目录

[前言 2](#_Toc428091743)

[缘起 2](#_Toc428091744)

[版权申明 2](#_Toc428091745)

[上篇：Ceph模块开发篇 3](#_Toc428091746)

[背景介绍 3](#_Toc428091747)

[Ceph平台初探 3](#_Toc428091748)

[初探Ceph架构 3](#_Toc428091749)

[Ceph基础概念 3](#_Toc428091750)

[Ceph基础数据结构 3](#_Toc428091751)

[Ceph的配置系统 3](#_Toc428091752)

[Ceph的模块化体系结构 3](#_Toc428091753)

[Ceph的请求处理 3](#_Toc428091754)

[ObjectStore模块 3](#_Toc428091755)

[模块简介 3](#_Toc428091756)

[模块的基本结构 3](#_Toc428091757)

[模块代码分析 3](#_Toc428091758)

[rgw模块 3](#_Toc428091759)

[其他模块 4](#_Toc428091760)

[下篇：Ceph原理解析篇 4](#_Toc428091761)

[Ceph架构详解 4](#_Toc428091762)

[Ceph的源码目录结构 4](#_Toc428091763)

[Ceph的进程机制 6](#_Toc428091764)

[Ceph基础设施 6](#_Toc428091765)

[Ceph的启动阶段 6](#_Toc428091766)

[Ceph的请求处理解析 6](#_Toc428091767)

[Ceph monitor原理解析 6](#_Toc428091768)

[附录 6](#_Toc428091769)

# 前言

# 缘起

# 版权申明

本书的著作权归作者华三统一存储服务器平台组成员所有。你可以：

* 下载、保存以及打印本书
* 网络链接、转载本书的部分或者全部内容，但是必须在明显处提供读者访问本书发布网站的链接
* 在你的程序中任意使用本书所附的程序代码，但是由本书的程序所引起的任何问题，作者不承担任何责任

你不可以：

* 以任何形式出售本书的电子版或者打印版
* 擅自印刷、出版本书
* 以纸媒出版为目的，改写、改编以及摘抄本书的内容

# 上篇：Ceph模块开发篇

## 背景介绍

## Ceph平台初探

### 初探Ceph架构

### Ceph基础概念

### Ceph基础数据结构

### Ceph的配置系统

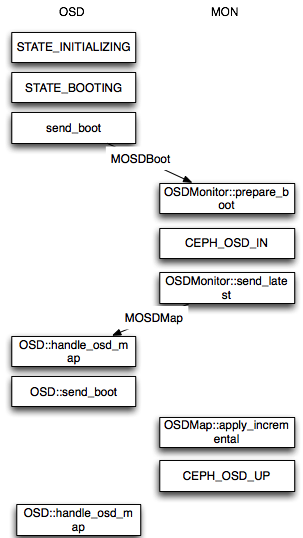
### Ceph的模块化体系结构

### Ceph的请求处理

#### 请求处理的流程

1. 新的osd加入流程

当一个新的 OSD 启动时，这时 Monitor 所掌握的最新 OSDMap 并没有该 OSD 的情况，因此该 OSD 会向 Monitor 申请加入，Monitor 再验证其信息后会将其加入 OSDMap 并标记为IN，并且将其放在 Pending Proposal 中会在下一次 Monitor “讨论”中提出，OSD 在得到 Monitor 的回复信息后发现自己仍然没在 OSDMap 中会继续尝试申请加入，接下来 Monitor 会发起一个 Proposal ，申请将这个 OSD 加入 OSDMap 并且标记为 UP 。然后按照 Paxos 的流程，从 proposal->accept->commit 到最后达成一致，OSD 最后成功加入 OSDMap 。当新的 OSD 获得最新 OSDMap 发现它已经在其中时。这时，OSD 才真正开始建立与其他OSD的连接，Monitor 接下来会开始给他分配PG。



## ObjectStore模块

### 模块简介

### 模块的基本结构

### 模块代码分析

## rgw模块

## mon模块

### 模块简介

### 模块的基本结构

### 模块代码分析

1. Client->mon业务消息

|  |  |  |
| --- | --- | --- |
| 消息类型 | 消息值 | 消息结构体 |
| CEPH\_MSG\_MON\_MAP | 4 | MMonMap |
| CEPH\_MSG\_MON\_GET\_MAP | 5 | MMonGetMap |
| CEPH\_MSG\_MON\_GET\_OSDMAP | 6 | MMonGetOSDMap |
| CEPH\_MSG\_MON\_METADATA | 7 | MMonMetadata |
| CEPH\_MSG\_STATFS | 13 | MStatfs |
| CEPH\_MSG\_STATFS\_REPLY | 14 | MStatfsReply |
| CEPH\_MSG\_MON\_SUBSCRIBE | 15 | MMonSubscribe |
| CEPH\_MSG\_MON\_SUBSCRIBE\_ACK | 16 | MMonSubscribeAck |
| CEPH\_MSG\_AUTH | 17 | MAuth |
| CEPH\_MSG\_AUTH\_REPLY | 18 | MAuthReply |
| CEPH\_MSG\_MON\_GET\_VERSION | 19 | MMonGetVersion |
| CEPH\_MSG\_MON\_GET\_VERSION\_REPLY | 20 | MMonGetVersionReply |

所有业务消息类型解析的源码为Message \*decode\_message（switch优化）；

Monitor处理业务消息接口实现为Monitor::dispatch

## 其他模块

# 下篇：Ceph原理解析篇

## Ceph架构详解

Ceph的下篇将会更加深入的介绍ceph的实现原理。

### Ceph的源码目录结构

1. 根目录
   1. [src]：各功能模块的源代码
   2. [qa]：各个模块的功能测试（测试脚本和测试代码）
   3. [wireshark]：#wireshark的ceph插件
   4. [admin]：管理工具，编译生成文档工具
   5. [debian]：用于制作debian（Ubuntu）安装包的相关脚本和文件
   6. [doc]：用于生成项目文档，生成结果参考http://ceph.com/docs/master/
   7. [man]：ceph各命令行工具的man文件
   8. [examples]：一些使用示例
   9. configure.ac：用于生成configure的脚本
   10. Makefile.am：用于生成Makefile的脚本
   11. autogen.sh：负责生成configure
   12. do\_autogen.sh：生成configure的脚本，实际上通过调用autogen.sh实现
   13. ceph.spec.in：RPM包制作文件
2. src目录
   1. [include]：头文件，包含各种基本类型的定义，简单通用功能等。
   2. [common]：共有模块，包含各类共有机制的实现，例如线程池、管理端口、节流阀等。
   3. [log]：日志模块，主要负责记录本地log信息（默认/var/log/ceph/目录）
   4. [global]：全局模块，主要是声明和初始化各类全局变量（全局上下文）、构建驻留进程、信号处理等
   5. [auth]：授权模块，实现了三方认知机制
   6. [crush]：Crush模块，Ceph的数据分布算法
   7. [msg]：消息通讯模块，包括用于定义通讯功能的抽象类Messenger以及目前的实现SimpleMessager
   8. [messages]：消息模块，定义了Ceph各节点之间消息通讯中用到的消息类型
   9. [os]：对象（Object Store）模块，用于实现本地的对象存储功能
   10. [osdc]：OSD客户端（OSD Client），封装了各类访问OSD的方法
   11. [mon]：mon模块
   12. [osd]：osd部分
   13. [mds]：mds模块
   14. [rgw]：rgw模块的
   15. [librados]：rados库模块的代码
   16. [librdb]：libbd库模块的代码
   17. [client]：client模块，实现了用户态的CephFS客户端
   18. [mount]：mount模块
   19. [tools]：各类工具
   20. [test]：单元测试
   21. [perfglue]：与性能优化相关的源代码
   22. [json\_spirit]：外部项目json\_spirit
   23. [leveldb]：外部项目leveldb from google
   24. [gtest]：gtest单元测试框架
   25. [doc]：关于代码的一些说明文档
   26. [bash\_completion]：部分bash脚本的实现
   27. [pybind]：python的包装器
   28. [script]：各种python脚本
   29. [upstart]：各种配置文件
   30. ceph\_mds.cc：驻留程序mds
   31. ceph\_mon.cc：驻留程序mon
   32. ceph\_osd.cc：驻留程序osd
   33. libcephfs.cc：cephfs库
   34. librdb.cc：rdb库
   35. ceph\_authtool.cc：工具ceph\_authtool
   36. ceph\_conf.cc：工具ceph\_conf
   37. ceph\_fuse.cc：工具ceph\_fuse
   38. ceph\_syn.cc：工具ceph\_syn
   39. cephfs.cc：工具cephfs
   40. crushtool.cc：工具crushtool
   41. dupstore.cc：工具dupstore
   42. librados-config.cc：rados库配置工具
   43. monmaptool.cc：工具monmap
   44. osdmaptool.cc：工具osdmap
   45. psim.cc：工具psim
   46. rados.cc：工具rados
   47. rdb.cc：工具rdb
   48. rados\_export.cc：rados工具相关类
   49. rados\_import.cc：rados工具相关类
   50. rados\_sync.cc：rados工具相关类
   51. rados\_sync.h：rados工具相关类
   52. sample.ceph.conf：配置文件样例
   53. ceph.conf.twoosds：配置文件样例
   54. Makefile.am：makefile的源文件
   55. valgrind.supp：内存检查工具valgrind的配置文件
   56. init-ceph.in：启动和停止ceph的脚本
   57. mkcephfs.in：cephfs部署脚本

### Ceph的进程机制

## Ceph基础设施

## Ceph的启动阶段

## Ceph的请求处理解析

## Ceph monitor原理解析

# 附录