Cloud storage 可靠性方面的在:software security and reliability

FAST会议：可能是专门讲存储的会议

4月20日

今天所查到的会议：

FAST：File and Storage Technologies

IEEE International Conference on Big Data

International Conference for High Performance Computing, Networking, Storage and Analysis

ACM Symposium on Cloud Computing

IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

IEEE International Parallel and Distributed Processing Symposium

FAST/OSDI/SOSP？是什么

异构式存储

YARN

Hadoop

HDFS

RAMDISK是什么？：RAM盘是通过使用软件将[RAM](https://zh.wikipedia.org/wiki/%E9%9A%A8%E6%A9%9F%E5%AD%98%E5%8F%96%E8%A8%98%E6%86%B6%E9%AB%94" \o "随机存取内存)模拟当做[硬盘](https://zh.wikipedia.org/wiki/%E7%A1%AC%E7%9B%98" \o "硬盘)来使用的一种技术

分布式文件系统：HDFS

可研究的方向

1 Flash寿命的延长

2 RAM，RAM-DISK，

4月21日

问题：

NAS存储是什么？

想法：

分布式存储系统中哈希分布是不是可以在混合式存储上提高效率

NAND FLASH寿命有限，有专门研究FLASH专用文件系统的。

服务器端的存储系统典型的应用可以是videoserver, fileserver, varmail,oltp

A number of studies focus on separating hot and cold data. Wang and Hu [28] proposed to distinguish active and inactive data in the buffer cache, instead of writing them to a single log and separating them during cleaning.

4月27日

HDFS trace, HDFS workload分析？

想法：针对某一类型的应用首先可以分析其工作负荷，找到可以改进的地方。

5月11日

1 第一步查看别人研究大数据的数据来源。可查的论文来源可以说ieee on big data或是与云计算、storage中有关机器学习的论文，统计其来源。

2 第二步查找挖数据的方法。Apache Nutch可以作为数据挖掘的工具。

3 想出新的机器学习的算法，分析数据的存储并提升存储效率。

4 不是先去找benchmark,而是应该先找到数据或是某种应用中存在的特征，而后再提出算法，之后再验证是否成功。

A study of application performance with non-volatile main memory ieee

# AnalyzeThis: an analysis workflow-aware storage system acm

BD-CATS: big data clustering at trillion particle scale acm

# When computing meets heterogeneous cluster: Workload assignment in graph computation ieee

5月12日

5月19日

网络爬虫可以用Python和http://scrapy.org/

http://aljun.me/post/22

明天看聚类分析和、情感语义分析的应用，然后找一个主题。

6月1日

机器学习基础知识：

1 多元统计分析

2 贝叶斯统计

3 非参数统计

4 概率论与数理统计

5 矩阵理论

6月14日

矩阵理论书：矩阵分析(matrix analysis)，矩阵计算(matrix computations)

# 贝叶斯统计书：现代统计研究基础，Bayesian Data Analysis