Introduction to Cloud Computing

Qloud Demonstration

15-319, spring 2010

3rd Lecture, Jan 19th

Suhail Rehman

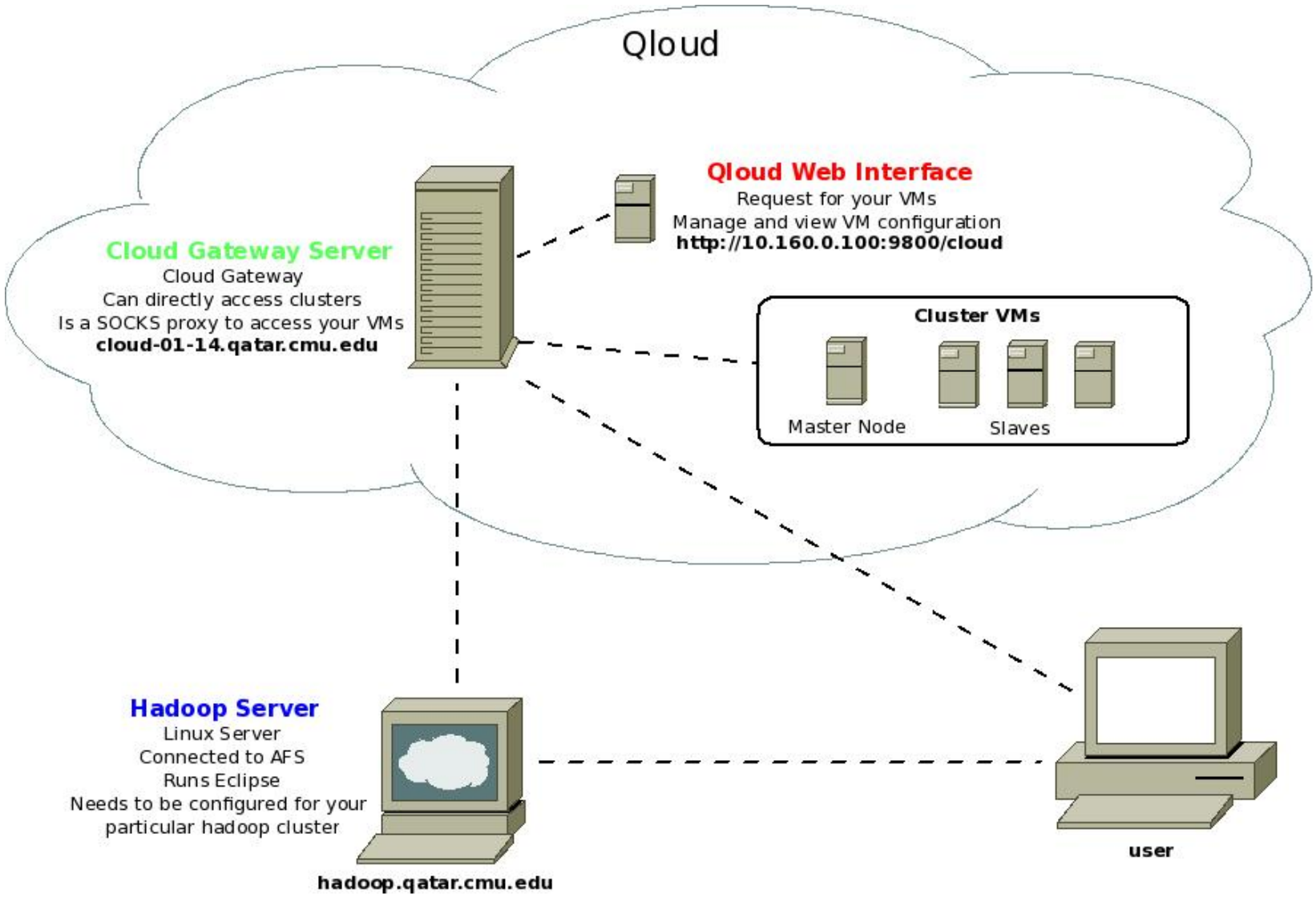
Time to check out the Qloud!

- Enough Talk!
- Time for some Action!
- Finally you can have your own Cloud (Virtual Machines)!
- Get your own Cloud from Qloud!

Time to check out the Qloud!

- Enough Talk!
- Time for some Action!
- Finally you can have your own Cloud (Virtual Machines)!
- Get your own Cloud from Qloud!

User's Qloud Perspective



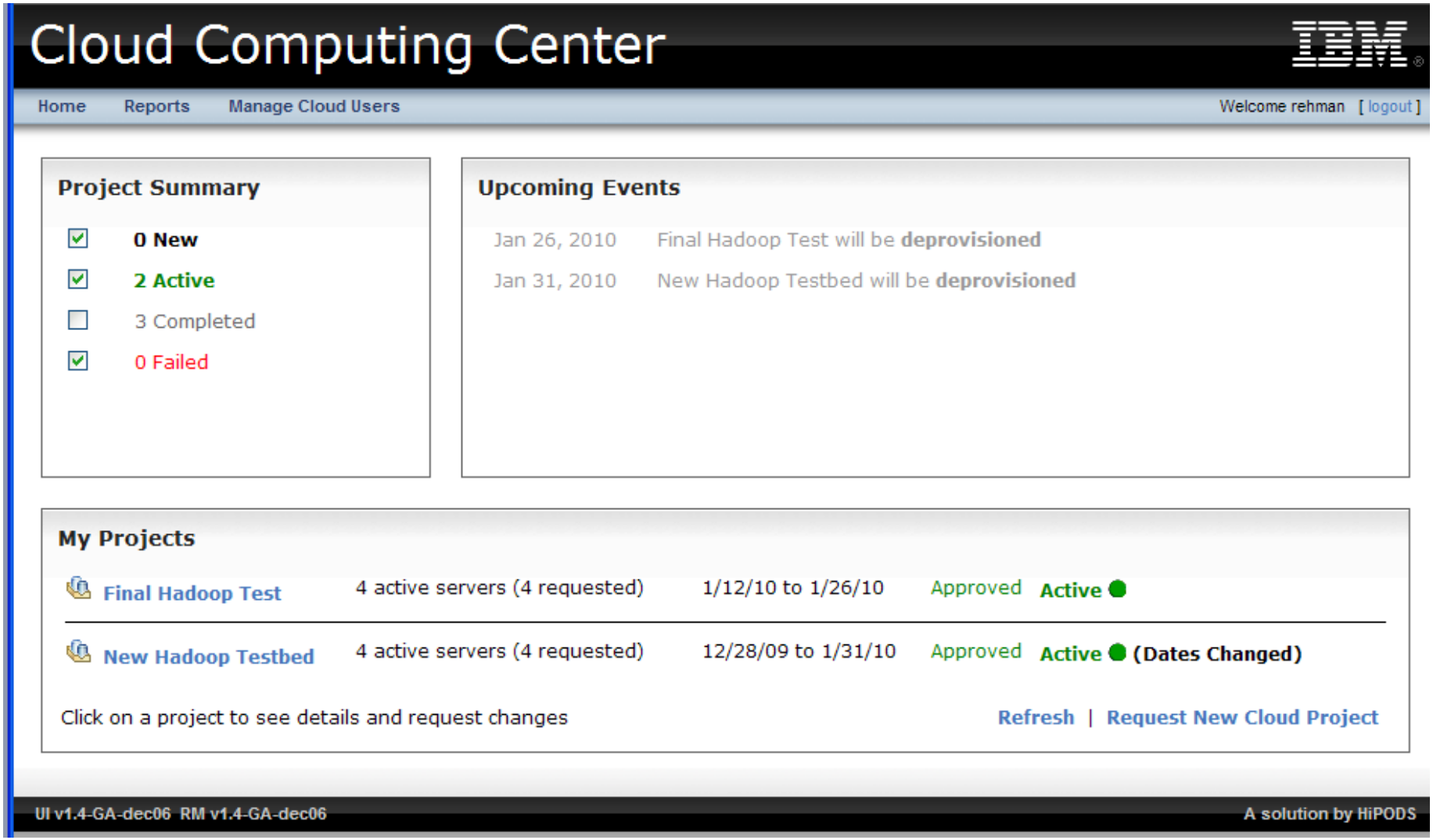
Important Qloud servers and interfaces

- **Hadoop Server**
  hadoop.qatar.cmu.edu
  - User workspace (Hadoop/Eclipse)
  - AFS access and login
- **Cloud Gateway Server**
  cloud-01-14.qatar.cmu.edu
  - Gives you access the virtualized resources of the cloud
  - Will be a SOCKS proxy for all your Cloud and Hadoop tasks
- **Qloud Web Interface**
  http://10.160.0.100:9080/cloud/
  - Easy web interface to request your Cloud
  - Once provisioned, you can checkout the vital stats of your cloud

Steps to get your own Cloud

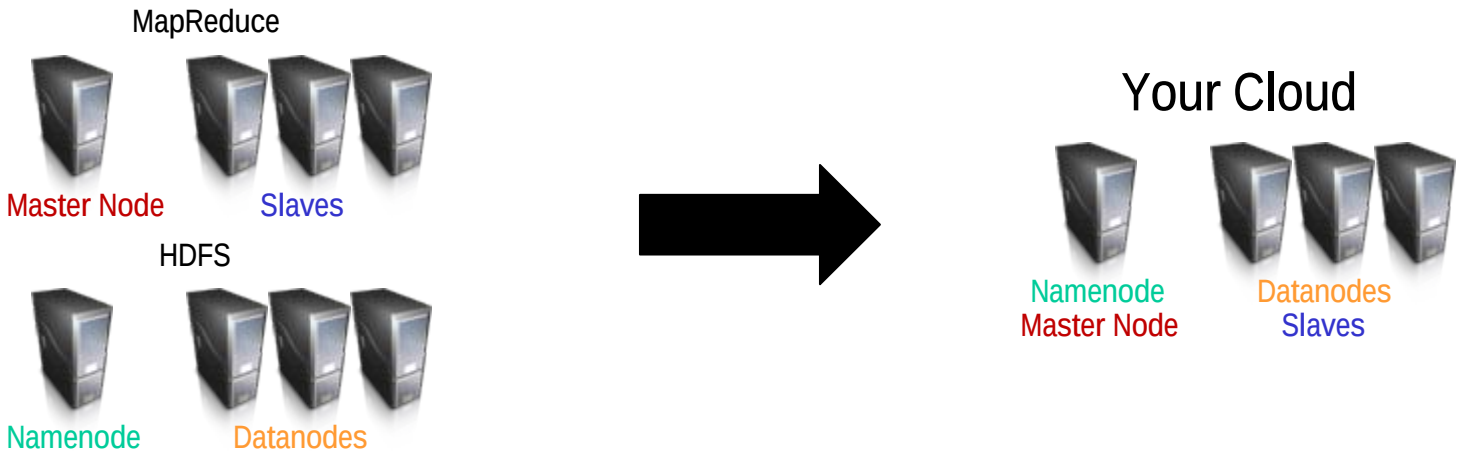
- Set the Cloud Gateway Server as the SOCKS proxy in your Browser
- Log on to the Qloud Web Interface and request your Cloud
- Wait for our uber-geek (aka Brian) to approve
- Once Brian approves it, you'll have your cloud in 2 hours
- The entire process should take less than 24 hours ☺
  - You cannot request a cloud at 2am and expect it to be ready at 4am

Qloud Web Interface

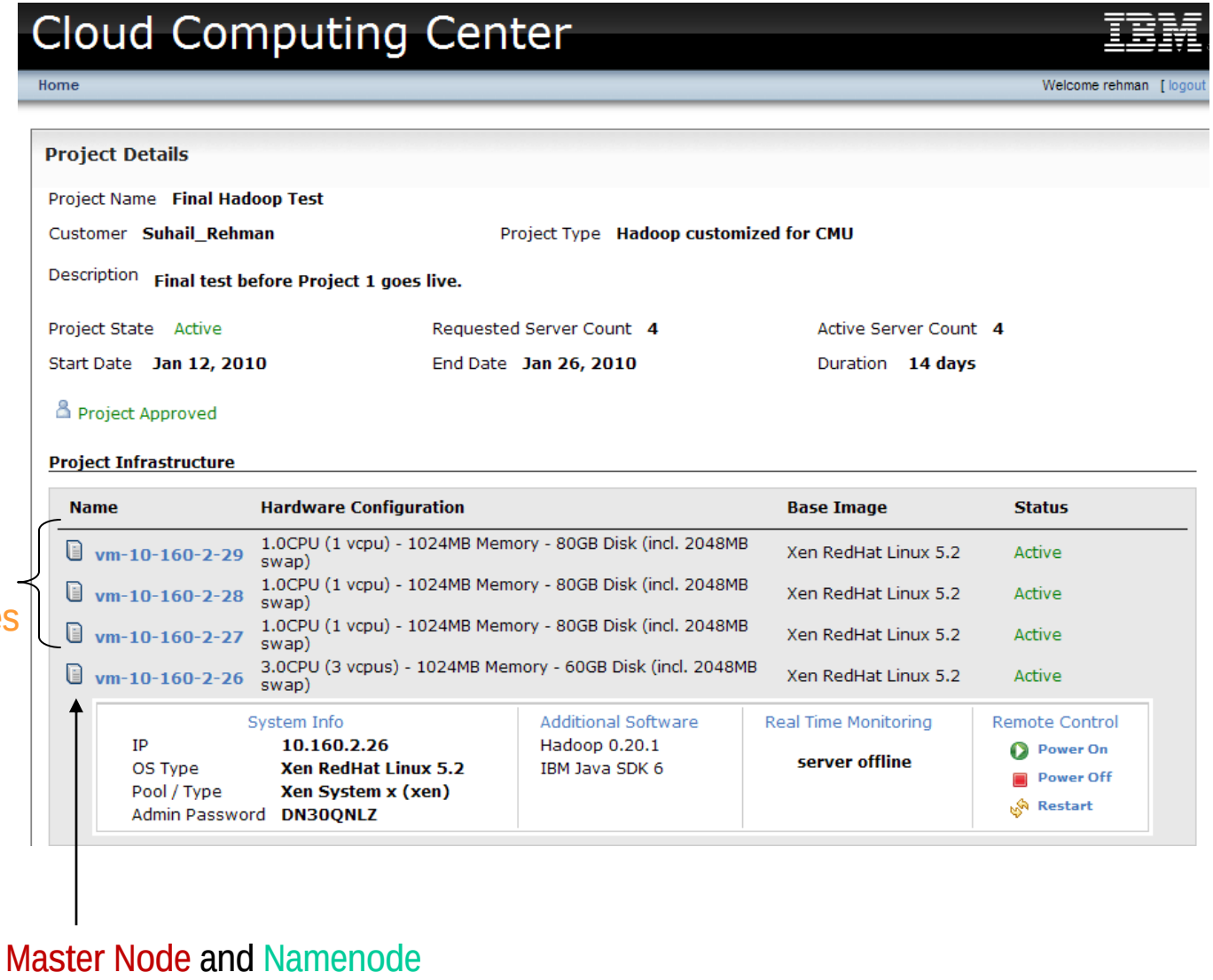


It is time for Hadoop ☺

- The Hadoop infrastructure allows you to run map-reduce jobs distributed over your virtual machines
- In Hadoop MapReduce, one node is designated as the *Master Node*, and the rest are slaves.
- HDFS requires one Namenode and several Datanodes.
- In our setup, the Master Node and Namenode are the same machine.



Hadoop on Your Cloud

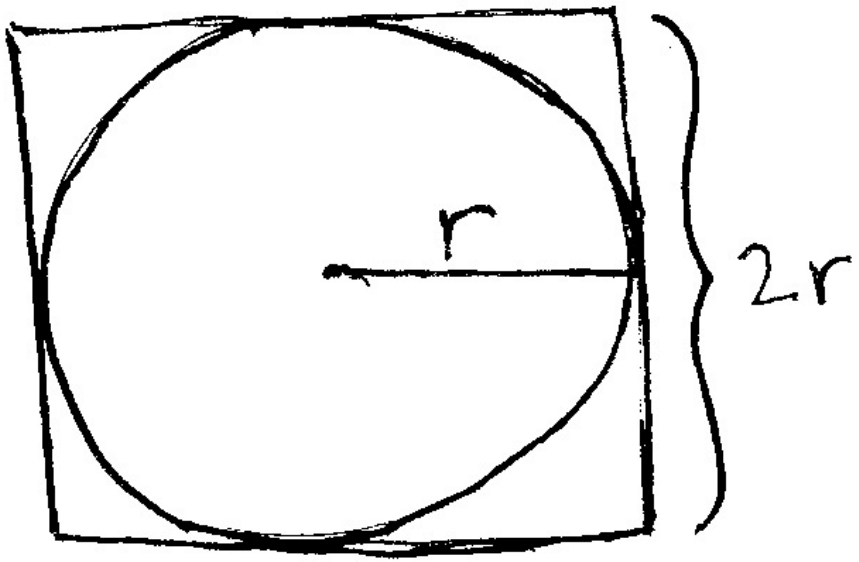


Where to go from here?

- **Logon to your Master Node**
  - ssh to cloud-01-14.qatar.cmu.edu and then ssh to your **master node**
- **Setup Hadoop**
  - Fortunately, your VM's automatically have the correct configuration files for Hadoop the moment they are provisioned (Thanks to Brian!)
  - All you need to do is format HDFS and start the Hadoop services.
- **Lets try running some sample code on Hadoop**

Sample MapReduce Code- Estimate π

- Estimating π by random sampling
- Imagine you have a dart board like so:



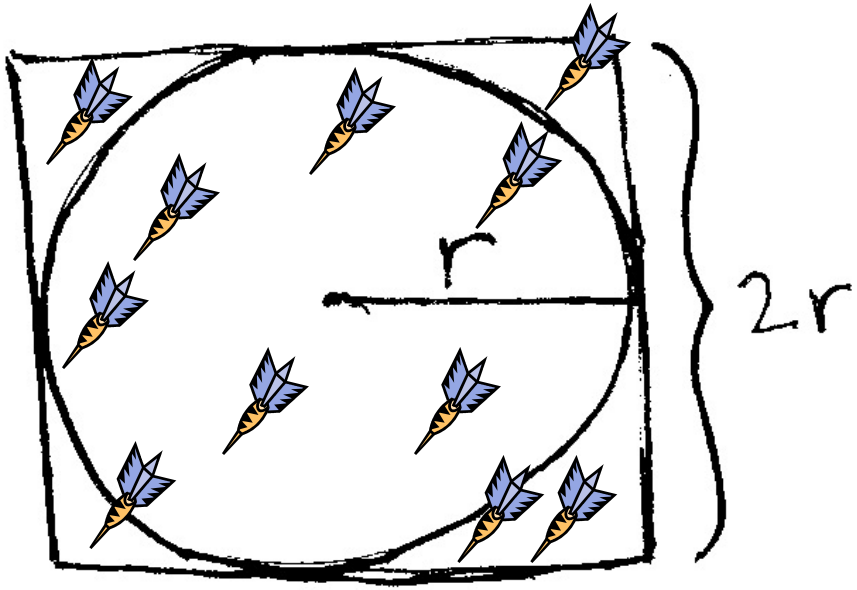
$$P(dart\ in\ circle) = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$

$$\Rightarrow \pi = 4P(dart\ in\ circle)$$

- π is simply the (ratio of darts that land inside the circle to the total number of darts thrown) times 4

Writing this as a Serial Program

- Throw N darts on the board. Each dart lands at a random position (x,y) on the board.



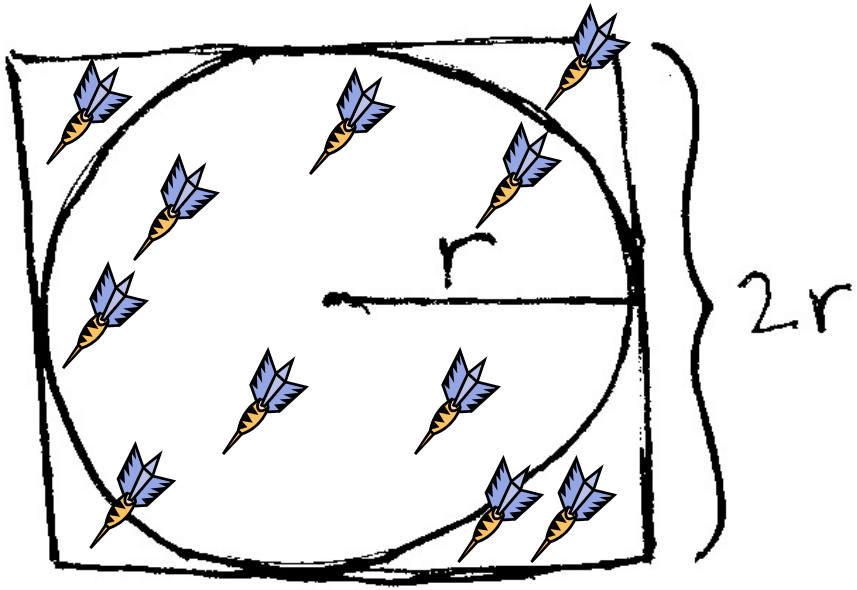
- Note if each dart landed inside the circle or not
  - Check if $x^2+y^2<r$
- Take the total number of darts that landed in the circle as S

$$4\left(\frac{S}{N}\right)=\pi$$

But I have Millions of Darts!

- If you want to get an accurate estimate of Pi, you need a large number of random samples.
- Notice that each dart can be thrown at any time and it's position can be evaluated independently
- With one person throwing all the darts, it will take a long time to finish
- If we had N people throwing a dart each, this would be much faster!

But I have Millions of Darts!



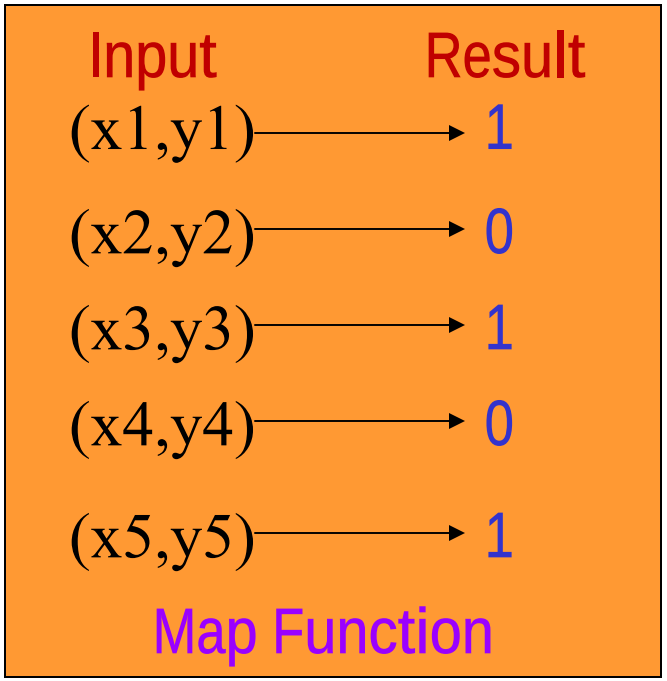
How do you do this in Parallel?

- Let *(x,y)* be a random position of the dart inside the square.
- Each *(x,y)* pair can be evaluated independently.
- Let us "map" each *(x,y)* pair to a result – the result being whether it is inside the circle (1) or not (0).

Input		Result
(x1,y1)	→	1
(x2,y2)	→	0
(x3,y3)	→	1
(x4,y4)	→	0
(x5,y5)	→	1

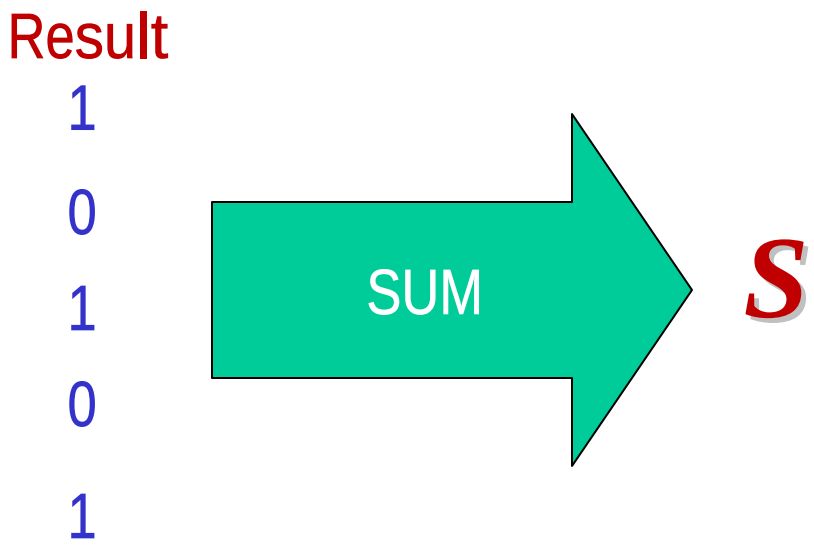
The Map function

- A Map function takes input values and produces an output for each input value in parallel.



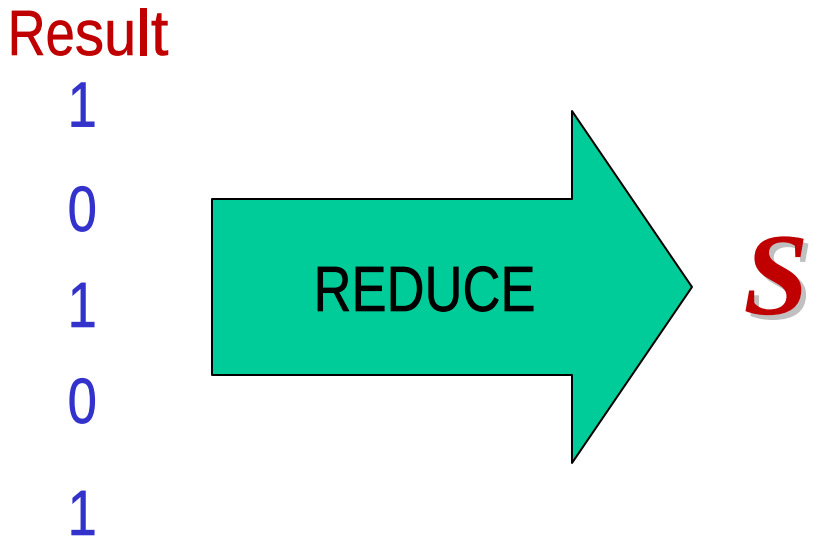
…..and then?

- So we have results of each (x,y) pair – lots of them
- We need to find the number of points inside the circle. We need to sum up the values



The Reduce Function

- A Reduce function takes input values from the Map functions and produces output using a user defined operation.



- In this case, addition is the reduce operation.

What about Pi?

- Now that we have the total number of points inside the circle, S and the total number of points N we've sampled…

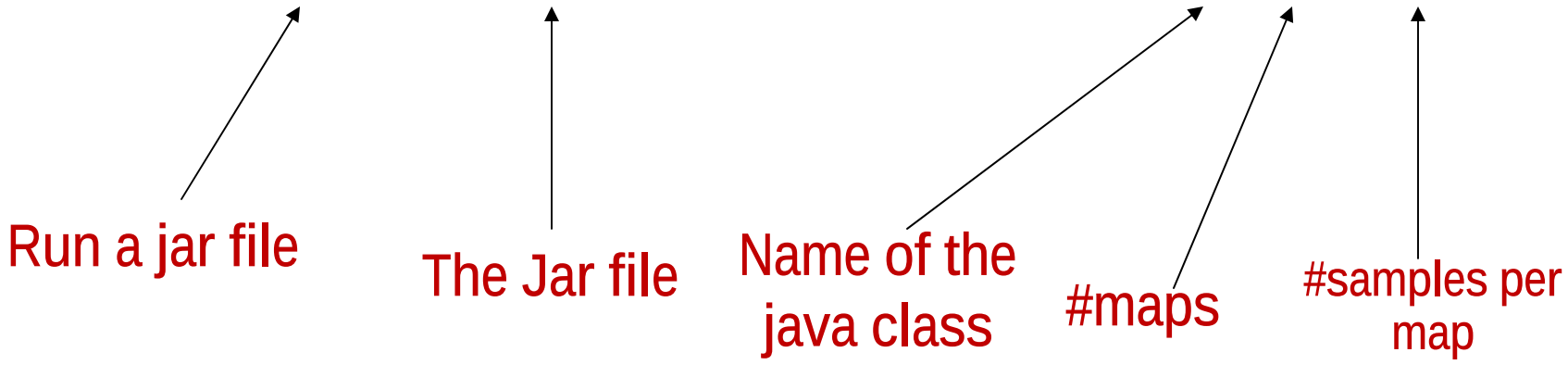
$$4\left(\frac{S}{N}\right) = \pi^{*}$$

*Subject to Terms and Conditions

1. N should be large
2. Points should be chosen uniformly at random

Running PI MapReduce Code

- The MapReduce code creates random (x,y) pair values
- It gives each node a number of (x,y) pairs and evaluates if it's in the circle or not (MAP)
- Then some nodes will collect the results of these samples, evaluate the percentage and calculate π (REDUCE)
- Running the hadoop example:
  - `hadoop jar hadoop-0.20.1-examples.jar pi 10 100`

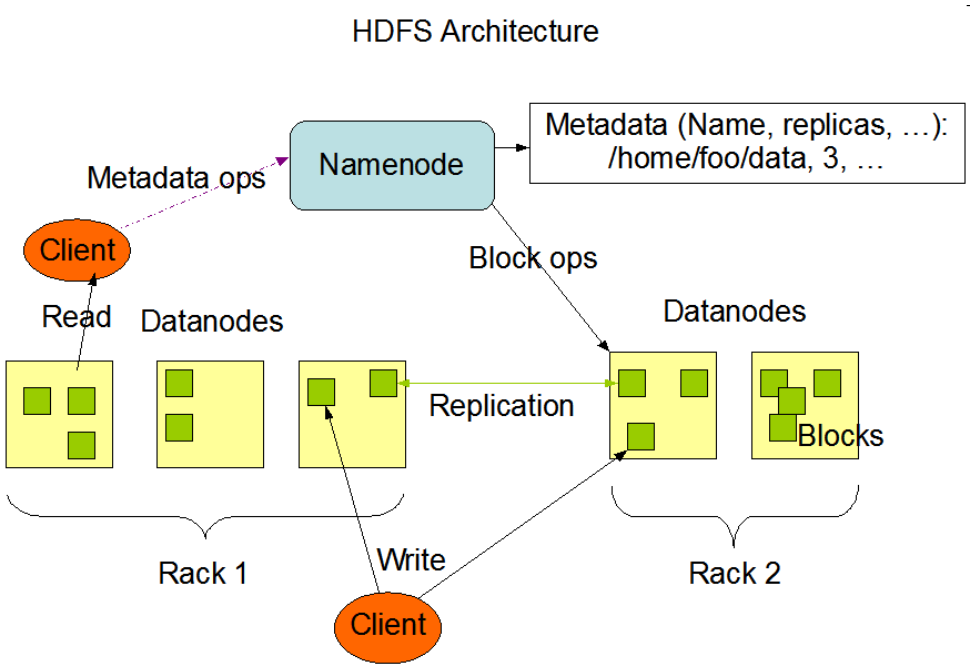


Working with Files in Hadoop

- Notice that the Pi example randomly generates input, it does not require any user files.
- Hadoop is mainly used to work with large data, and large data is always in a file.
- HDFS to the rescue!

HDFS Basics

- **HDFS is the Hadoop Distributed File System.**
- **Files are distributed over all four nodes and are triple-replicated, by default, to tolerate failure.**



HDFS Commands

- All commands begin with `hadoop dfs`

UNIX command	Hadoop HDFS Command
`ls /`	`hadoop dfs –ls /`
`cat /dir/filename`	`hadoop dfs –cat /dir/filename`
`mkdir dir1`	`hadoop dfs –mkdir /dir1`
`rm /dir/filename`	`hadoop dfs –rm /dir/filename`
`rm –r /dir`	`hadoop dfs –rmr /dir`

Handling Files in HDFS

- **To add files to HDFS:**
  - hadoop dfs –put localfilename /hdfs_dir/remotefilename
- **To copy files from HDFS to local filesystem**
  - hadoop dfs –get /hdfs_dir/remotefilename localfilename
- **To copy files inside HDFS filesystem**
  - hadoop dfs –cp /hdfs_dir/file1 /hdfs_dir/file2

Keeping track of your Hadoop & HDFS

- **Hadoop MapReduce has a JobTracker web interface**
  - Keeps Track of the submitted jobs, time taken, errors, logs etc.
  - http://MASTER_NODE_IP:50030

- **The HDFS Namenode also maintains a web interface**
  - Browse your HDFS files
  - See how much disk space you have remaining in your HDFS.
  - http://NAME_NODE_IP:50070

Setting up Eclipse

- **Might be easier to work with an IDE when developing large applications in Hadoop.**
- **Eclipse is available on hadoop.qatar.cmu.edu with the MapReduce plugin**
- **Setup and Run eclipse @ hadoop.qatar.cmu.edu**
  - Use xwin32 on windows machines to run eclipse remotely
  - Configure Eclipse to use your cloud
  - Start developing MapReduce applications