Grid'5000

a scientific instrument for experiment-driven research on parallel, large-scale and distributed systems

Lucas Nussbaum

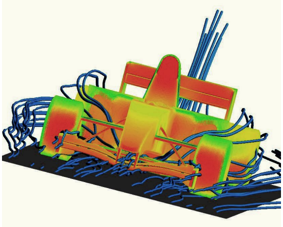
`lucas.nussbaum@loria.fr`

Grid'5000 executive committee member
in charge of following the technical team

Experimentation for distributed systems

Experimentation for distributed systems

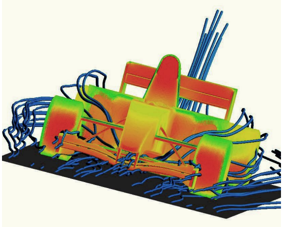
Simulation



1. Model application
2. Model environment
3. **Compute** interactions

Experimentation for distributed systems

Simulation



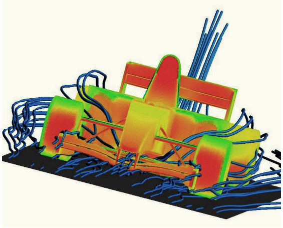
1. Model application
2. Model environment
3. **Compute** interactions

Real-scale experiments

Execute the **real** application on **real** machines

Experimentation for distributed systems

Simulation



1. Model application
2. Model environment
3. **Compute** interactions

Real-scale experiments

Execute the **real** application on **real** machines

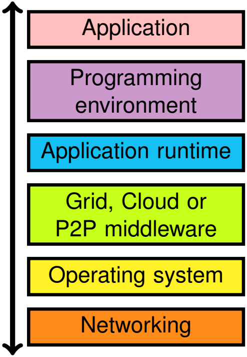
Complementary solutions:

☺ Work on algorithms
☺ Scalable, more user-friendly

☺ Work on applications
☺ Closer to production use

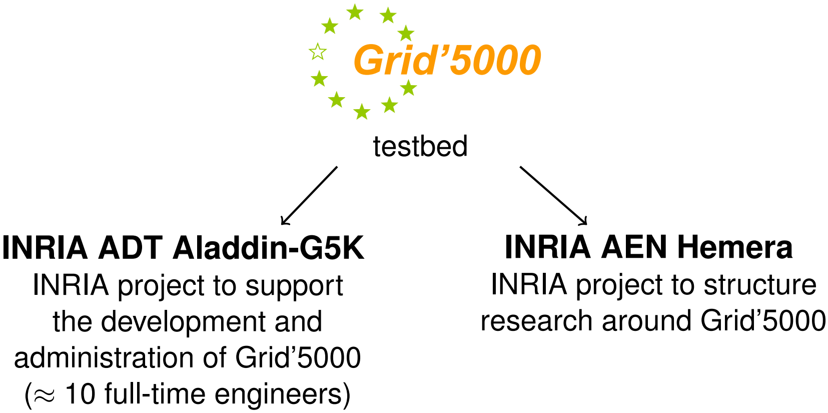
Grid'5000

- **Testbed for research on distributed systems**
  - High Performance Computing
  - Grids
  - Peer-to-peer systems
  - Cloud computing
- History:
  - 2003: Project started (ACI GRID)
  - 2005: Opened to users



- Funding: Inria, CNRS and many local entities (regions, universities)
- Only for research on distributed systems → no production usage
  Litmus test: *are you interested in the result of the computation?*
  - Free nodes during daytime to prepare experiments
  - Large-scale experiments during nights and week-ends

Organization



testbed

INRIA ADT Aladdin-G5K
INRIA project to support
the development and
administration of Grid'5000
($\approx$ 10 full-time engineers)

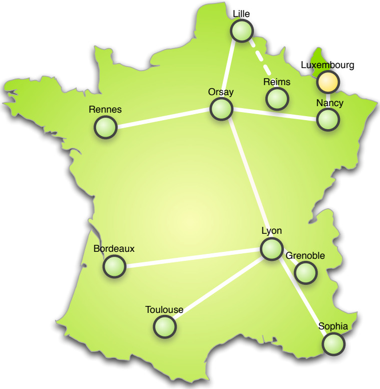
INRIA AEN Hemera
INRIA project to structure
research around Grid'5000

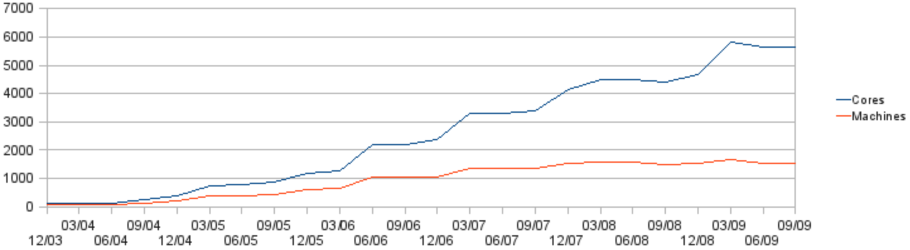
People:

- Scientific director: Frédéric Desprez
- Technical director: David Margery
- Hemera director: Christian Perez

Current status

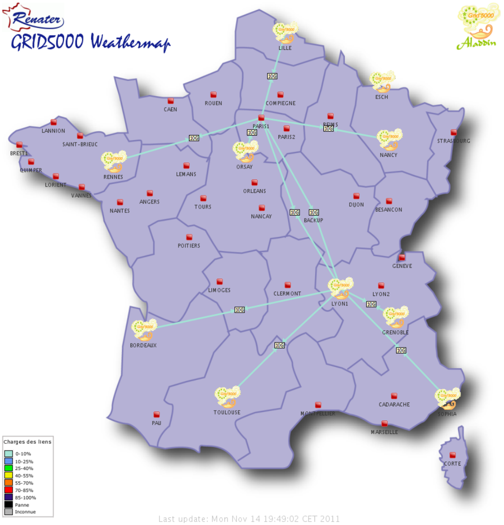
- 11 sites (1 outside France)
- 26 clusters
- 1700 nodes
- 7400 cores
- Diverse technologies:
  - Intel (60%), AMD (40%)
  - CPUs from one to 12 cores
  - Myrinet, Infiniband {S,D,Q}DR
  - Two GPU clusters
- **500+ users per year**



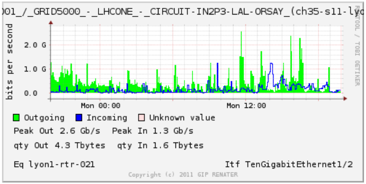


Backbone network

Dedicated 10 Gbps backbone provided by RENATER (french NREN)



Last update: Mon Nov 14 19:49:02 CET 2011



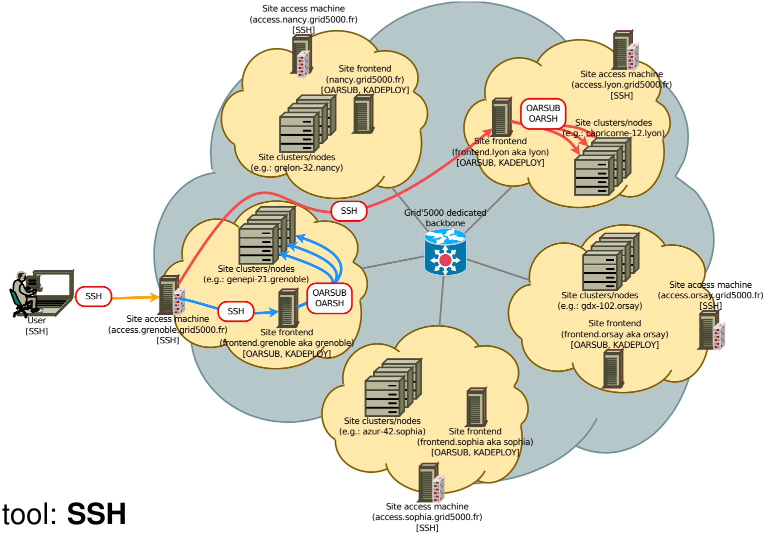
```
)01_/_GRID5000_-_LHCONE_-_CIRCUIT-IN2P3-LAL-ORSAY_(ch35-s11-lyo
■ Outgoing  ■ Incoming  □ Unknown value
Peak Out 2.6 Gb/s   Peak In 1.3 Gb/s
qty Out 4.3 Tbytes   qty In 1.6 Tbytes

Eq lyon1-rtr-021                    Itf TenGigabitEthernet1/2
```

Work in progress:

- packet-level and flow-level monitoring
- bandwidth reservation and limitation

Using Grid’5000: the user’s point of view



Site access machine
(access.nancy.grid5000.fr)
[SSH]

Site frontend
(nancy.grid5000.fr)
[OARSUB, KADEPLOY]

Site clusters/nodes
(e.g.: grelon-32.nancy)

Site access machine
(access.lyon.grid5000.fr)
[SSH]

OARSUB
OARSH

Site clusters/nodes
(e.g.: capricorne-12.lyon)

Site frontend
(frontend.lyon aka lyon)
[OARSUB, KADEPLOY]

SSH

Grid'5000 dedicated
backbone

Site clusters/nodes
(e.g.: genepi-21.grenoble)

OARSUB
OARSH

User
[SSH]

SSH

Site access machine
(access.grenoble.grid5000.fr)
[SSH]

SSH

Site frontend
(frontend.grenoble aka grenoble)
[OARSUB, KADEPLOY]

Site clusters/nodes
(e.g.: gdx-102.orsay)

Site access machine
(access.orsay.grid5000.fr)
[SSH]

Site frontend
(frontend.orsay aka orsay)
[OARSUB, KADEPLOY]

Site clusters/nodes
(e.g.: azur-42.sophia)

Site frontend
(frontend.sophia aka sophia)
[OARSUB, KADEPLOY]

Site access machine
(access.sophia.grid5000.fr)
[SSH]

- Key tool: **SSH**
- Private network: connect through access machines
- Data storage: **NFS** (one server per Grid’5000 site)

Grid'5000 software stack

- Resource management: **OAR**
- System reconfiguration: **Kadeploy**
- Network isolation: **KaVLAN**
- Monitoring: **Ganglia**, **Kaspied**, **Energy**
- Putting it all together: **Grid'5000 API**

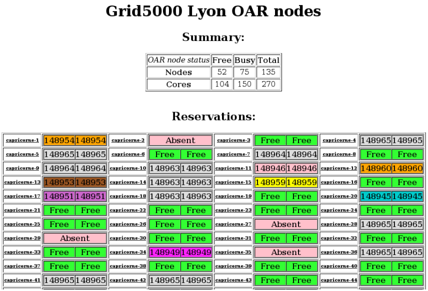
Resource management: OAR

- Batch scheduler with specific features
  - interactive jobs
  - advance reservations
  - powerful resource matching
- Resources hierarchy: cluster / switch / node / cpu / core
- Properties: memory size, disk type & size, hardware capabilities, network interfaces, . . .
- Other kind of resources: VLANs, IP ranges for virtualization

I want 1 core on 2 nodes of the same cluster with
4096 GB of memory and Infiniband 10G +
1 cpu on 2 nodes of the same switch with dualcore processors
for a walltime of 4 hours. . .

```
oarsub -I -l "{memnode=4096 and
ib10g=’YES’}/cluster=1/nodes=2/core=1
+{cpucore=2}/switch=1/nodes=2/cpu=1,walltime=4:0:0"
```

Resource management: OAR - visualization



Grid5000 Lyon OAR nodes

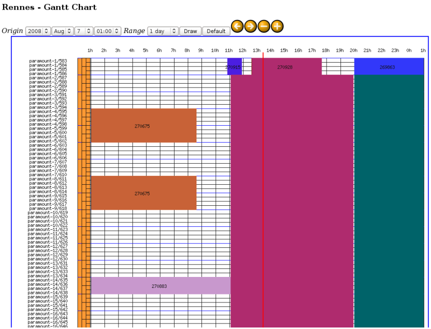
Summary:

OAR node status	Free	Busy	Total
Nodes	52	75	135
Cores	104	150	270

Reservations:

capricorne-1	148954	148954	capricorne-2	Absent		capricorne-3	Free	Free	capricorne-4	148965	148965
capricorne-5	148965	148965	capricorne-6	Free	Free	capricorne-7	148964	148964	capricorne-8	Free	Free
capricorne-9	148964	148964	capricorne-10	148963	148963	capricorne-11	148946	148946	capricorne-12	148960	148960
capricorne-13	148953	148953	capricorne-14	148963	148963	capricorne-15	148959	148959	capricorne-16	Free	Free
capricorne-17	148951	148951	capricorne-18	148963	148963	capricorne-19	Free	Free	capricorne-20	148945	148945
capricorne-21	Free	Free	capricorne-22	Free	Free	capricorne-23	Free	Free	capricorne-24	Free	Free
capricorne-25	Free	Free	capricorne-26	Free	Free	capricorne-27	Absent		capricorne-28	148965	148965
capricorne-29	Absent		capricorne-30	Free	Free	capricorne-31	Free	Free	capricorne-32	Free	Free
capricorne-33	Free	Free	capricorne-34	148949	148949	capricorne-35	Absent		capricorne-36	148965	148965
capricorne-37	Free	Free	capricorne-38	Free	Free	capricorne-39	Free	Free	capricorne-40	Free	Free
capricorne-41	148965	148965	capricorne-42	148965	148965	capricorne-43	Free	Free	capricorne-44	Free	Free

Resources status



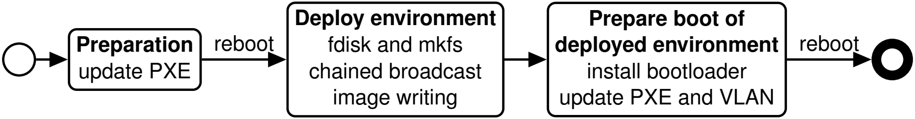
Rennes - Gantt Chart

Origin 2008 Aug 7 01:00 *Range* 1 day

Gantt chart

Kadeploy – scalable cluster deployment tool

- Provides a *Hardware-as-a-Service* Cloud infrastructure
- Built on top of PXE, DHCP, TFTP
- **Scalable, efficient, reliable and flexible**:
  - Chain-based and BitTorrent environment broadcast
  - **255 nodes deployed in 7 minutes**
- Support of a **broad range of systems** (Linux, Xen, *BSD, etc.)
- Command-line interface & asynchronous interface (REST API)

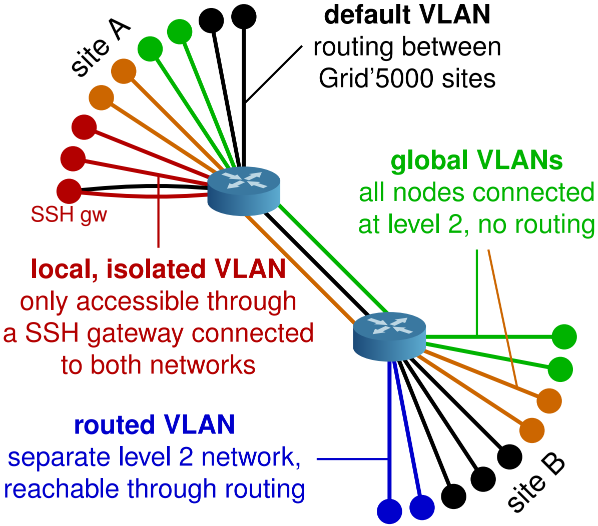


http://kadeploy3.gforge.inria.fr/

Network isolation: KaVLAN

- Reconfigures switches for the duration of a user experiment to achieve **complete level 2 isolation**:
  - Avoid network pollution (broadcast, unsolicited connections)
  - Enable users to start their own DHCP servers
  - Experiment on ethernet-based protocols
  - Interconnect nodes with another testbed without compromising the security of Grid'5000
- Relies on **802.1q (VLANs)**
- Compatible with many network equipments
  - Can use SNMP, SSH or telnet to connect to switches
  - Supports Cisco, HP, 3Com, Extreme Networks and Brocade
- Controlled with a command-line client or a REST API

KaVLAN - different VLAN types



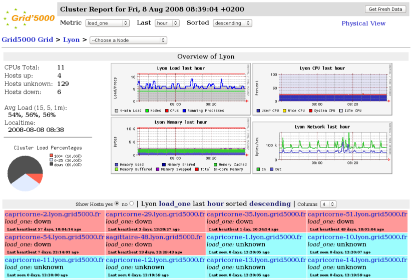
default VLAN
routing between
Grid’5000 sites

global VLANs
all nodes connected
at level 2, no routing

local, isolated VLAN
only accessible through
a SSH gateway connected
to both networks

routed VLAN
separate level 2 network,
reachable through routing

Monitoring: Ganglia, Kaspied, Energy



Grid'5000

Cluster Report for Fri, 8 Aug 2008 08:39:04 +0200 [Get Fresh Data]

Metric [load_one] Last [hour] Sorted [descending] Physical View

Grid5000 Grid > Lyon > [--Choose a Node]

Overview of Lyon

CPUs Total:	11
Hosts up:	4
Hosts unknown:	129
Hosts down:	6

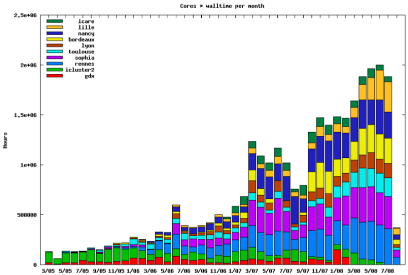
Avg Load (15, 5, 1m):
54%, 56%, 56%
Localtime:
2008-08-08 08:38

Show Hosts: yes ● no ○ | Lyon **load_one** last **hour** sorted **descending** | Columns [4]

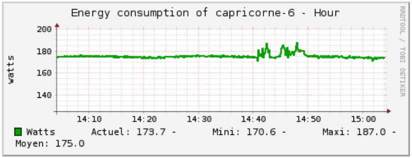
capricorne-2.lyon.grid5000.fr *load_one:* down **Last heartbeat 57 days, 16:04:14 ago**	capricorne-29.lyon.grid5000.fr *load_one:* down **Last heartbeat 3 days, 13:30:27 ago**	capricorne-35.lyon.grid5000.fr *load_one:* down **Last heartbeat 1 day, 20:34:54 ago**	capricorne-51.lyon.grid5000.fr *load_one:* down **Last heartbeat 49 days, 18:01:04 ago**
capricorne-54.lyon.grid5000.fr *load_one:* down **Last heartbeat 7 days, 22:14:01 ago**	sagittaire-48.lyon.grid5000.fr *load_one:* down **Last heartbeat 13 days, 22:29:43 ago**	capricorne-1.lyon.grid5000.fr *load_one:* unknown **Last seen 0 days, 12:20:05 ago**	capricorne-10.lyon.grid5000.fr *load_one:* unknown **Last seen 0 days, 12:20:07 ago**
capricorne-11.lyon.grid5000.fr *load_one:* unknown **Last seen 0 days, 12:20:00 ago**	capricorne-12.lyon.grid5000.fr *load_one:* unknown **Last seen 0 days, 12:19:59 ago**	capricorne-13.lyon.grid5000.fr *load_one:* unknown **Last seen 0 days, 12:20:05 ago**	capricorne-14.lyon.grid5000.fr *load_one:* unknown **Last seen 0 days, 12:19:59 ago**

Ganglia

Usage per month per cluster



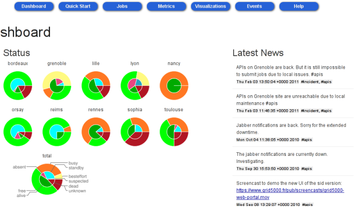
Kaspied
(Grid'5000 usage over time)



Power consumption

Putting it all together: Grid'5000 API

- Individual services & command-line interfaces are painful
- REST API for each Grid'5000 service:
  - **Reference API**: versioned description of Grid'5000 resources
  - **Monitoring API**: state of Grid'5000 resources
  - **Metrology API**: Ganglia data
  - **Jobs API**: OAR interface
  - **Deployments API**: Kadeploy interface
  - ...
- Also some nice Web interfaces on `https://api.grid5000.fr/`



Dashboard

Grid Status

Latest News

APIs on Grenoble are back. But it is still impossible to submit jobs due to local issues. #apis
Thu Feb 03 13:50:04 +0000 2011 #incident, #apis

APIs on Grenoble site are unreachable due to local maintenance #apis
Thu Feb 03 11:46:35 +0000 2011 #incident, #apis

Jabber notifications are back. Sorry for the extended downtime.
Mon Oct 04 11:38:05 +0000 2010 #apis

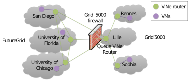
The jabber notifications are currently down. Investigating.
Thu Sep 30 15:53:50 +0000 2010 #apis

Screencast to demo the new UI of the sid version: https://www.grid5000.fr/pub/screencasts/grid5000-web-portal.mov
Wed Sep 08 13:29:07 +0000 2010 #apis

Leading to results in several fields

Cloud: Sky computing on FutureGrid and Grid’5000

- Nimbus cloud deployed on 450+ nodes
- Grid’5000 and FutureGrid connected using ViNe



HPC: factorization of RSA-768

- Feasibility study: prove that it can be done
- Different hardware ⇝ understand the performance characteristics of the algorithms

Grid: evaluation of the gLite grid middleware

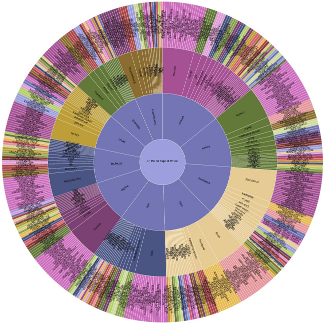
- Fully automated deployment and configuration on 1000 nodes (9 sites, 17 clusters)

gLite
Lightweight Middleware for Grid Computing

Open challenges

Testbeds optimize for experimental capabilities, not performance

- **Access** to the modern architectures / technologies
  - Not necessarily the fastest CPUs
  - But still expensive ⇝ funding!
- Ability to **trust** results
  - Regular checks of testbed for bugs
- Ability to **understand** results
  - Documentation of the infrastructure
  - Instrumentation & monitoring tools
    network, energy consumption
  - Evolution of the testbed
    maintenance logs, configuration history
- **Empower** users to perform complex experiments
  - Facilitate access to advanced software tools



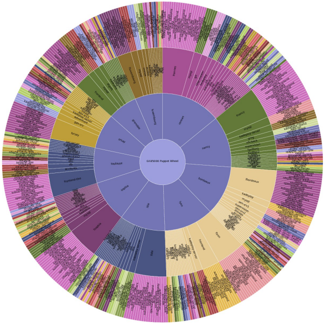
Open challenges

Testbeds optimize for experimental capabilities, not performance

- **Access** to the modern architectures / technologies
  - Not necessarily the fastest CPUs
  - But still expensive ⇝ funding!

- Ability to **trust** results
  - Regular checks of testbed for bugs

- Ability to **understand** results
  - Documentation of the infrastructure
  - Instrumentation & monitoring tools
    network, energy consumption
  - Evolution of the testbed
    maintenance logs, configuration history



- **Empower** users to perform complex experiments
  - Facilitate access to advanced software tools ← this afternoon

Conclusions

- Grid'5000: a testbed for experimentation on distributed systems
- With a unique combination of features
  - *Hardware-as-a-Service* cloud: redeployment of operating system on the bare hardware by users
  - Access to various technologies (CPUs, high performance networks, etc.)
  - Networking: dedicated backbone, monitoring, isolation
  - Programmable through an API

Interested in trying it? Contact us!

`https://www.grid5000.fr/`

`lucas.nussbaum@loria.fr`