三位Google资深数据中心专家精心编写
数据中心系统结构学科的“圣经”

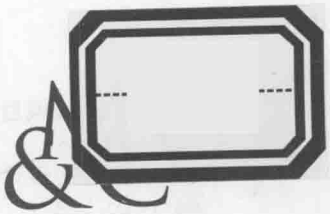
数据中心设计与运营实战

【美】Luiz André Barroso 【美】Jimmy Clidaras 【瑞士】Urs Hölzle 著
陈实 李典林 李艳 译 王海峰 曲海峰 审校

The Datacenter as a Computer

An Introduction to the Design of Warehouse-Scale Machines
Second Edition

人民邮电出版社
POSTS & TELECOM PRESS



数据中心设计与运营实战

【美】Luiz André Barroso 【美】Jimmy Clidaras 【瑞士】Urs Hölzle 著
陈实 李典林 李艳 译　　王海峰 曲海峰 审校

The Datacenter as a Computer

An Introduction to the Design of Warehouse-Scale Machines
Second Edition

人民邮电出版社
北京

图书在版编目（CIP）数据

数据中心设计与运营实战 /（美）巴罗索（Barroso, L. A.），（美）克利德瑞斯（Clidaras, J.），（瑞士）霍泽尔（Holzle, U.）著；陈实，李典林，李艳译. -- 北京：人民邮电出版社，2014.11
ISBN 978-7-115-36806-5

Ⅰ. ①数… Ⅱ. ①巴… ②克… ③霍… ④陈… ⑤李… ⑥李… Ⅲ. ①机房－建设 Ⅳ. ①TP308

中国版本图书馆CIP数据核字(2014)第215389号

◆ 著 [美] Luiz André Barroso Jimmy Clidaras
[瑞士] Urs Hölzle
译 陈 实 李典林 李 艳
审 校 王海峰 曲海峰
责任编辑 赵 轩
责任印制 彭志环 杨林杰

◆ 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
北京铭成印刷有限公司印刷

◆ 开本：720×960 1/16
印张：11.25
字数：139 千字 2014 年 11 月第 1 版
印数：1－3 500 册 2014 年 11 月北京第 1 次印刷
著作权合同登记号 图字：01-2014-2390 号

定价：49.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

运行大规模服务所需的计算平台已经不再是十多年前的那种比萨饼盒大小的服务器或者冰箱大小的高端多处理器系统了。这样一个平台的硬件是由成千上万的独立计算节点和与之对应的网络和存储子系统、配电、空调设备和巨大的制冷系统组成。这些系统所在的建筑也已经成为系统的一部分，和一个大型仓库没有什么区别。

本书的中心思想很简单：这一计算平台不能简单地看作一堆放在一起的机器。数据中心的软硬件资源必须协同工作，以提供高水平的互联网服务，而高水平的互联网服务只能通过整体的设计和部署来实现。换句话说，我们必须把数据中心本身看作是一台庞大的计算机。

计算正在移动到云端，进入仓储式数据中心（WSC，Warehouse-Scale Computer），软件和硬件架构师必须了解端到端的整个系统才能设计良好的解决方案。我们不再设计单独的“比萨饼盒”，或单服务器应用程序，我们也不能忽视装满服务器的一个大仓库背后的物理和经济机制。但事实上，建立一个具有成本效益且可靠的巨型规模的计算平台，并使其满足下一代云计算工作负载的可编程性要求是非常困难的。本书将帮助读者去了解新的领域，在未来攻克更多难题。

序

传统的数据中心设计往往只关注风火水电，考虑地理位置、建筑面积、供多大电量、冗余做到多少级，等等，很少关注承载的 IT 设备的特性，更不用说上层具体的业务应用。因此，很多的数据中心从投入生产那天起就落伍了，要么是跟不上业务的发展节奏，不得不进行很多的生产中的改造优化等工作；要么是做了过多的可靠性考虑，系统富余非常多的容量，投产后实际只能利用设计的几分之一，甚至更少。互联网数据中心由于支撑的计算规模和存储规模较传统企业的数据中心更大，甚至超过几个数量级，如果只是沿用过去的数据中心设计方法和思路，在效率和同业务的适配层面会表现得更差。腾讯公司在一开始都是租用电信运营商采用传统思路和方法设计的数据中心来托管自己的业务服务器，发现有很多的问题：机架供给慢，扩展能力差；能源效率低，一个机架只能放几台服务器，大量的空间被浪费；多个层级、多个模块的冗余设计，很少被用到；规模小，一个物理的机房连一个业务的一个功能模块有时都不能放下，等等。后来不得不自己设计、建设了一些数据中心，以使得同业务的发展和架构有更好的匹配。

本书总结了过去几年互联网数据中心的设计、建设与运行实践，特别是 Google 公司这家拥有世界上最大数据中心规模的互联网公司的经验教训，提出了把数据中心当成一台仓储式计算机（WSC）来规划设计，就像以前 IBM 设计大型计算机一样来考虑数据中心的各个模块，

不仅仅是最底层的供电、制冷、结构布局等，也包括承载的计算系统、存储系统以及互连通信需要的网络连接，更重要的是还会考虑数据中心里各个软的东西：系统层软件、平台型软件，以及最上层的业务应用软件，等等。垂直一体化设计，各模块组件水平对齐耦合，模块组件极致创新等一系列新的实践方法，使得数据中心从整体上做到端到端系统化地与互联网行业承载的服务保持最佳的适配，在能源效率以及总体效率方面大大优于传统的做法。

数据中心作为基础架构的主要载体，创新的设计思路、方法将在降低投资成本、运行成本以及提升效率等方面带来很大的好处。本书的多位译者在互联网行业、数据中心行业都有着丰富的理论与多年实践经验，原著内容案例丰富，方法可操作性强，翻译也非常恰当。相信此书的翻译出版对于中国互联网行业的架构设计、数据中心行业的建设发展都会有极大的推动作用。本书对于有志于从事这类工作的专家学者、工程人员也能起到非常大的借鉴作用。

侯金刚

腾讯网络平台部副总经理，总架构师

致谢

过去几年我们直接参与了 Google 基础设施的设计和运营，据此在这里的总结、汇报是整个 Google 团队同事艰苦工作、洞察和创造的结果。我们基础架构技术团队的职责范围涵盖了本书的所有内容，在此对他们的分享表示特别感谢。感谢 Google 的 Kristin Berdan 和 Morgan & Claypool 出版社的 Michael Morgan 提供的帮助。Gerry Kane 在技术写作方面的才华，使得本书质量大幅提高。我们对 Catherine Warner 在各个不同阶段的校阅和修订工作也表示感谢。

感谢 Mark Hill 和 Michael Morgan 邀请我们参与这个项目。他们不断地给予我们鼓励和督促，同时保持着无限的耐心。

本书受益于 David Andersen、Partha Ranganathan 和 Christos Kozyrakis 的仔细审阅，以及 Tor Aamodt、Dilip Agrawal、Remzi Arpaci-Dusseau、Mike Bennett、Liqun Chen、Xiaobo Fan、David Guild、Matthew Harris、Mark Hennecke、Mark Hill、Thomas Olavson、Jack Palevich、Pete Pellerzi、John Reese、Ankit Somani 和 Amin Vahdat 的更正和补充。

衷心感谢你们的帮助。

我们同样感谢 Vijay Rao、Robert Hundt、Mike Marty、David Konerding、Jeremy Dion、Juan Vargas、Artur Klauser、Pedro Reviriego Vasallo、Amund Tveit、Xiau Yu、Bartosz Prybylski、Laurie Doyle、Marcus

Fontoura、Steve Jenkin、Evan Jones、Chuck Newman、Taro Tokuhiro、Jordi Torres 和 Christian Belady 对第一版的反馈和更正。

Ricardo Bianchini、Fred Chong、Jeff Dean 和 Mark Hill 在第一版不成熟的初稿阶段提供了极其有用的反馈。感谢你们。

目录

第 1 章
介绍

ARPANET 问世已经超过 40 年了，World Wide Web 最近刚刚庆祝完其走过了 20 周年。被这两个有杰出意义的里程碑所引领的互联网技术继续改变着各行各业和现今人们的生活习惯，时至今日依然势头不减。诸如网页邮箱、搜索、社交网络等流行互联网服务的出现，加之高速互联网络在世界各地的普及，使互联网服务日渐呈现出向服务器端以及云端转移的趋势。

越来越多的计算和存储需求开始从类 PC 客户端向更小、更适合移动设备，并结合了大型互联网服务的方向迁移。早期的互联网服务大多是用来提供资讯，而今许多 Web 应用提供了以前客户端承载的服务，例如电子邮件、照片、视频存储和办公应用。驱动这种计算向服务器端转移的不仅是提升用户体验的需求，诸如无需配置或备份的便捷管理和无缝接入，软件供应商自身可以从中受益也是一个重要的驱动力。软件即服务允许更快的应用开发节奏，因为它使得供应商可以更快捷地改变和提升软件。供应商无需维护拥有特定硬件和软件配置的数以百万计的客户端，他们只需在自己的数据中心里就可以完成协同改进和修复，并且能够让他们的硬件以最佳配置部署。

此外，数据中心的经济性使许多应用服务降低了单用户成本。例如，服务器可能会为成千上万的活跃用户和更多不活跃用户提供服务共

享。同样的，计算自身也可以通过共享服务来降低成本，例如，对于一个收件人为多个用户的电子邮件附件，仅需存储一次，而不是多次。最后，放置在数据中心的服务器和存储设备比同等规模的台式机或笔记本电脑更容易管理，因为它们由单一组织进行管理。

有些工作负载需要强大的计算能力，大规模计算集群显然比客户端计算更适合用于这种情况。搜索服务（网页、图片等）是此类工作负载的最好案例。但是对于诸如语言翻译类应用而言，大规模集群计算依然更有效率，因为翻译依赖于对大规模语言模型的分析。

计算向服务器端转移的趋势和互联网服务的爆炸式流行创造了一类新的计算系统，我们将其命名为仓储式数据中心，即 WSC（Warehouse-Scale Computer）。这样命名是为了突出这些机器最显著的特点：拥有适应大规模基础架构的软件、数据仓库和硬件平台。这种系统使人们对计算技术沿袭多年的“单一程序运行在单一机器上”的这一认知成为历史。在 WSC 中，程序被定义为一个可能包括由数十个甚至更多独立程序交互实现的复杂用户服务，诸如电子邮件、搜索和地图。这些独立程序可能由不同的甚至跨越组织、地域和公司的工程师团队部署和维护，例如 Mashups（利用外部数据源检索到的内容来创建全新的服务的工具）。

运行大规模服务所需的计算平台已经不再是十多年前的那种一个比萨饼盒大小的服务器或者冰箱大小的高端多处理器系统了。这样一个平台的硬件由成千上万的独立计算节点，和与之对应的网络和存储子系统、配电、空调设备和巨大的冷却系统组成。这些系统所在的建筑也已经成为系统的一部分，和一个大型仓库没有什么区别。

1.1 仓储式数据中心

这些系统的显著特点在于规模，我们可以简单地称之为数据中心。

数据中心是部署了许多服务器和通信设备的专用建筑物，因为这些服务器和通信设备具有相同的环境和物理安全要求，并且需要易于维护。从这个意义上讲，WSC 是数据中心的一种类型。然而，传统数据中心通常在主机上大量运行着相对小型或中型的应用，每一个程序运行在一个专用的硬件基础设施上且高度耦合，并且在相同基础设施中进行隔离保护。这些数据中心为不同组织和公司提供硬件和软件服务，存在于这种数据中心里的不同计算系统在硬件、软件，或维护上几乎没有相同之处，而且彼此之间趋向于没有通信。

为诸如 Google、Amazon、Facebook 和 Microsoft 的在线服务部门提供服务的 WSC 数据中心，明显区别于传统数据中心：它们属于一个组织，使用互相兼容的硬件和系统软件平台，共享一个系统管理层。通常，相比采用第三方软件运行的传统数据中心，大多数应用、中间件和系统软件都是组织内部编写的。更重要的是，WSC 运行着数量少但规模大的应用（或者互联网服务），且通用的资源管理基础架构带来了巨大的部署灵活性[1]。同质性的要求，单一组织控制和对成本有效性的增长的关注都激励着设计师们采取新的方法来建设和运营这些系统。

互联网服务必须做到高可用，典型目标是至少 99.99%的正常运行时间（大约每年有一小时停机时间）。实现在大量软硬件和系统软件上无故障运行是相当困难的，而引入大量服务器将使其变得更加困难，虽然理论上在 10000 台服务器的集合中防止硬件故障是可能的，但成本极高。因此，WSC 必须被设计成能够进行大量组件容错，使之极少甚至不影响服务级别的性能和可用性。

1 公有云，如 AWS、Google Cloud Engine，会和普通的数据中心一样，运行很多小的应用。然而在供应商看来，所有这些应用都是同一个 VM，它们需要大量共同的服务，如块、数据库存储和负载均衡等。

1.2 规模成本效率

建设和运营如此巨大的计算平台成本极高，并且提供服务的质量很大程度上依赖于处理和存储能力，这将导致成本上升，因此需要重点关注成本效率。例如，对信息检索领域的网页搜索而言，如下三个主要因素驱动着计算需求的增长。

- 不断增长的服务普及转换成更高的负载需求。
- 问题规模持续增长——互联网每天有数以百万计的新网页产生，这使得建设和服务于网页索引的成本不断增加。
- 即使吞吐量和数据存储库可以保持不变，市场的竞争本质也会不断推动技术创新，以提升检索结果的质量和索引被更新的频率。虽然有些质量改进可以通过更智能的算法来实现，但最实质性的改进往往伴随着每个请求需要额外的计算资源。例如，在一个搜索系统考虑了被搜索词的同义词查询或语义关系后，产生检索结果将有更高的成本——或许是搜索需要查询那些匹配更复杂的查询语句包含了同义词的文档，亦或一个术语的同义词需要被复制到对应术语的索引数据结构中。

对更多计算能力的索求使得成本经济性成为 WSC 设计中的主要要素。成本经济性必须广泛涵盖成本的主要组成，包括托管成本和运营开销（包括了电力供应和能源开销）、硬件、软件、管理和维修成本。

1.3 不只是服务器的集合

我们核心的观点是，驱动现今很多成功的互联网服务的数据中心已经不再是将五花八门的服务器放在一起，连上网线这么简单的了。运行在这些系统上的软件，诸如 Gmail 和 Web 搜索服务，是使用着超越