Deployment of a Multi-Site Cloud Environment for Molecular Virtual Screenings

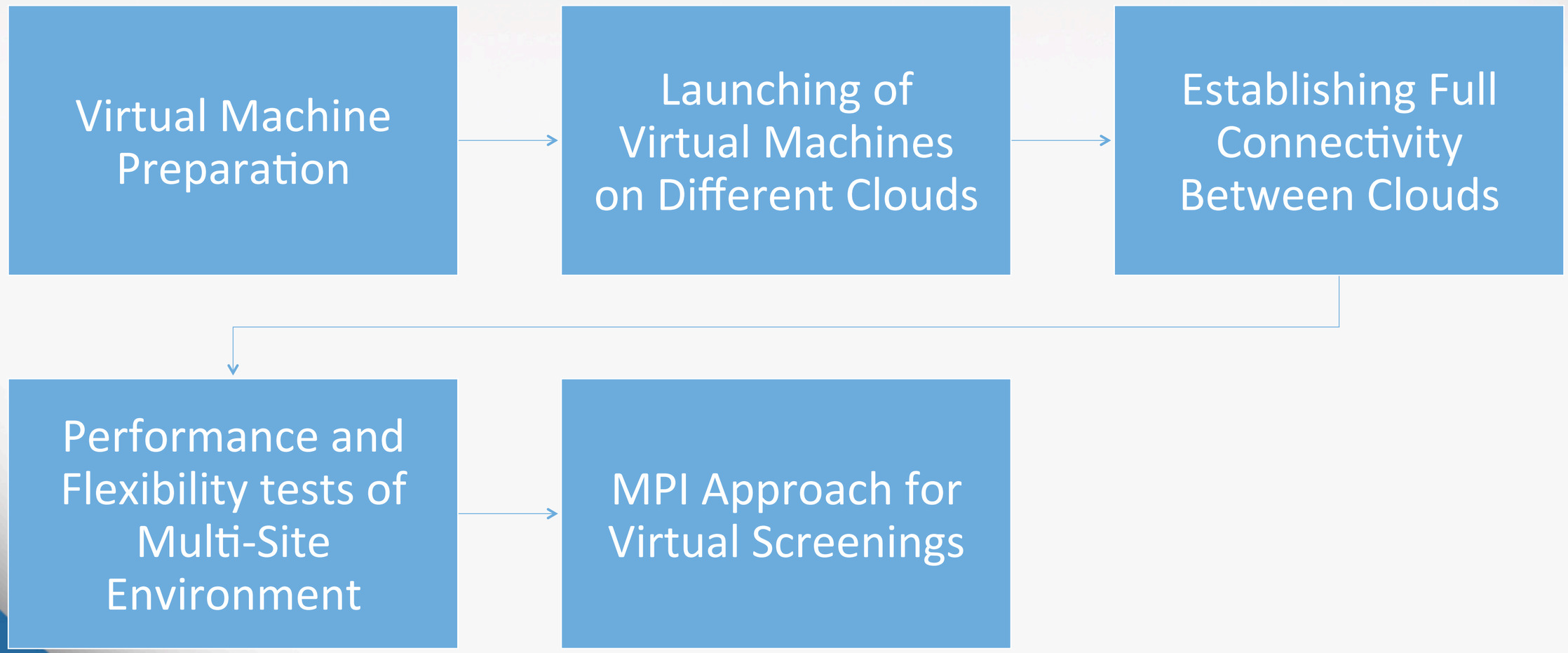
Anthony Nguyen, Andréa Matsunaga, Kohei Ichikawa, Susumu Date, Maurício Tsugawa, Jason H. Haga

Cyber-Physical Cloud Research Group, AIST
Software Design and Analysis Lab, NAIST
Advanced Computing and Information Systems Laboratory, University of Florida

Background

- Drug Discovery and Development
  - Process is driven by wet-lab tests
    - Provide understanding protein-ligand interactions
    - Very costly, time consuming process
  - Simulation programs such as DOCK act as a screening method prior to wet-lab
    - Can greatly reduce overall cost
- Multi-Site Cloud Environment
  - Restrictions of DOCK on grid-computing
    - Variablity in results due to heterogeneity of computing environment
    - Hardware limits system scalablity
  - Virtual Machines and Cloud Computing
    - Allow for homogeneity in computing environment
    - Access to virtually unlimited resources
  - Using multiple providers (i.e. multi-site cloud)
    - Optimize screenings under time and money constraints

Methods and Approach



Virtual Machine Deployment

- Scientific code fingerprinting tool (UCSD)
  - Analyzed dependencies of DOCK software on 32-bit CentOS5 and packaged it to be executed on a 64-bit CentOS6 environment
- DOCK output on both environments consistent
- Cloud locations utilized
  - NAIST (Nara Institute of Science and Technology)
  - UF (University of Florida)
  - UCSD (University of California, San Diego)
  - Microsoft Azure (WestUS Resource)
- Cloud selection requirements
  - Free cost
  - Easy accessibility

Azure Pricing Analysis

- A-series should be avoided
- D-series if cost is a factor
- G-series if time is a factor

TABLE II
AZURE WESTUS STANDARD TIER INSTANCES TESTED

VM Type	Cores	RAM	HD Space	VM Pricing (US dollars)	Time to complete screening (sec)
A3	4	7 GB	285 GB	$0.24/hr	1427
D1	1	3.5 GB	50 GB	$0.094/hr	889
D3	4	14 GB	200 GB	$0.376/hr	865
G1	2	28 GB	384 GB	$0.61/hr	510
G2	4	56 GB	768 GB	$1.22/hr	536

Pricing information from the Microsoft Azure website as of March 10, 2015 [26].

Virtual Machine Deployment

- General information hardware configuration at various cloud provider sites

TABLE I
RESOURCES AT EACH CLOUD PROVIDER SITE

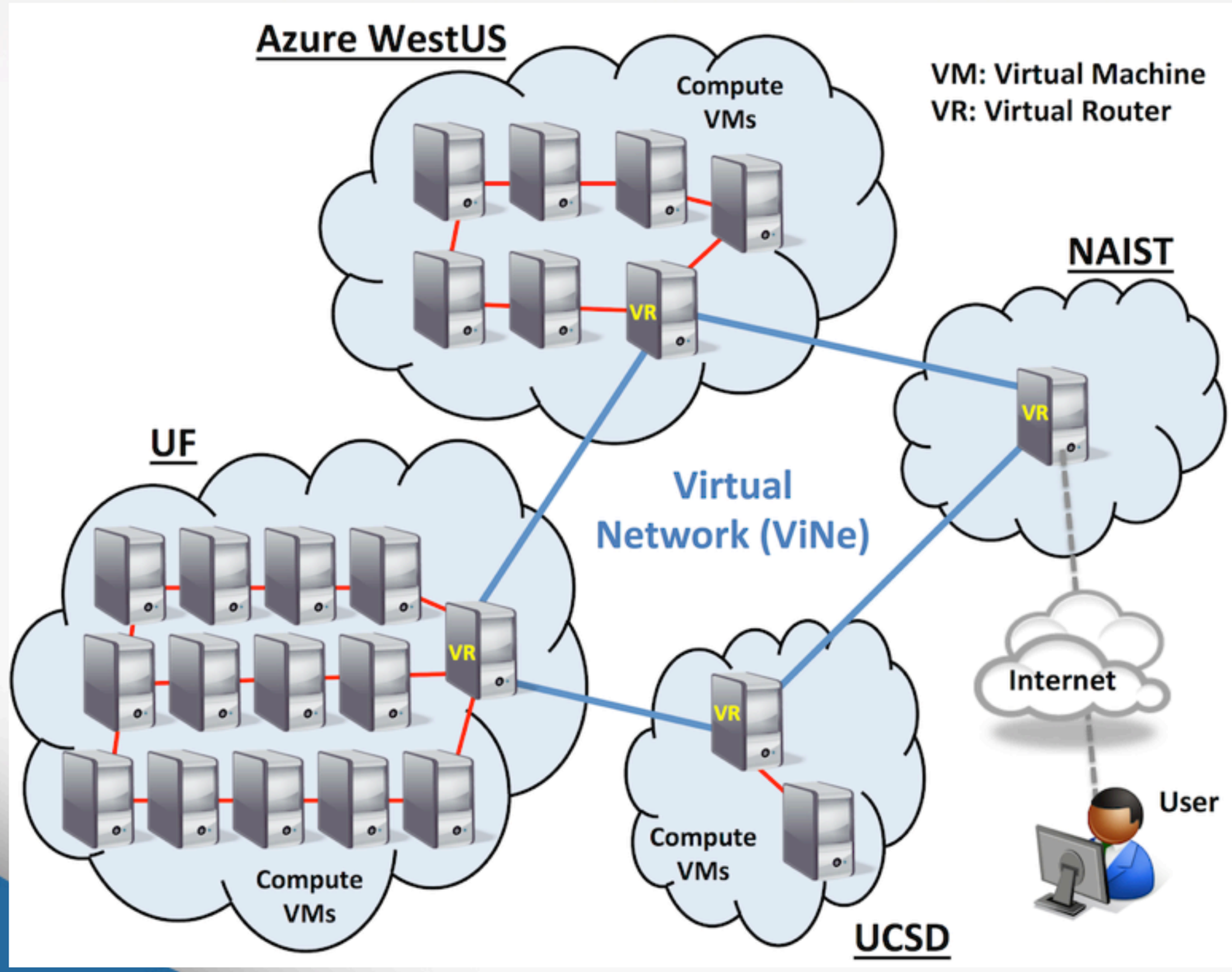
	Number of cores	Number of VMs	RAM per VM	Processor
NAIST	2	1	2GB	Xeon W3565
UCSD	16	2	8GB	Xeon E5520
Azure	28	7	56GB	Xeon E5-2698B v3
UF	56	14	6GB	Xeon 5140

Multi-Site Virtual Cluster Connectivity

- ViNe (UF)
  - Requires no change to specifications of the VM
  - Allows connection beyond boundaries like firewalls
- Deployment of ViNe
  - Virtual router (VR) host formed on each cloud
    - Accompanied by list of virtual IP addresses
  - Each VM receives a virtual IP address
    - Creates connectivity within one cloud through the VR
  - All VRs in setup join the same virtual network
    - Creates connectivity across entire environment through VRs

Multi-Site Virtual Cluster Connectivity

- Overall multi-site cloud system connected with ViNe



Molecule Preparation and Tests

- Molecule prepared using molecular visualization program Chimera
  - Used SSH-2
- SSH-2 tested with random selection of chemical compounds from ZINC database
  - “Clean Drug-Like” Subset
  - 30,000 compounds

Molecule Preparation and Tests

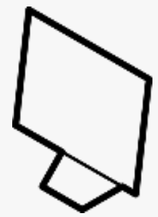
- Test Parameters
  - Correspond to a low accuracy screening
  - Docking completion times 10-20 seconds/compound
    - Will be referred to as effective “Processing Rate”
- Tests (using mpi)
  - Effect of geographic location of master node
  - Effect of how system effected by different loads
  - Scalability of the system

How mpi Enabled DOCK Works

Diagram Key

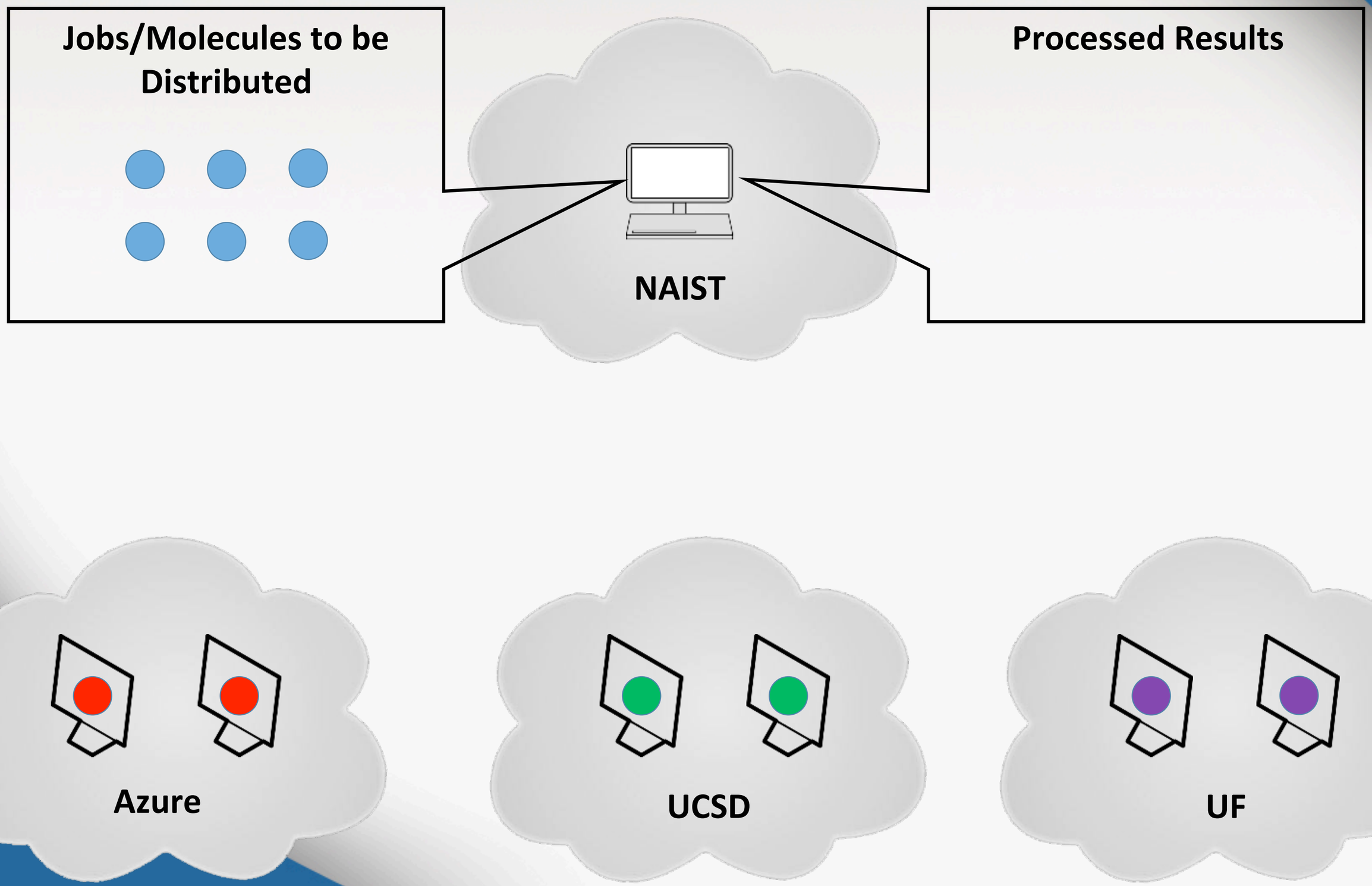
- Blue Circle: Job that has not yet been processed
- Red/Green/Purple Circle: Processed Job (Result)
  - Red: Azure
  - Green: UCSD
  - Purple: UF
- Circle Movement Time: Latency (i.e. time it takes for jab/result to send)

Master Node

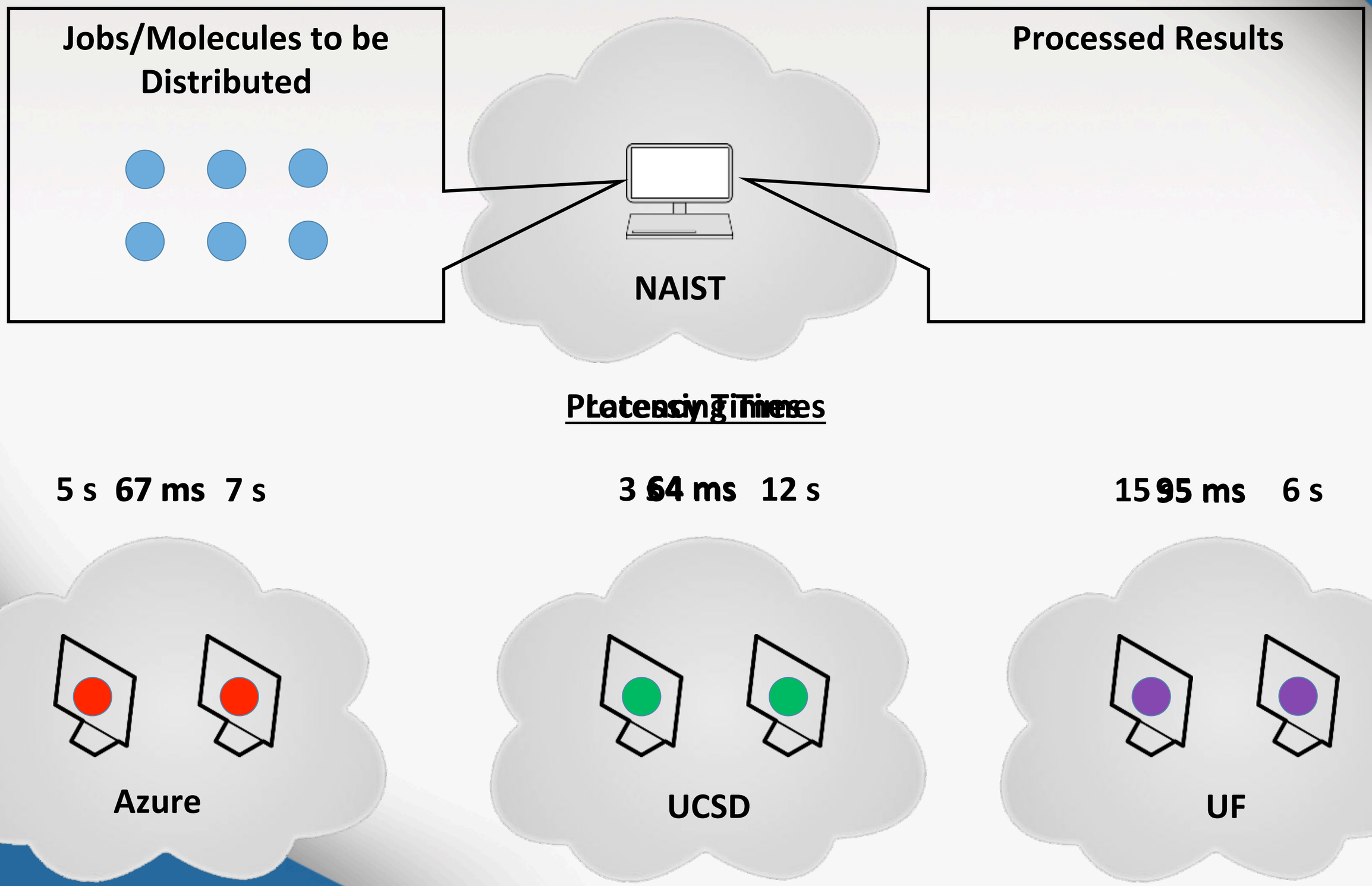


Compute Node

How mpi Enabled DOCK Works



How mpi Enabled DOCK Works



How mpi Enabled DOCK Works

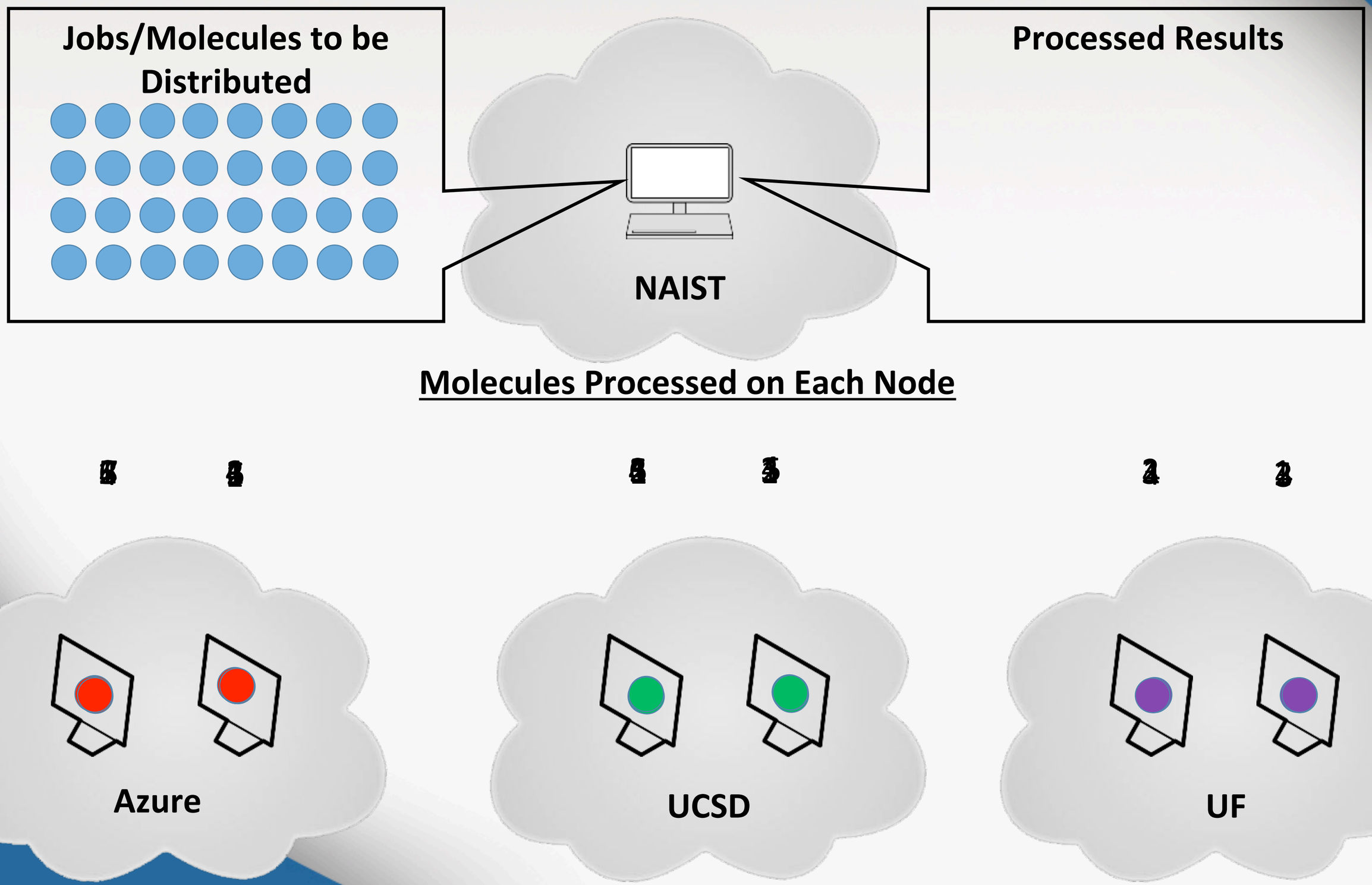
- Latency Times

TABLE III
ROUND-TRIP LATENCY (MS)

	NAIST	UCSD	Azure	UF
NAIST	-	129	133	190
UCSD	129	-	20	60
Azure	133	20	-	71
UF	190	60	71	-

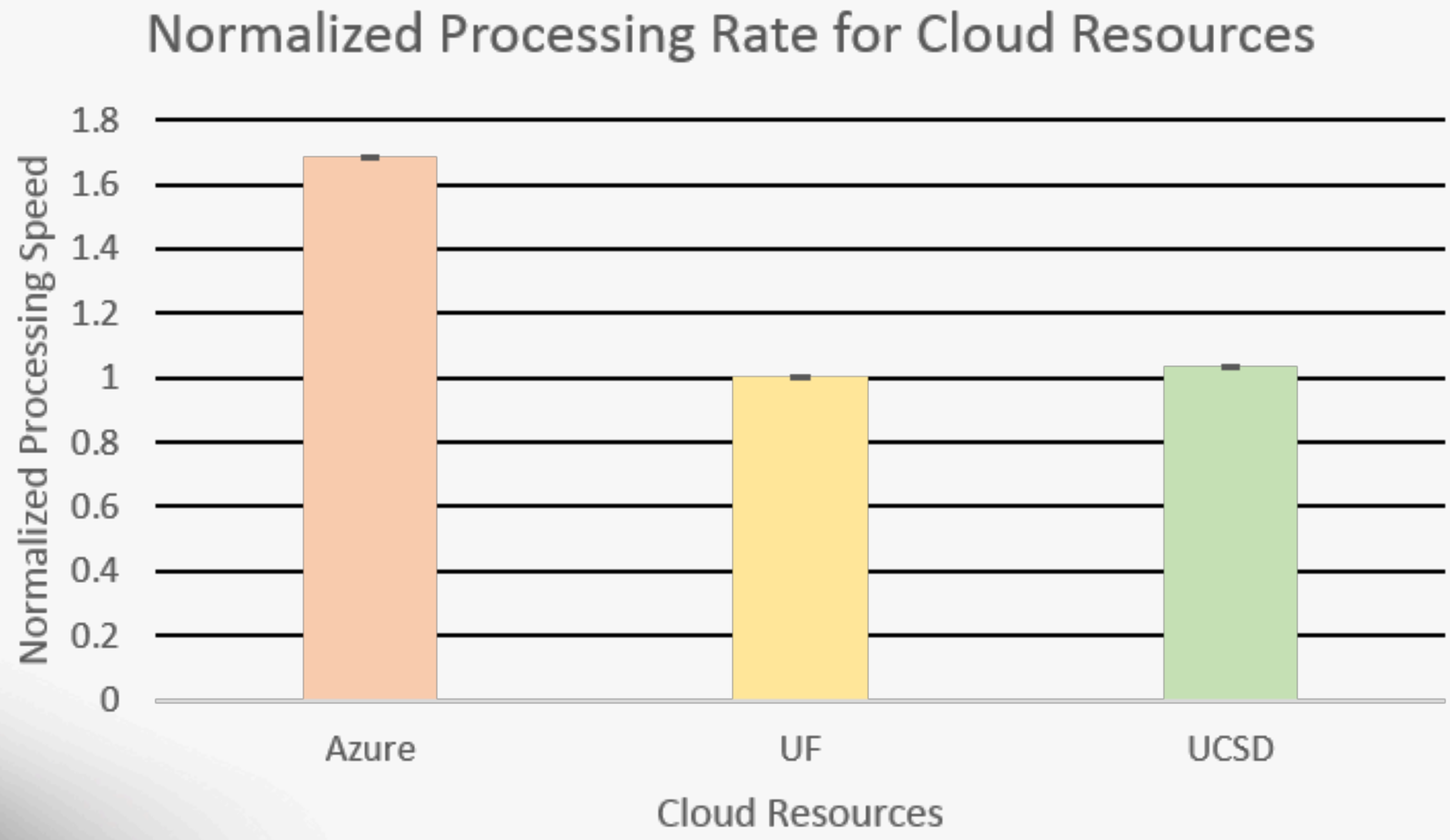
These values were obtained via ViNe overlay monitoring data.

Example Run with 32 Molecules



Results

- Individual Site Performance
  - UF and UCSD process at the same rate, but Azure is noticeably faster



Results

- Multi-site Performance Test
  - Geographic location of Master Node

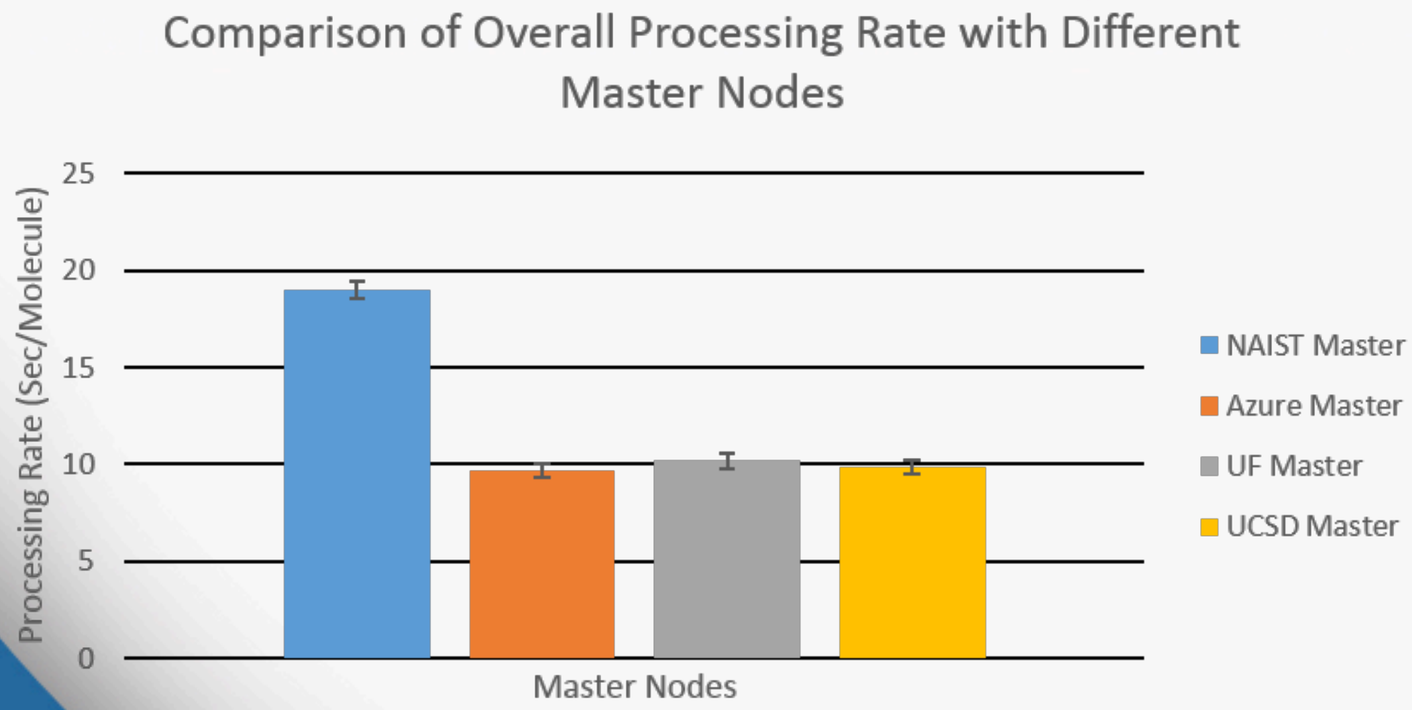


TABLE III
ROUND-TRIP LATENCY (MS)

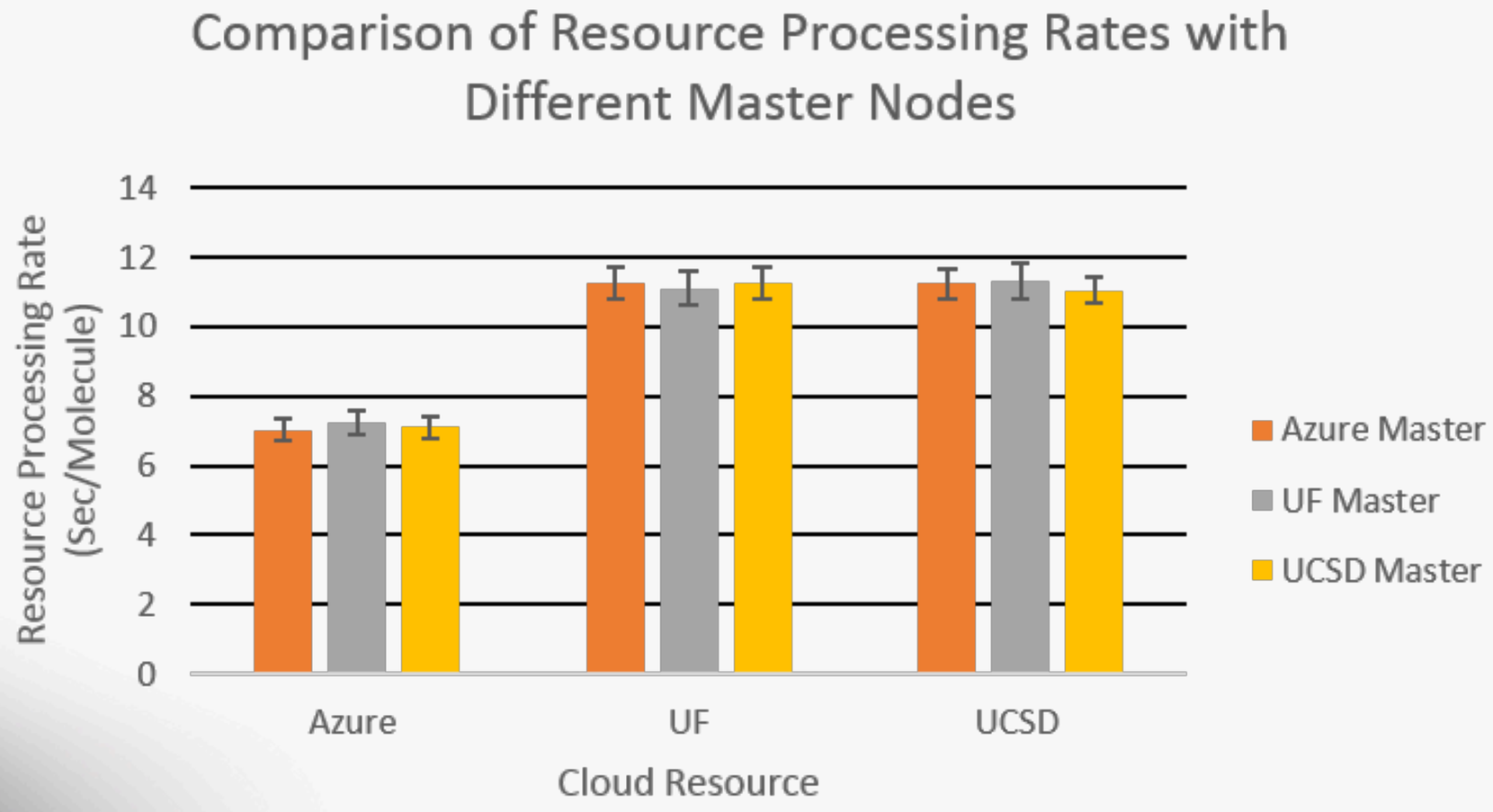
	NAIST	UCSD	Azure	UF
NAIST	-	129	133	190
UCSD	129	-	20	60
Azure	133	20	-	71
UF	190	60	71	-

These values were obtained via ViNe overlay monitoring data.

- United States locations had higher network links with lower latency
- Traveling network links from Japan to the United States resulted in significant latency

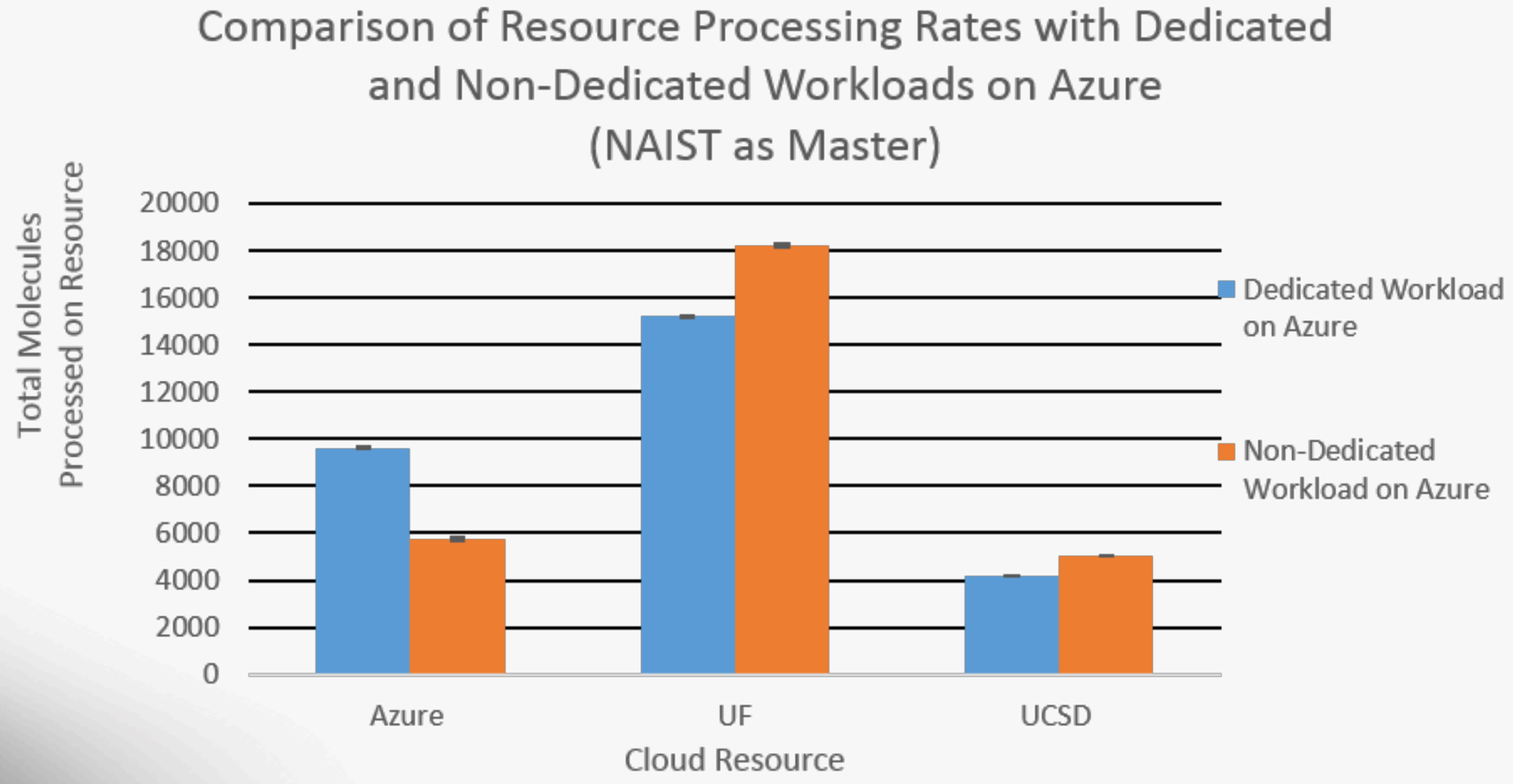
Results

- Multi-site Performance Test
  - Geographic location of Master Node
    - For United States resources, no significant difference in performance of each resource for different Master Nodes



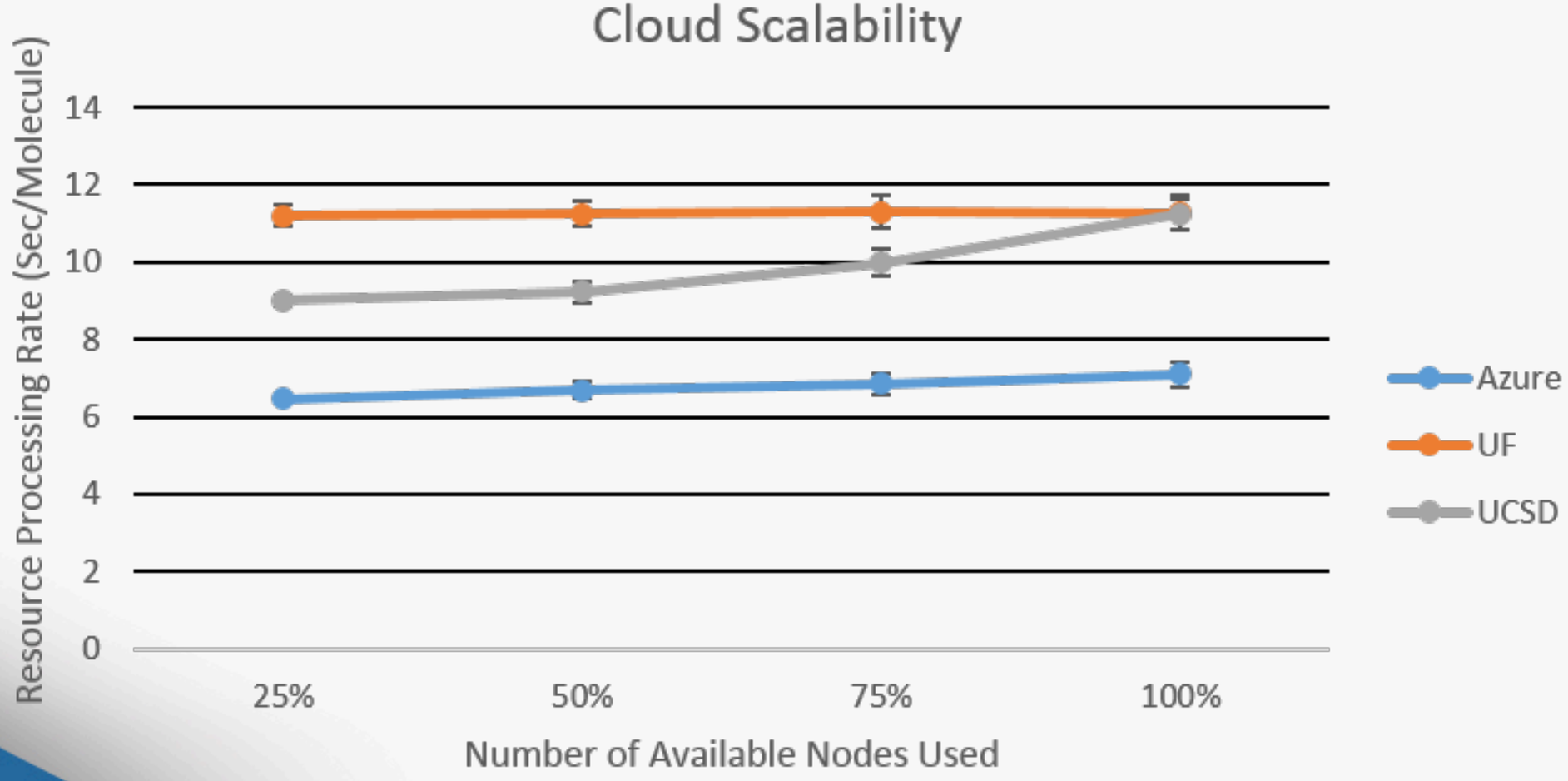
Results

- Multi-site Performance Test
  - Workload Test (Dedicated/Non-Dedicated Resource)
    - When Azure is “Non-Dedicated”, the resources processes less molecules while UF and UCSD process more



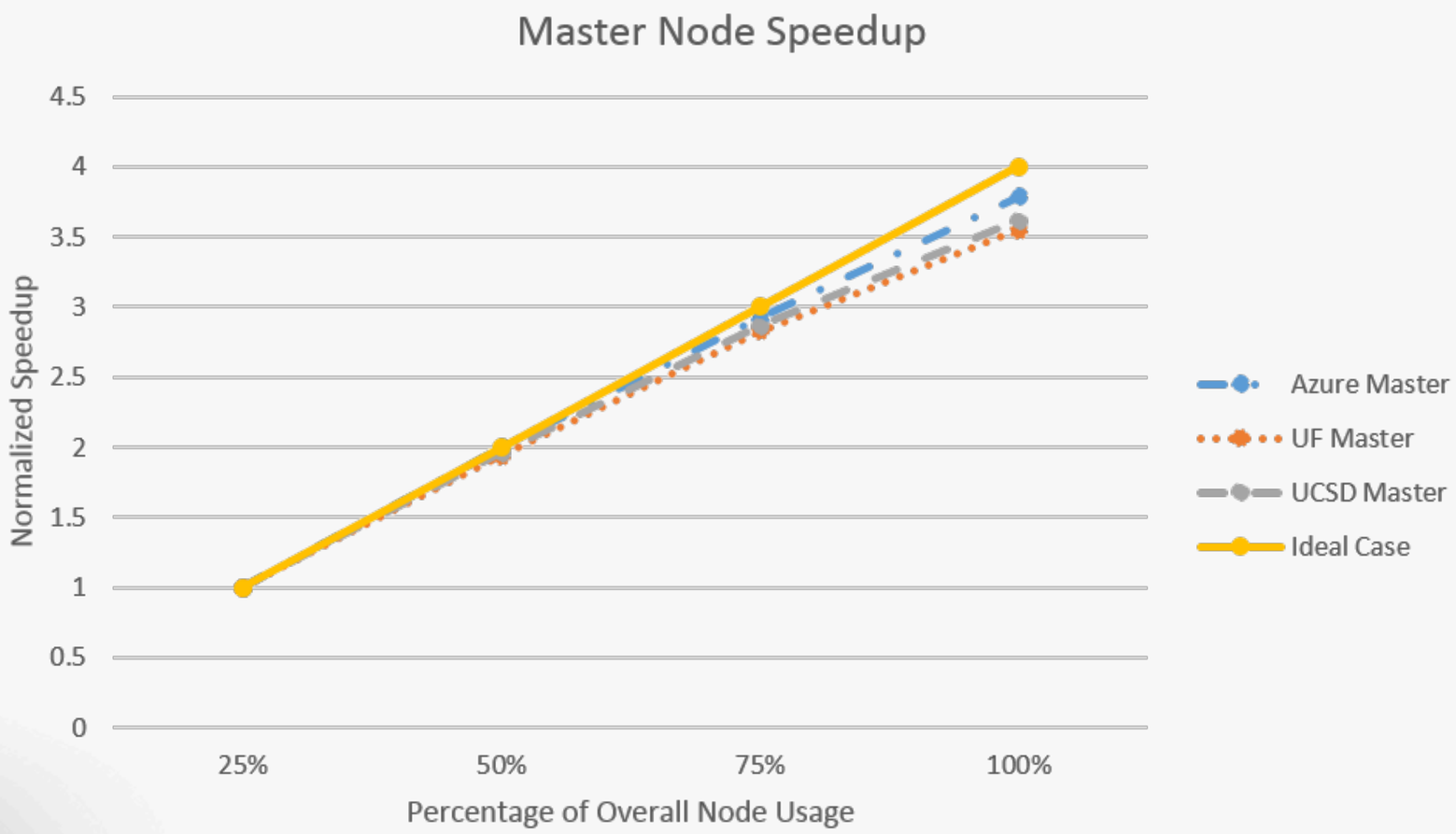
Results

- Multi-site Performance Test
  - Scalability
    - Processing rate maintained as more nodes used except on UCSD



Results

- Multi-site Performance Test
  - Scalability
    - All resources show a scalability trend similar to the ideal scalability case (based on Master Node Output)



Conclusion

- Multi-site cloud environment using private and commercial cloud deployed
- Results show that individual resources have different workloads and performances
  - However, overall environment minimizes the differences
- Alongside mpi, Hadoop as also been incorporated and used for multi-site testing with DOCK