

# Examples of SAAS on cloud: dynamically scaling R, Galaxy and Matlab

[tom.visser@sara.nl](mailto:tom.visser@sara.nl)

[linkedin.com/in/visservisser](https://www.linkedin.com/in/visservisser)

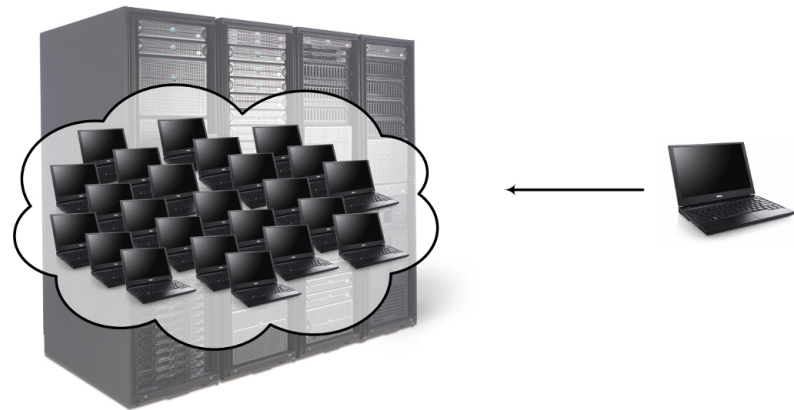
## **BiG**Grid

*the dutch e-science grid*

# Our vision

Cloud computing is not about new technology, it is about new uses of technology

Self Service Dynamically Scalable  
Computing Facilities



# Our experience in a nutshell\*



People bring their existing problems and ideas

They are creatively stimulated

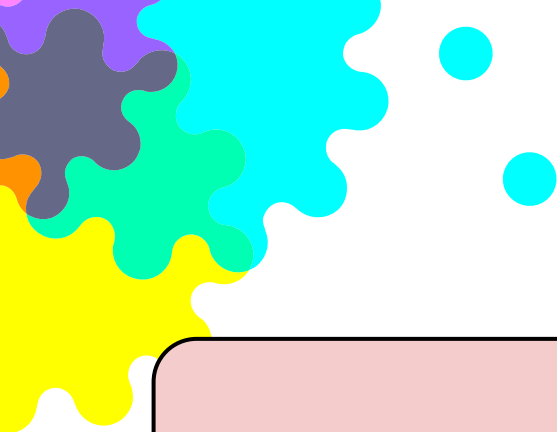
A world of new possibilities. A proof of concept environment

Very popular offer

\* see Munich march 2012 presentation <https://indico.eji.eu/indico/contributionDisplay.py?contribId=39&confId=679>

**BiG Grid**  
the dutch e-science grid





# Self service, how do you support that?

Software as a service

We are an e-science group, so why stop at self service...

Platform as a service

Hands-free to experiment with user communities

Infrastructure as a self-service



**BiG Grid**

the dutch e-science grid

# Preconfigured VM wizard

The screenshot shows the SARA HPC-Cloud web interface. The browser address bar displays <https://ui.cloud.sara.nl>. The page has a dark sidebar on the left with navigation links: Dashboard, Virtual Machines, Templates, Images, **Create VM** (highlighted), Network filter, Users, and Upload images. At the bottom of the sidebar is a 'Project Quota Usage' gauge showing 75% usage. The main content area is titled 'SARA HPC-Cloud' and includes links for Documentation, Support, Community, and a Welcome message. The wizard consists of four steps: 1 System (Operating system), 2 Size (Disk, CPU and memory), 3 Internet (Internet and Services), and 4 Overview (Configuration summary). The '1 System' step is active, showing a 'Choose Operating System' section with two options: Linux - CentOS and Linux - Ubuntu. At the bottom of the wizard, there is a link to 'cloud documentation' and 'Previous' and 'Next' buttons.

Dashboard  
Virtual Machines  
Templates  
Images  
**Create VM**  
Network filter  
Users  
Upload images

Project Quota Usage:  
75%

SARA HPC-Cloud Documentation | Support | Community Welcome

**1 System** Operating system  
**2 Size** Disk, CPU and memory  
**3 Internet** Internet and Services  
**4 Overview** Configuration summary

**Choose Operating System**

- Linux - CentOS
- Linux - Ubuntu

For advanced options see [cloud documentation](#).

Previous Next



# Cases for today

- Galaxy workflow server (slides thanks to Leon Mei & Matthias den Hollander at NBIC / NIOO)
- Using R the statistical software
- Matlab (MDCS and parallel computing toolkit).

# NBIC Galaxy server

The screenshot shows the NBIC Galaxy server interface. The browser address bar displays 'galaxy.nbic.nl/galaxy/'. The top navigation bar includes 'File', 'Edit', 'View', 'History', 'Bookmarks', 'Tools', and 'Help'. Below this, a dark blue bar contains 'Galaxy / Netherlands B' and various tool categories: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. A status bar on the right indicates 'Using 4%'. The main interface is divided into three panels:

- Tools panel (left):** A list of tools categorized by type. A green box highlights the 'Tools' section, which includes links like 'NGS: Peak Calling', 'NGS: Simulation', 'SNP/WGA: Data; Filters', 'SNP/WGA: QC; LD; Plots', 'SNP/WGA: Statistical Models', 'Human Genome Variation', 'Genome Diversity', 'NGS: VCF Tools', 'NGS: Bedtools', 'NGS Taskforce: Hubrecht - Alignment tool benchmarking', 'NGS Taskforce: WUR denovo benchmarking', 'NGS Taskforce: LUMC - GAPSS v2', 'NGS Taskforce: LUMC - GAPSS v3', 'NGS Taskforce: LUMC - deepSAGE', 'PC: msCompare', 'PC: Proteomics Tools', 'Taverna Workflows', 'Test tool for CTMM TraIT', and 'Workflows'. The text 'NBIC Tools' is overlaid in green.
- Control panel (center):** A large red box highlights the main workspace. It features a 'Welcome to Sombbrero - the NBIC Galaxy' message, a workflow diagram titled 'WWFSMD? grow noodly appendages...' from 'usegalaxy.org', and a notice about server hosting by the Netherlands Bioinformatics Centre. Below the notice, a list of quotas is provided: 'For registered users, you will get a 10GB quota' and 'For anonymous users, you will get a 10MB quota'. The text 'Control panel' is overlaid in red.
- History panel (right):** A red box highlights the 'History' panel, which shows 'Unnamed history' with '0 bytes'. A message states: 'Your history is empty. Click "Get Data" on the left pane to start'. The text 'History panel' is overlaid in red.

NBIC  
Tools

Control panel

History  
panel

BiG Grid

the dutch e-science grid



# Strong User Community

Galaxy is widely used for analyzing Next Generation Sequencing data

- PennState University, BSD like license
- Very active user community (about 200 participants to Galaxy Community Conference in 2012, and 150 in 2011)

**galaxy.nbic.nl**

- Started in 2010
- >240 registered users
- Used in a number of courses, trainings



**BiG Grid**

the dutch e-science grid

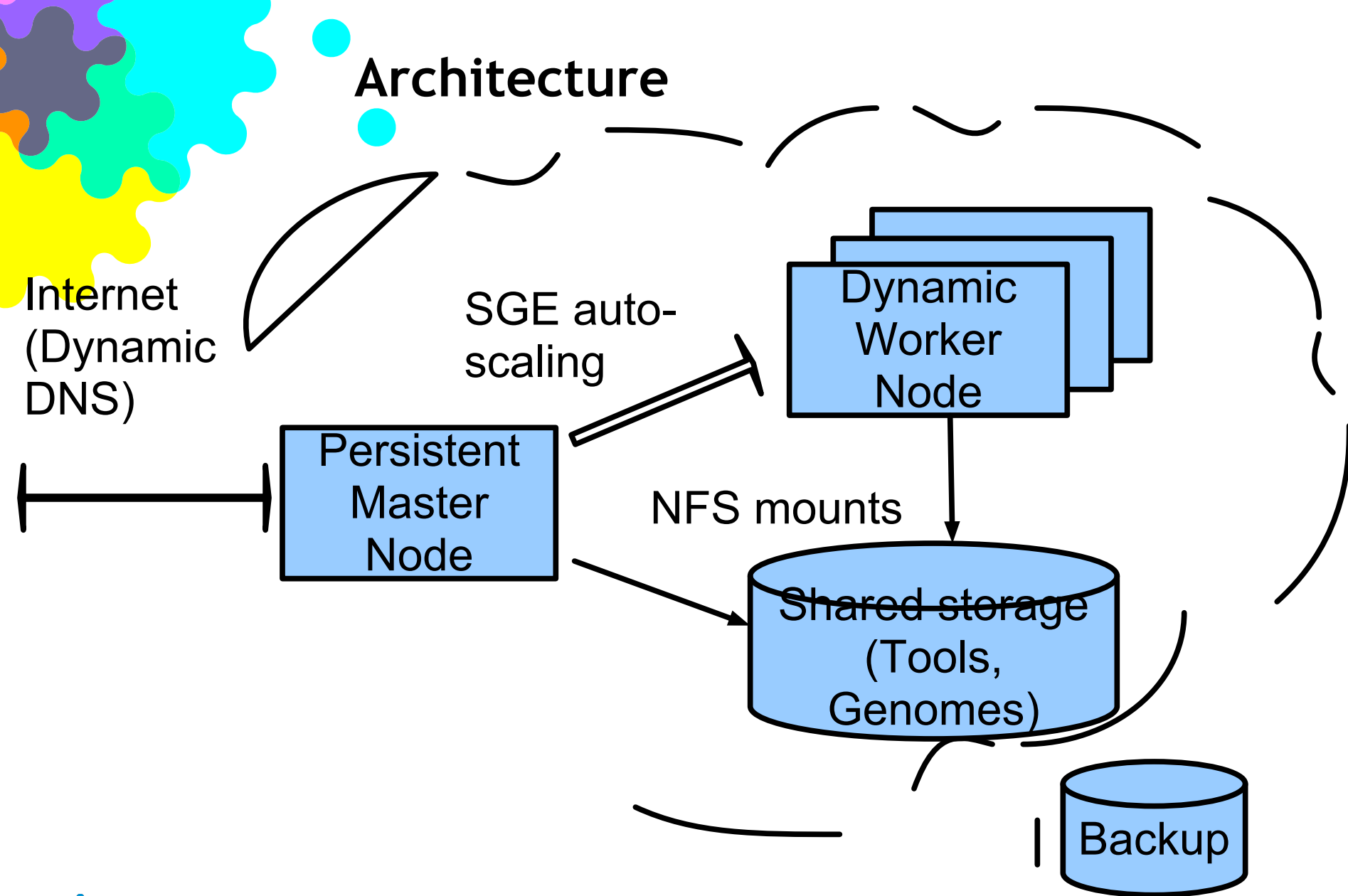




# NBIC Galaxy @HPC Cloud

- Project started in July 2012
- Supported by BiG Grid, SARA, NBIC, NIOO
- Planned launch in September 2012
- Aim to reach 500 users by the end of 2012
- Will be used as the base for other project-specific Galaxy server deployments in the HPC cloud

# Architecture



**BiG Grid**

the dutch e-science grid



# Tool Installation Automation

- . <http://usecloudman.org>
  - Developed by the PennState Galaxy team
  - MIT license
  - Support Amazon EC2 and OpenStack
- . Fabric installation scripts
  - Galaxy itself
  - Postgres
  - Sun Grid Engine
  - Common NGS tools, e.g. BWA, bowtie, samtools, etc.



**BiG Grid**

the dutch e-science grid



# Data Installation Automation

- <http://cloudbiolinux.org/>
  - Developed by a team consists members from Harvard Univ., J. Craig Venter Institute, the Galaxy team
  - MIT license
- Fabric installation script
  - Common genome builds, hg18, hg19, mm9, tair10, etc.
  - Tool specific genome indexes for bowtie, BWA, etc.tc.



**BiG Grid**

the dutch e-science grid

# Using R

- Using R for transcriptomics
  - Statistical package -> <http://www.r-project.org/>
  - Existing cluster installation
  - Ported to cloud via always-on headnode spawning R workers, ref: Han Rauwerda and Timo Breit
    - <http://www.ebiogrid.nl/generic-infrastructure.html>
    - <http://www.biggrid.nl/hpc-cloud-day-4-october-2011/>
- Using R (2) for economics analysis
  - R studio project - [r-studio.org](http://r-studio.org)
  - Interactive R session via web-browser on big virtual machine; work done by Lykle Voort at SARA.nl see also: <https://www.cloud.sara.nl/projects/ceff>



**BiG Grid**

the dutch e-science grid



# Matlab distributed computing service

- Preliminary investigations with Mathworks company using their MDCS solution
- Scenario; user has valid client license for parallel computing toolbox and spawns x workers on cloud cluster of service provider (BiG Grid)
  - licensing issue is fixed this way ??
  - user can work in existing environment
  - enormous potential because of broad user-base for matlab
- multi tenancy / accounting and dynamic scaling still have to be solved



**BiG Grid**

the dutch e.science grid



# Some observations

- Scientist have there own preferred tools and ways of working
  - There's a lot of hidden programmers / technically skilled *and or* ambitious people out there
  - Labs and institutes have there own clusters and computing and solutions
  - Local ICT departments are less facilitating in experimentation -> limited capacity

You can seduce and enable the scientific community by offering this type of infrastructure and striving for proper **integration**



**BiG Grid**

the dutch e-science grid



# Outlook / thoughts

- Coping with autoscaling mechanisms
- Data-locality -> does cloud solution actually fit the problem?
- Integration, automation.
- Infrastructure should be there; without friction! IAAS!
- Offering cluster solutions to the masses but there's no such thing as an infinite resource..
- Computing resources cost money, can we share the investments?



**BiG Grid**

the dutch e-science grid





**BiG Grid**  
the dutch e-science grid

# References

Galaxy - <http://galaxy.nbic.nl> - <https://www.cloud.sara.nl/projects/mattiasdehollander-project/wiki>

R cluster setup - [http://www.biggrid.nl/fileadmin/documents/Han\\_Rauwerda\\_HPC\\_Cloud\\_Day\\_20111004.pdf](http://www.biggrid.nl/fileadmin/documents/Han_Rauwerda_HPC_Cloud_Day_20111004.pdf)

R-studio <http://rstudio.org/>

BiG Grid project <http://www.biggrid.nl>

SARA <http://www.sara.nl>

Fed cloud task force <https://wiki.egi.eu/wiki/Fedcloud-tf:FederatedCloudsTaskForce>

EGI Tech Forum 2012 BiG Grid presentation

<https://indico.egi.eu/indico/contributionDisplay.py?sessionId=12&contribId=50&confId=1019>

EGI Community Forum presentations on BiG Grid cloud

<https://indico.egi.eu/indico/contributionDisplay.py?contribId=39&confId=679>

Picture courtesy <http://www.flickr.com/photos/wongjunhao/424826584/> ,  
<http://www.flickr.com/photos/rummenigge/2225696954/>



**BiG Grid**

the dutch e-science grid