



软件架构探索

The Fenix Project Alpha

Watch 6 Star 50 Follow 60

Release v1.0.20200608 License Apache 2.0 Doc License CC 4.0 Words 148,600 Author IcyFenix

周志明

icyfenix@gmail.com

发行日期：2020-06-08

这是什么？

这是一部以高级程序员、系统架构师为目标读者的技术手册，是一部以软件设计、架构工作中“要考虑哪些因素、需解决哪些问题、有哪些行业标准的解决方案”为主题的开源文档。文章《[什么是“The Fenix Project”](#)》详细阐述了此项目主旨、目标与名字的来由，文章《[如何开始](#)》简述了文档每章讨论的主要话题与内容详略分布，供阅前参考。

笔者出于以下目的，撰写这部文档：

- 笔者从事大型企业级软件的架构研发工作，借此机会，系统性地整理自己的知识，查缺补漏，将它们都融入既有的知识框架之中。
- 笔者正式出版过七本计算机技术书籍，遗憾的是没有一本与自己本职工作直接相关。能按照自己的兴趣去写作，还能获得不菲的经济报酬是一件很快乐的事情；撰写一部工作中能直接使用的、能随时更新、与人交流的在线文档，也是一件颇为实用、颇具价值的事情。
- 笔者认为技术人员成长有一“捷径”，学技术不仅要去看、去读、去想、去用，更要去说、去写。将自己“认为掌握了的”知识叙述出来，能够说得有条理清晰，讲得理直气壮；能够让他人听得明白，释去心中疑惑；能够把自己的观点交予别人的审视，乃至质疑，在此过程之中，会挖掘出很多潜藏在“已知”背后的“未知”。未有知而不行者，知而不行，只是未知。

除文档部分外，笔者同时还建立了若干配套的代码工程，这是针对不同架构、技术方案（如单体架构、微服务、服务网格、无服务架构，等等）的演示程序（[PetStore-Like-Project](#)）。它们既是文档中所述知识的实践示例，亦可作为实际项目新创建时的可参考引用的基础代码。

如何使用？

根据“使用”的所指含义的不同，笔者列举以下几种情况：

- **在线阅读**：本文档在线阅读地址为：<https://icyfenix.cn>。
网站由[GitHub Pages](#) 提供网站存储空间；由[Travis-CI](#) 提供的持续集成服务实时把

Git 仓库的 Markdown 文档编译同步至网站；由[腾讯云 CDN](#) 提供国内访问的缓存支持。

- **离线阅读：**

- 部署离线站点：文档基于[Vuepress](#) 构建，如你希望在企业内部搭建文档站点，请使用如下命令：

```
# 克隆获取源码  
$ git clone https://github.com/fenixsoft/awesome-fenix.git && cd  
awesome-fenix  
  
# 安装工程依赖  
$ npm install  
  
# 运行网站，地址默认为http://localhost:8080  
$ npm run dev
```

sh

- 生成PDF文件：工程源码中已带有基于[vuepress-plugin-export](#) 改造（针对本文档定制过）的PDF导出插件，如你希望生成全文 PDF 文件，请使用如下命令：

```
# 编译PDF，结果将输出在网站根目录  
$ npm run export
```

sh

PDF 全文编译时间较长，在笔者机器上约耗时25分钟，在 Travis-CI 上约需要约6分钟。PDF 中文字体采用阿里巴巴普惠字体渲染，此字体被允许免费使用与传播。

- **二次演绎、传播和发行：**本文档中所有的内容，如引用其他资料，均在文档中明确列出资料来源，一切权利归属原作者。除此以外的所有内容，包括但不限于文字、图片、表格，等等，均属笔者原创，这些原创内容，笔者声明以[知识共享署名 4.0](#) 发行，只要遵循许可协议条款中**署名、非商业性使用、相同方式共享**的条件，你可以在任何地方、以任何形式、向任何人使用、修改、演绎、传播本文档中任何部分的内容。详细可见本文档的“协议”一节。
- **运行技术演示工程：**笔者专门在探索起步中的“[技术演示工程](#)”详细介绍了配套工程的使用方法，如果你对构建运行环境也有所疑问，在附录中的“[环境依赖](#)”部分也已包括了详细的环境搭建步骤。此外，这些配套工程也均有使用 Travis-CI 提供的持续集成服务，

自动输出到 Docker 镜像库，如果你只关心运行效果，或只想了解部分运维方面的知识，可以直接运行 Docker 镜像而无需关心代码部分。你可以通过下面所列的地址，查看到本文档所有工程代码和在线演示的地址：

- 文档工程：

- 软件架构探索：<https://icyfenix.cn>
- Vuepress 支持的文档工程：<https://github.com/fenixsoft/awesome-fenix>

- 前端工程：

- Mock.js 支持的纯前端演示：<https://bookstore.icyfenix.cn>
- Vue.js 2 实现前端工程：<https://github.com/fenixsoft/fenix-bookstore-frontend>

- 后端工程：

- Spring Boot 实现单体架构：https://github.com/fenixsoft/monolithic_arch_springboot
- Spring Cloud 实现微服务架构：https://github.com/fenixsoft/microservice_arch_springcloud
- Kubernetes 为基础设施的微服务架构：https://github.com/fenixsoft/microservice_arch_kubernetes
- Istio 为基础设施的服务网格架构：https://github.com/fenixsoft/servicemesh_arch_istio
- 基于云端的无服务架构：https://github.com/fenixsoft/serverless_arch

协议

- 本文档代码部分采用 [Apache 2.0 协议](#) 进行许可。遵循许可的前提下，你可以自由地对代码进行修改，再发布，可以将代码用作商业用途。但要求你：
 - 署名：在原有代码和衍生代码中，保留原作者署名及代码来源信息。
 - 保留许可证：在原有代码和衍生代码中，保留 Apache 2.0 协议文件。
- 本作品文档部分采用 [知识共享署名 4.0 国际许可协议](#) 进行许可。遵循许可的前提下，你可以自由地共享，包括在任何媒介上以任何形式复制、发行本作品，亦可以自由地演绎、修改、转换或以本作品为基础进行二次创作。但要求你：
 - 署名：应在使用本文档的全部或部分内容时候，注明原作者及来源信息。

- **非商业性使用**：不得用于商业出版或其他任何带有商业性质的行为。如需商业使用，请联系作者。
- **相同方式共享的条件**：在本文档基础上演绎、修改的作品，应当继续以知识共享署名4.0国际许可协议进行许可。

目录

数据统计

列入目录文章 128 篇，目前已完成 54 篇，合计总字数 148,600 字，最后更新日期 2020-06-08。

1. 目录	54 字
2. 前言	
2.1. 关于作者	642 字
2.2. 什么是“The Fenix Project”	3,222 字
3. 探索起步	
3.1. 阅读指引	
3.1.1. 更新日志	176 字
3.1.2. 如何开始	2,634 字
3.2. 技术演示工程	327 字
3.2.1. 前端工程	2,243 字
3.2.2. 单体架构：Spring Boot	2,538 字
3.2.3. 微服务：Spring Cloud	3,454 字
3.2.4. 微服务：Kubernetes	3,534 字
3.2.5. 服务网格：Istio	7 字
3.2.6. 无服务：Serverless	6 字
4. 演进中的架构	
4.1. 服务架构演进史	579 字
4.1.1. 原始分布式时代	1,388 字

4.1.2. 单体系统时代	1,137 字
4.1.3. SOA时代	2,539 字
4.1.4. 微服务时代	2,681 字
4.1.5. 后微服务时代	2,178 字
4.1.6. 无服务时代	1,352 字

5. 设计者的视角

5.1. 架构的普适问题	491 字
5.1.1. 服务设计风格	664 字
5.1.1.1. 远程服务调用	1,805 字
5.1.1.2. RESTful服务	10,435 字
5.1.1.3. 异步服务调用	7 字
5.1.2. 事务处理	990 字
5.1.2.1. 本地事务	625 字
5.1.2.2. 全局事务	3,454 字
5.1.2.3. 共享事务	1,182 字
5.1.2.4. 分布式事务	8,574 字
5.1.3. 透明多级分流系统	1,262 字
5.1.3.1. 客户端缓存	3,522 字
5.1.3.2. 域名解析	1,893 字
5.1.3.3. 链路优化	4,731 字
5.1.3.4. 内容分发网络	3,652 字
5.1.3.5. 负载均衡	7,664 字
5.1.3.6. 缓存中间件	55 字
5.1.3.7. 数据库扩展	25 字
5.1.4. 安全架构	683 字
5.1.4.1. 认证	3,167 字
5.1.4.2. 授权	7,410 字
5.1.4.3. 凭证	4,997 字

5.1.4.4. 保密	3,770 字
5.1.4.5. 传输	7,658 字
5.1.4.6. 验证	2,678 字
5.1.4.7. 漏洞利用	51 字
5.1.5. 高效并发	5 字
5.1.5.1. 进程、线程与协程	9 字
5.1.5.2. 线程安全	5 字
5.1.5.3. 同步机制	40 字
5.1.5.4. 硬件并发机制	7 字
5.2. 设计方法论	
5.2.1. 系统分层	5 字
5.2.2. 容量规划	5 字
5.2.3. 非功能属性设计	48 字
6. 分布式的基石	
6.1. 分布式共识算法	1,847 字
6.1.1. Paxos	4,401 字
6.1.2. Multi Paxos与Raft	3,216 字
6.1.3. Gossip协议	1,803 字
6.2. 服务互联	
6.2.1. 服务发现	4,743 字
6.2.2. 路由与网关	6 字
6.2.3. 进程内负载均衡	8 字
6.2.4. 服务编排	5 字
6.2.5. 中心化配置	6 字
6.3. 流量管控	
6.3.1. 隔离	3 字
6.3.2. 熔断	3 字
6.3.3. 超时	3 字

6.3.4. 流控	3 字
6.3.5. 降级	3 字
6.3.6. 异常注入	5 字
6.4. 可靠通讯	
6.4.1. 流量加密	5 字
6.4.2. 访问策略	5 字
6.4.3. 事件审计	5 字
6.5. 可观测性	
6.5.1. 日志聚合	5 字
6.5.2. 链路跟踪	5 字
6.5.3. 应用性能管理	7 字
7. 不可变基础设施	
7.1. 虚拟化的概念与历史	10 字
7.2. 虚拟化容器	6 字
7.2.1. CRI接口	4 字
7.2.2. 资源隔离	18 字
7.2.3. 资源对象	5 字
7.2.4. 容器管理	5 字
7.3. 容器间网络	6 字
7.3.1. CNI接口	4 字
7.3.2. 网络策略	5 字
7.3.3. 网络插件	5 字
7.3.4. 容器负载均衡	7 字
7.4. 配置与数据持久化	9 字
7.4.1. CSI接口	4 字
7.4.2. 分布式文件系统	8 字
7.4.3. 共享存储插件	7 字
7.5. GPU虚拟化	5 字

7.5.1. Device Plugin机制	5 字
7.5.2. 调度GPU	4 字
7.5.3. Nvidia插件	4 字
7.6. 扩展基础设施	7 字
7.6.1. CRD定义	4 字
7.6.2. 自定义API Server	6 字
7.7. 硬件资源调度	7 字

8. 技巧与专题

8.1. Graal VM	1,327 字
8.1.1. 新一代即时编译器	851 字
8.1.2. 向原生迈进	1,729 字
8.1.3. 没有虚拟机的Java	1,846 字
8.1.4. Spring over Graal	2,678 字
8.2. 响应式编程	6 字
8.3. 函数式接口与流式编程思想	13 字
8.4. 并行、异步、非阻塞	10 字

9. 附录

9.1. 构建发布脚本	7 字
9.2. 持续集成	5 字
9.3. 灰度发布	5 字
9.4. 部署环境	72 字
9.4.1. 部署Docker CE容器环境	2,123 字
9.4.2. 部署Kubernetes集群	486 字
9.4.2.1. 使用Kubeadm部署	4,563 字
9.4.2.2. 使用Rancher部署	1,647 字
9.4.2.3. 使用Minikube部署	634 字
9.5. 运维环境	5 字

9.5.1. 在K8S上部署ELK/EFK日志监控	11 字
9.5.2. 在K8S上部署DevOps	7 字

关于作者

周志明

- 80后 程序员

职业上是上市软件公司高层管理人员，但自己不愿离开技术领域，不愿脱离一线程序员的行列。

职业上是从事偏宏观的大型企业级软件的架构研发，自己对高级语言虚拟机、程序语言设计、编译原理等偏底层、微观的方向也很感兴趣。

- 远光研究院 院长

博士，现任[远光软件](#)研究院院长。研究方向为机器学习，特征选择自动化。

- 计算机技术作家

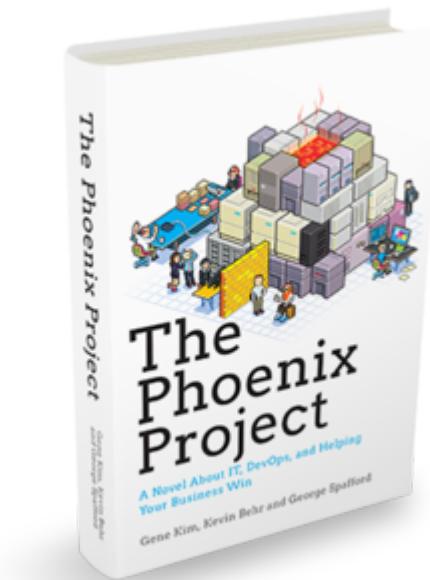
已正式出版过七部计算机技术书籍，撰写过两部开源文档，口碑和销量均有幸得到读者的认可。其中四本书在[豆瓣](#)上获得了9.0分或以上的评价，“深入理解Java虚拟机”系列重印超过40次，总销量逾30万册。

- 2020年 《软件架构探索：The Fenix Project》（Open Document，进行中）
- 2019年 《深入理解Java虚拟机：JVM高级特性与最佳实践（第三版）》（豆瓣 9.6）
- 2018年 《智慧的疆界：从图灵机到人工智能》（豆瓣 9.1）
- 2016年 《深入理解Java虚拟机：JVM高级特性与最佳实践（第二版）》（豆瓣 9.0）
- 2015年 《Java虚拟机规范（Java SE 8中文版）》（官方授权第二译者，豆瓣 8.0）
- 2014年 《Java虚拟机规范（Java SE 7中文版）》（官方授权第一译者，豆瓣 9.0）
- 2013年 《深入理解OSGi：Equinox原理、应用与最佳实践》（豆瓣 7.7）
- 2011年 《深入理解Java虚拟机：JVM高级特性与最佳实践（第一版）》（豆瓣 8.6）

- 2011年 《Java虚拟机规范（Java SE 7中文版）》（Open Document，第一译者）
- 技术布道师
开源技术的积极倡导者和推动者，对计算机科学相关的多个领域都有持续跟进。
 - 腾讯云最有价值技术专家（TVP）
 - 阿里云最有价值技术专家（MVP）
 - InfoQ.CN专栏撰稿人

什么是“The Fenix Project”

“Phoenix”这个词东方人不常用，但在西方的软件工程读物——尤其是关于Agile、DevOps话题的作品中时常出现。软件工程小说《The Phoenix Project》讲述了徘徊在死亡边缘的Phoenix项目在精益方法下浴火重生的故事；马丁·福勒（Martin Fowler）对《Continuous Delivery》的诠释里，曾多次提到“Phoenix Server”（取其能够“涅槃重生”之意）与“Snowflake Server”（取其“世界上没有相同的两片雪花”之意）的优劣比对。也许是东西方的文化的差异，尽管有“失败是成功之母”这样的谚语，但我们东方人的骨子里更注重的还是一次把事做对做好，尽量别出乱子；而西方人则要“更看得开”一些，把出错看做正常甚至是必须的发展过程，只要出了问题能够兜底使其重回正轨便好。



The Phoenix Project

在软件工程里，任何产品的研发，只要时间尺度足够长，人就总会疏忽犯错，代码就总会携有缺陷，电脑就总会宕机崩溃，网络就总会堵塞中断……如果一项工程需要大量的人员，共同去研发某个大规模的软件产品，并使其分布在网络中大量的服务器节点中同时运行，随着项目规模的增大、运作时间变长，其必然会受到墨菲定律的无情打击。

Murphy's Law : Anything that can go wrong will go wrong

为了得到高质量的软件产品，我们是应该把精力更多地集中在提升其中每一个人员、过程、产出物的能力和质量上，还是该把更多精力放在整体流程和架构上？

笔者先给这个问题一个“合稀泥”式的回答：这两者都重要。前者重术，后者重道；前者更多与编码能力相关，后者更多与软件架构相关；前者主要由开发者个体水平决定，后者主要由技术决策者水平决定；

然而，笔者也必须强调此问题的另外一面：这两者的理解路径和抽象程度是不一样的。如何学习一项具体的语言、框架，工具，譬如Java、Spring、Vue.js……都是相对具象的，不论其蕴含的内容多少，复杂程度高低，它是至少能看得见摸得着。而如何学习某一种风格的架构方法，譬如单体、微服务、服务网格、无服务、云原生……则是相对抽象的，谈论它们可能要面临则“一百个人眼中有一百个哈姆雷特”的困境。谈这方面的话题，若要言之有物，就不能是单纯的经验陈述。笔者想来，回到这些架构根本的出发点和问题上，真正去使用这些不同风格的架构方法来实现某些需求，解决某些问题，然后在实践中观察它们的异同优劣，会是一种很好的，也许是最好的讲述方式。笔者想说一下这些架构，而且还想说得透彻明白，这需要代码与文字的配合，不适合直接以传统书籍的形式撰写，于是，便有了这个项目。

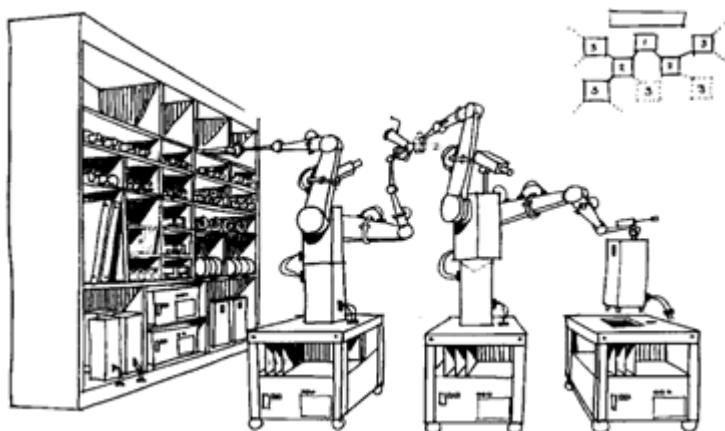
可靠的系统

让我们再来思考一个问题，构建一个大规模但依然可靠的软件系统，是否是可行的？

这个问题令人听起来的第一感觉也许会有点荒谬：废话。如果这个事情从理论上来说就是根本不可能的话，那我们这些软件开发从业人员现在还在瞎忙活些什么？但你再仔细想想，前面才提到的“墨菲定律”和在“大规模”这个前提下必然会遇到的各种不靠谱的人员、代码、硬件、网络等因素，从中能得出的一个听起来颇为符合逻辑直觉的推论：如果一项工作要经过多个“不靠谱”的过程相互协作来完成，其中的误差应会不断地累积叠加，导致最终结果必然不能收敛稳定才对。

这个问题也并非杞人忧天庸人自扰式的瞎操心，计算机之父冯·诺依曼（John von Neumann）在1940年代末期，曾经花费了大约两年时间，研究这个问题并且得出了一门理论《自复制自动机》（Theory of Self-Reproducing Automata），这个理论以机器应该如何从基本的部件中构造出与自身相同的另一台机器引出，其目的并不是想单纯地模拟或者理解生

物体的自我复制，也并不是简单想制造自我复制的计算机，他的最终目的就是想回答一个理论问题：如何用一些不可靠的部件来构造出一个可靠的系统。



当时自复制机的艺术表示

自复制机恰好就是一个最好的用不可靠部件构造的可靠的系统例子。这里，“不可靠部件”可以理解为构成生命的大量细胞、甚至是分子。由于热力学扰动、生物复制差错等因素干扰，这些分子本身并不可靠。但是生命系统之所以可靠的本质，恰是因为它可以使用不可靠的部件来完成遗传迭代。这其中的关键点便是承认细胞等这些零部件可能会出错，某个具体的零部件可能会崩溃消亡，但在存续生命的微生态系统中一定会有其后代的出现，重新代替该零部件的作用，以维持系统的整体稳定。在这个微生态里，每一个部件都可以看作一只不死鸟（Phoenix），它会老迈，而之后又能涅槃重生。

架构的演进

软件架构风格从大型机（Mainframe），到多层单体架构（Monolithic），到分布式（Distributed），到微服务（Microservices），到服务网格（Service Mesh），到无服务（Serverless）……技术架构上确实呈现出“从大到小”的发展趋势。当近年来微服务兴起以后，涌现出各类文章去总结、赞美微服务带来的种种好处，诸如简化部署、逻辑拆分更清晰、便于技术异构、易于伸缩拓展应对更高的性能等等，这些当然都是重要优点和动力。可是，如果不拘泥于特定系统或特定某个问题，以更宏观的角度来看，前面所列这种种好处却都只能算是“锦上添花”、是属于让系统“活得更好”的动因，肯定比不上系统如何“确保生存”的需求来得关键、本质。笔者看来，架构演变最重要的驱动力，或者说这种“从大到小”趋势的最根本的驱动力，始终都是为了方便某个服务能够顺利地“死去”与“重生”而设计的，个体服务的生死更迭，是关系到整个系统能否可靠续存的关键因素。

举个例子，譬如某企业中应用的单体架构的Java系统，其更新、升级都必须要有固定的停机计划，必须在特定的时间窗口内才能按时开始，必须按时结束。如果出现了非计划的宕机，那便是生产事故。但是软件的缺陷不会遵循领导定下的停机计划来“安排时间出错”，为了应对缺陷与变化，做到不停机地检修，Java曾经搞出了OSGi和JVMTI Instrumentation等这样复杂的HotSwap方案，以实现给奔跑中的汽车更换轮胎这种匪夷所思却又无可奈何的需求；而在微服务架构的视角下，所谓系统检修，不过只是一次在线服务更新而已，先停掉1/3的机器，升级新的软件版本，再有条不紊地导流、测试、做金丝雀发布，一切都是显得如此理所当然、平淡寻常；而在无服务架构的视角下，我们甚至都不可能去关心服务所运行的基础设施，连机器是哪台都不必知道，停机升级什么的就根本无从谈起了。

流水不腐，有老朽，有消亡，有重生，有更迭才是正常生态的运作合理规律。请设想一下，如果你的系统中每个部件都符合“Phoenix”的特性，哪怕其中某些部件采用了由极不靠谱的人员所开发的极不靠谱程序代码，哪怕存有严重的内存泄漏问题，最多只能服务三分钟就一定会崩溃。而即便这样，只要在整体架构设计有恰当的、自动化的错误熔断、服务淘汰和重建的机制，在系统外部来观察，整体上仍然有可能表现出稳定和健壮的服务能力。

The Fenix Project

在企业软件开发的历史中，一项新技术发布时，常有伴以该技术开发的“宠物店（PetStore）”作为演示的传统（如[J2EE PetStore](#)、[.NET PetShop](#)、[Spring PetClinic](#)等）。作为不同架构风格的演示时，笔者本也希望能遵循此传统，却无奈从来没养过宠物，遂改行开了书店（Fenix's Bookstore），里面出售了几本笔者撰写过的书籍，算是夹带一点私货，同时也避免了使用素材时可能的版权问题。

尽管相信没有人会误解，但笔者最后还是多强调一句，Oracle、Microsoft、Pivotal等公司设计宠物店的目的绝不是为了日后能在网上贩卖小猫小狗，只是纯粹的演示技术。所以也请勿以“实现这种学生毕业设计复杂度的需求，引入如此规模的架构或框架，纯属大炮打苍蝇，肯定是过度设计”的眼光来看待接下来的“Fenix's Bookstore”项目。相反，如果可能的话，笔者会在有新的技术、框架发布出来时，持续更新，以恰当的形式添加到项目的不同版本中，可能使其技术栈越来越复杂。笔者希望把这些新的、不断发展的知识，融合进已有的知识框架之中，让自己学习、理解、思考，然后将这些技术连同自己的观点看法，传播给感兴趣的人。

也算是缘分，网名“IcyFenix”在二十多年前我的中学时代开始使用，最初它是来源于暴雪公司的即时战略游戏《星际争霸》的Protoss英雄Fenix——如名字所预示的那样，他曾经是Zealot，牺牲后以Dragoon的形式重生，带领Protoss与刀锋女王Kerrigan继续抗争。尽管中学时期我已经笃定自己未来肯定会从事信息技术相关的工作，但显然不可能预计到二十年后我会写下这些文字。

所以，既然我们要开始一段关于“Phoenix”的代码与故事，那便叫它“The Fenix Project”，如何？

更新日志

2020年5月25日

- 提供了新的架构演示“[基于Kubernetes实现的后端工程](#)”

2020年5月15日

- 完成“[服务发现](#)”章节

2020年5月9日

- 完成“[分布式共识算法](#)”章节

2020年5月5日

- 创建更新日志页面
- 在[目录](#)中增加根据Git提交时间生成的内容更新标识

2020年5月2日

- 完成“[服务架构演进史](#)”章节
- 查了Git文档是在2019年12月23日创建的，今天在[微博](#)上开始小范围公开

如何开始

《软件架构探索》这个开源文档项目的是笔者对自己在软件架构方面知识的总结，它是完全免费开放的，但免费的、开源的文档并不意味着你使用它时就没有成本，也不见得这个文档中所有的内容对每一个开发人员来说都是必要的。为了避免浪费阅读者的时间和精力，笔者除了自身力求在知识点准确性和叙述流畅性方面保证质量之外，同时也在本文中简要介绍每一章的主题和所面向的读者类型，本文档各章节之间并没有明显的前后依赖关系，阅读时有针对性的查阅是完全可行的，无需一篇不漏地顺序阅读，在[此目录](#)中列出了各文章的明细及字数，希望有助于你制定阅读计划。

探索起步

字数: 14,743 字

本章面向于准备对文档介绍的内容亲身实践的探索者。

本章可以视为这个文档附带示例工程的说明书，以及相应运行环境的部署手册。之所以将它安排在目录的第一位，是因为笔者相信如果你是一名驾驶初学者，最合理的学习路径应该是先把汽车发动，然后慢慢行驶起来，而不是马上从“引擎动力原理”、“变速箱构造”入手去设法深刻地了解一台汽车。相信计算机技术也是同理，先从运行程序，看看效果，搭建好开发、调试环境，对即将进行的工作有一个整体的认知开始是很有好处的。

提示

本文档所涉及到的工程均在GitHub上存有独立的项目，以方便构建、阅读、运行和fork。

本章中的部分内容，是由这些工程的README.md文件人工同步而来，并没有通过持续集成工具自动处理，所以可能有偶尔更新不一致的情况，如可能，建议到这些项目的GitHub页面上查看最新情况，在[右上角](#)有相应的超链接。如这些工程对你有用，望请不吝给个

 Star 50。

本章所提供的工程是作为后面所述知识的演示样例，由于数量确实不少，笔者建议并无必要一次性地把上面所有的工程都运行起来，这样很无聊。因为它们是采用不同的技术来解

决同一个问题，所以每个工程执行后，最终看到的界面效果均是一样的，只是实现的架构不同。笔者的建议是不妨选择一种你目前关注的架构风格去运行起来（第一章 探索起步）、思考这种架构涉及到哪些标准方案（第二章 设计者的视角）、现在提倡的微服务、云原生解决了以前存在的哪些问题（第三章 演进中的架构），它们是如何解决这些问题的（第四、五章 核心技术支撑点与不可变基础设施）。把一个架构风格的所有内容都阅读完毕之后，再开始另一个。

设计者的视角

字数: 81,569 字

本章面向于技术架构师、系统设计、开发人员。

本章是整部文档集中分析理论的章节，作为后续实践的基础，将介绍的是普适的架构技术、技巧与方法论，无论你是否关注微服务、云原生这些概念，无论你是从事架构设计还是从事编码开发，了解这里所列的基础知识，对每一个技术人员都是有价值的。

“架构师”这个词的外延非常宽泛，不同语境中有不同所指，本章中的技术架构师特指的是企业架构中面向技术模型的系统设计者，这意味着本章讨论范围不会涉及到贴近于企业战略、业务流程的系统分析、信息战略设计等内容，而是聚焦于贴近一线研发人员的技术方案设计者。本章将介绍作为一个设计者，你应该在做架构设计时思考哪些问题，有哪些主流的解决方案和行业标准做法，各种方案有什么优点、缺点，不同的解决方法会带来什么不同的影响，等等。以达到将“架构设计”这种听起来抽象的工作具体化、具象化的目的。

演进中的架构

字数: 11,931 字

本章面向于技术架构师，尤其是刚刚从单体架构向微服务架构转型的架构师。

架构并不是“发明”出来的，是持续进化的结果。“服务架构演进史”这部分，借讨论历史之名，来梳理微服务发展里程中出现的大量名词、概念，借着微服务的演变过程，我们将从这些概念起源的最初，去分析它们是什么、它们取代了什么、以及它们为什么能够在斗争中取得成功，为什么变得不可或缺的支撑，又或者它们为什么会失败，在竞争中被淘汰，或逐渐湮灭于历史的烟尘当中。“从架构到实现”这部分，介绍了各种目前仍然常用架构将通过什么技术框架，以什么方式落地，这部分内容我们将分析、借鉴业界一些优秀框架解决技术问题的思路。

分布式的基石

字数: 16,097 字

本章面向于使用分布式架构的开发人员。

只要选择了分布式架构，无论是SOA、微服务、服务网格或者其他架构风格，涉及与远程服务交互时，服务的注册发现、跟踪治理、负载均衡、故障隔离、认证授权、伸缩扩展、传输通讯、事务处理，等等，这一系列问题都是无可避免的。不同的架构风格，其区别是到底要在技术规范上提供统一的解决方案，还是由应用系统自行去解决，又或者在基础设施层面将一类问题隔离掉，本章将会讨论这类问题的解决思路、方法和常见工具。

不可变基础设施

字数: 145 字

本章面向于基础设施运维人员、技术平台的开发者。

“不可变基础设施”这个概念由来已久。2012年Martin Flower设想的“[凤凰服务器](#)”与2013年Chad Fowler正式提出的“[不可变基础设施](#)”，都阐明了基础设施不变性所能带来的益处。在[CNCF](#)定义的“云原生”概念中，“不可变基础设施”提升到了与微服务平级的重要程度，此时它的内涵已不再局限于方便运维、程序升级和部署的手段，而是升华为向应用代码隐藏分布式架构复杂度、让分布式架构得以成为一种可普遍推广的普适架构风格的必要前提。在[云原生时代](#)、[后微服务时代](#)中，软件与硬件之间的界线已经彻底模糊，无论是基础设施的运维人员，抑或是技术平台的开发人员，都有必要深入理解基础设施不变性的目的、原理与实现途径。

技巧与专题

字数: 8,460 字

本章无特定读者对象，内容全凭笔者心情。

这部分是一些笔者所了解的开发、设计中常见技巧和编程模式的集合，由于它们还不具备足够的系统性，没有安排入前面的知识框架之中。但有一些或精彩，或有价值，或实用的技巧，笔者不想错过，所以安排了这一章相对独立的内容。

附录

字数: 9,565 字

本章面向刚刚开始接触云原生环境的设计者、开发者。

这一章内容主要是云原生环境搭建和程序发布过程，原本它们并不属于笔者准备讨论的重点话题，至少没有到单独开一章的必要程度。但由于容器化的服务编排环境本身构建、管理和运维都有一定的复杂性，尤其是在国内特殊的网络环境下，无法直接访问到Google等国外的代码仓库，以至于不得不通过手工预载镜像或者代理的方式来完成环境搭建。为了避免刚刚接触这一领域的读者在入门第一步就受到不必要的心理打击，笔者专门设置了这个目录章节。这章与其他几章讨论设计思想、实现原理的风格差异很大，它是整部文档唯一的讨论具体如何操作的内容。

市面上介绍如何安装环境的书籍、资料已经不计其数，肯定有相当一部分读者这章的内容本身就是了解的，已掌握的读者建议无需仔细阅读，在有需要的时候，可当作工具查阅。

技术演示工程

除文档部分外，笔者同时还建立了若干配套的代码工程，这是针对不同架构、技术方案（如单体架构、微服务、服务网格、无服务架构、云原生，等等）的演示程序（[PetStore-Like-Project](#)）。它们即是文档中所述知识的实践示例，亦可作为实际项目新创建时的可参考引用的基础代码。

本小节内容是由这些工程的README.md文件同步而来，由于未经过持续集成工具自动处理，所以可能有偶尔更新不一致的情况，如可能，建议到这些项目的GitHub页面上查看最新情况。

- 文档工程：
 - 软件架构探索：<https://icyfenix.cn>
 - Vuepress支持的文档工程：<https://github.com/fenixsoft/awesome-fenix>
- 前端工程：
 - Mock.js支持的纯前端演示：<https://bookstore.icyfenix.cn>
 - Vue.js 2实现前端工程：<https://github.com/fenixsoft/fenix-bookstore-frontend>
- 后端工程：
 - Spring Boot 实现单体架构：https://github.com/fenixsoft/monolithic_arch_springboot
 - Spring Cloud 实现微服务架构：https://github.com/fenixsoft/microservice_arch_springcloud
 - Kubernetes 为基础设施的微服务架构：https://github.com/fenixsoft/microservice_arch_kubernetes
 - Istio 为基础设施的服务网格架构：https://github.com/fenixsoft/servicemesh_arch_istio
 - 基于云端的无服务架构：https://github.com/fenixsoft/serverless_arch

前端工程



Release v1.0 build passing Doc License CC 4.0 License Apache 2.0 Author IcyFenix

如果你此时并不曾了解过什么是“The Fenix Project”，建议先阅读[这部分内容](#)。

Fenix Project的主要目的是展示不同的后端技术架构，相对而言，前端并非其重点。不过，前端的页面是比起后端各种服务来要直观得多，能让使用者更容易理解我们将要做的是一件什么事情。假设你是一名驾驶初学者，合理的学习路径肯定应该是把汽车发动，然后慢慢行驶起来，而不是马上从“引擎动力原理”、“变速箱构造”入手去设法深刻地了解一台汽车。所以，先来运行程序，看看最终的效果是什么样子吧。

运行程序

以下几种途径，可以马上浏览最终的效果：

- 从互联网已部署（由提供Travis-CI支持）的网站（由GitHub Pages提供主机，由腾讯云CDN提供国内加速）访问：

直接在浏览器访问：<http://bookstore.icyfenix.cn/>

- 通过Docker容器方式运行：

```
$ docker run -d -p 80:80 --name bookstore icyfenix/bookstore:frontend
```

sh

然后在浏览器访问：<http://localhost>

- 通过Git上的源码，以开发模式运行：

```
# 克隆获取源码  
$ git clone https://github.com/fenixsoft/fenix-bookstore-frontend.git  
  
# 进入工程根目录  
$ cd fenix-bookstore-frontend  
  
# 安装工程依赖  
$ npm install  
  
# 以开发模式运行，地址为localhost:8080  
$ npm run dev
```

然后在浏览器访问：<http://localhost:8080>



也许你已注意到，以上这些运行方式，均没有涉及到任何的服务端、数据库的部署。现代软件工程里，基于MVVM的工程结构使得前、后端的开发可以完全分离，只要互相约定好服务的位置及模型即可。Fenix's BookStore以开发模式运行时，会自动使用Mock.js拦截住所有的远程服务请求，并以事项准备好的数据来完成对这些请求的响应。

同时，你也应当注意到，以纯前端方式运行的时候，所有对数据的修改请求实际都是无效的。譬如用户注册，无论你输入何种用户名、密码，由于请求的响应是静态预置的，所以最终都会以同一个预设的用户登陆。也是因此，我并没有提供“默认用户”、“默认密码”一类的信息供用户使用，你可以随意输入即可登陆。

不过，那些只维护在前端的状态依然可以变动的，典型的如对购物车、收藏夹的增删改。让后端服务保持无状态，而把状态维持在前端中的设计，对服务的伸缩性和系统的鲁棒性都有着极大的益处，多数情况下都是值得倡导的良好设计。而其伴随而来的状态数据导致请求头变大、链路安全性等问题，都会在服务端部分专门讨论和解决。

构建产品

当你将程序用于正式部署时，一般不应部署开发阶段的程序，而是要进行产品化（producti
on）与精简化（minification），你可以通过以下命令，由node.js驱动webpack来自动完
成：

```
# 编译前端代码  
$ npm run build
```

sh

或者使用--report参数，同时输出依赖分析报告：

```
# 编译前端代码并生成报告  
$ npm run build --report
```

sh

编译结果存放在/dist目录中，应将其拷贝至Web服务器的根目录使用。对于Fenix Project
的各个服务端而言，则通常是拷贝到网关工程中静态资源目录下。

与后端联调

同样出于前后端分离的目的，理论上后端通常只应当依据约定的服务协议（接口定位、访
问传输方式、参数及模型结构、服务水平协议等）提供服务，并以此为依据进行不依赖前
端的独立测试，最终集成时使用的是编译后的前端产品。

不过，在开发期就进行的前后端联合在现今许多企业之中仍是主流形式，由一个人“全栈
式”地开发某个功能时更是如此，因此，当要在开发模式中进行联调时，需要修改项目根目
录下的main.js文件，使其不导入Mock.js，即如下代码所示的条件语句判断为假

```
/**  
 * 默认在开发模式中启用mock.js代替服务端请求  
 * 如需要同时调试服务端，请修改此处判断条件  
 */  
// eslint-disable-next-line no-constant-condition
```

js

```
if (process.env.MOCK) {  
    require('./api/mock')  
}
```

也有其他一些相反的情况，需要在生产包中仍然继续使用Mock.js提供服务时（譬如Docker镜像icyfenix/bookstore:frontend就是如此），同样应修改该条件，使其结果为真，在开发模式依然导入了Mock.js即可。

工程结构

Fenix's BookStore的工程结构完全符合vue.js工程的典型习惯，事实上它在建立时就是通过vue-cli初始化的。此工程的结构与其中各个目录的作用主要如下所示：

----build	webpack编译配置，该目录的内容一般不做改动
----config	webpack编译配置，用户需改动的内容提取至此
----dist	编译输出结果存放的位置
----markdown 片)	与项目无关，用于支持markdown的资源（如图片）
----src	
----api	本地与远程的API接口
----local	本地服务，如localStorage、加密等
----mock	远程API接口的Mock
----json	Mock返回的数据
----remote	远程服务
----assets	资源文件，会被webpack哈希和压缩
----components	vue.js的组件目录，按照使用页面的结构放置
----home	
----cart	
----detail	
----main	
----login	
----pages	vue.js的视图目录，存放页面级组件
----home	
----plugins	vue.js的插件，如全局异常处理器
----router	vue-router路由配置
----store	vuex状态配置
----modules	vuex状态按名空间分隔存放

\---static
 缩

静态资源，编译时原样打包，不会做哈希和压缩

组件

Fenix's BookStore前端部分基于以下开源组件和免费资源构建：

- [Vue.js](#)
渐进式JavaScript框架
- [Element](#)
一套为开发者、设计师和产品经理准备的基于Vue 2.0的桌面端组件库
- [Axios](#)
Promise based HTTP client for the browser and node.js
- [Mock.js](#)
生成随机数据，拦截 Ajax 请求
- [DesignEvo](#)
一款由PearlMountain有限公司设计研发的logo设计软件

协议

- 本文档代码部分采用[Apache 2.0协议](#)进行许可。遵循许可的前提下，你可以自由地对代码进行修改，再发布，可以将代码用作商业用途。但要求你：
 - 署名：在原有代码和衍生代码中，保留原作者署名及代码来源信息。
 - 保留许可证：在原有代码和衍生代码中，保留Apache 2.0协议文件。
- 本作品文档部分采用[知识共享署名 4.0 国际许可协议](#)进行许可。遵循许可的前提下，你可以自由地共享，包括在任何媒介上以任何形式复制、发行本作品，亦可以自由地演绎、修改、转换或以本作品为基础进行二次创作。但要求你：
 - 署名：应在使用本文档的全部或部分内容时候，注明原作者及来源信息。
 - 非商业性使用：不得用于商业出版或其他任何带有商业性质的行为。如需商业使用，请联系作者。

- **相同方式共享的条件**：在本文档基础上演绎、修改的作品，应当继续以知识共享署名4.0国际许可协议进行许可。

单体架构：Spring Boot



Release v1.0 build passing coverage 90% License Apache 2.0 Doc License CC 4.0
Author IcyFenix

如果你此时并不曾了解过什么是“The Fenix Project”，建议先阅读[这部分内容](#)。

单体架构是Fenix's Bookstore'第一个版本的服务端实现，它与此后基于微服务（Spring Cloud、Kubernetes）、无服务（Knative）架构风格实现的其他版本，在功能需求上的表现是完全一致的。如果你不是针对性地带着解决某个具体问题、了解某项具体工具、技术的目的而来，而是时间充裕，希望了解软件架构的全貌与发展的话，笔者推荐以此工程入手来了解现代软件架构，因为单体架构的结构是相对直观的，易于理解的架构，对后面接触的其他架构风格也起良好的铺垫作用。此外，笔者在对应的文档中详细分析了作为一个架构设计者，会考虑哪些的通用问题，希望把抽象的“架构”一词具象化出来。

运行程序

以下几种途径，可以运行程序，浏览最终的效果：

- 通过Docker容器方式运行：

```
$ docker run -d -p 8080:8080 --name bookstore  
icyfenix/bookstore:monolithic
```

sh

然后在浏览器访问：<http://localhost:8080>，系统预置了一个用户（user:icyfenix，pw:123456），也可以注册新用户来测试。

默认会使用HSQLDB的内存模式作为数据库，并在系统启动时自动初始化好了Schema，完全开箱即用。但这同时也意味着当程序运行结束时，所有的数据都将不会被保留。

如果希望使用HSQLDB的文件模式，或者其他非嵌入式的独立的数据库支持的话，也是很简单的。以常用的MySQL/MariaDB为例，程序中也已内置了MySQL的表结构初始化脚本，你可以使用环境变量“PROFILES”来激活Spring Boot中针对MySQL所提供的配置，命令如下所示：

```
$ docker run -d -p 8080:8080 --name bookstore  
icyfenix/bookstore:monolithic -e PROFILES=mysql
```

sh

此时你需要通过Docker link、Docker Compose或者直接在主机的Host文件中提供一个名为“mysql_lan”的DNS映射，使程序能顺利链接到数据库，关于数据库的更多配置，可参考源码中的[application-mysql.yml](#)。

- 通过Git上的源码，以Maven运行：

```
# 克隆获取源码  
$ git clone https://github.com/fenixsoft/monolithic_arch_springboot.git  
  
# 进入工程根目录  
$ cd monolithic_arch_springboot  
  
# 编译打包  
# 采用Maven Wrapper，此方式只需要机器安装有JDK 8或以上版本即可，无需包括Maven在内的其他任何依赖  
# 如在Windows下应使用mvnw.cmd package代替以下命令  
$ ./mvnw package  
  
# 运行程序，地址为localhost:8080  
$ java -jar target/bookstore-1.0.0-Monolithic-SNAPSHOT.jar
```

sh

然后在浏览器访问：<http://localhost:8080>，系统预置了一个用户（user:icyfenix，pw:123456），也可以注册新用户来测试。

- 通过Git上的源码，在IDE环境中运行：

- 以IntelliJ IDEA为例，Git克隆本项目后，在File -> Open菜单选择本项目所在的目录，或者pom.xml文件，以Maven方式导入工程。
- IDEA将自动识别出这是一个SpringBoot工程，并定位启动入口为BookstoreApplication，待IDEA内置的Maven自动下载完所有的依赖包后，运行该类即可启动。
- 如你使用其他的IDE，没有对SpringBoot的直接支持，亦可自行定位到BookstoreApplication，这是一个带有main()方法的Java类，运行即可。
- 可通过IDEA的Maven面板中Lifecycle里面的package来对项目进行打包、发布。
- 在IDE环境中修改配置（如数据库等）会更加简单，具体可以参考工程中application.yml和application-mysql.yml中的内容。

技术组件

Fenix's BookStore单体架构后端尽可能采用标准的技术组件进行构建，不依赖与具体的实现，包括：

- [JSR 370 : Java API for RESTful Web Services 2.1](#) (JAX-RS 2.1)
RESTful服务方面，采用的实现为Jersey 2，亦可替换为Apache CXF、RESTEasy、WebSphere、WebLogic等
- [JSR 330 : Dependency Injection for Java 1.0](#)
依赖注入方面，采用的实现为SpringBoot 2中内置的Spring Framework 5。虽然在多数场合中尽可能地使用了JSR 330的标准注解，但仍有少量地方由于Spring对@Named、@Inject等注解的支持表现上与本身提供的注解差异，使用了Spring的私有注解。如替换成其他的CDI实现，如HK2，需要较大的改动
- [JSR 338 : Java Persistence 2.2](#)
持久化方面，采用的实现为Spring Data JPA。可替换为Batoo JPA、EclipseLink、OpenJPA等实现，只需将使用CrudRepository所省略的代码手动补全回来即可，无需其他改动。
- [JSR 380 : Bean Validation 2.0](#)
数据验证方面，采用的实现为Hibernate Validator 6，可替换为Apache BVal等其他验证

框架

- [JSR 315 : Java Servlet 3.0](#)

Web访问方面，采用的实现为SpringBoot 2中默认的Tomcat 9 Embed，可替换为Jetty、Undertow等其他Web服务器

有以下组件仍然依赖了非标准化的技术实现，包括：

- [JSR 375 : Java EE Security API specification 1.0](#)

认证/授权方面，在2017年才发布的JSR 375中仍然没有直接包含OAuth2和JWT的直接支持，因后续实现微服务架构时对比的需要，单体架构中选择了Spring Security 5作为认证服务，Spring Security OAuth 2.3作为授权服务，Spring Security JWT作为JWT令牌支持，并未采用标准的JSR 375实现，如Soteria。

- [JSR 353/367 : Java API for JSON Processing/Binding](#)

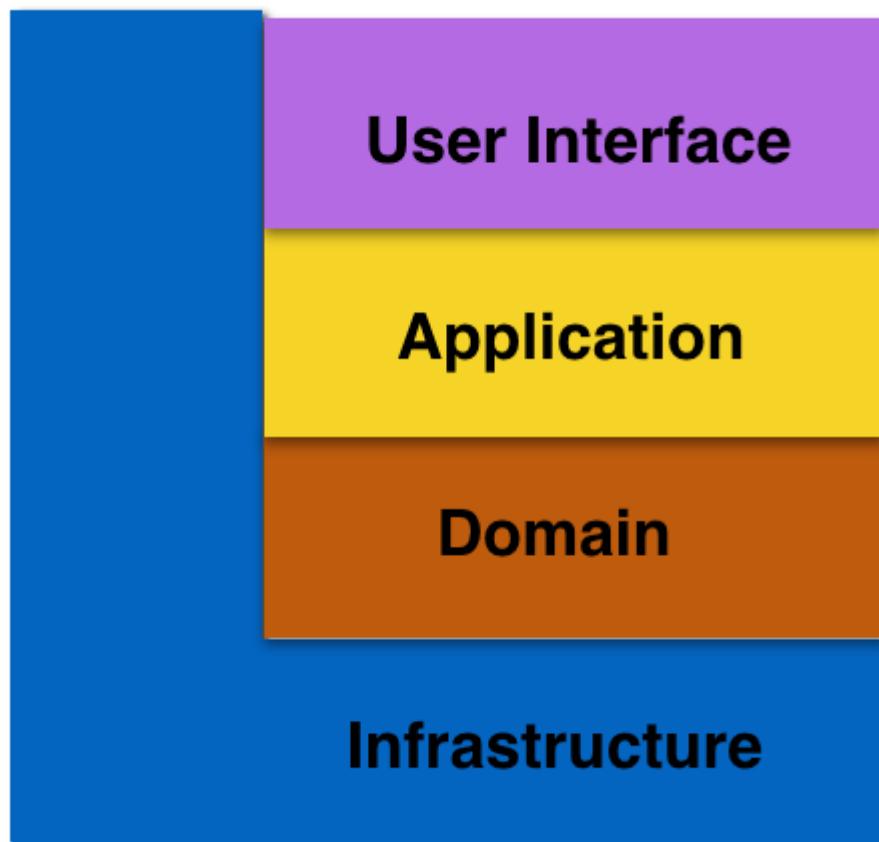
在JSON序列化/反序列化方面，由于Spring Security OAuth的限制（使用JSON-B作为反序列化器时的结果与Jackson等有差异），采用了Spring Security OAuth默认的Jackson，并未采用标准的JSR 353/367实现，如Apache Johnzon、Eclipse Yasson等。

工程结构

Fenix's BookStore单体架构后端参考（并未完全遵循）了DDD的分层模式和设计原则，整体分为以下四层：

1. Resource：对应DDD中的User Interface层，负责向用户显示信息或者解释用户发出的命令。请注意，这里指的“用户”不一定是使用用户界面的人，可以是位于另一个进程或计算机的服务。由于本工程采用了MVVM前后端分离模式，这里所指的用户实际上是前端的服务消费者，所以这里以RESTFul中的核心概念“资源”（Resource）来命名。
2. Application：对应DDD中的Application层，负责定义软件本身对外暴露的能力，即软件本身可以完成哪些任务，并负责对内协调领域对象来解决问题。根据DDD的原则，应用层要尽量简单，不包含任何业务规则或者知识，而只为下一层中的领域对象协调任务，分配工作，使它们互相协作，这一点在代码上表现为Application层中一般不会存在任何的条件判断语句。在许多项目中，Application层都会被选为包裹事务（代码进入此层事务开始，退出此层事务提交或者回滚）的载体。

3. Domain：对应DDD中的Domain层，负责实现业务逻辑，即表达业务概念，处理业务状态信息以及业务规则这些行为，此层是整个项目的特点。
4. Infrastructure：对应DDD中的Infrastructure层，向其他层提供通用的技术能力，譬如持久化能力、远程服务通讯、工具集，等等。



协议

- 本文档代码部分采用[Apache 2.0 协议](#)进行许可。遵循许可的前提下，你可以自由地对代码进行修改，再发布，可以将代码用作商业用途。但要求你：
 - 署名：在原有代码和衍生代码中，保留原作者署名及代码来源信息。
 - 保留许可证：在原有代码和衍生代码中，保留Apache 2.0协议文件。
- 本作品文档部分采用[知识共享署名 4.0 国际许可协议](#)进行许可。遵循许可的前提下，你可以自由地共享，包括在任何媒介上以任何形式复制、发行本作品，亦可以自由地演绎、修改、转换或以本作品为基础进行二次创作。但要求你：
 - 署名：应在使用本文档的全部或部分内容时候，注明原作者及来源信息。

- **非商业性使用**：不得用于商业出版或其他任何带有商业性质的行为。如需商业使用，请联系作者。
- **相同方式共享的条件**：在本文档基础上演绎、修改的作品，应当继续以知识共享署名4.0国际许可协议进行许可。

微服务 : Spring Cloud



Release v1.0 build passing coverage 81% License Apache 2.0 Doc License CC 4.0

Author IcyFenix

如果你此时并不曾了解过什么是“The Fenix Project”，建议先阅读[这部分内容](#)。

至少到目前，基于Spring Cloud的微服务解决方案仍是以Java为运行平台的微服务中，使用者数量最多的一个分支。这个结果即是Java在服务端应用中长久积累的深厚基础的体现，也是Spring在Java应用中统治性的地位的体现。Spring Cloud令现存数量极为庞大的、基于Spring和Spring Boot的单体系统，得以平滑地迁移到微服务架构中，令这些系统的大部分代码都能够无需或少量修改即可保留重用。微服务兴起的早期，Spring Cloud就集成了[Netflix OSS](#)（以及Spring Cloud Netflix进入维护期后对应的替代组件）所开发的体系化的微服务套件，基本上算“半透明地”解决了在微服务环境中必然会面临的服务发现、远程调用、负载均衡、集中配置等基础问题。

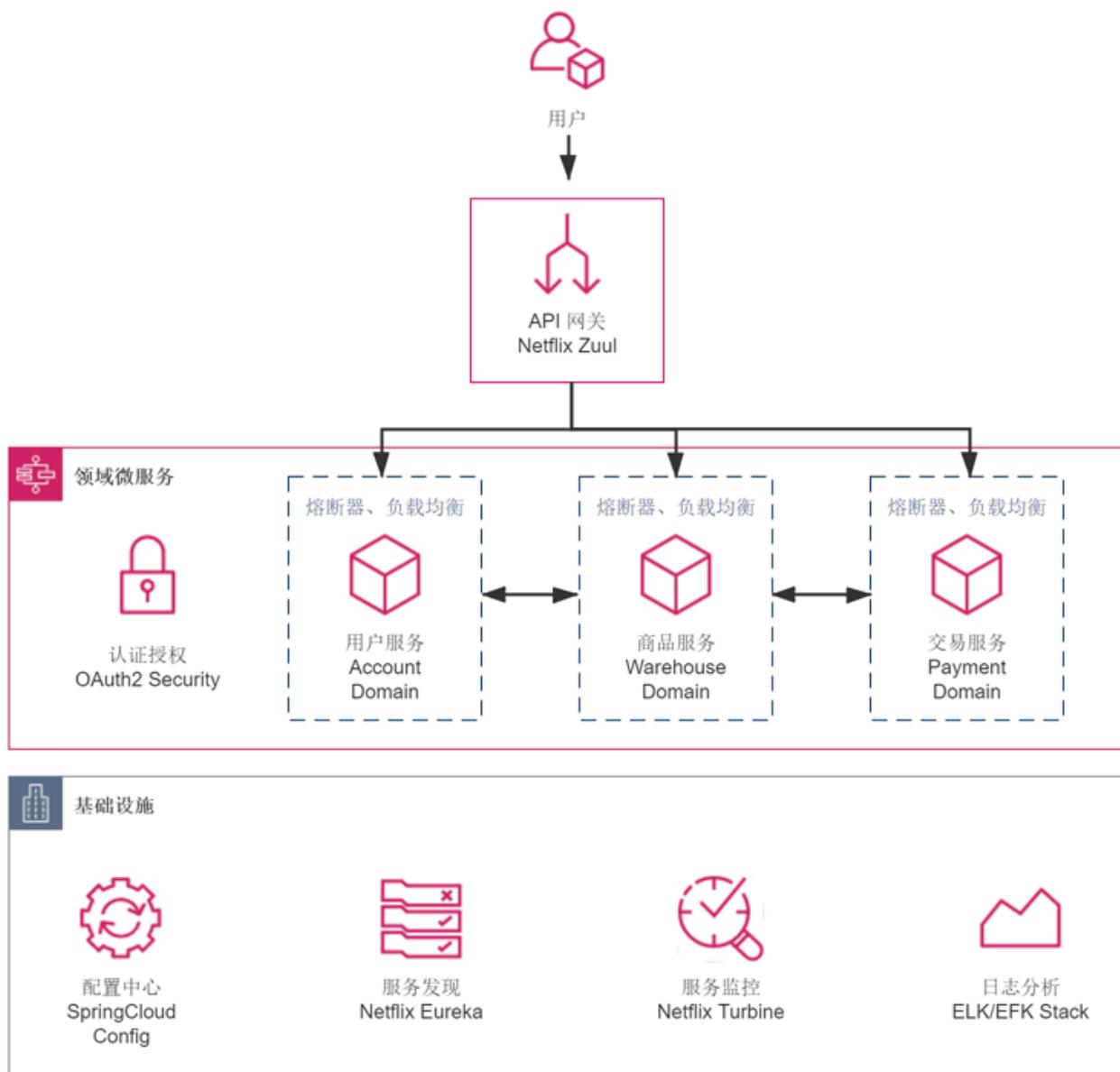
不过，笔者自己并不太认同Spring Cloud Netflix这种以应用代码去解决基础设施功能问题的“解题思路”，以笔者的观点看来，这既是容器化、原生化的微服务基础设施完全成熟之前必然会出现的应用形态，同时也决定了这是微服务进化过程中必然会被替代的过渡形态。无论笔者的看法如何，基于Spring Cloud Netflix的微服务在当前是主流，直至未来不算短的一段时间内仍会是主流，并且以应用的视角，自顶向下观察基础设施在微服务中面临的需求和挑战，用我们熟悉的Java代码来解释分析问题，也有利于对微服务的整体思想的深入理解，所以将它作为我们了解的第一种微服务架构的实现是十分适合的。

需求场景

小书店Fenix's Bookstore生意日益兴隆，客人、货物、营收都在持续增长，业务越发复杂，对信息系统并发与可用方面的要求也越来越高。由于业务属性和质量属性要求的提升，信息系统需要更多的硬件资源去支撑，这是合乎情理的，但是，如果我们把需求场景列的更具体些，便会发现“合理”下面的许多无可奈何之处：

- 譬如，恰逢双十一购物节，短时间内会有大批的交易事件发生，这时候运维的同学对系统进行扩容以应对流量洪峰。但此时增长的业务量并不是均衡的，只有商品交易的活动剧增，其他的活动，如新商品入库、新用户注册这类并未增加多少，此时，面对“铁板一块”的单体系统，运维在做扩容时，只能把“用不上的”商品管理代码、用户管理代码也一并扩容部署。
- 譬如，高性能硬件对性能的提升是有帮助，但对稳定性的提升通常无能为力。业务复杂度的增加促使系统的技术复杂度也在持续增长，当系统不可避免地滑向庞大臃肿时，总伴随有各种难以预料的问题出现；要维持一个庞然大物的健康生存，也对设计、开发、运维各方面的人员都提出越来越高的要求。人力终有穷时，迟早会面临“没有一个人能了解系统的所有细节”的情形；系统的复杂程度也总有极限，持续膨胀的代码终会有崩溃的一刻。
- 譬如，……

微服务的需求场景还可以列举很多，这里就不多列举了，总之，系统发展到一定程度，我们总能找到充分的理由去重构拆分它。在笔者设定的场景中，准备把单体的Fenix's Bookstore拆分为“用户”、“商品”、“交易”三个能够独立运行的子系统，它们将在一系列非功能性模块（认证授权等）和基础设施（配置中心、服务发现等）的支撑下互相协作，以统一的API网关对外提供与原来单体系统在功能上一致的服务，应用视图如下图所示：



运行程序

以下几种途径，可以运行程序，浏览最终的效果：

- 通过Docker容器方式运行：

微服务涉及到多个容器的协作，通过link单独运行容器已经被Docker官方声明为不提倡的方式，所以在工程中提供了专门的配置，以便使用docker-compose来运行：

```
# 下载docker-compose配置文件
$ curl -O
https://raw.githubusercontent.com/fenixsoft/microservice_arch_springcloud/master/docker-compose.yml
```

```
# 启动服务  
$ docker-compose up
```

然后在浏览器访问：<http://localhost:8080>，系统预置了一个用户（user:icyfenix，pw:123456），也可以注册新用户来测试。

- 通过Git上的源码，以Maven编译、运行：

由于笔者已经在配置文件中设置好了各个微服务的默认的地址和端口号，以便于本地调试。如果要在同一台机运行这些服务，并且每个微服务都只启动一个实例的话，那不加任何配置、参数即可正常以Maven编译、以Jar包形式运行。由于各个微服务需要从配置中心里获取具体的参数信息，因此唯一的要求只是“配置中心”的微服务必须作为第一个启动的服务进程，对其他的启动顺序则没有更多要求了。具体的操作过程如下所示：

```
# 克隆获取源码  
$ git clone  
https://github.com/fenixsoft/microservice_arch_springcloud.git  
  
# 进入工程根目录  
$ cd microservice_arch_springcloud  
  
# 编译打包  
# 采用Maven Wrapper，此方式只需要机器安装有JDK 8或以上版本即可，无需包括  
Maven在内的其他任何依赖  
# 克隆后你可能需要使用chmod给mvnw赋予执行权限，如在Windows下应使用mvnw.cmd  
package代替以下命令  
$ ./mvnw package  
  
# 工程将编译出七个SpringBoot Jar  
# 启动服务需要运行以下七个微服务组件  
# 配置中心微服务：localhost:8888  
$ java -jar ./bookstore-microservices-platform-  
configuration/target/bookstore-microservice-platform-configuration-  
1.0.0-SNAPSHOT.jar  
# 服务发现微服务：localhost:8761  
$ java -jar ./bookstore-microservices-platform-  
registry/target/bookstore-microservices-platform-registry-1.0.0-  
SNAPSHOT.jar  
# 服务网关微服务：localhost:8080  
$ java -jar ./bookstore-microservices-platform-  
gateway/target/bookstore-microservices-platform-gateway-1.0.0-  
SNAPSHOT.jar
```

```
# 安全认证微服务：localhost:8301
$ java -jar ./bookstore-microservices-domain-
security/target/bookstore-microservices-domain-security-1.0.0-
SNAPSHOT.jar
# 用户信息微服务：localhost:8401
$ java -jar ./bookstore-microservices-domain-
account/target/bookstore-microservices-domain-account-1.0.0-
SNAPSHOT.jar
# 商品仓库微服务：localhost:8501
$ java -jar ./bookstore-microservices-domain-
warehouse/target/bookstore-microservices-domain-warehouse-1.0.0-
SNAPSHOT.jar
# 商品交易微服务：localhost:8601
$ java -jar ./bookstore-microservices-domain-
payment/target/bookstore-microservices-domain-payment-1.0.0-
SNAPSHOT.jar
```

由于在命令行启动多个服务、通过容器实现各服务隔离、扩展等都较繁琐，笔者提供了一个docker-compose.dev.yml文件，便于开发期调试使用：

```
# 使用Maven编译出JAR包后，可使用以下命令直接在本地构建镜像运行
```

sh

```
$ docker-compose -f docker-compose.dev.yml up
```

以上两种本地运行的方式可任选其一，服务全部启动后，在浏览器访问：

[<http://localhost:8080>] (<http://localhost:8080>)，系统预置了一个用户
(user:icyfenix, pw:123456)，也可以注册新用户来测试

- 通过Git上的源码，在IDE环境中运行：

- 以IntelliJ IDEA为例，Git克隆本项目后，在File -> Open菜单选择本项目所在的目录，或者pom.xml文件，以Maven方式导入工程。
- 待Maven自动安装依赖后，即可在IDE或者Maven面板中编译全部子模块的程序。
- 本工程下面八个模块，其中除bookstore-microservices-library-infrastructure外，其余均是SpringBoot工程，将这七个工程的Application类加入到IDEA的Run Dashboard面板中。
- 在Run Dashboard中先启动“bookstore-microservices-platform-configuration”微服务，然后可一次性启动其余六个子模块的微服务。
- 配置与横向扩展
工程中预留了一些的环境变量，便于配置和扩展，譬如，对于热点模

块，往往需要启动多个微服务扩容，此时需要调整每个服务的端口号。预留的这类环境变量包括：

```
```bash
修改配置中心的主机和端口，默认为localhost:8888
CONFIG_HOST
CONFIG_PORT

修改服务发现的主机和端口，默认为localhost:8761
REGISTRY_HOST
REGISTRY_PORT

修改认证中心的主机和端口，默认为localhost:8301
AUTH_HOST
AUTH_PORT

修改当前微服务的端口号
譬如，你打算在一台机器上扩容四个支付微服务以应对促销活动的流量高峰
可将它们的端口设置为8601（默认）、8602、8603、8604等
真实环境中，它们可能是在不同的物理机、容器环境下，这时扩容可无需调整端口
PORT

SpringBoot所采用Profile配置文件，默认为default
譬如，服务默认使用HSQLDB的内存模式作为数据库，如需调整为MySQL，可将此环境变量调整为mysql
因为笔者默认预置了名为application-mysql.yml的配置，以及HSQLDB和MySQL的数据库脚本
如果你需要支持其他数据库、修改程序中其他的配置信息，可以在代码中自行加入另外的初始化脚本
PROFILES

Java虚拟机运行参数，默认为空
JAVA_OPTS
```

## 技术组件

Fenix's BookStore采用基于Spring Cloud微服务架构，微服务部分主要采用了Netflix OSS组件进行支持，它们包括：

- **配置中心**：默认采用[Spring Cloud Config](#)，亦可使用[Spring Cloud Consul](#)、[Spring Cloud Alibaba Nacos](#)代替。

- **服务发现**：默认采用[Netflix Eureka](#)，亦可使用[Spring Cloud Consul](#)、[Spring Cloud Zookeeper](#)、[etcd](#)等代替。
- **服务网关**：默认采用[Netflix Zuul](#)，亦可使用[Spring Cloud Gateway](#)代替。
- **服务熔断**：默认采用[Netflix Hystrix](#)，亦可使用[Sentinel](#)、[Resilience4j](#)代替。
- **进程内负载均衡**：默认采用[Netflix Ribbon](#)，亦可使用[Spring Cloud Loadbalancer](#)代替。
- **声明式HTTP客户端**：默认采用[Spring Cloud OpenFeign](#)。这个并没有代替的必要，访问远程服务而已，使用[RestTemplate](#)或者更底层的[OkHTTP](#)、[HttpClient](#)也能完成，多写点代码罢了。

尽管Netflix套件的使用人数很多，但由于Spring Cloud Netflix已进入维护模式，所以笔者均列出了上述组件的代替品。这些组件几乎都是声明式的，这保证了替代它们的成本相当低，只需要更换注解，修改配置，无需改动代码。你在阅读源码时也会发现，三个“platform”开头的服务，基本上没有任何实际代码的存在。

其他与微服务无关的技术组件（REST服务、安全、数据访问，等等），笔者已在[Fenix's BookStore单体架构](#)中介绍过，在此不再重复。

## 协议

- 本文档代码部分采用[Apache 2.0协议](#)进行许可。遵循许可的前提下，你可以自由地对代码进行修改，再发布，可以将代码用作商业用途。但要求你：
  - **署名**：在原有代码和衍生代码中，保留原作者署名及代码来源信息。
  - **保留许可证**：在原有代码和衍生代码中，保留Apache 2.0协议文件。
- 本作品文档部分采用[知识共享署名 4.0 国际许可协议](#)进行许可。遵循许可的前提下，你可以自由地共享，包括在任何媒介上以任何形式复制、发行本作品，亦可以自由地演绎、修改、转换或以本作品为基础进行二次创作。但要求你：
  - **署名**：应在使用本文档的全部或部分内容时候，注明原作者及来源信息。
  - **非商业性使用**：不得用于商业出版或其他任何带有商业性质的行为。如需商业使用，请联系作者。
  - **相同方式共享的条件**：在本文档基础上演绎、修改的作品，应当继续以知识共享署名4.0国际许可协议进行许可。



# 微服务：Kubernetes



Release v1.0 build passing License Apache 2.0 Doc License CC 4.0 Author IcyFenix

如果你此时并不曾了解过什么是“The Fenix Project”，建议先阅读[这部分内容](#)。

2017年，笔者曾在文章中描述其为“[后微服务时代](#)”的开端，这年是容器生态发展历史中具有里程碑意义的一年。在这一年，长期作为Docker竞争对手的[RKT容器](#)一派的领导者CoreOS宣布放弃自己的容器管理系统Fleet，未来将会把所有容器管理的功能移至Kubernetes之上实现。在这一年，容器管理领域的独角兽Rancher Labs宣布放弃其内置了数年的容器管理系统Cattle，提出了“All-in-Kubernetes”战略，从2.0版本开始把1.x版本能够支持多种容器管理工具的Rancher，“升级”为只支持Kubernetes一种容器管理系统。在这一年，Kubernetes的主要竞争者Apache Mesos在9月正式宣布了“[Kubernetes on Mesos](#)”集成计划，由竞争关系转为对Kubernetes提供支持，使其能够与Mesos的其他一级框架（如[HDFS](#)、[Spark](#) 和[Chronos](#)，等等）进行集群资源动态共享、分配与隔离。在这一年，Kubernetes的最大竞争者Docker Swarm的母公司Docker，终于在10月被迫宣布Docker要同时支持Swarm与Kubernetes两套容器管理系统，事实上承认了Kubernetes的统治地位。这场已经持续了三、四年时间，以Docker Swarm、Apache Mesos与Kubernetes为主要竞争者的“容器战争”终于有了明确的结果，Kubernetes登基加冕是容器发展中一个时代的终章，也将是软件架构发展下一个纪元的开端。

## 需求场景

当引入了微服务架构后，小书店Fenix's Bookstore解决了扩容缩容、独立部署、运维和管理等问题，满足了产品经理不断提出的日益复杂的业务需求。可是，对于团队的开发人员、设计人员、架构人员来说，并没有感觉到工作变得轻松，微服务中的各种新技术名词，如配置中心、服务发现、网关、熔断、负载均衡等等，就够一名新手学习好长一段时间；从产品角度来看，各种Spring Cloud的技术套件，如Config、Eureka、Zuul、Hystrix、Ribbon、Feign等，也占据了产品的大部分编译后的代码容量。之所以微服务架构里，我们选择在应用层面而不是基础设施层面去解决这些分布式问题，完全是因为由硬件构成的基础设施，跟不上由软件构成的应用服务的灵活性的无奈之举。当Kubernetes统一了容器编排管理系统之后，这些纯技术性的底层问题，便开始有了被广泛认可和采纳的基础设施层面的解决方案。为此，Fenix's Bookstore也迎来了它在“后微服务时代”中的下一次架构演进，这次升级的目标主要有如下两点：

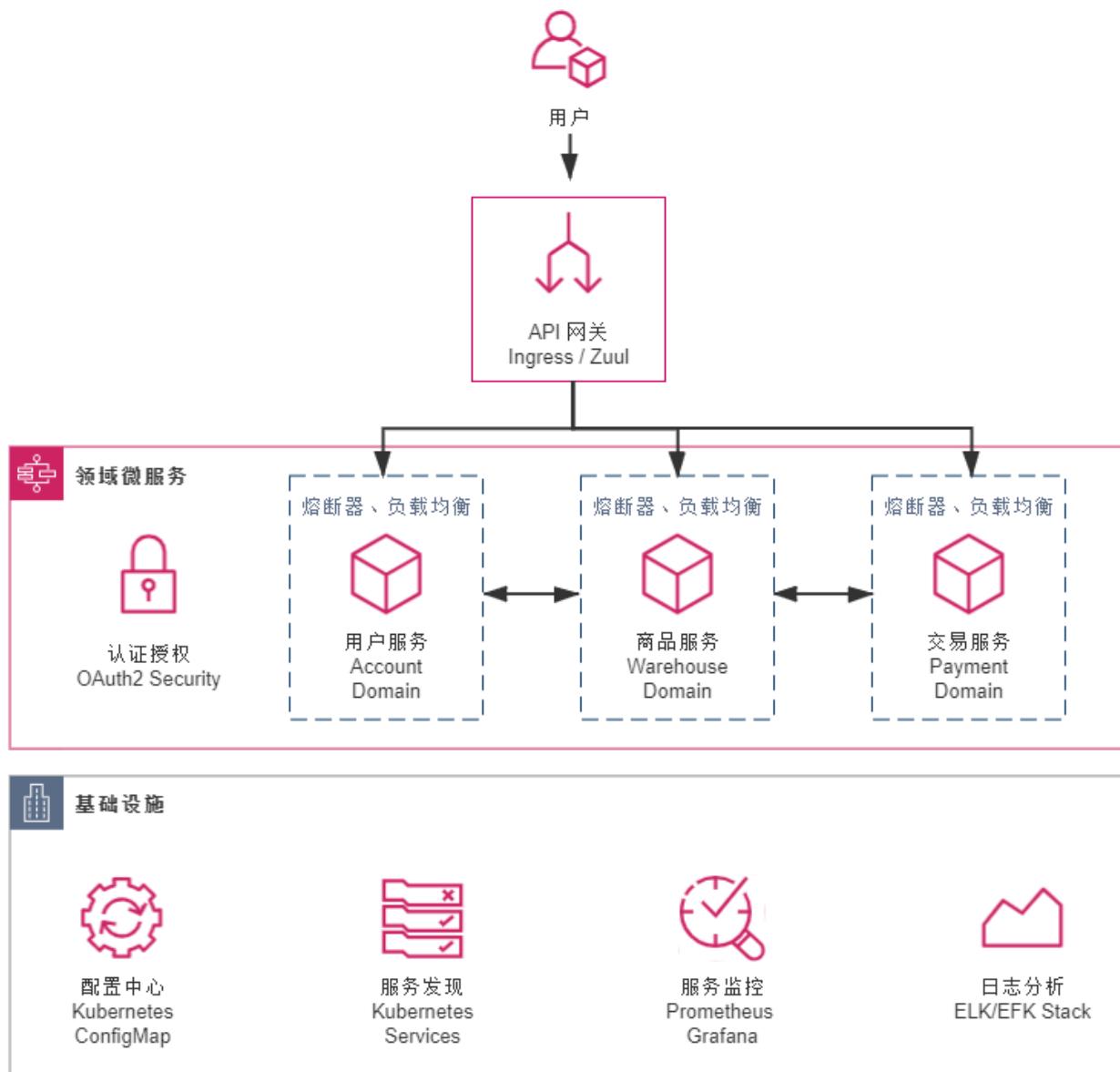
- **目标一**：尽可能缩减非业务功能代码的比例。

在Fenix's Bookstore中，用户服务（Account）、商品服务（Warehouse）、交易服务（Payment）三个工程是真正承载业务逻辑的，认证授权服务（Security）可以认为是同时涉及到了技术与业务，而配置中心（Configuration）、网关（Gateway）和服务注册中心（Registry）则是纯技术性。我们希望尽量消除这些纯技术的工程，以及那些依附在其他业务工程上的纯技术组件。

- **目标二**：尽可能在不影响原有的代码的前提下完成迁移。

得益于Spring Framework 4中的Conditional Bean等声明式特性的出现，近年来新发布的技术组件，**声明式编程**（Declarative Programming）已经逐步取代**命令式编程**（Imperative Programming）成为主流。这使得我们可以从目的而不是过程的角度去描述编码意图，使得代码几乎不会与具体技术实现产生耦合，若要更换一种技术实现，只需要调整配置中的声明便可做到。

从升级结果来看，如果仅以Java代码的角度来衡量，本工程与此前基于Spring Cloud的实现没有丝毫差异，两者的每一行Java代码都是一模一样的；其区别是Kubernetes的实现版本中直接删除了配置中心、服务注册中心的工程，在其他工程的pom.xml中也删除了如Eureka、Ribbon、Config等组件的依赖。取而代之的是新增了若干以YAML配置文件为载体的**Skaffold**和Kubernetes的资源描述，这些资源描述文件，将会动态构建出DNS服务器、服务负载均衡器等一系列虚拟化的基础设施，去代替原有的应用层面的技术组件。升级改造之后的应用架构如下图所示：



## 运行程序

在已经部署Kubernetes集群的前提下，通过以下几种途径，可以运行程序，浏览最终的效果：

- 直接在Kubernetes集群环境下运行：

工程在编译时已通过Kustomize产生出集成式的资源描述文件，可通过该文件直接在Kubernetes集群中运行程序：

```
资源描述文件
$ kubectl create -f
https://raw.githubusercontent.com/fenixsoft/microservice_arch_kubernetes/master/bookstore.yaml
```

当所有的Pod都处于正常工作状态后（这个过程一共需要下载几百MB的镜像，尤其是Docker中没有各层基础镜像缓存时，请根据自己的网速保持一定的耐心。未来GraalVM对Spring Cloud的支持更成熟一些后，可以考虑采用GraalVM来改善这一点），在浏览器访问：<http://localhost:30080>，系统预置了一个用户（user:icyfenix，pw:123456），也可以注册新用户来测试。

- 通过Skaffold在命令行或IDE中以调试方式运行：

一般开发基于Kubernetes的微服务应用，是在本地针对单个服务编码、调试完成后，通过CI/CD流水线部署到Kubernetes中进行集成的。如果只是针对集成测试，这并没有什么问题，但同样的做法应用在开发阶段就不十分不便了，我们不希望每做一处修改都要经过一次CI/CD流程，这将非常耗时且难以调试。

Skaffold是Google在2018年开源的一款加速应用在本地或远程Kubernetes集群中构建、推送、部署和调试的自动化命令行工具，对于Java应用来说，它可以帮助我们做到监视代码变动，自动打包出镜像，将镜像打上动态标签并更新部署到Kubernetes集群，为Java程序注入开放JDWP调试的参数，并根据Kubernetes的服务端口自动在本地生成端口转发。以上都是根据 `skaffold.yaml` 中的配置来进行的，开发时skaffold通过 `dev` 指令来执行这些配置，具体的操作过程如下所示：

```
克隆获取源码
$ git clone
https://github.com/fenixsoft/microservice_arch_kubernetes.git && cd
microservice_arch_kubernetes

编译打包
$./mvnw package

启动Skaffold
此时将会自动打包Docker镜像，并部署到Kubernetes中
$ skaffold dev
```

服务全部启动后，在浏览器访问：<http://localhost:30080>，系统预置了一个用户（user:icyfenix，pw:123456），也可以注册新用户来测试

由于面向的是开发环境，基于效率原因，笔者并没有像传统CI工程那样直接使用Maven的Docker镜像来打包Java源码，这决定了构建Dockerfile时，要监视的变动目标将是Jar

文件而不是Java源码，即Skaffold监视的是Jar包的变动，只当进行Maven编译、输出了新的Jar包后才会更新镜像。这样做一方面是考虑到在Maven镜像中打包不便于利用本地的仓库缓存，尤其在国内网络中，速度实在难以忍受；另外一方面，是笔者其实并不希望每保存一次源码时，都自动构建和更新一次镜像，毕竟比起传统的HotSwap或者Spring Devtool Reload来说，更新镜像重启Pod是一个更加重负载的操作。未来CNCF的Buildpack<sup>12</sup>成熟之后，应该可以绕过笨重的Dockerfile，对打包和容器热更新做更加精细化的控制。

另外，对于有IDE调试需求的同学，推荐采用Google Cloud Code<sup>13</sup>（Cloud Code同时提供了VS Code和IntelliJ Idea的插件）来配合Skaffold使用，毕竟是一个公司出品的产品，搭配起来能获得几乎与本地开发单体应用一致的体验。

## 技术组件

Fenix's BookStore采用基于Kubernetes的微服务架构，并采用Spring Cloud Kubernetes做了适配，其中主要的技术组件包括：

- **容器环境感知**：Spring Cloud Kubernetes本身引入了Fabric8的Kubernetes Client<sup>14</sup>作为容器环境感知，不过引用的版本相当陈旧，如Spring Cloud Kubernetes 1.1.2中采用的是Fabric8 Kubernetes Client 4.4.1，Fabric8提供的兼容性列表中该版本只支持到Kubernetes 1.14，实测在1.16上也能用，但是在1.18上无法识别到最新的Api-Server，因此Maven引入依赖时需要手工处理，排除旧版本，引入新版本（本工程采用的是4.10.1）。
- **配置中心**：采用Kubernetes的ConfigMap来管理，通过Spring Cloud Kubernetes Config<sup>15</sup>自动将ConfigMap的内容注入到Spring配置文件中，并实现动态更新。
- **服务发现**：采用Kubernetes的Service来管理，通过Spring Cloud Kubernetes Discovery<sup>16</sup>自动将HTTP访问中的服务转换为FQDN<sup>17</sup>。
- **负载均衡**：采用Kubernetes Service本身的负载均衡能力实现（就是DNS负载均衡），可以不再需要Ribbon这样的客户端负载均衡了。Spring Cloud Kubernetes从1.1.2开始也已经移除了对Ribbon的适配支持，也（暂时）没有对其代替品Spring Cloud LoadBalancer提供适配。
- **服务网关**：网关部分仍然保留了Zuul，未采用Ingress代替。这里有两点考虑，一是Ingress Controller不算是Kubernetes的自带组件，它可以有不同的选择（KONG、Nginx、Haproxy，等等），同时也需要独立安装，作为演示工程，出于环境复杂度最小化考虑

未使用Ingress；二是Fenix's Bookstore的前端工程是存放在网关中的，移除了Zuul之后也仍然要维持一个前端工程的存在，不能进一步缩减工程数量，也就削弱了移除Zuul的动力。

- **服务熔断**：仍然采用Hystrix，Kubernetes本身无法做到精细化的服务治理，包括熔断、流控、监视，等等，我们将在基于Istio的服务网格架构中解决这个问题。
- **认证授权**：仍然采用Spring Security OAuth 2，Kubernetes的RBAC授权可以解决服务间的安全访问问题，但Security是跨越了业务和技术的边界的，认证授权模块本身仍承担着对前端用户的认证、授权职责，这部分是与业务相关的。

## 协议

- 本文档代码部分采用[Apache 2.0协议](#)进行许可。遵循许可的前提下，你可以自由地对代码进行修改，再发布，可以将代码用作商业用途。但要求你：
  - **署名**：在原有代码和衍生代码中，保留原作者署名及代码来源信息。
  - **保留许可证**：在原有代码和衍生代码中，保留Apache 2.0协议文件。
- 本作品文档部分采用[知识共享署名 4.0 国际许可协议](#)进行许可。遵循许可的前提下，你可以自由地共享，包括在任何媒介上以任何形式复制、发行本作品，亦可以自由地演绎、修改、转换或以本作品为基础进行二次创作。但要求你：
  - **署名**：应在使用本文档的全部或部分内容时候，注明原作者及来源信息。
  - **非商业性使用**：不得用于商业出版或其他任何带有商业性质的行为。如需商业使用，请联系作者。
  - **相同方式共享的条件**：在本文档基础上演绎、修改的作品，应当继续以知识共享署名 4.0国际许可协议进行许可。

# 服务网格：Istio

# 无服务 : Serverless

# 服务架构演进史

服务架构的演进历史这一章，我们借讨论历史之名，来梳理微服务发展里程中出现的大量名词、概念，借着微服务的演变过程，我们将从这些概念起源的最初，去分析它们是什么、它们取代了什么、以及它们为什么能够在斗争中取得成功，为什么变得不可或缺的支撑，又或者它们为什么会失败，在竞争中被淘汰，或逐渐湮灭于历史的烟尘当中。

- **原始分布式时代**：使用多个独立的分布式服务共同构建一个更大型系统，尽可能促使服务交互透明与简单，令开发人员不必过份关注他们访问的方法或其他资源是位于本地还是远程。
- **单体系统时代**：“单体”只是表明系统中主要的过程调用都是进程内调用，不会发生进程间通讯，仅此而已。
- **SOA时代**：面向服务的架构是第一次系统性地成功解决分布式服务主要问题的架构模式。
- **微服务时代**：微服务是一种通过多个小型服务组合来构建单个应用的架构风格，这些服务围绕业务能力而非特定的技术标准来构建。各个服务可以采用不同的编程语言，不同的数据存储技术，运行在不同的进程之中。服务采取轻量级的通讯机制和自动化的部署机制实现通讯与运维。
- **后微服务时代**：从软件层面独力应对微服务架构问题，发展到软硬一体，合力应对架构问题的时代，此即为“后微服务时代”。
- **无服务时代**：如果说微服务架构是分布式系统这条路的极致，那无服务架构，也许就是“不分布式”的云端系统这条路的起点。

# 原始分布式时代

## Unix时代的分布式设计哲学

使用多个独立的分布式服务共同构建一个更大型系统，尽可能促使服务交互透明与简单，令开发人员不必过份关注他们访问的方法或其他资源是位于本地还是远程。

可能与绝大多数人心中的认知会有差异，“使用多个独立的分布式服务共同构建一个更大型系统”的设想与实际尝试，反而要比今天大家所了解的大型单体系统出现的时间更早。

在20世纪的70年代末期到80年代初，计算机科学刚经历了从以大型机为主向以微型机为主的蜕变，计算机逐渐从一种存在于研究机构、实验室当中的科研设备，转变为存在于商业企业中甚至家庭用户的生产设备。此时的计算机系统通常具有16位、不足5MHz时钟频率的处理器和128KB左右的内存空间，譬如，著名英特尔处理器的鼻祖，[Intel 8086处理器](#)就是在1978年发布，流行于80年代中期，甚至一直持续到90年代初期。当时计算机的硬件水平的局限性，已直接妨碍到了单台计算机上信息系统软件能够达到的最大规模，为了突破硬件算力限制，[Unix系统标准化组织开放软件基金会](#)（Open Software Foundation，OSF，也即后来的“国际开放标准组织”）制订了一种名为“[分布式运算环境](#)”（Distributed Computing Environment，DCE）的软件架构，其中包括了一整套完整的分布式服务组件与规范，DCE提出的很多技术、概念对\*nix系统后续的发展，甚至是今天计算机科学的诸多领域都产生了巨大而深远的影响，譬如远程服务调用（Remote Procedure Call，当时被称为[DCE/RPC](#)，后来Sun向IEFT提交了不局限于Unix系统的、基于TCP/IP、更通用的远程服务调用标准[ONC RPC](#)），分布式文件系统（Distributed File System，当时被称为[DCE/DFS](#)）、时间服务、授权服务，等等。

当时研究这些分布式技术，最主要的目标是实现分布式环境中的服务调用、资源访问、数据存储等操作尽可能的透明化，使开发人员不必过于关注他们访问的方法或其他资源是位于本地还是远程，这样的主旨非常符合一贯的[Unix设计哲学](#)（曾有过几个版本的不同说法，不过广为人知的“KISS”原则是最基础、无争议的一条），但这个过于理想化的目标背后其实蕴含着当时根本不可能完美解决的技术困难，研究最终结果是实现了远程服务的

调用，但远远没有能做到“透明”，本地与远程无论是编码、运行还是效率角度上看，都有着天壤之别。

这次研究是计算机科学中第一次有组织领导、有标准可循、有巨大投入的分布式计算的尝试，但无论是DCE还是稍后出现的CORBA，都不能算取得了成功，将一个系统直接拆分到不同的机器之中，这样做带来的服务的发现、跟踪、通讯、容错、隔离、配置、传输、数据一致性和编码复杂度等方面的问题，所付出的代价远远超过了分布式所取得的收益，这次尝试最大的收获就是对RPC、DFS等概念的开创，以及得到了一个价值千金的教训：“**某个功能能够进行分布式，并不意味着它就应该进行分布式，强行追求透明的分布式操作，只会自寻苦果**”。

### Observation about distributed computing

Just because something **can** be distributed doesn't mean it **should** be distributed. Trying to make a distributed call act like a local call always ends in tears

—— [Kyle Brown](#) , IBM Fellow , [Beyond buzzwords: A brief history of microservices pattern](#)  
[s](#) , 2016

上世纪80年代正是[摩尔定律](#)开始稳定发挥作用的黄金时期，微型计算机的性能以每两年增长一倍的速度提升。硬件算力束缚软件规模的链条很快变得松动，信息系统开始了以单台或少数几台微机即可作为服务器的单体系统时代，尽管如此，对分布式计算、远程服务调用的研究从未有过中断。关于远程服务调用这条分支早期的发展与现状，笔者已在服务设计风格中“[远程服务调用](#)”一节有过介绍。而在原始分布式时代中遭遇到的其他问题，还将会在后面几个时代中被反复提起。

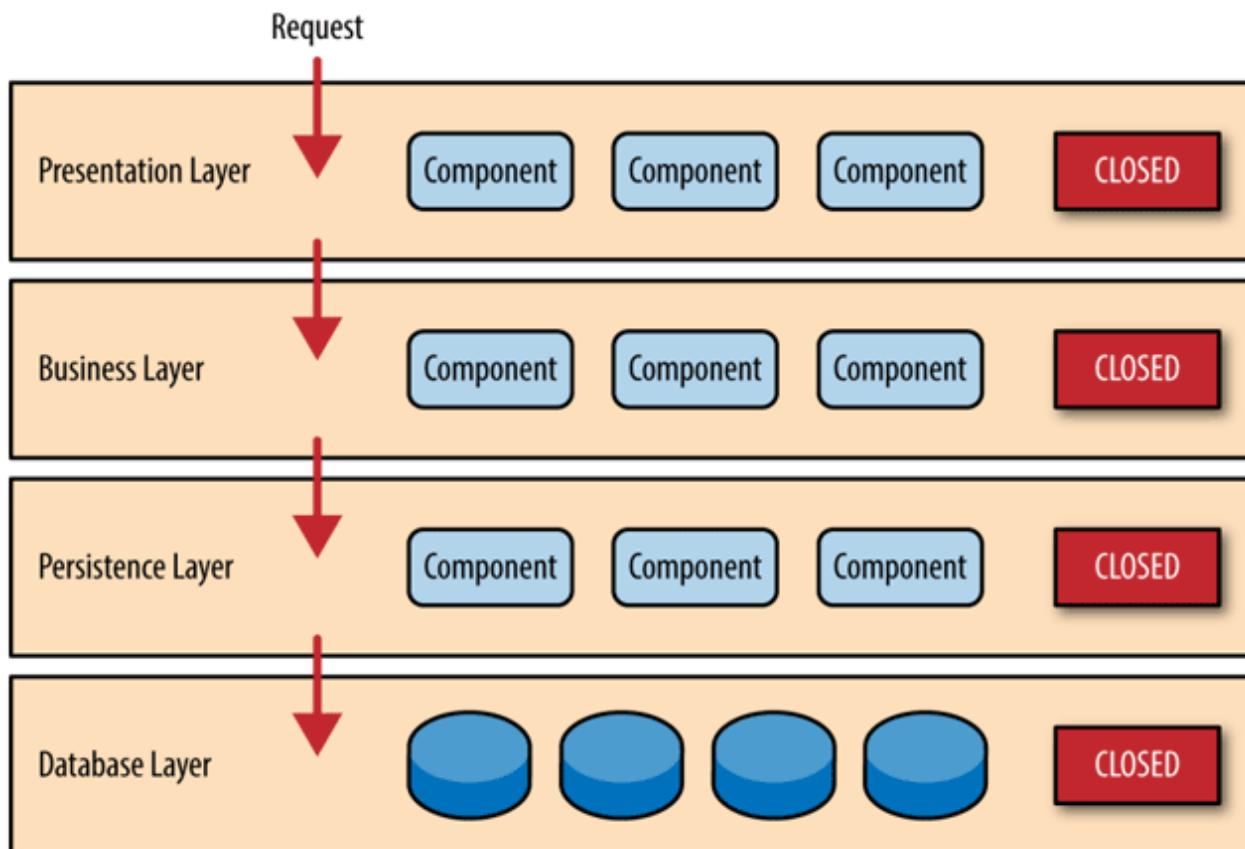
# 单体系统时代

## 单体架构 (Monolithic)

“单体”只是表明系统中主要的过程调用都是进程内调用，不会发生进程间通讯，仅此而已。

单体系统是绝大部分软件从业者都学习、实践过的架构形式，许多介绍微服务的书籍和技术资料中也把这种架构形式称作“[巨石系统](#)”（Monolithic Application），且某些文章中为对比展示出微服务架构的优点，往往会或有意或无意地强调、放大这种架构形式的缺点，以至于让人获得某种巨石系统就“人如其名”是铁板一块无可拆分也不可伸缩的，单体架构就不如微服务架构先进好用的潜在暗示。

如果说单体架构是一块巨石，不可拆分的显然有失偏颇，“单体”只是表明系统中主要的过程调用（不算数据库、文件、缓存等这类资源访问）都是进程内调用，不会发生[进程间通讯](#)（Inter-Process Communication，IPC。RPC属于IPC的一种特例，但请注意这里两个“PC”不是同个单词的缩写）。“Monolithic”一词在语言上确实最初是带有“单层”（Single-Tiered）的含义，但在现代信息系统中，笔者从未见过实际生产环境里的哪个大型的系统是完全不分层的。分层架构（Layered Architecture）已是现在大多数系统建设中普遍认可、普遍采用的软件设计方法，无论是单体还是微服务，抑或是其他架构风格，都会对代码进行横向拆分，收到的外部的请求在各层之间以不同形式的数据结构进行流转传递，触及最末端的数据库后依次返回响应。在这个意义上的“可拆分”，单体架构完全不会展露出丝毫的弱势，反而还可能会因更容易开发、部署、测试而获得一些便捷性上的好处。



图片来自O'Reilly的开放文档《Software Architecture Patterns》

至于比较微服务、单体架构哪种更先进，笔者认为“先进”不能是绝对的，这点可以举一个非常浅显的例子加以说明。譬如，沃尔玛将超市分为仓储部、采购部、安保部、库存管理部、巡检部、质量管理部、市场营销部，等等，可以划清职责，明确边界，让管理能力能支持企业的成长规模；但如果你家楼下开的小卖部，爸、妈加儿子，再算上管家的中华田园犬小黄一共也就只有四名员工，也去追求“先进管理”，划分仓储部、采购部、库存管理部……那纯粹是给自己找麻烦。

单体系统真正体现弱势的地方在于垂直切分上，它决定了哪怕是信息系统中两个相互毫无关联的子系统，也必须部署到一起。当系统规模小时这是优势，但系统规模大的时候，修改时候的部署成本、技术升级时的迁移成本都会变得高昂。按前面的例子来说，就是当公司小时，让安保部和质检部两个不相干的部门在同一栋大楼中办公是节约资源，但当公司人数增加，办公室已经拥挤不堪，也最多只能在楼顶加盖新楼层（相当于增强硬件性能），而不能让安保、质检分开地方办公，这才是缺陷所在。

不过，为了实现垂直拆分，并不意味着一定要依靠微服务架构才能解决，在新旧世纪之交，人们曾经探索过几种服务垂直拆分的方法，这些架构方法后来导致了面向服务架构（Service-Oriented Architecture）的一段兴盛期，我们称其为“SOA时代”。



# SOA时代

## SOA架构（Service-Oriented Architecture）

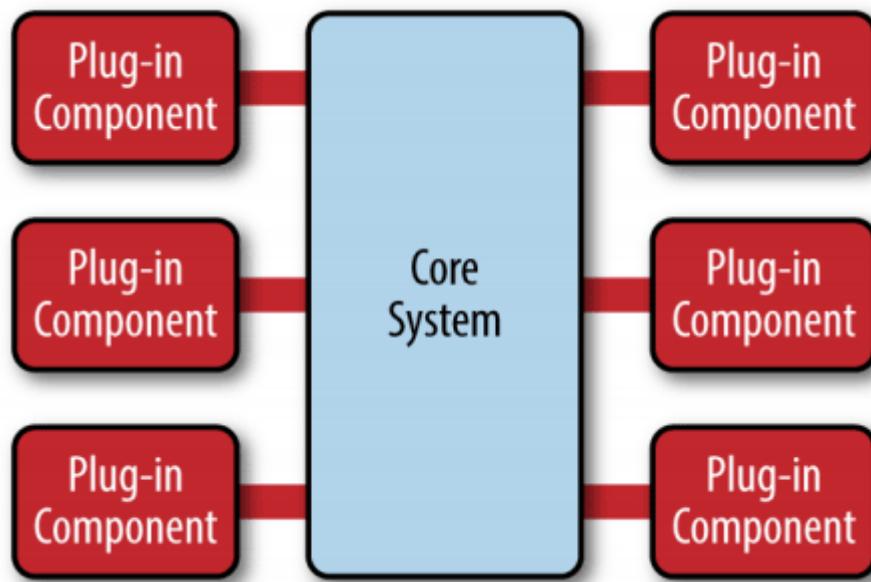
面向服务的架构是第一次系统性地成功解决分布式服务主要问题的架构模式。

当系统规模变大后，为了能对系统进行垂直拆分、复用，人们尝试过多种途径，笔者列举以下三种较有代表性的架构模式，分别为：

- **烟囱式架构**（Information Silo Architecture）：信息烟囱又名信息孤岛（Information Island），使用这种架构的系统也被称为孤岛式信息系统或者烟囱式信息系统。它指的是一种完全不与其他相关信息系统之间进行互操作或者说协调工作的信息系统。这样的系统其实并没有什么“架构设计”可言，如果两个部门真的完全不会发生交互，就并没有什么理由强迫把它们必须在一栋楼里办公；两个不发生交互的信息系统，让他它们使用独立的数据库、服务器即可完成拆分，而唯一的问题，也是致命的问题是，企业中真的存在完全不发生交互的部门？对于两个信息系统来说，哪怕真的毫无业务往来关系，但系统的人员、组织、权限等等主数据，会是完全独立、没有任何重叠的吗？这样“独立拆分”、“老死不相往来”的系统，显然不可能是企业所希望见到的。
- **微内核架构**（Microkernel Architecture）：微内核架构也被称为插件式架构（Plug-in Architecture）。既然烟囱式架构中，两个没有业务往来关系的系统也可能需要共享一部分的公共的主数据，那不妨就将这些主数据，连同其他可能被所有系统使用到的公共服务、数据、资源集中到一块，成为一个被所有业务系统共同依赖的核心系统（Kernel，也称为Core System），具体的业务系统就以插件模块（Plug-in Modules）的形式存在，这样便可提供可扩展的，灵活的，天然隔离的功能特性。

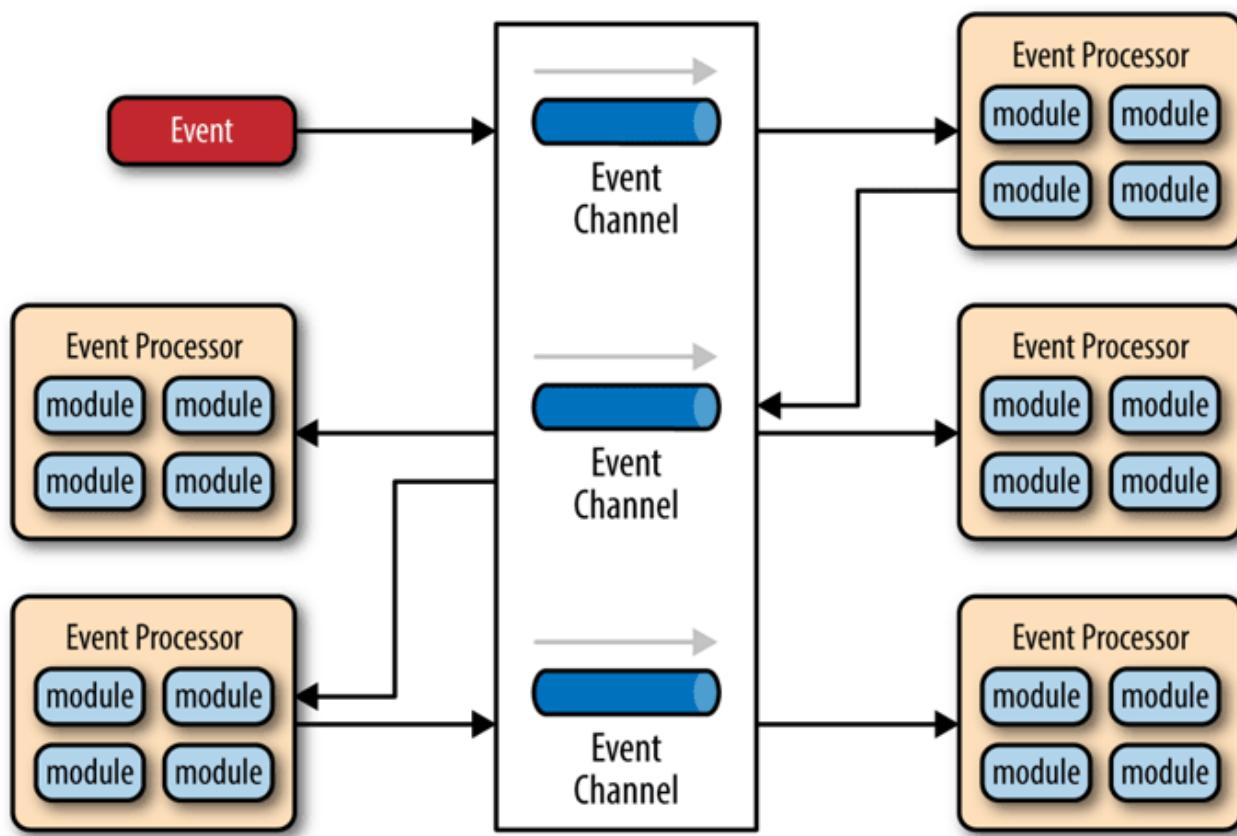
这种模式很适合桌面应用程序，也可以在Web应用程序中使用。事实上，本文列举的各种不同的架构模式一般都可视为整个系统的一种插件。对于产品型应用程序来说，如果我们想将新特性和功能及时加入系统，微内核架构是一种不错的选择。微内核的架构也可以嵌入到其它的架构模式之中，通过插件还可以提供逐步演化的功能和增量开发。所以如果你实现能够支持二次开发的软件系统，微内核是一种良好的架构模式。

不过，微内核架构也有它的局限和前提，这便是它假设各个插件模块之间是互不认识（不可预知系统会安装哪些模块），不会发生交互的，但无论在企业还是互联网，这一前提假设却通常并不成立，我们必须找到办法，既能垂直拆分系统，也能让拆分后的子系统之间可以互相调用通讯。



图片来自O'Reilly的开放文档《[Software Architecture Patterns](#)》

- **事件驱动架构** (Event-Driven Architecture)：为了能让子系统互相通讯，一种可行的方案是在子系统之间建立一套事件队列管道 (Event Queues)，来自系统外部的消息将以事件的形式发送至管道中，各个子系统可以从管道里获取自己感兴趣、可以处理的事件消息，可以为事件新增或者修改其中的附加信息，甚至可以自己发布一些新的事件到管道队列中去，如此，每一个消息的处理器都是独立的，高度解耦的，但又能与其他处理器（如果存在该消息处理器的话）通过事件管道进行互动。



图片来自O'Reilly的开放文档《Software Architecture Patterns》

当系统演化至事件驱动架构时，第一节提到的仍在并行发展的远程服务调用也来到了SOA时代（详见[远程服务调用](#)一文），此时SOA已经有了它登场所需要的全部前置条件。SOA的概念最初由Gartner公司在1994年提出，2006年，由IBM、Oracle、SAP等公司共同成立了OSOA联盟（Open Service Oriented Architecture），用于制定和推进SOA相关标准，在2007年，国际标准组织（OASIS）领导下，OSOA联盟的职能并入了新成立了[Open CSA](#)组织。

尽管SOA仍是抽象概念，而不是特指某一种具体的技术，但它已经比前面所说的三种架构模式要相对具体、充实了很多，已经不能简单视其为一种架构模式，可以称为是一套软件设计的方法论了。它拥有领导制定技术标准的组织Open CSA；有清晰软件设计的指导原则，譬如服务的封装性、自治、松耦合、可重用、可组合、无状态，等等；明确了采用SOAP作为远程调用的协议，依靠SOAP协议族（WSDL、UDDI和一大票WS-\*协议）来完成服务的发布、发现和治理；利用一个被称为[企业服务总线](#)（Enterprise Service Bus，ESB）的消息管道来实现各个子系统之间的通讯交互，令各服务间在ESB调度下无需相互依赖却能相互通讯，既带来了服务松耦合的好处，也为以后可以进一步实现[业务流程编排](#)（Business Process Management，BPM）提供了基础；使用[服务数据对象](#)（Service Data Object，SDO）提供了对各种数据源的统一访问能力。

ata Object , SDO ) 来访问和表示数据 , 使用服务组件架构 ( Service Component Architecture , SCA ) 来定义服务封装的形式和服务运行的容器 , 等等。

当软件架构发展至SOA时代 , 其中的许多思想都已经能在今天微服务中找到对应的身影了。服务之间的松散耦合、注册、发现、治理 , 隔离、编排 , 等等。这些今天微服务中耳熟能详的名词概念 , 多数是在分布式服务刚被提出时就已经存在的难题 , 这些难题在SOA时代进行过系统性的探索 , 才形成了前面所列的这些概念 ; 今天的微服务架构依然要面临这些问题 , 但它们在SOA架构中就曾经被解决过一次 , 甚至如果仅从技术可行性这一个角度来评判的话 , SOA可以算基本成功地解决了这些问题。

但是 , SOA并没有能真正解决的关键性问题反而是三十年前原始分布式时代时提出的“如何使用多个独立的分布式服务共同构建一个更大型系统 ? ”——这本该是SOA或者任何一个分布式架构的首要目标。笔者曾在[远程服务调用](#)一文中提到SOAP协议被逐渐边缘化的本质原因 : 过于严格的规范定义带来过度的复杂性。而构建在SOAP基础之上的ESB、BPM、SCA、SDO等诸多上层建筑 , 进一步加剧了这种复杂性。SOA诞生的那一天起 , 就已经注定了它只能是少数系统阳春白雪式的精致奢侈品 , 它可以实现多个异构大型系统之间的复杂集成交互 , 却很难作为一种具有广泛普适性的软件架构风格来推广。SOA最终没有获得成功的致命伤与当年的EJB如出一辙 , 尽管有Sun Microsystems和IBM等一众巨头在背后力挺 , EJB仍然败于以Spring、Hibernate为代表的“草根框架” , 可见一旦脱离人民群众 , 终究会淹没在群众的海洋之中 , 连信息技术也不曾例外过。

当你读到这一段的时候 , 不妨重新翻到开头 , 回头想一想“如何使用多个独立的分布式服务共同构建一个更大型系统”这个问题 , 再回顾下“原始分布式时代”一节中Unix DCE提出的分布式服务的主旨 : “让开发人员不必关心服务是远程还是本地 , 都能够透明地调用服务或者访问资源”。经过了三十年的技术进步 , 信息系统经历了巨石、烟囱、微内核、事件驱动、SOA等等的架构模式 , 应用受架构复杂度的牵绊却是越来越大 , 已经距离“透明”二字越来越远了 , 这是否算不自觉间忘记掉了当年的初心 ? 接下来我们所谈论的微服务时代 , 似乎正是带着这样的自省式的问句而开启的。

# 微服务时代

## 微服务架构（Microservices）

微服务是一种通过多个小型服务组合来构建单个应用的架构风格，这些服务围绕业务能力而非特定的技术标准来构建。各个服务可以采用不同的编程语言，不同的数据存储技术，运行在不同的进程之中。服务采取轻量级的通讯机制和自动化的部署机制实现通讯与运维。

“微服务”这个技术名词最早在2005年就已经被提出，它是由Peter Rodgers博士在2005年度的云计算博览会（Web Services Edge 2005）上首次使用，当时的说法是“Micro-Web-Service”，指的是一种专注于单一职责的、语言无关的、细粒度Web服务（Granular Web Services）。“微服务”一词并不是Peter Rodgers直接凭空创造出来的概念，初生的微服务可以说是SOA发展时催生的产物，就如同EJB推广过程中催生了Spring和Hibernate那样。这一阶段的微服务是作为一种SOA的轻量化的补救方案而被提出的。时至今日，在英文版的维基百科上，仍然将微服务定义为一种SOA的变种形式，所以微服务在最初阶段与SOA、Web Service这些概念有所牵扯也完全可以理解，但现在来看，维基百科对微服务的定义已经颇有些过时了。

### What is microservices

Microservices is a software development technique — a variant of the service-oriented architecture (SOA) structural style.

—— Wikipedia , [Microservices](#) ↗

微服务的概念提出后，将近10年的时间里面，都并没有受到太多的追捧，如果只是对现有SOA架构的修修补补，确实是难以唤起广大技术人员的更多激情了。不过，在这10年时间里，微服务本身也在思考、蜕变。2012年，在波兰克拉科夫举行的“33rd Degree Conference”大会上，Thoughtworks首席咨询师James Lewis做了题为《Microservices - Java, the U

nix Way》的主题演讲，其中提到了单一服务职责、康威定律、自动扩展、领域驱动设计等原则，却只字未提SOA，反而提倡应该重拾Unix的设计哲学（The Unix Philosophy），这点仿佛与笔者在前一节所说的“初心与自省”在遥相呼应。微服务已经迫不及待地要脱离SOA的附庸，成为一种独立的架构风格，也许，还将会是SOA的革命者。

微服务真正的崛起是在2014年，相信阅读此文的大多数读者，也是从Martin Flower与James Lewis合写的文章《Microservices:a definition of this new architectural term》中首次了解到微服务的，并不是指各位一定读过这篇文章，或者准确地说，今天各位所了解的“微服务”是这篇文章中提出的“微服务”。在此文中，定义了现代微服务的概念：“**微服务是一种通过多个小型服务组合来构建单个应用的架构风格，这些服务围绕业务能力而非特定的技术标准来构建。各个服务可以采用不同的编程语言，不同的数据存储技术，运行在不同的进程之中。服务采取轻量级的通讯机制和自动化的部署机制实现通讯与运维**”。此外，文中给出了微服务的九个核心的业务与技术特征，包括：

- **围绕业务能力构建**（Organized around Business Capabilities），这里再次强调了康威定律的重要性
- **分散治理**（Decentralized Governance），这是表达“谁家孩子谁来管”的意思，服务对应的开发团队有直接对服务运行质量负责的责任，也有着不受干预地掌控服务各个方面的权力，譬如选择与其他服务异构的技术来实现自己的服务
- **可独立替换升级的组件**（Componentization via Services），之所以通过“服务”（Service）而不是“类库”（Library）来构建组件，就是为了获得独立升级替换的能力
- **产品化思维**（Products not Projects），避免把软件研发视作要完成某种功能，而是视作一种持续改进、提升的过程
- **数据去中心化**（Decentralized Data Management），提倡数据按领域分散管理、更新、维护、存储
- **基础设施自动化**（Infrastructure Automation），由于服务变多且分散，微服务对CI/CD等基础设施的依赖程度更高，这一点在后续出现的“云原生”概念中，甚至被提升到与微服务本身平行的程度
- **轻量级通讯机制**（Smart Endpoints and Dumb Pipes），弱管道（Dumb Pipes）直接指名道姓地批评ESB那种复杂而刻板的管道通讯机制
- **容错性设计**（Design for Failure），不再虚幻地追求服务永远稳定，而是接受服务会出错的现实，笔者认为这是微服务最大的价值所在，也是这部开源文档标题“The Fenix Project”的含义

- **演进式设计** ( Evolutionary Design ) , 承认要设计一个靠谱的微服务 , 对架构者的能力与经验要求会比单体系统更高 , 因此建议依赖微服务所获得的灵活性 , 在有必要“微服务”的地方再进行“微服务”处理

此文中定义的微服务已经明确地与SOA划清了界线 , 拒绝再贴上任何SOA的标签。如此 , 微服务的概念才算是种真正丰满、独立、具体的架构风格 , 为它在未来的几年时间里如明星一般闪耀崛起于技术舞台铺下了厚实基础。

### Microservices and SOA

This common manifestation of SOA has led some microservice advocates to reject the SOA label entirely, although others consider microservices to be one form of SOA , perhaps service orientation done right.

—— Martin Flower / James Lewis , [Microservices](#)

从以上微服务的定义和特征中还可以明显地感觉到 , 微服务是相对自由的架构风格 , 摒弃了几乎所有可以抛弃的约束和规定 , 提倡以“实践标准”代替“规范标准”。可是 , 如果没有了统一的规范和约束 , 以前SOA所解决的那些分布式服务的问题 , 不也就一下子都重新都出现了吗 ? 的确如此 , 服务的注册发现、跟踪治理、负载均衡、故障隔离、认证授权、伸缩扩展、传输通讯、事务处理 , 等等 , 这些问题 , 微服务中不再会有统一的解决方案 , 即使只讨论Java范围内会使用到的微服务 , 光一个服务间通讯问题 , 可以列入解决方案的候选清单的就有 : RMI ( Sun/Oracle ) 、 Thrift ( Facebook ) 、 Dubbo ( 阿里巴巴 ) 、 gRPC ( Google ) 、 Motan2 ( 新浪 ) 、 Finagle ( Twitter ) 、 brpc ( 百度 ) 、 Arvo ( Hadoop ) 、 JSON-RPC、 REST , 等等 ; 光一个服务发现问题 , 可以选择的就有 : Eureka ( Netflix ) 、 Consul ( HashiCorp ) 、 Nacos ( 阿里巴巴 ) 、 Zookeeper ( Apache ) 、 etcd ( Core OS ) 、 CoreDNS ( CNCF ) , 等等。其他领域的情况也是与此类似 , 总之 , 完全是八仙过海 , 各显神通的局面。

微服务所带来的自由是一把双刃开锋的宝剑 , 当软件架构者拿起这把宝剑 , 一刀指向SOA定下的复杂技术标准 , 将选择的权力夺回的同一时刻 , 另外一刀也正朝向着自己映出冷冷的寒光。微服务时代中 , 软件研发本身的复杂度应该说是有所降低 , 一个简单服务 , 并不见得就会同时面临分布式中所有的问题 , 也就没有必要背上SOA那百宝袋般沉重的技术包袱。需要解决什么问题 , 就引入什么工具 ; 团队熟悉什么技术 , 就使用什么框架。此外 , 像Spring Cloud这样的胶水式的全家桶工具集 , 通过一致的接口、声明和配置 , 进一步屏

蔽了源自于具体工具、框架的复杂性，降低了在不同工具、框架之间切换的成本，所以，作为一个普通的服务开发者，作为一个“螺丝钉”式的程序员，微服务架构是友善的。可是，微服务对架构者是满满的恶意，对架构能力要求已提升到史无前例的程度，笔者在这部文档的多处反复强调过，技术架构者的第一职责就是做决策权衡，有利有弊才需要决策，有取有舍才需要权衡，如果架构者本身的知识面不足以覆盖所需要决策的内容，不清楚其中利弊，恐怕也就无可避免地陷入选择困难症的困境之中。

微服务时代充满着自由的气息，微服务时代充斥着迷茫的选择。软件架构不会止步于自由，微服务仍不是架构探索终点，如果有下一个时代，我希望是信息系统能同时拥有微服务的自由权利，围绕业务能力构建自己的服务而不受技术规范管束，但同时又不必以承担自行解决分布式的问题的责任为代价。管他什么利弊权衡！小孩子才做选择题，成年人全部都要！

# 后微服务时代

## 后微服务时代

从软件层面独力应对微服务架构问题，发展到软硬一体，合力应对架构问题的时代，此即为“后微服务时代”。

在微服务中面临的问题，自SOA时代，甚至可以说自原始分布式时代以来就一直存在，注册发现、跟踪治理、负载均衡，传输通讯，等等，这些问题，在分布式系统中都是无可避免的，但这些问题非得由分布式系统自己来解决吗？

我们先不纠结于微服务或者什么别的架构，直接来看待这些问题。如果某个系统需要伸缩扩容，通常会购买新的服务器；如果某个系统需要解决负载均衡问题，通常会布置负载均衡器；如果需要解决传输安全问题，通常会启用TLS传输链路；如果需要解决服务发现问题，通常会设置DNS服务器，等等。之所以微服务时代我们不得不在应用服务层面而不是基础设施层面去解决这些分布式问题，完全是因为由硬件构成的基础设施，跟不上由软件构成的应用服务的灵活性的无奈之举。软件可以只使用键盘就能拆分出不同的服务，只通过拷贝就能够伸缩扩容服务，硬件难道也可以通过敲键盘就变出相应的应用服务器、负载均衡器、DNS服务器、网络链路设施吗！嗯？好像也可以啊？

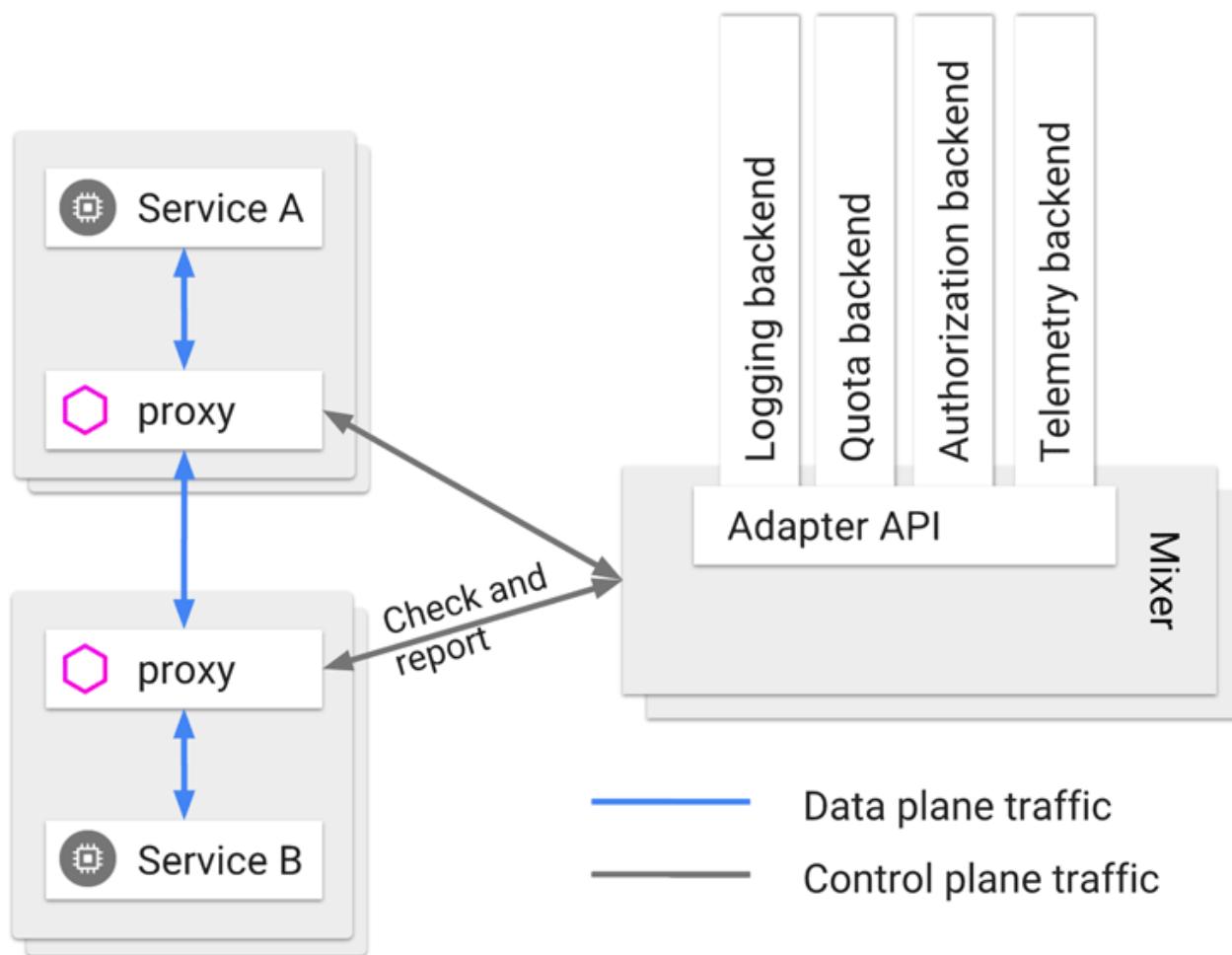
行至此，估计大家已经听出下面要说的是[虚拟化](#)技术和[容器化](#)技术了。微服务时代所取得的成就，本身就离不开以Docker为代表的早期容器化技术的巨大贡献。在此之前，笔者从来没有提过“容器”二字，这并不是刻意冷落，而是早期的容器只被简单地视为一种可快速启动的服务运行环境，目的是方便于程序的分发部署，这个阶段针对单个服务的容器并未真正参与到分布式问题的解决之中。尽管2014年微服务兴起的时候，Docker Swarm（2013年）和Apache Mesos（2012年）已经存在，更早之前也出现过[软件定义网络](#)（Software-Defined Networking，SDN）、[软件定义存储](#)（Software-Defined Storage，SDS）等技术，但是，被业界广泛认可、普遍采用的通过虚拟化的基础设施去解决分布式架构问题的方案，应该要从2017年Kubernetes赢得容器编排战争的胜利开始算起。

	Kubernetes	Spring Cloud
弹性伸缩	Autoscaling	N/A
服务发现	KubeDNS / CoreDNS	Spring Cloud Eureka
配置中心	ConfigMap / Secret	Spring Cloud Config
服务网关	Ingress Controller	Spring Cloud Zuul
负载均衡	Load Balancer	Spring Cloud Ribbon
服务安全	RBAC API	Spring Cloud Security
跟踪监控	Metrics API / Dashboard	Spring Cloud Turbine
降级熔断	N/A	Spring Cloud Hystrix

上表列出了在同一个分布式服务的问题在Spring Cloud中提供的应用层面的解决方案与在Kubernetes中提供的基础设施层面的解决方案，尽管因为各自出发点不同，解决问题的方法和效果都有所差异，但这无疑是提供了一条全新的、前途更加广阔的解题思路。“前途广阔”不仅仅是恭维赞赏，当虚拟化的基础设施从单个服务的容器发展至整个服务集群的所有硬件设施时，软件与硬件的界限便开始模糊。一旦硬件能够跟上软件的灵活性，那些与业务无关的技术性问题便有可能能从软件层面剥离，悄无声息地解决于硬件基础设施之内，让软件得以只专注业务，真正“围绕业务能力构建”。如此，DCE中未能实现的“透明的分布式应用”成为可能，Martin Flower设想的“[凤凰服务器](#)”成为可能，Chad Fowler提出的“[不可变基础设施](#)”成为可能，从软件层面独力应对分布式架构所带来的各种问题，发展到应用代码与基础设施软硬一体，合力应对架构问题的时代，现在常被媒体冠以“云原生”这个颇为抽象的名字加以宣传。云原生时代与此前微服务时代中追求的目标并没有本质改变，笔者更愿意称其为“后微服务时代”。

Kubernetes成为容器战争胜利者标志着后微服务时代的开端，但Kubernetes并没有能够解决全部的分布式问题，这是因为有一些问题处于应用系统与基础设施的边缘，使得完全在基础设施层面中很难完美地解决。举个例子，譬如微服务A调用了微服务B中发布的两个服务，称为B1和B2，假设B1表现正常但B2出现了持续的500错，那在达到一定阈值之后就应该对B2进行熔断，以避免产生[雪崩效应](#)。如果仅在基础设施层面来处理，这会遇到一个两难问题，切断A到B的网络通路则会影响到B1的正常调用，不切断的话则持续受B2的错误影响。

为了解决这一类问题，微服务基础设施很快进行了第二次进化，引入了今天被称为“[服务网格](#)”（Service Mesh）的“边车代理模式”（Sidecar Proxy）。所谓的“边车”是一种带垮斗的三轮摩托，我小时候还算常见，现在基本就只在抗日神剧中才会看到了。这里指的意思是会由系统自动在服务容器中注入一个通讯代理服务器（相当于那个垮斗），以类似网络安全里中间人攻击的方式，在应用无感知的情况下，悄然接管掉应用所有对外通讯。这个代理除了实现正常的服务调用通讯外（称为数据平面通讯），同时还接受控制器的指令（称为控制平面通讯），对数据平面通讯的内容进行分析，以实现熔断、认证、度量、监控、负载均衡等各种附加功能。



图片来自Istio的[配置文档](#)

很难从概念上判定清楚一个与应用系统运行于同一容器之内的代理服务到底应该算软件还是算基础设施，但它对应用是透明的，不需要改动任何软件代码就可以实现的服务治理，这便足够了。服务网格在2018年才火起来，今天它仍然是个新潮的概念，仍然未完全成熟，甚至连Kubernetes也还算是个新生事物（以它开源来计算）。但笔者相信，未来几年Kubernetes将会成为服务器端标准的运行环境，如同在此之前Linux；服务网格将会成为微服务之间通讯交互的主流模式，把“选择什么通讯协议”、“如何做认证授权”之类的技术问题

隔离于应用软件之外，取代今天Spring Cloud全家桶中大部分组件的功能，微服务只需要考虑业务本身的逻辑。

上帝的归上帝，凯撒的归凯撒，业务与技术完全分离，远程与本地完全透明，也许这就是最好的时代了吧？

# 无服务时代

## 无服务架构 ( Serverless )

如果说微服务架构是分布式系统这条路的极致，那无服务架构，也许就是“不分布式”的云端系统这条路的起点。

进行分布式的目的是由于单台机器的性能无法满足系统的运行需要，尽管后来架构演进过程中，容错能力、技术异构、职责划分等各方面因素都成为架构需要考虑的问题，但其中获得性能的需求在架构中比重依然很大。对软件研发而言，不做去分布式无疑是最简单的，如果单台服务器的性能可以是无限的，那架构演进的结果肯定会与今天有很大的差别，分布式也好，容器化也好，微服务也好，恐怕都未必会出现。

绝对意义上的无限性能必然是不存在的，但在云服务已经落地十年的今天，相对意义的无限性能已经成为了现实。2012年，iron.io公司率先提出了“无服务”（ Serverless，应该翻译为“无服务器”更合适）的概念，2014年开始，AWS发布Lambda的商业化无服务应用，在2019年，中国的阿里云、腾讯云等厂商也相应跟进了无服务的产品，一时间“无服务”又成为了技术届的“网红”。

无服务的概念并没有前面各种架构那么复杂，本来无服务也是以“简单”为主要卖点的，它只涉及两块内容：后端设施（ Backend ）和函数（ Function ）。后端设施是指数据库、消息队列、日志、存储，等等这一类用于支撑业务逻辑运行，但本身无业务含义的技术组件，这些后端设施都运行在云中，无服务中称其为“后端即服务”（ Backend as a Service , BaaS ）。函数就是指的业务逻辑代码，这里函数的概念与粒度，都已经很接近于程序编码角度的函数了，其区别是无服务中的函数运行在云端，不必考虑算力问题，不必考虑容量规划（从技术角度可以不考虑，从计费的角度你还是要掂量的），无服务中称其为“函数即服务”（ Function as a Service , FaaS ）。

无服务的愿景是让开发者只需要纯粹地考虑业务，不需要考虑技术组件，后端的技术组件是现成的，可以直接取用，不需要考虑如何部署，不需要考虑算力，也不需要操心运维。然而，与单体架构、微服务架构不同，无服务现在不是，以后估计也很难成为一种普适性

的架构模式，它对一些适合的应用确实能够大幅降低开发和运维环节的成本，譬如多数资讯类网站都适合于短链接、无状态的服务，但对于许多信息管理系统，或者说所有具有业务逻辑复杂，依赖服务端状态，响应速度要求较高，需要长链接，等等这些特征的应用，无服务架构至少目前是相对并不合适的。这是因为无服务天生“无限算力”的假设就决定了它必须要按使用量（函数运算的时间和内存）计费以控制消耗算力的规模，因而函数不会一直以活动状态常驻服务器，请求到了才会开始运行，这导致了函数不能依赖服务端状态，也导致能函数会有冷启动时间，响应的性能不可能太好（百毫秒到秒的级别）。不过，云计算毕竟是大势所趋，今天信息系统建设的概念和观念，在（较长尺度的）明天都是会转变适应云端的，届时无服务可能会有更广阔的应用空间。

如果说微服务架构是分布式系统这条路的极致，那无服务架构，也许就是“不分布式”的云端系统这条路的起点。笔者很难预想在架构演进之路上，微服务和无服务之后还会有什么，尽管目光所及之处，只是不远的前方，即使如此，依然可以看到那里有许多值得去完成的工作在等待我们。

### 架构演进之路

We can only see a short distance ahead, but we can see plenty there that needs to be done.

尽管目光所及之处，只是不远的前方，即使如此，依然可以看到那里有许多值得去完成的工作在等待我们。

—— Alan Turing , Computing Machinery and Intelligence [↗](#) , 1950

# 架构的普适问题

“设计者的视角”主要讲述作为一个架构师，你应该在做架构设计时思考哪些问题，有哪些主流的解决方案和行业标准做法，各种方案有什么优点或者缺点，不同的解决方法会带来什么不同的影响，等等。“架构”总是充满权衡博弈的，如果一件事情只有好处或者只有坏处，没有利弊优劣上的选择，也就无需架构师去做技术决策了。你所做的一个决定，可能关系到未来的系统在功能、质量属性上的高低，也关系着团队的成员工作、成长中的幸福感，请多思多想，慎重决定！

本章内容始终以业界标准方案和博弈为主线，代码怎样写、工具怎样用并不是我们讨论的主题。即使有部分内容会涉及到一些具体工具、类库的使用片段，这些代码也不足以成为它们的应用指南，起码不足以让一个完全不了解该工具的人学会如何使用，而是笔者用于说清楚某种解决方案的途径，仅此而已。

“架构的普适问题”、“设计方法论”部分，会讨论与架构风格（如单体、微服务、无服务等）无关或者关系不太密切的通用性话题。这些话题对于任何一个系统设计者来说都可能涉及到，建议通读。在“技巧与专题”部分，则会围绕某一个具体技术或领域，展开独立的讨论，建议根据兴趣和实际需要来选读。。

# 服务设计风格

在软件业发展的初期，程序编写都是以算法为核心的，程序员会把数据和过程分别作为独立的部分来考虑，数据代表问题空间中的客体，程序代码则用于处理这些数据，这种思维方式直接站在计算机的角度去抽象问题和解决问题，被称为面向过程的编程思想。与此相对，面向对象的编程思想则站在现实世界的角度去抽象和解决问题，它把数据和行为都看作是对象的一部分，这样可以让程序员能以符合现实世界的思维方式来编写和组织程序。

这两种思想出现的时间有先后，但在人类使用计算机语言来处理数据的工作中，无论提倡以计算机的思维还是提倡以人类的思维来抽象问题，都是合乎逻辑的，并不应该是评价它们先进性的标准。

12年一轮回，经过了上世纪90年代末到21世纪初期面向对象编程的火热之后，又出现了另一种考虑如何对内封装逻辑、对外重用服务的新思想：面向资源的编程思想。这种思想是把问题空间中的数据对象作为抽象的主体，把解决问题时从输入数据到输出结果的处理过程，看作是一个（组）数据资源的状态不断发生变换而导致的结果。这种思想有其生根的土壤基础：在跨越进程、跨越网络主机、跨越编程语言的分布式系统中，人们尝试过将之前在单进程应用里行之有效的面向过程、面向对象的服务设计方法改造迁移，使之适应分布式环境，这项工作总体上获得了成功，但在分布式环境里多少还是出现了一些新瑕疵，所以为另一种服务设计风格，即面向资源的编程思想留出了成长的空间。

尽管在2020年还谈论什么RESTful、RPC，大概是确实有点落伍了，可这个问题是一个架构设计者必须有明确取舍权衡的重要技术决策，今天笔者仍准备来谈一下这个话题。

# 远程服务调用

分布式系统各个节点中的机器大都通过特定的网络协议（HTTP、TCP等公有协议或JRMP<sup>1</sup>、GIOP<sup>2</sup>这样专有协议）相互访问，但网络协议只是负责往目标机器发送了一段文本或二进制的数据，为了建立可靠的服务，还有很多问题需要考虑：

- 服务所需的参数，服务返回的结果以什么格式传输？
- 服务变化了，如何兼容前后不同版本的格式？
- 如何提高网络利用的效率，譬如连接是否可被多个请求复用以减少开销？多个请求是否可以同时发往一个连接上？
- 如何提高数据序列化的效率？
- 如何保证网络的可靠性？譬如调用期间某个链接忽然断开了怎么办？
- 怎样进行异常处理？异常该如何让调用者获知？
- 万一发送的请求服务端不回复该怎么办？
- .....

早在1988年，绝大多数人都对分布式、远程服务没有什么概念的时候（这话轻了，说那时候多数人对计算机没什么概念都不嫌过分），Sun Microsystems就起草并向IETF提交了RFC 1050<sup>3</sup>规范，正式提出了远程服务调用（Remote Procedure Call，RPC）的概念，并设计了一套通用的、基于TCP/IP网络的、面向C语言的RPC协议，后被称为ONC RPC<sup>4</sup>（用以区别于Unix系统下专有的DEC/RPC<sup>5</sup>）。

## 远程服务调用

Remote Procedure Call is a protocol that one program can use to request a service from a program located in another computer on a network without having to understand the network's details. A procedure call is also sometimes known as a function call or a subroutine call.

1991年，万维网还没正式诞生的年代，对象管理组织<sup>6</sup>（Object Management Group，OMG）发布了跨进程、面向异构语言的服务调用协议：CORBA 1.0（Common Object Request Broker Architecture）。

uest Broker Architecture，1.0版本只提供了C语言的调用）。到1997年发布的CORBA 2.0版本，CORBA支持了C、C++、Java（1998年新加入的Java语言映射）等主流编程语言，这是第一套由国际标准组织牵头，多个主流软件提供商共同参与的分布式规范，当时影响力只有微软私有的[DCOM](#)可以与之媲美。

不过，CORBA与DCOM都没有获得最终的胜利，在1999年末，SOAP 1.0（Simple Object Access Protocol）规范的发布。SOAP是由微软和DevelopMentor共同起草的远程服务标准，随后提交给W3C成为国际标准，SOAP使用XML作为远程过程调用的编码载体（实际上并不绑定于XML-RPC，有SOAP over UDP这类其他载体的应用），当时XML是计算机工业最新的银弹，只要是定义为XML的东西几乎就都是好的，连微软自己都主动放弃了DCOM转投SOAP。

SOAP没有天生属于哪家公司的烙印，商业运作非常成功，很受市场欢迎，大量的厂商都想分一杯羹。但从技术角度来看，SOAP设计得并不优秀，甚至可以说是有显著缺陷的。对于开发者而言，SOAP最大的缺点是它那过于严格的规范定义，需要专门的客户端去调用和解析SOAP，也需要专门的服务去部署SOAP（如Apache Axis/CXF）。SOAP协议家族中，除它本身外包括了服务描述的[Web服务描述标准](#)（Web Service Description Language，WSDL）协议、服务发现的[统一描述、发现和集成](#)（Universal Description / Discovery and Integration，UDDI）协议、还有一堆几乎谁都说不清有多少个的[WS-\\*](#)的子功能协议，对开发者来说都是很大的学习负担。

人们对SOAP的热情迅速兴起，又逐渐冷却之后，远程服务器调用这个小小领域，开始进入了群雄混战、百家争鸣的战国时代，延续至今。相继出现了RMI（Sun/Oracle）、Thrift（Facebook）、Dubbo（阿里巴巴）、gRPC（Google）、Motan2（新浪）、Finagle（Twitter）、brpc（百度）、Arvo（Hadoop）、JSON-RPC 2.0（公开规范，JSON-RPC工作组）等一系列的协议/框架。这些框架功能、特点各不相同，有的是某种语言私有，有的能支持跨多门语言，有的运行在HTTP协议之上，有的能直接运行于TCP/UDP之上，但总体而言，RPC在朝着三个主要方向发展：

- 朝着**对象**发展，不满足于RPC将面向过程的编码方式带到分布式，希望在分布式系统中也能够进行跨进程的面向对象编程，代表为RMI、.NET Remoting，之前的CORBA和DCOM也可以归入这类，这条线有一个别名叫做[分布式对象](#)（Distributed Object）。
- 朝着**效率**发展，代表为gRPC和Thrift，传输效率（主要是Payload所占传输数据的比例大小，使用的传输协议和协议的设计都会影响到这点）和序列化效率的影响是最大的因

素，gRPC和Thrift都有自己优秀的私有序列化器，传输协议一个是HTTP2，支持多路复用和Header压缩，另一个直接基于TCP。

- 朝着**简化**发展，代表为JSON-RPC，说要选速度最快的RPC可能会有争议，但选速度最慢的，JSON-RPC大概是逃不了的。牺牲了功能和效率，换来的是协议的简单，接口与格式都更为通用。

不同的RPC框架所提供的不同特性多少是有矛盾的，很难有某一种框架说“我全部都要”。

譬如，要把面向对象那套全搬过来，就注定不会太简单（如建Stub、Skeleton就很烦了）；功能多起来，协议就要弄得复杂，效率一般就会受影响；要简单易用，那很多事情就必须遵循约定而不是配置才行；要重视效率，那就需要采用二进制的序列化器和较底层的传输协议，支持的语言范围容易受限。

也正是每一种RPC框架都有不完美的地方，所以才导致不断有新的RPC出现，也导致了跳出RPC的新想法出现，REST便有了它诞生的土壤。

# RESTful服务

REST无论是思想上、概念上、还是应用目标上，它与各种RPC协议只能算是有所牵连，但本质上并不是同一类型的东西。思想上的不同在上一节已经讨论过，就是面向过程的编程思想与面向资源的编程思想，至于什么是面向资源编程，稍后我们再详谈。

而概念上不同主要是指REST并不是一种远程服务调用协议，甚至可以把定语去掉，它就不是一种协议。协议都带有一定的规范性和强制性，至少也得有个文档吧，譬如JSON-RPC，它再简单，也要有个《[JSON-RPC Specification](#)》来规定它的格式细节、异常、响应码等信息，但REST并没有这些东西，尽管有一些指导原则，实际上并不受任何强制约束。常有人批评某个系统“设计得不是RESTful”，其实这句话本身就有争议，RESTful只是风格而不是规范，并且能完全达到REST所有指导原则的系统也是很少见的，这一点我们同样将在稍后详细讨论。

至于应用目标，REST与RPC在范围上是确有重合的，但实际上重合的区域并不大。上一节列举的RPC三个方向中，分布式对象这一条线的应用与REST可以说是毫无关系；而重视“效率”这个方向的应用，基本上就限制了只能是后端应用（前端应用对于网络协议、序列化器这两点都没有选择的余地，想要高效率也有心无力），在分布式服务各个后端节点之间通讯这一块，REST虽然照样可以用于任何语言（只要有个HTTP Client就可以用）之间的调用，但其实在需要追求“效率”的纯后端应用场景里REST使用率真算不得高。我们开发的REST服务，大多数的是提供给前端或效率不处于主要矛盾的部分后端场景去消费的。在前端这一块，一众RPC里最多也就是JSON-RPC有机会与REST产生竞争，其他所有RPC协议、框架，哪怕是支持HTTP协议，哪怕提供了JavaScript版本的客户端（如gRPC-Web），也只是存在[理论可行性](#)，很少见有实际项目把它们用到浏览器上的。

但尽管有如此多的不同，这两者还是产生了很多的比较与争论，就如同当年面向对象与面向过程一样，非得争出个高低不可。网上许多REST vs RPC的口水仗中说REST不好的，通常也并不是支持哪个RPC框架/协议比它好用，大多都只是不赞成REST的设计风格，心中说的本意其实是“面向资源编程”的思想不好，不如“面向过程编程”来得好用好理解。

## 理解REST

个人会有好恶偏爱，但计算机科学是务实的，有了面向过程之后，还能产生面向资源，并引起广泛的关注、使用和讨论，后者一定是一些面向过程没有的闪光点，或者解决/避免了一些面向过程中的缺陷。我们不妨先去理解REST为什么出现、解决什么问题、方法是什么，然后再来评价它。

许多人都知道REST源于Roy Thomas Fielding在2000年发表的博士论文：《Architectural Styles and the Design of Network-based Software Architectures》[»](#)，此文的确是REST的源头，但我们不能忽略Fielding的身份和之前工作的背景，这对理解REST的设计思想至关重要。

首先，Fielding是一名很优秀的软件工程师，他是Apache服务器的核心开发者，后来成为了著名的Apache软件基金会的合作创始人；同时，Fielding也是HTTP 1.0协议（1996年发布）的专家组成员，后来还成为了HTTP 1.1协议（1999年发布）的负责人。HTTP 1.1协议设计的极为成功，以至于发布之后长达十年的时间里，都没有多少人认为有修订的必要。用来指导HTTP 1.1协议设计的理论和思想，最初是以备忘录的形式在专家组成员之间交流，除了IETF、W3C的专家外，并没有在外界广泛流传。



Roy Thomas Fielding

从时间上看，对HTTP 1.1协议的设计工作贯穿了Fielding的整个博士研究生生涯，当起草HTTP 1.1协议的工作完成后，Fielding回到了加州大学欧文分校继续攻读自己的博士学位。第二年，他更为系统、严谨地阐述了这套理论框架，并且以这套理论框架导出了一种新的编程风格，他为这种风格取了一个很多人难以理解，但是今天已经广为人知的名字REST（**R**epresentational **S**tate **T**ransfer），即“表征状态转移”的缩写。

哪怕对编程和网络都很熟悉的同学，只从标题中也不太可能直接弄明白什么叫“表征”、啥东西的“状态”、从哪“转移”到哪。尽管在论文原文中确有论述这些概念，但写得确实相当晦涩（不想读英文的同学从此[获得中文版本](#)），我推荐一种比较好的方式是先理解什么是HTTP，再配合一些实际例子来进行类比，你会发现“REST”实际上是“HTT”（**H**yper **T**ext **T**ransfer）的进一步抽象，两者就如同接口与实现类之间的关系一般。

HTTP中使用的“超文本”一词是美国社会学家Theodor Holm Nelson在1967年于《[Brief Words on the Hypertext](#)》一文里提出的，下面引用的是他本人在1992年修正后的定义：

### Hypertext

By now the word "hypertext" has become generally accepted for branching and responding text, but the corresponding word "hypermedia", meaning complexes of branching and responding graphics, movies and sound – as well as text – is much less used. Instead they use the strange term "interactive multimedia": this is four syllables longer, and does not express the idea of extending hypertext.

—— Theodor Holm Nelson [Literary Machines](#), 1992

以上定义描述的“超文本（或超媒体）”是一种“能够对操作进行判断和响应的文本（或声音、图像等）”，这个概念在上世纪60年代提出时应该还属于科幻的范畴，但是今天大众已经完全接受了它，互联网中一段文字可以点击、可以触发脚本执行、可以调用服务端，这一切已稀松平常，毫不稀奇。那我们继续尝试从“超文本”或者“超媒体”的含义来理解什么是“表征”以及REST中其他关键概念，笔者使用一个具体事例来将其描述如下：

- **资源（Resource）**：譬如你现在正在阅读一篇名为《服务设计风格》的文章，这篇文章中的内容本身（你将其视作是某种信息、数据）我们称之为“资源”。无论你是在网上看的网页、是打印出来看的文字稿、是在电脑屏幕上阅读抑或是手机上浏览，尽管呈现的样子各不相同，但其中的信息是不变的，你所阅读的仍是同一个“资源”。

- **表征** ( Representation ) : 当你通过电脑浏览器阅读此文章时 , 浏览器向服务端发出请求“我需要这个资源的HTML格式” , 服务端向浏览器返回的这个HTML就被称为“表征” , 你可能通过其他方式拿到本文的PDF、Markdown、RSS等其他形式的版本 , 它们也同样是一个资源的多种表征。可见“表征”这个概念是指信息与用户交互时的表示形式 , 这与我们应用分层中常说的“表示层” ( Presentation Layer ) 的语义其实是一致的。
- **状态** ( State ) : 当你把这篇文章阅读完毕 , 想看下一篇文章是什么内容的时候 , 你向服务器请求“给我下一篇” , 但是“下一篇”是个相对概念 , 必须依赖“当前你正在阅读的文章是哪一篇”才能正确回应 , 这类在特定语境中才能产生的上下文信息即被称为“状态”。我们所说的有状态 ( Stateful ) 还是无状态 ( Stateless ) , 都是只相对于服务端来说的 , 服务器要完成“取下一篇”的请求 , 要么自己记住用户的状态 ( 这个用户现在阅读的是哪一篇文章 , 这是有状态 ) , 要么客户端来记住状态 , 在请求的时候明确告诉服务器 ( 我正在阅读某某文章 , 现在要读下一篇 , 这是无状态 ) 。
- **转移** ( Transfer ) : 无论状态是由服务端还是客户端来提供的 , “取下一篇”这个行为逻辑必然只能由服务端来提供。服务器通过某种方式 , 把“用户当前阅读的文章”转变成“下一篇” , 这就被称为“表征状态转移”

借着这个故事的上下文 , 笔者顺便再介绍几个现在不涉及但稍后要用到的概念名词 :

- **统一接口** ( Uniform Interface ) : 上面说的“服务器通过某种方式”具体是什么方式 ? 请把本文拉到结尾处 , 右下角有下一篇的URI超链接地址 , 这是服务端渲染这篇文章时就预置好的 , 点击它让页面跳转到下一篇 , 就是一种所谓的“某种方式”。但URI的含义是统一资源标识符 , 如何能表达出“转移”的含义呢 ? HTTP协议中提前约定好了一套“统一接口” , 包括 : GET、HEAD、POST、PUT、DELETE、TRACE、OPTIONS七种操作 , 任何一个支持HTTP协议的服务器都会遵守这套规定 , 对特定的URI采取这些操作 , 服务器自然就会触发相应的表征状态转移。
- **超文本驱动** ( Hypertext Driven ) : 尽管表征状态转移是由浏览器主动向服务器发出请求 , 该请求导致了“在我们浏览器的屏幕上显示出了下一篇文章的内容”这个结果的出现 , 但浏览器其实根本不知道系统中这套转移逻辑。它根据是用户输入的URI地址请求网站首页 , 服务器给予的首页超文本内容 , 我们是通过内部的超链接导航到了这篇文章 , 阅读结束时再导航到下一篇。浏览器作为所有网站的通用的客户端 , 任何网站的导航 ( 状态转移 ) 行为都是不可能预置于浏览器之中 , 而是由服务器每一个请求中的返回信息 ( 超文本 ) 来驱动的。这点大家习以为常 , 但其实与其他带有客户端的软件有很本

质的区别，在那些软件中，业务逻辑往往是预置于客户端之中的，有专门的页面控制器（无论在服务端还是在客户端中）来驱动页面的状态转移。

- **自描述消息** ( Self-Descriptive Messages ) : 由于资源的表征可能存在多种不同形态，在消息中应当有明确的信息来告知客户端该消息的类型以及该如何处理这条消息。一种被广泛采用的自描述方法是在名为“Content-Type”的HTTP Header中标识出[互联网媒体类型](#) ( MIME type )，譬如“Content-Type : application/json; charset=utf-8”，则说明该资源会以JSON的格式来返回，请使用UTF-8字符集进行处理。

建立了上面这些概念之后，我们就可以开始讨论面向资源的编程思想与REST所提出的几个具体的软件架构设计原则了。请注意，Fielding提出REST时所谈论的范围是“架构风格与网络的软件架构设计” ( Architectural Styles and Design of Network-based Software Architectures )，而不是现在被人们所狭义理解的一种“服务 ( API ) 设计风格”，这两者的范围差别就好比本站全站所谈论的话题“现代软件架构探索”与本篇文章谈论的“服务设计风格”一般，前者是后者的一个很大的超集（但是基于本文的主题和多数人的关注点，后文还是会从着重于“服务设计”的视角出发的）。

Fielding认为，一套理想的、完全满足REST的系统应该满足以下六个原则：

## 1. 服务端与客户端分离 ( Client-Server )

将用户界面所关注的逻辑和数据存储所关注的逻辑分离开来有助于提高用户界面的跨平台的可移植性，这一点正越来越受到广大开发者所认可，以前完全基于服务端控制和渲染的JSF这类框架实际用户已甚少，而在服务端进行界面控制 ( Controller )，通过服务端或者客户端的模版渲染引擎来进行界面渲染 ( Render ) 的框架 ( Struts、SpringMVC ) 也受到了颇大的冲击。这一点主要推动力量与REST关系并不大，前端技术（从ES规范，到语言实现，到前端框架等）的近年来的高速发展，使得前端表达能力大幅度加强才是真正的幕后推手。

## 2. 无状态 ( Stateless )

这是REST的一条关键原则，部分开发者在做服务接口规划时，觉得RESTful风格的API怎么设计都别扭，很有可能的一种原因是在服务端持有着比较重的状态。REST希望服务器能不负责维护状态，每一次从客户端发送的请求中，应包括所有的必要的上下文信息，会话信息也由客户端保存维护，服务器端依据客户端传递的状态信息来进行业务处理，并且驱动整个应用的状态变迁。至于客户端承担状态维护职责后的认证、授权等各方面的可信问题，都有针对性的解决方案（详见下一篇：[安全架构](#)）

但必须承认的现状是，目前大多数的系统是达不到这个要求的，越复杂、越大型的系统

越是如此。服务端无状态可以在分布式环境中获得非常高价值的好处，但大型系统的上下文状态数量完全可能膨胀到让客户端在每次请求时提供变得不切实际的程度，在服务端的内存、会话、数据库或者集中式缓存等地方持有一定的状态成为一种是事实上被广泛使用的主流的方案。

### 3. 可缓存 ( Cacheability )

无状态服务虽然提升了系统的可见性、可靠性和可伸缩性，但降低了系统的网络性。这句话通俗的解释就是，某个功能使用有状态的架构只需要一次请求就能完成，而无状态的服务则可能会需要多个请求才行。为了缓解这个矛盾，REST希望软件系统能够如同万维网一样，客户端和中间的通讯传递者（代理）可以将部分服务端的应答缓存起来。当然，应答中必须明确地或者间接地表明本身是否可以进行缓存，以避免客户端在将来进行请求的时候得到过时的数据。运作良好的缓存机制可以减少客户端、服务器之间的交互，甚至有些场景中可以完全避免交互，这就进一步提了高性能。

### 4. 分层系统 ( Layered System )

这里所指的并不是表示层、服务层、持久层这种意义上的应用分层。而是指客户端一般不需要知道是否直接连接到了最终的服务器，抑或是路径上的中间服务器。中间服务器可以通过负载均衡和共享缓存的机制提高系统的可扩展性，这样也可也便于缓存、伸缩和安全策略的部署。譬如，一种典型的应用是内容分发网络（CDN），如你现在访问这个站点，你所发出的请求一般（假设你在中国国境内的话）并不是直接访问位于GitHub Pages的源服务器，而是访问了位于腾讯云的CDN，但你并不需要感知到这一点。我们将在“透明多级分流系统”中讨论如何构建可缓存的分层系统。

### 5. 统一接口 ( Uniform Interface )

这是REST的另一条关键原则，REST希望开发者面向资源编程，希望设计软件系统的核心放在抽象系统该有哪些资源，而不是抽象系统该有哪些行为（服务）。对资源的操作是可数的、固定的、统一的，由于REST并没有设计新的协议，所以这些操作都借用了HTTP协议中固有的操作命令来完成。

这一点也是REST最容易陷入争论的地方，基于网络的软件系统，到底是面向资源更好，还是面向服务更好，这事情哪怕到了今天仍然是没有个定论，也许永远都没有。但是，有一个基本清晰的结论是，面向资源编程的抽象程度通常更高，这意味着坏处是往往距离人类的思维方式更远，而好处是往往通用程度会更好。这样诠释REST大概本身就挺抽象的，还是举个例子来说明：譬如几乎每个系统都有的登录和注销功能，如果你理解成登录对应于login()服务，注销对应于logout()服务这样两个服务，这是“符合人类思维”的；如果你理解成登录是CREATE Session，注销是REMOVE Session，这样你只需要设计一种“Session资源”即可满足需求，甚至以后对Session的其他需求，譬如查询

或者修改登陆用户的信息，都可以在这一套设计中囊括在内，这便是“抽象程度更高”带来的好处。

想要在架构设计中合理恰当地利用统一接口，Fielding建议系统应能做到每次请求中都包含资源的ID，所有操作均通过资源ID来进行；建议每个资源都应该是自描述的消息；建议通过超文本来驱动应用状态的转移。

## 6. 按需代码 ( Code-On-Demand )

这被Fielding列为一条可选原则。按需代码指任何按照客户端软件（譬如浏览器）的请求，将可执行的软件程序从服务器计算机发送到客户端的技术。这是可选的原因并非是它特别难以达到，而更多是出于必要性和性价比的考虑。举个例子，譬如你使用Element-UI组件库开发一个Web应用，但其实只用了里面一两个组件，却没有好好配置babel-plugin-component来做按需引入，一下子把几十个组件都打包进脚本中，这便是没有贯彻好按需代码的原则。这类事情（引入一个类库可能只使用其中很少量的一部分代码）是相当普遍的，但我个人并不赞成不考虑实际场景的唯性能论，在关键场景肯定要抠细节，但所有场景都无限度的“精益求精”并无必要。

REST的基本思想是面向资源来抽象问题，基本手段是尽可能复用HTTP协议中已经定义的语义和相关基础支持来解决问题，以上六个原则都是在这个指导思路下设计的。因为HTTP本来就是面向资源而设计的网络协议，只要面向资源的软件架构确实行得通的话，本文开篇中所列的“远程服务调用需要考虑的问题”便几乎不再需要独立考虑了，HTTP协议已经有效运作了30年，其相关的技术基础设施已是千锤百炼，无比成熟，这些问题早已解决过无数遍。唯一需要权衡的是你的软件系统、设计和开发人员是否能够适应面向资源的思想来设计软件，来编写代码。

# RMM成熟度

前面我们花费大量篇幅讨论了REST的思想、概念和指导原则等理论方面的内容，在这个小节里，我们把重心放在实践上，同时把目光从整个软件架构设计聚焦到REST服务接口，以切合本节的题目“服务设计风格”，也顺带填了前面埋下的“如何评价服务是否RESTful”的坑。

《RESTful Web APIs》和《RESTful Web Services》的作者Leonard Richardson曾提出过一个衡量“服务有多么REST”的Richardson成熟度模型（Richardson Maturity Model）

) , 便于那些原本不使用REST的服务 , 能够逐步地导入REST。Richardson将服务接口“REST的程度”从低到高 , 分为0至4级 :

0. The Swamp of Plain Old XML : 完全不REST。另外 , 关于POX这说法 , SOAP表示感觉有被冒犯到。
1. Resources : 开始引入资源的概念。
2. HTTP Verbs : 引入统一接口 , 映射到HTTP协议的方法上。
3. Hypermedia Controls : 在本文里面的说法是“超文本驱动” , 在Fielding论文里的说法是“Hypertext As The Engine Of Application State , HATEOAS” , 都是指同一件事情。

我们借用Martin Fowler撰写的关于RMM成熟度模型的文章中的实际例子 ( 原文是XML写的 , 我简化了一下 ) , 来实际看一下四种不同程度的REST反应到实际API是怎样的。假设你是一名软件工程师 , 接到需求 ( 也被我尽量简化了 ) 的UserStory是这样的 :

### 医生预约系统

作为一名病人 , 我想要从系统中得知指定日期内我熟悉的医生是否具有空闲时间 , 以便于我向该医生预约就诊。

## 第0级

医院开放了一个/appointmentService的Web API , 传入日期、医生姓名作为参数 , 可以得到该时间段该名医生的空闲时间 , 该API的一次HTTP调用如下所示 :

```
POST /appointmentService?action=query HTTP/1.1
{date: "2020-03-04", doctor: "mjones"}
```

然后服务器会传回一个包含了所需信息的回应 :

```
HTTP/1.1 200 OK
[
 {start:"14:00", end: "14:50", doctor: "mjones"},
 {start:"16:00", end: "16:50", doctor: "mjones"}
]
```

得到了医生空闲的结果后，我觉得14:00的时间比较合适，于是进行预约确认，并提交了我的基本信息：

```
POST /appointmentService?action=confirm HTTP/1.1

{
 appointment: {date: "2020-03-04", start:"14:00", doctor: "mjones"},
 patient: {name: xx, age: 30,}
}
```

如果预约成功，那我能够收到一个预约成功的响应：

```
HTTP/1.1 200 OK

{
 code: 0,
 message: "Successful confirmation of appointment"
}
```

如果发生了问题，譬如有人在我前面抢先预约了，那么我会在响应中收到某种错误信息：

```
HTTP/1.1 200 OK

{
 code: 1
 message: "doctor not available"
}
```

到此，整个预约服务宣告完成，直接明了，我们采用的是非常直观的基于RPC风格的服务设计似乎很容易就解决了所有问题……吗？

## 第1级

通往REST的第一步是引入资源的概念，在API中基本的体现是围绕着资源而不是过程来设计服务，说的直白一点，可以理解为服务的Endpoint应该是一个名词而不是动词。此外，每次请求中都应包含资源的ID，所有操作均通过资源ID来进行。

```
POST /doctors/mjones HTTP/1.1
```

```
{date: "2020-03-04"}
```

然后服务器传回一个包含了ID信息，注意，ID是资源的唯一编号，有ID即代表“医生的档期”被视为一种资源：

```
HTTP/1.1 200 OK
```

```
[
 {id: 1234, start:"14:00", end: "14:50", doctor: "mjones"},
 {id: 5678, start:"16:00", end: "16:50", doctor: "mjones"}
]
```

我还是觉得14:00的时间比较合适，于是又进行预约确认，并提交了我的基本信息：

```
POST /schedules/1234 HTTP/1.1
```

```
{name: xx, age: 30,}
```

后面预约成功或者失败的响应消息在这个级别里面与之前一致，就不重复了。比起第0级，第1级的服务抽象程度有所提高，但至少还有三个问题并没有解决，一是只处理了查询和预约，如果我临时想换个时间，要调整预约，或者我的病忽然好了，想删除预约，这都需要提供新的服务接口。二是处理结果响应时，只能靠着结果中的code、message这些字段做分支判断，每一套服务都要设计可能发生错误的code，这很难考虑全面，而且也不利于对某些通用的错误做统一处理；三是并没有考虑认证授权等安全方面的内容，譬如要求只有登陆用户才允许查询医生档期时间，某些医生可能只对VIP开放，需要特定级别的病人才能预约等等。

## 第2级

第1级遗留三个问题都可以靠引入统一接口来解决。HTTP协议的七个标准方法是经过精心设计的，几乎能涵盖资源可能遇到的所有操作场景（这其实更取决于架构师的抽象能力）。REST的做法是把不同业务需求抽象为对资源的增加、修改、删除等操作来解决第

一个问题；使用HTTP协议的Status Code，可以涵盖大多数资源操作可能出现的异常（而且也是可以自定义扩展的），以此解决第二个问题；依靠HTTP Header中携带的额外认证、授权信息来解决第三个问题（这个在实战中并没有体现，请参考安全架构中的“凭证”相关内容）。

按这个思路，获取医生档期，应采用具有查询语义的GET操作进行：

```
GET /doctors/mjones/schedule?date=2020-03-04&status=open HTTP/1.1
```

然后服务器会传回一个包含了所需信息的回应：

```
HTTP/1.1 200 OK

[
 {"id": 1234, "start": "14:00", "end": "14:50", "doctor": "mjones"},

 {"id": 5678, "start": "16:00", "end": "16:50", "doctor": "mjones"}
]
```

我仍然觉得14:00的时间比较合适，于是双进行预约确认，并提交了我的基本信息，用以创建预约，这是符合POST的语义的：

```
POST /schedules/1234 HTTP/1.1

{name: xx, age: 30,}
```

如果预约成功，那我能够收到一个预约成功的响应：

```
HTTP/1.1 201 Created

Successful confirmation of appointment
```

如果发生了问题，譬如有人在我前面抢先预约了，那么我会在响应中收到某种错误信息：

```
HTTP/1.1 409 Conflict
```

```
doctor not available
```

## 第3级

第2级是目前绝大多数系统所到达的REST级别，但仍不是不够完美的，至少还存在一个问题：你是如何知道预约mjones医生的档期是需要访问“/schedules/1234”这个服务Endpoint的？也许你甚至第一时间无法理解为何我会有这样的疑问，这当然是程序代码写的呀！但REST并不认同这种已烙在程序员脑海中许久的想法。RMM中的Hypermedia Controls、Fielding论文中的HATEOAS和现在提的比较多的“超文本驱动”，所希望的是除了第一个请求是有你在浏览器地址栏输入所驱动之外，其他的请求都应该能够自描述清楚后续可能发生的状态转移，由超文本自身来驱动。所以，当你输入了查询的指令之后：

```
GET /doctors/mjones/schedule?date=2020-03-04&status=open HTTP/1.1
```

服务器传回的响应信息应该包括诸如如何预约档期、如何了解医生信息等可能的后续操作：

```
HTTP/1.1 200 OK

{
 schedules: [
 {
 id: 1234, start:"14:00", end: "14:50", doctor: "mjones",
 links: [
 {rel: "comfirm schedule", href: "/schedules/1234"}
]
 },
 {
 id: 5678, start:"16:00", end: "16:50", doctor: "mjones",
 links: [
 {rel: "comfirm schedule", href: "/schedules/5678"}
]
 }
],
 links: [
 {rel: "doctor info", href: "/doctors/mjones/info"}
]
}
```

如果做到了第3级REST，那服务端的API和客户端也是完全解耦的，你要调整服务数量，或者同一个服务做API升级将会变得非常简单。

## 不足与争议

以下是笔者所见过的怀疑REST能否在实践中真正良好应用的争议问题，笔者将自己的观点总结如下：

- **面向资源的编程思想只适合做CRUD，不适合用来处理真正复杂的业务逻辑**

这是遇到最多的一个问题。HTTP的四个最基础的命令POST、GET、PUT和DELETE很容易让人直接联想到CRUD操作，以至于在脑海中自然产生了直接的对应。REST所能涵盖的范围当然远不止于此，不过要说POST、GET、PUT和DELETE对应于CRUD其实也没什么不对，但这个CRUD必须泛化去理解，它们涵盖了信息在客户端与服务端之间如何流动的几种主要方式，所有基于网络的操作逻辑，都可以对应到信息在服务端与客户端之间如何流动来理解，只是有的场景里比较直观，而另一些场景中可能比较抽象。面向资源的编程思想与另外两种主流编程思想只是抽象问题时所处的立场不同，只有选择问题，没有高下之分：

- 面向过程编程时，为什么要以算法和处理过程为中心，输入数据，输出结果？当然是为了符合计算机世界中主流的交互方式。
- 面向对象编程时，为什么要将数据和行为统一起来、封装成对象？当然是为了符合现实世界的主流的交互方式。
- 面向资源编程时，为什么要将数据（资源）作为抽象的主体，把行为看作是统一的接口？当然是为了符合网络世界的主流的交互方式。

- **REST与HTTP完全绑定，不适合应用于要求高性能传输的场景中**

我个人很大程度上赞同此观点，但并不认为这是REST的缺陷，锤子不能当扳手用并不是锤子的质量有问题。面向资源编程与协议无关，但是REST（特指Fielding论文中所定义的REST，而不是泛指面向资源的思想）的确依赖着HTTP协议的标准方法、状态码、协议头等各个方面。HTTP并不是传输层协议，它是应用层协议，如果仅将HTTP当作传输是不恰当的（SOAP：再次感觉有被冒犯到）。对于需要直接控制传输（如二进制细节/编码形式/报文格式/连接方式等）细节的场景中，REST确实不合适，这些场景往往

存在于服务集群的内部节点之间，这也是之前我曾提及的，REST和RPC尽管应用有所重合，但重合的范围其实并不大。

- **REST不利于事务支持**

这个问题首先要看你怎么看待“事务（ Transaction ）”这个概念。如果“事务”指的是数据库那种的狭义的刚性ACID事务，那分布式系统本身与此就是有矛盾的（ CAP不可兼得 ），这是分布式的问题而不是REST的问题。如果“事务”是指通过服务协议或架构，获得对多个分布式服务中数据提交进行统一协调的支持（ 2PC/3PC ）能力，譬如[WS-AtomicTransaction](#)、[WS-Coordination](#)这样的功能性协议，这REST确实不支持，假如你已经理解了这样做的代价，仍决定要这样做的话，SOAP是比较好的选择。如果“事务”是指希望保障数据的最终一致性，说明你已经放弃刚性事务了，这才是分布式系统中的主流，使用REST肯定不会有阻碍，谈不上“不利于”（当然，对此REST也并没有什么帮助，这完全取决于你系统的事务设计，在[事务一致性](#)中再详细讨论）

- **REST没有传输可靠性支持**

是的，并没有。在HTTP中你发送出去一个请求，通常会收到一个与之相对的响应，譬如HTTP/1.1 200 OK或者HTTP/1.1 404 Not Found诸如此类的。但如果你没有收到任何响应，那就无法确定消息到底是没有发送出去，抑或是没有从服务端返回回来，这其中的关键差别是服务端到底是否被触发了某些处理？应对传输可靠性最简单粗暴的做法是把消息再重发一遍。这种简单处理能够成立的前提是服务应具有[幂等性](#)（ Idempotency ），即服务被重复执行多次的效果与执行一次是相等的。HTTP协议要求GET、PUT和DELETE应具有幂等性，我们把REST服务映射到这些方法时，也应当保证幂等性。对于POST方法，曾经有过一些专门的提案（如[POE](#)，POST Once Exactly），但并未得到IETF的通过。对于POST的重复提交，浏览器会出现相应警告，如Chrome中“确认重新提交表单”的提示，对于服务端，就应该做预校验，如果发现可能重复，返回HTTP/1.1 425 Too Early。另，SOAP中有[WS-ReliableMessaging](#)功能协议用于支持消息可靠投递。

- **REST缺乏对资源进行“部分”和“批量”的处理能力**

这个观点我是认同的，我甚至认为这将是未来面向资源的思想和API设计风格的发展方向。REST开创了面向资源的服务风格，却肯定仍并不完美。以HTTP协议为基础给REST带来了极大的便捷（不需要额外协议，不需要重复解决一堆基础网络问题，等等），但也是HTTP本身成了束缚REST的无形牢笼。我仍通过具体例子来解释REST这方面的局限性：譬如你仅仅想获得某个用户的姓名，RPC风格中可以设计一个“getUsernameB

yld”的服务，返回一个字符串，尽管这种服务的通用性实在称不上“设计”二字，但确实可以工作；而REST风格中你将向服务端请求整个用户对象，然后丢弃掉返回的结果中该用户的其他属性，这便是一种Overfetching。REST的应对手段是通过位于中间节点或客户端缓存来缓解这种问题，但此缺陷的本质是由于HTTP协议完全没有对请求资源的结构化描述能力（但有非结构化的部分内容获取能力，即今天多用于端点续传的[Range Header](#)），所以返回资源的哪些内容、以什么数据类型返回等等，都不可能得到协议层面的支持，要做你就只能自己在GET方法的Endpoint上设计各种参数来实现。而另外一方面于此相对的缺陷是对资源的批量操作的支持，有时候我们不得不为此而专门设计一些抽象的资源才能应对。譬如你准备把某个用户的名字增加一个“VIP”前缀，提交一个PUT请求修改这个用户的名称即可，而你要给1000个用户加VIP时，就不得不先创建一个（名为“VIP-Modify-Task”）任务资源，把1000个用户的ID交给这个任务，最后驱动任务进入执行状态（你可以去调用1000次PUT的，[HTTP Status 429](#)与你不见不散）。又譬如你去网店买东西，下单、冻结库存、支付、加积分、扣减库存这一系列步骤会涉及到多个资源的变化，你可能面临不得不创建一种“事务”的抽象资源，或者用某种具体的资源（譬如“结算单”）贯穿这个过程的始终，每次操作其他资源时都带着事务或者结算单的ID。HTTP协议由于本身的无状态性，会相对不适应（并非不能够）处理这类业务场景。

解决这类问题，目前看起来一种理论上较优秀的解决方案是[GraphQL](#)，这是由Facebook提出并开源的一种面向资源API的数据查询语言（如同SQL一样，挂了个“查询语言”的名字，但CRUD都做）。比起依赖HTTP无协议的REST，GraphQL可以说是另一种“有协议”的、更彻底的面向资源的服务方式。然而凡事都有两面，离开了HTTP，它又面临着几乎所有RPC框架所遇到的那个如何推广这种交互接口的问题。

# 异步服务调用

# 事务处理

事务处理几乎是每一个信息系统中都会涉及到的问题，它存在的意义就是为了保证系统中数据是正确的，不同数据间不会产生矛盾（一致性，Consistency，请注意“一致性”在数据科学中是有严肃定义、且有多种细分类型的概念，后续介绍分布式共识算法时讨论的一致性与数据库状态的一致性严格来说并不能直接等同）。理论上，达成这个目标需要三方面共同努力来保障：

- **原子性（Atomic）**：在同一项业务处理过程中，事务保证了多个对数据的修改，要么同时成功，要么一起被撤销。
- **隔离性（Isolation）**：在不同的业务处理过程中，事务保证了各自业务正在读、写的数据互相独立，不会彼此影响。
- **持久性（Durability）**：事务应当保证所有成功被提交的数据修改都能够正确地被持久化，不丢失数据。

以上即事务的“ACID”的概念提法，笔者自己对这种已经形成习惯的“ACID”的提法是不太认同的，上述四种特性并不正交，A、I、D是手段，C是目的，完全是为了拼凑个单词缩写才弄到一块去，误导的弊端已经超过了易于传播的好处。

事务的概念最初是源于数据库，但今天的信息系统中已经不再局限于数据库本身，所有需要保证数据正确性（一致性）的场景中，包括但不限于数据库、缓存、[事务内存](#)、消息、队列、对象文件存储，等等，都有可能会涉及到事务处理。当一个服务只操作一个数据源时，通过A、I、D来获得一致性是相对容易的，但当一个服务涉及到多个不同的数据源，甚至多个不同服务同时涉及到多个不同的数据源时，这件事情就变得很困难，有时需要付出很大乃至是不切实际的代价，因此业界探索过许多其他方案，在确保可操作的前提下获得尽可能高的一致性保障，事务处理由此才从一个具体操作上的“编程问题”上升成一个需要仔细权衡的“架构问题”。

人们在探索这些事务方案的过程中，产生了许多新的思路和概念，有一些概念看上去并不那么直观，在本章里，笔者会通过同一个具体事例在不同的事务方案中如何处理来贯穿、理顺这些概念。

## 场景事例

Fenix's Bookstore是一个在线书店。当一份商品成功售出时，需要确保以下三件事情被正确地处理：

- 用户的账号扣减相应的商品款项
- 商品仓库中扣减库存，将商品标识为待配送状态
- 商家的账号增加相应的商品款项

接下来，笔者将逐一介绍在“单个服务使用单个数据源”、“单个服务使用多个数据源”、“多个服务使用单个数据源”以及“多个服务使用多个数据源”的不同场景下，我们可以采用哪些手段来保证以上场景实例的正确性。

# 本地事务

本地事务（Local Transactions，其实应该翻译成局部事务才好与稍后的全局事务相对应，但现在几乎所有人都这么叫了）即独立的、不需要“事务管理器（稍后解释这是啥）”进行协调的事务，这是最基础的一种事务处理方案，只能适用于单个服务使用单个数据源的场景。本地事务其实是直接依赖于数据源（通常是DBMS，下面均以JDBC为例）本身的事务能力来实现的，在服务层面，最多只能说是对事务接口做了一层薄包装而已，它对真正的事务的运作并不能产生多少影响。

为了解释“薄包装”和后续讨论方便，我们将事例场景进一步具体化：假设书店的用户、商家、仓库所涉及的数据表都存储于同一个数据库，它们的服务运行于同一个JVM实例之上，使用Spring来进行程序组织，所有服务的事务传播<sup>1</sup>都是默认的“需要事务”。按照现在主流的开发习惯，其代码大致应如下所示：

```
@Transactional
public void buyBook(PaymentBill bill) {
 userAccountService.pay(bill.getMoney());
 warehouseService.deliver(bill.getItems());
 businessAccountService.receipt(bill.getMoney());
}
```

java

我们将声明式事务手工还原回编程式事务：

```
public void buyBook(PaymentBill bill) {
 transaction.begin();
 try {
 userAccountService.pay(bill.getMoney());
 warehouseService.deliver(bill.getItems());
 businessAccountService.receipt(bill.getMoney());
 transaction.commit();
 } catch(Exception e) {
 transaction.rollback();
 }
}
```

java

代码的语义非常直白，但却不一定如字面所示那般严谨，看起来如果操作不出错，肯定会在commit()中提交事务，如果出错了，肯定会在rollback()中回滚事务。但并非绝对如此，譬如其中数据表采用引擎的是MyISAM，那即使调用了rollback()方法也无法回滚数据，原子性就无法得到保障。同理，对于隔离性，尽管Spring可以将用户所期望的隔离级别传递给数据库，但是具体数据库会不会按照所设置的参数调整隔离级别，如何进行事务隔离，Spring也是完全无法知晓且无法改变的。因此，本地事务具体能够提供怎样的能力，其实取决于底层的数据库本身。

# 全局事务

与本地事务相对的是全局事务（Global Transactions），有一些资料中也将其称为外部事务（External Transactions），在本文中，全局事务被限定为一种适用于单个服务使用多个数据源场景的事务解决方案。请注意，理论上真正的全局事务并没有“单个服务”的约束，它本来就是DTP（[Distributed Transaction Processing](#)）模型中的概念，但本节所讨论的内容——一种在分布式环境中仍追求强一致性的事务处理方案，对于多节点而且互相调用彼此服务的场合（典型的就是现在的微服务）中是极不合适的，今天它几乎只实际应用于单服务多数据源的场合中，为了避免与后续介绍的放弃了ACID的弱一致性事务处理方式相互混淆，所以这里的全局事务所指范围有所缩减，后续涉及多服务多数据源的事务，笔者将称其为“分布式事务”。

1991年，为了解决分布式事务的一致性问题，[X/Open](#)组织（后来并入了[TOG](#)）提出了一套名为[X/Open XA](#)（XA是eXtended Architecture的缩写）的处理事务架构，其核心内容是定义了全局的事务管理器（Transaction Manager，用于协调全局事务）和局部的资源管理器（Resource Manager，用于驱动本地事务）之间的通讯接口。XA接口是双向的，能在在一个事务管理器和多个资源管理器（Resource Manager）之间形成通信桥梁，通过协调多个数据源的一致动作，实现全局事务的统一提交或者统一回滚，现在我们在Java代码中还偶尔能看见的XADataSource、XAResource这些名字都源于此。

不过，XA并不是Java规范（那时候还没有Java），而是一套通用技术规范，所以Java中专门定义了[JSR 907 Java Transaction API](#)，基于XA模式在Java语言中的实现了一套全局事务处理的标准，这也就是我们现在所熟知的JTA。JTA最主要的两个接口是：

- 事务管理器的接口：javax.transaction.TransactionManager。这套接口是给Java EE服务器提供容器事务（由容器自动负责事务管理）使用的，还提供了另外一套javax.transaction.UserTransaction接口，用于通过程序代码手动开启、提交和回滚事务。
- 满足XA规范的资源定义接口：javax.transaction.xa.XAResource，任何资源（JDBC、JMS等等）如果需要支持JTA，只要实现XAResource接口中的方法即可。

JTA原本是Java EE中的技术，一般情况下应该由JBoss、WebSphere、WebLogic这些Java EE容器来提供支持，但现在Bitronix<sup>↗</sup>、Atomikos<sup>↗</sup>和JBossTM<sup>↗</sup>（以前叫Arjuna）都以JAR包的形式实现了JTA的接口，称为JOTM（Java Open Transaction Manager），使得我们能够在Tomcat、Jetty这样的Java SE环境下也能使用JTA。

现在，我们对示例场景做另外一种假设：如果书店的用户、商家、仓库分别处于不同的数据库中，其他条件仍与之前相同，那情况会发生什么变化？如果我们以声明式事务来编码，代码可能一个字都不会改变，但仍手工还原回编程式事务的话，其语义将如下所示：

```
public void buyBook(PaymentBill bill) {
 userTransaction.begin();
 warehouseTransaction.begin();
 businessTransaction.begin();
 try {
 userAccountService.pay(bill.getMoney());
 warehouseService.deliver(bill.getItems());
 businessAccountService.receipt(bill.getMoney());
 userTransaction.commit();
 warehouseTransaction.commit();
 businessTransaction.commit();
 } catch(Exception e) {
 userTransaction.rollback();
 warehouseTransaction.rollback();
 businessTransaction.rollback();
 }
}
```

java

事情是要做三次提交，但实际上代码是并不能这样写的，试想一下，如果在businessTransaction.commit()中出现错误，代码转到catch块中执行，此时userTransaction和warehouse Transaction已经完成提交，再调用rollback()方法也无济于事，导致一部分数据被提交，另一部分被回滚，整个事务的一致性也就无法保证了。为了解决这个问题，XA将事务提交拆分为两阶段过程：

- 准备阶段（又叫做投票阶段）：在这一阶段，协调者询问所有参与的是否准备好提交，参与者如果已经准备好提交则回复Prepared，否则回复Non-Prepared。所谓的准备操作，对于数据库来说，其逻辑是在重做日志（Redo Log）中记录全部事务提交操作所要做的内容，只是与本地事务提交时的区别是不写入最后一条commit命令而已，相当于在

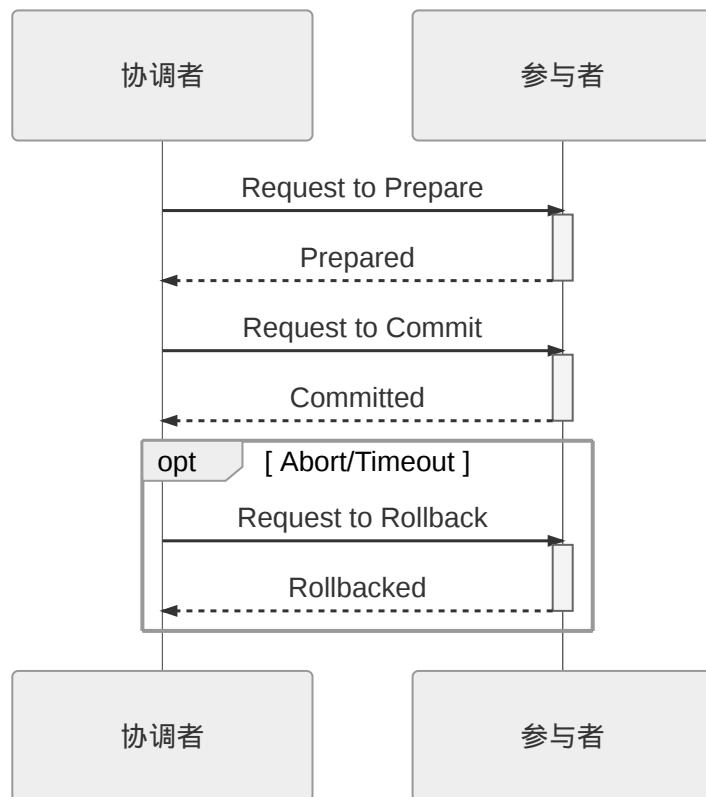
做完数据持久化后并不立即释放隔离性，即仍继续持有着锁和其他相关资源，维持数据对其他非事务内观察者的隔离状态。

- 提交阶段（又叫做执行阶段）：协调者如果在上一阶段收到所有参与者回复的Prepare d，则先自己在本地持久化事务状态为Commit，在此操作完成后向所有参与者发送Commit指令，所有参与者立即执行提交操作；否则，任意一个参与者回复Non-Prepared，或任意一个参与者超时未回复，协调者将在自己完成事务状态为Abort持久化后，向所有参与者发送Abort指令，参与者立即执行回滚操作。对于数据库来说，提交操作应是很轻量快速的，仅仅是持久化一条commit指令而已，只有收到Abort指令时，才需要清理已提交的数据，这可能是相对重负载操作。

以上这两个过程被称为“[两段式提交](#)”（2 Phase Commit，2PC）协议，而它能够成功保证一致性还要求有其他前提条件：

- 必须假设网络（在提交阶段的短时间内）是可靠的，XA的设计目标并不是解决诸如[拜占庭问题](#)的网络问题。两段式提交中投票阶段失败了可以补救（回滚），而提交阶段失败了无法补救（不再改变提交或回滚的结果，只能等失败的节点重新恢复），但此阶段耗时很短，这也是为了尽量控制网络风险的考虑。
- 必须假设因为网络、机器崩溃或者其他原因而导致失联的节点最终能够恢复，不会永久性地处于崩溃状态。由于在准备阶段已经写入了完整的重做日志，所以当失联机器一旦恢复，就能够从日志中找出已准备妥当但并未提交的事务数据，再而向协调者查询该事务的状态，确定下一步应该进行提交还是回滚操作。

请注意，上面所说的协调者、参与者通常都是数据库的角色，协调者一般是在参与者之间选举产生的，而应用服务器相对于数据库来说是客户端的角色。两段式提交的交互时序如下图所示：



两段式提交原理简单，易于实现，但其缺点也是显而易见的：

- **单点问题：**

协调者在两段提交中具有举足轻重的作用，协调者等待参与者回复时可以有超时机制，允许参与者宕机，但参与者等待协调者指令时无法做超时处理。一旦宕机的不是其中某个参与者，而是协调者的话，所有参与者都会受到影响，譬如，协调者没有正常发送Commit或者Rollback的指令，那所有参与者都将一直等待。

- **性能问题：**

两段提交过程中，所有参与者相当于被绑定成为一个统一调度的整体，期间要经过两次远程服务调用，三次数据持久化（准备阶段写重做日志，协调者做状态持久化，提交阶段在日志写入commit命令），整个过程将持续到参与者集群中最慢的那个处理操作结束为止，这决定了两段式提交对性能影响通常都会比较差。

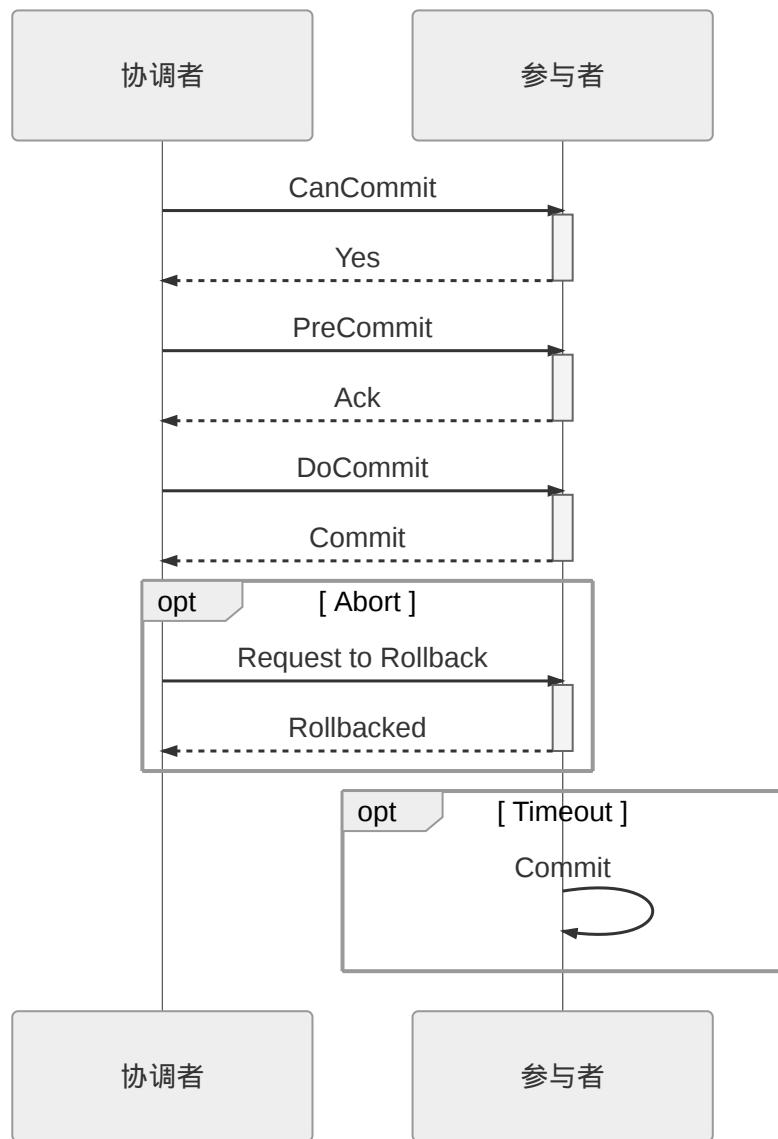
- **一致性风险：**

前面已经提到，两段式提交的成立是前提条件的，网络稳定性和宕机恢复能力的假设不成立时，仍可能出现一致性问题。宕机恢复能力这一点不必多谈，1985年Fischer、Lynch、Paterson提出了定理证明了如果宕机最后不能恢复，那就不存在任何一种分布式协议可以正确地达成一致性结果（被称为[FLP不可能原理](#)，以前可能不太著名，目前在区块链共识机制中用的挺多）。而网络稳定性带来的一致性风险是指：尽管提交阶段时间很短，但这仍是一段明确存在的危险期，如果协调者在发出准备指令后，根据收到各

个参与者发回的信息确定事务状态是可以提交的，协调者会先持久化事物状态，并提交自己的事务，如果这时候网络忽然被断开，无法再通过网络向参与者发出Commit指令的话，就会导致部分数据（协调者的）已提交，但部分数据（参与者的）还未提交（也没有回滚），产生了数据不一致的问题。

为了缓解两段式提交协议的头两点缺陷——即单点问题和性能问题，后续发展出了“**三段式提交**”（3 Phase Commit，3PC）协议。三段式提交把原本的两段式提交的准备阶段再细分为两个阶段，分别称为CanCommit、PreCommit，把的提交阶段称为DoCommit阶段。其中，新增的CanCommit是一个询问阶段，协调者让每个参与的数据库根据自身状态，评估该事务是否有可能顺利完成。将准备阶段一分为二的理由是这个阶段是重负载的操作，一旦协调者发出开始准备的消息，每个参与者都将马上开始写重做日志，它们所涉及的资源即被锁住，如果此时某一个参与者宣告无法完成提交，相当于大家都做了一轮无用功。所以，增加一轮询问阶段，如果都得到了正面的响应，那事务能够成功提交的把握就很大了，这意味着因某个参与者提交失败而导致大家全部回滚的风险变小。因此，在事务需要回滚的场景中，三段式的性能是要比两段式好很多的，但在事务能够正常提交的场景中，两者的性能都依然很差（三段式的多了一次询问，还要更差一些）

同样也是基于事务失败回滚概率变小的原因，三段式提交中，如果在PreCommit阶段之后发生了协调者宕机，参与者没有能等到DoCommit的消息的话，默认的操作策略将是提交事务（而不是回滚），这就相当于避免了协调者单点问题的风险。三段式提交的操作时序如下图所示。

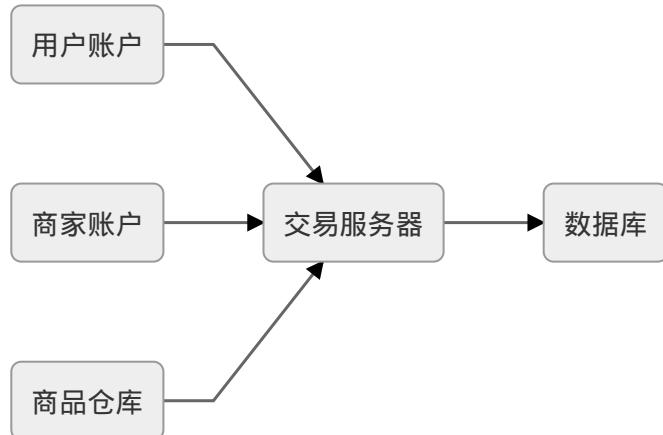


从以上过程可以看出，三段式提交对单点、和回滚时的性能问题有所改善，但是它对一致性问题并未有任何改进，在这方面它甚至反而是略有下降了的。譬如，进入PreCommit阶段之后，协调者发出的指令不是Ack而是Abort，而此时因网络问题，有部分参与者直至超时都未能收到协调者的Abort指令的话，这些参与者将会错误地提交事务，这就产生了数据一致性问题。

# 共享事务

与前面全局事务所指的单个服务使用多个数据源正好相反，共享事务（Share Transaction s）是指多个服务使用同一个数据源。请注意此语境里“数据源”与“数据库”的区别，我们在部署应用集群时一种典型模式是将同一套程序部署到多个中间件服务器的节点，它们连接了同一个数据库，但每个节点配有自己的专属的数据源JNDI，所有节点的数据访问都是完全独立的，并没有任何交集，此时所执行的是简单的本地事务。而本节讨论的是多个服务会产生交集的场景，譬如我们书店的事例中，如果用户账户、商家账户和商品仓库都存储于同一个数据库之中，但用户、商户和仓库每个领域都有独立的微服务，此时业务操作贯穿了三者，如果我们直接将不同数据源就视为是不同数据库，那上一节所讲的全局事务和下一节要讲的分布式事务都是可行的，不过，针对这种特例场景，共享事务则有机会成为另一条可能提高性能降低复杂度的途径（但更有可能是个伪需求）。

一种理论可行的方案是直接让各个服务共享数据库连接，同一个服务进程中的不同持久化工具（JDBC、ORM、JMS等）要共享数据库连接很容易，但由于数据库连接是与IP地址绑定的，字面意义上的“不同服务共享数据库连接”很难做到，所以这种方案里需要新增一个“交易服务器”的角色，无论是用户服务、商家服务还是仓库服务，它们都通过同一台交易服务器来与数据库打交道。如果你将交易服务器的对外接口实现为JDBC规范，那它完全可以视为是一个独立于各个服务的远程连接池或者数据库代理来看待，此时三个服务所发出的交易请求就有可能做到交由同一个数据库连接通过本地事务的方式完成。譬如，交易服务器根据传来的同一个事务ID，使用同一个数据库连接来处理不同服务的交易请求。



之所以强调理论上，是因为这是与实际生产系统中的压力方向相悖的，一个集群中数据库往往才是压力最大而又最不容易伸缩拓展的重灾区，所以现实中只有类似[ProxySQL](#)、[MaxScale](#)这样用于对多个数据库实例做负载均衡的代理（ProxySQL代理单个数据库，再启用Connection Multiplexing，其实已经挺接近于前面所提及的方案了），而几乎没有反过来代理一个数据库为多个应用提供事务协调的交易服务代理。这也是我说它更有可能是个伪需求的原因，连数据库都不拆分的话，你必须找到十分站得住脚的理由来向团队解释做微服务的意义是什么才行。

以上方案还有另外一种本质上是同样思路变种应用的形式：使用JMS服务器的来代替交易服务器，用户、商家、仓库的服务操作业务时，通过消息将所有对数据库的改动传送到JMS服务器，通过JMS来统一完成有事务保障的持久化操作。这被称作是“[单个数据库的消息驱动更新](#)”（Message-Driven Update of a Single Database）。“共享事务”的提法和这里所列的两种处理方式在实际应用中均不常见，笔者查询到的资料几乎都发源于十余年前的[这篇文章](#)，考虑到它并不契合于现在的技术趋势，我们也不花费更多的篇幅了。

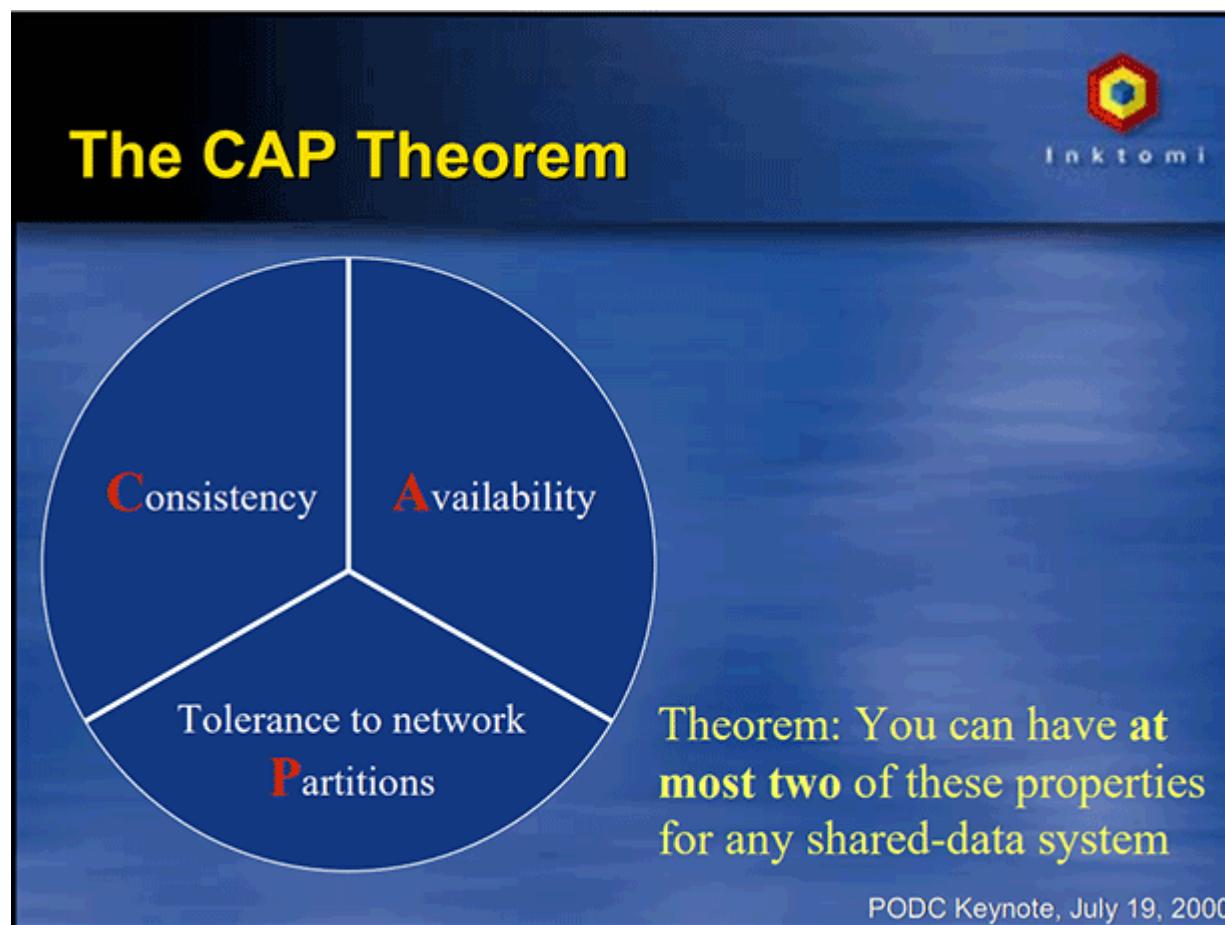
# 分布式事务

本节所说的分布式事务（Distributed Transactions）指的是多个服务同时访问多个数据源的事务处理机制，请注意它与[DTP模型](#)中“分布式事务”的差异，DTP模型所指的“分布式”是相对于数据源而言的，并不涉及服务，这部分内容已经在“全局事务”一节里进行过讨论。本节所指的“分布式”是相对于服务而言的，如果严谨地说，它更应该被称为“在分布式服务环境下的事务处理机制”。

曾经（在2000年以前），人们寄希望于XA的事务机制可以在本节所说的分布式环境中也能良好地应用，但这个美好的愿望今天已经被CAP理论彻底地击碎了，这节的话题就从CAP与ACID的矛盾说起。

## CAP与ACID

CAP理论，也被称为Brewer理论，是在2000年7月，加州大学伯克利分校的Eric Brewer教授于“ACM分布式计算原理研讨会（PODC）”上所提出的一个猜想：

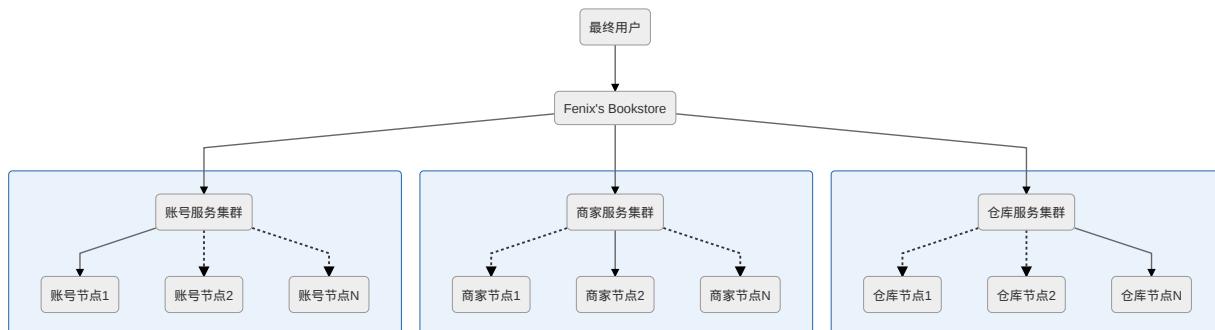


CAP理论原稿 (那时候还只是猜想)

2002年，麻省理工学院的Seth Gilbert和Nancy Lynch以严谨的数学推理上证明了CAP猜想。自此之后，CAP理论正式成为分布式计算领域所公认的著名定理。这个定理里描述了一个分布式的系统中，涉及到共享数据问题时，以下三个特性最多只能满足其二：

- **一致性 (Consistency)**：代表数据在任何时刻、任何分布式节点中所看到的都是没有矛盾的。这与前面所提的ACID中的C是相同的单词又有不同的定义（分别指Replication的一致性和数据库状态的一致性）。但分布式事务中，ACID的C要以满足CAP中的C为前提。
- **可用性 (Availability)**：代表系统**不间断地**提供服务的能力。
- **分区容忍性 (Partition Tolerance)**：代表分布式环境中部分节点因网络原因而彼此失联（即与其他节点形成“网络分区”）时，系统仍能**正确地**提供服务的能力。

单纯只列概念，CAP是比较抽象的，笔者仍以本章开头所列的事例场景来说明这三种特性对分布式系统来说将意味着什么。假设Bookstore的服务拓扑如下图所示，一个来自最终用户的交易请求，将交由账号、商家和仓库服务集群中某一个节点来完成响应：



在这套系统中，每一个单独的服务节点都有着自己的数据库，假设某次交易请求分别由“账号节点1”、“商家节点2”、“仓库节点N”来进行响应。当用户购买一件价值100元的商品后，账号节点1首先应给该用户账号扣减100元货款，它在自己数据库扣减100元很容易，但它还要把这次交易变动告知节点2到N，以及确保能正确变更商家和仓库集群其他账号节点中的关联数据，此时将面临以下情况：

- 如果该变动信息没有及时同步给其他账号节点，将导致有可能发生用户购买另一商品时，被分配给到另一个节点处理，由于看到账户上有不正确的余额而错误地发生了原本无法进行的交易，此为一致性问题。
- 如果由于要把该变动信息同步给其他账号节点，必须暂时停止对该用户的交易服务，直至数据同步一致后再重新恢复，将可能导致用户在下一次购买商品时，因系统暂时无法提供服务而被拒绝交易，此为可用性问题。
- 如果由于账号服务集群中某一部分节点，因出现网络问题，无法正常与另一部分节点交换账号变动信息，那此时服务集群中无论哪一部分节点对外提供的服务都可能是不正确的，能否接受由于部分节点之间的连接中断而影响整个集群的正确性，此为分区容忍性。

以上还仅是涉及到了账号服务集群自身的CAP问题，对于整个Bookstore站点来说，它更是面临着来自于账号、商家和仓库服务集群带来的CAP问题，譬如，用户账号扣款后，由于未及时通知仓库服务，导致另一次交易中看到仓库中有不正确的库存数据而发生超售。又譬如因涉及到仓库中某个商品的交易正在进行，为了同步用户、商家和仓库的交易变动，而暂时锁定该商品的交易服务，导致了的可用性问题，等等。

既然已有数学证明，我们就不去讨论为何CAP不可兼得，接下来直接分析如何权衡取舍CAP，以及不同取舍所带来的问题。

- 如果放弃分区容错性（CA without P），这意味着我们将假设节点之间通讯永远是可靠的。永远可靠的通讯在分布式系统中必定不成立的，这不是你想不想的问题，而是网络

分区现象始终会存在。在现实中，主流的RDBMS集群通常就是放弃分区容错性的工作模式，以Oracle的RAC集群为例，它的每一个节点均有自己的SGA、重做日志、回滚日志等，但各个节点是共享磁盘中的同一份数据文件和控制文件的，是通过共享磁盘的方式来避免网络分区的出现。

- 如果放弃可用性（CP without A），这意味着我们将假设一旦发生分区，节点之间的信息同步时间可以无限制地延长，此时问题相当于退化到前面“全局事务”中讨论的一个系统多个数据源的场景之中，我们可以通过2PC/3PC等手段，同时获得分区容错性和一致性。在现实中，除了DTP模型的分布式数据库事务外，著名的HBase也是属于CP系统，以它的集群为例，假如某个RegionServer宕机了，这个RegionServer持有的所有键值范围都将离线，直到数据恢复过程完成为止，这个时间通常会是很长的。
- 如果放弃一致性（AP without C），这意味着我们将假设一旦发生分区，节点之间所提供的数据可能不一致。AP系统目前是分布式系统设计的主流选择，因为P是分布式网络的天然属性，你不想要也无法丢弃；而A通常是建设分布式的目的，如果可用性随着节点数量增加反而降低的话，很多分布式系统可能就没有存在的价值了（除非银行这些涉及到金钱交易的服务，宁可中断也不能出错）。目前大多数NoSQL库和支持分布式的缓存都是AP系统，以Redis集群为例，如果某个Redis节点出现网络分区，那仍不妨碍每个节点以自己本地的数据对外提供服务，但这时有可能出现请求分配到不同节点时返回给客户端的是不同的数据。

行文至此，不知道你是否感受到一丝无奈，本章讨论的话题“事务”原本的目的就是获得“一致性”，而在分布式环境中，“一致性”却不得不成为了通常被牺牲、被放弃的那一项属性。但无论如何，我们建设信息系统，终究还是要保证操作结果（在最终被交付的时候）是正确的，为此，人们又重新给一致性下了定义，将前面我们在CAP、ACID中讨论的一致性称为“[强一致性](#)”（Strong Consistency），有时也称为“[线性一致性](#)”（Linearizability，通常是在讨论[共识算法](#)的场景中），而把牺牲了C的AP系统又要尽可能获得正确的结果的行为称为追求“弱一致性”，不过，如果单纯只说“弱一致性”那其实就是“不保证一致性”的意思……人类语言这东西真是博大精深。为此，在弱一致性中，人们又总结出了一种特例，被称为“[最终一致性](#)”（Eventual Consistency），它是指：如果数据在一段时间之内没有被另外的操作所更改，那它最终将会达到与强一致性过程相同的结果，有时候面向最终一致性的算法也被称为“[乐观复制算法](#)”。

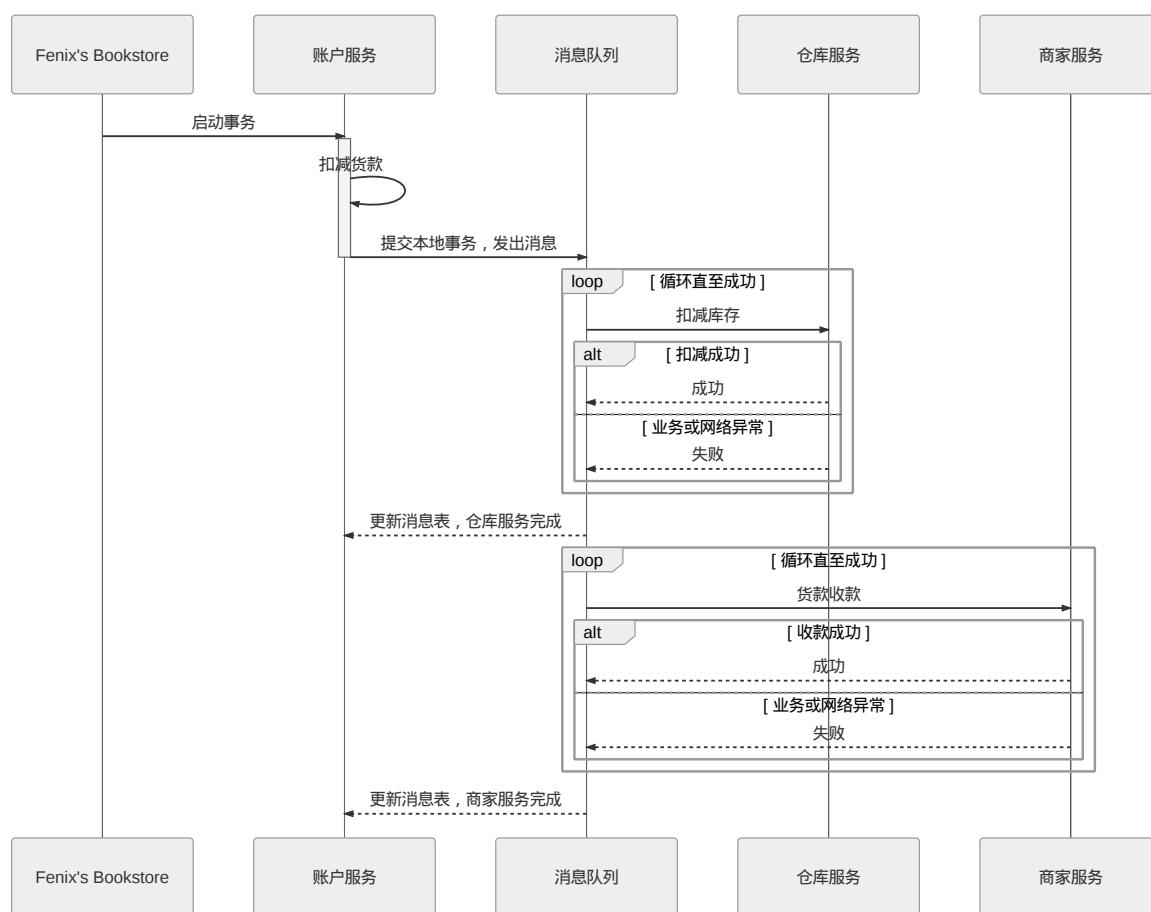
在本节讨论的主题“分布式事务”中，目标同样也不得不从前面的获得强一致性，降低为获得“最终一致性”，在这个意义上，其实“事务”一词的含义也已经被拓宽了，人们把之前追求

ACID的事务称为“刚性事务”，而把笔者下面将要介绍几种分布式事务的常见做法统称为“柔性事务”。

## 可靠事件队列

最终一致性的概念是eBay的系统架构师Dan Pritchett在2008年发表于ACM的论文《[Base: An Acid Alternative](#)》中提出的，该文中总结了另外一种独立于ACID获得的强一致性之外的、通过BASE来达成一致性目的的途径，最终一致性就是其中的“E”。BASE这提法比起ACID凑缩写的痕迹更重，不过有ACID vs BASE（酸 vs 碱）这个朗朗上口的梗，这篇文章传播得足够快，在这里笔者就不多谈BASE中的概念了，但这篇论文本身作为最终一致性的概念起源，并系统性地总结了一种在分布式事务的技术手段，还是非常有价值的。

我们继续以本章的事例场景来解释Dan Pritchett提出的“可靠事件队列”的具体做法，下图为操作时序：



1. 最终用户向Bookstore发送交易请求：购买一本价值100元的《深入理解Java虚拟机》。

2. Bookstore应该对用户账户扣款、商家账户收款、库存商品出库这三个操作有一个出错概率的先验评估，根据出错概率的大小来安排它们的操作顺序（这个一般体现在程序代码中，有一些大型系统也可能动态排序）。譬如，最有可能的出错的是用户购买了，但是不同意扣款，或者账户余额不足；其次是商品库存不足；最后商家收款，一般收款不会遇到什么意外。那顺序就应该是最容易出错的最先进行，即：账户扣款 → 仓库出库 → 商家收款。
3. 账户服务进行扣款业务，如扣款成功，则在自己的数据库建立一张消息表，里面存入一条消息：“事务ID：UUID，扣款：100元（状态：已完成），仓库出库《深入理解Java虚拟机》：1本（状态：进行中），某商家收款：100元（状态：进行中）”，注意，这个步骤中“扣款业务”和“写入消息”是依靠同一个本地事务写入自身数据库的。
4. 系统建立一个消息服务，定时轮询消息表，将状态是“进行中”的消息同时发送到库存和商家服务节点中去（可以串行地，即一个成功后再发送另一个，但在我门讨论的场景中没必要）。这时候可能产生以下几种可能的情况：
  1. 商家和仓库服务成功完成了收款和出库工作，向用户账户服务器返回执行结果，用户账户服务把消息状态从“进行中”更新为“已完成”。整个事务宣告顺利结束，达到最终一致性的状态。
  2. 商家或仓库服务有某个或全部因网络原因，未能收到来自用户账户服务的消息。此时，由于用户账户服务器中存储的消息状态一直处于“进行中”，所以消息服务器将在每次轮训的时候持续地向对应的服务重复发送消息。这个步骤可重复性决定了所有被消息服务器发送的消息都必须具备幂等性，通常的设计是让消息带上一个唯一的事务ID，以保证一个事务中的出库、收款动作只会被处理一次。
  3. 商家或仓库服务有某个或全部无法完成工作，譬如仓库发现《深入理解Java虚拟机》没有库存了，此时，仍然是持续自动重发消息，直至操作成功（譬如补充了库存），或者被人工介入为止。
  4. 商家和仓库服务成功完成了收款和出库工作，但回复的应答消息因网络原因丢失，此时，用户账户服务仍会重新发出下一条消息，但因消息幂等，所以不会导致重复出库和收款，只会导致商家、仓库服务器重新发送一条应答消息，此过程重复直至双方网络恢复。
  5. 也有一些支持分布式事务的消息框架，如RocketMQ，原生就支持分布式事务操作，这时候上述情况2、4也可以交由消息框架来保障。

以上这种靠着持续重试来保证可靠性的操作，在计算机中非常常见，它有个专门的名字叫做“[最大努力交付](#)”（Best-Effort Delivery），譬如TCP协议中的可靠性保障就属于最大努

力交付。而“可靠事件队列”有一种更普通的形式，被称为“最大努力一次提交”（ Best-Effort 1PC ），所指的就是将最有可能出错的业务以本地事务的方式完成后，通过不断重试的方式（不限于消息系统）来促使同个事务的其他关联业务完成，

## TCC事务

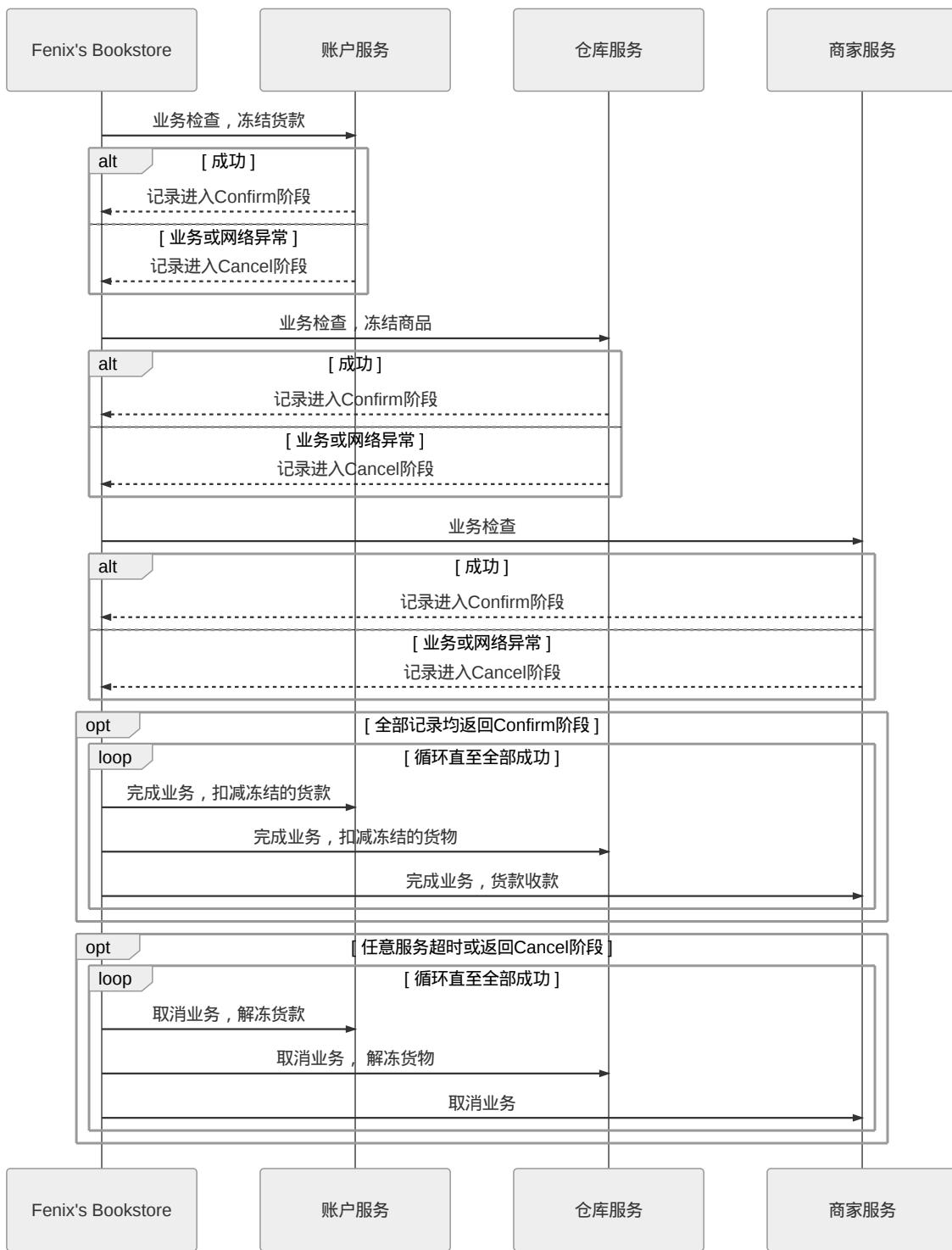
TCC是另一种常见的分布式事务机制，它是“Try-Confirm-Cancel”三个单词的缩写，是由数据库专家Pat Helland在2007年撰写的论文《Life beyond Distributed Transactions: an Apo state's Opinion》中提出。

前面介绍的可靠消息队列虽然能保证最终的结果是相对可靠的，过程也简单（相对于TCC来说），但整个过程完全没有任何隔离性可言，有一些业务中隔离性是无关紧要的，但有一些业务中缺乏隔离性就会带来许多麻烦。譬如我们的事例场景中，缺乏隔离性带来的一个显而易见的问题便是“超售”：完全有可能两个客户在短时间内都成功购买了同一件商品，而且他们各自购买的数量都不超过目前的库存，但他们购买的数量之和却超过了库存。如果这件事情处于刚性事务，且隔离级别足够的情况下是可以避免的，譬如，处理“第二类丢失更新的问题”（ Second Lost Update ）需要“可重复读”（ Repeatable Read ）的隔离剂别（这部分属于数据库基础常识，就不展开了），以保证后面提交的事务会因为无法获得锁而导致更新失败，但用可靠消息队列就无法做到这一点，这时候就可以考虑TCC方案了，它比较适合用于需要较强隔离性的分布式事务中。

TCC是一种业务侵入式较强的事务方案，它要求业务处理过程必须拆分为“预留业务资源”和“确认/释放消费资源”两个子过程。如同TCC的名字所示，它分为以下三个阶段：

- **Try**：尝试执行阶段，完成所有业务可执行性的检查（保障一致性），并且预留好全部需要用到的业务资源（保障隔离性）。
- **Confirm**：确认执行阶段，不进行任何业务检查，直接使用Try阶段准备的资源来完成业务处理。Confirm阶段可能会重复执行，需要满足幂等性。
- **Cancel**：取消执行阶段，释放Try阶段预留的业务资源。Cancel阶段可能会重复执行，需要满足幂等性。

按照我们的示例场景，TCC的执行过程应该是这样的：



1. 最终用户向Bookstore发送交易请求：购买一本价值100元的《深入理解Java虚拟机》。
2. 创建事务，生成事务ID，记录在活动日志中，进入Try阶段：
  - 用户服务：检查业务可行性，可行的话，将该用户的100元设置为“冻结”状态，通知下一步进入Confirm阶段；不可行的话，通知下一步进入Cancel阶段。
  - 仓库服务：检查业务可行性，可行的话，将该仓库的1本《深入理解Java虚拟机》设置为“冻结”状态，通知下一步进入Confirm阶段；不可行的话，通知下一步进入Cancel阶段。

- 商家服务：检查业务可行性，不需要冻结资源。
3. 如果第2步所有业务均反馈业务可行，将活动日志中的状态记录为Confirm，进入Confirm阶段：
- 用户服务：完成业务操作（扣减那被冻结的100元）
  - 仓库服务：完成业务操作（标记那1本冻结的书为出库状态，扣减相应库存）
  - 商家服务：完成业务操作（收款100元）
4. 第3步如果全部完成，事务宣告正常结束，如果第3步中任何一方出现异常（业务异常或者网络异常），将根据活动日志中的记录，重复执行该服务的Confirm操作（即最大努力交付）。
5. 如果第2步有任意一方反馈业务不可行，或任意一方超时，将活动日志的状态记录为Cancel，进入Cancel阶段：
- 用户服务：取消业务操作（释放被冻结的100元）
  - 仓库服务：取消业务操作（释放被冻结的1本书）
  - 商家服务：取消业务操作（大哭一场后安慰商家谋生不易）
6. 第5步如果全部完成，事务宣应回滚结束，如果第5步中任何一方出现异常（业务异常或者网络异常），将根据活动日志中的记录，重复执行该服务的Cancel操作（即最大努力交付）。

由上述操作过程可见，TCC其实有点类似于2PC的准备阶段和提交阶段，但TCC是位于用户代码层面，而不是基础设施层面，这为它的实现带来了一定的灵活性，可以根据需要设计资源锁定的粒度。同时，这也带来了更高的开发成本和业务侵入性（主要影响到可控性和更换事务实现方案的成本），所以，通常我们并不会裸编码来做TCC，而是基于某些分布式事务中间件（譬如阿里开源的Seata）基础之上完成。

## SAGA事务

TCC事务具有较强的隔离性，避免了“超售”的问题，而且其性能一般来说是本篇提及的几种柔性事务模式中最高的（只操作预留资源，几乎不会涉及到锁和资源的争用），但它仍不能满足所有的场景。TCC的最主要限制是它的业务侵入性很强，这里并不是说它需要开发编码配合所带来的工作量，而更多的是指它所要求的技术可控性上的约束。譬如，把我们的事例场景修改如下：由于中国网络支付日益盛行，现在用户和商家在书店系统中可以选择不在开设账号，至少不会强求一定要从银行充值到系统中才能进行消费，可以直接在购物时通过网络支付在银行账号中划转货款。这里面就给系统施加了限制，用户、商家的

账户在银行的话，其操作权限和数据结构就不可能再随心所欲的地设计，通常也就无法完成冻结款项、解冻、扣减这样的操作（银行一般不会配合你的操作）。所以TCC中第一步Try阶段往往就已经无法施行。这时候我们就可以考虑一下采用另外一种柔性事务方案：SAGA事务（SAGA在英文中是“长篇故事、长篇记叙、一长串事件”的意思）。

SAGA事务模式的历史很久，最早源于1987年普林斯顿大学的Hector Garcia-Molina和Kenneth Salem在ACM发表的一篇论文《SAGA<sup>1</sup>》（这就是论文的名字，居然这也能过审稿！）。文中提出了一种如何提升“长时间事务”（Long Lived Transaction）运作效率的方法，大致思路是把一个大事务分解为可以交错运行的一系列子事务集合。原本SAGA目的是为了避免大事务长时间锁定数据库的资源，后来发展成将一个分布式环境中的大事务分解为一系列本地事务的设计模式。SAGA由两部分操作组成：

- 每个分布式事务对数据的操作，分解为N个子事务，命名为 $T_1, T_2, \dots, T_i, \dots, T_n$ 。每个子事务都应该是或者能被视为是原子行为。如果分布式事务能够正常提交，其对数据的影响（最终一致性）应与连续按顺序成功提交 $T_i$ 等价。
- 为每一个子事务设计补偿动作，命名为 $C_1, C_2, \dots, C_i, \dots, C_n$ 。 $T_i$ 与 $C_i$ 满足以下条件：
  - $T_i$ 与 $C_i$ 都具备幂等性。
  - $T_i$ 与 $C_i$ 满足交换律（Commutative），即先执行 $T_i$ 还是先执行 $C_i$ ，其效果都是一样的。
  - $C_i$ 必须能成功提交，不考虑 $C_i$ 本身提交失败被回滚的情形，此时需要人工介入。

如果 $T_1$ 到 $T_n$ 均成功提交，那事务顺利完成，否则，要采取以下两种恢复策略之一：

- **正向恢复**（Forward Recovery）：如果 $T_i$ 事务提交失败，则一直对 $T_i$ 进行重试，直至成功为止（最大努力交付）。这种恢复方式不需要补偿，适用于事务最终都要成功的场景（譬如扣了款，就一定要给别人发货）。正向恢复的执行模式为： $T_1, T_2, \dots, T_i$ （失败）， $T_i$ （重试）， $\dots, T_n$ 。
- **反向恢复**（Backward Recovery）：如果 $T_i$ 事务提交失败，则一直执行 $C_i$ 对 $T_i$ 进行补偿，直至成功为止（最大努力交付）。这里要求 $C_i$ 必须（持续重试后）执行成功。反向恢复的执行模式为： $T_1, T_2, \dots, T_i$ （失败）， $C_i$ （补偿）， $\dots, T_2, T_1$ 。

与TCC相比，SAGA不需要为资源设计冻结状态和撤销冻结的操作，补偿操作往往要容易实现得多。譬如，前面提到的账户直接开设在银行的场景，从银行划转货款到Bookstore系

统中，这步是经由用户支付操作（扫码、U盾）来促使银行提供服务；如果后续业务操作失败，尽管我们无法要求银行撤销掉之前用户转账的操作，但是由Bookstore系统将货款转回到用户账上作为补偿措施确是完全可行的。

SAGA必须保证所有子事务都得以提交或者补偿，但SAGA系统本身也有可能会崩溃，所以它必须设计与数据库类似的日志机制（被称为SAGA Log）以保证系统恢复后可以追踪到子事务的执行情况，譬如执行至哪一步或者补偿至哪一步了。另外，尽管补偿操作通常比冻结/撤销容易实现，但保证正向、反向恢复过程的能严谨地进行也需要花费不少的工夫（譬如通过服务编排、可靠事件队列等方式完成），所以，SAGA事务通常也不会完全裸编码来实现，一般也是在事务中间件的基础上完成，前面提到的Seata同样支持SAGA模式。

基于数据补偿来代替回滚的思路，可以应用在其他事务方案上，这个笔者就不开独立小节，放在这里一起来解释。举个例子，譬如阿里的GTS（Global Transaction Service，Seata由GTS开源而来）所提出的“[AT事务模式](#)”就是这样的一种应用。

从整体上看是AT事务是参照了XA两段提交协议实现的，但针对XA 2PC的缺陷，即在准备阶段必须等待所有数据源都返回成功后，协调者才能统一发出Commit命令而导致的木桶效应（所有涉及到的锁和资源都需要等待到最慢的事务完成后才能统一释放），设计了针对性的解决方案。大致的做法是在业务数据提交时自动拦截所有SQL，将SQL对数据修改前、修改后的结果分别保存快照，生成行锁，通过本地事务一起提交到操作的数据源中（相当于记录了重做和回滚日志）。如果分布式事务成功提交，那后续清理每个数据源中对应日志数据即可；如果分布式事务需要回滚，就根据日志数据自动产生用于补偿的“逆向SQL”。基于这种补偿方式，分布式事务中所涉及的每一个数据源都可以单独提交，然后立刻释放锁和资源。这种异步提交的模式，相比起2PC极大地提升了系统的吞吐量水平。而其代价就是大幅度地牺牲了隔离性，在缺乏隔离性的前提下，以补偿代替回滚并不一定是总能成功的。譬如，当本地事务提交之后、分布式事务完成之前，该数据被补偿之前又被其他操作修改过，即出现了脏写（Dirty Write），这时候一旦出现分布式事务需要回滚，就不可能再通过自动的逆向SQL来实现补偿，只能由人工介入处理了。

通常来说，脏写是一定要避免的（几乎所有DBMS在最低的隔离级别上都仍然要加锁以避免脏写），实际上这种情况人工也很难进行有效处理。所以GTS增加了一个“全局锁”（Global Lock）的机制来实现写隔离，要求本地事务提交之前，一定要先拿到针对修改记录的全局锁后才允许提交，没有获得全局锁之前就必须一直等待，这避免了有两个分布式事务

中包含的本地事务修改了同一个数据，从而避免脏写。在读隔离方面，AT事务默认的隔离级别是Read Uncommitted，这意味着可能产生脏读（Dirty Read）。读隔离也可以采用全局锁的方案解决，但直接阻塞读取的话，代价就非常大了，通常并不会这样做。由此可见，分布式事务中没有一揽子包治百病的解决办法，因地制宜地选用合适的事务处理方案才是唯一有效的做法。

# 透明多级分流系统

用户使用信息系统的过程中，请求从浏览器出发，通过网络，触及存储到最后端的数据库服务器中的信息，然后再返回到用户的浏览器，这其中要经过许许多多的技术基础设施。作为系统的设计者，我们应该意识到：不同的设施、部件在系统中有各自不同的价值。它们有一些位于网络的边缘，能够迅速响应用户的请求，避免给后端网络带来压力；有一些易于伸缩拓展，可以使用较小的代价，譬如堆叠机器来获得与用户数量相匹配的处理能力；有一些时刻保持着主从热备，为系统容灾容错，维护着高可用性；但也有一些设施是难以扩展的单点部件，只能依靠堆砌机器本身的性能来提升处理能力，典型的单点部件是传统RDBMS，在事务处理的CAP部分中，我们曾讨论过传统数据库为了同时具备可用性和一致性，放弃了分区容错性。

在进行系统设计时，我们应该充分理解这些部件的价值差异，一个普适的原则是尽可能减少单点部件，有一些单点是无可避免的，则应尽最大限度减少到达单点部件的流量。举个例子，许多的用户请求（如获取一张图片）在系统中往往会有多个部件能够处理（如浏览器缓存、CDN、反向代理、Web服务器、文件服务器、数据库都有可能提供这张图片），而恰如其分地将请求分流至最合适的组件中，避免所有流量都汇集到单点（如数据库），同时仍能够（在绝大多数时候）保证处理结果的准确性，仍能在单点系统出现故障时自动而迅速地实施补救措施，这便架构设计中多级分流的原则。缓存、节流、主备、负载均衡等这类措施，都是为了达成该原则所采用的工具与手段，而高可用架构、高并发架构则是通过该原则达成的目标。

一个现代的企业或互联网系统，其中所涉及到的分流手段数量之多、场景之广，可能连它的开发者本身都未必能全部意识到程度。这听起来似乎并不合理，但笔者认为这恰好是优秀架构设计的一种体现，分布广阔谓之“多级”，意识不到谓之“透明”，也就是本章我们要讨论的话题“透明多级分流系统”（Transparent Multi-Level Diversion System）的来由。笔者将信息系统中我们可能使用到的分流手段，按从前（用户端）到后（服务端）的顺序列举如下，稍后将逐一讨论：

- **客户端缓存**（Client Cache）：HTTP协议的无状态性决定了它必须依靠客户端缓存来解决网络传输效率上的缺陷。

- **域名解析** ( DNS Lookup ) : DNS也许是全世界最大、使用最频繁的信息查询系统，如果没有适当的分流机制，DNS将会成为整个网络的瓶颈。
- **链路优化** ( Transmission Optimization ) : 今天的链路优化原则，在若干年后的未来再回头看它们时，其中多数已经成了奇技淫巧，有些甚至成了反模式。
- **内容分发网络** ( Content Distribution Network ) : CDN是一种十分古老而又十分透明，没什么存在感的分流系统，许多人都说听过它，但真正了解过它的人却很少。
- **负载均衡** ( Load Balancing ) : 调度后方的多台机器，以统一的接口对外提供服务，承担此职责的技术组件被称为“负载均衡”。
- **缓存中间件** ( Cache Middleware ) 编写中 : 讨论数据缓存、方法缓存、进程内/外、集中式/分布式缓存等等。
- **数据库扩展** ( Database Expansion ) 编写中 : (传统)数据库必须保证一致性与高可用，它与分布式天生就存在矛盾，我们要想一些别的办法来提升它的可扩展性。

# 客户端缓存

## 客户端缓存 (Client Cache)

HTTP协议的无状态性决定了它必须依靠客户端缓存来解决网络传输效率上的缺陷。

浏览器的缓存机制几乎是在万维网刚刚出现就已经存在，在HTTP协议设计之初，便确定了服务端与客户端之间“无状态”（Stateless）的交互原则，即要求每次请求是独立的，每次请求无法感知和依赖另一个请求的存在，这既简化了HTTP服务器的设计，也为其水平扩展能力留下了广袤的空间。但无状态并不只有好的一面，由于每次请求都是独立的，服务端不保存此前请求的状态和资源，所以也不可避免地导致其携带有重复的数据，造成网络性能降低。HTTP协议对此的解决方案就是客户端缓存，在HTTP从1.0到最新2.0版本的每次演进中，都提出过现在被称为“状态缓存”、“强制缓存”（许多资料中简称为“强缓存”）和“协商缓存”的缓存机制。

其中，状态缓存是指不经过服务器，客户端直接根据缓存信息对目标网站的状态判断，以前只有301/Moved Permanently（永久重定向）这一个；后来在[RFC6797](#)中增加了HSTS（HTTP Strict Transport Security）机制，用于避免依赖301/302跳转HTTPS时可能产生的降级中间人劫持（详细可见安全架构中的“[传输](#)”），这也属于另一种状态缓存。由于状态缓存所涉内容就只有这一点，后续我们就只聚焦于强制缓存与协商缓存两种机制。

## 强制缓存

只要是缓存，几乎都不可避免地会遇到一致性的问题。强制缓存对一致性处理就如它的名字一样，显得十分的直接粗暴，假设在某个时间点（譬如10分钟）之内，资源的内容和状态一定不会被改变，因此客户端可以无需经过任何浏览器请求，在该时间点来临前一直持有和使用该资源的本地缓存副本。

根据约定，强制缓存在用户在浏览器输入地址、页面链接跳转、新开窗口、前进/后退中均可生效，但在使用F5刷新页面时应当失效。有以下两类HTTP Header可以实现强缓存：

- **Expires**：Expires是HTTP/1.0协议中提供的Header（当然，在HTTP/1.1中同样存在），后面跟随一个截至时间参数。当服务器返回某个资源时带有该Header的话，意味着服务器承诺截止时间之前资源不会发生变动，浏览器可直接缓存该数据，不再重新发请求，示例：

```
HTTP/1.1 200 OK
Expires: Wed, 8 Apr 2020 07:28:00 GMT
```

Expires是HTTP协议最初版本的缓存机制，设计非常直观易懂，但考虑得并不够周全，它至少存在以下显而易见的问题：

- 受限于客户端的本地时间。譬如，客户端修改了本地时间，可能会造成缓存提前失效或超期持有。
- 无法处理涉及到用户身份的私有资源，譬如，某些资源被登录用户缓存在自己的浏览器上是合理的，但如果被CDN服务器缓存起来，则可能被其他未认证的用户所获取。
- 无法描述“不缓存”的语义。譬如，浏览器为了提高性能，往往会自动在当次会话中缓存某些MIME类型的资源，在HTTP/1.0的服务器中就缺乏手段强制浏览器不允许缓存某个资源。以前为了实现这类功能，通常不得不使用Script脚本，在资源后面增加时间戳（如“xx.js?t=1586359920”）来保证每次资源都会重新获取。

关于“不缓存”的语义，在HTTP/1.0中其实设计了“Pragma: no-cache”来实现，但Pragma在HTTP响应中的行为没有确切描述，随后就被HTTP/1.1中出现过的Cache-Control所替代，现在，尽管主流浏览器通常都会支持Pragma，但实际并没有什么使用价值了。

- **Cache-Control**：Cache-Control是HTTP/1.1协议中定义的强制缓存Header，它的语义比起Expires来说就丰富了很多，如果Cache-Control和Expires同时存在，并且语义存在冲突（Expires与max-age / s-maxage冲突）的话，必须以Cache-Control为准。Cache-Control的示例如下：

```
HTTP/1.1 200 OK
Cache-Control: max-age=600
```

Cache-Control在客户端的请求头或服务器的响应头中都可以使用，它定义了一系列的参数，且允许扩展（不在标准RFC协议中，由浏览器自行支持），其标准的参数主要包括有：

- max-age / s-maxage：max-age后面跟随一个以秒为单位的数字，表明相对于请求时间（Date Header中也会注明请求时间），多少秒以内缓存是有效的，资源不需要重新从服务器中获取。相对时间避免了Expires中采用的绝对时间可能受客户端时钟影响的尴尬。s-maxage中的s是“Share”的缩写，意味“共享缓存”（即被CDN、代理等持有的缓存）有效时间，用于提示CDN这类服务器如何对缓存进行失效。
- public / private：指明是否涉及到用户身份的私有资源，如果是public，着可以被代理、CDN等缓存，如果是private，着只能由客户端进行私有缓存。
- no-cache / no-store：no-cache指明该资源不应该被缓存，哪怕是同一个会话中对同一个URL地址的请求，也必须从服务端获取（但协商缓存机制依然是生效的）；no-store不强制会话中相同URL资源的重复获取，但禁止浏览器、CDN等以任何形式保存该资源。
- no-transform：禁止资源被任何形式地修改。譬如，某些CDN、透明代理支持自动GZIP压缩图片或文本，以提升网络性能，而no-transform就禁止了这样的行为，它要求Content-Encoding、Content-Range、Content-Type均不允许进行任何形式的修改。
- min-fresh / only-if-cached：这两个参数是仅用于客户端的请求Header。min-fresh后跟一个以秒为单位的数字，用于建议服务器能返回一个不少于该时间的缓存资源（即包含max-age且不少于min-fresh的数字）。only-if-cached表示要求客户端要求不发送网络请求，只使用缓存来进行响应，若缓存不能命中，就直接返回503/Service Unavailable错误。
- must-revalidate / proxy-revalidate：must-revalidate表示在资源过期后，一定需要从服务器中进行验证（即超过了max-age的时间，就等同于no-cache的行为），proxy-revalidate用于提示代理、CDN等缓存服务，语义与must-revalidate一致。

## 协商缓存

强制缓存是基于时效性的，但无论是人还是服务器，其实多数情况下都并没有什么把握去承诺某项资源多久不会发生变化。另外一种基于变化检测的缓存机制，在一致性上会有比强制缓存更好的表现，但需要一次变化检测的交互开销，性能上就会略差一些，这种基于检测的缓存机制，通常被称为“协商缓存”。另外，应注意在HTTP中协商缓存与强制缓存并

没有排他性，这两套机制是并行工作的，譬如，当强制缓存存在时，直接从强制缓存中返回资源，无需进行变动检查；而当强制缓存超过时效，或者被禁止（no-cache / must-revalidate），协商缓存仍可以正常地工作。协商缓存主要有根据资源的修改时间或根据资源唯一标识是否发生变化来进行变动检查的机制，这都是靠一组成对出现的请求、响应Header来实现的：

- **Last-Modified和If-Modified-Since**：Last-Modified是服务器的响应Header，用于告诉客户端这个资源的最后修改时间。对于带有这个Header的资源，当客户端需要在此请求时，会通过If-Modified-Since把之前收到的资源最后修改时间发送回服务端。如果此时服务端发现资源在该时间后没有被修改过，就只要返回一个304/Not Modified的响应即可，无需附带消息体，如下所示：

```
HTTP/1.1 304 Not Modified
Cache-Control: public, max-age=600
Last-Modified: Wed, 8 Apr 2020 15:31:30 GMT
```

如果此时服务端发现资源在该时间之后有变动，就会返回200/OK的完整响应，在消息体中包含最新的资源，如下所示：

```
HTTP/1.1 200 OK
Cache-Control: public, max-age=600
Last-Modified: Wed, 8 Apr 2020 15:31:30 GMT

Content
```

- **Etag和If-None-Match**：Etag是服务器的响应Header，用于告诉客户端这个资源的唯一标识（HTTP服务器可以根据自己的意愿来选择如何生成这个标识，譬如Apache服务器的Etag值，默认是对文件的索引节点（INode），大小（Size）和最后修改时间（MTime）进行哈希计算后得到的），对于带有这个Header的资源，当客户端需要在此请求时，会通过If-None-Match把之前收到的资源唯一标识发送回服务端。如果此时服务端计算后发现资源的唯一标识与上传回来的一致，说明资源没有被修改过，就只要返回一个304/Not Modified的响应即可，无需附带消息体，如下所示：

```
HTTP/1.1 304 Not Modified
Cache-Control: public, max-age=600
```

```
Last-Modified: Wed, 8 Apr 2020 15:31:30 GMT
```

如果此时服务端发现资源的唯一标识有变动，就会返回200/OK的完整响应，在消息体中包含最新的资源，如下所示：

```
HTTP/1.1 200 OK
Cache-Control: public, max-age=600
Last-Modified: Wed, 8 Apr 2020 15:31:30 GMT
```

```
Content
```

Etag是HTTP中一致性最强的缓存机制，譬如，Last-Modified标注的最后修改只能精确到秒级，如果某些文件在1秒钟以内，被修改多次的话，它将不能准确标注文件的修改时间；又或者如果某些文件会被定期生成，可能内容并没有任何变化，但Last-Modified却改变了，导致文件无法有效使用缓存，这些情况Last-Modified都有可能产生一致性问题，只能使用Etag解决。

Etag却又是HTTP中性能最差的缓存机制，体现在每次请求时，服务端都必须对资源进行哈希计算，这比起简单获取一下修改时间，开销要大了很多。Etag和Last-Modified是允许一起使用的，服务器会优先验证Etag，在Etag一致的情况下，再去对比Last-Modified，这是为了防止有一些HTTP服务器未将文件修改日期纳入哈希范围内。

到这里为止，HTTP的协商缓存机制已经能很好地处理通过URL获取单个资源的场景，“单个资源”是什么意思？在HTTP协议的设计中，一个URL地址有可能能够提供多份不同版本的资源，譬如，一段文字的不同语言版本，一个文件的不同编码格式版本，一份数据的不同压缩方式版本，等等。HTTP协议设计了Accept\*（Accept、Accept-Language、Accept-Charset、Accept-Encoding）的一套请求Header和对应的Content-\*（Content-Language、Content-Type、Content-Encoding）的响应Header，这被称为HTTP的内容协商机制。与之对应的，对于一个URL能够获取多个资源的场景中，缓存也同样也需要有明确的标识来获知根据什么内容来对同一个URL返回给用户正确的资源。这个就是Vary Header的作用，Vary后面可以跟随其他Header的名字，譬如：

```
HTTP/1.1 200 OK
Vary: Accept, User-Agent
```

以上说明应该根据MINE类型和浏览器类型来缓存资源，获取资源时也需要根据请求头中对应的字段来筛选出适合的资源版本。

根据约定，协商缓存不仅在用户在浏览器输入地址、页面链接跳转、新开窗口、前进/后退中生效，而且在使用F5刷新页面时也同样是生效的，只有用户强制刷新（Ctrl+F5）或者禁用缓存（譬如在DevTools中设定）时才会失效，此时客户端向服务端发出的请求会自动带有“Cache-Control: no-cache”。

# 域名解析

## 域名缓存 (DNS Lookup)

DNS也许是全世界最大、使用最频繁的信息查询系统，如果没有适当的分流机制，DNS将会成为整个网络的瓶颈。

我们都知道DNS的作用是将便于人类理解的域名地址转换为便于计算机处理的IP地址，也许你会觉得好笑：笔者在接触计算机网络的开头一段不短的时间里面，都把DNS想像成一个部署在全世界某个神秘机房中的大型电话本式的翻译服务。后来，当笔者第一次了解到DNS的工作原理，并知世界根域名服务器的ZONE文件只有2MB大小（甚至可以打印出来物理备份）的时候，对DNS系统的设计是非常惊讶的。域名解析这个话题同样涉及缓存等因素，虽然它并不算本篇讨论的重点，但其本身就是堪称示范性的透明多级分流系统，很值得我们借鉴。

假设我们访问域名：[www.icyfenix.com.cn](http://www.icyfenix.com.cn)，DNS并不是一次性地将“[www.icyfenix.com.cn](http://www.icyfenix.com.cn)”解析成IP地址的，这需要经历一个递归解析的过程。首先DNS会将域名还原为“[www.icyfenix.com.cn.](http://www.icyfenix.com.cn.)”，注意最后多了一个点“.”，它是“.root”的含义（早期的域名必须带有这个点DNS才能够正确解析，如今DNS服务器已经可以自动补上结尾的点号），然后开始如下过程：

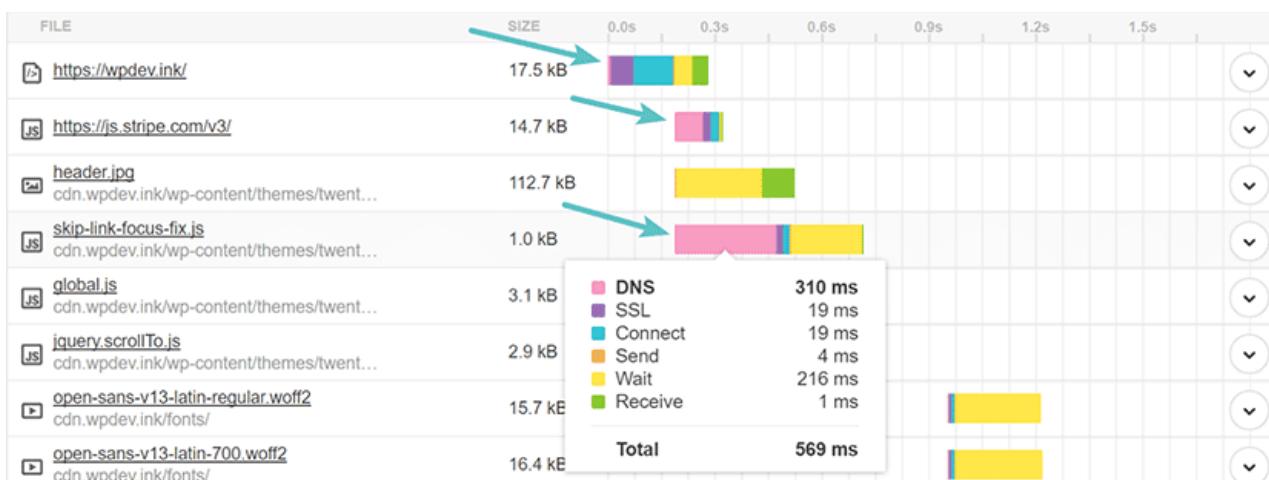
1. 客户端检查本地DNS缓存，查看是否存在并存活着的该域名的地址记录，DNS是以存活时间（Time to Live，TTL）来衡量缓存的存活情况的。后续每一级DNS查询的过程都会有类似的缓存查询操作，将不再重复叙述。
2. 客户端将地址发送给本机系统中设置的本地DNS（Local DNS，这个服务器可以通过手工设置，在路由做DHCP分配时或者在拨号时从PPP服务器中也会自动获取到）。
3. 本地DNS收到查询后，会按照“是否有[www.icyfenix.com.cn](http://www.icyfenix.com.cn)的权威服务器”→“是否有[www.icyfenix.com](http://www.icyfenix.com)的权威服务器”→“是否有[com.cn](http://com.cn)的权威服务器”→“是否有[cn](http://cn)的权威服务器”的顺序，查询自己的地址记录，如果都没有查询到，就会一直找到最后点号代表的根域名服务器为止。这里涉及了两个名词：

- **权威域名服务器** ( Authoritative DNS ) : 是指负责翻译指定域名的DNS服务器 , “权威”意味着指定域名应该翻译出怎样的结果是由它来决定。DNS翻译域名时无需像查电话本一样刻板地机械翻译 , 根据来访机器、网络链路、服务内容等各种信息 , 可以玩出很多花样。
- **根域名服务器** ( Root DNS ) 是指固定的、无需查询的 ( 可以默认为已内置 ) **顶级域名** ( Top-Level Domain ) 服务器。全世界一共有13个根域名服务器 ( 但并不是13台 , 每一个根域名都通过任播的方式建立了一大群镜像 , 根据维基百科的数据 , 迄今已经超过1000台根域名服务器的镜像了 ) 。13这个数字是由于DNS主要采用UDP传输协议 ( 在需要稳定性保证的时候也可以采用TCP ) 来进行数据交换 , 未分片的UDP数据包在IPv4下最大有效值为512字节 , 由此而来的限制。

4. 我们假设本地DNS是新开张的 , 上述权威服务器的记录它都没有 , 一直查到根域名服务器后 , 它将会得到“cn的权威服务器”的记录 , 然后通过“cn的权威服务器” , 得到“com.cn的权威服务器” , 以此类推 , 最后找到“www.icyfenix.com.cn的权威服务器”。
5. 通过“www.icyfenix.com.cn的权威服务器” , 查询www.icyfenix.com.cn的地址记录 ( 有RFC定义的地址记录有数十种类型 ) , 譬如IPv4下的IP地址为A记录 , IPv6下的AAAA记录、主机别名CNAME记录 , 等等 ) , 选择一条合适的返回给客户端。

一个域名可以配置多条不同的A记录 , 此时权威服务器可以根据自己的策略来进行选择。一种典型的应用是智能线路 : 根据访问者所处的不同地区 ( 譬如华北、华南、东北、港澳台、国外 ) 、不同服务商 ( 譬如电信、联通、移动 ) 等因素来确定返回的A记录。

DNS系统多级分流的设计使得DNS系统能够经受住全球网络流量不间断的冲击 , 但也并非全无缺点。譬如 , 当极端情况 ( 各级服务器均无缓存 ) 下的域名解析可能导致后续递归的多次查询而显著影响响应速度 , 譬如下图所示。



首次DNS请求耗时 ( 图片来自网络 )

专门有一种被称为“DNS预取”（ DNS Prefetching ）的前端优化手段：如果网站后续要使用来自于其他域的资源，那就在网页加载时便生成一个link请求，促使浏览器对该域名进行预解释，譬如下面所示：

```
<link rel="dns-prefetch" href="//domain.not-icyfenx.cn">
```

html

而另一种可能更严重的缺陷是DNS的分级查询意味着每一级都有可能受到中间人攻击的威胁，产生被劫持的风险。要攻陷位于递归链条顶层的（譬如根域名服务器，cn权威服务器）服务器和链路是非常困难的，但很多位于递归链底层的、本地运营商的Local DNS服务器的安全防护则相对松懈，甚至不少地区的运行商自己就会进行劫持，专门返回一个错的IP，在这个IP上代理用户请求，以便给特定资源（主要是HTML）注入广告，以此牟利。

为此，最近几年出现了另一种新的DNS应用形式：[HTTPDNS](#)（也称为DNS over HTTP S，DoH）。它将DNS服务开放为一个HTTPS服务，替代基于UDP传输协议的DNS域名解析，直接从权威DNS或者可靠Local DNS获取解析数据，从而绕过传统Local DNS。这种做法的好处是避免了底层的域名劫持（遇到顶层劫持是往往是政府行为，这是没办法的），能够有效解决Local DNS不可靠导致的域名生效缓慢、来源IP不准确产生的智能线路切换错误等问题。

# 链路优化

## 链路优化 (Transmission Optimization)

今天的链路优化原则，在若干年后的未来再回头看它们时，其中多数已经成了奇技淫巧，有些甚至成了反模式<sup>1</sup>。

在开始本节的讨论前，笔者先列一些在网络上很容易就能找到的，对Web进行链路性能优化的原则（譬如[雅虎YSlow23条规则](#)），这些原则在今天大多仍是（暂时）有一定价值的，至少也算是曾经（可能现在也还算是）广泛地流行过，但大概率在若干年后的未来再回头看它们时，其中多数已经成了奇技淫巧，有些甚至成了反模式。趁着当今的Web在传输链路这一块正处于新老交替之际，我们来说一下两代HTTP协议下的链路优化的问题。

1. 利用客户端缓存：缓存总是有益的，这点第一节中详细介绍过，本节不再涉及。
2. 减少请求数量：请求每次都需要建立通信链路进行数据传输，这些开销很昂贵，减少请求数量可有效的提高访问性能。
  - 雪碧图 ([CSS Sprites](#))
  - CSS、JS文件合并/内联 (Concatenation / Inline)
  - 分段文档 ([Multipart Document](#))
  - 媒体（图片、音频）内联 ([Data Base64 URI](#))
  - 异步请求合并 (Batch Ajax Request)
  - .....
3. 扩大并发请求数：现代浏览器一般对每个域名支持6个（IE为8-13个）并发请求，如果希望更快地加载大量图片或其他资源，需要进行域名分片（Domain Sharding），将图片同步到不同主机或者同一个主机的不同域名上（YSlow：Split Components Across Domains）。
4. 避免页面重定向：当页面发生了重定向，就会延迟整个文档的传输。在HTML文档到达之前，页面中不会呈现任何东西，降低了用户体验。
5. 按重要性调节资源优先级：将重要的、马上就要使用的、对客户端展示影响大的资源，放在HTML的头部，以便优先下载。

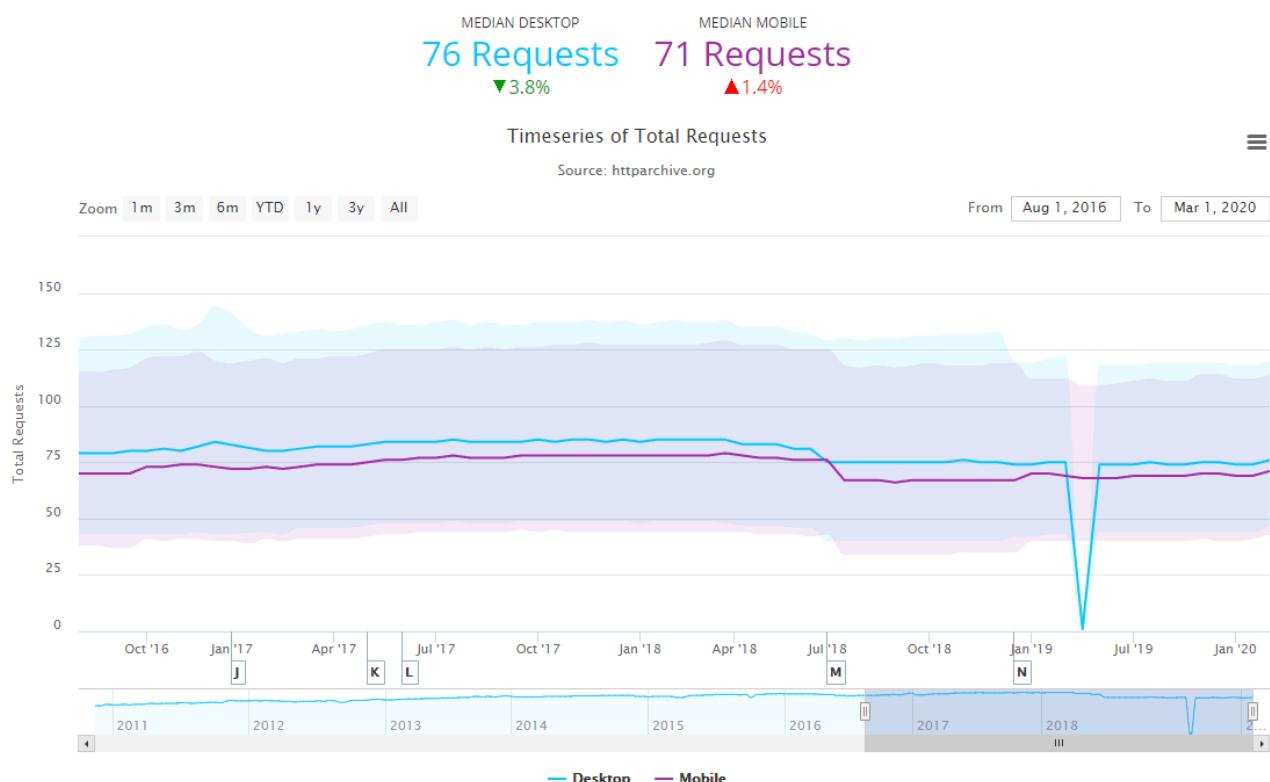
6. 启用压缩传输：启用压缩能够大幅度减少需要在网络上传输内容的大小，节省网络流量。

7. .....

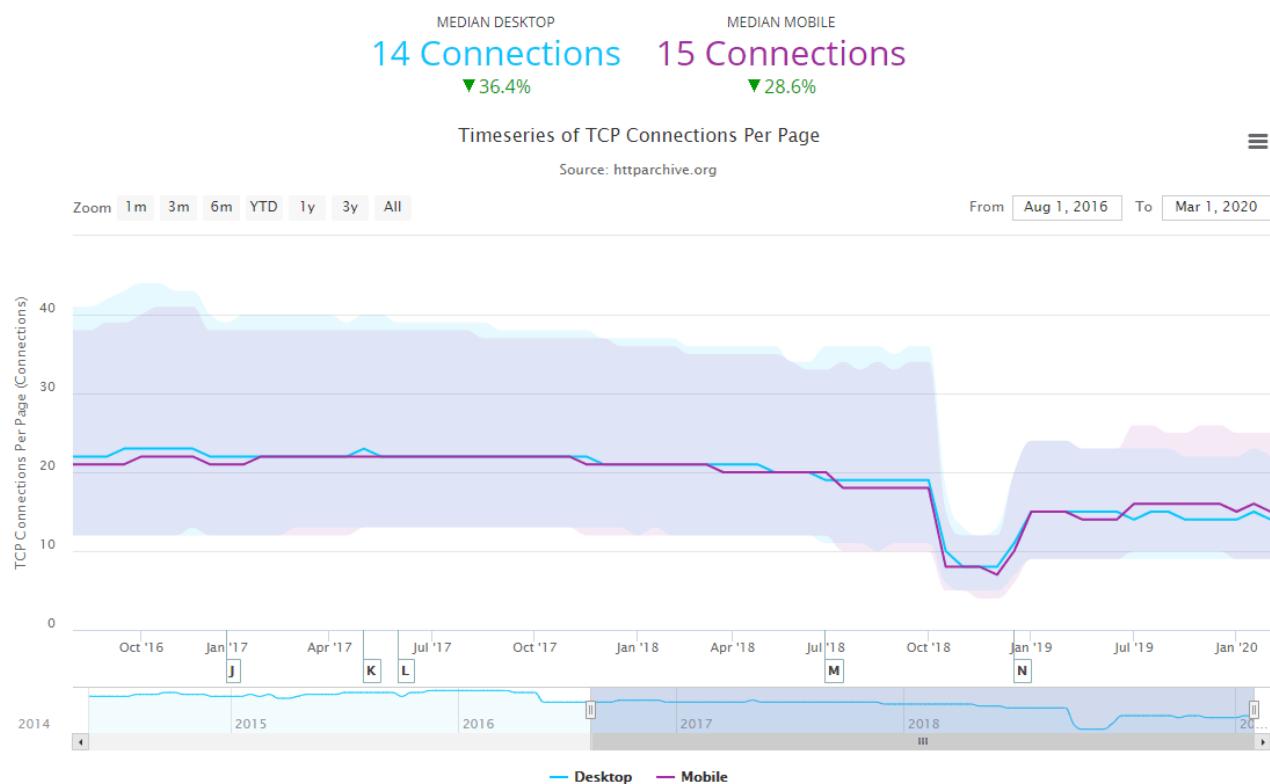
如同之前介绍客户端缓存时提到的那样，HTTP要得到无状态的好处，就必须相应承受网络效率降低的代价。在其他方面，HTTP协议设计和应用中也经历过了类似的权衡取舍，现在看来那些需要用户去优化的内容，往往都是当时技术现状下权衡取舍的结果。我们就从优化原则中条目最多的针对HTTP请求数量的措施说起。

## 连接数优化

我们都知道HTTP是基于TCP协议的，必须在[TCP三次握手](#)完成之后才能进行数据传输，这是一个通常以“百毫秒”为计时尺度的事件；此外，TCP还有[慢启动](#)的特性，使得刚刚建立连接时传输速度是最低的，后面再逐步加快直至稳定。由于TCP协议本身是面向于长时间、大数据传输来设计的，在长时间尺度下，它连接建立的成本高昂才不至于成为瓶颈，它的稳定性和可靠性的优势才能展现出来，那显然HTTP over TCP这种搭配，在目标倾向于上就多少产生了一些矛盾，以至于HTTP/1.x时代，大量短而小的TCP连接确实造成了网络性能的瓶颈。为了缓解HTTP在这个问题上的缺陷，聪明的程序员们一面致力于减少发出的请求数量，另外一方面也致力于增加客户端到服务端的连接数量，就是上面2、3点所提到的优化措施。这些Tricks的确减少消耗TCP连接数量，下面两张图片是来自于[HTTP Archive](#)对最近五年来数百万个URL地址采样得出的结论，页面平均请求没有改变的情况下，TCP连接在持续地下降（当然，后面说的HTTP/2.0其实占了很大功劳）。



HTTP平均请求数量，70余个，没有明显变化



TCP连接数量，约15个，有明显下降趋势

但是，上述这些节省TCP连接的优化措施但也带来了诸多不良的副作用：

- 如果你用CSS Sprites将多张图片合并，意味着任何场景下哪怕只用到其中一张小图，也必须完整加载整个大图片；任何场景下哪怕一张小图要进行修改，都会导致整个缓存失

效，类似地，样式、脚本等其他文件的合并也会造成同样的问题。

- 如果你使用了媒体内嵌，除了要承受Base64编码导致提及膨胀1/3的代价外，也将无法有效利用缓存。
- 如果你合并了异步请求，这就会导致所有请求返回时间都受最慢的那个请求的拖累，整体响应速度下降。
- 如果你把图片放到不同子域下面，将会导致更大的DNS解析负担，而且浏览器对两个不同子域下的同一图片必须持有两份缓存，也使得缓存效率的下降。
- .....

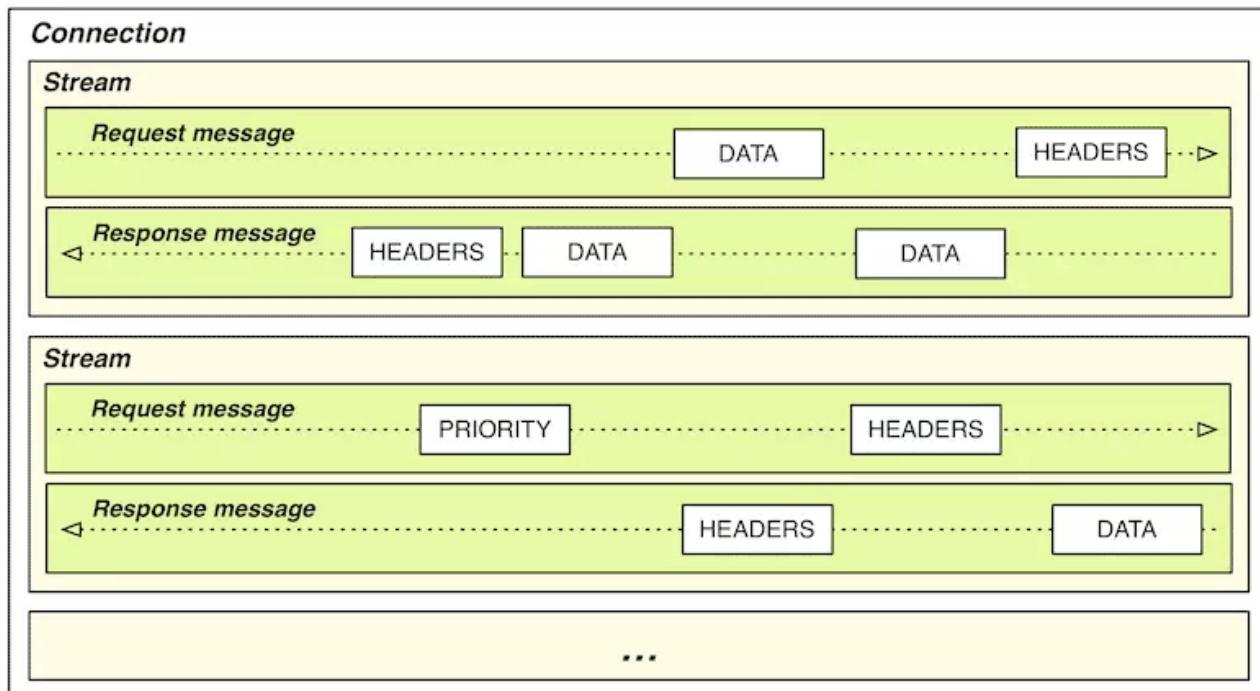
由此可见，一旦技术根基上出现的缺陷，依赖使用者通过各种Tricks去解决，无论如何都难以摆脱“两害相权取其轻”的权衡困境，否则这就不是Tricks而是会成为一种标准的设计模式了。

在另一方面，HTTP的设计者们并非没有尝试过在基础设施层面去解决连接成本过高的问题，即使是HTTP协议的最初版本（指HTTP/1.0，忽略非正式的HTTP/0.9版本）也是支持（不是默认，HTTP/1.1中变为默认）连接复用的，即今天大家所熟知的[持久连接](#)（Persistent Connection）或者叫连接[Keep-Alive机制](#)。其大致原理是让客户端可以对一个域名长期持有一个（或多个）TCP连接，在客户端维护一个FIFO队列，每次取完数据（如何在不断开连接下判断取完数据将会放到稍后压缩部分去讨论）之后不断开连接，以便下一个资源需要获取时备用，避免创建TCP连接的成本。而在2014年，IETF发布的[RFC 7230](#)中提出了名为“[HTTP管道](#)”（HTTP Pipelining）复用技术试图在服务端也建立类似的队列，以进一步提高效率，客户端一次过将所有请求发给服务端，由服务端来管理队列的话，可以保证队列中两项工作之间没有空隙，甚至可能进行并行化处理，提升了服务端的效率。不过，HTTP管道需要多方共同支持，推广得并不算成功。

不幸的是，连接复用仍然存在它的副作用，最主要的一项副作用是“[队首阻塞](#)”（Head-of-Line Blocking）问题，请设想以下场景：浏览器有10个资源需要从服务器中获取，此时它将10个资源放入队列，入列顺序只能是按照浏览器预见这些资源的先后顺序来决定的。但如果这10个资源中的第1个就让服务器陷入长时间运算状态那会怎样？当它的请求被发送到服务端之后，服务端开始计算，而运算结果出来之前TCP连接中并没有任何数据返回，此时后面9个资源都必须阻塞等待。无论队列维护在服务端还是客户端，其实都无法解决这个问题，因为服务端虽然很可能可以并行处理另外9个请求（譬如第一个是复杂运算请求，消耗CPU资源，第二个是数据库访问，消耗数据库资源，第三个是访问某张图片，消耗磁盘I/O资源，等等，这就很适合并行），但处理结果却无法发回给客户端，服务端既不能哪个

请求先完成就返回哪个，更不可能将所有要返回的资源混杂到一起交叉传输……显然，TCP连接带来的问题，本质上是传输链路上的问题，无论在服务端还是客户端，涉及到传输方面都显得无能为力。

队首阻塞问题一直持续到第二代的HTTP协议，即HTTP/2.0发布后才算是被比较完美地解决。在HTTP/1.x中，“请求”就是传输过程中最小粒度的信息单位了，所以如果将多个请求切碎，再混杂在一块传输，客户端势必难以分辨重组出有效信息。而在HTTP/2.0中，帧（Frame）才是最小粒度的信息单位，它可以用来描述各种数据，譬如请求的Header、Body，或者用来做控制标识，譬如打开流、关闭流。这里说的流（Stream）是一个逻辑数据通道的概念，每个帧都附带有一个流ID以标识这个帧属于哪个流。这样，在同一个TCP连接中传输的多个数据帧就可以根据流ID轻易区分出来，在客户端毫不费力地将不同流中的数据重组出HTTP的请求、响应报文来。这项设计是HTTP/2.0的重点技术特征之一，被称为[HTTP/2.0 多路复用](#)（HTTP/2.0 Multiplexing）



HTTP2的多路复用（图片来自：<https://hpbn.co/http2>）

有了多路复用的支持，HTTP/2.0就可以对每个域名只维持一个TCP连接（One Connection Per Origin），既减轻了服务器的连接压力，开发者也不用去考虑域名分片这种事情来突破浏览器对每个域名最多6个连接数限制了。而更重要的是，没有了TCP连接数的逼迫，所有通过合并/内联文件（无论是图片、样式、脚本）以减少请求数的需求就不再成立了，甚至反而是徒增副作用的反模式了——可能还有人会反驳说：不至于吧，减少请求数量，不是至少还减少了传输中耗费的Header吗？先得承认一个事实，在HTTP协议中，Header的成

本所占的比重相当的大，以至于在HTTP/2.0中需要专门考虑如何进行Header压缩的问题。但是，以下几个因素导致了通过合并资源文件减少请求数，对节省Header成本也几乎没有帮助：

- Header的传输成本在Ajax（尤其是只返回少量数据的请求）请求中可能是比重很大的开销，但在图片、样式、脚本这些静态资源的请求中，通常并不占主要。
- 在HTTP/2.0中Header压缩的原理是基于字典编码的信息复用，简而言之是同一个连接上产生的请求和响应越多，动态字典积累得越全，头部压缩效果也就越好。所以HTTP/2.0是单域名单连接的机制，合并资源和域名分片反而对性能提升不利。
- 与HTTP/1.x相反，HTTP/2.0本身反而变得更适合传输小资源了，譬如传输1000张10K的小图，HTTP/2.0要比HTTP/1.x快，但传输10张1000K的大图，则应该HTTP/1.x会更快。这一方面是TCP连接数量（相当于多点下载）的影响，更多的是由于TCP协议丢包重传机制导致的，一个丢失的TCP包会导致所有的流都必须等待这个包重传成功，这个问题就是HTTP/3.0要解决的目标了。因此，把小文件合并成大文件，在HTTP/2.0下是毫无好处的。

## 传输压缩

我们接下来再花一点点篇幅来讨论链路优化中除了缓存、连接之外另一个主要话题：压缩。很多人都知道HTTP协议是支持[GZip](#)压缩的，由于HTTP传输的主要内容，譬如HTML、CSS、Script等，都是文本数据，对于这些文本数据启用压缩的收益是非常高的，传输量一般会降至原有的20%左右。而对于那些不适合压缩的资源，Web服务器则能根据MIME类型来自动判断是否对响应进行压缩，这样，已经采用过压缩算法存储的资源，如JPEG、PNG图片，便不会被二次压缩，空耗性能。

不过，大概就没有多少人想过压缩与之前提到的用于节约TCP的持久连接机制是存在一些冲突的。在古代，服务器处理能力还很差的时候，通常是把静态资源先预先压缩为.gz文件的形式存放起来，当客户端可以接受压缩版本的资源时（请求的Header中包含Accept-Encoding: gzip）就返回压缩后的版本（响应的Header中包含Content-Encoding: gzip），否则就返回未压缩的原版，这种方式被称为“[静态预压缩](#)”（Static Pre-compression）。而现代的Web服务器处理能力有了大幅提升，已经没有人再采用麻烦的预压缩方式了，都是由服务器对符合条件的请求将在输出时进行“[即时压缩](#)”（On-The-Fly Compression），整个压缩过程全部在内存的数据流中完成，不必等资源压缩完成再返回响应，这样可以显著

提高“首字节时间”（ Time To First Byte , TTFB ） , 改善Web性能体验。而这个过程中唯一不好的地方就是服务器再没有办法给出Content-Length这个响应Header了 , 因为输出Header时服务器还不知道压缩后资源的确切大小。

到这里 , 大家想明白即时压缩与持久链接的冲突在哪了吗 ? 持久链接机制不再依靠TCP连接是否关闭来判断资源请求是否结束 , 它会重用同一个连接以便向同一个域名请求多个资源 , 这样 , 客户端就必须要有除了关闭连接之外的其他机制来判断一个资源什么时候算传递完毕 , 这个机制最初 ( 在HTTP/1.0时 ) 就只有Content-Length , 即靠着请求头中明确给出资源的长度 , 传输到达该长度即宣告一个请求响应的结束。由于启用即时压缩后就无法给出Content-Length了 , 如果是HTTP/1.0的话 , 持久链接和即时压缩只能二选其一 ( HTTP/1.0中两者默认都是不开启的 ) 。其实Content-Length的缺陷不仅仅在于即时压缩这一种场景 , 譬如对于动态内容 ( Ajax、 PHP、 JSP等输出 ) , 服务器也同样无法事项得知Content-Length。

HTTP/1.1版本中修复了这个缺陷 , 增加了另一种“分块传输编码” ( Chunked Transfer Encoding ) 的资源结束判断机制 , 解决Content-Length与持久链接的冲突问题。分块编码原理相当简单 : 在响应Header中加入“Transfer-Encoding: chunked”之后 , 就代表这个响应报文将采用分块编码。此时 , 报文中的Body需要改为用一系列“分块”来传输。每个分块包含十六进制的长度值和对应长度的数据内容 , 长度值独占一行 , 数据从下一行开始。最后以一个长度值为0的分块来表示资源结束。举个例子 ( 来自于前面维基百科中的页面 , 为便于观察 , 只分块 , 未压缩 ) :

```
HTTP/1.1 200 OK
Date: Sat, 11 Apr 2020 04:44:00 GMT
Transfer-Encoding: chunked
Connection: keep-alive

25
This is the data in the first chunk

1C
and this is the second one

3
con

8
sequence
```

0

根据分块长度可知，前两个分块包含显式的回车换行符（CRLF，即\r\n字符）

```
"This is the data in the first chunk\r\n" (37 字符 => 十六进制:
0x25)
"and this is the second one\r\n" (28 字符 => 十六进制:
0x1C)
"con" (3 字符 => 十六进制:
0x03)
"sequence" (8 字符 => 十六进制:
0x08)
```

所以解码后的内容为：

```
This is the data in the first chunk
and this is the second one
consequence
```

一般来说，Web服务器给出的数据分块大小是一致的（但并不强制），而不是如例子中那样随意。HTTP/1.1通过分块传输解决了即时压缩与持久连接并存的问题，到了HTTP/2.0，由于多路复用和单域名单连接的设计，已经无需再刻意强去提久链接机制了，但数据压缩仍然有节约传输带宽的重要价值。

# 内容分发网络

## 内容分发网络 (Content Distribution Network)

CDN是一种十分古老而又十分透明，没什么存在感的分流系统，许多人都说听过它，但真正了解过它的人却很少。

前面几个小节介绍了缓存、域名解析、链路优化，这节我们来讨论它们的一个经典的综合运用案例：内容分发网络 (Content Distribution Network, CDN)。

CDN是一种十分古老的应用，以至于笔者相信阅读本文的受众至少有八、九成应该对它有不同程度的了解的——起码是听说过它的名字的。如果把某个互联网系统比喻为一家开门营业的企业，那CDN就是它遍布世界各地的分支销售机构，客户要买一块CPU就订机票飞到美国加州Intel总部去那肯定是不合适的，到本地电脑城找个装机铺才是正常人类的做法，CDN就相当于电脑城那吆喝着CPU三十块钱一斤的本地经销商。

CDN又是一种十分透明的应用，以至于笔者相信阅读本文的受众至少有八、九成应该对它是如何为互联网站点分流、对它的工作原理并没有什么系统性的概念——起码没有自己亲自使用过。如果抛却其他影响服务质量的因素，仅从网络角度看，一个互联网系统的速度快慢取决于以下四点因素：

1. 网站服务器接入网络运营商的链路所能提供的出口带宽。
2. 用户客户端接入网络运营商的链路所能提供的入口带宽。
3. 从网站到用户之间经过的不同运营商之间互联节点的带宽，一般来说两个运营商之间只有固定的若干个点是互通的，所有跨运营商之间的交互都要经过这些点。
4. 从网站到用户之间的物理链路传输时延。打游戏的同学都清楚，ping比流量更重要。

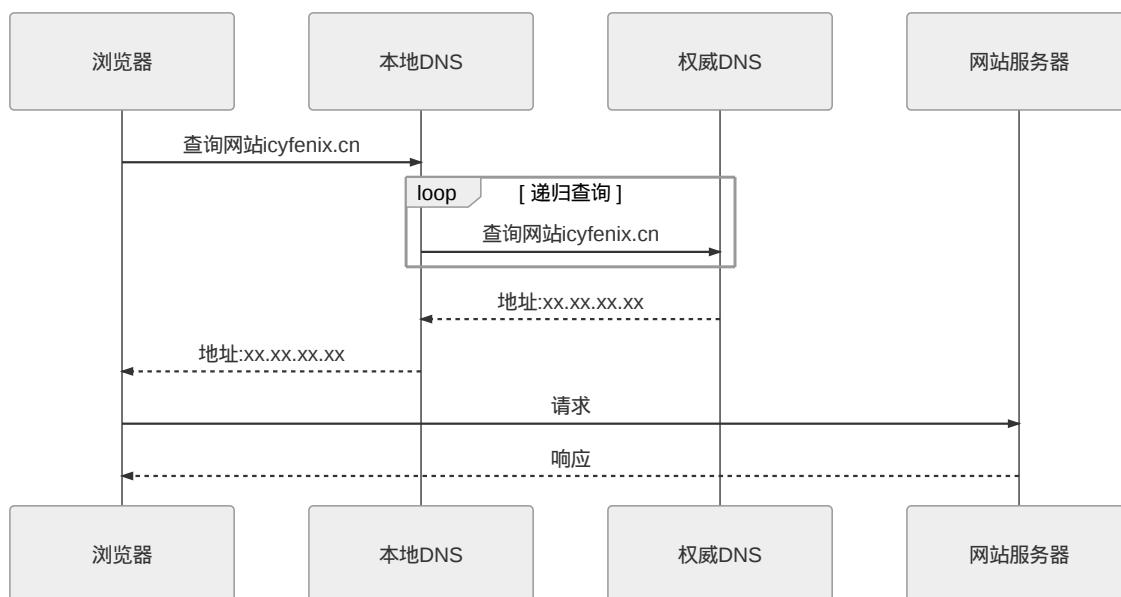
以上四个网络问题，除了第二个只能由用户掏腰包装个更好的宽带才能够解决之外，其余三个都能通过内容分发网络来显著改善的。一个工作良好的CDN，能为互联网系统解决跨运营商、跨地域物理距离所导致的时延问题，能为网站流量带宽起到分流、减负的作用。

如果不是有遍布全国的阿里云CDN网络支持，哪怕把整个杭州所有市民上网的权力都剥夺，带宽全部让给淘宝，恐怕也撑不住双十一全国乃至全球用户的疯狂围攻。

CDN的工作过程，主要涉及到路由解析、内容分发、负载均衡（由于后面专门有一节讨论负载均衡的内容，所以这部分在CDN中就暂不涉及）和所能支持的应用内容四个方面，下面我们就逐一了解。

## 路由解析

根据我们在第二节中对DNS系统的介绍，一个未使用CDN的用户访问网站的过程应该是这样的：



以上时序所表达的内容跟第二节中讲述的没有差异，这里仅列作对比，不再赘述。下面分析使用了CDN的DNS查询过程之前，我们先来看一段对本站进行DNS查询的实际应答。通过dig或者host命令，可以很方便地得到DNS服务器的返回结果（结果中头4个IP的地址是我手工加入的，后面的就不一个一个查了），如下所示：

```

$ dig icyfenix.cn
sh

; <>> DiG 9.11.3-1ubuntu1.8-Ubuntu <>> icyfenix.cn
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 60630
;; flags: qr rd ra; QUERY: 1, ANSWER: 17, AUTHORITY: 0, ADDITIONAL: 1

```

```

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 65494
;; QUESTION SECTION:
;icyfenix.cn. IN A

;; ANSWER SECTION:
icyfenix.cn. 600 IN CNAME
icyfenix.cn.cdn.dnsv1.com.
icyfenix.cn.cdn.dnsv1.com. 599 IN CNAME
4yi4q4z6.dispatch.spcdntip.com.
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 101.71.72.192 #浙江宁波市
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 113.200.16.234 #陕西省榆林市
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 116.95.25.196 #内蒙古自治区呼和浩特市
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 116.178.66.65 #新疆维吾尔自治区乌鲁木齐市
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 118.212.234.144
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 211.91.160.228
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 211.97.73.224
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 218.11.8.232
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 221.204.166.70
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 14.204.74.140
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 43.242.166.88
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 59.80.39.110
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 59.83.204.12
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 59.83.204.14
4yi4q4z6.dispatch.spcdntip.com. 60 IN A 59.83.218.235

;; Query time: 74 msec
;; SERVER: 127.0.0.53#53(127.0.0.53)
;; WHEN: Sat Apr 11 22:33:56 CST 2020
;; MSG SIZE rcvd: 152

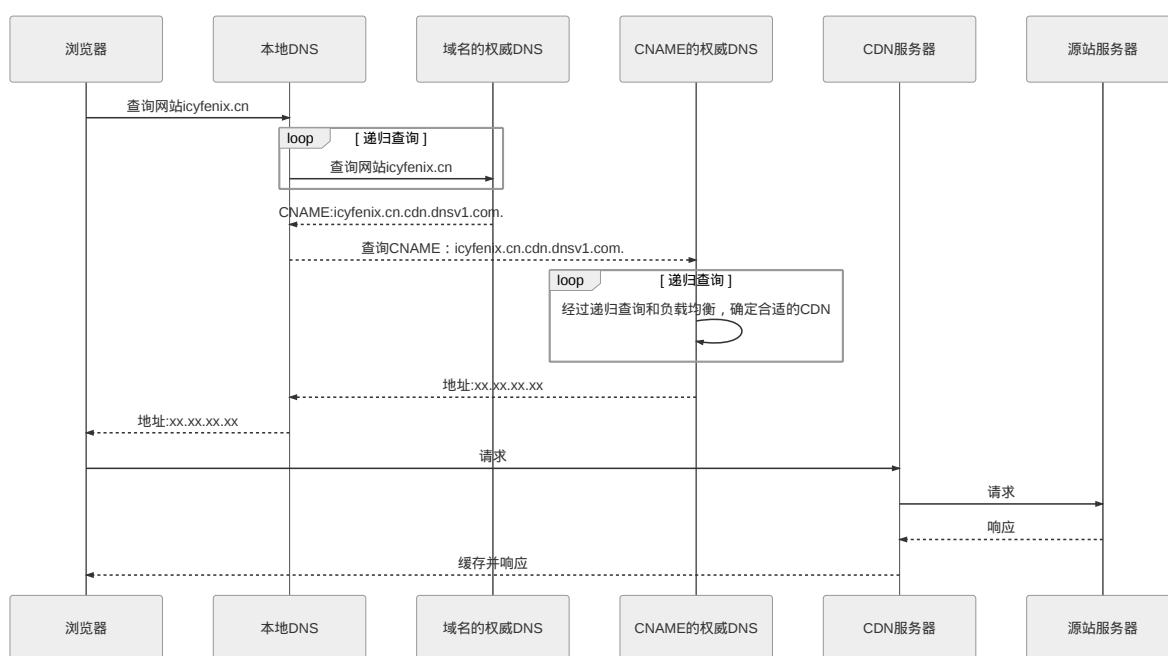
```

根据以上信息，查询“icyfenix.cn.”的查询结果首先返回了一个CNAME记录（icyfenix.cn.cdn.dnsv1.com.），递归查询该CNAME时候，返回了另一个看起来更奇怪的CNAME（4yi4q4z6.dispatch.spcdntip.com.），最后，这个CNAME返回了十几个位于全国不同地区的A记录，很明显，那些A记录就是存有本站缓存的CDN节点。CDN路由解析的工作过程是：

1. 架设好服务器后，将服务器的IP地址在你的CDN服务商上注册为“源站”，注册后你会得到一个CNAME，即本例中的“icyfenix.cn.cdn.dnsv1.com.”。

2. 将得到的CNAME在你购买域名的DNS服务商上注册为一条CNAME记录。
3. 当发生一次未命中缓存的DNS查询时，域名服务商解析出CNAME后，返回给本地DNS，至此之后链路解析的主导权就开始由CDN的调度服务接管了。
4. 本地DNS查询CNAME时，能解析该CNAME的权威服务器只有CDN服务商的权威DNS，该DNS会根据一定的均衡策略和参数，如拓扑结构、容量、时延等，在全国各地能提供服务的CDN节点中挑选一个适合的，将它的IP返回给本地DNS。
5. 浏览器从本地DNS拿到IP，将该IP当作源站服务器来进行访问，此时该IP的CDN服务上可能有，也可能没有缓存源站的资源，这点将在稍后“内容分发”部分讨论。
6. CDN代替源站向用户提供所需的资源。

以上步骤反映在时序图上，将如下图所示：



## 内容分发

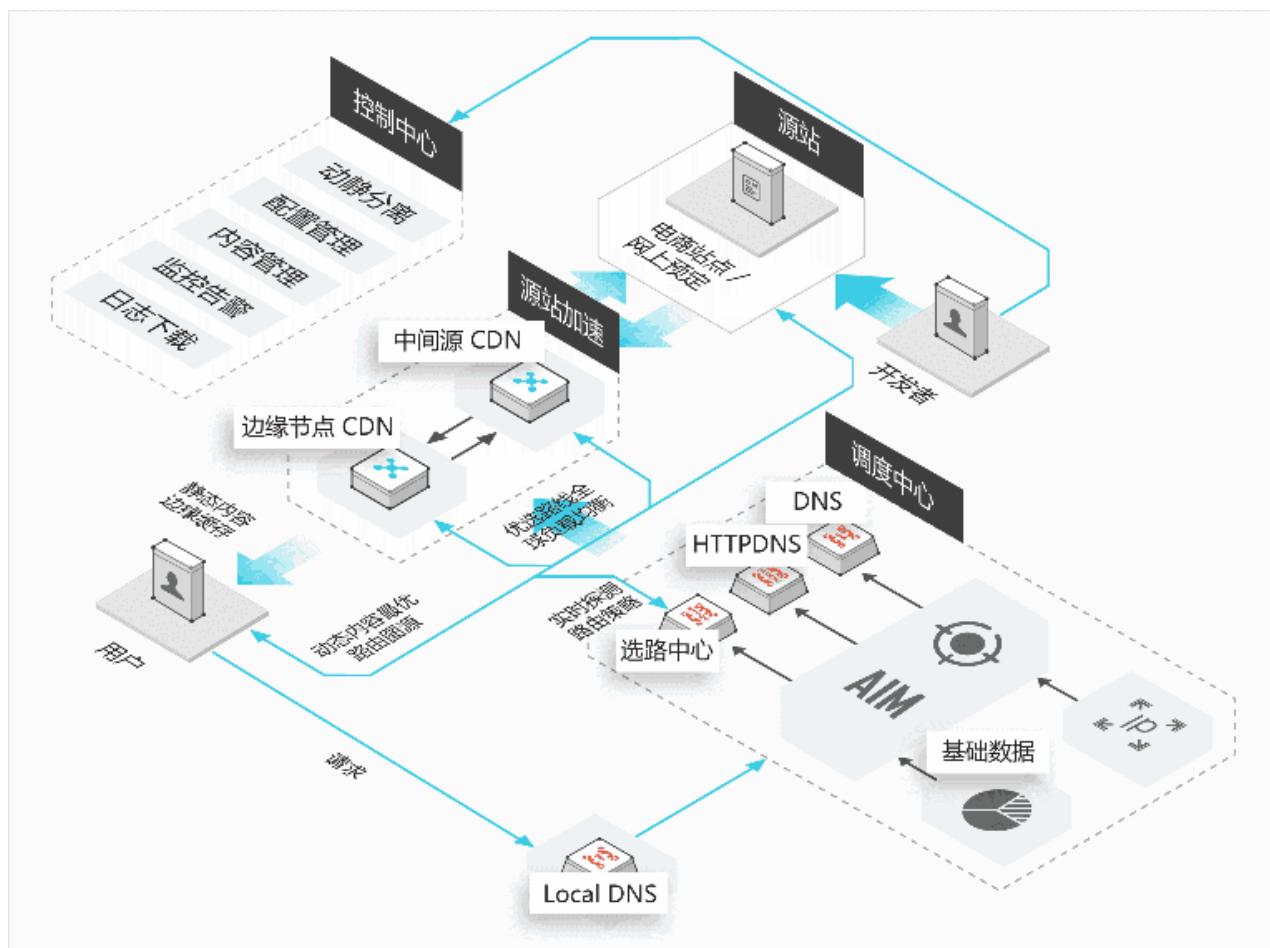
通过智能化的路由解析，CDN节点实现了无论是对用户端还是服务端，都完全透明地中间接管了用户向服务发出的资源请求，后面随之而来的下一个问题是CDN中必须缓存有用户想要请求的资源，然后才能代替源站来满足用户的资源请求。这包括了两个大的问题：“如何获取源站资源”和“如何管理（更新）资源”。

CDN获取源站资源的过程被称为“内容分发”，目前主要有两种主流的内容分发方式：

- **主动分发（Push）**：分发由源站主动发起，将内容从源站或者其他资源库推送到各边缘的CDN节点上。这个推送的过程没有什么技术标准，可以采用任何传输方式（HTTP、F

TP、P2P，等等）、任何推送策略（满足特定条件、定时、人工，等等）、任何推送时间，只要与后面说的更新策略先匹配即可。由于主动分发（通常）需要源站、CDN服务双方提供程序API接层面的配合，它并不是透明的，一般用于系统预载大量数据资源。譬如双十一之前一段时间内，淘宝、京东等各个网络商城就会开始把未来活动中所需用到的资源推送到CDN上，特别常用的资源甚至会直接缓存到你的手机APP、你浏览器的localStorage上。

- **被动回源（Pull）**：回源由用户访问触发，由CDN服务发起。当某资源第一次被用户访问的时候，CDN会实时从源站获取，这时资源的响应时间可粗略认为是资源从源站到CDN的时间加上CDN到用户的时间之和，因此，被动回源的首次访问通常是比较慢（但并不一定比你直接访问源站慢）的，不适合应用于大量的数据资源。但被动回源可以做到完全透明，不需要源站在程序上做任何的配合，使用起来较为方便，这种是小型站点使用CDN服务的主流选择（不是自建CDN，购买阿里云、腾讯云的CDN服务，多数就是这种方式）。



阿里云官网上的的CDN介绍

对于“CDN如何管理（更新）资源”这个问题，同样没有统一的标准可言，尽管在HTTP协议中关于缓存的Header定义中确实是有对CDN这类共享缓存的一些指引性参数（如Cache-C

ontrol的s-maxage），但是否要遵循，完全取决于CDN本身策略。更令人无奈的是，由于大多数网站的开发和运维本身并不了解HTTP缓存机制，所以CDN如果完全照着HTTP Header来控制缓存失效和更新，效果也会相当的差，还可能引发其他问题。因此，CDN缓存的管理并不存在通用的准则。

## CDN应用

CDN是为了快速分发静态资源而设计的，但今天的CDN所能做的事情已经远远超越了开始建设时的目标，这部分无法展开逐一细说，只对现在CDN可以做的事情简要列举，以便有个总体认知：

- 加速静态资源：这是CDN本职工作。
- 安全防御：CDN在广义上可以视作你网站的堡垒机，源站只对CDN提供服务，由CDN来对外界其他用户服务，这样恶意攻击者就不容易直接威胁源站，CDN对防御某些攻击手段，如[DDoS攻击](#)的尤其有效。但需注意，将安全都寄托在CDN上本身是不安全的，一旦源站真实IP被泄漏，就会面临很高风险。
- 协议升级：不少CDN提供商都同时对接（代售CA的）SSL证书服务，可以实现源站是HTTP的，而对外开放的网站是基于HTTPS的。同理，可以实现源站到CDN是HTTP/1.x协议，CDN提供的外部服务是HTTPS/2.0协议、实现源站是基于IPv4网络的，CDN提供的外部服务支持IPv6网络，等等。
- 状态缓存：第一节介绍缓存时简要提到了一下状态缓存，CDN不仅可以缓存源站的资源，还可以缓存源站的状态，譬如源站的301/302转向，可以缓存起来，让客户端直接跳转、可以通过CDN开启HSTS、可以通过CDN进行[OCSP装订](#)加速SSL证书访问，等等。有一些情况下甚至可以配置CDN对任意状态码（譬如404）进行一定时间的缓存，以减轻源站压力，但这个操作应当慎重。
- 修改资源：CDN可以在返回资源给用户的时候修改它的任何内容，以实现不同的目的。譬如，可以对源站未压缩的资源自动压缩并修改Content-Encoding，以节省用户的网络带宽消耗、可以对源站未启用客户端缓存的内容加上缓存Header，以启用客户端缓存，可以修改[CORS](#)的相关Header，将源站不支持跨域的资源提供跨域能力，等等。
- 访问控制：CDN可以实现IP黑/白名单，根据不同的来访IP提供不同的响应结果，根据IP的访问流量来实现QoS控制、根据HTTP的Referer来实现防盗链，等等。
- 注入功能：CDN可以在不修改源站代码的前提下，为源站注入各种功能，下图是国际CDN巨头CloudFlare提供的Google Analytics、PACE、Hardenize等第三方应用，均无需

修改源站任何代码。

## 1. Install your first app

We recommend getting started by installing one of these great apps. It only takes a few

Install



**Drift**

Automatically turn your website traffic into qualified sales me

Install



**Google Analytics**

Deep insights into your site's traffic to improve user retention

Install



**Hardenize**

Security reports that help safeguard your site from attacks.

Install



**PACE**

A loading bar for your site to make it feel faster while improv

- 绕过某些不存在的网络措施，这也是在国内申请CDN也必须实名备案的原因，就不细说了。
- .....

# 负载均衡

## 负载均衡 ( Load Balancing )

调度后方的多台机器，以统一的接口对外提供服务，承担此职责的技术组件被称为“负载均衡”。

互联网早期，业务流量比较小并且业务逻辑比较简单，单台服务器便可以满足基本的需求，但时至今日，互联网也好，企业也好，一般实际用于生产的系统，几乎都离不开集群了。信息系统不论是分布式单体架构还是微服务架构，不论是为了实现高可用还是为了获得高性能，都需要使用到多台机器来扩展服务能力，用户的请求不管连接到哪台机器上，都能得到相同的处理。另一方面，如何构建、调度服务集群这种事情，又应当对用户一侧保持足够的透明，即使用户的请求背后是由一千台、一万台机器来共同响应的，也并非用户所关心之事，他需记住的只有一个域名地址而已。调度后方的多台机器，以统一的接口对外提供服务，承担此职责的技术组件被称为“负载均衡” ( Load Balancing )

真正大型系统的负载均衡过程往往是多级的。譬如，在各地建有多个机房，或机房有不同网络链路入口的大型互联网站，会从DNS解析开始，通过“域名” → “CNAME” → “负载调度服务” → “就近的数据中心入口”，先将来访地用户根据IP（或者其他条件）分配到一个合适的数据中心中，然后才到后续将要讨论的各式负载均衡。在DNS层面的负载均衡与前一节介绍的CDN，在工作原理上是类似的，其差别只是数据中心能提供的不仅有缓存，而是全方位的服务能力。由于这种方式此前已经详细介绍过，后续我们所讨论的“负载均衡”将聚焦于网络请求进入数据中心入口之后的其他级别的负载均衡。

无论在网关内部建立了多少级的负载均衡，从形式上来说都可以分为两种：四层负载均衡和七层负载均衡。在详细介绍它们是什么、如何工作之前，我们先来建立两个大致的、概念性的印象：

- 四层负载均衡优势是性能高，七层负载均衡的优势功能强。
- 做多级混合负载均衡，通常应是低层的负载均衡在前，高层的负载均衡在后（想一想为什么？）。

我们所说的“四层”、“七层”，指的是经典的[OSI七层模型](#)中第四层传输层和第七层应用层，下表是来自于维基百科上对OSI七层模型的介绍，这部分属于网络基础知识，笔者就不多解释了（英文也偷懒不翻译了）。请注意到表中各个名词特意保留的超链接，后面我们会多次使用到这张表，如你对网络知识并不是特别了解的，可通过这些连接获得进一步的资料。下面我们直接从四层负载均衡具体工作过程开始说起。

	<b>Layer</b>	<b>Protocol Data Unit</b>	<b>Function</b>
7	<a href="#">Application</a>	Data	High-level APIs, including resource sharing, remote file access
6	<a href="#">Presentation</a>	Data	Translation of data between a networking service and an application; including character encoding, data compression and encryption/decryption
5	<a href="#">Session</a>	Data	Managing communication sessions, i.e., continuous exchange of information in the form of multiple back-and-forth transmissions between two nodes
4	<a href="#">Transport</a>	Segment	Reliable transmission of data segments between points on a network, including segmentation, acknowledgement and multiplexing
3	<a href="#">Network</a>	Packet	Structuring and managing a multi-node network, including addressing, routing and traffic control
2	<a href="#">Data link</a>	Frame	Reliable transmission of data frames between two nodes connected by a physical layer
1	<a href="#">Physical</a>	Symbol	Transmission and reception of raw bit streams over a physical medium

现在所说的四层负载均衡是多种工作模式的统称，“四层”的意思是说这些工作模式的共同特点是都维持着同一个TCP连接，而不是说它工作在第四层。事实上，这些模式其实都是工作在二层（数据链路层，改写MAC地址）和三层（网络层，改写IP地址）上的，单纯只处理第四层（传输层，可以改写TCP、UDP等协议的内容和端口）的数据无法做到负载均衡的转发，因为OSI的下三层是媒体层（Media Layers），上四层是主机层（Host Layers），既然流量都已经到达目标主机上了，也就谈不上什么流量转发，最多只能做代理了。但出于习惯和方便，现在几乎所有的资料都把它们统称为四层负载均衡，笔者也同样称呼

它为四层负载均衡，在这里提示一下，以免读者在某些资料上看见“二层负载均衡”、“三层负载均衡”这样的表述，误以为和这里说的“四层负载均衡”是一类意思。下面笔者介绍几种常见的四层负载均衡的工作模式。

## 数据链路层负载均衡

参考上面OSI模型的表格，数据链路层传输的内容是数据帧（Frame），譬如以常见的太网帧、ADSL宽带的PPP帧等。我们讨论的上下文场景里，目标肯定就是以太网帧，按照IEE E 802.3标准，以最典型的1500 Bytes MTU的以太网帧结构为例：

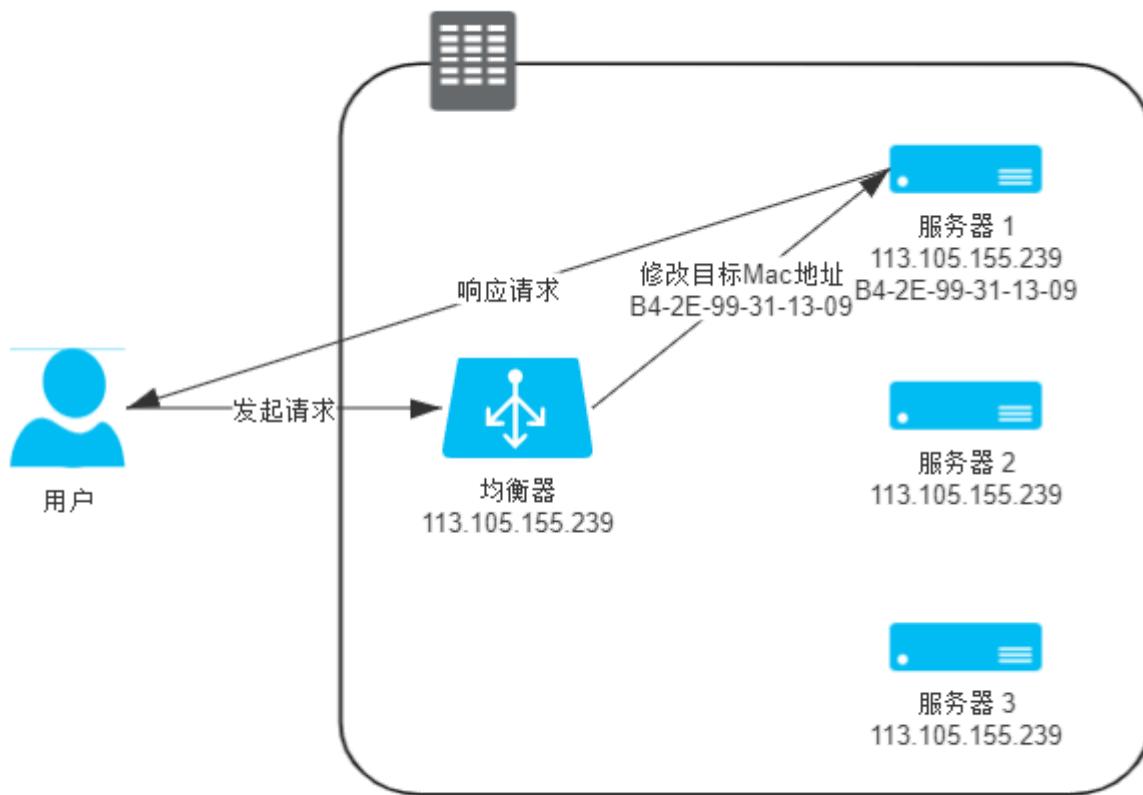
前导码	帧开始符	MAC目标地址	MAC源地址	802.1Q <u>标签(可选)</u>	以太类型	有效负载	冗余校验	帧间距
101 010 10 7 Bytes	101 010 11 1 Bytes	6 Bytes	6 Bytes	(4 Bytes)	2 Bytes	1500 Bytes	4 Bytes	12 Bytes

帧中其他结构项的含义我们可以不去理会，只需注意到“MAC目标地址”和“MAC源地址”两项即可。我们知道每一块网卡都有独立的MAC地址，以太帧上这两个地址告诉了交换机，此帧应该是从连接在交换机上的哪个端口的网卡发出，送至哪块网卡的。

数据链路层负载均衡所做的事情，是修改请求的数据帧中的MAC目标地址，让用户原本是发送给负载均衡器的请求的数据帧，被二层交换机根据新的MAC目标地址转发到服务器集群中对应的服务器（后文称为“真实服务器”，Real Server）的网卡上，这样真实服务器就获得了一个原本目标并不是发送给它的请求。

由于二层负载均衡器在转发请求过程中只修改了帧的MAC目标地址，不涉及更上层协议（没有修改Payload的数据），所以在更上层（第三层）看来，所有数据都是未曾被改变过的。由于第三层的数据包（IP数据包，下节会介绍）中包含了源（客户端）和目标（均衡器）的IP地址，只有真实服务器保证自己的IP地址与数据包中的目标IP地址一致，这个数据包才能被正确处理。因此，使用这种负载均衡模式时，将会把真实物理服务器集群所有机器的虚拟IP地址（Virtual IP Address，VIP）配置成与负载均衡服务器虚拟IP一样，这样经转发后的数据就能在真实服务器中顺利地使用。也正是因为实际处理请求的真实物理

服务器IP和数据请求中的目的IP是一致的，所以响应结果便不需要通过负载均衡服务器进行地址交换，可将响应结果的数据包直接从真实服务器返回给用户，避免负载均衡服务器网卡带宽成为瓶颈，数据链路层的负载均衡效率也是最高的。整个请求到响应的过程如下图所示：



### 数据链路层负载均衡

上述只有请求经过负载均衡器，而服务响应无需从负载均衡器原路返回的工作模式，整个请求、转发、响应的链路形成一个“三角关系”，所以这种负载均衡模式也常被很形象地称为“三角传输模式”（Direct Server Return，DSR），也有叫“单臂模式”（Single Legged Mode）或者“直接路由”（Direct Routing）。

虽然数据链路层负载均衡效率很高，但它并不能适用于所有的场合，除了那些需要感知应用层协议信息的负载均衡场景它无法胜任外（所有的四层负载均衡器都无法胜任，将在后续介绍七层均衡器时一并解释），它在网络一侧受到的约束也很大。它直接改写目标MAC地址的工作原理决定了它与真实的服务器的通讯必须是二层可达的，说白了就是必须在同一个子网当中，无法跨VLAN。优势（效率高）和劣势（不能跨子网）决定了数据链路层负载均衡最适合用来做数据中心的第一级（这里只谈负载均衡设备，并没有把ECMP等价路由算进去）负载均衡。

## 网络层负载均衡

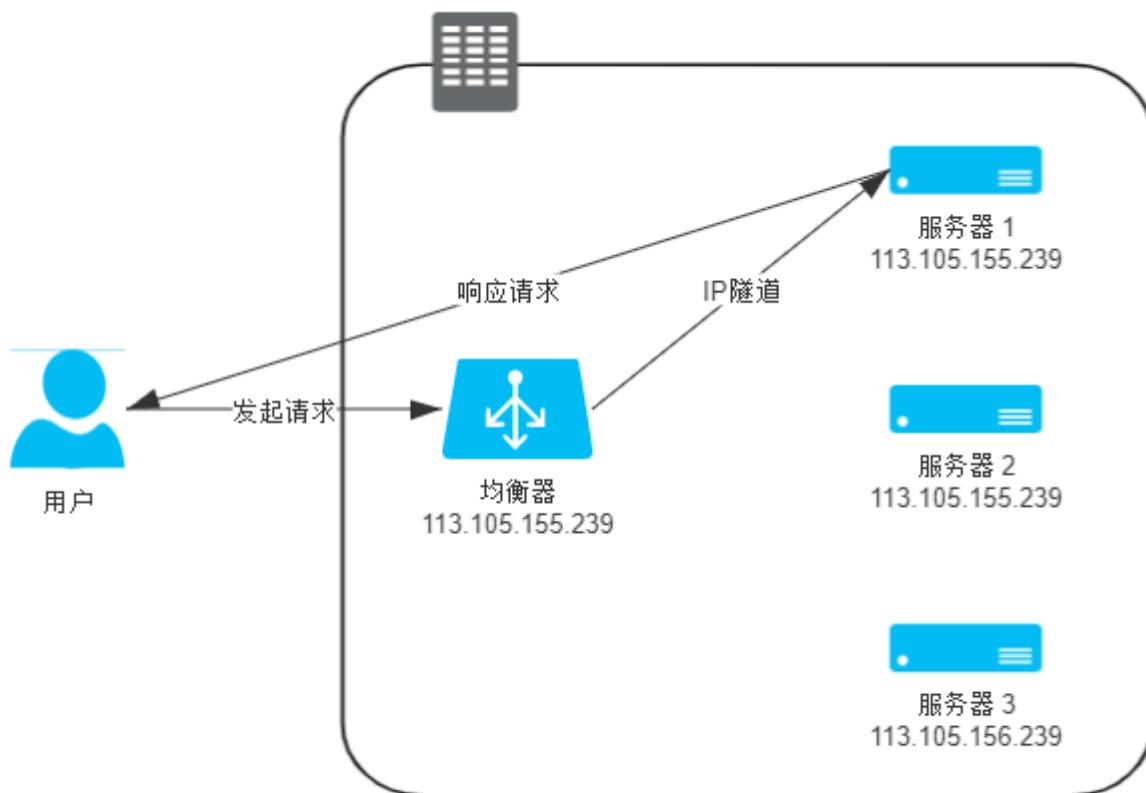
根据OSI七层模型，在第三层网络层传输的就是分组数据包（Packet），这是一种在分组交换网络（Packet Switching Network，PSN）中传输的结构化数据单位。以IP协议为例，一个IP数据包由Header和Payload两部分组成，其中Header长64 Bytes，其中包括了24 Bytes的固定数据和最长不超过40 Bytes的可选数据组成。按照IPv4标准，一个典型的分组数据包的Header部分如下表所示：

长度	存储信息
0-4 Bytes	版本号（4 Bits）、首部长度（4 Bits）、分区类型（8 Bits）、总长度（16 Bits）
5-8 Bytes	报文计数标识（16 Bits）、标志位（4 Bits）、片偏移（12 Bits）
9-12 Bytes	TTL生存时间（8 Bits）、上层协议代号（8 Bits）、首部校验和（16 Bits）
13-20 Bytes	源地址（32 Bits）
21-24 Bytes	目标地址（32 Bits）
25-64 Bytes	可选字段和空白填充

同样，我们对于表格中其他信息无需过多关心，只要知道在IP分组数据包的头部带有源和目标的IP地址即可。源和目标IP地址代表了数据是从分组交换网络中哪台机器发送到哪台机器，那我们就可以用之前同样的思路，通过改变这里面的地址来实现数据包的转发。有两种很容易想到的修改方式，其一是保持原来的数据包不变，新创建一个数据包，把原来数据包的Header和Payload整体作为另一个新的数据包的Payload，在这个新数据包的Header中写入真实服务器的IP作为目标地址，再把这个数据包发送出去。经过三层交换机的转发，真实服务器收到数据包后，要在接收入口处设计一个针对性的拆包机制，把由负载均衡器加入的那层Header扔掉，还原出原来的数据包来使用。这样，这台服务器就同样拿到了一个原本不是发给它（目标IP不是它）的数据包，达到了流量转发的目的。估计是那时

候还没有流行起“[禁止套娃](#)”的梗，所以设计者给这种“套娃式”的传输起名叫做“[IP隧道](#)”（ IP Tunnel ）传输，也还是相当的形象。

尽管由于要封装新的数据包，IP隧道的转发模式比起直接路由模式效率会有所下降，但由于并没有修改原有数据包中的任何信息，所以IP隧道的转发模式仍然具备三角传输的特性，即负载均衡器转发来的请求，可以由真实服务器直接应答，无需在经过均衡器原路返回。而且由于IP隧道工作在网络层，所以可以跨越VLAN，因此摆脱了直接路由模式中网络侧的约束。此模式从请求到响应的过程如下图所示：

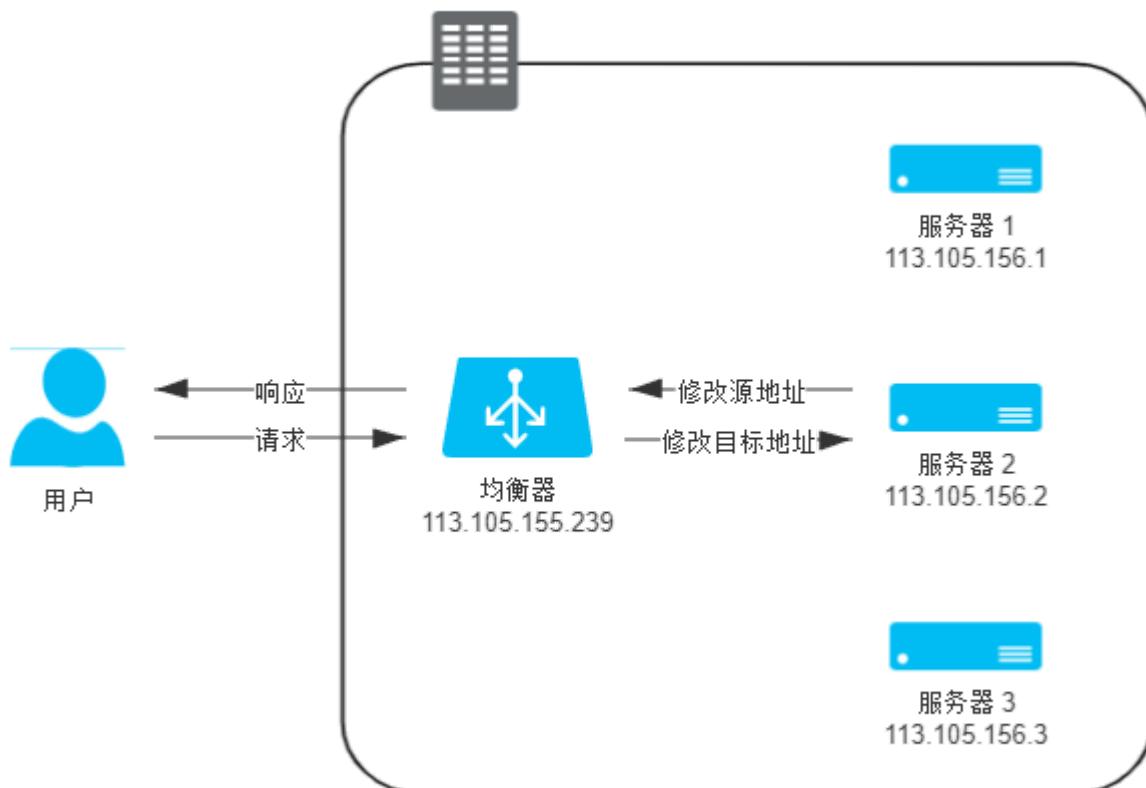


IP隧道模式的负载均衡

而这种方式的缺点是它要求真实服务器必须得支持[IP隧道协议](#)（ IP Encapsulation ）协议，就是它得会自己拆包扔掉一层Header，这个其实并非什么大问题，现在几乎所有的\*nix系统都支持IP隧道。而另外一个问题是这种模式仍必须通过专门的配置，保证所有的真实服务器与均衡器有着相同的虚拟IP地址，因为回复该数据包时，需要使用这个虚拟IP作为响应数据包的源地址，这样客户端收到这个数据包时才能正确解析。这个限制就比较讨厌了，它与“透明”这个原则有冲突。

对服务器进行虚拟IP的配置并不是在任何情况下都可行的，尤其是当有好几个服务共用一台物理服务器的时候。此时就需要考虑另一种改变目标数据包的方式：直接把数据包Head

er中的目标地址改掉。这样原本由用户发给均衡器的数据包，也会被三层交换机转发到真实服务器手上，而且因为没有经过IP隧道的额外包装，也就不再需要再拆包了。但现在问题是由于这种模式修改了目的IP地址才到达真实服务器的，如果真实服务器直接将应答包发回给客户端的话，这个应答数据包的源IP是真实服务器的IP，也即是均衡器修改后的那个IP地址，客户端肯定就无法正常处理这个应答。因此，只能让应答流量继续回到负载均衡，负载均衡把应答包的源IP改回自己的IP再发到客户端，这样才可以保证正常通信。如果你对网络知识有了解的话，肯定会觉得这种处理似曾相识，没错，这不就是在家里、公司、学校上网时，由一台路由器带着一群内网机器上网的“[网络地址转换](#)”（ Network Address Translation , NAT ）操作吗？此时，负载均衡器就是充当了家里、公司、学校的上网路由器的作用，所以，只要机器将自己的网关地址设置为均衡器地址，就无需再进行任何设置了。这种负载均衡的模式的确就被称为是NAT模式，此模式从请求到响应的过程如下图所示：



NAT模式的负载均衡

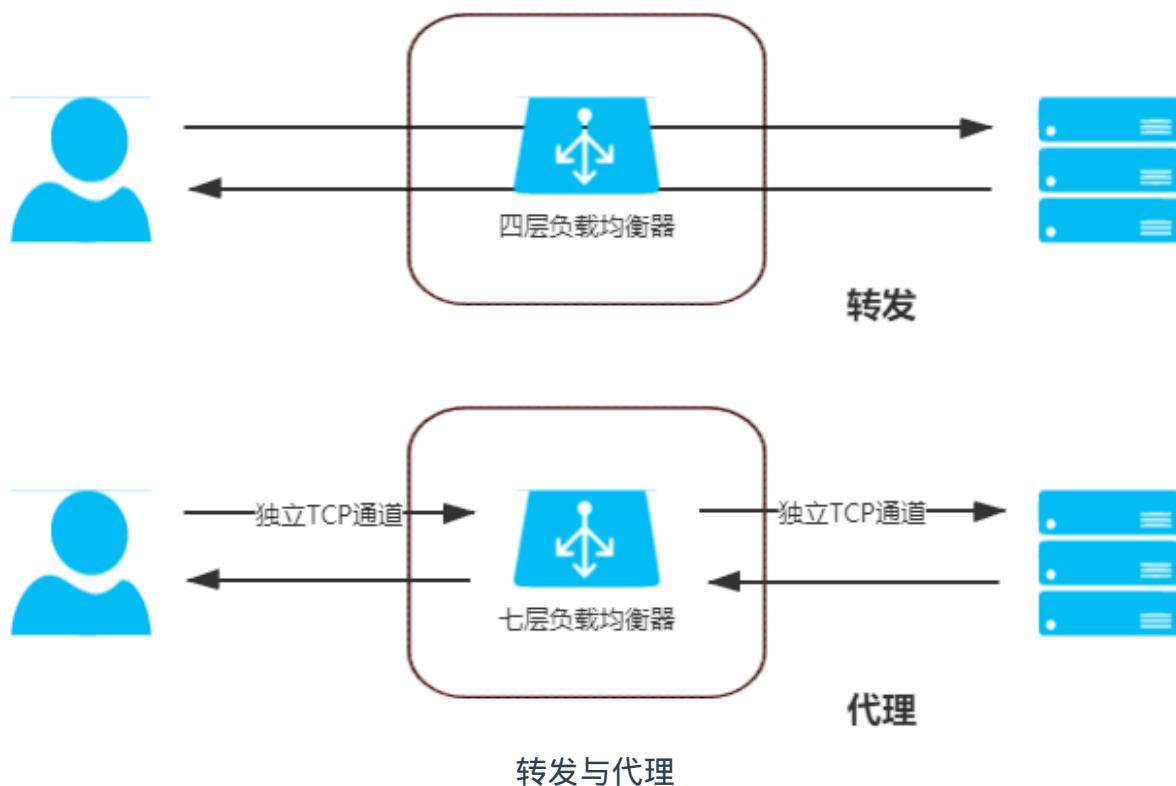
NAT模式的负载均衡会带来较大的（比起直接路由和IP隧道模式，往往是数量级上的下降）性能损失，这点是显而易见的，由负载均衡器代表整个服务集群来进行应答，各个服务器的响应数据都会互相争抢均衡器的出口带宽，这就好比在家里用NAT上网的话，如果

由人在下载，你打游戏可能就会觉得卡是一个道理，此时整个系统的瓶颈很容易就出现在负载均衡器上。

还有一种更加彻底的NAT模式，均衡器在转发时，不仅修改目标IP地址，连源IP地址也一起改了，源地址就改成均衡器自己的IP，称作Source NAT（SNAT）。这样好处是真实服务器连网关都无需配置了，可以让应答流量经过正常的三层路由回到负载均衡器上，做到了彻底的透明。但是缺点是由于做了SNAT，真实服务器处理请求时就无法拿到客户端的IP地址，在真实服务器的视角看来，所有的流量都来自于负载均衡器，这样有一些需要根据目标IP进行控制的业务逻辑就无法进行了。

## 应用层负载均衡

前面介绍的四层负载均衡工作模式都属于“转发”，即直接将承载着TCP报文的底层数据（IP数据包或以太帧）转发到真实服务器上，此时客户端到响应请求的真实服务器维持着同一条TCP通道。但工作在四层之后的负载均衡模式就无法再进行转发了，只能进行代理，转发与代理的区别如下图所示：



“代理”这个词，根据“哪一方能感知到”的原则，可以分为“正向代理”、“反向代理”和“透明代理”三类。正向代理就是我们通常简称的代理，指在客户端设置的、代表客户端与服务器通

讯的代理服务，它是客户端可知，而对服务器透明的。反向代理是指在设置在服务器这一侧，代表真实服务器来与客户端通讯的代理服务，此时它对客户端来说是透明的。至于透明代理是指对双方都透明的，配置在网络中间设备上（譬如，架设在路由器上的透明翻墙代理）的代理服务。

根据以上定义，很显然，七层负载均衡器它就属于一种反向代理，如果只论网络性能，七层均衡器肯定是无论如何比不过四层均衡器的，它比四层均衡器至少多一轮TCP握手，有着跟NAT模式一样的带宽问题，而且通常要耗费更多的CPU（因为可用的解析规则远比四层丰富），所以如果用七层均衡器去做下载站、视频站这种流量应用是不合适的（起码不能作为第一级均衡器）。但是，如果网站的性能短板并不在于网络性能，要论整个服务集群对外所体现出来的服务性能，七层均衡器就有它的用武之地了。这里面七层均衡器的底气就是来源于它工作在应用层，可以感知应用层通讯的内容，往往能够做出更明智的决策，玩出更多的花样来。

举个生活中的例子，四层均衡器就像银行的自助排号机，转发效率高且不知疲倦，每一个达到银行的客户根据排号机的顺序，选择对应的窗口接受服务；而七层均衡器就像银行大堂经理，他会先确认客户需要办理的业务，再安排排号。这样办理理财、存取款等业务的客户，会根据银行内部资源得到统一协调处理，加快客户业务办理流程，有一些无需柜台办理的业务，甚至大堂经理直接就可以解决了（譬如，反向代理的静态资源缓存）。

相信代理的工作模式大家应该是比较熟悉的，因此关于七层均衡器如何工作的就不作详细介绍了，笔者列举了一些七层代理可以实现的功能，以便读者对它“功能强大”有个直观的感受：

- 前面介绍CDN应用时，所有CDN可以做的缓存类工作（就是除去CDN就近返回这种优化链路的工作外），七层均衡器都可以实现，譬如静态资源缓存、协议升级、安全防护、访问控制，等等。
- 七层均衡器可以实现更智能化的路由。譬如，根据Session路由，以实现亲和性的集群；根据URL路由，实现专职化服务（如Kubernetes Ingress均衡器就属于这类）；甚至根据用户身份路由，实现对部分用户的特殊服务（如某些站点的贵宾服务器），等等。
- 某些安全攻击可以由七层均衡器来解决，譬如一种常见的DDoS手段是SYN Flood攻击，即攻击者控制众多客户端，使用虚假IP地址对同一目标大量发送SYN报文。从技术原理上看，由于四层均衡器无法感知上层协议的内容，这些SYN攻击都会被转发到后端

的真实服务器上；而七层均衡器下这些SYN攻击自然在负载均衡设备上就截止，不会影响到后面服务器的正常运行。类似地，可以在七层均衡器上设定多种策略，譬如过滤特定报文，以防御如SQL注入等应用层面的特定攻击手段。

- 多数微服务中的链路治理措施，都需要在七层中进行，譬如服务降级、熔断、异常注入等等。譬如，一台服务器只有出现物理层面或者系统层面的故障，导致无法应答TCP请求才能被四层均衡器所感知，进而剔除出服务集群，如果一台服务器一直在报500错，那四层均衡器对此是完全无能力的，只能由七层均衡器来解决。

## 进程内负载均衡

在七层负载均衡中，有另外一类非独立的负载均衡模式，在这轮架构风格的演变中占据了越来越重要的作用。所谓“非独立”即不在独立的反向代理服务器来调度请求，而是由真实服务器的服务进程内自己实现负载均衡。这种典型的应用便是Spring Cloud Netflix全家桶中的Ribbon。与此前谈到的独立于具体服务逻辑的均衡器不同，进程内负载均衡（也被称为客户端负载均衡）是与具体服务逻辑相关的，它要到服务注册中心中找到多个可以完成该业务的微服务Endpoint地址，然后根据自己的均衡逻辑选择其中一个合适的来响应请求。这部分我们将在微服务核心技术支撑点中的“[进程内负载均衡](#)”再详细介绍，在此就不去涉及。

## 均衡器实现与策略

负载均衡的两大职责是“选择谁来处理用户请求”和“将用户请求转发过去”。到此我们仅介绍了后者，即请求的转发或代理过程。前者是指均衡器所采取的均衡策略，这一块较为接近于实现细节，笔者就不展开了，常见的均衡策略有：

- **轮循均衡** ( Round Robin )：每一次来自网络的请求轮流分配给内部中的服务器，从1至N然后重新开始。此种均衡算法适合于服务器组中的所有服务器都有相同的软硬件配置并且平均服务请求相对均衡的情况。
- **权重轮循均衡** ( Weighted Round Robin )：根据服务器的不同处理能力，给每个服务器分配不同的权值，使其能够接受相应权值数的服务请求。譬如：服务器A的权值被设计成1，B的权值是3，C的权值是6，则服务器A、B、C将分别接受到10%、30%、60%的服务请求。此种均衡算法能确保高性能的服务器得到更多的使用率，避免低性能的服务器负载过重。

- **随机均衡** ( Random ) : 把来自客户端的请求随机分配给内部中的多个服务器，在数据足够大的场景能达到一个均衡分布。
- **权重随机均衡** ( Weighted Random ) : 此种均衡算法类似于权重轮循算法，不过在处理请求分担时是个随机选择的过程。
- **一致性哈希均衡** ( Consistency Hash ) : 根据请求中某一些数据（可以是MAC、IP地址，也可以是高层协议中的某些信息）作为特征值来计算需要落在的结点上，可以保证一个同一个特征值一定落在相同的服务器上。一致性的意思是保证当服务集群某个真实服务器出现故障，只影响该服务器的哈希，而不会导致整个服务集群的哈希键值重新分布。
- **响应速度均衡** ( Response Time ) : 负载均衡设备对内部各服务器发出一个探测请求（例如Ping），然后根据内部中各服务器对探测请求的最快响应时间来决定哪一台服务器来响应客户端的服务请求。此种均衡算法能较好的反映服务器的当前运行状态，但这最快响应时间仅仅指的是负载均衡设备与服务器间的最快响应时间，而不是客户端与服务器间的最快响应时间。
- **最少连接数均衡** ( Least Connection ) : 客户端的每一次请求服务在服务器停留的时间可能会有较大的差异，随着工作时间加长，如果采用简单的轮循或随机均衡算法，每一台服务器上的连接进程可能会产生极大的不同，并没有达到真正的负载均衡。最少连接数均衡算法对内部中需负载的每一台服务器都有一个数据记录，记录当前该服务器正在处理的连接数量，当有新的服务连接请求时，将把当前请求分配给连接数最少的服务器，使均衡更加符合实际情况，负载更加均衡。此种均衡策略适合长时处理的请求服务，如FTP。
- .....

负载均衡器的实现有“软件均衡器”和“硬件均衡器”两类。在软件均衡器方面，又分为直接建设在操作系统内核的均衡器和应用程序形式的均衡器两种。前者的代表是LVS ( Linux Virtual Server )，后者的代表有Nginx、HAProxy、KeepAlived等，前者性能会更好（数据包从网卡开始到达应用为止这段路径，会经过层层的协议处理和运行钩子的过程），后者选择广泛，使用方便。在硬件均衡器方面，往往会直接采用[应用专用集成电路](#) ( Application Specific Integrated Circuit , ASIC ) 来实现，有专用处理芯片的支持，避免操作系统层面的损耗，得以达到最高的性能。这类的代表是F5和A10公司的负载均衡产品。

# 缓存中间件

编写中

## 缓存中间件 (Cache Middleware)

讨论数据缓存、方法缓存、进程内/外、集中式/分布式缓存等等。

# 数据库扩展

读写分离

分片

双主架构

数据库代理

# 安全架构

即使只限定在“软件架构设计”这个语境下，系统安全仍然是一个很大的话题。我们谈论的计算机系统安全，远不仅指“防御系统被黑客攻击”这样狭隘的“安全”。架构安全性至少应包括了（不限于）以下这些问题的具体解决方案：

- **认证** (Authentication)：系统如何正确分辨出操作用户的真实身份？
- **授权** (Authorization)：系统如何控制一个用户该看到哪些数据、能操作哪些功能？
- **凭证** (Credentials)：系统如何保证它与用户之间的承诺是双方当时真实意图的体现，是准确、完整且不可抵赖的？
- **保密** (Confidentiality)：系统如何保证敏感数据无法被包括系统管理员在内的内外部人员所窃取、滥用？
- **传输** (Transport Security)：系统如何保证通过网络传输的信息无法被第三方窃听、篡改和冒充？
- **验证** (Verification)：系统如何确保提交到每项服务中的数据是合乎规则的，不会对系统稳定性、数据一致性、正确性产生风险？
- **漏洞利用** (Exploit) 编写中：系统如何避免在基础设施和应用程序中出现弱点，被攻击者利用？
- .....

上面这些安全相关的问题，解决起来确实是既繁琐复杂，又难以或缺。值得庆幸的是这一部分内容基本上都是与具体系统、具体业务无关的通用性问题、这意味着它们会存在着业界通行的，已被验证过是行之有效的解决方案，乃至已经形成某一些行业标准，不需要我们自己从头去构思如何解决。后面我们将会通过标准的方案，逐一探讨以上问题的主流处理方法。

还有其他一些安全相关的内容，主要由管理、运维、审计方面负责，尽管软件架构也需要配合参与，但不列入本文的讨论范围之中，譬如：安全审计、系统备份与恢复、防治病毒、信息系统安全法规与制度、计算机防病毒制度、保护私有信息规则，等等。



# 认证

## 认证 (Authentication)

系统如何正确分辨出操作用户的真实身份？

“认证”可以说是一个系统中最基础的安全设计，再简陋的系统大概也不大可能省略掉“用户登录”功能。但“认证”这件事情又并不如大多数人所认为的那样，校验一下用户名、密码是否正确这么简单。尤其是在基于Java的软件系统里，尝试去接触了解Java安全标准的人往往会对一些今天看起来很别扭的概念产生疑惑。在这一部分，将简要概览一下关于认证的主流行业规范、标准；项目中具体如何认证、授权的内容放到下一节去介绍。

最初的Java系统里，安全中的“认证”其实是特指“代码级安全”（你是否信任要在你的电脑中运行的代码），这是由“Java 2”之前它的主要应用形式Applets所决定的：从远端下载一段Java代码，以Applet的形式在用户的浏览器中运行，当然要保证这些代码不会损害用户的计算机才行。这一阶段的安全催生了今天仍然存在于Java体系中的“安全管理器”（java.lang.SecurityManager）、“代码权限许可”（java.lang.RuntimePermission）这些概念。

不久之后，Java迎来了互联网的迅速兴起，进入了Java第一次快速发展时期，基于超文本的Web应用迅速盖过了“Java 2”时代之前的Applet，此时“安全认证”的重点逐渐转为“用户级安全”（你是否信任正在操作的用户）。在1999年随着J2EE 1.2（它是J2EE的首个版本，版本号直接就是1.2）所发布的Servlet 2.2中增加了一系列认证的API，诸如：

- HttpServletRequest.isUserInRole()
- HttpServletRequest.getUserPrincipal()
- 还内置支持了四种硬编码、不可扩展的认证机制：BASIC、FORM、CLIENT-CERT和DIGEST。

到Java 1.3时代中，Sun公司提出了同时面向与代码级安全和用户级安全的认证授权服务JAAS（Java Authentication and Authorization Service，1.3处于扩展包中，1.4纳入标准包），不过相对而言，在JAAS中代码级安全仍然是占更主要的地位。

由于用户数据可能来自于各种不同的数据源（譬如RDBMS、JNDI、LDAP等等），JAAS设计了一种插入式（Pluggable）的认证和授权模型，以适配各种环境。在今天仍然活跃的主流安全框架中的许多概念，譬如用户叫做“Subject / Principal”、密码存在“Credentials”之中、登陆后从安全上下文“Context”中获取状态等都可以追溯到这一时期所设计的API：

- LoginModule ( javax.security.auth.spi.LoginModule )
- LoginContext ( javax.security.auth.login.LoginContext )
- Subject ( javax.security.auth.Subject )
- Principal ( java.security.Principal )
- Credentials ( javax.security.auth.Destroyable、 javax.security.auth.Refreshable )

但是，尽管JAAS开创了许多沿用至今的安全概念，实质上并没有得到广泛的应用。这里有两大原因，一方面是由于JAAS同时面向代码级和用户级的安全机制，使得它过度复杂化，难以推广。在这里问题上JCP一直在做着持续的增强和补救，譬如Java EE 6中的JASPI C、Java EE 8中的EE Security：

- JSR 115 : Java Authorization Contract for Containers [↗](#) ( JACC )
- JSR 196 : Java Authentication Service Provider Interface for Containers [↗](#) ( JASPI )
- JSR 375 : Java EE Security API [↗](#) ( EE Security )

而另一方面，可能是更重要的一个原因是在21世纪的第一个十年里，以EJB为代表的容器化J2EE与以“Without EJB”为口号、以Spring、Hibernate等为代表的轻量化企业级开发框架之争，以后者的胜利而结束。这也使得依赖于容器安全的JAAS无法得到大多数人的认可。

在今时今日，实际活跃于Java届的两大私有的（私有的意思是不由JSR所规范的，即没有java/javax.\*作为包名的）的安全框架：

- Apache Shiro [↗](#)
- Spring Security [↗](#)

相较而言，Shiro使用更为便捷易用，而Spring Security的功能则要复杂强大一些。在我们的项目中（无论是单体架构还是微服务架构），均选择了Spring Security作为安全框架。当然，这里面也有很大一部分是因为Spring Boot/Cloud全家桶的原因。这两大安全框架都解决的问题都很类似，大致可以分为四类：

- 认证：以HTTP协议中定义的各种认证、表单等认证方式确认用户身份，这是本节的主要话题。
- 授权：主要是授权结果，即访问控制（Access Control），稍后讲的“授权”将聚焦在授权的过程，尤其是多方授权中。这部分内容会放到下一节一起讨论。
- 密码的存储：就是字面意思，我们会放到“保密”这节去一起讨论。
- 安全上下文：用户获得认证之后，需要有API可以得知该用户的基本资料、用户拥有的权限、角色等。

介绍了一大段关于Java中安全标准的历史，我们最终还是要切入到如何处理认证的话题上，这可是随着网络出现就有的一个东西，所以，IETF的最初想法是基于Web的验证就应该在HTTP协议层面来解决。

### 互联网工程任务组（Internet Engineering Task Force，IETF）

管理和发布互联网标准的组织，其标准以RFC即“征求意见稿”Request for Comments的形式发出。不仅是HTTP，几乎目前所有的主要网络协议，如IP、TCP、UDP、FTP、CMI P、SOCKS，等等都是以RFC形式定义的。

IETF给HTTP 1.1协议定义了401（Unauthorized，未授权）状态码，当服务端向客户端返回此状态码时，应在Header中附带一个WWW-Authenticate项，此项目通过跟随的一个可扩展的Schema，告诉客户端应该采取怎样的方式来开始验证，例如：

```
HTTP/1.1 401 Unauthorized
Date: Mon, 24 Feb 2020 16:50:53 GMT
WWW-Authenticate: Basic realm="From icyfenix.cn"
```

同时，IETF也定义了几种标准的Schema，对应了一些预定义好的认证方式，包括：

- **Basic**：[RFC 7617](#)，HTTP基础认证，弹出一个输入框，把用户名和密码Base64之后发送出去
- **Digest**：[RFC 7616](#)，HTTP摘要认证，弹出一个输入框，把用户名和密码加盐后再通过MD5/SHA等哈希算法摘要后发送出去
- **Bearer**：[RFC 6750](#)，OAuth 2.0令牌（OAuth2是一个授权协议，但同时也涉及到认证的内容，下一节的主角）

- **HOBA** : RFC 7486 [↗](#) , HTTP Origin-Bound Authentication的缩写，一种基于数字签名的认证。

因为Scheme是允许自定义扩展的，很多厂商也加入了自己的认证方式，譬如：

- **AWS4-HMAC-SHA256** : 简单粗暴的名字，一看就是亚马逊AWS基于HMAC-SHA256哈希算法的认证
- **NTLM / Negotiate** : 微软公司NT LAN Manager ( NTLM ) 用到的两种认证方式
- **Windows Live ID** : 这个不需要解释了
- **Twitter Basic** : 一个不存在的网站所改良的HTTP基础认证
- .....

现在主流的信息系统，直接采用上面这些认证方式比例不算太高，目前的主流仍是Form表单认证，即我们通常所说的“登陆页面”。表单认证并没有什么行业标准可循，表单中的用户字段、密码字段、验证码字段、是否要在客户端加密、加密的方式、接受表单的服务入口等都可由服务端、客户端自行协商决定。

在Fenix's Bookstore项目中，我们所设计的登录实质上也是一种表单认证，借用了Spring Security的认证管理器。Spring Security中提供了默认的登陆表单界面和配套的服务，只要在Spring Security的Web安全中简单配置即可启用：

```
java
@Configuration
@EnableWebSecurity
public class WebSecurityConfig extends WebSecurityConfigurerAdapter {
 @Override
 protected void configure(HttpSecurity http) throws Exception {
 http.authorizeRequests()
 .antMatchers("/").permitAll() // 首页地址'/'的请求都允许访问
 .anyRequest().authenticated() // 任何请求, 登录后才可以访问
 .and()
 .formLogin() // 启用表单登录认证, 还有另一种httpBasic()方法
 代表了HTTP基础认证
 .permitAll(); // 登录页面用户任意访问
 .and()
 .logout().permitAll(); // 注销的服务任意访问
 }
}
```

Spring Security的权限控制措施在两个层面进行，一种Web级别的访问控制，这是在Web服务器中附加的过滤器（FilterSecurityInterceptor）实现的，另一种是方法级权限控制，是通过动态代理实现的。第二种将在下一节“授权”部分中提及，这里先来说第一种。

当Spring Security被启动时（在Spring Boot中通过@EnableWebSecurity注解启动），将会在Web服务器中附加十几个不同作用的过滤器，譬如上面代码就直接涉及到其中三个：

- SecurityContextPersistenceFilter：用于维护安全上下文，“上下文”说白了就是如果用户登陆了系统，那服务的代码中总该有个地方可以取到当前登陆用户是谁这类信息
- UsernamePasswordAuthenticationFilter：用于完成用户名、密码的验证过程
- LogoutFilter：用于注销
- FilterSecurityInterceptor：用于Web级别的访问控制，如果设置了指定地址需要登陆而实际未登陆，或者设定了需要某些权限才能访问而实际用户并没有，那将抛出AuthenticationException与AccessDeniedException异常

让我们再回到上面的代码，这段简单的工作流程是：

1. 启用过滤器UsernamePasswordAuthenticationFilter，在其attemptAuthentication()方法中，会从Request中获取用户名和密码，传递给认证管理器AuthenticationManager的authenticate()方法
2. 认证管理器的目的是协调不同的用户来源，譬如来自数据库、来自LDAP、来自OAuth等等，每一个用户来源都应该有一个实现了AuthenticationProvider接口并注册到认证管理器的实现类所代表，认证管理器将根据需要，调用对应Provider的authenticate()方法实际完成认证操作。
3. Spring Security默认的Provider是DaoAuthenticationProvider，它在Bookstore项目中并未被采用，而是另外实现了一个UsernamePasswordAuthenticationProvider。但是两者实际逻辑是相似的，都是调用UserDetailsService接口里的loadUserByUsername()来获取用户信息，UserDetailsService是读取用户明细数据的接口，Spring Security并不关心用户系统的实际存储结构，但认证时肯定也必须使用到用户信息，默认使用InMemoryUserDetailsManager，也就是从内存中写死一些用户数据来完成。
4. 在AuthenticationProvider中比较传入的用户密码与数据库中的用户密码是否一致（具体怎么个比较法将在“保密”这一节中说明），返回结果，完成认证。

以上流程是大多数系统，尤其是单体系统中主流的认证方式，哪怕不基于Apache Shiro或Spring Security来实现，其思路很可能也是与上面描述的差不多的。但我们的Bookstore却

并未直接应用这种认证方式，而是借用了OAuth2授权协议中的密码授权模式，在此过程中完成认证。为何会选择这种方式，以及具体实现部分的内容，将在下一部分“授权”中继续介绍。

# 授权

## 授权 ( Authorization )

系统如何控制一个用户该看到哪些数据、能操作哪些功能？

“授权”这个行为通常伴随着“认证”、“账号”共同出现，并称为AAA ( Authentication、Authorization、Account，也有把Account理解为计费的意思 )。授权行为在程序中其实非常普遍，我们给一个类、一个方法设置范围控制符 ( public、protected、private、<Package >)，这其实也是一种授权 ( 访问控制 ) 行为。授权涉及到了两个相对独立的问题：

- 确保授权的过程可靠：对于单一系统来说，授权的过程是比较容易做到可控的，以前很多语境上提到授权，实质上讲的都是访问控制，理论上两者是应该分开的。而在涉及多方的系统中，授权过程就是一个必须严肃对待的问题：如何既让第三方系统能够访问到所需的资源，又能保证其不泄露用户的敏感数据？现在常用的多方授权协议主要有OAuth2和SAML 2.0（注意这两个协议涵盖的功能并不是直接对等的）。
- 确保授权的结果可控：授权的结果往往是用于对程序功能或者资源的访问控制 ( Access Control )，形成理论的权限控制模型有：自主访问控制 ( Discretionary Access Control , DAC )、强制访问控制 ( Mandatory Access Control , MAC )、基于属性的权限验证 ( Attribute-Based Access Control , ABAC ) 还有最为常用，也相对通用的是基于角色的权限模型 ( Role-Based Access Control , RBAC )。

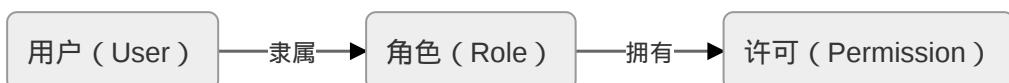
由于篇幅原因，在这个小节里我们只介绍 ( 将要 ) 使用到的，也是最常用到的RBAC和OAuth2。先来说较为简单的RBAC。

## RBAC

所有的访问控制模型，实质上都是在解决同一个问题：“谁 ( User ) ”拥有什么“权限 ( Authority ) ”去操作哪些“资源 ( Resource ) ”

这个问题看起来并不难，最直观的解决方案就是在用户对象上，设定一些操作权限，在使用资源时，检查是否有对应的操作权限即可。是的，请不要因太过简单直接而产生疑惑——Spring Security的访问控制本质上就是这么做的。不过，这种把操作权限直接关联在用户身上的简单设计，在复杂系统上确实会导致比较繁琐的操作。试想一下，如果某个系统涉及到成百上千的资源，又有成千上万的用户，要为每个用户分配合适的权限将带来务必庞大的操作量和极高的出错概率，这也即是RBAC所要解决的问题。

为了避免对每一个用户设定权限，RBAC将权限从用户身上剥离，改为绑定到“**角色（Role）**”上，一种我们常见的RBAC应用就是操作系统权限中的“用户组”，这就是一种角色。用户可以隶属与一个或者多个角色，某个角色中也会包含有多个用户，角色之间还可以有继承性（父、子角色的权限继承，RBAC1）。这样，资源的操作就只需按照有限且相对固定的角色去分配操作权限，而不去面对随时会动态增加的用户去分配。当用户的职责发生变化时，在系统中就体现为改变他所隶属的角色，譬如将“普通用户角色”改变“管理员角色”，就可以迅速完成其权限的调整，降低了权限分配错误的风险。RBAC的主要元素之间的关系可以以下图来表示：



上图中出现了一个新的名词“**许可（Permission）**”。所谓的许可，就是抽象权限的具体化体现。权限在系统中的含义应该是“允许何种**操作**作用于哪些**数据**之上”，这个即为“许可”。举个具体的例子，譬如某个文章管理系统的UserStory中，与访问控制相关的Backlog可能会是这样描述的：

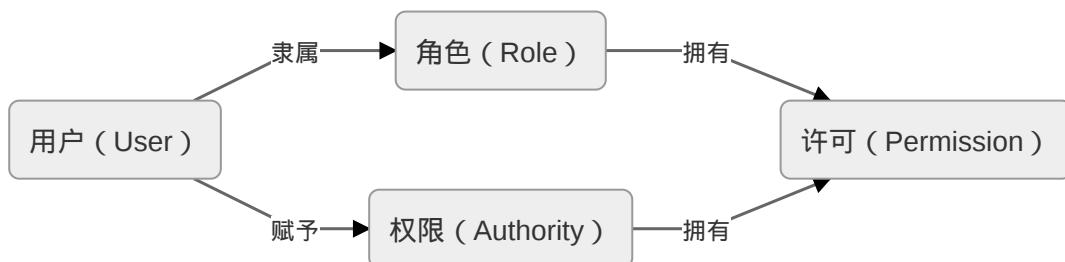
#### Backlog :

周同学（User）是某SCI杂志的审稿人（Role），职责之一是在系统中审核论文（Authority）。在审稿过程（Session）中，当他认为某篇论文（Resource）达到了可以公开发表标准时，就会在后台点击通过按钮（Operation）来完成审核。

以上，“给论文点击通过按钮”就是一种许可（Permission），它是“审核论文”这项权限（Authority）的具体化体现。

与微服务架构中的完全遵循RBAC进行访问控制的Kubernetes不同，我们在单体架构中使用的Spring Security参考了但并没有完全按照RBAC来进行设计。Spring Security的设计里用户和角色都可以拥有权限，譬如在HttpSecurity对象上，就同时有着hasRole()和hasAuth

ority()方法，可能有不少刚接触的人会疑惑，混淆它们之间的关系。在Spring Security的访问控制模型可以认为是下图所示这样的：



站在代码实现的角度来看，Spring Security中Role和Authority的差异很小，它们共同存储在同一位置，唯一的差别仅是Role会在存储时自动带上“ROLE\_”前缀（可以配置的）罢了。

但在使用者的角度来看，Role和Authority的差异可以很大，你可以执行决定你的系统中到底Permission只能对应到角色身上，还是可以让用户也拥有某些角色中没有的权限。这个观点，在Spring Security自己的文档上说的很清楚：这取决于你自己如何使用。

**The core difference between these two is the semantics we attach to how we use the feature.** For the framework, the difference is minimal – and it basically deals with these in exactly the same way.

使用RBAC，你可以控制最终用户在广义和精细级别上可以做什么。您可以指定用户是管理员，专家用户还是普通用户，并使角色和访问权限与组织中员工的身份职位保持一致。仅根据需要为员工完成工作的足够访问权限来分配权限。

## OAuth2

简要介绍过RBAC，下面我们再来看看相对要复杂繁琐写的OAuth2授权协议（顺带说一下，OAuth1.0已经完全废弃了）。先明确一件事情，OAuth2是一个多方系统中的授权协议，如果你的系统并不涉及到第三方（譬如我们单体架构的Bookstore，即不为第三方提供服务，也不使用第三方的服务），引入OAuth2其实并无必要。我们之所以把OAuth2提前引入，主要是为了给微服务架构做铺垫。

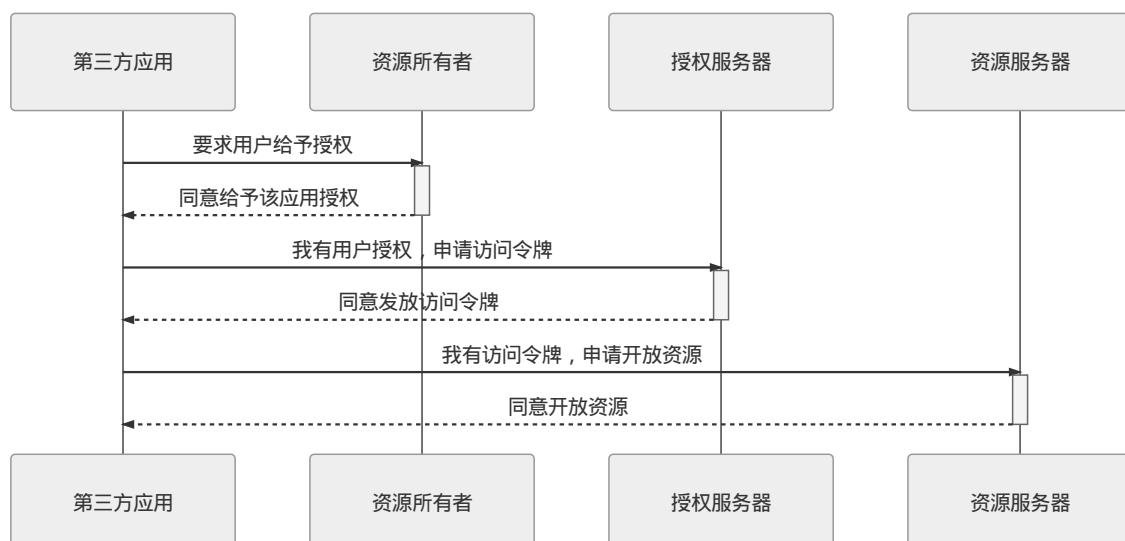
OAuth2是在[RFC 6749](#)中定义授权协议，在RFC 6749正文的第一句就明确了OAuth2是解决第三方应用（Third-Party Application）的授权协议。前面也说到，如果只是单方系

统，授权过程是比较容易解决的，至于多方系统授权过程会有什么问题，这里举个现实的例子来说明。

譬如你现在正在阅读的这个网站（<https://icyfenix.cn>），它的建设和更新大致流程是：笔者以Markdown形式写好了某篇文章，上传到由GitHub提供的代码仓库，接着由Travis-CI提供的持续集成服务会检测到该仓库发生了变化，触发一次Vuepress编译活动，生成目录和静态的HTML页面，然后推送回GitHub Pages，再触发腾讯云CDN的缓存刷新。这个过程要能顺利进行，就存在一些必须解决的授权问题，Travis-CI只有得到了我的明确授权，GitHub才能同意它读取我代码仓库中的内容，问题是它该如何获得我的授权呢？一种简单粗暴的方案是我把我的用户账号和密码都告诉Travis-CI，但这显然导致了以下这些问题：

- 密码泄漏**：如果Travis-CI被黑客攻破，将导致我GitHub的密码也同时被泄漏
- 访问范围**：Travis-CI将有能力读取、修改、删除、更新我放在GitHub上的所有代码仓库
- 授权回收**：我只有修改密码才能回收授予给Travis-CI的权力，可是我在GitHub的密码只有一个，修改了意味着所有别的第三方的应用程序会全部失效

以上出现的这些问题，也就是OAuth2所要解决的问题，尤其是没有HTTPS支持传输安全的环境下依然可以解决这些问题。OAuth2提出的解决办法是通过一个令牌（Token）代替用户密码作为授权的凭证，有了令牌之后，哪怕令牌被泄漏，也不会导致密码的泄漏，令牌上可以设定访问资源的范围以及时效性，每个应用都持有独立的令牌，哪个失效都不会波及其他，一下子上面提出的三个问题都解决了，有了一层令牌之后，整个授权的流程如下图所示：



这个时序图里面涉及到了OAuth2中几个关键术语，我们通过前面那个具体的上下文语境来解释其含义，这对理解后续几种认证流程十分重要：

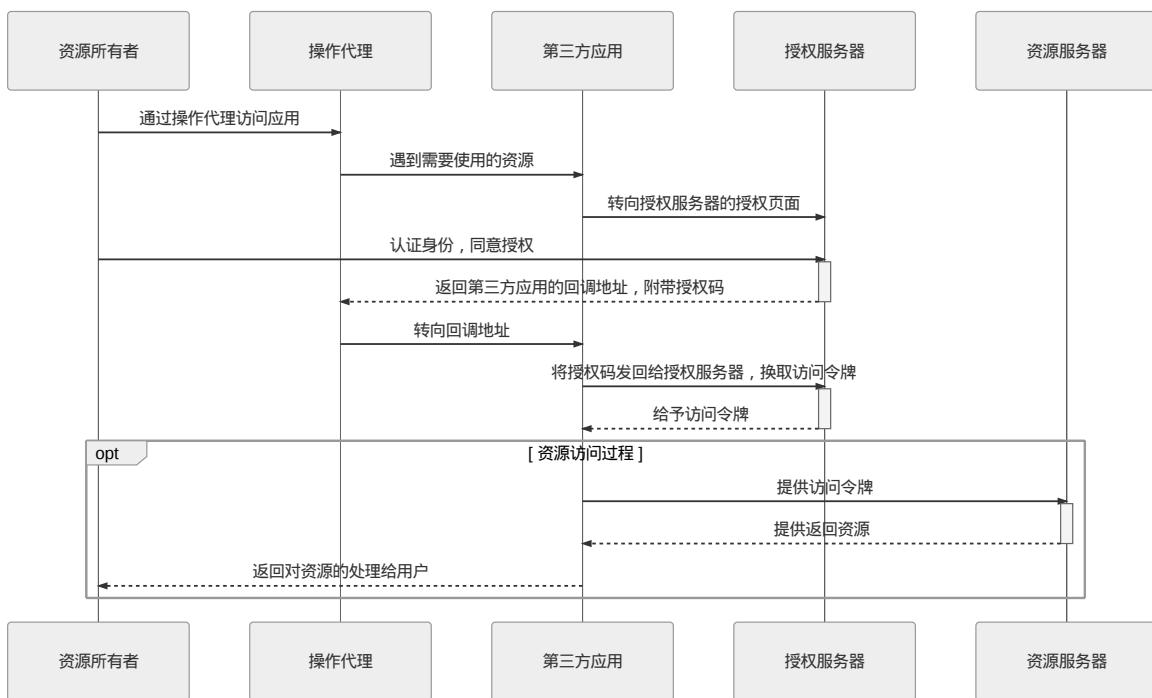
- **第三方应用** ( Third-Party Application )：需要得到授权访问我资源的那个应用，即“Travis-CI”
- **授权服务器** ( Authorization Server )：能够根据我的意愿提供授权（授权之前肯定已经进行了必要的认证过程，但这在技术上与授权可以没有直接关系）的服务，即“GitHub”
- **资源服务器** ( Resource Server )：能够提供第三方应用所需资源的服务（它与认证服务可以是相同的服务器，也可以是不同的服务器），即“代码仓库”
- **资源所有者** ( Resource Owner )：拥有授权权限的人，这里即是“我”
- **操作代理** ( User Agent )：指用户用来访问服务器的工具，对于指代人类的“用户”来说这个通常就是浏览器，但在微服务中一个服务经常会作为另一个服务的“用户”，此时指的可能就是HttpClient、RPCClient或者其他访问途径。

看来“用令牌代替密码”确实是解决问题的好方法，但这最多只能算个思路，距离执行步骤还是不够具体的，时序图中的“要求/同意授权”、“要求/同意发放令牌”、“要求/同意开放资源”几个服务请求、响应应该如何设计，这就是执行步骤的关键了。对此，OAuth2一共提出了四种不同的授权方式（这就是我说OAuth2复杂繁琐的原因，摊手），分别为：

- 授权码模式 ( Authorization Code )
- 简化模式 ( Implicit )
- 密码模式 ( Resource Owner Password Credentials )
- 客户端模式 ( Client Credentials )

## 授权码模式

授权码模式是四种模式中最严谨（繁琐）的，它考虑到了几乎所有敏感信息泄漏的预防和后果。具体步骤的时序如下：



在开始完成整个授权过程以前，第三方应用先要到授权服务器上进行注册，所谓注册，是指向认证服务器提供一个域名地址，从授权服务器中获取ClientID和ClientSecret，然后便可以开始如下授权过程：

1. 第三方应用将资源所有者（用户）导向授权服务器的授权页面，并向授权服务器提供ClientID及同意授权后的回调URI，这是一次客户端页面转向。
2. 授权服务器根据ClientID确认第三方应用的身份，用户在授权服务器中决定是否同意向该身份的应用进行授权（认证的过程在此之前应该已经完成）。
3. 如果用户同意授权，授权服务器将转向第三方应用在第1步调用中提供的回调地址URI，并附带上一个授权码和获取令牌的地址作为参数，这也是一次客户端页面转向。
4. 第三方应用通过回调地址收到授权码，然后将授权码与自己的ClientSecret一起作为参数，**通过服务端**向授权服务器提供的获取令牌的服务地址发起请求，换取令牌。该服务端应与注册时提供的域名一致。
5. 授权服务器核对授权码和ClientSecret，确认无误后，向第三方应用授予令牌。令牌可以是一个或者两个，其中必定要有的是访问令牌（Access Token），可选的是刷新令牌（Refresh Token）。访问令牌用于到资源服务器获取资源，有效期较短，刷新令牌用于在访问令牌失效后重新获取，有效期较长。
6. 资源服务器根据访问令牌所允许的权限，向第三方应用提供资源。

这个过程设计，已经考虑到了几乎所有合理的意外情况，举例几个容易想到的：

- 会不会有其他应用冒充第三方应用骗取授权？

ClientID代表一个第三方应用的“用户名”，这个是可以完全公开的。但ClientSecret应当只有应用自己才知道，这个代表了第三方应用的“密码”。在第5步发放令牌时，调用者必须能够提供ClientSecret才能成功完成。只要第三方应用妥善保管好ClientSecret，就没有人能够冒充它。

- 为什么要先发放授权码，再用授权码换令牌？

这是因为客户端转向（通常就是一次HTTP 302重定向）对于用户是可见的，换而言之，授权码完全可能会暴露给用户（以及用户机器上的其他程序），但由于用户并没有ClientSecret，光有授权码也是无法换取到令牌的，所以避免了令牌在传输转向过程中泄漏的风险。

- 为什么要设计一个时限较长的刷新令牌和时限较短的访问令牌？不能直接把访问令牌的时间调长吗？

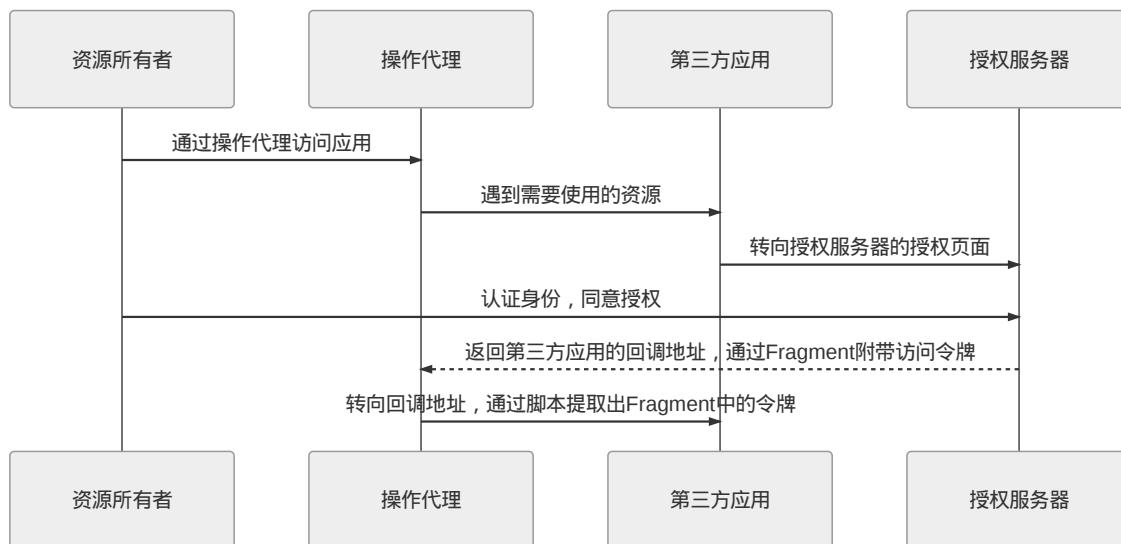
这是为了缓解OAuth2在**实际应用**中的一个主要缺陷，通常访问令牌一旦发放，除非超过了令牌中的有效期，否则很难（需要付出较大代价）有其他方式让它失效，所以访问令牌的时效性一般设计的比较短（譬如几个小时），如果还需要继续用，那就定期用刷新令牌去更新，授权服务器就可以在更新过程中决定是否还要继续给予授权。至于为什么说很难让它失效，我们将放到下一节“凭证”中解释这一点。

尽管授权码模式是严谨的，但是它并不够好用，这不仅仅体现在它那繁复的调用过程上，还体现在它对第三方应用提出了一个具体的要求：必须有服务端（因为第4步要发起服务端转向，而且服务端的地址必须与注册时提供的回调URI在同一个域内）。不要觉得要求一个系统要有服务端是天经地义理所当然的事情，本站的示例程序（<http://bookstore.icyfenix.cn>）就没有服务端支持，里面使用到了GitHub Issue作为留言板，对GitHub来说照样是第三方应用，需要OAuth2授权来解决。除浏览器外，现在越来越普遍的是移动或桌面端的Client-Side Web Applications，譬如现在大量的基于Cordova、Electron、Node-Webkit.js的PWA应用。所以在此需求里，引出了OAuth2的第二种授权模式：隐式授权。

## 隐式授权

隐式授权省略掉了通过授权码换取令牌的步骤，整个授权过程都不需要服务端支持，一步到位。其代价是在隐式授权中，授权服务器不会再去验证第三方应用的身份（因为没有服务器了，ClientSecret没有人保管，就没有意义了。但其实还是会限制第三方应用的回调URI地址必须与注册时提供的域名一致，有可能被DNS污染之类的攻击所攻破，但仍算是尽

人事努力一下）；也不能避免令牌暴露给资源所有者（以及用户机器上可能意图不轨的其他程序、HTTP的中间人攻击等）了。隐私授权的调用时序如下图（从此之后的授权模式，时序中我就不画资源访问部分的内容了，就是前面opt框中的那一部分，以便更聚焦重点）所示：



在以上过程设计中，与授权码模式模式的显著区别是授权服务器在得到用户授权后，直接返回了访问令牌，这显然降低了安全性，但OAuth2仍然努力尽可能地做到相对安全，譬如在前面提到的隐私授权中，尽管不需要用到服务端，但仍然需要在注册时提供回调域名，此时会要求该域名与接受令牌的域名处于同一个域内。此外，在隐私模式中明确禁止发放刷新令牌。

还有一点，在RFC 6749对隐式授权的描述中，特别强调了令牌是“通过Fragment带回”的。部分对超文本协议没有了解的读者，可能不知道Fragment是个什么东西？

### 额外知识

In computer [hypertext](#), a **fragment identifier** is a [string](#) of [characters](#) that refers to a [resource](#) that is subordinate to another, primary resource. The primary resource is identified by a [Uniform Resource Identifier](#) (URI), and the fragment identifier points to the subordinate resource.

不想看英文，或者看了觉得概念不好的话，我简单告诉你，Fragment就是地址中 "#"号后面的部分，譬如这个地址：

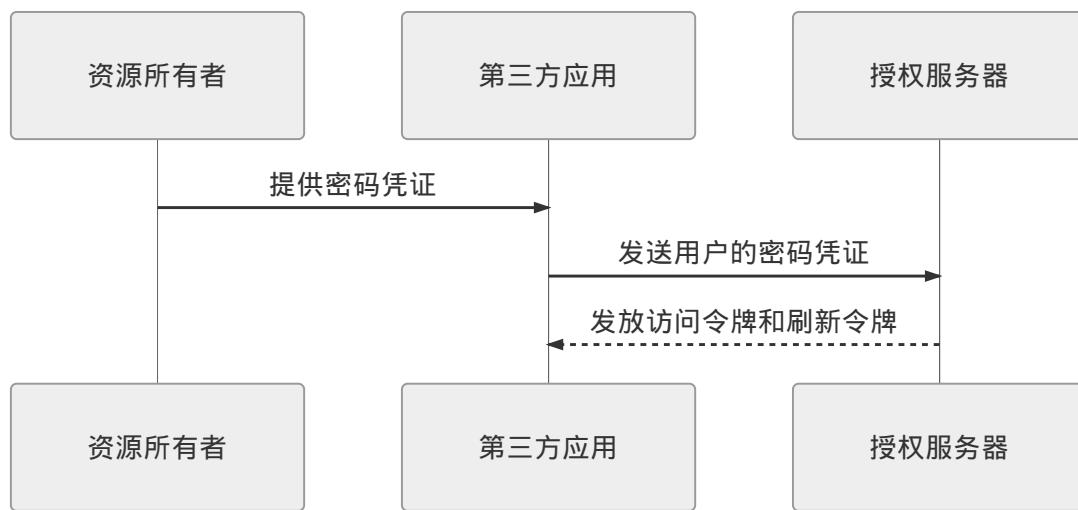
<http://bookstore.icyfenix.cn/#/detail/1>

后面的“/detail/1”便是Fragment，这个语法是在[RFC 3986](#)中定义的标准，规范中解释了这是用于客户端定位的URI从属资源，譬如HTML中就可以使用Fragment来做文档内的跳转（你现在可以点击一下这篇文章左边菜单中的几个子标题，看看浏览器地址的变化）而不会发起服务端请求。此外，如果浏览器对一个带有Fragment的地址发出Ajax请求，那Fragment是不会跟随请求被发送到服务端的，只能在客户端通过Script脚本来读取。所以隐式授权巧妙地利用这个特性，尽最大努力地避免了令牌从操作代理到第三方服务之间的链路存在被攻击的可能性，而被泄漏出去。而认证服务器到操作代理之间的这一段链路的安全，则可以通过TLS（即HTTPS）来保证没有中间没有受到攻击的，我们可以要求认证服务器都是基于HTTPS的，但无法要求第三方应用都是基于HTTPS。

## 密码模式

前面所说的授权码模式和隐私模式，是纯粹的授权模式，它与认证没有直接的关系，如何认证用户的真实身份这是与进行授权互相独立的过程。但在密码模式里，认证和授权就被整合成了同一个过程了。

这一种模式原本是只提供给用户对第三方应用是高度可信任的场景之中，譬如第三方应用是操作系统，本应该是不太常见的。但是近年来微服务风潮兴起，反而涌现出了密码模式的一种常见应用形式，譬如微服务群中有一些应用服务与授权服务都是由同一个服务商所搭建的，这自然就可以信任它们了。在单体服务的Fenix's Bookstore实现里，就直接采用了密码模式将认证和授权统一起来，我并不需要担心通过前端代码输入用户名、密码时，前端代码会对这些敏感信息做出什么不轨的行为，因为前端代码虽然在OAuth2中相当于第三方应用的角色，但它也是我本人所提供的，所以不存在信任问题（同时再次说明，如果不是出于方便与其他架构对比的目的，那也不存在引入OAuth2把它当作第三方看待的必要）。密码模式的调用时序就很简单了，如下图所示：

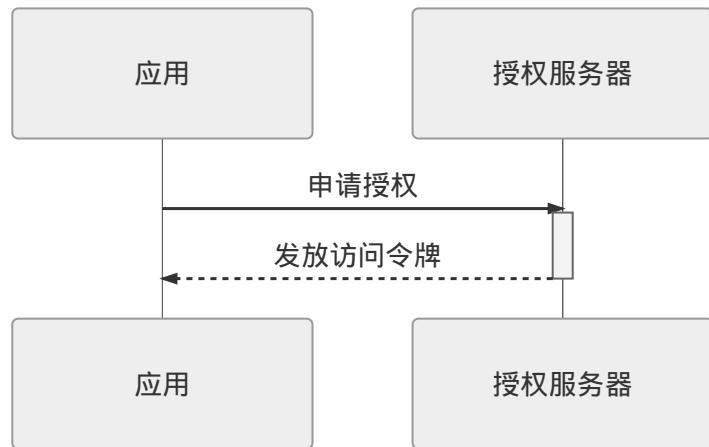


显然，在这种模式下，“如何保障安全”的职责无法由OAuth2的过程设计来承担，应是由用户和第三方应用来自行保障了，尽管OAuth2在规范中强调到“此模式下，第三方应用不得保存用户的密码”，但这并没有任何的约束力。

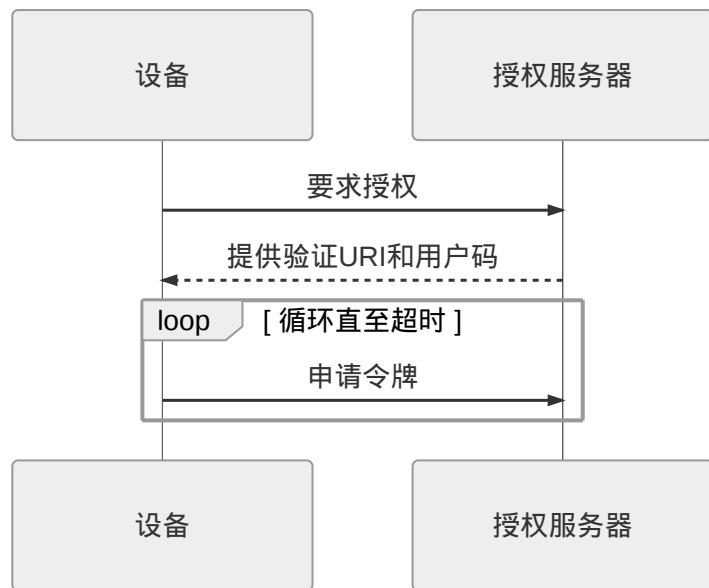
## 客户端模式

客户端模式是四种模式中最简单的，它只涉及到两个主体，第三方应用和授权服务器。严谨一点说，现在叫第三方应用已经不合适的，因为这里已经没有了“第二方”的存在，资源所有者、操作代理都是不存在的。甚至于叫“授权”都不太恰当，资源所有者都没有了，自然也不会有谁授予谁权限的过程。

客户端模式是指应用（就不写第三方了）以自己的名义，向授权服务器申请资源许可。这通常用在一些管理或者自动处理形场景之中。举个例子，譬如我开了一家网上书店，因为小本经营，不像京东那样全国多个仓库可以调货，我得保证只要客户成功购买，我就必须有货可发，不能超卖。但经常有人下了订单又拖着不付款，导致部分货物处于冻结状态。所以我写了一个订单清理的定时服务，自动清理掉超过2分钟的未付款的订单。这件UserStory里，订单肯定属于用户自己的资源，如果把订单清理服务看作一个独立的第三方应用的话，他就不应该向用户去申请授权，而应该直接以自己的名义向授权服务器申请一个能清理所有用户订单的授权。客户端模式的时序如下图所示：



还有一种与客户端模式类似的授权模式，在[RFC 8628](#)中定义为“设备码模式（Device Code）”，这里顺便简单提一下。设备码模式用于在无输入的情况下区分设备是否允许，典型的应用便是手机锁网解锁（锁网在国内较少，但在国外很常见）或者激活（譬如某游戏机注册到某个游戏平台）的过程。时序如下图所示：



进行验证时，设备需要从授权服务器获取一个URI地址和一个用户码，然后需要用户手动或设备自动地到验证URI中输入用户码。在这个过程中，设备会一直循环，尝试去获取令牌，直到拿到令牌或者用户码过期为止。

# 凭证

## 凭证 ( Credentials )

系统如何保证它与用户之间的承诺是双方当时真实意图的体现，是准确、完整且不可抵赖的？

在前面介绍OAuth2的内容中，每一种授权模式的目的都是拿到访问令牌，但从未涉及过拿回来的令牌应该长什么样子？反而还挖了一些坑（为何说OAuth2的一个主要缺陷是令牌难以主动失效）还没有填。这节我们讨论凭证，此话题中令牌必须得是主角了，此外，我们还要在这节讨论不使用OAuth2、最传统的方式是如何完成前面所讨论的认证、授权的。

## Cookie-Session

我们知道，HTTP协议是一种无状态的传输协议，无状态是指协议对事务处理没有上下文的记忆能力，每一个请求都是完全独立的，但是我们中肯定有许多人并没有意识到HTTP协议无状态的重要性。假如你做了一个简单的网页，其中包含了1个HTML、2个Script脚本、3个CSS、还有10张图片，这个网页成功展示在用户屏幕前，需要完成16次与服务端的交互，由于服务器响应的顺序与发送请求的先后没有直接联系，按照可能出现的响应顺序，一共会有 $P(16,16) = 20922789888000$ 种可能性。试想一下，如果HTML协议不是设计成无状态的，这16次请求各个有依赖关联，先调用哪一个、先返回哪一个，都会对结果产生影响的话，那协调工作会有多么复杂。

可是，HTTP协议的无状态特性又有悖于我们最常见的网络应用，譬如认证、授权方面，系统总得要获知用户身份才能提供服务，因此，我们也希望HTTP能有一种手段，让服务器至少有办法能够区分出发送请求的用户是谁。为了实现这个目的，[RFC 6265](#)规范中定义了HTTP的状态管理机制，在HTTP协议中设计了Set-Cookie指令，该指令的含义是以K/V值对的方式向客户端发送信息，此信息将在此后一定时间内的每次HTTP请求中，以名为Cookie的Header中附带着重新发回服务端，一个典型的Set-Cookie指令如下所示：

```
Set-Cookie: id=icyfenix; Expires=Wed, 21 Feb 2020 07:28:00 GMT; Secure;
HttpOnly
```

从此以后，当客户端对同一个域名（或者Path）的请求中都会带有值对信息“id=icyfenix”，例如以下所示：

```
GET /index.html HTTP/2.0
Host: icyfenix.cn
Cookie: id=icyfenix; sessionid=38afes7a8
```

根据每次请求传到服务端的Cookie，服务器就能分辨出请求来自于哪一个用户。由于Cookie是放在请求头上的负载（Payload，这个词后面还要频繁用到），不可能存储太大量的数据，放在Cookie中传输也不安全（被窃取，被篡改），所以通常是不会像例子中“id=icyfenix”这样的直接携带数据的。一般来说，Cookie中一般传输的是一个无意义的不重复的字符串，通常以sessionid或者jsessionid为名，服务器拿这个字符串为Key，再在内存中开辟一块空间，以Key/Entity的结构存储每一个在线用户的上下文状态，并辅以一些超时自动清理的管理措施，这种服务端的状态管理机制就是今天大家耳熟能详的Session，Cookie-Session就是在今天广泛应用于大量系统中的、服务端与客户端联动的状态管理机制。

Cookie-Session的方案在本章的主题“安全”上其实多少是占有一定优势的：信息都存储于服务器，不易遭遇传输中被泄漏、篡改的风险，只要通过域保护机制和传输层安全，保证Cookie中的键值不被窃取（如在“漏洞利用”小节中介绍的CSRF、XSS攻击）导致被冒认身份即可。Cookie-Session方案另一大优点是服务端有主动的管理能力，可根据自己的意愿随时修改、清除任意上下文状态，如实现强制某用户下线的功能就很容易。

Session-Cookie在单节点单体服务环境中是非常合适的方案，但当服务能力需要水平扩展，要部署集群时就开始面临一些麻烦了，由于Session建立在服务器的内存中，当服务器水平拓展成多节点时，我们必须在以下三种方案中选择其一：

- 要么就牺牲集群的一致性（Consistency）能力，让均衡器采用亲和式的负载均衡算法（譬如根据用户IP或者sessionid来分配节点），每一个特定用户发出的所有请求都一直被分配到其中某一个节点来提供服务，每个节点都不重复地保存着一部分用户的状态，如果这个节点崩溃了，里面的用户状态便完全丢失。

- 要么就牺牲集群的可用性（ Availability ）能力，让各个节点之间采用复制式的Session，每一个节点中的Session变动都会发送到组播地址的其他服务器上，这样某个节点崩溃了，不会中断都某个用户的服务，但Session之间组播复制的同步代价高昂，节点越多时越是如此。
- 要么就牺牲集群的分区容错（ Partition Tolerance ）能力，让普通的服务节点中不再保留状态，将上下文集中放在一个所有服务节点都能访问到的数据节点中进行存储。此时的矛盾是数据节点就成为了单点，一旦数据节点损坏，整个集群都不能提供服务。（多说一句，现在数据节点常见以Redis来搭建，本身Redis通常也会做集群，但将大集群的CAP问题放到小集群里，并不会让问题消失，简而言之就是：[禁止套娃](#)）

以后，我们在微服务架构中还会遇到更多分布式的问题，还会经常受到CAP理论（C、A、P必须牺牲一个）的打击，这是一个很值得深入探讨的技术权衡，但毕竟与本章的“安全”关系不大，这里就不再展开了。现在我只想知道一个问题的答案：前面三种方案都有缺陷，那在分布式应用中，就没有能绕过这些问题的解决方案吗？

我的答案是：有，也没有。如果说要解决分布式环境下的共享数据的CAP矛盾，这是被数学严格证明了不可能的，所以分布式环境中的状态管理一定会受到CAP的限制。但如果是在解决分布式下的认证授权问题，那确实还有一些别的法子可想。前面这句话的言外之意是提醒读者，接下来的JWT令牌与Cookie-Session并不是对等的技术方案，它只解决认证授权问题，充其量能携带少量非敏感的信息，只是Cookie-Session在认证授权问题上的替代品，而不会成为Cookie-Session本身的革命者与继承人。

## JWT

前面介绍的Cookie-Session机制在分布式环境下遇到一些问题，在多方系统中，就更不可能谈什么Session层面的数据共享了，而且Cookie也没法跨域。看来，服务器多了，确实不好解决，那就换个思路吧，客户端是唯一的，把数据存储在客户端，每次随着请求发回服务器——JWT就是这种思路的典型代表。

JSON Web Token（JWT），定义于[RFC 7519](#)的令牌格式，是目前广泛使用的一种令牌，尤其是与OAuth2配合应用于分布式的、涉及多方的应用系统之中。介绍JWT的具体构成之前，我们先来看一下它是什么样子的，一个JWT的例子如下图所示：

Encoded PASTE A TOKEN HERE

```
eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJ
1c2VyX25hbWUiOiJpY31mZW5peCIsInNjb3BlIjp
bIkFMTCJdLCJleHAiOjE10DQyNTA3MDQsImF1dGh
vcml0aWVzIjpBI1JPTEVfVVNFUiIsI1JPTEVfQUR
NSU4iXSianRpIjoiMTNmNGN1MWQtNmY20C00NzQ
xLWI5YzYtMzkyNzU10GQ5NzR1IiwiY2xpZW50X21
kIjoiYm9va3N0b3J1X2Zyb250ZW5kIiwidXNlcm5
hbWUiOiJpY31mZW5peCJ9.82awQU4IcLVXr7w6px
cUCWrcEHKq-LRT7ggPT_ZPhE0
```

Decoded EDIT THE PAYLOAD AND SECRET

HEADER: ALGORITHM & TOKEN TYPE

```
{
 "alg": "HS256",
 "typ": "JWT"
}
```

PAYOUT: DATA

```
{
 "user_name": "icyfenix",
 "scope": [
 "ALL"
],
 "exp": 1584250704,
 "authorities": [
 "ROLE_USER",
 "ROLE_ADMIN"
],
 "jti": "13f4ce1d-6f68-4741-b9c6-3927558d974e",
 "client_id": "bookstore_frontend",
 "username": "icyfenix"
}
```

VERIFY SIGNATURE

```
HMACSHA256(
 base64UrlEncode(header) + "." +
 base64UrlEncode(payload),
 your-256-bit-secret
) secret base64 encoded
```

## JWT令牌结构

以上截图来自于网站<https://jwt.io/>，当然，数据是我自己编的。左边的是JWT的本体，它通过名为Authorization的Header发送给服务端，前缀是在[RFC 6750](#)中定义的bearer，这点在之前关于“认证”的小节中提到过，一个完整的HTTP请求实例如下所示：

```
GET /restful/products/1 HTTP/1.1
Host: icyfenix.cn
Connection: keep-alive
Authorization: bearer
eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJ1c2VyX25hbWUiOiJpY31mZW5peCIsIn
Njb3BlIjpBIkFMTCJdLCJleHAiOjE10DQ5NDg5NDcsImF1dGhvcmI0aWVzIjpBI1JPTEVfV
VNFiIsI1JPTEVfQURNSU4iXSianRpIjoiOWQ3NzU4NmEtM2Y0Zi00Y2JiLTk5MjQtZmUy
Zjc3ZGZhMzNkIiwiY2xpZW50X2lkIjoiYm9va3N0b3J1X2Zyb250ZW5kIiwidXNlcm5hbWU
i0iJpY31mZW5peCJ9.539WMzbjv63wBtx4ytYYw_Fo1ECG_9vsgAn8bheflL8
```

图中右边的内容是经过Base64URL转码之后的令牌明文，是的，明文，JWT令牌默认是不加密的（你自己要加密也行就，接收时自己解密即可）。从明文中可以看到JWT令牌是以JSON结构（毕竟叫JSON Web Token）存储的，结构上可划分为三个部分，每个部分间用点号“.”分隔开。

第一部是令牌头 ( Header ) , 内容如下所示 :

```
{
 "alg": "HS256",
 "typ": "JWT"
}
json
```

它描述了令牌的类型 ( 统一为 typ:JWT ) 和令牌签名的算法 , 示例中 HS256 为 HMAC SHA256 算法的缩写 , 其他各种系统所支持的签名算法可以参考<https://jwt.io/> 网站所列。

第二部分是负载 ( Payload ) , 是令牌真正需要向服务端传递的信息 , 在认证问题中 , 至少应该包括告诉服务端“我是谁”的信息 , 在授权问题中 , 至少应该包括告诉服务端“我属于什么角色/权限 , 有哪些许可”。负载部分是可以完全自定义的 , 根据具体要解决的问题不同 , 设计自己所需要的信息 ( 但不能太多 , 毕竟受 HTTP Header 大小的限制 ) 。一个 JWT 负载的示例如下所示 :

```
{
 "username": "icyfenix",
 "authorities": [
 "ROLE_USER",
 "ROLE_ADMIN"
],
 "scope": [
 "ALL"
],
 "exp": 1584948947,
 "jti": "9d77586a-3f4f-4cbb-9924-fe2f77dfa33d",
 "client_id": "bookstore_frontend"
}
json
```

而 JWT 在 RFC 7519 中推荐 ( 无强制约束 ) 了 7 个声明名称 ( Claim Name ) , 如有需要用到这些内容 , 建议字段名与官方的保持一致 :

- iss ( Issuer ) : 签发人
- exp ( Expiration Time ) : 令牌过期时间
- sub ( Subject ) : 主题
- aud ( Audience ) : 令牌受众

- nbf（Not Before）：令牌生效时间
- iat（Issued At）：令牌签发时间
- jti（JWT ID）：令牌编号

此外在RFC 8225、RFC 8417、RFC 8485等规范文档，以及OpenID中都定义有约定公有含义的名称，比较多我就不贴出来了，可以参考[IANA JSON Web Token Registry](#)。

第三部分是**签名**（Signature），签名的意思是，使用特定的签名算法（在对象头中公开），使用特定的密钥（Secret，由服务器进行保密，不能公开）对前面两部分内容进行加密计算，以例子中JWT默认的HMAC SHA256算法为例，将通过以下公式产生签名值：

```
HMACSHA256(base64UrlEncode(header) + "." + base64UrlEncode(payload) ,
secret) java
```

签名的意义在于确保负载中的信息是可信的、没有被篡改的，也没有在传输过程中丢失。因为被签名的内容哪怕发生了一个字节的变动，也会导致整个签名发生显著变化。此外，由于这件事情只能由认证/授权服务器完成（只有它知道Secret），任何人都无法在篡改后重新计算出合法的签名值，所以服务端才能够完全信任客户端传上来的JWT中的负载信息。

之前提到了JWT默认采用的签名算法是HMAC SHA256，是一种哈希摘要算法，属于不可逆的“加密”，过程实质上是不用依赖密钥的，这时候的密钥实际上承担了加盐（Salt）的作用。在多方系统、授权服务与资源服务分离的实际应用中，通常会采用非对称加密算法（典型如RSA）来进行签名，这时候除了授权服务端持有的可以用于签名的私钥外，还会对其他服务器公开一个公钥，公钥不会用来签名，但是能被其他服务用于验证签名是否由私钥所签发的。这样其他服务器也能不依赖授权服务器独立判断JWT令牌中的信息的真伪。在Fenix's Bookstore的单体服务版本中，因为授权与资源服务在同一个服务端，采用了默认的哈希算法来加密签名；而Fenix's Bookstore的微体服务版本，认证授权单独划分出了一个独立的微服务，这时就采用了非对称加密来进行签名了。关于哈希、对称、非对称加密的讨论，将会放到“传输”一节中进行。

JWT令牌是多方系统中一种优秀的凭证载体，它不需要任何一个服务节点保留任何一点状态信息，就能够保障认证服务与用户之间的承诺是双方当时真实意图的体现，是准确、完整、不可篡改、且不可抵赖的。同时，由于JWT本身可以携带少量信息，这十分有利于RESTful API的设计，能够较容易地做成无状态服务，在做水平扩展时就不需要像前面Cooki

e-Session方案那样考虑如何部署的问题。现实中也确实有一些项目（譬如Fenix's Bookstore）直接采用JWT来承载上下文来实现完全无状态的服务端，这能获得很大的好处，譬如，在你调试Fenix's Bookstore的程序时，随时都可以停止、重启服务端程序，服务重启后客户端仍然是可以毫无感知地继续操作流程；而对于有状态的系统，一般就必须通过再次登录、进行前置业务操作来给服务端重建状态（以上这句话所指的“好处”不是开发时方便重启，而是指不必顾虑状态地增加或者减少服务来进行伸缩）。

目前，在大型系统中完全使用JWT来保存上下文状态，服务端完全不持有状态仍是不太现实的，不过将最热点的服务接口单独抽离出来，做成无状态的、幂等的服务，是一种很有效的提升系统吞吐能力的架构设计。这部分内容将在微服务架构的部分如何划分微服务的章节中进一步探讨。

JWT并不是没有缺点的完美方案，它存在着以下几个明显或者不明显的缺点：

- 令牌难以主动失效**：JWT令牌一旦签发，理论上就和认证服务器再没有什么瓜葛了，在到期之前就会始终有效，除非服务器部署额外的逻辑，这对某些管理功能的实现是很不利的。譬如，有一种颇为常见的需求是：要求一个用户只能在一台设备上登录，在B设备登陆后，之前已经登录过的A设备就应该自动退出。如果采用JWT，就必须设计一个“黑名单”的额外的逻辑，用来把要主动失效的令牌集中存储起来，而无论这个黑名单是实现在Session、Redis或者数据库中，都会让服务退化回有状态，降低了JWT本身的价值（但黑名单还是很常见的做法，需要维护的黑名单一般是很小的状态量，不少场景中是有存在意义的）。
- 更容易遭受重放攻击**：首先说明Cookie-Session也是有重放攻击问题的，只是因为Session中的数据控制在服务端手上，应对重放攻击会相对主动一些。要在JWT层面解决重放攻击需要付出比较大的代价，无论是加入全局序列号（HTTPS协议的思路）、Nonce字符串（HTTP Digest验证的思路）、挑战应答码（当下网银动态令牌的思路）、还是缩短令牌有效期强制频繁刷新令牌，在真正应用起来时其实都是很麻烦的，真要处理重放攻击，启用HTTPS是正道。
- 只能携带相当有限的数据**：HTTP协议并没有强制约束Header的最大长度，但是，各种服务器（甚至是浏览器）都会有约束，譬如Tomcat就要求Header最大不超过8KB，而在Nginx中则默认为4KB，因此在令牌中存储过多的数据不仅浪费带宽，还有额外的出错风险。
- 令牌在客户端如何存储**：严谨地说，这个并不是JWT的问题而是你的问题。如果授权之后，操作完了关掉浏览器这是结束了，那把令牌放到内存里面，压根不考虑持久化那是

最理想的。但并不是谁都能忍受一个网站关闭之后下次就一定强制要重新登陆的（大概也就银行的网站可以忍）。那这样的话，客户端该把令牌存放到哪里？Cookie？localStorage？Indexed DB？它们都有泄漏的可能，而令牌一旦泄漏，别人就可以冒充你的身份做任何事情。

- **无状态也不总是好处**：这个其实不也是JWT的问题。如果不能想像无状态会有什么不好的话，我给你提个需求：请基于无状态JWT的方案，做一个在线用户统计功能。兄弟，难搞哦。

我在写这篇文章的时候，在网上搜索资料，发现JWT的争议和吹捧都不少。技术只是工具而已，无论是迷信还是使劲黑它，都并无必要。

# 保密

## 保密 (Confidentiality)

系统如何保证敏感数据无法被包括系统管理员在内的内外部人员所窃取、滥用？

保密是加密和解密的统称，是以某种特殊的算法改变原有的信息数据，使得未授权的用户即使获得了已加密的信息，但因不知解密的方法，仍然无法了解信息的内容。

保密这个话题，按照需要保密信息所处的环节不同，可以划分为“信息在客户端时的保密”、“信息在传输时的保密”和“信息在服务端时的保密”，又或者进一步概括为“端的保密”和“链路的保密”。我们把最复杂、最有效，但却又早就有了标准解决方案的“传输环节”单独提取出来，放到下一个节去讨论。在本小节中，只讨论两个端的环节，即在客户端和服务端中的信息保密问题，谈一下笔者的几个观点。

## 安全的强度

首先来说一下“安全”的程度问题，保密的安全与否不应该被视为一个离散的二元选项，不是仅有“安全”或者“不安全”的差别，而是随着你的应用所要求的保密程度不同，应该有着不同的安全强度与之对应。这里面说的意思与很多口号中强调的“安全无小事”、“99%安全加1%的漏洞等于零”并不是一码事。我通过以下这些逐步提升的攻击手段和应对措施来解释“安全强度”是意味着什么：

1. 给密码做最简单的MD5，如果你的密码本身比较复杂，那一次简单的MD5至少可以保证密码不会被逆推出明文，密码在一个系统中泄漏了不至于威胁到其他系统的使用，但这不能阻止弱密码被彩虹表攻击所逆推。
2. 给密码加上固定的盐值，如果给密码加上盐值，可以替弱密码建立一道防御屏障，一定程度上防御已有的彩虹表攻击，但不能阻止加密结果被窃取后（譬如在链路上被抓包了），攻击者直接发送加密结果给服务端进行冒认。

3. 给密码加上动态的盐值，如果每次密码向服务端传输时都加入了动态的盐值，让每次加密的结果都不同，那即时传输给服务端的加密密码被窃取了，也无法用来冒认，但这只能保护登录这一个操作，无法阻止对其他功能的重放攻击。
4. 采用动态令牌与服务端的逻辑配合，可以做到防止重放攻击，依然无法抵御传输过程中被嗅探而泄漏信息的问题（如前面说的在链路上被抓包了）。
5. 启用HTTPS（且恰当选择支持的密码学算法、保护好证书），可以防御链路上的恶意嗅探，也在协议层面解决了重放攻击的问题，但它依然存在有被中间人攻击的可能性、有被证书攻击导致握手失败等风险。
6. .....

到了第5点，只要做法规范，已经可以抵御不少安全风险了，但也意味着你需要为它付出一些代价（包括加解密的算力，也包括购买证书的费用）。而安全的强度还可以用不同途径继续往上提升，如许多网站会使用手机验证码开辟另一条独立的信息传输渠道来保障安全、如银行会使用有专门物理存储的证书（就是俗称的U盾）来保障安全、如国家电网那样建设遍布全国各地的与公网物理隔离的专用网络来保障安全，等等。显然追求安全强度同时也意味着付出更多代价，肯定不是任何一个网站、系统都需要无限拔高的安全强度。

另一个问题是安全强度有尽头吗？存不存在某种绝对安全的保密方式？答案可能出乎多数人的意料，确实是有的。信息论之父香农严格证明了一次性密码（One Time Password）的绝对安全性。但是使用一次性密码必须有个前提，就是先把安全的把密码（密码列表）传达给对方。譬如，给你的朋友（人肉）送去一本存储了完全随机密码的密码本，然后每次使用其中一条密码来进行加密通讯，用完一条丢弃一条，理论上这是绝对安全的，但显然这对于公众互联网是没有任何的可执行性。

## 客户端加密

客户端在用户登录、注册一类场景里是否需要对密码进行加密，这个问题一直存有争议。我的观点是，为了保证密码不被黑客窃取而做客户端加密没有太多意义，上HTTPS可以说是唯一的普通系统实际可行的解决方案。但是！为了保证密码不在服务端被滥用，在客户端就开始加密是很有意义的。大网站被拖库的事情层出不穷，密码明文被写入数据库、被输出到日志中之类的事情屡见不鲜，做系统设计时就应该把明文密码这种东西当成是最烫手的山芋来看待，越早消灭掉越好。

关于第一个“没有太多意义”，有人不理解为什么为什么客户端加密对防御黑客会没有意义，我举个例子，在极端情况下，客户端可能被整个架空掉，这样上面无论做了什么防御措施都成“马其诺防线”了。典型的就是之前已经提到的中间人攻击，它可以通过劫持掉了客户端到服务端之间的某个节点，包括但不限于代理（通过HTTP代理返回赝品）、路由器（通过路由导向赝品）、DNS服务（直接将你机器的DNS查询结果替换为赝品地址）等等，把你想要访问的登陆页面整个给替换掉（全替换掉工作量太大，一般不会去做，都是注入一段恶意的JavaScript代码到正版的页面里）。最简单的劫持路由器，在局域网内其他机器释放ARP病毒便有可能做到这一点。这部分内容属于链路安全，我们将在下一节来讲如何防御，这里附带Mozilla[对中间人攻击的一段介绍以供参考。](#)

### 中间人攻击 (Man-in-the-Middle Attack , MitM )

在消息发出方和接收方之间拦截双方通讯。用日常生活中的写信来类比的话：你给朋友写了一封信，邮递员可以把每一份你寄出去的信都拆开看，甚至把信的内容改掉，然后重新封起来，再寄出去给你的朋友。朋友收到信之后给你回信，邮递员又可以拆开看，看完随便改，改完封好再送到你手上。你全程都不知道自己的信件和收到的信件都经过邮递员这个“中间人”转手和处理——换句话说，对于你和你朋友来讲，邮递员这个“中间人”角色是不可见的。

关于第二个“很有意义”，居然也有人会抬杠。一种是说涉及到密码等敏感信息的都会由靠谱的人完成，或者就是他本人做的，所以不会出问题，我觉得这个就没什么必要反驳了，开心就好。另一种的观点是保存明文密码（把不含盐的哈希结果也作明文看待）的目的是为了便于客户端做动态盐值，因为这需要服务端存储了明文才能每次用新的盐值重新加密来与客户端传上来的加密结果进行比较。我的观点是每次从服务端请求盐值在客户端动态加盐往往得不偿失，应在真正防御性的密码加密存储应该在服务端进行，因为客户端无论是否动态加盐，都不能代替HTTPS。

## 密码存储和验证

下面以Fenix's Bookstore的实现为具体样例，介绍从密码如何从客户端传输到服务端，存储进数据库的全过程。在保障一定安全强度的同时，避免消耗过多的运算资源，验证起来也比较便捷。这套过程对于一般的系统，配合一定的约束（如密码要求长度、特殊字符等），再配合HTTPS传输应该是够用的。即使在客户采用了弱密码、客户端盐值泄漏（本

来就不是保密的）、服务端被拖库泄漏了存储的密文和动态盐值这些问题同时发生，也没有用户明文密码被逆推出来的风险。

以下为密码创建的过程，

1. 用户在客户端注册，输入明文密码：123456。

```
password = 123456
```

java

2. 客户端对用户密码进行简单Hash，可选的算法有MD2/4/5、SHA1/256/512、BCrypt、PBKDF1/2，等等。

```
client_hash = MD5(password) // e10adc3949ba59abbe56e057f20f883e
```

java

3. 为了防御彩虹表攻击，应加盐处理，客户端加盐可取固定的字符串，或者伪动态（日期、用户名加上固定字符串，反正就是服务端不需要额外通讯可以得到的值）的盐值。

```
client_hash = MD5(MD5(password) + salt) // SALT =
$2a$10$o5L.dWYEjZjaej0mN3x4Qu
```

java

4. 我们假设攻击者截获了传输，把哈希值和盐值都拿到了，那他可以枚举遍历所有10位数以内（10位数只是举个例子，反正就是弱密码，你拿1024位随机字符当密码用，加不加盐，彩虹表都跟你没关系）的弱密码，然后对每个密码再加盐计算，得到一个固定盐值的对照彩虹表。为了应对这种暴力破解，我们需要引入慢哈希函数来代替MD5来加强安全性。

慢哈希函数是指这个函数执行时间（准确地说是运算次数）是可以调节的，BCrypt算法就是一种慢哈希函数，在做哈希时接收盐值salt和执行成本cost两个参数（代码层面cost一般是混入在salt中，譬如上面例子中的salt就是混入了10轮运算的盐值，10轮的意思是 $2^{10}$ 次哈希，cost参数是放在指数上的，最大取值就31）。如果我们控制BCrypt的执行时间大概是0.1秒完成一次哈希计算的话，按照1秒生成10个哈希的速度，算完所有的10位大小写字母和数字组成的弱密码大概需要 $P(62,10)/(3600*24*365)/0.1=1,237,204,169$ 年。

```
client_hash = BCrypt(MD5(password) + salt) //
MFfTW3uNI4eqhwDkG7HP9p2mzEUu/r2
```

java

## 5. 现在将哈希传输到服务端，链路这段安全在下一节去探讨。

服务端接受到哈希值后，对每一个密码都动态生成一个随机盐值。比较主流的建议是采用“[密码学安全伪随机数生成器](#)”（ Cryptographically Secure Pseudo-Random Number Generator , CSPRNG ）来产生一个长度与哈希值相等的随机字符串。对于Java语言，从Java SE 7起提供了java.security.SecureRandom类，用于支持CSPRNG字符串生成。

```
SecureRandom random = new SecureRandom();
byte server_salt[] = new byte[36];
random.nextBytes(server_salt); // tq2pdxb1kbgp8vt8kbdpmzdh1w8bex
```

java

## 6. 将盐混入客户端传来的哈希值，生成要存入数据库的密文，并将随机生成的盐值一并写入到同一条记录中。在服务端中就不建议采用慢哈希算法，对CPU占用率的影响较大，Spring Security 5的StandardPasswordEncoder提供了SHA256哈希算法的实现，就以此为例。

```
server_hash = SHA256(client_hash + server_salt); //
55b4b5815c216cf80599990e781cd8974a1e384d49fbde7776d096e1dd436f67
DB.save(server_hash, server_salt);
```

java

## 7. ( 可选 ) 出于对SHA256安全性的不信任，Spring Security 5中StandardPasswordEncoder已被@Deprecated，介意或者想偷懒简化操作的话，推荐第5、6步采用BCryptPasswordEncoder来替代。尽管使用的是BCrypt算法，但默认构造函数中的cost是-1，即进行 $2^{-1}=1$ 次哈希计算，这并不会造成服务端压力。说可以偷懒是因为用BCryptPasswordEncoder的话就不需要专门传入盐值，它本身就会调用CSPRNG产生盐值，也不需要给数据库添加盐值字段了，在它生成密码的前32位自动存储了盐值。

以下为密码验证的过程：

1. 客户端，用户在登陆页面中输入密码明文：123456，经过与注册相同的加密过程，向服务端传输加密后的结果。

```
authentication_hash = MFFTW3uNI4eqhwDkG7HP9p2mzEUu/r2
```

java

2. 服务端，接受到客户端传输上来的哈希值，从数据库中取出登陆用户对应的密文和盐值，采用服务端的哈希算法，对客户端传来的哈希值、服务端盐值计算出哈希结果。

```
result = SHA256(authentication_hash + server_salt); //
55b4b5815c216cf80599990e781cd8974a1e384d49fbde7776d096e1dd436f67
```

java

3. 比较上一步的结果和数据库储存的哈希值是否相同，如果相同那么密码正确，反之密码错误。

```
authentication = compare(result, server_hash) // yes
```

java

# 传输

## 传输 (Transport Security)

系统如何保证通过网络传输的信息无法被第三方窃听、篡改和冒充？

这节的主角是签名、证书、TLS等，但不会涉及“到哪找免费的CA证书？”、“如何生成数字证书？”、“如何把证书置入Web服务器？”这一类操作性的话题，而更多是对整套传输安全层原理的讲述。尽管这部分内容相对较难，但如果前面你已经阅读过并理解了认证、授权、凭证、保密的内容，而又对SSL/TLS本身没有什么了解的话，那这一节可能会是最容易理解的讲述传输安全层工作原理的方式。笔者将从“假设传输层安全得不到保障，攻击者如何摧毁之前认证、授权、凭证、保密中所提到的种种安全机制”为具体场景来讲解传输层安全所面临的问题和它的解决方案。

## 摘要、加密与签名

我们从JWT令牌的一小段“题外话”来引出整套现代加密通讯体系，以便于阐述哈希摘要、对称/非对称加密的特点与局限。我们知道，JWT中携带信息的价值来自于它是被签名的、不可篡改的信息。这一点之前介绍到：

签名的意义在于确保负载中的信息是可信的、没有被篡改的，也没有在传输过程中丢失。因为被签名的内容哪怕发生了一个字节的变动，也会导致整个签名发生显著变化。此外，由于这件事情只能由认证/授权服务器完成（只有它知道Secret），任何人都无法在篡改后重新计算出合法的签名值，所以服务端才能够完全信任客户端传上来的JWT中的负载信息。

我们来深入分析一下，“签名”具体是如何让负载中的信息变得“不可篡改”的。以默认的SHA 256哈希算法为例，进行签名，相当于进行如下计算过程：

```
signature = SHA256(base64UrlEncode(header) + "." +
base64UrlEncode(payload), secret)
```

java

理想的哈希算法通常都会具备两个特性：一是**易变性**，这是指算法的输入发生了任何一点变动，都会导致**雪崩效应**（Avalanche Effect），使得输出结果发生极大的变化。这个特性常被用来做校验，譬如互联网上下载大文件，常会附有一个哈希校验码，以确保下载下来的文件没有因网络或其他原因与原文件产生哪怕一个字节的偏差。二是**不可逆性**，这是指算法根据输入计算输出结果耗费的算力资源极小，但根据输出结果反过来推算原本的输入，耗费的算力就极大。一个经常被人们用来讲解不可逆性的例子是**大数分解**，我们可以轻而易举的地（以O(nlogn)的复杂度）计算出两个大素数的乘积：

```
97667323933 * 128764321253 = 12576066674829627448049
```

java

根据**算术基本定理**，质因数的分解形式是唯一的，且前面笔者所举例的运算因子已经都是素数，所以12576066674829627448049的分解形式只有唯一的上面所示的一种答案，但是如何对大数进行因数分解，迄今没有找到多项式时间的解法（24位十进制数的因数分解完全在现代计算机的暴力处理能力范围内，这里只是举例。但目前很多计算机科学家都相信大数分解问题就是一种P!=NP的证例，尽管也并没有人能证明它一定不存在多项式时间的解法）。不可逆性常被人用来做数字签名，利用的就是如果你知道密钥，很容易通过明文算出签名值，但知道明文和签名值，几乎不可能逆推出密钥，这就实现了签名易于验证，难以破解的特点。

必须注意，签名“易于验证、难以破解”是建立在密钥不会泄漏，也不会被篡改的基础上的。当授权服务器与资源服务器是同一个服务时，JWT运作不会遭遇什么风险。而当授权服务器与资源服务器并不是同一个，他们之间就涉及到资源服务其如何验证的问题，无论是资源服务器对每个收到的令牌都请求授权服务器验证一下，还是资源服务器自己也拿到密钥来自行验证令牌真伪都是不可行的。这种情况的解决方案前面讨论中已提到过：

在多方系统、授权服务与资源服务分离的实际应用中，通常会采用非对称加密算法（典型如RSA）来进行签名，这时候除了授权服务端持有的可以用于签名的私钥外，还会对其他服务器公开一个公钥，公钥不会用来签名，但是能被其他服务用于验证签名是否由私钥所签发的。这样其他服务器也能不依赖授权服务器独立判断JWT令牌中的信息的真伪。

非对称加密就是加密和解密使用的是不同的密钥的算法，那自然对称加密就是指加密是指加密和解密是一样的密钥的算法。不知道上面看这段话的时候，你心中是否会想“这里写JWT通常会采用非对称加密算法，那改用对称加密行不行呢？”之类的疑问。答案是除非有其他传递或者动态协商密钥的途径，否则这个场景中对称加密是不可行的，因为对称加密

只有一个密钥，授权和资源服务不在同一台服务器的话，如何将这个密钥传送给资源服务器？再加密一次传送的话就成了“[蛋鸡悖论](#)”了。

事实上，在分布式环境中，真正能够用来签名的，通常都只有非对称加密算法。在它的密钥对中，其中一个密钥是对外公开的，所有人都可以获取到，称为公钥，另外一个密钥是不公开的称为私钥。这两个密钥谁加密、谁解密构成了两种不同的用途：

1. 公钥加密，私钥解密，这种就是**加密**，用于向公钥所有者发布信息，这个信息可能被他人篡改，但是无法被他人获得。如果甲想给乙发一个安全的保密的数据，那么应该甲乙各自有一个私钥，甲先用乙的公钥加密这段数据，再用自己的私钥加密这段加密后的数据。最后再发给乙，这样确保了内容即不会被读取，也不会被篡改。
2. 私钥加密，公钥解密，这种就是**签名**，用于让所有公钥所有者验证私钥所有者的身份并且用来防止私钥所有者发布的内容被篡改。但是不用来保证内容不被他人获得。

看到这里，可能有人在想只用“非对称加密”行不行？为什么还需要对称加密？答案是非对称加密算法对加密内容的长度有限制，不能超过公钥长度。譬如说现在常用的公钥是长度是2048 bits，意味着明文不能超过256 bytes。此外，由于对称加密的设计难度相对较小，其加密的效率一般远高于非对称加密，这决定了在大数据量的加密数据传输中，通常是两种加密算法结合使用的，用非对称加密来传递密钥，收到密钥后用对称加密来加密内容。

下表把前面涉及到的三种算法放到一块，列举了它们的主要特征、用途和局限性：

类型	特点	常见实现	主要用途	主要局限
哈希摘要	不可逆，即不能解密，所以并不是加密算法，只是一些场景把它当作加密算法使用 易变性，输入发生1Bit变动，就可能导致输出结果50%的内容发生改变。 无论输入长度多少，输出长度固定（2的N次幂）	MD2/4/ 5/6、SH A0/1/25 6/512	摘要	无法解密
对称加密	加密是指加密和解密是一样的密钥 设计难度相对较小，执行速度相对较块 加宽数明文长度不受限制	DES、A ES、RC 4、IDEA	加密	要解决如何把密钥安全地传递给解密者

类型	特点	常见实现	主要用途	主要局限
非对称加密	加密和解密使用的是不同的密钥 明文长度不能超过公钥长度	RSA、B CDSA、 ElGamal	签名、 传递 密钥	加宽数字长 度受限

现在我们再回到多方系统如何验证令牌的问题中来。嗯？有了非对称加密，公钥可以不需要加密地公开了，那问题难道还没有解决了吗？并没有，还存在一个明显的漏洞，公钥虽然是公开的，但如何保证要获取公钥的资源服务，拿到的公钥就是授权服务所希望它拿到的呢？如果公钥在网络传输过程中，获取公钥的这一步被攻击者截获并篡改了，返回了攻击者自己提供的公钥，那以后攻击者就可以用自己的私钥签名，让资源服务器无条件信任自己的所有动作了。这里公钥显然也无法再用加密来传输，否则也是一个蛋鸡问题。

## 数字证书

当我们无法以“签名”的手段来达成信任时，就只能求助于其他途径。不妨想想真实的世界中，我们是如何达成信任的，其实不外乎以下两种：

- **基于共同私密信息的信任**

譬如某个陌生号码找你，说是你的老同学，生病了要找你借钱。你能够信任他得方式是向对方询问一些你们两个应该知道，且只有你们两个知道的私密信息，如果对方能够回答上来，他有可能真的是你的老同学。

- **基于权威公证人的信任**

如果有个陌生人找你，说他是警察，让你把存款转到他们的安全账号上。你能够信任他的方式是找到公安局，如果公安局担保他确实是个警察，那他有可能真的是警察。

回到网络世界中，我们并不能假设授权服务器和资源服务器是互相认识的，所以通常不太会采用第一种方式，而第二种就是目前标准的保证公钥可信分发的标准，这个标准一个名字：[公开密钥基础设施](#)（Public Key Infrastructure，PKI）。

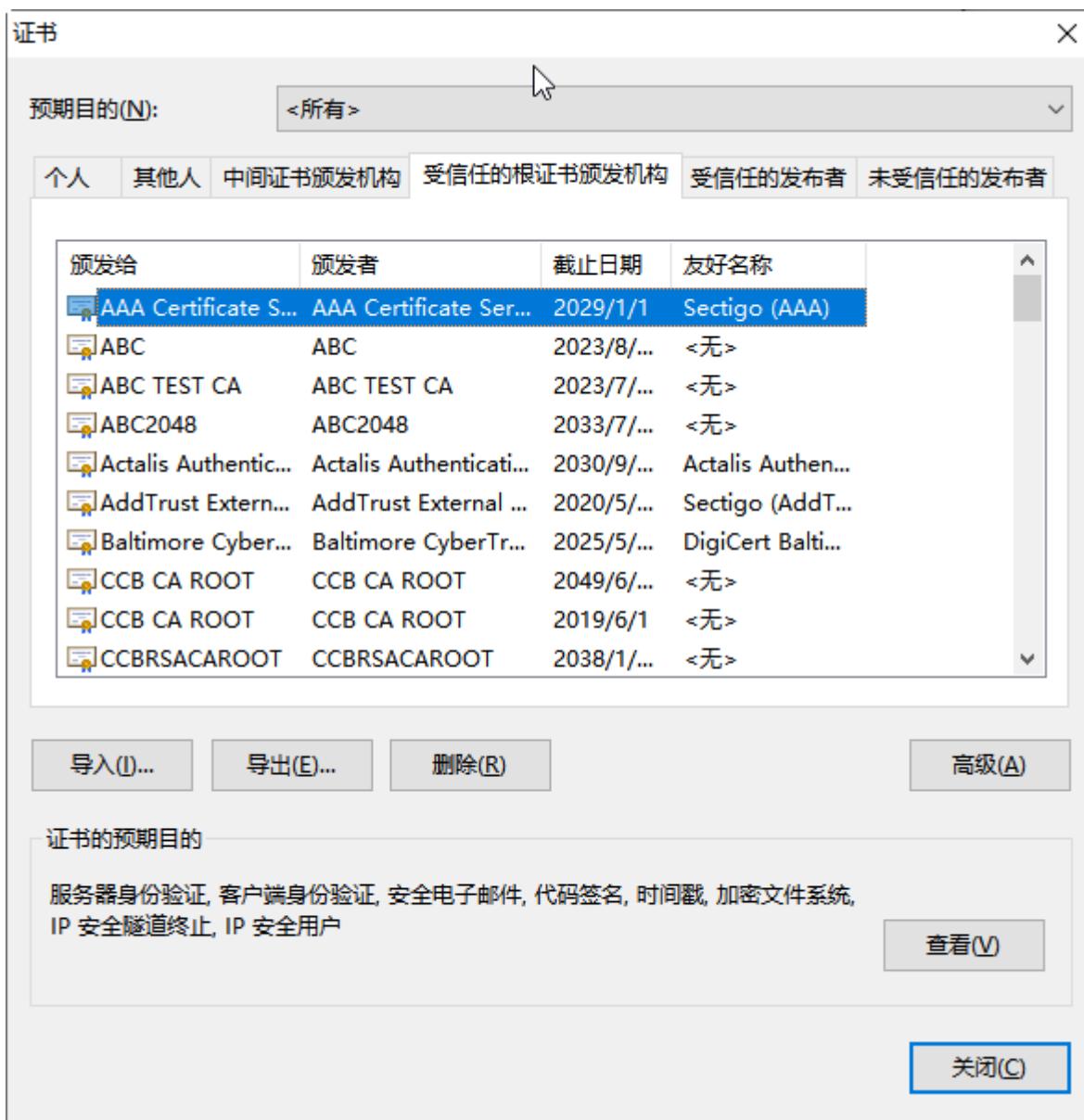
### 公开密钥基础设施（Public Key Infrastructure，PKI）

又称公开密钥基础架构、公钥基础建设、公钥基础设施、公开密码匙基础建设或公钥基础架构，是一组由硬件、软件、参与者、管理政策与流程组成的基础架构，其目的在于创

造、管理、分配、使用、存储以及撤销数字证书。

密码学上，公开密钥基础建设借着数字证书认证中心（ Certificate Authority , CA ）将用户的个人身份跟公开密钥链接在一起。对每个证书中心用户的身份必须是唯一的。链接关系通过注册和发布过程创建，取决于担保级别，链接关系可能由CA的各种软件或在人为监督下完成。PKI的确定链接关系的这一角色称为注册管理中心（ Registration Authority , RA ）。RA确保公开密钥和个人身份链接，可以防抵赖。

我们不纠缠于PKI概念上的内容，只要知道里面定义了数字证书认证中心便相当于前面例子中“权威公证人”的角色，负责发放和管理数字证书的权威机构（你也可以签发证书，不权威罢了），它作为受信任的第三方，承担公钥体系中公钥的合法性检验的责任。可是，这里和现实世界仍然有一些区别，现实世界你去找的公安局，那大楼不大可能是剧场布景冒认的；而网络世界，在假设所有网络传输都有可能被截获、冒认的前提下，“去CA中心进行认证”本身也是一种网络操作，这与之前的“去获取去公钥”本质上不是没什么差别吗？其实还是有差别的，公钥成千上万不可数，而权威的CA中心则应是可数的，“可数的”意味着可以不通过网络，而在浏览器、操作系统出厂时预置好，或者在专门安装（如银行的证书），下图为我机器上的现存的根证书。



Windows系统的CA证书

到这里终于出现了一个这节的关键词之一：证书（Certificate），证书是权威CA中心对特定公钥信息的一层公证载体，由于客户的机器上已经预置了这些权威CA中心本身的证书（称为根证书），使得我们能够在不依靠网络的前提下，使用里面的公钥信息对其所签发的证书中的签名进行确认。到此终于打破了鸡生蛋、蛋生鸡的循环，使得整套数字签名体系有了逻辑基础。

PKI中采用的证书格式是[X.509标准格式](#)，它定义了证书中应该包含哪些信息，并描述了这些信息是如何编码的，里面最关键的就是认证机构的数字签名和公钥信息两项内容。一个证书具体包含以下内容：

- 版本号（Version）**：指出该证书使用了哪种版本的X.509标准（版本1、版本2或是版本3），版本号会影响证书中的一些特定信息，目前的版本为3。

2. **序列号 ( Serial Number )** : 标识证书的唯一整数 , 由证书颁发者分配的本证书的唯一标识符。
3. **签名算法标识符** : 用于签证书的算法标识 , 由对象标识符加上相关的参数组成 , 用于说明本证书所用的数字签名算法。例如 , SHA1和RSA的对象标识符就用来说明该数字签名是利用RSA对SHA1杂凑加密。
4. **认证机构的数字签名** : 这是使用证书发布者私钥生成的签名 , 以确保这个证书在发放之后没有被篡改过。
5. **认证机构** : 证书颁发者的可识别名 , 是签发该证书的实体唯一的CA的X.500名字。使用该证书意味着信任签发证书的实体 ( 注意 : 在某些情况下 , 比如根或顶级CA证书 , 发布者自己签发证书 )。
6. **有效期限 ( Validity )** : 证书起始日期和时间以及终止日期和时间 ; 指明证书在这两个时间内有效。
7. **主题信息 ( Subject )** : 证书持有人唯一的标识符 ( Distinguished Name ) 这个名字在互联网上应该是唯一的。
8. **公钥信息 ( Public-Key )** : 包括证书持有人的公钥、 算法(指明密钥属于哪种密码系统)的标识符和其他相关的密钥参数。
9. **颁发者唯一标识符 ( Issuer )** : 标识符—证书颁发者的唯一标识符 , 仅在版本2和版本3中有要求 , 属于可选项。

## 传输安全层

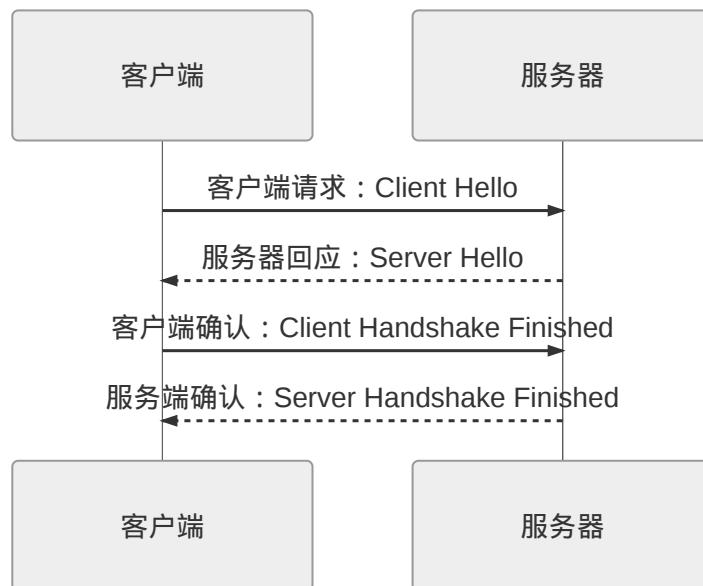
尽管到此为止 , 数字签名的安全性已经可以自洽了 , 但相信你也感受到了这套信任链的繁琐 , 如果从确定加密算法、 生成密钥、 公钥分发、 CA认证、 核验公钥、 签名、 验证签名这些步骤都要由用户来承担的话 , 这样意义的“安全”估计只能一直是存于实验室中的阳春白雪。如何把这一套繁琐的技术体系自动化地应用于无处不在的网络通讯之中 , 是这一节要讨论的话题。

在计算机科学里 , 隔离复杂性的常用手段之一就是分层 , 在网络中更是如此 , OSI模型、 TCP/IP模型将网络从物理特性 ( 比特流 ) 开始 , 逐层封装隔离 , 到了HTTP协议这种面向应用的协议里 , 就已经不会去关心网卡/交换机如何处理数据帧、 MAC地址 ; 不会去关心ARP如何做地址转换 ; 不会去关心IP寻址、 TCP传输等等 ; 那要在网络中让用户无感知地、 自动地安全通讯 , 最合理的做法就是在传输层之上、 应用层之下加入一层安全层来实现 , 这样对上层原本基于HTTP的Web应用来说 , 几乎可以是毫无感知的。而构建传输安全层这件

事情，可以说是和万维网的历史一样长，早在1994年，就已经有公司开始着手去尝试了，这里先简单回顾这将近30年来的进展：

- 1994年，网景（Netscape）公司开发了SSL协议（Secure Sockets Layer）的1.0版，这是构建传输安全层的起源，但是SSL 1.0并未正式对外发布。
- 1995年，Netscape把SSL升级到2.0版，正式对外发布，但是刚刚发布不久就被发现有严重漏洞，所以并未大规模使用。
- 1996年，修补好漏洞的SSL 3.0对外发布，这个版本得到了广泛的应用，成为Web网络安全层的事实标准。
- 1999年，互联网标准化组织接替Netscape，将SSL改名TLS（Transport Layer Security）后推进为国际标准，第一个正式的版本是[RFC 2246](#) 定义的TLS 1.0。这个版本的TLS生存时间极长，直至我写下这段文字的2020年3月，主流浏览器（Chrome、Firefox、IE、Safari）才刚刚共同宣布停止TLS 1.0/1.1的支持。而讽刺的是，由于停止后许多政府网站被无法被浏览，此时又正值新冠病毒（COVID-19）爆发期，Firefox紧急发布公告[宣布撤回该改动](#)，TLS 1.0的生命还在顽强延续。
- 2006年，TLS的第一个升级版1.1发布（[RFC 4346](#)），但却沦为了被遗忘的孩子，很少人使用TLS 1.1，甚至到了因此该版本没有已知的协议漏洞被提出的程度。
- 2008年，TLS 1.1发布2年之后，TLS 1.2标准发布（[RFC 5246](#)），迄今超过90%的互联网HTTPS流量是由TLS 1.2所支持的，现在仍在使用的浏览器几乎都完美支持了该协议。
- 2018年，最新的TLS 1.3（[RFC 8446](#)）发布，比起前面版本相对温和的升级，TLS 1.3做了出了一些激烈的改动，包括修改了从1.0起一直没有大变化的2轮4次（2-RTT）握手，首次连接仅需1轮（1-RTT）即可完成，在连接复用时甚至将TLS 1.2原本的1-RTT下降到了0-RTT，显著提升了访问速度。

下面笔者将以TLS 1.2为例，介绍传输安全层是如何保障所有信息都是第三方无法窃听（加密传输）、无法篡改（一旦篡改通讯算法会立刻发现）、无法冒充（证书验证身份）的。TLS 1.2在传输之前的握手过程一共需要进行上下2轮、4次信息发送，时序如下所示：



在介绍这四次握手具体会做什么之前，先推荐一个制作很用心的网站（<https://tls.ulfheim.net/>），上面以网页的方式详细解释了每一次握手过程中所做的事情、发送的数据、收到的响应等内容。然后，让我们开始一段相对要枯燥困难一些的握手过程：

### 1. 客户端请求：Client Hello

客户端（一般就是浏览器了）向服务器请求进行加密通讯，在这个请求里面，它会以明文的形式，向服务端提供以下信息：

- 支持的协议版本，譬如TLS 1.2。但是要注意，1.0-3.0分别代表SSL1.0-3.0，TLS1.0则是3.1，一直到TLS1.3的3.4。
- 一个客户端生成的32 bytes随机数，这个随机数将稍后用于产生加密的密钥。
- 一个SessionID（可选），不要和前面Cookie-Session那套混淆了，这个SessionID是只链接的SessionID，是为了TLS的链接复用。
- 一系列支持的密码学算法套件，它应该是一组算法的组合，例如TLS\_RSA\_WITH\_AES\_128\_GCM\_SHA256，代表着密钥交换算法是RSA，加密算法是AES128-GCM，消息认证码算法是SHA256
- 一系列支持的压缩算法。
- 其他可扩展的信息，为了保证协议的稳定，后续对协议的功能扩展大多都添加到这个变长结构中。譬如TLS 1.0中由于发送的数据并不包含服务器的域名地址，导致了一台服务器只能安装一张数字证书，这对虚拟主机来说就很不方便，所以TLS 1.1起就增加了名为“Server Name”的扩展，以便一台服务器给不同的站点安装不同的证书。

## 2. 服务器回应 : Server Hello

服务端收到客户单的通讯请求后，如果客户端支持的协议版本、加密算法组合在服务端中能找到一致的，将向客户端发出回应。如果不行，将会返回一个握手失败的警告提示。这次回应同样是明文的，包括以下信息：

- 服务端确认使用的协议版本。
- 第二个32 bytes的随机数，稍后用于产生加密的密钥。
- 一个SessionID，以后链接复用可以减少一轮握手。
- 服务端选定的密码学算法套件。
- 服务端选定的压缩方法。
- 其他可扩展的信息。
- 如果协商出的加密算法组合是依赖证书认证的，服务端要发送出自己的X.509证书，而证书中的公钥是什么，也必须根据协商的加密算法组合来决定。
- 密钥协商消息，这部对于不同密码学套件有不同的价值，譬如对于ECDH + anon这样得密钥协商算法组合（基于椭圆曲线的ECDH算法可以在双方通讯都公开的情况下协商出一组只有通讯双方知道的密钥）就不依赖证书中的公钥，而是通过Server Key Exchange消息协商出密钥。

## 3. 客户端确认 : Client Handshake Finished

由于密码学套件的组合复杂多样，这里仅以RSA算法为密钥交换算法为例介绍后续过程。客户端收到服务器应答后，先要验证服务器证书。如果证书不是可信机构颁布、或者证书中信息存在问题（域名与实际域名不一致、或者证书已经过期、或通过[在线证书状态协议](#)得知证书已被吊销，等等），都会向访问者显示一个“证书不可信任”的警告，由其选择是否还要继续通信。如果证书没有问题，客户端就会从证书中取出服务器的公钥，并向服务器发送以下信息：

- 客户端证书（可选）。部分服务端并不是面向全公众，只对特定的客户端提供服务，此时客户端需要发送它自身的证书，如果不发送，或者验证不通过，服务端可执行决定是否要继续握手，或者返回一个握手失败的信息。
- 第三个32 bytes的随机数，这个随机数不再是明文发送，而是以服务端传过来的公钥加密的，它被称为PreMasterSecret，将与前两次发送的随机数一起，根据[特定算法](#)计算出48 bytes的MasterSecret，此即为后续内容传输时的对称加密算法所采用的私钥。
- 编码改变通知，表示随后的信息都将用双方商定的加密方法和密钥发送。

- 客户端握手结束通知，表示客户端的握手阶段已经结束。这一项同时也是前面发送的所有内容的哈希值，用来供服务器校验。

#### 4. 服务端确认：Server Handshake Finished

服务端向客户端回应最后的确认通知，包括以下信息：

- 编码改变通知，表示随后的信息都将用双方商定的加密方法和密钥发送。
- 服务器握手结束通知，表示服务器的握手阶段已经结束。这一项同时也是前面发送的所有内容的哈希值，用来供客户端校验。

至此，整个握手阶段全部结束，一个安全的链接已经建立，每一个链接建立时，客户端和服务端均通过上面的握手过程协商出了一个只有双方才知道的随机产生的密钥，以及后面传输过程中要采用的对称加密算法（如例子中的AES128），此后该链接的通讯将使用此密钥和加密算法进行加、解密。这种处理方式对上层协议的功能上完全没有影响（性能上当然有影响），建立在这层安全传输层之上的HTTP协议，就被称为“HTTP Over SSL/TLS”，即大家熟知的HTTPS。

从上面握手协商的过程中我们还可以得知，HTTPS同样并非是离散的二元选项，不是只有“启用了HTTPS”和“未启用HTTPS”的差别，采用不同的协议版本、不同的密码学套件、证书是否有效、服务端/客户端对面对无效证书时的处理策略如何都导致了不同HTTPS站点的安全强度的不同。你可以使用[亚洲诚信](#)的诊断服务查看以下几个网站的安全评分，以对安全强度有更加量化直观的理解：

- 亚洲诚信：<https://myssl.com/myssl.com>
- 腾讯网：<https://myssl.com/www.qq.com>
- 本站：<https://myssl.com/icyfenix.cn>
- 趣店：<https://myssl.com/www.quqianbao.com>

# 验证

## 验证 (Verification)

系统如何确保提交到每项服务中的数据是合乎规则的，不会对系统稳定性、数据一致性、正确性产生风险？

一般认为，数据验证不归属在安全这个话题中，但请相信我，从数量来讲，数据验证不严谨而导致的安全问题比其他安全攻击导致的要多得多；而风险上讲，由数据质量导致的问题，风险有高有低，真遇到高风险的数据问题，导致的损失不一定比被拖库什么来的小。

不过，相比起其他富有挑战性的安全措施，防御与攻击两者缠斗的精彩，数学、心理、社会工程和计算机等跨学科知识的结合运用，数据验证倒确实是不可否认地有些无聊枯燥的，这是一项非常常见的工作，在日常的开发中贯穿于代码的各个层次，每个程序员都肯定写过。以架构者的视角，这种常见的代码反而是迫切需要被架构约束的，缺失的校验影响数据质量，过度的校验不会使得系统更加健壮，某种意义上反而是垃圾代码，甚至有副作用。来看看下面这个段子：

前 端： 提交一份用户数据（姓名：某， 性别：男， 爱好：女， 签名：xxx， 手机：xxx， 邮箱：null）

控制器： 发现邮箱是空的，抛ValidationException("邮箱没填")

前 端： 已修改，重新提交

安 全： 发送验证码时发现手机号少一位，抛RemoteInvokeException("无法发送验证码")

前 端： 已修改，重新提交

服务层： 邮箱怎么有重复啊，抛BusinessRuntimeException("不允许开小号")

前 端： 已修改，重新提交

持久层： 签名字段超长了插不进去，抛SQLException("插入数据库失败，SQL : xxx")

.....

前 端： 你们这些坑爹挖不管埋的后端，各种异常都往前抛！

用 户： 这系统牙膏厂生产的？

最基础的数据问题可以在前端做表单校验来处理，但后端验证肯定是要做的，上面的段子看完了想一想，服务端应该在哪一层去做校验？可能会有这样的答案：

- 在Controller层做，在Service层不做。理由是从Service开始会有同级重用，出现Service A.foo(params)调用ServiceB.bar(params)时，相当于对params重复校验了两次。
- 在Service层做，在Controller层不做。理由是无业务含义的格式校验已在前端表单验证处理过，有业务含义的校验，放在Controller层无论如何不合适。
- 在Controller、Service层各做各的。Controller做格式校验，Service层做业务校验，就是上面那段子中的行为。
- 还有其他一些意见，譬如还有提在持久层做校验，理由是这是最终入口，把守好写入数据库的质量最重要。

上述的讨论大概是没有统一的正确结论，但是在Java里确实是有验证的标准做法，提倡的是把校验行为从分层中剥离出来，不是在哪一层做，而是在Bean上做。即Java Bean Validation。从2009年的[JSR 303](#)的1.0，到2013年的[JSR 349](#)更新的1.1，到目前最新的2017年发布的[JSR 380](#)，定义了Bean验证的全套规范。单独将验证提取、封装，可以获得不少好处：

- 对于无业务含义的格式验证，可以做到预置。
- 对于有业务含义的业务验证，可以做到重用。一个Bean适用于多个方法是非常常见的。
- 利于集中管理，譬如统一认证的异常体系，统一做国际化、统一给客户端的返回格式等等。
- 避免对输入数据的防御污染到业务代码，如果你的代码里面如果很多下面这样的条件判断，应该考虑重构

```
// 一些已执行的逻辑
if (someParam == null) {
 throw new RuntimeException("客官不可以！")
}
```

java

- 利于多个校验器统一执行，统一返回校验结果，避免用户踩地雷、挤牙膏式的试错体验。

其实，据我了解，国内的项目使用Bean Validation的还是不少的，但多数都只使用到它的Built-In Constraint，即下面这堆注解（含义我就不写了，用处基本上看类名就能明白）：

```
java
@Null、@NotNull、@AssertTrue、@AssertFalse、@Min、@Max、@DecimalMin、
@DecimalMax、@Negative、@NegativeOrZero、@Positive、@PositiveOrZero、
@Szie、@Digits、@Pass、@PassOrPresent、@Future、@FutureOrPresent、
@Pattern、@NotEmpty、@NotBlank、@Email
```

一般实现会采用Hibernate Validator，另外一个非主流选择是Apache BVal，它们都扩展了自己的私有注解。其中有一些注解，像@Email、@NotEmpty、@NotBlank，从以前Hibernate Validator私有注解，随着版本升级转正成为标准。

但是其中多数项目对Bean Validation的使用就到此为止了，带业务含义的代码都还是习惯写到方法体内，导致完全没法管理。其实这部分带有复杂逻辑的校验，才是最需要约束的，更加应该借助Bean Validation来完成。以Fenix's Bookstore的在用户资源上的两个方法为例：

```
java
/**
 * 创建新的用户
 */
@POST
public Response createUser(@Valid @UniqueAccount Account user) {
 return CommonResponse.op(() -> service.createAccount(user));
}

/**
 * 更新用户信息
 */
@PUT
@CacheEvict(key = "#user.username")
public Response updateUser(@Valid @AuthenticatedAccount
@NotConflictAccount Account user) {
 return CommonResponse.op(() -> service.updateAccount(user));
}
```

注意其中的三个自定义校验注解，它们的含义分别是：

- @UniqueAccount：传入的用户对象必须是唯一的，不与数据库中任何已有用户的名称、手机、邮箱产生重复。
- @AuthenticatedAccount：传入的用户对象必须与当前登陆的用户一致。
- @NotConflictAccount：传入的用户对象中的信息与其他用户是无冲突的，譬如将一个注册用户的邮箱，修改成与另外一个已存在的注册用户一致的值，这便是冲突。

这里的需求很容易想明白，注册新用户时，应约束不与任何已有用户的关键信息重复；而修改自己的信息时，只能与自己的信息重复，而且只能修改当前登陆用户的信息。这些约束规则不仅仅为这两个方法服务，它们可能会在用户资源中的其他入口被使用到，甚至在其他分层的代码中被使用到。下面是这三个自定义注解对应校验器的实现类：

```
java
public static class AuthenticatedAccountValidator extends
AccountValidation<AuthenticatedAccount> {
 public void initialize(AuthenticatedAccount constraintAnnotation) {
 predicate = c -> {
 AuthenticAccount loginUser = (AuthenticAccount)
SecurityContextHolder.getContext().getAuthentication().getPrincipal();
 return c.getId().equals(loginUser.getId());
 };
 }
}

public static class UniqueAccountValidator extends
AccountValidation<UniqueAccount> {
 public void initialize(UniqueAccount constraintAnnotation) {
 predicate = c ->
!repository.existsByUsernameOrEmailOrTelephone(c.getUsername(),
c.getEmail(), c.getTelephone());
 }
}

public static class NotConflictAccountValidator extends
AccountValidation<NotConflictAccount> {
 public void initialize(NotConflictAccount constraintAnnotation) {
 predicate = c -> {
 Collection<Account> collection =
repository.findByUsernameOrEmailOrTelephone(c.getUsername(),
c.getEmail(), c.getTelephone());
 // 将用户名、邮件、电话改成与现有完全不重复的，或者只与自己重复的，就
不算冲突
 return collection.isEmpty() || (collection.size() == 1 &&

```

```

 collection.iterator().next().getId().equals(c.getId())));
 }
}
}

```

这样业务校验便和业务逻辑分离开来，在需要使用时用@Valid注解自动或者通过代码手动触发执行，可根据你们公司的要求，使用于控制器、服务层、持久层等任何层次之中。此外，校验结果不满足时的提示信息，也便于统一处理，如提供默认值、提供国际化支持（这里没做）、提供统一的客户端返回格式（创建一个用于ConstraintViolationException的异常处理器），以及批量执行全部校验避免挤牙膏等诸多好处。下面是预置默认提示信息的例子：

```

/**
 * 表示一个用户的信息是无冲突的
 *
 * “无冲突”是指该用户的敏感信息与其他用户不重合，譬如将一个注册用户的邮箱，修改成
与另外一个已存在的注册用户一致的值，这便是冲突
 */
@Documented
@Retention(RUNTIME)
@Target({FIELD, METHOD, PARAMETER, TYPE})
@Constraint(validatedBy =
AccountValidation.NotConflictAccountValidator.class)
public @interface NotConflictAccount {
 String message() default "用户名、邮箱、手机号码与现存用户产生重复";
 Class<?>[] groups() default {};
 Class<? extends Payload>[] payload() default {};
}

```

另外一条建议是将不带业务含义的格式校验注解放到类上，将带业务含义的注解放到外面。譬如用户账号实体中的部分代码为：

```

public class Account extends BaseEntity {
 @NotEmpty(message = "用户名不允许为空")
 private String username;

 @NotEmpty(message = "用户姓名不允许为空")
 private String name;
}

```

```
private String avatar;

@Pattern(regexp = "1\\d{10}", message = "手机号格式不正确")
private String telephone;

@email(message = "邮箱格式不正确")
private String email;
}
```

把校验注解放在类定义中，意味着所有执行校验的时候它们都会被运行（譬如Insert、Update的时候，Hibernate都会自动执行DO上的校验注解）。而不带业务含义的注解运行是不需要其他外部资源（譬如数据库）参与的，这种重复执行通常并无坏处（系统的压力往往不在CPU，闲着也是闲着）。

如果真的遇到一些非典型情况，譬如“新增”操作A需要执行全部校验规则，“修改”操作B中希望不校验某个字段，“删除”操作C中希望改变某一条校验规则，这时候要就要启用分组校验来处理，设计一套“新增”、“修改”、“删除”这样的标识类，置入到校验注解的groups参数中。

# 漏洞利用

编写中

## 漏洞利用 (Exploit)

系统如何避免在基础设施和应用程序中出现弱点，被攻击者利用？

# 高效并发

# 进程、线程与协程

# 线程安全

# 同步机制

## 阻塞同步

### 锁的属性

公平性

互斥性

可重入性

## 非阻塞同步

## 无同步机制

# 硬件并发机制

# 系统分层

# 容量规划

# 非功能属性设计

可测试性

---

可约束性

---

高可用设计

---

高并发设计

---

一致性设计

---

反脆弱设计

---

扩展性设计

---

# 分布式共识算法

## 前置知识

关于分布式中CAP问题，请先阅读“[分布式事务](#)”中的介绍，后文中提及的一致性、可用性、网络分区等概念，均在此文中有过介绍。

在本章正式开始探讨各种分布式环境中面临的技术问题和解决方案之前，笔者先安排一篇“纯理论”的文章，来分析分布式环境中对共享数据操作的本质。分布式系统里，如果准备在各个分布式节点中进行一致的操作，并且期望获得一致的结果，均可以理解为是一种“[状态机复制](#)”（State Machine Replication）过程，无论这个操作是新增、修改、删除抑或是其他可能的程序行为，都可以理解为要将一连串的操作日志正确地复制到各个分布式节点上，如果分布式系统各个节点的初始状态一致，接受到的操作序列都相同，那各个节点最终都能得到一致的状态。这句话听起来颇为抽象，如果你现在暂时不能理解的话，不妨先在心中回想一下经典数据库中的重做和回滚日志的做法，然后跟后续的讲解进行类比。

为了解释清楚分布式环境中共享数据所面临的问题，笔者先从一个最浅显的场景开始说起：

如果你有一份很重要的数据，要确保它长期存储在电脑上不会丢失，你会怎么做？

这不是什么脑筋急转弯的古怪问题，答案就是去买几块硬盘，把数据在不同磁盘上多备份几个副本。假设一块硬盘每年损坏的概率是5%，那把文件复制到另一块备份盘上，由于两块磁盘同时损坏而丢失数据的概率就只有0.25%，如果使用三块硬盘存储则是0.00125%，四块是0.0000625%，换而言之，这已经保证了数据超过99.9999%的概率是不会丢失的。

在软件系统里，要保障系统的可靠性，采用的办法也和那几个备份磁盘大体上并无区别，单个节点的系统宕机无法提供服务的原因可能有很多，譬如程序出错、硬件损坏、网络分区、电源断电，等等，往往一年中出现系统宕机的概率要远高于5%，这更加决定了软件系统也必须有多台机器能够拥有一致的数据，才能对外提供一致的服务。但分布式的软件系

统与备份磁盘又有着本质的区别，磁盘之间是孤立的，不需要互相通讯，备份数据初始化后状态就是不变的，由人工完成的文件复制操作保障了数据各个副本的一致；而分布式系统里面，我们必须考虑数据如何在可能出现分区的网络环境下在各个节点之间正确复制的问题：

如果你有一份很重要的数据，要确保它正确地存储于网络中的几台不同机器之上，你会怎么做？

一个最容易想到的答案是“同步”（Synchronous）：每当数据有变化，把变化情况在各个节点间的复制视作一种原子性的操作，只有系统里每一台机器都反馈成功地完成磁盘写入后，数据的变化才宣告成功，我们在前文中曾经讲解过，可以使用2PC/3PC来实现这种同步操作。同步的其中一种真实应用场景是数据库中的主从全同步复制（Fully Synchronous Replication），譬如MySQL Cluster，进行全同步复制时，会等待所有Slave的Binlog都完成写入后，Master的事务才进行提交。这里有一个显而易见的问题，尽管可以确保Master和Slave中的数据是绝对一致的，但任何一个Slave节点因为任何原因未响应均都会阻塞整个事务，每增加一个Slave，都导致造成整个系统可用性风险增加一分。显然这种简单的一致性保障手段，是以完全牺牲可用性为代价的，我们在建设分布式系统的时候，往往不能承受这种代价，一些关键系统，在要求数据正确的前提下，对可用性的要求也非常苛刻，譬如要达到99.99999%可用的程度，这就引出了我们的第三个问题：

如果你有一份很重要的数据，要确保它正确地存储于网络中的几台不同机器之上，并且要尽可能保证及时地应用到正确的数据，你会怎么做？

在网络分区不可能消除的前提下，一致性与可用性的矛盾造成了增加机器数量反而带来可用性的降低，为缓解这个矛盾，我们不再追求系统内所有节点在任何情况下的数据状态都一致，改为采用“少数服从多数”的原则，一旦数据变化在系统中过半数的节点中完成了复制，就认为数据的变化已经正确地存储在系统当中，这样就可以容忍少数（不超过半数）的机器形成网络分区，使得增加机器数量对系统整体的可用性变成是有益的。此模式在分布式中被称为“Quorum机制”，或者可以直接形象地称其为“多数派机制”（Majority）。在这个前提下，我们需要设计一种算法，能够让分布式系统内部可以暂时容忍存在不同的状态，但最终全部节点的状态均能够达成一致；同时能够让分布式系统外部看来，始终表现出整体一致的结果，这个过程，我们称其为“协商共识”（Consensus）。

请注意共识与一致性（Consistency）的区别，一致性指的是数据不同副本之间的差异，而共识是指达成一致性的方法与过程。由于翻译的关系，很多中文资料把Consensus同样翻

译为“一致性”，导致网络上大量的二手中文资料把这两个概念混淆起来，如果你在网上看到“分布式一致性算法”，应明白其所指其实是“Distributed Consensus Algorithm”。

# Paxos

## Distributed Consensus Algorithm

There is only one consensus protocol, and that's "Paxos" — all other approaches are just broken versions of Paxos

世界上只有一种共识协议，就是Paxos，其他所有共识算法都是Paxos的退化版本。

—— Mike Burrows [↗](#) , Inventor of Google Chubby

Paxos是由[Leslie Lamport](#) [↗](#)（就是大名鼎鼎的[LaTeX](#) [↗](#)中的“La”）提出的一种基于消息传递的协商共识算法，现已是当今分布式系统最重要的理论基础，几乎就是“共识”二字的代名词（这句话是Raft作者在论文中说的）。尽管不像Mike Burrows说的“世界上只有Paxos一种分布式共识算法”那么夸张，但是如果沒有Paxos，那后续的Raft、ZAB算法，ZooKeeper、etcd这些分布式协调框架、Hadoop、Consul等在此基础上的各类分布式应用都很可能会延后几年面世。

Lamport虚构了一个名为“Paxos”的希腊城邦，这个城邦按照民主制度制定法律，却又不存在一个中心化的专职立法机构，立法靠着“兼职议会”（Part-Time Parliament）来完成，无法保证所有城邦居民都能够及时地了解新的法律提案、也无法保证居民会及时为提案投票。Paxos算法的目标就是让城邦能够在每一位居民都无法承诺一定会及时参与的情况下，依然可以按照少数服从多数的原则，最终达成一致意见（但是并不考虑[拜占庭将军问题](#) [↗](#)，即假设信息可能丢失也可能延迟，但不会被错误传递）。

Lamport最初在1990年首次发表了Paxos算法，选的题目就是“The Part-Time Parliament”[↗](#)。由于算法本身较为复杂，用希腊城邦作为比喻反而使得描述更为晦涩，论文的三个审稿人一致要求他把希腊城邦的故事删除掉，这令Lamport感觉颇为不爽，然后干脆就撤稿不发了，所以Paxos刚刚被提出的时候并没有引起什么反响。八年之后（1998年），Lamport再次将此文章重新整理后投到《ACM Transactions on Computer Systems》[↗](#)，这次论文

成功发表，吸引了一些人去研究，结果是并没有什么人能弄懂。时间又过去了三年（2001年），Lamport认为前两次是同行们无法理解他以“希腊城邦”来讲故事的幽默感，第三次以“[Paxos Made Simple](#)”为题，在《[SIGACT News](#)》杂志上发表文章，终于放弃了“希腊城邦”的比喻，尽可能用（他认为）简单直接、（他认为）可读性较强的方式去介绍Paxos算法，情况虽然比前两次要好，但以Paxos本应获得的重视程度来说，这次依然只能算是应者寥寥。这段跟网络段子一般的经历被Lamport以自嘲的形式放到了[他自己的个人网站](#)上。尽管我们作为后辈应该尊重Lamport老爷子，但当笔者翻开“Paxos Made Simple”读到只有一句话的“摘要”时，心里实在是不得不怀疑Lamport这样写论文是不是在恶搞审稿人和读者，在嘲讽“你们这些愚蠢的人类！”。

### Abstract

The Paxos algorithm, when presented in plain English, is very simple.

Paxos Made Simple ↗

虽然Lamport本人连发三篇文章都没能让大多数同行理解Paxos，但是到了2006年，Google的Chubby、Megastore以及Spanner等分布式系统都使用Paxos解决了分布式共识的问题，并将其整理成正式的论文发表之后，得益于Google的行业影响力，辅以Chubby作者Mike Burrows那略显夸张但足够吸引眼球的评价推波助澜，致使Paxos一夜间成为计算机科学分布式这条分支中最炙手可热网红概念，从这时起被学术界众人争相研究。2013年，La

import本人因其对分布式系统的杰出理论贡献获得了2013年的图灵奖，足可见技术圈里即使再有本事，也还是需要好好包装一下的。

讲完段子吃过西瓜，希望你没有被这些对Paxos的“困难”做的铺垫所吓倒，反正又不让你去实现它，假如放弃些许严谨性，并简化分支细节和特殊情况的话，Paxos是完全可能去通俗地理解的，Lamport在论文中也只用两段话就描述“清楚”了它的工作流程，下面我们正式来学习Paxos算法（在本节中均特指Basic Paxos算法）。Paxos算法将分布式系统中的节点分为三类：

- **提案节点**（称为Proposer）：提出对某个值进行设置操作的节点，设置值这个行为就被称为“提案”（Proposal）。请注意，这里的“设置值”不要类比成程序中变量赋值操作，应该类比成日志记录操作，值一旦设置成功，就是不丢失、不可变的，在后面介绍的Raft算法中就索性直接把“提案”叫做“日志”了。
- **决策节点**（称为Acceptor）：应答提案节点该提案是否可被投票、是否可被接受。提案一旦得到过半数决策节点的接受，即称该提案被批准，提案被批准即意味着该值不能再被更改，也不会丢失，且最终所有节点都会接受该它。
- **记录节点**（被称为Learner）：不参与提案，也不参与决策，只是单纯地从提案、决策节点中学习已经达成一致的提案，譬如少数派节点从网络分区中恢复时，将会进入这种状态。

使用Paxos的分布式系统里的，所有的节点都是平等的，它们都可以承担以上某一种或者多种的角色，不过为了便于确保有明确的多数派，决策节点的数量应该被设定为奇数个，且在初始化时，网络中每个节点都知道整个网络所有决策节点的数量、地址等信息。

分布式环境下，我们如果说各个节点“就某个值（提案）达成一致”，所指的意思是“不存在某个时刻有一个值为A，另一个时刻该值又为B的情景”。解决这个问题的复杂度主要来源于以下两个方面因素的共同作用：

- 系统内部各个节点通讯是不可靠的，不论对于系统中企图设置数据的提案节点抑或是决定是否批准设置行为的决策节点，其发出、收到的信息可能延迟送达、也可能丢失，但不去考虑消息传递错误的情况。
- 系统外部各个用户访问是可并发的，如果系统只会有一个用户，如果每次只对系统进行串行访问，那单纯的应用Quorum机制已经足以保证值被正确地读写。

第一点是网络通讯中客观存在的现象，也是我们要重点解决的问题，而第二点其实也很好理解：现在我们讨论的是“分布式环境下操作的共享数据”的问题，而即使是在非分布式的环境下，有一个变量i当前在系统中存储的数值为2，如果同时有外部请求A、B分别对系统发送指令“把i的值加1”和“把i的值乘3”，如果不加任何并发控制的话，将可能得到“(2+1)\*3=9”与“2\*3+1=7”两种可能的结果。为此，对同一个变量的修改请求必须先“加锁”（实际操作上并不是并发控制中互斥量的加锁），不能让A、B的请求被交替处理。在分布式的环境下，由于还要同时考虑到分布式系统内可能在任何时刻出现的通讯问题，如果一个节点在取得同步锁之后，在释放锁之前发生失联，这将导致无限期的阻塞等待，因此还必须提供一个其他节点能抢占锁的机制，以避免死锁。

Paxos算法包括两个阶段，其中第一阶段“准备”（Prepare）就相当于上面抢占锁的过程。如果某个提案节点准备发起提案，必须先向所有的决策节点广播一个许可申请（称为Prepare请求）。提案节点的Prepare请求中会附带一个全局唯一的数字n作为提案ID，决策节点收到后，将会给予提案节点两个承诺和一个应答。

两个承诺是：

- 承诺不会再接受提案ID小于或等于n的Prepare请求。
- 承诺不会再接受提案ID小于n的Accept请求。

一个应答是：

- 不违背以前作出的承诺的前提下，回复已经批准过的提案中提案ID最大的那个提案所设定的值和提案ID，如果该值从来没有被任何提案设定过，则返回空值。如果违反此前做出的承诺，即收到的提案ID并不是决策节点收到过的最大的，那可以直接对此Prepare请求不予理会。

当提案节点收到了多数派决策节点的应答（称为Promise应答）后，可以开始第二阶段“批准”（Accept）过程，这时有如下两种可能：

- 如果提案节点发现所有响应的决策节点此前都没有批准过该值（即为空），那说明它是第一个设置值的节点，可以随意地决定要设定的值，将自己选定的值与提案ID，构成一个二元组“(n, value)”，再次广播给全部的决策节点（称为Accept请求）。
- 如果提案节点发现响应的决策节点中，已经有至少一个节点的应答中包含有值了，那它就不能够随意取值了，必须无条件地从应答中找出提案ID最大的那个值并接受，构成一个二元组“(n, maxAcceptValue)”，再次广播给全部的决策节点（称为Accept请求）。

当每一个决策节点收到Accept请求时，都会在不违背以前作出的承诺的前提下，接收并持久化对当前提案ID和提案附带的值。如果违反此前做出的承诺，即收到的提案ID并不是决策节点收到过的最大的，那可以直接对此Accept请求不予理会。

当提案节点收到了多数派决策节点的应答（称为Accepted应答）后，协商结束，共识决议形成，将形成的决议发送给所有记录节点进行学习。整个过程的时序如下图所示：

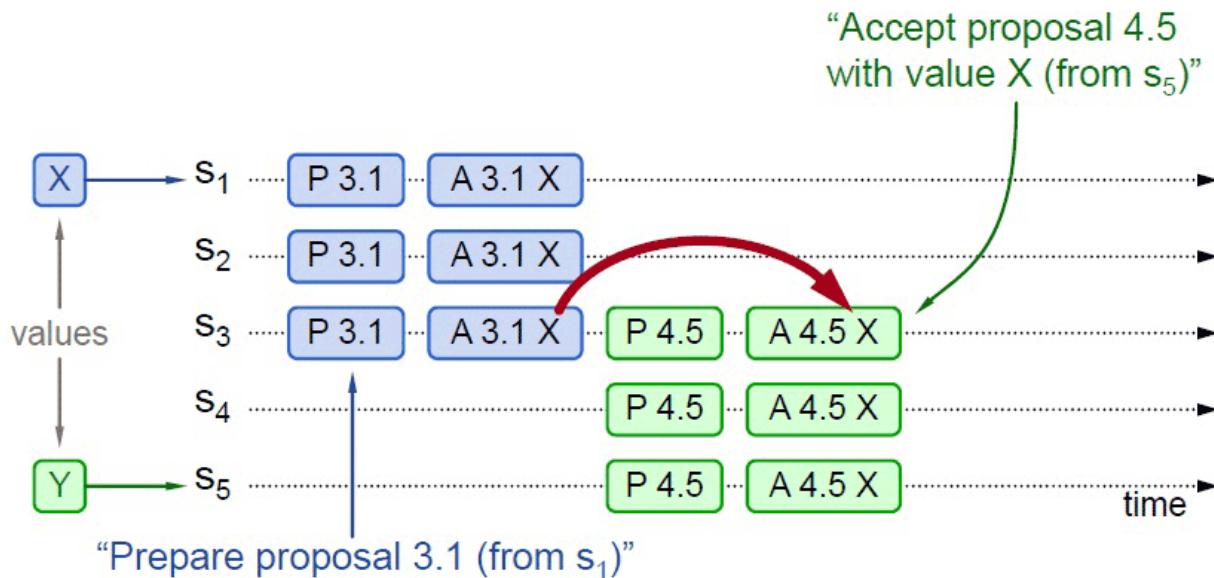


整个Paxos算法的工作流程确实并不复杂，如果你此前并未专门学习过分布式的知识，相信阅读到这里，不一定会对操作过程有疑惑，但估计还是不能对Paxos究竟是如何解决协商共识的形成具体的概念的，下面笔者将举一个具体的例子来讲解，例子与截图来源于《Implementing Replicated Logs with Paxos》[\[1\]](#)，在此统一注明，后面就不单独列出了。

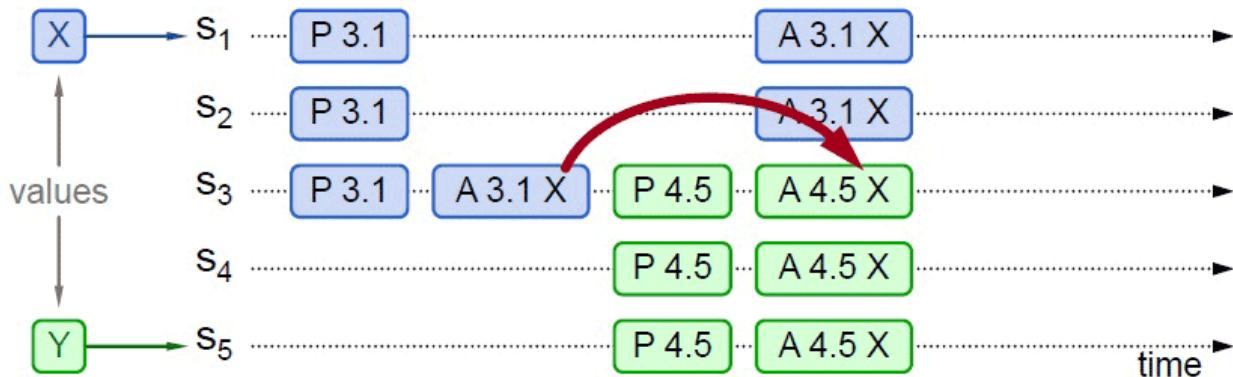
假设一个分布式系统有五个节点，分别命名为S<sub>1</sub>、S<sub>2</sub>、S<sub>3</sub>、S<sub>4</sub>、S<sub>5</sub>，只讨论正常场景，不会涉及到网络分区。全部节点都同时扮演着提案节点和决策节点的身份。此时，有两个并发的请求分别希望将同一个值分别设定为X（由S<sub>1</sub>作为提案节点提出）和Y（由S<sub>5</sub>作为提案节点提出），以P代表准备阶段，以A代表批准阶段，这时候可能发生以下情况：

- 情况一：譬如，S<sub>1</sub>选定的提案ID是3.1（全局唯一ID加上节点编号），先取得了多数派决策节点的Promise和Accepted应答，此时S<sub>5</sub>选定提案ID是4.5，发起Prepare请求，收到的多数派应答中至少会包含1个此前应答过S<sub>1</sub>的决策节点，假设是S<sub>3</sub>，那么S<sub>3</sub>提供的

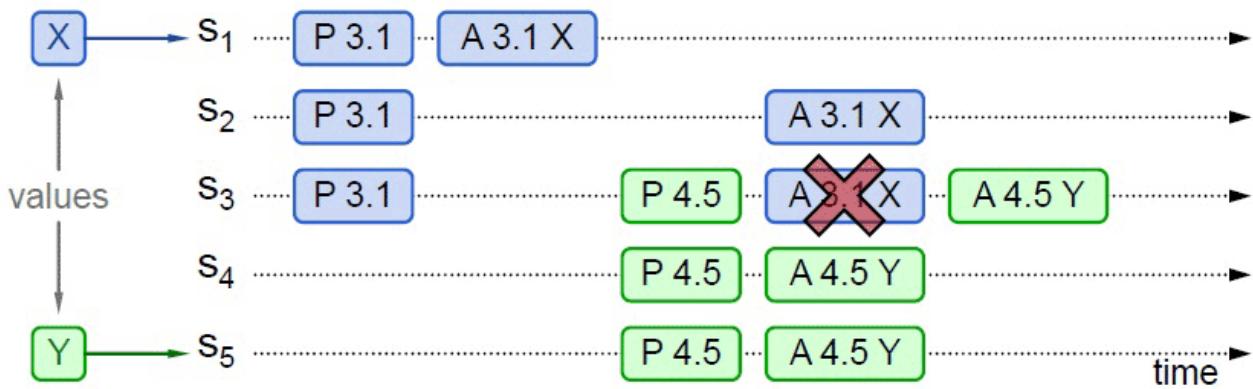
Promise中必将包含S<sub>1</sub>已设定好的值X，S<sub>5</sub>就必须无条件地用X代替Y作为自己提案的值，由此整个系统对“取值为X”这个事实达成了一致。如下图所示：



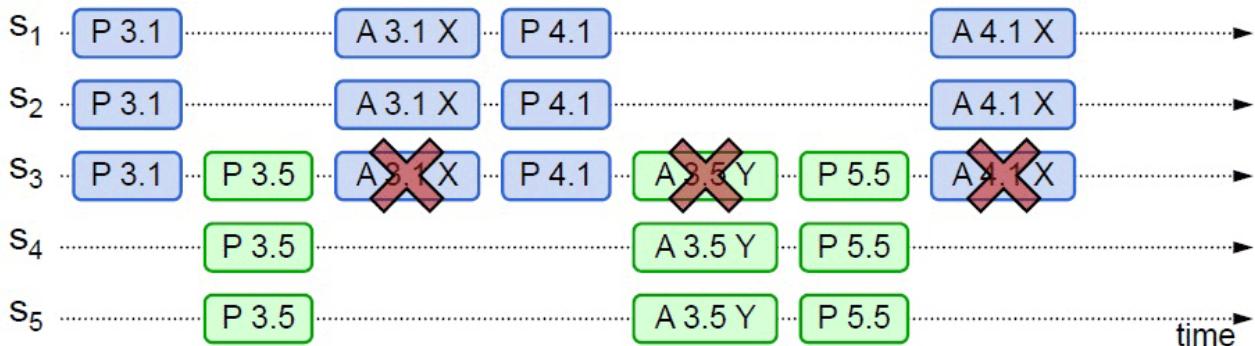
- 情况二：事实上，对于情况一，X被选定为最终结果是必然结果，但从图中可以看出，X被选定为最终值并不是必定需要多数派的共同批准，只取决于S<sub>5</sub>提案时Promise应答中是否已包含了批准过X的决策节点，譬如下图所示，S<sub>5</sub>发起提案的Prepare请求时，X并未获得多数派批准，但由于S<sub>3</sub>已经批准的关系，最终共识的结果仍然是X。



- 情况三：当然，另外一种可能的结果是S<sub>5</sub>提案时Promise应答中并未包含批准过X的决策节点，譬如应答S<sub>5</sub>提案时，节点S<sub>1</sub>已经批准了X，节点S<sub>2</sub>、S<sub>3</sub>未批准但返回了Promise应答，此时S<sub>5</sub>以更大的提案ID获得了S<sub>3</sub>、S<sub>4</sub>、S<sub>5</sub>的Promise，这三个节点均未批准过任何值，那么S<sub>3</sub>将不会再接受来自S<sub>1</sub>的Accept请求，因为它的提案ID已经不是最大的了，这三个节点将批准Y的取值，整个系统最终会对“取值为Y”达成一致。



- 情况四：从以上情况三可以推导出另一种极端的情况，如果两个提案节点交替使用更大的提案ID使得准备阶段成功，但是批准阶段失败的话，这个过程理论上可以无限持续下去，形成活锁（Livelock）。在算法实现中会引入随机超时时间来避免活锁的产生。



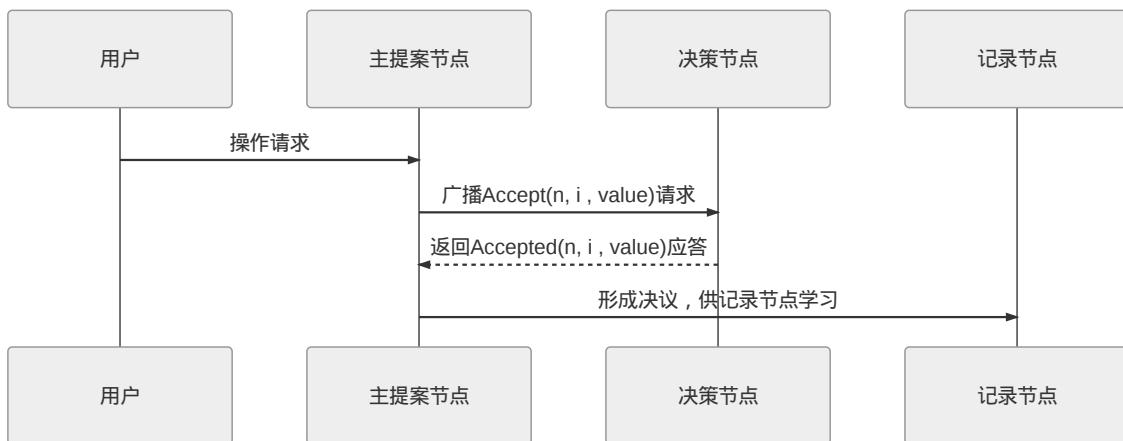
以上介绍是基于Basic Paxos、以正常流程（未出现网络分区等异常）、以通俗的方式介绍的Paxos算法，并未涉及到严谨的逻辑和数学原理，也未讨论Paxos的推导证明过程，对于普通的技术人员，理解起来应该并不算困难的。

本节介绍的Basic Paxos只能对单个值形成决议，并且决议的形成至少需要两次网络请求和应答（准备和批准阶段各一次），高并发情况下将产生较大的网络开销，极端情况下甚至可能形成活锁。总之，Basic Paxos是一种很学术化但对工程并不友好的算法，现在几乎只用来做理论研究，并不直接应用在实际软件研发当中。实际的应用都是基于Multi Paxos和Fast Paxos算法的，接下来我们将会了解Multi Paxos与它的理论等价算法Raft和ZAB算法。

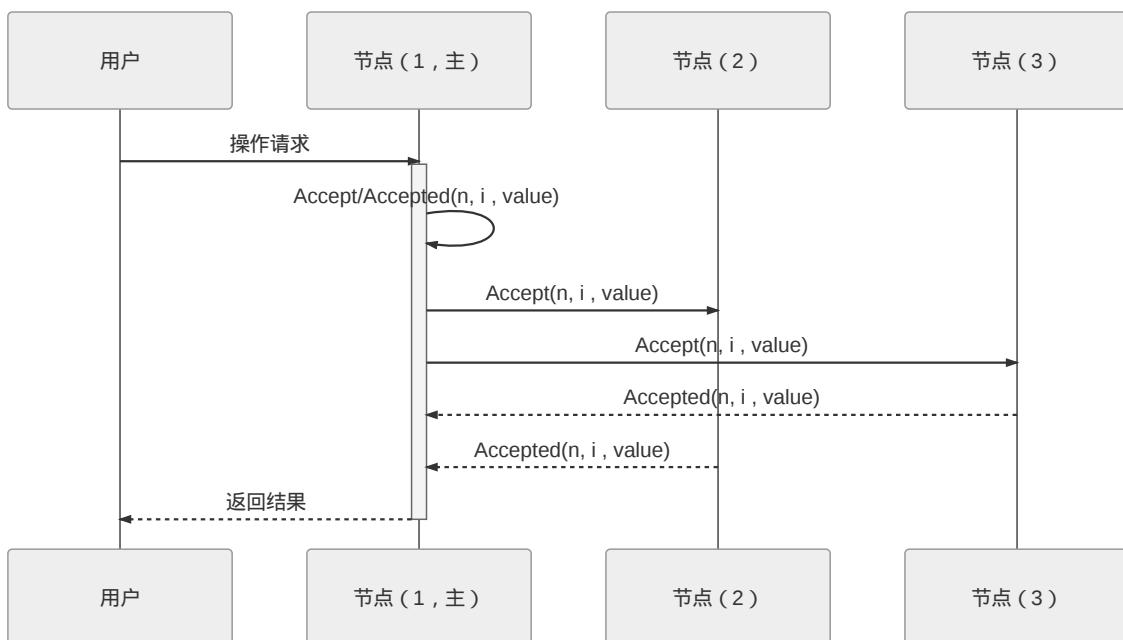
# Multi Paxos与Raft

上一节的最后，笔者举例介绍了Basic Paxos的活锁问题，两个提案节点互不相让地争相提出自己的提案，抢占同一个值的修改权限，导致整个系统在持续性地“反复横跳”，外部看起来就像被锁住了一样。而在上一节的开头，笔者还陈述过一个观点，分布式共识的复杂性，主要来源于网络的不可靠与请求的可并发两大因素，活锁问题与许多Basic Paxos异常场景中所遭遇的麻烦，都可以看作是由于任何一个提案节点都能够完全平等地、与其他节点并发地提出提案而带来的复杂问题。为此，专门有一种Paxos的主要改进版本“Multi Paxos”算法被设计（设计的意思：在Lamport的论文中随意提了几句可以这么做）出来，希望能够找到一种两全其美的办法，既不破坏Paxos中“众节点平等”的原则，又能在提案节点中实现主次之分，限制每个节点都有不受控的提案权利，这两个目标听起来似乎是矛盾的，但现实世界中的选举，就很符合这种在平等节点中挑选意见领袖的场景。

Multi Paxos对Basic Paxos的关键改进是有了“选主”的过程，提案节点会通过定时轮询（心跳），确定当前网络中的所有节点里是否存在有一个主提案节点，一旦没有发现主节点存在，节点就会在心跳超时后使用Basic Paxos中定义的准备、批准的两轮网络交互过程，向所有其他节点广播自己希望竞选主节点的请求，希望整个分布式系统对“由我作为主节点”这件事情协商达成一致共识，如果得到了决策节点中多数派的批准，便宣告竞选成功。当选主完成之后，除非主节点失联之后发起重新竞选，否则从此往后，就只有主节点本身才能够提出提案。此时，无论哪个提案节点接收到客户端的操作请求，都会将请求转发给主节点来完成提案，而主节点提案的时候，也就无需再次经过准备过程，因为可以视作是经过选举时的那一次准备之后，后续的提案都是对相同提案ID的一连串的批准过程。也可以通俗理解为选主过后，就不会再有其他节点与它竞争，相当于是处于无并发的环境当中进行的有序操作，所以此时系统中要对某个值达成一致，只需要进行一次批准的交互即可，具体如下序列所示：



可能有人注意到这时候的二元组( $n, \text{value}$ )已经变成了三元组( $n, i, \text{value}$ )，这是因为需要为主节点增加一个“任期编号”，这个编号必须是严格单调递增的，以应付主节点陷入网络分区后重新恢复，但另外一部分节点仍然有多数派，且已经完成了重新选主的情况，此时必须以任期编号大的主节点为准。当节点有了选主机制的支持，在整体来看，就可以进一步简化节点角色，不去区分提案、决策和记录节点了，统统以“节点”来代替，节点只有主 (Leader) 和从 (Follower) 的区别，此时协商共识的时序如下：



在这个理解的基础上，我们换一个角度来重新思考“分布式系统中如何对某个值达成一致”这个问题，可以把该问题划分做三个子问题来考虑，可以证明（具体证明就不写了，参考文末的论文）当以下三个问题同时被解决时，即等价于达成共识：

- 如何选主 (Leader Election)
- 如何把数据复制到各个节点上 (Entity Replication)
- 如何保证过程是安全的 (Safety)

选主问题尽管还涉及到许多工程上的细节，譬如心跳、随机超时、并行竞选，等等，但要只论原理的话，如果你已经理解了Paxos算法的介绍，相信对选主并不会有什么疑惑，因为这本质上仅仅是分布式系统对“谁来当主节点”这件事情的达成共识而已，我们在前一节已经花了几千字来讲述分布式系统该如何对一件事情达成共识，这里就不继续展开了，下面直接介绍数据（Paxos中的提案、Raft中的日志）在网络各节点间的复制问题。

在正常情况下，当客户端向主节点发起一个操作，譬如提出“将某个值设置为X”，此时主节点将X写入自己的变更日志，但先不提交，接着把变更X的信息在下一次心跳包中广播给所有的从节点，并要求从节点回复确认收到的消息，从节点收到信息后，将操作写入自己的变更日志，然后给主节点发送确认签收的消息，主节点收到过半数的签收消息后，提交自己的变更、应答客户端并且给从节点广播可以提交的消息，从节点收到提交消息后提交自己得变更，数据在节点间的复制宣告完成。

在异常情况下，网络出现了分区，部分节点失联，但只要仍能正常工作的节点的数量能够满足多数派，分布式系统就仍然可以正常工作，这时候数据复制过程如下：

- 假设有 $S_1$ 、 $S_2$ 、 $S_3$ 、 $S_4$ 、 $S_5$ 五个节点， $S_1$ 是主节点，由于网络故障，导致 $S_1$ 、 $S_2$ 和 $S_3$ 、 $S_4$ 、 $S_5$ 之间彼此无法通讯，形成网络分区。
- 一段时间后， $S_3$ 、 $S_4$ 、 $S_5$ 三个节点中的某一个（譬如是 $S_3$ ）最先达到心跳超时的阈值，获知当前分区中已经不存在主节点了，它向所有节点发出自己要竞选的广播，并收到了 $S_4$ 、 $S_5$ 节点的批准响应，加上自己一共三票，即得到了多数派的批准，竞选成功，此时系统中同时存在 $S_1$ 和 $S_3$ 两个主节点，但由于网络分区，它们不会知道对方的存在。
- 这种情况下，客户端发起操作请求：
  - 如果客户端连接到了 $S_1$ 、 $S_2$ 之一，都将由 $S_1$ 处理，但由于操作只能获得最多两个节点的响应，不构成多数派的批准，所以任何变更都无法成功提交。
  - 如果客户端连接到了 $S_3$ 、 $S_4$ 、 $S_5$ 之一，都将由 $S_3$ 处理，此时操作可以获得最多三个节点的响应，构成多数派的批准，是有效的，变更可以被提交，即系统可以继续提供服务。
  - 事实上，以上两种“如果”情景很少机会能够并存。网络分区是由于软、硬件或者网络故障而导致的，内部网络出现了分区，但两个分区仍然能分别与外部网络的客户端正常通讯的情况甚为少见。通常网络中下线了一部分节点，按照这个例子来说，如果下

线了两个节点，系统正常工作，下线了三个节点，那剩余的两个节点也不可能继续提供服务了。

- 假设现在故障恢复，分区解除，五个节点重新可以通讯了：

- $S_1$ 和 $S_3$ 都向所有节点发送心跳包，从各自的心跳中可以得知两个主节点里 $S_3$ 的任期编号更大，它是最新的，此时五个节点均只承认 $S_3$ 是唯一的主节点。
- $S_1$ 、 $S_2$ 回滚它们所有未被提交的变更。
- $S_1$ 、 $S_2$ 从主节点发送的心跳包中获得它们失联期间发生的所有变更，将变更提交写入本地磁盘。
- 此时分布式系统各节点的状态达成最终一致。

下面我们来看第三个问题：“如何保证过程是安全的”，你是否感受到这个问题与前两点的存在一点差异？选主、数据复制都是很具体的行为，但是“安全”就很模糊，什么算是安全或者不安全？

在分布式理论中，Safety 和 Liveness 两种属性是有预定义的，在专业的书籍中一般翻译成“协定性”和“终止性”，它们的概念也是由Lamport最先提出，当时给出的定义是：

- 协定性（Safety）：所有的坏事都不会发生（something “bad” will **never** happen）
- 终止性（Liveness）：所有的好事都终将发生，但不知道是啥时候（something “good” will **must** happen, but we don't know when）

这种就算解释了你也看不明白的定义，是不是很符合Lamport老爷子一贯的写作风格？

（笔者无奈地摊摊手），我们不纠结严谨的定义，仍通过举例来说明，譬如以选主问题为例，Safety 保证了选主的结果一定是有且只有一个主节点，不可能同时出现两个主节点；而 Liveness 则要保证选主过程是一定可以在某个时刻能够结束的。由前面对活锁的介绍可以得知，在 Liveness 这个属性上选主问题是存在理论上的瑕疵的，可能会由于活锁而导致一直无法选出明确的主节点，所以Raft论文中只写了对 Safety 的保证，但由于工程实现上的处理，现实中是几乎不可能会出现终止性的问题。

最后，以上这种把共识问题分解为“Leader Election”、“Entity Replication”和“Safety”三个问题来思考、解决的解题思路，即是本节主题中的“Raft算法”，这篇以“一种可以让人理解的共识算法”（In Search of an Understandable Consensus Algorithm，Lamport：好像有人在论文标题中对我有意见？）为题的论文提出了Raft算法，获得了USENIX ATC 2014的Best Paper，后来更是成为了日后Etcd、LogCabin、Consul等重要分布式程序的实现基

础，ZooKeeper的ZAB算法与Raft的思路也非常类似，这些算法都被认为是与Multi Paxos的等价派生实现。

# Gossip协议

## Gossip

Trying to squash a rumor is like trying to unring a bell.

—— Shana Alexander↗，American Journalist

Paxos、Raft、ZAB这些分布式算法经常会被称作是“强一致性”的分布式共识协议，这样的描述扣细节概念的话是很别扭的，有语病嫌疑，但我们都明白它的意思其实是在说“尽管系统内部节点可以存在不一致的状态，内部是最终一致的，但从系统外部看来，不一致的情况并不会被观察到，所以整体上看系统是强一致性的”。与它们相对的，还有一类被冠以“最终一致性”的分布式共识协议，这表明系统中不一致的状态有可能能够在一定时间内被外部观察到。一种典型而又极为常见的最终一致的分布式系统就是DNS系统，在各节点缓存的TTL到期之前，都有可能与真实的域名翻译结果存在不一致，在本节中，我们将介绍在比特币网络和许多重要分布式框架中都有应用的另一种具有代表性的“最终一致性”的分布式共识协议：Gossip协议。

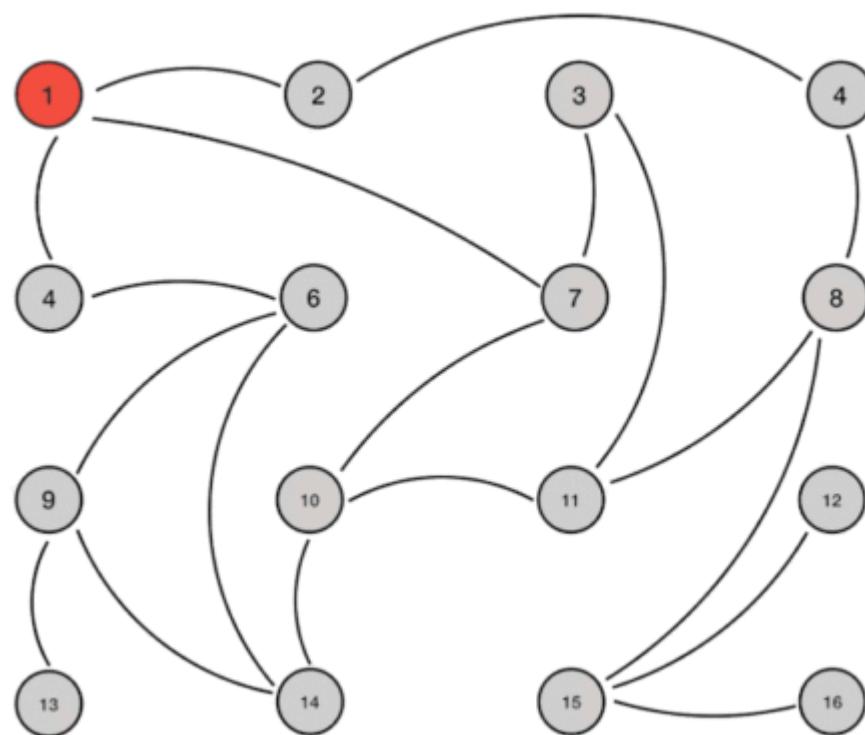
Gossip最早由施乐公司↗（Xerox，现在可能很多人不了解施乐了，或只把施乐当一家复印产品公司看待，这家公司是计算机许多关键技术的鼻祖）Palo Alto研究中心（世界首个图形界面的发明者）在论文“Epidemic Algorithms for Replicated Database Maintenance↗”中提出的一种用于分布式数据库在多节点间复制同步数据的算法。从论文题目中可以看出，最初它是被称作“流行病算法”（Epidemic Algorithm）的，而今天Gossip这个名字使用得更为普遍。除此以外，它还有“流言算法”、“八卦算法”、“瘟疫算法”等别名，这些名字都是很形象化的描述，反应了Gossip的特点：要同步的信息如同流言一般传播、病毒一般扩散。

尽管笔者按照习惯也把Gossip也称作是“共识协议”，但首先必须强调它所解决的问题并不是直接与Paxos、Raft这些共识协议等价的，只是基于Gossip之上可以通过某些方法去实现与Paxos、Raft一致的目标。一个最典型的例子是比特币网络中使用到了Gossip协议，用它来在各个分布式节点中互相同步区块头和区块体的信息，这是整个网络能够正常交换

信息的基础，但并不能称作共识；比特币使用工作量证明（Proof of Work，PoW）来对“这个区块由谁来记账”这一件事情在全网达成共识，这个目标才可以认为与Paxos、Raft的目标是一致的。

下面，我们来了解Gossip的具体算法过程。相比起Paxos、Raft，Gossip的算法过程可算是十分简单了，它可以看作是以下两个步骤的简单循环：

- 如果有某一项信息需要在整个网络中所有节点中传播，那从信息源开始，选择一个固定的传播周期（譬如1秒），随机选择它相连接的k个节点（称为Fan-Out）来传播消息。
- 每一个节点收到消息后，如果这个消息是它之前没有收到过的，将在下一个周期内，选择除了发送消息给它的那个节点外的其他相邻k个节点发送相同的消息，直到最终网络中所有节点都收到了消息，尽管这个过程需要一定时间，但是理论上最终网络的所有节点都会拥有相同的消息。



Gossip传播示意图（[图片来源](#)）

上图是Gossip传播过程的示意图，根据示意图和Gossip的过程描述，我们很容易发现Gossip对网络节点的连通性和稳定性几乎没有任何要求，它将网络某些节点只能与一部分节点部分连通（Partially Connected Network）而不是以全连通网络（Fully Connected Network）作为前提；能够容忍网络上节点的随意地增加或者减少，随意地宕机或者重启，新

增加或者重启的节点的状态最终会与其他节点同步达成一致。Gossip把网络上所有节点都视为平等的、普通的，没有任何中心化节点或者主节点的概念，这些特点使得Gossip具有很强的鲁棒性，而且极为适合在公众互联网中应用。

同时我们也很容易找到Gossip的缺点，消息最终是通过多个轮次的散播而到达全网的，因此它必然会产生全网各节点状态不一致的情况，而且由于是随机选取发送消息的节点，所以尽管可以在整体上测算出传播速率，但对于个体消息来说，无法准确地预计到需要多长时间才能达成全网一致。另外一个缺点是消息的冗余，同样是由于随机选取发送消息的节点，也就不可避免的存在消息重复发送给同一节点的情况，增加了网络的传输的压力，也给消息节点带来额外的处理负载。

Gossip传播消息时，有两种可能的传播方式：反熵（Anti-Entropy）和传谣（Rumor-Mongering），这两个名字都挺文艺的。熵（Entropy）是生活中少见但科学中很常用的概念，它代表着事物的混乱程度。反熵的意思就是反混乱，以提升网络各个节点之间的相似度为目标，所以在反熵模式下，会同步节点的全部数据，以消除各节点之间的差异，目标是整个网络各节点完全的一致。但是，在节点本身就会发生变动的前提下，这个目标将使得整个网络中消息的数量会非常庞大，给网络带来巨大的传输开销。而传谣模式是以传播消息为目标，仅仅发送新到达节点的数据，即只对外发送变更信息，这样消息数据量将显著缩减，网络开销也较小。

# 服务发现

自“子程序”诞生之日起，计算机实现了通过方法调用来组装复用指令序列，打开了软件达到更大规模一扇大门。无论是编译期链接的C/CPP，抑或是运行期链接的Java，都要通过[链接器](#)（Linker）将代码里的[符号引用](#)转换为模块入口或进程内存地址的直接引用。自“远程服务调用”诞生之日起，通过分布于网络中不同机器的互相协作来复用功能，是软件发展规模的第二次飞跃，此时如何确定目标方法的确切位置，便是与编译链接有着等同意义的问题，解决该问题的过程就被称作“[服务发现](#)”（Service Discovery）。

所有的远程服务调用都是使用“IP地址、端口号、服务标识”构成的三元组来确定一个远程服务的精确坐标的。其中“IP地址、端口号”的含义在各种远程服务中都一致，这源于TCP/IP协议在计算机网络中统治性地位，IP和端口代表了某台服务器上的某个应用程序所提供的数据交换接口。而“服务标识”则与应用层协议相关，可以是多样的，譬如HTTP的远程服务，标识是URL地址；RMI的远程服务，标识是Stub类中的方法；SOAP的远程服务，标识是WSDL中的定义，等等。远程服务的多样性导致了“服务发现”也会有两种不同的理解，一种是以UDDI为代表的“百科全书式”的服务发现，上至提供服务的企业信息（企业实体、联系地址、分类目录等等），下至服务的程序接口细节（方法名称、参数、返回值、技术规范等等）都在服务发现的管辖范围之内；另一种是类似于DNS这样“门牌号码式”的服务发现，只满足从某个代表服务提供者的符号（如域名、服务ID）到服务实际主机的翻译转换，并不关心服务具体是哪个厂家提供的，也不关心服务由几个方法，各自有什么参数所构成，默认这些细节信息是服务消费者本身所了解的，此时服务坐标就可以退化为简单的“IP+端口”。当今，后一种服务发现占主流地位，本文后续所说的服务发现，如无说明，均是特指的是后者。

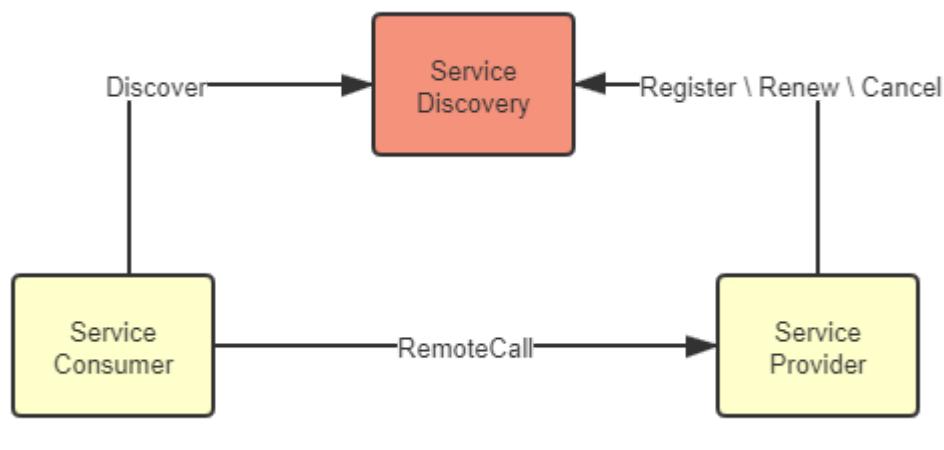
原本服务发现只依赖DNS将一个网络名称翻译为一个或者多个IP地址（或者SRV等其他记录）便可实现，后来负载均衡器也承担了一部分服务发现的职责，这些内容我们在“[透明多级分流系统](#)”一节中曾经详细解析过，这种方式在软件追求不间断长时间运行的时代是合适的。但随着微服务的逐渐流行，服务的非正常宕机重启和正常的上线下线变得更加频繁，仅靠着DNS服务器和负载均衡器等基础设施就显得有些逐渐疲于应对，无法跟上服务变动的步伐了。人们开始尝试使用ZooKeeper这样的分布式K/V框架，通过软件自身来完成服务

注册与发现，ZooKeeper曾短暂统治过远程服务发现，是微服务早期对服务发现的主流选择，但毕竟ZooKeeper是很底层的分布式工具，用户自己还需要做相当多的工作才能满足服务发现的需求。到了2014年，在Netflix内部经受过长时间实际考验的、专门用于服务发现的Eureka宣布开源，并很快被纳入Spring Cloud，成为Spring默认的远程服务发现的解决方案。从此Java程序员再无需再在服务注册这件事情上花费太多的力气。到2018年，Spring Cloud Eureka进入维护模式以后，HashiCorp的Consul和阿里巴巴的Nacos很快就从Eureka手上接过传承的衣钵。此时的服务发现框架已经发展得相当成熟，考虑到几乎方方面面的问题，譬如支持通过DNS或者HTTP请求进行符号与实际地址的转换，支持各种各样的服务健康检查方式，支持集中配置、K/V存储、跨数据中心的数据交换等多种功能，可算是应用自身去解决服务发现的一个顶峰。如今，云原生时代来临，基础设施的灵活性得到大幅度的增强，最初的使用基础设施来透明化地做服务发现的方式又重新被人们所重视，如何在基础设施和网络协议层面，对应用尽可能无感知、尽可能方便地实现服务发现是目前一个主要的发展方向。

本文中，我们将会分析服务发现的几个关键的子问题，并且探讨、对比时下最常见的用作服务发现的几种形式。首先，第一个问题是“服务发现”具体是指进行过什么操作？这里面其实包含了三个必须的过程：

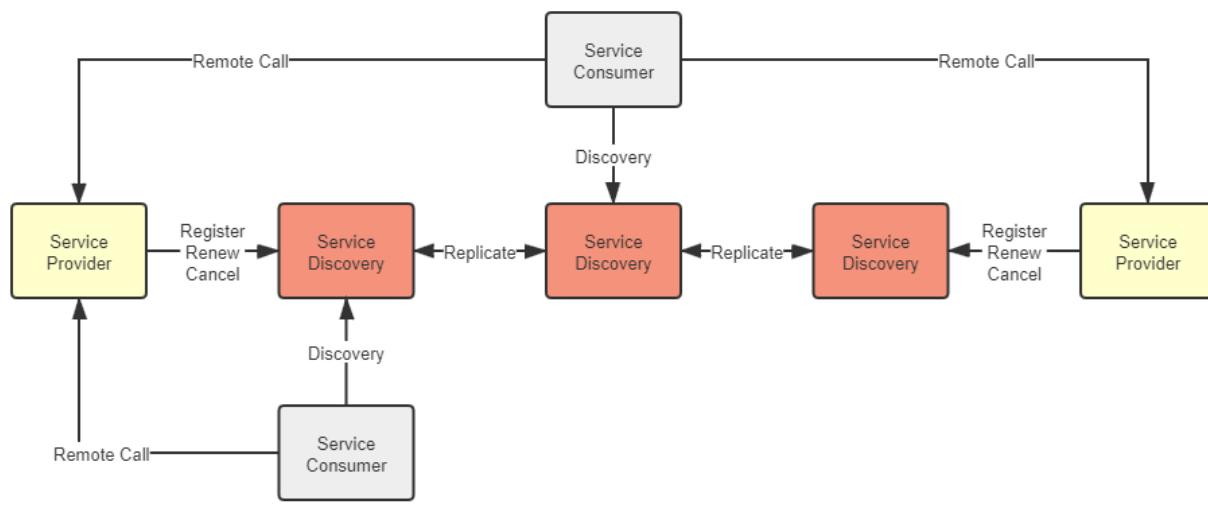
- **服务的注册**（Service Registration）：当服务启动的时候，它应该通过某些形式（譬如调用API、产生事件消息、在zk/etc的指定位置记录、存入数据库，等等）将自己的坐标信息通知到服务注册中心，这个过程可能由应用程序来完成（譬如Spring Cloud的@EnableDiscoveryClient注解），也可能有容器框架（譬如Kubernetes）来完成。
- **服务的维护**（Service Maintaining）：尽管服务发现框架通常都有提供下线机制，但并没有什么办法保证每次服务都能优雅地下线（Graceful Shutdown）而不是由于宕机、断网等原因突然失联。所以服务发现框架必须要自己去保证所维护的服务列表的正确性，以避免告知消费者服务的坐标后，得到的服务却不能使用的尴尬情况。现在的服务发现框架，往往都能支持多种协议（HTTP、TCP等）、多种方式（长连接、心跳、探针、进程状态等）去监控服务是否健康存活，将不健康的服务自动下线。
- **服务的发现**（Service Discovery）：这里的发现是狭义特指消费者从服务发现框架中，把一个符号（譬如Eureka中的ServiceID、Nacos中的服务名、或者通用的FDQN）转换为服务实际坐标的过程，这个过程现在一般是通过HTTP API请求或者通过DNS Look up操作来完成（还有一些相对少用的方式，如Kubernetes也支持注入环境变量）。

以上三点只列举了必须的过程，在此之余还会有一些可选的功能，譬如在服务发现时进行的负载均衡、流量管控、K/V存储、元数据管理、业务分组，等等，这部分后续会有专门介绍，就不再展开。我们来讨论另一个很常见的问题，说起服务发现的文章，总是无可避免地会先扯到“CP”还是“AP”的问题上。为什么服务发现对CAP如此关注、如此敏感呢？我们可以从服务发现在整个系统中所处的角色来着手分析这个问题，在概念模型中，服务中心所处的地位是如下图所示这样的：提供者在服务发现中注册、续约和下线自己的真实坐标，消费者根据某种符号从服务发现中获取到真实坐标，它们都可以视为系统中平等的微服务，如下图所示：



概念模型

但在真实的系统中，服务发现的地位还是有一些特殊，并不能为完全视其为一个普通的服。务发现是整个系统中所有其他服务都直接依赖的最基础服务（类似相同待遇的大概就数配置中心了，现在服务发现框架也开始同时提供配置中心的功能，以避免配置中心又去专门搞出一集群的节点来），几乎没有办法在业务层面进行容错处理。服务注册中心一旦崩溃，整个系统都受波及，因此必须尽最大可能在技术层面保证可用性。所以，分布式系统中，服务注册中心一般会以内部小集群的方式部署，提供三个或者五个节点（通常最多七个，一般也不会更多了，否则日志复制的开销太高）来保证高可用性，如下图所示：



真实系统

同时，也请注意到上图中各服务发现节点之间的“Replicate”字样，作为用户，我们当然期望服务注册一直可用永远健康的同时，也能够在访问每一个节点中都能取到一致的数据，这两个需求就构成了CAP矛盾。以AP、CP两种取舍作为选择维度，以最有代表性的Eureka和Consul为例，Consul采用Raft协议，要求多数派节点写入成功后服务的注册或变动才算完成，严格地保证了在集群外部读取到的服务发现结果一定是一致的；Eureka的各个节点间采用异步复制来交换服务注册信息，服务注册或变动时，并不需要等待信息在其他节点复制完成，而是马上在该服务发现节点就宣告可见（但其他节点是否可见并不保证）。这两点差异带来的影响并不在于服务注册的快慢（当然，快慢确实是有差别），而在于你如何看待以下这件事情：

假设系统形成了A、B两个网络分区后，A区的服务只能从区域内的服务发现节点获取到A区的服务坐标，B区的服务只能取到在B区的服务坐标，这对你的系统会有什么影响？

- 如果这件事情对你并没有什么影响，甚至有可能还是有益的，就应该倾向于选择AP的服务发现。譬如假设A、B就是不同的机房，是机房间的网络交换机导致服务发现集群出现的分区问题，但每个分区中的服务仍然能独立提供完整且正确的服务能力，此时尽管不是有意而为，但网络分区在事实上避免了跨机房的服务请求，反而还带来了服务调用链路优化的效果。
- 如果这件事情也可能对你影响非常大，甚至可能带来比整个系统宕机更坏的结果，就应该倾向于选择CP的服务发现。譬如系统中大量依赖了集中式缓存、消息总线、或者其他有状态的服务，一旦这些服务全部或者部分被分隔到某一个分区中，会对整个系统的操作的正确性产生直接影响的话，那与其搞出一堆数据错误，还不如停机来得痛快。

数据一致性是分布式系统永恒的话题，在服务发现这个场景里，权衡的主要关注点是一旦出现分区所带来的后果，其他在正常运行过程中的速度问题都是次要的。最后，我们再来讨论一个很“务实”的话题，现在那么多的服务发现框架，哪一款最好？或者说应该如何挑选适合的？

现在直接以服务发现、服务注册中心为目标，或者间接用来实现这个目标的方式主要有以下三类：

- 在分布式K/V存储框架上自己实现的服务发现，这类的代表是ZooKeeper、Doozerd、Etcd

这些K/V框架提供了分布式环境下读写操作的共识保证，Etcd采用的是我们学习过的Raft算法，ZooKeeper采用的是ZAB算法（一种Multi Paxos的派生算法），所以采用这种方案，就不必纠结CP还是AP的问题，它们都是CP的。这类框架的宣传语中往往会主动提及“高可用性”，潜台词其实是“在保证一致性和分区容错性的前提下，尽最大努力实现最高的可用性”，譬如Etcd的宣传语就是“高可用的集中配置和服务发现”（**Highly-Available Key Value Store for Shared Configuration and Service Discovery**）。这些K/V框架的另一个共同特点是在整体较高复杂度的架构和算法的外部，维持着极为简单的应用接口，只有基本的CRUD和Watch等少量API，所以要在上面完成功能齐全的服务发现，很多基础的能力，譬如服务如何注册、如何做健康检查，等等都必须自己实现，如今一般也只有“大厂”才会直接这些框架去做服务发现了。

- 以基础设施（主要就是DNS服务器）来实现服务发现，这类的代表是SkyDNS、KubeDNS、CoreDNS

在Kubernetes 1.3之前的版本使用SkyDNS作为默认的DNS服务，其工作原理是从kube-apiserver中监听集群服务的变化，然后根据服务生成NS、SRV等DNS记录存放到Etcd中，kubelet会在每个Pod内部设置DNS服务的地址为SkyDNS的地址，需要调用服务时，只需查询DNS把域名转换成IP列表便可实现分布式的服务发现。在Kubernetes 1.3之后，SkyDNS不再是默认的DNS服务器，由不使用Etcd而是只将DNS记录存储在内存中的KubeDNS代替，到了1.11版，就更推荐采用扩展性很强的CoreDNS，此时可以通过各种插件来决定是否要采用Etcd存储、重定向、定制DNS记录、记录日志，等等。采用这种方案，是CP还是AP就取决于后端采用何种存储，如果是基于Etcd实现的，那自然是CP的，如果是基于内存异步复制的方案实现的，那就是AP的。以基础设施来做服务发现，好处是对应用透明，任何语言、框架、工具都肯定是支持HTTP、DNS的，所以完全不受程序技术选型的约束，但坏处是透明的并不一定是简单的，你必须自己考

虑如何去做客户端负载均衡、如何调用远程方法等这些问题，而且必须遵循或者说受限于这些基础设施本身所采用的实现机制，譬如服务健康检查里，服务的缓存期限就必须采用TTL（Time to Live）来决定，这是DNS协议所规定的，如果想改用KeepAlive长连接来实时判断服务是否存活就很麻烦。

- 专门用于服务发现的框架和工具，这类的代表是Eureka、Consul和Nacos  
这一类框架中，你可以自己决定是CP还是AP的问题，譬如CP的Consul、AP的Eureka，还有同时支持CP和AP的Nacos（Nacos采用类Raft协议做的CP，采用自研的Distro协议做的AP，这里“同时是”都支持”的意思，它们必须二取其一，不是说CAP全能满足）。另外，还有很重要一点是它们对应用并不是透明的，尽管Consul、Nacos也支持基于DNS的服务发现，尽管这些框架都基本上做到了以声明代替编码（譬如在Spring Cloud中只改动pom.xml、配置文件和注解即可实现），但它们依然是应用程序有感知的。所以或多或少还需要考虑你所用的程序语言、技术框架的集成问题。但这一点其实并不见得就是坏处，譬如采用Eureka做服务注册，那在远程调用服务时你就可以用OpenFeign做客户端，写个声明式接口就能跑，相当能偷懒；在做负载均衡时你就可以采用Ribbon做客户端，要换均衡算法改个配置就成，这些“不透明”实际上都为编码开发带来了一定便捷，而前提是你选用的语言和框架支持。如果老板提出要在Rust上用Eureka，你就只能无奈叹息了（原本这里我写的是Node、Go、Python等，查了一下这些居然都有非官方的Eureka客户端，用的人多就是有好处啊）。

# 路由与网关

# 进程内负载均衡

# 服务编排

# 中心化配置

# 隔离

# 熔断

# 超时

# 流控

# 降级

# 异常注入

# 流量加密

# 访问策略

# 事件审计

# 日志聚合

# 链路跟踪

# 应用性能管理

# 虚拟化的概念与历史

# 虚拟化容器

# CRI接口

# 资源隔离

## 内核

### Cgroups

### 用户空间运行时

# 资源对象

# 容器管理

# 容器间网络

# CNI接口

# 网络策略

# 网络插件

# 容器负载均衡

# 配置与数据持久化

# CSI接口

# 分布式文件系统

# 共享存储插件

# GPU虚拟化

# Device Plugin机制

# 调度GPU

# Nvidia插件

# 扩展基础设施

# CRD定义

# 自定义API Server

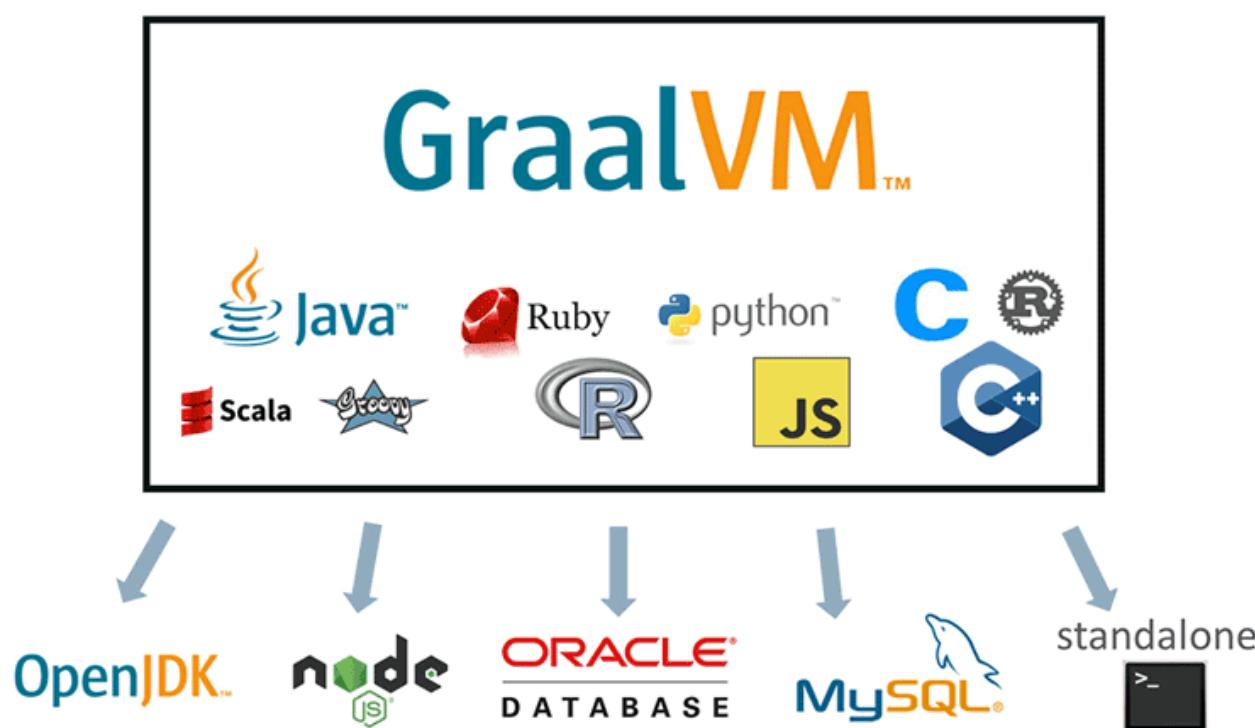
# 硬件资源调度

# Graal VM

网上每隔一段时间就能见到几条“未来X语言将会取代Java”的新闻，此处“X”可以用Kotlin、Golang、Dart、JavaScript、Python……等各种编程语言来代入。这大概就是长期占据[编程语言榜单](#)第一位的烦恼，天下第一总避免不了挑战者相伴。

如果Java有拟人化的思维，它应该从来没有惧怕过被哪一门语言所取代，Java“天下第一”的底气不在于语法多么先进好用，而是来自它庞大的用户群和极其成熟的软件生态，这在朝夕之间难以撼动。不过，既然有那么多新、旧编程语言的兴起躁动，说明必然有其需求动力所在，譬如互联网之于JavaScript、人工智能之于Python，微服务风潮之于Golang等等。大家都清楚不太可能有哪门语言能在每一个领域都尽占优势，Java已是距离这个目标最接近的选项，但若“天下第一”还要百尺竿头更进一步的话，似乎就只能忘掉Java语言本身，踏入无招胜有招的境界。

2018年4月，Oracle Labs新公开了一项黑科技：[Graal VM](#)，从它的口号“Run Programs Faster Anywhere”就能感觉到一颗蓬勃的野心，这句话显然是与1995年Java刚诞生时的“Write Once，Run Anywhere”在遥相呼应。



## Graal VM

Graal VM被官方称为“Universal VM”和“Polyglot VM”，这是一个在HotSpot虚拟机基础上增强而成的跨语言全栈虚拟机，可以作为“任何语言”的运行平台使用，这里“任何语言”包括了Java、Scala、Groovy、Kotlin等基于Java虚拟机之上的语言，还包括了C、C++、Rust等基于LLVM的语言，同时支持其他像JavaScript、Ruby、Python和R语言等等。Graal VM可以无额外开销地混合使用这些编程语言，支持不同语言中混用对方的接口和对象，也能够支持这些语言使用已经编写好的本地库文件。

Graal VM的基本工作原理是将这些语言的源代码（例如JavaScript）或源代码编译后的中间格式（例如LLVM字节码）通过解释器转换为能被Graal VM接受的中间表示<sup>[1]</sup>（Intermediate Representation，IR），譬如设计一个解释器专门对LLVM输出的字节码进行转换来支持C和C++语言，这个过程称为“程序特化<sup>[2]</sup>”（Specialized，也常称为Partial Evaluation）。Graal VM提供了Truffle工具集<sup>[3]</sup>来快速构建面向一种新语言的解释器，并用它构建了一个称为Sulong<sup>[4]</sup>的高性能LLVM字节码解释器。

以更严格的角度来看，Graal VM才是真正意义上与物理计算机相对应的高级语言虚拟机，理由是它与物理硬件的指令集一样，做到了只与机器特性相关而不与某种高级语言特性相关。Oracle Labs的研究总监Thomas Wuerthinger在接受InfoQ采访<sup>[5]</sup>时谈到：“随着Graal VM 1.0的发布，我们已经证明了拥有高性能的多语言虚拟机是可能的，并且实现这个目标的最佳方式不是通过类似Java虚拟机和微软CLR那样带有语言特性的字节码”。对于一些本来就不以速度见长的语言运行环境，由于Graal VM本身能够对输入的中间表示进行自动优化，在运行时还能进行即时编译优化，往往使用Graal VM实现能够获得比原生编译器更优秀的执行效率，譬如Graal.js要优于Node.js、Graal.Python要优于CPython，TruffleRuby要优于Ruby MRI，FastR要优于R语言等等。

针对Java而言，Graal VM本来就是在HotSpot基础上诞生的，天生就可作为一套完整的符合Java SE 8标准Java虚拟机来使用。它和标准的HotSpot差异主要在即时编译器上，其执行效率、编译质量目前与标准版的HotSpot相比也是互有胜负。但现在Oracle Labs和美国大学里面的研究院所做的最新即时编译技术的研究全部都迁移至基于Graal VM之上进行了，其发展潜力令人期待。如果Java语言或者HotSpot虚拟机真的有被取代的一天，那从现在看来Graal VM是希望最大的一个候选项，这场革命很可能会在Java使用者没有明显感觉的情况下悄然而来，Java世界所有的软件生态都没有发生丝毫变化，但天下第一的位置已经悄然更迭。



# 新一代即时编译器

对需要长时间运行的应用来说，由于经过充分预热，热点代码会被HotSpot的探测机制准确定位捕获，并将其编译为物理硬件可直接执行的机器码，在这类应用中Java的运行效率很大程度上是取决于即时编译器所输出的代码质量。

HotSpot虚拟机中包含有两个即时编译器，分别是编译时间较短但输出代码优化程度较低的客户端编译器（简称为C1）以及编译耗时长但输出代码优化质量也更高的服务端编译器（简称为C2），通常它们会在分层编译机制下与解释器互相配合来共同构成HotSpot虚拟机的执行子系统的。

自JDK 10起，HotSpot中又加入了一个全新的即时编译器：Graal编译器，看名字就可以联想到它是来自于前一节提到的Graal VM。Graal编译器是作为C2编译器替代者的角色登场的。C2的历史已经非常长了，可以追溯到Cliff Click大神读博士期间的作品，这个由C++写成的编译器尽管目前依然效果拔群，但已经复杂到连Cliff Click本人都不愿意继续维护的程度。而Graal编译器本身就是由Java语言写成，实现时又刻意与C2采用了同一种名为“Sea-of-Nodes”的高级中间表示（High IR）形式，使其能够更容易借鉴C2的优点。Graal编译器比C2编译器晚了足足二十年面世，有着极其充沛的后发优势，在保持能输出相近质量的编译代码的同时，开发效率和扩展性上都要显著优于C2编译器，这决定了C2编译器中优秀的代码优化技术可以轻易地移植到Graal编译器上，但是反过来Graal编译器中行之有效的优化在C2编译器里实现起来则异常艰难。这种情况下，Graal的编译效果短短几年间迅速追平了C2，甚至某些测试项中开始逐渐反超C2编译器。Graal能够做比C2更加复杂的优化，如“[部分逃逸分析](#)”（Partial Escape Analysis），也拥有比C2更容易使用“[激进预测性优化](#)”（Aggressive Speculative Optimization）的策略，支持自定义的预测性假设等等。

今天的Graal编译器尚且年幼，还未经过足够多的实践验证，所以仍然带着“实验状态”的标签，需要用开关参数去激活，这让笔者不禁联想起JDK 1.3时代，HotSpot虚拟机刚刚横空出世时的场景，同样也是需要用开关激活，也是作为Classic虚拟机的替代品的一段历史。

Graal编译器未来的前途可期，作为Java虚拟机执行代码的最新引擎，它的持续改进，会同时为HotSpot与Graal VM注入更快更强的驱动力。



# 向原生迈进

对不需要长时间运行的，或者小型化的应用而言，Java（而不是指Java ME）天生就带有一些劣势，这里并不光是指跑个HelloWorld也需要百兆的JRE之类的问题，而更重要的是指近几年从大型单体应用架构向小型微服务应用架构发展的技术潮流下，Java表现出来的不适应。

在微服务架构的视角下，应用拆分后，单个微服务很可能就不再需要再面对数十、数百GB乃至TB的内存，有了高可用的服务集群，也无须追求单个服务要7×24小时不可间断地运行，它们随时可以中断和更新；但相应地，Java的启动时间相对较长、需要预热才能达到最高性能等特点就显得相悖于这样的应用场景。在无服务架构中，矛盾则可能会更加突出，比起服务，一个函数的规模通常会更小，执行时间会更短，当前最热门的无服务运行环境AWS Lambda所允许的最长运行时间仅有15分钟。

一直把软件服务作为重点领域的Java自然不可能对此视而不见，在最新的几个JDK版本的功能清单中，已经陆续推出了跨进程的、可以面向用户程序的类型信息共享（Application Class Data Sharing，AppCDS，允许把加载解析后的类型信息缓存起来，从而提升下次启动速度，原本CDS只支持Java标准库，在JDK 10时的AppCDS开始支持用户的程序代码）、无操作的垃圾收集器（Epsilon，只做内存分配而不做回收的收集器，对于运行完就退出的应用十分合适）等改善措施。而酝酿中的一个更彻底的解决方案，是逐步开始对提前编译（Ahead of Time Compilation，AOT）提供支持。

提前编译是相对于即时编译的概念，提前编译能带来的最大好处是Java虚拟机加载这些已经预编译成二进制库之后就能够直接调用，而无须再等待即时编译器在运行时将其编译成二进制机器码。理论上，提前编译可以减少即时编译带来的预热时间，减少Java应用长期给人带来的“第一次运行慢”不良体验，可以放心地进行很多全程序的分析行为，可以使用时间压力更大的优化措施。

但是提前编译的坏处也很明显，它破坏了Java“一次编写，到处运行”的承诺，必须为每个不同的硬件、操作系统去编译对应的发行包。也显著降低了Java链接过程的动态性，必须

要求加载的代码在编译期就是全部已知的，而不能再是运行期才确定，否则就只能舍弃掉已经提前编译好的版本，退回到原来的即时编译执行状态。

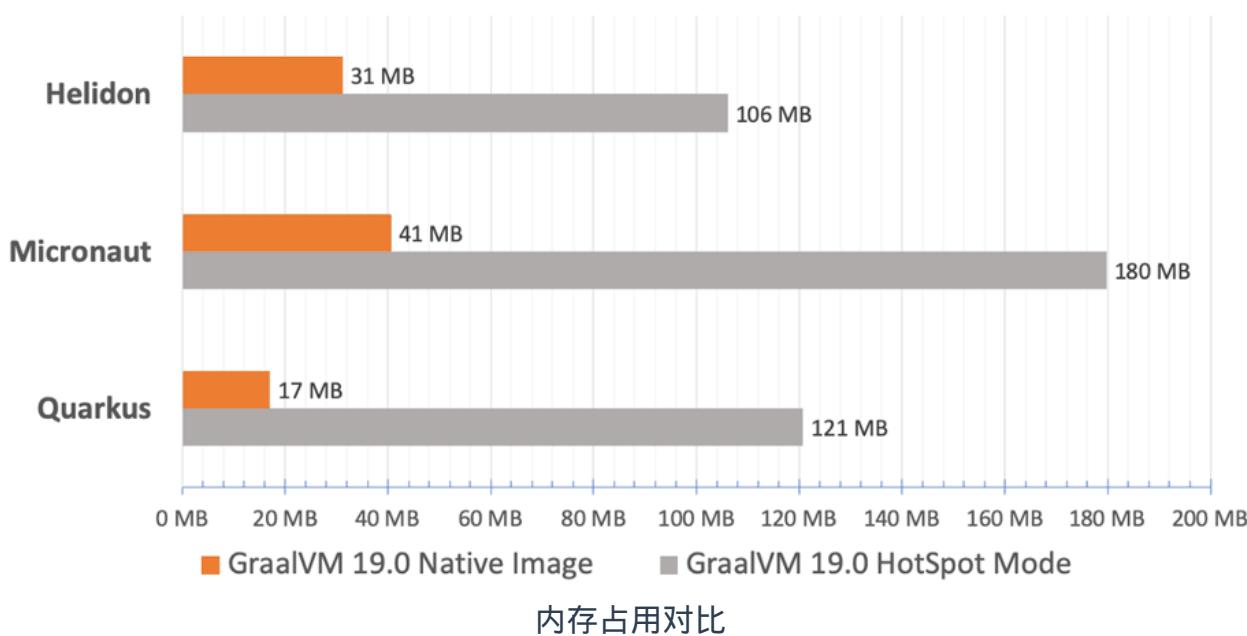
早在JDK 9时期，Java 就提供了实验性的Jaotc命令来进行提前编译，不过多数人试用过后都颇感失望，大家原本期望的是类似于Excelsior JET那样的编译过后能生成本地代码完全脱离Java虚拟机运行的解决方案，但Jaotc其实仅仅是代替掉即时编译的一部分作用而已，仍需要运行于HotSpot之上。

直到Substrate VM<sup>1</sup>出现，才算是满足了人们心中对Java提前编译的全部期待。Substrate VM是在Graal VM 0.20版本里新出现的一个极小型的运行时环境，包括了独立的异常处理、同步调度、线程管理、内存管理（垃圾收集）和JNI访问等组件，目标是代替HotSpot用来支持提前编译后的程序执行。它还包含了一个本地镜像的构造器（Native Image Generator）用于为用户程序建立基于Substrate VM的本地运行时镜像。这个构造器采用指针分析（Points-To Analysis）技术，从用户提供的程序入口出发，搜索所有可达的代码。在搜索的同时，它还将执行初始化代码，并在最终生成可执行文件时，将已初始化的堆保存至一个堆快照之中。这样一来，Substrate VM就可以直接从目标程序开始运行，而无须重复进行Java虚拟机的初始化过程。但相应地，原理上也决定了Substrate VM必须要求目标程序是完全封闭的，即不能动态加载其他编译期不可知的代码和类库。基于这个假设，Substrate VM才能探索整个编译空间，并通过静态分析推算出所有虚方法调用的目标方法。

Substrate VM带来的好处是能显著降低了内存占用及启动时间，由于HotSpot本身就会有一定的内存消耗（通常约几十MB），这对最低也从几GB内存起步的大型单体应用来说并不算什么，但在微服务下就是一笔不可忽视的成本。根据Oracle官方给出的测试数据<sup>2</sup>，运行在Substrate VM上的小规模应用，其内存占用和启动时间与运行在HotSpot相比有了5倍到50倍的下降，具体结果如下图所示：

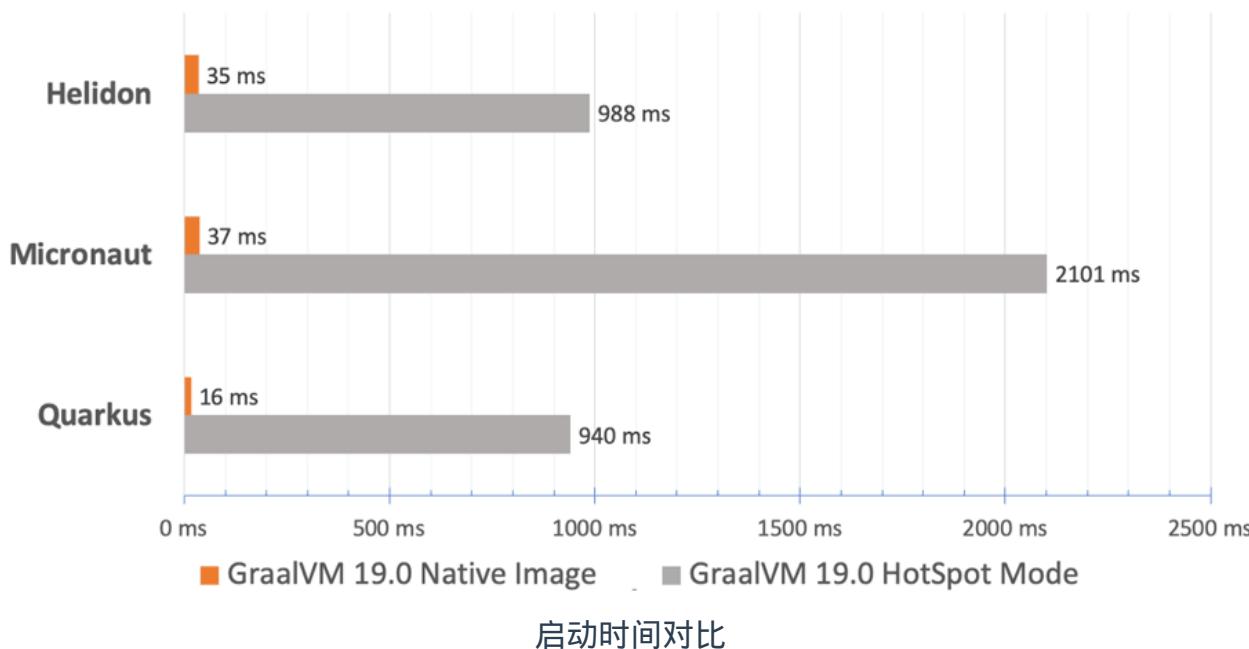
## Java Microservice: Memory Footprint

~5x lower



## Java Microservice: Startup Time

~50x faster



Substrate VM补全了Graal VM“Run Programs Faster Anywhere”愿景蓝图里最后的一块拼图，让Graal VM支持其他语言时不会有重量级的运行负担。譬如运行JavaScript代码，Node.js的V8引擎执行效率非常高，但即使是最简单的HelloWorld，它也要使用约20MB的内存，而运行在Substrate VM上的Graal.js，跑一个HelloWorld则只需要4.2MB内存而已，且运行速度与V8持平。Substrate VM 的轻量特性，使得它十分适合于嵌入至其他系统之中，譬如Oracle自家的数据库就已经开始使用这种方式支持用不同的语言代替PL/SQL来编写存储过程。



# 没有虚拟机的Java

尽管Java已经看清楚了在微服务时代的前进目标，但是，Java语言和生态在微服务、微应用环境中的天生的劣势并不会一蹴而就地被解决，通往这个目标的道路注定会充满荆棘；尽管已经有了放弃“一次编写，到处运行”、放弃语言动态性的思想准备，但是，这些特性并不单纯是宣传口号，它们在Java语言诞生之初就被植入到基因之中，当Graal VM试图打破这些规则的同时，也受到了Java语言和在其之上的生态生态的强烈反噬，笔者选择其中最主要的一些困列举如下：

- 某些Java语言的特性，使得Graal VM编译本地镜像的过程变得极为艰难。譬如常见的反射，除非使用[安全管理器](#)去专门进行认证许可，否则反射机制具有在运行期动态调用几乎所有API接口的能力，且具体会调用哪些接口，在程序不会真正运行起来的编译期是无法获知的。反射显然是Java不能放弃不能妥协的重要特性，为此，只能由程序的开发者明确地告知Graal VM有哪些代码可能被反射调用（通过JSON配置文件的形式），Graal VM才能在编译本地程序时将它们囊括进来。

```
[
 {
 name: "com.github.fenixsoft.SomeClass",
 allDeclaredConstructors: true,
 allPublicMethods: true
 },
 {
 name: "com.github.fenixsoft.AnotherClass",
 fields: [{name: "foo"}, {name: "bar"}],
 methods: [
 {
 name: "<init>",
 parameterTypes: ["char[]"]
 }]
 },
 // something else
]
```

这是一种可操作性极其低下却又无可奈何的解决方案，即使开发者接受不厌其烦地列举出自己代码中所用到的反射API，但他们又如何能保证程序所引用的其他类库的反射行为都已全部被获知，其中没有任何遗漏？与此类似的还有另外一些语言特性，如动态代理等。另外，一切非代码性质的资源，如最典型的配置文件等，也都必须明确加入配置中才能被Graal VM编译打包。这导致了如果没有专门的工具去协助，使用Graal VM编译Java的遗留系统即使理论可行，实际操作也将是极度的繁琐。

- 大多数运行期对字节码的生成和修改操作，在Graal VM看来都是无法接受的，因为Substrate VM里面不再包含即时编译器和字节码执行引擎，所以一切可能被运行的字节码，都必须经过AOT编译成为原生代码。请不要觉得运行期直接生成字节码会很罕见，误以为导致的影响应该不算很大。事实上，多数实际用于生产的Java系统都或直接或讲解、或多或少引用了ASM、CGLIB、Javassist这类字节码库。举个例子，CGLIB是通过运行时产生字节码（生成代理类的子类）来做动态代理的，长期以来这都是Java世界里进行类增强的主流形式，因为面向接口的增强可以使用JDK自带的动态代理，但对类的增强则并没有多少选择的余地。CGLIB也是Spring用来做类增强的选择，但Graal VM明确表示是不可能支持CGLIB的，因此，这点就必须由用户（面向接口编程）、框架（Spring这些DI框架放弃CGLIB增强）和Graal VM（起码得支持JDK的动态代理，留条活路可走）来共同解决。自Spring Framework 5.2起，@Configuration注解中加入了一个新的proxyBeanMethods参数，设置为false则可避免Spring对与非接口类型的Bean进行代理。同样地，对应在Spring Boot 2.2中，@SpringBootApplication注解也增加了proxyBeanMethods参数，通常采用Graal VM去构建的Spring Boot本地应用都需要设置该参数。
- 一切HotSpot虚拟机本身的内部接口，譬如JVMTI、JVMCI等，在都将不复存在了——在本地镜像中，连HotSpot本身都被消灭了，这些接口自然成了无根之木。这对使用者一侧的最大影响是再也无法进行Java语言层次的远程调试了，最多只能进行汇编层次的调试。在生产系统中一般也没有人这样做，开发环境就没必要采用Graal VM编译，这点的实际影响并不算大。
- Graal VM放弃了一部分可以妥协的语言和平台层面的特性，譬如Finalizer、安全管理器、InvokeDynamic指令和MethodHandles，等等，在Graal VM中都被声明为不支持的，这些妥协的内容大多倒并非全然无法解决，主要是基于工作量性价比的原因。能够被放弃的语言特性，说明确实是影响范围非常小的，所以这个对使用者来说一般是可以接受的。

- .....

以上，是Graal VM在Java语言中面临的部分困难，在整个Java的生态系统中，数量庞大的第三方库才是真正最棘手的难题。可以预料，这些第三方库一旦脱离了Java虚拟机，在原生环境中肯定会暴露出无数千奇百怪的异常行为。Graal VM团队对此的态度非常务实，并没有直接硬啃。要建设可持续、可维护的Graal VM，就不能为了兼容现有JVM生态，做出过多的会影响性能、优化空间和未来拓展的妥协牺牲，为此，应该也只能反过来由Java生态去适应Graal VM，这是Graal VM团队明确传递出对第三方库的态度：

### 3rd party libraries

Graal VM native support needs to be sustainable and maintainable, that's why we do not want to maintain fragile patches for the whole JVM ecosystem.

The ecosystem of libraries needs to support it natively.

—— Sébastien Deleuze , DEVOXX 2019 ↗

为了推进Java生态向Graal VM兼容，Graal VM主动拉拢了Java生态中最庞大的一个派系：Spring。从2018年起，来自Oracle的Graal VM团队与来自Pivotal的Spring团队已经紧密合作了很长的一段时间，共同创建了[Spring Graal Native](#)项目来解决Spring全家桶在Graal VM上的运行适配问题，在不久的将来（预计应该是2020年10月左右），下一个大的Spring版本（Spring Framework 5.3、Spring Boot 2.3）的其中一项主要改进就是能够开箱即用地支持Graal VM，这样，用于微服务环境的Spring Cloud便会获得不受Java虚拟机束缚的更广阔舞台空间。

# Spring over Graal

前面几部分，我们以定性的角度分析了Graal VM诞生的背景与它的价值，在最后这部分，我们尝试进行一些实践和定量的讨论，介绍具体如何使用Graal VM之余，也希望能以更加量化的角度去理解程序运行在Graal VM之上，会有哪些具体的收益和代价。

尽管需要到2020年10月正式发布之后，Spring对Graal VM的支持才会正式提供，但现在的我们其实已经可以使用Graal VM来（实验性地）运行Spring、Spring Boot、Spring Data、Netty、JPA等等的一系列组件（不过SpringCloud中的组件暂时还不行）。接下来，我们将尝试使用Graal VM来编译一个标准的Spring Boot应用：

- **环境准备：**

- 安装Graal VM，你可以选择直接[下载](#)安装（版本选择Graal VM CE 20.0.0），然后配置好PATH和JAVA\_HOME环境变量即可；也可以选择使用[SDKMAN](#)来快速切换环境。个人推荐后者，毕竟目前还不适合长期基于Graal VM环境下工作，经常手工切换会很麻烦。

```
安装SDKMAN
$ curl -s "https://get.sdkman.io" | bash

安装Graal VM
$ sdk install java 20.0.0.r8-grl
```

- 安装本地镜像编译依赖的LLVM工具链。

```
gu命令来源于Graal VM的bin目录
$ gu install native-image
```

请注意，这里已经假设你机器上已有基础的GCC编译环境，即已安装过build-essential、libz-dev等套件。没有的话请先行安装。对于Windows环境来说，这步是需要Windows SDK 7.1中的C++编译环境来支持。我个人并不建议在Windows上进行Java应

用的本地化操作，如果说在Linux中编译一个本地镜像，通常是为了打包到Docker，然后发布到服务器中使用。那在Windows上编译一个本地镜像，你打算用它来干什么呢？

- **编译准备：**

- 首先，我们先假设你准备编译的代码是“符合要求”的，即没有使用到Graal VM不支持的特性，譬如前面提到的Finalizer、CGLIB、InvokeDynamic这类功能。然后，由于我们用的是Graal VM的Java 8版本，也必须假设你编译使用Java语言级别在Java 8以内。
- 然后，我们需要用到尚未正式对外发布的Spring Boot 2.3，目前最新的版本是Spring Boot 2.3.0.M4。请将你的pom.xml中的Spring Boot版本修改如下（假设你编译用的是Maven，用Gradle的请自行调整）：

```
<parent>
 <groupId>org.springframework.boot</groupId>
 <artifactId>spring-boot-starter-parent</artifactId>
 <version>2.3.0.M4</version>
 <relativePath/>
</parent>
```

由于是未发布的Spring Boot版本，所以它在Maven的中央仓库中是找不到的，需要手动加入Spring的私有仓库，如下所示：

```
<repositories>
 <repository>
 <id>spring-milestone</id>
 <name>Spring milestone</name>
 <url>https://repo.spring.io/milestone</url>
 </repository>
</repositories>
```

- 最后，尽管我们可以通过命令行（使用native-image命令）来直接进行编译，这对于没有什么依赖的普通Jar包、写一个Helloworld来说都是可行的，但对于Spring Boot，光是在命令行中写Classpath上都忙活一阵的，建议还是使用[Maven插件](#)来驱动Graal VM编译，这个插件能够根据Maven的依赖信息自动组织好Classpath，你只需

要填其他命令行参数就行了。因为并不是每次编译都需要构建一次本地镜像，为了不干扰使用普通Java虚拟机的编译，建议在Maven中独立建一个Profile来调用Graal VM插件，具体如下所示：

```
<profiles>
 <profile>
 <id>graal</id>
 <build>
 <plugins>
 <plugin>
 <groupId>org.graalvm.nativeimage</groupId>
 <artifactId>native-image-maven-plugin</artifactId>
 <version>20.0.0</version>
 <configuration>
 <buildArgs>-Dspring.graal.remove-unused-autoconfig=true
--no-fallback -H:+ReportExceptionStackTraces --no-
server</buildArgs>
 </configuration>
 <executions>
 <execution>
 <goals>
 <goal>native-image</goal>
 </goals>
 <phase>package</phase>
 </execution>
 </executions>
 </plugin>
 <plugin>
 <groupId>org.springframework.boot</groupId>
 <artifactId>spring-boot-maven-plugin</artifactId>
 </plugin>
 </plugins>
 </build>
 </profile>
</profiles>
```

这个插件同样在Maven中央仓库中不存在，所以也得加上前面Spring的私有库：

```
<pluginRepositories>
 <pluginRepository>
 <id>spring-milestone</id>
```

```
<name>Spring milestone</name>
<url>https://repo.spring.io/milestone</url>
</pluginRepository>
</pluginRepositories>
```

至此，编译环境的准备顺利完成。

- 程序调整：

- 首先，前面提到了Graal VM不支持CGLIB，只能使用JDK动态代理，所以应当把Spring对普通类的Bean增强给关闭掉：

```
@SpringBootApplication(proxyBeanMethods = false)
public class ExampleApplication {

 public static void main(String[] args) {
 SpringApplication.run(ExampleApplication.class, args);
 }

}
```

- 然后，这是最麻烦的一个步骤，你程序里反射调用过哪些API、用到哪些资源、动态代理，还有哪些类型需要在编译期初始化的，都必须使用JSON配置文件逐一告知Graal VM。前面也说过了，这事情只有理论上的可行性，实际做起来完全不可操作。Graal VM的开发团队当然也清楚这一点，所以这个步骤实际的处理途径有两种，第一种是假设你依赖的第三方包，全部都在Jar包中内置了以上编译所需的配置信息，这样你只要提供你程序里用户代码中用到的配置即可，如果你程序里没写过反射、没用过动态代理什么的，那就什么配置都无需提供。第二种途径是Graal VM计划提供一个Native Image Agent的代理，只要将它挂载在在程序中，以普通Java虚拟机运行一遍，把所有可能的代码路径都操作覆盖到，这个Agent就能自动帮你根据程序实际运行情况来生成编译所需要的配置，这样无论是你自己的代码还是第三方的代码，都不需要做预先的配置。目前，第二种方式中的Agent尚未正式发布，只有方式一是可用的。幸好，Spring与Graal VM共同维护的在[Spring Graal Native](#)项目已经提供了大多数Spring Boot组件的配置信息（以及一些需要在代码层面处理的Patch），我们只需要简单依赖该工程即可。

```
<dependencies>
 <dependency>
 <groupId>org.springframework.experimental</groupId>
 <artifactId>spring-graal-native</artifactId>
 <version>0.6.1.RELEASE</version>
 </dependency>
 <dependency>
 <groupId>org.springframework</groupId>
 <artifactId>spring-context-indexer</artifactId>
 </dependency>
</dependencies>
```

另外还有一个小问题，由于目前Spring Boot嵌入的Tomcat中，WebSocket部分在JMX反射上还有一些瑕疵，在[修正该问题的PR](#)被Merge之前，暂时需要手工去除掉这个依赖：

```
<dependencies>
 <dependency>
 <groupId>org.springframework.boot</groupId>
 <artifactId>spring-boot-starter-web</artifactId>
 <exclusions>
 <exclusion>
 <groupId>org.apache.tomcat.embed</groupId>
 <artifactId>tomcat-embed-websocket</artifactId>
 </exclusion>
 </exclusions>
 </dependency>
</dependencies>
```

- 最后，在Maven中给出程序的启动类的路径：

```
<properties>
 <start-class>com.example.ExampleApplication</start-class>
</properties>
```

- **开始编译：**

- 到此一切准备就绪，通过Maven进行编译：

```
$ mvn -Pgraal clean package
```

sh

编译的结果默认输出在target目录，以启动类的名字命名。

- 因为AOT编译可以放心大胆地进行大量全程序的重负载优化，所以无论是编译时间还是空间占用都非常可观。笔者在intel 9900K、64GB内存的机器上，编译了一个只引用了org.springframework.boot:spring-boot-starter-web的Helloworld类型的工程，大约耗费了两分钟时间。

```
[com.example.exampleapplication:9839] (typeflow): 22,093.72 ms,
6.48 GB
[com.example.exampleapplication:9839] (objects): 34,528.09 ms,
6.48 GB
[com.example.exampleapplication:9839] (features): 6,488.74 ms,
6.48 GB
[com.example.exampleapplication:9839] analysis: 65,465.65 ms,
6.48 GB
[com.example.exampleapplication:9839] (clinit): 2,135.25 ms,
6.48 GB
[com.example.exampleapplication:9839] universe: 4,449.61 ms,
6.48 GB
[com.example.exampleapplication:9839] (parse): 2,161.78 ms,
6.32 GB
[com.example.exampleapplication:9839] (inline): 3,113.77 ms,
6.25 GB
[com.example.exampleapplication:9839] (compile): 15,892.88 ms,
6.56 GB
[com.example.exampleapplication:9839] compile: 25,044.34 ms,
6.56 GB
[com.example.exampleapplication:9839] image: 6,580.71 ms,
6.63 GB
[com.example.exampleapplication:9839] write: 1,362.73 ms,
6.63 GB
[com.example.exampleapplication:9839] [total]: 120,410.26 ms,
6.63 GB
[INFO]
[INFO] --- spring-boot-maven-plugin:2.3.0.M4:repackage (repackage)
@ exampleapplication ---
[INFO] Replacing main artifact with repackaged archive
[INFO] -----

```

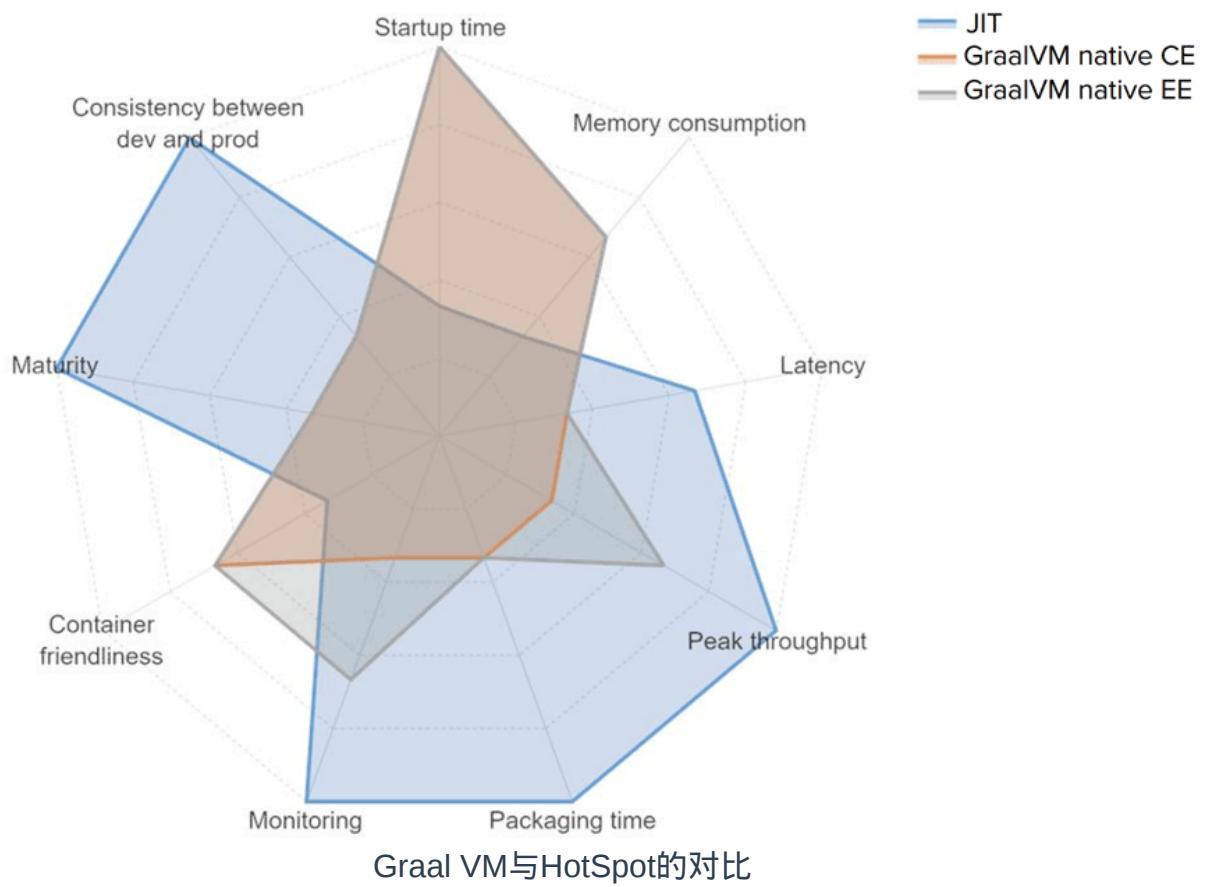
```
[INFO] BUILD SUCCESS
[INFO] -----

[INFO] Total time: 02:08 min
[INFO] Finished at: 2020-04-25T22:18:14+08:00
[INFO] Final Memory: 38M/599M
[INFO] -----

```

- 效果评估：

- 笔者使用Graal VM编译一个最简单的HelloWorld程序（就只在控制台输出个HelloWorld，什么都不依赖），最终输出的结果大约3.6MB，启动时间能低至2ms左右。如果用这个程序去生成Docker镜像（不基于任何基础镜像，即使用FROM scratch打包），产生的镜像还不到3.8MB。而OpenJDK官方提供的Docker镜像，即使是slim版，其大小也在200MB到300MB之间。
- 使用Graal VM编译一个简单的Spring Boot Web应用，仅导入Spring Boot的Web Starter的依赖的话，编译结果有77MB，原始的Fat Jar包大约是16MB，这样打包出来的Docker镜像可以不依赖任何基础镜像，大小仍然是78MB左右（实际使用时最好至少也要基于alpine吧，不差那几MB）。相比起空间上的收益，启动时间上的改进是更主要的，Graal VM的本地镜像启动时间比起基于虚拟机的启动时间有着绝对的优势，一个普通Spring Boot的Web应用启动一般2、3秒之间，而本地镜像只要100毫秒左右即可完成启动，这确实有了数量级的差距。
- 不过，必须客观地说明一点，尽管Graal VM在启动时间、空间占用、内存消耗等容器化环境中比较看重的方面确实比HotSpot有明显的改进，尽管Graal VM可以放心大胆地使用重负载的优化手段，但如果是处于长时间运行这个前提下，至少到目前为止，没有任何迹象表明它能够超越经过充分预热后的HotSpot。在延迟、吞吐量、可监控性等方面，仍然是HotSpot占据较大优势，下图引用了DEVOXX 2019中Graal VM团队自己给出的Graal VM与HotSpot JIT在各个方面的对比评估：



Graal VM团队同时也说了，Graal VM有望在2020年之内，在延迟和吞吐量这些关键指标上追评HotSpot现在的表现。Graal VM毕竟是一个2018年才正式公布的新生事物，我们能看到它这两三年间在可用性、易用性和性能上持续地改进，Graal VM有望成为Java在微服务时代里的最重要的基础设施变革者，这项改进的结果如何，甚至可能与Java的前途命运息息相关。

# 响应式编程

# 函数式接口与流式编程思想

# 并行、异步、非阻塞

# 构建发布脚本

# 持续集成

# 灰度发布

# 部署环境

这一部分介绍了Docker容器环境和Kubernetes集群的Step-By-Step的安装流程，严格来讲，环境依赖这部分内容与主题并无直接关系，对这部分已有了解的读者，完全可以略过。

# 部署Docker CE容器环境

本文为Linux系统安装Docker容器环境的简要说明，主要包括：

1. 安装稳定最新发行版（Stable Release）的命令及含义。
2. 针对国内网络环境的必要镜像加速或者代理设置工作。

若需了解Docker安装其他方面的内容，如安装Nightly/Test版本、Backporting、软件版权和支持等信息，可参考官方的部署指南：<https://docs.docker.com/install/>

文中涉及到的Debian系和Redhat系的包管理工具，主要包括：

- Debian系：Debian、Ubuntu、Deepin、Mint
- Redhat系：RHEL、Fedora、CentOS

如用的其他Linux发行版，如Gentoo、Archlinux、OpenSUSE等，建议自行安装二进制包。

## 移除旧版本Docker

如果以前已经安装过旧版本的Docker（可能会被称为docker，docker.io 或 docker-engine），需先行卸载。

Debian系：

```
$ sudo apt-get remove docker docker-engine docker.io containerd runc
docker-ce docker-ce-cli containerd.io
```

sh

RedHat系：

```
$ sudo yum remove docker \
docker-client \
docker-client-latest \
dockerd
```

sh

```
docker-common \
docker-latest \
docker-latest-logrotate \
docker-logrotate \
docker-engine
```

## 安装Docker依赖工具链及软件源

在Debian上主要是为了apt能够正确使用HTTPS协议，并将Docker官方的GPG Key ( GNU Privacy Guard，包的签名机制 ) 和软件源地址注册到软件源中。

在RHEL上是为了devicemapper获得yum-config-manager、device-mapper-persistent-data、lvm2的支持。

Debian系：

```
$ sudo apt-get install apt-transport-https \
ca-certificates \
curl \
software-properties-common

注册Docker官方GPG公钥
$ sudo curl -fsSL https://download.docker.com/linux/debian/gpg | sudo
apt-key add -

检查Docker官方GPG公钥指纹是否正确
$ sudo apt-key fingerprint 0EBFCD88

pub 4096R/0EBFCD88 2017-02-22
 Key fingerprint = 9DC8 5822 9FC7 DD38 854A E2D8 8D81 803C 0EBF
CD88
uid Docker Release (CE deb) <docker@docker.com>
sub 4096R/F273FCD8 2017-02-22

将Docker地址注册到软件源中
注意$(lsb_release -cs)是返回当前发行版的版本代号，例如Ubuntu 18.04是bionic，19.10是eoan
但在Ubuntu 19.10发布一段时间后，Docker官方并未在源地址中增加eoan目录，导致此命令安装失败，日后在最新的系统上安装Docker，需要注意排查此问题，手动更改版本代号完成安装
$ sudo add-apt-repository \
"deb [arch=amd64] https://download.docker.com/linux/ubuntu \
```

```
$(lsb_release -cs) \
stable"
```

RedHat系：

```
$ sudo yum install -y yum-utils \
device-mapper-persistent-data \
lvm2

将Docker地址注册到软件源中
$ sudo yum-config-manager \
--add-repo \
https://download.docker.com/linux/centos/docker-ce.repo
```

sh

## 更新系统软件仓库

Debian系：

```
$ sudo apt-get update
```

sh

RedHat系：

```
$ sudo yum update
```

sh

## 安装Docker-Engine Community

Debian系：

```
$ sudo apt-get install docker-ce docker-ce-cli containerd.io
```

sh

RedHat系：

```
$ sudo yum install docker-ce docker-ce-cli containerd.io
```

sh

# 确认Docker安装是否成功

直接运行官方的hello-world镜像测试安装是否成功

```
$ sudo docker run hello-world
```

sh

## 配置国内镜像库 可选

由于Docker官方镜像在国内访问缓慢，官方提供了在国内的镜像库：<https://registry.docker-cn.com>，以加快访问速度（但其实体验也并不快）。

```
该配置文件及目录，在Docker安装后并不会自动创建
$ sudo mkdir -p /etc/docker

配置加速地址
$ sudo tee /etc/docker/daemon.json <<- 'EOF'
{
 "registry-mirrors": ["https://registry.docker-cn.com"]
}
EOF

重启服务
$ sudo systemctl daemon-reload
$ sudo systemctl restart docker
```

sh

### 注意

以上操作有两点提醒读者重点关注：

1. 必须保证daemon.json文件中完全符合JSON格式，如果错了，Docker不会给提示，直接起来。
2. 如果Docker是作为systemd管理的服务的，daemon.json文件会处于锁定状态，应先关闭后再修改配置；

这两点出了问题都会导致Docker服务直接无法启动，如果出现该情况，可以通过systemd status命令检查，看是否有类似如下的错误提示：

```
Drop-In: /etc/systemd/system/docker.service.d
 └─mirror.conf
Active: inactive (dead) (Result: exit-code) since 五 2017-09-15
13:25:28 CST; 7min ago
 Docs: https://docs.docker.com
Main PID: 21151 (code=exited, status=1/FAILURE)
```

如果是，修改daemon.json后重新启动即可。另，关闭systemd服务的方法是：

```
$ sudo systemctl stop docker
$ sudo rm -rf /etc/systemd/system/docker.service.d
```

最后，Docker的官方国内镜像库的速度只能说比起访问国外好了一丢丢，聊胜于无。国内还有一些公开的镜像库，如微软的、网易的等，但要么是不稳定，要么也是慢。比较靠谱的是阿里云的镜像库，但这个服务并不是公开的，需要使用者先到阿里云去申请开发者账户，再使用加速服务，申请后会得一个类似于“<https://yourname.mirror.aliyuncs.com>”的私有地址，把它设置到daemon.json中即可使用。

## 为Docker设置代理 可选

另外一种解决Docker镜像下载速度慢的方案就是使用代理，Docker的代理可以直接读取系统的全局代理，即系统中的HTTP\_PROXY、HTTPS\_PROXY两个环境变量。不过，如果设置这两个变量，其他大量Linux下的其他工具也会受到影响，所以建议的方式是给Docker服务设置专有的环境变量，我们使用Systemd来管理Docker服务，那直接给这个服务设置一个额外配置即可，操作如下：

```
sudo mkdir -p /etc/systemd/system/docker.service.d
配置代理地址，支持http、https、socks、socks5等协议
$ sudo tee /etc/systemd/system/docker.service.d/http-proxy.conf <<-
'EOF'
[Service]
```

```
Environment="HTTP_PROXY=socks5://192.168.31.125:2012"
EOF

#重启docker
$ sudo systemctl restart docker
```

设置后可以通过systemctl检查一下环境变量，看看是否有设置成功：

```
$ systemctl show --property=Environment docker
```

sh

输出：

```
Environment=HTTP_PROXY=socks5://192.168.31.125:2012
```

sh

## 开放Docker远程服务

可选

如果需要在其他机器上管理Docker——譬如典型的如在IntelliJ IDEA这类IDE环境中给远程Docker部署镜像，那可以开启Docker的远程管理端口，这步没有设置任何安全访问措施，请不要在生产环境中进行。

具体做法是修改Docker的服务配置：

Debian系：

```
$ sudo vim /lib/systemd/system/docker.service
```

sh

RedHat系：

```
$ sudo vim /usr/lib/systemd/system/docker.service
```

sh

在ExecStart后面增加以下参数（2375端口可以自定义）：

```
-H tcp://0.0.0.0:2375 -H unix://var/run/docker.sock
```

sh

譬如，默认安装完Docker，修改之后完整的ExecStart应当如下所示：

```
sh
ExecStart=/usr/bin/dockerd -H fd:// --
containerd=/run/containerd/containerd.sock -H tcp://0.0.0.0:2375 -H
unix://var/run/docker.sock
```

最后重启Docker服务即可：

```
sh
#重启docker
$ sudo systemctl daemon-reload
$ sudo systemctl restart docker
```

## 启用Docker命令行自动补全功能 可选

在控制台输入docker命令时可以获得自动补全能力，提高效率。

Docker自带了bash的命令行补全，用其他shell，如zsh，则需采用zsh的插件或者自行获取补全信息

bash：

```
sh
$ echo 'source /usr/share/bash-completion/completions/docker' >>
~/.bashrc
```

zsh：

```
sh
$ mkdir -p ~/.zsh/completion
$ curl -L
https://raw.githubusercontent.com/docker/cli/master/contrib/completion/zs
h/_docker > ~/.zsh/completion/_docker

$ echo 'fpath=(~/zsh/completion $fpath)' >> ~/.zshrc
$ echo 'autoload -Uz compinit && compinit -u' >> ~/.zshrc
```

## 将Docker设置为开机启动

可选

一般使用systemd来管理启动状态

```
设置为开机启动
$ sudo systemctl enable docker

立刻启动Docker服务
$ sudo systemctl start docker
```

sh

## 安装Docker-Compose

在开发和部署微服务应用时，经常要使用Docker-Compose来组织多个镜像，对于Windows系统它是默认安装的，在Linux下需要另外下载一下，下载后直接扔到bin目录，加上执行权限即可使用

```
从GitHub下载
sudo curl -L
"https://github.com/docker/compose/releases/download/1.25.5/docker-
compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
从国内镜像下载
sudo curl -L
"https://get.daocloud.io/docker/compose/releases/download/1.25.5/docker-
compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose

sudo chmod +x /usr/local/bin/docker-compose
```

sh

## 卸载Docker

Debian系：

```
$ sudo apt-get purge docker-ce
```

sh

```
清理Docker容器缓存和自定义配置
$ sudo rm -rf /var/lib/docker
```

RedHat系：

```
$ sudo yum remove docker-ce

清理Docker容器缓存和自定义配置
$ sudo rm -rf /var/lib/docker
```

sh

# 部署Kubernetes集群

Kubernetes是一个由Google发起的开源自动化部署，缩放，以及容器化管理应用程序的容器编排系统。

部署Kubernetes曾经是一件比较麻烦的事情，kubelet、Api-Server、etcd、controller-manager等每一个组件都需要自己部署，还要创建自签名证书来保证各个组件之间的网络访问。但程序员大概是最爱偷懒最怕麻烦的群体，随着Kubernetes的后续版本不断改进（如提供了自动生成证书、Api-Server等组件改为默认静态Pod部署方式），使得部署和管理Kubernetes集群正在变得越来越简单。目前主流的方式大致有：

- [使用Kubeadm部署Kubernetes集群](#)  
其他如KubeSphere等在Kubernetes基础上构建的工具均归入此类
- [使用Rancher部署、管理Kubernetes集群](#)  
其他如KubeSphere等在Kubernetes基础上构建的工具均归入此类
- [使用Minikube在本地单节点部署Kubernetes集群](#)  
其他如Microk8s等本地环境的工具均归入此类
- [在Google Kubernetes Engine云原生环境中部署](#)  
其他如AWS、阿里云、腾讯云等提供的Kubernetes云主机均归入此类

以上集中部署方式都有很明显的针对性，个人开发环境以Minikube最简单，生产环境以Rancher最简单，在云原生环境中，自然是使用环境提供的相应工具。不过笔者推荐首次接触Kubernetes的同学最好还是选择Kubeadm来部署，毕竟这是官方提供的集群管理工具，是相对更底层、基础的方式，充分熟悉了之后再接触其他简化的方式会快速融会贯通。以上部署方式无需全部阅读，根据自己环境的情况选择其一即可。

# 使用Kubeadm部署

尽管使用Rancher或者KubeSphere这样更高层次的管理工具，可以更“傻瓜式”地部署和管理Kubernetes集群，但kubeadm作为官方提供的用于快速安装Kubernetes的命令行工具，仍然是应该掌握的基础技能。kubeadm随着新版的Kubernetes同步更新，时效性也会比其他更高层次的管理工具来的更好。

随着Kubernetes不断成熟，kuberadm无论是部署单控制平面（Single Control-Plane，单Master节点）集群还是高可用（High-Availability，多Master节点）集群，都已经有了更优秀的易用性，现在手工部署Kubernetes集群已经不是什么太复杂、困难的事情了。本文以Debian系的Linux为例，介绍通过kuberadm部署集群的全过程。

## 注意事项

1. 安装Kubernetes集群，需要从谷歌的仓库中拉取镜像，由于国内访问谷歌的网络受阻，需要通过科学上网或者在Docker中预先拉取好所需镜像等方式解决。
2. 集群中每台机器的Hostname不要重复，否则Kubernetes从不同机器收集状态信息时会产生干扰，被认为是同一台机器。
3. 安装Kubernetes最小需要2核CPU、2GB内存，且为x86架构（暂不支持ARM架构）。对于物理机来说，今时今日要找一台不满足以上条件的机器很困难，但对于云主机来说，尤其是购买网站上最低配置的同学，要注意一下是否达到了最低要求，不清楚的话请在/proc/cpuinfo、/proc/meminfo中确认一下。
4. 确保网络通畅的——这听起来像是废话，但确实有相当一部分的云主机默认不对selinux、iptable、安全组、防火墙进行设置的话，内网各个节点之间、与外网之间会存在访问障碍，导致部署失败。

## 注册apt软件源

由于Kubernetes并不在主流Debian系统自带的软件源中，所以要手工注册，然后才能使用apt-get安装。

官方的GPG Key地址为：<https://packages.cloud.google.com/apt/doc/apt-key.gpg>，其中包括的软件源的地址为：<https://apt.kubernetes.io/>（该地址最终又会被重定向至：<https://packages.cloud.google.com/apt/>）。如果能访问google.com域名的机器，采用以下方法注册apt软件源是最佳的方式：

```
添加GPG Key
$ sudo curl -fsSL https://packages.cloud.google.com/apt/doc/apt-key.gpg
| sudo apt-key add -

添加K8S软件源
$ sudo add-apt-repository "deb https://apt.kubernetes.io/ kubernetes-xenial main"
```

sh

对于不能访问google.com的机器，就要借助国内的镜像源来安装了。虽然在这些镜像源中我已遇到过不止一次同步不及时的问题了——就是官方源中已经发布了软件的更新版本，而镜像源中还是旧版的，除了时效性问题外，还出现过其他的一些一致性问题，但是总归比没有的强。国内常见用的apt源有阿里云的、中科大的等，具体为：

阿里云：

- GPG Key：<http://mirrors.aliyun.com/kubernetes/apt/doc/apt-key.gpg>
- 软件源：<http://mirrors.aliyun.com/kubernetes/apt/>

中科大：

- GPG Key：[https://raw.githubusercontent.com/EagleChen/kubernetes\\_init/master/kube\\_apt\\_key.gpg](https://raw.githubusercontent.com/EagleChen/kubernetes_init/master/kube_apt_key.gpg)
- 软件源：<http://mirrors.ustc.edu.cn/kubernetes/apt/>

它们的使用方式与官方源注册过程是一样的，只需替换里面的GPG Key和软件源的URL地址即可，譬如阿里云：

```
添加GPG Key
$ curl -fsSL http://mirrors.aliyun.com/kubernetes/apt/doc/apt-key.gpg |
sudo apt-key add -

添加K8S软件源
```

sh

```
$ sudo add-apt-repository "deb http://mirrors.aliyun.com/kubernetes/apt kubernetes-xenial main"
```

添加源后记得执行一次更新：

```
$ sudo apt-get update
```

sh

## 安装kubelet、kubectl、kubeadm

其实并不需要在每个节点都装上kubectl，但是，我缺的是哪点磁盘空间？

下面简要列出了这三个工具/组件的作用，现在看不看得懂都没有关系，以后用到它们的机会多得是，要相信日久总会生情的。

- kubeadm: 引导启动Kubernetes集群的命令行工具。
- kubelet: 在群集中的所有计算机上运行的组件，并用来执行如启动pods和containers等操作。
- kubectl: 用于操作运行中的集群的命令行工具。

```
$ sudo apt-get install kubelet kubeadm kubectl
```

sh

## 初始化集群前的准备

在使用kubeadm初始化集群之前，还有一些必须的前置工作要妥善处理：

首先，基于安全性（如在文档中承诺的Secret只会在内存中读写）、利于保证节点同步一致性等原因，从1.8版开始，Kubernetes就在它的文档中明确声明了它**默认不支持Swap分区**，在未关闭Swap分区时，集群将直接无法启动。关闭Swap的命令为：

```
$ sudo swapoff -a
```

sh

上面这个命令是一次性的，只在当前这次启动中生效，要彻底关闭Swap分区，需要在文件系统分区表的配置文件中去直接除掉Swap分区。使用vim打开/etc/fstab，注释其中带有sw

ap的行即可，或使用以下命令直接完成修改：

```
还是先备份一下
$ yes | sudo cp /etc/fstab /etc/fstab_bak

进行修改
$ sudo cat /etc/fstab_bak | grep -v swap > /etc/fstab
```

sh

### 可选操作

当然，在服务器上使用的话，关闭Swap影响还是很大的，如果服务器除了Kubernetes还有其他用途的话（除非实在太穷，否则建议不要这样混用；一定要混用的话，宁可把其他服务搬到Kubernetes上）。关闭Swap有可能会对其他服务产生不良的影响，这时需要修改每个节点的kubelet配置，去掉必须关闭Swap的默认限制，具体操作为：

```
$ echo "KUBELET_EXTRA_ARGS=--fail-swap-on=false" >>
/etc/sysconfig/kubelet
```

sh

其次，由于Kubernetes与Docker默认的cgroup（root控制组）驱动程序并不一致，Kubernetes默认为systemd，而Docker默认为cgroupfs。

### 更新信息

从1.18开始，Kubernetes默认的cgroup驱动已经默认修改成cgroupfs了，这时候再进行改动反而会不一致

### 额外知识

Kubernetes是在Docker之上做容器编排的，为什么它的cgroup驱动会被设计成与Docker的不一致？

尽管可能绝大多数的Kubernetes都是使用Docker作为容器配合使用的，但这两者并没有什么绝对绑定的依赖关系，Kubernetes对其管理的容器发布了一套名为”容器运行时接口“（Container Runtime Interface，CRI）的API，这套API在设计上，刻意兼容了”容器开放联盟“（Open Container Initiative，OCI）所制定的容器运行时标准，其他符合OCI标准的容器，同样也是可以与Kubernetes配合工作的，常见的有以下四种：

- CRI-O：由Kubernetes自己发布的ORI参考实现

- [rktlet](#) : rkt容器运行时
- [Frakti](#) : 一种基于Hypervisor的容器运行时
- [Docker CRI shim](#) : 支持Docker直接充当CRI适配器

在这里我们要修改Docker或者Kubernetes其中一个的cgroup驱动，以便两者统一。根据官方文档《[CRI installation](#)》中的建议，对于使用systemd作为引导系统的Linux的发行版，使用systemd作为Docker的cgroup驱动程序可以服务器节点在资源紧张的情况下表现得更为稳定。

### 额外知识

cgroups是Linux内核提供的一种可以限制单个进程或者多个进程所使用资源的机制，可以对cpu，内存等资源实现精细化的控制。

这里选择修改各个节点上Docker的cgroup驱动为systemd，具体操作为编辑（无则新增）/etc/docker/daemon.json文件，加入以下内容即可：

```
{
 "exec-opts": ["native.cgroupdriver=systemd"]
}
```

然后重新启动Docker容器：

```
$ systemctl daemon-reload
$ systemctl restart docker
```

sh

## 预拉取镜像

可选

预拉取镜像并不是必须的，本来初始化集群的时候系统就会自动拉取Kubernetes中要使用到的Docker镜像组件，也提供了一个“kubeadm config images pull”命令来一次性的完成拉取，这都是因为如果要手工来进行这项工作，实在非常非常非常的繁琐。

但对于许多人来说这项工作往往又是无可奈何的，Kubernetes的镜像都存储在k8s.gcr.io上，如果您的机器无法直接或通过代理访问到gcr.io（Google Container Registry，敲黑

板：这是属于谷歌的网址）的话，初始化集群时自动拉取就无法顺利进行，所以就不得不手工预拉取。

预拉取的意思是，由于Docker只要查询到本地有相同（名称和tag完全相同、哈希相同）的镜像，就不会访问远程仓库，那只要从GitHub上拉取到所需的镜像，再将tag修改成官方的一致，就可以跳过网络访问阶段。

首先使用以下命令查询当前版本需要哪些镜像：

```
$ kubeadm config images list --kubernetes-version v1.17.3
sh
k8s.gcr.io/kube-apiserver:v1.17.3
k8s.gcr.io/kube-controller-manager:v1.17.3
k8s.gcr.io/kube-scheduler:v1.17.3
k8s.gcr.io/kube-proxy:v1.17.3
k8s.gcr.io/pause:3.1
k8s.gcr.io/etcd:3.4.3-0
k8s.gcr.io/coredns:1.6.5
.....
```

这里必须使用“--kubernetes-version”参数指定具体版本，因为尽管每个版本需要的镜像信息在本地是有存储的，但如果不去加的话，Kubernetes将向远程GCR仓库查询最新的版本号，会因网络无法访问而导致问题。但加版本号的时候切记不能照抄上面的命令中的“v1.17.3”，应该与你安装的kubelet版本保持一致，否则在初始化集群控制平面的时候会提示控制平面版本与kubectl版本不符。

得到这些镜像名称和tag后，可以从[DockerHub](#)上找存有相同镜像的仓库来拉取，至于具体哪些公开仓库有，考虑到以后阅读本文时Kubernetes的版本应该会有所差别，所以需要自行到网站上查询一下。笔者比较常用的是一个名为“anjia0532”的仓库，有机器人自动跟官方同步，相对比较及时。

```
#以k8s.gcr.io/coredns:1.6.5为例，每个镜像都要这样处理一次
$ docker pull anjia0532/google-containers.coredns:1.6.5

#修改tag
$ docker tag anjia0532/google-containers.coredns:1.6.5
k8s.gcr.io/coredns:1.6.5
sh
```

```
#修改完tag后就可以删除掉旧镜像了
$ docker rmi anjia0532/google-containers.coredns:1.6.5
```

## 初始化集群控制平面

到了这里，终于可以开始Master节点的部署了，先确保kubelet是开机启动的：

```
$ sudo systemctl start kubelet
$ sudo systemctl enable kubelet
```

sh

接下来使用su直接切换到root用户（而不是使用sudo），然后使用以下命令开始部署：

```
$ kubeadm init --kubernetes-version v1.17.3 --pod-network-cidr=10.244.0.0/16
```

sh

这里使用“--kubernetes-version”参数（要注意版本号与kubelet一致）的原因与前面预拉取是一样的，避免额外的网络访问；另外一个参数“--pod-network-cidr”着在稍后介绍完CNI网络插件时会去说明。

当看到下面信息之后，说明集群主节点已经安装完毕了。

```
Your Kubernetes control-plane has initialized successfully! If kubeadm can't pull the image, the following log appears:
To start using your cluster, you need to run the following as a regular user:
 mkdir -p $HOME/.kube
 sudo cp -i /etc/kubernetes/admin.conf $HOME/.kube/config
 sudo chown $(id -u):$(id -g) $HOME/.kube/config

You should now deploy a pod network to the cluster.
Run "kubectl apply -f [podnetwork].yaml" with one of the options listed at:
 https://kubernetes.io/docs/concepts/cluster-administration/addons/
Then you can join any number of worker nodes by running the following on each as root:
 kubeadm join 10.3.7.5:6443 --token ejg4tt.y08moym055dn9i32 \
 --discovery-token-ca-cert-hash sha256:9d2079d2844fa2953d33cc0da57ab15f571e974aa40ccb50edde12c5e906d513
```

这信息先恭喜你已经把控制平面安装成功了，但还有三行“you need.....”、“you should.....”、“you can.....”开头的内容，这是三项后续的“可选”工作，下面继续介绍。

## 为当前用户生成kubeconfig

使用Kubernetes前需要为当前用户先配置好admin.conf文件。切换至需配置的用户后，进行如下操作：

```
$ mkdir -p $HOME/.kube
$ sudo cp -i /etc/kubernetes/admin.conf $HOME/.kube/config
$ sudo chown $(id -u):$(id -g) $HOME/.kube/config
```

sh

## 安装CNI插件 可选

CNI即“容器网络接口”，在2016年，CoreOS发布了CNI规范。2017年5月，CNI被CNCF技术监督委员会投票决定接受为托管项目，从此成为不同容器编排工具（Kubernetes、Mesos、OpenShift）可以共同使用的、解决容器之间网络通讯的统一接口规范。

部署Kubernetes时，我们可以有两种网络方案使得以后受管理的容器之间进行网络通讯：

- 使用Kubernetes的默认网络
- 使用CNI及其插件

第一种方案，尤其不在GCP或者AWS的云主机上，没有它们的命令行管理工具时，需要大量的手工配置，基本上是反人类的。实际通常都会采用第二种方案，使用CNI插件来处理容器之间的网络通讯，所以本节所标识的“[可选]”其实也也没什么选择不安装CNI插件的余地。

Kubernetes目前支持的CNI插件有：Calico、Cilium、Contiv-VPP、Flannel、Kube-router、Weave Net等六种，每种网络提供了不同的管理特性（如MTU自动检测）、安全特性（如是否支持加密通讯）、网络策略（如Ingress、Egress规则）、传输性能（甚至对TCP、UDP、HTTP、FTP、SCP等不同协议来说也有不同的性能表现）以及主机的性能消耗。后续我们将专门对不同CNI插件进行测试对比，在环境部署这部分，对于初学者来说，使用Flannel是较为合适的，它是最精简的CNI，没有安全特性的支持，主机压力小，安装便捷，效率也不错，使用以下命令安装Flannel网络：

```
$ curl --insecure -sfL
https://raw.githubusercontent.com/coreos/flannel/master/Documentation/kube-flannel.yml | kubectl apply -f -
```

sh

使用Flannel的话，要注意要在创建集群时加入“--pod-network-cidr”参数，指明网段划分。

## 移除Master节点上的污点 可选

污点（Taint）是Kubernetes Pod调度中的概念，在这里通俗地理解就是Kubernetes决定在集群中的哪一个节点建立新的容器时，要先排除掉带有特定污点的节点，以避免容器在Kubernetes不希望运行的节点中创建、运行。默认情况下，集群的Master节点是会带有污点的，以避免容器分配到Master中创建。但对于许多学习Kubernetes的同学来说，并没有多宽裕的机器数量，往往是建立单节点集群或者最多只有两、三个节点，这样Master节点不能运行容器就显得十分浪费了。需要移除掉Master节点上所有的污点，在Master节点上执行以下命令即可：

```
$ kubectl taint nodes --all node-role.kubernetes.io/master-
```

sh

做到这步，如果你只有一台机器的话，那Kubernetes的安装已经宣告结束了，可以使用此环境来完成后续所有的部署。你还可以通过cluster-info和get nodes子命令来查看一下集群的状态，类似如下所示：

```
ubuntu @ linux in ~ [15:37:45]
$ kubectl cluster-info
Kubernetes master is running at https://10.3.7.5:6443
KubeDNS is running at https://10.3.7.5:6443/api/v1/namespaces/kube-system/services/kube-dns:dns/proxy

To further debug and diagnose cluster problems, use 'kubectl cluster-info dump'.

ubuntu @ linux in ~ [15:37:53]
$ kubectl get nodes
NAME STATUS ROLES AGE VERSION
lab.server Ready master 29m v1.17.3
```

## 调整NodePort范围 可选

Kubernetes默认的NodePort范围为30000-32767，为了方便使用低端口，可能需要修改此范围，这需要调整Api-Server的启动参数，具体操作如下（如过是高可用部署，需要对每一个Master节点进行修改）：

- 修改 /etc/kubernetes/manifests/kube-apiserver.yaml 文件，添加一个参数在 spec.containers.command 中增加一个参数 --service-node-port-range=1-32767

- 重启Api-Server，现在Kubernetes基本都是以静态Pods模式部署，Api-Server是一个直接由kubelet控制的静态Pod，删除后它会自动重启：

```
获得 apiserver 的 pod 名字
export apiserver_pods=$(kubectl get pods --selector=component=kube-
apiserver -n kube-system --output=jsonpath={.items..metadata.name})
删除 apiserver 的 pod
kubectl delete pod $apiserver_pods -n kube-system
```

sh

- 验证修改结果：可以在pod中看到该参数即可

```
kubectl describe pod $apiserver_pods -n kube-system
```

sh

## 启用kubectl命令自动补全功能

可选

由于kubectl命令在后面十分常用，而且Kubernetes许多资源名称都带有随机字符，要手工照着敲很容易出错，强烈推荐启用命令自动补全的功能，这里仅以bash和笔者常用的zsh为例，如果您使用其他shell，需自行调整：

bash：

```
$ echo 'source <(kubectl completion bash)' >> ~/.bashrc
$ echo 'source /usr/share/bash-completion/bash_completion' >> ~/.bashrc
```

sh

zsh：

```
$ echo 'source <(kubectl completion zsh)' >> ~/.zshrc
```

sh

## 将其他Node节点加入到Kubernetes集群中

在安装Master节点时候，输出的最后一部分内容会类似如下所示：

Then you can join any number of worker nodes by running the following on each as root:

```
kubeadm join 10.3.7.5:6443 --token ejg4tt.y08moym055dh9i32 \
--discovery-token-ca-cert-hash
sha256:9d2079d2844fa2953d33cc0da57ab15f571e974aa40ccb50edde12c5e906d513
```

这部分内容是告诉用户，集群的Master节点已经建立完毕，其他节点的机器可以使用“kubeadm join”命令加入集群。这些机器只要完成kubeadm、kubelet、kubectl的安装即可、其他的所有步骤，如拉取镜像、初始化集群等等都不需要去做，就是可以使用该命令加入集群了。需要注意的是，该Token的有效时间为24小时，如果超时，使用以下命令重新获取：

```
$ kubeadm token create --print-join-command
```

sh

# 使用Rancher部署

Rancher是在Kubernetes更上层的管理框架，Rancher是图形化的，有着比较傻瓜式的操作，只有少量一两处地方（如导入集群）需要用到Kubernetes命令行。也由于它提供了一系列容器模版、应用商店等的高层功能，使得要在Kubernetes上部署一个新应用，简化到甚至只需要点几下鼠标即可，因此用户们都爱使用它。

Rancher还推出了RancherOS（极致精简专为容器定制的Linux，尤其适合边缘计算环境）、K3S（Kubernetes as a Service，5 Less Than K8S，一个大约只有40MB，可以运行在x86和ARM架构上的极小型Kubernetes发行版）这样的定制产品，用以在用户心中暗示、强化比K8S更小、更简单、更易用的主观印象。

不过也由于Rancher入门容易，基础性的应用需求解决起来很方便，也导致了不少人一开始使用它之后，就陷入了先入为主的印象，后期再接触Kubernetes时，便觉得学习曲线特别陡峭，反而限制了某些用户对底层问题的进一步深入。

在本文中，笔者以截图为主，展示如何使用Rancher来导入或者创建Kubernetes集群的过程。

## 安装Rancher

前置条件：已经安装好Docker。

使用Docker执行Rancher镜像，执行以下命令即可：

```
$ sudo docker run -d --restart=unless-stopped -p 8080:80 -p 8443:443
rancher/rancher
```

sh

## 使用Rancher管理现有Kubernetes集群

前置条件：已经安装好了Kubernetes集群。

使用Rancher的导入功能将已部署的Kubernetes集群纳入其管理。登陆Rancher主界面（首次登陆会要求设置admin密码和Rancher在集群中可访问的路径，后者尤其不能乱设，否则Kubernetes无法访问到Rancher会一直处于Pending等待状态）之后，点击右上角的Add Cluster，然后有下面几个添加集群的选择：

Add Cluster - Select Cluster Type

From existing nodes (Custom)  
Create a new Kubernetes cluster using RKE, out of existing bare-metal servers or virtual machines.

Import an existing cluster  
Import an existing Kubernetes cluster. The provider that created it will continue to manage the provisioning and configuration of the cluster.

With RKE and new nodes in an infrastructure provider

- Amazon EC2
- Azure
- DigitalOcean
- Linode
- vSphere

With a hosted Kubernetes provider

- Amazon EKS
- Azure AKS
- Google GKE

- 要从某台机器中新安装Kubernetes集群选择“From existing nodes (Custom)”
- 要导入某个已经安装好的Kubernetes集群选择“Import an existing cluster”
- 要从各种云服务商的RKE ( Rancher Kubernetes Engine ) 环境中创建，就选择下面那排厂商的按钮，没有的话（譬如国内的阿里云之类的），请先到Tools->Driver中安装对应云服务厂商的驱动。

这里选择“Import an existing cluster”，然后给集群起个名字以便区分（由于Rancher支持多集群管理，所以集群得有个名字以示区别），之后就看见这个界面：

**Note:** If you want to import a Google Kubernetes Engine (GKE) cluster (or any cluster that does not supply you with a `kubectl` configuration file with the ClusterRole `cluster-admin` bound to it), you need to bind the ClusterRole `cluster-admin` using the command below.

Replace `[USER_ACCOUNT]` with your Google account address (you can retrieve this using `gcloud config get-value account`). If you are not importing a Google Kubernetes Engine cluster, replace `[USER_ACCOUNT]` with the executing user configured in your `kubectl` configuration file.

```
kubectl create clusterrolebinding cluster-admin-binding --clusterrole cluster-admin --user [USER_ACCOUNT]
```

Run the `kubectl` command below on an existing Kubernetes cluster running a supported Kubernetes version to import it into Rancher:

```
kubectl apply -f https://localhost:8443/v3/import/vgkj5tzphj9vzg6157krdc9gfc4b4zsfp419prrf6sb7z9d2wvbbh5.yaml
```

If you get an error about 'certificate signed by unknown authority' because your Rancher installation is running with an untrusted/self-signed SSL certificate, run the command below instead to bypass the certificate check:

```
curl --insecure -sfL https://localhost:8443/v3/import/vgkj5tzphj9vzg6157krdc9gfc4b4zsfp419prrf6sb7z9d2wvbbh5.yaml | kubectl apply -f -
```

Rancher自动生成了加入集群的命令，这行命令其实就是部署一个运行在Kubernetes中的代理（Agent），在Kubernetes的命令行中执行以上自动生成的命令。

最后那条命令意思是怕由于部署的Rancher服务没有申请SSL证书，导致HTTPS域名验证过不去，`kubectl`下载不下来yaml。如果你的Rancher部署在已经申请了证书的HTTPS地址上那可以用前面的，否则还是直接用`curl --insecure`命令来绕过HTTPS证书查验吧，譬如以下命令所示：

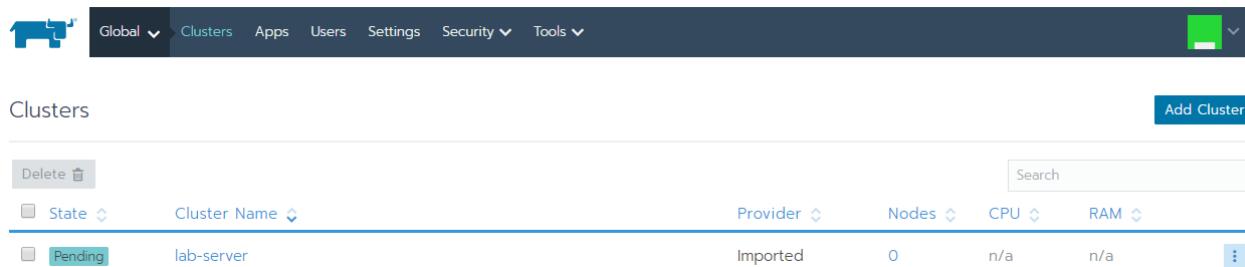
```
$ curl --insecure -sfL
https://localhost:8443/v3/import/vgkj5tzphj9vzg6157krdc9gfc4b4zsfp419pr
rf6sb7z9d2wvbbh5.yaml | kubectl apply -f -
```

多说一句，用哪条命令安装的Agent只决定了yaml文件是如何下载获得的，对后续其他事情是毫无影响的，所以怎么简单怎么来，别折腾。

执行结果类似如下所示，一堆secret、deployment、daementset创建成功，就代表顺利完成了：

```
ubuntu @ linux in ~ [16:41:24]
$ curl --insecure -sfL https://localhost:8443/v3/import/vgkj5tzphj9vzg6157krdc9gfc4b4zsfp419prrf6sb7z9d2wvbbh5.yaml | kubectl apply -f -
clusterrole.rbac.authorization.k8s.io/proxy-clusterrole-kubeapiserver created
clusterrolebinding.rbac.authorization.k8s.io/proxy-role-binding-kubernetes-master created
namespace/cattle-system created
serviceaccount/cattle created
clusterrolebinding.rbac.authorization.k8s.io/cattle-admin-binding created
secret/cattle-credentials-5d6cb35 created
clusterrole.rbac.authorization.k8s.io/cattle-admin created
deployment.apps/cattle-cluster-agent created
daemonset.apps/cattle-node-agent created
```

然后回到Rancher网页，点击界面上的“Done”按钮。可以看到集群正处于Pending状态：



Cluster Name	Provider	Nodes	CPU	RAM
lab-server	Imported	0	n/a	n/a

如果Agent成功到达Running状态的话，这里也会很快就变成Waiting状态，然后再变为Active状态，导入工作即宣告胜利结束。

而如果一直持续Pending状态，说明安装的Agent运行失败。典型的原因是无法访问到Rancher的服务器，这时可以通过kubectl logs命令查看一下cattle-cluster-agent-xxx的日志，通常会看见"XXX is not accessible"，其中的XXX是Rancher第一次进入时跟你确认过的访问地址，假如你乱填了，或者该地址被防火墙挡掉，又或者因为证书限制等其他原因导致Agent无法访问，Rancher就会一直Pending。

最后再提一句，Rancher与Kubernetes集群之间是被动链接的，即由Kubernetes去主动找Rancher，这意味着部署在外网的Rancher，可以无障碍地管理处于内网（譬如NAT后）的Kubernetes集群，这对于大量没有公网IP的集群来说是很方便的事情。

## 使用Rancher创建Kubernetes集群

也可以直接使用Rancher直接在裸金属服务器上创建Kubernetes集群，此时在添加集群中选择From existing nodes (Custom)，在自定义界面中，设置要安装的集群名称、Kubernetes版本、CNI网络驱动、私有镜像库以及其他一些集群的参数选项。

添加集群 - Custom

集群名称 \*

hk

添加描述

成员角色

控制哪些用户可以访问集群，以及他们拥有的对其进行更改的权限。

标签/注释

为集群配置标签和注释。

无

集群选项

编辑 YAML

全部展开

Kubernetes选项

自定义集群功能

Kubernetes版本

v117.2-rancher1-2

网络驱动

Flannel

Windows支持

启用

禁用

项目网络隔离

启用

禁用

网络 MTU

0

Only applied if the value is non-zero. When applied, the MTU value is explicitly configured for the chosen network provider (disabling auto-discovery). The override must be calculated from the host's MTU minus the CNI plugin's required overhead.

下一步确认该主机在Kubernetes中扮演的角色，每台主机可以扮演多个角色。但至少要保证每个集群都有一个Etcd角色、一个Control角色、一个Worker角色。

**添加集群 - Custom**

**集群选项**

**添加主机命令**

选择主机角色,端口映射请参考: <https://rancher.com/docs/rancher/v2.x/en/installation/references/>

角色选择 (每台主机可以运行多个角色。每个集群至少需要一个Etcd角色、一个Control角色、一个Worker角色)

Etcd     Control     Worker

**主机地址**

为主机配置公网地址和内网地址, 如果为VPC网络的云服务器, 如果不指定公网地址节点将无法获取到对应公网IP。

公网地址: 例如: 12.3.4    内网地址: 例如: 12.3.4

**节点名称**

(可选) 自定义节点显示的名称, 不显示实际的主机名

例如: My-worker-node

**主机标签**

(可选) 添加到节点的标签

+ 添加标签

**节点污点 (Taints)**

(可选) 添加到节点的污点 (taints)

+ 添加污点 (Taint)

**2 复制以下命令在主机的SSH终端运行。**

```
sudo docker run -d --privileged --restart=unless-stopped --net=host -v /etc/kubernetes:/etc/kubernetes -v /var/run:/var/run
rancher/rancher-agent:v2.3.5 --server https://k8s.icyfenix.cn:444 --token 8snnt4mmj4hfwld892cl8f7knwj1v8824hwtm05grqm7gg7ftzkz2 --ca-
checksum 825812c06dea6cf75008f91df2ccabe23b177b7d3cbd522585af025cdbe5ec4b --etcd --controlplane --worker
```

复制生成的命令，在要安装集群的每一台主机的SSH中执行。此时Docker会下载运行Rancher的Agent镜像，当执行成功后，Rancher界面会有提示新主机注册成功。



点击完成，将会在集群列表中看见正在Provisioning的新集群，稍后将变为Active状态。

状态	集群名称	供应商	主机数	处理器	内存
Provisioning	hk	自定义	0	n/a	n/a

Waiting for etcd and controlplane nodes to be registered

安装完成后你就可以在Rancher的图形界面管理Kubernetes集群了，如果还需要在命令行中工作，kubectl、kubeadm等工具是没有安装的，可参考“[使用Kubeadm部署Kubernetes集群](#)”的内容安装使用。



# 使用Minikube部署

Minikube是Kubernetes官方提供的专门针对本地单节点集群的Kubernetes集群管理工具。针对本地环境对Kubernetes使用有一定的简化和针对性的补强。这里简要介绍其安装过程

## 安装Minikube

Minikube是一个单文件的二进制包，安装十分简单，在已经完成Docker安装的前提下，使用以下命令可以下载并安装最新版的Minikube。

```
$ curl -Lo minikube
https://storage.googleapis.com/minikube/releases/latest/minikube-linux-
amd64 && chmod +x minikube && sudo mv minikube /usr/local/bin/
sh
```

## 安装Kubectl工具

Minikube中很多提供了许多子命令以代替Kubectl的功能，安装Minikube时并不会一并安装Kubectl。但是Kubectl作为集群管理的命令行，要了解Kubernetes是无论如何绕不过去的，通过以下命令可以独立安装Kubectl工具。

```
$ curl -LO https://storage.googleapis.com/kubernetes-
release/release/$(curl -s https://storage.googleapis.com/kubernetes-
release/release/stable.txt)/bin/linux/amd64/kubectl && chmod +x kubectl
&& sudo mv kubectl /usr/local/bin/
sh
```

## 启动Kubernetes集群

有了Minikube，通过start子命令就可以一键部署和启动Kubernetes集群了，具体命令如下：

```
$ minikube start --iso-url=https://kubernetes.oss-cn-hangzhou.aliyuncs.com/minikube/iso/minikube-v1.6.0.iso
 --registry-mirror=https://registry.docker-cn.com
 --image-mirror-country=cn
 --image-repository=registry.cn-hangzhou.aliyuncs.com/google_containers
 --vm-driver=none
 --memory=4096
```

sh

以上命令中，明确要求Minikube从指定的地址下载虚拟机镜像、Kubernetes各个服务Pod的Docker镜像，并指定了使用Docker官方节点作为国内的镜像加速服务。

“vm-drvier”参数是指Minikube所采用的虚拟机，根据不同操作系统，不同的虚拟机可以有以下选项：

操作系统	支持虚拟机	参数值
Windows	Hyper-V	hyperv
Windows	VirtualBox	virtualbox
Linux	KVM	kvm2
Linux	VirtualBox	virtualbox
MacOS	HyperKit	hyperkit
MacOS	VirtualBox	virtualbox
MacOS	Parallels Desktop	parallels
MacOS	VMware Fusion	vmware

特别需要提一下的是如果读者使用的并非物理机器，而是云主机环境——现在流行将其成为“裸金属”（Bare Metal）服务器，那在上面很可能是无法再部署虚拟机环境的，这时候应该将vm-drvier参数设为none。也可以使用以下命令设置虚拟机驱动的默认值：

```
$ minikube config set vm-driver none
```

sh

至此，整个Kubernetes就一键启动完毕了。其他工作，如命令行的自动补全，可参考使用Kubeadm安装Kubernetes集群中相关内容。



# 运维环境

# 在K8S上部署ELK/EFK日志监控

# 在K8S上部署DevOps