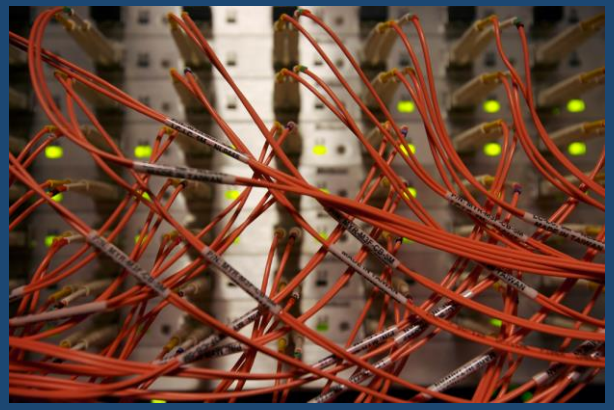


Grid'5000

Grid'5000: a Large Instrument for Parallel and Distributed Computing Experiments

Lucas Nussbaum

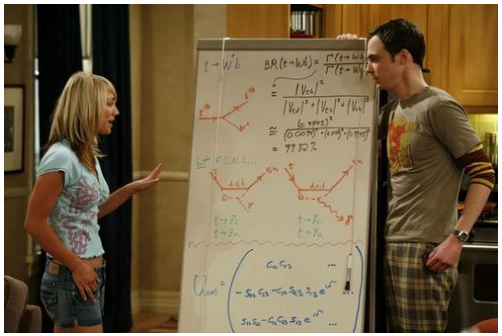
Université de Lorraine - LORIA - Madynes team

Joint work with G. Antoniu, F. Desprez, Y. Georgiou, D. Glesser, A. Lebre, L. Lefèvre, M. Liroz, D. Margery, C. Perez, L. Pouillioux

Validation in (Computer) Science

Two classical approaches for validation

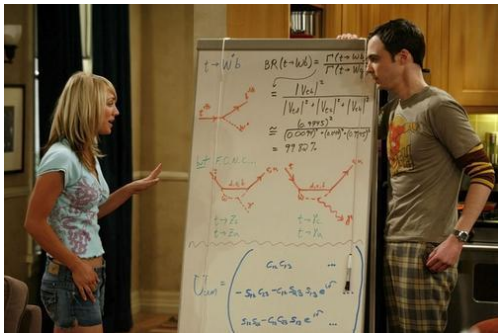
- **Formal**: equations, proofs, etc.
- **Experimental**: on a scientific instrument



Validation in (Computer) Science

Two classical approaches for validation

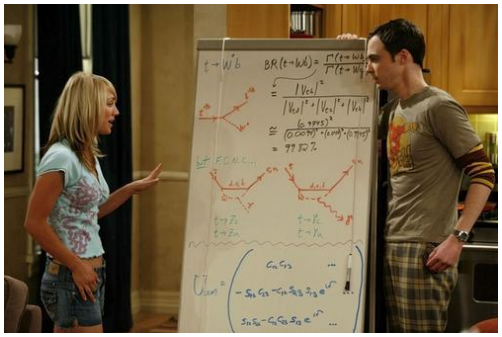
- **Formal**: equations, proofs, etc.
- **Experimental**: on a scientific instrument



Often a mix of both

- In Physics
- In Computer Science

Validation in (Computer) Science



Two classical approaches for validation

- **Formal**: equations, proofs, etc.
- **Experimental**: on a scientific instrument

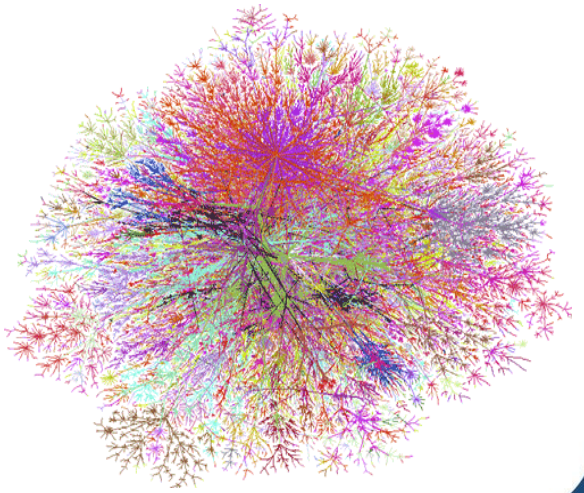
Often a mix of both

- In Physics
- In Computer Science

Very little formal validation in distributed computing research

- Our scientific objects often cannot be attacked theoretically
  - Too complex, dynamic, heterogeneous, large

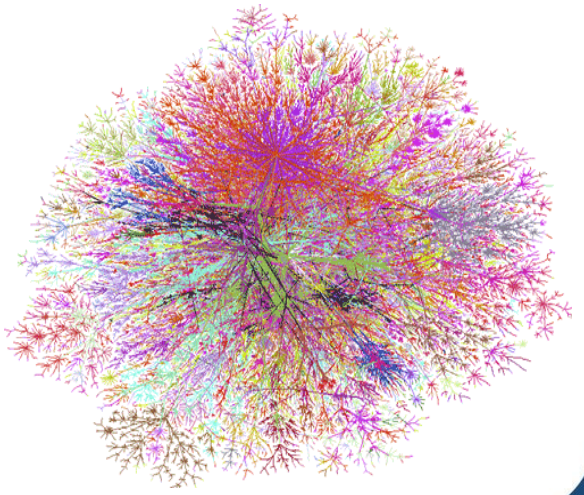
Computer science: an experimental science



Computer science: an experimental science

The reality of computer science

- not just information and algorithms
- also computers, network, programs, etc.



Computer science: an experimental science

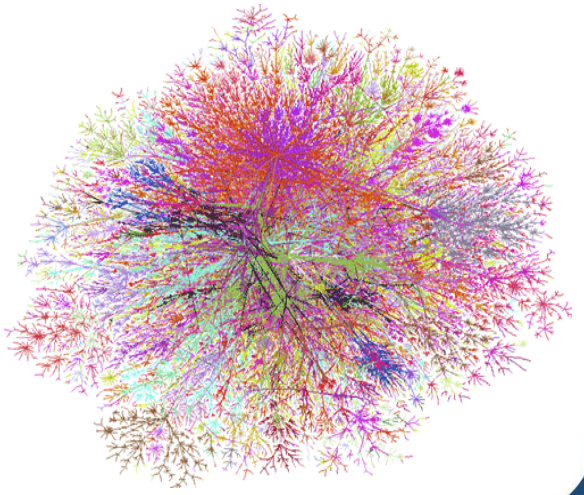
Grid'5000

The reality of computer science

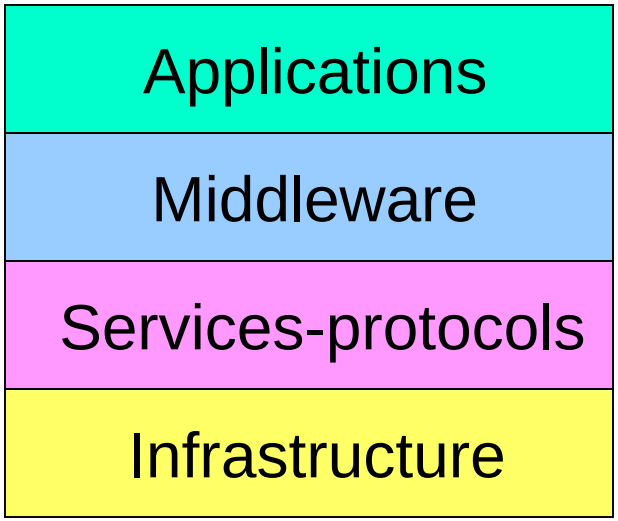
- not just information and algorithms
- also computers, network, programs, etc.

With a huge impact on performance

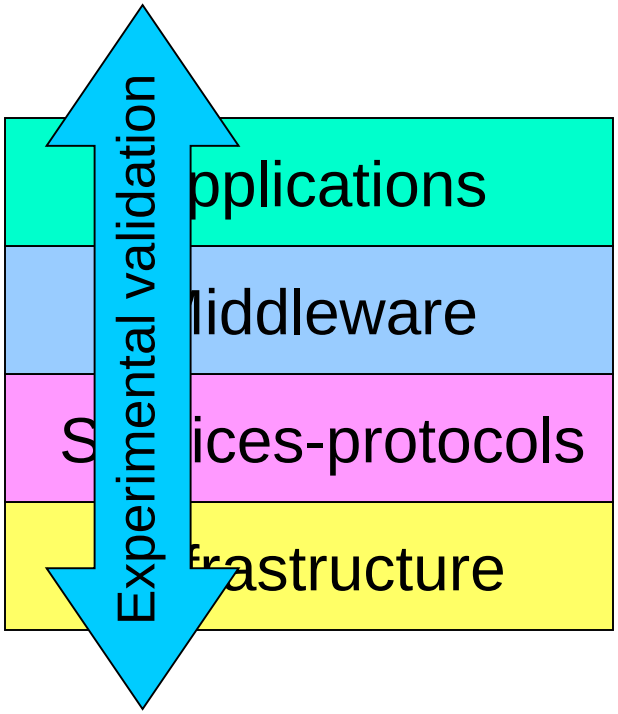
- Processors: caches, hyperthreading, multi-core
- Operating system: process scheduling, socket implementation, etc.
- Runtime environment: MPICH ≠ OPENMPI
- Middleware
- Various parallel architectures that can be heterogeneous, hierarchical, distributed, dynamic



Research issues at each layer of the stack



Research issues at each layer of the stack



Experimentation is hard!

- What is a good experiment ?
- Which methodologies, testbeds, tools ?

GRID'5000

- **Testbed for research on distributed systems**
  - Born (2003) from the observation that we need a better and larger testbed
  - High Performance Computing, Grids, Peer-to-peer systems, Cloud computing, Big Data
  - A complete access to the nodes' hardware in an exclusive mode
  - RIaaS : Real Infrastructure as a Service ! ?
  - **Not a Grid**, more like a meta-Grid, or a meta-Cloud: infrastructure to instanciate Grids and Clouds and experiment on them.

GRID'5000

- **Testbed for research on distributed systems**
  - Born (2003) from the observation that we need a better and larger testbed
  - High Performance Computing, Grids, Peer-to-peer systems, Cloud computing, Big Data
  - A complete access to the nodes' hardware in an exclusive mode
  - RIaaS : Real Infrastructure as a Service ! ?
  - **Not a Grid**, more like a meta-Grid, or a meta-Cloud: infrastructure to instanciate Grids and Clouds and experiment on them.

- **Funding**
  - INRIA, CNRS, and many local entities (regions, universities)

GRID'5000

- **Testbed for research on distributed systems**
  - Born (2003) from the observation that we need a better and larger testbed
  - High Performance Computing, Grids, Peer-to-peer systems, Cloud computing, Big Data
  - A complete access to the nodes' hardware in an exclusive mode
  - RIaaS : Real Infrastructure as a Service ! ?
  - **Not a Grid**, more like a meta-Grid, or a meta-Cloud: infrastructure to instanciate Grids and Clouds and experiment on them.

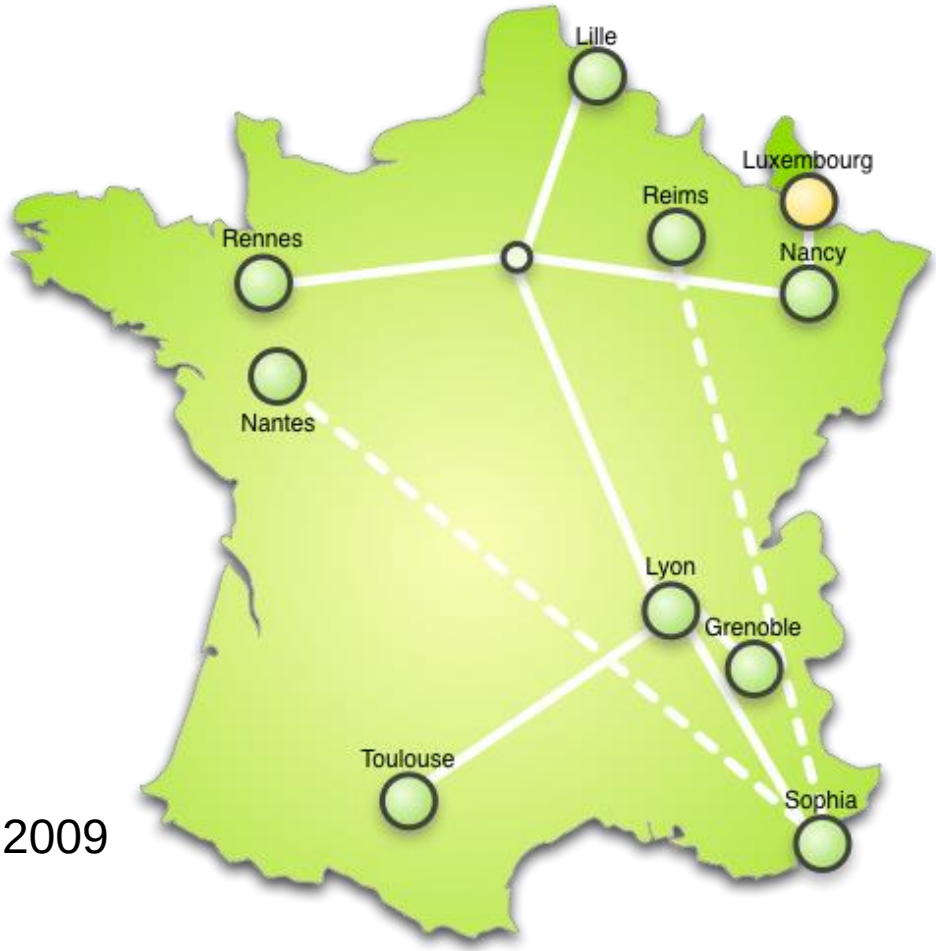
- **Funding**
  - INRIA, CNRS, and many local entities (regions, universities)

- For research in computer science

  → focus on how the computation/processing was done, not on the result
  - Free nodes during daytime to prepare experiments
  - Large-scale experiments during nights and week-ends

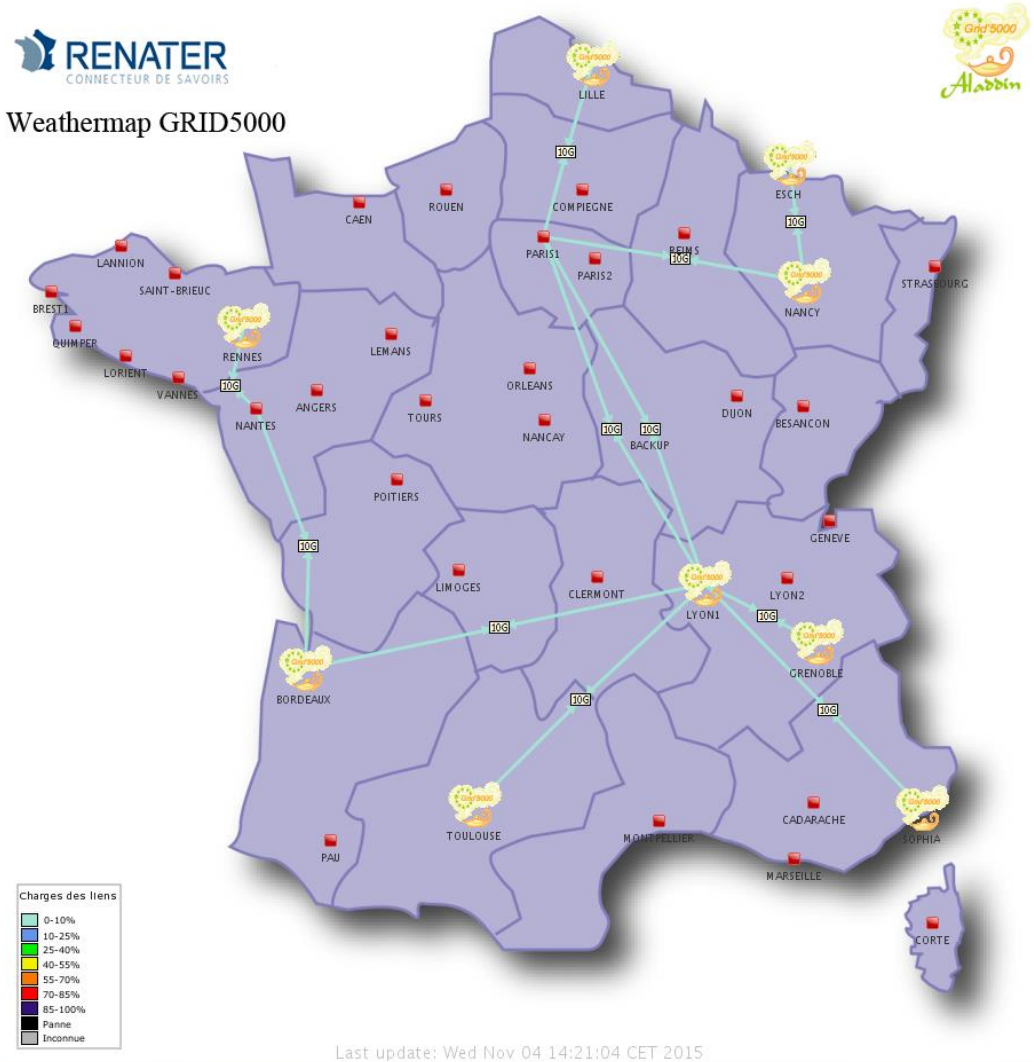
Current Status

- 10 sites (1 outside France), 25 clusters, 1000 nodes, 8000 cores
- **Diverse technologies**
  - Intel, AMD
  - CPUs from one to 12 cores
  - Ethernet 1G, 10G,
  - Infiniband {S, D, Q}DR
  - Two GPU clusters
  - One Xeon Phi cluster
  - 3 data clusters (3-5 disks/node)
- Hardware renewed regularly
- Widely used since 2005
  - More than **500 users** per year
  - More than **750 publications** since 2009



Backbone Network

Dedicated 10 Gbps backbone provided by RENATER (french NREN)

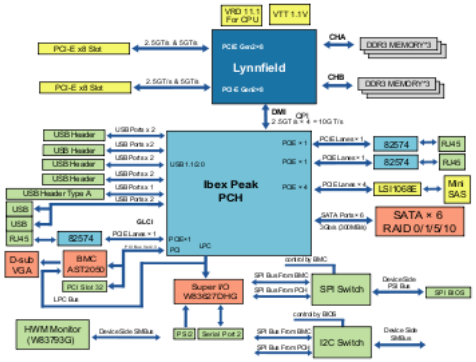


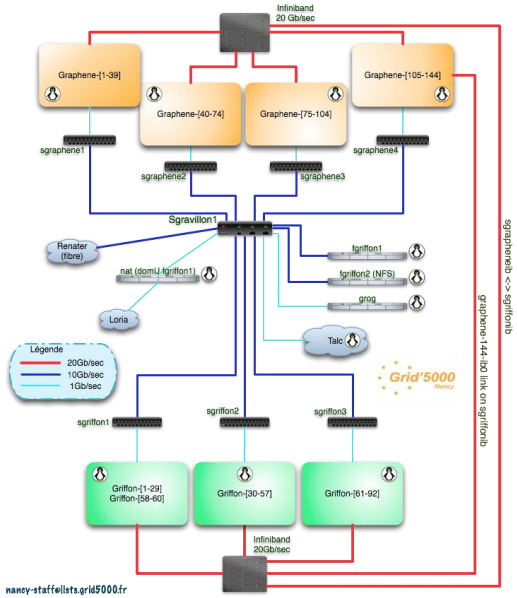
Facets of an Experiment on Grid'5000

- Description and verification of the environment
- Reconfiguring the testbed to meet experimental needs
- Monitoring experiments, extracting and analyzing data
- Improving description and control of experiments

Description and selection of resources

- **Describing resources to understand results**
  - Detailed description on the Grid'5000 wiki
  - Machine-parsable format (JSON)
  - Archived (State of testbed 6 months ago?)






```
"processor": {
  "cache_l2": 8388608,
  "cache_l1": null,
  "model": "Intel Xeon",
  "instruction_set": "",
  "other_description": "",
  "version": "X3440",
  "vendor": "Intel",
  "cache_l1i": null,
  "cache_l1d": null,
  "clock_speed": 2530000000.0
},
"uid": "graphene-1",
"type": "node",
"architecture": {
  "platform_type": "x86_64",
  "smt_size": 4,
  "smp_size": 1
},
"main_memory": {
  "ram_size": 17179869184,
  "virtual_size": null
},
"storage_devices": [
  {
    "model": "Hitachi HDS72103",
    "size": 298023223876.953,
    "driver": "ahci",
    "interface": "SATA II",
    "rev": "JPFO",
    "device": "sda"
  }
],
```

- Selecting resources
  - OAR database filled from JSON

```
oarsub -p "wattmeter='YES' and gpu='YES' »
oarsub -l "cluster='a'/nodes=1+cluster='b' and eth10g='Y'/nodes=2,walltime=2"
```

Verification of resources

Inaccuracies in resources descriptions ➔ dramatic consequences

- Happen frequently: maintenance, broken hardware (e.g. RAM)
- Our solution: g5k-checks
  - Runs at node boot (can also be run manually by users)
  - Retrieves current description of node in Reference API
  - Acquire information on node using OHAI, ethtool, etc.
  - Compare with Reference API

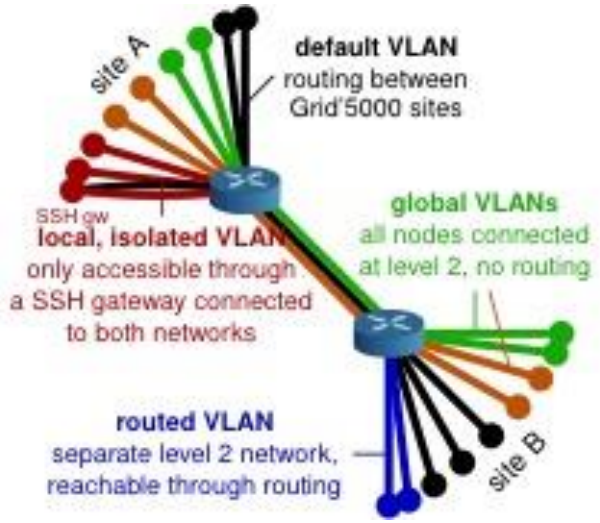
Reconfiguring the testbed

- **Typical needs**
  - How can I install $SOFTWARE on my nodes?
  - How can I add $PATCH to the kernel running on my nodes?
  - Can I run a custom MPI to test my fault tolerance work?
  - How can I experiment with that Cloud/Grid middleware?

- Likely answer on any production facility: impossible
  - Or: use virtual machines ➔ experimental bias

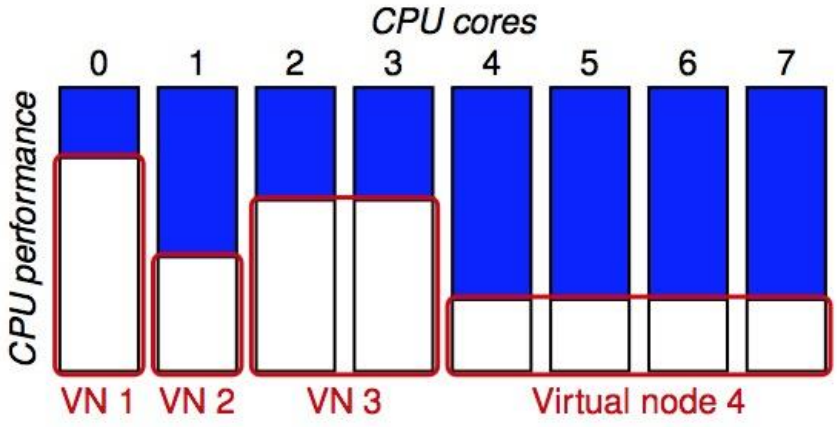
- **On Grid'5000**
  - Operating System reconfiguration with Kadeploy
    - Hardware-as-a-service Cloud!
  - Customize networking environment with KaVLAN
    - To isolate your experiment

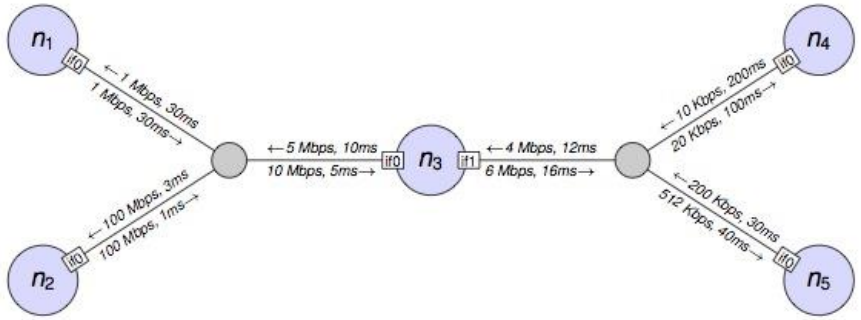
KADEPLOY



Changing experimental conditions

- **Reconfigure experimental conditions with Distem**
  - Introduce heterogeneity in an homogeneous cluster
  - Emulate complex network topologies
  - Introduce faults, varying concurrent load



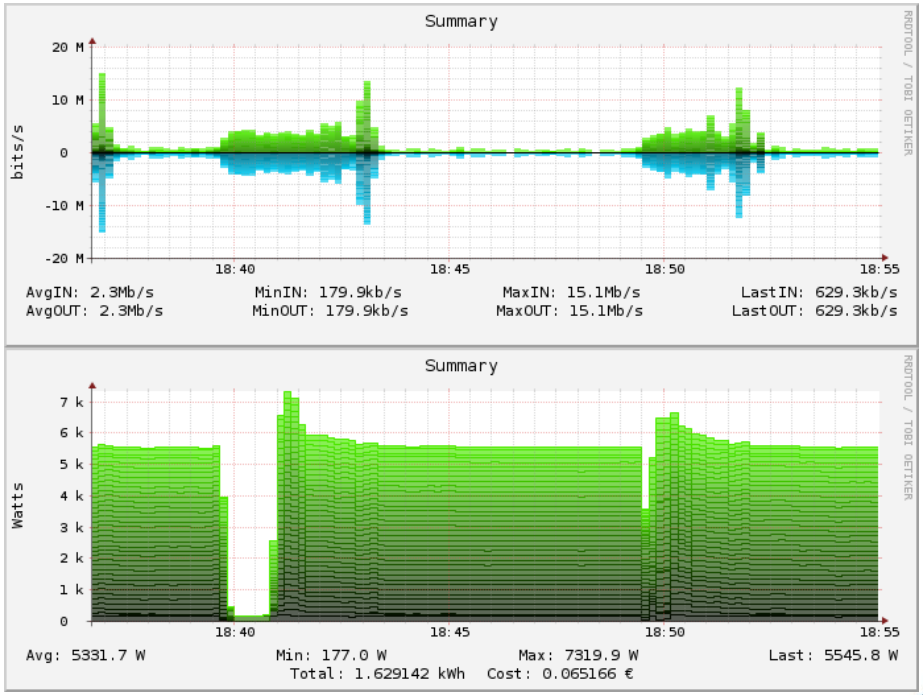


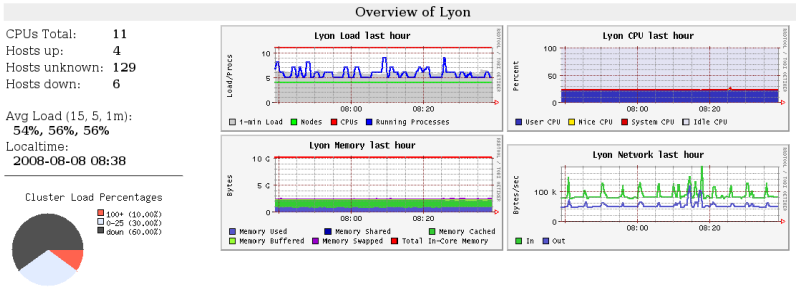
http://distem.gforge.inria.fr/

Monitoring experiments

Goal: enable users to understand what happens during their experiment

- System-level probes (usage of CPU, memory, disk, with Ganglia)
- Infrastructure-level probes
  - Network, power consumption
  - Captured at high frequency (1 Hz)
  - Live visualization
  - REST API
  - Long-term storage





Improving description and control of experiments

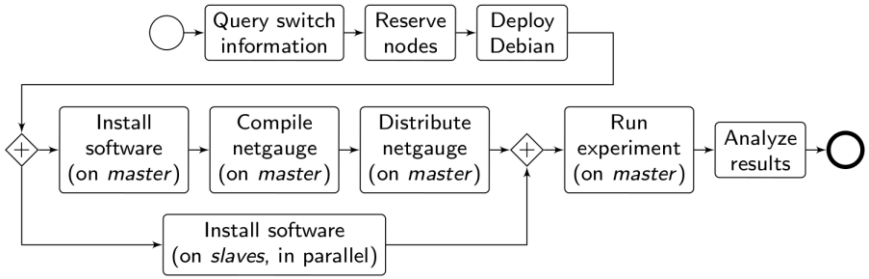
- Legacy way of performing experiments: shell commands
  - time-consuming
  - error-prone
  - details tend to be forgotten over time
- Promising solution: automation of experiments
  - Executable description of experiments
- Support from the testbed: Grid'5000 RESTful API
  - Resource selection, reservation, deployment, monitoring
- Several projects around Grid'5000 (but not specific to Grid'5000)
  - g5k-campaign, Expo, Execo, XPFlow
  - Facilitate scripting of experiments in high-level languages (Ruby, Python)
    - Testbed management
    - Local & remote execution of commands
    - Data management

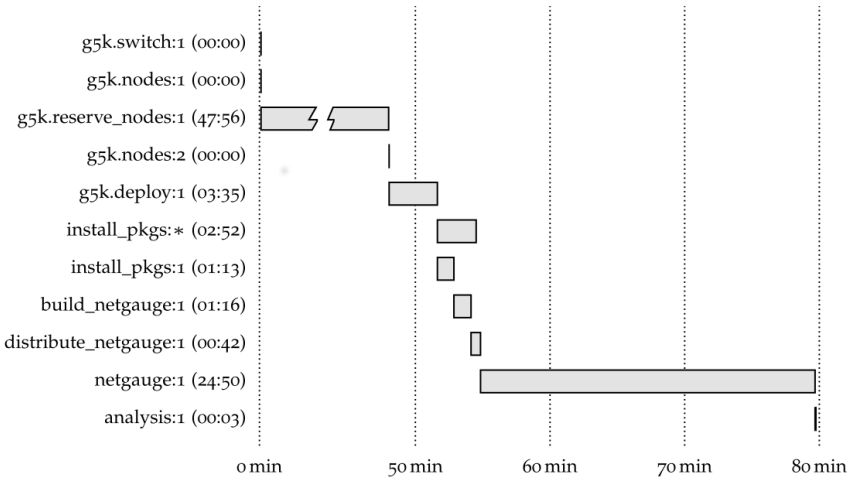
XPFlow

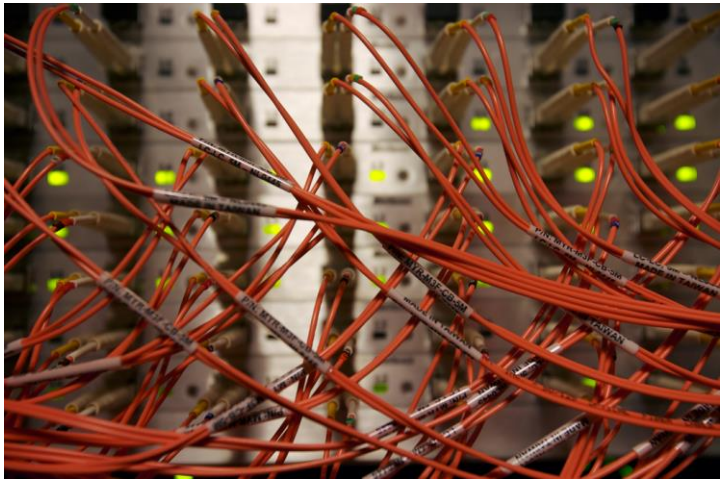
Experiments as a Business Workflow

Supports error handling, checkpointing, built-in logging and provenance collection

```
engine.process :exp do |site, switch|
    s = run g5k.switch, site, switch
    ns = run g5k.nodes, s
    r = run g5k.reserve_nodes,
        :nodes => ns, :time => '2h',
        :site => site, :type => :deploy
    master = (first_of ns)
    rest = (tail_of ns)
    run g5k.deploy,
        r, :env => 'squeeze-x64-nfs'
    checkpoint :deployed
    parallel :retry => true do
        forall rest do |slave|
            run :install_pkgs, slave
        end
        sequence do
            run :install_pkgs, master
            run :build_netgauge, master
            run :dist_netgauge,
                master, rest
        end
    end
    checkpoint :prepared
    output = run :netgauge, master, ns
    checkpoint :finished
    run :analysis, output, switch
end
```







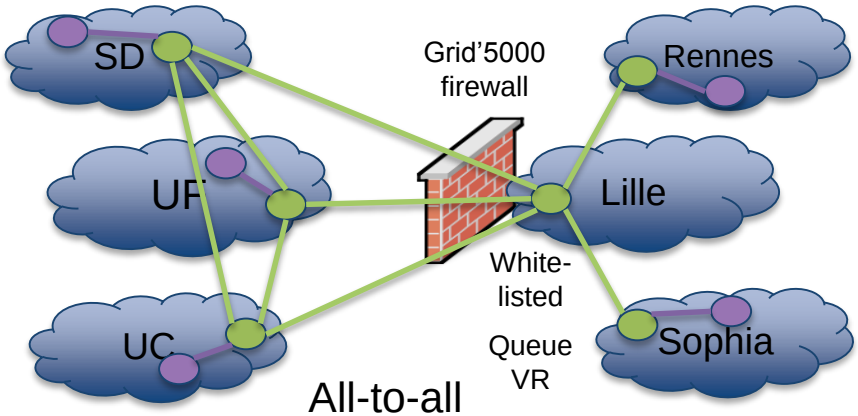
GRID'5000 EXPERIMENTS

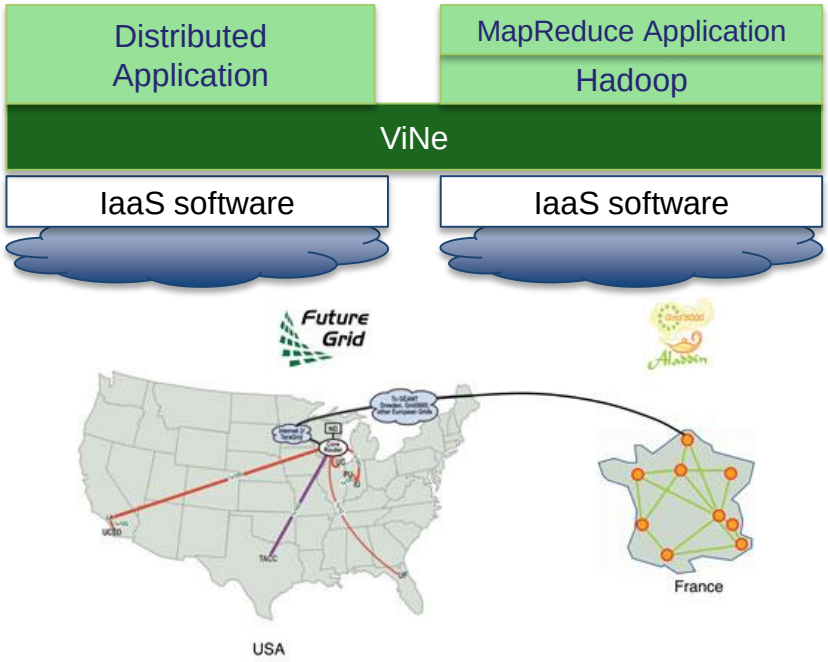
VIRTUALIZATION AND CLOUDS

GRID'5000, Virtualization and Clouds: Sky computing use-case

Experiments between USA and France

- Nimbus (resource management, contextualization)/ViNe (connectivity)/Hadoop (task distribution, fault-tolerance, dynamicity)
- FutureGrid (3 sites) and Grid'5000 (3 sites) platforms
- Optimization of creation and propagation of VMs



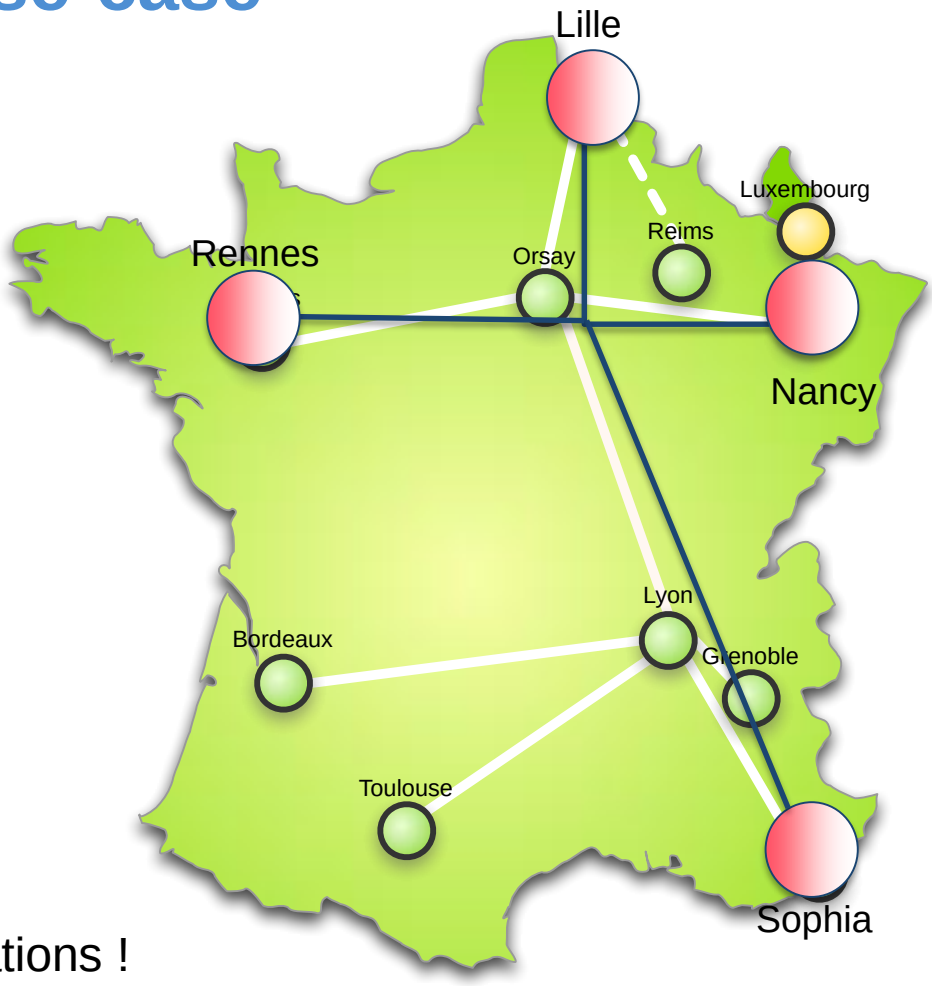


Large-Scale Cloud Computing Research: Sky Computing on FutureGrid and Grid'5000, by Pierre Riteau, Maurício Tsugawa, Andréa Matsunaga, José Fortes and Kate Keahey, ERCIM News 83, Oct. 2010.

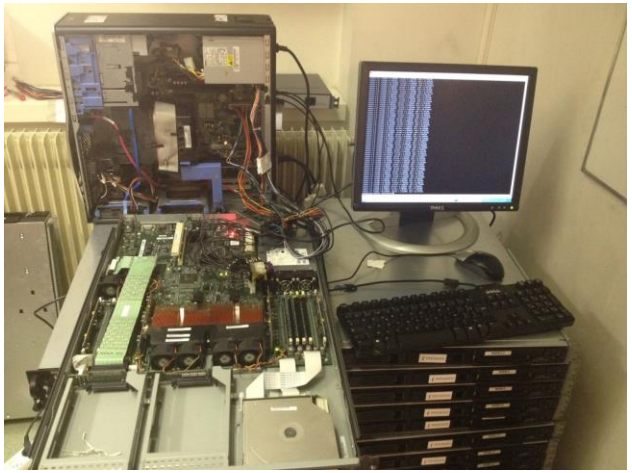
GRID'5000, Virtualization and Clouds: Dynamic VM placement use-case

Deploy 10240 VMs upon 512 PMs

- Prepare the experiment
  - Book resources
    512 PMs with Hard. Virtualization
    A global VLAN
    A /18 for IP ranges
  - Deploy KVM images and put PMs in the global VLAN
- Launch/Configure VMs
  - A dedicated script leveraging Taktuk utility to interact with each PM
  - G5K-subnet to get booked IPs and assign them to VMs
- Start the experiment and make publications !

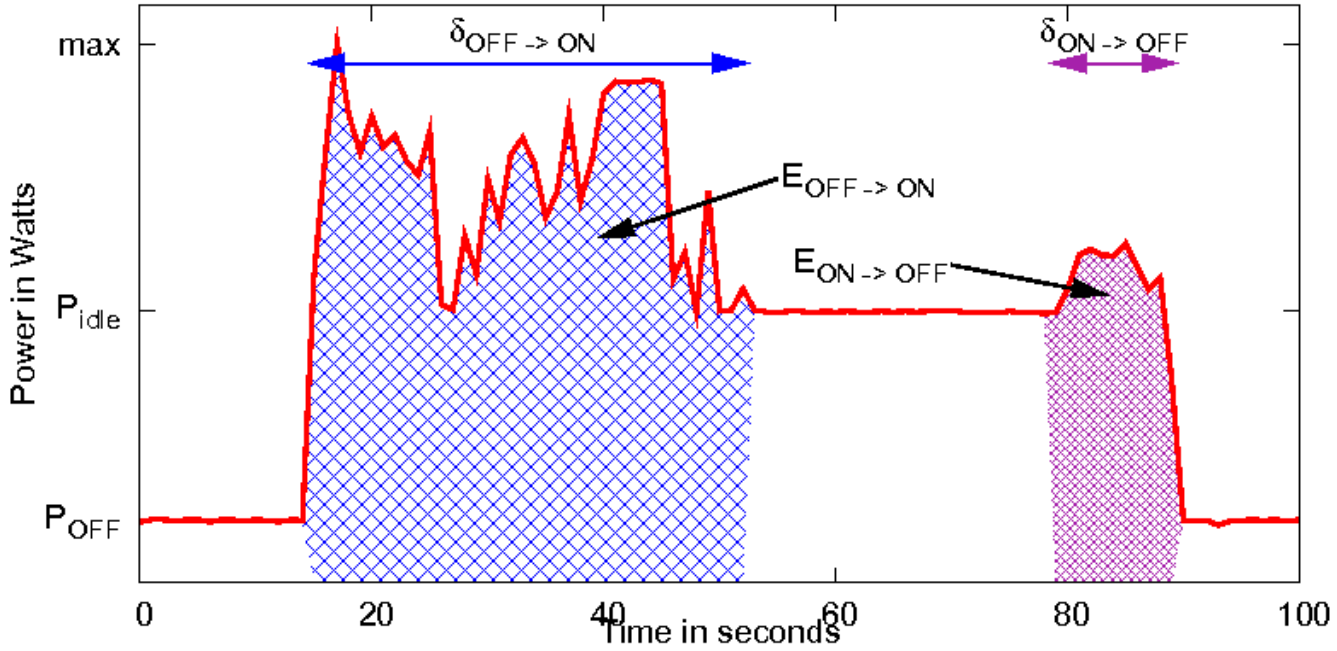


F. Quesnel, D. Balouek, and A. Lebre. **Deploying and Scheduling Thousands of Virtual Machines on Hundreds of Nodes Distributed Geographically**. In IEEE International Scalable Computing Challenge (SCALE 2013) (colocated with CCGRID 2013), Netherlands, May 2013



ENERGY MANAGEMENT

Aggressive ON/OFF is not always the best solution

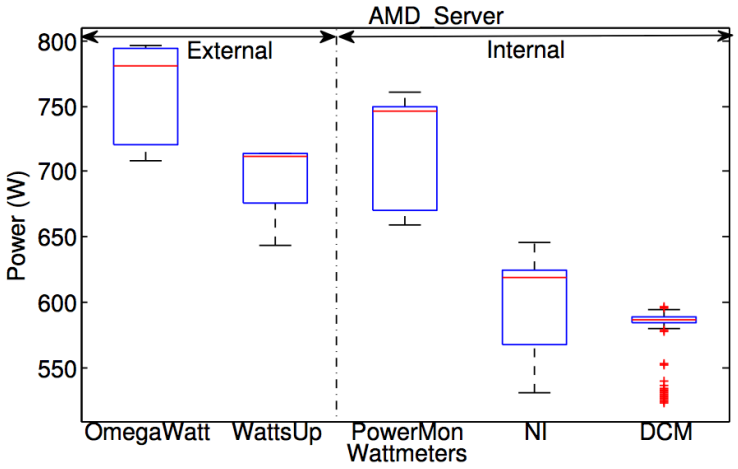


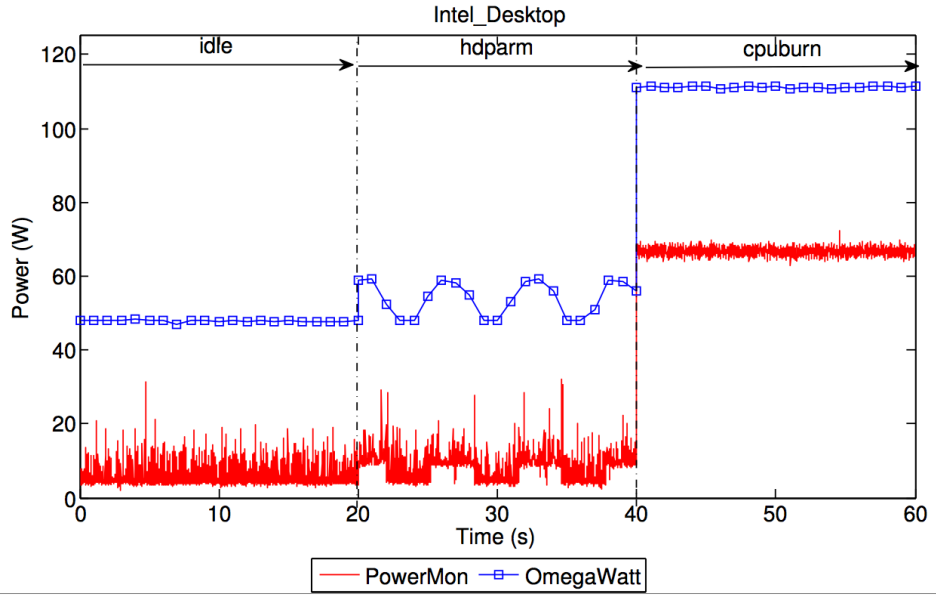
- Exploiting the gaps between activities
- Reducing unused plugged ressources number
- Only switiching off if potential energy saving

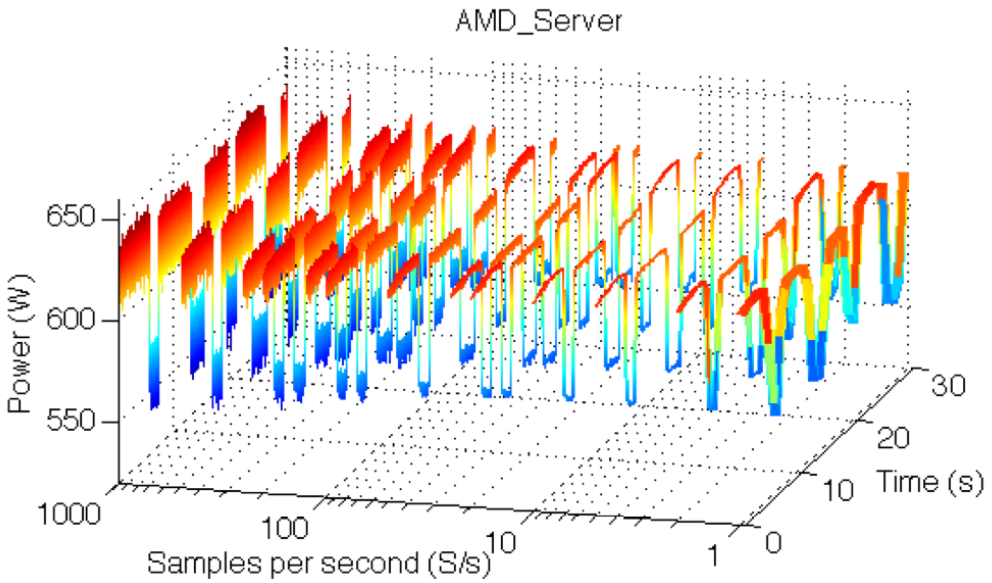
Anne-Cecile Orgerie, Laurent Lefevre, and Jean-Patrick Gelas. "**Save Watts in your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems**", ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems, Melbourne, Australia, December 2008

To understand energy measurements : take care of your wattmeters !

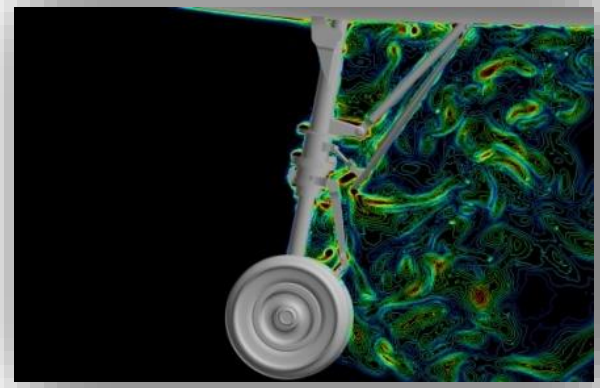
Frequency / precision







M. Diouri, M. Dolz, O. Glück, L. Lefevre, P. Alonso, S. Catalan, R. Mayo, E. Quintan-Orti. **Solving some Mysteries in Power Monitoring of Servers: Take Care of your Wattmeters!**, *EE-LSDS 2013 : Energy Efficiency in Large Scale Distributed Systems conference* , Vienna, Austria, April 22-24, 2013



HIGH PERFORMANCE COMPUTING

Riplay: A Tool to Replay HPC Workloads

- **RJMS : Ressource and Job Management System**
  - It manages resources and schedule jobs on High-Performance Clusters
  - Most famous ones : Maui/Moab, OAR, PBS, SLURM

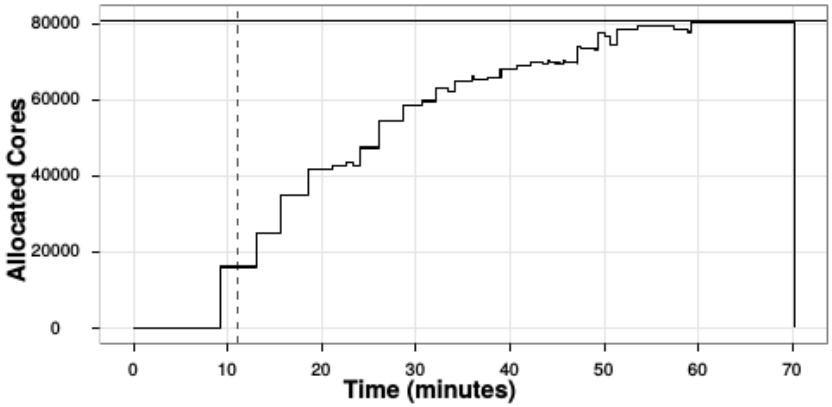
- ***Riplay***
  - Replay traces on a real RJMS in an emulated environment
  - 2 RJMS supported (OAR and SLURM)
  - Jobs replaced by *sleep commands*
  - Can replay a full or an interval of a workload

- **On Grid'5000**
  - 630 emulated cores need 1 physical core to run

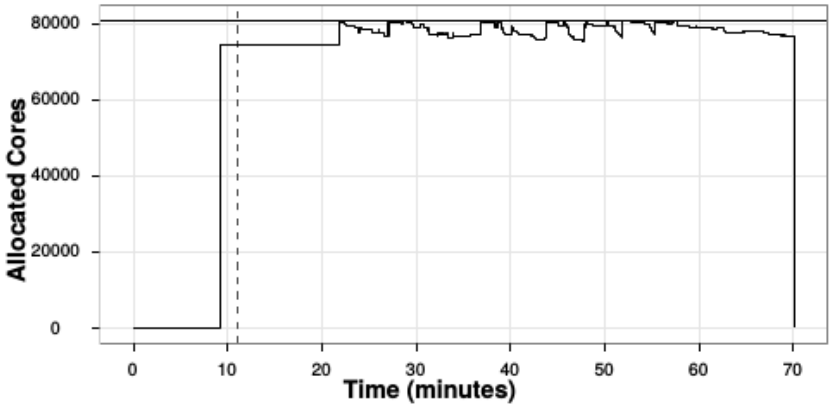
- **Curie (rank 26th on last Top500, 80640 cores)**
  - Curie's RJMS can be ran on 128 Grid'5000 cores

Riplay: A Tool to Replay HPC Workloads

- Test RJMS scalability
  - Without the need of the actual cluster.
  - Test a huge cluster fully loaded on a RJMS in minutes.



OAR before optimizations



OAR after optimizations

Large Scale Experimentation Methodology for Resource and Job Management Systems on HPC Clusters, Joseph Emeras, David Glesser, Yiannis Georgiou and Olivier Richard

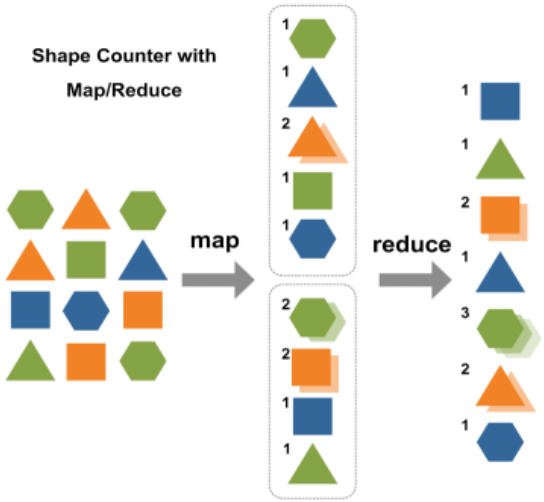
https://forge.imag.fr/projects/evalys-tools/

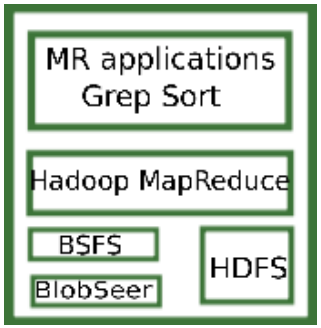
DATA MANAGEMENT

Scalable Map-Reduce Processing

Goal: High-performance Map-Reduce processing through concurrency-optimized data processing

- **Some results**
  - Versioning-based concurrency management for increased data throughput (BlobSeer approach)
  - Efficient intermediate data storage in pipelines
  - Substantial improvements with respect to Hadoop
  - Application to efficient VM deployment
- **Intensive, long-run experiments done on Grid'5000**
  - Up to 300 nodes/500 cores
  - Plans: validation within the IBM environment with IBM MapReduce Benchmarks





- ANR Project Map-Reduce (ARPEGE, 2010-2014)
- Partners: Inria (teams : KerData - leader, AVALON, Grand Large), Argonne National Lab, UIUC, JLPC, IBM, IBCP

mapreduce.inria.fr

Damaris: A Middleware-Level Approach to I/O on Multicore HPC Systems

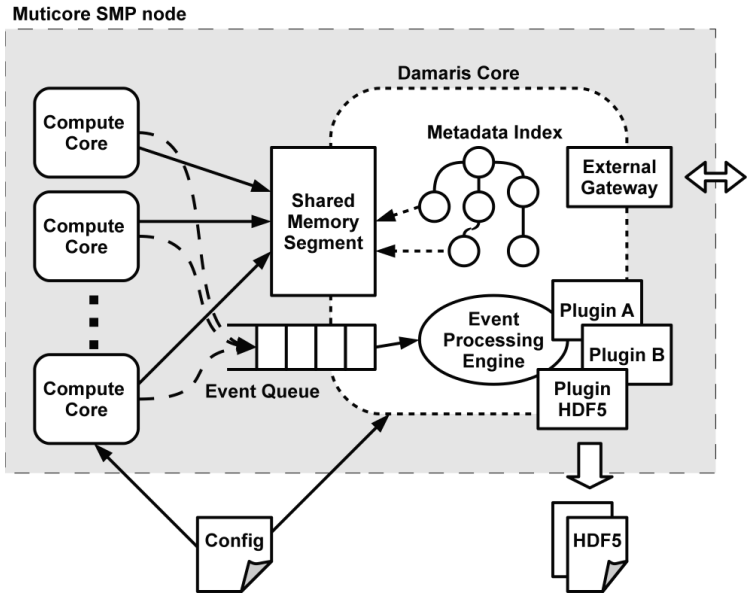
Idea : one dedicated I/O core per multicore node

Originality : shared memory, asynchronous processing

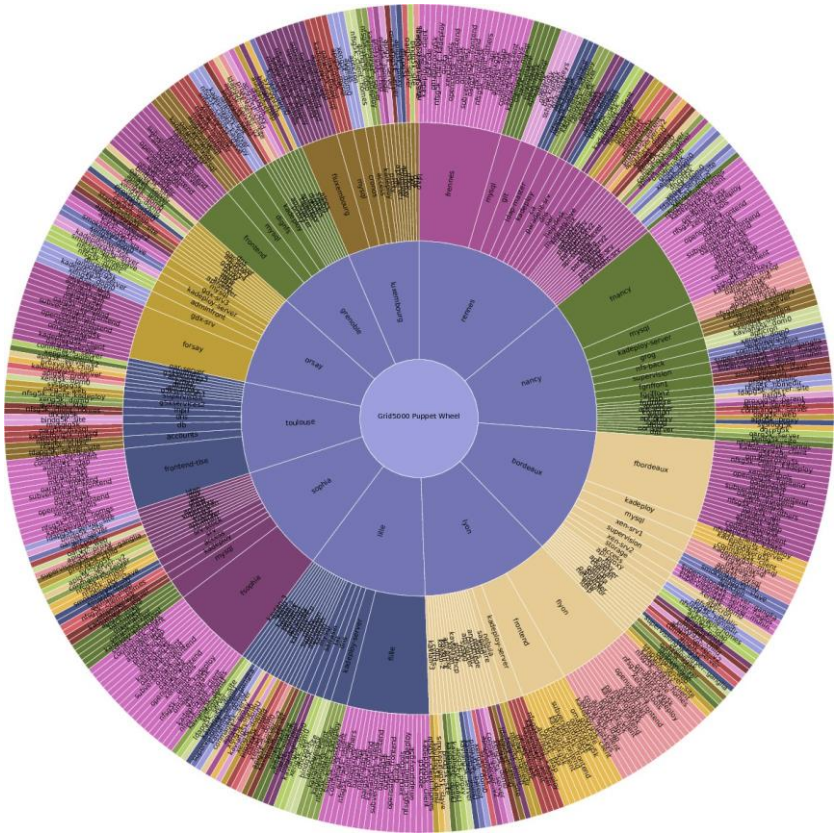
Implementation: software library

Applications: climate simulations (Blue Waters)

Preliminary experiments on Grid'5000



http://damaris.gforge.inria.fr/



CONCLUSIONS

Conclusions

- Computer-Science is also an experimental science
- There are different and complementary approaches for doing experiments in computer-science
- Computer-science is not yet at the same level than other sciences
- But things are improving…

Conclusions

- Computer-Science is also an experimental science
- There are different and complementary approaches for doing experiments in computer-science
- Computer-science is not yet at the same level than other sciences
- But things are improving…

- Grid'5000: a test-bed 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
  - Energy consumption monitoring
- Useful and used platform
  - More than 750 publications with Grid'5000 in their tag (HAL)
  - Between 500 and 600 users per year since 2006

In 2016:

Grid’5000 school

Grenoble, February 2-5

QUESTIONS ?

Special thanks to
G. Antoniu, F. Desprez,
Y.Georgiou, D. Glesser, A. Lebre,
L. Lefèvre, M. Liroz, D. Margery,
L. Nussbaum, C. Perez,
L.Pouillioux
and the Grid'5000 technical team

www.grid5000.fr