Queueing Theory

A mathematically rigorous yet intuitive introduction to queues.

Rishabh Pomaje — 210020036@iitdh.ac.in

Samyak Sanjay Parakh — 210020043@iitdh.ac.in

Instructor: Prof. Naveen M. B.

Autumn 2024-25

Indian Institute of Technology Dharwad

Contents

Chapter 1

Motivation and Background

Queues arise in nature by the virtue of there being limitations on the quantity and efficiency of resources. At first glance, the study of queues may seem unremarkable. I mean, have you ever looked forward to waiting in any kind of line? However, we will adopt an intuitive approach to make the subject engaging while ensuring we cover the rigorous mathematical details, which often provide valuable insights. Ready? Let's begin...

1.1   Queueing Theory

I like the statement made in [2], that Queueing Theory is the discipline that conducts *Study of Waiting*. What? I hear you ask. Whats there to study about waiting? Let me again assure you, a lot. Just to show how often scenarios arise where we have to wait, recollect the instances when you waited hours at the bank just to get your passbook updated or the ATM to get some cash. What about the long queues at the shopping center. We can also go beyond human queues. All of us use a plethora of networks on a day to day basis. The most familiar network is one associated with WiFi. All of our devices, smartphones, laptops, PCs, and now a days even TVs, refrigerators, and wrist watches connect to it. By it, I mean the router. When wanting to communicate to some other device in some other locations, these devices are essentially lining up in a queue at the router waiting for their work, in this case their data to be processed. Hence, we see that queues are ubiquitous. We just have to look for them!

1.2   Groundwork

Now we will start by making a few things concrete. The entities forming a queue vary depending on the application area. For a service based company like banks, it will be humans, for a computer scientist, it might be 'jobs' at a server. Similarly, in telecommunications context, it will be number of calls or packets of data. The number of objects in a queue at a given time will form a reference for us.

1.2.1   Block diagram

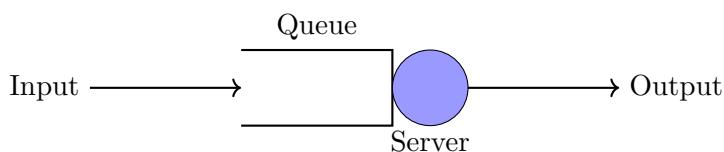To visualize a queue we will use a depiction shown below.

Figure 1.1: Block diagram of a simple, 1-D queue with one server and infinite capacity.

1.2.2 Terminology

The above block diagram has a few terms used in it. A queue is formed by entities. What these entities are depends on the context provided by the application of the theory. For example, it could be jobs waiting to be executed at a server in a data center, or it could be humans waiting at a bank, or it could even be calls lined up at a call center. To address the requirements of the entity, the requirements again depend on the

Whenever a new entity is added to the queue, we call it as an arrival event.

1.2.3 Kendall Notation

A queue is characterized by how entities arrive to it, how they depart from it, how many servers are present, and if existant, what is the maximum length of the queue limited to.

Kendall's Notation provides a condensed way to denote a queue. The general form is given by

$$X/Y/x/y \tag{1.1a}$$

where,

$$X = \text{Describes arrival statistics.}$$
$$Y = \text{Describes the departure statistics.}$$
$$x = \text{Number of Servers.}$$
$$y = \text{Maximum length of the queue.}$$

A frequently used statistics is *Markovian*, where the process(es) of arrival and/or departure are Markov. 'M' is used to abbreviate a Markovian statistics, 'D' to denote Deterministic Timing, 'G' for general statistics and 'Geom' for Geometric.

1.2.4 Assumptions

1.

Chapter 2

M/M/1 Queue

In this chapter, we will study the simplest kind of queue, the $M/M/1$ or also known as *Markovian* queue. While the queue, once we get to its details may not seem realistics, it does provide valuable insights due to its mathematical tractability.

> **Definition 2.0.1: M/M/1 (or Markovian) Queue**
>
> A $M/M/1$ queue, also known as *Markovian* queue is characterized as follows:
>
> 1. The arrival process is a Poisson Random process.
>
> 2. There is a single server with the serving times being exponentially distributed.
>
> 3. There is no limit on the size of the queue. Also, the state of the queue is given by the number of arrivals/ customers in the queue at a given moment.

Chapter 3

Appendix

Only the bare minimum concepts have been covered here. I strongly advise that you go through the relevant material in [1]

3.1 Random Processes

> **Definition 3.1.1: Random Process**
>
> A random process is a collection (or a sequence) of variables usually indexed by time.

If the indexing variable is continuous, we refer to the process as a continuous-time random process and if the indexing variable is discrete, we call the process a discrete-time process. Thus, sampling a random process at a time instant gives a random variable. If this random variable is discrete, we call the process a discrete-valued process. Similarly, we define continuous-valued process.

Bibliography

[1] H. Pishro-Nik. *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC, 2014.

[2] Thomas G. Robertazzi. *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. Springer-Verlag, Berlin, Heidelberg, 3rd edition, 2000.