Grid'5000

Olivier RICHARD
Associate Professor
Grid'5000 Technical and Steering Committee
Université Joseph Fourier Grenoble
<Olivier.Richard@imag.fr>

Action Concertée Incitative
[ACI]
Globalisation de Ressources
Informatiques et des Données
[GRID]

Outline

- Motivations
- Status
- Technical aspects
- Experiments
- Next step
- Conclusion

Motivations

Genesis: 2001, ACI Grid

- Peer-to-Peer
  - CGP2P (F. Cappello, LRI/CNRS)
- Application Service Provider
  - ASP (F. Desprez, ENS Lyon/INRIA)
- Algorithms
  - TAG (S. Genaud, LSIIT)
  - ANCG (N. Emad, PRISM)
  - DOC-G (V-D. Cung, UVSQ)
- Compiler techniques
  - Métacompil (G-A. Silbert, ENMP)
- Networks and communication
  - RESAM (C. Pham, ENS Lyon)
  - ALTA (C. Pérez, IRISA/INRIA)
- Visualisation
  - EPSN (O. Coulaud, INRIA)
- Data management
  - PADOUE (A. Doucet, LIP6)
  - MEDIAGRID (C. Collet, IMAG)
- Tools
  - DARTS (S. Frénot, INSA-Lyon)
  - Grid-TLSE (M. Dayde, ENSEEIHT)
- Code coupling
  - RMI (C. Pérez, IRISA)
  - CONCERTO (Y. Maheo, VALORIA)
  - CARAML (G. Hains, LIFO)
- Applications
  - COUMEHY (C. Messager, LTHE) - Climate
  - GenoGrid (D. Lavenier, IRISA) - Bioinformatics
  - GeoGrid (J-C. Paul, LORIA) - Oil reservoir
  - IDHA (F. Genova, CDAS) - Astronomy
  - Guirlande-fr (L. Romary, LORIA) - Language
  - GriPPS (C. Blanchet, IBCP) -Bioinformatics
  - HydroGrid (M. Kern, INRIA) - Environment
  - Medigrid (J. Montagnat, INSA-Lyon) - Medical
- Grid Testbeds
  - CiGri-CIMENT (L. Desbat, UjF)
  - Mecagrid (H. Guillard, INRIA)
  - GLOP (V. Breton, IN2P3)
  - GRID5000 (F. Cappello, INRIA)
- Support for disseminations
  - ARGE (A. Schaff, LORIA)
  - GRID2 (J-L. Pazat, IRISA/INSA)
  - DataGRAAL (Y. Denneulin, IMAG)

Existing research tools

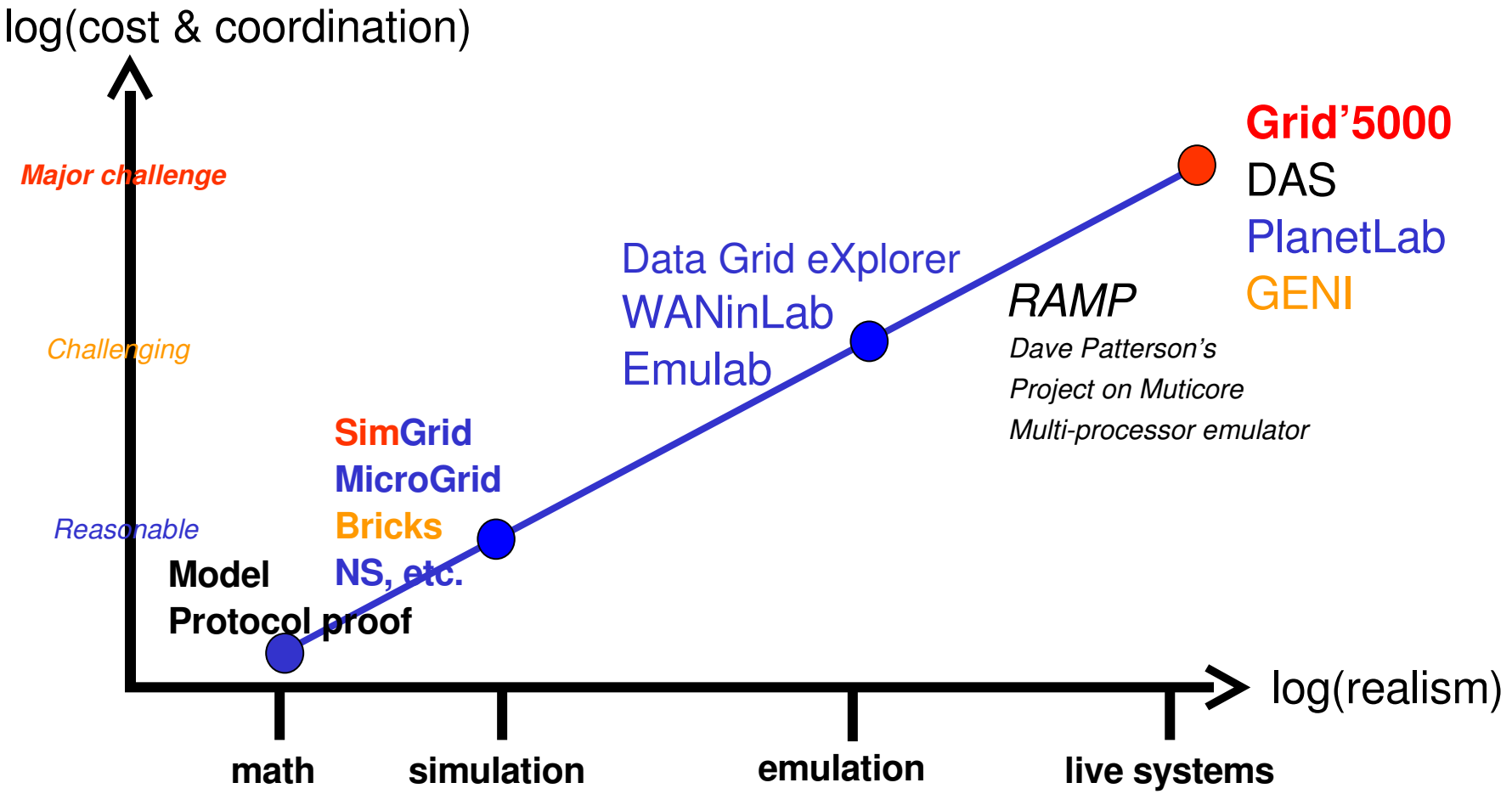
- SimGRid and SimGrid2
  - Discrete event simulation with trace injection
  - Originally dedicated to scheduling studies
  - Single user, multiple servers
- GridSim
  - Dedicated to scheduling (with deadline), DES (Java)
  - Multi-clients, Multi-brokers, Multi-servers
- Titech Bricks
  - Discrete event simulation for scheduling and replication studies
- GangSim
  - Scheduling inside and between VOs
- MicroGrid,
  - Emulator, Dedicated to Globus, Virtualizes resources and time, Network (MaSSf)

France
USA
Australia
Japan

→Nowhere to test networking/OS/middleware ideas, to measure real application performance, and Simulation and Emulation are quite slow.

Need a real life testbed

In 2003, the Grid'5000 project is launched



The Grid'5000 project

1) Building a nation wide experimental platform for Large scale Grid & P2P experiments
   - 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
   - Port and test applications, develop new algorithms
   - Address critical issues of Grid system/middleware:
     - Programming, Scalability, Fault Tolerance, Scheduling
   - Address critical issues of Grid Networking
     - High performance transport protocols, QoS
   - Investigate original mechanisms
     - P2P resources discovery, Desktop Grids

Domains of experiment

- 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
- Programming
  - Component programming for the Grid (Java, Corba)
  - GRID-RPC
  - GRID-MPI
  - Code Coupling
- Middleware / OS
  - Resource management / Scheduling / data distribution in Grid
  - Fault tolerance in Grid
  - Grid SSI OS and Grid I/O
  - Desktop Grid/P2P systems
- Networking
  - End host communication layer (interference with local communications)
  - High performance long distance protocols (improved TCP)
  - High Speed Network Emulation

Allow experiments at any level of the software stack

Testbed for experiments

Quantitative metrics :

- Performance: Execution time, throughput, overhead, QoS (Batch, interactive, soft real time, real time).
- Scalability:Resource occupation (CPU, memory, disc, network), Applications algorithms, Number of users, Number of resources.
- 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.

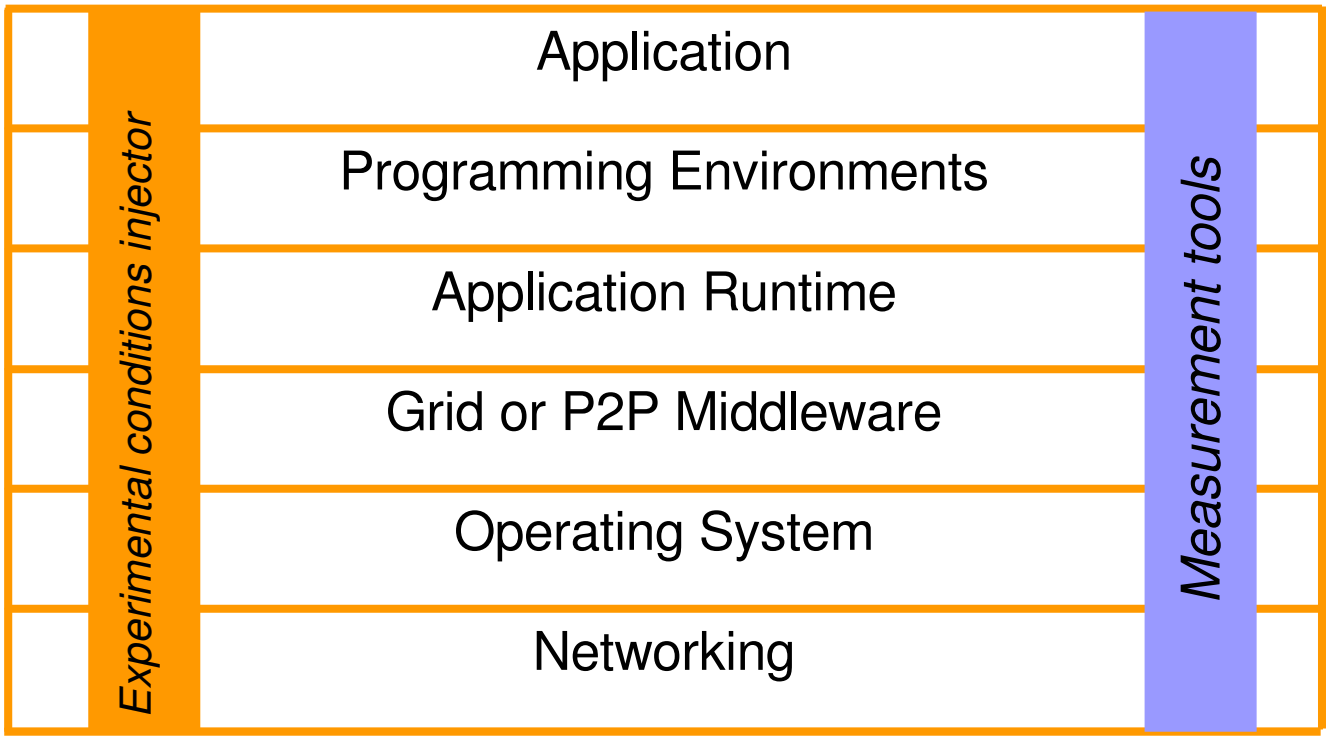
Experimental Condition injection :

- Background workloads: CPU, Memory, Disk, network, Traffic injection at the network edges.
- Stress: high number of clients, servers, tasks, data transfers,
- Perturbation: artificial faults (crash, intermittent failure, memory corruptions, Byzantine), rapid platform reduction/increase, slowdowns, etc.

Allow users to run their favorite measurement tools
and experimental condition injectors

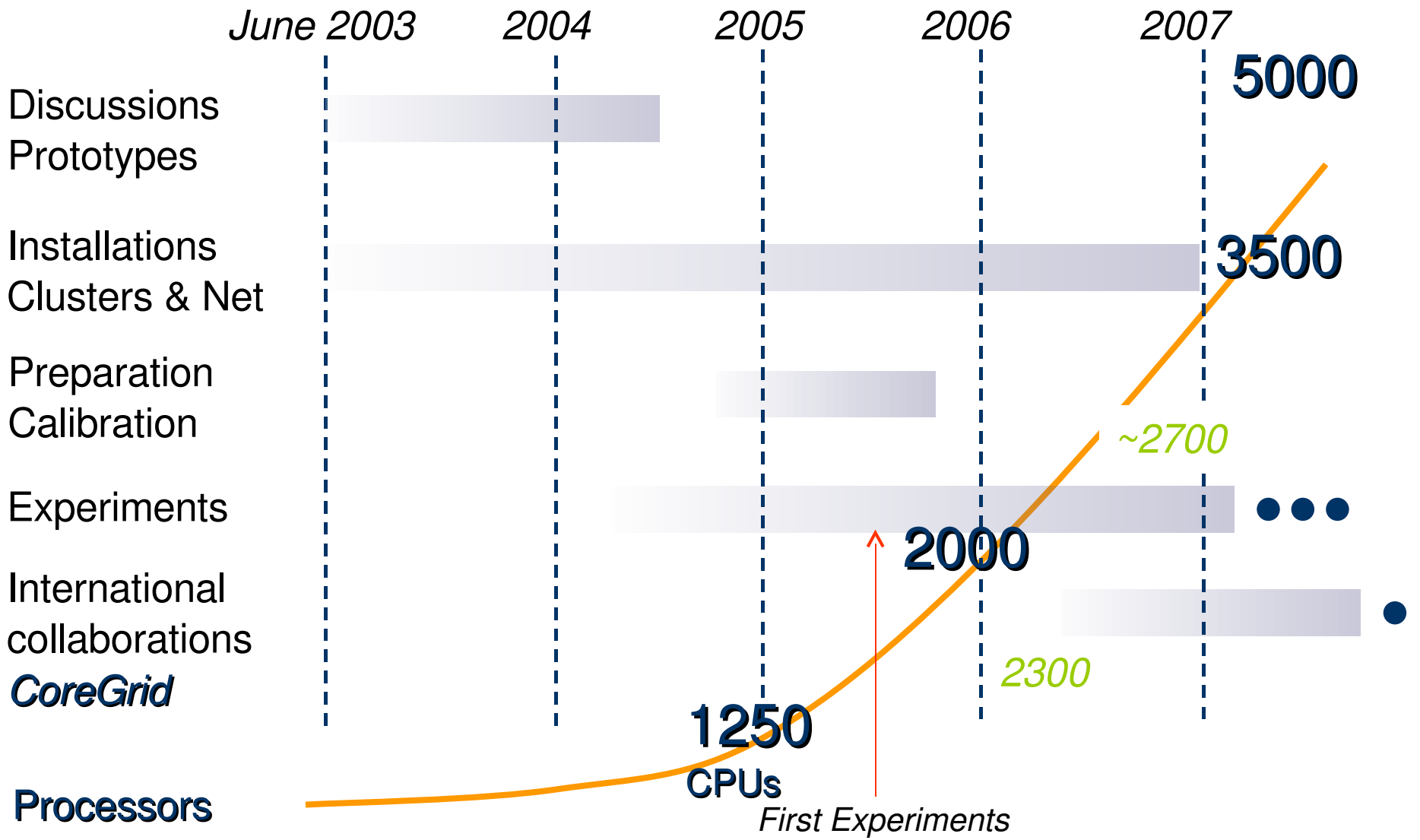
Grid'5000 principle

A highly reconfigurable experimental platform



Let users run ***(deploy)*** their complete customized software stack, including software providing measurement tools + experimental conditions injectors

Timeline


June 2003
2004
2005
2006
2007
Discussions
Prototypes
Installations
Clusters & Net
Preparation
Calibration
Experiments
International
collaborations
CoreGrid
Processors
5000
3500
~2700
2000
2300
1250
CPUs
First Experiments

Technical aspects

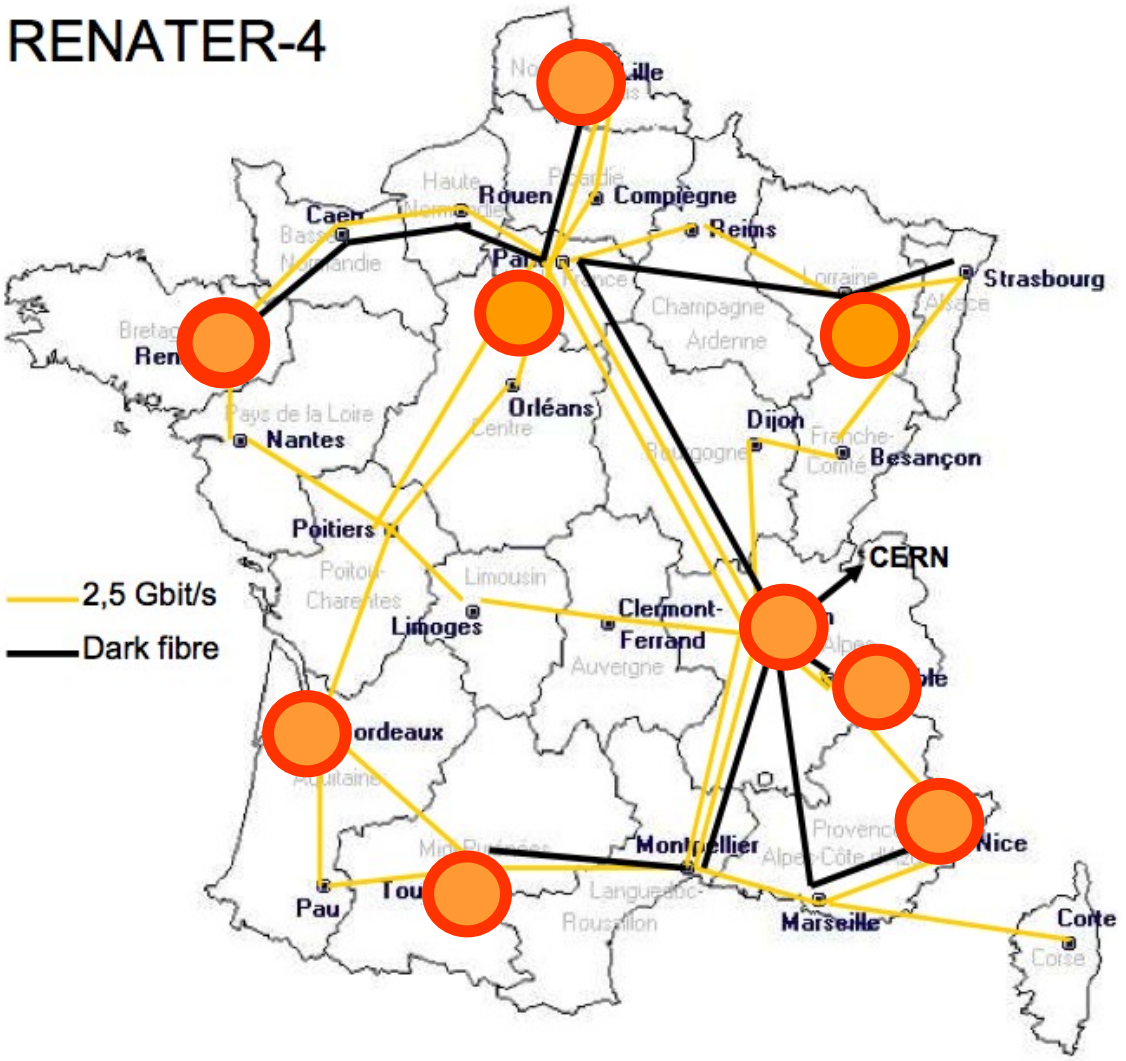
Grid'5000 components

- A nationwide backbone network
- A Security Architecture
- A uniform account management
- Several Grid middleware
  - OAR
  - Kadeploy
- A dedicated staff

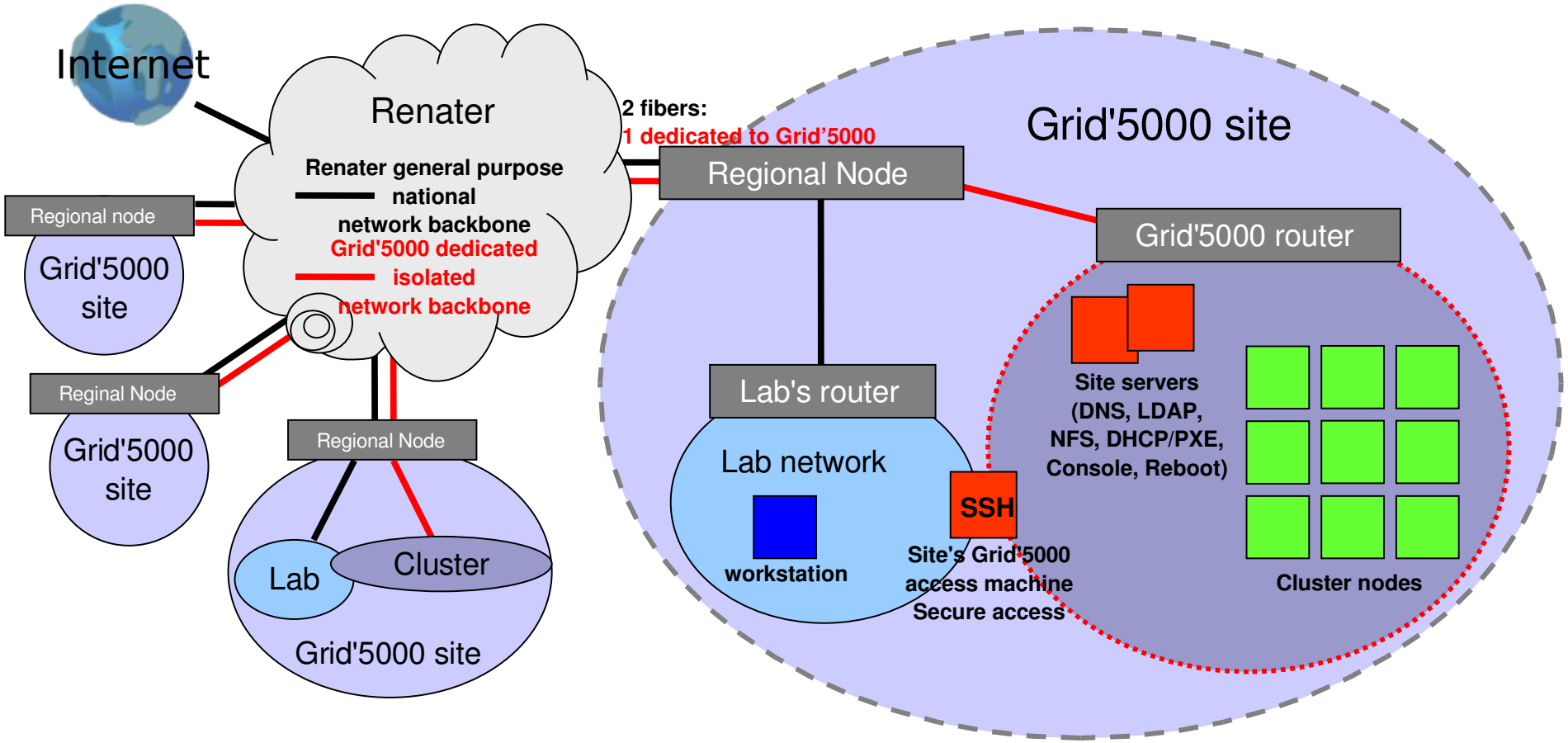
Backbone network

Nationwide backbone network through Renater-4

Renater-4 Dark Fiber Infrastructure provide Grid'5000 with a 10Gb/s dedicated and isolated intersite network



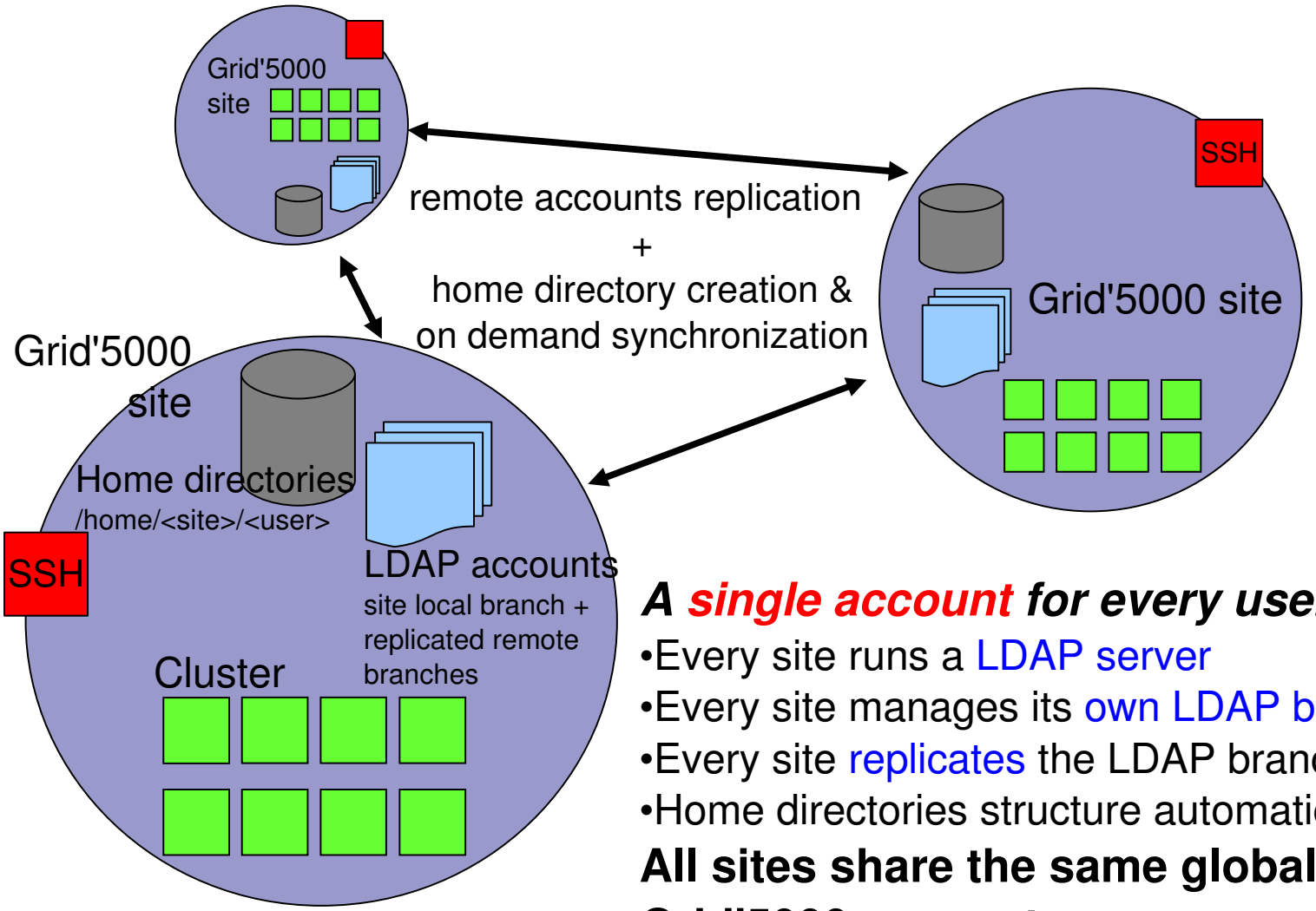
Security architecture



Protecting Grid'5000 AND Protecting Internet:

- Grid'5000 network is confined
- Access thru Secure Shell + restricted outbount traffic

Account management



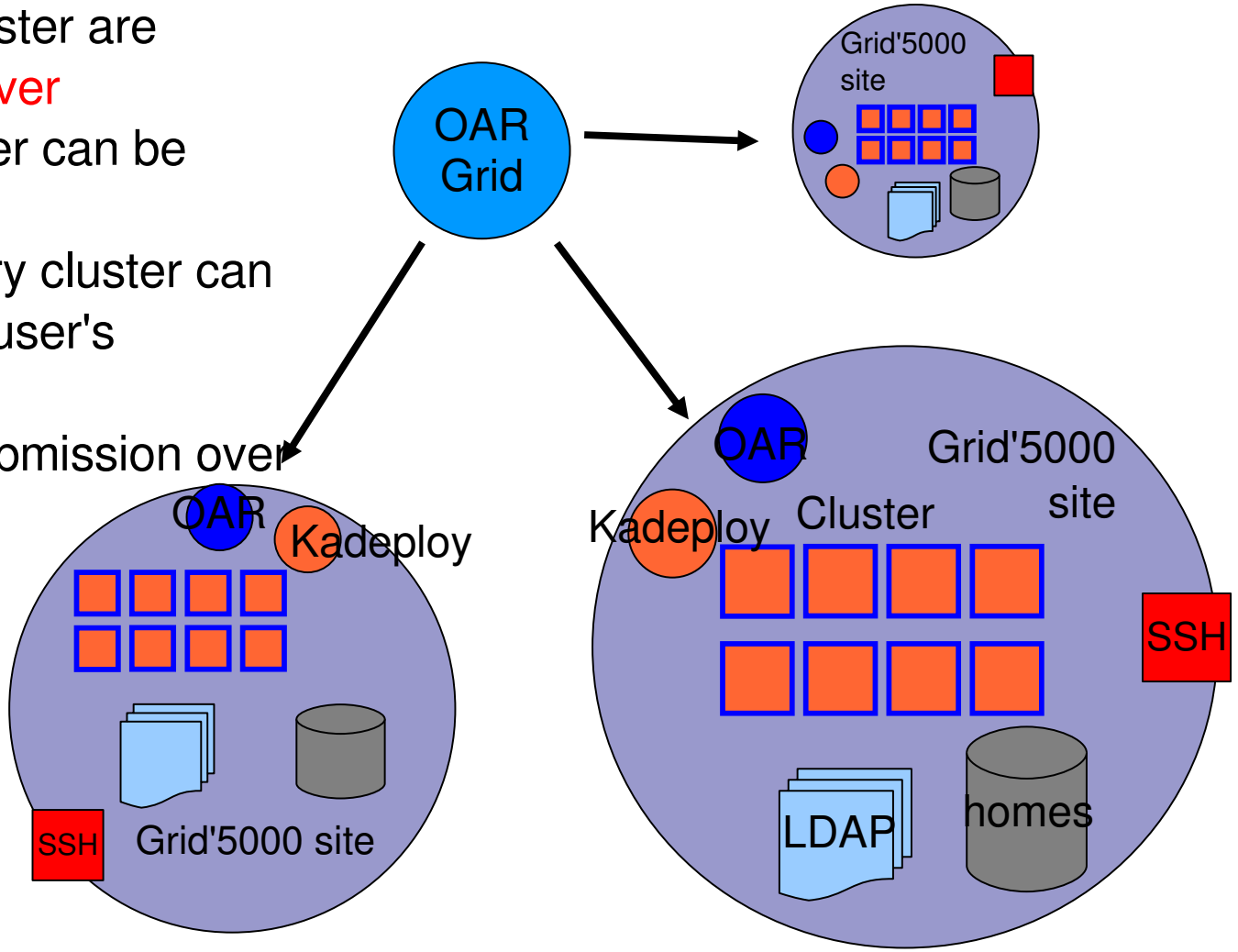
A single account for every user for every sites:

- Every site runs a LDAP server
- Every site manages its own LDAP branch of accounts
- Every site replicates the LDAP branches of the other sites
- Home directories structure automatic management

All sites share the same global directory of Grid'5000 accounts

Grid tools

- Resources on every cluster are managed by a OAR server
- Systems on every cluster can be managed by Kadeploy
- On user's demand, every cluster can be redeployed with the user's provided environment
- OAR grid allows grid submission over the whole grid



OAR Batch Scheduler

A classical but scalable, robust and flexible Batch Scheduler, with many features:

- Admission rules
- Flexible resources schema *
- Hierarchical resources *
- Multiple resources *
- Moldable job support *
- CPUSET support*
- Matching of resources
- Hold and resume jobs
- Multischedulers support
- Multiqueues with priority
- Besteffort queues (for exploiting idle resources)

Enhanced submission expression *

Checkpointing support *

ssh as remote execution protocols (Taktuk for large cluster)

Dynamic insertion/deletion of compute node

FirstFit Scheduler with matching resource

Advance Reservation

No specific daemon on compute nodes

Environment of Demand support (KaDeploy integration)

Check compute nodes before launching

Activity visualization tools (GanttChart)

* in version 2.0

OAR Grid

Main scheduling objective for G5K : co-allocation for simultaneous jobs start.

A very simple grid extension:

- submit an advance reservation on every selected clusters

```
for c in #clusters
        ssh c oarsub -r now
end_for
```

- if a reservation is rejected: 2 modes
  - default: stop submissions and delete accepted reservation(s)
  - forced: continue to submit other reservations
- a database keeps all information (job_grid id, list of job id for each cluster submission)

Kadeploy

The tool to providing the ***highly reconfigurable grid*** feature to Grid'5000 users:

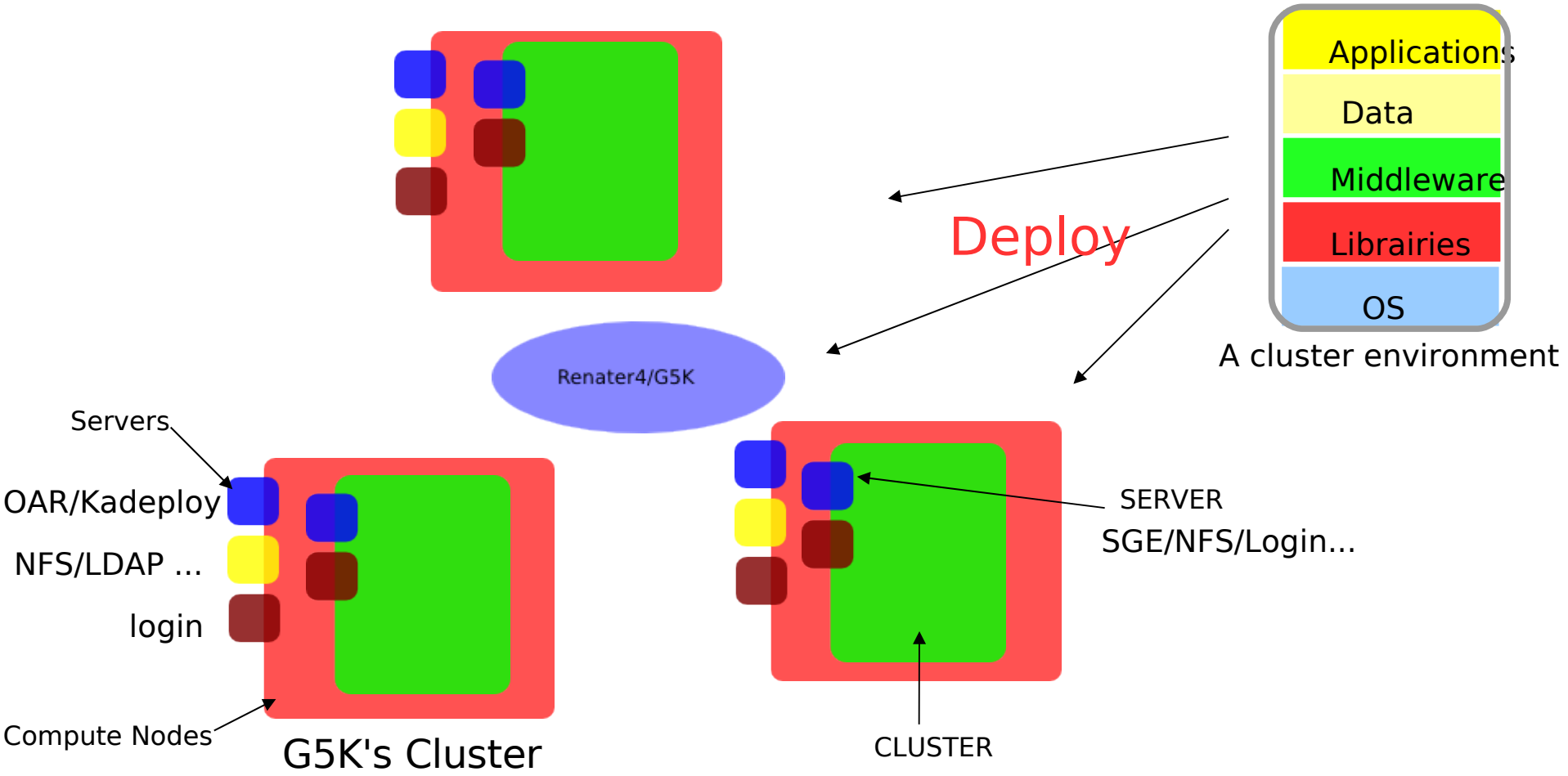
- Deep control (allow to reboot nodes, access to the console...)
- Unrestricted access (gain administrator privilege, kernel patch, modified network behavior)
- Unlimited configuration possibilities (complete customized grid environment deployment)

Rely on many low level mechanisms:

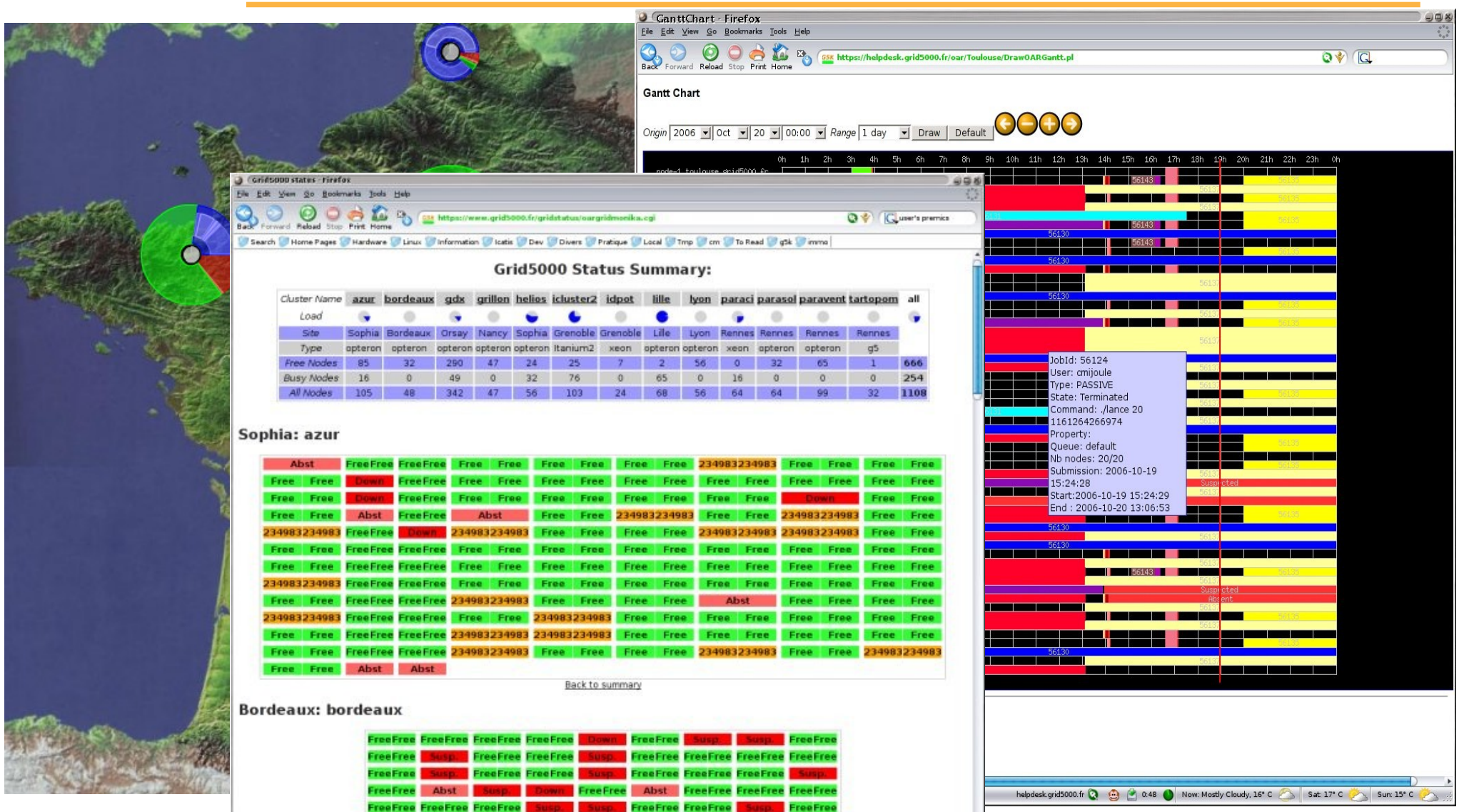
- remotely driven boot switch mechanism (PXE)
- node software/hardware remote reboot (IPMI/RSA management cards)
- remote console access
- massive system deployment (tar chain)
- many environments: any Linux distribution flavour, FreeBSD, Solaris, ...
- User rights management/OAR reservation coupling

Deployment example

- Goal : replay workload traces on real multicluster system



Grid visualization



Work in progress

Still working on improvements:

- Non regression/Qualification platform
- Improvement of kadeploy environments
- OAR 2.0 in validation phase
- Deployment thru virtualization
- Routing configuration
- Proxy infrastructure
- Network probes infrastructure

Grid'5000 Staff

Grid'5000 staff making all the components fit together:

- 1-2 engineer per site + coordination + researchers

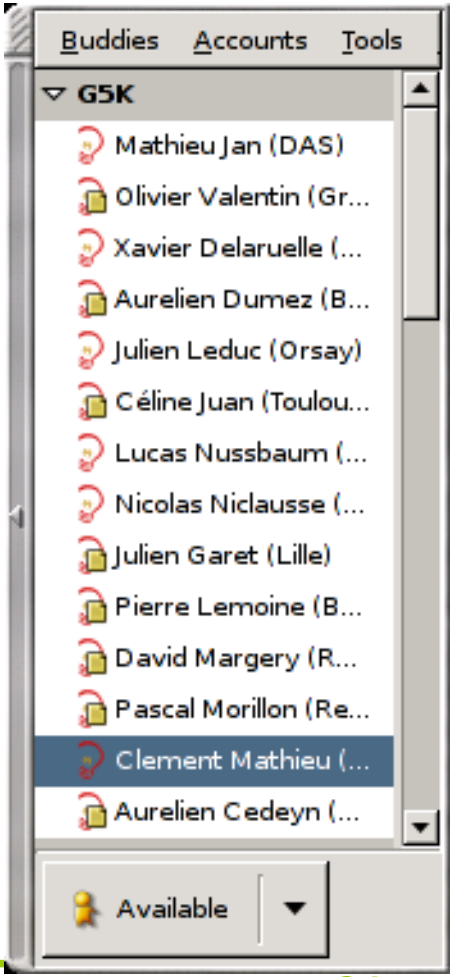
Collaboration between site is one challenge:

Communication means:

- *Mailing lists* (CP, CT, users, *site*-users, *site*-staff, network-staff)
- *Instant messaging* (Grid'5000 dedicated Jabber server)
- *Phone* + audio-conf (monthly CT audio-meeting + task dedicated meeting...)
- Physical meeting every 4 months

Tools aiding collaborative work:

- *Wiki*: public web site + user's portal + committees portal
- *Bugzilla*: ticket tracking, task assignment
- *SVN repository / GForge*
- *incident tracking tools/cross admin logs/monitoring*

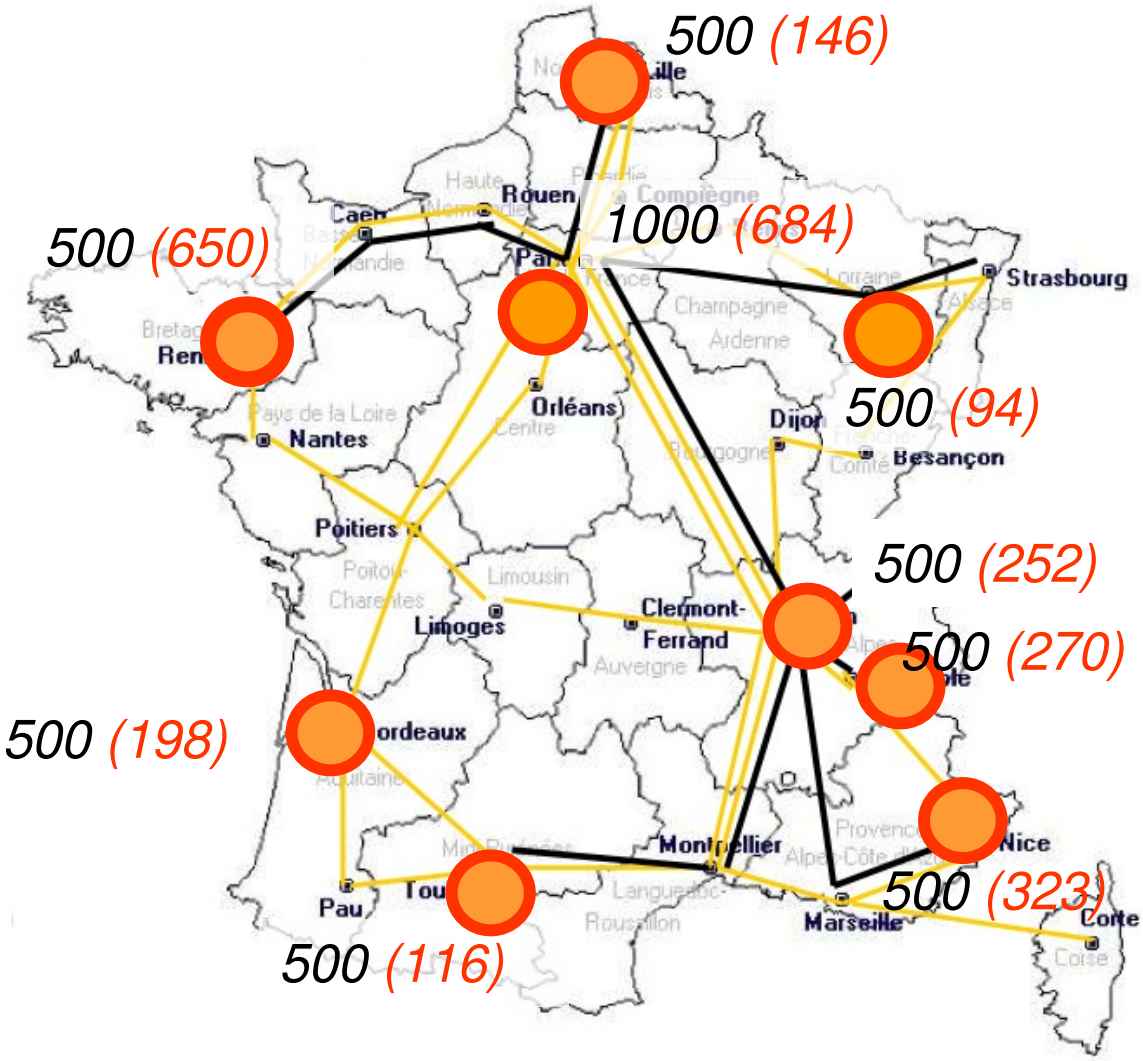


Status

Grid'5000 sites

- 9 sites
- Planned CPUs in *black*
- Current CPUs in *(red)*

2007: 3000+ CPU



Grid'5000 hardware

Node details

Sites \ Nodes	Apple Xserve G5	Dell PowerEdge 1600SC	Dell PowerEdge 1750	Dell PowerEdge 1855	Dell PowerEdge 1950	HP Integrity RX2600	HP ProLiant DL145G2	IBM eServer 325	IBM eServer 326	IBM eServer 326m	Sun Fire V20z	Sun Fire X4100	Site total
Bordeaux				51				48					**99**
Grenoble		32				103							**135**
Lille									53	20			**73**
Lyon													
Nancy													
Orsay													
Rennes	32												
Sophia													
Toulouse													
Nodes total	32												

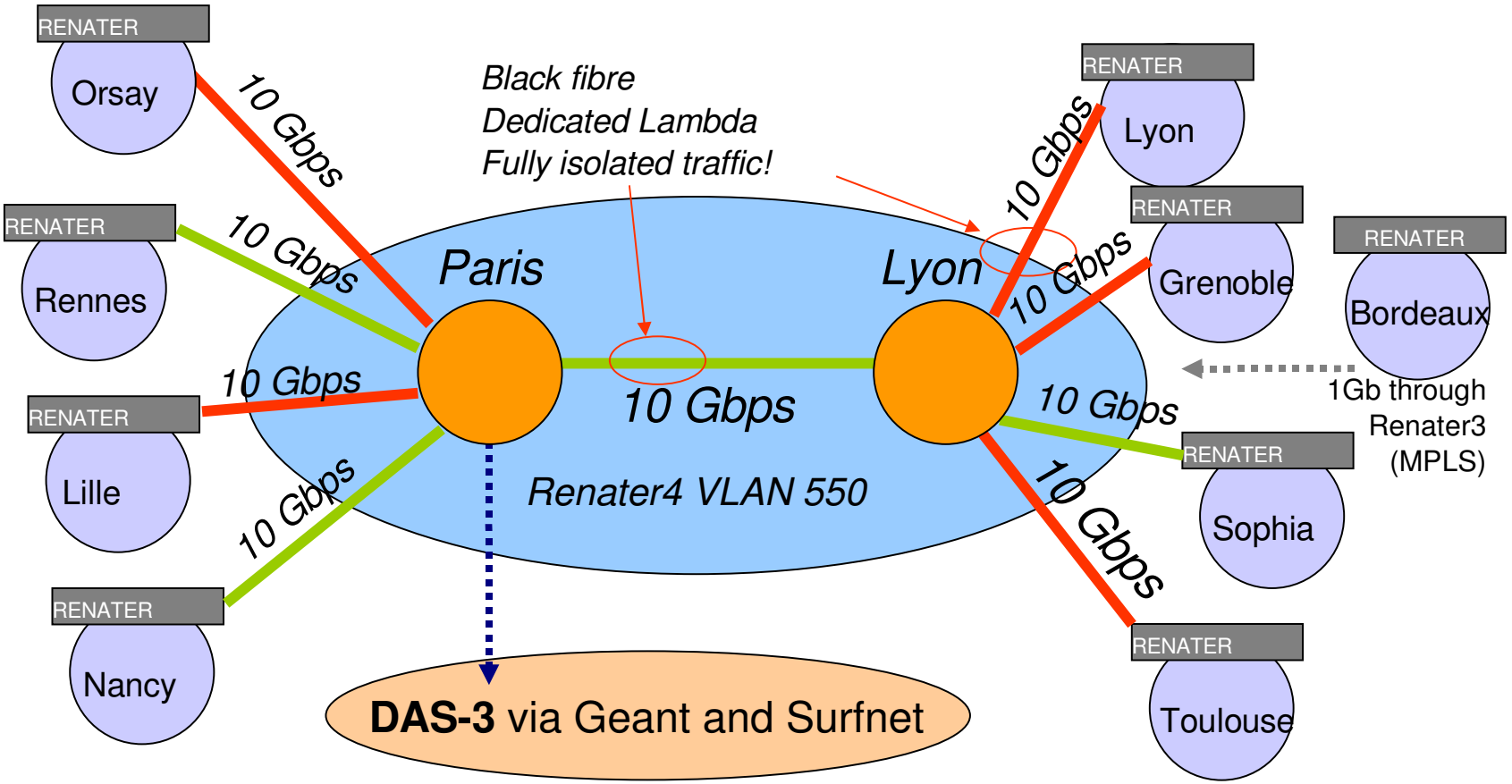
Processor details

Sites \ Processors	AMD Opteron 246	AMD Opteron 248	AMD Opteron 250	AMD Opteron 252	AMD Opteron 275	Intel Itanium 2	Intel Xeon	Intel Xeon 5148 LV	Intel Xeon IA32	PowerPC	Site total
Bordeaux		96					102				**198**
Grenoble						206			64		**270**
Lille		106		40							**146**
Lyon	112		140								**252**
Nancy	94										**94**
Orsay	432		252								**684**
Rennes	198	128						132	128	64	**650**
Sophia	210				112						**322**
Toulouse		116									**116**
Processors total	1046	446	392	40	112	206	102	132	192	64	**2732**

New clusters are currently being installed in Nancy, Toulouse, Sophia...

https://www.grid5000.fr/mediawiki/index.php/Special:HardwareAbstract

Network status

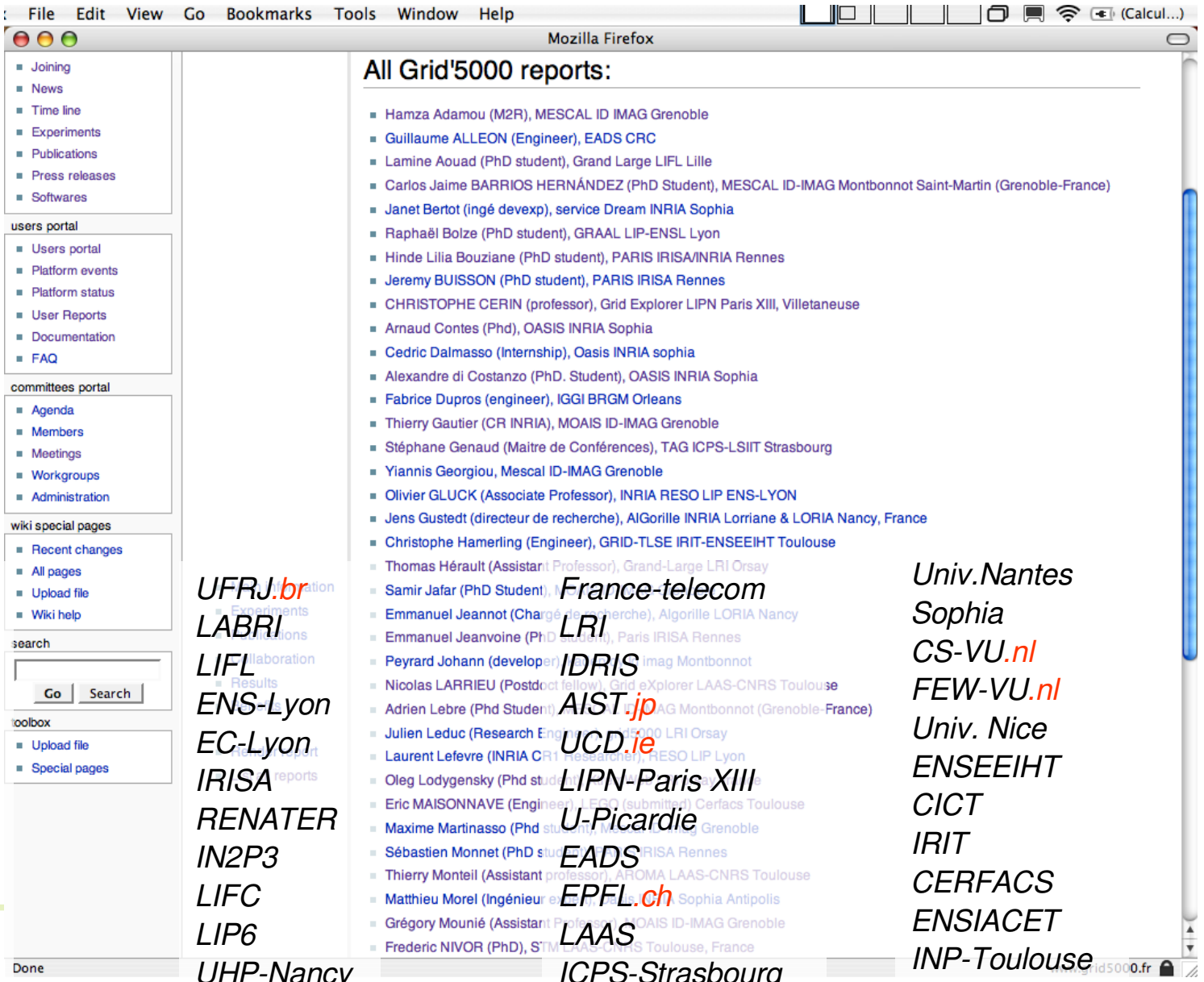


+ Myrinet 10G + Infiniband 10G available on some sites

Grid'5000 users

more than 200 active users

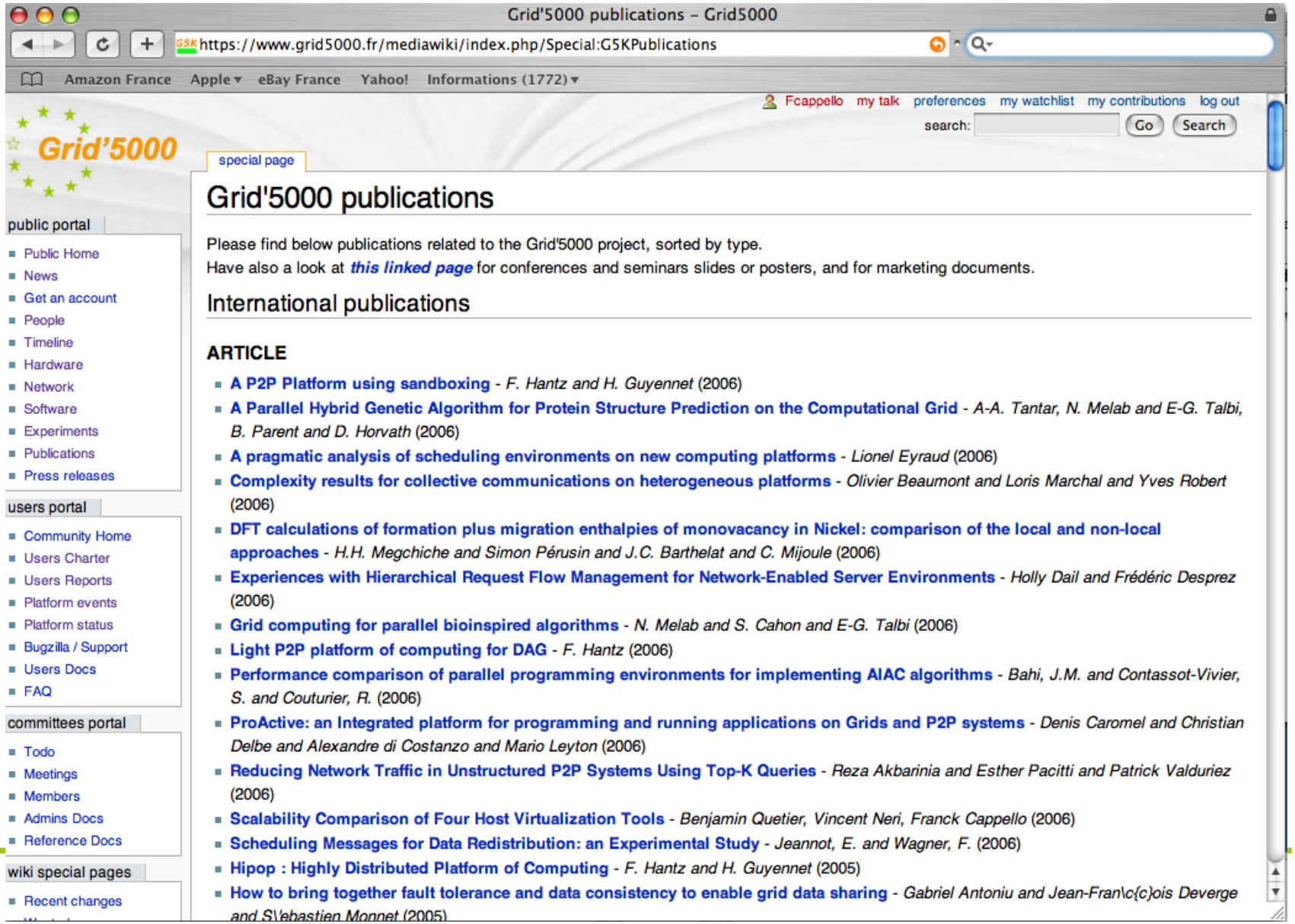
coming from more than 40 laboratories.



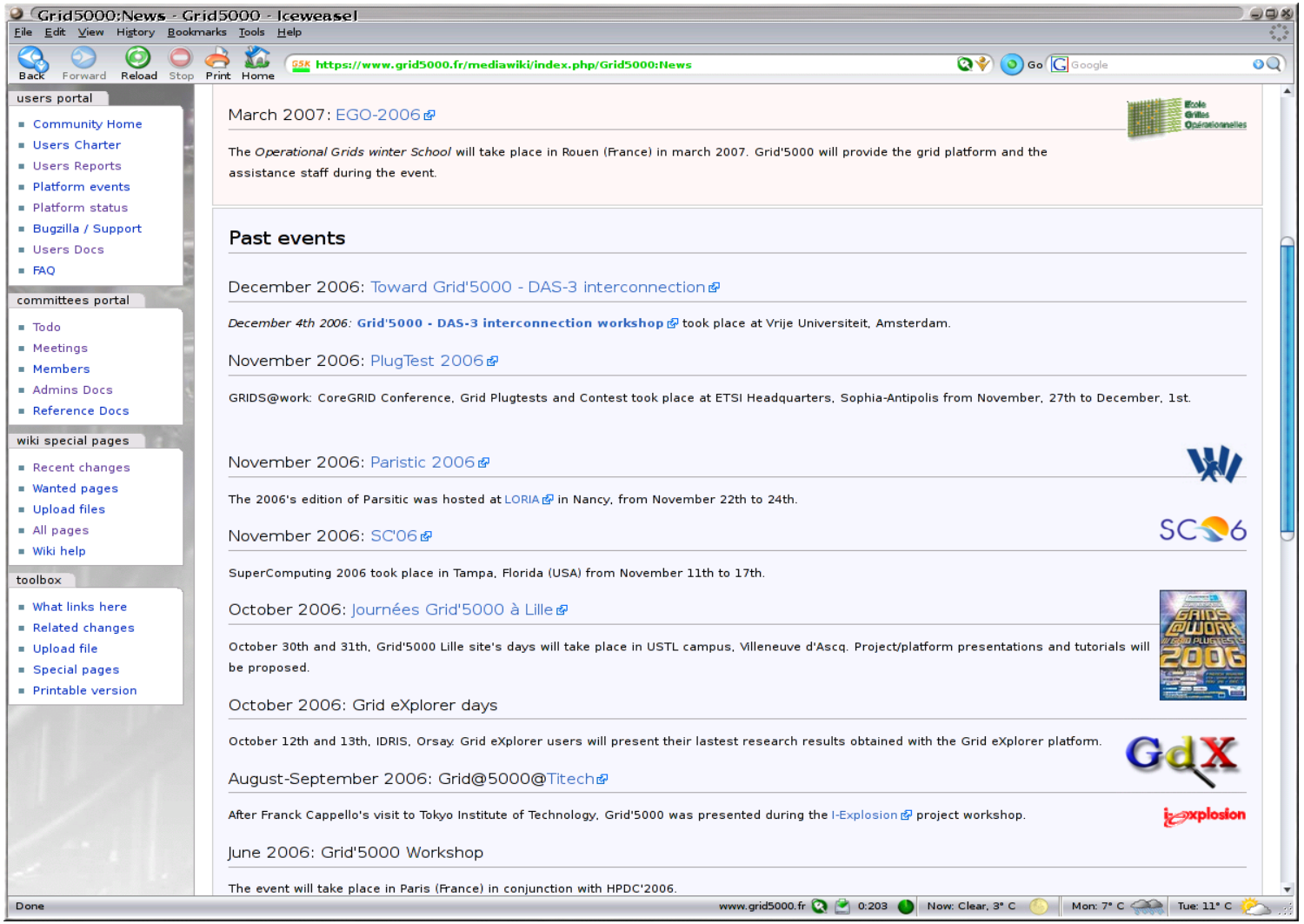
IBCP
IMAG
INRIA-Alpes
INSA-Lyon
Prism-Versailles
BRGM
INRIA
CEDRAT
IME/USP.br
INF/UFRGS.br
LORIA
UFRJ.br
LABRI
LIFL
ENS-Lyon
EC-Lyon
IRISA
RENATER
IN2P3
LIFC
LIP6
UHP-Nancy
France-telecom
LRI
IDRIS
AIST.jp
UCD.ie
LIPN-Paris XIII
U-Picardie
EADS
EPFL.ch
LAAS
ICPS-Strasbourg
Univ.Nantes
Sophia
CS-VU.nl
FEW-VU.nl
Univ. Nice
ENSEEIHT
CICT
IRIT
CERFACS
ENSIACET
INP-Toulouse
SUPELEC

250+ experiments

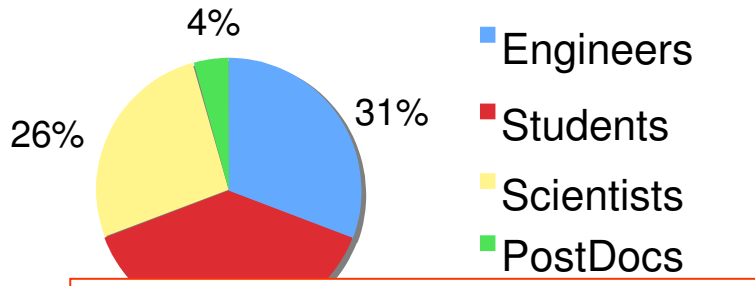
230+ publications

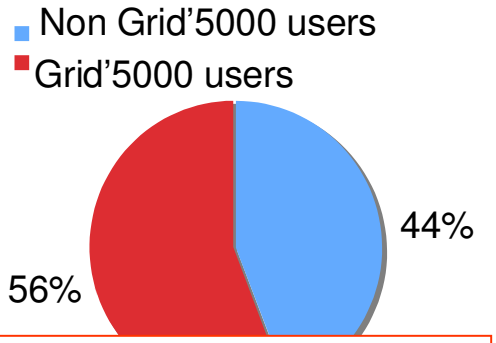


A series of events



117 participants (we tryed to limit to 100)

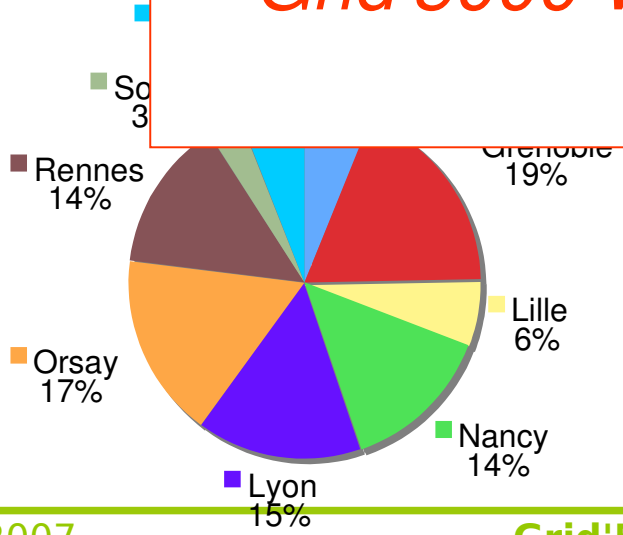


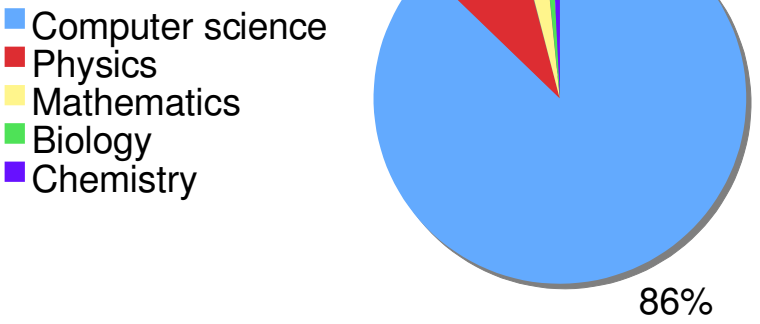


Don't miss the Second

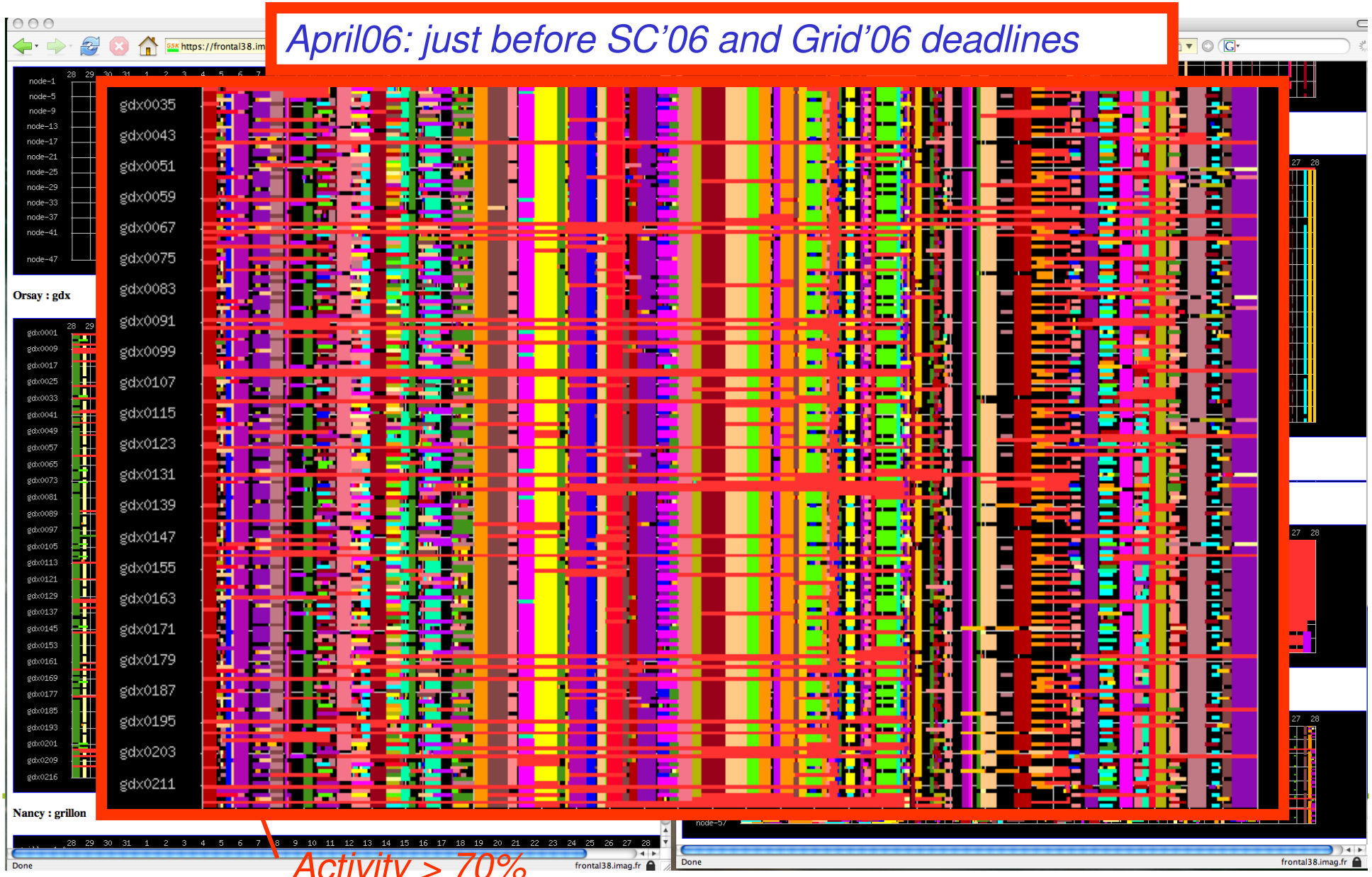
Grid'5000 Winter School in March 2007

http://ego-2006.renater.fr/





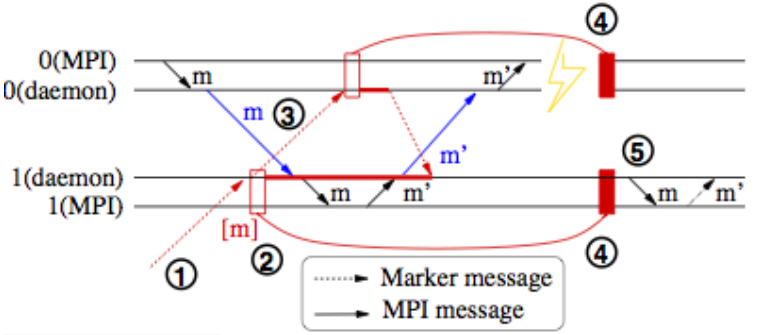
Resources usage



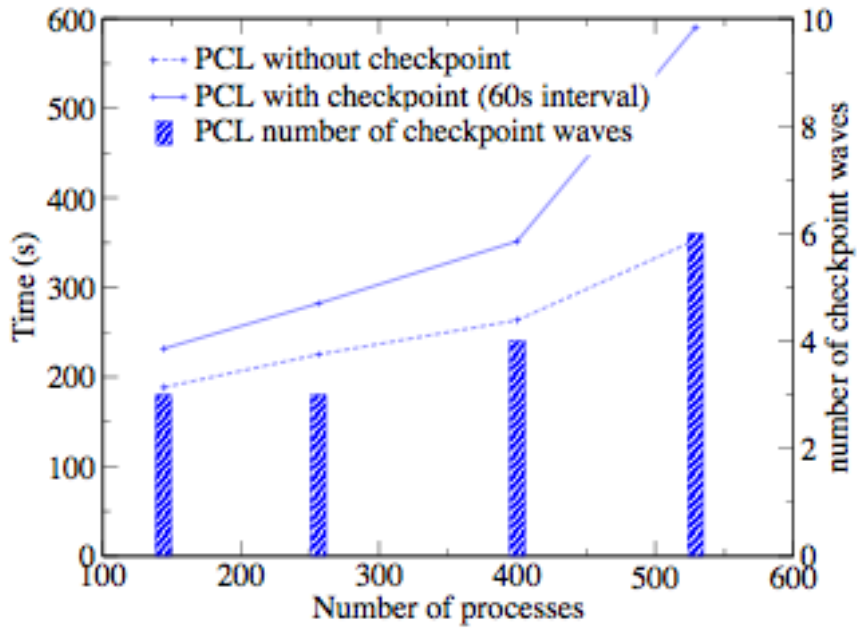
Experiments

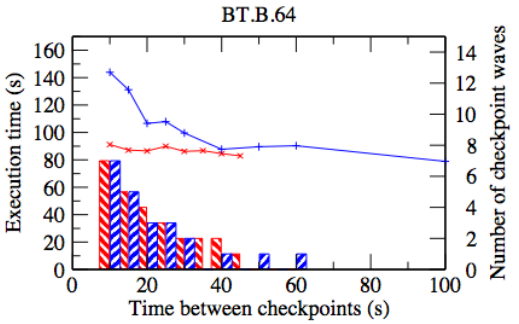
Fault tolerant MPI for the Grid

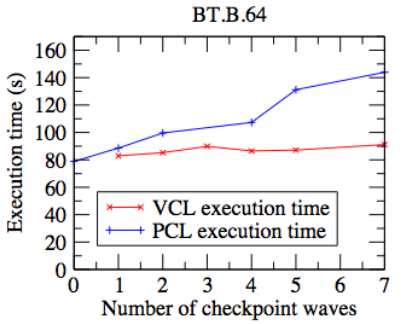
- **MPICH-V**: Fault tolerant MPI implementation
- Research context: large scale fault tolerance
- Research issue: Blocking or non Blocking Coordinated Checkpointing?
- Experiments on 6 sites up to 536 CPUs

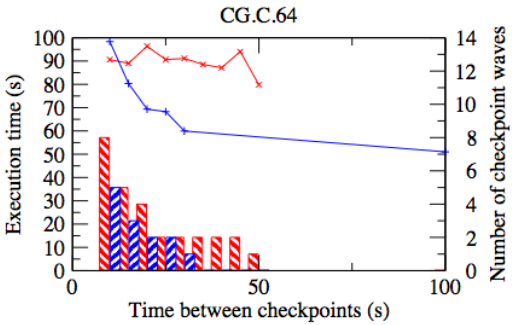


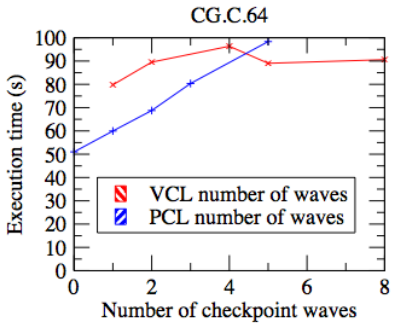
BW – Lat	Bordeaux	Orsay	Rennes	Sophia	Lille	Toulouse
Toulouse	190Mb/s – 1.5ms	59.81Mb/s – 5.51ms	34.79Mb/s – 9.92ms	83.7Mb/s – 3.74ms	26.97Mb/s – 13.04ms	930.4Mb/s – 0.04ms
Lille	30.23Mb/s – 11.62ms	132.55Mb/s – 2.25ms	56.84Mb/s – 5.83ms	37.41Mb/s – 9.22ms	938Mb/s – 0.04ms	
Sophia	68.1Mb/s – 5.2ms	42.4Mb/s – 8.6ms	40.2Mb/s – 9.1ms	940.5Mb/s – 0.04ms		
Rennes	110.1Mb/s – 4.0ms	95.4Mb/s – 4.7ms	940.4Mb/s – 0.04ms			
Orsay	108.1Mb/s – 4.1ms	930.4Mb/s – 0.06ms				
Bordeaux	940.2Mb/s – 0.04ms					



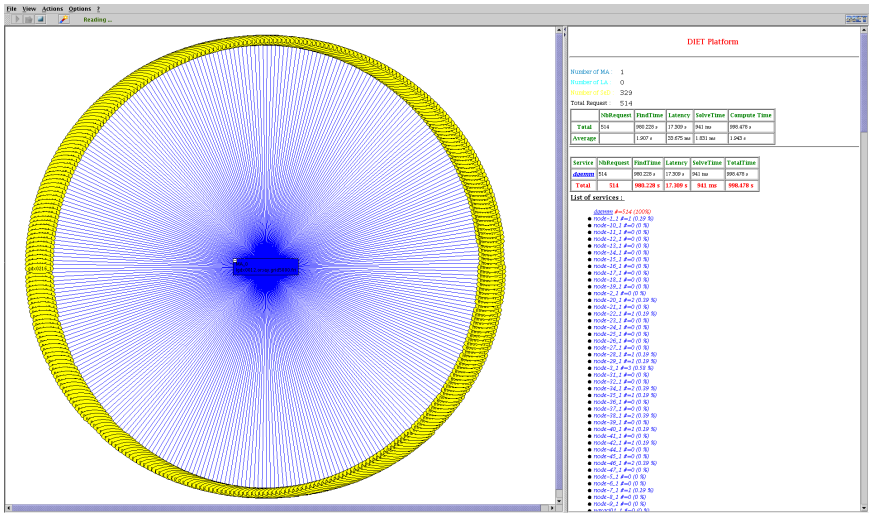


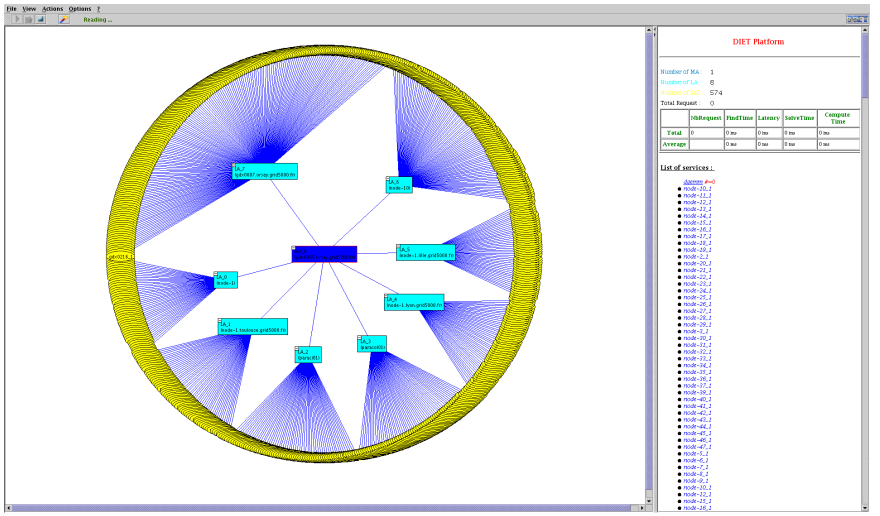






Large Scale experiment of DIET: A GridRPC environment





1120 clients submitted more than 45 000 REAL GridRPC requests (dgemm matrix multiply) to GridRPC servers

7 sites : Lyon, Orsay,Rennes, Lilles, Sophia,Toulouse,Bordeaux

8 clusters - 585 machines - 1170 CPUs.

Objectives :

- Proove that the DIET environment is scallable.

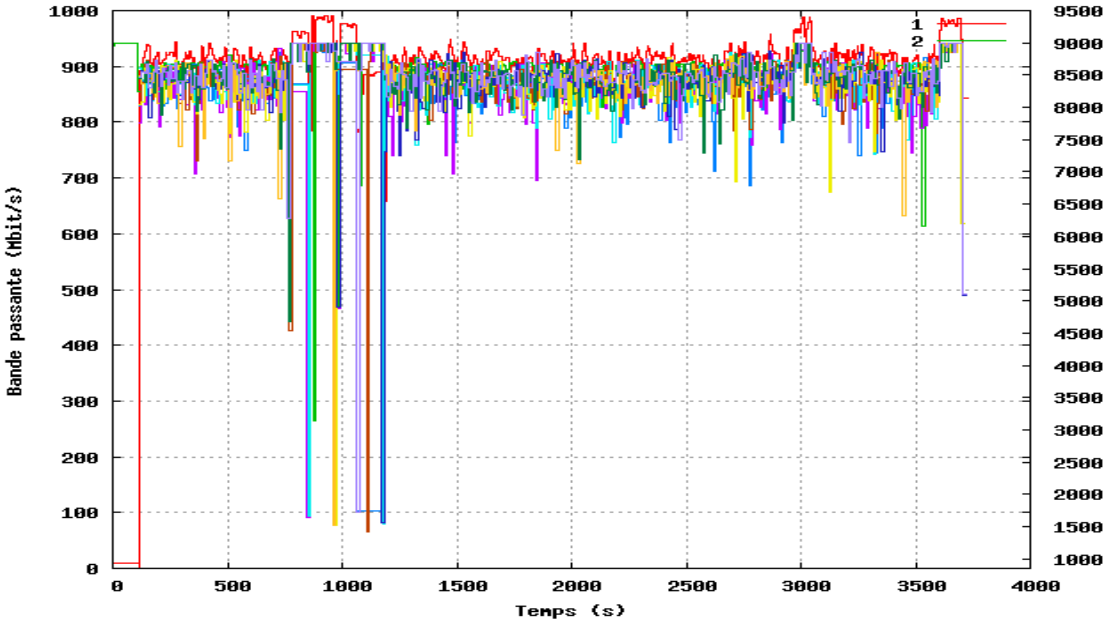
- Test the functionnalities of DIET at large scale

Raphaël Bolze

TCP limits over 10Gb/s links

- Highlighting TCP stream interaction issues in very high bandwidth links (congestion colapse) and poor bandwidth fairness
- Grid'5000 10Gb/s connections evaluation.
- Evaluation of TCP variants over Grid'5000 10Gb/s links (BIC TCP, H-TCP, weswood…)

Interaction of 10 1Gb/s TCP streams, over the 10Gb/s Rennes-Nancy link, during 1 hour.



Aggregated bandwidth of 9,3 Gb/s on a time interval of few minutes. Then a very high drop of the bandwidth on one of the connection.

Next Step

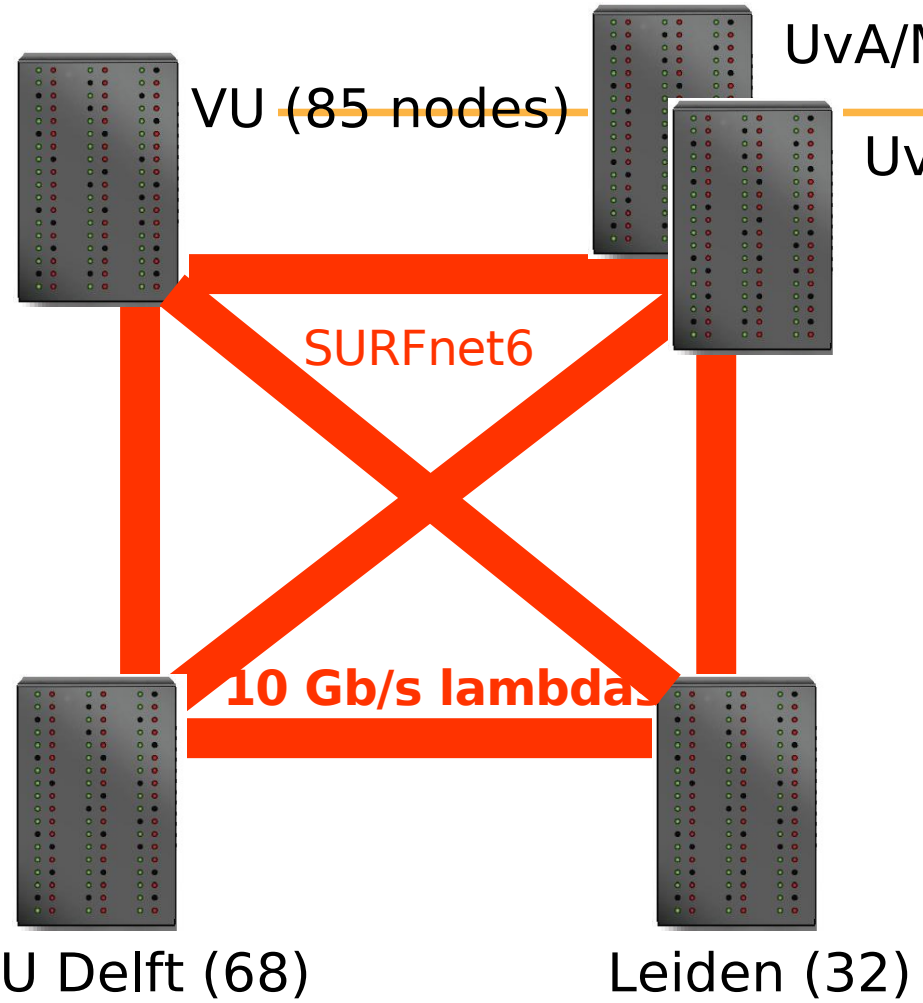
European (DAS)

and International collaborations

DAS is a *Computer Science* grid

- Motivation: CS needs its own infrastructure for
  - Systems research and experimentation
  - Application experiments
- DAS is simpler and more homogeneous than most production grids
  - Single operating system
  - “A simple grid that *works*’’

DAS-3: overall structure



Operational: oct. 2006

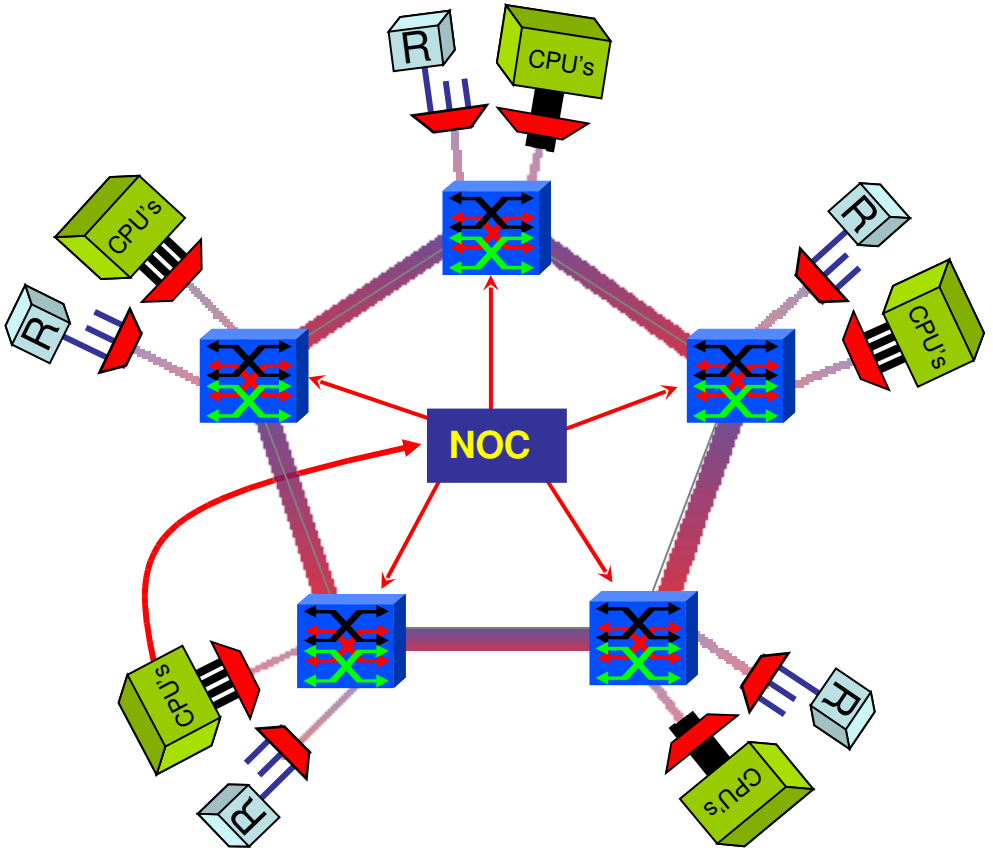
- 272 AMD Opteron nodes
- 792 cores, 1TB memory
- more heterogeneous:
  2.2-2.6 GHz
  single/dual core nodes
- Myrinet-10G (not in Delft)
- Gigabit Ethernet

Advanced School for Computing and Imaging

Netherlands Organisation for Scientific Research

Projects using DAS: StarPlane

- Key idea:
  - Applications can dynamically allocate light paths
  - Applications can change the topology of the wide-area network, possibly even at the sub-second timescale



DAS-3 /GRID'5000

- Vers l'interconnection des 2 plateformes
  - interconnection réseau via GEANT
  - utilisation de l'ordonnanceur pour grille KOALA
    - adaptateur OAR en cours de développement
  - Environnement DAS3 sur Grid'5000
  - Accès au réseau StartPlane par la communauté Grid'5000

Japon / Grid'5000

- Interconnexion réseau plus délicate
  - en cours
- Expérience sur réseau longue distance avec équipement réseau spécifique
  - Projet RESO (INRIA/LIP)

Liens avec les autres communautés

- Des liens naturels existent dans plusieurs sites:
  - Grenoble: communauté CIMENT
  - Lyon: IN2P3
  - Orsay: ...
- Echange avec EGEE, mais rien de vraiement concret pour l'instant

Conclusion

Derniers mots

- Grid'5000: une plate-forme pour

  experimental computer science
- *“En production”*
- Collaboration avec les autres communautés scientifiques
- Formation aux grilles
- *Charge de 25% dans DAS2 (Grid'5000 ?)*
- *Job Best-Effort*

ANNEXES

Environment switch

Reboot & deployment vs Virtual Machines

Reboot:

Remote control with IPMI, RSA, etc.

Disc repartitioning, if necessary

Reboot or Kernel switch (Kexec)

Virtual Machine:

No need for reboot

Virtual machine technology Selection not so easy

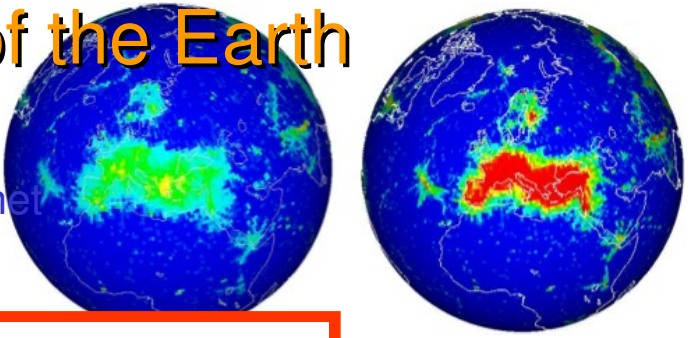
Xen has some limitations:
-Xen3 in "initial support" status for Intel VT
-Xen2 does not support x86/64
-Many patches not supported
-High overhead on high speed Net.

→ Currently we use Reboot, but Xen will be used in the default environment.

Let users select its experimental environment:
Fully dedicated or shared within virtual machine

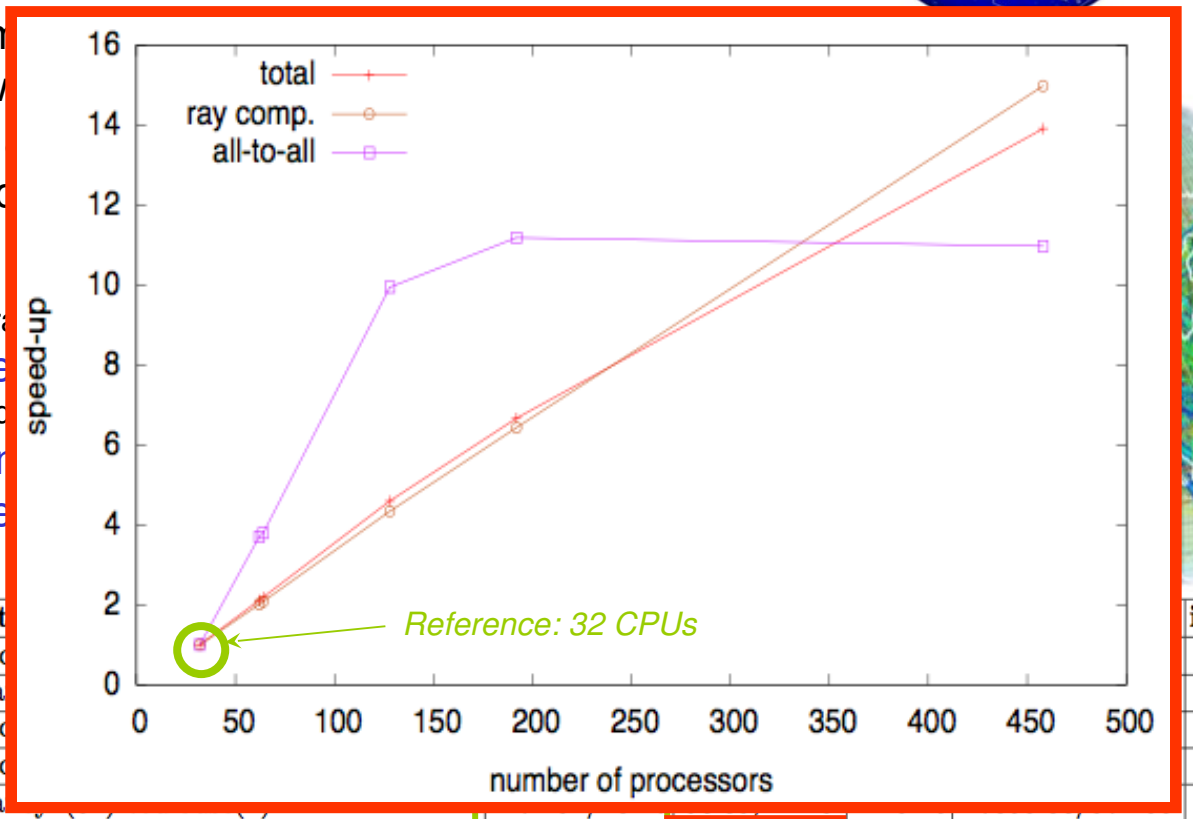
Experiment: Geophysics: Seismic Ray Tracing in 3D mesh of the Earth

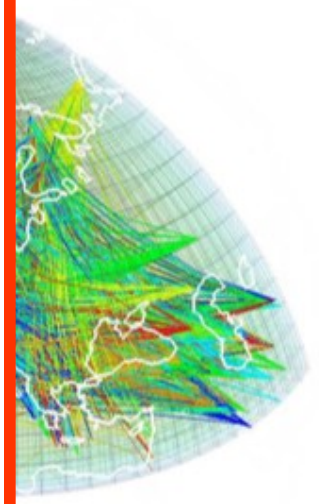
Stéphane Genaud , Marc Grunberg , and Catherine Mongenet
IPGS: “Institut de Physique du Globe de Strasbourg”



Building a seism
using seismic w
Seismic waves
Ray tracing algo
traced between

A MPI parallel progra
1) Master-worke
successively fetched
2) all-to all comm
3) merging of ce





sites	procs	sit					in transit
1	32	nic					7.29 GB
	62	na					9.02 GB
	62	nic					8.88 GB
	138	nic					15.47 GB
2	64	na					9.26 GB
	128	nancy*(62) toulouse(66)	610.91/20.40	34.68/3.08	688.28	8629.14/573.01	15.29 GB
3	128	rennes(42) nancy*(44) toulouse(42)	620.27/21.21	33.56/2.98	699.57	8629.14/648.83	15.47 GB
	192	rennes(64) nancy*(64) toulouse(64)	412.16/13.70	30.84/2.23	474.65	5737.70/410.90	16.77 GB
5	458	rennes(152) nancy*(32) orsay(184 nice(58) toulouse(32)	177.07/5.69	31.43/1.47	227.53	2398.03/221.91	20.82 GB

Solving the Flow-Shop

"one of the hardest challenge problems in combinatorial optimization"

- Schedule a set of jobs on a set of machines minimizing the *makespan*.
- Jobs order must be respected and machines can execute 1 job at a time.
- Complexity is very high for large size instances (possible schedules).
- Exhaustive enumeration of all combinations would take several years.
- The challenge is thus to reduce the number of explored solutions.
- But the problem cannot be efficiently solved without computational grids.

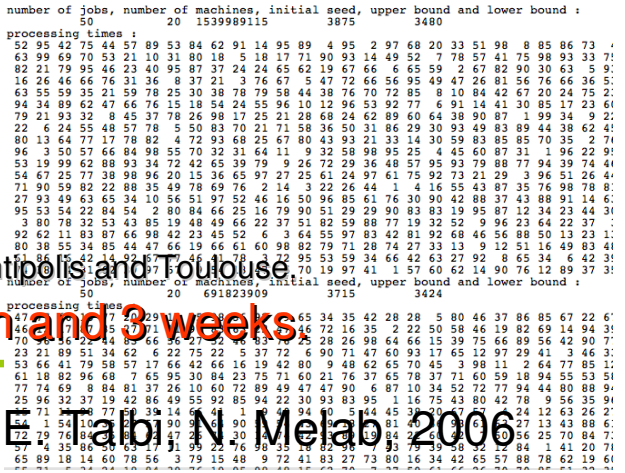
→New Grid exact method based on the Branch-and-Bound algorithm (Talbi, melab, et al.), combining new approaches of combinatorial algorithmic, grid computing, load balancing and fault tolerance.

→ Problem: 50 jobs on 20 machines, optimally solved for the 1st time, with 1245 CPUs (peak)

Using simultaneously Grid5000 and other clusters

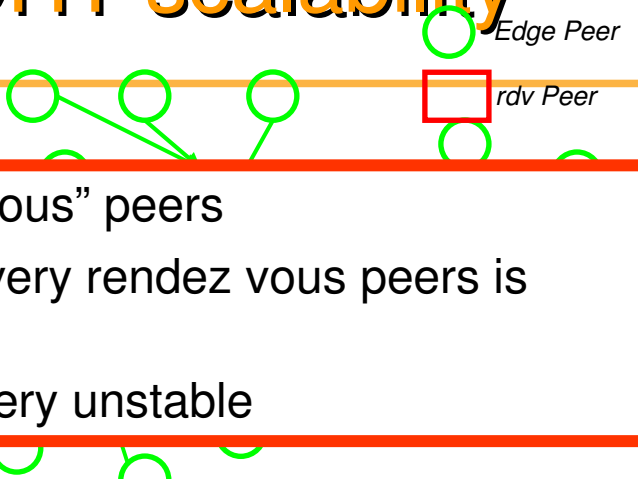
Involved Grid5000 sites (6): Bordeaux, Lille, Orsay, Rennes, Sophia-Antipolis and Toulouse.

The optimal solution required a wall-clock time of 1 month and 3 weeks.

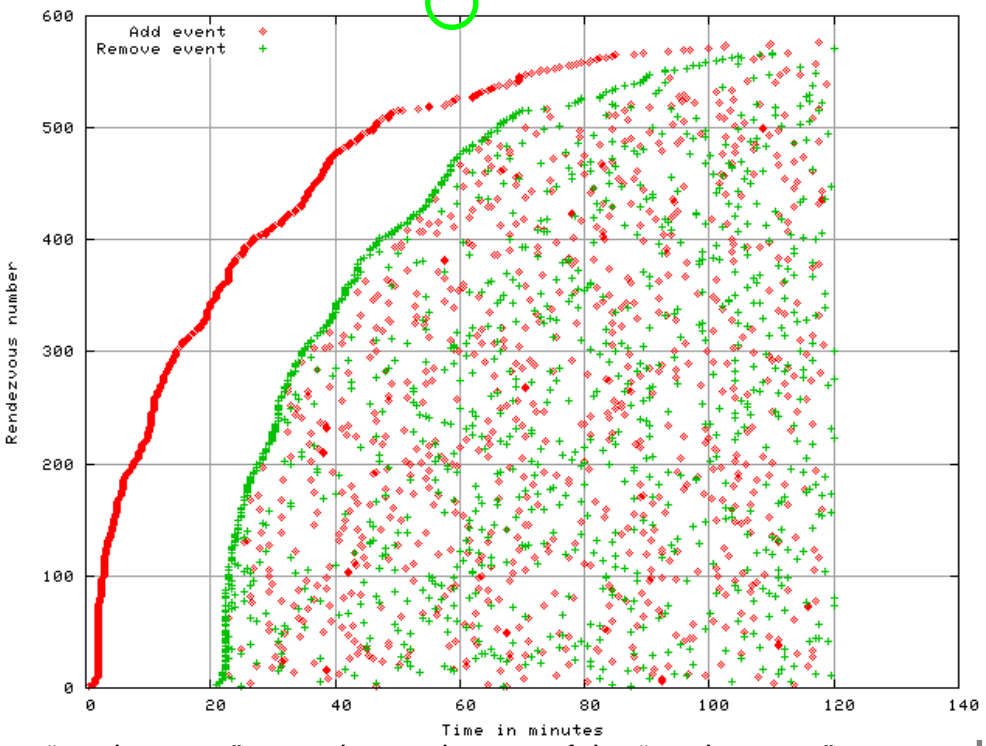


E.Talbi N.Melab 2006

Jxta DHT scalability



- Goals: study of a JXTA "DHT"
  - It requires 2 hours to contact all "rendez vous" peers
  - With the per default setting, the view of every rendez vous peers is limited to only 300 rendez vous peers
  - The view of every "rendez vous" peer is very unstable
- Organization of a JXTA overlay (peerview protocol)
  - Each rendezvous peer has a local view of other rendezvous peers
  - Loosely-Consistent DHT between rendezvous peers
  - Mechanism for ensuring convergence of local views
- Benchmark: time for local views to converge
  - Up to 580 nodes on 6 sites

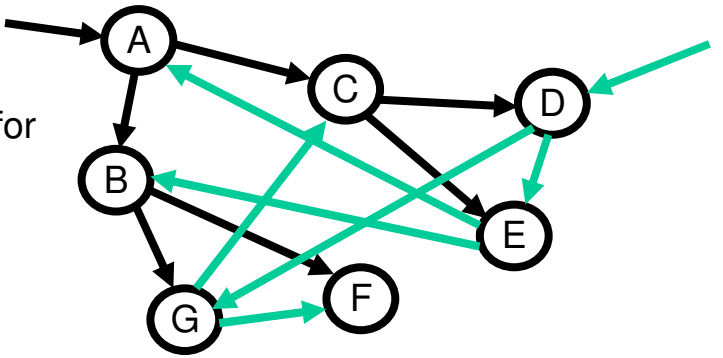


"rendez vous" peers known by one of the "rendez vous" peer

X axis: time ; Y axis: "rendez vous" peer ID

Fully Distributed Batch Scheduler

- **Motivation** : evaluation of a fully distributed resource allocation service (batch scheduler)
- **Vigne** : Unstructured network, flooding (random walk optimized for scheduling).
- **Experiment**: a bag of 944 homogeneous tasks / 944 CPU
  - Synthetic sequential code (monte carlo application).
  - Measure of the mean execution time for a task (computation time depends on the resource)
  - Measure the overhead compared with an ideal execution (central coordinator)
  - Objective: 1 task per CPU.



mean: 1972 s

- **Tested configuration**:
  - 944 CPUs
  - Bordeaux (82), Orsay(344), Rennes Paraci (98), Rennes Parasol (62), Rennes Paravent (198), Sophia (160)
  - Duration: 12 hours
- **Result :**

Grid@work 2005, 2006

- Series of conferences and tutorials including
- Grid PlugTest (N-Queens and Flowshop Contests).

The objective of this event was to bring together ProActive users, to present and discuss current and future features of the ProActive Grid platform, and to test the deployment and interoperability of ProActive Grid applications on various Grids.

The N-Queens Contest (4 teams) where the aim was to find the number of solutions to the N-queens problem, N being as big as possible, in a limited amount of time

The Flowshop Contest (3 teams)

- **2005:** a total of 1600 CPUs: 1200 provided by Grid'5000
- **2006:** a total of 4000 CPUs: 2300 provided by Grid'5000

Charter & Accounting

- A chart of the good Grid'5000 citizen
- Accounting tools Kaspied
- User Reports to record experiments/publication/collaboration