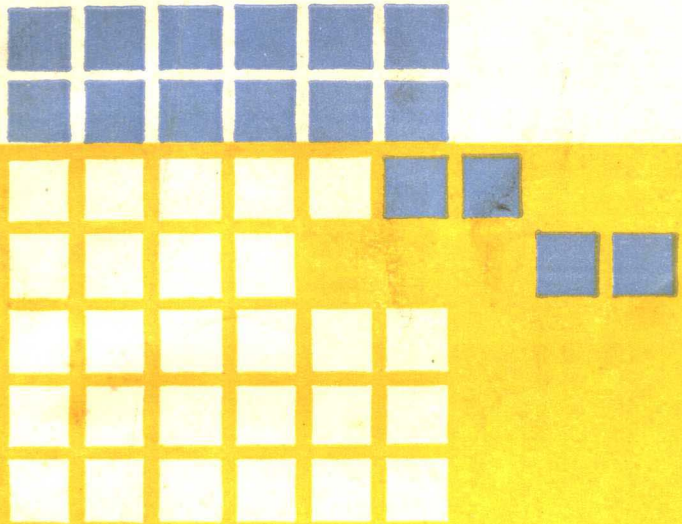


排队模型及其应用

罗荣桂

中理工大学出版社

排队模型及其应用

罗 荣 桂

责任编辑 李立鹏

*

华中理工大学出版社出版发行

（武昌喻家山）

新华书店湖北发行所经销

华中理工大学出版社沔阳印刷厂印刷

*

开本：787×1092 1/32 印张：9.25字数：200 000

1990年12月第1版 1990年12月第1次印刷

印数：1-1 000

ISBN 7-5609-0459-9/O·72

定价：1.85元

内容提要

本书从实用的角度出发介绍了排队论的基本理论，着重介绍了几种常用的排队模型。其内容包括单服务（多服务）机构指数服务模型、单服务机构一般服务模型、单服务机构一般到达型模型、有限源排队模型和爱尔兰排队模型等，同时还简要地介绍了排队系统的优化和排队论的应用。全书阐述严谨，深入浅出，读者只须有微积分和概率论的知识就可以阅读本书。

本书可作为运筹学、应用数学专业的高年级学生以及系统工程、管理工程、计算机等方向的研究生的教材或参考书，也可供有关的管理人员和工程技术人员阅读和参考。

序

排队论是研究拥挤和排队或者一般说的随机聚散现象的一门学科。所论对象从系统科学的观点看，可称之为随机服务系统，它是运筹学与系统科学的一个重要分支。

与任何一门有生命力的学科一样，它所研究的问题无不具有强烈的实际背景，致使其理论的构成与展开具有数理学科的一般特征，即首先从所选择研究现象的观察开始，然后由现象自在规律表现的某些固有特性，进行逐步的概念性升华，建立一些理论方法体系以数学模型及其求解为主要内容；将对象作定量而确切的刻划描述，并力求解释所论现象的某些较为全面的规律性，最后还要求这种理论所引出的结果尚能为某种新的观察所证实。也正如爱因斯坦说的，科学必须以事实始，以事实终，不管它建筑在两者之间的理论结构是什么。科学家首先是观察家；其次，他试图用完全的普遍性来描述他所看见的东西，以及他期望在将来看到的东西；然后，他根据他的理论作出预言，并再一次让事实来检验。

排队论在它问世以来的50年间所走过的道路也正是这样。自从丹麦数学家Erlang在研究电话话务理论获得第一批结果以来，这一理论已显著发展了；尤其是Ledermann等人在转入研究系统的瞬时性质以后，使这一领域的研究进入了一个崭新的阶段。在研究过程中引入和使用了诸如生灭过程、补充变量、积分微分方程、嵌入马氏链、半马氏分析、更新论、组合方法和扩散近似等理论中的方法概念与分析技巧之后，作为一种特殊随机过程论的排队过程理论便日趋成熟了，而且已经取

得了一系列完美的结果。与此同时，其应用已深入到交通运输、机器管理、计算机优化设计、电子对抗、水库设计、系统可靠性和军事等众多领域。目前，随着计算机与通讯网络的复杂化，这一学科的应用前景将更为广阔。特别是在它与最优化方法相结合后，就可成为上述诸领域内强有力的管理工具。

本书作者在多年的教学和科研的基础上，从更有益于初学和应用的角度出发，以《排队模型及其应用》为题，介绍了本学科的基本内容。全书内容比较系统全面，深入浅出，脉络分明；以模型的提炼和论证为主，易于初学。因此，无论是初学者、实际工作者，还是专家，都可以从本书中获得必要的知识，且有所裨益。

排队论是一个有丰富前途的运筹学方向，乐之为序。

李国平

1989年8月28日

前 言

排队论是20世纪初由丹麦数学家Erlang应用数学方法在研究电话话务理论过程中而发展起来的一门学科。因此它起源于电讯系统，是运筹学的一个重要分支。排队论亦名随机服务系统理论，主要研究具有随机性的排队模型的几个重要数量指标的统计规律，即研究排队过程的一些整体性质；同时，还探讨有关排队系统的最优设计和相应的随机过程的统计分析。

由于排队论所研究的问题具有很强的实际背景，因而它在国民经济的各个领域里都有着广泛的应用。在近80年的发展历史中，排队论不仅在理论方面而且在应用方面都已有了飞跃的进展；特别是60年代后，随着电子计算机的迅速发展，给这一学科展现了未来的更为广泛的应用前景。排队论和最优化方法与计算机的紧密结合必将成为一个强有力的现代化管理工具。在我国，排队论的理论研究虽然仅在50年代末才开始，但已有许多学者在排队过程的瞬时性质的研究中作出了很多漂亮的结果；尤其是在应用方面，它已应用于电讯、纺织、矿山、交通、机器维修、可靠性、计算机设计和军事等领域，并且取得了显著的成绩。随着我国改革的全面深化、决策的科学化和计算机的迅速发展，我们深信排队论的应用范围将会愈来愈广泛和深入，同时也必将进一步促进这一学科本身的理论发展。

本书试图对排队论中一些常用的排队模型作一概括性介绍。力求做到叙述简洁易懂，概念准确；对所引用的理论特殊的均有交代，定理的引入前后联系，有本有源，使读者便于接受而不致突然。但为不使篇幅过长，书中对于大多数排队模型

的瞬时性质未能涉及；同时，由于排队模型千变万化，其应用方面的参考文献数以万计．因此在本书中无论是理论或应用方面均不可能面面俱到，甚至挂一漏万，需要深入了解的读者请参阅其它有关的书籍和资料。

本书是作者根据多年来在给系统工程和管理工程等方向研究生讲课时所用的讲义和讲稿改编而成．全书共七章，第一章介绍排队论中的基本概念和有关的预备知识；第二至六章分别介绍几个常用的排队模型以及有关的数量指标；最后一章主要介绍排队系统的优化决策和应用举例．本书可作为高等工科院校管理工程、系统工程、自动控制、计算机等专业的研究生以及应用数学、运筹学专业的高年级学生的教材或参考书；也可供上述有关专业的教师、工程技术人员及企业管理干部阅读．阅读本书需要微积分、线性代数和概率论与随机过程等基础知识。

在原讲义编写过程中，作者曾得到中科院应用数学所已故研究员董泽清的帮助，本书不少地方参阅了他的著作《排队论及其应用》；同时书中还参阅了徐光辉研究员的名著《随机服务系统》．有些章节还采用了他们的写法和证明，这里不一一指出，作者谨此向他们表示深切的敬意．本书初稿承蒙冯文权教授的详细审阅，并提出了许多宝贵意见，中国科学院学部委员、前辈科学家李国平教授特为本书作序。对此，作者对他们谨致诚挚的谢意。最后，我特别要对华中理工大学出版社所付出的辛勤劳动表示衷心的感谢。

由于作者学识浅薄，书中错误或不妥之处自属难免，恳求读者惠予指正。

罗荣桂　1988年12月于武汉

目 录

第一章 导论及预备知识 …… (1)
§1 基本概念 …… (1)
§2 排队论中常见的分布与排队模型的符号表示 …… (7)
§3 负指数分布与最简单流 …… (13)
§4 一个基本公式及其证明 …… (32)
§5 生灭过程 …… (38)
第二章 $M/M/1$排队模型 …… (44)
§1 $M/M/1$排队模型的队长 …… (44)
§2 $M/M/1$排队模型的等待时间 …… (52)
§3 $M/M/1$排队模型的输出过程 …… (60)
§4 $M/M/1$排队模型的几个特例 …… (65)
§5 $M/M/1$排队模型的瞬态性质 …… (79)
第三章 $M/M/c$排队模型 …… (100)
§1 $M/M/c$等待制排队模型 …… (100)
§2 $M/M/c$排队模型的几个特例 …… (110)
§3 串联排队模型 …… (122)
§4 有限源的$M/M/c$排队模型 …… (131)
第四章 Erlang排队模型 …… (143)
§1 $M/E_k/1$排队模型 …… (143)
§2 $E_k/M/1$排队模型 …… (154)
§3 $E_m/E_k/1$排队模型 …… (169)
第五章 $M/G/1$排队模型 …… (176)
§1 $M/G/1$系统的嵌入马氏链 …… (176)
§2 $M/G/1$排队模型的稳态性质 …… (187)
§3 $M/G/1$排队模型的等待时间 …… (197)

§4 $M/G/1$排队模型的忙期 …… (208)
第六章 $GI/M/1$排队模型 …… (214)
§1 $GI/M/1$系统的嵌入马氏链 …… (214)
§2 $GI/M/1$排队模型的队长 …… (231)
§3 $GI/M/1$排队模型的等待时间与忙期 …… (237)
§4 “随机服务”的$GI/M/1$排队模型的等待时间 …… (242)
第七章 排队系统的优化及应用简述 …… (247)
§1 排队系统的优化决策模型 …… (247)
§2 排队论的应用与发展简述 …… (254)
§3 数据处理机的设计 …… (257)
§4 双机工作维修能力的配置 …… (270)
附表Ⅰ $M/M/1/k$系统ρ的最优值 …… (281)
附表Ⅱ $GI/M/1$系统β作为ρ的函数表 …… (282)
参考文献 …… (282)

第一章　导论及预备知识

§1　基本概念

1．引言

“排队”是指在服务机构处要求服务对象的一个等待队列；而“排队论”则是研究各种排队现象的理论，它是运筹学的一个重要分支．在排队论中，我们把要求服务的对象称为“顾客”，而将从事服务的机构或人称为“服务台”或“服务员”．在顾客到达服务台时，可能立即得到服务，也可能要等待到可以利用服务台的时候为止．

人们在日常工作和生活中，常常遇到各种各样的排队现象，构成顾客-服务台结构的排队系统．例如，上、下班的工人乘坐公共汽车，这里工人是顾客，公共汽车是服务员；病人去医院看病，病人是顾客，医生是服务员；一台发生故障的机器需要修理，机器是顾客，而修理工是服务员，如此等等．这些例子说明，“顾客”与“服务员”这两个名词可以从不同的角度去理解．

队列除了有形的还有无形的．例如，有许多顾客同时打电话给车站订购车票，当其中一个顾客正在通话时，其它顾客只好在各自的电话机旁等待通话，他们尽管分散在各处，但却形成一个无形的队列．其次，排队的顾客不一定是人，也可以是物；同样，服务员也不一定是人，可以是物．如急需降落的飞机

因跑道不空而在空中盘旋；大海中的船舶等待靠岸等等都是这种例子。

在上述顾客-服务员组成的排队系统中，顾客到来的时刻与服务员进行服务的时间一般说来都是随不同的时机与条件而变化的，往往无法预先确知。因此，系统的状态是随机的。为了强调其随机性，我们有时称排队论为“随机服务系统”。随机性是排队系统的一个共性。

2. 排队系统的基本特征

上述列举的排队现象，尽管各式各样，然而不难看出，它们有一些共同的基本特征。

(1) 输入过程：即各种类型的顾客以怎样的方式到达。如果到达的出现和服务的提供是严格的按照时间表进行的，那么排队现象就可以避免。然而，这类到达间隔实际上并非常见，大多数场合是顾客的到达由外部因素来决定的。因此，理想的方法是用随机变量来表示输入过程。顾客的到达源、到达的类型和相继到达的间隔时间等三者完全刻划了一个输入过程。

(2) 服务机构：即在同一时刻有多少个服务台（或服务员）为顾客服务；每一个顾客在此服务机构里接受服务需要多少时间。在服务机构中，包含不确定的因素是服务员的数目，在任意时刻接受服务的顾客数，服务时间和服务方式等。在网络流和排序问题中，常常需要处理多于一个服务台（员）的串联或并联形式的排队模型。对此，我们同样需要用随机变量来表示。

(3) 排队规划：即到达的顾客以怎样的规则接受服务。一般地采取“等待”、“消失”以及一种介乎二者之间的混合式方法加以解决。象“先来先服务”、“后来先服务”和“随

机服务”等这些规则是不解自明的．例如，公共汽车的乘客按其到达先后次序上车；存放在仓库中的物资按从上到下，后到先搬出仓库；电话用户随机地拨通电话．在许多场合，需要引进“优先权服务”，以使构成的系统更加符合实际．例如发送电报要让“加急”电报先发送；急诊病人要比一般病人优先得到服务．

关于排队规则，可分成三种情形．

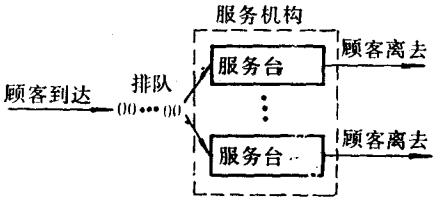
a） 损失制：当一个顾客到达时，若所有服务台均被占用，则该顾客以及更后的顾客都将自动离去且永不再来．例如，顾客拨电话，当线路被占用时就接不通，此时顾客也即自动消失．

b） 等待制：当一个顾客到达而所有服务台均被占用时，该顾客就列队等待服务．例如长途电话系统的顾客则是此种情形．

c） 混合制：这是上述两种排队规则的混合形式．当队长有限制时（例如系统只能容纳k个顾客），一顾客到达若队长小于k，则他列队等待；若队列已满，则他只好离去，且永不再来．当等待时间有限制时，例如顾客在队列中的等待时间不能超过t_0，若他到达时估计等待时间不会超过t_0，则他加入等待队列，否则他将离去且不再来．还有，当一个顾客的逗留时间（即等待时间加上服务时间）有限制时，若他估计到达至离开系统时不会超过此时间，则他列队等候服务，否则离去且不再来．

需要指出，损失制和等待制可以看成是混合制的特殊情形．正是因为任何排队系统具有上述三个共同特征（见图1.1），才使我们有可能建立处理这些问题的统一理论，即排队论．

(*a*)　单服务台



(*b*)　多服务台

图1.1　排队系统的结构

3. 排队系统的主要数量指标

排队系统的最重要数量指标有以下三个。

(1) 队长：队长是指系统中顾客（包括排队等候和正在接受服务的）的数目；排队长度则仅指在队列中排队等待的顾客数。无论队长或排队长度，都是顾客和服务台最关心的数量指标，特别是对系统的设计者来说，尤为重要。因为它涉及到系统等待空间的大小。

(2) 等待时间：等待时间是指从顾客到达时刻算起到他开始接受服务时止的这段时间；而逗留时间则是从顾客到达时刻算起到他接受服务完毕为止所需要的时间，即是顾客在系统中所花费的总时间。显然，一个顾客的逗留时间等于其等待时间与接受服务的时间之和。等待时间与逗留时间对顾客来说是最关心的，因为每个顾客都希望自己用于排队等待的时间愈

短愈好；同样，有时也希望自己在系统中的停留时间愈短愈好。例如，当一个顾客想利用乘飞机之前去机场商店购买礼物，他自然最关心的是买完礼品后能否赶得上班机。

(3) 忙期：忙期是指服务台连续繁忙的时间，即指顾客从到达空闲服务台起到服务台再次变为空闲时止的这段时间。这是服务台最关心的数量指标，因为它直接关系到服务员的工作强度。与忙期相对应的是闲期，这是指服务台连续保持空闲的时间长度。很显然，在排队系统中，忙期与闲期是交替出现的。

排队系统除了上述三个主要数量指标外，服务台的利用率(即服务员忙碌的时间在总时间中所占的比例)在排队论的研究中也是很重要的指标。在特殊的排队模型中，甚至还有一些特殊的数量指标。由于在输入过程和服务机构中随机性的假设，因此上述各种指标均是相应的状态和参数空间的随机过程或随机变量。这些随机过程和随机变量的分布性质和矩是排队模型中研究的主要目标。

需要指出，在排队系统中只要有需要服务的顾客，服务台就得保持工作。当一个顾客接受服务，就直到服务完成。排队规则不象某些优先权那样的动态型的情形，其排队长度与排队规则无关。然而，等待时间显然是依赖于排队规则的。

4. 研究的内容、目的与方法

由于排队现象具有随机性，因此研究排队系统的首要内容是排队模型的数量指标的概率规律，即研究排队过程的一些整体性质，进而研究系统的最优化问题及相应的随机过程的统计推断工作。

在一个服务系统中，由于要求服务的顾客数目常常总是多

于可用的服务台的数目，因而才会产生排队现象；如果服务员的人数（或服务速率）能有充分保证，那末排队现象就可以避免，至少队伍将不至于排得冗长。因此，对顾客来说，服务台（或服务员）越多，接受服务就愈方便；而另一方面，服务台越多或服务速率越高，服务机构的人力、物力和财力的支出就会越大，甚至造成不必要的浪费，这也是不经济的。于是，在服务系统中将出现顾客的等待与服务台的数量（或服务速率快慢）之间的均衡协调问题。系统分析的任务是根据有关系统的运行特征，调整系统的有关参数（如服务速率、服务台数量或等待空间大小等）以保证从顾客和服务员两方面来看都能得到更有效的利用。对于许多实际排队模型，其关键在于确定服务率或服务台数目或选取顾客接受服务的规则，使得在某种意义（通常以某种成本分析）下达到最优；当然，也可以是确定顾客的到达率或别的量的某种组合。总之，其目的是要寻求排队模型中某些参数的最优值，以使系统在某种意义下达到最优。

关于排队系统的优化问题有两类：一类是在系统设置之前，根据过去已有的资料（顾客的输入与服务过程）对系统未来的发展作出正确的估计或预测，以便使设计工作有所依据，此类优化称为**静态优化**。例如工厂厂房大小、码头泊位数目、城市交通路面尺寸等的设计就是如此。另一类是对现有的服务系统，设计者根据顾客输入的变化而对服务机构进行适当的调整，这种对已有系统的最优运营称为**动态优化**（或称为系统的实时控制）。例如，春节时间因旅客流量突增时，车站就设法增加服务员和临时增开加班客车，以减少旅客的等车时间。

如何合理地设计和最优运营一个服务系统，使之既能尽量满足顾客的要求，又能使系统的花费最为经济（或系统的收益最大），这个问题是现代管理决策者面临的一个极其重要的问

题，也是我们研究排队论的最终目的。

排队论本身不是一种最优化方法，而是一种分析工具，它可为有关问题提供更有效的资料。因此，处理排队的程序可以概括为如下四步：

（1） 确定表达排队问题的各个变量，并建立它们之间的相互关系；

（2） 根据现有的数据，运用适当的统计检验以得到有关的分布；

（3） 应用已得到的概率分布，确定描述整个系统的运行特征；

（4） 根据系统的运行特征，通过应用适当的决策模型，改进系统的功能。

随机因素在排队论中起着决定性的影响。因为在一般的服务系统中，顾客的到达与服务机构的服务时间都具有一定的随机性，也正是由于这样，才使排队论的研究变得复杂了。自然，在排队模型的研究中我们需要有研究随机现象规律性的一门数学理论，即“概率统计与随机过程”。

§2 排队论中常见的分布与排队模型的符号表示

1. 顾客到达的间隔时间分布

若以T_n表示第n（$n\geqslant 1$）个顾客到达系统的时刻（$T_0=0$），则在顾客单个到达的情形

$$0=T_0<T_1<T_2<\cdots<T_{n-1}<T_n<\cdots$$

设$\tau_n=T_n-T_{n-1}$（$n=1,2,\cdots$），则τ_n是第n与第$n-1$个顾客到达系统的时刻之差，我们称它为第n个到达间隔时间。一

般地，假定$\{\tau_n\}$是相互独立、同分布的随机变量序列，其分布密度与分布函数分别记为$a(t)$和$A(t)$。

在排队问题中，顾客的到达间隔时间常常具有以下几种概率分布。

(1) 定长分布：顾客有规则地等距时间到达，例如每隔时间α到达一个顾客，即有

$$A(t)=P\{\tau_n\leqslant t\}=\begin{cases}1, & \text{若}t\geqslant\alpha;\\ 0, & \text{若}t<\alpha.\end{cases} \tag{1.1}$$

这种输入的例子很多，例如产品通过传送带然后进入包装箱就是定长输入的典型例子。

(2) 负指数分布：顾客到达间隔时间服从负指数分布，即

$$a(t)=\begin{cases}\lambda e^{-\lambda t}, & \text{若}t\geqslant 0;\\ 0, & \text{若}t<0,\end{cases} \tag{1.2}$$

或

$$A(t)=P\{\tau_n\leqslant t\}=\begin{cases}1-e^{-\lambda t}, & \text{若}t\geqslant 0;\\ 0, & \text{若}t<0.\end{cases} \tag{1.3}$$

若$\{\tau_n\}$独立，且同为负指数分布(1.2)或(1.3)，则称顾客到达过程为服从参数λ（$\lambda>0$）的最简单流（或称Poisson流）。

(3) k阶爱尔兰（Erlang）分布E_k：顾客到达间隔时间$\{\tau_n\}$相互独立，且具有相同的Erlang分布密度

$$a(t)=\frac{\lambda(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t}, \quad t\geqslant 0, \tag{1.4}$$

则称$\{\tau_n\}$为服从k阶Erlang分布。Erlang分布的均值为$E[\tau_n]=k/\lambda$，方差为$D_k[\tau_n]=k/\lambda^2$。所以

$$k=\frac{E^2[\tau_n]}{D_k[\tau_n]}.$$

可以证明，Erlang分布E_k是k个独立，服从同一负指数分布