Building a scalable Galaxy cluster for biomedical research in The Netherlands

David van Enckevort[1], Anthony Potappel[2], Niek Bosch[3], Jeroen Beliën[4], Rita Azevedo[5], Rob Hooft[5], Sander Ruiter[2], Sanne Abeln[7], members of TraIT WP4, Irene Nooren[3], Jan-Willem Boiten[6]

[1]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, [2]Vancis, Amsterdam, The Netherlands, [3]SURFsara, Amsterdam, The Netherlands, [4], Department of Pathology, VU university medical center, Amsterdam, The Netherlands, [5] Netherlands eScience Center, Amsterdam, The Netherlands, [6] Center for Translational Molecular Medicine, Eindhoven, The Netherlands, [7]VU university, Amsterdam, The Netherlands

E-mail: david.van.enckevort@umcg.nl, Anthony.Potappel@vancis.nl

1. Introduction

Galaxy[1] is a popular web-based bioinformatics workflow framework that is excellent for explorative research. In recent years several institutes have installed their own Galaxy instance including the Netherlands Bioinformatics Centre (NBIC), which has maintained a public instance[2] since 2010.

For the national translational IT project CTMM/TraIT Galaxy has been selected as one of the tools in the experimental domain. The TraIT partners (among others NBIC and SURFsara) have developed a vision how to make Galaxy available to the research community in The Netherlands.

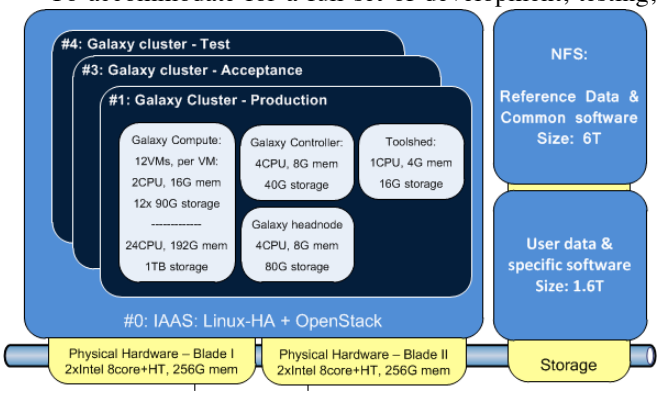
The scalable Galaxy cluster on the SURFsara HPC Cloud will be transferred to Vancis to provide a sustainable production-level Galaxy cluster. In the design of this environment Vancis has made use of the knowledge and experience of NBIC and SURFsara hosting the public NBIC instance on the SURFsara HPC Cloud.

2. Materials & Methods

To assess the minimal requirements for the infrastructure we used metrics collected while running the NBIC Galaxy on the HPC Cloud. Next we drafted a set of use cases the infrastructure should be able to fulfil, such as the ability to run Omics-pipelines and the ability to scale to handle peak demand.

We identified I/O performance as a major bottleneck, since many Galaxy tools are I/O intensive, while Galaxy has a shared data design. Memory was also recognized as a critical factor, since typical datasets are in the order of the tens of gigabytes. We also built upon the experiences from SURFsara in operating the HPC Cloud and other HPC.

To accommodate for a full set of development, testing, acceptance & production environments, as well as private installations, the infrastructure should support multiple Galaxy clusters.



The chosen architecture will use a Linux High Availability environment with OpenStack, which will run on two large-size blades. Storage is split into multiple tiers with different characteristics to support both high I/O workloads and a reliable large storage. The chosen setup is horizontally scalable in a cost-efficient manner.

3. Results

From May to September 2014 we will pilot the new architecture within the TraIT project. For this pilot we have selected a few TraIT NGS tools and pipelines to stress test the system under different workload scenarios.

Furthermore we have established a process to ensure the quality of the tools required for a stable production environment. We have formulated acceptance criteria for the deployment of a tool in Galaxy: 1) we require that each tool is properly licensed; 2) tools should be distributed through a Galaxy Tool Shed and 3) we expect that developers abide by a set of standards for publishing a tool. For example, each tool needs to come with a brief description explaining the tool and full documentation. 4) Tools must be testable within the Galaxy functional testing framework and provide example data for testing. By setting these criteria we can verify that a tool works correctly on our platform.

4. Discussion

The model and architecture used to build the infrastructure is chosen to be scalable across different use cases and allows for efficient sharing of resources between users with disparate requirements. We envision that we can not only offer a stable environment to be used by CTMM studies but also make this service available to other researchers.

To keep the environment sustainable we seek a model where the base infrastructure is financed by institutes, but where individual studies or researchers can buy additional services or resources for their project. So far this model is largely untested in the Dutch research community and we actively seek the input from interested parties.

References

1. http://getgalaxy.org/
2. http://galaxy.nbic.nl/