Alif Rahi
Cloud Computing Midterm

1. Describe the benefits of virtualization for cloud computing.

1) Flexibility: Using a VM we can easily move one physical server to another physical server.
2) Resource optimization: Dynamic load balancing: move hotspot to under-utilizaed hardware.
3) Better server consolidation, 10 to 1 in many cases.
4) Disaster recovery: move affected VMs to other hardware.
5) Easy management: With a virtualized interface, you can manage the resources and provide better levels of the facilitation for sharing resources.
6) Cheaper to migrate virtualized services rather than physical ones on your computer.

2. Describe (at least five) features of cloud computing. Justify your answer with details.

Cloud computing is soon going to take over almost every company's infrastructure and systems. From learning in class and from person use of AWS services, I can mention 5 features and benefits of those features.

1) One feature is **Scalability and elasticity**. Cloud computing provided this ability to scale resources up or down in response to changes in demand with quickness. This is crucial for companies when they need to increase their infrastructure or scale it down.
2) **Resource reusability** is another feature. Cloud computing provides a lot of computing resources such as servers, storage and equipment. This feature allows multiple users to share the same resources which improves resource utilization and reduces costs.
3) Cloud computing also allows you to be **anywhere in the world** as long as there is a reliable internet connection available. Cloud computing relies on remote servers and data centers so as long as the user has access to the internet, they can access cloud services from anywhere from their phones, desktops or laptops etc…
4) Another thing is that **you can service yourself**. You can literally make a full stack application and deploy it completely by yourself by using the cloud. You can put your node.js server onto an AWS EC2 instance. You can deploy your database into AWS RDS and deploy your front end into AWS Amplify. It's up to you to scale up or down your infrastructure based on your needs, without having to go through a lengthy procurement process.
5) This brings me to my last feature which is the **measured services** and pay as you go feature. Cloud computing providers typically offer a usage-based pricing model. This allows you to pay based on what you actually used eliminating the need to over scale your applications like it was 30 years ago.

3. What is PUE? How is PUE calculated? Describe the factors that may affect PUE.
PUE is power usage effectiveness and is a metric used to measure the energy efficiency of a data center (Which is total power divided by server power). 1.0 is a perfect PUE. The efficiency of equipment, the climate, the design of the data center itself and power distribution are all ways that the PUE can change. We generally want lower values.

4. Compare the concepts of region, availability zone, and data center.
Availability zones are unique physical locations within regions that are isolated from other availability zones. Each of these zones are made up of one or more data centers; which are physical facilities where cloud providers house their computing infrastructure, such as servers, storage, and other networking equipment. Data centers are usually located in places where electricity is cheap and cooling is free to keep costs down. They are also designed to maintain high levels of reliability, scalability, and security. Project Natik was an example of a submarine data center which is very very cheap, very unlikely to have malfunctions and also less chances of human faults since it was under water.

5. Why is the container more lightweight than the VM?
There are numerous reasons for why a container is more lightweight than a VM but they all stem from the fact that a VM requires an operating system while containers share the operating system kernel with the host machine, which means they do not need to include an operating system like a VM does. This makes containers much lighter making them more portable and faster to start up. Another thing that makes containers advantageous is that they are platform agnostic meaning they can work on any system that supports the container. For example Docker desktop is a popular containerizing software that runs on all OS.

6. Why is Kubernetes neither IaaS nor PaaS?
Kubernetes is neither an IaaS nor a PaaS. It provides a platform for deploying and managing containerized applications while also allowing users to configure and manage the infrastructure resources. It's a container orchestration engine which makes it more like a Container As A Service or CaaS. You need a IaaS layer below kubernetes to provide VMs such as an AWS EC2 server or other bare metal servers.

7. Why does MapReduce need to ensure that all key-value pairs with the same key are shuffled to the same reducer? How is it implemented in Hadoop?
It does this for multiple reasons. One reason is, when a MapReduce Job is sending all of the Key value pairs to its specific reducer, it can perform aggregated or grouped operations on it. If the input values are not sorted by keys, It would have to scan all of the Mapper outputs to pick up every instance of that key and value. If the Mapper output is sorted by as soon as it is picked up, you would have that group/set of keys ready to be sent to its designated reducer for processing. The benefit of this is that you won't have to wait for every reducer to be ready in order for each of them to start reducing as it does this in parallel.

8. Given a file of 10 GB saved in HDFS with 256 MB as its chunk size (1 GB = 1024 MB), how many map tasks are there in a MapReduce job taking this file as its input? Assume that this job performs word counting with 4 reduce tasks. If the size of this job's output is 1 GB and each word appears 50 times on average, what is the average amount of data that a reduce tasks receives from a map task?
Block size 256 MB
= 10,240 MB / 256 MB = 40
MapReduce job will have **40 map tasks.**
Each map task will output data to 4 reducer tasks. If the data is 1 GB, which means each reducer will have a partition size of 1GB / 4. (256 MB)
Each partition contains 256 MB of data
Thus each reducer receives data of the size 256 / 40 **or approximately 6.4 MB** of data from each map task.

9. Compare the difference between the transformation and the action in Spark.
Transformations define a new RDD based on current RDDs while actions return values by triggering the computation of RDDs. Transformations are evaluated using lazy evaluation while actions, on the other hand, are evaluated using an eager evaluation.

10) Calculate 1 + 2^3 + . . . + n^3 in Spark.

```
// Create an RDD of integers from 1 to n
nums = sc.parallelize( [1, 2, 3, 4, … n] )

// map each integer to its cube
cubes = nums.map(lambda x: x*x*x)

// reduce the RDD to get the sum
sum = nums.reduce(lambda x, y: x + y)

// Print the result
println(sum)
```