How to Cloud for Earth Scientists: An Introduction

Chris Lynnes, EOSDIS System Architect

Outline

- Cloud Basics
- What good is cloud computing to an Earth Scientist?
- What's the catch?
- Getting Started...

What Cloud Computing Is

- “Someone else’s computer”
- But also someone else’s problem
- Rent instead of own, like:
  - A box truck (Bigger than your SUV)
  - A seat on an airplane (Faster than your sports car)
- Computing a la carte
- Service-based computing

What Cloud Computing Isn’t

1. It’s not the solution for everything
2. It’s not the solution for everyone
3. It’s not a silver bullet

Why Should We Care?

1. More and bigger data are coming
2. Scientists may achieve competitive advantage in proposals
3. We are hearing direction from NASA HQ to buy fewer computers and use more cloud
4. NASA’s Earth Observation data are moving to the cloud

Cloud Fundamentals - Elasticity

- Elastic = scaling up, down or sideways instantly
  - CPU
    - More or less
    - Memory-optimized or compute-optimized or GPU or FPGA
  - Storage:
    - more or less,
    - faster or slower
    - local or web-accessible
- Elastic = pay for only what you use
  - (Remember to turn off when not using!)

“Undifferentiated Heavy Lifting”

Stuff for which you need Earth science expertise
Radiative Transfer Modeling
Atmospheric correction
Bias assessment
Geophysical parameter retrievals
EOF Analyses

“Undifferentiated Heavy Lifting”

Stuff for which you need Earth science expertise	*Stuff for which you **DON’T** need Earth science expertise*
Radiative Transfer Modeling	Ordering a computer
Atmospheric correction	Installing and patching operating systems
Bias assessment	Partitioning and formatting disks
Geophysical parameter retrievals	Calculating power and cooling requirements
EOF Analyses	Locating available floor space for computers

Cloud is “Service - Based”

<X> as a Service	Translation
Infrastructure as a Service	Virtual Machine
Platform as a Service	Virtual Machine pre-loaded with useful software
Software as a Service	Software accessed via the Web, like Google Docs (who cares about the machine?)
Data as a Service	Data accessed through a software interface

Service-based also means...

1. Everything can be coded
2. Everything can be automated
3. The details of How Things Really Work Underneath are often hidden--on purpose


<Interlude: Automation is your friend...>

What Good Is Cloud Computing to an Earth Scientist???

Go Faster

- Commercial cloud CPUs are usually faster than ours...
- ...*And* you can use as many as you want
- Uses
  - Near-real-time processing
  - Massive reprocessing
  - Compute-intensive analysis
  - Deep learning

Pop Quiz!

If a compute-optimized CPU with 16 cores costs 80 ¢ / hr...

And you need 1000 CPU-hours to compute your calculation…

Which of these is cheaper?

1. 1 CPU running for 1000 hours
2. 1000 CPUs running for 1 hour

Answer:

1. 1 CPU * 1000 Hrs * 0.8 = $800
2. 1000 CPU * 1 Hr * 0.8 = $800

Go Bigger

- Many levels of storage
  - Fast and easy but expensive: 30 ¢ / GB-month (EFS)
  - Slow but dirt-cheap: 0.25 ¢ / GB-month (Glacier)
- You can have as much as you want
- Uses
  - Short-term storage of large interim results
  - Long-term storage of data that you *might* need again some day

Go Cheaper

- Pay only for what you use
  - CPU
  - Storage
- Uses
  - Short bursts of lots of processing
  - Lots of storage needed for a short time

Go Simpler

Data as a Service	Quick access to pre-processed data	E.g., Google Earth Engine
Software as a Service	Run analysis on data via the Web	E.g., ArcGIS
Platform as a Service	Use a virtual machine with analysis software pre-installed	E.g., GSFC ADAPT cloud
Infrastructure as a Service	Spin up a machine cluster or get a huge storage system in minutes	Exploratory processing before you win that proposal

Sharing with Cloud

- Store data or results in web-accessible storage (Dropbox, Google Drive) and share URL
- Invoke Software as a Service via URL and share URL

The Reprocessing Example: OCO-2

Example: ABoVE Science Cloud + ADAPT

- ABoVE: Arctic Boreal Vulnerability Experiment
- ADAPT: Advanced Data Analytics **Platform** adjunct to NASA Center for Climate Simulation
- Goals:
  - Bring analysis to the data
  - Enable collaboration among investigators
- Key Cloud / Platform Elements
  - Persistent Data Services (GDS, OPeNDAP, ArcGIS, uvcdat…)
  - Purpose-built VMs for projects
  - Shared high performance filesystem
  - Rapid provisioning of resources

What's the Catch?

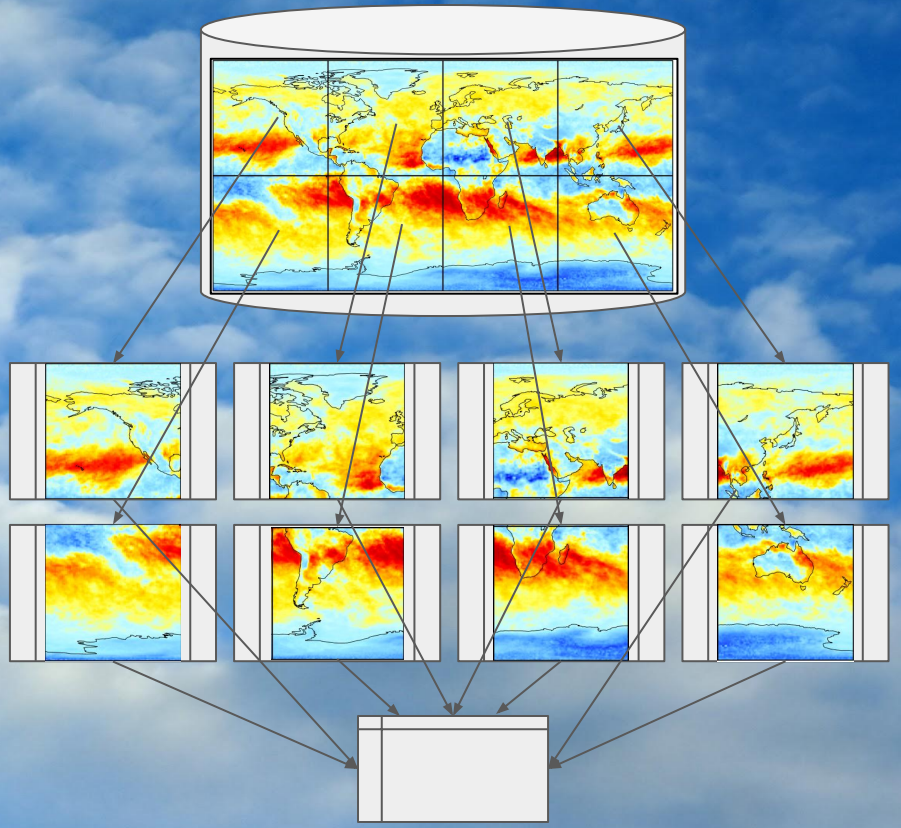
1. New processing paradigm 🤔
2. Failures 💣
3. Policy 🙄
4. Egress charges 🚚

Catch #1: New Processing Paradigm

Bad News:

To get the speedup, you must:

1. Spread input data around
2. Go analyze the pieces
3. Reassemble final result



Catch #1: New Processing Paradigm

Good News:

LOTS of packages and frameworks to help with this

1. Distributed Data Stores
   a. Scalable Databases (Cassandra, Athena...)
   b. Distributed Filesystems (HDFS...)
2. Processing frameworks (MapReduce, Spark)

Pssst….think about learning Python (just sayin')

Catch #2: Failures

Bad News:

thousands of computers

\+ thousands of “disks”

---

s*** happens

Catch #2: Failures

Good News:

- Many of the technologies underlying cloud are there to provide resiliency to node, disk and software failure
- BUT our programs need to be able to pick themselves up and/or restart on another node.

<Interlude: ChaosMonkey>

Image from https://github.com/Netflix/chaosmonkey, open source license Apache 2.0

Catch #3: Policy Stuff

- How/where do I buy/obtain cloud resources?
- Where can I learn how to use cloud--safely?
- What about security policies?

Stay tuned: we are working on these even as we speak...

Catch #4: “Egress” charges

Moving results from cloud to your machine costs money:

First 1 GB/mo	Free!
Up to 10 TB/mo	9 ¢/GB
Next 40 TB/mo	8.5 ¢/GB
Next 100 TB/mo	7 ¢/GB
Next 350 TB/mo	5 ¢/GB

Analyze as much as you can in the cloud to reduce output size

Getting Started with Cloud...

- Many vendors offer free tiers for learning
- There is lots of online training
- There will be more “How to Cloud” seminars to come:
  - Short-learning-ramp ways to use cloud
  - Examples from colleagues
  - What would YOU like to see?

Pssst….don’t forget about the Python thing.

Questions? Requests?

Send requests for cloud seminars to:

chris.lynnes@nasa.gov, EOSDIS