



Netherlands
Bioinformatics
Centre

NBIC Galaxy@HPC Cloud

Mattias de Hollander
David van Enckevort
Leon Mei
Rob Hooft



SURFsara HPC Cloud

- 19 nodes, 32 cores and 256 GB RAM each
 - Intel 2.13 GHz 32 cores (Xeon-E7 "Westmere-EX")
- 400 TB storage in total
- Nodes are connected by four 10 Gb interfaces to a non-blocking switch
- So maximum 40 GB access to storage (per node)
- <https://www.surfsara.nl/systems/hpc-cloud>

Cloud Management Console

File Edit View History Bookmarks Tools Help

HPC-Cloud Sara - Web user int... +

https://ui.cloud.sara.nl

SARA HPC-Cloud Documentation | Support | Community Welcome halliang.mei%40nbic.nl | Sign out

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

1 System
Operating system

2 Size
Disk, CPU and memory

3 Internet
Internet and Services

4 Overview
Configuration summary

Configure system

System size

Small 1 CPU 8 GB memory	Medium 4 CPU 32 Gb memory	Large 8 CPU 64 Gb memory	Extra large 16 CPU 128 Gb memory
--------------------------------------	--	---------------------------------------	---

Disk size

5 Gb	10 Gb	25 Gb	50 Gb
-------------	--------------	--------------	--------------

Project Quota Usage: 51% used

51
51%

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

Previous Next Create VM

Copyright 2002-2012 © OpenNebula Project Leads ([OpenNebula.org](#)). All Rights Reserved. OpenNebula 3.2.1

NBIC Galaxy Server

The screenshot shows the NBIC Galaxy Server interface in Mozilla Firefox. The browser address bar shows `http://galaxy.nbic.nl/galaxy/`. The interface is divided into three main panels:

- NBIC Tools (Left Panel):** A list of tools under the 'Tools' tab. The 'NGS: Tools LUMC' category is expanded, showing a list of tools including 'Map with Bowtie for Illumina', 'GAPSS - FASTA to FASTQ', 'GAPSS - FASTQ to FASTA', and 'GAPSS - SCARF to FASTQ'. The 'Map with Bowtie for Illumina' tool is highlighted with a green box.
- Control panel (Center Panel):** The configuration interface for the 'Map with Bowtie for Illumina' tool. It includes the following settings:
 - Reference genome: 'Human_UCSC_hg19_complete' (selected from a dropdown)
 - Library type: 'Single-end' (selected from a dropdown)
 - FASTQ file: '22: FASTQ Groomer on data 2' (selected from a dropdown)
 - Bowtie settings: 'Commonly used' (selected from a dropdown)
 - Suppress the header in the output SAM file:
 - Output in SAM format:
 - An 'Execute' button is visible at the bottom.
- History panel (Right Panel):** A list of recent jobs. The top job is '38: GAPSS - FASTQ to FASTA on data 22'. Other jobs include '37: GAPSS - FASTQ to FASTA on data 22', '36: Map with Bowtie for Illumina on data 22', '35: Map with Bowtie for Illumina on data 22', '34: Map with Bowtie for Illumina on data 22', '31: VarScan - pileup2snp on data 30', '30: Generate pileup on data 29', '29: SAM-to-BAM on data 28', and '28: Map with Bowtie for Illumina on data 22' (308 lines, format: sam, database:).

Share Scripts & Pipelines

- GAPSS (Variant calling, annotations)
 - DeepSAGE
 - RNA-seq
 - msCompare
 - Chip-seq
 - Circos visualization
 - ...
-
- <https://trac.nbic.nl/galaxytools/>

Portal To Other Resource

The screenshot shows the Galaxy / NBIC web portal in a Mozilla Firefox browser. The browser's address bar shows the URL `galaxy.nbic.nl`. The page title is "Galaxy / NBIC". The navigation menu includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Admin", "Help", and "User". A status bar at the top right indicates "Using 6.0 Gb".

The main content area is titled "Cluster Status" and displays a table with the following data:

Chromosome	Job ID	Status	Progress	Runtime (min)
3	7121414	Running	<div style="width: 23%; background-color: green; height: 10px;"></div> 23%	00:14

Below the table, the job status is summarized: "Job status (1 running, 0 queued, 0 finished, 1 total) (data is refreshed every 90 seconds)".

The left sidebar contains a "Tools" panel with various tool categories, including "VCF Tools", "NGS: Bedtools", "NGS Taskforce: Hubrecht - Alignment tool benchmarking", "NGS Taskforce: WUR denovo benchmarking", "NGS Taskforce: LUMC - GAPSS v2", "Haplotype sharing", "NGS Taskforce: LUMC - GAPSS v3", "NGS Taskforce: LUMC - deepSAGE", "NPC: msCompare", and "CTMM TraIT". The "GRONCROSS" tool is highlighted with a mouse cursor.

The right sidebar contains a "History" panel showing a list of recent jobs:

- genallice meeting (7.1 Mb)
- 104: GRONCROSS Combined Output on data 102 and data 91 (Job is currently running)
- 103: GRONCROSS on data 102 and data 91 (html report) (Job is currently running)
- 102: Logfile from data 91 (2 lines, format: tabular, database: ?)

The browser's address bar at the bottom shows the URL `galaxy.nbic.nl/tool_runner?tool_id=CROSS`.

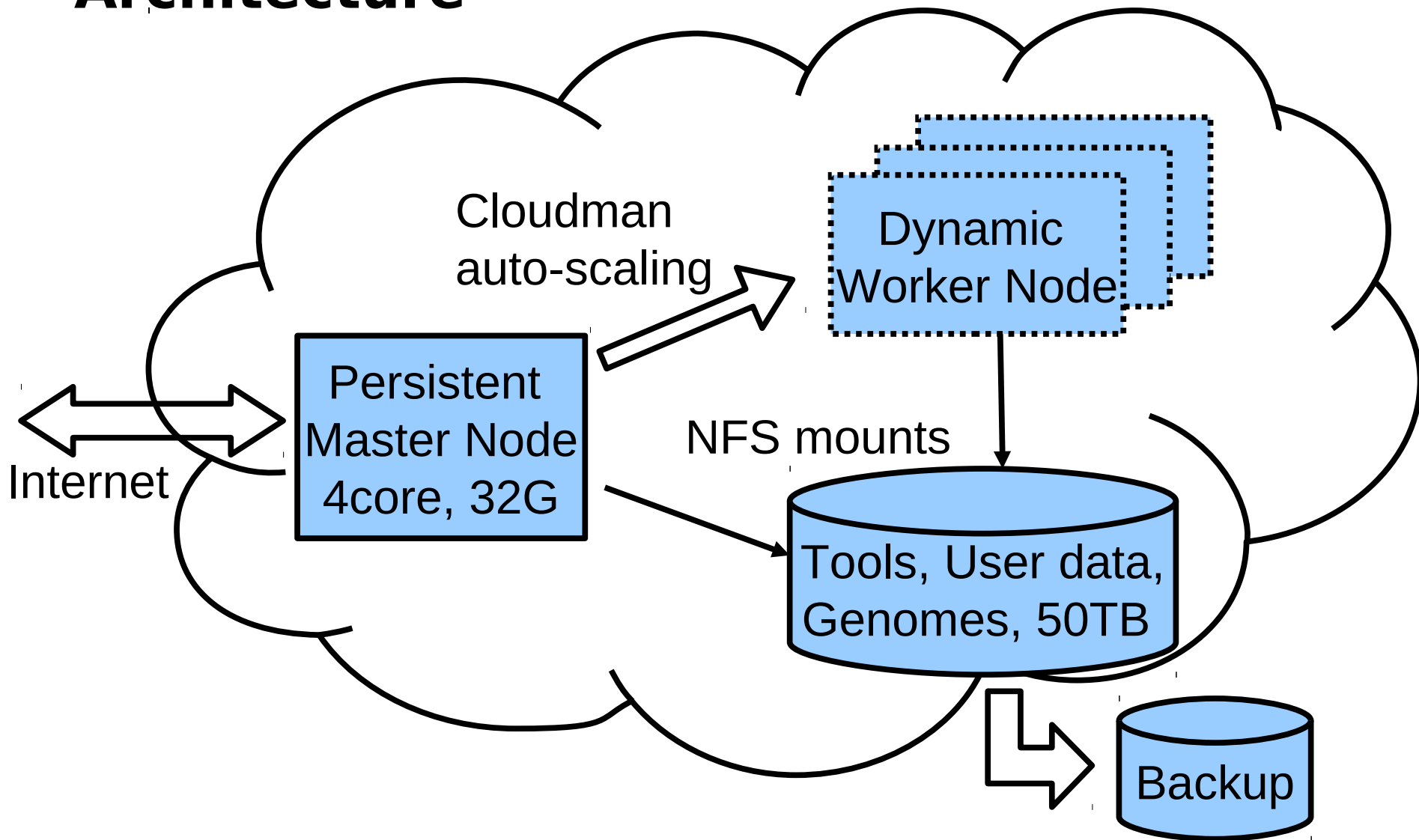
Strong User Community

- 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)
- Galaxy2.nbiceng.net
 - Started in 2010
 - >240 registered users
 - Used in a number of courses
 - However, only a demo server

Migration to Cloud

- Galaxy2.nbiceng.net
 - 4 CPU, 24G RAM, 1.5T HD, 100Mbit Internet
- Migration project started in July 2012
- Supported by BiGGrid, SURFsara, NBIC, NIOO
- Production server launched in September 2012
- Lightpath connecting to Cloud
- Will be used as the base for other project specific Galaxy servers in the HPC cloud
 - TraIT NGS Galaxy will contain several CompleteGenomics specific tools
 - NBIC Galaxy VM

Architecture



File Structure

- Master VM
 - CloudMan scripts: /mnt/cm
 - SGE: /opt/sge
 - ProFTP: /opt/galaxy/pkg/proftpd
 - Nginx: /opt/galaxy/pkg/nginx
- NFS shared
 - /mnt/galaxyTools: galaxy, 3rd party tools, 2.4GB
 - /mnt/galaxyData: galaxy user data, postgres DB, 784GB
 - /mnt/galaxyIndices: genomes, liftover chains, 72GB

Tool Installation Automation

- CloudMan scripts <http://usecloudman.org>
 - Developed by the PennState Galaxy team
 - Support Amazon EC2 and OpenStack
- CloudMan script customized for OpenNebula
 - Developed by Mattias de Hollander (NIOO)
 - Hosted at <http://downloads.nbiceng.net/cloudman-on>
- Fabric installation scripts
 - Galaxy itself
 - PostgreSQL
 - Sun Grid Engine
 - Common NGS tools, e.g. BWA, bowtie, samtools, etc.

Genomes Installation Automation

- <http://cloudbiolinux.org/>
 - Cloud version of BioLinux
 - Developed by a team consists of members from Harvard Univ., J. Craig Venter Institute, the Galaxy team
 - “Using Cloud Computing Infrastructure with CloudBioLinux, CloudMan, and Galaxy”, June 2012
- Fabric installation script
 - Common genome builds, hg18, hg19, mm9, tair10, etc.
 - Tool specific genome indexes for bowtie, BWA, etc.



Some Hurdles

- Installation
 - Fabric scripts use hardcoded versions, some of them are outdated
 - PostgreSQL version conflict
- Too many layers so hard to tell where is the root cause
 - VM or Cloud
- Running
 - SGE failed to start

Cloud Issues

- Everything on the NFS share is owned by the same user.
However:
 - PostgreSQL requires the ownership of its data directory.
 - CloudMan tries to chown directories.
 - Work-around: disable chown / use idmapd
- Open Nebula issues
 - VIRTIO network driver bug caused network connection unstable
 - Worker node start up problem

DB Migration

- MySQL to PostgreSQL
 - Many tools are difficult to set up or do not work
 - Fortunately, we found `py-mysql2pgsql`

Remaining Issues

- Data transfer bottleneck (~40MB/s) between master/worker node and the NFS storage.
- Installation of additional tools
 - [/usr/local/bin => /mnt/galaxyTools](#)
- Enabling test framework
- NBIC Toolshed

Remaining Issues

- Data transfer bottleneck (~40MB/s) between master/worker node and the NFS storage.
- Installation of additional tools
 - [/usr/local/bin => /mnt/galaxyTools](#)
- Enabling test framework
- NBIC Toolshed

Acknowledgement

- BiGGrid/SURFsara/NBIC
 - Niek Bosch, Marc van Driel, Irene Nooren, Machiel Jansen, the cloud-support team
- NBIC Galaxy admin team