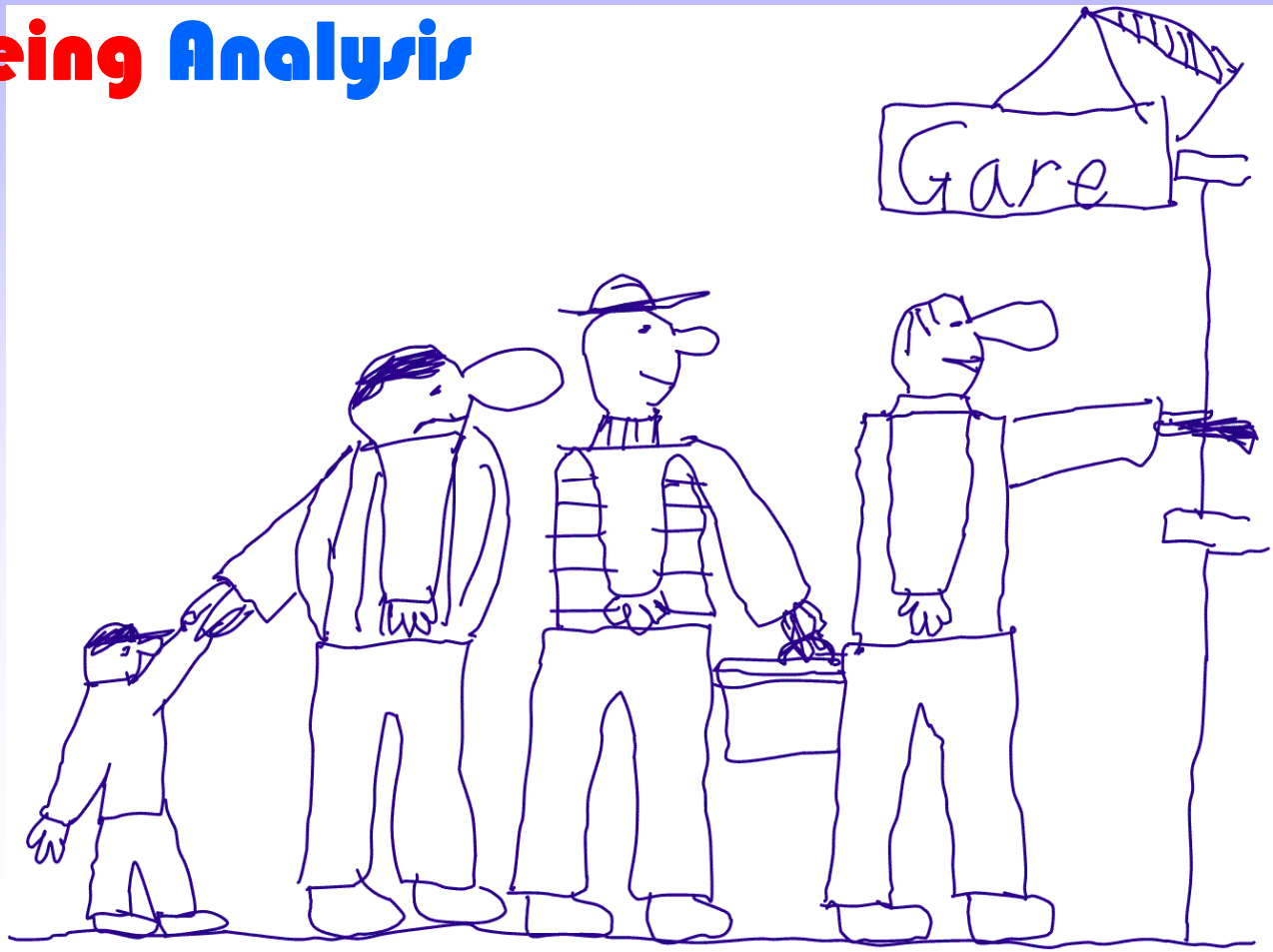
# Ch7. 网络性能建模与分析

## 7.1 Queueing Analysis



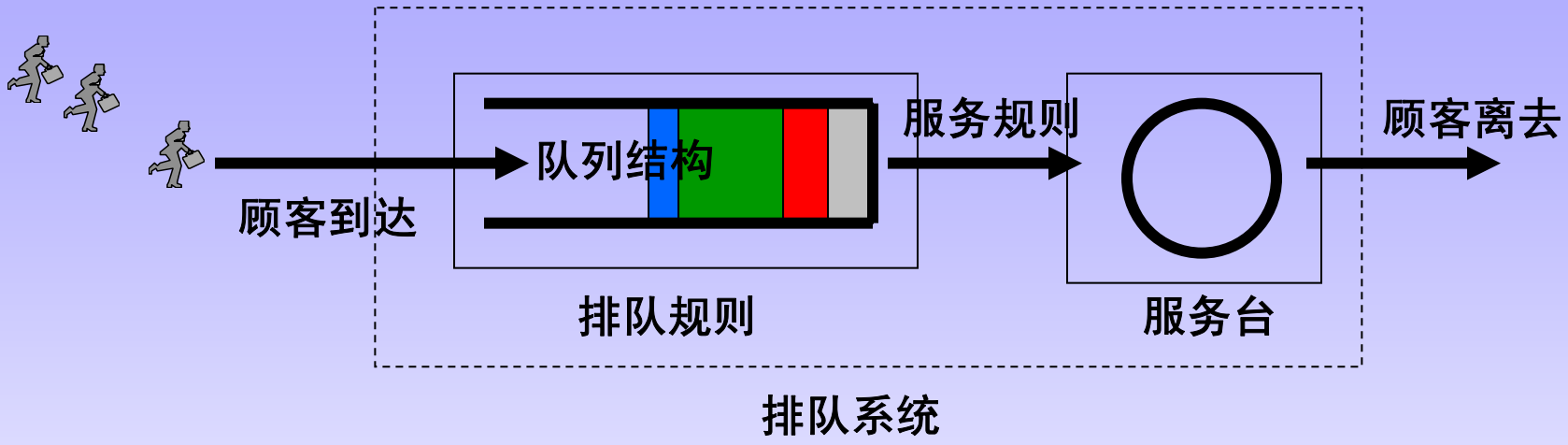
# 7.1.1 基本概念

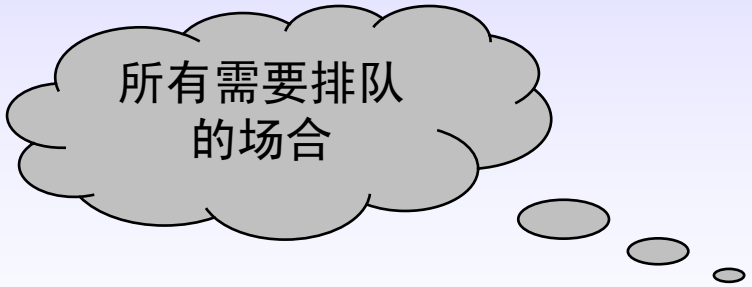
- 排队论或称**随机服务**系统理论
  - 为接受某种服务而排队等待。
  - 抽象出物理模型，建立数学论。
  - 性能评价中占有相当重要地位。

- 理论**成熟**，应用**广泛**
  - 已在电话交换网、公路、铁路、航空运输、工程管理、公共服务、货物存储和生产流水线过程等方面得到了广泛的应用。
  - **通信流量**研究的基础理论，定量研究往往借助排对论。

- 解决三大问题
  - **推断**：某排队系统属于**何种**模型？由此**分析**其行为，**评价**性能
  - **性态**：排队系统**概率**规律，**队长、等待时间和忙期**分布，瞬/稳态
  - **优化**：**静**/**动**态最优，即**设计**优化，**运营**优化

# 排队系统的一般表示



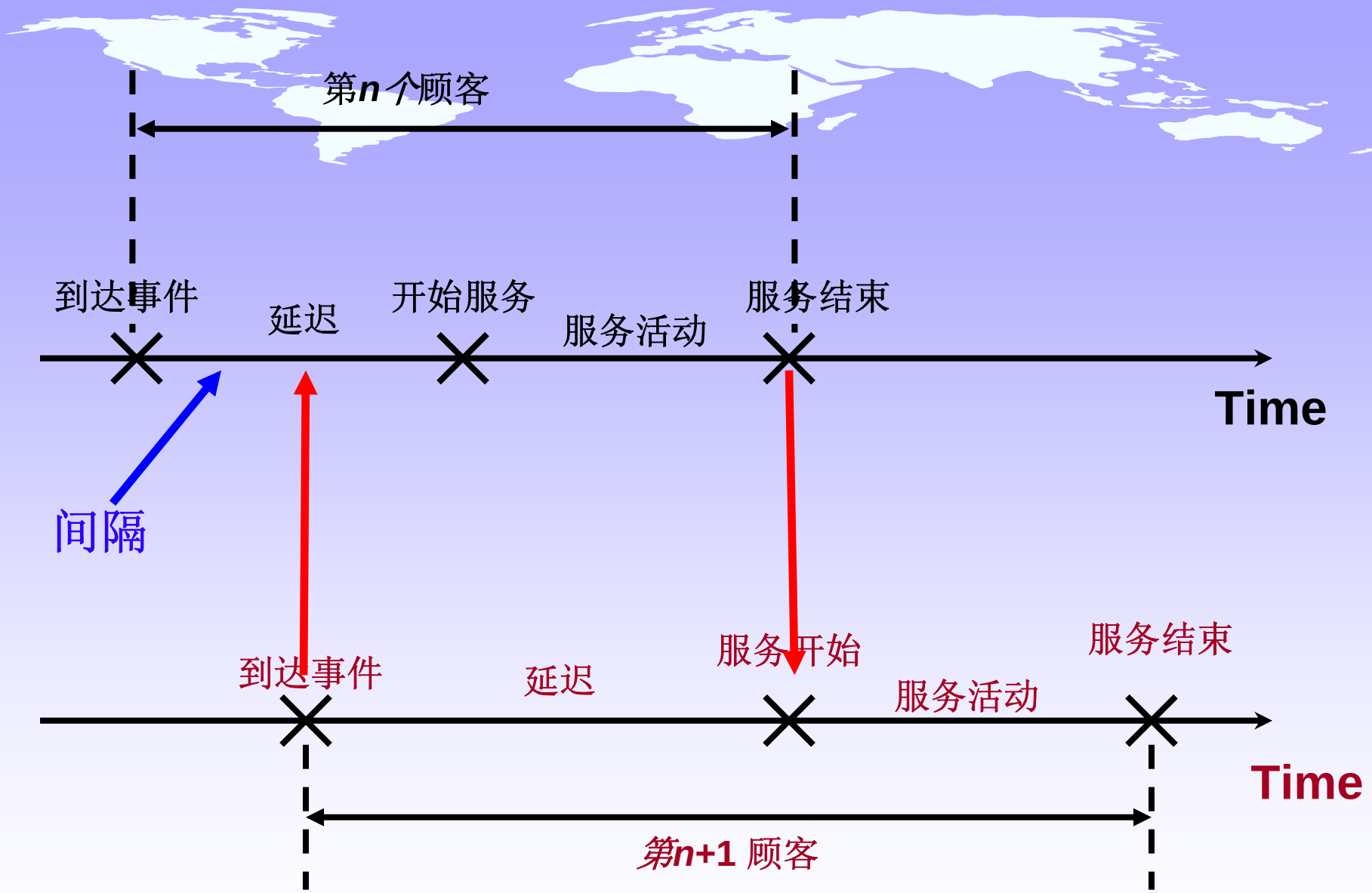

| 到达的顾客 | 服务内容 | 服务机构 |
|---|---|---|
| 不能运转的机器 | 修理 | 修理工 |
| 病人 | 诊断或手术 | 医生或手术台 |
| 电话呼叫 | 通话 | 交换台 |
| 文件稿 | 打字 | 打字员 |
| 达到机场飞机 | 降落 | 跑道 |
| 驶入港口货船 | 装卸货物 | 码头泊位 |
| 上游水入库 | 放水、调水 | 水闸 |
| **数据包达到入端口** | **转发到其出口** | **交换机/路由器** |
| 数据流达到总线入口 | 交换到目的设备 | 总线 |

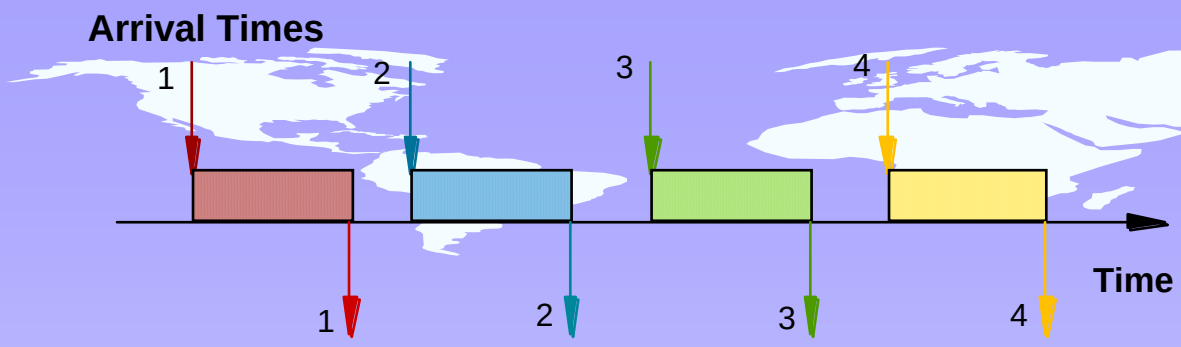
# 7.1.2 排队系统的组成和特征

- 输入过程（顾客按照怎样的规律到达）
  - 顾客**源**：有限（待修机器），或无限（上游水）
  - 到达**方式**：一个一个到达的，或成批的。只研究前者
  - 到达**间隔**：可是确定的（装配线），可是随机的（商店顾客）。符合一定的分布，称到达分布。
  - **到达时间**:假设**独立**同分布的随机变量。
  - 过程**平稳**，或时间齐次的，间隔分布参数（期望、方差）**与时间无关**
- 排队规则(顾客按照一定规则排队等待服务)
  - **等待制**：等待制：服务台忙则等待，损失制：忙则离开
  - **队列容量**：有限/无限。有限队列时顾客可能丢失
  - **服务协议**：决定排队方式，单队列、并联式多队列、串联式多队列及杂乱队列
- 服务机构（服务台数量,服务方式,服务时间分布）
  - **服务规律**:一个一个进行服务，每顾客服务时间长短不一。
  - **服务时间**：是随机变量，**独立同分布**。

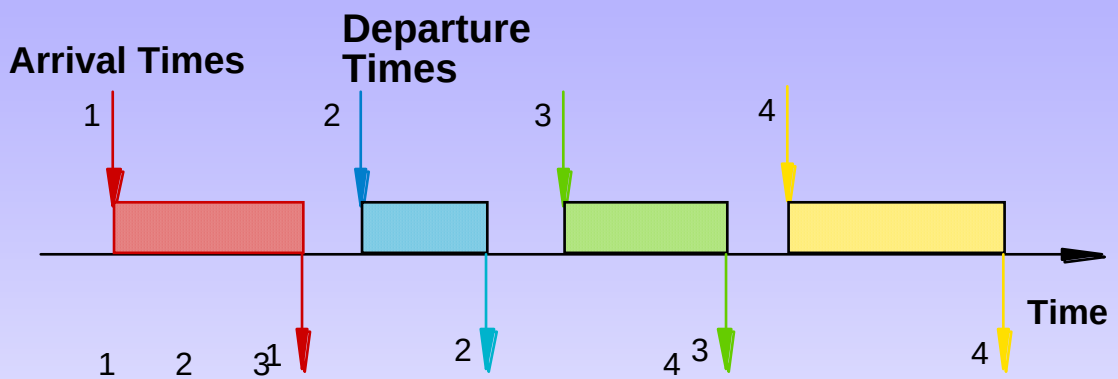
# 服务协议/服务员数量/质量

- 先来先服务：FCFS（first come first served）
- 后来先服务：LCFS（last come first served）堆栈形式
- 随机选择服务 ：从等待队列中随意抽取一个进行服务
- 优先服务和动态优先服务：如计算机中断优先级
  - 预先给到达顾客规定优先顺序类别，再按FCFS进行服务
  - 三类优先权：排队、中断、动态
- 共享服务（processor sharing）：网络服务系统
  - 服务能力平均分配给队列中所有顾客，无排队出现
  - 当顾数量增加时，只是顾客服务时间变长
- 无限服务员（infinite server）
  - 队列中的每个顾客接受完全相同的服务，似是唯一的一个顾客一样
  - 为每个顾客都可“克隆”出一个新服务员，且克隆数目无限。
- 服务员数量和质量
  - 单服务员系统、多服务员系统、无限服务员系统。

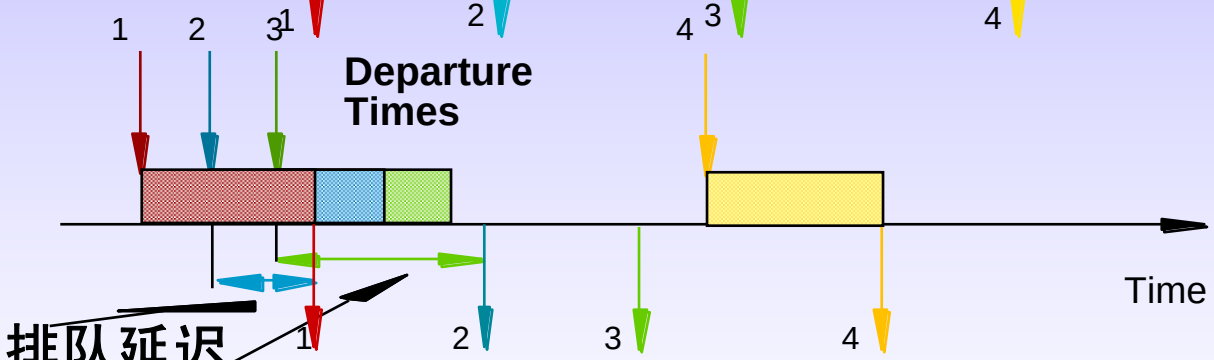

第n个顾客
到达事件
延迟
开始服务
服务活动
服务结束
Time
间隔
到达事件
延迟
服务开始
服务活动
服务结束
Time
第n+1 顾客



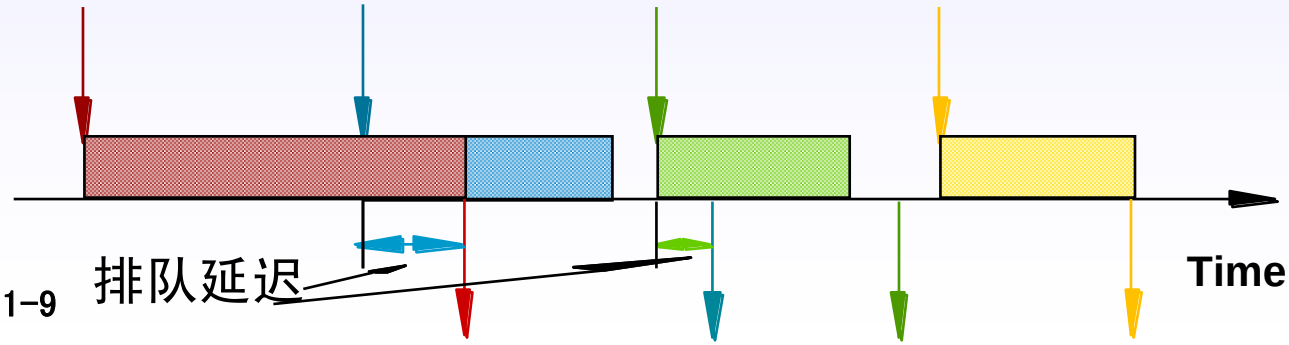

规则流－不需排队




不规则
但不冲突可发出

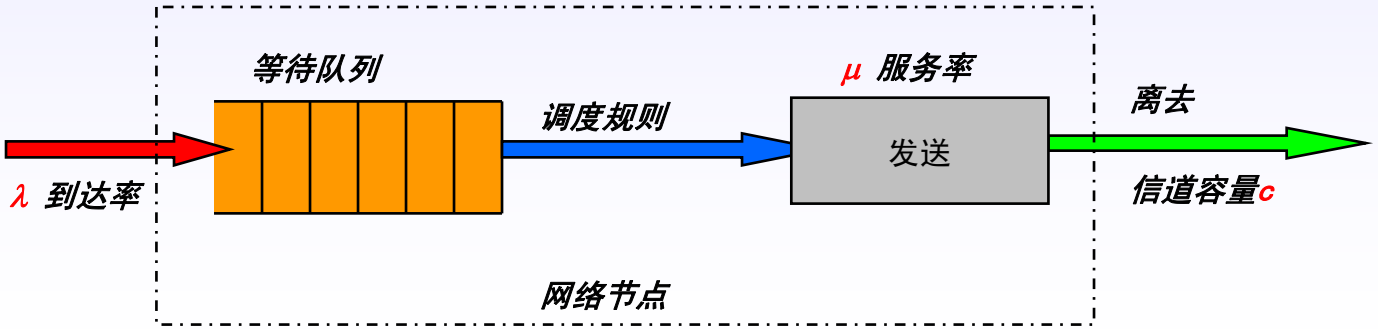



突发流量




不规则包长

- 定义
  - 报文平均到达率λ （mag/s）
  - 输出信道容量c(bit/s)
  - 报文长度随机，平均为1/μ(bit)
- 则有
  - 每个报文的平均发送时间[1/μ]/c =1/μc秒
  - 亦即信道发送能力为μc（mag/s）
- 稳态时
  - 平均输出率 = 平均输入率
  - λ = P[输出信道忙]×μc + P[输出信道闲]×0 = ρμc
  - ρ = λ/μc;  ρ = λ/μ，if c = 1
  - λ≤μc ；因为 0≤ ρ ≤1



# 7.1.3 排队系统的分类与标记

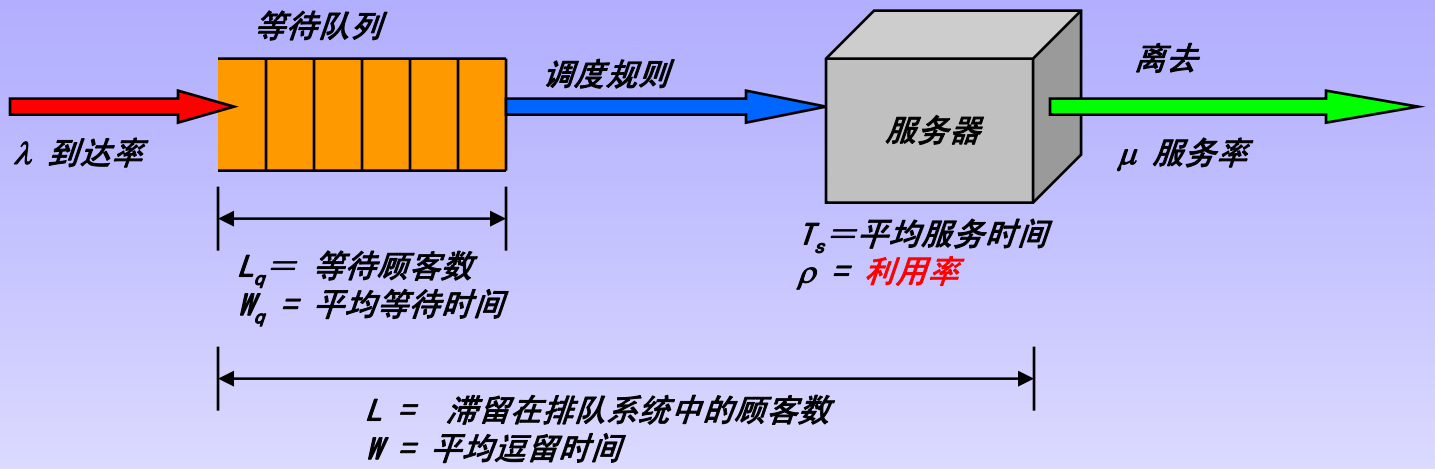
- D.G Kendall分类（主要三特征）
  - 相继顾客达到间隔时间的分布
  - 服务时间的分布
  - 服务台的个数
- Kendall记号： X／Y／Z／A／B/C：
  - X：顾客到达的规律；
  - Y：服务时间分布；
  - Z：服务员的数目；
  - A：系统容量限制N；
  - B：顾客源数目m。
  - C：服务规则（FCFS/LCFS）
- :X／Y／Z
  - 即指略去后三项的 X／Y／Z/∞/∞/FCFS

- 到达间隔与服务时间的分布类型
  - M：Markov首字母。泊松到达过程，指数分布；
  - G：General首字母。一般分布。
  - $E_k$：k- Erlang，k阶爱尔朗分布
  - H：超几何分布。
  - L：H项式分布。
- 服务规程典型的有：
  - FCFS：先来先服务；
  - LCFS：后来先服务；
  - RSS：随机选择服务；
  - PR：优先权服务；
  - GD：一般规约服务；
  - Ba：批量服务

# 7.1.4 排队问题的求解

- **系统**队长：系统中的总顾客数，其期望值**L**
- **排队**队长：排队等待服务的顾客数，其期望值**$L_q$**
  - $L = L_q$ + 正在被服务的顾客数
  - $L/L_q$越大，系统服务率越低，排队长龙，顾客厌烦
- **逗留**时间：顾客在系统中滞留的时间，其期望值**W**
- **等待**时间：顾客排队等待的时间，其期望值**$W_q$**
  - $W = W_q$ + 服务时间
- 忙期：顾客到达空闲服务台起，到服务台再次空闲的这段时间长度，即**服务台持续繁忙的时间长度**
  - 忙期关系到服务员的**工作强度**
  - 忙期中服务完成的平均顾客数是服务台**效率**的评价指标

# 队列参数图示




- λ ：每秒平均到达的顾客数（平均到达率）
- μ ：每秒平均服务的顾客数（平均离开率、服务率）
- Lq：平均等待队列长度
- Wq ：每个顾客的平均等待时间，包括没有排队的顾客
- ρ ：平均利用率，一段时间内测得
- L：系统中平均顾客数，包括正在被服务（若有）和等待（若有）的
- W = 平均等待时间 ＝ 平均等待时间＋平均服务时间 mean residence time

# 到达间隔分布和服务时间分布

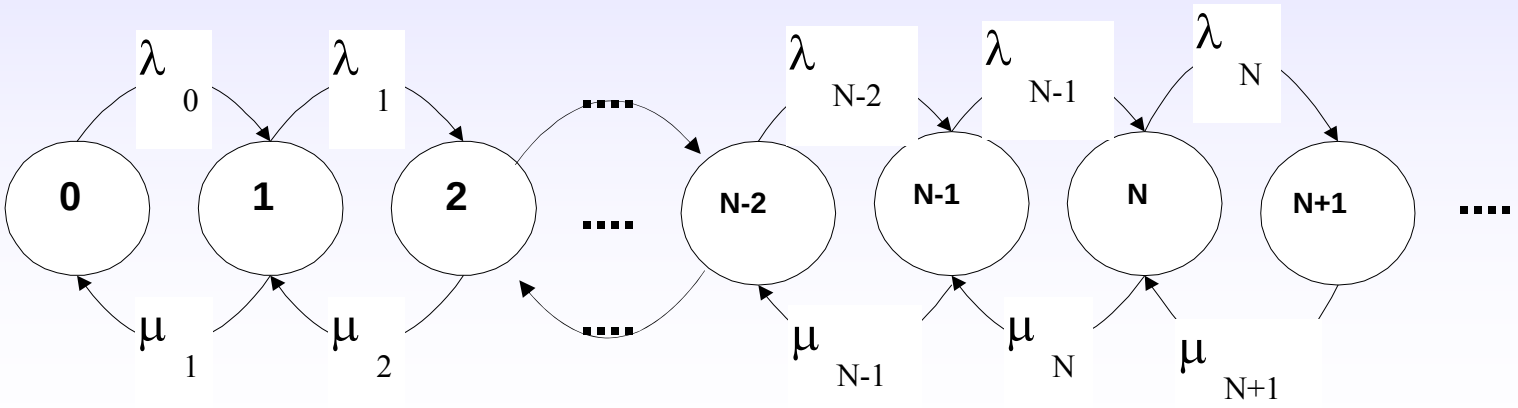
- 顾客泊松流到达（Poisson）：
  - t时间区间内到达n个顾客的概率
  - $P_n(t) = (\lambda t)^n e^{-\lambda t}/ n!$　　$n=0,1,2,\ldots,$　$t>0,$　$\lambda>0$
  - 其均值和方差分别是：$\lambda t$和$\lambda t$
- 服务时间的负指数分布：
  - $f(t)=\mu e^{-\mu x}$；$x\geqslant 0$;其分布函数：$F(x)=1-e^{-\mu x}$　　$x\geqslant 0$
  - 其数学期望等于标准差：$=1/\mu$
- 泊松流与负指数分布的关系：
  - 如果顾客以泊松到达；　则顾客到达的时间间隔$T_a$服从指数分布：
  - 因为对泊松流，在[0,t)区间至少一个顾客到达的概率是（代入n=0）

$$1-P_0(t)=P\{T_a<t\}=1-e^{-\lambda t}\ ,\ E[T_a]=1/\lambda$$

  - 平均到达时间间隔是到达速率的倒数。

# M/M/1生灭过程的状态转移

- M/M/1模型：求系统在任意时刻t的状态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
- 稳态情况下
  - $P_n(t)$与t无关；$P_n(t) = P_n$
  - 状态转移平衡方程：转入率 = 转出率
  - 状态0转移到状态1的转移率$\lambda_0 P_0$；
  - 状态1转移到状态0的转移率$\mu_1 P_1$



| 状态/State | 平衡方程：Rate In = Rate Out |
| --- | --- |
| 0 | $\mu_1 P_1 = \lambda_0 P_0$ |
| 1 | $\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$ |
| 2 | $\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$ |
| …. | …… |
| N-1 | $\lambda_{N-2} P_{N-2} + \mu_N P_N = (\lambda_{N-1} + \mu_{N-1}) P_{N-1}$ |
| N | $\lambda_{N-1} P_{N-1} + \mu_{N+1} P_{N+1} = (\lambda_N + \mu_N) P_N$ |
| …. | …… |

- 求稳态状态解

**State**

$0$: $P_1 = (\lambda_0 / \mu_1) P_0$

$1$: $P_2 = (\lambda_1 / \mu_2) P_1 + (\mu_1 P_1 - \lambda_0 P_0) / \mu_2$

$= (\lambda_1 / \mu_2) P_1 + (\mu_1 P_1 - \mu_1 P_1) / \mu_2$

$= (\lambda_1 / \mu_2) P_1$

$$= \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$$

***State***

***n-1:*** $P_n = (\lambda_{n-1} / \mu_n) P_{n-1} + (\mu_{n-1} P_{n-1} - \lambda_{n-2} P_{n-2}) / \mu_n$

$= (\lambda_{n-1} / \mu_n) P_{n-1} + (\mu_{n-1} P_{n-1} - \mu_{n-1} P_{n-1}) / \mu_n$

$= (\lambda_{n-1} / \mu_n) P_{n-1}$

$$= \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} P_0$$

***n:*** $P_{n+1} = (\lambda_n / \mu_{n+1}) P_n + (\mu_n P_n - \lambda_{n-1} P_{n-1}) / \mu_{n+1}$

$= (\lambda_n / \mu_{n+1}) P_n$

$$= \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} P_0$$

简化：

*Let* $C = (\lambda_{n-1}\ \lambda_{n-2} \ldots\ \lambda_0) / (\mu_n\ \mu_{n-1} \ldots\ \mu_1)$

*Then* $P_n = C_n P_0$, $n = 1, 2, \ldots$

$\rho = \lambda / \mu < 1$ (for steady-state)

$C_n = (\lambda / \mu)^n = \rho^n$, for n = 1, 2, ...

因为： $\Sigma^{\infty}_{n=0} P_n = 1 = P_0 + \Sigma^{\infty}_{n=1} P_n = P_0 + \Sigma^{\infty}_{n=1} C_n P_0$

$\Rightarrow [1 + \Sigma^{\infty}_{n=1} C_n] P_0 = 1$ ； （$c_0 = \rho^0 = 1$）

$\Rightarrow P_0 = 1 / (1 + \Sigma^{\infty}_{n=1} C_n)$

$= 1 / (1 + \Sigma^{\infty}_{n=1} \rho^n)$

$= 1 / (\rho^0 + \Sigma^{\infty}_{n=1} \rho^n)$ $(\rho^0 = 1)$

$$P_0 = 1 / (\Sigma^{\infty}_{n=0} \rho^n)$$

$$= (\Sigma^{\infty}_{n=0} \rho^n)^{-1}$$

$$= \{1 / (1 - \rho)\}^{-1}$$

$$P_0 = 1 - \rho$$

所以, $P_n = C_n P_0 = (1 - \rho)\rho^n$ , for n = 0, 1, 2,...

$$P_0 = 1 - \rho$$

**Note:**

1) $\Sigma^{n}_{i=0} x^i = (1 - x^{n+1}) / (1 - x)$, for any x,

2) $\Sigma^{\infty}_{n=0} x^n = 1 / (1 - x)$, if $|x| < 1$.

# M/M/1的系统队长

$L=\sum_{n=0}^{\infty} n(1-\rho)\rho^n = n\rho^n - n\rho^{n+1}$ 为变队长及其概率的积之和

$=(\rho+2\rho^2+3\rho^3+\ldots+n\rho^n)-(\rho^2+2\rho^3+3\rho^4\ldots+n\rho^{n+1})$

$=\rho+\rho^2+\rho^3+\ldots\ldots+\rho^n+n\rho^{n+1}$

$=\rho/(1-\rho)$

$$=(1-\rho)\rho\sum_{n=0}^{\infty}\frac{d}{d\rho}\rho^n=(1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right)$$

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

$$=\rho/(1-\rho)$$

$$or=\lambda/(\mu-\lambda)$$

# 排队队列长度期望值

$L_q = \Sigma^{\infty}_{n=1} (n - 1) P_n$ …(第n个顾客到来时的队长为n-1,发生概率为$P_n$)

$= \Sigma^{\infty}_{n=1} nP_n - \Sigma^{\infty}_{n=1} P_n$

$= \Sigma^{\infty}_{n=0} nP_n - (\Sigma^{\infty}_{n=0} P_n - P_0)$

$= L - 1(1 - P_0)$

$= \rho / (1 - \rho) - 1 + (1 - \rho)$

$= \rho^2 / (1 - \rho)$ or

$= \lambda^2 / \mu(\mu - \lambda)$

# 系统逗留和等待时间的期望值

- 系统逗留时间w在M/M/1下服从(μ - λ)的负指数分布

$$f(w) = (\mu - \lambda)e^{-(\mu-\lambda)w}; w \geq 0$$

$$F(w) = 1 - e^{-(\mu-\lambda)w}$$

$$w = E(w) = \frac{1}{\mu - \lambda}$$

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

- 或由ittle公式Wq=Lq/ λ = $\lambda^2$ / λ μ(μ - λ) = = ρ/ (μ - λ)

◆ little公式

$$L = \lambda w$$

$$L_q = \lambda w_q$$

$$w = w_q + \frac{1}{\mu}$$

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

◆ 排队论公式

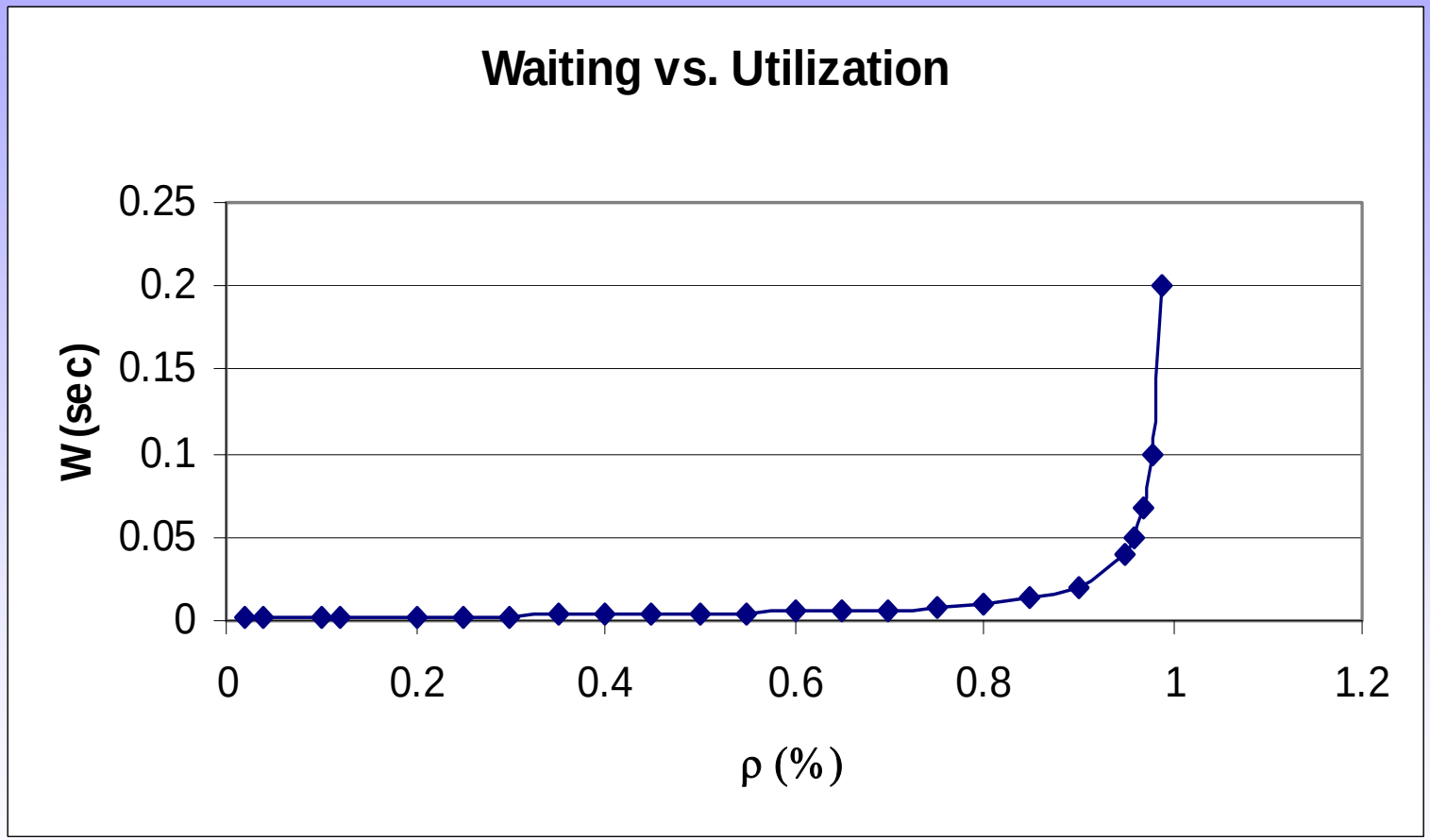
$$\rho = \frac{\lambda}{\mu}$$

$$L_q = \frac{\rho^2}{1-\rho}, \quad L = \frac{\rho}{1-\rho}$$

$$W_q = \frac{\rho}{\mu(1-\rho)}, \quad W = \frac{1}{\mu(1-\rho)}$$

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

# 等待时间与ρ的关系




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

# 7.1.5 排队问题求解之例

**服务排队问题的求解**

- 理想：1个3W服务器，
  - 假定平均1ms就有1个请求，服务器处理每个请求的时间也是精确到1ms。这样如果1000次/s(即1ms一次)速率到达，似乎能处理这样的负载？
  - 潜在理想假定：到一个请求，处理完一个，又到一个……
- 实际：假定平均到达率是1次/ms
  - 但非均匀到达，在给定的1ms间隔中，可能没有请求，也可能多个请求，但平均仍然是1次/ms.
- 问题：多于1个请求时
  - 把来不及处理的请求放在缓冲区（或进入等待服务队列）
  - 闲时就可以服务并清除缓冲区中的请求！
  - 缓存应该多大？队列可能多长？

表1：归一化到达率为0.5时队列的行为

| 时间 | 输入 | 输出 | 队列 | 时间 | 输入 | 输出 | 队列 |
|---|---|---|---|---|---|---|---|
| 00 | 0 | 0 | 0 | 26 | 190 | 190 | 0 |
| 01 | 88 | 88 | 0 | 27 | 500 | 500 | 0 |
| 02 | 796 | 796 | 0 | 28 | 96 | 96 | 0 |
| 03 | 1627 | 1000 | 627 | 29 | 943 | 943 | 0 |
| 04 | 51 | 678 | 0 | 30 | 105 | 105 | 0 |
| 05 | 34 | 34 | 0 | 31 | 183 | 183 | 0 |
| 06 | 96 | 966 | 0 | 32 | 447 | 447 | 0 |
| 07 | 714 | 714 | 0 | 33 | 542 | 542 | 0 |
| 08 | 1276 | 1000 | 276 | 34 | 166 | 166 | 0 |
| 09 | 494 | 769 | 0 | 35 | 165 | 165 | 0 |
| 10 | 933 | 933 | 0 | 36 | 490 | 490 | 0 |
| 11 | 107 | 107 | 0 | 37 | 510 | 510 | 0 |
| 12 | 241 | 241 | 0 | 38 | 877 | 877 | 0 |
| 13 | 16 | 16 | 0 | 39 | 37 | 37 | 0 |
| 14 | 671 | 671 | 0 | 40 | 163 | 163 | 0 |
| 15 | 643 | 643 | 0 | 41 | 104 | 104 | 0 |
| 16 | 812 | 812 | 0 | 42 | 42 | 42 | 0 |
| 17 | 262 | 262 | 0 | 43 | 291 | 291 | 0 |
| 18 | 218 | 218 | 0 | 44 | 645 | 645 | 0 |
| 19 | 1378 | 1000 | 378 | 45 | 363 | 363 | 0 |
| 20 | 507 | 885 | 0 | 46 | 134 | 134 | 0 |
| 21 | 15 | 15 | 0 | 47 | 920 | 920 | 0 |
| 22 | 820 | 820 | 0 | 48 | 1507 | 1000 | 507 |
| 23 | 1253 | 1000 | 253 | 49 | 598 | 1000 | 105 |
| 24 | 307 | 559 | 0 | 50 | 172 | 277 | 0 |
| 25 | 540 | 540 | 0 | 平均 | 499 | 499 | 43 |

- 假定表1中的到达率是500次/s;这是服务器处理能力的50%
- 50s后缓存平均有43个请求
- 请求队列中最多有627个

表2：归一化到达率为0.95时队列的行为

| 时间 | 输入 | 输出 | 队列 | 时间 | 输入 | 输出 | 队列 |
|---|---|---|---|---|---|---|---|
| 00 | 0 | 0 | 0 | 26 | 361 | 1000 | 3255 |
| 01 | 167 | 167 | 0 | 27 | 950 | 1000 | 3205 |
| 02 | 1512 | 1000 | 512 | 28 | 182 | 1000 | 2387 |
| 03 | 3091 | 1000 | 2604 | 29 | 1792 | 1000 | 3179 |
| 04 | 97 | 1000 | 1701 | 30 | 200 | 1000 | 2378 |
| 05 | 65 | 1000 | 765 | 31 | 348 | 1000 | 1726 |
| 06 | 1835 | 1000 | 1601 | 32 | 849 | 1000 | 1575 |
| 07 | 1357 | 1000 | 1957 | 33 | 1030 | 1000 | 1605 |
| 08 | 2424 | 1000 | 3382 | 34 | 315 | 1000 | 921 |
| 09 | 939 | 1000 | 3320 | 35 | 314 | 1000 | 234 |
| 10 | 1773 | 1000 | 4093 | 36 | 931 | 1000 | 165 |
| 11 | 203 | 1000 | 3296 | 37 | 969 | 1000 | 134 |
| 12 | 458 | 1000 | 2754 | 38 | 1666 | 1000 | 800 |
| 13 | 30 | 1000 | 1784 | 39 | 70 | 871 | 0 |
| 14 | 1275 | 1000 | 2059 | 40 | 310 | 310 | 0 |
| 15 | 1222 | 1000 | 2281 | 41 | 198 | 198 | 0 |
| 16 | 1543 | 1000 | 2824 | 42 | 80 | 80 | 0 |
| 17 | 498 | 1000 | 2322 | 43 | 553 | 553 | 0 |
| 18 | 414 | 1000 | 1736 | 44 | 1226 | 1000 | 226 |
| 19 | 2618 | 1000 | 3354 | 45 | 690 | 915 | 0 |
| 20 | 963 | 1000 | 3317 | 46 | 255 | 255 | 0 |
| 21 | 29 | 1000 | 2346 | 47 | 1748 | 1000 | 748 |
| 22 | 1558 | 1000 | 2904 | 48 | 2863 | 1000 | 2611 |
| 23 | 2581 | 1000 | 4285 | 49 | 1136 | 1000 | 2748 |
| 24 | 583 | 1000 | 3868 | 50 | 327 | 1000 | 2074 |
| 25 | 1026 | 1000 | 3894 | 平均 | 948 | 907 | 1859 |

- 假定到达率是950次/s；是服务器处理能力的95%
- 50s后缓存平均1859个请求
- 请求队列中最多有4093个
- 结果似乎出乎意外：
  - 到达率上升不到2倍，
  - 而缓存上升了40倍

表3：归一化到达率为0.99时队列的行为

| 时间 | 输入 | 输出 | 队列 | 时间 | 输入 | 输出 | 队列 |
|---|---|---|---|---|---|---|---|
| 00 | 0 | 0 | 0 | 26 | 376 | 1000 | 4445 |
| 01 | 174 | 174 | 0 | 27 | 990 | 1000 | 4435 |
| 02 | 1576 | 1000 | 576 | 28 | 190 | 1000 | 3625 |
| 03 | 3221 | 1000 | 2798 | 29 | 1867 | 1000 | 4492 |
| 04 | 101 | 1000 | 1899 | 30 | 208 | 1000 | 3700 |
| 05 | 67 | 1000 | 966 | 31 | 362 | 1000 | 3062 |
| 06 | 1913 | 1000 | 1879 | 32 | 885 | 1000 | 2947 |
| 07 | 1414 | 1000 | 2292 | 33 | 1073 | 1000 | 3020 |
| 08 | 2526 | 1000 | 3819 | 34 | 329 | 1000 | 2439 |
| 09 | 978 | 1000 | 3797 | 35 | 327 | 1000 | 1676 |
| 10 | 1847 | 1000 | 4644 | 36 | 970 | 1000 | 1646 |
| 11 | 212 | 1000 | 3856 | 37 | 1010 | 1000 | 1656 |
| 12 | 477 | 1000 | 3333 | 38 | 1736 | 1000 | 2392 |
| 13 | 32 | 1000 | 2365 | 39 | 73 | 1000 | 1465 |
| 14 | 1329 | 1000 | 2693 | 40 | 323 | 1000 | 788 |
| 15 | 1273 | 1000 | 2967 | 41 | 206 | 994 | 0 |
| 16 | 1608 | 1000 | 3574 | 42 | 83 | 83 | 0 |
| 17 | 519 | 1000 | 3093 | 43 | 576 | 576 | 0 |
| 18 | 432 | 1000 | 2525 | 44 | 1277 | 1000 | 277 |
| 19 | 2728 | 1000 | 4253 | 45 | 719 | 996 | 0 |
| 20 | 1004 | 1000 | 4257 | 46 | 265 | 265 | 0 |
| 21 | 30 | 1000 | 3287 | 47 | 1822 | 1000 | 822 |
| 22 | 1624 | 1000 | 3910 | 48 | 2984 | 1000 | 2805 |
| 23 | 2481 | 1000 | 5391 | 49 | 1184 | 1000 | 2990 |
| 24 | 608 | 1000 | 4999 | 50 | 341 | 1000 | 2330 |
| 25 | 1069 | 1000 | 5068 | 平均 | 988 | 943 | 2583 |

- 假定到达率略有增加
- 是990次/s;这是服务器处理能力的99%
- 50s后缓存平均有2583个请求
- 请求队列中最多有5068个

# 排队分析的任务和假设条件

- 队列分析的基本任务是：
  - **已知**如下输入信息（概率分布）：
    - 到达速率（ $\lambda$ ）
    - 服务时间（Ts）
  - **求出**如下输出信息（均值、标准差）：
    - 队列中等待顾客的数量（ $L_q$ ； $\sigma_{Lq}$ ）
    - 队列中顾客的等待时间（ $W_q$， $\sigma_{wq}$ ）
    - 系统中滞留顾客的数量（ L， $\sigma_L$ ）
    - 顾客在系统中的滞留时间（ W， $\sigma_W$ ）
- 排队论中的假设：
  - 到达**顾客**服从**泊松**分布，或到达**间隔**时间服从**指数**分布，这又等价于说到达顾客是随机的并彼此独立。我们几乎一直要作这一假定。
  - 在这个假定的条件下，仅仅知道到达速率和**服务时间的均值**和标准差就可以得到许多有用的结果。

**问题1：**

- 某修理店只有一个修理工，来修理的顾客到达过程为Poission流，平均4人/小时；修理时间服从负指数分布，平均需要6分钟。试求：

（1）修理店空闲的概率；　　（2）店内恰有三个顾客的概率

（3）店内至少有一个顾客的概率　　（4）在店内的平均顾客数

（5）每位顾客在店内的平均逗留时间　　（6）等待服务的平均顾客数

（7）每位顾客平均等待服务时间　　（8）顾客在店内等待时间超过10分钟的概率

**解答1：**

- M/M/1: $\lambda$ =4/60min 1/15; $\mu$ = 1/6min ; $\rho = \lambda / \mu = 2/5 = 0.4$
- (1) $P_0 = 1-\rho = 0.6$　　因为：( $P_n = (1-\rho)\ \rho^n$ )
- (2) $P_3 = (1-\rho)\ \rho^3 = 0.6*0.4^3 =$
- (3) $1- P_0 = 1-1+\rho = \rho$
- (4) $L = \rho/(1-\rho)$
- (5) $w = 1/\mu(1-\rho)$
- (6) $L_q = \rho^2/(1-\rho)$
- (7) $w_q = \rho/\mu(1-\rho)$
- (8) $P(T>10) = e^{-(\mu-\lambda)t}$

- 顾客在系统中的逗留时间T，服从参数为 $\mu-\lambda$ 的负指数分布，即：$P\{T>t\} = e^{-(\mu-\lambda)t}$
- $P\{T>10 \text{ min}\} = e^{-(1/16 - 1/15)10} = e^{1/24}$

**问题2：**一个书店平均每分钟有3个顾客到达，正常情况有48个顾客在书店中，求每一个顾客在商店花费的平均时间？

**问题3：**一条通信线路带宽是2000位/秒，该线路用来传8位一个的字符，故线路的最大速率是250字符/秒,来自应用要求是12000字符/分。求

- （1）等待被传输的平均字符数w
- （2）每个字符平均传输时间Tq

**问题4：**假定一个电话通话的持续时间平均3分钟，一个人等待电话平均最大可以忍耐3分钟，求可以支持的最大呼叫量？

(1) **解答2：** M/M/1: $\lambda$ = 3/min = 1/15; L = 48

- $\rho = \lambda/\mu = L/(1+L)$ $\mu = (49/48)\lambda = 3.08$; $W = 1/\mu(1-\rho) = 1/(3.08)(1-3/3.08) = 12.5$ min

(2) **解答3：** $\lambda$ =12000/min; $\mu$ = 250/(1/60)min=15000/min

- (1)$L_q = \rho^2/(1-\rho) = 0.8^2/(1-0.8) = 3.2$;
- (2):$W_q = \rho/\mu(1-\rho) = 0.8/15000*0.2 = (0.26/1000)60 = 0.156$min

(3) **解答4：**

- Ts = $1/\mu$ = 3 min , Max Wq = $\rho/\mu(1-\rho)$ = 3 min
- 求可支持的最大呼叫量即 $L = \rho/(1-\rho)$;
- $L = \rho/(1-\rho)$; 0.5/0.5 = 1 ;因为 $\rho$ = 1/2

**问题5：**

- 一局域网连有100台个人计算机，合一个存放支持查询的共用数据库服务器。服务器响应一次查询的平均时间是0.6s，估计标准差等于均值。最忙时间里，局域网内发生的查询达到20次/min；求
  - 若忽略线路开销，响应时间的平均值？
  - 若可接受的最长响应时间是1.5s （90%最长响应即可），报文负载增加多大的百分比才使影响时间达到最大值？
  - 若利用率上升20%，响应时间的增加使大于还是小于20%？

**解答5：忽略网络影响，M/M/1模型**

- $\rho = \lambda Ts = 20/min * 0.6s/(60s/min) = 0.2$
- $W = 1/\mu(1-\rho) = Ts/(1-\rho) = 0.6/(1-02) = 0.75\ s$；1.0s（$\rho=0.4$）
- $Max\ W\ y = Max\ W(90) = W*\ln(100/(100-y)) = W*\ln 10$
- $= Ts/(1-\rho)*2.3 = 1.5s$
- 求解 $\rho=0.08$；实际上，要使响应时间在90%时间内都小于1.5s，利用率就必须从20%降到8%
- 第三，求负载与响应的关系：因0.2的利用率处于曲线下部，响应时间增长很慢。故$\rho$从20%到40%，增加100%，W从0.75增长到1.0s，增长33%

**问题6**：局域网到路由器的通信量是5个包/s;平均包长144字节，包长指数分布，路由器到广域网的速率9600bps:

- 包到路由器的平均滞留时间？
- 路由器中平均有多少个包（等待和正传输的）？
- 同2，要求90%时间里路由器中的包数？
- 同2，95%时间里路由器中的包数？

**解答6：（计算百分位）**

- $\lambda$ =5个包/s;Ts=144*8/9600bps=0.12s
- $\rho = \lambda Ts = 5*0.12 = 0.6$;
- 平均滞留时间：$W = Ts/(1-\rho) = 0.3s$
- 平均滞留包数 $L = \rho/(1-\rho) = 1.5$个包
- 为得百分比，利用$P[R=N] = (1-\rho)\rho^n$；计算百分之y时间内的队列长度：$y/100 = \sum_{i=0}^{m_L(y)} (1-\rho)\rho^i = 1-\rho^{1+m_L(y)}$；（$0 \leqslant i \leqslant m_L(y)$）
- $m_L(y)$表示百分之y的时间里最多的包数，即$m_L(y)$是这样一个数，L 低于它的时间百分比是y。反过来，给定y，求$m_L(y)$，于是对等式两边取对数
- $m_L(y) = [\ \ln(1-y/100)/\ln\rho\ ] - 1$
- $m_L(y)$是分数，则取下一个较大的整数，若为负，则取0。本例$\rho = 0.6$；要求$m_L(90)$和$m_L(95)$
- $m_L(90) = 【\ln(1-0.9)/\ln0.6】-1 = 3.5 = 4$
- $m_L(95) = 【\ln(1-0.95)/\ln0.6】-1 = 4.8 = 5$
- 即若要设计一个到达95%的便准，必须提供能至少存储5个包的缓存。

# 计算满足条件x的百分位

- **问题5：求90%的响应时间小于1.5s？**

$m_x(y) \rightarrow x$小于$m_x(y)$的百分比是$y$

$$\Pr[\ W\ \le T\ ] = 1 - e^{-(1-\rho)t/T_s}$$

$$\frac{y}{100} = \Pr[\ W\ \le T\ ] = 1 - e^{-(1-\rho)m_W(y)/T_s}$$

$$1-\frac{y}{100} = e^{-(1-\rho)m_W(y)/T_s}$$

$$m_W(y)\frac{(1-\rho)}{T_S} = \ln(\frac{100}{100-y});\text{因为等待时间}W = \frac{T_S}{1-\rho} = \frac{1}{\mu(1-\rho)}$$

所以：$m_W(y) = W \times \ln(\frac{100}{100-y})$

$$m_W(90) = \frac{T_S}{1-\rho} \times \ln(10) = 0.75 \times 2.3 = 1.725s < 1.5s$$

# 计算满足条件x的百分位

- **问题6：95%时间里路由器中的包数？**

因为队长为 $N$ 的概率为： $\Pr[\ R \le N\ ] = \sum_{i=0}^{N} (1-\rho)\rho^{i}$

为计算百分之 $y$ 时间内的队列长度 ：

$$\frac{y}{100} = \sum_{i=0}^{m_r(y)} (1-\rho)\rho^{i} = 1-\rho^{1+m_r(y)};$$

所以： $$m_r(y) = \frac{\ln(1-\frac{y}{100})}{\ln\rho} - 1$$

故： $$m_r(90) = \frac{\ln(1-0.90)}{\ln(0.6)} - 1 = 3.5$$

而： $$m_r(95) = \frac{\ln(1-0.95)}{\ln(0.6)} - 1 = 4.8$$

## 问题7：

- 测得一网络路由器的平均到达率为125 pps（packets per second），路由器转发出每个包的平均时间是2 ms（毫秒），假设上述过程符合M/M/1模型，提示：

$$\text{排队队长为}n\text{的概率} = P_n = (1-\rho)\ \rho^n$$

  - 若路由器仅有13个包大小的队列缓冲区，求该队列缓冲区溢出概率？
  - 为保持路由器的丢包率在每100万个包丢一个包以下，则需要多大队列缓冲区？

## 解答7：

- 到达率 $\lambda$ = 125 pps
- 服务率 $\mu$ = 1/0.002 = 500 pps
- 路由器利用率 $\rho = \lambda / \mu = 0.25$
- 路由器中有n个包的概率 = $(1-\rho)\rho^n = 0.75(0.25)^n$
- 路由器中平均包数（平均队长L） = $\frac{\rho}{1-\rho} = \frac{0.25}{0.57} = 0.33$
- 缓冲区溢出的概率： P（路由器中大等于 13 包）

$$1-\sum_{i=0}^{12}(1-\rho)\rho^i = 1-\sum_{i=0}^{12}(\rho^i+\rho^{i+1})$$

$$= 1-(1+\rho+\rho^2+\ldots+\rho^{12}-\rho-\rho^2-\ldots-\rho^{12}-\rho^{13}) = \rho^{13}$$

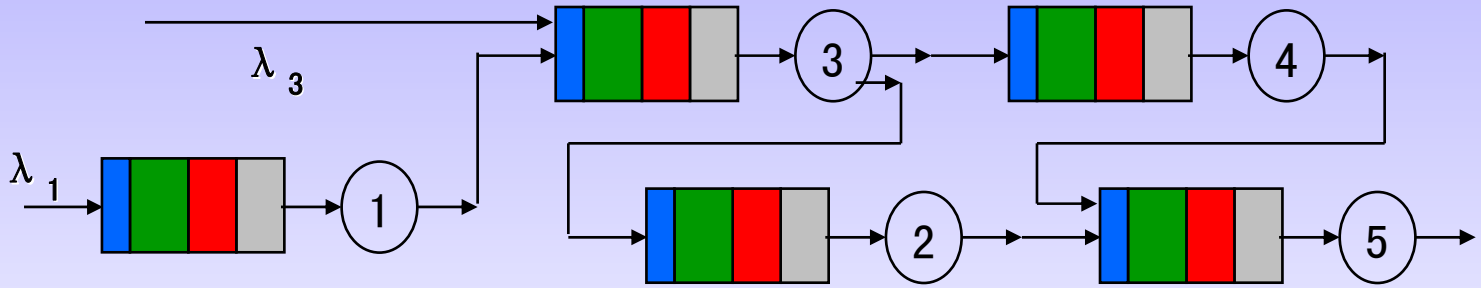
$= \rho^{13} = 0.25^{13} = 1.49\text{x}10^{-8}$ = 大约每百万15个包溢出。

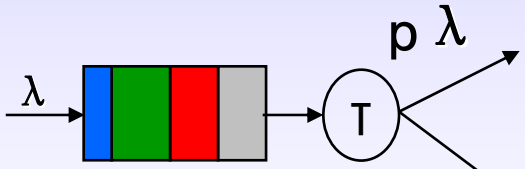
- 限制丢包概率小于$10^{-6}$，则有 $\rho^n \le 10^{-6}$

$$9.96 = n > \log(10^{-6}) / \log(0.25)$$

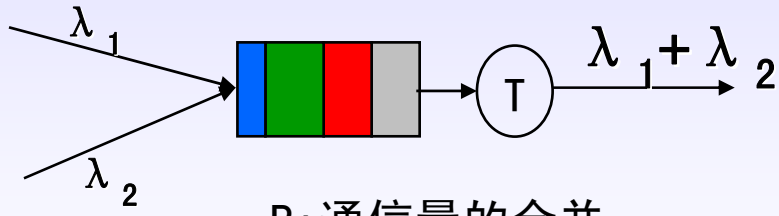
# 排队网络

- 排队网络的复杂化
  - 通信量的分开和合并
  - 队列组成一串

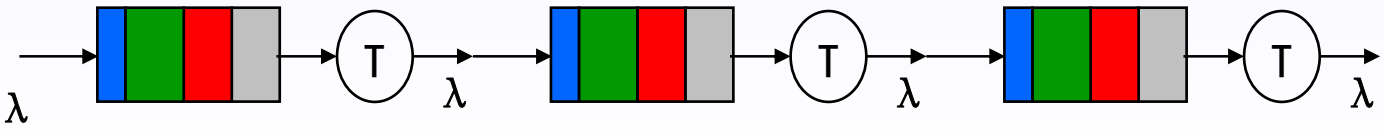



A:通信量的分开




B:通信量的合并




C:简单串联队列

# Jackson定理

- 3个假设条件：
  - 排队网络由m个节点组成，每个节点都具有独立的指数分布时间；
  - 从系统外到达系统中任何一个节点的对象都以泊松速率达到；
  - 一旦在一个节点上得到服务，对象要么立即以固定的概率达到另外的某个节点，要么离开系统。
- 定理：
  - 在这样一个排队网络中，每个节点都是一个独立的排队系统，其达到过程是泊松的，这个泊松过程则由分开、合并和串联排队原理确定。因而每个节点都可用独立于其它节点，用M/M/1或M/M/N模型来分析。
  - 所得结果可用普通统计方法组合在一起，每个节点上的平均时延可以加起来得到系统时延，但系统时延的高阶矩（如标准差），则不能通过这种方法得到。

# 分组交换网络的复杂网络排队

- 模型：N个源端和N个宿端由多跳传输链路连接起来：
  - 则网络总负载为：

$$\gamma = \sum_{j=1}^{N}\sum_{k=1}^{N}\gamma_{jk};\text{其中}$$

$\gamma =$ 网络总负载（每秒分组数$pps$）

$\gamma_{jk} =$ 源端$j$和宿端$k$之间的负载，

$N =$ 源端和宿端的总数

  - 分组从源到宿过程中可能会经过多条链路，中的内部负载会比供给的负载高：

$$\lambda = \sum_{i=1}^{L}\lambda_i;\text{其中}$$

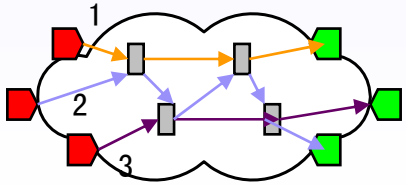
$\lambda =$ 网络上所有链路的总负载，

$\lambda_i =$ 链路$i$上的负载

$L =$ 链路总数

  - 给定路由算法，可通过负载 $\gamma_{jk}$ 确定各个链路上的负载 $\lambda_i$。对给定路由，一分组将经过的平均链路数为：

$$E[\text{一条路径中的链路个数 }] = \frac{\lambda}{\gamma}$$




$$\frac{\lambda}{\gamma} = \frac{1\times 3 + 2\times 5 + 3\times 3}{1+2+3} = 3.6\text{条}$$

- 目标：确定1个分组经过网络时所经历的评价时延；
  - 由Little公式，在每条链路上等待和正在被服务的平均数是：

$$r_i = \lambda_i T_{ri};$$

$r_i$是正在被服务的平均对象数

$T_{ri}$是有待确定的排队时延

  - 如果对每条链路上的平均服务对象求和，就可得到在网络的所有队列处等待的总的分组平均数：
  - 大系统也符合Little公式，故网络中等待和正被服务的分组数：

$$T\gamma = \sum_{i=1}^{L} \lambda_i T_{ri}; \text{或} T = \frac{1}{\gamma} \sum_{i=1}^{L} \lambda_i T_{ri}$$

$T$是每个分组经历的全部时延

  - 要确定每个队列i分别的时延$T_{ri}$；由每个队列是独立M/M/1可得：

$$T_{ri} = \frac{T_{si}}{1-\rho_i} = \frac{T_{si}}{1-\lambda_i T_{si}}$$

  - 链路i上的服务时间$T_{si}$是M与$R_i$之比;M是分组的平均长度（bit）；$R_i$是链路的数据率(bps)，于是：

$$T_{ri} = \frac{\frac{M}{R_i}}{1-\frac{M\lambda_i}{R_i}} = \frac{M}{R_i - M\lambda_i}$$

  - 分组经过网络的平均时延T:

$$T = \frac{1}{\gamma} \sum_{i=1}^{L} \frac{M}{R_i - M\lambda_i}$$

# Zipf's Law

- 对正数序列$f_1,f_2,\ldots f_s$,对1-s间的任意正整数i,j，如果有$i\times f_i=j\times f_j=C$（$C\neq 0$）,则称$f_1,f_2,\ldots f_s$服从Zipf's Law
- 显然若有数字序列$f_1,f_2,\ldots f_s$服从齐普夫定律，则$f_i=C/i$ $(1\leqslant i\leqslant s)$

# 排队论基本概念练习

1。指出下列排队系统中的顾客和服务员：

（1）自行车修理店；（2）按客户订单进行加工的加工车间

（3）机场起飞的客机 （4）十字路口红灯前的车辆

2。判断正误

（1）若到达排队系统的顾客人数服从泊松分布，则依次到达的两名顾客之间的间隔时间服从指数分布。

（2）在一个排队系统中，不管顾客到达和服务时间的情况如何，只要运行时间足够长的时间后，系统将进入稳定状态。

（3）在排队系统中，顾客等待时间的分布不受排队规则的影响。

# Thank you!