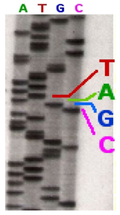
Incremental DNA Sequence Analysis in the Cloud

Romeo Kienzler and Nesime Tatbul, ETH Zurich, Switzerland
Rémy Bruggmann, University of Bern, Switzerland
Anand Ranganathan, IBM T.J. Watson Research Center, USA

Use Case: DNA Sequence Analysis

Data explosion due to revolution in DNA sequencing technology

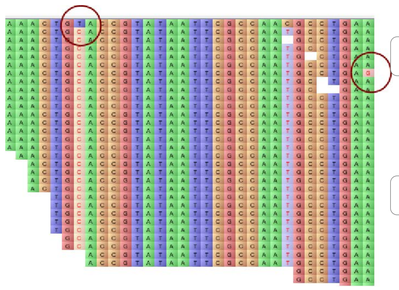
Sanger Sequencing → (2004 / 2007) → Next-Generation Sequencing (NGS)



Throughput for 80 kbp:
1 day vs. 10 seconds

Cost for 80 kbp:
$150 vs. $0.01

120 GB/day → ~11 Mbit/s



Read Alignment

Bring reads into correct order.

SNP Detection

Find single letter mutations.

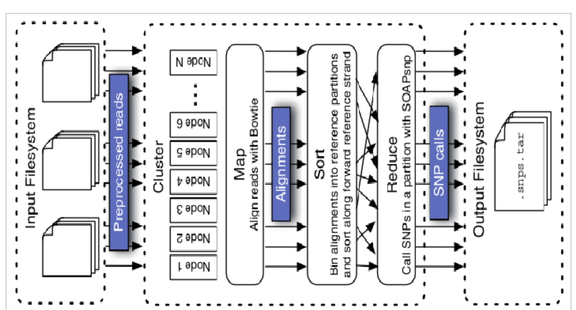
Applications

- medical treatment (personalized medicine)
- cancer research and diagnostics
- gene expression experiments (through RNA-seq)

State of the Art

Hadoop-based SNP Detection with Crossbow

- provides linear scale-out and fault tolerance
- but suffers from data transfer latency

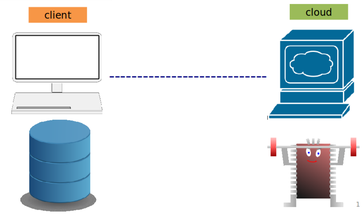


Source: http://bowtie-bio.sourceforge.net/crossbow

The Problem

Big Data: Large amounts of data stored on the Client

Big Processing Power: High number of CPU cores available in the Cloud

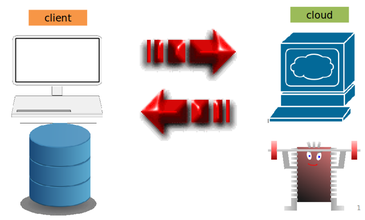


Bottleneck: Limited network bandwidth between the Client and the Cloud

The Stream-As-You-Go Approach

Key Idea

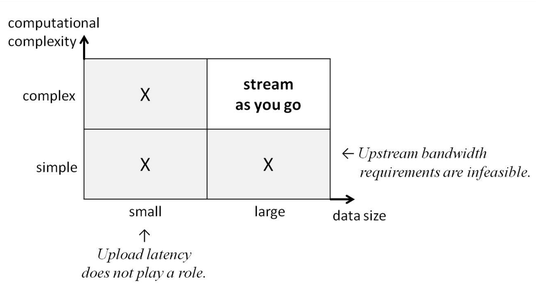
- Data is streamed into the cloud.
- Incremental processing starts right away.
- Results are streamed back to the client as they become available.



Key Benefits

- hides the data transfer latency, significantly reducing the total round trip time
- enables in-memory processing, saving from data access time and cloud storage costs
- enables pipelined parallelism in addition to partitioned parallelism, leading to early results and shorter completion time

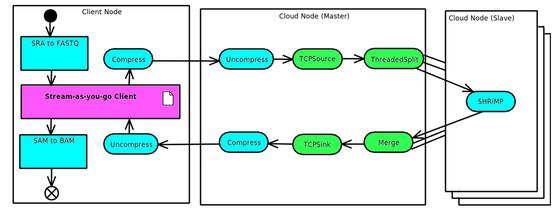
Good Fit for Data- and Compute-intensive Cloud Applications



Stream-As-You-Go in Action

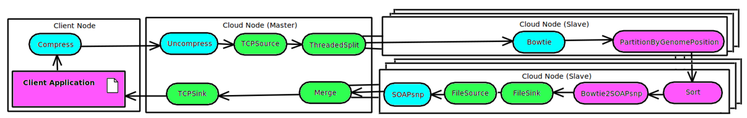
Incremental Read Alignment Process

1. Data is compressed and streamed into the cloud.
2. Data gets uncompressed, split, and aligned by SHRiMP in parallel on multiple nodes.
3. Results get merged, compressed, and sent back to the client as soon as they become available.



Incremental SNP Detection Process

1. Data is compressed and streamed into the cloud.
2. Data gets uncompressed, split, and aligned by Bowtie in parallel on multiple nodes.
3. Aligned reads get partitioned by their genome positions.
4. They are then sorted in parallel using a distributed in-memory insertion sort based on red-black trees.
5. Data is coverted into SOAPsnp input format.
6. SNP calling is performed using SOAPsnp.
7. Results get merged and sent back to the client as soon as they become available.



Experimental Results