# Distributed Storage and Parallel Computing

Homework 1

12112627 李乐平

**1.1** Suppose we stored a file with size 1.1 GB in linux, where the block size of HDFS is 128MB. How large is the usage of disks caused by the storage of this file? Please analyze the possible usage of the disks.

**Solution:**

#Block = #Replica * Ceil(1.1 GB / Block Size) = 27

Metadata Usage = #Block * Unit Metadata Usage (of average 150B) << 1MB

Usage = File size + Metadata Usage ⪆ 3.3GB

**1.2** How do you think the NameNode can be high-available? How do the state-of-the-arts work?

**Solution:**

High-availability (HA) is designed to eliminate the Single Point of Failure (SPOF) of the NameNode. To achieve HA, one or more 「Standby NameNode(s)」 is setup instead of the SecondaryNameNode. The original NameNode is referred to as the 「Active NameNode」.

The Standby NameNode is expected to perform the same tasks as SecondaryNameNode, including recording EditLog and updating FsImage. In the event of an unexpected Active NameNode failure, Secondary NameNodes (if multiple) will vote for a new Active NameNode, which is typically facilitated by a ZooKeeper cluster. However, it is important to note that the metadata on SecondaryNameNode typically lags behind that of the Active NameNode, so new technology may be required to address this issue.

One solution to the lag problem is to have all NameNodes share the same EditLog. Only the Active NameNode has write access to the EditLog, while Standby NameNodes can only read from it. There are 2 ways to achieve that: NFS (Network File System) File Sharing and QJM (Quorum Jornal Maneger). QJM is generally preferred since it can help mitigate the split-brain problem (multiple Active NameNodes) to some extent.

Moreover, a NameNode Federation can be established to fit arbitrarily large-scale DataNodes. Within a Federation, NameNodes can be configured for high-availability, and a pair or a cluster of Active/Standby NameNodes that can communicate with each other are referred to as a Namespace. Each Namespace maintains its own content tree and is isolated from other Namespaces, while all Namespaces share the same DataNodes.

**1.3**： Suppose there are 3 files being stored in HDFS, where file A is 1GB, file B is 500GB, file C is 328GB. The replica factor is by default. The block size is 128MB. Suppose after writing all the files into HDFS, there are two datanodes are corrupted.

**1.3.1** How many data blocks are needed for storing these files?

**Solution:**

A -> 8 blocks; B -> 4000 blocks; C -> 2624 blocks;

Total = 3 * (8 + 4000 + 2624) blocks = 19896 blocks

**1.3.2** What is the minimal number of racks that are needed to keep the redundancy storage of「Never lose all data if entire rack fails」?

**Solution:**

2.

**1.3.3** What is size of the network traffic for completing all the above events?

**Solution:**

Because there is lack of information about the data of corrupted Datanodes, we cannot infer how much network traffic will cost to the recovery of data.

Network Traffic = Data Size * Replication + Recovery Traffic = 2487 GB + Unknown size

$\gtrapprox$ 2487GB