Building a scalable Galaxy cluster for biomedical research in The Netherlands

NBIC 2014

David van Enckevort
Technical Project Manager @ UMCG

Anthony Potappel
Big Data Engineer @ Vancis

trait
a ctmm project

Vancis
Advanced ICT Services

Outline

- Introduction
- The Challenge
  - What, Why & How?
- Technology Transfer
- Solution
  - Approach
  - Architecture & design
  - Roadmap
- Credits

Introduction

David van Enckevort
Technical Project Manager @ UMCG

Anthony Potappel
Big Data Engineer @ Vancis

trait
a ctmm project

Vancis
Advanced ICT Services

Introduction

CTMM/TraIT
Sustainable infrastructure for medical research projects.
http://www.ctmm-trait.nl/

Vancis
Delivering cloud solutions, infrastructure and platforms as a service.
http://www.vancis.nl/

What?

Provide a workbench for experimental research.

- Enable researchers to:
  - Run standardized pipelines
  - Engage in explorative research
  - Track the provenance of their data
- Reliable & secure platform
- Including good support
- Scalable: in terms of applications & performance
- Supports multi-site research

Why?

- Research is the work of large, multi-national consortia
- Translational research seeks fast conversion of research results in improved healthcare
- Complexity of research & tools requires specialized knowledge
- Omics-research is a big data challenge

How?

CTMM/TraIT selected Galaxy as the workbench solution. Galaxy offers:

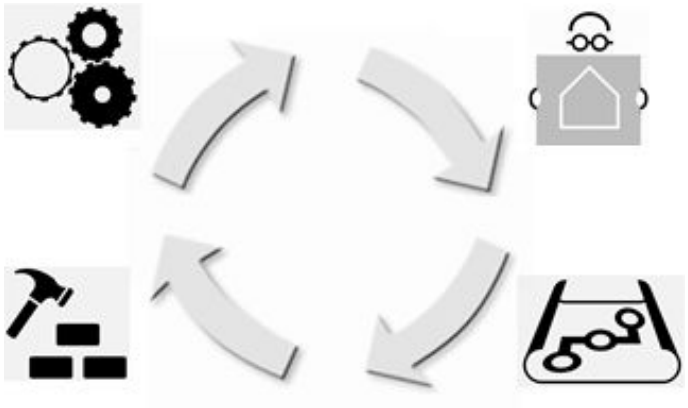
- Web-based, scalable platform
- Large (inter)national community
- Proven technology: NBIC has experience running an instance of Galaxy at SURFsara's HPC Cloud

Technology Transfer

- User requirements:
  - Education:
    - Small jobs, many users
    - Availability
  - Research needs:
    - Large jobs, fewer users
    - Integrity & confidentiality of data
- Technical know-how:
  - Performance metrics: large memory footprint & intensive disk i/o
  - Usage patterns: Erratic & Big Data alike

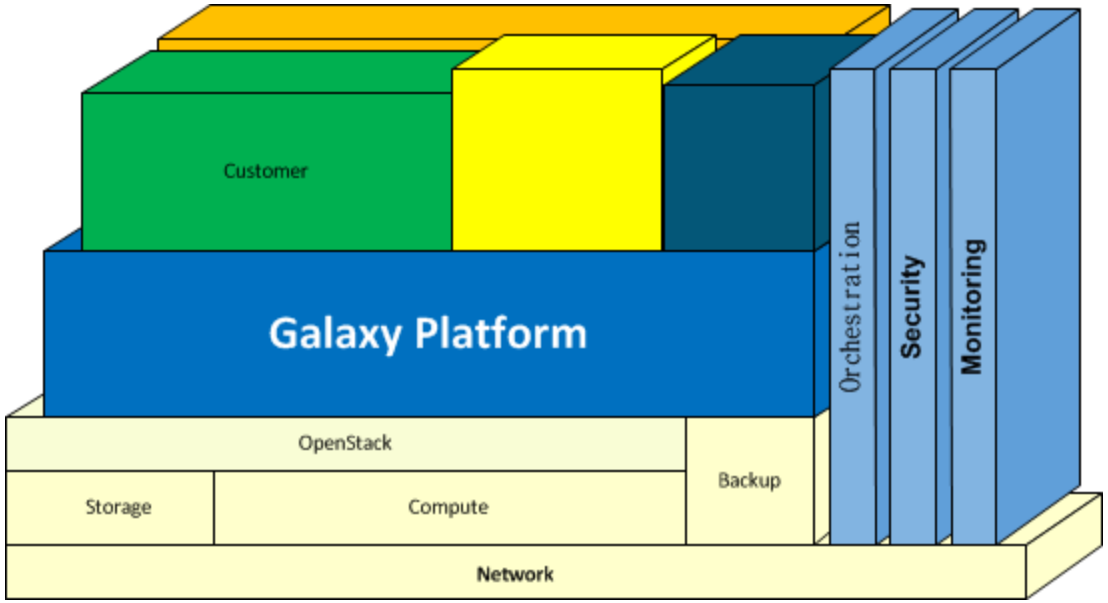
Approach

- Architecture
  - Meeting users' requirements
  - Standard building block
  - Integrated and connected
- Design
  - Create blueprints
  - Define bill-of-materials
- Build
  - Configure hardware
  - Create templates and scripts
  - Connect to services
  - Testing: functionality & performance
- Improve iteratively based on user feedback



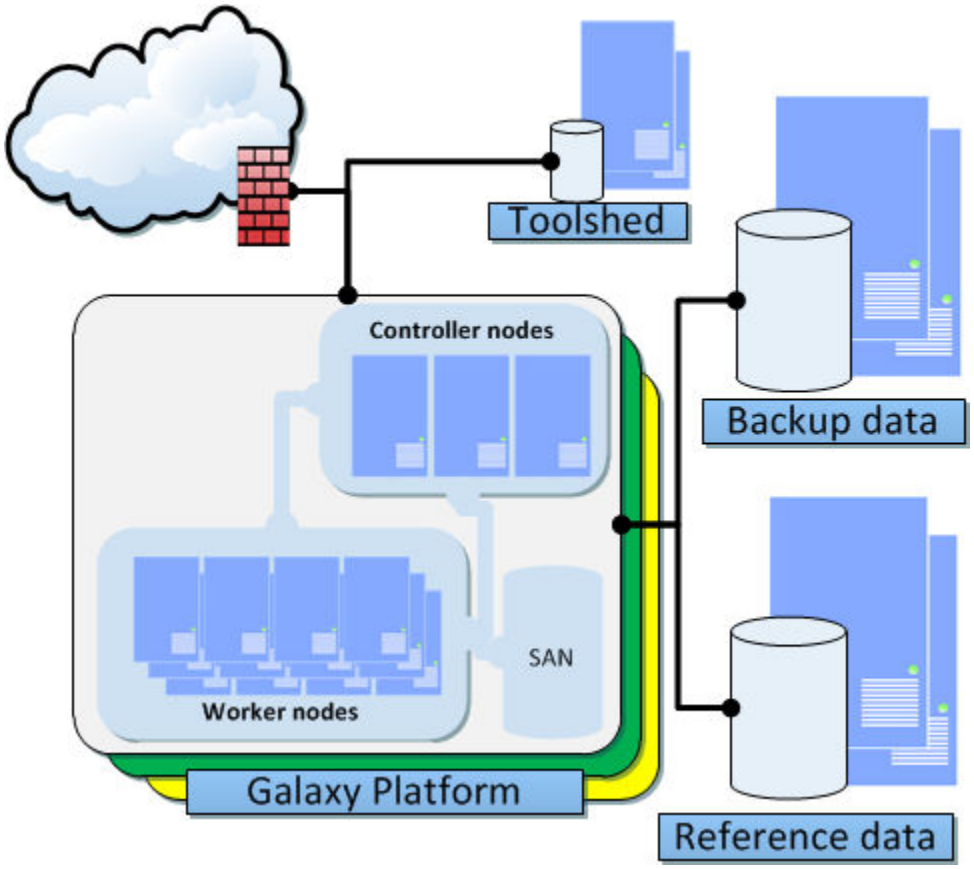
Architecture

- Must have's
  - Security: Integrity, Availability and Confidentiality
  - Rapid deployment
  - Multi-tenant
  - Scalable
- Platform "building block"
- Integrated
  - Compute, Storage, Network, Backup.
  - Orchestration
  - Monitoring
  - Security
  - Technical support



Design

- Galaxy Platform
  - Fast provisioning
  - Controller-, worker nodes
  - Network, connectors and Services
  - Scalable, scalable, scalable
- Storage
  - SAN: direct I/O
  - NFS: shared reference data
  - Tape-backup
- Excellent connectivity
  - AMS-IX, NL-IX
  - NetherLight
  - 150+ carriers in data-center



Roadmap

Phase	Timeframe	Status
Initiation	2013: Q3-Q4	✓
Architecture & Design	2014: Q1	✓
Proof-of-Concept	2014: Q2-Q3	build-phase
Production	2014: Q4	

Credits

- CTMM: Jan-Willem Boiten
- Netherlands eScience Center: Rita Azevedo, Rob Hooft
- SURFsara: Niek Bosch, Irene Nooren
- Vancis: Sander Ruiter
- VU university: Sanne Abeln
- VU university Medical Center: Jeroen Beliën

And especially everyone within the CTMM TraIT and NBIC community who shared their experiences!

trait
a ctmm project

Vancis
Advanced ICT Services