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.

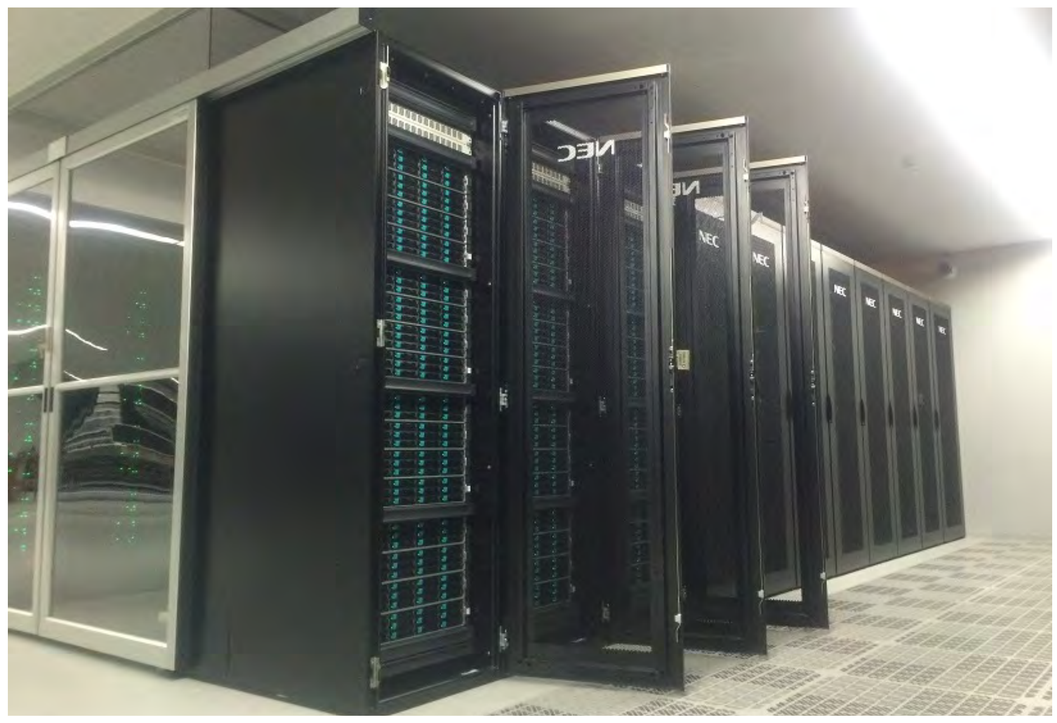
Multimedia Language Education Research Division seeks to create an ideal environment for language education by developing an innovative, user-friendly learning management system for all language teachers/learners, and self-learning software for foreign-language learners. It also supports the operation 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 super computer 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.

Vector supercomputer



Vector supercomputer

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.

Location

Cybermedia Center
Osaka University

New Supercomputer System SX-ACE at the Cybermedia Center

Cybermedia Center, Osaka University, Japan

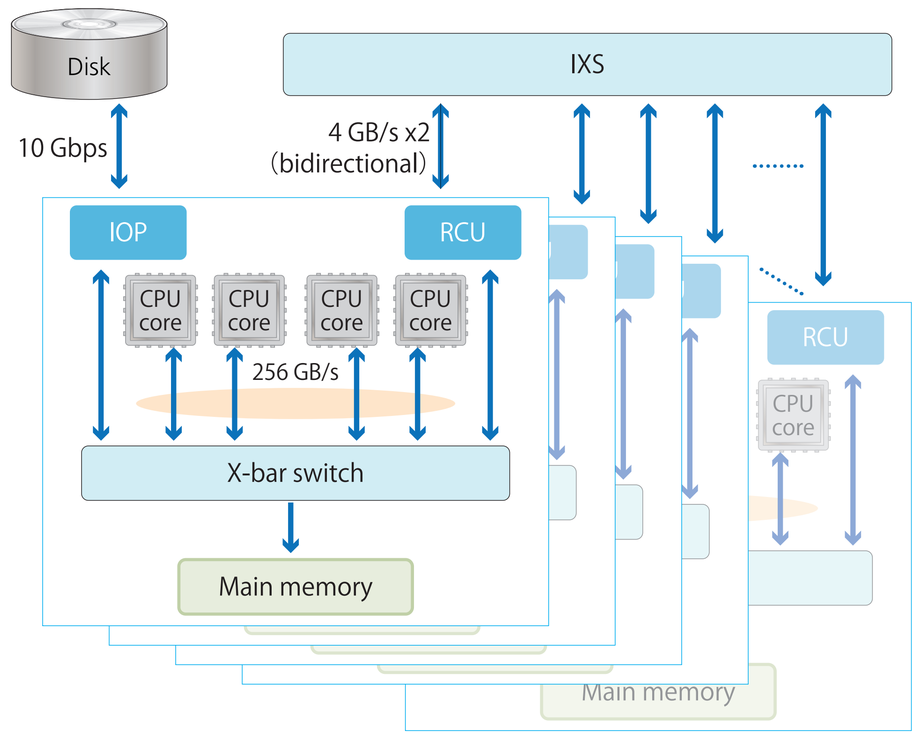
System Overview

The SX-ACE, newly introduced by the Cybermedia Center (CMC) is a "clusterized" vector-typed supercomputer, composed of 3 clusters, each of which is composed of 512 nodes. Each node is equipped with 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. In the CMC, 2 Peta-byte storage is managed on NEC Scalable Technology File System (ScateFS), NEC-developed fast distributed and parallel file system, so that it can be accessed from the large-scale computing systems including the SX-ACE at the CMC

Node Performance

As a single SX-ACE node has a multi-core vector-typed processor composed of 4 cores, each of which exhibits 64 GFlops vector performance, and a 64 GB main memory, the vector performance per node becomes 256 GFlops. On the other hand, the maximum transfer between processor and main memory is performed with 256 GB/s. This fact indicates that a single SX-ACE node achieves high memory-bandwidth performance of 1 Byte/Flops, taking higher CPU performance into consideration.

Internode communication enables 4GB/s x 2 (bi-directional) high-bandwidth data communication with a specialized internode communication control unit named RCU connected to IXS. Also, communication between node and disk storage enables 10 Gbps data communication with I/O control unit called IOP.



System Performance

	SX-ACE		
	Per-node	1 cluster	3 cluster
# of CPU	1	512	1536
# of core	4	2048	6144
Performance	276 GFLOPS	141 TFLOPS	423 TFLOPS
Vector performance	256 GFLOPS	131 TFLOPS	393 TFLOPS
Main memory	64 GB	32 TB	96 TB
Storage	2 PB		

The SX-ACE which the CMC has introduced is composed of 3 clusters (1536 nodes). Theoretical peak performance of the SX-ACE at the CMC is derived as the left table indicates.

Importantly, note that performance is the sum of vector-typed processor and scalar processor on SX-ACE. SX-ACE has a 4-core multi-core vector-typed processor and a single scalar processor.

IT Core Annex

IT Core Annex is a new datacenter that aims to aggregate and accommodate computer systems and supercomputer systems on campus for energy-efficient administration and management. It was designed and built based on the careful consideration on air flow and circulation for efficient cooling. The SX-ACE at the CMC is set up at the IT Core Annex.

A remarkable feature of IT Core Annex has two indirect evaporative air cooling systems to improve PUE for more energy-efficient driving of the newly introduced supercomputer system SX-ACE at the CMC.

OSAKA UNIVERSITY

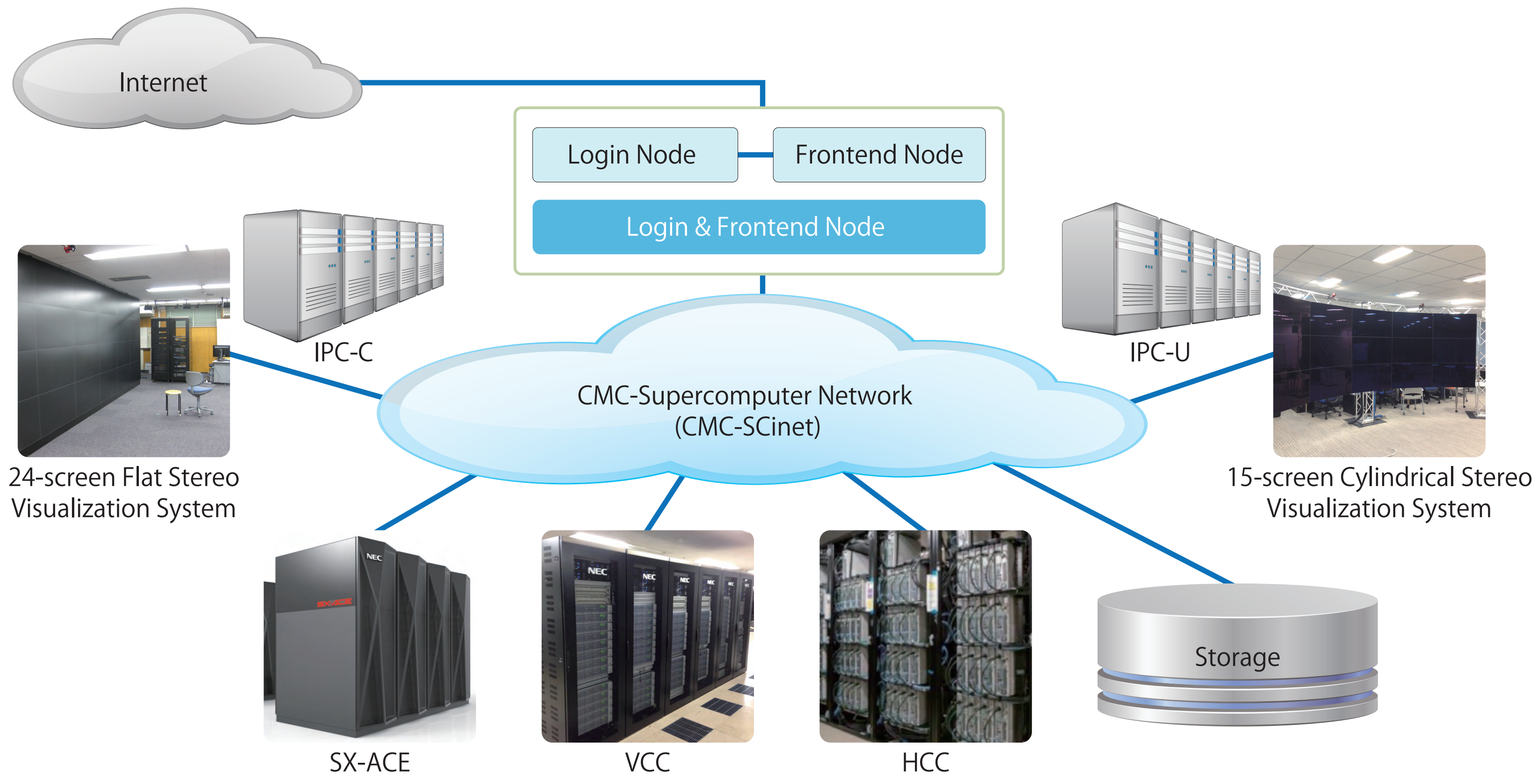
Contact : system@cmc.osaka-u.ac.jp
http://www.hpc.cmc.osaka-u.ac.jp/

Cybermedia Center
Osaka University

Large-scale Computing and Visualization Systems at the Cybermedia Center

Cybermedia Center, Osaka University, Japan

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 on our large-scale visualization system without loosing any important information.

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
The newly introduced SX-ACE by the CMC is a "clusterized" vector-typed supercomputer, composed of 3 clusters, each of which is composed of 512 nodes. Each node has 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.

Type: Scalar
OS: Linux
# of nodes: 56
# of cores: 1120
Total memory: 3.584 TB
Peak performance: 22.4 TFlops
Accelerator: NVIDIA Tesla K20 × 48

VCC (PC Cluster for large-scale visualization)
PC cluster for large-scale visualization (VCC) is a cluster system composed of 56 nodes. Each node has 2 Intel Xeon E5-2670v2 processors and a 64 GB main memory. These 56 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 20Gbps 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
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 × 7

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

Large-scale Visualization System

The large-scale visualization systems at the CMC are set up on Osaka U. 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

This visualization system is composed of 24 50-inch Full HD (1920x1080) 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 (1366x768) 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 very high-definition stereo display.

OSAKA UNIVERSITY

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

Cybermedia Center
Osaka University

Visualization Services at Cybermedia Center

Cybermedia Center, Osaka University, Japan

Visualization Service at Cybermedia Center

High Resolution Stereo Visualization Systems

Cybermedia Center at Osaka University installed two types of high-resolution stereo visualization systems, one on Toyonaka campus and the other at Umekita office near JR Osaka station. These systems are used for scientific visualization, information visualization, visual analytics, and other research activities as well as showcasing, exhibitions and other outreach activities to general public with large-scale scientific data processed by high performance computing. A variety of visualization and virtual reality software packages also are available including AVS Express, EasyVR, Fusion VR and VR4Max.

24-screen Flat Stereo Visualization System @ Toyonaka Campus

* Full HD (1920x1080) 50-inch Stereo Projection Module x 24 (approx. 50 million pixels)
* Image Processing PC with NVIDIA K5000 x 7
* HD Video Conferencing System x 1
* Motion Capturing System x 1

15-screen Cylindrical Stereo Visualization System @ Umekita Office

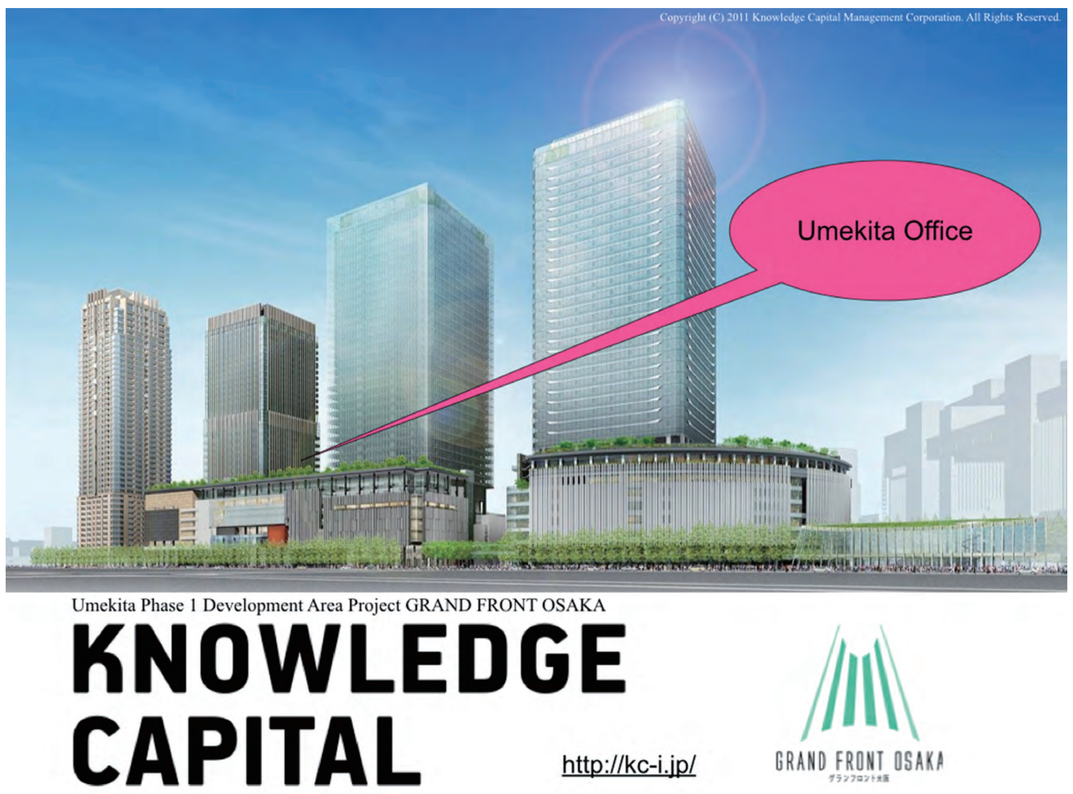
* WXGA (1366x768) 46-inch LCD x 15 (approx. 16 million pixels)
* Image Processing PC with NVIDIA K5000 x 6
* HD Video Conferencing System x 1
* Motion Capturing System x 1

Visualization Services at Cybermedia Center

Following visualization services are now available through HPCI (High-Performance Computing Infrastructure (https://www.hpci-office.jp/folders/english)) and JHPCN (Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (http://jhpcn-kyoten.itc.u-tokyo.ac.jp/en/)).

(1) Training sessions, seminars, and workshops: Opportunities are given to users to learn how to take advantage of our visualization systems. Also, opportunities to learn and share the cutting-edge visualization technologies and techniques is offered.

(2) Consultation: Technical consultations are offered through which users can learn best suitable visualization techniques for their problems from a diversity of solutions which dramatically reduce time and effort required to produce satisfactory visualization results.



Umekita Office space: Umekita Office of Cybermedia Center near JR Osaka station is provided for a variety of research and outreach activities such as discussions, seminars and workshops on HPC and visualization. Reservation can be made through a web system.

Use Case: Kumikomi Tekijuku (http://www.kansai-kumikomi.net/en/) is a technical seminar on embedded software development, which was held by tele-conference between Umekita Office and Tohoku University using Video Conferencing System.

OSAKA UNIVERSITY

Contact : vis-support@cmc.osaka-u.ac.jp
http://vis.cmc.osaka-u.ac.jp

An MPI Framework Enhanced for SDN Architecture Cluster

Cybermedia Center, Osaka University, Japan

An MPI framework Enhanced for SDN Architecture Cluster

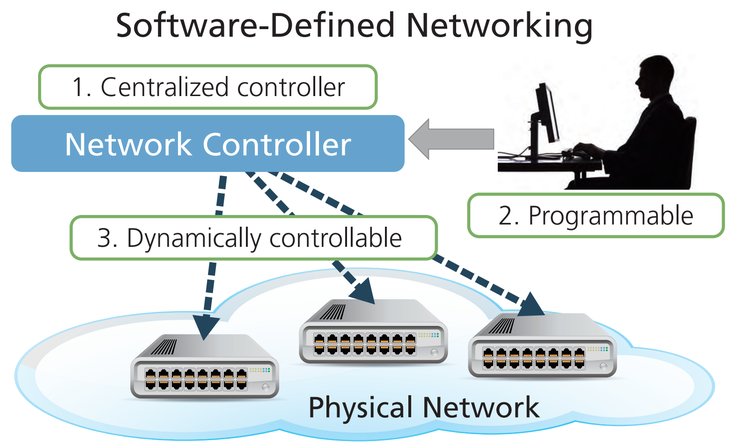
Message Passing Interface

Message Passing Interface (MPI) has been a *de facto* standard programming model in parallel computing for around two decades. MPI offers two types of communications:

1. *One-to-one communication:* used for low-level description of communication pattern.
2. *Collective communication:* used for high-level and human-friendly description.

Software-Defined Networking

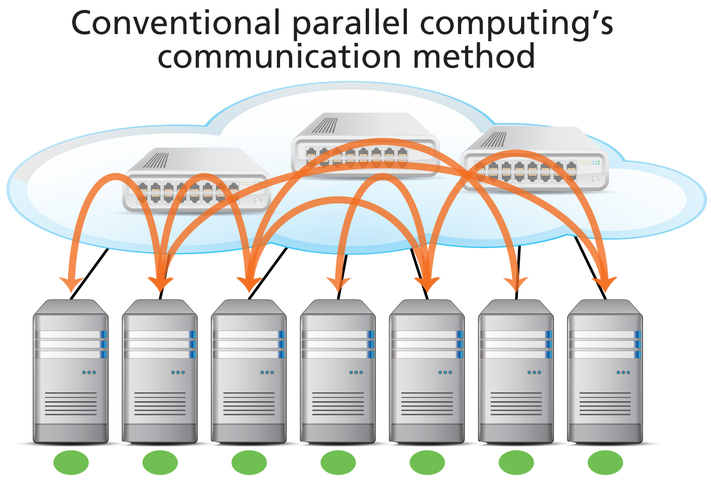
Software-Defined Networking (SDN) is a concept of network architecture that decouples conventional networking function into a programmable control plane (network controller) and a data plane (physical network).



Problem

However, collective communication of MPI often suffers from performance degradation. One of the main causes is that most of the MPI implementations are designed on the assumption that:

1. MPI programs are not able to acquire network topology information.
2. Network resources are statically allocated.

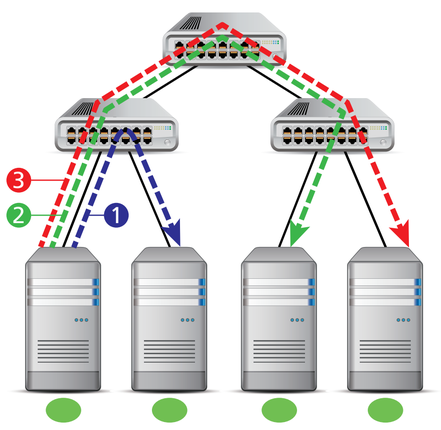


Research Goal

Integrate the *dynamic* controlling ability of Software-Defined Networking into MPI in order to optimize *collective communication* by overturning the assumption that network is a static resource. In the future, we'd like to implement a new MPI library which cuts down communication latency and traffic amount by programmatically controlling the underlying network.

Fast MPI_Bcast Design Enhanced with SDN

MPI_Bcast in Conventional Methods

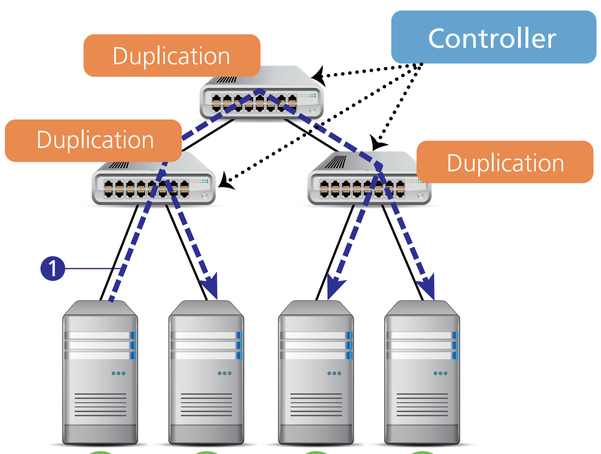


MPI_Bcast uses combination of multiple one-to-one communciations.
Intra- and Inter-node communications does not overlap with each others well.

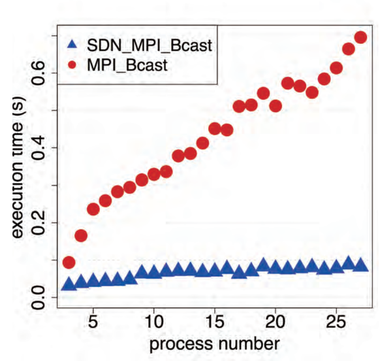
SDN MPI_Bcast Method

SDN_MPI_Bcast uses SDN-enabled switches as packet copier.

Stage 1 – Broadcast Data



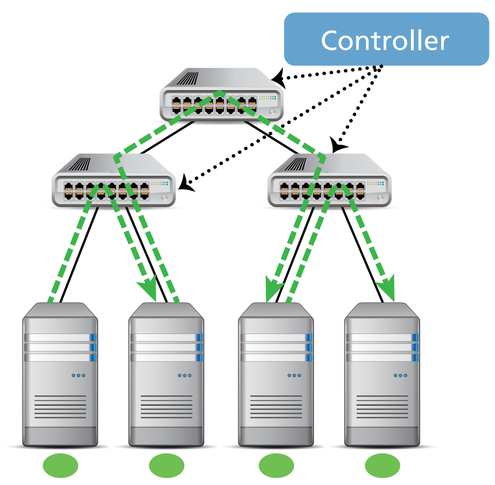
Execution time of SDN_MPI_Bcast



Stage 2 – Assure Reliability

MPI processors send data to "next" processor using one-to-one communication.

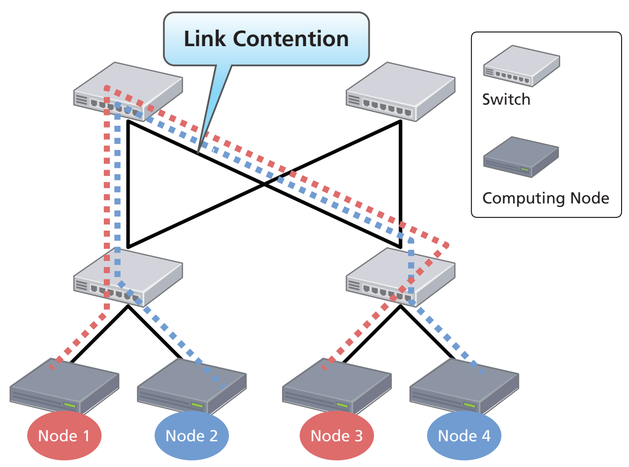
SDN Controller calculates "next" processor considering network topology and physical location of processes. Intra-node communications occur during Stage 2.



Efficient Bandwidth Utilization of Clusters

Link Contentions in Conventional Clusters

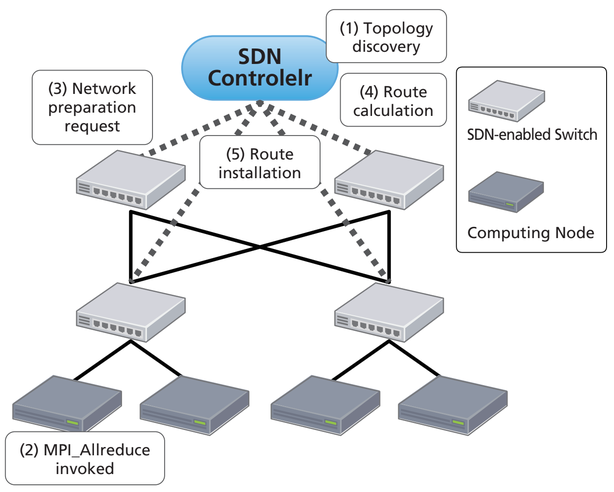
In conventional computer clusters without full bisection bandwidth interconnect, link congestion could happen due to the deviation of packet flow.

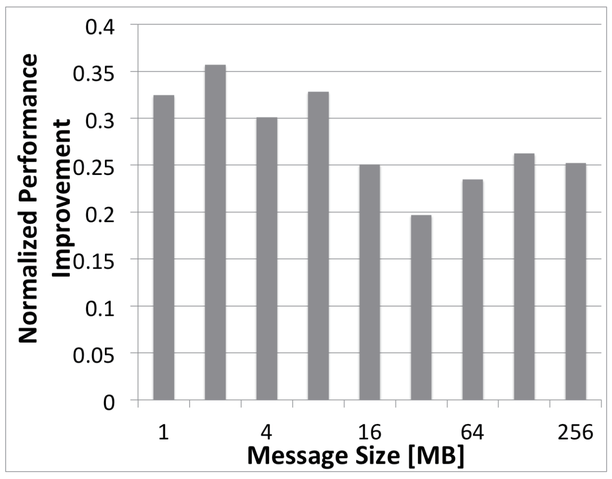


Cooperation of MPI Library and SDN Controller

We propose a new computer cluster architecture consisting of a customized MPI library, SDN controller and LLDP daemon. This system dynamically reconfigures the underlying network depending on the MPI communication request so that link contention will not happen.

An experiment conducted on a computer cluster with fat-tree interconnect revealed that our method increased the performance of MPI_Allreduce for 36 % at maximum.





[1] Khureltulga Dashdavaa, Susumu Date, Hiroaki Yamanaka, Eiji Kawai, Yasuhiro Watashiba, Kohei Ichikawa, Hirotake Abe and Shinji Shimojo. Architecture of a High-speed MPI_Bcast Leveraging Software-Defined Network. In UCHPC2013: The 6th Workshop on UnConventional High Performance Computing 2013, Archen, Germany, Aug. 2013.

[2] Keichi Takahashi, Dashdavaa Khureltulga, Yasuhiro Watashiba, Yoshiyuki Kido, Susumu Date, Shinji Shimojo, "Performance Evaluation of SDN-enhanced MPI_Allreduce on a Cluster System with Fat-tree Interconnect", The International Conference on High Performance Computing and Simulations (HPCS2014), Bologna, Italy, Jul. 2014.

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

Cybermedia Center
Osaka University

Toward efficient and flexible resources provisioning on SDN-enhaned Job Management System Framework

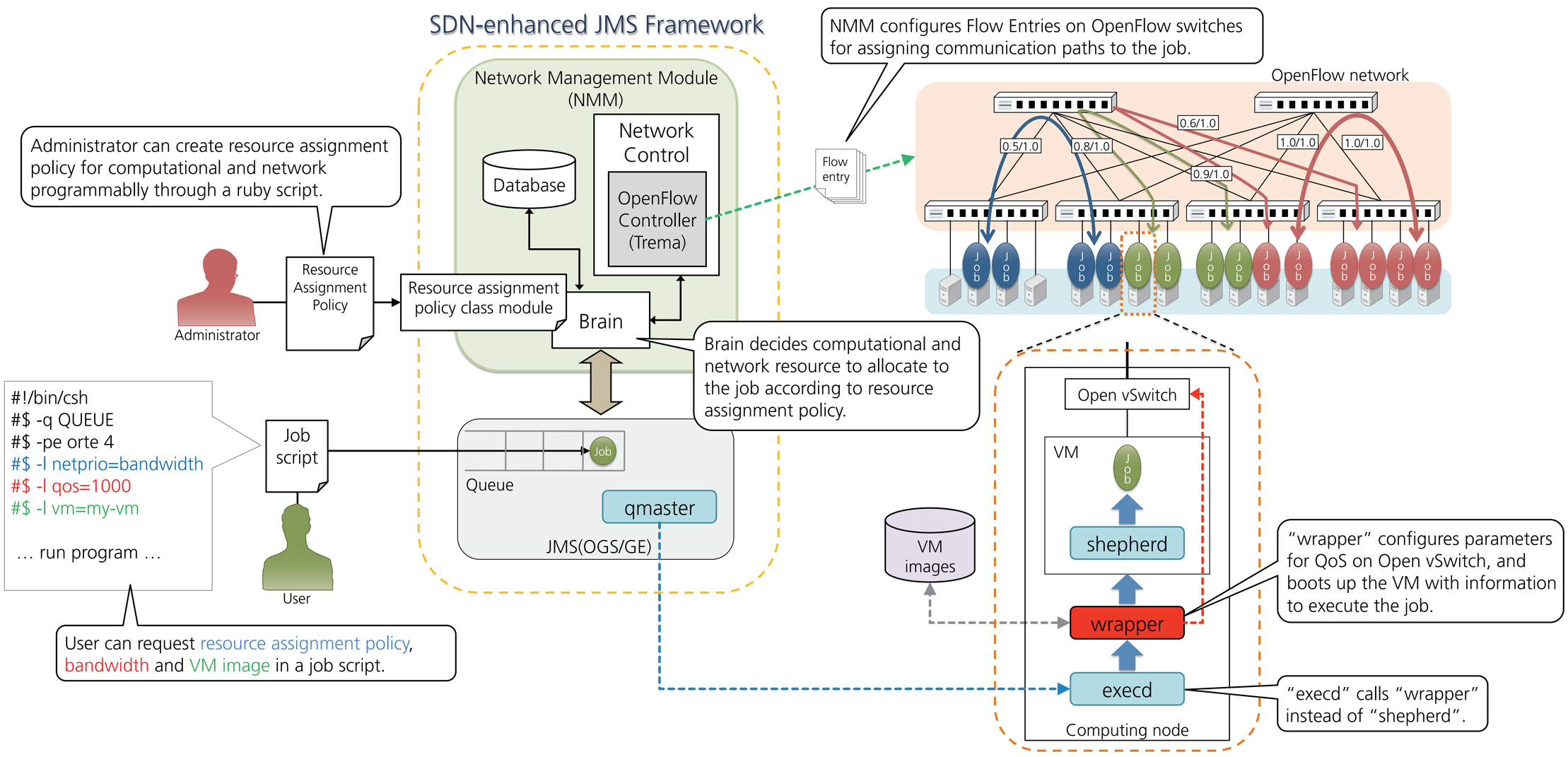
Cybermedia Center, Osaka University, Japan

Motivation and Objectives

Nowadays, usage patterns of users' computations on high-performance cluster system have been diversified for large-scale simulations and analyses in the various science fields. Since an HPC cluster system needs to accommodate multiple jobs concurrently, efficient and flexible resource management is essential for providing high performance computing capabilities for multiple users. However, most Job Management Systems (JMSs) available today, which are deployed on HPC cluster system for computational workload distribution and balancing purposes, determine resource allocation to each jobs only based on computational resources such as CPU and memory. In this research, we realize a novel JMS framework for handling various kinds of resources such as network resources and virtualized computational resources.

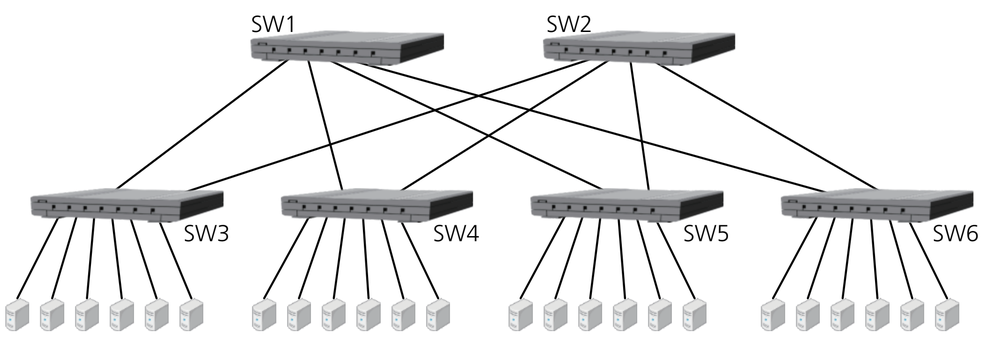
Proposed SDN-enhanced Job Management System Framework

We have been studying and developing a novel network-aware JMS integrated *Software-Defined Networking (SDN)* concept, which can dynamically control an entire network in a centralized manner, into a traditional JMS [1]. The mechanisms to manage the network resources are designed and implemented as *Network Management module (NMM)* leveraging OpenFlow, which is an implementation of the SDN concept. The SDN-enhanced JMS can allocate both computational and network resources to each job according to the resource usage on a cluster system and *Resource Assignment Policy* defined by administrator.
Moreover, we have also been developing a mechanism for deploying job's processes to virtual machines (VMs) on computing nodes, and guaranteeing available bandwidth on communication paths allocated to a job by using QoS functions of Open vSwitches (OVSs) connected with VMs.

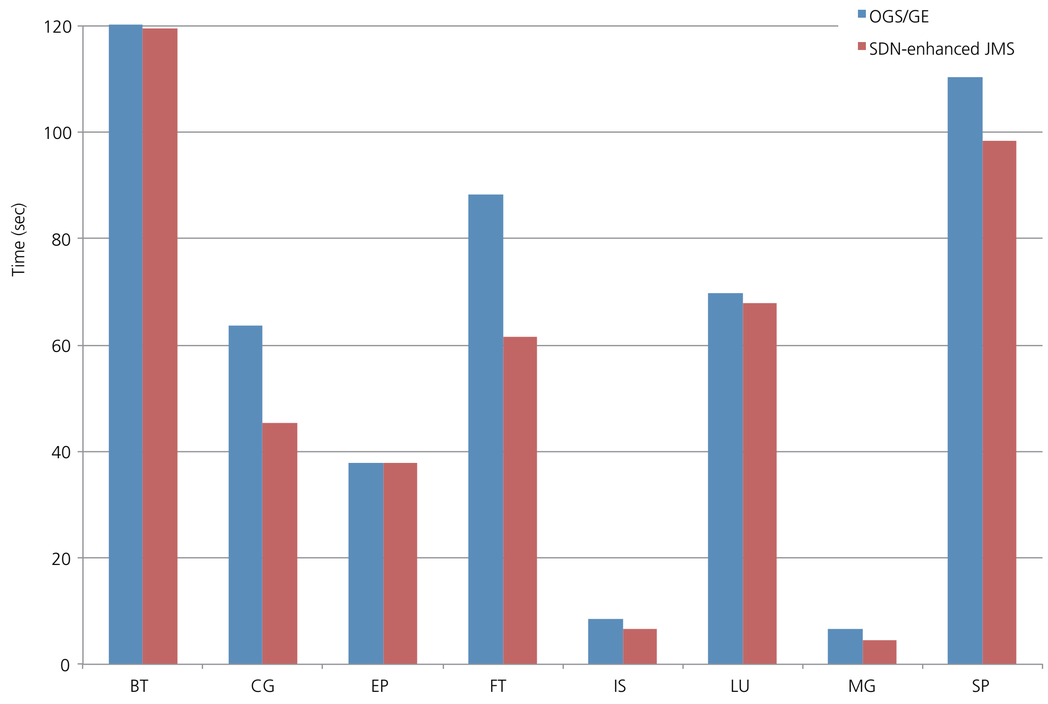


Evaluation

To assess the efficiency of explicitly allocating communication paths to each jobs by the SDN-enhanced JMS, we conducted an experiment to compare the average execution time of jobs on a fat-tree cluster system. Note that processes of jobs were allocated to real computational resources and a single computing node accommodated only one process. In the experiment, we submitted multiple jobs, each of which ran NAS Parallel Benchmarks with class B and generated 4 processes. As a result, our proposed SDN-enhanced JMS succeeded to reduce the average execution time of jobs.



2-tier fat-tree cluster system:
The number of computing nodes is 28 and every link is connected on a Gigabit Ethernet LAN.



Average execution time of NAS Parallel Benchmarks

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, Y. Kido, S. Date, H. Abe, K. Ichikawa, H. Yamanaka, E. Kawai, H. Takemura, "Prototyping and Evaluation of a Network-aware Job Management System on a Cluster System Leveraging OpenFlow", The 19th IEEE International Conference On Networks (ICON 2013), Dec. 2013.

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