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 #btg I I

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

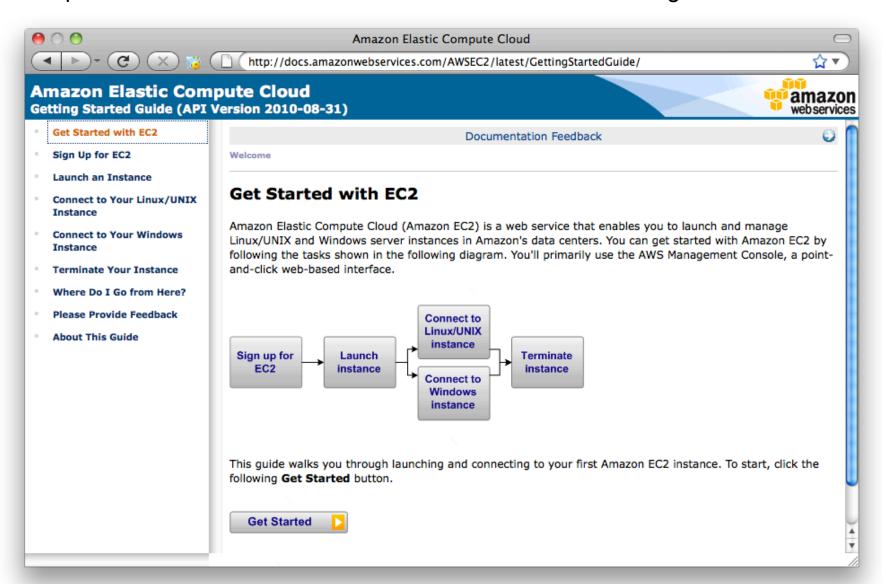
Amazon S3

- 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 I4¢ / 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

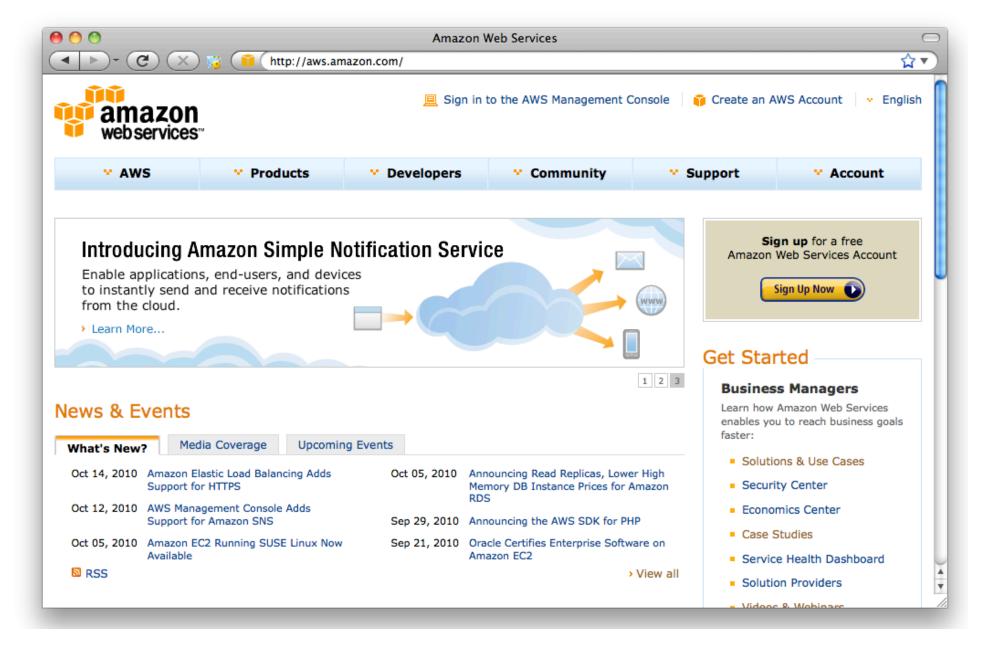


Getting Started

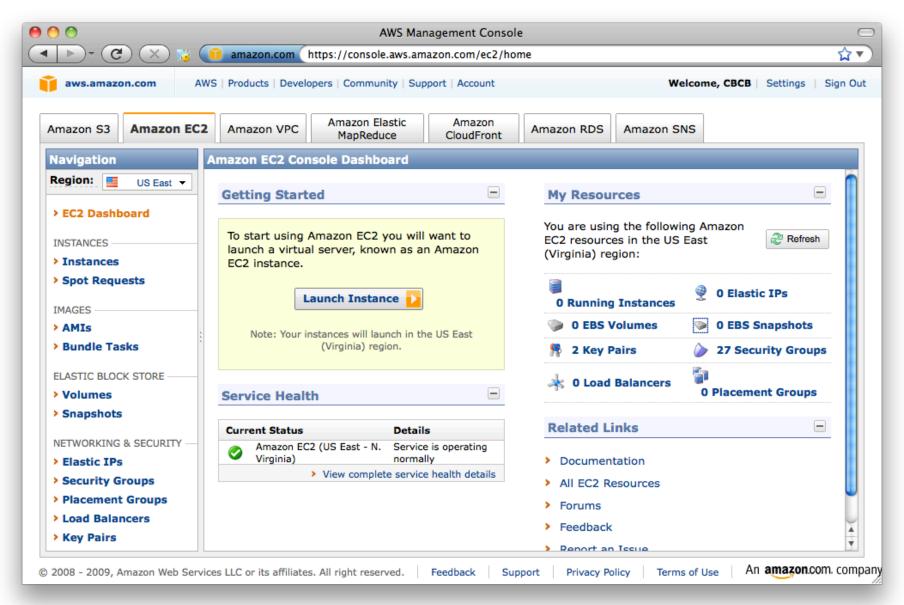
http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/



Signing Up



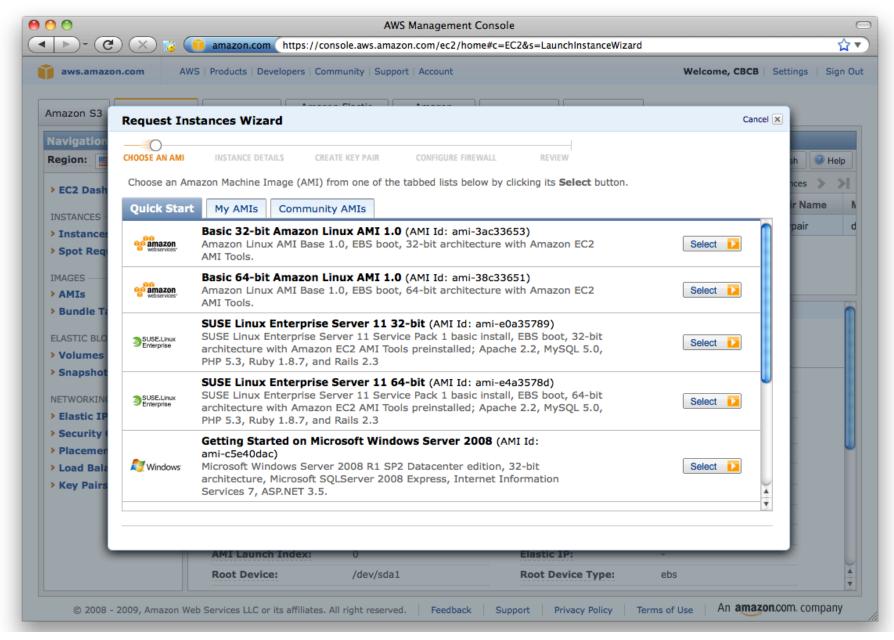
AWS Management Console



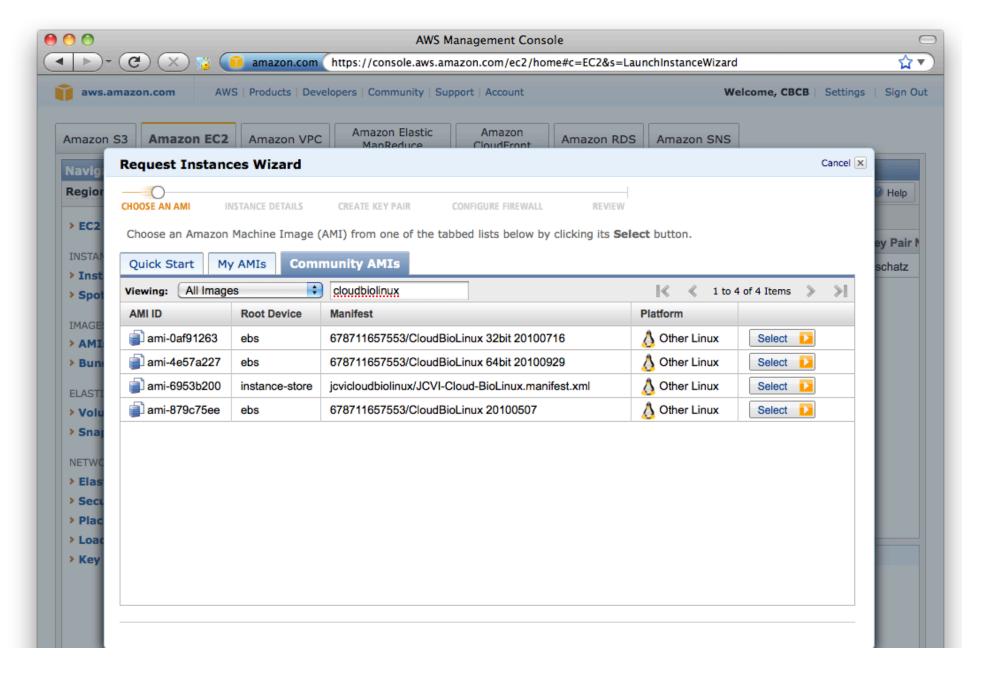
Running your First Cloud Analysis

- I. 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!

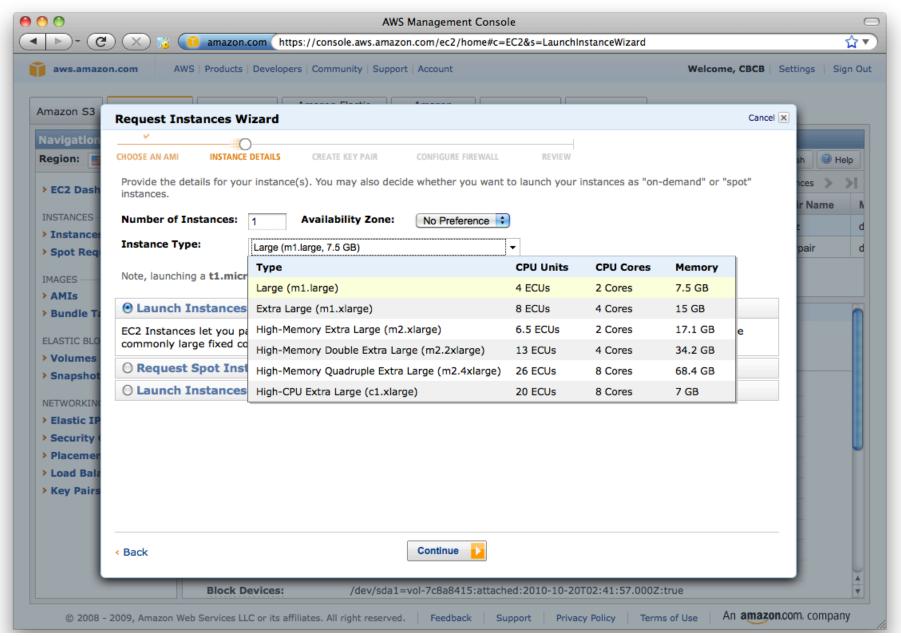
I. Pick your AMIs



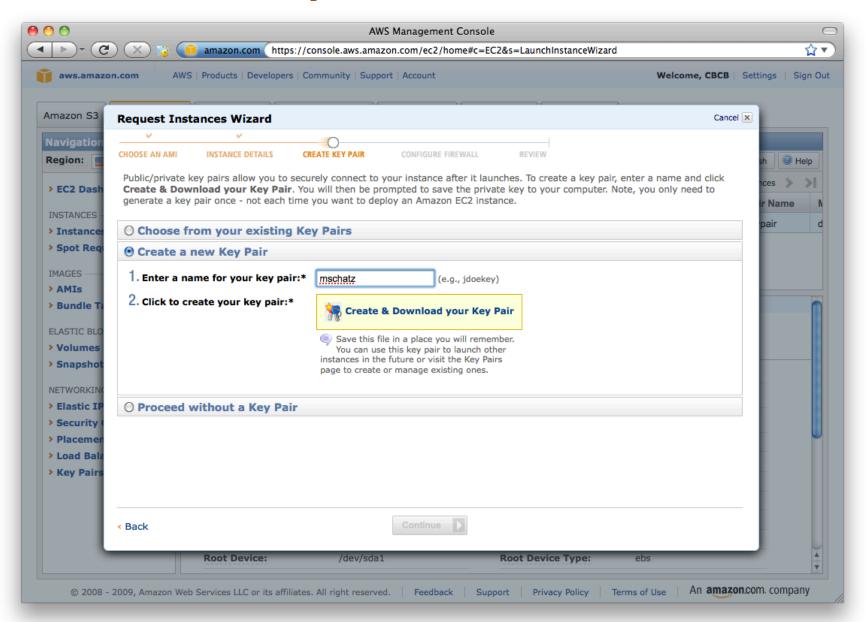
CloudBioLinux



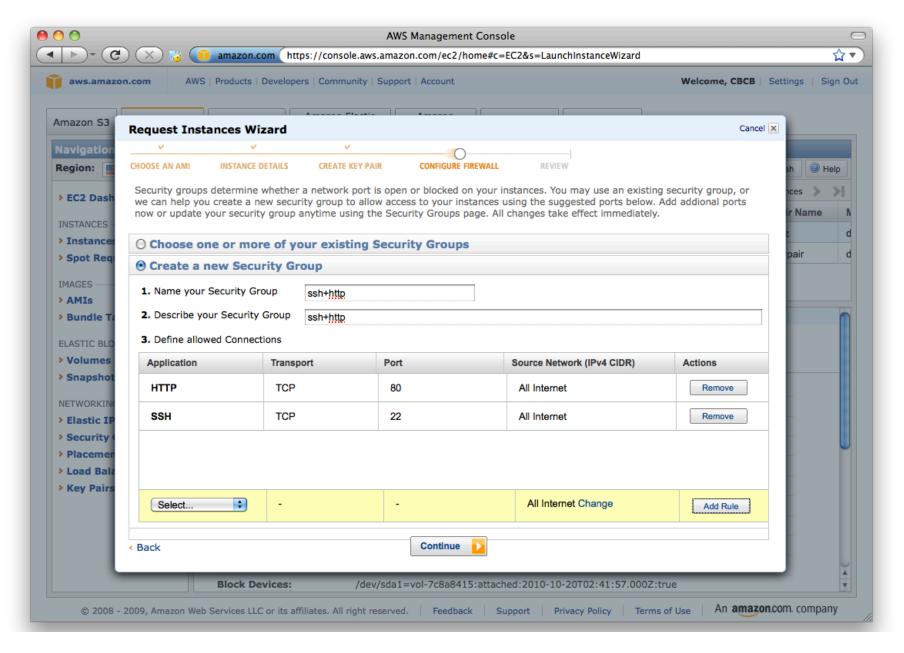
2. Pick your Instance Type



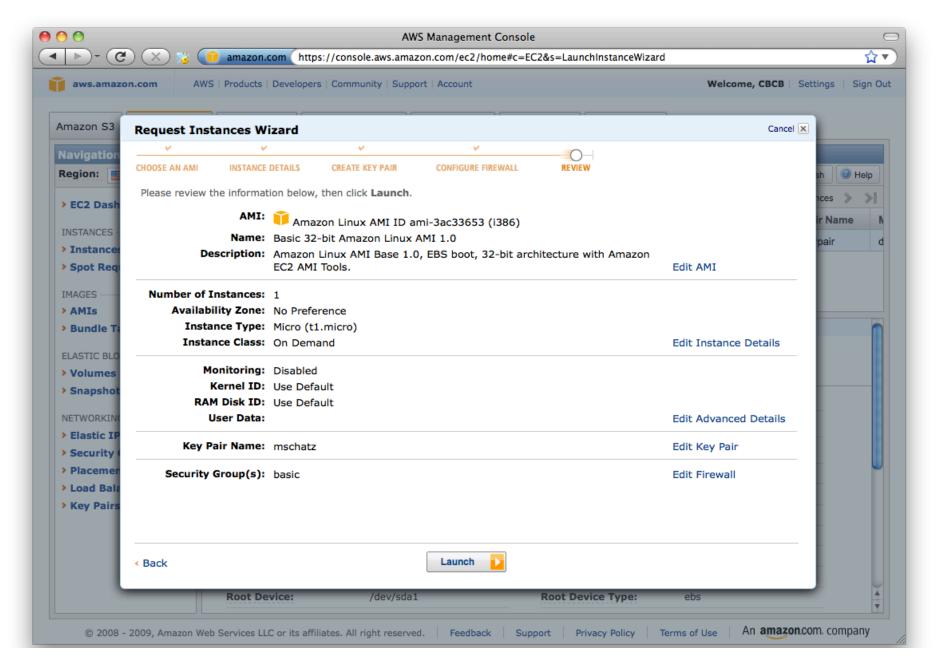
3. Pick your Credentials



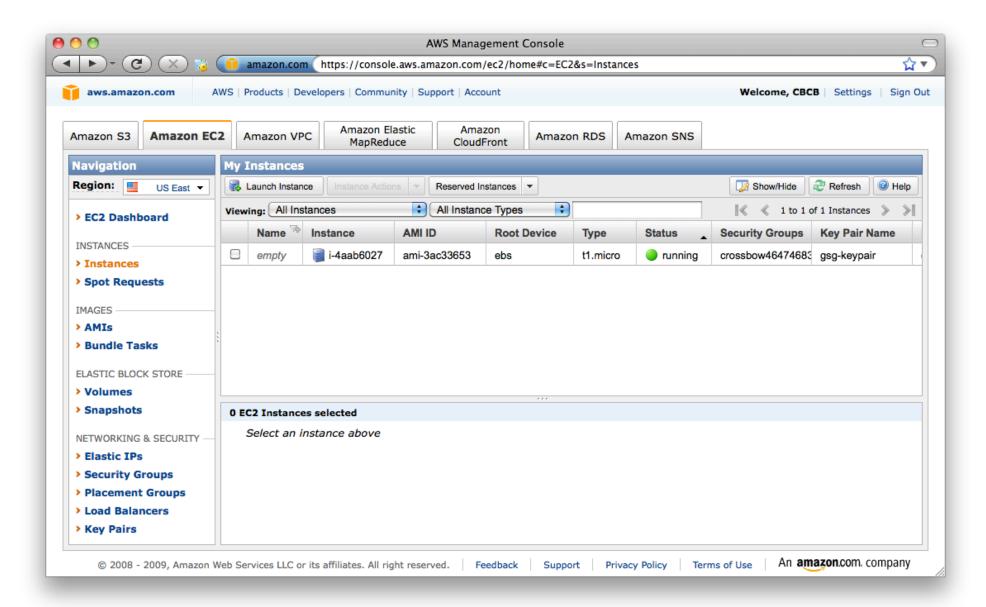
4. Configure your Firewall



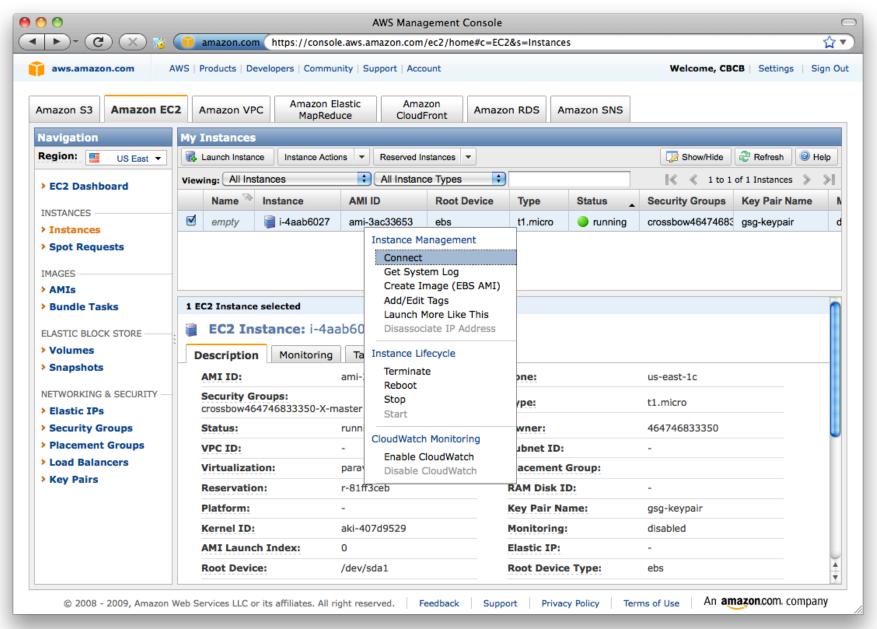
5. Launch!



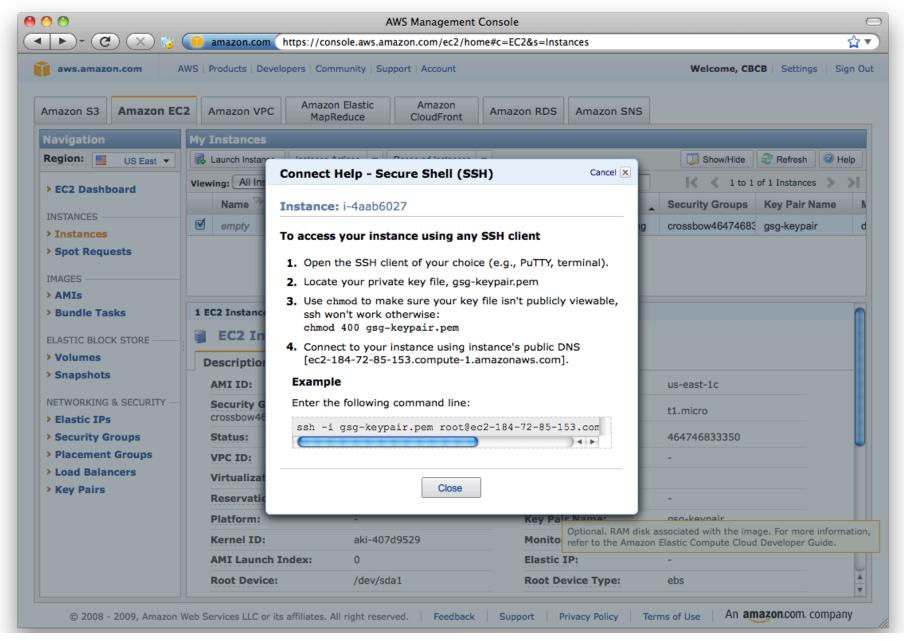
Monitoring your Server



Connecting (I)



Connecting (2)



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>
1s
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

```
# cp data/.s3cfg .

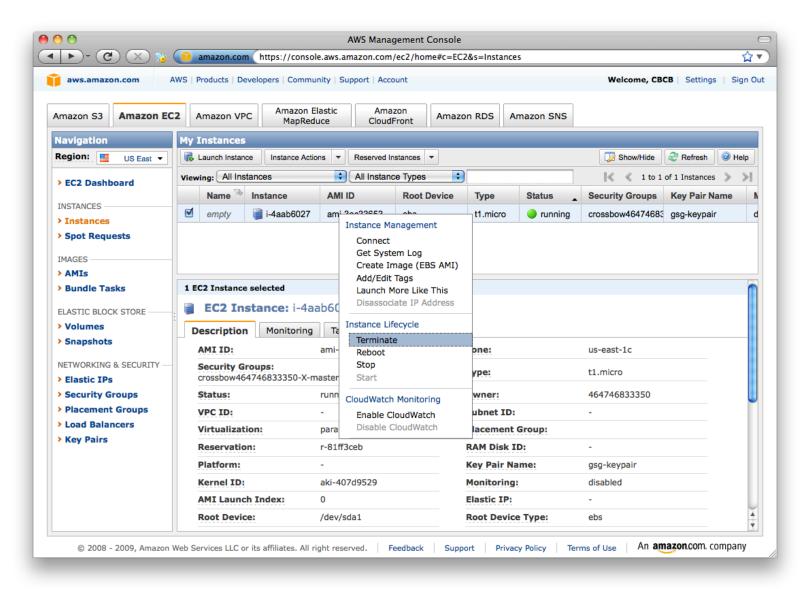
s3cmd ls s3://1000genomes

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
 - I week CPU = I 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