DOI :CNKI :50-1165/N. 20120115. 1809. 011

基于马尔科夫状态转移过程的 $M/M/m$ 排队模型仿真*

曹永荣[1,2], 韩瑞霞[1,2], 胡 伟[3]
(1. 上海交通大学 人文艺术研究院;2. 安泰经济与管理学院;3. 国际与公共事务学院, 上海 200052)

摘要:马尔科夫链是研究排队系统的主要方法,本文在现有 $M/M/m$ 排队理论和排队系统仿真理论基础上,利用 Matlab 建立基于马尔科夫状态转移过程的 $M/M/m$ 排队模型仿真程序。仿真程序在产生初始化参数设定后,利用时钟推进法来模拟空闲服务台和繁忙服务台情况下的服务流程,最后通过 $M/M/m$ 模型特征描述的仿真计算,获得平均等待时间($E[W]$)、平均停机时间($E[DT]$)、平均排队队长 $E[Q]$、系统中的平均客户数($E[L]$) 和可能延迟的概率(Π)5 项重要的特征描述。模拟次数设定为 20 000 次,模拟客户服务率和客户到达率相同,服务台在 3 ~ 6 个的排队系统,并将仿真结果与理论值以及 Queue2. 0 的模拟结果相比较。最终结果显示 $E[W]$、$[DT]$ 和 Π 3 项最重要指标的仿真结果和理论值都极为相近,误差范围小,本研究将为优先权排队系统的仿真研究提供理论依据。

关键词:马尔科夫状态转移过程;$M/M/m$ 排队模型;仿真
中图分类号:O211. 6 **文献标志码**:A **文章编号**:1672-6693(2012)01-0061-06

$M/M/m$ 排队模型是在现实生活中经常遇到的一种随机服务系统,如银行的存取款服务、超市的付款服务和图书馆的借阅台服务等。分析 $M/M/m$ 排队模型的重要工具是马尔科夫过程,它的特点是当过程在时刻 t_0 所处状态已知时,t_0 以后过程所处状态与 t_0 以前过程所处状态无关,这个特性叫无后效性,也叫马尔科夫性。马尔科夫链在数理分析和模拟仿真方面都是研究排队系统的主要方法和手段。$M/M/m$ 模型的马尔科夫链在一维空间无限延伸[1-7]。目前已经有一些单服务台或多服务台排队系统的仿真研究[8-12],离散事件系统仿真通常基于两种时钟推进方式:面向事件和面向时间的仿真时钟推进方式[13-14],所采用的仿真软件有 Extend、Matlab、Witness、Java 等[15-20]。本文在现有研究的基础上,形成基于马尔科夫状态转移的仿真研究方法,并利用 Matlab 进行仿真试验,该研究成果将为优先权排队系统仿真提供理论依据。

1 基于马尔科夫状态转移过程的排队系统仿真过程

$M/M/m$ 排队系统的生灭过程存在平稳分布的充要条件为 $\rho < 1$,状态转移过程如图 1 所示,在分析排队系统时候从状态间的转移关系开始。状态 1 转移到状态 0,即系统中有 1 名客户被服务完了(离去)的转移率为 μp_1;状态 2 转移到状态 1 时,这就是在两个服务台上被服务的客户中有 1 个被服务完成而离去,那么此时的状态转移率为 $2\mu p_2$。同理考虑状态 n 转移到 $n-1$ 的情况,当 $n \leqslant m$ 时,状态转移率为 $n\mu p_n$,当 $n > m$ 时,因为只有 m 个服务台,最多有 m 个客户在被服务,$n-m$ 个客户在等待,因此状态转移率应为 $m\mu p_n$。

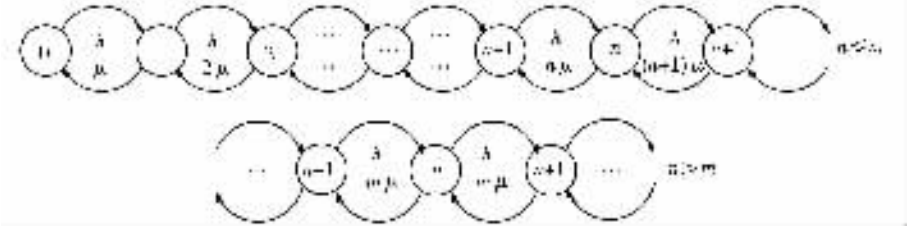

图 1 状态转移过程示意图

* 收稿日期:2011-08-11 网络出版时间:2012-01-15 18:09:00
作者简介:曹永荣,男,博士后,研究方向为工商管理;通讯作者:韩瑞霞,E-mail:annahan@hotmail.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20120115.1809.201201.61_011.html

仿真过程中涉及到的参数可以分为系统状态变量、实体属性、集合变量、活动持续时间变量和累计汇总统计变量等,详细的变量名和变量名代表的意义见表1。

表1 $M/M/m$ 模型仿真参数设定变量名表(部分)

变量名	代表意义	变量名	代表意义
n	程序仿真总次数	m,k	m 为服务台数量,k 为任一服务台
$arr\_interval$	客户平均到达时间间隔	$S(i)$	第 i 客户的服务时间
cus_served	在服务中的顾客数	$csr\_stat(k,i)$	第 i 期第 k 个服务台的忙闲状态
$fin\_time(i)$	第 i 客户服务完毕时刻	$DT(i)$	第 i 客户的停机时间
$E[L]$	客户在系统中的平均数量	$arr\_time(i)$	第 i 客户设备故障出现时刻
$E[Q]$	平均排队队长	$csr\_left(k,i-1)$	第 i 期第 k 个服务台服务完成时刻
$E[W]$	服务台平均响应时间	$sta\_time(k,i-1)$	第 i 期第 k 个服务台服务开始时刻
$E[DT]$	客户的平均设备停机时间	$W(i)$	第 i 客户的服务台响应时间
Π	服务可能延迟的概率	$q(i)$	第 i 期系统中的排队等待的客户

首先,仿真开始需要系统初始化设置和产生随机数矩阵(平均到达时间间隔、平均服务时间和仿真次数)。设 $A=|A_1,A_2,\cdots,A_n|$ 为客户平均到达时间间隔向量,其中 n 为仿真次数,则 n 个客户到达排队系统的时间 $arr\_time(i)$ 应该为

$$arr\_time=\left|A_1,\sum_1^2 A_i,\cdots,\sum_1^n A_i\right|,i=1,2,\cdots,n \tag{1}$$

仿真时钟($time$)由客户到达时间和 m 个服务台的服务完成时间确定,由于有排队现象,假设第 i 次仿真的队长为 $q(i)$,则第 i 次仿真的时钟可以由(2)式和(3)式获得。

$$time(i)=\min\{arr\_time(i),csr\_left(1,i-q(i-1)-1),\cdots,csr\_left(m,i-q(i-1)-1)\} \tag{2}$$

$$time(i)=\min\{arr\_time(i),csr\_left(1,i-q(i)-1),\cdots,csr\_left(m,i-q(i)-1)\} \tag{3}$$

其次,为了能够对所有的服务台进行动态监控,需要对每个服务台进行编号,按各服务台的忙闲状况将它们分为两类:1类和 m 类,l 类为空闲的服务台,m 类为服务繁忙的服务台,k 是从 l 类中抽出来为客户服务的服务台。服务台的分类和编号的目的是在仿真过程中可以跟踪到每个服务台的服务状况。仿真推进基于排队系统的马尔科夫状态转移过程。假设在第 i 次仿真开始时刻排队系统内有 $q(i-1)$ 名等待服务的顾客,则第 i 个客户到达后可能面临以下一系列系统状态:1)至少一个服务台空闲,第 $i-q(i-1)$ 名客户直接接受服务;2)所有的服务台都处于繁忙中,第 i 个客户到排队队列中等待,并继续推进仿真;3)检查是否达到仿真次数,如果达到则终止仿真程序。

最后,仿真结果的统计计算和输出。仿真结束后输出仿真结果,如平均排队队长、系统中的顾客数、平均等待时间(平均响应时间)和停机时间(总服务时间)等。

2 $M/M/m$ 模型特征描述仿真计算

研究排队系统的目的是为了获得排队系统的性能指标,普遍使用的性能指标有:平均响应时间 $E[W]$、平均停机时间 $E[DT]$、平均排队队长 $E[Q]$、系统中客户数 $E[L]$ 和服务可能延迟的概率 Π 等。

稳态平均延误时间为 $E[W]$,其中 $W(i)$ 为第 i 个实体的延误时间,n 为接受服务的实体数,平均延误时间为实体在队列中的平均等待时间。

$$E[W]=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}W(i),i=1,2,\cdots,n \tag{4}$$

平均停机(滞留)时间为 $E[DT]$,其中 $DT(i)$ 为第 i 个实体通过系统时的滞留时间,它等于实体在队列中的等待时间 $W(i)$ 与该实体接受服务的时间 $S(i)$ 之和。

$$E[DT] = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}DT(i) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(W(i)+S(i)) \tag{5}$$

稳态平均队长为 $E[Q]$,其中 $Q(t)$ 为 t 时刻系统中的实体数,T 为系统运行时间。

$$E[Q] = \frac{1}{T}\lim_{T\to\infty}\int_0^T Q(t)\,\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{n}q(i) \tag{6}$$

系统中稳态平均实体数为 $E[L]$,其中 $L(t)$ 为 t 时刻系统中的实体数,它是在队列中的实体 $Q(t)$ 与正在接受服务的实体 $S(t)$ 之和。

$$E[L] = \frac{1}{T}\lim_{T\to\infty}\int_0^T L(t)\,\mathrm{d}t = \frac{1}{T}\lim_{T\to\infty}\int_0^T (Q(t)+S(t))\,\mathrm{d}t \tag{7}$$

系统中接受服务的平均客户数可以由服务台的忙闲状态进一步简化为

$$E[B] = \frac{1}{n}\sum_{i-1}^{n}\sum_{k=1}^{m}csr\_stat(k,i),\ i=1,2,\cdots,n,\quad k=1,2,\cdots,m \tag{8}$$

系统中客户数为系统中接受服务的平均客户数和平均排队队长之和

$$E[L] = E[Q] + E[B] \tag{9}$$

如果 $q(i)=0$,令 $a(i)=1$,可能延迟的概率(Π)可以由下式得到。

$$\Pi = 1 - \sum_{k=0}^{m-1}p_k = 1 - \frac{\sum_{i=1}^{n}\{a(i)\}}{n},\ i=1,2,\cdots,n \tag{10}$$

3 $M/M/m$ 模型仿真程序设计

主程序运行包括 初始化参数设定和时钟推进,空闲服务台服务流程,繁忙服务台服务流程,仿真运算和结果输出 4 个子程序。

1) 初始化参数设定和时钟推进

初始化设置产生 $M/M/m$ 排队系统仿真过程需要的原始参数设置,客户平均到达率为符合 $\lambda = 1/arr\_interval$ 的泊松分布 ;客户平均服务率为符合 $\mu = 1/ser\_interval$ 的负指数分布 ;服务台个数为 m ;仿真次数为 n 等。仿真时钟按下一个最早发生事件的发生时间来推进,见(2) 式和(3) 式。

2) 空闲服务台服务流程

当客户到达时,如果 $q(i)=0$ 且有空闲服务台,从 l 类服务台中找一名空闲的服务台(k) 为客户提供服务(同时从 l 类中剔除),更新各类服务台的状态。

l 类(空闲) 服务台的忙闲状态和服务完成时间分别为

$$csr\_stat(l,i-q(i-1)) = 0 \tag{11}$$

$$csr\_left(l,i-q(i-1)) = csr\_left(l,i-q(i-1)-1) \tag{12}$$

m 类(繁忙) 服务台的忙闲状态和服务完成时间为

$$csr\_stat(m,i-q(i-1)) = \begin{cases}1, & \text{如果 } csr\_left(m,i-q(i-1)-1) > time(i)\\ 0, & \text{如果 } csr\_left(m,i-q(i-1)-1) \leqslant time(i)\end{cases} \tag{13}$$

$$csr\_left(m,i-q(i-1)) = csr\_left(m,i-q(i-1)-1) \tag{14}$$

k 服务台的忙闲状态、服务完成时间、响应时间和设备的停机时间分别为

$$csr\_stat(k,i-q(i-1)) = 1 \tag{15}$$

$$csr\_left(k,i-q(i-1)) = time(i) + ser\_time(i-q(i-1)) \tag{16}$$

$$W(i-q(i-1)) = csr\_left(k,i-q(i-1)) - arr\_time(i-q(i-1)) - ser\_time(i-q(i-1)) \tag{17}$$

$$DT(i-q(i-1)) = csr\_left(k,i-q(i-1)) - arr\_time(i-q(i-1)) \tag{18}$$

3) 繁忙服务台服务流程

当客户到达时,如果没有空闲服务台,则比较第 i 期客户到达的时间和 m 个服务台完成服务离开的时间((19)式),m 个服务台的服务完成时间都大于第 i 期客户到达的时间,则到达客户到队列中等待

$$arr\_time(i) > \min\{csr\_left(1,i-q(i-1)-1),csr\_left(2,i-q(i-1)-1),\cdots,csr\_left(m,i-q(i-1)-1)\} \tag{19}$$

否则,可以寻找最早完成服务的服务台(k)为最早等待的客户服务。

$$k = \min_{i=1\to m}\{csr\_left(1,i-q(i-1)-1),\cdots,csr\_left(m,i-q(i-1)-1)\}\ \&$$
$$csr\_left(k,i-q(i-1)-1) < arr\_time(i) \tag{20}$$

此时的服务台只有两类,k 服务台是可以提供服务的(1名),m 类服务台为工作繁忙的共有 m-1 名。m 类(繁忙)服务台的忙闲状态和服务完成时间,k 服务台的忙闲状态和服务完成时间同上((11)~(18)式),排队队列减少1名,并且更新时钟((3)式)。

如果此时 $q(i)>0$ 成立,而且 m 个服务台的服务完成时间不全都大于第 i 期客户到达的时间,即

$$arr\_time(i) > csr\_left(k,i-q(i)) = \min_{i=1\to m}\{csr\_left(1,i-q(i)),\cdots,csr\_left(m,i-q(i))\} \tag{21}$$

k 服务台的忙闲状态、服务完成时间、响应时间和设备的停机时间分别为

$$csr\_stat(k,i-q(i)+1) = 1 \tag{22}$$
$$csr\_left(k,i-q(i)+1) = time(i) + ser\_time(i-q(i)+1) \tag{23}$$
$$W(i-q(i)+1) = csr\_left(k,i-q(i)+1) - arr\_time(i-q(i)+1) - ser\_time(i-q(i)+1) \tag{24}$$
$$DT(i-q(i)+1) = csr\_left(k,i-q(i)+1) - arr\_time(i-q(i)+1) \tag{25}$$

m 类服务台为工作繁忙的服务台共有 $m-1$ 名,m 类(繁忙)服务台的忙闲状态和服务完成时间分别为

$$csr\_stat(m,i-q(i)+1) = \begin{cases} 1, \text{如果 } csr\_left(m,i-q(i)) > time(i) \\ 0, \text{如果 } csr\_left(m,i-q(i)) \leqslant time(i) \end{cases} \tag{26}$$
$$csr\_left(m,i-q(i)+1) = csr\_left(m,i-q(i)) \tag{27}$$

服务队列减去1名,并且更新时钟((3)式),继续上述循环直到到达时间小于所有服务台的服务完成时间或者排队队长为零,则跳出循环进行下一次仿真,如(28)式所示。

$$arr\_time(i) < \min\{csr\_left(1,i-q(i)),\cdots,csr\_left(m,i-q(i))\} \text{ 或 } q(i)=0 \tag{28}$$

4)仿真运算和结果输出,根据(1)~(10)式计算并输出结果。

4 $M/M/m$ 排队模型仿真实验和结论

在排队系统 $\lambda = 1/10, \mu = 1/21$ 的情况下,设定仿真次数为20 000次,模拟服务台为3~6的排队系统,各运行程序10次,取得各种状况下排队系统特征描述的平均值,仿真结果见表2。同时罗列出 Lingo 11.0 软件和 Queue 2.0 软件计算出的对应排队系统特征描述的理论值,Queue 2.0 不仅有计算排队模型理论值的功能,而且可以模拟排队系统,它的模拟次数为1 000次,本文将 Queue 2.0 的模拟结果一并列于表中。

表2 $M/M/m$ 模型在 $\lambda = 1/10, \mu = 1/21$ 的仿真结果

	本文结果仿真结果				Queue 2.0 仿真结果		
	平均值	标准差	绝对误差	相对误差	理论值	仿真值	标准差
		$\lambda = 1/10, \mu = 1/21, c = 3$					
$E[Q]$	1.647 5	0.082 1	0.498 7	0.302 7	1.148 8	--	--
$E[L]$	3.747 5	--	--	--	3.248 8[b]	3.251	2.851
$E[W]$	11.504	0.670 1	0.016	0.001 4	11.488[a,b]	11.629	19.782
$E[DT]$	32.489 9	0.734 2	0.001 9	0.000 1	32.488[b]	32.723	28.943

Π	0.492 8	0.009	0.000 5	0.001	0.492 3[a]	--	--
		$\lambda = 1/10, \mu = 1/21, c = 4$					
$E[Q]$	0.430 8	0.035 1	0.210 4	0.488 4	0.220 4	--	--
$E[L]$	2.530 8	--	--	--	2.320 4[b]	2.315	1.771
$E[W]$	2.261 1	0.2	0.057 1	0.025 3	2.204[a, b]	2.128	6.686
$E[DT]$	23.273 6	0.303 3	0.069 6	0.003	23.204[b]	23.188	22.188
Π	0.201 2	0.008 8	0.001 8	0.008 9	0.199 4[a]	--	--
		$\lambda = 1/10, \mu = 1/21, c = 5$					
$E[Q]$	0.117 6	0.010 6	0.066 1	0.562 1	0.051 5	--	--
$E[L]$	2.217 6	--	--	--	2.151 5[b]	2.13	1.56 2
$E[W]$	0.487 3	0.059 3	-0.027 7	0.056 8	0.515[a, b]	0.568	2.801
$E[DT]$	21.490 3	0.152 2	-0.024 7	0.001 1	21.515[b]	21.413	21.123
Π	0.069 1	0.004 1	-0.002	0.028 9	0.071 1[a]	--	--
		$\lambda = 1/10, \mu = 1/21, c = 6$					
$E[Q]$	0.035 7	0.003 7	0.023 6	0.661 1	0.112 1	--	--
$E[L]$	2.135 7	--	--	--	2.112 1[b]	2.131	1.469
$E[W]$	0.128	0.017 1	0.007 4	0.057 8	0.120 6[a, b]	0.102	0.891
$E[DT]$	21.126 6	0.086 5	0.006	0.000 3	21.120 6[b]	21.265	21.17
Π	0.023	0.001 6	-0.048 1	2.091 3	0.071 1[a]	--	--

注:a 表示用 Lingo 11 的 QMMC 模型的求解结果;b 表示采用 Queue 2.0 的计算结果(Queue 2.0 为在线 Java 软件 http://www.win.tue.nl/cow/Q2/)。

本文在经典 $M/M/m$ 模型的理论基础上建立基于马尔科夫状态转移过程的排队模型仿真理论,利用 Matlab 强大的矩阵运算功能建立 $M/M/m$ 排队仿真程序。随后在客户到达率和服务率相同、不同服务台数量情况下计算 $M/M/m$ 排队系统的特征描述,并与 Lingo 11.0 和 Queue 2.0 的结果比较。仿真结果显示几个重要指标 $E[W]$、$E[DT]$ 和 Π 的仿真结果和理论值都极为相近。模型在 20 000 次的仿真结果和 Queue 2.0 的模拟结果也比较接近,但是本文构建的模型仿真结果的标准差更小。仿真结果与理论值以及与 Queue 2.0 的模拟结果的比较显示,本文的仿真模型有一定的优越性,本研究将为后期优先权排队系统的研究提供理论基础。

参考文献:

[1] 孙荣恒,李建平. 排队论基础[M]. 北京:科学出版社,2002.

[2] 陆传赉. 排队论[M]. 北京:北京邮电学院出版社,1989.

[3] 孟玉珂. 排队论基础及应用[M]. 上海:同济大学出版社,1989.

[4] 唐应辉,唐小我. 排队论—基础及应用[M]. 成都:电子科技大学出版社,2000.

[5] 运筹学教材编写组. 运筹学[M]. 北京:清华大学出版社,2005.

[6] 华兴. 排队论与随机服务系统[M]. 上海:上海翻译出版公司,1987.

[7] 徐光辉. 随机服务系统[M]. 第 2 版. 北京:科学出版社,1988.

[8] 贾小娇,方红雨,李晓辉. 基于 OPNET 的 $M/M/m$ 队列仿真[J]. 通信技术,2008,41(12):183-185.

[9] 唐彦,王志坚,吴吟. 基于 Java 的排队系统仿真研究[J]. 计算机工程,2006,32(13):26-29.

[10] 朱军,李晓辉,罗长青. 排队系统仿真及应用[J]. 微机发展,2002(3):47-49.

[11] 林峰,符涛,黄生叶. 大型 $M/P/C/C$ 排队系统仿真研究[J]. 计算机仿真,2007,24(5):131-134.
[12] 张英,郭劲添. 基于 Extend 的排队规则仿真研究[J]. 武汉理工大学学报:信息与管理工程版,2007,29(5):20-24.
[13] 叶宗文. $M/M/C$ 排队模型在理发服务行业中的应用[J]. 重庆师范大学学报:自然科学版,2009,26(2):75-78.
[14] 曹永荣,胡伟. 基于售后现场服务排队近似 $M/G/m$ 模型的服务代表配置[J]. 重庆师范大学学报:自然科学版,2010,27(4):36-40.
[15] 肖田元,张燕云,陈加栋. 系统仿真导论[M]. 北京:清华大学出版社,2000.
[16] 曹永荣,韩传峰. 售后现场服务排队近似 $M/G/m$ 模型仿真[J]. 工业工程与管理,2009,14(5):103-107.
[17] 侯冬倩,高世泽. 服务率可变且窗口能力不等的 $M/M/n$ 排队模型研究[J]. 重庆师范大学学报:自然科学版,2010,27(2):46-48.
[18] 宣慧玉,张发. 复杂系统仿真及应用[M]. 北京:清华大学出版社,2008.
[19] 李焕. 具有可变输入率和不耐烦顾客 $M/M/n$ 的排队模型[J]. 重庆师范大学学报:自然科学版,2011,28(3):49-52.
[20] 张景晓. 整数矩阵的性质及应用[J]. 重庆理工大学学报:自然科学版,2010,24(4):117-119.

Simulation of $M/M/m$ Queuing Model Based on Markov State Transition Process

CAO Yong-rong[1,2], *HAN Rui-xia*[1,2], *HU Wei*[3]

(1. Institute of Arts and Humanities; 2. Antai College of Economics & Management;
3. School Of International and Public Affairs Shanghai Jiao Tong University, Shanghai 200052, China)

Abstract: Markov chain is the main method for the study of queuing systems. This paper integrates the existing theories of $M/M/m$ queuing system and theories of queuing system simulation, and builds simulation program of $M/M/m$ Queuing Model according to the Markov state transition process using Matlab. The simulation process is as follows. First of all, simulation program initializes the parameter settings, such as service time, the interval of customer arrival, the number of server etc. Secondly, promotes the program used time clock which is based on the arrival time of customers and the end time of service. Thirdly, simulates the free servers and busy servers process when a customer arrived, and recodes the corresponding data. Finally, calculate the $M/M/m$ model's characterized descriptions, included in the average down time ($E[DT]$), the average waiting time ($E[W]$), the average number of queuing customer ($E[Q]$), the average number of customers in the queuing system($E[L]$) and delay probability (Π), based on the simulation formula. Sets the iteration times on 20,000 times, at the same service rate and arrival rate, but the servers are from 3-6, and runs the simulation program. The results show that the most important indexes ($E[W]$, $E[DT]$ and Π) are closed to the theoretical value and Queue2.0 simulation results, and it is more comparably accurate than Queue2.0. These studies will provide with theoretical basis for the simulation of priority queuing system.

Key words: Markov state transition process; $M/M/m$ queuing model; simulation

(责任编辑 游中胜)