Queueing Theory - Seminar Notes

A Mathematically-Rigorous yet Intuitive Introduction to Queues

Rishabh Pomaje
210020036@iitdh.ac.in

Autumn 2024-25

Indian Institute of Technology Dharwad

Contents

Chapter 1

Motivation and Background

Queues arise in nature by the virtue of limited quantity and efficiency of resources being available. At the 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 [5], that Queueing Theory is the *Study of Waiting*. What? I hear you ask. What's there to study about waiting? Let me assure you, a lot. Just to show how often scenarios arise where we have to wait, recall the instances when you waited hours at the bank just to get your passbook updated or at an 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 is Wi-Fi. All of our devices—smartphones, laptops, PCs, and nowadays even TVs, refrigerators, and wristwatches—connect to it. When wanting to communicate with another device in some other location, these devices are essentially lining up in a queue at the router, waiting for their work—in this case, their data—to be processed. Then there are queues formed by interrupts and faults in electrical systems, especially those with an operating system or real-time computing systems. We see that queues are ubiquitous. We just have to look for them!

1.2 Modeling

An immediate question that might arise is: what is there to model in a queue? The arrivals in a queue are spaced in time, and these intervals often appear random to us. Moreover, the demand that each arrival places on the "server" is also of random size. Therefore, we can model the stochastic nature of arrival times and service requirements using probability distributions.

Modeling a queueing system, as it is commonly known, can provide valuable insights that help us make informed decisions about resource planning and allocation, ultimately optimizing the output of a system. Later, we will provide examples of how some basic statistics can be used in practice.

1.3 Groundwork

Now we will start by making a few definitions and concepts concrete.

1.3.1 Block diagram

A queue is frequently depicted in literature as shown in Fig.1.1.

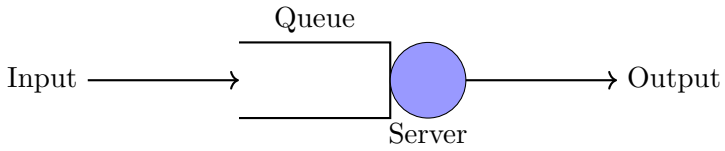


Figure 1.1: Block diagram of a simple, 1-D queue with one server and infinite capacity.

1.3.2 Terminology

The Fig.1.1 has a few terms used in it. A queue is formed by entities, the nature of which 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, humans waiting at a bank, or airplanes at an airport awaiting takeoff. To address the requirements of the entities, there is an object we call a server. This server might be the processor, the bank manager, or the air traffic controller, respectively, in the aforementioned queues.

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

The state of the queue at any given instant is defined by the number of customers at that moment. This includes those waiting in the queue for their turn at the server, as well as the one currently being served by the server. 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 to the relevant chapters in [4].

1.3.3 Assumptions

To make our life easy we make a few seemingly trivial yet important assumptions [6]:

1. If a server is free, an arriving customer directly goes into service. If a server is not free, the customer waits in the queue waiting for its turn at the server.
2. When a server becomes free, the next customer is chosen according to a *scheduling policy*.
3. Customers remain in the "queueing facility" once they have been accepted in it and do not become impatient and leave.

1.3.4 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, etc. *Kendall's Notation* provides a condensed way to represent a queue. The general form is given by

$$\boldsymbol{A/B}/c/\boldsymbol{X/Y/Z} \tag{1.1a}$$

where,

$\boldsymbol{A}$ = Describes arrival statistics.
$\boldsymbol{B}$ = Describes the departure statistics.
c = Number of Servers.
$\boldsymbol{X}$ = Maximum length of the queue, more commonly known as *system capacity*.
$\boldsymbol{Y}$ = Size of the customer population.
$\boldsymbol{Z}$ = Queue scheduling discipline.

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, 'E' for Erlang,'G' for general statistics and 'Geom' for Geometric.

1.3.5 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.

1.4 Service Discipline

One thing we haven't addressed so far, is the order in which the customers in a queue are called for service.

Definition 1.4.1: Scheduling Discipline

The rule that decides which customer in the queue will be serviced after a departure is known as the scheduling discipline.
This rule is also referred to as the scheduling algorithm or scheduling policy in the literature.

One obvious rule is First Come First Serve (FCFS), also known as First-In-First-Out (FIFO). However, there are scenarios where different disciplines are implemented. For example, in mobile communication networks, emergency notifications regarding disasters must be prioritized over all other messages. In this case, the system might follow Last-In-First-Out-Pre-Emptive-Resume (LIFOPR). Readers associated with embedded systems or computer architecture may be familiar with the "memory stack" which is used to manage interrupts or function calls. 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 (LIFO is the most commonly used discipline).

For our purposes, we will keep it simple and limit ourselves to the FIFO service discipline. Now will look at a few queues with models that provide feasible and tractable mathematical models.

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 delve into its details, may not seem entirely realistic, it provides 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 follows a Poisson random process.
2. There is a single server, with the service times for each customer being independent and exponentially distributed.
3. There is no limit on the size of the queue. Additionally, the state of the queue is given by the number of arrivals/customers in the queue at a given moment.

Its our first queue under study, let's understand this queue slowly, building the model step by step.

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. However, you may refer [4] for more details.

2.1.1 Poisson Random Arrival Process

Consider an experiment where you are observing a queue and tracking its movement. For simplicity, you divide the time axis into smaller intervals of length δt, such that at most one arrival can occur within each interval, or there may be no arrival at all. What decides whether there is an arrival or not? You toss a magical coin that is weirdly biased. If the toss results in heads, there is an arrival at the queue; otherwise, there is no arrival. The bias of the coin landing on heads is proportional to the length of the time interval, i.e.,

$$\mathbb{P}(\text{Heads}) = \lambda \times \delta t,$$

where λ is a constant. Now, if we consider an interval of length T, the number of slots (or small δt intervals) within that time interval is approximately,

$$n \approx \frac{T}{\delta t}.$$

Therefore, the experiment reduces to n coin flips, each with a bias of $\lambda \delta t$.

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 $\delta t \to 0$ and $\delta t \approx t/n$ (this implies that $n \to \infty$),

$$\lim_{\delta t \to 0} P_{N(t)}(k) = \lim_{\delta t \to 0} \frac{n!}{k!(n-k)!} \frac{\lambda^k t^k}{n^k} \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{2.3}$$

$$= \lim_{n \to \infty} \frac{n \times (n-1) \times (n-2) \times \cdots \times (n-(k-1))}{k!} \frac{\lambda^k t^k}{n^k} \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{2.4}$$

$$= \frac{1}{k!} \lim_{n \to \infty} \left(\frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-(k-1)}{n}\right) \lambda^k t^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{2.5}$$

$$= \frac{1}{k!} \lim_{n \to \infty} \left(1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{k-1}{n}\right)\right) \lambda^k t^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{2.6}$$

$$= \frac{\lambda^k t^k}{k!} \lim_{n \to \infty} \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{2.7}$$

$$= \frac{\lambda^k t^k}{k!} \lim_{n \to \infty} \left(1 - \frac{\lambda t}{n}\right)^{n} \tag{2.8}$$

$$= \frac{\lambda^k t^k}{k!} e^{-\lambda t} \tag{2.9}$$

Thus, we get the result,

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

Lets make a few observations.

$$P_{N(\delta t)}(0) = e^{-\lambda\delta t} \tag{2.11a}$$

$$= 1 - \lambda\delta t + (\lambda\delta t)^2 - \ldots \qquad \ldots \text{Taylor series expansion of } e^x. \tag{2.11b}$$

$$\approx 1 - \lambda\delta t \qquad \ldots \text{Neglecting higher order terms.} \tag{2.11c}$$

$$\tag{2.11d}$$

$$P_{N(\delta t)}(1) = \lambda\delta t e^{-\lambda\delta t} \tag{2.11e}$$

$$= \lambda\delta t(1 - \lambda\delta t) \tag{2.11f}$$

$$= \lambda\delta t \qquad \ldots \text{Ignoring higher order terms.} \tag{2.11g}$$

$$\tag{2.11h}$$

$$P_{N(t)}(k \geqslant 1) \approx 0 \tag{2.11i}$$

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.

2.1.2 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

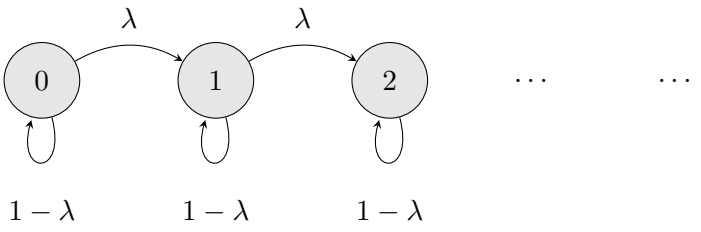


Figure 2.1: State Transition Diagram for only arrivals

of going from state i to j in a δt interval.

$$P_n(t+\delta t) = P_n(t)p_{n,n} + P_{n-1}(t)p_{n-1,n} \tag{2.12a}$$
$$= P_n(t)(1-\lambda\delta t) + P_{n-1}(t)(\lambda\delta t) \quad \ldots \text{ See Fig.2.1} \tag{2.12b}$$
$$P_0(t+\delta t) = P_0(t)p_{0,0} \tag{2.12c}$$
$$= P_0(t)(1-\lambda\delta t) \tag{2.12d}$$

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 2.12,

$$\frac{P_n(t+\delta t) - P_n(t)}{\delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t) \tag{2.13a}$$
$$\frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t) \quad \ldots\ \delta t \to 0. \tag{2.13b}$$

Case $n = 0$:

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \tag{2.13c}$$

Now, those who are comfortable with and accustomed to working with differential equations may solve the above equation by visual inspection. Others may verify the solution by plugging the alleged function into the differential equation and checking that it satisfies the equation. The solution is,

$$P_0(t) = e^{-\lambda t} \tag{2.13d}$$

Similarly, Case $n = 1$:

$$\frac{dP_1(t)}{dt} = \lambda P_1(t) + \lambda e^{-\lambda t} \tag{2.13e}$$
$$P_1(t) = \lambda t e^{-\lambda t} \tag{2.13f}$$

Recognizing a pattern, we can generalize the result without explicit proof (provided by induction in [6]) as,

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2.13g}$$

Thus, we reach the same result that we derived in the previous section. The difference is that we started from two different, yet equivalent, definitions of the Poisson random process.

Note:-

The Poisson random process has *independent increments* and *stationary increments*[4]. This, means that the number of arrivals in two disjoint time intervals are independent and the associated statistics are time-invariant, i.e., they depend only on the length of the interval.

Modeling a queue using such statistics has a fair share of advantages. Other than easy mathematical derivations, it is useful that splitting (thinning) and aggregation of Poisson random processes results in other independent Poisson random processes. See the appendix if you are unfamiliar with splitting and merging of Poisson random processes.

Question 1: [5]

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.

Now that we have the distribution of the state of the queue, we can find a few statistics that might help us in making decisions. We begin by finding the mean number of customers in the queue, denoted by $\bar{N}$.

2.1.3 Mean number of arrivals in an interval $[0, t]$

$$\bar{N}(t) = \mathbb{E}[N(t)] \tag{2.14a}$$
$$= \sum_{n=0}^{\infty} nP_n(t) \tag{2.14b}$$
$$= \sum_{n=0}^{\infty} n \times \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2.14c}$$
$$= e^{-\lambda t}(\lambda t) \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} \tag{2.14d}$$
$$= e^{-\lambda t}(\lambda t) \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \tag{2.14e}$$
$$= e^{-\lambda t}(\lambda t)e^{\lambda t} \quad \ldots \text{ Taylor Series expansion of } e^x. \tag{2.14f}$$
$$\therefore \bar{N}(t) = \lambda t \tag{2.14g}$$

Note:-

- The result obtained is intuitively satisfactory, since it states that, on average, the number of customers in a queue is proportional to the time interval of interest. Let us now see what this proportionality constant means.

Question 2: What does the proportionality constant λ signify?

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

$$\bar{N} = \frac{1}{t}\mathbb{E}[N(t)] \tag{2.15a}$$
$$\bar{N} = \frac{1}{t}\sum_{k=0}^{\infty} kP_{N(t)}(k) \tag{2.15b}$$
$$= \frac{1}{t}\lambda t \tag{2.15c}$$
$$= \lambda \tag{2.15d}$$

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

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

Consider,

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

2.1.5 Inter-arrival times

Definition 2.1.1: Interarrival Time

The time elasped between two consecutive arrival events is called the Inter-arrival time between those two events. Note that this is a stochastic quantity in our model.

Let T denote the interarrival time. Thus we begin by finding the cumulative distribution function (CDF) of T,

$$F_T(t) = \mathbb{P}(T \leqslant t) = \mathbb{P}(\text{Interarrival time is less than or equal to } t) \tag{2.17a}$$
$$= 1 - \mathbb{P}(\text{Interarrival time is more than } t) \tag{2.17b}$$
$$= 1 - \mathbb{P}(\text{No arrival in time interval of length } t) \tag{2.17c}$$
$$= 1 - P_0(t) \tag{2.17d}$$
$$= 1 - e^{-\lambda t} \tag{2.17e}$$

We recognize this CDF as that of an *exponentially* distributed random variable. If you don't, we can get the PDF (which hopefully you will be more familiar with) by differentiating as,

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

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.

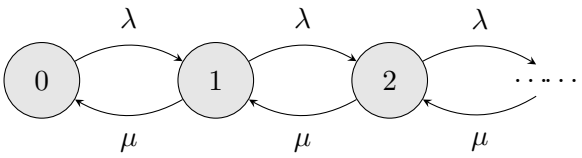


Figure 2.2: State Diagram for an $M/M/1$ system.

Note:-

1. The inter-arrival times are exponentially distributed, meaning that the system is memoryless (see [4]). 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 minutes since we started observing. Then, the time after which we expect an arrival is still 10 minutes from now, not 4 minutes! This means it is as if we are starting a new observation window. I recommend that you go through the proof of the memorylessness property of the exponential distribution in [4], as it is quite simple.
2. Due to this, the state of the system at any given instant is completely determined by the number of customers in the queue at that instant; there is no conditional dependence on the past.
3. A discrete distribution with the memoryless property is the Geometric distribution.

2.2 Service

We have ignored the server for a long time. Let's include it 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 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[5, 6]. The state transition diagram can be now updated to include the departures as well. See Fig.2.2.

2.2.1 State of the System

Using the previous notation and the differential equation approach, we now include the possibility of a departure as well. Remember, we can only jump between adjacent states in a single δt interval. Also, when we are in state 0, we can either stay in state 0 or depart to state 1. Thus, we get the equations,

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

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 for a first order analysis. Also, we cannot have a departure in state 0.

Similar to the previous derivations, we obtain the following differential equations:

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) \tag{2.20a}$$
$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \tag{2.20b}$$

Solving these equations can be arduous especially for the transient phase of the system i.e., when the state probabilities have not settled or become *stationary*. For a first-cut analysis we are satisfied with the stationary distribution or the state probabilities when the system has reached equilibrium.

Question 3: When does a Markov chain achieve equilibrium?

Solution: It is apparent from the state diagram, Fig.2.2, that we are dealing with a discrete time Markov chain. A Markov chain is said to be in equilibrium or stationary when the state distribution is no longer time-variant. It is this distribution that we are most interested about. Moving forward, we will only concentrate on finding the stationary distributions.

The formation of differential equations though easy was time consuming. Now we define a new quantity called *Probability Flux* that will help us form recursive equations of the stationary distributions which we can easily solve.

Definition 2.2.1: Probability Flux [5]

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.
It can be physically understood as the mean number of transitions that occur per unit time. For example, if $\lambda = 10$ sec^{-1} and probability of being in originating state is 0.5, then on an average the systems makes that transition five times every second.

Probability flux can be understood better if we make an analogy. However, beware that the following analogy is only a loose one and does not extrapolate to other results. We can relate a node in the state transition diagram to a node in an electrical circuit. Just as Kirchhoff's Current Law states that the current into a node must equal the current out of it, similarly, the flow of transitions into a node of a state transition diagram must equal the flow out of it in equilibrium. The resultant equations that one gets are collectively known as *Global Balance Equations*.

For those not yet convinced, lets just go by the definition of Probability flux. Take a node say node 1 in Fig.2.2. If we add the probability fluxes together with the signs representing the direction,

we get,

$$\Phi_P(\text{node 1}) = \text{Flow out of node 1} + \text{Flow into node 1} \tag{2.21}$$
$$= -(\lambda+\mu)p_1 + \lambda p_0 + \mu p_2 \tag{2.22}$$

If you observe carefully, this is the right-hand side of Eq. (2.20a) for $n = 1$. So, if we are at equilibrium, we expect that the state probabilities be time-invariant, i.e., $\frac{dp_n(t)}{dt} = 0$. Thus, $\Phi_P(\text{node1}) = 0$, i.e., the total flux into a node equals the total flux out of a node. We can use these to form recursive equations for the state probabilities at equilibrium and obtain the stationary distribution.

In conclusion, when the system is in equilibrium, 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 the balancing equations.

To further simplify this process, we can form the local balance equations which essentially say that the flow toward the right, across a boundary separating two states of a system, equals the flow towards the left[1]

Going back to our queue model, we get the following local balance equations,

$$\lambda p_0 = \mu p_1 \qquad \implies p_1 = \frac{\lambda}{\mu} p_0 \tag{2.23a}$$
$$\lambda p_1 = \mu p_2 \qquad \implies p_2 = \frac{\lambda}{\mu} p_1 \tag{2.23b}$$
$$\vdots \tag{2.23c}$$
$$\lambda p_{n-1} = \mu p_n \qquad \implies p_n = \frac{\lambda}{\mu} p_{n-1} \tag{2.23d}$$
$$\vdots \tag{2.23e}$$
$$\tag{2.23f}$$

Thus, we get the general equation[2],

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \tag{2.23g}$$

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

$$\sum_{i=0}^{\text{num}_{\text{states}}} p_i = 1. \tag{2.24}$$

Defining $\rho = \frac{\lambda}{\mu}$ for our $M/M/1$ system with an infinite number of states, we get:

$$p_0 + \rho p_0 + \rho^2 p_0 + \ldots = 1 \tag{2.25}$$
$$\therefore p_0 = \frac{1}{\sum_{i=0}^{\infty} \rho^i} \tag{2.26}$$

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

$$p_0 = \frac{1}{\frac{1}{1-\rho}} \tag{2.27}$$
$$= 1 - \rho. \tag{2.28}$$

[1] This process of forming the local balance equations is easy only when the states are 1-D. But when the states are drawn in higher dimensions, it is claimed in [6] that a cluster-based approach would be easier. These structures being beyond the scope of these notes have not been verified or even looked into by us.

[2] [6] provides an arguably better proof by principle of mathematical induction in our opinion which you may refer.

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 a very light load or zero load. Alternatively, if $\lambda \to \mu$, we say we have a heavy load.

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

Solution: No (atleast for reaching equilibrium). 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.

Definition 2.2.2: Utilization

In a single server system, Utilization is defined as the fraction of time the server is busy.
In case of multiple servers, we Utilization as the average fraction of available servers that are busy.

Question 5: What is utilization for $M/M/1$ queue? Is it always the same?

Solution: Consider a time interval of length T. Then the mean arrivals in this time interval will be λT. The server serves at an average rate of μ per unit time. Thus, the amount of time the server will be busy is given by $\lambda T/\mu$. Normalizing to find the fraction of time the server is busy, Utilization $= \frac{1}{T}(\lambda T/\mu) = \lambda/\mu$ which is why ρ is also called utilization.

In some cases, where customers are denied access to the queue due to a finite buffer size, the utilization is not exactly ρ, but slightly less than it, as there are fewer than λ mean arrivals due to excess customers being dropped.

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

We will now use the updated distribution (for steady state) that includes the service time distribution as well.

$$\bar{N} = \mathbb{E}[N] = \sum_{n=0}^{\infty} n p_n \tag{2.29a}$$

$$= \sum_{n=0}^{\infty} n \rho^n p_0 \tag{2.29b}$$

$$= (1-\rho) \sum_{n=0}^{\infty} n \rho^n \tag{2.29c}$$

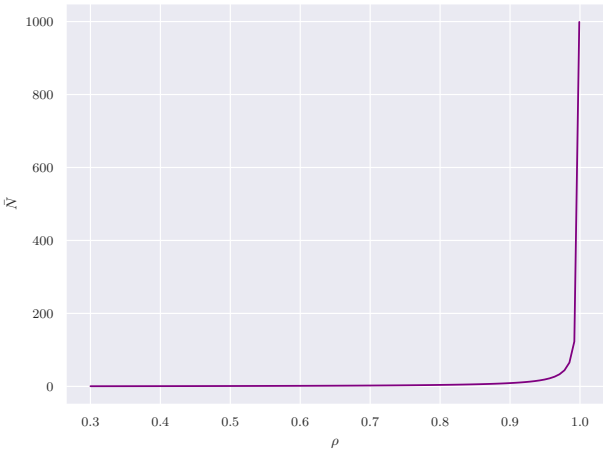
$$= \rho(1-\rho) \sum_{n=0}^{\infty} n \rho^{n-1} \tag{2.29d}$$

$$= \rho(1-\rho) \sum_{n=0}^{\infty} \frac{d\rho^n}{d\rho} \tag{2.29e}$$

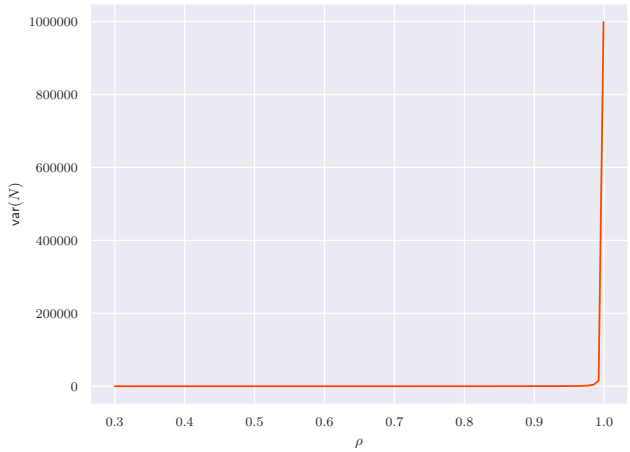
$$= \rho(1-\rho) \frac{d}{d\rho} \left(\sum_{n=0}^{\infty} \rho^n \right) \tag{2.29f}$$

$$= \rho(1-\rho) \frac{d}{d\rho} \left(\frac{1}{1-\rho} \right) \tag{2.29g}$$

$$= \frac{\rho}{1-\rho} \tag{2.29h}$$



(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,

$$\mathbb{E}[N^2] = \sum_{n=0}^{\infty} n^2(1-\rho)(\rho^n) \tag{2.30a}$$

$$= (1-\rho)\big[\sum_{n=0}^{\infty} n(n+1-1)\rho^n\big] \tag{2.30b}$$

$$= (1-\rho)\big[\sum_{n=0}^{\infty} n(n+1)\rho^n - \sum_{n=0}^{\infty} n\rho^n\big] \tag{2.30c}$$

$$= (1-\rho)\big[\rho^2 \sum_{n=0}^{\infty} n(n+1)\rho^{n-1} - \rho\sum_{n=0}^{\infty} n\rho^{n-1}\big] \tag{2.30d}$$

$$= (1-\rho)\big[\rho^2 \sum_{n=0}^{\infty} \frac{d^2\rho^{n+1}}{d\rho^2} - \rho\sum_{n=0}^{\infty} \frac{d\rho^n}{d\rho}\big] \tag{2.30e}$$

$$= (1-\rho)\big[\rho^2 \frac{d^2\big(\sum_{n=0}^{\infty}\rho^n\big)}{d\rho^2} - \rho\frac{d\big(\sum_{n=0}^{\infty}\rho^n\big)}{d\rho}\big] \tag{2.30f}$$

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

$$= (1-\rho)\big[\rho^2\frac{2}{(1-\rho)^3} + \frac{\rho}{(1-\rho)^2}\big] \tag{2.30g}$$

$$= \frac{\rho(1+\rho)}{(1-\rho)^2} \tag{2.30h}$$

$$\therefore \text{var}(N) = \frac{\rho(1+\rho)}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} \tag{2.30i}$$

$$= \frac{\rho}{(1-\rho)^2}. \tag{2.30j}$$

Please see Figs.2.3a and 2.3b. Note how the statistics blow up as the Utilization factor approaches 1. Due to the high variance as $\rho \to 1$, the randomness in the system increases i.e., the unpredictability in the queue is very high. How do you think this might affect analysis of such a queue?

Question 6: Where might we use these statistics?

Solution: As I promised, we will take an approach that will make things clear intuitively. What better way than using a real-life example?

In a call center, a clever worker who is a queueing theory aficionado could estimate the time between the calls he receives. This would allow him to plan his next coffee break and decide how much time he can spend conversing with a colleague.

More seriously, these statistics are important when it comes to scheduling other regular, low-priority tasks in microprocessor- or microcontroller-based systems. They can be used to plan resources as well. A hospital might decided whether or not a doctor needs an associate. A internet service provider might decide whether or not his infrastucture needs an upgrade.

2.3 Little's Law

Now that we have come 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 $[0, T]$.

$$\begin{aligned}
&\mathbb{P}(\text{n customers in the queue after one departure}) && \text{(2.31a)}\\
&= \mathbb{P}(\text{n customers arrive during the time spent in the queue by the depart customer.}) && \text{(2.31b)}\\
&= \int_0^\infty \mathbb{P}(\text{n arrivals in during the interval } [0,T] \mid T = t) f_T(t) dt && \text{(2.31c)}\\
&= \int_0^\infty \mathbb{P}(\text{n arrivals in during the interval } [0,t]) f_T(t) dt && \text{(2.31d)}\\
&= \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_T(t) dt && \text{(2.31e)}
\end{aligned}$$

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

$$\bar{N} = \mathbb{E}[N] \tag{2.31f}$$
$$= \sum_{n=0}^{\infty} nP_n(t) \tag{2.31g}$$
$$= \sum_{n=0}^{\infty} n \int_0^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda} f_T(t)dt \tag{2.31h}$$
$$= \int_0^{\infty} \sum_{n=1}^{\infty} n \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_T(t)dt \tag{2.31i}$$
$$= \int_0^{\infty} \sum_{n=1}^{\infty} \lambda t \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t} f_T(t)dt \tag{2.31j}$$
$$= \int_0^{\infty} \lambda t e^{-\lambda t} f_T(t) \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} dt \tag{2.31k}$$
$$= \int_0^{\infty} \lambda t e^{-\lambda t} e^{\lambda t} f_T(t)dt \qquad \ldots \text{ Taylor series expansion of } e^x. \tag{2.31l}$$
$$= \lambda \int_0^{\infty} t f_T(t)dt \tag{2.31m}$$
$$= \lambda \mathbb{E}[T] \tag{2.31n}$$
$$= \lambda \bar{T}. \tag{2.31o}$$

😊

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.

With this result, we will conclude our study on $M/M/1$ queue. We end with a summary

2.4 Summary

Random Variable	Distribution
Number of arrivals in interval (0, t]	Poisson(λt)
Inter-arrival times	Exponential(λ)
Number of departures in the interval (0, t]	Poisson(μt)
Inter-departure times	Exponential(μ)

At equilibrium, $\lambda < \mu$,

State Probabilities	$p_0 = 1 - \rho$ $p_n = \rho^n p_0$
Average number of customers in the queue	$\rho/(1-\rho)$
Variance of number of customers in the queue	$\rho/(1-\rho)^2$

Chapter 3

Other Queues

3.1 Limitations of $M/M/1$ model

While the previous queueing model was simple enough that we could derive the nice results we obtained, it isn't quite realistic. Recall the assumptions we made before embarking on the derivations. One of them was that the coin tosses, which 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 received 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. This means that the calls are not completely uncorrelated. However, this assumption is acceptable for a first-order analysis.

The next assumption we made was that the queue could, in theory, grow without bound. This is quite unrealistic, simply due to the limited availability of resources and (the doctor's waiting room can't possibly be big enough to have place for infinite patients), not to mention, the patience one would need to endure an infinite human queue. Hence, for the next part of our analysis, we consider a model with finite buffer length.

After this analysis, we ask: why limit our system to a single-server queue? After all, efficiency lies in parallelism. Nowadays, almost all the microprocessors are equipped with multiple cores, essentially multiple compute units. 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 with, 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 turned away or dropped. To maintain the independence of arrivals, we further assume that the dropped customers do not return.

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

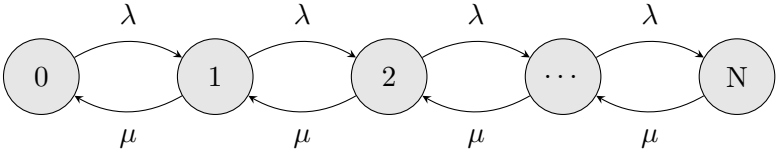


Figure 3.1: State transition diagram for $M/M/1/N$ queue.

Recollecting the local balance equations, we have under equilibrium,

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

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

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

Substituting in 3.1a,

$$p_n = (\frac{1-\rho}{1-\rho^{N+1}})\rho^n \qquad \dots\ 0 \leqslant n \leqslant N. \tag{3.1g}$$

Definition 3.2.1: Blocking Probability

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

Note:-

Using the blocking probability, we can find the mean number of customers turned away per unit time as λP_N.

Question 7: What is p_n when $\lambda/\mu = 1$?

Solution: When $\lambda/\mu = 1$, p_n takes the form $0/0$. We apply L'Hospital's rule:

$$\lim_{\rho\to 1} p_n = \lim_{\rho\to 1} \frac{1-\rho}{1-\rho^{N+1}}\rho^n \tag{3.2}$$
$$= \lim_{\rho\to 1} \frac{1-\rho}{1-\rho^{N+1}} \tag{3.3}$$
$$= \lim_{\rho\to 1} \frac{-1}{-(N+1)\rho^N} \tag{3.4}$$
$$= \frac{1}{N+1}. \tag{3.5}$$

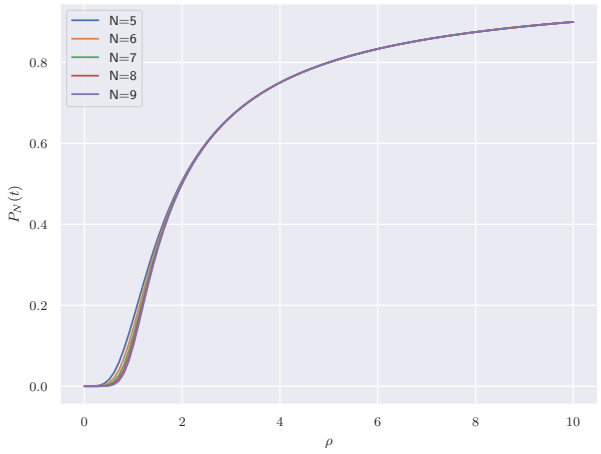
Note:-

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 and the queue size will not grow without bounds.

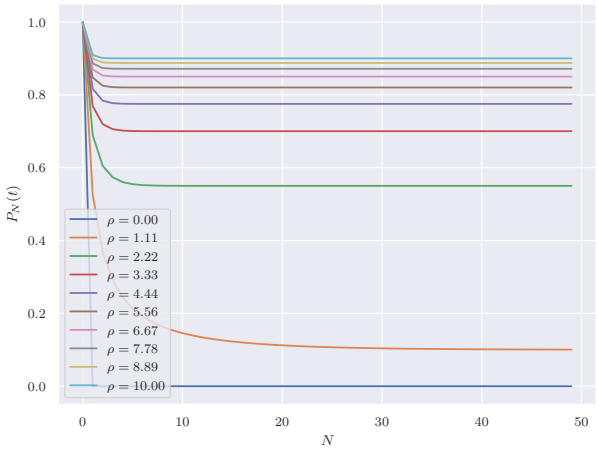
We plot the blocking probabilities in Fig.3.2a and 3.2b. It concurs with our intuition that the blocking probability increases with increasing utilization (ρ) and decreases with increasing buffer length, N.

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

Now, we move on to the case of multiple server models. However, to begin with, it is easier to consider the infinite server case, as we will soon see. The setup we now have is such that every



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

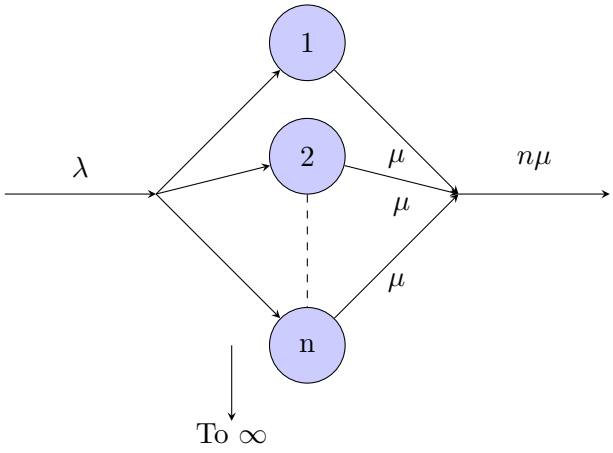


(b) Blocking probability as a function of N (ρ fixed).

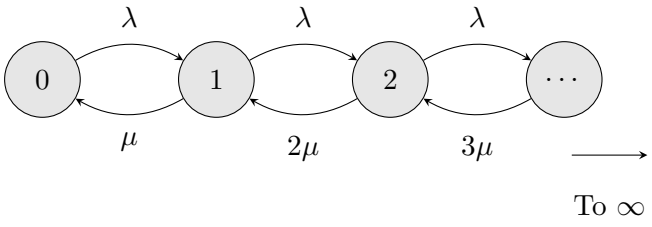
arriving customer is assigned a personal server with rate μ. Since we have an infinite number of servers, we don't mind if the queue size increases without bound.

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) State transition diagram of an $M/M/\infty$ queue.

Using this setup, and referring to the state transition diagram shown in Fig. 3.3b, we get the following local balance equations:

$$\lambda p_0 = \mu p_1 \implies p_1 = \frac{\lambda}{\mu} p_0 \tag{3.6a}$$

$$\lambda p_1 = 2\mu p_2 \implies p_2 = \frac{\lambda}{2\mu} p_1 \tag{3.6b}$$

$$\lambda p_2 = 3\mu p_3 \implies p_3 = \frac{\lambda}{3\mu} p_2 \tag{3.6c}$$

$$\tag{3.6d}$$

Generalizing, we get:

$$p_n = \frac{\lambda}{n\mu} p_{n-1} \tag{3.6e}$$

$$= \frac{\lambda}{n\mu} \cdot \frac{\lambda}{(n-1)\mu} p_{n-2} \tag{3.6f}$$

$$\vdots \tag{3.6g}$$

$$= \frac{\lambda^n}{n!\mu^n} p_0 \tag{3.6h}$$

Using the axioms of probability,

$$1 = p_0 + \frac{\lambda}{\mu} p_0 + \frac{\lambda^2}{2!\mu^2} p_0 + \frac{\lambda^3}{3!\mu^3} p_0 + \dots \tag{3.6i}$$

$$\implies 1 = p_0 \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2!\mu^2} + \frac{\lambda^3}{3!\mu^3} + \dots \right) \tag{3.6j}$$

$$\implies 1 = p_0 \left(1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n \right) \tag{3.6k}$$

$$\implies 1 = p_0 \times e^{\lambda/\mu} \qquad \dots \text{ Taylor series expansion of } e^x. \tag{3.6l}$$

In summary,

$$p_0 = e^{-\lambda/\mu} \tag{3.6m}$$

$$p_n = \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n e^{-\lambda/\mu} \tag{3.6n}$$

Defining $\rho = \frac{\lambda}{\mu}$, we get:

$$p_n = \frac{\rho^n}{n!} e^{-\rho} \tag{3.6o}$$

$$\therefore N \sim \text{Poisson}(\rho) \tag{3.6p}$$

Note:-

An $M/M/\infty$ queue can also be interpreted in a slightly different manner. From the state transition diagram, Fig.3.3b, that the aggregate output rate of the system is proportional to the state of the system. Thus, one can say that this a queue with a single server but with a *load-dependent service rate*. See Fig.3.4a

3.4 $M/M/m$ - m Parallel servers with a queue

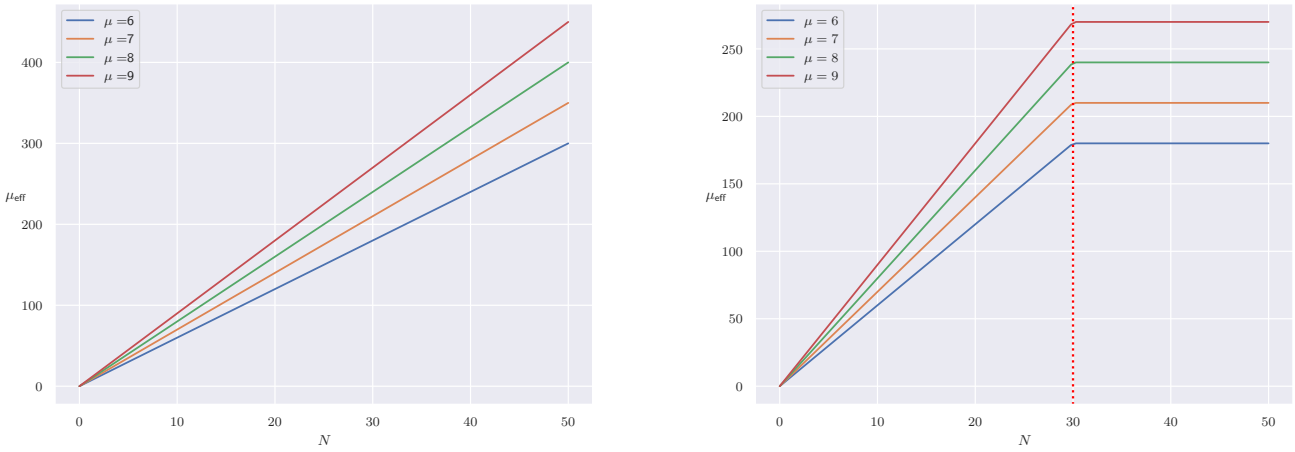
Now, we make a realistic assumption and consider a model with only m parallel servers. This results in a flow diagram, shown in Fig.3.5, and a state transition diagram, shown in Fig.3.6.

The transition rates in this case are given by:

$$\lambda(n) = \lambda, \qquad n = 0, 1, 2, \dots, \tag{3.7}$$

$$\mu(n) = n\mu, \qquad n = 1, 2, \dots, m-1, \tag{3.8}$$

$$\mu(n) = m\mu, \qquad n = m, m+1, m+2, \dots. \tag{3.9}$$



(a) Effective (aggregate) output rate in a $M/M/\infty$ system. (b) Effective (aggregate) output rate in a $M/M/30$ system.

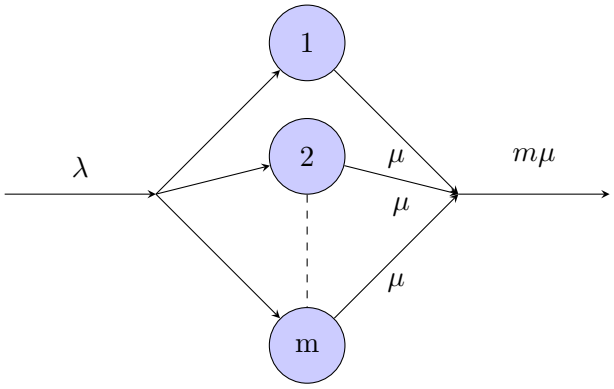


Figure 3.5: Flow diagram of an $M/M/m$ queue.

Using these rates, we get the following local balance equations,

$$p_1 = \left(\frac{\lambda}{\mu}\right) p_0, \tag{3.10a}$$

$$p_2 = \left(\frac{\lambda}{2\mu}\right) p_1 = \left(\frac{\lambda^2}{2\mu^2}\right) p_0, \tag{3.10b}$$

$$p_3 = \left(\frac{\lambda}{3\mu}\right) p_2 = \left(\frac{\lambda^3}{3!\mu^3}\right) p_0, \tag{3.10c}$$

$$\vdots \tag{3.10d}$$

$$p_m = \left(\frac{\lambda}{m\mu}\right) p_{m-1} = \left(\frac{\lambda^m}{m!\mu^m}\right) p_0, \tag{3.10e}$$

$$p_{m+1} = \left(\frac{\lambda}{m\mu}\right) p_m = \left(\frac{\lambda^{m+1}}{mm!\mu^{m+1}}\right) p_0, \tag{3.10f}$$

$$p_{m+2} = \left(\frac{\lambda}{m\mu}\right) p_{m+1} = \left(\frac{\lambda^{m+2}}{m^2 m!\mu^{m+2}}\right) p_0, \tag{3.10g}$$

$$\vdots \tag{3.10h}$$

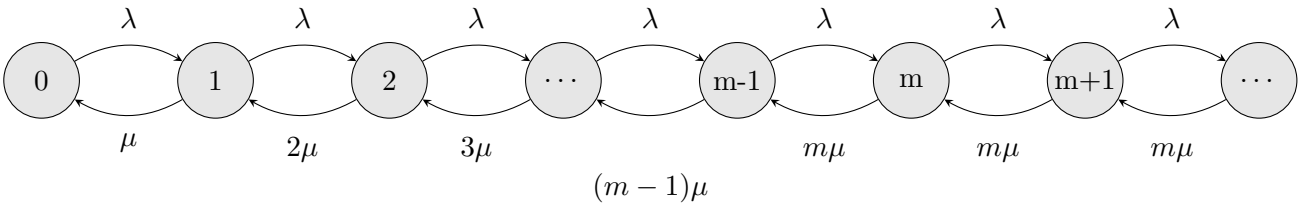


Figure 3.6: State transition diagram of an $M/M/m$ queue.

Generalising, we can write in short as,

$$p_n = \frac{\lambda^n}{n!\mu^n} p_0 \qquad \text{if } 1 \leqslant n < m \tag{3.10i}$$

$$= \frac{\lambda^n}{m^{n-m} m! \mu^n} p_0 \qquad \text{if } n \geqslant m. \tag{3.10j}$$

Using the axioms of probability,

$$\sum_{n=0}^{\infty} p_n = 1, \tag{3.10k}$$

$$\sum_{n=0}^{m-1} p_n + \sum_{n=m}^{\infty} p_n = 1, \tag{3.10l}$$

$$p_0 \left(1 + \sum_{n=1}^{m-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=m}^{\infty} \frac{1}{m^{n-m}} \frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^n \right) = 1, \tag{3.10m}$$

$$\therefore p_0 = \left[1 + \sum_{n=1}^{m-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=m}^{\infty} \frac{1}{m^{n-m}} \frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}. \tag{3.10n}$$

Thus, we can find other state probabilities in equilibrium. One important probability that we can consider is the probability of the event that there will be a queue. There will be a queue whenever, all of the m servers are busy and there are arrivals.

Question 8: What is the *queueing* probability in an $M/M/m$ queue?

Solution:

$$\mathbb{P}(\text{queueing}) = \mathbb{P}(n \geqslant m) \tag{3.11}$$

$$= \sum_{n=m}^{\infty} p_n \tag{3.12}$$

$$= \sum_{n=m}^{\infty} \frac{1}{m^{n-m}} \frac{1}{m!} (\frac{\lambda}{\mu})^n p_0 \tag{3.13}$$

$$= \frac{1}{m!} (\frac{\lambda}{\mu})^m \sum_{n=0}^{\infty} (\frac{\lambda}{m\mu})^n p_0 \tag{3.14}$$

$$= \frac{1}{m!} (\frac{\lambda}{\mu})^m (\frac{1}{1 - \frac{\lambda}{m\mu}}) p_0 \tag{3.15}$$

Letting $\rho \coloneqq \frac{\lambda}{m\mu}$,

$$\mathbb{P}(\text{queueing}) = \frac{1}{m!}(\frac{\lambda}{\mu})^m \frac{1}{1-\rho} p_0 \tag{3.16}$$

Note:-

Like in previous, model we can observe that the aggregate output rate of the system is proportional to the state of the system. However, now there is a limit on how much this rate can increase. See Fig.3.4b.

This kind of queue can provide a lot of insights in telecommunications field as this model fits many use cases in the field very well. The *Queueing probability* found in the previous question represents the probability that all the servers are busy. The closed form expression of this probability is called *Erlang-C formula*. Note that by substituting $\rho = 1$, we go back to the $M/M/1$ queue. Our intuition tells us that as m, number of servers increases, the queueing probability should go down. To verify this we have plotted this probability as a function of m. See Fig.

This metric can be used as a measure of performance of a communication system and to take decisions in the field.

Chapter 4

Summary and other models

Finally, we have covered all the basic concepts and models that one can cover in an introductory lecture. As always, you may have a question in your mind popping up as we reach the end of this discussion: What next? Well, as we have been mentioning periodically, there are several applications of queueing theory in industrial operations, computer science, retail markets, business operations, and more. You are now ready to apply the concepts learned here and demonstrate your creativity.

We will begin by first providing a summary of all the models discussed so far. Then, we will briefly touch upon other, more realistic models and finally conclude with a practical example where queueing theory has been applied.

4.1 Summary

We understand that reading through these notes might be exhausting. Therefore, for your quick reference, we have tabulated the results below, focusing primarily on the equilibrium state probability distributions.

Queueing Model	State Probability Distribution
$M/M/1$	$p_0 = 1-\rho$ $p_n = \rho^n p_0$
$M/M/1/N$	$p_n = \left(\frac{1-\rho}{1-\rho^{N+1}}\right)\rho^n \quad \dots \ 0 \leqslant n \leqslant N.$
$M/M/\infty$	$p_n = \frac{\rho^n}{n!}e^{-\rho}$ $\therefore N \sim \text{Poisson}(\rho)$
$M/M/m$	$p_0 = \left[1 + \sum_{n=1}^{m-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=m}^{\infty} \frac{1}{m^{n-m}} \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$ $p_n = \frac{\lambda^n}{n!\mu^n}p_0 \quad \text{if } 1 \leqslant n < m$ $p_n = \frac{\lambda^n}{m^{n-m}m!\mu^n}p_0 \quad \text{if } n \geqslant m.$

4.2 Other models

The models we have encountered so far all assumed Markovian statistics for both the arrival and departure processes. However, we need not stick to these processes. Other distributions can be used to model scenarios more appropriately.

For instance, at a bus or train ticket counter, the time required to satisfy the demands of each customer will be more or less the same. Thus, the service times can be modeled as deterministic.

[5, 6] also provide results for the $M/G/1$ queue. Although the method for deriving these results differs from what we have done so far, it is simple enough to follow, with some *Transform Theory* involved. Those interested may refer to these sources.

More general models, such as $G/M/1$ and $G/G/1$, have analytical results derived in [2, 3].

4.3 Critical Application of Queueing Theory - A case study[1]

The U.S. Military's Air Force fleet includes 20 B-2 bombers, which require frequent maintenance. A key maintenance procedure, *Low Observable* (LO), restores the special coating on the aircraft. However, maintenance scheduling and manpower distribution introduce unpredictability into the process.

The B-2's *Flying Hour Program* has two main goals: maintaining operational readiness, including weapons preparedness, and ensuring wartime posture. Initially, each aircraft underwent a 200-flying hour post-flight inspection followed by *heavy LO* restoration. This deterministic plan, however, resulted in poor Aircraft Availability (AA), with some B-2s completing the process in weeks, while others took up to six months.

Data analysis revealed that, on average, 4.75 B-2s were grounded for LO maintenance at any time, reducing AA to 75%. Additionally, about 3 aircraft were grounded for other maintenance, further reducing AA to 60%, leaving only 12 B-2s available.

To improve AA, maintenance procedures were updated. An acceptable AA range of 80-85% (i.e., 17 B-2s) was set. Data analysis revealed that every seven days one B2 required maintenance.

We can now model this scenario using queueing theory. At any time, there is (or rather must be) a queue of, on average, at most 3 planes. Thus, we have:

$$\bar{N} = 3.$$

Additionally, one plane enters service every week, so the arrival rate is:

$$\lambda = 1 \text{ per week} \quad \text{or} \quad \lambda = \frac{1}{7} \text{ per day.}$$

By Little's Law, the average service time $\bar{T}$ is given by:

$$\begin{aligned} \bar{T} &= \frac{\bar{N}}{\lambda} && (4.1)\\ &= \frac{3}{\frac{1}{7}} && (4.2)\\ &= 21 \text{ days.} && (4.3) \end{aligned}$$

Thus, to maintain this system, the lead time for heavy LO maintenance must be 21 days. Using this estimate, the maintenance group was able to devise a schedule and routine to ensure the required AA. In this way, queueing theory played a key role in resource planning in a critical field.

Chapter 5

Appendix

Only the bare minimum concepts have been covered here. I strongly advise that you go through the relevant material in [4]

5.1 Random Processes

Definition 5.1.1: 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.

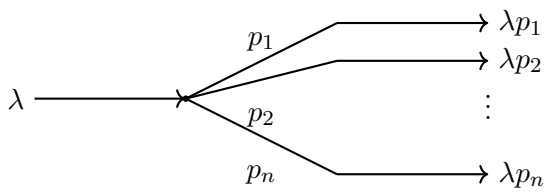
Definition 5.1.2: Splitting (Thinning) of Possion Process

Consider a parent Poisson random process with rate λ. Then, if the arrivals in this process are split into n different children with probabilities $p_1, p_2, \ldots, p_n$, 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.5.1a.

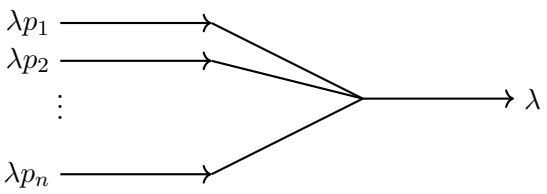
Definition 5.1.3: Merging of Poisson Process

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.5.1b.

See Fig.5.1a and Fig.5.1b to understand how a process is split into multiple ones or how multiple processes are combined into a single process.



(a) Splitting (Thinning of Poisson Process)



(b) Merging of Poisson Process

Bibliography

[1] Robert Borries. Usaf uses continuous process improvement on the b-2 bomber: Part 1.

[2] David G. Kendall. Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *The Annals of Mathematical Statistics*, 24(3):338 – 354, 1953.

[3] L. Kleinrock. *Queueing Systems: Theory*. A Wiley-Interscience publication. Wiley, 1974.

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

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

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