Use of Cloud Computing for Bioinformatics

A collaboration between the San Francisco State University CS Department, SFSU Center for Computing and Life Sciences and Stanford Helix Group

Jinesh Lalan, Mike Wong and Dr. Dragutin Petkovic

Use of cloud computing for bioinformatics: case study of Amazon Elastic Compute Cloud EC2

Supported in part by NIH grant LM05652, SFSU Center for Computing for Life Sciences and Amazon Educational Grant

Background

- Genomic and proteomic sequencing data have grown tremendously in size because of advancing in sequencing technology.
- Consequently, analysis is intractably computationally intensive for common PC architectures and building and operating even a relatively small cluster can be a formidable undertaking.
- Recently, virtual computing clouds are drawing attentions for its flexibility since users can create a high performance cluster of any number of virtual computing units.

Purpose

- Investigate applicability of Amazon Cloud Computing EC2 for typical bioinformatics applications, specifically for performance and ease of use.
- Provide tutorials and setup SW for non-expert users.

Methods

- We downloaded the Anopheles Gambiae Gene data from Vector Base.
- We ran BLAST on EC2 with 30, 40 and 50 instances and measured the run time of each. At least three runs of each number of instances were carried out. Instance type used for the research was “Small” (1.7 GB of Memory, 1 -1.2 GHz Opteron or Xen Processor, 1 virtual core).
- We also carried out BLAST runs on a local cluster of 36 cores, 152 GB of Memory and 72 GB of Disks.

System Overview

Cloud Computing Vs. Local Cluster

- Cheap - Zero cost of equipments, physical space, Cooling, Power, maintenance.
- Highly Scalable and Elastic - Can scale Resources up and down as and when you want.
- Your Jobs don’t have to wait in the Queue.
- Root access to your Nodes.
- Researcher with no access to a local cluster can have their results cheaply within hours.

Discussion

- Looking at the graphs, the performance seems to be stable and scales linearly.
- Noisy Neighbor Problem: A noisy neighbor here means a Virtual Machine that uses the shared resources disproportionately. Sadly, you cannot choose your neighbors.
- Network Bandwidth issues (Internal as well as external).
- Provides Simple and Easy tutorials to get started using AWS services.
- Various alternative Instance types. Instance types specially designed for bioinformatics. Various instance types have different compute capacity.
- Public Datasets on AWS can be accessed with no charge.

Results

Number of Nodes

Genome Analysis Resources on AWS

- Publicly available Amazon Machine Images and Volumes for Bioinformatics
  a. ami-4e57a227 – 64 bit image
  b. ami-a0b912de – 32 bit image
  c. snap-84e77d1f – Data Volume [5]
- Some of the tools installed on the above AMI’s: PRIMER3, MUSCLE, BLAST2, EMBOSS, HMMER, READSEQ, BOWTIE, NOVOALIGN [5]
- Databases available on the above mentioned volume
  a. Genome sequences pre-indexed for search with next-gen aligners like BOWTIE, NOVOALIGN
  b. UnRef protein databases, indexed for searching with BLAST+ [5]

Amazon Web Services

References

3. EC2. [http://aws.amazon.com/ec2/]

Acknowledgements

I would like to thank the gracious support and guidance of Dr. Russ Altman, Dr. Dragutin Petkovic, Mike Wong, Ljubomir Buturovic.