CONTRASTING REPLICATION AND ERASURE CODING

CS555 ASSIGNMENT1

PRIYANK BAMBHROLIA

Abstract

The ever-increasing storage need of the data humankind produces collectively is increasing day-by-day at an exponential rate. While the storage of data is trivial, the correct retrieval of the stored data is equally trivial. We will examine some aspects of two technologies used to achieve this namely, *Erasure Coding (Reed-Solomon)* and *Replication.* For the purposes of this report, we will assume that the replication level is 3.

AVAILABILITY

Reed-Solomon encoding generates the data which is half the size of original data, after appending it, the resultant data is split into shards. These shards are in turn stored on different disks. Out of these shards even if 1/3rd of the shards is corrupted, the original data can be recovered. While if we replicate data into 3 disks, the original data can be recovered even if 2/3rd of the data gets corrupted, the original data can be recovered.

STORAGE

Storage refers to Storage Space required and Redundancy. The Reed-Solomon Erasure Coding encodes the data into parity bits to achieve fault-tolerance, while Replication simply replicates the data. The encoded parity bits in Reed-Solomon are half the size of actual data. These bits are appended to the data and then split into shards and stored on different disks. Thus Reed-Solomon only takes 1.5 times the storage space of actual data, while Replication takes 3 times the storage space.

COMPUTATION COST

Reed-Solomon encodes the data into parity bits while storing and decodes it back while retrieving to generate missing shards. With increasing amount of data, the processing overhead also increases. On the other hand, there is no processing required for replication, the data is simply replicated.

TIME CONTRAINT

Since there is a processing overhead while encoding and decoding, Reed-Solomon Erasure Coding also take much more time than Replication. The latency of this processing is introduced while storing and fetching data. On the contrary, Replication does not introduce any latency other than storing and retrieving latency.

TRANSPARENCY

Erasure Coding is performed on client side, thus reducing transparency. Client also has encode and split data while storing and decode and merge data while retrieving. The Client is also responsible of storing the shards created after encoding on different disks. While, in Replication the redundancy can be transparent from the Client.

BANDWIDTH

Each time a read or save operation is performed, Reed-Solomon fetches or sends more data in form of parity bits in addition to original data. While in Replication only the original data is send and fetched from the disk.