Japanese Computational Grid Research Project: NAREGI

SATOSHI MATSUOKA, MEMBER, IEEE, SINJI SHIMOJO, MEMBER, IEEE, MUTSUMI AOYAGI, SATOSHI SEKIGUCHI, MEMBER, IEEE, HITOHIDE USAMI, AND KENICHI MIURA

Invited Paper

The National Research Grid Initiative (NAREGI) is one of the major Japanese national IT projects currently being conducted. NAREGI will cover the period 2003–2007, and collaboration among industry, academia, and the government will play a key role in its success. The Center for Grid Research and Development has been established as a center for R&D of high-performance, scalable grid middleware technologies, which are aimed at enabling major computing centers to host grids over high-speed networks to provide a future computational infrastructure for scientific and engineering research in the 21st century. As an example of utilizing such grid computing technologies, the Center for Application Research and Development is conducting research on leading-edge, grid-enabled nanoscience and nanotechnology simulation applications, which will lead to the discovery and development of new materials and next-generation nanodevices. These two centers are collaborating to establish daily research use of a multiteraflop grid testbed infrastructure, which will be built to demonstrate the advantages of grid technologies for future applications in all areas of science and engineering.

Keywords**—***Computational grid, nanoscience, research grid, virtual organization hosting service.*

Manuscript received March 1, 2004; revised June 1, 2004. This work was supported by a grant from the Ministry of Education, Sports, Culture, Science and Technology (MEXT) of Japan.

S. Matsuoka is with the Global Scientific Information and Computing Center, Tokyo Institute of Technology, Tokyo 152-8550, Japan, and also with the the NAREGI Project, National Institute of Informatics, Tokyo 101-0051, Japan (e-mail: matsu@is.titech.ac.jp).

S. Simojo is with the Cybermedia Center, Osaka University, Osaka 567-0047, Japan; E-mail: shimojo@cmc.osaka-u.ac.jp.

M. Aoyagi is with the Computing and Communications Center at Kyushu University, Fukuoka 812-8581, Japan (e-mail: aoyagi@cc.kyushu-u.ac.jp).

S. Sekiguchi is with Grid Technology Research Center, the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan (e-mail: s.sekiguchi@aist.go.jp).

H. Usami is with the Computational Science and Engineering Center, Fujitsu Ltd., Chiba 261-8588, Japan, and also with the Center for Grid Research and Development, National Institute of Informatics (NII), Tokyo 101-0051, Japan (e-mail: usami@grid.nii.ac.jp).

K. Miura was with Fujitsu Laboratory, Kanagawa 211-8588, Japan. He is now with the Center for Grid Research and Development, National Institute of Informatics (NII), Tokyo 101-0051, Japan (e-mail: kenmiura@grid.nii.ac.jp).

Digital Object Identifier 10.1109/JPROC.2004.842748

I. INTRODUCTION

The NAREGI project is a five-year project that was instituted in fiscal 2003 in Japan as part of a “leading project for economic activation”[1]. NAREGI aims to research and develop high-performance, scalable grid middleware for the national scientific computational infrastructure. Such middleware will help facilitate computing centers within Japan as well as worldwide in constructing a large-scale scientific “research grid” for all areas of science and engineering, to construct a “National Research Grid.”

Another essential part of NAREGI is the inclusion of nanoscience as a representative application area, and large-scale nanoscience simulation on the grid being is another of the project’s primary objectives. We assume that the future computational environment for scientific research will have a computational scale well beyond 100 teraflops and tens of thousands of users online. As such, the grid-enabled nanoscience applications associated with NAREGI will serve as hallmarks of the project to evaluate the effectiveness of the grid middleware that we will develop. The experimental deployment of these applications will be significant in terms of the scale of the computational requirements. It will also provide a virtual distributed computing environment with a large number of users in nanoscience and nanotechnology, from academia as well as industry.

The middleware R&D work is being conducted at the newly established Center for Grid Research and Development, hosted by the National Institute of Informatics (NII), Tokyo, Japan. The grid-enabled nanoscience application work is under the auspice of the Center for Applications Research and Development, hosted by the Institute for Molecular Science (IMS), Okazaki, Japan. These two centers are collaborating to establish and operate a dedicated NAREGI testbed with Japan’s SuperSINET as the underlying network infrastructure. The testbed will facilitate nearly 18 teraflops of computing power distributed over nearly 3000 processors. Both the developed grid middleware

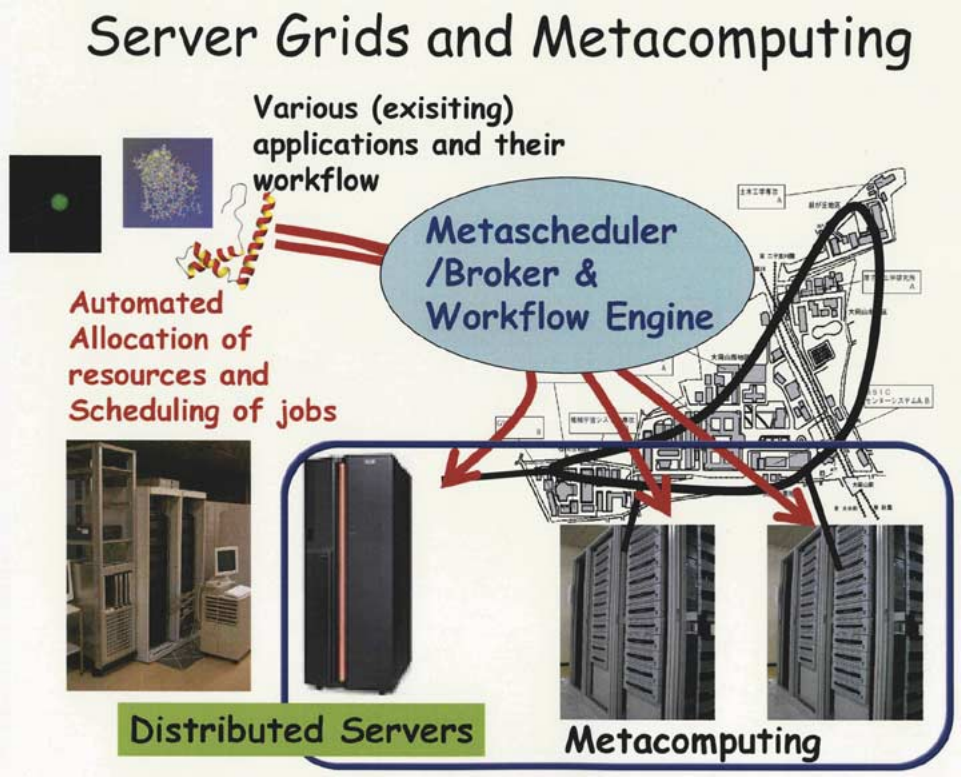

Server Grids and Metacomputing
Various (exisiting) applications and their workflow
Metascheduler /Broker & Workflow Engine
Automated Allocation of resources and Scheduling of jobs
Distributed Servers
Metacomputing

Fig. 1. Server grid.

and the grid-enabled nanoscience applications will be under scrutiny and expected to achieve performance over a scale, as well as serving to test the stability and manageability of future grids hosted by the two centers and utilized by various application domains. This article mostly focuses on the middleware R&D activities conducted at NII's Center for Grid Research and Development.

II. NAREGI Middleware Framework

A. Overview

The overall goal of the NAREGI project is to develop the technological foundations for "research grids," i.e., grids that support large-scale scientific and technological R&D. More specifically, we research develop grid middleware foundations for research grids. In addition, we investigate the architecture of large, distributed, terascale computing infrastructures, high-speed networks to support research grids, and methodologies for effective application development on the underlying distributed infrastructure. In the long term, the project deliverables are expected to enable scientists and engineers to use grids on a daily basis, through the aggregation of large amounts of resources from supercomputers and large archival storage facilities to desktop resources, such as workstations, personal computers, and possibly PDAs, thus accelerating the research conducted by multiple different institutions and companies as members of "virtual organizations" (VOs) [2].

Before examining the NAREGI project in more detail, we first must answer the question of what is meant by a "research grid," especially with respect to the deliverable goal of the project. The general definition of a grid and its possible applications have experienced a widening trend for the past several years, but for the purposes of NAREGI's five-year period, there are two types of objectives: short-term and long-term.

B. Short-Term Objectives

As a first step, we aim to construct a virtual computing environment encompassing computing centers at various universities and research institutions along with their users. By aggregating these extensive computing resources and facilitating seamless access to them, we can implement the so-called Server Grid, as shown in Fig. 1. The NAREGI Security Infrastructure will serve as the basis of a single sign-on to the entire grid, and users will be able to seamlessly navigate in the grid environment by utilizing the Grid Problem-Solving Environment (GridPSE).

C. Long-Term Objectives

As a longer term goal, we aim to establish a grid middleware stack in which computing centers and other large organizations can serve as hosting centers for multiple VOs, as proposed by Foster *et al.* [2]. Currently, most grids in offered production today are either actually intraorganizational or operated by application domains, such as the High-Energy Physics (HEP) Grid. We feel that such situations are wasteful, with duplicated efforts as well as difficulty in providing intergrid interoperability. Instead, analogously to the Internet, centers and providers should act on behalf of VOs of various sizes, ranging from large application areas to specific projects managing their infrastructures in various ways.

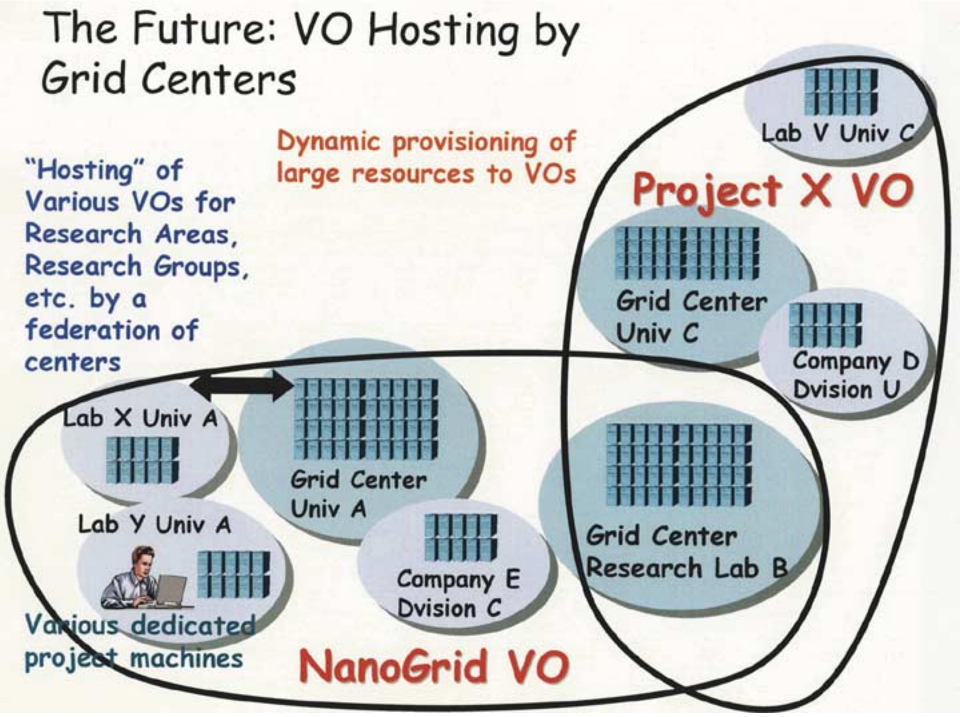

The Future: VO Hosting by Grid Centers
"Hosting" of Various VOs for Research Areas, Research Groups, etc. by a federation of centers
Dynamic provisioning of large resources to VOs
Lab V Univ C
Project X VO
Grid Center Univ C
Company D Dvision U
Lab X Univ A
Grid Center Univ A
Lab Y Univ A
Company E Dvision C
Grid Center Research Lab B
Various dedicated project machines
NanoGrid VO

Fig. 2. Future VO hosting by multiple centers.

As Fig. 2 depicts, the resources owned by different VOs may or may not be shared, while access to a pool of resources on the provider side will be granted on a VO-by-VO basis according to service-level agreements. The current crop of grid middleware partially supports VO management, but for real operation many issues remain, including authorization enforcement, VO-based accounting, integrating multiple overlapping VO policies, and hosting the management of multiple large-scale VOs.

III. Grid Middleware

The grid middleware R&D work consists of six research and development groups, as shown in Fig. 3, which are referred to as "Work Packages" (WPs). WP-1 focuses on lower and middle-tier middleware for resource management, such as a superscheduler, GridVM (providing local resource controllers), and information services on the grid. WP-2 covers basic parallel programming tools for the grid, mainly consisting of two key middleware pieces: GridRPC (for task-parallel applications) and GridMPI (for data-parallel applications). WP-3 works on grid tools for end users, including workflow, problem-solving environment (PSE), and visualization tools. WP-4 deals with packaging and configuration management of the software products from the project, while WP-5 investigates networking, security, and user management issues for high-performance grid infrastructures, such as real-time traffic measurement, QoS provisioning, and optimal routing for VOs and robust file and data transfer protocols. Finally, WP-6 acts as a liaison with the Center for Applications Research and Development, developing application-specific middleware components in order to grid-enable large-scale nanoscience applications, including those that require coupling of multiple applications on the Grid.

A. Lower and Middle-Tier Middleware (WP-1)

The requirements for a scheduler that can handle the widely distributed computing resources of a grid environment include a high level of scalability, fault tolerance, and collaborative scheduling functions coordinating among multiple sites. This area of research and development covers such components as a "superscheduler" that can manage all scheduling over a wide area, a broker that can secure computational resources meeting user requirements, such as the number of CPUs, urgency, and cost, a scheduler for the cluster environment, middleware for computational resources, networks, and grids, and tools for monitoring information and managing system configurations for the various applications.

1) Superscheduler: This is a scheduling system for large-scale control and management of a wide variety of resources shared by different organizations in the grid environment. The system will be aimed primarily at identifying resources that can meet requests from batch job users and allocating these resources to specific jobs.

2) GridVM (Local Resource Controllers): This is a new grid middleware that deploys a virtual layer of computing resources in the grid environment and facilitates resource utilization, resource protection, and fault tolerance.

3) Information Services: A secure, scalable resource information management service will be established for the purpose of running a large-scale, multidiscipline grid computing environment.

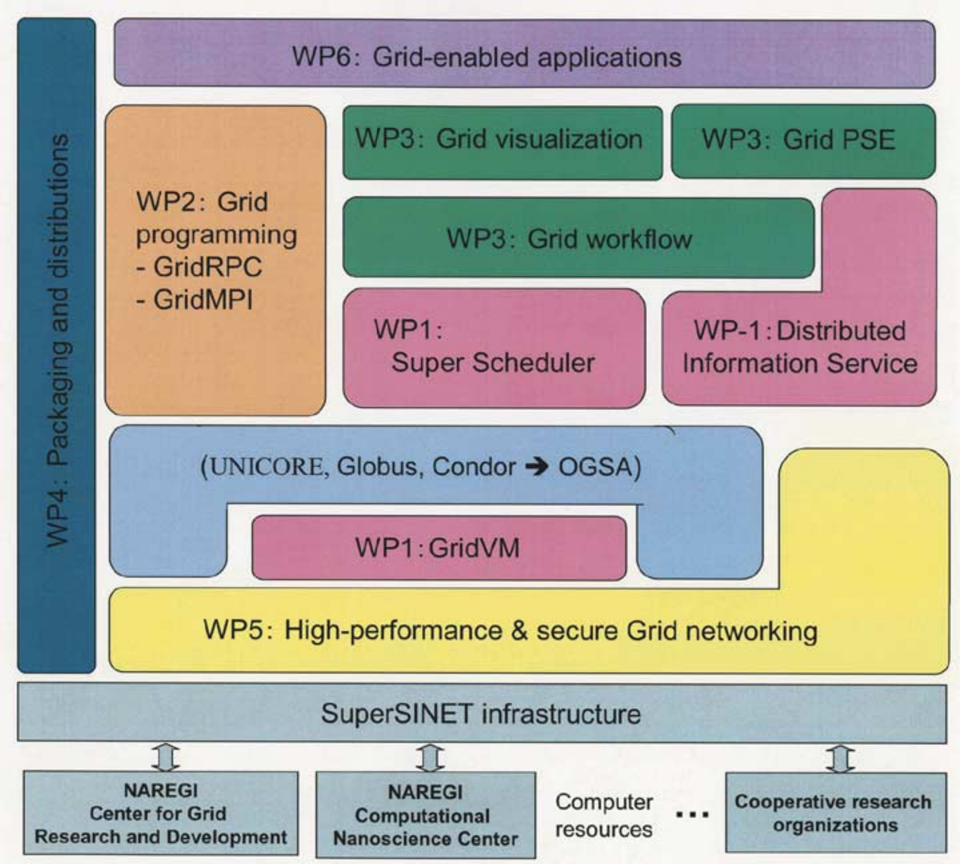

WP6: Grid-enabled applications
WP4: Packaging and distributions
WP2: Grid programming
- GridRPC
- GridMPI
WP3: Grid visualization
WP3: Grid PSE
WP3: Grid workflow
WP1: Super Scheduler
WP-1: Distributed Information Service
(UNICORE, Globus, Condor ➔ OGSA)
WP1: GridVM
WP5: High-performance & secure Grid networking
SuperSINET infrastructure
NAREGI Center for Grid Research and Development
NAREGI Computational Nanoscience Center
Computer resources
...
Cooperative research organizations

Fig. 3. NAREGI grid middleware stack.

B. Grid Programming Environment (WP-2)

There have been various attempts to provide a programming model and a corresponding system or language appropriate for grid computing. Many such efforts have been collected and catalogued by the Advanced Programming Models Research Group (APM-RG) of the Global Grid Forum (GGF). Two particular programming models that have proven viable are a Remote Procedure Call (RPC) mechanism tailored for the grid, called GridRPC, and a grid-enabled Message Passing Interface (MPI). Although from a very high level viewpoint, the programming model provided by GridRPC is that of standard RPC combined with asynchronous, coarse-grained parallel tasking, in practice, there are a variety of features that will largely hide the dynamic, insecure, and unstable aspects of the grid from programmers.

1) GridRPC: GridRPC not only enables individual applications to be distributed ut also can serve as the basis for even higher level software substrata, such as distributed scientific components on the grid. Ninf-G is a reference implementation of the GridRPC API and a proposed GGF standard. Ninf-G aims to support the development and execution of grid applications that will run efficiently on a large-scale computational grid. Here, the large-scale computational grid in question is a cluster of about ten geographically distributed cluster systems, each consisting of tens to hundreds of processors. Ninf-G has been designed to provide: 1) high performance in a large-scale computational grid; 2) the rich functionalities required to compensate for the heterogeneity and unreliability of the grid environment; and 3) an application programming interface (API) supporting easy development and execution of grid applications.

Ninf-G has been implemented to work with basic grid services, such as GSI, GRAM, and MDS in the Globus Toolkit, version 2. Ninf-G employs the following components from the Globus Toolkit: the Grid Resource Allocation Manager (GRAM) invokes remote executables; the Monitoring and Discovery Service (MDS) publishes interface information and the pathnames of GridRPC components; Globus-IO is used for communication between clients and remote executables; and Global Access to Secondary Storage (GASS) redirects stdout and stderr of the GridRPC component to the client tty [3]–[6].

We evaluated the performance of Ninf-G by using a weather forecasting system developed with it [7]. The experimental results showed that Ninf-G enables stable, efficient utilization of a large-scale cluster of clusters. We were, thus, able to demonstrate the feasibility of Ninf-G for this type of environment and to confirm that it can run on a computational grid with realistic performance for relatively fine-grain, task-parallel applications, which are considered unattractive applications on a grid. Ninf-G has been released as open-source software and is available at the Ninf project home page.

2) GridMPI: GridMPI provides users an environment for running MPI applications efficiently in the grid. It has a layer called the latency-aware communication topology, which optimizes communication with nonuniform latency and hides the various lower-level communication libraries. These libraries include the socket library, PMv2 of the SCore cluster system software, and vendor-proprietary communication libraries, including a vendor MPI enabling seamless connection of the clusters and commercial parallel computers. To achieve interoperability among the MPI implementations, GridMPI also supports the interoperable MPI (IMPI) protocol. The GridMPI checkpoint facility is provided for not only fault tolerance but also process migration. One of the key issues with the checkpoint facility that has not yet been addressed is the file consistency mechanism, by which file contents being read or written in an application are consistently saved with process images. The GridMPI checkpoint facility does provide such a file consistency mechanism.

C. Grid Application Environment (WP-3)

For this grid to be widely accepted by researchers, who are the end users, the grid software environment must be easy for them to use. To this end, R&D will be conducted in areas such as the design of a workflow description language for controlling jobs on the grid, workflow tools for executing jobs in cooperation with the resource management mechanism, software tools for visualizing computational results on the grid, and a PSE to act as a software environment that can easily enable the execution, linkage, and coordination of the applications, computational modules, data, and other resources used by researchers over a wide area.

1) Grid Workflow: This is a visual tool for seamlessly preparing, submitting, and querying distributed jobs running on remote computing resources. It handles programs and data explicitly and is independent of specific Grid middleware. Complex workflow descriptions such as loops and conditional branches are supported for nanoscience applications. Graphically described workflow jobs are converted to an enhanced workflow language based on Grid Services Flow Language (GSFL), which may be a common interface with other systems such as the PSE.

2) GridPSE: This is a workflow-based software platform for executing and coordinating collaboration among simulation applications distributed in the grid environment. The applications developed by researchers and engineers are incorporated into the grid via this software platform [8].

3) Grid Visualization: This is a real-time, postprocessing visualization system for nanosimulation, capable of reducing network loads that may interfere with smooth visualization, through flexible distribution of visualization tasks in the grid environment. This system is also characterized by having functions for large-scale parallel visualization, visualization for coupled simulation, and collaboration.

Packaging (WP-4) provides a secure, scalable resource information management service that will be established for the purpose of running a large-scale, multidiscipline grid computing environment. High-performance secure grid networking (WP-5) and grid-enabled nanoapplications (WP-6) are described in more detail in the following sections.

The grid middleware platform outlined in Fig. 3 will not be developed alone or in isolation. As a matter of fact, we are planning on fostering as much international collaboration as possible and utilizing the open-source products of various grid projects worldwide. As the core grid middleware, the NAREGI project currently employs the triad of Globus, Unicore, and Condor as the underlying foundation for basic security checking, as well as basic job launching and file transfer features. In fact, NAREGI has established strong bilateral R&D relationships with the Unicore team at Fujitsu Labs Europe and the Condor Team at the University of Wisconsin, Madison, through our collaborative Unicondore project, which attempts to facilitate bilateral job submission and control between the Unicore and Condor systems. We are also in the process of becoming one of the academic affiliates for the Globus alliance. As the Open Grid Services Architecture is developed, along with core middleware that replaces or improves on the current software, we will migrate technologies in NAREGI to comply with the standard architecture.

For "traditional" applications that have been programmed without grids in mind, a user specifies preferences for where and how his job should be executed in a declarative fashion; for compound jobs, he will also specify a workflow amongst the tasks by using the NAREGI Workflow Language. The workflow might also be generated by the GridPSE. In turn, the superscheduler, based on measurements from the Grid Information Service (such as the utilization of servers on the grid, accounting/charge/authorization information, network information such as the ratio of bandwidth to staging file size, etc.), will automatically decide the sites and servers where the individual tasks should be executed. The enforcement of individual resource authorizations and quotas may be provided by the GridVM, a virtual machine tailored for the grid to implement such control as well as other sandboxing and virtualization features. The execution status of each workflow can be observed by users via a grid portal interface, which is part of the GridPSE. Some results may automatically be visualized effectively, utilizing the grid resources in parallel in the grid-enabled VisualizationX Environment.

New programs assuming large-scale parallel execution on the grid will be facilitated with generic parallel programming libraries, namely, GridRPC, which enables easy task parallelization of scientific programs, and GridMPI, which allows efficient programming and porting of programs by using the MPI standard. The GridPSE may provide additional APIs to effectively grid-enable applications, such as simple file transfer, program installation, and so forth. Such metacomputing applications will work across the grid in a highly parallel fashion; in fact, the IMS, in collaboration with the NAREGI project, will conduct extensive metacomputing application implementation, porting, and benchmarking studies of large-scale nanoscience applications for the grid, whose results will be fed back to the process of middleware development.

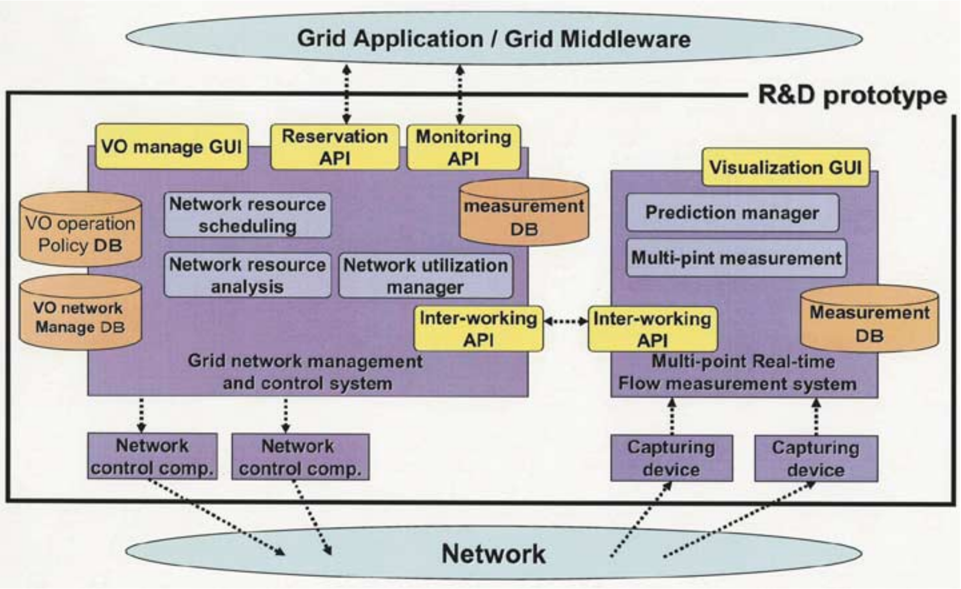


Fig. 4. Internal structure of the network management system.

IV. High-Performance, Secure Grid Networking

This section discusses the R&D of high-performance, secure grid networking in the NAREGI project. In the last decade, network infrastructure has become a very complicated combination of different subinfrastructures, with wide ranges of throughput, delay, error rate, and jitter supported by different technologies, such as MPLS, diffserv, and optical networking. Because of the nature of distributed computing, the performance of grid computing may be considerably degraded by certain conditions of the underlying networks, such as poor bandwidth, long delay, or temporal failures. Thus, we should be aware of network resources as well as computing resources [9]. The goal of the project's high-performance, secure grid networking subgroup is to develop a reliable, easy-to-use, high-performance, secure networking infrastructure for grid computing by considering the requirements of various applications; that is, the goal is to develop a high-performance "managed network." For this purpose, we have set up three subgroups. The network function infrastructure subgroup will develop a high-performance managed network in terms of VOs and applications through fine-grained, network-wide measurement. The communication protocol infrastructure subgroup will establish a methodology for analyzing and evaluating the large-scale grid network and improving the end-to-end performance of TCP. Finally, the grid security infrastructure subgroup will provide a security infrastructure for VO management.

A. *Network Function Infrastructure*

In developing high-performance networks, overprovisioning and strict, static reservation of network resources are potential solutions but neither scalable nor cost effective. Therefore, our goal for the network function infrastructure is to develop a system of measurement, management, and control for adaptively using and assigning network resources in order to avoid resource conflicts and cost-effectively maintain the network quality for grid computing. Our system consists of two parts, as illustrated in Fig. 4: multipoint fine-grained real-time measurement of network-internal traffic and status, and dynamic bandwidth control and QoS routing based on both operational policies and the network measurement. The most notable feature of the measurement capability is real-time fine-grained on-demand flow measurement, which is achieved through a distributed architecture of capturing devices and a hierarchical data structure of pattern matching [10]. At the same time, the most notable feature of the management function is resource control based on VO operational policies. Each VO is treated as a dynamic collection of distributed resources in order to manage the various organizations. In terms of network resources and VOs, the network management system enables grid applications to efficiently utilize the network infrastructure shared by multiple organizations [2], [11]. The system provides the measured information not only to the grid middleware or applications (through an API) but also to operators (through a GUI) for troubleshooting and performance tuning.

B. *Communication Protocol Infrastructure*

The currently deployed version of TCP cannot detect congestion in a network until a packet loss occurs, so that many packets will be discarded. As either the network bandwidth or the router buffer size increases, the number of lost packets becomes large, significantly degrading the TCP throughput. Therefore, our goal for the communication protocol infrastructure is to develop a communication protocol optimized for grid computing and a method of evaluating network performance. We have, thus, designed such a communication protocol providing scalability and compatibility with the existing TCP, and we have established a methodology for analyzing and evaluating large-scale grid networks [12].

Fig. 5 shows an example of modeling an entire network with our analysis technique. TCP is a sort of feedback-based control which dynamically changes its window size according to the packet loss probability in the network.

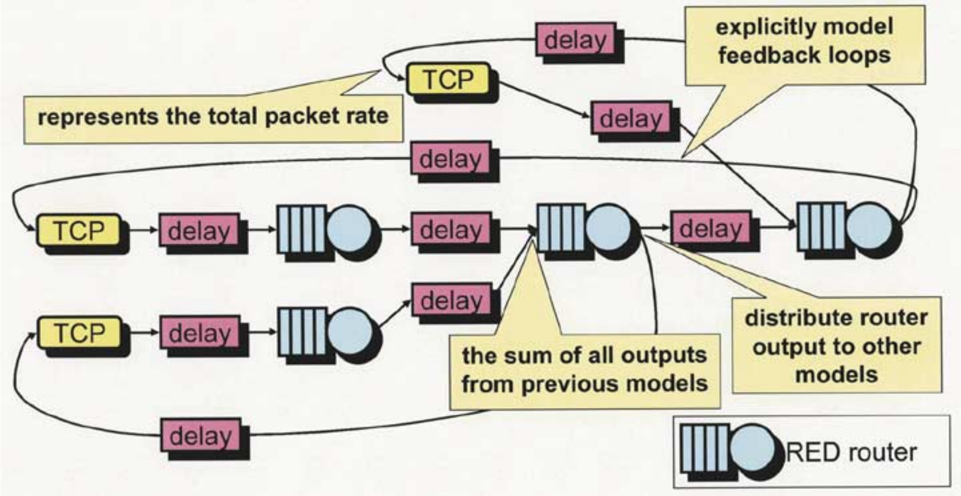

explicitly model feedback loops
delay
TCP
represents the total packet rate
delay
delay
TCP
delay
delay
delay
TCP
delay
delay
the sum of all outputs from previous models
distribute router output to other models
delay
RED router

Fig. 5. Model for performance analysis of multiple TCP connections.

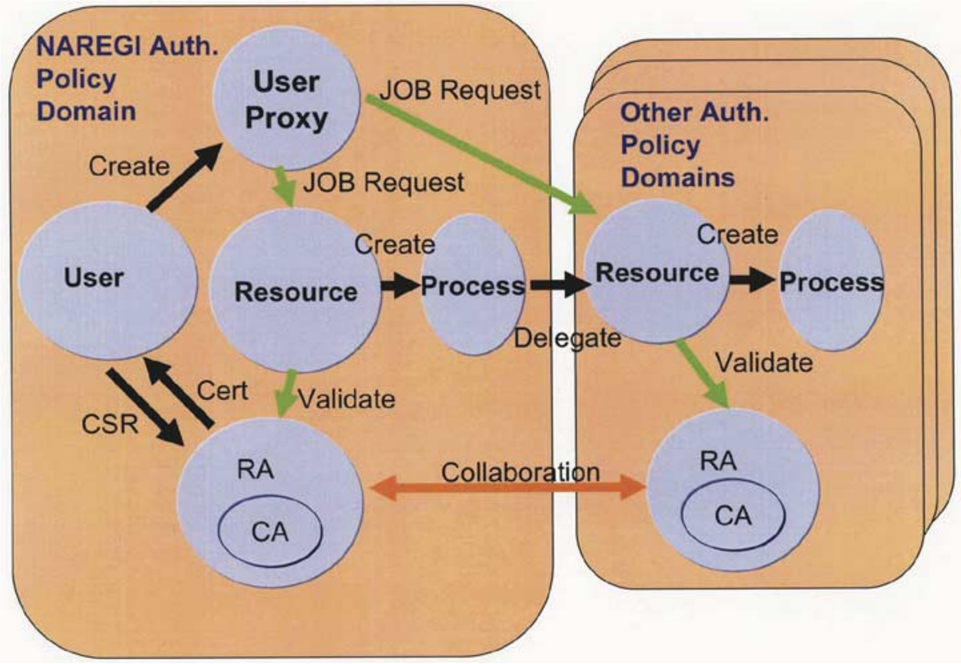


Fig. 6. Model for secure grid communication over multiple domains.

Therefore, we model multiple TCP connections as independent continuous-time systems and the bottleneck router in the network as another single continuous-time system. Using our model for GridFTP by combining these continuous-time systems, we quantitatively investigated the optimal parameter configuration of GridFTP, particularly in terms of the number of TCP connections and the TCP socket buffer size.

C. Grid Security Infrastructure

Security problems in grid computing may occur in accessing distributed resources over the network. Our goals for the grid security infrastructure are to develop a security model for grid computing based on public key infrastructure (PKI) and to implement authentication and VO management across multiple organizations [13], [14]. We have developed authentication services for Unicore and Globus with a Certificate Policy at the basic assurance level defined by GGF [15]. The Certificate Authority (CA) and Registration Authority (RA) will be developed in stages, and a prototype system is now under operational testing as part of the NAREGI project. Fig. 6 shows the NAREGI authentication services collaborating with each other to provide a uniform grid-computing environment over heterogeneous policy domains.

V. Nanoscience Applications

WP6 is developing application-specific middleware components to grid-enable large-scale nanoscience applications, including those that require coupling of multiple applications on the grid. One example of such applications is multiscale simulation, where each application component utilizes different mathematical and physical modeling approaches and cooperates on spatially or temporally different calculations. To advance such multiscale applications and, more generally,

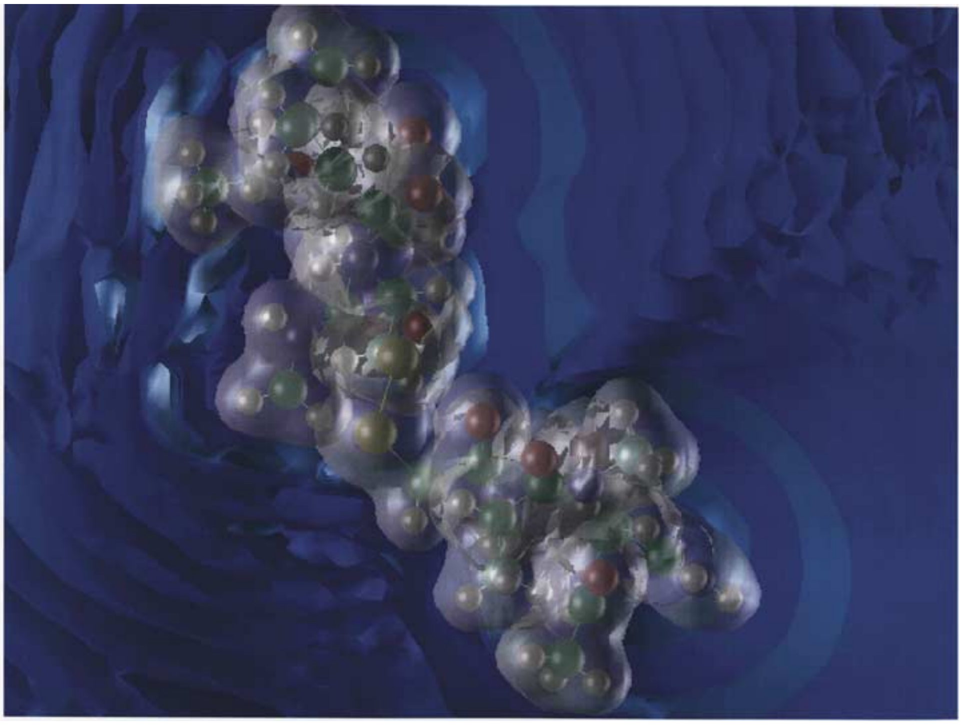


Fig. 7. Example of the RISM-FMO coupled simulation on the grid.

multiple applications, users have wasted large amounts of effort in developing custom codes and decomposing original codes for semantic-level communication between heterogeneous scientific application components.

We are instead developing a middleware system, called a "Mediator," that provides high-level transparency in automatically transferring and transforming data between heterogeneous application components. The Mediator focuses on a data-handling specification that correlates different discrete points in finite-difference method (FDM), finite-element method (FEM), or particle simulations in the unified way. It supports a variety of techniques for semantically transforming the values associated with the correlated points, e.g., in-sphere, first nearest neighbors, and nearest points [16]. To facilitate easier usage and minimize customization of original user programs, the Mediator provides three types of API, which: 1) manage task identification and construct association between Mediator processes and application processes in parallel programming style such as Single Program Multiple Data and Master-Worker; 2) register different kinds of discrete points, search the correlated discrete points, and determine processes with which communication is required by building a correlation table according to the specification; and 3) transfer messages incorporating the extraction and transformation of the values associated with the correlated points. The prototype system has been applied to semiconductor device simulation and tested on an LAN environment [17].

To demonstrate and enhance the efficiency and functionalities of the middleware system on grids, Mediator based on a grid-enabled MPI is developed for multiscale simulations in nanoscience, in which Reference Site Model (RISM) and Fragment Molecular Orbital (FMO), as shown in Fig. 7, are coupled to analyze the entire electric structure of large-scale molecules immersed in infinite solvent. RISM, originally developed by Dr. Hirata (IMS), is employed to analyze the pair correlation functions of molecular sites between a solvent and solute, while FMO is used to calculate the total electronic energy and molecular structure of the solute. The FMO method, originally developed by Dr. Kitaura (National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan), is a very efficient technique in which a large target molecule is decomposed into fragments and the total electronic energy is evaluated from the energies of the fragments and fragment pairs.

A number of other systems allow integration of scientific simulation components. JACO3 is a grid environment that supports the execution of coupled simulations based on CORBA [18]. The Common Component Architecture (CCA) provides a component architecture and interoperability standard oriented toward scientific computation and to overcome the weaknesses of existing component architectures to support high performance parallel computing [19]. These frameworks are efficient to couple functional decomposed codes wrapped into component by using scientific interface definition language (SIDL). On the other hand, to minimize impact on original and parallel simulation codes, Mediator APIs are provided to build for plug-in components and Mediator libraries based on MPI support asynchronous communication between simulation components. In practice, it needs 44 steps of Mediator APIs to couple RISM and FMO without any task of code decomposition, which demonstrates comprehensive applicability of the Mediator in nanoscience regimes. A unique set of Mediator APIs is provided for grid

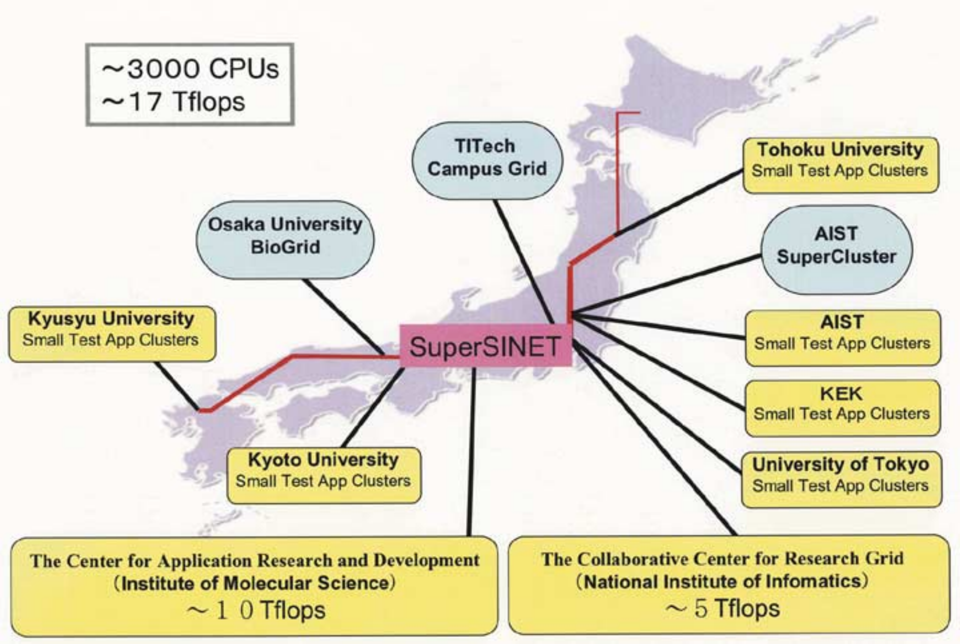

~3000 CPUs
~17 Tflops
TITech Campus Grid
Osaka University BioGrid
Kyusyu University Small Test App Clusters
SuperSINET
Kyoto University Small Test App Clusters
Tohoku University Small Test App Clusters
AIST SuperCluster
AIST Small Test App Clusters
KEK Small Test App Clusters
University of Tokyo Small Test App Clusters
The Center for Application Research and Development (Institute of Molecular Science) ~1 0 Tflops
The Collaborative Center for Research Grid (National Institute of Infomatics) ~5 Tflops

Fig. 8. NAREGI grid testbed.

and LAN versions that facilitate portability of coupled simulations on grids. Our present system has been developed based on the MPICH-G2 and Globus grid environment, and we are planning to use GridMPI in our future work.

Interoperability between nanoscience applications on the grid requires the ability not only to retrieve and transport data but also to reuse data from one application domain to another. We are also developing application-specific middleware for grid-enabled, large-scale nanoscience applications to analyze and visualize linked data sets from related domains, such as Monte Carlo calculations, molecular dynamics, electronic structure studies, and further cross-disciplinary data mining. Although this kind of interoperability is typically achieved with specialized data adapters, we have been cooperating with the Center for Applications Research and Development team. Together, we have identified common data semantics across different application domains and proposed standard intermediate representations based on the XML data format. We expect the specification of scientific metadata semantics to be a key component of the grid's data infrastructure.

To demonstrate the efficiency of intermediate representations, a prototype system based on a loosely coupled FMO simulation model, called Grid-FMO, has been developed. In this system, the data communications between the fragment calculations and the data transformations for the external visualization tool proceed through a standard intermediate format on the grid. For example, as an "initial guess" for a fragment MO, Grid-FMO could reuse previously calculated data from an external application, which is transformed into a standard intermediate representation. This greatly benefits not only the data-handling processes of various kinds of nano-applications, but also application developers by allowing them to guarantee the independence of the IO formats of specific applications.

VI. NAREGI Grid Testbed

To support effective R&D spanning five years and aiming to create production quality software for future grid computing, we decided early on in the project that NAREGI will require a large grid testbed infrastructure, which will allow scaling of the middleware and application testing to realistic scales. At NII, we are facilitating a middleware R&D testbed of over 900 processors and 5 teraflops, configured as an aggregation of ten machines of various sizes and types ranging from SMPs to PC clusters, interconnected by multigigabit networks that can simulate long-latency WAN connections and firewalls. At IMS, there is a larger nanoscience application testbed of approximately 1600 processors and 10 teraflops. The individual machines in this testbed are both larger and tailored for application benchmarking rather than middleware development. Combining the smaller clusters facilitated at partner sites, the dedicated NAREGI testbed has a capacity of approximately 2900 processors and 17.9 teraflops, interconnected at 10 Gb/s via SuperSINET, which is Japan's national research backbone network featuring advanced lambda switching for extremely high bandwidth. In addition, we have several partner sites with large grid node installations, such as AIST and Titech. Combined, the NAREGI testbed, as illustrated in Fig. 8, includes over 7000 processors and is one of the largest nonproduction grids in the world. We envision partnering with related U.S., European Union (EU), and other global grid computing efforts, such as the TeraGrid and the EU's EEGE/DEISA.

A. Computational Resources for Developing Grid Middleware (NII)

The computational resources currently set up at NII consist of the following:

- *SMP systems:*
  - —Primepower HPC2500 SPARC64V, 1.3 GHz/64 CPUs, 128 GB;
  - —SGI Altix3700 Itanium2, 1.3 GHz/32 CPUs, 32 GB;
  - —IBM pSeries690 Power4, 1.3 GHz/32 CPUs, 64 GB;
- *PC clusters:*
  - —Primergy RX200 Xeon, 3.06 GHz/128 CPUs, 130 GB, InfiniBand 4X (8 Gb/s);
  - —Primergy RX200 Xeon, 3.06 GHz/128 CPUs, 65 GB, InfiniBand 4X (8 Gb/s);
  - —Express 5800 Xeon, 2.8 GHz/128 CPUs, 65 GB, GbE (1 Gb/s), × 2 sets;
  - —HPC LinuxNetworx cluster Xeon, 2.8 GHz/128 CPUs, 65 GB, GbE (1 Gb/s) × 2 sets;
- *File server:*
  - —Primepower 900 Eternus 3000 Eternus LT160 10-TB RAID, 5 disks).

With this heterogeneous configuration, on top of which the NAREGI middleware is being developed, we expect to be able to emulate real, heterogeneous grid environments. Note that the testbed consists of SMPs and PC clusters, with each machine configured with a different OS, hardware capacity, and architecture.

In addition, to emulate long-latency network connections, we have installed a hardware gigabit network emulator called GNET-1 [20]. It allows the network time delay to be configured up to 134 ms with gigabit wire-rate transfer. By appropriately setting the latency between NII computation resources, even though they are physically connected in a gigabit-level network, we can emulate a diverse range of WAN environments. For example, a broadband network connection between the United States and Japan (i.e., a connection over a distance of 10 000 kilometers with several routers) can easily be emulated.

B. Computational Resources for Developing Nanoscience Applications (IMS)

The large-scale computational resources set up at IMS for benchmarking runs of nanoscience applications consist of the following:

- *SMP system:* SR11000 Power4, 1.7 GHz/800 CPUs (16 ways × 50 nodes), 3072 GB;
- *PC cluster:* HA8000 Xeon, 3.06 GHz/812 CPUs, 1600 GB, Myrinet2000 (2 Gb/s);
- *File server:* Primepower 900+ Eternus 3000+ Eternus LT160 30-TB RAID, 5 disks)

The SR11000 is a late-model parallel computer with high-performance internode network connections using crossbar switches, providing a very high level of computation power.

C. Network Infrastructure

The NAREGI network infrastructure employs Super-SINET, which has been managed by NII since 4 January 2002. Super-SINET is an information communication network dedicated to academic research. It connects nationwide connection points (nodes) throughout Japan via 10-Gb/s communication lines. Super-SINET began operating with the Internet for the purpose of researching high-level optical communication technology. It consists of a 10-Gb/s photonic backbone with a total length of 6000 km and very low latency through the use of dark fiber, with a Tokyo–Tsukuba round trip time of 3–4 ms. Super-SINET is connected between IMS and NII and provides a very important network infrastructure for researching grid performance.

Super-SINET is mutually connected with the Inter-Ministry Research Information Network (IMnet) and commercial ISPs in order to promote the international exchange of information, as well as the exchange of research data among the industrial, governmental, and academic sectors. NAREGI had already started joint research with many organizations in foreign countries, enabling use of the NAREGI testbed on a worldwide scale.

VII. Conclusion

The NAREGI project is now in the middle of its second year and is now fully engaged in software development for delivering the alpha version of NAREGI integrated Grid middleware at the end of March 2005. The research projects under the NAREGI project have made considerable progress in the developing the prototype software, and some of them have already produced preliminary results, including GridRPC, GridMPI and NAREGI-CA. In the NAREGI project, developing the grid middleware for seamless federation of heterogeneous resources is the primary objective. There are also some grid R&D topics that are not covered by the project, which focuses on the computational aspects of the Grids hosted by a federation of R&D centers. These topics include data grid issues, utilization of desktop resources (desktop grids), and collaborative human interface grids, such as the Access Grid. The NAREGI project will certainly stay informed on the R&D in such areas, and we may collaborate with such projects or mutually include their efforts as part of the final software distribution.

Finally, we regard Grid computing as one of the fundamental technologies of the IT infrastructure in the 21st century, and expect that the results of NAREGI project will greatly advance the R&D in scientific fields, improve Japan's international competitiveness, and have a major economic impact.

Acknowledgment

The authors would like to thank Prof. Y. Ishikawa (University of Tokyo), Dr. Y. Tanaka (AIST), Dr. H. Osaki (Osaka University), Dr. M. Imase (Osaka University), Dr. M. Tsuru (Osaka University), Mr. K. Nakamura (Kyushu Institute of Technology), Prof. T. Ikenaga (Kyushu Institute of Technology), Dr. Y. Kitatsuji (KDDI Laboratory), Dr. K. Yamazaki (KDDI Laboratory), S. Mineo (NEC), Dr. S. Hwang (NII), Dr. Y. Kodama (AIST), Y. Tanaka (Hitach), and S. Ho (HITACH) for writing and reviewing part of this paper. The authors would also like to thank collaborating members of the NAREGI project, including Prof. Y. Oie (Kyushu

Institute of Technology), Prof. S. Kawata (Utsunomiya University), Prof. M. Sato (Tsukuba University), Prof. K. Aida (Tokyo Institute of Technology), Prof. K. Kono (University of Electro-Communications), Dr. H. Nakada (AIST), A. Komatsu (NEC), M. Fujisaki (Fujitsu), M. Yamada (Fujitsu), Y. Miyahara (Fujitsu), F. Hayashi (NII), T. Otsubo (NII), Y. Koeda (NII), Dr. Y. Saeki (NII), Y. Shiji (NII), A. Yoshida (NII), T. Hirano (NII), K. Saga (NII), Dr. M. Kanamori (NII), Dr. S. Akioka (NII), and other members for R&D of this project and for valuable discussions and useful comments.

References

[1] NAREGI: National Research Grid Initiative. Nat. Inst. Informatics; Inst. Mol. Sci.. [Online]. Available: http://www.naregi.org; http://nanogrid.ims.ac.jp/nanogrid

[2] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the Grid: Enabling scalable virtual organizations," *Int. J. Supercomput. Appl.*, vol. 15, no. 3, 2001.

[3] K. Symour, H. Nakada, S. Matsuoka, J. Dongarra, C. Lee, and H. Casanova, "Overview of GridRPC: A remote procedure call API for grid computing," in *Proc. 3rd Int. Workshop Grid Computing*, 2002, pp. 274–278.

[4] K. Symour, C. Lee, F. Desprez, H. Nakada, and Y. Tanaka, "The end-user and middleware APIs for GridRPC," presented at the Workshop Grid Application Programming Interfaces, Brussels, Belgium, 2004.

[5] Y. Tanaka, H. Nakada, S. Sekiguchi, T. Suzumura, and S. Matsuoka, "Ninf-G: A reference implementation of rpc-based programming middleware for grid computing," *J. Grid Comput.*, vol. 1, no. 1, pp. 41–51, 2003.

[6] Y. Tanaka, H. Takemiya, H. Nakada, and S. Sekiguchi, "Design, implementation, and performance evaluation of GridRPC programming middleware for a large-scale computational grid," in *Proc. IEEE/ACM Int. Workshop Grid Computing*, 2004, pp. 298–305.

[7] H. Takemiya, K. Shudo, Y. Tanaka, and S. Sekiguchi, "Constructing grid applications using standard grid middleware," *J. Grid Comput.*, vol. 1, no. 2, pp. 117–131, 2004.

[8] S. Kawata, H. Usami, Y. Hayase, Y. Miyahara, M. Yamada, M. Fujisaki, Y. Numata, S. Nakamura, N. Ohi, M. Matsumoto, T. Teramoto, M. Inaba, R. Kitamuki, H. Fuju, Y. Senda, Y. Tago, and Y. Umetani, "A problem solving environment (PSE) for distributed computing," *Int. J. High Perform. Comput. Netw.*, to be published.

[9] I. Foster and C. Kesselman, *The GRID Blueprint for a New Computing Infrastructure*. San Francisco, : Morgan Kaufmann, 1998.

[10] Y. Kitatsuji and K. Yamazaki, "A distributed real-time tool for IP-flow measurement," in *Proc. 2004 Int. Symp. Applications and the Internet*, 2004, pp. 91–98.

[11] G. Wasson and M. Humphrey, "Toward explicit policy management for virtual organization," in *Proc. IEEE 4th Int. Workshop Policies for Distributed Systems and Networks*, 2003, pp. 173–182.

[12] H. Ohsaki and M. Imase, "On modeling GridFTP using fluid-flow approximation for high speed grid networking," in *Proc. 2004 Int. Symp. Applications and the Internet*, pp. 638–644.

[13] R. Housley, W. Polk, W. Ford, and D. Solo. (2002, Apr.) Request for comments 3280: Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile. [Online]. Available: http://www.ietf.org/rfc/rfc3280.txt

[14] C. Adams and S. Farrell. (1999, Mar.) Request for comments 2510: Internet X.509 public key infrastructure certificate management protocols. [Online]. Available: http://www.ietf.org/rfc/rfc2510.txt

[15] R. Butler and T. J. Genovese. (2002, Oct.) draft-gridforum-CP.txt. [Online]. Available: http://www.gridforum.org/Meetings/ggf5/pdf/GGF Certificate Policy Version 6.pdf

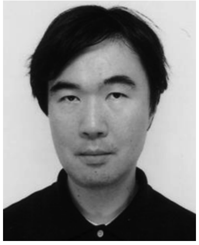
[16] S. Ho, S. Itoh, S. Ihara, and R. Schlichting, "Agent middleware for heterogeneous scientific simulations," in *Proc. 1998 ACM/IEEE Supercomputing'98 Conf.*, p. 15.

[17] S. Ho, Y. Ohkura, T. Maruizumi, P. Joshi, N. Nakamura, S. Kubo, and S. Ihara, "Hot carrier induced degradation due to multi-phonon mechanism analyzed by lattice and device Monte Carlo coupled simulation," *IEICE Trans. Electron.*, vol. E86-C, no. 3, pp. 336–349, 2003.

[18] JACO3 Project [Online]. Available: http://sourceforge.net/projects/jaco3

[19] Common Component Architecture Forum [Online]. Available: http://www.cca-forum.org

[20] Y. Kodama, T. Kudoh, R. Takano, H. Sato, O. Tatebe, and S. Sekiguchi. GNET-1: Gigabit ethernet network testbed. presented at IEEE Int. Conf. Cluster Computing [Online]



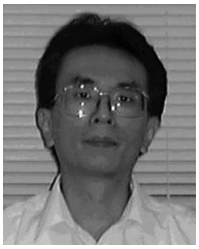
Satoshi Matsuoka (Member, IEEE) received the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in 1993.

He has been a Professor in the Global Scientific Information and Computing Center (GSIC) at Tokyo Institute of Technology, Tokyo (Titech), since 2001, leading the Problem Solving Environment Group that oversees the supercomputing needs of the Titech campus. He has been involved with software for large-scale cluster and grid computing since the mid-1990s, and he currently serves as one of the Principal Investigators of the Japanese National Research Grid Initiative (NAREGI) project, which aims to develop essential middleware and a large-scale testbed for a next-generation research infrastructure using grid technologies. He also serves in several editorial positions for international journals, including *Concurrency: Practice and Experience* and the *Journal of Grid Computing*.

Prof. Matsuoka has been program and general chair of several international conferences, including ACM OOPSLA'2002 and IEEE CCGrid 2003. He has been a Steering Group member and an Area Director of the Global Grid Forum since 1999, and he now cochairs the Grid Research Oversight Committee therein. He has won several awards, including the Sakai Award for research excellence from the Information Processing Society of Japan in 1999.

Sinji Shimojo (Member, IEEE) received the M.E. and D.E. degrees from Osaka University, Osaka, Japan, in 1983 and 1986, respectively.

He became an Assistant Professor with the Department of Information and Computer Sciences, Faculty of Engineering Science, Osaka University, in 1986, and he served as an Associate Professor with the university's Computation Center from 1991 to 1998. During that period, he also worked as a Visiting Researcher at the University of California, Irvine, for one year. He has been a Professor with Cybermedia Center (formerly the Computation Center) at Osaka University since 1998. His current research focuses on a wide variety of multimedia applications, peer-to-peer communication networks, ubiquitous network systems, and grid technologies.



Mutsumi Aoyagi received the M.S. degree from Keio University, Tokyo, Japan, in 1983 and the Ph.D. degree from Nagoya University, Nagoya, Japan, in 1987, respectively.

He became a Postdoctoral Fellow in the chemistry division of Argonne National Laboratory in 1990, and he moved to Institute for Molecular Science as an Associate Professor. He has been a Professor at Kyushu University, Fukuoka, Japan, since 2002. His current research focuses on a wide variety of computational science and computational molecular science, utilizing grid technologies.

Satoshi Sekiguchi (Member, IEEE) was born in 1959. He received the B.S. degree from the Department of Information Science, Faculty of Science, University of Tokyo, Tokyo, Japan, in 1982 and the M.Se. degree from the University of Tsukuba, Tsukuba, Japan, in 1984.

He joined the Electrotechnical Laboratory, Agency of Industrial Science and Technology, in 1984 to engage research in high-performance and parallel computing widely from the computer architecture, compiler, numerical algorithm, performance evaluation, as well as its applications. He served as the Deputy Director of the Research Institute of Information Technology, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, in 2001, and is currently the founding director of Grid Technology Research Center (GTRC), AIST. Since the dawn of the grid era, he has been one of technology and community leaders, who is in particular one of the Principal Investigators of the Ninf project since 1995 being developed as a reference implementation of the GridRPC model, the founder of the Asia Pacific Grid partnership (ApGrid), and the chair of the Japan Grid Consortium (JpGrid).

Dr. Sekiguchi is a Member of the Society for Industrial and Applied Mathematics (SIAM) and the Information Processing Society of Japan (IPSJ) and is a Chair of the SIGHPC. He also served as one of the steering committee members of the Global Grid Forum (GGF) till 2003 and is now a Member of the GGF Advisory Committee.

Hitohide Usami was born in Tokyo, Japan. He received the Ph.D. degree from the Graduate School of Engineering, Tohoku University, Sendai, Japan, in 1978.

He has worked for Fujitsu, Ltd. since graduating and now works temporarily with the National Institute of Informatics (NII) on the Japanese National Research Grid Initiative (NAREGI) project. His research interests are in knowledge-based systems, semantic grids, and problem-solving environments (PSEs).

Dr. Usami has been a Member of the Japan Society for Computational Engineering and Science, the Institute of Electronics, Information and Communication Engineers, and the Japan Society for Artificial Intelligence.

Kenichi Miura received the Ph.D. degree in computer science from the University of Illinois, Urbana-Champaign, in 1973.

He joined Fujitsu in 1973 and has since been engaged in high-end computing. From 1992 to 1997, he was Vice President and General Manager of the Supercomputer Group at Fujitsu America, Inc., where he was responsible for all supercomputer-related activities in the United States. He also served as a Visiting Professor at the Computer and Communications Center of Kyushu University from 1990 to 1993, and also as a Visiting Professor at the National Institute of Informatics (NII) in 2003. He is currently a Professor in High-End Computing at NII and the Project Leader of the Japanese National Research Grid Initiative (NAREGI) project. Since June 2002, he has also been a Fellow of Fujitsu Laboratories, Ltd. His research interests include grid computing, supercomputer architecture, vector and parallel numerical algorithms, and computational physics.