



链滴

SeaWeedfs 分布式文件服务

作者: [liuzhiheng126](#)

原文链接: <https://ld246.com/article/1572511045115>

来源网站: [链滴](#)

许可协议: [署名-相同方式共享 4.0 国际 \(CC BY-SA 4.0\)](#)

SeaWeedfs 分布式网络文件存储介绍

引子

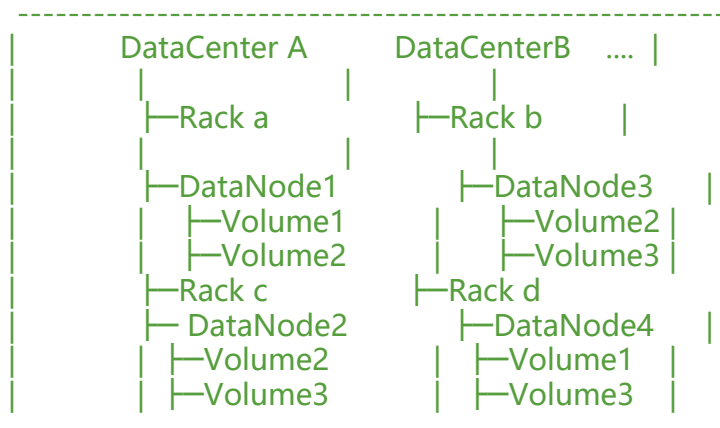
随着业务量增长，一个系统需要存储上百万文件的情况越来越多，尤其是互联网网站。在这种情况下使用传统磁盘/共享存储的方式进行支持会有以下问题：

- 文件的备份、恢复困难，大量文件的copy 耗时耗力
- 文件数量暴增占满操作系统文件系统inode，导致磁盘空间虽然没有用完但是因为inode用尽无法用
- 文件读取效率太低，无法应对高并发读取要求

针对以上问题，facebook 提出了自己的方案 Facebook' s Haystack design paper 。之后各种实出现，如tfs、MogileFS、GlusterFS等，其中Seaweedfs是一个比较优秀的实现。具有效率高、结构单、代码清晰等优点。本文针对seaweesfs进行简要的分析，在开始之前推荐阅读facebook的论文以习理解NoFS（Net Object File System）相关的理念。

拓扑结构

111 模式下的复写



在逻辑上Seaweedfs的几个概念：

- Node 系统抽象的节点，抽象为DataCenter、Rack、DataNode
- DataCenter 数据中心，对应现实中的不同机房
- Rack 机架，对应现实中的机柜
- Datanode 存储节点，用于管理、存储逻辑卷
- Volume 逻辑卷，存储的逻辑结构，逻辑卷下存储Needle
- Needle 逻辑卷中的Object，对应存储的文件
- Collection 文件集，可以分布在多个逻辑卷上

在实际实现中

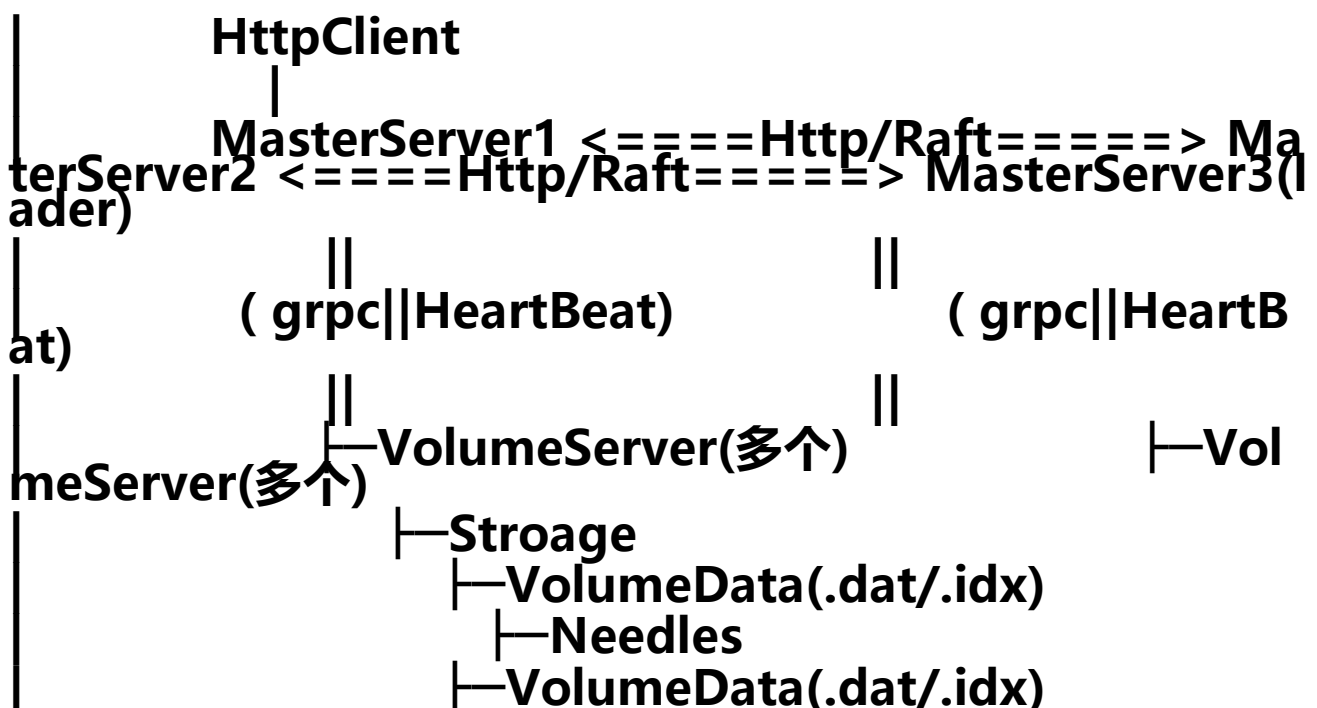
代码结构

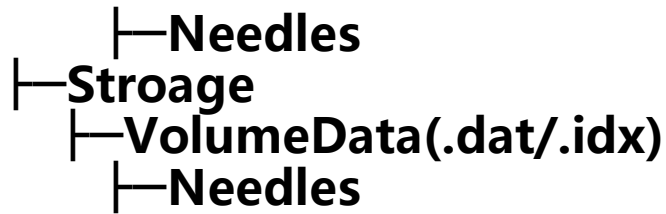
文件目录 逻辑层级 用途

weed

└─ command	1	weed的命令抽象
└─ filer	2	
└─ filer2	2	Filer 实现
└─ filesystem	2	FUSE mount到本地操作系统的实现
└─ glide.yaml	0	包依赖关系
└─ glog	4	日志组件
└─ images	3	对图片的伸缩/矫正操作
└─ operation volume	4	对于查询/读写的一些原子操作工具类，如通过http请求查询master下指定
└─ pb	4	MasterServer和VolumeServer之间 gRpc 心跳实现，主要是两个pb 文件以自动生成的go-grpc文件
└─ security	4	安全模块，主要是黑白名单实现，JWT实现
└─ sequence	3	序列号生成
└─ server	2	服务器实现，包括MasterServer/VolumeServer/FilerServer的实现
└─ stats	3	监控状态，主要是主机内存、磁盘等
└─ storage	3	物理存储的部分，包括物理Needle/物理卷Volume 的读写
└─ tools	1	实用的工具
└─ topology	2	逻辑拓扑结构，Datacenter、Rack、DataNode，主要是全局卷伸缩/查找操作
└─ util	4	工具类
└─ weed.go	0	主程序

部署结构





使用指南

详见: [Seaweedfs wiki](#)

1

不足与问题

- seaweedfs 采用的是同步式复写有以下几个问题:
 - a. 当在某个volume-server 下线又上线恢复的情况下, 没有自动的同步机制
 - b. 同步复写需要等待每个节点都复写成功, 效率相对较低
 - c. 虽然节点的上下线会快速通过心跳通知master节点, 但是仍然存在一定的延迟, 期间Volume-Servr在复写的时候可能会出现因为复写已经下线的volume-server导致上传失败的情况
- seaweedfs目前在权限管理方面还相对比较弱, 目前仅有一个白名单控制机制, 来控制外部的读写限/恶意删除。