Elementary Queuing Analysis

Basics and M/M/1 Analysis

Of all the analysis techniques that enable us to do "back of the envelope" estimation of computing systems performance, queuing analysis is by far the most important.

Introduction to Queuing Systems

The basic entities in a queuing system are:

Customers:

These are the individual requests for service (e.g. a request for I/O, or a request by a customer in a bank, etc.)

Queues:

These are waiting areas where requests for service wait for server(s) (e.g. the ready queue of processes waiting for CPU, or the waiting room at a doctor's office).

Servers:

These are the entities or resources that are capable of satisfying the service requests (e.g. CPU, disk, bank teller, etc.)

In addition to the above entities, we must discuss a number of other issues, including:

- **Dispatching Discipline:** Once a server is done serving a customer, it must pick the next customer out of some queue. The algorithm used to do so is termed the dispatching discipline. Possibilities (e.g. for scheduling purposes) include First-Come-First-Serve (FCFS) also called First-In-First-Out (FIFO), Shortest-Job-First, Earliest-Deadline-First, etc. We will see the impact of these scheduling techniques later in the course. For the purposes of this course we will restrict our analysis to FCFS.
- **Distribution of Arrivals:** For the purposes of this class we will restrict our analysis to a Poisson process for the arrival of customers (from the outside world) to the system. The rate of arrivals is λ. In other words, the number of customers coming into the system every period T is λT. As we have discussed in earlier lectures, a Poisson arrival process implies that the arrivals are independent.

- **Distribution of Service Time:** How long does it take a server to service a customer's request? Well, the service time may be the same for all customers (e.g. all customers request exactly the same service and it takes exactly the same time to serve each and every one of them). Alternatively, and more realistically, the service time is likely to be variable. For the purposes of this class we will restrict our analysis to the case in which the service time is a random variable that is exponentially distributed with a mean service time T_s. If the average time it takes a server to service a request is T_s, then it follows that the average rate of service (if the server has an infinite supply of requests to work on) would be $\mu=1/T_s$.

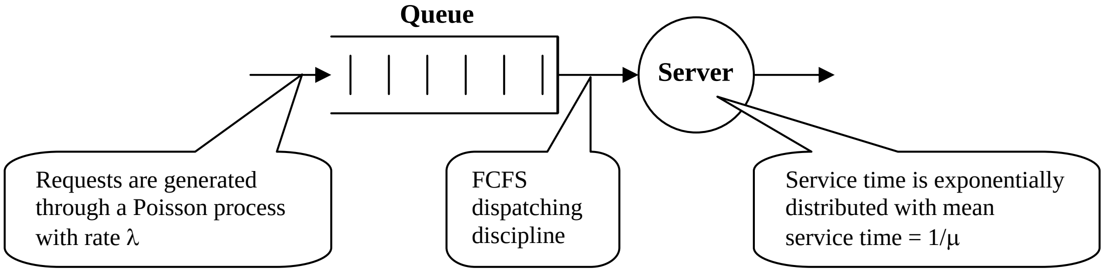
For the purposes of our introductory treatment of queuing theory, we make the following assumptions:

- **Population:** We assume that Requests for service are generated from an infinite population. The significance of this assumption is that the arrival of a request to the system does not influence "future" arrivals. For example, the likelihood of a new customer walking into the bank in a given interval of time has nothing to do with other customers walking in and out of the bank.

- **Queue Size:** We assume that queues have infinite capacity. The significance of this assumption is that it will never be the case (in our analysis, that is) that requests for service will be "lost" or will affect the likelihood of other requests joining the queue. For example, we exclude the likelihood that a patient will not be able to see a doctor because there was no waiting space for them in the doctor's office. A more computing-system-related example would be the arrival of IP packets to the buffer of a router. Under our assumption, packets can always be buffered and thus cannot be lost! Of course, any real system must have a finite capacity (e.g. network buffers, etc.) and it is possible (albeit more complex) to analyze queuing systems with finite queues. However, this assumption simplifies analysis greatly and provides an acceptable approximation.

Single-Queue Single-Server System

We are interested in studying the performance of the system below *in the steady state.*


Queue
Server
Requests are generated through a Poisson process with rate λ
FCFS dispatching discipline
Service time is exponentially distributed with mean service time = 1/μ

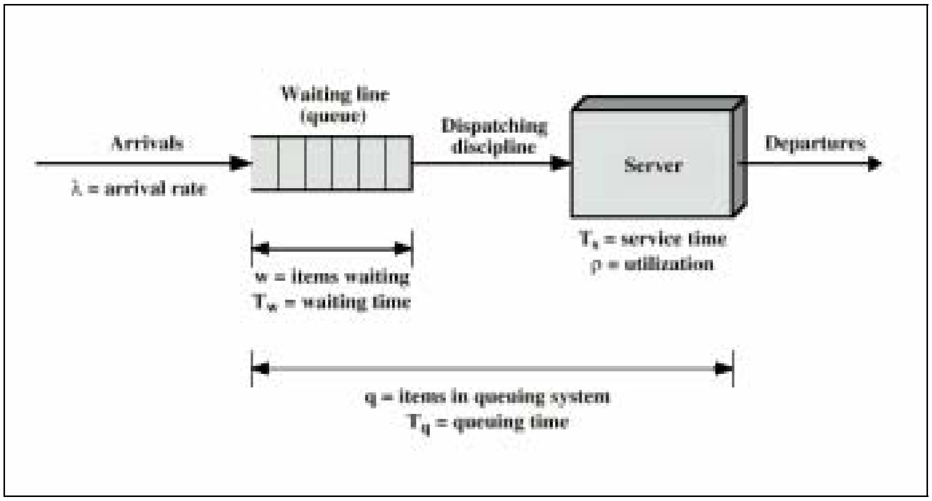
Notation:

- ❑ Let w denote the number of customers waiting in the queue
- ❑ Let q denote the total number of customers in the system (waiting + being served)
- ❑ Let T_s denote the service time (the time it takes the server to serve a customer).
- ❑ Let T_w denote the waiting time in the queue.
- ❑ Let T_q denote the turnaround time (waiting time + service time).
- ❑ Let ρ denote the utilization of the system, which is the ratio between the rate of arrivals and the rate of service $\rho = \lambda/\mu$.

Obviously, in the steady state, the rate at which requests are queued cannot exceed the rate at which the server is able to serve them. Thus, we have:

$$\lambda < \mu$$

$$\rho < 1$$



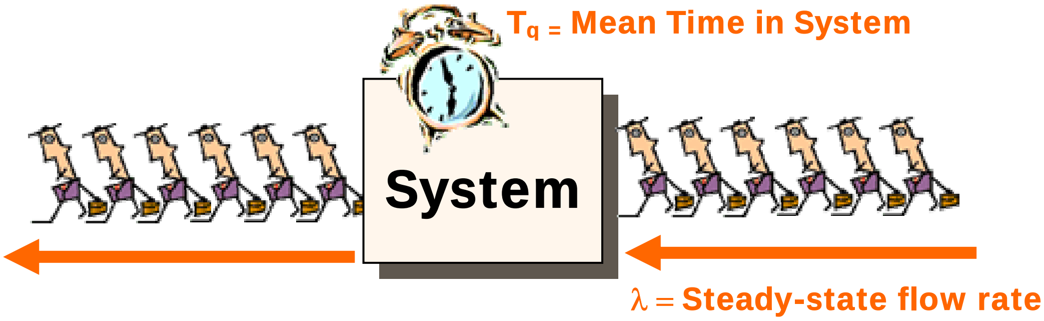
Some Basic Queuing Relationships

Little's Formula

The following two relationships are true of any "steady state" queuing system (i.e. a queuing system in equilibrium). They are known as Little's Formulae (or Little's Theorem, or Little's Law).

$$\mathbf{q} = \lambda\,.\mathrm{T}_{\mathbf{q}}$$

$$\mathbf{w} = \lambda\,.\mathrm{T}_{\mathbf{w}}$$



The intuition behind these two formulas is that over a period of time T, the number of arrival in the system is λ T. At equilibrium, every customer will wait (on the average) an amount of time T_q in the system. From the instant a customer arrives and until that customer leaves (i.e. after T_q units of time) λT_q "new" customers would have joined the system (on the average). Thus at the time the customer leaves the system a total of λT_q customers would be in the system. Obviously, the same applies at the time any other customer leaves the system (and in general at any instant of time!)

Relationship between Turnaround, Queuing, and Service Times

In a queuing system, a customer's time is spent either waiting for service or getting service. Thus, we get the following additional (obvious) relationship:

$$\mathbf{T_q = T_w + T_s}$$

Multiplying the above equation by the arrival rate λ and applying Little's formula, we get:

$$\mathbf{q = w + \lambda T_s = w + \lambda\,/\mu = w + \rho}$$

Analysis of an M/M/1 Queuing System

M/M/1 Queues

An M/M/1 queuing system is a single-queue single-server queuing system in which arrivals are Poisson and service time is exponential. The notation M/M/1 describes the "queue" in the system as having a **M**arkovian arrival process (i.e. Poisson) and a **M**arkovian (i.e. exponential) service discipline with **1** server.

Birth and death probabilities for M/M/1

Consider a very small interval of time of length *h*. Assume that this interval of time (*h*) is so small that a maximum of one arrival can realistically occur in that period of time. Since the rate of arrival is λ requests per unit time, then it follows that the rate of arrival per interval is λh.

During an interval *h* one of two things can happen: either no requests arrive during that small interval of time, or one request does arrive. We call the event of an arrival of a request to the system a "*birth*" event.

We know that the probability density of the Poisson distribution is:

$$f(x) = (\frac{\lambda^x}{x!})e^{-\lambda}, x = 0,1,2,...$$

Given that the rate of arrival per interval *h* is λh, the probability of *x* arrivals per interval *h* is

$$f(x) = (\frac{\lambda h^x}{x!})e^{-\lambda h}, x = 0,1,2,...$$

According to the above, the probability that there will be *no* arrivals during a given interval *h* is $e^{-\lambda h}$ and, thus, the probability that at least[1] one arrival (i.e. a birth) will occur is:

[1] Recall that we assume that the period is so small that we can assume that a maximum of one arrival is possible during that period of time.

Pr(birth) = 1 - Pr[no arrivals]
Pr(birth) = 1- $e^{-\lambda h}$
Pr(birth) = 1 - (1 - λh + $(\lambda h)^2$/2! - $(\lambda h)^3$/3! - …)

For very small period *h*, we can use the first order approximation, which results in the following probability:

Pr(birth) = λh

Similarly, we can show that the probability that a customer will leave the system (i.e. a customer for whom service was finished) given that somebody is in the system in the first place is μh. We call such an event a *"death"* event.

Pr(death) = μh

State (rate) transition diagram for M/M/1

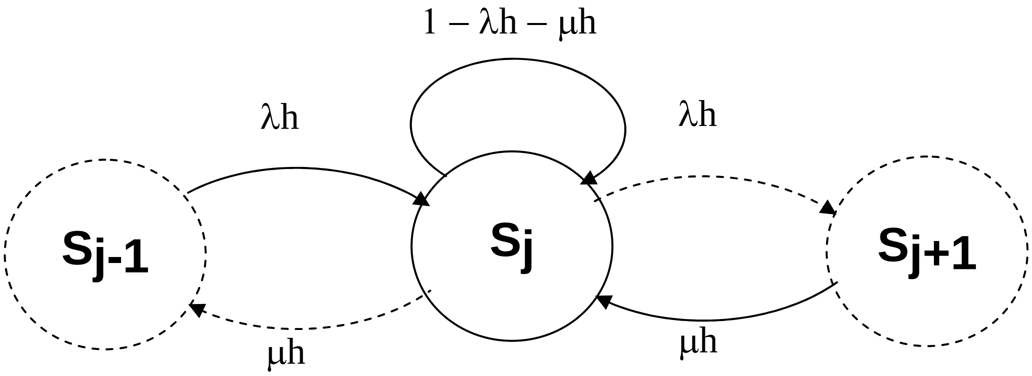
Consider a M/M/1 system at steady state (i.e. equilibrium). Such a system will have a variable number of customers. In particular, at any point of time, a customer may be added to the system through a birth event, or a customer may be removed from the system due to a death event.

Consider the state of the system when exactly j customers are in the system. We denote such a state by Sj. In order to compute many important properties of an M/M/1 queuing system, it will be necessary to calculate the probability that *at steady state*, the system is in a state Sj (for any j=0, 1, 2, …)

Consider the system at a given instant. Let the state of the system at that instant be Sj. How could the system be in such a state? Well, to answer this question, consider the system at an interval *h* earlier (where *h* is very small). There are three scenarios that would result in the system moving into state Sj (see the figure below):

(a) The system was in state Sj-1 and a birth occurred. The probability of that happening is λh.
(b) The system was in state Sj+1 and a death occurred. The probability of that happening is μh.
(c) The system was in state Sj and, neither a birth nor a death occurred. The probability of that happening is $1-\lambda h-\mu h$.

The diagram below shows the above transitions. Solid arrows denote the transitions that result into entering state Sj.



From the above, we have the following relationship:

$$\Pr(S_j) = \lambda h \Pr(S_{j-1}) + \mu h \Pr(S_{j+1}) + (1-\lambda h-\mu h) \Pr(S_j)$$

Rearranging terms, we get

$$\mu \text{Prob}(S_{j+1}) = (\mu+\lambda) \text{Prob}(S_j) - \lambda \text{Prob}(S_{j-1})$$

$$\text{Prob}(S_{j+1}) = (1+\rho) \text{Prob}(S_j) - \rho \text{Prob}(S_{j-1})$$

S_0 is obviously a special case (since there is no S_{-1}). Thus we get:

$$\text{Prob}(S_0) = \mu \text{Prob}(S_1) + (1-\lambda) \text{Prob}(S_0)$$

$$\text{Prob}(S_1) = \rho \text{Prob}(S_0)$$

By successive substitution of $\Pr(S_1)$ and $\Pr(S_0)$ in the equation for $\Pr(S_{j+1})$, we obtain $\Pr(S_2)$, $\Pr(S_3)$, $\Pr(S_4)$, … Namely:

$$\boxed{\mathbf{Pr(S_j) = \rho^j Pr(S_0)}}$$

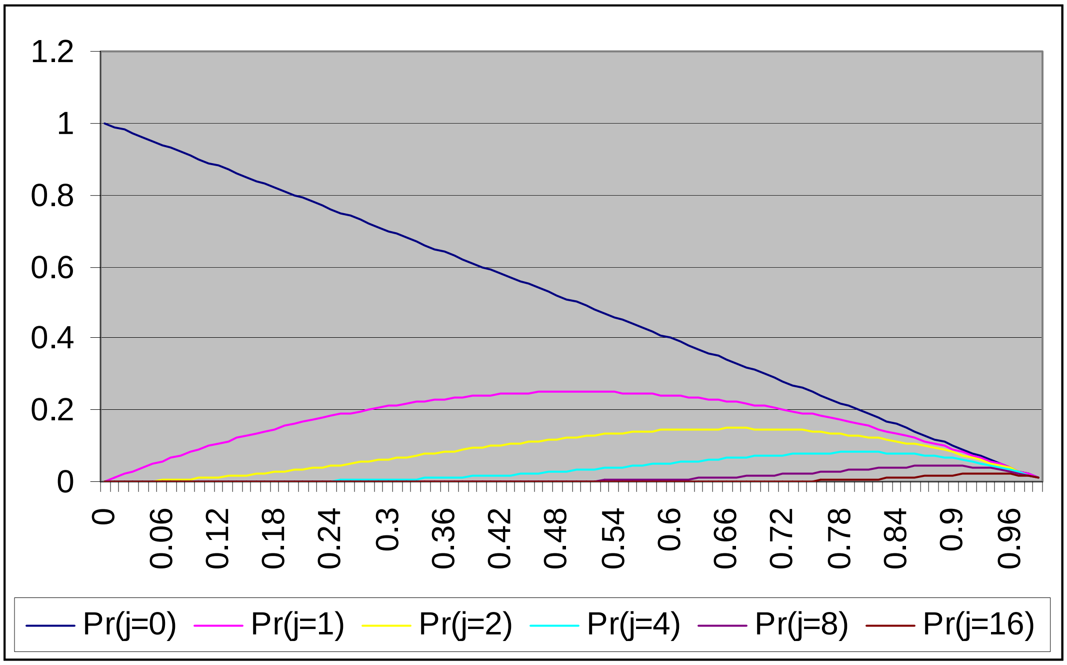
Moreover, since the overall probability density must add up to 1, we get:

$$\Pr(S_0) + \rho \Pr(S_0) + \rho^2 \Pr(S_0) + \rho^3 \Pr(S_0) + \rho^4 \Pr(S_0) \ldots = 1$$

$$\Pr(S_0) [1 + \rho + \rho^2 + \rho^3 + \rho^4 \ldots] = 1$$

$$\boxed{\mathbf{Pr(S_0) = (1-\rho)}}$$

To appreciate the above relationship, let us look at the probability of having j customers in the system (i.e. $\Pr(S_j)$) as a function of the utilization. This is shown in the figure below for j=0, 1, 2, 4, 8, and 16.



Probably, a better visualization of the relationship between the number of pending requests (i.e. number of customers in the system) and utilization is that relating the probability that less than n customers are in the system. We derive this next.

Average number of customer in a M/M/1 System

The average number of customers in a M/M/1 system could be calculated as follows:

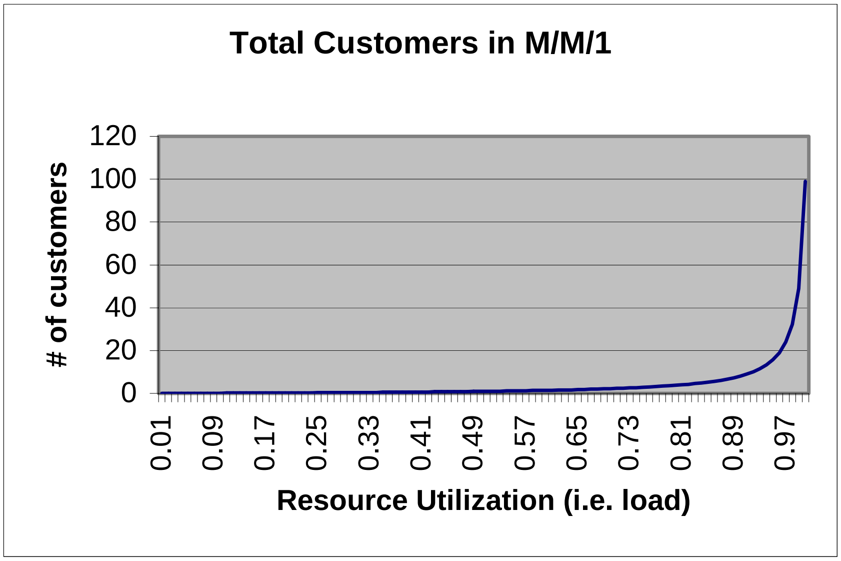
$$q = E[\text{Number of customers}] = 0 * \Pr(S_0) + 1 * \Pr(S_1) + 2 * \Pr(S_2) + \ldots$$

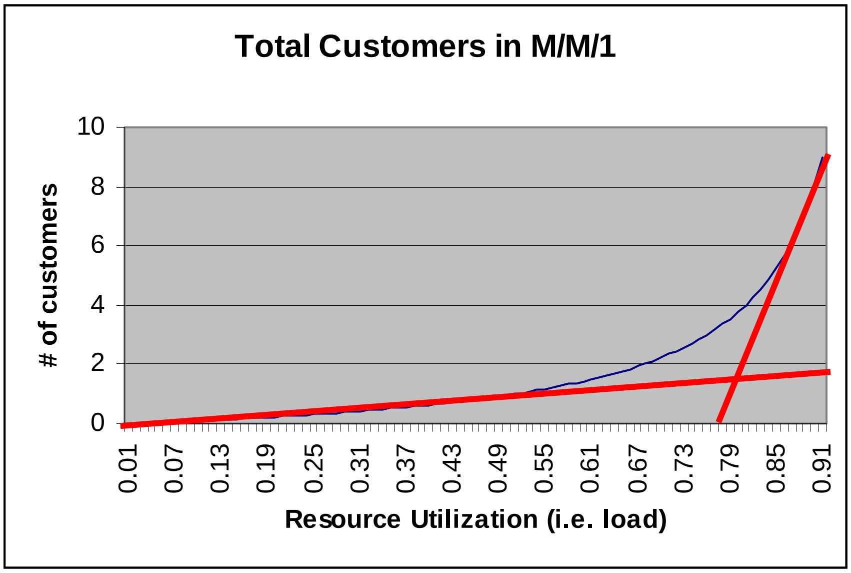
Substituting from the formulas above, we obtain:

$$q = (\rho + 2\rho^2 + 3\rho^3 + 4\rho^4)*(1-\rho)$$

$$\boxed{\mathbf{q} = \rho/(1-\rho)}$$

To appreciate the above relationship, let us plot the number of pending requests (i.e. q) as a function of the utilization of the system. This is shown below.


Total Customers in M/M/1
# of customers
120
100
80
60
40
20
0
0.01
0.09
0.17
0.25
0.33
0.41
0.49
0.57
0.65
0.73
0.81
0.89
0.97
Resource Utilization (i.e. load)


Total Customers in M/M/1
# of customers
10
8
6
4
2
0
0.01
0.07
0.13
0.19
0.25
0.31
0.37
0.43
0.49
0.55
0.61
0.67
0.73
0.79
0.85
0.91
Resource Utilization (i.e. load)

Average number of customers waiting for service in a M/M/1 system

We know that $q = w + \rho$. Thus,

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

$$\boxed{\mathbf{w} = \rho^2/(1-\rho)}$$

Average Time in a M/M/1 system

Using Little's formula, we get

$$T_q = q/\lambda$$

$$\boxed{\mathbf{T_q} = 1/\mu(1-\rho)}$$

Average Time waiting in a M/M/1 queue

Using Little's formula, we get

$$T_w = w/\lambda$$

$$\boxed{\mathbf{T_w} = \rho/\mu(1-\rho)}$$

Variability Measures

The above measures average-case behaviors. As we explained earlier, it would be interesting to estimate the variability around these averages. Using similar derivations, one can derive the standard deviations of the above metrics. In particular, we can calculate the standard deviation in the number of customers in an M/M/1 system (i.e. q)

$$\boxed{\sigma_q = \frac{\sqrt{\rho}}{1-\rho}}$$

$$\boxed{\sigma_{Tq} = \frac{T_s}{1-\rho}}$$