ALADDIN-G5K

The Future of Grid'5000

David Margery

Directeur technique, ALADDIN-G5K

David.Margery@inria.fr

(talk by Lucas Nussbaum)

INRIA

The big picture

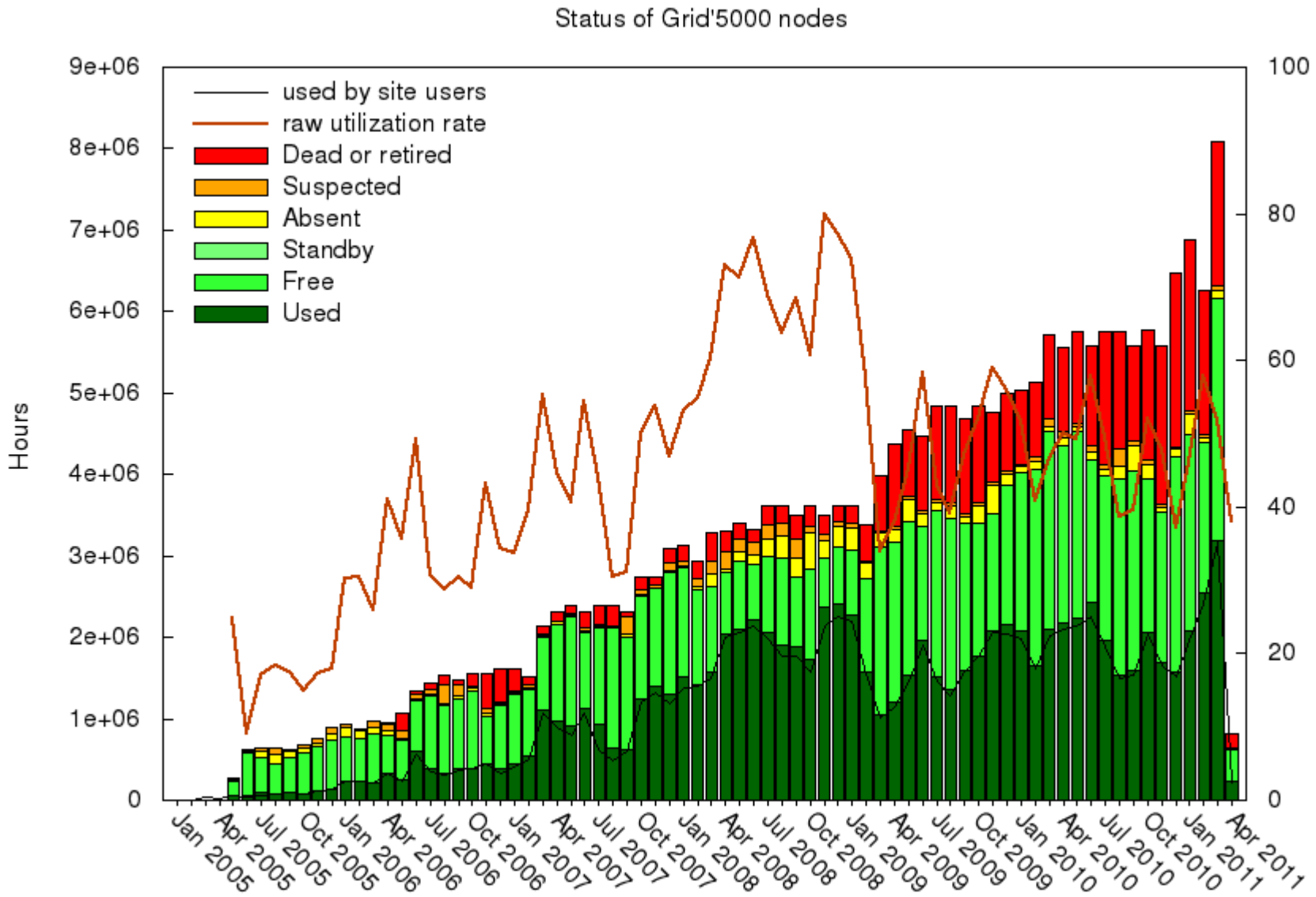
« The need for scientific tools to support experiment-driven research will not disappear »

- Four Grid'5000 sites in need of new hardware (Lyon (07/06), Orsay (05/06), Toulouse (11/07) and Bordeaux (11/07))
  - Executive Committee in active discussions about how this can be financed
    - Grid'5000 might evolve
- Two sites working on technical integration
  - Reims and Luxembourg
  - Porto Alegre not integrated in the shared admin, therefore laging
  - Nantes active candidate for a new site

Key figures

- https://www.grid5000.fr/stats_and_graphs/
- Active users since the beginning of Grid'5000 (1487 differents users)
  - Active users in year 2006 (362 different users)
  - Active users in year 2007 (380 different users)
  - Active users in year 2008 (487 different users)
  - Active users in year 2009 (542 different users)
  - Active users in year 2010 (585 different users)
  - Active users in year 2011 (307 different users so far this year)
- Spring school participants
  - 2006: 100 registered participants
  - 2009: 72 registered participants
  - 2010: 80 registered participants
  - 2011: 68 registered participants

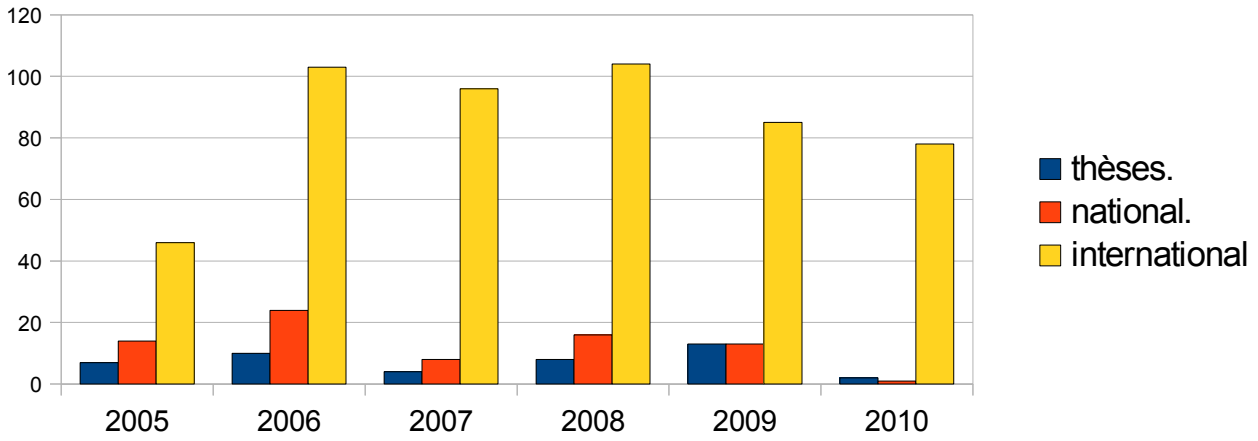
Usage


Status of Grid'5000 nodes
used by site users
raw utilization rate
Dead or retired
Suspected
Absent
Standby
Free
Used
Hours
Raw utilization rate
9e+06
8e+06
7e+06
6e+06
5e+06
4e+06
3e+06
2e+06
1e+06
0
100
80
60
40
20
0
Jan 2005
Apr 2005
Jul 2005
Oct 2005
Jan 2006
Apr 2006
Jul 2006
Oct 2006
Jan 2007
Apr 2007
Jul 2007
Oct 2007
Jan 2008
Apr 2008
Jul 2008
Oct 2008
Jan 2009
Apr 2009
Jul 2009
Oct 2009
Jan 2010
Apr 2010
Jul 2010
Oct 2010
Jan 2011
Apr 2011

Publications

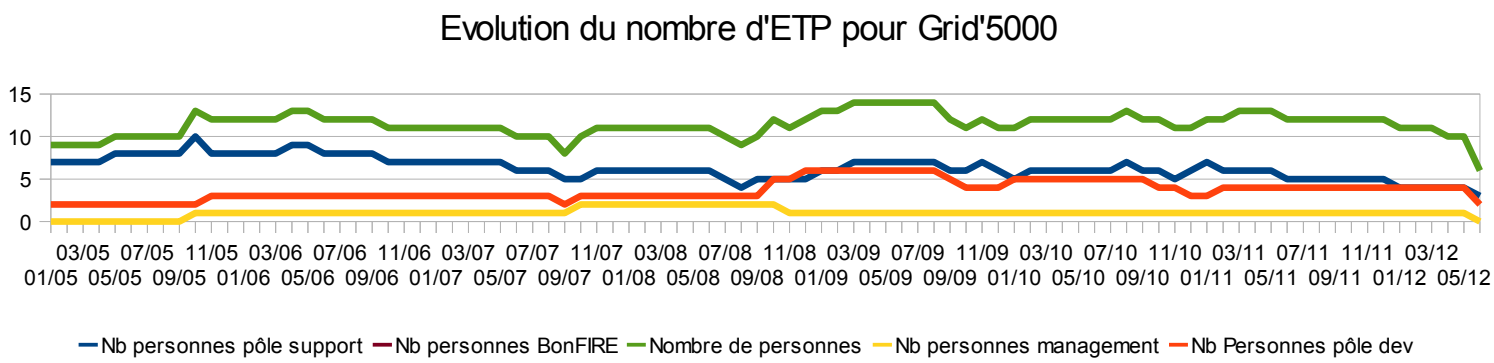
https://www.grid5000.fr/mediawiki/index.php/Special:G5KPublications

- Users are strongly encouraged to upload a description of any publication with results coming from their usage of Grid'5000.
  - Indexed :
    - 544 international publications
    - 79 national publications
    - 45 PHDs, 7 HDR
  - Profil



Human resources

- In total, since 1/01/2005, 721.4 M.Months
  - 441,8 M.M for sysadmin
  - 202 M.M for development
  - 70,6 M.M for management
  - 7 M.M for European project
- By partner, since 1/1/2005
  - 549,4 M.M provided by INRIA
  - 71 MM by université de Rennes 1, 41 M.M by ENS Lyon



Outline

- Some facts
- Planned work in the technical team
- Focus on the network

List of work in progress

- Experiment campaign management
- Storage and Quota management
- Energy Saving
- Reporting
- User accessible recipes
- Virtualisation on production images
- Faster first deployment
- Up-to-date reference images
- Unique global ssh gateways
- Shorter delays to bring nodes back up
- Experiment engines (experimental)

Network related stuff under development

(special focus of second part of talk)

- Metrology of inter-site links
- Bandwidth on demand between sites
- Network Golden rules and Kavlan on all sites
- Reference description of the network
- Ips for virtual machines and subnet reservation

Experiment campaign management

- Problems to solve
  - Best-effort usage requires cooperation between best-effort campaigns
  - Best-effort usage not compatible with parallel jobs (complete job lost if one node lost)
  - No tools to help users supervise a complete set of experiments made of multiple jobs
  - No tools to help users use Grid'5000 no more than X%
  - Submitted jobs might run out of chart even if submitted friday evening, because user has no control on scheduling
- Ideas
  - Build on the experience gained from cigri
  - Offer a higher level submission point than OAR

Storage and Quota management

- Quota management
  - Today
    - Quota is managed globally and approved by the manager of the user
    - Hard limit used to save the site from 'disk full' from over consumption by a single user
  - Tomorrow
    - Quota space managed storage space by storage space
    - Hard limit linked to the space left on the storage space
    - Quota extensions approved by the storage manager
- Storage
  - Today : no other storage space than home directories
  - Tomorrow :
    - reservable (for periods in weeks) storage spaces, in 500G blocks ?
    - Deployable storage clusters ?
    - Uses cases please!

Energy Saving

- Today
  - In Rennes, some nodes are shutdown to save energy
    - The number of nodes kept alive is important
  - Some Absent nodes in standby mode
  - Not all tools are ready for the standby mode
  - Users can shut nodes down/bring them up with kapower3 on all sites
- Tomorrow
  - All sites implement energy saving
  - The number of nodes kept alive even if not used will decrease

Reporting

- Today
  - User Management Service (UMS) managing accounts
  - User report managing experiment description and publication references
  - OAR/Kaspied keeping usage records
- Tomorrow
  - UMS and User reports reconciled
  - Experiment description and oar linked ?
    - On a voluntary base, using the -p <project> option of OAR
    - After initial resource consumption has reached a cap
  - Kaspied and other statistics linked to affiliation described in UMS
    - We are missing the info about which institution is using Grid'5000

User accessible recipes

- Today
  - Grid5000-code used so that users can share there tools
  - Staff code kept in the git.grid5000.fr codebase, away from users
- Tomorrow
  - Legaleese sorted out for code and recipe sharing
  - Community contributions and creation around specific topics easier
  - Grid5000.github.com as a tool for sharing know-how

Virtualization on production images

- Today
  - Users need to deploy to start a virtual machine
- Tomorrow
  - Users could start kvm commands on the production environment
  - The study for this was user contributed

Faster first deployment

- Today
  - Kadeploy3 cannot use the production environment as a deployment environment
    - One reboot cycle is used to boot into a deployment environment
- Tomorrow
  - The production environment could be blessed as a working deployment environment
    - First kadeploy3 command half as long

Up-to-date reference images

- Until Today
  - Reference image manually generated and kept up-to-date
  - Outdated reference images still very visible
- Today
  - Reference images generated
    - Lenny 2.3 on all sites
    - Squeeze nearly ready
  - New policy for the contents of these reference images
- Tomorrow
  - User access to recipes used to generate the images
  - Shorter cycles to update the images

Unique global ssh gateways

- Today
  - An access machine on each site
    - Not always open to the world
    - With varying configurations
      - Not always correct
      - Not always remotely logged/automatically updated
- Tomorrow
  - An access.grid5000.fr
    - With redundant network access (north/south)
    - With high availability links between 2 implementations
    - With a hardened security configuration
  - With open issues
    - How ssh keys will be uploaded to these machines
    - With what homedir for users, to transfer files in and out of Grid'5000

Shorter delays to bring nodes back up

- Today,
  - Nodes are checked for conformance with their description in the API
    - Still some properties to check
  - Nodes are checked for correct filesystem
    - Nfs mounts
    - Partition scheme

    Bad nodes put in suspected state
  - The phoenix script detects nodes stuck in Absent or suspected state
    - One attempt to detect/correct the problem every 15mn
- Tomorrow
  - Phoenix could become a deamon
    - Delays to bring resource back online would be shortened, as it could be notified of the nature of the problems and react faster

Experiment engines

- Past: Expo [Videau], NXE [Guillier]
- Other testbeds are better than us: Emulab, Plush (PlanetLab), GENI
- Recent attempt by the technical team: G5Kcampaign
http://g5k-campaign.gforge.inria.fr/
- Still an open research question
- More work to be expected in this area in order to be able to perform large-scale and/or complex experiments:
  - Hemera WG Methodology
  - Hemera Challenges

The BonFIRE European project

- INRIA, for Grid'5000, is a partner of the BonFIRE IP projects, lasting 42 Months
  - Building service testbeds for Future Internet Research and Experimentation
    - The Project will design, build and operate a multi-site cloud facility to support applications, services and systems research targeting the Internet of Services community within the Future Internet.
  - 10 old paraquad machines in Rennes part of the testbed
  - Possible benefits for Grid'5000
    - Tools for experiment-driven research ?
    - OCCI or higher level access to Grid'5000/Experience in cloud based testbeds
    - Links to other testbeds and participation in the European testbed community

Links to other projects/testbeds ?

- No evident business model for that
  - Interconnecting costs, in hardware and in personnel
- No user-driven use-cases
  - Last year, nobody answered the call for a Grid'5000/planetlab Use-Case
- Promising contacts with FutureGrid
  - Reciprocal accounts policy, joint workshop planned
- Please come with use-cases !!!
  - ANR PetaFlow is using Grid'5000 to link to Japan
  - ANR DALIA has used Grid'5000 to link virtual reality hardware from Grenoble, Bordeaux and Orleans.
  - ...

Conclusions

- Future work should always be discussed with users on devel@lists.grid5000.fr
  - This is the way to get involved and influence development
  - The minority who speaks influences Grid'5000 to ease their own work
- A lot to do in parallel for the technical team
  - New clusters, new sites, hardware failures, maintenance operations (electric, cooling, etc...)
  - User support, bugs and corner cases, whose solution must be propagated to 9 (soon 11 or 12 ?) sites
  - Developing the new stuff and putting it into production
  - Manage renewal of participants in the technical team
  - Document, update the wiki and the tutorials

Yet to find a good way to provide an overview of the ongoing work

- What seems simple to test and implement on one cluster is never simple on 23 clusters/9 sites. Patience is required from users.