hw4

Tuesday, 16 March 2021 09:36

hw4

Homework 4: GFS

Introducon to Big Data Systems course

Due: March 21, 2021 23:59 China me. Late submission results in lower (or even no) scores.

For quesons or concerns, contact TA (Jiping Yu) by WeChat. Or send an email to yjp19@mails.tsinghua.edu.cn if you could not use WeChat.

Overview

Read the GFS paper and answer the following questions.

Submit a PDF report to Tsinghua web learning, or by email if you can't access web learning.

Part 1

The master stores three major types of metadata: the file and chunk namespaces, the mapping from files to chunks, and the locations of each chunk's replicas. While the first two type of data are persisted by master, the locations of each chunk are not persisted in the master side.

Q1

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

At startup, the master node polls each chunkserver to get the chunk locations.

Q2

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

The benefit is that resources do not have to be spent sychronising the master node and the chunkservers regarding chunk locations, since many events can occur at the chunkservers that will result in a change in chunk locations that are not within the master node's control.

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)?

Need to replicate and redistribute $10 \times 10TB = 100TB$ across remaining 999 servers.
Includes comms betw servers and writing into disks.
For min time, assume all 999 servers and all disks share load equally.

$\therefore$ recovery time = server comms time + disk write time

$= \frac{100TB/999}{1Gbps} + \frac{100TB/999/10}{100MBps} = \frac{100 \times 10^{12}/999}{125 \times 10^6} + \frac{100 \times 10^{12}/9990}{100 \times 10^6}$

$= 800.80 + 100.10$

$= 900.9s$

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 time required to recover a failure node?

Server comms. bandwidth for recovery now decreased from 1Gbps to 100Mbps

$\frac{100 \times 10^{12}/999}{\frac{100}{8} \times 10^6} + \frac{100 \times 10^{12}/999}{100 \times 10^6} \div 10 = 8108.1s$

$= 2.252\ h$

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?

Mean time betw node failure in cluster $= \frac{\text{MTBF of one node}}{\text{no. of nodes in cluster}} = \frac{10k}{1000} = 10h$

Q4

No. of server failures in a year in cluster $= \frac{\text{No. of hrs in a year}}{\text{Mean time betw node failure}} = \frac{365 \times 24}{10} = 876$

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

Q2: Long recovery time of $\sim 2h$

Q3: High failure rate of once per 10h on avg

Since 2h is close to 10h, the chance that {before one replica has been created, another failure involving the other replica would occur} is not very small. In this case, 2 replicas would be insufficient.

So GFS choice of 3 replicas good.