

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
 - You 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.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/>

The screenshot shows a web browser window titled "Amazon Elastic Compute Cloud" with the URL <http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/>. The page header includes the Amazon logo and "amazon web services". The main content area is titled "Amazon Elastic Compute Cloud Getting Started Guide (API Version 2010-08-31)". A left sidebar contains a table of contents with items like "Get Started with EC2", "Sign Up for EC2", "Launch an Instance", "Connect to Your Linux/UNIX Instance", "Connect to Your Windows Instance", "Terminate Your Instance", "Where Do I Go from Here?", "Please Provide Feedback", and "About This Guide". The main content area has a "Documentation Feedback" link and a "Welcome" message. The main heading is "Get Started with EC2". Below this is a paragraph explaining that Amazon Elastic Compute Cloud (Amazon EC2) is a web service for launching and managing Linux/UNIX and Windows server instances. A flowchart illustrates the process: "Sign up for EC2" leads to "Launch instance", which then branches into "Connect to Linux/UNIX instance" and "Connect to Windows instance", both of which lead to "Terminate instance". Below the flowchart is a paragraph stating that the guide walks through launching and connecting to the first Amazon EC2 instance, and a "Get Started" button with a right-pointing arrow.

Amazon Elastic Compute Cloud
Getting Started Guide (API Version 2010-08-31)

Documentation Feedback

Welcome

Get Started with EC2

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.

```
graph LR; A[Sign up for EC2] --> B[Launch instance]; B --> C[Connect to Linux/UNIX instance]; B --> D[Connect to Windows instance]; C --> E[Terminate instance]; D --> E;
```

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

The screenshot shows the Amazon Web Services website in a browser window. The address bar displays `http://aws.amazon.com/`. The page features the AWS logo and navigation links for **AWS**, **Products**, **Developers**, **Community**, **Support**, and **Account**. At the top right, there are links for **Sign in to the AWS Management Console** and **Create an AWS Account**, along with a language selector set to **English**.

The main content area highlights the **Introducing Amazon Simple Notification Service** with the text: "Enable applications, end-users, and devices to instantly send and receive notifications from the cloud." Below this is a diagram of a cloud with arrows pointing to various devices and a "Learn More..." link.

On the right side, a green box contains the text: "Sign up for a free Amazon Web Services Account" and a prominent **Sign Up Now** button.

The **Get Started** section includes a **Business Managers** heading and a list of resources: Solutions & Use Cases, Security Center, Economics Center, Case Studies, Service Health Dashboard, and Solution Providers.

The **News & Events** section has tabs for **What's New?**, **Media Coverage**, and **Upcoming Events**. Under **What's New?**, there are three entries:

Date	Event/Update	Date	Event/Update
Oct 14, 2010	Amazon Elastic Load Balancing Adds Support for HTTPS	Oct 05, 2010	Announcing Read Replicas, Lower High Memory DB Instance Prices for Amazon RDS
Oct 12, 2010	AWS Management Console Adds Support for Amazon SNS	Sep 29, 2010	Announcing the AWS SDK for PHP
Oct 05, 2010	Amazon EC2 Running SUSE Linux Now Available	Sep 21, 2010	Oracle Certifies Enterprise Software on Amazon EC2

At the bottom left of the news section is an **RSS** icon, and at the bottom right is a **View all** link.

AWS Management Console

The screenshot displays the AWS Management Console interface for the Amazon EC2 service. The browser address bar shows the URL `https://console.aws.amazon.com/ec2/home`. The page header includes navigation links for AWS, Products, Developers, Community, Support, and Account, along with a user greeting "Welcome, CBCB" and links for Settings and Sign Out.

The main navigation pane on the left lists various AWS services, with Amazon EC2 selected. The EC2 navigation menu includes sections for INSTANCES (EC2 Dashboard, Instances, Spot Requests), IMAGES (AMIs, Bundle Tasks), ELASTIC BLOCK STORE (Volumes, Snapshots), and NETWORKING & SECURITY (Elastic IPs, Security Groups, Placement Groups, Load Balancers, Key Pairs). The region is set to US East.

The central content area, titled "Amazon EC2 Console Dashboard", features several key sections:

- Getting Started:** A yellow box with the text "To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance." and a prominent "Launch Instance" button. A note below states: "Note: Your instances will launch in the US East (Virginia) region."
- My Resources:** A summary of resources in the US East (Virginia) region, including:
 - 0 Running Instances
 - 0 Elastic IPs
 - 0 EBS Volumes
 - 0 EBS Snapshots
 - 2 Key Pairs
 - 27 Security Groups
 - 0 Load Balancers
 - 0 Placement Groups
- Service Health:** A table showing the current status of the Amazon EC2 service in the US East - N. Virginia region, which is "operating normally".
- Related Links:** A list of links for documentation, all EC2 resources, forums, feedback, and a "Banned an Issue" link.

The footer of the page contains copyright information for 2008-2009, Amazon Web Services LLC, and links for Feedback, Support, Privacy Policy, and Terms of Use, along with the Amazon logo and the text "An amazon.com company".

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!

I. Pick your AMIs

The screenshot shows the AWS Management Console interface. The main window is titled "Request Instances Wizard" and is currently on the "CHOOSE AN AMI" step. The wizard has a progress bar at the top with five steps: "CHOOSE AN AMI", "INSTANCED DETAILS", "ELASTIC NET PAIR", "EBS VOLUME INFORMATION", and "BIOS". Below the progress bar, there is a instruction: "Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its Select button." There are three tabs: "Quick Start", "My AMIs", and "Community AMIs". The "Quick Start" tab is selected and displays a list of five AMIs, each with a logo, a title, a description, and a "Select" button. The AMIs listed are:

- 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 1.8.7, and Rails 2.3.
- SUSE Linux Enterprise Server 11 64-bit** (AMI Id: ami-e4a3578d): 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 1.8.7, 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 SQLServer 2008 Express, Internet Information Services 7, ASP.NET 3.5.

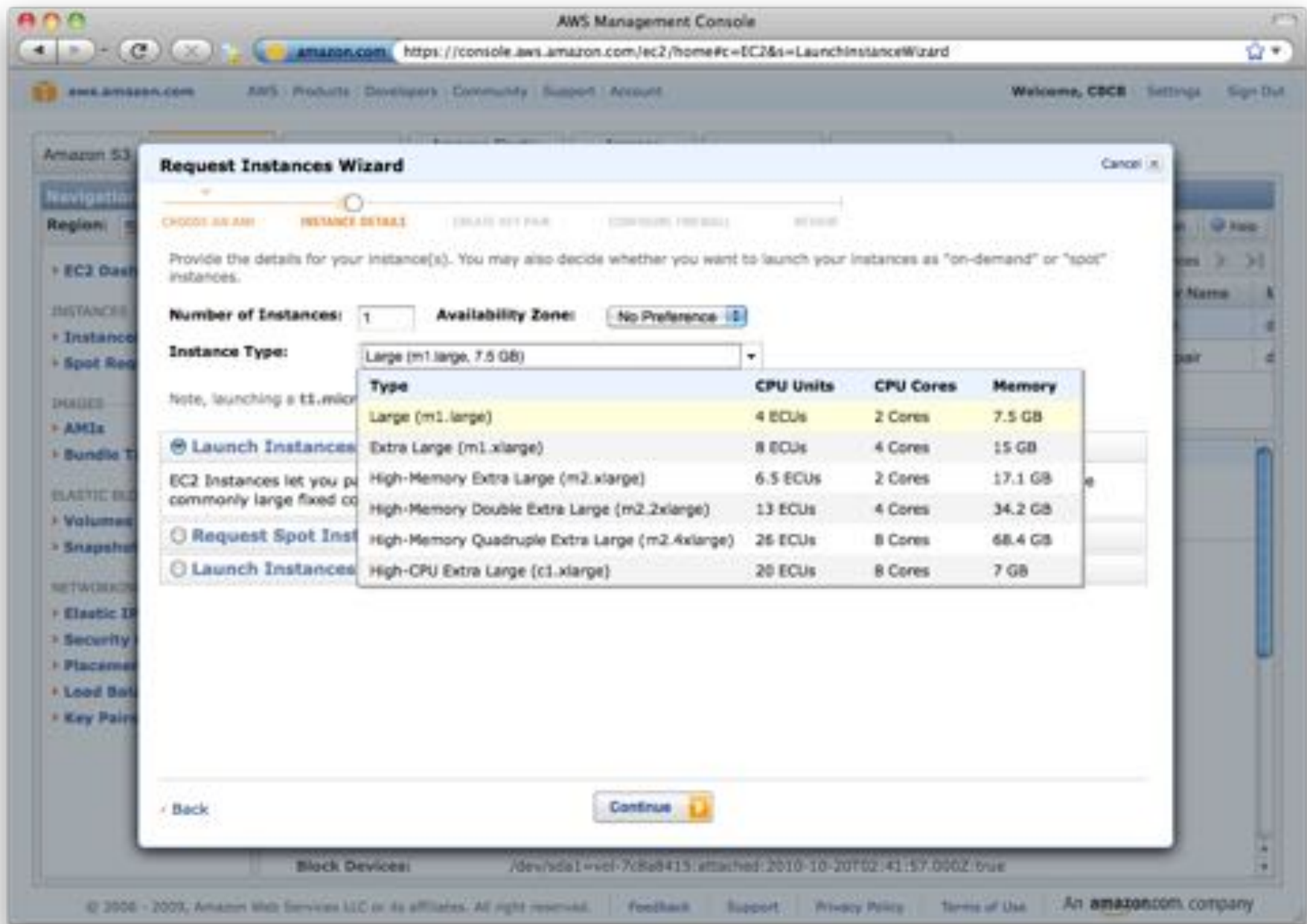
At the bottom of the wizard, there are fields for "AMI Launch Check" and "Instance IP". The "AMI Launch Check" field shows "Root Device: /dev/sda1" and "Root Device Type: sbs". The footer of the console shows copyright information: "© 2006 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved." and links for "Feedback", "Support", "Privacy Policy", and "Terms of Use". The Amazon logo and "An amazon.com company" are also present.

CloudBioLinux

The screenshot shows the AWS Management Console interface. The browser address bar displays the URL `https://console.aws.amazon.com/ec2/home#fc=EC2&s=LaunchInstanceWizard`. The page title is "Request Instances Wizard". The wizard is currently on the "CHOOSE AN AMI" step. Below the step indicator, there is a text prompt: "Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its Select button." There are three tabs: "Quick Start", "My AMIs", and "Community AMIs". The "Community AMIs" tab is selected. Below the tabs, there is a search bar with the text "cloudbiolinux" and a dropdown menu set to "All Images". Below the search bar is a table of AMIs:

AMI ID	Root Device	Manifest	Platform	
ami-0af91263	ebs	678711657553/CloudBioLinux 32bit 20100716	Other Linux	Select
ami-4e57a227	ebs	678711657553/CloudBioLinux 64bit 20100929	Other Linux	Select
ami-6953b200	instance-store	jcvi-cloudbiolinux/JCVI-Cloud-BioLinux.manifest.xml	Other Linux	Select
ami-879c75ee	ebs	678711657553/CloudBioLinux 20100507	Other Linux	Select

2. Pick your Instance Type



The screenshot shows the AWS Management Console's 'Request Instances Wizard' in the 'INSTANCE DETAILS' step. The wizard is titled 'Request Instances Wizard' and has a progress bar with four steps: 'CHOOSE AN AMI', 'INSTANCE DETAILS', 'SELECT KEY PAIR', and 'CONFIGURE FIREWALL'. The 'INSTANCE DETAILS' step is currently active.

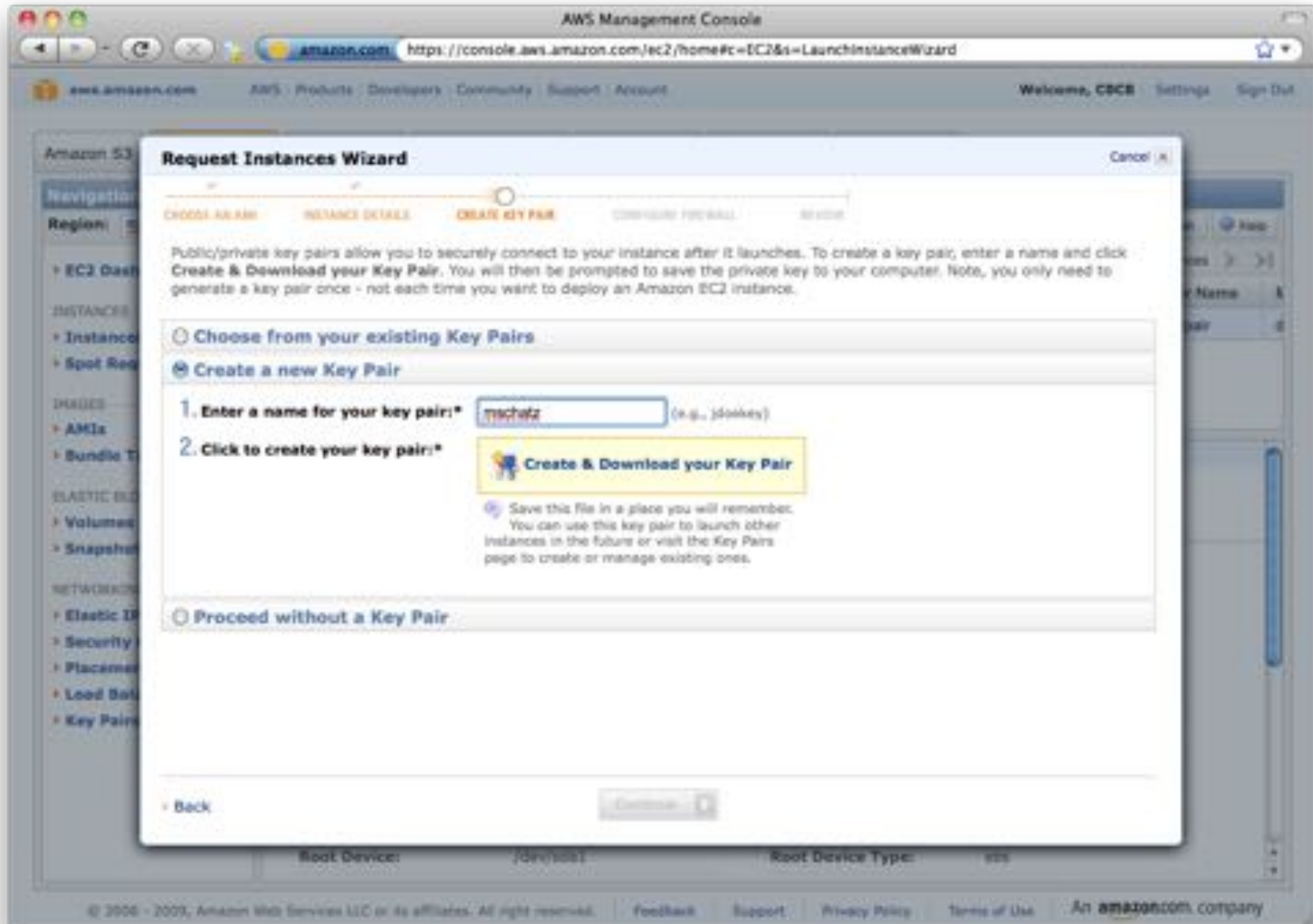
Below the progress bar, there is a text prompt: 'Provide the details for your instance(s). You may also decide whether you want to launch your instances as "on-demand" or "spot" instances.'

The 'Number of Instances' is set to 1, and the 'Availability Zone' is set to 'No Preference'. The 'Instance Type' dropdown menu is open, showing a list of instance types with their respective specifications:

Type	CPU Units	CPU Cores	Memory
Large (m1.large)	4 ECUs	2 Cores	7.5 GB
Extra Large (m1.xlarge)	8 ECUs	4 Cores	15 GB
High-Memory Extra Large (m2.xlarge)	6.5 ECUs	2 Cores	17.1 GB
High-Memory Double Extra Large (m2.2xlarge)	13 ECUs	4 Cores	34.2 GB
High-Memory Quadruple Extra Large (m2.4xlarge)	26 ECUs	8 Cores	68.4 GB
High-CPU Extra Large (c1.xlarge)	20 ECUs	8 Cores	7 GB

The 'Continue' button is highlighted in yellow. At the bottom of the wizard, there is a 'Back' button and a 'Continue' button with a yellow arrow. The footer of the console shows the copyright information: '© 2006 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved.' and the text 'An amazon.com company'.

3. Pick your Credentials



The screenshot shows the AWS Management Console interface. The main content area is a modal window titled "Request Instances Wizard" with a "Cancel" button in the top right. The wizard has five steps: "CHOOSE AN AMI", "INSTANCE DETAILS", "CREATE KEY PAIR" (which is the active step), "CONFIGURE FIREWALL", and "START".

Below the step indicators, there is a paragraph of text: "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."

There are three radio button options:

- Choose from your existing Key Pairs
- Create a new Key Pair
- Proceed without a Key Pair

Under "Create a new Key Pair", there are two numbered steps:

1. Enter a name for your key pair:* (e.g., jdonkey)
2. Click to create your key pair:*

Below step 2, there is a note: "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."

At the bottom of the wizard, there are "Back" and "Continue" buttons. Below the wizard, the "Root Device" is set to "/dev/sda1" and the "Root Device Type" is "efs".

The footer of the console contains: "© 2008 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved. Feedback Support Privacy Policy Terms of Use An amazon.com company"

4. Configure your Firewall

Request Instances Wizard

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | **CONFIGURE FIREWALL** | REVIEW

Security groups determine whether a network port is open or blocked on your instances. You may use an existing security group, or we can help you create a new security group to allow access to your instances using the suggested ports below. Add additional ports now or update your security group anytime using the Security Groups page. All changes take effect immediately.

Choose one or more of your existing Security Groups

Create a new Security Group

1. Name your Security Group:

2. Describe your Security Group:

3. Define allowed Connections

Application	Transport	Port	Source Network (IPv4 CIDR)	Actions
HTTP	TCP	80	All Internet	<input type="button" value="Remove"/>
SSH	TCP	22	All Internet	<input type="button" value="Remove"/>
Select...	-	-	All Internet Change	<input type="button" value="Add Rule"/>

Block Devices: /dev/sda1=vol-7c8a8415,attached:2010-10-20T02:41:57.000Z,true

© 2006 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved. [Feedback](#) [Support](#) [Privacy Policy](#) [Terms of Use](#) An **amazon.com** company

5. Launch!

The screenshot shows the AWS Management Console interface with the 'Request Instances Wizard' modal open. The wizard is in the 'Review' step, indicated by a progress bar at the top. The configuration details are as follows:

- AMI:** Amazon Linux AMI ID ami-3ac33653 (1386)
Name: Basic 32-bit Amazon Linux AMI 1.0
Description: Amazon Linux AMI Base 1.0, EBS boot, 32-bit architecture with Amazon EC2 AMI Tools. [Edit AMI](#)
- Number of Instances:** 1
Availability Zone: No Preference
Instance Type: Micro (t1.micro)
Instance Class: On Demand [Edit Instance Details](#)
- Monitoring:** Disabled
Kernel ID: Use Default
RAM Disk ID: Use Default
User Data: [Edit Advanced Details](#)
- Key Pair Name:** mschatz [Edit Key Pair](#)
- Security Group(s):** basic [Edit Firewall](#)

At the bottom of the wizard, there is a 'Back' link and a prominent 'Launch' button with a lightning bolt icon. Below the wizard, the 'Root Device' is set to '/dev/sda1' and the 'Root Device Type' is 'efs'. The footer of the console includes copyright information and links for Feedback, Support, Privacy Policy, Terms of Use, and the Amazon logo.

Monitoring your Server

The screenshot shows the AWS Management Console interface. At the top, the browser address bar displays the URL `https://console.aws.amazon.com/ec2/home#ec2:instances`. The navigation bar includes the AWS logo, navigation links for Products, Developers, Community, Support, and Account, and a user profile section for 'Welcome, CBCB' with links for Settings and Sign Out.

Service tabs are visible for Amazon S3, Amazon EC2 (selected), Amazon VPC, Amazon Elastic MapReduce, Amazon CloudFront, Amazon RDS, and Amazon SNS.

The left-hand navigation pane is expanded to show the 'My Instances' section. Under 'INSTANCES', the 'Instances' link is selected. Other options include Spot Requests, AMIs, Bundle Tasks, Volumes, and Snapshots. Under 'NETWORKING & SECURITY', options include Elastic IPs, Security Groups, Placement Groups, Load Balancers, and Key Pairs.

The main content area, titled 'My Instances', shows a table of instances. The table has columns for Name, Instance, AMI ID, Root Device, Type, Status, Security Groups, and Key Pair Name. One instance is listed:

Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
empty	i-4aab6027	ami-3ac3853	ebs	t1.micro	running	crossbow45474681	gag-keypair

Below the table, a summary bar indicates '0 EC2 Instances selected' and prompts the user to 'Select an instance above'.

The footer contains copyright information: '© 2008 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved.' and links for Feedback, Support, Privacy Policy, and Terms of Use. It also states 'An amazon.com company'.

Connecting (I)

The screenshot displays the AWS Management Console interface. At the top, the browser address bar shows the URL `https://console.aws.amazon.com/ec2/home#ec2:instances`. The console header includes navigation links for various AWS services: Amazon S3, Amazon EC2 (selected), Amazon VPC, Amazon Elastic MapReduce, Amazon CloudFront, Amazon RDS, and Amazon SNS. A navigation sidebar on the left lists categories like EC2 Dashboard, INSTANCES, IMAGES, ELASTIC BLOCK STORE, and NETWORKING & SECURITY. The main content area is titled 'My Instances' and shows a table with one instance named 'empty' (ID: i-4aab6027). A context menu is open over this instance, with the 'Connect' option highlighted. Below the table, the instance details are shown in a key-value format.

Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
empty	i-4aab6027	ami-3ac33653	ebs	t1.micro	running	crossbow464746833350	gsg-keypair

1 EC2 Instance selected

EC2 Instance: i-4aab6027

Description | Monitoring | Tags

AMI ID: ami-3ac33653
Security Groups: crossbow464746833350-X-master
Status: running
VPC ID: -
Virtualization: para
Reservation: r-81f3c6b
Platform: -
Kernel ID: aki-407d9529
AMI Launch Index: 0
Root Device: /dev/sda1

Instance Management

- Connect
- Get System Log
- Create Image (EBS AMI)
- Add/Edit Tags
- Launch More Like This
- Dissociate IP Address

Instance Lifecycle

- Terminate
- Reboot
- Stop
- Start

CloudWatch Monitoring

- Enable CloudWatch
- Disable CloudWatch

Region: us-east-1c
Instance Type: t1.micro
Owner: 464746833350
Subnet ID: -
Placement Group: -
RAM Disk ID: -
Key Pair Name: gsg-keypair
Monitoring: disabled
Elastic IP: -
Root Device Type: ebs

© 2008 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved. | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

Connecting (2)

The screenshot shows the AWS Management Console interface. A modal dialog box titled "Connect Help - Secure Shell (SSH)" is open, providing instructions for connecting to an EC2 instance. The dialog includes the following content:

Instance: i-4aab6027

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, ssh won't work otherwise:
`chmod 400 gsg-keypair.pem`
4. Connect to your instance using instance's public DNS [ec2-184-72-85-153.compute-1.amazonaws.com].

Example

Enter the following command line:

```
ssh -i gsg-keypair.pem root@ec2-184-72-85-153.com
```

The background shows the "My Instances" page with a table of instances. The instance "t1.micro" is highlighted. The table columns include Name, Security Groups, Key Pair Name, and Public IP Address.

Name	Security Groups	Key Pair Name	Public IP Address
crossbow46474683	sg-1a111111	gsg-keypair	184.72.85.153
t1.micro	sg-1a111111	gsg-keypair	184.72.85.153

At the bottom of the console, there is a footer with copyright information: © 2006 - 2009, Amazon Web Services LLC or its affiliates. All rights reserved. Links for Feedback, Support, Privacy Policy, and Terms of Use are also present.

Calling SNPs in the Cloud 😊

```
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 .
```


1000Genomes in the Cloud

```
s3cmd --configure
```

```
# cp data/.s3cfg .
```

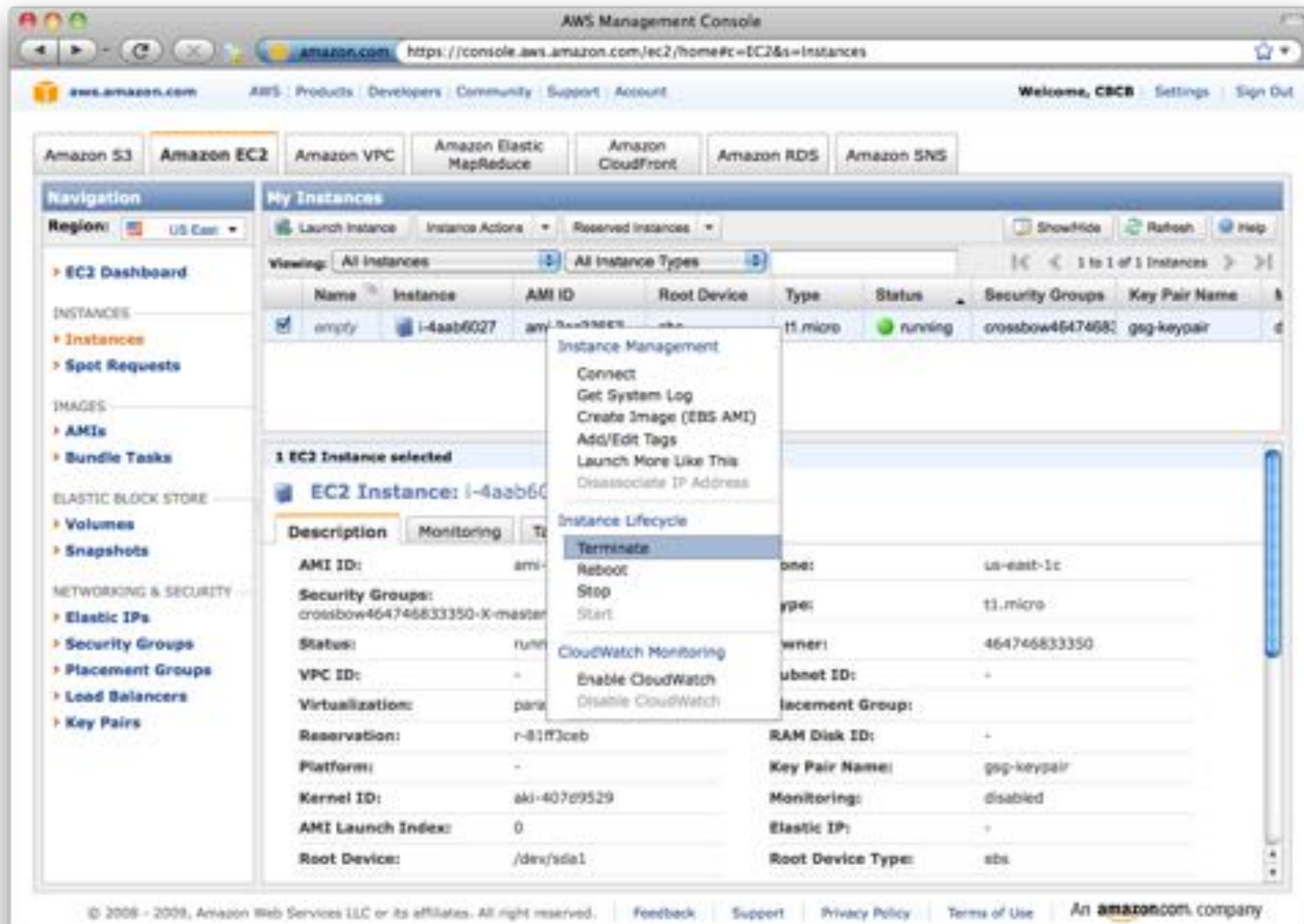
```
s3cmd ls s3://1000genomes
```

```
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](#)