# Building the CyberScience Infrastructure Community in NAREGI

Satoshi Matsuoka, Professor/Dr.Sci.

Global Scientific Information and Computing Center (GSIC)
Tokyo Inst. Technology
& NAREGI Project National Inst. Informatics

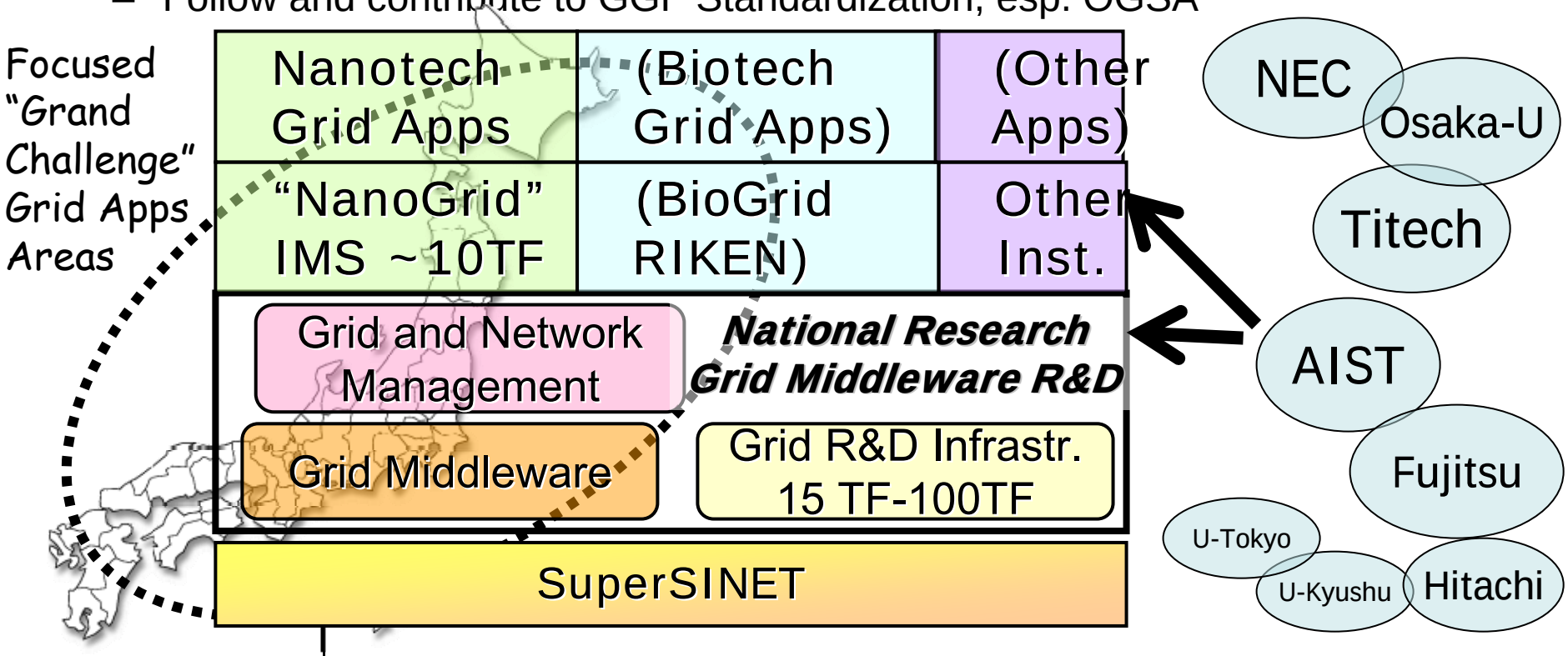
May 7, 2007
OGF20 Campus & Community Grid Workshop
Manchester, UK

**TSUBAME (Swallow)**

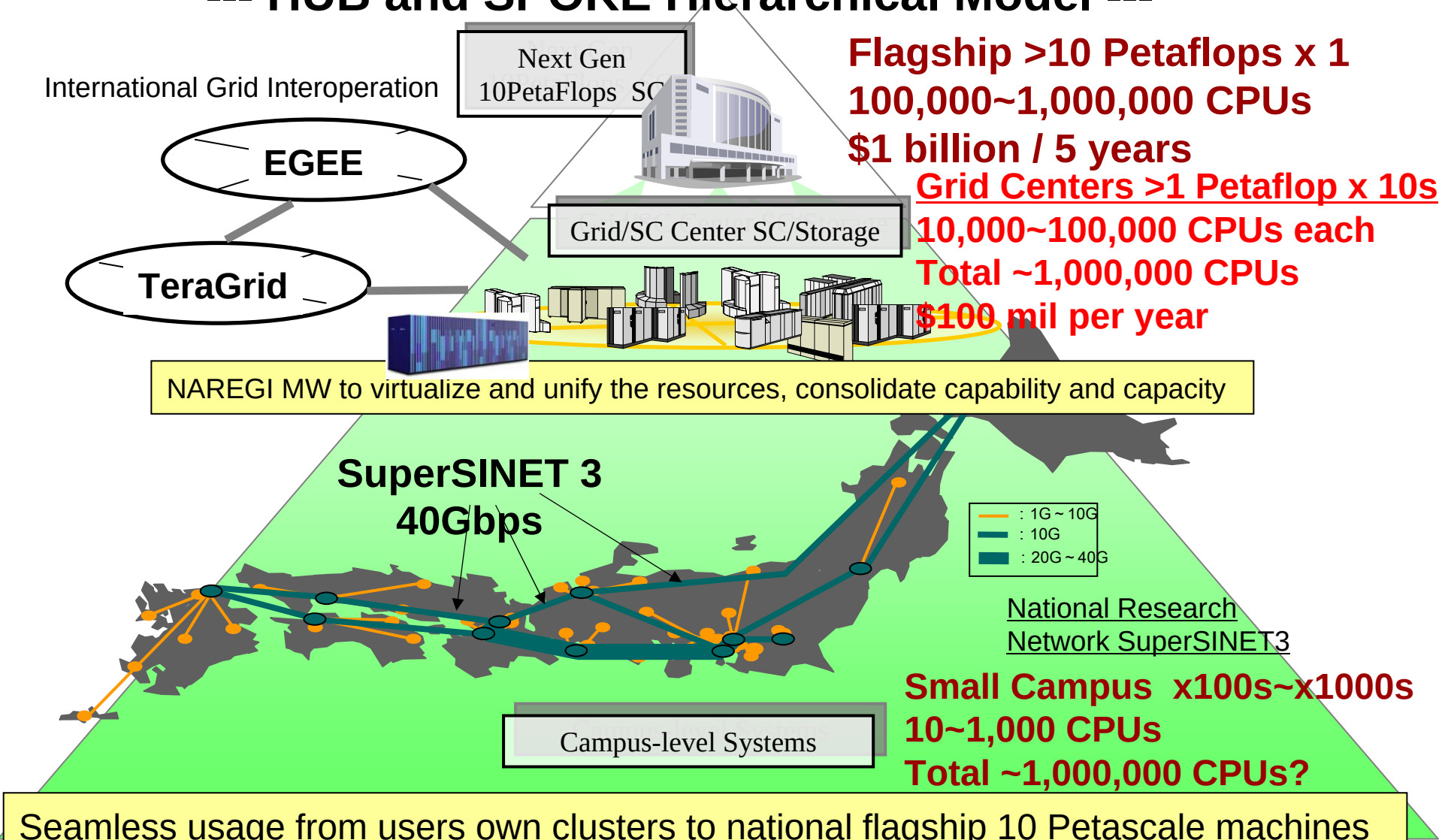
# National Research Grid Infrastructure (NAREGI) 2003-~~2007~~ 2010

## --- A Core of Japanese CyberScience Infrastucture (CSI) ---

- 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 (NII), Co-PI: Kento Aida (new) (NII), S. Sekiguchi(AIST), S. Matsuoka(Titech), S. Shimojo(Osaka-U), M. Aoyagi (Kyushu-U)…
  - Participation by multiple (>= 3) vendors, Fujitsu, NEC, Hitachi, NTT, etc.
  - Follow and contribute to GGF Standardization, esp. OGSA



# CSI Tier-Model of Next Generation Grid/SC Infrastructure
## --- HUB and SPOKE Hierarchical Model ---




**Flagship >10 Petaflops x 1**
**100,000~1,000,000 CPUs**
**$1 billion / 5 years**

**Grid Centers >1 Petaflop x 10s**
**10,000~100,000 CPUs each**
**Total ~1,000,000 CPUs**
**$100 mil per year**

**Small Campus x100s~x1000s**
**10~1,000 CPUs**
**Total ~1,000,000 CPUs?**

Seamless usage from users own clusters to national flagship 10 Petascale machines

**A ~1,000,000 User CyberScience Infrastructure**

# The TSUBAME Production "Supercomputing Grid Cluster", Spring 2006 @ Tokyo Tech GSIC Center

Voltaire ISR9288 Infiniband 10Gbps x2
(DDR next ver.)
~1310+50 Ports
~13.5Terabits/s (3Tbits bisection)

*"Fastest Supercomputer in Asia" 9th on the 28th Top500@47.38TF*

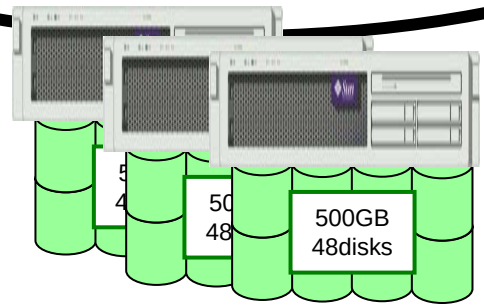
*(Peak 85 TeraFlops)*

Sun Galaxy 4 (Opteron Dual core 8-socket)
10480cores/655Nodes
21.4Terabytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

10Gbps+External Network

Unified IB network

NEC SX-8i
(for porting)





Storage
~~1.0 Petabyte~~ **1.5PB** (Sun "Thumper")
0.1Petabyte (NEC iStore)
**Lustre FS, NFS, CIF, WebDAV (over IP)**
~~50GB/s~~ **70GB/s** aggregate I/O BW

ClearSpeed CSX600
SIMD accelerator
360 boards,
35TeraFlops(Current))


Titech TSUBAME
~80 racks
350m^2 floor area
80 tons
1.2 MW (peak)

# みんなのスパコン

TITECH GRID

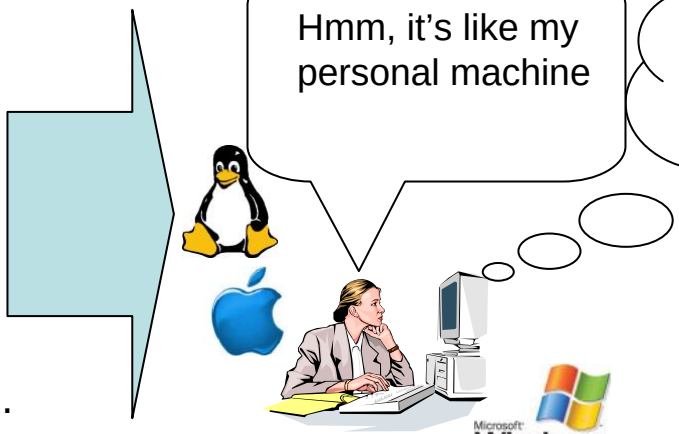
## "Everybody's Supercomputer" as core of Campus Grid and IT Consolidation

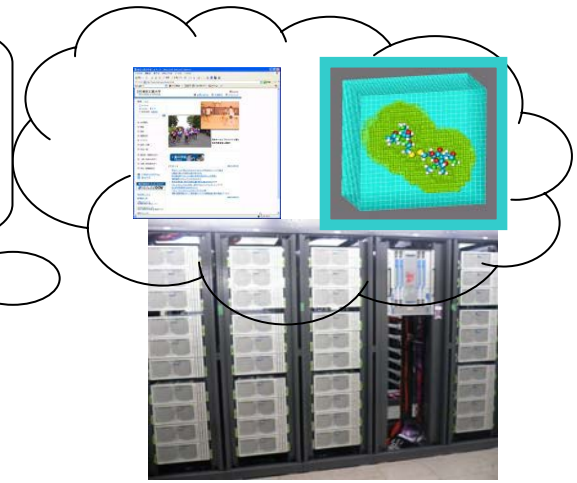
Isolated High-End

Massive Usage Env. Gap

- Different usage env. from
- No HP sharing with client's PC
- Special HW/SW, lack of ISV support
- Lack of common development env. (e.g. Visual Studio)
- Simple batch based, no interactive usage, good UI


**Seamless integration of SCs with *end-user and Enterprise Env.***



*"Everybody's Supercomputer"*

Seamless,Ubiquitous access and usage

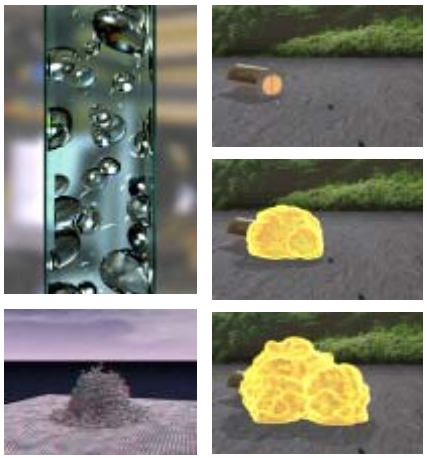
=>Breakthrough Science through Commoditization of Supercomputing and Grid Technologies

# みんなのスパコン

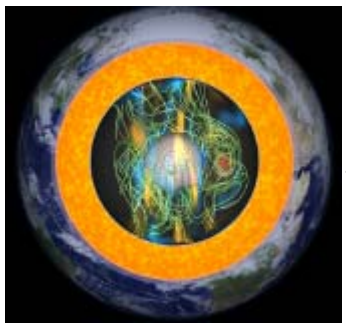
## Grand Challenge Supercomputing @ Titech

## 100 Teraflops-scale computing with Petascale Storage

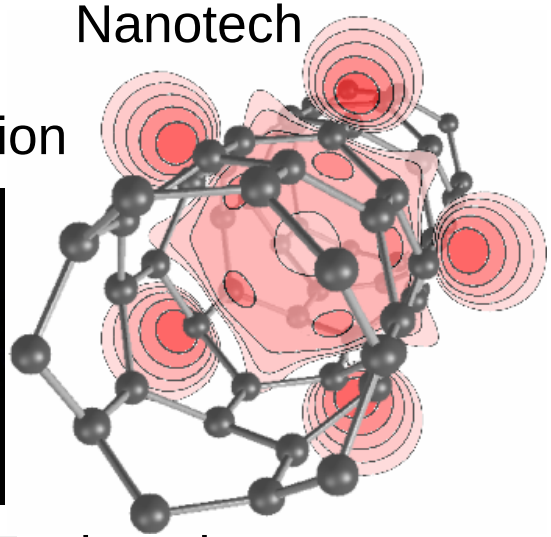
CFD



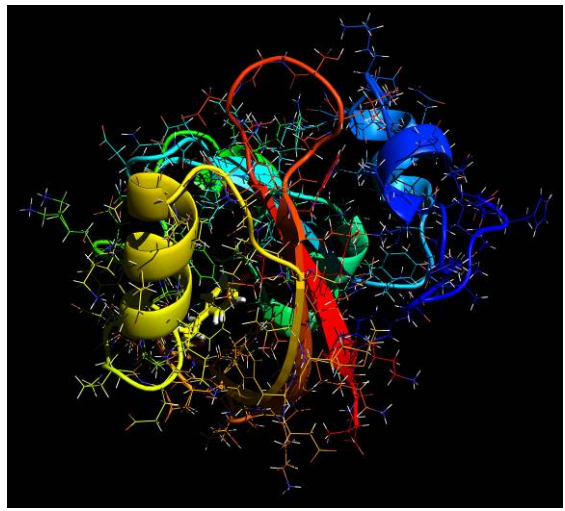
EMF Simulation



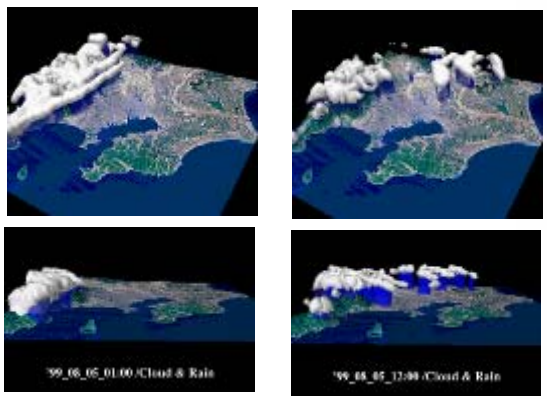
Nanotech



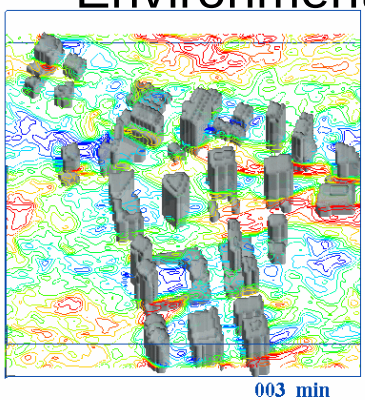
Bioinformatics

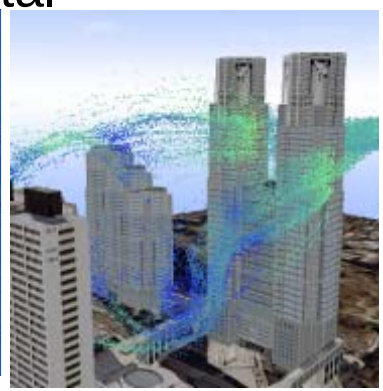


Weather Prediction

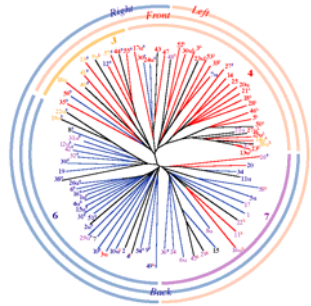



Civil Engineering
Environmental



Bio-simulation
+ Bioinformatics



# みんなのスパコン

TITECH GRID

## TSUBAME General Purpose DataCenter Hosting

***As a core of IT Consolidation***

***All University Members == Users***

***No more private servers & clusters in closets***

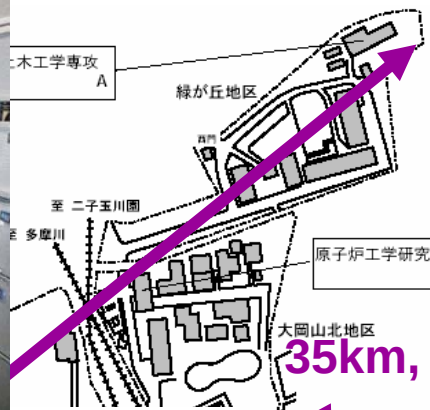
***> 10,000 users on campus***

- Various Application Portals for Edu and Research
- Campus-wide AAA Sytem (April 2006)
  - 50TB (for email), 9 Galaxy1 nodes
- Campus-wide Storage Service (NEST)
  - 10s GBs per everyone, Research Reposit
- CAI, On-line Courses (OCW)
- Administrative Hosting (VEST)

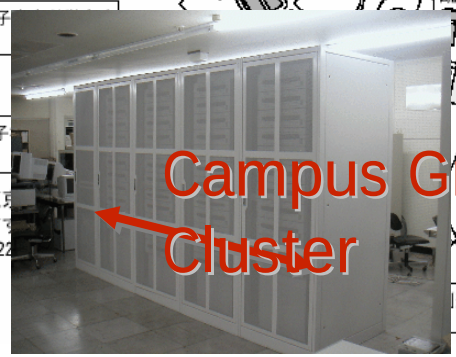

# Titech Campus Grid 2006
# - An x86 “DataCenter” Grid -

- **~13,000 CPUs, 90 TFlops, ~26 TBytes Mem, ~1.1 PBytes HDD**
- ***All Hosted at GSIC: No more private servers & clusters in closets, same as the modern Internet***





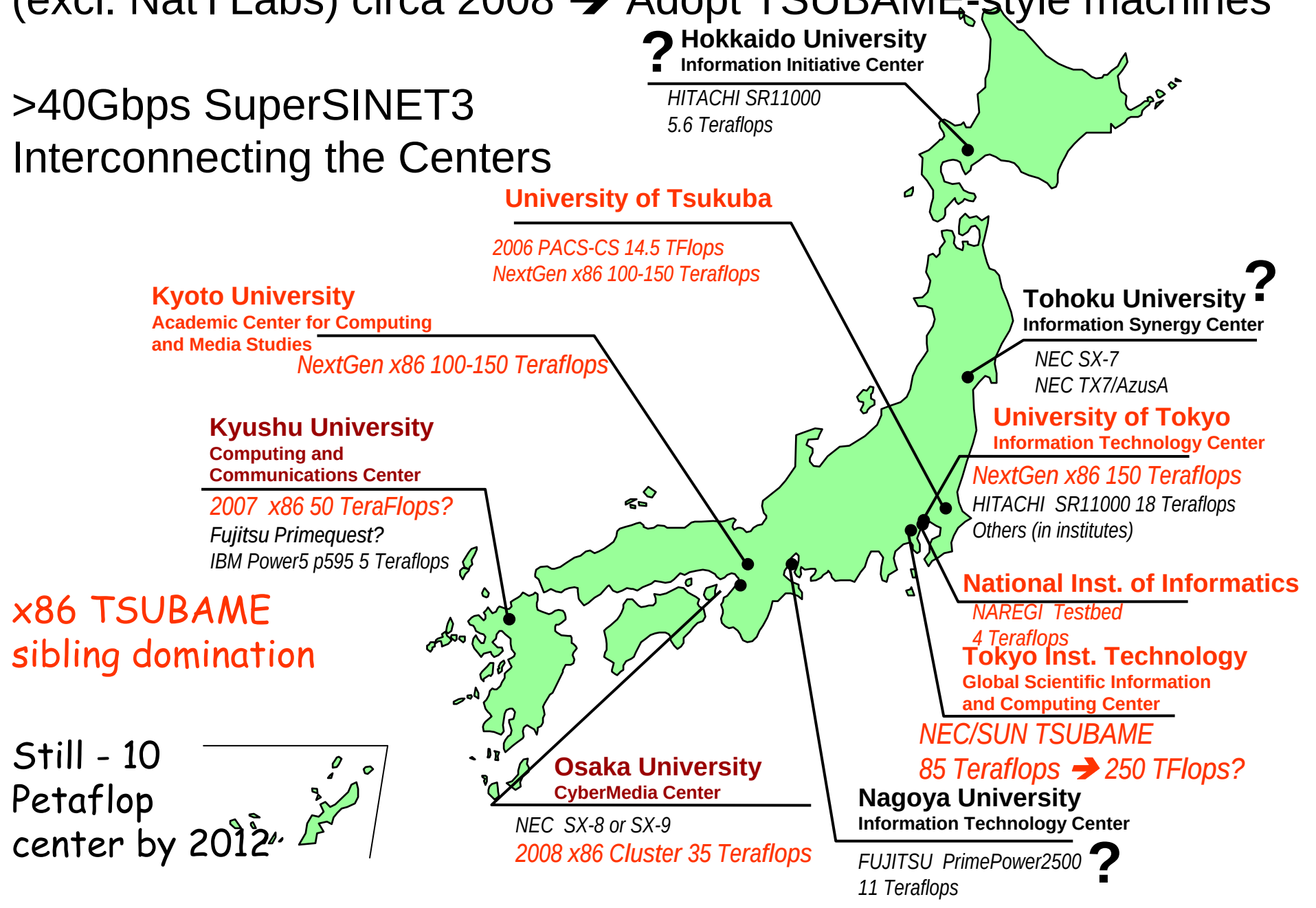





すずかけ台

# Japan's 9 Major University Computer Centers (excl. Nat'l Labs) circa 2008 ➔ Adopt TSUBAME-style machines

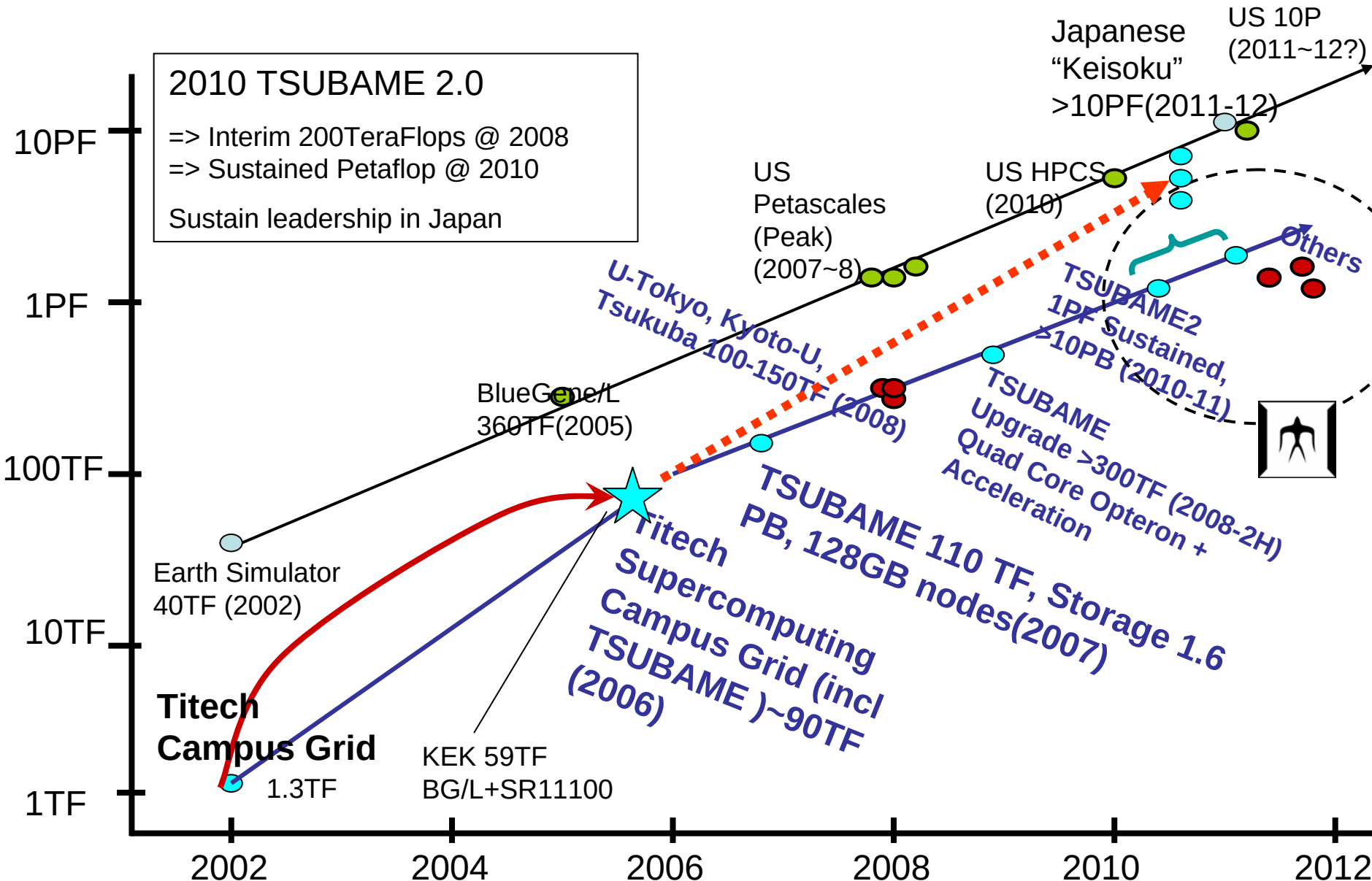
>40Gbps SuperSINET3
Interconnecting the Centers




x86 TSUBAME
sibling domination

Still - 10
Petaflop
center by 2012

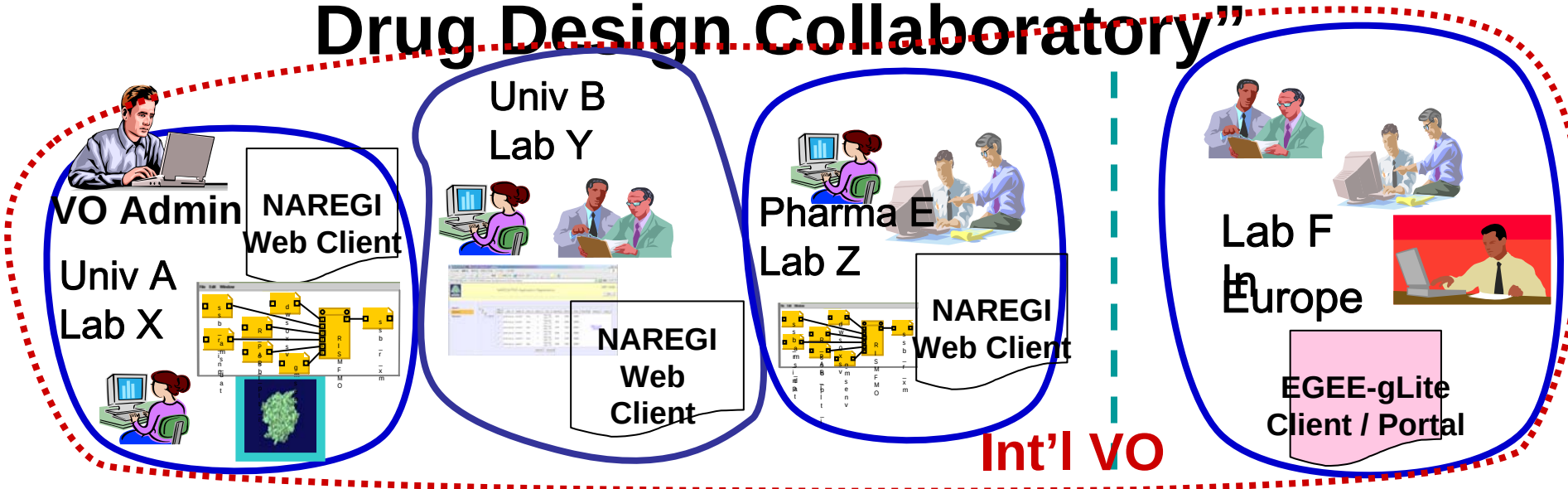
# Scaling Towards Petaflops



# VO Operations in the CyberScience Infrastructure (CSI)

- A individual-globally identifiable ID
  - Everyone MUST have a verifiable electrical ID
  - UPKI (University PKI) effort in CSI, Shibboleth, etc.
- VO's register with a designated grid centers, just as one would register domains
  - Most cases, Grid centersw would host VO services
- I.e., SC/grid centers are resource providers as well as service providers for generic & VO-specific VO services (Web Services, Web DB, Web Portalなど)
  - Web 2 and beyond community hosting vision

# CSI VO Vision Examle: “Int’l Computational Drug Design Collaboratory”




**Virtualization of Resoruces and Services via NAREGI MW / Web Services**

Univ A Lab X Private Cluster

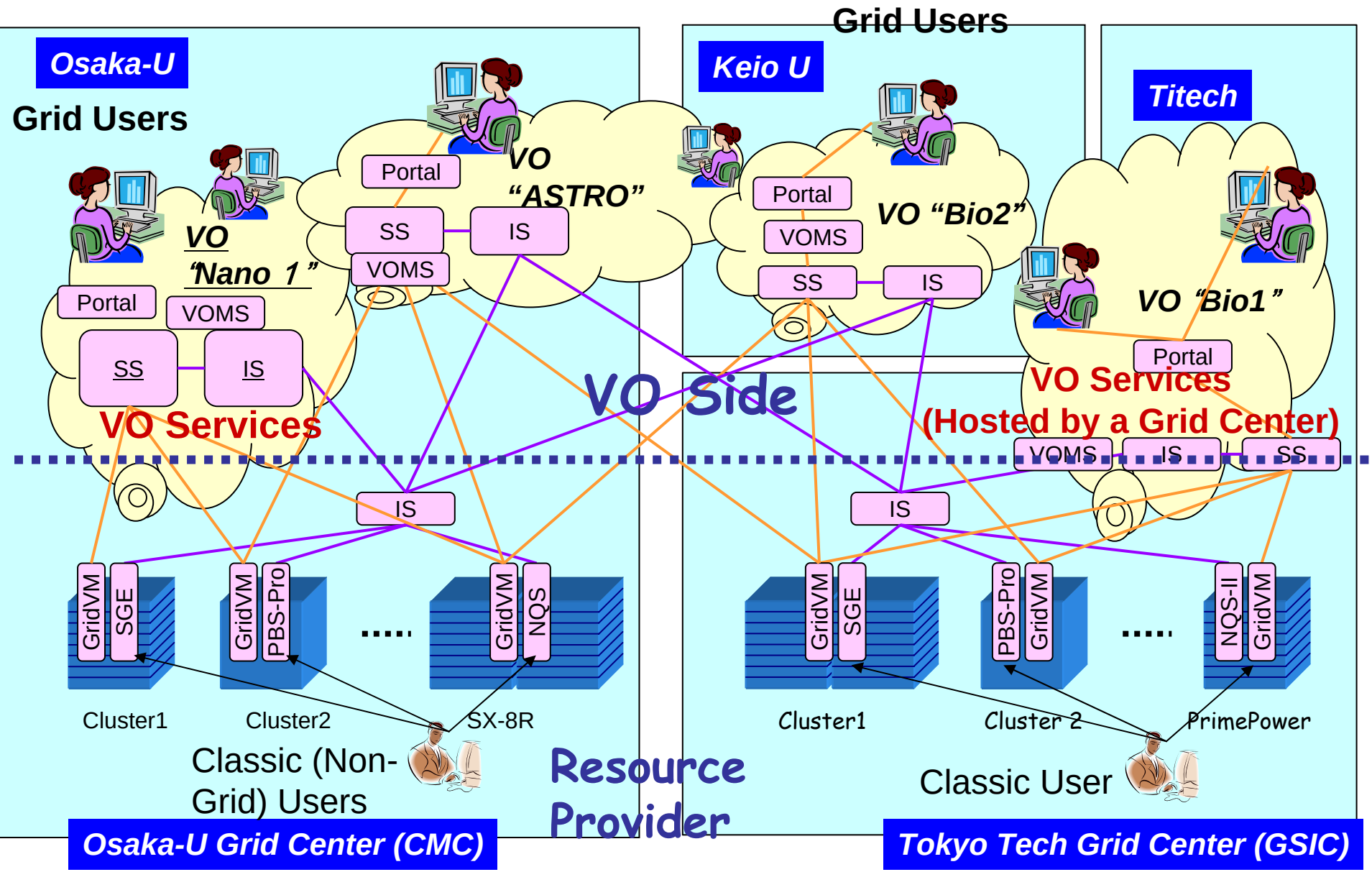
National SC/grid center at Univ A

National Lab C SC/grid center

European Grid

**CSI Virtual Computation, Resource providers**

# NAREGI β2 Operational Model



# NAREGI β1 – Release May, 2006

- **Objective：Functional test for ver.1.0**
- Platform Shift (Unicore=> Globus 4-WSRF)
  - Set of WSRF (Web Services Resource Framework) Components
- OGF OGSA-EMS based resource management
  - Reservation/Co-Allocation/Co-Scheduling framework
- VO management built on EGEE-VOMS
- OGF-ACS WS application deployment
- Data grid features based on grid filesystem (GFarm)
- NAREGI-WFML description of complex workflow
  - Including co-scheduled resources
- GridMPI/GridRPC and other programming frameworks
- MyProxy-based Security and ID management w/session management, IGTF “ready”
- Being test-deployed, currently release 1.0.2
- Many many patches / bug fixes applied ☺

# NAREGI Beta 2 - v.1.0 Highlights

- Beta2-Production Release Candidate (2Q 2007)
- Lots of bug, performance & stability fixes
- Stable WS(RF) components and APIs (+ Globus 4.0.3)
- RPM and Dynamic, VM-based deployment
- VO and “Resource Provider” decoupling for multiple VO management by VOs and Centers
- Integration of NAREGI WF and Ninf-G GridRPC
- More BQ and systems support (+PBS Pro, LoadLeveler)
  - NEC SX-NQS, SGE, Fujitsu NQS II… (Condor?)
- Flexible Job submission and WF management
  - Non-grid jobs, non-reserved jobs, various WF tools
- EGEE-GIN Interoperation (new)
- Various Administration and Logging Tools
- Support from dedicated NAREGI support team
- Large-scale deployments @ Osaka-U, Titech... (beta2)

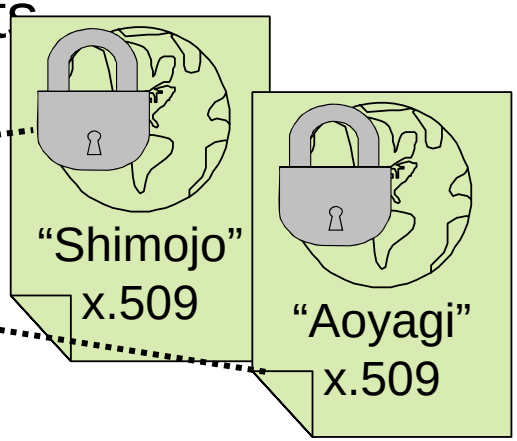
# NAREGI β2 Mapping to "Groups" center accounts

**VOs are mapped to local account groups**

**Centers retain their local accounting policy**

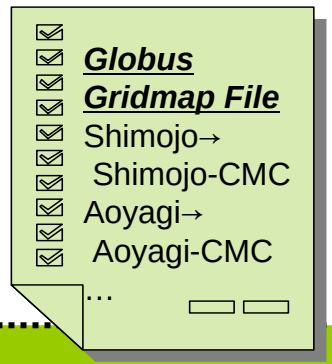
Virtual Org. "Nano1"
Leader (PI): Shimojo
Member1: Shimojo
Member2: Aoyagi
Member3: Okabe
Member4: Ishikawa
Hosted by: Osaka-U CMC / NII

**VO**




**Cert from IGTF-ceritfied CA**

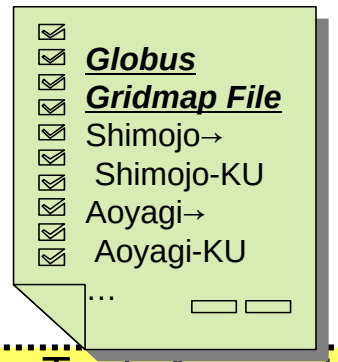
**Delegation of Funds**




Osaka-U CMC Account Group
"Nano1-CMC"
Local PI: Aoyagi-CMC
Member1: Shimojo-CMC (local Unix ID)
Member2: Aoyagi-CMC
Member3: Okabe-CMC
Member4: Ishikawa-CMC

**LOCAL Acct Group**

**Delegation of Funds**

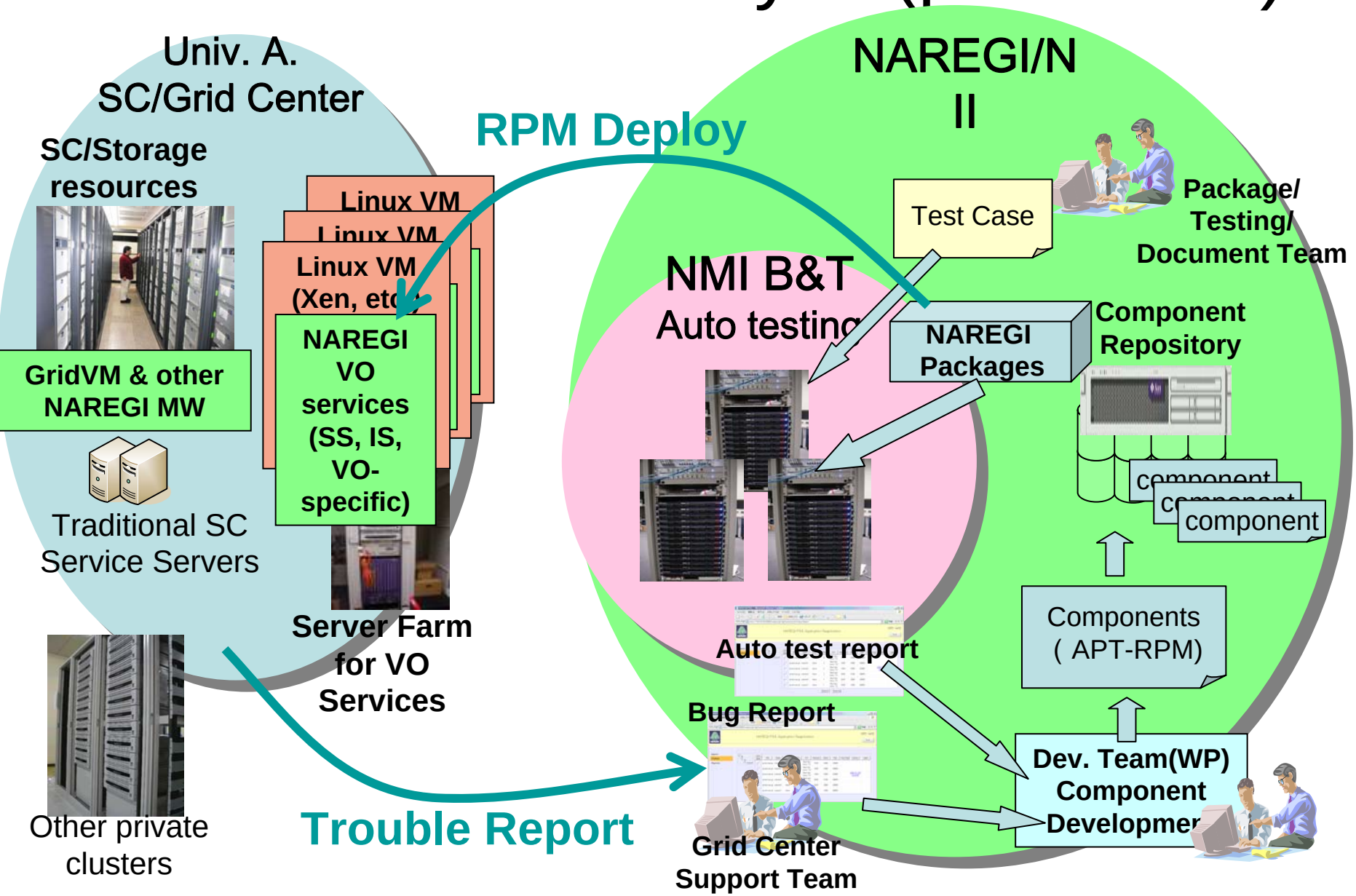



Centers maintain VO→Account Group mapping (missing feature in Globus)

Tokyo Tech Account Group
"Nano1-Titech"
Local PI: Okabe-Titech
Member1: Shimojo-Titech
(local Unix ID)
Member2: Aoyagi-Titech
Member3: Okabe-Titech
Member4: Ishikawa-TItech

**LOCAL Acct Group**

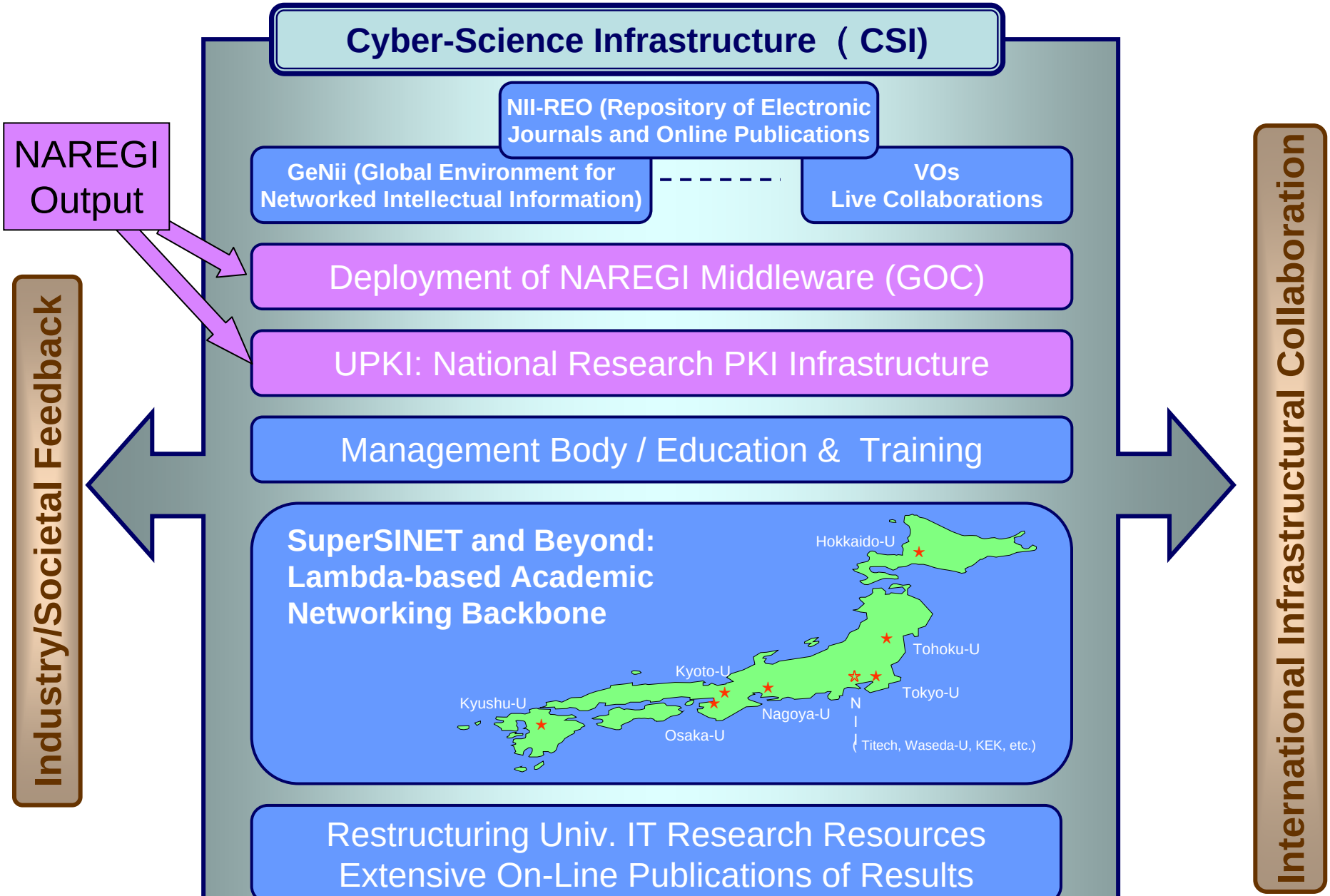
# NAREGI MW Lifecycle(β2 and v1)



# NAREGI EGEE Interoperation

- GLite ⇔ NAREGI MW Interoperation component as standard in NAREGI beta2 distribution
- Results of GIN Interoperation Efforts

# Japanese CyberScience Infrastructure Project



# Upscaling the Resources to a Petascale Grid

~10TF

Capacity

#Users

~1000

~1GF

> x1000

~10PF

Capacity

x10^6

x10^7

#Users

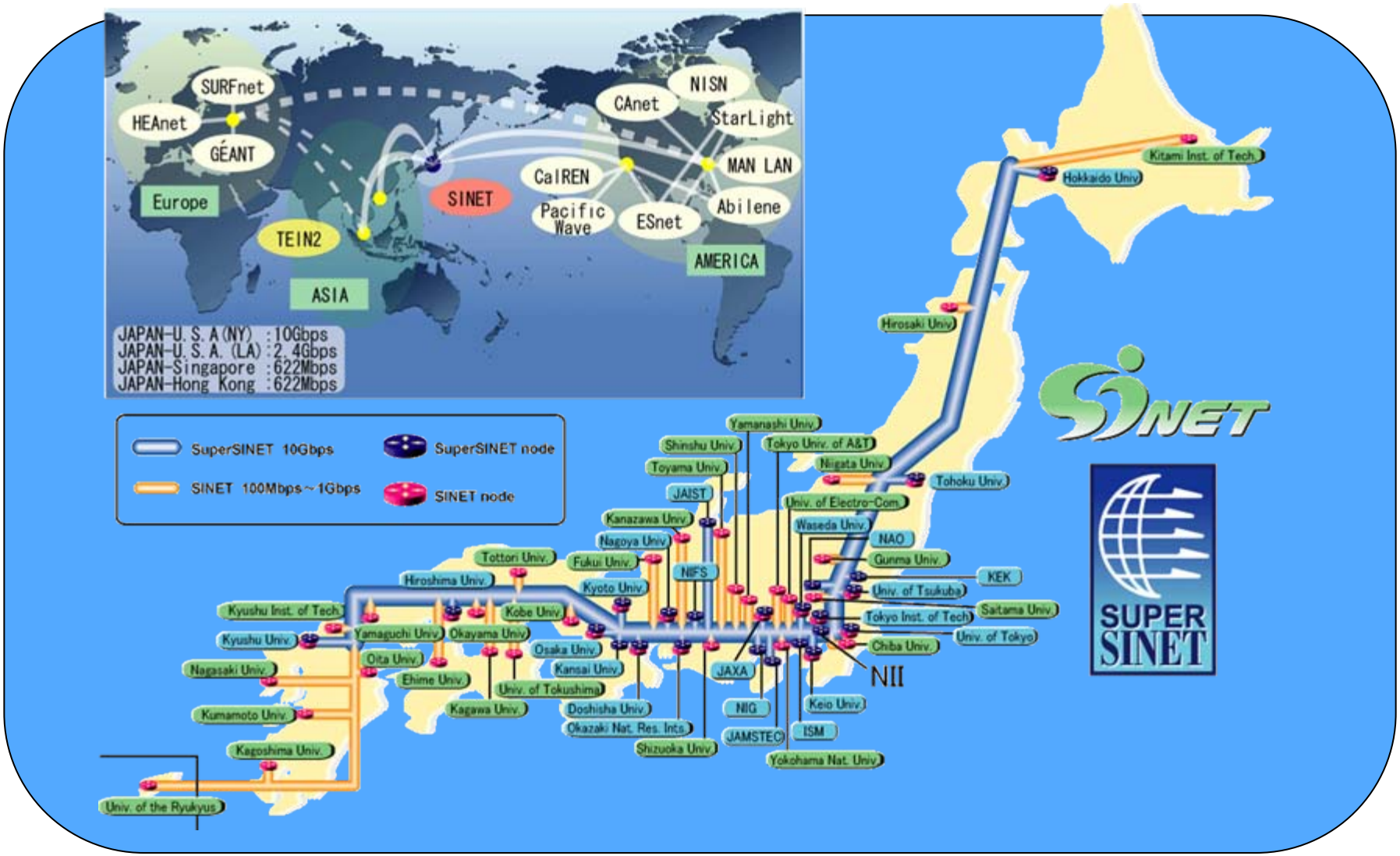
~1,000,000

~1GF

# Backup Slides

# Super SINET3 (new!)

## Dynamic L1/L2/L3 provisioning 40 Gbps Backbone

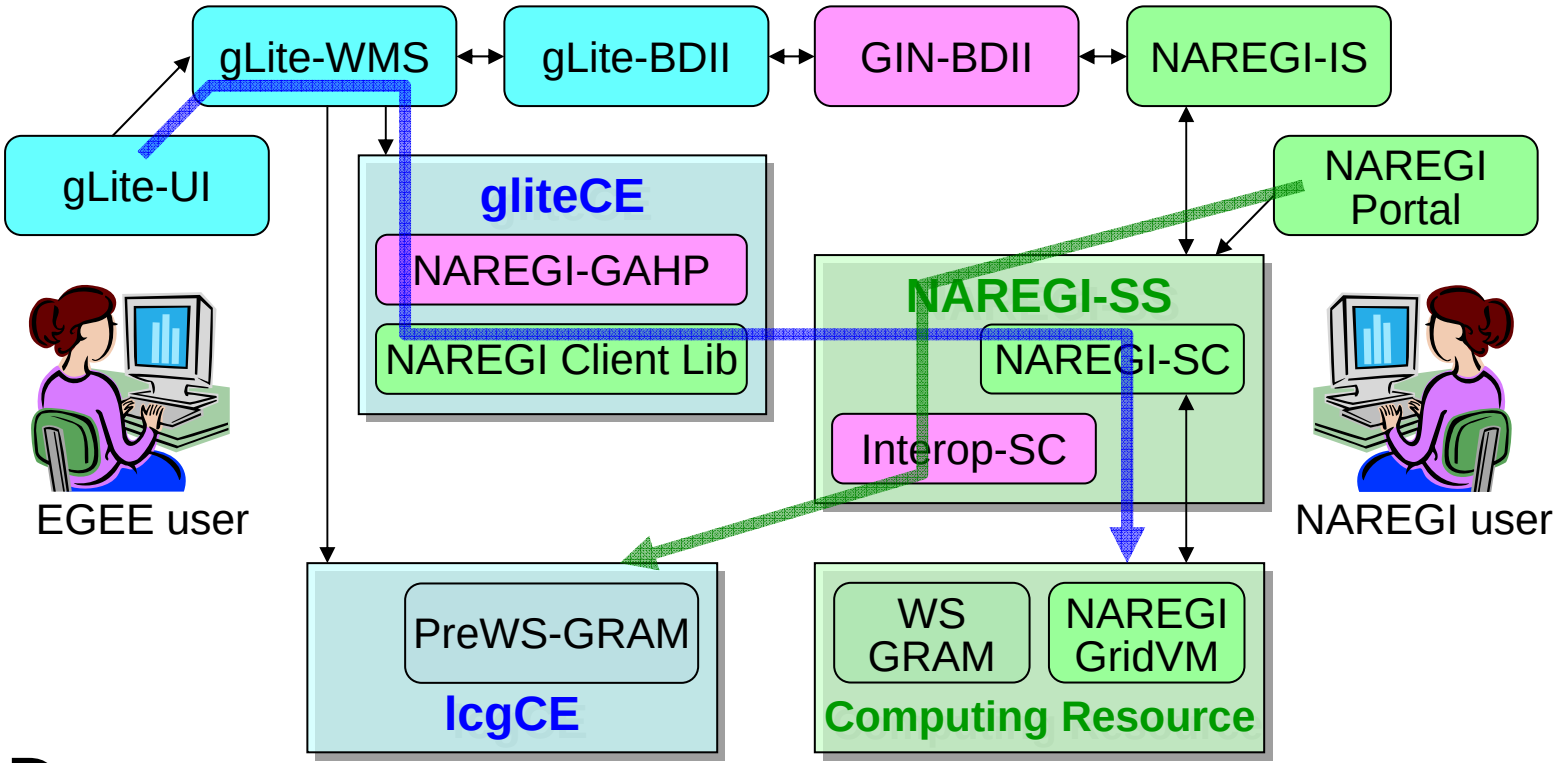


# GIN (Grid Interoperation Now)

- An activity of OGF for interoperation among production grids
- Major grid projects are participating
  - EGEE, NAREGI, UK National Grid Service, NorduGrid, OSG, PRAGMA, TeraGrid, ...
- Trying to identify islands of interoperation between production grids and grow those islands
- Areas
  - GIN-auth: Authorization and Identity Management
  - GIN-data: Data Management and Movement
  - GIN-jobs: Job Description and Submission
  - GIN-info: Information Services and Schema
  - GIN-ops: Operations Experience of Pilot Test Applications

# GIN-jobs: NAREGI-EGEE Architecture

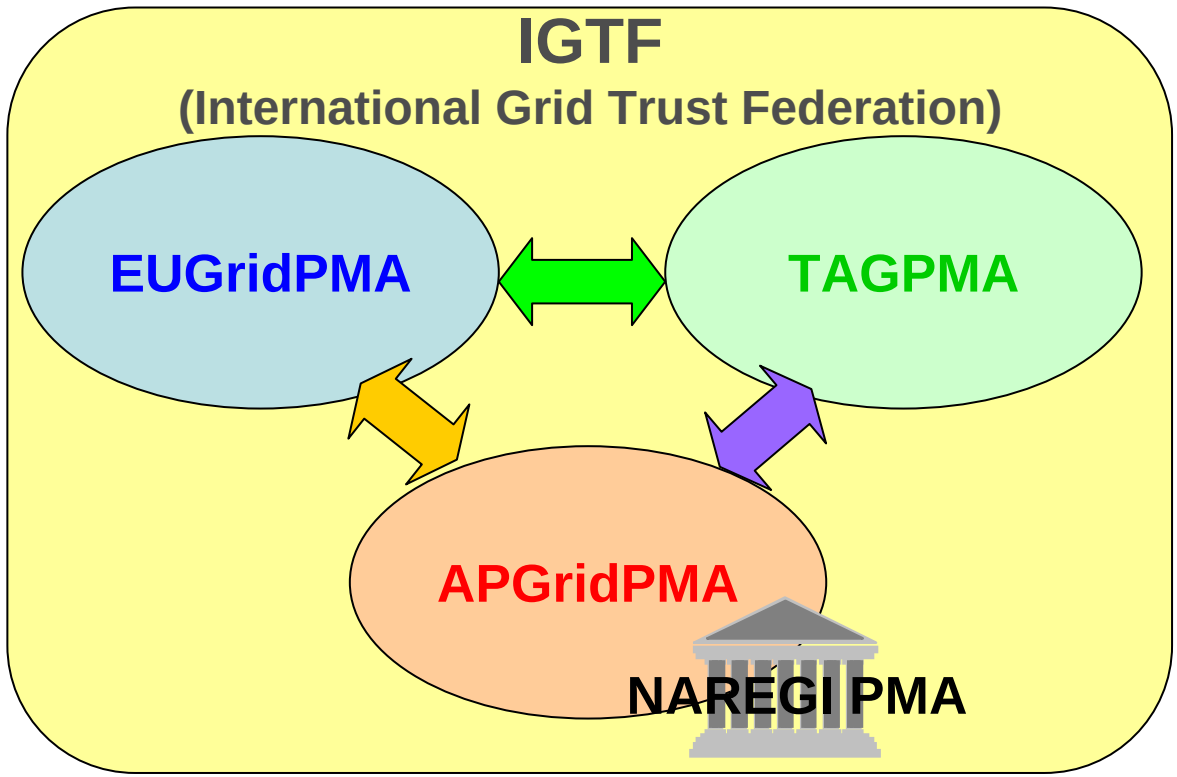
- Architecture




- Demo
  - NAREGI → EGEE: using NAREGI Workflow
  - EGEE → NAREGI: using glite WMS commands

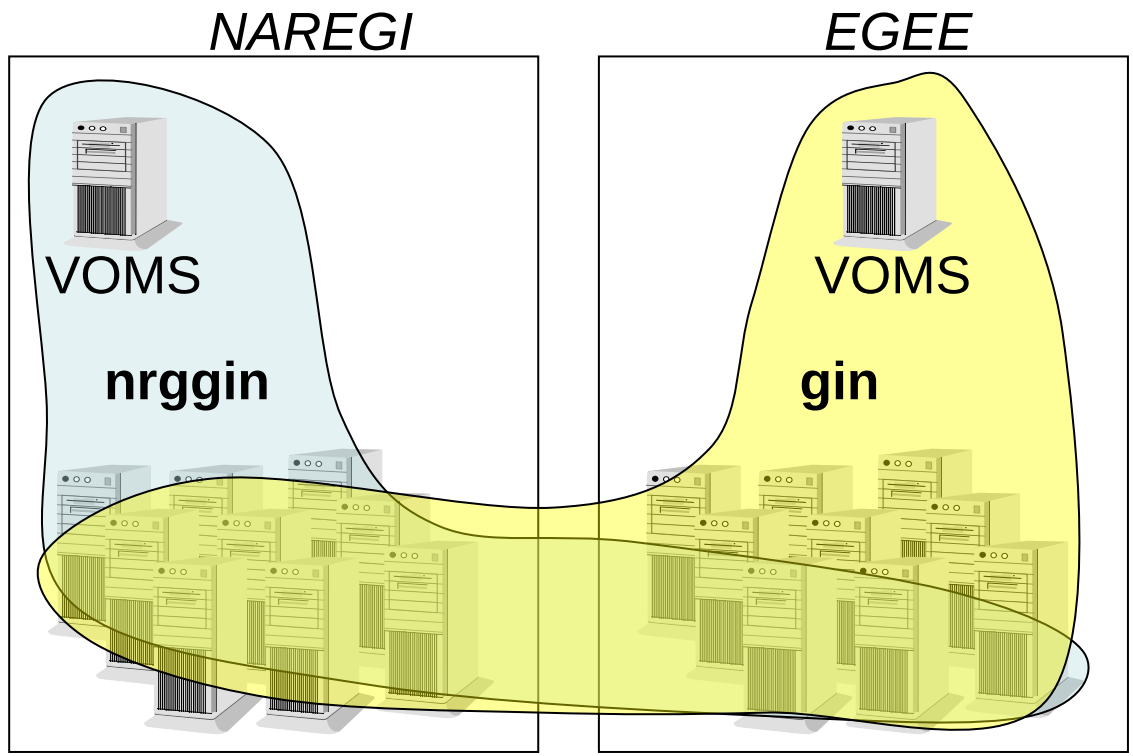
# Authentication

- IGTF is framework of International Grid Trust Federation.
- IGTF consists of APGridPMA, EUGridPMA and TAGPMA.
- NAREGI CA joined the APGrid PMA.
- NAREGI CA has been approved as a production-level CA by APGridPMA.
- GSI compliant with x.509 proxy certificates for authentication.
- It has become available to use grid computing easily on the worldwide Internet by IGTF.



# VO Management

- The GIN VO is a VOMS service.
- NAREGI uses VOMS as  VO management system.
- Transport of supported authorization attributes via VOMS extensions.
- VO names are expected to abide by the VO naming conventions described in GIN VO Naming in order to avoid name conflicts between grids.
- All members of GIN VO should observe AUP(Acceptable Use Policy).

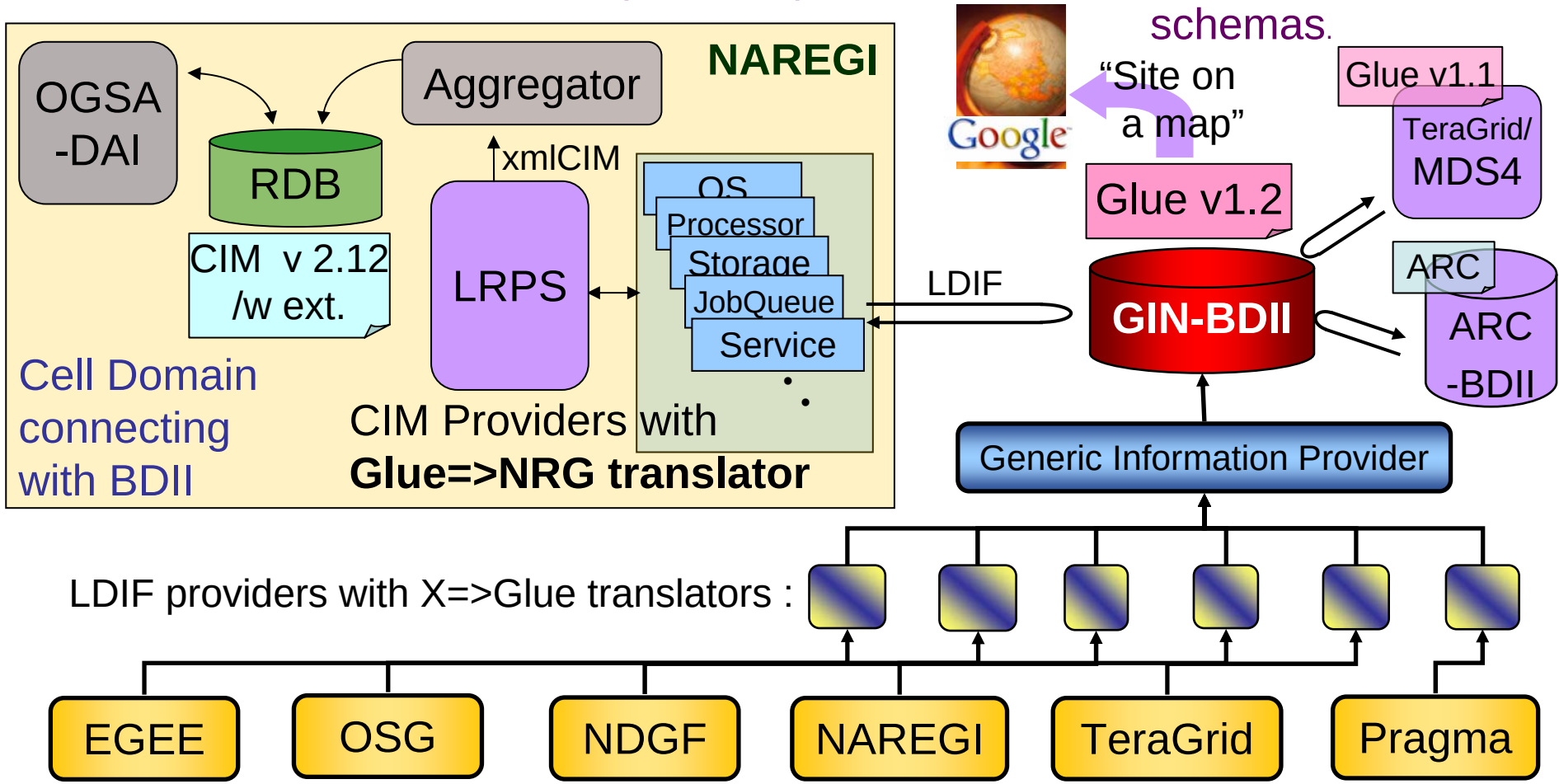



reference
*http://forge.gridforum.org/sf/wiki/do/viewPage/projects.gin/wiki/GINAuth*

# GIN-info: Architecture

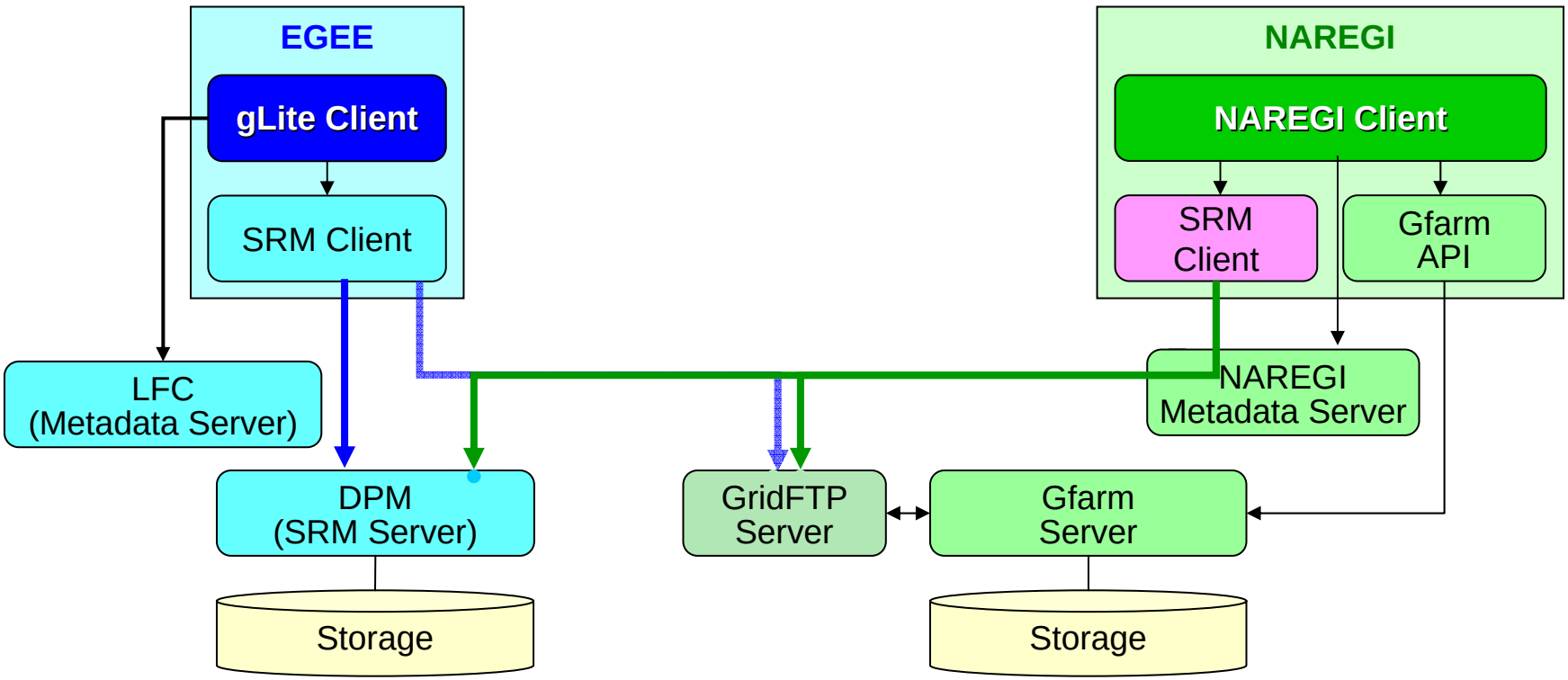
All of grid information can be retrieved by each of grid in its fashion WRT resource description schema, data format, query language, client API, …

Each information service in grid acts as an information provider for the other and translator embedded in the provider performs conversion between different schemas.



# GIN-data: Architecture

- NAREGI and EGEE gLite clients can access to both data resources (e.g., bi-directional file copy) using SRM interface.
- GridFTP is used as its underlying file transfer protocol.
- File catalog (metadata) exchange is planned.



# NAREGI GIN Summary

- NAREGI developed EGEE-NAREGI island as an activity of GIN
  - Bilateral information exchange
  - Bilateral job submission
  - Bilateral file exchange
  - Interoperable security properties
- Next steps
  - Improve interoperation interfaces and functions
    - WS-GRAM, BES, JSDL, …
  - Grow the island with other EGEE partners
  - KEK will use NAREGI-EGEE interoperation environment for their high energy physics calculations

# CSI / NAREGI VO Communities

- ~= User Groups, unit of resource policies
- Various Vo Examples
  - International Research Consortium
  - A research area and SC user group thereof
  - Members of (large) research grants
  - Industry-Academia collaboration group
  - User group of a particular application hosted by a center
  - A research lab in a university
  - Students of across-Institutional class
  - ｡ ｡ ｡

**Full Blown Reseach SNS**

**SC Usage bu Individuals**

↓

**Group based social networks**

# NAREGI's outreach to all research communities