

Early Cloud Experiences with the Kepler Scientific Workflow System



Jianwu Wang, Ilkay Altintas

*San Diego Supercomputer Center
University of California, San Diego*

SDSC



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



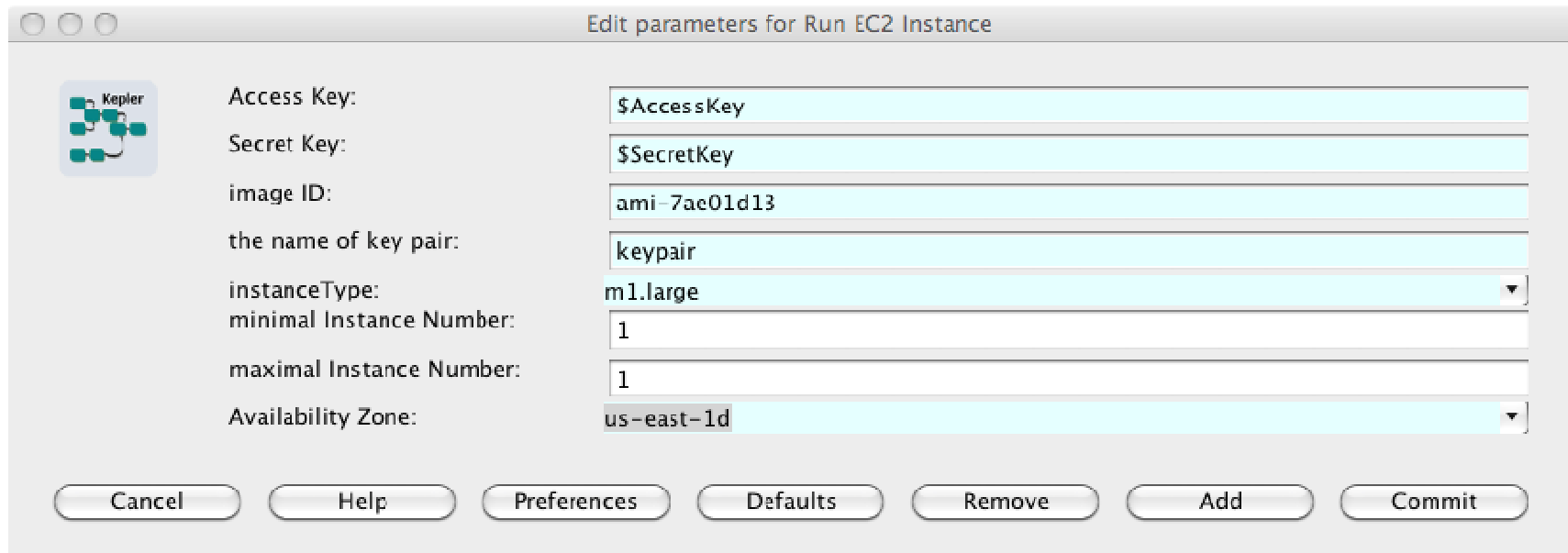
Cloud Computing Requirements for Scientific Workflows

- **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.



Kepler

Access Key: \$AccessKey

Secret Key: \$SecretKey

image ID: ami-7ac01d13

the name of key pair: keypair

instanceType: m1.large

minimal Instance Number: 1

maximal Instance Number: 1

Availability Zone: us-east-1d

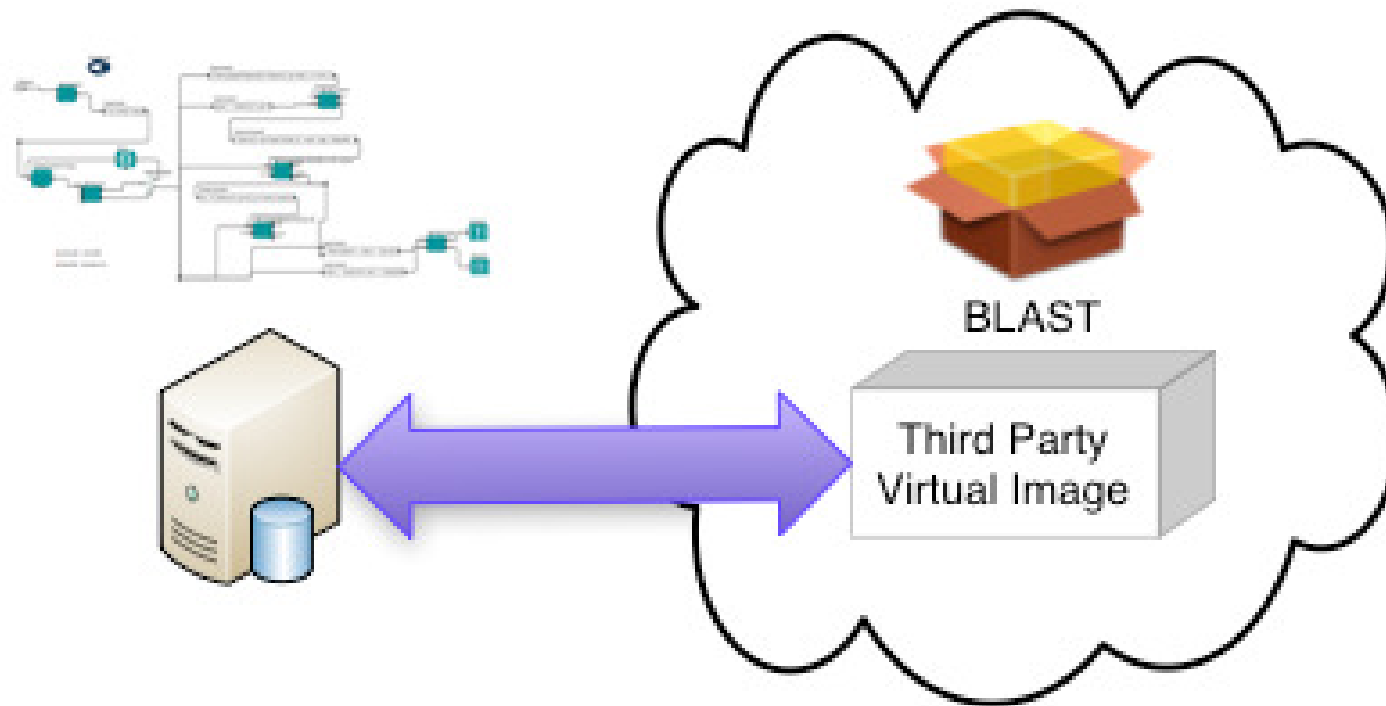
Cancel Help Preferences Defaults Remove Add Commit

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

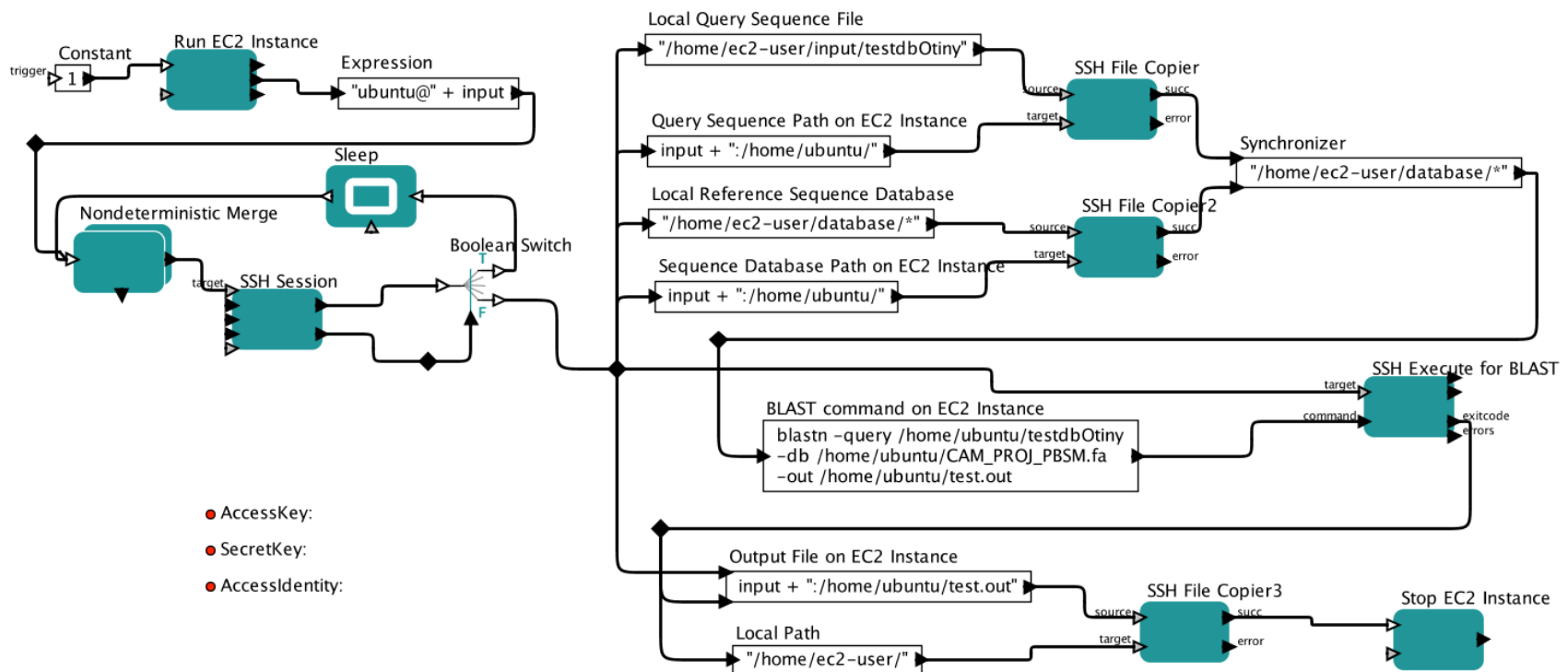


Usage Mode 1



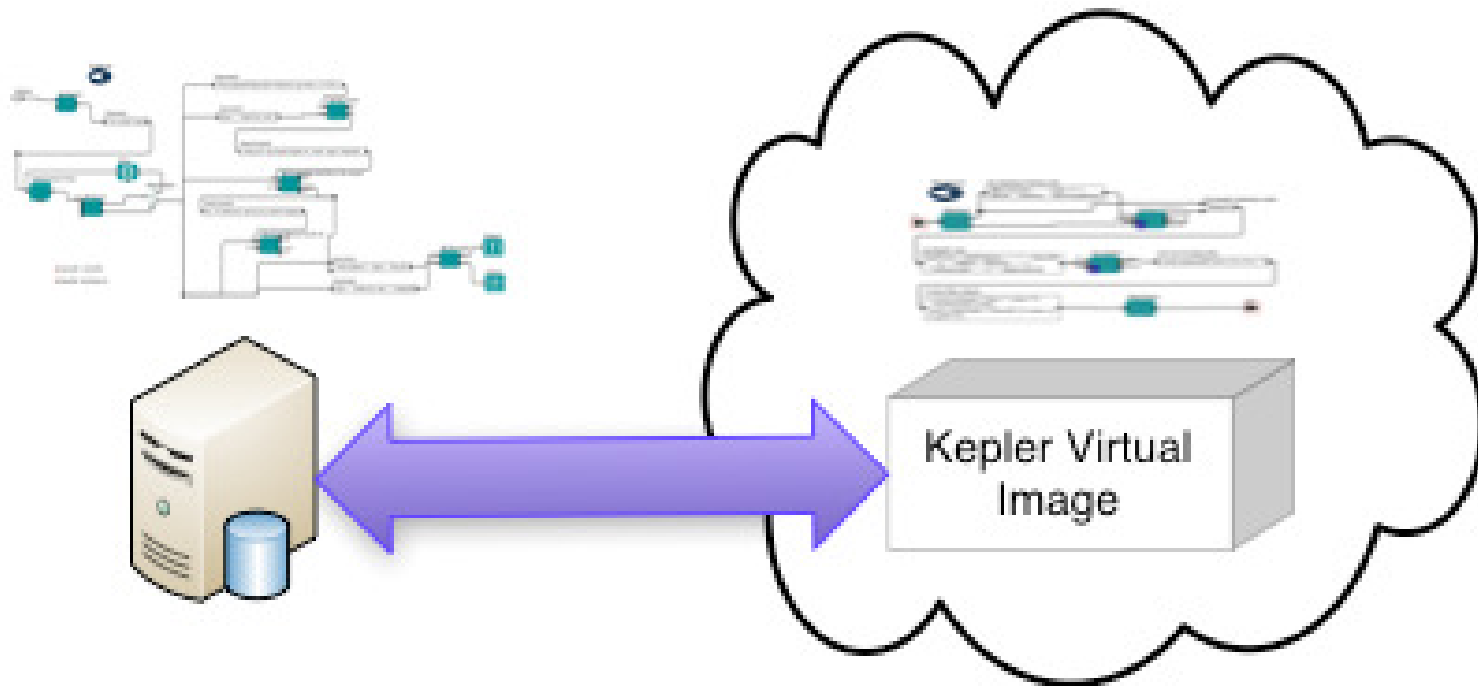
Kepler EC2 Actors + Third Party AMIs

Sample Workflow for Usage Mode 1



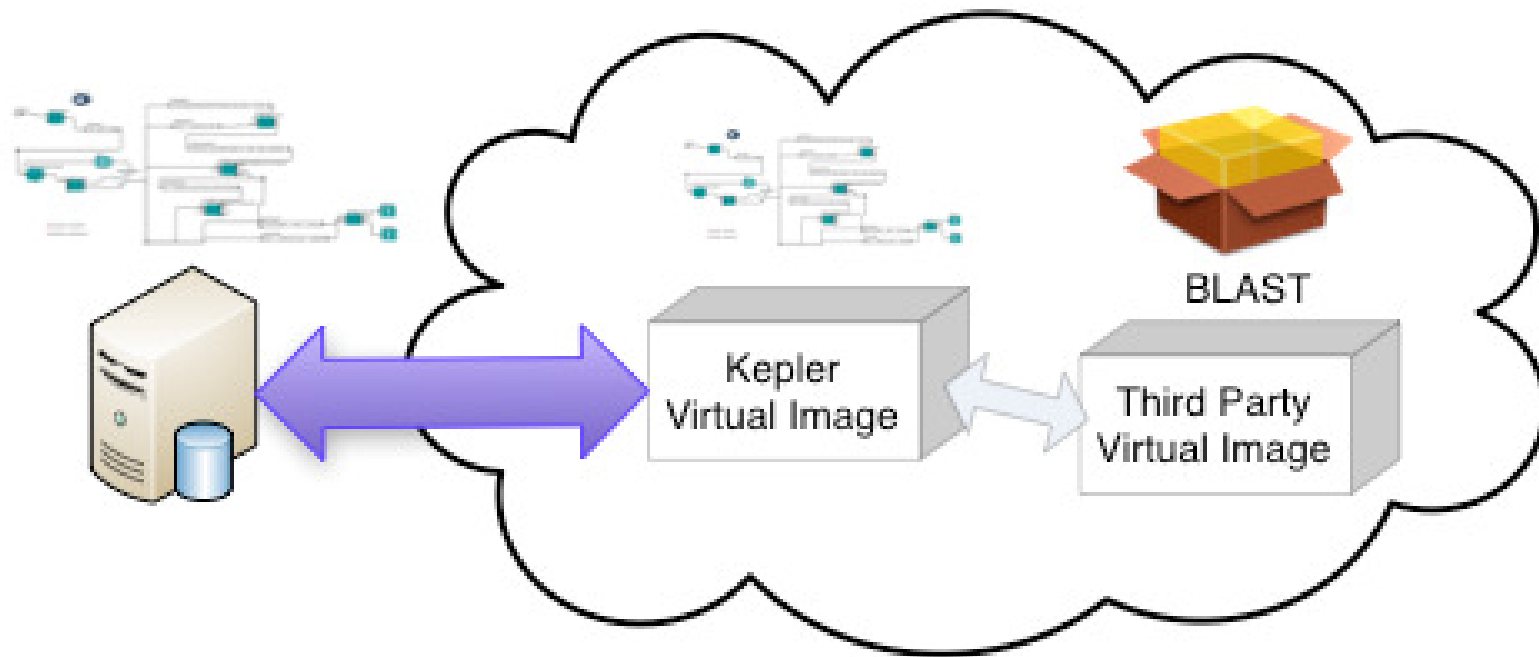
- AccessKey:
- SecretKey:
- AccessIdentity:

Usage Mode 2



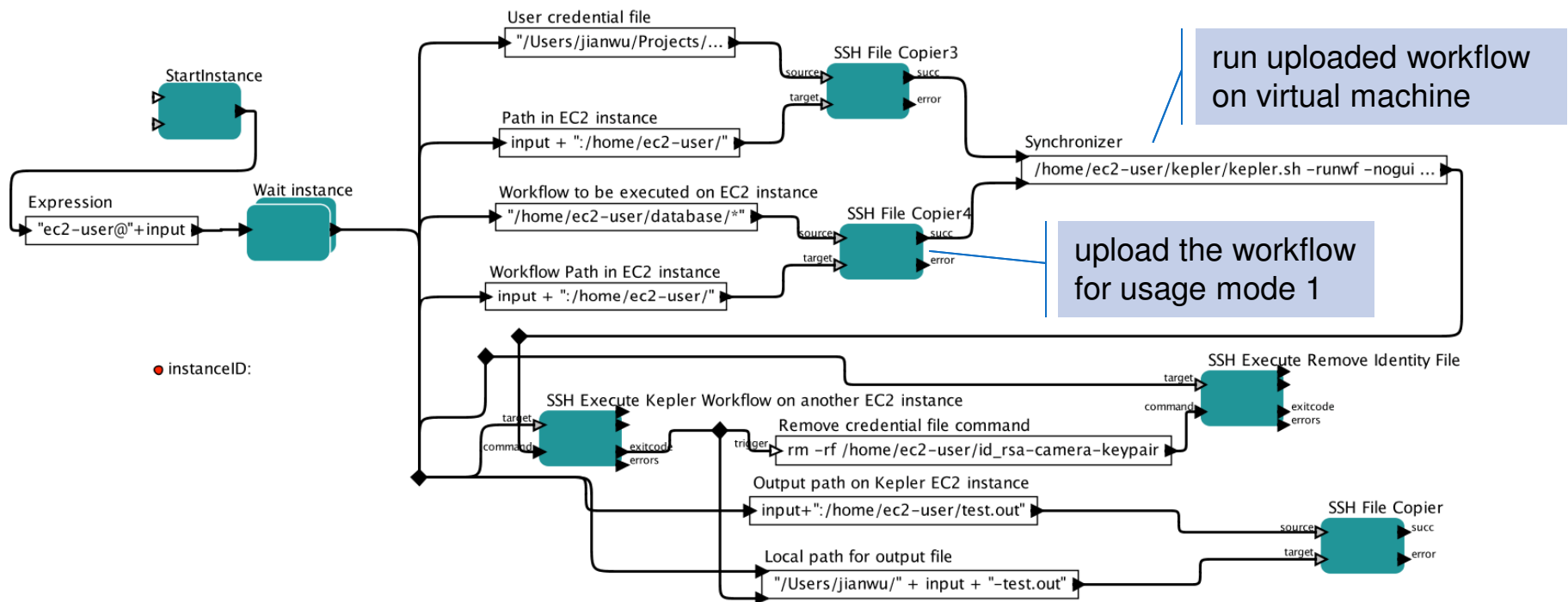
Kepler EC2 Actors + Kepler AMI

Usage Mode 3

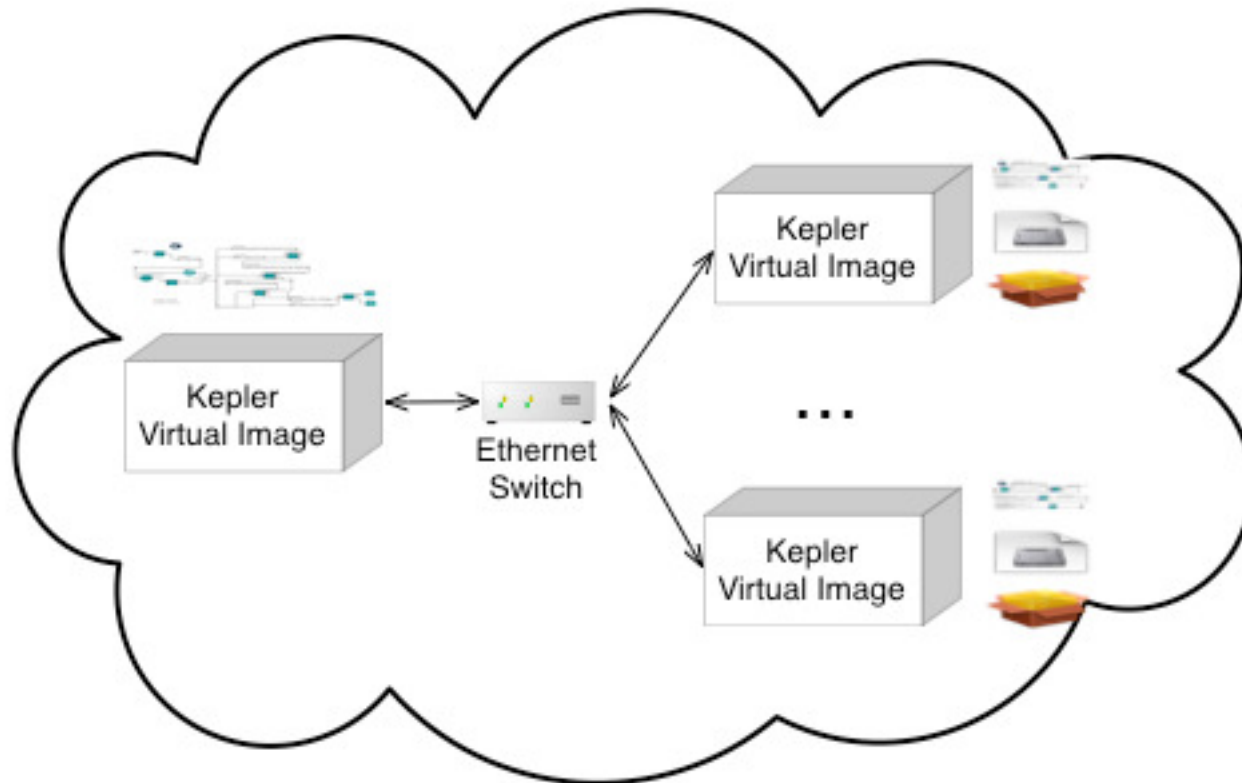


**Kepler EC2 Actors + Kepler AMI
+ Third Party AMIs**

Sample Workflow for Usage Mode 3



Usage Mode 4



Virtual Cluster based on Kepler AMI

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

