Notes on Queuing Theory

Mark S. Daskin
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208

1. Introduction to the problem

A **queue** is a line for service. We are all familiar with queues from our everyday experiences. When we call a doctor, we may be placed in a queue until the receptionist can answer our call. If we need to see a doctor, we usually must wait in his/her office until (s)he can see us. When we go to a grocery store, we must join a queue to pay for our food. We wait for busses in cities and for lights to change when we drive our cars. And so on….

Queuing theory is concerned with the mathematical description of queues. Two basic approaches to queueing theory have developed: models based on fluid approximations (e.g., Newell, 1971) and probabilistic queuing analysis. These notes concentrate on the latter, which has received the bulk of the attention to date. In simplistic terms, the fluid approximation approach is a deterministic view of queues and is particularly useful in analyzing queues in which the average arrival rate exceeds the average service rate for *extended* periods of time. Probabilistic queuing theory, as its name implies, adopts a stochastic view of queues and is most useful in analyzing queues in which the arrival rate is less than the service rate for *extended* periods of time.

Queuing theory, be it deterministic or stochastic, requires several **inputs**. These include:

a) A description of the way in which customers arrive at the system. This is termed the **arrival process**. Of particular interest is the distribution of the time between customer arrivals.
b) A description of the way in which customers are served or the **service process**. Of particular concern are estimates of the mean and variance of the time needed to serve a customer.
c) The number of servers
d) The maximum number of customers that can be in the system
e) The size of the pool of customers
f) The way in which waiting customers are chosen for service, or the **service discipline**.

Inputs (a), (b), and (c) are always needed. Kendall has developed a standard notation for these inputs. The notation is written as X/Y/Z where

X and Y are letters used to describe the arrival and service processes respectively, and Z is an integer (may be ∞) stating the number of servers.

To be more specific, X and Y are used to describe the probability distributions used in modeling the **interarrival times** and **service times** of customers. Frequently used symbols include:

M	E**xponential** distribution. Note that as shown below, an exponential interarrival time distribution corresponds to **Poisson** arrivals
E_k	**Erlang-k** distribution. Recall that an Erlang-1 distribution is an exponential distribution.
HE	**Hyperexponential** distribution
D	**Deterministic**
G, GI	Any **general** distribution with a finite mean and variance. GI is usually used for arrivals and denotes general independent; G is usually used for service times. In both cases, we assume that successive interarrival or service times are independent random variables.

The **outputs** of queueing models include:

a) The mean number in the system (in the queue or line and in service)
b) The mean number in the queue (waiting for service)
c) The mean time in the system or in the queue.
d) The distribution of time in the queue or in the system.

2. Relations between key probability distributions

Before proceeding, we derive several key relationships between the Poisson, Exponential, and Erlang-k distributions as well as properties of the distributions.

If the number of arrivals in time *t*, *N(t)*, follows a **Poisson process**, we have

$$P(N(t)=n)=\frac{(\lambda t)^n e^{-\lambda t}}{n!} \quad n=0,1,2,... \tag{1}$$

and in particular,

$$P(N(t)=0)=e^{-\lambda t}. \tag{2}$$

If the time between arrivals is Exponential, we have

$$P(\text{interarrival time} \le t) = \int_0^t \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda t} \qquad t \ge 0 \tag{3}$$

and $$P(\text{interarrival time} > t) = \int_t^\infty \lambda e^{-\lambda x}\, dx = e^{-\lambda t} \quad t \ge 0 \tag{4}$$

where in (1)-(4), λ is the rate of customer arrivals per unit time.

From equations (2) and (4), we see that **Poisson arrivals imply that the time between arrivals is Exponential and vice versa.**

The exponential distribution has a key property, the **memoryless property**, that makes it particularly useful in queuing theory. In words, the property states that if interarrival times (for example) are Exponential, then the probability distribution of the remaining time until an arrival given that we have already waited T_0 minutes is also Exponential with the same (original) parameter. To see this, we note that if t=interarrival time, for $K > T_0$,

$$P(t > K | t > T_0) = \frac{P(t > K \text{ AND } t > T_0)}{P(t > T_0)} = \frac{P(t > K)}{P(t > T_0)} = \frac{e^{-\lambda K}}{e^{-\lambda T_0}} = e^{-\lambda(K - T_0)} \tag{5}$$

but $K - T_0$ is the remaining time. So the

$$P(\text{remaining time} > K - T_0 = R | t > T_0) = e^{-\lambda(K - T_0)} = e^{-\lambda R} \tag{6}$$

which is Exponential with the original parameter λ. In other words, if buss arrivals follow a Poisson process with a mean of 6 per hour (one every 10 minutes on average), the expected additional waiting time given we have been waiting 8 minutes is 10 more minutes, **not** 2 minutes. Note that an estimate of 2 more minutes would be wrong for virtually all distributions of interarrival times.

This property means that, in modeling a queue with Poisson arrivals, we do not need to know when the last person arrived to characterize the state of the system. Similarly, we do not need to know how long the current customers have been in service if service times are Exponentially distributed. Qualitatively speaking this means that the state space described by the number of people in the system is **Markovian**, meaning we do not have to worry about how we got to the state in order to fully describe the probability distribution of the state space at some future point in time. This leads us to study such queues. In section 4 below, we are more specific about what is needed to have a Markovian state space.

If $X_1, X_2, ..., X_k$ are k independent identically distributed random variables, each with an Exponential distribution given by $f_{X_i}(x_i) = \lambda e^{-\lambda x_i} \qquad x_i \geq 0 \quad i = 1,2,...,k$, then $S_k = \sum_{i=1}^{k} X_i$ is a random variable with an **Erlang-k** distribution:

$$f_{S_k}(s) = \frac{\lambda(\lambda s)^{k-1} e^{-\lambda s}}{(k-1)!} \quad s \geq 0 \qquad k = 1,2,... \tag{7a}$$

Sometimes we write

$$f_{S_k}(s) = \frac{k\nu(k\nu s)^{k-1} e^{-k\nu s}}{(k-1)!} \qquad s \geq 0 \qquad k = 1,2,... \tag{7b}$$

where $\lambda = k\nu$. We have

$$E(S_k) = \frac{k}{\lambda} = \frac{1}{\nu} \tag{8}$$

$$Var(S_k) = \frac{k}{\lambda^2} = \frac{1}{k\nu^2} \tag{9}$$

So the Erlang-k distribution is the distribution of the sum of k i.i.d. Exponential random variables. The cumulative Erlang-k may be found by noting that

$$\begin{aligned} P(S_k > s) &= P(\text{sum of k i.i.d. Exponential random variables} > \text{s}) \\ &= P(k-1 \text{ or fewer Poisson arrivals in time s}) \\ &= \sum_{n=0}^{k-1} \frac{(\lambda s)^n e^{-\lambda s}}{n!} \end{aligned} \tag{10}$$

So, the Poisson and Erlang-k are related. In particular,

$$P(k \text{ Poisson arrivals in time } s) = P(S_k < s) = \text{cumulative Erlang - k}$$

Finally, we note that the probability of no Poisson arrivals in time Δt is

$$P(N(\Delta t) = 0) = e^{-\lambda \Delta t} \approx 1 - \lambda \Delta t + o\left((\Delta t)^2\right) \tag{11}$$

where $o\left((\Delta t)^2\right)$ are terms of order $(\Delta t)^2$ or smaller. For <u>small</u> Δt we have

$$P(N(\Delta t) = 0) \approx 1 - \lambda \Delta t \tag{12a}$$

$$P(N(\Delta t) = 1) \approx \lambda \Delta t \tag{12b}$$

$$P(N(\Delta t) > 1) \approx 0 \tag{12c}$$

We will make use of equations (12) in Section 4 below.

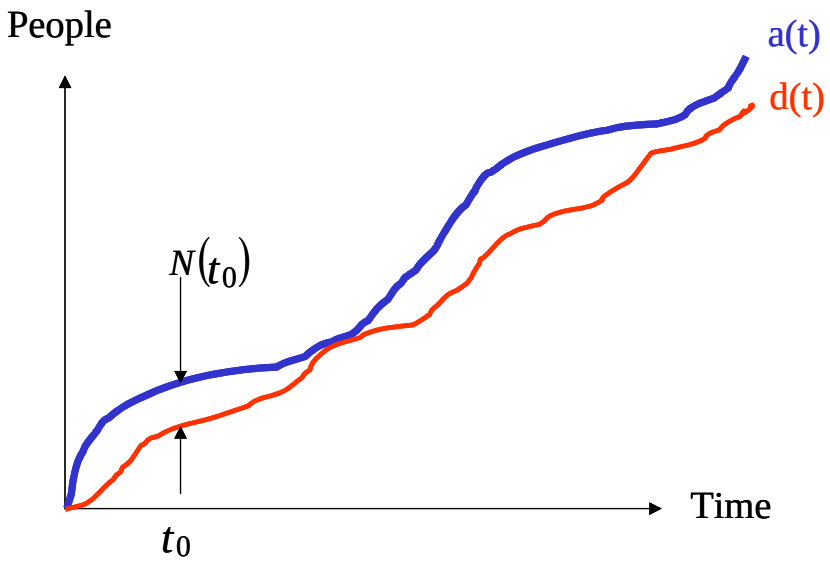
3. A few basic relations

We define the following quantities

L = average number of customers in the system
L_q = average number waiting to be served
W = average time in the system
W_q = average time in the queue waiting to be served
λ = average arrival rate
μ = average service rate
$\frac{1}{\mu}$ = average service time

Also let

$a(t)$ = number of people to arrive at the sytem in time $[0, t]$
$d(t)$ = number of people to depart from the system in time $[0, t]$
$N(t)$ = number of people in the system at time $t = a(t) - d(t)$



The total number of person-minutes of time in the system is given by the area between *a(t)* and *d(t)*. That is, let

$\hbar(t)$ = total accumulated person - minutes during the interval $[0, t]$

and

$$\hbar(t) = \int_0^t N(\tau)d\tau = \int_0^t [a(\tau) - d(\tau)]d\tau \tag{13}$$

The average waiting time in the interval [0,t] is $\frac{\hbar(t)}{a(t)} = W(t)$ and the average number in the system is $\frac{\hbar(t)}{t} = L(t)$. So

$$L(t) = \frac{\hbar(t)}{t} = \frac{\hbar(t)}{a(t)} \bullet \frac{a(t)}{t} \tag{14}$$

Taking the limit of (14) as $t \to \infty$, we have

$$\begin{aligned} &\lim_{t \leftarrow \infty} L(t) = L \\ &\lim_{t \leftarrow \infty} \frac{\hbar(t)}{a(t)} = W \\ &\lim_{t \leftarrow \infty} \frac{\hbar(t)}{t} = \lambda \end{aligned} \tag{15}$$

So

$$\boxed{L = \lambda W} \tag{16}$$

This is known as **Little's Formula** and it holds under very general circumstances. We can similarly show that

$$\boxed{L_q = \lambda W_q} \tag{17}$$

and

$$\boxed{L_s = \lambda W_s = \lambda \frac{1}{\mu}} \tag{18}$$

where L_s is the average number in service and W_s is the average service time. Finally we also have,

$$\boxed{W = W_q + \frac{1}{\mu}} \tag{19}$$

or the average time spent in the system is the sum of the average time spent waiting for service to begin plus the average service time.

4. A framework for Markovian queues

A **Markovian** stochastic process is one in which the conditional probability of being in any state at some future time given the present and past states equals the probability of being in the state in the future given only the present state. That is, the past history of the system does not provide any information needed to predict future states.

If the arrivals are Poisson and the service times are Exponential, the underlying process is **Markovian**.

Let

λ_n = (Poisson) arrival rate with *n* people in the system

μ_n = service rate with *n* people in the system

Then we can write, letting $P_i(t)$ be the probability of being in state *i* at time *t*,

$$P_0(t+\Delta t) = (1-\lambda_0 \Delta t) P_0(t) + (\mu_1 \Delta t) P_1(t) \tag{20}$$

and

$$P_i(t+\Delta t) = (\lambda_{i-1} \Delta t) P_{i-1}(t) + (1-\lambda_i \Delta t)(1-\mu_i \Delta t) P_i(t) + (\mu_{i+1} \Delta t) P_{i+1}(t) \qquad i = 1,2,... \tag{21}$$

If we bring $P_i(t)$ to the left hand side of (20) and (21) and divide by Δt, we have

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \tag{22}$$

$$\frac{P_i(t+\Delta t) - P_i(t)}{\Delta t} = \lambda_{i-1} P_{i-1}(t) - \lambda_i P_i(t) - \mu_i P_i(t) + \lambda_i \mu_i P_i(t)\Delta t + \mu_{i+1} P_{i+1}(t) \qquad i = 1,2,... \tag{23}$$

Taking the limits as $\Delta t \to 0$, we have

$$\frac{d P_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \tag{24}$$

$$\frac{d P_i(t)}{dt} = \lambda_{i-1} P_{i-1}(t) - (\lambda_i + \mu_i) P_i(t) + \mu_{i+1} P_{i+1}(t) \quad i = 1,2,... \tag{25}$$

which are known as the **Chapman-Kolmogorov equations.**

Now, if we are in steady-state, the state probabilities do not depend on time; i.e., $P_i(t) = P_i \qquad \forall i; \forall t$ and $\frac{d\,P_i(t)}{dt} = 0 \qquad \forall i; \forall t$. Therefore, we can write

$$0 = -\lambda_0 P_0 + \mu_1 P_1 \tag{26}$$

$$0 = \lambda_{i-1} P_{i-1} - (\lambda_i + \mu_i) P_i + \mu_{i+1} P_{i+1} \qquad i = 1,2,... \tag{27}$$

Solving (26) for P_1 in terms of P_0

$$\boxed{P_1 = \frac{\lambda_0}{\mu_1} P_0} \tag{28}$$

Let us now write (27) for *i=1*.

$$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1 = (\lambda_1 + \mu_1) \frac{\lambda_0}{\mu_1} P_0 = \frac{\lambda_0 \lambda_1}{\mu_1} P_0 + \lambda_0 P_0$$

or

$$\mu_2 P_2 = \frac{\lambda_0 \lambda_1}{\mu_1} P_0 \text{ or } \boxed{P_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0} \tag{29}$$

Let us generally assume that

$$\boxed{P_j = \frac{\prod_{i=0}^{j-1} \lambda_i}{\prod_{i=1}^{j} \mu_i} P_0} \tag{30}$$

We can now verify this by showing that it holds for *j+1*

$$\begin{aligned} \mu_{j+1} P_{j+1} &= (\lambda_j + \mu_j) P_j - \lambda_{j-1} P_{j-1} \\ &= (\lambda_j + \mu_j) \frac{\lambda_{j-1}}{\mu_j} P_{j-1} - \lambda_{j-1} P_{j-1} \\ &= \frac{\lambda_{j-1} \lambda_j}{\mu_j} P_{j-1} \end{aligned}$$

So

$$
\begin{aligned}
P_{j+1} &= \frac{\lambda_{j-1}\lambda_j}{\mu_j \mu_{j+1}} P_{j-1} \\
&= \frac{\lambda_{j-1}\lambda_j}{\mu_j \mu_{j+1}} \frac{\prod_{i=0}^{j-1} \lambda_i}{\prod_{i=1}^{j-1} \mu_i} P_0 \\
&= \frac{\prod_{i=0}^{j} \lambda_i}{\prod_{i=1}^{j+1} \mu_i} P_0
\end{aligned}
$$

Q.E.D.

Equation (30) combined with the condition

$$\sum_{j=0}^{\infty} P_j = 1 \tag{31}$$

enables us to find all the state probabilities. From these, additional quantities of interest may be found. For example,

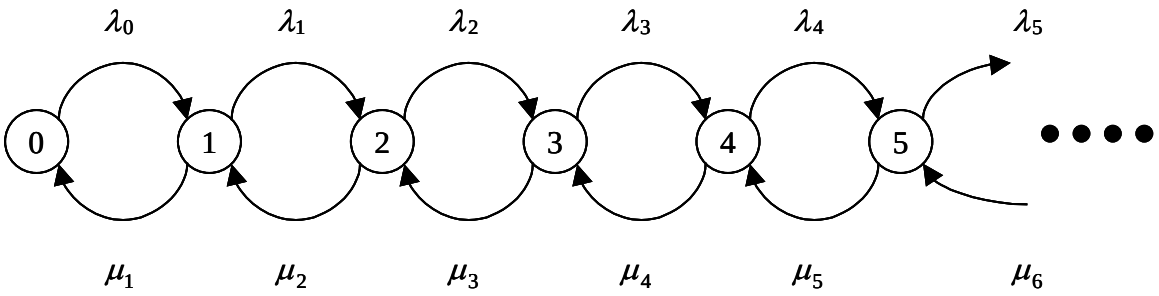
$$L = \sum_{j=0}^{\infty} j\, P_j \tag{32}$$

and

$$L_q = \sum_{j=s}^{\infty} (j-s) P_j \tag{33}$$

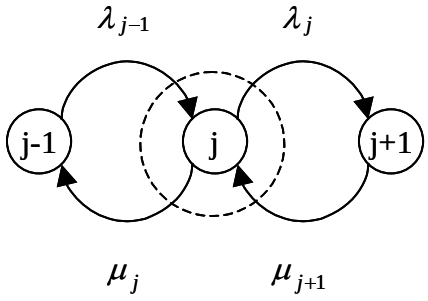
where s is the number of servers.

Before discussing some of the more common examples of the use of equation (30), we will show two alternate approaches to developing the **steady-state balance equations** (26) and (27). Consider the following state-diagram.



In which λ_j and μ_j may be thought of as the rates at which we move from state *j* upward or downward respectively.

In steady state, the rate at which probability flux (if you will) leaves state *j* must equal the rate at which probability flux enters state *j* If we now isolate state *j* (*j=1, 2,* …), we have



The rate of probability flux out of the dashed circle must equal the rate in

Rate out = rate in (34)

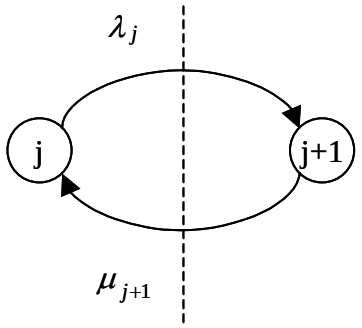
$$(\lambda_j + \mu_j)P_j = \lambda_{j-1}P_{j-1} + \mu_{j+1}P_{j+1} \qquad j = 1,2,... \tag{35}$$

which is identical to equation (27). Similarly, for state 0, we have

$$\lambda_0 P_0 = \mu_1 P_1 \tag{36}$$

which is identical to equation (26).

An alternative approach is to consider an imaginary slice between states *j* and *j+1* (*j=0,1,…).*



In steady state, the net rate of probability flux across any such cut must be zero. This implies,

Rate to the right = Rate to the left (37)

Or

$$\lambda_j P_j = \mu_{j+1} P_{j+1} \tag{38}$$

But (38) simply says

$$P_{j+1} = \frac{\lambda_j}{\mu_{j+1}} P_j \qquad j = 0,1,... \tag{39}$$

which aggress with (30).

5. Examples

M/M/1 Queue

Consider a queue with a single server, Poisson arrivals (independent of the state of the system) and Exponential service times, also state independent. That is,

$$\begin{aligned} \lambda_i &= \lambda \qquad \forall i \\ \mu_i &= \mu \qquad \forall i \end{aligned} \tag{40}$$

Then

$$P_j = \frac{\lambda^j}{\mu^j} P_0 = \rho^j P_0 \qquad \forall j \tag{41}$$

where

$$\boxed{\rho = \frac{\lambda}{\mu} = \text{utilization ratio}} \tag{42}$$

Combining equations (14) and (31) we have

$$\sum_{j=0}^{\infty} P_0 \rho^j = 1 \text{ or } P_0 \sum_{j=0}^{\infty} \rho^j = 1 \tag{43}$$

For $\rho < 1$ we have

$$\sum_{j=0}^{\infty} \rho^j = \frac{1}{1-\rho} \tag{44}$$

or

$$\boxed{\begin{aligned} &P_0 = 1-\rho \\ &P_j = (1-\rho)\rho^j \qquad\qquad j = 0,1,2,... \end{aligned}} \tag{45}$$

These are the **M/M/1 Queue State Probabilities**. Note that this is a Geometric Distribution. We therefore have:

$$L = \sum_{j=0}^{\infty} j\, P_j = P_0 \sum_{j=0}^{\infty} j\, \rho^j = (1-\rho)\sum_{j=0}^{\infty} j\, \rho^j \tag{46}$$

For $\rho < 1$ we have

$$\begin{aligned} \sum_{j=0}^{\infty} j\, \rho^j &= \rho + 2\rho^2 + \cdots \\ &= \frac{\rho}{(1-\rho)^2} \end{aligned} \tag{47}$$

So

$$\boxed{L = \frac{\rho}{1-\rho}} \tag{48}$$

$$\boxed{W = \frac{L}{\lambda} = \frac{1}{\lambda}\frac{\lambda}{\mu}\frac{1}{1-\rho} = \frac{1}{\mu(1-\rho)}} \tag{49}$$

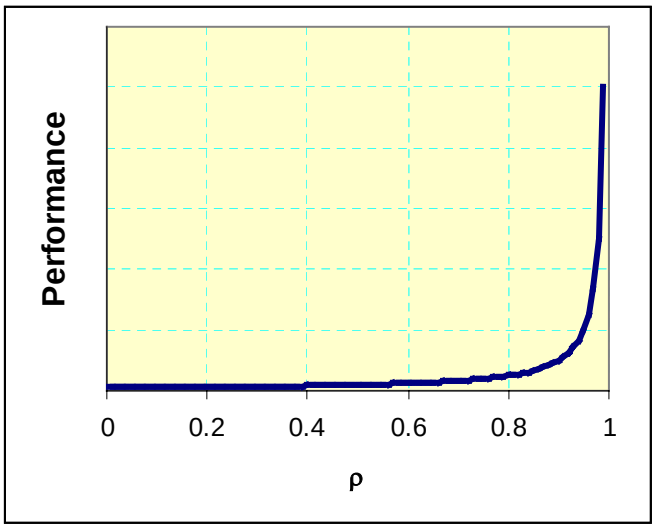
$$W_q = W - \frac{1}{\mu} = \frac{1}{\mu(1-\rho)} - \frac{1-\rho}{\mu(1-\rho)} = \frac{\rho}{\mu(1-\rho)} \tag{50}$$

$$L_q = \lambda W_q = \frac{\rho^2}{1-\rho} \tag{51}$$

The condition $\rho < 1$, needed for equations (44) and (47) is known as the **steady-state condition**. It says that the arrival rate must be strictly less than the service rate if steady-state conditions are to hold. If $\rho \geq 1$, the queue grows without bound.

Note that equations (48)-(51) all are proportional to $\frac{1}{1-\rho}$. As the utilization ratio, ρ, approaches 1.0, performance gets **very** bad. This is illustrated below:

For example, consider a single toll booth that can serve 360 people (cars) per hour, or one every 10 seconds, on average. We have the values shown in the table below. Note that if the arrival rate is 300 cars per hour (5/6 of the capacity) it takes a minute to get through the toll. If the arrival rate increase by 10 percent, it takes 2 minutes. At 345 cars per hour, it takes 4 minutes to get through the toll. Can you hear the complaints to the state government yet?



arrivals per hour			Times in seconds		
λ	ρ	L	W	W_q	L_q
120	0.333	0.5	15	5	0.167
180	0.500	1	20	10	0.500
240	0.667	2	30	20	1.333
270	0.750	3	40	30	2.250
300	0.833	5	60	50	4.167
330	0.917	11	120	110	10.083
345	0.958	23	240	230	22.042

Now let us compute the variance of the number in the system, denoted by $Var(N)$

$$Var(N) = \sum_{j=0}^{\infty} j^2 P_j - L^2 \tag{52}$$

and

$$\begin{aligned}
\sum_{j=0}^{\infty} j^2 P_j &= (1-\rho)\sum_{j=0}^{\infty} j^2 \rho^j \\
&= (1-\rho)\left\{\rho^2 \sum_{j=0}^{\infty} j(j-1)\rho^{j-2} + \sum_{j=0}^{\infty} j\,\rho^j\right\} \\
&= (1-\rho)\left\{\rho^2 \sum_{j=0}^{\infty} \frac{d^2 \rho^j}{d\,\rho^2} + \frac{\rho}{(1-\rho)^2}\right\} \\
&= (1-\rho)\left\{\rho^2 \frac{d^2 \sum_{j=0}^{\infty} \rho^j}{d\,\rho^2} + \frac{\rho}{(1-\rho)^2}\right\} \\
&= (1-\rho)\left\{\rho^2 \frac{d^2\left(\frac{1}{1-\rho}\right)}{d\,\rho^2} + \frac{\rho}{(1-\rho)^2}\right\} \\
&= (1-\rho)\left\{\frac{2\rho^2}{(1-\rho)^3} + \frac{\rho(1-\rho)}{(1-\rho)^3}\right\} \\
&= \frac{\rho+\rho^2}{(1-\rho)^2}
\end{aligned} \tag{53}$$

so

$$\begin{aligned}
Var(N) &= \sum_{j=0}^{\infty} j^2 P_j - L^2 \\
&= \frac{\rho+\rho^2}{(1-\rho)^2} - \left[\frac{\rho}{1-\rho}\right]^2 \\
&= \frac{\rho}{(1-\rho)^2}
\end{aligned} \tag{54}$$

So $Var(N)$ goes up even faster than does L as ρ approaches 1.0 To continue with the example, we get

Note that the standard deviation of the number in the system is approximately equal to the number in the system itself.

λ	ρ	L	Var(N)	STD(N)
120	0.333	0.5	0.750	0.866
180	0.500	1	2.000	1.414
240	0.667	2	6.000	2.449
270	0.750	3	12.000	3.464
300	0.833	5	30.000	5.477
330	0.917	11	132.000	11.489
345	0.958	23	552.000	23.495

Now let us compute the probability of finding *m* or more people in the system.

$$\begin{aligned} P(m \text{ or more in system}) &= \sum_{j=0}^{\infty} (1-\rho)\rho^j \\ &= (1-\rho)\frac{\rho^{\mathrm{m}}}{(1-\rho)} \\ &= \rho^{\mathrm{m}} \end{aligned} \tag{55}$$

and in particular

$$P(1 \text{ or more}) = P(wait) = \rho \tag{56}$$

We can also compute the probability density function of the time in the system. If you arrive to find *m* people ahead of you, which occurs with probability $(1-\rho)\rho^m$, your time in the system will be the sum of *m+1* independent, identically distributed (iid) Exponential random variables, or will have an **Erlang m+1** distribution.[1] The moment generating function (MGF) of an Erlang m+1 distribution (with parameter λ) is

---

[1] This derivation and that which follows for the waiting time distribution assume **a first come first served** (FCFS) queue discipline. The moment generating function of a random variable X is given by $E(e^{sX})$ where *s* is a parameter of the MGF **which should not be confused with the *s* used below for the number of servers**. For the exponential distribution we have $E(e^{sX}) = \int_0^{\infty} e^{sx}\,\lambda\, e^{-\lambda x}\,dx = \frac{\lambda}{\lambda - s}\int_0^{\infty}(\lambda - s)e^{-(\lambda - s)x}\,dx = \frac{\lambda}{\lambda - s}$. There is a one-to-one relationship between a moment generating function and a probability mass function. Knowing one automatically gives you the other. If $T = X + Y$ and *X* and *Y* are independent random variables then $E(e^{sT}) = E(e^{s(X+Y)}) = E(e^{sX})E(e^{sY})$. So the MGF of the sum of independent random variables is the product of the MGFs of the individual random variables. Thus, the MGF of the Erlang m+1 distribution is the product of *m+1* MGFs of exponential distributions. Finally, we note that since $e^{sX} = 1 + (sX) + \frac{(sX)^2}{2!} + \frac{(sX)^3}{3!} + \frac{(sX)^4}{4!} + \cdots$, we have $\left.\frac{d^n MGF(X)}{d X^n}\right|_{s=0} = E(X^n)$, hence the name **moment generating function.**

$$\left(\frac{\lambda}{\lambda-s}\right)^{m+1}$$

So the unconditional MGF of the time in the system is

$$\begin{aligned}\sum_{j=0}^{\infty}(1-\rho)\rho^m\left(\frac{\mu}{\mu-s}\right)^{m+1} &= \frac{(1-\rho)\mu}{\mu-s}\sum_{j=0}^{\infty}\left(\frac{\rho\mu}{\mu-s}\right)^{m+1}\\ &= \frac{(1-\rho)\mu}{\mu-s}\frac{1}{1-\frac{\rho\mu}{\mu-s}}\\ &= \frac{(1-\rho)\mu}{\mu-s}\frac{\mu-s}{\mu-s-\rho\mu}\\ &= \frac{(1-\rho)\mu}{(1-\rho)\mu-s}\end{aligned} \tag{57}$$

But this is just the MGF of an Exponential distribution

$$\begin{aligned}f_W(w) &= (1-\rho)\mu e^{-(1-\rho)\mu w} \qquad w\ge 0\\ &= (\mu-\lambda)e^{-(\mu-\lambda)w} \qquad w\ge 0\end{aligned} \tag{58}$$

which has a mean of

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

in agreement with (49).

Similarly, if you arrive to find *m* people ahead of you, $m\ge 1$, your <u>wait</u> time has an Erlang-m distribution. The probability of *m* people ahead of you, **conditional** on $m\ge 1$ is $\frac{P_j}{P(\text{1 or more in the system})} = \frac{P_j}{\rho} = (1-\rho)\rho^{j-1}, \quad j=1,2,...$ So the MGF of the waiting time **conditional** on at least one ahead of you is

$$\sum_{j=1}^{\infty}(1-\rho)\rho^{j-1}\left(\frac{\mu}{\mu-s}\right)^{j} = (1-\rho)\sum_{j=0}^{\infty}\rho^{j}\left(\frac{\mu}{\mu-s}\right)^{j+1} = \frac{(1-\rho)\mu}{(1-\rho)\mu-s}$$

as before. Therefore, the **unconditional probability distribution of waiting time** is

$$f_{W_q}(w_q) = \begin{cases} 1-\rho & w_q = 0\\ \rho(\mu-\lambda)e^{-(\mu-\lambda)w_q} & w_q > 0\end{cases} \tag{59}$$

with a mean

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

as in (50).
Finally, let us consider the distribution of times between departures from the system. To do so, we need to consider two cases:

a) If the system is empty, the time until the next departure is the sum of an Exponential random variable with mean $1/\lambda$ (the time until the next arrival) **plus** the service time which is Exponential with mean $1/\mu$. In this case, the MGF of the time until the next departure is $\left(\frac{\lambda}{\lambda - s}\right)\left(\frac{\mu}{\mu - s}\right)$.
b) If the system is not empty, the time until the next departure is Exponential with mean $1/\mu$ and the MGF of this time is simply $\left(\frac{\mu}{\mu - s}\right)$.

So the unconditional MGF of the time until the next departure is

$$\begin{aligned}(1-\rho)\left(\frac{\lambda}{\lambda - s}\right)\left(\frac{\mu}{\mu - s}\right) + (1-\rho)\left(\frac{\mu}{\mu - s}\right)\frac{\rho}{1-\rho} &= \frac{\lambda(\mu-\lambda)}{(\lambda - s)(\mu - s)} + \frac{\lambda(\lambda - s)}{(\lambda - s)(\mu - s)} \\ &= \frac{\lambda\mu - \lambda s}{(\lambda - s)(\mu - s)} \\ &= \frac{\lambda}{\lambda - s}\end{aligned} \tag{60}$$

So the inter-departure times are Exponential or the departure process is **Poisson** with mean rate λ, just like the arrival process. Note that while it should not be surprising that the mean arrival rate and the mean departure rates are equal (since what goes into the queue in the long run must come out), what is perhaps more surprising is that the *distribution* of the departure process is the same as the *distribution* of the input process for this queue. This result is most useful in modeling a **series** of queues. Burke has shown that this holds for an M/M/s queue. In both cases, there can be no limit on the number in the system for this result to be true.

M/M/1 Queue with a restricted queue length

Suppose now that we limit the number in the system to M. In other words,

$$\lambda_n = \begin{cases} \lambda & n = 0,1,\ldots, M-1 \\ 0 & n = M, M+1,\ldots \end{cases}$$

$$\mu_n = \begin{cases} \mu & n = 1,\ldots, M \\ 0 & n = M+1,\ldots \end{cases}$$

Then

$$P_j = \rho^j P_0 \qquad j = 0,1,\ldots, M \tag{61}$$

by equation (30) and by (31) we have

$$P_0 \sum_{j=0}^{M} \rho^j = P_0 \left(\frac{1-\rho^{M+1}}{1-\rho} \right) = 1$$

so

$$P_0 = \frac{1-\rho}{1-\rho^{M+1}}$$
$$P_j = \frac{(1-\rho)\rho^j}{1-\rho^{M+1}} \quad j = 1,\ldots, M \tag{62}$$

and

$$
\begin{aligned}
L &= \sum_{j=0}^{M} j \frac{(1-\rho)\rho^j}{1-\rho^{M+1}} \\
&= \frac{(1-\rho)}{1-\rho^{M+1}} \sum_{j=0}^{M} j\,\rho^j \\
&= \frac{(1-\rho)\rho}{1-\rho^{M+1}} \frac{d\sum_{j=0}^{M} \rho^j}{d\rho} \\
&= \frac{(1-\rho)\rho}{1-\rho^{M+1}} \frac{d\left(\frac{1-\rho^{M+1}}{1-\rho}\right)}{d\rho} \\
&= \frac{(1-\rho)\rho}{1-\rho^{M+1}} \left\{ \frac{1-(M+1)\rho^M + M\,\rho^{M+1}}{(1-\rho)^2} \right\} \\
&= \frac{\rho\left[1-(M+1)\rho^M + M\,\rho^{M+1}\right]}{(1-\rho)\left(1-\rho^{M+1}\right)}
\end{aligned}
\tag{63}
$$

This is clearly messy. The table below gives values of L for various values of M and ρ.

	M					
ρ	1	2	3	10	100	10000000
0.1	0.091	0.108	0.111	0.111	0.111	0.111
0.3	0.231	0.345	0.396	0.429	0.429	0.429
0.5	0.333	0.571	0.733	0.995	1.000	1.000
0.7	0.412	0.767	1.069	2.111	2.333	2.333
0.9	0.474	0.930	1.369	3.969	8.998	9.000

As ρ goes up, the average number in the system goes up. As M goes up, L approaches that of an *M/M/1* queue (as shown in the final column) with no restriction on the queue length. For small values of ρ, the limit on the number in the system has little effect; as ρ goes up, the effect of M increases.

To compute W, we need the **effective** arrival rate, λ_{eff}. Note that when there are M customers in the system, any additional arrivals are tuned away or lost. So,

$$\begin{aligned}\lambda_{eff} &= \lambda\left(1-P_M\right)=\lambda\left[1-\frac{\left(1-\rho\right)\rho^M}{1-\rho^{M+1}}\right]\\ &=\lambda\left[\frac{1-\rho^{M+1}-\rho^M+\rho^{M+1}}{1-\rho^{M+1}}\right]\\ &=\lambda\left[\frac{1-\rho^M}{1-\rho^{M+1}}\right]\end{aligned} \tag{64}$$

Now

$$W=\frac{L}{\lambda_{eff}}=\frac{1-\left(M+1\right)\rho^M+M\,\rho^{M+1}}{\mu\left(1-\rho\right)\left(1-\rho^M\right)} \tag{65}$$

and

$$W_q=W-\frac{1}{\mu}\qquad\text{and}\qquad L_q=\lambda_{eff}W_q \tag{66}$$

The equations above assume that $\rho\neq 1$. For $\rho=1$, we get the following revised equations:

$$P_0=\frac{1}{M+1}$$
$$P_j=\frac{1}{M+1}\qquad j=1,\ldots,M \tag{62a}$$

$$\begin{aligned}L&=\sum_{j=0}^{M} j\frac{1}{M+1}\\ &=\frac{1}{M+1}\sum_{j=0}^{M} j\\ &=\frac{1}{M+1}\frac{M\left(M+1\right)}{2}\\ &=\frac{M}{2}\end{aligned} \tag{63a}$$

$$\begin{aligned}\lambda_{eff} &= \lambda\left(1-P_M\right)=\lambda\left[1-\frac{1}{M+1}\right]\\ &=\lambda\left[\frac{M}{M+1}\right]\end{aligned} \tag{64a}$$

$$W=\frac{L}{\lambda_{eff}}=\frac{M}{2}\left(\frac{M+1}{M}\bullet\frac{1}{\lambda}\right)=\frac{1}{\lambda}\frac{M+1}{2} \tag{65a}$$

$$W_q=W-\frac{1}{\mu}=\frac{1}{\lambda}\left(\frac{M+1}{2}\right)-\frac{1}{\mu}\qquad\text{and}\qquad L_q=\lambda_{eff}W_q=\frac{M}{2}-\frac{M}{M+1} \tag{66a}$$

M/M/s Queue

We now consider a queue with Poisson arrivals (independent of the state) *s* identical servers each operating with an Exponential service time distribution and no limit on the queue lengths. We assume a single queue (see the diagram below).

This yields

$$\lambda_n = \lambda$$
$$\mu_n = \begin{cases} n\mu & n = 1,2,...,s \\ s\mu & n = s, s+1,... \end{cases} \tag{67}$$

Substituting into equation (30), we get

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & n = 0,1,...,s \\ \dfrac{(\lambda/\mu)^n}{s!\, s^{n-s}} P_0 & n = s+1,... \end{cases} \tag{68}$$

and

$$P_0 = \left[\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^s}{s!} \sum_{j=1}^{\infty} \left(\frac{\lambda}{s\mu} \right)^j \right]^{-1}$$
$$= \left[\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^s}{s!} \frac{\lambda}{s\mu - \lambda} \right]^{-1} \tag{69a}$$

$$P_0 = \left[\sum_{j=0}^{s-1} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^s}{s!} \sum_{j=0}^{\infty} \left(\frac{\lambda}{s\mu} \right)^j \right]^{-1}$$
$$= \left[\sum_{j=0}^{s-1} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^s}{s!} \frac{s\mu}{s\mu - \lambda} \right]^{-1} \tag{69b}$$

for

$$\frac{\lambda}{s\mu} < 1, \text{ the } \textbf{steady state condition} \tag{70}$$

We can also derive

$$P(wait) = \sum_{j=s}^{\infty} P_j = \frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu-\lambda}\right)P_0 \tag{71}$$

$$L_q = \frac{\lambda\mu}{(s-1)!}\frac{(\lambda/\mu)^s}{(s\mu-\lambda)^2}P_0 \tag{72}$$

and again

$$W_q = \frac{L_q}{\lambda} \tag{73a}$$

$$W = W_q + \frac{1}{\mu} \tag{73b}$$

$$L = \lambda W \tag{73c}$$

Note that if *s=1*, we have

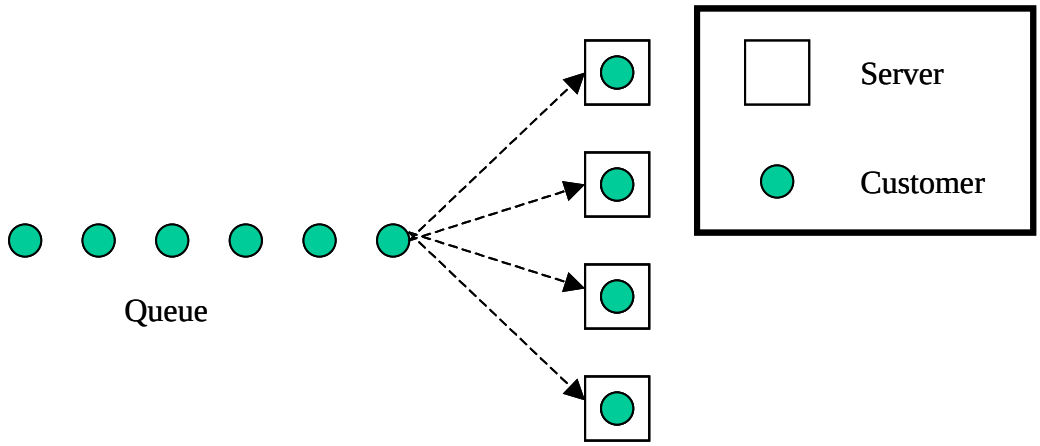
$$P_0 = \left[1+\frac{\lambda}{\mu}\frac{\mu}{\mu-\lambda}\right]^{-1} = \left[\frac{\mu-\lambda+\lambda}{\mu-\lambda}\right]^{-1} = 1-\frac{\lambda}{\mu}$$

and

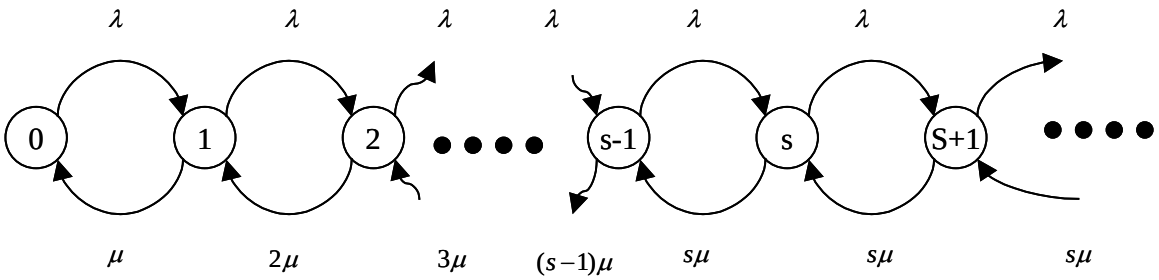
$$L_q = \frac{\lambda\mu\left(\frac{\lambda}{\mu}\right)}{(\mu-\lambda)^2}\frac{\mu-\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\lambda^2/\mu^2}{(\mu-\lambda)/\mu} = \frac{\rho^2}{1-\rho}$$

which agree with (45) and (51) respectively.

Note that the configuration assumed here is given by the figure below



The state transition diagram is given by



Consider, for example, an airport check-in line with 4 servers. Let $\frac{1}{\mu} = 2 \min$, so $\mu = 0.5$ and we require $\lambda < 2$ or less than 120 passengers per hour arriving for checkin at the four servers. The following table gives the results for this case

λ (per min)	λ (per hr)	μ	P_0	L_q	W_q	W	L
0.5	30	0.5	0.367	0.007	0.014	2.014	1.007
1	60	0.5	0.130	0.174	0.174	2.174	2.174
1.5	90	0.5	0.038	1.528	1.019	3.019	4.528
1.75	105	0.5	0.015	5.165	2.951	4.951	8.665
1.9	114	0.5	0.005	16.937	8.914	10.914	20.737
1.95	117	0.5	0.002	36.859	18.902	20.902	40.759

Now suppose $\lambda = 1.95$. Consider adding servers

s	λ (per min)	λ (per hr)	μ	L_q	W_q	W	L	L
4	1.95	117	0.5	0.002	36.859	18.902	20.902	40.759
5	1.95	117	0.5	0.015	1.830	0.939	2.939	5.730
6	1.95	117	0.5	0.019	0.485	0.249	2.249	4.385

Note that

$$\rho = \frac{\lambda}{s\mu} = P(\text{randomly selected server is busy})$$
$$= \sum_{j=0}^{s} \frac{j P_j}{s} + \sum_{j=s+1}^{\infty} P_j$$

In the *M/M/s* case we have

$$\sum_{j=0}^{s}\frac{j\,P_j}{s}+\sum_{j=s+1}^{\infty}P_j=\frac{P_o}{s}\sum_{j=0}^{s}\frac{(\lambda/\mu)^j}{(j-1)!}+P_0\frac{(\lambda/\mu)^s}{s!}\left(\frac{\lambda}{s\mu-\lambda}\right)$$

$$=\frac{P_o}{s}\frac{\lambda}{\mu}\left\{\sum_{j=0}^{s-1}\frac{(\lambda/\mu)^j}{j!}+\frac{(\lambda/\mu)^{s-1}}{(s-1)!}\left(\frac{\lambda}{s\mu-\lambda}\right)\frac{\lambda/\mu}{s}\frac{s\mu}{\lambda}\right\}$$

$$=\frac{P_o}{s}\frac{\lambda}{\mu}\left\{\sum_{j=0}^{s-1}\frac{(\lambda/\mu)^j}{j!}+\frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu-\lambda}\right)\right\}$$

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

So $\rho=\frac{\lambda}{s\mu}$ which is the **utilization ratio** or the **fraction of time each server is busy.**
Note that this is **not equal to** the probability that all servers are busy. This is given by

$$P(\text{all servers are busy})=\sum_{j=s}^{\infty}P_j=\frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu-\lambda}\right)P_0$$

M/M/∞ queue, or the Self-Service Queue

In this case, one never has to wait. We have

$$\begin{aligned}&\lambda_n=\lambda && n=0,1,2,...\\ &\mu_n=n\mu && n=1,2,...\end{aligned}\qquad(74)$$

and

$$P_j=\frac{(\lambda/\mu)^j}{j!}P_0 \qquad j=0,1,...$$

$$P_0\sum_{j=0}^{\infty}\frac{(\lambda/\mu)^j}{j!}=P_0e^{\lambda/\mu}=1$$

so

$$P_0=e^{-\lambda/\mu}=e^{-\rho}\qquad(75)$$

and

$$P_j = \frac{\rho^j e^{-\rho}}{j!} \qquad j = 0,1,2,... \tag{76}$$

which is just the **Poisson** distribution. So,

$$L = \rho = \frac{\lambda}{\mu} \tag{77}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu} \qquad \text{(as expected)} \tag{78}$$

and

$$W_q = L_q = 0 \tag{79}$$

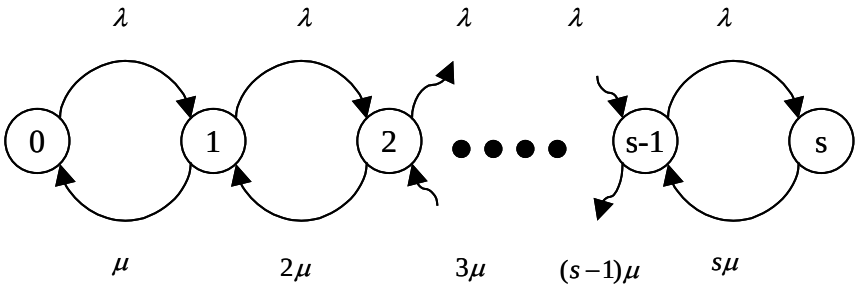
As it turns out, (76) holds for the *M/G/∞* queue as well (when we have any general service time distribution with a finite mean and variance and the service times are independent and identically distributed). The *M/G/∞* results are useful when queuing delays are unlikely, as in the case of an EMS (emergency medical services) system.

M/M/s Queue with no waiting room

Now consider an *M/M/s* queue in which there is no room to wait. Here we have:

$$\begin{aligned} \lambda_n &= \lambda \qquad n = 0,1,2,...s-1 \\ \mu_n &= n\mu \qquad n = 1,2,...s \end{aligned} \tag{80}$$

and the steady state transition diagram looks like



And

$$P_j = \frac{(\lambda/\mu)^j}{j!} P_0 \qquad j = 0,1,\ldots, s \tag{81}$$

with

$$P_0 = \frac{1}{\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!}} \tag{82}$$

so

$$P_j = \frac{(\lambda/\mu)^j}{j!} \left[\frac{1}{\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!}} \right] \qquad j = 0,1,\ldots, s \tag{83}$$

from which L may be computed numerically. Alternatively, we can derive

$$\begin{aligned}
L &= \sum_{j=0}^{s} j\, P_j \\
&= \frac{1}{\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!}} \sum_{j=0}^{s} j \frac{(\lambda/\mu)^j}{j!} \\
&= \frac{1}{\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!}} \sum_{j=1}^{s} j \frac{(\lambda/\mu)^j}{j!} \\
&= \frac{1}{\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!}} \sum_{j=1}^{s} \frac{(\lambda/\mu)^j}{(j-1)!} \\
&= \frac{1}{\sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!}} \rho \sum_{j=1}^{s-1} \frac{\rho^j}{j!} \\
&= \frac{\rho \sum_{j=1}^{s-1} \frac{\rho^j}{j!}}{\sum_{j=0}^{s} \frac{\rho^j}{j!}}
\end{aligned} \tag{84}$$

Again, to compute *W*, we need λ_{eff} which is given by

$$\lambda_{eff} = \frac{\lambda \sum_{j=1}^{s-1} \frac{\rho^j}{j!}}{\sum_{j=0}^{s} \frac{\rho^j}{j!}} \tag{85}$$

and as expected

$$\begin{aligned} W &= \frac{L}{\lambda_{eff}} = \frac{1}{\mu} \\ W_q &= 0 \\ L_q &= 0 \end{aligned} \tag{86}$$

For example, consider a parking lot with 10 spaces and an average stay of $\frac{1}{\mu} = 2 \text{ hours}$ so $\mu = 0.5$. In this case we obtain:

λ (per hr)	L	P(full)
1	2.00	3.82E-05
2	3.98	0.005
3	5.74	0.043
5	7.85	0.215
7	8.72	0.377
10	9.24	0.538
20	9.69	0.758

Note that we do not require $\lambda < s\mu = 5$ since we have a finite queue. Also note that $\lambda_{eff} = \lambda\{1 - P(full)\}$.

Again, it turns out that equations (83) and (84) and (85) hold for an *M/G/s* queue with no waiting area, where $\frac{1}{\mu} = E(\text{service time})$. Note that as $s \to \infty$ we get the results for the *M/G/*∞ queue, as expected.

Equation (83) is known as **Erlang's Loss Formula**