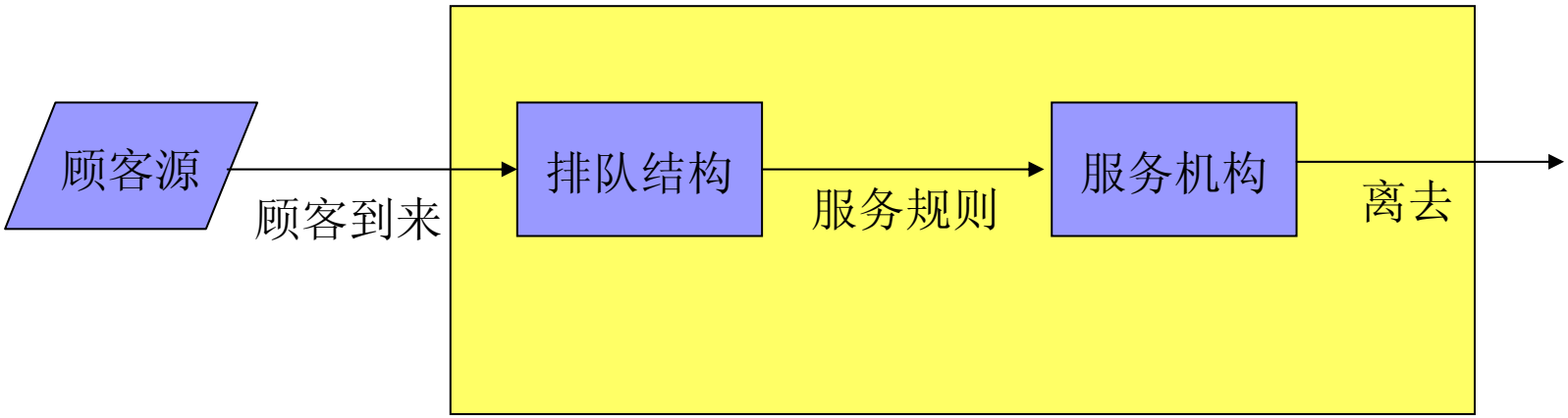
排队论
(Queueing Theory)

南京大学计算机系　黄皓教授

2007年 月 日星期

1. 基本概念

(1) 排队系统的一般表示



(2) 排队系统的组成和特征 — 输入过程

- 顾客的总体：有限或者无限
- 顾客到达方式：单个或成批
- 顾客到达的时间间隔：确定、随机
- 输入过程：平稳的（与时间无关）、非平稳的。

(2) 排队系统的组成和特征 — 排队规则

- 损失制
  - 顾客到达时服务台被占用，顾客即离去。
- 等待制
  - 先到先服务
  - 后到先服务
  - 随机服务
  - 有优先权
- 队列
  - 单列
  - 多列

(2) 排队系统的组成和特征 — 服务机构

- 服务台数量
  - 单个服务台、多个服务台
- 服务台排列
  - 平行、服务台串行
- 服务时间
  - 确定、随机
- 服务时间分布
  - 平稳、非平稳

(3) 排队模型的分类

- 相继顾客到达的时间间隔分布
- 服务时间分布
- 服务台个数

X / Y / Z

- M： 负指数时间分布
- E_k：K阶爱尔朗(Erlang)分布
- GI：一般相互独立的时间间隔分布
- G：一般服务时间分布

M/M/1

M/M/c

(3) 排队模型的分类

X / Y / Z / A / B / C

- X：相继顾客到达的时间间隔分布
- Y：服务时间分布
- Z：服务台个数
- A：系统容量限制
- B：顾客源数目
- C：服务规则（FCFS，LCFS）

(4) 排队问题的求解

- 解决排队问题的目的
  - 排队系统的运行效率
  - 估计服务质量
  - 确定系统参数的最优值
- 基本数量指标
  - 队长：在系统中的顾客数L_s
  - 排队长：在系统中排队等待服务的顾客数L_q

    系统中顾客数 = 在队列中等待服务的顾客数+正在被服务的顾客数
  - 逗留时间：顾客在系统中逗留的时间W_s
  - 等待时间：顾客在系统中排队等待的时间Wq

    逗留时间= 等待时间 + 服务时间
  - 忙期：顾客到达空闲的服务系统直到系统再次空闲的时间长度。

2. 到达间隔分布和服务时间分布

(1) 泊松(Poisson) 分布

- N(t)：在时间区间[0, t)内到达的顾客数量；
- $P_n(t_1, t_2)$：在时间区间$[t_1, t_2)$内有n 个顾客到达的概率。

$$P_n(t_1, t_2)= P\{ N(t_2)-N(t_1)=n \}， \quad t_1> t_2， \quad n\geqslant 0$$

- 当$P_n(t_1, t_2)$符合以下三个条件时，我们说顾客的到达形成泊松流：

（1） 在不相重叠的时间间隔内顾客到达数是相互独立的，无后效性。

（2） 对与充分小的时间间隔 Δt内，在时间区间 $[t, t+ \Delta t)$内有一个顾客到达的概率与t 无关，而大约与 Δt成正比：

$$P_1(t, t+ \Delta t) = \lambda \bullet \Delta t + o(\Delta t)$$

（3） 对与充分小的时间间隔 Δt内，有两个或两个以上的顾客到达的概率极小：

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

(1) 泊松(Poisson) 分布

- 简记： $P_n(0,t) = P_n(t)$
- 由条件(2)、(3),推出在区间$[t, t+\Delta t)$内没有顾客到达的概率

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

区间 / 情况	[0， t)		[t, t+Δt)		[0, t+Δt)	
	个数	概率	个数	概率	个数	概率
(A)	n	$P_n(t)$	0	$1-\lambda \bullet \Delta t + o(\Delta t)$	n	$P_n(t)(1-\lambda\ \Delta t + o(\Delta t))$
(B)	n-1	$P_{n-1}(t)$	1	$\lambda \bullet \Delta t$	n	$P_{n-1}(t)\ \lambda\ \Delta t$
(C)	n-2	$P_{n-2}(t)$	2	$o(\Delta t)$	n	$o(\Delta t)$
(C)	n-3	$P_{n-3}(t)$	3	$o(\Delta t)$	n	$o(\Delta t)$
(C)	…	…	…	…	…	…
(C)	0	$P_0(t)$		$o(\Delta t)$	n	$o(\Delta t)$

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

(1) 泊松(Poisson) 分布

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

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

$$P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t}$$

$$t>0, n=0,1,2,\cdots$$

$$\begin{cases}\dfrac{dP_0(t)}{dt}=-\lambda P_0(t)\\ P_0(0)=1\end{cases}$$

E(N(t))= λ t

Var(N(t)) = λ t

(2) 负指数分布

- 随机变量T的概率密度是

$$f_T(t)=\begin{cases}\lambda e^{-\lambda t}, & t\geq 0\\ 0, & t<0\end{cases}\qquad F_T(t)=\begin{cases}1-e^{-\lambda t}, & t\geq 0\\ 0, & t<0\end{cases}$$

则称T服从负指数分布。

$$E(T)=\int_{-\infty}^{\infty} t\bullet f_T(t)dt=\int_{0}^{\infty} t\bullet e^{-\lambda t}dt=\frac{1}{\lambda}$$

$$Var\ (T)=\int_{-\infty}^{\infty}(t-E(T))^2\bullet f_T(t)dt=\frac{1}{\lambda^2}$$

由条件概率公式容易证明：

$$P(\ T>\ t+s\mid T>s\ \}\ =P(T>t)$$

这个性质称为无记忆性或马尔柯夫性。

(2) 负指数分布

- 如果顾客的输入过程是Poisson流时，则顾客相继到达时间间隔T 就服从负指数分布。

- 在[0，t)时间内到达至少一个顾客的概率是

$$1-P_0(t) = 1- e^{-\lambda t}, \quad t > 0$$

也就是 $F_T(t) = P(T \leqslant t) = 1- e^{-\lambda t}, \quad t > 0$

- 也就是说顾客到达的时间间隔具有Markov性，用M表示。

(3) Erlang 分布

- 设v_1, v_2, v_k是k个相互独立的随机变量，服从相同参数 $k\mu$ 的负指数分布，那么：

$$T = v_1 + v_2 + \cdots + v_k$$

的概率密度是

$$b_k(t) = \frac{\mu k(\mu k t)^{k-1}}{(k-1)!} e^{-\mu k t}, t > 0$$

3. 单服务台负指数分布排队系统

M / M / 1 （M / M / 1 /∞ / ∞）

情况	在时刻t的顾客数	在区间 [t, t+ Δt)		在时刻t+ Δt的顾客数
		到达	离去	
A	n	0	0	n
B	n+1	0	1	n
C	n-1	1	0	n
D	n	1	1	n

- 在区间 [t, t+ Δt)内有一个顾客到达的概率为 λ • Δt + o(Δt)，没有顾客到达的概率 1- λ • Δt + o(Δt)
- 当有一个顾客在接受服务时，1个顾客被服务完离去的概率是 μ • Δt + o(Δt)，没有离去的概率是1- μ • Δt + o(Δt)
- 多于一个顾客到达和离去的概率为： o(Δt)。

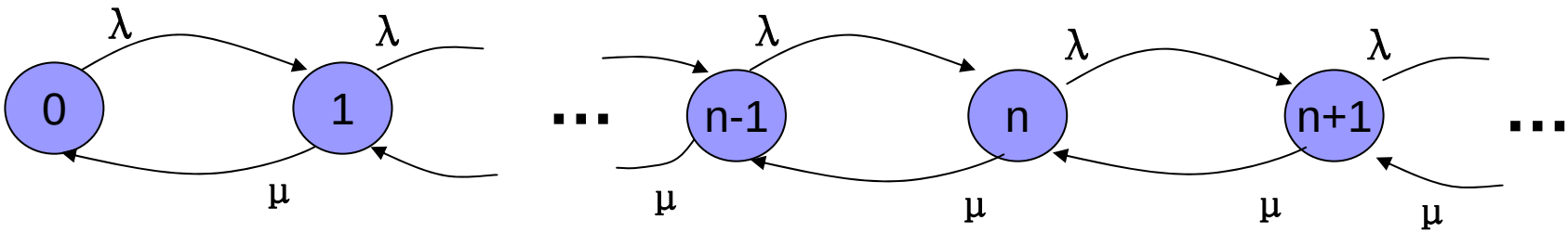
M / M / 1 （M / M / 1 /∞ / ∞）

- A：$P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t)$
- B：$P_{n+1}(t)(1-\lambda\bullet\Delta t)\bullet\mu\Delta t$
- C：$P_{n-1}(t)\ \lambda\Delta t(1-\mu\Delta t)$
- D：$P_n(t)\bullet\lambda\Delta t\bullet\mu\Delta t$

- **$P_n(t+\Delta t)=P_n(t)(1-\lambda\Delta t-\mu\Delta t)+P_{n+1}(t)\ \mu\Delta t+P_{n-1}(t)\bullet\lambda\Delta t$**
- **$d\,P_n(t)/dt=\lambda P_{n-1}(t)+\mu P_{n+1}(t)-P_n(t)(\lambda+\mu)$**
- **$d\,P_0(t)/dt=\mu P_1(t)-\lambda P_0(t)$**

M / M / 1 （M / M / 1 /∞ / ∞）

- 在稳态的情况下：$P_n(t)$与t无关，可以写成P_n，导数为0。
  - $0 = \lambda P_{n-1} + \mu P_{n+1} - P_n(\lambda + \mu), \quad n \geqslant 1$
  - $0 = \lambda P_0 + \mu P_1$



- $P_1=(\lambda / \mu)P_0$
- $P_n= =(\lambda / \mu)^n P_0$

M / M / 1 （M / M / 1 /∞ / ∞）

令 ρ = λ / μ　　ρ =(1/ μ)/(1/ λ)：服务时间与到达时间之比

$P_1=(\lambda/\mu)P_0=\rho P_0$

$P_n=(\lambda/\mu)^nP_0=\rho^nP_0$

$$1=\sum_{i=0}^{\infty}P_i=\sum_{i=0}^{\infty}\rho^iP_0=P_0\frac{1}{1-\rho}$$

$P_0=1-\rho$

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

M / M / 1 （M / M / 1 /∞ / ∞）系统指标

（1） 系统中的平均顾客数

$$Ls = \sum_{i=1}^{\infty} nP_n = \sum_{i=1}^{\infty} n(1-\rho)\rho^n$$

$$= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

（2） 在队列中等待的平均顾客数

$$L_q = \rho \frac{\lambda}{\mu - \lambda}$$

（3）$W_s = 1/(\mu - \lambda)$

（4） $W_q = \rho /(\mu - \lambda)$

M / M / 1 （M / M / 1 /∞ /∞）系统指标

$$L_s = \frac{\lambda}{\mu - \lambda}$$

$$L_q = \rho \frac{\lambda}{\mu - \lambda}$$

$$W_s = \frac{1}{\mu - \lambda}$$

$$W_q = \frac{\rho}{\mu - \lambda}$$

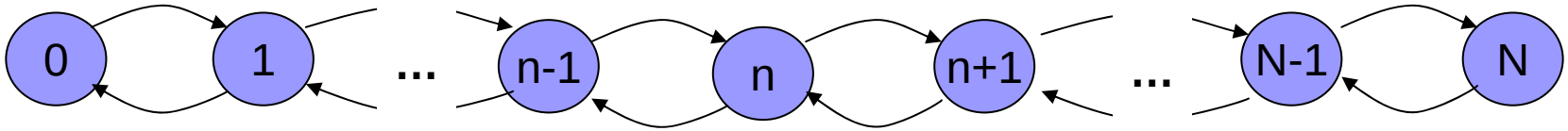
$$L_s = L_q + \frac{1}{\mu}$$

$$W_s = W_q + \frac{1}{\mu}$$

$$L_s = \lambda W_s$$

$$L_q = \lambda W_q$$

系统有容量限制的情况 M / M / 1 / N / ∞



- $\lambda P_0 = \mu P_1$
- $\lambda P_{n-1} + \mu P_{n+1} = (\lambda + \mu) P_n, \quad n \leqslant N-1$
- $\lambda P_{N-1} = \mu P_N$

$$P_0 = \frac{1-\rho}{1-\rho^{N+1}} \qquad \rho \neq 1$$

$$P_n = \frac{1-\rho}{1-\rho^{N+1}} \rho^n \qquad n \leq N$$

系统有容量限制的情况 M / M / 1 / N / ∞

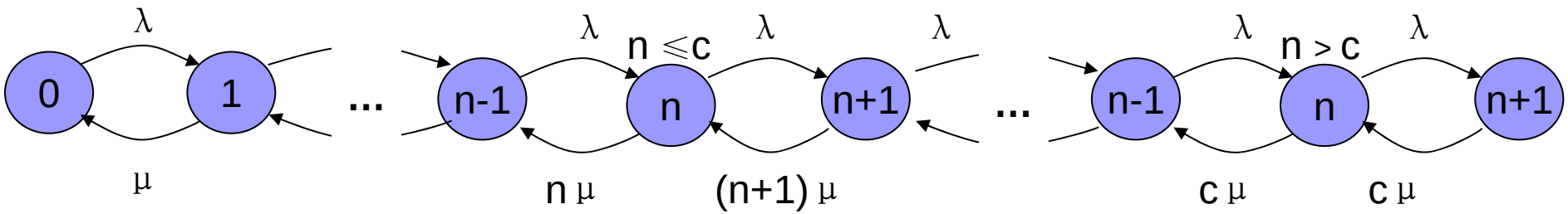
$$L_s = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}} \qquad \rho = \frac{\lambda}{\mu} < 1$$

$$L_q = L_s - (1 - P_0)$$

$$W_s = \frac{L_s}{\mu(1-P_0)}$$

$$W_q = W_s - \frac{1}{\mu}$$

多服务台复指数分布 M / M / c



- $\lambda P_0 = \mu P_1$
- $\lambda P_{n-1} + (n+1)\mu P_{n+1} = (\lambda + n\mu) P_n, \quad 1 \leqslant n \leqslant c$
- $\lambda P_{N-1} = \mu P_N \qquad n>c$

$$\sum_{i=0}^{\infty} P_i = 1 \qquad \rho = \frac{\lambda}{c\mu} < 1$$

多服务台复指数分布 M / M / c

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

$$P_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & (n \le c) \\ \dfrac{1}{c!\,c^{n-c}}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & (n > c) \end{cases}$$

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

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

$$W_q = \frac{L_q}{\lambda}, \qquad W_s = \frac{L_s}{\lambda}$$

多服务台复指数分布 M / M / c

- 到达率： λ =0.9; 平均服务率 μ =0. 4； c=3
- **L_q=1.70； L_s=3.95**
- **W_q=1.89(分钟)； W_s=4.39(分钟)**

- 到达率： λ =0.9/3=0.3; 平均服务率 μ =0. 4； c=3
- **L_q=2.25； L_s=9**
- **W_q=7.5(分钟)； W_s=10(分钟)**

- 一队比三队有明显的优势。

一般服务时间 M / G / 1

$$L_s = \rho + \frac{\rho^2 + \lambda_2 Var(T)}{2(1-\rho)} \qquad \rho = \lambda \cdot E(T)$$

$$L_s = L_q + L_{se} = L_q + \rho$$

$$W_s = L_s / \lambda$$

$$W_q = L_q / \lambda$$

定长服务时间 M / D / 1

- T = 1/ μ， Var(T) = 0

$$L_s = \rho + \frac{\rho^2}{2(1-\rho)} \qquad \rho = \lambda / \mu$$

$$L_q = \frac{\rho^2}{2(1-\rho)}$$

$$W_s = L_s / \lambda$$

$$W_q = L_q / \lambda$$

M / D / 1的例

- 某实验室有一台自动检验机器性能的仪器，要求检验及其按Poisson分布到达，平均每小时到达4台要求检验的机器，检验每台机器所需时间为6分钟。求：
  - 在检验室内要求检验的机器的平均数量L_s；
  - 等候检验的机器的平均数量L_q；
  - 每台机器在实验室中耗费的时间W_s；
  - 每台机器平均等待检验的时间W_q
- λ =4， E(T) = 0.1, ρ = λ E(T) =0.4，
- L_s=0.4+ 0.4^2 / 2 · (1−0. 4) =0. 533
- L_q= Ls − ρ = 0.533-0.4 = 0.133
- W_s= Ls/ λ = 0.533 /4 = 0.133 (=8分钟)
- W_q= W_s/ λ = 0.133/4 = 0.033 (=4分钟)