排队论方法

基本概念
排队模型问题分类
排队问题求解
排队系统的优化

排队论(Queuing Theory)，又称随机服务系统理论(Random Service System Theory),是一门研究拥 挤现象(排队、等待)的学科。

排队论内容主要有三部分:

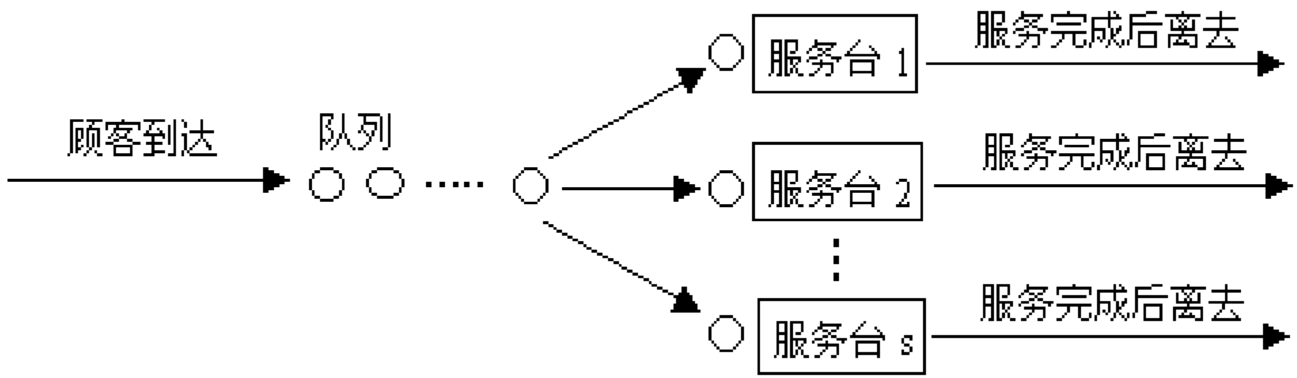
(1)性态问题:研究排队系统的概率分布规律,主要研究**队长分布,等待时间分布和忙期分布**等;

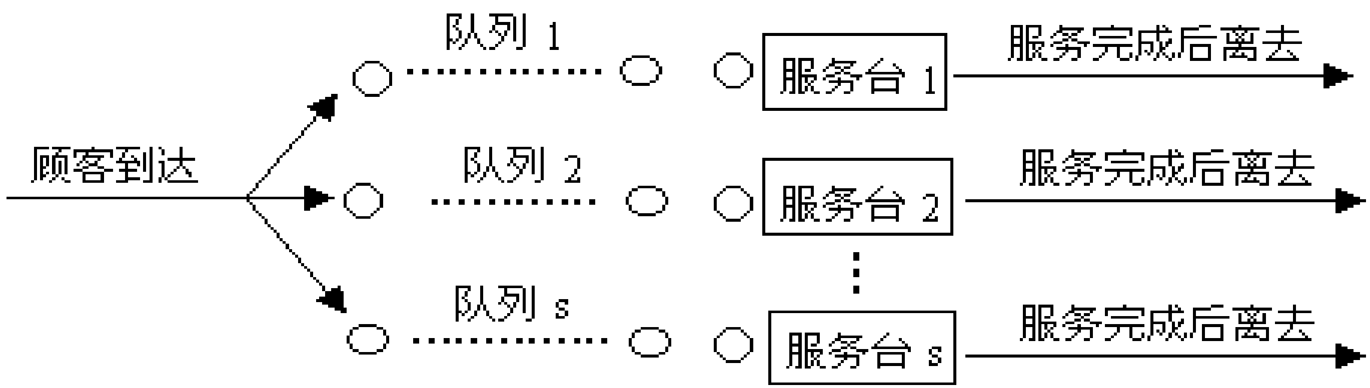
(2)最优化问题:分为静态最优化和动态最优化,即系统的**最优设计和最优运营问题**;

(3)排队系统的统计推断:判断一个给定的排队系统属于哪种类型,以进行研究.

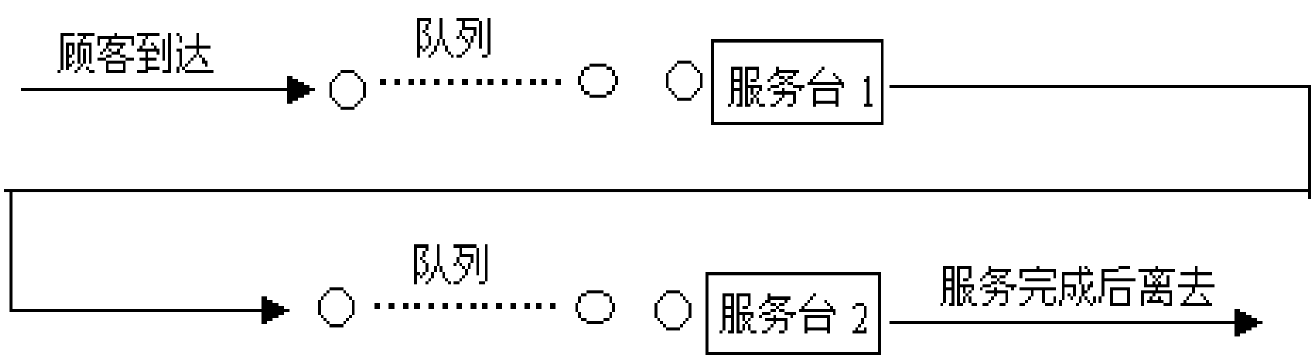
1.排队论基本概念

实际上排队的过程可由下面几个图理解:

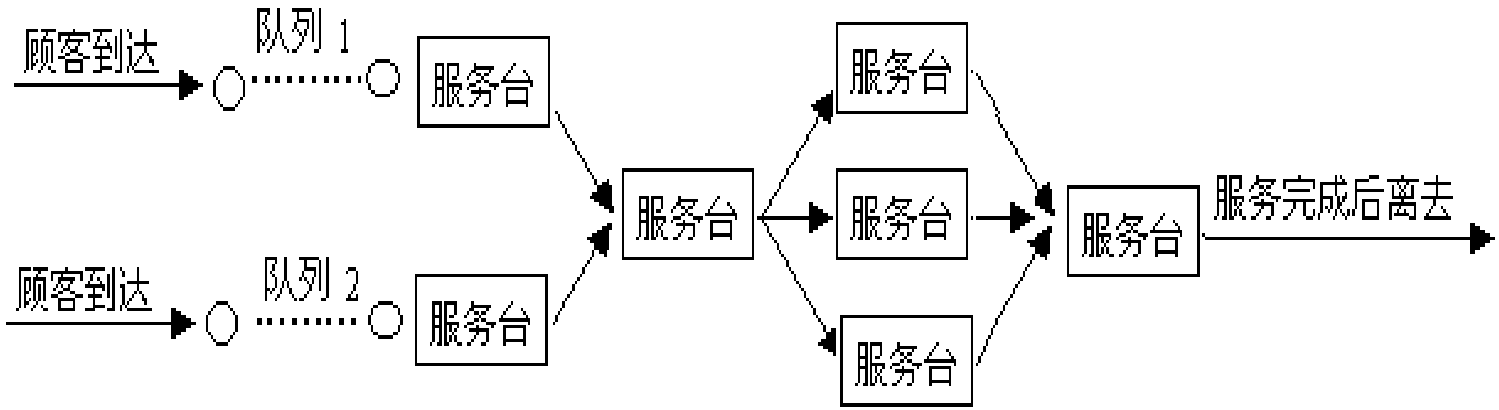


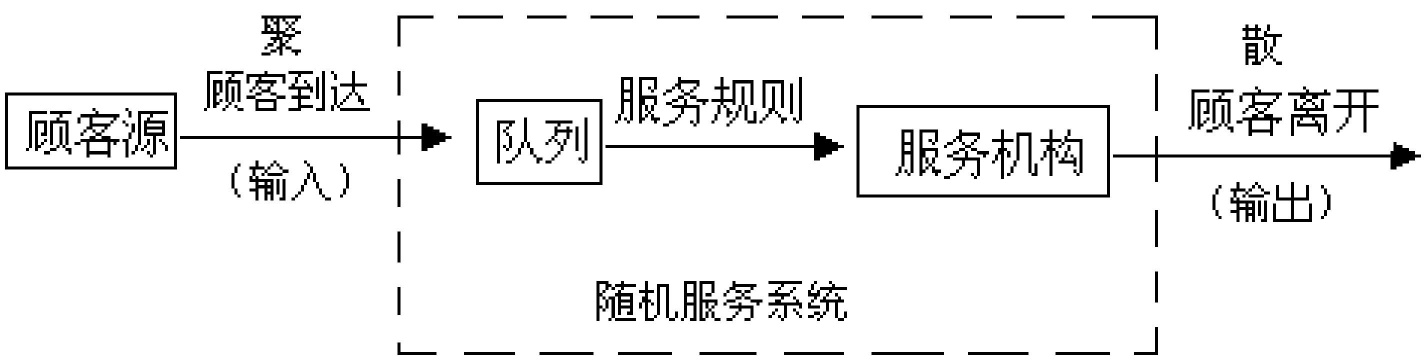


队列 1
队列 2
队列 s
顾客到达
服务台 1
服务台 2
服务台 s
服务完成后离去
服务完成后离去
服务完成后离去



顾客到达
队列
服务台 1
队列
服务台 2
服务完成后离去


顾客到达
队列 1
服务台
顾客到达
队列 2
服务台
服务台
服务台
服务台
服务台
服务台
服务完成后离去


聚
顾客到达
顾客源
（输入）
队列
服务规则
服务机构
散
顾客离开
（输出）
随机服务系统

1. 排队论基本概念

1．输入过程－－顾客到达排队系统的过程.

(1) 顾客总体数，又称顾客源、输入源。这是指顾客的来源。

顾客源可以是有限的，也可以是无限的。

例如，到售票处购票的顾客总数可以认为是无限的，而某个工厂因故障待修的机床则是有限的。

(2)**顾客到达方式。**这是描述顾客是怎样来到系统的，他们是单个到达，还是成批到达。病人到医院看病是顾客单个到达的例子。在库存问题中如将生产器材进货或产品入库看作是顾客，那么这种顾客则是成批到达的

(3) 顾客相继到达的时间间隔可能是确定型的，也可能是随机型的.

(4)顾客到达是相互独立的.

(5)相继到达的时间间隔与时间无关

2.排队规则

(1) 即时制也称损失制。这是指如果顾客到达排队系统时，所有服务台都已被先来的顾客占用，那么他们就自动离开系统永不再来。典型例子是，如电话拨号后出现忙音，顾客不愿等待而自动挂断电话，如要再打，就需重新拨号，这种服务规则即为即时制

(2) 等待制。这是指当顾客来到系统时，所有服务台都不空，顾客加入排队行列等待服务。例如，排队等待售票，故障设备等待维修等。等待制中，服务台在选择顾客进行服务时，常有如下四种规则：

①先到先服务。按顾客到达的先后顺序对顾客进行服务，这是最普遍的情形。

②后到先服务。仓库中迭放的钢材，后迭放上去的都先被领走，就属于这种情况。

③**随机服务**。即当服务台空闲时，不按照排队序列而随意指定某个顾客去接受服务，如电话交换台接通呼叫电话就是一例。

④**优先权服务**。如老人、儿童先进车站；危重病员先就诊；遇到重要数据需要处理计算机立即中断其他数据的处理等，均属于此种服务规则。

(3) **混合制**．这是等待制与损失制相结合的一种服务规则，一般是指允许排队，但又不允许队列无限长下去。具体说来，大致有三种：

① 队长有限。当排队等待服务的顾客人数超过规定数量时，后来的顾客就自动离去，另求服务，即系统的等待空间是有限的。例如最多只能容纳K个顾客在系统中，当新顾客到达时，若系统中的顾客数(又称为队长)小于K，则可进入系统排队或接受服务；否则，便离开系统，并不再回来。如水库的库容是有限的，旅馆的床位是有限的。

② 等待时间有限。即顾客在系统中的等待时间不超过某一给定的长度 T，当等待时间超过 T 时，顾客将自动离去，并不再回来。如易损坏的电子元器件的库存问题，超过一定存储时间的元器件被自动认为失效。又如顾客到饭馆就餐，等了一定时间后不愿再等而自动离去另找饭店用餐。

③ 逗留时间(等待时间与服务时间之和)有限。例如用高射炮射击敌机，当敌机飞越高射炮射击有效区域的时间为t时，若在这个时间内未被击落，也就不可能再被击落了。

不难注意到，损失制和等待制可看成是混合制的特殊情形.

服务机构

①单队——单服务台式；
②单队——多服务台并联式；
③多队——多服务台并联式；
④单队——多服务台串联式；
⑤单队——多服务台并串联混合式, 以及
多队——多服务台并串联混合式等等。

见前面图1至图5所示。

（三）排队系统的描述符号与分类

为了区别各种排队系统，根据输入过程、排队规则和服务机制的变化对排队模型进行描述或分类，可给出很多排队模型。为了方便对众多模型的描述，肯道尔（D．G．Kendall）提出了一种目前在排队论中被广泛采用的“Kendall记号”，完整的表达方式通常用到6个符号并取如下固定格式：

X/Y/Z/A/B/C

各符号的意义为：

X表示相继到达间隔时间的分布，Y表示服务时间的分布，X,Y取值有以下几种情况：

M——表示泊松过程或负指数分布；

D——表示确定型分布；

E_k——表示k阶爱尔朗分布；

G——表示一般服务时间分布;

GI——表示一般相互独立的时间间隔分布.

Z—表示服务台(员)个数：“1”则表示单个服务台，“s”。($s>1$)表示多个服务台。

A—表示系统中容量限额，或称等待空间容量；

B—表示顾客源数目，分有限与无限两种，∞表示顾客源无限，此时一般∞也可省略不写。

C—表示服务规则

C—表示服务规则，常用下列符号：

(1)FCFS：表示先到先服务的排队规则；

(2)LCFS：表示后到先服务的排队规则；

(3)PR：表示优先权服务的排队规则。

(4)随机服务

省略时代表的是FCFS（先到先服务）。

二、排队系统的主要数量指标

研究排队系统的目的是通过了解系统运行的状况，对系统进行调整和控制，使系统处于最优运行状态。因此，首先需要弄清系统的运行状况。描述一个排队系统运行状况的主要数量指标有：

1. 队长和排队长（队列长）

队长是指系统中的平均顾客数(排队等待的顾客数与正在接受服务的顾客数之和)，

排队长是指系统中正在排队等待服务的平均顾客数。

队长和排队长一般都是随机变量。我们希望能确定它们的分布，或至少能确定它们的平均值(即平均队长和平均排队长)及有关的矩(如方差等)。队长的分布是顾客和服务员都关心的，特别是对系统设计人员来说，如果能知道队长的分布，就能确定队长超过某个数的概率，从而确定合理的等待空间。

2．等待时间和逗留时间

从顾客到达时刻起到他开始接受服务止这段时间称为**等待时间**，是随机变量，也是顾客最关心的指标，因为顾客通常希望等待时间越短越好。从顾客到达时刻起到他接受服务完成止这段时间称为**逗留时间**，也是随机变量，同样为顾客非常关心。对这两个指标的研究当然是希望能确定它们的分布，或至少能知道顾客的平均等待时间和平均逗留时间。

3．忙期和闲期

忙期是指从顾客到达空闲着的服务机构起，到服务机构再次成为空闲止的这段时间，即服务机构连续忙的时间。这是个随机变量，是服务员最为关心的指标，因为它关系到服务员的服务强度。与忙期相对的是闲期，即服务机构连续保持空闲的时间。在排队系统中，忙期和闲期总是交替出现的。

4.损失率和服务强度

损失率:由于系统条件限制,使顾客被拒绝服务而使服务部门受到损失的概率.

服务强度:

绝对通过能力A,表示单位时间内被服务完顾客的均值，或称为平均服务率.

相对通过能力Q,表示单位时间内被服务完客户数于请求服务客户数的比值.

1.4系统状态的概率

系统状态是求运行指标的基础,所谓系统状态是指系统中顾客的数量.如果系统中有n个顾客,则说系统的状态为n,即可能取值为:

(1) 当队长无限制时,则n=0,1,2,…..;

(2) 当队长有限制,且最大值为N时,则n=0,1,2,….N;

(3) 当服务台个数为c,且服务为即时制时，则n=0,1,2,…c

一般来说状态取值与时间t有关,因此在时刻t系统状态取值为n的概率记为Pn(t) ,若Pn(t) →Pn 则称为稳态(或统计平衡状态)解.实际中多数平衡问题都是属于稳态的情况,并不是真正的t→∞,即过一段时间以后就有Pn(t) →Pn.

2.到达时间的间隔分布和服务时间分布

到达时间的间隔分布和服务时间分布一般分为泊松分布,负指数分布和爱尔朗分布.

2.1 泊松分布

设N(t)表示在时间段[0,t)内的到达的顾客数,Pn(t1,t2)表示在时间段[t1,t2)(t2＞t1)内有n(n≥0)个顾客达到的概率,即Pn(t1,t2)=P{N(t2)-N(t1)=n},当Pn(t1,t2)满足如下三个条件时,则称顾客到达形成泊松流;

(1) 无后效性:在不相交的时间区间内顾客到达是相互独立的,即在[t,t+△t]到达的顾客与时刻t前到达的顾客数无关.

(2)平稳性：对于充分小的△t，在时间段[t,t+△t]内有1个顾客到达的概率只与时间段长度△t有关,而与起始时间t无关,且$P_1(t,t+\triangle t)=\lambda t+o(\triangle t)$,其中$\lambda>0$称为概率强度(或平稳流强度),即单位时间内有一个顾客到达的概率.

（3）普通性:对于充分小的△t,在[t,t+△t]内有2个或2个以上顾客到达的概率极小,即

$$\sum_{n=2}^{\infty} P_n(t, t+\Delta t) = o(\Delta t)$$

下面研究系统状态概率n的分布.

如果取时间段的 初始时间为t=0,则可记为 $P_n(0,t)=P_n(t)$,在[t,t+△t)内,由于

$$\sum_{n=0}^{\infty} P_n(t,t+\Delta t) = P_0(t,t+\Delta t)+P_1(t,t+\Delta t)$$

$$+\sum_{n=2}^{\infty} P_n(t,t+\Delta t) = 1$$

故在[t,t+△t)内没有顾客到达的概率为

$$P_0(t,t+\Delta t)=1-P_1(t,t+\Delta t)-\sum_{n=2}^{\infty}P_n(t,t+\Delta t)$$

$$=1-\lambda\Delta t+o(\Delta t)$$

将 [0,t+△t)分为[0,t)和[t,t+ △t),则在时间段[0,t+△t)达到n个顾客的概率为

$$
\begin{aligned}
P_n(t+\Delta t) &= \{N(t+\Delta t) - N(0) = n\} \\
&= \sum_{k=0}^{n} P\{N(t+\Delta t) - N(t) = k\} \\
&\qquad\qquad \cdot P\{N(t) - N(0) = n-k\} \\
&= \sum_{k=0}^{n} P_{n-k}(t) \cdot P_k(t, t+\Delta t) \\
&= P_n(t)(1-\lambda\Delta t) + P_{n-1}(t)\lambda\Delta t + o(\Delta t),
\end{aligned}
$$

即

$$\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}=-\lambda P_n(t)+\lambda P_{n-1}(t)+\frac{o(\Delta t)}{\Delta t},$$

令$\Delta t\to 0$,则

$$\begin{cases}\dfrac{dP_n(t)}{dt}=-\lambda P_n(t)+\lambda P_{n-1}(t)\\ P_0(0)=1\end{cases}$$

由上式解得$P_0(t)=e^{-\lambda t}$,代入上两式解得$P_1(t)=\lambda t e^{-\lambda t}$

$$P_2(t)=\frac{(\lambda t)^2}{2!}e^{-\lambda t}$$

一般的有 $P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t}(n=0,1,2,\cdots t>0)$

表示在长为t的时间内到达n个顾客的概率, 即泊松分布,
数学期望方差分别为$E[N(t)]=\lambda t$, $D[N(t)]=\lambda t$.

2.2 负指数分布

当顾客流为泊松分布时,用T表示两个相继到达的时间间隔,是一个随机变量,其分布函数

$$F_T(t) = P\{T \le t\}$$

$$= 1 - P\{T > t\} = 1 - P_0(t)$$

由上可知$P_0(t)=e^{-\lambda t}$，于是$F_T(t)=1-e^{-\lambda t}, t\geq 0$.这里$\lambda$表示单位时间内平均到达的顾客数，则$E(T)=\frac{1}{\lambda}$表示相继顾客到达平均间隔时间

类似的,设系统对一个顾客的服务时间为v(即在忙期内相继离开系统的两个顾客的间隔时间)服从于负指数分布,分布函数为

$F_v(\mathrm{t}) = 1 - \mathrm{e}^{-\mu \mathrm{t}}, \mathrm{t} \geq 0$.分布密度为$f_v(t) = \mu \mathrm{e}^{-\mu \mathrm{t}}, \mathrm{t} \geq 0$.其中$\mu$为平均服务率(即单位时间内被服务完的顾客数),其期望值为$E(v) = \dfrac{1}{\mu}$,表示一个顾客的服务时间.

因此T服从于负指数分布，即与概率强度为λ的泊松流等价. 并且注意到，由条件概率可知：$P\{T>t+s|T>s\}=P\{T>t\}$即说明后一个顾客到来所需时间与前一个顾客所需时间无关，故T无记忆性. 于是，在排队模型的记号中都用M表示，且$E(T)=\frac{1}{\lambda}$，$D(T)=\frac{1}{\lambda^2}$.

2.3爱尔朗分布

设有如下顾客流,记k个顾客达到系统的时间间隔 $v_1, v_2 \cdots v_n$

同时服从于ku的负指数分布,则随机变量

$$T = \sum_{i=1}^{k} v_i$$

服从k阶爱尔朗分布.

当k=1时爱尔朗分布为负指数分布

爱尔朗分布的密度函数：

$$f_k(t) = \frac{\mu k(ukt)^{k-1}}{(k-1)!} e^{-ukt}$$

注： $\mathrm{E}[\mathrm{v}_{\mathrm{i}}]=\frac{1}{\mathrm{k}\mu}(i=1,2,\cdots,k)$,则$\mathrm{E}(T)=$

$\sum_{i=1}^{k} E[\mathrm{v}_{\mathrm{i}}]=\frac{1}{\mu}$,且$\mathrm{D(T)}==\frac{1}{\mathrm{k}\mu^2}$。

当k=1时，即为负指数分布，因此爱尔朗分布式比负指数分布更广泛。

3.单服务台模型

单服务台模型分为3种

设系统输入过程服从泊松流,服务时间服从负指数分布,单服务台:

(1)标准型:*M/M/1/* ∞ /∞

(2)系统容量有限型:*M/M/1/* N /∞

(3)顾客源有限型:*M/M/1/* ∞ /m

3.1 标准型：M/M/1

标准型表示顾客源为无限的，顾客的到达相互独立，到达规律服从参数为λ的泊松分布：单服务台、对长无限、先到先服务；各顾客的服务时间相互独立，且同服从于参数为μ的负指数分布。

1.确定系统在任意时刻t状态为n的概率

因为到达规律服从于参数为λ泊松分布,服务时间服从于参数为μ的负指数分布于是在［t, t+Δt］内有:

(1)有一个顾客到达的概率$\lambda\Delta t+o(\Delta t)$;

(2)没有一个顾客到达的概率$1-\lambda\Delta t+o(\Delta t)$;

(3)有一个顾客被服务完的概率$\mu\Delta t+o(\Delta t)$;

(4)没有一个顾客被服务完的概率$1-\mu\Delta t+o(\Delta t)$;

(5)多于一个顾客到达或被服务完的概率$o(\Delta t)$;

现在考虑在t+△t时刻系统中有n个顾客（即状态为n）的概率Pn(t +△t)，如下表：

情况	时刻t的顾客数	在区间(t,t+△t)		时刻(t,t+△t)的顾客数	$P_n(t+\triangle t)$
		到达	离去		
A	n	0	0	n	$P_n(t)(1-\lambda\triangle t)(1-u\triangle t)$
B	n+1	0	1	n	$P_{n+1}(t)(1-\lambda\triangle t)(1-u\triangle t)$
C	n-1	1	0	n	$P_{n-1}(t)(\lambda\triangle t)(1-u\triangle t)$
D	n	1	1	n	$P_n(t)(\lambda\triangle t)(u\triangle t)$

于是我们得到

$$P_n(t+\Delta t) = P_n(t)(1-\lambda\Delta t-\mu\Delta t)+P_{n+1}(t)\mu\Delta t$$

$$+P_{n-1}(t)\lambda\Delta t+o(\Delta t)$$

两端移项除以Δt,当$\Delta t \to 0$时,

$$\frac{\mathrm{d}P_n(t)}{dt} = P_{n-1}(t)\lambda + P_{n+1}(t)\mu - (\lambda+\mu)P_n(t)$$

当$\mathrm{n}=0$时，类似地可有$\dfrac{\mathrm{d}P_0(t)}{\mathrm{d}t} = -\lambda P_0(t)+\mu P_1(t)$

由于$P_n(t)$与时间t无关, 故

其稳态概率方程为

$$\begin{cases} -\lambda P_0 + \mu P_1 = 0 \\ \lambda P_{n-1} + \mu P_{n+1} - (\lambda + \mu) P_n = 0, n > 1 \end{cases}$$

由递推关系 $\mathrm{P_n} = (\frac{\lambda}{\mu})^{\mathrm{n}} P_0$

令$\rho = \dfrac{\lambda}{\mu} < 1$即平均到达率与平均服务率之比,

由概率性质:$\sum\limits_{n=0}^{\infty} \mathrm{P}_n = 1$即$\mathrm{P}_0 \sum\limits_{n=0}^{\infty} \rho^n = \dfrac{\mathrm{P}_0}{1-\rho} = 1$

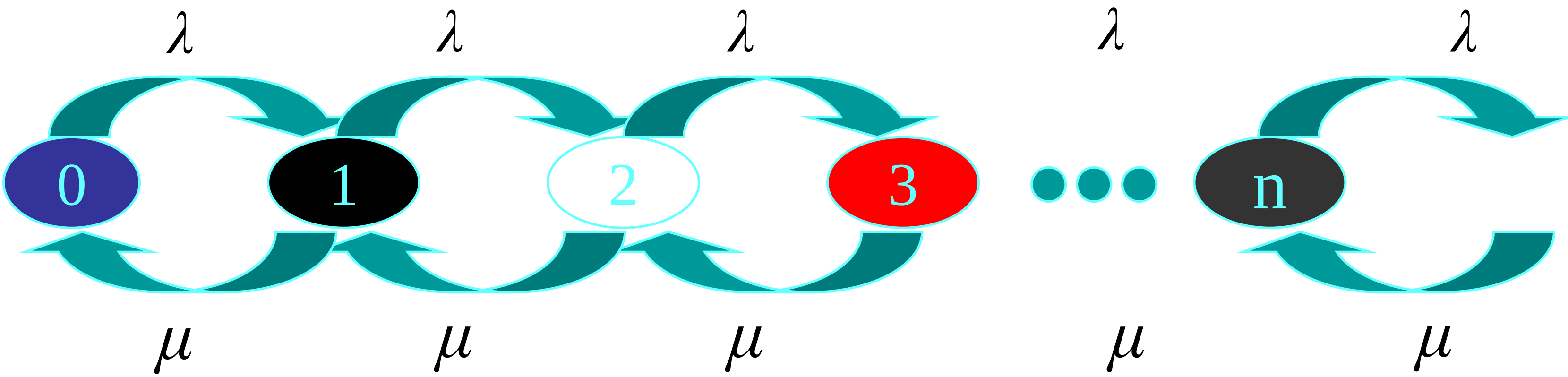
所以$\mathrm{P}_0 = 1 - \rho$

$P_n = (1-\rho)\rho^n, n \geq 1$

这就是系统状态为n的概率。

M/M/1　无限源系统

系统状态转移关系图：



2.系统运行指标

队长: $L_s = \sum_{n=0}^{\infty} nP_n = \frac{\rho}{1-\rho} = \frac{\lambda}{1-\lambda}$

队列长: $L_q = \sum_{n=1}^{\infty} (n-1)P_n = L_s - \rho = \frac{\rho^2}{1-\rho} = \frac{\rho\lambda}{\mu-\lambda}$

顾客逗留时间: $w_s = E(w) = \frac{1}{\mu-\lambda}$ (逗留时间 $F(w) = 1 - e^{-(\mu-\lambda)w}$)

顾客等待时间: $w_q = w_s - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}$

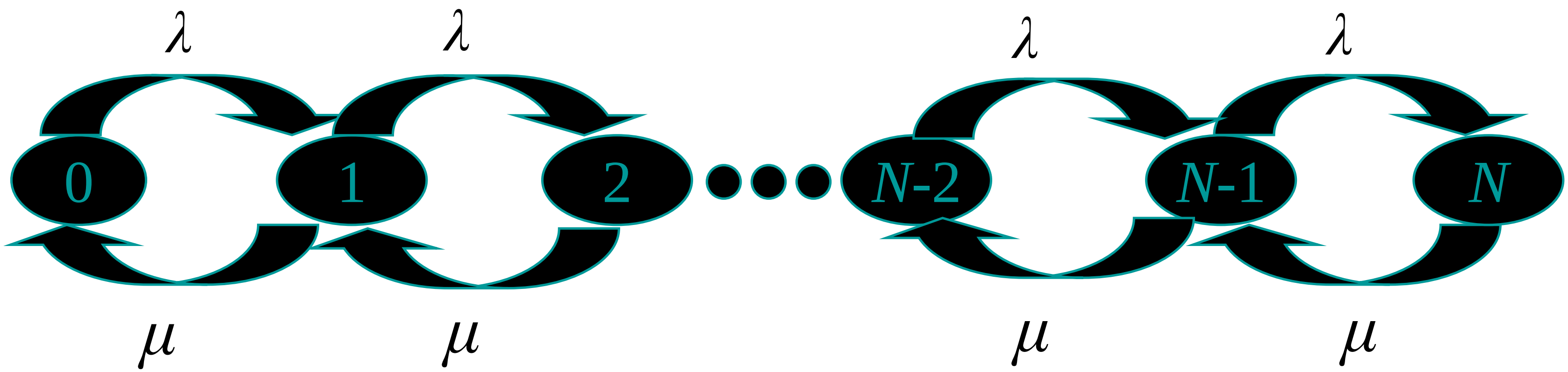
运行指标之间的关系

$$L_s = \lambda w_s, L_q = \lambda w_q, w_s = w_q + \frac{1}{\mu}, L_s = L_q + \frac{\lambda}{\mu}$$

这些关系式称为Little公式.

3.2系统容量有限制:M/M/1/N/∞

•



3.2系统容量有限制:M/M/1/N/∞

1．系统状态概率

(1)稳态方程为

$$
\begin{cases}
\mu P_1 = \lambda P_0 \\
\mu P_{n+1} + \lambda P_{n-1} = (\lambda + \mu) P_n (1 \le n \le N-1), \\
\mu P_N = \lambda P_{N-1}
\end{cases}
$$

注意到$\sum_{n=0}^{N} P_n = 1$由递推关系得

$$\begin{cases} P_0 = \dfrac{1-\rho}{1-\rho^{N+1}}, \rho \neq 1, \\ P_n = \dfrac{1-\rho}{1-\rho^{N+1}} \rho^n, 1 \leq \rho \leq N \end{cases}$$

注意如果$\rho = 1$说明达到率和服务率相等,不会出现排队现象.

当$\lambda > \mu$时,表示到达率比服务率大, 系统损失增加.

2.系统运行指标

(1) 队长：$L_s = \sum_{n=0}^{N} nP_n = \frac{1}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}, \rho \neq 1$

(2) 队列长：$\mathrm{L_q} = \sum_{n=1}^{N} (n-1)P_n = L_s - (1-P_0);$

(3)顾客逗留时间:首先要注意到一个事实,因为$\mathrm{w_s}$与平均到达率λ有关,而λ表示系统有空时的平均到达率, 当系统满员时, 则到达率为0, 为此引入有效到达率λ_{e},表示有空时平均到达率λ减去满员后拒绝顾客的平均数λP_N,即$\lambda_{\mathrm{e}} = \lambda(1-P_N)$,

由于$\lambda_{\mathrm{e}}=\lambda(1-P_N)=\lambda\left(1-\frac{1-\rho}{1-\rho^{N+1}}\rho^N\right)=\lambda\frac{1-\rho^N}{1-\rho^{N+1}}$

$\mu=\frac{\lambda P_0}{P_1}=\frac{1-\rho^{N+1}}{(1-\rho)\rho}\lambda P_0=\frac{\lambda}{\rho}$

所以有效服务强度:$\frac{\lambda_{\mathrm{e}}}{\mu}=\lambda\frac{1-\rho^N}{1-\rho^{N+1}}\frac{\rho}{\lambda}=\frac{\rho-\rho^{N+1}}{1-\rho^{N+1}}=1-P_0$

于是顾客逗留时间:$\mathrm{w_s}=\frac{L_{\mathrm{s}}}{\lambda_{\mathrm{e}}}=\frac{L_{\mathrm{s}}}{\lambda(1-P_0)}=\frac{L_{\mathrm{q}}}{\lambda(1-P_N)}+\frac{1}{\mu}$

(4)顾客等待时间:$\mathrm{w_q}=\frac{L_{\mathrm{q}}}{\lambda_{\mathrm{e}}}=\frac{L_{\mathrm{q}}}{\lambda(1-P_N)}=\mathrm{w_s}-\frac{1}{\mu}$.

3.3顾客源为有限的:M/M/1/∞/m

对顾客总体只有m个顾客,但每个顾客来到并接受服务后,仍然回到顾客总体,即可以再次到来,所以对顾客的容量是没有限制的,实际上系统中顾客永远不会超过m,即与模型.

M/M/1/m/m是相同的,于是我们得到系统概率平衡方程:

$$\begin{cases} \mu P_1 = m\lambda P_0 \\ \mu P_{n+1} + (m-n+1)\lambda P_{n-1} = [(m-n)\lambda + \mu] P_n (1 \le n \le m-1) \\ \mu P_m = \lambda P_{m-1} \end{cases}$$

注意到$\sum_{n=0}^{m} P_n = 1$,由递推关系可得系统状态概率：

$$\begin{cases} P_0 = \dfrac{1}{\sum_{i=0}^{m} \dfrac{m!}{(m-i)!} (\dfrac{\lambda}{\mu})^i}, \\ P_N = \dfrac{m!}{(m-n)!} (\dfrac{\lambda}{\mu})^n P_0, 1 \le n \le m \end{cases}$$

系统运行指标为:

$$L_s=\sum_{n=0}^{m} nP_n=m-\frac{u}{\lambda}(1-P_0),$$

$$w_s=\frac{m}{\mu(1-P_0)}-\frac{1}{\lambda},$$

$$w_q=w_s-\frac{1}{\mu},$$

$$L_q=\sum_{n=1}^{\infty}(n-1)P_n=m-\frac{(\lambda+\mu)(1-P_0)}{\lambda}=L_s-(1-P_0).$$

3.3 多服务台的排队模型

多服务台主要分为三种情况:

(1)标准型：M/M/C

(2)系统容量有限型：M/M/C/N/∞

(3)顾客源为有限的：M/M/C/ ∞/m

4.1标准型：M/M/C

前提假设同单服务台型,另外假设顾客流为泊松流,平均到达率为λ, 各服务台的服务时间满足负指数分布, 而各服务台服务是相互独立的, 单个服务台的平均服务率是μ, 则整个服务机构的平均服务率是$c\mu(\mathrm{n} < c)$,令$\rho = \dfrac{\lambda}{c\mu}$,称为系统服务强度,当$\rho > 1$时,会出现排队现象.

类似的系统状态平衡方程为:

$$\begin{cases} \mu P_1 = \lambda P_0 \\ (n+1)\mu P_{n+1} + \lambda P_{n-1} = (\lambda + n\mu) P_n (1 \le n \le c) \\ c\mu P_{n+1} + \lambda P_{n-1} = (\lambda + c\mu) P_n (n > c) \end{cases}$$

其中$\sum_{n=0}^{\infty} P_n = 1$,且$\rho = \frac{1}{c\mu} \le 1$,由递推关系可求得

系统状态概率$\mathrm{P}_0 = [\sum_{k=0}^{c-1} \frac{1}{k!} (\frac{\lambda}{\mu})^k + \frac{1}{c!} \frac{1}{1-\rho} (\frac{\lambda}{\mu})^c]^{-1}$,

$$P_n = \begin{cases} \frac{1}{n!} (\frac{\lambda}{\mu})^n P_0, \text{当} n \le c, \\ \frac{1}{c!} \frac{1}{c^{n-c}} (\frac{\lambda}{\mu})^n P_0, \text{当} n > c. \end{cases}$$

系统运行指标:

$$\mathrm{L_s} = L_q + \frac{\lambda}{\mu}$$

$$L_q = \sum_{n=c+1}^{\infty}(n-c)P_n = \sum_{k=1}^{\infty} kP_{k+c} = \sum_{k=1}^{\infty}\frac{k}{c!c^k}(cp)^{k+c}P_0 = \frac{(c\rho)^c\rho}{c!(1+\rho)^2}P_0$$

其中$\mathrm{c}\rho = \frac{\lambda}{\mu}, k = n-c, w_s = \frac{\mathrm{L_s}}{\lambda} = \frac{(cp)^c\rho}{c!(1-\rho)^2\lambda}P_0 + \frac{1}{\mu}$

$$w_q = \frac{L_q}{\lambda} = \frac{(cp)^c\rho}{c!(1-\rho)^2\lambda}P_0.$$

4.2系统容量有限制

假设系统内有c个服务台,顾客流为泊松流,平均到达率为λ. 各服务台的服务时间服从负指数分布,而工作是独立的,平均服务率为μ. 系统最大容量为N(N≥c),当系统客满时有c个接受服务,N-c在排队,再有顾客来被拒绝服务. 当系统的状态为n时,每个服务台的服务率为μ,则系统的总服务率:

当0<n<c时为nμ;当$n \geq c$时为cμ,令$\rho = \dfrac{\lambda}{c\mu}$为系统服务强度.

类似的,可以得到系统状态概率方程:

$$\begin{cases} \mu P_1 = \lambda P_0 \\ (n+1)\mu P_{n+1} + \lambda P_{n-1} = (\lambda + n\mu)P_n (1 \le n \le c) \\ c\mu P_{n+1} + \lambda P_{n-1} = (\lambda + n\mu)P_n (c \le n < N) \\ \mu P_{N-1} = c\mu P_N \end{cases}$$

其中$\sum_{n=0}^{N} P_n = 1$,且$\rho = \dfrac{\lambda}{c\mu} \le 1$

由递推关系可得系统状态概率为:

$$P_0=\begin{cases}[\sum_{k=0}^{c-1}\frac{1}{k!}(c\rho^k)+\frac{c^c}{c!}\frac{\rho(\rho^c-\rho^N)}{1-\rho}]^{-1},\rho\neq 1\\ \sum_{k=0}^{c-1}\frac{1}{k!}(c^k)+\frac{c^c}{c!}(N-c+1),\rho=1\end{cases}$$

$$P_n=\begin{cases}\frac{1}{n!}(cp)^n P_0,当0\leq n\leq c\\ \frac{c^c}{c!}\rho^n P_0,c<n\leq N\end{cases}$$

系统运行指标：

$$L_s = L_q + c\rho(1-P_N), w_s = w_q + \frac{1}{\mu}, w_q = \frac{L_q}{\lambda(1-P_N)},$$

$$L_q = \sum_{n=c+1}^{N}(n-c)P_N = \frac{(c\rho)^c\rho}{c!(1-\rho)2}P_0[1-\rho^{N-c}(1-\rho)\rho^{N-c}]$$

系统损失率为 $\mathrm{P}_{损} = P_N = \frac{c^c}{c!}\rho^N P_0.$

特别的,当n=c时, 即M/M/C/C/∞,此时系统为即时制服务,不允许顾客在系统内排队,亦即系统的状态概率为:

$$P_0=[\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!}]^{-1},P_n=\frac{(c\rho)^n}{c!}P_0,$$

相应的运行指标为$L_q=W_q=0,W_s=\frac{1}{\mu},L_s=\sum_{n=1}^{c}nP_n=c\rho(1-P_c).$

4.3顾客源有限

假设与前面问题相同, 即c, m, λ, μ, ρ的意义相同,

为系统状态概率平衡方程为:

$$
\begin{cases}
\mu \mathrm{P}_1 = \mathrm{m}\lambda \mathrm{P}_0 \\
(\mathrm{n}+1)\,\mu P_{n+1} + [(m-n+1)\lambda P_{n-1}] = [(m-n)\lambda + n\mu]P_n, (1 \le n \le c) \\
c\mu P_{n+1} + (m-c+1)\lambda P_{n-1} = [(m-c)\lambda + c\mu]P_n \\
\lambda P_{m-1} = c\mu P_m
\end{cases}
$$

由递推关系可得系统状态概率

$$P_0=\frac{1}{m!}[\sum_{k=0}^{c}\frac{1}{k!(m-k)!}(\frac{c\rho}{m})^k+\frac{c^c}{c!}\sum_{k=0}^{c}\frac{1}{(m-k)!}(\frac{\rho}{m})^k]^{-1},\rho=\frac{m\lambda}{c\mu},$$

$$P_n=\begin{cases}\dfrac{m!}{(m-n)!n!}(\dfrac{\lambda}{\mu})^n P_0,\text{当}0\le n\le c,\\[2ex] \dfrac{m!}{(m-n)!c!c^{n-c}}(\dfrac{\lambda}{\mu})^n P_0,\text{当}c<n\le m\end{cases}$$

系统运行指标为:

$$\mathrm{L_s} = \sum_{n=1}^{m} nP_n, \mathrm{L_q} = \sum_{n=c+1}^{m} (n-c)P_n, w_s = \frac{\mathrm{L_s}}{\lambda_c}, w_q = \frac{\mathrm{L_q}}{\lambda_c}$$

有效到达率为$\lambda_e = \lambda(m - \mathrm{L_s})$,且$\mathrm{L_s} = \mathrm{L_q} + \dfrac{\lambda}{\mu}(m - \mathrm{L_s})$.

5．排队系统的最优化问题

1 最优化问题分为俩类:

(1)静态最优化,指服务机构根据一定的质量指标,找出参数最优值,使系统设计最经济.

(2)动态最优化,对已有的队伍系统寻求某一没表函数达到最优的运营机制.

2.费用模型,我们的目标使服务机构成本和等待费用之和最小.

M/M/1 中最优服务率

标准型

设目标函数: $z = c_s\mu + c_w L_s$,即为单位时间内服务成本与顾客等待时间之和的期望值,其中c_s表示当$\mu = 1$(单位时间内完成一个顾客)时服务机构的服务费用,由$L_S = \dfrac{\lambda}{\mu - \lambda}$,则$z = c_s\mu + c_w \dfrac{\lambda}{\mu - \lambda}$,求其最小值.即令$\dfrac{dz}{d\mu} = 0$,则$c_s - c_w \dfrac{\lambda}{(\mu - \lambda)^2} = 0$解出最优解$\mu^* = \lambda + \sqrt{\dfrac{c_w}{c_s}\lambda}$

即为最优服务率

5.2系统容量有限

如果系统中已有N个顾客,则后来的将被拒绝,于是可设P_N为被拒绝的概率,$1-P_N$为接受服务的概率.$\lambda(1-P_N)$表示单位时间内实际进入的客户数, 在稳定状态下即为单位时间内服务完的客户数.

设系统服务完1个顾客收入G元, 于是单位时间内收入的期望值为$\lambda(1-P_N)G$, 则系统的纯利润为:

$z=\lambda(1-P_N)G-c_s\mu=\lambda G\dfrac{1-\rho^N}{1-\rho^{N+1}}-c_s\mu$令$\dfrac{dz}{d\mu}=0$,可解得

$$\rho^{N+1}\frac{N-(N+1)\rho+\rho^{N+1}}{(1-\rho^{N+1})^2}=\frac{c_s}{G}$$

其中$P_N=\dfrac{\rho^N-\rho^{N+1}}{1-\rho^{N+1}},\rho=\dfrac{\lambda}{\mu}$,而$c_s,G,\lambda,N$均为已知,用数值方法求得$\mu^*$

5.3顾客源为有限

设顾客数为m, 单个服务台, 服务时间服从负指数分布, 当服务率μ=1时, 服务机构的成本费为c_s,单位时间内服务完一个顾客收入G元, 单位时间服务完的顾客数为$m-L_s$,则单位时间内的纯利润为

$$z=(m-L_s)G-c_s\mu=\frac{mG}{\rho}\bullet\frac{E_{m-1}(\frac{m}{\rho})}{E_m^2(\frac{m}{\rho})}-c_s\mu$$

$E_m(\frac{m}{\rho}) = \sum_{k=1}^{m} \frac{(\frac{m}{\rho})^k}{k!} e^{-(\frac{m}{\rho})}$为泊松和, $\rho = \frac{\mathrm{m}\lambda}{\mu}$, 令$\frac{dz}{d\mu} = 0$,

则得:

$$\frac{E_{m-1}(\frac{m}{\rho})E_m(\frac{m}{\rho}) + \frac{m}{\rho}[E_{m-1}(\frac{m}{\rho})E_m(\frac{m}{\rho}) - E_m^2(\frac{m}{\rho})]}{E_m^2(\frac{m}{\rho})} = \frac{c_s\lambda}{G}$$

当给定c_s, G, λ, m后, 利用泊松分布表和数值计算可求得最有服务率μ^*

5.3M/M/c中最优服务台数

仅对标准型讨论,在稳态的假设下,单位时间内每个服务台的成本费为c_s,每个顾客停留单位时间的费用为c_w,则单位时间费用的期望值:$z = c_s c + c_w L_s$,其中$L_s = L_s(c)$,即与服务台个数有关. 因此总费用为z=z(c), 记c的最优值c^*则$z(c^*)$为最小费用. 由于c只能取整数, 故z(c)离散函数, 所以只能用边际分析法求解.

事实上, 根据$z(c^*)$为最小值, 可有:

$$\begin{cases} z(c^*) \le z(c^*-1) \\ z(c^*) \le z(c^*+1) \end{cases}$$

由$z = c_s c + c_w L_s$则有

$$\begin{cases} c_s c^* + c_w L_s(c^*) \le c_s(c^*-1) + c_w L_s(c^*-1) \\ c_s c^* + c_w L_s(c^*) \le c_s(c^*+1) + c_w L_s(c^*+1) \end{cases}$$

化简整理得

$$L_s(c^*) - L_s(c^*+1) \le \frac{c_s}{c_w} \le L_s(c^*-1) - L_s(c^*)$$

应用举例

例:兴建一座港口码头，只有一个装卸船只的泊位。要求设计装卸能力。装卸能力单位为（只/日）船数。已知：单位装卸能力的平均生产费用a=2千元，船只逗留每日损失b=1.5千元。船只到达服从泊松分布，平均速率λ=3只/日。船只装卸时间服从负指数分布。目标是每日总支出最少。

解：　　$\lambda=3$　μ待定　模型 $M/M/1/\infty/\infty$

队长　　$L_s=\lambda/(\mu-\lambda)$

总费用 $C=a\mu+bL_s=a\mu+b\lambda/(\mu-\lambda)$

求极值（最小值）

求 $\mathrm{d}c/\mathrm{d}u = a + (-b\lambda)/(\mu-\lambda)^2 = 0$

得: $\mu-\lambda = \pm(b\lambda/a)^{1/2}$（根据题意舍）

$\mu=\lambda+(b\lambda/a)^{1/2}=3+(2.25)^{1/2}=4.5$(只/日)

例6.4: 建造一口码头，要求设计装卸船只的泊位数。已知:预计到达λ=3只/天，泊松流；装卸μ=2只/天，负指数分布。装卸费每泊位每天a=2千元，停留损失费b=1.5千元/日。目标是总费用最少。

- 解:模型 $M/M/c/\infty/\infty$ c待定
- 总费用：$F=ac+bL_s$（c）离散，无法用求导来解。

考虑: $M/M/c/\infty/\infty$

要求:$\rho=\lambda/(c\mu) <1$ 即$c>(\lambda/\mu)=1.5$

讨论　$c=2,3,4\ldots$

$$P_0=\left[\sum_{n=0}^{c-1} c^n/n!\,\rho^n + c^c/c!\,\rho^c/(1-\rho)\right]^{-1}$$

$$L_q = c^c \rho^{c+1}/c!(1-\rho)^2\, P_0$$

$$L = L_q + \lambda/\mu$$

	$\rho=3/4$	$\rho=1/2$	$\rho=3/8$
P_0	1/7=0.14286	4/19=0.21053	40/181=0.22099
Lq	27/14=1.92857	9/83=0.23684	81/1810=0.04475
Ls	24/7=3.42857	33/19=1.73684	1398/905=1.54475
F	64/7=9.14286	327/38=8.60526	9337/905=10.31713