Queuing theory

models systems with servers and clients (presumably waiting in **queues**).

Notation: there are many standard symbols like $\lambda, \mu, \rho, A_n, W_n, L(t), L, L_Q, w, w_Q$ etc. These represent the **actual** properties of the system. When a model is used to **estimate** them, they appear with a "hat" as in: $\widehat{L}$.

The interpretation is simple: **estimator** $\widehat{L}$ is an approximate value of L. Note that the exact value of L will never be known.

Basic terms and estimators

- **System** denotes the sum of queues and servers.
- The average number of customers in the system $\widehat{L}$ and number of customers waiting in queues $\widehat{L}_Q$.
- The average time spent in the system $\widehat{w}$ and average time spent waiting in queues $\widehat{w_Q}$.
- Arrival rate λ and service rate μ (both expressed in $time^{-1}$.
- Server utilisation $\widehat{\rho}$.
- Probability of having i customers in the system P_i with P_0 being the most important.

The system and its model

The standard assumption is that the “system” starts its operation at time 0 and works forever. Thus we will not know the true properties of the system until the end of time (whatever that means).

In a model, we use the information gathered so far, i.e. from time 0 to time T (T is just a symbolic notation). The model yields some statistics; the larger T is, the closer these statistics are to the real ones.

Mathematically speaking: if L is the average number of customers in the system and $\widehat{L}$ is the number that came out of the model, then:

$$\lim_{T \to \infty} \widehat{L} = L$$

The same applies to all estimators.

Average number of customers

In the actual system the number of customers in the system is given by the function $L(t)$. Obviously we do not know this function in its analytical form. The average number of customers in the real system is:

$$L = \lim_{t \to \infty} \int_0^t L(x)dx$$

In the model, we know what happened in the period $[0, T]$. The number of customers varied in time, but their number at any moment can be recorded and then tallied up. For each number of customers i (where $0 \le i < \infty$) the lengths of the time intervals when there were precisely i customers in the system are added together giving a total time T_i Clearly:

$$\sum_{i=0}^{\infty} T_i = T$$

The estimator $\widehat{L}$ for the average number of customers is a weighted average (weights are time durations):

$$\widehat{L} = \frac{1}{T} \sum_{i=0}^{\infty} i T_i = \sum_{i=0}^{\infty} i \frac{T_i}{T}$$

Note that the right formula looks like the mean of a pdf.

What we are interested in is L. What we have is $\widehat{L}$.

It is easy to see that

$$\widehat{L} = \frac{1}{T} \int_0^T L(t)dt$$

(even though we do not know how $L(t)$ looks like).

Since

$$L = \lim_{t \to \infty} \int_0^t L(x)dx$$

it is obvious that:

$$\lim_{T \to \infty} \widehat{L} = L$$

Average time spent in the system

The average time spent in the system is denoted by w and is impossible to determine in finite time.

In the model, we can record the times spent in the system of all the customers that left before time T. Let these times be $W_1, W_2, \cdots, W_n$ where n is the number of arrivals (or departures) in the period $[0, T]$.

An estimator for w is:

$$\widehat{w} = \frac{1}{n} \sum_{i=1}^{n} W_i$$

As in the case of the number of customers,

$$\lim_{n \to \infty} \widehat{w} = w$$

This is true for all “real” cases, but not “mathematically” as there are some weird (and unrealistic) counterexamples.

Little's law

In the model, the arrival rate is $\frac{n}{T}$. We denote this rate by $\widehat{\lambda}$ because it is an estimator of the true arrival rate λ.

The **conservation law** (Little's) gives the following relationship:

$$\widehat{L} = \widehat{\lambda}\widehat{w}$$

Not surprisingly, Little's law holds for the real system:

$$L = \lambda w$$

but this is purely for the record, because we have no means to compute L or w.

Server utilisation

Let μ be the service rate (as in “can handle μ customers per minute”). Note that $\frac{1}{\mu}$ is the average service time.

If $\mu < \lambda$ the system eventually explodes, because the waiting queue grows to infinity.

A system can be stable only if $\lambda < \mu$. In that case, the server rate should be split into two parts: the part when the server is “busy” and the remainder, when the server is “idle.” The busy part obviously is equal to the arrival rate λ because th e server is busy if and only if there is a customer around.

Hence the server utilisation ρ (“the busy part”) equals:

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

If the system has c identical servers, the same rule holds:

$$\rho = \frac{\lambda}{c\mu} \qquad\qquad \lambda < c\mu$$

Types of queuing systems

Kendall proposed a unified notation describing the properties of a given queuing system: $A/B/c/N/K$, where the letters represent:

A interarrival–time distribution

B service–time distribution

c the number of parallel servers

N the system capacity

K the size of the customer population.

Details

The following symbols are commonly used for the two distribution parameters:

M: exponential (named after Markov).

D: deterministic (fixed rate).

E$_k$: Erlang of order k.

G: general, i.e. not specified.

There are other possibilities, less common.

The other three parameters are integers; they are commonly omitted if equal to ∞.

Everything is known about $M/M/1$ queues; almost nothing about $G/G/c$ queues (beyond the conservation law).

$M/M/1$

The most important formula for $M/M/1$ queues is:

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

and its companion:

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

They tell us what happens when $\lambda \longrightarrow \mu$ (i.e. $\rho \longrightarrow 1$).

The average number of customers in the system is:

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

Also, $L_Q = \rho L$ and $w_Q = \rho w$.

An example

A tool crib has exponential interarrival and service times and serves a very large group of mechanics.

Eureka! $M/M/1$ *queue.*

The mean time between arrivals is 4 minutes. It takes 3 minutes on average for a tool–crib attendant to service a mechanic.

Aha. $\lambda = \frac{1}{4}, \mu = \frac{1}{3}$. $\lambda < \mu \rightarrow \rho = \frac{\lambda}{\mu} = \frac{3}{4}$

The attendant in paid \$10/hour and the mechanic is paid \$15/hour. Would it be advisable to have a second tool–crib attendant?

Hmmm. $M/M/2$ *queue?*

Single attendant

The tool crib is a system with one server. The average waiting time is $\frac{1}{\mu(1-\rho)} = 12$ minutes.

12 minutes of a mechanic's time are worth \$3.

15 mechanics arrive per hour, hence the hourly cost of the system is: \$10 + 15 $\times$ \$3 = \$55/hour.

$M/M/2$

We use the formula for $M/M/c$ from the book, using $c = 2$:

$$P_0 = 0.4545, L = 0.8727, w = 3.4908$$

i.e. the cost is 3.49/60 $\times$ \$15 = \$0.8727 per hour per mechanic.

The total hourly cost is: 2 $\times$ \$10 + 15 $\times$ \$0.8727 = \$33.09/hour.

Queue confusion

$\mathcal{A}$ and $\mathcal{B}$ applied for a loan manager position. The bank manager is obsessed with minimising the average queue length. The customer arrival rate is $\lambda = 1/30$ (2 per hour).

	$\mathcal{A}$	$\mathcal{B}$
Service time $1/\mu$	24	25
Standard deviation of s.t. σ	20	2

Note that $\mathcal{A}$ is "almost" an exponentially–distributed server while $\mathcal{B}$ is close to deterministic.

Which of them should be hired?

M/M/1

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

$$L_Q = \rho L$$

$$L_Q = \frac{\rho^2}{1-\rho}$$

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

	ρ	L_Q
$\mathcal{A}$	$\frac{1}{30} \times 24 = 4/5$	$\frac{4^2}{5^2(1/5)} = 16/5 = 3.2$
$\mathcal{B}$	$\frac{1}{30} \times 25 = 5/6$	$\frac{5^2}{6^2(1/6)} = 25/6 = 4.17$

Clearly, $\mathcal{A}$ should be hired.

M/G/1 queue

$$L_Q = \frac{\rho^2}{1-\rho} \times \frac{1 + (\sigma\mu)^2}{2}$$

(looks like an "average" between an M/M/1 queue and $\frac{(\lambda\sigma)^2}{1-\rho}$ with the variability of service being the perturbation).

	ρ	σ	L_Q
$\mathcal{A}$	$4/5$	20	2.711
$\mathcal{B}$	$5/6$	2	2.097

It appears that $\mathcal{B}$ is the better candidate (after replacing the exponential service time with an unclear service distribution).

M/M/c

There are c identical servers ("channels") used by an unlimited population of customers.

There are several ways of implementing an M/M/c system. The main two:

Channel division with no multiplexing: the c channels are separate each with its own input queue. Used in Telecommunications as **TDMA** and **FDMA**.

Statistical multiplexing: arrivals join a single queue and enter the first available channel (Internet's **best effort**).

The M/M/c model describes statistical multiplexing.

	M/M/1	M/M/c
ρ	$\frac{\lambda}{\mu}$	$\frac{\lambda}{c\mu}$
L	$\frac{\rho}{1-\rho}$	$c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2}$
w	$\frac{1}{\mu-\lambda}$	$\frac{L}{\lambda}$
L_Q	$\frac{\rho^2}{1-\rho}$	$L - c\rho$
P_n	$(1-\rho)\rho^n$	impossibly complicated

M/M/∞ and M/G/∞ systems

This seemingly absurd system is used to model the performance of some Internet services.

As usual, λ is the customer arrival rate and μ is the service rate (hence, $\frac{1}{\mu}$ is the mean service time). Server utilisation makes no sense in this context, but let us use $\rho = \frac{\lambda}{\mu}$ anyway.

$$P_0 = e^{-\rho} = e^{-\lambda/\mu}$$

$$P_n = P_0 \frac{\rho^n}{n!}$$

$$L = \rho$$

$$L_Q = w_Q = 0$$

$$w = 1/\mu$$

These equations apply to any M/G/∞ systems, including M/M/∞.

M/M/c/N

The key property is that no more than N customers can be in the system. Hence the key importance of P_N, the probability that there are N customers in the system:

$$P_N = P_0 \frac{\rho^N}{c^{N-c}}$$

where $\rho = \frac{\lambda}{c\mu}$ (utilisation rate). Although λ is the customer arrival rate, the rate at which customers are entering the system is different, λ_e (“effective”):

$$\lambda_e = \lambda(1 - P_N)$$

This impacts Little’s equality:

$$L = \lambda_e w$$

$$L_Q = \lambda_e w_Q$$

The formula for L is unprintable but not unusable.