排队论在系统工程中的应用

（专题译丛）

第四机械工业部 一〇二八研究所

排队论在系统工程中的应用

（专题译丛）

1 9 8 0 年 1 2 月

四机部1028研究所情报室编

印　　刷　　河 南 新 乡

（南京1406信箱324分箱）

编 者 按

排队论这个题目，早期称之为公用事业的理论，它用概率论的方法，通过对各个随机服务对象的统计研究，找出反应这些随机过程的平均特性的规律。从而，在可以接受的性能/价格比条件下，改进系统工作能力，达到所谓高业务量的目的。它在各种各样的系统工程中有着广泛的应用。同样，在指挥控制系统中，为解决各类系统中断排队，系统响应时间，所设通道数目及缓冲器容量等是否满足实时控制要求等问题，都需要排队问题。

这里汇集的一组文章，大多是译文，是从一些比较有名的原版书（如Harry H.Goode和Robert E.Machol合著的“Systen Engineering”等）中选译的章节，有排队论的基本概念和方法，有应用的计算公式和图表，也有应用的实例。可供从事系统工程设计的工作者参考。

这组文章的错漏不妥之处，请同志们在阅读中指正。

四机部一〇二八研究所情报研究室

一九八〇年十月

目　录

排队论

这一章研究的题目有各种各样的名称：等待线理论、排队理论、拥挤问题中继理论。在第21章我们讲过系统内设计的第二阶段是研究和说明系统的"高业务量"方面。在任何"高业务量"系统里，如果作某种服务的通道数目太少，那么常常会形成特别长的等待线。因此，确定各处通道的个数——即"高业务量"设计的第一阶段，就要求解有关的排队问题。

在第21章还指出，不必详细说明单线设计就能研究系统"高业务量"的要求。应用排队论必须知道的事情有三件：

系统中每一"大框"的输入的频率（或时间分布）；

"大框"进行该操作所需的时间（或时间分布）；

当某一输入到达时，"大框"正服务于前一个输入，那么系统将按照什么样的规则去处理（例如：形成等待线或输入丢失）。

第一件事由系统外设计（Exterior system design）确定，后两件事在系统内设计（Interior system design）中确定（特别是在单线设计中确定）。其中规则由系统逻辑得来，处理时间的分布来自予先的设备（即组件）研究。

23—1输入源、排队、通道

我们讨论的基本系统如图23—1所示，可以有很多源供给一个排队，一个源供给一个排队，或一个源供给很多排队，也可以有很多排队供给一个通道，一个排队供给一个通道，或者一个排队供给很多通道。这似乎引出九种不同的概型。实际上它可以简化成两三种概型，如果几个源供给一个通道，一般地可以和单通道的情况一样考虑。当然从组合源而来的输入时间分布，一般地，不同于一个源的分布。当一个源供给几个排队时，必须确定一个特殊的输入进入那一个排队，一旦确定了这规则，每一个排队就可以认为有一个单一的源。类似地，如果几个排队供给一个通道，那么一般地也可以把它们看成单一的排队。因此，我们假设一个单一的源供给单一的排队，一个排队再供给一个或几个通道。如果有几个排队供给几个通道，那么和以前一样，必须存在一个规则。如果一个排队中的顾客，只能进入特殊的通道，那么事实上那种排队供给一个单一通道；如果一个排队中的顾客可以进入任一空着的通道，那么事实上供给多通道的所有的排队是一个单一排队。我们把要详细讨论的这两种情况分别叫做单通道和多通道问题。其它作者已经分别给它们起了不同的名字，例如分别叫做单干通道，合作通道。

本文译自"System Engineering" Harry H.Goode，Roberf E.Machol 1957.第五部分，第23章。 吴健译 孙光第校

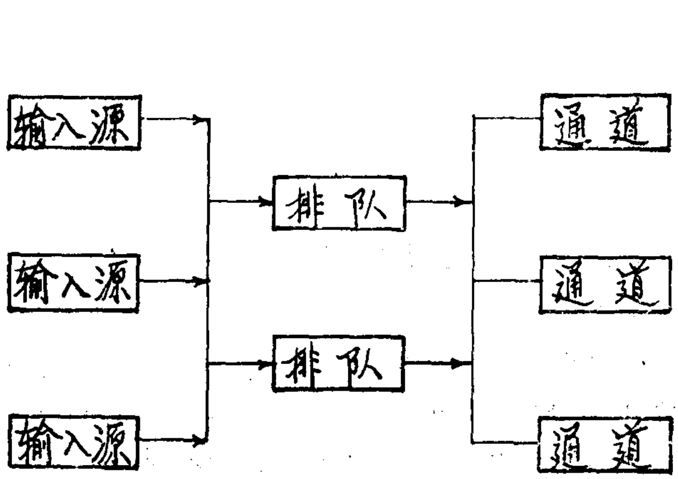

图23—1 输入源、排队、通道的基本系统

输入源 我们要分清输入源、输出、顾客。输入源给出输入，当输入去排队时，称它为顾客。给出输入的时间分布是输入源的特征。偶而顾客之间也要区分，这在“排队”中讨论。

讨论的最简单源是输入在特定的相等的时间间隔发出。另外，如果通道处理工作的时间像自动生产线一样为一个常数，那么容易保证没有等待线形成，不存在什么问题。但是如果通道处理时间是散布着的，那么等待线可能形成。这情况出现在医生办公室里，在那里病历诊断表在前面排列着，但看病时间都是散布着的。反之，散布着的输入的时间间隔，常数处理时间，也出现在地下铁道的旋转栅门的情况下，也可能形成等待线。最容易处理的分布是普阿松分布，它在某种意义上是最随机的[1]。这是这一章我们将处理的唯一的分布。

输入源的一个重要特性是它为有限的还是无限的。例如在电话系统里，一个中继组为有限个用户服务。虽然我们可以安全地假定每一个用户以等概率进行一次呼唤（所以输入是普阿松分布的），但占线的用户不能产生呼唤。所以在占线期间可能产生的呼唤数目要比假如已占线的用户产生呼唤的数目小得可观。前面我们讲了多重源可以看做是单一源。但是，当源的数目不比通道数目大很多时，当前一个输入未从通道中清除之前，如果源不能产生新的输入，那么表示一组多重源的假想单一源产生的输入不是等概率的，这后者的概率正比于有用源的数目

（23—5）节里将研究一个多重源和多重通道的例子。那里输入是机器的故障。但是在很多的分析里我们将假设输入率不依赖排队的长短。特别是为了在电路上应用，已经研究发展了很多排队理论，在有限输入源的排队问题文献中有不少材料。

排队 虽然一“排队”和等待线一般用做同一个意义，但是我们还是有必要区分一下。为了方便，应用一个表示队长的符号 n，并且包括正被服务的顾客（或者，在多重通道的情况下，正被服务的顾客们），这样定义的排队，包括在通道中顾客和顾客们。我们把等待线定义为包括在排队中而不在通道里的顾客，所以队长比等待线长一个通道（或n个通道），除非有一个或多个通道是空闲的。

当没有等待线形成时，至少有两种不同的情况是有趣的，而且用排队理论研究了。第一种情况是，顾客未立即被服务就离开了。例如：在电话总机值班室中，所有来话中继线和来

话记录器都占线，那么在电话系统中的一个局间呼唤，就是这种情况。此时，用户听一个“忙音”，他除了挂上电话以外，没有别的选择，他可以重新挂电话，但这一特定的呼唤“丢失”了，从电话系统中被清除掉了。在这种情况下。要求把丢失顾客数保持在最小，排队理论的中心问题就是要研究有多少数量的顾客被丢掉。

另一个极端情况是，如果一架飞机到达一个机场，发现所有的跑道是“占线”的，它除了排队之外，没有别的选择。用户打自动电话是一种中间情形，发现所有的设备都忙着，他没有听到拨号音，他可以等，或者不等，随他便。未被立即服务的顾客被丢掉的情况在数学上要比它排队简单。另一个数学上简单的情况是未被服务的顾客等待（在等待线上，或在通道上，假如它总是到达通道的）一个等于它立即被服务的服务时间。这两个情况在[33]里被详细讨论了，它们是这一章技巧的简单的推广，下面我们不再研究它们了。

不形成等待线的另一个简单情形是无穷多通道的情形。当然这种假设是不现实的。虽然如此，它也不仅仅是数学上简单而且是相当有兴趣的。因为我们可以确定无穷多通道中任何数目通道是忙的时间比率（或概率），如果多于10个（假设）通道是忙的概率仅仅是0.001，那么我们可以只构造10个通道，且知道非零长的等待线仅仅在千分之一的时间内形成。这在数学上不是严格精确的，但是如果概率迅速减少的话（在实际上一般是这样），它是很好的近似。无穷多通道是下面我们所要研究的多通道的第一个情形。

另外有一类型排队不满足图23—1所示的框图，那就是“双边排队”，这种情况下排队可以认为是通道，而通道也可以认做为排队。例如：有一个出租汽车站，那里如车比人多则车排队，人比车多则人排队。在一定意义上，所有的排队都能看作双边排队。我们把闲着的通道看成是在排队。例如，考虑上面提到的机器及机械师，当有很多机器发生故障。其中有的机器闲着等着修理（排队），而修理工人修理其它的。另一方面可能有时某一些机械师甚至全体都闲起来，而所有的机器都工作着，这种情况下机械师在排队。但这个排队没有输入源，除非机械师们放弃他们的职责，它们的长度是严格有限的。

所谓“排队规矩”。通常我们假设排队规矩是每一个顾客在线上保持它固有的位置。例如当等待旅客时的出租汽车就是这种情况，另外一种极端的情况是忙着的电话交换机的情况，当用户打电话时，没有听到拨号音，但在线上没有位置，当通道有空时，随机地从等待线上选取顾客。最后一个到达和最先一个到达是完全一样的。当具有如下性质时出现一种中间的情况，即每一个顾客在线上有一个位置，而某些顾客如有某些可以优先于其它顾客的高一级性质。

这种类型的排队影响我们希望估计的某些因素，不影响其它一些因素。例如：它不影响排队长度的分布，但它影响等待时间的分布，当排队规矩是重要的时候，我们在这一章里永远假设位置是固定的。但是在很多实际系统里存在有优先的性质。例如，在机场上的排队，燃料快用尽的飞机显然优先于带有充足燃料的飞机。在很多情况下，系统设计者希望给出系统优先性。有一个优先性准则，它似乎显然，不值得一提，但常常被忘记：当一个顾客先到线上则所有其他顾客实际上往后排。如果大部分顾客给了优先性，其余的必须等待一段时间。

通道 在很多情况下，当顾客排完了队之后，最终得到某一有限时间的服务，提供这种服务的人或设备被叫做通道或服务员。在本书仅用第一个名称。顾客在通道里所花的时间称

为处理时间或服务时间，这里也仅用第一个名称，处理时间可以是常量或是散布着的。

我们研究前叙的出租汽车和乘客的情况，如果出租汽车是通道，那么出租汽车到达时间间隔是处理时间。线上最前面的人必须被假设是在通道中。当出租汽车是顾客，那么出租汽车的到达时间间隔确定输入分布。这样人们对输入分布和处理时间分布的紧密关系不感到惊奇。如果一个“繁忙”通道的处理时间端点是普阿松分布，则处理时间本身是一指数分布（反之亦然）。处理时间的指数分布与输入的普阿松分布类似也是随机的（见后面注θ）。而且它最易进行数字处理，在本章我们只研究这种处理时间分布。

存储器及缓冲存储器 缓冲存储器，顾名思意是为保存排队而提供的存储器，例如建筑一个剧场时，为了在恶劣的天气下，观众在可能的排队时不受损失，必须提供一个象前庭式的广阔的存储空间，这个昂贵的存储空间，可以看做是缓冲存储器，一个更典型的系统设计例子是通道的编码系统，就象在28.1节所讨论的，如某些符号（如英语中的字母q.z)比某些符号（如字母e.t）更常用，我们可对常用字母用短编码，对不常用的长编码。但是如果译码器的输入和输出均为常数率，则可能在短期内出现很多不常用字母，这时通道（译码器）将不能处理它们，于是形成等待线，为保存这些“顾客”必须提供缓冲器，排队论告诉我们缓冲存储器将必须是多大（即排队可以有多长），如果我们要求一星期内不从缓冲存储器溢出，系统逻辑上怎么办呢？或者亻止输入源，或者允许信号丢失，或者允许人工干予。

在原来的存储器和缓冲存储器之间划一条并不是容易的。我们研究一个仓库，它是一种排队，它的输入是工厂的产品，它的通道是出卖，处理时间（两次销售之间）或多或少随机分布着。一个更重要的例子（这里再重复强调一下，我们断言所有的关系都是信息系统）是空军地面系统的数据存储设备。有一百万个信息，55种天然分类。问题是55个存储器，每一个该作多大。这里要求高速存储，所以加一点容量也是昂贵的。所以处理时间的时间长度是当一片数据在存储中写入直到擦掉。在这种情况下，可以更方便地认为存储器由多重通道组成（在存储器中，每位是一个单一通道）。问题是为了把在某一时间，如一年内任何时间等待线形成的概率减少到某个指定值以下所必须提供的通道数目。利用排队论公式估值，应用系统设计的基本原则（21—6节）可以确定通道个数，缓冲区大小等问题。这就是说，第Ⅰ类错误和第Ⅱ类错误的概率可以从排队论计算，确定的第Ⅰ类和第Ⅱ类错误的花费可以从其它一些考虑来计算，对不同选择花费期望值可以计算出来并且极小化。

未知量 从上面的讨论已经很清楚，我们希望在排队论中计算的主要量是排队长度和等待时间。应该注意，这里一个是离散的，另一个是连续的，所以在数学上是不同的，一个是差分方程，另一个是微分方程。

有时只要求期望（即平均）排队长度和（或）等待时间，这常常是很容易计算的，特别是当允许进行某种程度近似时，在另外一些场合，我们要求知道排队（等待时间）将超过某长度的概率。为求这些量常常必须计算分布率数。而且要求出有兴趣的参数，常常需要把上述的某些量或全部看做通道的函数来计算。下面将给出说明这些的例子。

在特殊情况下，可能要求计算分布的某些特殊性质。譬如：如果一个通道是（或应用）一个电动机，可能不仅要求知道应用它的时间比率，而且也要知道在应用时每两次闲置时间之间长度有多长，就是应用电动机的时间分布。这等价于排队是非零长的时间分布，为了论述“忙时”分布要用贝塞尔函数，我们省略了。

当通道昂贵时，人们常常计算它的"繁忙率"，它是"忙时"的平均比率，但是读者不要误解为一定要保持一个低的"繁忙率"。象在本章未尾数值例子中所看到的一样，为减少50%或更少一些繁忙率，常常必须提供充分多的通道，但是等待线仍以可观的频率形成。如果一个昂贵的通道仅仅有一个低的"繁忙率"，则可以看一看是否可以改善，但不必要注意它的无效性。在同一机场安置两个全套的 GCA 设备，即使它们两个有99%的时间是闲置的。

23—2 单通道

如上所述，当输入具有常数的时间间隔，且处理时间也是常量，则不存在排队论的问题。当一个为常量另一个随机地遵从某一分布，或两者都随机地遵从某一分布时，则至少在某些时间内存在等待线问题。首先我们感兴趣的是对于当这分布为任意时的排队，能确定些什么。

在全章中我们将应用如下符号：

t = 相继输入产生的时间间隔（当上下文不引起混淆时t也用来表示产生的时间变量）。

γ = 在某一有兴趣的区间内所产生的输入的数目。

m = 每单位时间的平均输入数。

T = 处理时间。

R = 在某个有兴趣的时间内从一个通道输出的数目。

M = 每单位时间从一个"繁忙"通道的平均输出数，

$\frac{1}{M}$是T的期望值。

$$\rho=\frac{m}{M}\cdot$$

ω = 等待时间 = 输入产生到它进入通道之间的时间。

n = 排队长度（包括在一个或几个通道中的"顾客"或"顾客们"）。

ν = 通道数目。

j = 等待线长度（$j=n-\nu$，当$n\geqslant\nu$；$j=0$，当$n\leqslant\nu$）。

β = 源的数目（除23—5节，永远无限）。

比值ρ在排队论里是相当重要的②。容易看到，对单一通道，如$\rho>1$，则这排队必然越来越长没有限制。任何特定的有限的排队长度的概率随时间的增加而趋于0；任何特定的有限的等待时间的概率也同样。很清楚，这时我们求不出有兴趣的概率的表示。设计者的第一个要求就是探知"线"是稳定的条件。过程是"平稳"时，这条件将出现。对于我们的目的，用如下条件确定平稳是足够的。即

当
$$\frac{dP(n)}{dt}=0 \tag{23—1}$$

对所有的n成立（这里$P(n)$为队长为n的概率——译者）。在单一通道里，当$n\rho>1$，这条件绝不会出现，同样，当$\rho=1$也不会出现，这下面会指明。在所有有兴趣的实际情况，如果m和M保持常量且当$\rho<1$（或，一般地，如$\rho<\nu$）在充分长的时间后将会出现这个条件。除非相反的声明，在下面所有讨论中，我们永远假设$\rho<\nu$且过程是平稳的。

任意输入和处理时间分布 我们首先指出在单一通道的情况下，通道在“忙”的概率等于ρ。在一定意义上讲，这是直观的。假设，每两分钟平均产生一个输入，平均处理时间是1/4分钟。那么，$M=4$，$m=1/2$，且$\rho=1/8$。在10个小时的期间之后产生约300个输入。该通道将忙$1/4\times300=75$分钟或1.25小时，即时间的八分之一。下面的方法是Kendall[80]发明的。

研究当某一“顾客”C_0刚刚离开通道的情况。我们假定在它离开后排队长度是n。（可以是0），当$n_0\neq0$在“线”上的输入是C_1，C_2，……。C_1刚刚开始它的处理时间T_1。如当$n_0=0$，下一个到达的输入C_1将马上被服务。在T_1每一种情况下，在T_1期间内，当C_1正被服务时，将有很多输入到达（其数目可为0，是C_1的处理时间)。我们将称这个数目为r_1。在第一种情况下，即$n_0\neq0$，当C_1离开时，新的排队长度将是$n_0+\gamma_1-1$，在第二种情况下是γ_1（因为我们考虑离开的那一无穷小瞬间，另一个输入产生的概率几乎为0）。这两种情况可以表示成如下的一个方程：

$$n_1=n_0+\gamma_1-1+\delta_0 \tag{23—2}$$

其中

$$\delta_0=\begin{cases}1 & 当n_0=0,\\ 0 & 当n_0\neq0.\end{cases}$$

量δ_0是仅取0或1的一个数，它的期望值在这两者之间。注意$n_0\delta_0=0$且$\delta_0{}^2=\delta_0$。

现在我们对（23—2）两端，首先对γ_1，然后对T_1取期望值。从我们的平稳性假设$E(n_0)=E(n_1)$。所以对于γ_1有

$$E(\delta_0)=1-E(r_1)=1-mT_1.$$

因为当C_1在通道里时的期望输入数目是每单位时间到达的平均值乘处理时间。对T_1取期望，给出δ的期望：

$$E(\delta)=1-mE(T_1)=1-\frac{m}{M}=1-\rho.$$

这是因为$E(T_1)=1/M$已被定义。但$E(\delta)$是$n=0$的概率，即为通道在空闲的概率。所以：

$$P(n=0)=p(0)=1-\rho \tag{23—3a}$$

$$P(n>0)=p(>0)=\rho \tag{23—3b}$$

当不知道输入时间分布和处理时间分布的任何特性时，我们还能进展一点。把（23—2两端平方，我们得到

$$n_1{}^2=n_0{}^2+(\gamma_1-1)^2+\delta_0{}^2+2n_0(\gamma_1-1)+2\delta_0(\gamma_1-1)+2n_0\delta_0.$$

右边最后一项等于0。代替$\delta_0{}^2=\delta_0$且取期望值，得到：

$$E(n^2)=E(n_0{}^2)+E((\gamma_1-1)^2)+E(\delta_0)+2E(n_0(\gamma_1-1))+2E(\delta_0(\gamma_1-1)).$$

同前，由于平稳性，右边的第一项等于左边等一项。更进一步因为到达数目γ_1被假定不依赖于在“线”上的数n_0，也一定不依赖于仅依赖于n_0的δ_0，利用多元分布的有关公式，可以通过右端两个乘积项计算期望：

$$0=E[(\gamma_1-1)^2]+E(\delta_0)+2E(n_0)E(\gamma_1-1)+2E(\delta_0)E(\gamma_1-1).$$

对γ_1和T_1取期望值。替代入上边求出来的值，就是

$$E(\delta)=1-\rho 和 E(\gamma)=\rho,$$

$$O = E(\gamma^2) - 2\rho + 1 + (1-\rho) + 2E(n)(\rho - 1) + 2(1-\rho)(\rho-1)$$

$$2E(n)(1-\rho) = E(\gamma^2) - 3\rho + 2 - 2 + 4\rho - 2\rho^2 = E(\gamma^2) - 2\rho^2 + \rho$$

$$E(n) = \frac{E(\gamma^2) - 2\rho^2 + \rho}{2(1-\rho)} = \rho + \frac{E(\gamma^2) - \rho}{2(1-\rho)} \quad \rho \neq 1 \tag{23—4}$$

公式（23—4）对任何输入和处理时间分布都成立，只要这些分布不依赖 n，不随时间而改变，且 $\rho<1$。注意，虽然 $E(\gamma)=\rho$，$E(\gamma^2)$ 一般不等于 ρ，所以式（23—4）表明当 $\rho\to1$ 时，$E(n)\to\infty$。没有到达时间分布的某些知识，我们就不能求出 $E(\gamma^2)$ 的值。

任意处理时间分布，普阿松输入分布

本章后面，除特别声明以外，我们永设输入是普阿松分布的。由于普阿松分布在下面几页应用很多，我们抄本书第5章两个表达式于下：

$$p(k) = \frac{e^{-\mu}\mu^k}{k!} \tag{5.20}$$

$$p(k) = \frac{e^{-\mu}(\mu t)^k}{k!} \tag{5.27}$$

在（5.20）中，$p(k)$ 是单位时间恰出现 k 次的概率；在（5.27）中，$p(k)$ 是一个时间区间 t 恰出现 k 次的概率。在每种情况下，每单位时间出现的期望数是 μ。而（5.20）和（5.27）的均值、方差分别是 μ 和 μt。

在求出（23—4）的值时，应用第二个形式

$$P(\gamma_1) = \frac{e^{-mT_1}(mT_1)^{\gamma_1}}{\gamma_1!}$$

因为 γ_1 是处理时间 T_2 期间内输入的数目，它的均值 mT_1 是一个常量，它的方差也是 mT_1。但由（5.10）有

$$\sigma_{\gamma_1}{}^2 = E(\gamma_1{}^2) - (E(\gamma_1))^2$$

所以 $$E(\gamma_1{}^2) = \sigma_{\gamma_1}{}^2 + (E(\gamma_1))^2 = mT_1 + (mT_1)^2.$$

现在对所有的求 T_1 的期望值，得到：

$$E(\gamma^2) = mE(T) + m^2E(T^2).$$

由（5.10）有 $E(T^2) = \sigma_T{}^2 + [E(T)]^2 = \sigma_T{}^2 + 1/M^2$

所以 $$E(\gamma^2) = \frac{m}{M} + m^2\left(\sigma_T{}^2 + \frac{1}{M^2}\right) = \rho + m^2\sigma_T{}^2 + \rho^2$$

最后代入（23—4）

$$E(n) = \rho + \frac{\rho^2 + m^2\sigma_T{}^2}{2(1-\rho)} \tag{23—5}$$

从（23—5）看到，由于这个分数的分子永远是正的，不管处理时间是什么分布，当 $\rho\to1$ 时，n 的期望趋于无穷。这公式对任何处理时间分布都成立。对任一特定的分布可以把 σ_T 代入而得到 $E(n)$。在指数分布的情况，$\sigma_T{}^2 = 1/M^2$（6-36），这一表达式简化成

$$E(n) = \frac{\rho}{1-\rho}. \tag{23—6}$$

为求期望等待时间，我们研究输入C_1产生到它离开通道的区间段，这整个排队时间等于等待时间ω_1加上处理时间T_1。当C_1离开通道时，这排队长度为n_1（它可为0），在区间ω_1+T_1的期间内输入平均数是$m(\omega_1+T_1)$，所以期望排队长度是

$$E(n_1)=m(E(\omega_1)+E(T_1))$$

$$E(n)=mE(\omega)+\rho$$

代入（23—5）中E(n)的值，得

$$\rho+\frac{\rho^2+m^2\sigma_T^2}{2(1-\rho)}=mE(\omega)+\rho$$

$$E(\omega)=\frac{\rho^2+m^2\sigma_T^2}{2m(1-\rho)} \qquad (23—7)$$

在（23—7）两端除以$E(T)=1/M$，得到一个很有用的形式

$$\frac{E(\omega)}{E(T)}=\frac{\rho^2+m^2\sigma_T^2}{2\rho(1-\rho)}=\frac{\rho}{2(1-\rho)}(1+M^2\sigma_T^2) \qquad (23—8)$$

（23—8）的比值对于任何排队系统都是一个很有价值而没有量纲的数，它给出依赖处理时间的期望等待时间，例如图23—2所画的就是这么个数。从（23—8）的一般形式，我们可以导出一些定性的推论，因为括号内的量永远为正。当ρ趋于1时，期望等待时间趋于无穷大。更进一步，当$\sigma_T=0$时，期望等待时间最小，这种情况仅当处理时间是一个常数时出现（即不是分散分布的）。注意，可用这种方法获得一定程度节约。假定处理时间是指数分布的，那么$\sigma_T^2=1/M^2$，期望等待时间是

$$\frac{E(\omega)}{E(T)}=\frac{\rho}{1-\rho} \qquad (23—9)$$

它是同一均值的常数分布的两倍，可以指出[81]用控制输入到达时间（而不是处理时间）为不同常数的办法，可以把期望等待时间减少到相当的程度。如果不用这些办法，那么，一般说来，减少等待时间，只有两个办法：减少ρ(减少m或增加M或者两者都进行）或外加通道。

指数处理时间分布，任意输入分布

对任意输入分布和指数处理时间分布有许多公式被研究出来了。仅仅在上面导出期望值的许多情况，就有人[130]已经导出了它们的分布公式。这些公式一般地说类似于我们下面对普阿松输入分布所导出的那些公式。例如（23—17），只要用一个复杂的函数代替参数ρ。所有上面的结论（下面还有很多）基于必须在实践中验证的一些假设，特别，如m，M和n都是独立的。这些假设将在23—6节中讨论。

指数处理时间分布，普阿松输入分布

本章后面部分，除非另外声明，我们将假定输入分布如（5.20）、（5.27）所描述。而处理时间如第六章所讨论的，由

$$P(T)dT=Me^{-MT}dT \qquad (6.34)$$

所描述。

排队长度 平均排队长度是由（23—6）给出的。而我们还希望知道队长的分布，为此将利用时间t、排队是某一长度的概率之间的差分方程。特别地，如果在时间t+dt队长是

n $n \neq 0$，那么下三个事件必须发生其中之一。

1． 在时间 t，队长是 $n-1$，在时间区间dt内，有一个输入产生，且设有一个通道服务完毕。

2． 在时间t，队长是n；在时间区间dt内，设有输入产生，且没有一个通道服务完毕。

3． 在时间 t，队长是 $n+1$。在时间区间 dt 内，没有输入产生。且有一个通道服务完毕。

在任何使得 $t+dt$ 队长为 n 的可以想得到的其它事件序列中，在一个无穷小的区间内，至少包括两个事件（即两个输入，两个完毕，或一个输入，一个完毕）。由于我们有普阿松分布的假定，这样组合出现的概率是0。在dt内恰有一个输入的概率是mdt，在dt内0个输入的概率是$1-mdt$；在dt内恰有一个服务完毕的概率是Mdt，在 dt 内没有服务完毕的概率是 $1-Mdt$，所以，

$$P(n;t+dt)=p(n-1;t)mdt(1-Mdt)+p(n;t)(1-mdt)(1-Mdt)+p(n+1;t)(1-mdt)Mdt.\ n\neq 0 \qquad (23—10)$$

其中，符号$P(n;t)$表示在时间t队长为n的概率。展开，消去所有的$(dt)^2$项，得到

$$P(n;t+dt)p(n-1;t)mdt+p(n;t)-p(n;t)mdt-p(n;t)Mdt+p(n+1;t)Mdt.\ n\neq 0 \qquad (23—10')$$

减去$P(n;t)$除以dt得到

$$\frac{p(n;t+dt)-p(n;t)}{dt}=mp(n-1;t)-(m+M)p(n;t)+Mp(n+1;t).\ n\neq 0$$

令dt趋于0，则由微分定义，方程的左边成为$dp(n)/dt$，但由（23—1）这个导数等于0。利用$p(n+1,t)=p(n+1)$我们得到

$$p(n)=\frac{m+M}{M}p(n)-\frac{m}{M}p(n-1).\ n\neq 0 \qquad (23—11a)$$

对$n=0$，原来的差分方程改为：

$$P(0;t+dt)=p(0;t)(1-mdt)+p(1;t)(1-mdt)Mdt,$$

得到

$$\frac{P(0;t+dt)-p(0;t)}{dt}=-mp(0;t)+Mp(1;t)=0.$$

和
$$P(1)=\frac{m}{M}p(0). \qquad (23—11b)$$

方程（23—11）也可写作

$$P(n+1)=(\rho+1)p(n)-\rho p(n-1).\ n\neq 0$$
$$P(1)=\rho p(0).$$

用考察头几项的办法解这差分方程：

$$P(2)=(\rho+1)p(1)-\rho p(0)=(\rho^2+\rho)p(0)-\rho p(o)=\rho^2 p(0).$$
$$P(3)=(\rho+1)p(2)-\rho p(1)=(\rho^3+\rho^2)p(0)-\rho^2 p(0)=\rho^3 p(0).$$
$$P(n)=\rho^n p(0).$$

由（23—3）给出p(0)的值是$1-\rho$，所以

$$\frac{\rho}{1-\rho}=\frac{m}{M-m}\text{.}$$

方差是

$$\frac{\rho}{(1-\rho)^2}\text{.}$$

等待时间 正如早些时候指出的那样，如果处理时间是指数分布的，那么一个“繁忙”通道的处理时间的尾端是普阿松分布。这就是说，如果通道是“繁忙”的，那么服务将在d^t区间内完毕的概率是Mdt，所以从连续“繁忙”通道中输出数目的分布是由（5.27）给出：

$$P(R)=\frac{e^{-Mt}(Mt)^R}{R!}\text{.}$$

当然空闲通道的输出由

$$P(R)=0$$

给出。

现在我们希望求出概率密度 $p_n(\omega)d\omega$，这是下面事件的条件概率：一个输入产生时队长为n，而这输入等待时间大于ω小于$\omega+d\omega$。这是下两个事件的乘积：一个是在时间区间ω内通道恰产生了n-1个输出（这个可以由上述的P(R)当$R=n-1$的表示式给出）。另一个是在dω时间内通道恰恰有一个输出。所以

$$P_n(\omega)d\omega=\frac{e^{-M\omega}(M\omega)^{n-1}}{(n-1)!}Md\omega\text{.}$$

当输入产生时，队长为n的概率由（23—12）给出，这两个事件联合出现的概率——即由（4.11）表示队长为n且输入将等待大于ω而小于$\omega+d\omega$时间的概率，是这两者之积。

$$P(n,\omega)d\omega=(1-\rho)\rho^n\frac{e^{-M\omega}(M\omega)^{n-1}}{(n-1)!}Md\omega\text{.}\quad n\neq 0$$

为了从这个表达式中消去n得到ω的概率密度函数，我们从0到无穷对所有n的值求和。当然，对$n=0$等待时间恒等于0。

$$P(\omega)d\omega=\sum_{n=1}^{\infty}(1-\rho)\rho^n\frac{e^{-M\omega}(M\omega)^{n-1}}{(n-1)!}Md\omega=(1-\rho)\rho e^{-M\omega}Md\omega\sum_{n=1}^{\infty}\frac{\rho^{n-1}(M\omega)^{n-1}}{(n-1)!}$$

$$=(1-\rho)\rho Me^{-M\omega}d\omega\sum_{h=0}^{\infty}\frac{(\rho M\omega^h)}{h!},$$

其中$h=n-1$，求和结果是指数$e^{\rho M\omega}$，于是

$$P(\omega)d\omega=(1-\rho)\rho Me^{-M\omega(1-\rho)}d\omega \tag{23-13}$$

这个表达式的均值已经由（23—9）给出是$\rho/1-\rho$，平均处理时间（每个等于1/M），即

$$\frac{\rho}{(1-\rho)M}=\frac{m}{M(M-m)}\text{.}$$

在进入通道之前至少等待W时间的概率由W到∞积分（23—13）式而得

$$\int_W^{\infty}p(\omega)d\omega=\rho\int_{MW(1-\rho)}^{\infty}e^{-M\omega(1-\rho)}d[M\omega(1-\rho)]=\rho e^{-MW(1-\rho)}$$

因为（23—13）定义一个概率密度函数，因为ω不能取负值，它从0到无穷积分为1。但

是如果我们代替 $W=0$，在最后一个表达式里，我们得到ρ，显然，（23—3）式给出等待时间等于0的概率是 $1-\rho$，所有等待时间在0与∞之间的概率一定是ρ。等待时间概率密度函数在 $W=0$ 有一个大小为 $1-\rho$ 的所谓δ函数值，如图（23—2）所示，在这一点上密度函数本身是无限的，在这0宽度、无限高的狭条上面积是 $1-\rho$。

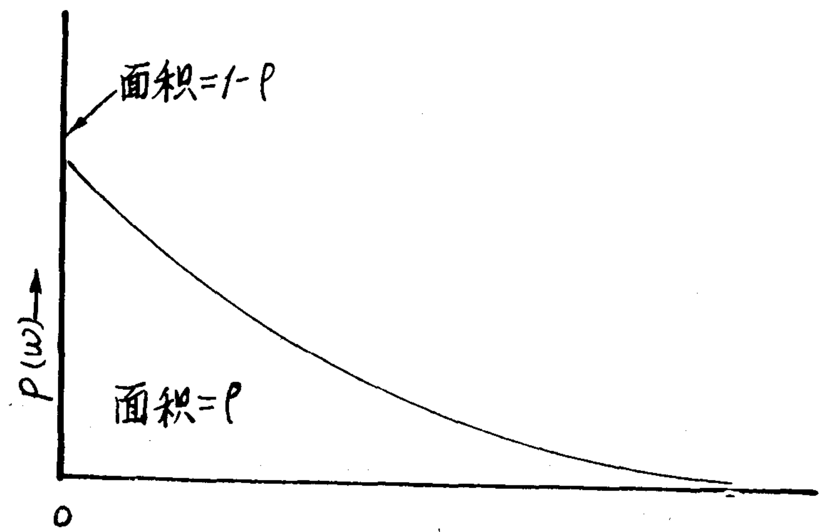

图23—2　等待时间分布

23—3　多（平行）通道

很多系统设计兴趣在于提供通道的数目。我们设想一个排队供给一组通道，在等待线前面的顾客进到第一个通道里，这样它就被利用了。象以前一样，考虑去排队的输入是单位时间平均m个输入的普阿松分布，每个通道有彼此独立（并独立于队长）的处理时间，分布带有均值1/M的指数分布。队长n包括通道中的和等待线中的所有顾客，当等待线中的长度是非负时，我们以 $j=n-\nu$ 表示，其中ν是通道数。

我们将发现只要做一些适当的变更，在（23—2)中发展起来的很多方程可以直接应用。这是因为在区间dt内有一个输出的概率独立于排队长度。为指出这一事实，用 M_n（或 M_{n-1}、或 M_{n+1} 随方便）替代（23—10）中的M。后面将处理如下情况：在区间dt中输入的数目也独立于排队长度。在（23—10）中将用 m_n（或 m_{n-1}、或 m_{n+1}）替代m，象以前一样，在解这方程时，求出导数

$\frac{dP(n)}{dt}$，设它等于0，代入（23—11）中得到 $P(n+1)=\frac{m_n+M_n}{M_{n+1}}P(n)-\frac{m_{n-1}}{M_{n+1}}P(n-1)$

$$P(n+1)=\frac{m_n+M_n}{M_{n+1}}P(n)-\frac{m_{n-1}}{M_{n+1}}P(n-1) \qquad (23\text{—}14a)$$

$$P(1)=\frac{m_0}{M_1}P(0) \qquad (23\text{—}14b)$$

这更一般的形式包括（23—11）的特殊情况，即当 $m_n=m$ 和 $M_n=M$ 时。

无穷多通道　我们首先假设 $\nu=\infty$，当然，这不对应于任何实际情况，永远不形成等待线。但这个问题容易解，而且结果有用处。例如，假设在无穷多通道的特殊情况下，求得应用多于6个通道的概率仅为0.01。如果构造一个6个通道的系统，我们知道：（1）如没有被

服务的顾客马上丢失的话，将损失大约1/100的顾客。（2）如果没有被服务的顾客立即排队的话，则仅有百分之一的时间形成等待线。

在这种情况下，在dt区间内有一个输出的概率是

$$M_n dt = nMdt,$$

这是因为在区间dt内两个或多于两个通道为空闲的概率包含高阶无穷小，代入 $m_n = m_{n-1} = m$，$M_n = nM$和$M_{n+1} = (n+1)M$，（23-24）成为

$$P(n+1) = \frac{m+nM}{(n+1)M}P(n) - \frac{m}{(n+1)M}P(n-1), \quad n \neq 0$$

$$P(1) = \frac{m}{M} p(0).$$

用前述的方法可以解得

$$P(2) = \frac{m+M}{2M}P(1) - \frac{m}{2M}p(0) = \frac{m+M}{2M} \frac{m}{M}P(0) - \frac{mM}{2M^2}P(0) = \frac{m^2}{2M^2}P(0).$$

$$P(3) = \frac{m+2M}{3M}P(2) - \frac{m}{3M}P(1) = \frac{m+2M}{3M} \cdot \frac{m^2}{2M^2}P(0) - \frac{m}{3M} \frac{m}{M}P(0)$$

$$= \frac{m^3+2Mm^2}{6M^3}P(0) - \frac{2Mm^2}{6M^3}P(0) = \frac{m^3}{6M^3}P(0).$$

$$P(n) = \frac{m^n}{n!M^n}P(0) = \frac{\rho^n}{n!}P(0).$$

为求出P(0)，对n从0到∞求和

$$\sum_{n=0}^{\infty}P(n) = 1 = p(0)\sum_{n=0}^{\infty} \frac{\rho^n}{n!} = e^{\rho}p(0),$$

所以

$$P(0) = e^{-\rho};$$

且

$$P(n) = \frac{e^{-\rho}\rho^n}{n!} \qquad (23—15)$$

这样，队长是带均值 $\rho = m/M$的普阿松分布。在这种情况下，队长只不过就是在使用的通道数目。等待时间是零。要求多于某一定量通道数的概率（假定如果有无穷多通道被使用的话），可以从普阿松累积分布表来确定。

有限数目的通道，排队长度 在某一时间区间dt内当 $n \leqslant \nu$ 时某个通道的一个输出的概率是nMdt；当$n \geqslant \nu$时是νMdt，方程（23—16a）和（23—16b）同前，但方程（23—16c）不同。

$$P(1) = \frac{m}{M}p(0) \qquad (23—16a)$$

$$P(n+1) = \frac{m+nM}{(n+1)M}p(n) - \frac{m}{(n+1)M}p(n-1) \quad 0<n<\nu \qquad (23—16b)$$

$$P(n+1) = \frac{m+\nu M}{\nu M}p(n) - \frac{m}{\nu M}(n-1) \quad n \geqslant \nu \qquad (23—16c)$$

（23—16b）的解同前面一样，就是

$$P(n)=\frac{\rho^{n}}{n!}p(0)。0<n<\nu \qquad (23—17a)$$

为解方程（23—16c），我们用M除分子、分母：

$$P(n+1)=\frac{\rho+\nu}{\nu}p(n)-\frac{\rho}{\nu}p(n-1), \quad n\geqslant\nu$$

对$n=\nu$

$$P(\nu+1)=\frac{\rho+\nu}{\nu}p(\nu)-\frac{\rho}{\nu}\ \frac{\rho^{\nu-1}}{(\nu-1)!}\ p(0)=\frac{\rho+\nu}{\nu}p(\nu)-\frac{\rho^{\nu}}{\nu!}p(0)$$

$$=\frac{\rho+\nu}{\nu}p(\nu)-p(\nu)=\frac{\rho}{\nu}p(\nu)。$$

对$n=\nu+1$继续

$$P(\nu+2)=\frac{\rho+\nu}{\nu}p(\nu+1)-\frac{\rho}{\nu}p(\nu)=\frac{\rho+\nu}{\nu}\ \frac{\rho}{\nu}p(\nu)-\frac{\rho}{\nu}p(\nu)=\frac{\rho^{2}}{\nu^{2}}p(\nu),$$

且
$$P(\nu+j)=\left(\frac{\rho}{\nu}\right)^{j}p(\nu)=\left(\frac{\rho}{\nu}\right)^{n-\nu}\cdot\frac{\rho^{\nu}}{\nu!}p(0)$$

$$p(n)=\frac{\rho^{n}}{\nu^{n-\nu}\nu!}p(0)。\quad n\geqslant\nu \qquad (23—17b)$$

n从0到∞所有值的概率和一定等于1

$$\sum_{n=0}^{\nu-1}\frac{\rho^{n}}{n!}p(0)+\frac{\rho^{\nu}}{\nu!}\sum_{j=0}^{\infty}\left(\frac{\rho}{\nu}\right)^{j}p(0)=1。$$

如果$\rho\geqslant\nu$，除$P(0)=0$外，第二个和是无限的，这就是说，每单位时间输入的数目大于或等于每单位时间的服务器处理的数目，队长趋于无穷。而任何有穷长队长的概率趋于0。所以象以前一样假定$\rho<\nu$：

$$P(0)\sum_{n=0}^{\nu-1}\frac{\rho^{n}}{n!}+P(0)\ \frac{\rho^{\nu}}{\nu!}\sum_{j=0}^{\infty}\left(\frac{\rho}{\nu}\right)^{j}=1$$

第二个和的首项为1，公比为ρ/ν的无穷几何级数。我们记得，它的和为

$$\frac{1}{1-\frac{\rho}{\nu}}。$$

$$P(0)\left(\sum_{n=0}^{\nu-1}\frac{\rho^{n}}{n!}+\frac{\rho^{\nu}}{\nu!}\cdot\frac{1}{1-\frac{\rho}{\nu}}\right)=1$$

$$P(0)=\frac{1}{\sum_{n=0}^{\nu-1}\frac{\rho^{n}}{n!}+\frac{\rho^{\nu}}{\nu!}\left[\frac{\nu}{\nu-\rho}\right]}。 \qquad (23—18)$$