Toward an International Computer Science Grid

DAS3 Workshop, Delft

Franck Cappello
INRIA
Director of Grid'5000

Email: fci@lri.fr

http://www.lri.fr/~fci

6/6/2007
Agenda

Motivations

Grid’5000

DAS3

International Computer Science Grid
Large Scale Distributed Systems raise many research challenges

LSDS (Grid, P2P, etc.) are complex systems:
- Large scale, dynamic distributed systems
- Deep stack of complicated software

Main Large Scale Distributed Systems research issues:
- Security, Performance, Fault tolerance, Scalability, Load Balancing,
- Coordination, deployment, accounting, Data storage, Programming,
- Communication protocols, Accounting, etc.
Research in Large Scale Distributed systems raise methodological challenges

How designers of applications, application runtimes, middleware, system, networking protocols, etc.

Can test and compare
- Fault tolerance protocols
- Security mechanisms
- Networking protocols
- etc.

Following a strict scientific procedure,

Knowing that all these components run simultaneously during the execution in a complex and dynamic environment?
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

2) Strong interaction among these research tools

- Abstraction
  - math
  - simulation
  - emulation
  - live systems

validation
Existing Grid Research Tools

- **SimGrid3**
  - Discrete event simulation with trace injection
  - Originally dedicated to scheduling studies
  - Single user, multiple servers

- **GridSim**
  - Dedicated to scheduling (with deadline, budget), DES (Java)
  - Multi-clients, Multi-brokers, Multi-servers
  - Support for SLAs, DataGrids, Service Model

- **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)

--> Simulators and Emulators are quite slow and do not scale well
What about production platforms used for experimental purpose?

Not reconfigurable:
• Many projects require experiments on OS and networks,
• Some projects require the installation of specific hardware

No reproducible experimental conditions:
• Scientific studies require reproducible experimental conditions

Not designed for experiments:
• Many researchers run short length, highly parallel & distributed algos
• Preparation and execution of experiments are highly interactive

Not optimized for experiments
• Experimental platforms should exhibit a low utilization rate to allow researchers executing large collocated experiments

→ Nowhere to test networking/OS/middleware ideas, to measure real application performance,
We need experimental tools

In 2002 (DAS2) and 2003 (Grid’5000), the design and development of an experimental platform for Grid researchers was initiated:

- **Model** proof
- **Simulation**
- **Emulation**
- **Live systems**

**log(cost & coordination)**

**Major challenge**

**Challenging**

**Reasonable**

- **Data Grid eXplorer**
- **WANinLab**
- **Emulab**

**Grid’5000**

- **DAS3**
- **PlanetLab**
- **GENI**

**RAMP**

- Dave Patterson’s Project on Municore
- Multi-processor emulator

6/6/2007
Computer Science Grids

Grid’5000 and DAS:
- Designed by computer scientists
- for computer scientists

- Not production platforms for Physics or Biology
- Production platforms for Computer Science

-Large Scale Scientific Instruments:
  - Researchers share experimental resources,
  - Reserve resources,
  - Configure resources,
  - Run experiments,
  - Conduct precise measurements,
  - Reproduce experimental conditions

-More than testbeds:
  - Researchers share experiences, results, skills
  - access an environment with supporting engineers
Is this really Original, New?

In Parallel Computing and HPC --> Not really

- Most of the evaluations for new methods, algorithms, optimizations in parallel computing and HPC are conducted on REAL computers NOT simulators
- Why: because Parallel Computers and HPC machines are easy to access and it’s easy to build and run a test, users need trustable results!

In Grid and Large Scale Distributed systems --> YES

- It’s difficult to get access to a Grid and a Large Scale Distributed Systems
- Simulators are easy to build
- Results are rarely confronted to reality

→ Computer Science Grid and Experimental Facilities for Large Scale Distributed systems should change the situation!
Agenda

Motivations

**Grid’5000**
  Design
  Status
  Results

DAS3

An International CSG
Grid’5000

One of the 40+ ACI Grid projects
ACI GRID projects

- 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)
1) Building a nationwide experimental platform for Large scale Grid & P2P experiments
   - 9 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 testing and repeating 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 novel mechanisms
     - P2P resources discovery, Desktop Grids
Grid’5000 foundations: Collection of experiments to be done

- **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**
  - Scheduling / data distribution in Grid
  - Fault tolerance in Grid
  - Resource management
  - 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

**Grid’5000 foundations:**
Measurements and condition injection

Quantitative metrics:
- **Performance:** Execution time, throughput, overhead, QoS (Batch, interactive, soft real time, real time).
- **Scalability:** Resource load (CPU, memory, disk, 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 running their preferred measurement tools and experimental condition injectors
**Grid’5000 principle:**

A highly reconfigurable experimental platform

<table>
<thead>
<tr>
<th>Experimental conditions injector</th>
<th>Application</th>
<th>Programming Environments</th>
<th>Application Runtime</th>
<th>Grid or P2P Middleware</th>
<th>Operating System</th>
<th>Networking</th>
<th>Measurement tools</th>
</tr>
</thead>
</table>

Let users create, deploy and run their software stack, including the software to test and evaluate using measurement tools + experimental conditions injectors.
Experiment workflow

1. Reserve 1 node
2. Reboot node (existing env.*)
3. Adapt env.
4. Reboot node
5. Env. OK?
   - yes: Go to step 8
   - no: Go to step 6

6. Reserve nodes corresponding to the experiment

7. Reboot the nodes in the user experimental environment (optional)

8. Log into Grid’5000
9. Import data/codes
10. Build an env.? (yes/no)
11. Transfer params + Run the experiment
12. Collect experiment results
13. Exit Grid’5000

*Available on all sites: Fedora4all, Ubuntu4all, Debian4all
Hardware Configuration

QuickTime™ et un décompresseur TIFF (LZW) sont requis pour visionner cette image.
Grid’5000 network

Renater connections

10 Gbps

Dark fiber
Dedicated Lambda
Fully isolated traffic!
Grid’5000 Team

ACI Grid Director + ACI Grid Scientific Committee President

At National Level:
1 Director (D)
1 Technical Director (TD)
1 Steering committee
1 Technical committee

Per Site:
1 Principal Investigator (PI)
(site scientific,
Administrative and
financial
Manager)
1 Technical Advisor
1 Engineer (about) (E)
Grid’5000 as an Instrument

4 main features:

• A high security for Grid’5000 and the Internet, despite the deep reconfiguration feature
  --> Grid’5000 is confined: communications between sites are isolated from the Internet and Vice versa (Dedicated lambda).

• A software infrastructure allowing users to access Grid’5000 from any Grid’5000 site and have a simple view of the system
  --> A user has a single account on Grid’5000, Grid’5000 is seen as a cluster of clusters

• A reservation/scheduling tools allowing users to select nodes and schedule experiments
  → a reservation engine + batch scheduler (1 per site) + OAR Grid (a co-reservation scheduling system)

• A user toolkit to reconfigure the nodes
  --> software image deployment and node reconfiguration tool
Experimental Condition injectors

Network traffic generator

Main TCP applications throughputs (Renater)

\[ \Delta = 10 \text{ms} \]

\[ \Delta = 32 \text{ms} \]

\[ \Delta = 400 \text{ms} \]

A non Gaussian long memory model

Gamma + Farima:

\[ \Gamma_{\alpha,\beta} - \text{farima}\left(\phi, d, \theta\right) \]

Normal \[ \Delta = 10 \text{ms} \]

DOS attack \[ \Delta = 32 \text{ms} \]

Flash Crowd \[ \Delta = 2 \text{ms} \]

6/6/2007

DAS3 Workshop, Delft
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 tomography model of the Earth geology using seismic wave propagation characteristics in the Earth.

Seismic waves are modeled from events detected by sensors. Ray tracing algorithm: waves are reconstructed from rays traced between sensors.

A MPI parallel program composed of 3 steps:
1) Master-worker: ray tracing and mesh update by each process with blocks of rays successively fetched from the master process,
2) all-to all communications to exchange submesh information between the processes,
3) merging of cell information of the submesh associated with each process.

Reference: 32 CPUs
Solving the Flow-Shop Scheduling Problem

“one of the hardest challenge problems in combinatorial optimization”

Solving large instances of combinatorial optimization problems using a parallel Branch and Bound algorithm

Flow-shop:
• Schedule a set of jobs on a set of machines minimizing makespan.
• Exhaustive enumeration of all combinations would take several years.
• The challenge is thus to reduce the number of explored solutions.


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

Involved Grid5000 sites (6): Bordeaux, Lille, Orsay, Rennes, Sophia-Antipolis and Toulouse. The optimal solution required a wall-clock time of 25 days.
Jxta DHT scalability

- Goals: study of a JXTA “DHT”
  - “Rendez vous” peers form the JXTA DHT
  - Performance of this DHT?
  - Scalability of this DHT?

- 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

G. Antoniu, M. Jan, 2006
TCP limits on 10Gb/s links

[Guillier - Primet - HSN workshop of Infocom2007 - INRIA RESO]

- In Grids, TCP is used by most applications & libraries (MPI, GridFTP, ...)
- Long distance High speed networks are challenging for TCP & transport protocols
  Designing new schemes is not straightforward (fairness, stability, friendliness)
  New variants are proposed/implemented in Linux (Bic) & Windows (Compound)
  Experimental investigations in real high speed networks are highly required

Without reverse traffic
- Around 470Mb/s per flow
- 9200 Mb/s global

With reverse traffic
- Less efficiency
- More instability
- Around 400Mb/s per flow
- 8200 Mb/s global

6/6/2007 19 flows forward /19 flows reverse Affects flow performance & global throughput by 20%, global instability
Community: Grid’5000 users

~300 active Users
Coming from 60 Laboratories.

2DRMP .uk
AIST .jp
BRGM
CEDRAT
CERFACS
CDS
CICT
CIRIMAT
CS-VU.nl
EADS CCR
EC-Lyon
eDAM
ENS-Lyon
ENSEEIHT
ENSIACET
EPFL.ch
FEW-VU.nl
France-telecom

HPCS .jp
IBCP
ICPS-Strasbourg
IDRIS
IECN
IMAG
IME/USP.br
IN2P3
INF/UFRRGS.br
INP-Toulouse
INRIA
INSAS-Lyon
INT
IRISA
IRIT
LAAS
LABI
LABR
LBP
LEN7
LGCT
LICA
LIFC
LIFL
LIFO
LINA
LIP6
LIPN-Paris XIII
LIRIS
LMA
LORIA
LPC
LSR
LSR

Prism-Versailles
RENATER
SUPELEC
UCD.ie
UFRJ.br
UHP-Nancy
UTK .edu

Hamza Adamou (MSc), MESCAL, ID-MAG Grenoble
Guillaume ALLIEN (Engineer), EADS CAR
Lamine Aouad (PhD student), Grand Lare LIFL Lille
Carlos Jaime BARRIOS HERNÁNDEZ (PhD student), MESCAL ID-MAG Montbonnot Saint-Martin (Grenoble-France)
Janet Bernt (Ing devepx), service Dream INRIA Sophia
Raphaël Bolze (PhD student), GRAAL LIP-ENS Lyon
Hinde Lilia Bouziane (PhD student), PARIS IRIS/INRIA Rennes
Jeremy BUISSON (PhD student), PARIS IRIS/INRIA Rennes
CHRISTOPHE CERIN (prof associé), LIR PP Paris XIII, Villetaine, Strasbourg
Arnaud Contes (PhD), OASE
Cédric Dalmaso (Interim Prof), CNRS Sophia Antipolis
Alexandre di Costanzo (PhD), MESCAL ID-MAG INRIA Sophia Antipolis
Fabrice Dupros (engineer), ISEP, Université d'Orléans
Thierry Gautier (Prof), LIPN, Paris XIII
Stéphane Genaud (Maitre des Recherches), TAG ICPS-LSIT Strasbourg
Yiannis Georgiou, Mescal ID-MAG Grenoble
Olivier GLUCK (Associado Professor), EAFES POIC LIP-ENS-Lyon
Jens Gustedt (directeur de recherche, INRIA Lorraine & LORIA Nancy, France)
Christoph Hamerling (Eng Devel), LIRIS-IRIT-ENSEEIHT Toulouse
Thomas Héraut (Assistant Professor), ENS-Lyon
Samer Jafar (PhD student), MESCAL ID-MAG Grenoble
Emmanuel Jeannot (Chargé de recherche), EAFES POIC LIP-ENS-Lyon
Emmanuel Jeannot (PhD), LORIS Rennes
Paymond Johann (developer), TAG ICPS-LSIT Strasbourg
Nicolas LARRIUE (Postdoc), LIRIS-IRIT-ENSEEIHT Toulouse
Adrien Labre (PhD student), LIP6
Julien Leduc (Research Engineer), EAFES POIC LIP-ENS Lyon
Laurent Lefèvre (INRIA CR, MESCAL ID-MAG Grenoble)
Oleg Lodygorsky (PhD student), LIP6
Eric MAISONNAVE (Engineer), LIRIS-IRIT-ENSEEIHT Toulouse
Maxime Martinasso (PhD student), MESCAL ID-MAG Grenoble
Sébastien Monnet (PhD student), LORIS Rennes
Thierry Monteil (Assistant Professor), EAFES POIC LIP-ENS Lyon
Matthieu Morel (Ingénieur expérimental), INRIA Sophia Antipolis
Grégory Moullaut (Assistant Professor), EAFES POIC LIP-ENS Lyon
Pascal NIVOT (PhD), STM
Frédéric NIVOT (PhD), STM
Yves Lévy-Langevin (Prof),美学 D-MAG Montbonnot (Grenoble-France)
Ludovic Montesano (PhD student), LIP6
About 340 Experiments

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 platforms (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]
- aLOL - I/O Scheduler for High Performance Computing [in progress]
- Study of peer-to-peer systems using emulation [in progress]
- Tamosoire 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 paracell and paracellnet [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]
- NFS: A Non-Intrusive Parallel NFS Server [in progress]
- KadeployFS [in progress]
- Développement d’une méthode de rejet de trafic réaliste [in progress]
Main purpose:
- Evaluate the performance of applications ported to the Grid,
- Test alternatives,
- Design new algorithms and new methods

Application domains:
- Life science (mammogram comparison, protein sequencing, Gene prediction, virtual screening, conformation sampling, etc.)
- Physics (seismic imaging, parallel solvers, hydrogeology, Self-propelled solids, seismic tomography, etc.)
- Applied Mathematics (sparse matrix computation, combinatorial Optimization, parallel model checkers, PDE problem solvers, etc.)
- Chemistry (molecular simulation, estimation of thickness on Thin films),
- Industrial processes,
- Financial computing
300 Publications

Including:
- 5 HDR
- 14 Ph.D.
- 200 international publications
Resource usage: activity

March’07

Activity > 70%
Reconfiguration stats.
Agenda

Motivations

Grid’5000

**DAS3**
  Design
  Status
  Results

A European CSG
DAS-2 (2002)

- Geographically distributed cluster-based system
  - 5 clusters with 200 nodes in total
- Homogeneous system:
  - same processors, network, OS
  - Eases collaboration, software exchange, systems management, experimentation
- Funded mainly by Dutch national science foundation (NWO)
DAS-1/2 Results & lessons

- 200 users in total
- Used for 32 Ph.D. theses
- Used for many publications, including 11 in ACM/IEEE journals/transactions and 1 in Nature
- Used to solve Awari: 3500-year old game
- Homogeneous structure eases software exchange and system administration
  - *DAS-2 is a grid that works!*
- Ideal for experimental research: “laboratory-like grid”
  - Measure performance of parallel applications
- Allows new types of Grid research
  - Distributed supercomputing (languages, scheduling)
  - High-level interactive steering/visualization from VR
  - Virtual laboratories
DAS-3 “going optical” (2006)

- Proposed next generation grid in the Netherlands
  - 4 sites, 5 clusters connected by optical network (SURFnet-6)
  - Interconnected with 4 to 8 dedicated lambdas of 10 Gb/s each
  - Separate band on academic internet backbone
DAS-3 Challenges

- Application can *dynamically allocate* light paths, of 10 Gbit/sec each
- Application can *control topology* through the Network Operations Center (subsecond lambda switching times)
- Gives flexible, dynamic, high-bandwidth links
- Research questions (proposed *StarPlane* project)
  - How to provide this flexibility (across domains)?
  - How to integrate optical networks with applications?
Projects Similar to DAS3

Optical networks for grids

- **G-lambda (Japan)**

- **Enlightened (USA)**
  - views the network as a grid resource, same level as compute and storage resources
  - Grid framework for 1) dynamic application request (computing, storage high bandwidth, secure network) 2) Software tools and protocols (fast network reconfiguration, on-demand or in-advance provisioning of lightpaths)
  - Determine: how to abstract the network resources & how to distribute the network intelligence among [network control plane, management plane, grid middleware]

- **Phosphorus (EU)**
A Prototype of an European Computer Science Grid

Renater-Geant-Surfnet (a dedicated lambda at 10G)

Grid’5000

3000 CPUs

DAS3

2007 (July 27)

3000 CPUs

1500 CPUs

1000 CPUs

Nadeira (Portugal)

Canary Islands (Spain)

Gibraltar Strait

Ibiza, Mallorca, Menorca (Spain)

Corsica, Sardinia (Italy)

Sicily (Italy)

Malta

Crete (Greece)

Iberian Peninsula

Mediterranean Sea

Cyprus

Macedonia

Greece

Bulgaria

Rumelia

Bosnia

Romania

Moldova

Ukraine

Baltic States

Netherlands

Germany

2007

DAS3 Workshop, Delft
Recent long range connection

Connection of Grid’5000 to the NAREGI R&D lab @ Tokyo
OS Reconfiguration in both side (first tests on transport Protocols for large data sets)
Conclusion

• Large scale and highly reconfigurable Grid experimental platforms
• Used by Master’s and Ph.D. students, PostDoc and researchers (and results are presented in their reports, thesis, papers, etc.)

• Grid’5000 and DAS-3 offer together in 2007:
  – 19 clusters distributed over 13 sites in France and The Netherlands,
  – about 10 Gigabit/s (directional) of bandwidth
  – The capability for all users to reconfigure the platform (G5K) [protocols/OS/Middleware/Runtime/Application]
  – The capability for all users to reconfigure the network topology (DAS-3)

• Grid’5000 and DAS results in 2007:
  – 280 users + 200 users
  – 12 Ph. D. + 30 Ph.D.
  – 300 publications + 11 in ACM/IEEE journals/transactions and 1 in Nature
  – 340 planned experiments
  – Tens of developed and experimented software
  – Participation to tens of research grants (ACI, ANR, Strep, etc.)

• Connection of the two platforms is already started

• Towards an international “Computer Science Grid”!

• More than a Test bed, “Computer Science Large Scale Instrument”
Questions?

Acknowledgements

Grid’5000
Michel Cosnard, Thierry Priol and Brigitte Plateau
Steering committee: Michel Dayde, Frédéric Desprez, Emmanuel Jeannot, Yvon Jegou, Stéphane Lanteri, Nouredine Melab, Raymond Namyst, Olivier Richard, Pascale Vicat-Blanc Primet and Dany Vandromme, Pierre Neyron

DAS3
Andy Tanenbaum, Bob Hertzberger, and Henk Sips
Steering group members: Lex Wolters, Dick Epema, Cees de Laat and Frank Seinstra. Kees Verstoep

www.grid5000.fr
### Grid’5000 versus PlanetLab

**Grid’5000**

- Cluster of clusters: V
- Distributed PC: -
- Capability to reproduce experimental conditions: V
- Capability for dedicated usage for precise measurement: V
- Experiments on virtual machines technologies: V
- Precise network monitoring: V
- Planet wide: soon

**PlanetLab**

- Cluster of clusters: -
- Distributed PC: V
- Capability to reproduce experimental conditions: -
- Capability for dedicated usage for precise measurement: -
- Experiments on virtual machines technologies: -
- Precise network monitoring: -
- Planet wide: V

*An open platform for developing, deploying, and accessing planetary-scale services*
• 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:** 1600 CPUs in total: 1200 provided by Grid’5000 + 50 by the other Grids (EGEE, DEISA, NorduGrid) + 350 CPUs on clusters.
# DAS-3: Cluster configurations

<table>
<thead>
<tr>
<th></th>
<th>LU</th>
<th>TUD</th>
<th>UvA-VLe</th>
<th>UvA-MN</th>
<th>VU</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Head</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* storage</td>
<td>10TB</td>
<td>5TB</td>
<td>2TB</td>
<td>2TB</td>
<td>10TB</td>
<td>29TB</td>
</tr>
<tr>
<td>* CPU</td>
<td>2x2.4GHz DC</td>
<td>2x2.4GHz DC</td>
<td>2x2.2GHz DC</td>
<td>2x2.2GHz DC</td>
<td>2x2.4GHz DC</td>
<td></td>
</tr>
<tr>
<td>* memory</td>
<td>16GB</td>
<td>16GB</td>
<td>8GB</td>
<td>16GB</td>
<td>8GB</td>
<td>64GB</td>
</tr>
<tr>
<td>* Myri 10G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>* 10GE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Compute</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* storage</td>
<td>400GB</td>
<td>250GB</td>
<td>250GB</td>
<td>2x250GB</td>
<td>250GB</td>
<td>84 TB</td>
</tr>
<tr>
<td>* CPU</td>
<td>2x2.6GHz DC</td>
<td>2x2.4GHz DC</td>
<td>2x2.2GHz DC</td>
<td>2x2.4GHz DC</td>
<td>2x2.4GHz DC</td>
<td>1.9 THz</td>
</tr>
<tr>
<td>* memory</td>
<td>4GB</td>
<td>4GB</td>
<td>4GB</td>
<td>4GB</td>
<td>4GB</td>
<td>1048 GB</td>
</tr>
<tr>
<td>* Myri 10G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Myrinet</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* 10G ports</td>
<td>33 (7)</td>
<td>41</td>
<td>47</td>
<td>86 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>* 10GE ports</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>320 Gb/s</td>
<td></td>
</tr>
<tr>
<td><strong>Nortel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* 1GE ports</td>
<td>32 (16)</td>
<td>136 (8)</td>
<td>40 (8)</td>
<td>46 (2)</td>
<td>85 (11)</td>
<td>339 Gb/s</td>
</tr>
<tr>
<td>* 10GE ports</td>
<td>1 (1)</td>
<td>9 (3)</td>
<td>2</td>
<td>2</td>
<td>1 (1)</td>
<td></td>
</tr>
</tbody>
</table>
## DAS Results

<table>
<thead>
<tr>
<th></th>
<th>DAS-1</th>
<th>DAS-2</th>
<th>DAS-3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sites</strong></td>
<td>4</td>
<td>5</td>
<td>4 sites, 5 clusters</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>200 MHz Pentium Pro</td>
<td>1 GHz Pentium3</td>
<td>2.2+ GHz Opteron</td>
</tr>
<tr>
<td><strong>Networking</strong></td>
<td>Myrinet → BSD</td>
<td>Myrinet</td>
<td>Myrinet + WAN</td>
</tr>
<tr>
<td><strong>Operating System</strong></td>
<td>Linux</td>
<td>Red Hat Linux</td>
<td>Uniform</td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>6 Mb/s ATM (full mesh)</td>
<td>1 Gb/s SURFnet routed</td>
<td>8×10 Gb/s dedicated</td>
</tr>
</tbody>
</table>

- 200 users in total
- Used for over 32 Ph.D. theses
- Used for many publications, including 11 in ACM/IEEE journals/transactions and 1 in Nature
- Used to solve Awari: 3500-year old game
OS Reconfiguration techniques
Reboot OR 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 vtx
- Xen2 does not support x86/6
- 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
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.
TCP limits on 10Gb/s links

[Guillier - Primet - HSN workshop of Infocom2007]

- In Grids, TCP is used by most applications & libraries (MPI, GridFTP....)
- Long distance High speed networks are challenging for TCP & transport protocols
- Designing new schemes is not straightforward (fairness, stability, friendliness)
- New variants are proposed/implemented in Linux (Bic) & Windows (Compound)
- Experimental investigations in real high speed networks are highly required

Without reverse traffic:
Efficiency, Fairness & equilibrium
Around 470Mb/s per flow
9200 Mb/s global

With reverse traffic:
No efficiency, Fairness & instability
Around 400Mb/s per flow
8200 Mb/s global
**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.

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

**Result**:

<table>
<thead>
<tr>
<th>Submission interval</th>
<th>5s</th>
<th>10s</th>
<th>20s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average execution time</td>
<td>2057s</td>
<td>2042s</td>
<td>2039s</td>
</tr>
<tr>
<td>Overhead</td>
<td>4.80%</td>
<td>4.00%</td>
<td>3.90%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grid'5000 site</th>
<th>#CPU used</th>
<th>Execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bordeaux</td>
<td>82</td>
<td>1740s</td>
</tr>
<tr>
<td>Orsay</td>
<td>344</td>
<td>1959s</td>
</tr>
<tr>
<td>Rennes (paraci)</td>
<td>98</td>
<td>1594s</td>
</tr>
<tr>
<td>Rennes (parasol)</td>
<td>62</td>
<td>2689s</td>
</tr>
<tr>
<td>Rennes (paravent)</td>
<td>198</td>
<td>2062s</td>
</tr>
<tr>
<td>Sophia</td>
<td>160</td>
<td>1905s</td>
</tr>
</tbody>
</table>

mean: 1972 s
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

<table>
<thead>
<tr>
<th>BW – Lat</th>
<th>Bordeaux</th>
<th>Orsay</th>
<th>Rennes</th>
<th>Sophia</th>
<th>Lille</th>
<th>Toulouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toulouse</td>
<td>190Mb/s – 1.5ms</td>
<td>59.81Mb/s – 5.51ms</td>
<td>34.79Mb/s – 9.92ms</td>
<td>83.7Mb/s – 3.74ms</td>
<td>26.97Mb/s – 13.04ms</td>
<td>930.4Mb/s – 0.04ms</td>
</tr>
<tr>
<td>Lille</td>
<td>30.23Mb/s – 11.62ms</td>
<td>132.55Mb/s – 2.25ms</td>
<td>56.84Mb/s – 5.83ms</td>
<td>37.41Mb/s – 9.22ms</td>
<td>938Mb/s – 0.04ms</td>
<td></td>
</tr>
<tr>
<td>Sophia</td>
<td>68.1Mb/s – 5.2ms</td>
<td>42.4Mb/s – 8.6ms</td>
<td>40.2Mb/s – 9.1ms</td>
<td>940.5Mb/s – 0.04ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rennes</td>
<td>110.1Mb/s – 4.0ms</td>
<td>95.4Mb/s – 4.7ms</td>
<td>940.4Mb/s – 0.04ms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orsay</td>
<td>108.1Mb/s – 4.1ms</td>
<td>930.4Mb/s – 0.06ms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bordeaux</td>
<td>940.2Mb/s – 0.04ms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
117 participants (we tried to limit to 100)

- Engineers
- Students
- Scientists
- PostDocs

Topics and exercises:
- Reservation
- Reconfiguration
- MPI on the Cluster of Clusters
- Virtual Grid based on Globus GT4

Non Grid’5000 users: 44%
Grid’5000 users: 56%

Locations:
- Bordeaux: 6%
- Grenoble: 19%
- Lille: 6%
- Nancy: 14%
- Lyon: 15%
- Orsay: 17%
- Sophia: 3%
- Toulouse: 6%

Fields:
- Computer science: 86%
- Physics: 9%
- Mathematics: 1%
- Biology: 1%
- Chemistry: 3%