Beyond the Genome: Cloud-scale computing demo
Michael Schatz, Ben Langmead, & James Taylor

Sept. 19, 2011
Beyond the Genome
Beyond the Genome Challenge

http://schatzlab.cshl.edu/data/btg11.tgz

http://aws.amazon.com/awscredits

The goal is to identify a viral sequence insertion into a human cancer exome. To keep it tractable, we will only use genes on chromosome 22, and only exons > 500bp long.

If you have questions, tweet #btg11

Submit your solution to: mschatz@cshl.edu

The subject line should be: BTG2011 human_gene virus_name

The body should contain all the steps you took to identify the gene and virus. If at all possible, please include the exact commands used. Winners will be selected by first correct answer (name of gene, name of virus) and for reproducibility. You must be registered and present at Beyond the Genome 2011 to win. The judges decisions are final. Rules are subject to change at anytime.
Amazon Web Services
http://aws.amazon.com

• All you need is a credit card, and you can immediately start using one of the largest datacenters in the world

• Elastic Compute Cloud (EC2)
  – On demand computing power
    • Support for Windows, Linux, & OpenSolaris
    • Starting at 8.5¢ / core / hour

• Simple Storage Service (S3)
  – Scalable data storage
    • 10¢ / GB upload fee, 15¢ / GB monthly fee

• Plus many others
EC2 Architecture

• Very large pool of machines
  – Effectively infinite resources
  – High-end servers with many cores and many GB RAM

• Machines run in a virtualized environment
  – Amazon can subdivide large nodes into smaller instances
  – You are 100% protected from other users on the machine
  – You get to pick the operating system, all installed software
Amazon Machine Images

- A few Amazon sponsored images
  - Suse Linux, Windows

- Many Community Images & Appliances
  - CloudBioLinux: Genomics Appliance
  - Crossbow: Hadoop, Bowtie, SOAPsnp
  - Galaxy: CloudMan

- Build you own
  - Completely customize your environment
  - Your results could be totally reproducible
Amazon S3

- S3 provides persistent storage for large volumes of data
  - Very high speed connection from S3 to EC2 compute nodes
  - Public data sets include s3://1000genomes

- Tiered pricing by volume
  - Pricing starts at 14¢ / GB / month
  - 5.5¢ / GB / month for over 5 PB
  - Pay for transfer out of Amazon

- Import/Export service for large volumes
  - FedEx your drives to Amazon
Getting Started

http://docs.amazonaws.com/AWSEC2/latest/GettingStartedGuide/

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that enables you to launch and manage Linux/UNIX and Windows server instances in Amazon’s data centers. You can get started with Amazon EC2 by following the tasks shown in the following diagram. You’ll primarily use the AWS Management Console, a point-and-click web-based interface.

Sign up for EC2 → Launch Instance → Connect to Linux/UNIX instance → Connect to Windows instance → Terminate Instance

This guide walks you through launching and connecting to your first Amazon EC2 instance. To start, click the following Get Started button.

Get Started
Signing Up

Introducing Amazon Simple Notification Service
Enable applications, end-users, and devices to instantly send and receive notifications from the cloud.
› Learn More...

News & Events

What's New  Media Coverage  Upcoming Events

Oct 14, 2010  Amazon Elastic Load Balancing Adds Support for HTTPS
Oct 12, 2010  AWS Management Console Adds Support for Amazon SNS
Oct 05, 2010  Amazon EC2 Running SUSE Linux Now Available

Oct 05, 2010  Announcing Read Replicas, Lower High Memory DB Instance Prices for Amazon RDS
Sep 29, 2010  Announcing the AWS SDK for PHP
Sep 21, 2010  Oracle Certifies Enterprise Software on Amazon EC2

Get Started

Business Managers
Learn how Amazon Web Services enables you to reach business goals faster:
• Solutions & Use Cases
• Security Center
• Economics Center
• Case Studies
• Service Health Dashboard
• Solution Providers
• Videos & Webinars
AWS Management Console
Running your First Cloud Analysis

1. Pick your AMI
   - Machine Image: Operating System & Tools

2. Pick your instance type & quantity
   - Micro - High-Memory Quadruple Extra Large

3. Pick your credentials
   - SSH Keys

4. Configure your Firewall
   - Protect your servers

5. Launch!
1. Pick your AMIs

- **Basic 32-bit Amazon Linux AMI 1.0** (AMI Id: ami-3ac33653)
  - Amazon Linux AMI Base 1.0, EBS boot, 32-bit architecture with Amazon EC2 AMI Tools.

- **Basic 64-bit Amazon Linux AMI 1.0** (AMI Id: ami-38c33651)
  - Amazon Linux AMI Base 1.0, EBS boot, 64-bit architecture with Amazon EC2 AMI Tools.

- **SUSE Linux Enterprise Server 11 32-bit** (AMI Id: ami-e0a35789)
  - SUSE Linux Enterprise Server 11 Service Pack 1 basic install, EBS boot, 32-bit architecture with Amazon EC2 AMI Tools preinstalled; Apache 2.2, MySQL 5.0, PHP 5.3, Ruby 1.7, and Rails 2.3

- **SUSE Linux Enterprise Server 11 64-bit** (AMI Id: ami-e04a3578d)
  - SUSE Linux Enterprise Server 11 Service Pack 1 basic install, EBS boot, 64-bit architecture with Amazon EC2 AMI Tools preinstalled; Apache 2.2, MySQL 5.0, PHP 5.3, Ruby 1.7, and Rails 2.3

- **Getting Started on Microsoft Windows Server 2008** (AMI Id: ami-c5e40dac)
  - Microsoft Windows Server 2008 R1 SP2 Datacenter edition, 32-bit architecture, Microsoft SQL Server 2008 Express, Internet Information Services 7, ASP.NET 3.5.
CloudBioLinux
2. Pick your Instance Type
3. Pick your Credentials

Public/private key pairs allow you to securely connect to your instance after it launches. To create a key pair, enter a name and click Create & Download your Key Pair. You will then be prompted to save the private key to your computer. Note, you only need to generate a key pair once - not each time you want to deploy an Amazon EC2 Instance.

Choose from your existing Key Pairs

Create a new Key Pair

1. Enter a name for your key pair: *
   - Name: adminkey (e.g., jdoekey)

2. Click to create your key pair:
   - Create & Download your Key Pair

Proceed without a Key Pair

Save this file in a place you will remember. You can use this key pair to launch other instances in the future or visit the Key Pairs page to create or manage existing ones.
4. Configure your Firewall
5. Launch!
Monitoring your Server

![AWS Management Console](image)

<table>
<thead>
<tr>
<th>Name</th>
<th>Instance</th>
<th>AMI ID</th>
<th>Root Device</th>
<th>Type</th>
<th>Status</th>
<th>Security Groups</th>
<th>Key Pair Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>empty</td>
<td>i-4aab6027</td>
<td>ami-3ac33653</td>
<td>ebs</td>
<td>t1.micro</td>
<td>running</td>
<td>crossbow4847465</td>
<td>ssg-keypair</td>
</tr>
</tbody>
</table>

0 EC2 Instances selected

*Select an instance above*
Connecting (1)
Connecting (2)

To access your instance using any SSH client:

1. Open the SSH client of your choice (e.g., PuTTY, terminal).
2. Locate your private key file, gsg-keypair.pem.
3. Use `chmod` to make sure your key file isn't publicly viewable, e.g., `chmod 400 gsg-keypair.pem`.
4. Connect to your instance using instance's public DNS, e.g., `ssh -i gsg-keypair.pem root@ec2-184-72-85-153.compute-1.amazonaws.com`.

Example:

Enter the following command line:

```bash
ssh -i gsg-keypair.pem root@ec2-184-72-85-153.compute-1.amazonaws.com
```
chmod 400 mschatz.pem

scp -r -i mschatz.pem data.tgz ubuntu@ec2-174-129-123-73.compute-1.amazonaws.com:
ssh -i mschatz.pem ubuntu@ec2-174-129-123-73.compute-1.amazonaws.com

<remote>

ls

tar xzvf data.tgz
bowtie -S data/genomes/e_coli data/reads/e_coli_10000snp.fq ec_snp.sam
samtools view -bS -o ec_snp.bam ec_snp.sam
samtools sort ec_snp.bam ec_snp.sorted

samtools pileup -cv -f data/genomes/NC_008253.fna ec_snp.sorted.bam > snps

samtools index ec_snp.sorted.bam
samtools tview ec_snp.sorted.bam data/genomes/NC_008253.fna

exit

<local>

scp -i mschatz.pem ubuntu@ec2-174-129-123-73.compute-1.amazonaws.com:snps .
s3cmd --configure

# cp data/.s3cfg .

s3cmd ls s3://1000genomes

s3cmd ls s3://1000genomes/Pilots_Bam/NA20828/

s3cmd get s3://1000genomes/Pilots_Bam/NA20828/*chr22* .

samtools view NA20828.SLX.maq.SRP000033.2009_09.chr22_1_49691432.bam
Terminating

Total cost: 8.5¢
Reflections

• Launching and managing virtual clusters with the AWS Console is quick and easy
  – Entirely scriptable using ec2 tools
  – iPhone App also available

• Things get really interesting on 168 cores
  – 1 week CPU = 1 hour wall

Just 3 commands to bring up a 168 core (21 node) cluster & crunch terabytes:

$HADOOP/src/contrib/ec2/bin/hadoop-ec2 launch-cluster HADOOP 21
$HADOOP/src/contrib/ec2/bin/hadoop-ec2 <hadoop cmd> HADOOP
$HADOOP/src/contrib/ec2/bin/hadoop-ec2 terminate-cluster HADOOP
Thank You!

http://schatzlab.cshl.edu
@mike_schatz / #btg