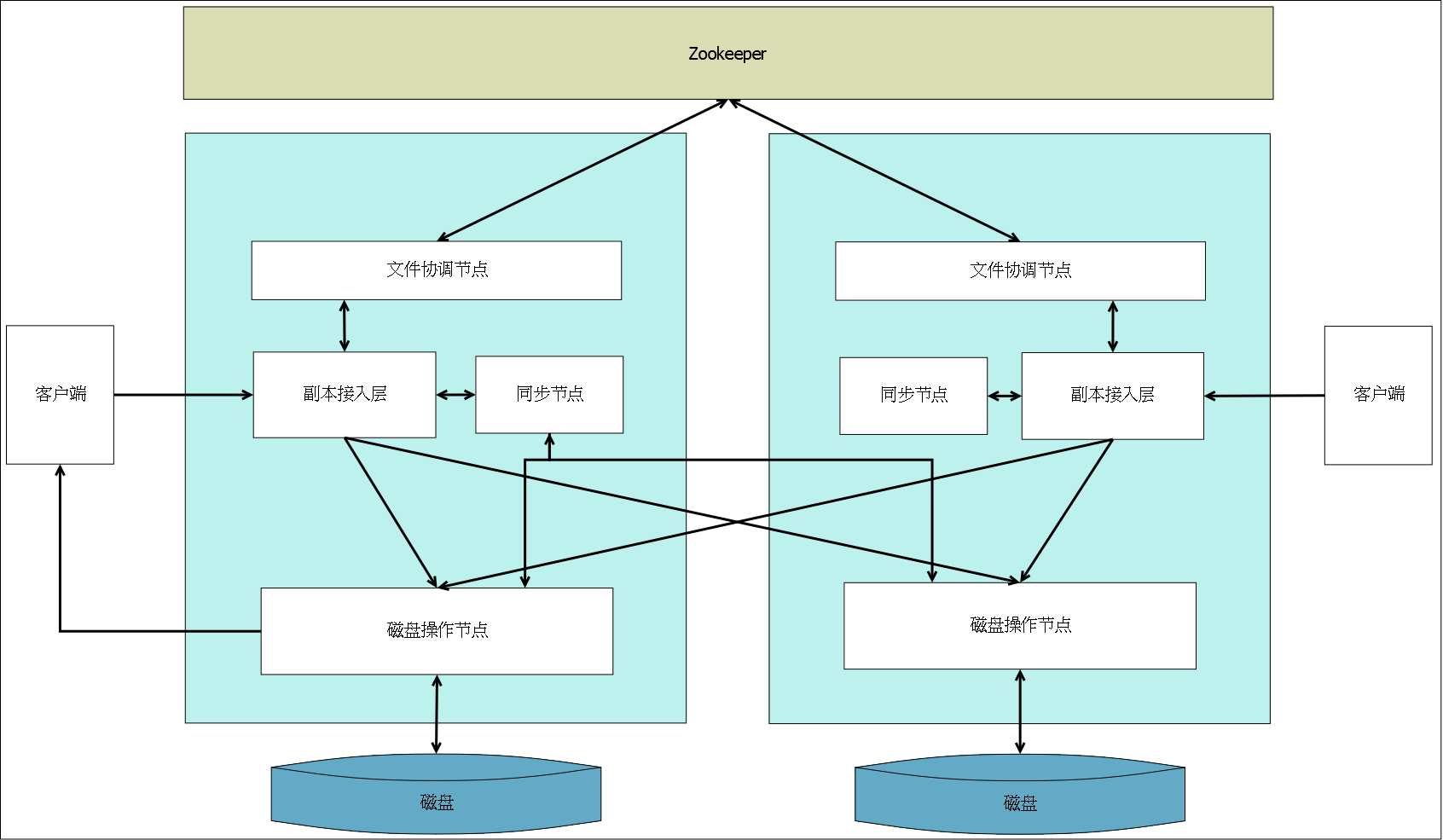
**整体轮廓**



数据写入流程中的三个主要节点：

1、客户端；

2、副本接入节点；

3、磁盘操作节点；

其中每个节点在数据处理上都有自己的职责范围：

（1）客户端：

负责关联StorageName和用户数据，即

（StorageName, UserData）；

（2）副本接入节点：

其负责3种映射关系：

①StorageName到文件名之间的映射，即

（StorageName, [FileName1, FileName2, …]）

②用户数据到文件之间的映射，即

（UserData, FileName）

③文件到主机之间的映射，即

（FileName, [Server1, Server2, …]）

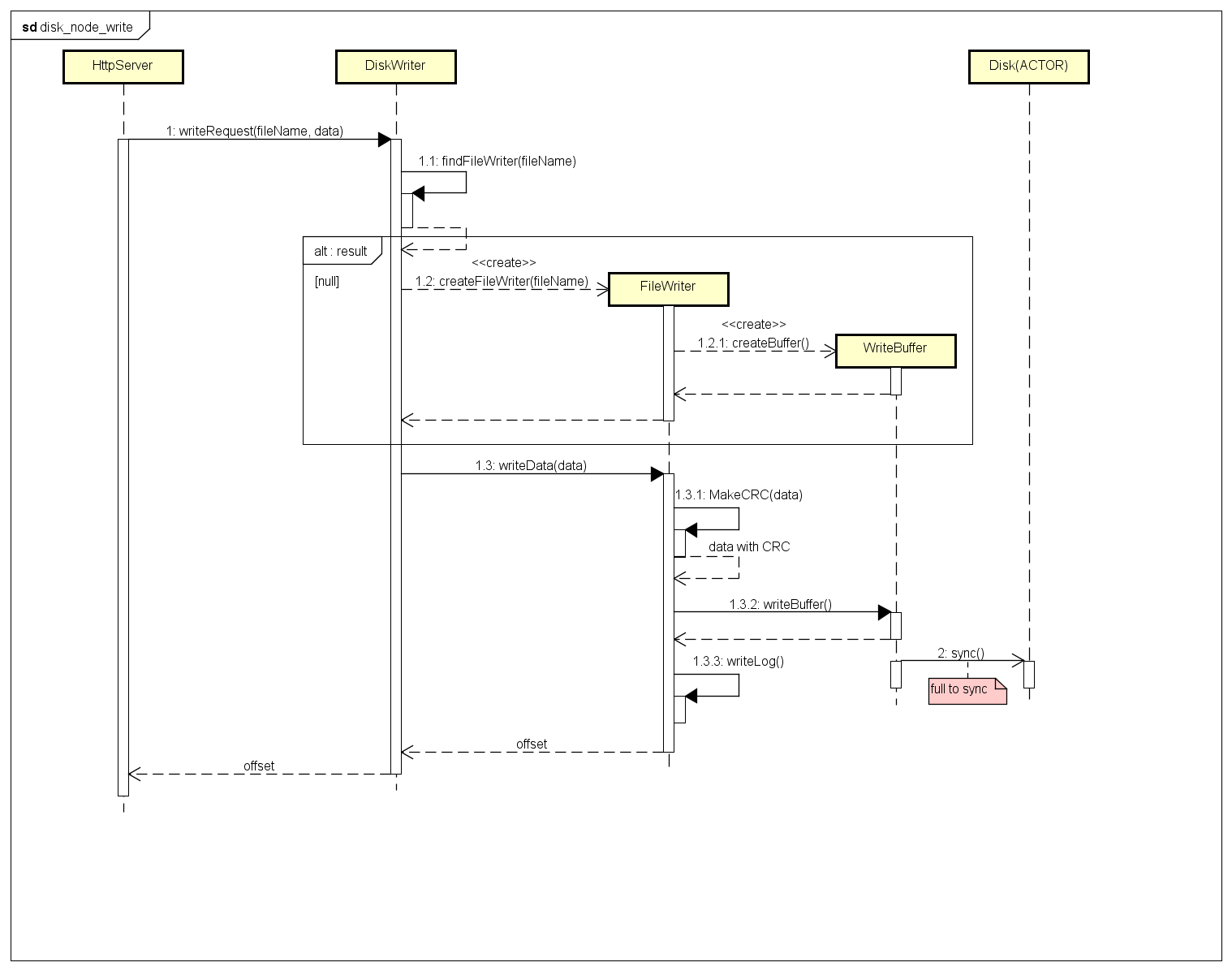
3、磁盘操作节点：

负责文件的读写操作，维护文件的偏移量：

（FileName， offset）

**数据读写流程**

（1）磁盘写数据



1、磁盘写入过程先写入缓存，然后再同步到磁盘，每个缓存大小可配置。经测试，当缓存设置为1M时，本地机器最高写速率可达到50M/s；外网云主机上可达到240M/s。

2、CRC校验和写数据日志是为了保证文件数据的完整性和不同副本间数据一致性的检测和恢复。

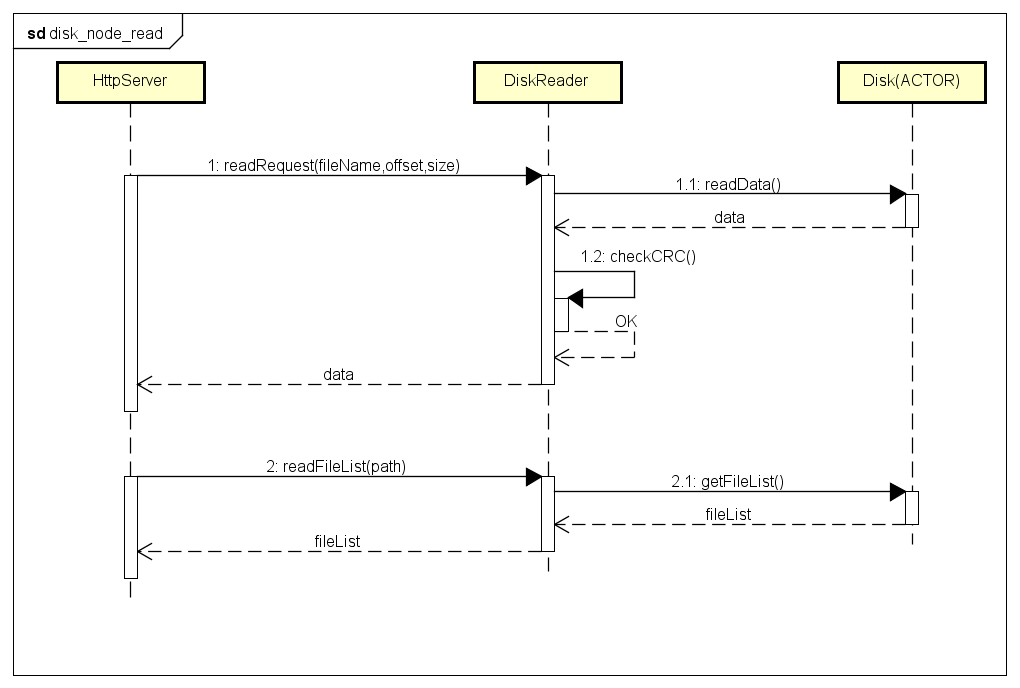
3、每一条用户数据的写入都会相应的在日志中写一条记录，记录格式为：

offset， size， crc

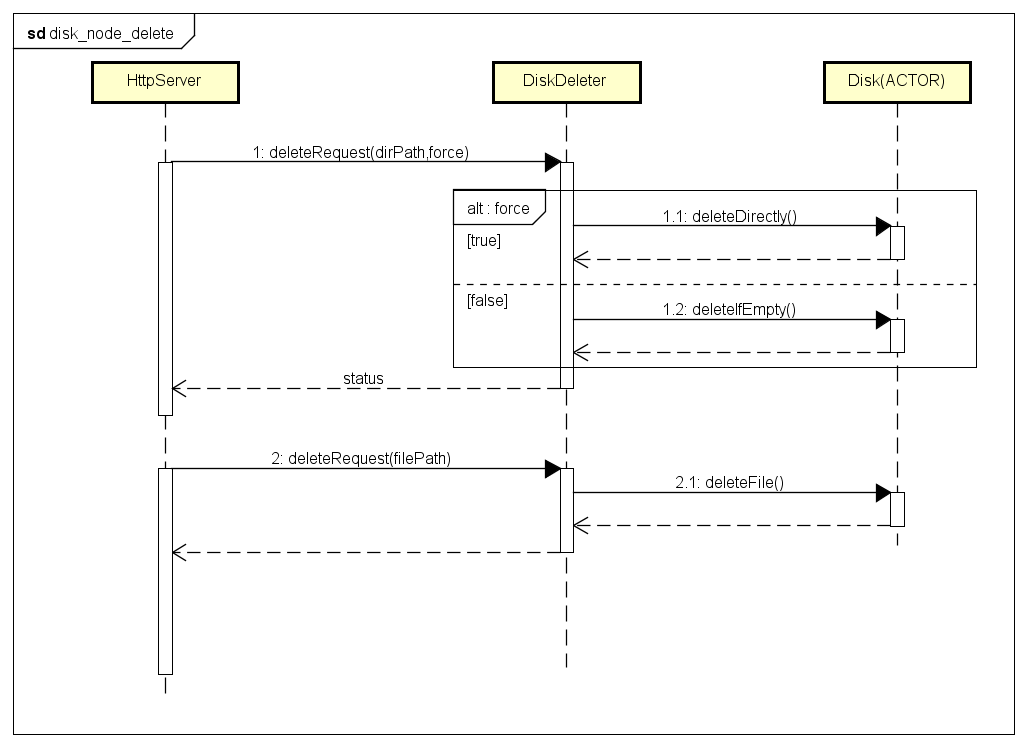
文件的中数据的有效性会根据日志中的记录进行验证。

4、流程图中只展示主要接口，比如文件同步的过程可能还需要文件truncate的接口。

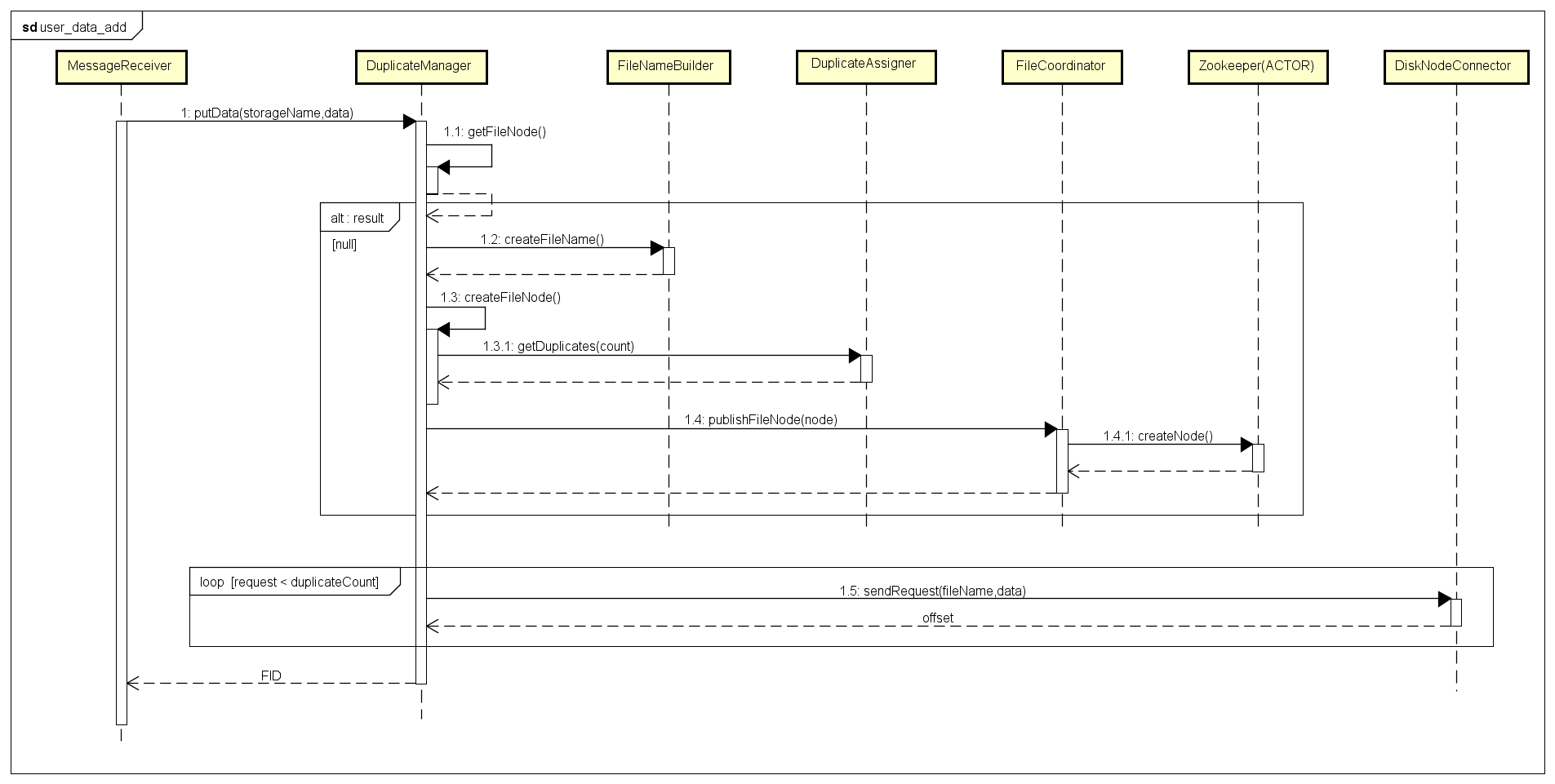
（2）磁盘读文件



（3）磁盘数据删除



（4）副本节点写数据



（5）用户数据读取

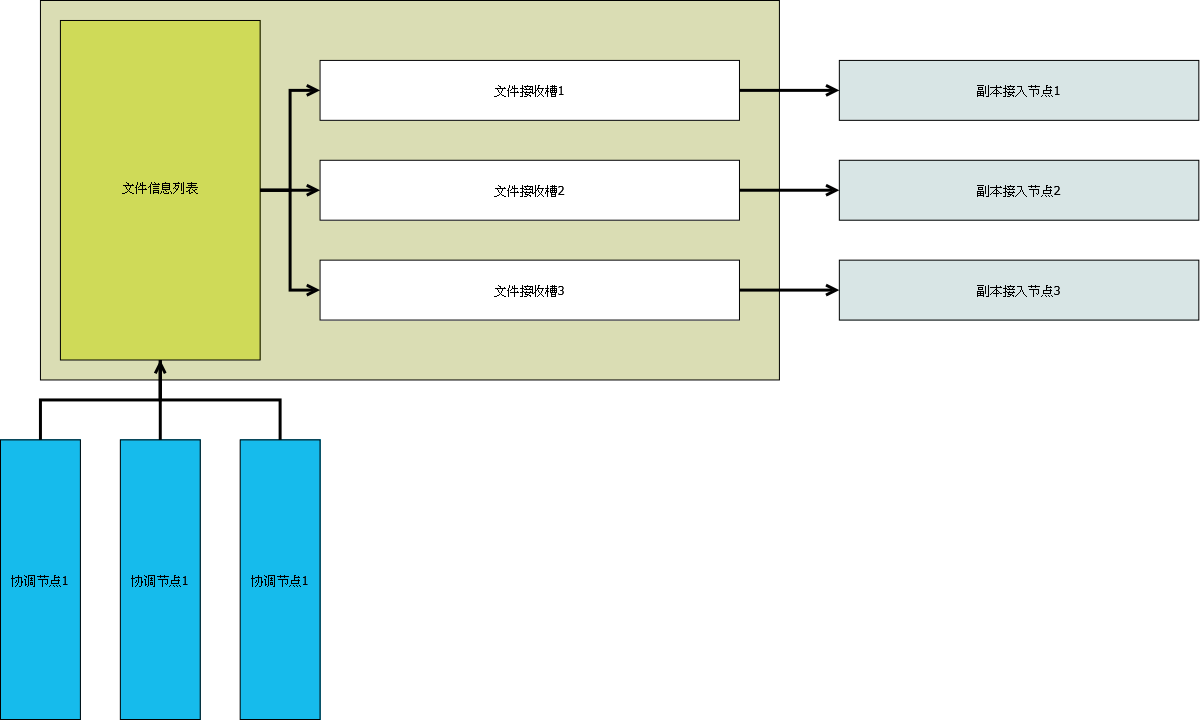
客户端可以直接解析FID，然后根据文件名、偏移量、size直接从磁盘管理节点读取文件内容。

（6）用户数据删除

数据删除目前采用任务提交、执行的方式处理，客户端可以直接提交任务。

（7）文件协调节点

协调节点和副本管理节点配合使用，它们在Zookeeper上的存储结构如下：



1、副本节点负责创建文件，并向协调节点发布文件信息；

2、协调节点收到副本节点传输的文件信息后会把其信息放到zookeeper中，其区域为图中的文件信息列表；

3、当一个文件正常写满并关闭后，副本接入节点会向协调节点发送关闭请求，之后文件从文件信息列表删除；

4、当一个Server出现异常，协调节点会扫描文件信息列表，筛选出异常Server拥有的所有文件，然后根据特定策略将其分配至正常Server对应的文件接收槽（只有文件对应的副本Server至少有一个存活时才会对其进行分配）；

5、正常Server会监听到文件接收槽中的内容变化，从槽中取出文件，放到自己维护的写入列表中。

**文件同步**

同步开始条件：

1、文件写过程中数据不一致，例如Disknode返回的结果不一致；

2、文件对应的某个副本Server恢复正常时；

同步结束条件：

1、文件对应的正常Server上的副本数据达到一致（以文件有效数据最多那个为准）。

同步中的文件是不可写的。同步完成后文件可继续写入。