A nation wide Experimental Grid

Franck Cappello
INRIA
fci@lri.fr
With all participants

INRIA FUTURS
MINISTERE DE LA RECHERCHE
ACI GRID
Action Concertée Incitative [ACI]
Globalisation des Ressources Informatiques et des Données [GRID]
ministère jeunesse éducation recherche
ministère délégué recherche et nouvelles technologies
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Agenda

Motivation

Grid'5000 project

Grid'5000 design

Grid'5000 developments

Conclusion

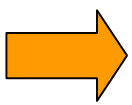
Grid raises research issues but also methodological challenges

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, etc.



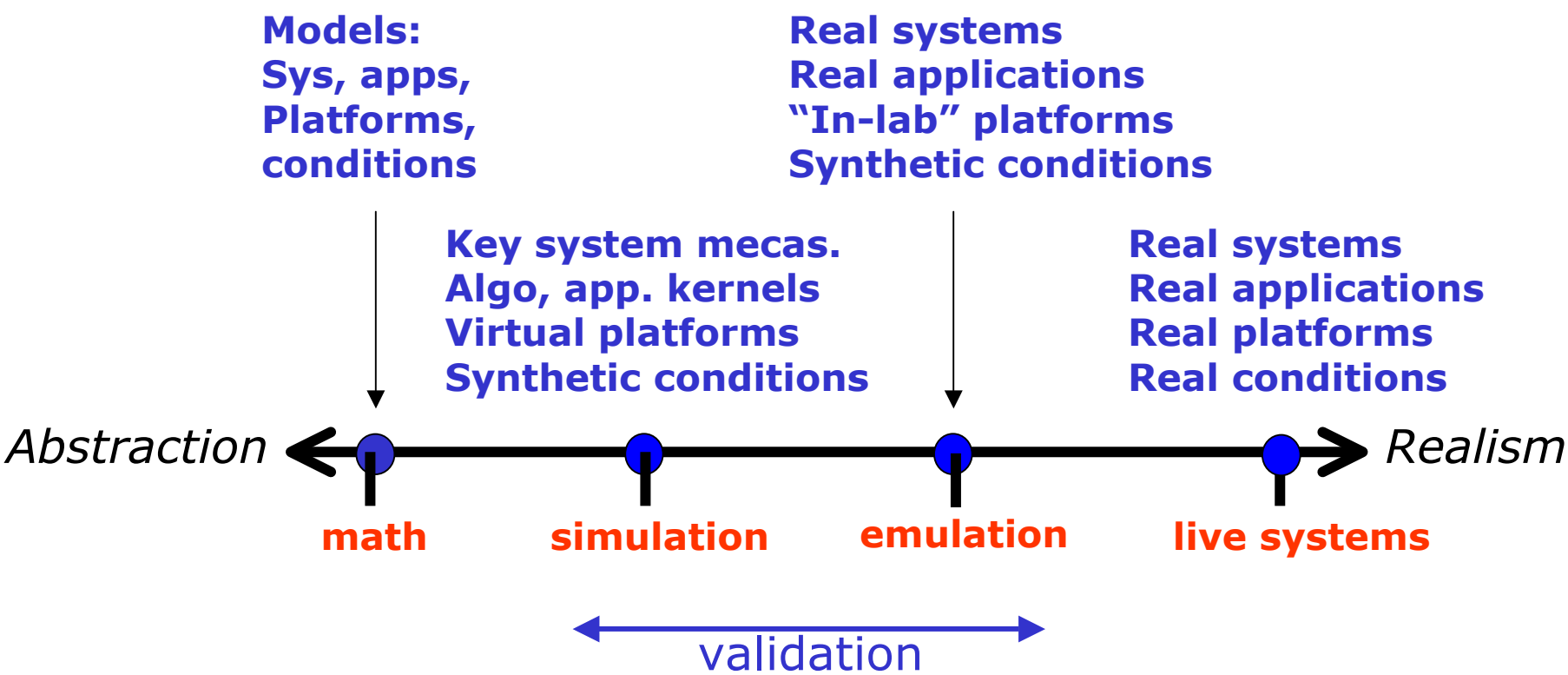
How to test and compare?

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

Tools for Distributed System Studies

To investigate Distributed System issues, we need:

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



2) Strong interaction between these research tools

Existing Grid Research Tools

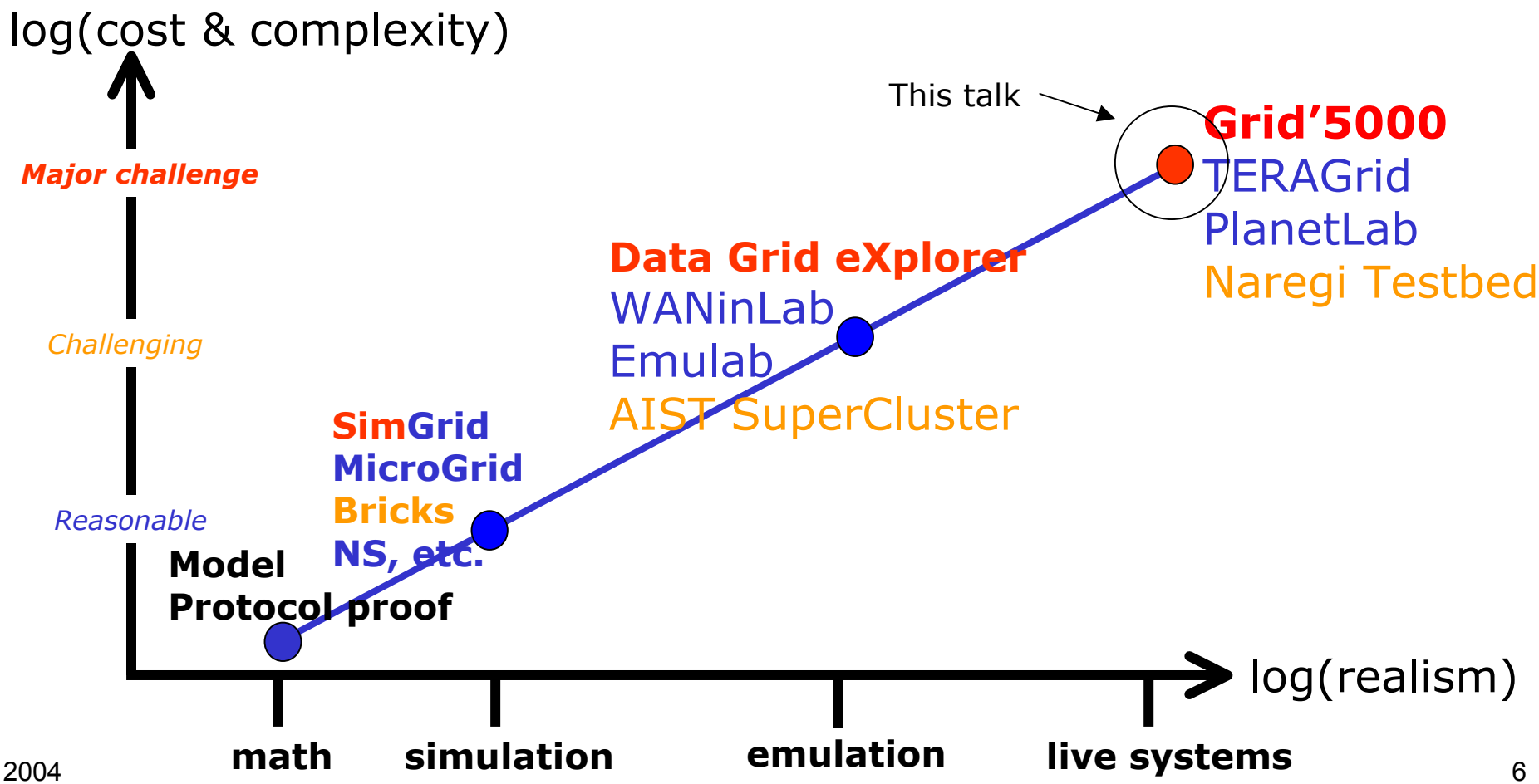
- SimGRid and SimGrid2
  - Discrete event simulation with trace injection
  - Originally dedicated to scheduling studies
- GridSim
  - Australian competitor of SimGrid
  - Dedicated to scheduling (with deadline)
- Titech Bricks
  - Discrete event simulation for scheduling and replication studies
- MicroGrid
  - 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



NAREGI Middleware Development Infrastructure

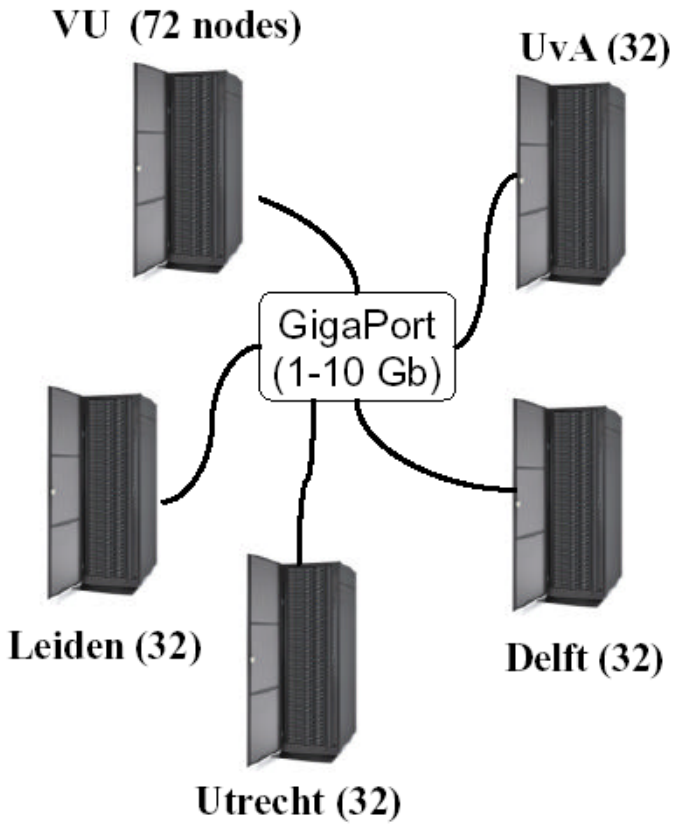
- Installation in Dec. 2003
  - 3 SMPs, 128 procs total
  - 6 x 128-proc clusters, with different interconnects
  - 1 File Server
  - Multi-gigabit networking to simulate Grid Environment
  - NOT a production system (c.f. TeraGrid) – Mainly geared towards R&D, but could be used partially for experimental production
  - ~5 Teraflops
  - To form a Grid with the IMS NAREGI application testbed infrastructure (~ 10 Teraflops, March 2004), and other national centers(voluntary basis) via SuperSINET

Some Grid testbeds in Europe

Grid Lab

GridLab testbed

DAS2 (2002) :



DAS3 →2005

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:

1) ACI GRID (Hardware)
2) INRIA (Hardware, Engineers)
3) CNRS (AS, Engineers, etc.)
4) Regional councils (Hardware)

Steering Committee (11) :

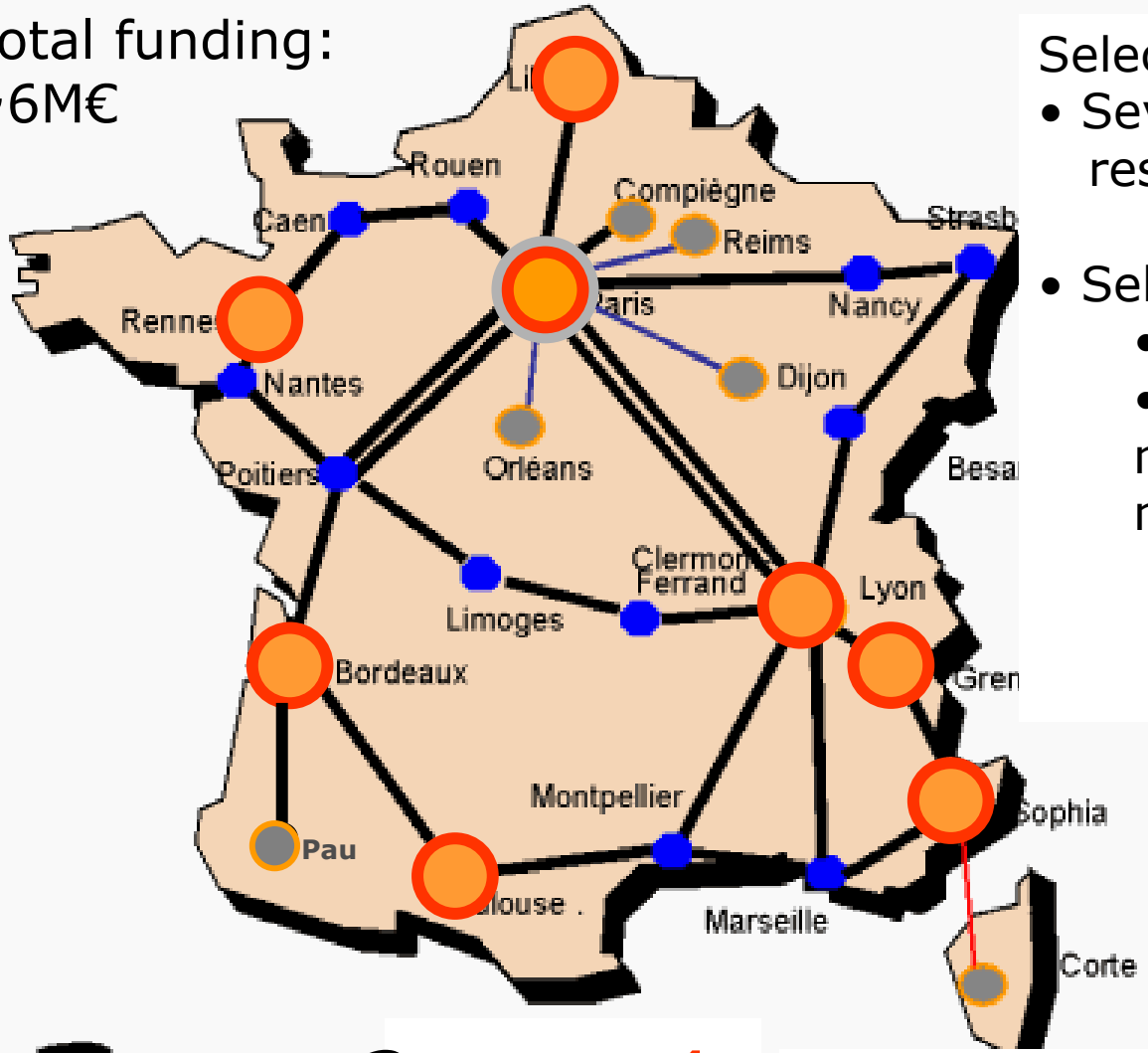
-**Franck Cappello** (animateur)
-**Thierry Priol** (Directeur directeur ACI Grid)
-**Brigitte Plateau** (Directrice CS ACI Grid)
-**Dany Vandrome** (Renater)
-Frédéric Desprez (Lyon)
-Michel Daydé (Toulouse)
-Yvon Jégou (Rennes)
-Stéphane Lantéri (Sophia)
-Raymond Namyst (Bordeaux)
-Pascale Primet (Lyon)
-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
Claude Inglebert
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 Sites

Total funding:
~6M€



Selection by the ACI Grid SC:

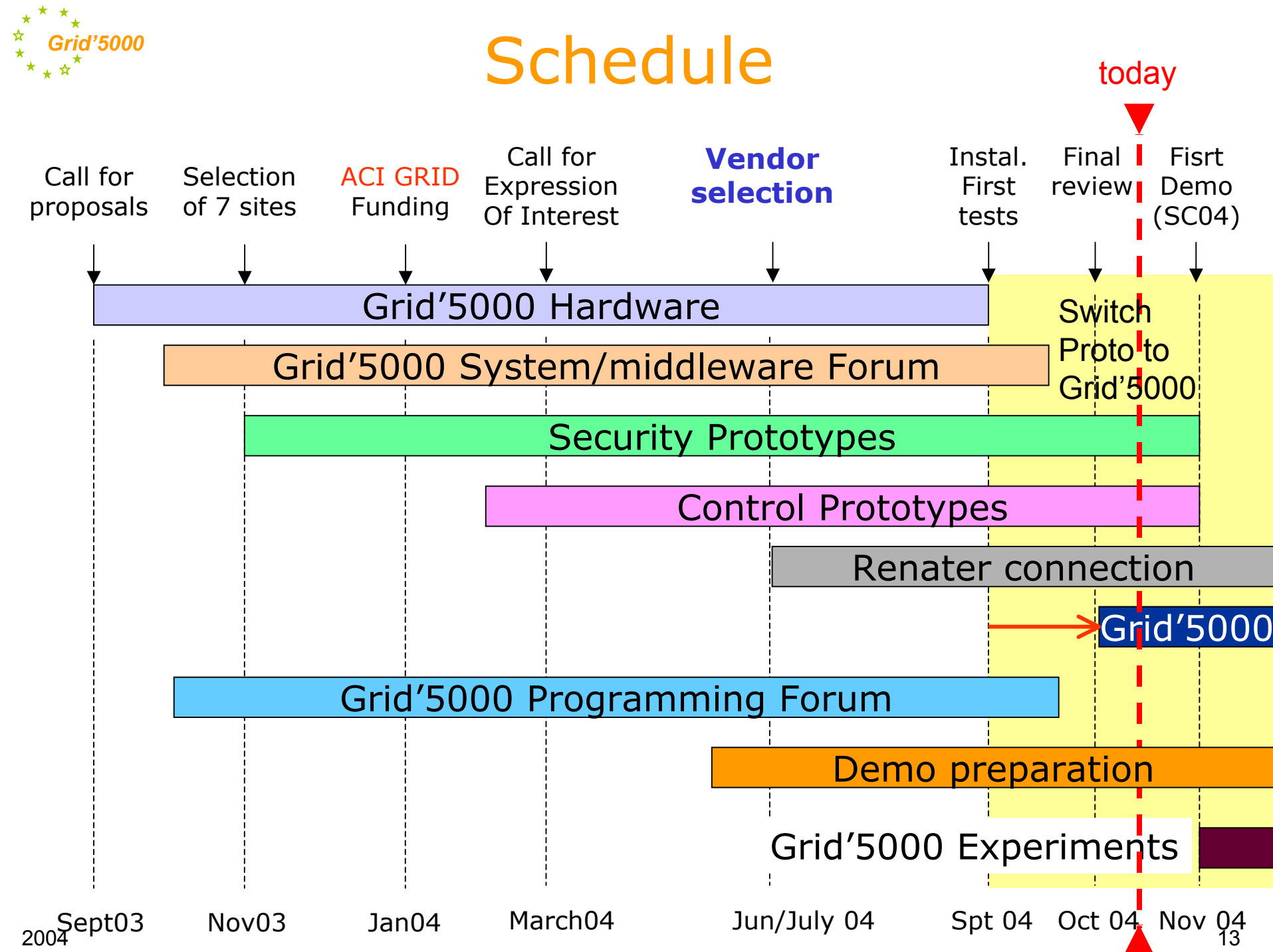
- Several sites submitted research proposals
- Selection considered:
  - Scientific aspects
  - AND capacity to manage a Grid'5000 nodes

Data Grid eXplorer

Grid 5000 nodes

Renater 3 (soon 4) 1 Gb/sec ; may evolve to 10 Gb/s

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
Switch Proto to Grid'5000
Grid'5000 System/middleware Forum
Security Prototypes
Control Prototypes
Renater connection
Grid'5000
Grid'5000 Programming Forum
Demo preparation
Grid'5000 Experiments
Sept03
Nov03
Jan04
March04
Jun/July 04
Spt 04
Oct 04
Nov 04

Agenda

Rational

Grid'5000 project

Grid'5000 design

Technical developments

Conclusion

Grid'5000 foundations: Collection of experiments to be done

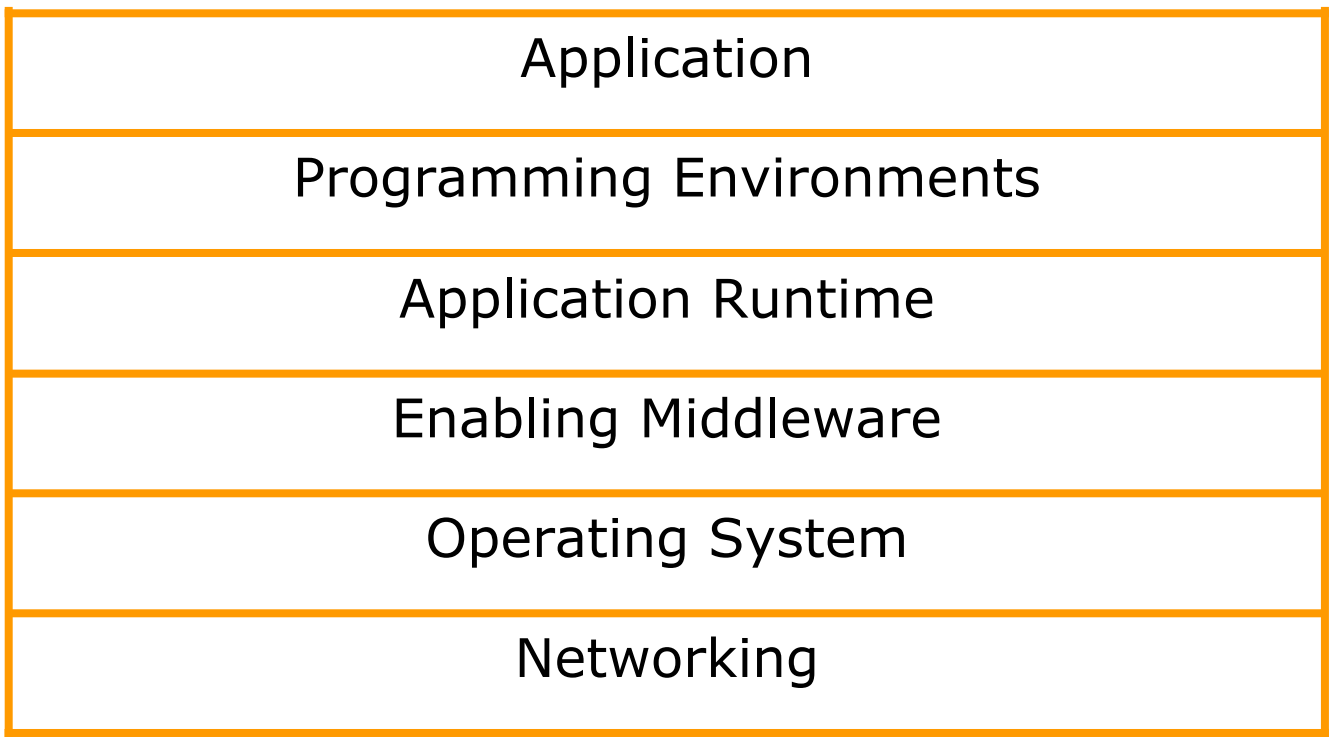
- Networking
  - End host communication layer (interference with local communications)
  - High performance long distance protocols (improved TCP)
  - High Speed Network Emulation
- Middleware / OS
  - Scheduling / data distribution in Grid
  - Fault tolerance in Grid
  - Resource management
  - Grid 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 Design goal:
Experimenting all layers of the Grid software stack



Grid'5000 Vision

Grid'5000 is NOT a production Grid!

Grid'5000 should be:

- an instrument
  to experiment and observe phenomena
  in 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 (deep control)

Grid'5000 as an Instrument

Technical issues:

- Remotely controllable Grid nodes (installed in geographically distributed laboratories)
- A « Controllable » and « Monitorable » Network between the Grid nodes → (may be unrealistic in some cases)

3) A middleware infrastructure allowing users to access, reserve and share the Grid nodes

4) A user toolkit to deploy, run, monitor, control experiments and collect results

Scientific issues:

1) Monitorable experimental conditions

2) Control of a running experiment (suspend/restart)

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)

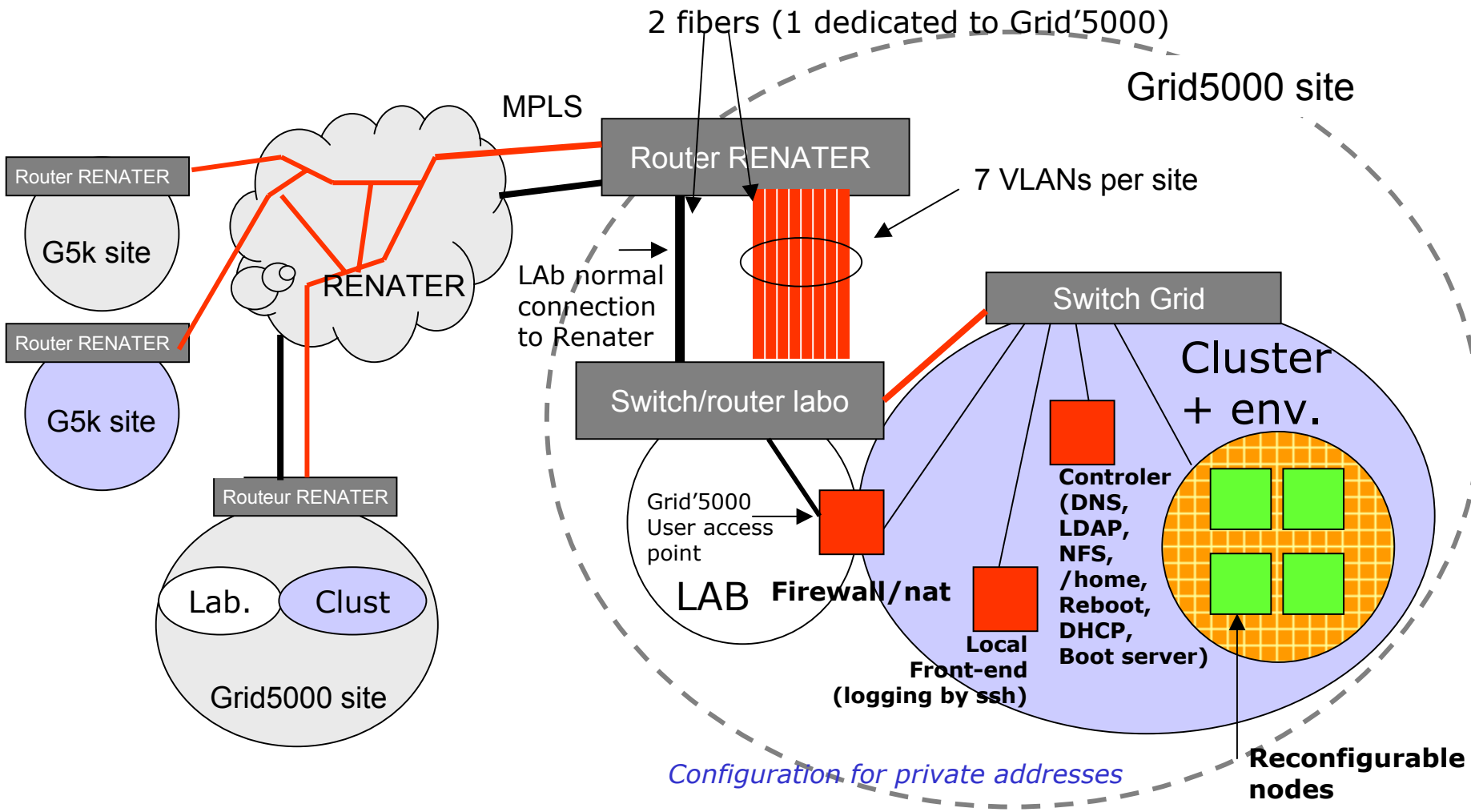
- Some sites want private addresses, some others want public addresses

- 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



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 in actions

Usage modes

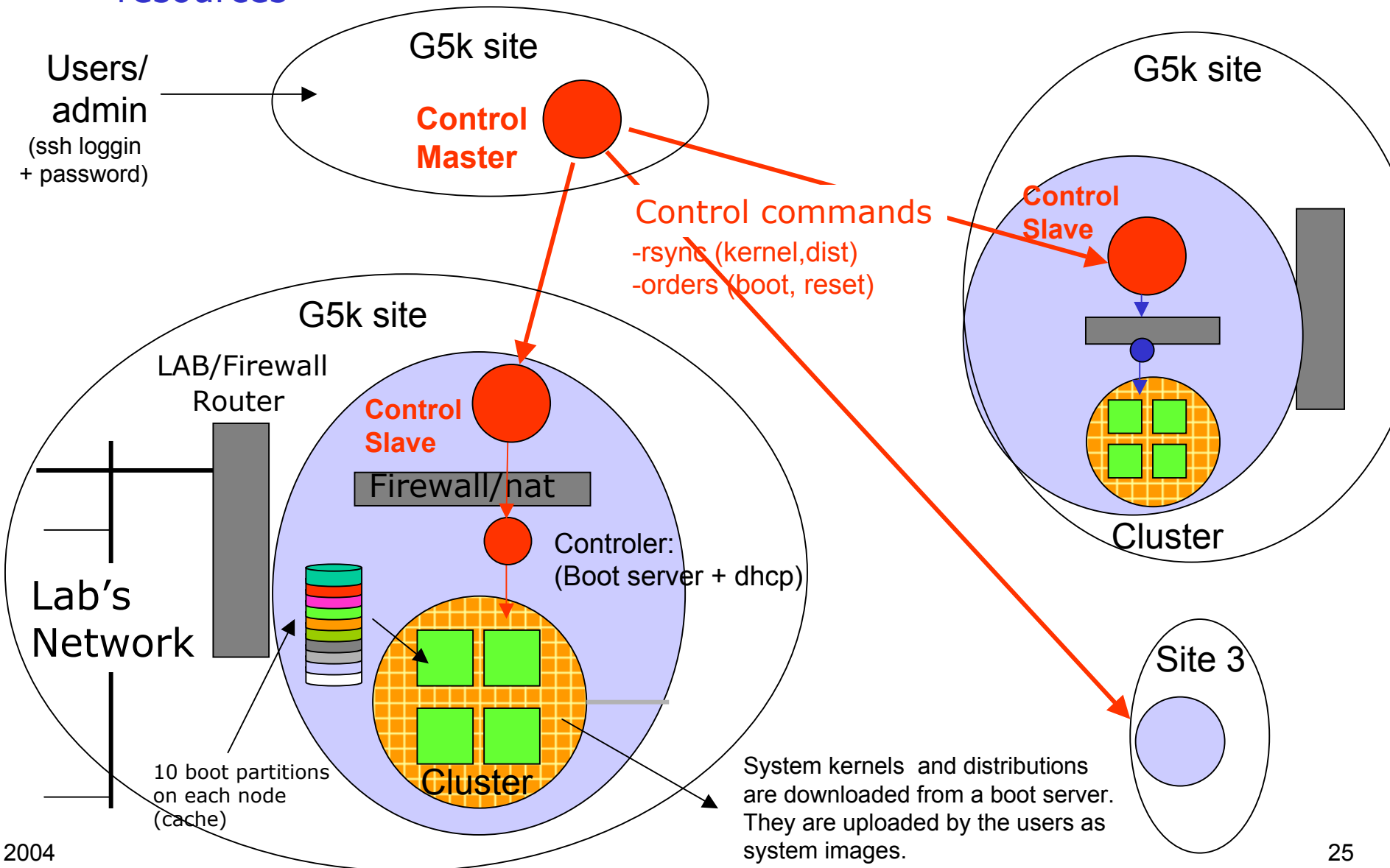
- Shared (preparing experiments, size S)
  - No dedicated resources (users log in nodes and use default settings, etc.)
- Reserved (similar to Planet lab, size M)
  - Reserved resources but un coordinated (Users may change node's OS on reserved ones)
- Batch (automatic, size L ou XL)
  - Reserved and coordinated resources experiments run under batch/automatic mode)

- All these modes with calendar scheduling

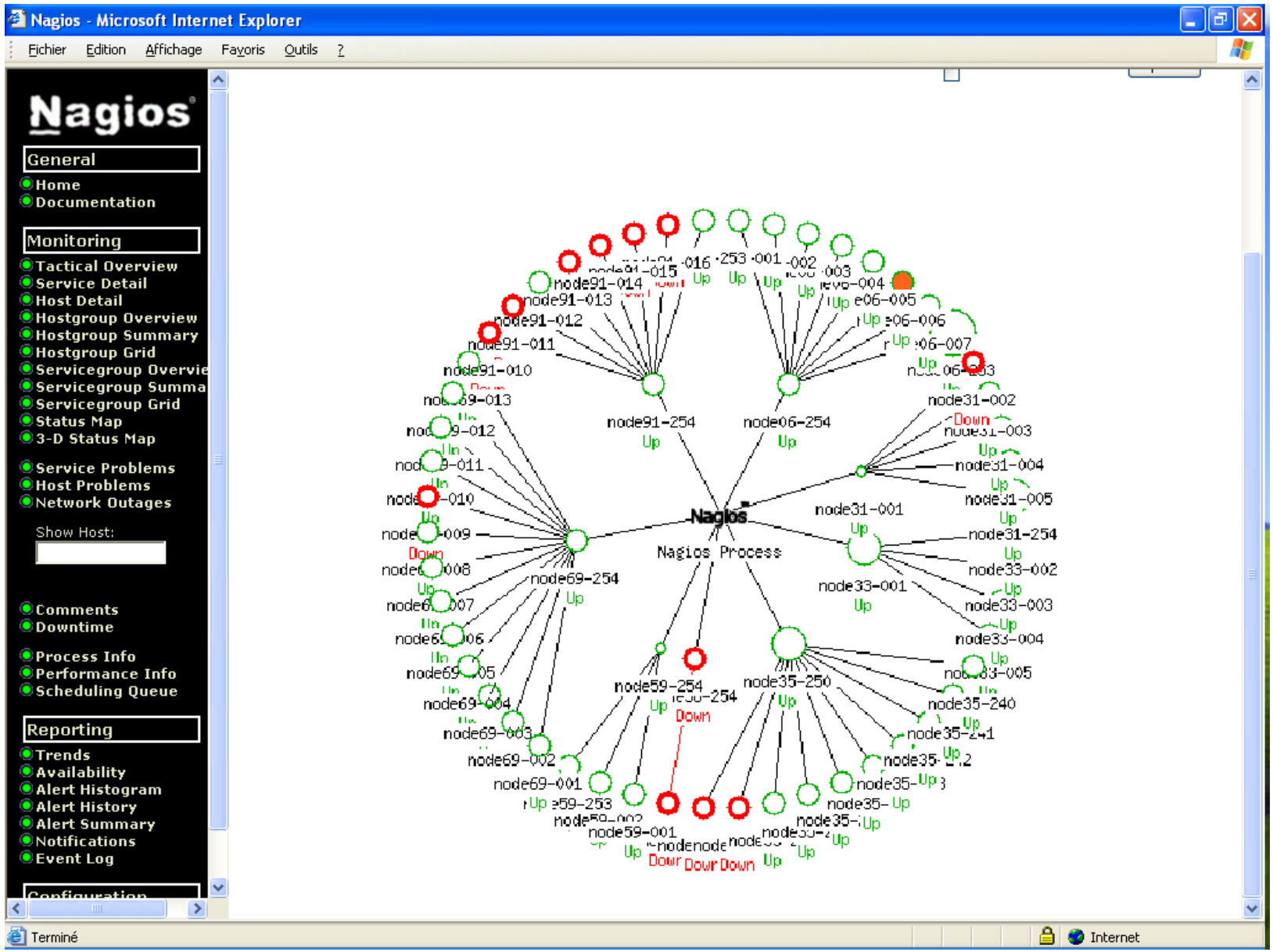
+ compliance with local usages (almost every cluster receives funds from different institutions and several projects)

Control Architecture

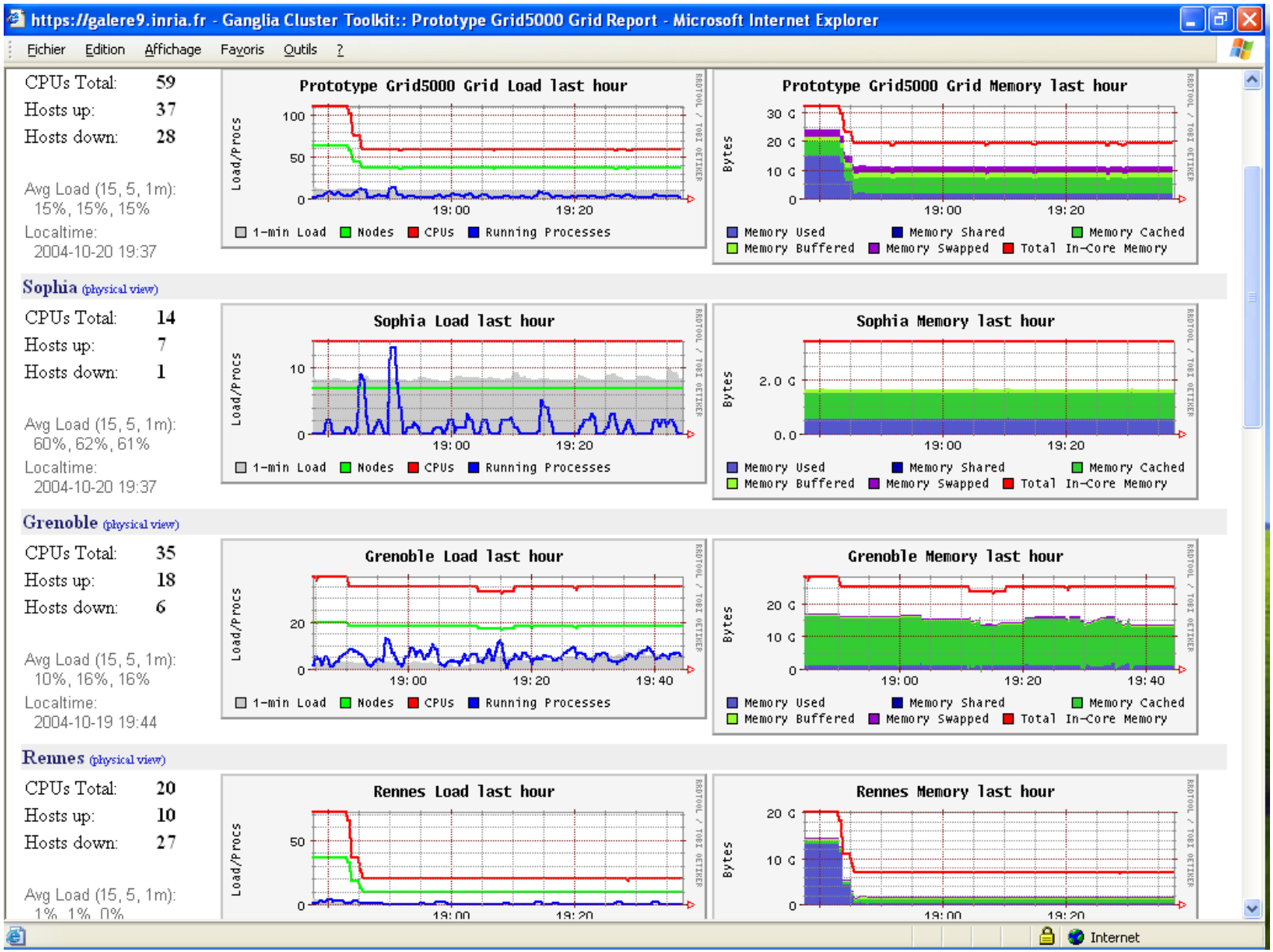
In reserved and batch modes, admins and users can control their resources



Grid'5000 prototype

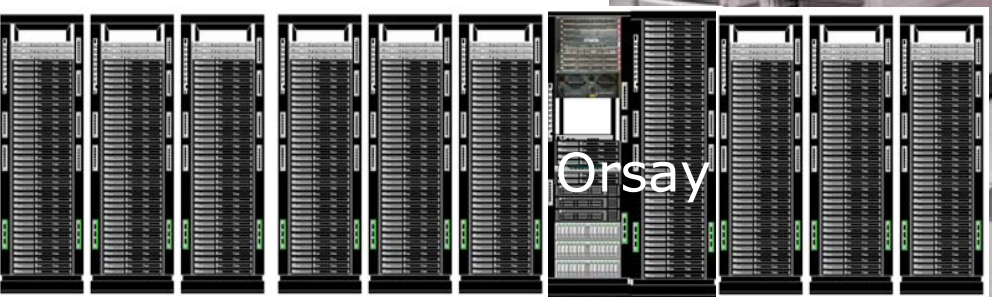


Grid'5000 prototype



Installing the real stuff

Installing the real stuff



Agenda

Rational

Grid'5000 project

Grid'5000 design

Grid'5000 developments

Conclusion

Summary

- 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 uses to reconfigure the platform [protocols/OS/Middleware/Runtime/Application]
- Grid'5000 will be opened to Grid researchers in early 2005
- Could be opened to other researchers (ACI « Data Masse », CoreGrid European project members, etc.)
- Beyond an Instrument Grid'5000 has federated a strong community: this is a human adventure!

Q&A