

CS554 Project Ideas

FusionFS:FHFS: Improving Hadoop through FusionFS

Overview

HDFS [1] is one of the most popular distributed file system for cloud computing. However, one concern has long existed: its metadata is stored in a limited number of metadata servers, which is a potential bottleneck (and point of failure). We have proposed a new distributed file system, namely FusionFS [2], for high-performance computing, and we plan to extend this new architecture to cloud computing. In this project, you are to adopt the HDFS interface and implement it on top of FusionFS. The HDFS-like interface should bypass the FUSE interface currently used in FusionFS. You will then modify Hadoop to operate on top of FusionFS (with the new HDFS-like interface). You will evaluate it with micro benchmarks and real applications, including fine granular workloads. This project will be open-source, and will be merged into the next release of FusionFS.

Relevant Systems and Reading Material

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

[1] Shvachko, Konstantin and Kuang, Hairong and Radia, Sanjay and Chansler, Robert. The Hadoop Distributed File System, *IEEE 26th Symposium on Mass Storage Systems and Technologies*, 2000. Available online: <http://dl.acm.org/citation.cfm?id=1914427>

[2] Dongfang Zhao, Chen Shou, Zhao Zhang, Iman Sadooghi, Xiaobing Zhou, Tonglin Li and Ioan Raicu. FusionFS: a distributed file system for large scale data-intensive computing, *2nd Greater Chicago Area System Research Workshop*, 2013. Available online: http://datasys.cs.iit.edu/~dongfang/download/poster_fusionfs_revised.pdf

Preferred/Required Skills

Principles: operating system, distributed systems, computer network, key-value stores

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

Operating systems: Linux

Project Mentor

Dongfang Zhao

Email: dzhao8@hawk.iit.edu