调研报告——校园规模的分布式文件共享系统

一、项目背景

- 在一所大学中，校园中的同学需要频繁的从网络上下载各种文件。这些文件中可能包含音频文件、影音文件，游戏文件；也有可能诸如电子图书、教学课件的文档文件。由于同一校园内同学们所需要的从网络上下载的资源有很大的相同，因此同学们分别到网上去搜索并下载自己需要的资源肯定不如从一个校园内的文件系统上下载文件有效率。因此需要一个在校园内使用的校园规模的文件共享系统。

二、立项依据

1. 根据上述背景及需求，目前有以下几种解决方案：

- 学校官方搭建资源网站

  可以搭建一个校园官方的资源站，同学们都访问该网站来下载自己所需要的资源。这种解决方案的优点是架构简单，方便搭建；但存在着以下缺点：

  - 开销大

    这种方案需要一个中心化的服务器来存储的数据，而海量的数据需要大量的存储服务器来存储数据。这对学校而言将是一笔不小的开销。

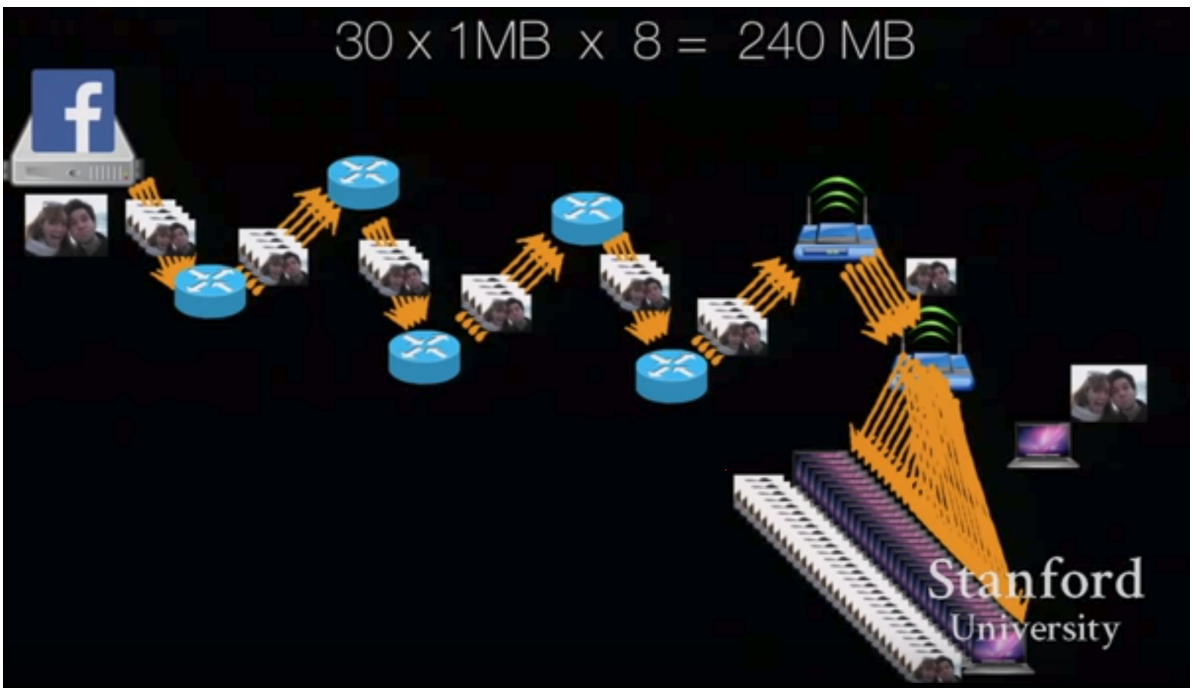
  - 扩展性差

    如果当前的存储容量不够时，需要购入新的服务器来存储数据，可扩展性差

  - 隐私性差

    由于是中心化的文件系统，必然有学校官方的机构来对该系统进行管理和维护，这就意味着该文件系统中的资源完全由学校官方进行监控，隐私性差

  - 对网络带宽要求高

由于是中心化的文件系统，当大量用户同时从这个文件系统下载资源时，网络带宽的压力很大，从而导致下载速度很慢

    - 存在单点故障问题
- 网盘
    - 私密性差

      由于存储在网盘上的资源都被网盘服务提供商所控制，服务商可能查看并监管资源，私密性较差。
    - 传输速度慢

      当前市面上的网盘产品下载速度都比较慢，甚至可能对用户下载进行限速

2. 基于以上几点说明，我们发现理想的用于学生之间资源交流的网络文件系统需要具备以下几个特点：

- 去中心化
- 传输速度快
- 可扩展性高
- 容错性高

3. 框架设计及依据

- 文件存储：每个用户贡献出一定的磁盘存储空间，将文件分块后分散存储到各个用户的硬盘上
    - 纠删码(Erasure Code)对文件进行分块

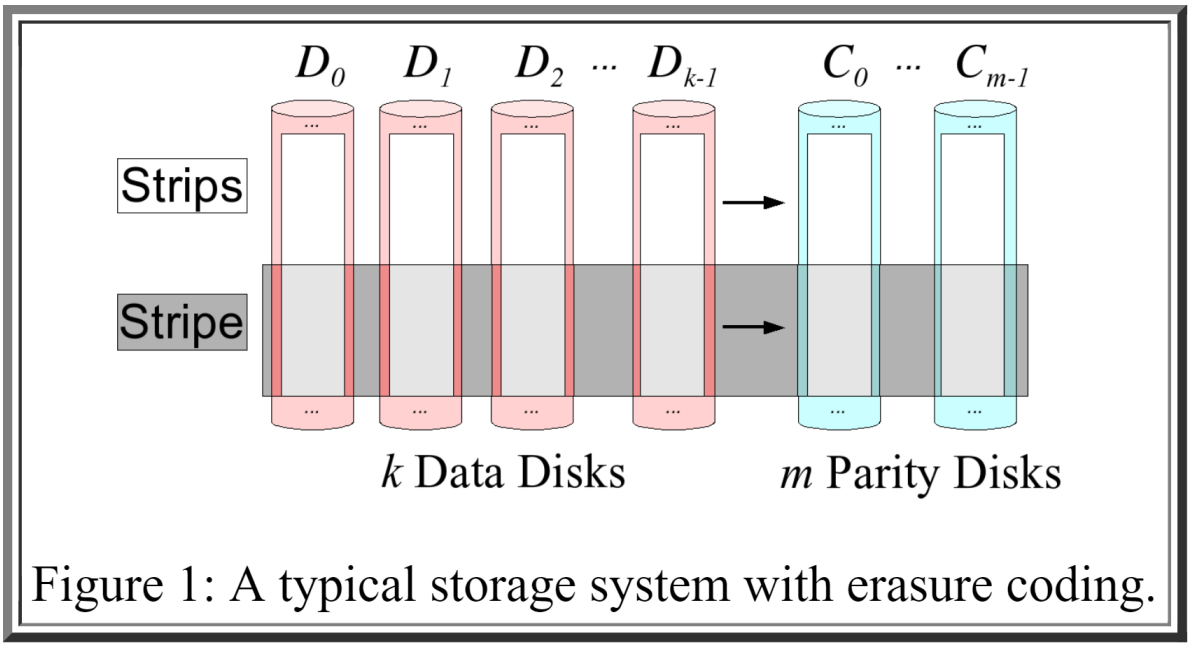


Figure 1: A typical storage system with erasure coding.

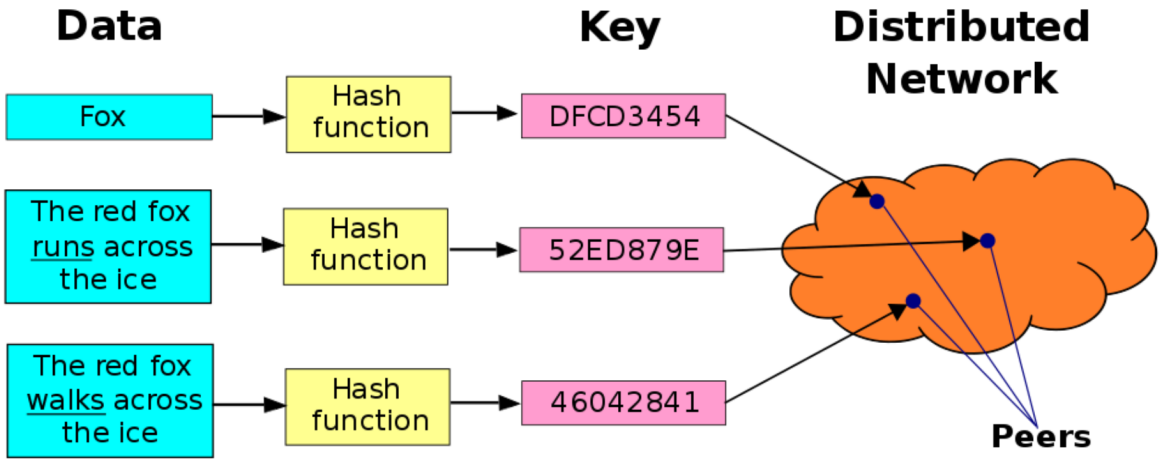
Erasure Coding的主要原理就是将待存储的文件首先分成若干个块，之后利用算法计算出维持存储冗余所需要的额外几块校验块；最后将数据块和校验快分散存储到各个存储服务器上

可以选择存储文件被分割成的存储块的数目以及校验块的数目，从而获得不同的容错性能

当某个存储服务器发生错误时，Erasure Coding可以根据校验块以及当前可用的存储块重建数据并恢复冗余

Erasure Coding通常需要50%的额外空间就能在5~6个磁盘发生错误的情况下恢复文件；这比传统的RAID需要一倍以上的额外存储空间但只能容许1~2个磁盘发生故障性能要强得多

- 分布式哈希表将文件分布存储



DHT的结构可以分为几个部分，其中最重要的基础就是散列值范围。

考虑到DHT的应用场景，其需要承载的数据量通常比较大，为了防止冲突，一般将散列值范围(keyspace)设定得尽可能大；通常采用长度大于128bit的散列值(2^128已经比地球上所有电子文档的总述还要高出好几个数量级，因此是完全够用的)。

每个结点分配一个长为128bit的Node ID

将每一个文件进行hash产生一个Key，并给每个Key分配一个Key ID

void put (KEY k, VALUE v); //保存“键-值对”

VALUE get(KEY k); //根据Key值获取数据

当某个节点得到了加入的新数据的(key, value)后，它会先根据自己的ID和新数据的key计算自己与新数据之间的“距离”，然后再计算他知道的其他节点与这个该数据的距离。

若计算结果表明该节点距该数据最近，就将该数据保存再自己这里；否则将数据发送给距离最小的节点。

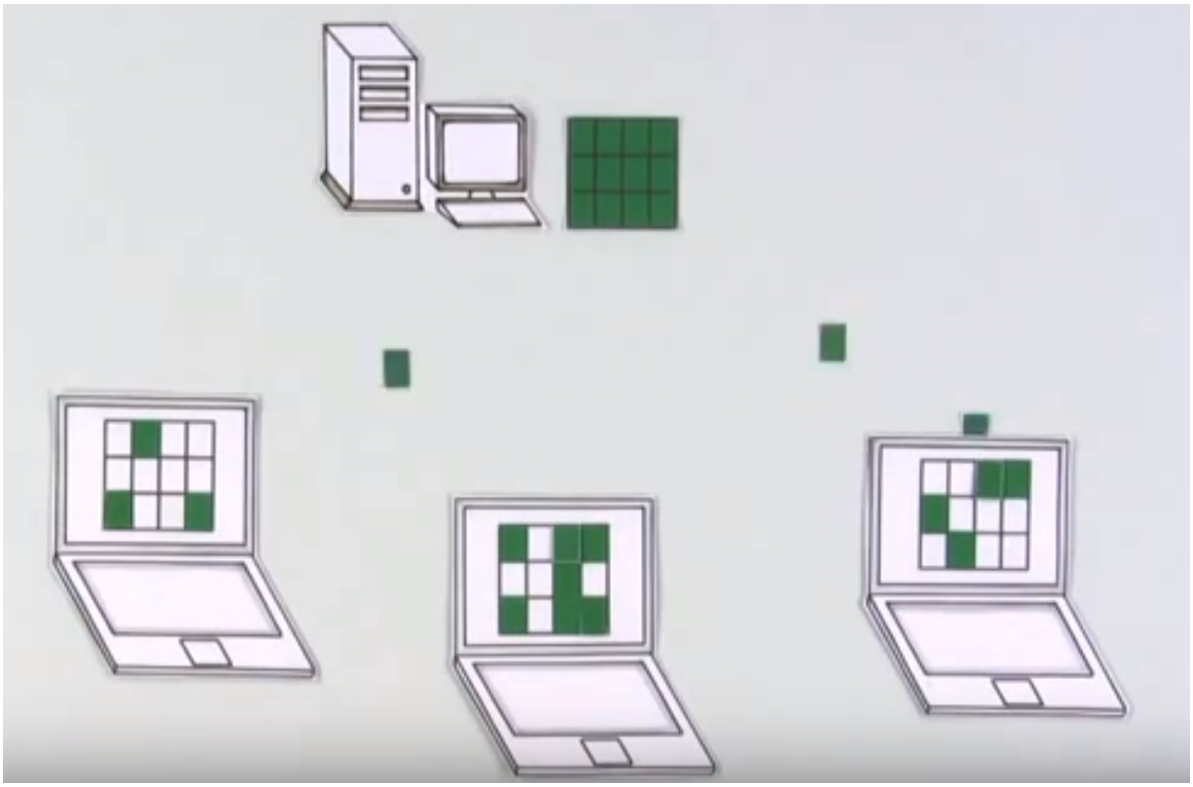
收到该数据的节点重复以上两个步骤直至找到最近的结点

- 文件定位
  - DHT路由算法

    当某个节点接收到查询数据的请求，也即接收到一个key值后，它首先根据自己的ID以及这个key值计算自己与该数据之间的距离，然后在计算它所知道的其他节点与这个key的距离。

    若计算结果表明该结点据该数据最近，该结点就在自己的存储空间中寻找该key值所对应的数据。如果没有找到则报错。

    ③若计算结果表明另一个节点N'据该数据最近，则该结点就把这个key值发送给节点N'。N'在接收到key值之后，重复上述的过程，直至找到文件。
- 文件下载
  - Bit Torrent



用户从Tracker服务器获取到若干其他下载者(peer)的ip和port信息，会进行请求并维持跟每一个peer的连接状态

当一个用户对一个远程peer感兴趣并且那个远程peer没有choke这个用户，那么这个用户就可以从远程peer下载块(block)

第一次通信会先发送握手报文，告诉远程客户端本客户端的一些信息，包括info_hash和peer_id

接下来经过别的一些报文在本地用户和若干个远程用户之间的来回传递，就能够获取到资源文件

三、前瞻性/重要性

- 校园规模的分布式文件共享系统可以满足同学对于资源共享及存储的平台的需求，同时由于其去中心化的设计，一定程度上减少了学校相关机构对同学之间共享的数据被进行监管和控制。
- 这个文件系统特点在于系统中的每个用户既是文件的贡献者也是文件的享有者同时也是文件的存储者，这样的特点使得整个系统具有较好的可扩展性
- 目前传统的中心化文件存储需要投入专用的存储设备，设备的投入和更换代价都很大，并且可扩展性差，可能造成单点故障。可见去中心化的分布式文件系统的重要性

四、相关工作

- 目前主流的分布式文件系统有很多，例如Lustre，HDFS，GFS，NFS等等，但是这些文件系统主要针对的是互联网应用的海量数据存储，同时对网络的带宽和时延要求很高，不适合用于校园网这种带宽较低时延较大的环境
- 国内专门针对校园网设计的专用分布式文件系统主要有Corsair FS，但是Corsair FS的存储服务器集群仍然是由学校相关人员进行维护，并不是一个去中心化的系统

*参考文献

[1] 刘立坤, 武永卫, 徐鹏志,等. CorsairFS:一种面向校园网的分布式文件系统[J]. 西安交通大学学报, 2009, 43(8):43-47.

[2] Stoica I, Morris R, Karger D, et al. Chord: A scalable peer-to-peer lookup service for internet applications[C]// Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM, 2001:149-160.

[3] Lakshman A, Malik P. Cassandra:a decentralized structured storage system[J]. Acm Sigops Operating Systems Review, 2010, 44(2):35-40.