Early Cloud Experiences with the Kepler Scientific Workflow System

Jianwu Wang, Ilkay Altintas
San Diego Supercomputer Center
University of California, San Diego
Background

- **Advantages of Cloud computing**
  - Virtualization, abundance and scalability
  - IaaS, like Amazon EC2, gets virtualized hardware and pre-configured software stack
  - Use available Cloud resources for compute and storage instantly and pay as you go

- **Abundant domain-specific toolkits on Cloud**
  - Bio-Linux AMI: over 500 bioinformatics programs
Cloud Computing Requirements for Scientific Workflows

• Utilize available Cloud resources and packages via workflow
  – An application may execute toolkits on multiple AMIs with dependencies
  – Workflows can help on Cloud instance management, data transfer between instances, program execution on instances, and dependency control

• Data-intensive workflow application on Cloud
  – Localities of data and programs could be achieved for good performance
Kepler Amazon EC2 Actors

- A set of actors to manage EC2 virtual instances on Amazon Cloud and attach EBS Volumes.

![Edit parameters for Run EC2 Instance]

- Access Key: $AccessKey
- Secret Key: $SecretKey
- Image ID: ami-7ae01d13
- Key pair: keypair
- Instance Type: m1.large
- Minimal Instance Number: 1
- Maximal Instance Number: 1
- Availability Zone: us-east-1d
Kepler Amazon Machine Images and EBS Volumes

• **Contents of Kepler Images/Volumes**
  – Kepler system, Kepler workflows, and third-party tools like BLAST

• **Difference between Kepler Images and Volumes**
  – Kepler Images can be used directly to run virtual instances, while Kepler Volumes have to be attached to running instances

• **Virtual clusters via Kepler Images/Volumes**
  – Scalable and virtualized workflow execution
Usage Mode 1

Kepler EC2 Actors + Third Party AMIs
Sample Workflow for Usage Mode 1

AccessKey: 
SecretKey: 
AccessIdentity: 

BLAST command on EC2 Instance
blastn -query /home/ubuntu/testdbTiny -db /home/ubuntu/CAM_PROJ_PBSM.fq -out /home/ubuntu/test.out

Output File on EC2 Instance
input + "/home/ubuntu/test.out"

Local Path
"/home/ec2-user/"

SSH File Copier
"/home/ec2-user/input/testdbTiny"

Query Sequence Path on EC2 Instance
input + "/home/ubuntu/"

Local Reference Sequence Database
"/home/ec2-user/database/"

Sequence Database Path on EC2 Instance
input + "/home/ubuntu/"

SSH Execute for BLAST
Stop EC2 Instance
Usage Mode 2

Kepler EC2 Actors + Kepler AMI

http://biokepler.org/
Usage Mode 3

Kepler EC2 Actors + Kepler AMI
+ Third Party AMIs

http://biokepler.org/
Sample Workflow for Usage Mode 3

- StartInstance
- Expression "ec2-user"+input
- Wait instance
- User credential file "/Users/jianwu/Projects/"
- Path in EC2 instance input + ":/home/ec2-user/"
- Workflow to be executed on EC2 instance "/home/ec2-user/database/"
- Workflow Path in EC2 instance input + ":/home/ec2-user/"
- SSH File Copier3
- SSH Execute Remove Identity File
- SSH Execute Kepler Workflow on another EC2 instance
- Remove credential file command "rm -rf /home/ec2-user/id_rsa-camera-keypair"
- Output path on Kepler EC2 instance input + ":/home/ec2-user/test.out"
- Local path for output file "/Users/jianwu/ + input + ":-test.out"
- Synchronizer "/home/ec2-user/kepler/kepler.sh -runw -nogui...

run uploaded workflow on virtual machine
upload the workflow for usage mode 1

6/2/12
http://biokepler.org/
Usage Mode 4

Virtual Cluster based on Kepler AMI

SDSC 6/2/12 http://biokepler.org/
Summary

• Early Cloud Experiences with Kepler
  – A set of EC2 actors
  – Kepler AMI and EBS Volumes
  – Different usage modes

• Future Work
  – Compare the usage modes through experiments
  – Optimize data-intensive workflow execution on EC2
Questions?

• More Information

  {jianwu, altintas}@sdsc.edu
  http://www.kepler-project.org
  http://www.bioKepler.org

• Acknowledgements
  – NSF OCI-0722079 for Kepler/CORE, DBI-1062565 for bioKepler
  – Gordon and Betty Moore Foundation for CAMERA
  – UCSD Triton Research Opportunities Grant
  – AWS in Education Research Grants from Amazon.com, Inc for EC2 usage credit