MW Activities in the NAREGI Japanese (Research) Grid Project

Satoshi Matsuoka

Professor, Global Scientific Information and Computing Center,

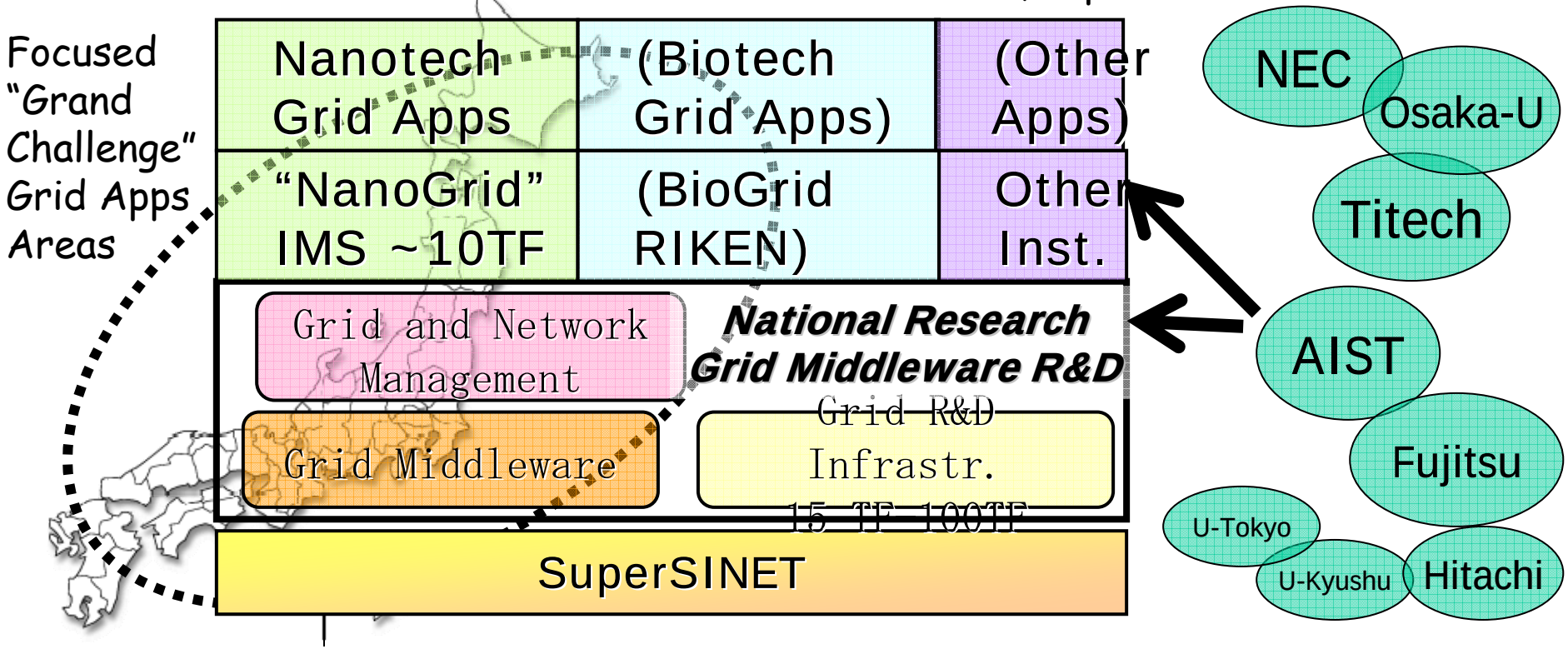
Deputy Director, NAREGI Project

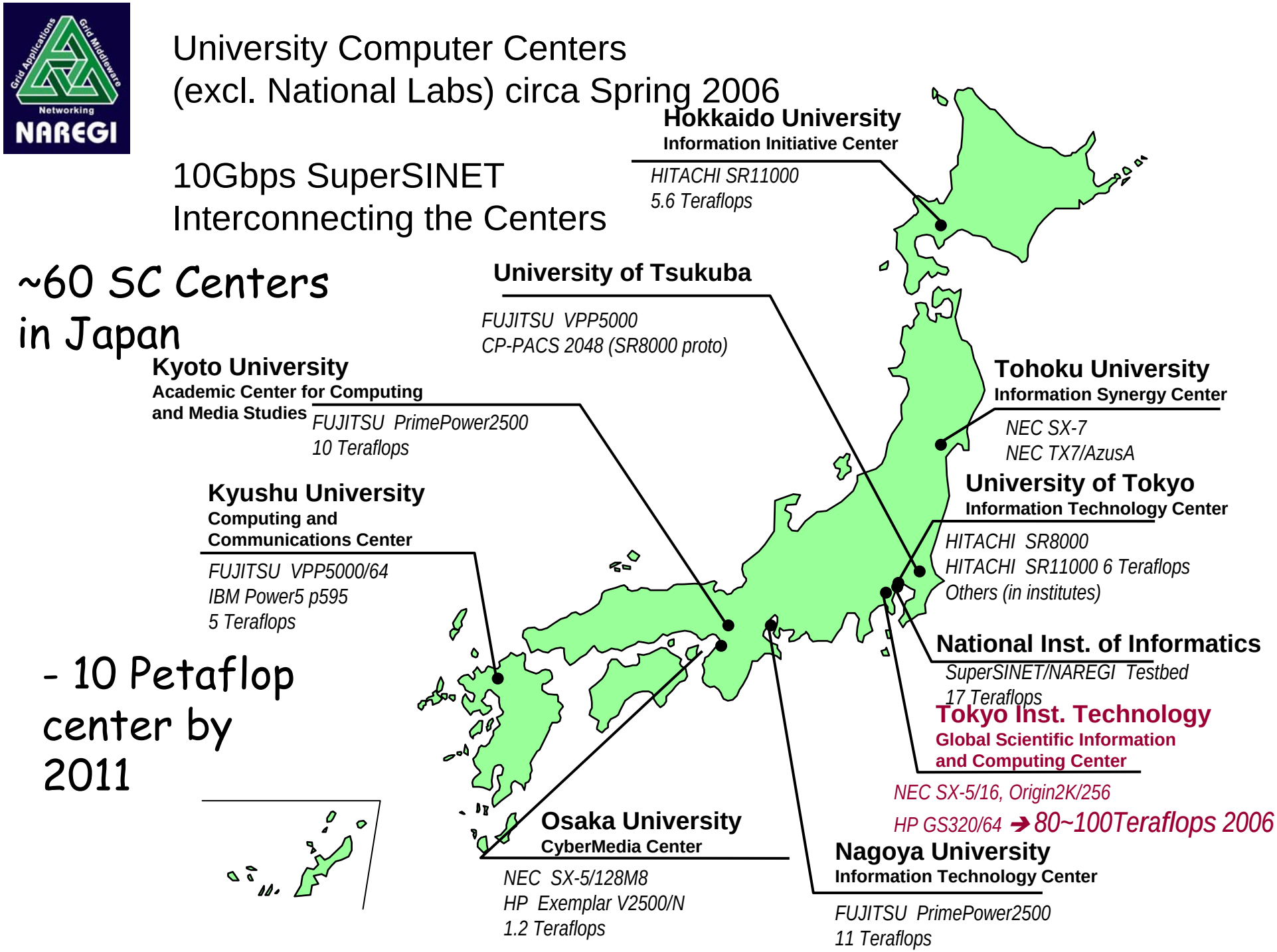
Tokyo Institute of Technology / NII

http://www.naregi.org

National Research Grid Infrastructure (NAREGI) 2003-2007

- 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, Fujitsu, NEC, Hitachi, NTT, etc.
  - Follow and contribute to GGF Standardization, esp. OGSA




University Computer Centers
(excl. National Labs) circa Spring 2006
10Gbps SuperSINET
Interconnecting the Centers
~60 SC Centers in Japan
- 10 Petaflop 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

The New "Supercomputing Campus Grid" Core System, Spring 2006

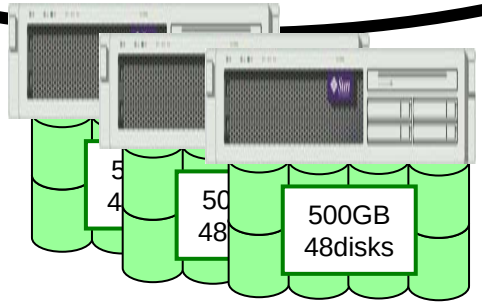
Voltaire Infiniband
10Gbps x DDR x 2 Lanes
x ~700Ports
28 Terabits/s
Bisection BW

655 Sun Galaxy 4 (Opteron
Dual core 16-Way)
10480core/655Nodes
50TeraFlops
OS Linux
(Future) Solaris, Windows
NAREGI Grid MW

10Gbps+External
Network

85 Teraflops
1.1 PB Storage
28 Tb/s BW
120,000 FP Exec units

NEC SX-8
Small Vector
Nodes (under
plan)



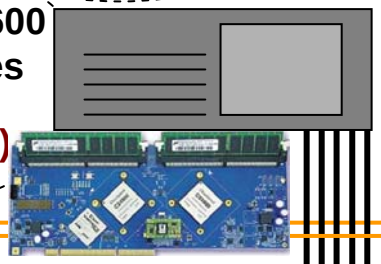
Storage
1 Petabyte (Sun "Thumper")
0.1Petabyte (NEC iStore)
Lustre FS, NFS (v4?)

ClearSpeed CSX600
SIMD accelerator
360 boards
35TeraFlops
➔
60TeraFlops(future)

NEC/Sun Campus Supercomputing Grid: Core Supercomputer Infrastructure @ Titech GSIC - operational late Spring 2006 -

SunFire (TBA)
655nodes
16CPU/node
10480CPU/50TFlops (Peak)
Memory: **21.4TB**

ClearSpeed CSX600
Initially 360nodes
96GFlops/Node
35TFlops (Peak)



85 TeraFlops (50 Scalar + 35 SIMD-Vector)
> 100,000 FP execution units, 70 cabinets

max config ~4500 Cards, ~900,000 exec. units,
500Teraflops possible,
Still 655 nodes, ~1MW

InfiniBand Network Voltaire ISR 9288 x 6

1400Gbps Unified & Redundunt Interconnect

External
10Gbps Switch
Fabric

External
Grid
Connectivity

24+24Gbps
bidirectional

200+200Gbps
bidirectional

42 units

500GB
48disks

500GB
48disks

500GB
48disks

FileServer

FileServer

Storage Server A
Sun Storage (TBA), all HDD
Physical Capacity 1PB, 40GB/s

Storage B
NEC iStorage S1800AT
Phys. Capacity 96TB RAID6
All HDD, Ultra Reliable

Total 1.1PB

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

Grid Regional Infrastructural Efforts
Collaborative talks on PMA, etc.

Different software stacks but interoperable

Europe: EGEE, UK e-Science, …

US: TeraGrid, OSG,

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

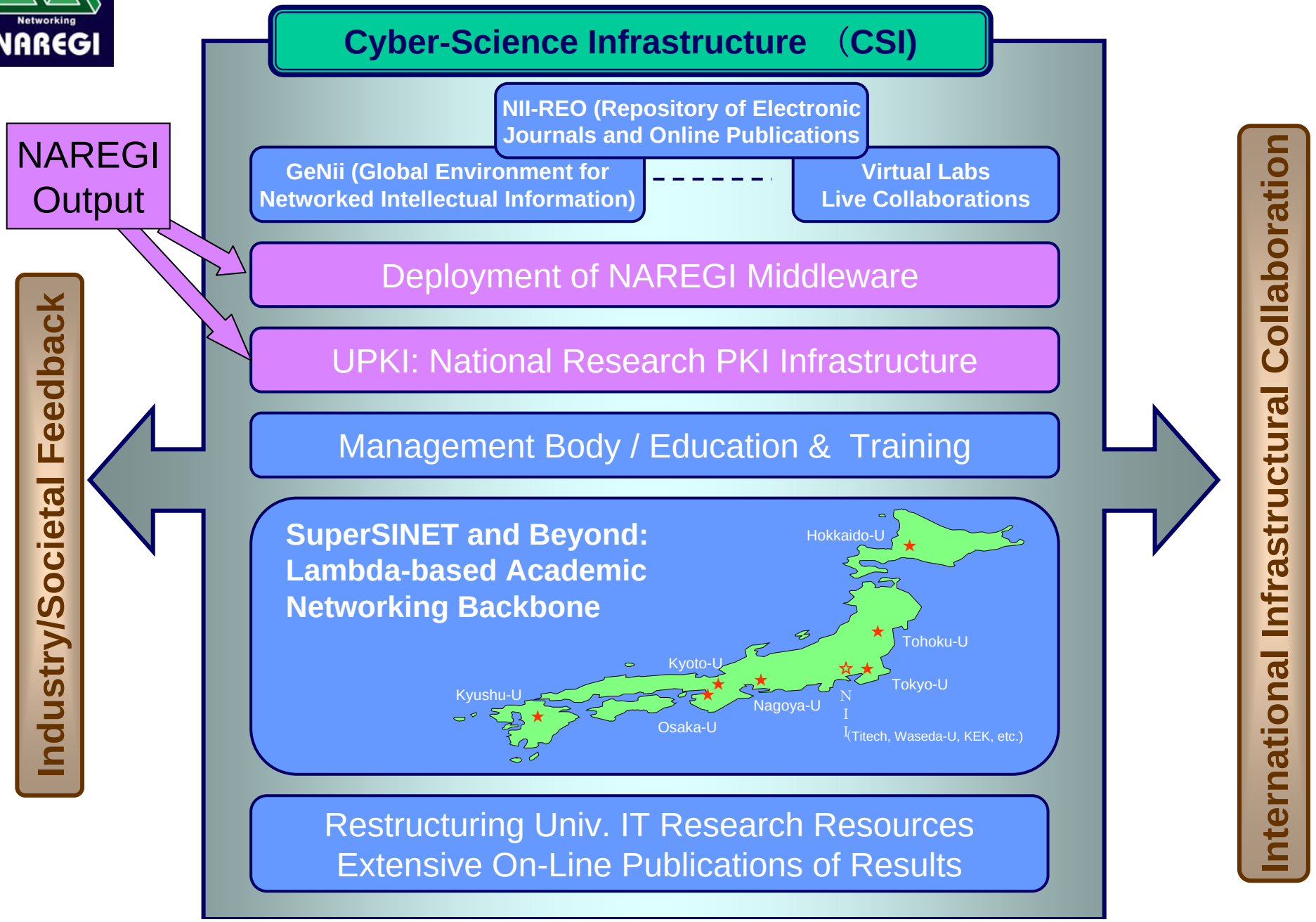
HEP Grid VO

NEES -ED Grid VO

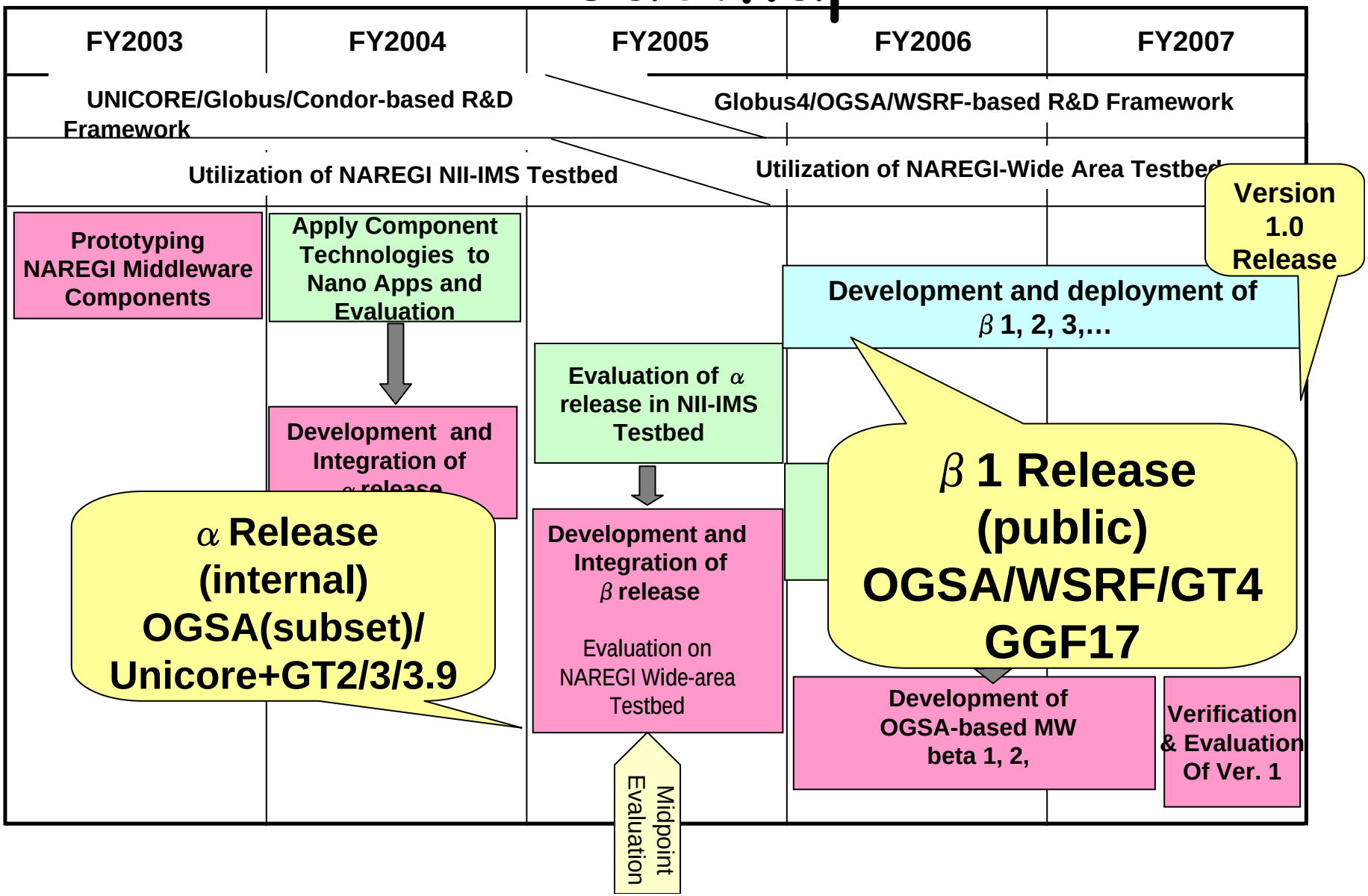
Astro IVO

Standardization, commonality in software platforms will realize this

Towards a Cyber-Science Infrastructure for R & D



NAREGI Middleware Roadmap



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 Appls
  - 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, NTT, …
  - No students involved in development
- Will NOT develop all software by itself
  - Will rely on external components in some cases
  - Must be WS and other industry standards compliant
- Will NOT deploy its own production Grid
  - Although there is a 3000-CPU testbed
  - Work with National Centers for CSI deployment
- Will NOT hinder industry adoption at all costs
  - Intricate open source copyright and IP policies
  - We want people to save/make money using NAREGI MW

NAREGI Programming Models

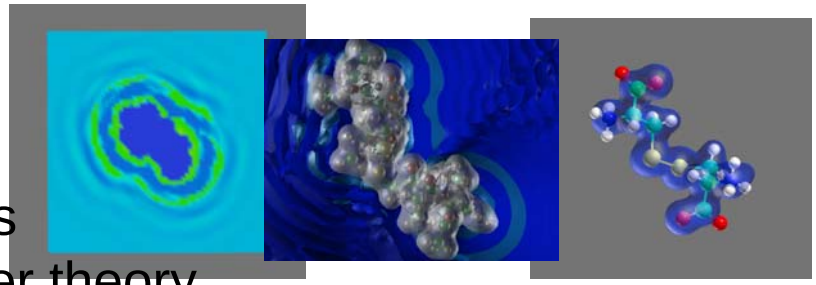
- High-Throughput Computing
  - But with complex data exchange inbetween
  - NAREGI Workflow or GridRPC
- Metacomputing (cross-machine parallel)
  - Workflow (w/co-scheduling) + GridMPI
  - GridRPC (for task-parallel or task/data-parallel)
- Coupled Multi-Resolution Simulation
  - Workflow (w/co-scheduling) + GridMPI + Coupling Components
    - Mediator (coupled simulation framework)
    - GIANT (coupled simulation data exchange framework)

Nano-Science : coupled 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

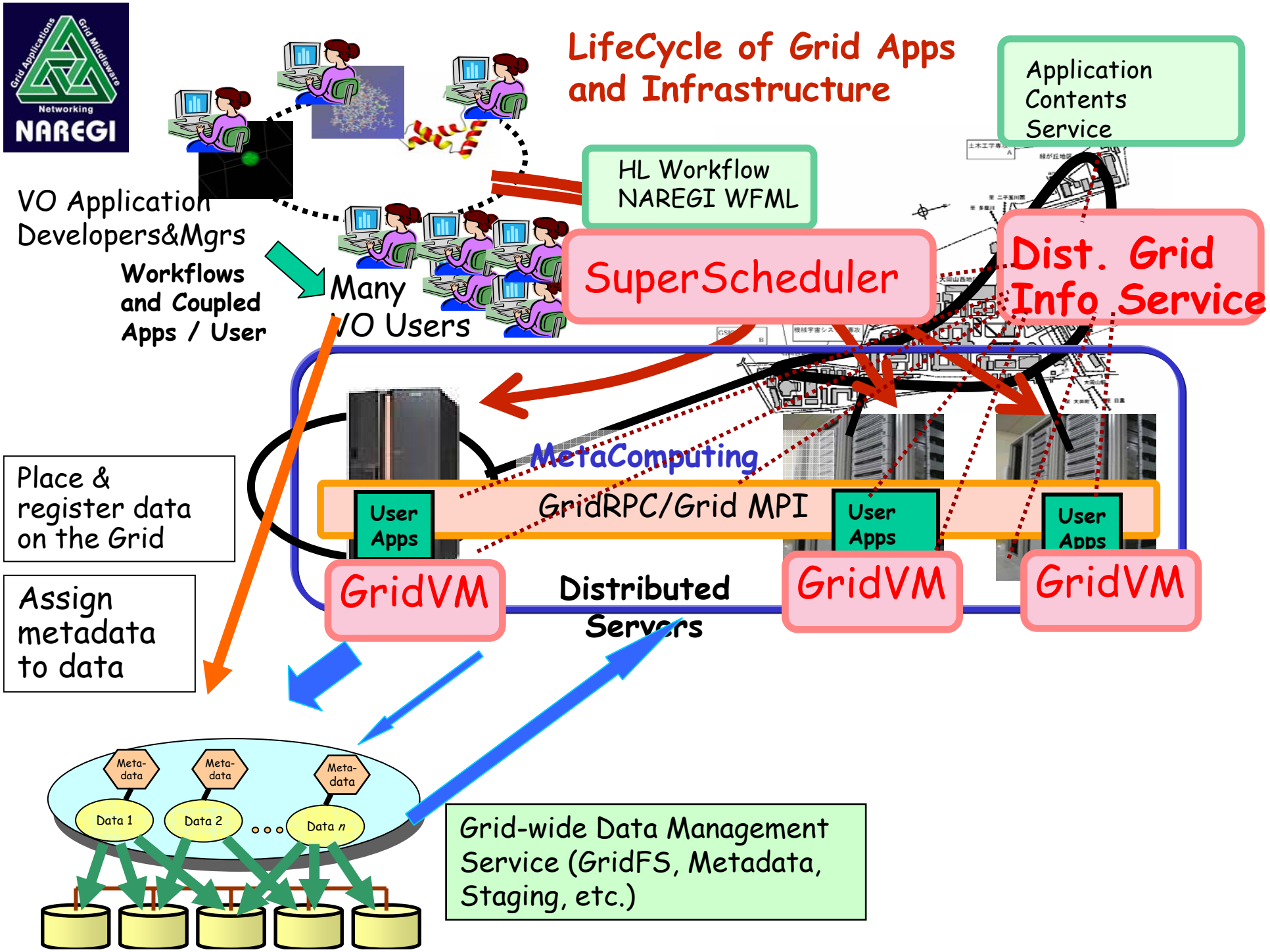
Limit of Computing Capability

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

Coordinates decoupled resources;

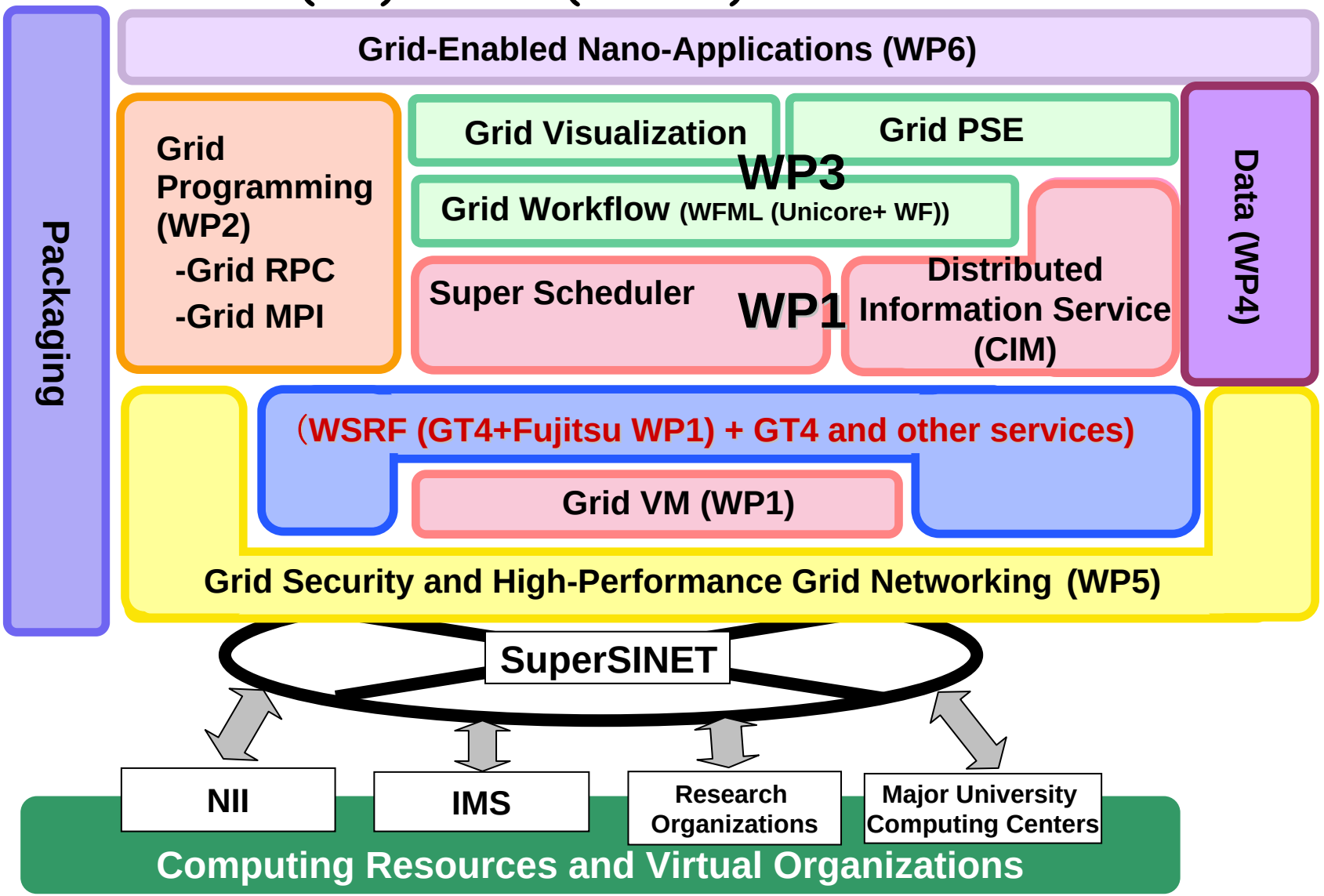
Meta-computing,
High throughput computing,
Multi-Physics simulation
w/ components and data from different groups
within VO composed in real-time

The only way to achieve true scalability!


NAREGI
Networking
Grid Applications
Grid Middleware
LifeCycle of Grid Apps and Infrastructure
Application Contents Service
HL Workflow NAREGI WFML
VO Application Developers&Mgrs
Workflows and Coupled Apps / User
Many VO Users
SuperScheduler
Dist. Grid Info Service
MetaComputing
GridRPC/Grid MPI
User Apps
User Apps
User Apps
GridVM
GridVM
GridVM
Distributed Servers
Place & register data on the Grid
Assign metadata to data
Meta-data
Meta-data
Meta-data
Data 1
Data 2
Data n
Grid-wide Data Management Service (GridFS, Metadata, Staging, etc.)

NAREGI Software Stack (beta 1 2006)

- WS(RF) based (OGSA) SW Stack -



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

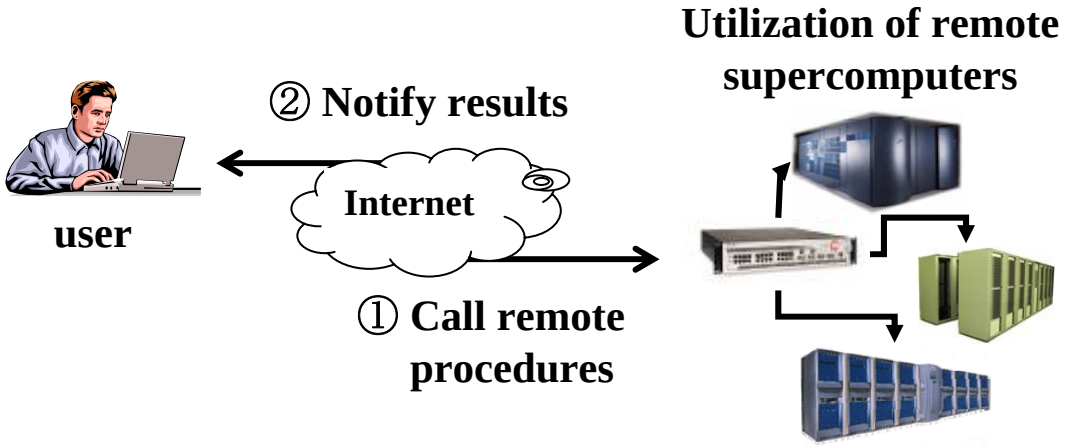
Necessary for Longevity and Vendor Buy-In
Metric of WP Evaluation

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
  - Full WS-Agreement brokering and co-allocation
  - GGF JSDL1.0-based 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)

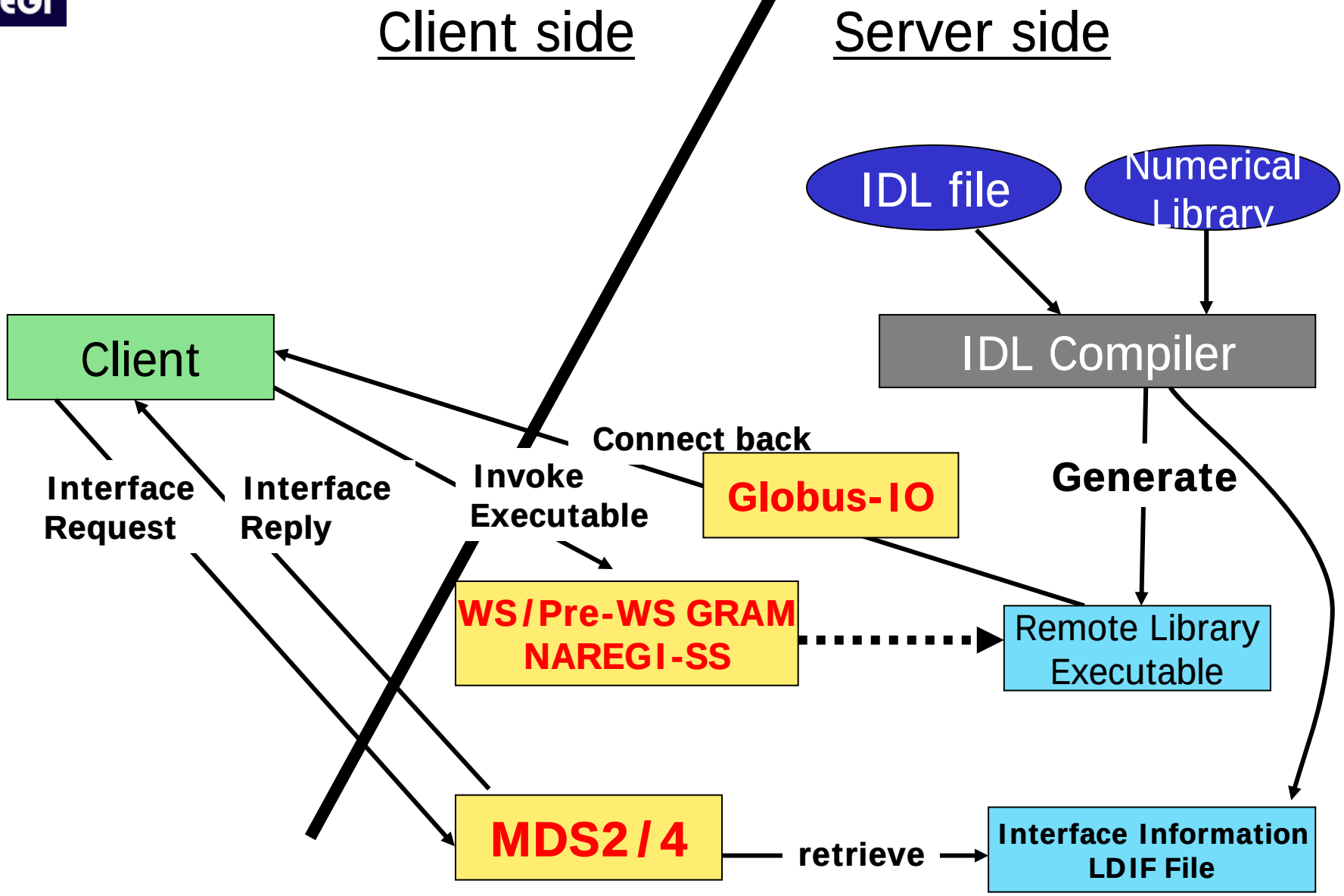
Ninf-G: A Reference Implementation of the GGF GridRPC API

- 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



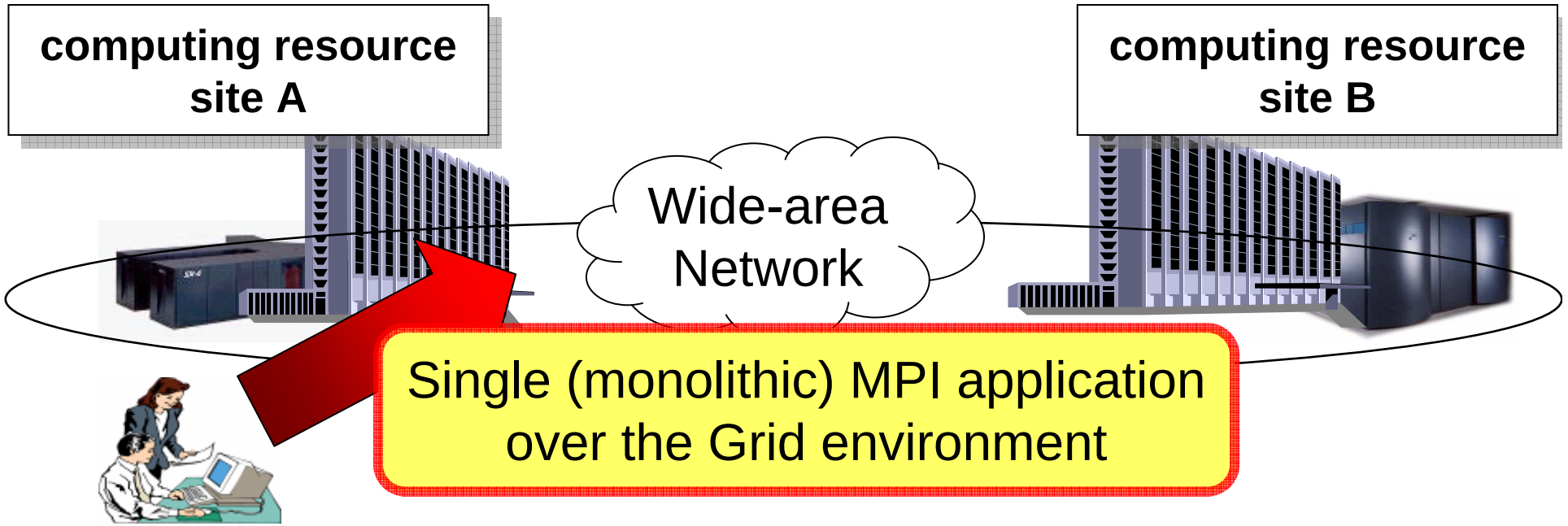
Large scale computing across supercomputers on the Grid

Architecture of Ninf-G

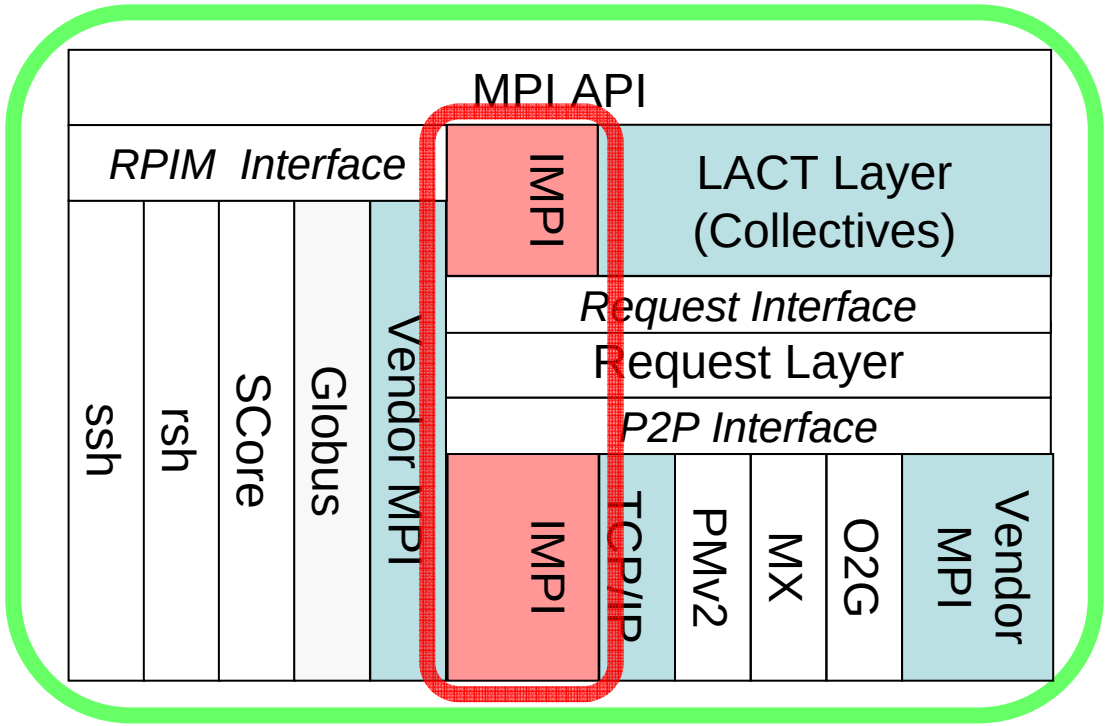


GridMPI

- MPI applications run on the Grid environment
- Metropolitan area, high-bandwidth environment: ≥ 10 Gpbs, ≤ 500 miles (smaller than 10ms one-way latency)
  - Parallel Computation
- Larger than metropolitan area
  - MPI-IO

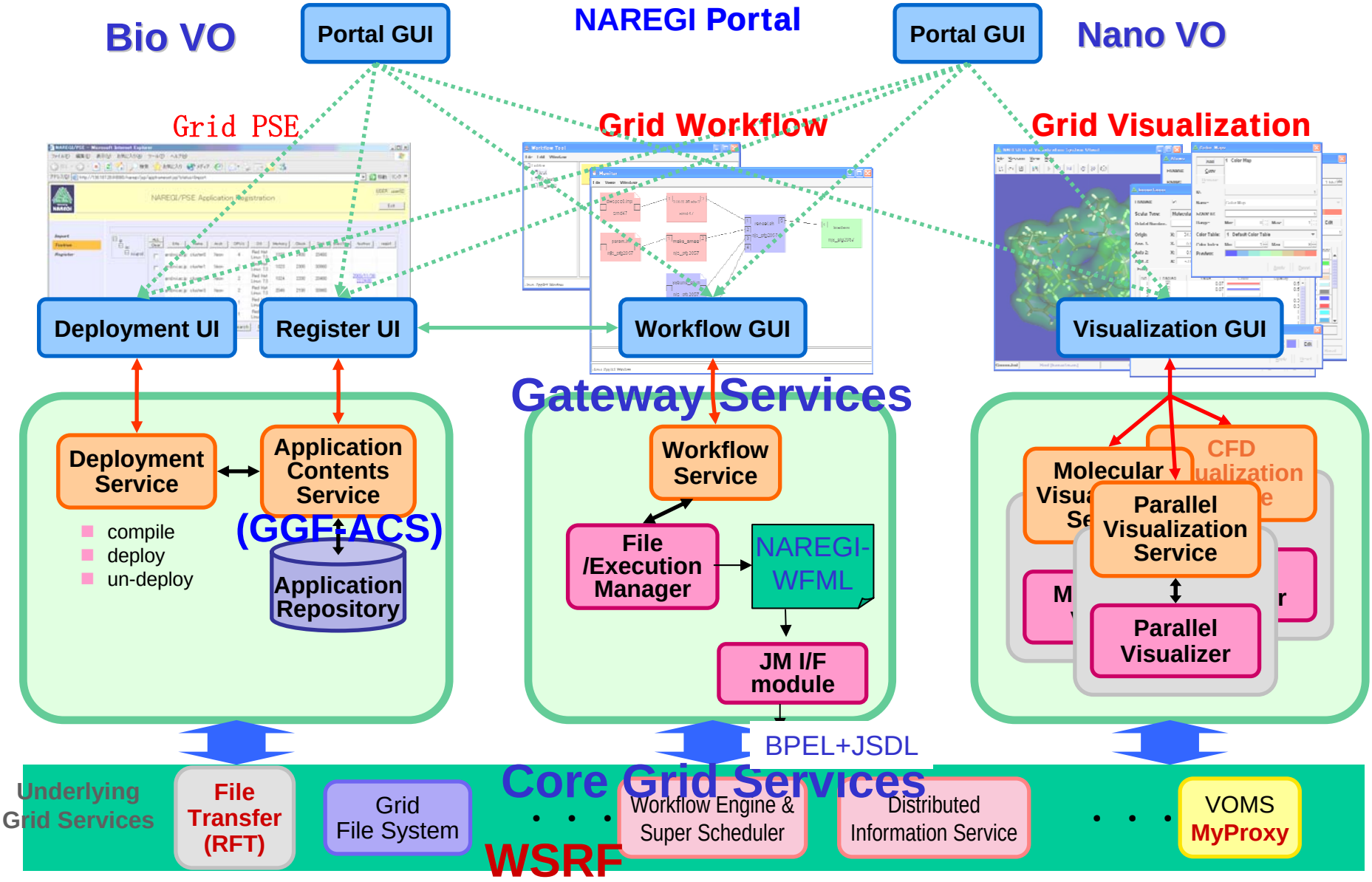


GridMPI Software Architecture and Standards

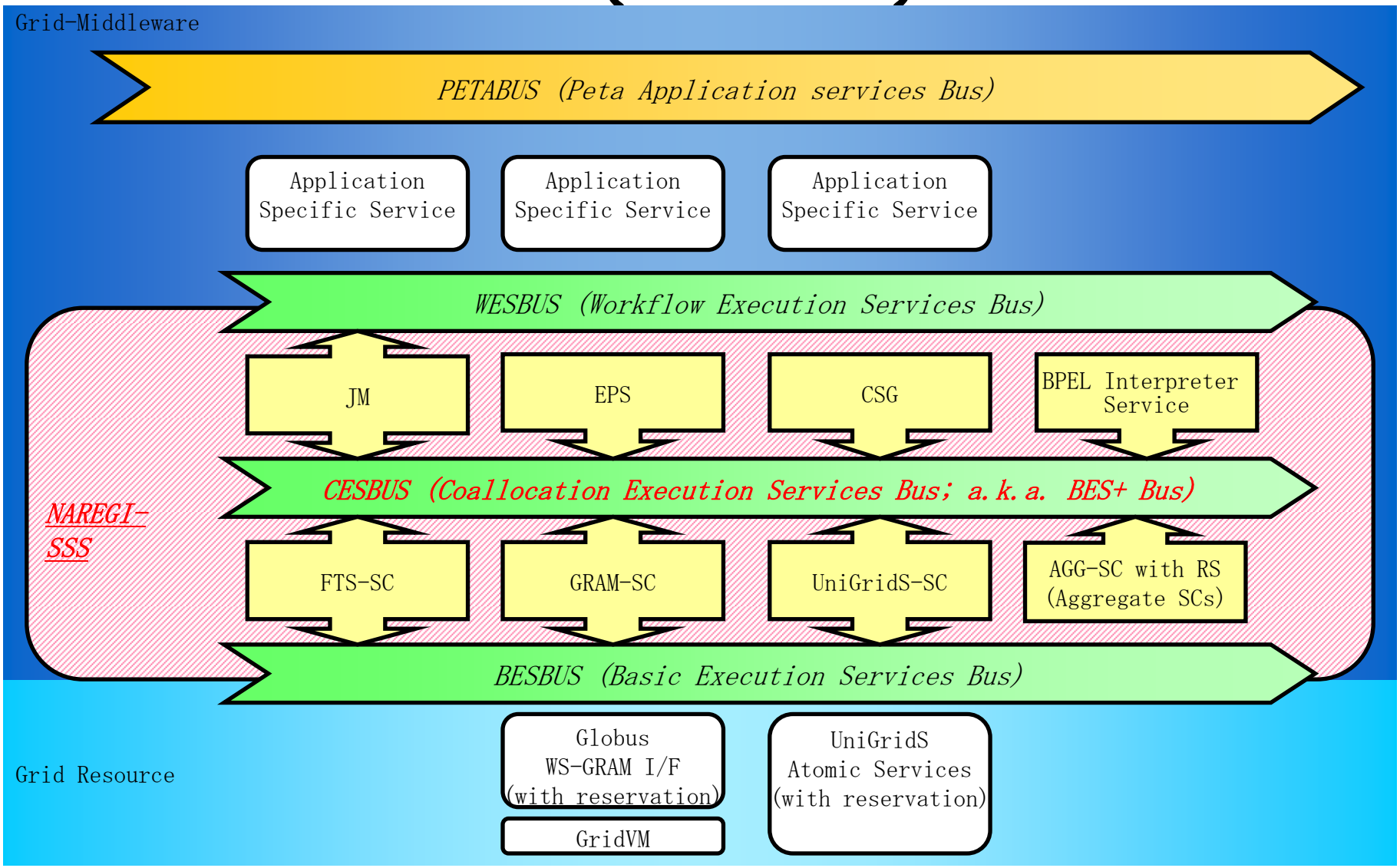


- *MPI 2.0 (test suite compliant)*
- *IMPI (Interoperable MPI)*
  - The original IMPI is defined only for the MPI-1.2 feature
  - Extension for MPI-2
- Porting the extended IMPI protocol to Open MPI
- Planning to submit the protocol to NIST

Grid Application Environment (WP3)

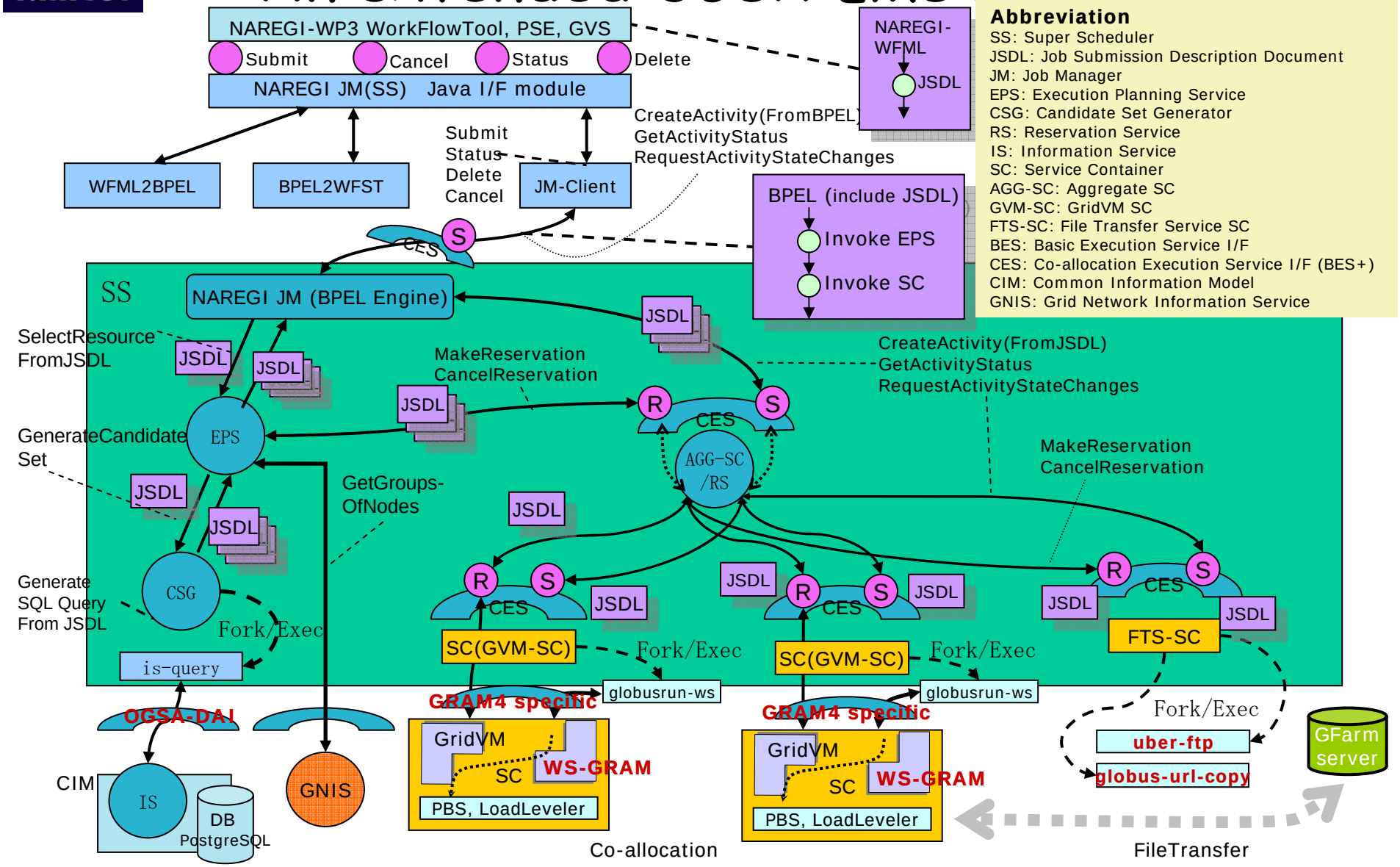


The NAREGI SSS Architecture (2007/3)



NAREGI beta 1 SSS Architecture

An extended OGSA-EMS Incarnation



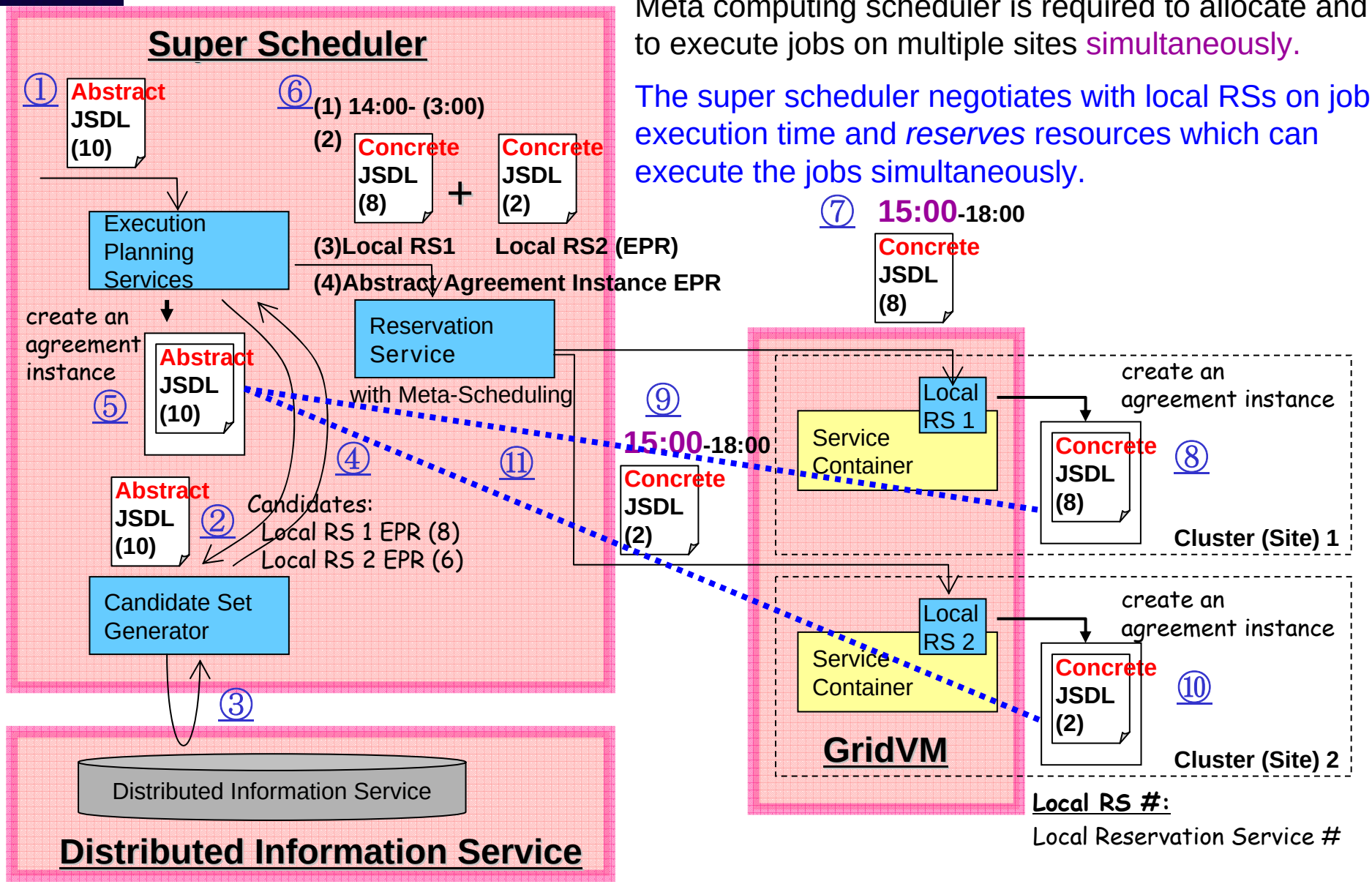
Abbreviation

SS: Super Scheduler
JSDL: Job Submission Description Document
JM: Job Manager
EPS: Execution Planning Service
CSG: Candidate Set Generator
RS: Reservation Service
IS: Information Service
SC: Service Container
AGG-SC: Aggregate SC
GVM-SC: GridVM SC
FTS-SC: File Transfer Service SC
BES: Basic Execution Service I/F
CES: Co-allocation Execution Service I/F (BES+)
CIM: Common Information Model
GNIS: Grid Network Information Service

3, 4: Co-allocation and Reservation

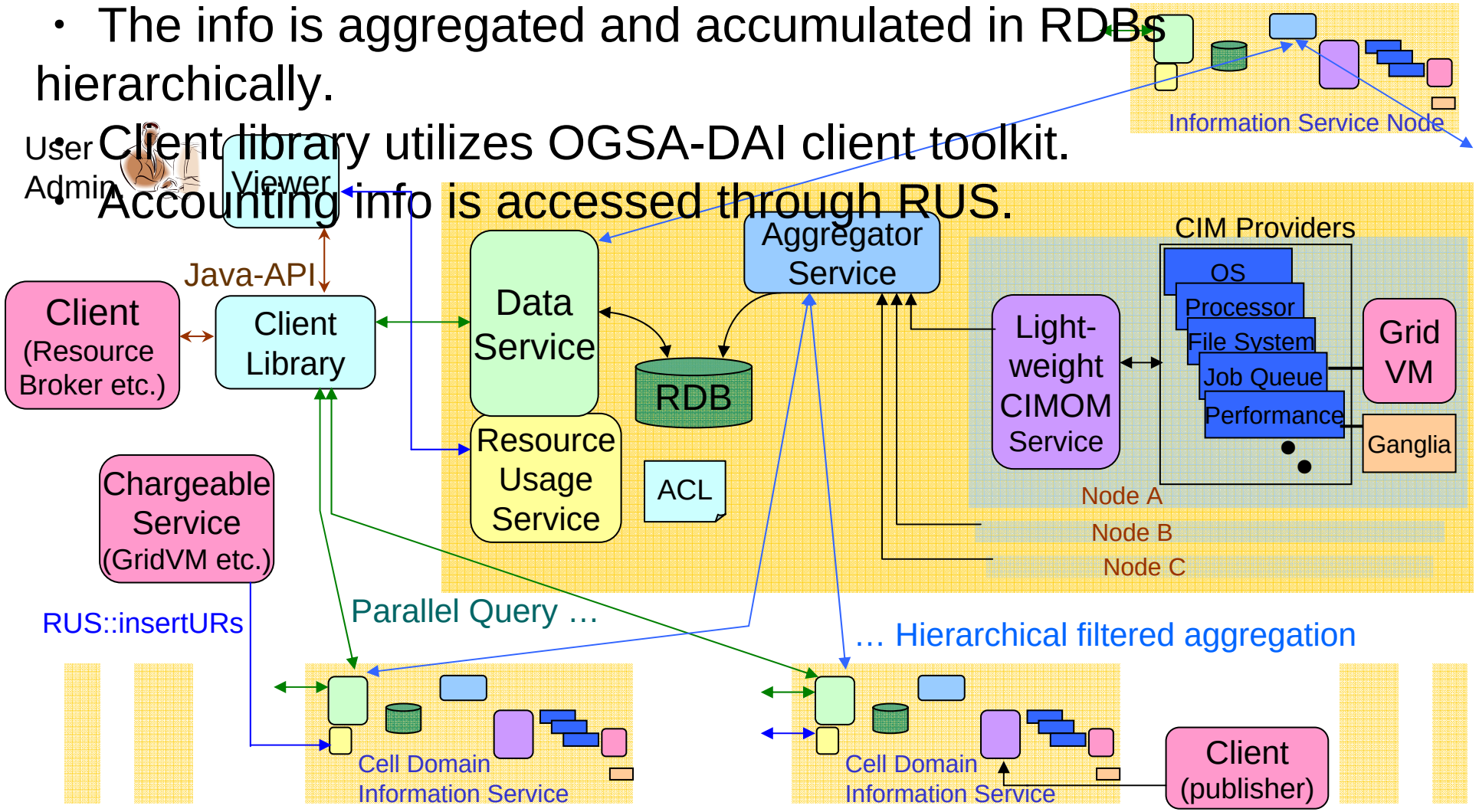
Meta computing scheduler is required to allocate and to execute jobs on multiple sites simultaneously.

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

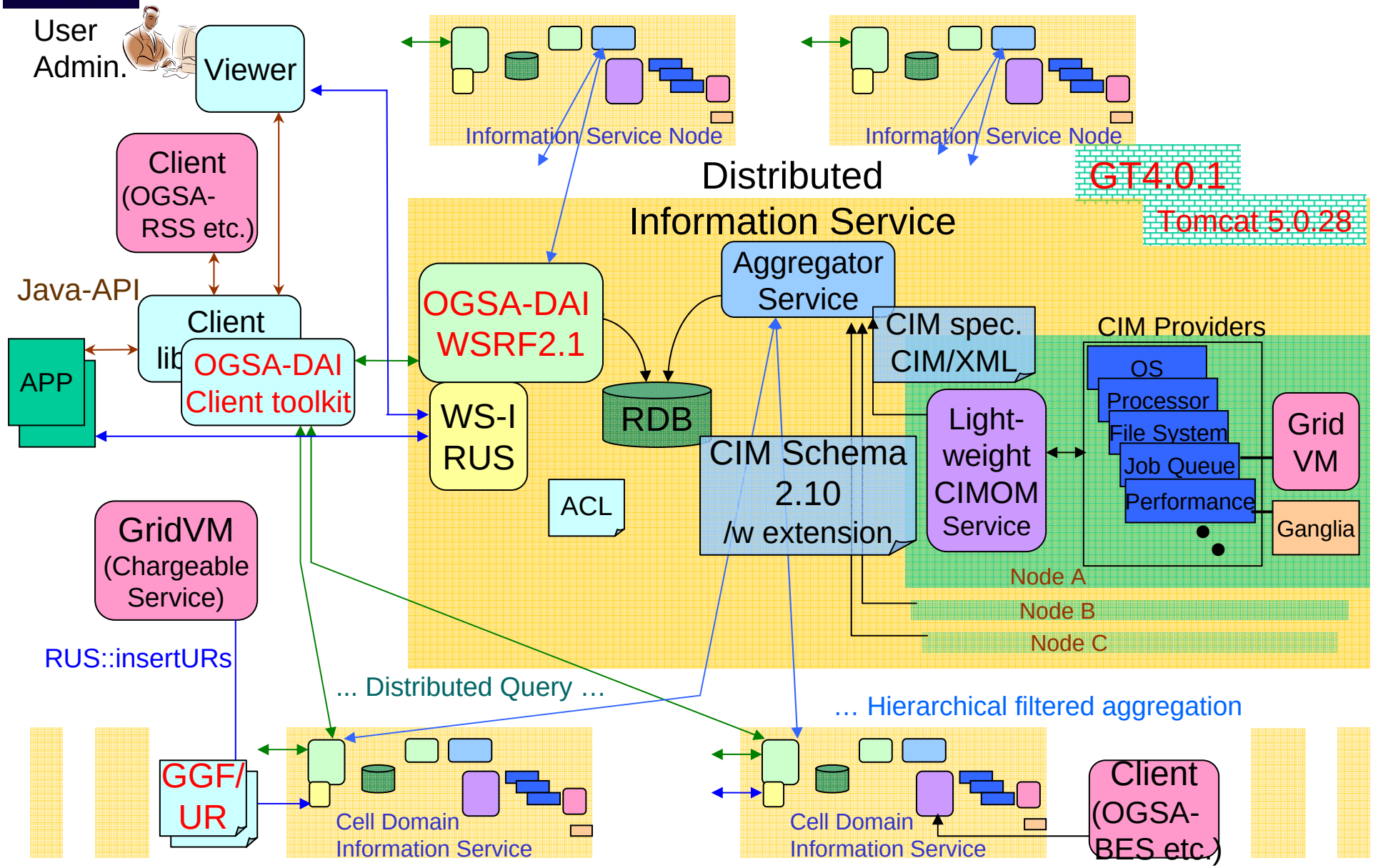


NAREGI Info Service (beta) Architecture

- 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.



NAREGI IS: Standards Employed in the Architecture

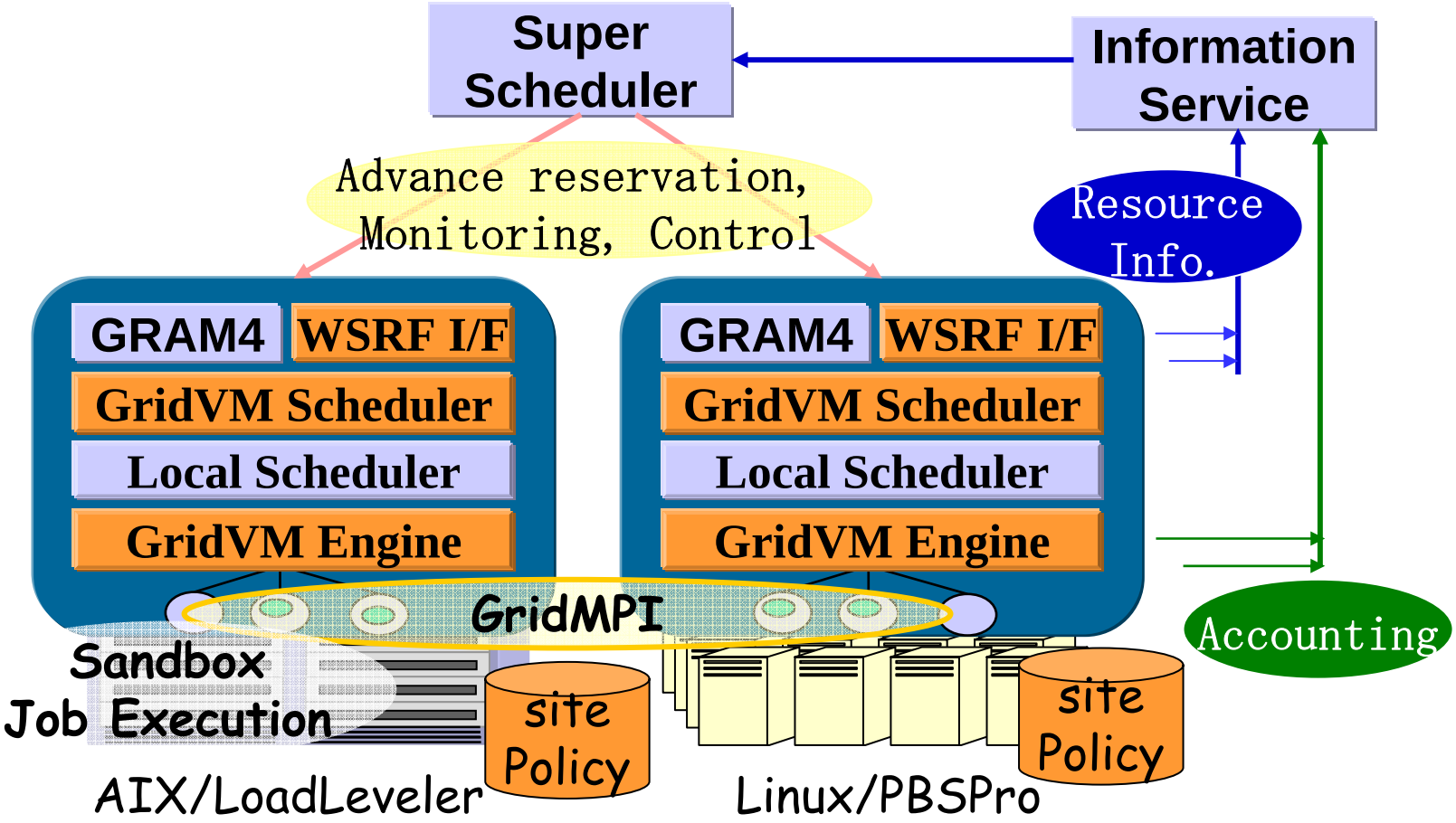


GridVM Features

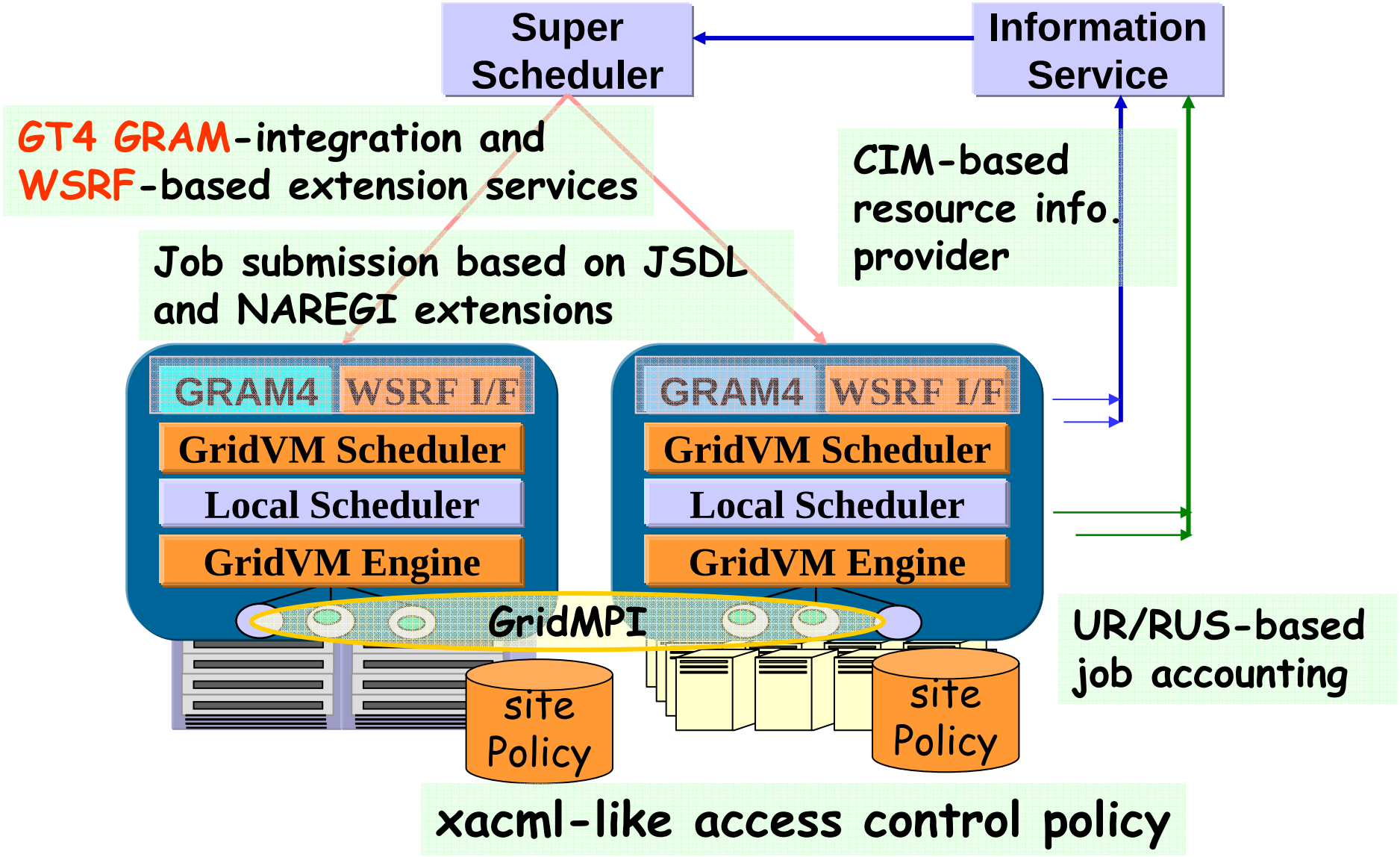
- ✓ Platform independence as OGSA-EMS SC
  - WSRF OGSA-EMS Service Container interface for heterogeneous platforms and local schedulers
  - "Extends" Globus4 WS-GRAM
  - Job submission using JSDL
  - Job accounting using UR/RUS
  - CIM provider for resource information
- ✓ Meta-computing and Coupled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

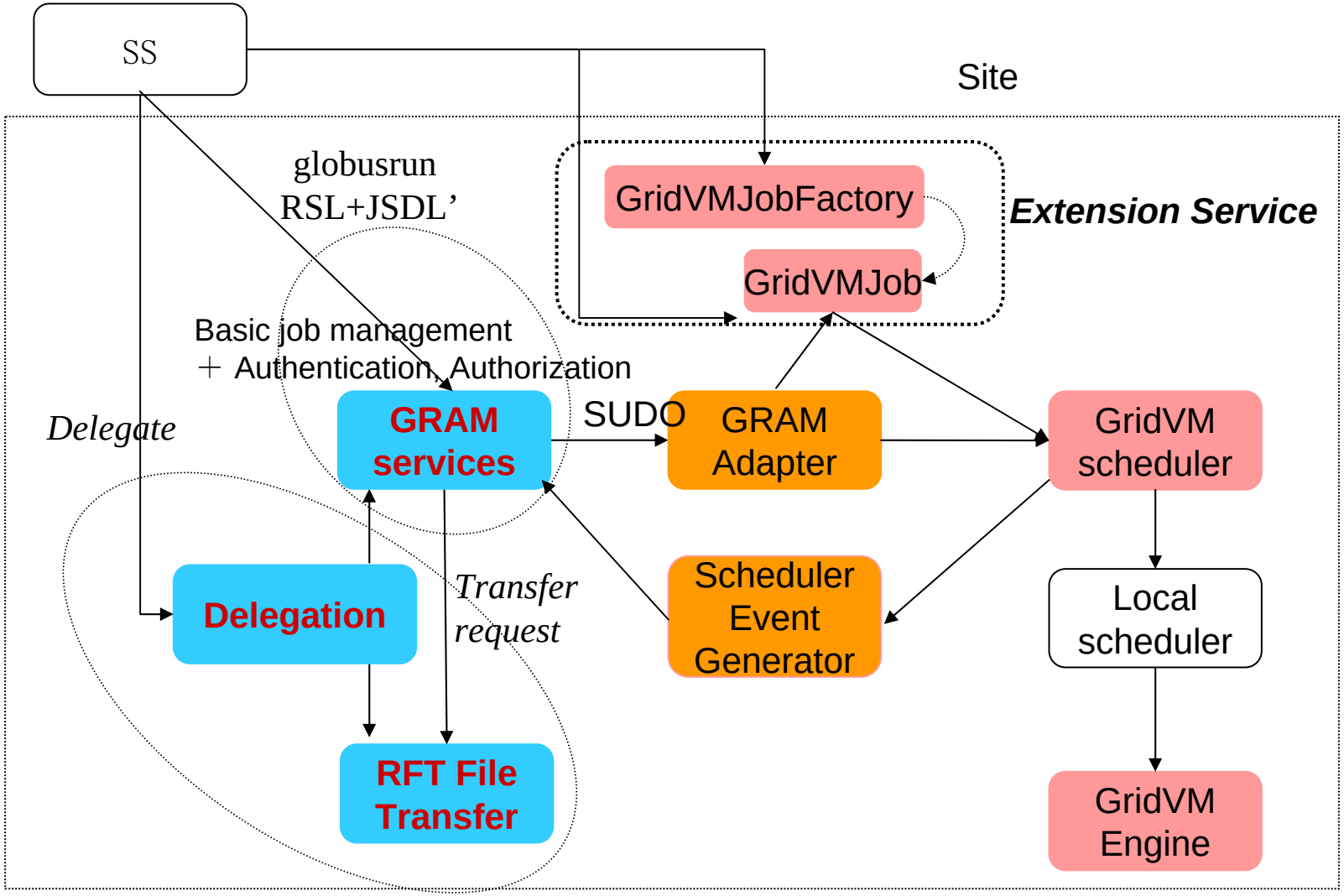


NAREGI GridVM: Standards Employed in the Architecture

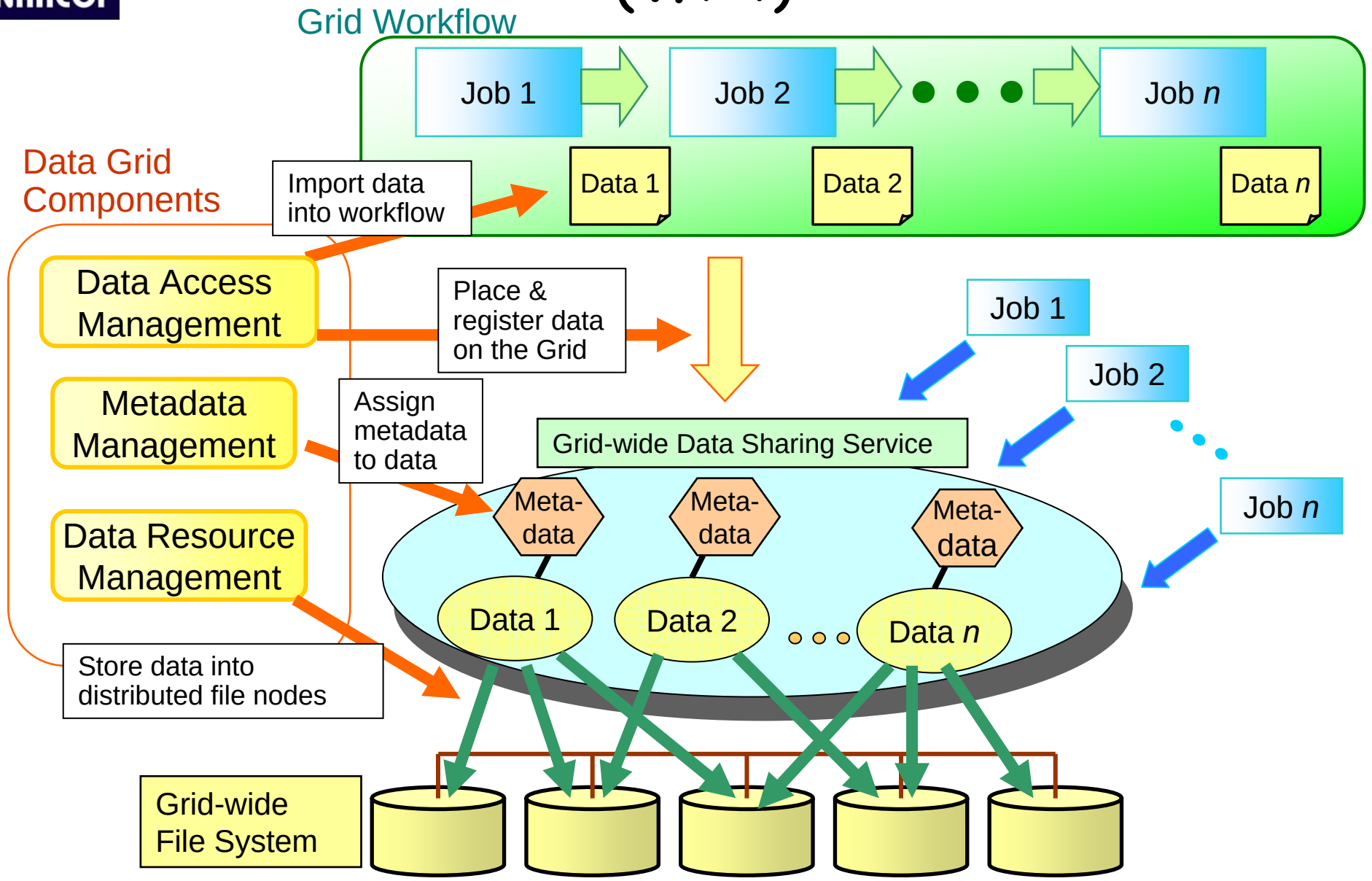


GT4 GRAM-GridVM Integration

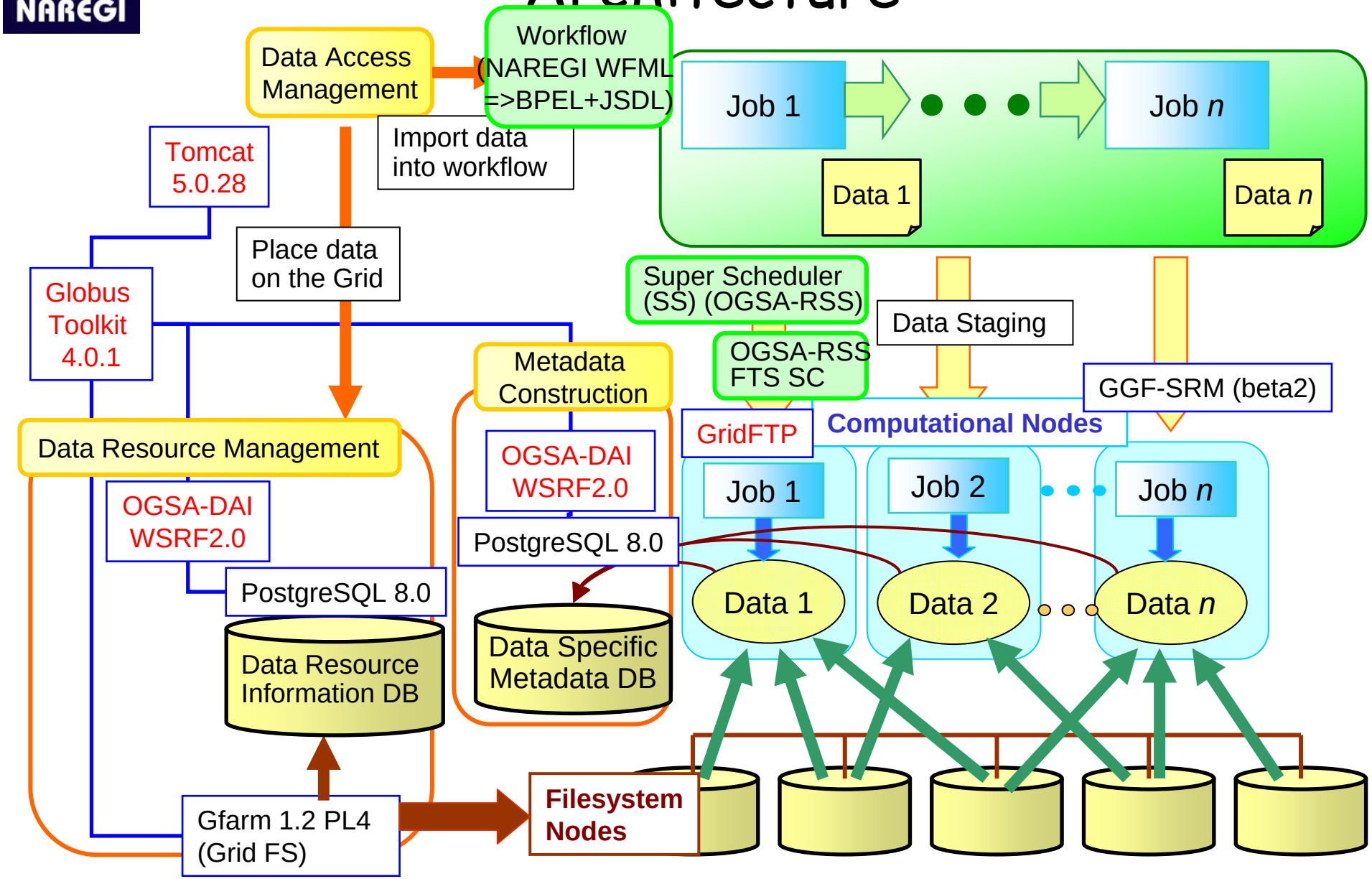
- ✓ Integrated as an extension module to GT4 GRAM
- ✓ Aim to make the both functionalities available



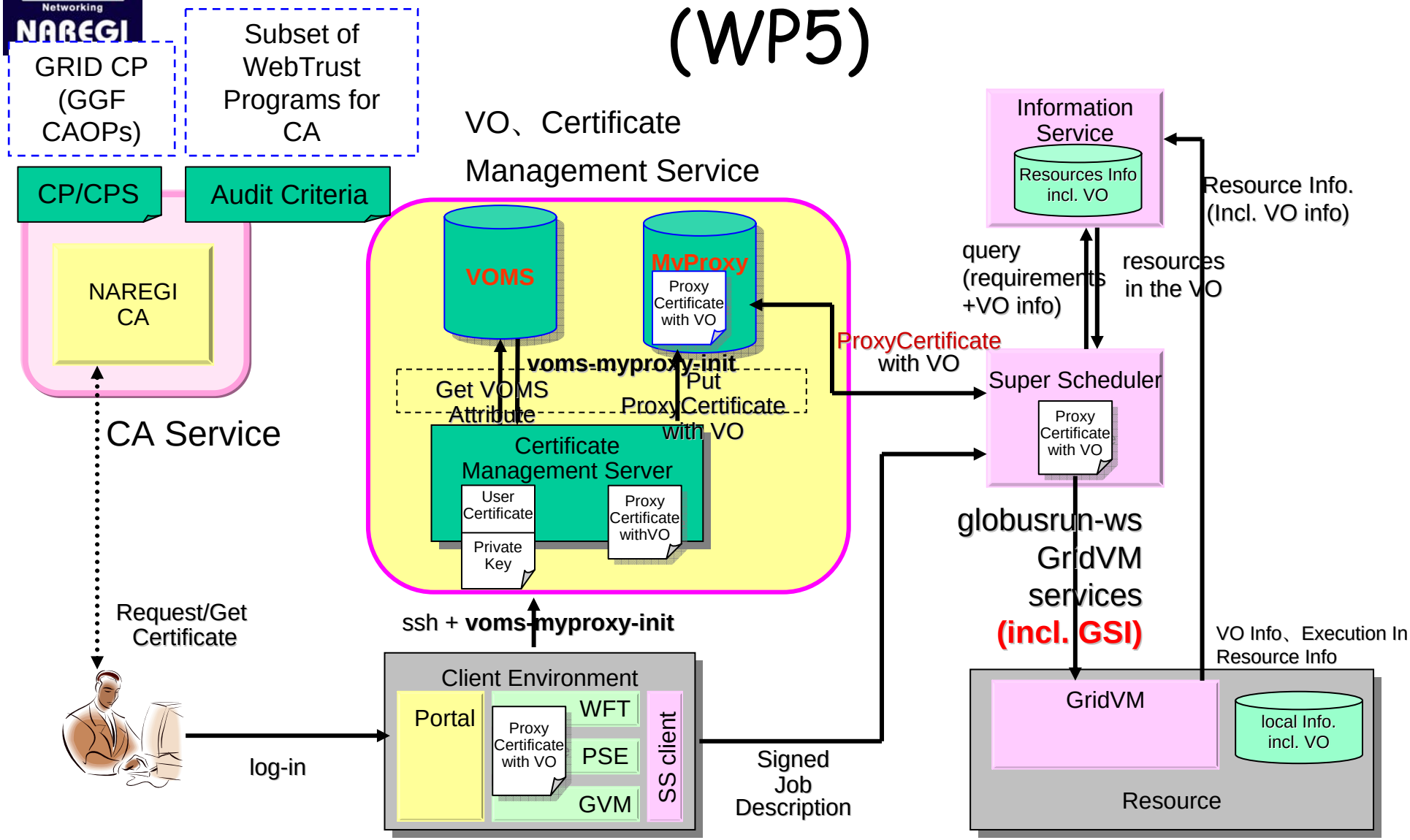
NAREGI Data Grid beta1 Architecture (WP4)



NAREGI WP4: Standards Employed in the Architecture



NAREGI-beta1 Security Architecture (WP5)



NAREGI Application Mediator (WP6) for Coupled Applications

Mediator Components

Support data exchange between coupled simulation

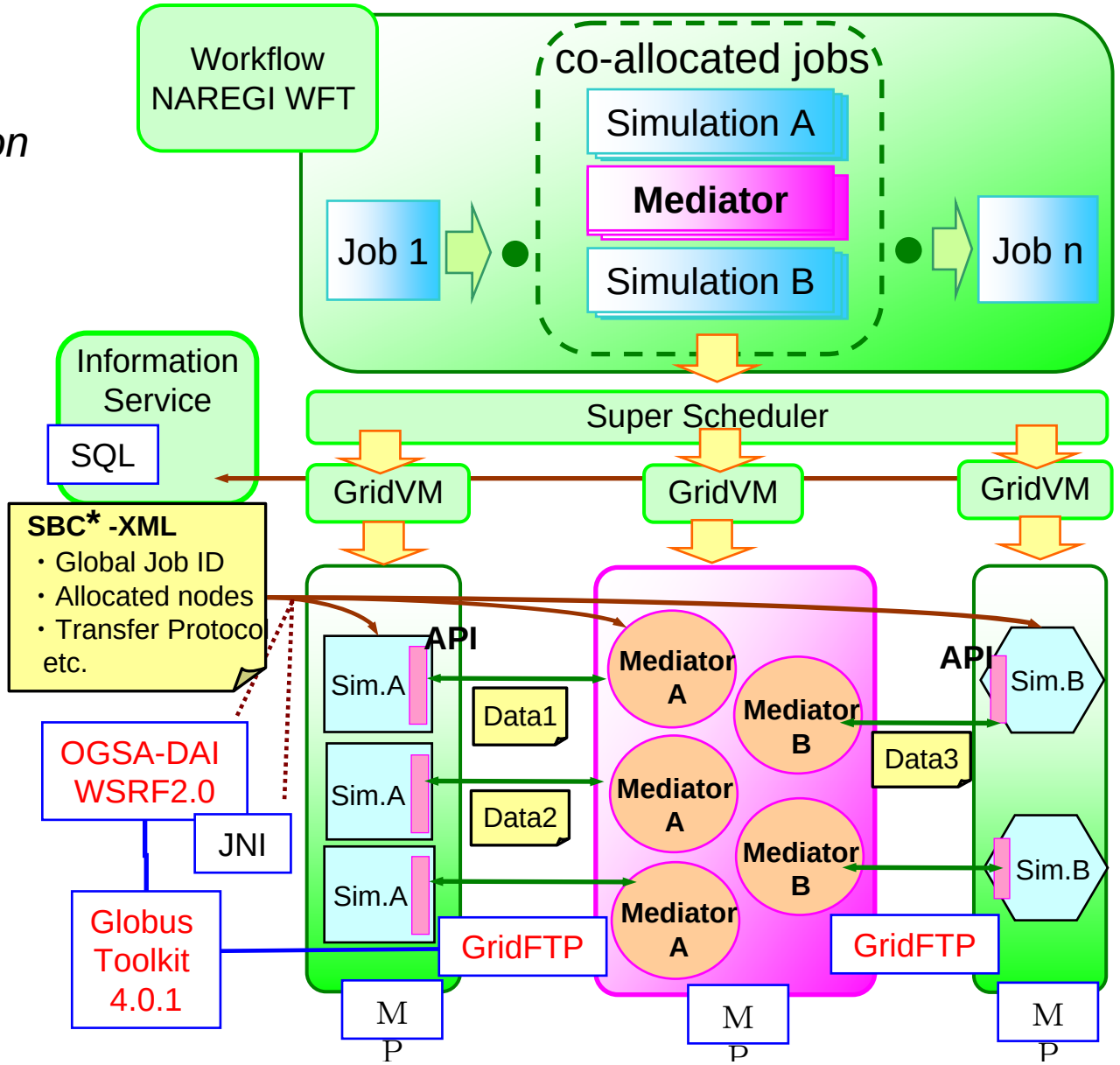
Data transfer management

- Synchronized file transfer
- Multiple protocol GridFTP/MPI

Data transformation management

- Semantic transform-ation libraries for different simulations
- Coupled accelerator

*SBC: Storage-based communication



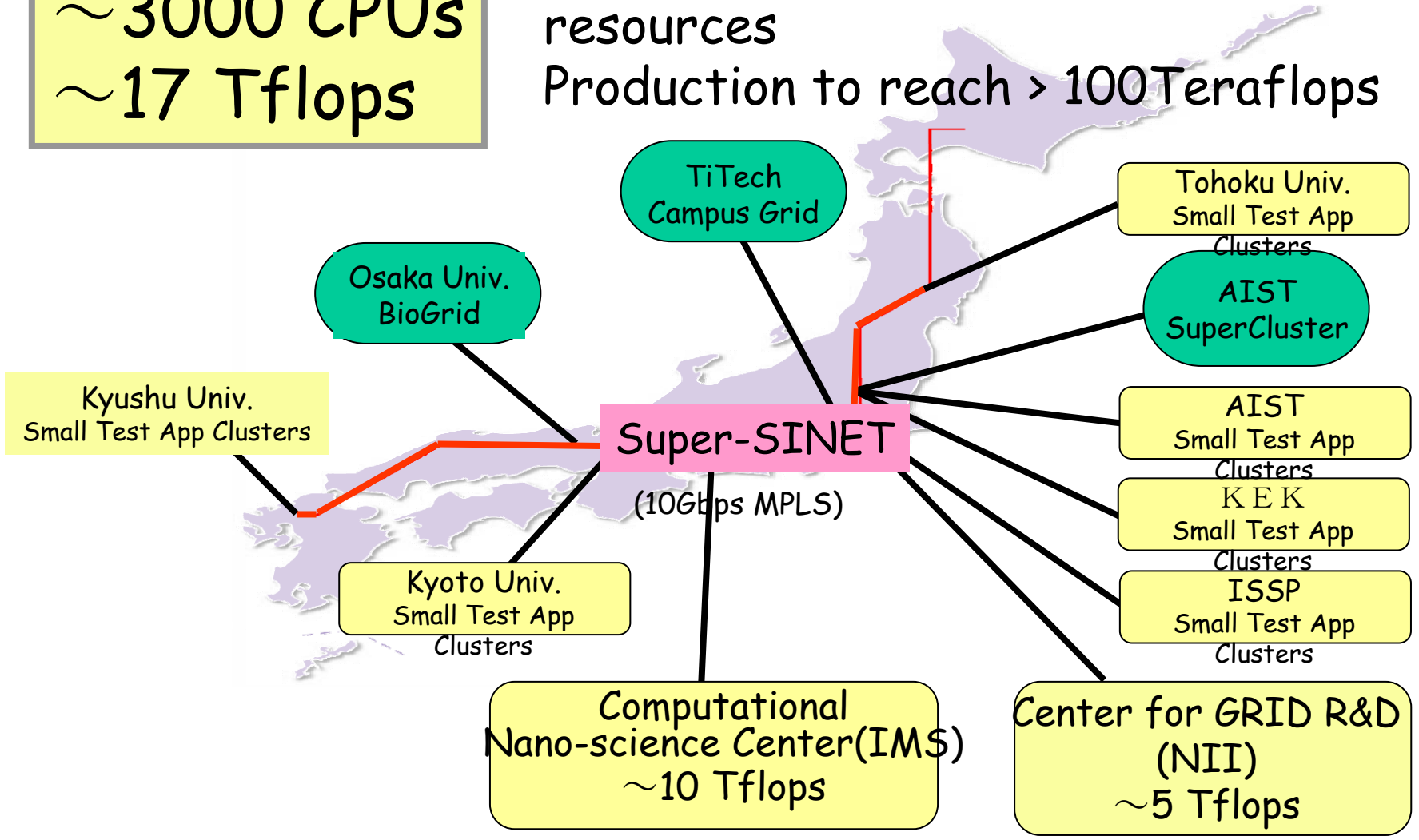
NAREGI Phase 1 Testbed

Dedicated Testbed

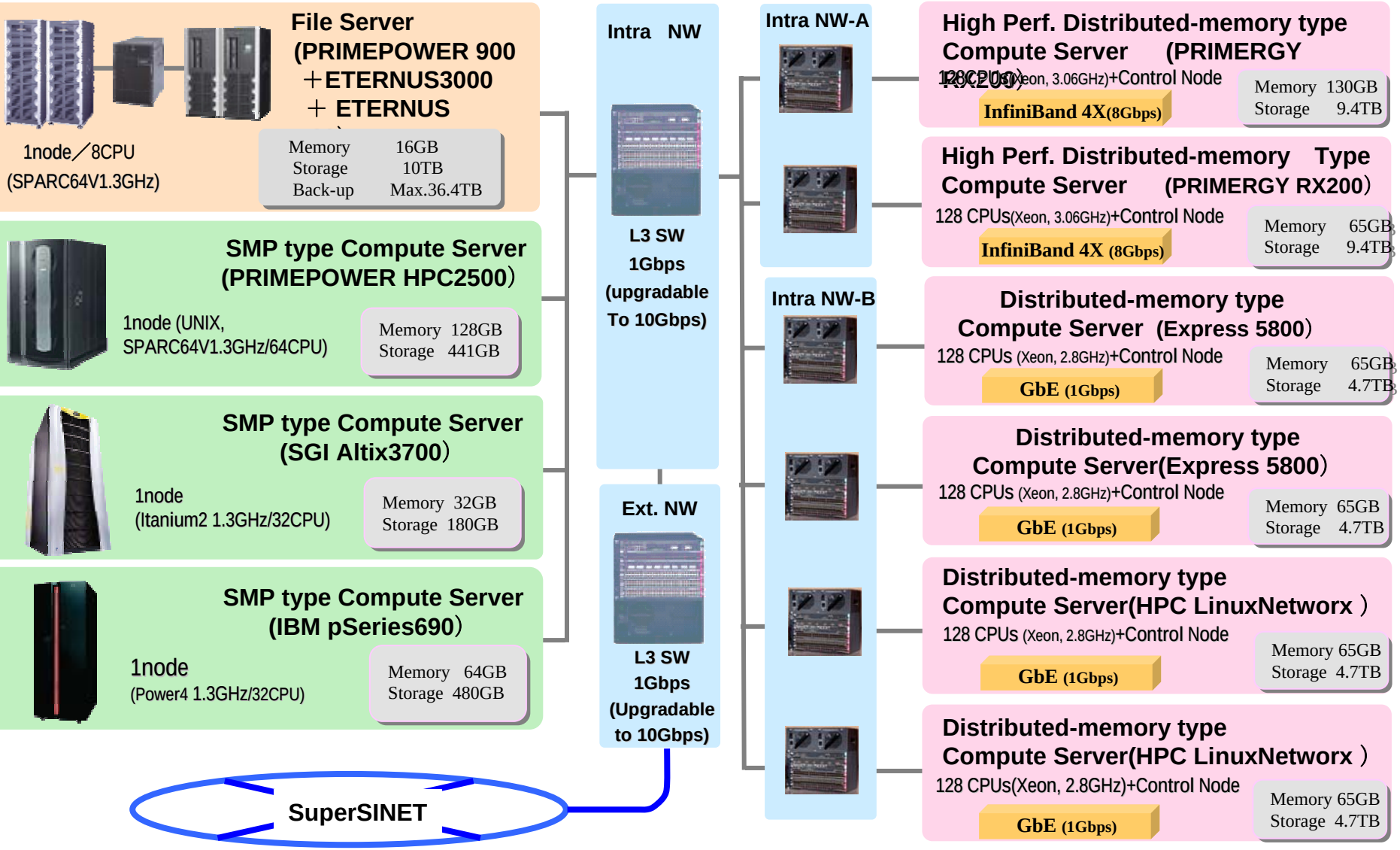
No "ballooning" w/production resources

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)



Computer System for Nano Application R & D

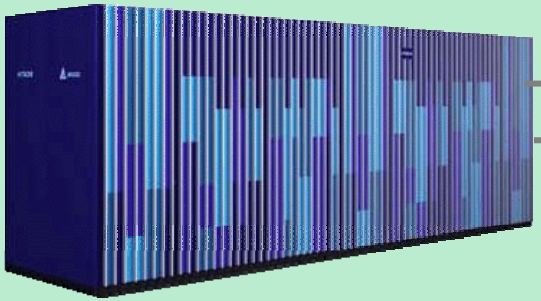
Computational Nano science Center （10Ｔflops，5TB)

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

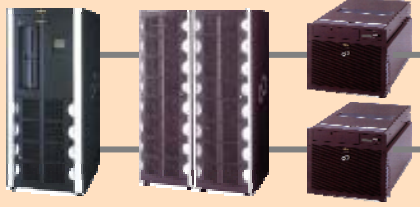
L3 SW

1Gbps

(Upgradable to 10Gbps)

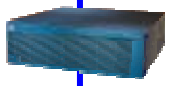
File Server

16CPUs (SPARC64 GP, 675MHz)



Memory 8GB
Storage 30TB
Back-up 25TB

CA/RA Server



Firewall

VPN

Center for Grid R & D

SuperSINET