系统工程丛刊

运筹学

下册

[美] FREDERICK S. HILLIER
GERALD J. LIEBERMAN 著

赵孟养 主译

中国系统工程学会

《系统工程丛刊》出版说明

一、我们编辑出版《系统工程丛刊》，是为了及时交流国内外有关系统工程的理论研究和学术著作、系统工程的教育与普及用书和资料、系统工程在各方面应用的经验。内容兼顾学术研究、实际应用、教育和普及的需要。

二、本丛刊作为内部交流。稿件的著者、译者和编者可另行联系公开出版。

三、丛刊编辑工作由中国系统工程学会组织的编辑委员会负责，印刷出版由中国系统工程学会教育与普及工作委员会的挂靠单位上海机械学院负责。编委会设于上海机械学院系统工程研究所。

四、读者对本丛刊有什么意见和要求，欢迎写信给中国系统工程学会《系统工程丛刊》编委会。地址是：上海市军工路516号上海机械学院系统工程研究所。

中国系统工程学会

下 册 目 录

第二篇　概率性模型（续）

第二篇　概率性模型(续)

第9章　排队论

排队论无非是“蛇队”亦即等待线的数学研究。等待线的形成自然是一种常见的现象；只要当前对一项服务的需求超过提供该项服务的当前能力，就会出现这种现象。在工业及其他地方，常常必须就所要提供的能力的大小作出决策。可是，因为寻求服务的各单元将在何时到达及（或）提供该项服务将需要多少时间，往往不可能精确预测，所以这些决策往往是困难的。提供过多的服务不免要支出过大的费用。另一方面，不提供足够的服务能力有时会使等待线变得过长。过分的等待在某种意义上也是费用大的，无论这是社会费用、丧失顾客的代价、还是闲散雇员的损失、等等。所以最终目的是要在服务的费用与有关等待该项服务的费用之间达到经济上的平衡。排队论本身并不直接解决此问题；可是，凭借预测诸如平均等待时间之类的等待线特征，排队论的确提供这样一种决策所需要的重要资料。

排队论提出许多备选择的数学模型供等待线场合的描述。这些模型往往有预测等待线某些特征的数学结果可利用。这一章在一些概括的讨论之后给出大多数较初等的模型及基本结果。第10章则讨论怎样可应用排队论所提供的资料来作出决策。

9.1　典范实例

郡医院的急诊室对于用救护车或私人汽车送至医院的急症患者提供迅速的治疗。任何时候总有一个医生在急诊室值勤。可是，由于急症患者有日益增长的趋势即宁愿使用这些急救设施而不去看私人医师，该医院已察觉到急诊室就诊人数正在逐年增加。结果是在高峰使用时间（黄昏）到达的病人，不得不挨次等待医生来诊治，这已成为很常见的事情。所以有人建议在这段时间内应当增派第二个医生至急诊室，以便同时可看两个急症病号。此问题已指定医院的管理工程师加以研究。①

管理工程师在开始时先搜集有关的过去数据，然后把这些数据投入下一年的计划。认识到急诊室是一个排队系统，他就应用若干备选择排队论模型来预测该系统有一个医生及两个医生的等待特征，如读者将在本章较后几节中所看到的。

9.2　排队模型的基本结构

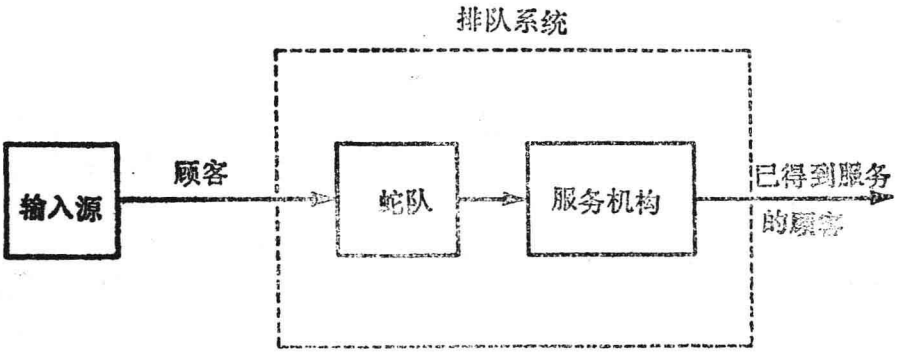
基本排队过程

下面是大多数排队模型所呈现的基本过程。需要服务的“顾客”是由一个“输入源”随着时

① 关于这种问题的一项实际个案研究，见 W. Blaker Bolling, “Queueing Model of a Hospital Emergency Room（医院急诊室的排队模型）”，《Industrial Engineering》，26—31页，1972年九月。

间产生的。这些顾客进入排队系统并参加一个蛇队。在某些时刻根据称为服务纪律的法则选出蛇队的一个成员作为服务对象。于是由服务机构为该顾客完成所需要的服务，该顾客随即离开排队系统。此过程已在图 9.1 中绘出。

图 9.1 基本排队过程



关于排队过程的各种要素，可作出许多备选择的假定；下面就来讨论这些要素。

输入源（呼叫总体）

输入源或呼叫总体的一个特征是它的“大小”。大小就是随时可能需要服务的顾客的总数，即各别潜在顾客的总数。可假定此数或无限或有限。由于无限情况的计算远更容易，即使在实际大小是某一相当大的有限数时也往往假定为无限。有限的情况在分析上较为困难，因为在任何时候排队系统内的顾客数影响到系统外的潜在顾客数。可是，如果输入源产生新顾客的速率受到排队系统内顾客数的影响，那么必须采用有限的假定。

还必须指明顾客随着时间而产生的统计模式。通常的假定是他们按照 Poisson 过程产生，即截至任何特定时间为止所产生的顾客数具有 Poisson 分布（见 8.6 节）。当排队系统的到达者“随机地”但按一定平均速率出现时，情况就是这样。一种等价的假定是介于相继到达者的时间具有指数概率分布（见 8.7 节）。介于相继到达者的时间称为**到达际时间**。

还必须指明关于顾客行为的任何异常的假定。一个例子是逡巡不前，即如果蛇队太长，顾客不愿进入系统，因而被放过。

蛇队

一个蛇队是以其所能包含的最大可允许的顾客数来刻划的。蛇队叫做无限或有限，视此数为无限或有限而定。

服务纪律

服务纪律是指选取蛇队成员给以服务的次序。例如，这可以是先到先服务的、随机的、或按照某种优先过程的、等等。排队模型通常不言而喻地假定先到先服务，除非另有说明。

服务机构

服务机构由一处或几处服务设施组成，每一设施含有一条或几条并联服务通道，叫做**服务者**。如果服务设施不止一处，顾客可从一系列设施（串联服务通道）得到服务。在一处给定的设施上，顾客进入并联服务通道之一且完全由该服务者得到服务。一个排队模型必须指明各设施的布置及在每一设施上服务者（并联通道）的数目。大多数初等模型假定一处服务设施，内有一个或有限个服务者。

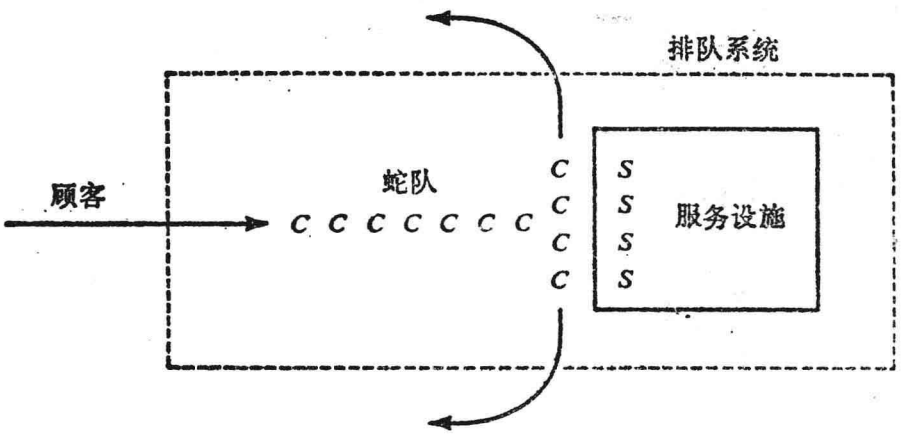
在一处设施上对一个顾客从服务开始至其结束所消逝的时间称为服务时间或占用时间。一个排队模型必须就每一服务者（且可能就不同类型的顾客）指明服务时间的概率分布，纵然通常对于所有服务者假定同样的分布。常常被选用的分布是 Erlang（加默）分布或其一种特殊情况、指数分布、及退化分布（恒定服务时间）。

一个初等排队过程

正如上面各小节所启示，排队论已探究了许多不同类型的等待线场合。然而它所集中注意的主要是下述场合：在驻有一个或几个服务者的单独一处服务设施前面形成（有时可能是空的）单独一条等待线。由输入源所产生的每一顾客大概在蛇队（等待线）中稍经等待后就由服务者之一得到服务。所涉及的排队系统已在图 9.2 中绘出。

图 9.2 一个初等排队系统

（顾客由C，服务者由S表示）



注意 9.1 节示范实例中的排队过程正是此类型的。输入源所产生的顾客就是需要治疗的急症患者。急诊室就是服务设施，而医生就是服务者。

服务者不一定是单独的个人；它可以是一群人，例如同时合力为一个顾客进行所需要服务的修理班组。而且，服务者甚至不一定是人。在许多情况下服务者可能是一台机器或一件设备，例如一辆随叫随到来完成某项服务的铲车（虽然不免要有人来指引）。同理，等待线

中的顾客也不一定是人。例如，它们可以是等待某种机器来进行某项加工的工件，也可以是守候在收税亭前的汽车。

也并非一定要实际上有一条物质的等待线排列在一个组成服务设施的物质结构的前面。这就是说，蛇队的各成员可以分散在一个区域内等候服务者来到它们那里，例如等待修理的一些机器。分派至给定区域的一个或一群服务者便组成该区域的服务设施。排队论将仍然给出平均等待中个数、平均等待时间、等等，因为顾客结成一群来等待与否，无关紧要。为使排队论可适用，唯一必不可少的要求是等待某项服务的顾客的数目应当发生变化，恰如图9.2所描述的物理场合（或类似的合法场合）。

所有在9.6至9.8节中讨论的排队模型都属于本节所描述的初等类型。

术语与记法

除非另有说明，今后将使用下述标准术语与记法：

系统的状态 = 排队系统内的顾客数。

蛇队长度 = 等待服务的顾客数

= 系统的状态减去正在得到服务的顾客数。

$N(t)$ = 在时间 t（$t \geq 0$）排队系统内的顾客数。

$P_n(t)$ = 在时间 t 排队系统内正好有 n 个顾客的概率，在时间0的个数已给定。

s = 排队系统内服务者（并联服务通道）的数目。

λ_n = 当系统内有 n 个顾客时新顾客的平均到达速率（期望每单位时间到达数）。

μ_n = 当系统内有 n 个顾客时全系统的平均服务速率（期望每单位时间完成服务顾客数）。注：μ_n 表示所有忙碌（服务着顾客的）服务者合并的完成服务速率。

如果 λ_n 对于所有 n 为常数，此常数便记作 λ；如果每忙碌服务者平均服务速率对于所有 $n \geq 1$ 为常数，此常数便记作 μ（于是当 $n \geq s$ 时 $\mu_n = s\mu$，因而所有 s 个服务者是忙碌的）。在这些情况下，$1/\lambda$ 与 $1/\mu$ 分别是期望到达际时间与期望服务时间。又 $\rho = \lambda/s\mu$ 是服务设施的利用系数，即各服务者都忙碌的期望时间比数，因为 $\lambda/s\mu$ 表示系统的服务能力（$s\mu$）中正在平均被到达顾客（λ）所利用的比数。

还需要一些记法来描述稳态结果。如果一个排队系统最近已开始运转，该系统的状态（系统内的顾客数）将大大受到初始状态及此后所消逝时间的影响。此时我们说系统处于瞬变情况。然而在已有足够的时间消逝后系统的状态成为实质上跟初始状态及消逝时间无关（异常的情况除外①）。系统现在实质上已到达稳态情况。排队论势必大都集中注意于稳态情况，部分的理由是瞬变情况在分析上更为困难。（有某些瞬变结果存在，但一般说来已超出本书的技术范围。）下述记法假定系统处于稳态情况：

P_n = 排队系统内正好有 n 个顾客的概率。

L = 排队系统内的期望顾客数。

① 当 λ 与 μ 已确定时，通常的要求是 $\rho < 1$，否则，系统的状态势必随着时间的过去而愈来愈大。

L_q = 期望蛇队长度。

W = 系统内的期望等待时间（包括服务时间）。

W_q = 蛇队内的期望等待时间（不包括服务时间）。

L与W之间的关系

假定 λ_n 对于所有的 n 为常数 λ。现已证明[①] 在稳态排队过程中

$$L = \lambda W。$$

而且，此同一证明也显示

$$L_q = \lambda W_q。$$

如果各 λ_n 不相等，那么在这些等式中 λ 可由长远平均到达速率 $\bar{\lambda}$ 来替换。（稍后我们将显示怎样可就某些基本情况确定 $\bar{\lambda}$。）

现在假定平均服务时间对于所有 $n \geq 1$ 为常数 $1/\mu$。于是可得出

$$W = W_q + \frac{1}{\mu}。$$

这些关系式是极重要的，因为它们使得所有四个基本量 L，W，L_q，及 W_q 中，只要有一个以分析法求得，便全部立即确定。这是幸运的，因为从基本原理来索解一个排队模型时，这些量中的某一些往往远比其他一些更易求得。

9.3 现实排队系统举例

我们在上一节中关于排队系统的描述可能显得比较抽象且仅适用于相当特殊的实际场合。恰恰相反，排队系统是在各种各样的现实环境中意想不到地处处可见的。为使读者扩大在排队论适用性上的视野，我们将简短地举出现实排队系统的各种例子。

我们在日常生活中谁都碰到的一类重要排队系统就是**商业服务系统**，其中各顾客得到商业组织所提供的服务。在这些系统中有许多涉及在固定地点上人对人的服务，如理发店（理发师是服务者）、银行出纳窗口、食品杂货铺的收款台、以及自动食堂线(串联服务通道)。然而另外有许多不是这样的，例如家庭用具修理（修理服务者上门到顾客家里）、自动售货机（因而服务者是机器）、及加油站（可把汽车看作顾客）。

另一类重要系统是**运输服务系统**。对这些系统的有一些说来，运输工具就是顾客，如等候在收税亭或交通管制灯前的汽车、等待搬运工(服务者)来装卸的卡车或船舶、及行将在跑道（服务者）降落或起飞的飞机。（这种系统的一个特例是停车场，在那里汽车是顾客，而停车场地是服务者，但并无蛇队，因为在场地已满时后到的顾客便驶向别处）。在其他情况下运输工具是服务者，如出租汽车、消防车、及电梯。

近年来应用排队论最多的，或许是**工商服务系统**。这些系统包括由材料输送装置（服务者）移动负载物（顾客）的材料输送系统；由维护人员（服务者）来修理机器（顾客）的维护系统；以及由质量检验员（服务者）来检验产品（顾客）的检验站。雇员福利设施及打字

① John D. C. Little, "A Proof for the Queueing Formula: L=λW（排队公式 L=λW 的一个证明）", «Operations Research», 9(3): 383—387, 1961; Shaler Stidham, Jr., "A Last Word on L=λW（关于 L=λW 的定论）", «Operations Research», 22(2)417—421, 1974.

合作社也都吻合这类系统。此外，还可以把机器当作服务者，其顾客则是被加工的工件。一个相关的极重要例子是把计算机看作服务者的计算机设施。

现在人们日益认识到排队论也可适用于**社会服务系统**。例如，司法系统是一个排队网络，其中法院是服务设施，审判员（或审判团）是服务者，而等候审讯的案件是顾客。立法系统是一个类似的排队网络，其中顾客是等待处理的国会议案。各种保健系统也是排队系统。读者早已在9.1节中看到一个实例（医院急诊室），但是还可以把救护车、X射线机、及医院病床看作其各自排队系统中的服务者。类似地，可把等待低收入与中等收入住房或其他社会服务的家庭看作排队系统中的顾客。

以上这些是四大类排队系统，但仍未把排队系统包罗无遗。事实上，排队论是在本世纪初最先从电话工程的应用开始的，而这仍然是一个重要的应用领域。此外，我们都有自己的个人蛇队，如指定的家庭作业、要阅读的书籍、等等。然而，这些例子已足以启示，排队系统确实渗透到社会的各个不同领域。

9.4 指数分布的作用

排队系统的运转特征在很大程度上取决于两个统计性质，即，到达际时间的概率分布（见9.2节输入源）及服务时间的概率分布（见9.2节服务机构）。在现实排队系统中这些分布几乎可取任何形式。（唯一的限制是不能出现负值。）可是，要构成一个排队论模型来表示现实系统，便必须指明这些分布各自所呈现的形式。所呈现的形式，要有效用，应当充分现实借以使模型提供合理的预测，而同时又充分简单借以使模型在数学上易于处理。在这些基础上，排队论中最重要的概率分布是指数分布。

假设随机变量T要末表示到达际时间，要末表示服务时间。（我们将把标志这些时间结束的现象——到达或服务完毕——称为事变。）应记得（见8.7节指数分布）如果T有概率密度函数

$$f_T(t)=\begin{cases}\alpha e^{-\alpha t}, & t\geq 0\\ 0, & t<0,\end{cases}$$

如图8.9所示，我们就说T具有参数为α的指数分布。① 此时，累积概率为

$$P\{T\leq t\}=1-e^{-\alpha t},\qquad (t\geq 0)$$

$$P\{T>t\}=e^{-\alpha t},$$

而T的期望值与方差为

$$E(T)=\frac{1}{\alpha},\qquad var(T)=\frac{1}{\alpha^2}.$$

假定在一个排队模型中T具有指数分布，这有怎样的含义？为探究此事，让我们来考察指数分布的五个关键性质。

性质1　$f_T(t)$ 是 t（$t\geq 0$）的一个严格减函数。

①为了跟排队论的习惯记法一致，参数α是8.7节指数分布中所使用的参数θ的倒数。

性质 1 的一个后果是对于 Δt 与 t 的任何严格正值有

$$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\}.$$

〔此关系式来自下述事实：这些概率是在曲线 $f_T(t)$ 的下面，位于长度为 Δt 的所指出区间之内的面积，而就曲线的平均高度而论第二个概率要比第一个小。〕所以 T 将取一个接近零的小值，不仅可能，而且比较有希望。事实上，

$$P\left\{0 \leq T \leq \frac{1}{2}\frac{1}{\alpha}\right\} = 0.393,$$

而

$$P\left\{\frac{1}{2}\frac{1}{\alpha} \leq T \leq \frac{3}{2}\frac{1}{\alpha}\right\} = 0.383,$$

因此 T 所取的值是“小”值〔即小于 E(T)的一半〕要比“接近”其期望值〔即并不离 E(T)更远些〕更有希望。

在排队模型中这果真是 T 的一个合理性质吗？如果 T 表示服务时间，答案取决于所涉及的服务的一般性质，有如下述。

如果所需要的服务对每一顾客基本上相同而服务者总进行同样的一系列服务操作，那么实际的服务时间势必接近期望服务时间。可能出现小的均值离差，但通常这只是由于服务者的效能稍有差别。远在均值之下的短服务时间实质上是不可能的，因为即使服务者正在以最高速度工作，也得有一定的极小量时间来进行所需要的服务操作。在此种场合，指数分布显然并不提供服务时间分布的精确逼近。

另一方面，考虑那种对服务者所要求的具体任务因顾客而异的场合。服务的大致性质可能相同，但具体的服务种类与份量却有区别。例如，在 9.1 节所讨论的郡医院急诊室问题中情况就是这样。医生们遇到各种各样的医疗问题。在大多数情况下他们能相当迅速地提供所需要的治疗，但偶尔也有病人需要延长的照料。类似地，银行出纳员与食品杂货铺收款员是其他属于此一般类型的服务者，其所需要的服务往往是简短的，但偶尔必须延长。对于这种服务场合，指数的服务时间分布看来十分可取。

如果 T 表示到达际时间，性质 1 排除了如下的场合，即趋向排队系统的潜在顾客因看到另一个顾客已先一步进入系统而迟迟不想进去。反之，它是跟“随机地”发生的普通到达现象，如随后各性质所描述，完全一致的。

性质 2　缺乏记忆。

此性质可用数学式表述为对于任何正量 t 与 Δt 有

$$P\{T > t + \Delta t \mid T > \Delta t\} = P\{T > t\}.$$

换句话说，不管有多少时间（Δt）早已过去，截至事变（到达或服务完毕）出现为止的剩余时间总具有相同的概率分布。实际上，该过程“不记得”它的过去。（在指数分布中出现此意想不到的现象是因为

$$P\{T > t + \Delta t \mid T > \Delta t\} = \frac{P\{T > \Delta t, T > t + \Delta t\}}{P\{T > \Delta t\}}$$

$$= \frac{P\{T>t+\Delta t\}}{P\{T>\Delta t\}} = \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha\Delta t}} = e^{-\alpha t}。\Big)$$

对到达际时间说来，此性质描述这样的普通场合，其中无论上一个到达者何时出现，完全不影响截至下一个到达者为止的时间。对服务时间说来，这性质较难解释。我们不应当期望它适用于服务者必须为各顾客进行同样固定的一系列操作的场合，因为在这种情况下一项历时长的服务应当意味着大概已无事可做。反之，在那种所需服务操作因顾客而异的场合，这性质的数学表述可能是十分现实的。就此而论，如果对某一顾客有大量的服务早已消逝，那么仅有的含义可能是此特定顾客需要比大多数顾客更延长的服务。

性质 3　若干独立指数随机变量的极小值具有指数分布。

为了数学地来叙述此性质，设 T_1，T_2，…，T_n 为分别有参数 α_1，α_2，…，α_n 的独立指数随机变量。又设U为这样的随机变量，其所取值等于 T_1，T_2，…，T_n 实际所取各值中的极小值，即

$$U = \min\{T_1, T_2, \cdots, T_n\}。$$

这样，如果 T_i 表示截至将有某种事变出现为止的时间，那么U表示截至将有n个不同事变的第一个出现为止的时间。现在注意对于任何 $t \geq 0$，

$$\begin{aligned} P\{U>t\} &= P\{T_1>t, T_2>t, \cdots, T_n>t\} \\ &= P\{T_1>t\}P\{T_2>t\}\cdots P\{T_n>t\} \\ &= e^{-\alpha_1 t} e^{-\alpha_2 t} \cdots e^{-\alpha_n t} = \exp\left\{-\sum_{i=1}^{n}\alpha_i t\right\}, \end{aligned}$$

因而U确实具有参数为

$$\alpha = \sum_{i=1}^{n}\alpha_i$$

的指数分布。

此性质在排队模型中对于到达际时间有某些含义。详细说来，假设有若干（n）不同类型的顾客，但每一类型（类型 i）的到达际时间具有参数为 α_i 的指数分布（$i = 1, 2, \cdots, n$）。根据性质 2，从任何规定时刻至下一个类型 i 顾客到达为止的剩余时间具有此相同的分布。所以设 T_i 为此剩余时间，从一个任何类型顾客到达的时刻算起。于是性质 3 告诉我们，整个排队系统的到达际时间U具有参数 α 由上列等式来定义的指数分布。结果是读者可决定不管顾客之间的区别而排队模型依旧有指数到达际时间。

然而，在不止一个服务者的排队模型中关于服务时间的含义甚至更为重要。例如，考虑所有服务者具有参数为 μ 的相同指数服务时间分布的场合。在此设 n 为当前正在提供服务的服务者的数目，并设 T_i 为服务者 i（$i = 1, 2, \cdots, n$）的剩余服务时间，也具有参数 $\alpha_i = \mu$ 的指数分布。于是可知，截至这些服务者的任何一个下一次服务完毕为止的时间U具有参数 $\alpha = n\mu$ 的指数分布。实际上，该排队系统当前正象一个单服务者系统那样在运行，其中服务时间具有参数为 $n\mu$ 的指数分布。在本章中，后面我们将常常使用这种含义来分析多服务者模型。

性质 4　跟 Poisson 分布的关系。

假设介于某种事变（例如，顾客的到达，或连续忙碌服务者的服务完毕）相继出现的时间具有参数为 α 的指数分布。于是性质 4 所涉及的是关于这种事变在某段时间内出现次数的概率分布所产生的含义。详细说来，设 $X(t)$ 为截至时间 t（$t>0$）的出现次数，而时间 0 是指计数开始的瞬间。此含义是

$$P\{X(t)=n\}=\frac{(\alpha t)^n e^{-\alpha t}}{n!},\quad n=0,\ 1,\ 2,\ \cdots;$$

即，$X(t)$ 具有参数为 αt 的 Poisson 分布（见 8.6 节）。例如，当 $n=0$ 时，

$$P\{X(t)=0\}=e^{-\alpha t},$$

这正是根据指数分布在时间 t 之后出现第一个事变的概率。此 Poisson 分布的均值是

$$E\{X(t)\}=\alpha t,$$

因而期望每单位时间事变数为 α。这样，我们就说 α 是事变出现的平均速率。当事变在连续的基础上来计数时，我们把计数过程 $\{X(t);\ t>0\}$ 说成是参数（平均速率）为 α 的 Poisson 过程。

当服务时间具有参数为 μ 的指数分布时，此性质提供了关于服务完毕的有用资料。为此，我们把 $X(t)$ 定义为一个连续忙碌服务者在消逝时间 t 内所达成的服务完毕数，其中 $\alpha=\mu$。对于多服务者排队模型，也可以把 $X(t)$ 定义为 n 个连续忙碌服务者在消逝时间 t 内所达成的服务完毕数，其中 $\alpha=n\mu$。

当到达际时间具有参数为 λ 的指数分布时，该性质对于描述到达者的概率性行为特别有用。此时 $X(t)$ 将是在消逝时间 t 内到达者的数目，其中 $\alpha=\lambda$ 是平均到达速率。所以到达者是按照 Poisson 输入过程出现的。这样的排队模型也被描述为呈现 Poisson 输入。

有时到达者被说成随机地出现，意思就是他们按照 Poisson 输入过程出现。此现象的一种直观解释是在长短固定的每一时期中有相同的机会看到一个到达者，不管前一个到达者何时出现，如下述性质所启示。

性质 5　对于 t 的所有正值，当 Δt 取小值时，$P\{T\leq t+\Delta t\mid T>t\}\approx\alpha\Delta t$。

当随机变量 T 具有参数为 α 的指数分布时，性质 2 意味着对于任何正量 t 与 Δt，

$$P\{T\leq t+\Delta t\mid T>t\}=P\{T\leq\Delta t\}=1-e^{-\alpha\Delta t}.$$

所以，由于 x 为任何指数时 e^x 有级数展开式：

$$e^x=1+x+\sum_{n=2}^{\infty}\frac{x^n}{n!},$$

可知当 Δt 取小值时，①

①更精确地，

$$\lim_{\Delta t\to 0}\frac{P\{T\leq t+\Delta t\mid T>t\}}{\Delta t}=\alpha.$$

$$P\{T\leq t+\Delta t \mid T>t\} = 1 - 1 + \alpha\Delta t - \sum_{n=2}^{\infty}\frac{(-\alpha\Delta t)^n}{n!} \approx \alpha\Delta t,$$

因为对于 $\alpha\Delta t$ 的充分小值，连加号后的各项可相对地忽略不计。注意 t 的值实际上丝毫不影响此概率。

前已指出，在排队模型中 T 可能表示到达际时间，也可能表示服务时间。所以此性质给所关注的事变（到达或服务完毕）在下一小段时间（Δt）内出现的概率提供了一个简便的近似值。（取 $\Delta t \to 0$ 时的适当极限也可使根据此近似值的分析变为精确。）该性质又表明对于小的 Δt 值此概率实质上跟 Δt 成比例。

9.5 增消过程

大多数初等排队模型假定排队系统的输入（到达的顾客）与输出（离去的顾客）按照增殖与消亡过程*出现。概率论中的这个重要过程在各种不同领域里都有应用。然而，在排队论的现实环境中，**增殖**一词是指排队系统里新顾客的到达，而**消亡**是指已得到服务顾客的离去。在时间 t（$t \geq 0$）系统的状态给定为 N(t)（如 9.2 节术语与记法中所定义）。这样，增消过程概率性地描述 N(t) 怎样随着 t 的增大而变化。概括说来，它表明个别的增殖与消亡随机地出现，其平均出现速率仅取决于系统的当前状态。更确切说来，增消过程的各项假定有如下述：

假定 1　给定 $N(t)=n$，截至下一个增殖（到达）为止的剩余时间的当前概率分布是参数为 λ_n（$n=0, 1, 2, \cdots$）的指数分布。

假定 2　给定 $N(t)=n$，截至下一个消亡（服务完毕）为止的剩余时间的当前概率分布是参数为 μ_n（$n=1, 2, \cdots$）的指数分布。

假定 3　一次只能出现一个增殖或消亡。

由于指数分布的性质 4（见 9.4 节）意味着 λ_n 与 μ_n 是平均速率，我们可用图 9.3 所示速率图把这些假定总括起来。此图中的各箭号显示系统状态的仅有可能转移（如假定 3 所指明），而每一箭号的登入项给出系统处于箭底的状态时该转移的平均速率（如假定 1 与 2 所指明）。

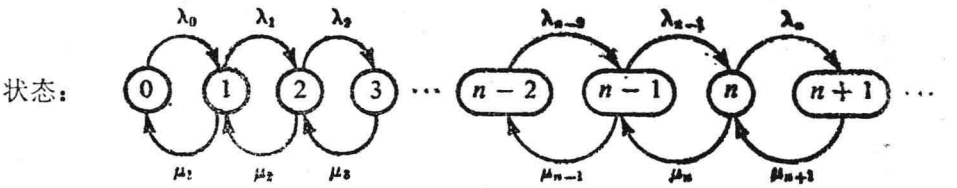
除少数特例之外，增消过程的分析在系统处于瞬变情况时是很困难的。关于 N(t) 的概率分布已获得某些结果，① 但它们因过于复杂而不大有实用。另一方面，在系统已到达稳态情况之后（假定此情况能到达），导出此分布是比较直捷了当的。这可直接从速率图做到，如下面所概述。

---

* 通常简称增消过程——译注。

① S. Karlin 与 J. McGregor, "Many Server Queueing Processes with Poisson Input and Exponential Service Times（具有 Poisson 输入与指数服务时间的多服务者排队过程）", 《Pacific Journal of Mathematics》 8: 87—118, 1958.

图 9.3 增消过程速率图



考虑系统的任何特定状态 n（n = 0，1，2，…）。假设我们要开始来计数增殖过程进入此状态的次数及离去此状态的次数。由于两种事变（进入与离去）必须交替出现，此两数必定要末相等，要末正好相差 1。此可能的差值 1 最终将使这两种事变出现的平均速率（每单位时间出现的总数）只有可忽略不计的差值（即当 $t \to \infty$ 时 $1/t \to 0$）。所以这两个速率从长远看来必定是相等的。这就产生下述关键的原则：

进速率 = 出速率原则　对于系统的任何状态 n（n = 0，1，2，…），进入事变出现的平均速率（期望每单位时间出现数）必定等于离去事变出现的平均速率。

表达此原则的方程叫做状态 n 的平衡方程。以未知的 P_n 概率建立所有状态的平衡方程后，便可解此方程组来求出这些概率。

为了具体说明平衡方程，考虑状态 0。增消过程仅从状态 1 进入此状态。这样，处于状态 1 的稳态概率（P_1）表示过程可能进入状态 0 的时间比例。给定过程处于状态 1，进入状态 0 的平均速率是 μ_1。（换句话说，对于过程消耗在状态 1 的每一累积时间单位，过程离开状态 1 而进入状态 0 的期望次数是 μ_1。）从任何其他状态，此平均速率为 0。所以过程离开其当前状态而进入状态 0 的总平均速率（进入事变的平均出现速率）是

$$\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1.$$

根据同样的推理，离去事变的平均出现速率必定是 $\lambda_0 P_0$，因此状态 0 的平衡方程是

$$\mu_1 P_1 = \lambda_0 P_0.$$

其他每一种状态有出入该状态的两个可能转移。所以这些状态的平衡方程的每一端表示两个有关转移的平均速率之和。否则，推理方法跟状态 0 所用的并无差别。这些平衡方程已在表 9.1 中总括起来。

表 9.1 增消过程平衡方程

状态	进速率 = 出速率
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$
⋮	⋮
n − 1	$\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$
⋮	⋮

注意第一个平衡方程含有两个待解的变量（P_0 与 P_1）；最初两个方程含有三个变量（P_0，P_1，及 P_2），等等，因而始终有一个“多余”的变量。所以求解这些方程的办法是使各变量都由其中的某一个表出，最方便的一个就是 P_0。这样，从第一个方程用 P_0 解出 P_1，再从此结果与第二个方程用 P_0 解出 P_2，等等。最后，可根据所有概率之和必须等于 1 这个条件来算出 P_0 的值。

应用此法，遂得下列结果：

状态

$$0 : P_1 = \frac{\lambda_0}{\mu_1} P_0$$

$$1 : P_2 = \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$$

$$2 : P_3 = \frac{\lambda_2}{\mu_3} P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) = \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$$

$$\vdots \qquad \vdots$$

$$n-1 : P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} + \frac{1}{\mu_n}(\mu_{n-1} P_{n-1} - \lambda_{n-2} P_{n-2}) = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} P_0$$

$$n : P_{n+1} = \frac{\lambda_n}{\mu_{n+1}} P_n + \frac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1} P_{n-1}) = \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} P_0$$

$$\vdots \qquad \vdots$$

为简化记法，设

$$C_n = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1}, \quad n = 1, 2, \cdots.$$

这样，稳态概率是

$$P_n = C_n P_0, \quad n = 1, 2, \cdots.$$

条件 $\sum_{n=0}^{\infty} P_n = 1$ 意味着

$$\left[1 + \sum_{n=1}^{\infty} C_n \right] P_0 = 1,$$

因而

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} C_n}.$$

给定此项资料，

$$L = \sum_{n=0}^{\infty} n P_n.$$

又由于服务者的数目 s 表示可同时得到服务（并因此脱离蛇队）的顾客数，

$$L_q = \sum_{n=s}^{\infty} (n - s) P_n.$$

而且，由 9.2 节中所示关系式还得出

$$W = \frac{L}{\bar{\lambda}}, \quad W_q = \frac{L_q}{\bar{\lambda}},$$

式中 $\bar{\lambda}$ 是长远平均到达速率。由于 λ_n 是系统处于状态 n（n = 0，1，2，…）时的平均到达速率，而 P_n 是系统处于此状态的时间比例，便有

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n。$$

上面给出的表达式有几个包含无限项的求和。幸而这些求和在若干有趣的特例中都有分析的解，① 如将在下一节中看到。否则，可在电子计算机上求有限项的和来得到近似解。

在导出这些稳态结果时曾假定各参数 λ_n 与 μ_n 有使过程能实际到达稳态情况的值。如果在 n 取某一值时有 $\lambda_n = 0$，此假定总是成立的，因而只可能有有限的几（即小于此 n）种状态。当 λ 与 μ 已有定义（见 9.2 节中术语与记法）且 $\rho = \lambda/s\mu < 1$ 时，也总是成立的。如果 $\sum_{n=1}^{\infty} C_n = \infty$，则不成立，

9.6 根据增消过程的排队模型

由于增消过程的平均速率 λ_0，λ_1，…及 μ_1，μ_2，…可各取任何非负值，这就在塑造排队系统的模型时提供巨大的灵活性。大概在排队论中使用最广泛的模型都是直接根据此过程的。由于假定 1 与 2（及指数分布的性质 4），我们说这些模型具有 **Poisson 输入与指数服务时间。** 这些模型仅在各 λ_n 与各 μ_n 怎样随 n 而变的假定上有差别。在这一节里我们就四种重要的排队系统提出这些模型中的四个。

基本模型（恒定到达速率及服务速率）

一种很常见的情况是排队系统的平均到达速率及每忙碌服务者平均服务速率，不管系统的状态怎样，实质上都恒定（分别为 λ 与 μ）。所以基本模型便作此假定。如果系统正好有一个单服务者（s = 1），这意味着增消过程的各参数是 $\lambda_n = \lambda$（n = 0，1，2，…）及 $\mu_n = \mu$（n = 1，2，…）。由此得出的速率图已在图 9.4 a 示出。

然而，如果系统有多服务者（s > 1），则不能如此简单地来表达 μ_n。要记得，当系统内目前有 n 个顾客时 μ_n 表示全排队系统的平均服务速率（即出现服务完毕因而顾客离开系统的平均速率）。在指数分布的性质 4 中已提到（见 9.4 节），如果每忙碌服务者平均服务速率为 μ，n 个忙碌服务者的总平均速率必定是 $n\mu$。所以当 $n \leq s$ 时 $\mu_n = n\mu$，而当 $n \geq s$ 时 $\mu_n = s\mu$，因而所有 s 个服务者都忙碌。此情况的速率图已在图 9.4 b 中示出。

①这些解根据下列众所周知的几何级数求和公式：

$$\sum_{n=0}^{N} x^n = \frac{1 - x^{N+1}}{1 - x}, \quad \text{任何 } x,$$

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1 - x}, \quad |x| < 1.$$