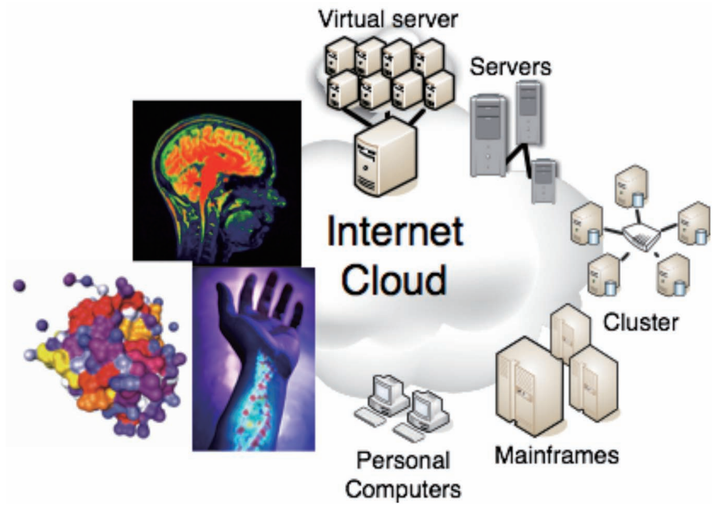
Throughput Evaluation for Improving Predictability over the Internet

Cybermedia Center, Osaka University, Japan

Motivation

Applications in Grid environment are required to share distributed large blocks of data among distributed multiple sites.
Network throughput for data transfer affects total processing time as well as the task processing.

The predicted network throughput would be a useful parameter on scheduling tasks to improve the total processing performance.

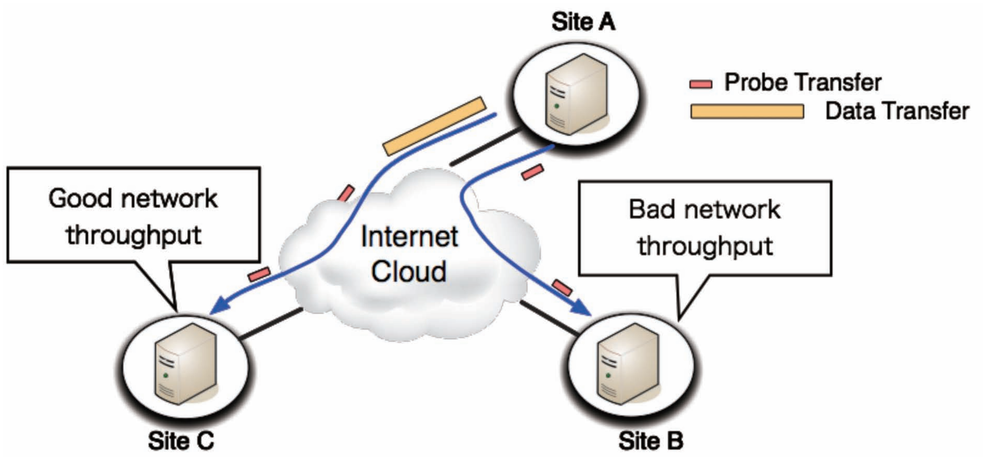


Purpose of our research

Improving precision of the throughput prediction method called “Network Weather Service” [1].
Adapting the prediction method to virtualized hosting environment, which shows anomalous behavior more frequently than physical nodes.

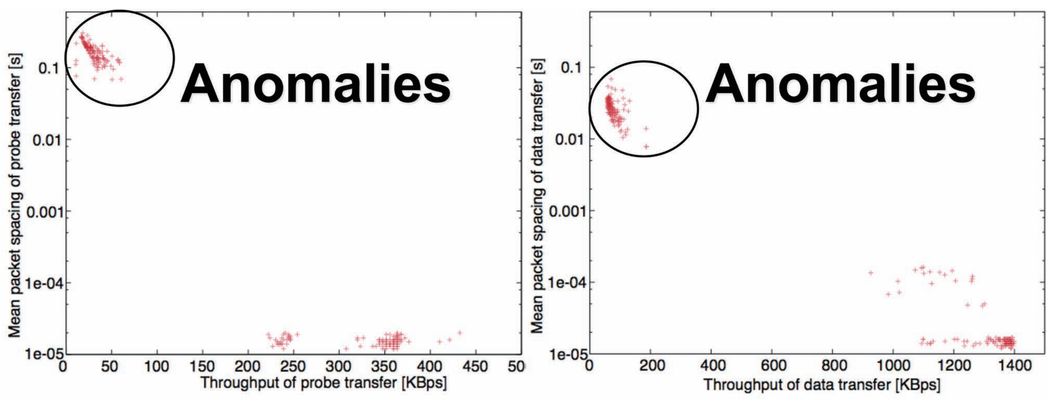
Connection pair

A network throughput prediction has been a challenging issue due to the dynamics of network traffic and no guarantee for bandwidth reservation.
Connection pair uses a small size of probe transfer to predict the throughput of a large size of data transfer.
In previous work [1], the restricted sets of pairs on probe and data size were examined on limited network environment.



Traffic anomalies

The evaluation results were affected by oversize packet spacings, which are caused by CPU scheduling latency.
The packet spacing which is larger than the TCP transmission period involves packet transmissions, which results in severe throughput[2].



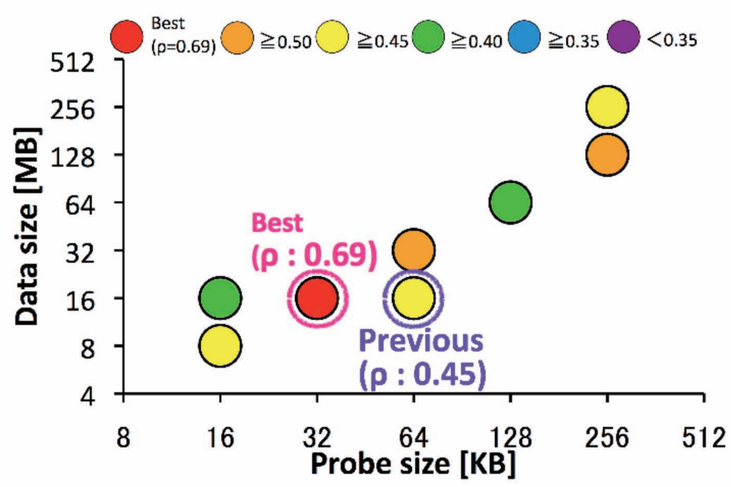
< Mean packet spacing and throughput on the best condition >

Experimental settings

We used PlanetLab nodes, equipped with a virtualization mechanism called V-Server (http://www.linux-vserver.org).
Various sizes for both probe and data transfer are used.
Correlation between both probes is evaluated by Spearman’ s rank correlation coefficient, one of non-parametric metrics.

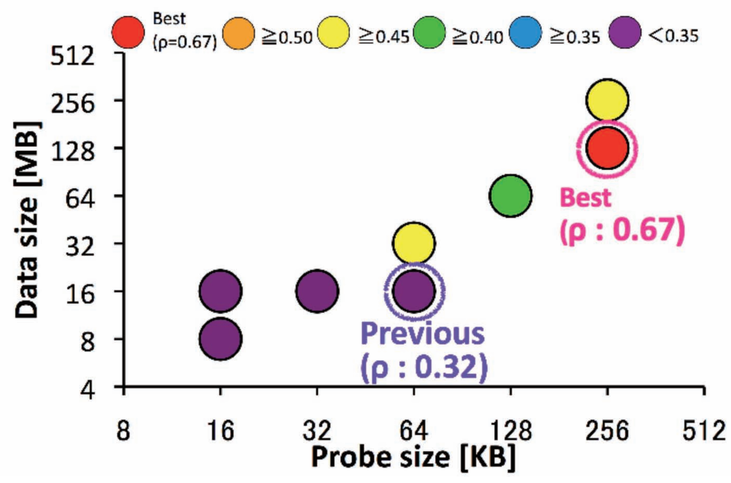
Original result

Smaller-size probes had better conditions than larger-size probes.



Results without anomalous cases

We re-evaluate the results without the anomalies, and a larger-size probe is required for improving predictability.
If throughput is decreased by the anomalies, we should carefully review measurement results.



Conclusion and future work

Anomalies from virtualized hosting environment have great impacts on the prediction results.
We re-evaluate the results without the anomalies, and found that larger-size probe is required for improving predictability.
Our future works are to measure throughput with various probe size and to devise an anomaly estimation method.

References

[1] M. Swany and R. Wolski, Multivariate Resource Performance Forecasting in the Network Weather Service, in Proceedings of IEEE/ACM Conference on High-Performance Computing and Networking, pp 1-10, November 2002,.
[2] C. Lee, H. Abe, T. Hirotsu, and kyoji Umemura, “Analysis of anomalies on a virtualized network testbed,” in Proceedings of the 10th IEEE International Conference on Computer and Information Technology, pp. 297–304, June 2010.

Chunghan Lee：Toyohashi University of Tech.Dept. of Electronic and Information Eng.
Hirotake Abe：Osaka University Cybermedia Center
Toshio Hirotsu ：Hosei University Faculty of Computer and Information Sciences
Kyoji Umemura ：Toyohashi University of Tech.Dept. of Computer Science and Eng.

Contact: Hirotake Abe
E-mail: habe@cmc.osaka-u.ac.jp

About Us: Cybermedia Center, Osaka University

Cybermedia Center
Osaka University

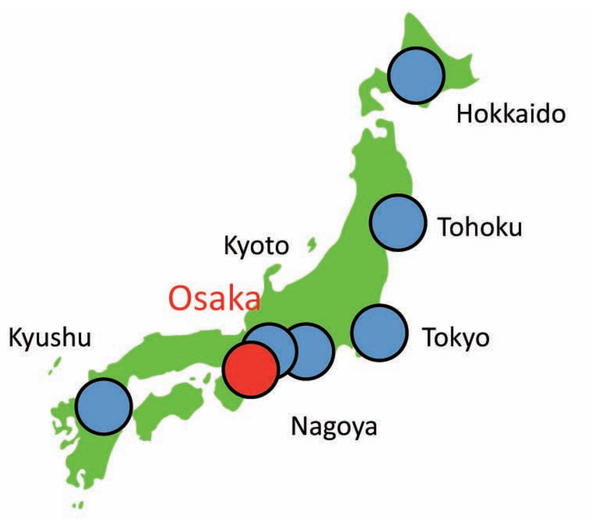
Cybermedia Center, Osaka University, Japan

In April 1969, the Computation Center (CC) of Osaka University was established as a laboratory that provides researchers of universities and other institutes computation and information processing services indispensable to their academic researches and education. CC of Osaka University was a part of a grand-scale endeavor by the Japanese government to found seven such supercomputer centers across the nation after accepting the Japan Science Council's suggestion to facilitate the collaborative use of information technology amongst researchers. The other supercomputer centers are located in Hokkaido University, Tohoku University, Tokyo University, Nagoya University, Kyoto University and Kyushu University.

Subsequently, CC of Osaka University and the supercomputer centers of Tokyo University and Kyushu University were reorganized into the Information Infrastructure Center for the collaborative use of information technology for researchers in Japan and to serve additional functions, including conducting practical researches and reinforcing information infrastructure, all of which are aimed at disseminating information and computing technology.

In April of 2000, Osaka University expanded and reorganized CC to form its branch of the Information Infrastructure Center, which was named the Cybermedia Center (CMC). In the expansion, the Education Center for Information Processing and a part of the university library were merged into CMC. While CC continues to provide computers for advanced scientific techniques and media services, the Education Center for Information Processing contributes by promoting education in information processing and the university library by providing digital contents.

Information Infrastructure Center



HPC System

The aim of CMC is to achieve remarkable evolution in information and computing infrastructure by complementarily and systematically integrating functions of computing technology-related organizations, as well as to provide an advanced infrastructure for the accumulation and the dissemination of digital contents and for the efficiency of the high level utilization.

CMC provides a powerful high performance computing environment for university researchers across Japan. It plays the role of the nation's hub in teaching and diffusing advanced information technology. In addition, the center assumes the responsibility of facilitating campus IT infrastructure and promoting its effective use. CMC also provides facilities for advanced education to Osaka University students. It operates an Information Education system and Computer Assisted Language Learning (CALL) system, connected by the Osaka Daigaku Information Network System (ODINS). The center offers a consistent information education curriculum, covering areas from basic usage of e-mail communication and the Internet to advanced computing technologies, such as programming, as well as provides foreign language and culture education support on various levels through a comprehensive application of multimedia techniques.

As a long-term goal, CMC supports educational and research activities by making comprehensive use of computer-related technology. Specifically, CMC develops various electronic and multimedia functions to improve the efficiency of educational activities. For research activities, CMC provides facilities to improve their scalability.

In next-generation computer applications, it is necessary to effectively "digitalize" ideas of researchers, meaning that CMC will need to deal with every stage of the research process, including information input, retrieval and collection, reports and discussion, analysis and modeling and finally visualization.

Research Divisions

Informedia Education Division develops the advanced environment for information processing education, offers educational programs on information processing and information ethics, and also conducts educational research, including faculty development programs for teaching staffs in information processing.

Multimedia Language Education Division division develops an environment for language education using multimedia, provides assistance in the development of multimedia-based language education materials, such as internationalized education using networks and foreign language programs as common subjects in Osaka University.

Large-Scale Computational Science Division supports the operation of CMC's supercomputer system, disseminates technology for visual presentation of computational results, and facilitates the advanced utilization technology of large-scale computing systems. This also offers educational programs and studies on computing science and related subjects.

Computer Assisted Science Division supports the operation of general-purpose computer systems, makes faculty developments to improve efficient computer applications for setting up and solving scientific problems, and it also offers educational programs and does research on subjects related to learning process for setting up and solving scientific problems.

Cybercommunity Division supports the planning and operation of SCS-based distance learning, plans and operates distance training in the field of advanced technology, and studies on the operation and promotion of cybercommunity plans.

Advanced Network Environment Division supports the operation and utilization of ODINS (Osaka Daigaku Information Network System), to introduce new network technologies such as high-speed networks and mobile computing environments, to facilitate the utilization technologies of large-scale wide-area computer networks, and to carry out research on network-related education.

Applied Information Systems Division develops and provides education on utilization technology for large-scale information systems, to digitalize libraries, to support the management of various databases, to implement education on information systems and multimedia systems, and to undertake education on information explorer.

Communication Network Analysis, NTT DOCOMO Collaborative Research Division has been founded this year.

Supercomputing Contest for High School Students (co-hosted by Tokyo Institute of Technology)

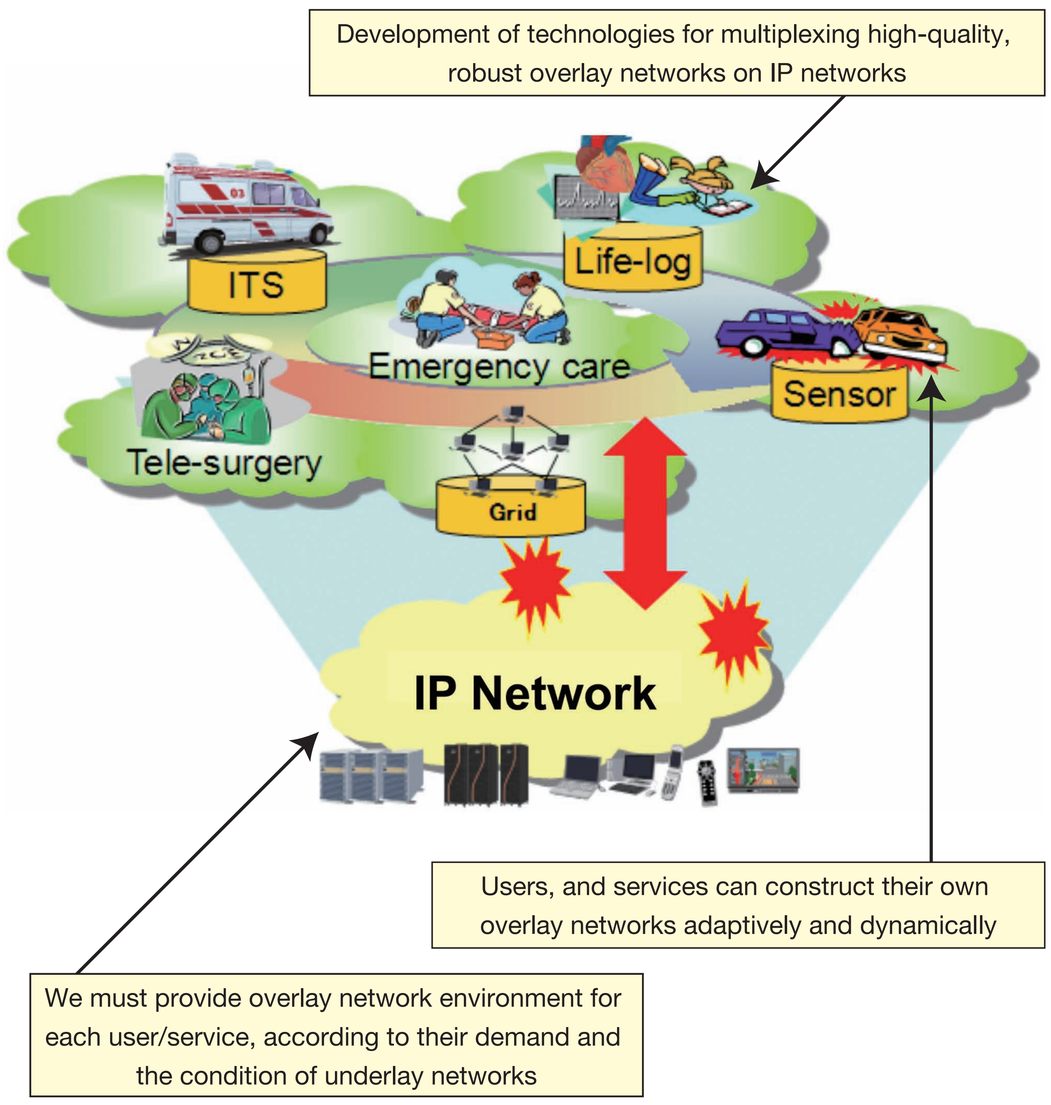
Management of overlay network performance: End-to-end network measurement strategy and quick failure recovery

Cybermedia Center
Osaka University

Cybermedia Center, Osaka University, Japan

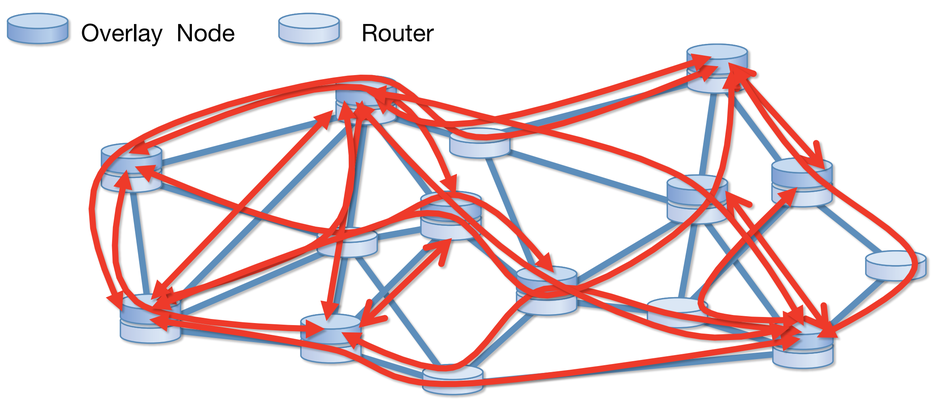
Joint works of Osaka University and NEC Corporation, Japan, supported in part by the National Institute of Information and Communications Technology (NICT) of Japan

Overlay Networks

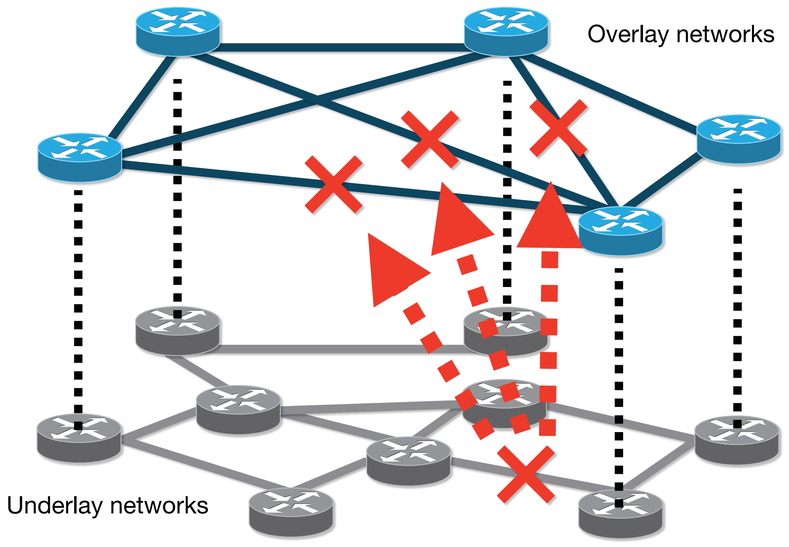


Management of overlay networks

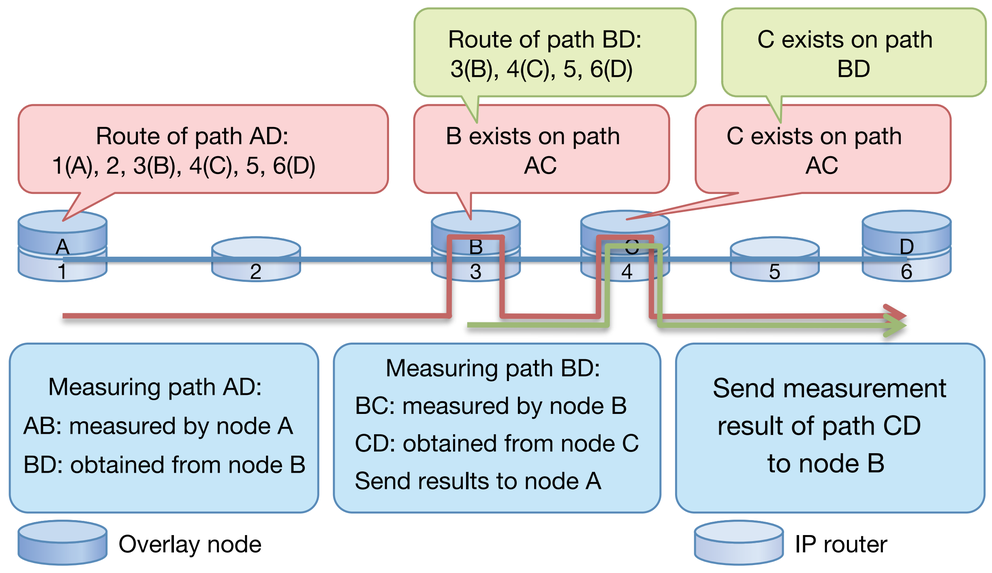
Network measurements: Simple full-mesh measurement has the O (N2) overhead. So, we need simple, lightweight, and scalable measurement method.



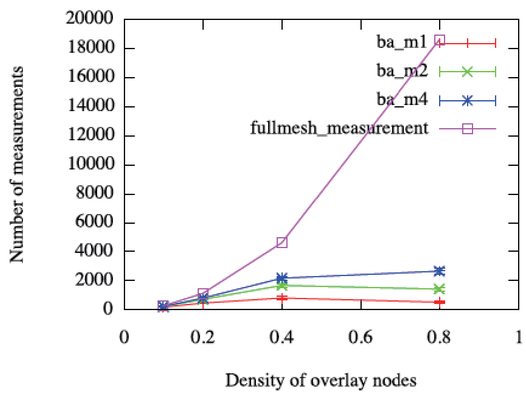
Failure recovery: Since overlay networks share underlay network equipments, a single network failure would bring multiple, simultaneous failures in overlay networks. So, we need failure recovery method from simultaneous failures in overlay networks.



Measurement strategies for overlay networks

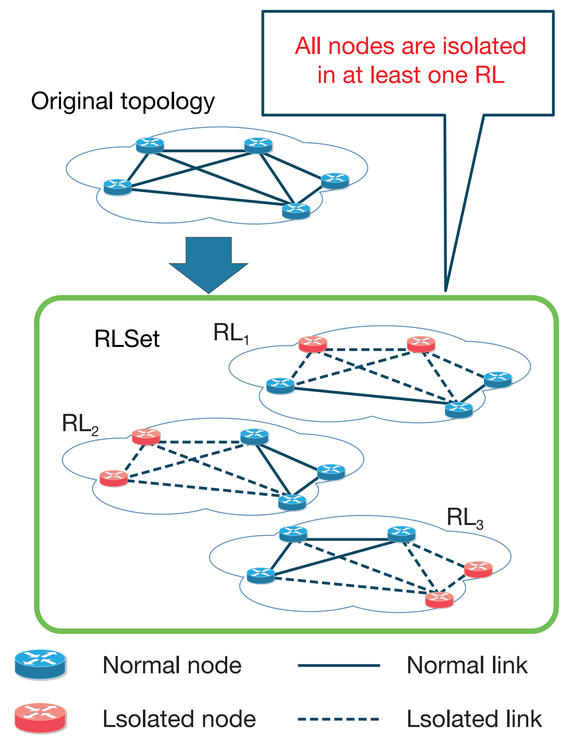


-Each overlay node conduct traceroute commands to other overlay nodes
-Intermediate overlay nodes capture them and record src/dst nodes
-All overlay nodes understand path overlapping status
-Measurement of longer path will be omitted and measurement result is estimated from results of shorter paths
-- Delay: D=d1+d2, bandwidth: B=min(b1, b2)
-- Packet loss ratio: P=1-(1-p1)(1-p2)



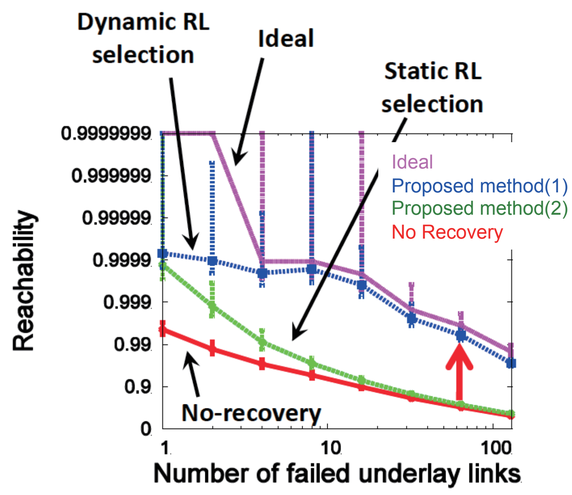
We can decrease the number of required measurement paths to 1/30 – 1/50, regardless of the underlay network topologies and the number of overlay nodes.

Proactive recovery method for overlay networks



-We construct multiple topologies for failure recovery from the original topology
-Each topology has isolated nodes. When a node failure occurs, we utilize the topology which isolates the failed nodes
-Multiple simultaneous node failures can be recovered, when such nodes isolated in one topology

We can improve reachability of the overlay network from 69% to 99%, when 5% links in the underlay network fail simultaneously.



Contact: Go Hasegawa
E-mail: hasegawa@cmc.osaka-u.ac.jp

OSAKA UNIVERSITY

A Handheld User Interface for GPU-Accelerated Large-scale Volume Visualization

Cybermedia Center
Osaka University

Cybermedia Center, Osaka University, Japan

Figure 1: Prototype System

Introduction

The ever-increasing sizes of volumetric data produced from a variety of scientific studies post a formidable challenge for the subsequent real-time large-scale volume visualization and analysis. While steady advances in graphics hardware enable faster rendering, real-time rendering of large-scale volume data is still a tough problem. Unlike polygonal data, volume data must take care of transparency of internal values within the volume that thus requires time-consuming voxel sorting for conventional volume rendering algorithms.

Such large-scale volume visualization must also take into account interactive transfer function control for data mining and scientific discovery. A high-resolution large screen is often used for displaying large-scale volume data, in which conventional input devices such as a mouse and a keyboard is not a practical solution.

In this poster, presented is an ongoing work on an interactive large-scale volume visualization system using a handheld device and a GPU-accelerated particle-based volume rendering algorithm (see Figure 1).

Particle-based Volume Rendering

To treat large-scale volume data, a particle-based volume rendering algorithm is employed in our visualization system, which has been proposed by Kyoto University. In the algorithm, a set of tiny opaque particles are generated from a given 3D scalar field. The final image is then generated by projecting these particles onto an image plane. A semi-transparency effect is realized by sub-pixel processing and averaging and thus the necessity of voxel sorting is eliminated. This algorithm is further accelerated by GPU programming.

Interactive Transfer Function Control

For assisting data mining and scientific discovery, a handheld device is provided to control the transfer function in detail in real-time. Our visualization system consists of a server machine which renders volume data and a client handheld device (i.e. iPad) that are connected to each other via WiFi or Bluetooth (see Figure 2). The hardware specification in our system is summarized in Table 1.

The GPU-based rendering algorithm has been carefully modified to accept user input to be able to control a variety of visualization parameters with negligible perform degradation.

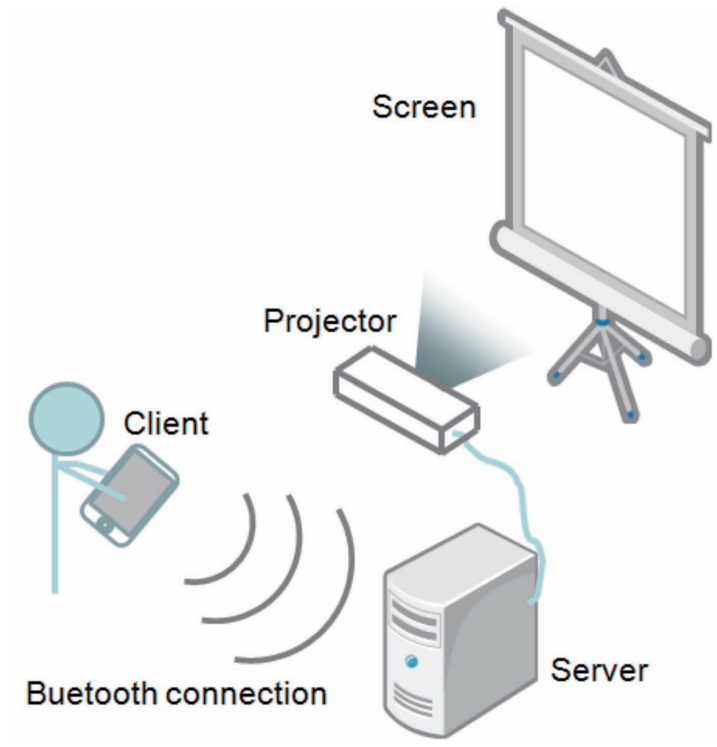


Figure 2: System configuration

Table 1: Specification

Server CPU：AMD Opteron 2350 2.0GHz RAM：32GB GPU:nVIDIA Quadro FX 4600 OS:Ubuntu 8.10 (64bit)	Client Multi-touch Display: 9.7 in Resolution: 768 x 1024

A set of graphical user interfaces (GUI) are under development (see Figure 3) for the handheld device. A user will be able to control the visualization parameters interactively. Example interactions implemented are:

- Color curve editing. Mapping between a scalar value of voxel data and its rendered color can be changed simply by deforming the color curves.
- Particle filtering. Particles with specified values can be hidden or highlighted.

Any change in visualization parameters will be immediately reflected to the rendering result.

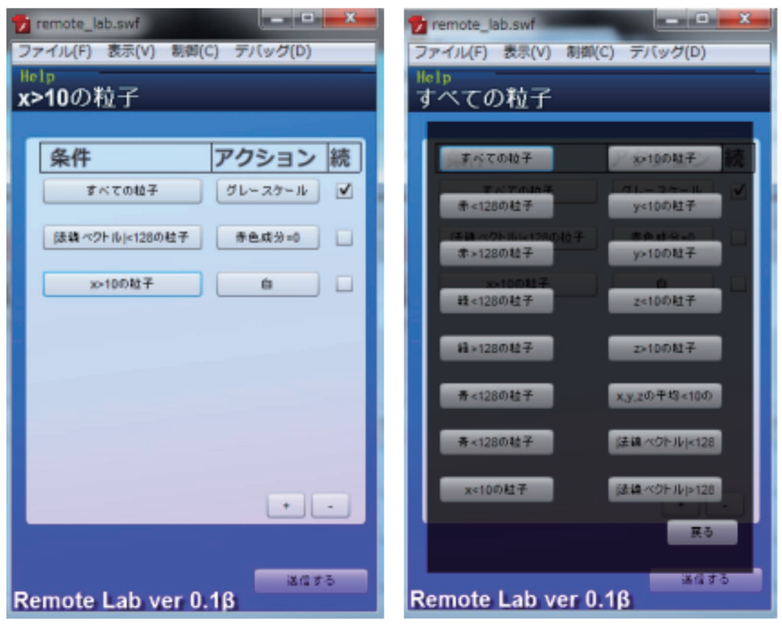


Figure 3: User Interface

Conclusion

An interactive visualization system for large-scale volume data has been proposed. Large-scale volume visualization has been realized thanks to the GPU-accelerated particle-based volume rendering. Interactive transfer function control has been realized by a dedicated graphical user interface on a handheld device.

Kentaro Oita : Graduate School of Information Science and Technology, Osaka University, Japan
Kiyoshi Kiyokawa : Cybermedia Center, Osaka University, Japan
Haruo Takemura : Cybermedia Center, Osaka University, Japan

OSAKA UNIVERSITY

Contact: Haruo Takemura and Kiyoshi Kiyokawa
E-mail: kiyo@ime.cmc.osaka-u.ac.jp

A Handheld User Interface for GPU-Accelerated Large-scale Volume Visualization

A Virtual Cluster over Multiple Physical Clusters Using P2P Overlay Network

Cybermedia Center, Osaka University, Japan

Introduction

Recently, virtual cluster technology has attracted attention in scientific research field. A virtual cluster system allows scientists to build their private computational environment which is customized as they like on one physical cluster.

However a virtual cluster over multiple physical clusters is still difficult to realize. This is because each physical cluster is commonly isolated by Firewall or NAT and so VMs deployed on different clusters cannot connect directly with each other. A virtual cluster over multiple physical clusters needs to virtualize network to hide disconnect of network.

We develop such a cluster over multi physical cluster leverage cluster install toolkit software with P2P overlay network software.

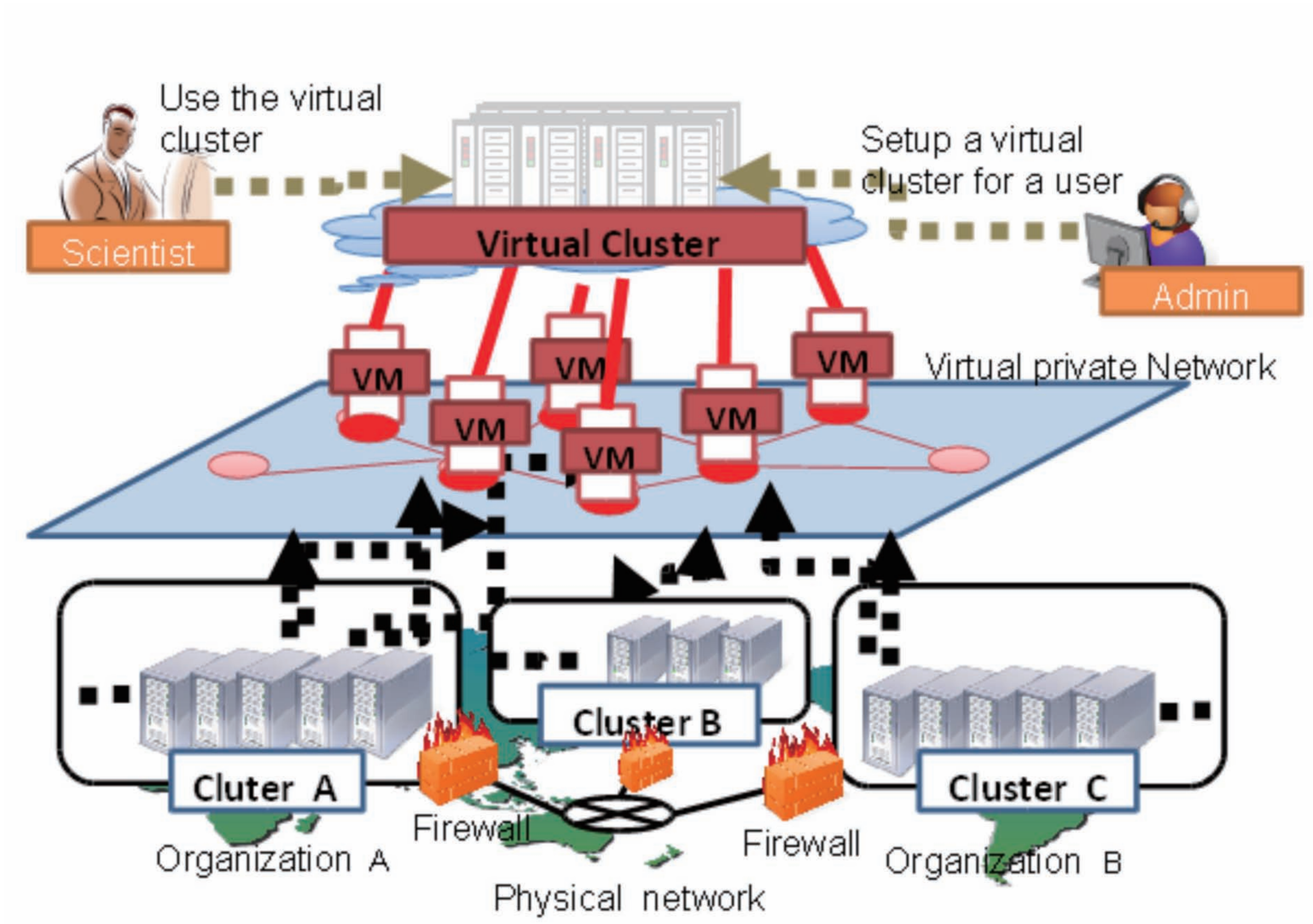


Figure 1 : Concept of our virtual culster

Our Virtual Cluster

Utilizes Two key technologies

N2N : A P2P software which establishes an encrypted layer-two virtual private network based on a P2P protocol implemented by ntop.org.

Rocks : A cluster installing toolkit including virtual cluster implemented by UCSD.

The process of developing a virtual cluster

Our virtual cluster is set up on multiple Rocks physical clusters. N2N virtual private network are established between VMs which are deployed by Rocks. The setup processes are performed by a admin of one Rocks physical cluster. Rocks clusters which allow admins to set up a virtual cluster are known in advance. The process of developing a virtual cluster is shown in the following.

1.Register selected vm-containers to each Rocks DB.

2.Establish a P2P virtual private network among the selected vm-containers and make configuration for VMs.

3.Start a Rocks virtual Frontend on a physical frontend.

4.Start VMs on selected vm-containers.

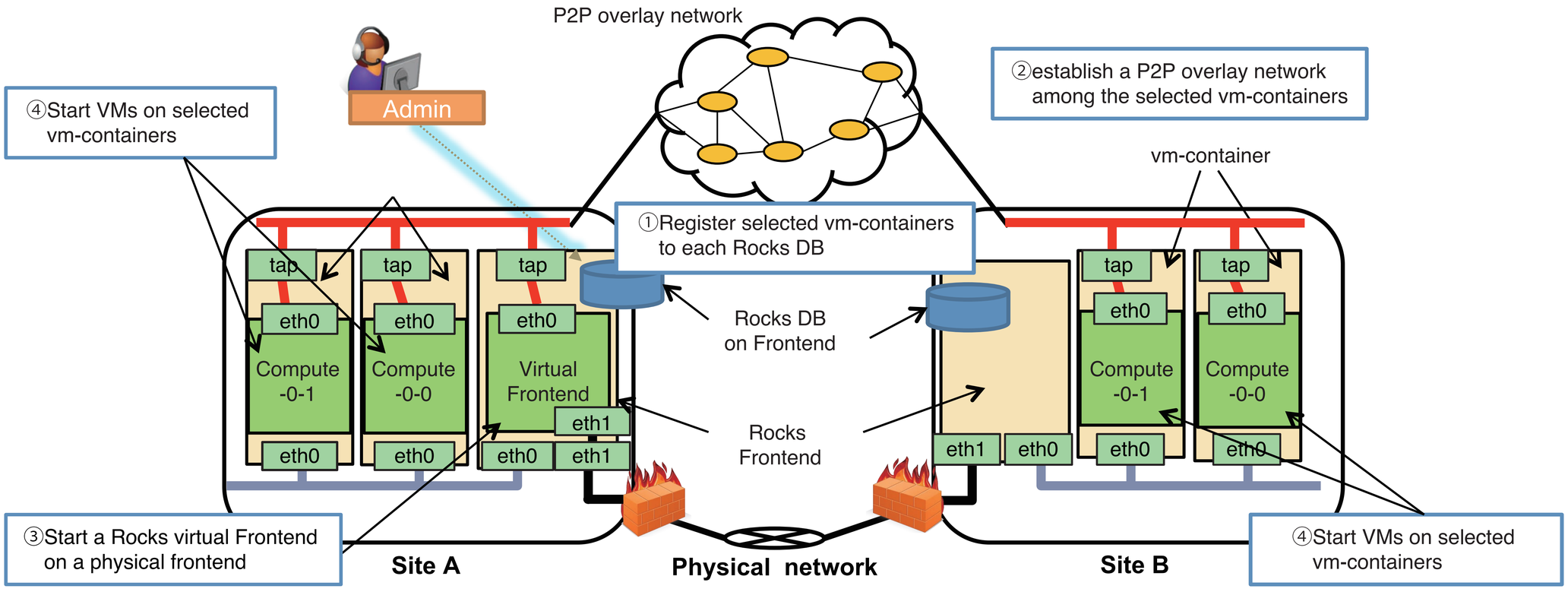


Figure 2 : Detaildesign of our proposed virtual cluster

Contact: Kei Kokubo, Adrian Ho, Kohei Ichikawa, Date Susumu and Jason Haga
E-mail: kokubo.kei@ais.cmc.osaka-u.ac.jp

PETFLOW
a project towards an ultra parallel synergy internet system in scientific applications

Cybermedia Center
Osaka University

Cybermedia Center, Osaka University, Japan

BACKGROUND

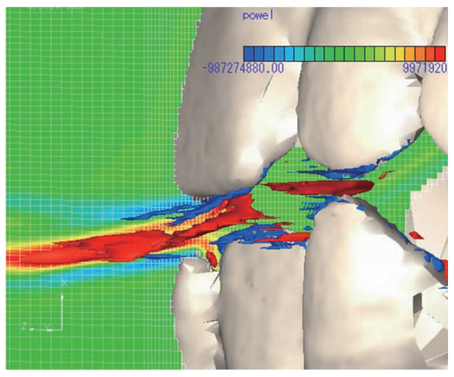
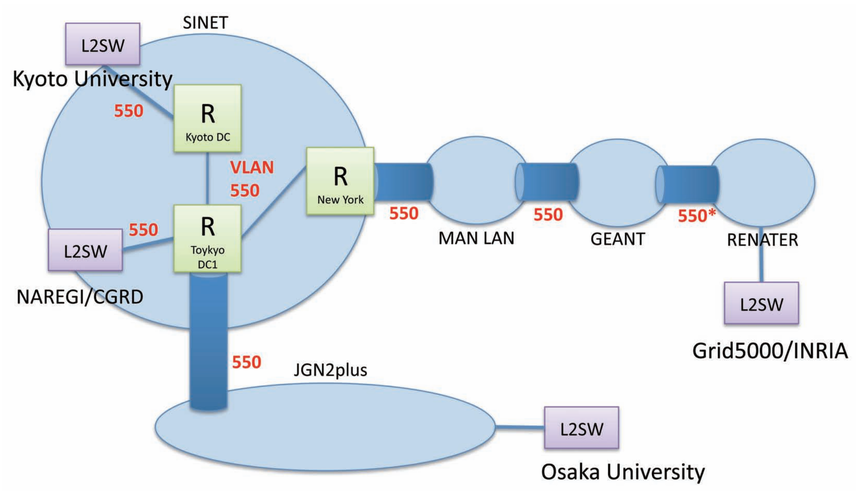


Fig.1 (Top) :
The oral flow-sound simulation in pronunciation of "/s/"

It is no falsehood to state that 'current society and science attempt to deal with increasing amounts of data' . Today, peta-scale data are commonly gathered as well as generated thanks to the continuous development of measurement technologies and computational resources in diverse fields of science and society. Efficient processing or generation of peta-scale data requires high performance computational (HPC) resources which should be made remotely accessible through long-distance high performance networking and might be represented thanks to the petaflow-project, interactive scientific visualization. Consequently, generation or processing of peta-scale data benefits from the emergence of adequate 'Information and communication technologies (ICT)' with respect to high performance 'computing-networking-visualization' and their mutual 'awareness' . It is aimed to develop and validate such ICT solutions using a transnational high speed research network between Japan and France connecting 'GRID5000' (France) to the 'Naregi' (Japan) testbed. Data-transfer protocols are aimed to be validated on data obtained for a real scientific problem involving peta-scale data. Due to the medical relevance as well as basic scientific interest, peta-scale data are obtained from HPC Computational Fluid Dynamics (CFD) simulations on a vector supercomputer (NEC SX9 Japan) aiming to predict the airflow through the upper airways. In addition, CFD simulation outcome is used as an input for aero-acoustic computations (CAA) for prediction of noise production. Besides the international transfer of the generated peta-scale data, scientific visualization of peta-scale data is aimed on a single PC as well as on a tiled display wall for 3D interactive reconstruction of the flow and noise data. In summary, the petaflow project aims to contribute to the state-of-theart of HPC, networking, scientific visualization and their mutual interactions for peta-scale data, while at the same time it is aimed to contribute to basic research in the fields of CFD and CAA applied to flow through the upper airways.

PETAFLOW NETWORK

PetaFlow network testbed is a Layer-2 Virtual Private Network. It has been developed from NAREGI-Grid5000 network testbed (2006--2009) and constructed with the collaboration of SINET, JGN2plus, RENATER, GEANT, and MAN LAN. Figure XX shows the topology of the PetaFlow network testbed. Japanese-side network is composed of SINET and JGN2plus networks connected at Tokyo. NII and Kyoto University connect with SINET, and Osaka University connects with JGN2plus. In order to connect Japanese research foothold with Grid5000, the international network operated by SINET is used and extends to MAN LAN. On the other hand, Grid5000 backbone network is provided by RENATER. The network from Grid5000 extends to MAN LAN via GEANT.



PETAFLOW CLOUD

There are many kinds of applications which perform numerical simulations, for example, Virtual Physiological Human (VPH), Real time Space Weather Simulater, Numerical Whether Simulation, Nuclear Fusion Simulations, Aerodynamic Research for the Next Generation Supersonic Transport and Numerical Simulation to Tsunami Disaster Prevention. It is necessary to increase the total amount of computational resources to perform those applications. It has not been sufficient, however, to do them at all.
This problem arise from such a structure that each researchers or groups own their own computational resources. This structure of scientific society is not efficient in terms of several aspects. We hereby propose "PetaFlow cloud". PetaFlow cloud enables researchers or groups to obtain the solutions for their concrete objects that can be required by several numerical simulation codes. PetaFlow cloud has three main components, first "Peta-scale Networking-Storage", second "Peta-scale Computing" and "Peta-scale Visualization".

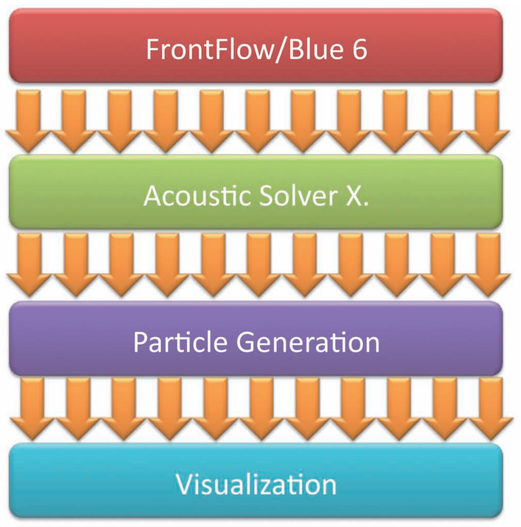
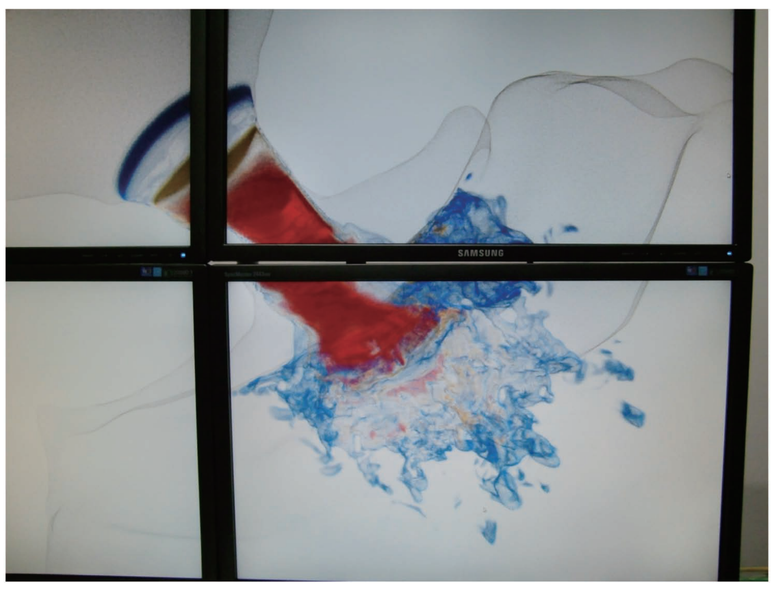


Fig.6(c) : Vision of PetaFlow:
- Peta-scale Networking – Storage
- Peta-scale Computing
- Peta-scale Visualization

PETASCALE VISUALIZATION

Oral flow-sound simulation results in higher-level complexities of 3D phenomena and huge volume datasets. Visualizing huge data often requires use of high-resolution display such that important fine structures are not missed. It is because the number of elements forming the huge volume data exceeds resolutions of the normal LCD displays.
The particle-based volume rendering (PBVR) is one of the effective rendering techniques applicable to huge volume data. It is based on the Sabella' s density emitter model, in which the scalar field is characterized as a cloud of opaque and self-emitting particles with the single-level scattering. PBVR does not require sorting of elements, being different from the ray-casting method, and so enables us to treat over giga-byte huge data easily.



Paulo Goncalves : INRIA, ENS Lyon, Universit´e de Lyon
Xavier Grandchamp : Gipsa-lab, UMR CNRS 5216, Grenoble Universities
Xavier Pelorson : Gipsa-lab, UMR CNRS 5216, Grenoble Universities
Bruno Raffin : INRIA Grenoble, France
Annemie Van Hirtum : Gipsa-lab, UMR CNRS 5216, Grenoble Universities
Pascale Vicat-Blanc : INRIA, ENS Lyon, Universit´e de Lyon
Ken-ichi Baba : Vizlab, Kyoto University
Julien Cisonni : The Center for Advanced Medical Engineering and Informatics
Yasuo Ebara : Cybermedia Center, Osaka University
Kazunori Nozaki : The Center for Advanced Medical Engineering and Informatics
Hiroyuki Ohsaki : Graduate School of Information Science and Technology
Shigeo Wada : The Center for Advanced Medical Engineering and Informatics
Takuma Kawamura : Vizlab, Kyoto University
Kohji Koyamada : Vizlab, Kyoto University
Eisaku Sakane : National Information and Communication Technology, Japan
Naohisa Sakamoto : Vizlab, Kyoto University
Shinji Shimojo : National Institute of Informatics, Japan

Contact: Petaflow
E-mail: petaflow@ispj.jp

A Handheld User Interface for GPU-Accelerated Large-scale Volume Visualization

Cybermedia Center, Osaka University, Japan

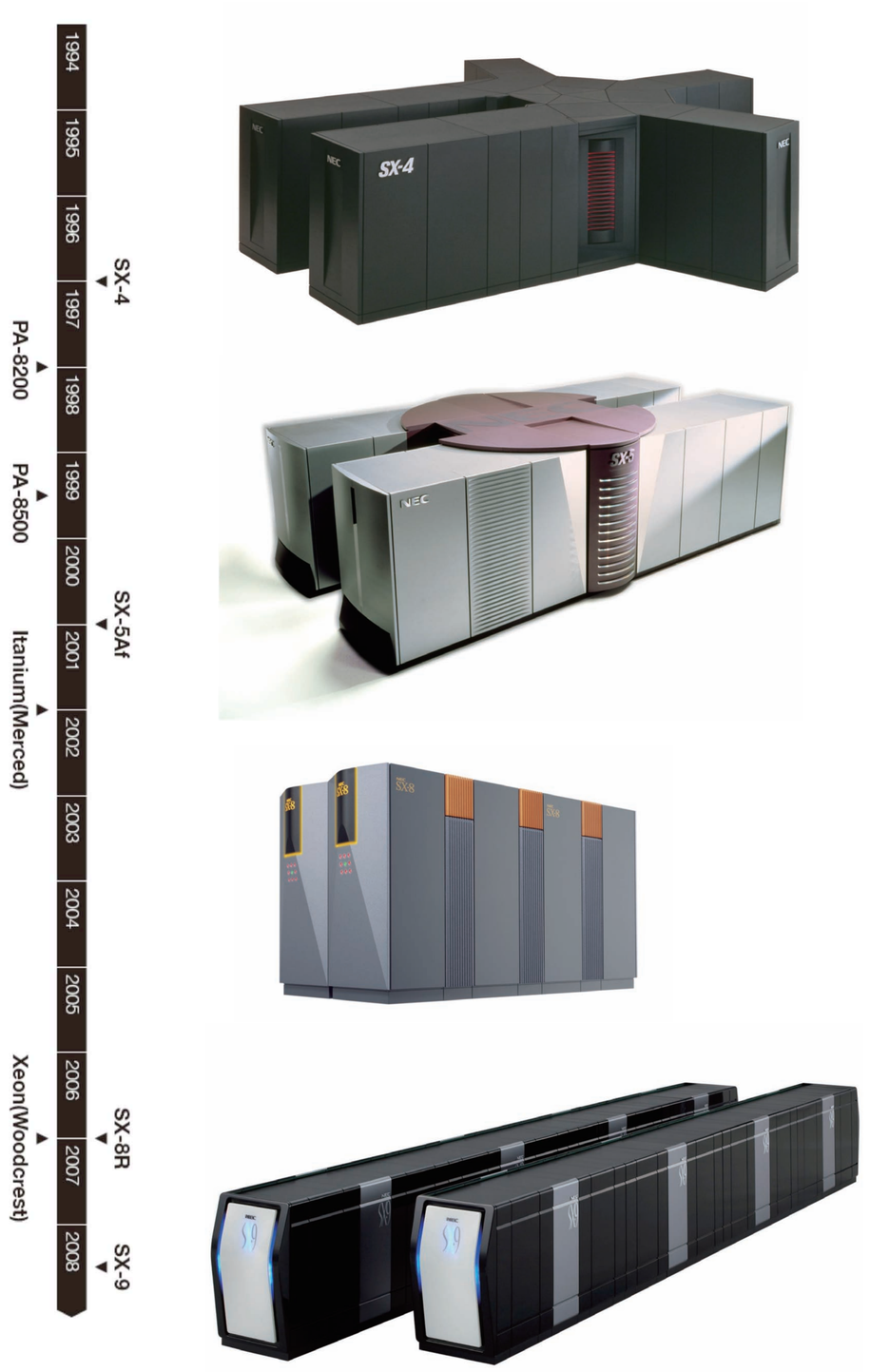
The Cybermedia Center at Osaka University was founded by merging the former Computation Center, the former Education Center for Information Processing and part of the university library in April 2000. Such reorganization was conducted in order to comprehensively promote educational study in view of rapid developments in the field of information technology.

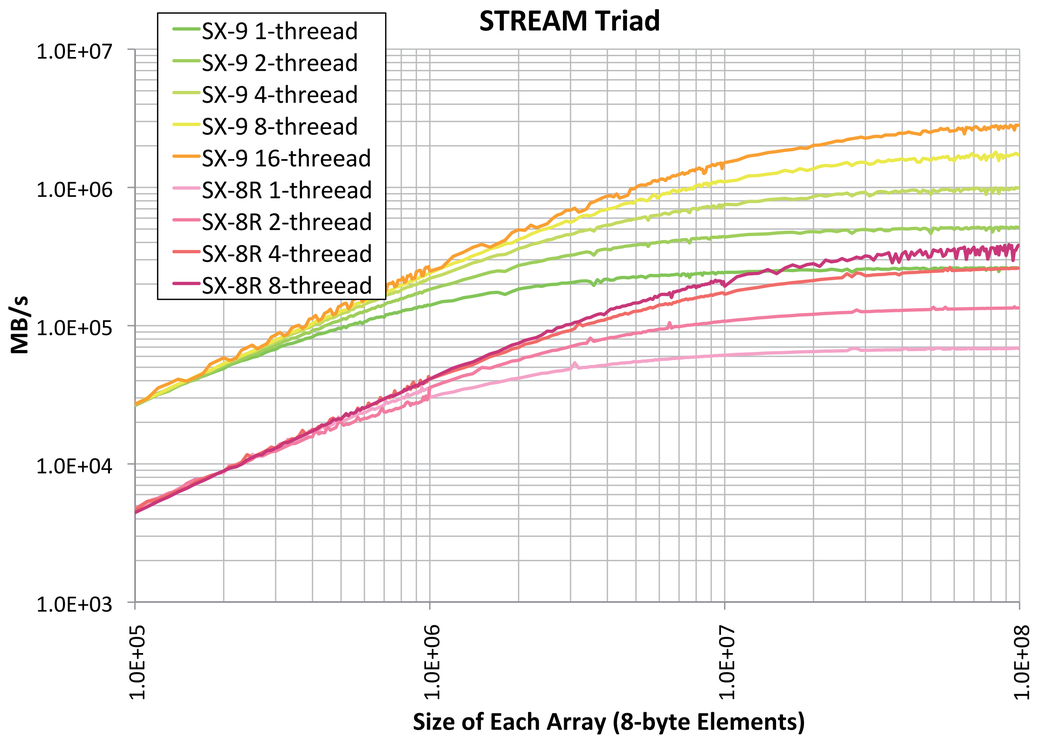
The goal of our center is twofold: 1) to continue providing stable infrastructure services as well as technical knowledge about supercomputers, information education systems and networks used around the world, and 2) to pursue research that enables the most advanced infrastructure services.

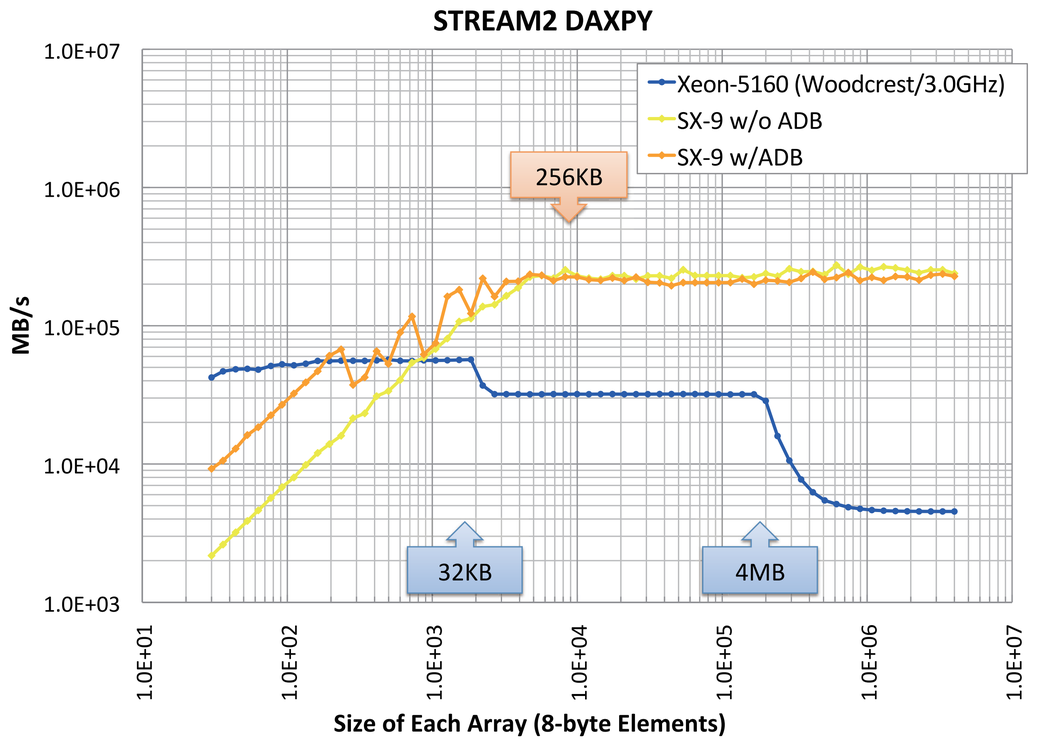
Advantages of Historical Vector Computing Maintained

At our center, we introduced 20-nodes of SX-8R in January, 2007. It replaced SX-5/128M8 which had stunned the HPC community with its peak performance of more than 1TFLOPS for the first time ever as a vector-type supercomputer and the 8th rank on the TOP500 list in 2001. While we have witnessed a phenomenal increase in the computational performance on the TOP500 list after this SX-5, we now give first priority to the users' benefits gained through the continuous improvement in performance rather than mere performance index. In line with such a policy, we decided to upgrade this system in a phased manner with an additional 10-node SX-9 system for July, 2008.

The performance in running real application programs has been improved by the sophisticated compiler technology. While the recent microprocessors can realize the improved performance as long as their caches are effective, the supremacy of vector machines is still remarkable. Especially, multi-threaded performance with automatic parallelization of the SX-9 is outstanding.







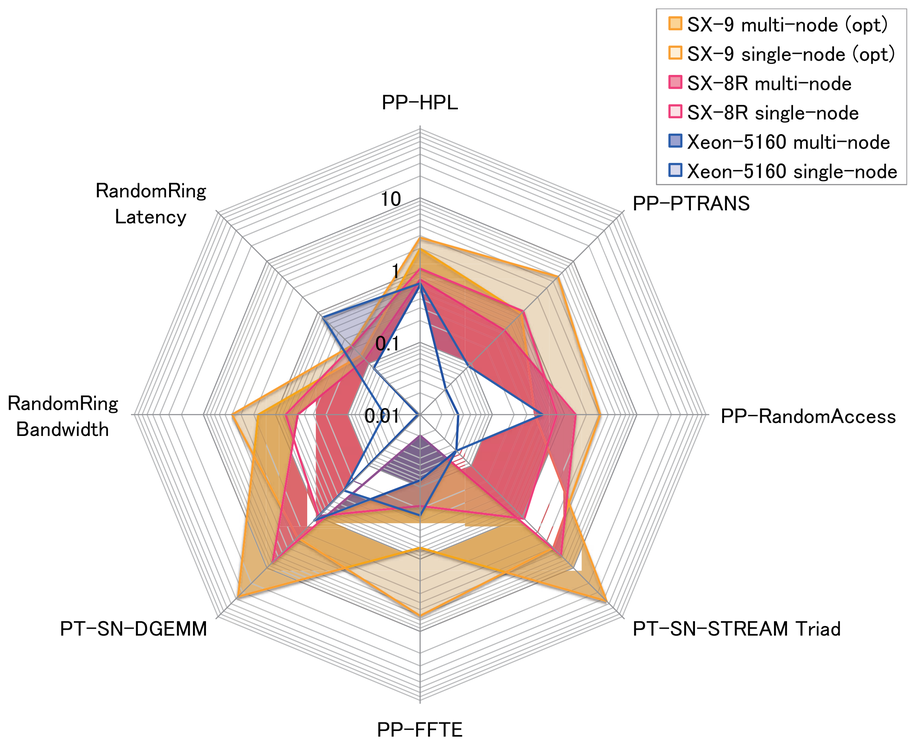
In these ten years, the technology of past high-end microprocessors has been inherited to budget-price products, and the low power consumption technology has been spread simultaneously. Although the large-scale cluster system based on PC is getting popular, it has also left the issues surrounding operation and maintenance. We built a cluster system that can closely be co-operated with the vector machine in order to gain benefits from both architectures. It became a good example which demonstrates a synergistic effect by different architectures.

In terms of the STREAM2 benchmark, the performance of vector computers is rather excellent even for very short loop lengths partly due to the automatic loop collapse by the compiler. For the SX-9, 256KB of ADB (Assignable Data Buffer) acts like a secondary cache to increase the bandwidth for short loop lengths. The efficiency of the cache of a microprocessor can now be grasped. In fact, it became clear that the vector machine is superior to conventional scalar processors also for the case with short loop lengths that was thought to be tailored to microprocessors: vector operations can be effective even for the range of very short array lengths where the L1 cache of the microprocessor is effective.

HPC Challenge Benchmark: SX vs. PC Cluster

The HPCC benchmark is gaining popularity as a comprehensive measure on the performance of HPC systems. While it is not realistic to make a straightforward comparison among different systems based on this benchmark, unlike the Linpack benchmark, the HPCC benchmark can give a certain insight into the performance characteristics of HPC systems through careful consideration.

Let us use the Kiviat Diagram for the comparative performance among different systems based on the results submitted to the HPCC benchmark site as of October, 2008. At this time, we measured the performance by setting the number of MPI processes to be identical to the number of CPUs. For the multiple-node configuration, the number of MPI processes was set to the number of nodes with non-OpenMP based parallelization within a node by utilizing the automatic parallelizable BLAS library for the SX and multi-threaded GotoBLAS for Xeon-5160.



These measurements show that the SX series has a excellent single-node performance with balanced scores for many performance measures. In contrast, microprocessor-based systems show relatively poor performance numbers depending on performance index.

System - Processor - Speed - Processors Count – OpenMP Threads – MPI Processes					G-HPL	G-PTRANS	G-Random Access	G-FFTE	EP-STREAM Sys	EP-STREAM Triad	EP-DGEMM	RandomRing Bandwidth	RandomRing Latency
					Tflop/s	GB/s	Gup/s	Gflop/s	GB/s	GB/s	Gflop/s	GB/s	usec
NEC SX-9 (opt)	3.2GHz	16	1	16	1.38215	334.561	0.402194	241.7130	2,723.90	170.24	88.73	62.8236	3.53181
		32	1	2	1.94299	128.303	0.100641	55.3206	5,630.32	2,815.16	1,305.26	27.1893	5.45687
NEC SX-8R	2.2GHz	8	1	8	0.25778	35.176	0.050399	3.6489	373.924	46.74	34.21	11.4181	3.63782
		16	1	2	0.36489	29.911	0.183769	0.7758	721.54	360.77	278.19	7.7624	6.56754
NEC Express 120Rg-1 Intel Xeon 5160	3.0GHz	2	1	4	0.04034	0.743	0.008019	1.2338	5.50	1.37	10.80	0.4898	0.97902
		32	1	16	0.61089	3.297	0.008736	6.4515	51.81	3.24	43.10	0.1685	9.67466

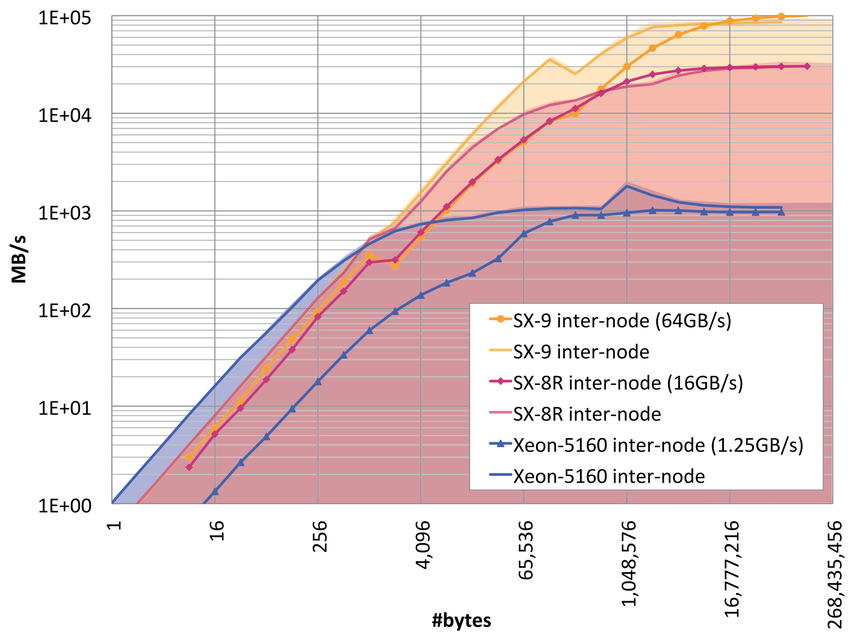
Here let us take a close look at each of the measures. While there is an increasingly narrower gap between the SX-8R and the Xeon system with respect to HPL (High Performance Linpack) and DGEMM (matrix-matrix multiply) due to the improved cache blocking and SIMD-extension/multi-threading techniques getting pervasive for commodity processors, the SX-8R still maintains competitiveness over scalar systems on STREAM and RandomRing Bandwidth. The SX-9 shows an excellent performance even for FFT with vector optimization, but the performance in HPL is limited due to the limitation resulting from 32-bit indexing in coding of HPL. Here we should be aware of the controversy over HPL as a neutral performance indicator, since the recent commodity processor-based systems can easily achieve a relatively high HPL performance even with low-speed interconnects. On the other hand, there are also indexes subject to change depending on the actual number of nodes used for the measurement. It can be suggested that for certain application codes, the hybrid system of a vector machine and a PC cluster might be appropriate.

The results of the HPCC benchmark shows that per-processor performance of SX-9 is the world fastest and the per-thread performance for a single-node by the automatic parallelization is also the world fastest. It is proven that ease of use combined with high performance can greatly improve the efficiency of user's research.

Of the total 20 nodes of the SX-8R system introduced at last time, 8 nodes are interconnected by bidirectional 16 GB/s through the IXS (Internode X-bar Switch) equipment. According to the Intel MPI Benchmark 3.0, the latency time measured for Ping-Pong is 3.79 microseconds, and the maximum bidirectional throughput for Send-Recv is 29.6 GB/s. On the Linpack HPC benchmark, 2.056TFLOPS (N= 352,256) with its peak performance ratio of 91.3% was achieved.

Of the total 10 nodes of the SX-9 system introduced at this time, 8 nodes are interconnected by bidirectional 64 GB/s through the IXS crossbar equipment. The latency time measured for Ping-Pong is 4.38 microseconds, and the maximum bidirectional throughput for Send-Recv is 98.0 GB/s.

All the PCs comprising our cluster system with more than 600 nodes have already been introduced and are now inter-connected with the Chelsio’ s T310-CX (10GBase-CX4, ToE-enabled). The latency time with Ping-Pong is 10.49 microseconds and the maximum bidirectional throughput for Send-Recv is 1.39GB/s. It seems that the throughput is limited due to the overhead arising from the internal processing of the sock driver of Intel’ s MPI at present. Priority is given to the flexibility of cluster construction and utility on GridMPI, although it is also possible to transpose a stack to RDMA and to improve the throughput.



OSAKA UNIVERSITY

Contact: Manabu Higashida, Ph. D.
E-mail: manabu@cmc.osaka-u.ac.jp

Development of Pedestrian Crowd Simulation on GPU and Application to Evacuation from Large-scale Underground Shopping Mall

Cybermedia Center, Osaka University, Japan

INTRODUCTION

Pedestrian crowd simulations have been developed and applied to verification of the safety in buildings or urban environments. Agent based simulation are particularly well suited to pedestrian crowd behaviors from a set of simple individual rules. However, computing and visualizing crowd behaviors in real-time is a computationally intensive task because this intensity mostly comes from the O (n2) complexity of the algorithm needed for the interactions of all agents. By using special data structures such as grids, relevant previous works reduces this complexity.

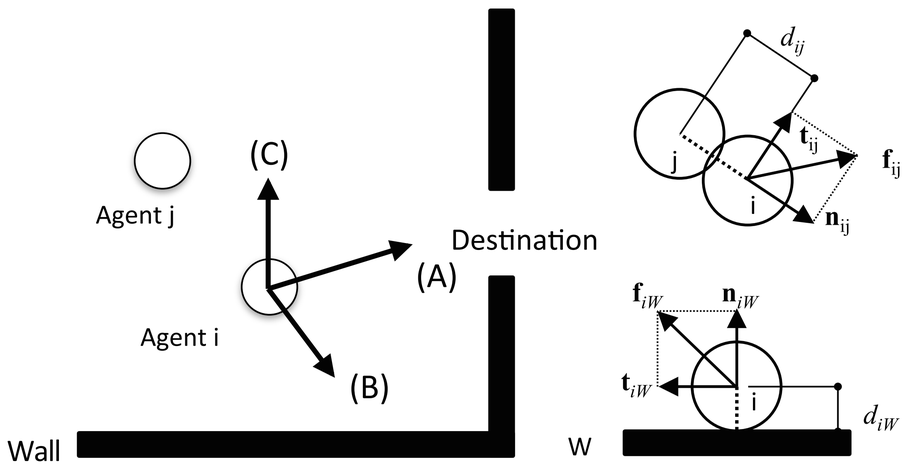
On the other hand, researchers demonstrated significantly increased speed-up after adapting existing CPU-oriented algorithms to parallel processing. Furthermore, modern GPUs have many cores; they offer large performance benefits for parallel processing at low cost. In 2007, NVIDIA released CUDA parallel-processing architecture for next-generation GPUs, letting programmers use C. It is becoming easier to develop agent based simulation on GPU.

In this study, we present a GPU based implementation of agent-based pedestrian crowd simulation and an application to the evacuation from a large-scale underground shopping mall.

IMPLEMENTATION

Agent-Based Modeling

We implemented the social force model in order to model pedestrian crowd behavior. The social force model solves the motion equation of an agent, which is represented by a moving disc (Figure 1). An agent that has mass and a constant radius is subjected to a force exerted by other agents and obstacles and exit position. These steps perform all operations on a GPU using CUDA (Figure 2). Parallel processing and the GPU' s many cores produce huge computational power, enough to update massive pedestrian crowds.



$$m_i \frac{d\mathbf{v}_i}{dt} = m_i \frac{v_i^0(t)\mathbf{e}_i^0(t) - \mathbf{v}_i(t)}{\tau_i} + \sum_{j(\neq i)} \mathbf{f}_{ij} + \sum_W \mathbf{f}_{iW}$$

$$\mathbf{f}_{ij} = \left\{A_i \exp[(r_{ij} - d_{ij})/B_i] + kg(r_{ij} - d_{ij})\right\}\mathbf{n}_{ij} + \kappa g(r_{ij} - d_{ij})\Delta v_{ji}^t \mathbf{t}_{ij}$$

$$\mathbf{f}_{ib} = \left\{A_i \exp[(r_i - d_{iW})/B_i] + kg(r_i - d_{iW})\right\}\mathbf{n}_{iW} - \kappa g(r_i - d_{iW})(\mathbf{v}_i \cdot \mathbf{t}_{iW})\mathbf{t}_{iW}$$

Fig.1 Social Force Model

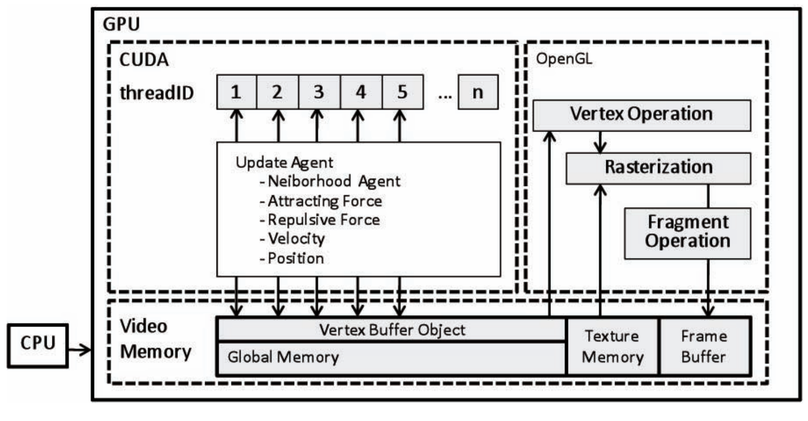


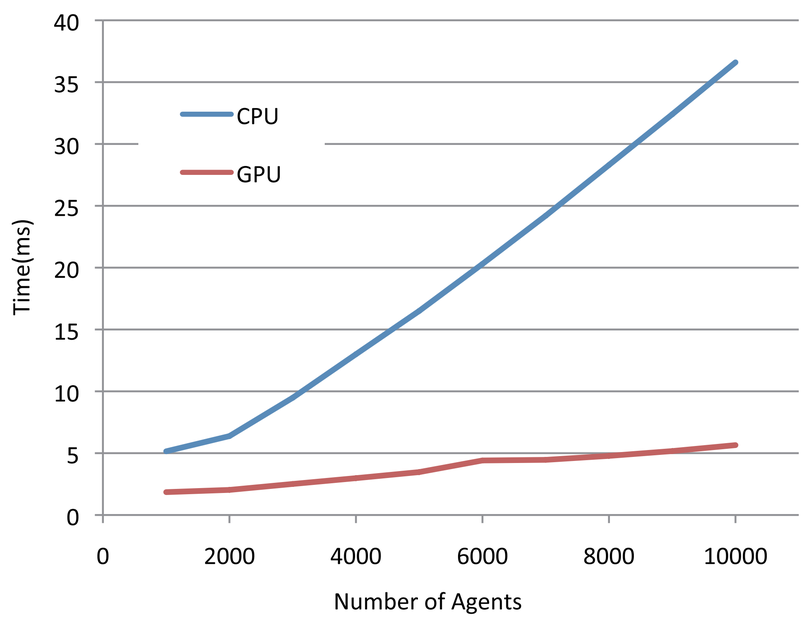
Fig2. Operations on GPU

CASE STUDY AND PERFORMANCE

As a case study of the pedestrian crowd simulation, we configured an evacuation situation in a large-scale underground shopping mall in Osaka, Japan in Figure 3.

As a performance testing, we evaluated two versions of the simulation. While the first was executed on the GPU, the second was executed on the CPU. Both versions used the same algorithm and data structure. The average time to compute the social force of all agents without rendering process is shown in Figure 4. These simulations were executed for each implementation type varying only the number of agents, ranging from 1,000 to 10,000. As a result, the GPU version presents better scalability than the CPU version. When the number of agent is 10,000, the GPU version is approximately seven times faster than the CPU version. Moreover, it is possible to see that the performance of the GPU implementation sustained interactive frame rates with rendering of complex models of agents and buildings.

Fig3. Application to Large-scale Underground Shopping Mall



These tests were performed on an Intel Core i7 CPU 930@2.80GHz with a NVIDIA GeForce GTX 460 GPU

Fig. 4 Performance Result of CPU and GPU

Conclusion

We showed an implementation capable of running up to 10,000 abents at an interactive frame rate using current graphics hardware and CUDA technology

Contact: Kensuke Yasufuku, Ph.D.
E-mail: yasufuku@cmc.osaka-u.ac.jp

Development of Pedestrian Crowd Simulation on GPU and Application to Evacuation from Large-scale Underground Shopping Mall