NAREGI and CSI (and UNICORE): The Japanese National Research Grid & Networking Infrastructure

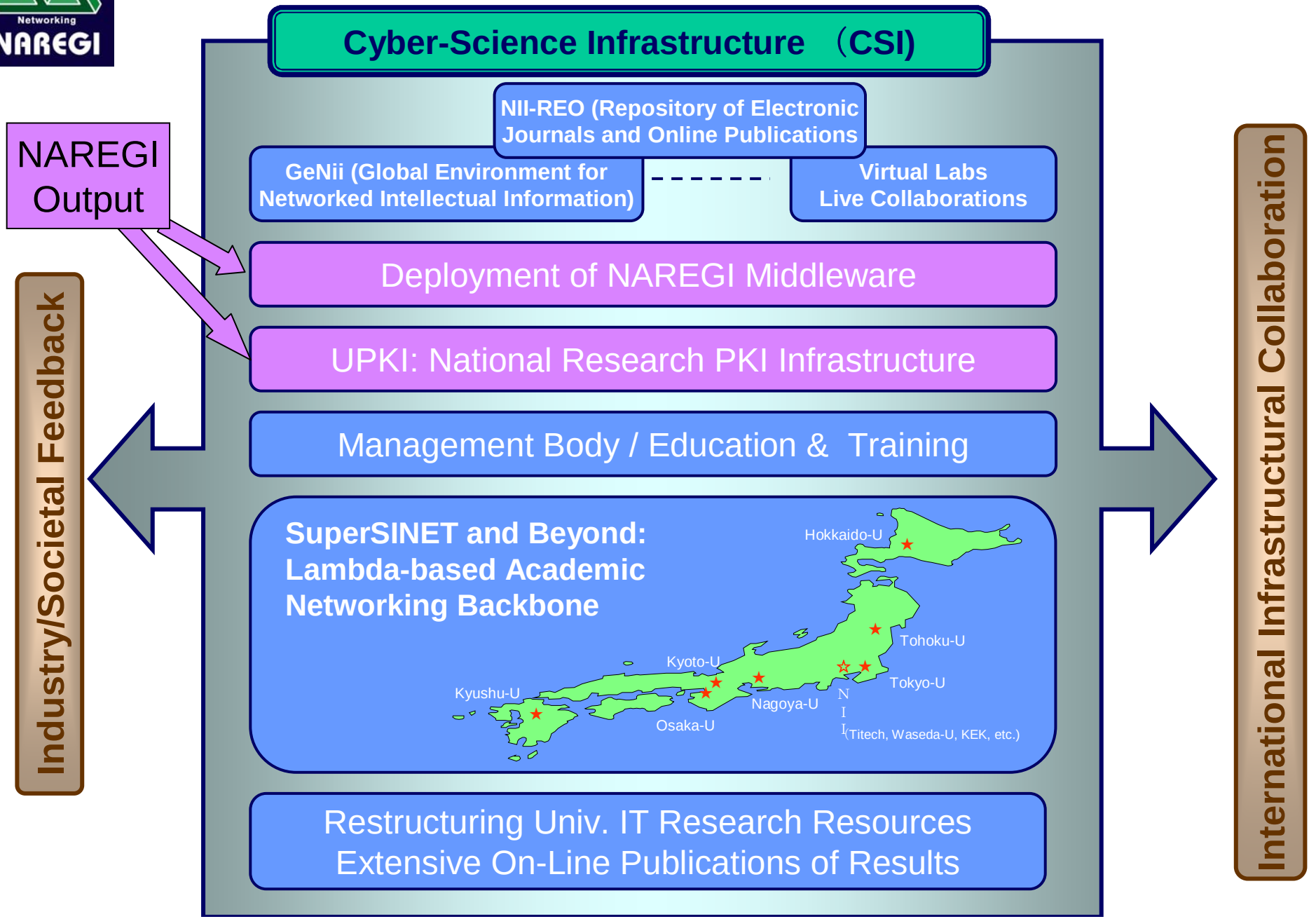
Satoshi Matsuoka

Professor, Global Scientific Information and Computing Center,

Deputy Director, NAREGI Project

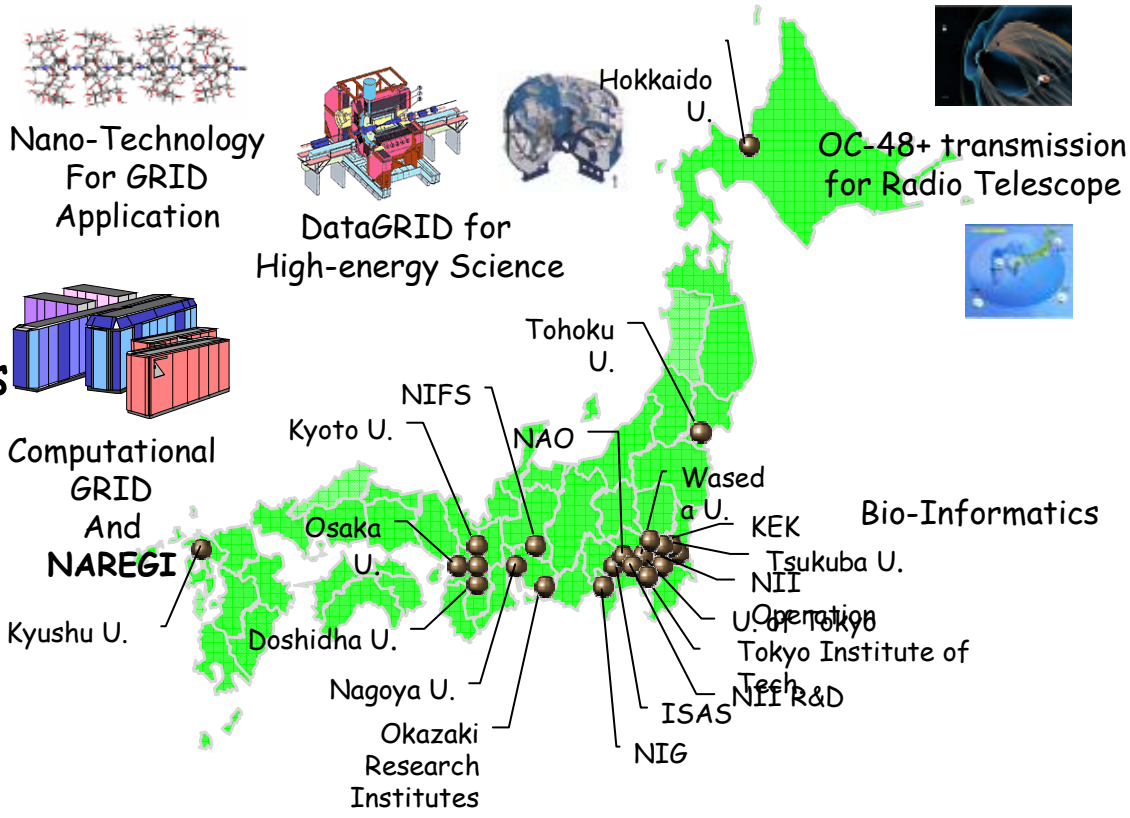
Tokyo Institute of Technology / NII

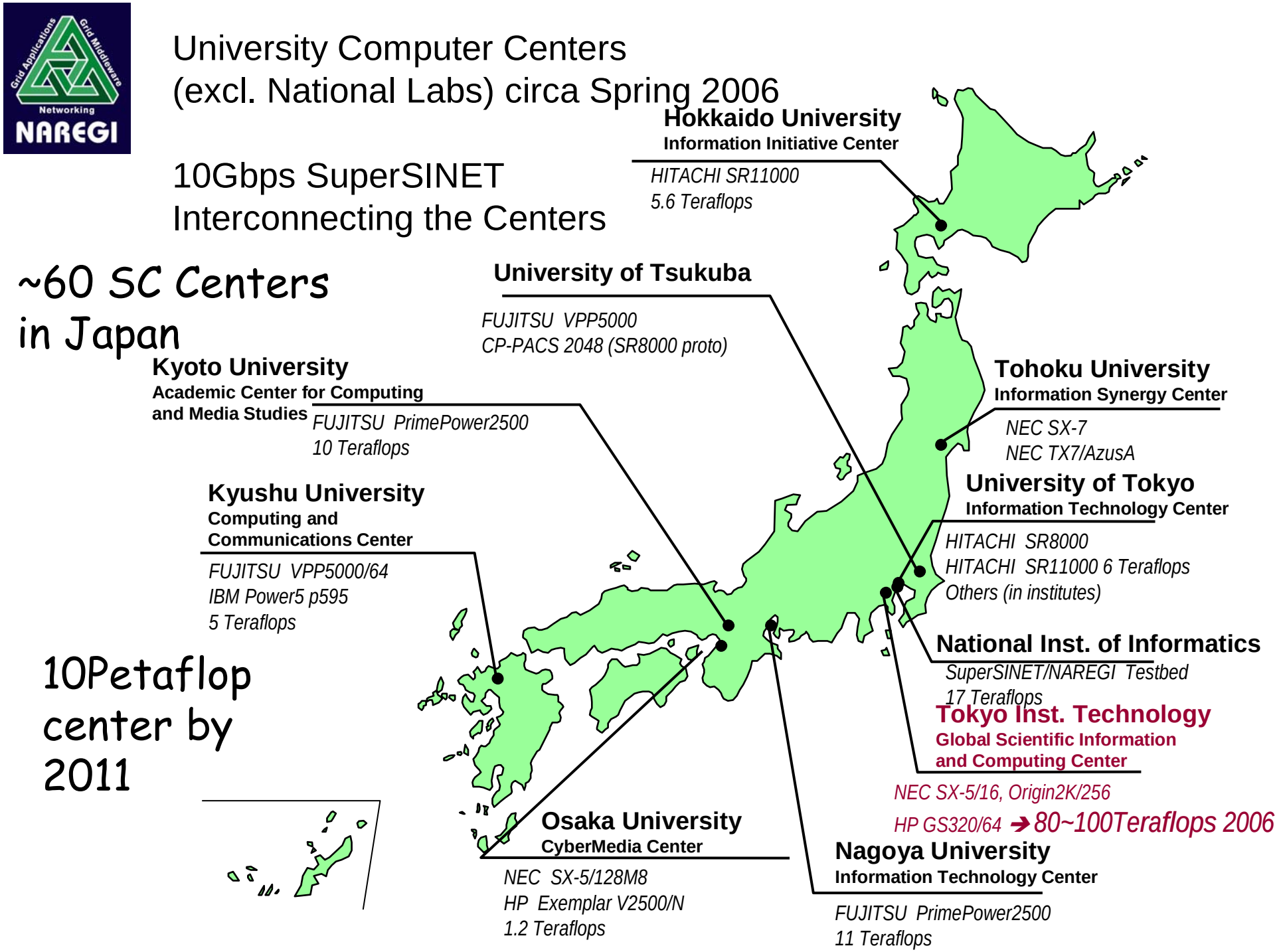
Towards a Cyber-Science Infrastructure for R & D



SuperSINET: All Optical Production Research Network in Japan

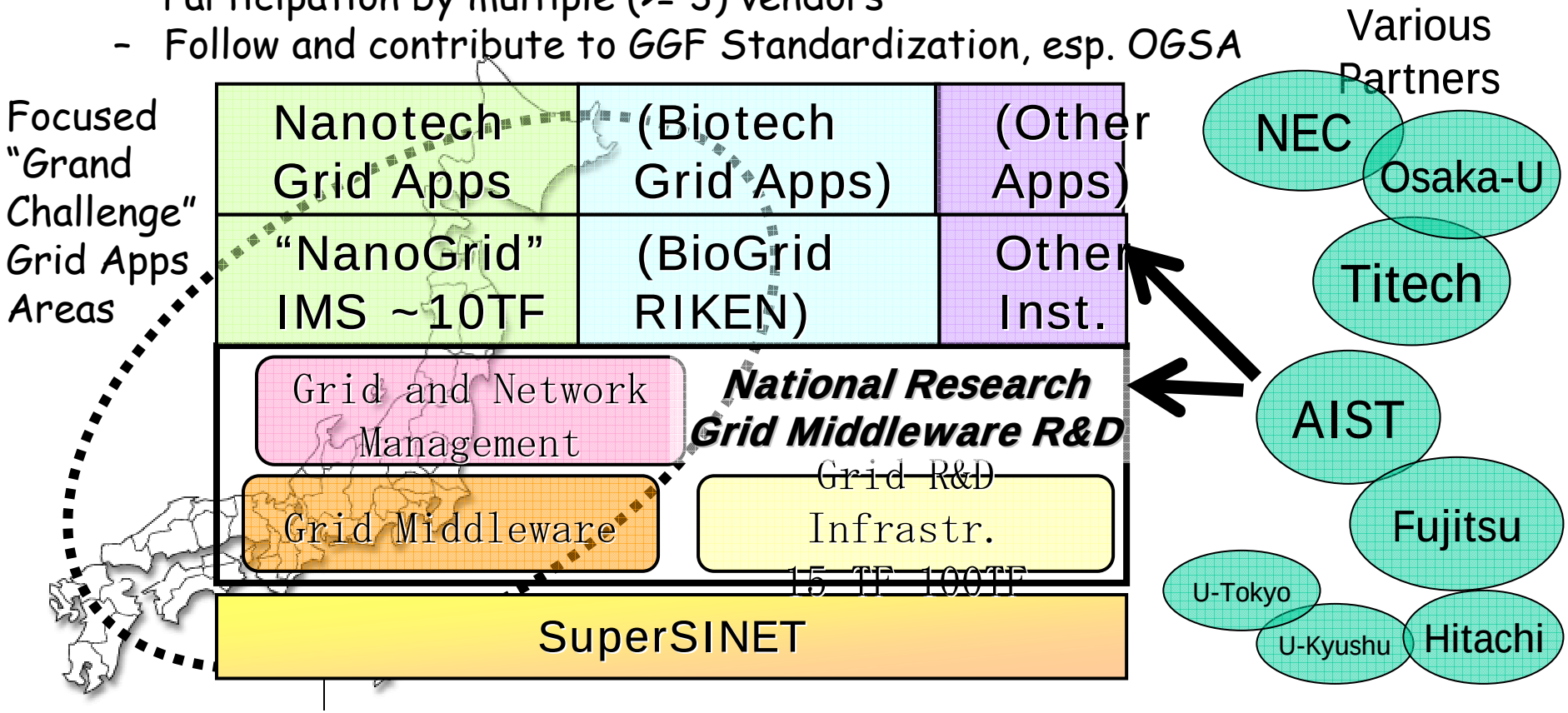
- 10Gbps Photonic Backbone
- GbEther Bridges for peer-connection
- Very low latency – Titech-Tsukuba 3-4ms roundtrip
- Operation of Photonic Cross Connect (OXC) for fiber/wavelength switching
- 6,000+km dark fiber, many e-e lambdas, 10Gb
- Operational since January, 2002




University Computer Centers
(excl. National Labs) circa Spring 2006
10Gbps SuperSINET
Interconnecting the Centers
~60 SC Centers
in Japan
10Petaflop
center by
2011
Hokkaido University
Information Initiative Center
HITACHI SR11000
5.6 Teraflops
University of Tsukuba
FUJITSU VPP5000
CP-PACS 2048 (SR8000 proto)
Kyoto University
Academic Center for Computing
and Media Studies
FUJITSU PrimePower2500
10 Teraflops
Kyushu University
Computing and
Communications Center
FUJITSU VPP5000/64
IBM Power5 p595
5 Teraflops
Tohoku University
Information Synergy Center
NEC SX-7
NEC TX7/AzusA
University of Tokyo
Information Technology Center
HITACHI SR8000
HITACHI SR11000 6 Teraflops
Others (in institutes)
National Inst. of Informatics
SuperSINET/NAREGI Testbed
17 Teraflops
Tokyo Inst. Technology
Global Scientific Information
and Computing Center
NEC SX-5/16, Origin2K/256
HP GS320/64 ➔ 80~100Teraflops 2006
Osaka University
CyberMedia Center
NEC SX-5/128M8
HP Exemplar V2500/N
1.2 Teraflops
Nagoya University
Information Technology Center
FUJITSU PrimePower2500
11 Teraflops

National Research Grid Infrastructure (NAREGI) 2003-2007

- Petascale Grid Infrastructure R&D for Future Deployment
  - $45 mil (US) + $16 mil x 5 (2003-2007) = $125 mil total
  - Hosted by National Institute of Informatics (NII) and Institute of Molecular Science (IMS)
  - PL: Ken Miura (Fujitsu➔NII)
    - Sekiguchi(AIST), Matsuoka(Titech), Shimojo(Osaka-U), Aoyagi (Kyushu-U)…
  - Participation by multiple (>= 3) vendors
  - Follow and contribute to GGF Standardization, esp. OGSA



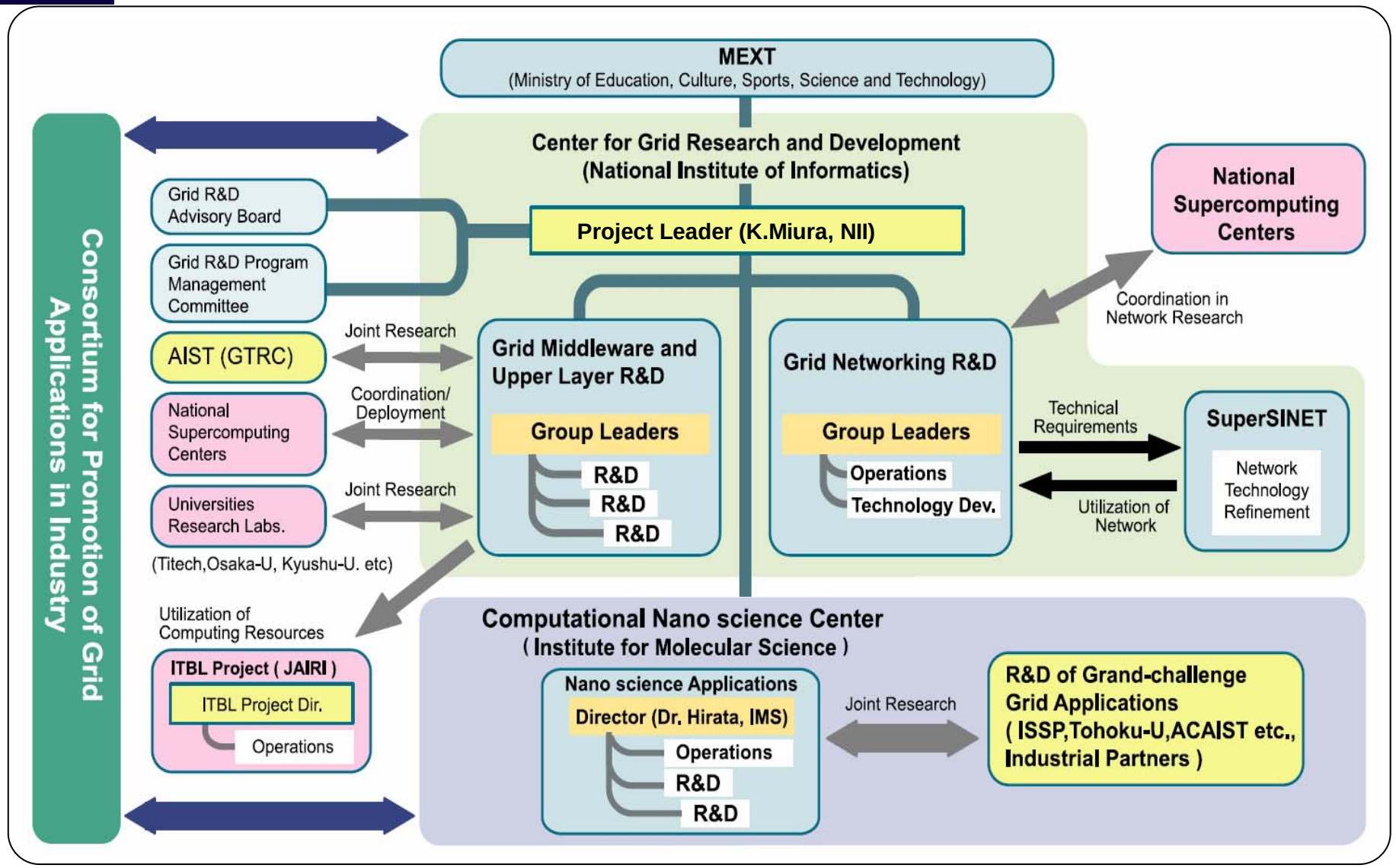
NAREGI R&D Assumptions & Goals

- Future Research Grid Metrics for Petascale
  - 10s of Institutions/Centers, various Project VOs
  - > 100,000 users, > 100,000~1,000,000 CPUs
    - Machines very heterogeneous, SCs, clusters, desktops
  - 24/7 usage, production deployment
  - Server Grid, Data Grid, Metacomputing…
- High Emphasis on Standards
  - Start with Globus, Unicore, Condor, extensive collaboration
  - GGF contributions, esp. OGSA reference implementation (Globus 4 != OGSA (!))
- Win support of users
  - Application and experimental deployment essential
  - R&D for production quality (free) software
  - Nano-science (and now Bio) involvement, large testbed

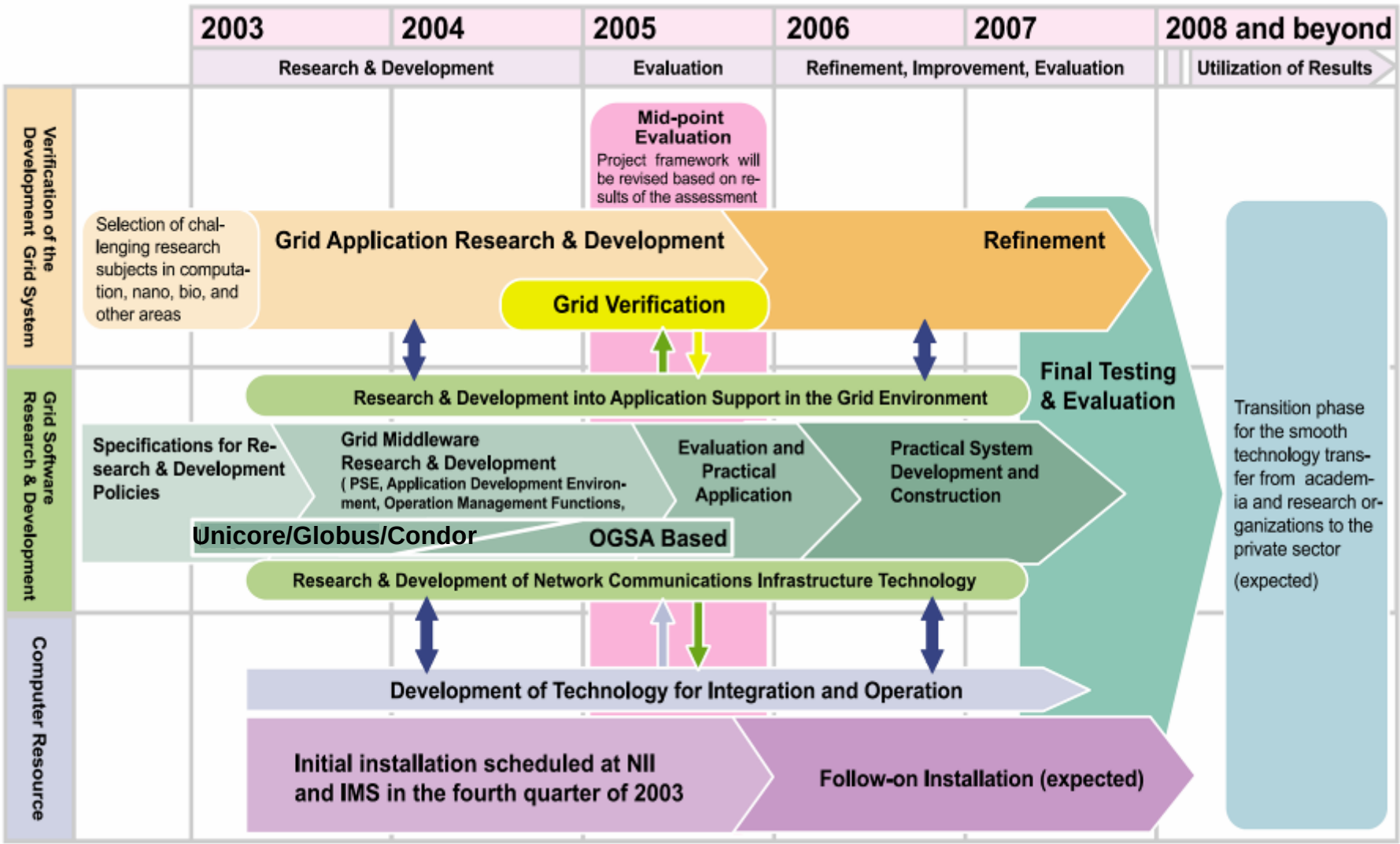
Participating Organizations

- National Institute of Informatics (NII)
  (Center for Grid Research & Development)
- Institute for Molecular Science (IMS)
  (Computational Nano – science Center)
- Universities and National Labs (Joint R&D)
  (AIST Grid Tech Research Center, Titech, Osaka-u, Kyushu-u, Kyushu Inst. Tech., etc.)
  (ITBL Project, National Supecomputing Centers etc.)
- Participating Vendors (IT as well as Chemicals/Materials)
  - Consortium for Promotion of Grid Applications in Industry

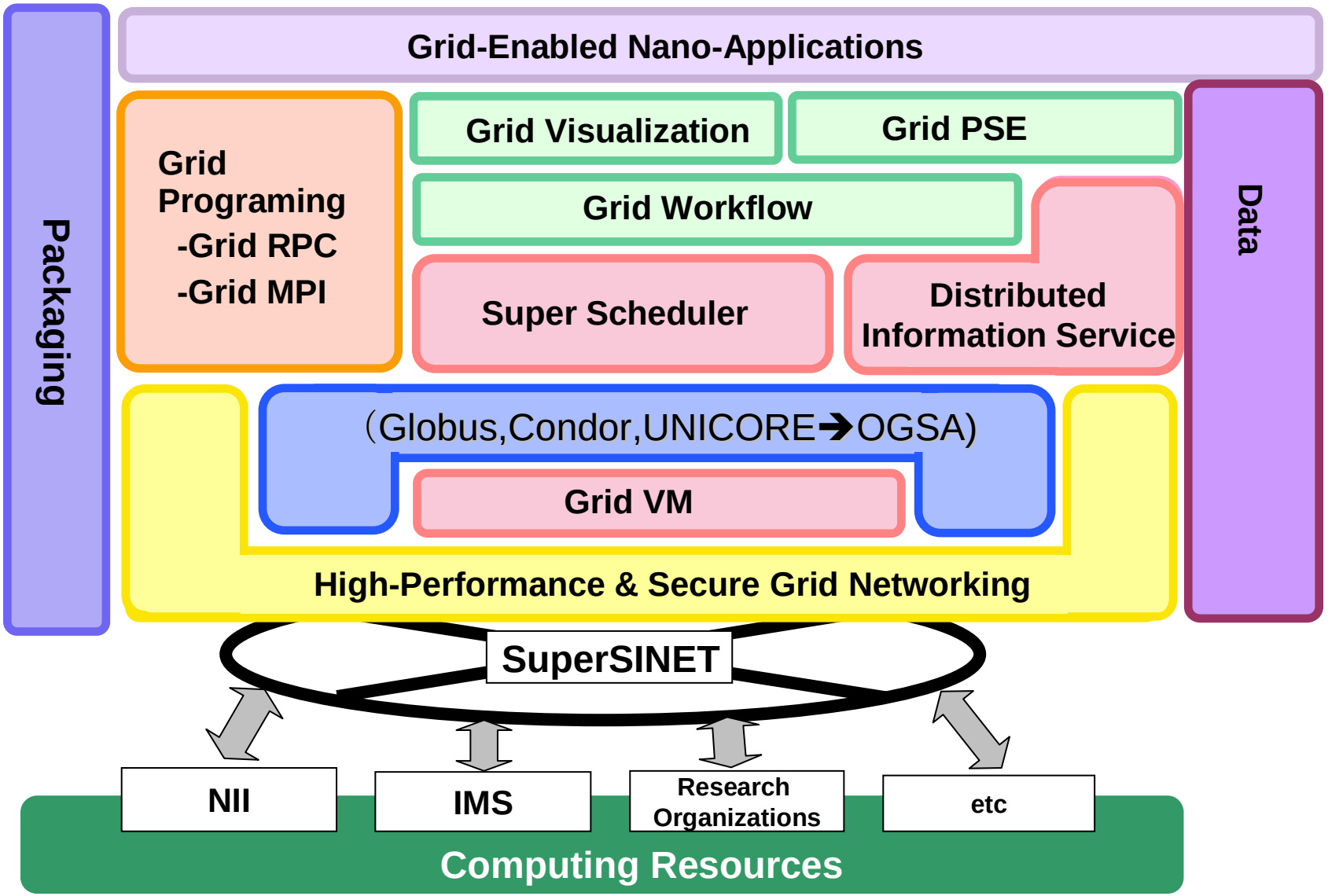
NAREGI Research Organization and Collaborations



NAREGI Five-year Plan



NAREGI Software Stack (Beta Ver. 2006)



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

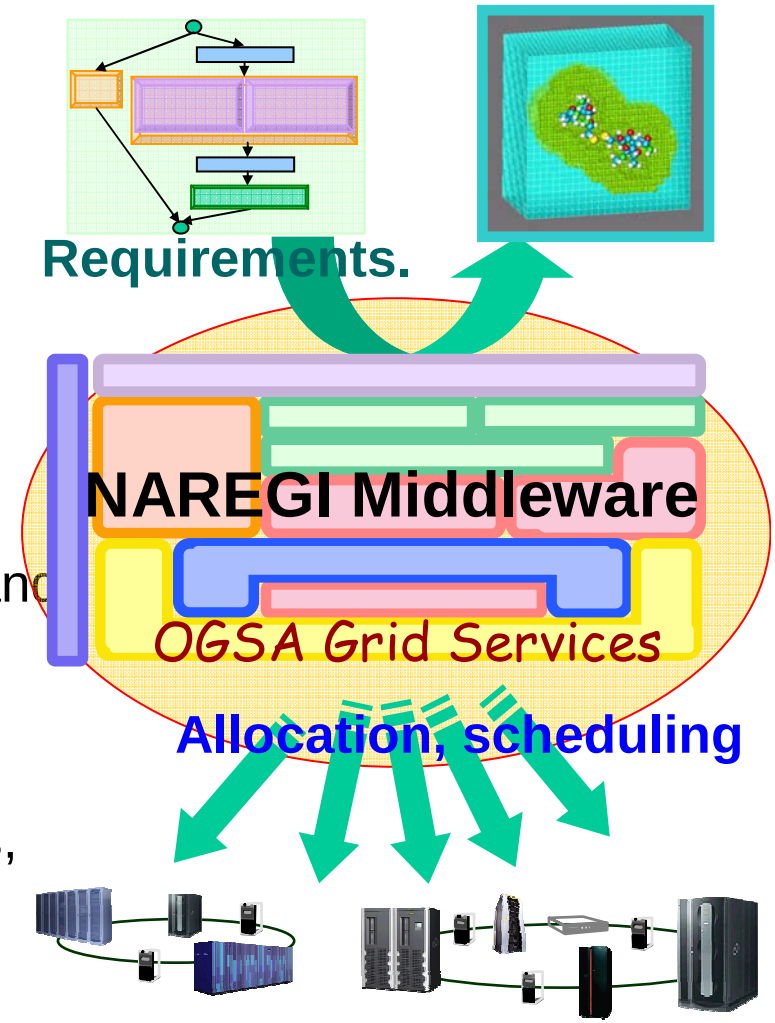
- WP-1: Resource Management:
  - Matsuoka(Titech), Nakada(AIST/Titech)
- WP-2: Programming Middleware:
  - Sekiguchi(AIST), Ishikawa(U-Tokyo), Tanaka(AIST)
- WP-3: Application Grid Tools:
  - Usami (new FY2005, NII), Kawata(Utsunomiya-u)
- WP-4: Data Management (new FY 2005, Beta):
  - Matsuda (Osaka-U), Date (Osaka-U)
- WP-5: Networking & Security
  - Shimojo(Osaka-u), Oie(Kyushu Tech.)
- WP-6: Grid-enabling Nanoscience Appls
  - Aoyagi(Kyushu-u)

NAREGI Middleware Objectives

Different Work Package deliverables implement Grid Services that combine to provide the followings:

- Allow users to execute complex jobs with various interactions on resources across multiple sites on the Grid
  - E.g., nano-science multi-physics coupled simulations w/execution components & data spread across multiple groups within the nano-VO
- Stable set of middleware to allow scalable and sustainable operations of centers as resource and VO hosting providers
- Widely adopt and contribute to grid standards, and provide open-source reference implementations, esp. GGF/OGSA.

⇒ *Sustainable Research Grid Infrastructure.*

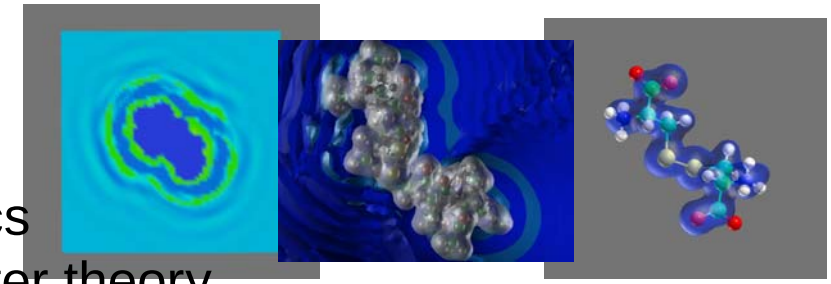


Nano-Science : coupled simluations on the Grid as the sole future for true scalability

… between Continuum & Quanta.

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

…



Molecular Science
- Quantum chemistry
- Molecular Orbital metho
- Molecular Dynamics

…

10^{-6} m 10^{-9}

Limit of Idealization

Multi-Physics

Limit of Computing Capability

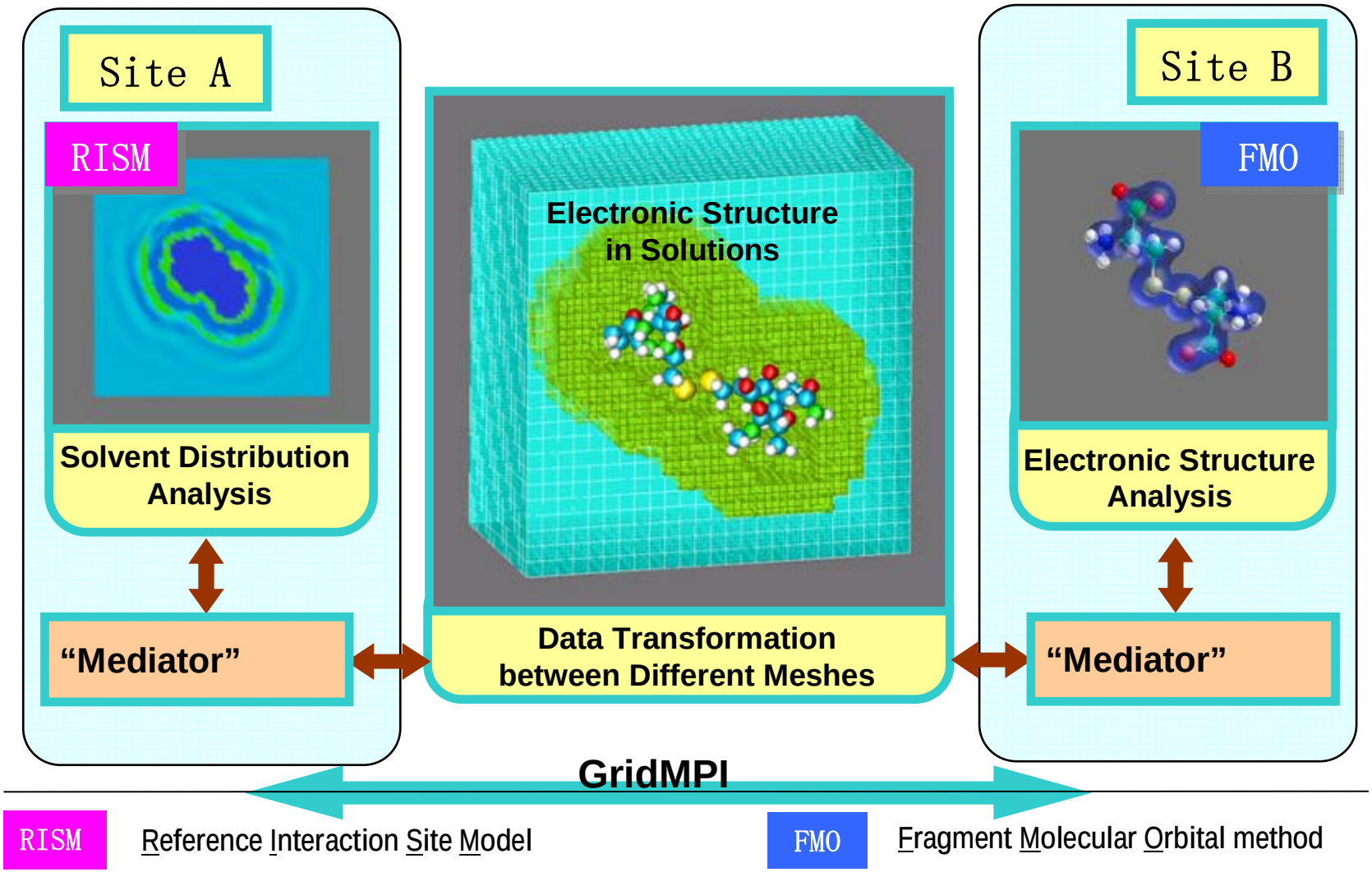
Old HPC environment:
- decoupled resources,
- limited users,
- special software, ...

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!

The NAREGI WP6 RISM-FMO-Mediator Coupled Application



NAREGI Alpha Middleware (2004-5)

1. Resource Management （WP1）

- **Unicore/Globus/Condor as “Skeletons” and underlying service provider**
- **OGSA-EMS Job Management, Brokering, Scheduling**
- Coupled Simulation on Grid (Co-Allocation/Scheduling)
- WS(RF)-based Monitoring/Accounting Federation

First OGSA-EMS based impl. In the world

2. Programming (WP2）

- High-Performance Standards-based GridMPI（MPI-1/2, IMPI）
  - Highly tuned for Grid environment (esp. collectives)
- **Ninf-G: Scientific RPC for Grids**

GridRPC GGF Programming Standard

3. Applications Environment（WP3）

- Application Discovery and Installation Management （PSE）
- Workflow Tool for coupled application interactions
- **WSRF-based Large Scale Grid-based Visualization**

WSRF-based Terabyte interactive visualization

4. Networking and Security（WP5）

- **Production-Grade CA**
- **Unicore and GSI security federation**
- **Out-of-band Real-time network monitor/control**

Drop-in Replacement for SimpleCA

5. Grid-enabling Nano Simulations （WP6）

- **Framework for coupled simulations**

Large scale coupled simluation of proteins in a solvent on a Grid

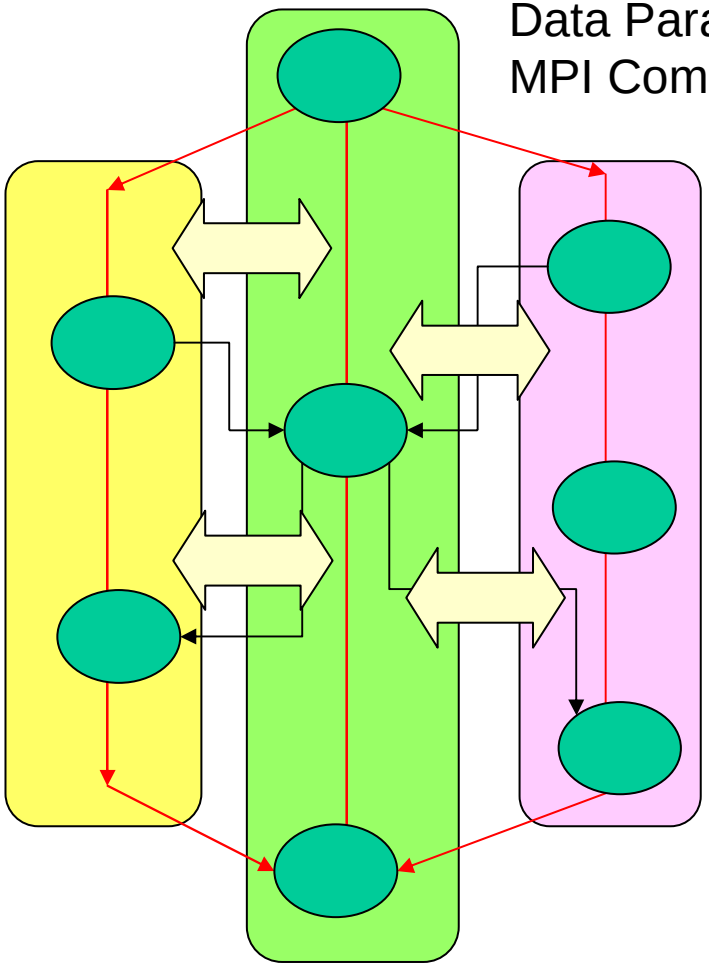
NAREGI Programming Models

- 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 + Mediator (coupled nanoscience simulation framework) / GIANT (coupled simulation data exchange framework)

NAREGI Parallel Programming Models

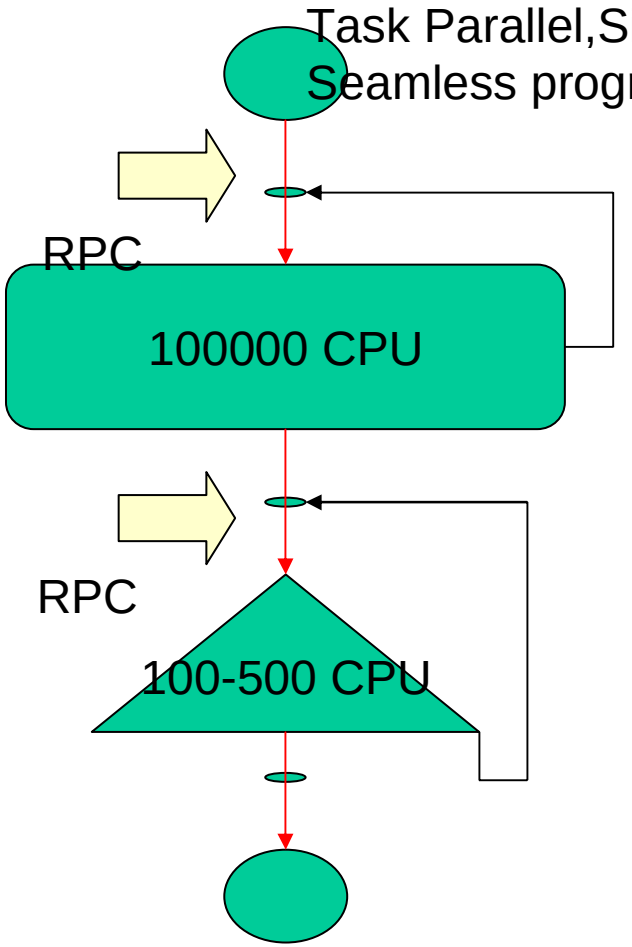
GridMPI

Data Parallel
MPI Compatibility



GridRPC (Ninf-G2)

Task Parallel,Simple
Seamless programming

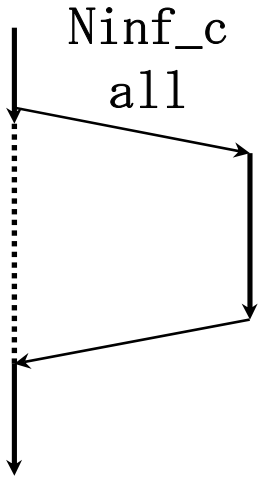


Basic Ninf-G API

- Ninf_call(FUNC_NAME, args....);
- FUNC_NAME=ninf://HOST/ENTRY_NAME

```
double A[n][n],B[n][n],C[n][n]; /* Data Decl.*/
dmmul(n,A,B,C);                 /* Call local function*/

Ninf_call("dmmul",n,A,B,C); /*  Call Ninf Func */
```



Note that this implies call-by-reference of arrays A, B, C, whose size is dynamically dictated by n

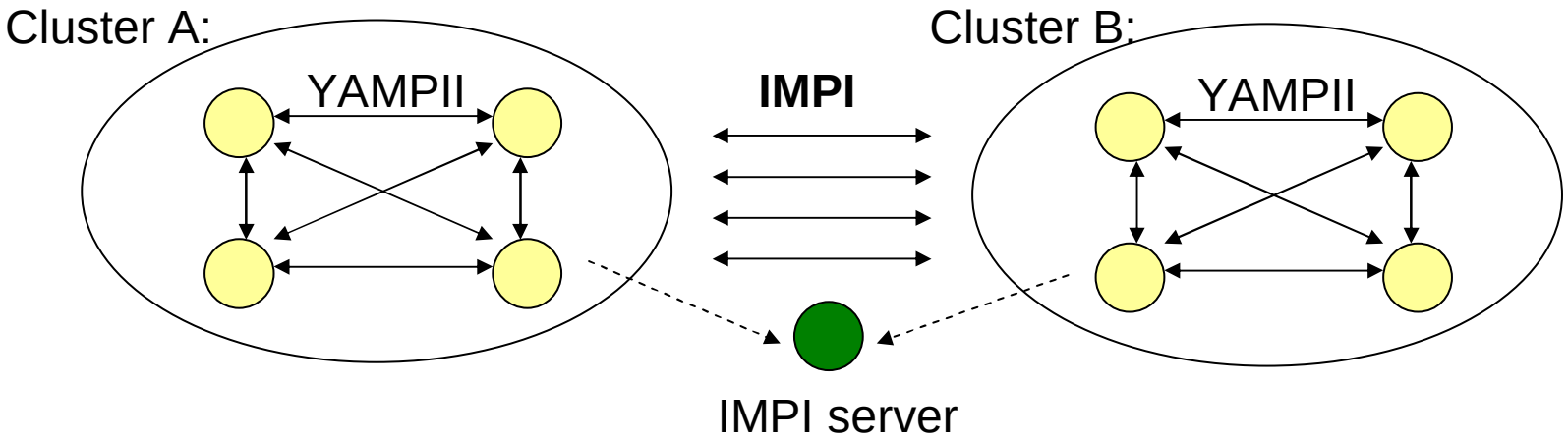
=> not trivial, info embedded in IDL, interpreted by the Ninf-G runtime

GridMPI

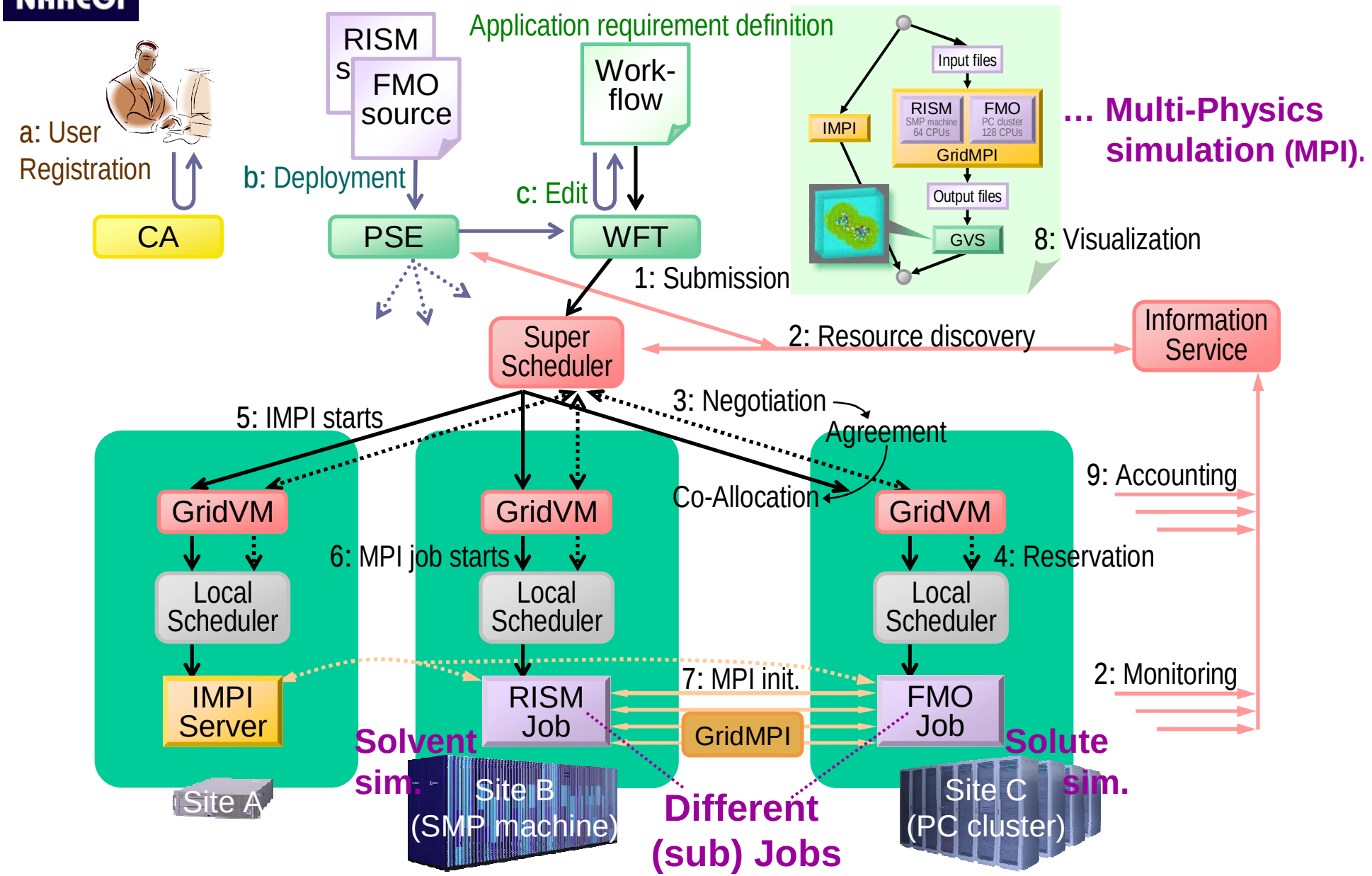
- GridMPI: a new implementation of Grid-aware MPI
  - Huge data size jobs which cannot be executed in single cluster system
  - Coupled apps spanning heterogeneous resources

① Interoperability:
- IMPI（Interoperable MPI） compliance communication protocol
- Strict implementation of MPI standard (better than MPICH!)
- Can be used standalone or plugged into various Grid middleware stack

② High performance:
- Extremely Grid Aware and Efficient, from TCP drivers to Collectives
- Can use and tested on vendor MPIs for local communication



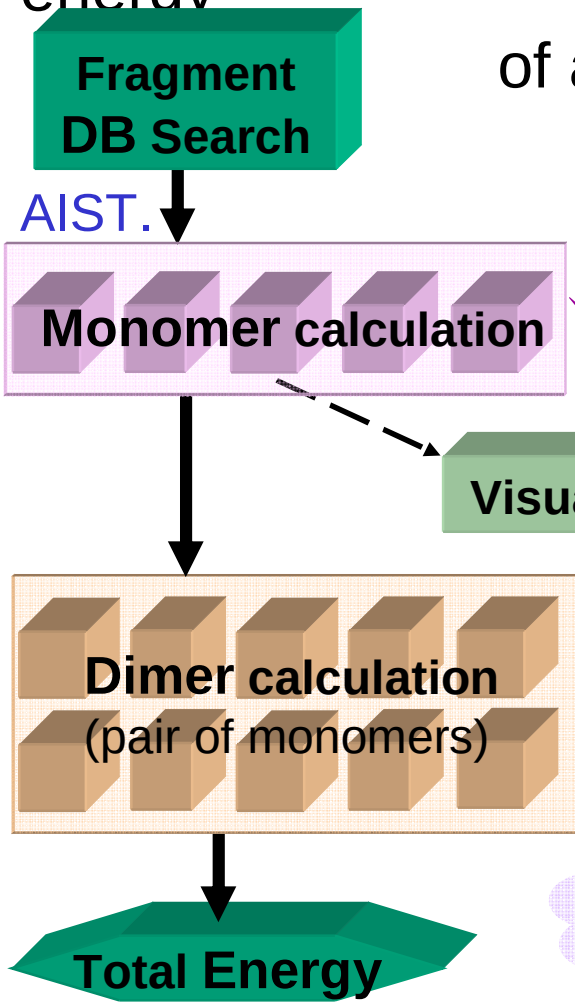
Coupled Simulation & NAREGI M/W Flow



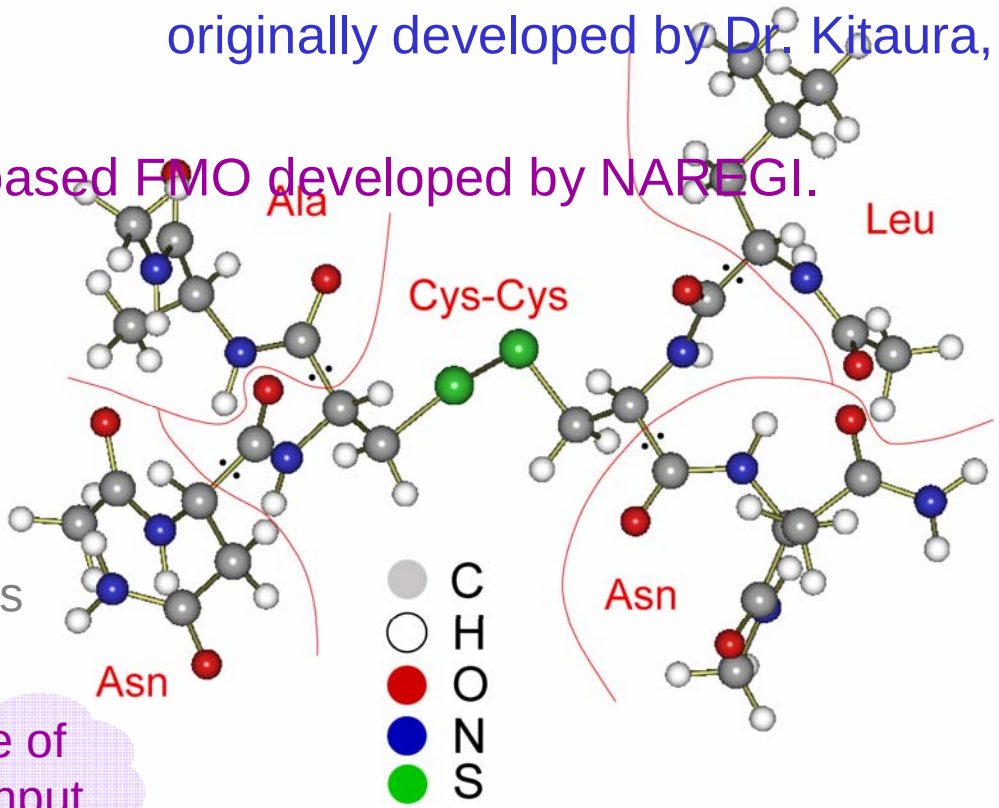
Run a job; Nano-Application

■ **Job: Grid enabled FMO** (Fragment Molecular Orbital method)
: a method of calculating electron density and energy

of a big molecule, divided to small fragments,
originally developed by Dr. Kitaura, AIST.

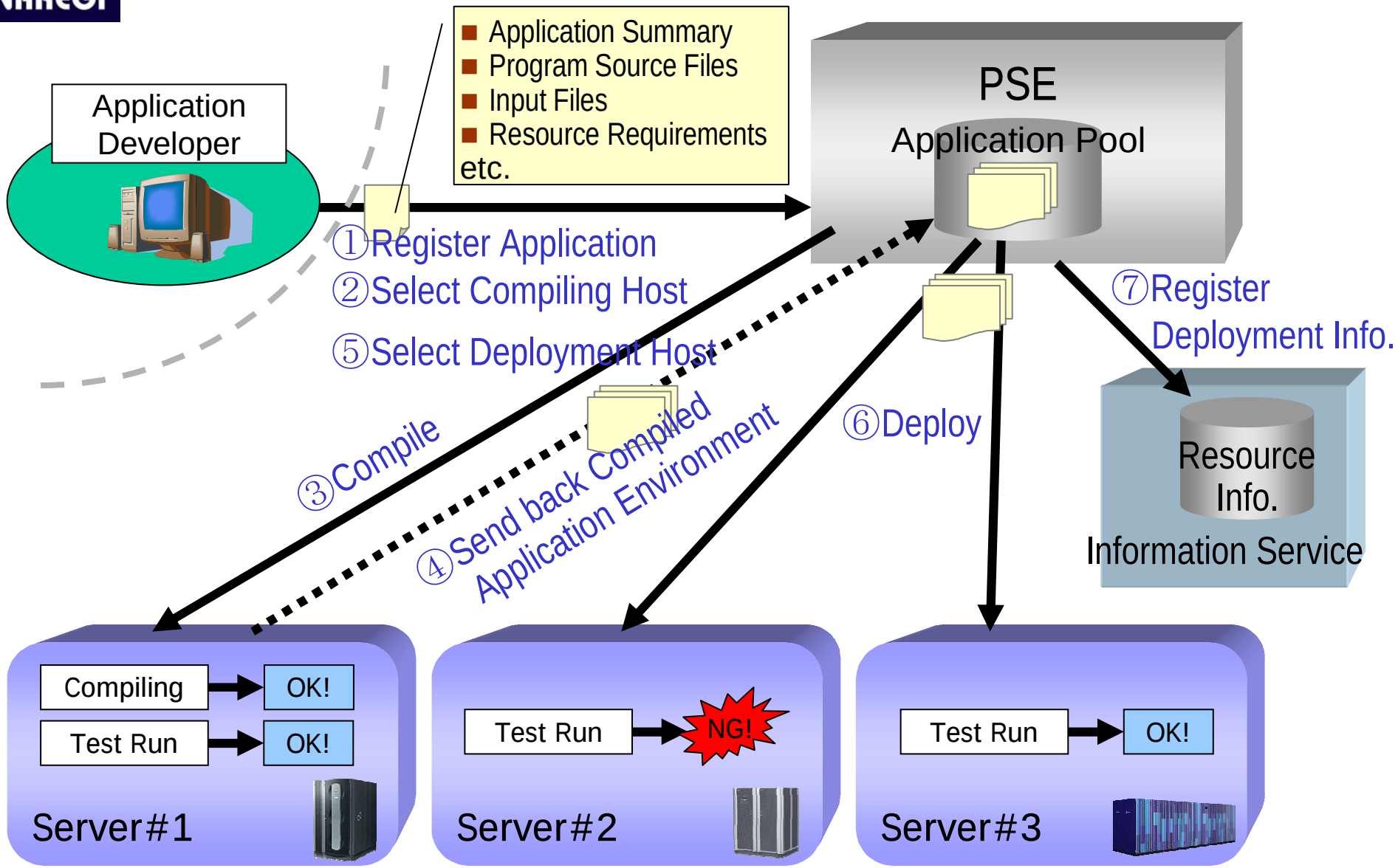


↓ Workflow based FMO developed by NAREGI.

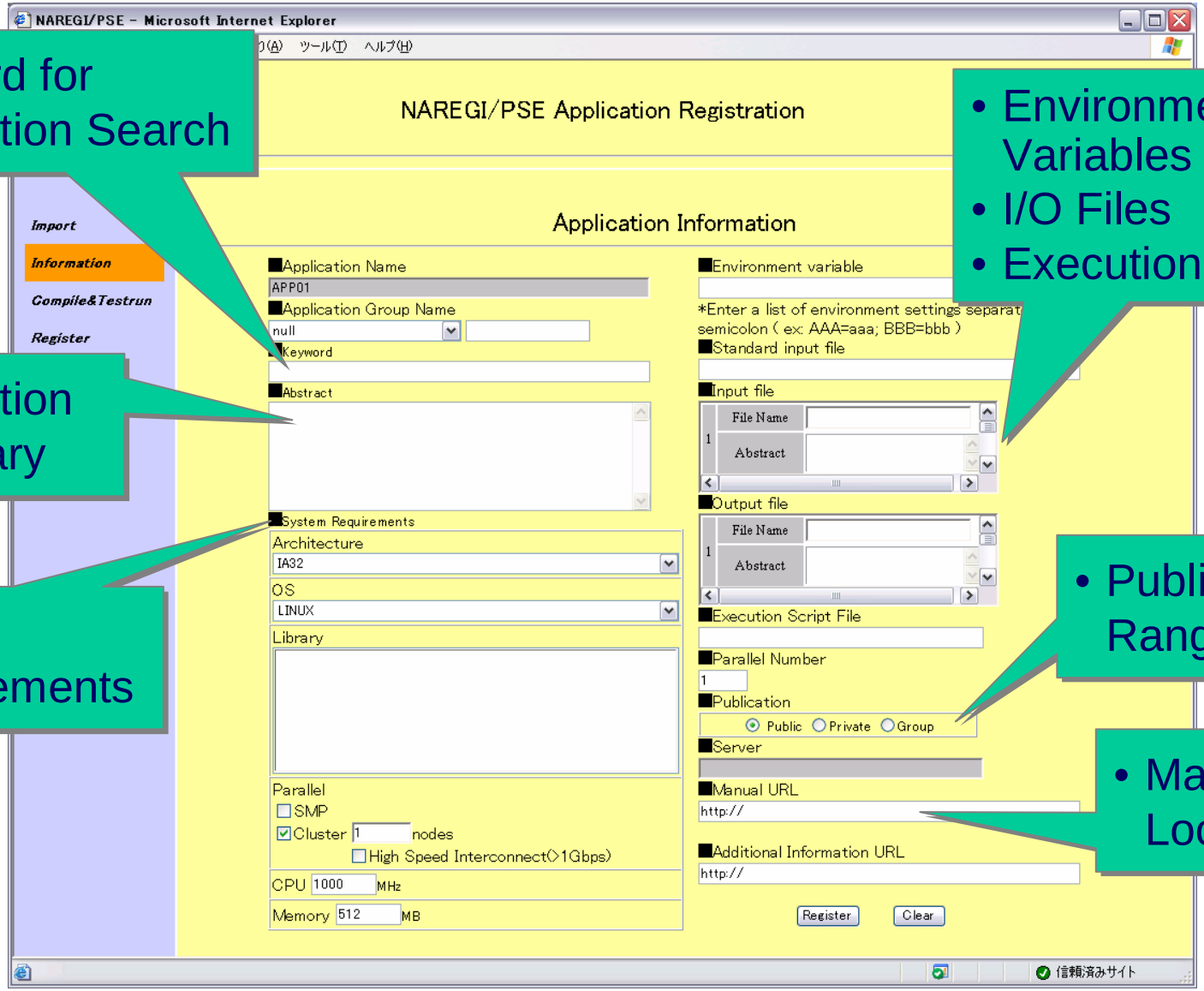


An example of High throughput computing.

b: Registration & Deployment of Applications

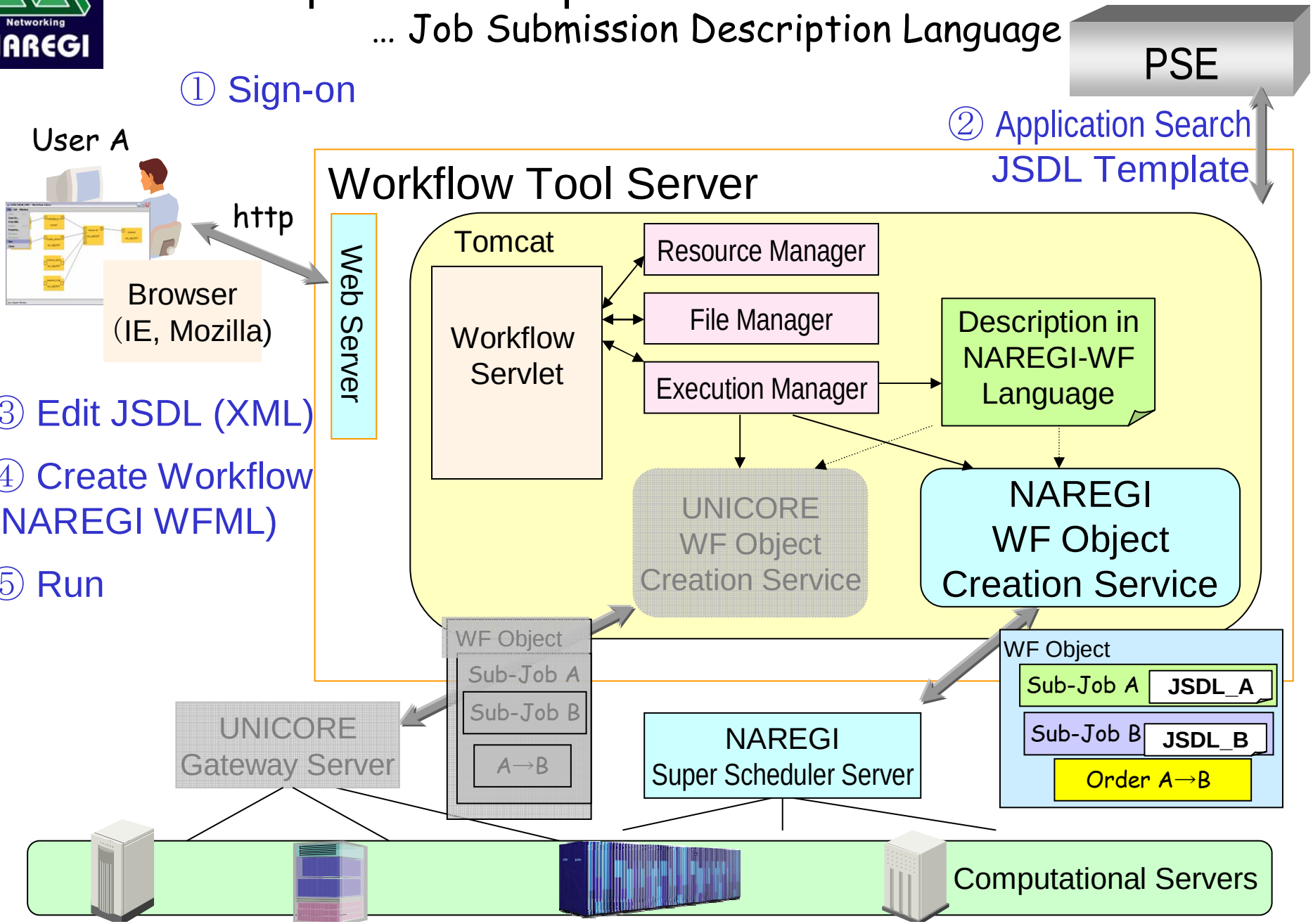


Application Registration: Screenshot



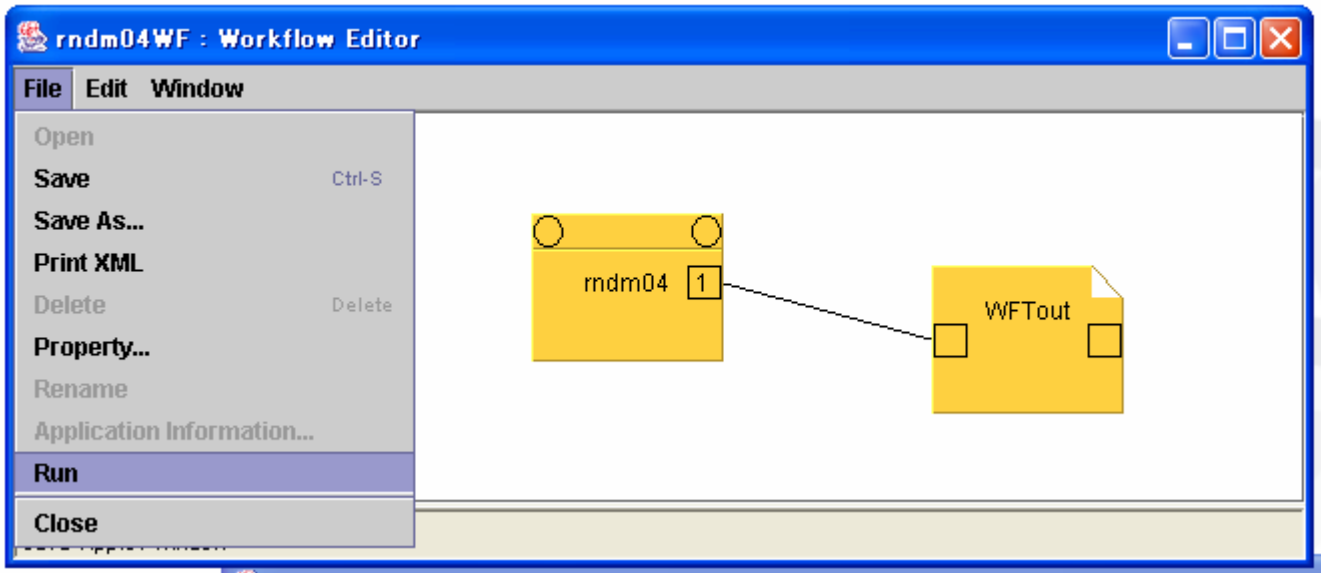
c: Description of Requirements for Workflow Execution

… Job Submission Description Language

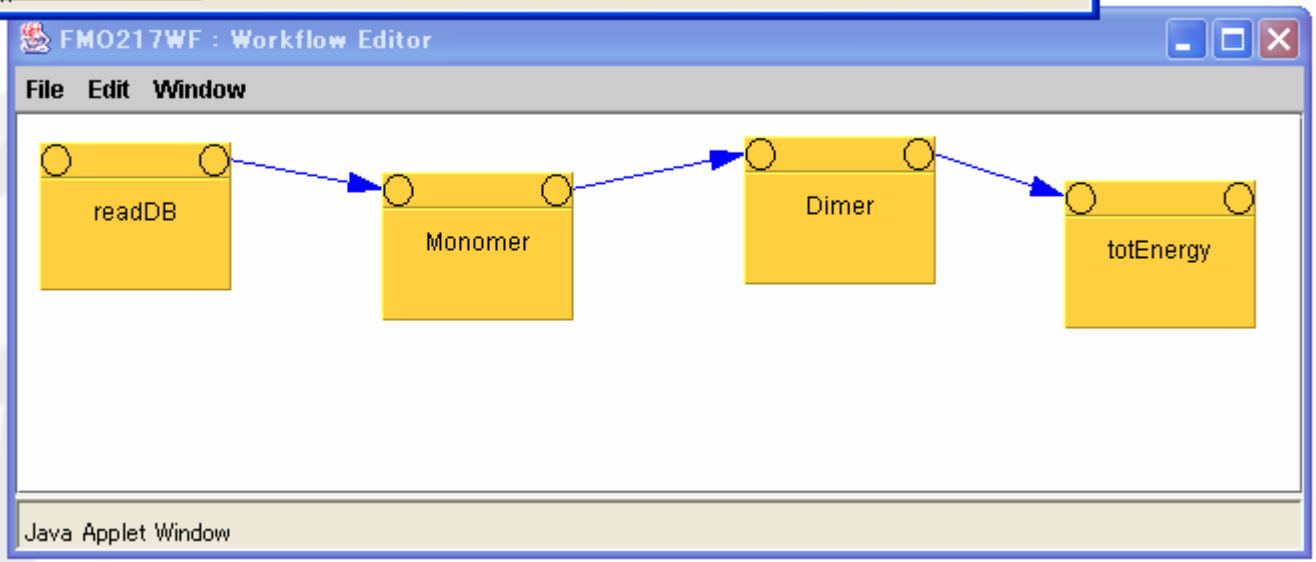


Workflow Edit & Execution: Screenshot

MPI Job:



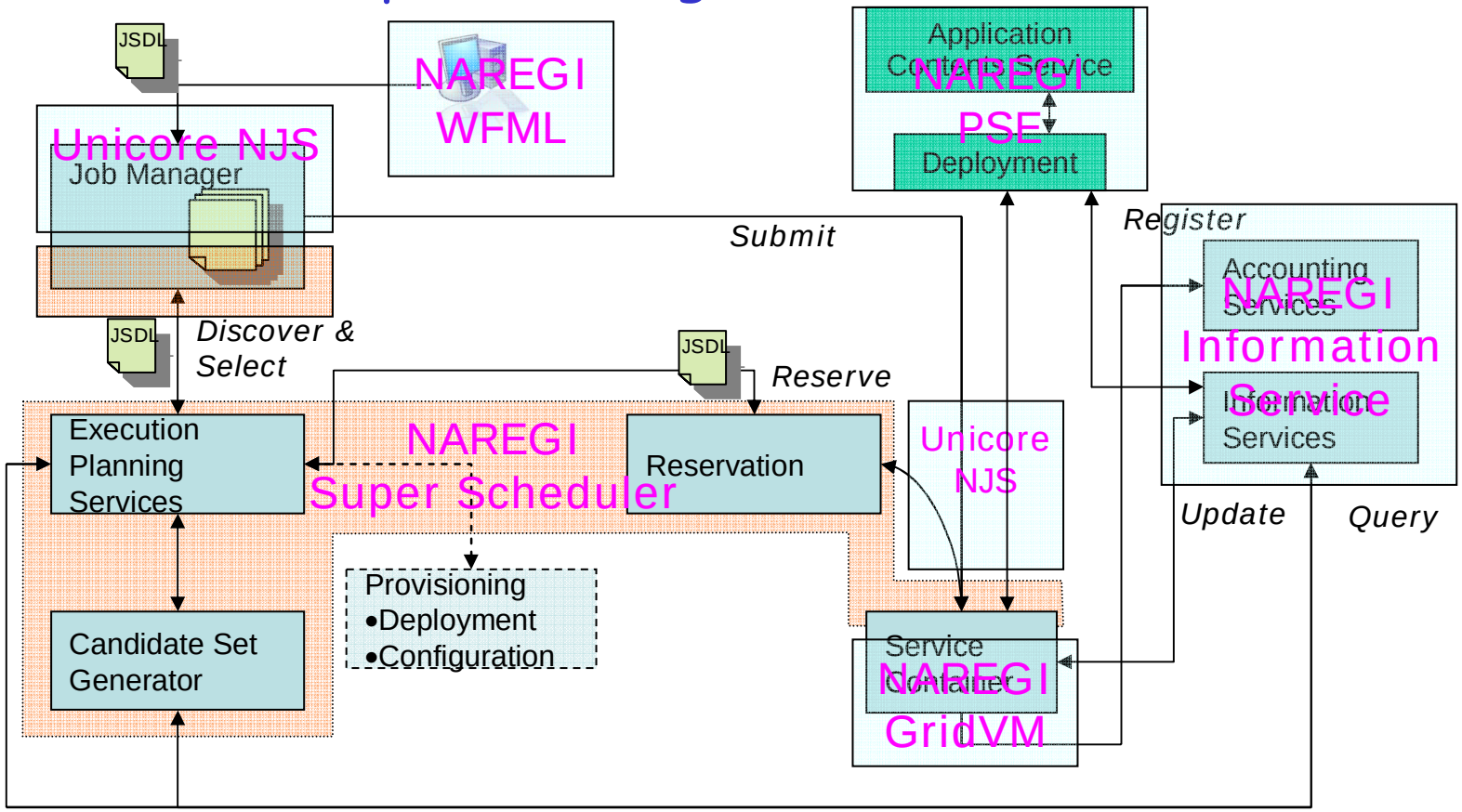
Loosely Coupled FMO Job:



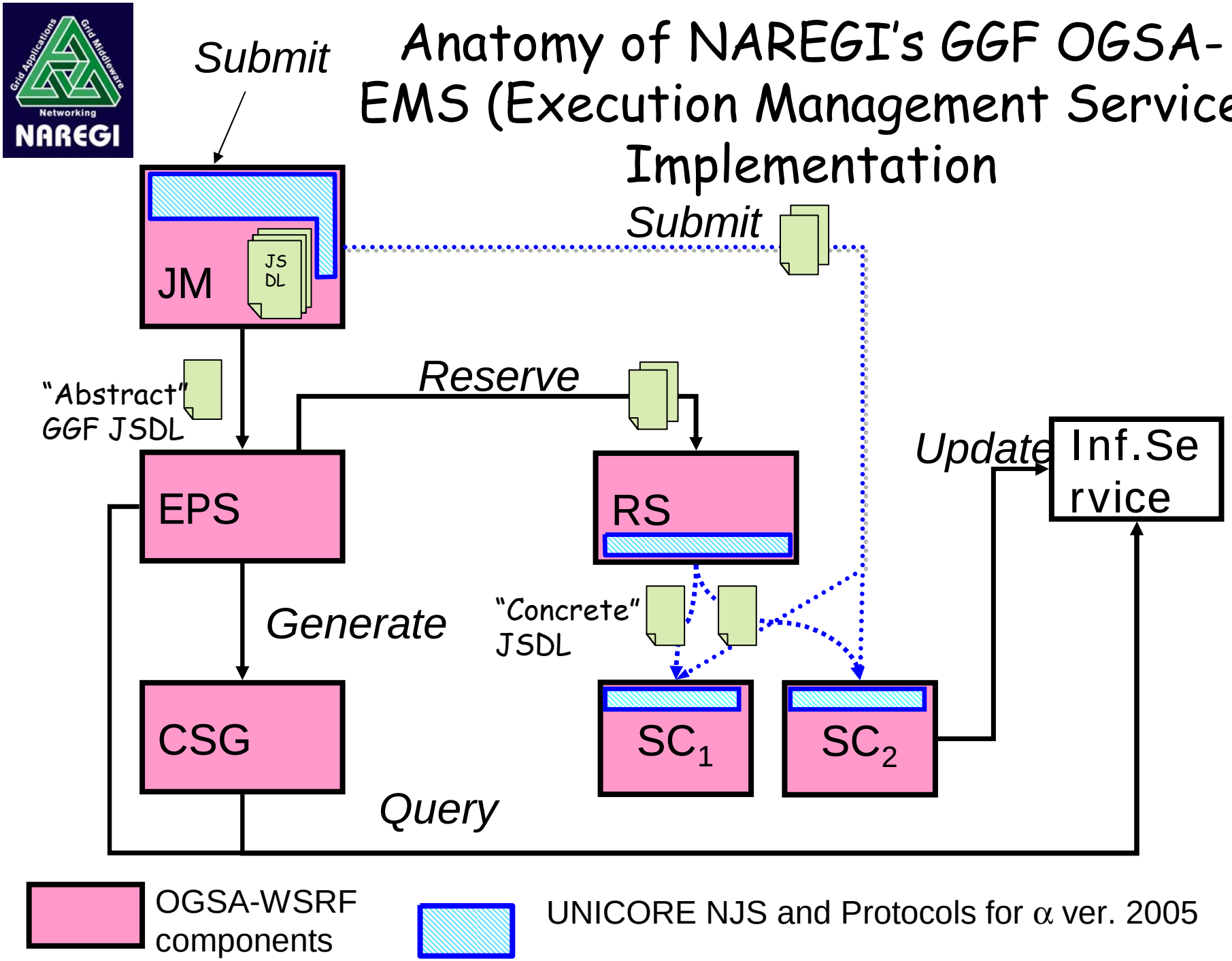
NAREGI Super Scheduler and OGSA-EMS

Open Grid Service Architecture – Execution Management Service will be standardized by OGSA EMS-WG

- NAREGI SS also serves as a test reference implementation of OGSA-EMS, contributing to the GGF standards process (e.g., OGSA-RSS, JSDL, …)



Anatomy of NAREGI's GGF OGSA-EMS (Execution Management Service) Implementation

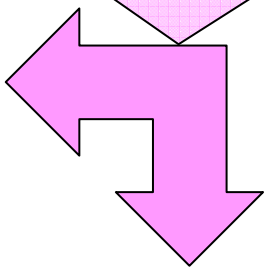


1: Job Submission

- Resource requirements of the demonstration job

CPU:	Intel Xeon
# of CPUs:	18(# of MPI tasks)
# of nodes:	9
Physical Memory:	>= 32MB/CPU
Walltime limit:	15min
…	

Cannot be allocated the job on single site !!



- Demonstration system resources

Site α :

CPU:	Intel Xeon
# of CPUs:	16
# of nodes:	8
Pmemory:	512MB/CPU
Walltime limt:	non
…	

Site μ :

CPU:	Intel Xeon
# of CPUs:	12
# of nodes:	6
Pmemory:	512MB/CPU
Walltime limt:	non
…	

GGF Job Submission Description Language

- Job Submission with JDSL

Under standardization in GGF/JSDL-WG

NAREGI JSDL (parts)

属性	説明
/naregi-jsdl:SubJobID	Subjob ID
/naregi-jsdl:CPUCount	# of total CPUs
/naregi-jsdl:TasksPerHost	# of tasks per host
/naregi-jsdl:TotalTasks	# of total MPI tasks
/naregi-jsdl:NumberOfNodes	# of nodes
/naregi-jdsl:CheckpointablePeriod	Check point period
/jsdl:PhysicalMemory	Physical memory for each process
/jsdl:ProcessVirtualMemoryLimit	Virtual memory for each process
/naregi-jsdl:JobSatrtTrigger	Job start time
/jsdl:WallTimeLimit	Wall time limit
/jsdl:CPUTimeLimit	CPU time limit
/jsdl:FileSizeLimit	Maximum file size
/jsdl:Queue	Queue name

NAREGI extensions for parallel jobs

NAREGI extension for co-allocation

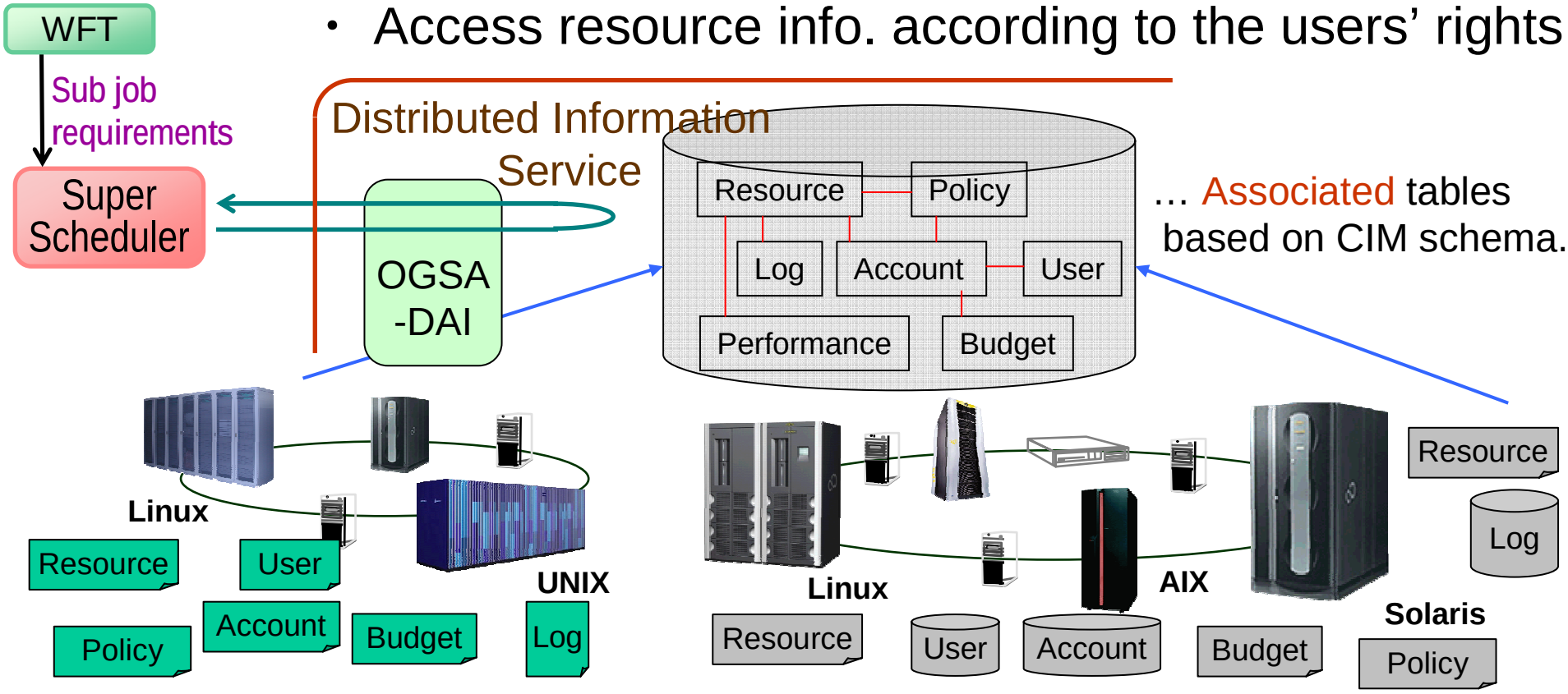
2: Resource Discovery

Distributed Info.Services maintain various kind of information;

CPU, Memory, OS, Job Queue, Account, Usage Record, etc.etc.

across multiple administrative domains,

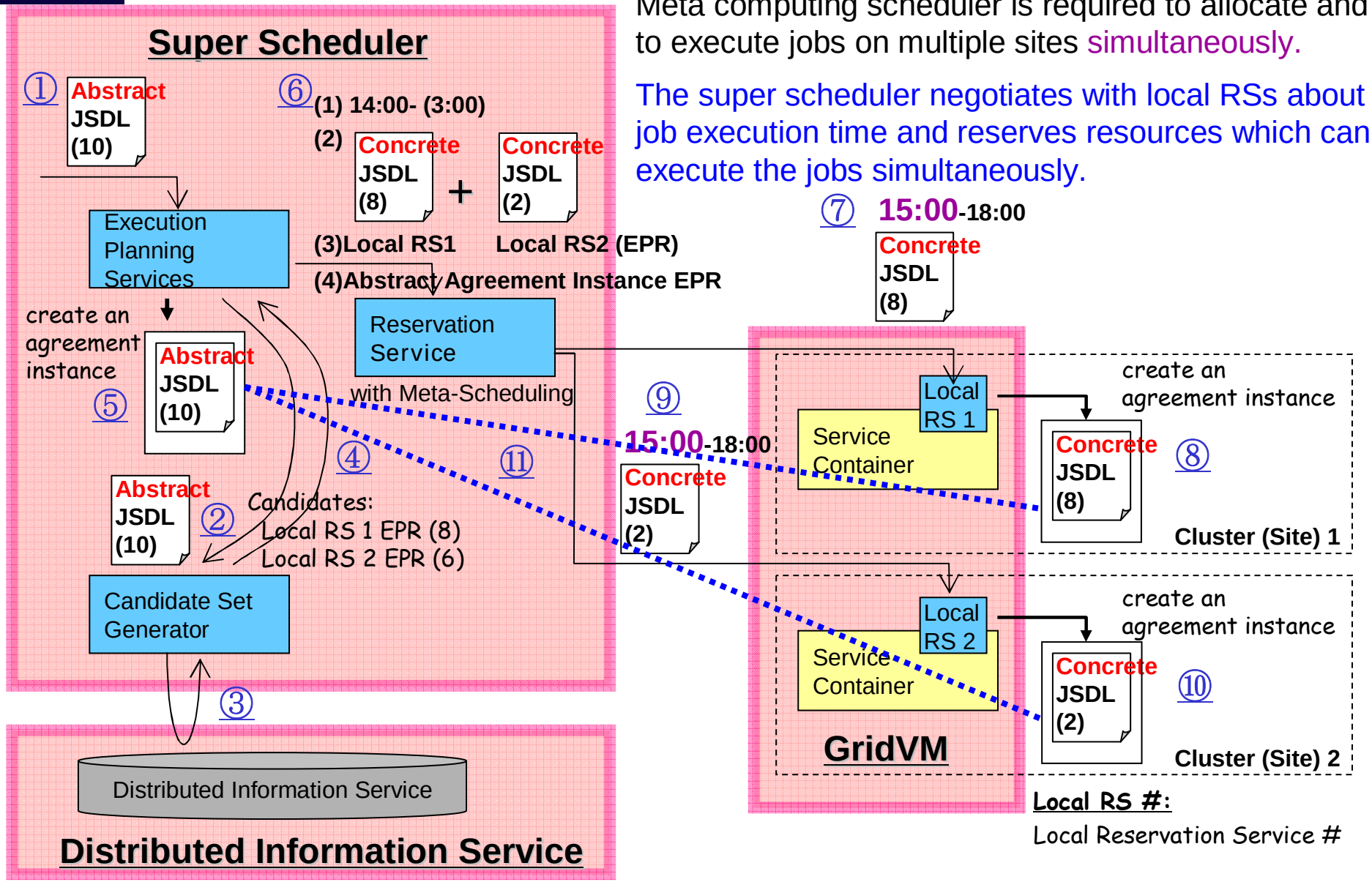
- Abstract heterogeneous resources (CIM schema) → RI
- Retrieve resource DB through Grid Service(OGSA-DA
- Access resource info. according to the users' rights.



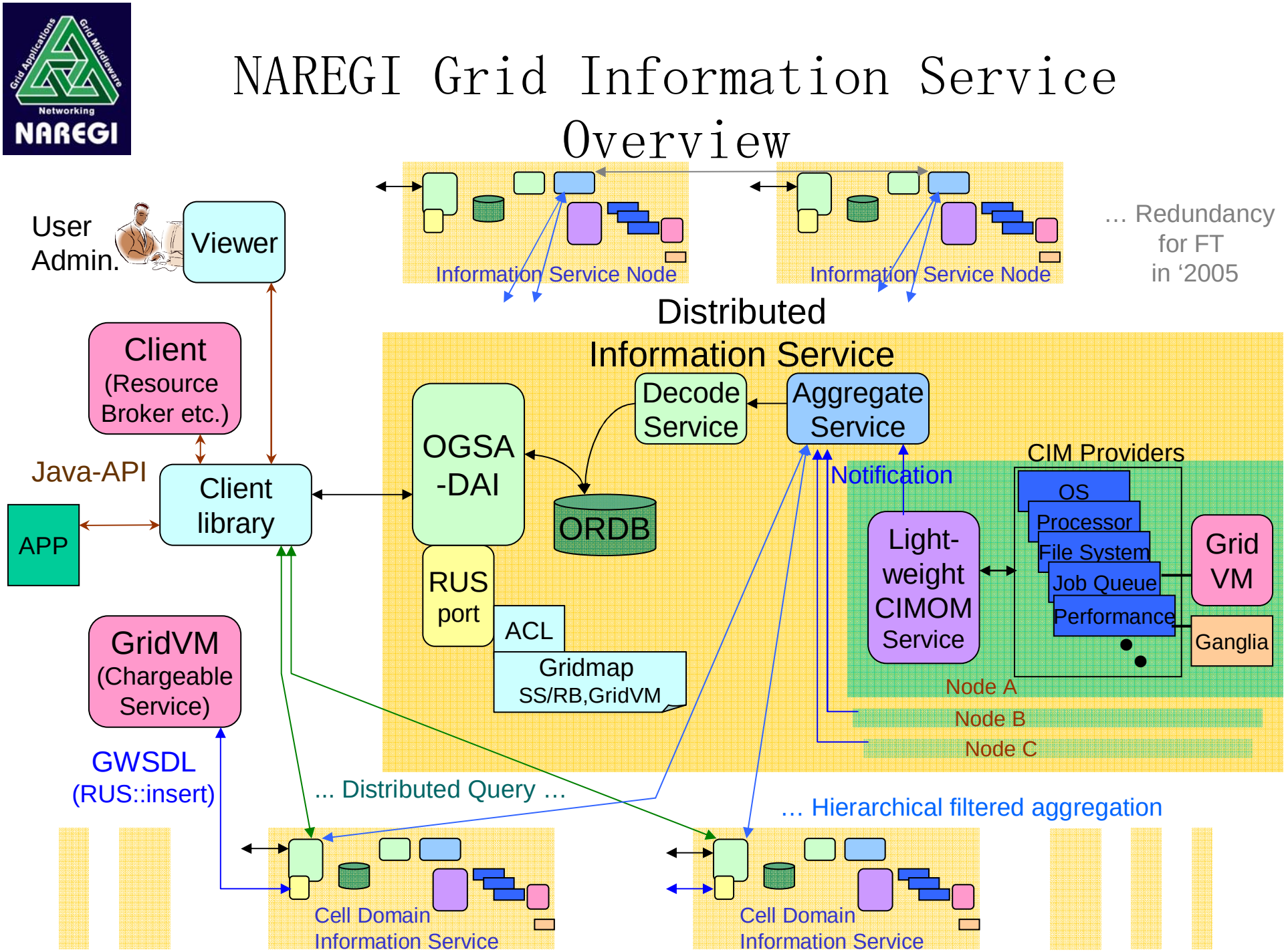
3, 4: Co-allocation and Reservation

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

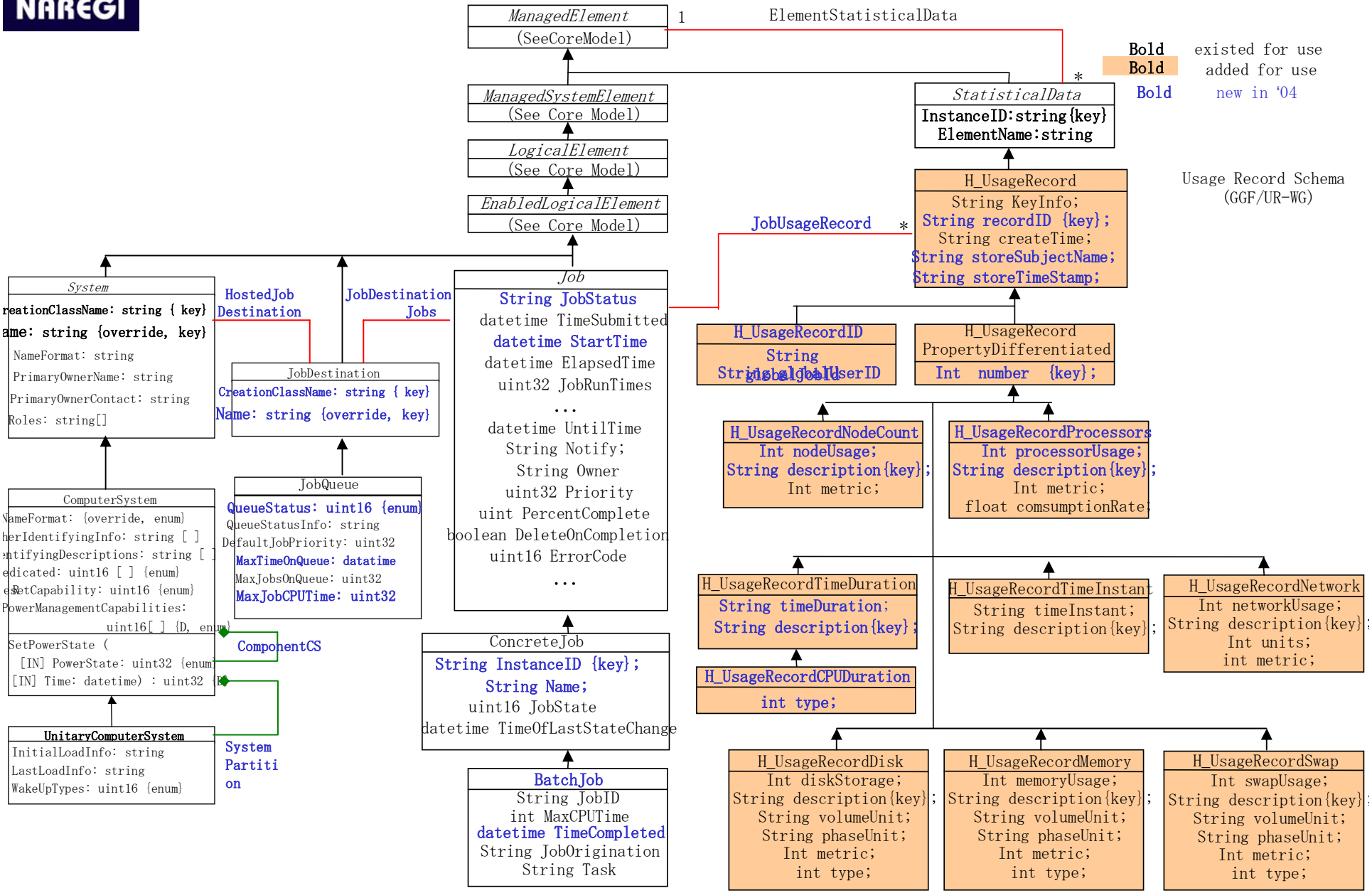
The super scheduler negotiates with local RSs about job execution time and reserves resources which can execute the jobs simultaneously.



NAREGI Grid Information Service Overview

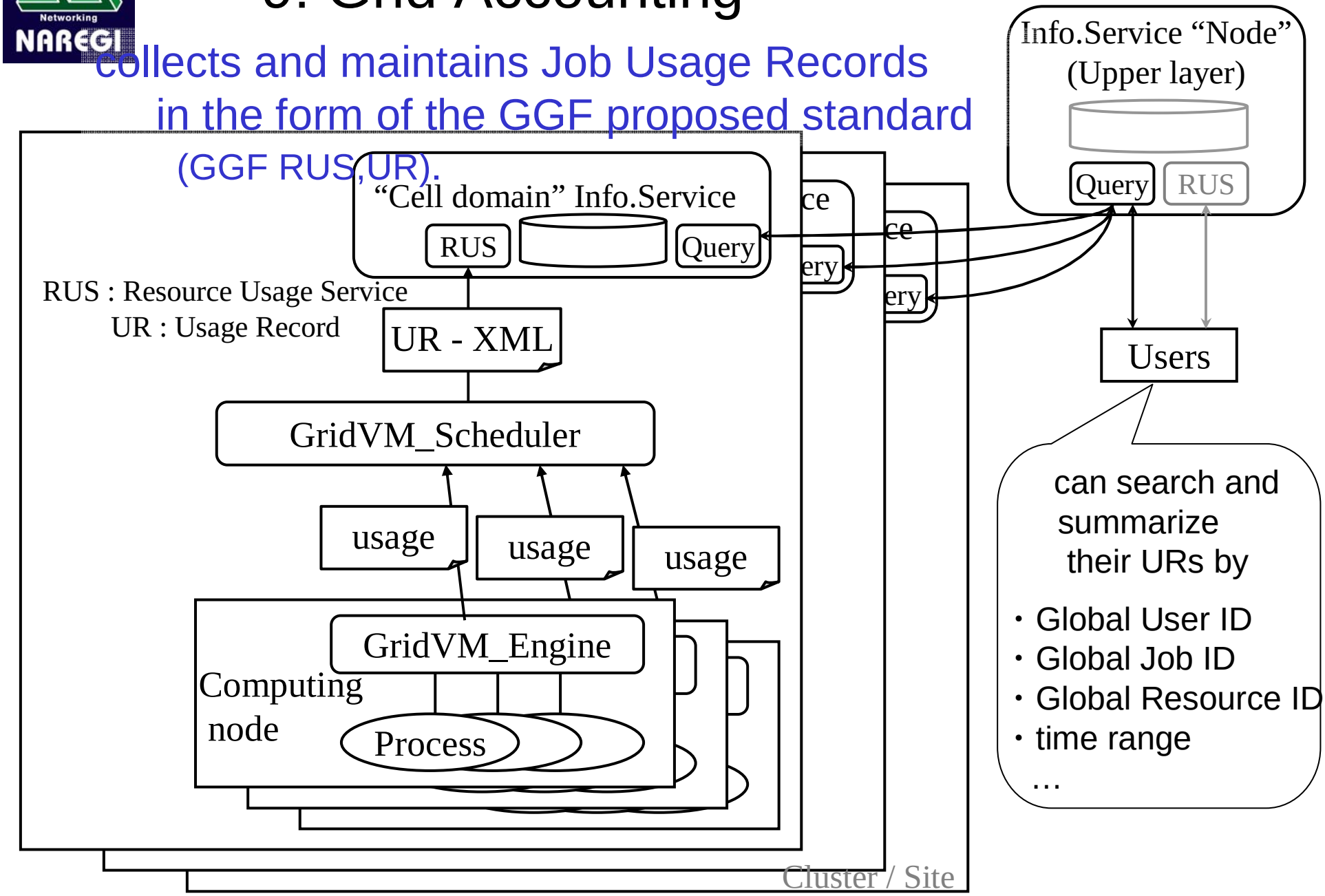


Schema for Usage Record

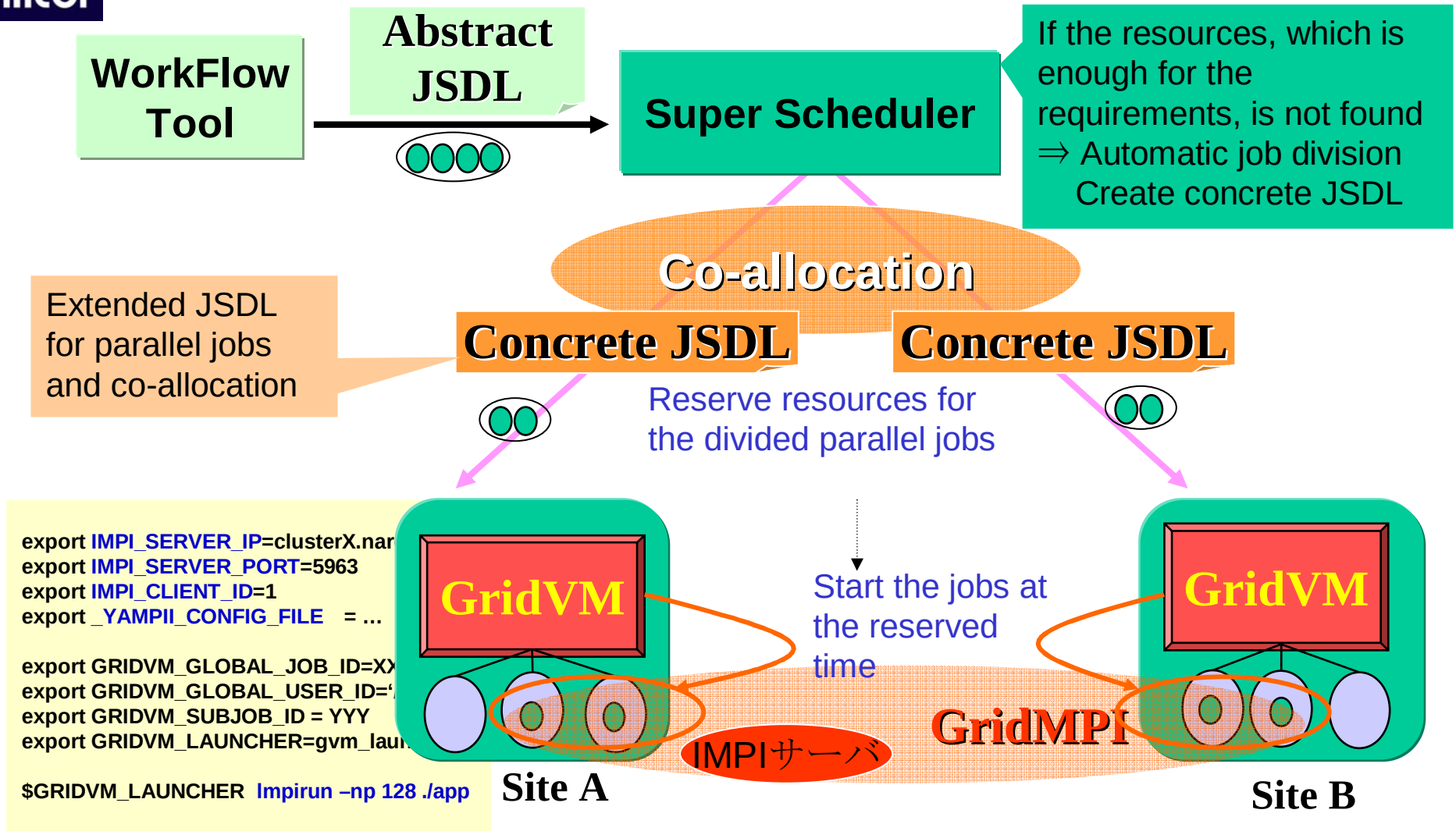


9: Grid Accounting

collects and maintains Job Usage Records
in the form of the GGF proposed standard
(GGF RUS,UR).



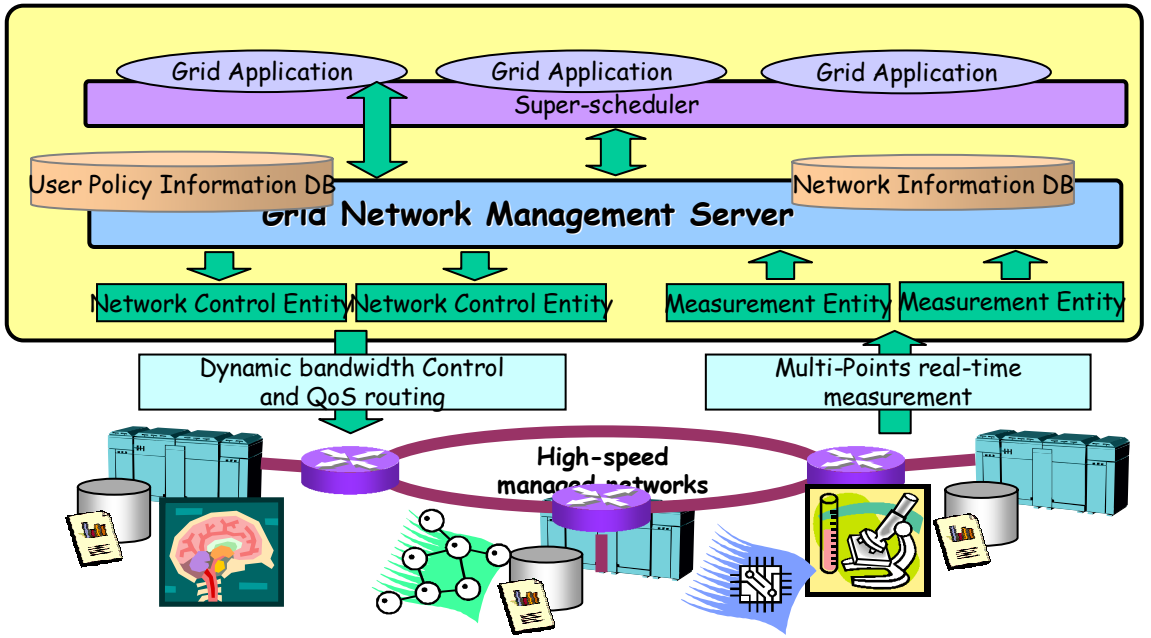
GridMPI submission



✓ Generate a shell script for the local MPI job according to JSDL and run the script.
✓ Configure the IMPI server and the execution nodes, etc…
✓PBS-Pro underneath GridVM for 2005 Alpha ver. due to reservation capabilities

WP-5: Network Measurement, Management & Control in Grid

- Real-time, Fine-Grained Traffic measurement on SuperSINET/APAN
- QoS and federated resource management of networks
- Robust and Fast Transfer Protocol for Grids
- Grid CA/User Grid Account Management and Deployment



WP5: NAREGI CA

- A full-fledged, industry-strength, Open sourceCA (Certificate Authority) for PKI
- Originally developed as Grid CA, but can be used as a general purpose CA (for UPKI)
- Free open source software

  Version 1.0.1 is available at:
  http://www.naregi.org/download/

  Can be used and customized for commercial use without restrictions

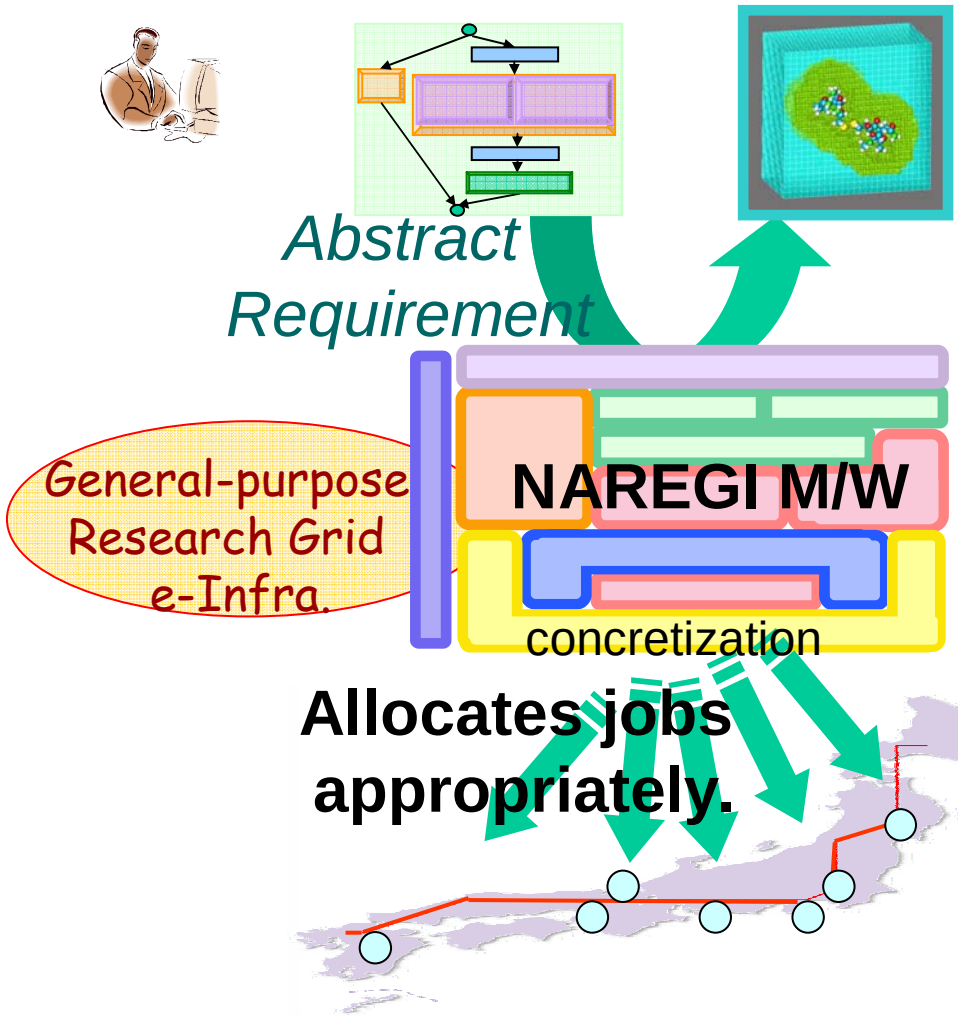
Comparison Among Various CAs

Producut name	Issue of Certif.	CRL periodical	LDAP	HSM	Multiple CA	Profile management	HW token	Operator	Logging
NAREGI CA (WP5)	file, bulk, WEB, LCMP	O	O	O	O	O	O	O	O
OpenSSL	file	-	-	-	O	-	-	-	-
Microsoft Certificate Server	WEB, LDAP	O	* (Active Directory only)	* (Domain Controller onlu)	-	* (Domain Controller only)	O	-	* (Event logging)
Commercial Entrust Authority	CMP, bulk, LDAP,WEB, SCEP	O	O	O	-	O	O	O	O

O：available、-：not available、*：some restrictions

Summary: NAREGI Middleware [Alpha]

・Users can execute complex jobs across resource sites, w/mulitple components and data from groups with a VO such as coupled simulations in nano-science.
・NAREGI M/W allocates the jobs to grid resources appropriately.



Users describe
- Component availability within VO
- Workflow being executed with abstract requirement of sub jobs.

NAREGI M/W
- discovers appropriate grid resources,
- allocates resources executes on virtualized environments
  - as a ref. impl. of OGSA-EMS and others
- Programming Model : GridMPI & Grid RPC
- User Interface : Deployment, Visualization, WFT

will be evaluated on NAREGI testbed.
- ～15TF, 3000 CPUs
- SuperSINET 10Gbps national backbone

Highlights of NAREGI Beta (2005-2006)

- "Full" OGSA-EMS incarnation
  - OGSA-EMS/RSS WSRF components --- no legacy (pre-WS) Unicore/Globus dependencies
  - WS-Agreement brokering and co-allocation
  - JSDL-based job submission to WS-GRAM
  - 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
- Grid-wide file system (GFarm) and other data management tools
- Complex workflow for various coupled simulations
- Overall stability/speed/functional improvements for real deployment
- To be interoperable with WSRF/OGSA components in EGEE, UniGrids, etc.

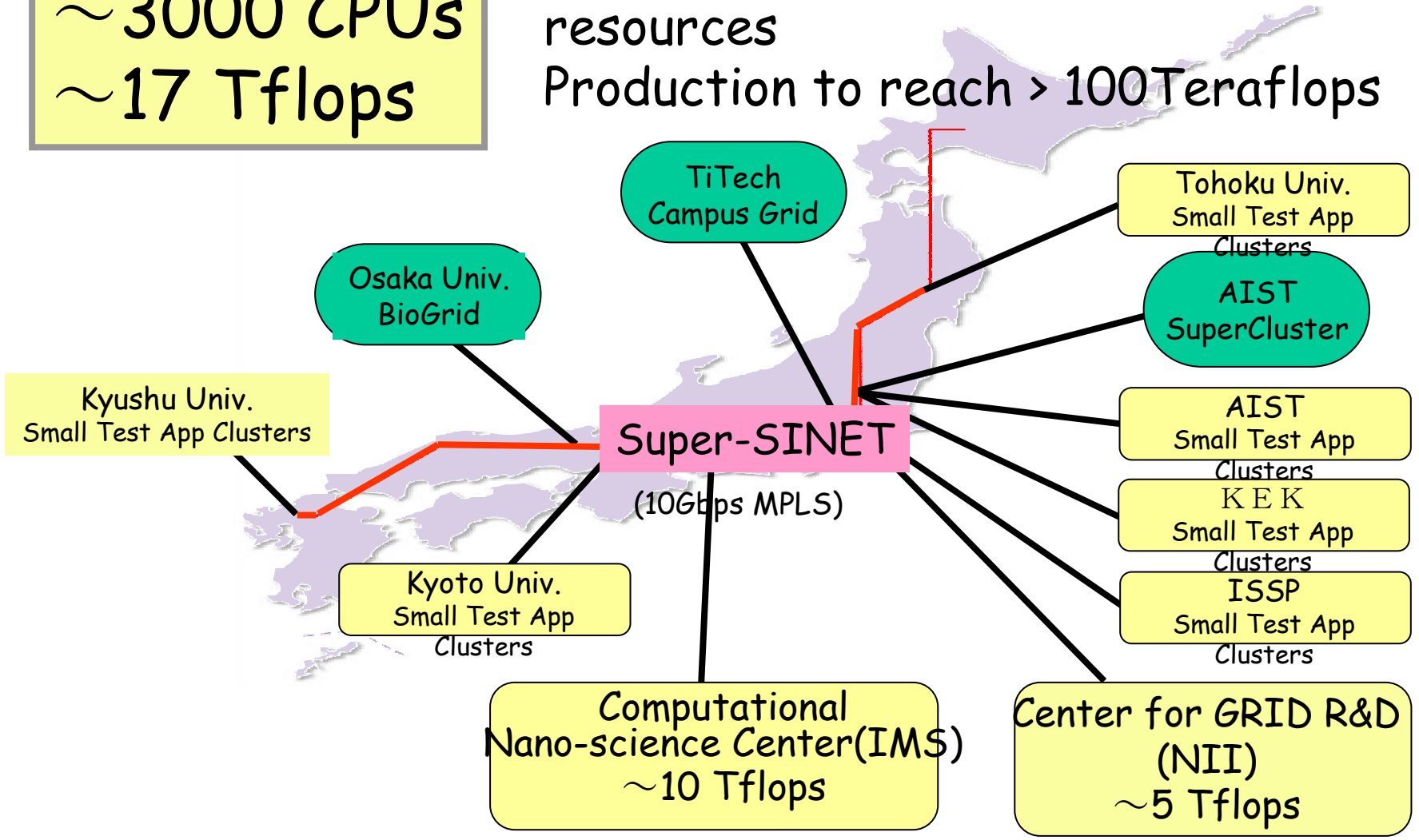
NAREGI Phase 1 Testbed

Dedicated Testbed

No "ballooning" w/production resources

Production to reach > 100Teraflops

~3000 CPUs
~17 Tflops



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

File Server (PRIMEPOWER 900 ＋ETERNUS3000 ＋ ETERNUS

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

SMP type Compute Server (SGI Altix3700)

1node
(Itanium2 1.3GHz/32CPU)

Memory 32GB
Storage 180GB

SMP type Compute Server (IBM pSeries690)

1node
(Power4 1.3GHz/32CPU)

Memory 64GB
Storage 480GB

Intra NW

L3 SW
1Gbps
(upgradable
To 10Gbps)

Ext. NW

L3 SW
1Gbps
(Upgradable
to 10Gbps)

SuperSINET

Intra NW-A

High Perf. Distributed-memory type Compute Server (PRIMERGY RX200)

128 CPUs(Xeon, 3.06GHz)+Control Node

InfiniBand 4X(8Gbps)

Memory 130GB
Storage 9.4TB

High Perf. Distributed-memory Type Compute Server (PRIMERGY RX200)

128 CPUs(Xeon, 3.06GHz)+Control Node

InfiniBand 4X (8Gbps)

Memory 65GB
Storage 9.4TB

Intra NW-B

Distributed-memory type Compute Server (Express 5800)

128 CPUs (Xeon, 2.8GHz)+Control Node

GbE (1Gbps)

Memory 65GB
Storage 4.7TB

Distributed-memory type Compute Server(Express 5800)

128 CPUs (Xeon, 2.8GHz)+Control Node

GbE (1Gbps)

Memory 65GB
Storage 4.7TB

Distributed-memory type Compute Server(HPC LinuxNetworx)

128 CPUs (Xeon, 2.8GHz)+Control Node

GbE (1Gbps)

Memory 65GB
Storage 4.7TB

Distributed-memory type Compute Server(HPC LinuxNetworx)

128 CPUs(Xeon, 2.8GHz)+Control Node

GbE (1Gbps)

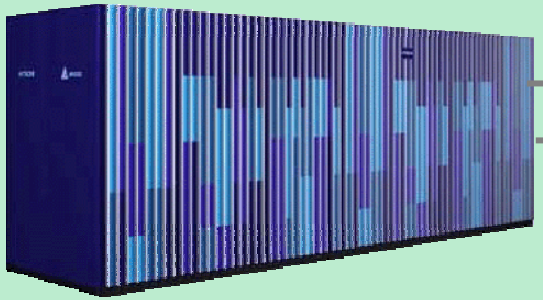
Memory 65GB
Storage 4.7TB

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



Memory 3072GB
Storage 2.2TB

Front-end Server

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

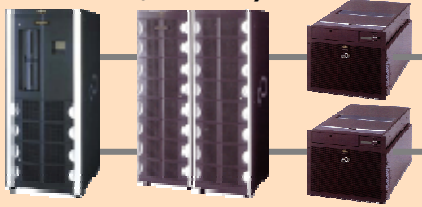
Memory 1.6TB
Storage 1.1TB/unit

Front-end Server

L3 SW
1Gbps
(Upgradable to 10Gbps)

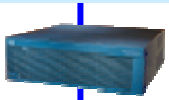
File Server

16CPUs (SPARC64 GP, 675MHz)



Memory 8GB
Storage 30TB
Back-up 25TB

CA/RA Server



Firewall

VPN

Center for Grid R & D

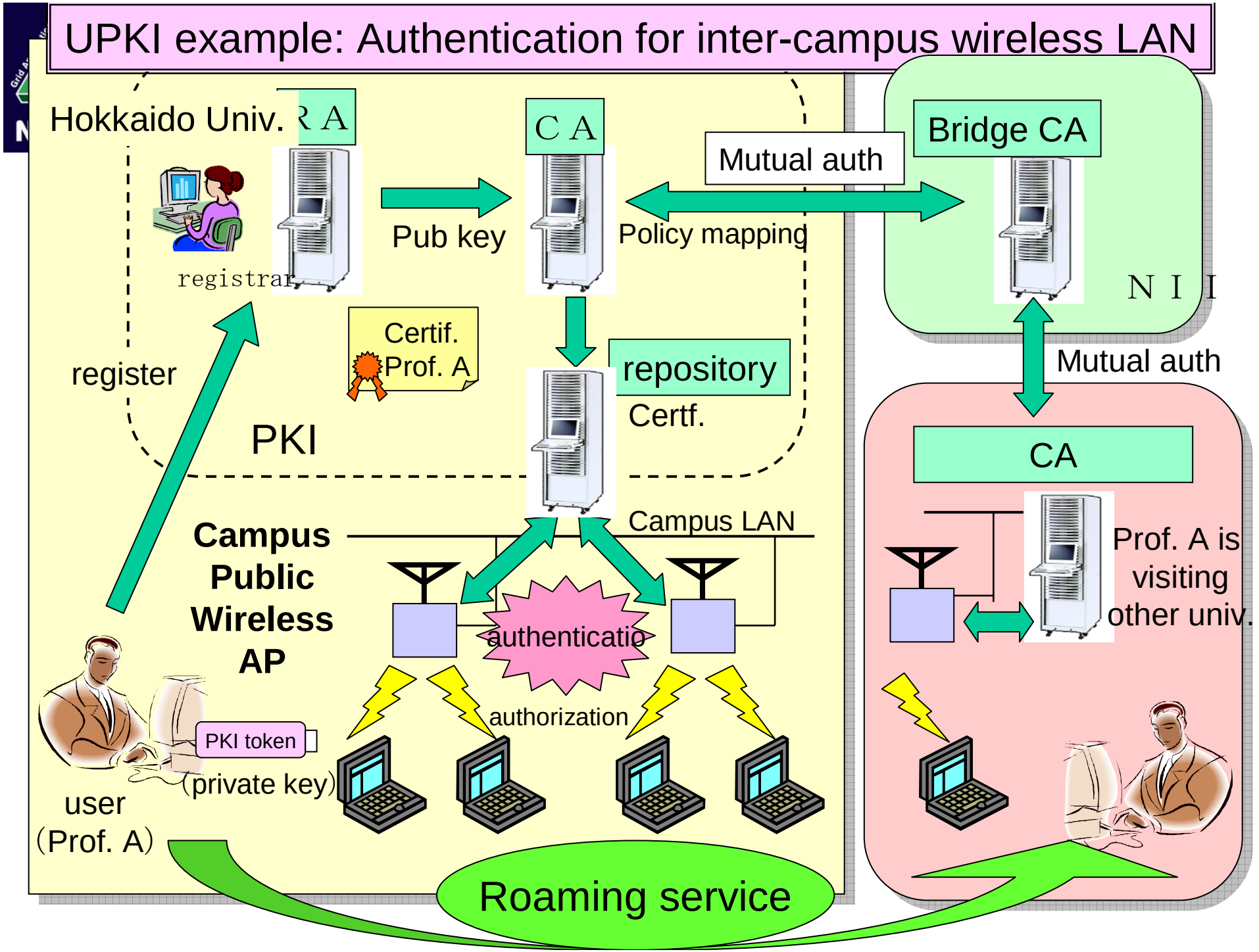
SuperSINET

UPKI (University PKI): concept

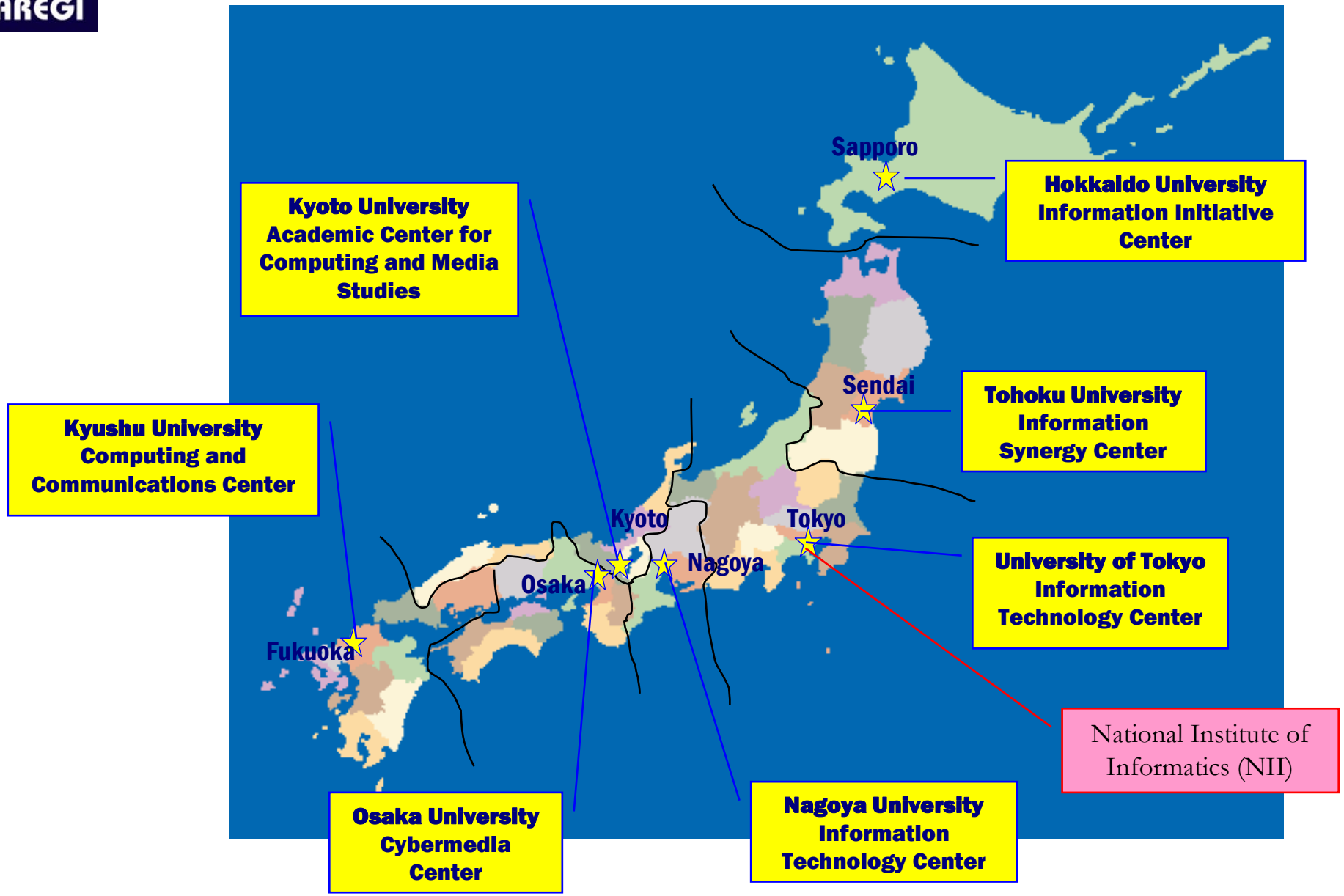
Authentication and Authorization platform for Cyber-Science Infrastructure in Japan

- Targets various applications
  - SSO of Web and Web Services
  - Network Servicec
    - Wireless LAN roaming, VPN, Public IP phone/Web terminals
  - Unified Grid account
- Utilize PKI and Grid/WS Security and Identity Mgmt Technologies

UPKI example: Authentication for inter-campus wireless LAN



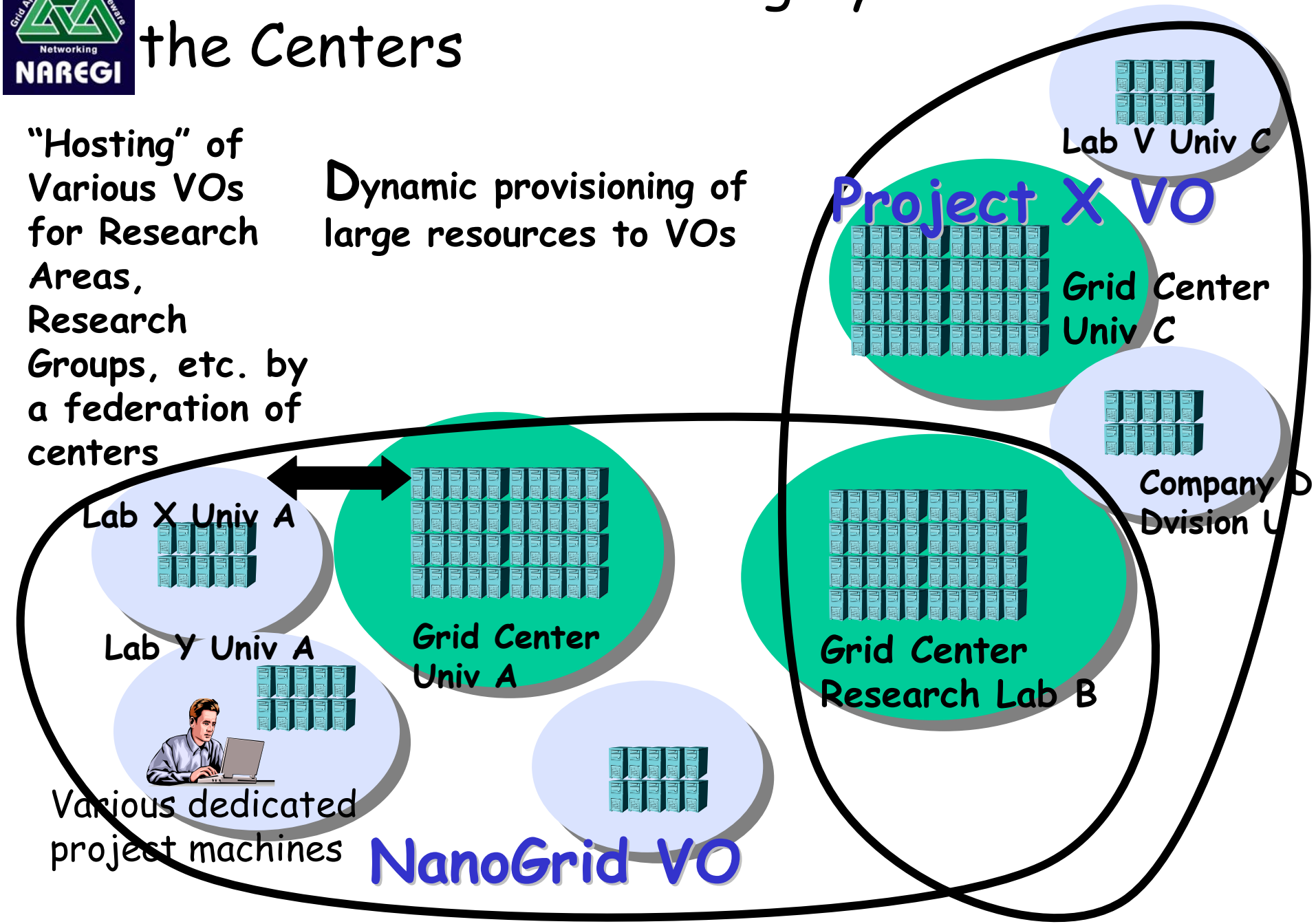
Information Infrastructure Centers in the Seven Universities in JAPAN



The Future: VO Hosting by the Centers

"Hosting" of Various VOs for Research Areas, Research Groups, etc. by a federation of centers

Dynamic provisioning of large resources to VOs



Future Plans: Importance in unifying Grid middleware, esp. VO & user management for international e-Science

Grid Regional Infrastructural Efforts
Collaborative talks on PMA, etc.

All different software stacks

Europe: EGEE, UK e-Science, …

US: TeraGrid, OSG,

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

HEP Grid VO

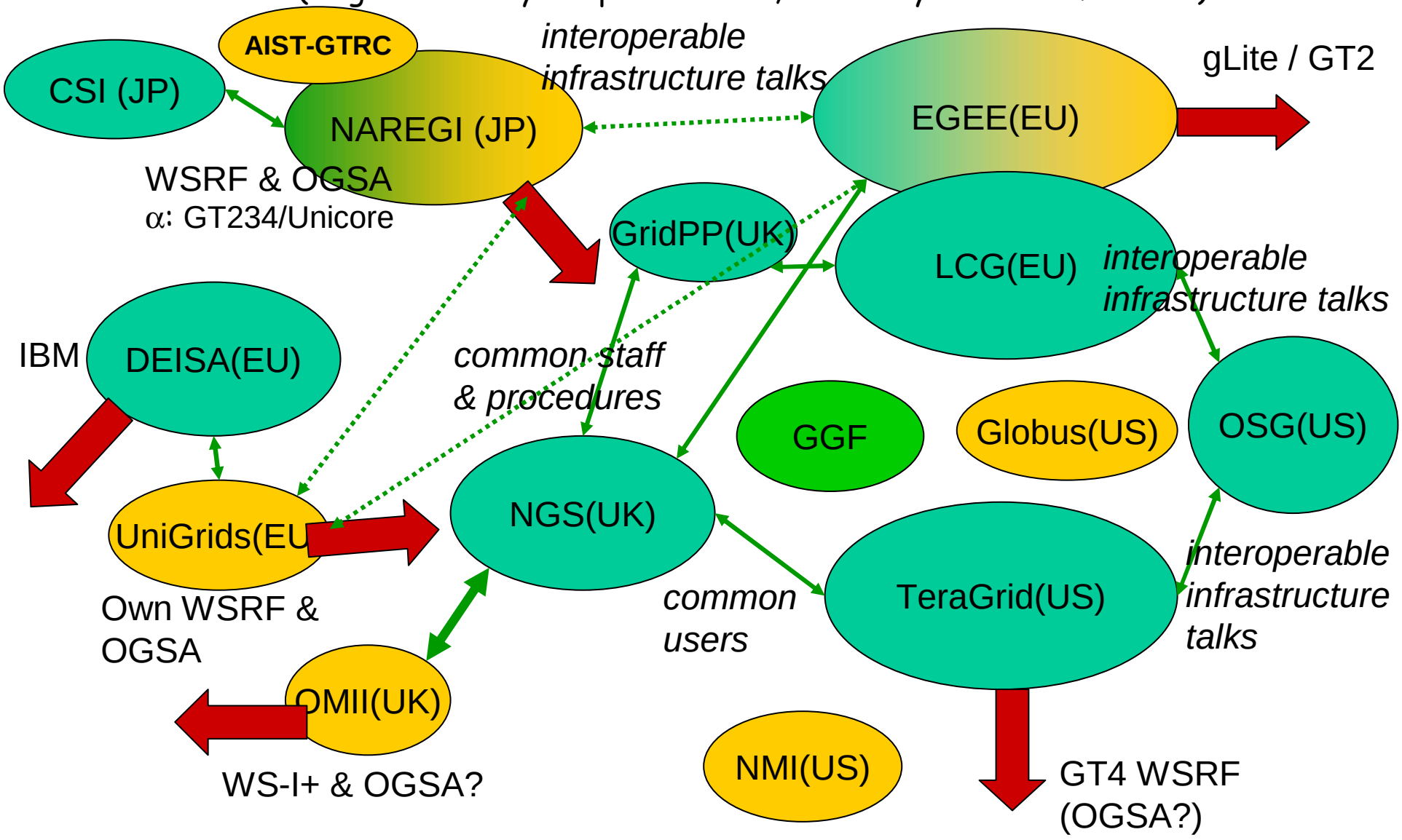
NEES -ED Grid VO

Astro IVO

Requires standardization and some commonality in software platforms

Convergence/Divergence of Project Forces

(original slide by Stephen Pickles, edited by Satoshi Matsuoka)



Summary

- NAREGI & CSI are the basis next generation Japanese Research Grid Infrastructure
- Coordination with EU, US, and AP projects essential (already with EGEE, and continuing with UniGrids)
- Special thanks to David Snelling and his team for help in designing NAREGI Alpha