



Faculty of Engineering & Applied Science

SOFE4790U – Distributed Systems

Homework: HDFS as a Distributed System

Due Date: 09/18/2022

First Name	Last Name	Student ID
Abdul	Bhutta	100785884

1. Reflect on your previous course on OS. Can you provide a few paragraphs of a summary of what you studied in the course?

In the operating system course, we learned different aspects of operating systems in terms of functionality and performance. The course was broken down into various components to help us understand operating system functions from hardware to software related. The purpose was to design and implement simple tasks of an operating system while understanding how each concept was applied. The ubuntu/Linux OS and C language was used to demonstrate how processes were created and deleted while learning how the communication between the processes worked. Synchronization with OS concepts was introduced such as mutex locks, semaphores, and deadlocks. CPU scheduling concepts were introduced such as FIFO, LIFO, Round Robin, Shortest Job First, Queues, and Shortest Remaining Time First. Lastly, file management concepts were introduced such as virtual memory, memory allocation, paging, and page table structure. A summary of topics is provided below.


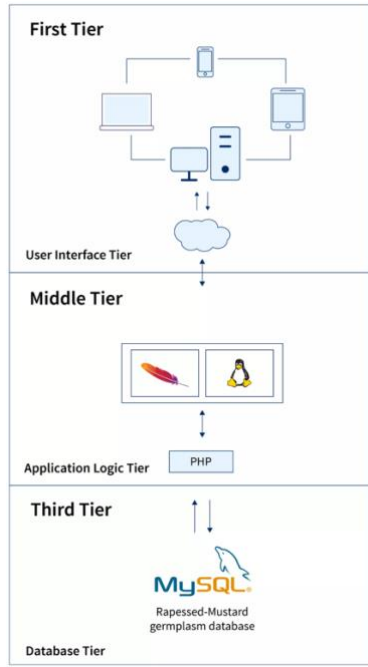
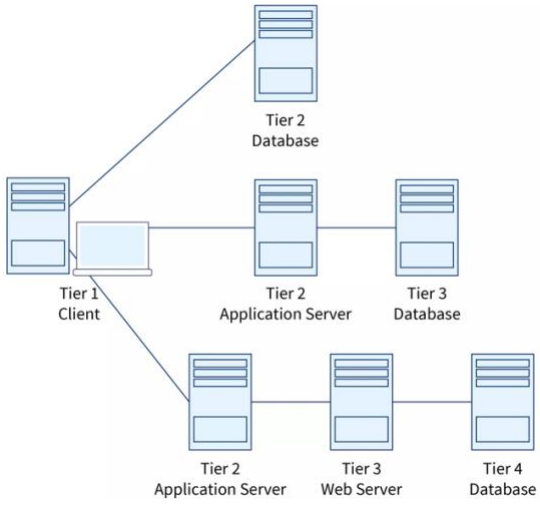
Topic	Concepts
Processes	Creation, Deletion, Inter-process communication
Threads	Creation, Deletion, Shared Memory
Synchronization	MUTEX locks, Semaphore, Deadlocks
CPU Scheduling	FIFO, Round Robin, Shortest Job first, Queue, Shortest Remaining Time First
File Management	Virtual Memory, Allocation, Paging, Page Table Structure

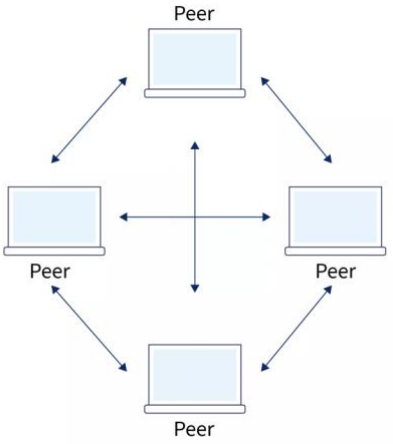
2. What are the main functions of a distributed OS?

A distributed operating system performs as a normal operating system but runs on multiple CPUs where the system has its CPU and memory while interacting with each other. The distributed operating system uses multiple CPUs for real-time applications to help process the data. Distributed systems allow the users for resource sharing, scalability, performance, and fault tolerance.

3. How do you implement them?

Most distributed systems rely on network calls therefore a good network is required to construct a distributed system. We need to explore the four types of distributed systems to implement the system. The four systems are shown below.

Type	Description	Implementation
Client-Server Model	Mainly used for resource sharing where the server handles data and resources requested by the client. An example is Netflix.	
Three – Tier	The three-tier architecture has three layers: presentation, application (business layer), and data. The presentation tier is where the user will access the application to make a request. The application layer is used for business logic. The data tier is the database tier where the application data is stored. An example of a three-tier architecture is a typical business application.	
Multi – Tier	Used for when application needs to forward data or requests to multiple networks	

Peer – to - peer	Each system is a node while acting as a client or server. An example of P2P is downloading torrents.	
------------------	--	--

4. Compare the functionalities provided by a centralized file system to that of a distributed file system?

A centralized file system is where all the data is stored on one server or location while limiting the number of users accessing the data. The data is stored on a single database and accessed through the local area network. The data stored is secured since all the data is stored in one database and retrieving the data will take less time. A distributed file system is where the data is stored on multiple databases in various locations that are interconnected with each other. The distributed file system performs better as there are more systems which reduce the network's load and allows the network to be accessed from more networks. A centralized file system is cheaper than a distributed file system since all the data is in one location and on one system while the distributed system will have many nodes/systems at different locations. Although in a distributed system, if one of the databases fails, not all users will be impacted but in a centralized file system, all users will be impacted.

5. What are the components of the distributed system listed in the article? Can you predict other components required that are not mentioned in the article?

Components	Description
NameNode	A dedicated server to store metadata
DataNodes	A dedicated server to store application data.

TCP-Based Protocols	All servers are interconnected and communicating through a TCP based protocol
HDFS Client	Users access the files through the HDFS client.
CheckpointNode	It combines the existing checkpoint to create a new checkpoint and an empty journal.
BackupNode	A new feature integrated incase of a a failure to the NameNode which keeps a backup of the NameNode.
Block Placement	It used for large clusters to distribute the nodes across multiple racks.
Replication Management	It used to verify all block have the number of intended replicas.
Balancer	Used to balance the disk space utilization

6. How are middleware transparencies provisioned within the described image-sharing system?

A middleware is a virtual middle layer between the applications and the hardware which runs the processes above the layer and all the resources are below the layer, this allows for resources and networks to be hidden away.

7. A HDFS is a distributed system, why do we build distributed systems? What are the advantages? What are the disadvantages?

We build distribution systems because they are more reliable and efficient that a monolithic system. Horizontal scalability is cheaper than vertical scalability after a certain threshold which also allows for fault tolerance and faster response time from the systems. The advantages and disadvantages of a distributed system is shown below.

Advantages	Disadvantages
Availability as the system is always available to the user.	A distributed system is complex and hard to deploy.
Scalability through horizontal scalability which allows the adding of new servers to the existing resources to meet the requirements	The data may be lost within the network as it moves from node to node.
Performance – The system allows the users to run queries faster due to the database being close to the users since the databases are at various locations.	It is difficult to manage security as the systems/node and the network need to be secured.

8. What applications distributed systems are most suitable for?

A few examples of distributed systems are Netflix, social network websites, online banking, and online gaming where resource sharing, availability, performance, and reliability is the main priority.

9. Why distributed systems are hard?

Designing and implementing a distributed system is hard because the overall system is complex, and many failures can occur during the initial phase in hardware or software. The software written by the developer may not perform the intended way it was designed which may lead to failure. There may be a network failure or glitches which may cause the overall system to fail.