



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, LSIT)
 - 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. Messenger, 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

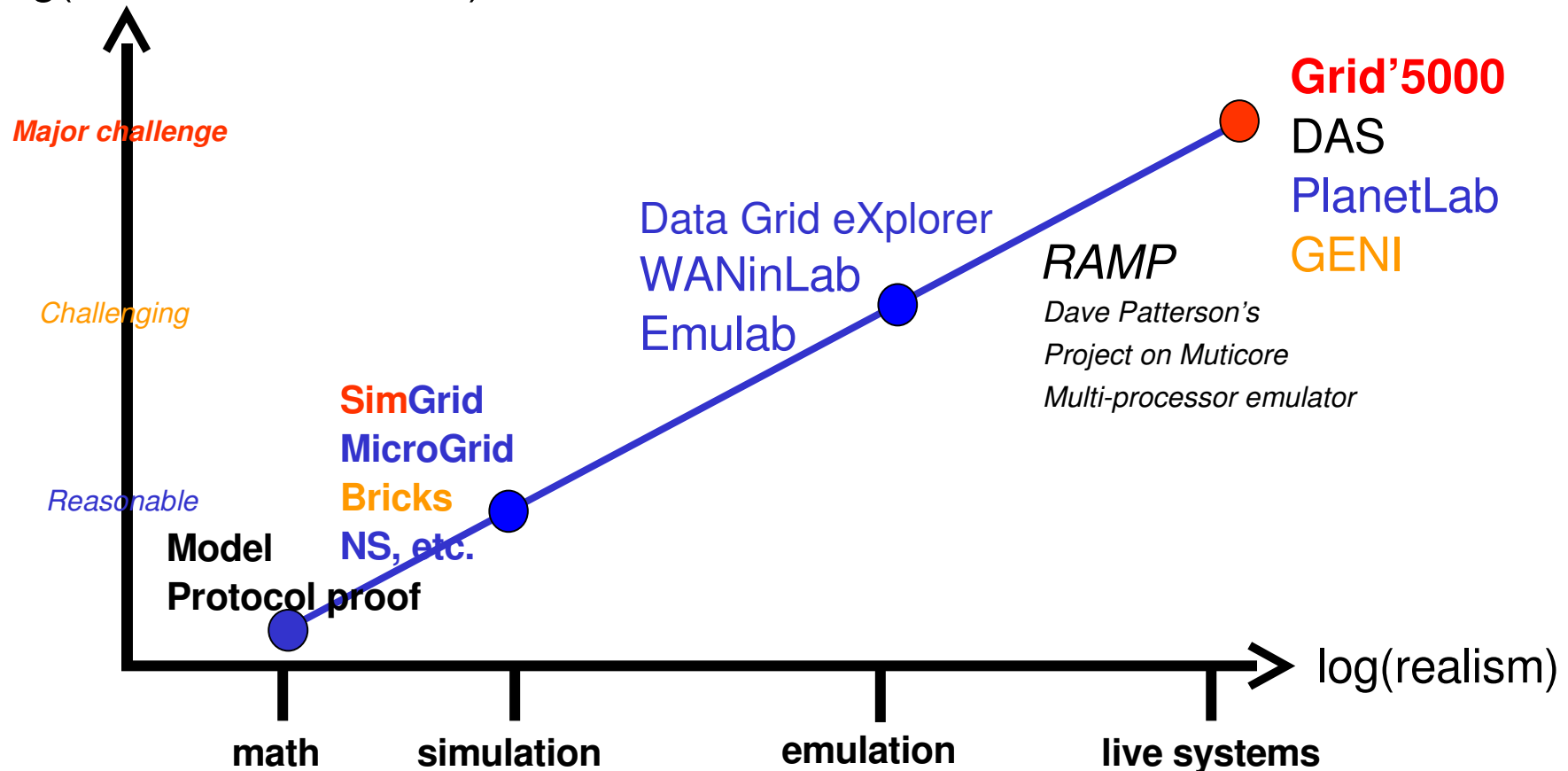
- **SimGRid and SimGrid2**
 - Discrete event simulation with trace injection
 - Originally dedicated to scheduling studies
 - Single user, multiple servers
 - **GridSim**
 - Dedicated to scheduling (with deadline), DES (Java)
 - Multi-clients, Multi-brokers, Multi-servers
 - **Titech Bricks**
 - Discrete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

log(cost & coordination)



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.

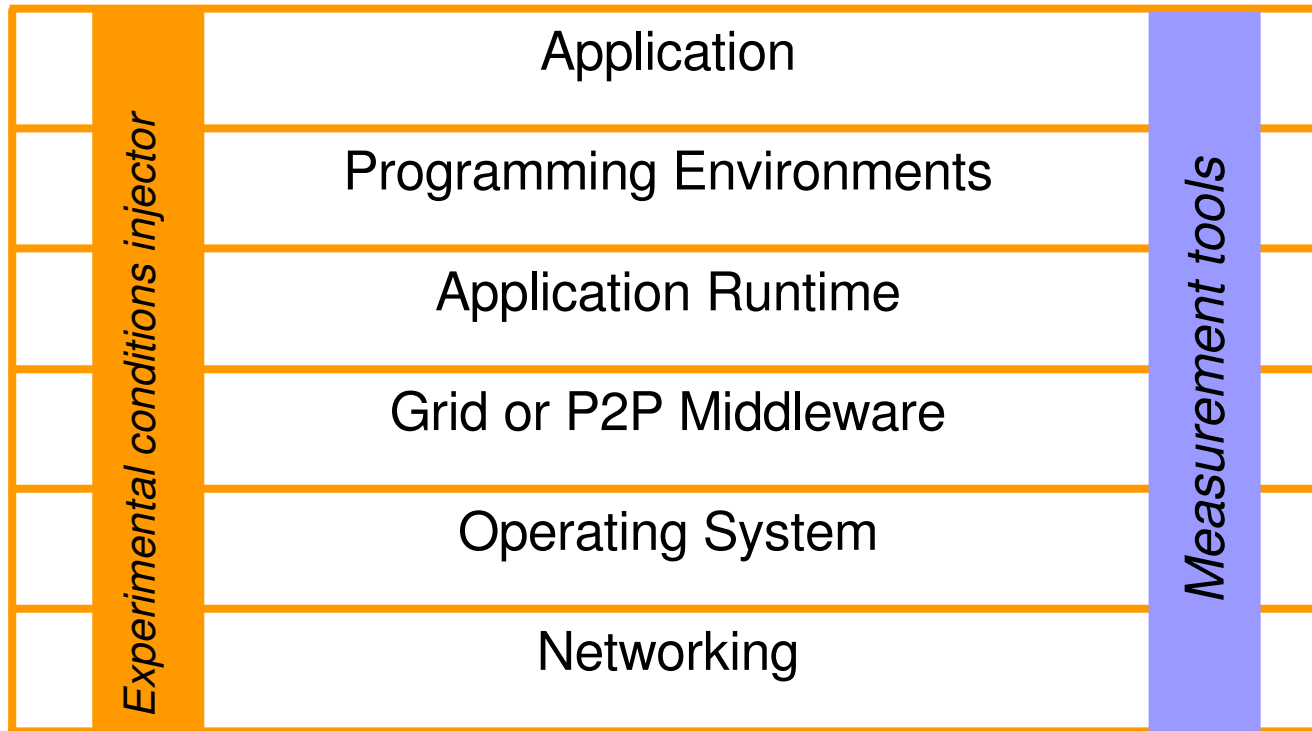
Experimental Condition injection :

- **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

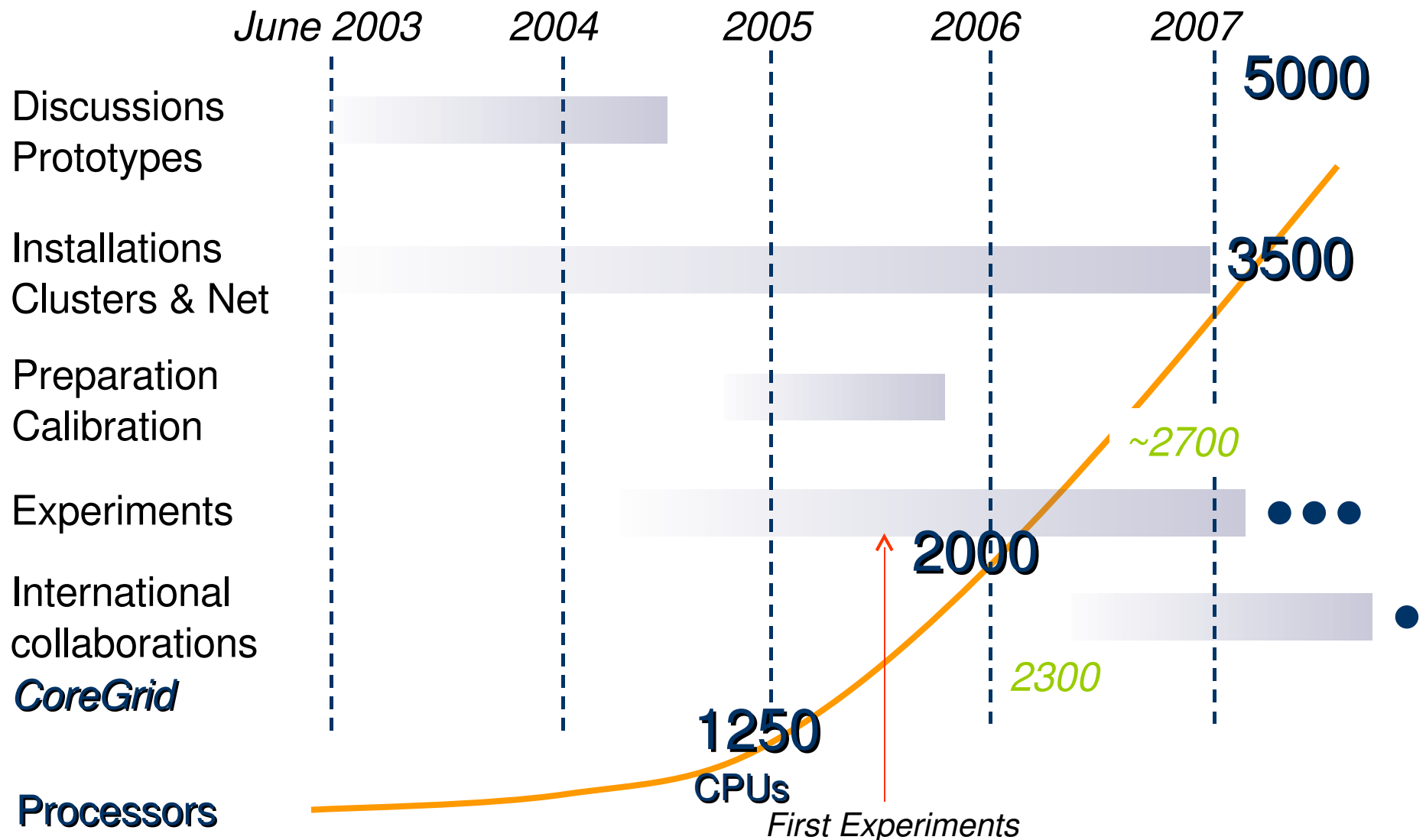
Grid'5000 principle

A highly reconfigurable experimental platform



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

Timeline





Technical aspects

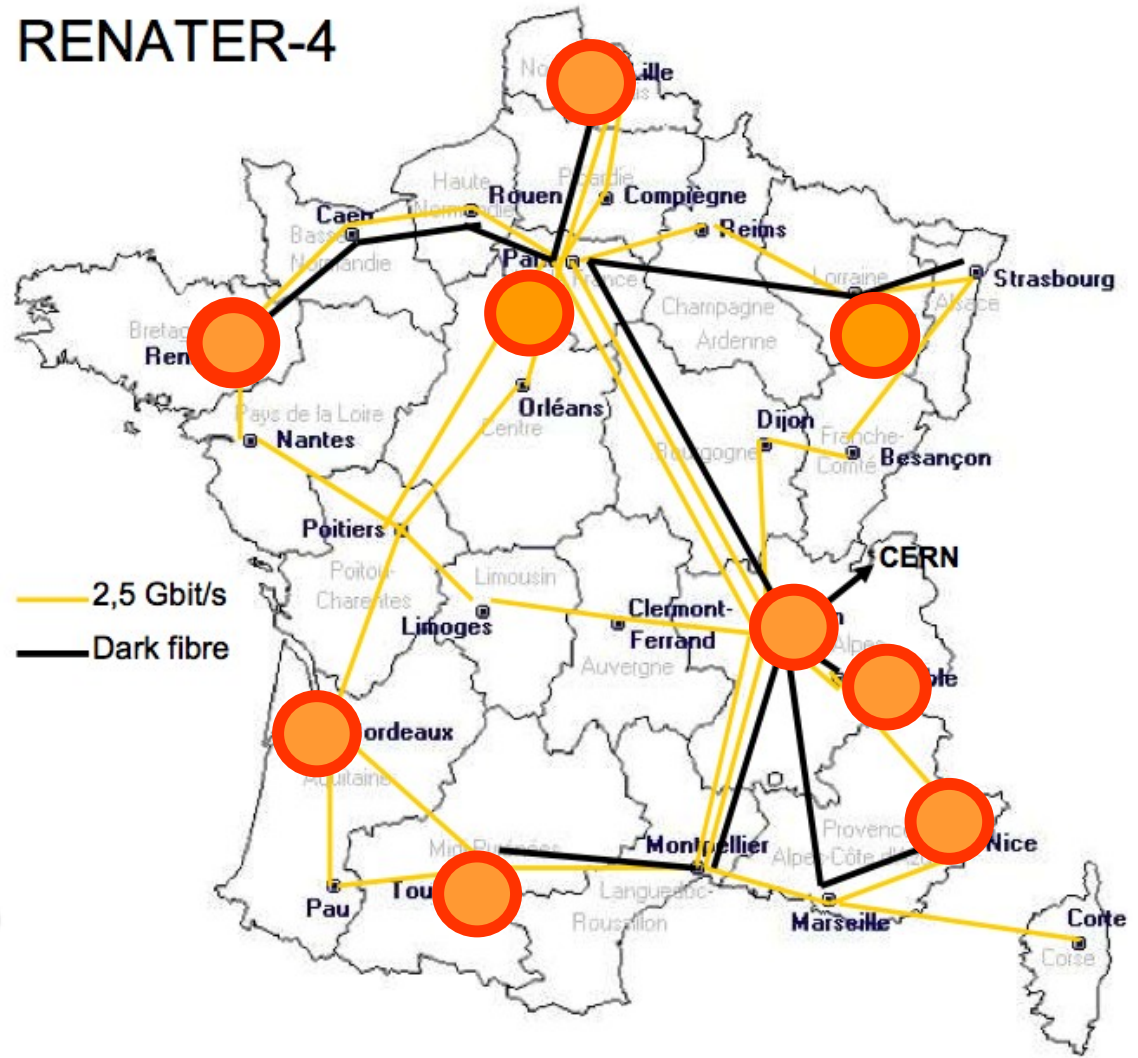
Grid'5000 components

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

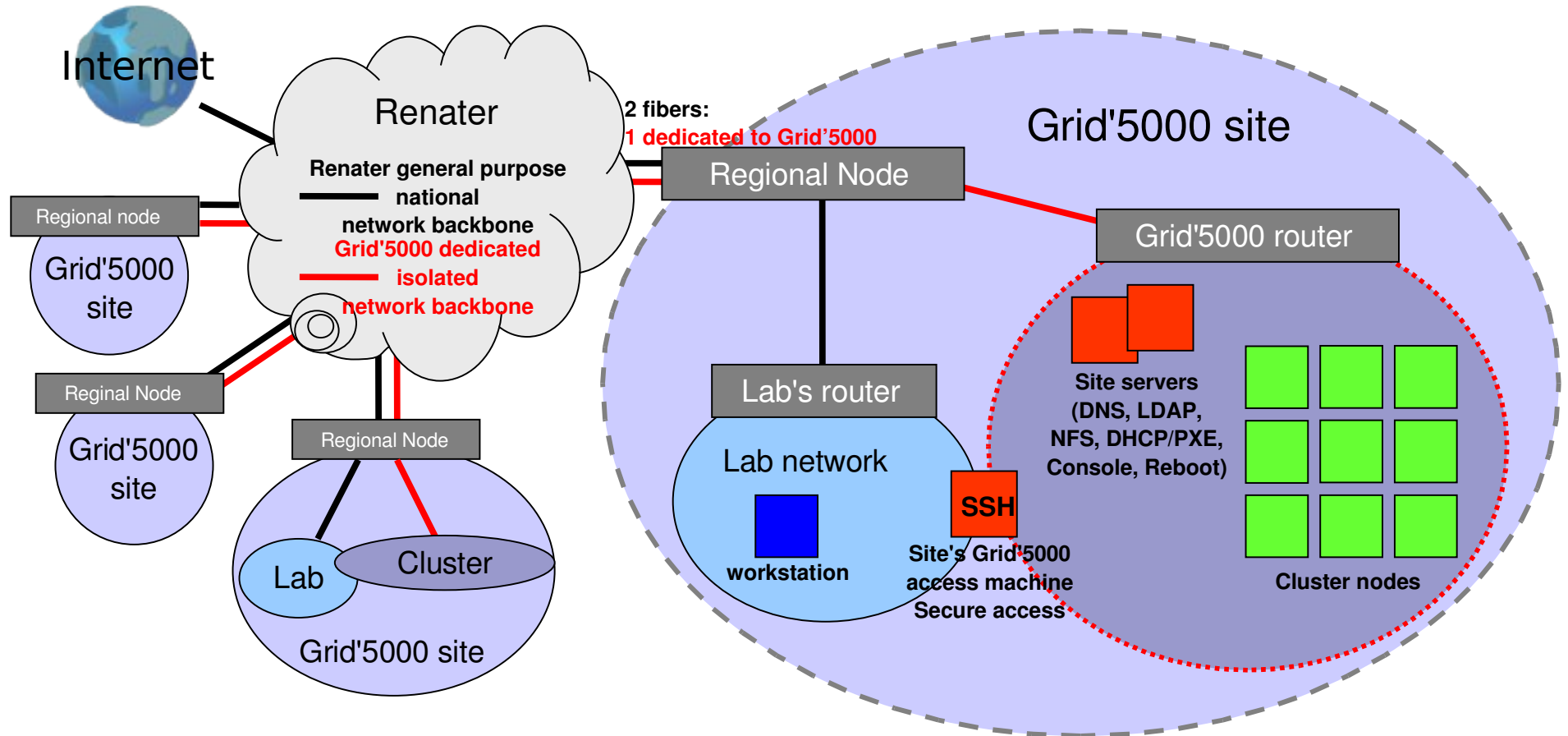
Backbone network

Nationwide backbone network through **Renater-4**

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



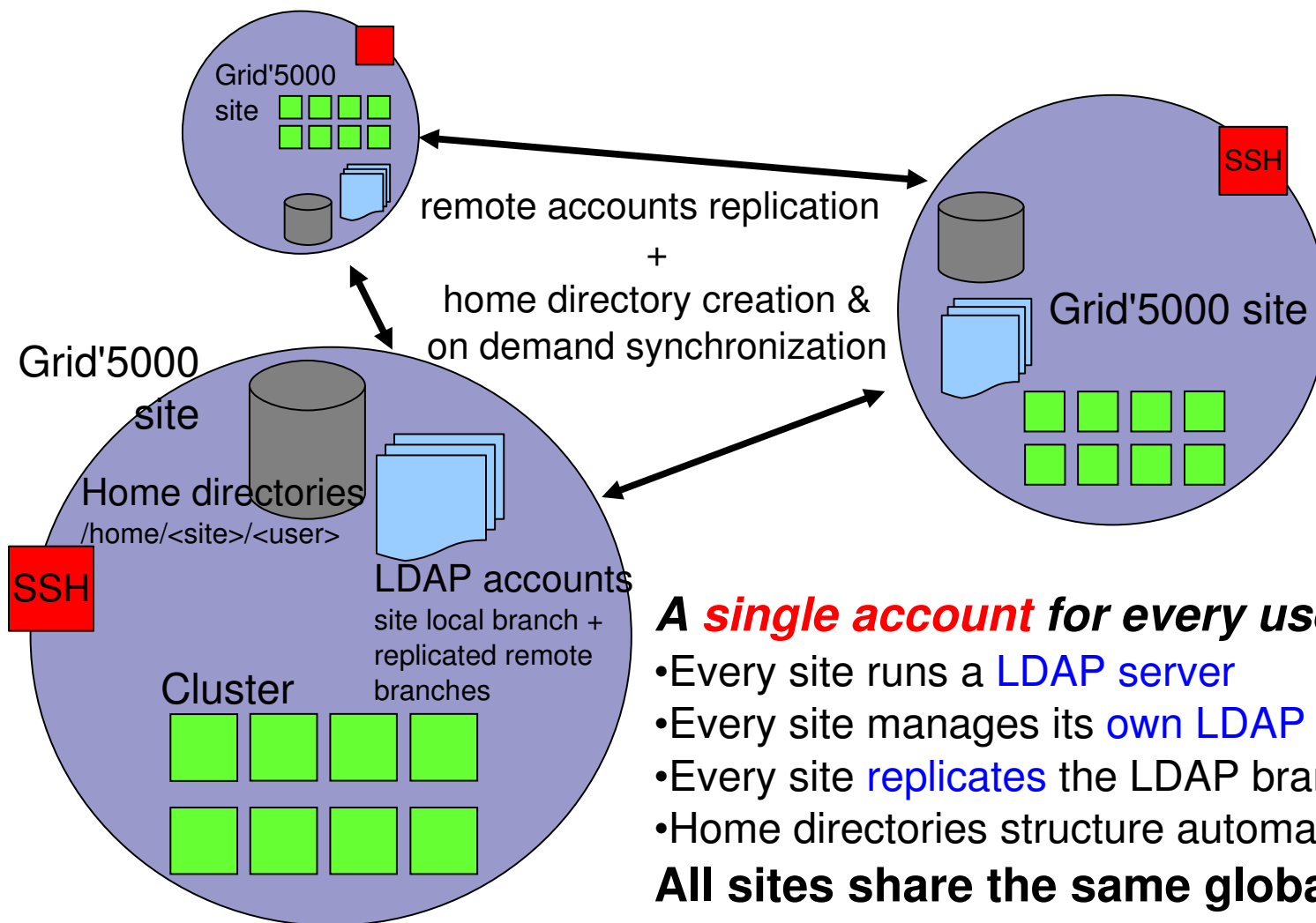
Security architecture



Protecting **Grid'5000** AND Protecting **Internet**:

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

Account management



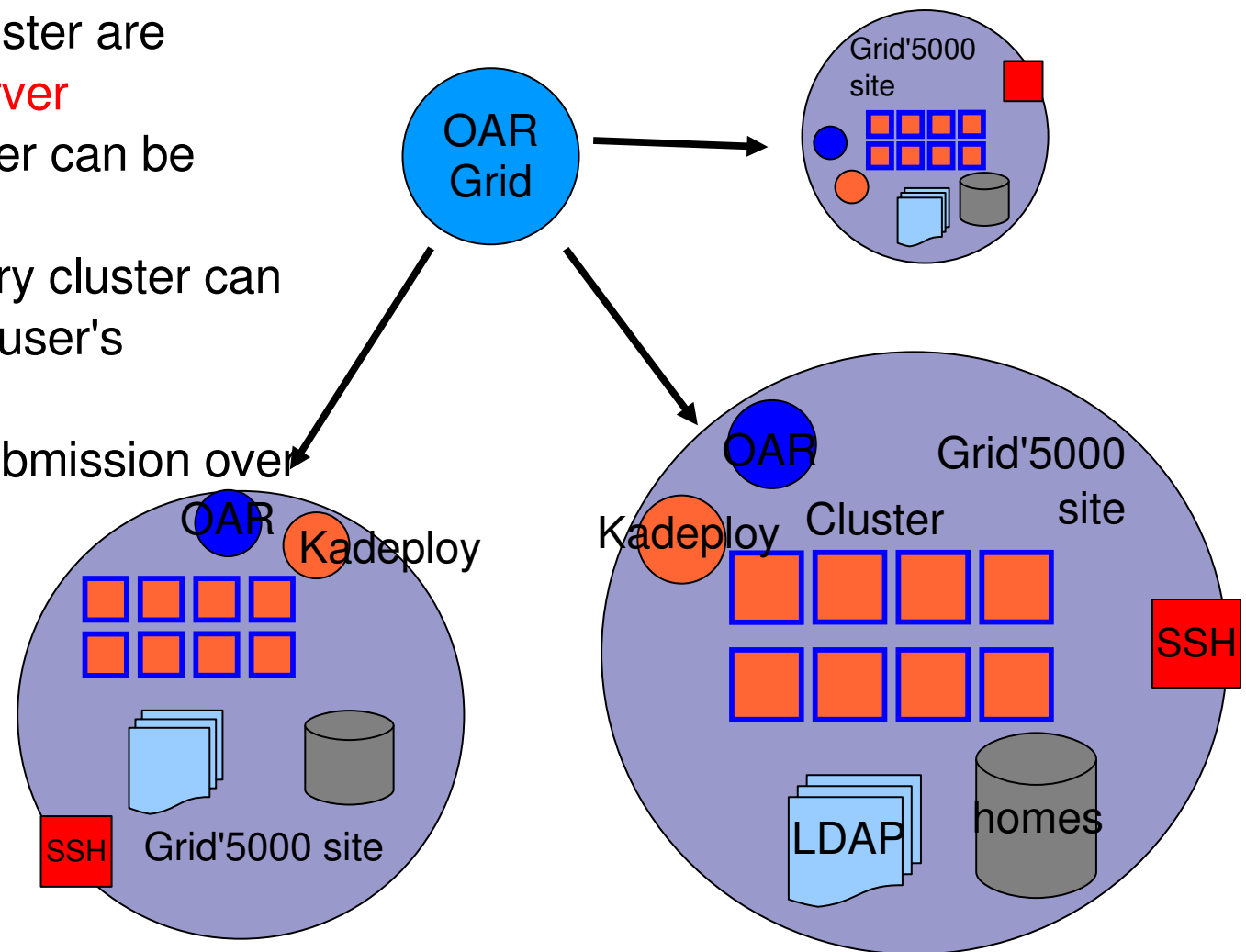
A single account for every user for every sites:

- 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:

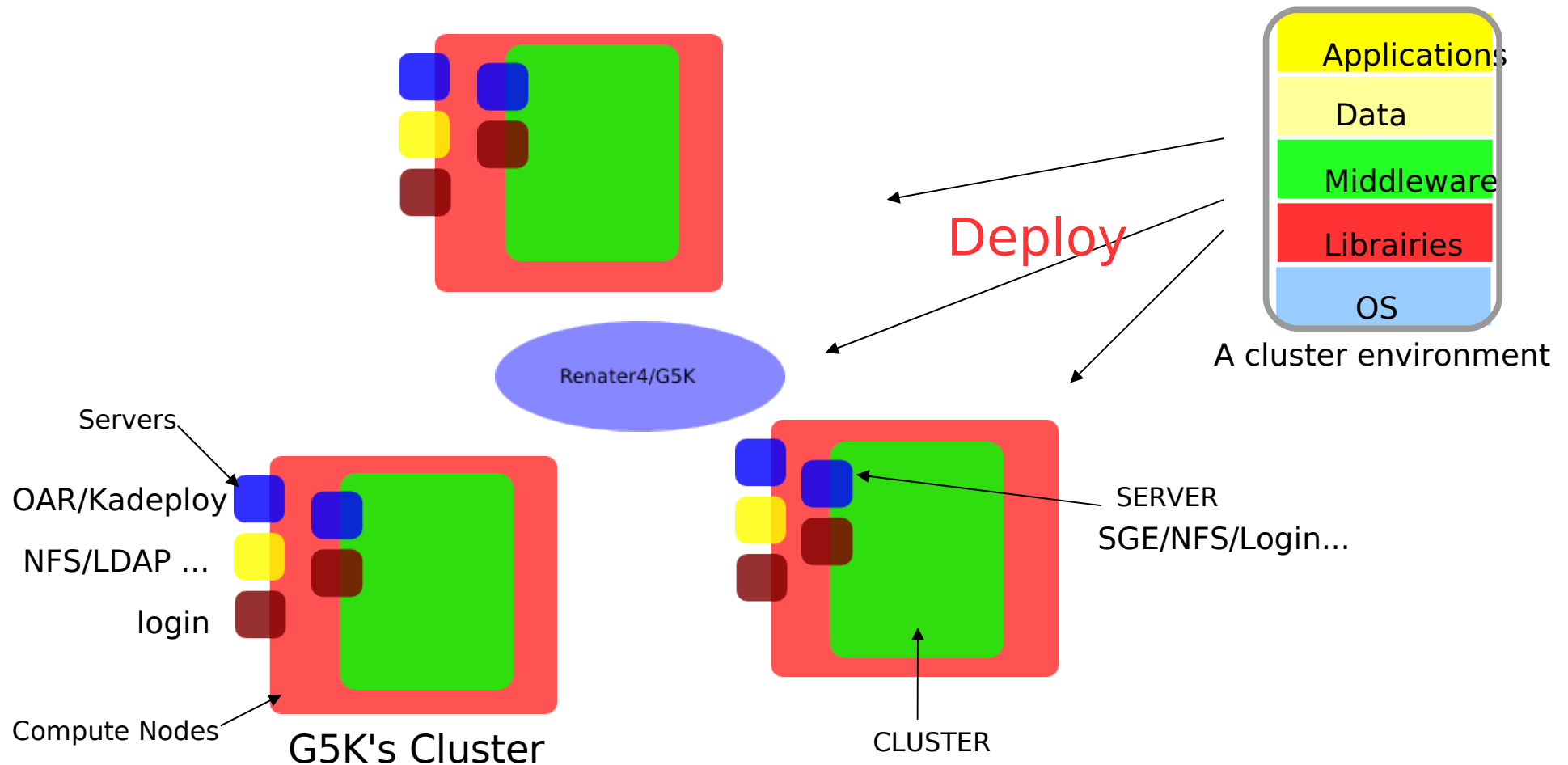
- **Deep control** (allow to reboot nodes, access to the console...)
- **Unrestricted access** (gain administrator privilege, kernel patch, modified network behavior)
- **Unlimited configuration possibilities** (complete customized grid environment deployment)

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



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

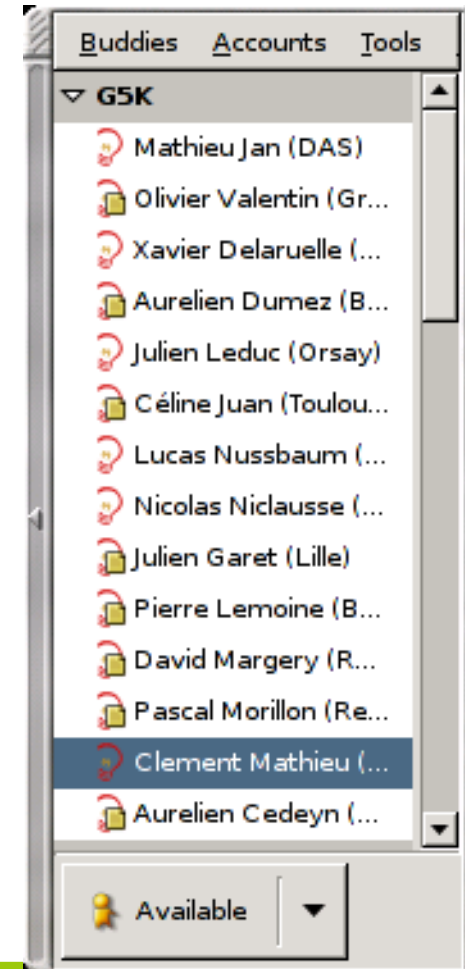
Collaboration between site is one challenge:

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*



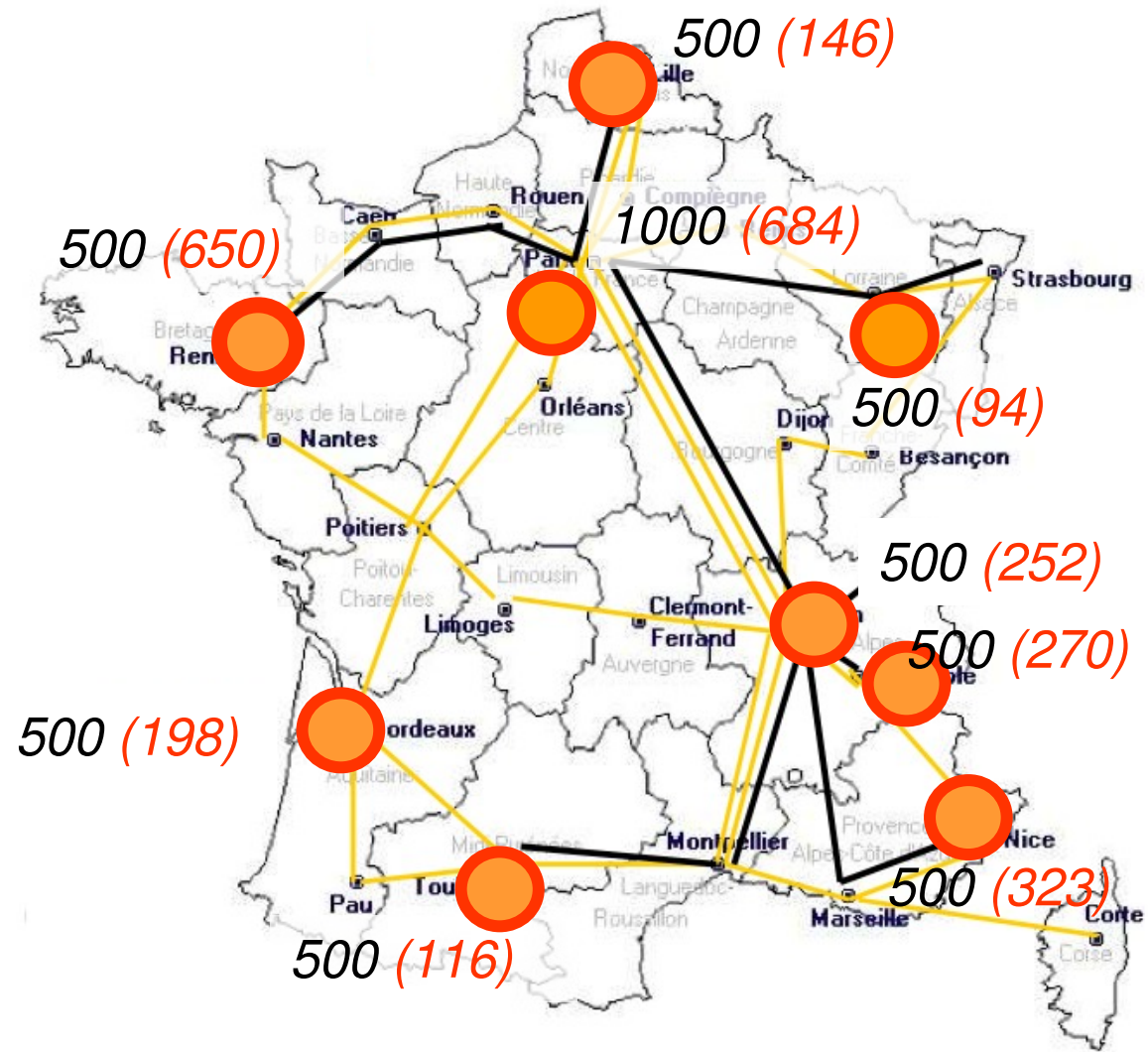


Status

Grid'5000 sites

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

2007: 3000+ CPU



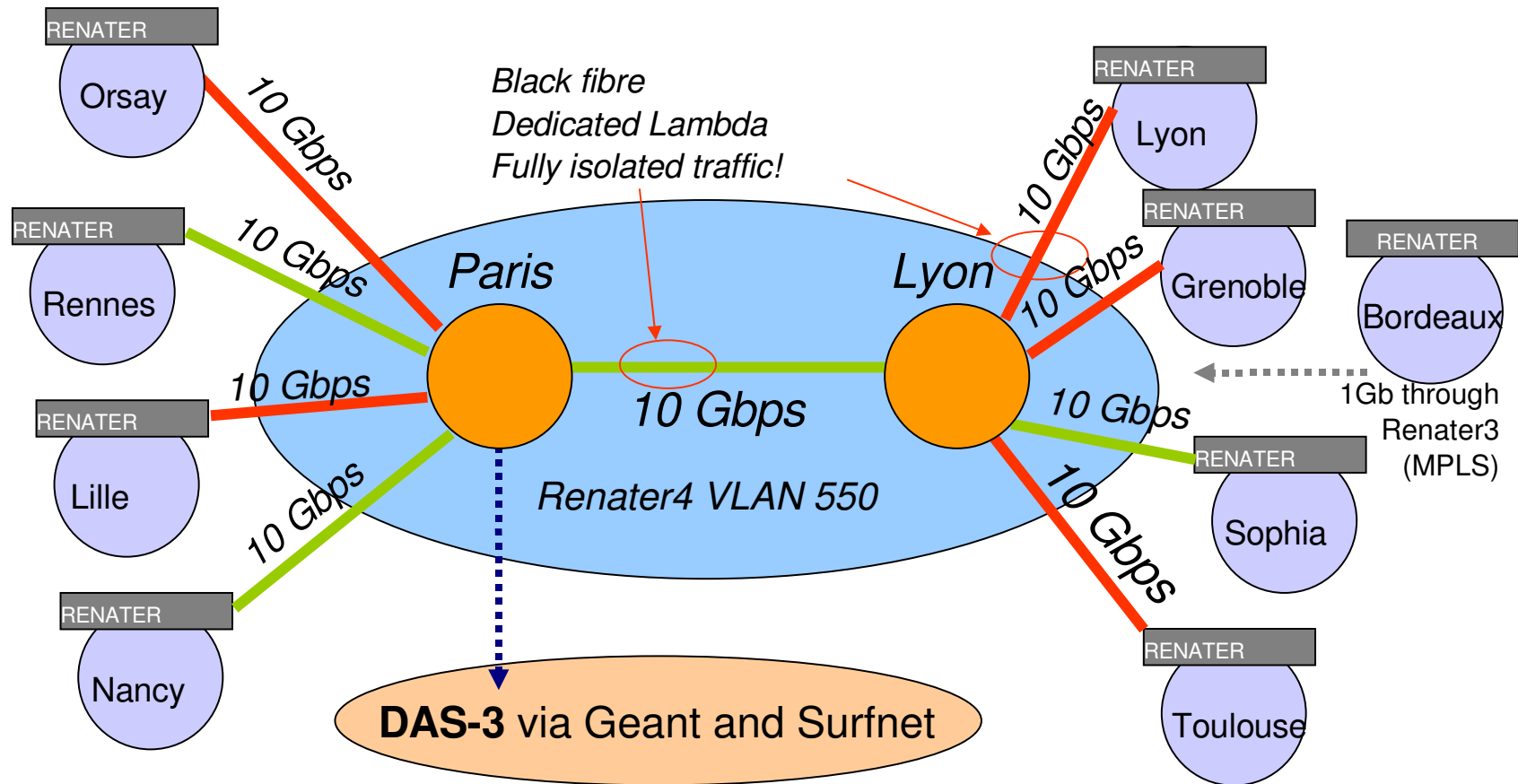
Grid'5000 hardware

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	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		
Rennes	32												
Sophia			96					102			198		
Toulouse							206			64	270		
Nodes total	32		106		40						146		
		112		140							252		
		94									94		
		432		252							684		
		198	128					132	128	64	650		
		210			112						322		
			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

9/01/2007

File Edit View Go Bookmarks Tools Window Help

Mozilla Firefox

All Grid'5000 reports:

- Hamza Adamou (M2R), MESCAL ID IMAG Grenoble
- Guillaume ALLEON (Engineer), EADS CRC
- Lamine Aouad (PhD student), Grand Large LIFL Lille
- Carlos Jaime BARRIOS HERNÁNDEZ (PhD Student), MESCAL ID-IMAG Montbonnot Saint-Martin (Grenoble-France)
- Janet Bertot (ingé devexp), service Dream INRIA Sophia
- Raphaël Bolze (PhD student), GRAAL LIP-ENSL Lyon
- Hinde Lilia Bouziane (PhD student), PARIS IRISA/INRIA Rennes
- Jeremy BUISSON (PhD student), PARIS IRISA Rennes
- CHRISTOPHE CERIN (professor), Grid Explorer LIPN Paris XIII, Villetaneuse
- Arnaud Contes (Phd), OASIS INRIA Sophia
- Cedric Dalmasso (Internship), Oasis INRIA sophia
- Alexandre di Costanzo (PhD. Student), OASIS INRIA Sophia
- Fabrice Dupros (engineer), IGGI BRGM Orleans
- Thierry Gautier (CR INRIA), MOAIS ID-IMAG Grenoble
- Stéphane Genaud (Maitre de Conférences), TAG ICPS-LSIIT Strasbourg
- Yiannis Georgiou, Mescal ID-IMAG Grenoble
- Olivier GLUCK (Associate Professor), INRIA RESO LIP ENS-LYON
- Jens Gustedt (directeur de recherche), AlGorille INRIA Lorraine & LORIA Nancy, France
- Christophe Hamerling (Engineer), GRID-TLSE IRIT-ENSEEIH Toulouse
- Thomas Hérault (Assistant Professor), Grand-Large LRI Orsay
- Samir Jafar (PhD Student), MESCAL ID-IMAG Montbonnot
- Emmanuel Jeannot (Chargé de recherche), Algorille LORIA Nancy
- Emmanuel Jeanvoine (PhD Student), PARIS IRISA Rennes
- Peyrard Johann (developer), MESCAL ID-IMAG Montbonnot
- Nicolas LARRIEU (Postdoctoral fellow), Grid Explorer LAAS-CNRS Toulouse
- Adrien Lebre (PhD Student), MESCAL ID-IMAG Montbonnot (Grenoble-France)
- Julien Leduc (Research Engineer), MESCAL ID-IMAG Montbonnot
- Laurent Lefevre (INRIA CR1 Researcher), RESO LIP Lyon
- Oleg Lodyginsky (PhD student), MESCAL ID-IMAG Montbonnot
- Eric MAISONNAVE (Engineer), IEGO (submitter), Cerfacs Toulouse
- Maxime Martinasso (Phd student), MESCAL ID-IMAG Grenoble
- Sébastien Monnet (PhD student), PARIS IRISA Rennes
- Thierry Monteil (Assistant professor), AROMA LAAS-CNRS Toulouse
- Matthieu Morel (Ingénieur expérimental), INRIA Sophia Antipolis
- Grégory Mounié (Assistant Professor), MOAIS ID-IMAG Grenoble
- Frederic NIVOR (PhD), STM LAAS-CNRS Toulouse, France

users portal

- Joining
- News
- Time line
- Experiments
- Publications
- Press releases
- Softwares

users portal

- Users portal
- Platform events
- Platform status
- User Reports
- Documentation
- FAQ

committees portal

- Agenda
- Members
- Meetings
- Workgroups
- Administration

wiki special pages

- Recent changes
- All pages
- Upload file
- Wiki help

search

Go Search

toolbox

- Upload file
- Special pages

UFRJ.br
LABRI
LIFL
ENS-Lyon
EC-Lyon
IRISA
RENATER
IN2P3
LIFC
LIP6
UHP-Nancy

France-telecom
LRI
IDRIS
AIST.jp
UCD.le
LIPN-Paris XIII
U-Picardie
EADS
EPFL.ch
LAAS
ICPS-Strasbourg

Univ.Nantes
Sophia
CS-VU.nl
FEW-VU.nl
Univ. Nice
ENSEEIH
CICT
IRIT
CERFACS
ENSIACET
INP-Toulouse
SUPELEC

Done

http://grid5000.fr

250+ experiments

The screenshot shows a web browser window with the title "Grid'5000 experiments - Grid5000". The address bar contains the URL "https://www.grid5000.fr/mediawiki/index.php/Special:G5KExperiments". The browser's search bar shows "Google". The page content includes a navigation menu on the left with sections for "public portal", "users portal", "committees portal", and "wiki special pages". The main content area is titled "Grid'5000 experiments" and contains a summary paragraph, a "Networking" section header, and a list of 20 experimental projects, each with a status indicator in brackets.

Grid'5000 experiments

A summary of the domains of experiment which Grid'5000 is providing a research platform for can be found on [this page](#). Experiments actually performed on the platform are listed below.

Networking

- [Benchmarking of network management plateforms \(SNMP, JMX\)](#)
- [A distributed GRID monitoring architecture driven by models](#)
- [LSCAN \(Large Scale Programmable Networking\) \[planned\]](#)
- [A Distributed Network Measurement System \[planned\]](#)
- [PadicoTM \[in progress\]](#)
- [ALTA \[in progress\]](#)
- [MPICH/Madeleine \[in progress\]](#)
- [Isolation réseau sur la grille par l'attribution dynamique de VLAN \[in progress\]](#)
- [OAR Fault management \[in progress\]](#)
- [alOLi - I/O Scheduler for High Performance Computing \[in progress\]](#)
- [Study of peer-to-peer systems using emulation \[in progress\]](#)
- [Tamanoir tests \(grid simulation\) \[in progress\]](#)
- [Stress of 10G interconnection link \[in progress\]](#)
- [DIET-FD \(Fault detection\) \[in progress\]](#)
- [Network telescope analysis \[in progress\]](#)
- [Data redistribution \[in progress\]](#)
- [benchmarking of par::cell and par::cellnet \[in progress\]](#)
- [Data Transfer Time Forecasting \(Network monitoring\) \[in progress\]](#)
- [Optimization of Long-distance communications for MPICH-Madeleine \[in progress\]](#)
- [Pipelined Broadcasts \[in progress\]](#)
- [Grid Gateway using Intel IXP2400 Network Processors \[in progress\]](#)
- [NFSp: A Non-Intrusive Parallel NFS Server \[in progress\]](#)
- [KadeployFS \[in progress\]](#)
- [Developpement d'une méthode de rejeu de trafic réaliste \[in progress\]](#)

9/01/2007

230+ publications

Grid'5000 publications - Grid5000

https://www.grid5000.fr/mediawiki/index.php/Special:G5KPublications

Amazon France Apple eBay France Yahoo! Informations (1772)

Fcappello my talk preferences my watchlist my contributions log out

search: Go Search

Grid'5000

special page

Grid'5000 publications

Please find below publications related to the Grid'5000 project, sorted by type.
Have also a look at [this linked page](#) for conferences and seminars slides or posters, and for marketing documents.

International publications

ARTICLE

- [A P2P Platform using sandboxing](#) - *F. Hantz and H. Guyennet* (2006)
- [A Parallel Hybrid Genetic Algorithm for Protein Structure Prediction on the Computational Grid](#) - *A-A. Tantar, N. Melab and E-G. Talbi, B. Parent and D. Horvath* (2006)
- [A pragmatic analysis of scheduling environments on new computing platforms](#) - *Lionel Eyraud* (2006)
- [Complexity results for collective communications on heterogeneous platforms](#) - *Olivier Beaumont and Loris Marchal and Yves Robert* (2006)
- [DFT calculations of formation plus migration enthalpies of monovacancy in Nickel: comparison of the local and non-local approaches](#) - *H.H. Megchiche and Simon Pérusin and J.C. Barthelat and C. Mijoule* (2006)
- [Experiences with Hierarchical Request Flow Management for Network-Enabled Server Environments](#) - *Holly Dail and Frédéric Desprez* (2006)
- [Grid computing for parallel bioinspired algorithms](#) - *N. Melab and S. Cahon and E-G. Talbi* (2006)
- [Light P2P platform of computing for DAG](#) - *F. Hantz* (2006)
- [Performance comparison of parallel programming environments for implementing AIAC algorithms](#) - *Bahi, J.M. and Contassot-Vivier, S. and Couturier, R.* (2006)
- [ProActive: an Integrated platform for programming and running applications on Grids and P2P systems](#) - *Denis Caromel and Christian Delbe and Alexandre di Costanzo and Mario Leyton* (2006)
- [Reducing Network Traffic in Unstructured P2P Systems Using Top-K Queries](#) - *Reza Akbarinia and Esther Pacitti and Patrick Valduriez* (2006)
- [Scalability Comparison of Four Host Virtualization Tools](#) - *Benjamin Quetier, Vincent Neri, Franck Cappello* (2006)
- [Scheduling Messages for Data Redistribution: an Experimental Study](#) - *Jeannot, E. and Wagner, F.* (2006)
- [Hipop : Highly Distributed Platform of Computing](#) - *F. Hantz and H. Guyennet* (2005)
- [How to bring together fault tolerance and data consistency to enable grid data sharing](#) - *Gabriel Antoniu and Jean-François Deverge and Sébastien Monnet* (2005)

9/01/2007

A series of events

The screenshot shows a web browser window displaying the Grid5000 News website. The browser's address bar shows the URL <https://www.grid5000.fr/mediawiki/index.php/Grid5000:News>. The page content is organized into several sections:

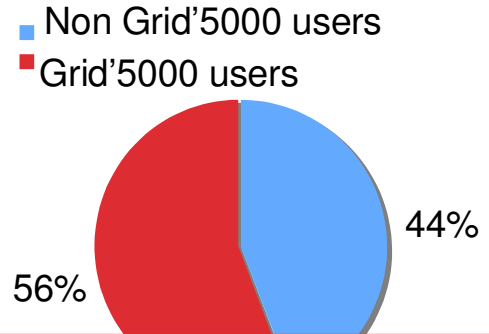
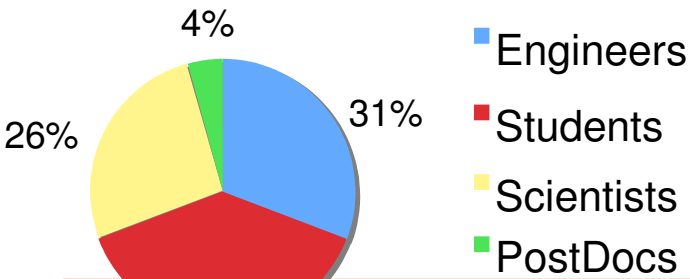
- users portal**: A sidebar menu with links to Community Home, Users Charter, Users Reports, Platform events, Platform status, Bugzilla / Support, Users Docs, and FAQ.
- committees portal**: A sidebar menu with links to Todo, Meetings, Members, Admins Docs, and Reference Docs.
- wiki special pages**: A sidebar menu with links to Recent changes, Wanted pages, Upload files, All pages, and Wiki help.
- toolbox**: A sidebar menu with links to What links here, Related changes, Upload file, Special pages, and Printable version.

The main content area lists events in chronological order from top to bottom:

- March 2007: EGO-2006**: The *Operational Grids winter School* will take place in Rouen (France) in march 2007. Grid'5000 will provide the grid platform and the assistance staff during the event.
- Past events**: A section header for previous events.
- December 2006: Toward Grid'5000 - DAS-3 interconnection**: December 4th 2006: **Grid'5000 - DAS-3 interconnection workshop** took place at Vrije Universiteit, Amsterdam.
- November 2006: PlugTest 2006**: GRIDS@work: CoreGRID Conference, Grid Plugtests and Contest took place at ETSI Headquarters, Sophia-Antipolis from November, 27th to December, 1st.
- November 2006: Paristic 2006**: The 2006's edition of Parsitic was hosted at **LORIA** in Nancy, from November 22th to 24th.
- November 2006: SC'06**: SuperComputing 2006 took place in Tampa, Florida (USA) from November 11th to 17th.
- October 2006: Journées Grid'5000 à Lille**: October 30th and 31th, Grid'5000 Lille site's days will take place in USTL campus, Villeneuve d'Ascq. Project/platform presentations and tutorials will be proposed.
- October 2006: Grid eXplorer days**: October 12th and 13th, IDRIS, Orsay. Grid eXplorer users will present their latest research results obtained with the Grid eXplorer platform.
- August-September 2006: Grid@5000@Titech**: After Franck Cappello's visit to Tokyo Institute of Technology, Grid'5000 was presented during the **I-Explosion** project workshop.
- June 2006: Grid'5000 Workshop**: The event will take place in Paris (France) in conjunction with HPDC'2006.

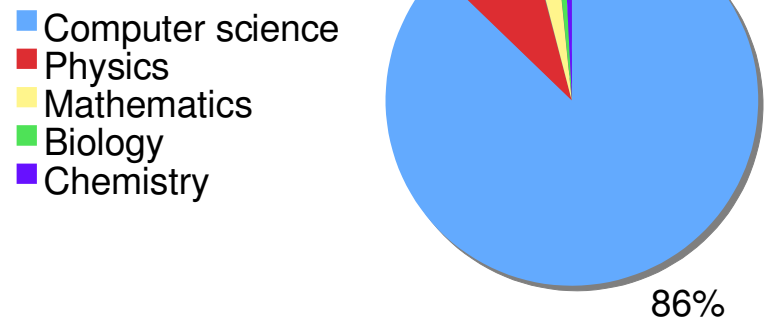
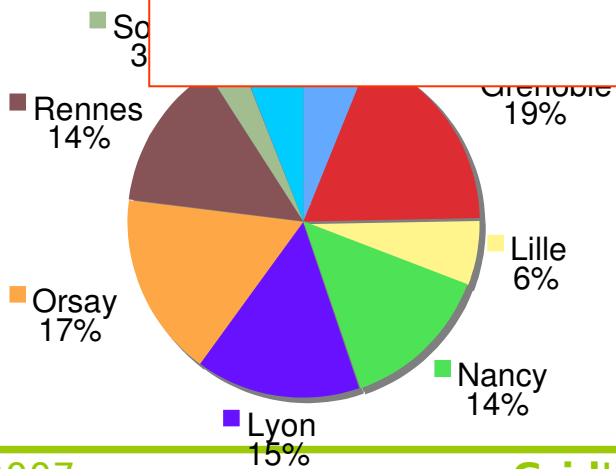
The browser's status bar at the bottom shows the website address, a small icon, the number 0.203, and weather information: Now: Clear, 3° C, Mon: 7° C, Tue: 11° C.

117 participants (we tried to limit to 100)



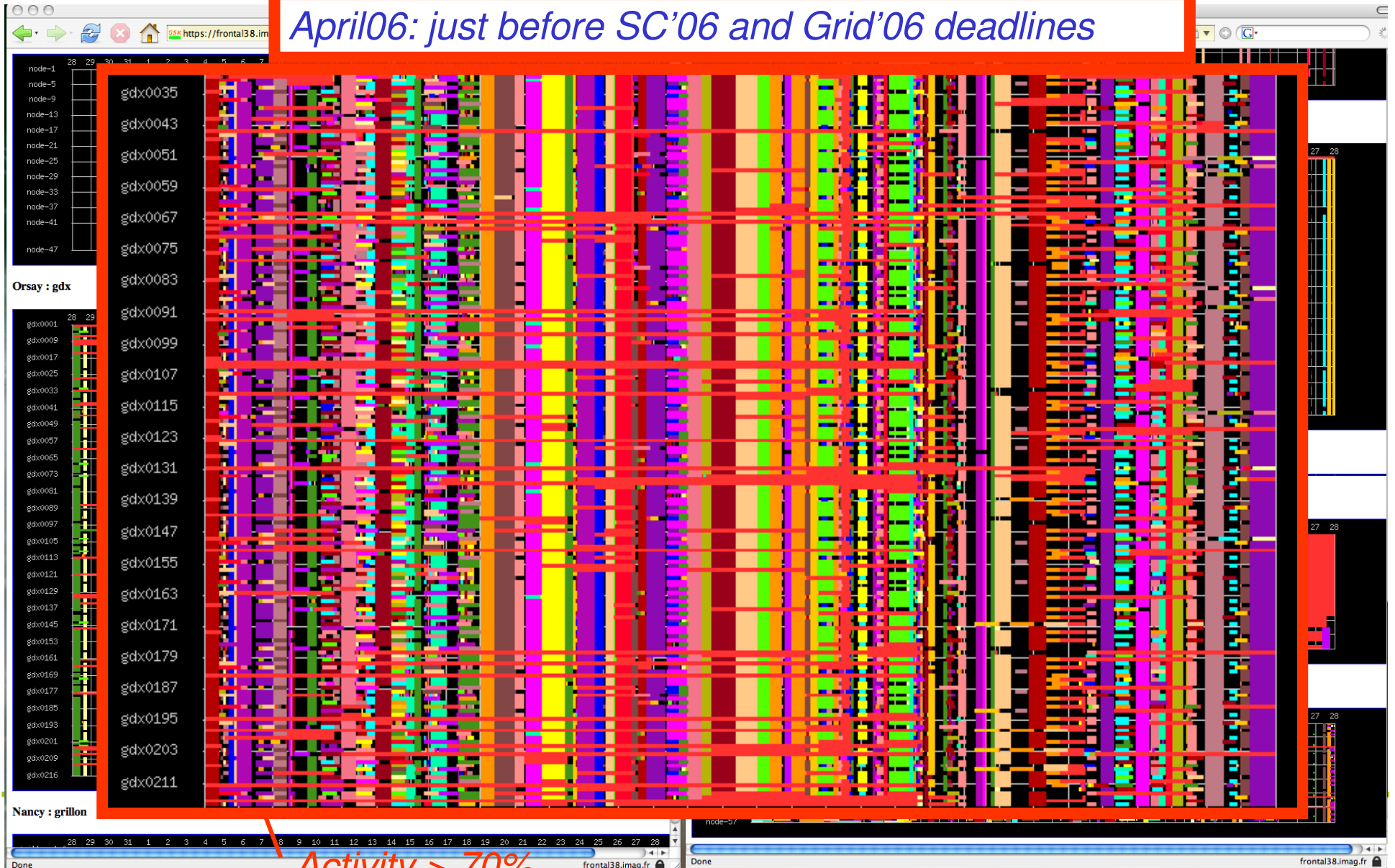
*Don't miss the Second
Grid'5000 Winter School in March 2007*

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



Resources usage

April 06: just before SC'06 and Grid'06 deadlines

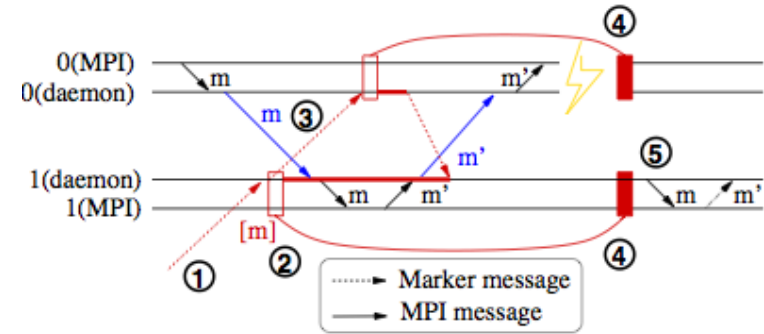




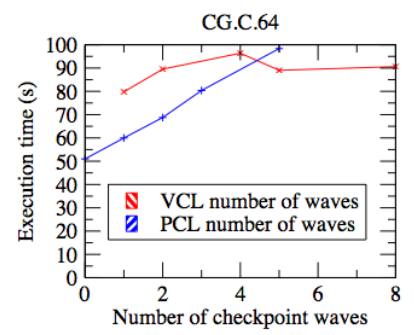
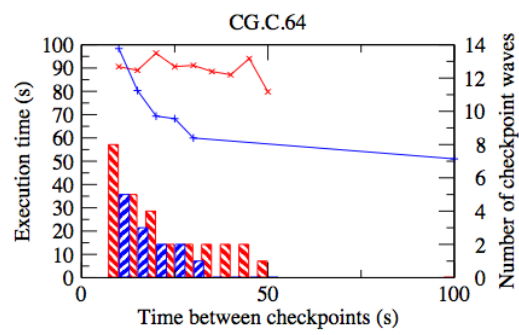
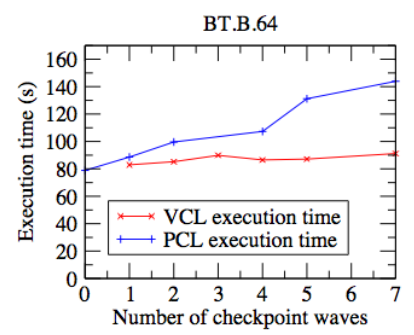
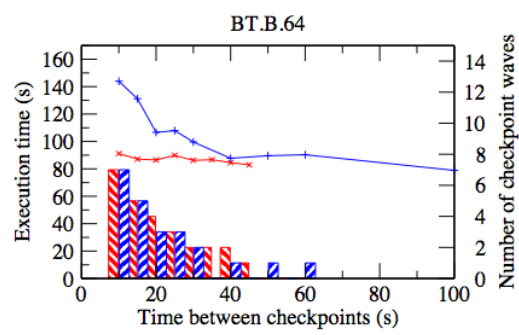
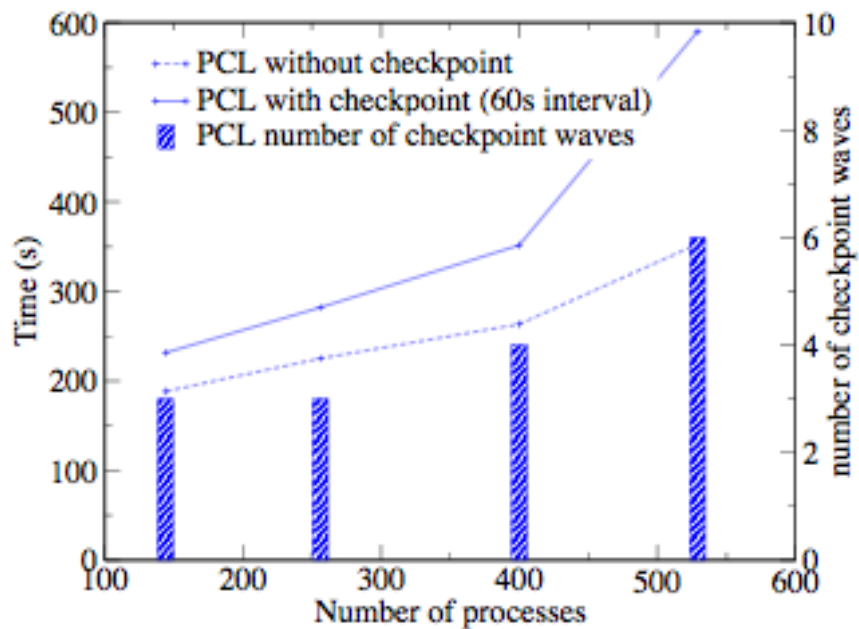
Experiments

Fault tolerant MPI for the Grid

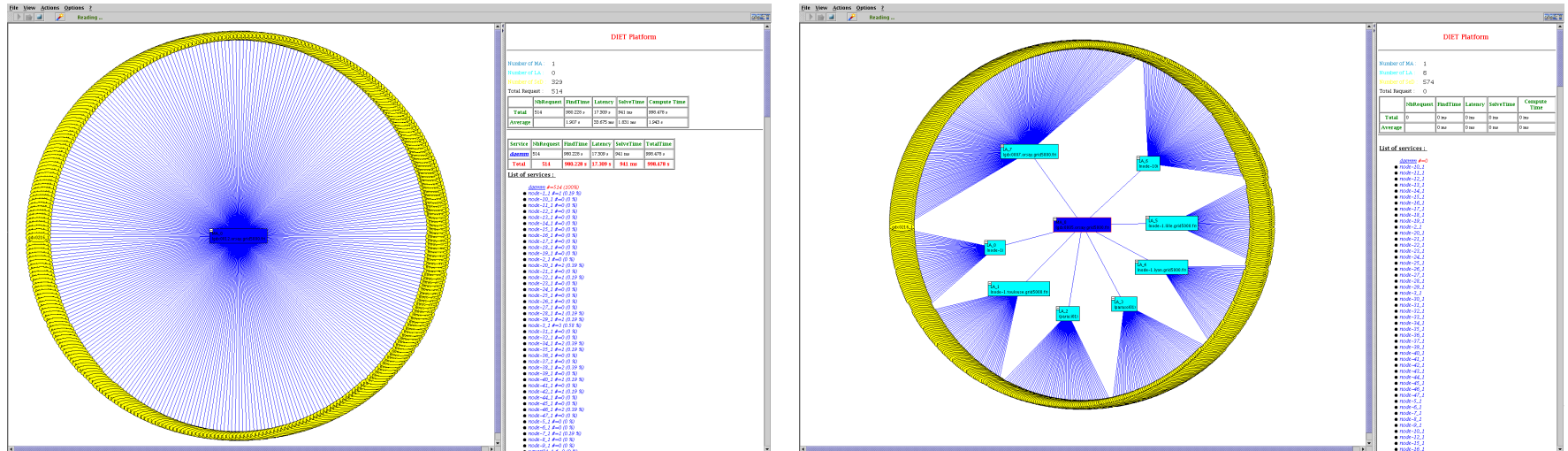
- **MPICH-V**: Fault tolerant MPI implementation
- Research context: large scale fault tolerance
- Research issue: Blocking or non Blocking
Coordinated Checkpointing?
- Experiments on 6 sites up to 536 CPUs



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 :

- Prove that the DIET environment is scalable.
- Test the functionalities of DIET at large scale

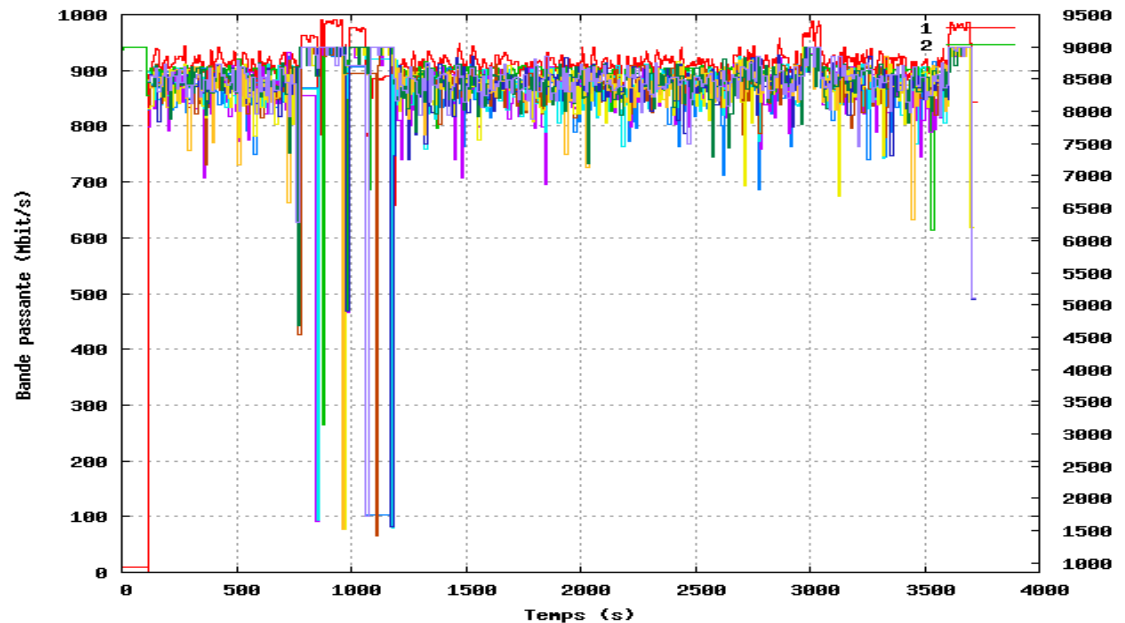


Raphaël Bolze

TCP limits over 10Gb/s links

- Highlighting TCP stream interaction issues in very high bandwidth links (congestion collapse) 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.



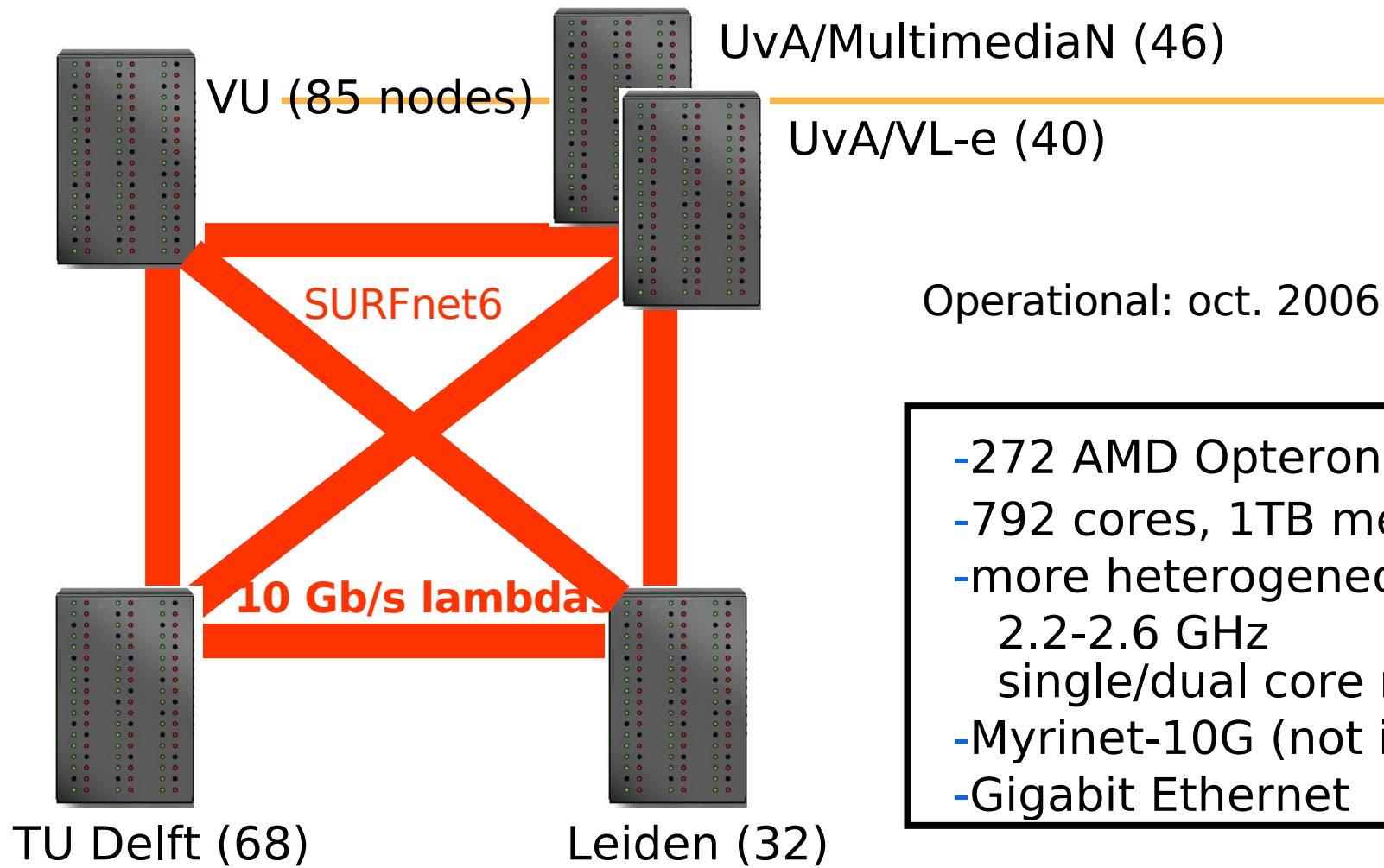
Next Step

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*”

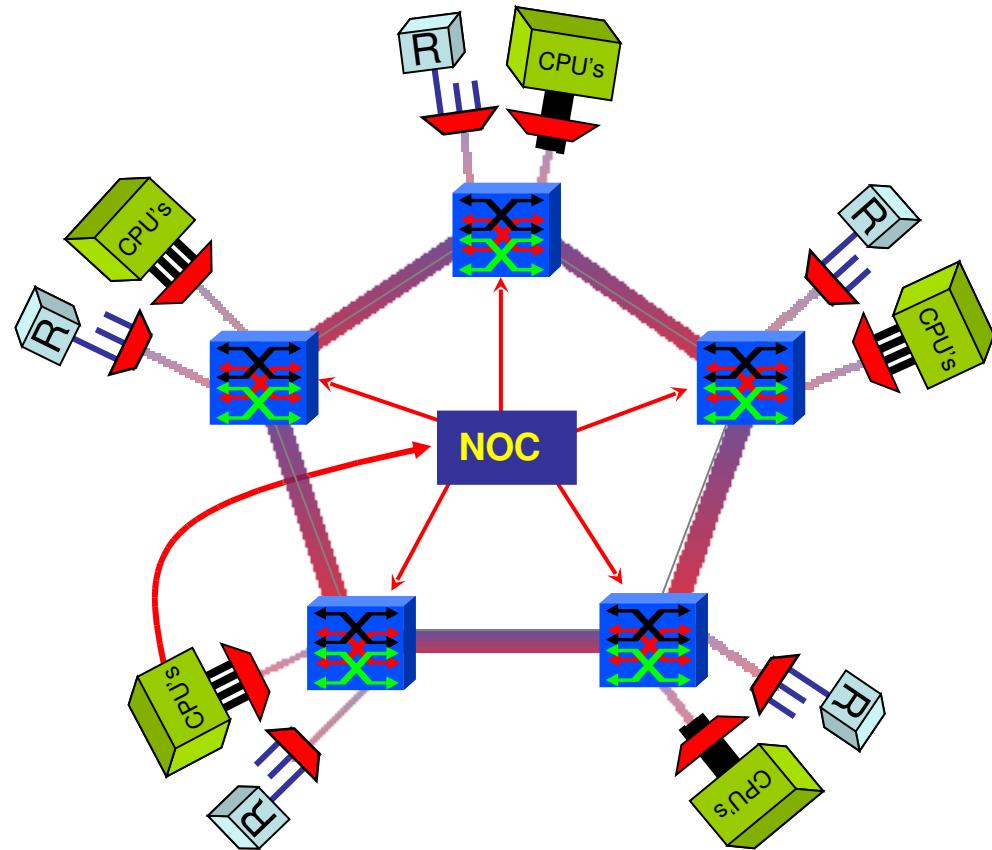
DAS-3: overall structure



- 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

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 vraiment 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.*

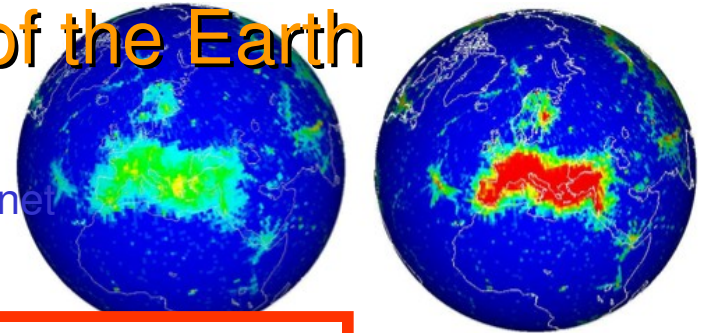


Currently we use Reboot, but Xen will be used in
the default environment.

Let users select its experimental environment:

Fully dedicated or shared within virtual machine

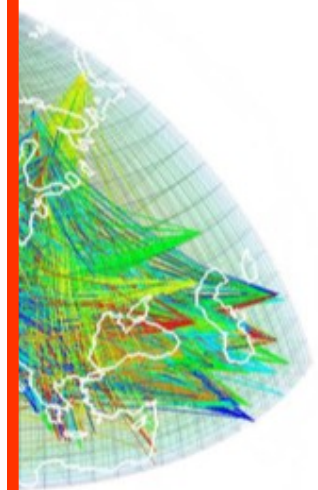
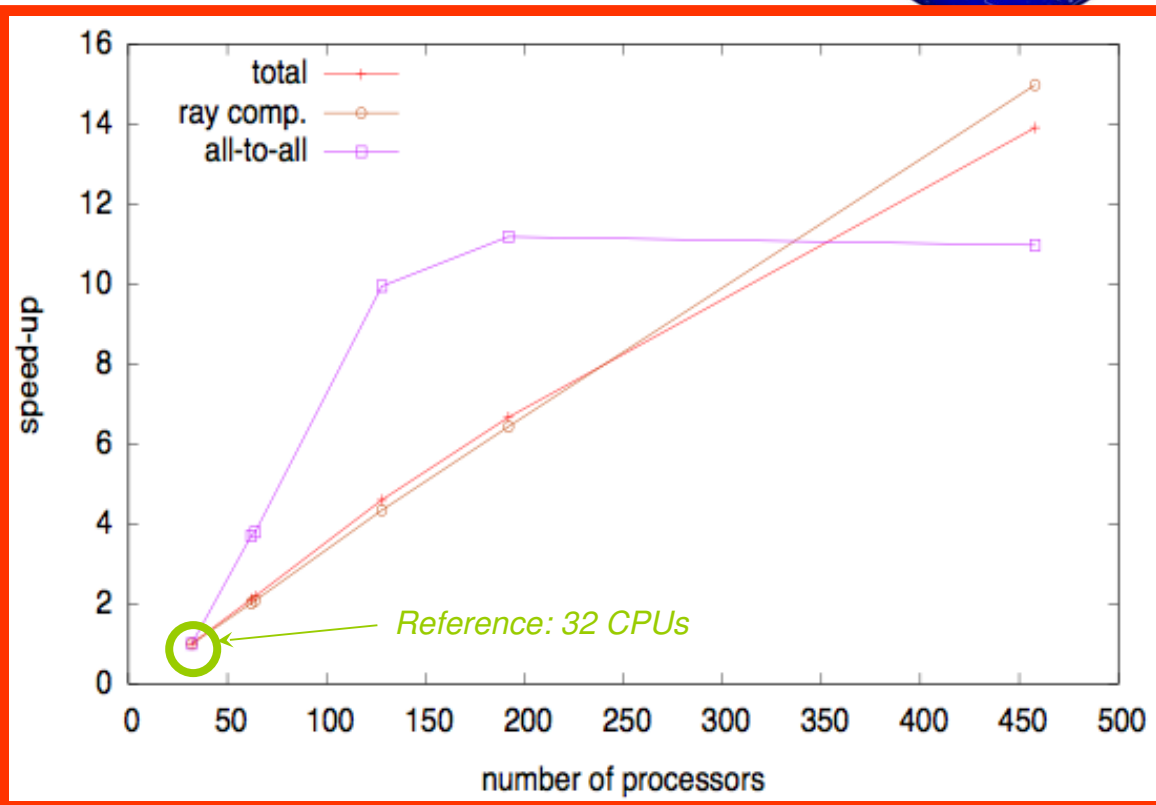
Experiment: Geophysics: Seismic Ray Tracing in 3D mesh of the Earth



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

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

- A MPI parallel progra
- 1) Master-work
 - 2) all-to all com
 - 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)

```
number of jobs, number of machines, initial seed, upper bound and lower bound :
50 20 1539989115 3875 3480
processing times :
52 95 42 75 44 57 89 53 84 62 91 14 95 89 4 95 2 97 60 20 33 51 98 8 85 86 73 4
63 99 69 70 53 21 10 31 80 18 5 18 17 71 90 93 14 49 52 7 78 57 41 75 98 93 33 7
82 21 79 95 46 23 40 95 87 37 24 24 65 62 19 67 66 6 65 59 2 67 82 90 30 63 5 9
16 26 46 66 76 31 26 8 37 21 3 76 67 5 47 72 66 56 95 49 47 26 81 56 76 66 26 5
63 55 59 35 21 59 78 25 30 38 78 79 58 44 38 76 70 72 85 8 10 84 42 67 20 24 75 2
94 34 89 62 47 66 76 15 18 54 24 55 96 10 12 96 53 92 77 6 91 14 41 30 85 17 23 6
79 21 93 32 8 45 37 78 26 98 17 25 21 28 68 24 62 89 60 64 38 90 87 1 99 34 9 2
22 6 24 55 48 57 78 5 50 83 70 21 71 58 36 50 31 86 29 30 93 49 63 89 44 38 62 4
80 13 64 77 17 78 82 4 72 93 68 25 67 80 43 93 21 33 14 30 59 83 85 85 70 35 2 7
96 3 50 57 66 84 98 55 70 32 31 64 11 9 32 58 98 95 25 4 45 60 87 31 1 96 22 9
53 19 99 62 88 93 34 72 42 65 39 79 9 26 72 29 36 48 57 95 93 79 88 77 94 39 74 4
54 67 25 77 38 98 96 20 15 36 65 97 27 25 61 24 97 61 75 92 73 21 29 3 96 51 26 4
71 90 59 82 22 88 35 49 78 69 76 2 14 3 22 26 44 1 4 16 55 43 87 35 76 98 78 8
27 93 49 63 65 34 10 56 51 97 52 46 16 50 96 85 61 76 30 90 42 88 37 43 88 91 14 6
95 53 54 22 84 54 2 80 84 66 25 16 79 90 51 29 29 90 83 83 19 95 87 12 34 23 44 3
3 80 78 32 53 43 85 19 48 49 66 22 37 51 82 59 88 77 19 32 52 9 96 23 64 22 37
92 62 11 83 87 66 98 42 23 45 52 6 3 64 55 97 83 42 81 92 68 46 56 88 50 13 23 1
80 38 55 34 85 44 47 66 19 66 61 60 98 82 79 71 28 74 27 33 13 9 12 51 16 49 83 4
51 98 89 42 14 82 87 35 22 78 8 35 95 53 59 34 66 42 63 27 92 8 65 34 6 42 3
number of jobs, number of machines, initial seed, upper bound and lower bound :
50 20 691823909 3715 3424
processing times :
47 71 46 50 52 29 14 44 8 71 8 65 34 35 42 28 28 5 80 46 37 86 85 67 22 6
46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46
70 76 16 22 44 83 66 36 27 22 16 83 70 25 28 26 98 64 66 15 39 75 66 89 56 42 90 7
23 21 89 51 34 62 6 22 75 22 5 37 72 6 90 71 47 60 93 17 65 12 97 29 41 3 46 3
53 66 41 79 58 57 17 66 42 66 16 19 42 80 9 48 62 65 70 45 3 98 11 2 64 77 85 1
61 18 82 96 68 7 65 95 30 84 23 75 71 60 21 76 37 65 78 37 71 60 59 18 94 55 53 5
77 74 69 8 84 81 37 26 10 60 72 89 49 47 90 6 87 10 34 52 72 77 94 44 80 88 9
25 96 32 37 19 42 86 49 55 92 85 94 22 30 93 83 95 1 16 75 43 80 42 78 9 56 35 9
54 71 46 50 52 29 14 44 8 71 8 65 34 35 42 28 28 5 80 46 37 86 85 67 22 6
54 1 54 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
72 29 76 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88
27 4 35 86 50 65 17 11 59 22 76 38 35 18 82 98 83 75 39 58 32 84 1 41 20 7
65 89 18 14 60 78 56 3 79 15 48 9 72 41 83 27 73 80 16 34 42 65 57 88 78 62 19 6
```

Using simultaneously Grid5000 and other clusters

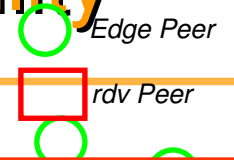
Involved Grid5000 sites (6): Bordeaux, Lille, Orsay, Rennes, Sophia-Antipolis and Toulouse

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

ASOV/Grid meeting

Talbi, N. Melab 2006

Jxta DHT scalability



- Goals: study of a JXTA “DHT”

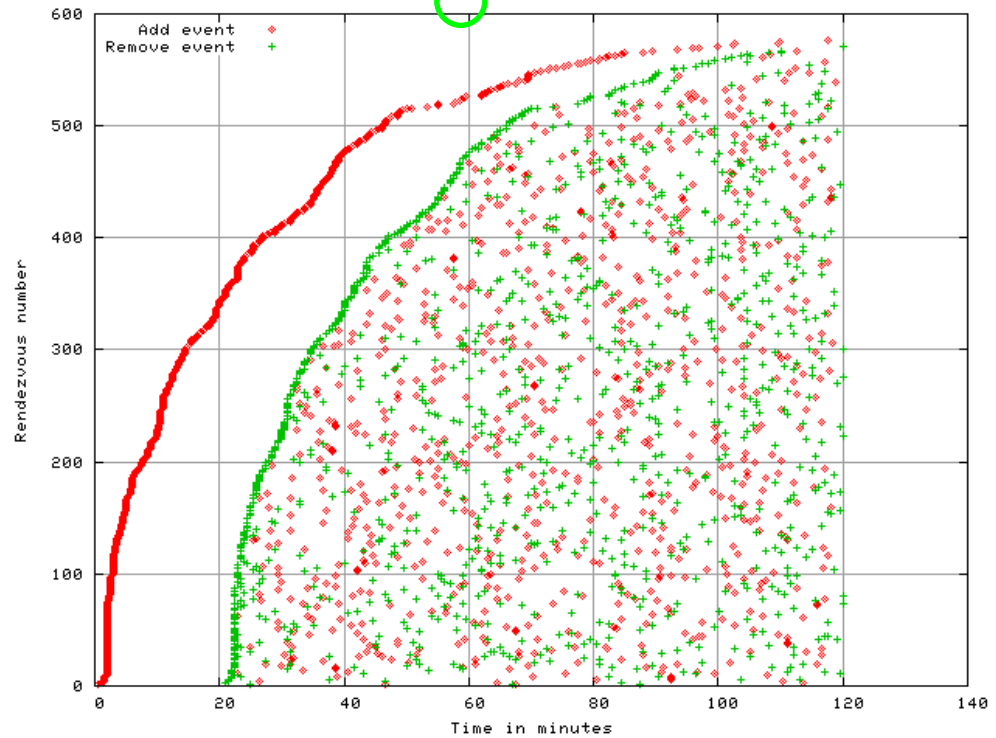
- It requires 2 hours to contact all “rendez vous” peers
- With the per default setting, the view of every rendez vous peers is limited to only 300 rendez vous peers
- The view of every “rendez vous” peer is very unstable

- Organization of a JXTA overlay (peer view 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

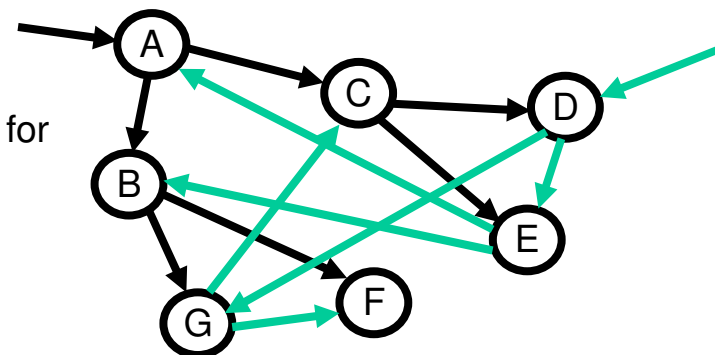


“rendez vous” peers known by one of the “rendez vous” peer
X axis: time ; Y axis: “rendez vous” peer ID



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.

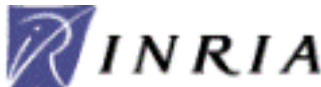


The **N-Queens Contest** (4 teams) where the aim was to find the number of solutions to the N-queens problem, N being as big as possible, in a limited amount of time

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

9/01/2007



Charter & Accounting

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