* 1 of the 30+ projects of ACI Grid

Grid'5000*

A nation wide Experimental Grid

Franck Cappello
INRIA
fci@lri.fr

A very brief overview

ministère éducation nationale enseignement supérieur recherche

ministère délégué à la recherche

ACI GRID

INRIA

Renater

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Agenda

Motivation

Grid'5000 project

Grid'5000 design

Grid'5000 developments

Conclusion

Grid raises research issues but also methodological challenges

Grid are complex systems:

Large scale, Deep stack of complicated software

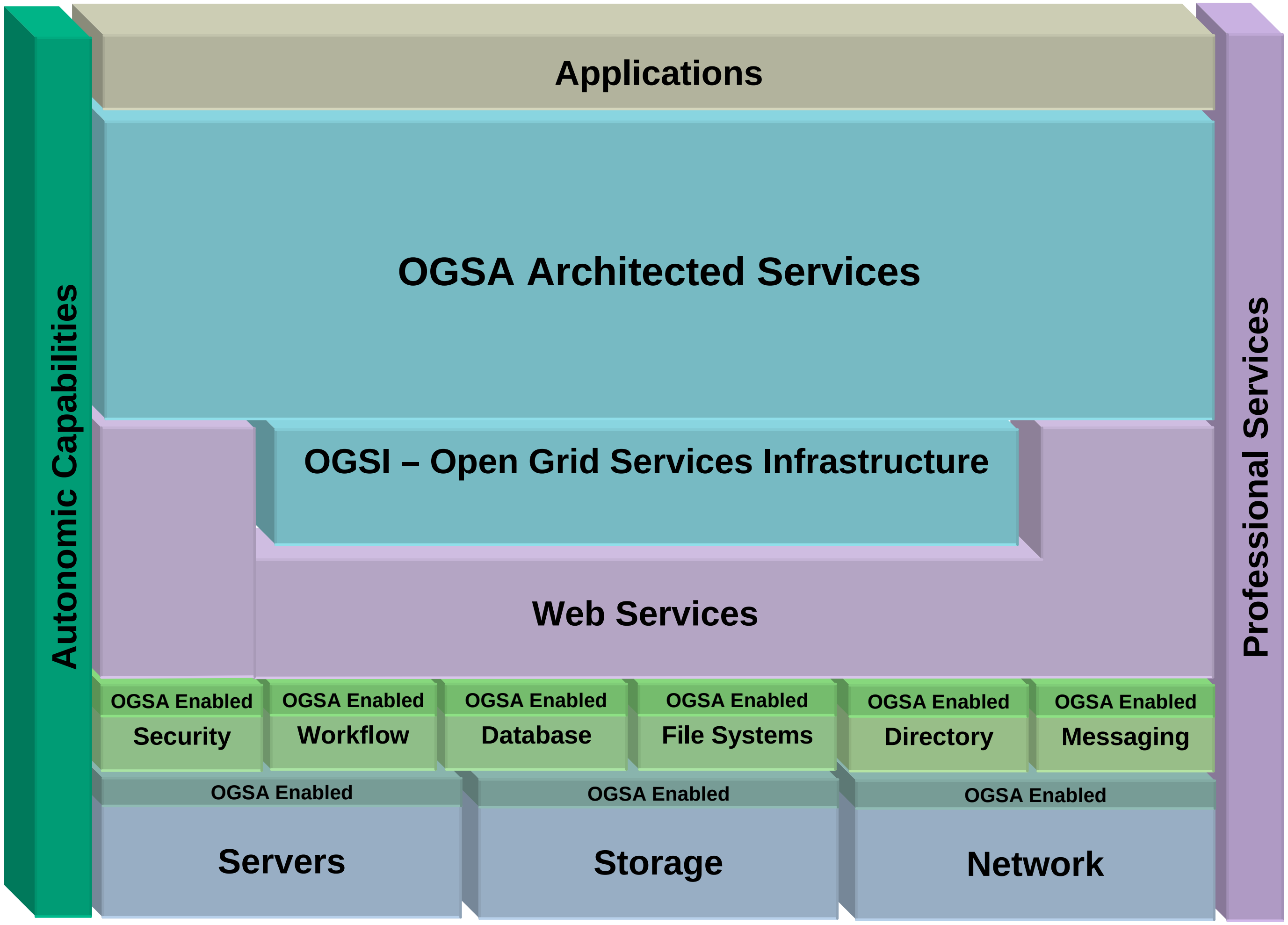
Grid raises a lot of research issues:

Security, Performance, Fault tolerance, Scalability, Load Balancing, Coordination, Message passing, Data storage, Programming, Algorithms, Communication protocols and architecture, Deployment, Accounting, etc.

➡ How to test and compare?

- Fault tolerance protocols
- Security mechanisms
- Networking protocols
- etc.

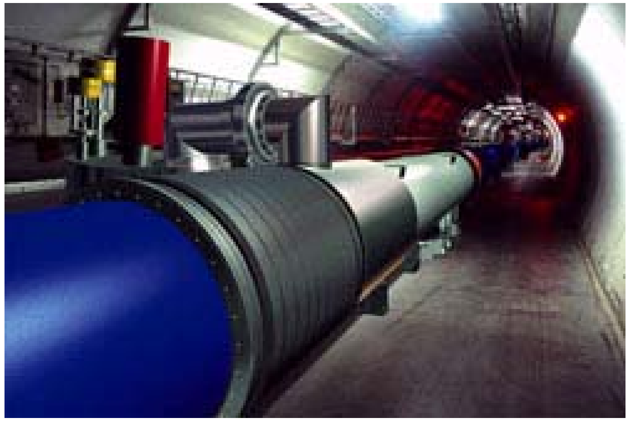
Service oriented approach



Reconfiguration oriented approach

The CERN openlab for DataGrid Applications

Openlab is a collaboration between CERN and industrial partners to develop data-intensive grid technology to be used by a worldwide community of scientists working at the next-generation Large Hadron Collider.



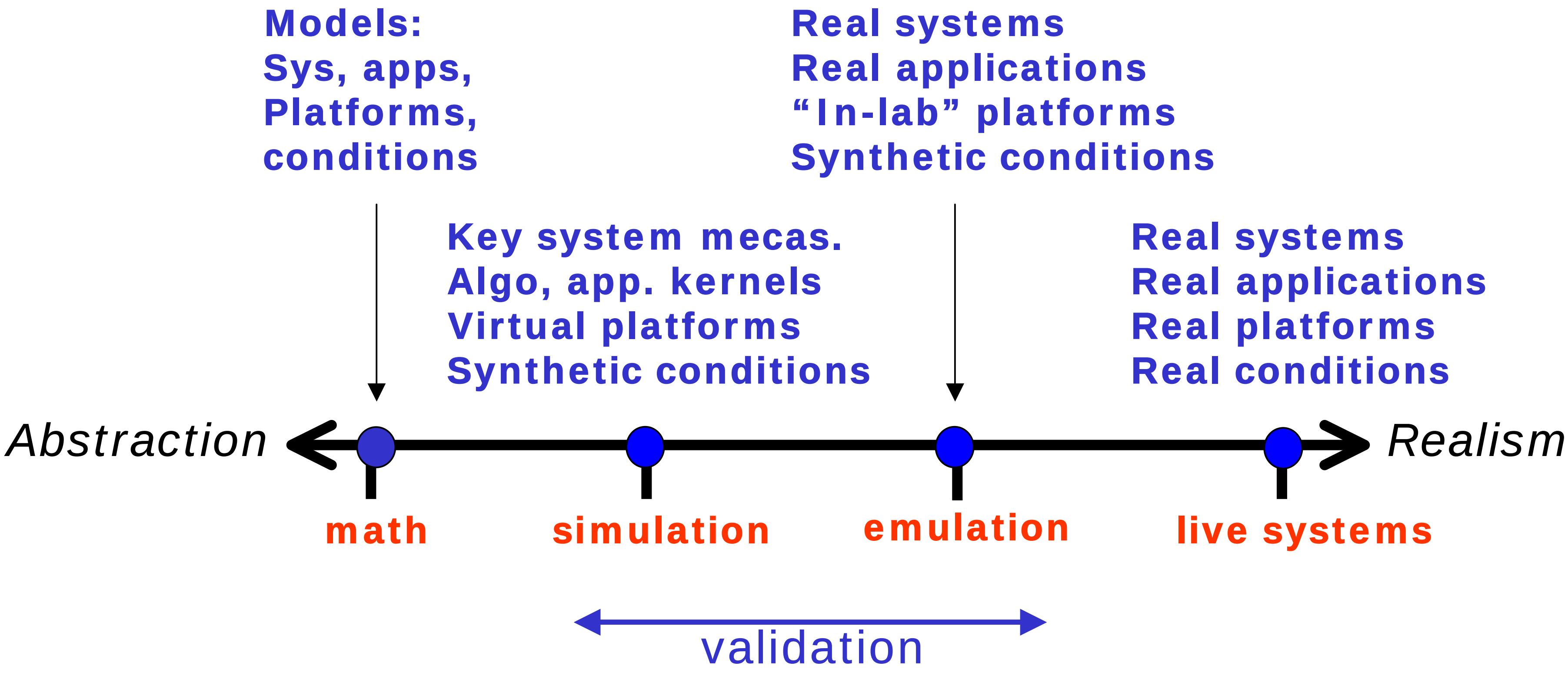
"Scientific software is usually distributed in form of optimized binaries for every platform and sometimes even tightly coupled to specific versions of the operating system."

"A grid node executing a task should thus be able to provide exactly the environment needed by the application."

Tools for Distributed System Studies

To investigate Distributed System issues, we need:

1) Tools (model, simulators, emulators, experi. Platforms)



Models:
Sys, apps,
Platforms,
conditions

Real systems
Real applications
"In-lab" platforms
Synthetic conditions

Key system mecas.
Algo, app. kernels
Virtual platforms
Synthetic conditions

Real systems
Real applications
Real platforms
Real conditions

Abstraction

Realism

math

simulation

emulation

live systems

validation

2) Strong interaction between these research tools

Existing Grid Research Tools

- 🟥 *France*
- 🟦 *USA*
- 🟩 *Australia*
- 🟧 *Japan*

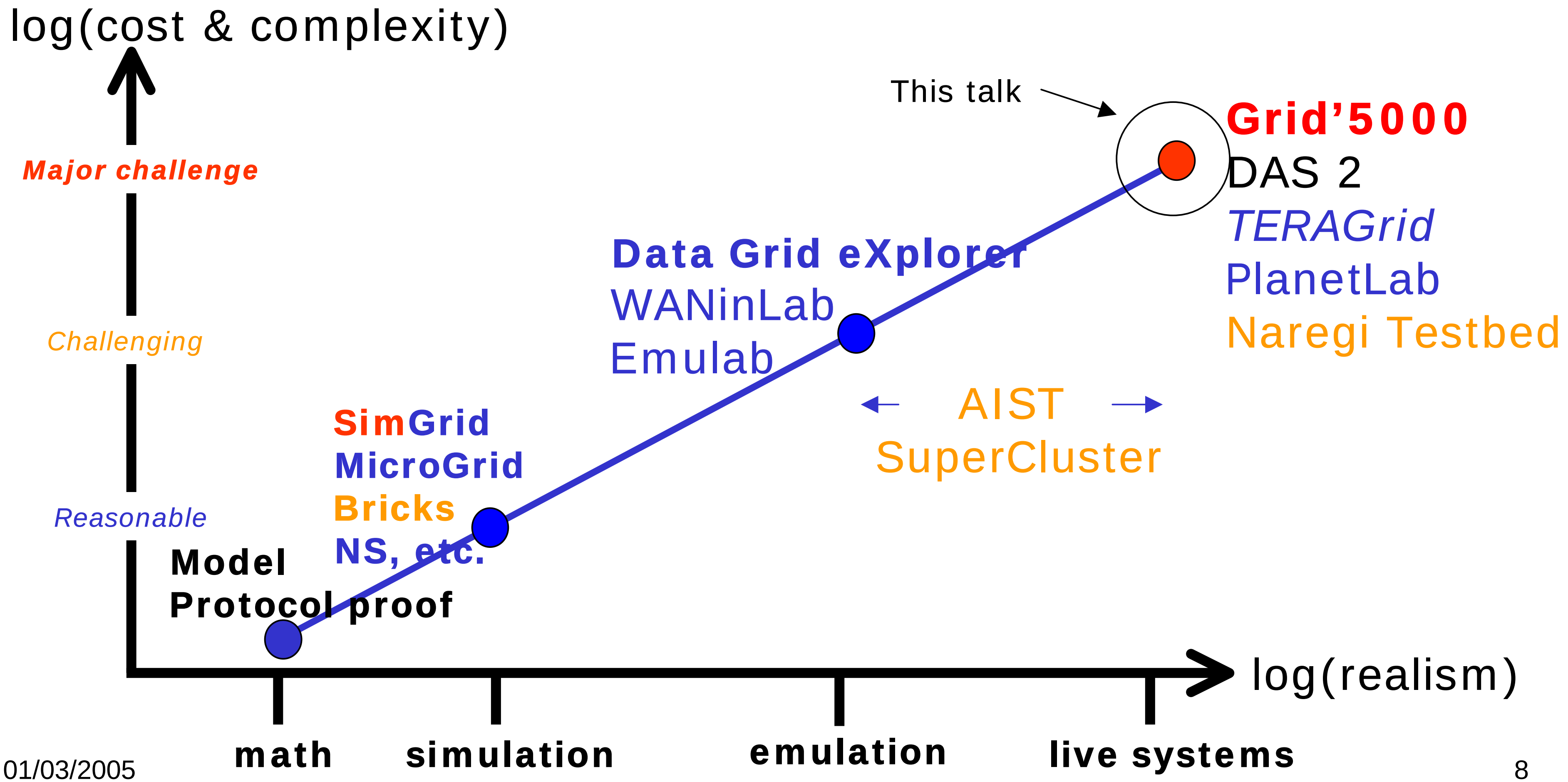
- SimGRid and SimGrid2
  - Discrete event simulation with trace injection
  - Originally dedicated to scheduling studies
- GridSim
  - Australian competitor of SimGrid
  - Dedicated to scheduling (with deadline)
- Titech Bricks
  - Discrete event simulation for scheduling and replication studies
- MicroGrid
  - Emulator with MPI communications
  - Not dynamic

→No emulator or real life experimental platform
→These tools do not scale (limited to ~100 grid nodes)
→They do not consider the network issues (almost)

We need Grid experimental tools

In the first ½ of 2003, the design and development of two Grid experimental platforms was decided:

→ Grid’5000 as a real life system
→ Data Grid eXplorer as a large scale emulator



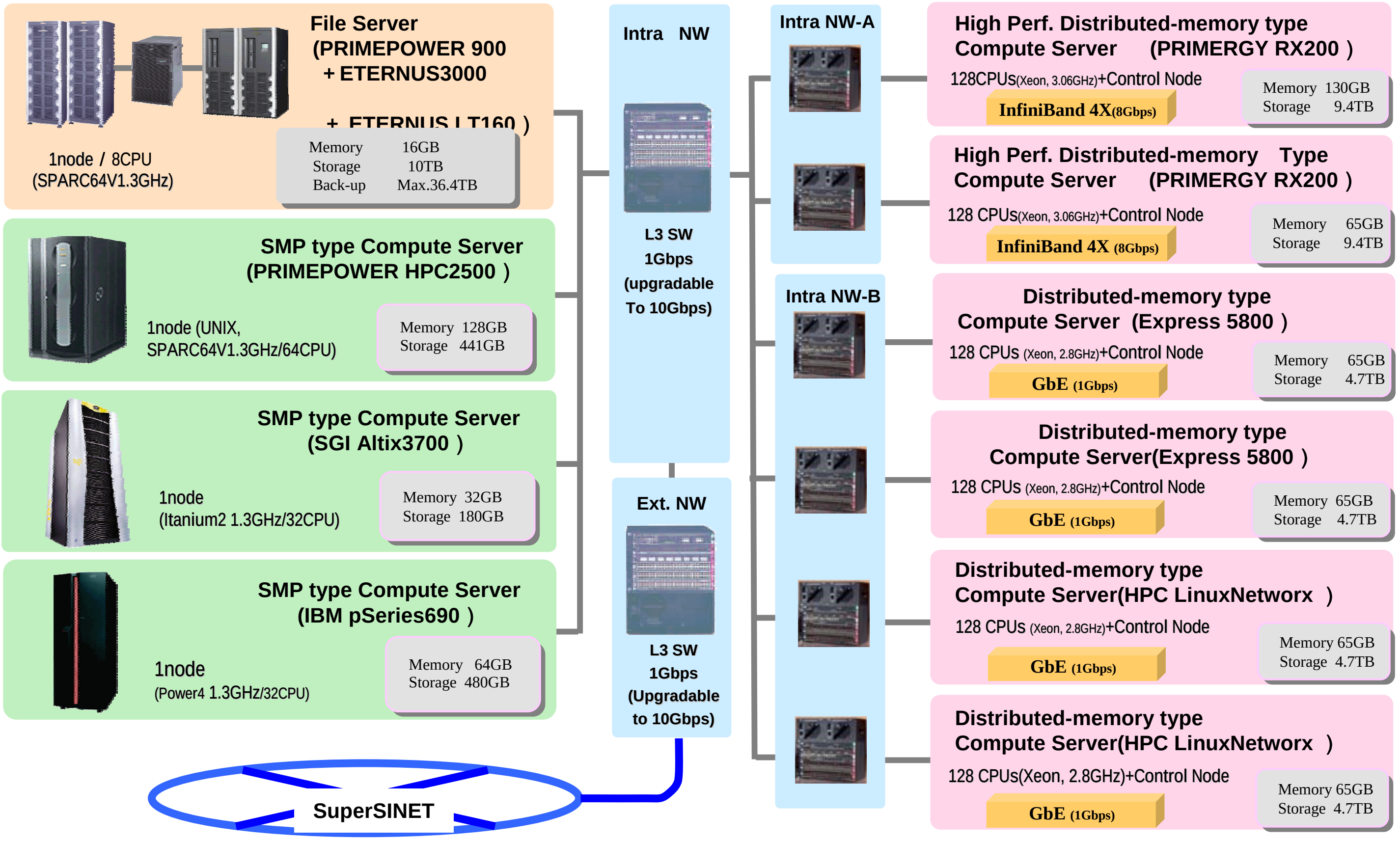
Ken
Miura

NAREGI Middleware Development Infrastructure

2003 - 2007

NOT a production system (c.f. TeraGrid) – Mainly geared towards R&D, but could be used partially for experimental production

To form a Grid testbed infrastructure (~ 10 Teraflops, March 2004)



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

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

Intra NW-B

High Perf. Distributed-memory type
Compute Server (PRIMERGY RX200)
128CPUs(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

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

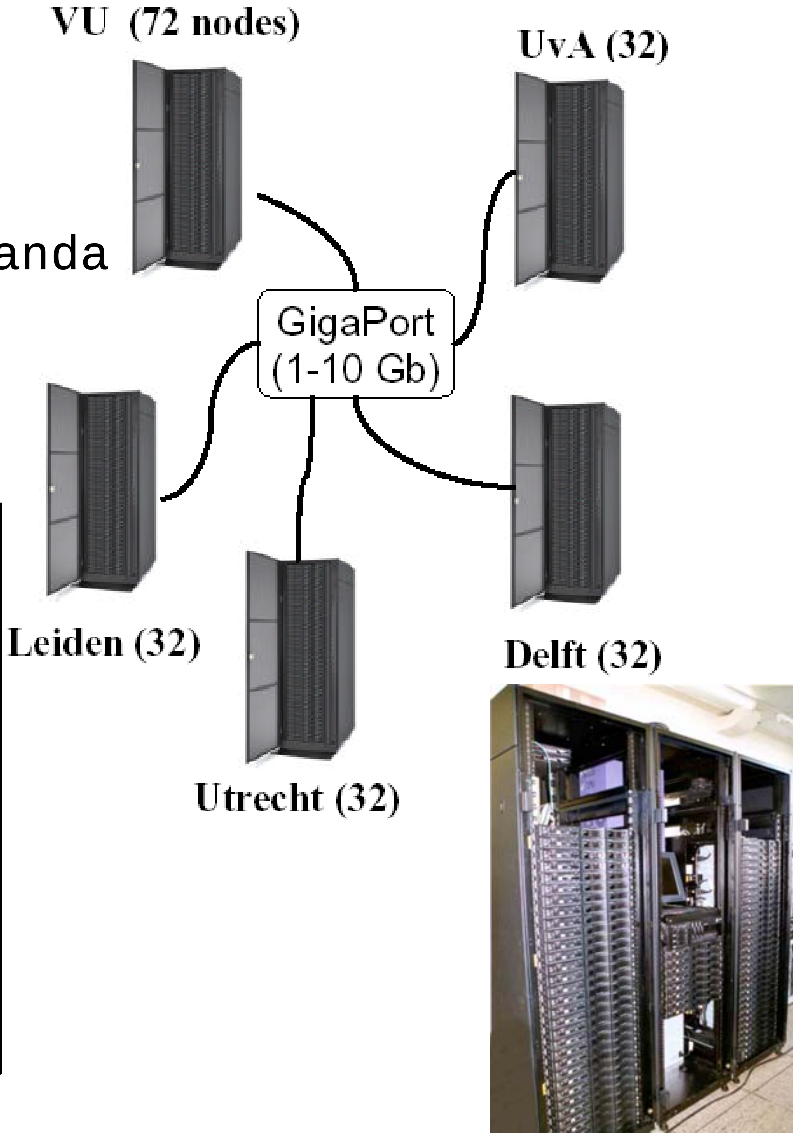
Distributed-memory type
Compute Server(HPC LinuxNetworx)
128 CPUs(Xeon, 2.8GHz)+Control Node
GbE (1Gbps)
Memory 65GB
Storage 4.7TB

01/03/2005

DAS2: 400 CPUs exp. Grid

- Homogeneous nodes!
- Grid middleware
  - Globus 3.2 toolkit
  - PBS+Maui scheduler
- Parallel programming support
  - MPI (MPICH-GM, MPICH-G2), PVM, Panda
  - Pthreads
- Programming languages
  - C, C++, Java, Fortran 77/90/95

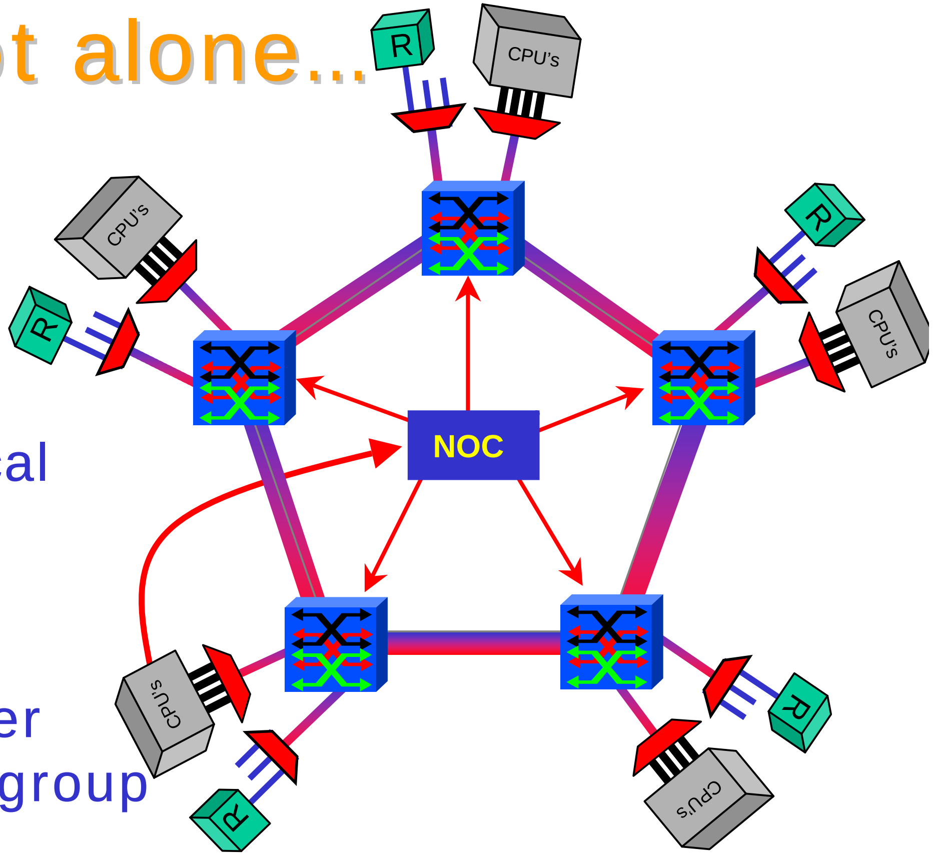
DAS2 (2002) :



	VU	UvA	Leiden	Delft	Utrecht
#nodes	72	32	32	32	32
Memory (GB)	1	1.5	1.5	1	1
Local disks (GB)	20	80	60	20	20
File server (GB)	6 * 36	6 * 36	6 * 36	2 * 18	2 * 18

We are not alone...

DAS-3



- Proposed next generation grid in the Netherlands
  - 5 clusters connected by optical network (SURFnet-6)
- Partners:
  - ASCI, VL-e, MultimediaN Gigaport-NG: DWDM computer backplane (dedicated optical group of 8 lambdas)
  - Application can dynamically allocate light paths, of 10 Gbit/sec
  - Application control topology through Network Operations Center
  - Gives flexible, dynamic, high-bandwidth links
  - Research questions :

    How to provide this flexibility (across domains)?

    How to integrate optical networks with applications?

➡ Collaboration with Grid'5000

Agenda

Rational

Grid’5000 project

Grid’5000 design

Grid’5000 developments

Conclusion

The Grid'5000 Project

1) Building a nation wide experimental platform for Grid researches (like a particle accelerator for the computer scientists)
   - 8 geographically distributed sites
   - every site hosts a cluster (from 256 CPUs to 1K CPUs)
   - All sites are connected by RENATER (French Res. and Edu. Net.)
   - RENATER hosts probes to trace network load conditions
   - Design and develop a system/middleware environment for safely test and repeat experiments

2) Use the platform for Grid experiments in real life conditions
   - Address critical issues of Grid system/middleware:
     - Programming, Scalability, Fault Tolerance, Scheduling
   - Address critical issues of Grid Networking
     - High performance transport protocols, Qos
   - Port and test applications
   - Investigate original mechanisms
     - P2P resources discovery, Desktop Grids

Funding & Participants

Funding (~7,6M€):

1) Ministry or Research
2) ACI GRID and MD (Hardware, Engineers)
3) INRIA (Hardware, Engineers)
4) CNRS (AS, Engineers, etc.)
5) Regional councils (Hardware)

Steering Committee (11) :

-**Franck Cappello** (Director)
-**Thierry Priol** (Director ACI Grid)
-**Brigitte Plateau** (Director CS ACI Grid)
-**Dany Vandrome** (Renater)
-Frédéric Desprez (Lyon)
-Michel Daydé (Toulouse)
-Yvon Jégou (Rennes)
-Stéphane Lantéri (Sophia)
-Raymond Namyst (Bordeaux)
-Pascale Primet (Lyon)
-Olivier Richard (Grenoble)

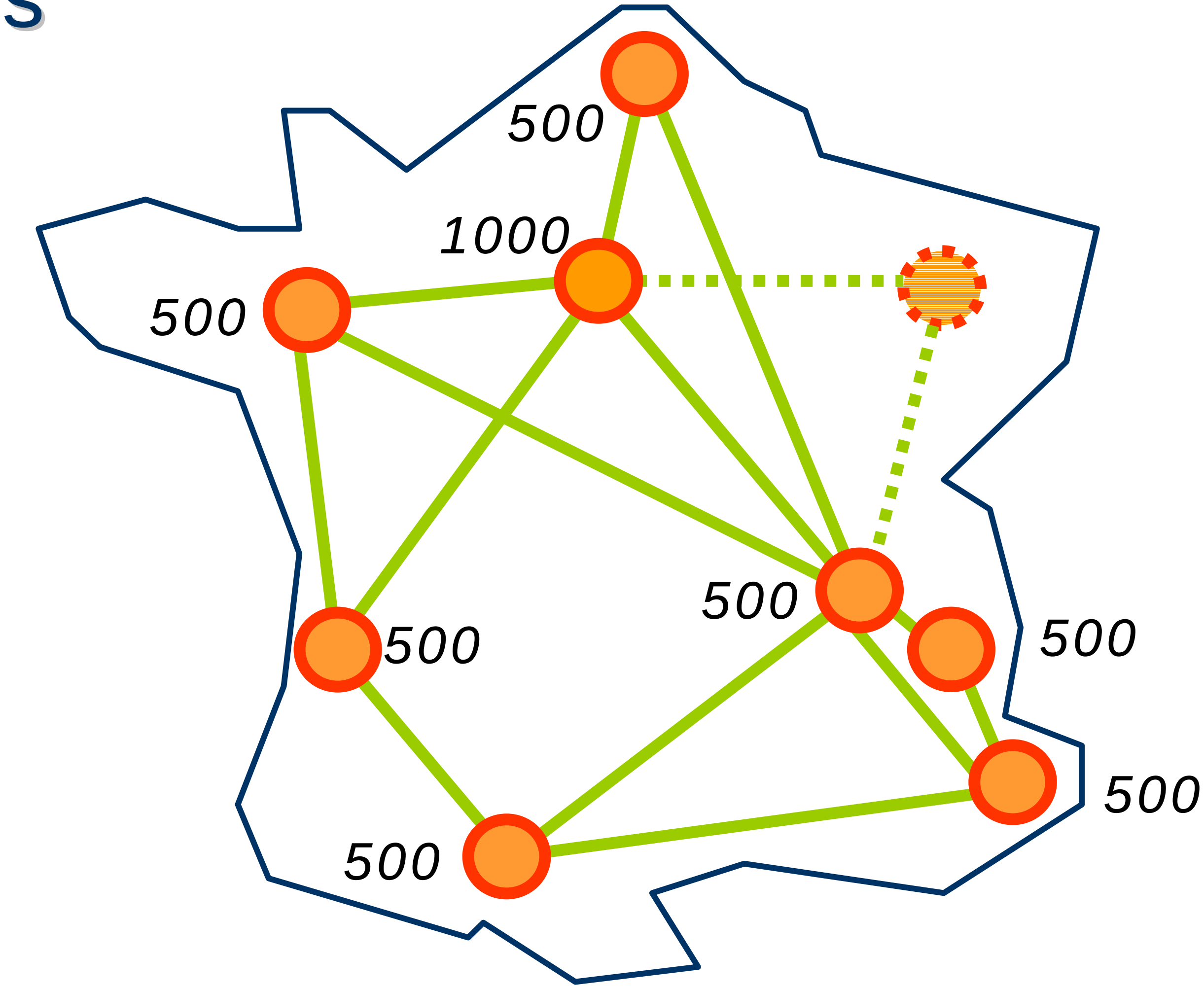
Technical Committee (28) :

Jean-Luc ANTHOINE
Jean-Claude Barbet
Pierrette Barbaresco
Nicolas Capit
Eddy Caron
Christophe Cérin
Olivier Coulaud
Georges Da-Costa
Yves Denneulin
Benjamin Dexheimer
Aurélien Dumez
Gilles Gallot
David Geldreich
Sébastien Georget
Olivier Gluck
Julien Leduc
Cyrille Martin
Jean-Francois Méhaut
Jean-Christophe Mignot
Thierry Monteil
Guillaume Mornet
Alain Naud
Vincent Néri
Gaetan Peaquin
Franck Simon
Sebastien Varrette
Jean-Marc Vincent

Grid'5000

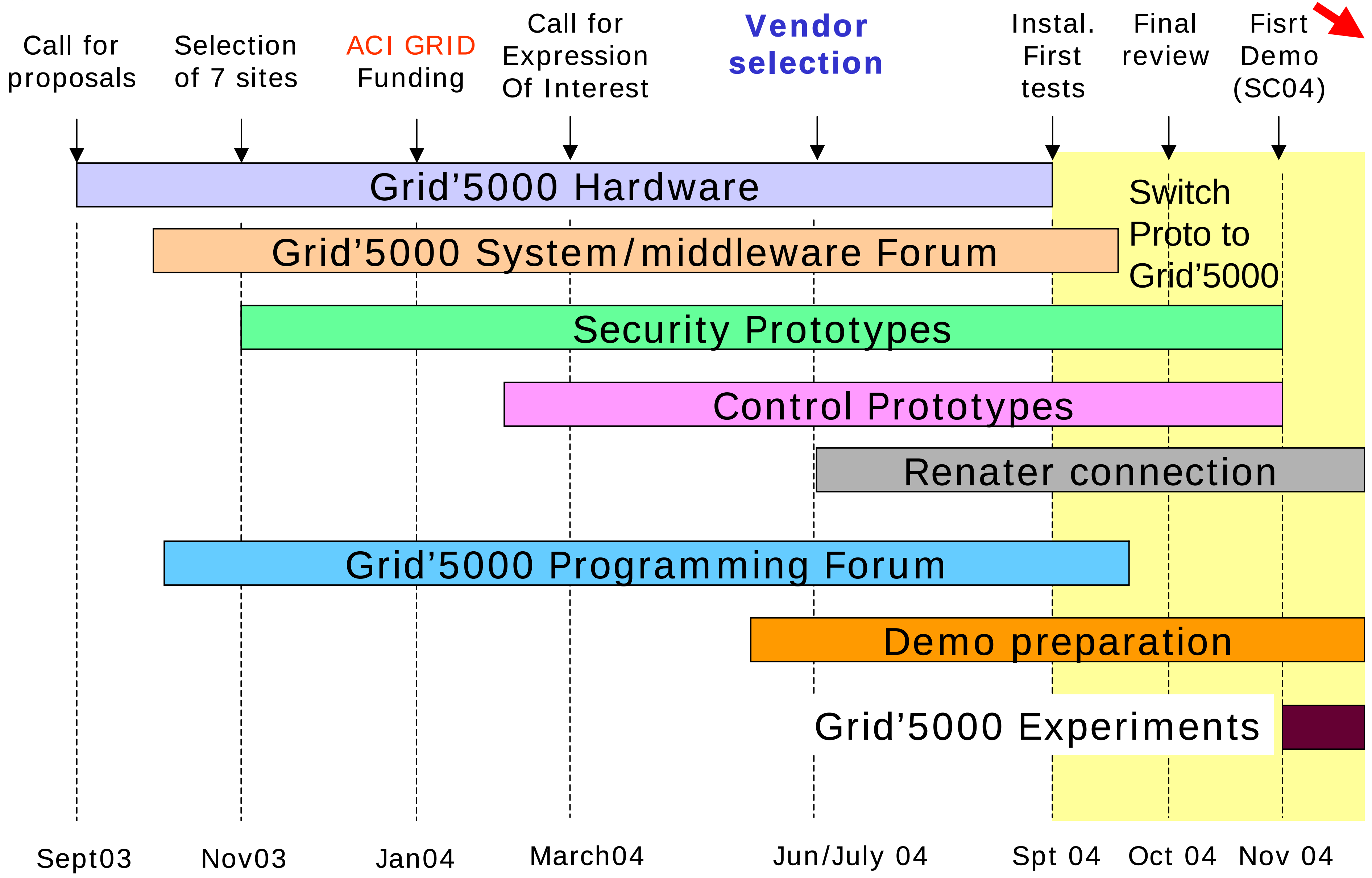
Grid'5000 map

The largest Instrument to study Grid issues



Schedule

today



Call for proposals

Selection of 7 sites

ACI GRID Funding

Call for Expression Of Interest

Vendor selection

Instal. First tests

Final review

Fisrt Demo (SC04)

Grid'5000 Hardware

Grid'5000 System/middleware Forum

Switch Proto to Grid'5000

Security Prototypes

Control Prototypes

Renater connection

Grid'5000 Programming Forum

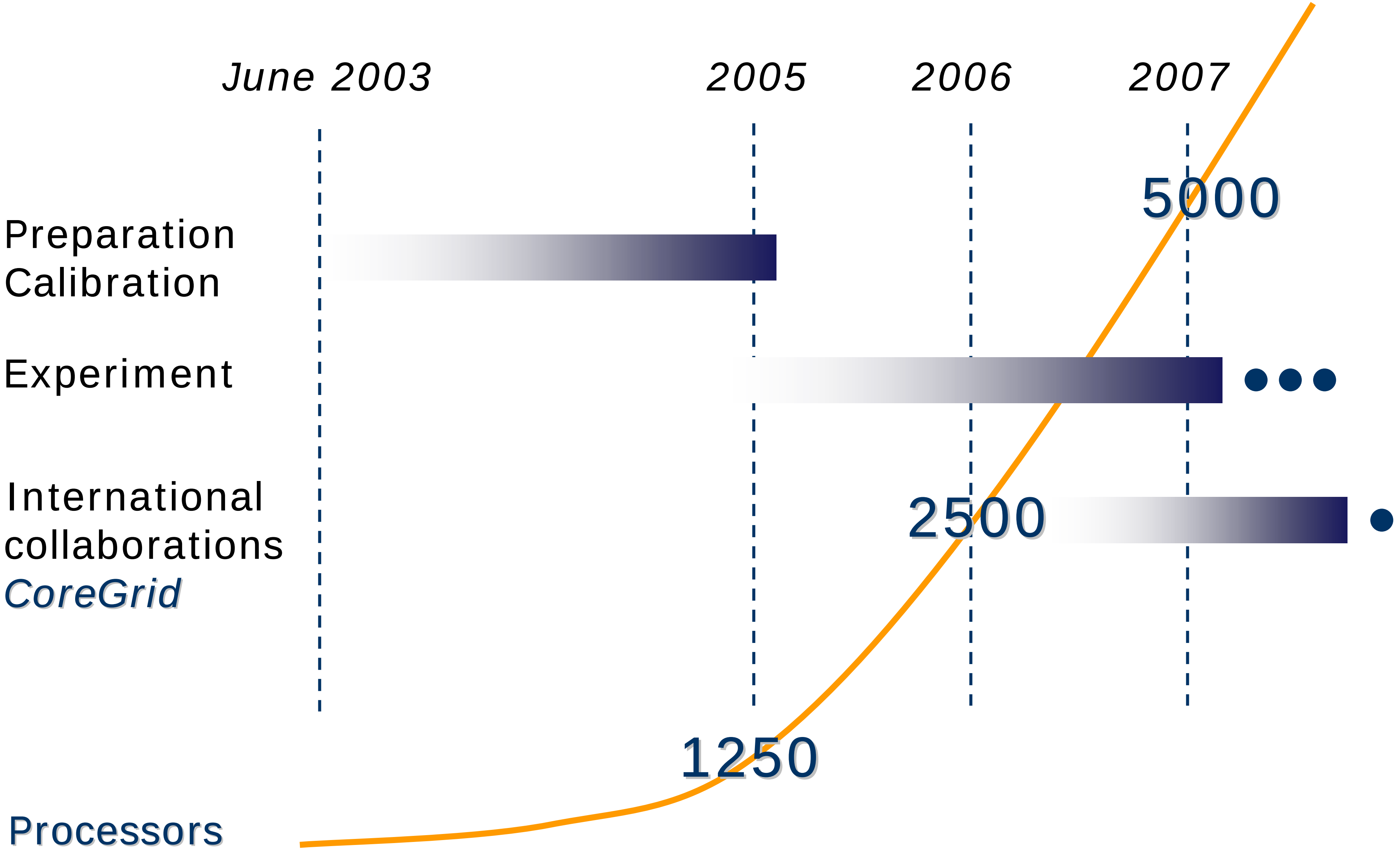
Demo preparation

Grid'5000 Experiments

Sept03 Nov03 Jan04 March04 Jun/July 04 Spt 04 Oct 04 Nov 04

Today: still switching + configuring. First runs on February 9 2005

Planning


June 2003
2005
2006
2007
Preparation
Calibration
Experiment
International
collaborations
CoreGrid
5000
2500
1250
Processors

Agenda

Rational

Grid'5000 project

Grid'5000 design

Grid'5000 developments

Conclusion

Grid'5000 foundations: Collection of experiments to be done

- Networking
  - End host communication layer (interference with local communications)
  - High performance long distance protocols (improved TCP)
  - High Speed Network Emulation
- Middleware / OS
  - Scheduling / data distribution in Grid
  - Fault tolerance in Grid
  - Resource management
  - Grid SSI OS and Grid I/O
  - Desktop Grid/P2P systems
- Programming
  - Component programming for the Grid (Java, Corba)
  - GRID-RPC
  - GRID-MPI
  - Code Coupling
- Applications
  - Multi-parametric applications (Climate modeling/Functional Genomic)
  - Large scale experimentation of distributed applications (Electromagnetism, multi-material fluid mechanics, parallel optimization algorithms, CFD, astrophysics
  - Medical images, Collaborating tools in virtual 3D environment

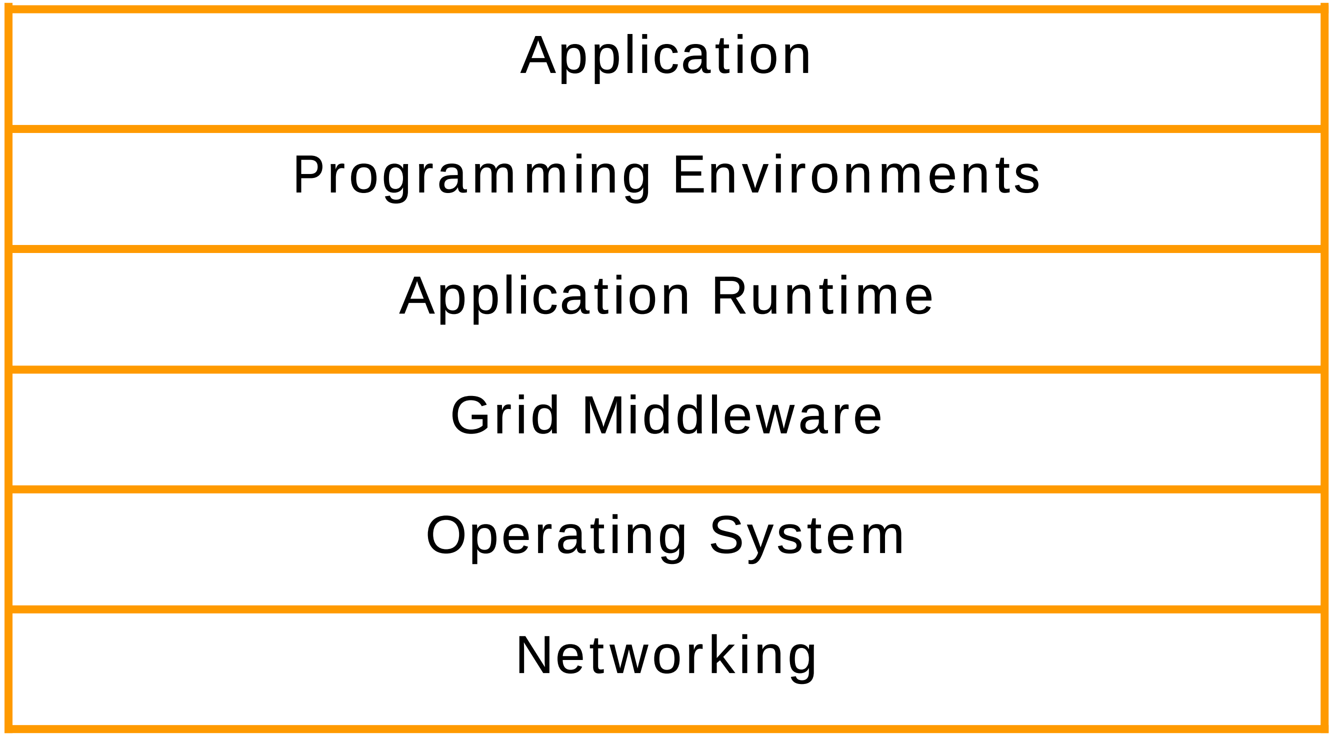
Grid'5000 foundations: Collection of properties to evaluate

Quantitative metrics :

- Performance
  - Execution time, throughput, overhead
- Scalability
  - Resource occupation (CPU, memory, disc, network)
  - Applications algorithms
  - Number of users
- Fault-tolerance
  - Tolerance to very frequent failures (volatility), tolerance to massive failures (a large fraction of the system disconnects)
  - Fault tolerance consistency across the software stack.

Grid'5000 goal:

Experimenting all layers of the Grid software stack



Application

Programming Environments

Application Runtime

Grid Middleware

Operating System

Networking

→ A highly reconfigurable experimental platform

Grid'5000 Vision

Grid'5000 is NOT a production Grid!

Grid'5000 should be:

- an instrument

  to experiment all levels of the software stack involved in Grid.

Grid'5000 will be:

- a low level testbed harnessing clusters (a nation wide cluster of clusters),

  allowing users to fully configure the cluster nodes (including the OS) for their experiments (strong control)

Agenda

Rational

Grid'5000 project

Grid'5000 design

Grid'5000 developments

Conclusion

Security design

- Grid'5000 nodes will be rebooted and configured at kernel level by users (very high privileges for every users);

  → Users may configure incorrectly the cluster nodes opening security holes

- How to secure the local site and Internet?

→ A confined system (no way to get out; access only through strong authentication and via a dedicated gateway)

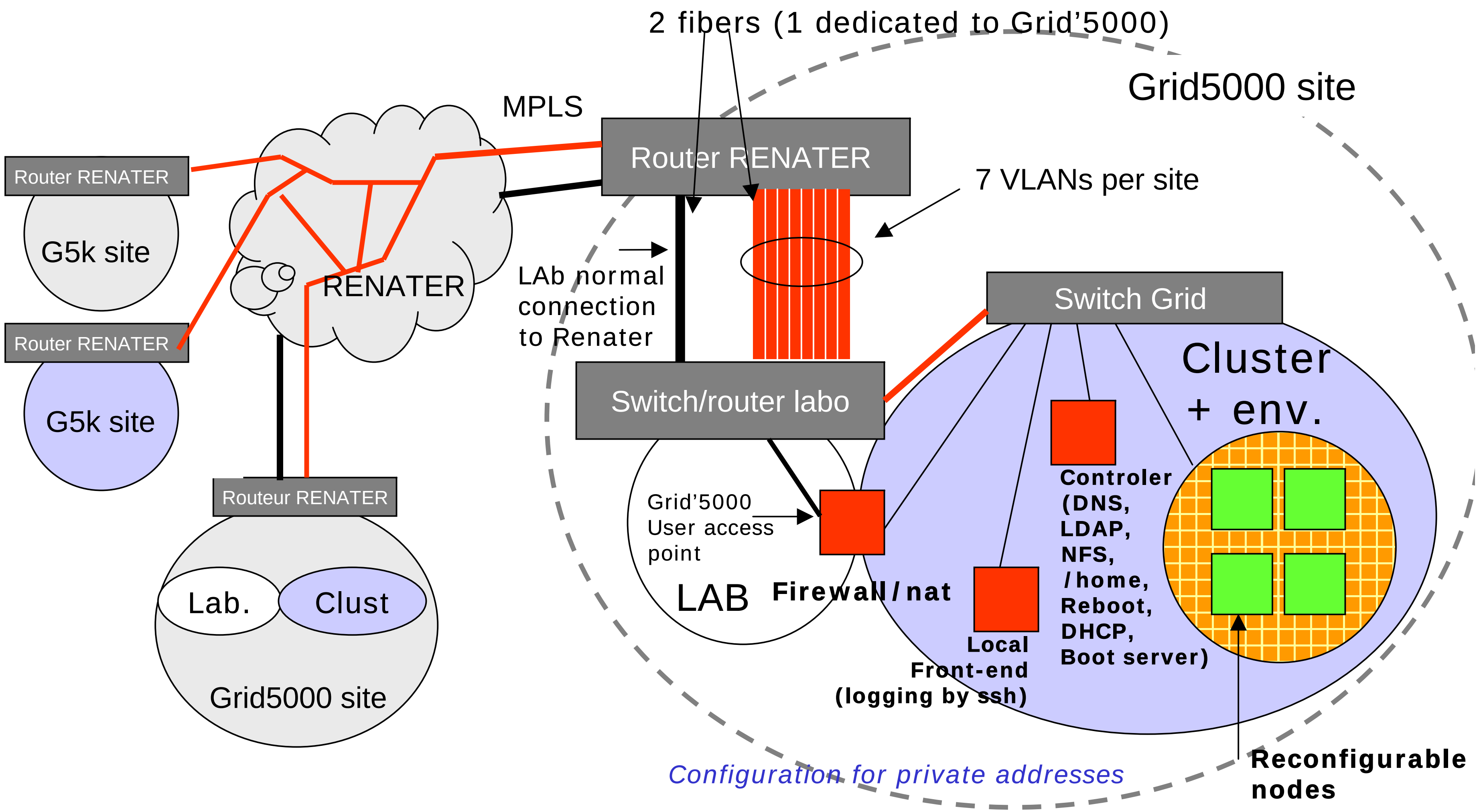
- Some sites want private addresses, some others want public addresses

- Some sites want to connect satellite machines

→ Access is granted only from sites

→ Every site is responsible to following the confinement rules

Grid'5000 Security architecture: A confined system



2 fibers (1 dedicated to Grid'5000)
Grid5000 site
MPLS
Router RENATER
7 VLANs per site
Router RENATER
G5k site
RENATER
LAb normal connection to Renater
Switch Grid
Cluster + env.
Router RENATER
G5k site
Switch/router labo
Controler (DNS, LDAP, NFS, /home, Reboot, DHCP, Boot server)
Routeur RENATER
Grid'5000 User access point
Lab.
Clust
LAB
Firewall/nat
Local Front-end (logging by ssh)
Grid5000 site
Configuration for private addresses
Reconfigurable nodes

8 x 7 VLANs in Grid'5000 (1 VLAN per tunnel)

Control design

- User want to be able to install on all Grid'5000 nodes some specific software stack from network protocols to applications (possibly including kernel)

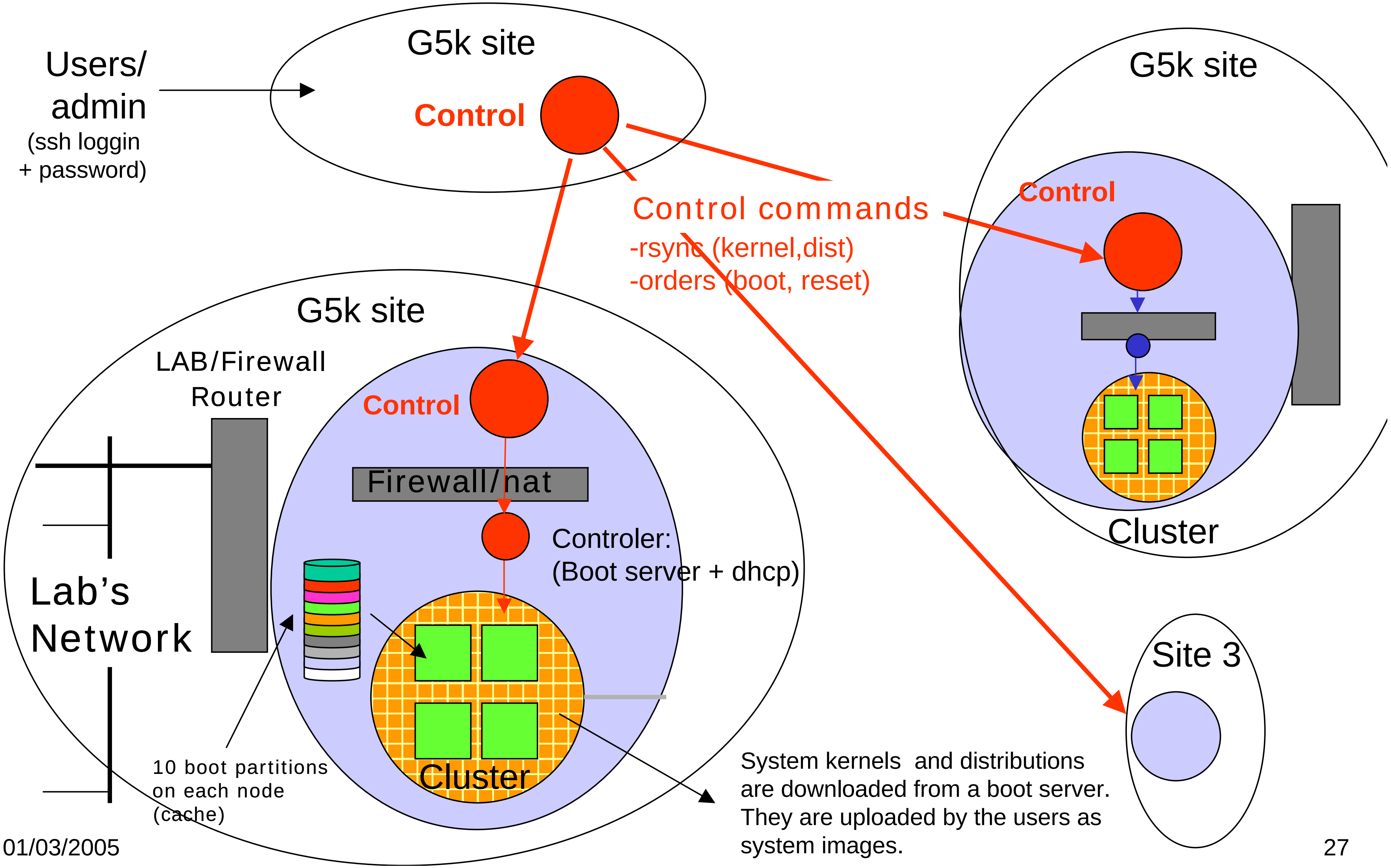
- Administrators want to be able to reset/reboot distant nodes in case of troubles

- Grid'5000 developers want to develop control mechanisms in order to help debugging, such as "step" by "step" execution (relying on checkpoint/restart mechanisms)

→ A control architecture allowing to broadcast orders from one site to the others with local relays to convert the order into actions

Control Architecture

In reserved and batch modes, admins and users can control their resources

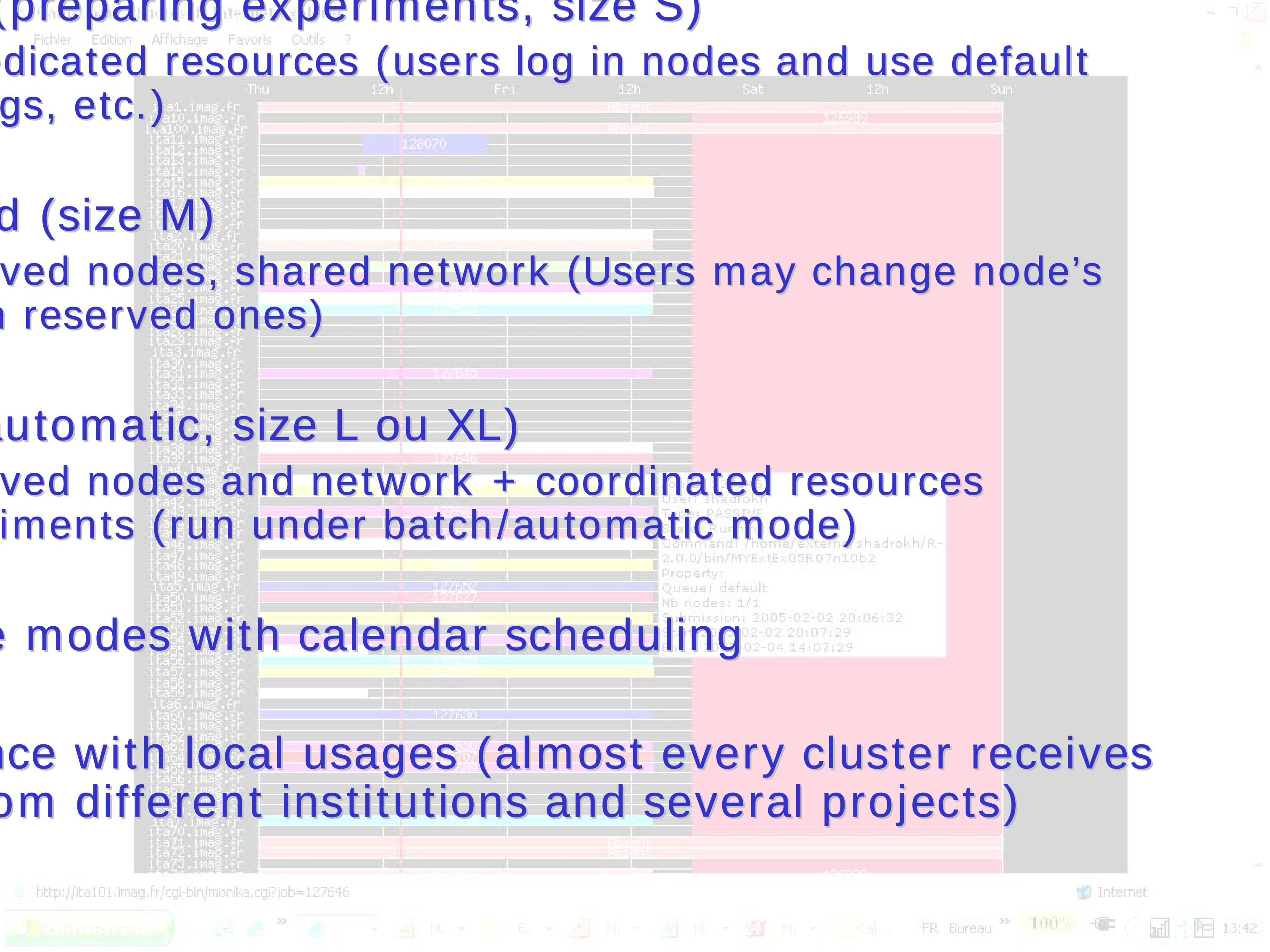


System kernels and distributions are downloaded from a boot server. They are uploaded by the users as system images.

Usage modes

- Shared (preparing experiments, size S)
  - No dedicated resources (users log in nodes and use default settings, etc.)
- Reserved (size M)
  - Reserved nodes, shared network (Users may change node's OS on reserved ones)
- Batch (automatic, size L ou XL)
  - Reserved nodes and network + coordinated resources experiments (run under batch/automatic mode)
- All these modes with calendar scheduling

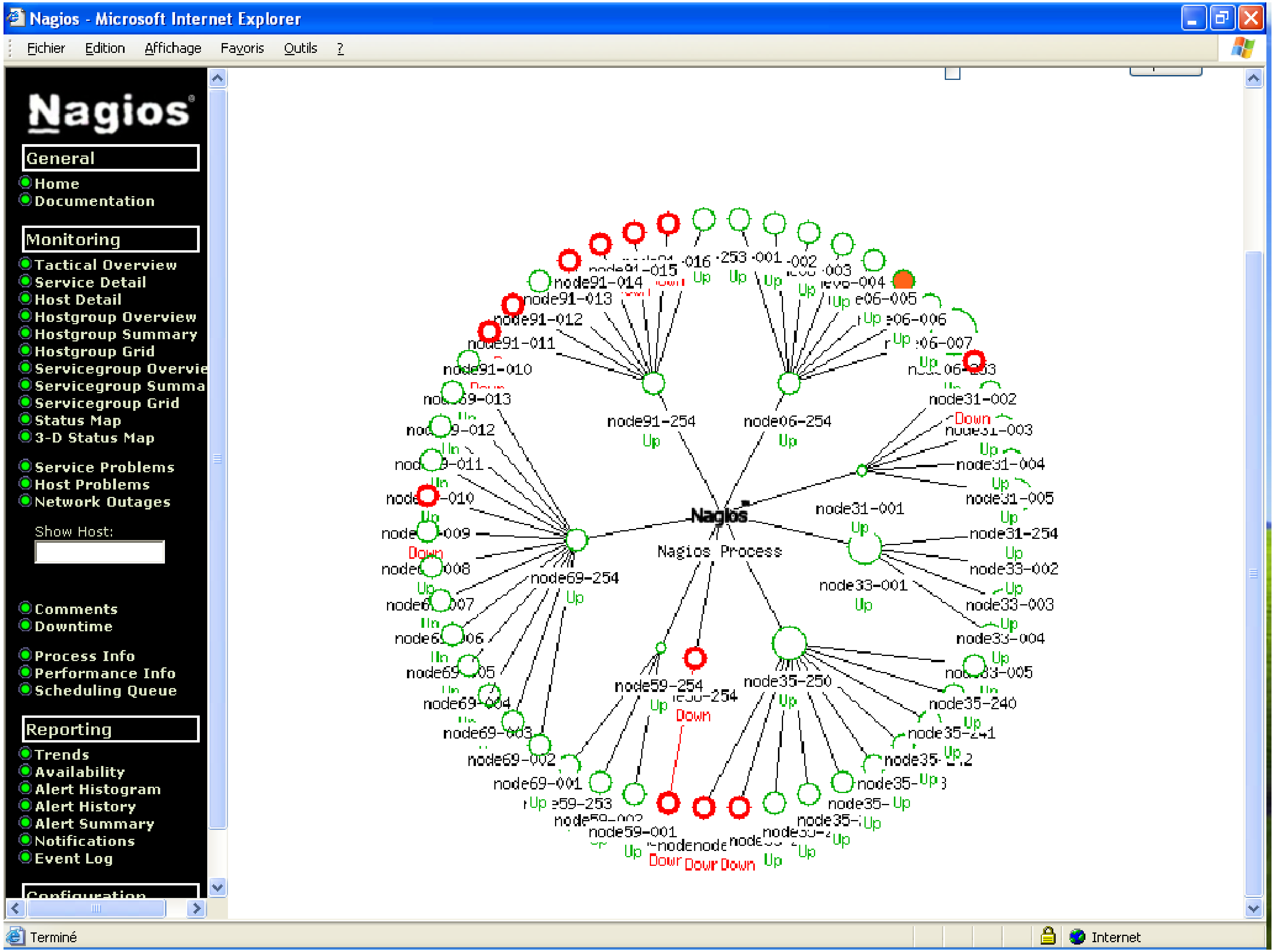
+ compliance with local usages (almost every cluster receives funds from different institutions and several projects)

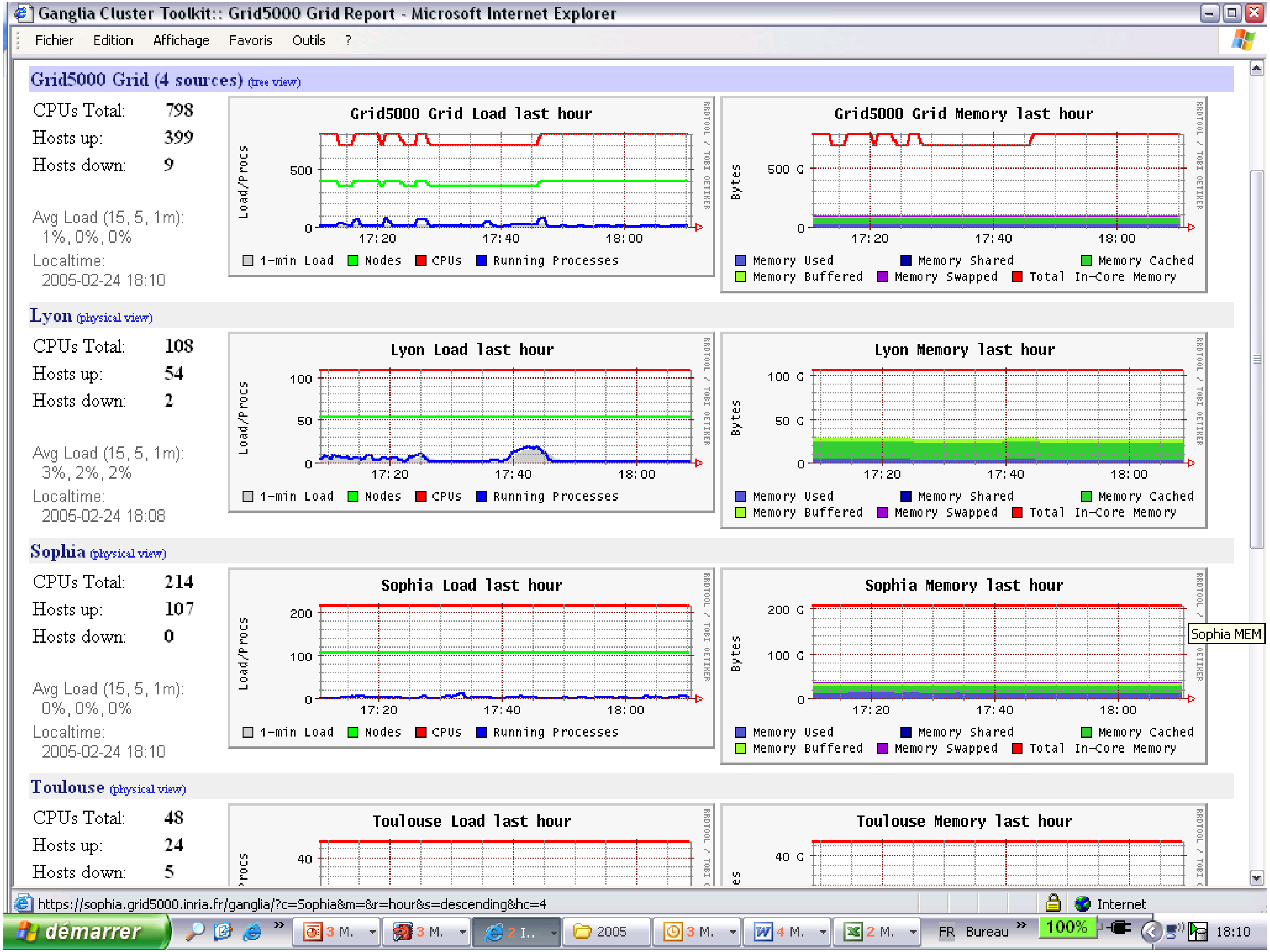


User: shadrokh
Type: PASSIVE
Command: /home/externe/shadrokh/R-2.0.0/bin/MYExtEx05R07n10b2
Property:
Queue: default
Nb nodes: 1/1
Submission: 2005-02-02 20:06:32
02-02 20:07:29
02-04 14:07:29

Rennes
Lyon
Sophia
Grenoble
Orsay
Toulouse

Grid'5000


Nagios - Microsoft Internet Explorer
Fichier
Edition
Affichage
Favoris
Outils
?
Nagios
General
Home
Documentation
Monitoring
Tactical Overview
Service Detail
Host Detail
Hostgroup Overview
Hostgroup Summary
Hostgroup Grid
Servicegroup Overvie
Servicegroup Summa
Servicegroup Grid
Status Map
3-D Status Map
Service Problems
Host Problems
Network Outages
Show Host:
Comments
Downtime
Process Info
Performance Info
Scheduling Queue
Reporting
Trends
Availability
Alert Histogram
Alert History
Alert Summary
Notifications
Event Log
Configuration
Nagios
Nagios Process
node91-254
Up
node06-254
Up
node31-001
Up
node33-001
Up
node35-250
Up
node59-254
Up
node50-254
Down
node69-254
Up
node91-015
node91-014
node91-013
node91-012
node91-011
node91-010
node69-013
node69-012
node69-011
node69-010
node69-009
node69-008
node69-007
node69-006
node69-005
node69-004
node69-003
node69-002
node69-001
node59-253
node59-002
node59-001
node31-002
node31-003
node31-004
node31-005
node31-254
node33-002
node33-003
node33-004
node33-005
node35-240
node35-241
node06-005
node06-006
node06-007
Up
Down
Terminé
Internet


Ganglia Cluster Toolkit:: Grid5000 Grid Report - Microsoft Internet Explorer
Fichier
Edition
Affichage
Favoris
Outils
?

Grid5000 Grid (4 sources) (tree view)

CPUs Total: 798
Hosts up: 399
Hosts down: 9

Avg Load (15, 5, 1m):
1%, 0%, 0%

Localtime:
2005-02-24 18:10

Lyon (physical view)

CPUs Total: 108
Hosts up: 54
Hosts down: 2

Avg Load (15, 5, 1m):
3%, 2%, 2%

Localtime:
2005-02-24 18:08

Sophia (physical view)

CPUs Total: 214
Hosts up: 107
Hosts down: 0

Avg Load (15, 5, 1m):
0%, 0%, 0%

Localtime:
2005-02-24 18:10

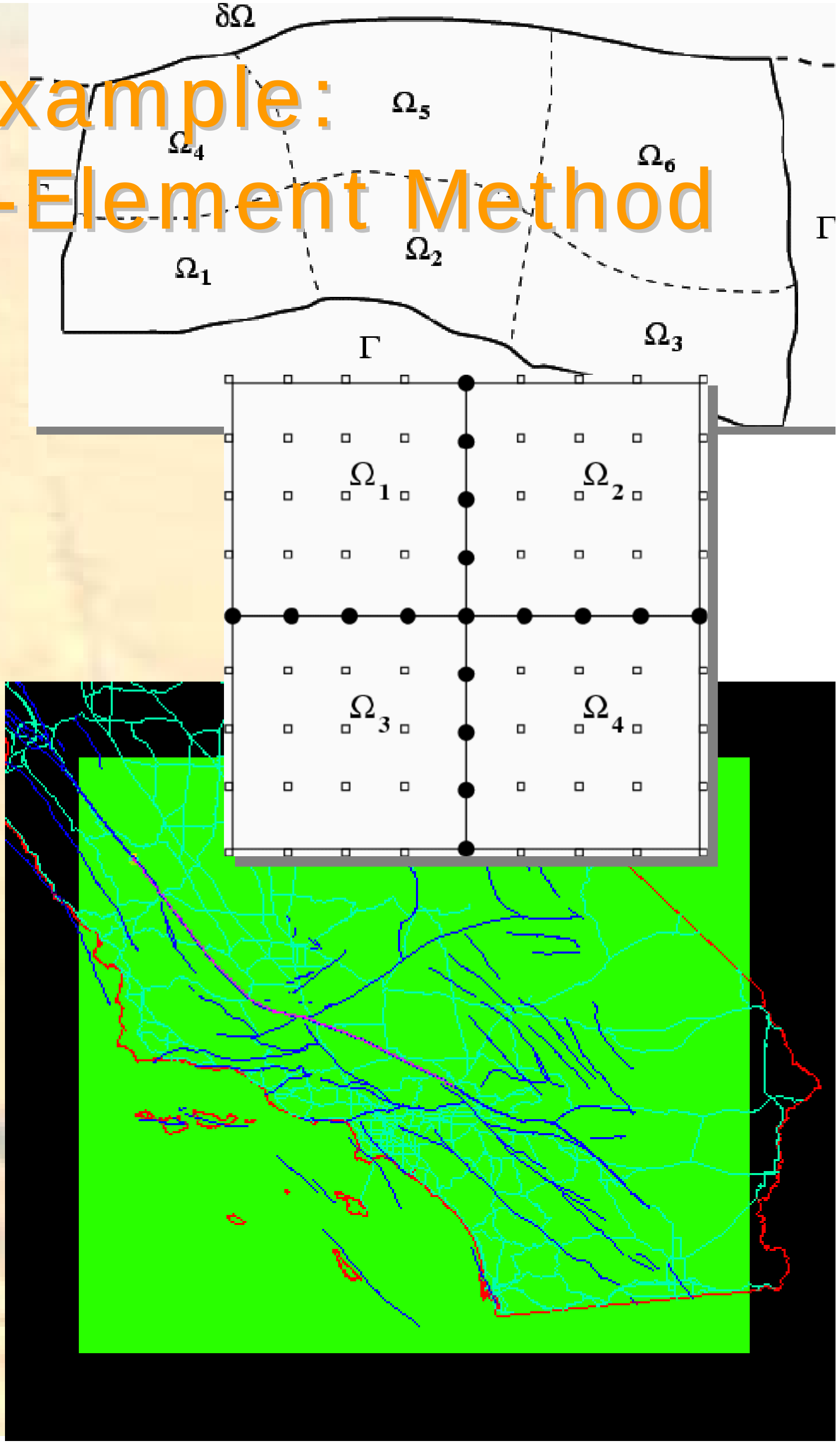
Toulouse (physical view)

CPUs Total: 48
Hosts up: 24
Hosts down: 5

https://sophia.grid5000.inria.fr/ganglia/?c=Sophia&m=&r=hour&s=descending&hc=4

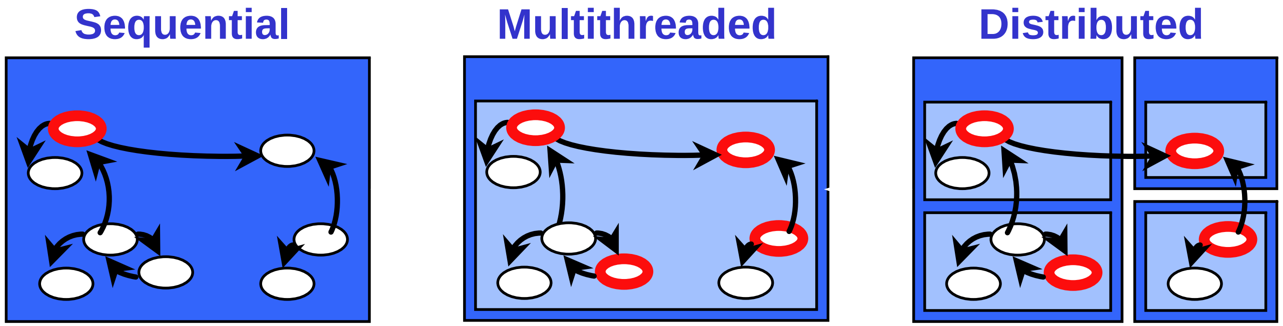
Experiment example:
SPECFEM3D: Spectral-Element Method

- Developed in Computational Fluid Dynamics (Patera 1984)
- Introduced by Chaljub (2000) at IPG Paris
- Extended by Komatitsch and Tromp, Capdeville et al.
- 5120 CPUs (640 x 8), 10 terabytes of mem. (Earthsim.)
- SPECFEM3D wan Gordon Bell price at SuperComputing’2003
- How to adapt it for the Grid?



Experiment example: testing Grid programming models

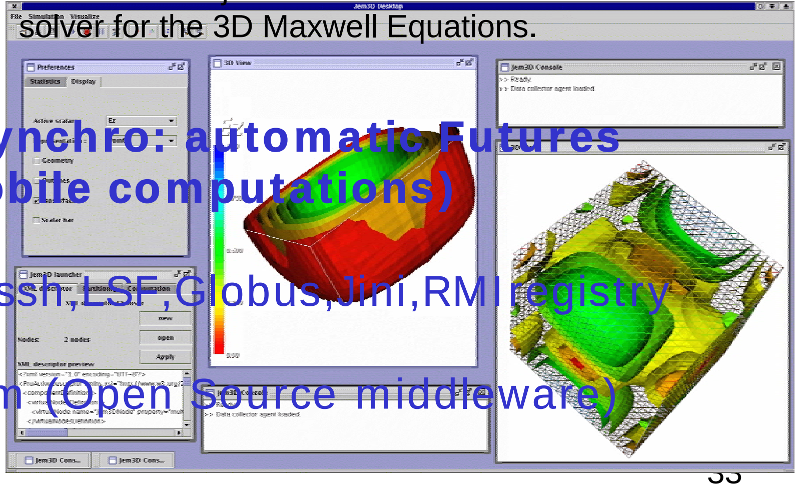
A Java API + Tools for Parallel, Distributed Computing



- **A uniform framework:** **An Active Object pattern**
- **A formal model behind:** **Prop. Determinism**

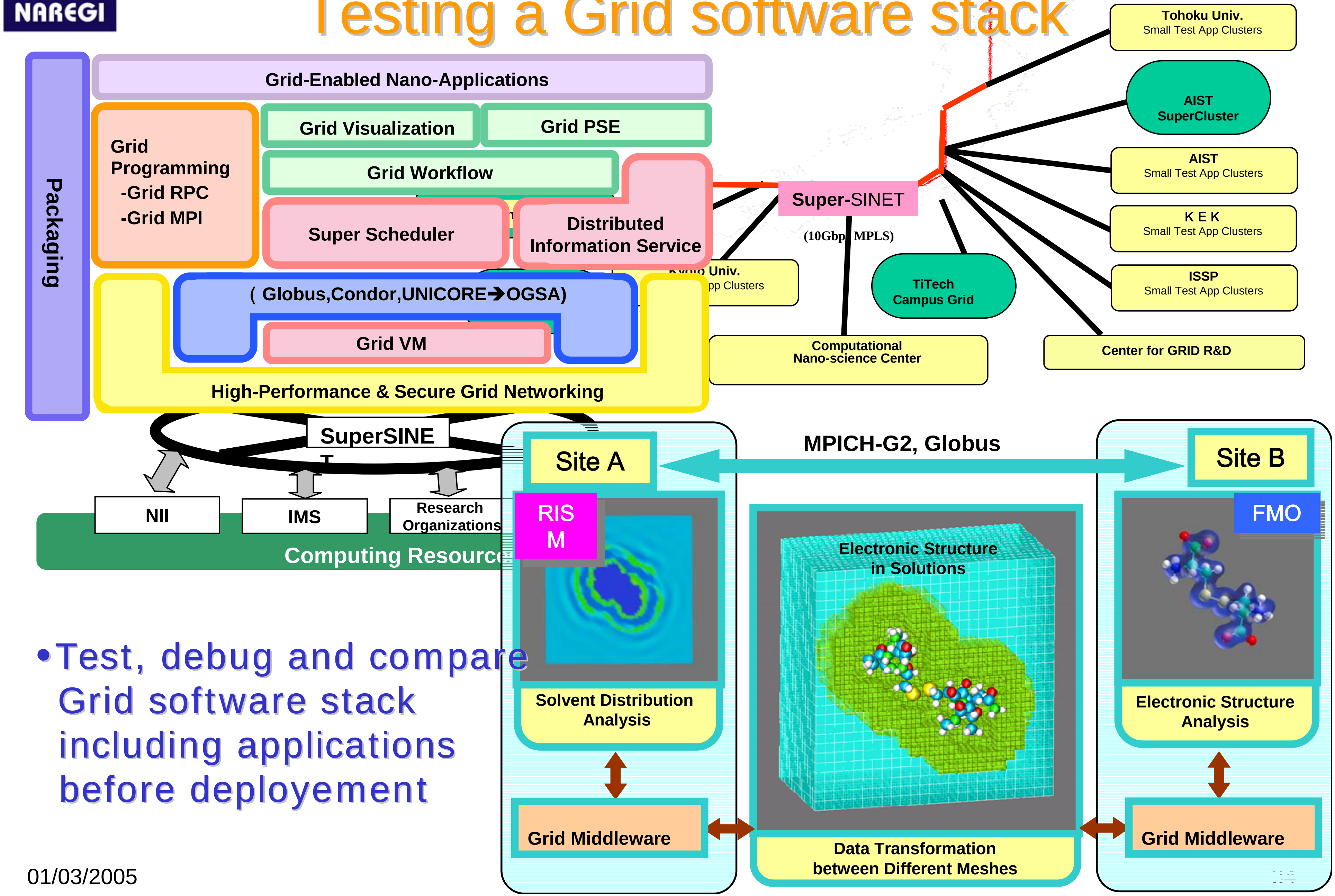
- **Main features:**
- **Remotely accessible Objects**
- **Asynchronous Communications with synchro: automatic Futures**
- **Group Communications, Migration (mobile computations)**
- **XML Deployment Descriptors**
- **Interfaced with various protocols:** rsh,ssh,LSF,Globus,Jini,RMIregistry
- **Visualization and monitoring: IC2D**
- In the www. ObjectWeb .org Consortium (Open Source middleware)
- since April 2002 (**LGPL license)**

JEM3D: an object-oriented time domain finite volume solver for the 3D Maxwell Equations.



```
>> Ready
>> Data collector agent loaded.
```

Experiment example: Testing a Grid software stack



- Test, debug and compare Grid software stack including applications before deployement

01/03/2005

Agenda

Rational

Grid’5000 project

Grid’5000 design

Grid’5000 developments

Conclusion

Summary

- The largest Instrument for research in Grid Computing

- Grid'5000 will offer in 2005:
  - 8 clusters distributed over 8 sites in France,
  - about 2500 CPUs,
  - about 2,5 TB memory,
  - about 100 TB Disc,
  - about 8 Gigabit/s (directional) of bandwidth
  - about 5 à 10 Tera operations / sec
  - the capability for all users to reconfigure the platform [protocols/OS/Middleware/Runtime/Application]

- Grid'5000 will be opened to Grid researchers in early 2005

- International extension currently under discussion (Netherlands, Japan)