DOCTORAL THESIS

ESetStore: an erasure-coding based distributed storage system with fast data recovery

Liu, Chengjian

Date of Award:
2018

Abstract

The past decade has witnessed the rapid growth of data in large-scale distributed storage systems. Triplication, a reliability mechanism with 3x storage overhead and adopted by large-scale distributed storage systems, introduces heavy storage cost as data amount in storage systems keep growing. Consequently, erasure codes have been introduced in many storage systems because they can provide a higher storage efficiency and fault tolerance than data replication. However, erasure coding has many performance degradation factors in both I/O and computation operations, resulting in great performance degradation in large-scale erasure-coded storage systems.

In this thesis, we investigate how to eliminate some key performance issues in I/O and computation operations for applying erasure coding in large-scale storage systems. We also propose a prototype named ESetStore to improve the recovery performance of erasure-coded storage systems. We introduce our studies as follows.

First, we study the encoding and decoding performance of the erasure coding, which can be a key bottleneck with the state-of-the-art disk I/O throughput and network bandwidth. We propose a graphics processing unit (GPU)-based implementation of erasure coding named G-CRS, which employs the Cauchy Reed-Solomon (CRS) code, to improve the encoding and decoding performance. To maximize the coding performance of G-CRS by fully utilizing the GPU computational power, we designed and implemented a set of optimization strategies. Our evaluation results demonstrated that G-CRS is 10 times faster than most of the other coding libraries.

Second, we investigate the performance degradation introduced by intensive I/O

operations in recovery for large-scale erasure-coded storage systems. To improve the recovery performance, we propose a data placement algorithm named ESet. We define a configurable parameter named overlapping factor for system administrators to easily achieve desirable recovery I/O parallelism. Our simulation results show that ESet can significantly improve the data recovery performance without violating the reliability requirement by distributing data and code blocks across different failure domains.

Third, we take a look at the performance of applying coding techniques to in-memory storage. A reliable in-memory cache for key-value stores named R-Memcached is designed and proposed. This work can be served as a prelude of applying erasure coding to in-memory metadata storage. R-Memcached exploits coding techniques to achieve reliability, and can tolerate up to two node failures. Our experimental results show that R-Memcached can maintain very good latency and throughput performance even during the period of node failures.

At last, we design and implement a prototype named ESetStore for erasure-coded storage systems. The ESetStore integrates our data placement algorithm ESet to bring fast data recovery for storage systems.

Keywords: Erasure coding, Storage System, ESet, R-Memcached, ESetStore

Table of Contents