CIS 508 (Spring 2014) Assignment 2

Name: Yanan Xiao

Masdar Email Account: yxiao@masdar.ac.ae

Deadline: **3 Mar 2013, 11:59pm (Late submissions of assignments will not be marked)**

Submission Instructions:

1. Fill in this document
2. Save as a pdf file, with filename as "cis508-2014-asg2-[MasdarEmailAccount].pdf" (replace [MasdarEmailAccount] by your own Masdar Email Account)
3. Send as an attachment to sidckchau@gmail.com, with email subject as "[CIS 508] Assignment 2 (2014)" before the above deadline

Questions:

Part 1: Fault tolerance.

1. What is Hadoop Distributed File System? Compare it with Google File System (including the system designs and architectures). (2 Marks)

Definition: It is a storage component of the Hadoop distributed high performance computing platform, and features in fault tolerance, scalability and distributivity.
Comparison: Since Hadoop is inspired by GFS, there is no wonder that these two share a lot in common.
Similarities:
Both are designed for a high redundancy and suitable to run on commodity hardwares.
For a single cluster, both file systems have the master/slave architecture. The master node(or server) does not operate on the files directly, but records all metadata of files.
Differences:
There is no record append operation in HDFS.
Hadoop is written in Java as an open source software. At the same time, it is reported that its real-time performance is not as good as claimed.

2. How does the concept of Google File System revolutionize the designs of data center (e.g., the supported applications, the management system, and costs)? (2 Marks)

As reported in Google's paper, the benchmark result for a small cluster of servers are not promising. It is barely comparable to a single disk with even worse write and append performance. But in real-life the number of servers at data center is much larger. Therefore with the help of distributed file system like GFS, the overall performance of a data center would be dramatically benefited, even running on commodity hardwares.

Supported applications: The GFS is built to support the “unique” need of Google services and applications. In this manner, we can see the growth of data centers built to provide more than just the computing service. Video streaming and storage services are benefited from GFS alike due to the I/O throughputs and redundancy.

Management system: The duty on management system is heavily reduced mainly because of the redundancy of these distributed file systems.

Cost: One feature inspired by GFS is that when tunned well, even commodity hardwares can do wonders.

Part 2: Wireless communications.

3. One of the solutions that addresses the scarcity of wireless spectrum is based on “white spaces”. What is the main idea behind white spaces? Discuss its advantages and disadvantages. (2 Marks)

Main idea: If things as valuable as wireless spectrum is not made the most of, it is in fact a waste for people and things related to the communication industry. To be more accurate, the white spaces are frequencies that are not used locally though already allocated to a broadcasting service. In this question, it mainly refers to the spectrum left over after TV signals went digital.

Advantages:
High efficiency in using wireless spectrum
In the U.S., this range of spectrum is unlicensed. Therefore any person or organization is free to use it.

Disadvantages:
No clear report has been made on whether interference would be caused after completely open this spectrum.

4. Google aims to advance white spaces. What has Google done? How can white spaces contribute to Google's business? (2 Marks)

Google:
Becoming a member of White Space Coalition which is consisted of eight large tech companies.
Sponsored a campaign named Free the Airwaves trying to tempting more people & organizations into using this spectrum.

Benefits:
If this range of spectrum is officially opened, and better if unlicensed, more Wi-Fi vendors will produced devices that support it. Take Chromecast as an example, the Wi-Fi communication it's using may enjoy a better QoS (quality of service).
This spectrum may be used in the communication between cars like the Google driverless cars to which a stable wireless communication is important.

Part 3: Cloud Computing.

5. Compare Google compute engine with AWS (including the pricing schemes, service models, system platforms). (2 Marks)

Pricing:
Both paid service of GAE and AWS support usage based pricing.
#GAE: The free version is enough for experimental or small scale usage. Whereas the paid version enjoys highend service like "infinitely scalable".
#AWS: One Amazon account enjoys free tier of all web services, most of them are free for up to one year.

Service model:
From the pricing policy we can see that both companies provide "try before you pay & pay as you go" service to be more customer-friendly.
#GAE: Currently the Google cloud services are computing-based, providing backend for apps that need network communication & computation.
#AWS: As a forerunner, Amazon's cloud services are diversified. From storage to DNS, Amazon tries to provide companies one-stop experience with all those "products".

Platforms:
Both of them are based on each company's own data centers. Virtualization plays an important role in making those service highly scalable and reliable.

6. MapReduce is designed to support a wide range of tasks in Google. Discuss how MapReduce can support PageRank computations. (2 Marks)

The idea and conceptual details are discussed in a book by some Google senior engineer.
PageRank: To be brief, this algorithm takes votes from all other sites in the Internet to "vote" how important a site is. If one site has a lot of links linking to it, then its rank should be high when searched.
The initial rank for all sites are assumed to be equal. But it is mathematically proven that no matter what value the initial rank (B=[b1,b2,b3...,bn]) take, it will converge eventually after several iterations of voting.
Let's assume **sparse matrix** A=[a11,...,amm] is a matrix where amn represents the number of links linking from site (or more practically, page) m to site n.
Then what we need to do is, Bi = A * Bi-1. (* is the matrix dot product.)
- - - - - - -
MapReduce: It is just a fancy name for divide-and-conquer algorithm.
As we can see from the "simple" equation above, doing the PageRank is mathematically equal to doing an iteration of matrix product. In practice, these two matrix are huge and sparse. (The handling for sparse matrix is omitted in this answer.)
Divide matrix A by its rows, we get something like A1, A2, A3... And divide matrix B by its columns, we get B1, B2, B3...
Then the huge dot product is mapped into tens of thousands of small dot product of matrix, like A1B1, A2B3...
These divided tasks could then be computed by normal or commodity computers, which are conceptually regarded as nodes in clusters.
"Reduce" the divided results from those nodes, and then we can get the value we want.
- - - - - - -
The technical details like how to divide a huge matrix automatically, ensure the load average of each node, and successfully "reduce" the value to get what we want are not revealed by Google.