UPYUN：用 Erlang 开发的对象存储系统

作者：sai

2014-02-20

本文字数：3394 字 阅读完需约分钟

在国内的几家云计算创业公司当中，UPYUN（又拍云）选择了一个比较独特的定位：面向开发者提供非结构化数据云存储服务。非结构化数据存储服务一个很重要的卖点是要提供快速的静态文件访问能力，这对底层的存储系统性能和上层的 CDN 系统提出了较高的要求。

黄慧攀（ @oneoo ）是 UPYUN 技术总监。在 QCon 上海 2013 大会上，黄慧攀介绍了 UPYUN 的 CDN 系统架构，包括 Nginx 的二次开发经验、防盗链服务的实现、海量小文件的性能处理等；在 QCon 北京 2014 大会上，他将对 UPYUN 底层的对象存储系统的研发经验进行分享。

在本次采访中，黄慧攀介绍了 UPYUN 对象存储系统的一些历史，团队的分工，以及做测试方面的一些思路。

InfoQ：先介绍一下你自己吧。你关于计算机的知识都是自学，从底层网络、操作系统到上层的 Java、PHP 都玩，Lua 也玩。你对技术的选择有什么标准吗？如果有，是怎样的标准？

> 黄慧攀：我是出身于广东一个小城市“鹤山”人，最早是在 95 年接触电脑，98 年开始使用互联网，那时网易还只是做邮箱服务的，我非常感谢我的初中母校，使我能这么早期接触到互联网，影响一生 :) 也因为当年电脑、互联网才刚刚起步，学校也缺乏较好的教育能力，所以很多知识需要自学。也因为这个兴趣太浓，搞得其他学科基本都挂科了，也就没考上高中和大学。到现在还是有点小后悔，起码得把英文学好。
>
> 2001 年，18 岁的我第一份工作是市里一个集团公司的 B2B 门户网站，负责程序开发工作。那时用的语言是 PHP，边学边做的折腾了 3 年时间。
>
> 2004 年，项目因为市场、资金的原因结束了。在我们的小城市互联网就业机会基本为零，只好转到一个做弱电工程的公司任技术工程师，负责网络系统方案设计、智能灯光系统等等。
>
> 2006 年，压抑不住互联网的心，就出来创办 yo2.cn 优博网，国内第一个基于 WordPress 的博客服务平台，这个创业经历使我的技术能力提升很大，因为没人嘛，所以整个网站的事情都得自己做，开发、运维、客服，甚至设计等等。记得当时网站被人吐槽最多的就是用户体验，我想如果能把这块也做好，可以做 UED 了，哈哈。
>
> 2009 年，机缘巧合来到杭州，跟朋友做了几次创业，虽然也是失败告终，但在其中的过程使自己成长了很多，因为创业嘛，所以很多事情都必须自己做的，这就奠定了我的技术层面比较广的基础。

2011 年，收到又拍网的邀请，开始又拍云的开发工作至今。

经历这么多年和多次创业，积累到比较丰富的技术经验。知识比较全面，在看待技术选型方面的把握还是比较准的。比如：Java 的优点是适合大型项目、团队协作开发，缺点也很明显，开发周期长、人员成本相对 PHP 高一些；而 PHP 的优点则是适合中小型项目、开发周期短、人员成本低，当然弊端也很明显，不支持多线程、系统资源占用高。每个语言都有自己的优缺点，要根据项目实际情况来选择。后来一个偶然的机会接触到 Lua 语言，发现它跟 PHP 很像，但又没了 PHP 几个大缺点，非常棒。所以现在我主要使用 C 和 Lua 这两个开发语言。

InfoQ：你自己做过博客平台，也在企业网站、网游等网站做过，现在在 UPYUN，可以说是从面向消费者的.com 公司转移到了一个更加基础一些的服务。你觉得在 UPYUN 做的事情跟以前有什么不一样吗？

黄慧攀：我觉得做 UPYUN 这件事，是之前几个项目的升华吧。因为这些面向消费者的项目让我知道在开发过程中产生的痛点，从而挖掘出开发者的需求。我很高兴能为开发者服务，帮助大家更快的把项目做好。

InfoQ：能不能简单介绍一下 UPYUN 这套对象存储系统的研发历程？比如是什么时候开始做的，最初的设计者是什么背景，借鉴过哪些思路，研发的过程中有没有什么好玩的故事等等。

黄慧攀：UPYUN 的对象存储系统其实早在 08 年就开始设计的，当初用的是 MogileFS，为又拍网服务。因为早期的 MogileFS 的设计本身有一定限制，tracker 角色的元信息使用单个 MySQL 实例存储，无法满足我们日益增长的存储量，所以在 2010 年转为使用 Erlang 语言开发。设计目标是提供 PB 级别的存储服务，经历 1 年多的业务测试才正式对外开放存储服务。

选择 Erlang 语言进行开发，主要是语言本身就支持分布式，这可以节省很多开发工作。且 Erlang 语言在电信行业的应用非常广泛，稳定性有保障。

在分布式算法的选型上是参考 Dynamo 方案。而在具体的数据存储结构方面则是自主研发的一致性哈希算法，以实现多机柜、多服务器和多磁盘之间的数据备份工作。做到每文件的对应备份点在不同机柜、不同服务器上，避免某台服务器甚至某个机柜的服务器宕机而影响到文件的读写操作。

至于测试周期长达 1 年多，是因我们本身又拍网（照片社区）的数据量就非常庞大，从老的 MogileFS 集群迁移到新的云存储服务器占大部分时间，另外是因分布式存储服务的容灾测试过程比起应用测试要漫长得多，主要的测试点会有：某磁盘故障、某服务器故障、某机柜故障等好几种灾难测试，且每个故障都会产生一定量的数据迁移，文件会在集群内部自动寻找合适的备份点再建备份，所以说测试周期需要很长时间。也只有做到充分的测试，我们才放心的在集群上存储大量数据。否则等遇到无法排除的问题，要考虑新建集群的话，迁移成本和周期都会非常巨大。比如 10PB 的数据要从 A 集群迁移到 N 集群，网络传输就要 100pb，基于 10gb 网络也得耗时半年；且要保障迁移期间内不再发生新故障，这是很难做到的。所以我们选择前期测试做得非常充分，来保障日后服务的可持续性。

InfoQ：又拍云专注于做图片的存储，你们提供了一些很有特色的服务（如缩略图、防盗链），同时非常专注于服务质量。相对于文件备份类的应用场景，海量小图片存储是非常吃资源的，你们在存储系统的设计上做了哪些工作以确保在资源占用高的情况下仍然能保持图片访问的服务质量？

> 黄慧攀：是的，UPYUN 主要面向小于 100MB 的小文件提供服务，目前我们的存储集群已存有超过 2PB 的数据。面向海量小文件所面临的主要问题是：随机读取非常高、磁盘性能低；大家都知道缓存系统可以解决这类问题，而 CDN 其实就是个巨大的缓存系统，所以我们自建了 CDN 并对外提供服务。不仅能解决海量小文件所产生的磁盘性能问题，还能加速文件在互联网上的传输，一举两得。

InfoQ：UPYUN 系统的测试是如何做的？

> 黄慧攀：我们团队还比较小，目前未专门设立测试部门，所有测试工作均由项目开发者来完成，毕竟开发人员更清楚会有哪些潜在问题，并制定自动化测试的样例。下面是我们一个项目的开发、测试与发布流程：
>
> 1. 项目策划、文档和方案撰写
> 2. 开发（过程中会有两名以上开发人员交叉 review）
> 3. 本地测试（主要测试该项目的功能是否正常和程序稳定性、资源占用率等等）
> 4. 模拟平台测试（主要测试该项目的功能上线是否对原平台上其他子系统产生不良影响，这里会有我们自己编写的一批批量测试脚本，以验证平台每项功能逻辑是否正确）
> 5. 灰度测试（业务环境中抽取 1% 的服务器更新或指定某个别客户可使用该功能来进行测试）
> 6. 全网发布
>
> 从整个流程来看，我们的测试周期是比较长的，测试工作占整个项目周期 50% 以上，甚至个别影响范围大的项目，测试周期会长达半年以上。

InfoQ：你们的团队是怎样分工的？研发跟产品运营、系统运维的同学又是如何沟通的？

> 黄慧攀：大家从我们的产品介绍上会知道我们主要提供 3 块服务，
>
> 1. 云存储
> 2. 云分发
> 3. 云处理
>
> 所以我们的开发团队主要是根据这 3 个方向进行分组。现在我们团队分应用开发组和核心研发组，而在核心研发组中又分存储、分发、处理 3 个小组，分得比较细。因此我们的小组成员之间会有交叉分工，以便大家对整体系统能有充分的了解。
>
> 我们的产品服务与一般互联网服务不太一样，我们是以产品为主导而非运营主导，且我们的产品经理也是开发出身的，所以在与开发团队的协作沟通上不会存在什么问题。另外的运维部门则是更加紧密，因我们正在开始整个平台的自动化运维系统开发，我们的开发人员已走到一线，跟运维人员一起探讨运维自动化系统的功能性问题，开发人员能亲身了解运维工作和痛点，并以此来驱动运维自动化系统的开发工作。

InfoQ：这次 QCon 北京，你希望面向哪些人群进行分享？他们能从你的分享中获得什么？

> 黄慧攀：很感谢 QCon 能让我们来继续跟大家做些云计算方面的分享。在上一次的 QCon 上海大会我跟大家分享了又拍云的 CDN 技术，按我们公司的服务层次划分，这次的分享主题是我们在云存储系统的研发和构

> 建过程中遇到的一些问题和经验。希望大家能通过我们这次的分享，对云存储能有更深入的了解，比如分布式算法、存储结构和日常维护等等。

发布于：2014-02-20 00:50

QCon Erlang 服务革新 音视频（后端） DevOps & 平台工程 语言 & 开发 性能优化 操作系统 编程语言 技术选型

轻点一下，留下你的鼓励

评论

快抢沙发！虚位以待

发布

· 暂无评论 ·

更多内容推荐

Node.js 实现存储服务的上传功能【包含前后端代码】

上传和下载功能是存储服务非常基础的功能，也是存储服务日常使用过程中最常用的功能，比如阿里云的OSS、腾讯云的COS、百度云…

2021-08-09

不忘初心，砥砺前行|暨 InfoQ 写作平台一周年

新的机遇和挑战即将开始，珍惜每一次选择的机会，不断夯实基础，保持所在领域技术和业务的敏感度，相信未来一定能有所收获。 开…

2021-04-20

全链路压测平台（Quake）在美团中的实践

本文来自美团点评技术文章系列。

文化 & 方法，性能优化，中间件，操作系统，编程语言，框架，微服务，在离线混部，实时计算

Apache Pulsar：云原生时代的消息服务｜QCon

当新生事物出现时，人们总是有两种角度去观察它，要么把它看小，要么把它放大。

QCon，视频，方法论，最佳实践，云原生，开源，性能优化，多云/混合云，在离线混部

朱建平：如何架构海量存储系统

5月25日，互联网架构技术沙龙圆满落幕。本期沙龙特邀请腾讯的技术专家分享关于技术架构、落地实践案例、无服务器云函数架构、海…

文化 & 方法，架构，方法论，性能优化，微服务，在离线混部，技术选型

mac 安装特定版本 php-redis

mac中的brew是一个方便的工具，但是有时候也不方便。比如：服务器线上的服务可能使用的版本比较特殊，本地也需要这个特殊的版本…

2020-05-26

羽量级实现灵活通用的微服务流量分发

伴随着业务的飞速发展，达达集团内部的微服务数量和节点个数也都在不断增长。当业务逻辑和运行环境越来越复杂，简单的服务发现和…

架构，微服务，最佳实践，性能优化，中间件，框架，在离线混部

什么是大数据：从 GFS 到 Dataflow，12 年大数据生态演化图

要想学好大数据，我们需要先正本清源，弄清楚大数据在技术上到底涵盖了些什么。所以今天这节课，我就从大数据技术的核心理念和历…

2021-09-15

The Google File System （二）： 如何应对网络瓶颈?

在“大数据”爆发之后，数据中心的大量数据传输变成了数据中心的服务器横向之间的传输，而这个也让工程师们开始重新基于需求，重新…

2021-09-27

InfoQ 专访腾讯高级工程师：小程序云开发即将发布实时数据推送服务

可以监听云数据库的数据变更，实时推送到小程序端。的成本，是小程序中实时推送的高效实践方案。本文是InfoQ前端之巅对腾讯高级…

文化 & 方法，方法论，腾讯，性能优化，框架

按时上下班的程序员，做出来的东西没有“弹性” | DIVE 基础软件大会专访

编辑｜辛晓亮采访嘉宾｜刘新铭 开发效率跟每个开发者和开发团队息息相关，高效率意味着可以在更快的时间内更好的完成更多的内容。

文化 & 方法，语言 & 开发，技术管理，InfoQ大会-大咖说，方法论，技术选型，性能优化，中间件，操作系统，编程语言，芯片，数字化转型

百分点大规模文件存储 OSS 技术与实践

本文介绍百分点基于实践探索自主研发出的大规模文件存储OSS技术

安全，架构，硬件，软件工程，最佳实践，性能优化，音视频（前端），音视频（后端），编程语言，实时计算

CDN 搭配 OSS 最佳实践 ——搭建动静态分离的应用架构

传统的网站产品应用架构，所有资源部署在应用服务器本地存储或挂载的数据存储区，对于动静态资源不作分离

架构，服务革新，软件工程，方法论，性能优化，在离线混部，实时计算

海量小文件存储系统 HOS 探索与实践

对象存储业界较为普遍解决方案，一是对小文件进行合并处理，二是构建高速缓存；HBase2.0之后支持的MOB新特性可以满足中小对象…

2020-12-19

14｜中国芯片现状与机会（下）

这一讲，聊聊中国芯片行业的就业机会

2021-06-18

加餐（七） | 从微博的 Redis 实践中，我们可以学到哪些经验？

俗话说“他山之石，可以攻玉”，学习掌握这些经验，可以帮助我们在自己的业务场景中更好地应用Redis。

2020-11-30

01｜拨云见日——云上架构一点儿也不神秘

2022-09-21

Dubbo 的服务注册与调用

作业：根据微服务框架 Dubbo 的架构图，画出 Dubbo 进行一次微服务调用的时序图

2020-08-12

图解大型网站技术架构的历史演化过程

开篇明义：【大型网站技术架构笔记】系列是阅读《大型网站技术架构核心原理与实践》一书的一些笔记，记录了原书的一些重要内容以...

文化 & 方法，架构，方法论，性能优化，中间件，框架

以史鉴今：监控是如何一步步发展而来的？

可观测性是怎样发展而来的？让我们从监控的源头讲起。

2022-09-14

发现更多内容