High-Performance Big Data Processing Tools for Neuroscience and A Demo on Chameleon Cloud

Breakout Session at ACNN (Sept. 2016)

Xiaoyi Lu

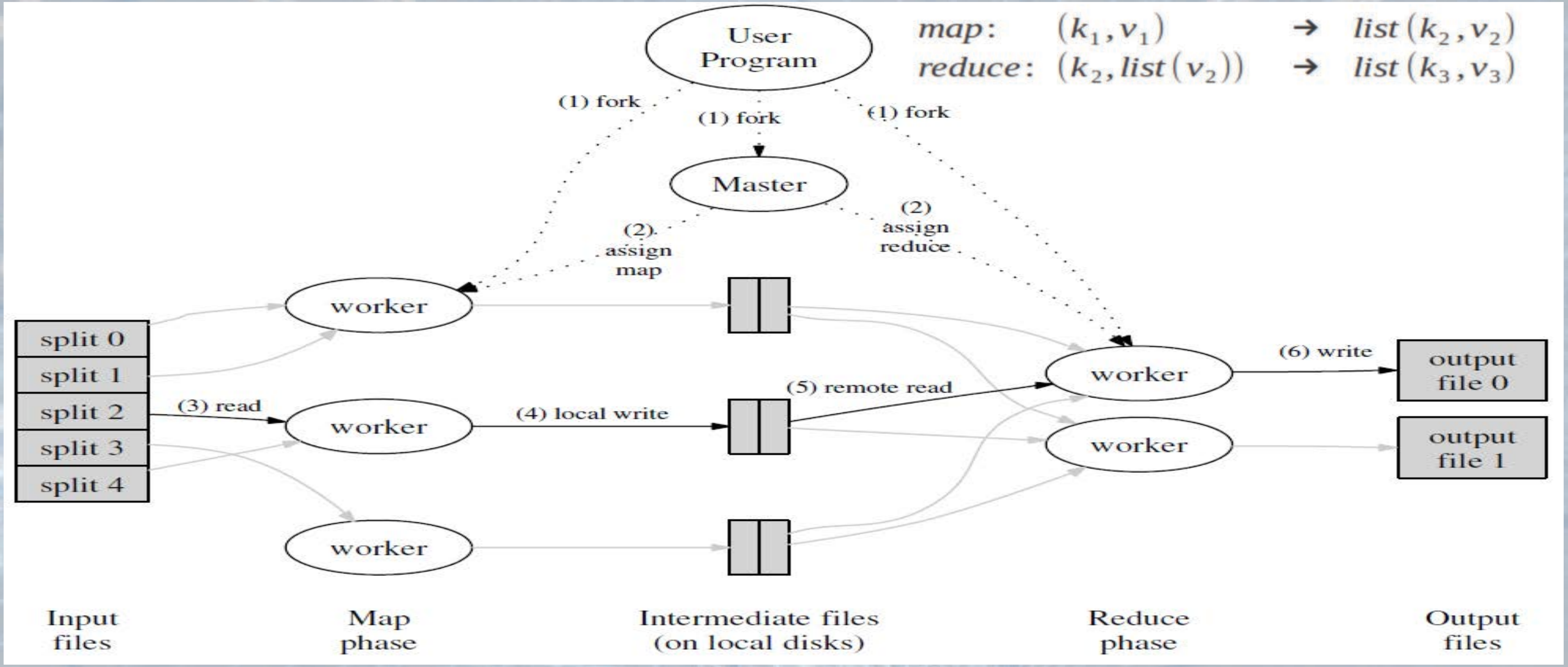
The Ohio State University

luxi@cse.ohio-state.edu

http://web.cse.ohio-state.edu/~luxi/

THE OHIO STATE UNIVERSITY

The MapReduce Model

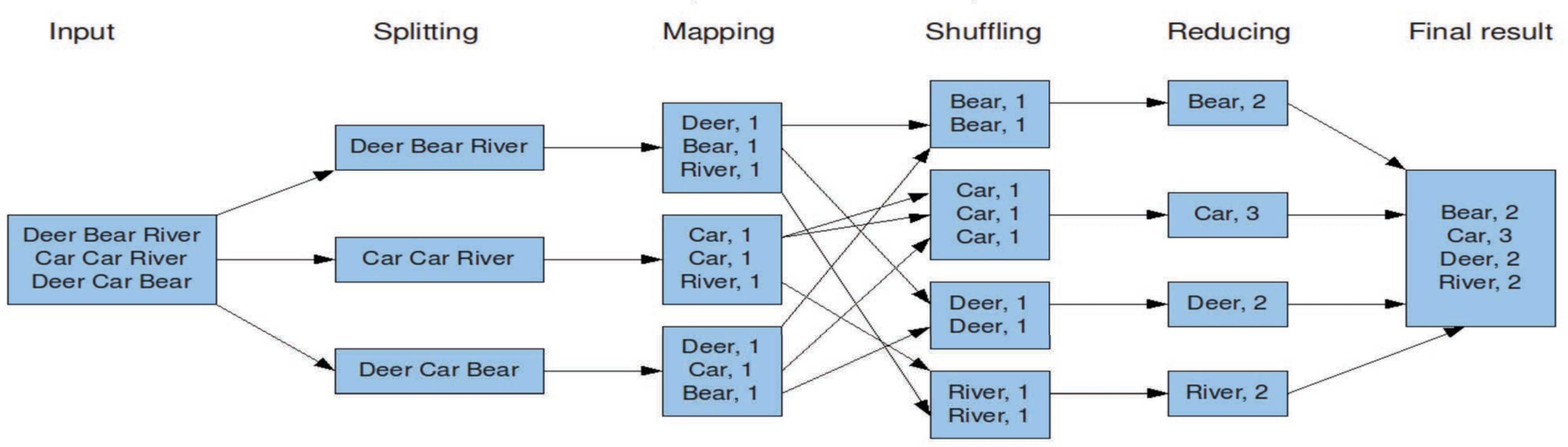


$$map: \quad (k_1, v_1) \rightarrow list(k_2, v_2)$$
$$reduce: \quad (k_2, list(v_2)) \rightarrow list(k_3, v_3)$$

J. Dean and S. Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters*. In *Proceedings of the 6th Symposium on Operating Systems Design & Implementation (OSDI'04)*, 2004.

WordCount Execution

- The overall execution process of WordCount in MapReduce



Word Count in Hadoop!

```
public class WordCount {
 public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(LongWritable key, Text valu   Contex  ontext) throws IOException, InterruptedException {
        String line = value.toString()
        StringTokenizer tokenizer =
        while (tokenizer.hasMoreToke
            word.set(tokenizer.nex        ()
            context.write(word, one);
        }
    }
 }
 public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterator<IntWritable> values, Context context)
      throws IOException, Inte   uptedException {
        in      = 0;
        whil

        }
        context   rit    y,    IntWr   ble(sum));
    }
 }
}
```

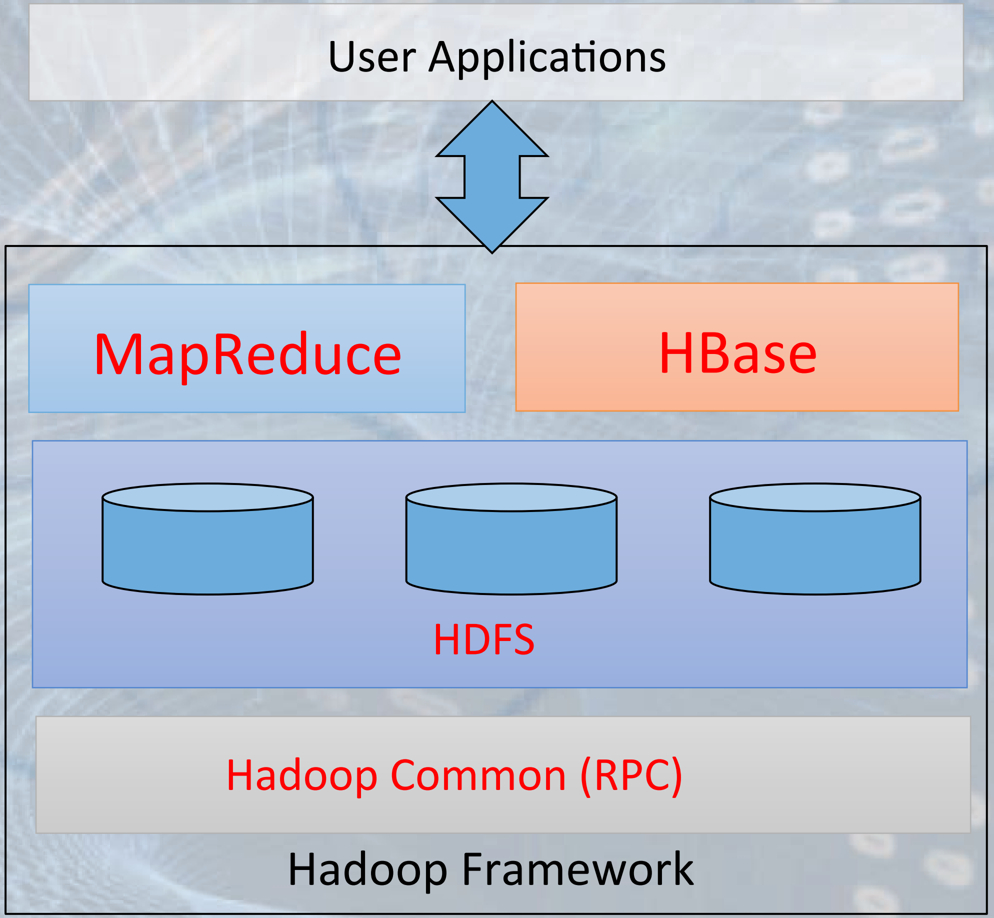
Productive

Scalable

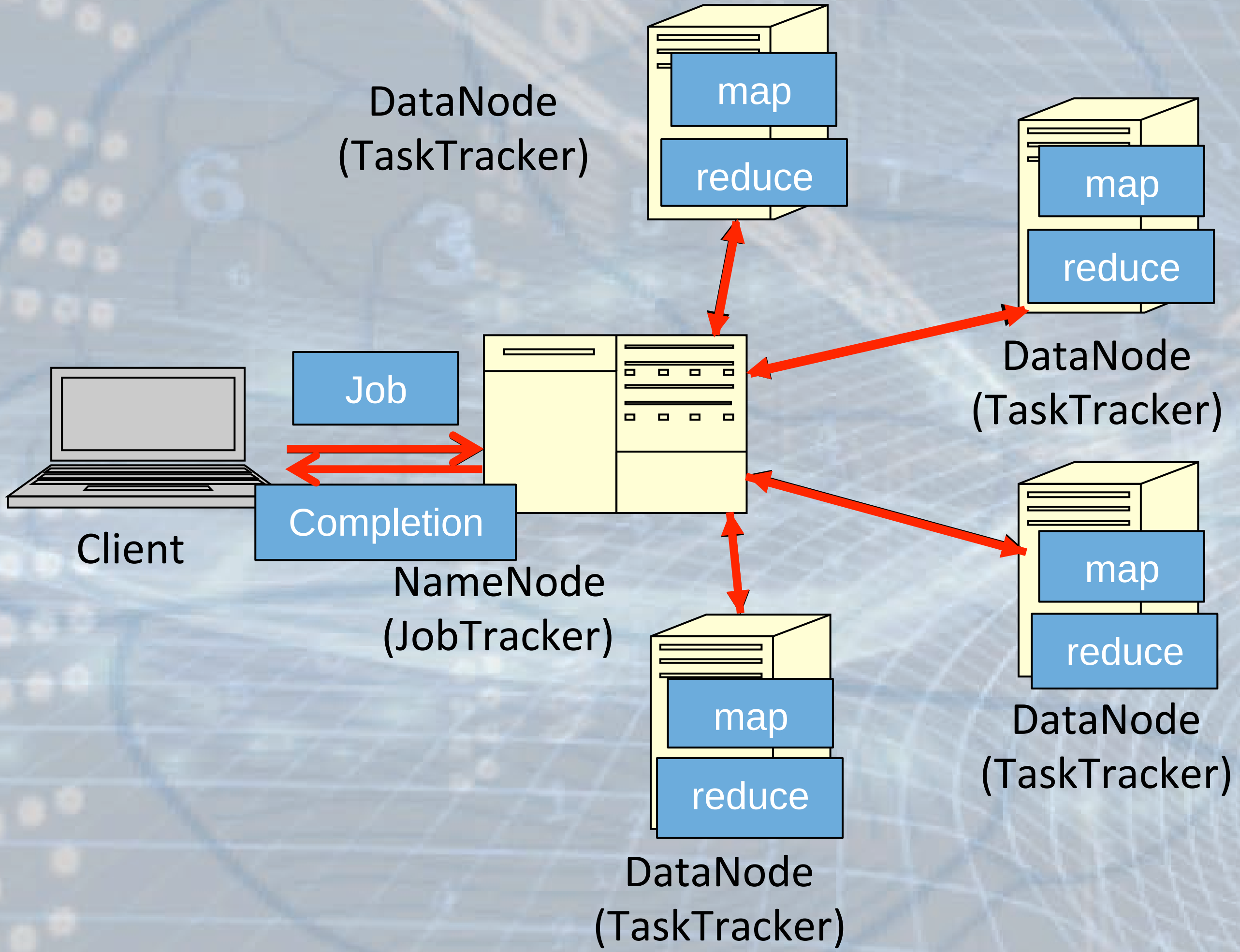
Fault-Tolerant

Big Data Processing with Hadoop Components

- Major components included in this tutorial:
  - MapReduce (Batch)
  - HBase (Query)
  - HDFS (Storage)
  - RPC (Inter-process communication)
- Underlying Hadoop Distributed File System (HDFS) used by both MapReduce and HBase
- Model scales but high amount of communication during intermediate phases can be further optimized

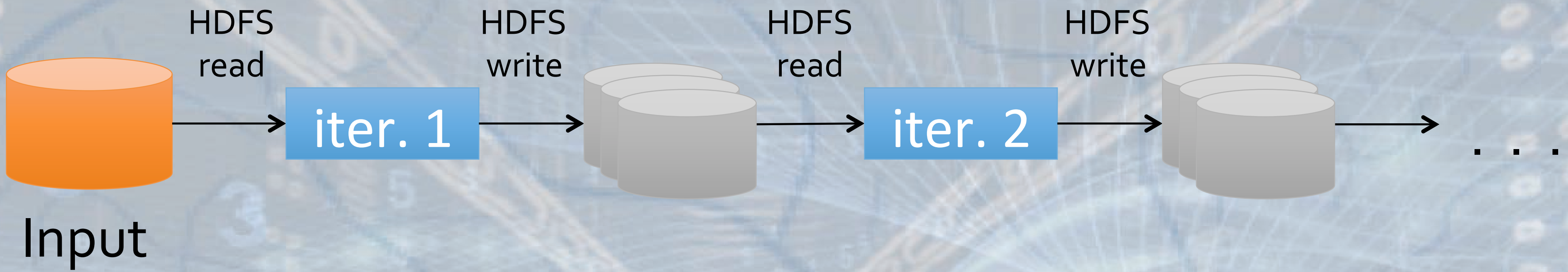


MapReduce Job Execution on Hadoop

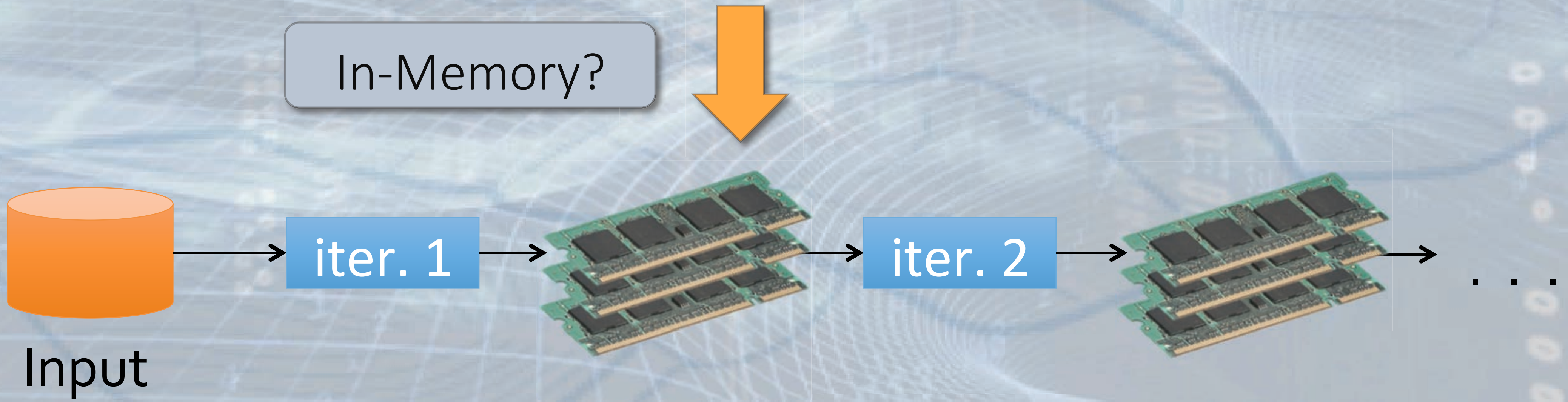


- Main Features
  - Replication (e.g. 3)
  - Data locality for Maps
  - HTTP-based Shuffle
  - Speculative execution
  - Independence among tasks
  - …
- Goals
  - Fault Tolerance
  - Scalability

Data Sharing Problems in MapReduce



Slow due to replication, serialization, and disk IO



10-100× faster than network and disk

RDD Programming Model in Spark

- Key idea: *Resilient Distributed Datasets* ***(RDDs)***
  - Immutable distributed collections of objects that can be cached in memory across cluster nodes
  - Created by transforming data in stable storage using data flow operators (map, filter, groupBy, …)
  - Manipulated through various parallel operators
  - Automatically rebuilt on failure
    - rebuilt if a partition is lost
- Interface
  - Clean language-integrated API in Scala (Python & Java)
  - Support R, distributed machine learning using MLlib
  - Can be used *interactively* from Scala console

RDD Operations

Transformations
(define a new RDD)

map
filter
sample
union
groupByKey
reduceByKey
sortByKey
join
…

Actions
(return a result to driver)

reduce
collect
count
first
Take
countByKey
saveAsTextFile
saveAsSequenceFile
…

More Information:

- https://spark.apache.org/docs/latest/programming-guide.html#transformations
- https://spark.apache.org/docs/latest/programming-guide.html#actions

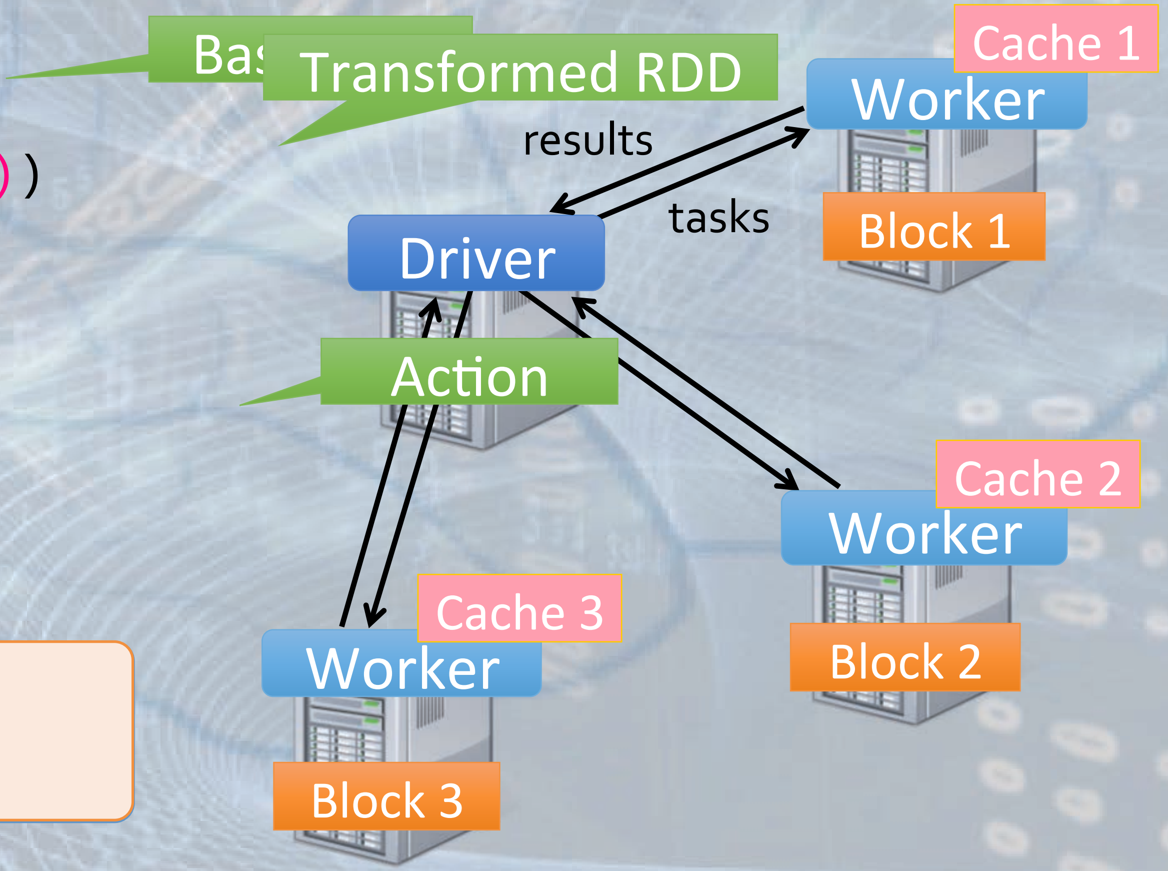
Example: Log Mining

Load error messages from a log into memory, then interactively search for various patterns

```
lines = spark.textFile("hdfs://...")
errors = lines.filter(_.startsWith("ERROR"))
messages = errors.map(_.split('\t')(2))
cachedMsgs = messages.cache()

cachedMsgs.filter(_.contains("foo")).count
cachedMsgs.filter(_.contains("bar")).count
. . .
```

Result: scaled to 1 TB data in 5-7 sec
(vs 170 sec for on-disk data)



Courtesy: https://spark.apache.org/

Example: Word Count in Spark!

```
val fi           extFile("hdfs://   )

val cou       le.flatMap(line =>           " "))
                  .map(word => (word, 1))
                  .reduceByKey(_ + _)

counts.saveAsT xtFile("hdfs://...")
```

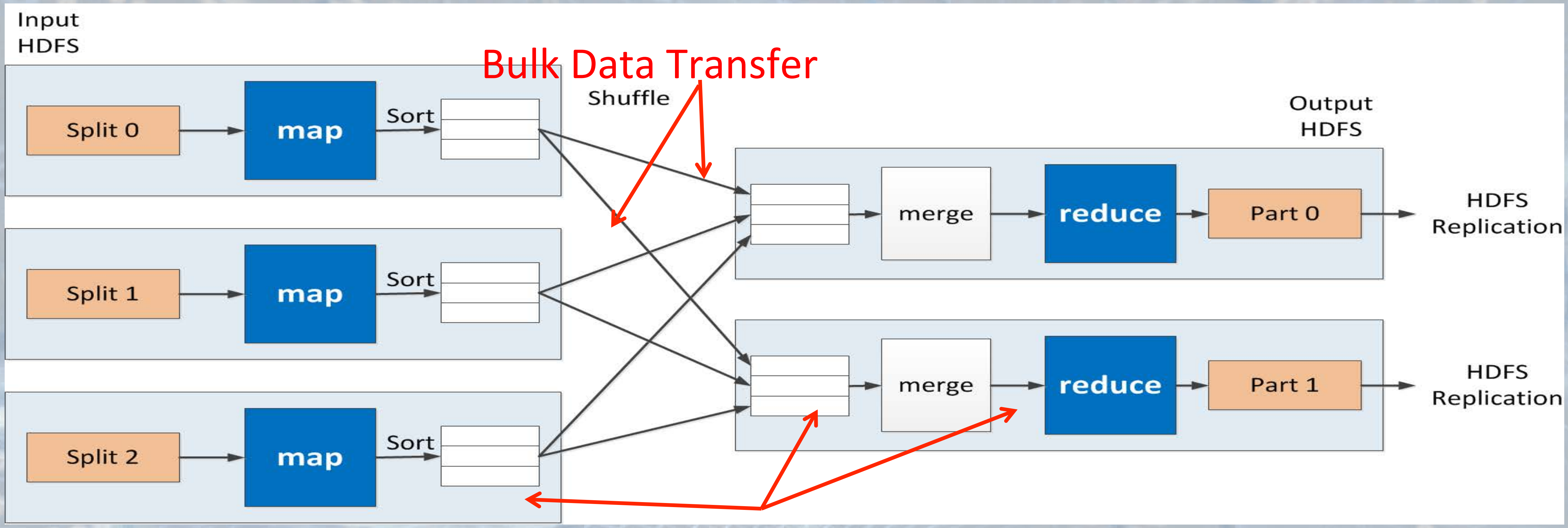
Productive

High-Performance

Scalable

Fault-Tolerant

Data Movement in MapReduce



Bulk Data Transfer

Disk Operations

- Map and Reduce Tasks carry out the total job execution
  - Map tasks read from HDFS, operate on it, and write the intermediate data to local disk
  - Reduce tasks get these data by shuffle from TaskTrackers, operate on it and write to HDFS
- Communication in shuffle phase uses HTTP over Java Sockets
- Similar bottlenecks exist in HDFS, RPC, Spark, HBase, Memcached, ec.

The High-Performance Big Data (HiBD) Project

- **http://hibd.cse.ohio-state.edu**
- RDMA for Apache Spark
- RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)
  - Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions
- RDMA for Apache HBase
- RDMA for Memcached (RDMA-Memcached)
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- OSU HiBD-Benchmarks (OHB)
  - HDFS, Memcached, and HBase Micro-benchmarks
- Users Base: 190 organizations from 26 countries
- More than 17,800 downloads from the project site
- RDMA for Impala (upcoming)

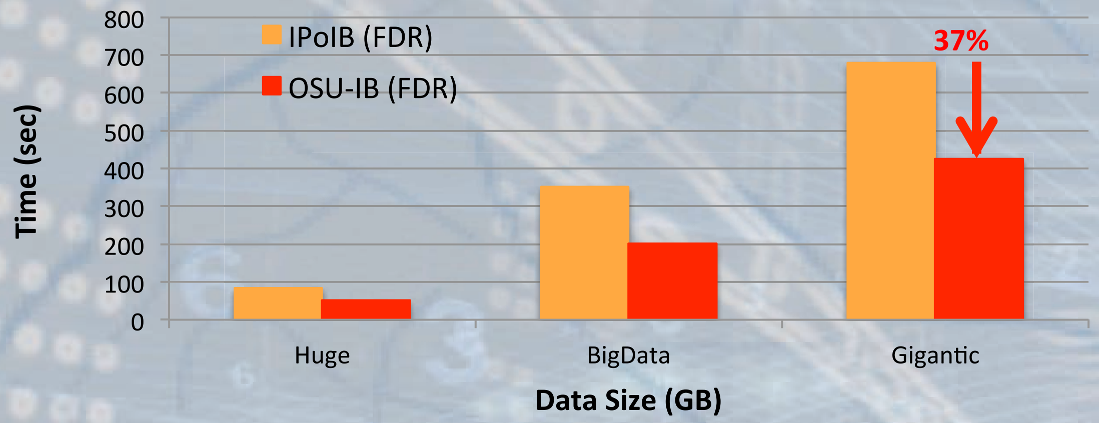
Available for InfiniBand and RoCE

HiBD
High-Performance
Big Data

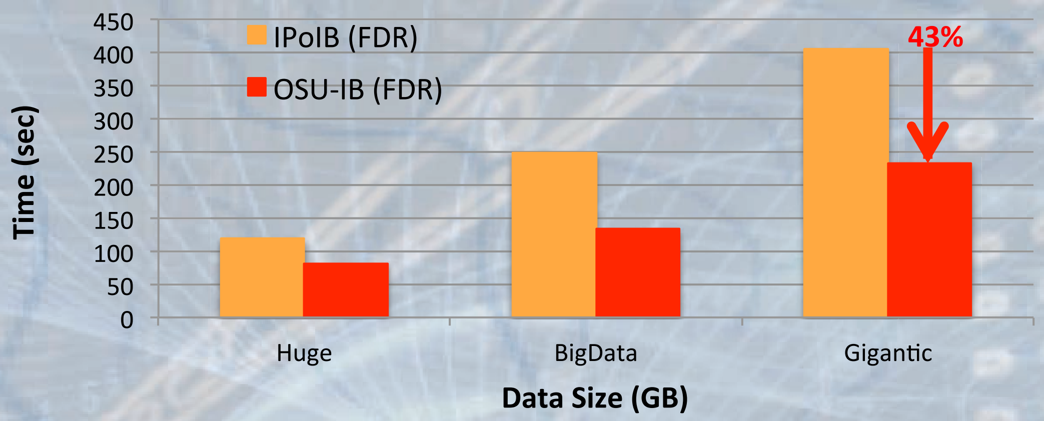
Network Based Computing
Laboratory

THE OHIO STATE
UNIVERSITY

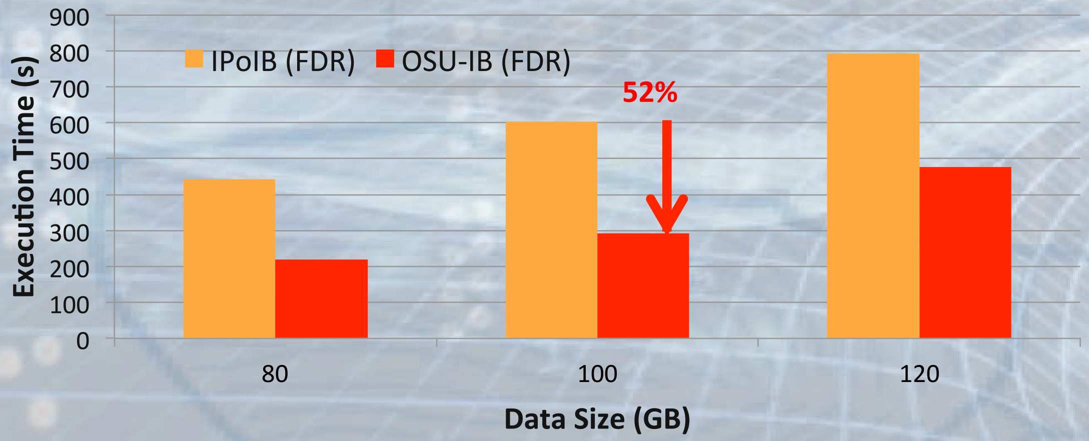
Performance Benefits of HiBD on PageRank, Sort, and TeraSort



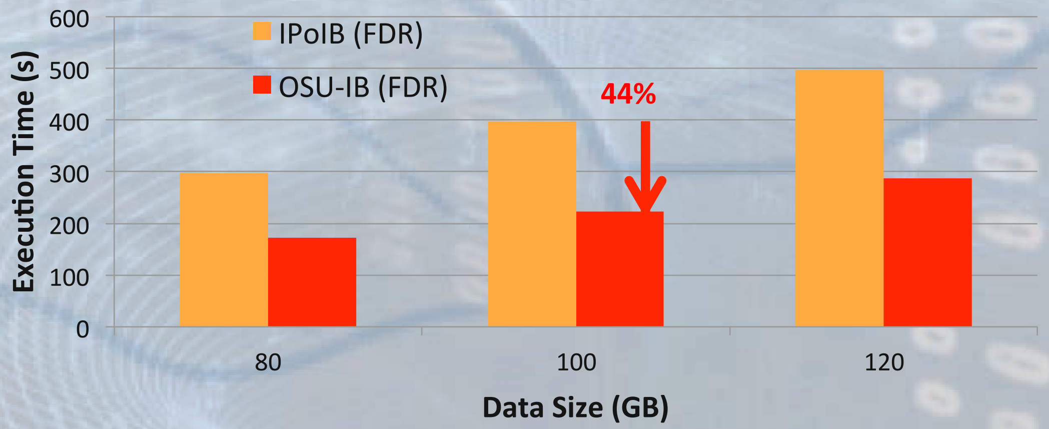
SDSC Comet 32 Worker Nodes, 768 cores
PageRank Total Time



SDSC Comet 64 Worker Nodes, 1536 cores
PageRank Total Time



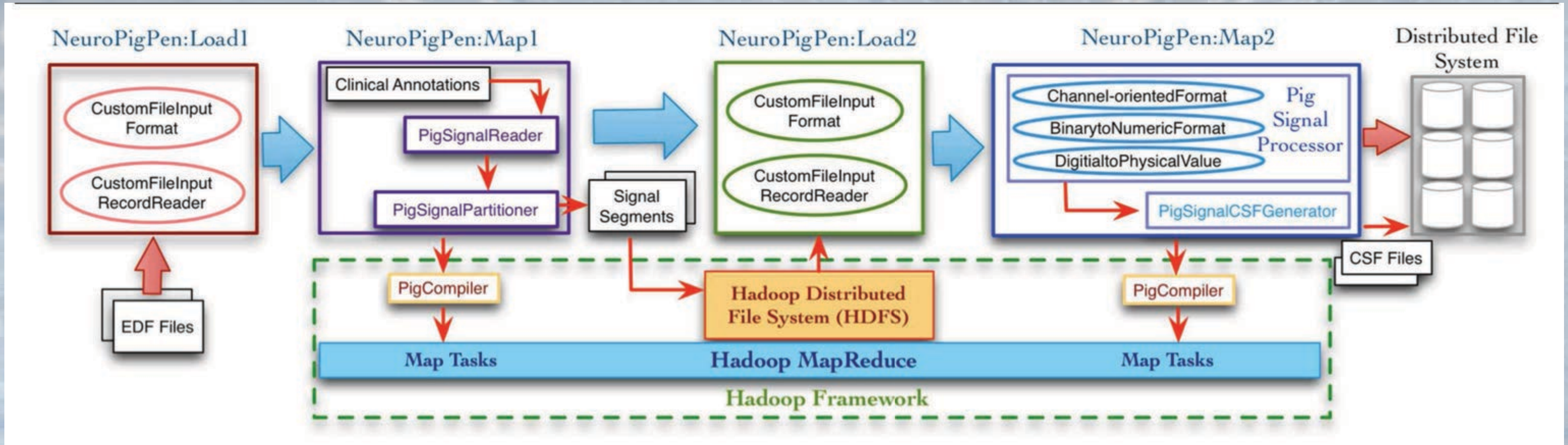
TACC Stampede 32 Worker Nodes,
Sort Total Time



TACC Stampede 32 Worker Nodes,
TeraSort Total Time

A Case Study with NeuroPigPen

- **NeuroPigPen Toolkit[1]:** a data management toolkit using Hadoop Pig for processing electrophysiological signals in Neuroscience applications
  - Dr. Sahoo, etc., Case Western Reserve University



The data processing workflow supported by the NeuroPigPen toolkit modules consists of multiple steps with EDF files as input and CSF files as output. The Load functions in the toolkit extend the Hadoop FileInputFormat and FileInputRecordReader classes to support signal data. The Map functions in the toolkit are automatically compiled into MapReduce tasks by the Apache Pig compiler. The intermediate and final results are stored in Hadoop Distributed File System (HDFS), which provides a reliable and scalable storage platform for signal data.

[1] Sahoo SS, Wei A, Valdez J, Wang L, Zonjy B, Tatsuoka C, Loparo KA, Lhatoo SD. NeuroPigPen: a Data Management Toolkit using Hadoop Pig for Processing Electrophysiological Signals in Neuroscience Applications, http://www.ncbi.nlm.nih.gov/pubmed/27375472

Demo on NSF Supported Chameleon Cloud

- Five InfiniBand bare-metal nodes (with SR-IOV support) on Chameleon Cloud
- 25 VMs in total, 6 VMs per node
- RDMA-Hadoop appliance available on Chameleon Cloud
  - Pre-installed OS, drivers, and software packages
  - Automatically configuration
- Deployment of virtual machines
- RDMA-Hadoop examples
  - RandomWriter
  - Sort

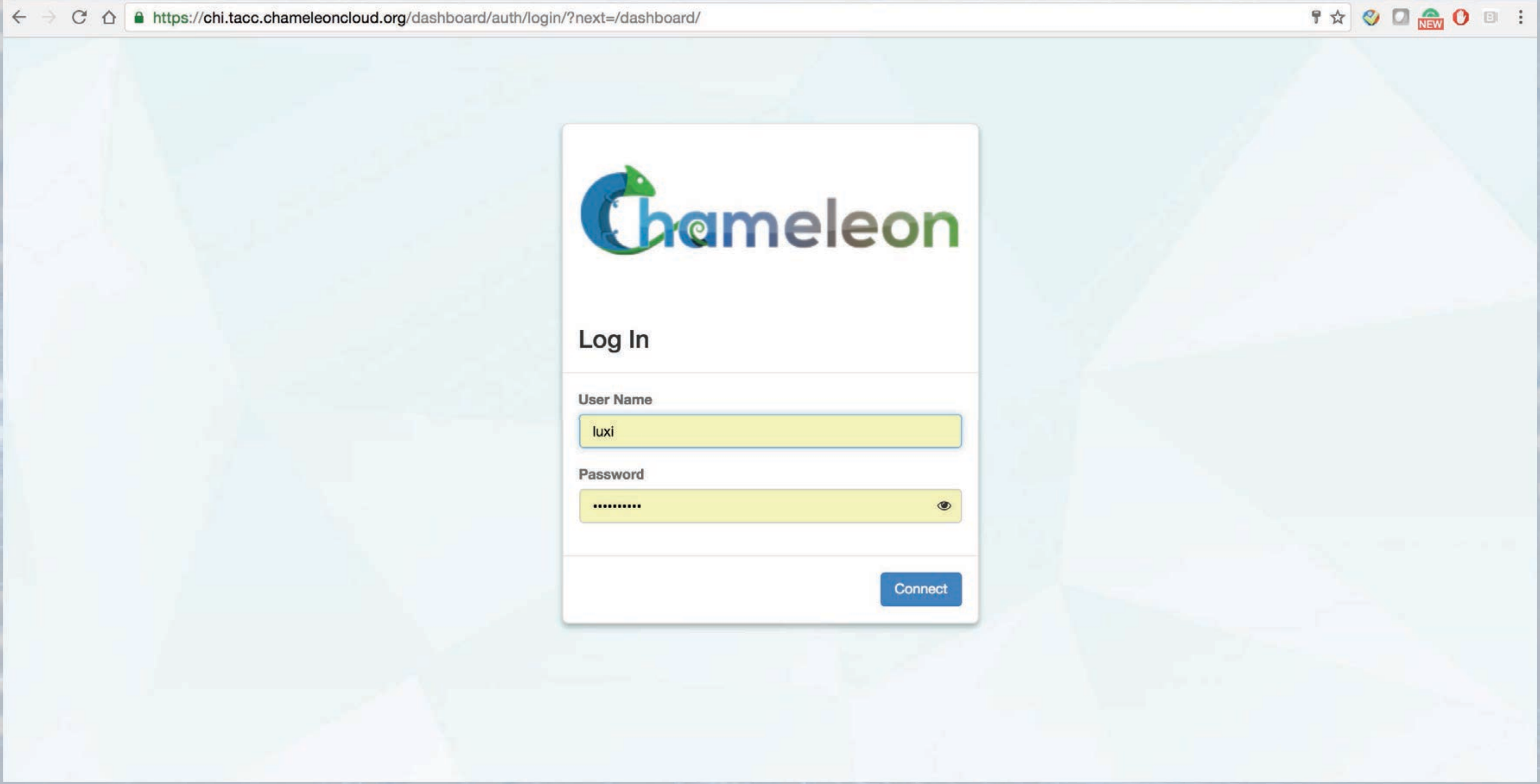
NSF Chameleon Cloud: A Powerful and Flexible Experimental Instrument

- Large-scale instrument
  - Targeting Big Data, Big Compute, Big Instrument research
  - ~650 nodes (~14,500 cores), 5 PB disk over two sites, 2 sites connected with 100G network
- Reconfigurable instrument
  - Bare metal reconfiguration, operated as single instrument, graduated approach for ease-of-use
- Connected instrument
  - Workload and Trace Archive
  - Partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Partnerships with users
- Complementary instrument
  - Complementing GENI, Grid'5000, and other testbeds
- Sustainable instrument
  - Industry connections

http://www.chameleoncloud.org/

Computation Institute TACC iCAIR The Ohio State University UTSA

Demo - Login to Chameleon



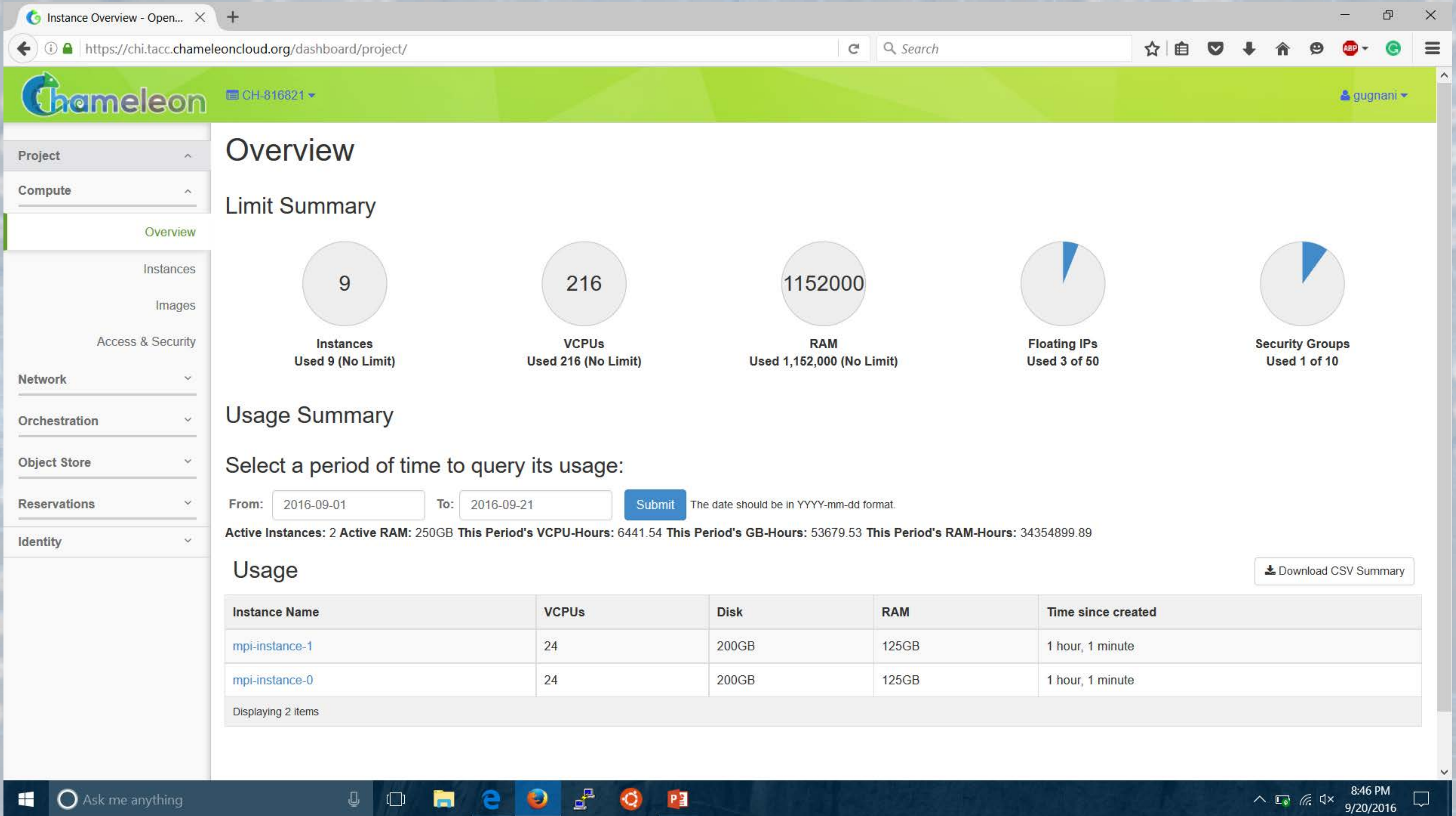
https://chi.tacc.chameleoncloud.org/dashboard/auth/login/?next=/dashboard/

Log In

User Name: luxi

Password: ••••••••••

Demo – Overview of Your Chameleon Resources



Overview

Limit Summary

9	216	1152000		
Instances Used 9 (No Limit)	VCPUs Used 216 (No Limit)	RAM Used 1,152,000 (No Limit)	Floating IPs Used 3 of 50	Security Groups Used 1 of 10

Usage Summary

Select a period of time to query its usage:

From: 2016-09-01 To: 2016-09-21 Submit The date should be in YYYY-mm-dd format.

Active Instances: 2 **Active RAM:** 250GB **This Period's VCPU-Hours:** 6441.54 **This Period's GB-Hours:** 53679.53 **This Period's RAM-Hours:** 34354899.89

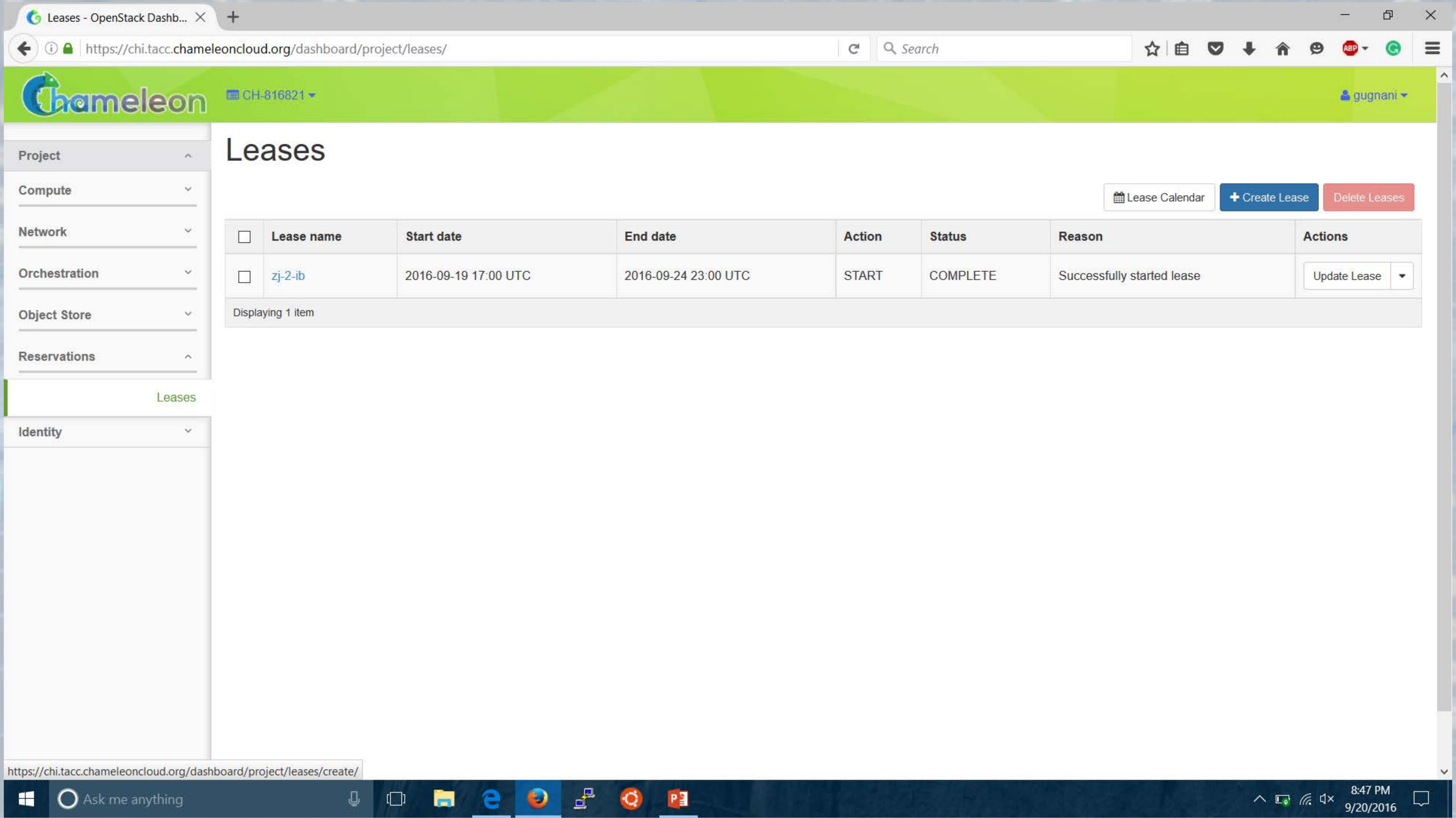
Usage

Download CSV Summary

Instance Name	VCPUs	Disk	RAM	Time since created
mpi-instance-1	24	200GB	125GB	1 hour, 1 minute
mpi-instance-0	24	200GB	125GB	1 hour, 1 minute

Displaying 2 items

Demo – Create Your Lease

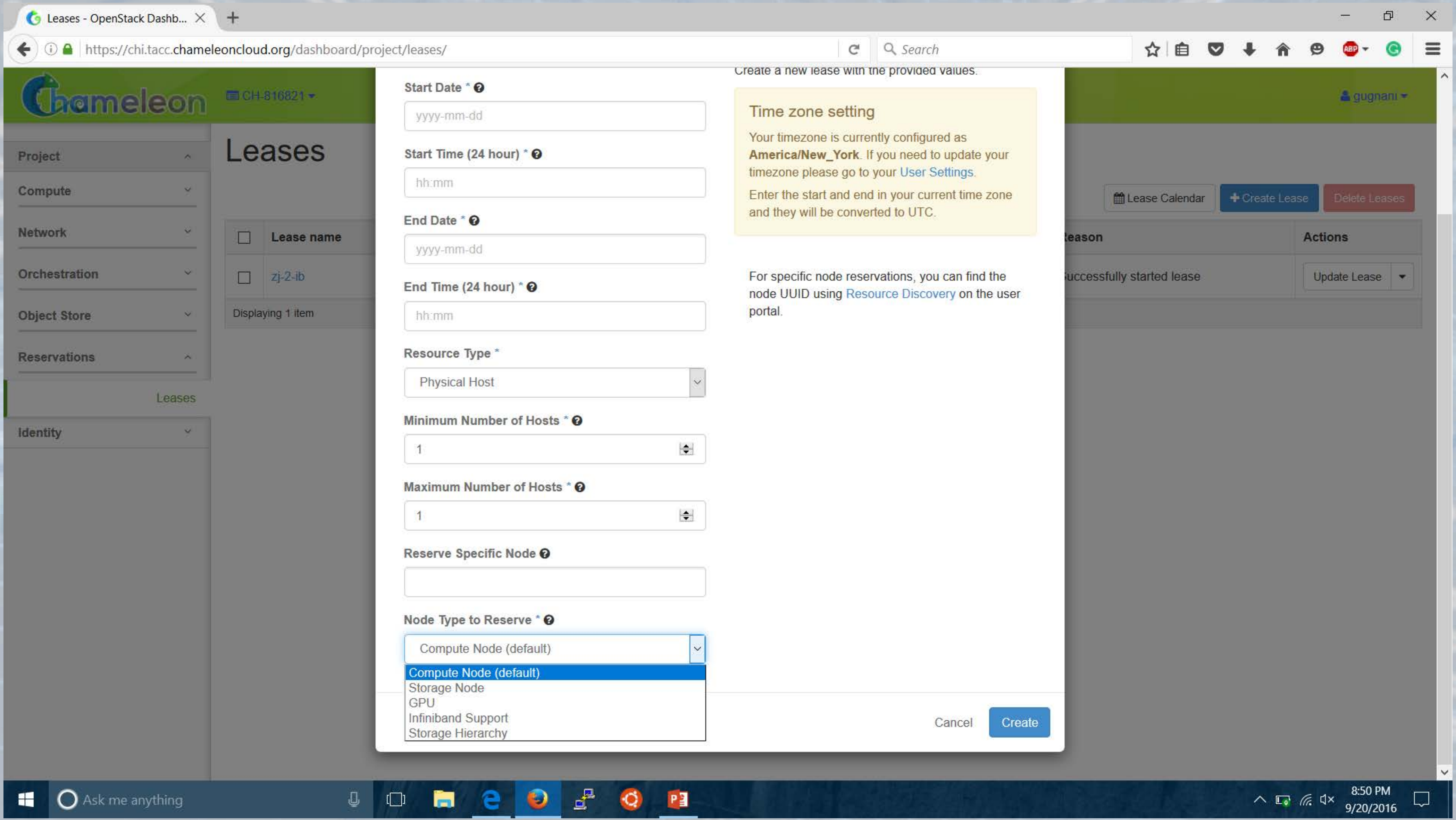


Leases

□	Lease name	Start date	End date	Action	Status	Reason	Actions
□	zj-2-ib	2016-09-19 17:00 UTC	2016-09-24 23:00 UTC	START	COMPLETE	Successfully started lease	Update Lease

Displaying 1 item

Demo – Create Your Lease



Create a new lease with the provided values.

Start Date * yyyy-mm-dd

Start Time (24 hour) * hh:mm

End Date * yyyy-mm-dd

End Time (24 hour) * hh:mm

Resource Type * Physical Host

Minimum Number of Hosts * 1

Maximum Number of Hosts * 1

Reserve Specific Node

Node Type to Reserve * Compute Node (default)

- Compute Node (default)
- Storage Node
- GPU
- Infiniband Support
- Storage Hierarchy

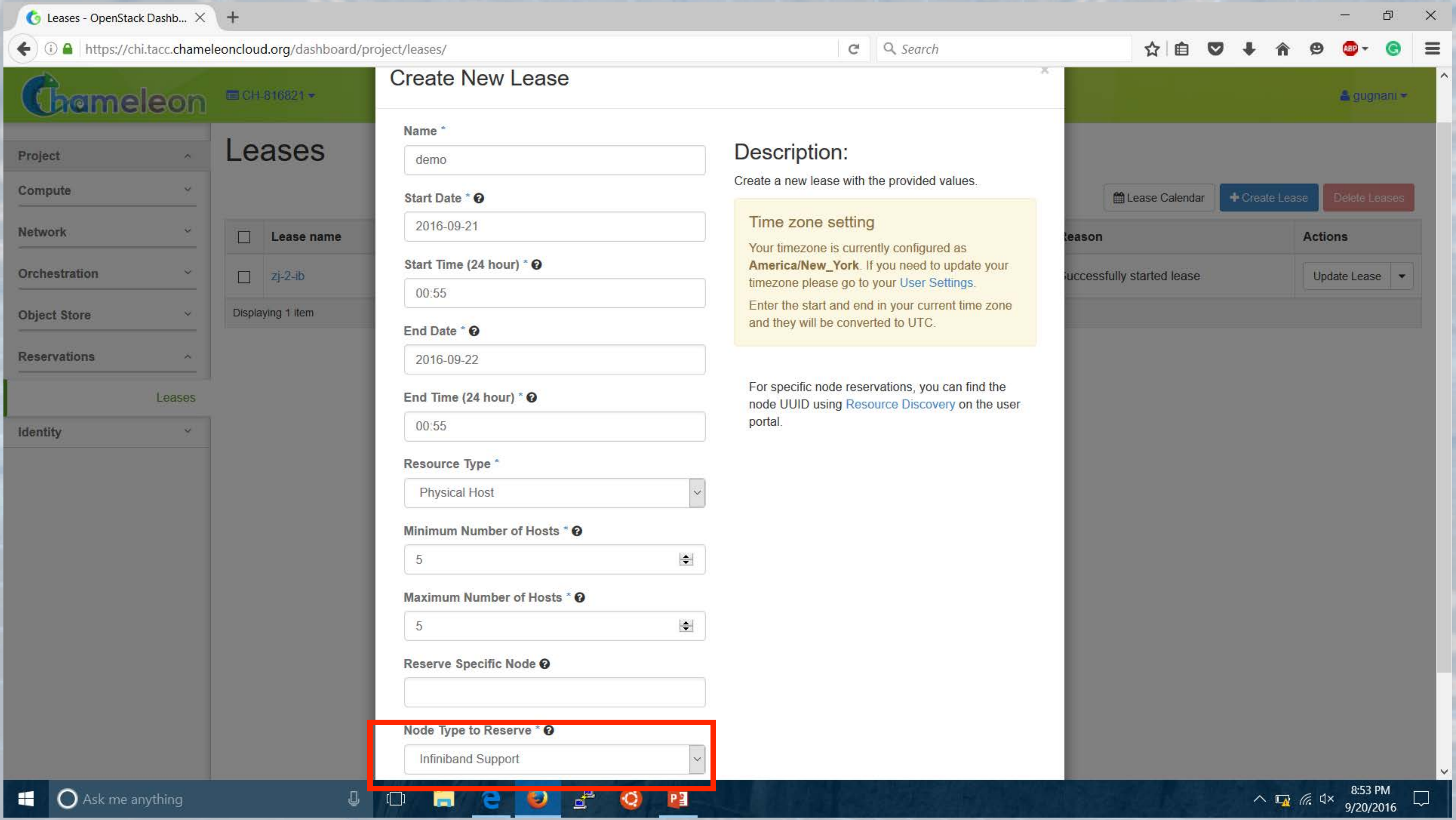
Time zone setting

Your timezone is currently configured as **America/New_York**. If you need to update your timezone please go to your User Settings.

Enter the start and end in your current time zone and they will be converted to UTC.

For specific node reservations, you can find the node UUID using Resource Discovery on the user portal.

Demo – Create Your Lease



Create New Lease

Name * demo

Start Date * 2016-09-21

Start Time (24 hour) * 00:55

End Date * 2016-09-22

End Time (24 hour) * 00:55

Resource Type * Physical Host

Minimum Number of Hosts * 5

Maximum Number of Hosts * 5

Reserve Specific Node

Node Type to Reserve * Infiniband Support

Description:

Create a new lease with the provided values.

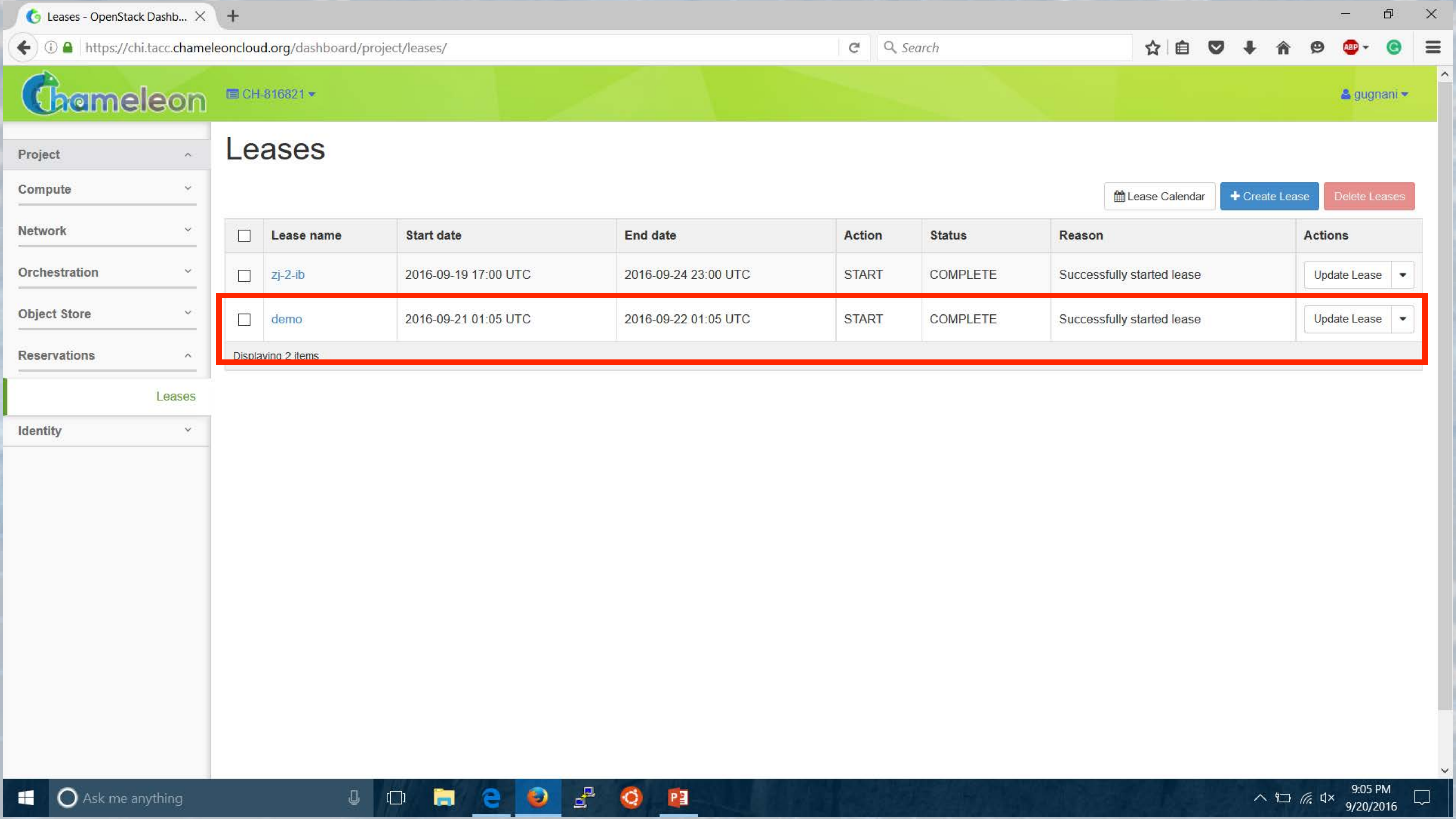
Time zone setting

Your timezone is currently configured as **America/New_York**. If you need to update your timezone please go to your User Settings.

Enter the start and end in your current time zone and they will be converted to UTC.

For specific node reservations, you can find the node UUID using Resource Discovery on the user portal.

Demo – Create Your Lease

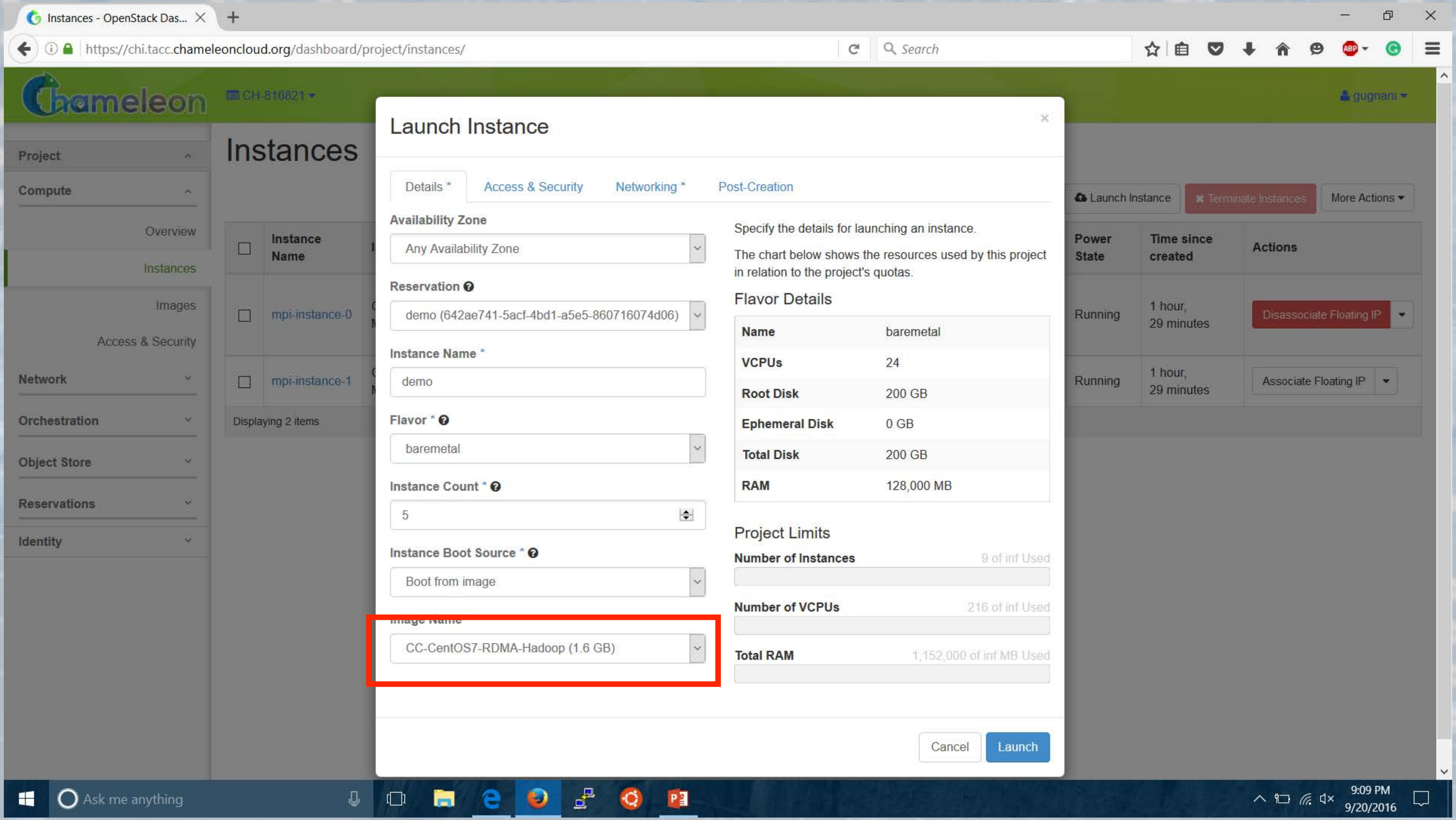


Leases

□	Lease name	Start date	End date	Action	Status	Reason	Actions
□	zj-2-ib	2016-09-19 17:00 UTC	2016-09-24 23:00 UTC	START	COMPLETE	Successfully started lease	Update Lease
□	demo	2016-09-21 01:05 UTC	2016-09-22 01:05 UTC	START	COMPLETE	Successfully started lease	Update Lease

Displaying 2 items

Demo – Launch Your Instances



Instances

Instance Name	Power State	Time since created
mpi-instance-0	Running	1 hour, 29 minutes
mpi-instance-1	Running	1 hour, 29 minutes

Displaying 2 items

Launch Instance

Availability Zone: Any Availability Zone

Reservation: demo (642ae741-5acf-4bd1-a5e5-860716074d06)

Instance Name *: demo

Flavor *: baremetal

Instance Count *: 5

Instance Boot Source *: Boot from image

Image Name: CC-CentOS7-RDMA-Hadoop (1.6 GB)

Specify the details for launching an instance.

The chart below shows the resources used by this project in relation to the project's quotas.

Flavor Details

Name	baremetal
VCPUs	24
Root Disk	200 GB
Ephemeral Disk	0 GB
Total Disk	200 GB
RAM	128,000 MB

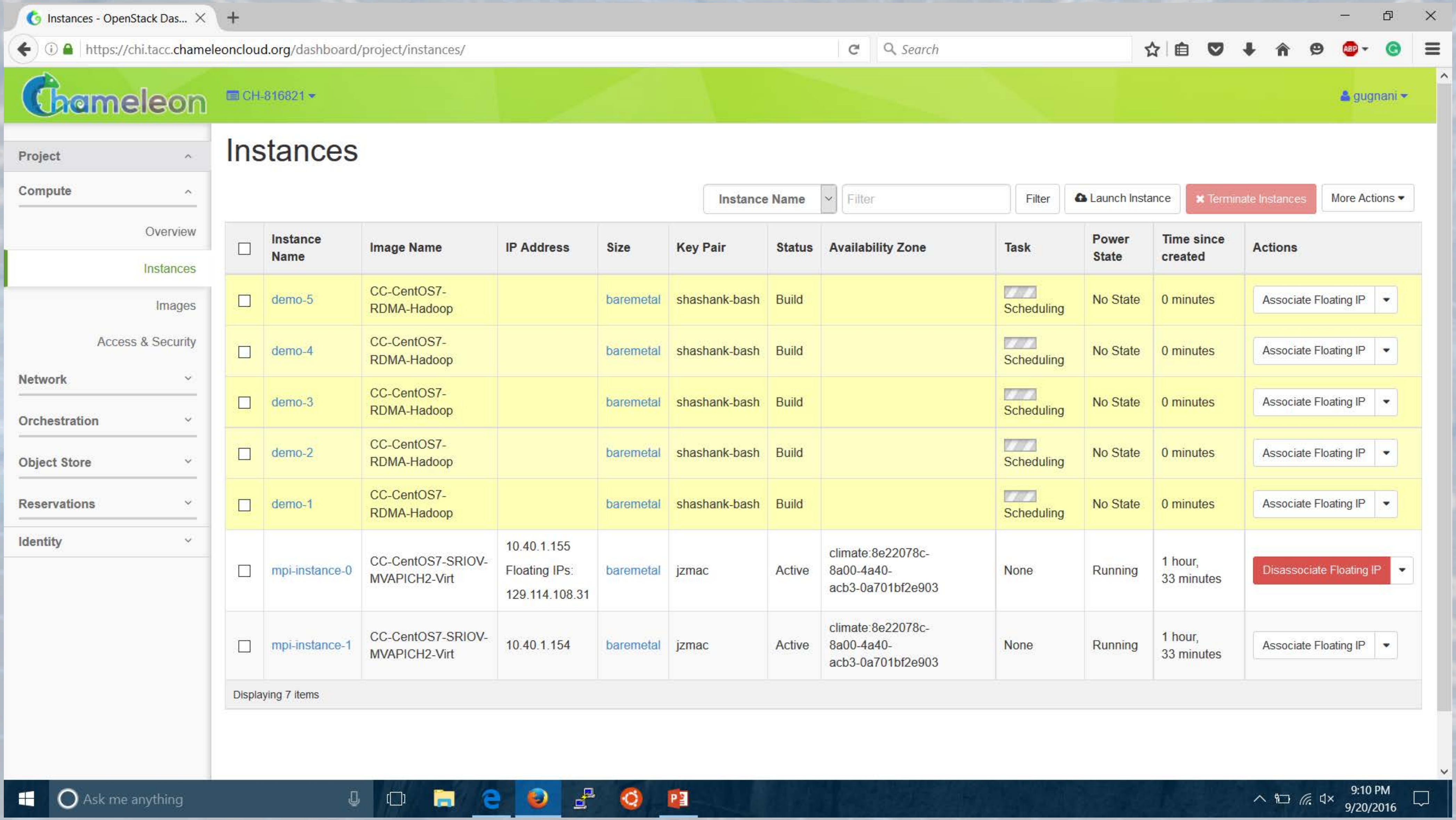
Project Limits

Number of Instances — 9 of inf Used

Number of VCPUs — 216 of inf Used

Total RAM — 1,152,000 of inf MB Used

Demo – Launch Your Instances

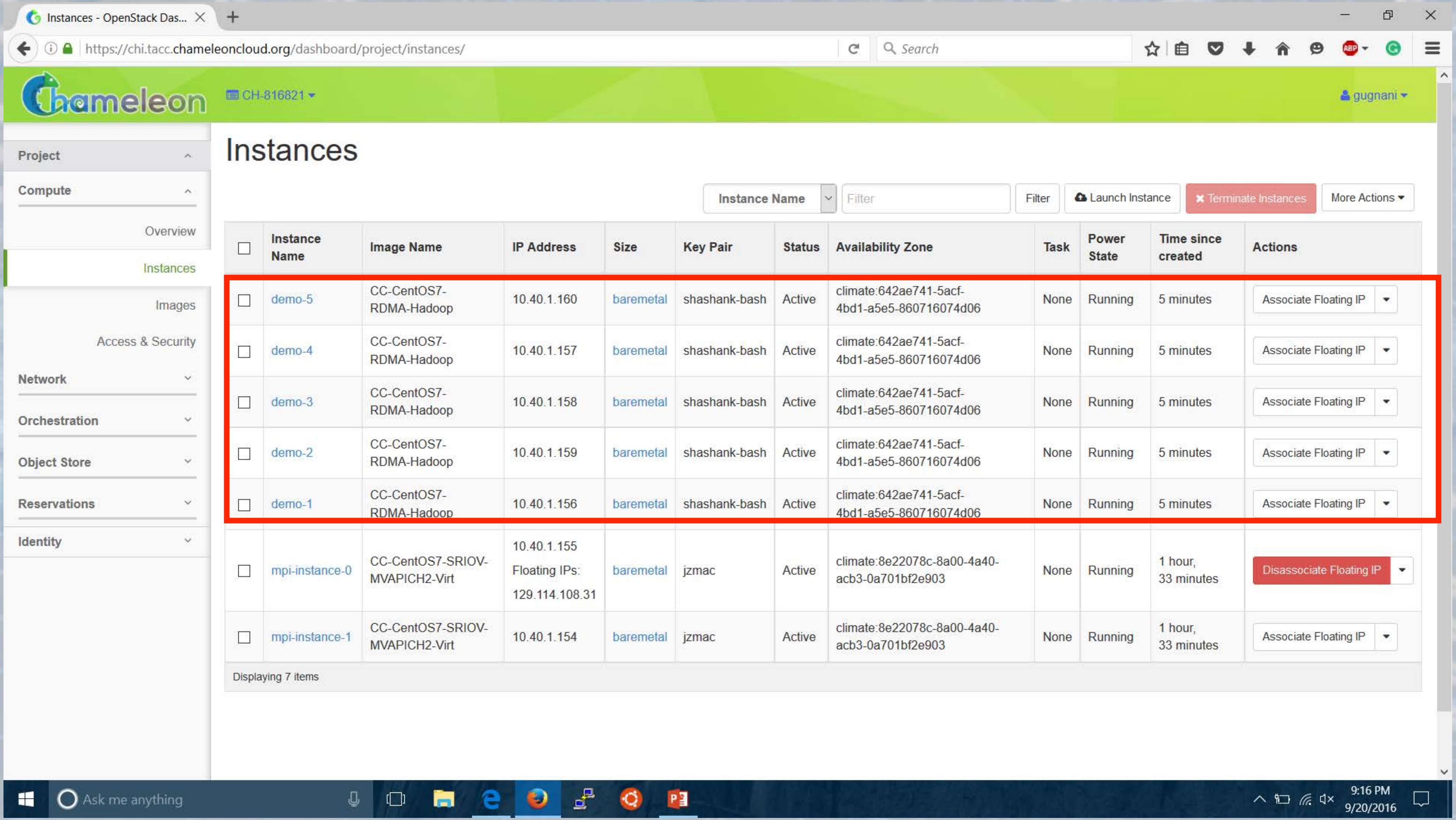


Instances

☐	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Time since created	Actions
☐	demo-5	CC-CentOS7-RDMA-Hadoop		baremetal	shashank-bash	Build		Scheduling	No State	0 minutes	Associate Floating IP ▾
☐	demo-4	CC-CentOS7-RDMA-Hadoop		baremetal	shashank-bash	Build		Scheduling	No State	0 minutes	Associate Floating IP ▾
☐	demo-3	CC-CentOS7-RDMA-Hadoop		baremetal	shashank-bash	Build		Scheduling	No State	0 minutes	Associate Floating IP ▾
☐	demo-2	CC-CentOS7-RDMA-Hadoop		baremetal	shashank-bash	Build		Scheduling	No State	0 minutes	Associate Floating IP ▾
☐	demo-1	CC-CentOS7-RDMA-Hadoop		baremetal	shashank-bash	Build		Scheduling	No State	0 minutes	Associate Floating IP ▾
☐	mpi-instance-0	CC-CentOS7-SRIOV-MVAPICH2-Virt	10.40.1.155 Floating IPs: 129.114.108.31	baremetal	jzmac	Active	climate:8e22078c-8a00-4a40-acb3-0a701bf2e903	None	Running	1 hour, 33 minutes	Disassociate Floating IP ▾
☐	mpi-instance-1	CC-CentOS7-SRIOV-MVAPICH2-Virt	10.40.1.154	baremetal	jzmac	Active	climate:8e22078c-8a00-4a40-acb3-0a701bf2e903	None	Running	1 hour, 33 minutes	Associate Floating IP ▾

Displaying 7 items

Demo – Launch Your Instances

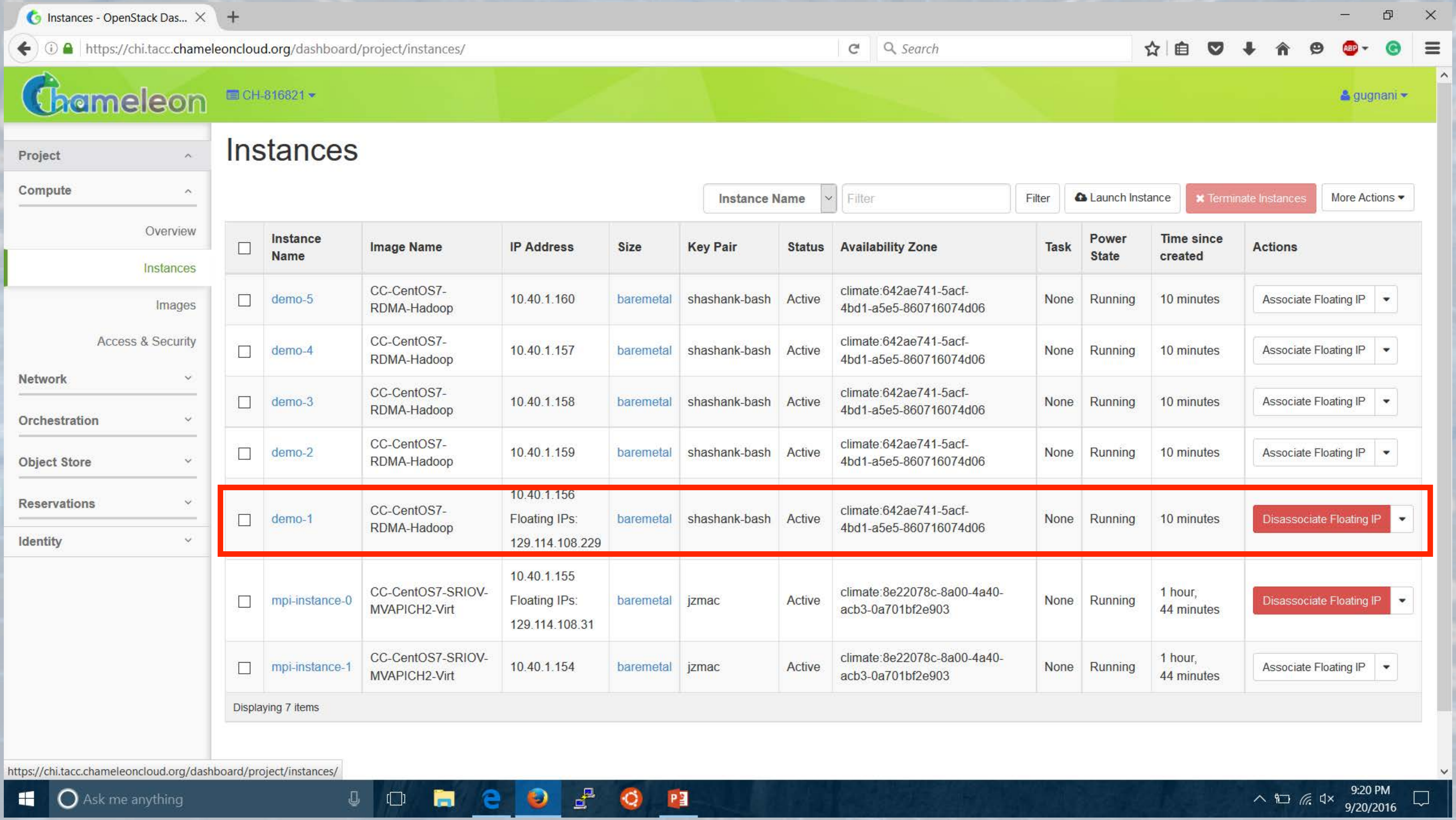


Instances

□	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Time since created	Actions
□	demo-5	CC-CentOS7-RDMA-Hadoop	10.40.1.160	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	5 minutes	Associate Floating IP
□	demo-4	CC-CentOS7-RDMA-Hadoop	10.40.1.157	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	5 minutes	Associate Floating IP
□	demo-3	CC-CentOS7-RDMA-Hadoop	10.40.1.158	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	5 minutes	Associate Floating IP
□	demo-2	CC-CentOS7-RDMA-Hadoop	10.40.1.159	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	5 minutes	Associate Floating IP
□	demo-1	CC-CentOS7-RDMA-Hadoop	10.40.1.156	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	5 minutes	Associate Floating IP
□	mpi-instance-0	CC-CentOS7-SRIOV-MVAPICH2-Virt	10.40.1.155 Floating IPs: 129.114.108.31	baremetal	jzmac	Active	climate:8e22078c-8a00-4a40-acb3-0a701bf2e903	None	Running	1 hour, 33 minutes	Disassociate Floating IP
□	mpi-instance-1	CC-CentOS7-SRIOV-MVAPICH2-Virt	10.40.1.154	baremetal	jzmac	Active	climate:8e22078c-8a00-4a40-acb3-0a701bf2e903	None	Running	1 hour, 33 minutes	Associate Floating IP

Displaying 7 items

Demo – Associate Floating IP

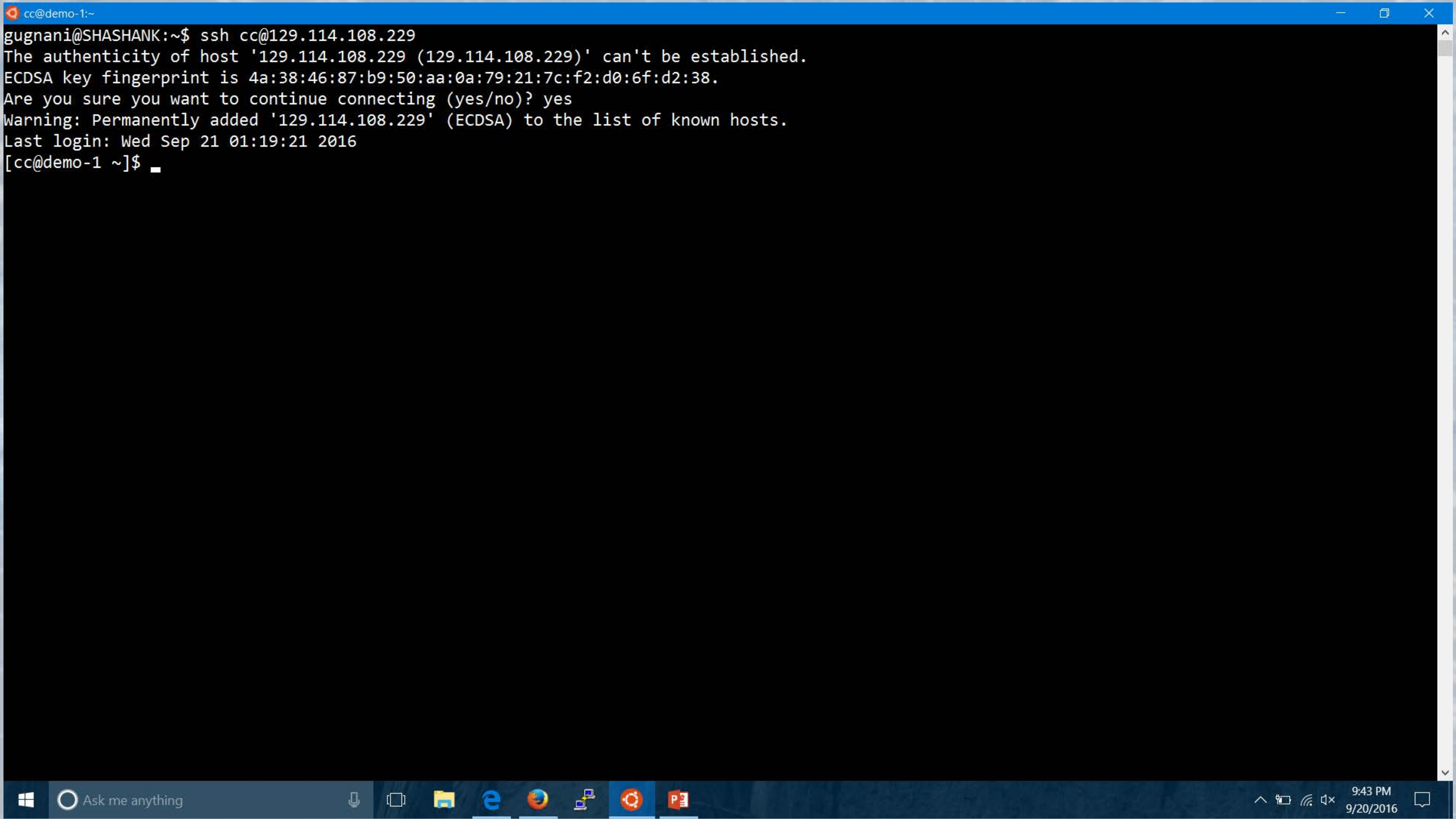


Instances

□	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Time since created	Actions
□	demo-5	CC-CentOS7-RDMA-Hadoop	10.40.1.160	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	10 minutes	Associate Floating IP
□	demo-4	CC-CentOS7-RDMA-Hadoop	10.40.1.157	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	10 minutes	Associate Floating IP
□	demo-3	CC-CentOS7-RDMA-Hadoop	10.40.1.158	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	10 minutes	Associate Floating IP
□	demo-2	CC-CentOS7-RDMA-Hadoop	10.40.1.159	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	10 minutes	Associate Floating IP
□	demo-1	CC-CentOS7-RDMA-Hadoop	10.40.1.156 Floating IPs: 129.114.108.229	baremetal	shashank-bash	Active	climate:642ae741-5acf-4bd1-a5e5-860716074d06	None	Running	10 minutes	Disassociate Floating IP
□	mpi-instance-0	CC-CentOS7-SRIOV-MVAPICH2-Virt	10.40.1.155 Floating IPs: 129.114.108.31	baremetal	jzmac	Active	climate:8e22078c-8a00-4a40-acb3-0a701bf2e903	None	Running	1 hour, 44 minutes	Disassociate Floating IP
□	mpi-instance-1	CC-CentOS7-SRIOV-MVAPICH2-Virt	10.40.1.154	baremetal	jzmac	Active	climate:8e22078c-8a00-4a40-acb3-0a701bf2e903	None	Running	1 hour, 44 minutes	Associate Floating IP

Displaying 7 items

Demo – Login to Your Instances



```
gugnani@SHASHANK:~$ ssh cc@129.114.108.229
The authenticity of host '129.114.108.229 (129.114.108.229)' can't be established.
ECDSA key fingerprint is 4a:38:46:87:b9:50:aa:0a:79:21:7c:f2:d0:6f:d2:38.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '129.114.108.229' (ECDSA) to the list of known hosts.
Last login: Wed Sep 21 01:19:21 2016
[cc@demo-1 ~]$
```

Demo – Play with Your Hadoop on Chameleon Cloud

- ssh cc@129.114.108.229
- cd rdma-hadoop-2.x-0.9.8/
- ./sbin/start-all.sh
- bin/hdfs dfsadmin –report
- bin/mapred job -list-active-trackers
- ./bin/hdfs dfs -ls /
- bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar randomwriter -Dmapreduce.randomwriter.mapsperhost=4 -Dmapreduce.randomwriter.bytespermap=1000 /rw-output
- ./bin/hdfs dfs -ls /
- bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar sort /rw-output /sort-output
- ./bin/hdfs dfs -ls /
- ./sbin/stop-all.sh

Conclusion

- Overview of MapReduce Programming Model and Hadoop Architecture
- Overview of RDD Programming Model and Spark Architecture
- Overview of HiBD project: accelerating Big Data processing middleware (e.g., Spark, Hadoop, Memcached) by taking advantage of HPC technologies, such as InfiniBand/RDMA, SSD, etc.
- A case study with NeuroPigPen
- Demo of RDMA-Hadoop on NSF supported Chameleon Cloud
- Soliciting collaboration with other groups
  - Interested in exploring BigData processing requirements from Neuroscience researchers
  - We can explore how to collaborate further

Thank You!

luxi@cse.ohio-state.edu

Network Based Computing Laboratory

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

MVAPICH

HiBD
High-Performance
Big Data

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

Computational Neuroscience Network (ACNN)

http://www.NeuroscienceNetwork.org/ACNN_Workshop_2016.html