e-infrastructures planning ahead

Knowledge café

Unstructured notes by Ola Spjuth

Challenge of storage capacity

- If you pay money for sequencing, you want it all
  - FASTQ, BAM, forever – not sustainable
- Bio is young field computing-wise
  - We will run out of space
- Users need to pay for storage
  - Individual user groups universities
  - Cite data
- File formats need to evolve and become accepted/supported
- Some cases: Save the blood and re-run is one option
  - What percentage? 20%?
  - Store data in DNA
  - 10 years: Checking for fake and fraud, need the original data
- Types of data: Working, backup, archive
  - Users: REALLY want to save it
- For long term – can we use data? Data curation, can we read FASTQ?
- Planning: 3 years ago: no funding for IT, now: Funding for IT (HW + pipelines) but no understanding of timing. Has to be acknowledged in future.
- Re-use data, open data: Individual researchers will not do this. Need central initiatives and funding.
- NGI Today: More costly to support snapshots of all versions than to store raw data + results
- Yesterday: Save all images – times are changing - data reduction is there (CERN don't cry) – bio is not mature yet to accept this
- Only way to be efficient: Huge central storage, standardized and open format. Like GenBank.

Challenge of computing capacity

- Previously: Driven by physics
- Now: Driven by gaming
  - We adapt bio-software to this
  - We are going to be stuck in this for a while (glue with workflows)
    - Prev: Geeks.
    - Now: everyone uses computers
    - Future: Tablet-science
- Towards microservices/virtualization
- End users and data production have different needs
- Cloud-provisioned bare metal
  - One huge OpenStack. Sysadmin can be customer, running a slurm cluster on bare metal from OpenStack
- High performance computing vs high-throughput computing
- BIO are not big CPU users per se compared to e.g. physics (MD, FLOW)
- Re-run jobs with settings defined by previous finished jobs

Will bioinformatics embrace HPC? Clouds? Hadoop? Spark?

- Inefficient bioinfo-jobs
  - Virtualization to overcommit nodes?
- A mix of both will persist for a while
- Hadoop/Spark: Massively parallel
  - is less trusted, questioned
- GenAlice: Optimize HPC
- Hadoop and HPC
  - Now: Need resources for both in parallel
- Difficult to optimize on the forefront
  - We are not at standstill but things are slowing down (alignment)