Broadview
www.broadview.com.cn

MANNING

Mesos 实战

[美] Roger Ignazio 著
余何 陈秋浩 杨永帮 译

Mesos in Action

Florian Leibert 为本书作序

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Mesos实战

[美] Roger Ignazio 著

余何 陈秋浩 杨永帮 译

Mesos in Action

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

《Mesos 实战》为读者介绍 Apache Mesos 集群管理器及以应用程序为中心的基础架构概念。本书充满了有用的数据图表及实践指导，它将指引你迈出创建一个高可用的 Mesos 集群的第一步，接着在生产环境中部署应用程序，最后编写适合自己数据中心的“本地”Mesos framework（计算框架）。你将学习到如何对数千个节点进行弹性伸缩，同时通过 Linux 和 Docker 容器保证不同的进程间能实现资源隔离。你也将学习到如何使用热门主流的 framework 来部署应用程序的实践技术。

本书包含的主要内容有：搭建启动你的第一个 Mesos 集群；Mesos 的调度、资源管理及日志记录；使用 Marathon、Chronos 和 Aurora 部署容器化的应用程序；使用 Python 编写 Mesos framework。

阅读本书的读者需要熟悉数据中心管理的核心理念，也需要了解 Python 或者类似编程语言的基础知识。

版权贸易合同登记号 图字：01-2016-6365

图书在版编目（CIP）数据

Mesos实战 / (美) 罗杰·英格纳齐奥(Roger Ignazio) 著；余何，陈秋浩，杨永帮译. —北京：电子工业出版社，2017.5
书名原文：Mesos in Action
ISBN 978-7-121-31164-2

Ⅰ. ①M… Ⅱ. ①罗… ②余… ③陈… ④杨… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第060460号

策划编辑：符隆美
责任编辑：徐津平
特约编辑：顾慧芳
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：16.25　字数：364 千字
版　　次：2017 年 5 月第 1 版
印　　次：2017 年 5 月第 1 次印刷
定　　价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

推荐序一

世界著名的曼宁出版社（Manning）出版了不少广受欢迎的计算机丛书，如搜索领域的 *Lucene in Action*、*Elasticsearch in Action*，现在，他们又出版了这本云计算领域的 *Mesos in Action*。

Mesos 是一个开源的集群任务调度管理系统。现在随着分布式系统的广泛应用，越来越多的任务运行在集群上，而不是在单台服务器上。在 x86 PC 服务器集群上运行任务的好处是：单台服务器成本低，集群可以随着负载的增加添加服务器，水平扩展 scale out，而不是过去使用昂贵服务器的 scale up。随着集群规模的扩大，节点数越多，某个节点出现问题的概率就越大，当某个节点出现问题时，如何保证在这个节点上运行的任务能够顺利执行完成，成为一个技术难题。另外，如何管理集群，如何分发任务、监控任务执行过程等都是挑战。如果对于运行在集群上的任务，工程师还是需要在各台服务器上部署和管理，工作量将非常大，现在有些大规模集群的服务器数量已经超过万台。理想的情况是，工程师不需要关心集群里具体的每台服务器，而是把整个集群看成是一个计算、存储资源，把任务提交给集群的管理系统，由集群的管理系统去分发任务、监控任务执行，当某台服务器出现故障时，集群管理系统自动把任务派发到其他服务器上运行。这样的集群管理系统可以看作集群操作系统，甚至是数据中心操作系统。Google 在这方面做了大量的实践，在 2009 年发表的 *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale*

Machines 文章中，把整个数据中心看成一台计算机，所有的资源都由数据中心操作系统进行调度管理。

在 20 世纪 80 年代，学术界就开始了集群任务调度管理系统的研究，美国威斯康辛大学的研究人员开发了 Condor 系统，后来演进成开源的 HTCondor 系统。Google 在 1998 年成立之初，就使用 PC 服务器抓取、索引、检索全世界的网页，服务器数量巨大，他们先是开发了 WorkQueue 系统，也就是一个任务队列，工程师把需要集群运行的任务提交给这个任务队列，由任务队列把任务下发到集群的服务器，并监控任务的运行，如果服务器出现故障，就把任务重新下发到新的服务器上。后来，他们在 WorkQueue 的基础上开发了 Borg 系统，Borg 就是 Google 的集群任务调度管理系统。10 年前我在 Google 工作的时候，每天需要在集群上运行的任务，都是通过 Borg 来提交、管理的，非常方便。那时候，Google 的一个集群就已经有 2 万台服务器的规模，一个数据中心有 10 个这样的集群，20 万台服务器。我加入腾讯后，也开发了这样的系统。这种系统的资源隔离采用了容器技术（操作系统之上的资源隔离），而不是虚拟机（物理服务器之上的资源隔离），以实现更小的系统开销，更方便的管理。现在容器技术也正在被广泛使用，成为虚拟机之外的可选技术路线。

Mesos 也是一个这样的系统。Mesos 由著名的美国加州大学伯克利分校的 AMPLab 发明，AMPLab 是 Algorithms，Machines and People（算法、机器和人）实验室的缩写。AMPLab 的研究成果非常多，现在应用广泛的大数据处理框架 Spark 就是 AMPLab 的发明，AMPLab 的研究人员还包括了著名的美国科学院及工程院院士、发明了 RAID（磁盘阵列）和 RISC（简约指令架构）的 David Patterson 教授。Mesos 系统也在 Twitter、Airbnb 和苹果公司得到应用。

本书译者之一余何是 IT 专家，他曾在平安科技工作多年，有着丰富的大规模集群系统的开发、运维、管理经验，经历了多个云计算、大数据系统在金融行业的应用，在 2015 年出版了《PaaS 实现与运维管理》一书，由他作为经验丰富的实战者来翻译这本书是最合适不过的了。

相信这本书会为对大规模分布式系统、集群任务调度管理、云计算和大数据感兴趣的读者带来受益。

陈军[1]

日志易 CEO

1 陈军先生拥有18年IT及互联网研发管理经验，曾就职于Cisco、Google、腾讯和高德软件，历任高级软件工程师、专家工程师、技术总监、技术副总裁等职务，负责过Cisco路由器研发、Google数据中心系统及搜索系统研发、腾讯数据中心系统和集群任务调度系统研发、高德软件云平台系统研发及管理，对数据中心自动化运维和监控、云计算、搜索、大数据和日志分析具有丰富的经验。他拥有美国南加州大学计算机硕士学位，发明了4项计算机网络及分布式系统的美国专利。

推荐序二

余何兄是我的老朋友了，在运维领域耕耘多年，不仅运维能力强，而且善于总结乐于分享，也是 GOPS 全球运维大会金牌讲师。我和余何兄最开始是在圈子中互相关注，之后在一次论坛上偶遇，一番交流下来，一拍即合，一同走上了追求快乐运维的道路。我们的共识是“一起愉快地玩耍”，让运维变得更加轻松，让运维人员更加健康地生活。

但凡一个好的产品，都是从 0 到 1，而不是从 0 到无穷、从 0 到包罗万象，这其中的道理就是专注。专注于做好一件事，提供稳定兼容的接口，这就是一个好产品的伊始。在运维平台领域，由于其需求范围广、组织差异大，很难有那么一个产品能够满足一切，而短时间满足一切又很可能意味着 bug 多多，因此，我们应该问自己，你到底需要什么？

刚好，Mesos 就是这样一个产品：专注于做好一件事，专注于资源管理。关于其上的应用任务调度，无论是服务、批处理还是大数据，它都提供了稳定而兼容的接口，从而让用户在其上按照自己的需求，不断迭代实现，最后形成自己的产品。其在提高集群资源利用率，服务自动化部署等方面的表现，尤其令人称赞。

希望余何兄组织翻译的本书，能得到您的喜爱，为 Mesos 在中国的继续壮大添砖加瓦。为了我们共同的运维事业，一起加油！

萧田国

高效运维社区发起人开放运维联盟主席

国内第一个 DevOps Master

推荐序三

我一直在关注国内 IT 运维的发展，在不同的业务领域、企业规模下，运维的标准与规则参差不齐，真正掌握运维真知的人绝不仅意味着玩转创新技术、流行框架，更重要的是如何真正解决企业问题，如何保证落地与实现。国内第一本 PaaS 原创畅销书作者“众生的大师兄”余何有着十多年运维实战经验，经历了中国运维发展的各个阶段，在电信、金融以及物流领域都有所耕耘，今天由他领衔翻译的《Mesos 实战》同样保持了高水准，相信一定会广受欢迎。

Mesos 并没有什么吸引眼球的华丽外表，在起步阶段由于具备一定门槛，同时效能只能在具备一定规模后才能体现，因此一直没有快速地在运维领域流行起来。倒是 OpenStack 借助云计算概念、K8s 背靠谷歌这个亲爹在社区内刮起了一股大的旋风，回头再来看看近几年商业化的过度炒作，连一向低调的运维领域也产生了泡沫。回到问题的本身，为什么要使用 Mesos，我会站在和译者一样的角度考虑，我们必须考虑绕不过的环节，上层的应用，我们需要对遗留应用架构系统进行兼容吗？需要。Mesos 是一个通用性资源管理框架，能够适配各种计算类型的服务，正因为如此，向上的任务调度才稍显复杂，为了做好兼容，我们必须做更多工作。

资源管理策略 Dominant Resource Fairness(DRF) 是 Mesos 的核心，是将 Mesos 比作分布式系统 Kernel 的根本所在。Mesos 能够保证集群内所有用户都能够平等地使用集群内资源，这里的资源包括 CPU、内存、磁盘等。在通用性方面，Mesos

只负责提供资源给上层任务调度 framework，而不负责具体任务管理，于是可让各种类型计算任务使用集群资源。关于准入门槛，如果仅部署一套 Mesos，我们几乎什么也干不了，为了使用好它，我们需要不同的 Mesos framework，像 Marathon，chronos 等，在特殊场景下，甚至需要开发自己的 framework，除此之外磁盘、网络等问题也需要着重考虑。如果没有强大的意志力，初学者大多会望而却步。还好有“众生的大师兄”这样一位谆谆不倦的运维发展践行者，带着对运维未来美好发展的憧憬，坚持不懈的推行运维理念与实践，我也希望本书能够帮助 Mesos 在运维社区中快速流行与成熟，大家都能有所贡献与分享，真正解决我们企业内部遇到的运维问题。

肖力
云技术社区创始人

译者序

云计算时代，对开源产品的选择，很容易陷入一个误区，用商业化方式进行判断，选择最“热门畅销”的产品，而忽视了自身需求及组织能力，从而导致目标的偏离。2014 年，我与小伙伴们开始考察一系列平台产品并进行研究，当时我们的需求很明确，提升资源利用率与运维效率，没有其他（专注，专注，再专注）。我们的组织能力也很明确，深入理解操作系统，能够像使用黑魔法一样改变操作系统的行为，具有工程设计能力，能很快地驯服与改造各种开源产品。基于以上两点，经历了一段考察期后，我们最终选择了 Mesos。

Mesos 很难让人一下子就亲近，她绝不是让你一见钟情的那种，她没有华丽外表让你如痴如醉，也没有什么直接功能让你耍酷摆炫，你需要花很长一段时间去理解她，你需要有应用场景，有资源节约型需求，有大规模机器集群管理，有多种计算类型，之后在不断的运用中，才慢慢发现 Mesos 的奥妙之处，一步步坠入到她的爱河中。Mesos 专注于数据中心资源管理，专注于做好一件事，干净、简洁，如同操作系统内核一样，它是数据中心资源管理的 Kernel。

很多人问我 Mesos 是否可以解决运维领域的可靠性问题、资源管理问题，从此天下太平、安枕无忧，作者在书中也有一点这种态度。但很遗憾，实际运维环境是相当复杂的，这种神一样的工具几乎不可能存在。不同的管理结构有着不同的权限层级，不同的应用类型有着不同的服务级别，不同的组织环境有着不同的流程规范，

一言以蔽之，没有什么固定的工具可以解决不断变化的运维问题。

要保持运维水平的不断提升，最后发现这是一个组织行为学问题，是理念、人、流程、工具的结合体。组织的理念是什么？人的专业要求有哪些？对人的投入有哪些？组织理念文化又是如何影响着人，匹配的流程与工具又是什么？是这一切决定了最终的运维水平。精益思想对运维理念是一种启示，工程师文化强调了所需要的人才，ITIL 是运维流程上的最佳实践，工具则是连接人与流程的桥梁。Mesos 能否在企业发挥最大效应，看的不是 Mesos，而是组织自身。

余何

2016.11 于深圳软件产业基地

序

如果你想看到一个人抓狂的样子，你可以走到一个在数据中心手动配置并供应数十台服务器的人面前，然后说道："哇！持续追踪在那些机器运行了什么东西，肯定非常容易，也非常好玩。"

或者找一个身上常年带着传呼机以便响应服务器中断的人，并说："这听起来像一份毫无压力的工作呀。至少它保证你夜晚能睡个好觉。"

当然，事实上，管理服务器和其他数据中心基础架构一直以来都很困难和沉闷，给负责配置这些机器并响应机器故障的可怜男女带来了无数个不眠之夜。由于过去20年来公司越来越依赖于信息技术，经常会在每个服务器（或近些年的虚拟机）上配备一个应用程序，因此实际操作变得越来越困难。服务器数量动辄从一位数升级到两位数，有时甚至上升到三位数。

接着由Google、Facebook和Twitter等热门服务助燃的互联网呈爆炸式增长。由数以十亿计的智能手机、平板和其他设备助燃的移动互联网也立刻随后跟上。在任何特定时间里，数以百万计的用户可能同时在一个网站或APP中，而旧式计算技术无法再切入这样的世界。

在数据中心内，单一服务器的数据库（甚至所有单一服务器的服务）很快被分布式系统取代，其能以之前无法想象的容量来处理数据和流量。复杂庞大的应用程序也经常被微服务取代——把多组单一用途的服务分开管理，接着通过API进行连

接，最终构造成用户端的应用。虽然伸缩性提升了，但构建这些系统的学习曲线和管理系统复杂性也都随之提升。

Google 有个极好的方法，即在它自己的数据中心内使用一个叫 Borg 的系统来解决这个问题，表面上让大多数员工——如系统管理员和开发者——像管理一台大计算机的方式来管理数以万计的服务器。在 Borg 简化 Google 的操作几年后，开源的 Apache Mesos 项目粉墨登场并以相似的方式改变了其用户的生活。忽然之间，部署、运行和管理复杂分布式系统的过程变得非常简单。所有东西共享同样的机器组，而 Mesos 只需要轻松处理跑腿活儿——以可用的资源来匹配工作负载的需求。

最初我是作为一名 Twitter 的软件工程师来亲身体验这第一手转变的，在那里 Mesos 帮助攻克难堪的“失败之鲸”（fail whale，Twitter 服务宕机标志）并帮助 Twitter 达到伸缩性和可靠性的新高度。当 2012 年我到 Airbnb 工作时，它还只是一家建立了 4 年的创业公司，Mesos 又一次帮助我们的基础架构随着用户基数扩大成长——但是并不包括它的复杂性。Mesos 和其前景深深触动了我，于是我决定建立 Mesosphere 这个公司，致力于让 Mesos 为主流企业所用。

随着 Mesos 趋于热门和 Mesosphere 的扩张，我们把目标设为雇用最好的 Mesos 工程师和从业者。当我们看到 Roger Ingnazio 在 Pupet 实验室建立一个基于 Mesos 的持续集成平台的工作成果时，我们知道我们必须拥有它。在知名公司运行可伸缩的生产系统是非常宝贵的经验，而且自从加入 Mesosphere 后，Roger 的经验给我们基于 Mesos 的数据中心操作系统技术和我们的用户体验带来很多帮助。

Roger 的《Mesos 实战》让对 Mesos 和其技术生态系统感兴趣的人皆受益于他的经验。这本书是运行 Mesos 集群和安装你的第一个 framework（计算框架）的极佳指导，它同时也探索了更高级的主题，例如掌握强大的 Mesos framework（包括容器编排用的 Marathon 和大数据分析用的 Spark），甚至包括构建你自己的 framework。

无论你是在准备部署 Mesos，还是你已经运行了它，并想进一步提升你的知识，你都难以找到一个比 Roger 更好的导师，或者一本比《Mesos 实战》更好的书。

FLORIAN LEIBERT
Mesosphere 联合创始人、CEO

自序

Benjamin Hindman 主导的团队于 2009 年在加州大学伯克利分校创立了 Apache Mesos 这个研究项目。Ben 和他的团队想要通过允许多个应用共享一个单一的计算集群来提升数据中心的效率，就像多个应用可以在你的笔记本或工作站共享处理器、内存和硬盘驱动器一样。但他们想要在由许多服务器组成的现代数据中心上实现这个想法。经过 10000 行 C++ 代码的最初实现后，他们在 2010 年发布了一篇论文《Mesos：一个数据中心内部细粒度资源共享平台》（*Mesos:A Platform for Fine-Grained Resource Sharing in the Data Center*）。

没过多久，Ben 加入 Twitter 并使用 Mesos 来更好地扩展其基础架构，终结了 Twitter 难堪的“fail whale”时期——成名后 Twitter 的服务器处理能力跟不上用户的需求增长。虽然 Twitter 没有公开披露其庞大基础架构中的服务器数量，但根据其展示的在线资源和第一手资料，这个数量大约在每个集群有 10000 个 Mesos 节点。

2010 年 12 月，Mesos 项目进入 Apache 孵化器，作为 Apache 软件基金会的分支，使得项目获得 ASF 付出的全面支持。Apache Mesos 项目于 2013 年 6 月从孵化器中得以完善，现在已经是一个高等级的项目。

2013 年，Ben 和 Florian Leibert 及 Tobi Knaup 一起创立了 Mesosphere 公司。Mesosphere 的旗舰产品，数据中心操作系统（DCOS），通过为那些想要像 Airbnb、Apple 和 Netflix 一样轻易地使用 Mesos 来部署应用和扩展基础架构的企业提供全方

位解决方案，将开源项目成功地商业化。Mesos 持续成为开源 Mesos 项目的主要贡献者，并为开源社区提供 Mesos 安装包和工具。

我对于 Mesos 生态系统和大规模基础架构的初次涉足始于 2014 年，那时我开始想要使用 Mesos 在多个 Jenkins（热门的持续集成框架）实例之间共享资源。其时，Mesos 看起来像是留给那些已经知晓它的人去使用的，因为虽然那时有大量可用的线上资源，但是很难被找到，而且没有一个权威可信的来源。彼时也没有任何书籍涉及 Mesos。我写了几篇关于我的工作经验的博客，其他人看起来好像跟我一样：对这个项目都想了解多一些，但不知道从何入手。

2015 年 1 月，Manning 出版社找到我并问我是否有兴趣写一本关于 Mesos 的书。我之前从未写过书，这个请求一开始让我有点不知所措。但我也把它当作一个好机会来写一本工具书，因为当我刚开始使用 Mesos 时，我确实希望拥有这么一本书来指导我。幸运的是，Manning 的团队给我这份自由来实现这个想法。

希望你在使用《Mesos 实战》的过程中，能发现它是一个部署和管理 Mesos 集群和提升你基础架构整体效率的宝贵资源，而且它能够帮助你的团队或你的客户更快捷便利地部署应用到生产中。

Roger Ignazio
于美国俄勒冈州，波特兰

鸣谢

你正在阅读的是一本历经一年努力而产出的，关于 Apache Mesos 项目和生态系统的深度书籍。尽管我的名字在封面上，但还有许多为此书最终出版贡献力量的人。如果我不在这里感谢他们的话，那他们将继续匿名，无人知晓。我确定我的家人、朋友和我的妻子早已知晓在这次努力中，我对他们给予的支持怀有无比感激之情。

首先，我想要感谢 Mesos 社区。在每次交流中，无论是会议上、邮件列表上还是 IRC 上，每个人都给予了我极大帮助和善意。在撰写本书时，对 Mesos 代码库有超过 100 个贡献者，甚至有更多的志愿者在 Mesos 邮件列表和聊天室中回答问题和提供帮助。除了我要感谢有幸在会议中一起探讨和日常中一起工作的所有人，我还要感谢 Ben Hindman，Florian Leibert，Thomas Rampelberg，Dave Lester，Christian Bogeberg 和 Michael Hausenblas 他们提供的所有帮助。另外我还要感谢 Florian 为这本书写的前言。

接着，我在写作过程中改变了大约三分之二的工作方式，让写作此书仿佛不是一个有压力和花费时间的任务。在 Puppet 实验室，我要感谢 Scott Schneider，Colin Creeden，Cody Herriges，Eric Zounes 和 Alanna Brown 他们的支持。在我与 Manning 签约之前，我回想那时候当 Scott 问我是否真的认为自己能够写一整本关于 Mesos 的书。结果就是，真的可以！

很多人在幕后帮忙检查这本书的不同阶段并提供反馈，包括 Al Rahimi，Clive

Harber，John Guthrie，Luis Moux Dominguez，Mohsen Mostafar Jokar，Morgan Nelson，Nitin Gode，Odysseas Pentakalos 和 Thomas Peklak。特别鸣谢 Jerry Kuch 和 Chris Schaefer 的技术审阅及文字编辑 Sharon Wilkey 对原稿数不尽的修改。

最后，但肯定不止这些，我需要感谢我在 Manning 出版社的了不起的团队。我的编辑，Mike Stevens，帮助我从一开始“听起来像一堆胡言乱语”到获得一个正式的计划和签好的合同。开发编辑 Cynthia Kane 确保我总是提供正确数量的上下文（文字和图表），并帮助我成为一个更好的写作者和沟通者。最后感谢我的出版者 Marjan Bace，他不仅在编辑审阅时帮忙塑造这本书，并且完全给予了我自由去写这本书，这是我一开始使用 Mesos 梦想拥有的一本工具书。感谢你！

我对所有帮助我写成此书的人深深感激！若有遗漏未致谢，我在此致歉。

关于本书

《Mesos 实战》是一本在现实环境中学习和部署 Apache Mesos 的实用指南。在书中我将带你巡游整个项目——从对 Mesos 和容器的基础介绍，到满足生产发布的包含高可用和 framework（计算框架）认证的应用部署。我也将介绍热门（和开源）的 Mesos 应用的实际用法，来帮助你在 Mesos 集群里部署应用程序和计划作业。

尽管《Mesos 实战》是专门为中高级的系统管理员定制的书，它也适用于不同的读者。我写这本书时，尽量让系统管理员、DevOps 人员、应用管理员及软件工程师等类似从业者在通读全文时都能感到自在。尽管有些应用部署和软件开发的知识很吸引人，但我只会提供足够的，且非严格要求的背景资料。我会选择教你新的技巧来帮你自己的团队更聪明、而非更辛苦地工作。

路线图

如果你是一名想要部署首个 Mesos 集群的系统管理员或者 DevOps 人员，你得特别关注第 1 章到第 8 章。这些章节涵盖了你需要知道的安装和运行集群的所有事情，也涵盖了一些部署应用程序和计划作业的方法。第 10 章也能帮你了解如何编写自有的 Mesos 可用的应用程序。另外，本书分为三个部分。

第 1 部分介绍了 Apache Mesos 项目，比较了容器和虚拟机的区别，并且展示了部署 Mesos 集群的实际用例。