SACC 2016 第八届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2016

架构创新之路

专场14：高性能存储及文件系统

主持人：徐海峰 阅文集团首席架构师

加入专场微信群，参与群中抢红包活动，可获赠技术图书和SACC2016定制版路由器！

阅文集团自主分布式文件系统

大嘴
xuhaifeng@yuewen.com

当时环境

腾讯文学 Tencent Literature

XINHUA

SDL盛大文学

@TECH2IPO

阅文集团 CHINA READING LIMITED

创世文学

创世中文网 文学精品阅读

SACC 2016 第八届中国系统架构师大会

SequeMedia 盛拓传媒

IT168

ChinaUnix

ITPUB

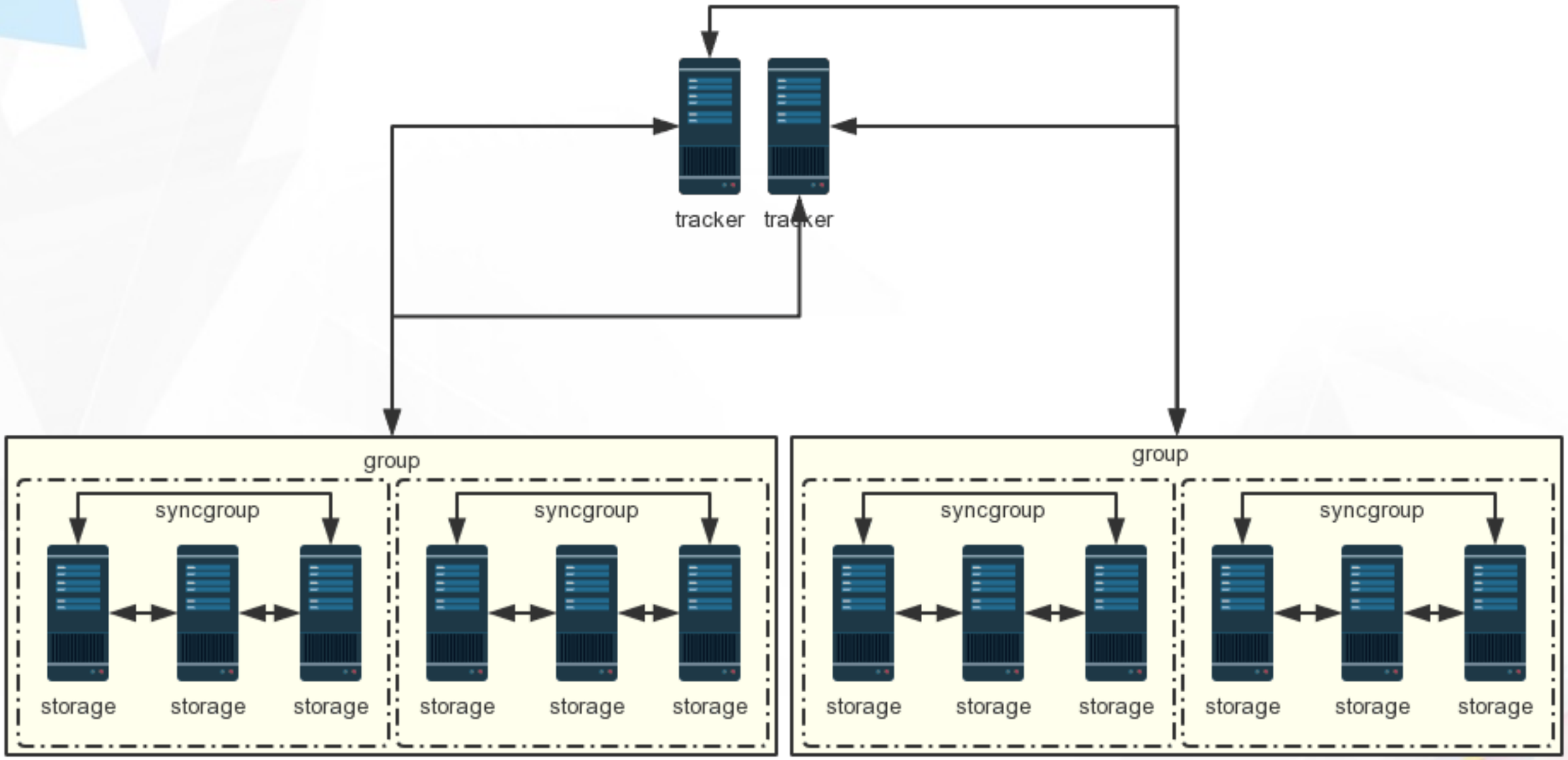
动机

- 需要存储日志、文本、单文件，并且可以提供查询，分析等服务
- 需要一个简单的k-v存储，没有内存和存储的限制
- 存储的扩展性要足够好，尽量简单的扩容操作，数据自平衡
- 特别的业务：小文件，随机IO，短时间操作
- 可支持频繁更改，频繁的数据长度变更
- 高可靠性，无单点故障

寻找

- Fastdfs：
  - chunkfile的版本控制不满意
  - 同group镜像粒度太大，运维也不方便
- CFS（腾讯内部）：
  - 大文件849mb 写 13.777 读 25.45 del 5.565
  - 小文件 478mb 写28m37s
  - 挂载运行，没有管理层
- 剩下的：
  - redis，mongodb…
- 所有问题：DFS对更改支持都不友好

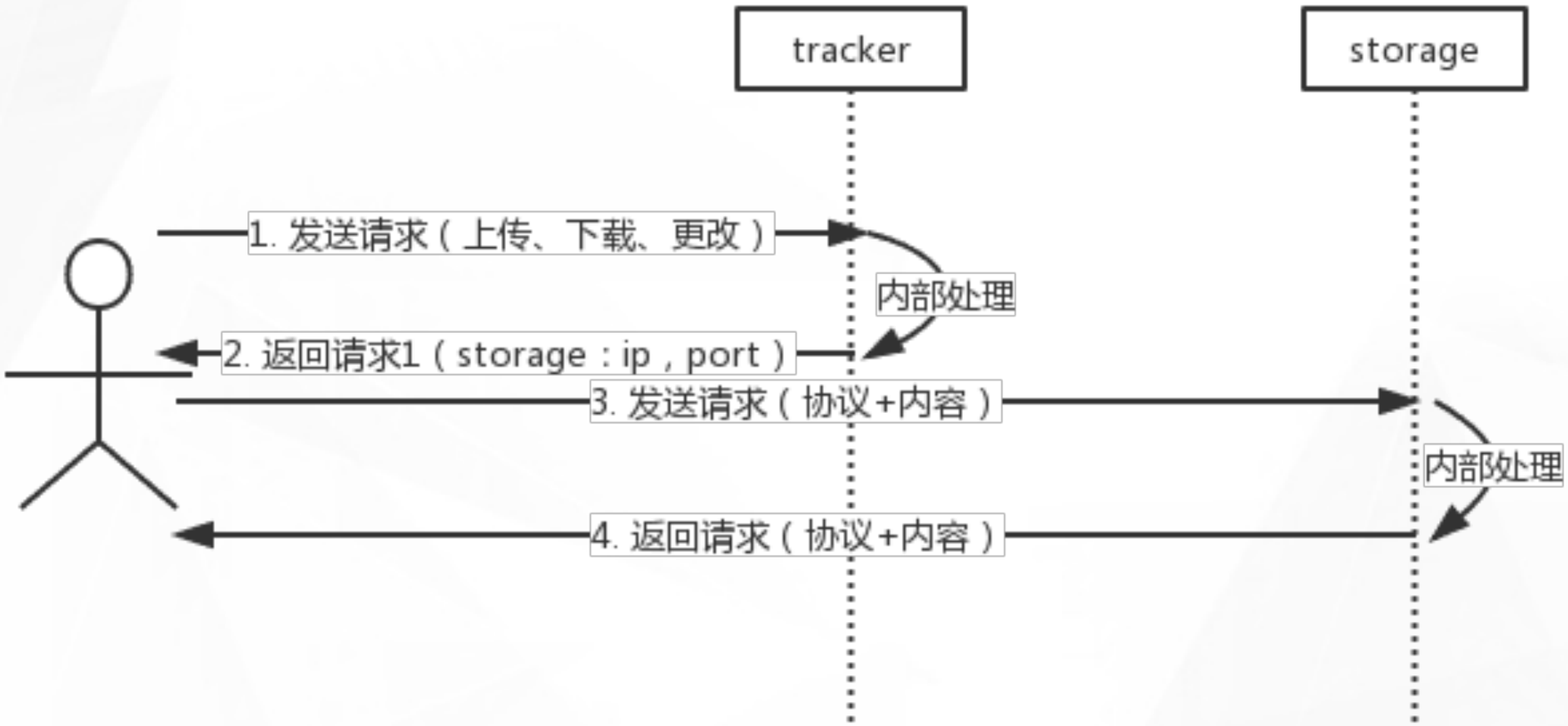
架构


tracker
tracker
group
syncgroup
syncgroup
storage
storage
storage
storage
storage
storage
group
syncgroup
syncgroup
storage
storage
storage
storage
storage
storage

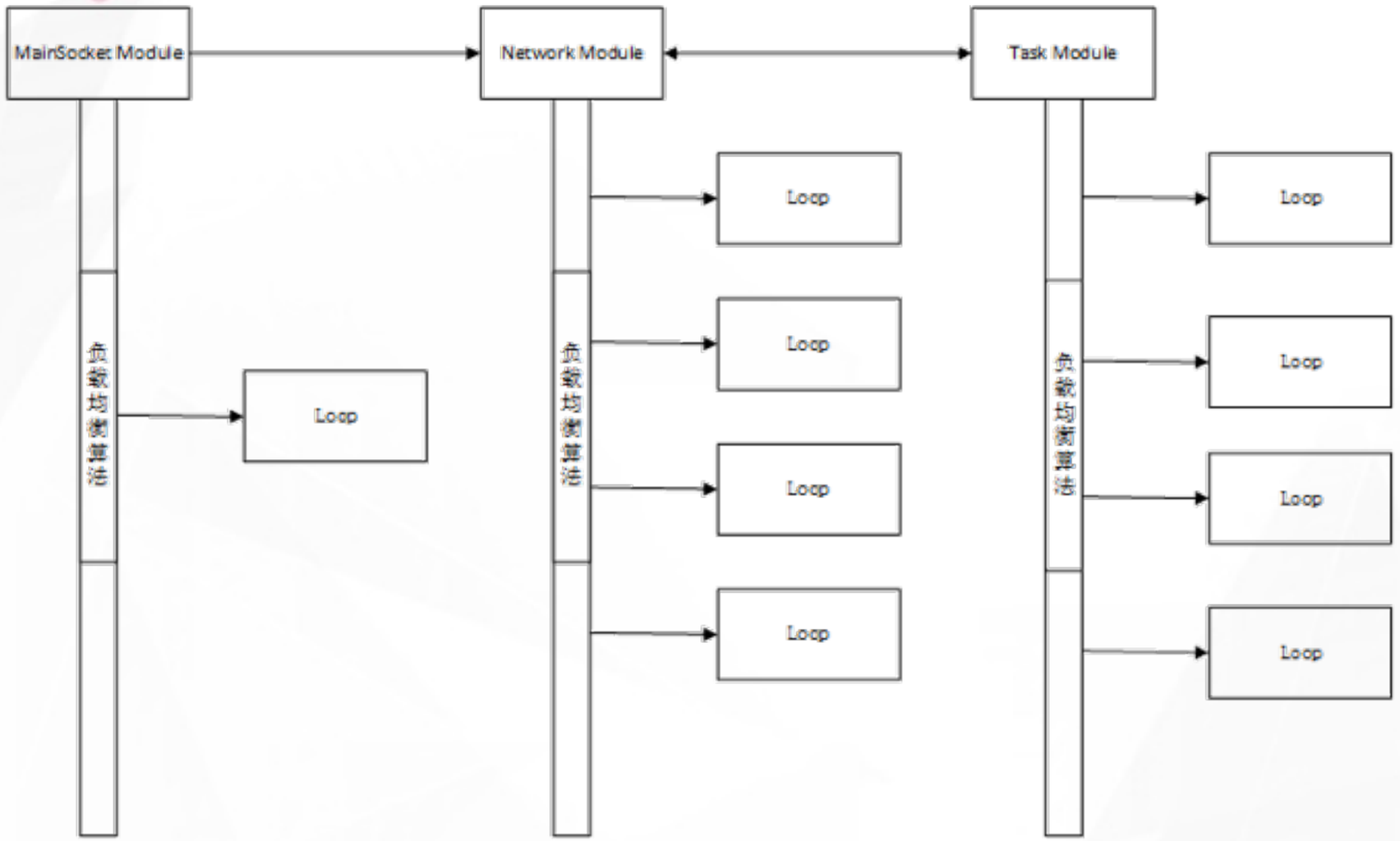
主要功能

- tracker：
  - 对于请求storage的负载均衡
  - 维持心跳状态
  - 请求重定向
- storage：
  - 存储文件
  - 同步文件
  - 磁盘的负载均衡

面对client的流程



进程内部结构



- 清楚的逻辑分层：职责清楚
- 线程调度算法（无锁编程算法）

存储模型

- 文件空洞：长度 * 120
- 多master引申的版本问题
  - 文件版本
  - context版本
  - 操作版本问题
- vector clock算法
  - 每个storage都有一个版本算法
  - 每个操作和磁盘context都有一个单独版本
  - 时间戳也是版本号的一部分
  - 修改的时候，只能靠这个算法维持版本统一
- hashcode控制签名

同步

- 单盘恢复
- 一致性同步

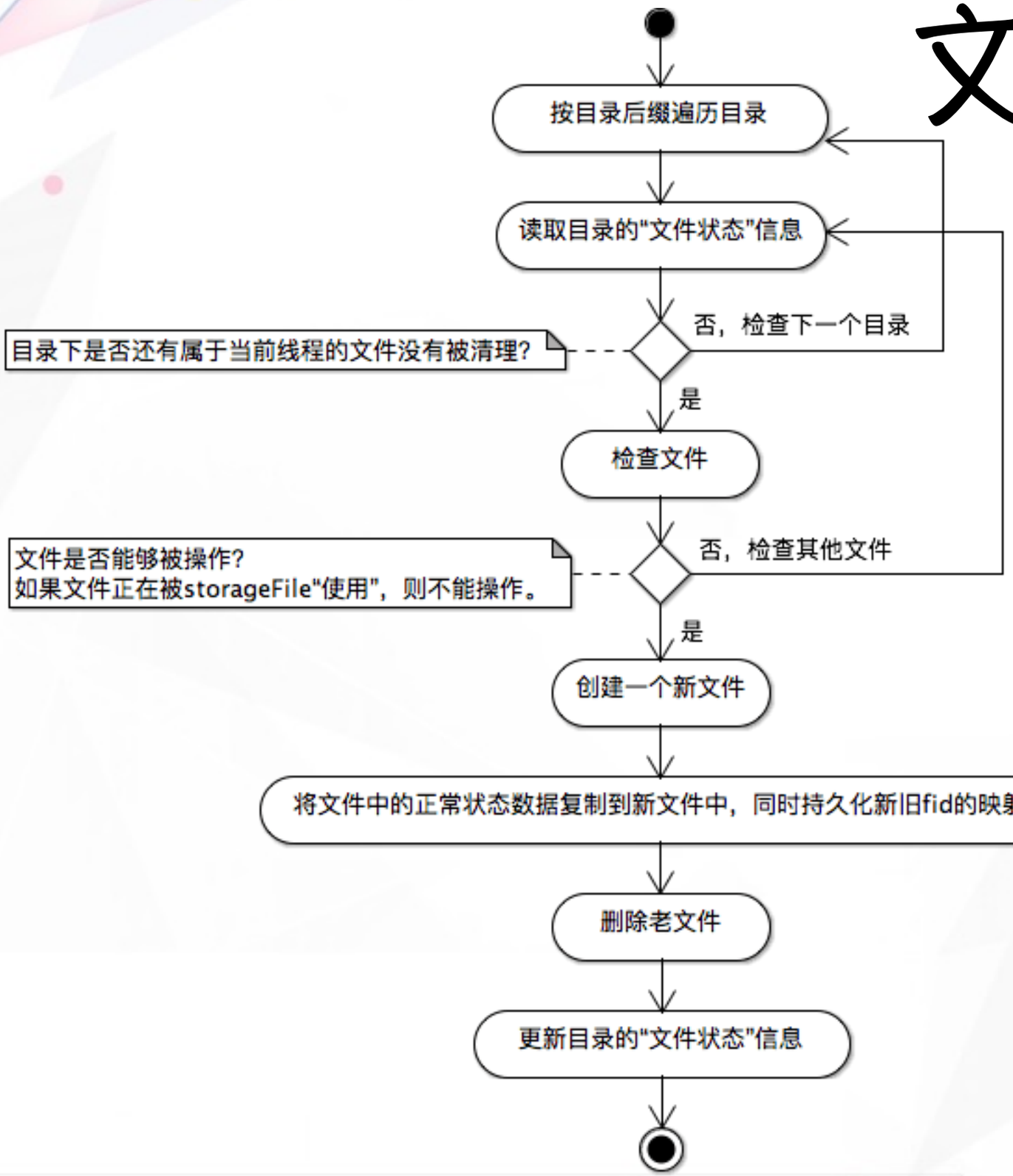
单盘同步

- 时机
  - 新加入机器
  - 新加入磁盘
  - 磁盘挂掉
- 条件
  - 无状态文件
  - 磁盘大小异常
- 基于日志文件的一致性算法
- 先同步日志文件，再同步数据，最后负责和实时同步连接
- master主动拉取数据

一致性同步

- 时机
  - 单盘恢复结束
  - 日常同步
- 基于gossip协议和binlog日志文件
- 标记synclog
- master主动推数据
- 同步状态机：marklog

文件回收


按目录后缀遍历目录
读取目录的"文件状态"信息
目录下是否还有属于当前线程的文件没有被清理?
否，检查下一个目录
是
检查文件
文件是否能够被操作?
如果文件正在被storageFile"使用"，则不能操作。
否，检查其他文件
是
创建一个新文件
将文件中的正常状态数据复制到新文件中，同时持久化新旧fid的映射
删除老文件
更新目录的"文件状态"信息

测试结果

- 单机测试状态 170+mb/s
- 多备份状态下 120+mb/s
- 同步1s-

线上情况

- 使用业务
  - 文章章节内容
  - 作家编辑历史库
  - 多媒体内容
- 使用量
  - 3个group集群，5个syncgroup，15台+机器
  - 访问量：2400w+ q/d
  - 容量20T+

V2

- 块存储->对象存储
- 对象索引
- 对象搜索
- 定时合并->写入合并、定时合并
- LRU缓存

THANKS

SequeMedia 盛拓传媒 | IT168 专注导购16年 | ChinaUnix | ITPUB www.itpub.net