Enabling immediate access to GEOS-Chem on Amazon Web Service (AWS) cloud

Jiawei Zhuang

5/6/2019

The 9th International GEOS-Chem Meeting (IGC9)

Although GEOS-Chem aims to be user-friendly, new users and existing users can still face difficulties

New users often have a steep learning curve:

- Need to configure a software environment from scratch and fight with mysterious compile errors.
- Need to spend days and weeks downloading large volumes of input meteorology and emission data
- Small research groups often lack computing resources.

Existing users often cannot catch up with the rapidly-updating model versions and input datasets:

- Need to pull new data, reconfigure/debug new model code, and even re-learn the model as if you are a new user.
- For example, it took very long for early users to transition from v9 to v10, due to the new HEMCO emission code and data.
- Similar challenges can happen for the new NetCDF diagnostics module, custom nesting with FlexGrid, and the high-performance GEOS-Chem (GCHP).

Cloud computing provides easy and immediate access to the latest, standard GEOS-Chem and the complete input datasets

- Request "virtual machines" on the cloud, with pre-configured environment ready to run GEOS-Chem
  -- no need to worry about **software**

- Perform computation near the data repository inside cloud, avoiding slow data transfer
  -- no need to worry about **data**

- Virtually unlimited resources provided by AWS and other commercial cloud vendors
  -- no need to worry about **computing resources**

30+ TB of complete GEOS-Chem input data have been available on AWS since Feb 2018

Registry of Open Data on AWS

GEOS-Chem Input Data

climate | weather | meteorological | environmental | air quality | sustainability

Description

Input data for the GEOS-Chem Chemical Transport Model. Including the NASA/GMAO MERRA-2 and GEOS-FP meteorological products, the HEMCO emission inventories, and other small data such as model initial conditions.

Update Frequency

New meteorological and emission data will be added when available.

License

http://acmg.seas.harvard.edu/geos/geos_licensing.html

Documentation

http://cloud-gc.readthedocs.io

Contact

http://acmg.seas.harvard.edu/geos/

Usage Examples

Resources on AWS

Description
Top-level directory for all GEOS-Chem data.

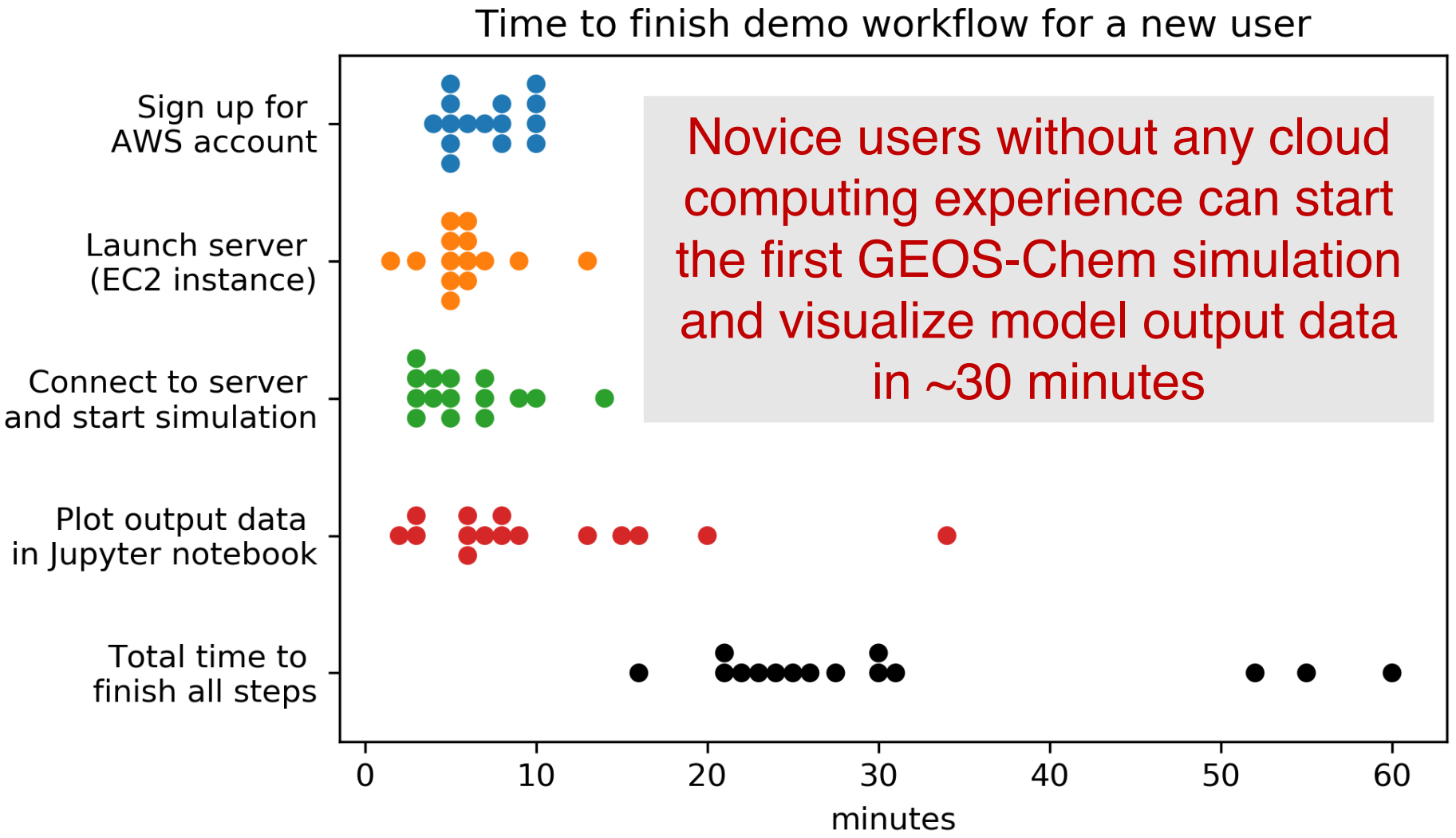
Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::gcgrid`

AWS Region
`us-east-1`

https://registry.opendata.aws/geoschem-input-data/

It is very easy to get started with cloud computing by following our tutorial at http://cloud.geos-chem.org.



(Zhuang et al., 2019, BAMS)

How much does the cloud cost?

For small simulations, the costs are low:

- $30 USD for 1-year global 4° × 5° simulation[1]
- $120 USD for 1-year global 2° × 2.5° simulation
- Can be largely covered by the $100/year student credit[2]

For expensive simulations (e.g. 20-year 2° × 2.5° runs, or 1-year ~0.5° GCHP runs), see AWS Research Credit Program[3], and contact the AWS representatives for your university to work out the proposal.

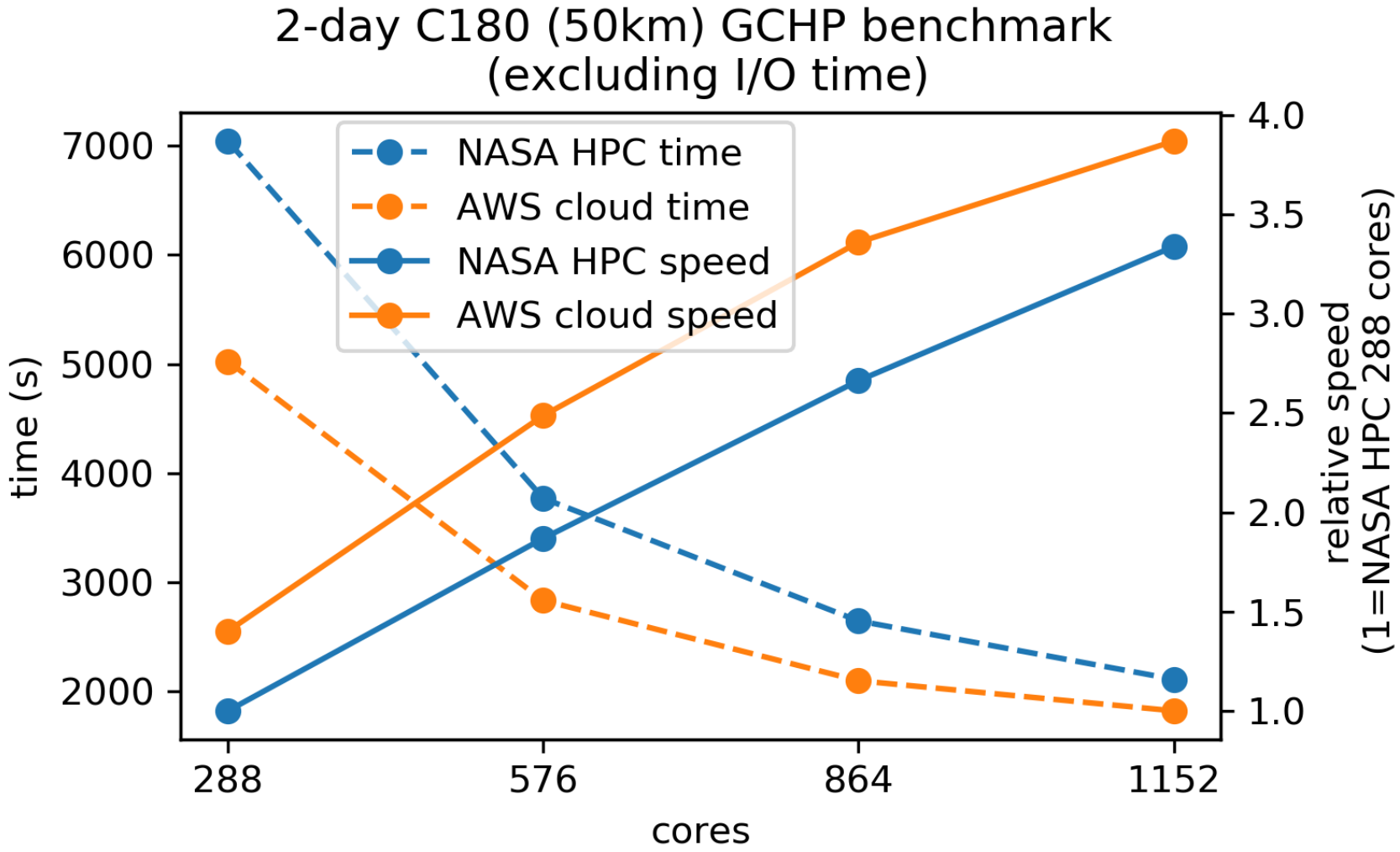
---

1. Assuming the standard tropospheric-stratospheric chemistry, with the default 10min/20 min time step for advection/chemistry. “Tropchem”-only simulations or longer time steps are even cheaper.
2. https://aws.amazon.com/education/awseducate
3. https://aws.amazon.com/research-credits/

For existing users with local clusters, treat the cloud as an additional utility, not a replacement

- It is the quickest way to test the latest model and input datasets without re-configuring your local environment.
- Easily migrate the GEOS-Chem environment across cloud platforms and local clusters using “software containers” such as Docker and Singularity.
- Easily share your model configuration and input data with colleagues, and ensure that the exact simulation can be reproduced by anyone.
- Deal with temporary surges in computing demand.
- The cloud will be an easy way to use the seemingly-complicated High-Performance GEOS-Chem (GCHP)

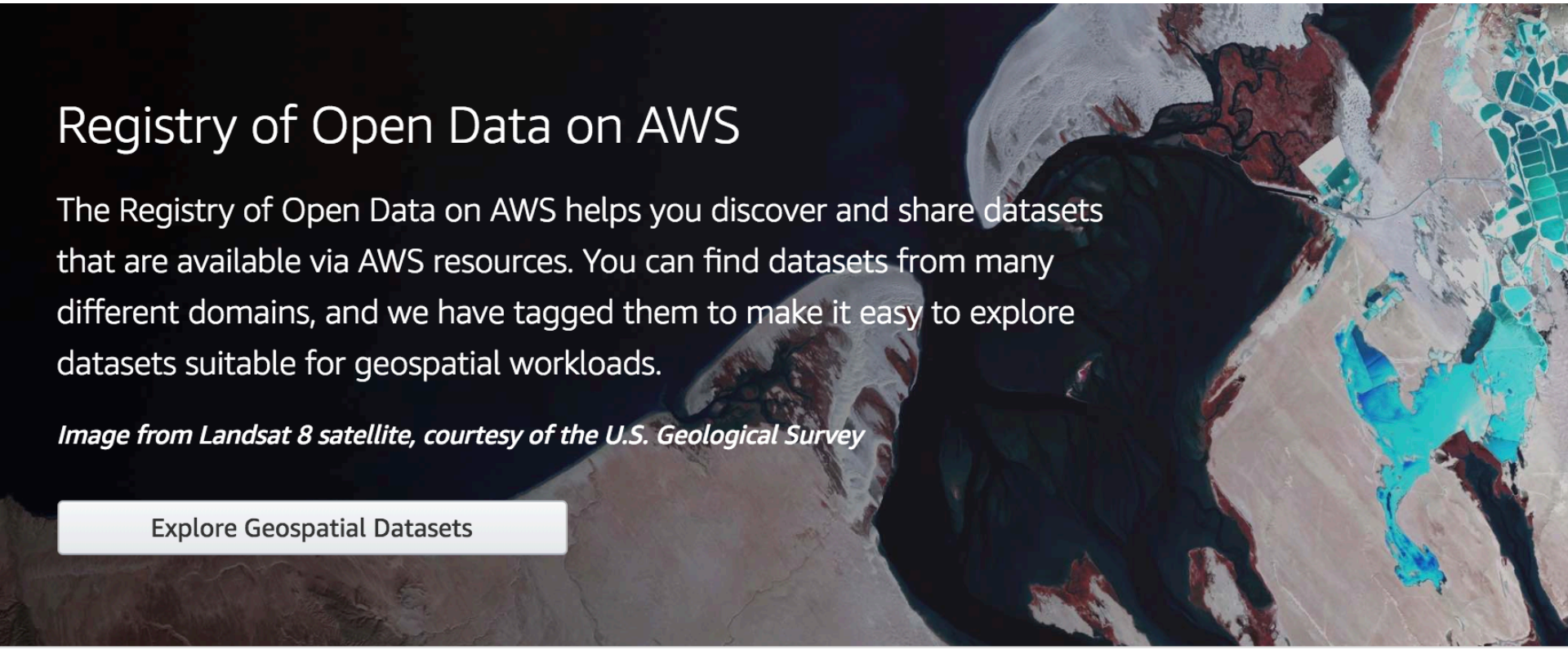
GCHP scales efficiently to ~1000 cores on AWS



- I/O is currently a bottleneck on AWS, still being investigated.
- The new AWS EFA[1] will further improve performance.
- A mature cloud-GCHP capability should be available in a year.

1. https://aws.amazon.com/blogs/aws/now-available-elastic-fabric-adapter-efa-for-tightly-coupled-hpc-workloads/

Cloud computing opens new research opportunities -- government agencies (e.g. NASA, NOAA, ESA) start to share huge amounts of data through the cloud



- Earth science data on AWS: https://aws.amazon.com/earth/
- Registry of Open Data on AWS: https://registry.opendata.aws/

See more in the upcoming BAMS paper:

Zhuang, J., D.J. Jacob, J. Flo-Gaya, R.M. Yantosca, E.W. Lundgren, M.P. Sulprizio, and S.D. Eastham, *Enabling immediate access to Earth science models through cloud computing: application to the GEOS-Chem model*, under review on *the Bulletin of the American Meteorological Society*, 2019.
http://acmg.seas.harvard.edu/publications/2019/zhuang2019.pdf

Attend cloud-computing-related sessions:

- **Model Clinic 4: Using GEOS-Chem on the AWS Cloud**
  Thursday morning (May 9), 10:45-11:15, Maxwell-Dworkin G125
  Presentation on cloud technical concepts and research workflow

- **Hands-on AWS workshop**
  Thursday afternoon (May 9), 3:00-5:00, Maxwell-Dworkin G115
  Free temporary AWS accounts will be provided.
  Will use the just-released GEOS-Chem version 12.3.2