H3ABioNet
Pan African Bioinformatics Network for H3Africa

Cloud Computing

Scott Hazelhurst

University of the Witwatersrand, Johannesburg

8th H3A Consortium Meeting

H3ABioNet
Pan African Bioinformatics Network for H3Africa

What is cloud computing?

Cloud computing: *other people's computers*

- Packaged in a way that you can use easily
- Usually charged for
- Different types:
  - e.g. Dropbox, or Gmail (Software as a Service)
  - e.g. Amazon Elastic Computing Cloud (EC2): (Infrastructure as a Service)


Why? Why not?

- Turn capital cost into operating cost
- Costs predictable: power, maintenance is someone else's responsibility
- Elastic, pay for what you use
- Networking an issue
- May be legal, regulatory issues

H3ABioNet

H3ABioNet


Who?

Several large infrastructure providers

- Amazon Web Services
- Google
- Microsoft
- Rackspace

H3ABioNet

H3ABioNet

How?

1. Replicate your physical set up in the cloud

How?

1. Replicate your physical set up in the cloud
2. Many systems to assist (e.g., StarCluster)


How?

1. Replicate your physical set up in the cloud
2. Many systems to assist (e.g., StarCluster)
3. Excellent pipelines for bioinformatics – either by infrastructure providers or companies with their systems that use cloud infrastructure

How?

1. Replicate your physical set up in the cloud
2. Many systems to assist (e.g., StarCluster)
3. Excellent pipelines for bioinformatics – either by infrastructure providers or companies with their systems that use cloud infrastructure

How?

1. Replicate your physical set up in the cloud
2. Many systems to assist (e.g., StarCluster)
3. Excellent pipelines for bioinformatics – either by infrastructure providers or companies with their systems that use cloud infrastructure

Many specialised bioinformatics cloud services built on top of this


How much?

How much do you want to pay?


How much?

How much do you want to pay? . . . Is it cost effective?


How much?

How much do you want to pay? . . . Is it cost effective?

1. Huge advantage is elasticity. . .
   Paradox of many clusters: idle lots of the time, far too small when your data arrives


How much?

How much do you want to pay? . . . Is it cost effective?

1. Huge advantage is elasticity. . .
   Paradox of many clusters: idle lots of the time, far too small when your data arrives
2. Depends on your usage pattern.

If you buy 20 twelve-core computers you pay for each core whether idle or not.

- If all 240 cores busy 80% of the time, much cheaper than using the cloud. But this means $80\% \times 240 \times 24 \times 365$!
- If cores only busy 10% of the time (240k CPU hours) it's cheaper to use the cloud.

- Computational costs competitive with what you can buy for yourself, elastic.
- ≈USD0.05/core/hour: ($1500 for a 200 core cluster for a week)

- Computational costs competitive with what you can buy for yourself, elastic.
- Storage costs probably 3-4 times more expensive

- ≈USD0.05/core/hour: ($1500 for a 200 core cluster for a week)
- S3: $0.03/GBmonth. ($360/TB year)

witsGWAS: Example use of cloud services

Wits has a small cluster: 300 cores or so

- Good for most production work but limited at peak times
- Can we use the cloud?
- Need to do production work: first few projects were hand-crafted but we now need to replicate and automate

Solution:

- Use nextflow as a workflow language that allows us to produce reliable, automated pipelines for work
  - e.g., QC, convert from image to call, upstream QC, statistical analyses, population admixture.

Solution:

- Use nextflow as a workflow language that allows us to produce reliable, automated pipelines for work
  - e.g., QC, convert from image to call, upstream QC, statistical analyses, population admixture.
- Use Docker as a lightweight virtualisation system that is portable

Credit: Lerato Magosi and Rob Clucas


Docker

Uses *containers* to isolate applications from underlying operating system:

- light-weight virtualisation
- runs across range of operating systems (mileage varies)
- very well supported on Linux

Major cloud vendors support it

Now can run our pipeline on

1. Our own cluster
2. At the SA Centre for High Performance Computing
3. On the Amazon EC2 Docker Service

H3ABioNet funded by NHGRI grant number U41HG006941