Hadoop on C8 cluster

Yida Wang
What is Hadoop

• An open-source software framework that supports data-intensive distributed applications
• Running of applications on large clusters of commodity hardware
• Using MapReduce abstraction, based on HDFS
C8 cluster

- 66 nodes, 2.66GHz Intel Xeon processors, 8 cores/16GB DRAM per node, Gig-E interconnect

Structure of one node

Structure of the cluster
Hadoop on C8 nodes

• 1 NameNode (master): e0/c8-0-0
• 1 Secondary NameNode: (backup): e1/c8-0-1
• 16 DataNode (slaves): e6~e21/c8-0-0-6~c8-0-21
• ~50TB disk space
  – Block replication: 3
• Administrator: fuse
Toy examples

• **WordCount**: provided Java code
  – Take jar package as the code input
  – `hadoop jar /opt/hadoop/hadoop-examples-1.0.4.jar wordcount /user/fuse/essays output`

• **Hadoop Streaming**: if you don’t write in Java
  – Run map/reduce jobs with any executable or script
  – `hadoop jar /opt/hadoop/contrib/streaming/hadoop-streaming-1.0.4.jar -input /user/fuse/essays -output output -mapper cat -reducer wc`
Exercise

• Refine the given WordCount to clean up the documents and output the frequent words
  – Source code: HDFS:/user/fuse/WordCount.java
  – Corpus: HDFS:/user/fuse/essays
• No Java? You can use Hadoop Streaming
More info

• Get an account: write to me at yidawang@cs

• Access the cluster:
  1. ssh to yourNetID@mmx.cs.princeton.edu
  2. ssh to the cluster, usually go to the NameNode at e0: ssh e0 from mmx.cs.princeton.edu

• Hadoop-related questions
  – http://hadoop.apache.org/
  – google it

• Some interesting data at /dd/2/home/project_archive/2_ferret/data