EE 699 - Next Generation Wireless Networks

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** Page 3

1.1 Queueing Theory 3

1.2 Groundwork 3

Block diagram — 3 • Terminology — 4 • Kendall Notation — 4 • Assumptions — 4

1.3 Service Discipline 4

Chapter 2 **M/M/1 Queue** Page 6

2.1 Arrivals 6

State of the System - Alternative approach — 8 • Mean arrivals in an interval $[0,t]$ — 9 • Variance of Number of arrivals in an interval $[0,t]$ — 10 • Inter-arrival times — 10

2.2 Service 11

State of the System — 11 • Average number of customers in $M/M/1$ queue — 13 • Variance of number of customers in $M/M/1$ queue — 14

2.3 Little's Law 15

Chapter 3 **Other Queues** Page 16

3.1 Limitations of $M/M/1$ model 16

3.2 $M/M/1/N$ - Finite Buffer Queue 16

3.3 $M/M/\infty$ - Infinite servers 17

Chapter 4 **Appendix** Page 19

4.1 Random Processes 19

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.

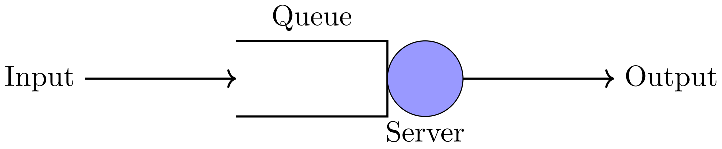


Figure 1.1: Block diagram of a simple, 1-D queue with one server and infinite capacity.

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 situation, there is an object we call a server. This server might be the processor, or the bank manager or the router.

Whenever a new entity is added to the queue, we call it as an arrival event. When the server addresses an entity, we call the event a departure event. In these notes we will refer to the entities in the queues as either arrivals or customers.

The state of the queue at any time instant is given by the number of customers at that time instant. Diagramatically, we will be using a state diagram quite frequently and hope that the reader is comfortable with it. If not, it is recommended that you refer [1].

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

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

where,

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

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

For our purposes of analysis, we make the following assumptions:

1. If the server is not serving a customer i.e., it is free, then the arriving (next in line) customer is immediately assigned that server.

2. Unless mentioned otherwise, if a server is busy, any new arrival joins the queue and waits for its chance.

3. The time between the departure of a serviced customer and the start of the next customer is zero.

Also, for simplicity, we will be sticking to a first-order analysis, thus, trading modeling accuracy for tractability and thereby insights.

1.3 Service Discipline

One thing we haven't addressed is the order in which the customers in a queue are called for service. One obvious rule is *First Come First Serve* (FCFS) also known as *First-In-First-Out* (FIFO). However, there arise scenarios where different disciplines are implemented. For example, in case of mobile communication networks, emergency notifications regarding disasters must be prioritized over all other messages. Thus, in this case, the system might follow *Last-In-First-Out-Pre-Emptive-Resume* (LIFOPR). For those working with embedded systems or computer architecture may be familiar with ‘memory stack’ which is used to manage interrupts. In this case, the order in which different routines are serviced depends on the priorities of the interrupts in addition to the order in which they are triggered. For our purposes, we keep it simple and limit to FIFO service discipline.

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 realistic, 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.

Let's understand this queue slowly, one thing at a time.

2.1 Arrivals

By definition, the arrivals in an $M/M/1$ queue are a Poisson random process. If you are not familiar, its okay, since we will derive the entire framework from basic probability and a little bit of imagination.

Consider an experiment. You are watching a queue and tracking its movement. For ease, you divide the time axis into smaller intervals of length δt such that there can atmost be a single arrival or no arrival. What decides whether there is a arrival or not? You toss a magical coin that is weirdly biased. If a toss results in heads, there is an arrival at the queue, else there is no arrival. The bias of the coin to land on its heads is proportional to the time interval i.e., δt. Thus, $\mathbb{P}(\text{Heads}) = \lambda \times \delta t$. If we consider an interval of length T, the number of slots $n \approx T/\delta t$. The experiment then reduces to n coin flips with a coin with bias $\lambda \delta t$.

The state of the system is given by the number of customers in the queue, $N(t)$ at some time instant t. From the coin analogy, we see that $N(t) \sim \text{Binomial}(n, p = \lambda \delta t)$

$$P_{N(t)}(k) = \mathbb{P}(N(t) = k) = \mathbb{P}(\text{k arrivals in the interval [0, t]}) \tag{2.1}$$

$$= \binom{n}{k} (\lambda \delta t)^k (1 - \lambda \delta t)^{n-k} \tag{2.2}$$

Taking limit as $\delta t \longrightarrow 0$ and $\delta t \approx t/n$

$$\begin{aligned}
\lim_{\delta t \to 0} P_{N(t)}(k) &= \lim_{\delta t \to 0} \frac{n!}{k!(n-k)!} \frac{\lambda^k t^k}{n^k} (1 - \frac{\lambda t}{n})^{n-k} && (2.3)\\
&= \lim_{n \to \infty} \frac{n \times n-1 \times n-2 \times \cdots \times n-(k-1)}{k!} \frac{\lambda^k t^k}{n^k} (1 - \frac{\lambda t}{n})^{n-k} && (2.4)\\
&= \frac{1}{k!} \lim_{n \to \infty} (n/n \times (n-1)/n \times (n-2)/n \times \cdots \times (n-(k-1))/n) \lambda^k t^k (1 - \frac{\lambda t}{n})^{n-k} && (2.5)\\
&= \frac{1}{k!} \lim_{n \to \infty} (1 \times (1-1/n) \times (1-2/n) \times \cdots \times (1-(k-1)/n)) \times \lambda^k t^k (1 - \frac{\lambda t}{n})^{n-k} && (2.6)\\
&= \frac{\lambda^k t^k}{k!} \lim_{n \to \infty} (1 - \frac{\lambda t}{n})^{n-k} && (2.7)\\
&= \frac{\lambda^k t^k}{k!} \lim_{n \to \infty} (1 - \frac{\lambda t}{n})^{n} && (2.8)\\
&= \frac{\lambda^k t^k}{k!} e^{-\lambda t} && (2.9)
\end{aligned}$$

Thus, we get the result,

$$P_{N(t)}(k) \sim \text{Poisson}(\lambda t) \quad (2.10)$$

Lets make a few observations.

$$\begin{aligned}
P_{N(\delta t)}(0) &= e^{-\lambda \delta t} && (2.11a)\\
&= 1 - \lambda \delta t + (\lambda \delta t)^2 - \ldots && \ldots \text{Taylor series expansion of } e^x. \quad (2.11b)\\
&\approx 1 - \lambda \delta t && \ldots \text{Neglecting higher order terms.} \quad (2.11c)\\
& && (2.11d)\\
P_{N(\delta t)}(1) &= \lambda \delta t e^{-\lambda \delta t} && (2.11e)\\
&= \lambda \delta t (1 - \lambda \delta t) && (2.11f)\\
&= \lambda \delta t && (2.11g)\\
& && (2.11h)\\
P_{N(t)}(k \geqslant 1) &\approx 0 && (2.11i)
\end{aligned}$$

Thus, we see that in the small interval δt, there is at most one arrival with probability of $\lambda \delta t$. Also there cannot be more than one arrival in the small interval.

Question 1: What does the proportionality constant λ signify?

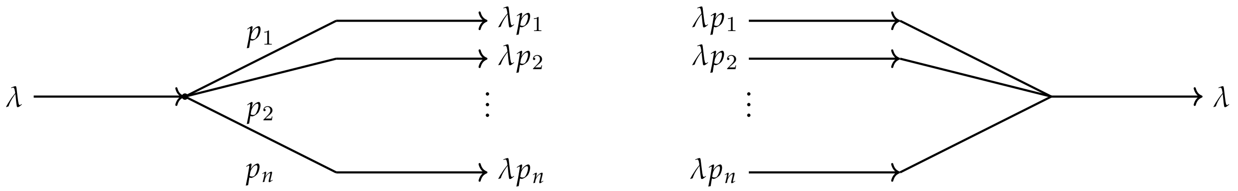
Consider the average number of arrivals per unit time i.e.,

$$\begin{aligned}
\bar{N} &= \frac{1}{t} \mathbb{E}[N(t)] && (2.12a)\\
\bar{N} &= \frac{1}{t} \sum_{k=0}^{\infty} k P_{N(t)}(k) && (2.12b)\\
&= \frac{1}{t} \lambda t && (2.12c)\\
&= \lambda && (2.12d)
\end{aligned}$$

Hence, λ is the rate of arrivals or mean arrivals per unit time.

Note:-

The Poisson random process has *independent increments* and *stationary increments*.



(a) Splitting (Thinning of Poisson Process)

(b) Merging of Poisson Process

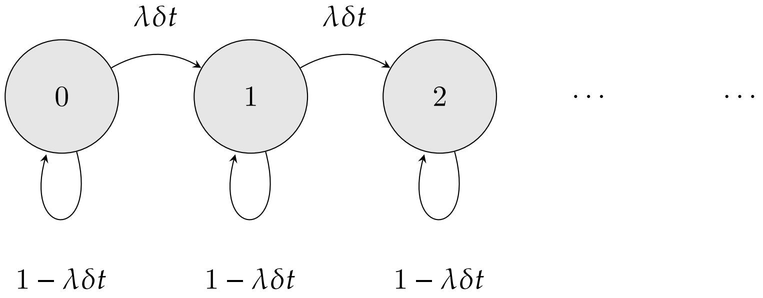


Figure 2.2: State Transition Diagram for only arrivals

Modeling a queue using such statistics has a fair share of advantages. Other than mathematically easy calculations, it is useful that splitting (thinning) and aggregation of Posson random processes results in other Poisson random processes.

Definition 2.1.1: Splitting (Thinning) of Possion Process

Consider a parent Poisson Random process with rate λ. Then, if the arrivals in this processes ar e split into n different children with probabilities $p_1, p_2, \ldots, p_n$, then the result is n Poisson processes with rates $\lambda p_1, \lambda p_2, \ldots, \lambda p_n$. Note that we must have $\sum p_i = 1$. See Fig.2.1a.

Definition 2.1.2: Merging of Poisson Process

Let $N_1(t)$ and $N_2(t)$ be two Poisson processes with rates λ_1 and λ_2 respectively. Then, the process $N(t)$ defined as $N(t) = N_1(t) + N_2(t)$ is also a Poisson random process with rate $\lambda_1 + \lambda_2$. See Fig.2.1b.

2.1.1 State of the System - Alternative approach

I would like to emphasize again that we are still considering only the arrivals and the state of the system is given by the number of customers in the system at a paticular time.

We can derive the distribution alternatively starting from Equations 2.11 as the basic setup or assumptions.

If $\mathbb{P}(\text{Number of customers in the queue} = k, \text{at time } t) = P_k(t)$, then in a single δt interval we can reach a state by either a single arrival or no arrival. We denote $p_{i,j}$ as the transition probability of going from state i to j in a δt interval.

$$\begin{aligned}
P_n(t+\delta t) &= P_n(t)p_{n,n} + P_{n-1}(t)p_{n-1,n} && \text{(2.13a)}\\
&= P_n(t)(1-\lambda\delta t) + P_{n-1}(t)(\lambda\delta t) \qquad \ldots \text{ See Fig.2.2} && \text{(2.13b)}\\
P_0(t+\delta t) &= P_0(t)p_{0,0} && \text{(2.13c)}\\
&= P_0(t)(1-\lambda\delta t) && \text{(2.13d)}
\end{aligned}$$

Thus, we arrive at a recursive equation. However, we still need a starting (boundary) condition in order to get a solution. For this, note that to be at state 0, the system must have no arrival starting at state 0 (only arrivals remember?). Reorganizing Equations 2.13,

$$\frac{P_n(t+\delta t)-P_n(t)}{\delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t) \tag{2.14a}$$

$$\frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t) \qquad \ldots\ \delta t \to 0. \tag{2.14b}$$

Case $n = 0$:

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \tag{2.14c}$$

Now those that are comfortable and in habit of working with differential equations may solve the above equation by visual inspection. Others may verify the solution by plugging the alleged function in the differential equation and verify that it satisfies the same. The solution is,

$$P_0(t) = e^{-\lambda t} \tag{2.14d}$$

Similarly, Case $n = 1$:

$$\frac{dP_1(t)}{dt} = \lambda P_1(t) + \lambda e^{-\lambda t} \tag{2.14e}$$

$$P_1(t) = \lambda t e^{-\lambda t} \tag{2.14f}$$

Recognizing a pattern, we can generalize the result without explicit proof as,

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2.14g}$$

Question 2: [2]

If a telephone exchange is known to receive 100 calls a minute on average, what is the probability, that it gets 0 calls in 5 seconds.
Solution: 0.00024.

2.1.2 Mean arrivals in an interval $[0, t]$

$$\bar{N} = \mathbb{E}[N(t)] \tag{2.15a}$$

$$= \sum_{n=0}^{\infty} n P_n(t) \tag{2.15b}$$

$$= \sum_{n=0}^{\infty} n \times \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2.15c}$$

$$= e^{-\lambda t}(\lambda t) \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} \tag{2.15d}$$

$$= e^{-\lambda t}(\lambda t) \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \tag{2.15e}$$

$$= e^{-\lambda t}(\lambda t) e^{\lambda t} \qquad \ldots \text{ Taylor Series expansion of } e^x. \tag{2.15f}$$

$$\therefore \bar{N} = \lambda t \tag{2.15g}$$

Note:-

- The results obtained is intuitively satisfactory since it states that on an average, the number of customers in a queue is proportional to the time interval of interest and the proportionality constant is given by the arrivals per unit time.
- This statistic can be used by a call center to determine a rough estimate of employees it may require to handle the calls.

2.1.3 Variance of Number of arrivals in an interval $[0, t]$

Consider,

$$\begin{aligned}
\mathbb{E}[N(t)^2] &= \sum_{n=0}^{\infty} n^2 \frac{(\lambda t)^n}{n!} e^{-\lambda t} && (2.16a)\\
&= \sum_{n=1}^{\infty} n^2 \frac{(\lambda t)^n}{n!} e^{-\lambda t} && (2.16b)\\
&= \sum_{n=1}^{\infty} n \frac{(\lambda t)^n}{(n-1)!} e^{-\lambda t} && (2.16c)\\
&= (\lambda t) e^{-\lambda t} \sum_{n=1}^{\infty} n \frac{(\lambda t)^{n-1}}{(n-1)!} && (2.16d)\\
&= (\lambda t) e^{-\lambda t} \sum_{n=0}^{\infty} (n+1) \frac{(\lambda t)^n}{(n)!} && (2.16e)\\
&= (\lambda t) e^{-\lambda t} \Big[\sum_{n=0}^{\infty} n \frac{(\lambda t)^n}{(n)!} + \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{(n)!} \Big] && (2.16f)\\
&= \lambda t e^{-\lambda t} [\lambda t e^{\lambda t} + e^{\lambda t}] && (2.16g)\\
&= \lambda(1 + \lambda t) && (2.16h)\\
&= \lambda t + \lambda^2 t^2 && (2.16i)\\
\therefore var(N) &= \lambda t + \lambda^2 t^2 - (\lambda t)^2 && (2.16j)\\
&= \lambda t. && (2.16k)
\end{aligned}$$

2.1.4 Inter-arrival times

Definition 2.1.3: Interarrival Time

The time elasped between two consecutive arrival events is called the Interarrival time between those two events. Note that this quantity is a random quantity due to our setup.

Let T denote the interarrival time. Thus,

$$\begin{aligned}
F_P(t) = \mathbb{P}(T < t) &= \mathbb{P}(\text{Interarrival time is less than} t) && (2.17a)\\
&= 1 - \mathbb{P}(\text{Interarrival time is more than} t) && (2.17b)\\
&= 1 - \mathbb{P}(\text{No arrival in time interval of length} t) && (2.17c)\\
&= 1 - P_0(t) && (2.17d)\\
&= 1 - e^{-\lambda t} && (2.17e)
\end{aligned}$$

We recognize this CDF as that of an Exponentially distributed random variable. If not, we can get the PDF by differentiating as,

$$f_T(t) = \frac{dF_T(t)}{dt} = \lambda e^{-\lambda t} \qquad (2.17f)$$

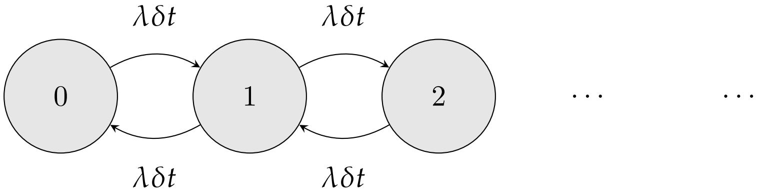


Figure 2.3: State Diagram for an $M/M/1$ system.

Thus, the interarrival time T is distributed as

$$T \sim \text{Exponential}(\lambda) \tag{2.17g}$$

Now's a good time to make some observations.

Note:-

1. The interarrival times are exponentially distributed meaning that the system memoryless. Intuitively, the current state of the system does not depend on the past. More concretely, say $\lambda = 10$ min and we have had no arrival for 6 mins. Then the time after which we expect an arrival is still 10 mins and not 4 mins! Meaning it is as good as starting a new observation window. I recommend you to go through the proof of memorylessness property of Exponential distribution.

2. Due to this, the state of the system at an instant is completely determined by the number of customers in the queue at that instant, there is no conditional dependence on the past.

3. Discrete distribution with memoryless property is the Geometric distribution.

Question 3: Where might we use this statistic?

In a call center a worker could estimate the time between the calls he would receive. Then he can plan the next coffee break and how much he can converse with a colleague.
More seriously, this statistic is important when it comes to scheduling other low priority tasks in a micro-processor or micro-controller based systems.

2.2 Service

We have ignored the poor server for a long time. Let's include him in our model now. In our $M/M/1$ model, we make the assumption that the service times for every customer are independent and identically distributed Exponential random variables with some rate μ.

Thus, we make the following assumptions in addition to our previous ones.

$$\mathbb{P}(\text{exactly 1 service in}[t, t+\delta t]) = \mu\delta t \tag{2.18a}$$
$$\mathbb{P}(\text{no service in}[t, t+\delta t]) = 1 - \mu\delta t \tag{2.18b}$$
$$\mathbb{P}(\text{more than 1 service in}[t, t+\delta t]) = 0. \tag{2.18c}$$

Now the state of the system will be given by the number of customers in the queue as well as the one in service.

The process resulting from such a system is called a *birth-death* process in the literature.[2, 3]. The state transition diagram can be now updated to include the departures as well. See Fig.2.3.

2.2.1 State of the System

Using the previous notation and differential equation approach, we now include the possiblity of a departure as well. Remember, we can jump only between adjacent states in δt. Also, we can either stay at zero to be in state 0 or depart from state 1.

$$P_n(t+\delta t) = P_n(t)p_{n,n} + P_{n-1}(t)p_{n-1,n} + P_{n+1}(t)p_{n+1,n} \tag{2.19a}$$
$$P_0(t+\delta t) = P_0(t)p_{0,0} + P_1(t)p_{1,0} \tag{2.19b}$$

Now,

$$p_{n,n} = \mathbb{P}(\text{(No arrival and no departure) or (One arrival and One departure)}) \tag{2.19c}$$
$$= (1-\lambda\delta t)(1-\mu\delta t) + (\lambda\delta t)(\mu\delta t) \tag{2.19d}$$
$$\approx 1 - \lambda\delta t - \mu\delta t \tag{2.19e}$$
$$p_{n-1,n} = \mathbb{P}(\text{One arrival and no departure}) \tag{2.19f}$$
$$= (\lambda\delta t)(1-\mu\delta t) \tag{2.19g}$$
$$\approx \lambda\delta t \tag{2.19h}$$
$$p_{n+1,n} = \mathbb{P}(\text{One departure and no arrival}) \tag{2.19i}$$
$$= (\mu\delta t)(1-\lambda\delta t) \tag{2.19j}$$
$$\approx \mu\delta t. \tag{2.19k}$$

Note:-

Higher order terms have been neglected. Also, We cannot have a departure in state 0.

Similar to previous derivations, we get the following equations:

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) \tag{2.20a}$$
$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P(t) \tag{2.20b}$$

Solving these equations can be arduous. Thus, [2] defines a new quantity. Let's see how this will come to our aid.

Definition 2.2.1: Probability Flux

Probability Flux is defined as the product of the probability of being in the state at which a transition originates and the transition rate to which the state travels next.

When the system is in equilibrium, we can say that the probability flux into a state equals the probability flux out of the state. This is called flow balancing and leads to what are known as local(global) balancing equations.

When in equilibrium we assume that $\frac{dP_n(t)}{t} = 0$ Thus, we get the following "local balance equations",

$$\lambda p_0 = \mu p_1 \qquad \Longrightarrow p_1 = (\lambda/\mu)p_0 \tag{2.21a}$$
$$\lambda p_1 = \mu p_2 \qquad \Longrightarrow p_2 = (\lambda/\mu)p_1 \tag{2.21b}$$
$$\vdots \tag{2.21c}$$
$$\lambda p_{n-1} = \mu p_n \qquad \Longrightarrow p_{n-1} = (\lambda/\mu)p_n \tag{2.21d}$$
$$\vdots \tag{2.21e}$$
$$\tag{2.21f}$$

Thus, we get the equation,

$$p_n = (\lambda/\mu)^n p_0 \tag{2.21g}$$

Since the system must be in atleast one the possible states, the axioms of probability give us another equation,

$$\sum_{i=0}^{num_{states}} p_i = 1. \tag{2.22}$$

Defining $\rho = \lambda/\mu$ for our $M/M/1$ with infinite states, we get,

$$p_0 + \rho p_0 + \rho^2 p_0 + \ldots = 1 \tag{2.23}$$

$$\therefore p_0 = \frac{1}{\sum_{i=0}^{\infty} \rho^i} \tag{2.24}$$

Consider the case when $0 \leqslant \rho < 1$,

$$p_0 = \frac{1}{\frac{1}{1-\rho}} \tag{2.25}$$

$$= 1 - \rho. \tag{2.26}$$

Note:-

Since p_0 is the probability that the queue is empty or unutilized, we call ρ Utilization. When λ is close to 0, we say we have very light load or zero load. Alternatively, if $\lambda \longrightarrow \mu$, we say we have heavy load.

Question 4: Is it possible that $\rho > 1$?

Solution: No. From the above derivation we see that the length of the queue will balloon up and blow up to be infinite in size. Hence, the system will not reach equilibrium, invalidating our above assumptions.

2.2.2 Average number of customers in $M/M/1$ queue

We will now use the updated distribution.

$$\bar{N} = \mathbb{E}[N(t)] = \sum_{n=0}^{\infty} nP_n(t) \tag{2.27a}$$

$$= \sum_{n=0}^{\infty} n\rho^n p_0 \tag{2.27b}$$

$$= (1-\rho)\sum_{n=0}^{\infty} n\rho^n \tag{2.27c}$$

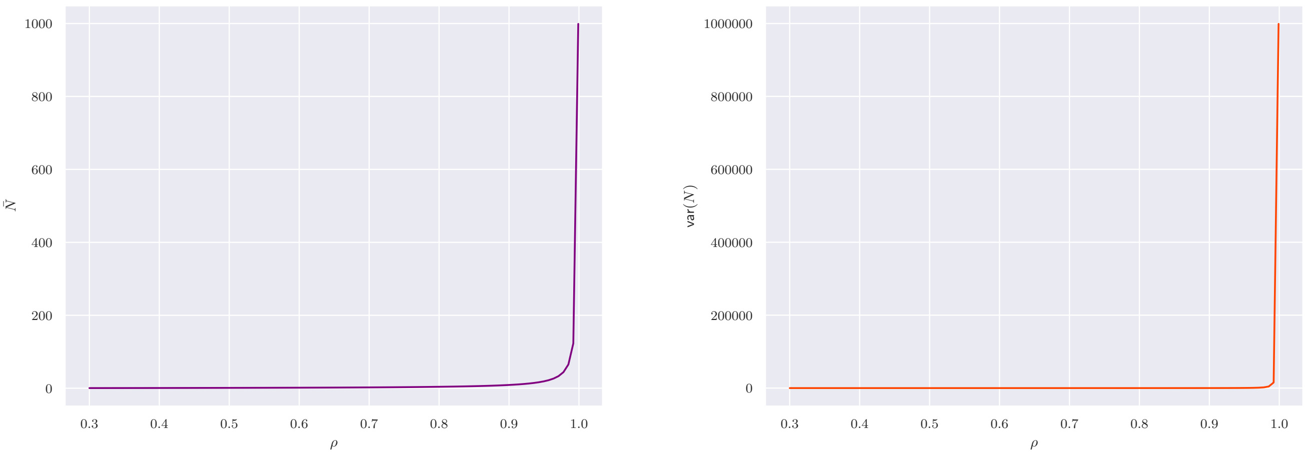
$$= \rho(1-\rho)\sum_{n=0}^{\infty} n\rho^{n-1} \tag{2.27d}$$

$$= \rho(1-\rho)\sum_{n=0}^{\infty} \frac{d\rho^n}{d\rho} \tag{2.27e}$$

$$= \rho(1-\rho)\frac{d\sum_{n=0}^{\infty} \rho^n}{d\rho} \tag{2.27f}$$

$$= \rho(1-\rho)\frac{d1/(1-\rho)}{d\rho} \tag{2.27g}$$

$$= \frac{\rho}{1-\rho} \tag{2.27h}$$



(a) Average number of customers in an $M/M/1$ queue.

(b) Variance of number of customers in an $M/M/1$ queue.

2.2.3 Variance of number of customers in $M/M/1$ queue

Consider,

$$\mathbb{E}[N^2] = \sum_{n=0}^{\infty} n^2(1-\rho)(\rho^n) \tag{2.28a}$$

$$= (1-\rho)\big[\sum_{n=0}^{\infty} n(n+1-1)\rho^n\big] \tag{2.28b}$$

$$= (1-\rho)\big[\sum_{n=0}^{\infty} n(n+1)\rho^n - \sum_{n=0}^{\infty} n\rho^n\big] \tag{2.28c}$$

$$= (1-\rho)\big[\rho^2\sum_{n=0}^{\infty} n(n+1)\rho^{n-1} - \rho\sum_{n=0}^{\infty} n\rho^{n-1}\big] \tag{2.28d}$$

$$= (1-\rho)\big[\rho^2\sum_{n=0}^{\infty} \frac{d^2\rho^{n+1}}{d\rho^2} - \rho\sum_{n=0}^{\infty} \frac{d\rho^n}{d\rho}\big] \tag{2.28e}$$

$$= (1-\rho)\big[\rho^2\frac{d^2\big(\sum_{n=1}^{\infty}\rho^n\big)}{d\rho^2} - \rho\frac{d\big(\sum_{n=0}^{\infty}\rho^n\big)}{d\rho}\big] \tag{2.28f}$$

Using the formula for sum of Geometric Progression (for common ratio < 1),

$$= (1-\rho)\big[\rho^2\frac{2}{(1-\rho)^3} + \frac{\rho}{(1-\rho)^2}\big] \tag{2.28g}$$

$$= \frac{\rho(1+\rho)}{(1-\rho)^2} \tag{2.28h}$$

$$\therefore \text{var}(N) = \frac{\rho(1+\rho)}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} \tag{2.28i}$$

$$= \frac{\rho}{(1-\rho)^2}. \tag{2.28j}$$

Please see Figs.2.4a and 2.4b. Note how the statistics blow up as the Utilization factor approaches 1.

2.3 Little's Law

Now that we have come up so far, lets try to relate the number in the queue with the time dimension.

Theorem 2.3.1 Little's Law

The mean number of customers in the queue, $\bar{N}$ is related to the rate of arrivals, λ, and the mean service time, $\bar{T}$ as

$$\bar{N} = \lambda \bar{T}.$$

Proof: Consider a time interval of $[0, T]$.

$$\mathbb{P}(\text{n customers in the queue after one departure}) \tag{2.29a}$$

$$= \mathbb{P}(\text{n customers arrive during the time spent in the queue by the depart customer.}) \tag{2.29b}$$

$$= \int_0^\infty \mathbb{P}(\text{n arrivals in during the interval}[0, T] \mid T = t) f_T(t) dt \tag{2.29c}$$

$$= \int_0^\infty \mathbb{P}(\text{n arrivals in during the interval}[0, t]) f_T(t) dt \tag{2.29d}$$

$$= \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda} f_T(t) dt \tag{2.29e}$$

Now, using formula for expectation of a discrete random variable,

$$\bar{N} = \mathbb{E}[N] \tag{2.29f}$$

$$= \sum_{n=0}^{\infty} n P_n(t) \tag{2.29g}$$

$$= \sum_{n=0}^{\infty} n \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda} f_T(t) dt \tag{2.29h}$$

$$= \int_0^\infty \sum_{n=1}^{\infty} n \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_T(t) dt \tag{2.29i}$$

$$= \int_0^\infty \sum_{n=1}^{\infty} \lambda t \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t} f_T(t) dt \tag{2.29j}$$

$$= \int_0^\infty \lambda t e^{-\lambda t} f_T(t) \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} dt \tag{2.29k}$$

$$= \int_0^\infty \lambda t e^{-\lambda t} e^{\lambda t} f_T(t) dt \qquad \ldots \text{ Taylor series expansion of } e^x. \tag{2.29l}$$

$$= \lambda \int_0^\infty t f_T(T) dt \tag{2.29m}$$

$$= \lambda \mathbb{E}[T] \tag{2.29n}$$

$$= \lambda \bar{T}. \tag{2.29o}$$

😄

Note:-

This result can be interpreted in simple way. The average waiting, $\bar{T}$, is the amount of time a customer is expected to wait once it arrives at the queue and until it leaves after being serviced. Moreover, λ is the mean number of arrivals per unit time. Thus, in the time that a customer is in a queue, is likely to see $\lambda \bar{T}$ customers on an average.

Chapter 3

Other Queues

3.1 Limitations of $M/M/1$ model

While the previous queueing model was friendly enough that we could derive the nice results that we got, it isn't all that realistic. Recollect the assumptions that we made prior to embarking on the derivations. One of them was that the coin tosses that decide whether there is an arrival or not are independent and thus uncorrelated. However, how realistic is this assumption?

Consider the case of a call center. In my experience at least, I have hardly ever got any resolution on my first call. What I mean is that once you make a call to a place, it is quite likely that you will call again. That means that the calls are not completely uncorrelated. However, this assumption is ok to make for a first-order analysis.

Next assumption that we made was that the queue could in theory grow limitlessly. This is quite impossible simply due to availability of limited resources and not to mention the patience one would require to stand in an infinite human queue. Hence, for next part of our analysis we consider a model with finite buffer length.

After this analysis, we ask why to limit our system to a single server queue? After all, the efficiency lies in parallelism. Thus, we consider the cases of 'm' servers and a special case where $m = \infty$.

The analysis in the following sections will be similar to the previous ones and will heavily borrow from those sections. We request, that you are comfortable and up-to-speed on the concepts covered so far.

3.2 $M/M/1/N$ - Finite Buffer Queue

In this model, the queue size is limited to N. If there are N customers in the queue, including one in the server, then any new arrivals are turnew away or dropped. To maintain the independence of arrivals, we further assume that the dropped customers do not arrive again.

Due to the finite states we get a new state transition diagram shown in Fig.3.1.

Recollecting the local balance equations, we have under equilibrium,

$$p_n = (\frac{\lambda}{\nu})^n p_0 = \rho^n p_0. \tag{3.1a}$$

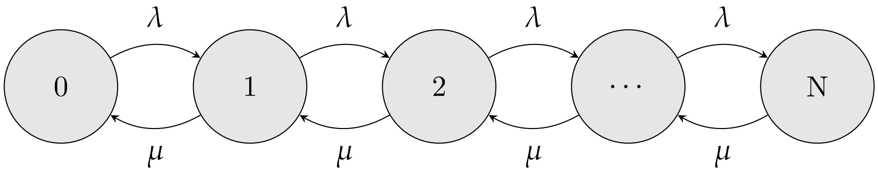


Figure 3.1: State transition diagram for $M/M/1/N$ queue.

Unlike previous model however, $0 \leqslant n \leqslant N$. Thus, to satisfy the axioms of probability, we must have,

$$
\begin{aligned}
p_0 + p_1 + \cdots + p_N &= 1 &\text{(3.1b)}\\
\implies p_0 + \rho p_0 + \rho^2 p_0 + \cdots + \rho^N p_0 &= 1 &\text{(3.1c)}\\
p_0(1 + \rho + \rho^2 + \cdots + \rho^N) &= 1 &\text{(3.1d)}\\
p_0(\frac{1-\rho^{N+1}}{1-\rho}) &= 1 &\text{(3.1e)}\\
\therefore p_0 &= \frac{1-\rho}{1-\rho^{N+1}} &\text{(3.1f)}
\end{aligned}
$$

Substituting in 3.1a, for $0 \leqslant n \leqslant N$,

$$p_n = (\frac{1-\rho}{1-\rho^{N+1}})\rho^n. \tag{3.1g}$$

Definition 3.2.1: Blocking Probability

The blocking probability is the probability that the queue buffer is full, i.e., $P_N(t)$. Using the derived equations, Blocking Probability is $\frac{(1-\rho)}{(1-\rho^{N+1})}\rho^N$.

Note:-

Using the blocking probability, we can find the mean number of customers turned away per unit time as $\lambda P_N(t)$.

Question 5: What is p_n when $\lambda/\mu = 1$?

At $\lambda/\mu = 1$, p_n assumes a $0/0$ form. We use L' Hopital's rule.

$$
\begin{aligned}
\lim_{\rho\to 1} p_n &= lim_{\rho\to 1}\frac{1-\rho}{1-\rho^{N+1}}\rho^n &\text{(3.2)}\\
&= \lim_{\rho\to 1}\frac{-1}{-(N+1)\rho^N} &\text{(3.3)}\\
&= \frac{1}{N+1}. &\text{(3.4)}
\end{aligned}
$$

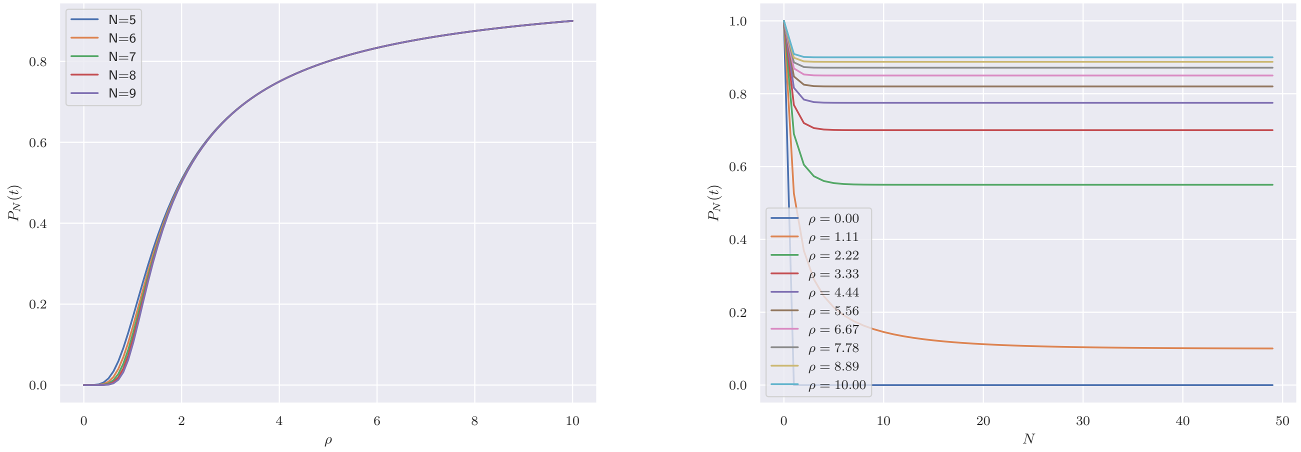
Note:-

In this case, since the buffer size is limited, we can have $\lambda > \mu$ as the excess arrivals will simply be turned down or blocked.

We plot the blocking probabilities in Fig.3.2a and 3.2b. It concurs with our intuition that the blocking probability increases with increasing utilization (ρ) and decreases with increasing buffer length, N.

3.3 $M/M/\infty$ - Infinite servers

Now, we move on to the case of multiple server models. However, to warm, it is easier to consider the infinite server case as we will soon see. The setup we now have to work with is such that every arriving customer is assigned a personal server of rate μ. Since, we have infinite servers, we don't mind it if the queue size increases limitlessly.

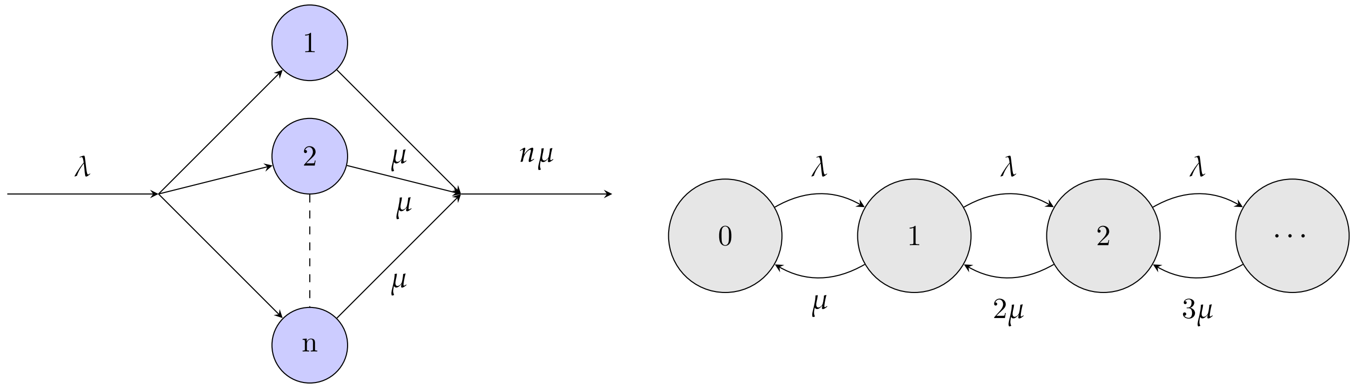


(a) Blocking probability as a function of ρ (N fixed).

(b) Blocking probability as a function of N (ρ fixed).

Note:-

If the system has n customers at a time instant, as a whole the system has an aggregate output rate of $n\mu$.



(a) Flow diagram of an $M/M/\infty$ queue.

(b) Flow diagram of an $M/M/\infty$ queue.

Chapter 4

Appendix

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

4.1 Random Processes

Definition 4.1.1: 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.

[3] William J. Stewart. *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling.* Princeton University Press, 2009.