Cloud and virtualization in NGS

Notes collected by Aleksi Kallio

Knowledge café
19.1.2015

DEFINITIONS

NGS data analysis?

- Key points for infrastructure
  - Large data sizes
  - Data can be confidential
  - Deep software stacks
  - Quickly developing field
  - Incoming data quite uniform

Cloud and virtualization?

- Virtualization
  - Hardware virtualization (virtual machines)
  - Software containers (Docker etc.)
- Cloud
  - Infrastructure-as-a-Service, typically implemented with virtualization
  - Platform-as-a-Service, e.g. computing frameworks such as Hadoop
  - Software-as-a-Service, NGS specific/ related software

CLOUD SWOT QUESTIONS

Strengths

- Could cloud be used to create more scalable infrastructure?
    - Solution for intermittent workloads
    - Dynamic scaling (Heat etc.)
- Good way to achieve reproducibility
- Flexibility, dependency management
    - Example: Docker container inside VM for Galaxy tools, makes updates easier

Weaknesses

- Is cloud HPC too expensive?
  - Can be with major workloads
- How will data transfer work?
  - There are concerns
  - Solution: shipping hard drives?
- Privacy can be major blocker

Opportunities

- Open cloud platforms for open science?
  - Moving computation instead of data because of confidentiality
- Enable new kind of NGS analyses?
- Training environments
  - Easy capacity, pay as you go
  - Reproducible environment
  - Separation of student environments

Threats

- Could confidentiality and security kill NGS in cloud?
    - Yes it could, depending on the case and country
- Could there be vendor lock-ins?
    - Potentially serious lock-in: BaseSpace (SaaS)
- Responsibility cannot be transferred to cloud provider (Sweden)
- Question: Data cannot leave organisation, but how do you define organisation?
- Badly maintained VMs security threat