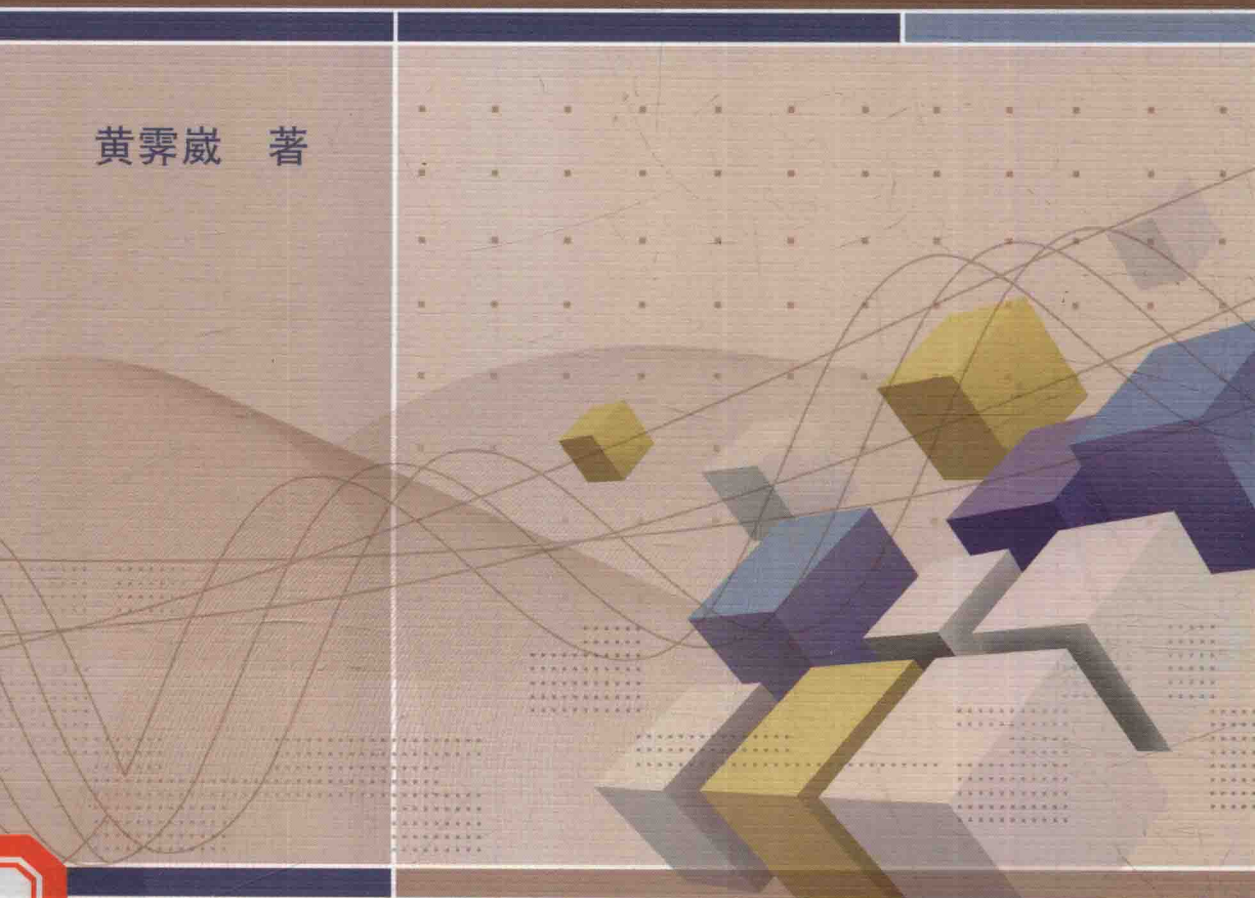
基于随机模型的服务质量多属性评价与优化

Multi-Attribute QoS Evaluation and Optimization Based on Stochastic Models

黄霁崴 著



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

基于随机模型的服务质量多属性评价与优化

黄霖崴 著

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

内 容 简 介

随着服务计算技术的蓬勃发展，服务质量越发成为重要的需求，并呈现多样化特性。本书以理论模型、量化分析和优化技术为重点，从性能、可信赖性和能耗三个维度开展服务质量的多属性评价和优化研究。本书以随机模型为主线，从服务内部的特性出发，描述服务的动态行为，通过数学建模和分析方法，建立服务质量多维度属性的量化评价，进一步分析讨论模型参数和评价结果之间的量化关系，并通过优化方法和技术得到最佳的系统参数或配置方案。在服务质量评价方面，本书介绍了基于排队模型的性能指标评价方法、基于半马尔可夫模型的可信赖性评价方法和基于随机游走模型的服务质量排序方法；在服务质量优化方面，本书介绍了基于折扣马尔可夫决策的性能－能耗综合优化方法，以及基于平均时间马尔可夫决策的可信赖性－能耗综合优化方法。

本书旨在为系统性能评价和服务计算领域的研究人员提供专业参考。本书可作为高等学校相关专业研究生的教学用书，也可作为相关领域研究人员的理论和技术指导书。

图书在版编目（CIP）数据

基于随机模型的服务质量多属性评价与优化 / 黄霁崴著. —北京：电子工业出版社，2017.8
ISBN 978-7-121-32523-6

Ⅰ.①基… Ⅱ.①黄… Ⅲ.①互联网络－网络服务器 Ⅳ.①TP368.5

中国版本图书馆 CIP 数据核字（2017）第 203392 号

策划编辑：张 剑（zhang@phei.com.cn）
责任编辑：张 剑
印 刷：北京京华虎彩印刷有限公司
装 订：北京京华虎彩印刷有限公司
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：787×1092 1/16 印张：6.75 字数：173 千字
版 次：2017 年 8 月第 1 版
印 次：2017 年 8 月第 1 次印刷
定 价：68.00 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：zhang@phei.com.cn。

前　言

随着计算机技术和网络技术的不断发展，服务计算（Services Computing）应运而生。服务计算是一种新兴的计算模式，它使服务提供商将特定的功能封装成为遵循统一标准的、交互式的计算机程序，将其以服务的形式提供给不同商业领域中的多个用户使用。服务计算具有灵活、高效的特点，构成了连接商业服务与信息服务的桥梁，带来了商业模式和计算模式的转型，受到了学术界和工业界的青睐。随着服务计算技术的发展和广泛应用，互联网上各类服务的数目逐渐增多。越来越多的第三方在互联网上提供了功能相同或相似的服务，为同样的服务流程提供了多个候选服务。因此，用户对服务的关注逐渐从功能性需求（服务标准的制定、服务流程的设计、服务系统的实现等）向非功能性需求（服务质量）转变。服务质量（Quality of Service，QoS）体现的是消费者对服务提供者所提供的服务的满意程度，是对服务者服务水平的一种度量和评价。目前，在服务计算中按照用户的要求提供 QoS 保障是一个普遍的需求，也是服务计算发展的重要挑战，已成为当今服务计算领域的核心研究方向和热点问题之一。

随着服务计算应用范围的扩展，用户对服务质量的需求逐渐呈现多样化的特点。首先，性能是服务质量首要关注的指标，它反映了服务的效率，代表服务的处理能力，对于计算密集型服务或实时服务具有重要意义，是其核心的服务质量评价指标。服务质量在性能维度上又包括响应时间、吞吐率、利用率等属性。其次，随着服务计算逐渐应用于医疗卫生、交通管理或军事作战等领域，用户对服务正常持续运行的能力提出了越来越高的要求，因此可信赖性作为其评价指标成为服务质量中的另一类重要需求。可信赖性代表了服务系统避免或减少发生服务失效的能力，包括可用性、完整性、保险性、可维护性和可靠性等属性。此外，随着支撑服务的计算机和网络系统的规模逐渐增大，能源消耗越来越大，对环境造成的影响越来越严重。近年来，能耗问题已成为服务质量研究中的热点问题，是服务设计和运行中需要关注的重要因素。

服务质量研究的一个重要理论基础是服务质量的评价和优化。它可以定量描述服务质量多个维度的属性指标，评价各种服务设计方案或运行策略在不同维度指标上的优劣程度，并从理论上指导服务系统的构建和服务运营方案的选择。目前，大部分的服务质量优化工作都是以数据作为驱动，根据对已有服务或系统的服务质量的测量结果，预测未知方案的服务质量表现，或者直接对新方案实现后进行测量，进而选取最优的方案。在这种方法中，原始数据的获取需要对服务系统进行实现和部署，开销巨大；基于数据的预测受限于数据规模和质量，可能存在偏差，不能反映服务系统的本质行为。因此，需要从服务内部的特性出发，描述服务的动态行为，通过数学建模和分析方法，建立服务质量多维度属性的量化评价体系，进一步分析讨论模型参数和评价结果之间的量化关系，以便得到最优的系统参数或配置方案。随机建模分析方法在服务质量评价和优化方面提供了可行的新思路和新技术，为服务质

量的保障和优化提供了理论依据和支撑技术，这将是一个重要的充满前景的研究方向。

本书作者在基于随机模型的服务质量评价和优化领域进行了一系列深入而系统的研究工作。本书主要以马尔可夫模型、排队模型等为基础，以最优化方法、马尔可夫决策等为技术支撑，探讨服务质量的建模和优化方法，并结合服务计算中的具体问题进行探讨和验证。本书绝大部分内容取材于作者近期在国际一流学术期刊和会议上发表的论文，全面而系统地展示了很多新的研究成果和进展。

作者的研究工作得到了国家自然科学基金项目（No. 61502043）、北京市自然科学基金项目（No. 4162042）和北京市优秀人才培养资助项目（No. 2015000020124G082）的资助。

由于作者水平有限，加之基于随机模型的服务质量评价和优化的研究仍处于不断发展之中，书中存在疏漏和不足之处在所难免，恳请专家、读者予以批评指正。

黄霁崴

2017 年 4 月于北京

目　录

前言
第 1 章　绪论 …… 1
1.1　研究背景 …… 1
1.1.1　服务计算 …… 1
1.1.2　服务质量的评价与优化 …… 3
1.2　研究内容 …… 3
1.3　研究难点与创新点 …… 4
1.4　章节组织结构 …… 5
第 2 章　相关研究综述 …… 7
2.1　服务质量的评价指标 …… 7
2.1.1　性能指标 …… 7
2.1.2　可信赖性指标 …… 7
2.1.3　能耗指标 …… 8
2.2　服务计算的评价方法 …… 8
2.2.1　模型方法 …… 8
2.2.2　测量方法 …… 11
2.2.3　预测方法 …… 11
2.3　服务计算中的多属性问题 …… 12
2.3.1　多属性之间的相互关系 …… 12
2.3.2　多属性优化 …… 12
2.4　讨论与总结 …… 13
第 3 章　基于排队模型的性能评价 …… 15
3.1　服务和服务系统的性能模型 …… 15
3.1.1　原子服务的性能模型 …… 15
3.1.2　服务系统的性能模型 …… 16
3.2　模型求解与量化分析 …… 17
3.2.1　马尔可夫排队模型的求解 …… 18
3.2.2　半马尔可夫排队模型的求解 …… 19
3.3　性能优化举例 …… 20
3.3.1　服务节点的资源管理 …… 20
3.3.2　分层系统的任务调度 …… 22

3.4 案例分析 …… 23
3.4.1 数据集介绍 …… 23
3.4.2 原子服务的模拟实验 …… 24
3.4.3 服务系统的模拟实验 …… 26
3.5 本章小结 …… 27
第4章 基于半马尔可夫模型的可信赖性评价 …… 29
4.1 服务计算系统的可信赖性建模 …… 29
4.1.1 基于半马尔可夫过程的系统状态建模 …… 29
4.1.2 可信赖性形式化定义 …… 31
4.2 模型分析与可信赖性评价 …… 32
4.2.1 稳态概率分析 …… 32
4.2.2 平均失效时间分析 …… 33
4.3 可信赖性属性之间的相互关系 …… 37
4.4 服务选择和服务组合的可信赖性建模与分析 …… 41
4.4.1 基本框图模式的可信赖性分析 …… 41
4.4.2 基本服务模式的可信赖性分析 …… 42
4.5 案例分析 …… 44
4.5.1 系统简介 …… 44
4.5.2 可信赖性分析 …… 44
4.5.3 敏感度分析 …… 46
4.6 本章小结 …… 48
第5章 基于随机游走模型的服务质量排序评价 …… 49
5.1 协同分布式服务评价框架 …… 50
5.2 服务评价排序模型 …… 50
5.2.1 基于比较的概率模型 …… 51
5.2.2 全局排序的随机游走模型 …… 51
5.2.3 随机游走的马尔可夫模型 …… 52
5.3 协同服务评价算法 …… 52
5.3.1 算法概述 …… 52
5.3.2 比较聚合方法 …… 53
5.3.3 求解马尔可夫链算法 …… 53
5.4 算法分析 …… 55
5.4.1 误差分析 …… 55
5.4.2 时间复杂度分析 …… 55
5.5 案例分析 …… 56
5.5.1 数据集简介 …… 56
5.5.2 评价指标 …… 57
5.5.3 实验结果 …… 57
5.6 本章小结 …… 59

第 6 章　基于折扣马尔可夫决策的性能—能耗综合优化 …… 61
6.1　问题描述 …… 61
6.2　服务系统的性能—能耗评价模型 …… 62
6.2.1　原子服务的性能模型 …… 62
6.2.2　嵌入离散时间马尔可夫回报模型 …… 64
6.2.3　模型分析的更新技术 …… 65
6.3　基于马尔可夫决策的动态功耗控制 …… 66
6.3.1　问题描述和求解方法 …… 66
6.3.2　马尔可夫决策对马尔可夫回报优化的有效性证明 …… 68
6.4　分布式的动态服务选择与功耗控制综合优化 …… 68
6.4.1　优化框架 …… 68
6.4.2　优化算法 …… 69
6.5　案例分析 …… 71
6.6　本章小结 …… 76
第 7 章　基于平均时间马尔可夫决策的可信赖性 - 能耗综合优化 …… 77
7.1　服务系统的可依赖性—能耗评价模型 …… 78
7.1.1　可信赖性模型 …… 78
7.1.2　能耗模型 …… 79
7.1.3　嵌入离散时间马尔可夫模型 …… 79
7.2　基于回报值等价的状态聚合方法 …… 80
7.2.1　可信赖性模型中的状态聚合 …… 80
7.2.2　能耗模型中的状态聚合 …… 82
7.3　基于马尔可夫决策的可信赖功耗控制算法 …… 84
7.3.1　优化建模与算法 …… 84
7.3.2　最优性证明 …… 85
7.4　案例分析 …… 86
7.5　本章小结 …… 87
第 8 章　总结与展望 …… 89
8.1　总结 …… 89
8.2　研究展望 …… 90
参考文献 …… 93

第1章

绪　论

1.1　研究背景

1.1.1　服务计算

近年来，互联网的发展与普及为计算机软件的设计模式带来了新的思路与挑战。随着软件工程理念和技术的不断进步，传统的基于硬件和系统的架构模式逐渐转变为面向服务的设计模式。服务业已成为IT系统的基础组成部分[1]。一种新兴的计算模式——服务计算（Services Computing）应运而生，并被越来越广泛地应用于各个领域。这里，服务（Service）是指服务提供商与用户之间为实现特定的商业目标或解决方案而形成的一种契约关系；服务计算则是指服务提供商将特定的功能封装成遵循统一标准的、交互式的计算机程序，并将其以服务的形式提供给不同商业领域中的多个用户使用的一种计算模式[2]，它构成了连接商业服务与信息服务的桥梁。服务计算带来了商业模式和计算模式的转型。对服务计算的研究已成为目前学术界和工业界的研究热点，形成了涉及服务科学与信息技术的交叉学科。

服务计算是在预先定义好的标准框架下对服务过程进行设计、操作、管理和优化，以满足日益增长的动态的业务需求[3]。它涉及面向服务的体系结构（Service Oriented Architecture，SOA）、Web服务（Web Service，WS）、网格计算（Grid Computing）、云计算（Cloud Computing）等多项技术，涵盖了服务的整个生命周期的所有阶段，包括服务咨询与企划（Consulting and Strategic Planning）、服务契约（Services Engagement）、服务发布（Services Delivery）、服务运营（Services Operation）、服务计费（Services Billing）、服务管理（Services Management）等，由此也产生了服务标准制定、架构设计、服务发现、服务推荐、服务选择、服务组合、监控与管理、服务质量评价和保障等一系列科研问题[4]。

服务计算系统是基于IT技术，为实现特定的业务服务需求而构建的计算机系统，为服务的持续性工作提供硬件保障和技术支撑[4]。一个服务计算系统包括3层结构，底层由基础设施构成，基于服务计算技术构建的大规模数据中心是提供服务的基础设施，它可以看做是一个通过网络互联的大型服务器集群，承载着各种应用和服务[5]；第2层是管理模块和中间件，包括虚拟机管理、服务质量管理、服务组合、服务监控、计费系统等模块，它将服务层的应用和底层的硬件基础设施联系起来，并具有松耦合的特点，使整个系统具有更灵活的特

性；第3层是服务层，面向用户，通过统一的标准化接口，为用户提供满足需求的各种类型的服务。借鉴典型的服务计算实例——云计算系统作为参考[6]，结合服务计算普遍的特性，可以得到典型服务计算系统架构示意图，如图1.1所示。

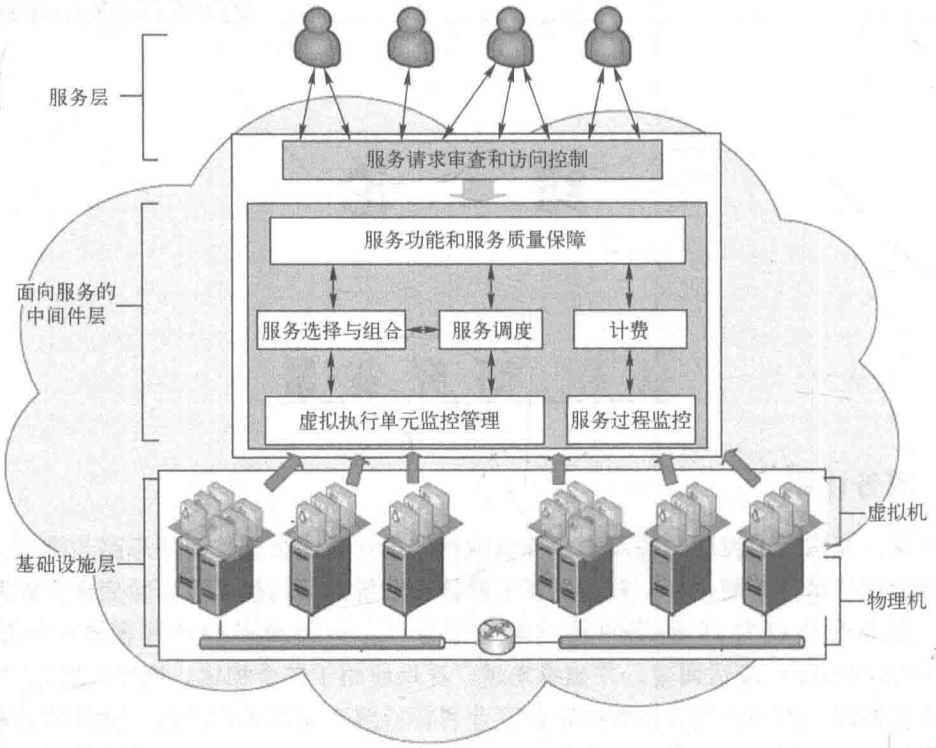


图1.1　典型服务计算系统架构示意图

随着服务计算的逐渐普及，用户需求的不断提高，服务计算及服务计算系统呈现如下3个显著特点。

1）动态性　为了满足用户日益增长的多样化需求，提高服务的利用率，服务之间可以通过服务组合[7]的形式调用已有的服务，并且相互协同[8]，从具有简单功能的服务构建复杂的服务工作流。服务组合和服务协同使得服务的过程具有高度动态性，一个服务过程可能会涉及一系列的服务和系统组件。

2）松耦合　服务计算中，通过服务描述语言（Web Services Description Language，WSDL）[9]定义统一的服务接口，降低了上层服务与下层实现机制的依赖程度（即耦合性）。另一方面，虚拟化是服务计算系统构建中的常用技术[10]，硬件资源的虚拟化使得软件级的服务实现与系统底层硬件实现了松耦合，服务计算中的上层服务与底层硬件之间不再具有固定的对应关系。虚拟机迁移[11]和虚拟资源共享[12]机制则进一步增加了耦合的灵活性，并提高了资源利用的效率。

3）大规模　在服务计算中，服务被部署在分布式的硬件基础设施上。随着服务种类和数目的不断增多，基础设施的规模也在不断增大[2]。服务计算中的底层数据中心往往包括成千上万的服务器和大量的网络设备[5]。服务和硬件设备的大规模特性，为系统的设计、服务的部署、资源的管理带来了严峻的要求和巨大的挑战。

1.1.2 服务质量的评价与优化

服务计算兴起之初，绝大多数的研究主要关注于服务计算功能性（Functional）需求的满足，研究内容包括服务标准的制定、服务流程的设计和服务系统的实现等。随着服务计算被越来越广泛地应用，各种类型的应用对系统和服务提出了越来越高的需求，其中非功能性（Nonfunctional）的需求，即服务质量（Quality of Service，QoS），逐渐被关注和研究[13]。这里，服务质量可以包括多种指标，每类指标又包含多个属性。其中，第一类是传统的性能指标，它反映了系统处理能力或服务效率，包括响应时间、吞吐率、利用率等属性；第二类是服务的可信赖性[14]，它描述了服务计算系统避免或容忍组件或服务失效，以持续提供可用服务的能力，包括可靠性、可用性、可维护性、完整性、保险性等；第三类是服务过程中所消耗的能源或系统运行的能耗效率，即能耗与功率[15]。

目前，在很多服务过程中，用户均对服务提供商提出了 QoS 的需求，这些需求往往通过规范化的服务等级协议（Service - Level Agreement，SLA）得以体现[16]。它是在一定开销下为保障服务的性能和可靠性等 QoS 指标，服务提供商与用户之间定义的一种双方认可的协定。SLA 协议已成为服务计算 QoS 保障中的一个重要方面。

在服务设计和系统实现过程中，为了保障服务的 QoS，首先需要研究系统配置、负载对 QoS 指标的影响，从而给出 QoS 的准确评价；其次，需要对服务计算系统进行优化，选择最优的系统配置和服务解决方案，以适应不同的环境与需求。优化过程覆盖服务系统的整个生命周期，涉及多个学科，如运筹学、复杂系统建模、系统工程等[2]。服务质量的评价与优化是服务计算领域中的一个热点研究问题。

1.2 研究内容

性能评价（Performance Evaluation）[17]是指对系统的动态行为进行分析和优化，包括对实际系统的行为进行建模，对系统性能指标进行测量，对已有系统的性能缺陷或瓶颈进行改进，根据性能需求对系统实施方案进行设计或选择。性能评价是计算机系统和计算机网络研究与应用的重要理论基础和支撑技术，也是当今通信和计算机科学领域的重要研究方向[18]。

基于随机模型的方法是性能评价的重要手段，其核心思路是对要评价的服务或系统建立一个适当的数学模型，然后根据模型参数求解评价指标。模型参数的确定往往依赖于对实际评价对象的测量结果或其参数的估计[19]。基于随机模型的评价方法可以清晰地描述各种因素之间的关系，并且可以应用于尚未存在的系统的性能预测，普适性好，评价成本低，因而在服务计算的 QoS 评价中被广泛使用。根据模型分析结果，可以利用数学方法或计算机方法调整系统参数或配置策略，在保证用户得到约定或希望的服务质量的前提下，使资源利用率最大化，提供服务质量最优的服务。

随着服务计算中 QoS 受到越来越广泛的关注，如何对服务计算系统进行评价、比较和改进，是服务计算中一个重要的问题。性能评价理论与技术逐渐发挥出其巨大的价值，它包括如下 5 个方面。

1）多维度的评价指标 考虑服务计算的特点和需求，建立完整统一的多维度、多属性评价指标体系，可以全面、客观地对系统进行评价。

2）随机评价模型 根据评价指标，建立系统的随机理论模型，可以对系统的动态行为、资源管理和任务调度策略进行形式化的描述。

3）量化分析 在模型基础上，对系统进行量化分析，给出指标的形式化解析解，以便对不同的系统进行分析和比较。

4）属性之间相互关系的研究 各属性之间存在相互影响的关系，对其进行深入研究可以完善整个评价和优化的理论体系，并为多指标优化提供理论上的依据。

5）多属性综合优化 研究服务计算系统的多属性综合优化方法，对系统设计改进和最优控制具有重要的指导意义。

本书针对服务计算背景下的性能评价展开研究，从性能、可信赖性和能耗3个维度，考虑每个维度下的多个属性，建立服务计算的随机评价模型；研究模型的分析求解方法，给出各属性的形式化的量化表达；研究属性之间的相互关系，结合随机评价模型给出其数学表达，用以指导系统优化设计；在定量分析的基础上，研究优化方法，结合随机模型理论，给出服务质量动态保障机制和优化策略。

1.3 研究难点与创新点

服务计算环境的动态性、复杂性、大规模、资源的异构性等，给建模、评价和优化工作带来了一系列困难和挑战。本书的研究工作分别针对以下难点，进行了创新性的研究。

1）服务质量的多维度、多属性的评价指标体系 服务计算中，对性能、可信赖性和能耗3个维度均提出了严格的要求，每个维度又都包含多个属性，从而构成了一个多维度、多属性的评价指标体系。如何全面、客观、普适地建立评价指标体系，用以对服务计算中的计算流程、系统实现进行分析和比较，是一个研究难点和热点。

【创新点1】 本研究从不同层面研究系统建模的理论框架和评价体系，分析和解决相关科学问题，建立服务计算系统中性能、可信赖性和能耗的多维度、多属性评价指标体系。在此基础上，进一步研究多指标之间的相互关系，完善多属性评价的理论框架，为改善服务系统提供科学支撑。

2）服务计算的建模方法 服务计算中的动态性、松耦合和大规模的特点，使得传统的基于组件的建模方法面临了巨大的困难。用户得到的服务和底层硬件不再具有固定的对应关系，单纯地进行系统组件的建模和分析难以反映服务的特性。服务计算中亟需新的建模思路和方法，以有效地对系统行为和特性进行描述。

【创新点2】 本研究从面向服务的角度出发，研究服务计算的特性，建立其评价模型。该面向服务的建模方法从服务层出发，自上而下对服务计算系统进行描述和分析，对其进行综合度量，有利于实时监控系统状态，深入理解系统特性，指导建立相应的保障机制，并丰富建模理论。

3）服务计算评价模型的分析方法 服务计算的大规模特性，使得分析评价的过程中可能会遇到状态空间过大，模型求解过程过于复杂，服务评价成本过大等问题。朴素的模型求解和评价方法在大规模服务的分析过程中可能不再有效。高效的服务分析方法和评价技术，是服务计算中亟待解决的问题。

【创新点3】 本研究从服务计算的模型出发，针对其马尔可夫模型和半马尔可夫评价模

型，分别给出了状态聚合方法。从理论上可以严格证明，聚合后的模型和原模型在评价指标上等价。这些方法的提出，可以有效降低模型状态的规模，并能更清晰地反映评价对象的特性，直观描述系统的动态行为。

【创新点4】 本研究针对大规模服务评价中的服务排序问题，提出了高效的分布式服务评价技术，及其相关的排序聚合方法。该方法可以显著地降低服务评价的成本，并且将误差控制在可接受的范围内。分布式的服务评价技术可以为大规模的服务选择、服务组合、服务推荐等提供理论支持和技术保障。

4）服务计算多属性动态优化方法 服务计算的动态性对其参数配置和流程优化带来了挑战。考虑服务计算中对多维度多属性的需求，如何在理论上保证优化方法的最优性，同时在技术上保证算法的效率，是本研究中的核心问题。

【创新点5】 本研究采用马尔可夫决策过程对服务计算中的动态优化进行形式化建模。在理论上证明了马尔可夫决策过程与服务计算的评价模型（马尔可夫回报模型）优化的一致性，并量化地给出了二者在优化目标上的对应关系。在技术上，借鉴马尔可夫决策中的前沿技术和相关算法，给出了高效可行的优化方法。

【创新点6】 为了在大规模待选服务中选取最优解，本研究提出了基于多 Agent 技术的分布式协同优化方法，给出了优化框架及其相关的优化决策算法。利用 Agent 间相互通信和协作的特性，将问题分而治之，并行地寻找局部最优解，进而从局部最优解中找到可行的全局最优解。

1.4 章节组织结构

本书围绕性能、可信赖性和能耗3个维度开展了随机建模、量化分析和优化方法研究，并针对服务排序这一类特殊的评价问题，进行了普适性的建模理论和评价方法的探讨。

全书内容共分为8章。其中，第1章为绪论；第2章介绍相关研究工作，对服务质量的评价指标、评价方法和多属性优化等方面的已有研究进行综述；第3章研究服务计算系统的性能评价方法，基于排队论建模理论，提出了原子服务和服务系统的性能模型，给出了模型的量化分析方法，并介绍了资源管理和任务调度等典型优化问题的解决方法；第4章研究服务计算系统的可信赖性评价，提出了面向服务的可信赖性评价模型，给出了可信赖性各属性的分析方法，讨论了可信赖性属性间的相互关系；第5章研究服务排序评价的整合测量技术，旨在大规模服务环境下，针对排序评价结果，有效地对评价结果进行聚合，从而减小评价的开销，并介绍了基于比较的排序模型，给出了模型分析算法，通过理论分析和实验验证证明了模型和算法的有效性；第6章研究服务计算中性能和能耗的综合评价和优化，提出了基于马尔可夫回报模型的系统建模方法，给出了性能和能耗的量化分析，以及基于回报值的综合优化框架和算法；第7章综合考虑性能收益、可信赖性和能耗，给出了普适的评价模型，介绍了基于平均时间马尔可夫决策理论的综合优化方法，并在理论上证明了优化方法的有效性；第8章总结全文，并对未来的研究方向进行展望。

第2章

相关研究综述

随着服务概念的推广，服务计算逐渐成为一个活跃的研究领域。服务质量的评价和优化，是服务计算中的一个核心问题，关系到用户需求的满足、服务计算系统的优化设计和有效管理，因此引起了相关领域专家和学者们极大的研究兴趣。

2.1 服务质量的评价指标

服务质量的评价是对服务和系统优劣的客观量化描述，是对系统设计、服务管理方案进行比较和选择的基准。随着服务计算的兴起和广泛应用，用户多样化的需求对服务质量的多个维度、多个指标提出了严格的要求。目前，服务质量的评价可以从3个维度展开，即性能、可信赖性和能耗。

2.1.1 性能指标

性能反映了服务的能力或效率，是服务质量中的一个重要方面。性能指标是一个综合性的指标，包括响应时间（Response Time）、吞吐率（Throughput）、利用率（Utilization）等属性。

响应时间是从服务请求发出到执行完成的时间，包括请求发送延迟、等待时间、服务执行时间等。在服务计算中，响应时间是性能的代表性指标，与服务器负载、服务调度策略、数据带宽等系统参数有密切的联系[20]。

吞吐率是指系统单位时间内可以完成的服务的数目，是评价服务计算系统最大服务能力的参考指标[21]。

利用率反映了服务或系统资源被使用的情况，是整个服务过程中该服务或承载服务的软/硬件资源被使用的比率。服务和系统资源的利用率是系统和服务管理、资源分配、任务调度等过程的一个关键因素，并直接关系到服务提供商的利益，在服务计算的评价和优化过程中需要加以关注[22]。

2.1.2 可信赖性指标

随着服务计算的广泛应用，服务种类的增多对服务质量提出了更高的需求。除传统的性能指标外，用户对持续服务、可靠运行的需求成为了关键需求，尤其是在一些特定的服务应

用中，如交通管制、医疗卫生监护、军事应用等。可信赖性（Dependability）是指一个系统避免发生严重的服务失效或过于频繁超出可接受范围的故障的能力[14,23]，它已逐渐成为了学术界和工业界的研究热点。

可信赖性是一个综合性的概念，它包含如下5个不同的属性[14,24]。

1）**可用性** 是指当用户需要时，系统可以向用户提供正常服务的能力。

2）**完整性** 是指抵御不适当的系统或数据修改的能力。

3）**保险性** 是指服务在生命周期内，不会对用户和环境造成灾难性后果的能力。

4）**可维护性** 是指系统进行修复和容错的能力。

5）**可靠性** 是指服务能够持续正常工作的能力。

在服务计算系统的生命周期中，其可信赖性面临的威胁包括3类，即缺陷（Fault）、错误（Error）和故障（Failure）；而增强可信赖性的手段包括缺陷的预防（Fault Prevention）、容忍（Fault Tolerance）、修复（Fault Removal）和预测（Fault Forecasting）[14]。

2.1.3 能耗指标

近年来，计算系统中的能量消耗受到了越来越多的关注。有报告指出，计算机服务器的能耗开销呈逐年增长的趋势[25]。在Web服务、云计算、数据中心等大规模服务计算系统中，能源消耗的问题更加严峻[26]。2006年，仅在美国境内，IT基础设施的运营和维护就消耗了610×10^8kW·h的电量，造成了超过3200万美元的开销[27]。2010年，仅Google一家公司就消耗了22.6×10^8kW·h的电量[28]。截至2011年，为了承载互联网上广泛发布的各类服务，底层数据中心消耗的电量占据了全球耗电总量的1.3%；据推测，该比例将在2020年增长至8%[29]。能耗问题已经成为服务计算设计和维护中不可或缺的关键因素[15,30]。

能耗指标主要包括两个属性，即功率（或功耗，Power）与能耗（Energy）[26]。系统的功率$P(\tau)$是指系统在运行过程中的某个时间τ消耗能量的速率，而能耗是指一段时间t内系统消耗的总能量。功率与能耗之间的对应关系为

$$E(t)=\int_0^t P(\tau)\mathrm{d}\tau \tag{2.1}$$

2.2 服务计算的评价方法

针对多维度的评价指标，如何得到指标的数值，是服务计算评价中需要解决的核心问题之一。服务计算的评价方法可以分为3类，即模型方法、测量方法和预测方法。

2.2.1 模型方法

基于模型的方法是性能评价的重要手段，其核心思路是对要评价的服务或系统建立一个适当的模型，然后根据模型参数求解出评价指标。这里，模型参数的确定往往依赖于对实际评价对象的测量结果或对其参数的估计[19]。基于模型的评价方法可以清晰地描述各种因素之间的关系，并且可以应用于尚未存在的系统的性能预测，普适性好，评价成本低，因而在服务计算的性能评价中被广泛使用。

1. 马尔可夫模型

马尔可夫模型（Markov Model）是性能评价常用的模型工具。它利用马尔可夫过程

(Markov Process) 对系统的动态行为进行描述。马尔可夫建模方法要求所描述的系统或服务的状态变化具有无后效性（或称为马尔可夫特性），即对于系统状态变化的随机过程$\{X(t), t \in T\}$，如果对于任意的时间参数$t_0 < t_1 < \cdots < t_n < t$，在$X(t_0), X(t_1), \cdots, X(t_n)$已知的情况下，$X(t)$的条件分布只与$X(t_n)$状态相关，即

$$\Pr\{X(t) \leqslant x \mid X(t_0) = x_0, X(t_1) = x_1, \cdots, X(t_n) = x_n\} = \Pr\{X(t) \leqslant x \mid X(t_n) = x_n\} \tag{2.2}$$

一般地，在性能评价研究中可以假定系统的行为不依赖于观测时间，符合时齐性(Time Homogeneous)，即认为马尔可夫过程的条件概率分布函数与观察起始时刻无关[19]。在评价中，可以任意选取时间轴起点$t_n < t$，时齐性保证下式始终成立：

$$\Pr\{X(t) \leqslant x \mid X(t_n) = x_n\} = \Pr\{X(t - t_n) \leqslant x \mid X(0) = x_n\} \tag{2.3}$$

具体地，在性能评价中使用的马尔可夫模型也可进一步细分为3类，即马尔可夫链(Markov Chain)、马尔可夫回报模型（Markov Reward Model）和半马尔可夫模型（Semi-Markov Model）。

1）马尔可夫链 考虑到绝大多数计算机系统的状态空间均为离散的，这样的马尔可夫过程称为马尔可夫链（Markov Chain，MC）。根据时间参数t的取值范围T，MC又可以进一步分为离散时间马尔可夫链（Discrete-Time Markov Chain，DTMC）和连续时间马尔可夫链(Continuous-Time Markov Chain，CTMC）两种。

马尔可夫链被广泛用于计算机系统的性能建模。文献［31］采用CTMC对云计算中的服务器进行了建模，将服务器缓存中的虚拟机迁移请求数目定义为CTMC的状态，综合考虑虚拟机迁移请求到达和服务过程，应用马尔可夫模型理论和验证技术对系统性能进行评价。文献［32］针对大规模网格服务系统建立了系统状态转移的DTMC模型，描述了系统在大规模请求到达或出现故障情况下的动态响应和状态变化，用以评价系统的可用性和可靠性属性。文献［33］对服务计算中的虚拟机集群进行了分析，根据虚拟机状态建立了CTMC模型，通过CTMC稳态概率的求解，给出了系统可用性的量化分析。文献［34］建立了服务器状态的DTMC模型，研究系统的能耗状态，用以实现性能、可靠性和能耗平衡。

2）马尔可夫回报模型 在马尔可夫链中，若对每个状态定义一个回报值，代表系统在利润、性能、可靠性等方面的收益，则该MC称为马尔可夫回报模型（Markov Reward Model，MRM）。与MC类似，MRM同样可以分为离散时间和连续时间两种时间维度。

回报值的引入为性能评价提供了巨大的便利。MRM可以将模型结构和系统需求紧密联系起来，并且可以和马尔可夫决策理论中的优化过程相结合，是性能、可靠性等分析评价的有力工具[35]。文献［36］提出了生产服务系统的离散时间MRM建模方法，用以评价系统的性能，在该模型中，状态回报值可以具有随机性，以适应实际系统中更为复杂的情形。文献［37］采用连续时间的MRM对服务共享系统的状态转移进行建模，通过模型分析给出了系统可用性的评价方法。文献［38］从离散时间和连续时间两个维度建立了服务组合和系统资源的MRM，对服务的可靠性进行了高效而准确的分析。

3）半马尔可夫模型 半马尔可夫模型是对经典马尔可夫模型的推广。考虑一个有限状态、连续时间、齐次的随机过程$\{X(t), t \geqslant 0\}$，在时刻$t_0 < t_1 < \cdots < t_n$状态发生变化，并且在时间间隔$[t_n, t_{n+1})$有值Y_n，满足：

$$\begin{aligned}&\Pr\{Y_{n+1} = j, t_{n+1} - t_n \leqslant \tau \mid Y_0 = y_0, Y_1 = y_1, \cdots, Y_n = i; t_0 = \tau_0, t_1 = \tau_1, \cdots, t_n = \tau_n\}\\ &= \Pr\{Y_{n+1} = j, t_{n+1} - t_n \leqslant \tau \mid Y_n = i\} = H_{ij}(\tau)\end{aligned} \tag{2.4}$$