


(/apps/redirect?utm_source=side-banner-click)

深入浅出Google File System

 超级个体颢项 (/u/b7f94092fc21) [+ 关注](#)
2016.09.30 16:49* 字数 796 阅读 1476 评论 0 喜欢 3
(/u/b7f94092fc21)

GFS是什么

GFS，顾名思义就是谷歌文件系统，和Big Table，Map Reduce并称谷歌三驾马车。大部分谷歌服务的基石（Search, Cloud Drive, Gmail etc.）

架构的层次

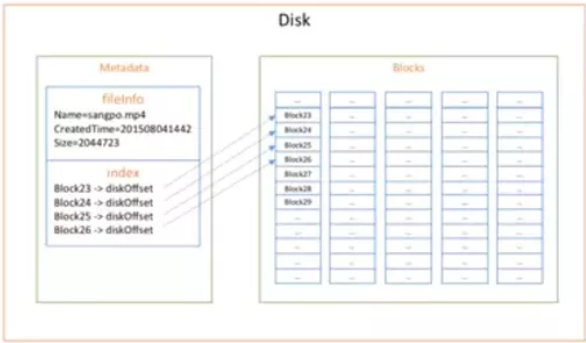


图片

最底层是文件系统，在之上是将数据模型抽象出来，便于很好的使用，这就是bigTable，在之上是算法，算法除了访问数据模型外，还能够直接访问文件系统，最上面就是各类应用了

gfs从哪里来

源头是如何保存一个文件？



关键点
• 1 block = 1024Byte

图片

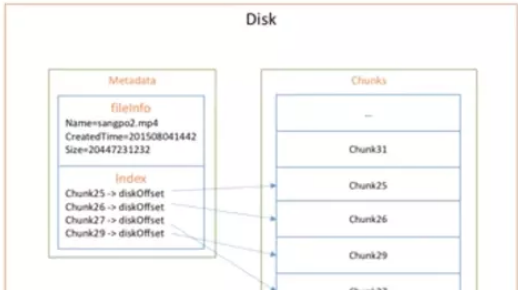
保存文件需要两部分：

- metadata：包括文件信息和索引
- file content：具体的文件内容

进一步如何保存大文件



保存一个大文件



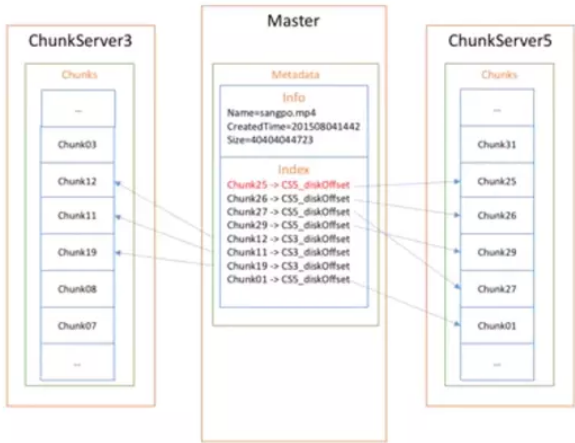
- 关键点
 - 1 chunk = 64MB
 - = 64*1024
 - = 65,536 blocks
- 优点
 - 减少元数据
 - 减少流量
- 缺点

(/apps/redirect?utm_source=side-banner-click)

图片

此时索引信息会保存的粒度更粗，存的是chunk，每个chunk是64M

再进一步，怎么保存超大文件
保存超大文件



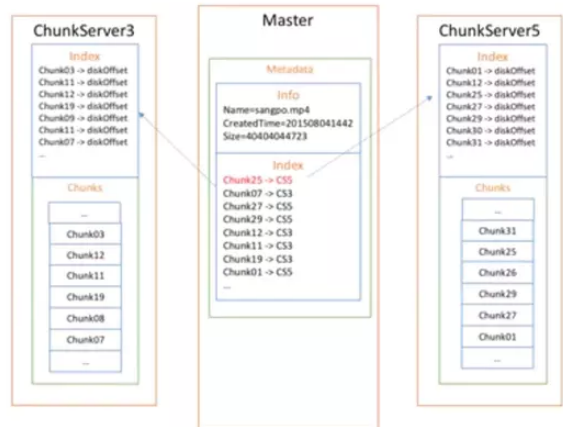
- 关键点
 - Master + many ChunkServers
- 缺点
 - ChunkServer数据的任何改变都需要通知Master

图片

确定啊很明显：chunkServer的变化都需要将其告诉master

怎么进行改进？

减少Master的数据和流量



- 关键点
 - Master不记录每块数据的偏移量（diskOffset）
- 优点
 - 减少Master的元数据信息
 - 减少Master和Chunkserver之间的通信

图片

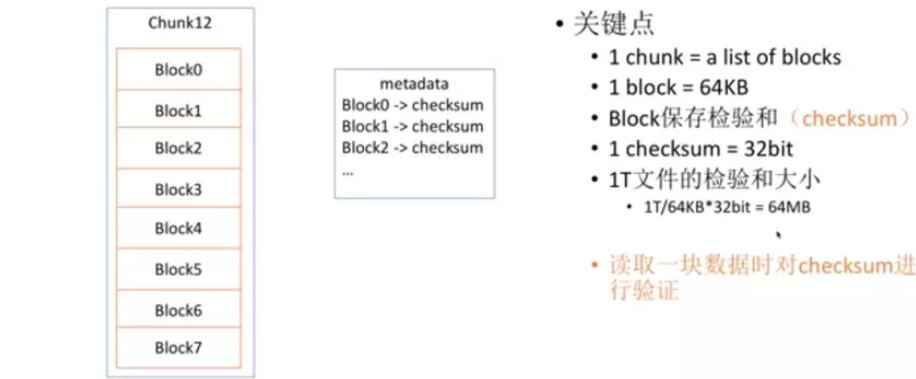
系统设计中非常关键的点：**耦合和聚合**，将属于它的放到它那，不属于的放到其他地方

将master保存每一块在哪个服务器上，每个服务器的索引放到chunkServer中

GFS容错机制

- 怎么发现数据损坏

发现数据损坏

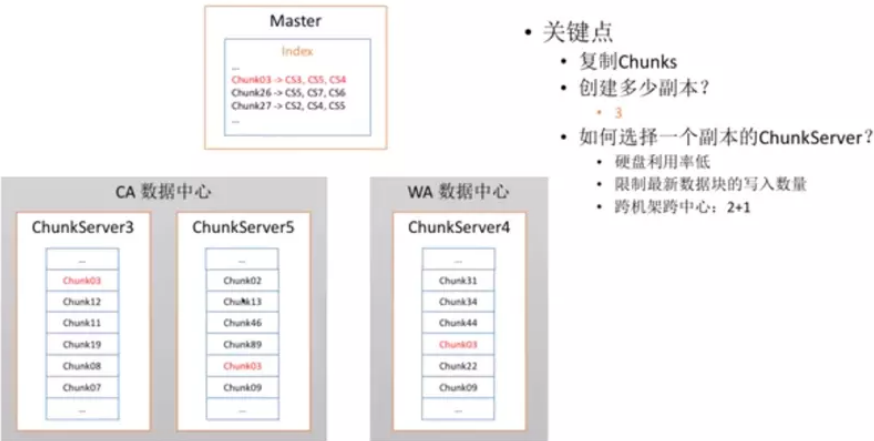


图片

可以对每个block保存个checksum，对于1T的数据，只有64M，完全可以放到内存中

如果数据损坏的话呢，Chunk Server就找Master恢复数据

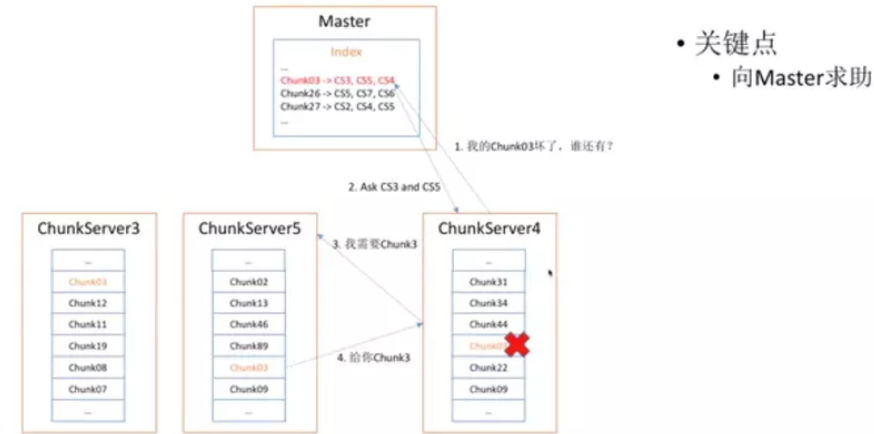
减少ChunkServer挂掉带来的损失



图片

为了防止数据的丢失，就做冗余存储，每个chunk存3份，在chunkServer的选择上，尽可能放到不同的机房，然后同机房也放到不同的机架上

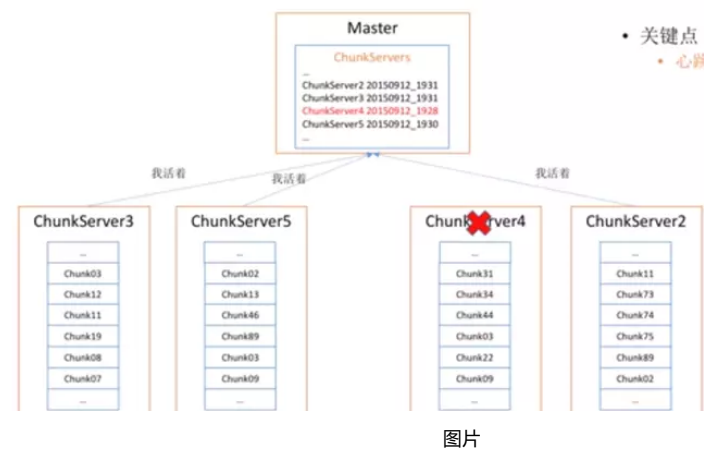
恢复损坏的chunk



图片

master是关键，会同时发送心跳检查Chunk Server是否运行正常。如果有服务器挂掉的话就向Master申请恢复

发现ChunkServer挂掉

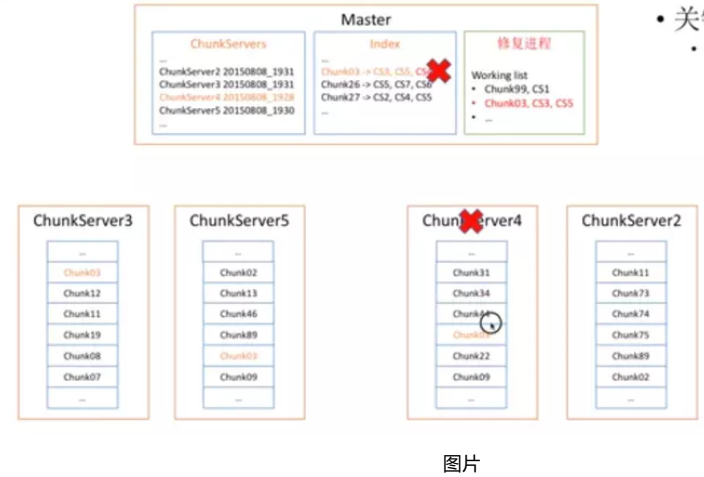


- 关键点
- 心跳

(/apps/redirect?utm_source=side-banner-click)

心跳的设计：可能是由于master和server之间网络不通，这个时候，master会求助其他的server，让他们再去ping下失联的server

ChunkServer挂掉后恢复数据

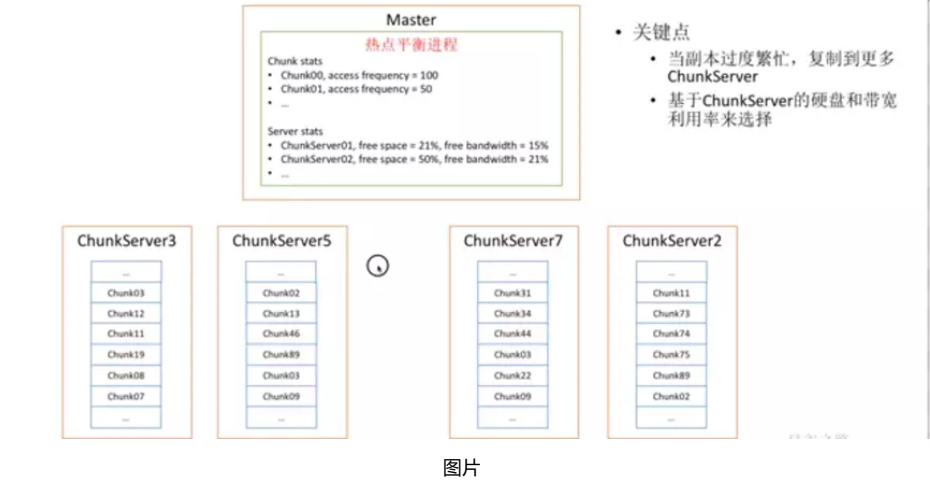


- 关键点
- 基于存活副本数的恢复策略

当发现副本数小于3个，会启动修复进程进行修复，修复的优先级

怎么应对热点

应对热点



- 关键点
- 当副本过度繁忙，复制到更多ChunkServer
- 基于ChunkServer的硬盘和带宽利用率来选择

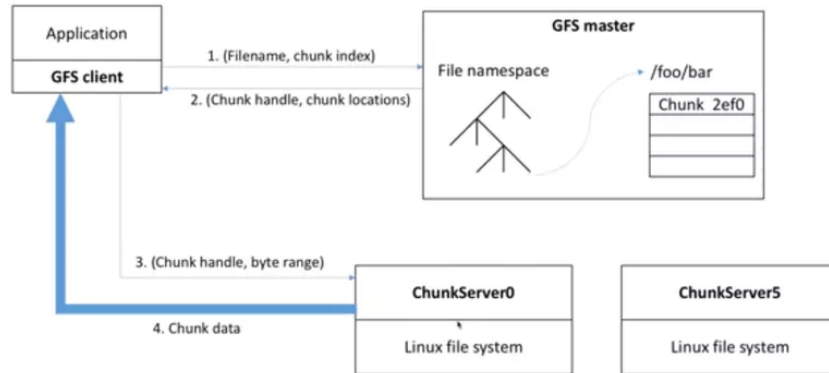
负载均衡

核心读写操作

好了，我们构建这么一个庞大的系统最后不就是要读和写嘛，现在我们看看GFS是如何读写的。读数据时Client先向Master要到Chunk信息，然后去ChunkServer取数据

读文件过程

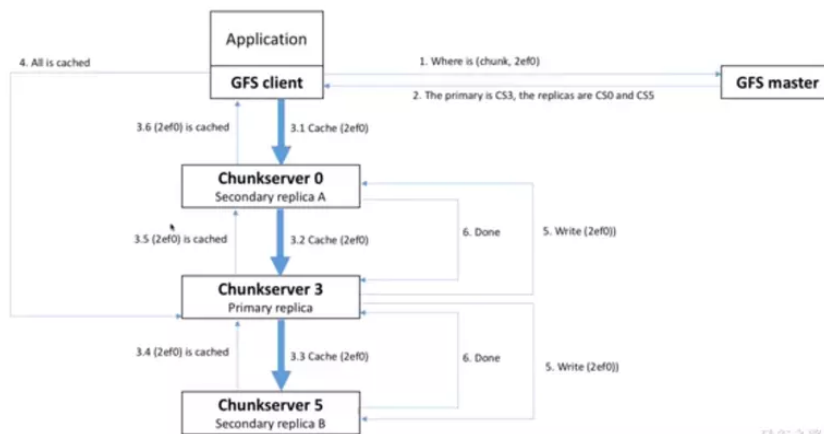
(/apps/redirect?
utm_source=side-
banner-click)



图片

写文件的时候呢，也是先找到Master Server要到信息，然后找到距离最近的Chunk Server。由其带领其他ChunkServer一起写数据。如果图中有任何一步出现错误则中止写入，返回错误

写文件过程



图片

写的时候是往最近的server写，然后server再接收到数据后就往其他server发送，在写入的时候呢，是先缓存下来，最后都缓存好了，再写入到磁盘，好处是减少出错的概率，因为一旦缓存好，再写出错的几率就大大减少了。

第3步是cache，都cached后，由primary server负责协调开始写入，都写成功后，告诉客户端

如果写入出错了怎么办？如果引入出错处理机制，会引入更多的问题，往往解决一个问题会带来更多的问题，因此系统在设计过程中，尽可能只提供最简单的功能，由客户端来负责重试

参考

深入浅出Google File System: 视频 (<https://link.jianshu.com?t=http://www.bittiger.io/videos/QPQAY2DFkqLwHBS4K/qtFZHdaf6JJQxyMCB>)

硅谷之路9: 深入浅出理解GFS: 文字 (<https://link.jianshu.com?t=https://zhuanlan.zhihu.com/p/20673524?refer=bittiger>)

您的每一次打赏，都是对我最大的鼓励，期待我们共同进步

赞赏支持

(/apps/redirect?utm_source=side-banner-click)

点滴成长 (/nb/3685950) 举报文章 © 著作权归作者所有




超级个体颢项 (/u/b7f94092fc21)

写了 109918 字，被 454 人关注，获得了 449 个喜欢 (/u/b7f94092fc21)

专注大规模分布式系统开发，紧跟人工智能浪潮

+ 关注

喜欢 | 3



更多分享

开发10年
全记在这本Java进阶宝典了

Spring源码分析

分布式架构

微服务架构

JVM性能优化

高效DevOps

多线程并发编程

点击领取



(/p/428251ede1aa)







登录 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-comment-form)

评论

智慧如你，不想发表一点想法 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-nocomments-text)咩~

被以下专题收入，发现更多相似内容

-  互联网科技 (/c/93d58e9169cb?utm_source=desktop&utm_medium=notes-included-collection)
-  首页投稿 (暂停... (/c/bDHhpK?utm_source=desktop&utm_medium=notes-included-collection)
-  程序员 (/c/NEt52a?utm_source=desktop&utm_medium=notes-included-collection)
-  互联网通用技术 (/c/cd8ddb694f4c?)

^



utm_source=desktop&utm_medium=notes-included-collection)

推荐阅读

更多精彩内容 > (/)

成功学是种“瘾”，得戒 (/p/ed91dbc2b300?utm_cam...

(/p/ed91dbc2b300?

(/apps/redirect?
utm_source=side-
utm_campaign=maleskine&utm_content=note&utm
banner-click)

作者:任争气(醉美古都) 年过30, 请扔掉你手里的成功学。草根逆袭的事, 有, 但是不常有。所以, 成功学是一种瘾, 得戒。若大一个城市、丢进去一个渺...

醉美长安 (/u/7886cd7da46e?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

“渣女能渣到什么程度?” (/p/8c0224dccf0e?utm_ca...

(/p/8c0224dccf0e?

utm_campaign=maleskine&utm_content=note&utm

最近在网上看到有人在讨论这个问题“渣女能渣到什么程度?” 点开一看, 简单的两张截图却写满了精彩的故事: 中国古代智慧告诉我们, 男人是土捏烤制的...

浪迹情感TV (/u/4a6caa28b3e3?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

也说年味 | 大寒 (/p/64bf25311313?utm_campaign=...

(/p/64bf25311313?

utm_campaign=maleskine&utm_content=note&utm

么舅电话约回老家, 知为大寒祭扫事, 却因扶贫迎检, 脱身不能, 只得婉拒推脱, 心中愧疚难言。前两天屡见大寒文, 多言节气时令, 于老家川北, 大寒...

山居散人 (/u/f0c4683449cb?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

不要让低效的自律, 成为拖垮你的理由! (/p/fd126137...

(/p/fd126137f02f?

utm_campaign=maleskine&utm_content=note&utm

填坑进度: 15/1001 01 不知道大家有没有注意到这么一个奇怪的现象: 比如在读高中的时候, 班上总有那么几个同学, 课间在学习, 午休在学习, 回到宿舍...

一言不合填个坑 (/u/098c923e4f43?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

小说设计·上篇 (/p/fd87ad1f86ee?utm_campaign=ma...

(/p/fd87ad1f86ee?

utm_campaign=maleskine&utm_content=note&utm

【寻找写作的真相】目录【上一篇: 小说解剖·下篇】小说设计是什么? 小说设计并不高深, 大家可以简单理解它是写作之前的构思过程。记录一个灵感点...

一鸣 (/u/dc22650a4033?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

(/p/540ae70c0500?

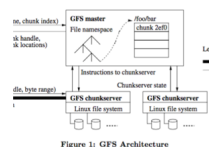


Figure 1: GFS Architecture

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

GFS(Google File System)读书笔记 (/p/540ae70c0500?utm_campaign=...

引言 GFS是谷歌2003年提出的一个文件系统。虽然GFS比较古老, 但是后来的HDFS, 是受到了GFS的启发, 是GFS的一种开源实现。因此熟悉与理解GFS的设计原理, 会对理解整个hadoop生态系统有更好的帮...

炸茄盒 (/u/bf5ca0e36a87?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/1d7d0e4d41a1?

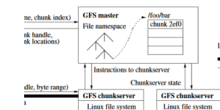


Figure 1: GPS Architecture

```
(/apps/redirect?  
utm_source=side-  
ommendation)  
banner-click)
```

GFS-Google文件系统 (/p/1d7d0e4d41a1?utm_campaign=maleskine&ut...

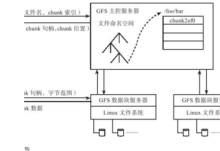
GFS-Google文件系统 GFS是一个可扩展的分布式文件系统，用于大型的、分布式的、对大量数据进行访问的应用。它运行于廉价的普通硬件上，但可以提供容错功能。它可以给大量的用户提供总体性能较高的服...



全能程序猿 (/u/082645cc19c6?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/e9a477ee27c1?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

笔记-GFS (/p/e9a477ee27c1?utm_campaign=maleskine&utm_content=...

分布式文件系统的主要功能有两个：一个是存储文档、图像、视频之类的Blob类型数据；另外一个作为分布式表格系统的持久化层。分布式文件系统中最为著名的莫过于Google File System (GFS)，它构建在...



olostin (/u/2c508e07ea8d?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/ad77b167f5c6?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

GFS--Google File System (/p/ad77b167f5c6?utm_campaign=maleskine...

sina Google File System, 一个适用于大规模分布式数据处理相关应用的, 可扩展的分布式文件系统。GFS 提供了常见的文件系统的接口, 文件是通过...



一颗鲜橙子 (/u/a83d32b815fc?



utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/492ecffbe614?

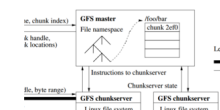


Figure 1: GPS Architecture

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

GFS-Google 论文阅读笔记 (/p/492ecffbe614?utm_campaign=maleskine...

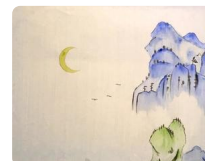
众所周知，Hadoop的存储基础，HDFS分布式文件系统，是按照GFS的思想实现的。本文参考：Google File System 中文版 1.0 版 译者 alex，原文地址 <http://blademaster.ixiezi.com/> GFS是面向大规模数据密集型应...



SmileySure (/u/349176f0d34e?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/57d38ef6c3f3?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

人闲桂花落 (/p/57d38ef6c3f3?utm_campaign=maleskine&utm_content...



空森林 (/u/7baa508faa39?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/d7d8c3cdb46d?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

关于古法黑糖 (/p/d7d8c3cdb46d?utm_campaign=m...

关于古法黑糖 如果感兴趣 有需要 可以加我微信 244412661

 BigBig张大大 (/u/f43776665b21?)



云南，
在中国西南边陲有一个
大理以其秀丽的自然风
优美的民族风情为特色
无数中外游客。

(/apps/redirect?
utm_source=side-
banner-click)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)


(/p/2f33ed883e82?)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

童话不是神话 (/p/2f33ed883e82?utm_campaign=maleskine&utm_conte...


写在2016年春节..... 春节，是在说春天已经要来了， 前些日回到家乡， 很多熟悉又陌生的感觉不时地冲刷着自己。忽然想起两句诗：“吟诗日日待春风，及至桃花开后却匆匆”。想想自己多年，一些一直期盼和追...

 水木言 (/u/d7bef5258eb2?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

觉察日记 (/p/caf1e74a67c9?utm_campaign=maleskine&utm_content=n...

昨天和老公吵架啦，看到了旧有模式的延续与张牙舞爪，仿佛对我说你想消除我，没那么容易，想起奇异博士里的一句话，心魔不会被消灭，学会和他相处。我对旧模式是说，你挺好的，我之前是挺嫌弃你的，对...

 flying0227 (/u/81ebc89de065?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)


(/p/61cdc1266de0?)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

抑郁症是一条黑狗，咬死了无数人才 (/p/61cdc1266de0?utm_campaign=...

本兮，创作才女，于平安夜抑郁症跳楼。我并不是本兮的粉丝，只是觉得她的歌词让我比较有感觉。最近我有很多朋友都在怀念她，看着这件事，不由叹息。又是这个叫做抑郁症的东西祸害的。近些年熟知的公众...

 heartGeraldine (/u/82a580146c7a?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

