

Paradoxes and Counter-Intuitive Examples in Analysis of Queues

Here we present a small set of examples from queueing theory where the results go against intuition. If you know of other examples pertaining to queueing systems, we would immensely appreciate (and acknowledge) your contribution. Comments and suggestions are absolutely welcome. We first present the example in words and give technical details subsequently.

Example 1 *Consider a queue where customers arrive one by one. There is a single server that serves each customer on a first-come-first-served basis. Define the sojourn time experienced by a customer as the time he/she spends in the system: waiting (if any) for service plus the duration of the service at the server. The server toggles between being idle for some time when there are no customers in the system and being busy for a while serving customers one after the other. Define the busy period of the server as the time spent continuously serving customers. Notice that an arbitrary busy period is never smaller than the sojourn time of any of the customers that was served during that busy period. Yet, the average busy period could be smaller than the average sojourn time!*

Technical Details. Consider a stable $M/G/1$ queue with coefficient of variation for service times greater than 1. For this queue in steady state, any customer's time in the system (sojourn time) is smaller than the duration of the busy period (consecutive time a server is busy) during which that sojourn time occurs. In steady state, however, the average sojourn time is larger than the average busy period duration. We show that result analytically (but we do not explain why).

The average busy period in a stable $M/G/1$ queue in steady-state lasts for $1/(\mu - \lambda)$ time, where λ and μ are respectively the average arrival and service rates. Also, the average time a customer spends in that $M/G/1$ queue in steady-state is $\frac{1}{\mu} + \frac{\lambda E(S^2)}{2(1 - \lambda/\mu)}$ where $E(S^2)$ is the second moment of the service time. Notice that if the service time has coefficient of variation greater than 1, then

$$1/(\mu - \lambda) \leq \frac{1}{\mu} + \frac{\lambda E(S^2)}{2(1 - \lambda/\mu)}$$

i.e., the average busy period is shorter than the average time a customer is in the system. ■

Example 2 *This is like a strange dual to Example 1 but has a much easier explanation. Consider a queue where customers arrive one by one. There is a single server that serves each customer on a first-come-first-served basis. Define idle time as a consecutive stretch of time when there are no customers in the system (notice from Example 1 that the server essentially toggles between idle times and busy periods). Also define inter-arrival time as the time between customer arrivals. Notice that an arbitrary idle time is always smaller than the inter-arrival time during which the idle period occurred. Yet, the average idle period could be as big as the average inter-arrival time!*

Technical Details. Consider a stable $M/G/1$ queue. The idle times are indeed according to an exponential distribution with parameter λ (where λ is the arrival rate). The inter-arrival times are also according to an exponential distribution with parameter λ . Hence among other things, the average idle period is equal to the average inter-arrival time. It may be possible to show for the $G/G/1$ queue where the inter-arrival times have a coefficient of variation greater than 1, that the idle times are actually longer on average than the inter-arrival times. This is very similar to the inspection paradox in Example 4 and can be explained in a similar manner. ■

Example 3 *Braess' paradox: Consider a network. Does adding extra capacity always improve the system in terms of performance? Although it appears intuitive, adding extra capacity to a network, when the moving entities selfishly choose their routes, can in some cases worsen overall performance!*

Technical Details. Consider a network with nodes A , B , C and D . There are directed arcs from A to B , B to D , A to C and C to D . Customers arrive into node A according to a Poisson process with mean rate 2λ . The customers need to reach node D and they have two paths, one through B and the other through C as shown in Figure 1. Along the arc from A to B there is a single server queue with exponentially distributed service times (and mean $1/\mu$). Likewise there is an identical queue along the arc from C to D . In addition, it takes a deterministic time of 2 units to traverse arcs AC and BD . Assume that

$$\mu > \lambda + 1.$$

In equilibrium each arriving customer to node A would select either of the two paths with equal

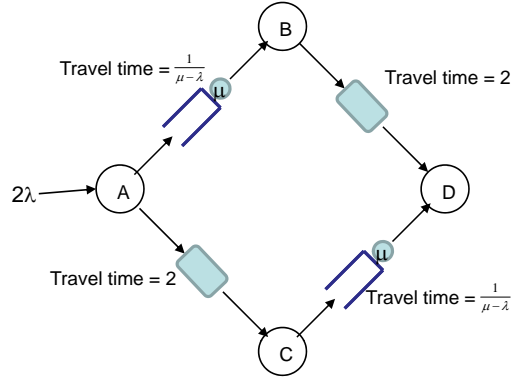


Figure 1: Travel times along arcs in equilibrium

probability. Thus the average travel time from A to D in equilibrium is

$$2 + \frac{1}{\mu - \lambda}.$$

Now a new path from B to C is constructed along which it would take a deterministic time of 1 unit to traverse. For the first customer that arrives into this new system described in Figure 2, this would be a short-cut because the new expected travel time would be $1 + \frac{2}{\mu - \lambda}$ which is smaller than the old expected travel time given above under the assumption $\mu > \lambda + 1$. Soon customers would selfishly choose their routes so that in equilibrium, all three paths $A - B - D$, $A - C - D$ and $A - B - C - D$ have identical mean travel times. Actually the equilibrium splits would not be necessary to calculate, instead notice that each of the three routes would take 3 times units to traverse on average. But the old travel time before the new capacity was added $2 + \frac{1}{\mu - \lambda}$ is actually less than 3 units under the assumption $\mu > \lambda + 1$. Thus adding extra capacity has actually worsened the average travel times! ■

Example 4 *Inspection paradox:* Consider arrivals into a queue according to a renewal process. That means that the inter-arrival times are independent and identically distributed. Let the mean and variance of the inter-arrival times be τ and σ^2 respectively. Say we observe the system at an arbitrary time in steady state. The expected value of the time remaining for the next arrival is $\frac{\sigma^2 + \tau^2}{2\tau}$. Also, the expected value of the time when the previous arrival occurred is $\frac{\sigma^2 + \tau^2}{2\tau}$ time units ago. Thus the total expected inter-arrival time wrapped around this observation is $\frac{\sigma^2 + \tau^2}{\tau}$ which is greater than the generic inter-arrival time's expected value of τ !

Technical Details. If the system is observed at time t , then define $R(t)$ as the time remaining for the next arrival, and $A(t)$ as the time since the previous arrival. From the results in renewal

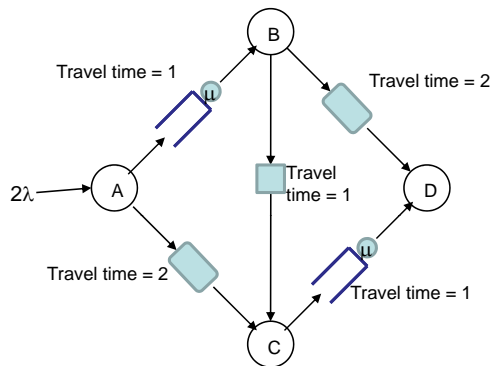


Figure 2: New travel times along arcs in equilibrium

theory it can be shown that

$$\lim_{t \rightarrow \infty} E[A(t)] = \lim_{t \rightarrow \infty} E[R(t)] = \frac{\sigma^2 + \tau^2}{2\tau}.$$

Therefore

$$\lim_{t \rightarrow \infty} E[A(t) + R(t)] = \frac{\sigma^2 + \tau^2}{\tau} \geq \tau.$$

In other words, the inter-arrival time wrapped around an observation is longer than a generic inter-arrival time. An explanation is to consider n sticks of lengths l_1, l_2, \dots, l_n that are arranged in a pipeline fashion so that we have a single line of length $\ell = l_1 + l_2 + \dots + l_n$. If we pick a point that is uniformly distributed on this line of length ℓ , then the probability the point will fall on a longer line is higher than that of a shorter line (in other words the probability that line i is selected is l_i/ℓ which is proportional to its length l_i). Our case essentially is when n grows to infinite. ■

Example 5 *Can the computation of waiting times in a queueing system depend on the model?*

Technical Details. Consider a stable queue that gets customer-arrivals externally according to a Poisson process with mean rate λ . There is a single server and infinite waiting room. The service times are exponentially distributed with mean $1/\mu$. At the end of service each customer exits the system with probability p and re-enters the queue with probability $(1 - p)$. The system is depicted in Figure 3, for now ignore A, B, C and D . We consider two models:

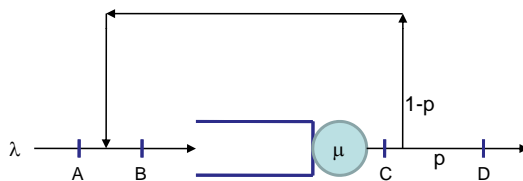


Figure 3: Points of reference

1. If the system is modeled as a birth and death process with birth rates λ and death rates $p\mu$, then $L = \frac{\lambda}{p\mu - \lambda}$ and $W = L/\lambda = \frac{1}{p\mu - \lambda}$.

2. If the system is modeled as a Jackson network with 1 node and effective arrival rate λ/p and service rate μ , then $L = \frac{\lambda/p}{\mu - \lambda/p}$ and $W = \frac{L}{\lambda/p} = \frac{p}{p\mu - \lambda}$.

Clearly, the two methods give the same L but the W values are different! Although this appears to be a paradox, that is really not the case. Let us revisit Figure 3 but now let us consider A , B , C and D . The W from the first method (birth and death model) is measured between A and D , which is the total time spent by a customer in the system (going through one or more rounds of service). The W from the second method (Jackson network model) is measured between B and C , which is the time spent by a customer from the time he/she entered the queue until one round of service is completed. Note that the customer does a geometric number of such services (with mean $1/p$). Therefore the total time spent on average would indeed be the same in either methods if we used the same points of reference, i.e. A and D . ■