/Linux
# of nodes: 7
# of cores : 84
Total memory: 448 GB
Peak performance: 1.68 TFlops
Accelerator: NVIDIA Quadro K5000 × 6

IPC-U (Image Processing PC Cluster on Umekita)
Type: Scalar
OS: Windows/Linux
# of nodes: 6
# of cores : 72
Total memory: 384 GB
Peak performance: 1.44 TFlops
Accelerator: NVIDIA Quadro K5000 × 6

Library

Intel MKL (BLAS, LAPACK, etc)
intelMPI, OpenMPI

Application

GROMACS	Gaussian09	Marc / Mentat	Nastran
GROMACS for GPU	IDL	Dytran	AVS/Express (DEV/PCE/MPE)
LAMMPS	NEC Remote Debugger	Patran	OpenFOAM
LAMMPS for GPU	NEC Ftrace Viewer	Adams	

Large-scale Visualization System

The large-scale visualization systems at the CMC are set up on Campus and on CMC's Umekita Office. Large-scale and interactive visualization processing becomes possible through the dedicated use of PC cluster for large-scale visualization (VCC) on these systems.

24-screen Flat Stereo Visualization System

The visualization system is composed of 24 50-inch Full HD (1920 x 1080) stereo projection module (Barco OLS-521), Image-Processing PC cluster (IPC-C) driving visualization processing on 24 screens. A notable feature of this visualization system is that it enables approximately 50 million high-definition stereo display with horizontal 150 degree view angle.

15-screen Cylindrical Stereo Visualization System

This visualization system is composed of 15 46-inch WXGA (1366 x 768) LCD, and Image-Processing PC Cluster (IPC-U) driving visualization processing on 15 screens. A notable characteristic of this visualization system is that it enables approximately 16-million-pixel high-definition stereo display.

OSAKA UNIVERSITY

Contact : system@cmc.osaka-u.ac.jp
http://www.hpc.cmc.osaka-u.ac.jp/

Scalable and Low-latency Communication Method for Reliability Improvement of SDN MPI_Bcast

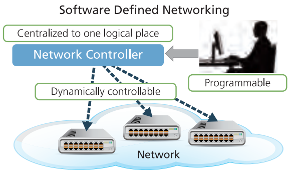
Cybermedia Center, Osaka University, Japan

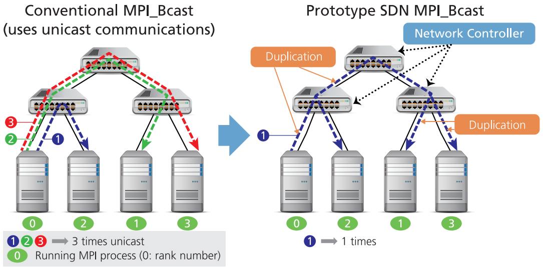
Scalable and Low-latency Communication Method for Reliability Improvement of SDN MPI_Bcast

Communication time of MPI_Bcast collective tends to get longer on a large-scaled cluster.

Previous Work

Our previous work implements MPI_Bcast through duplication of broadcast data on the fly from source process to others leveraging SDN. As the result, source process sends data only once for broadcasting data.


Software Defined Networking



Problem of Previous Work

Data delivery from source process to others is not guaranteed in prototype SDN MPI_Bcast.

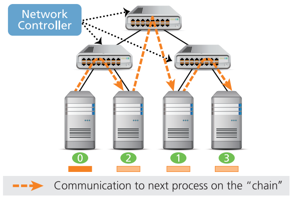
Research Goal

To implement scalable and low-latency "chain" communication method for improving reliability of prototype SDN MPI_Bcast.

- All receiving processes need to let source process know they received data.

Proposal

Each process sends data to next process on the "chain" for the acknowledgement of data receiving.



"chain": series of all processes placed in a line.
Eg. 0→2→3→1, 0→1→2→3

Reliable SDN MPI_Bcast has two stages.
1. Source process sends data using prototype SDN MPI_Bcast.
2. Each process sends data to next process on the "chain" as soon as receives it.

Low-latency: Network controller generates the "chain" considering network topology and process placement
Scalable: Each process responsible for only one process' data delivery

Khureltulga Dashdavaa*, Munkhdorj Baatarsuren†, Keichi Takahashi*, Susumu Date*, Yoshiyuki Kido*, and Shinji Shimojo* *Osaka University, Japan, †The University of Tokyo, Japan

Contact : Khureltulga Dashdavaa huchka@ais.cmc.osaka-u.ac.jp

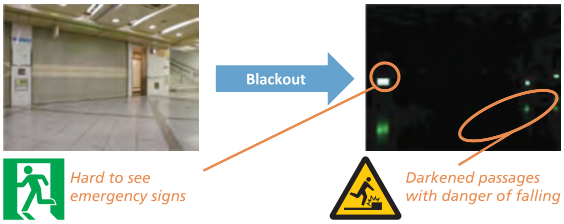
Indoor Evacuation System With Smartphones That Helps Evacuees In a Blacked Out Building

Cybermedia Center, Osaka University, Japan

Indoor Evacuation System With Smartphones That Helps Evacuees In a Blacked Out Building

1. Introduction

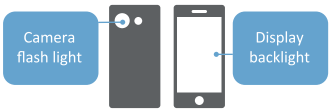
When a disaster happens, people in a bulding have to escape as quickly as possible.
However, *Power failure* may occur and prevent people from escaping.


Blackout
Hard to see emergency signs
Darkened passages with danger of falling

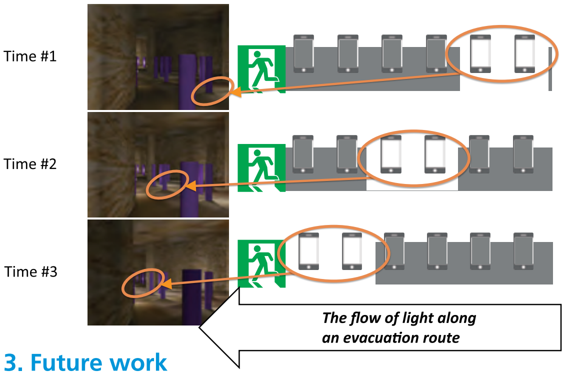
As a result, people take long time to escape and are in danger of losing them lives.

2. Our proposal

Our research group has proposed *an indoor evacuation system that utilizes people's smartphones*



Our system provides information of the evacuation routes and illuminates the passages around people at the same time.
・Illuminates passages by *turning smartphone's lights on*
・Indicates evacuation routes by *making each smartphone's light blinks*



3. Future work

In order to implement our system, *an indoor positioning method for smartphones* is needed.
A method that *finds the most appropriate evacuation route* is also needed.

Contact: Takuya Yamada yamada.takuya@ais.cmc.osaka-u.ac.jp

Cybermedia Center
Osaka University

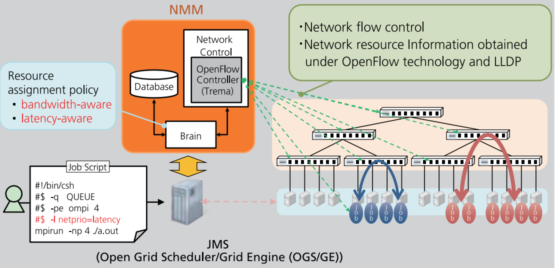
Does SDN Technology Make HPC Guys Happy?

R&D and Empirical Studies Towards HPC With Enhanced Network Controllability

Cybermedia Center, Osaka University, Japan

Does SDN Technology Make HPC Guys Happy?

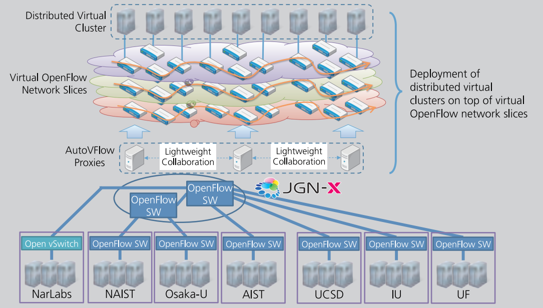
SDN-enhanced Job Management System overturns the common sense that network is static and uncontrollable resource


NMM
Network Control
OpenFlow Controller (Trema)
Database
Brain
Resource assignment policy
▪ bandwidth-aware
▪ latency-aware
·Network flow control
·Network resource Information obtained under OpenFlow technology and LLDP
Job Script
Job

```
#!/bin/csh
#$  -q   QUEUE
#$  -pe  ompi  4
#$  -l netprio=latency
mpirun  -np 4 ./a.out
```

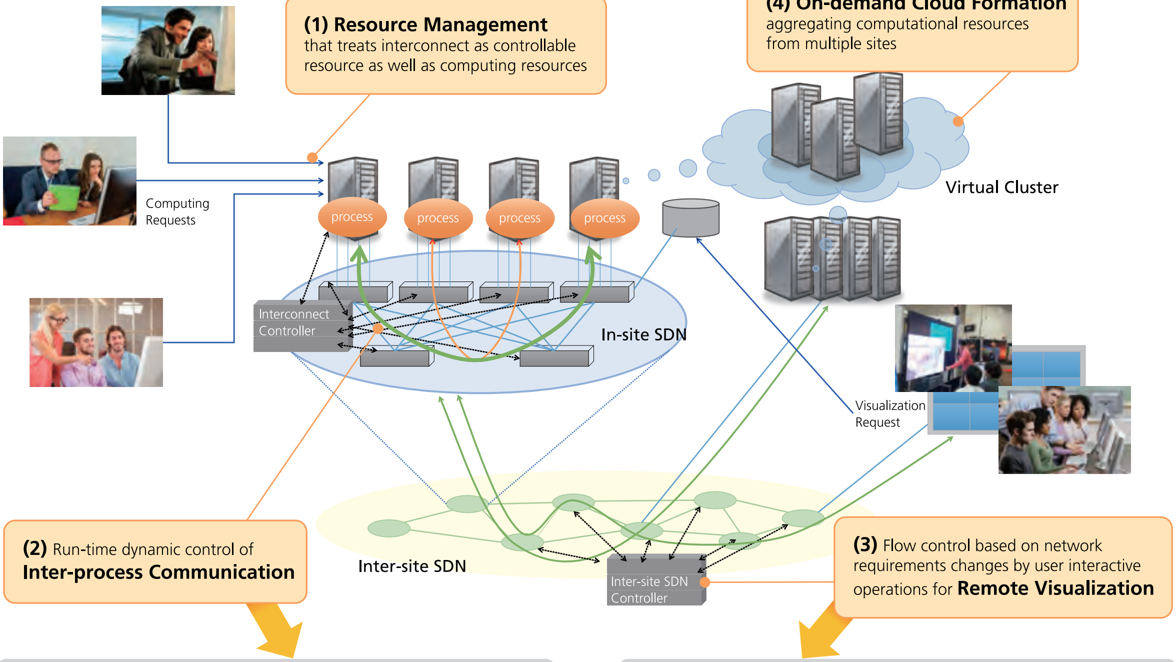
JMS
(Open Grid Scheduler/Grid Engine (OGS/GE))

Multi-site Virtual Cluster aggregates a set of computational and network resources into a HPC cloud


Distributed Virtual Cluster
Virtual OpenFlow Network Slices
Deployment of distributed virtual clusters on top of virtual OpenFlow network slices
AutoVFlow Proxies
Lightweight Collaboration
Lightweight Collaboration
OpenFlow SW
OpenFlow SW
JGN-X
Open vSwitch
OpenFlow SW
OpenFlow SW
OpenFlow SW
OpenFlow SW
OpenFlow SW
OpenFlow SW
NarLabs
NAIST
Osaka-U
AIST
UCSD
IU
UF

(1) Resource Management that treats interconnect as controllable resource as well as computing resources

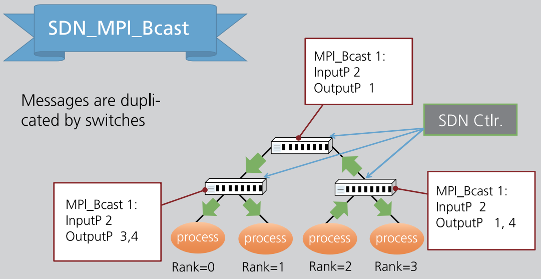
(4) On-demand Cloud Formation aggregating computational resources from multiple sites



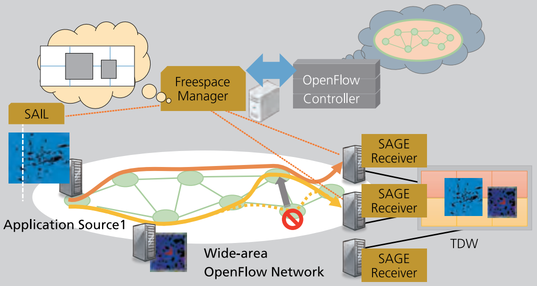
(2) Run-time dynamic control of **Inter-process Communication**

(3) Flow control based on network requirements changes by user interactive operations for **Remote Visualization**

SDN-MPI reduces the execution time of MPI program by shortening MPI communication time occurred during parallel computation



SDN-enhanced SAGE manages both tiled display walls and network resources for smooth high-res scientific visualization



OSAKA UNIVERSITY
NICT

Visit us at www.cmc.osaka-u.ac.jp

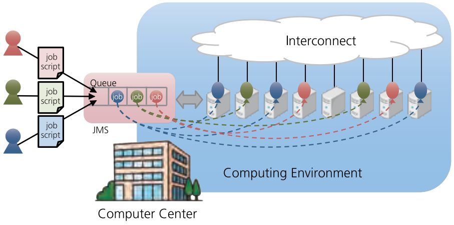
Mechanism for Handling Network and Virtualized Computational Resources on SDN-enhanced Job Management System

Cybermedia Center, Osaka University, Japan

Mechanism for Handling Network and Virtualized Computational Resources on SDN-enhanced Job Management System

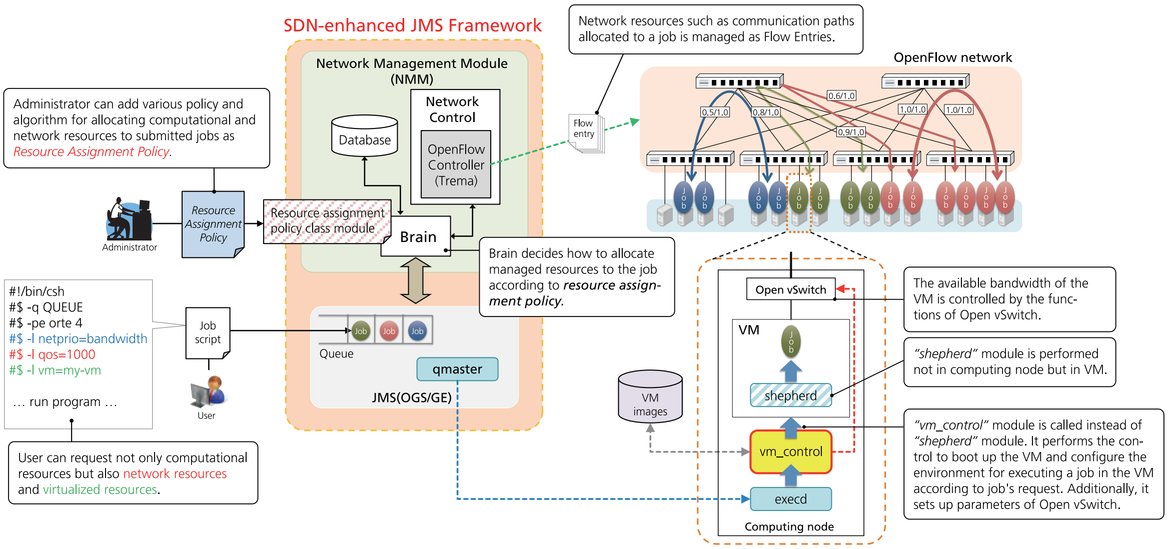
Motivation and Objectives

Nowadays, users' computation requests to a high-performance computing (HPC) environment have been increasing and diversifying for performing large-scale simulations and analysis in the various science fields. Since computer center flexibly complies such computation requests, efficient and flexible resource management system is essential for guaranteeing high performance computing capabilities for multiple users and gaining high job-throughput in computing environment.



Mechanism on SDN-enhanced JMS Framework

We have been studying and developing a novel Job Management System (JMS) for various resources of computing environment. For handling interconnect as network resources, the *SDN-enhanced JMS Framework* has been implemented by leveraging Software-Defined Networking (SDN) concept, which can dynamically control an entire network in a centralized manner [1]. Moreover, we have also been developing a mechanism for deploying job's processes to virtual machines (VMs) on computing nodes, and controlling available bandwidth on communication paths allocated to a job by using QoS functions of Open vSwitches (OVSs) connected with VMs.



Administrator can add various policy and algorithm for allocating computational and network resources to submitted jobs as *Resource Assignment Policy.*

```
#!/bin/csh
#$ -q QUEUE
#$ -pe orte 4
#$ -l netprio=bandwidth
#$ -l qos=1000
#$ -l vm=my-vm

 … run program …
```

User can request not only computational resources but also network resources and virtualized resources.

Network resources such as communication paths allocated to a job is managed as Flow Entries.

Brain decides how to allocate managed resources to the job according to *resource assignment policy.*

The available bandwidth of the VM is controlled by the functions of Open vSwitch.

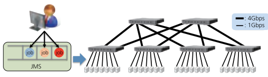
"shepherd" module is performed not in computing node but in VM.

"vm_control" module is called instead of *"shepherd"* module. It performs the control to boot up the VM and configure the environment for executing a job in the VM according to job's request. Additionally, it sets up parameters of Open vSwitch.

Evaluation

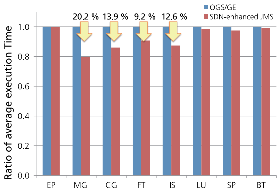
1) Effectiveness of job execution time by network resource management

An experiment was conducted to assess the effectiveness of network resource management. In the experiment, multiple pararell jobs, each of which generates four processes for executing NAS Parallel Benchmarks (Class B) were submitted. As a result, the SDN-enhanced JMS Framework achieved the reduction of average job execution time, even if a cluster system with fat-tree interconnect has enough bandwidth capacity.



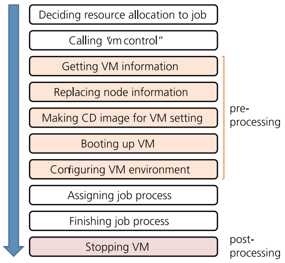
Specifiation of Cluster System

Node	OS	CentOS 6.3
	CPU	Intel Xeon E5-2620(2.00GHz) x2
	Memory	64GB (DDR3-1600 8GB DIMM x8)
	Network	on board Intel I350 GbE
OpenFlow Switch		NEC UNIVERGE PF5240



2) Overhead to handle virtualized computational resources

We measured the overhead caused by the *vm_control* for managing VMs on the SDN-enhanced JMS Framework. The overhead of the *vm_control* has the pre-processing and the post-processing: the configuration process to prepare the environment in a VM before it starting, and removing additional setting for the VM after it stopping individually.



Average overhead time (sec)

Number of processes	Pre-processing	Post-processing
1	1.299	0.004
2	1.315	0.004
3	1.14	0.004
4	1.196	0.004

Acknowledgments

This research was supported in part by the collaborative research of the National Institute of Information and Communications Technology (NICT) and Osaka University (Research on High Functional Network Platform Technology for Large-scale Distributed Computing).

[1] Y. Watashiba, S. Date, H. Abe, Y. Kido, K. Ichikawa, H. Yamanaka, E. Kawai, S. Shimojo, and H. Takemura, "Performance Characteristics of an SDN-enhanced Job Management System for Cluster Systems with Fat-tree Interconnect", Emerging Issues in Cloud (EIC) Workshop, The 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2014), pp. 781-786, December 2014.

Contact : Yasuhiro Watashiba, E-mail : watashiba-y@cmc.osaka-u.ac.jp
Masaharu Shimizu, E-mail : shimizu.masaharu@ais.cmc.osaka-u.ac.jp
Web : http://hpc-sdn.ime.cmc.osaka-u.ac.jp/sdnjms/

Cybermedia Center
Osaka University

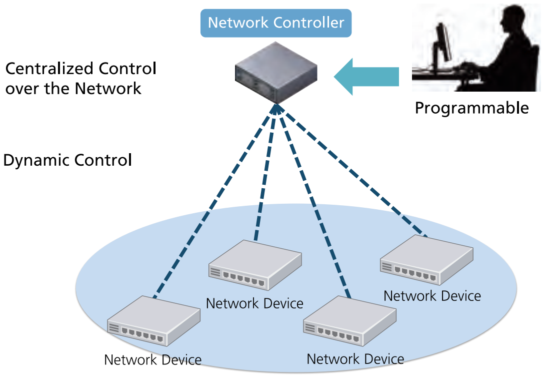
Architecture of SDN-enhanced MPI Framework

Cybermedia Center, Osaka University, Japan

Architecture of SDN-enhanced MPI Framework

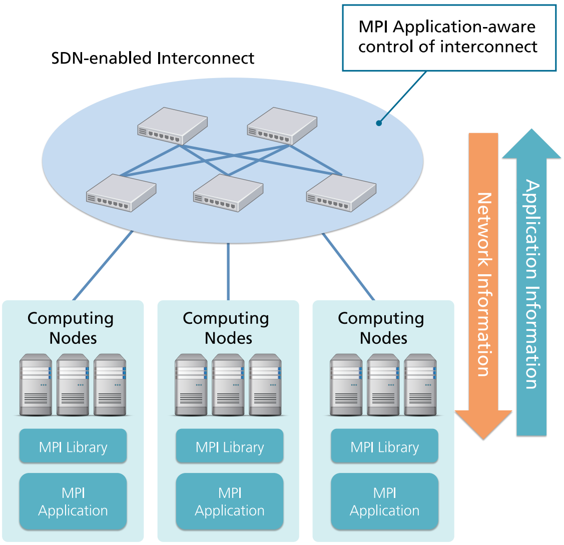
1. Software-Defined Networking (SDN)

Software-Defined Networking (SDN) is a new concept of network architecture that decouples conventional networking function into a programmable control plane (responsible for deciding how to control the packets) and a data plane (responsible for the actual packet delivery). Currently, OpenFlow is the most common implementation of SDN, which enables to dynamically control the forwarding functionality of network from a centralized controller.

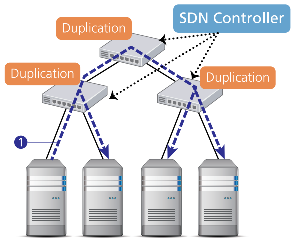


2. Basic Idea Behind the SDN-enhanced MPI Framework

Practical HPC systems are often deployed with an exceedingly low-latency and high-throughput network. However, this approach is getting increasingly difficult and expensive as a result of the recent rapid scale-out in node number. We have been developing SDN-enhanced MPI based on the idea that a mechanism that configures and controls the network of a cluster system depending on the requirement of each application is essential. The key concept of SDN-enhanced MPI is to utilize the underlying network of a computer cluster to its maximum capacity by fully leveraging the flexible network controllability of SDN.

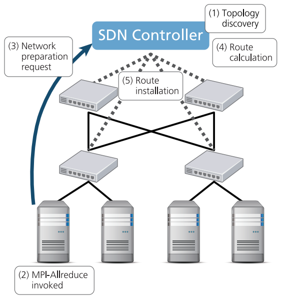


3. SDN-enhanced MPI Communication Functions



A. SDN_MPI_Bcast

SDN_MPI_Bcast is an SDN-enhanced version of MPI_Bcast, which is the broadcasting function in MPI. SDN_MPI_Bcast offloads packet duplication operations during the broadcast onto SDN switches. As a result, SDN_MPI_Bcast has successfully decreased the number of communications and communication latency of MPI_Bcast.

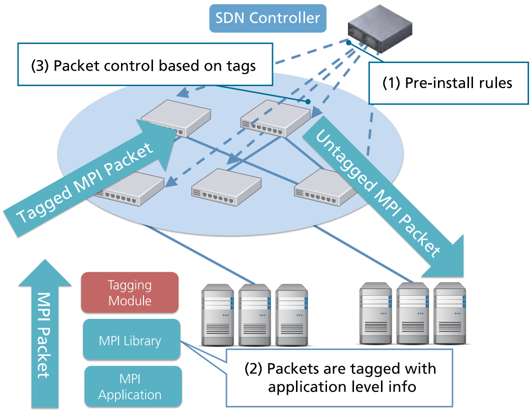


B. SDN_MPI_Allreduce

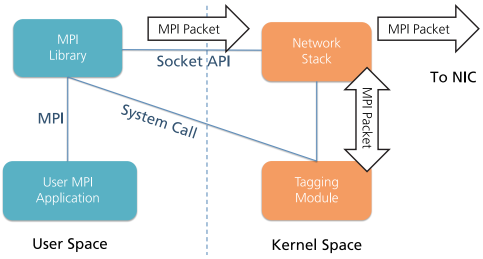
SDN_MPI_Allreduce is an SDN-enhanced version of MPI_ Allreeduce. Since MPI_Allreduce requires multiple simultaneous communication between nodes, congestion may happen on a interconnect without full bisection bandwidth. We employ a real-time traffic load balancing method using SDN to solve this problem.

4. Architecture of SDN-enhanced MPI Framework

We propose an integrated framework to combine SDN-MPI components that we have developed in our previous works. In this framework, MPI packets are tagged with MPI-layer information which are used by the SDN switches to determine how to control the packets.



Our implementation embeds a tag into the L2 header. The figure below illustrates the packet tagging mechanism.



Contact: Keichi Takahashi takahashi.keichi@ais.cmc.osaka-u.ac.jp
Khureltulga Dashdavaa huchka@ais.cmc.osaka-u.ac.jp

OSAKA UNIVERSITY
NICT