Access to Big Data in Bioinformatics

Brendan Ball, Andrew van Rooyen

Proposer - Michelle Kuttel

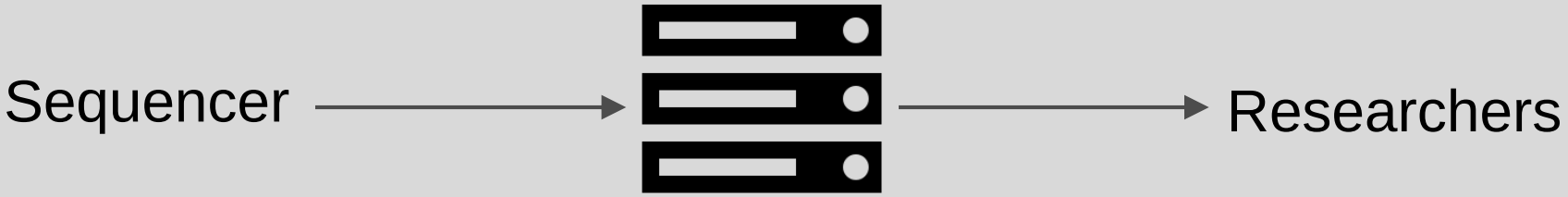
Collaboration (UCT - UWC)

External advisors:

- Antoine Bagula (CS,UWC)
- Alan Christoffels (SANBI,UWC)
- Peter van Heusden (SANBI,UWC)

How things work right now

- Raw genomic data is sequenced
- Up to tens of gigabytes per dataset
- Hosted at specific locations (like UCSC)

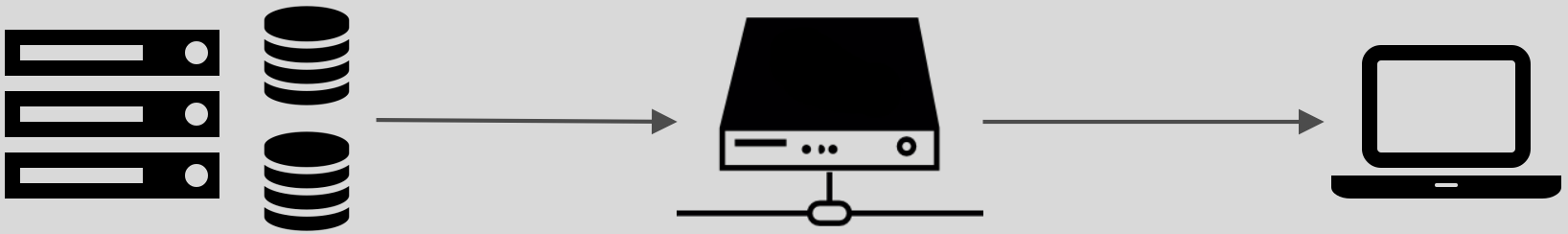


How things work right now

- Researchers use this in various areas
- The raw data is filtered and mapped to something which is relevant to them

How things work right now

- Before running these processes, they need to have the data
- Gigabytes of data which is 95% irrelevant to them



How we plan to improve this (Part 1)

- Transfer of these datasets should be optimised
- Especially on links between organisations
- We already have the infrastructure

SANReN
South African National Research Network

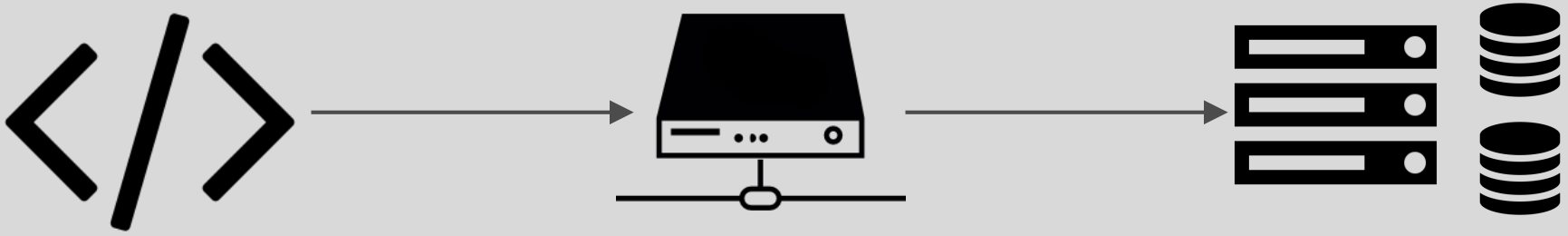

SANReN
Existing SANReN backbone
Construction in progress
Proposed extensions
TENET/SEACOM 10Gbps
Makhado
Polokwane
Turfloop
Pretoria
Middelburg
Nelspruit
SAC
Rustenburg
Mafikeng
eMalahleni
Johannesburg
Potchefstroom
Vanderbijlpark
Welkom
Kimberley
Bloemfontein
Europe/Ubuntunet
Mtunzini
Pietermaritzburg
Durban
SKA
SALT
Mthatha
Butterworth
Alice
Bisho
Potsdam
Grahamstown
East London
Cape Town
Port Elizabeth
Simeon Miteff, 2011/02/10

How we plan to improve this

- Unfortunately, relevant links in the WC are getting nowhere near these speeds
- We will be testing GridFTP, HPN-SSH as alternatives
- Aim to hit full 10Gb/s utilisation

Cloud as a logical next step (Part 2)

- Already have data in a datacenter
- Instead of downloading data, why not do processing remotely?



Micro-Cloud platform

- Use the “cloud” ideology
- Deploy local micro-cloud on-site
  - Allows analysis of Big Data
  - More maintainable
  - Scalable

Community Cloud

- Connect micro clouds from different universities
- Allows remote data analysis
  - Move data to code vs
  - Move code to data

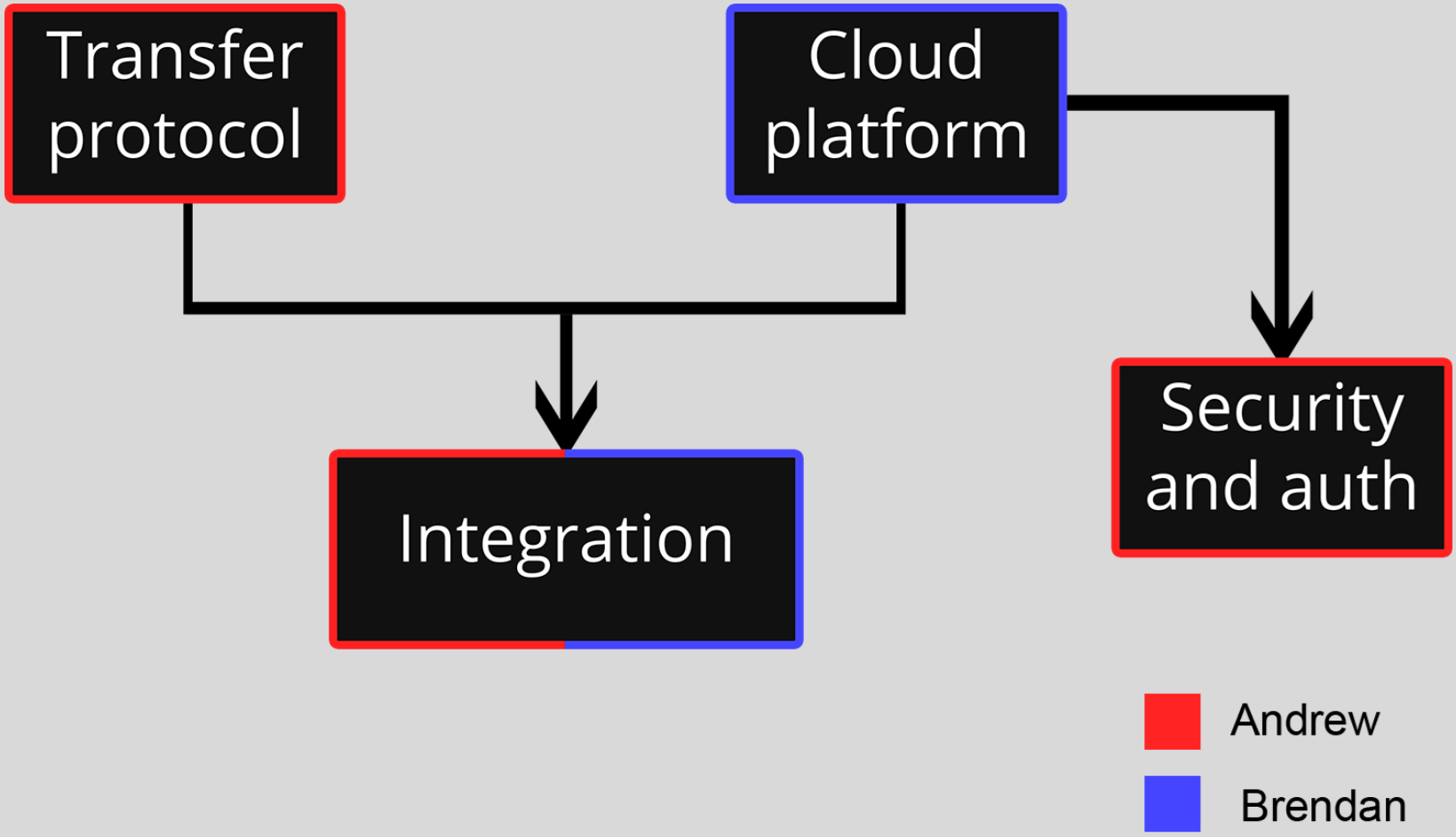
Authentication & Security

- Only authorized users can access cloud
- community cloud authentication
  - User from one cloud must have access to another connected cloud
- Encryption between micro-clouds

What are we doing?

- Survey existing cloud computing infrastructures
- Design and implement a micro cloud solution
- Prototype community cloud
- Implement a secure platform for running code remotely
- Integrate with fastest protocol over SANReN

Work Allocation



Anticipated Outcomes

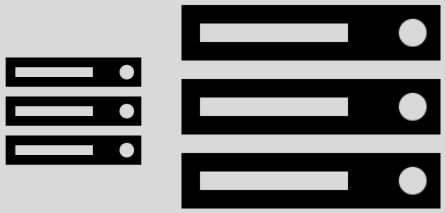
- Choice of best transfer protocol for big data in our context of bioinformatics in South Africa
- A micro cloud solution providing capability of creating a community cloud

Evaluation

- Metrics for protocols (throughput, overhead etc)
- Community cloud usability tests with expert users
- Successful remote processing of example bioinformatics data

Resources Required

- UCT network and computers (Hons lab)
- Access to external servers (UWC)
- Access to SANReN



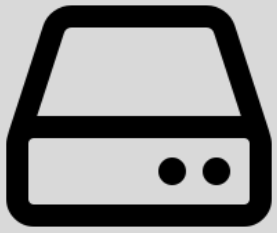
SANReN
South African National Research Network

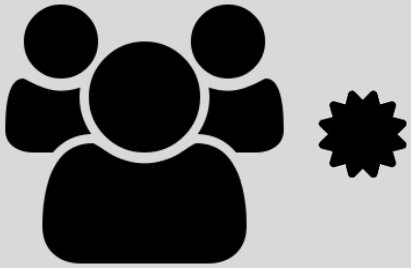
Related Work

- File formats and compression
- *fasp* protocol by AsperaSoft
- Cloud service providers (Amazon, Google)

Possible Future Work

- Optimisation on different levels of the stack (disk reads etc)
- Linking user identity for micro clouds to existing databases





Milestones

- Initial experiment with GridFTP and HPN-SSH
- Decide what a micro cloud agent would look like
- Best transfer protocol configured optimally
- Implement micro cloud
- Implement a management and deployment system for the agents
- Authentication and security worked into all moving parts

BigBInf

Brendan Ball, Andrew van Rooyen

Proposer - Michelle Kuttel

External advisors:

- Antoine Bagula (CS,UWC)
- Alan Christoffels (SANBI,UWC)
- Peter van Heusden (SANBI,UWC)