

分布式系统为什么需要 Tracing?



凡墙 (/u/7ac19e0b777c) [+ 关注](#)
2016.02.03 10:28* 字数 2983 阅读 3334 评论 2 喜欢 12
(/u/7ac19e0b777c)

来源: http://www.cnblogs.com/zhengyun_ustc/p/55solution2.html
(https://link.jianshu.com?t=http://www.cnblogs.com/zhengyun_ustc/p/55solution2.html)

先介绍一个概念: 分布式跟踪, 或分布式追踪。

电商平台由数以百计的分布式服务构成, 每一个请求路由过来后, 会经过多个业务系统并留下足迹, 并产生对各种Cache或DB的访问, 但是这些分散的数据对于问题排查, 或是流程优化都帮助有限。对于这么一个跨进程/跨线程的场景, 汇总收集并分析海量日志就显得尤为重要。要能做到追踪每个请求的完整调用链路, 收集调用链路上每个服务的性能数据, 计算性能数据和比对性能指标 (SLA), 甚至在更远的未来能够再反馈到服务治理中, 那么这就是分布式跟踪的目标了。在业界, twitter 的 zipkin 和淘宝的鹰眼就是类似的系统, 它们都起源于 Google Dapper 论文, 就像历史上 Hadoop 发源于 Google Map/Reduce 论文, HBase 源自 Google BigTable 论文一样。

好了, 整理一下, Google叫Dapper, 淘宝叫鹰眼, Twitter叫ZipKin, 京东商城叫Hydra, eBay叫Centralized Activity Logging (CAL), 大众点评网叫CAT, 我们叫Tracing。

这样的系统通常有几个设计目标:

- (1) 低侵入性——作为非业务组件, 应当尽可能少侵入或者无侵入其他业务系统, 对于使用方透明, 减少开发人员的负担;
- (2) 灵活的应用策略——可以 (最好随时) 决定所收集数据的范围和粒度;
- (3) 时效性——从数据的收集和产生, 到数据计算和处理, 再到最终展现, 都要求尽可能快;
- (4) 决策支持——这些数据是否能在决策支持层面发挥作用, 特别是从 DevOps 的角度;
- (5) 可视化才是王道。

先来一个直观感受:
下面依次展示了 ZipKin、鹰眼、窝窝的调用链绘制界面。



图1 twitter zipkin 调用链

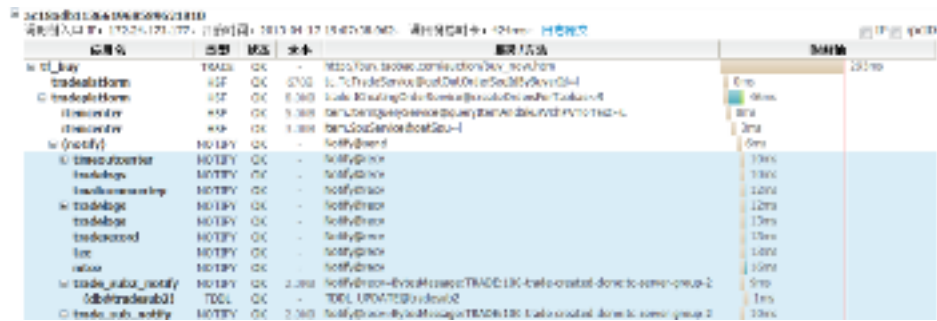


图2 淘宝鹰眼的调用链



图3 京东商城hydra调用链



图4 窝窝tracing调用链

鼠标移动到调用链的每一层点击，可以看到执行时长、宿主机IP、数据库操作、传入参数甚至错误堆栈等等具体信息。

淘宝如何实现的：

同一次请求的所有相关调用的情况，在淘宝 EagleEye 里称作 **调用链**。同一个时刻某一台服务器并行发起的网络调用有很多，怎么识别这个调用是属于哪个调用链的呢？可以在各个发起网络调用的中间件上下手。在前端请求到达服务器时，应用容器在执行实际业务处理之前，会先执行 EagleEye 的埋点逻辑（类似 Filter 的机制），埋点逻辑为这个前端请求分配一个全局唯一的调用链ID。这个ID在 EagleEye 里面被称为 Traceld，埋点逻辑把 Traceld 放在一个调用上下文对象里面，而调用上下文对象会存储在 ThreadLocal 里面。调用上下文里还有一个ID非常重要，在 EagleEye 里面被称作 RpcId。RpcId 用于区分同一个调用链下的多个网络调用的发生顺序和嵌套层次关系。对于前端收到请求，生成的 RpcId 固定都是0。

当这个前端执行业务处理需要发起 RPC 调用时，淘宝的 RPC 调用客户端 HSF 会首先从当前线程 ThreadLocal 上面获取之前 EagleEye 设置的调用上下文。然后，把 RpcId 递增一个序号。在 EagleEye 里使用多级序号来表示 RpcId，比如前端刚接到请求之后的 RpcId 是0，那么 它第一次调用 RPC 服务A时，会把 RpcId 改成 0.1。之后，调用上下文会作为附件随这次请求一起发送到远程的 HSF 服务器。

HSF 服务端收到这个请求之后，会从请求附件里取出调用上下文，并放到当前线程 ThreadLocal 上面。如果服务A在处理时，需要调用另一个服务，这个时候它会重复之前提到的操作，唯一的差别就是 Rpclid 会先改成 0.1.1 再传过去。服务A的逻辑全部处理完

毕之后，HSF 在返回响应对象之前，会把这次调用情况以及 Traceld、Rpclid 都打印到它的访问日志之中，同时，会从 ThreadLocal 清理掉调用上下文。如图6-1展示了一个浏览器请求可能触发的系统间调用。

(/apps/d
utm_sou

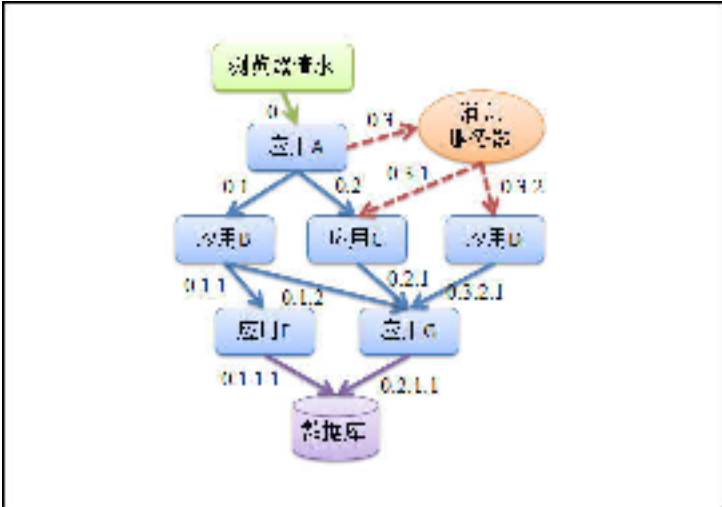


图6-1-一个浏览器请求可能触发的系统间调用

图6-1描述了 EagleEye 在一个非常简单的分布式调用场景里做的事情，就是为每次调用分配 Traceld、Rpclid，放在 ThreadLocal 的调用上下文上面，调用结束的时候，把 Traceld、Rpclid 打印到访问日志。类似的其他网络调用中间件的调用过程也都比较类似，这里不再赘述了。访问日志里面，一般会记录调用时间、远端IP地址、结果状态码、调用耗时之类，也会记录与这次调用类型相关的一些信息，如URL、服务名、消息topic等。很多调用场景会比上面说的完全同步的调用更为复杂，比如会遇到异步、单向、广播、并发、批处理等等，这时候需要妥善处理好 ThreadLocal 上的调用上下文，避免调用上下文混乱和无法正确释放。另外，采用多级序号的 Rpclid 设计方案会比单级序号递增更容易准确还原当时的调用情况。

最后，EagleEye 分析系统把调用链相关的所有访问日志都收集上来，按 Traceld 汇总在一起之后，就可以准确还原调用当时的情况了。



图6-2-一个典型的调用链

如图6-2所示，就是采集自淘宝线上环境的某一条实际调用链。调用链通过树形展现了调用情况。调用链可以清晰地看到当前请求的调用情况，帮助问题定位。如上图，mtop应用发生错误时，在调用链上可以直接看出这是因为第四层的一个(tair@1)请求导致网络超时，使最上层页面出现超时问题。这种调用链，可以在 EagleEye 系统监测到包含异常的访问日志后，把当前的错误与整个调用链关联起来。问题排查人员在发现入口错误量上涨或耗时上升时，通过 EagleEye 查找出这种包含错误的调用链采样，提高故障定位速度。

调用链数据在容量规划和稳定性方面的分析



如果对同一个前端入口的多条调用链做汇总统计，也就是说，把这个入口URL下面的所有调用按照调用链的树形结构全部叠加在一起，就可以得到一个新的树结构（如图6-3所示）。这就是入口下面的所有依赖的调用路径情况。

(/apps/d
utm_sou

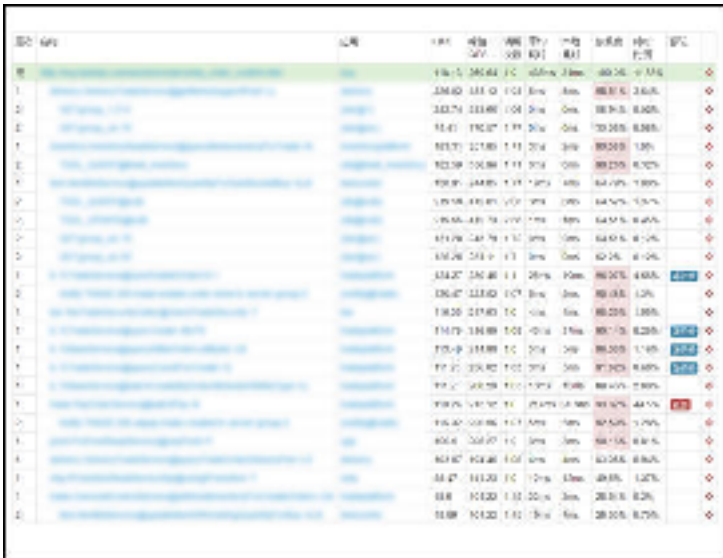


图6-3-对某个入口的调用链做统计之后得到的依赖分析

这种分析能力对于复杂的分布式环境的调用关系梳理尤为重要。传统的调用统计日志是按固定时间窗口预先做了统计的日志，上面缺少了链路细节导致没办法对超过两层以上的调用情况进行分析。例如，后端数据库就无法评估数据库访问是来源于最上层的哪些入口；每个前端系统也无法清楚确定当前入口由于双十一活动流量翻倍，会对后端哪些系统造成多大的压力，需要分别准备多少机器。有了 EagleEye 的数据，这些问题就迎刃而解了。

下图6-4展示了数据流过程。

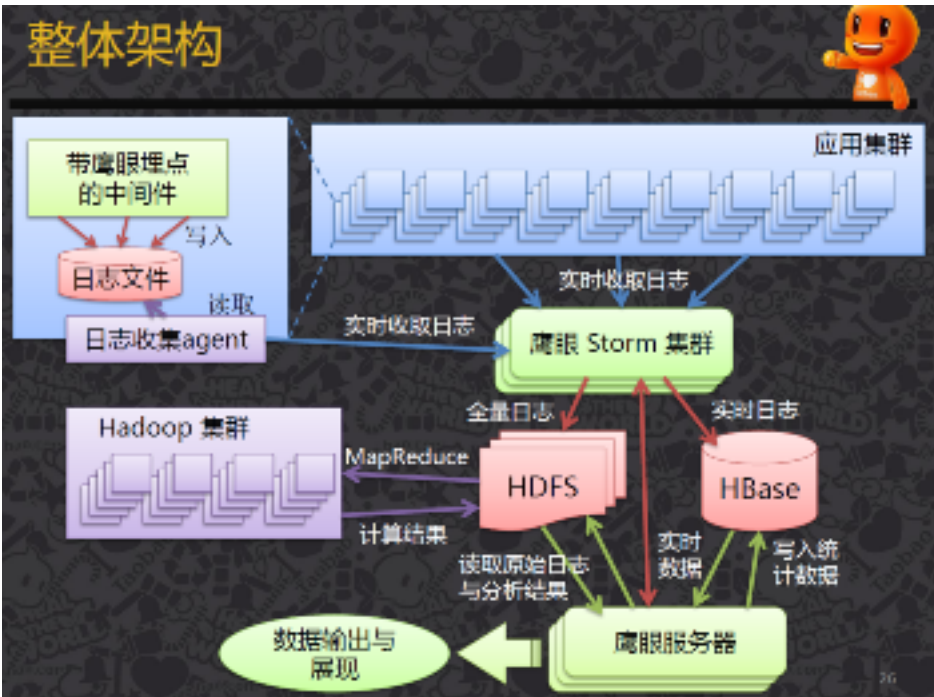
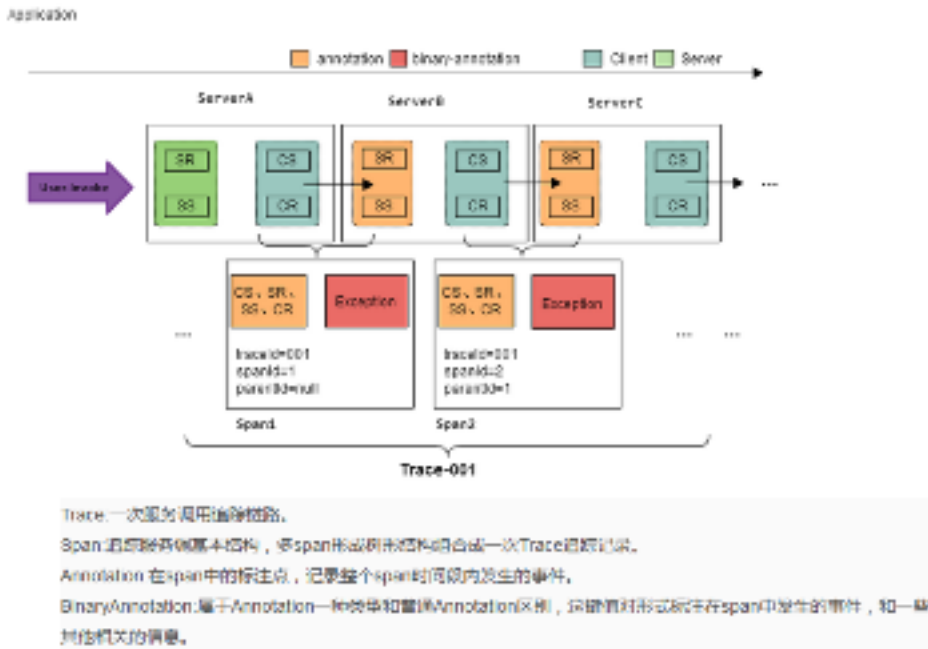


图6-4 鹰眼的收集和数据存储

京东如何实现的：

京东商城引入了阿里开源的服务治理中间件 Dubbo，所以它的分布式跟踪 Hydra 基于 Dubbo 就能做到对业务系统几乎无侵入了。

Hydra 的领域模型如下图7所示：



(/apps/d
utm_so

图7 hydra 领域模型以及解释
hydra 数据存储是 HBase，如下图8所示：

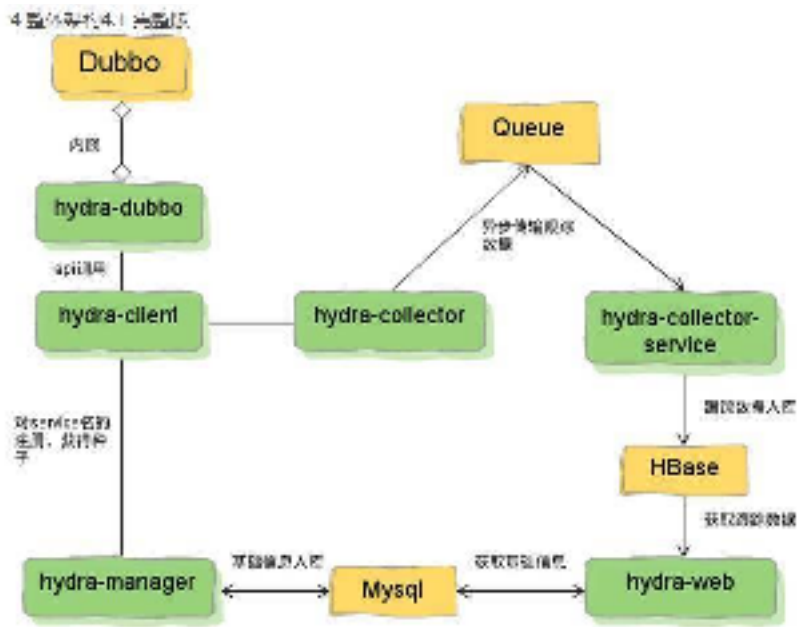


图8 hydra 架构

窝窝如何实现的：

2012年，逐渐看到自建分布式跟踪系统的重要性，但随即意识到如果没有对 RPC 调用框架做统一封装，就可能侵入到每一个业务工程里去写埋点日志，于是推广 Dubbo 也提上日程。2013年，确定系统建设目标，开始动手。由于 tracing 跟 DevOps 息息相关，所以数据聚合、存储、分析和展示由运维部向荣牵头开发，各个业务工程数据埋点和上报由研发部国玺负责。

经过后续向荣、刘卓、国玺、明斌等人的不断改进，技术选型大致如下所示。

埋点

实现线程内 trace 上下文传递，即服务器内部的方法互调时不需要强制在方法形参中加 Message 参数；

实现 trace 埋点逻辑自动织入功能，即业务开发人员不需要在方法中打印 trace 日志，只需要给该方法加注解标识；




```
@TraceClass(projectName = "ucenter")
public class UserServiceImpl extends AbstractServiceImpl implements UserService {

    @TraceMethod
    public UserResult findUserInfoByNick(final String nick, final RpcCallArg rpcCallArg) {
        //如果该接口有 message 对象，则从该 message 中取出 rootid 或 parentid
        //从而进行过滤
    }

    @TraceMethod(messageType = MessageTypeEnum.CENTIS)
    public UserResult findUserInfoByNick(final String nick) {
        //你的业务逻辑
    }
}
```

(/apps/d
utm_sou

原理：利用 Javaagent 机制，执行 main 方法之前，会先执行 premain 方法，在该方法中将字节码转换器载入 instrumentation，而后 jvm 在加载 class 文件之前都会先执行字节码转换器。

字节码转换器中的逻辑为，识别出注解 trace 的类及方法，并修改该方法字节码，织入埋点逻辑。进入方法时会初始 trace 上下文信息，并存储在线程的 threadLocals 中，退出方法会打印 trace 日志并清空该方法的上下文。

数据聚合应用层 trace 日志通过 flume agents 实时发送至 flume collector；

数据存储服务端分别通过 hdfs-sink 和 hbase-sink，实时录入至 hbase、hdfs；
hdfs 有 tmp 临时文件存放实时聚合过来的数据，每5分钟生成一个 done 文件；

数据分析和统计load 程序每 4 分钟检查 done 文件并存放至 hive 表 hkymessage 指定分区；
分析程序每5分钟执行一次，将生成统计数据入库，结果集数据如下：数据格式：{5个分层的5个响应时段请求个数合集} {5个分层5-10s和大于10s散点数据合集} 当前5分钟最后一次请求rootid 统计时间

数据展示基于 Python 的 Django

基于这些数据分析和统计，我们就能绘制性能曲线图，从中可以发现哪些时间点哪些层有性能问题，然后一路点进去，直到找到到底是哪一个调用链里的哪一个环节慢。

图9 性能曲线默认图形

还可以从每一次调用结果分析出各层的异常曲线，并按照 memcached/redis/mongodb/mysql/runtime/fail 分类查看。

图10 异常曲线默认图形

还可以进一步统计各个业务工程的访问量、访问质量和平均访问时长，并于历史同期对比，从而快速理解系统服务质量。

小礼物走一走，来简书关注我

赞赏支持

 WEB架构 (/nb/2922031)

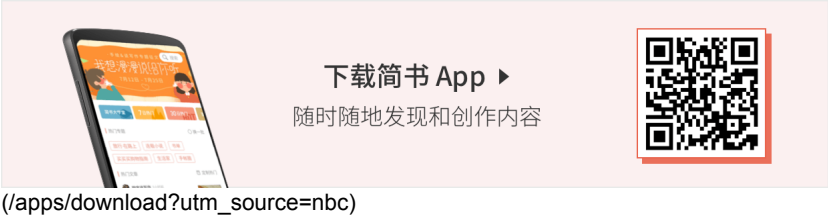
举报文章

© 著作权归作者所有

 凡墙 (/u/7ac19e0b777c)
写了 8979 字，被 62 人关注，获得了 79 个喜欢
(/u/7ac19e0b777c)

+ 关注





登录 (/sign后发表评论source=desktop&utm_medium=not-signed-in-comment-form)

2条评论

只看作者

按喜欢排序 按时间正序 按时间倒序

荆人七十 (/u/5d5645b61be3)

2楼 · 2016.07.12 17:52

(/u/5d5645b61be3)

为什么没有"zipkin是如何实现的"...

赞

回复

yiming_906e (/u/594d9cf15fc1)

3楼 · 2017.09.17 21:25

(/u/594d9cf15fc1)

很多调用场景会比上面说的完全同步的调用更为复杂，比如会遇到异步、单向、广播、并发、批处理等等，这时候需要妥善处理好 ThreadLocal 上的调用上下文，避免调用上下文混乱和无法正确释放。-- 异步和消息怎么处理，以保证获取到同一个traceID？又避免业务侵入？

赞

回复

被以下专题收入，发现更多相似内容

架构之美 (/c/164f34a02ca4?utm_source=desktop&utm_medium=notes-included-collection)

推荐阅读

更多精彩内容 > (/)

《PHP与中间件整合开发最佳实践》— 目录 (/p/e3545a56b2ff?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

多进程pcntiswooleStark 集中配置管理系统Qconf 数据库中间件php-cplCPswoolemycat 缓存系统 memcachedistair 分布式锁memcachedredis 消息系统memcachqredisHTTPSQS rpc 任务调度系统...

凡墙 (/u/7ac19e0b777c?

PHP系统声明式事务处理 (/p/34261804bc45?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

1.数据库事务 事务（Transaction）是并发控制的基本单位。所谓的事务，它是一个操作序列，这些操作要么都执行，要么都不执行，它是一个不可分割的工作单位。例如，银行转账工作：从一个账号扣款并使另一...

凡墙 (/u/7ac19e0b777c?

女子30岁不想结婚：“自己什么都能做，要男的干嘛？”... (/p/d326d1f78982?utm_campaign=maleskine&utm_content=note&utm

01 微博上有一条视频，从事会计工作的吴小姐，从26岁开始相亲，已经从排斥

到被动接受。如今年近30岁的她，仍然不想结婚。她觉得面对两个家庭跟一...

袁曲无闻 (/u/deeea9e09cbc?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation) (/apps/dutm_soi

大问题 | 不知道这些，你不可能学好英语 (/p/b19972d3... (/p/b19972d39afe?utm_campaign=maleskine&utm_content=note&utm

1. 经历了30年的英语学渣噩梦之后，我在2017年的10月份进行了一次特殊的课程。历时7天，每天12个小时的封闭式英语学习训练。2018年的第2天，我和...

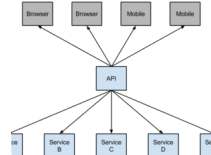
彭小六 (/u/1441f4ae075d?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

目标管理，如何制定一份高效的年度计划，实现开挂？ ... (/p/4d472cc5bd35?utm_campaign=maleskine&utm_content=note&utm

01 每年的时候我们都会去制定一年的年度计划，但是你制定的不一定会实现，去年网上有一个这样的段子就是我2017要努力实现2016未完成的2015年没有...

晓多 (/u/fee4b4b0b89e?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

(/p/46fd0faecac1?

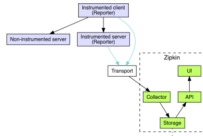


utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) Spring Cloud (/p/46fd0faecac1?utm_campaign=maleskine&utm_conte...

Spring Cloud为开发人员提供了快速构建分布式系统中一些常见模式的工具（例如配置管理，服务发现，断路器，智能路由，微代理，控制总线）。分布式系统的协调导致了样板模式，使用Spring Cloud开发人员可...

卡卡罗2017 (/u/d90908cb0d85?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/e02972487e00?

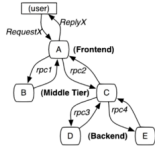


utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) 分布式调用跟踪系统调研笔记 (/p/e02972487e00?utm_campaign=maleski...

分布式调用链跟踪系统通常有几个设计目标 低侵入性 -- 作为非业务组件，应当尽可能少侵入或者无侵入其他业务系统，对于使用方透明，减少开发人员的负担； 灵活的应用策略 -- 可以（最好随时）决定所收集数据...

ginobefun (/u/0ffaa3601861?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/cdefc9971951?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) Dapper，大规模分布式系统的跟踪系统 (/p/cdefc9971951?utm_campaign=...


作者： Benjamin H. Sigelman, Luiz Andr’e Barroso, Mike Burrows, Pat Stephenson, Manoj Plakal, Donald Beaver, Saul Jaspan, Chandan Shanbhag 概述...

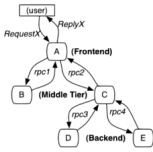
Josh_Song (/u/ab5623b0fcc4?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/f3bf4d39f1fe?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) Dapper-大规模分布式链路跟踪系统 (/p/f3bf4d39f1fe?utm_campaign=mal...

http://bigbully.github.io/Dapper-translation/ Dapper，大规模分布式系统的跟踪系统作者： Benjamin H. Sigelman, Luiz Andr’e Barroso, Mike Burrows, Pat...

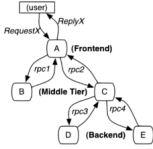
 jerrik (/u/29c8146c9b23?



(/apps/d
utm_so

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

(/p/7c719a75df8c?



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

Dapper分布式跟踪系统-翻译 (/p/7c719a75df8c?utm_campaign=maleski...

概述 当代的互联网的服务，通常都是用复杂的、大规模分布式集群来实现的。互联网应用构建在不同的软件模块集上，这些软件模块，有可能是由不同的团队开发、可能使用不同的编程语言来实现、有可能布在了...

 咖灰 (/u/56a16fd41213?

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


(/p/15d70164d4ad?



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

双12火了！吃喝玩乐背后，线下商业借互联网智能升级 (/p/15d70164d4ad?...


双12来了，这个被称为“线下版双11”的节日，每年都掀起了巨大的线下消费高潮。2014年开始，每年的“双12”成了线下商业的狂欢节。12月10日，双12第一天，早上9点多，全国各地的超市、蛋糕店、早餐店就排...

 经理人分享 (/u/aea9614b6476?

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

D151 (/p/61ab58c1a260?utm_campaign=maleskine&utm_content=note...


20171116 【幸福三朵玫瑰】 昨日 3朵玫瑰 1.早起 2.阅读✔ 3.接当当✔ 今日3朵玫瑰 1.早起 2.阅读 3.拍照
【幸福实修60天目标】 1.接纳父母，每天做父母臣服 2.让自己柔软放下来，关注自己，每天半小时冥...

 幸福实修08班叶青 (/u/e9b97ebaecf9?

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

【走心】十五年专注、一个产品的坚守！ (/p/e178f1cfe9ae?utm_campaign...


蓝凌，作为国内第一家研究与推动知识管理实践的服务商，从2001创立开始，就一直在探索知识管理的最佳模式与成功之道，2004年至今，凭借领先的知识管理市场占有率，一直被业界公认为知识管理领军品牌。 ...

 雯丽 (/u/639014b92fe1?

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

Server-Sent Events (/p/5b2f534a364e?utm_campaign=maleskine&utm...


什么是 SSE 我们先看一看一个标准的 HTTP Request / Response 的过程。 客户端与服务器建立连接，发送 HTTP request 给服务器 服务器收到客户端的 HTTP request，发送 HTTP response 给客户端 完成 HTTP...

 吞噬悲伤的影子 (/u/e73be636e7c5?

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

原来我还是很渺小 (/p/98353749ea9e?utm_campaign=maleskine&utm_c...

初中，就想梦一样飘走了，而我却傻到可怜，在等你的消息，每次你总说天下没有不散的宴席，才发现你一开始都在打算了，所以，我能怎么办，傻傻的，傻傻的，付出了那么多，换来了毕业，我们也散了，我能...

 烟花似萌 (/u/2e03e0a60416?

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



(/apps/d
utm_sou

