# 结构矩阵方法与工作休假排队

李继红 著

科学出版社

# 结构矩阵方法与工作休假排队

李继红　著



科学出版社
北京

## 内 容 简 介

本书基于结构矩阵方法，系统论述工作休假排队的思想原理和主要结果，着重介绍嵌入 Markov 链(MC)、拟生灭过程与结构矩阵在工作休假排队分析中的应用，内容包括 M/M/1 型、GI/M/1 型和 M/G/1 型经典工作休假排队的建模和分析，并简要介绍休假中断策略、成批到达、门限策略与工作休假排队结合的研究成果. 书中包含各类工作休假排队的详细数值分析及在通信网络性能分析的应用实例.

阅读本书只需要矩阵分析、概率论和随机过程的基本知识. 本书可供随机运筹学及管理科学领域的研究人员、高校教师和相关专业研究生参考使用，也可供计算机系统和通信网络的工程技术人员阅读参考.

**图书在版编目(CIP)数据**

结构矩阵方法与工作休假排队/李继红著. —北京：科学出版社, 2016.3

ISBN 978-7-03-047640-1

Ⅰ. ①结… Ⅱ. ①李… Ⅲ. ①矩阵法分析 Ⅳ. ①O342

中国版本图书馆 CIP 数据核字(2016) 第 049132 号

责任编辑：李 欣 / 责任校对：张凤琴

责任印制：张 伟 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 3 月第 一 版 开本：720 × 1000 B5

2016 年 3 月第一次印刷 印张：13 3/4

字数：267 000

**定价：88.00 元**

(如有印装质量问题，我社负责调换)

# 前　言

随着电子商务和现代技术的迅猛发展, 管理机构和通信网络的运行越来越复杂, 对服务质量和系统性能要求也越来越高. 排队论, 作为解决系统理论分析和应用研究的有效工具, 需要不断引入新策略, 以适应和解决不同的排队拥塞问题. 休假排队, 作为解决服务运行的有效机制, 可以充分利用闲期, 从事其他工作, 减少资源损耗, 降低系统成本, 获得更多关注.

休假排队研究中, 通常假定休假的服务员完全停止为顾客服务, 或者理解为完全停止原有工作, 但可以从事其他辅助工作或进行保养维修. 在一些情况下, 这是比较合理的, 但随着计算机通信网络、柔性制造系统、异步传输模式及电子商务等高新技术的发展, 出现了大量的复杂系统设计和控制问题. 休假的存在, 使得系统在休假期的负载过大, 如在管理机构中, 完全休假可能造成大批或大宗业务办理的延误; 通信网络中数据和信号的传输和接收也会由于波长或通道的限制而造成阻塞. 计算机科学和信息技术的发展, 激发了人们对休假排队系统的不断关注和探索, 工作休假排队研究正是在这种背景下产生和发展起来的.

文献 Servi 和 Finn(2002) 是工作休假 (Working Vacation, WV) 排队分析的起点. 与经典休假排队不同, 工作休假排队是一类半休假策略, 指在休假期间服务员将以较低的速度接待顾客, 而不是完全停止对顾客的服务. 这意味着, 在休假期间仍保持一部分服务能力用于接待顾客, 另一部分服务能力可进行缓冲调整或用于从事其他辅助任务. 这种系统实质上是两种不同服务速率交替运行的系统, 通过设立高速与低速期, 有效地解决了完全休假所造成的工作延误, 减轻了系统负载, 又保持了休假的原有效果.

工作休假排队研究迄今有十几年时间, 由于其更适用于服务系统建模和性能分析, 引起了大批排队论专家的关注, 并迅速产生了一些关于理论分析及实际应用的研究文献, 而田乃硕等的著作《离散时间排队论》(2008), 以一章的篇幅亦对离散时间工作休假排队的早期成果给出了处理. 然而, 或因研究时间较短, 或因模型的复杂, 工作休假排队的大量研究成果散见在各种杂志上, 并且通常是使用不同方法给出的, 相关文献和著作均未能反映工作休假排队研究的全貌, 更没有一本以建立完整理论体系为目标的专著.

本书的目的是对工作休假排队构建一个较完整的理论框架, 同时较全面地反映该领域的研究成果和方法. 20 世纪 70 年代以来, Neuts 等系统地发展了结构矩阵

分析方法, 为复杂随机模型的分析提供了强有力的工具, 在排队模型的研究中广泛使用矩阵分析方法, 已经成为当代流行趋势. 作者对国内外相关成果进行梳理和加工, 结合自己的研究工作, 基于结构矩阵分析方法, 给出了一个由浅入深、层次分明、有机联系的理论体系. 在内容上, 本书包括国内外最常见三类排队系统: M/M/1 型、GI/M/1 型和 M/G/1 型; 既包括经典模型的分析, 也包括带有各种休假策略、成批 Markov 到达、门限策略等新近研究成果的系统处理; 既包括连续时间排队, 也包括离散时间变体模型. 对各类型模型, 使用嵌入 Markov 链 (MC)、拟生灭 (QBD) 过程、GI/M/1 型和 M/G/1 型结构矩阵分析方法处理, 力争在研究内容和分析方法上反映工作休假排队研究的结构性和整体性. 同时, 这些复杂排队系统的分析也充分显示了结构矩阵分析方法的有效性. 书中还对每类工作休假排队模型给出详细的数值分析, 并包含通信网络性能分析的例子, 努力做到理论分析难度适中, 图文结合, 易于理解.

作者从事排队论研究已有十多年, 特别是在休假排队和离散时间排队分析中取得了诸多研究成果, 获得了国际和国内专家的认可. 书中大量内容均来源于作者本人的研究工作.

全书共 8 章. 第 1 章介绍排队论的基本内容和工作休假排队中出现的一些新现象和独有的特性. 第 2 章简要介绍本书中需要的预备知识, 包括泊松过程、离散时间 MC 和连续时间 MC、半 Markov 过程等内容. 第 3 章基于拟生灭过程方法处理各种工作休假的 M/M/1 型排队系统, 其中包括多重工作休假、单重工作休假、休假中断及门限策略. 第 4 章致力于工作休假的 GI/M/1 排队系统的分析, 采用矩阵几何解方法, 给出了多重、单重工作休假和休假中断排队的理论结果. 平行于第 4 章, 第 5 章处理了各种工作休假的 GI/Geo/1 型离散时间工作休假排队, 包括 Geo/Geo/1 型的拟生灭链排队. 第 6 章和第 7 章展示 M/G/1 型工作休假排队系统的分析, 采用 M/G/1 型结构矩阵方法, 给出单到达和批到达机制下工作休假排队的研究结果. 第 8 章介绍工作休假排队在光接入网络中的应用, 应用问题是尚处于发展过程中的工作休假研究领域, 成果相对较少, 但本章仍结合通信网络的应用例子, 基于工作休假机制, 提出解决网络运行的一个有效策略. 每一章的最后都设置文献评述一节, 给出本章内容和相关工作的出处, 便于读者阅读和研究进一步的文献.

2010 年以来, 作者关于排队论的研究及本书内容直接相关的成果, 得到国家自然科学基金 (编号: 71301091) 和教育部人文社科基金 (编号: 10YJC630114) 两个项目的资助, 作者本人入选山西省高等学校优秀青年学术带头人, 本书的出版也得到这三个项目的支持, 特此表示衷心的感谢.

燕山大学田乃硕教授、山西大学刘维奇教授, 始终关注着作者的研究进展和书

稿撰写, 给了很大的帮助, 谨向他们致以真诚的感谢! 感谢我的家人在书稿撰写中的付出, 他们的支持为我今后发展、前进的动力和希望!

由于作者水平有限, 不足之处在所难免, 恳请读者批评指正!

作　者

2015 年 9 月

于太原山西大学

# 目　录

# 第1章 引 论

## 1.1 排队模型

随着人类城市化和信息化进程的加快, 人们会发现自己经常陷入排队等待的烦恼之中. 路途中交通阻塞, 我们不得不在队伍中等待交警的疏通; 超市收银台繁忙, 我们不得不在队伍中等待收银员结账, 等等, 诸如此类. 除上述有形的排队之外, 还有大量的所谓“无形”排队现象, 如网络服务堵塞; 车站、码头等交通枢纽的车船堵塞和疏导; 通信卫星与地面若干待传递的信息, 等等, 均造成系统暂时的排队和等待.

排队现象面临的共同问题是: 增加服务设施无疑可以减少排队时间, 消除拥挤现象, 但可能会导致设备闲置, 造成设备投资浪费. 如果减少服务设施, 则常常造成排队和拥挤, 或者顾客会自动离去, 使系统丧失服务机会, 影响到经济效益的提高. 所以, 如何配备服务员的数量, 如何确定顾客的平均等待时间以及队伍的长度, 如何在服务配置方面达到平衡, 使得既不出现拥挤排队, 又不致发生设备闲置和浪费, 从而达到既提高服务质量, 又降低服务成本的目的, 就构成了排队论研究的目的.

排队论 (Queueing Theory) 又名随机服务系统理论 (Random Service System Theory), 是研究服务系统在运行过程中所产生的排队等待现象的一门数学理论, 是运筹学的一个重要分支. 具体地说, 它是在研究各种排队系统概率规律性的基础上, 通过研究各种服务系统在排队中的概率特性, 得到队长、等待时间等数量指标的变化规律, 解决相应排队系统的最优设计和最优控制问题.

在各种排队系统中, 起根本性作用的是它们的一个共同性质: 随机性. 顾客的到达间隔时间和顾客接受服务的时间中, 至少有一个具有随机性. 尽管排队系统是多种多样的, 但是从主要决定因素看, 它由三个部分组成: 输入过程、排队规则和服务机构; 为简便, 它采用 Kendell 引入的记法表示.

(1) 输入过程, 反映顾客来源以及顾客抵达排队系统的规律. 例如, 顾客来源是有限还是无限, 顾客是单个到达还是成批到达, 顾客相继到达的间隔时间之间是否独立, 顾客的到达间隔 $\{T_n, n \subseteq N\}$ 服从怎样的概率分布以及分布参数, 一般有指数分布、几何分布、一般分布等.

若在每个时隙点上最多到达一个顾客, 称为单个到达; 如果允许一个点处有许多顾客同时到达, 称为成批到达. 在更一般情况下, 还需要引入到达间隔有相依性

结构的到达过程. 例如, 连续时间 Markov 到达过程 (MAP).

(2) 排队规则, 是指服务系统中顾客接受服务的规则. 若顾客不愿意排队, 到达时系统只要有顾客, 即离开, 称为完全损失制排队. 若顾客排队等待, 又分为先到先服务 (First Come First Served, FCFS)、后到先服务 (Last Come First Served, LCFS)、随机服务 (Random Served, RS)、处理器共享 (Processor Sharing, PS)、有优先权的服务等. 通常采用等待制先到先服务 (FCFS) 的排队规则.

(3) 服务机构, 其刻画主要包括: ①服务台的数目 $k$(单个或多个, 在多个的情况下, 服务台是串联还是并联); ②服务机构的容量, 即最多可容纳的排队顾客数; ③是成批服务还是单个服务; ④每位顾客所需的服务时间 $\{S_n, n \subseteq N\}$ 是否独立, 服从什么样的概率分布以及分布参数, 一般有指数分布、几何分布、一般分布等. 本书集中于单服务台工作休假排队分析.

**例 1.1.1**　单服务台排队.

一个典型的单服务台排队, 顾客将依据一定的排队规则进行排队等待服务, 并由一个服务台服务顾客. 由于诸多服务系统的运行可以分解地看成多个单服务排队, 如通信网络中单条信道的传输、呼叫中心每个接线台的业务等, 单服务台排队在理论研究和应用领域获得最多关注 (图 1.1).

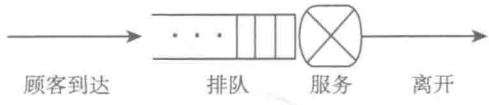


图 1.1　单服务台排队系统

**例 1.1.2**　多服务台排队.

典型的多服务台排队与单服务台排队的不同在于服务台的数量, 多服务台排队常用于模式服务台可提供相同服务, 即是无差别的, 如银行窗口、超市收银台、多台机器加工、边境海关等, 其更关注服务台数量的设计和控制问题 (图 1.2).

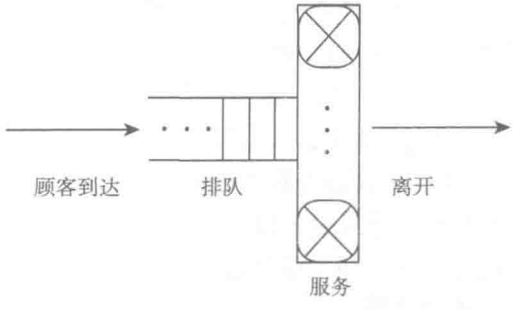


图 1.2　多服务台排队系统

(4) 符号表示. 一个排队系统是由许多条件决定的, 为了简明, 常采用 3 ~ 6 个

英文字母表示, 字母间用斜线隔开, 如

$$A/B/C/N/Y/Z$$

其中 A 表示输入分布类型, B 表示服务时间的分布类型, C 表示服务台的数目, N 表示系统的容量 (默认为 $\infty$), Y 表示顾客源中的顾客数目 (默认为 $\infty$), Z 表示服务规则 (默认为先到先服务 FCFS), 后三个一般省略不写, 其中 A 与 B 将用一些特定的符号取代, 通用记号如下:

D 表示定长分布, A=D 表明到达间隔为一个确定的长度 (正整数), B=D 表明顾客服务时间为定长;

M 表示指数分析, A=M 表明到达是泊松过程, B=M 表明服务时间服务指数分布;

Geo 表示几何分布, A=Geo 表明到达是 Bernoulli 过程, B=Geo 表明服务时间服从几何分布;

G 表示一般分布, 只假定到达间隔或服务时间为一般随机变量, 对其分布不加具体限制;

X 表示批变量, 根据批到达或批服务加于相应标记的右上角, 如 $M^X/G/1$ 表示到达过程是成批到达, 每批到达形成泊松过程.

常见类型排队系统如下:

M/M/1 表示输入过程是泊松流, 顾客来到间隔独立、同服从指数分布, 所需服务时间独立、同服从指数分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

Geo/Geo/1 表示输入过程是 Bernoulli 过程, 顾客来到间隔独立、同服从几何分布, 所需服务时间独立、同服从几何分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

M/G/1 表示输入过程是泊松流, 顾客来到间隔独立、同服从指数分布, 所需服务时间独立、同服从一般分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

Geo/G/1 表示输入过程是 Bernoulli 过程, 顾客来到间隔独立、同服从几何分布, 所需服务时间独立、同服从一般离散分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

GI/M/1 表示输入过程是一般过程, 顾客来到间隔独立、同服从一般连续分布, 所需服务时间独立、同服从指数分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

GI/Geo/1 表示输入过程是一般离散过程, 顾客来到间隔独立、同服从一般离散分布, 所需服务时间独立、同服从几何分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

GI/G/1 表示输入过程为一般过程, 顾客来到间隔独立、同服从一般分布, 所需服务时间独立、同服从一般分布, 系统中只有一个服务台, 顾客源容量为无穷的等待制排队系统.

M/M/$c$ 表示输入过程是泊松流, 顾客来到间隔独立、同服从指数分布, 所需服务时间独立、同服从指数分布, 系统中有 $c$ 个服务台, 顾客源容量为无穷的等待制排队系统.

M/G/$c$ 表示输入过程是泊松流, 顾客来到间隔独立、同服从指数分布, 所需服务时间独立、同服从一般分布, 系统中有 $c$ 个服务台, 顾客源容量为无穷的等待制排队系统.

由于 GI/G/1 和 M/G/$c$ 型排队的一般性与复杂性, 研究成果相对较少, 本书的内容也是仅集中于单服务台排队模型.

(5) 数量指标. 顾客与服务机构考虑到自己的利益, 对排队系统中的指标: 队长与等待队长、等待时间与逗留时间、服务台的忙期与闲期比较关系. 因此, 这三组指标就成了排队论的主要研究内容.

(i) 队长与等待队长. 令 $\{L(t), t \geqslant 0\}$ 为时刻 $t$ 系统中的顾客数, $L$ 为其稳态极限, 是指达到稳定后在系统中的顾客数 (包括正在接受服务的顾客), 称为队长, 而等待队长 $L_w$ 是指系统中排队等待的顾客数, 它们都是随机变量. 显然, 队长等于等待队长加上正在被服务的顾客数.

(ii) 等待时间与逗留时间. 顾客在系统中的等待时间 $W$ 是指从顾客进入系统到开始接受服务的这段时间, 而逗留时间 $S$ 是顾客在系统中所用时间的总和, 即等待时间与服务时间之和.

(iii) 忙期与忙循环. 系统的忙期 $D$ 是指从顾客到达空闲的系统时, 服务马上开始, 直到系统再次变为空闲为止的这段时间, 它是系统连续处于繁忙状态的时间长度, 反映系统中服务员的工作强度. 而系统的闲期 $I$ 与忙期对应, 指系统连续保持空闲状态的时间长度. 通常在统计平衡状态下, 忙期与闲期交替出现. 因为服务过程是循环进行的, 忙期和闲期组成了一次忙循环 $C$.

为方便, 提供几种在本书中经常用到的概率分布及相关参数 (表 1.1), 其中

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad 0 \leqslant k \leqslant n.$$

**表 1.1　几种常用概率分布**

| 分布 | 密度函数 | 取值范围 | 参数 |
|---|---|---|---|
| 指数分布 | $\lambda \mathrm{e}^{-\lambda t}$ | $t>0$ | $\lambda>0$ |
| $\Gamma$ 分布 | $\dfrac{\lambda(\lambda t)^{k-1}}{(k-1)!}\mathrm{e}^{-\lambda t}$ | $t>0$ | $k \in N, \lambda>0$ |

续表

| 分布 | 密度函数 | 取值范围 | 参数 |
|---|---|---|---|
| 泊松分布 | $\frac{\lambda^n}{n!}e^{-\lambda}$ | $n \in N_0$ | $\lambda > 0$ |
| 几何分布 | $(1-p)p^n$ | $n \in N_0$ | $p \in [0,1]$ |
| 二项分布 | $C(n,k)p^k(1-p)^{n-k}$ | $0 \leqslant k \leqslant n$ | $n \in N_0, p \in [0,1]$ |

## 1.2 休假排队

排队论的工作起源于丹麦工程师 Erlang(1909) 关于电话交换机使用情况的研究. 之后随着高新技术的发展, 提出了大量的复杂系统设计和控制问题. 这些系统的行为通常依赖于随状态而变化的参数, 经典排队模型在处理这类问题时表现出极大的局限性. 大部分研究都集中于休假排队, 即当服务系统中仅有少量顾客时, 停止原有工作, 引入一段通常理解的休假期, 休假完成再开始正常工作. 利用闲期对服务设施进行调整维修, 或者服务员在闲期中休假, 或者从事辅助性工作的排队系统都称为休假服务系统. 它泛指服务台在某些时候不能被顾客利用, 而暂时不能用于接待顾客的那些时间统称为休假. 诱发休假排队的问题多种多样, 如传统意义上的休假、设备保养与机器故障、启动时间、辅助工作、轮询服务等. 这种休假排队可以充分利用闲期, 从事其他工作, 减少资源损耗, 降低系统成本, 进一步达到了排队论研究的目的.

休假排队的概念最初产生于 20 世纪 70 年代, Levy 和 Yechiali(1975) 从有效利用系统闲期的观点出发, 引入了 "休假" 和 "休假策略" 等术语, 首先研究了 M/G/1 型休假排队系统. 到目前为止, 研究较多的是单服务台的 M/G/1 型和 GI/M/1 型排队系统. 关于经典休假的论述, 具体可见田乃硕 (2001).

### 1.2.1 休假机制

休假排队最重要的是将休假机制引入到经典排队中. 按照休假开始的规则, 可将休假策略分为空竭服务和非空竭服务两类. 前者表示服务一旦开始, 就要持续到系统中没有顾客为止, 休假只能从系统变为空闲状态时开始, 而后者是指系统中有顾客时也可以休假的情况. 本书考虑以下相关机制.

1. 多重休假 (Multiple Vacation, MV)

多重休假规则, 是指当系统内无顾客时, 服务员开始休假, 当休假结束时, 若系统中至少有一位顾客等待, 则服务员立即开始为顾客服务, 直到系统无顾客时又去进行新的休假; 若系统中仍没有顾客等待, 服务员就接着开始再一次新的休假. 其特点是任何时刻服务员只处于 "工作" 或 "休假" 两种状态之一.

2. 单重休假 (Single Vacation, SV)

单重休假是指系统变空时, 服务员开始休假, 当休假结束时, 若系统中至少有一位顾客等待, 服务员就立即开始为顾客服务, 直到系统再次变空; 若系统中仍然没有顾客, 服务员进入闲期, 直到顾客到来. 在单重休假机制下, 服务员处于忙期、假期和闲期三种状态之一. 虽然服务员 "休假" 和 "空闲" 时, 均不为顾客服务, 但这两种状态是不同的. 若服务员处在休假状态, 即使有顾客到达, 也不接待顾客, 要等到休假结束后服务才开始. 而若服务员处于空闲状态, 一有顾客到达就立刻开始服务.

3. 休假中断 (Vacation Interruption, VI)

休假中断是指休假期间, 系统的稳态指标达到特定值, 服务员可以随时转回到正常工作, 而非继续休假, 休假发生了中断. 之前的研究一般都假设系统只能在完成一次完整休假的前提下, 才能回到正常工作. 实际上, 系统经常有突发现象的发生, 比如, 银行等各类机构经常发生大宗业务到达, 完整休假可能会造成这些大宗业务无法及时有效的处理.

4. 门限策略

门限策略是指系统的顾客数达到特定值时, 即设定了一个门限值, 服务员从闲状态或休假转化为正常工作, 为休假中断策略的一种特殊情形. 通过设立门限策略, 可以更有效地利用服务系统.

诸多研究者给出很多与休假策略相关的其他策略, 如启动策略是指服务员需要一段启动时间才能转回正常工作, 见 Tian 和 Zhang(2006); 重试策略是指服务员忙时, 顾客进入重试区域. 一段时间后再次要求服务不断进行重试, 直到重试成功见 Artalejo 和 Gómez(2008) 与 Wang 等 (2001); 可修策略是指服务台出现故障进行修理, 见 Wang 等 (2002); 负顾客策略是指到达的负顾客抵消队尾的正顾客, 见朱翼隽等 (2004).

### 1.2.2　随机分解规律

休假排队系统理论的核心内容是随机分解 (Stochastic Decomposition), 即休假排队系统中队长、等待时间等稳态指标, 通常可以分解成两个独立随机变量之和. 其中一个是对应经典无休假系统中的同名指标, 另一个是由休假引起的附加随机变量, 称这样的结果为随机分解规律. 各种随机分解结果及导致这些结果的方法, 是休假排队研究的一个突出特色. 随机分解使休假排队与经典排队的比较一目了然, 便于分析各种休假策略对经典排队模型的影响.

对一个经典的单服务台 GI/G/1 排队系统中, 以 $L, Q, W$ 分别表示其稳态下系

统中的顾客数、排队等待的顾客数及等待时间. 引入某种休假策略后, 用添加下标的 $L_v$, $Q_v$, $W_v$ 表示休假排队系统中的对应稳态随机变量. 并以 $L(z)$, $Q(z)$, $L_v(z)$, $Q_v(z)$, $W^*(s)$, $W_v^*(s)$ 表示对应的母函数和拉普拉斯–斯蒂尔切斯变换 (Laplace-Stieltjes Transform, LST). 在这些符号下, 随机分解结果可表示为

$$\begin{aligned} L_v &= L + L_d, \quad & L_v(z) &= L(z)L_d(z); \\ Q_v &= Q + Q_d, \quad & Q_v(z) &= Q(z)Q_d(z); \\ W_v &= W + W_d, \quad & W_v^*(s) &= W^*(s)W_d^*(s), \end{aligned}$$

其中 $L_d, Q_d$ 统称为 (由休假引起的) 附加延长, $W_d$ 称为 (由休假引起的) 附加延迟, 而 $L_d(z), Q_d(z)$ 及 $W_d^*(s)$ 是对应的母函数和 LST.

随机分解规律揭示了排队稳态指标的特征和性质, 在休假排队研究中占有非常重要的地位. 本书将展示各种各样的随机分解定理, 作为工作休假排队的基础理论.

### 1.2.3 工作休假排队

2002 年, Servi 和 Finn 引入了一类半休假策略, 指的是在休假期间服务员将以较低的速度接待顾客, 而不是完全停止对顾客的服务, 称为工作休假 (Working Vacation, WV). 这意味着, 在休假期间仍保持一部分服务能力用于接待顾客, 另一部分服务能力可进行缓冲调整或用于从事其他辅助任务. 从实践的角度看, 与经典的完全休假相比, 工作休假可以更有效地保证系统运行, 达到减少系统损耗和减轻系统负载的双重作用, 即系统中主要工作相对较少时, 可从事其他辅助工作.

工作休假与经典休假存在一些实质性的差异. 在经典休假排队中, 休假的服务员完全停止为顾客服务, 因此, 休假期间没有顾客输出. 在工作休假排队中, 休假的服务员慢速接待顾客, 休假期间也可有顾客因完成服务而离去. 这导致工作休假排队要比经典休假排队表现出更加复杂的结构和行为, 其建模和分析也更加困难. 如果工作休假期间服务速率退化为零, 就回到了经典休假排队. 因此, 工作休假排队是经典休假排队的一种推广, 工作休假策略允许服务员灵活地在高低速服务转换, 使得服务系统更加接近实际运行, 应用更广泛. 下面列举几个可以模式成工作休假排队的实际问题.

**例 1.2.1** 网络数据传输.

Servi 和 Finn(2002) 指出在光纤通信系统中, IP 网关数据传输可采用轮询队列模式: 令牌在网络中轮流经过 $N$ 个队列. 对于无令牌的队列, 具有固定波长, 数据将以正常速率传输. 如果队列 $i$ 具有令牌, 此队列将被赋予一段附加波长, 将以更高的速率传输数据. 一旦令牌转移到别的队列, 此队列将转为正常运作水平. 某一队列的运作模式即为一个工作休假排队. 类似地, 在信号系统中, 可采用预留部分信道率的方式, 设立高速与低速期, 减少系统损耗. 具体应用分析将在第 8 章展示.

**例 1.2.2**　服务窗口设置.

银行及呼叫中心等服务类型机构经常需要不断地调整服务窗口的数量. 当业务量或呼叫需求相对较多时, 即高峰期, 服务窗口或呼叫台全部开放或大部分开放, 反之可以开放较少, 其余服务窗口或呼叫台从事机构内部的核算或整理等业务. 如果把机构看成一个排队系统, 开放较多的窗口相对有较高的服务率, 较少的窗口数量相对有较低的服务率, 构成了一个工作休假排队. 类似的服务设置问题也发生在电子商务、城市公交系统及高速收费站等.

**例 1.2.3**　服务运行设计.

在经典排队系统中, 无论是否存在顾客, 服务机构将以固定速率提供服务. 显然, 从经济的角度, 这并非是最佳选择. 如果存在较多的顾客时, 服务系统需要提高服务速率, 反之降低服务速率以达到较高的资源利用率和节省成本的目的. 当然, 降低服务速率可能会造成等待时间的增大及顾客满意度的降低等问题, 因此如何选择恰当的服务速率是机构关心的问题, 也是服务运作的设计问题. 这类以不同服务速率运行的排队系统的稳态设计, 可对城市供电系统、管理机构、交通系统等随机网络的分析和设计提供新的工具.

## 1.3　文献评述

经典休假排队经过了大约三十多年的发展, 建立了以随机分解结构为核心的完整稳态理论框架, 并形成了几种有效的稳态理论分析方法: 一是 Kendall(1953) 提出的嵌入 Markov 链 (Embedded Markov Chain, EMC) 方法并被发展到马氏更新过程. 这种方法的关键是寻找过程的再生点或嵌入点, 运用马氏链的技巧或建立更新方程来得到系统的统计特征量. 二是 Cox(1955) 提出的补充变量方法. 该方法通过增加变量, 构造向量马氏过程, 从而建立密度演化方程, 并求解各种统计特征量. 三是 Neuts 提出的矩阵解析方法, 其中包括拟生灭过程理论、M/G/1 型和 GI/M/1 型结构矩阵理论, 并将生灭过程的方法加以推广和引申, 逐渐形成了一套可以处理一系列相关于多服务台和 PH, MAP(马氏到达) 及 BMAP(批马氏到达) 分布所构成的排队模型, 具体方法可见 Neuts(1981,1989) 及 Breuer 和 Baum(2005). 各种方法在具体分析休假排队时交替使用, 取得了丰富的成果, 而这些理论结果也不断地应用在现实生活中, 对实际问题起到非常大的指导作用.

Neuts(1981) 给出的矩阵几何解方法, 可以非常有效地解决 M/M/1 型和 GI/M/1 型工作休假排队系统, Neuts(1989) 给出了解决 M/G/1 型排队系统的矩阵解析方法, 但一般情况下很难得到系统的解析表达式, 而在 M/G/1 型工作休假排队系统中, 采用此种矩阵解析方法, 可以得到非线性矩阵方程的最小非负解, 从而确保了系统各种稳态指标的解析表达式的得到. 矩阵解析方法的理论将在每种不同