trait
a ctmm project

Center for Translational Molecular Medicine

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[6], Irene Nooren[3], Jan-Willem Boiten[7]
[1] University Medical Center Groningen, University of Groningen, Groningen; [2] Vancis, Amsterdam; [3] SURFsara, Amsterdam; [4] VU university medical center, Amsterdam; [5] Netherlands eScience Center, Amsterdam; [6] VU university, Amsterdam; [7] Center for Translational Molecular Medicine, Eindhoven. All affiliations: The Netherlands.

Introduction

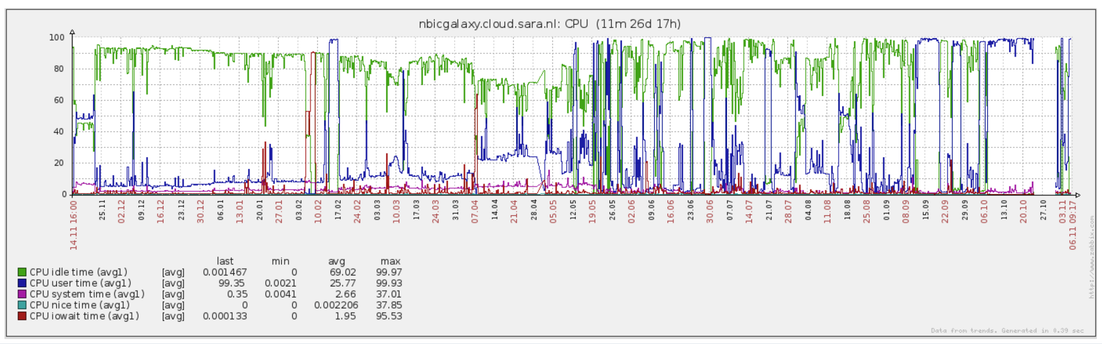
Galaxy has been selected as an important tool to analyze (combinations of) genomics and proteomics data for Dutch medical research projects. The TraIT partners (among others NBIC and SURFsara) have decided how they will make Galaxy available to the research community in The Netherlands.

Our scalable Galaxy cluster was developed on the academic High Performance Computing (HPC) cloud hosted by SURFsara. We will now transfer it to a production grade cluster supported by the commercial hosting partner Vancis. In the design of the new system, we have used our collective experience.

Metrics

To assess the minimal requirements for the infrastructure we used metrics collected with the system monitoring software Zabbix while running the NBIC Galaxy on the SURFsara HPC Cloud. These metrics provided us with real world usage insights from a public Galaxy instance with over 1500 registered users.

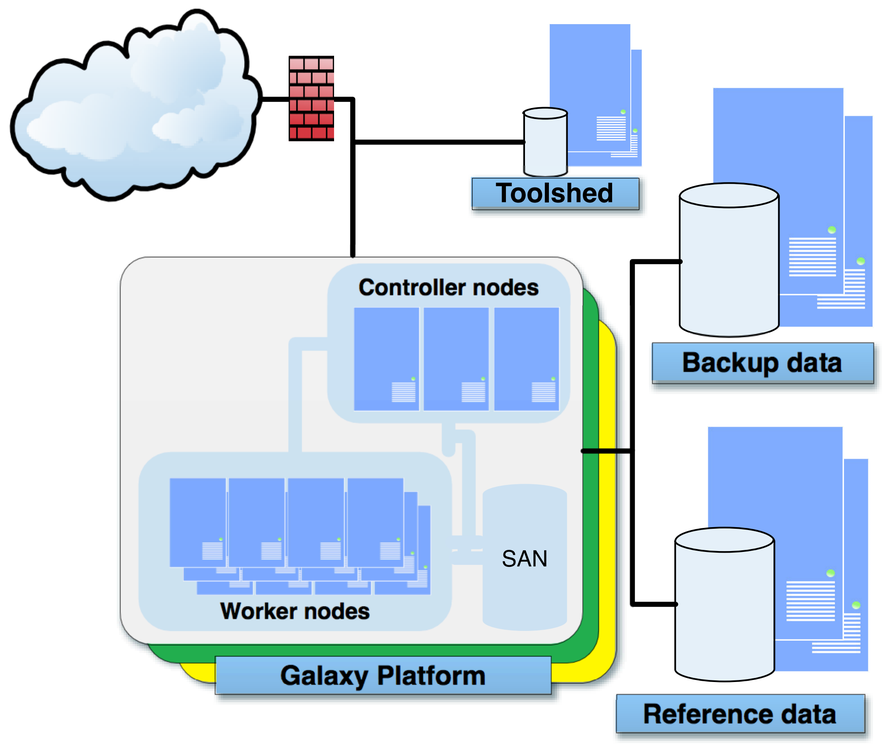
We identified I/O performance as a major bottleneck, since many tools in Galaxy are I/O intensive, while Galaxy has a shared data design. Memory was also recognized as a critical factor, because tools often load large parts of a dataset in memory and typical datasets are in the order of the tens of gigabytes.



Architecture

Based on these requirements an architecture was created from predefined cloud building blocks. Our solution uses multiple tiers of storage, each with different characteristics and costs, to support both high I/O workloads and reliable large storage. By using cloud technology we can elastically scale compute resources based on demand. Also, we can allow different users to each have their own private cluster with their own unique tools and customizations, and backup it completely for later use.

The multi-tenant setup facilitates a very efficient and robust testing-, acceptance- and deployment process. Environments can be quickly verified (or rejected) before moving into final production.



Use cases & requirements

Based on the use cases from NBIC and CTMM we identified different requirements ranging from security and availability to functional and procedural requirements. These must be addressed in different parts of the architecture of the platform and by hosting by professionals in a certified data center.

Requirement	Addressed by	Remarks
Ease of use	Galaxy	
Provenance	Galaxy	
Ease of administration	Tool Shed / hosting company	
Stable system	DTAP instances, Tool Shed, stable releases	
Service levels	Hosting company, skilled helpdesk / redundant architecture	
Certified security of data	Architecture / hosting company certification	Under investigation
Scalability	Scalable architecture / CloudMan	
Performance	Architecture with HPC components	
Resource accounting	Galaxy reporting module Quota & scheduler	Partially implemented
Single Sign On	OpenConext (SAML)	Under investigation

Current status

In the initiation phase CTMM-TraIT and NBIC collected use cases and defined the requirements for the platform. Vancis as preferred vendor and partner in the CTMM-TraIT project was approached to design an architecture that can meet those requirements.

Currently we are building the proof of concept, which will be validated by using selected TraIT NGS tools and pipelines to stress test the system under different workload scenarios that mimic real world use cases.

Once in production the platform will be available for studies that make use of the CTMM-TraIT architecture. Vancis will commercially offer Galaxy as a “building block” in the Vancis portfolio for other parties.

Phase	Time frame	Status
Initiation	2013: Q3-Q4	Completed
Architecture & Design	2014: Q1	Completed
Proof of Concept	2014: Q2-Q3	Build phase
Production	2014: Q4	Planned

Acknowledgements

The authors would like to acknowledge the valuable input of CTMM-TraIT and the NBIC community.

Contact

David van Enckevort
UMCG
Department of Genetics
david.van.enckevort@umcg.nl

Participants

TraIT is a Dutch public/private partnership between University Medical Centers, several other public institutions, charities, and companies: