The Cloud for Biologists

using bioinformatics tools

Mattias de Hollander

Netherlands Institute of
Ecology (NIOO-KNAW)

NIOO

Why choose for the Cloud?

Why choose for the Cloud?

- It's **flexible**

Why choose for the Cloud?

- It's **flexible**
- You have full **control**

Why choose for the Cloud?

- It's **flexible**
- You have full **control**
- Perfect for **small labs**

Why choose for the Cloud?

- It's **flexible**
- You have full **control**
- Perfect for **small labs**
- It's **fancy** (Google and Amazon are using it)

Why choose for the Cloud?

- It's **flexible**
- You have full **control**
- Perfect for **small labs**
- It's **fancy** (Google and Amazon are using it)
- It's **environmental friendly** (Gmail: Its cooler in the cloud)

How do we use the Cloud?

Galaxy

a web-based genome analysis platform[1]

[1] Slide by Anton Nekrutenko, Galaxy Developer Conference 2011, Lunteren (NL)

Galaxy

a web-based genome analysis platform[1]

- A free (for everyone) **web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

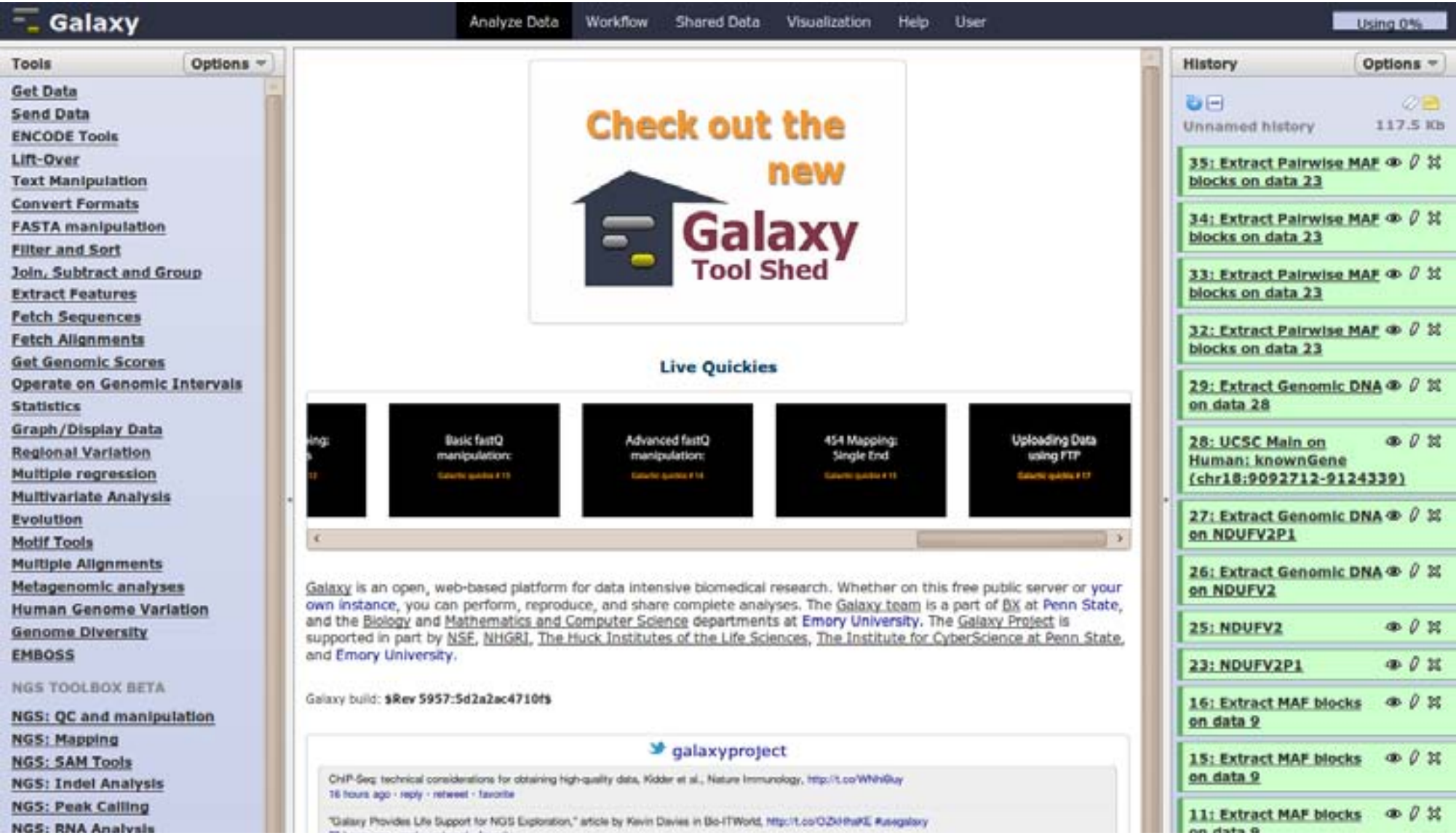
[1] Slide by Anton Nekrutenko, Galaxy Developer Conference 2011, Lunteren (NL)

Galaxy

a web-based genome analysis platform[1]

- A free (for everyone) **web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple

[1] Slide by Anton Nekrutenko, Galaxy Developer Conference 2011, Lunteren (NL)


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
Check out the new
Galaxy
Tool Shed
Live Quickies
Basic fastQ manipulation:
Advanced fastQ manipulation:
454 Mapping: Single End
Uploading Data using FTP
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
Galaxy build: $Rev 5957:5d2a2ac4710f$
galaxyproject
History
Options
Unnamed history
117.5 Kb
35: Extract Pairwise MAF blocks on data 23
34: Extract Pairwise MAF blocks on data 23
33: Extract Pairwise MAF blocks on data 23
32: Extract Pairwise MAF blocks on data 23
29: Extract Genomic DNA on data 28
28: UCSC Main on Human: knownGene (chr18:9092712-9124339)
27: Extract Genomic DNA on NDUFV2P1
26: Extract Genomic DNA on NDUFV2
25: NDUFV2
23: NDUFV2P1
16: Extract MAF blocks on data 9
15: Extract MAF blocks on data 9
11: Extract MAF blocks

Most biologists don't write code

Most biologists don't write code

- Analyze
  - Interactively **manipulate** genomic data with a comprehensive and expanding 'best-practices' toolset

Most biologists don't write code

- Analyze
  - Interactively **manipulate** genomic data with a comprehensive and expanding 'best-practices' toolset
- Publish and Share
  - Results and step-by-step analysis record (**Data Libraries** and **Histories**)
  - Customizable pipelines (**Workflows**)
  - Share workflows with other users

Cloudman

What is Cloudman?

What is Cloudman?

- Cloudman is written by Enis Afghan *et.al.*, Emory University and provides a ready-to-run, dynamically scalable version of Galaxy on **Amazon AWS**

What is Cloudman?

- Cloudman is written by Enis Afghan *et.al.*, Emory University and provides a ready-to-run, dynamically scalable version of Galaxy on **Amazon AWS**
- Now it's possible to run it also on the **SARA HPC Cloud / Opennebula** (with some limitations)

How does it work?

How does it work?

- A **master node** contains all the data and tools

How does it work?

- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load

How does it work?

- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS

How does it work?

- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS
- RabbitMQ is used for **communication** between cluster nodes

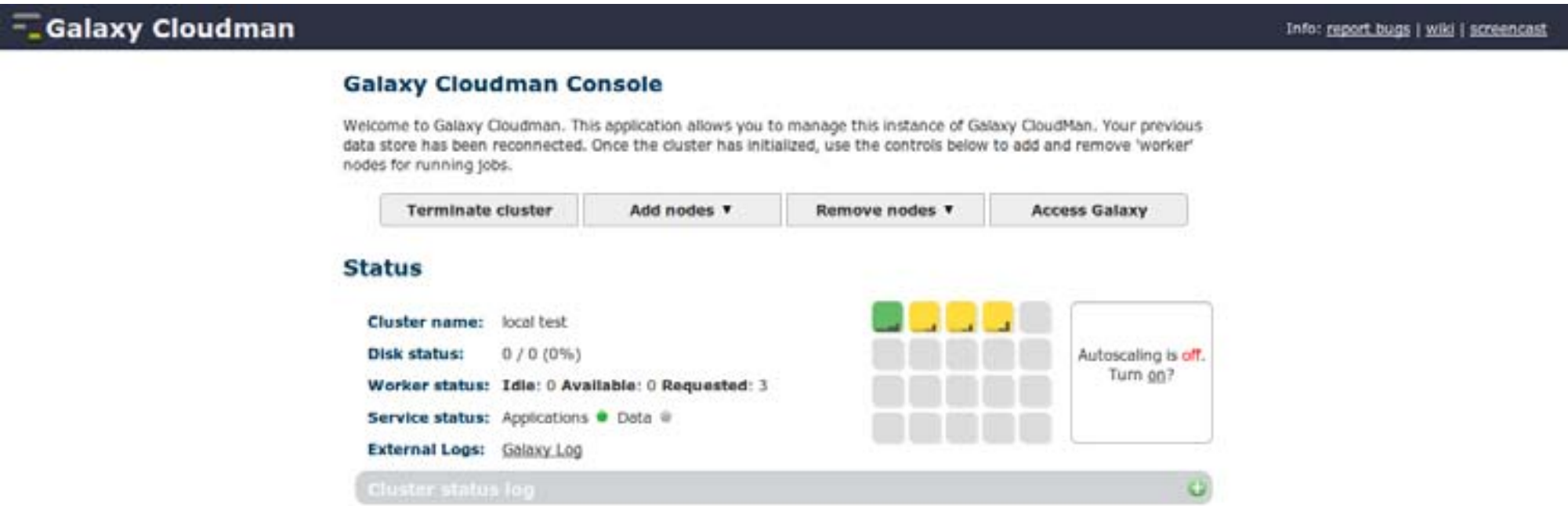
How does it work?

- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS
- RabbitMQ is used for **communication** between cluster nodes
- Jobs are queued using **SGE**

How does it work?

- A **master node** contains all the data and tools
- Initiate **worker nodes** based on needs/load
- Data is available on all nodes using a **shared filesystem**: NFS
- RabbitMQ is used for **communication** between cluster nodes
- Jobs are queued using **SGE**
- Galaxy is served using nginx **webserver**

Workers instances are being configured



Galaxy is accessible

How is Galaxy used at the NIOO?

How is Galaxy used at the NIOO?

How is Galaxy used at the NIOO?

- Analyzing high-throughput community sequencing data with QIIME

How is Galaxy used at the NIOO?

- Analyzing high-throughput community sequencing data with QIIME
  - Denoising (**CPU-intensive**)

How is Galaxy used at the NIOO?

- Analyzing high-throughput community sequencing data with QIIME
  - Denoising (**CPU-intensive**)
  - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools

How is Galaxy used at the NIOO?

- Analyzing high-throughput community sequencing data with QIIME
  - Denoising (**CPU-intensive**)
  - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools
  - Taxonomy assignment with BLAST or the RDP classifier (**CPU-intensive**)

How is Galaxy used at the NIOO?

- Analyzing high-throughput community sequencing data with QIIME
  - Denoising (**CPU-intensive**)
  - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools
  - Taxonomy assignment with BLAST or the RDP classifier (**CPU-intensive**)
  - Sequence alignment with PyNAST, muscle, infernal, or other tools (**CPU-intensive**)

How is Galaxy used at the NIOO?

- Analyzing high-throughput community sequencing data with QIIME
  - Denoising (**CPU-intensive**)
  - OTU and representative set picking using uclust, cdhit, mothur BLAST or other tools
  - Taxonomy assignment with BLAST or the RDP classifier (**CPU-intensive**)
  - Sequence alignment with PyNAST, muscle, infernal, or other tools (**CPU-intensive**)
  - and more!

Thanks!

Thanks to the Galaxy Cloud Team

Questions?

Extra slides

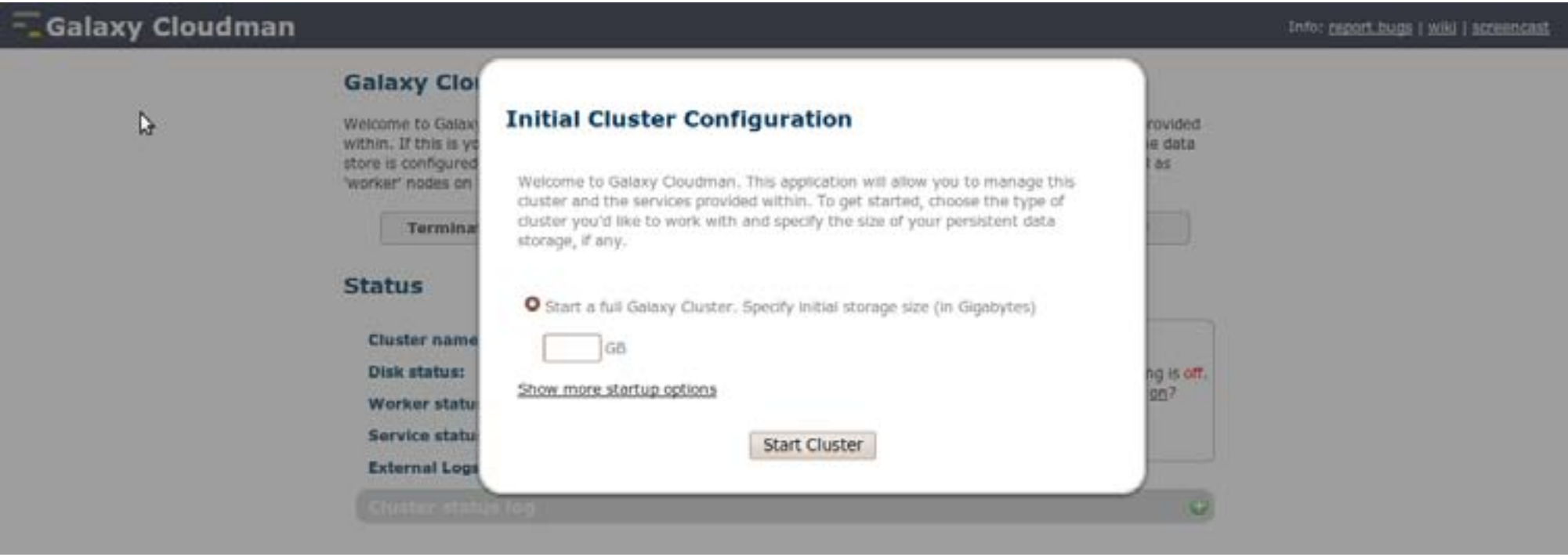
Limitations of Opennebula

- Create instances providing user data (available in production cloud?)
- No support for growing qcow filesystem
- Would be create to access the cloud the ON API from outside
- Cloned instances have not a working network

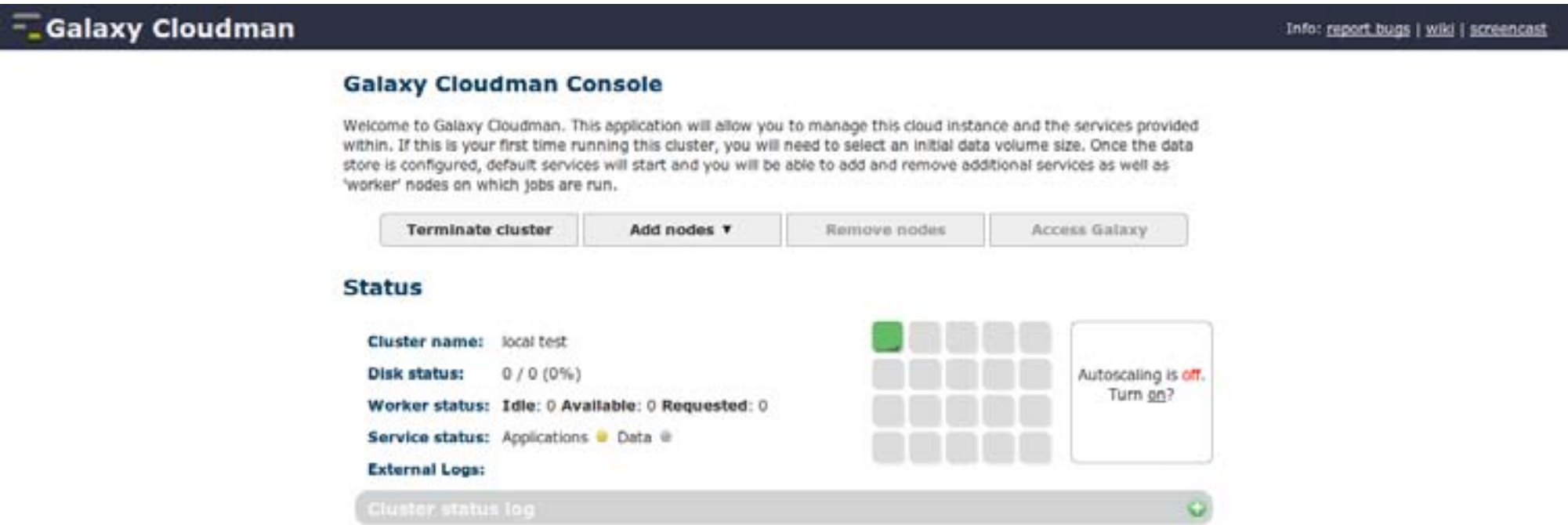
More info at

- My notes: https://www.cloud.sara.nl/projects/galaxy/wiki
- Galaxy Cloud on Amazon: http://usegalaxy.org/cloud
- Cloudman scripts: https://bitbucket.org/galaxy/cloudman/
- Install tools: https://bitbucket.org/afgane/mi-deployment
- Bio-linux repository: http://nebc.nerc.ac.uk/tools/bio-linux/bio-linux-6.0

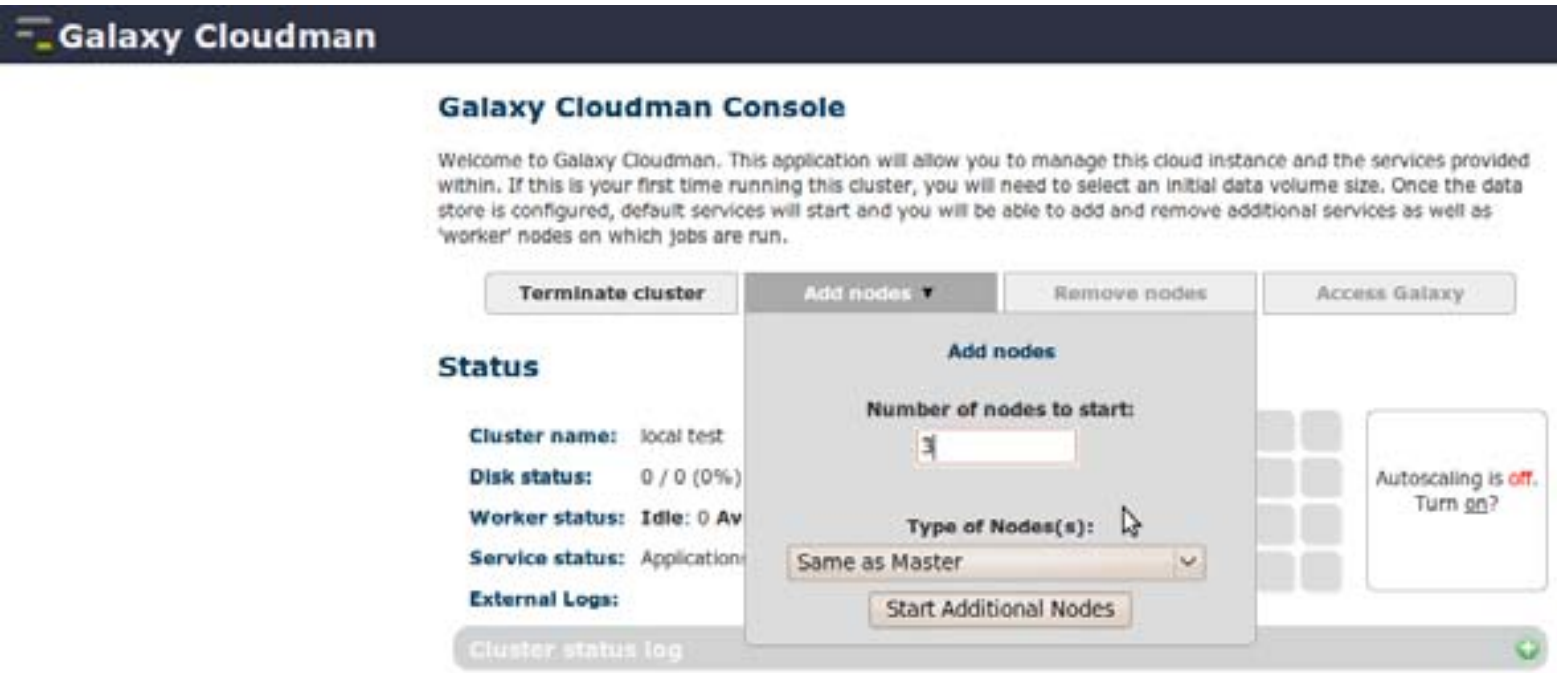
Launch Cloudman Console



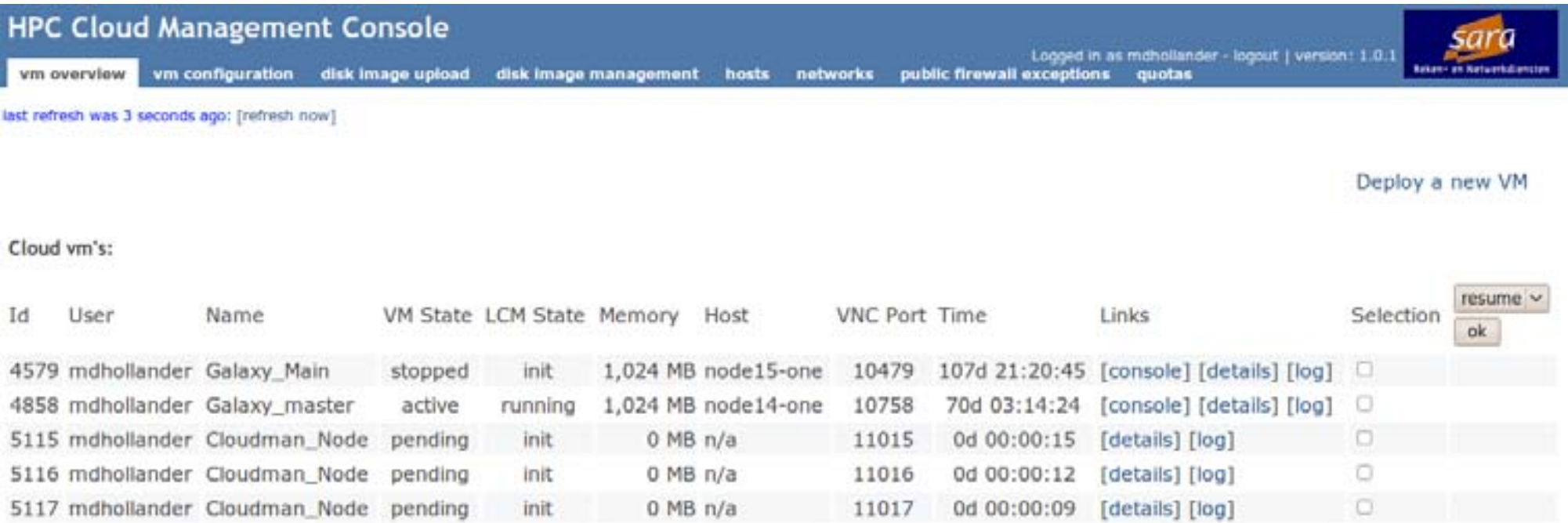
Master node is online



Add extra worker nodes

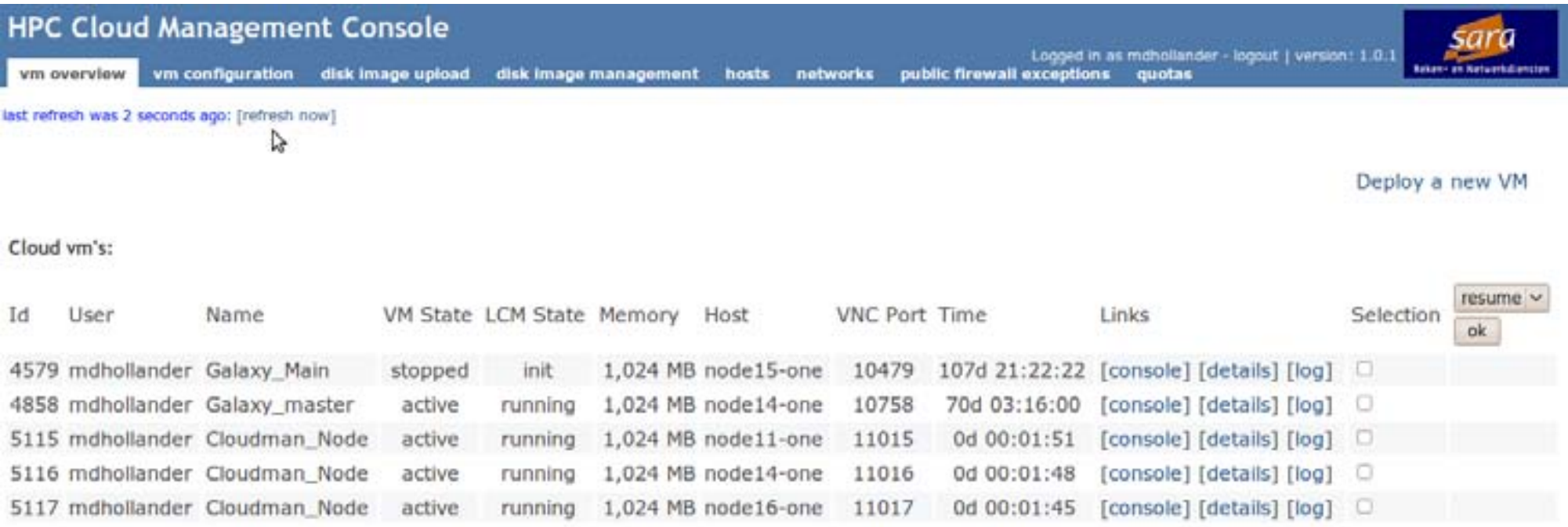


New instances are pending

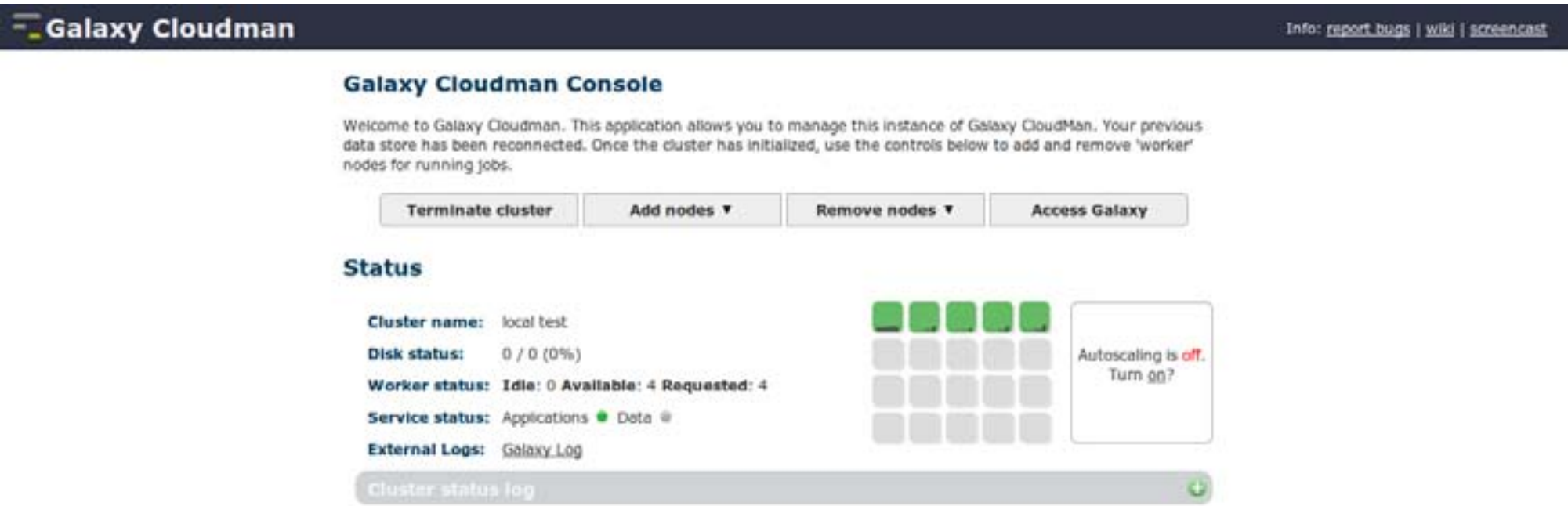


New instances are pending #2

New instances are running



New instances are online



Galaxy is accessible

```
galaxy@ubuntu:/home/cloud$ qstat -f
queuename                      qtype resv/used/tot. load_avg arch          states
---------------------------------------------------------------------------------
all.q@ubuntu                   BIP   0/0/1          0.26     lx24-amd64
---------------------------------------------------------------------------------
all.q@worker-5118              BIP   0/0/1          0.05     lx24-amd64
---------------------------------------------------------------------------------
all.q@worker-5119              BIP   0/0/1          0.00     lx24-amd64
---------------------------------------------------------------------------------
all.q@worker-5120              BIP   0/0/1          0.08     lx24-amd64
---------------------------------------------------------------------------------
all.q@worker-5121              BIP   0/0/1          0.04     lx24-amd64
```

```
galaxy@ubuntu:/home/cloud$ qstat
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
      1 0.55500 galaxy_117 galaxy       r     09/01/2011 15:49:02 all.q@ubuntu                       4
```