e-Infrastructure for NGS - a Finnish perspective

Aleksi Kallio

19.1.2015

CSC-TIETEEN TIETOTEKNIIKAN KESKUS

Outline

- Finnish model for scientific computing
- CSC's infrastructures for different user groups
- NGS data management
- IT systems
- Future plans

Finnish model for scientific computing

- CSC is the national service provider
- University IT centers are local service providers
- Service provided in tiers
- CSC provides: supercomputing capacity, cluster capacity, cloud capacity, data storage, software, support (tier 1)
- Universities and sequencing centers provide: cluster capacity, data storage, software, support (tier 2)

CSC's role

- National supercomputing center
  - Serves all universities in Finland and all scientific disciplines
  - Not dedicated for NGS data analysis
- CSC's users
  - A lot of users
  - Different types of users
  - Beginner / advanced
  - A wide range of analysis topics, even within NGS
- Need to cater for everybody and hence provide flexible services

NGS users at CSC

- Sequencing centers
- Individual groups with bioinformaticians
- Individual life scientists
- Different types of users, different services provided

Infrastructure for sequencing centers

- Dedicated cloud capacity (BMI)
  - Dedicated network connection also available (lightpath)
- (Co-located or hosted cluster capacity)
- Developing Hadoop and Spark environments (PaaS)
  - Hadoop-BAM and SeqPig with Aalto University

Infrastructure for groups

- HPC cluster capacity (Taito)
  - Wide selection of analysis software and databases installed
- Cloud capacity (Pouta)
  - Allows to install own software environment
  - Allows to build custom systems (web server, grid engine, Hadoop, Spark, Storm…)
  - Tool support for environment provisioning
- (Co-located or hosted cluster capacity)

Infrastructure for life scientists

- Cluster and cloud capacity for technically capable users (minority)
- Chipster GUI (SaaS)
    - Over 340 analysis tools for NGS, microarray and "traditional" sequence analysis
    - Eija will talk about this tomorrow
- Customer survey 12/2014: large need for easy to use interfaces

NGS data management

- Raw data stored by sequencing centers or research groups
- Working data storage
  - At university (NFS)
  - At CSC (Lustre)
- Result data archival
  - HPC archive at CSC
  - IDA service via CSC (iRODS)
- Data moves inside Funet network
  - Lightpath to largest sequencing center FIMM

IT systems

- Taito HPC cluster
  - HP SL cluster with 19,000 cores, 16-1,500 GB/node, InfiniBand, Linux, SLURM
- Pouta cloud
  - OpenStack cloud environment (IaaS)
  - Physically runs in Taito
- BMI cloud
  - OpenNebula dedicated cloud environment (IaaS)
  - For confidential data
- Hardware located in industrial site at Kajaani (Northern Finland)
- Connected to university network backbone (Funet)

Related projects

- ELIXIR
  - IaaS, lightpaths, training
- Tryggve
  - Confidential patient data
- SeqAhead
  - Collaboration on data intensive computing and infrastructures
- D2I
  - Data intensive computing for NGS (Hadoop, Spark)

Future plans

- Chipster: easy to use tools for sharing large datasets and accessing external storage, tool support for new NGS applications
- Better tools for users who use CSC cloud to serve other users
- More comprehensive selection of productized virtual machines
- Specialized cloud hardware for data intensive tasks
    - In-memory computing and Spark
- Universal object storage for working data
- Dedicated cloud: Replacing BMI (OpenNebula) with specialized ePouta (OpenStack)

Summary

- Tiered model: CSC is the national provider, university IT centers are local providers
- To avoid overlap between tiers, flexible cloud solutions are used
- Different services for different NGS user groups
- Future development: strengthening and consolidating cloud platform (IaaS, PaaS, SaaS), easy to use interfaces