Netherlands
Bioinformatics
Centre

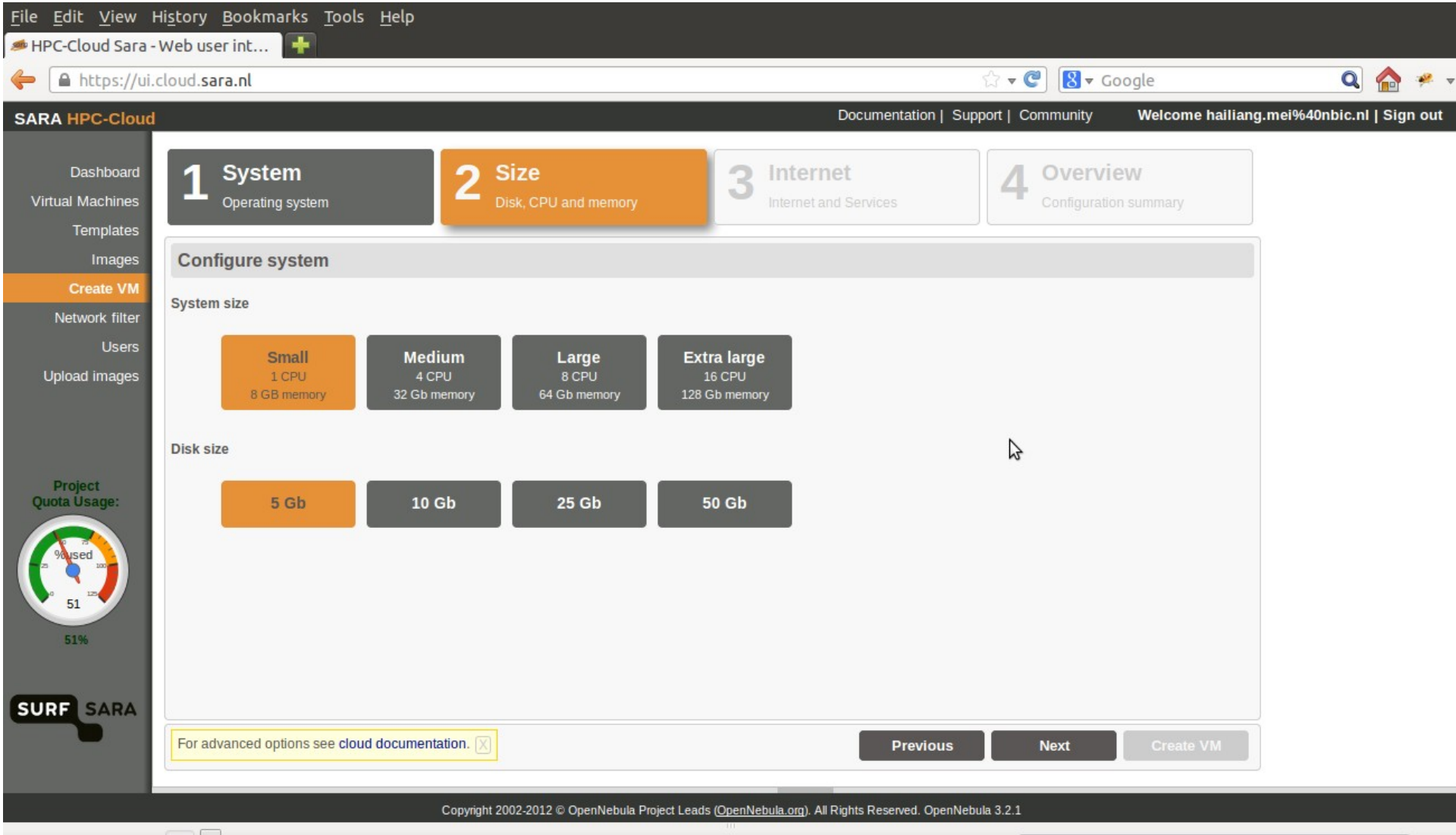
NBIC Galaxy@HPC Cloud

Mattias de Hollander
David van Enckevort
Leon Mei
Rob Hooft

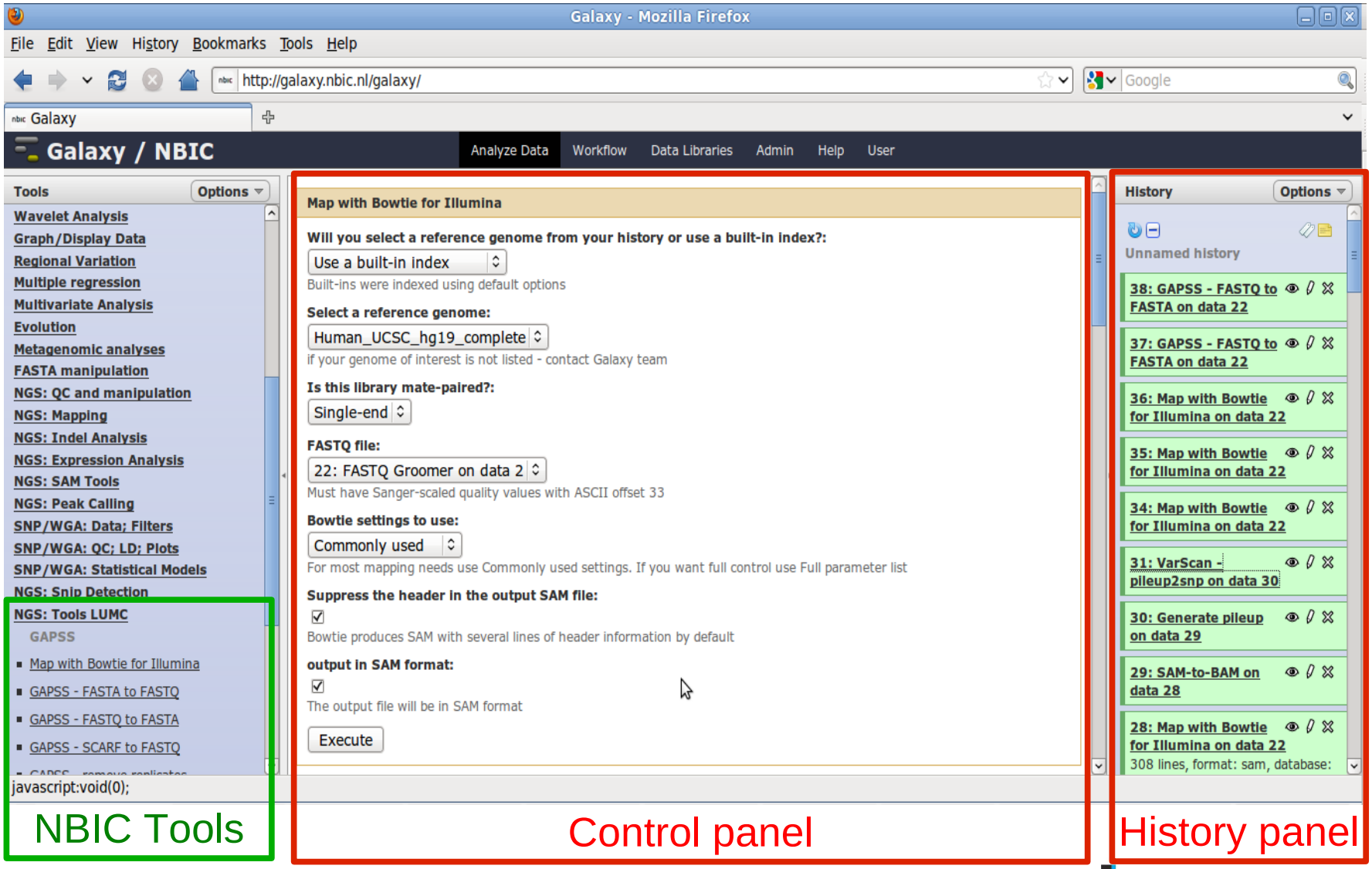
SURFsara HPC Cloud

- 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



NBIC Galaxy Server



Netherlands Bioinformatics Centre

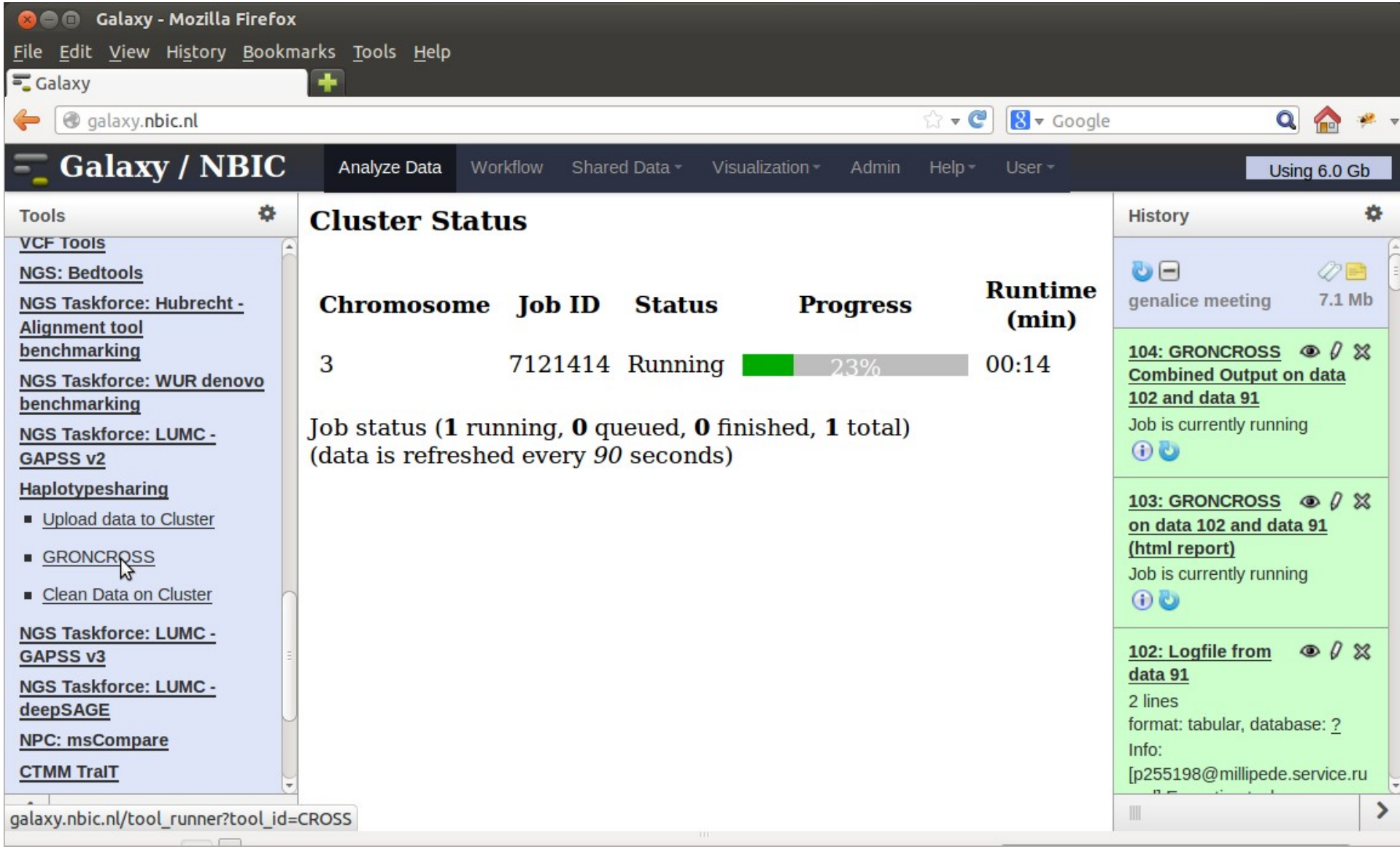
Share Scripts & Pipelines

- GAPSS (Variant calling, annotations)
- DeepSAGE
- RNA-seq
- msCompare
- Chip-seq
- Circos visualization
- …

- https://trac.nbic.nl/galaxytools/

nbic Netherlands Bioinformatics Centre

Portal To Other Resource



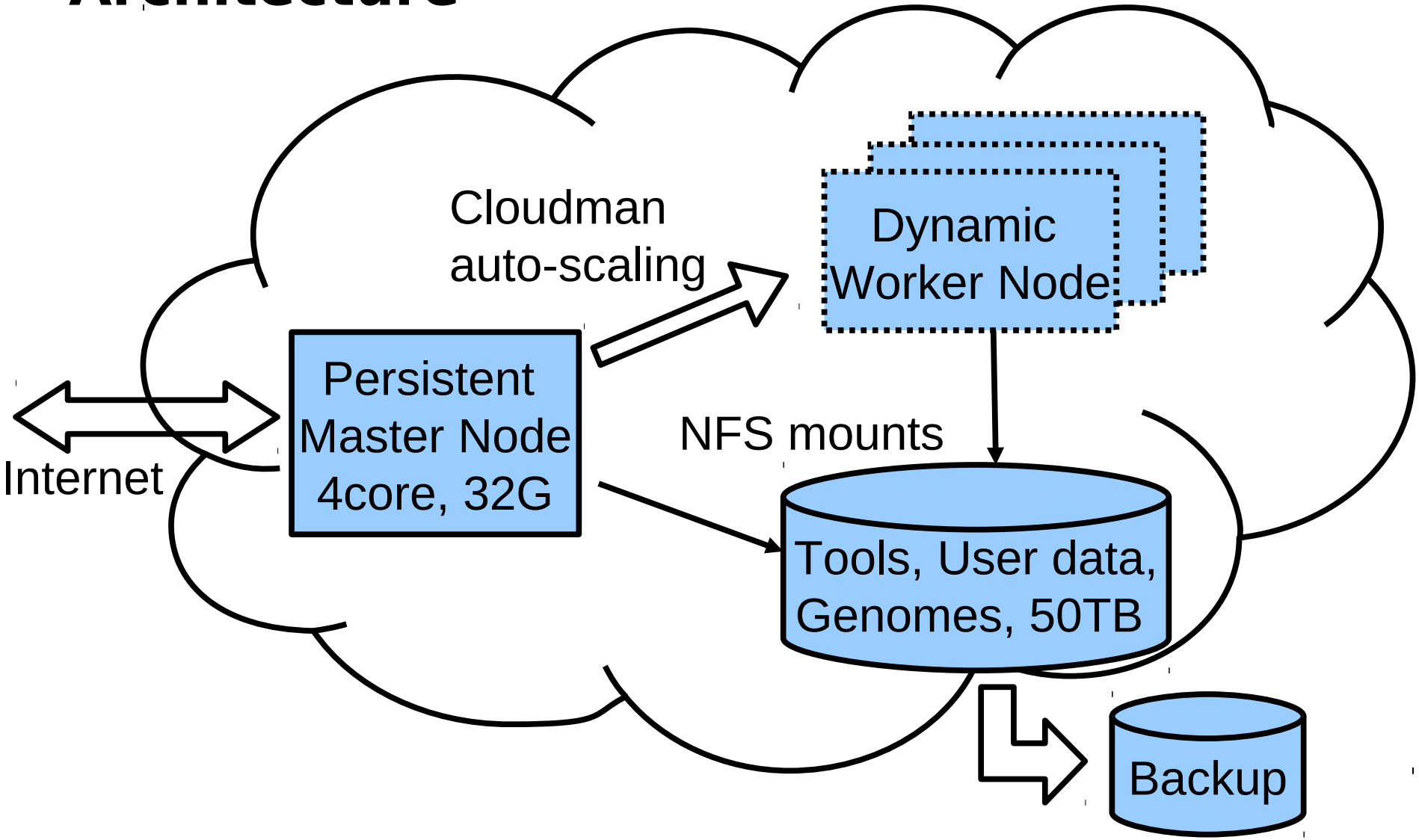
Strong User Community

- 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



nbic Netherlands Bioinformatics Centre

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.

nbic Netherlands Bioinformatics Centre

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

nbic Netherlands Bioinformatics Centre