CS554 Project Ideas

FusionFS:IDA - Towards Storage-Efficient Data Reliability in Distributed File Systems

Overview

Data replication is the state-of-the-art technique to achieve high availability in distributed systems. The major issue of this method is the low space efficiency. To address that, we have implemented a software RAID with existing Erasure Coding libraries [1] (also known as information dispersal algorithms) into FusionFS [2]. However, these libraries are not optimized for distributed file systems. In this project you will design and implement new information dispersal algorithms, and integrate them into FusionFS. We hope to achieve better performance than the current algorithms. The implementation will be merged to the next release of FusionFS.

Relevant Systems and Reading Material

Please read the following papers (and their references) before submitting your proposal:

[1] Plank, James S. and Luo, Jianqiang and Schuman, Catherine D. and Xu, Lihao and Wilcox-O'Hearn, Zooko. A performance evaluation and examination of open-source erasure coding libraries for storage, *7th conference on File and storage technologies*, 2009. Available online: http://dl.acm.org/citation.cfm?id=1525927

[2] Dongfang Zhao, Kent Burlingame, Corentin Debains, Pedro Alvarez-Tabio and Ioan Raicu. Towards High-Performance and Cost-Effective Distributed Storage Systems with Information Dispersal Algorithms, *IEEE International Conference on Cluster Computing*, 2013. Available online: http://datasys.cs.iit.edu/~dongfang/download/IDA_Storage_crc.pdf

Preferred/Required Skills

Principles: operating system, distributed systems, computer network, RAID disks

Programming: Shell Script, Perl/Python, C, C++, PThread, sockets, FUSE

Operating systems: Linux

Project Mentor

Dongfang Zhao

Email: dzhao8@hawk.iit.edu