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

23andMe

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.

BioMed Central
The Open Access Publisher

Oxford NANOPORE Technologies®

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

amazon webservices™

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

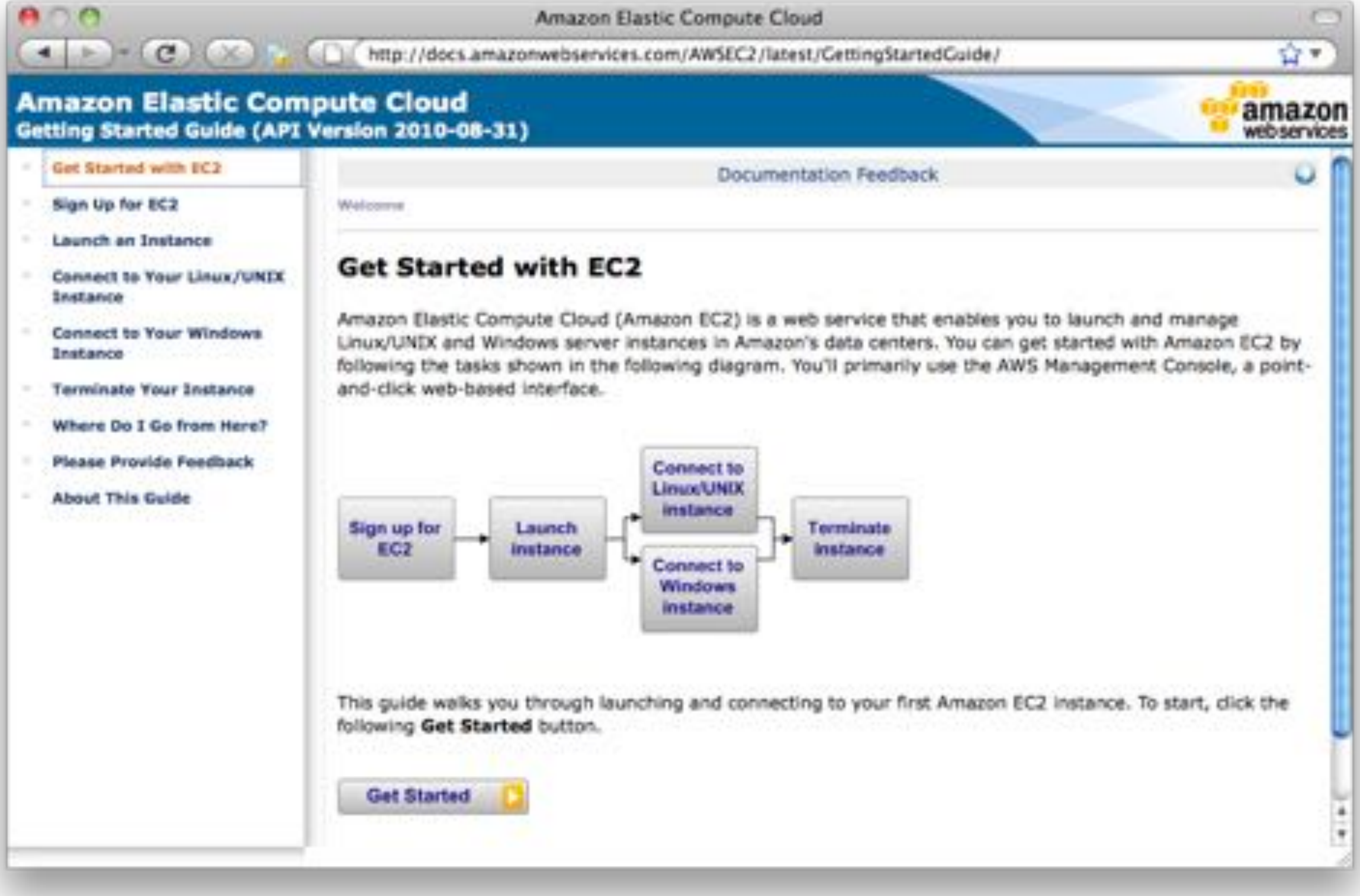
- S3 provides persistent storage for large volumes of data
  - Very high speed connection from S3 to EC2 compute nodes
  - Public data sets include s3://1000genomes

– 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/



Amazon Elastic Compute Cloud

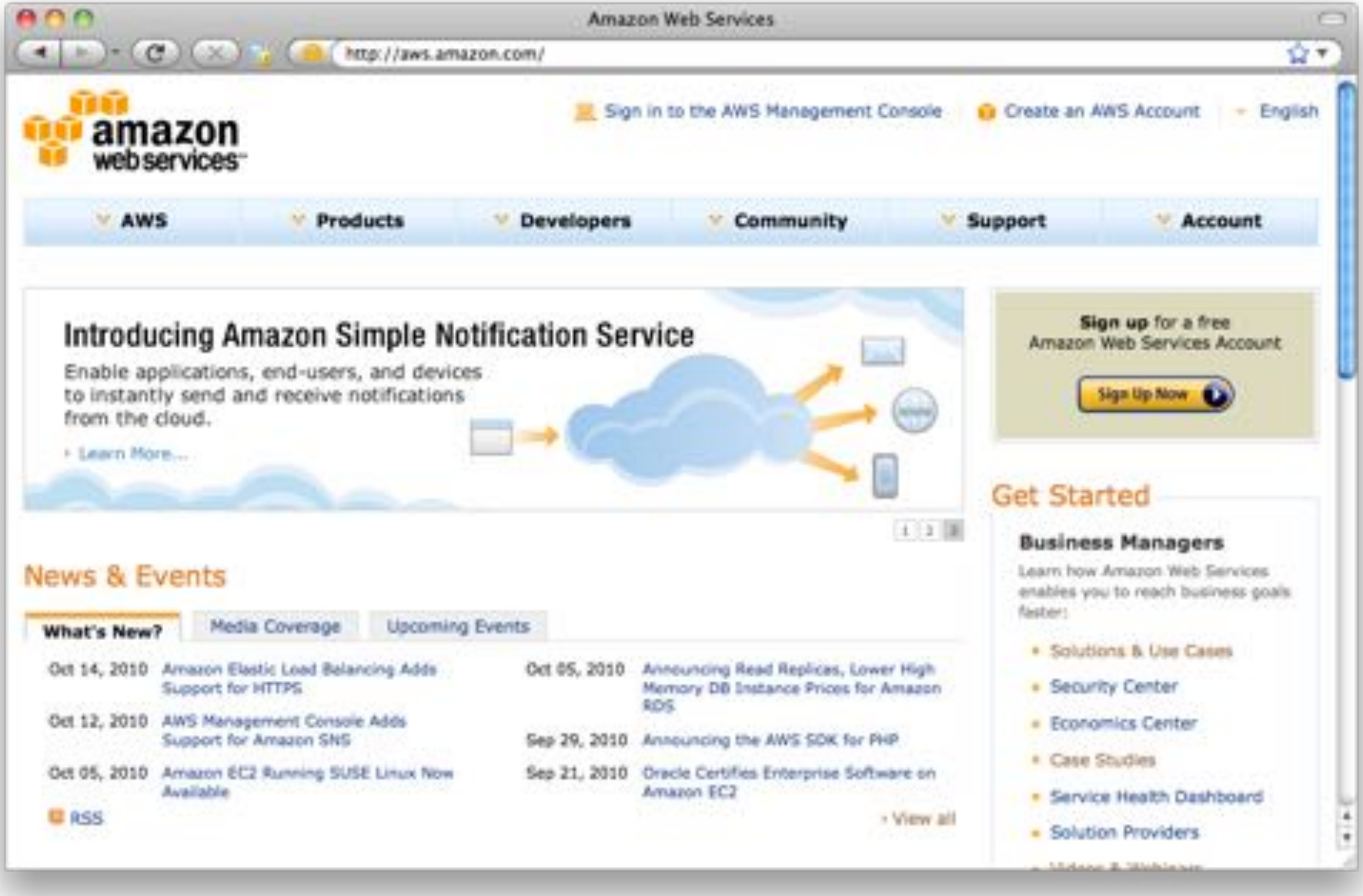
Getting Started Guide (API Version 2010-08-31)

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.

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

Signing Up



Introducing Amazon Simple Notification Service

Enable applications, end-users, and devices to instantly send and receive notifications from the cloud.

Learn More...

Sign up for a free Amazon Web Services Account

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

RSS

View all

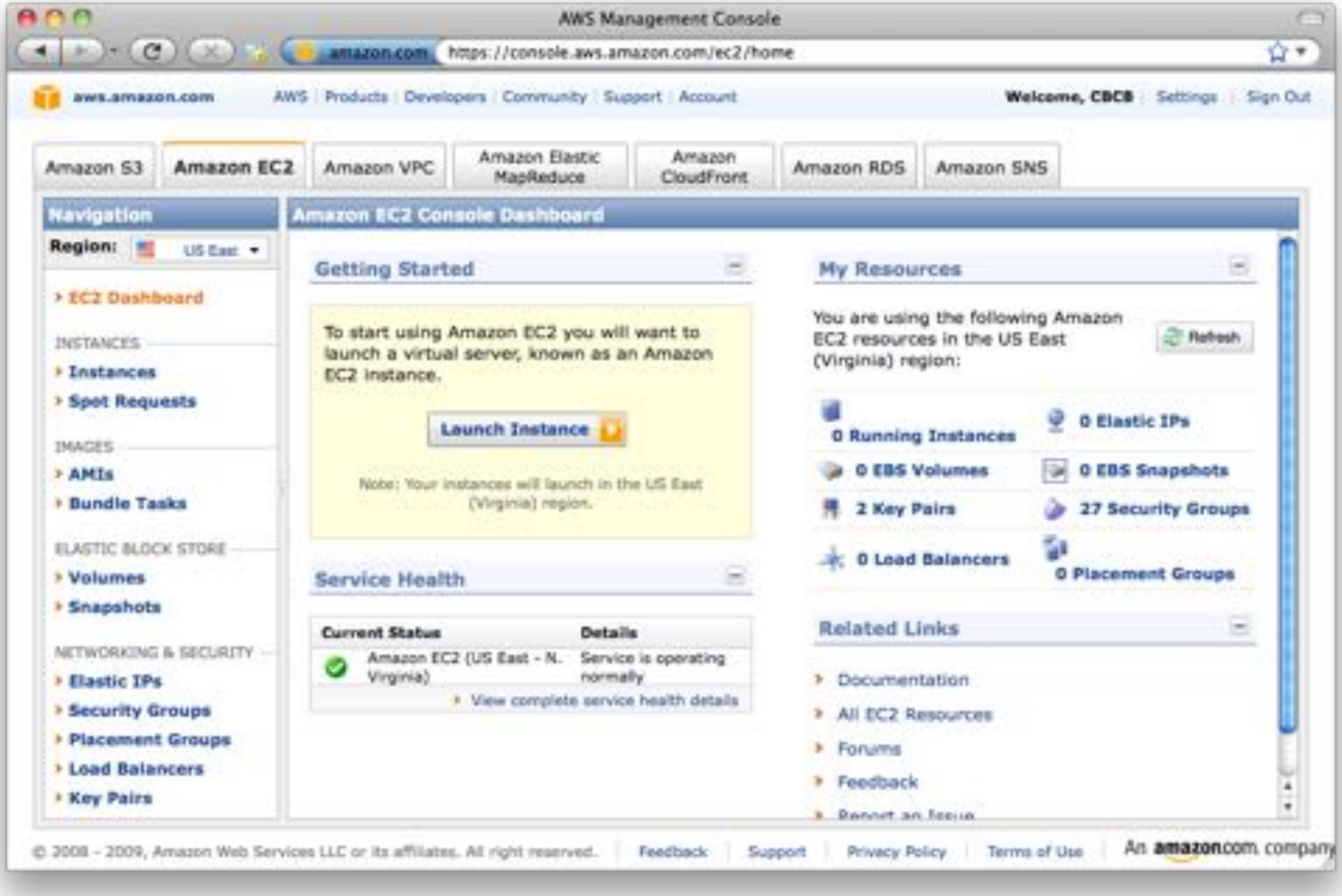
Get Started

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

AWS Management Console



To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Note: Your instances will launch in the US East (Virginia) region.

You are using the following Amazon EC2 resources in the US East (Virginia) region:

- 0 Running Instances
- 0 Elastic IPs
- 0 EBS Volumes
- 0 EBS Snapshots
- 2 Key Pairs
- 27 Security Groups
- 0 Load Balancers
- 0 Placement Groups

Current Status	Details
✓ Amazon EC2 (US East - N. Virginia)	Service is operating normally

› View complete service health details

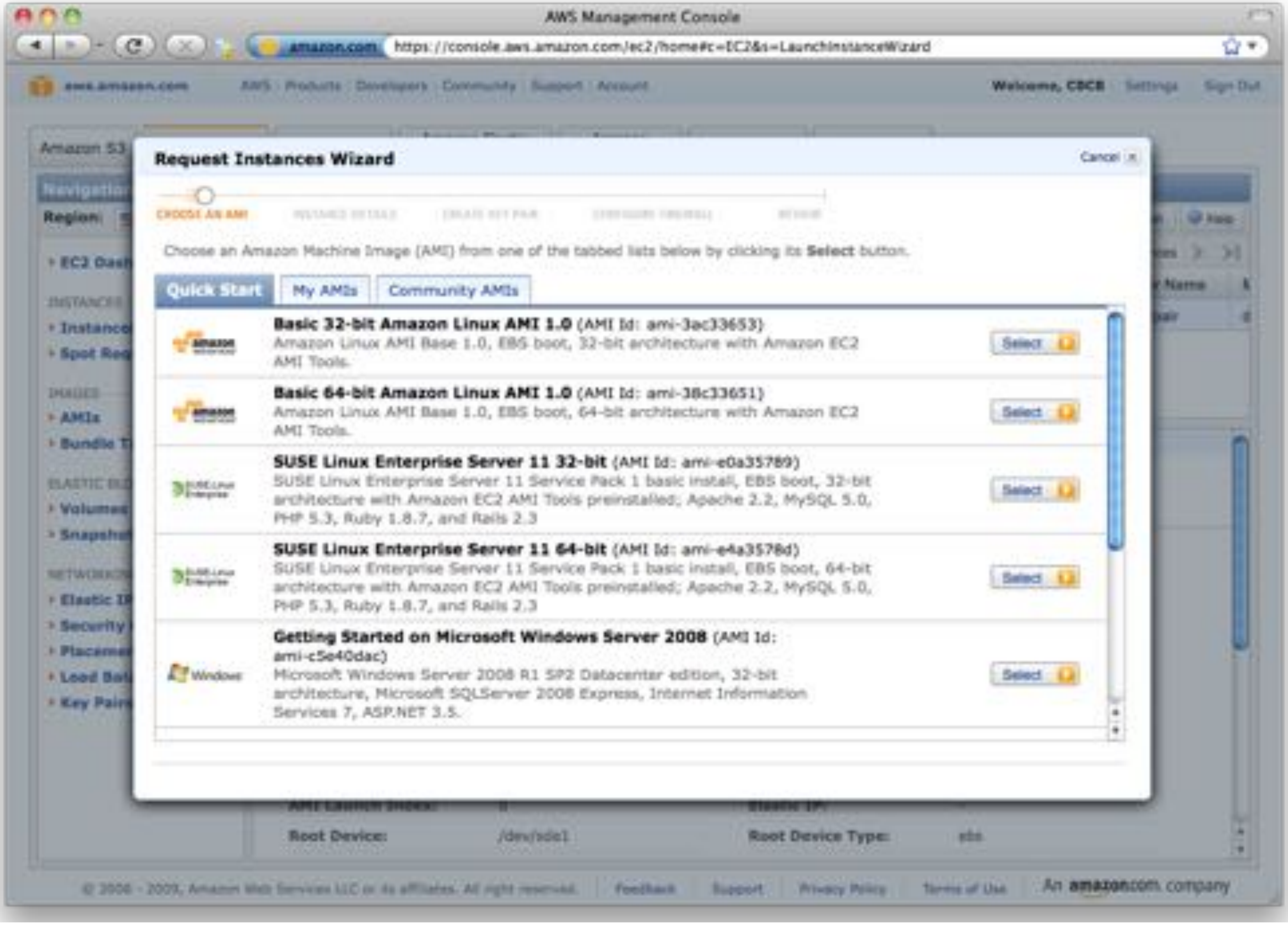
- Documentation
- All EC2 Resources
- Forums
- Feedback
- Report an Issue

© 2008 - 2009, Amazon Web Services LLC or its affiliates. All right reserved. | Feedback | Support | Privacy Policy | Terms of Use | 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!

1. Pick your AMIs



Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its **Select** button.

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.

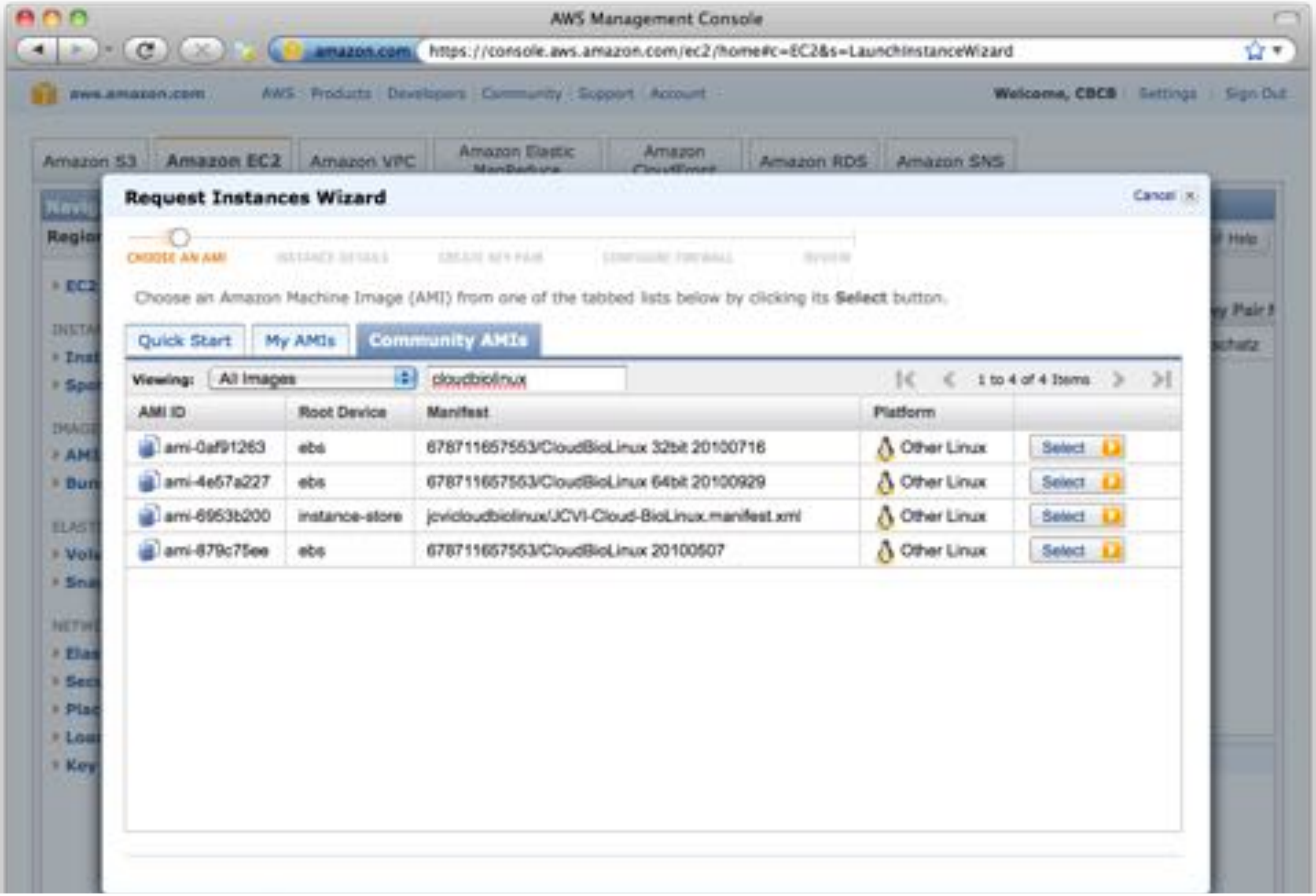
SUSE Linux Enterprise Server 11 32-bit (AMI Id: ami-e0a35789)
SUSE Linux Enterprise Server 11 Service Pack 1 basic install, EBS boot, 32-bit architecture with Amazon EC2 AMI Tools preinstalled; Apache 2.2, MySQL 5.0, PHP 5.3, Ruby 1.8.7, and Rails 2.3

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.

Root Device: /dev/sda1
Root Device Type: ebs

CloudBioLinux



https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard

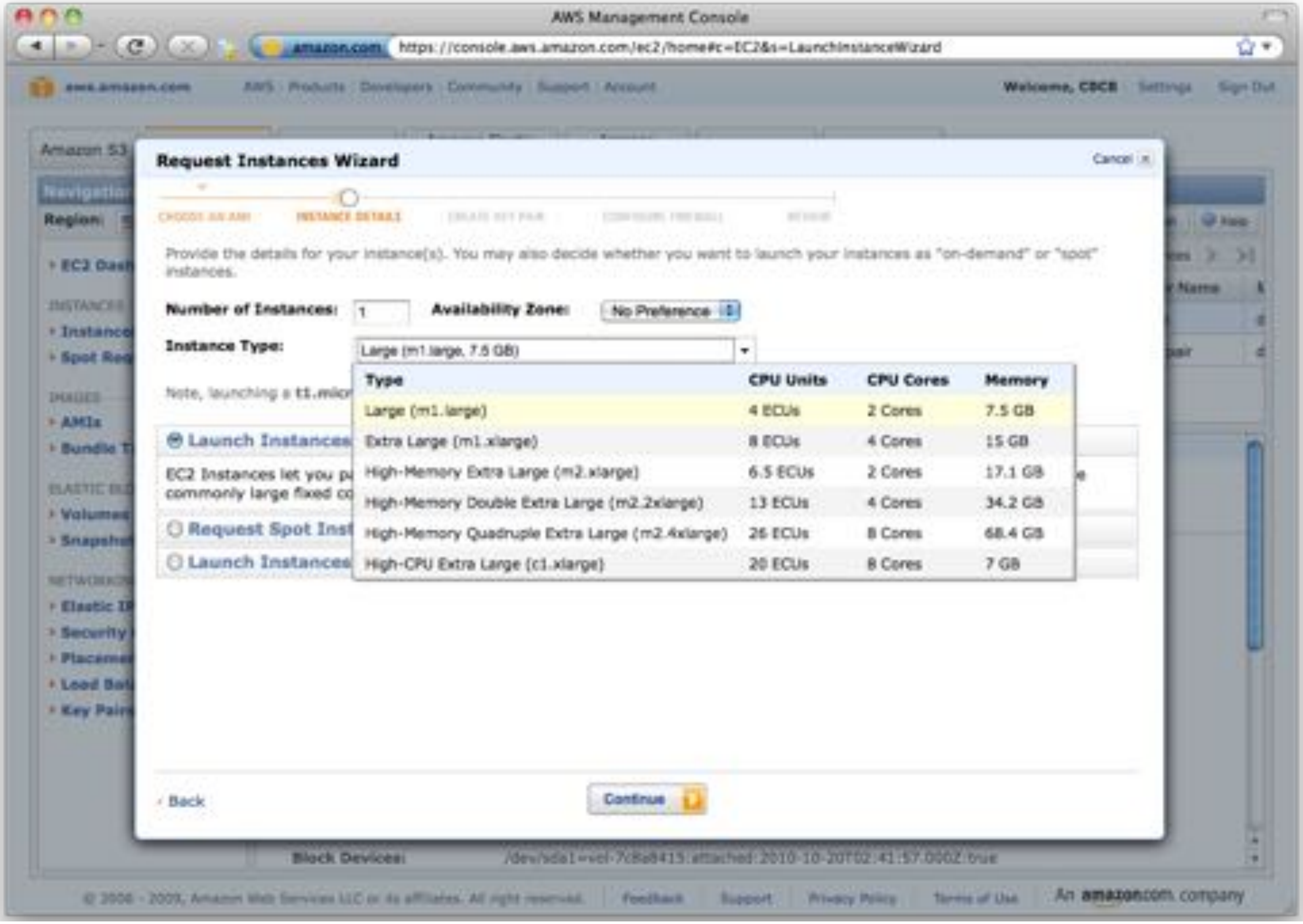
Request Instances Wizard

Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its **Select** button.

Viewing: All Images cloudbiolinux

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	jcvicloudbiolinux/JCVI-Cloud-BioLinux.manifest.xml	Other Linux	Select
ami-879c75ee	ebs	678711657553/CloudBioLinux 20100507	Other Linux	Select

2. Pick your Instance Type


AWS Management Console
https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard
aws.amazon.com
AWS | Products | Developers | Community | Support | Account
Welcome, CBCB | Settings | Sign Out
CHOOSE AN AMI
INSTANCE DETAILS
CREATE KEY PAIR
CONFIGURE FIREWALL
REVIEW
Cancel ✕
Back
Continue

Request Instances Wizard

Provide the details for your instance(s). You may also decide whether you want to launch your instances as "on-demand" or "spot" instances.

Number of Instances: 1 **Availability Zone:** No Preference

Instance Type: Large (m1.large, 7.5 GB)

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

Note, launching a t1.micr

● Launch Instances

EC2 Instances let you pa
commonly large fixed co

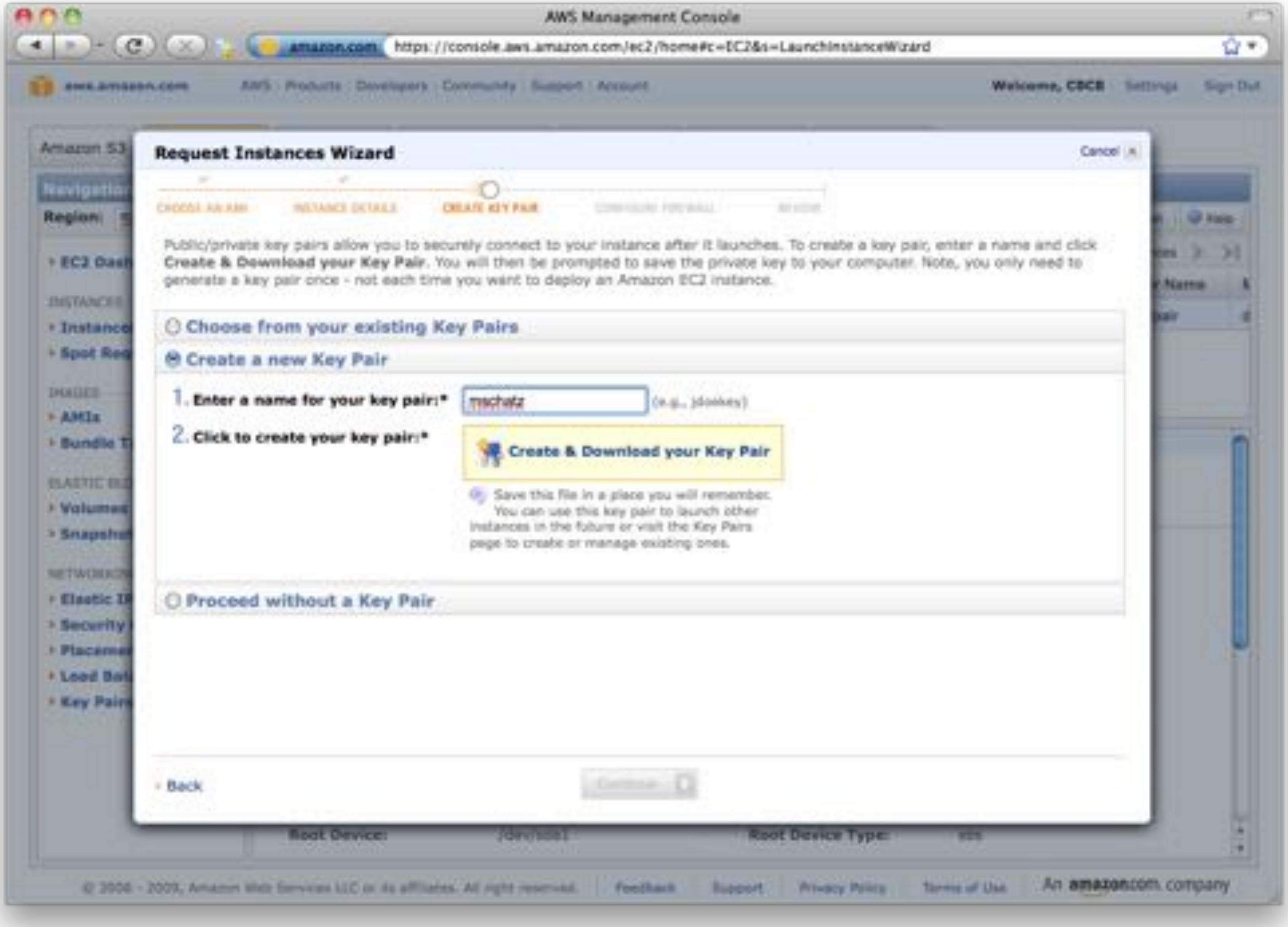
○ Request Spot Inst

○ Launch Instances

Block Devices: /dev/sda1=vol-7d8a8415:attached:2010-10-20T02:41:57.000Z:true

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

3. Pick your Credentials



Request Instances Wizard

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

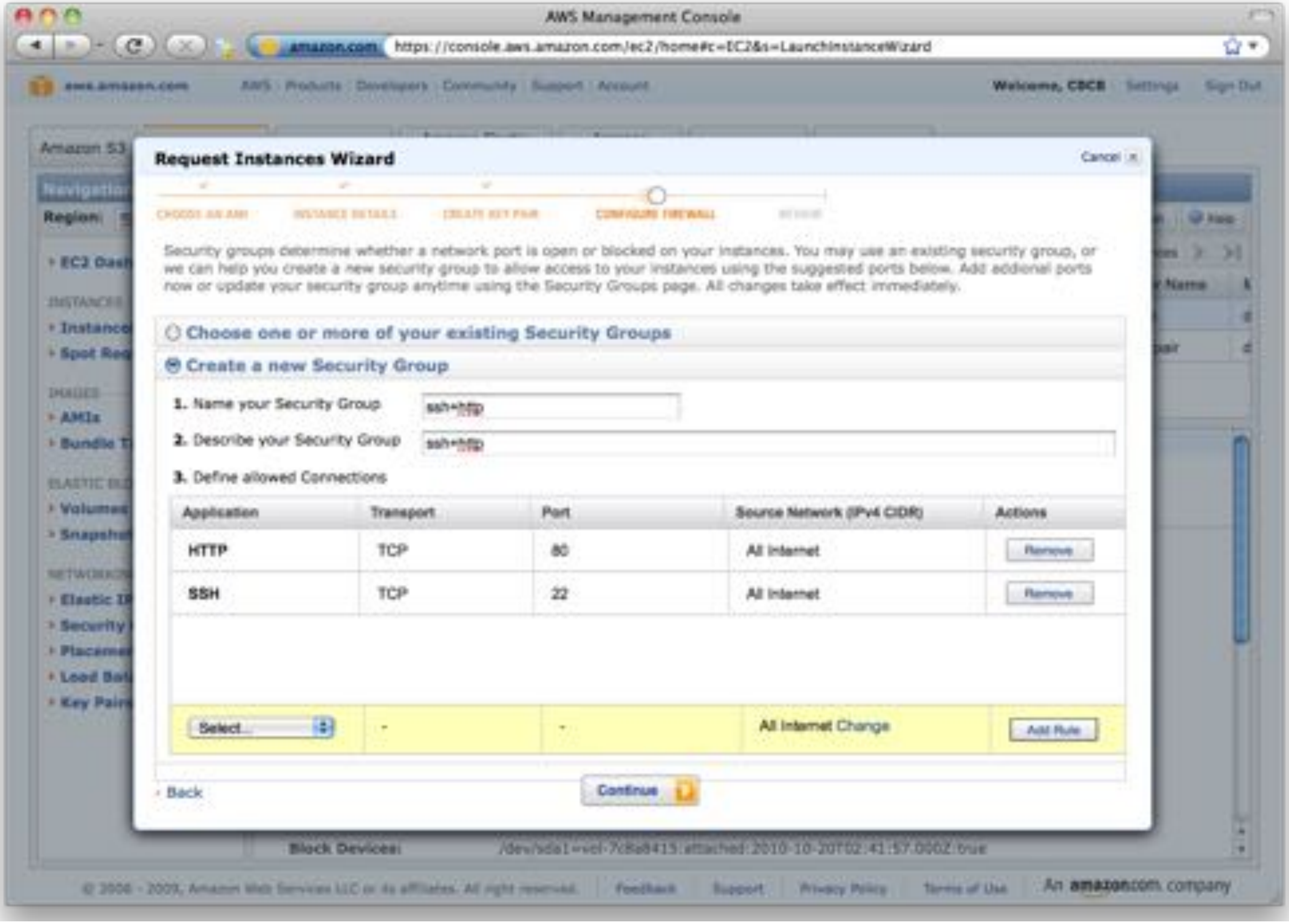
● Create a new Key Pair

1. Enter a name for your key pair:* mschatz (e.g., jdoekey)
2. Click to create your key pair:* Create & Download your 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.

○ Proceed without a Key Pair

4. Configure your Firewall



https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard

Request Instances Wizard

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 addional ports now or update your security group anytime using the Security Groups page. All changes take effect immediately.

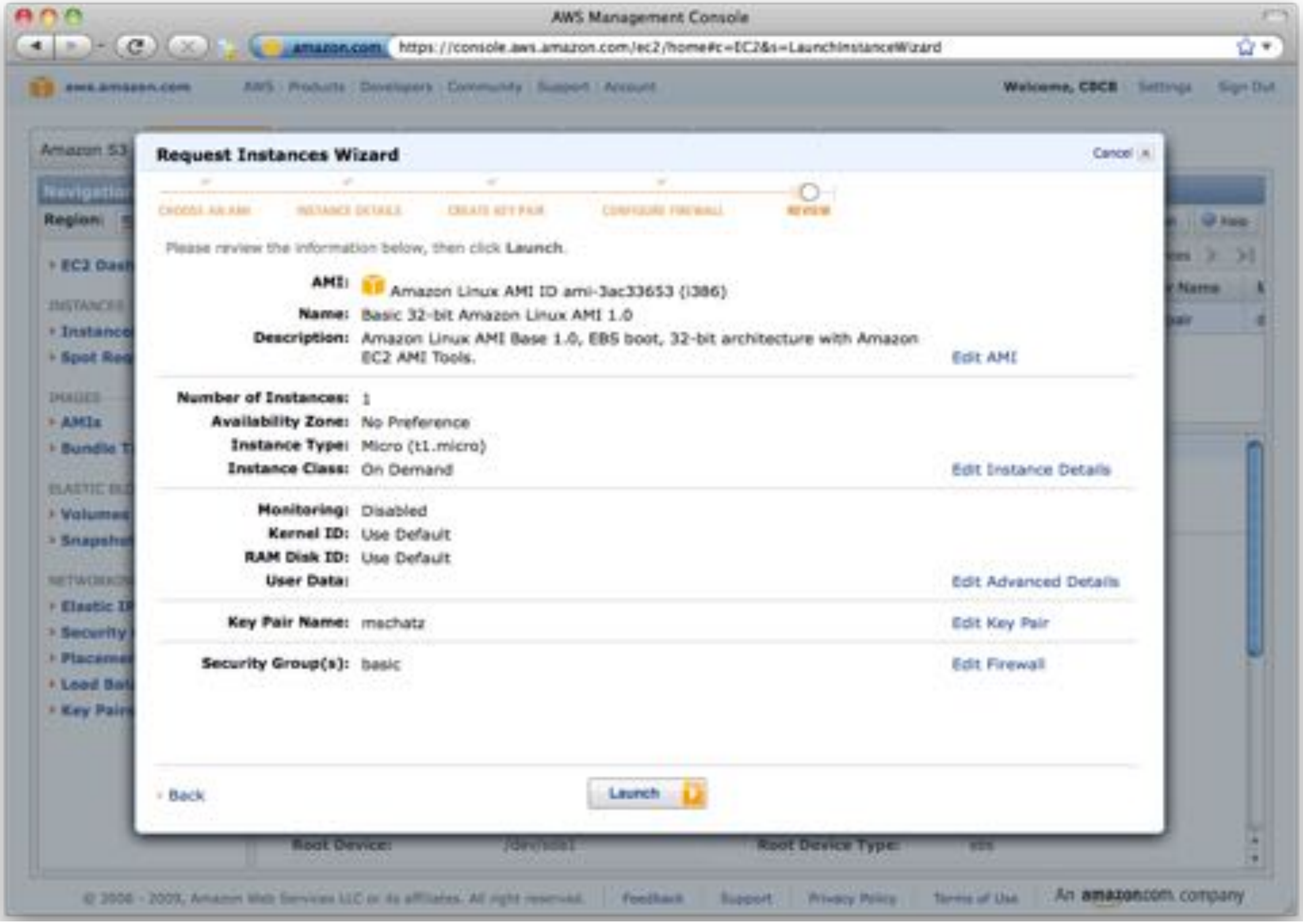
○ Choose one or more of your existing Security Groups

● Create a new Security Group

1. Name your Security Group: ssh+http
2. Describe your Security Group: ssh+http
3. Define allowed Connections

Application	Transport	Port	Source Network (IPv4 CIDR)	Actions
HTTP	TCP	80	All Internet	Remove
SSH	TCP	22	All Internet	Remove
Select...	-	-	All Internet Change	Add Rule

5. Launch!



Please review the information below, then click **Launch**.

AMI: Amazon Linux AMI ID ami-3ac33653 (i386)
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.

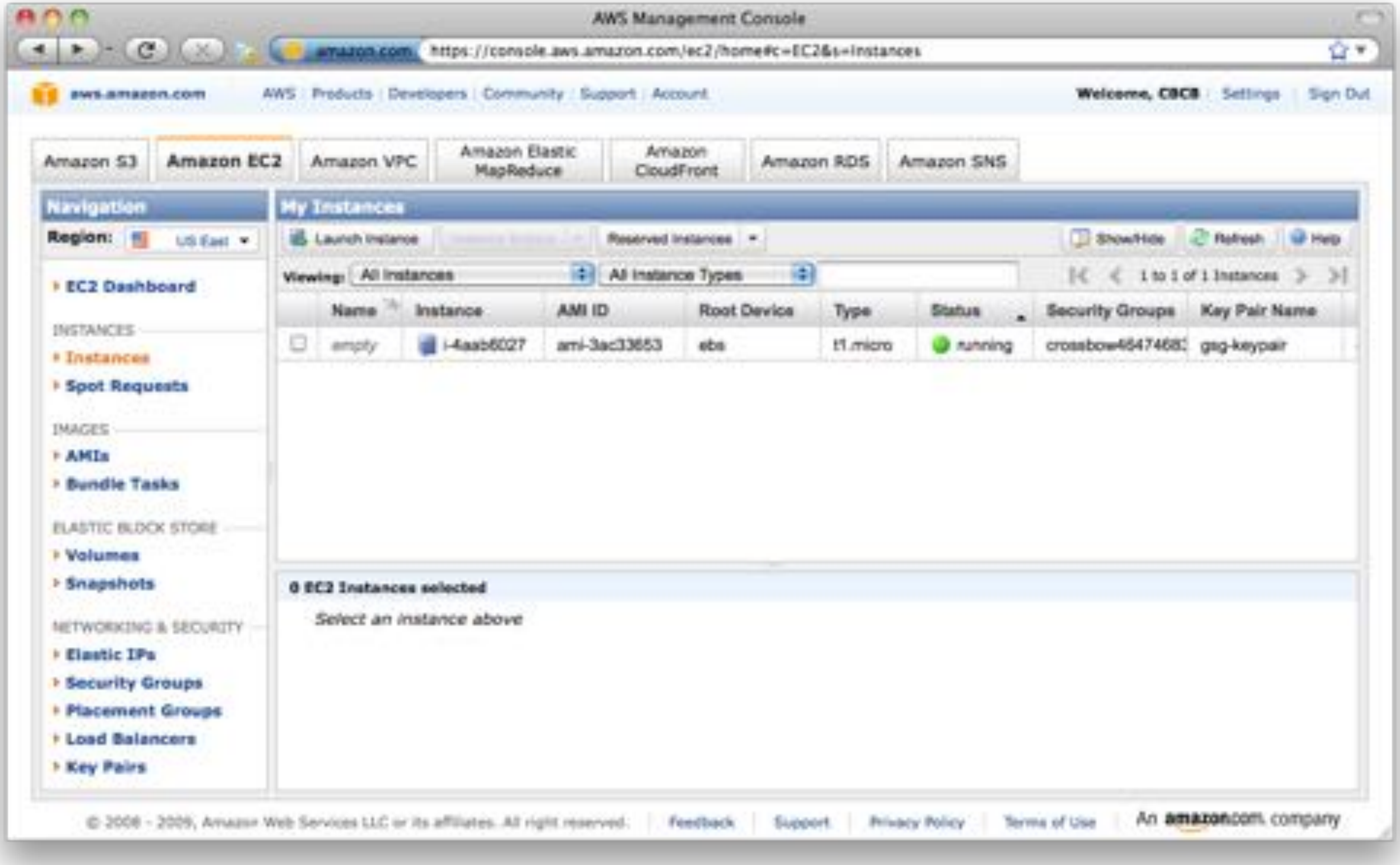
Number of Instances: 1
Availability Zone: No Preference
Instance Type: Micro (t1.micro)
Instance Class: On Demand

Monitoring: Disabled
Kernel ID: Use Default
RAM Disk ID: Use Default
User Data:

Key Pair Name: mschatz

Security Group(s): basic

Monitoring your Server



https://console.aws.amazon.com/ec2/home#c=EC2&s=Instances

Region: US East

Viewing: All Instances | All Instance Types

1 to 1 of 1 Instances

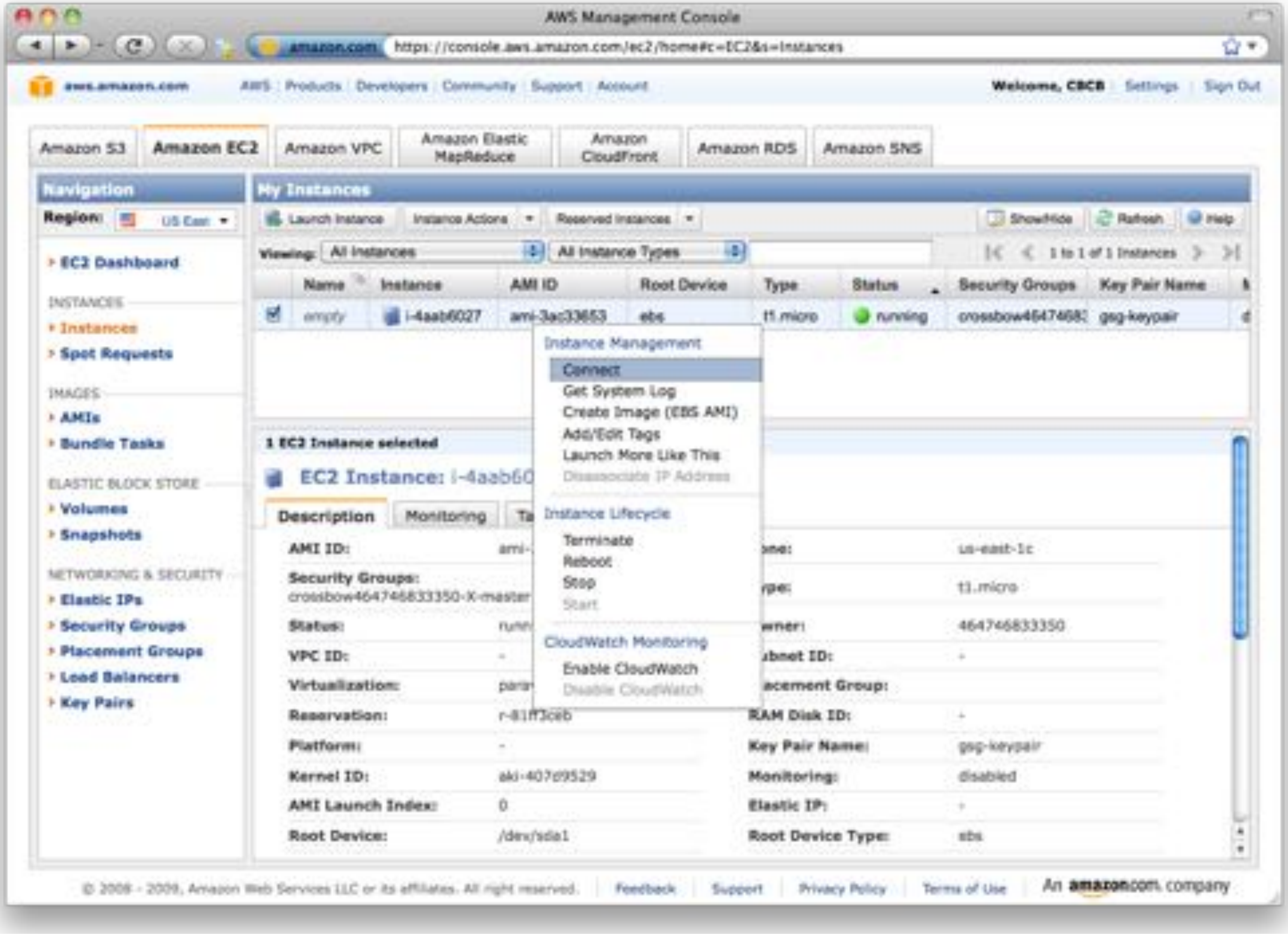
	Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
□	empty	i-4aab6027	ami-3ac33653	ebs	t1.micro	● running	crossbow4647468:	gsg-keypair

0 EC2 Instances selected

Select an instance above

© 2008 - 2009, Amazon Web Services LLC or its affiliates. All right reserved.

Connecting (1)



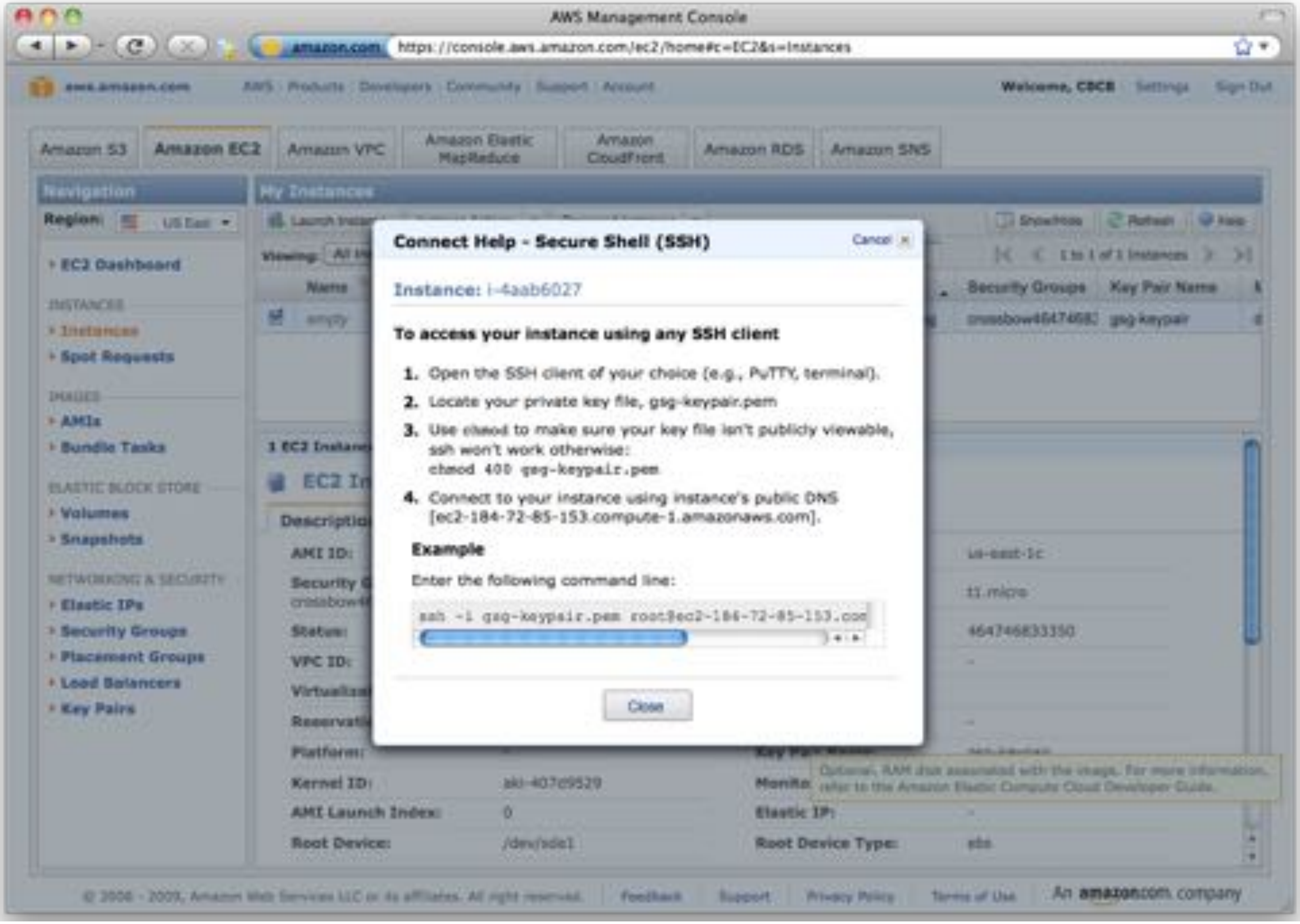
	Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
☑	empty	i-4aab6027	ami-3ac33653	ebs	t1.micro	running	crossbow46474683	gsg-keypair

1 EC2 Instance selected

EC2 Instance: i-4aab60

AMI ID:	ami-	Zone:	us-east-1c
Security Groups: crossbow464746833350-X-master		Type:	t1.micro
Status:	runn	Owner:	464746833350
VPC ID:	-	Subnet ID:	-
Virtualization:	para	Placement Group:	
Reservation:	r-81ff3ceb	RAM Disk ID:	-
Platform:	-	Key Pair Name:	gsg-keypair
Kernel ID:	aki-407d9529	Monitoring:	disabled
AMI Launch Index:	0	Elastic IP:	-
Root Device:	/dev/sda1	Root Device Type:	ebs

Connecting (2)



Connect Help - Secure Shell (SSH)

Cancel

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
```

Close

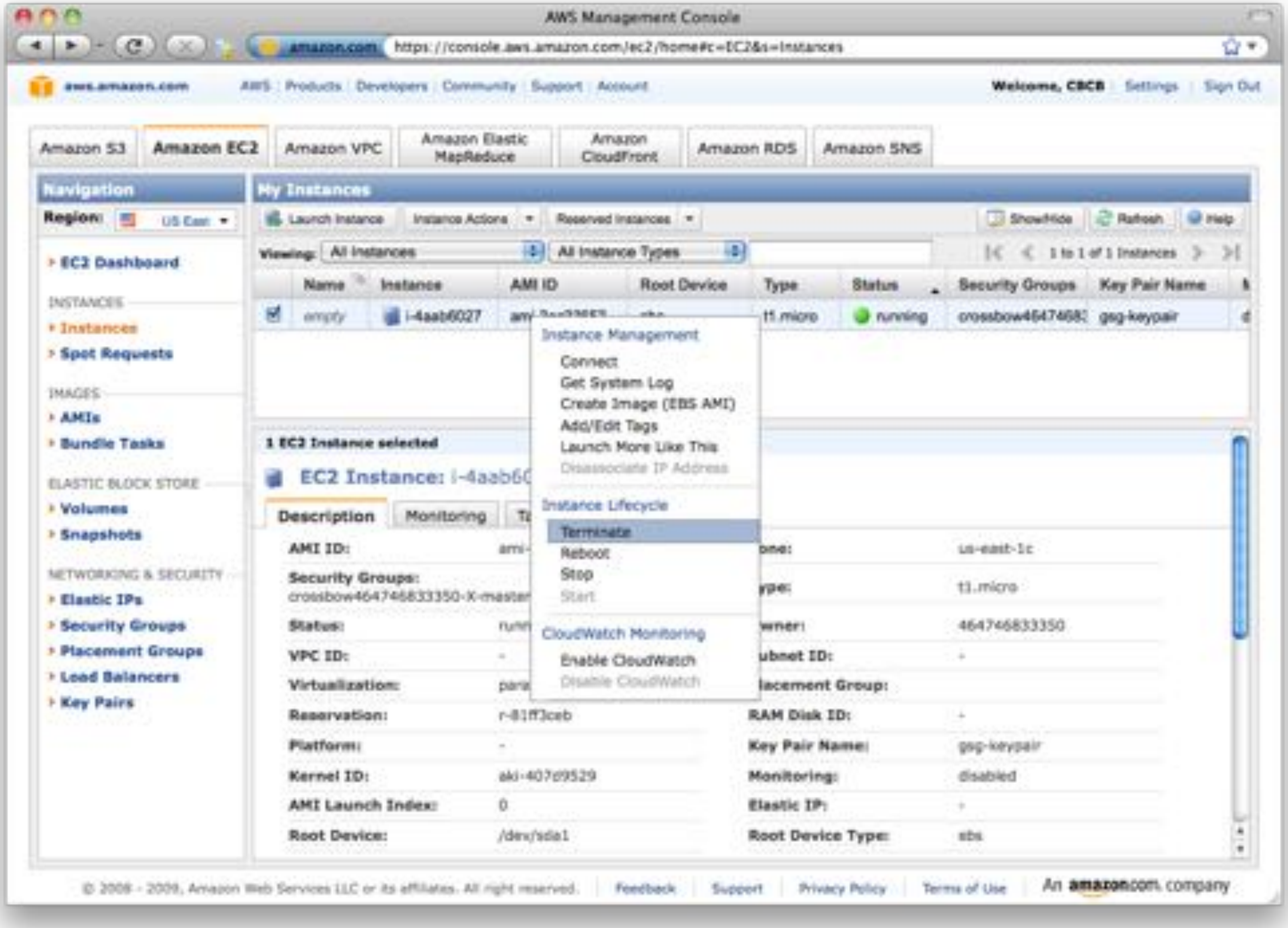
Calling SNPs in the Cloud ☺

```
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



	Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name
☑	empty	i-4aab6027	ami	ebs	t1.micro	● running	crossbow46474683	gsg-keypair

1 EC2 Instance selected

EC2 Instance: i-4aab6

AMI ID:	ami-	one:	us-east-1c
Security Groups: crossbow464746833350-X-master		ype:	t1.micro
Status:	runn	wner:	464746833350
VPC ID:	-	ubnet ID:	-
Virtualization:	para	lacement Group:	
Reservation:	r-81ff3ceb	RAM Disk ID:	-
Platform:	-	Key Pair Name:	gsg-keypair
Kernel ID:	aki-407d9529	Monitoring:	disabled
AMI Launch Index:	0	Elastic IP:	-
Root Device:	/dev/sda1	Root Device Type:	ebs

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