Model Clinic: Using GEOS-Chem on Amazon Web Service (AWS) cloud

Jiawei Zhuang
5/9/2019
The 9th International GEOS-Chem Meeting (IGC9)

Get familiar with cloud computing concepts – read our overview paper

Zhuang, J., et al. *Enabling immediate access to Earth science models through cloud computing: application to the GEOS-Chem model*, under review on Bull. Amer. Met. Soc., 2019. http://acmg.seas.harvard.edu/publications/2019/zhuang2019.pdf

The paper covers:

- The strong motivation of moving to the cloud
- How we make the cloud easily accessible to scientists
- High-level overview of research workflow on AWS cloud
- Performance and cost analysis of AWS versus local cluster

Actually start using cloud computing – follow the online step-by-step tutorial at http://cloud.geos-chem.org

The tutorial covers the complete research workflow, including

- Starting the first GEOS-Chem simulation in a few minutes
- Analyzing output data with Python and Jupyter notebooks
- Setting up custom model versions and configurations
- Using software containers such as Docker and Singularity
- Configuring your custom system/compilers if you wish

Major AWS cloud concepts to remember

Elastic Compute Cloud (EC2)

- The most essential computing service on AWS
- Similar to your local Linux server that you can “ssh” to.
- Can have custom hardware capacity (CPU, memory, disk, network), as well as custom software environment defined by an Amazon Machine Image (AMI)

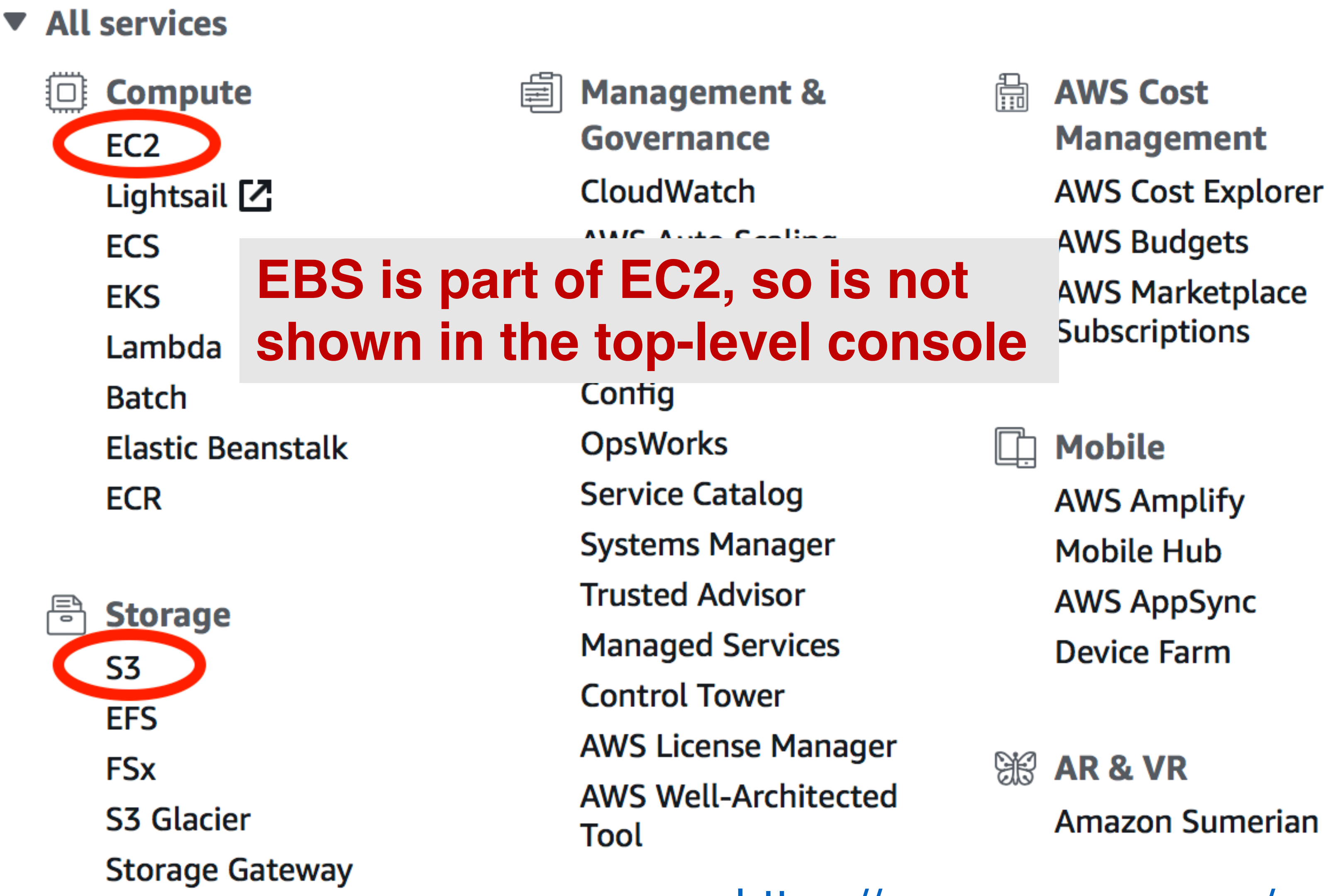
Elastic Block Storage (EBS)

- The disk storage directly used by EC2
- Similar to the hard drive for your computer
- Mostly used for temporary storage during computation

Simple Storage Service (S3)

- The most essential storage service on AWS
- The closest analogy is Dropbox or Google Drive
- Mostly used for long-term persistent storage
- Cost 50~90% less than EBS

You will see hundreds of services in the AWS web console. Don't be scared. Only knowing EC2 and S3 is enough to get almost everything done.



▼ All services

Compute

EC2

Lightsail

ECS

EKS

Lambda

Batch

Elastic Beanstalk

ECR

Storage

S3

EFS

FSx

S3 Glacier

Storage Gateway

Management & Governance

CloudWatch

AWS Auto Scaling

Config

OpsWorks

Service Catalog

Systems Manager

Trusted Advisor

Managed Services

Control Tower

AWS License Manager

AWS Well-Architected Tool

AWS Cost Management

AWS Cost Explorer

AWS Budgets

AWS Marketplace Subscriptions

Mobile

AWS Amplify

Mobile Hub

AWS AppSync

Device Farm

AR & VR

Amazon Sumerian

EBS is part of EC2, so is not shown in the top-level console

https://aws.amazon.com/console/

Step 0: Sign up for AWS accounts

- Sign up at http://aws.amazon.com/. The initial account verification can take ~30 minutes.

- The only requirement is a credit card. There are also “educational accounts” without credit card requirements[1] .

- For students, remember to apply for the $100 per year educational credit[1] .

[1] All education-related contents are at https://aws.amazon.com/education/awseducate

Step 1: Launch virtual servers (EC2 instances)

1. Choose the proper **software environment**, by selecting an Amazon Machine Image (AMI) that contains pre-configured GEOS-Chem and sample input data.

2. Choose the proper **hardware capacity**, by selecting the EC2 instance type (https://aws.amazon.com/ec2/instance-types).

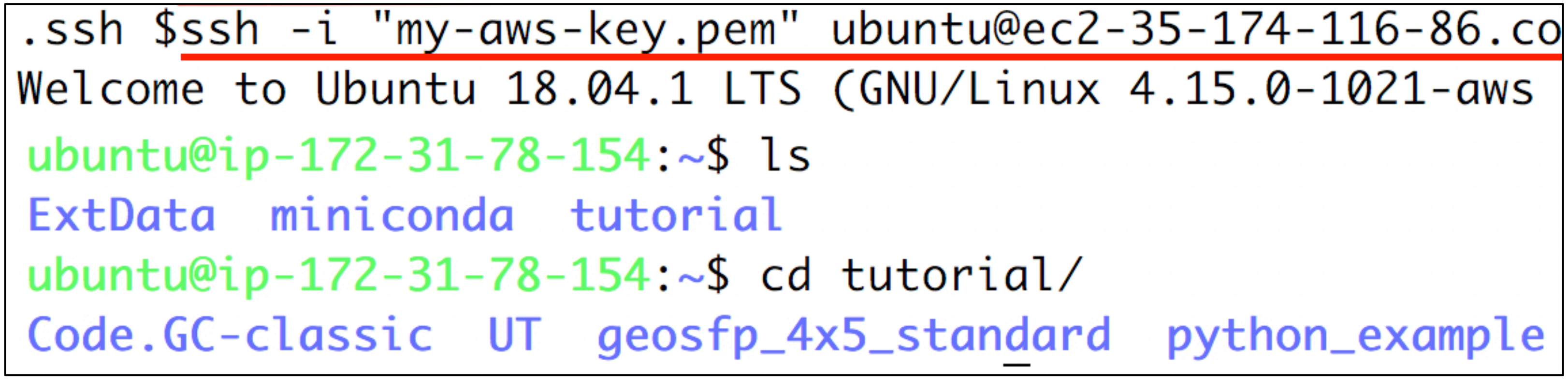
3. Launch the EC2 instance, and your virtual server will be running in a few seconds.

4. Connect to your server via the Secure Shell (SSH) in your computer terminal, just like for normal Linux servers.

See concrete instructions and screenshots at
https://cloud-gc.readthedocs.io/en/latest/chapter02_beginner-tutorial/quick-start.html

Step 2a: Start the first demo simulation

- After logging-in to the instance, you will immediately see a pre-configured GEOS-Chem run directory, for demo purpose.



```
.ssh $ssh -i "my-aws-key.pem" ubuntu@ec2-35-174-116-86.co
Welcome to Ubuntu 18.04.1 LTS (GNU/Linux 4.15.0-1021-aws
 ubuntu@ip-172-31-78-154:~$ ls
 ExtData  miniconda  tutorial
 ubuntu@ip-172-31-78-154:~$ cd tutorial/
 Code.GC-classic  UT  geosfp_4x5_standard  python_example
```

- Simply running the “geos” executable will correctly start a short simulation which generates sample output data.
- You don’t need to submit jobs (with “qsub” or “sbatch” commands) to the job scheduler. Use “tmux” or GNU Screen to manage long-running jobs[1].

[1] See instructions at https://cloud-gc.readthedocs.io/en/latest/chapter06_appendix/keep-running.html

Step 2b: Analyze output data with Python and Jupyter

- If you are new to Python, spend an hour studying our Python tutorial[1]. It is much easier than IDL/MATLAB.

- Jupyter notebooks always display the user interface, Python code, and output graphics in the web browser, no matter running locally or on the cloud. No more slow "X11 forwarding" when working with remote servers as with IDL.

- We recommend using NetCDF output format [2], and open NC files using Xarray (http://xarray.pydata.org). The old BPCH format can also be read by Xarray[3], if you really want to.

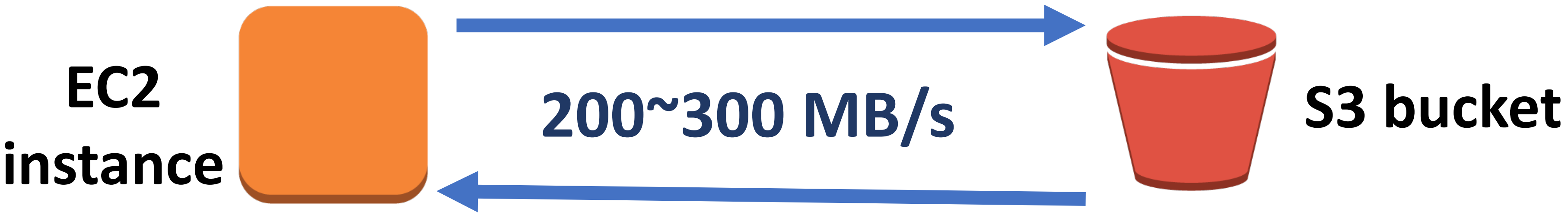
[1] https://github.com/geoschem/GEOSChem-python-tutorial
[2] http://wiki.seas.harvard.edu/geos-chem/index.php/List_of_diagnostics_archived_to_netCDF_format
[3] Using xbpch https://xbpch.readthedocs.io

Step 3: Working with persistent storage in S3

- The major difference between cloud and local servers is how to manage long-term data storage.
- You should upload important data to “S3 buckets” before terminating an EC2 instance, otherwise the data will be deleted as well.
- The complete 30+ TB GEOS-Chem input data also live in S3.

Upload files from EC2 to S3

(via command “aws s3 cp ec2-file-location s3://s3-file-location”)



Download files from S3 to EC2

(via command “aws s3 cp s3://s3-file-location ec2-file-location”)

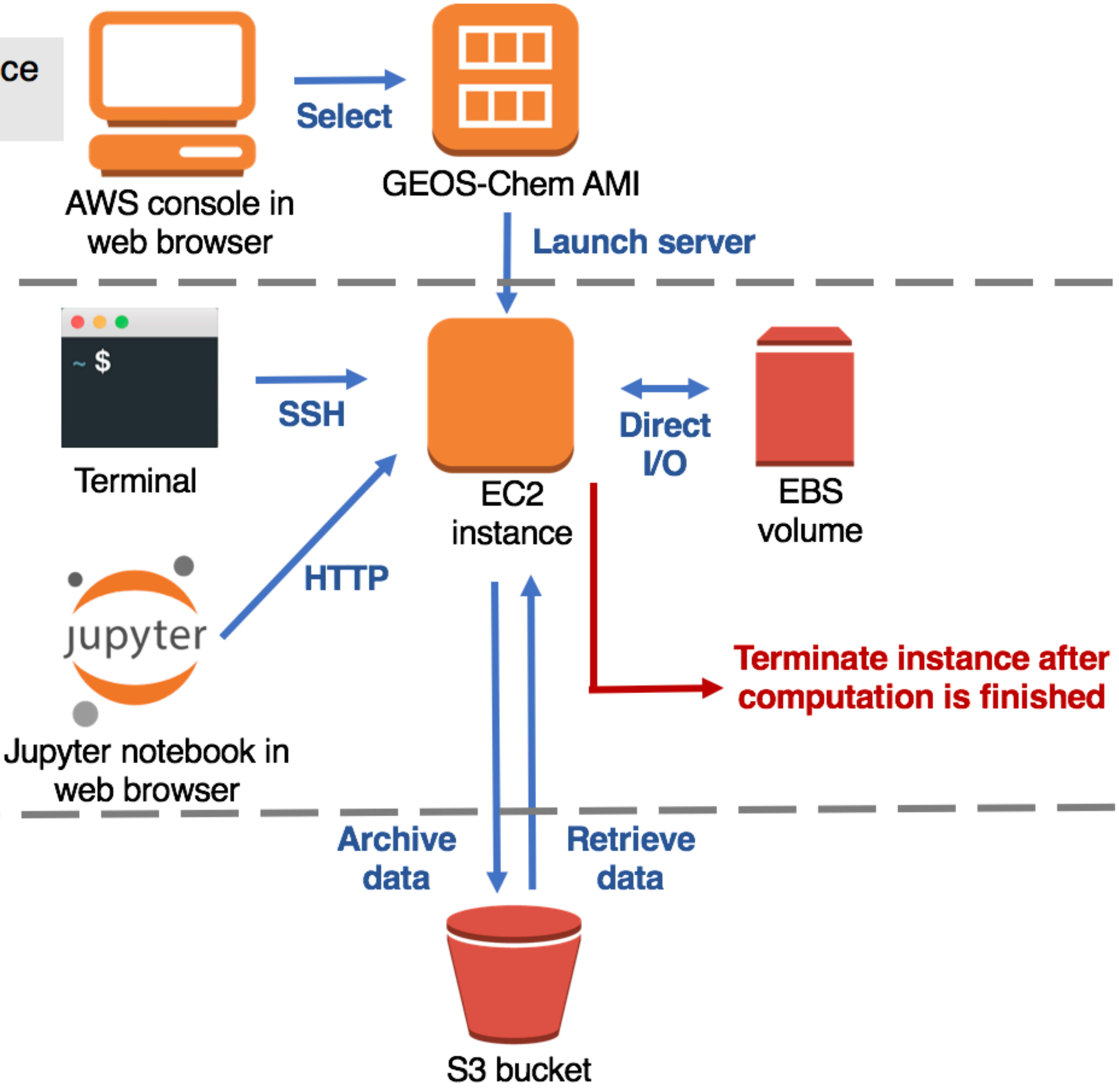
Summary of research workflow on cloud

Step 1. Launch EC2 instance from GEOS-Chem AMI

Step 2a. Use terminal to login to EC2 instance and run simulation

Step 2b. Use Jupyter notebook to analyze output data

Step 3. Use S3 as persistent storage



(Zhuang et al., 2019, BAMS)

Cost and billing models

Service	Cost
EC2	• ~$0.1 / core / hour, depending on hardware • 60%~70% cheaper with “spot instances”
EBS	• $0.1 / GB / month, for solid-state drive (SSD). • Cheaper options with different I/O characteristics are available.
S3	• $0.023 / GB / month. • Cheaper options for infrequent access patterns are available.
Data Egress	• $0.09 per GB of data transferred out of cloud • Research institutes can waive some egress fee

(Zhuang et al., 2019, BAMS)

Cost of an example project

(1-year 2° × 2.5° simulation with standard chemistry, saving out 3-D daily concentration fields for 168 species.)

Service	Resources needed	Total cost (USD)
EC2	330 hours on EC2 “c5.4xlarge” type	$224 with standard price $90 with spot price
EBS	300 GB disk to host input and output data files during simulation	$14 with standard solid-state drive (SSD)
S3 or download	150 GB output data files	$3.50 per month on S3, or $13.50 to download

(Zhuang et al., 2019, BAMS)

Advanced topic: software containers

docker

Singularity

Charliecloud

- Containers can ship the complete software environment (libraries and pre-compiled models) across different cloud platforms and local servers.
- Containers offer a super quick way to install GEOS-Chem on local servers, as long as some container software (e.g. Docker, Singularity, Charliecloud) is installed.
- Multi-node MPI runs with containers are not very mature yet and remain an active research topic.

See instructions at
https://cloud-gc.readthedocs.io/en/latest/chapter03_advanced-tutorial/container.html

Other useful resources

□ Cloud Computing for Science and Engineering, Ian Foster and Dennis B. Gannon, MIT Press, 2017

- Freely available online at https://cloud4scieng.org
- Probably the only cloud textbook written for scientists
- Touches AWS, Azure and Google cloud
- Touches a broad range of scientific applications

□ Researcher's Handbook by AWS

- Sigh up for the Research Cloud Program and download the handbook PDF at https://aws.amazon.com/government-education/research-and-technical-computing/research-cloud-program/

□ Cloud Computing for Research, by University of Washington

- https://itconnect.uw.edu/research/cloud-computing-for-research/
- Gives a high-level overview of cloud computing
- Answers common questions such as research funding for cloud resources.