A cache is primarily used to improve:

**A)**Latency

**Explanation**

Caches are used to bring the data closer to where it is used, thus reducing the latency.

In a distributed environment, a network partition can cause which of the following problems? Select all that apply.

**A)**Split-brain

**B)**Each of the two servers might think that the other one is dead.

**D)**Message or data loss.

**Explanation**

E.g., in Dynamo, when a node is down, its data is written to an intermediate node, and if that intermediate node dies before transferring the data to the correct node, the data will be lost.

What is the benefit of using Piggybacked acknowledgments?

**B)**It efficiently uses the available channel bandwidth, thus improving the efficiency of the system.

**Explanation**

Piggybacking saves bandwidth as it removes the need of sending a separate acknowledgment message.

The CAP theorem tells us that:

**C)**Any distributed system which is highly-available will have problems with consistency.

**Explanation**

CAP theorem can be summarized as: “In the presence of a network partition, a distributed system must choose either Consistency or Availability.”

Following is the property of the Write-ahead Log (WAL). Select all that apply.

**A)**Data is always sequentially appended to the WAL.

**B)**Each log entry in the WAL has a unique identifier.

**c)**WAL does not allow updating the data once written.

As opposed to locks, leases:

**A)**Have expiration times

**Explanation**

Leases are just locks that expire.

What advantage does it give to a system to have separate read and write locks?

**A)**Enables concurrent reading of a resource by multiple entities

**Explanation**

Since read locks keep out writers, multiple readers can access a resource as no one will be modifying it.

For GFS, what are the benefits of a large chunk size? Select all that apply.

**B)**Large chunk size provides highly efficient sequential reads and appends of large amounts of data.

**C)**A large chunk size simplifies ChunkServer management.

**Explanation**

I.e., to check which ChunkServers are near capacity or which are overloaded.

**D)**A large chunk size reduces the size of the metadata stored on the master.

Consistent hashing ensures that:

###### C)Only a small set of keys move when servers are added or removed.

Dynamo uses virtual nodes for which of the following reasons? Select all that apply.

**A)**To avoid hotspots

**B)**To spread the load more evenly across the physical nodes on the cluster by dividing the hash ranges into smaller subranges

###### D)To better maintain a cluster containing heterogeneous machines

Of the following options, which is not a characteristic of Amazon’s Dynamo?

###### C)Strongly Consistent

Unlike BigTable, with Amazon Dynamo:

**C)**Two processes may write conflicting updates.

**Explanation**

Multiple processes may end up writing conflicting values to the same key with Dynamo. Vector timestamps identify concurrent updates.

Virtual nodes in Amazon Dynamo are designed to:

**C)**Improve load distribution when adding or removing nodes

**Explanation**

A newly available node accepts a roughly equivalent amount of load from each of the other available nodes.

Which of the following are in-memory structures where Cassandra buffers write?

###### D)Memtables

What are the disadvantages of using Tombstones in Cassandra? Select all that apply.

**A)**Slower reads

**Explanation**

When a table accumulates many tombstones, read queries on that table could become slow and can cause serious performance problems like timeouts.

**B)**Deleted data takes more storage until Tombstones are completely removed during compaction.

# Quiz II

Test your knowledge of different systems that you have learned in this course by completing this quiz.

In Google File System (GFS), when a client needs to access data, it contacts the master and gets the data from the master.

**B)**False

**Explanation**

The client gets the chunk location from the master and then reads data directly from ChunkServers.

The main responsibility of the GFS master is:

**D)**Store all the filesystem metadata, like file names, total parts, locations of the parts, etc.

**Explanation**

Metadata is stored at GFS master, whereas the file contents are stored across ChunkServers.

When scaling systems, mostly distributed system give up:

**D)**Data consistency in favor of availability

**Explanation**

Mostly, distributed systems choose higher availability and eventual consistency.

The motivation for an eventual consistency model was:

**Your Answer**

**A)**It is impossible to have highly available replicated data that is fully consistent in a system that can survive network partitioning.

**Explanation**

CAP theorem: if you want consistency, availability, and partition tolerance, you can get at most two out of three. The eventual consistency model favors high availability and partition tolerance – at the expense of consistency.)

In BigTable, a Tablet refers to:

###### B)A contiguous range of rows of a table

Split-brain is a condition that arises when:

###### B)A network is partitioned.

The main difference between the Google File System (GFS) and the Hadoop Distributed File System (HDFS) is that HDFS:

**D)**In HDFS, there are no concurrent writers and arbitrary file modifications.

**Explanation**

Contrary to GFS, multiple writers cannot concurrently write to an HDFS file. Furthermore, writes are always made at the end of the file, in an append-only fashion. There is no support for modifications at arbitrary offsets in a file.

Distributed systems use the following techniques for Fencing. Select all that apply.

###### A)Resource Fencing

###### C)Node Fencing

In distributed systems, Fencing is used to:

**C)**Isolate a faulty server to prevent it from doing any damage or causing corruption.

**Explanation**

By means of Fencing, we take a faulty server out of service.

A column family in BigTable is:

**Your Answer**

**A)**A group of columns within a single table.

**Explanation**

A column family is a storage unit for multiple columns in a table – usually related in function. From a performance point, data in a row is retrieved by column families.

Dynamo uses vector clocks to:

**C)**Determine causality between different updates on the same object.

**Explanation**

Dynamo uses vector clocks to find causality for concurrent updates on an object.

In distributed systems, replication gives following benefits. Select all that apply.

###### A)Availability

###### C)Reliability

###### D)Better read performance

In distributed systems, Heartbeat is used to:

###### A)Detect dead servers

What statements are true about the Google File System (GFS)? Select all that apply.

**B)**GFS can store multiple replicas of a file’s data on different servers.

**C)**Any file stored in GFS can be bigger than the capacity of a server.

What is the importance of choosing a good row key in BigTable?

**A)**Take advantage of the locality of adjacent rows.

**Explanation**

BigTable keeps its data sorted by the row key. Rows are stored next to each other in a tablet – except occasionally when we have to jump to the next tablet.