Deployment of Virtual Clusters in a Multi-Cloud Environment for Molecular Docking

Anthony Nguyen, Jason H. Haga, Mauricio Tsugawa, Kohei Ichikawa
October 17, 2014

Software Design and Analysis Lab, NAIST
Infraware Research Group, AIST
Advanced Computing and Information Systems Laboratory, University of Florida

Background

- Multi-cloud computing is both cost efficient and offers a high degree of versatility and flexibility.
  - No hardware cost (pay for time used)
  - Essentially limitless resources
  - Can easily scale and alter computational environment
  - Adds a degree of fault tolerance (If one cloud has high traffic or goes down for maintenance others still run)

Objective

- The goal is….
  - Develop a multi-cloud environment to run virtual screenings
    - Virtual screenings of molecular dockings performed via DOCK program
- Anticipating that the multi-cloud environment will…
  - Simplify process to scale the number of screenings
  - Maintain same performance level as grid computing

Run DOCK Test

Project Approach

Phase 1

- Launch VMs on Different Clouds

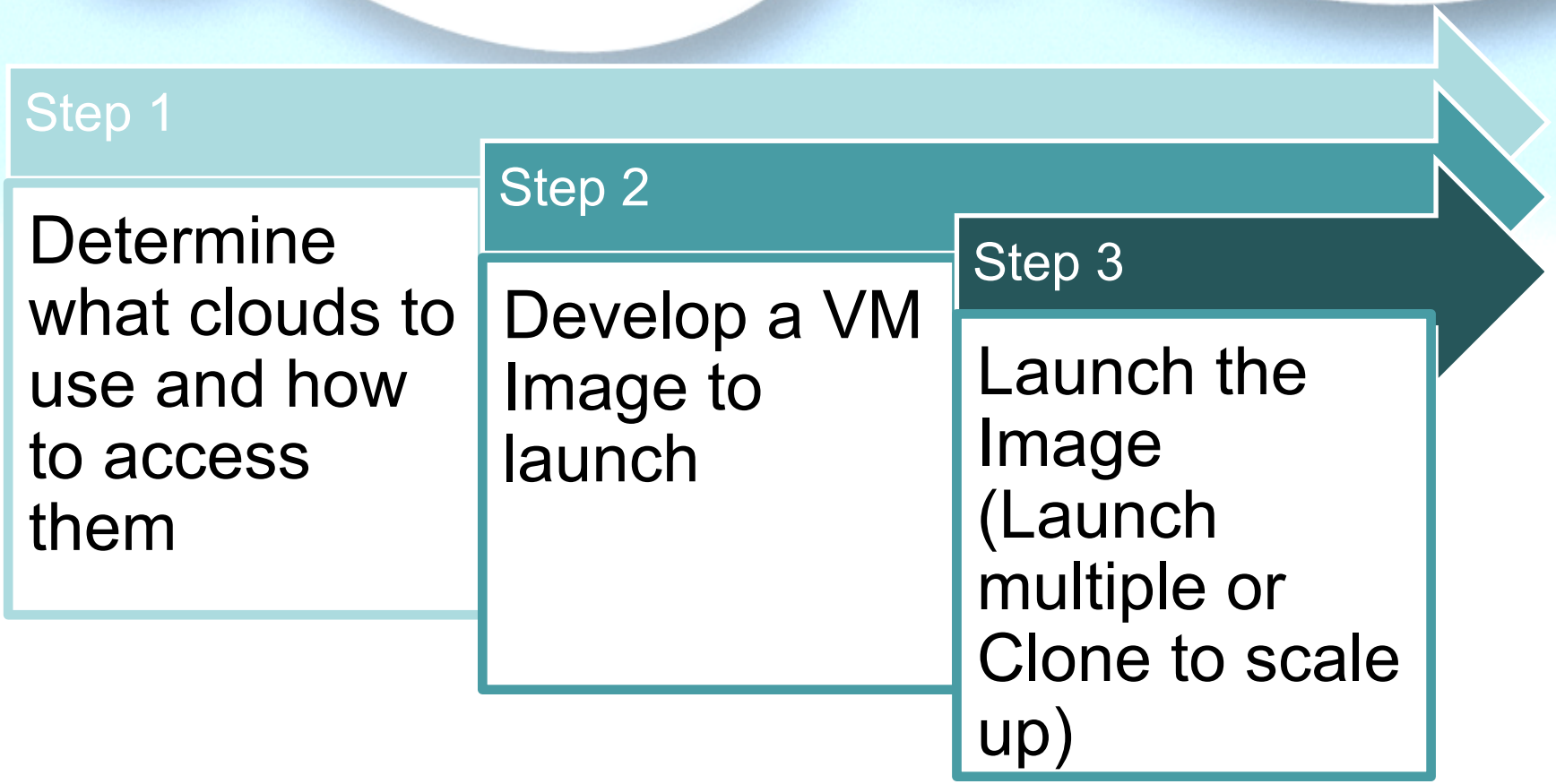
Phase 2

- Establish Connectivity Across Clouds

Phase 3

- Run DOCK Tests and Analyze Results

Phase 1: Launch VMs on Different Clouds



Step 1: Cloud Selection

- Factor
  - Low Cost or Free
- Clouds Selected
  - NAIST Private Cluster Environment
  - AIST Private Cluster Environment
  - FutureGrid Alamo Cloud Computing Resource
    - Proposal Submitted to attain access to FutureGrid resources.

Step 2: Developing VM Image

- Used a KVM Virtual Image designed by UCSD PRIME 2013 Interns
  - This Virtual Image was set up to optimize DOCK program performance with regards to consistency in results.
  - Some modification were necessary to match specific requirements of cloud (network configuration)

Step 3: Launch the Image

- Each Cluster uses a different Cloud Middleware System so image launch method varied from cloud to cloud
  - NAIST: Libvirt (KVM)
  - AIST: OpenNebula
  - FutureGrid Alamo: Nimbus

Project Approach

Phase 1

- Launch VMs on Different Clouds

Phase 2

- Establish Connectivity Across Clouds

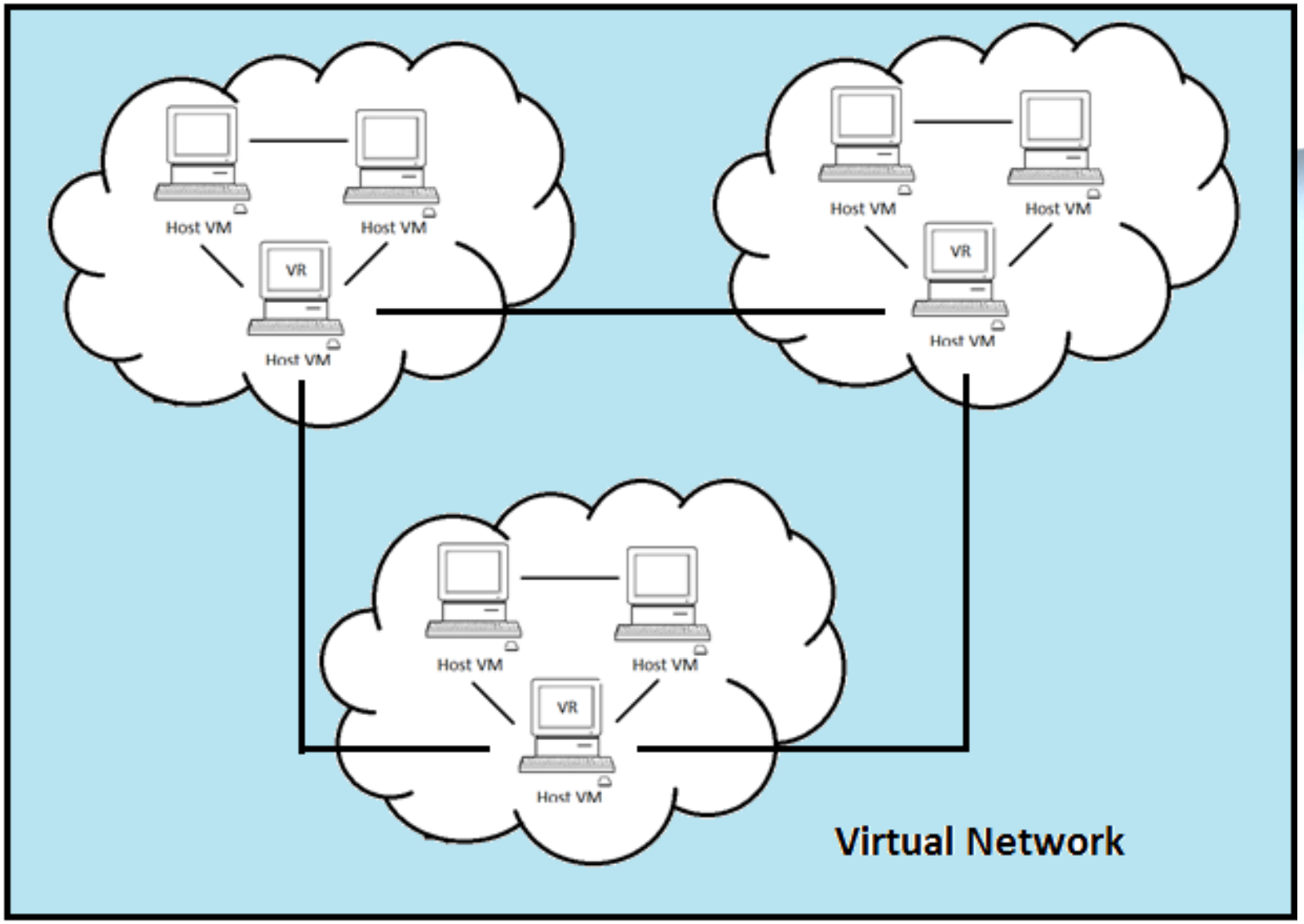
Phase 3

- Run DOCK Tests and Analyze Results

Phase 2: Establish Connectivity

- ViNe
  - Can be used to establish a virtual network supporting symmetric communication beyond boundaries like firewalls
  - Requires no change to physical network
  - Requires no change to machines operating system

How ViNe Works



Project Approach

Phase 1

- Launch VMs on Different Clouds

Phase 2

- Establish Connectivity Across Clouds

Phase 3

- Run DOCK Tests and Analyze Results

Phase 3: DOCK Tests and Results

- With full connectivity in Multi-Cloud Environment, DOCK jobs were run using mpi (message passing interface)
- The tests that were run were used to assess effectiveness and flexibility of a multi-cloud environment.

Our Environment

- Three VMs on NAIST
  - One is master, two run DOCK
- Three VMs on AIST
  - All three run DOCK
- Three VMs on FutureGrid
  - All three run DOCK

Results

- Processing times for DOCK tests in same order of magnitude for Grid Computing and Multi-Cloud Environment
  - Grid Computing DOCK Test
    - 82s per compound per processor
  - Multi-Cloud DOCK Test
    - 30s per compound per processor

Results

- Processing times longer with increased workloads
  - Anticipate that this will no longer be the case when using larger clouds with more resources
- Jobs evenly distributed across clouds and VMs when scaling environment
  - Jobs distributed on a per VM basis, not a per cloud basis

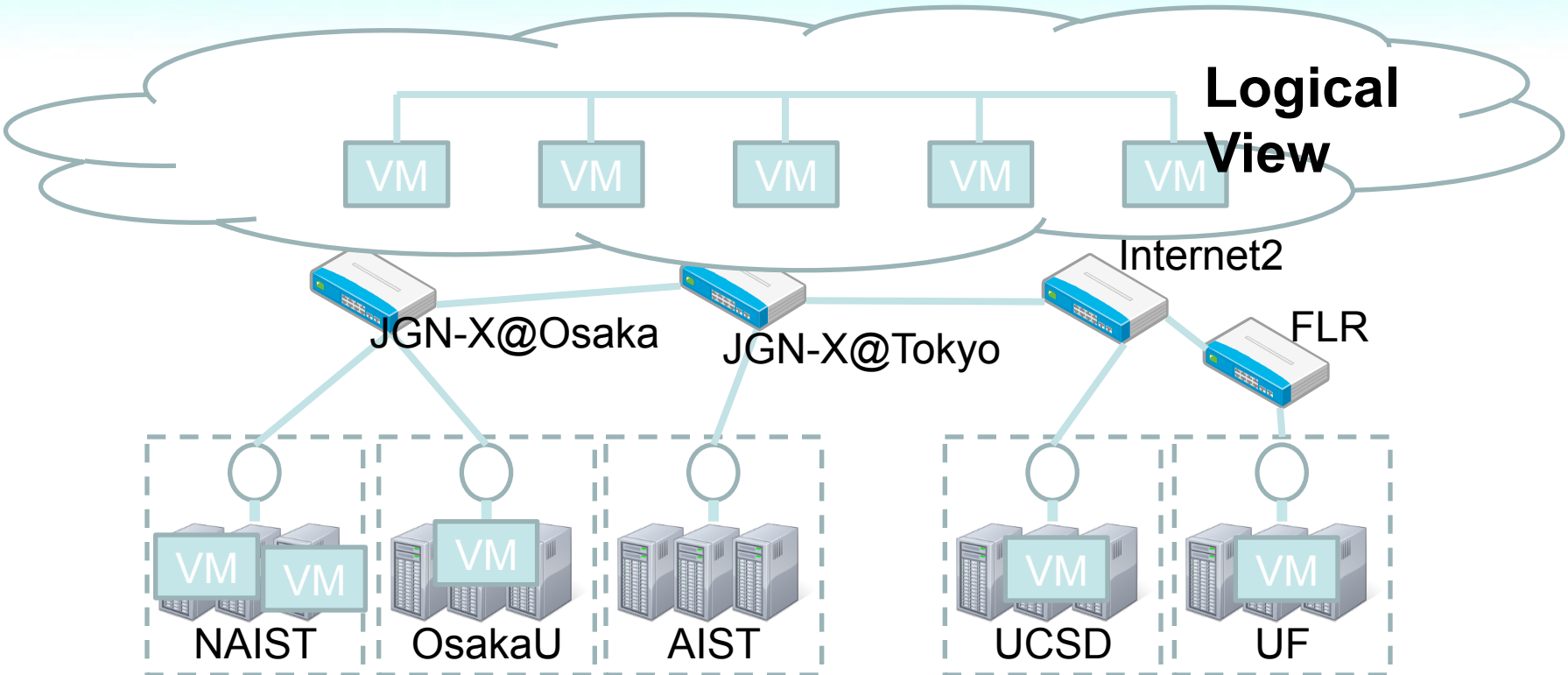
Our Demo Environment

- Two VMs on NAIST
  - One is master, one run DOCK
- One VM on Osaka U
- One VM on UF
- One VM on UCSD

*Connectivity through PRAGMA-ENT

PRAGMA-ENT

- Application using multisite resources
  - Pragma-ENT allows to provide a single L2 flat network for applications. This makes it easy for application users to use multisite resources simultaneously.



DOCK Demo Output

Conclusions and Future Work

- Scaling up of the number of virtual screenings and VMs on the environment was simple and quick
- Performance on the Multi-Cloud environment was equivalent to that on a grid computing environment

- Incorporate Hadoop for job distribution and fault tolerance
- Incorporate larger commercial clouds (Azure) for larger scale tests

Acknowledgements

- Previous PRIME students
  - Kevin Lam
  - Karen Rodriguez
- UCSD Pacific Rim Experiences for Undergraduates (PRIME)
  - Dr. Gabriele Wienhausen
  - Teri Simas
  - Jim Galvin
  - Madhvi Acharya
- Funding Sources
  - URS Ledell
  - PRIME Alumna Haley Hunter-Zinck
  - National Science Foundation

Rigid Test with Flexible Parameter Off

Molecules Processed per Host (100 total)			Test 1 (Small Workload)
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	10	10
AIST	13	11	14
FG	16	14	12

Molecules Processed per Host (100 total)			Test 2 (Large Workload)
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	14	13
AIST	10	12	12
FG	13	11	15

Rigid Test with Flexible Parameter Off

Average Processing Time per Molecule on each Host (sec) Test 1 (Small Workload)			Total Time: 1066 sec
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	96	140
AIST	79	90	70
FG	66	71	81

Average Processing Time per Molecule on each Host (sec) Test 2 (Large Workload)			Total Time: 3015 sec
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	197	222
AIST	280	237	233
FG	232	241	189

Rigid Comparison Test (100 Compounds)

Average Processing Time per Molecule on each Host (Multi-Cloud) (sec) Total Time: 24380			
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	1290	1561
AIST	3104	1184	1649
FG	1907	2603	1850

- Published Rate for Grid Computing DOCK Test
  - **82 seconds per compound per processor**
- Multi-Cloud DOCK Test
  - **30 seconds per compound per processor**

Unbalanced Rigid Test

Molecules Processed per Host (Balanced) (100 total)			Test 1
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	10	10
AIST	13	11	14
FG	16	14	12

Molecules Processed per Host (Unbalanced) (100 total)						
Cloud	**Host 1**	**Host 2**	**Host 3**	**Host 4**	**Host 5**	**Host 6**
NAIST	MASTER	8	9	-	-	-
AIST	9	8	7	-	-	-
FG	9	11	10	9	11	9

Unbalanced Rigid Test

Average Processing Time per Molecule on Each Host (Balanced) (sec) Total Time: 1066 sec			
Cloud	**Host 1**	**Host 2**	**Host 3**
NAIST	MASTER	96	140
AIST	79	90	70
FG	66	71	81

Average Processing Time per molecule on each Host (Unbalanced) (sec) Total Time: 766 sec						
Cloud	**Host 1**	**Host 2**	**Host 3**	**Host 4**	**Host 5**	**Host 6**
NAIST	MASTER	91	82	-	-	-
AIST	73	89	97	-	-	-
FG	78	60	65	80	64	85