第六章 排队论

（Queueing Theory）

6.1 基本概念

6.1.1 排队系统的组成

排队是日常生活中经常遇到的现象，如到商店买东西，到医院看病，均是顾客希望得到某种服务，而当某时刻要求服务的顾客数量超过服务机构的容量时，便出现排队现象。这时，排队等待的顾客与服务机构便构成一个排队系统，如图 6.1 所示，现实世界中存在着形形色色的排队系统，如表 6.1 所示。

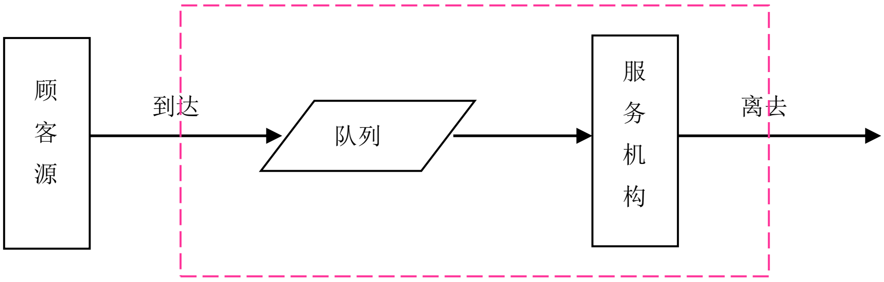

图 6.1 排队系统

表 6.1 现实世界中形形色色的排队系统

到达的顾客	要求服务的内容	服务机构
不能运转的机器	修理	修理技工
修理技工	领取修配零件	发放零件的管理员
电话呼唤	通话	交换台

如果服务设施过少或服务效率太低，便会加剧拥挤，排队成龙。但增加服务设施便会增加服务成本或造成系统空闲，而有些服务设施如机场、港口泊位等一旦建成就不易改动。因此，有必要对排队系统的结构和运行规律加以研究，为排队系统的设计和调控提供依据。

排队系统的基本组成部分主要有输入过程、排队规则、服务机构。

1．输入过程

输入过程是指顾客到达排队系统的过程，对此我们主要讨论两方面内容：

（1）顾客源：分为 • 无限 ∞ (如电话呼唤)

• 有限 m (如车间里待修理的机器)

（2）到达规律：指到达间隔时间 T 的分布

分为 • 定长 D

• 负指数 M

• k 阶爱尔朗 E_k

2．排队规则

排队规则是指顾客到达系统后排队等候服务的方式和规则。可分为三种类型：

（1）损失制

指顾客到达时若所有服务实施均被占用，则顾客自动离去。

（2）等待制

指顾客到达时若所有服务实施均被占用，则留下来等待，直至被服务完离去。等待制的排队规则又可按顾客被服务的次序分为以下几种：

- 先到先服务（FCFS）
- 后到先服务（LCFS）
- 具有优先权的服务（PS）

（3）混合制这是损失制和等待制的混合。这种排队规则既允许排队又不允许队列无限长，主要分为：

- 系统容量有限制
- 等待时间有限制

3．服务机构

对服务机构的研究内容主要有服务设施（称为服务台）的数量和服务的规律。

(1) 服务台个数

$$C\begin{cases}=1\\ >1\ (\text{并列多台})\end{cases}$$

(2) 服务规律：指服务时间v的分布分为：定长 D、负指数 M、k阶爱尔朗 E_k、一般分布 G

6.1.2 排队模型的表示

根据输入过程、排队规则和服务机构的不同情况对排队系统进行描述和归类，可以给出很多排队模型。为了方便对众多模型的描述，康道尔在 1953 年提出一种依据排队系统的三个基本特征对排队模型进行分类表示的方法，称为 Kendall 记号，后来，在 1971 年的一次关于排队符号标准化会议上决定，将 Kendalll 记号扩充为（X/Y/Z/A/B/C），其中

X：顾客到达时间间隔的分布

Y：服务时间的分布

Z：服务台个数

A：系统容量

B：顾客源数量

C：服务规则

（M / M / 1 / ∞/ ∞/ FCFS）表示：到达间隔为负指数分布，服务时间也为负指数分布，1 个服务台，顾客源无限，系统容量也无限，先到先服务。若只讨论先到先服务的情况，可略去第 6 项。

6.1.3 排队问题的求解

研究排队系统的目的是通过了解系统的运行的状况，对系统进行调整和控制，使系统处于最优运行状态。而描述系统运行状况的客观数量指标主要是系统中由于排队和被服务而滞留的顾客数量，以及顾客为等待服务而必须在系统中消耗的时间。因此，排队论中对排队问题的求解，一个基本内容就是研究和计算这些描述系统运行状态的数量指标，简称系统运行指标。这些数量指标一般来说都是随机变量，并且和系统运行的时间t有关。这里主要研究$t\to\infty$时的稳态情形。这时系统处于平衡状态，数量指标的分布等与时间无关，这时求得的结果称为系统处于统计平衡下的解。

1．队长和排队长

队长：系统中的顾客数；其概率分布称状态概率，记为 P_n ，表示系统中有 n 个顾客的概率；队长的平均值记为 L_s。

排队长：系统中正在排队等待的顾客数，记其均值为 L_q。

2．逗留时间和等待时间

逗留时间：一个顾客在系统中的停留时间，记为 W，其均值记为 W_s。

等待时间：一个顾客在系统中排队等待的时间，记其均值为 W_q 。

6.1.4 到达的规律

描述顾客到达规律可从两方面

$$\begin{cases} \text{到达间隔(时间)} \\ \text{到达数(人数)} \end{cases}$$

现实中有许多服务系统，其顾客的到达具有下述特征：（1）无后效性：任一时段的到达数不受前一时段的影响；（2）平稳性：顾客到达是均匀的；（3）稀有性：瞬时内只可能有 1 个顾客到达。

称具有上述特征的输入为泊松流，其在 t 时段内到达 n 个顾客的概率为

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, n = 0,1,\cdots \tag{6-1}$$

即参数为 λt 的泊松分布。

由概率论知识可知，泊松分布的参数即其均值。因此，λ 的含义是单位时间到达系统的平均顾客数，即到达率。下面考察，当顾客按泊松流到达时，其到达的间隔时间 T 是服从什么分布呢？

因为到达为泊松流，所以 t 时段内没有来顾客的概率为

$$P_0(t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \tag{6-2}$$

所以，t 时段内有顾客到来（即间隔 $T \le t$ ）的概率为

$$P(T \le t) = 1 - e^{-\lambda t}，\text{即} F(t) = 1 - e^{-\lambda t}$$

而这正是负指数分布的分布函数，说明 T 服从负指数分布，且参数同为 λ 。可证反之也成立。于是得到关于到达规律的重要性质：到达数为泊松流　　　　到达间隔服从负指数分布（同参数）。

由概率论知识可知，负指数分布的表达式（密度函数）为

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, t \ge 0 \\ 0, \quad\quad t < 0 \end{cases} \tag{6-3}$$

参数 λ 即其均值的倒数。因此，$\dfrac{1}{\lambda}$ 的含义是平均间隔时间，这与 λ 为单位时间到达系统的平均顾客数的含义一致。负指数分布有一个有趣的性质：无记忆性，即

$$P(T > t_0 + t | T > t_0) = P(T > t) \tag{6-4}$$

直观上看，在已知 $T > t_0$ 的条件下估计 $T > t$ 的概率，与无条件时估计 $T > t$ 的概率相同，

把以前的 t_0 时间给忘了。

证：由条件概率公式得

$$
\begin{aligned}
P(T>t_0+t|T>t_0) &= \frac{P(T>t_0+t)\bigcap(T>t_0))}{P(T>t_0)} \\
&= \frac{P(T>t_0+t)}{P(T>t_0)} = \frac{e^{-\lambda(t_0+t)}}{e^{-\lambda t_0}} = e^{-\lambda t} = P(T>t)
\end{aligned}
\tag{6-5}
$$

假若 T 表示某种电子元件的寿命，则当元件已使用了 t_0 时间后估计它还能再使用 t 时间的概率，与它全新时估计用 t 时间的概率一样，即它对已使用了的 t_0 时间无记忆。说明这种元件是高度耐磨损的。

6.1.5 服务的规律

描述排队系统的服务规律主要是采用系统对顾客服务时间 v 的分步。主要讨论服务时间 v 服从负指数分布的情形，参数为 μ，即

$$
f_v(t) = \begin{cases} \mu e^{-\mu t}, t \geq 0 \\ 0, \quad t < 0 \end{cases} \tag{6-6}
$$

由于 v 的均值为 $\frac{1}{\mu}$，即平均对每位顾客的服务时间为 $\frac{1}{\mu}$，可得，参数 μ 的含义：服务率，即单位时间平均服务完 μ 人。注：负指数分布的一般化——爱尔朗分布，可用于描述由 k 道程序组成的 1 个服务台的服务时间的分布。

6.2 M/M/1 排队系统

6.2.1 标准的 M/M/1 模型（M/M/1/∞/∞）

1．问题的一般提法

设：泊松输入/负指服务/单服务台/系统无限制/顾客源无限制

求：（1）系统状态概率 P_n；

（2）系统运行指标 L_s，L_q，W_s，W_q。

2．系统状态概率

（1）利用状态转移图列出平衡方程

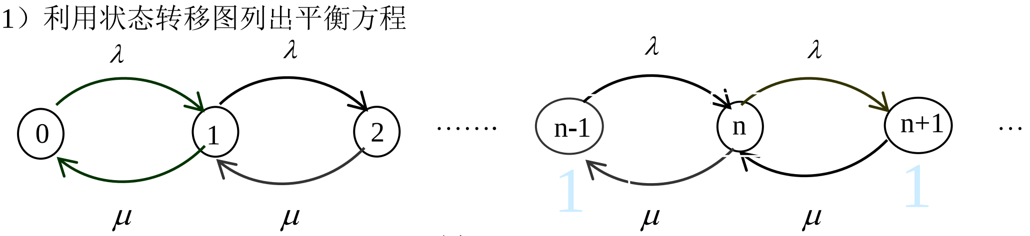

图 6.2

状态转移图是处理稳态 M/M/C 系统的一种工具，设到达与服务率分别为 λ 和 μ，则对于状态 n（$n=1,2,\ldots$），由于系统满足统计平衡，即流入应等于流出，故有：

$$
\lambda P_{n-1} + \mu P_{n+1} = \lambda P_n + \mu P_n \tag{6-7}
$$

对于状态 0，也应有

$$\lambda P_0 = \mu P_1 \tag{6-8}$$

由此列出平衡方程：

$$\begin{cases} \lambda P_0 = \mu P_1 \\ \lambda P_{n-1} + \mu P_{n+1} = (\lambda + \mu) P_n, n \geq 1 \end{cases} \tag{6-9}$$

（2）由平衡方程解得状态概率

由平衡方程$\begin{cases} \lambda P_0 = \mu P_1 \\ \lambda P_{n-1} + \mu P_{n+1} = (\lambda + \mu) P_n, n \geq 1 \end{cases}$

可解得状态概率：

$$\begin{cases} P_0 = 1 - \dfrac{\lambda}{\mu} \\ P_n = (\dfrac{\lambda}{\mu})^n (1 - \dfrac{\lambda}{\mu}), n \geq 1 \end{cases} \tag{6-10}$$

记$\dfrac{\lambda}{\mu} = \rho$，称为服务强度，规定$\rho<1$，则

$$\begin{cases} P_0 = 1 - \rho \\ P_n = \rho^n P_0 \end{cases} \tag{6-11}$$

式中ρ有明确的实际意义，由于$\rho = \dfrac{\lambda}{\mu}$，说明$\rho$是平均到达率与平均服务率之比，由于$\rho = \dfrac{1/\mu}{1/\lambda}$说明$\rho$是平均服务时间与平均到达间隔时间之比，反映了系统的服务强度。而由$\rho = 1 - P_0$说明ρ是系统忙着的概率，反映了系统的繁忙程度和利用率。因此，称ρ为服务强度或服务机构的利用率。

3．系统运行指标

(1) L_S 与 L_q

$\because L_S$ 表示系统中的平均顾客数，由期望定义

$$\begin{aligned} \therefore L_s &= \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n \rho^n (1-\rho) = \rho(1-\rho) \sum_{n=1}^{\infty} n \rho^{n-1} \\ &= \rho(1-\rho) \sum_{n=1}^{\infty} \frac{\mathrm{d}\rho^n}{\mathrm{d}\rho} = \rho(1-\rho) \frac{\mathrm{d}}{\mathrm{d}\rho} \sum_{n=0}^{\infty} \rho^n \\ &= \rho(1-\rho) \frac{\mathrm{d}}{\mathrm{d}\rho} \frac{1}{1-\rho} = \rho(1-\rho) \frac{1}{(1-\rho)^2} \\ &= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \end{aligned} \tag{6-12}$$

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty} (n-1) P_n = \sum_{n=0}^{\infty} n P_n - \sum_{n=1}^{\infty} P_n = L_s - (1 - P_0) \\ &= L_s - \rho \end{aligned} \tag{6-13}$$

其中$\rho=\dfrac{\lambda}{\mu}<1$。问题：为什么$L_s-L_q=\rho<1$（而不是=1）呢？——因为是均值。

（2）W_S 与 W_q

首先可证，逗留时间W服从参数为$\mu-\lambda$的负指数分布，而负指数分布的均值等于其参数的倒数，故平均逗留时间 $W_s=\dfrac{1}{\mu-\lambda}$。

平均等待时间等于平均逗留时间减去平均服务时间，即 $W_q=W_s-\dfrac{1}{\mu}$。

(3)上述 4 个指标之间的关系——里特公式

$L_s=\lambda W_s\quad L_q=\lambda W_q\quad L_s-L_q=\dfrac{\lambda}{\mu}\quad W_s-W_q=\dfrac{1}{\mu}$ 一般的里特公式中λ应为λ_e，称有效到达率，即实际进入系统率。本模型中因系统容量无限制，故$\lambda_e=\lambda$。

例 6.1 某修理店只有一个修理工人，来修理的顾客到达数服从泊松分布，平均每小时 4 人；修理时间服从负指数分布，平均需 6 分钟。求：（1）修理店空闲的概率；（2）店内有 3 个顾客的概率；（3）店内至少有 1 个顾客的概率；（4）店内顾客的平均数；（5）顾客在店内的平均逗留时间；（6）等待服务的顾客平均数；（7）平均等待修理时间；（8）必须在店内消耗 15 分钟以上的概率。

解：此为标准的 M/M/1 模型，λ=4 人/小时，$\mu=\dfrac{1}{6}$人/分钟=10 人/小时，

$$\rho=\frac{\lambda}{\mu}=\frac{2}{5}$$

（1）$P_0=1-\rho=\dfrac{3}{5}$；

（2）$P_3=\rho^3(1-\rho)=(\dfrac{2}{5})^3(\dfrac{3}{5})=0.0384$；

（3）$1-P_0=\dfrac{2}{5}$；

（4）$L_s=\dfrac{\lambda}{\mu-\lambda}=\dfrac{4}{6}=\dfrac{2}{3}$（人/小时）；

（5）$W_s=\dfrac{1}{\lambda}L_s=\dfrac{1}{6}$（人/小时）；

（6）$L_q=L_s-\rho=\dfrac{2}{3}-\dfrac{2}{5}=\dfrac{4}{15}$ （人/小时）；

（7）$W_q=W_s-\dfrac{1}{\mu}=\dfrac{1}{6}-\dfrac{1}{10}=\dfrac{1}{15}$（人/小时）；

（8）$P(W\geq\dfrac{1}{4})=1-P(W<\dfrac{1}{4})=1-F(\dfrac{1}{4})=e^{-(10-4)\frac{1}{4}}=e^{-1.5}=0.223$

6.2.2 系统容量有限的 M/M/1 模型（M/M/1/*N*/∞）

1．与（M/M/1/∞/∞）的区别

（1）系统状态只有 *N*+1 种（$n=0,1,\cdots,N$）；

（2）顾客的实际进入系统的速率$\begin{cases}\lambda, & 当\ \ n<N \\ 0, & 当 n\geq N\end{cases}$，

故平均到达率$\lambda_e=\lambda(1-P_N)+0P_N=\lambda(1-P_N)$

注：由于系统稳定时应达到统计平衡，即进入速率应等于离去速率，故 $\lambda(1-P_N)=\mu(1\text{-}P_0)$。

2．状态概率

首先画出系统的状态转移方程，如图 6.3 所示：

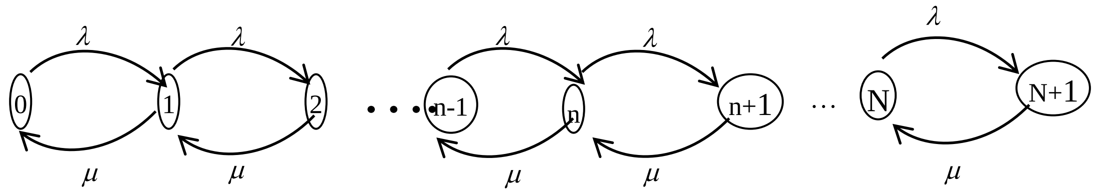

图 6.3

根据系统的状态转移图，由此列出平衡方程：

$$\begin{cases}\lambda P_0=\mu P_1 \\ \lambda P_{n-1}+\mu P_{n+1}=(\lambda+\mu)P_n, n=1,\cdots,N\text{-}1 \\ \lambda P_{N\text{-}1}=\mu P_N\end{cases} \tag{6-14}$$

先解得 $P_n=\rho^n P_0$，

再由$\sum_{n=0}^{N}P_n=P_0+\rho P_0+\cdots+\rho^N P_0=P_0\dfrac{1-\rho^{N+1}}{1-\rho}=1$可解得 P_0，故$\begin{cases}P_0=\dfrac{1-\rho}{1-\rho^{N+1}} \\ P_N=\rho^N P_0\end{cases}$。

当 ρ =1，由平衡方程组可得 P_0=P_1=⋯=P_N=$\dfrac{1}{N+1}$。

状态概率中的 P_N 表示系统满员的概率，因此也可称为系统对顾客的拒绝率或顾客的损失率。

3．系统运行指标

先求出系统的平均队长

$$L_s=\sum_{n=0}^{N}nP_n=\frac{\rho}{1-\rho}-\frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}, \tag{6-15}$$

再由里特公式，求出其他三个指标。

$$L_q=L_s-(1-P_0), \tag{6-16}$$

$$W_s=\frac{L_s}{\lambda_e}, \tag{6-17}$$

$$W_q=\frac{L_q}{\lambda_e}。 \tag{6-18}$$

$\lambda_e=\lambda(1-P_N)$为有效到达率，它的直观解释是，顾客到达且系统未满而实际进入系统率。由统计平衡知，它应等于实际服务率 μ (1—P_0)。它表示系统不空且以 μ 的速率服务着。$\dfrac{\lambda_e}{\mu}$称为本系统的服务强度。

例 6.2 某修理站只有 1 个修理工，且站内最多只能停放 3 台待修理的机器。设待修理的机器按泊松流到达，平均每小时到达 1 台；修理时间服从负指数分布，平均每 1.25 小时

可修理 1 台。试求：（1）站内空闲率；（2）顾客损失率；（3）有效到达率；（4）站内平均队长；（5）机器为修理而需等待的平均时间。

解：此为（$M/M/1/4/\infty$）排队系统，λ=1，μ=0.8，$\rho=\dfrac{1}{0.8}$=1.25。

(1) $P_0=\dfrac{1-\rho}{1-\rho^{4+1}}=\dfrac{1-1.25}{1-1.25^5}=0.122$；

(2) $P_4=\rho^4P_0=1.25^4\times0.122=0.298$；

(3) $\lambda_e=\lambda(1-P_4)=1\times(1-0.298)=0.702$；

(4) $L_s=\dfrac{\rho}{1-\rho}-\dfrac{(4+1)\rho^5}{1-\rho^5}=\dfrac{1.25}{1-1.25}-\dfrac{5\times1.25^5}{1-1.25^5}=2.44(台)$；

(5) $W_q=\dfrac{L_q}{\lambda_e}=\dfrac{L_s-(1-P_0)}{\lambda_e}=\dfrac{2.44-(1-0.122)}{0.702}=2.23(小时)$。

例 6.3 为开办一个小型汽车冲洗站，必须决定提供等待汽车使用的场地大小。设要冲洗的汽车到达服从泊松分布，平均每 4 分钟 1 辆，冲洗的时间服从负指数分布，平均每 3 分钟洗 1 辆。试计算当所提供的场地仅能容纳（a）1 辆；（b）3 辆；（c）5 辆（包括正在被冲洗的 1 辆）时，由于等待场地不足而转向其它冲洗站的汽车的比例。

解：$\dfrac{1}{\lambda}$=4，$\lambda=\dfrac{1}{4}$辆/分钟，$\mu=\dfrac{1}{3}$辆/分钟，$\rho=\dfrac{3}{4}$。

(a) $N=1$，$P_0=\dfrac{1-\rho}{1-\rho^2}=\dfrac{1-\frac{3}{4}}{1-\frac{9}{16}}=\dfrac{4}{7}$，$P_1=\rho P_0=\dfrac{3}{7}=0.428$

(b) $N=3$，$P_0=\dfrac{1-\rho}{1-\rho^4}=\dfrac{1-\frac{3}{4}}{1-\frac{81}{256}}=\dfrac{64}{175}$，$P_3=\rho^3P_0=\dfrac{27}{175}=0.154$

(c) $N=5$，$P_0=\dfrac{1-\rho}{1-\rho^6}=\dfrac{1-\frac{3}{4}}{1-0.178}=0.304$，$P_5=\rho^5P_0=0.237\times0.304=0.072$

6.2.3 顾客源有限的 M/M/1 模型（M/M/1/∞/m）

1. 与（M/M/1/∞/∞）的区别

(1) 系统的状态只有 m+1 种（n=0,1,⋯m）；

(2) λ 的含义 $\begin{cases}(\infty/\infty)\text{: 表全体顾客的平均到达率，} \\ \qquad\text{即单位时间平均到}\lambda\text{人（每人率不同）;} \\ (\infty/m)\text{: 表每个顾客的平均到达率，} \\ \qquad\text{即单位时间每人到}\lambda\text{次（每人率同}\lambda\text{）。}\end{cases}$

$$
\text{(3) 实际的进入率}\begin{cases}
(\infty/\infty)：\lambda\text{（与状态无关）；} \\
(\infty/m)：\begin{cases}(m-n)\lambda\text{（为什么？）}, n<m, \\ 0, n=m,\end{cases} \\
\qquad \text{故平均到达率}\lambda_e = E[(m-n)\lambda] \\
\qquad = E(m\lambda)-E(n\lambda) = m\lambda-\lambda E(n) \\
\qquad = \lambda(m-L_s).
\end{cases}
$$

说明（进入率与状态有关）：如 m=5，n=3，如下图所示

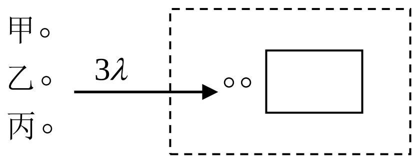

图 6.4

进入的或甲或乙或丙，故 3λ

2. 状态概率

首先画出系统的状态转移图，如图 6.5 所示:

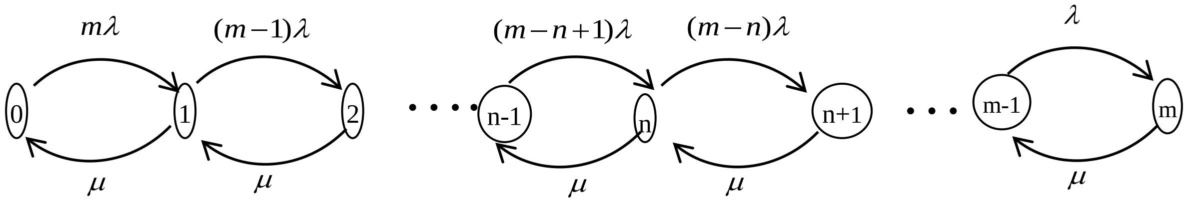

图 6.5 由此列出平衡方程：

$$
\begin{cases}
m\lambda P_0 = \mu P_1 \\
(m-n+1)\lambda P_{n-1} + \mu P_{n+1} = [(m\text{-}n)\lambda + \mu]P_n, n = 1,\cdots,m\text{-}1 \\
\lambda P_m = \mu P_{m-1}
\end{cases} \tag{6-19}
$$

解得：

$$
\begin{cases}
P_0 = \dfrac{1}{\sum\limits_{i=0}^{m} \dfrac{m!}{(m-i)!}(\dfrac{\lambda}{\mu})^i} \\
P_n = \dfrac{m!}{(m-i)!}(\dfrac{\lambda}{\mu})^n P_0, n = 1,\cdots,m
\end{cases} \qquad \text{（(6-20)}
$$

3. 系统运行指标

$\because \lambda_e = \lambda(m-L_s)$，

$\therefore$ 由里特公式，$L_s - L_q = \dfrac{\lambda_e}{\mu} = \dfrac{\lambda(m-L_s)}{\mu}$；

另一方面 $L_s - L_q = 1 - P_0$，

解得 $L_s = m - \dfrac{\mu}{\lambda}(1-P_0)$。

问题：L_s的直观意义为何？

$\lambda(m-L_s)=\mu(1-P_0)$，可以很直观地解释其实际意义：$m-L_s$为系统外的顾客数，$\lambda(m-L_s)$为系统的实际到达率；$1-P_0$为系统忙着的概率，$\mu(1-P_0)$为系统忙时的服务率，由统计平衡，二者应当相等。

例 6.4 某车间有 5 台机器，每台机器的连续运转时间服从负指数分布，平均连续运转时间为 15 分钟。有 1 个修理工，每次修理时间服从负指数分布，平均每次需 12 分钟。（1）修理工空闲的概率；（2）5 台机器都出故障的概率；（3）出故障机器的平均台数；（4）等待修理机器的平均台数；（5）每台机器的平均停工时间；（6）每台机器的平均等待修理时间。

解： 此为（M/M/1/∞/5）排队系统 $\lambda=\frac{1}{15}$，$\mu=\frac{1}{12}$，$\rho=\frac{12}{15}=0.8$。

$$\begin{cases} P_0=\dfrac{1}{\sum\limits_{i=0}^{m}\dfrac{m!}{(m-i)!}(\dfrac{\lambda}{\mu})^i} \\ P_n=\dfrac{m!}{(m-i)!}(\dfrac{\lambda}{\mu})^n P_0, n=1,\cdots,m \end{cases} \qquad (1)$$

$$P_0=[\frac{5!}{5!}(0.8)^0+\frac{5!}{4!}(0.8)^1+\frac{5!}{3!}(0.8)^2+\frac{5!}{2!}(0.8)^3+\frac{5!}{1!}(0.8)^4+\frac{5!}{0!}(0.8)^5]^{-1}=0.0073; \qquad (2)$$

$P_5=\frac{5!}{0!}(0.8)^5 P_0=0.287$；（3）$L_s=5-\frac{1}{0.8}(1-0.0073)=3.76(台)$；

（4）$L_q=3.76-(1-0.0073)=2.77$（台）；

（5）$W_s=\dfrac{5}{\frac{1}{12}(1-0.0073)}-15=46$（分钟）

$W_q=46-12=34$（分钟）。

由此可对该排队系统做何分析？——机器停工时间过长，修理工几乎没有空闲时间应当提高服务率或增加修理工，或购置高效机器减少需修理率。