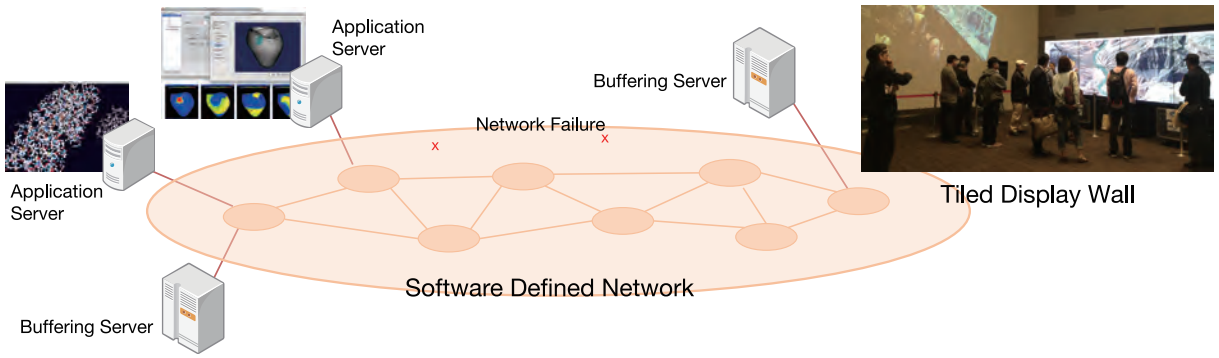
Tiled Display Wall with Network Failure Avoidance Mechanism using Software Defined Networking

Cybermedia Center, Osaka University, Japan

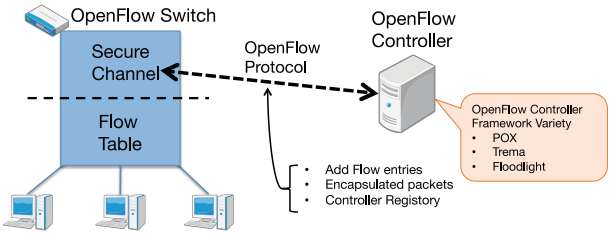


Summary

In this study, we seek a new design of tiled display wall middleware that are aware of network condition and tries to avoid instabilities.
We modified tiled display wall middleware to be able to detect network failures and packet buffering mechanism based on OpenFlow technology, one of the Software Defined Network implementation. With the modified version of SAGE, a visualization application can detect a network failure, and change network path according to the circumstances. And also the packet buffering mechanism can compensate dropping packets, that might lead to corruption of video frames.

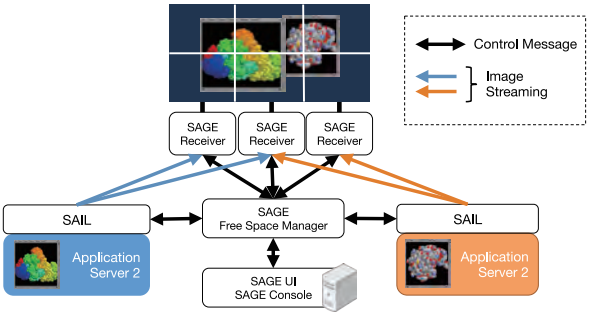
Key Technologies
SDN/OpenFlow

Software Defined Network (SDN) is a newly-emerged network concept that allows us to separate the network control plane from the data transfer plane. That can help network administrators to manage network resources in centralized manner. And also SDN enables us to realize flexible control of network resources, such as redundant pathway, packet buffering and so on.
OpenFlow is the most popular SDN implementation which is being standardized by the Open Networking Foundation [1]. The OpenFlow protocol, which is used to split the control plane from the data plane,
has been affected by SDN concept.



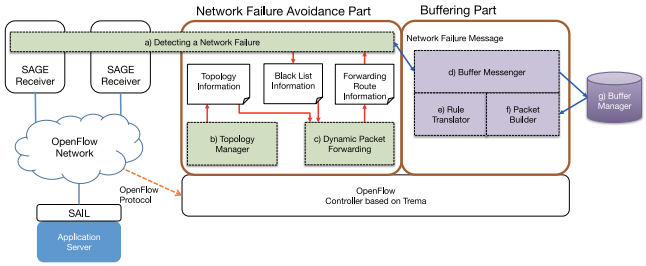
TDW/SAGE

SAGE (Scalable Adaptive Graphics Environment)[2] is a middleware designed to control tiled display walls. Notable features of SAGE include distributed rendering, display number scalability, and multiple viewing applications. We believe that these features are very important for spreading and encouraging e-Science movement.



System Overview

We implement detecting and avoidance a network failure with SAGE and Trema [3]. Trema is one of OpenFiow controller framework, which provides easy-to-use framework for developing OpenFlow controller in Ruby and C language[4].



Network Failure Avoidance Part

a) Detecting a Network Failure
If network flow is down, SAGE receiver stops the rendering image and adds failed network flow to the black list. Then, it sends a message to the Buffering Part.

b) Topology Manager
Topology Manager has been capturing the network topology.

c) Dynamic Packet Forwarding
If black list updates, this function rewrite flow entry as an un-failed network flow. Then, receiving a dropped packet from the Buffering Part and sending it to SAGE Receiver.

Buffering Part

d) Buffering Messanger
If network flow is down, SAGE receiver stops the rendering image and adds failed network flow to the black list. Then, it sends a message to the Buffering Part.

e) Rule Translator
It aims the function of packet filter for buffering. A user create filter rules which apply and filter a packet using TCP-dump.

f) Packet Builder
If network flow is down, Packet Builder exports the captured packets and recover failed packets.

g) Buffer Manager
Buffer Manger has buffer captures on a per-flow basis. Packet capture is continuous once a flow has been added.

[1]. The Open Flow Switch Specication, Version 1.1.0, http://www.openflow.org/documents/openflow-spec-v1.1.0.pdf, (2011).
[2]. Leigh, J., Johnson, A., Renambot, L., Peterka, T., Jeong, B., Sandin, D., Talandis, J., Jagodic, R., Nam, S., Hur, H. and Sun, Y.: Scalable Resolution Display Walls, Proc. of the IEEE, Vol. 101, Issue 1, pp.115-129, (2013).
[3]. Furuichi, T., Date, S., Yamanaka, H., Ichikawa, K., Abe, H., Takemura, H. and Kawai, E.: A Prototype of Network Failure Avoidance Functionality for SAGE using OpenFlow, Proc. of IEEE 36th International Conference on Computer Software and Applications Workshops, pp.88-93, (2012).
[4]. Trema, http://trema.github.io/trema/

Contact : Yoshiyuki Kido, kido@cmc.osaka-u.ac.jp

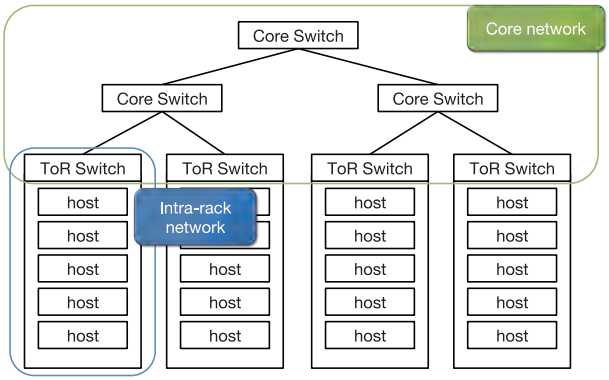
Tiled Display Wall with Network Failure Avoidance Mechanism using Software Defined Networking

Optical Path Scheduling Methods Considering Host Bandwidth in Data Center Networks

Cybermedia Center, Osaka University, Japan

Introduction

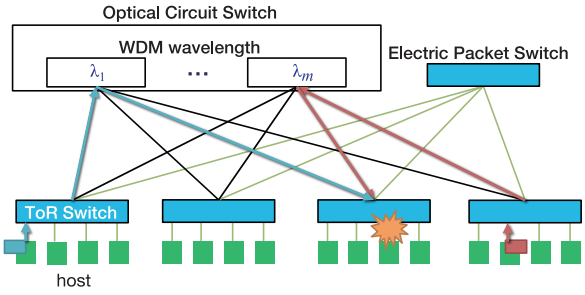
Data centers always need to process large amounts of data and to transfer it at high speeds. Data centers have a high-speed network and a large number of hosts and switches.



Current data center network architectures with electric packet switching consume large amounts of energy. Therefore, in recent years, all optical network technologies have been studied. By using wavelength division multiplexing (WDM), all-optical network is possible to transmit large amounts of data in low energy consumption.

Challenges

In all optical network, it is important to establish optical paths efficiency. We need the effective path establishment method which has a minimum and sufficient number of transmission times. Helios, a data center network architecture with optical circuit switching, does not consider collisions with connections to the same destination host. If one destination host has many connections, there is a possibility that collisions between connections occur and many packets are lost.

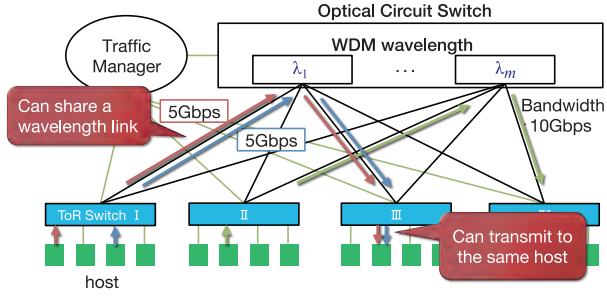


Our Target

We propose path establishment methods considering available bandwidths for each hosts, not only between racks, for data center network architectures with optical circuit switching.

Our Approach

Our proposed method can establishes paths as long as the total required bandwidth does not exceed the maximum bandwidth for all link between hosts.

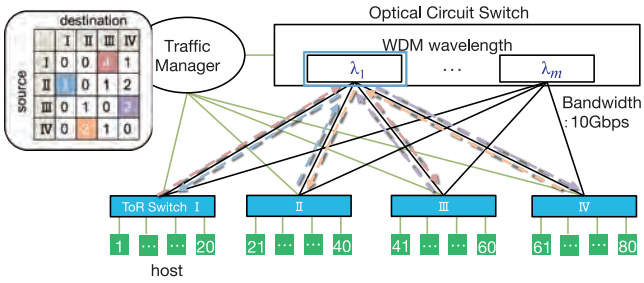


The path establishment method occupying wavelength

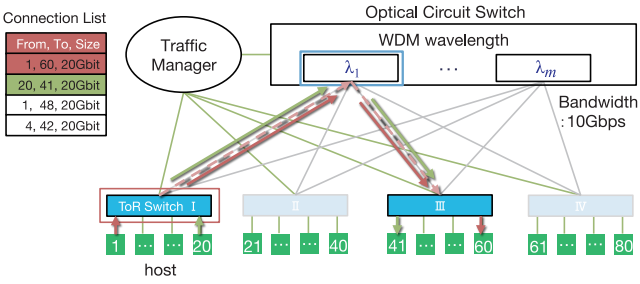
In this study, traffic is sent optically across source destination racks. The traffic manager calculates and establishes optical paths across racks from the number of connections which required path across each rack as well as Helios. In order to do this, the traffic manager has a traffic table that aggregates the total number of connections across all source-destination rack pairs from each ToR switches.

In the proposed method, the path is established through the two-step.

- First, the path between racks are established by using connection lists sent from Top of Rack switches. From connection table between racks, traffic manager establish the path between racks that can send the most connection.

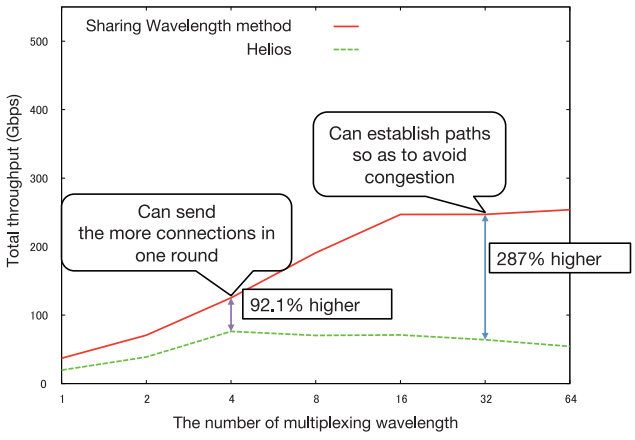


- Second, the traffic manager establishes the paths between hosts based on the path between racks. The traffic manager checks lists for all rack pairs with wavelength, all wavelengths. If the required bandwidth of a connection is less than the available bandwidth between hosts, the connection is established.



Evaluation

We evaluate these path establishment method to compare the proposed system with Helios through simulation. We use the value, the throughput which is calculated from the received data size.



If the number of multiplexed wavelength is small, the path establishment method sharing wavelength has higher throughput than the path establishment method of Helios because this method can send more connections in one round. On the other hand, if the number of multiplexed wavelength is large, the path establishment method wavelength has higher throughput than the Helios because this method can establish paths so as to avoid congestion.

Contact: Yukihiro Kunishige, kunishige.yukihiro@ais.cmc.osaka-u.ac.jp
Ken-ichi Baba, baba@cmc.osaka-u.ac.jp

Optical Path Scheduling Methods Considering Host Bandwidth in Data Center Networks

Towards SDN-enhanced Message Passing Interface

Cybermedia Center, Osaka University, Japan

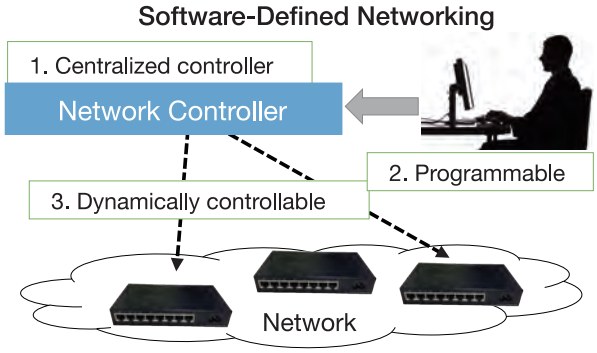
Towards SDN-enhanced Message Passing Interface

Background

Nowadays, parallel computing technique is an inseparable technology for High Performance Computing (HPC). Message Passing Interface (MPI) has been a *de facto* standard programming model in parallel computing for around two decades. MPI offers two types of communication; one is *one-to-one communication*, which is for low-level description of communication pattern, and the other one is *collective communication*, which is for high-level and human-friendly description.

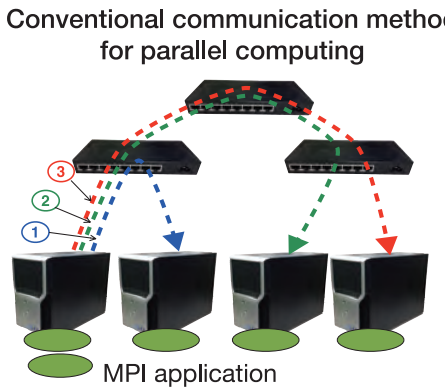
Software-Defined Networking

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 devices from a centralized controller.



Problem

However, collective communication of MPI often suffers from performance degradation when it is used on HPC environment based on commodity hardware, such as Gigabit Ethernet. One of the main cause is that most of the MPI implementations are not optimized for such hardware. For example, when a process wants to broadcast some data to other processes (MPI_Bcast : a process sends data to all other processes), multiple times of transmission will happen.



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. Ultimately, we'd like to implement a new MPI library, which cuts down communication latency and traffic amount by programmatically controlling the underlying network.

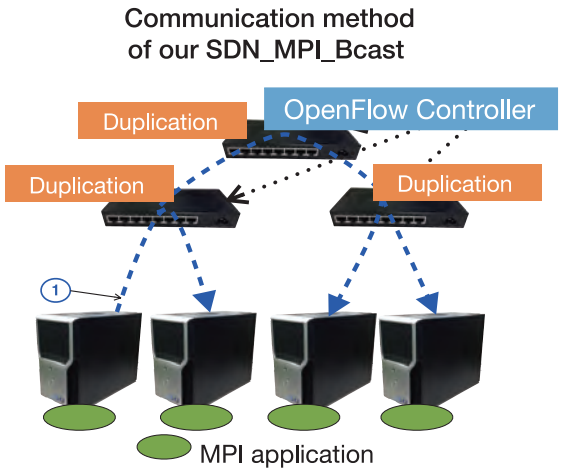
Progress Report

MPI_Bcast [1]

Design:

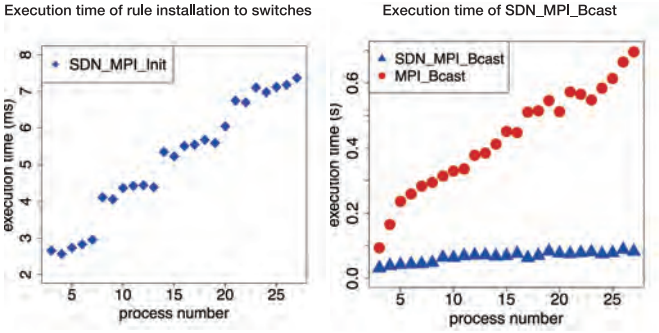
1. OpenFlow controller obtains MPI-cluster network's topology with LLDP.
2. IP addresses of the node which will broadcast and all other receiving nodes are sent to the controller.
3. Controller builds a broadcast tree with the information gathered in step 1 and 2. An algorithm based on Floyd-Warshall's method is used in this process.
4. A Set of packet duplication rules are generated, which instructs to flow packets from a source process to other processes on the broadcast tree in step 3. These rules are deployed to OpenFlow compatible network switches.
5. Broadcast node sends packets to its network.

Every packet being broadcasted includes a specific ID, which are unique to the combination of source node and destination nodes. Packet copy rules will only match if the incoming packet includes the corresponding ID



Current Result:

Reduced execution time of MPI_Bcast.



MPI_Reduce

Design:

1. Controller analyzes network topology with LLDP and traffic usage with OpenFlow's flow statistics. Network link usage information is used to select one route from multiple redundant routes in step 3.
2. Size of the array being reduced is examined, so that an optimal reducing algorithm can be selected from flat-tree, binary-tree, etc.
3. Using these data, an optimized "Reduction tree" is built.
4. Packet forwarding rules are generated and installed in the same way as MPI_Bcast.

[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, Aachen, Germany, August 27, 2013.

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

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

PC cluster

Location

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.

Visualization Service at Cybermedia Center

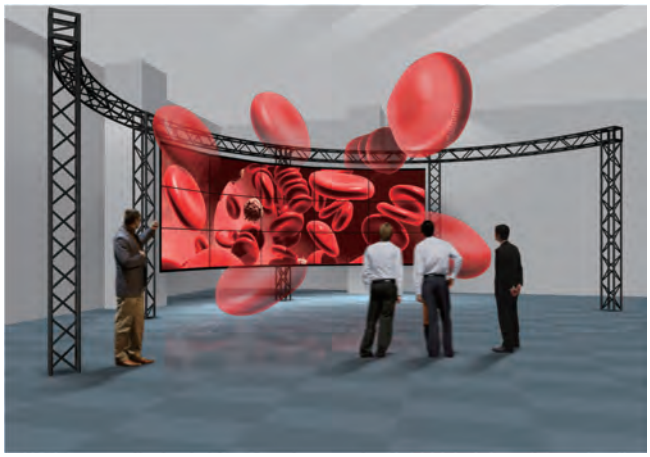
Cybermedia Center, Osaka University, Japan

High Resolution Stereo Visualization Systems

Cybermedia Center at Osaka University will introduce two types of high-resolution stereo visualization systems in 1Q/2014, one on Toyonaka campus and the other at Umekita office near JR Osaka station. These systems will be 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 will also be available including AVS Express/MPE VR, IDL, Gsharp, CAVELib, EasyVR MH 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



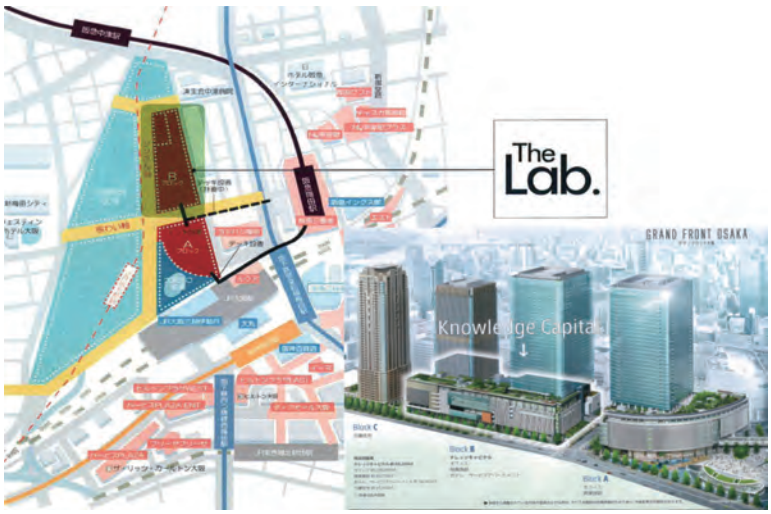
15-screen Cylindrical Stereo Visualization System @ Umekita Office

* WXGA (1366x768) 46-inch LCD x 15 (approx. 16 million pixels)
* Image Processing PC with NVIDIA K5000 x 5
* HD Video Conferencing System x 1
* Motion Capturing System x 1

Visualization Services at Cybermedia Center

Following visualization services will become available in 1Q/2014 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 will be 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 will be offered.
2) **Consultation:** Technical consultations will be offered through which users can learn best suitable visualization techniques for their problems from a diversity of solutions which will dramatically reduce time and effort required to produce satisfactory visualization results.
3) **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.



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

Architecture of Job Management System Framework Leveraging Software Defined Networking

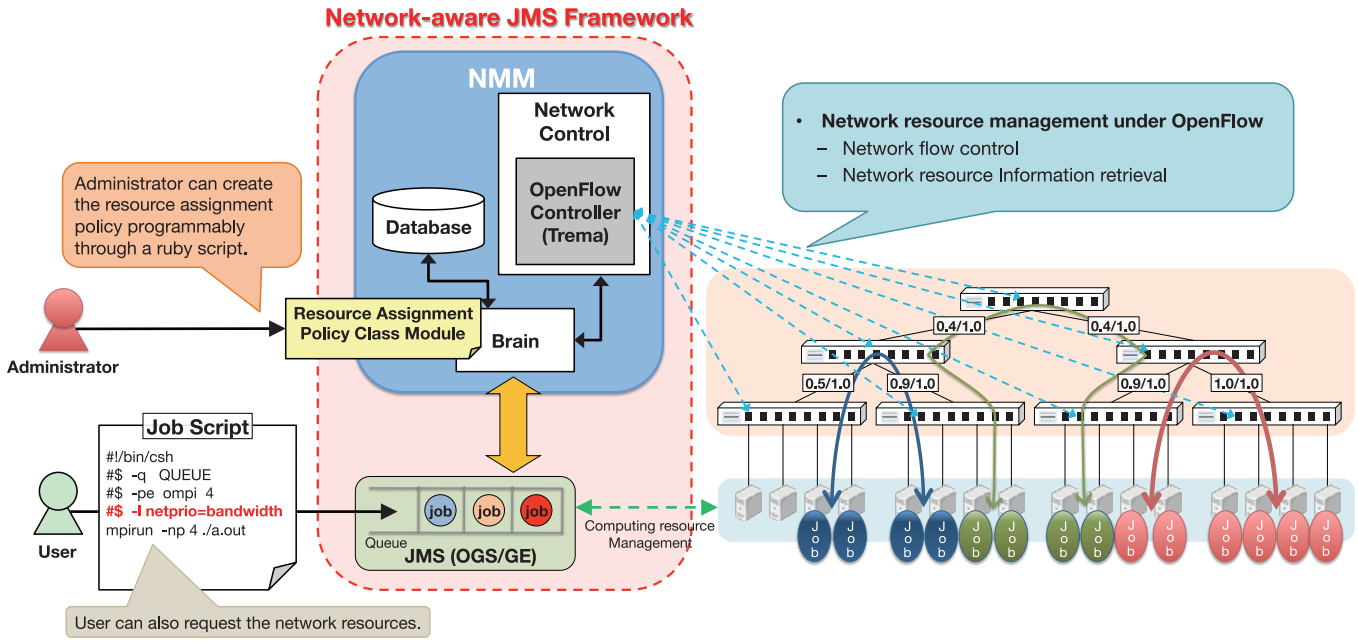
Cybermedia Center, Osaka University, Japan

Motivation and Objectives

Network communication performance in high-performance computing environment such as cluster systems plays more important role due to a fact that parallel computation executes on the distributed multiple computing hosts. Since the resources of such computing environment are shared by multiple user jobs, efficient allocation of both network and computing resources to each job is necessary. However, most Job Management Systems (JMSs) available today, which determine the resource allocation to jobs on the computing environment, are not designed to consider status information on network resources. Therefore, we aim to realize network-aware JMS framework that can design the rule of resource allocation for both network and computing resources.

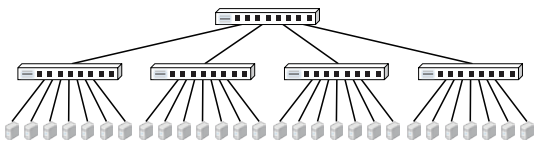
Proposed Network-aware Job Management System Framework

In order to realize a novel network-aware JMS Framework, we take advantage of ***Software Defined Networking (SDN)*** concept, which can dynamically administer an entire network in a centralized manner. 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. A function to create a rule of resource allocation is provided by *Resource Assignment Policy Class Module.*



Evaluation

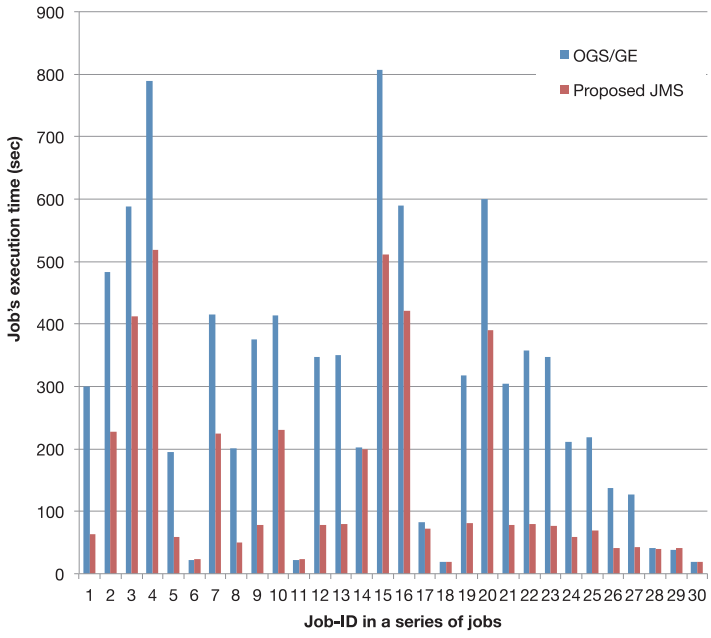
We conducted a measurement experiment to compare job's execution time in each jobs on a cluster system. In this experiment, we submitted a series of network-intensive jobs with the number of processes such as the following table. The resource assignment policy selected a set of computing hosts in which the total number of hops between each computing hosts is smallest. In the experimental result, our proposed network-aware JMS Framework succeeded in reducing the job's execution time by 44.8 percent on the average.



Cluster system configuration:
Each link are connected on a Gigabit Ethernet LAN.

The number of processes in each job.

Job-ID	1	2	3	4	5	6	7	8	9	10
Process	6	10	12	14	6	2	10	4	8	10
Job-ID	11	12	13	14	15	16	17	18	19	20
Process	2	8	8	4	14	12	4	2	8	12
Job-ID	21	22	23	24	25	26	27	28	29	30
Process	8	8	8	6	6	4	4	4	4	2



Contact : Yasuhiro Watashiba, watashiba-y@cmc.osaka-u.ac.jp