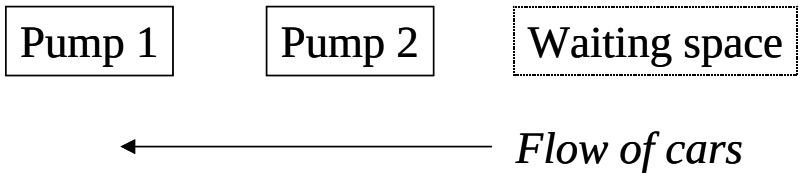
Additional Notes on Queuing Theory

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

1. Example of a more complex problem (a classic example)

Consider a gas station. Let us limit that the number of cars that may wait to one car. Assume that the station has 2 pumps as shown below. If a car arrives to find:



a) No one in the station, it goes to pump 1
b) Pump 2 occupied, it waits if there is room. Note that we assume that it cannot swing around the car at pump 2 even if pump 1 is empty due to space limitations.
c) Pump 1 busy and pump 2 idle, it goes to pump 2
d) Someone else waiting, it is lost (like a call to a busy phone line with no answering system).

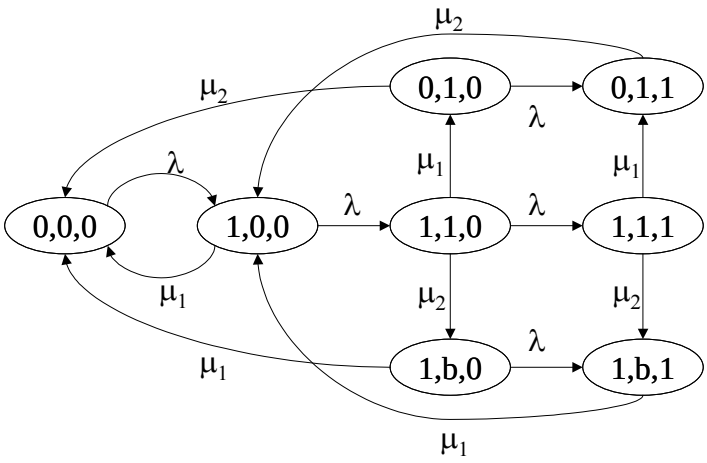
We assume that arrivals occur according to a Poisson process with rate λ, and that service times at the pumps are Exponentially distributed with rates μ_1 and μ_2, respectively.

From the above, we see that pump 1 may be in two states: occupied and empty. Pump 2 may be in one of 3 states: empty, occupied and pumping, and occupied and blocked (the car is finished, but cannot leave the station because a car is still at pump 1). Finally, the waiting space may be occupied or empty. Let us define a 3-dimensional state space described by

(state of pump 1, state of pump 2, number waiting)

So, state (0,1,0) indicates that pump 2 is occupied and pumping. State (1,b,1) indicates that pump 1 is occupied, pump 2 is blocked and one person is waiting. Some of the 2x3x2 = 12 states are not possible, e.g., (0,0,1), (0,b,0), (0,b,1), (1,0,1). **Why?**

Our state diagram now is shown below. Our steady state equations are:



Our steady state balance equations are

$$\begin{aligned}
\lambda P_{000} &= \mu_1 P_{100} + \mu_1 P_{1b0} + \mu_2 P_{010} \\
(\lambda + \mu_1) P_{100} &= \lambda P_{000} + \mu_1 P_{1b1} + \mu_2 P_{011} \\
(\lambda + \mu_1 + \mu_2) P_{110} &= \lambda P_{100} \\
(\mu_1 + \mu_2) P_{111} &= \lambda P_{110} \\
(\lambda + \mu_2) P_{010} &= \mu_1 P_{110} \\
\mu_2 P_{011} &= \lambda P_{010} + \mu_1 P_{111} \\
(\lambda + \mu_1) P_{1b0} &= \mu_2 P_{110} \\
\mu_1 P_{1b1} &= \lambda P_{1b0} + \mu_2 P_{111}
\end{aligned}$$

Now

$$P_{100} = \frac{\lambda + \mu_1 + \mu_2}{\lambda} P_{110}$$

$$P_{111} = \frac{\lambda}{\mu_1 + \mu_2} P_{110}$$

$$P_{010} = \frac{\mu_1}{\lambda + \mu_2} P_{110}$$

$$P_{1b0} = \frac{\mu_2}{\lambda + \mu_1} P_{110}$$

$$P_{011} = \frac{\lambda}{\mu_2} P_{010} + \frac{\mu_1}{\mu_2} P_{111} = \left[\frac{\lambda}{\mu_2} \left(\frac{\mu_1}{\lambda + \mu_2} \right) + \frac{\mu_1}{\mu_2} \left(\frac{\lambda}{\mu_1 + \mu_2} \right) \right] P_{110}$$

$$P_{1b1} = \frac{\mu_2}{\mu_1} P_{111} + \frac{\lambda}{\mu_1} P_{1b0} = \left[\frac{\mu_2}{\mu_1} \left(\frac{\lambda}{\mu_1 + \mu_2} \right) + \frac{\lambda}{\mu_1} \left(\frac{\mu_2}{\lambda + \mu_1} \right) \right] P_{110}$$

$$P_{000} = \frac{\mu_1}{\lambda} P_{100} + \frac{\mu_1}{\lambda} P_{1b0} + \frac{\mu_2}{\lambda} P_{010} = \left[\frac{\mu_1}{\lambda} \left(\frac{\lambda + \mu_1 + \mu_2}{\lambda} \right) + \frac{\mu_1}{\lambda} \left(\frac{\mu_2}{\lambda + \mu_1} \right) + \frac{\mu_2}{\lambda} \left(\frac{\mu_1}{\lambda + \mu_2} \right) \right] P_{110}$$

All of which are in terms of P_{110}, which becomes

$$P_{110} = \frac{1}{\left\{ \begin{array}{l} \left(\frac{\lambda + \mu_1 + \mu_2}{\lambda} \right) \left(\frac{\lambda + \mu_1}{\lambda} \right) + \left(\frac{\mu_2}{\lambda + \mu_1} \right) \left(\frac{\lambda \mu_1 + \mu_1^2 + \lambda^2}{\lambda \mu_1} \right) + 1 + \\ \left(\frac{\mu_1}{\lambda + \mu_2} \right) \left(\frac{\lambda \mu_2 + \mu_2^2 + \lambda^2}{\lambda \mu_2} \right) + \left(\frac{\lambda}{\mu_1 + \mu_2} \right) \left(\frac{\mu_1 \mu_2 + \mu_2^2 + \mu_1^2}{\mu_1 \mu_2} \right) \end{array} \right\}}$$

Consider the case in which $\lambda = 10, \quad \mu_1 = \mu_2 = 6$. We find

$$\left. \begin{array}{l} P_{110} = 0.1056 \\ P_{100} = 0.2323 \\ P_{111} = 0.0880 \\ P_{010} = 0.0396 \\ P_{1b0} = 0.0396 \\ P_{011} = 0.1540 \\ P_{1b1} = 0.1540 \\ P_{000} = 0.1869 \end{array} \right\} \text{ and }$$

$$P(\text{no one in system}) = P_{000} = 0.1869$$

$$P(1 \text{ in system}) = P_{100} + P_{010} = 0.2719$$

$$P(2 \text{ in system}) = P_{110} + P_{1b0} + P_{011} = 0.2992$$

$$P(3 \text{ in system}) = P_{111} + P_{1b1} = 0.2420$$

$$P(\text{blocked}) = P_{1b0} + P_{1b1} = 0.1936$$

$$P(\text{arrival cannot join queue}) = P_{111} + P_{011} + P_{1b1} = 0.3960$$

$$\text{Nominal arrival rate} = 10(1 - 0.3960) = 6.040$$

$$\text{Average number in system} = 1.5963$$
$$\text{Average number of occupied pumps} = P_{010} + P_{100} + P_{011} + 2\left(P_{110} + P_{111} + P_{1b0} + P_{1b1}\right) = 1.2003$$
$$\text{Average time in system} = \frac{1.5963}{6.040} = 0.2643$$

If λ and μ are measured in terms of hours, this comes to 15.86 minutes (which clearly is rather long).

Now suppose our station owner can invest in **<u>one</u>** faster pump to replace an existing one. The new pump has $\mu = 8$. She wants to place the pump at the location (pump 1 or pump 2) that maximizes his throughput or maximizes λ_{eff} or minimizes $P(\text{arrival cannot join queue})$. **Why is this a good measure for the pump owner to use?** For

$\lambda = 10 \quad \mu_1 = 8 \quad \mu_2 = 6$, we get $P(\text{arrival cannot join queue}) = 0.3394$
$\lambda = 10 \quad \mu_1 = 6 \quad \mu_2 = 8$, we get $P(\text{arrival cannot join queue}) = 0.3562$

So she should replace **<u>pump 1</u>** with the faster pump. **Justify this in your own mind.**

2. The M/G/1 queue

In this queue, we must be careful. We lose the Memoryless property of the Exponential service time distribution. To preserve the Markovian analysis, let us examine the queue only after (immediately after) a <u>departure</u>. In so doing, we do not need to be concerned with how long the person being served has been in service. These points in time are called <u>renewal</u> points (in the stochastic processes literature) because the system renews itself (in some sense) at these points in time. It turns out that the average number in the system just after a departure equals the average number in the system at a randomly selected point in time.

To analyze the queue, we will use a "trick" that applies or works well only in this case. A more rigorous approach involves a more complete understanding of stochastic processes and a <u>very</u> good familiarity with probability generating functions (PGFs) and moment generating functions (MGFs). (See Gross and Harris or Clarke and Disney.)

Let

N_i = number of people in the system just after the i - th departure
R_i = number of people to arrive during the i - th service time

Then if

$$N_i > 0, \;\; N_{i+1} = N_i + R_{i+1} - 1$$
$$N_i = 0, \;\; N_{i+1} = R_{i+1}$$

You should **justify these for yourself.**

Define

$$\delta = \begin{cases} 1 & N_i = 0 \\ 0 & N_i = 1 \end{cases}$$

Then

$$\boxed{N_{i+1} = N_i + R_{i+1} - 1 + \delta} \text{ for all } N_i \qquad (*)$$

Again, you should justify this for yourself.

In steady state, we should have $E(N_{i+1}) = E(N_i)$, so if we take the expectation of the equation above we get

$$E(N_{i+1}) = E(N_i) + E(R_{i+1}) - 1 + E(\delta)$$

or

$$E(\delta) = 1 - E(R_{i+1})$$

Now, also in steady state, $E(R_{i+1})$ should equal $E(R_i)$, or $E(R_{i+1})$ should be independent of the time notation *i+1*.

Given that the *i+1*st service time takes τ minutes, the <u>conditional</u> expected number of arrivals is $\lambda\tau$ (since the arrivals are Poisson). The <u>unconditional</u> expected number of arrivals is $\lambda E(S)$ where $E(S)$ is the mean service time. So,

$$E(\delta) = 1 - E(R_{i+1}) = 1 - \lambda E(S) = 1 - \rho$$

where we define $\rho = \lambda E(S)$ as before.

More formally, we can say

$$E\left(R_{i+1}\middle|\text{i+1}^{\text{st}}\text{ service time}=\tau\right)=\lambda\tau$$

$$E(R_{i+1})=\int \lambda\tau\, f_T(\tau)d\tau=\lambda E(S)=\frac{\lambda}{\mu}=\rho$$

$$E\left(R_{i+1}^2\middle|\text{i+1}^{\text{st}}\text{ service time}=\tau\right)=\lambda^2\tau^2+\lambda\tau$$

$$E(R_{i+1}^2)=\int\left(\lambda^2\tau^2+\lambda\tau\right)f_T(\tau)d\tau=\lambda^2E(S^2)+\lambda E(S)$$

$$=\lambda^2\left(\sigma^2+E^2(S)\right)+\lambda E(S)$$

$$=\lambda^2\sigma^2+\rho^2+\rho$$

where $\sigma^2=Var(S)$.

Finally, note that

1. $\delta^2=\delta$ so $E(\delta^2)=E(\delta)=1-\rho$
2. $\delta N=0$ so $E(\delta N)=0$

Now square both sides of (*)

$$N_{i+1}^2=N_i^2+R_{i+1}^2+1+\delta^2+2N_iR_{i+1}-2N_i+2\delta N_i-2R_{i+1}+2\delta R_{i+1}-2\delta$$

Taking expectations, noting that $E(N_{i+1})=E(N_i)=E(N)$ and $E(N_{i+1}^2)=E(N_i^2)=E(N^2)$ in steady state and N_i and R_{i+1} are independent as are δ and R_{i+1}, we get

$$E(N^2)=E(N^2)+E(R^2)+1+E(\delta^2)+2E(N)E(R_{i+1})-2E(N)+2E(\delta N)-2E(R_{i+1})+2E(\delta)E(R_{i+1})-2E(\delta)$$

Substituting we find

$$0=\lambda^2\sigma^2+\rho^2+\rho+1+1-\rho+2E(N)\rho-2E(N)-2\rho+2(1-\rho)\rho-2(1-\rho)$$

$$=\lambda^2\sigma^2+\rho^2-2\rho^2+\rho-\rho-2\rho+2\rho+2\rho+1+1-2+2E(N)[\rho-1]$$

or

$$\lambda^2\sigma^2-\rho^2+2\rho=2E(N)[1-\rho]$$

or

$$E(N) = \frac{\lambda^2\sigma^2 + \rho^2 + 2\rho - 2\rho^2}{2(1-\rho)}$$
$$= \frac{2\rho(1-\rho)}{2(1-\rho)} + \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

or

$$\boxed{E(N) = \rho + \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}}$$

This is the <u>Pollaczek-Khintchine</u> formula.

So far, this formula appears to apply only when we look at the queue just after a departure. However, we now show that

$$\pi_n = p_n$$

where

π_n = steady state probability of n in the system after a departure
p_n = steady state probability of n in the system at any randomly selected time

Following Gross and Harris (pp. 235-236), let

$A_n(t)$ be the number of arrivals in the interval $(0,t)$ when the system is in state n
$D_n(t)$ be the number of departures to state n in the interval $(0,t)$

Now since the system only moves up or down in unit steps,

$$|A_n(t) - D_n(t)| \le 1 \qquad \text{(a)}$$

That is, in the interval *(0,t)* $A_n(t)$ and $D_n(t)$ differ by <u>at most</u> 1. As long as the system is not saturated (ρ<1), then

$$\lim_{T\to\infty} \frac{D(T)}{A(T)} = \lim_{T\to\infty} \frac{A(T)}{D(T)} = 1 \qquad \text{(b)}$$

where *A(T)* and *D(T)* are the total number of arrivals and departures in time *(0,T)* respectively.

Now divide (a) by *A(T)* and take the limit to obtain

$$\lim_{T\to\infty}\left|\frac{A_n(T)}{A(T)}-\frac{D_n(T)}{A(T)}\right|\le\lim_{T\to\infty}\frac{1}{A(T)}=0\text{, since }\lim_{T\to\infty}A(T)=\infty \tag{c}$$

But (c) implies that

$$\lim_{T\to\infty}\left|\frac{A_n(T)}{A(T)}-\frac{D_n(T)}{A(T)}\right|=0$$

which in turn implies

$$\lim_{T\to\infty}\frac{A_n(T)}{A(T)}=\lim_{T\to\infty}\frac{D_n(T)}{A(T)} \tag{d}$$

Now

$$\begin{aligned}\lim_{T\to\infty}\frac{A_n(T)}{A(T)}&=\lim_{T\to\infty}\frac{D_n(T)}{A(T)}\lim_{T\to\infty}\frac{A(T)}{D(T)}\\&=\lim_{T\to\infty}\frac{D_n(T)}{A(T)}\frac{A(T)}{D(T)}\\&=\lim_{T\to\infty}\frac{D_n(T)}{D(T)}\end{aligned} \tag{e}$$

where the first equality uses equation (b). Since arrivals occur according to a Poisson process independent of time,

$$\lim_{T\to\infty}\frac{A_n(T)}{A(T)}=p_n \tag{f}$$

Also, by definition

$$\pi_n=\lim_{T\to\infty}\frac{D_n(T)}{D(T)} \tag{g}$$

Substituting (g) and (f) into (e), we get the desired result that

$$\pi_n=p_n\qquad\forall n \tag{h}$$

This implies that the Pollackek-Khintchine formula applies for <u>any</u> point in time. So, we may write

$$\boxed{L=\rho+\frac{\rho^2+\lambda^2\sigma^2}{2(1-\rho)}}$$

and Little's formula applies, so

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

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

$$L_q = \frac{\lambda^2 + \rho^2}{2(1-\rho)}$$

Using these equations, we can begin to get a feel for the **impact of variability** in the service time distribution on the performance of the queue. Consider two different queues: the *M/M/1* queue with Exponential service times and the *M/D/1* queue with deterministic service times. For the *M/M/1* queue, we have $\sigma^2 = 1/\mu^2$, while for the *M/D/1* queue we have $\sigma^2 = 0$. If we now look at the time in the queue before service, we find

$$W_q^{M/M/1} = \frac{\rho^2}{\lambda(1-\rho)}$$

$$W_q^{M/D/1} = \frac{\rho^2}{2\lambda(1-\rho)} = \frac{1}{2}W_q^{M/M/1}$$

Thus, the waiting time in the queue for the *M/M/1* queue is twice the waiting time in the queue for the *M/D/1* queue. Variability HURTS.

3. The G/G/1 Queue

In this brief section, we outline the results for an approximation of the behavior of a single server queue with any general inter-arrival time distribution and any general service time distribution. We can approximate the waiting time in the queue by:

$$W_q \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}$$

where

$\frac{1}{\mu}$ is the mean service time as before

ρ is the utilization ratio or λ/μ where λ is the arrival rate

c_a^2 is the squared coefficient of variation of the inter-arrival time distribution (the variance divided by the square of the mean)

c_e^2 is the squared coefficient of variation of the service time distribution.

Note that for the M/M/1 queue, we have $c_a^2 = c_e^2 = 1$ and so this equation boils down to $W_q = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}$ which is the exact result for the M/M/1 queue (see equation (50) of the Notes on Queuing Theory). Also note that once we have the value of the waiting time in the queue, we can get the other three key measures of queueing performance ($L, W,$ and L_q) quite readily.

4. An approximation of the M/M/s queue

Note that the equations for the M/M/s queue are rather messy and involve a rather difficult formula for computing P_0, the probability that the system is empty. We can approximate the performance of this system using the following equation:

$$W_q \approx \frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)} \bullet \frac{1}{\mu}$$

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

To see how well this approximation works, consider an M/M/5 queue with $\lambda = 500$ arrivals per hour and $\frac{1}{\mu} = 30$ seconds (or $\frac{1}{\mu} = \frac{30}{3600} = \frac{1}{120}$ hours). The exact formulas yield 22.4 seconds for the waiting time in the queue, while the approximation yields 23.0 seconds. If we consider an M/M/9 queue with twice the arrival rate, the exact formula yields 34.2 seconds while the approximation yields 34.4 seconds.

5. An approximation of the G/G/s queue

We can also approximate the G/G/s queue performance as follows:

$$W_q \approx \left(\frac{c_a^2 + c_e^2}{2}\right) \bullet \frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)} \bullet \frac{1}{\mu}$$ where all of the terms are as defined above.