Alm Lab AWS Workshop

March 4th, 2016

What is AWS?

- Amazon's cloud computing platform.
- Cloud refers to:
  - Delocalized data
  - Pay-per-use service model
  - Extra layer of tools for high performance computing (more advanced…)
- Bare bones: just a bunch of datacenters containing CPUs+RAM (EC2) and hard drives (EBS and S3) connected by ethernet cables. The added value is in the software, administrative tools, database management systems and general flexibility.

Why AWS for the Alm lab?

- Originally, was intended as computational infrastructure for Microbiome Center clinical activities

- Integrate all lab-developed and commonly used public software/tools into core, lab-maintained pipelines, e.g.
  - StrainFinder
  - 16S processing (best practices, built on scripts from Lawrence, Chris, Scott, etc.)
  - Metagenomics processing
  - Metabolomics processing

- Vision: non-proprietary multi-omics data analysis computing platform

Why AWS for the Alm lab?

- Coyote sucks:
  - No root access, need [greg] to install anything
  - Prohibitively long queue around grant submission deadlines
  - Constantly breaking
  - Insecure for housing patient data. AWS is HIPAA compliant.
- In anticipation of increased and more diverse bioinformatics needs:
  - Increasingly large datasets (storage, and computing)
  - Ability to trade time for money, based on urgency
  - Allows centralized and scalable management of all computing resources (web services, databases, computing clusters)

Why AWS for the Alm lab?

- If you have very specific needs, you can even have your own **machine image**, which is pre-loaded with all your favorite tools.

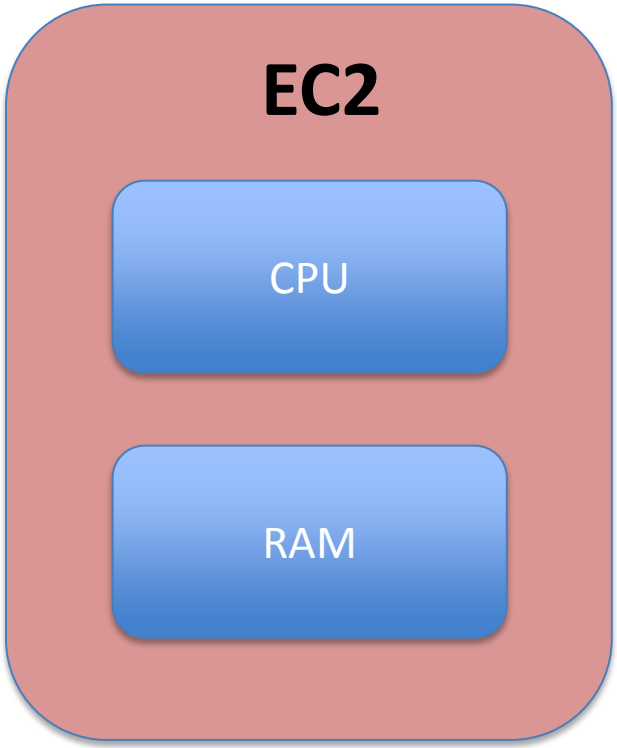
Do I have to use AWS? Why can't I just use a classical computing cluster?

- You of course do not **have** to do anything. Coyote will still exist, as will radiodurans, and frankly, your laptop is often enough for most purposes these days.

- Decision should be based on individual needs:
  - Are you a new user?
  - Are you familiar with the tools you need to do your computing, or would you like to use previously conceived pipelines?
  - Do you need a lot of storage space or extra RAM?
  - Do you potentially have some hardcore computing tasks lined up (e.g. large metagenomic dataset analyses)?
  - Do you want to learn about cloud computing/AWS?
  - E.g. Netflix

AWS is a collection of services
Elastic Cloud Compute (EC2)

- Service for computing *instances* (a.k.a. nodes in a cluster context)

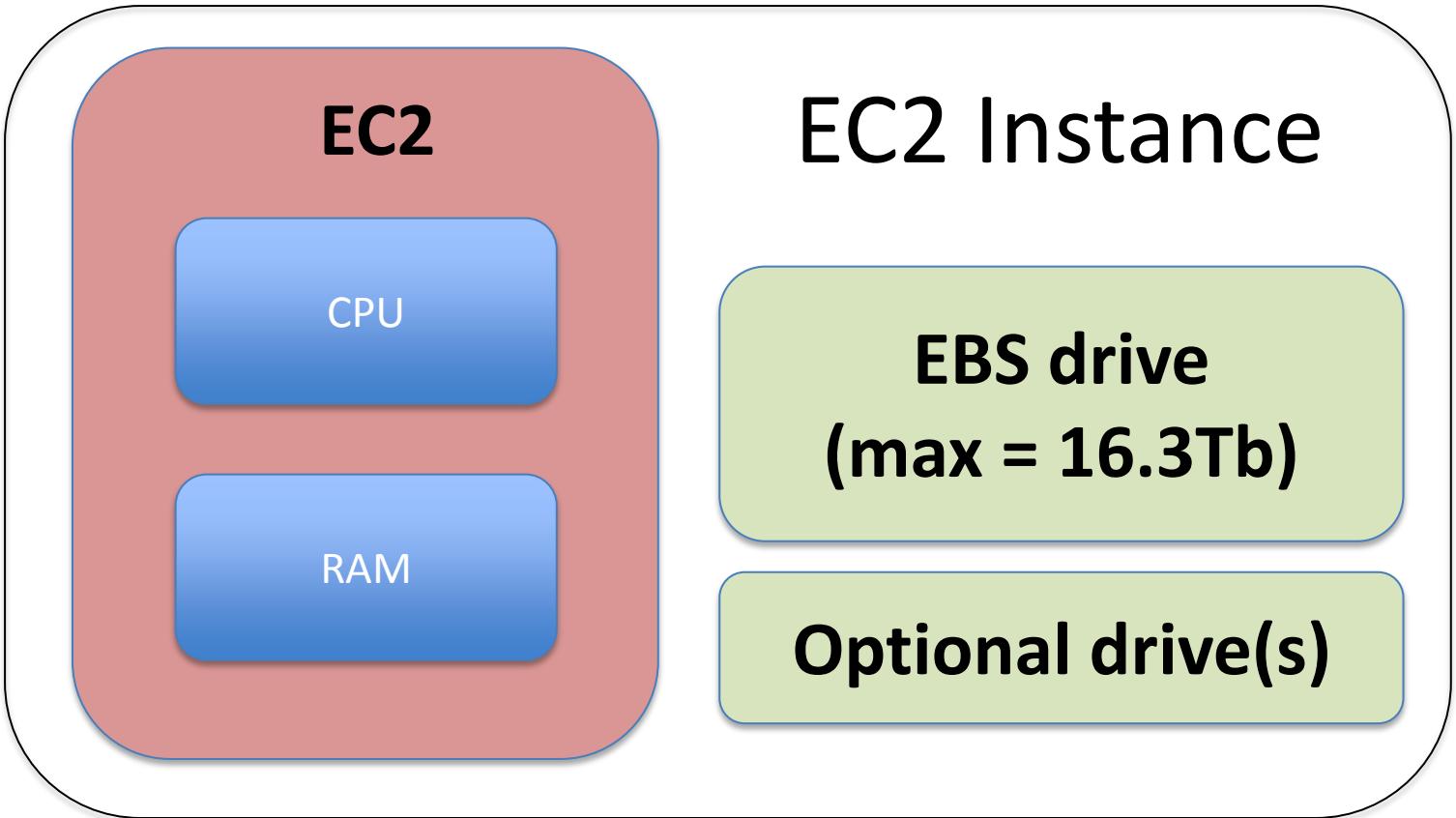


General Purpose - Current Generation

t2.nano	1	Variable	0.5	EBS Only	$0.0065 per Hour
t2.micro	1	Variable	1	EBS Only	$0.013 per Hour
t2.small	1	Variable	2	EBS Only	$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	$0.052 per Hour
t2.large	2	Variable	8	EBS Only	$0.104 per Hour
m4.large	2	6.5	8	EBS Only	$0.12 per Hour
m4.xlarge	4	13	16	EBS Only	$0.239 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.479 per Hour
m4.4xlarge	16	53.5	64	EBS Only	$0.958 per Hour
m4.10xlarge	40	124.5	160	EBS Only	$2.394 per Hour
m3.medium	1	3	3.75	1 x 4 SSD	$0.067 per Hour
m3.large	2	6.5	7.5	1 x 32 SSD	$0.133 per Hour
m3.xlarge	4	13	15	2 x 40 SSD	$0.266 per Hour
m3.2xlarge	8	26	30	2 x 80 SSD	$0.532 per Hour

AWS is a collection of services
Elastic Cloud Compute (EC2)

- Need to attach a hard drive (EBS)
- EBS storage is expensive



Instances can be:
- Running
- Stopped
- Terminated

AWS is a collection of services
Simple Storage Service (S3)

- Distributed set of hard-drives that essentially act like an infinite hard-drive, pay per use

- Comes in 2 flavors:
  - standard (always online)
  - Glacier (on physical disks; retrieval requests take 3-5h and data is available for 24h)

AWS disk space pricing

- EBS: $100/Tb/month
- S3 (standard): $30/Tb/month
- S3 (Glacier): $7/Tb/month

What storage does the Alm lab currently have set up?

- Alm lab S3 bucket:
  - Contains folders for each user, can put as much as you want there.
  - Permissions currently very open. Discuss at the end.
- 1 Tb scratch drive, currently mounted on ‘almlab 1’
  - Can theoretically have multiple such drives

How do I ‘connect to AWS’?

- You never really connect to AWS. You either **log on to specific instances** (which are computers in the cloud), or you **log on to the management console**
- Specific instances (via SSH using the SSH key):

ssh –i almlab_key.pem ubuntu@nodeIP

Alm lab general node IP: 52.4.58.59
Alm lab QIIME node IP: 52.1.17.174

- Management console: Username and password + link provided when your user account is created.

Basic troubleshooting

ssh -i almlab_key.pem ubuntu@52.7.213.89

@@@
@@@@@@@@@@@@
@   WARNING: UNPROTECTED PRIVATE KEY FILE!   @
@@@
@@@@@@@@@@@@
Permissions 0755 for '/Users/tornam/Dropbox/Personal analytics/AWS/admin/keys/PA_key.pem' are too open.
It is required that your private key files are NOT accessible by others.
This private key will be ignored.
bad permissions: ignore key: /Users/tornam/Dropbox/Personal analytics/AWS/admin/keys/PA_key.pem
Permission denied (publickey).

- Permissions of key need to be set to '400' on your computer when you receive the key:

  sudo chmod 400 almlab_key.pem

Basic troubleshooting

- Problem: I use Putty to connect, and it won't accept the almlab_key.pem!!

- Solution: Putty needs a .ppk key. You can make it yourself with PuttyGen from the almlab_key.pem

Basic troubleshooting

- Check your known hosts, the computer's hardware ID may have changed but still has the same IP

Transferring data to/from the cloud

- Need the SSH key on your computer (can't be done from Coyote, sorry – need to transfer via your computer or for larger transfers, we can arrange something)

scp –i almlab_key.pem /path/file.txt ubuntu@nodeIP

Alm lab general node IP: 52.4.58.59
Alm lab QIIME node IP: 52.1.17.174

Transferring data within the cloud

- From S3 to EC2: use the AWS S3 command line interface
  - For files:

```
aws s3 cp s3://almlab.bucket/thomas/somefolder/    somefile.txt
/path/somefile.txt
```

  - For folders:

```
aws s3 sync s3://almlab.bucket/thomas/somefolder/
      /path/newfolder/
```

Security

- With an AWS access key with root permissions, you can theoretically launch as many nodes as you want. This is managed through the Identity Access Management (IAM) system on AWS.
- Some clinical datasets may contain sensitive info (HIPAA)
- GitHub example
- almlab_key.pem or access credentials **should not be shared without permission from Eric or Thomas**
  - Please take this seriously, as consequences are severe; there is a reason [*insert_generic_sys_admin_here*] does not usually grant root access to users

Discussion: Permissions, administration, and sustainability

- S3 bucket permissions
- Key distribution protocol
- Future administration – will it break when Thomas leaves?