GridVM (Grid Virtual Machine)

**Overview**

A new grid middleware called GridVM is developed in the NAREGI project, which deploys an virtualized layer of computing resources in a grid environment and facilitates resource utilization, resource protection and fault tolerance.

**Features**

1. The GridVM abstracts the differences of local schedulers and provides common facilities for a large scale computation and resource utilization, which includes co-allocation and synchronous control such as gang scheduling for a parallel job over sites.

2. The GridVM protects computing resources from unauthorized accesses of external users by enables specified policies, which enables the confinement on file accesses and quantitative resource usage.

3. The GridVM facilitates the applicability of checkpoint/migration to various kinds of programs by virtualization of their accesses to an underlying operating system such as process IDs and network addresses.
Super Scheduler

**Overview**
A super-scheduling system is developed at the NAREGI project, which is for large-scale control and management of various resources that are shared by different organizations in a grid environment. The system will be primarily aimed at identifying resources that can meet requests from batch job users, and reserving/allocating these resources to specific jobs.

**Features**

1. **NAREGI Resource Schema**
   A model of various resources shared within the NAREGI grid has been built using the CIM (Common Information Model) approved by the DMTF (Distributed Management Task Force). The Super Scheduler uses this model for finding resources, allocating them to jobs, and monitoring and controlling them.

2. **Resource Broker**
   The resource broker collects information on the grid resources from the grid information system, including their present status, load condition, and future levels, and searches for available resources for particular requests.

3. **Job Execution Base**
   This employs UNICORE as one of the job executing environments in the Super Scheduler. With UNICORE, the resource broker starts with UNICORE NJS (Network Job Supervisor), receives given job requirements, and then if resource allocation is possible, it sends back the completed table of resources and information on NJS.
Grid Operation Supporting Software

Overview

A scalable and secure resource information management service will be established for the purpose of running a large-scale, multi-disciplined grid computing environment in the NAREGI Project.

Features

1. Collection, aggregation, and accumulation of information, required for proper operation of the system, such as information on resources, performance, users, budget and costs, as well as any system faults. The framework of common schema is leveraged to integrate heterogeneous computing resources in the grid environment.

2. Providing a resource information management service with scalability as well as access control functions for other resource management services and human systems based on OGSI and CIM.

3. Policy-based autonomous control, such as Account Mapping Service, for reducing system manager workload. This service will be investigated and established as sample implementation of an operation support service.
Grid Programming Environment

GridRPC (Remote Procedure Call)

- Programming Model using RPCs on a Grid
- The GridRPC API was proposed as a standard in the GGF GridRPC WG
- Ninf-G Version 2
  - A reference implementation of GridRPC API
  - Implemented on top of the Globus Toolkit
  - Provides C, Fortran, and Java APIs

GridMPI (Message Passing Interface)

GridMPI is a new MPI programming environment to run MPI applications efficiently over grids. GridMPI introduces the latency-aware communication topology layer, which optimizes the communication performance over links with non-uniform latency and bandwidth, and hides lower-level communication libraries.

The GridMPI demonstration using an emulated grid environment is shown in the AIST booth:739
**Visualization System**

A real-time and postprocessing visualization system for nano simulation, capable of reducing network loads that may interfere with smooth visualization, by means of flexible distribution of visualization tasks in a grid environment. It is also characterized in having functions of large scale parallel visualization, visualization for coupled simulation, and collaboration.

**GUI Workflow Tool**

A visual tool for preparing, submitting and querying distributed jobs in remote computing resources seamlessly. Workflow job is described by enhanced Workflow language based on grid service, that may be a common interface with other systems such as PSE.

**Problem Solving Environment**

A workflow based software platform for executing and coordinating the collaboration among simulation applications distributed in a grid environment. The simulation applications developed by researchers and engineers are incorporated into the grid via the facility provided by this software platform.
SuperSINET and Applications

SuperSINET is used for a network infrastructure in NAREGI project.

SuperSINET
- Research Project on advanced science and technologies
- 10Gbps Backbone
- GbE Bridges for peer-connection
- Operation of Photonic Cross Connect (PXC) for fiber / wavelength switching
- 6,000+km dark fiber, 300+Gb/s, including 60+e-e lambda
- Operational since January, 2002

SuperSINET Nodes
High Performance and Secure Network Infrastructure on Grid

The subjects include
- measurement, management, and control of underlying networks
- communication protocols optimized for Grid applications
- authentication mechanisms supporting secure Grid communications
to provide a fundamental middleware for the wide-area Grid computing.

Network Measurement, Management, and Control

These functions are essential for efficient use of network-internal resources on high-speed managed networks. A novel framework and its implementation of these functions adaptive to wide-area Grid computing environment are being developed. They are focusing on
- multi-points real-time measurement and prediction of fine-grained network-internal traffic and resource states, and
- dynamic bandwidth control and QoS routing based on both user policies and network measurement,
which allow the distributed computing performance to be fault-tolerant, predictable and globally optimal, in cooperation with the Super Scheduler.
Communication Protocol Infrastructure Development

A novel transport layer protocol for Grid computing is designed and implemented for cooperating with Grid middleware. Furthermore, a parallel network simulator for evaluating and designing a large-scale Grid network is developed.

Project Goals

1. Development of a communication protocol optimized for Grid
   - Support an ultra high-speed and wide-area network
   - Simply rely on end-to-end measurement information

2. Development of a parallel network simulator for a large-scale Grid network
   - Simulate a large-scale network (e.g., more than 10000 nodes)
   - Automatic collection of network information and traffic information

NAREGI Authentication Service (NAS)

Authentication services based on PKI are defined as NAREGI infrastructure assuming GT2 and GT3 operation. Certificate policy is defined in basic assurance level and Certificate Authority (CA)/Registration Authority (RA)/Validation Authority (VA) are developed to realize these services. NAS forms a single policy domain, and collaboration function is developed to realize uniform grid computing environment over heterogeneous policy domains.
Mediator-based Middleware for Multiscale Simulations

**Overview**

A new grid middleware in the form of mediators is developed that allows various kinds of Nano-simulations to be coupled efficiently for solving multiscale problems. The mediators provide high-level transparency in data communication between different discretization methods associated with model specific spatial and temporal scale based on physical requirements. A prototype system, in which RISM and FMO are coupled, is built for structure and functionality study of Nano-scale molecules immersed in solvent.

**Features**

1. Mediators provide semantic transformations to transfer physical values associated with discrete points in a unified way defined by correlative specifications such as in-sphere and first nearest neighbors.

2. The middleware is developed on MPICH-G2 and Globus so that coupled simulations are highly portable and can be executed simultaneously on different grid machines as well as LANs.

3. To wrap legacy codes as simulation components, library calls to mediators are provided for automatic process management and determination of processes to which communication is required.
Target of Each Group

Functional Nanomolecule Group
- Molecular Isomerization by Low-Energy Integration

Nano-Molecular Assembly Group
- Protein Folding

Nano-Magnetic System Group
- Super High-Density Magnetic Device Generated by Self-Organization of Surface
  - Applications:
    - Super High-Density (Tera-Bit) Memory
    - Super High-Resolution Magnetic Sensor
- Ferromagnetic Nano-Dot
- Arrangement-Control of Magnetic Nano-Dots by Self-Organization

Nano-System Designing Group
- Quantum Dot
- Quantum Wire
- Application: Half-Metal Magnetic Tunneling Element
- Manganese Oxide: Ferromagnetic Half-Metal
- Orbiton (Orbital Wave)
Real-time Flow Measurement System

The network research group at the NAREGI project is investigating and developing a fundamental middleware so that grid applications can utilize the network infrastructure efficiently. Our subgroup of the network research group focuses on dynamic bandwidth control and QoS routing based on both user policies and network measurement.

What is our demonstration?

1. Real-time flow measurement applicable to user defined flows
2. Architecture of multiple capturing devices applicable to fast networks
3. GUI imposing flow characteristics with a high resolution up to milliseconds

Real-time Flow Measurement System

Packets passing a switch are mirrored to a Distribution Device.
Top N-byte chunk of packet are collected and sent to one of Capturing Devices.
Every Capturing Device identifies flows from N-byte chunks by pattern matching. The flow definition (pattern) is advertised from Manager Device.

Manager Device maintains flow definitions and stores collected data counted by Capturing Devices.

CIM definition for specifying data
API for accessing Network Control & Management System based on OGSA Grid service interfaces
Future Research

Standardization Works
Our group is planning to contribute proposal drafts to the GGF/IETF working groups.

IP flow information exporting protocol
CIM definition for OGSA services
Relational Grid Monitoring Architecture by cooperating with DataGrid projects
Network Control System Based on User Policies

The network research group at the NAREGI project is investigating and developing a fundamental middleware so that grid applications can utilize the network infrastructure efficiently. Our subgroup of the network research group focuses on dynamic bandwidth control and QoS routing based on both user policies and network measurement.

What is our demonstration?

1. How to manage network-internal resources
2. How to make user policies regarding network-internal resources
3. How to control bandwidth and reroute dynamically

Multi-Points Real-time measurement Entity
Super-Scheduler
Future Research

High-speed Managed Networks

Network Resource Management Based on Virtual Organizations

Network Resource Control
- Bandwidth and Priority Control
- QoS Routing

USER

Application Program Interfaces
Policy Engine
Network Enforcer
Network Manager
Network Control & Management System

Web User Interfaces

Cooperation

Standardization Works
Our group is planning to contribute proposal drafts to the GGF working groups.

Policy Service Core of OGSA
Grid Advance Researevation API
OGSA fundamental Services: Autonomic Management Service (Provisioning)
Mediator-based Coupled Simulation Between RISM and FMO

(RISM: Reference Interaction Site Model, FMO: Fragment Molecular Orbital method)

1. Adaptive meshes of hydrate solvent are generated by mediators that search discretized meshes in RISM, using in-sphere correlation, inside the radius of van der Waals sphere centered at molecular sites of solute.

2. Pair correlation function of molecular sites between solvent and solute is obtained from RISM and electronic polarization and molecular structure of solute are calculated in FMO.

3. Solvent-solute pair distribution associated with the discretized meshes in RISM are transformed by mediators to the adaptive meshes employed in FMO using correlative specification of first nearest neighbors.

4. Electronic polarization of molecular sites due to the long-range Coulomb interaction according to the solvent distribution is obtained from FMO.

5. Our future plan targets on evaluation of the optimized solute structure immersed in infinite solvent in which variations of molecular structure and partial charge associated with molecular sites are calculated iteratively.
Computer System for Grid Software Infrastructure R & D (NII)

5Tflops, 700GB

Super SINET

High Perf. Distributed-memory type Compute Server 1
128 CPUs (Xeon, 3.06GHz) + Control Node
InfiniBand 4X(8Gbps)
Memory 130GB 9.4TB

High Perf. Distributed-memory type Compute Server 2
128 CPUs (Xeon, 3.06GHz) + Control Node
InfiniBand 4X(8Gbps)
Memory 65GB 9.4TB

Distributed-memory type Compute Server 1
128 CPUs (Xeon, 2.8GHz) + Control Node
Gb Ether (1Gbps)
Memory 65GB 4.7TB

Distributed-memory type Compute Server 2
128 CPUs (Xeon, 2.8GHz) + Control Node
Gb Ether (1Gbps)
Memory 65GB 4.7TB

Distributed-memory type Compute Server 3
128 CPUs (Xeon, 2.8GHz) + Control Node
Gb Ether (1Gbps)
Memory 65GB 4.7TB

Distributed-memory type Compute Server 4
128 CPUs (Xeon, 2.8GHz) + Control Node
Gb Ether (1Gbps)
Memory 65GB 4.7TB

Intra NW-A

L2 SW A

Intra NW-B

L2 SW A

L2 SW B

L2 SW B

L2 SW B

L2 SW B

Ext. NW

L3 SW 1Gbps
(upgradable to 10Gbps)
Intra NW

L3 SW 1Gbps
(upgradable to 10Gbps)

File Server
8CPUs (SPARC64V, 1.3GHz)
Memory 16GB
Storage 10TB
Back-up Max. 36.4TB

SMP type Compute Server 1
64 CPUs (SPARC64V, 1.3GHz)
Memory 128GB
Storage 441GB

SMP type Compute Server 2
32 CPUs (Itanium2, 1.3GHz)
Memory 32GB
Storage 180GB

SMP type Compute Server 3
32 CPUs (Power4, 1.3GHz)
Memory 64GB
Storage 480GB
Computer System for Grid Software Infrastructure R & D (IMS)

SMP type Computer
5.44 TFLOPS

16 ways × 50 nodes (POWER4+1.7GHz)
Multi-stage Crossbar Network

Memory 3072GB
Storage 6.8TB

Distributed-memory type Computer (4 units)
818 CPUs (Xeon, 3.06GHz) + Control Nodes
5.0 TFLOPS

Myrinet2000 (2Gbps)

L3 SW 1Gbps
(upgradable to 10Gbps)

Gb Ether (1Gbps)

Memory 1.6TB/unit
Storage 1.1TB/unit

× 6 Sites

File Server
16 CPUs (SPARC64 GP, 675MHz)

Memory 8GB
Storage 30TB
Back-up 25TB

Gb Ether (1Gbps)

Memory 64GB
Storage 1.2TB

User Side Cluster
32 CPUs (Xeon, 3.06GHz)
+ File SErvice Node

Super SINET