网络性能分析——排队分析

一、网络性能分析的目的和方法

㈠ 目的

- 对不同的网络设计策略、方案进行评价；
- 预测在给定输入负载下网络的性能；
- 对已经存在的网络，控制输入负载，从而得到需要的性能。

㈡ 方法

主要有两种：

1. 分析模型

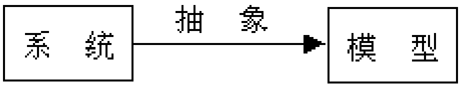
利用数学方法根据网络的基本特征构成模型，然后进行评价。

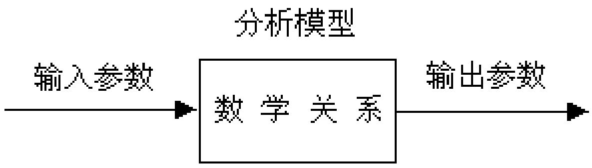
2. 模拟

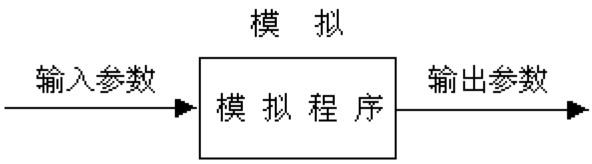
利用模拟程序根据网络的基本特征构成模型，然后进行评价。

3. 两者之间的关系

见下图:







【提示】

本补充介绍一些相关的数学知识。

二、相关概率论知识

㈠ 最简单流

设 N（t）表示在时间区间$[0,t)$内到达的顾客数（t>0），令 P_k（t1，t2）表示在时间$[t_1,t_2)$（t2>t1）内有 k 个顾客到达的概率，即：

P_k（t1，t2）＝P{N（t2）－N（t1）＝k}　（k=0、1、2、…，且 0≤t1<t2）

当 P_k（t1，t2）符合下列三个条件时，称顾客到达形成***<u>简单流</u>***:

条件一：

在不相交的时间区间内顾客到达数是相互独立的，称为无后效性。

条件二：

对于充分小的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：

$$P_1(t,t+\Delta t)=P\{N(t,t+\Delta t)=1\}=\lambda\Delta t+o(t)$$

其中，O（t），当 t→0 时是关于 t 的高阶无穷小量，常数λ>0 称为顾客平均到达率。

条件三：

对于充分小的Δt，在时间区间$[t, t+\Delta t)$内有 2 个以上顾客到达的概率极小，以致于可以忽略，即：

$$\sum_{j=2}^{\infty} P_j(t, t+\Delta t) = \sum_{j=2}^{\infty} P\{N(t, t+\Delta t) = j\} = o(t)$$

㈡ Poisson 过程（泊松过程）

1. 泊松过程

定理：假定有无穷个顾客，且顾客在时间间隔 t 内独立到达 k 个的概率为：

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad （k=0、1、2\cdots）$$

其中，λ是常数，为顾客平均到达率，则称这种到达为***泊松过程***。

2. 概率密度函数

可以分别求出服从 Poisson 分布的顾客相邻到达间隔时间 x 的概率密度函数、数学期望和标准方差（推导过程略）

⑴ 概率密度函数（负指数）

$$f(x) = \lambda e^{-\lambda x}$$

⑵ 数学期望

$$E[x] = \frac{1}{\lambda}$$

⑶ 标准方差

$$\sigma_x = \frac{1}{\lambda}$$

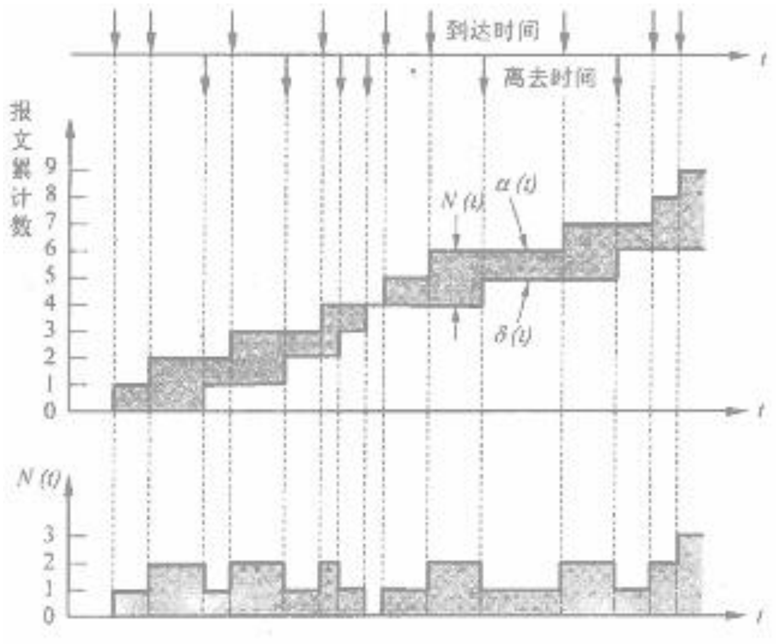
三、Little 定律（李特尔定律）

设在时间区间$[0,t)$内进入网络的报文数为α（t），离开网络的报文数为δ（t），存储在网络中的报文数N（t）为：

$$N(t) = \alpha(t) - \delta(t)$$

【提示】

典型的报文进入网络和离开网络的表示曲线*见下图*:



在稳定状态下累计进入网络和离开网络以及在网络中存储的报文数目

报文平均到达率为：

$$\lambda_t = \alpha(t)/t \tag{A-1}$$

所有报文在网络中经历时间为：[*即曲线α（t）和δ（t）之间的面积*]

$$\gamma(t) = \int_0^t N(x)dx$$

在时间区间$[0,t)$内网络中的平均报文数为：

$$N_t = \int_0^t N(x)dx / t = \gamma(t)/t \tag{A-2}$$

每一个报文在网络中所经历的平均时间为：

$$T_t = \gamma(t)/\alpha(t) \tag{A-3}$$

由（A-1）、（A-2）和（A-3）式可得：

$$N_t = \lambda_t \cdot T_t \qquad \text{(A-4)}$$

令：$\lambda = \lim\limits_{t\to\infty}\lambda_t$、$T = \lim\limits_{t\to\infty}T_t$ 和 $N = \lim\limits_{t\to\infty}N_t$，于是，（A-4）改写成：

$$\boxed{N = \lambda \cdot T} \qquad \text{(A-5)}$$

这就是 ***Little 定律***：在稳定状态下，存储在网络中的报文平均数 N，等于报文的平均到达率 λ 乘以这些报文在网络中经历的平均时间。

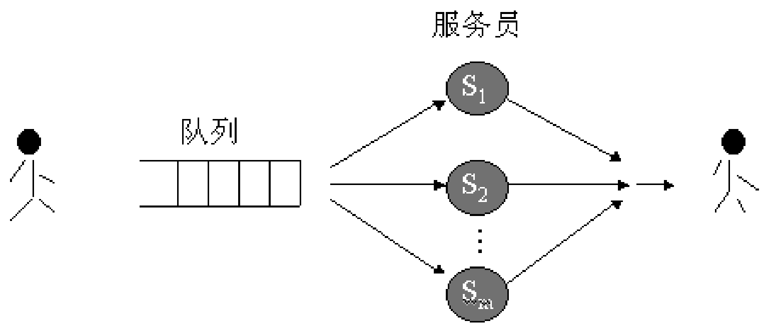
【提示】

网络边界可任意设定，但 N、λ 和 T 必须在同一个网络中。此外，报文输入或长度的规律如何均不影响 Little 定律成立。（这点很重要）

四、排队系统

㈠ 排队系统的基本模型

1. 排队系统模型要素



- 顾客总数是无限的；
- 顾客到达规律由到达时间间隔的概率密度函数描述；
- 服务员服务的规律由服务时间的概率密度函数描述；
- 服务员个数；
- 排队法则：通常有优先级、FIFO、最短（长）先服务等；
- 队列空间的大小，通常假定无穷大。

2. 排队系统的标识

用 A/B/m 标识某一排队系统，其中：

- A：为顾客到达时间间隔的概率密度（到达的规则）；
- B：为服务时间的概率密度（服务的规则）；
- m：为服务员个数。

【提示】对于 A、B 一般可取值为

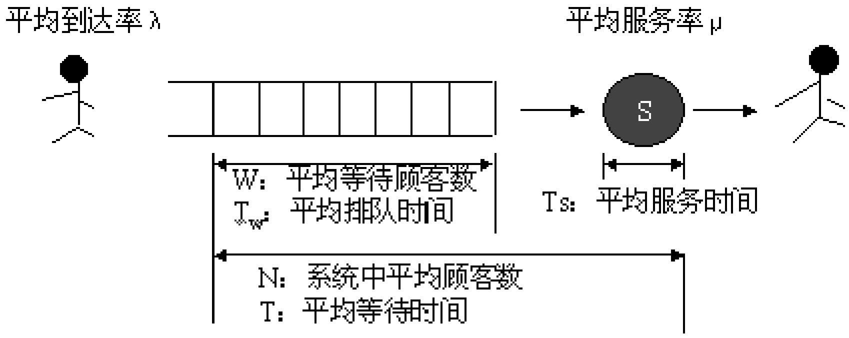
- M：表示指数分布（负指数概率密度）；
- D：表示不变的确定值；
- G：表示通用的，即任意概率密度。

㈡ M/M/1 排队系统

含义：到达规律是负指数概率密度，服务规则也是负指数概率密度，而输出信道只有一个。

1. 模型

见下图：



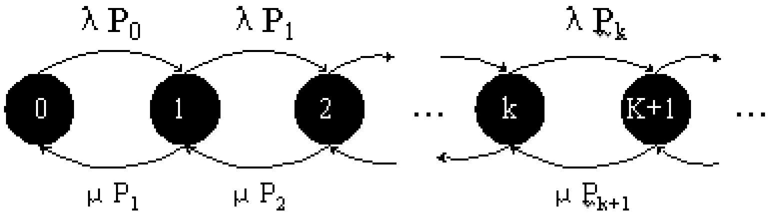
排队系统的两个参数：

- $T_w = T - T_s$
- W＝N－平均正被服务的顾客数（服务员忙的概率）

2. 平均队列长度

(1) 状态转移图

<u>*见下图*</u>:



平衡状态：处于 k 状态的概率 P_k 是一个与时间无关的常数

(2) 平衡方程

由平衡状态可得平衡方程：

$$\lambda P_k = \mu P_{k+1}\text{（k=0，1，2，…）}$$

令 $\rho = \dfrac{\lambda}{\mu}$，则：

$$P_k = \frac{\lambda}{\mu} P_{k-1} = \rho P_{k-1} = \rho^2 P_{k-2} = \rho^k P_0 \text{（k=1，2，…）} \qquad \text{（A-6）}$$

又因系统处于平衡状态，故 $\lambda < \mu$，即：$\rho = \dfrac{\lambda}{\mu} < 1$

由于 $\sum\limits_{k=0}^{\infty} P_k = 1$，将（A-6）式代入，得：$\sum\limits_{k=0}^{\infty} \rho^k P_0 = 1$

$$\Theta \sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho}$$

∴$P_0 = 1 - \rho$

从而得：$P_K = \rho^k P_0 = \rho^k (1 - \rho)$

(3) 平均顾客数

$$N = \sum_{k=0}^{\infty} k P_k = (1-\rho) \sum_{k=0}^{\infty} k \rho^k = (1-\rho) \frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}$$

⑷ 平均队列长度

∵P_0是系统为空的概率

∴ $\rho=1-P_0$为系统非空概率，即服务员忙的概率

从而得平均队列长度：

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

3. 平均等待时间（T）

由 Little 定律可知，在 M/M/1 系统中平均等待时间为：

$$\boxed{T = \frac{N}{\lambda} = \frac{\rho/\lambda}{1-\rho} = \frac{1/\mu}{1-\lambda/\mu} = \frac{1}{\mu-\lambda}} \qquad \text{(A-7)}$$

【注意】

这是在网络延迟时间分析中使用的关键结果。

4. 平均排队时间（T_w）

平均服务时间=1/μ，从而得平均排队时间为：

$$T_w = T - \frac{1}{\mu} = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$$

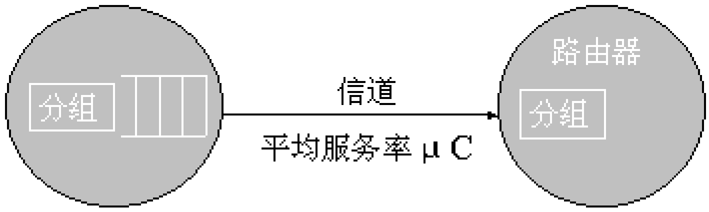
㈢ M/M/1 排队网络

1. 两种术语的对应

- 顾客—报文
- 服务员—信道
- 服务时间—报文发送时间

2. 网段平均延迟时间

见下图：



在网络文献中通常用 1/μ 表示平均分组长度，所以平均服务率为：

$$\frac{\text{容量}}{\text{平均分组长度}} = \frac{C}{1/\mu} = \mu C \quad \text{（分组/秒）}$$

所以一个分组通过该节点和输出信道，即通过一个网段存储转发的平均延迟时间：（由 A-7 可知）

$$\boxed{T = \frac{1}{\mu C - \lambda}}$$

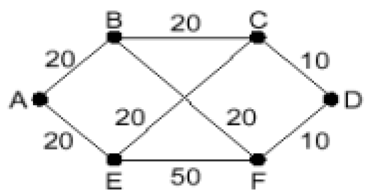
网络中可能各段 λ 不同，所以通过不同段的平均延迟时间不同，于是：

总平均延迟时间=所有段的平均延迟时间*平均段数

3. 举例

全双工通信：

- 通信子网：见下图



线上的数值表示信道容量 C_i（Kbps）。

- 通信量和路由矩阵：见下图

Source \ Destination	A	B	C	D	E	F
A		9 AB	4 ABC	1 ABFD	7 AE	4 AEF
B	9 BA		8 BC	3 BFD	2 BFE	4 BF
C	4 CBA	8 CB		3 CD	3 CE	2 CEF
D	1 DFBA	3 DFB	3 DC		3 DCE	4 DF
E	7 EA	2 EFB	3 EC	3 ECD		5 EF
F	4 FEA	4 FB	2 FEC	4 FD	5 FE	

数字表示通信量 r_{ij}（分组/s），字母表示路由。

- 网络分析：*见下图*

i	Line	λ_i (pkts/sec)	C_i (kbps)	μC_i (pkts/sec)	T_i (msec)	Weight
1	AB	14	20	25	91	0.171
2	BC	12	20	25	77	0.146
3	CD	6	10	12.5	154	0.073
4	AE	11	20	25	71	0.134
5	EF	13	50	62.5	20	0.159
6	FD	8	10	12.5	222	0.098
7	BF	10	20	25	67	0.122
8	EC	8	20	25	59	0.098

平均分组长度 1/μ=800 比特/分组，且反向通信量与正向通信量相同。

(1) 计算分组通过网络中一个网段的平均延迟时间

$$T' = \frac{\sum_{i=1}^{m}(\lambda_i T_i)}{\sum_{i=1}^{m}\lambda_i} = \sum_{i=1}^{m}\left(\frac{\lambda_i}{\lambda}T_i\right)$$（即所有网段的加权和）

其中，$\lambda = \sum_{i=1}^{m}\lambda_i$，本例 T'=86ms

⑵ 计算网络中所有端到端信息量总和

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}$$

本例 γ =62

⑶ 计算分组经过的平均网段数

$$\bar{n} = \frac{\lambda}{\gamma}$$

本例为 1.32

⑷ 计算网络的总平均延迟时间

$$T = \bar{n} \cdot T' = \frac{\lambda}{\gamma}\sum_{i=1}^{m}(\frac{\lambda_i}{\lambda}T_i) = \frac{1}{\gamma}\sum_{i=1}^{m}(\lambda_i T_i) = \frac{1}{\gamma}\sum_{i=1}^{m}\frac{\lambda_i}{\mu C - \lambda_i}$$

本例 T≈114ms