NAREGI the Japanese (Research) Grid Project

Satoshi Matsuoka
Professor, Global Scientific Information and Computing Center,
Deputy Director, NAREGI Project
Tokyo Institute of Technology / NII
http://www.naregi.org
**National Research Grid Infrastructure (NAREGI) 2003-2007 2010**

- **Petascale** Grid Infrastructure R&D for Future Deployment
  - > $120 mil total over 8 years
  - **Now part of Japanese 10 petascale computing initiative**
  - Hosted by National Institute of Informatics (NII)
  - PL: Ken Miura (Fujitsu→NII)
    - Sekiguchi(AIST), Matsuoka(Titech), Shimojo(Osaka-U), Aoyagi (Kyushu-U)...
  - Participation by multiple (≥ 3) vendors, Fujitsu, NEC, Hitachi, NTT, etc.
  - Follow and contribute to GGF Standardization, esp. OGSA

---

<table>
<thead>
<tr>
<th>Nanotech Grid Apps</th>
<th>(Biotech Grid Apps)</th>
<th>(Other Apps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“NanoGrid” IMS ~10TF</td>
<td>(BioGrid RIKEN)</td>
<td>Other Inst.</td>
</tr>
</tbody>
</table>

**Focused “Grand Challenge” Grid Apps Areas**

**National Research Grid Middleware R&D**

- Grid and Network Management
- Grid Middleware
- Grid R&D Infrastr. 15 TF–100TF

**SuperSINET**
The TSUBAME Supercomputing Cluster, Spring 2006

- **ClearSpeed CSX600 SIMD accelerator**
  - 360 boards
  - 35TeraFlops (Current)
  - 60TeraFlops (1 board per node)

- **Storage**
  - 1 Petabyte (Sun “Thumper”)
  - 0.1 Petabyte (NEC iStore)
  - Lustre FS, NFS (v4?)

- **NEC SX-8 Small Vector Nodes (under plan)**

- **Voltaire ISR9288 Infiniband**
  - 10Gbps x2 (xDDR) x ~700 Ports

- **Unified IB network**

- **Sun Galaxy 4 (Opteron Dual core 8-Way)**
  - 10480 core/655 Nodes
  - 50.4 TeraFlops
  - OS Linux (SuSE 9, 10)

- **NAREGI Grid MW**

- **10Gbps+External Network**
Titech TSUBAME

~80+ racks
350m² floor area
1.2 MW (peak)
Titech Supercomputing Grid 2006

- ~13,000 CPUs, 90 TeraFlops, ~26 TeraBytes Mem, ~1.1 Petabytes Disk
- CPU Cores: x86: TSUBAME (~10600), Campus Grid Cluster (~1000), COE-LKR cluster (~260), WinCCS (~300)
Scaling Towards Petaflops…

2010 Titech “PetaGrid”

=> Interim 200TeraFlops @ 2008
=> “Petascale” @ 2010
NORM for a typical Japanese center?

→ HPC Software is the key!

Earth Simulator 40TF (2002)
BlueGene/L 360TF (2005)
TSUBAME Upgrade >200TF (2008-2H)
KEK 59TF BG/L+SR11100
"Keisoku” >10PF(2011)

Titech Supercomputing Campus Grid (incl TSUBAME )~90TF (2006)

US HPCS (2010)
US Petascale (2007~8)
Chinese National Machine >100TF (2007~8)
Korean Machine >100TF (2006~7)

Titech Campus Grid

Next Gen “PetaGrid” 1PF (2010)
University Computer Centers (excl. National Labs) circa Spring 2006

10Gbps SuperSINET Interconnecting the Centers

~60 SC Centers in Japan

- 10 Petaflop center by 2011
What is a PetaScale (Research) Grid?

- A Grid Embodying Petaflops and Zetabytes of shared Resources, Inteconnected by networks, embodying millions of users
- Ubiquitously Accessible and Useable

- So that everyone can and will use it (researchers, students, engineers...), as they use the Internet
  - Next Gen CyberScience Infrastructure
PetaScale Grid Infrastructure
Assumptions

• As Massive and Sustainable Production Infrastructure
  - 100s of Institutions/Centers, Domain/Project VOs
  - > 100,000 users, > 100,000~1,000,000 CPUs
    • Machines very heterogeneous, SCs, clusters, desktops
  - 24/7 usage, production deployment, vendor supported
  - Server Grid, Data Grid, Metacomputing...

• Global Standards and Interoperability
  - Must be interoperable - Global Virtual Organizations
  - Global Standardization via GGF, IETF, OASIS

• Various Underlying Efforts and Infrastructures necessary
  - Global Identity and A&A availability, interoperability
  - High-performance networks
  - Professional R&D for production quality software
  - Support centers for hundreds of Virtual Organizations
Towards a Cyber-Science Infrastructure for R & D

Cyber-Science Infrastructure (CSI)

- NII-REO (Repository of Electronic Journals and Online Publications)
- Virtual Labs
- Live Collaborations
- NAREGI Middleware
- UPKI: National Research PKI Infrastructure
- Management Body / Education & Training
- SuperSINET and Beyond: Lambda-based Academic Networking Backbone
- Restructuring Univ. IT Research Resources
- Extensive On-Line Publications of Results

NAREGI Output

Industry/Societal Feedback

International Infrastructural Collaboration
## NAREGI Middleware Roadmap

|--------|--------|--------|--------|--------|

- **α Release (internal)**
  - OGSA(subset)/Unicore+GT2/3/3.9

- **β Release**
  - Development and Integration of β release
  - Evaluation on NAREGI Wide-area Testbed

- Development and deployment of β 1, 2, 3,

- **β 1 Release (public)**
  - OGSA/WSRF/GT4 GGF17

- Development of OGSA-based MW beta 1, 2,

- Verification & Evaluation Of Ver. 1

- Midpoint Evaluation
R&D in Grid Software and Networking Area (Work Packages)

• Work Package Structure (~150 FTEs):
  - Universities and National Labs: technology leadership
  - Vendors (Fujitsu, NEC, Hitachi, etc.): professional development

• WP-1: Resource Management:
  - Matsuoka(Titech), Nakada(AIST/Titech)

• WP-2: Programming Middleware:
  - Sekiguchi(AIST), Ishikawa(U-Tokyo), Tanaka(AIST)

• WP-3: Application Grid Tools:
  - Usami (new FY2005, NII), Kawata(Utsunomiya-u)

• WP-4: Data Management (new FY 2005, Beta):
  - Matsuda (Osaka-U)

• WP-5: Networking & Security
  - Shimojo(Osaka-u), Oie( Kyushu Tech.)

• WP-6: Grid-enabling Nanoscience Apps
  - Aoyagi(Kyushu-u)
NAREGI is/has/will...

• Is THE National Research Grid in Japan
  - Part of CSI and future Petascale initiatives
  - METI “Business Grid” counterpart 2003-2005

• Has extensive commitment to WS/GGF-OGSA
  - Entirely WS/Service Oriented Architecture
  - Set industry standards e.g. 1st impl. of OGSA-EMS

• Will work with EU/US/AP counterparts to realize a “global research grid”
  - Various talks have started, incl. SC05 interoperability meeting

• Will deliver first OS public beta in May 2006
  - To be distributed @ GGF17/GridWorld in Tokyo
NAREGI is not/doesn’t/won’t...

• Is NOT an academic research project
  - All professional developers from Fujitsu, NEC, Hitachi, (No students)
  - *IMPORTANT for Vendor buy-in and tech transfer*

• Will NOT develop all software by itself
  - Will rely on external components in some cases
  - Must be WS and other industry standards compliant
  - *IMPORTANT for Vendor buy-in and tech transfer*

• Will NOT hinder industry adoption at all costs
  - Intricate open source copyright and IP policies
  - We want companies to save/make money using NAREGI MW
  - *IMPORTANT for Vendor buy-in and tech transfer*
Standards: Setting & Adoption
Critical to Grid Adoption & Longevity

• Project lifetimes limited
• Most production centers, commercial entities, and users won’t commit their time & money unless longevity guaranteed
  - A few brave ones... (high-risk, high-return)
  - Success legacies overtold
• **Responsible Action** for major research grid infrastructural projects (e.g. NAREGI) to guarantee longevity beyond project lifetime
• Commercial Adoption by IT vendors: ISPs, ISVs, ASPs, etc. key for such longevity
  - But how do we achieve that?
Approach: “Bait and Catch”
--- Typical Japanese National Project Strategy ---

- IT vendors involved (seriously) from Day 1
- Public entity project benefactors should not solely expend its grants to sustain itself
  - Most money going to IT vendors to have them commit
- OTOH, vendors are later pressured by the government not to make direct profit
- (Exclusive IPR acquisition or) standardization used as bait for commercial adoption
  - Thus, even a research grid infrastructural project MUST FOCUS on standards, above all other priorities
- Then, standards & supporting technologies are DIRECTLY transferred to IT vendors
  - Productization and support => longevity => adoption!
A Picture is Worth a 1000 Words...
NAREGI Programming Models

- **High-Throughput Computing**
  - But with complex data exchange inbetween
  - NAREGI Workflow or GridRPC

- **Metacomputing (cross-machine parallel)**
  - Workflow (w/co-scheduling) + GridMPI
  - GridRPC (for task-parallel or task/data-parallel)

- **Coupled Multi-Resolution Simulation**
  - Workflow (w/co-scheduling) + GridMPI + Coupling Components
    - Mediator (coupled simulation framework)
    - GIANT (coupled simulation data exchange framework)
Nano-Science: coupled simulations on the Grid as the sole future for true scalability ... between Continuum & Quanta.

Material physics (Infinite system)
- Fluid dynamics
- Statistical physics
- Condensed matter theory

Molecular Science
- Quantum chemistry
- Molecular Orbital method
- Molecular Dynamics

Limit of Computing Capability
- Coordinates decoupled resources; Meta-computing,
  High throughput computing,
  Multi-Physics simulation w/ components and data from different groups within VO composed in real-time

The only way to achieve true scalability!

Old HPC environment:
- decoupled resources,
- limited users,
- special software, ...
Distributed Servers

LifeCycle of Grid Apps and Infrastructure

MetaComputing

Workflows and Coupled Apps / User

Many VO Users

HL Workflow NAREGI WFML

SuperScheduler

Dist. Grid Info Service

Application Contents Service

Place & register data on the Grid

Assign metadata to data

MetaData

Data 1

Data 2

Data n

Grid-wide Data Management Service (GridFS, Metadata, Staging, etc.)

GridRPC/Grid MPI

User Apps

GridVM

Distributed Servers

User Apps

GridVM

GridVM
NAREGI Software Stack (beta 1 2006)
- WS(RF) based (OGSA) SW Stack -

Grid-Enabled Nano-Applications (WP6)

Grid Programming (WP2)
- Grid RPC
- Grid MPI

Grid Visualization

Grid PSE

Grid Workflow (WFML (Unicore+ WF))

Super Scheduler

Distributed Information Service (CIM)

Grid VM (WP1)

SuperSINET

Grid Security and High-Performance Grid Networking (WP5)

(WSRF (GT4+Fujitsu WP1) + GT4 and other services)

Packaging

Data (WP4)

Computing Resources and Virtual Organizations

NII
IMS
Research Organizations
Major University Computing Centers
List of NAREGI “Standards”
(beta 1 and beyond)

- GGF Standards and Pseudo-standard Activities set/employed by NAREGI
  - GGF “OGSA CIM profile”
  - GGF AuthZ
  - GGF DAIS
  - GGF GFS (Grid Filesystems)
  - GGF Grid CP (GGF CAOPs)
  - GGF GridFTP
  - GGF GridRPC API (as Ninf-G2/G4)
  - GGF JSDL
  - GGF OGSA-BES
  - GGF OGSA-Byte-IO
  - GGF OGSA-DAI
  - GGF OGSA-EMS
  - GGF OGSA-RSS
  - GGF RUS
  - GGF SRM (planned for beta 2)
  - GGF UR
  - GGF WS-I RUS
  - GGF ACS
  - GGF CDDLM

- Other Industry Standards Employed by NAREGI
  - ANSI/ISO SQL
  - DMTF CIM
  - IETF OCSP/XKMS
  - MPI 2.0
  - OASIS SAML2.0
  - OASIS WS-Agreement
  - OASIS WS-BPEL
  - OASIS WSRF2.0
  - OASIS XACML

- De Facto Standards / Commonly Used Software Platforms Employed by NAREGI
  - Ganglia
  - GFarm 1.1
  - Globus 4 GRAM
  - Globus 4 GSI
  - Globus 4 WSRF (Also Fujitsu WSRF for C binding)
  - IMPI (as GridMPI)
  - Linux (RH8/9 etc.), Solaris (8/9/10), AIX, ...
  - MyProxy
  - OpenMPI
  - Tomcat (and associated WS/XML standards)
  - Unicore WF (as NAREGI WFML)
  - VOMS

Implement “Specs” early even if nascent if seemingly viable

Necessary for Longevity and Vendor Buy-In

Metric of WP Evaluation
Highlights of NAREGI Beta (May 2006, GGF17/GridWorld)

• Professionally developed and tested
• “Full” OGSA-EMS incarnation
  - Full C-based WSRF engine (Java -> Globus 4)
  - OGSA-EMS/RSS WSRF components
  - GGF JSDL1.0-extension job submission, authorization, etc.
  - Support for more OSes (AIX, Solaris, etc.) and BQs
• Sophisticated VO support for identity/security/monitoring/accounting (extensions of VOMS/MyProxy, WS-* adoption)
• WS- Application Deployment Support via GGF-ACS
• Comprehensive Data management w/Grid-wide FS
• Complex workflow (NAREGI-WFML) for various coupled simulations
• Overall stability/speed/functional improvements
• To be interoperable with EGEE, TeraGrid, etc. (beta2)
• Release next week at GGF17, press conferences, etc.
Ninf-G: A Reference Implementation of the GGF GridRPC API

- **What is GridRPC?**
  - Programming model using RPCs on a Grid
  - Provide easy and simple programming interface
  - The GridRPC API is published as a proposed recommendation (GFD-R.P 52)

- **What is Ninf-G?**
  - A reference implementation of the standard GridRPC API
  - Built on the Globus Toolkit
  - Now in NMI Release 8 (first non-US software in NMI)

- **Easy three steps to make your program Grid aware**
  - Write IDL file that specifies interface of your library
  - Compile it with an IDL compiler called ng_gen
  - Modify your client program to use GridRPC API

![Diagram showing call remote procedures, notify results, and call remote libraries](image)
Architecture of Ninf-G

Client side

Client
- Interface Request
- Interface Reply
- Invoke
- Executable
- Connect back

Server side

IDL file
- Numerical Library

IDL Compiler

Generate
- Remote Library Executable

MDS2/4

WS/ Pre-WS GRAM
- NAREGI-SS

Globus-IO

Interface Information
- LDIF File

retrieve
**GridMPI**

MPI applications run on the Grid environment

- **Metropolitan area, high-bandwidth environment**: \( \geq 10 \text{ Gpbs}, \leq 500 \text{ miles} \) (smaller than 10ms one-way latency)
  - Parallel Computation

- **Larger than metropolitan area**
  - MPI-IO

---

**computing resource**
- site A
- site B

**Wide-area Network**

**Single (monolithic) MPI application over the Grid environment**
GridMPI Software Architecture and Standards

- **MPI 2.0** (*test suite compliant*)
- **IMPI** (*Interoperable MPI*)
  - The original IMPI is defined only for the MPI-1.2 feature
  - Extension for MPI-2
- Porting the extended IMPI protocol to Open MPI
- Planning to submit the protocol to NIST
Grid Application Environment (WP3)

Gateway Services

Deployment Service
Application Contents Service
Application Repository

Workflow Service
File/Execution Manager
NAREGI-WFML
JM I/F module

Workflow GUI

Grid Workflow

Bio VO
Portal GUI

NAREGI Portal
Portal GUI

Nano VO
Registration GUI

VOMS
MyProxy

Underlying Grid Services

File Transfer (RFT)
Grid File System

Core Grid Services

Workflow Engine & Super Scheduler
Distributed Information Service

CFD Visualization Service
Molecular Visualization Service
Parallel Visualization Service

Gateway Services

NAREGI-Workflow Language
BPEL+JSDL

WSRF

WSIF

...
WP-3: User-Level Grid Tools & PSE

- **Grid PSE**
  - Deployment of applications on the Grid
  - Support for execution of deployed applications

- **Grid Workflow**
  - Workflow language independent of specific Grid middleware
  - GUI in task-flow representation

- **Grid Visualization**
  - Remote visualization of massive data distributed over the Grid
  - General Grid services for visualization
The NAREGI SSS Architecture (2007/3)

PETABUS (Peta Application services Bus)

Application Specific Service
Application Specific Service
Application Specific Service

WESBUS (Workflow Execution Services Bus)

JM
EPS
CSG
BPEL Interpreter Service

CESBUS (Coallocation Execution Services Bus; a.k.a. BES+ Bus)

FTS-SC
GRAM-SC
UniGridS-SC
AGG-SC with RS (Aggregate SCs)

BESBUS (Basic Execution Services Bus)

Globus
WS-GRAM I/F
(with reservation)

UniGridS
Atomic Services
(with reservation)

GridVM

Grid Resource

Grid-Middleware

NAREGI SSS

Network
Meta computing scheduler is required to allocate and to execute jobs on multiple sites simultaneously.

The super scheduler negotiates with local RSs on job execution time and reserves resources which can execute the jobs simultaneously.
NAREGI Info Service (beta) Architecture

- CIMOM Service classifies info according to CIM based schema.
- The info is aggregated and accumulated in RDBs hierarchically.
- Client library utilizes OGSA-DAI client toolkit.
- Accounting info is accessed through RUS.

CIMOM Service classifies info according to CIM based schema.
- The info is aggregated and accumulated in RDBs hierarchically.
- Client library utilizes OGSA-DAI client toolkit.
- Accounting info is accessed through RUS.
NAREGI IS: Standards Employed in the Architecture

- **OGSA-DAI WSRF 2.1**
- **Aggregator Service**
- **CIM Schema 2.10**
  - with extension
- **Light-weight CIMOM Service**
- **RDB**
- **CIM Providers**
  - OS
  - Processor
  - File System
  - Job Queue
  - Performance
- **Grid VM**
  - Ganglia
- **NodeA**
  - Node B
  - Node C
- **Cell Domain Information Service**
- **User Admin.**
- **Viewer**
- **Client (OGSA-RSS etc.)**
- **OGSA-DAI Client toolkit**
- **GridVM (Chargeable Service)**
- **RUS::insertURs**
- **GGF/UR**
- **GTF 4.0.1**
- **Distributed Information Service**
- **Client (OGSA-BES etc.)**
GridVM Features

✓ Platform independence as OGSA-EMS SC
  • WSRF OGSA-EMS Service Container interface for heterogeneous platforms and local schedulers
  • “Extends” Globus4 WS-GRAM
  • Job submission using JSDL
  • Job accounting using UR/RUS
  • CIM provider for resource information

✓ Meta-computing and Coupled Applications
  • Advanced reservation for co-Allocation

✓ Site Autonomy
  • WS-Agreement based job execution (beta 2)
  • XACML-based access control of resource usage

✓ Virtual Organization (VO) Management
  • Access control and job accounting based on VOs (VOMS & GGF-UR)
NAREGI GridVM (beta) Architecture

✓ Virtual execution environment on each site
  • Virtualization of heterogeneous resources
  • Resource and job management services with unified I/F
NAREGI GridVM: Standards Employed in the Architecture

- Super Scheduler
- Information Service
- GT4 GRAM-integration and WSRF-based extension services
- Job submission based on JSDL and NAREGI extensions
- CIM-based resource info. provider
- UR/RUS-based job accounting
- xacml-like access control policy
GT4 GRAM-GridVM Integration

✓ Integrated as an extension module to GT4 GRAM
✓ Aim to make the both functionalities available

SS → globusrun RSL+JSDL’ → GRAM services

Delegate → Delegation → RFT File Transfer

Delegate → Delegation → GRAM services

Extension Service

GridVM Job Factory

GridVMJob

GridVM scheduler → Local scheduler → GridVM Engine

Scheduler Event Generator

Basic job management + Authentication, Authorization

SS
NAREGI Data Grid beta1 Architecture (WP4)

Grid Workflow

Job 1 → Job 2 → ... → Job n

Data Grid Components

Data Access Management
- Import data into workflow

Metadata Management
- Place & register data on the Grid
- Assign metadata to data

Data Resource Management
- Store data into distributed file nodes

Grid-wide Data Sharing Service

Data 1 → Metadata → Data 1
Data 2 → Metadata → Data 2
... → Metadata → Data n

Grid-wide File System

NAREGI Data Grid beta1 Architecture (WP4)
NAREGI WP4: Standards Employed in the Architecture

Data Access Management
- Workflow (NAREGI WFML => BPEL+JSDL)

Job 1 \(\rightarrow\) Job \(n\)
- Data 1
- Data \(n\)

Job 1
- Import data into workflow

Tomcat 5.0.28
- Place data on the Grid

Globus Toolkit 4.0.1
- Data Resource Management

Data Resource Management
- OGSA-DAI WSRF2.0
- PostgreSQL 8.0
- Gfarm 1.2 PL4 (Grid FS)

Metadata Construction
- OGSA-DAI
- WSRF2.0
- PostgreSQL 8.0
- Data Specific Metadata DB

Metadata Construction
- GridFTP

Super Scheduler (SS) (OGSA-RSS)
- Data Staging

OGSA-RSS FTS SC
- Computational Nodes

GGF-SRM (beta2)
- PostgresQL 8.0

Workflow
- NAREGI WFML
- BPEL+JSDL

Job 1
- Data 1

Job \(n\)
- Data \(n\)

Filesystem Nodes
- Computational Nodes

Data Staging
- Import data into workflow
NAREGI Application Mediator (WP6) for Coupled Applications

Mediator Components
Support data exchange between coupled simulation

Data transfer management
- Synchronized file transfer
- Multiple protocol GridFTP/MPI

Data transformation management
- Semantic transformation libraries for different simulations
- Coupled accelerator

*SBC: Storage-based communication

Workflow
NAREGI WFT

co-allocated jobs

Simulation A
Mediator
Simulation B

Super Scheduler

GridVM

Information Service
SQL

OGSA-DAI
WSRF2.0

GridFTP
MPI

JNI

Globus Toolkit 4.0.1

MPI
NAREGI Phase 1 Testbed
Dedicated Testbed
No “ballooning” w/production resources
Production to reach > 100Teraflops

~3000 CPUs
~17 Tflops

TiTech
Campus Grid

Osaka Univ.
BioGrid

Tohoku Univ.
Small Test App
Clusters

AIST
SuperCluster

Kyushu Univ.
Small Test App Clusters

Kyoto Univ.
Small Test App
Clusters

AIST
Small Test App
Clusters

KEK
Small Test App
Clusters

ISSP
Small Test App
Clusters

Super-SINET
(10Gbps MPLS)

Computational
Nano-science Center(IMS)
~10 Tflops

Center for GRID R&D
(NII)
~5 Tflops
Computer System for Grid Software Infrastructure R & D
Center for Grid Research and Development (5Tflops, 700GB)

**File Server**
- (PRIMEPOWER 900 + ETERNUS3000 + ETERNUS LT160)
- 1node / 8CPU (SPARC64V1.3GHz)
  - Memory: 16GB
  - Storage: 10TB
  - Back-up: Max. 36.4TB

**SMP type Compute Server**
- (PRIMEPOWER HPC2500)
  - 1node (UNIX, SPARC64V1.3GHz/64CPU)
    - Memory: 128GB
    - Storage: 441GB

- (SGI Altix3700)
  - 1node (Itanium2 1.3GHz/32CPU)
    - Memory: 32GB
    - Storage: 180GB

- (IBM pSeries690)
  - 1node (Power4 1.3GHz/32CPU)
    - Memory: 64GB
    - Storage: 480GB

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

**Intra NW-A**
- L3 SW 1Gbps
- (Upgradable to 10Gbps)

**Intra NW-B**
- L3 SW 1Gbps
- (Upgradable to 10Gbps)

**Ext. NW**
- L3 SW 1Gbps

**High Perf. Distributed-memory type Compute Server**
- (PRIMERGY RX200)
  - 128 CPUs (Xeon, 3.06GHz) + Control Node
    - Memory: 130GB
    - Storage: 9.4TB
    - InfiniBand 4X (8Gbps)

- (Express 5800)
  - 128 CPUs (Xeon, 2.8GHz) + Control Node
    - Memory: 65GB
    - Storage: 4.7TB
    - GbE (1Gbps)

- (HPC LinuxNetworx)
  - 128 CPUs (Xeon, 2.8GHz) + Control Node
    - Memory: 65GB
    - Storage: 4.7TB
    - GbE (1Gbps)

**SuperSINET**

**Ext. NW**
- 1Gbps
- (upgradable to 10Gbps)

**L3 SW**
- 1Gbps
- (Upgradable to 10Gbps)

**Memory**
- 130GB
- 65GB
- 64GB
- 128GB
- 32GB
- 64GB
- 128GB

**Storage**
- 9.4TB
- 4.7TB
- 180GB
- 441GB
- 128GB
- 32GB
- 480GB
- 441GB

**Networks**
- Intra NW-A
- Intra NW-B
- Ext. NW
Computer System for Nano Application R & D
Computational Nano science Center (10Tflops, 5TB)

SMP type Computer Server
5.4 TFLOPS
16ways-50nodes (POWER4+ 1.7GHz)
Multi-stage Crossbar Network

Distributed-memory type
Computer Server (4 units) 5.0 TFLOPS
818 CPUs (Xeon, 3.06GHz) + Control Nodes
Myrinet2000 (2Gbps)

File Server
16 CPUs (SPARC64 GP, 675MHz)
Memory 8GB
Storage 30TB
Back-up 25TB

L3 SW
1Gbps
(Upgradable to 10Gbps)

Front-end Server

CA/RA Server

Startegic Information and Communications R & D Promotion (SCOPE)

SuperSINET

Center for Grid R & D

VPN

Firewall
The Ideal World: Ubiquitous VO & user management for international e-Science

Europe: EGEE, UK e-Science, ...

US: TeraGrid, OSG,

Japan: NII CyberScience (w/NAREGI), ...
Other Asian Efforts (GFK, China Grid, etc.)...

Grid Regional Infrastructural Efforts

Collaborative talks on PMA, etc.

Different software stacks but interoperable

Standardization, commonality in software platforms will realize this
From Interoperation to Interoperability
GGF16 “Grid Interoperations Now”

Charlie Catlett
Director, NSF TeraGrid
on Vacation

Satoshi Matsuoka
Sub Project Director, NAREGI Project
Tokyo Institute of Technology / NII
Interoperation Activities

- The GGF GIN (Grid Interoperations Now) effort
  - Real interoperation between major Grid projects
  - Four interoperation areas identified
    - Security, Data Mgmt, Information Service, Job Submission (not scheduling)

- EGEE/gLite - NAREGI interoperation
  - Based on the four GIN areas
  - Several discussions, including 3 day meeting at CERN mid March, email exchanges

- Updates at GGF17 Tokyo next week
- Some details in my talk tomorrow
The Ideal World: Ubiquitous VO & user management for international e-Science

Europe: EGEE, UK e-Science, ...

US: TeraGrid, OSG, ...

Japan: NII CyberScience (w/NAREGI), ...
Other Asian Efforts (GFK, China Grid, etc.)...

Different software stacks but interoperable

Standardization, commonality in software platforms will realize this

Grid Regional Infrastructural Efforts
Collaborative talks on PMA, etc.
Last Time We Met

• “Agreeing to Agree on what needs to be Agreed first”
  - Essential for interoperability, difficult to change later on

• Identified 4 Essential Key Common Services
  - Authentication, Authorization, Identity Management
    • Individuals, communities (VO’s)
  - Jobs: submission, auditing, tracking
    • Job submission interface, job description language, etc.
  - Data Management
    • Data movement, remote access, filesystems, metadata mgmt
  - Resource discovery and Information Service
    • Resource description schema, information services
“Interoperation” versus “Interoperability”

• **Interoperability**
  “The ability of software and hardware on multiple machines from multiple vendors to communicate”
  - Based on commonly agreed documented specifications and procedures

• **Interoperation**
  “Just make it work together”
  - Whatever it takes, could be ad-hoc, undocumented, fragile
  - Low hanging fruit, future interoperability
Interoperation Status

- GIN meetings GGF16 and GGF17
- 3-day meeting at CERN end of March
- Security
  - Common VOMS/GSI infrastructure
  - NAREGI more complicated use of GSI/Myproxy and proxy delegation but should be OK
- Data
  - SRM commonality and data catalog integration
  - GFarm and DCache consolidation
- Information Service
  - CIM vs. GLUE schema differences
  - Monitoring system differences fairly
  - Schema translation (see next slides)
- Job Submission
  - JDL vs. JSDL, Condor-C/CE vs. OGSA SS/SC-VM architectural differences, etc.
  - Simple job submission only (see next slides)
Information Service Characteristics

- **Basic syntax:**
  - Resource description schemas (e.g., GLUE, CIM)
  - Data representations (e.g., XML, LDIF)
  - Query languages (e.g., SQL, XPath)
  - Client query interfaces (e.g., WS Resource Properties queries, LDAP, OGSA-DAI)

- **Semantics:**
  - What pieces of data are needed by each Grid (various previous works & actual deployment experiences already)

- **Implementation:**
  - Information service software systems (e.g., MDS, BDII)
  - The ultimate sources of this information (e.g., PBS, Condor, Ganglia, WS-GRAM, GridVM, various grid monitoring systems, etc.).
An implementation of the GGF Grid Monitoring Architecture (GMA)

All data modelled as tables: a single schema gives the impression of one virtual database for VO
<table>
<thead>
<tr>
<th>Grid</th>
<th>Schema</th>
<th>Data</th>
<th>Query Lang</th>
<th>Client IF</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tera-Grid</td>
<td>GLUE</td>
<td>XML</td>
<td>XPath</td>
<td>WSRF RP Queries</td>
<td>MDS4</td>
</tr>
<tr>
<td>OSG</td>
<td>GLUE</td>
<td>LDIF</td>
<td>LDAP</td>
<td>LDAP</td>
<td>BDII</td>
</tr>
<tr>
<td>NAREGI</td>
<td>CIM 2.10+ext</td>
<td>Relational</td>
<td>SQL</td>
<td>OGSA-DAI WS-I RUS</td>
<td>CIMOM + OGSA-DAI</td>
</tr>
<tr>
<td>EGEE/LCG</td>
<td>GLUE</td>
<td>LDIF</td>
<td>LDAP</td>
<td>LDAP</td>
<td>BDII</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relational</td>
<td>R-GMA i/f</td>
<td>R-GMA</td>
</tr>
<tr>
<td>Nordu Grid</td>
<td>ARC</td>
<td>LDIF</td>
<td>LDAP</td>
<td>LDAP</td>
<td>GIIS</td>
</tr>
</tbody>
</table>
Low Hanging Fruit
“Just make it work by GLUEing”

- Identify the minimum common set of information required for interoperation in the respective information service
- Employ GLUE and extended CIM as the base schema for respective grids
- Each Info service in grid acts as a information provider for the other
- Embed schema translator to perform schema conversion
- Present data in a common fashion on each grid; WebMDS, NAREGI CIM Viewer, SCMSWeb, ...

Minimal Common Attributes

- Define minimal common set of attributes required
- Each system component in the grid will only access the translated information
GLUE→CIM translation

• Development of information providers with translation from GLUE data model to CIM about selected common attributes such as up/down status of grid services.

SQL “SELECT”

Publish Tuples
SQL “CREATE TABLE”
SQL “INSERT”

GLUE-CIM mapping; selected Minimal Attributes

Development of information providers with translation from GLUE data model to CIM about selected common attributes such as up/down status of grid services.
Interoperability: NAREGI Short Term Policy

• gLite
  - Simple/Single Job (up to SPMD)
  - Bi-Directional Submission
    • NAREGI \rightarrow gLite: GT2-GRAM
    • gLite \rightarrow NAREGI: Condor-C
  - Exchange Resource Information

• GIN
  - Simple/Single Job (up to SPMD)
  - NAREGI \rightarrow GIN Submission
  - WS-GRAM
  - Exchange Resource Information

• BES/ESI
  - TBD
    • Status somewhat Confusing
    • ESI middleware already developed?
      - Globus 4.X and/or UnicoreGS?
## Job Submission Standards

**Comparison: Goals**

<table>
<thead>
<tr>
<th>Feature</th>
<th>ESI</th>
<th>BES</th>
<th>NAREGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use JSDL</td>
<td>✔️*1</td>
<td>✔</td>
<td>✔️*1</td>
</tr>
<tr>
<td>WSRF OGSA Base Profile 1.0 Platform</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Job Management Service</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Extensible Support for Resource Models</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Reliability</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Use WS-RF modeling conventions</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Use WS-Agreement</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Advance reservation</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Bulk operations</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generic management frameworks (WSDM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Define alternative renderings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server-side workflow</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>

*1: Extended
## Job Factory Operations

<table>
<thead>
<tr>
<th></th>
<th>ESI (0.6)</th>
<th>BES (Draft v16)</th>
<th>NAREGI (β)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
<td>CreateManagedJob (there is a subscribe option)</td>
<td>CreateActivityFromJSDL</td>
<td>MakeReservations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GetActivityStatus</td>
<td>CommitReservations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RequestActivityStateChanges</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>StopAcceptingNewActivities</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>StartAcceptingNewActivities</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IsAcceptingNewActivities</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GetActivityJSDLDocuments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GetMultipleResourceProperties</td>
<td>GetMultipleResourceProperties</td>
<td></td>
</tr>
<tr>
<td></td>
<td>QueryResourceProperties</td>
<td>QueryResourceProperties</td>
<td></td>
</tr>
<tr>
<td><strong>WS-ResourceLifeTime</strong></td>
<td>ImmediateResourceDestruction</td>
<td>Destroy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ScheduledResourceDestruction</td>
<td>SetTerminationTime</td>
<td></td>
</tr>
<tr>
<td><strong>WS-BaseNotification</strong></td>
<td>NotificationProducer</td>
<td></td>
<td>Notify</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subscribe</td>
<td></td>
</tr>
</tbody>
</table>

NAREGI (β) = BES (Draft v16) = ESI (0.6)
Interoperation: gLite-NAREGI

- gLite-IS [GLUE]
- gLite-WMS [JDL]
- gLite-CE
- GT2-GRAM
- Interop-SC

- NAREGI-IS [CIM]
- NAREGI-SS [JSDL]
- NAREGI-SC
- NAREGI Portal
- NAREGI GridVM

- IS bridge
- gLite to NAREGI bridge
- gLite-NAREGI bridge

- NAREGI-WF generation
- Job Submit./Ctrl.
- Status propagation
- Certification?

- Job Submit./Ctrl.
- Status propagation

- GLUE ↔ CIM

- ClassAd → JSDL
- NAREGI-WF generation
- Job Submit./Ctrl.
- Status propagation
- Certification?
Interoperation: GIN (Short Term)

- anotherGrid-IS [CIM or GLUE] → IS bridge → NAREGI-IS [CIM]
- GLUE ↔ CIM
- GLUE ↔ anotherGrid-IS [CIM or GLUE]

- anotherGrid [JSDL]
- GT4

- NAREGI Portal
- WS-GRAM
- anotherGrid-CE
- NAREGI GridVM

- NAREGI-SS [JSDL]
- Interop-SC

- Job Submit./Ctrl.
- Status propagation

Sorry !!
One way now
Cyber Science Infrastructure toward Petascale Computing (planned 2006-2011)

Cyber-Science Infrastructure (CSI)
(ITU Infra. for Academic Research and Education)

Contents

International Collaboration
- EGEE
- UNIGRIDS
- Teragrid
- GGF etc.

NII Collaborative Operation Center

Cyber-Science Infrastructure (CSI)

R&D Collaboration

Middleware
Operation/Maintenance

Networking

Domain Specific VO
(e.g ITBL)

Project-oriented VO

NAREGI Site

Joint Project
AIST

Nano Proof of al.Concept
Eval.

IMS

Delivery

Univ./National Supercomputing
VO

Peta-scale System
VO

Industri
Project

Core Site

Customization
Operation/Maintenance

Joint Project

R&D Collaboration

Feedback

Delivery

Feedback

Feedback

Core Site

Networking Infrastructure (Super-SINET)

Note: names of VO are tentative)
Summary of NAREGI Project

NAREGI V1.0

Grid Middleware for Large Computer Centers
Productization of General-purpose Grid Middleware for Scientific Computing
Personnel Training (IT and Application Engineers)
Contribution to International Scientific Community and Standardization
Progress in the Latest Research and Development (Nano, Biotechnology)
Use in Industry (New Intellectual Product Development)

MEXT: Ministry of Education, Culture, Sports, Science and Technology
Thank you!