异步图书
www.epubit.com

分布式系统设计实践

李庆旭 ◎ 编著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

分布式系统设计实践

李庆旭◎编著

人民邮电出版社
北京

图书在版编目（CIP）数据

分布式系统设计实践 / 李庆旭编著. -- 北京 : 人民邮电出版社, 2019.11
ISBN 978-7-115-51945-0

Ⅰ. ①分… Ⅱ. ①李… Ⅲ. ①分布式操作系统－系统设计 Ⅳ. ①TP316.4

中国版本图书馆CIP数据核字(2019)第190630号

内 容 提 要

本书对近年来涌现出的各种主流分布式技术做了简要介绍和全面梳理。本书将分布式系统中涉及的技术分为前端构造技术、分布式中间件技术和分布式存储技术三大类，对每类技术都详细介绍了其原理、主流实现的设计思想和架构，以及相关应用场景。此外，本书还总结了分布式系统的构建思想，并分别针对业界几个非常成功的大型分布式系统（谷歌搜索系统、淘宝网电商平台、阿里云公有云平台、领英社交平台）进行了案例研究。

本书适合业界的架构师、工程师、项目管理人员，也适合大中专院校的高年级本科生和研究生参考和阅读。

◆ 编　　著　李庆旭
责任编辑　杨海玲
责任印制　马振武

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：14
字数：292 千字　　2019 年 11 月第 1 版
印数：1－2 500 册　　2019 年 11 月北京第 1 次印刷

定价：59.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

序

我很喜欢看书。从小到大，我看过很多书。

每看一本好书，我都会对书的作者充满感激和敬意，因为他愿意花那么多精力，写一本那么好的书给别人看，实在是非常值得尊敬。我知道，写书是赚不了多少钱的，一个人不为名、不为利，却愿意花很多精力写一本好书给别人去读、去欣赏，实在是难能可贵！

因此，我一直有一个想法，就是有一天，我也要尽力去写一本“好点儿”的书，一则对自己的所学做一个系统的整理，二则也为读者做些贡献。

我不知道是否有人会喜欢我的这本书，但无论如何，我已经尽力把它写好，因为我想和那些自己曾经读过的好书的作者一样，奉献给读者一本值得读的书。

我从 1991 年开始接触计算机，绝对算是一个“老码农”了。屈指一算，从那时到现在，计算机技术经历了几次重要的更新换代。

第一次发生在 20 世纪 90 年代初，是从 DOS 到 Windows 的换代。这一次换代淘汰了一批落后的公司和产品，亦有一批新的公司和产品崛起。当年 DOS 下的 WPS 如日中天，然而 Windows 一经推出，就难觅 WPS 的踪迹了。直到 21 世纪初，新版 Windows 下的金山 WPS 才姗姗而来。如果当年求伯君先生能及时推出新版的 Windows 下的 WPS，那么中文办公软件的市场也许就不是今天这样了。

第二次换代发生在 20 世纪 90 年代末到 21 世纪初，Java 和.NET 这样的基于虚拟机的开发技术开始兴起。在 DOS/Windows 下，C/C++一直是主流的开发语言，而随着计算机软件的复杂度越来越高，C/C++语言开发效率低、查错难的弊端越来越突出，于是 Java 和.NET 这样的基于虚拟机的开发技术日益受到开发者的青睐。这一波技术换代使 Java 的应用范围越来越广，也动摇了微软公司的技术在开发领域中无可替代的地位。

进入 21 世纪后，由于互联网技术，尤其是移动互联网技术的发展，诞生了一批大型互联网公司（谷歌、亚马逊、Facebook、百度、阿里巴巴、腾讯等）。在互联网发展的初期，这些公司在汹涌而来的大数据面前毫无准备，被折腾得筋疲力尽。痛定思痛后，以谷歌为代表的互联网新贵们，开发了许多优秀的技术（GFS、Bigtable、Hadoop 等），完美而优雅地解决了大多数曾经令它们头疼的问题。继谷歌将其技术以论文的形式发表后，业界就开始模仿其设计和架构，相继开源了许多优秀的产品（如 Apache Hadoop 系列产品）。从此，大型分布式技术的大门被轰然推开，一个又一个大型的分布式应用（淘宝网上商城、京东网上商城、微信、谷歌搜索、Twitter 社交平台、Facebook 社交平台等）相继诞生。这就是以移动互联网和分布式计算的大规模应用为代表的第三次技术换代。无疑，这一次换代又

是一次各大公司实力的大洗牌。

第四次大的技术换代是什么？也许它正在进行（如果它是人工智能的话），也许还没有开始……

每一次大的技术换代，都会有公司和技术被无情地淘汰，从 Novell 到 Borland，到 Netscape，再到 Nokia，昔日霸主早已风光不再。但被淘汰的不仅有大公司，同时被淘汰的还有不少从事技术工作的工程师和架构师。因此，为了不被技术浪潮无情地拍倒，每一个 IT 从业者都应该对新技术心存敬畏，都应该关注技术的发展。

今天，对于在 IT 行业从事技术工作的人，无论是工程师、架构师还是管理者，也无论从事的工作是否与分布式相关，都应该了解分布式技术，因为总有一天，你会遇到它、接触它、使用它、理解它、完善它。

然而，分布式技术涉及的方面（存储、计算、框架、中间件等）是如此之多，且迄今为止尚未见到一本书对其进行概括和梳理，要想对分布式技术有全面的了解，特别是对初学者而言，何其难哉！

因此，本书试图对近年来涌现出的各种主流分布式技术做一个简要介绍，以使不太熟悉这个领域的读者能了解其概貌、原理和根源。

本书共分为以下 6 部分。

- 第一部分对典型的分布式系统的组成及其中每个组件的功能进行简要介绍，以使读者对分布式系统有一个总体了解。
- 第二部分介绍分布式系统的前端经常使用的 Web 框架、反向代理及负载均衡技术。
- 第三部分对分布式系统中经常使用的各种中间件技术逐一进行介绍，包括分布式同步服务中间件、关系型数据库访问中间件、分布式服务调用中间件、分布式消息服务中间件和分布式跟踪服务中间件。
- 第四部分介绍分布式文件系统、各种 NoSQL 数据库技术（基于键值对的 NoSQL 技术、基于列的 NoSQL 技术、基于文档的 NoSQL 技术、基于图的 NoSQL 技术）和 NewSQL 数据库系统。
- 第五部分对业界在构建大型分布式系统的过程中的主要经验加以总结，使后来者避免重蹈覆辙。
- 第六部分介绍业界几个知名的大型分布式系统的主要设计思想和架构，包括谷歌搜索系统、淘宝网、阿里云和领英的社交应用。此外，还会探讨和思考分布式系统实现中的一些问题。

在本书写作的过程中，参考了许多网上的资料和书籍，感谢这些资料的作者们。由于涉及的资料太多，无法将其作者一一列出，谨表歉意。

感谢人民邮电出版社的杨海玲编辑，没有你的辛勤付出，也不会有本书的顺利出版。

另外，还要感谢我的好朋友申天雷，感谢你在繁忙的工作之余抽出时间来全面审阅这本书。

最后，感谢我的家人。没有父母的养育，我不可能有机会掌握这些知识；没有我岳父母帮我看管孩子，没有我爱人的关心与支持，我不可能有时间来写作；没有我儿子那双充满期待的眼睛，我也不可能一直坚持把本书写完！

由于分布式系统涉及的内容实在太多，我不可能对其中每个细节都有很深刻和精确的理解，因此，书中错误及疏漏之处在所难免，欢迎批评指正。

李庆旭

2019年夏于北京

资源与支持

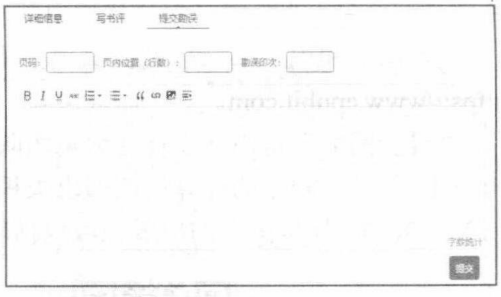
本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供源代码下载，要获得相关配套资源，请在异步社区本书页面中单击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。



当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您来自学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录

第一部分　分布式系统概述

第二部分　分布式系统的前端构造技术

第三部分　分布式中间件

第四部分 分布式存储技术

第五部分 分布式系统的构建思想

第六部分 大型分布式系统案例研究及分析

第一部分

分布式系统概述

之所以需要有分布式系统，最根本的原因还是单机的计算和存储能力不能满足系统的需要，但要把成百上千台计算机组织成一个有机的系统绝非易事。

这一部分会对典型的分布式系统的组成及其每个组件的功能做简要介绍，以便读者对分布式系统有一个总体的了解。

第一部分

第1章

分布式系统概述

1999年8月6日，CNN报道了一起eBay网站的事故：从7:30开始，整个网站崩溃，一直持续了9个多小时。下午5:30后，技术人员开始进行系统恢复，但搜索功能依然不能使用。

2011年4月21日至22日，亚马逊EC2（Elastic Computer Cloud）服务出现大面积事故，导致数以千计的初创公司受到影响，而且造成大约11小时的历史数据永久性丢失。

2013年4月27日，《大掌门》游戏的开发商玩蟹科技CEO叶凯在微博上吐槽，“我们在阿里云上用了20多台机器。半年时间，出现过1次所有机器全部断电，2次多个硬盘突然只读，3次硬盘I/O突然变满……”。

2013年12月28日，春运第一天，铁道部首次推出了网上订票系统，但很快就出现许多用户无法访问、响应缓慢甚至串号等事故。

2017年9月17日，谷歌的网盘服务Drive出现故障，成千上万用户受到影响。

上面的这几起事故，当时都闹得沸沸扬扬，不仅给受影响的用户带来了很大的损失，也极大地影响了厂商的形象。事实上，几乎每一家互联网公司的后台系统都曾经不止一次地经历过这样或那样的尴尬时刻。可以这样说，几乎每一家互联网公司的后台架构都是在发现问题、解决问题的循环中发展起来的。

即便是执分布式系统技术牛耳的谷歌，在2017年9月，也出现过分布式系统的故障。可见，开发并维护一个成功的分布式系统是多么不易！

最早得到广泛应用的分布式系统是诞生于20世纪70年代的以太网。尽管分布式系统存在的历史已经有近半个世纪，然而其大规模的发展和应用则是2000年以后的事情。

21世纪以来，随着雅虎、谷歌、亚马逊、eBay、Facebook、Twitter等众多互联网公司的崛起，其用户量以及要处理的数据量迅速增长，远远超过了传统的计算机系统能够处理的范围，因此，以谷歌为代表的互联网公司提出了许多新技术（如HDFS、Bigtable、

MapReduce 等)。以 BAT 为代表的中国互联网公司，也在 21 世纪整体崛起，在初期借鉴美国公司技术的基础上，他们也自行开发了许多新的技术（如淘宝的管理海量小文件的分布式存储系统 TFS、阿里巴巴开源的分布式调用框架 Dubbo、阿里巴巴开源的数据库中间件 Cobar 等)。

为了解决分布式系统中的各种各样的问题，各大互联网公司开发了各种各样的技术，当然，这也促进了当今分布式系统技术领域的飞速发展。为了存储大量的网站索引，谷歌设计了 GFS 分布式文件存储系统和基于列存储的 Bigtable NoSQL 数据库系统；为了计算 PageRank 算法中的页面 rank 值，谷歌又设计了 MapReduce 分布式计算系统；为了方便其分布式系统中不同主机间的协调，谷歌还设计了 Chubby 分布式锁系统；为了解决不同语言实现的组件间的通信问题，Facebook 设计了 Thrift；为了解决大量消息的快速传递问题，领英设计了 Kafka……这个列表可以很长很长。

为了“压榨”分布式系统中每个组件的性能，人们已经不再仅仅满足于在程序库（如网络编程库 Netty、内存管理库 TCMalloc 等)、程序框架（如 Spring）等“略显浅薄”的地方提高，而是已经渗透到了硬件（如谷歌为其计算中心专门设计了计算机)、网络（如 SDN)、操作系统(如各大互联网公司定制的 Linux 内核)、语言(如谷歌设计的 Go 语言)、数据库系统（如各种 NoSQL 系统)、算法（如人工智能领域的突飞猛进）等各种计算机基础领域。

毫无疑问，我们处于计算机技术发展最为迅猛的时代。在这个如火如荼的时代里，许多尘封多年的计算机技术（如人工智能、分布式系统、移动计算、虚拟计算等)，一改往日不温不火的模样，在互联网这片广袤的土地上如日中天，发展迅速。

今天的计算机领域，已经与 20 年前大为不同。20 年前，只需要对操作系统、数据库、网络、编译等领域有深刻的理解，再熟练掌握几门计算机语言，了解一些常见的软件架构（客户服务器架构、管道架构、分层架构等）和软件工程（主要是瀑布模型）的知识，基本上就能胜任大多数软件开发工作了。而今天，仅了解这些基础知识已经远远不够，因为在近 20 年内，人类创造了太多的新技术，而这些新技术又大都起源并服务于分布式计算领域。

1.1　分布式系统的组成

一个大型的分布式系统虽然非常复杂，但其设计目标却往往是非常简单的，例如，京东和淘宝这样的电商，其设计目标是卖东西；谷歌和百度这样的搜索引擎，其设计目标是帮助大家在网上找相关的内容；Facebook 和微信这样的社交应用，其设计目标是方便大家相互联系并分享自己生活中的点点滴滴。

如前文所述，之所以需要有分布式系统，最根本的原因还是单机的计算和存储能力不能满足系统的需要。但要把成百上千台计算机组织成一个有机的系统，绝非易事。在人类

社会中，其实也一样，找到1000个人容易，但要把这1000个人组织成一只能战斗的军队可就没那么简单了。

一个典型的分布式系统如图1-1所示。

- 分布式系统大都有一个Web前端，用户可以通过浏览器随时随地访问，当然，前端也可以是运行在Windows/Linux上的桌面程序或者运行在手机上的应用。
- 分布式系统还要有后端支撑。分布式系统的后端大都是基于Linux的集群[①]。之所以采用Linux，一是因为开源操作系统成本低，二是因为开源软件可以定制。
- 就像人类社会需要有一定的组织和管理一样，为了组成一个集群，在单机的操作系统之上，还需要集群管理系统。在集群管理系统中，一个非常重要的组件是分布式协调组件，用来协调不同机器之间的工作。这些协调系统大都基于一些著名的分布式一致性协议（如Paxos、Raft等）。有些超大型的后端还拥有专门的集群操作系统，这些系统不仅有分布式协调功能，还有资源的分配与管理功能。
- 为了满足大规模数据的存储需要[②]，需要有能够存储海量数据的后端存储系统。
- 为了满足大规模数据的计算需要[③]，还需要有能够分析海量数据的后端计算系统。

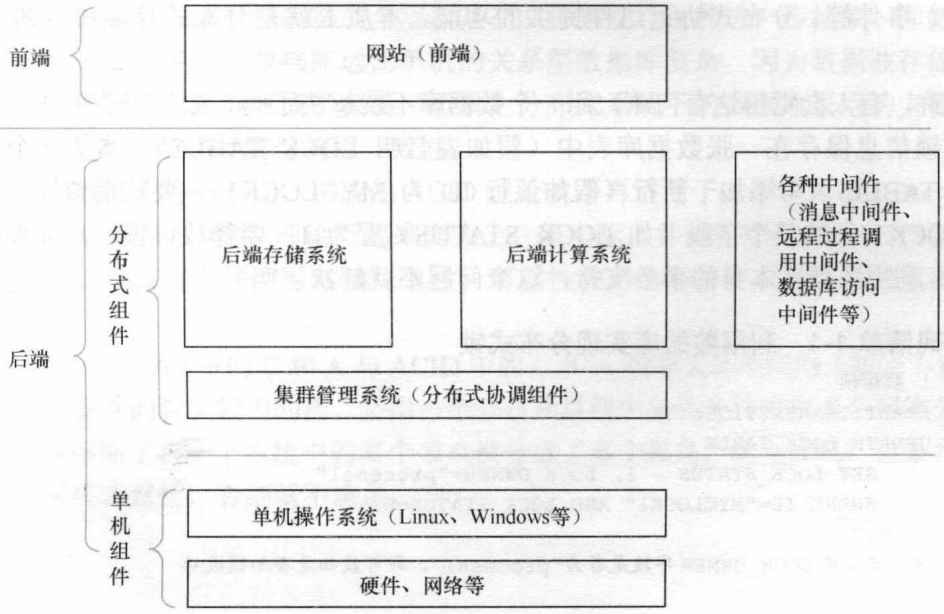


图1-1 一个典型的分布式系统

① 所谓集群，是指采用同样或类似配置的许多台机器，为了达到一个共同的目的而组成的系统。

② 像谷歌和百度这样的公司，因为索引的页面量非常庞大，需要很大的存储空间。微信和Facebook这样的社交应用亦然。

③ 例如谷歌，其PageRank算法就需要很大的计算量；再如京东和淘宝这样的电商，其商品推荐系统也需要很大的计算量。