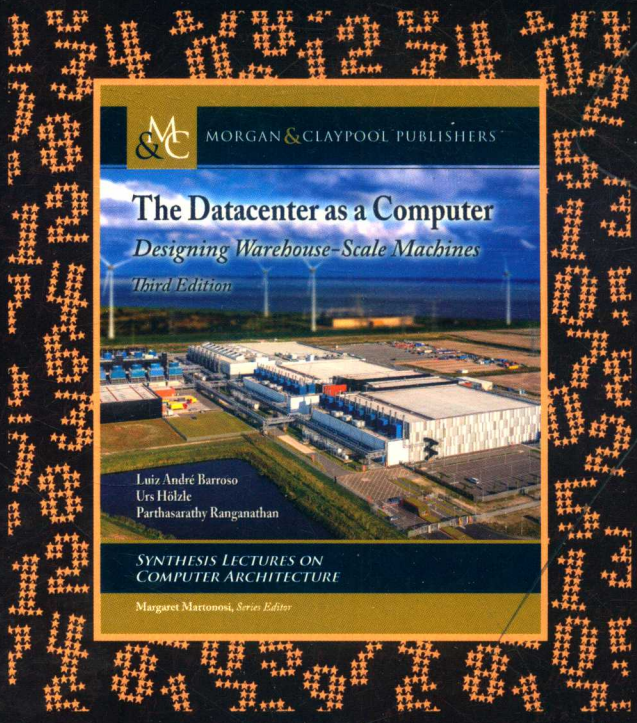
HZ BOOKS
华章IT

数据科学与工程技术丛书

数据中心一体化最佳实践
设计仓储级计算机
（原书第3版）

路易斯·安德烈·巴罗索（Luiz André Barroso）
[美] 乌尔斯·霍尔兹勒（Urs Hölzle） 著
帕塔萨拉蒂·兰加纳坦（Parthasarathy Ranganathan）
徐凌杰 译



THE DATACENTER AS A COMPUTER
DESIGNING WAREHOUSE-SCALE MACHINES
(THIRD EDITION)

机械工业出版社
China Machine Press

数据科学与工程技术丛书

THE DATACENTER AS A COMPUTER

DESIGNING WAREHOUSE-SCALE MACHINES

(THIRD EDITION)

数据中心一体化最佳实践

设计仓储级计算机

（原书第3版）

路易斯·安德烈·巴罗索（Luiz André Barroso）
[美] 乌尔斯·霍尔兹勒（Urs Hölzle） 著
帕塔萨拉蒂·兰加纳坦（Parthasarathy Ranganathan）
徐凌杰 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据中心一体化最佳实践：设计仓储级计算机（原书第3版）/（美）路易斯·安德烈·巴罗索（Luiz André Barroso）等著；徐凌杰译．—北京：机械工业出版社，2020.1
（数据科学与工程技术丛书）
书名原文：The Datacenter as a Computer: Designing Warehouse-Scale Machines (Third Edition)

ISBN 978-7-111-64486-6

Ⅰ.数… Ⅱ.①路… ②徐… Ⅲ.计算机中心－一体化 Ⅳ.TP308

中国版本图书馆CIP数据核字（2019）第284413号

本书版权登记号：图字 01-2019-2169

全书共分8章。第1章概述仓储级计算机及其架构；第2章介绍WSC中运行的应用，以及包括平台层软件、集群层基础软件、监控和管理软件在内的系统基础设施技术栈；第3章涵盖关键的硬件组件，重点讨论服务器和加速器组件、存储架构和数据中心网络设计，以及计算、存储和网络之间的相互作用；第4章着重点关注数据中心电力、冷却基础设施和建筑的设计；第5章讨论能耗和能效相关的话题；第6章讲解如何对WSC的TCO进行建模；第7章讨论正常运行时间和可用性；第8章总结历史趋势并展望未来。

本书适合对仓储级计算机系统感兴趣的架构师和程序开发人员阅读，也适合仅想了解互联网基础设施信息的读者阅读。

数据中心一体化最佳实践
设计仓储级计算机（原书第3版）

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：关 敏
责任校对：殷 虹
印 刷：北京瑞德印刷有限公司
版 次：2020年1月第1版第1次印刷
开 本：185mm×260mm 1/16
印 张：11
书 号：ISBN 978-7-111-64486-6
定 价：79.00元

客服电话：（010）88361066 88379833 68326294
投稿热线：（010）88379604
华章网站：www.hzbook.com
读者信箱：hzit@hzbook.com

本书介绍了仓储级计算机（WSC）的设计。这类计算平台是云计算的核心，支撑着我们每天都在使用的各种强大的互联网服务。书中讨论了此类新型系统如何将数据中心本身当作一台超大规模的仓储级计算机来使用，同时又能使软硬件充分协同并提供高性能的互联网服务。每一章都涵盖多个真实世界的案例，其中包括详尽的谷歌在线服务的基础设施信息。

全书共分8章。第1章为绪论，概述仓储级计算机及其架构；第2章首先总体介绍WSC中运行的应用，以及包括平台层软件、集群层基础软件、监控和管理软件在内的系统基础设施技术栈；第3章涵盖关键的硬件组件，重点讨论服务器和加速器组件、存储架构和数据中心网络设计，以及计算、存储和网络之间的相互作用；第4章着眼于更底层的系统设计，重点关注数据中心电力、冷却基础设施和建筑的设计；第5章讨论能耗和能效相关的话题，包括稳定测定能效的挑战、衡量数据中心能效的电力使用效率以及电力超额配置的设计和好处；第6章讲解如何对WSC的TCO进行建模，其中包括资本支出和运营支出，并通过案例比较传统计算机和WSC计算机；第7章讨论正常运行时间和可用性，包括如何对故障进行分类以及故障处理、维修优化的方法；第8章总结历史趋势并展望未来——WSC和云计算将成为主流和中心。

本书主要面向当今WSC系统的架构师和程序开发人员，希望能为有志于在此重要领域发展的人员打下一个坚实的基础，同时相关的内容也适用于那些仅想了解互联网基础设施信息的人群。

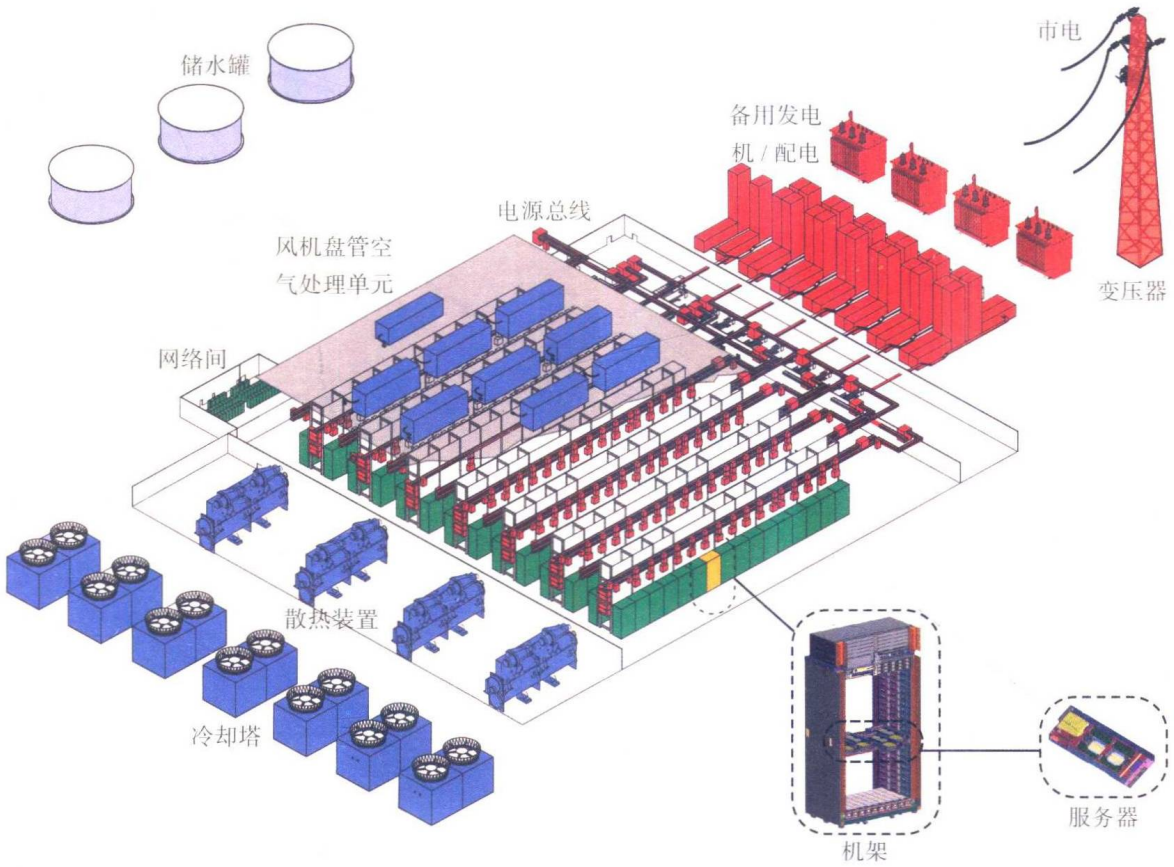


图 4-4　一个典型数据中心的主要部件

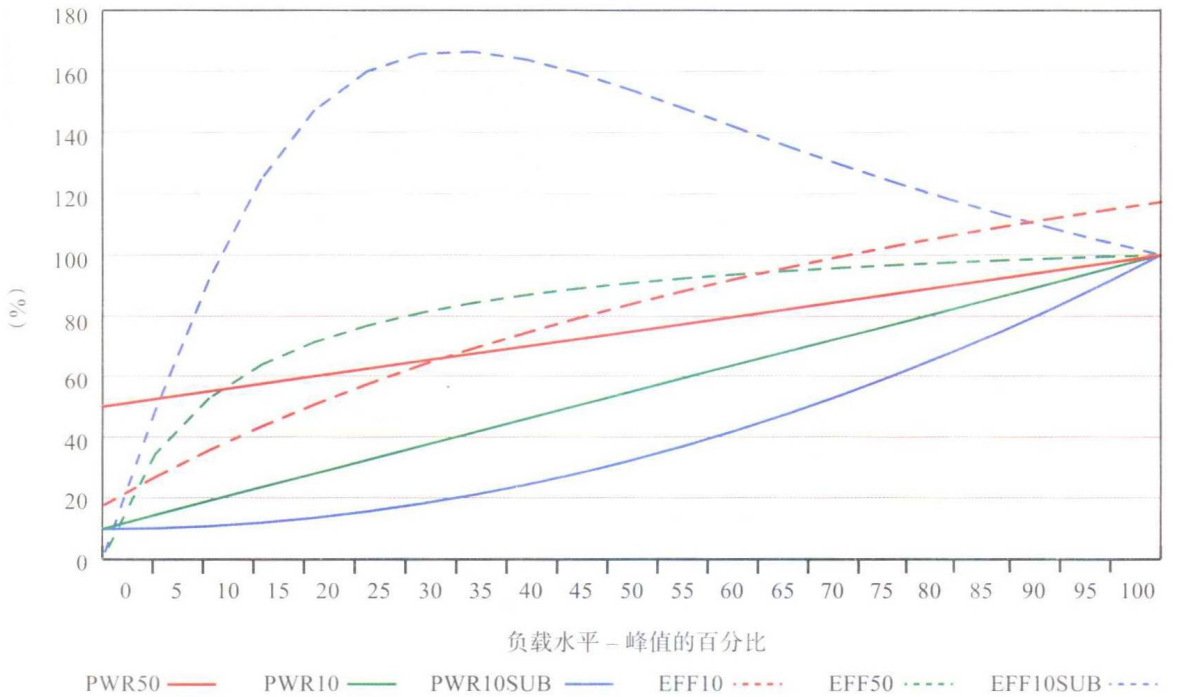


图 5-6　三个假想系统的功率和相应的功效：一个典型服务器，其空闲功率为峰值的 50%（Pwr50 和 Eff50）；一个能耗更成比例的服务器，空闲功率为峰值的 10%（Pwr10 和 Eff10）；一个能耗成次线性比例的服务器，空闲功率为峰值的 10%（Pwr10sub 和 Eff10sub）。实线表示功率 %（归一化为峰值功率），虚线表示功效占峰值的百分比

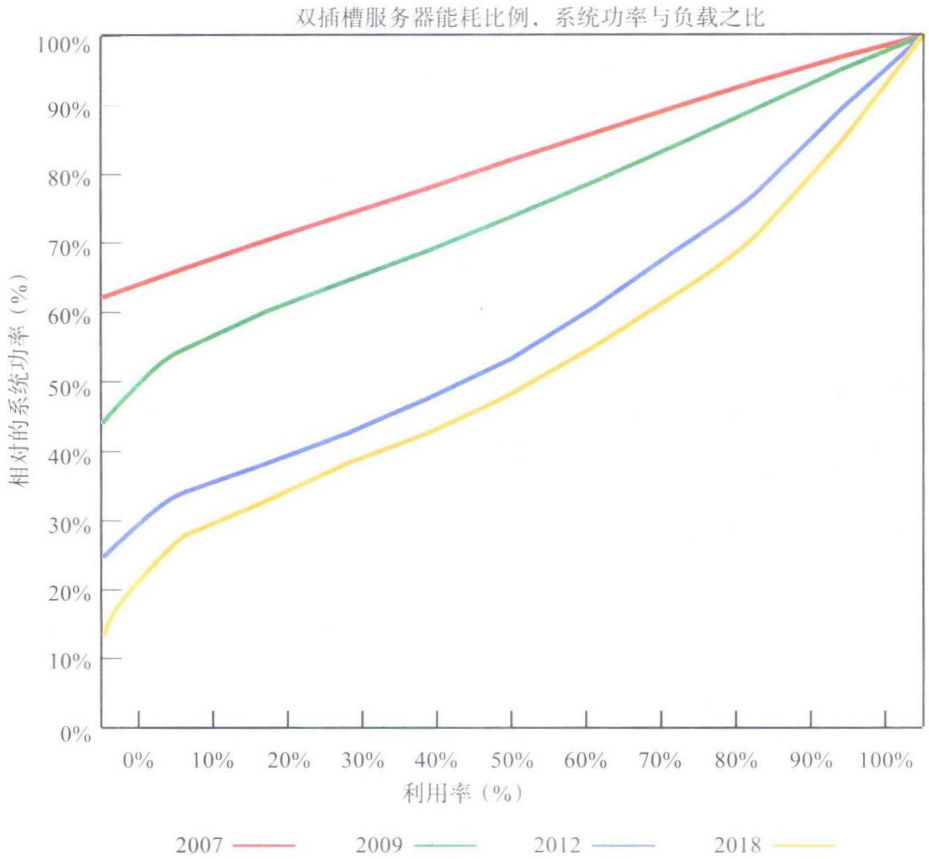


图 5-8　2007 ~ 2018 年间英特尔服务器按峰值归一化的功率与利用率图（感谢谷歌 David Lo 提供）。该图表明英特尔服务器在这 12 年间已经变得更为能耗成比例

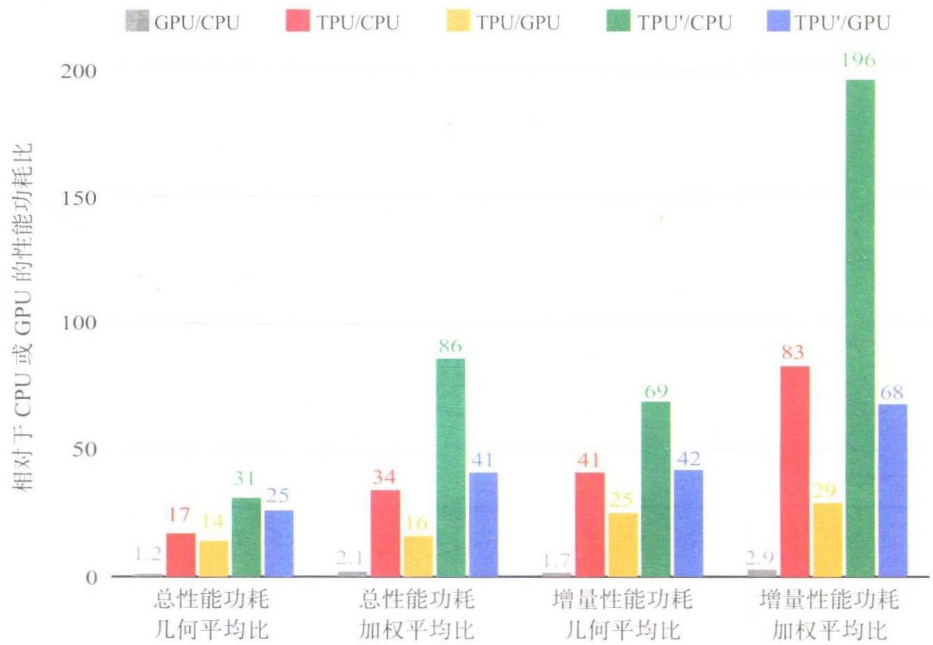


图 5-9　GPU 服务器相对 CPU 服务器（蓝色条）、TPU 服务器（红色条）相对于 CPU 服务器、TPU 服务器相对于 GPU 服务器（橙色条）的相对能效比（TDP 能耗）。TPU' 是一种改进的 TPU，绿色条显示其与 CPU 服务器的比率，蓝色条显示其与 GPU 服务器的关系。总的数据包括主机服务器功耗，但增量数据不包括

译 者 序

“数据中心”这一名词为人们所熟知已有几十年的历史，而云计算作为构建在此之上的一场产业革命，正在改造着社会的各行各业。现如今，云计算早已被应用于社会管理和民生服务的方方面面，而互联网公司作为云计算最早的推动者和受益者，正在依靠数据技术和算力能力把云计算打造成商业创新的基础设施。

IaaS（Infrastructure as a Service，基础设施即服务）无疑是当今云计算最重要的组成部分，它通过整合计算、存储和网络资源降低硬件和运维成本，同时利用云本身的弹性实现灵活性，从而为用户带来可观的经济收益。云计算将传统的企业 IT 转移到云上，把“共享”的概念体现得淋漓尽致，毫不夸张地说，这种“规模化下的经济效益”是商业模式创新的巨大成功。

然而，云计算更是一场技术革命，依托云 IaaS 的底层技术重构基础软件和顶层应用正是大势所趋，容器、微服务和函数计算所代表的正是云原生（Cloud Native）设计理念的核心。以互联网服务为代表的大型应用，由于其规模大、算力高，加之对延迟和可用性有要求，使得应用开发者从设计之初就必须考虑云基础设施的特性，把数据中心高效地当作一台计算机来使用，即 Datacenter as a Computer（数据中心即计算机），这是一个庞大的系统工程。

开放和互通则是云计算的另一张名片。放眼世界，无论是谷歌、亚马逊，还是阿里巴巴，这些互联网公司无一不是在体量达到了一定规模之后，把自有的技术体系开源、开放，再通过云计算的形式普惠给更多的中小企业。PaaS（Platform as a Service，平台即服务）不仅可以让云上的用户根据业务情况来快速构建最适合自己的系统，更能通过其强大的基础软件能力屏蔽底层的硬件差异和细节，让开发者更专注于顶层的业务逻辑。通过开源和开放打破自有的、封闭的技术体系，让全社会、全世界的经验积累变得更加通用，让地球变平，是云计算带来的更大变革。

AI 技术的出现和普及使得机器智能往前跨了一大步，这同时也对数据中心的算力能力提出了更高的要求。随着摩尔定律脚步的放缓，依赖传统的通用 CPU 已经无法满

足复杂的深度学习网络的算力要求，GPGPU 和专用 AI 芯片在数据中心的比重也因此变得越来越高，有效缓解了“算力赤字”的问题，而本书中谷歌 TPU 的系统设计从芯片到整机再到集群的完美契合正是当代数据中心一体化实践的最佳案例。

5G 时代的到来同样给数据中心的设计者提出了新的挑战，无处不在的移动互联网、AIOT 产生的源源不断的海量数据都必将让我们重新审视数据中心现有的计算、存储和网络架构——仓储级的数据中心是否仍然适用于边缘场景？未来的不确定性更意味着无限的可能，这也正是云计算黄金时代的开端。

中国经济的未来在于数据技术与释放生产力的结合，这将给激发内需带来跨越式的发展机遇，而云计算则将成为必不可少的商业基础设施资源。翻译本书的初衷也正是在于传播知识、分享观点和探讨未来，希望本书能通过诸多数据中心一体化的最佳实践给各位云计算行业的从业者带来启发。

最后，特别感谢在本书翻译过程中给予我帮助的各位同事和朋友，以及华章公司的关敏编辑，是你们的反馈和专业的指导使得本书能够更加贴合原著思想地与各位读者见面。

徐凌杰

2019 年 11 月 12 日 于杭州

致　谢

在过去的几年里，我们几位作者直接参与了谷歌基础设施的设计和运营，而我们在其中的大量所学并据此在本书里的总结汇报正是整个谷歌团队同事们辛勤工作、洞察力和创造性的结晶。谷歌基础设施技术团队的工作覆盖了本书的所有内容，我们在此对他们的经验分享表示特别感谢。

感谢谷歌的 Kristin Berdan 以及 Morgan & Claypool 出版公司的 Michael Morgan 让我们有机会将本书的英文电子版免费公之于众——这也正是我们接受本书编写工作的前提条件。非常感谢 Mark Hill 和 Michael Morgan 邀请我们参与这个项目，也感谢 Michael Martonosi 和 Margaret Martonosi 给予我们的鼓励和督促，是你们的耐心让第 3 版变成了可能。

特别感谢谷歌的同事 Jichuan Chang、Liqun Cheng、Mike Dahlin、Tim Hockin、Dave Landhuis、David Lo、Beckett Madden、Jon McCune、Tipp Moseley、Nishant Patil、Alex Ramirez、John Stanley、Varun Sakalkar 以及 Amin Vahdat 帮助审阅第 3 版并添加新的内容。感谢 Robert Stroud、Andrew Morgan 以及其他在过去的版本中帮助我们整理数据和图表的人。特别感谢 Jimmy Clidara 在第 2 版中为第 5 章撰写的关于冷却和配电系统的原始材料。Ana Lopez-Reynolds、Marcos Armenta、Gary Borella 和 Loreene Garcia 负责所有的插图，并创作了新的图表，保持了本书在整体风格上的一致性。感谢我们的审稿人 Christos Kozyrakis 和 Thomas Wenisch 提供了许多宝贵意见，Bob Silva 帮助我们进行了校对。同时也感谢 Morgan & Claypool 出版公司的 Brent Beckley、Deborah Gabriel 和 Christine Kiilerich 的支持与协作。

第 2 版极大地受益于 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 的贡献和指正，由衷地感谢他们的帮助。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 1 版的反馈和指正。Ricardo Bianchini、Fred Chong、Jeff Dean 和 Mark Hill 在第 1 版的草稿阶段提供了极其有用的反馈。第 1 版也受益于 Catherine Warner 所做的校对工作。

最后，我们要感谢读者对本书的支持以及对于前两版的所有反馈。我们一如既往地欢迎你们对本书第 3 版提出各种想法和建议，敬请在 https://goo.gl/HHqQ25 提交评论和勘误，我们在此提前表示感谢。

作者简介

Luiz André Barroso 对网页搜索、基础软件、存储可用性、能效和硬件设计多个工程领域都有涉猎。他曾是谷歌平台工程团队的第一任经理，负责设计公司的计算平台。目前他领导着谷歌地图的工程基础设施工作。加入谷歌之前，他曾在 DEC 公司（后被康柏收购）从事研究工作，他的团队在多核 CPU 处理器和内存系统设计领域做了一些开创性的工作。他拥有南加州大学计算机工程博士学位，在里约热内卢天主教大学获得电气工程本科和硕士学位。Luiz 是谷歌研究员、ACM 会士，也是 AAAS（美国科学促进会）会士。

Urs Hölzle 是谷歌首位工程副总裁，自 1999 年以来一直领导谷歌技术基础设施的开发。他目前的职责包括服务器、网络、数据中心以及基础软件的设计与运维，以支持谷歌内部和对外的云平台。他也因为他的红袜子和兰波格犬 Yoshka（谷歌第一个顶级犬）为人们所熟知。Urs 在瑞士长大，在苏黎世联邦理工学院获得计算机科学硕士学位，之后在斯坦福以富布赖特（Fulbright）学者的身份获得博士学位。在斯坦福期间（包括后来成立一个初创公司并被 Sun 收购），他发明了当今主流 Java 编译器广泛使用的底层技术。在加入谷歌之前，他在加州大学圣塔芭芭拉分校担任计算机科学教授。他是 ACM 和 AAAS 会士、瑞士技术科学院和国家工程院的成员，并在美国的世界自然基金会的董事会任职。

Parthasarathy Ranganathan 是谷歌计算和数据中心硬件领域的技术带头人。此前，他曾是惠普实验室的研究员和首席技术专家，领导系统和数据中心的研究。他曾参与了多个跨学科系统的项目，其中包括能耗感知的用户接口、异构多核处理器、高效能服务器、加速器，以及隔离的和以数据为中心的数据中心，这些创新无论是在学术界还是在工业界都有很高的影响力并被广泛采用。他发表了大量的论文，作为共同发明人的专利超过 100 项。他曾被 Business Insider 提名为 15 大企业技术明星，也是 MIT Tech Review 评出的全球 35 位青年发明者之一。他还是 ACM SIGARCH Maurice Wilkes 奖的获得者以及莱斯大学杰出青年工程校友奖得主。他目前是谷歌的杰出工程师、IEEE 和 ACM 会士。

译者简介

徐凌杰，阿里云资深技术专家，负责包括 GPU 和 AI 芯片在内的数据中心异构计算基础设施，专注于架构与应用的软硬件协同。他是国际机器学习权威组织 MLPerf 的创始成员，成功推出了首个被业界公认的 AI 训练和推理的基准框架。在加入阿里巴巴之前，他曾在 NVIDIA、AMD 和三星担任过多个大型 GPU 芯片项目的高级管理和架构师职位。徐凌杰本科就读于上海交通大学信息工程专业，后赴美在德州大学奥斯汀分校获计算机体系结构硕士学位，并拥有加州大学伯克利分校的 MBA 学位。

目　录