Cloud Computing for Earth Sciences: Deployment of GEOSS Clearinghouse on Amazon’s EC2

Qunying Huang¹, Chaowei Yang¹, Doug Nebert ²
Kai Liu¹, Huayi Wu¹

¹Joint Center of Intelligent Computing
George Mason University

²Federal Geographic Data Committee (FGDC)
ESIP 2011 Winter Meeting
Jan 4th, 2011
Outline

- Introduction
- Related Work
- Deploying GEOSS Clearinghouse onto Amazon EC2
- Conclusion
Introduction

- Data and computational intensive
  - Scientific problems
  - Spatial analysis /algorithm /applications

- Computing paradigm
  - Cluster computing
  - Grid computing
  - Cloud Computing

The growth of cloud computing
From http://www.zdnet.com/blog/hichecliffe
Project Objectives

- **GeoCloud**
  - Ten geospatial application projects in the Cloud environment
  - Common operating system and software suites
  - Deployment and management strategies
  - Usage and costing of Cloud services
  - Security

- **GEOSS Clearinghouse**
  - Metadata catalogues search facility for the Intergovernmental Group on Earth Observation (GEO).
Cloud Computing

- **Definition**
  - “A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction “(NIST, 2010)

- **Defining characteristics**
  - On-demand self-service
  - Multi-tenancy
  - Measured Services
  - Device and Location independent resource pooling
  - Rapid elasticity
Cloud Computing Services

**Software as a Service (SaaS)**
- Almost any IT services
- Users: End-user

**Platform as a Service (PaaS)**
- Platform for developing and delivering applications, abstracted from infrastructures
- Users: Developer

**Infrastructure as a Service (IaaS)**
- On-demand sharing physical infrastructures
- Users: System Administrator
Amazon Cloud Services

- Elastic Compute Cloud – EC2 (IaaS)
- Simple Storage Service – S3 (IaaS)
- Elastic Block Storage – EBS (IaaS)
- SimpleDB (SDB) (PaaS)
- Simple Queue Service – SQS (PaaS)
- Consistent AWS Web Services API (SaaS)
Amazon EC2

- A “Web service that provides resizable compute capacity in the cloud”
- Amazon Machine Image (AMI): a bootable VM image, which can be launched as a EC2 instance
Deployment of GEOSS Clearinghouse on Amazon EC2

1. Launch a CentOS AMI as an EC2 instance
2. Create an EBS volume
3. Authorize network access
4. SSH to the Amazon EC2 virtual server

Restore the GEOSS clearinghouse database

1. Transfer the GEOSS clearinghouse codes/data into the virtual server
2. Mount the EBS volume to the Data/log directory of Postgresql
3. Install Postgresql/Postgis

Install Tomcat, Jetty or other servlet container

1. Configure servlet container for GEOSS Clearinghouse
2. Start servlet container
3. Create a New AMI based on the running instance
Deployment of GEOSS Clearinghouse on Amazon EC2 - cont

- **Scalability**
  - Load balancer

- **Reliability**
  - Network
  - Disaster Recovery

- **Reducing duplicated efforts**
  - Infrastructure
  - Development
# Amazon EC2 Standard Linux Instance Types

<table>
<thead>
<tr>
<th>Type</th>
<th>CPU</th>
<th>Memory</th>
<th>Storage (unformatted)</th>
<th>Platform</th>
<th>I/O</th>
<th>AWS Name</th>
<th>Cost/ hour</th>
</tr>
</thead>
</table>
| Small      | 1 EC2-CU (1 virtual core with 1 EC2 Compute Unit)                    | 1.7 GB (917MB swap) | 170GB instance storage (160GB plus 10GB root partition, 1 spindle)                  | 32-bit   | Moderate  | m1.small | $0.085  
$747 a year or $490.30 a year Reserved         |
| Large      | 4 EC2-CU (2 virtual cores with 2 EC2 Compute Units each)             | 7.5 GB (No swap) | 910GB instance storage (2x450 GB plus 10GB root partition, 3 spindles)              | 64-bit   | High      | m1.large | $0.34   
$2978 a year or $1961 a year Reserved           |
| Extra Large| 8 EC2-CU (4 virtual cores with 2 EC2 Compute Units each)             | 15 GB (No swap) | 1810GB instance storage (4x450GB plus 10GB root partition, 5 spindles)              | 64-bit   | High      | m1.xlarge| $0.68   
$5957 a year or $3922 a year Reserved           |
## Amazon EC2 High-Memory Linux Instance Types

<table>
<thead>
<tr>
<th>Type</th>
<th>CPU</th>
<th>Memory</th>
<th>Storage</th>
<th>Platform</th>
<th>I/O</th>
<th>AWS Name</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-Memory Extra Large</td>
<td>6.5 ECU (2 virtual cores with 3.25 EC2 Compute Units each)</td>
<td>17.1 GB</td>
<td>420 GB</td>
<td>64-bit</td>
<td>High</td>
<td>m2.xlarge</td>
<td>$0.50 per hour</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Memory Double Extra Large</td>
<td>13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each)</td>
<td>34.2 GB</td>
<td>850 GB</td>
<td>64-bit</td>
<td>High</td>
<td>m2.2xlarge</td>
<td>$1.0 per hour</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Memory Quadruple Extra Large</td>
<td>26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each)</td>
<td>68.4 GB</td>
<td>1690 GB</td>
<td>64-bit</td>
<td>High</td>
<td>m2.4xlarge</td>
<td>$2.0 per hour</td>
</tr>
</tbody>
</table>
# Amazon EC2 High-CPU Linux Instance Types

<table>
<thead>
<tr>
<th>Type</th>
<th>CPU</th>
<th>Memory</th>
<th>Storage</th>
<th>Platform</th>
<th>I/O</th>
<th>AWS Name</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-CPU Medium</td>
<td>5 ECU (2 virtual cores with 2.5 EC2 Compute Units each)</td>
<td>1.7 GB</td>
<td>370 GB</td>
<td>32-bit</td>
<td>Medium</td>
<td>c1.medium</td>
<td>$0.17 per hour</td>
</tr>
<tr>
<td>High-CPU Extra Large</td>
<td>20 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each)</td>
<td>7.5 GB</td>
<td>1810 GB</td>
<td>64-bit</td>
<td>High</td>
<td>c1.xlarge</td>
<td>$0.68 per hour</td>
</tr>
</tbody>
</table>
Amazon EC2 New Instance Categories

- **Micro On-Demand Instances**
  - Micro $0.02 per hour

- **Cluster Compute Instances**
  - 10 Gigabit Ethernet
  - Quadruple Extra Large $1.60 per hour

- **Cluster GPU Instances**
  - Quadruple Extra Large $2.10 per hour
GetCapabilities request from different number of concurrent requests
Conclusion

- Cloud computing
- We are at a prescient time
  - Technologies
  - Cloud Architecture
  - Platform independent languages
  - Open data standards
- Spatial Cloud Computing
  - Geospatial Middleware


Thank You!

Qunying Huang  
qhuang1@gmu.edu

Chaowei Yang  
cyang3@gmu.edu

Doug Nebert  
ddnebert@fgdc.gov

CISC  
http://cisc.gmu.edu

Pointers

Portal  
http://aws.amazon.com

Blog  
http://aws.typepad.com

EC2  
http://aws.amazon.com/ec2

S3  
http://aws.amazon.com/s3

Resource Center  
http://aws.amazon.com/resources

Forums  
http://aws.amazon.com/forums