HW4: GFS

Date: 2022-10-12

Name: 陈英发

Student ID: 2022280387

Part 1

Q1

How does the master node get the locations of each chunks at startup?

Answer:

It asks each chunkserver for this information at startup or whenever a chunkserver joins the cluster.

Q2

What is the benefit of this approach comparing with the approach that the master persists this information?

Answer:

Firstly, this lessens the storage requirements for the master, it scales very slowly with the number of chunkservers. Moreover, the location of chunks is logically intertwined with the chunkserver, so it is natural for the chunkservers to manage such metadata. Also, the set of active chunkservers in the cluster can change at any time, so if the master persists this information, it will be difficult to maintain the consistency of the information.

Part 2

Assume in a cluster of GFS of 1000 servers. Each server has 10 disks with 10TB storage capacity and 100MB/s I/O bandwidth for each disk. The ethernet that connects servers has bandwidth of 1Gbps.

Q1

What is the minimum time required to recovery a node failure (i.e. distribute its replica to other survived server nodes)?

Answer:

Assume that all replica are distributed as soon as possible, so the I/O bandwith of the failure node is being fully used, and the minimum time required is therefore 1TB / 100MB/s = 10,000s = 2.78h.

Q2

For quality of service, usually the recovery traffic is throttled. If the bandwidth used for recovery is 100Mbps per machine, what is the roughly me required to recover a failure node?

Answer:

1TB / 100Mbps = 80000s = 22.2h.

Q3

Assume the server node has 10000 hours MTBF. How many server failures is likely to have in a year in this cluster? What is the mean time between node failure in this cluster?

Answer:

1 year has about 8766 hours. So each node is expected to have 8766 / 10000 = 0.8766 failure in a year. The expected number of failures in a year is 0.8766 * 1000 = 876.6. The MTBF for the entire server is 8766 / 876.6 = 10 hours.

Q4

Comparing the time you got from Q2 and Q3, what is the implication number of replicas that used in GFS?

Answer:

Since the time required to recover a failure node 2.2x the MTBF, it is reasonable to have at least 3 replicas for each chunk such that after a node failure, we will tolerate 2 more failures before the chunk is lost, during which time we will have been able to finish node failure recovery.