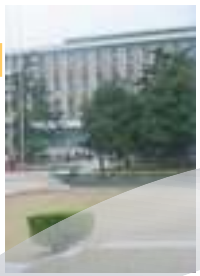
Computer Science and Technology



运 筹 学

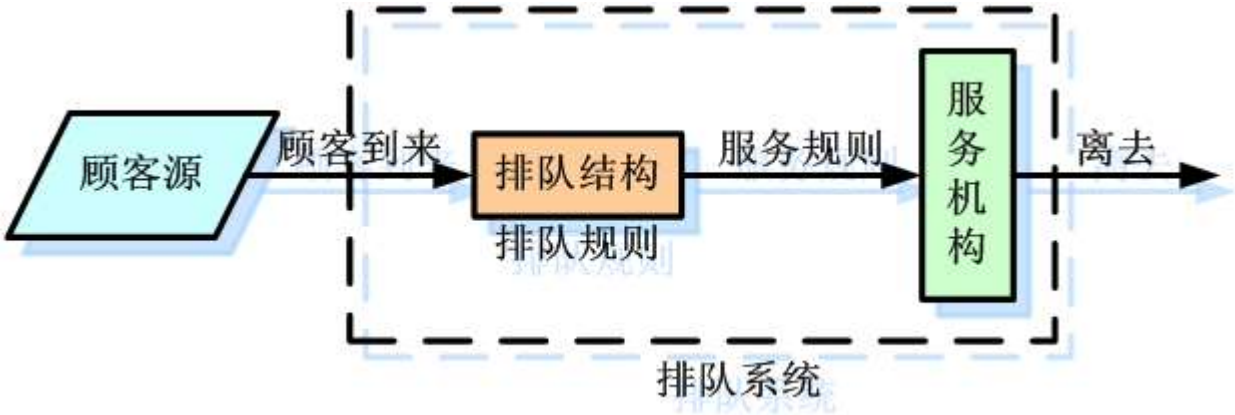
排队论

Queueing Theory

- 排队论（Queueing Theory）也称随机服务系统理论，它研究的内容有：
  - 性态问题：即研究各种排队系统的概率规律性，主要是研究队长分布、等待时间分布和忙期分布等，包括瞬态和稳态两种情形；
  - 最优化问题：又分静态最优和动态最优，前者指最优设计，后者指现有排队系统的最优运营；
  - 排队系统的统计推断：即判断一个给定的排队系统符合于哪种模型，以便根据排队理论进行分析研究。
- 排队过程的一般表示：
  - 有形排队现象：进餐馆就餐，到图书馆借书，车站等车，去医院看病，售票处售票，到工具房领物品等现象。

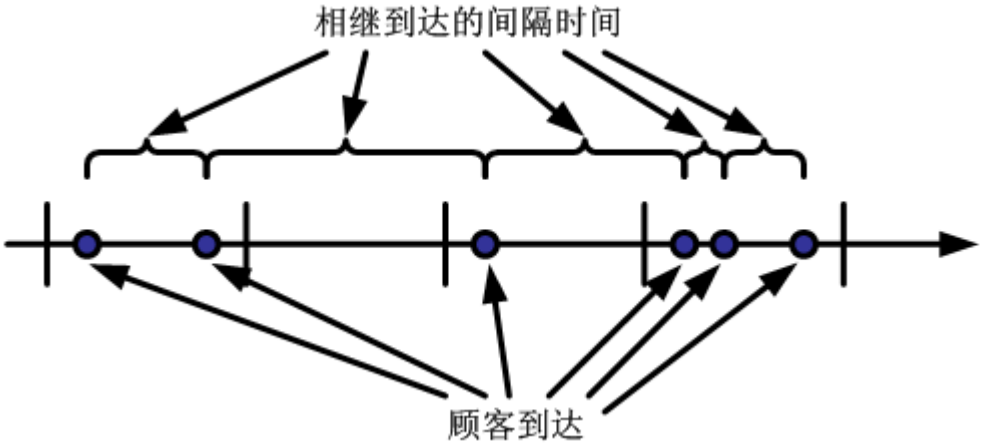
12.1 基本概念(cont.)

- 无形排队现象：如几个旅客同时打电话订车票；如果有一人正在通话，其他人只得在各自的电话机前等待，他们分散在不同的地方，形成一个无形的队列在等待通电话。
- 排队的不一定是人，也可以是物。如生产线上的原材料，半成品等待加工；因故障而停止运行的机器设备在等待修理；码头上的船只等待装货或卸货；要下降的飞机因跑道不空而在空中盘旋等。当然，进行服务的也不一定是人，可以是跑道，自动售货机，公共汽车等。



12.1 基本概念(cont.)

- 排队系统的组成和特征
  - 输入过程——即指顾客到达排队系统，可能有下列不同情况，这些情况并不彼此排斥：
    - 顾客的总体（称为顾客源）的组成可能是有限的，也可能是无限的。
    - 顾客到来的方式可能是一个一个的，也可能是成批的。
    - 顾客相继到达的间隔时间可以是确定型的，也可以是随机型的。



随机型中，单位时间内顾客到达数或相继到达间隔时间的概率分布

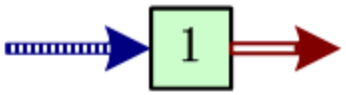
- 顾客的到达可以是相互独立的，即以前的到达情况对以后顾客的到来没有影响。或者是有关联的。
- 输入过程可以是平稳的，或称对时间是齐次的，指描述相继到达的间隔时间分布和所含参数（期望值、方差等）都是与时间无关的。否则就是非平稳的，非平稳情况的数学处理很难。

– 排队规则

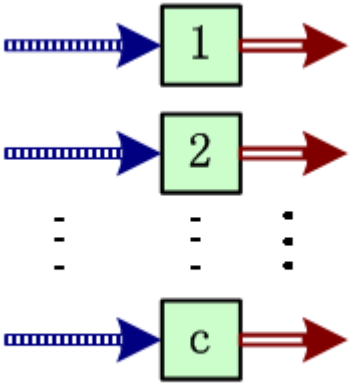
- 顾客到达时，如果所有服务台都被占用，在这种情况下顾客可以随即离去，也可以排队等候。随即离去的称为即时制或损失制，排队等候的称为等待制。等待制可以采用下列规则：先到先服务、后到先服务、随机服务、有优先权的服务。
- 从占有的空间看，队列可以排在具体的处所，也可以是抽象的。由于空间的限制或其它原因，有的系统要规定容量的最大限；有的没有限制。
- 从队列的数目看，可以是单列，也可以是多列。在多列情况下，各列间的顾客有的可以相互转移，有的不能；有的排队顾客因等候时间过长而中途退出，有的不能退出，必须坚持到被服务为止。

– 服务机构

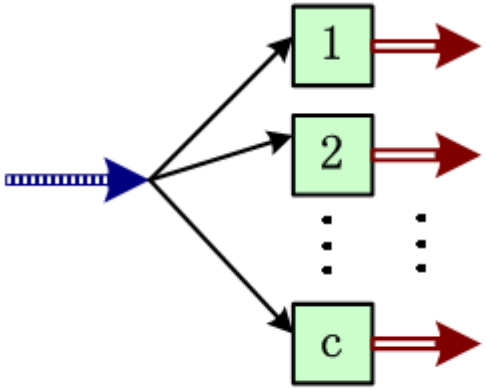
- 服务机构可以没有服务员，也可以有一个或多个服务员。
- 在有多个服务台的情况下，它们可以是平行排列（并列的），可以是前后排列（串列的），也可以是混合的。



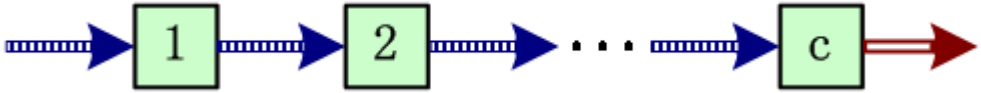
(a) 单队－单服务台



(b) 多队－多服务台（并列）

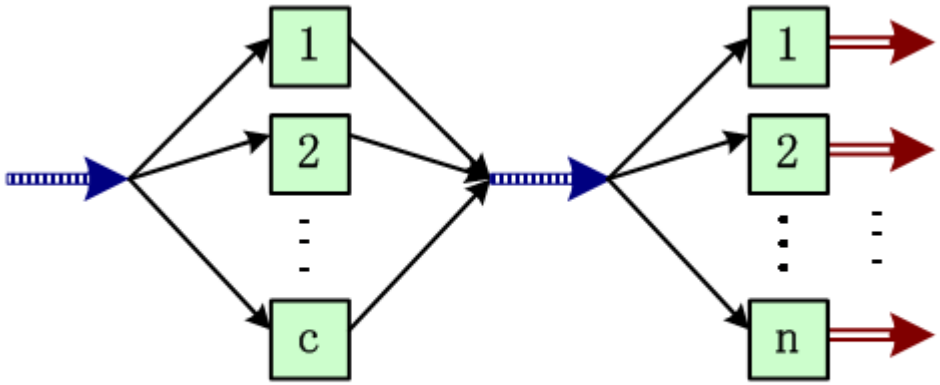


(c) 单队－多服务台（并列）



(d) 多服务台（串列）

12.1 基本概念(cont.)



(e) 多服务台（混合）

– 服务规则

- 服务方式可以对单个顾客进行，也可以对成批顾客进行。
- 和输入过程一样，服务时间也分确定型和随机型，对于随机型的服务时间，需要知道其概率分布。
- 和输入过程一样，服务时间的分布可以是平稳的，或非平稳的。

12.1 基本概念(cont.)

– 本书排队论研究中的几点假设：

- 假设顾客是单个到来的，即不研究成批到来情形；
- 假设顾客的到达是相对独立的，即以前的到达情况对以后顾客的到来没有影响；
- 假设输入过程和服务时间的分布是平稳的，即间隔时间分布和所含参数（如期望值、方差等）都是与时间无关的；
- 假设并列的各列间顾客不能相互转移、不能中途退出；
- 假设服务方式只对单个顾客进行，即不研究对成批顾客服务的情形；
- 假设排队规则采用先到先服务规则；
- 假设输入过程和服务时间至少有一个是随机型的情况，因为若二者都是确定型的则问题太过于简单。

12.1 基本概念(cont.)

- 排队模型的分类
  - 1953年，D.G.Kendall根据上面特征中最主要的、影响最大的三个，提出了肯德尔记号，该记号只针对并列服务台的情形，表示为“X/Y/Z”，分别表示：
    - 相继顾客到达间隔时间分布；
    - 服务时间的分布；
    - 服务台个数。
  - 1966年和1968年，A.M.李相继将肯德尔符号扩充为以下形式：

    $X/Y/Z/A/B/C$

    X处填写顾客相继到达间隔时间的分布；

    Y处填写表示服务时间的分布；

    Z处填写并列的服务台数目；

    A处填写系统容量限制N；

    B处填写顾客源数目m；

    C处填写服务规则，如先到先服务$FCFS$，后到先服务$LCFS$。

12.1 基本概念(cont.)

- 表示相继到达间隔时间和服务时间的各种分布的符号是:
  - **M**——负指数分布（**M**是**Markov**的字头，因为负指数具有无记忆性，即**Markov**性）；
  - **D**——确定型（**Deterministic**）；
  - E_k——**k**阶爱尔朗（**Erlang**）分布；
  - **GI**——一般相互独立（**General Independent**）的时间间隔的分布；
  - **G**——一般（**General**）服务时间的分布。
- 约定，如略去后三项，即指***X/Y/Z/∞/∞/FCFS***的情形。

- **排队问题的求解**
  - 在求解排队问题时需要研究它属于哪个模型，其中顾客到达的间隔时间分布和服务时间分布需要实测的数据来确定。

12.1 基本概念(cont.)

- 求解排队问题的目的是研究排队系统运行效率、估计服务质量、确定系统参数的最优值以决定系统结构是否合理、研究设计改进措施。
- 解排队问题首先要求出用以判断系统运行优劣的基本数量指标的概率分布或特征数：
  - 队长：指在系统中的顾客数，它的期望值记作L_s；
  - 排队长（队列长）：指在系统中排队等待服务的顾客数，它的期望值记作L_q；

$$\begin{bmatrix}\text{系统中}\\ \text{顾客数}\end{bmatrix}=\begin{bmatrix}\text{在队列中等待}\\ \text{服务的顾客数}\end{bmatrix}+\begin{bmatrix}\text{正被服务}\\ \text{的顾客数}\end{bmatrix}$$

一般情形，L_s（或L_q）越大，说明服务效率越低。

  - 逗留时间：指一个顾客在系统中的停留时间，它的期望值记作W_s；
  - 等待时间：指一个顾客在系统中排队等待的时间，它的期望值记作W_q；

$$[\text{逗留时间}]=[\text{等待时间}]+[\text{服务时间}]$$

12.1 基本概念(cont.)

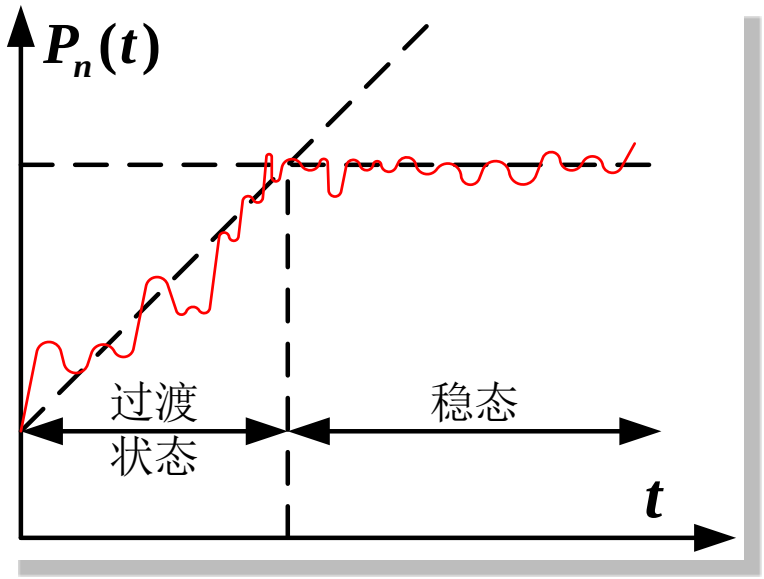
- 忙期（**Busy Period**）：指从顾客到达空闲服务机构起，到服务机构再次为空闲为止这段时间长度，即服务机构连续繁忙的时间长度。该指标关系到服务员的工作强度。忙期和一个忙期中平均完成服务顾客数都是衡量服务机构效率的指标。
- 在即时制或排队有限制的情形，由于顾客被拒绝而使企业受到损失的损失率和服务强度都是衡量服务机构效率的指标。

– 计算上述指标的基础是表达系统状态的概率。系统的状态指系统中顾客数，如果系统中有**n**个顾客就说系统的状态是**n**，它的可能值是：
  - 队长没有限制：**n=0,1,2,……**
  - 队长有限制，最大数为**N**时，**n=0,1,2,……,N**;
  - 即时制，服务台个数是**c**时，**n=0,1,2,……,c**。
– 在上述第三种情况中，状态**n**又表示正在工作（繁忙）的服务台数。这些状态一般是随时刻**t**而变化，在时刻**t**、系统状态为**n**的概率用$P_n(t)$表示。

12.1 基本概念(cont.)

– 求状态概率$P_n(t)$的方法：

- 建立含$P_n(t)$的关系式（如图），因t是连续变量，而n只取非负整数，所以建立的$P_n(t)$关系式一般是微分差分方程（关于t的微分方程，关于n的差分方程）。
- 方程的解称为瞬态（或称过渡状态）（Transient State）解。
- 一般情况下，用极限 $\lim_{t\to\infty} P_n(t) = P_n$ 称为稳态（Steady State），或称统计平衡状态（Statistical Equilibrium State）的解。



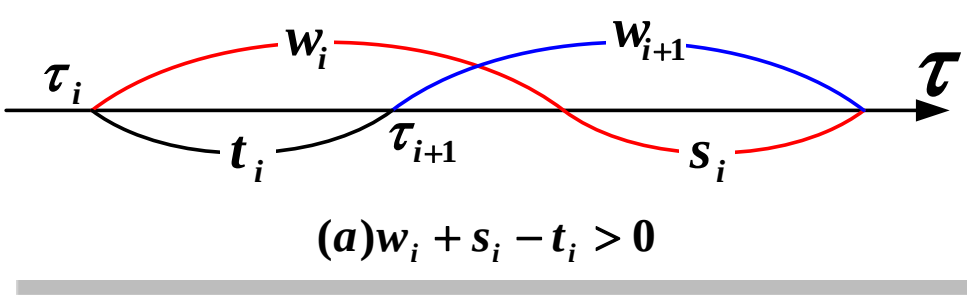
稳态的物理含义：

当系统运行了无限长的时间之后，初始（t=0）出发状态的概率分布（$P_n(t)$,n≥0）的影响将消失，而且系统的状态概率分布不会再随时间变化。

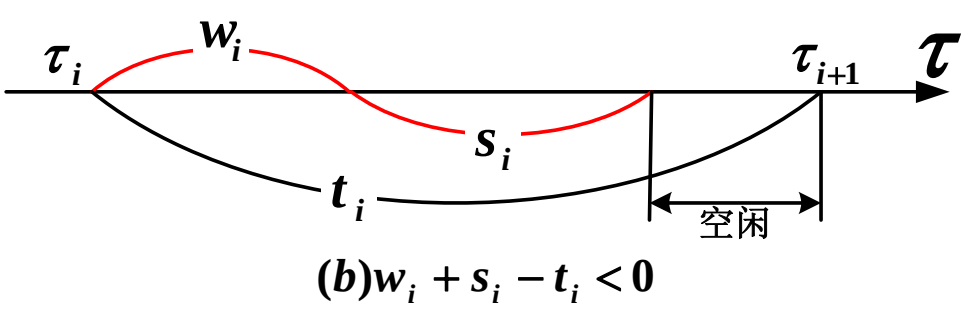
12.2 到达间隔和服务时间的分布

- 经验分布
  - 原始资料的整理
    - 原始资料记录各顾客到达的时刻和对各顾客的服务时间。

以τ_i表示第i号顾客到达的时间，以s_i表示对它的服务时间，可算出相继到达的间隔时间$t_i(t_i=\tau_{i+1}-\tau_i)$和排队等待时间$w_i$。



$(a)w_i+s_i-t_i>0$



$(b)w_i+s_i-t_i<0$

如图所示：

间隔

$$t_i=\tau_{i+1}-\tau_i$$

等待时间

$$w_{i+1}=\begin{cases}w_i+s_i-t_i, & \text{当}w_i+s_i-t_i>0\\ 0, & \text{当}w_i+s_i-t_i<0\end{cases}$$

12.2 到达间隔和服务时间的分布(cont.)

- 顾客到达或离开的泊松分布
  - 设$N(t)$表示在时间区间$(0,t)$内到达的顾客数$(t>0)$，令$P_n(t_1,t_2)$表示在时间区间$[t_1,t_2)(t_2>t_1)$内有$n(\geq 0)$个顾客到达（随机事件）的概率，即
    $P_n(t_1,t_2)=P\{N(t_2)-N(t_1)=n\}(t_2>t_1,n\geq 0)$
    当$P_n(t_1,t_2)$合于下列三个条件时，就称顾客的到达形成泊松流。

(1)在不相重叠的时间区间内的顾客到达数是相互独立的，称该性质为无后效性；

(2)对充分小的Δt，在时间区间$(t,t+\Delta t)$内有1个顾客到达的概率与t无关，而约与区间长Δt成正比，即$P_1(t,\ t+\Delta t)=\lambda\Delta t+o(\Delta t)$，其中$o(\Delta t)$，当$\Delta t\to 0$时，是关于$\Delta t$的高阶无穷小。$\lambda>0$是常数，表示单位时间有一个顾客到达的概率，称为概率强度。

(3)对于充分小的Δt，在时间区间$(t,t+\Delta t)$内有2个或2个以上顾客到达的概率极小，以致可以忽略，即$\sum_{n=2}^{\infty}P_n(t,\ t+\Delta t)=o(\Delta t)$

12.2 到达间隔和服务时间的分布(cont.)

– 顾客到达数n的概率分布

- 由条件(2)，取时间由0算起，简记$P_n(0,t)=P_n(t)$
- 由条件(2)(3)，在$[t,t+\Delta t)$区间内没有顾客到达的概率
  $P_0(t,t+\Delta t)=1-\lambda t+o(\Delta t)$
- 对于区间$[0,t+\Delta t)$可以分成两个不重叠的区间$[0,t)$和$[t,t+\Delta t)$，到达总数为n，分别出现在上面两个区间上，不外下表中的三种情况，见下表。

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

12.2 到达间隔的分布和服务时间的分布 (cont.)

- 在$[0,t+\Delta t)$内到达n个顾客应该是表中三种情况之一，所以概率$P_n(t+\Delta t)$是表中三个概率之和：

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

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

令$\Delta t\to 0$，得下列方程，并注意初始条件，有：

$$\begin{cases}\dfrac{dP_n(t)}{dt}=-\lambda P_n(t)+\lambda P_{n-1}(t), & n\geq 1 \quad (12.5)\\ P_n(0)=0;\end{cases}$$

当$n=0$时，没有$(B)(C)$两种情况，所以得：

$$\begin{cases}\dfrac{dP_0(t)}{dt}=-\lambda P_0(t) & (12.6)\\ P_0(0)=1\end{cases}$$

解(12.5)和(12.6)即得：

$$P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t},\quad t>0 \quad (12.7)$$
$$n=0,1,2,\cdots$$

12.2 到达间隔和服务时间的分布(cont.)

- $P_n(t)$表示长为t的时间区间内到达n个顾客的概率，由(12.7)式，随机变量$\{N(t)=N(s+t)-N(s)\}$服从泊松分布，其数学期望和方差是：

  $E[N(t)]=\lambda t;\ \ Var[N(t)]=\lambda t$

- 期望值和方差相等是泊松分布的一个重要特征。

- 到达时间间隔的负指数分布
  - 对于负指数分布：

    $$f(t)=\lambda e^{-\lambda t}, t>0$$

    $$E(t)=\frac{1}{\lambda}$$

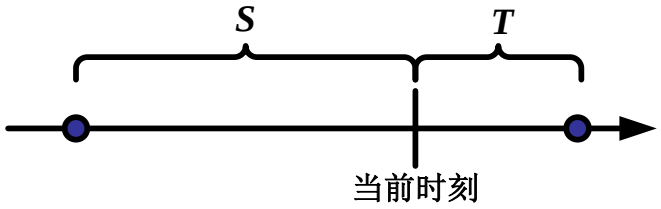
    $$P\{t\le T\}=\int_0^T \lambda e^{-\lambda t}dt=1-e^{-\lambda T}=1-P\{t>T\}$$

  - E{t}的定义表明，λ是每单位时间产生事件（到达或离去）的速率，P{t≤T}是0到T这段时间的累积概率。

12.2 到达间隔和服务时间的分布(cont.)

– 负指数分布的无记忆性：

- 该性质称为无记忆性或马尔柯夫性。若t表示排队系统中顾客到达的间隔时间，那么这个性质说明一个顾客到来所需的时间与上一个顾客到来以来的时间区间S无关，即该情形下顾客到达是纯随机的。



即 $P\{t > T + S \mid t > S\} = P\{t > T\}$ (12.11)

证明：

根据负指数分布定义有

$$P\{t > Y\} = 1 - P\{t \le Y\} = e^{-\lambda Y}$$

$$\therefore P\{t > T + S \mid t > S\} = \frac{P\{t > T + S, t > S\}}{P\{t > S\}} = \frac{P\{t > T + S\}}{P\{t > S\}}$$

$$= \frac{e^{-\lambda(T+S)}}{e^{-\lambda S}} = e^{-\lambda T} = 1 - (1 - e^{-\lambda T}) = 1 - P\{t \le T\} = P\{t > T\}$$

– 负指数分布与泊松分布的关系

- 定义 $p_0(t)$：t时间期间内没有到达的概率

已知到达时间是负指数分布，且每单位时间顾客到达率为λ，则：

$$p_0(t) = P\{\text{到达间隔时间} > t\}$$

$$= 1 - P(\text{到达间隔时间} \le T) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}$$

对一充分小的时间区间h>0，有：

$$p_0(h) = e^{-\lambda h} = 1 - \lambda h + \frac{(\lambda h)^2}{2!} - \cdots = 1 - \lambda h + o(h)$$

负指数分布基于假设：在充分小的h>0期间，最多有一个事件发生，即最多一个到达或一个离开。因此当h→0时，有：

$$p_1(h) = 1 - p_0(h) \approx \lambda h$$

这一结果表示，h期间一次到达的概率直接与h成正比，以到达率λ为比例常数。

- 定义 $p_n(t)=t$时间期间内有n个到达的概率

  对充分小的$h>0$，有

$$p_n(t+h)\approx\begin{cases}p_n(t)p_0(h)+p_{n-1}(t)p_1(h)=p_n(t)(1-\lambda h)+p_{n-1}(t)\lambda h, & n>0\\ p_0(t)p_0(h)=p_0(t)(1-\lambda h), & n=0\end{cases}$$

- 移项并取h→0，得到：

  对充分小的$h>0$，有

$$p_n'(t)=\lim_{h\to 0}\frac{p_n(t+h)-p_n(t)}{h}=-\lambda p_n(t)+\lambda p_{n-1}(t),\quad n>0$$

$$p_0'(t)=\lim_{h\to 0}\frac{p_0(t+h)-p_0(t)}{h}=-\lambda p_0(t),\quad n=0$$

- 求解得到：

$$p_n(t)=\frac{(\lambda t)^n e^{-\lambda t}}{n!},n=0,1,2,\cdots$$

- 上式正是t期间平均有λt个到达的泊松分布，即到达间隔时间服从平均值为1/ λ的负指数分布，则指定期间t内的到达数服从平均值λt的泊松分布。
- 上述模型为只有到达的“纯生模型”，反之，对于在0时刻初始状态为N的派对系统，只有离开的情况，称为“纯灭模型”。设离开的发生率为每单位时间μ个顾客，则可建立t时间单位后系统只剩下n个顾客的概率的差分-微分方程：

对充分小的$h>0$，有

$$p_n(t+h)\approx\begin{cases}p_n(t)(1-\mu h) & n=N\\ p_n(t)(1-\mu h)+p_{n+1}(t)\mu h, & 0<n<N\\ p_0(t)(1)+p_1(t)\mu h, & n=0\end{cases}$$

$h\to 0$时,有

$$\begin{cases}p'_N(t)=-\mu p_N(t)\\ p'_n(t)=-\mu p_n(t)+\mu p_{n+1}(t), \quad 0<n<N\\ p'_0(t)=\mu p_1(t)\end{cases}$$

- 解得截尾泊松分布：

$$p_n(t)=\frac{(\mu t)^{N-n}e^{-\mu t}}{(N-n)!},n=1,2,\cdots,N$$

$$p_0(t)=1-\sum_{n=1}^{N}p_n(t)$$

- 对一顾客的服务时间也就是在忙期相继离开系统的两顾客的间隔时间，即负指数分布。其中1/μ也就是平均服务时间。

– 负指数分布和泊松分布在Kendall记号中都用M表示

12.3 单服务台负指数分布排队系统的分析

- 标准的M/M/1模型（M/M/1/∞/∞/FCFS）
  - 标准的M/M/1模型是指适合下列条件的排队系统：
    - 1.输入过程——顾客源是无限的，顾客单个到来，相互独立，一定时间的到达数服从泊松分布，到达过程已是平稳的；
    - 2.排队规则——单队，且对队长没有限制，先到先服务；
    - 3.服务机构——单服务台，各顾客的服务时间是相互独立的，服从相同的负指数分步；
    - 4.假设到达间隔时间和服务时间是相互独立的。
  - 分析M/M/1模型
    - 首先要求出系统在任意时刻t的状态为n（系统中有n个顾客）的概率$P_n(t)$，它决定了系统运行的特征。
    - 已知到达规律服从参数为 λ 的泊松过程，服务时间服从参数为 μ 的负指数分步，所以在[t,t+Δt)时间区间内分为以下几种：

12.3 单服务台负指数分布排队系统的分析(cont.)

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

没有顾客到达的概率就是$1-\lambda\Delta t+o(\Delta t)$。

(2)当有顾客在接受服务时,1个顾客被服务完了（离去）的概率是$\mu\Delta t+o(\Delta t)$,

没有离去的概率就是$1-\mu\Delta t+o(\Delta t)$。

(3)多于一个顾客的到达或离去的概率是$o(\Delta t)$, 是可以忽略的。

- 在时刻t+Δt，系统中有n个顾客（n>0）存在下列四种情况（到达或离去是2个以上或的没有列入）：

情况	在时刻t顾客数	在区间(t,t+Δt)		在时刻t+Δt顾客数
		到达	离去	
(A)	n	×	×	n
(B)	n+1	×	○	n
(C)	n-1	○	×	n
(D)	n	○	○	n

12.3 单服务台负指数分布排队系统的分析(cont.)

- 四种情况的概率分别是：（略去o(Δt)）

(A): $P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t)$

(B): $P_{n+1}(t)(1-\lambda\Delta t)\cdot\mu\Delta t$

(C): $P_{n-1}(t)\cdot\lambda\Delta t(1-\mu\Delta t)$

(D): $P_n(t)\cdot\lambda\Delta t\cdot\mu\Delta t$

- 由于这四种情况是互不相容的，所以：

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

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

令$\Delta t\to 0$，得到关于$P_n(t)$的微分差分方程：

$$\frac{dP_n(t)}{dt}=\lambda P_{n-1}(t)+\mu P_{n+1}(t)-(\lambda+\mu)P_n(t)\quad n=1,2,\cdots\quad(12.15)$$

当$n=0$时，表中只有$(A)(B)$两种情况，即：$P_0(t+\Delta t)=P_0(t)(1-\lambda\Delta t)+P_1(t)(1-\lambda\Delta t)\mu\Delta t$

得到：

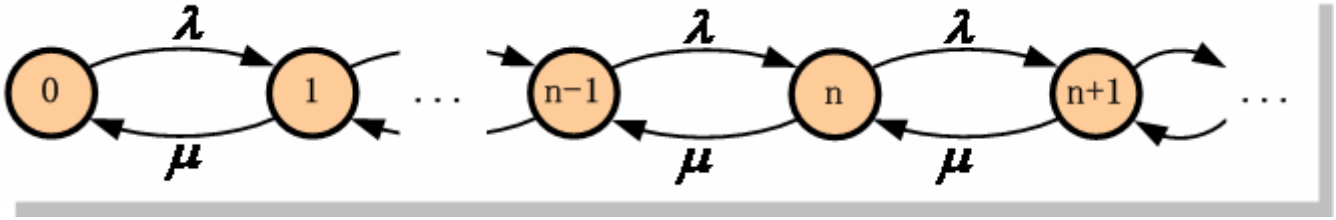
$$\frac{dP_0(t)}{dt}=-\lambda P_0(t)+\mu P_1(t)\quad(12.16)$$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 这样系统状态（n）随时间变换的过程是称为生灭过程的一个特殊情况。
- 因前面假设只针对稳态情况求解，$P_n(t)$与t无关，可写成P_n，它的导数为0。由上面的两个方程（12.15）（12.16）可得：

$$\begin{cases} -\lambda P_0 + \mu P_1 = 0 & (12.17) \\ \lambda P_{n-1} + \mu P_{n+1} - (\lambda + \mu) P_n = 0 \quad n \geq 1 & (12.18) \end{cases}$$

- 这是关于P_n的差分方程，表明了各状态间的转移关系，用下图表示：



- 由上图可见，状态0转移到状态1的转移率为λP_0，状态1转移到状态0的转移率为μP_1。对状态0必须满足以下平衡方程：$\lambda P_0 = \mu P_1$
- 同样，对于任何n≥1的状态，可得到（12.18）的平衡方程。求解（12.17）得：

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

- 将上式代入(12.18)，令n=1，则：

$$\mu P_2 = (\lambda + \mu)(\lambda / \mu) P_0 - \lambda P_0 \Rightarrow P_2 = (\lambda / \mu)^2 P_0$$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 同理依次推得： $P_n = (\lambda/\mu)^n P_0$
- 设$\rho = \frac{\lambda}{\mu} < 1$（否则队列将排至无限远），又由概率性质知 $\sum_{n=0}^{\infty} P_n = 1,$

  将P_n的关系代入：$P_0 \sum_{n=0}^{\infty} \rho^n = P_0 \cdot \frac{1}{1-\rho} = 1$
- 得：

$$P_0 = 1-\rho$$
$$P_n = (1-\rho)\rho^n, \quad n \geq 1 \qquad \rho < 1 \quad (12.19)$$

即为系统状态为n的概率。

– 上式中的ρ具有实际意义：

1）当ρ=λ/μ表达时，是平均到达率与平均服务率之比，即在相同时区内顾客到达的平均数与被服务的平均数之比；

2）当ρ=(1/μ)/(1/λ)表达时，是为一个顾客的服务时间与到达间隔时间之比，称为服务强度（traffic intensity），或话务强度；

3）ρ=1−P_0，刻划了服务机构的繁忙程度，又称服务机构的利用率。

12.3 单服务台负指数分布排队系统的分析(cont.)

– 以（12.19）为基础算出系统的运行指标：

（1）在系统中的平均顾客数L_s（队长期望值）：

$$L_s=\sum_{n=0}^{\infty}nP_n=\sum_{n=0}^{\infty}n(1-\rho)\rho^n=\rho+\rho^2+\rho^3+\cdots=\frac{\rho}{1-\rho}=\frac{\lambda}{\mu-\lambda}$$，其中$0<\rho<1$

（2）在队列中等待的平均顾客数L_q（队列长期望值）：

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

（3）在系统中顾客逗留时间W_s和等待时间的期望值W_q与L_s、L_q有以下关系：

$$L_s=\lambda_{eff}W_s$$

$$L_q=\lambda_{eff}W_q$$

其中λ_{eff}为有效到达率，在本模型中，因为系统容量无限，顾客源无限，因此所有到达顾客都能进入系统，即$\lambda_{eff}=\lambda$，因此有：

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

$$W_q=\frac{\lambda}{\mu(\mu-\lambda)}$$

- 可见顾客在系统中逗留的时间(随机变量)，在$M/M/1$情形下，服从参数为$\mu-\lambda$的负指数分布，即：

  分布函数$F(w)=1-e^{-(\mu-\lambda)w}$，$w\geq 0$

  概率密度$f(w)=(\mu-\lambda)e^{-(\mu-\lambda)w}$

– 将指标归纳如下：

$$(1)\ L_s=\frac{\lambda}{\mu-\lambda}\quad (2)\ L_q=\frac{\rho\lambda}{\mu-\lambda}$$
$$(3)\ W_s=\frac{1}{\mu-\lambda}\quad (4)\ W_q=\frac{\rho}{\mu-\lambda}\qquad (12.21)$$

*Little*公式：

$$(1)\ L_s=\lambda W_s\quad (2)\ L_q=\lambda W_q$$
$$(3)\ W_s=W_q+\frac{1}{\mu}\quad (4)\ L_s=L_q+\frac{\lambda}{\mu}\qquad (12.22)$$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 例：某医院手术室根据病人来诊和完成手术时间的记录，任意抽取100个工作小时和100个完成手术的病历，每小时来就诊的病人数n的出现次数和手术时间v（小时）出现的次数如下表：

到达的病人数n	出现次数f_n
0	10
1	28
2	29
3	16
4	10
5	6
6以上	1
合计	100

为病人完成手术时间v(小时)	出现次数f_v
0.0~0.2	38
0.2~0.4	25
0.4~0.6	17
0.6~0.8	9
0.8~1.0	6
1.0~1.2	5
1.2以上	0
合计	100

12.3 单服务台负指数分布排队系统的分析(cont.)

- 分析：

1. 每小时病人平均到达率 $= \frac{\sum nf_n}{100} = 2.1$(人/小时)

   每次手术平均时间 $= \frac{\sum vf_v}{100} = 0.4$(小时/人)

   每小时完成手术人数（平均服务率）$= \frac{1}{0.4} = 2.5$(人/小时)

2. 取$\lambda = 2.1,\ \mu = 2.5,$ 可以通过统计检验方法（如χ^2检验法），认为病人到达服从参数为2.1的泊松分布，手术时间服从参数为2.5的负指数分布。

3. $\rho = \frac{\lambda}{\mu} = \frac{2.1}{2.5} = 84\%$说明手术室有84%的时间是繁忙的，有16%的时间是空闲的。

4. 依次代入(12.21)算出各指标：

   在病房中病人数(期望值)$L_s = 5.25$(人)

   排队等待病人数(期望值)$L_q = 4.41$(人)

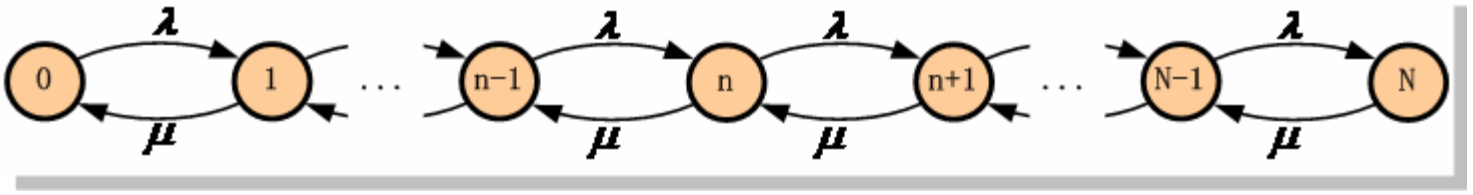
   病人在病房中逗留时间(期望值)$W_s = 2.5$(小时)

   病人排队等待时间(期望值)$W_q = 2.1$(小时)

12.3 单服务台负指数分布排队系统的分析(cont.)

- 系统的容量有限制的情形（M/M/1/N/∞）
  - 标准的M/M/1/N/∞模型指以下模型：
    - 如果系统的最大容量为N，对于单服务台的情形，排队等待的顾客最多为N-1，在某一时刻顾客到达时，如系统中已有N个顾客，那么这个顾客被拒绝进入系统。

    - 只考虑稳态情形，各状态间概率强度的转换关系见下图：



12.3 单服务台负指数分布排队系统的分析(cont.)

- 状态概率的稳态方程为：

$$\begin{cases} \mu P_1 = \lambda P_0 \\ \mu P_{n+1} + \lambda P_{n-1} = (\lambda + \mu) P_n, \quad n \le N-1 \\ \mu P_N = \lambda P_{N-1} \end{cases} \quad (12.23)$$

$$P_0 + P_1 + \cdots + P_N = 1$$

- 令 $\rho = \lambda / \mu$，得：

$$\begin{cases} P_0 = \dfrac{1-\rho}{1-\rho^{N+1}} & \rho \ne 1 \\ P_n = \dfrac{1-\rho}{1-\rho^{N+1}} \rho^n & n \le N \end{cases} \quad (12.24)$$

– 导出系统的各项指标：

- 队长（期望值）：$L_s = \sum_{n=0}^{N} n P_n = \dfrac{\rho}{1-\rho} - \dfrac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}, \quad \rho \ne 1$
- 队列长（期望值）：$L_q = \sum_{n=0}^{N} (n-1) P_n = L_s - (1 - P_0)$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 顾客逗留时间（期望值）：
  平均到达率λ是系统中有空时的平均到达率，系统已满($n=N$)时，到达率为0。
  因此有效到达率$\lambda_{eff}=\lambda(1-P_N)$，可以验证$1-P_0=\lambda_{eff}/\mu$。

  顾客逗留时间期望值$W_s=\dfrac{L_s}{\mu(1-P_0)}=\dfrac{L_q}{\lambda(1-P_N)}+\dfrac{1}{\mu}$

- 顾客等待时间（期望值）：$W_q=W_s-\dfrac{1}{\mu}$

– 将指标归纳如下($\rho\neq1$):

$$\left.\begin{aligned}
&(1)\quad L_s=\frac{\rho}{1-\rho}-\frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}\\
&(2)\quad L_q=L_s-(1-P_0)\\
&(3)\quad W_s=\frac{L_s}{\mu(1-P_0)}\\
&(4)\quad W_q=W_s-1/\mu
\end{aligned}\right\}\quad (12.25)$$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 例：单人理发馆有六个椅子接待人们排队等待。顾客平均到达率为3人/小时，理发需时平均15分钟。
- 分析： 1.N=7为系统中顾客最大数，λ=3人/小时，μ=4人/小时。

2.某顾客一到达就能理发的概率，即理发馆内没有顾客：$P_0=0.2778$。

3.需要等待的顾客数的期望值：

$$L_s=\frac{3/4}{1-3/4}-\frac{8(3/4)^8}{1-(3/4)^8}=2.11$$

$$L_q=L_s-(1-P_0)=2.11-(1-0.2778)=1.39$$

4.有效到达率：

$$\lambda_e=\mu(1-P_0)=4(1-0.2778)=2.89\text{人/小时}$$

5.求一顾客在理发馆内逗留的期望时间：

$$W_s=L_s/\lambda_e=2.11/2.89=0.73\text{小时}=43.8\text{分钟}$$

6.在可能到来的顾客中有百分之几不等待就离开，即系统中有7个顾客的概率：

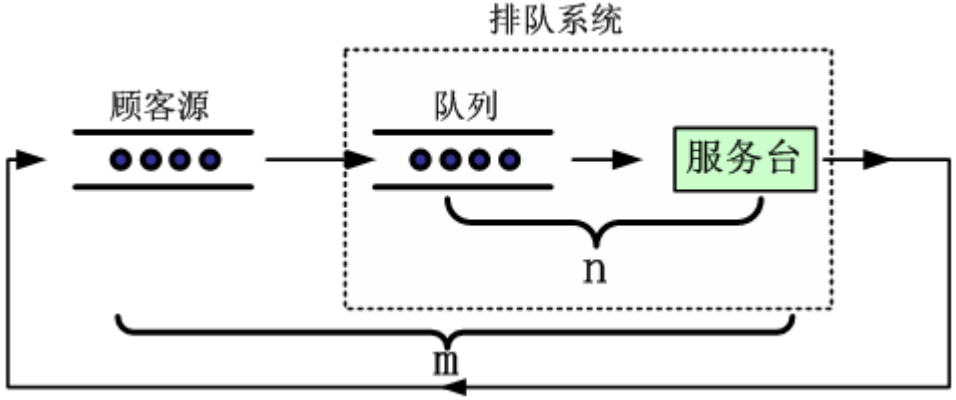
$$P_7=\left(\frac{\lambda}{\mu}\right)^7\left(\frac{1-\lambda/\mu}{1-(\lambda/\mu)^8}\right)\approx 3.7\%$$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 比较理发馆队长为无限和有限的两种结果：

λ=3人/小时 μ=4人/小时	L_s	L_q	W_s	W_q	P_0	可能到来的顾客中有百分之几离开
有限队长N=7	2.11	1.39	0.73	0.48	0.278	3.7%
无限队长	3	2.25	1.0	0.75	0.25	0

- **顾客源为有限的情形（M/M/1/∞/m）**
  - 顾客总体虽然只有m个，但每个顾客到来并经过服务后仍回到原来总体，所以仍然可以到来。虽然该模型中对系统容量没有限制，但实际上它永远不会超过m，所以和写成（M/M/1/m/m）的意义相同。



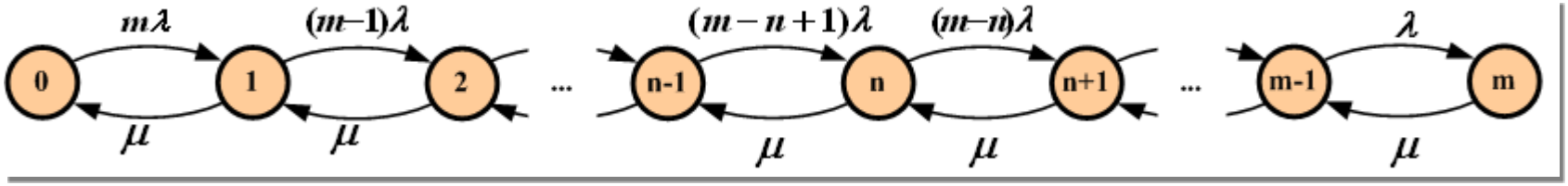
12.3 单服务台负指数分布排队系统的分析(cont.)

– 平均到达率

- 在无限源的情形按全体顾客来考虑，在有限源的情形下必须按每个顾客来考虑。设单个顾客的到达率都是相同的λ，系统的有效到达率为：

$$\lambda_{eff} = \lambda(m - L_s)$$

- 模型的状态转移关系：



- 各状态间的转移差分方程：

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

$$\sum_{i=0}^{m} P_i = 1$$

12.3 单服务台负指数分布排队系统的分析(cont.)

- 解上述方程得：

$$\left.\begin{aligned} P_0 &= \frac{1}{\sum_{i=0}^{m}\frac{m!}{(m-i)!}(\frac{\lambda}{\mu})^i} \\ P_n &= \frac{m!}{(m-n)!}(\frac{\lambda}{\mu})^n P_0 \quad (1\le n\le m) \end{aligned}\right\} \quad (12.27)$$

- 系统的各项指标为：

$$\left.\begin{aligned} &(1)\ \ L_s = m-\frac{\mu}{\lambda}(1-P_0) \\ &(2)\ \ L_q = m-\frac{(\lambda+\mu)(1-P_0)}{\lambda} = L_s-(1-P_0) \\ &(3)\ \ W_s = \frac{m}{\mu(1-P_0)}-\frac{1}{\lambda} \\ &(4)\ \ W_q = W_s - 1/\mu \end{aligned}\right\} \quad (12.28)$$

– 该模型的物理含义：

- 该模型可以用于表示m台机器中因故障而停机等待维修的问题，也可以用于表示m个打字员使用1台打印机等待打印服务的问题。在机器故障问题中L_s就是平均故障台数，$m-L_s=(\lambda/\mu)(1-P_0)$表示正常运转的台数。

12.3 单服务台负指数分布排队系统的分析(cont.)

- 例：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15分钟，有一个修理工，每次修理时间服从负指数分布，平均每次12分钟。
- 分析：$m=5,\lambda=1/15,\mu=1/12,\lambda/\mu=0.8$

(1)修理工空闲概率：

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

(2)五台机器都出故障的概率：$P_5=\frac{5!}{0!}(0.8)^5P_0=0.287$

(3)出故障的平均台数：$L_s=5-\frac{1}{0.8}(1-0.0073)=3.76$(台)

(4)等待修理的平均台数：$L_q=3.76-0.993=2.77$(台)

(5)平均停工时间：$W_s=\frac{5}{\frac{1}{12}(1-0.0073)}-15=46$(分钟)

(6)平均等待修理时间：$W_q=46-12=34$(分钟)

- 机器平均停工时间为46分钟，过长，修理工几乎没有空闲时间，应当提高服务率减少修理时间或增加修理工人。

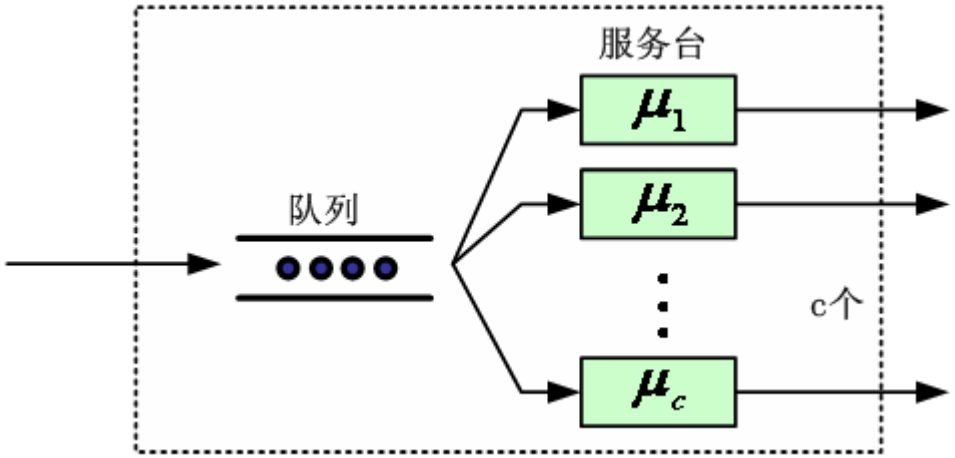
12.4 多服务台负指数分布排队系统的分析

- 标准的M/M/c模型（M/M/c/∞/∞）
  - 标准的M/M/c模型各种特征的规定与标准的M/M/1模型相同。另外规定各服务台是相互独立（不搞协作）且平均服务率相同，即：

$$\mu_1 = \mu_2 = \cdots = \mu_c = \mu$$

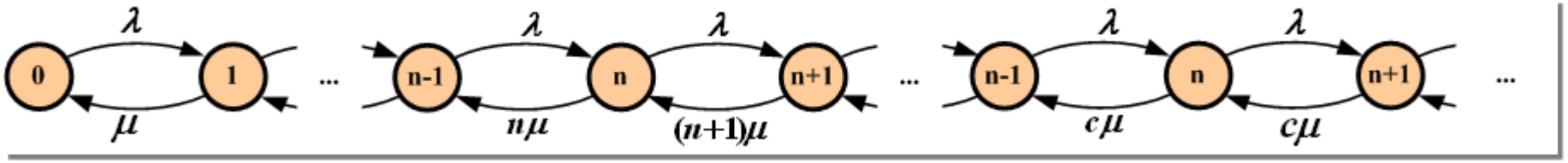
    - 整个服务机构的平均服务率为：当$n \geq c$时：$c\mu$

      当$n < c$时：$n\mu$
    - 令ρ=λ/(cμ)，只有ρ<1时才不会形成无限的队列，称之为这个系统的服务强度或称服务机构的平均利用率。



12.4 多服务台负指数分布排队系统的分析(cont.)

– 状态间的转移关系为：



– 状态转移方程为：

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

$$\sum_{i=0}^{\infty} P_i = 1, \quad \text{且} \rho \le 1$$

– 解方程得：

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

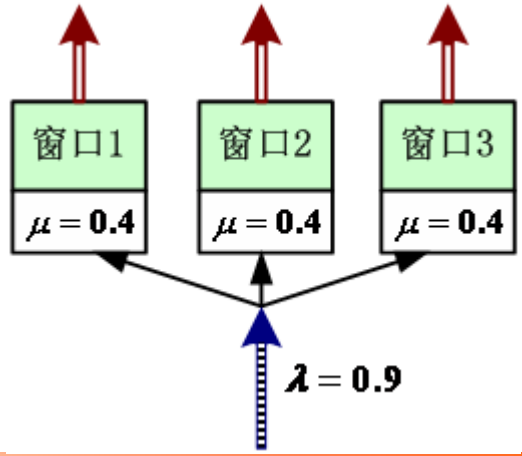
$$P_n = \begin{cases} \dfrac{1}{n!} \left(\dfrac{\lambda}{\mu}\right)^n P_0 & (n \le c) \\ \dfrac{1}{c! c^{n-c}} \left(\dfrac{\lambda}{\mu}\right)^n P_0 & (n > c) \end{cases} \qquad (12.29)$$

12.4 多服务台负指数分布排队系统的分析(cont.)

- 系统的各项运行指标为：

$$\left.\begin{aligned} L_s &= L_q + \frac{\lambda}{\mu} \\ L_q &= \sum_{n=c+1}^{\infty}(n-c)P_n = \frac{(c\rho)^c\rho}{c!(1-\rho)^2}P_0 \\ W_q &= \frac{L_q}{\lambda} \\ W_s &= \frac{L_s}{\lambda} \end{aligned}\right\} \quad (12.30)$$

- 例：某售票所有三个窗口，顾客的到达服从普阿松分布，平均到达率每分钟λ=0.9(人)，服务（售票）时间服从负指数分布，平均服务率每分钟μ=0.4(人)。现设顾客到达后排成一队，依次向空闲窗口购票（如图）。



12.4 多服务台负指数分布排队系统的分析(cont.)

- 分析:

$c=3,\frac{\lambda}{\mu}=2.25,\ \rho=\frac{\lambda}{c\mu}=\frac{2.25}{3}<1$符合$M/M/c$模型要求的条件。

(1)整个售票所空闲概率:$P_0=\left[\frac{(2.25)^0}{0!}+\frac{(2.25)^1}{1!}+\frac{(2.25)^2}{2!}+\frac{(2.25)^3}{3!}\cdot\frac{1}{1-2.25/3}\right]^{-1}=0.0748$

(2)平均队长:$L_q=\frac{(2.25)^3\cdot 3/4}{3!(1/4)^2}\times 0.0748=1.70$

$L_s=L_q+\frac{\lambda}{\mu}=3.95$

(3)平均等待时间和逗留时间:

$W_q=1.70/0.9=1.89$分钟

$W_s=1.89+1/0.4=4.39$分钟

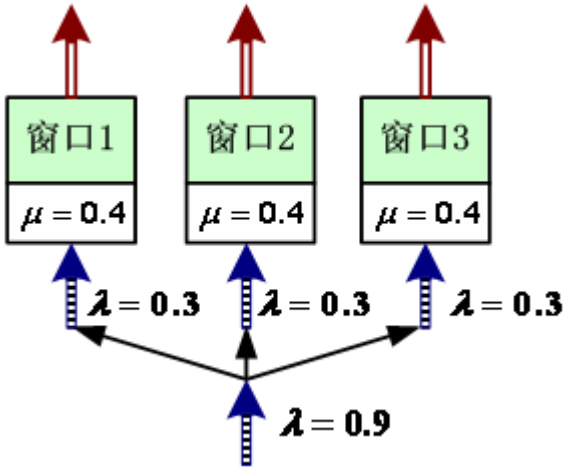
(4) 顾客到达后必须等待(即系统中顾客数已有3人、各服务台都没有空闲)的概率:

$P(n\geq 3)=\frac{(2.25)^3}{3!(1/4)}\times 0.0748=0.57$

12.4 多服务台负指数分布排队系统的分析(cont.)

– M/M/c型系统和c个M/M/1型系统的比较

- 以上例为例，上例可转化为右图形式：



- 按3个M/M/1型系统计算，并与上例结果对比：

指标 \ 模型	M/M/3	M/M/1
服务台空闲的概率P_0	0.0748	0.25（每个子系统）
顾客必须等待的概率	P(n≥3)=0.57	0.75
平均队列长L_q	1.70	2.25（每个子系统）
平均队长L_s	3.95	9.00（整个系统）
平均逗留时间W_q	4.39	10
平均等待时间W_s	1.89	7.5

12.4 多服务台负指数分布排队系统的分析(cont.)

– 多服务台的$W_q \cdot \mu$的数值表以方便计算：

λ/cμ	服务台数				
	c=1	c=2	c=3	c=4	c=5
0.1	0.1111	0.0101	0.0014	0.0002	小于0.00005
0.2	0.2500	0.0417	0.0103	0.0030	0.0010
0.3	0.4286	0.0989	0.0333	0.0132	0.0058
0.4	0.6667	0.1905	0.0784	0.0378	0.0199
0.5	1.0000	0.3333	0.1579	0.0870	0.0521
0.6	1.5000	0.5625	0.2956	0.1794	0.1181
0.7	2.3333	0.9608	0.5470	0.3572	0.2519
0.8	4.0000	1.7778	1.0787	0.7455	0.5541
0.9	9.0000	4.2632	2.7235	1.9694	1.5250
0.95	19.0000	9.2564	6.0467	4.4571	3.5112

- 如上例中，已知c=3，ρ=0.75，查表无此数，用线性插值法得：

$W_q \cdot \mu$=(1.0787+0.5470)/2=0.8129，因μ=0.4，所以W_q=2.03分，W_s=4.53分，L_q=2.2人，L_s=4.45人。

结果与前面的计算略有差异，是由于使用线性插值法引起的。

12.4 多服务台负指数分布排队系统的分析(cont.)

- 系统的容量有限制的情形（M/M/c/N/∞）
  - 当系统中顾客数n已达到N时，再来的顾客被拒绝，其他条件与标准的M/M/c型相同。
    - 系统的状态概率和运行指标，其中ρ=λ/(cμ)，但不对ρ加以限制：

$$P_0=\frac{1}{\sum_{k=0}^{c}\frac{(c\rho)^k}{k!}+\frac{c^c}{c!}\cdot\frac{\rho(\rho^c-\rho^N)}{1-\rho}}\qquad \rho\neq 1$$

$$P_n=\begin{cases}\dfrac{(c\rho)^n}{n!}P_0 & (0\le n\le c)\\ \dfrac{c^c}{c!}\rho^n P_0 & (c\le n\le N)\end{cases}\qquad (12.31)$$

$$\left.\begin{aligned}L_q&=\frac{P_0\rho(c\rho)^c}{c!(1-\rho)^2}[1-\rho^{N-c}-(N-c)\rho^{N-c}(1-\rho)]\\ L_s&=L_q+c\rho(1-P_N)\\ W_q&=\frac{L_q}{\lambda(1-P_N)}\\ W_s&=W_q+\frac{1}{\mu}\end{aligned}\right\}\qquad (12.32)$$

12.4 多服务台负指数分布排队系统的分析(cont.)

– 当c=N(即时制)的情形（例如街头的停车场不允许排队等待空位）：

$$\left.\begin{aligned} P_0 &= \frac{1}{\sum_{k=0}^{c}\frac{(c\rho)^k}{k!}} \\ P_n &= \frac{(c\rho)^n}{n!}P_0, \quad 0 \le n \le c \end{aligned}\right\} \quad (12.33)$$

- 当n=c时即关于Pc的公式，被称为爱尔朗呼唤公式，是A.K.Erlang在1917年发现的，广泛应用于电话系统的设计中。
- 此时运行指标如下：

$$\left.\begin{aligned} L_q &= 0, \quad W_q = 0, \quad W_s = \frac{1}{\mu} \\ L_s &= \sum_{n=1}^{c} nP_n = \frac{c\rho\sum_{n=0}^{c-1}\frac{(c\rho)^{n-1}}{n!}}{\sum_{n=0}^{c}\frac{(c\rho)^n}{n!}} = c\rho(1-P_c) \end{aligned}\right\} \quad (12.34)$$

12.4 多服务台负指数分布排队系统的分析(cont.)

- 例：某旅馆顾客到达数为泊松流，每天平均到λ=6人，顾客平均逗留时间1/μ=2天，试就该旅馆在具有c=1、2、3、……、8个房间的前提下分别计算每天客房平均占用数L_s及满员概率P_c。
- 分析：在客房满员条件下，旅客不能排队等待。

$$\lambda = 6, 1/\mu = 2,\ c\rho = \lambda/\mu = 12$$

- 计算步骤如下：

(1)n	(2)$(c\rho)^n = 12^n$	(3)$n!$	(4)$(c\rho)^n / n!$	(5)$\sum_{n=0}^{c} \frac{(c\rho)^n}{n!}$	(6)P_c	(7)$\sum_{n=0}^{c-1} \Big/ \sum_{n=0}^{c}$	(8)L_s
0	1	1	1	1	1	-	-
1	1.2×10	1	12	13	0.92	0.08	0.92
2	1.44×10^2	2	72	85	0.85	0.15	1.83
3	1.73×10^3	6	288	373	0.77	0.23	2.74
4	2.07×10^4	24	864	1.24×10^3	0.70	0.30	3.62
5	2.49×10^5	120	2.07×10^3	3.31×10^3	0.63	0.37	4.48
6	2.99×10^6	720	4.15×10^3	7.46×10^3	0.56	0.44	5.33
7	3.58×10^7	5.04×10^3	7.11×10^3	1.45×10^4	0.49	0.51	6.14
8	4.30×10^8	4.03×10^4	1.07×10^4	2.52×10^5	0.42	0.58	6.93

12.4 多服务台负指数分布排队系统的分析(cont.)

- 顾客源为有限的情形（M/M/c/∞/m）
  - 顾客总体为有限数m，且m>c，和单服务台情形一样，顾客到达率 λ 是按每个顾客来考虑的。

▪在机器管理问题中，就是共有m台机器，c个修理工人，每个顾客的到达率λ是指每台机器每单位运转时间出故障的期望次数。系统中顾客数n就是出故障的机器台数，当n≤c时，所有的故障机器都在修理，有c-n个修理工人在空闲；当c<n≤m时，有n-c台机器在停机等待修理，修理工人都处在繁忙状态。假定这c个工人修理技术相同，服务时间都服从参数为μ的负指数分布。

$$P_0=\frac{1}{m!}\cdot\frac{1}{\sum_{k=0}^{c}\frac{1}{k!(m-k)!}\left(\frac{c\rho}{m}\right)^k+\frac{c^c}{c!}\sum_{k=c+1}^{m}\frac{1}{(m-k)!}\left(\frac{\rho}{m}\right)^k}$$

$$P_n=\begin{cases}\dfrac{m!}{(m-n)!n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & (0\le n\le c)\\ \dfrac{m!}{(m-n)!c!c^{n-c}}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & (c+1\le n\le m)\end{cases} \qquad (12.35)$$

其中，$\rho=\dfrac{m\lambda}{c\mu}$

12.4 多服务台负指数分布排队系统的分析(cont.)

- 平均顾客数（即平均故障台数）

$$L_s = \sum_{n=1}^{m} nP_n$$

$$L_q = \sum_{n=c+1}^{m} (n-c)P_n$$

- 有效的到达率λ_e等于每个顾客的到达率λ乘以在系统外（即正常生产的）机器的期望数：

$$\lambda_{eff} = \lambda(m - L_s)$$

在机器故障问题中，它是每单位时间m台机器平均出现故障的次数。

- 各项指标之间的关系：

$$\left.\begin{aligned} L_s &= L_q + \frac{\lambda_e}{\mu} = L_q + \frac{\lambda}{\mu}(m - L_s) \\ M_s &= L_s / \lambda_e \\ W_q &= L_q / \lambda_e \end{aligned}\right\} \quad (12.36)$$

12.4 多服务台负指数分布排队系统的分析(cont.)

- 例：设有两个修理工人，负责5台机器的正常运行，每台机器平均损坏率为每运转1小时1次，两工人能以相同的平均修复率4（次/小时）修好机器。
- 分析：

$m=5,\ \lambda=1(\text{次/小时}),\ \mu=4(\text{台/小时}),\ c=2,\ c\rho/m=\lambda/\mu=1/4$

$$P_0=\frac{1}{5!}\left[\frac{1}{5!}\left(\frac{1}{4}\right)^0+\frac{1}{4!}\left(\frac{1}{4}\right)^1+\frac{1}{2!3!}\left(\frac{1}{4}\right)^2+\frac{2^2}{2!}\frac{1}{2!}\left(\frac{1}{8}\right)^3+\left(\frac{1}{8}\right)^4+\left(\frac{1}{8}\right)^5\right]^{-1}=0.3149$$

$P_1=0.394,\ P_2=0.197,\ P_3=0.074,\ P_4=0.018,\ P_5=0.002$

等待修理的机器平均数 $L_q=P_3+2P_4+3P_5=0.118$

需要修理的机器平均数 $L_s=\sum_{n=1}^{m}nP_n=L_q+c-2P_0-P_1=1.094$

有效损坏率：$\lambda_{eff}=1\times(5-1.094)=3.906$

等待修理时间：$W_q=0.118/3.906=0.03$小时

停工时间：$W_s=1.094/3.906=0.28$小时

12.5 一般服务时间M/G/1模型*

- 服务时间是任意分布的情形：
  - E[系统中顾客数]$=E$[队列中顾客数]$+E$[服务机构中顾客数]
    E[在系统中逗留时间]$=E$[排队等候时间]$+E$[服务时间]
    其中$E[\cdot]$表示求期望值，用符号表示即：

$$\left.\begin{aligned} L_s &= L_q + L_{se} \\ W_s &= W_q + E[T] \end{aligned}\right\} \quad (12.37)$$

  - T表示服务时间（随机变量），当T服从负指数分布时，$E[T]=1/\mu$。
- **Pollaczek-Khintchine(P-K)公式**
  - 对于M/G/1模型，服务时间T的分布是一般的（但要求期望值E[T]和方差Var[T]都存在），其他条件和标准的M/M/1型相同。为了达到稳态，要求ρ= λ E[T]<1。有：

$$L_s = \rho + \frac{\rho^2 + \lambda^2 Var[T]}{2(1-\rho)} \quad (12.38)$$

Pollaczek-Khintchine(P-K)公式

12.5 一般服务时间M/G/1模型(cont.)

- 由上式可知，不管T是什么分布，只要知道λ，E[T]和Var[T]，就可以求出L_s，并根据各项指标之间的关系式求出L_q、W_q和W_s。
- 例9：有一售票口，已知顾客按平均为2分30秒的时间间隔的负指数分布到达，顾客在售票口前服务时间平均为2分钟。在下面两种情况下求顾客购票的平均逗留时间和等待时间：(1)服务时间也服从负指数分布；(2)经过调查，顾客在售票窗口至少要占用1分钟，服从以下概率密度分布：

$$f(y)=\begin{cases} e^{-y+1}, & y\geq 1 \\ 0, & y<1 \end{cases}$$

- 分析：

$(1)\lambda=1/2.5=0.4,\ \mu=1/2=0.5,\ \rho=\lambda/\mu=0.8$

$W_s=\frac{1}{\mu-\lambda}=10$分，$W_q=\frac{\rho}{\mu-\lambda}=8$分

(2)令y为服务时间，那么$Y=1+X$，X服从均值为1的负指数分布，则

$E[Y]=2,\ Var[Y]=Var[1+X]=Var[X]=1$

$\rho=\lambda E[Y]=0.8$，代入$P-K$公式得：

$L_s=0.8+\frac{0.8^2+0.4^2\times 1}{2\times(1-0.8)}=2.8,\ L_q=L_s-\rho=2$

$W_s=L_s/\lambda=7$分，$W_q=L_q/\lambda=5$分

12.5 一般服务时间M/G/1模型(cont.)

- 定长服务时间M/D/1模型
  - 服务时间是确定的常数，此时：

$$T = 1/\mu$$
$$Var[T] = 0$$
$$L_s = \rho + \frac{\rho^2}{2(-\rho)} \quad (12.39)$$

- 例：某实验室有一台自动检验机器性能的仪器，检验机器的顾客按泊松分布到达，每小时平均4个顾客，检验每台机器需要6分钟。
- 分析：$\lambda = 4,\ E[T] = 1/10(\text{小时}),\ \rho = 4/10,\ Var[T] = 0$

在检验室内机器台数期望值 $L_s = 0.4 + \frac{(0.4)^2}{2(1-0.4)} = 0.533(\text{台})$

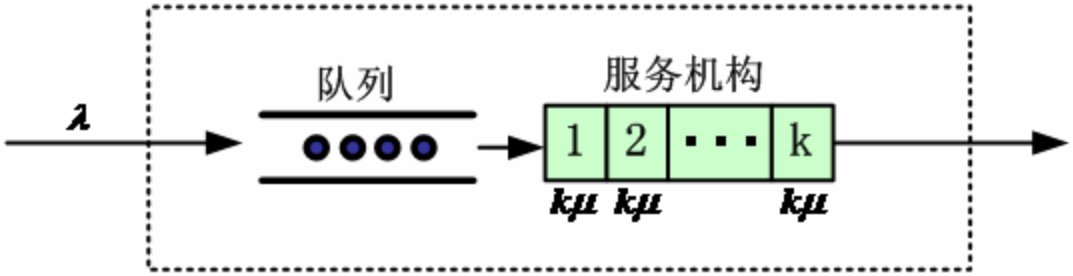
等候检验的机器台数期望值 $L_q = 0.533 - 0.4 = 0.133(\text{台})$

每台机器在室内逗留时间期望值 $W_s = \frac{0.533}{4} = 0.133\text{小时}$

每台机器平均等待检验的时间期望值 $W_q = \frac{0.133}{4} = 0.033\text{小时}$

12.5 一般服务时间M/G/1模型(cont.)

- 爱尔朗服务时间$M/E_k/1$模型
  - 如果顾客必须经过k个服务站，在每个服务站的服务时间T_i相互独立并服从相同的负指数分布（参数为kμ），则：$T=\sum_{i=1}^{k}T_i$ 服从k阶爱尔朗分布。



$$E[T_i]=\frac{1}{k\mu} \qquad Var[T_i]=\frac{1}{k^2\mu^2}$$

$$E[T]=\frac{1}{\mu} \qquad Var[T]=\frac{1}{k\mu^2}$$

  - 该模型的各项指标：

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

$$L_q=\frac{(k+1)\rho^2}{2k(1-\rho)} \qquad (12.40)$$

$$W_s=L_s/\lambda,\quad W_q=L_q/\lambda$$

12.5 一般服务时间M/G/1模型(cont.)

- 例：某单人裁缝店做西服，每套需经过4个不同的工序，4个工序完成后才能开始做另一套。每一个工序的时间服从负指数分布，期望值为2小时。顾客到来服从泊松分布，平均订货率为5.5套/周（设一周6天为工作日，每天8小时）。顾客为等到做好一套西服的期望时间有多长?
- 分析：$1/4\mu = 2$小时，$\mu = \frac{1}{8}$套/小时$= 6$套/周，$\rho = 5.5/6$

$$E[T_i] = 2, \quad Var[T_i] = \left(\frac{1}{4\times 6}\right)^2$$

$$E[T] = 8, \quad Var[T] = \frac{1}{4\times 6^2}$$

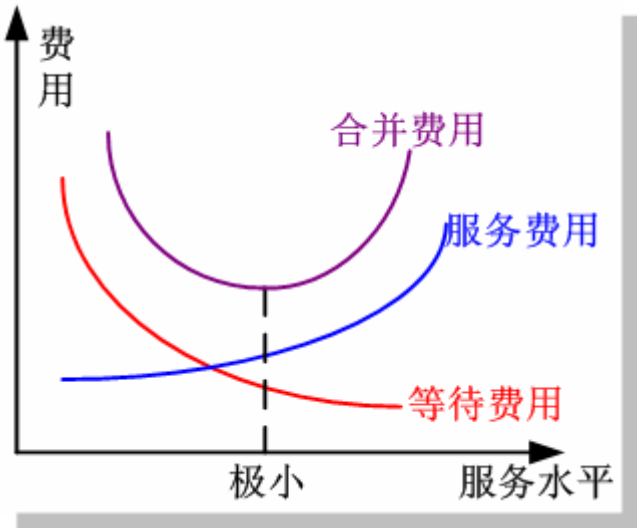
$$L_s = \frac{5.5}{6} + \frac{\left(\frac{5.5}{6}\right)^2 + (5.5)^2 \times \frac{1}{4\times 6^2}}{2\left(1-\frac{5.5}{6}\right)} = 7.2188$$

顾客为等到做好一套西服的期望时间：

$W_s = L_s/\lambda = 1.3$周

12.6 经济分析——系统的最优化

- 排队系统的最优化问题
  - 系统设计的最优化
    - 称为静态问题，目的在于使设备达到最大效益，在一定的质量指标下要求机构最为经济。
  - 系统控制的最优化
    - 称为动态问题，是指给定一个系统，如何运营可使某个目标函数得到最优。



- 在一般情形下，提供服务水平（数量、质量）自然会降低顾客的等待费用（损失），但却常常增加了服务机构的成本。
- 最优化的目标之一是使二者费用之和为最小；另一个常用的目标函数是使纯收入或使利润（服务收入与服务成本之差）最大。

12.6 经济分析——系统的最优化(cont.)

- 费用
  - 各种费用在稳态情形下都是按单位时间来考虑的，一般情形，服务费用（成本）可以确切计算或估计，但是顾客的等待费用（损失）有许多不同情况。
- 服务水平
  - 平均服务率**μ**（代表服务机构的服务能力和经验等）；
  - 服务设备，如服务台个数**c**;
  - 由队列所占空间大小所决定的队列最大限制数**N**;
  - 服务水平也可以通过服务强度 ρ 来表示。
- 常用的求解方法：
  - 对于离散变量常用边际分析法；
  - 对于连续变量常用经典的微分法；
  - 对于复杂的问题，也可以使用非线性规划或动态规划方法。

12.6 经济分析——系统的最优化(cont.)

- M/M/1模型中最优服务率μ
  - 标准的M/M/1模型

取目标函数z为单位时间服务成本与顾客在系统逗留费用之和的期望值：

$$z = c_s\mu + c_w L_s \quad (12.41)$$

其中c_s为当$\mu = 1$时服务机构单位时间的费用，

c_w为每个顾客在系统停留单位时间的费用。

又$L_s = \dfrac{\lambda}{\mu - \lambda}$，因此$z = c_s\mu + c_w \cdot \dfrac{\lambda}{\mu - \lambda}$，求极小值，令

$$\frac{dz}{d\mu} = c_s - c_w \cdot \frac{\lambda}{(\mu-\lambda)^2} = 0,$$ 解得：

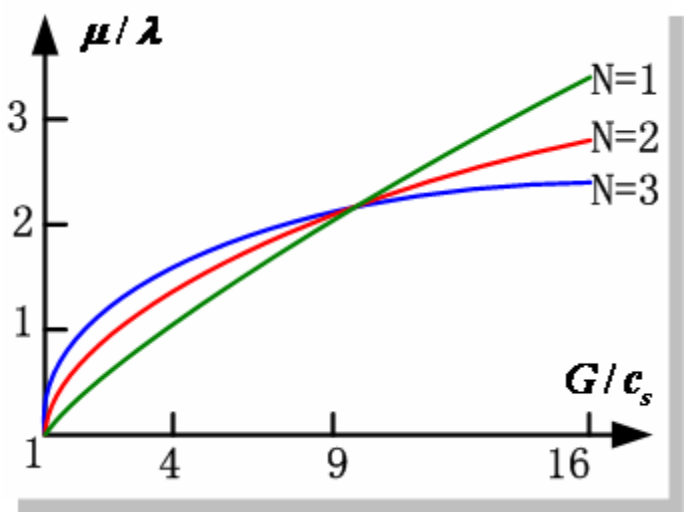
$$\mu^* = \lambda + \sqrt{\frac{c_w}{c_s}\lambda}$$

（为了保证$\rho < 1$，$\mu > \lambda$，根号前取+号）

12.6 经济分析——系统的最优化(cont.)

– 系统中顾客最大限制数为N的情形

- 系统中若有N个顾客，则后来的顾客被拒绝。P_N即为被拒绝的概率（呼损率），$1-P_N$即为能接受服务的概率，$\lambda(1-P_N)$即为单位时间实际进入服务机构顾客的平均数，在稳定状态下，等于单位时间内实际服务完成的平均顾客数。



设每服务1人能收入G元，于是单位时间收入的期望值是$\lambda(1-P_n)G$元，纯利润$z=\lambda(1-P_n)G-c_s\mu$

$$=\lambda G\cdot\frac{1-\rho^N}{1-\rho^{N+1}}-c_s\mu=\lambda\mu G\cdot\frac{\mu^N-\lambda^N}{\mu^{N+1}-\lambda^{N+1}}-c_s\mu$$

令$\frac{dz}{d\mu}=0$，得$\rho^{N+1}\cdot\frac{N-(N+1)\rho+\rho^{N+1}}{(1-\rho^{N+1})^2}=\frac{c_s}{G}$

最优的μ^*应满足上式。

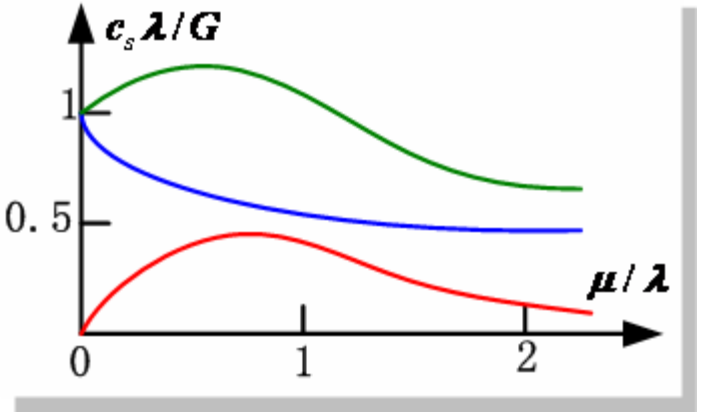
上式中，c_s、G、λ、N都是给定的，通常通过数值计算来求出μ^*，或将上式左方（对一定的N）作为ρ的函数做出图形，对于给定的G/c_s，根据图形可求出μ^*/λ。

12.6 经济分析——系统的最优化(cont.)

– 顾客源为有限的情形（以机器故障问题为例）

- 共有机器m台以及1个修理工人，各台连续运转时间和修理时间均服从负指数分布。当服务率μ=1的时的修理费用c_s，单位时间每台机器运转可得收入G元。平均运转台数为$m-L_s$，所以单位时间纯利润为：

$$z=(m-L_s)G-c_s\mu=\frac{mG}{\rho}\cdot\frac{E_{m-1}\left(\frac{m}{\rho}\right)}{E_m\left(\frac{m}{\rho}\right)}-c_s\mu$$



式中$E_m(x)=\sum_{k=0}^{m}\frac{x_k}{k!}e^{-x}$称为泊松部分和；$\rho=\frac{m\lambda}{\mu}$，而$\frac{d}{dx}E_m(x)=E_{m-1}(x)-E_m(x)$，令$\frac{dz}{dx}=0$，得：

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

，给定m、G、c_s、λ，求解μ^*很困难，通常利用泊松分布表通过数值计算获得，或将上式左方（对一定的m）作为ρ的函数做出图形，对于给定的$\frac{c_s\lambda}{G}$根据图形可求出μ^*/λ。

12.6 经济分析——系统的最优化(cont.)

- **M/M/c模型中最优的服务台数c**
  - 标准的M/M/c模型在稳态情形下单位时间全部费用（服务成本与等待费用之和）的期望值：$z = c_s' \cdot c + c_w \cdot L \quad (12.43)$

    其中c是服务台数，c_s'是每服务台单位时间的成本，

    c_w为每个顾客在系统停留单位时间的费用，

    L是系统中顾客平均数L_s或队列中等待的顾客平均数L_q（它们随z值的不同而不同）。

    因为c_s'和c_w都是给定的，唯一可能变动的是服务台数c，所以z是c的函数$z(c)$。

    - 因为c只取整数值，z(c)不是连续变量的函数，采用边际分析法(Marginal Analysis)，根据z(c*)最小的特点：

$$\begin{cases} z(c^*) \le z(c^*-1) \\ z(c^*) \le z(c^*+1) \end{cases} \Rightarrow \begin{cases} c_s' \cdot c^* + c_w L(c^*) \le c_s'(c^*-1) + c_w L(c^*-1) \\ c_s' \cdot c^* + c_w L(c^*) \le c_s'(c^*+1) + c_w L(c^*+1) \end{cases}$$

$$\Rightarrow L(c^*) - L(c^*+1) \le c_s'/c_w \le L(c^*-1) - L(c^*)$$

    - 依次求$c = 1,2,3,\cdots$时L的值，并作两相邻的L值之差，因c_s'/c_w是已知数，根据这个数落在哪个不等式区间里就可以确定出c^*。

12.6 经济分析——系统的最优化(cont.)

- 例：某检验中心为各工厂服务，要求做检验的工厂的到来服从泊松流，平均到达率λ为每天48次，每次来检验由于停工等原因损失为6元。服务时间服从负指数分布，平均服务率μ为每天25次，每设置一个检验员服务成本为每天4元，其他条件符合标准M/M/c模型。问应设置几个检验员才能使总费用的期望值最小？
- 分析：$c_s' = 4$元/每检验员，$c_w = 6$元/次，$\lambda = 48,\ \mu = 25,\ \lambda/\mu = 1.92$

  设检验员数为c，令c依次为1、2、3、4、5，根据已有的计算表求出L_s:

c	1	2	3	4	5
$\lambda/c\mu$	1.92	0.96	0.64	0.48	0.38
查多服务台的$W_q \cdot \mu$计算表	-	10.2550	0.3961	0.0772	0.0170
$L_s = \frac{\lambda}{\mu}(W_q \cdot \mu + 1)$	-	21.610	2.680	2.068	1.952

12.6 经济分析——系统的最优化(cont.)

- 将L_s代入得表:

检验员数 c	来检验顾客数 $L_s(c)$	L(c)-L(c+1)~ L(c)-L(c-1)	总费用(每天) z(c)
1	∞		∞
2	21.610	18.930~∞	154.94
3	2.680	0.612~18.930	27.87
4	2.068	0.116~0.612	28.38
5	1.952		31.71

$\frac{c_s'}{c_w} = 0.66$, 落在区间$(0.612 \sim 18.930)$内

$\therefore c^* = 3$

即设3个检验员使总费用最小, 此时$z(c^*) = 27.87$(元)

本章完

The end