Analysis of Queueing Systems with Sample Paths and Simulation

Nicky D. van Foreest

October 11, 2017

Contents

1 Introduction

Motivation and Examples Queueing systems abound, and the analysis and control of queueing systems are major topics in the control, performance evaluation and optimization of production and service systems.

At my local supermarket, for instance, any customer that joins a queue of 4 of more customers get his/her shopping for free. Of course, there are some constraints: at least one of the cashier facilities has to unoccupied by a server and the customers in queue should be equally divided over the cashiers that are open. (And perhaps there are some further rules, of which I am unaware.) The manager that controls the occupation of the cashier positions is focused on keeping $\pi(4)+\pi(5)+\cdots$, i.e., the fraction of customers that see upon arrival a queue length longer or equal than 4, very small. In a sense, this is easy enough: just hire many cashiers. However, the cost of personnel may then outweigh the yearly average cost of paying the customer penalties. Thus, the manager's problem becomes to plan and control the service capacity in such a way that both the penalties and the personnel cost are small.

Fast food restaurants also deal with many interesting queueing situations. Consider, for instance, the making of hamburgers. Typically, hamburgers are made-to-stock, in other words, they are prepared before the actual demand has arrived. Thus, hamburgers in stock can be interpreted as customers in queue waiting for service, where the service time is the time between the arrival of two customers that buy hamburgers. The hamburgers have a typical lifetime, and they have to be scrapped if they remain on the shelf longer than some amount of time. Thus, the waiting time of hamburgers has to be closely monitored. Of course, it is easy to achieve zero scrap cost, simply by keeping no stock at all. However, to prevent lost-sales it is very important to maintain a certain amount of hamburgers on stock. Thus, the manager has to balance the scrap cost against the cost of lost sales. In more formal terms, the problem is to choose a policy to prepare hamburgers such that the cost of excess waiting time (scrap) is balanced against the cost of an empty queue (lost sales).

Service systems, such as hospitals, call centers, courts, and so on, have a certain capacity available to serve customers. The performance of such systems is, in part, measured by the total number of jobs processed per year and the fraction of jobs processed within a certain time between receiving and closing the job. Here the problem is to organize the capacity such that the sojourn time, i.e., the typical time a job spends in the system, does not exceed some threshold, and such that the system achieves a certain throughput, i.e., jobs served per year.

Clearly, all the above systems can be seen as queueing systems that have to be monitored and controlled to achieve a certain performance. The performance analysis of such systems can, typically, be characterized by the following performance measures:

1. The fraction of time $p(n)$ that the system contains n customers. In particular, $1-p(0)$, i.e., the fraction of time the system contains jobs, is important, as this is a measure of the time-average occupancy of the servers, hence related to personnel cost.

2. The fraction of customers $\pi(n)$ that 'see upon arrival' the system with n customers. This measure relates to customer perception and lost sales, i.e., fractions of arriving customers

that do not enter the system.

3. The average, variance, and/or distribution of the waiting time.

4. The average, variance, and/or distribution of the number of customers in the system.

Here the system can be anything that is capable of holding jobs, such as a queue, the server(s), an entire court house, patients waiting for an MRI scan in a hospital, and so on.

It is important to realize that a queueing system can, typically, be decomposed into *two subsystems*, the queue itself and the service system. Thus, we are concerned with three types of waiting: waiting in queue, i.e., *queueing time*, waiting while being in service, i.e., the *service time*, and the total waiting time in the system, i.e., the *sojourn time*.

Organization In these notes we will be primarily concerned with making models of queueing systems such that we can compute or estimate the above performance measures. Part of our work is to derive analytic models. The benefit of such models is that they offer structural insights into the behavior of the system and scaling laws, such as that the average waiting time scales (more or less) linearly in the variance of the service times of individual customers. However, these models have severe shortcomings when it comes to analyzing real queueing systems, in particular when particular control rules have to be assessed. Consider, for example, the service process at a check-in desk of KLM. Business customers and economy customers are served by two separate queueing systems. The business customers are served by one server, server A say, while the economy class customers by three servers, say. What would happen to the sojourn time of the business customers if server A would be allowed to serve economy class customers when the business queue is empty? For the analysis of such cases simulation is a very useful and natural approach.

In the first part of these notes we concentrate on the analysis of *sample paths of a queueing process*. We assume that a typical sample path captures the 'normal' stochastic behavior of the system. This sample-path approach has two advantages. In the first place, most of the theoretical results follow from very concrete aspects of these sample paths. Second, the analysis of sample-paths carries over right away to simulation. In fact, simulation of a queueing system offers us one (or more) sample paths, and based on such sample paths we derive behavioral and statistical properties of the system. Thus, the performance measures defined for sample paths are precisely those used for simulation. Our aim is not to provide rigorous proofs for all results derived below; for the proofs and further background discussion we refer to **?**. As a consequence we tacitly assume in the remainder that results derived from the (long-run) analysis of a particular sample path are equal to their 'probabilistic counterpart'.

In the second part we construct algorithms to analyze open and closed queueing networks. Many of the sample path results developed for the single-station case can be applied to these networks. As such, theory, simulation and algorithms form a nicely round out part of work. For this part we refer to book of Prof. Zijm; the present set of notes augment the discussion there.

Exercises I urge you to make *all* exercises in this set of notes. Many exercises require many of the tools you learned previously in courses on calculus, probability, and linear algebra. Here you can see them applied. Moreover, many of these tools will be useful for other, future, courses. Thus, the investments made here will pay off for the rest of your (student) career. Moreover, the exercises are meant to *illustrate* the material and to force you to *think* about it. Thus, the main text does not contain many examples; the exercises form the examples.

You'll notice that many of these problems are quite difficult, often not because the problem itself is difficult, but because you need to combine a substantial amount of knowledge all at the same time. All this takes time and effort. Next to this, I did not include the exercises with the intention that you would find them easy. The problems should be doable, but hard.

The solution manual is meant to prevent you from getting stuck and to help you increase your knowledge of probability, linear algebra, programming (analysis with computer support), and queueing in particular. Thus, read the solutions very carefully.

As a guideline to making the exercises I recommend the following approach. First read the notes. Then attempt to make a exercises for 10 minutes or so by yourself. If by that time you have not obtained a good idea on how to approach the problem, check the solution manual. Once you have understood the solution, try to repeat the arguments *with the solution manual closed*.

Symbols The meaning of the symbols in the margin of pages are as follows:

- The symbol in the margin means that you have to memorize the *emphasized concepts*.
- The symbol in the margin means that this question has a *hint*.
- The symbol in the margin means that this question or its solution requires still some *work on my part*; you can skip it.

Acknowledgments I would like to acknowledge dr. J.W. Nieuwenhuis for our many discussions on the formal aspects of queueing theory and prof. dr. W.H.M. Zijm for allowing me to use the first few chapters of his book. Finally, without **?** I could not have written these notes.

2 Single-Station Queueing Systems

2.1 Exponential Distribution

As we will see in the sections to come, the modeling and analysis of any queueing system involves the specification of the (probability) distribution of the time between consecutive arrival epochs of jobs, or the specification of the distribution of the number of jobs that arrive in a certain interval. For the first case, the most common distribution is the exponential distribution, while for the second it is the Poisson distribution. For these reasons we start our discussion of the analysis of queueing system with these two exceedingly important distributions. In the ensuing sections we will use these distributions time and again.

As mentioned, one of the most useful models for the interarrival times of jobs assumes that the sequence $\{X_i\}$ of interarrival times is a set of *independent and identically distributed (i.i.d.)* random variables. Let us write X for the generic random time between two successive arrivals. For many queueing systems, measurements of the interarrival times between consecutive arrivals show that it is reasonable to model an interarrival X as an *exponentially distributed* random variable, i.e.,

$$\mathsf{P}\{X \le t\} = 1 - e^{-\lambda t} := G(t)$$

The constant λ is often called the *rate*. The reason behind this will be clarified once we relate the exponential distribution to the Poisson process in Section **??**. In the sequel we often write $X \sim \exp(\lambda)$ to mean that X is exponentially distributed with rate λ.

Let us show with simulation how the exponential distribution originates. Consider N people that regularly visit a shop. We assume that we can characterize the interarrival times $\{X_k^i, k = 1,2,\ldots\}$ of customer i by some distribution function, for instance the uniform distribution. Then, with $A_0^i = 0$ for all i,

$$A_k^i = A_{k-1}^i + X_k^i = \sum_{j=1}^{n} X_j^i,$$

is the arrival moment of the kth visit of customer i. Now the shop owner 'sees' the superposition of the arrivals of all customers. One way to compute the arrival moments of all customers together as seen by the shop is to put all the numbers $\{A_k^i, k = 1,\ldots,n, i = 1,\ldots,N\}$ into one set, and sort these numbers in increasing order. This results in the (sorted) set of arrival times $\{A_k, k = 1,2,\ldots\}$ at the shop, and then

$$X_k = A_k - A_{k-1},$$

with $A_0 = 0$, must be the interarrival time between the $k-1$th and kth visit to the shop. Thus, starting from interarrival times of individual customers we can construct interarrival times as seen by the shop.

To plot the *empirical distribution function*, or the histogram, of the interarrival times at the shop, we need to count the number of interarrival times smaller than time t for any t. For this, we introduce the *indicator function*:

$$\mathbb{1}_A = \begin{cases} 1, & \text{if the event } A \text{ is true,} \\ 0, & \text{if the event } A \text{ is false.} \end{cases}$$

With the indicator function we define the emperical distribution of the interarrival times for a simulation with a total of $n \cdot N$ as

$$\mathsf{P}_{nN}\{X \le t\} = \frac{1}{nN} \sum_{k=1}^{nN} \mathbb{1}_{X_k \le t},$$

where $\mathbb{1}_{X_k \le t} = 1$ if $X_k \le t$ and $\mathbb{1}_{X_k \le t} = 0$ if $X_k > t$.

Let us now compare the probability density as obtained for several simulation scenarios to the density of the exponential distribution, i.e., to $\lambda e^{-\lambda t}$. As a first example, take $N = 1$ customer and let the computer generate $n = 100$ uniformly distributed numbers on the set $[4,6]$. Thus, the time between two visits of this customer is somewhere between 4 and 6 hours, and the average interarrival times $\mathsf{E}\{X\} = 5$. In a second simulation we take $N = 3$ customers, and in the third, $N = 10$ customers. The emperical distributions are shown, from left to right, in the three panels in Figure 2.1. The continuous curve is the graph of $\lambda e^{-\lambda x}$ where $\lambda = N/\mathsf{E}\{X\} = N/5$. (In Eq. (2.3) we show that when 1 person visits the shop with an average interarrival time of 5 hours, it must be that the arrival rate is $1/\mathsf{E}\{X\} = 1/5$. Hence, when N customers visit the shop, each with an average interarrival time of 5 hours, the total arrival rate as seen by the shop must be $N/5$.) As a second example, we extend the simulation to $n = 1000$ visits to the shop, see Figure 2.2. In the third example we take the interarrival times to be normally distributed times with mean 5 and $\sigma = 1$, see Figure 2.3.

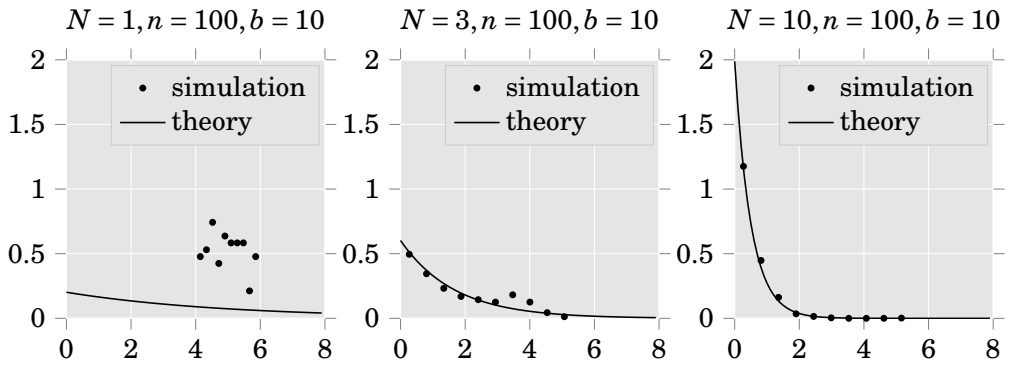

Figure 2.1: The interarrival process as seen by the shop owner. Observe that the density $\lambda e^{-\lambda x}$ intersects the y-axis at level $N/5$, which is equal to the arrival rate when N persons visit the shop. The parameter $n = 100$ is the simulation length, i.e., the number of visits per customer, and $b = 10$ is number of bins to collect the data.

As the graphs show, even when the customer population consists of 10 members, each visiting the shop with an interarrival time that is quite 'far' from exponential, the distribution of the interarrival times as observed by the shop is very well approximated by an exponential distribution. Thus, for a real shop, with many thousands of customers, or a hospital, call center, in fact for nearly every system that deals with random demand, it seems reasonable to use the exponential distribution to model interarrival times. In conclusion, the main conditions to use an exponential distribution are: 1) arrivals have to be drawn from a large population, and 2) each of the arriving customers decides, independent of the others, to visit the system.

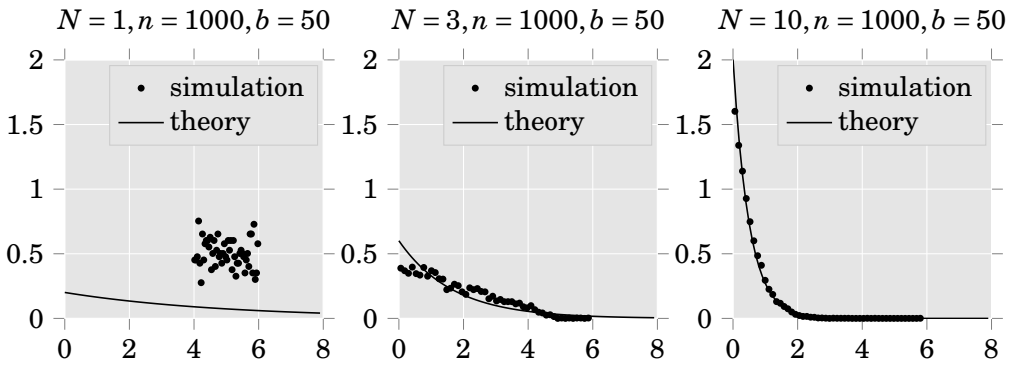

Figure 2.2: Each of the N customer visits the shop with exponentially distributed interarrival times, but now the number of visits is $n = 1000$.

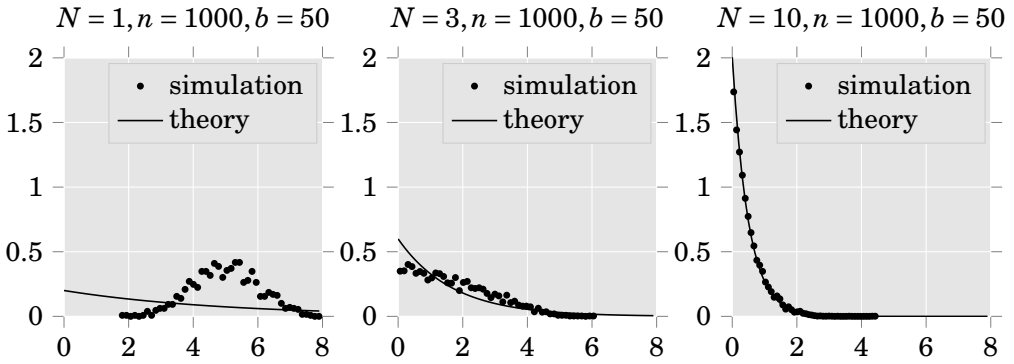

Figure 2.3: Each of the N customer visits the shop with normally distributed interarrival times with $\mu = 5$ and $\sigma = 1$.

Another, but theoretical, reason to use the exponential distribution is that an exponentially distributed random variable is *memoryless*, that is, X is memoryless if it satisfies the property that

$$\mathrm{P}\{X > t + h | X > t\} = \mathrm{P}\{X > h\}.$$

In words, the probability that X is larger than some time $t + h$, conditional on it being larger than a time t, is equal to the probability that X is larger than h. Thus, no matter how long we have been waiting for the next arrival to occur, the probability that it will occur in the next h seconds remains the same. This property seems to be vindicated also in practice: suppose that a patient with a broken arm just arrived at the emergency room of a hospital, what does that tell us about the time the next patient will be brought in? Not much, as most of us will agree.

It can be shown that only exponential random variables have the memoryless property. The proof of this fact requires quite some work; we refer the reader to the literature if s/he want to check this, see e.g. **?**, Appendix 3.

Finally, the reader should realize that it is simple, by means of computers, to generate exponentially distributed interarrival times. Thus, it is easy to use such exponentially distributed random variables to simulate queueing systems.

Exercises

Exercise 2.1.1. (Using conditional probability) We have to give one present to one of three children. As we cannot divide the present into parts, we decide to let 'fate decide'. That is, we choose a random number in the set $\{1,2,3\}$. The first child that guesses the number wins the present. Show that the probability of winning the present is the same for each child.

Exercise 2.1.2. Assume that the time X to fail of a machine is uniformly distributed on the interval $[0,10]$. If the machine fails at time t, the cost to repair it is $h(t)$. What is the expected repair cost?

Exercise 2.1.3. Show that an exponentially distributed random variable is memoryless.

Exercise 2.1.4. If the random variable $X \sim \exp(\lambda)$, show that

$$\mathsf{E}\{X\} = \int_0^\infty t \, \mathrm{d}F(t) = \int_0^\infty t f(t) \, \mathrm{d}t = \int_0^\infty t\lambda e^{-\lambda t} \, \mathrm{d}t = \frac{1}{\lambda},$$

where f is the density function of the distribution function F of X.

Exercise 2.1.5. If the random variable $X \sim \exp(\lambda)$, show that

$$\mathsf{E}\{X^2\} = \int_0^\infty t^2\lambda e^{-\lambda t} \, \mathrm{d}t = \frac{2}{\lambda^2}.$$

Exercise 2.1.6. If the random variable $X \sim \exp(\lambda)$, show that the *variance*

$$\mathsf{V}\{X\} = \mathsf{E}\{X\}^2 - (\mathsf{E}\{X\})^2 = \frac{1}{\lambda^2}.$$

Recall in particular this middle term to compute $\mathsf{V}\{X\}$; it is very practical.

Exercise 2.1.7. Define the *square coefficient of variation (SCV)* as

$$C_a^2 = \frac{\mathsf{V}\{X\}}{(\mathsf{E}\{X\})^2}. \tag{2.1}$$

Prove that when X is exponentially distributed, $C_a^2 = 1$. As will become clear later, the SCV is a very important concept in queueing theory. Memorize it as a measure of *relative variability*.

Exercise 2.1.8. If X is an exponentially distributed random variable with parameter λ, show that its moment generating function

$$M_X(t) = \mathsf{E}\left\{e^{tX}\right\} = \frac{\lambda}{\lambda - t}.$$

Exercise 2.1.9. Let A_i be the arrival time of customer i and set $A_0 = 0$. Assume that the interarrival times $\{X_i\}$ are i.i.d. with exponential distribution with mean $1/\lambda$ for some $\lambda > 0$. Prove that the density of $A_i = X_1 + X_2 + \cdots + X_i = \sum_{k=1}^{i} X_k$ with $i \geq 1$ is

$$f_{A_i}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{i-1}}{(i-1)!}.$$

Exercise 2.1.10. Assume that the interarrival times $\{X_i\}$ are i.i.d. and $X_i \sim \exp(\lambda)$. Let $A_i = X_1 + X_2 + \cdots + X_i = \sum_{k=1}^{i} X_k$ with $i \geq 1$. Use the density of A_i or the moment generating function of A_i to show that

$$\mathsf{E}\{A_i\} = \frac{i}{\lambda},$$

that is, the expected time to see i jobs is i/λ.

Exercise 2.1.11. If $X \sim \exp(\lambda)$ and $S \sim \exp(\mu)$ and X and S are independent, show that

$$Z = \min\{X, S\} \sim \exp(\lambda + \mu),$$

hence $\mathsf{E}\{Z\} = (\lambda + \mu)^{-1}$.

Exercise 2.1.12. If $X \sim \exp(\lambda)$, $S \sim \exp(\mu)$ and independent, show that

$$\mathsf{P}\{X \leq S\} = \frac{\lambda}{\lambda + \mu}.$$

Exercise 2.1.13. A machine serves two types of jobs. The processing time of jobs of type i, $i = 1, 2$, is exponentially distributed with parameter μ_i. The type T of job is random and independent of anything else, and such that $\mathsf{P}\{T = 1\} = p = 1 - q = 1 - \mathsf{P}\{T = 2\}$. (An example is a desk serving men and women, both requiring different average service times, and p is the probability that the customer in service is a man.) What is the expected processing time and what is the variance?

Exercise 2.1.14. Try to make Figure 2.2 with simulation.

Hints

Hint 2.1.1: For the second child, condition on the event that the first does not chose the right number.

Hint 2.1.3: Condition on the event $X > t$.

Hint 2.1.9: Check the result for $i = 1$ by filling in $i = 1$ (just to be sure that you have read the formula right), and compare the result to the exponential density. Then write $A_i = \sum_{i=1}^{k} X_k$, and compute the moment generating function for A_i and use that the interarrival times X_i are independent. Use the moment generating function of X_i.

Hint 2.1.10: Use that $\int_0^\infty (\lambda t)^i e^{-\lambda t}\,\mathrm{d}t = \frac{i!}{\lambda}$. Another way would be to use that, once you have the moment generating function of some random variable X, $\mathsf{E}\{X\} = \frac{\mathrm{d}}{\mathrm{d}t} M(t)|_{t=0}$.

Hint 2.1.11: Use that if $Z = \min\{X, S\} > x$ that then it must be that $X > x$ and $S > x$. Then use independence of X and S.

Hint 2.1.12: Define the joint distribution of X and S and carry out the computations, or use conditioning, or use the result of the previous exercise.

Solutions

Solution 2.1.1: Use the definition of conditional probability ($\mathsf{P}\{A|B\} = \mathsf{P}\{AB\}/\mathsf{P}\{B\}$, provided $\mathsf{P}\{B\} > 0$)

The probability that the first child to guess also wins is 1/3. What is the probability for child number two? Well, for him/her to win, it is necessary that child one does not win and that child two guesses the right number of the remaining numbers. Assume, without loss of generality that child 1 chooses 3 and that this is not the right number. Then

$$\begin{aligned}
&\mathsf{P}\{\text{Child 2 wins}\} \\
&= \mathsf{P}\{\text{Child 2 guesses the right number and child 1 does not win}\} \\
&= \mathsf{P}\{\text{Child 2 guesses the right number}|\ \text{child 1 does not win}\} \cdot \mathsf{P}\{\text{Child 1 does not win}\} \\
&= \mathsf{P}\{\text{Child 2 makes the right guess in the set } \{1,2\}\} \cdot \frac{2}{3} \\
&= \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.
\end{aligned}$$

Similar conditional reasoning gives that child 3 wins with probability 1/3.

Solution 2.1.2: Write for $F(x) = \mathsf{P}\{X \le x\}$ and $f(x) = \mathrm{d}F(x)/\mathrm{d}x$ for the density of F.

$$\begin{aligned}
\mathsf{E}\{h(X)\} &= \int_0^{10} \mathsf{E}\{h(X)\,|\,X = x\}\,\mathsf{P}\{X \in \mathrm{d}x\} \\
&= \int_0^{10} \mathsf{E}\{h(x)\,|\,X = x\}\,\mathrm{d}F(x) \\
&= \int_0^{10} \mathsf{E}\{h(x)\,|\,X = x\}\,F(\mathrm{d}x) \\
&= \int_0^{10} \mathsf{E}\{h(x)\,|\,X = x\}\,f(x)\mathrm{d}x \\
&= \int_0^{10} h(x)\frac{\mathrm{d}x}{10}.
\end{aligned}$$

Here we introduce some notation that is commonly used in the probabity literature to indicate the same conceptual idea, i.e, $\mathsf{P}\{X \in \mathrm{d}x\} = \mathrm{d}F(x) = F(\mathrm{d}x) = f(x)\mathrm{d}x$, where the last equality follows from the fact that F has a density f everywhere on $[0,10]$.

The concept of conditional expectation is of fundamental importance in probability theory. Any *good* probability book defines this concept as a random variable measurable with respect to some σ-algebra. In this course we will not deal with this elegant idea, due to lack of time.

Solution 2.1.3: This is easy, but be sure you can do it.

To see that an exponentially distributed is memoryless, use the definition of conditional probability ($\mathsf{P}\{A|B\} = \mathsf{P}\{AB\}/\mathsf{P}\{B\}$, provided $\mathsf{P}\{B\} > 0$):

$$\mathsf{P}\{X > t+h | X > t\} = \frac{\mathsf{P}\{X > t+h, X > t\}}{\mathsf{P}\{X > t\}} = \frac{\mathsf{P}\{X > t+h\}}{\mathsf{P}\{X > t\}} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = \mathsf{P}\{X > h\}.$$

Solution 2.1.4:

$$\begin{aligned}
\mathsf{E}\{X\} &= \int_0^\infty \mathsf{E}\{X \mid X = t\} f(t)\,\mathrm{d}t \\
&= \int_0^\infty t\lambda e^{-\lambda t}\,\mathrm{d}t, \quad \text{density is } \lambda e^{-\lambda t} \\
&= \lambda^{-1} \int_0^\infty u e^{-u}\,\mathrm{d}u, \quad \text{by change of variable } u = \lambda t, \\
&= -\lambda^{-1}\, t e^{-t}\big|_0^\infty + \lambda^{-1} \int_0^\infty e^{-t}\,\mathrm{d}t \\
&= -\lambda^{-1}\, e^{-t}\big|_0^\infty = \frac{1}{\lambda}.
\end{aligned}$$

Solution 2.1.5:

$$\begin{aligned}
\mathsf{E}\{X^2\} &= \int_0^\infty \mathsf{E}\{X^2 \,|\, X = t\} f(t)\,\mathrm{d}t \\
&= \int_0^\infty t^2 \lambda e^{-\lambda t}\,\mathrm{d}t \\
&= \lambda^{-2} \int_0^\infty u^2 e^{-u}\,\mathrm{d}u, \quad \text{by change of variable } u = \lambda t, \\
&= -\lambda^{-2}\, t^2 e^{-t}\big|_0^\infty + 2\lambda^{-2} \int_0^\infty t e^{-t}\,\mathrm{d}t \\
&= -2\lambda^{-2}\, t e^{-t}\big|_0^\infty + 2\lambda^{-2} \int_0^\infty e^{-t}\,\mathrm{d}t \\
&= -2\lambda^{-2}\, e^{-t}\big|_0^\infty \\
&= 2/\lambda^2.
\end{aligned}$$

Solution 2.1.6: By the previous problems, $\mathsf{E}\{X^2\} = 2/\lambda^2$ and $\mathsf{E}\{X\} = \lambda$.

Solution 2.1.7: By the previous problems, $\mathsf{V}\{X\} = 1/\lambda^2$ and $\mathsf{E}\{X\} = 1/\lambda$.

Solution 2.1.8:

$$\begin{aligned}
M_X(t) &= \mathsf{E}\{\exp(tX)\} = \int_0^\infty e^{tx}\,\mathrm{d}F(x) = \int_0^\infty e^{tx} f(x)\,\mathrm{d}x = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\,\mathrm{d}x \\
&= \lambda \int_0^\infty e^{(t-\lambda)x}\,\mathrm{d}x = \frac{\lambda}{\lambda - t}.
\end{aligned}$$

Solution 2.1.9: One way to find the distribution of A_i is by using the moment generating function $M_{A_i}(t) = \mathsf{E}\{e^{tA_i}\}$ of A_i. Let X_i be the interarrival time between customers i and $i-1$, and $M_X(t)$ the associated moment generating function. Using the i.i.d. property of the $\{X_i\}$,

$$\begin{aligned}
M_{A_i}(t) &= \mathsf{E}\left\{e^{tA_i}\right\} = \mathsf{E}\left\{\exp\left(t \sum_{j=1}^{i} X_j\right)\right\} \\
&= \prod_{j=1}^{i} \mathsf{E}\left\{e^{tX_j}\right\} = \prod_{j=1}^{i} M_{X_j}(t) = \prod_{j=1}^{i} \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^i.
\end{aligned}$$

From a table of moment generation functions it follows immediately that $A_i \sim \Gamma(n, \lambda)$, i.e., A_i is Gamma distributed.

Solution 2.1.10:

$$\mathsf{E}\{A_i\} = \int_0^\infty t\mathsf{P}\{A_i \in \mathrm{d}t\} = \int_0^\infty t f_{A_i}(t)\,\mathrm{d}t = \int_0^\infty t\lambda e^{-\lambda t}\frac{(\lambda t)^{i-1}}{(i-1)!}\,\mathrm{d}t.$$

Thus,

$$\mathsf{E}\{A_i\} = \frac{1}{(i-1)!}\int_0^\infty e^{-\lambda t}(\lambda t)^i\,\mathrm{d}t = \frac{i!}{(i-1)!\lambda} = \frac{i}{\lambda},$$

where we used the hint.

What if we would use the moment generating function?

$$\begin{aligned}
\mathsf{E}\{A_i\} &= \frac{\mathrm{d}}{\mathrm{d}t}M_{A_i}(t)\bigg|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\lambda}{\lambda-t}\right)^i\bigg|_{t=0} \\
&= i\left(\frac{\lambda}{\lambda-t}\right)^{i-1}\frac{\lambda}{(\lambda-t)^2}\bigg|_{t=0} \\
&= \frac{i}{\lambda}\left(\frac{\lambda}{\lambda-t}\right)^{i-1}\bigg|_{t=0} \\
&= \frac{i}{\lambda}.
\end{aligned}$$

And indeed, by the fact that $\mathsf{E}\{X+Y\} = \mathsf{E}\{X\} + \mathsf{E}\{Y\}$ for any r.v. X and Y,

$$\mathsf{E}\{A_i\} = \mathsf{E}\left\{\sum_{k=1}^{i} X_k\right\} = i\,\mathsf{E}\{X\} = \frac{1}{\lambda}.$$

We get the same answer.

Solution 2.1.11: Use that X and S are independent to get

$$\begin{aligned}
\mathsf{P}\{Z > x\} &= \mathsf{P}\{\min X, S > x\} = \mathsf{P}\{X > x \text{ and } S > x\} = \mathsf{P}\{X > x\}\,\mathsf{P}\{S > x\} \\
&= e^{-\lambda x}e^{-\mu x} = e^{-(\lambda+\mu)x}.
\end{aligned}$$

Solution 2.1.12: There is more than one way to show that $\mathsf{P}\{X \le S\} = \lambda/(\lambda+\mu)$.

Method 1. (I admit that, although the simplest, least technical, method, I did not think of this right away. I am 'conditioned' to use conditioning...) Observe first that X and S, being exponentially distributed, both have a density. Moreover, as they are independent, we can

sensibly speak of the joint density $f_{X,S}(x,y) = f_X(x)f_S(y) = \lambda\mu e^{-\lambda x}e^{-\mu y}$. With this,

$$\begin{aligned}
\mathsf{P}\{X \le S\} &= \mathsf{E}\{\mathbb{1}_{X\le S}\} \\
&= \int_0^\infty \int_0^\infty \mathbb{1}_{x\le y} f_{X,S}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \lambda\mu \int_0^\infty \int_0^\infty \mathbb{1}_{x\le y} e^{-\lambda x} e^{-\mu y}\,\mathrm{d}y\,\mathrm{d}x \\
&= \lambda\mu \int_0^\infty e^{-\mu y} \int_0^y e^{-\lambda x}\,\mathrm{d}x\,\mathrm{d}y \\
&= \mu \int_0^\infty e^{-\mu y}(1-e^{-\lambda y})\,\mathrm{d}y \\
&= \mu \int_0^\infty (e^{-\mu y} - e^{-(\lambda+\mu)y})\,\mathrm{d}y \\
&= \mu \int_0^\infty (e^{-\mu y} - e^{-(\lambda+\mu)y})\,\mathrm{d}y \\
&= 1 - \frac{\mu}{\lambda+\mu}
\end{aligned}$$

This argument is provided in the probability book you use in the first year.

Method 2. Applying a standard conditioning argument

$$\mathsf{P}\{X \le S\} = \int_0^\infty \mathsf{P}\{X \le S | S = s\}\,\mu e^{-\mu s}\,\mathrm{d}s.$$

Now, $\mathsf{P}\{X \le S|S = s\}$ is a conditional probability distribution. This is a bit of tricky object, but very useful once you get used to it. The tricky part is that $\mathsf{P}\{S = s\} = 0$. Therefore $\mathsf{P}\{X \le S \,|\, S = s\}$ cannot be defined as $\frac{\mathsf{P}\{X\le s, S=s\}}{\mathsf{P}\{S=s\}}$. However, if we proceed nonetheless and use the independence of S and X, we get

$$\mathsf{P}\{X \le S|S = s\} = \frac{\mathsf{P}\{X \le s, S = s\}}{\mathsf{P}\{S = s\}} = \frac{\mathsf{P}\{X \le s\}\,\mathsf{P}\{S = s\}}{\mathsf{P}\{S = s\}} = \mathsf{P}\{X \le s\}$$

and thus, indeed, $\mathsf{P}\{X \le S|S = s\} = \mathsf{P}\{X \le s\}$. Then,

$$\begin{aligned}
\mathsf{P}\{X \le S\} &= \int_0^\infty \mathsf{P}\{X \le S|S = s\}\,\mu e^{-\mu s}\,\mathrm{d}s \\
&= \int_0^\infty \mathsf{P}\{X \le s\}\,\mu e^{-\mu s}\,\mathrm{d}s \\
&= \int_0^\infty (1-e^{-\lambda s})\mu e^{-\mu s}\,\mathrm{d}s
\end{aligned}$$

and we arrive at the integral we have seen above.

So we get the correct answer, but by the wrong method. How can we repair this? As a first step, let's not fix S to a set of measure zero, but let's assume that $S \in [s,t]$ for $s < t$. Then it follows that

$$\mathbb{1}_{X\le s}\,\mathbb{1}_{S\in[s,t]} \le \mathbb{1}_{X\le S}\,\mathbb{1}_{S\in[s,t]} \le \mathbb{1}_{X\le t}\,\mathbb{1}_{S\in[s,t]}$$

As a second step, using that $\mathsf{P}\{S \in [s,t]\} > 0$ if $s < t$ and the independence of X and S,

$$\begin{aligned}
\mathsf{P}\{X \le s\} &= \frac{\mathsf{P}\{X \le s\}\,\mathsf{P}\{S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}} = \frac{\mathsf{P}\{X \le s, S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}} \\
\mathsf{P}\{X \le t\} &= \frac{\mathsf{P}\{X \le t\}\,\mathsf{P}\{S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}} = \frac{\mathsf{P}\{X \le t, S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}}
\end{aligned}$$

Now with the result of the first step

$$\begin{aligned}\mathsf{P}\{X \le s\} &= \frac{\mathsf{P}\{X \le s, S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}} \\ &\le \frac{\mathsf{P}\{X \le S, S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}} \\ &= \mathsf{P}\{X \le S | S \in [s,t]\} \\ &\le \frac{\mathsf{P}\{X \le t, S \in [s,t]\}}{\mathsf{P}\{S \in [s,t]\}} \\ &= \mathsf{P}\{X \le t\}.\end{aligned}$$

Hence,

$$\mathsf{P}\{X \le s\} \le \mathsf{P}\{X \le S | S \in [s,t]\} \le \mathsf{P}\{X \le t\}.$$

Finally, taking the limit $t \downarrow s$, and defining $\mathsf{P}\{X \le S | S = s\} = \lim_{t \downarrow s} \mathsf{P}\{X \le S | S \in [s,t]\}$, it follows that

$$\mathsf{P}\{X \le s\} = \mathsf{P}\{X \le S | S = s\}$$

A more direct way to properly define $\mathsf{P}\{X \le S | S = s\}$ is as follows. For any y such that $f_S(y) > 0$, we can define the conditional probability density function of X, given that $S = s$, as

$$f_{X|S}(x|s) = \frac{f_{X,S}(x,s)}{f_Y(s)},$$

where, as before, $f_{X,S}(x,s)$ is the joint density of X and S. Now that the conditional probability density is defined, we can properly define

$$\mathsf{E}\{X | S = s\} = \int_0^\infty x f_{X|S}(x|s)\,\mathrm{d}x$$

and also

$$\mathsf{P}\{X \le S | S = s\} = \mathsf{E}\{\mathbb{1}_{X \le S} | S = s\} = \int_0^\infty \mathbb{1}_{x \le s} f_{X|S}(x|s)\,\mathrm{d}x.$$

Using the definition of $f_{X|S}(x|s)$ and the independence of X and S it follows that

$$f_{X|S}(x|s) = \frac{f_{X,S}(x,s)}{f_Y(s)} = \frac{\lambda e^{-\lambda x} \mu e^{-\mu s}}{\mu e^{-\mu s}} = \lambda e^{-\lambda x}$$

from which we get that

$$\begin{aligned}\mathsf{E}\{\mathbb{1}_{X \le S} | S = s\} &= \int_0^\infty \mathbb{1}_{x \le s} f_{X|S}(x|s)\,\mathrm{d}x \\ &= \int_0^\infty \mathbb{1}_{x \le s} \lambda e^{-\lambda x}\,\mathrm{d}x \\ &= \int_0^s \lambda e^{-\lambda x}\,\mathrm{d}x \\ &= 1 - e^{-\lambda s},\end{aligned}$$

that is,

$$\mathsf{P}\{X \le S | S = s\} = \mathsf{E}\{\mathbb{1}_{X \le S} | S = s\} = 1 - e^{-\lambda s} = \mathsf{P}\{X \le s\}.$$

All of these problems can be put on solid ground by using measure theory. We do not pursue these matters any further, but trust on our intuition that all is well.

Solution 2.1.13: Let X be the processing (or service) time at the server, and X_i the service time of a type i job. Then,

$$X = \mathbb{1}_{T=1}X_1 + \mathbb{1}_{T=2}X_2,$$

where $\mathbb{1}$ is the indicator function, that is, $\mathbb{1}_A = 1$ if the event A is true, and $\mathbb{1}_A = 0$ if A is not true. With this,

$$\begin{aligned}
\mathsf{E}\{X\} &= \mathsf{E}\{\mathbb{1}_{T=1}X_1\} + \mathsf{E}\{\mathbb{1}_{T=2}X_2\} \\
&= \mathsf{E}\{\mathbb{1}_{T=1}\}\,\mathsf{E}\{X_1\} + \mathsf{E}\{\mathbb{1}_{T=2}\}\,\mathsf{E}\{X_2\}, \text{ by the independence of } T, \\
&= \mathsf{P}\{T=1\}/\mu_1 + \mathsf{P}\{T=2\}/\mu_2 \\
&= p/\mu_1 + q/\mu_2 \\
&= p\,\mathsf{E}\{X_1\} + q\,\mathsf{E}\{X_2\}.
\end{aligned}$$

(The next derivation may seem a bit long, but the algebra is standard. I include all steps so that you don't have to use pen and paper yourself if you want to check the result.) Next, using that

$$\mathbb{1}_{T=1}\,\mathbb{1}_{T=2} = 0 \text{ and } \mathbb{1}^2_{T=1} = \mathbb{1}_{T=1},$$

we get

$$\begin{aligned}
\mathsf{V}\{X\} &= \mathsf{E}\{X^2\} - (\mathsf{E}\{X\})^2 \\
&= \mathsf{E}\{(\mathbb{1}_{T=1}X_1 + \mathbb{1}_{T=2}X_2)^2\} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\
&= \mathsf{E}\{\mathbb{1}_{T=1}X_1^2 + \mathbb{1}_{T=2}X_2^2\} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\
&= p\,\mathsf{E}\{X_1^2\} + q\,\mathsf{E}\{X_2^2\} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\
&= p\,\mathsf{V}\{X_1\} + p(\mathsf{E}\{X_1\})^2 + q\,\mathsf{V}\{X_2\} + q(\mathsf{E}\{X_2\})^2 - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\
&= p\,\mathsf{V}\{X_1\} + \frac{p}{\mu_1^2} + q\,\mathsf{V}\{X_2\} + \frac{q}{\mu_2^2} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\
&= p\,\mathsf{V}\{X_1\} + q\,\mathsf{V}\{X_2\} + \frac{p}{\mu_1^2} + \frac{q}{\mu_2^2} - \frac{p^2}{\mu_1^2} - \frac{q^2}{\mu_2^2} - \frac{2pq}{\mu_1\mu_2} \\
&= p\,\mathsf{V}\{X_1\} + q\,\mathsf{V}\{X_2\} + \frac{p(1-p)}{\mu_1^2} + \frac{q(1-q)}{\mu_2^2} - \frac{2pq}{\mu_1\mu_2} \\
&= p\,\mathsf{V}\{X_1\} + q\,\mathsf{V}\{X_2\} + \frac{pq}{\mu_1^2} + \frac{qp}{\mu_2^2} - \frac{2pq}{\mu_1\mu_2} \\
&= p\,\mathsf{V}\{X_1\} + q\,\mathsf{V}\{X_2\} + pq(\mathsf{E}\{X_1\} - \mathsf{E}\{X_2\})^2.
\end{aligned}$$

Interestingly, we see that even if $\mathsf{V}\{X_1\} = \mathsf{V}\{X_2\} = 0$, $\mathsf{V}\{X\} > 0$ if $\mathsf{E}\{X_1\} \neq \mathsf{E}\{X_2\}$. Bear this in mind; we will use these ideas later when we discuss the effects of failures on the variance of service times of jobs.

Solution 2.1.14: The source code can be found in `progs/converge_to_exp.py`.

2.2 Rate Stability and Utilization

In the analysis of any queueing process the first step should be to check the relations between the arrival, service and departure rates. The concept of rate is crucial because it captures our intuition that when, on the long run, jobs arrive faster than they can leave, the system must 'explode'. Thus, the first performance measures we need to estimate when analyzing a queueing system are the arrival and departure rate, and then we need to check that the arrival rate is smaller than the departure rate. In particular, the load, defined as the ratio of the arrival rate and service rate is of importance. In this section we define and relate these concepts. As a reminder, we keep the discussion in these notes mostly at an intuitive level, and refer to **?** for proofs and further background.

We first formalize the *arrival rate* and *departure rate* in terms of the *counting processes* $\{A(t)\}$ and $\{D(t)\}$. The *arrival rate* is the long-run average number of jobs that arrive per unit time, i.e.,

$$\lambda = \lim_{t\to\infty} \frac{A(t)}{t}. \tag{2.2}$$

We remark in passing that this limit does not necessarily exist if $A(t)$ is some pathological function. If, however, the interarrival times $\{X_k\}$ are the basic data, and $\{X_k\}$ are i.i.d. and distributed as a generic random variable X with finite mean $\mathsf{E}\{X\}$, we can construct $\{A_k\}$ and $\{A(t)\}$ as described in Section **??**; the strong law of large numbers guarantees that the above limit exists.

Observe that at time $t = A_n$, precisely n arrivals occurred. Thus, by applying the definition of $A(t)$ at the epochs A_n, we see that $A(A_n) = n$. Thus,

$$\frac{1}{n}\sum_{k=1}^{n} X_k = \frac{A_n}{n} = \frac{A_n}{A(A_n)}.$$

But since $A_n \to \infty$ if $n \to \infty$, it follows from Eq. (2.2) that the average interarrival time between two consecutive jobs is

$$\mathsf{E}\{X\} = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \lim_{n\to\infty} \frac{A_n}{A(A_n)} = \lim_{t\to\infty} \frac{t}{A(t)} = \frac{1}{\lambda}, \tag{2.3}$$

where we take $t = A_n$ in the limit for $t \to \infty$. In words, the above states that the arrival rate λ is the inverse of the expected interarrival time.

The development of the departure times $\{D_k\}$ is entirely analogous to that of the arrival times; we leave it to the reader to provide the details. As a result we can define the *departure rate* as

$$\lim_{t\to\infty} \frac{D(t)}{t} = \gamma \tag{2.4}$$

Assume now that there is a single server. Let S_k be the required service time of the kth job to be served, and define

$$U_n = \sum_{k=1}^{n} S_k$$

as the total service time required by the first n jobs. With this, let

$$U(t) = \sup\{n : U_n \le t\}.$$

and define the *service* or *processing rate* as

$$\mu = \lim_{t\to\infty} \frac{U(t)}{t}.$$

In the same way as we derived that $\mathsf{E}\{X\} = 1/\lambda$, we obtain for the expected (or average) amount of service required by an individual job

$$\mathsf{E}\{S\} = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} S_k = \lim_{n\to\infty} \frac{U_n}{n} = \lim_{n\to\infty} \frac{U_n}{U(U_n)} = \lim_{t\to\infty} \frac{t}{U(t)} = \frac{1}{\mu}.$$

Now observe that, if the system is empty at time 0, it must be that at any time the number of departures must be smaller than the number of arrivals, i.e., $D(t) \le A(t)$ for all t. Therefore,

$$\gamma := \lim_{t} \frac{D(t)}{t} \le \lim_{t} \frac{A(t)}{t} = \lambda. \tag{2.5}$$

We call a system *(rate) stable* if

$$\lambda = \gamma,$$

in other words, the system is stable if, on the long run, jobs leave the system just as fast as they arrive. Of course, if $\lambda > \gamma$, the system length process $L(t) \to \infty$ as $t \to \infty$.

It is also evident that jobs cannot depart faster than they can be served, hence, $D(t) \le U(t)$ for all t. Combining this with the fact that $\gamma \le \lambda$, we get

$$\gamma \le \min\{\lambda, \mu\}.$$

When $\mu \ge \lambda$ the above inequality reduces to $\gamma = \lambda$ for rate-stable systems. (It is interesting to prove this.) As it turns out, when $\mu = \lambda$ and the variance of the service times $\mathsf{V}\{S\} > 0$ or $\mathsf{V}\{X\} > 0$ then $\lim_t L(t)/t$ does not necessarily exist. For this reason we henceforth require that $\mu > \lambda$.

The concept of *load* or *utilization*, denoted by the symbol ρ, is fundamental. One way to define it is as the limiting fraction of time the server is busy, i.e.,

$$\rho = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}_{L(s)>0}\,\mathrm{d}s.$$

Interestingly, we can express this in terms of the arrival rate λ and service rate μ. Observe that

$$\sum_{k=1}^{A(t)} S_k \ge \int_0^t \mathbb{1}_{L(s)>0}\,\mathrm{d}s \ge \sum_{k=1}^{D(t)} S_k,$$

since t can lie half way a service interval and $A(t) \ge D(t)$. As $A(t) \to \infty$ as $t \to \infty$,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{A(t)} S_k = \lim_{t\to\infty} \frac{A(t)}{t}\frac{1}{A(t)}\sum_{k=1}^{A(t)} S_k = \lim_{t\to\infty} \frac{A(t)}{t} \cdot \lim_{t\to\infty} \frac{1}{A(t)}\sum_{k=1}^{A(t)} S_k = \lambda\,\mathsf{E}\{S\}.$$

Applying similar limits to the other inequality gives

$$\lambda\,\mathsf{E}\{S\} \ge \rho \ge \gamma\,\mathsf{E}\{S\}.$$

Hence, if $\gamma = \lambda$, $\rho = \lambda\,\mathsf{E}\{S\}$.

From the identies $\lambda^{-1} = \mathsf{E}\{X\}$ and $\mu^{-1} = \mathsf{E}\{S\}$, we get a further set of relations:

$$\rho = \lambda\,\mathsf{E}\{S\} = \frac{\lambda}{\mu} = \frac{\mathsf{E}\{S\}}{\mathsf{E}\{X\}}.$$

Thus, the load has also the interpretation as the rate at which jobs arrive multiplied by the average amount of work per job. Finally, recall that for a system to be rate-stable, it is necessary that $\mu > \lambda$, implying in turn that $\rho < 1$. The relation $\rho = \mathsf{E}\{S\}/\mathsf{E}\{X\} < 1$ then tells us that the average time it takes to serve a job must be less than the average time between two consecutive arrivals, i.e., $\mathsf{E}\{S\} < \mathsf{E}\{X\}$.

Exercises

Exercise 2.2.1. Define the departure time D_k of the kth job in terms of $\{D(t)\}$.

Exercise 2.2.2. Define the random variables $\{\tilde{X}_k, k = 1,\ldots\}$ as

$$\tilde{X}_k = S_{k-1} - X_k.$$

For stability of the queueing process it is essential that $\tilde{X}_k$ has negative expectation, i.e., $\mathsf{E}\{\tilde{X}_k\} = \mathsf{E}\{S_{k-1} - X_k\} < 0$. What is the conceptual meaning of this inequality?

Exercise 2.2.3. Define $\tilde{X}_k = S_{k-1} - X_k$. Show that $\mathsf{E}\{\tilde{X}_k\} < 0$ implies that $\lambda < \mu$.

Exercise 2.2.4. Show that $\mathsf{E}\{S\}/\mathsf{E}\{X\}$ is the fraction of time the server is busy.

Exercise 2.2.5. Show that $\mathsf{E}\{X - S\}/\mathsf{E}\{X\}$ is the fraction of time the server is idle.

Exercise 2.2.6. If $\mathsf{E}\{B\}$ is the expected busy time and $\mathsf{E}\{I\}$ is the expected idle time, show that

$$\mathsf{E}\{B\} = \frac{\rho}{1-\rho}\mathsf{E}\{I\}.$$

is the fraction of time the server is busy.

Exercise 2.2.7. Consider a queueing system with c identical servers (identical in the sense that each server has the same production rate μ). What would be a reasonable stability criterion for this system?

Exercise 2.2.8. Consider a paint factory which contains a paint mixing machine that serves two classes of jobs, A and B. The processing times of jobs of types A and B are constant and require t_A and t_B hours. The job arrival rate is λ_A for type A and λ_B for type B jobs. It takes a setup time of S_s hours to clean the mixing station when changing from paint type A to type B, and there is no time required to change from type B to A.

To keep the system (rate) stable, it is necessary to produce the jobs in batches, for otherwise the server, i.e., the mixing machine, spends a too large fraction of time on setups, so that $\mu < \lambda$. Thus, it is necessary to identify minimal batch sizes to ensure that $\mu > \lambda$. Motivate that the linear program below can be used to determine the minimal batch sizes.

$$\text{minimize } T$$

such that

$$\begin{aligned} T &= k_A t_A + S + k_B t_B, \\ \lambda_A T &< k_A, \\ \lambda_B T &< k_B. \end{aligned}$$

Exercise 2.2.9. Can you make an arrival process such that $A(t)/t$ does not have a limit?

Exercise 2.2.10. This exercise is meant to give the reader some idea about what needs to be done to put everything on solid ground. If you are not interested in the maths, you can skip the problem.

1. In Eq. (2.3) we replaced the limit with respect to n by a limit with respect to t. But why is this actually allowed? Use the notation $A_{A(t)}$ to show that all is OK.
2. Show that the function $t \to A(t)$ as defined by Eqs. (**??**) is right-continuous.

Exercise 2.2.11. Check with simulation that when $\lambda > \mu$ the queue length grows roughly linearly with slope $\lambda - \mu$. Thus, if $\rho > 1$, 'we are in trouble'.

Hints

Hint 2.2.1: Use the analogy with Eq. (**??**).

Hint 2.2.3: Remember that $\{X_k\}$ and $\{S_k\}$ are sequences of i.i.d. random variables. What are the implications for the expectations?

Hint 2.2.4: Let $T_1 > A_1$ be the first time after the arrival of job 1 that arrives at an empty system. (Observe that job 1 also arrives at an empty system.) Suppose, for ease of writing, that this job is the $n+1$th job, so that up to time T_1 the number of arrivals is n. Since the first job arrived at time $A_1 = X_1$, the first n jobs arrived during $[X_1, T_1)$. The total amount of service that arrived during this period is $\sum_{i=1}^{n} S_i$. What is the fraction of time the server has been busy in this cycle?

Hint 2.2.7: What is the rate in, and what is the service capacity?

Hint 2.2.8: Here are some questions to help you interpret this formulation.

1. What are the decision variables for this problem? In other words, what are the 'things' we can control/change?
2. What are the interpretations of $k_A t_A$, and $S + k_B t_B$?
3. What is the meaning of the first constraint? Realize that T represents one production cycle. After the completion of one such cycle, we start another cycle. Hence, the start of every cycle can be seen as a restart of the entire system.
4. What is the meaning of the other two constraints?
5. Why do we minimize the cycle time T?
6. Solve for k_A and k_B in terms of S, λ_A, λ_B and t_A, t_B.
7. Generalize this to m job classes and such that the cleaning time between jobs of class i and j is given by S_{ij}. (Thus, the setup times are sequence dependent.)

Hint 2.2.9: As a start, the function $\sin(t)$ has not a limit as $t \to \infty$. However, the time-average $\sin(t)/t \to 0$. Now you need to make some function whose time-average does not converge, hence it should grow fast, or fluctuate wilder and wilder.

Hint 2.2.10: Use that $A_{A(t)} \le t < A_{A(t)+1}$. Divide by $A(t)$ and take suitable limits. BTW, such type of proof is used quite often to show that the existence of one limit implies, and is implied by, the existence of another type of limit.

Solutions

Solution 2.2.1:

$$D_k = \inf\{t; D(t) \ge k\}.$$

Solution 2.2.2: That the average time customers spend in service is smaller than the average time between the arrival of two subsequent jobs.

Solution 2.2.3: $0 > \mathsf{E}\{\tilde{X}_k\} = \mathsf{E}\{S_{k-1} - X_k\} = \mathsf{E}\{S_{k-1}\} - \mathsf{E}\{X_k\} = \mathsf{E}\{S\} - \mathsf{E}\{X\}$, where we use the fact that the $\{S_k\}$ and $\{X_k\}$ are i.i.d. sequences. Hence,

$$\mathsf{E}\{X\} > \mathsf{E}\{S\} \iff \frac{1}{\mathsf{E}\{S\}} > \frac{1}{\mathsf{E}\{X\}} \iff \mu > \lambda.$$

Solution 2.2.4: The fraction of time that the server has been busy during $[X_1, T_1)$ is

$$\frac{\sum_{i=1}^{n} S_i}{T_1 - X_1} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n+1} X_i - X_1} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=2}^{n+1} X_i}$$

Now use the assumption that the $\{X_i\}$ and $\{S_i\}$ are sequences of i.i.d. random variables distributed as the generic random variables X and S, respectively. Then by taking expectations the above becomes

$$\frac{\mathsf{E}\{\sum_{i=1}^{n} S_i\}}{\mathsf{E}\{\sum_{i=2}^{n+1} X_i\}} = \frac{n\,\mathsf{E}\{S\}}{n\,\mathsf{E}\{X\}} = \frac{\mathsf{E}\{S\}}{\mathsf{E}\{X\}}.$$

These busy cycles occur over and over again. Thus, the long-run average fraction of time the server is busy must also be $\mathsf{E}\{S\}/\mathsf{E}\{X\}$. (For the die-hards, there is a subtle point here: the arrival epochs of the $G/G/1$ queue are not real renewal moments, hence the epochs at which the busy times start also do not form a sequence of renewal times. But then it is not true, in general, that the busy times $\{B_i\}$ have the same distribution, neither do the idles times $\{I_n\}$. Showing that in the limit all is OK requires a substantial amount of mathematics. The above claim is still true however.)

Solution 2.2.5: Since the fraction of idle time is 1 minus the fraction of busy time, it follows that $1 - \mathsf{E}\{S\}/\mathsf{E}\{X\} = (\mathsf{E}\{X\} - \mathsf{E}\{S\})/\mathsf{E}\{X\}$ is the idle time fraction.

Solution 2.2.6: Consider a busy cycle, that is, a cycle that starts with the first job that sees an empty system at upon arrival up to the time another job sees an empty system upon arrival. In such one cycle, the server is busy for an expected duration $\mathsf{E}\{B\}$. The total expected length of the cycle is $\mathsf{E}\{B\} + \mathsf{E}\{I\}$, since after the last job of the cycle left, the expected time until the next job is $\mathsf{E}\{I\}$.

Since ρ is utilization of the server,

$$\rho = \frac{\mathsf{E}\{B\}}{\mathsf{E}\{B\} + \mathsf{E}\{I\}}.$$

With a bit of algebra the result follows.

Solution 2.2.7: The criterion is that c must be such that $\lambda < c\mu$. (Thus, we interpret the number of servers as a *control*, i.e., a 'thing' we can change, while we assume that λ and μ cannot be easily changed.) To see this, we can take two different points of view. Imagine that the c servers are replaced by one server that works c times as fast. The service capacity of these two systems (i.e., the system with c servers and the system with one fast server) is the same, i.e., $c\mu$, where μ is the rate of one server. For the system with the fast server the load is defined as $\rho = \lambda/c\mu$, and for stability we require $\rho < 1$. Another way to see it is to assume that the stream of jobs is split into c smaller streams, each with arrival rate λ/c. In this case, applying the condition that $(\lambda/c)/\mu < 1$ per server leads to the same condition that $\lambda/(c\mu) < 1$.

Solution 2.2.8: Realize that the machine works in cycles. A cycle starts with processing k_A jobs of type A, then does a setup, and processes k_B jobs of type B, and then a new cycle starts again. The time it takes to complete one such cycle is $T = k_A t_A + S + k_B t_B$. The number of jobs of type A processed during one such cycle is, of course, k_A. Observe next that the average number of jobs that arrive during one cycle is $\lambda_A T$. We of course want that $\lambda_A T < k_A$, i.e., less jobs of type A arrive on average per cycle than what we can process.

Solution 2.2.9: If $N(t) = 3t^2$, then clearly $N(t)/t = 3t$. This does not converge to a limit.

Another example, let the arrival rate $\lambda(t)$ be given as follows:

$$\lambda(t) = \begin{cases} 1 & \text{if } 2^{2k} \le t < 2^{2k+1} \\ 0 & \text{if } 2^{2k+1} \le t < 2^{2(k+1)}, \end{cases}$$

for $k = 0, 1, 2, \ldots$. Let $N(t) = \lambda(t)t$. Then $A(t)/t$ does not have limit. Of course, these examples are quite pathological, and are not representable for 'real life cases' (Although this is also quite vague. What, then, is a real life case?)

For the mathematically interested, we seek a function for which its Cesàro limit does not exist.

Solution 2.2.10: Observing that $A_{A(t)}$ is the arrival time of the last job before time t and that $A_{A(t)+1}$ is the arrival time of the first job after time t:

$$A_{A(t)} \le t < A_{A(t)+1} \Leftrightarrow \frac{A_{A(t)}}{A(t)} \le \frac{t}{A(t)} < \frac{A_{A(t)+1}}{A(t)} = \frac{A_{A(t)+1}}{A(t)+1}\frac{A(t)+1}{A(t)}$$

Now $A(t)$ is a counting process such that $A(t) \to \infty$ as $t \to \infty$. Therefore, $\lim_t A_{A(t)}/A(t) = \lim_n A_n/n$. Moreover, it is evident that $\lim_t A_{A(t)+1}/(A(t)+1) = \lim_t A_{A(t)}/A(t)$, and that $(A(t)+1)/A(t) \to 1$ as $t \to \infty$. Thus it follows from the above inequalities that $\lim_n A_n/n = \lim_t t/A(t)$.

Hopefully this problem, and its solution, clarifies that even such small details require attention. If we want to make some progress with respect to developing some queueing theory, we have to skip most of the proofs and mathematical problems; we simply don't have enough time in this course to be concerned with all theorems and proofs.

For the right-continuity of $A(t)$, define $f(t) = 1\{A_1 \le t\}$. Observe first that $f(t)$ is increasing, and $f(t) \in \{0, 1\}$. Thus, if $f(t) = 1$ then $f(u) = 1$ for all $u \ge t$, and if $f(t) = 0$ then $f(u) = 0$ for all $u \le t$.

You may skip the rest of the prove below, but the above is essential to memorize; make a plot of $f(t)$, in particular the behavior around A_1 is important.

We need to prove, for right-continuity, that $f(u) \to f(t)$ as $u \downarrow t$. When $f(t) = 1$, $f(u) = 1$ for any $u > 1$, by the definition of $f(x)$. When $f(t) = 0$ we have to do a bit more work. Formally, we have

to prove that, for fixed t and for all $\epsilon > 0$, there is a $\delta > 0$ such that $u \in (t, t+\delta) \Rightarrow |f(u) - f(t)| < \epsilon$. (Note the differences with the regular definition of continuity.) Since, by assumption, t is such that $f(t) = 0$, and $f \in \{0,1\}$ we need to show that $f(u) = 0$ for $u \in (t, t+\delta)$. Now, clearly, if $f(t) = 0$ only if $t < A_1$. But, then for any $u \in (t, A_1)$, we have that $f(u) = 0$. Thus, taking $\delta = A_1 - t$ suffices.

The next step is to observe that $A(t)$ is a sum of right-continuous functions whose steps do not overlap since by assumption $0 < A_1 < A_2 < \cdots$. As A is (almost surely) finite sum of bounded, increasing and right-continuous functions, it is also right-continuous.

If you like, you can try to prove this last step too.

Solution 2.2.11: See my website.