

# 第八章 排队理论概要

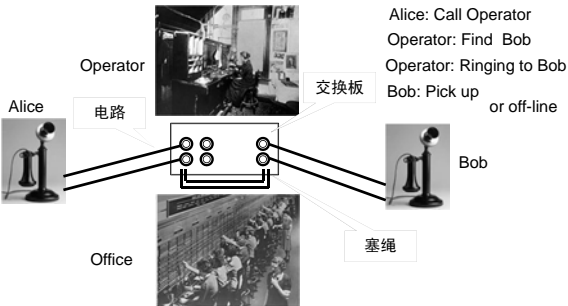
网络性能与服务质量评价方法

## 第八章 排队理论基础

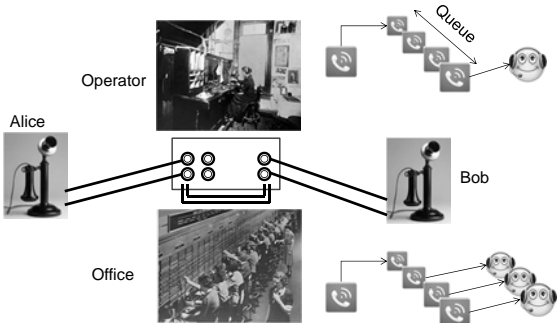
- 8.1 排队模型
- 8.2 M/M/1系统
- 8.3 M/G/1系统
- 8.4 M/M/n系统
- 8.5 排队网络

## 排队模型的来源

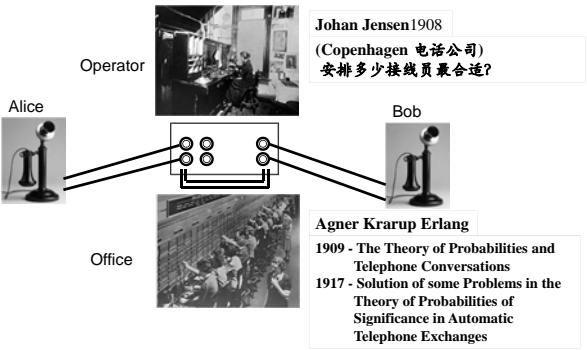
## 电话服务设施及呼叫控制



## 呼叫等待过程



## 排队的技术问题

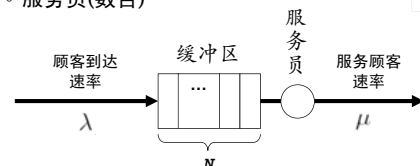


## 排队模型的表示

## 图形表示

### 直观表示

- 顾客到达和离去(过程)
- 缓冲区(容量)
- 服务员(数目)



## Kendall记法

### 形式化表示

- $A/B/C/D/E/F$
- $A/S/s/c/p/D$



- A:** 顾客到达 (Arrival) 的时间间隔分布  
**B:** 服务 (Service) 时长的分布  
**C:** 服务员 (server) 的数目  
**D:** 排队机的容量 (capacity)  
**E:** 顾客总数 (population)  
**F:** 队列调度规则 (Discipline)

## 排队系统的典型类型

## 到达过程, Arrival

Symbol	Name	Description	Examples
M	Markovian or memoryless	Poisson process (or random) arrival process.	M/M/1
M <sup>x</sup>	batch Markov	Poisson process with a random variable X for the number of arrivals at one time.	M <sup>x</sup> /M <sup>y</sup> /1
MAP	Markovian arrival process	Generalisation of the Poisson process.	
BMAP	Batch Markovian arrival process	Generalisation of the MAP with multiple arrivals	
MMPP	Markov modulated poisson process	Poisson process where arrivals are in "clusters".	
D	Degenerate distribution	A deterministic or fixed inter-arrival time.	D/M/1
E <sub>k</sub>	Erlang distribution	An Erlang distribution with k as the shape parameter.	
G	General distribution	Although G usually refers to independent arrivals, some authors prefer to use GI to be explicit.	
PH	Phase-type distribution	Some of the above distributions are special cases of the phase-type, often used in place of a general distribution.	

[http://en.wikipedia.org/wiki/Kendall\\_notation](http://en.wikipedia.org/wiki/Kendall_notation)

## 排队规则, Discipline

Symbol	Name	Description
FIFO/FCFS	First In First Out/First Come First Served	The customers are served in the order they arrived in.
LIFO/LCFS	Last in First Out/Last Come First Served	The customers are served in the reverse order to the order they arrived in.
SIRO	Service In Random Order	The customers are served in a random order with no regard to arrival order.
PNPN	Priority service	Priority service, including preemptive and non-preemptive.
PS	Processor Sharing	

第三章 通信网控制与信令

8.1 排队模型

8.2 M/M/1系统

8.3 M/G/1系统

8.4 M/M/m系统

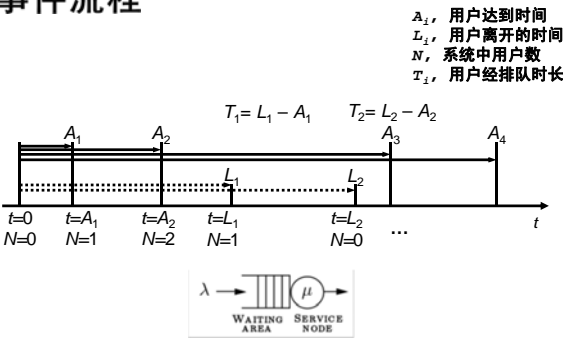
8.5 排队网络



M/M/1 排队系统示例



事件流程



时序表

$t$	事件	$\lambda$	$N$	$T_i$
0	开始	0	0	-
$A_1$	用户1到达	1	1	-
$A_2$	用户2到达	2	2	-
$L_1$	用户1离去	2	1	$T_1=L_1-A_1$
$L_2$	用户2离去	2	0	$T_2=L_2-A_2$
$A_3$	用户3到达			
$A_4$	用户3到达			
统计平均		$2/L_2$	?	$(T_1+T_2)/2$

加权统计

$t$	事件	$\lambda$	$N$	$T_i$
0	开始	0	0	-
$A_1$	用户1到达	1	1	-
$A_2$	用户2到达	2	2	-
$L_1$	用户1离去	2	1	$T_1=L_1-A_1$
$L_2$	用户2离去	2	0	$T_2=L_2-A_2$
...				

加权统计平均       $2/L_2$        $(T_1+T_2)/2$

$$\begin{aligned} & 0 \times A_1/L_2 + \\ & 1 \times (A_2-A_1)/L_2 + \\ & 2 \times (L_1-A_2)/L_2 + \\ & 1 \times (L_2-L_1)/L_2 \end{aligned}$$

平均值之间关系(Little's law)

$$\begin{aligned} \langle N \rangle &= [(A_2-A_1) + 2(L_1-A_2) + (L_2-L_1)]/L_2 \\ &= [-A_1 - A_2 + L_1 + L_2]/L_2 \\ &= (T_1+T_2)/L_2 \\ &= [(T_1+T_2)/2] \times [2/L_2] \\ &= \langle T \rangle \times \lambda \end{aligned}$$

从用户到达至离去的时间之内，  
<T>，  
所有后续到达的所有用户，  
<N>，  
都要滞留在排队系统中。

$\lambda$        $\langle N \rangle$        $\langle T \rangle$

$$\begin{aligned} & 0 \times A_1/L_2 + \\ & 1 \times (A_2-A_1)/L_2 + \\ & 2 \times (L_1-A_2)/L_2 + \\ & 1 \times (L_2-L_1)/L_2 \end{aligned} \quad (T_1+T_2)/2$$

## M/M/1平衡性能验算



## 结论

- |     |   |         |
|-----|---|---------|
| (1) | $N = \frac{\lambda}{\mu - \lambda}$         | 平均逗留用户数 |
| (2) | $T = N / \lambda = \frac{1}{\mu - \lambda}$ | 平均逗留延时  |
| (3) | $W = T - 1 / \mu$                           | 平均等待延时  |
| (4) | $N_Q = \lambda W = \frac{\rho^2}{1 - \rho}$ | 平均等待用户数 |
- $\frac{1}{\mu}$  平均服务时间       $\rho = \lambda / \mu$  利用率因子

$$P V / T = const. \quad \text{热力学宏观规律}$$

## 形式定义与稳定性条件

$N(t)$  = 逗留在系统中的用户数

$\alpha(t)$  = 进入到系统的累计用户数

$T_i$  = 第  $i$  用户逗留在系统的时长

$N = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t N(t) dt$  系统中用户数是限的

$\lambda = \lim_{t \rightarrow \infty} \frac{\alpha(t)}{t}$  到达的用户数是有限的

$T = \lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$  用户逗留在系统的时间是有限的

Little's Law       $N = \lambda T$

## Little's Law

$\gamma(t)$  用户逗留时间之和, 即  $\sum_{i=1}^{\alpha(t)} T_i$

$\alpha(t)$  进入到系统的累计用户数

所以, 用户的平均逗留时间

$$T_t = \frac{\gamma(t)}{\alpha(t)}$$

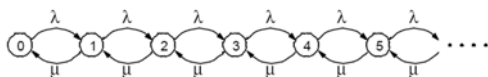
逗留在系统中的用户数

$$N_t = \frac{\gamma(t)}{t} = \frac{\alpha(t)}{t} T_t = \lambda_t T_t.$$

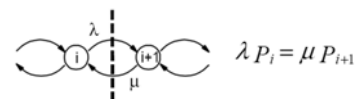
$$N = \lambda T$$

如果已知  $N$ , 可求得  $T$   
如果已知  $T$ , 可求得  $N$

## 系统状态与状态迁移



$$P_i = P\{\text{system in state } i\}$$



$$\lambda P_0 = \mu P_1$$

$$\lambda P_1 = \mu P_2$$

$$\lambda P_2 = \mu P_3$$

⋮

## 所有概率之和为1

$$P_1 = \frac{\lambda}{\mu} P_0 = \rho P_0$$

$$P_2 = \frac{\lambda}{\mu} P_1 = \rho(\rho P_0) = \rho^2 P_0$$

⋮

$$P_k = \rho^k P_0$$

$$\sum_{k=0}^{\infty} P_k = \sum_{k=0}^{\infty} \rho^k P_0 = 1$$

$$\frac{1}{1 - \rho} P_0 = 1 \Rightarrow P_0 = 1 - \rho.$$

□ That is,  $P_k = \rho^k (1 - \rho)$

□ Note that  $\rho$  must be less than 1, or else the system is unstable.

## 平均用户数

$$P_k = \rho^k (1 - \rho)$$

$$N = \sum_{k=0}^{\infty} k P_k = (1 - \rho) \sum_{k=0}^{\infty} k \rho^k = ?$$

$$I(x) = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \dots$$

$$x I(x) = \sum_{k=0}^{\infty} x^{k+1} = \sum_{k=0}^{\infty} x^k - 1$$

$$\Rightarrow I(x) = \frac{1}{1-x}$$

$$\frac{dI(x)}{dx} = \frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} k x^{k-1} = \sum_{k=0}^{\infty} (k+1) x^k$$

$$\frac{dI(x)}{dx} - I(x) = \sum_{k=0}^{\infty} k x^k$$

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

$$\Downarrow$$

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

□ Average delay per customer (time in queue plus service time):

$$T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda}$$

□ Average waiting time in queue:

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

□ Average number of customers in queue:

$$N_Q = \lambda W = \frac{\rho^2}{1 - \rho}$$

## LP中除法转加法的近似处理

$$T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda}$$

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

$$= \frac{1}{\mu} (1 + \rho + \rho^2 + \dots)$$

$$\approx \frac{1}{\mu} (1 + \rho) = T'$$

$$\lambda = 0.75\mu \quad T = 1.33/\mu$$

$$\lambda = 0.1\mu \quad T' = 1.25/\mu$$

$$\lambda = 0.1\mu \quad T = 1.11/\mu$$

$$\lambda = 0.1\mu \quad T' = 1.1/\mu$$

Prob.  $(1 - \rho) \ll 1$

## M/M/1应用示例

## 基本参数

- ▶ 24台计算机各自独立，平均每秒发出48个分组
- ▶ 分组长度符合负指数分布，平均为125Bytes
- ▶ 占用 T1(1.544Mb/s) 传输线路
  - 方案一：24台计算机，按TDM，各占1个时隙(8bit)
  - 方案二：24台计算机，按STDM，占24个时隙(192bit)

平均分组长度  $\langle L \rangle = 125 \times 8 = 1 \text{ kbit}$   
 平均分组发送速率  $\mu = 64 \text{ kps} / \langle L \rangle = 64$

## 方案一

□ The system can be considered as 24 M/M/1 queues:



We have  $\rho = \frac{48}{64} = 75\%$

$$N = \frac{0.75}{1 - 0.75} = 3$$

$$T = \frac{1}{64 - 48} = \frac{1}{24} \approx 42 \text{ msec}$$

$$64 - 48 = 16$$

$$1/16 = 0.0625$$



方案二

- The (aggregated) arrival rate is  $24 \times 48 = 1152$ .
- The service rate is  $24 \times 64 = 1536$ .
- We have

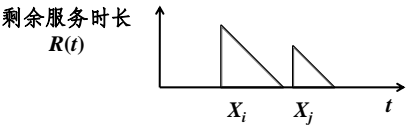
$\rho = 48 / 64 = 75\%$   
 $T = \frac{1}{1536 - 1152} = \frac{1}{384} \approx 2.4 \text{ msec}$  ~~X~~ 2.6 ms

STDM相比TDM，快了  $384 / 16 = 24$ 倍！  
Prob. 平均分组长为1500 Bytes？

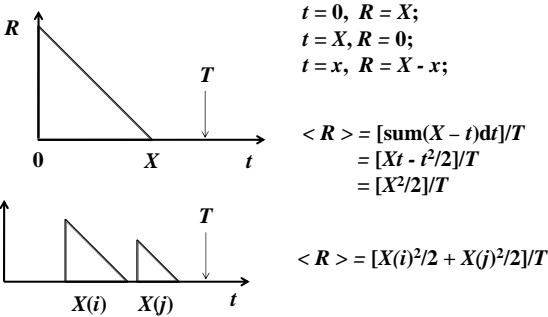
第八章 排队理论基础

- 8.1 排队模型
- 8.2 M/M/1系统
- 8.3 M/G/1系统
- 8.4 M/M/m系统
- 8.5 排队网络

M/G/1的用户服务过程



服务员的剩余服务时长

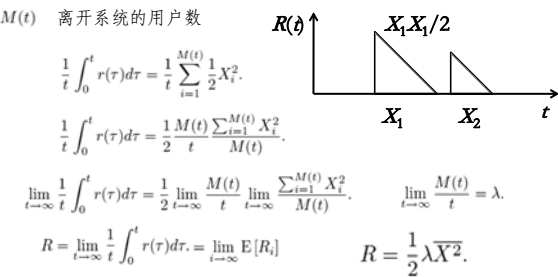


定义和结论

$E[X] = 1/\mu$  平均服务时长 (一阶矩)  
 $E[X^2]$  服务时长二阶矩  
 $X_i$  第  $i$  用户服务时长

$W = \frac{\lambda E[X^2]}{2(1 - \rho)}$   
 $N_Q = \lambda W = \frac{\lambda^2 E[X^2]}{2(1 - \rho)}$

剩余服务时长的平均



用户等待时长

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j.$$

$$N_i: \text{ 第 } i \text{ 用户进入队列时已有用户数}$$

$$E[W_i] = E[R_i] + E\left[\sum_{j=i-N_i}^{i-1} E[X_j|N_i]\right] = E[R_i] + \overline{X}E[N_i].$$

$$W = R + \frac{1}{\mu} N_Q$$

$$N_Q = \lambda W,$$

$$W = R + \frac{\lambda}{\mu} W.$$

$$R = \frac{1}{2} \lambda \overline{X}^2.$$

$$W = \frac{\lambda \overline{X}^2}{2(1-\rho)}$$

随机变量之和的平均

**Proposition: Sum of a Random Number of Random Variables**

- $N$ : random variable taking values  $0,1,2,\dots$ , with mean  $E[N]$
- $X_1, X_2, \dots, X_N$ : iid random variables with common mean  $E[X]$

Then:  $E[X_1 + \dots + X_N] = E[X] \cdot E[N]$

**Proof:** Given that  $N=n$  the expected value of the sum is

$$E\left[\sum_{j=1}^N X_j \mid N = n\right] = E\left[\sum_{j=1}^n X_j\right] = \sum_{j=1}^n E[X_j] = nE[X]$$

Then:

$$\begin{aligned} E\left[\sum_{j=1}^N X_j\right] &= \sum_{n=1}^{\infty} E\left[\sum_{j=1}^N X_j \mid N = n\right] \times P\{N = n\} = \sum_{n=1}^{\infty} nE[X] \cdot P\{N = n\} \\ &= E[X] \sum_{n=1}^{\infty} nP\{N = n\} = E[X]E[N] \end{aligned}$$

M/G/1的退化(Degenerate)

$$W = \frac{\lambda \overline{X}^2}{2(1-\rho)}$$

$$\text{M/D/1} \quad X_i = \frac{1}{\mu} \text{ and therefore } \overline{X}^2 = \frac{1}{\mu^2}$$

$$W = \frac{\rho}{2\mu(1-\rho)}.$$

$$\text{M/M/1} \quad \overline{X}^2 = 2/\mu^2$$

$$W = \frac{\rho}{\mu(1-\rho)}.$$

M/M/1的结论

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

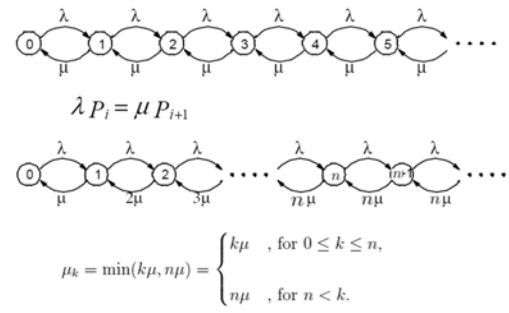
第八章 排队理论基础

- 8.1 排队模型
- 8.2 M/M/1系统
- 8.3 M/G/1系统
- 8.4 M/M/n系统
- 8.5 排队网络

M/M/n 排队系统



状态及平衡关系



## 归一化约束求解 $P_0$

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, \quad a = \frac{\lambda}{n\mu} = \frac{\rho}{n} < 1$$

$$P_k = P_0 \prod_{i=0}^{n-1} \frac{\lambda}{(i+1)\mu} \prod_{j=n}^{k-1} \frac{\lambda}{n\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{n! n^{k-n}}, \quad P_k = \begin{cases} P_0 \frac{\rho^k}{k!}, & \text{for } k \leq n, \\ P_0 \frac{\rho^k n^n}{n!}, & \text{for } k > n, \end{cases}$$

$$P_0 = \left(1 + \sum_{k=1}^{n-1} \frac{\rho^k}{k!} + \sum_{k=n}^{\infty} \frac{\rho^k}{n!} \frac{1}{n^{k-n}}\right)^{-1} = \left(\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n}{n!} \frac{1}{1-a}\right)^{-1}$$

## 用户必须等待的情况

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, \quad a = \frac{\lambda}{n\mu} = \frac{\rho}{n} < 1$$

$$P_k = P_0 \prod_{i=0}^{n-1} \frac{\lambda}{(i+1)\mu} \prod_{j=n}^{k-1} \frac{\lambda}{n\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{n! n^{k-n}}, \quad P_k = \begin{cases} P_0 \frac{\rho^k}{k!}, & \text{for } k \leq n, \\ P_0 \frac{\rho^k n^n}{n!}, & \text{for } k > n, \end{cases}$$


$$P_0 = \left(1 + \sum_{k=1}^{n-1} \frac{\rho^k}{k!} + \sum_{k=n}^{\infty} \frac{\rho^k}{n!} \frac{1}{n^{k-n}}\right)^{-1} = \left(\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n}{n!} \frac{1}{1-a}\right)^{-1}$$

$$P(\text{waiting}) = \sum_{k=n}^{\infty} P_k = \sum_{k=n}^{\infty} P_0 \frac{\rho^k}{n!} \frac{1}{n^{k-n}}$$

## 必须等待的概率

$$P(\text{waiting}) = \sum_{k=n}^{\infty} P_k = \sum_{k=n}^{\infty} P_0 \frac{\rho^k}{n!} \frac{1}{n^{k-n}}$$

$$= \frac{\rho^n}{n!} \frac{1}{1-a} = \frac{\rho^n}{n!} \frac{n}{n-\rho} = C(n, \rho)$$

$$= \frac{\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n}{n!} \frac{1}{1-a}}{\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n}{n!} \frac{n}{n-\rho}} = \text{Erlang's C formula}$$


## 无缓冲(M/M/n/n)的呼损概率

$$\lambda_k = \begin{cases} \lambda, & \text{if } k < n, \\ 0, & \text{if } k \geq n, \end{cases} \quad \mu_k = k\mu, \quad k = 1, 2, \dots, n.$$

$$P_k = \begin{cases} P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, & \text{if } k \leq n, \\ 0, & \text{if } k > n. \end{cases}$$

$$P_0 = \left(\sum_{k=0}^n \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\right)^{-1}$$

$$P_k = \frac{\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}{\sum_{i=0}^n \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}} = \frac{\frac{\rho^k}{k!}}{\sum_{i=0}^n \frac{\rho^i}{i!}}$$

$$P_n = \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^n \frac{\rho^i}{i!}} = B(n, \rho)$$

$$\text{Erlang's B-formula}$$



$$B(1, \rho) = \rho / (1 + \rho)$$

$$B(2, \rho) = ?$$

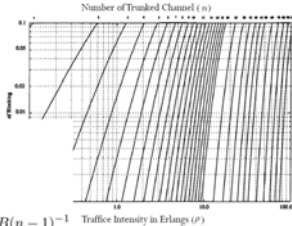
## 爱尔兰B公式的计算

- 在线计算
- 查图/表
- 迭代算法

$$B(n, \rho)^{-1} = \frac{\sum_{k=0}^n \rho^k / k!}{\rho^n / n!}$$

$$B(n)^{-1} = \frac{\rho^n / n! + \sum_{k=0}^{n-1} \rho^k / k!}{\rho^n / n!} = 1 + (n/\rho) B(n-1)^{-1}$$

$$B(n) = \frac{B(n-1)}{B(n-1) + n/\rho}$$

$$B(0) = 1$$


## POTS设计示例

- 忙时呼叫数(BHCA, CAPS)
- 接续时长
- 98%接通率, 或2%呼损

Johan Jensen 1908  
(Copenhagen 电话公司)  
安排多少接线员最合适?

百度百科: BHCA

单位时间内处理机用于呼叫处理的时间开销为

$$t = a + bN$$

信令处理时长 32ms

a: 固有开销

b: 处理一次呼叫的平均开销 (非固有开销)

N: 单位时间内所处理的呼叫总数, 即处理能力值 (BHCA)

t=0.85, a=0.29, b=32ms, 求其BHCA为多少?

$$N = 63000 \text{次/小时} \quad \rho = (1/N) / (1/b) = (32 \times 10^{-3}) / (63000 / 3600) < 0.002$$

呼叫完成时长 32 s



人工接续的求解示例

$$P_n = \frac{\frac{\rho^n}{n!}}{\sum_{k=0}^n \frac{\rho^k}{k!}} = B(n, \rho)$$

Johan Jensen1908  
(Copenhagen 电话公司)  
安排多少接线员最合适?

rho = (1/N)/(1/b) = (32\*E-3)/(63000/3600)<0.002  
B(1,rpo) = 0.002/(1+0.002)  
= 0.001996 < 2%  
rho = (1/N)/(1/b) = (32)/(63000/3600)=1.83  
B(1,rho) = 1.83/(1+1.83) = 0.65  
B(2,rho) = 0.65/(0.65+2/1.83) = 0.37  
B(3,rho) = 0.37/(0.37+3/1.83) = 0.18  
B(4,rho) = 0.18/(0.18+4/1.83) = 0.08  
B(5,rho) = 0.08/(0.08+5/1.83) = 0.028  
B(6,rho) = 0.028/(0.028+6/1.83) = 0.0085

答：需要安排6名接线员

第八章 排队理论基础

- 8.1 排队模型
- 8.2 M/M/1系统
- 8.3 M/G/1系统
- 8.4 M/M/n系统
- 8.5 排队网络

Jackson’s理论

由 k 个M/M/1排队系统任意组成的网络

$$P(n_1, n_2, ..., n_k) = P_1(n_1) P_2(n_2) ... P_k(n_k),$$

where

$$P_j(n_j) = \rho_j^{n_j} (1 - \rho_j) .$$

多任务操作系统的任务处理时间

$$P(i, j) = \rho_1^i (1 - \rho_1) \rho_2^j (1 - \rho_2) \quad \rho_1 = \lambda_1 / \mu_1 \quad \rho_2 = \lambda_2 / \mu_2$$
$$N = N_1 + N_2$$
$$N_1 = \frac{\rho_1}{1 - \rho_1}, \quad N_2 = \frac{\rho_2}{1 - \rho_2} \quad T = \frac{N}{\lambda} = \frac{\rho_1}{\lambda(1 - \rho_1)} + \frac{\rho_2}{\lambda(1 - \rho_2)}$$

第八章 排队理论基础思考题

- 8.1 32台计算机，采用TDM或STDM方式通过E1传输平均长1500字节的分组，计算机平均每秒产生48个分组，估算分组排队的平均时延。
- 8.2 求解M/G/1队列等待用户数的突发增长条件，依此分析M/M/1与M/D/1的队列长度差别。