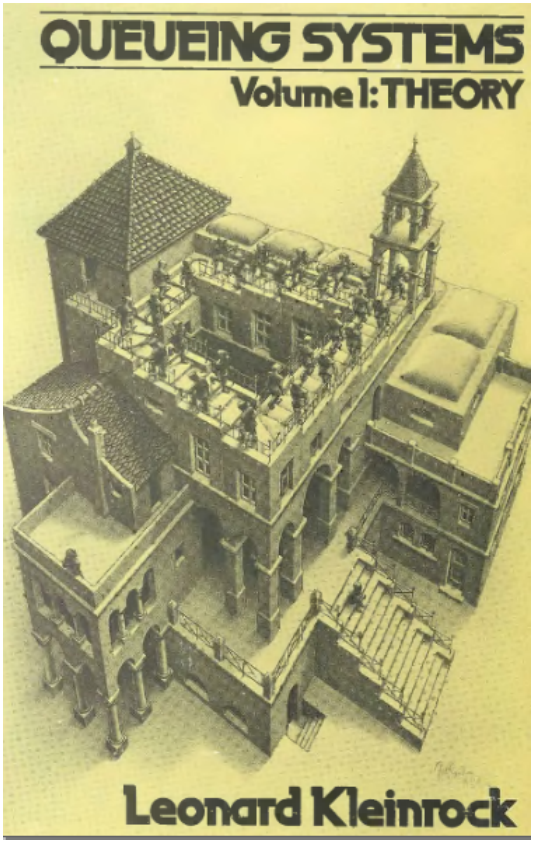
Queueing Systems - Volume I : Theory

RK

April 14, 2015

Abstract

The purpose of this document is to summarize the main points of the book written by Leonard Kleinrock, titled, 'Queueing Systems".



Contents

1 Queueing Systems

Queueing systems represent an example of much broader class of interesting dynamic systems, which can be referred to as *systems of flow*. A flow system is one in which some commodity flows, moves, or is transferred through one or more finite-capacity channels in order to go from one point to another. When one analyses systems of flow, they naturally break into two classes

1. Steady flow : In these kind of systems, there is some orderliness, i.e. quantity of flow is exactly known. There are a network of channels, each with a channel capacity. The math principles that are used in the analysis of such systems are graph theory, combinatorial mathematics, optimization theory, mathematical programming and heuristic programming
2. Unsteady flow : These are random or stochastic flow systems. The times at which demands for service arrive are uncertain, the size of demands themselves are unpredictable. Even a simplistic case of single channel means a non-trivial analysis. Analyzing a single channel in a steady flow is no brainer. In the case of stochastic flow, there are many questions that need specific mathematical tools to answer. In most of the cases, there are no closed form solutions for the questions. The tools necessary to solve these problems are described in queueing theory. Any computer network involves complex queueing problems and in order to understand and solve such problems, a thorough knowledge of basic tools is a must.

In order to completely specify a queueing system, one must identify stochastic processes that describe the arriving stream as well as the structure and discipline of the service facility. What are the elements needed to study a stochastic flow ?

- arrival process
- service time distribution or at least a few moments of its pdf
- storage capacity of the system to accommodate waiting customers
- number of service stations available
- queue discipline
- presence or absence of customer behavior such as jockeying, balking, bribing, cheating and defections

Once the above elements are identified, various math tools are used to answer questions such as

- average number of customers in the system
- distribution of number of customers in the system
- average waiting time of a customer in the system
- distribution of waiting time
- length of busy period
- length of an idle period
- current work backlog expressed in units of time

2 Some Important Random Processes

Structure for basic queueing system

The author starts off this chapter by defining symbolic and graphic notation used in the book. One can intuitively understand Little's law as follows:

> The average number that an arriving customer sees is same as the average number that a departing customer sees. The latter equals the average time spent in the system times the arrival rate of the customers

$$\overline{N} = \lambda T$$

The beauty of this intuition is that you can apply the law to any specific component of the queueing system. If we consider the queue, then the average number of customers in the queue is related to the average waiting time

$$\overline{N}_q = \lambda W$$

Or you can restrict to only the server, in which the average number of customers in the service facility is related to the average time spent at the server.

$$\overline{N}_s = \lambda \overline{x}$$

This existed as a "folk theorem" until J.D.C Little proved it in 1961. The best way to remember this law is

> The average arrival rate of customers to a *queueing system* times the average time spent by customers in that *system* is equal to the average number of customers in the *system* regardless of how we define that *system*

The utilization factor of a system denoted by ρ is ratio of the rate at which work enters the system to the maximum rate at which the system can perform this work. For a single server case, it is given by

$$\rho = \frac{\lambda}{\mu}$$

An intuitive way to determine the relationship between p_0, the probability that the system is idle and ρ is : If you consider a very long time interval τ, we can equate the number of arrivals to the number of departures and obtain the following equation

$$\lambda\tau = \frac{\tau - \tau p_0}{\overline{x}}$$

Hence the relationship is

$$\rho = 1 - p_0$$

Definition and Classification of Stochastic processes

A stochastic process is a family of random variable that are indexed by the time parameter t. The classification of random processes depend on three quantities

- state space - continuous or discrete
- index parameter t - continuous or discrete
- statistical dependencies among the random variables $X(t)$for different values of the the index parameter

By considering various combinations, there can be 4 types of stochastic processes

- discrete state - discrete parameter (the game of snakes and ladders)
- discrete state - continuous parameter (the number of customers in a queueing system)
- continuous state - continuous parameter (Brownian motion)
- continuous state - discrete parameter (a guy picks a random number at every discrete unit of time)

The truly distinguishing feature of a stochastic process is the relationship of the random variables $X(t)$ or X_n to other members of the same family. Ideally a joint distribution of all the variables needs to be specified to define a stochastic process. However many interesting processes permit a simpler description. The chapter provides a broad classification of stochastic processes :

- Stationary processes : Distribution is invariant to shifts in time. A subset of stationary processes are called wide sense stationary processes if the first two moments are time invariant
- Independent process : Joint distribution splits. In fact there is no structure whatsoever amongst the random variables
- Markov processes : In 1907 A.A.Markov published a paper in which he identified and investigated the properties of what are now known as Markov processes. In fact, what he created was a simple and highly useful form of dependency among the random variables forming a stochastic process. A set of random variables forms a Markov chain if the probability of next value depends only upon the current value and not upon the previous values. Thus backward dependency is limited to one. For a discrete time Markov chain, the process may remain in any given state for a time that must be geometrically distributed. For a continuous Markov chain, the process may remain in ay given state for a time that is exponentially distributed.
- Birth-Death process : This is a very special class of Markov processes where the state transitions are only to the neighboring states.
- Semi-Markov process : If we want to permit an arbitrary distribution for the time spent by process in a state, we obtain a semi-Markov process. If you consider only the instants of state transition, you obtain an imbedded Markov chain.
- Random walk : A special case of semi-Markov process. The following describes the state of the process. The time between state transitions is some other random variable

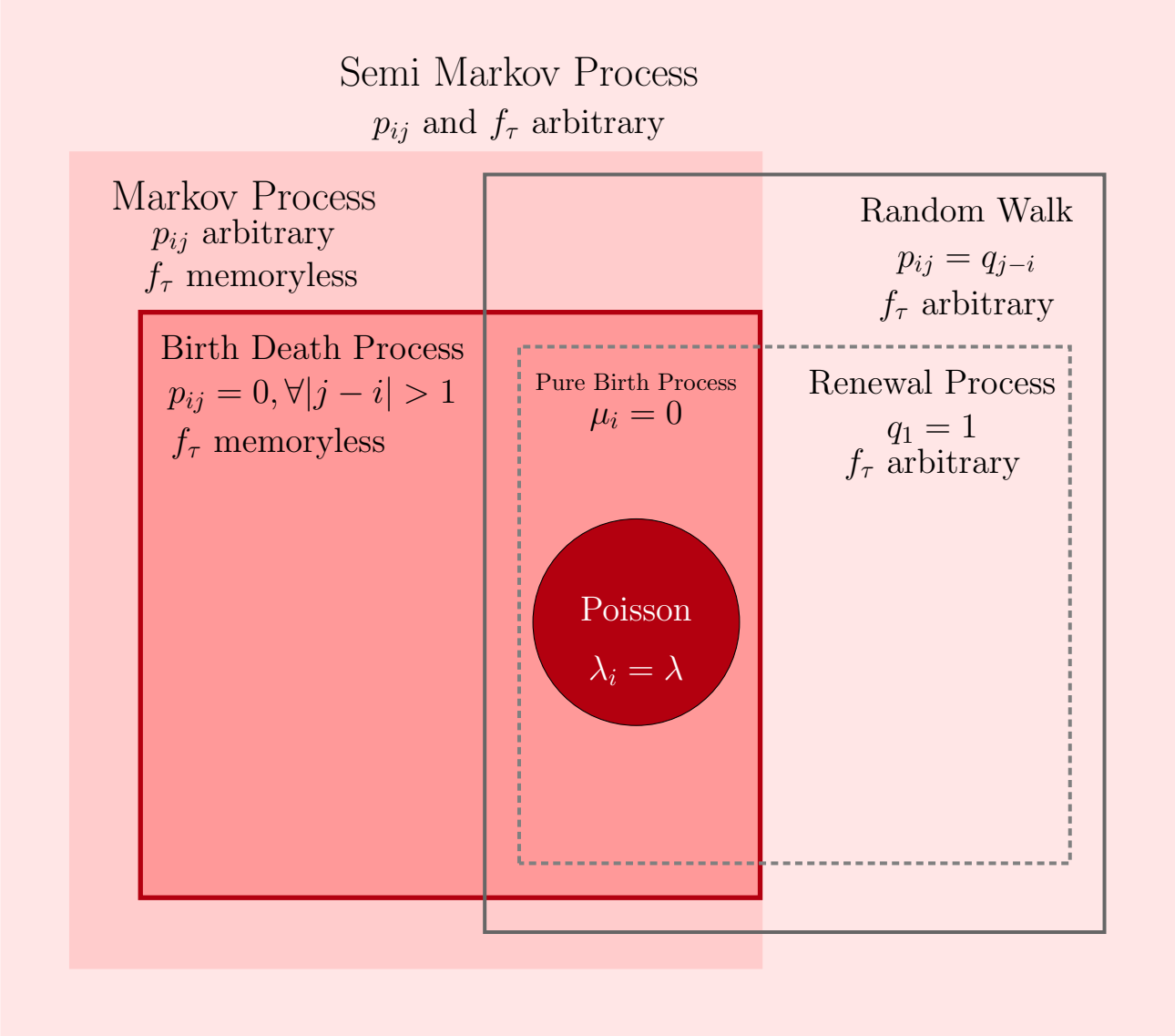
$$S_n = X_1 + X_2 + \ldots + X_n, \quad n = 1, 2, \ldots$$

- Renewal process : The following equation describes the time of n^{th} renewal.

$$S_n = X_1 + X_2 + \ldots + X_n, \quad n = 1, 2, \ldots$$

  This process can be thought of as one that starts from 0 that has transition probability $q_1 = 1, q_i = 0, \forall i \neq 1$.

The author provides a nice visual to categorize various stochastic processes. Here is a modified version of the same :



The process that lies at the intersection of all the six processes is the *Poisson processes*. It is in the sweet spot of all stochastic processes .

Discrete-time Markov chains

I have understood Discrete-time Markov chains by studying Snakes and Ladders game. Much of my understanding about Markov chains comes form writing code and answering questions about Snakes and Ladders. One might dismiss that this is a back-door approach and claim that the correct approach is to learn the math. May be. But this game based approach worked for me and slowly over time, I have thoroughly understood the math behind the Markov chains. The author does a great job of explaining the relevant concepts such as

- Traveling Hippie example to give an idea of what a discrete markov chain looks like
- Transition probabilities
- Homogeneous Markov chain
- Irreducible Markov chain
- Aperiodic Markov chain
- Ergodic Markov chain
- Null recurrent chain
- Positive recurrent chain
- Limiting probabilities
- Baricentric coordinates
- Use of generating function approach to solve limiting probabilities
- Chapman-Kolmogorov equation

Continuous-time Markov chains

I did not use a game analogy for understanding CTMC. However it is not difficult to imagine such a game. Let's say there are 16 bulbs in 16 squares. Each bulb fails with a Poisson rate of λ. Every time a bulb fails, you need to figure out the quickest route to the room and replace the bulb and you have to do it certain number of steps. May be if one spends some time on this, one can think of a more meaningful game. In learning through CTMC for the first time, one thing I remember about my slog was that I plunged in to the math with a leap of faith, assuming that DTMC slog will be good enough to understand CTMC. In the hindsight, I think I feel I should have begun with a M/M/1 system in mind so that I could have appreciated the math. One of my friends takes this approach. Before venturing out in to the math and details, he always has a game/scenario in mind. Subsequently he goes about understanding the math. May be that is a right approach.

Anyway, coming back to CTMC mentioned in the book. The hippie example in the book is also a good one. If the hippie decides to leave a city at any time of day or night and lands up in some other city, we have a CTMC. For a CTMC, the time spent by the variable in each state is exponentially distributed. Just like in the discrete markov chain case, we have a transition rate matrix $Q(t)$. Forward Chapman-Kolmogorov equation and Backward Chapman-Kolmogorov equation are derived in the section. For time dependent state probabilities, we have a differential equation

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\boldsymbol{H}(0,t)$$

where $H(s,t) \triangleq [p_{ij}(s,t)]$. The relationship between H matrix and Q matrix is

$$H(s,t) = \exp\left[\int_s^t Q(u)du\right]$$

Thus the main equation connecting the state probabilities and the Q matrix is

$$\frac{d\boldsymbol{\pi}(t)}{dt} = \boldsymbol{\pi}(t)\boldsymbol{Q}(t)$$

The forward and backward equations in matrix form are

$$\frac{d\boldsymbol{H}(t)}{dt} = \boldsymbol{H}(t)\boldsymbol{Q}$$

$$\frac{d\boldsymbol{H}(t)}{dt} = \boldsymbol{Q}\boldsymbol{H}(t)$$

The limiting distribution can be obtained by solving

$$\boldsymbol{\pi}\boldsymbol{Q} = 0$$

To simulate CTMC, one needs to come up with a jump matrix.This matrix is obtained by analyzing the imbedded Markov chain. For every Q matrix for a simple system, you can write the Jump matrix and simulate away to glory.

Birth-Death Processes

This is a special case of CTMC where the transition rate matrix is a banded matrix. By writing a difference equation between state probabilities and then converting in to differential-difference equations, one can solve for the transient probabilities in a birth death chain. In any book you turn to on Queueing theory, in all likelihood, you will see a state transition rate diagram. Each state is surrounded by an oval and states are connected by directed arrows displaying the rates. Since the diagonal of the Q matrix does not contain any additional information, self-loop is not indicated in the diagram. Also the rate at which the process returns to the state that it currently occupies is infinite. The labels on the links in the state transition rate diagram refers to birth and death rates and not to probabilities. If one wishes to convert these labels to probabilities, one must multiply each by the quantity dt to obtain the probabilities of such a transition occurring in the next interval of time whose duration is dt. In that case, it is also necessary to put self-loops on each state indicating the probability that in th next interval of time dt, the systems remains in the given state. One can also write the differential-difference equation by equating the rate of flow in minus rate of flow out to the effective probability flow rate.

$$\frac{dP_k(t)}{dt} = \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) - (\lambda_k + \mu_k)P_k(t)$$

One can use a generating function approach to solve the above equation to get transient probabilities. The final solution for even the simplistic case of M/M/1 system contains a complicated expression involving Bessel functions. At this point, the author makes a case for studying equilibrium solutions rather than transient solutions of any Markovian system.

3 The Queue M/M/1

This and the next chapter in the book deal with pure Markovian systems. The one thing that should strike your mind whenever you hear the work "Markovian" is that the state description is convenient and manageable. The key aspect of this chapter is that most of the equilibrium solutions to the system flow from one single equation.

Markov processes play a fundamental role in the study of queueing systems. A special form of Markov processes known as *birth-death* process are relevant to many elementary queueing systems. These processes have a convenient property that the time between births and the time between deaths are each exponentially distributed.

General Equilibrium Solution

Obtaining a transient solution for a queueing system is a good math exercise but it is not clear how useful the set of transient functions would aid in understanding the behaviour of the systems. Hence it is natural to ask whether the probabilities $P_k(t)$ eventually settle down as t gets large and displays no more "transient" behavior. For convenience, we denote

$$p_k \triangleq \lim_{t\to\infty} P_k(t)$$

It is important to understand that whereas $p_k(t)$ is no longer a function of t, we are not claiming that the process does not move from state to state in this limiting case; certainly, the number of members in the population will change with time, but the long-run probability of finding the system with k members will be properly described by p_k. The forward Kolmogorov equations are

$$\begin{aligned}\frac{dP_k(t)}{dt} &= -(\lambda_k+\mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t)\\ \frac{dP_0(t)}{dt} &= -(\lambda_0)P_0(t) + \mu_1 P_1(t)\end{aligned}$$

In the steady state, we have

$$\begin{aligned}0 &= -(\lambda_k+\mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t)\\ 0 &= -\lambda_0 P_0(t) + \mu_1 P_1(t)\end{aligned}$$

From flow conversation across a boundary, one can easily see that

$$p_k = \frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu_1\mu_2\ldots\mu_k}p_0$$

with

$$p_0 = \frac{1}{1+\sum_{k=1}^{\infty}\Pi_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}}$$

For the existence of the steady-state probabilities, the system should occasionally empty. Hence the various

possibilities are captured by imposing conditions on two sums

$$S_1 \triangleq \sum_{k=0}^{\infty} \Pi_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}$$
$$S_2 \triangleq \sum_{k=0}^{\infty} \left(\frac{1}{\lambda_k \Pi_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \right)$$

- Birth death process is ergodic $\Rightarrow S_1 < \infty$ and $S_2 = \infty$
- Birth death process is null recurrent $\Rightarrow S_1 = \infty$ and $S_2 = \infty$
- Birth death process is transient $\Rightarrow S_1 = \infty$ and $S_2 < \infty$

The condition of ergodicity meets whenever $\lambda_k/\mu_k < 1$. The birth-death process of permitting only nearest-neighbour transitions might appear too restrictive. But they can be also be applied to more general markovian systems.

Various Markovian systems

- M/M/1 - Poisson arrivals/Poisson departures/1 server
  - $\lambda_k = \lambda$
  - $\mu_k = \mu$
  - p_k : steady state probability of the system being in k state is

$$p_k = p_0 \left(\frac{\lambda}{\mu} \right)^k$$

  - p_k : probability of server being idle

$$p_0 = 1 - \rho$$

  - $\overline{N}$: average number of customers in the system

$$\overline{N} = \frac{\rho}{1-\rho}$$

  - T : Time in the system

$$T = \frac{1/\mu}{1-\rho}$$

- Discourage Arrivals
  - $\lambda_k = \frac{\alpha}{k+1}$, $k = 0, 1, 2, \ldots$
  - $\mu_k = \mu$
  - p_k : steady state probability of the system being in k state is

$$p_k = \text{Poisson}(\alpha/\mu)$$

  - p_k : probability of server being idle

$$p_0 = e^{-\alpha/\mu}$$

  - $\overline{N}$: average number of customers in the system

$$\overline{N} = \frac{\alpha}{\mu}$$

  - T : Time in the system
$$T = \frac{\alpha}{\mu^2(1 - e^{-\alpha/\mu})}$$

- M/M/∞
  - $\lambda_k = \lambda$, $k = 0, 1, 2, \ldots$
  - $\mu_k = k\mu$, $k = 0, 1, 2, \ldots$
  - p_k : steady state probability of the system being in k state is
$$p_k = \text{Poisson}(\lambda/\mu)$$
  - p_k : probability of server being idle
$$p_0 = e^{-\lambda/\mu}$$
  - $\overline{N}$: average number of customers in the system
$$\overline{N} = \frac{\lambda}{\mu}$$
  - T : Time in the system
$$T = \frac{1}{\mu}$$
- M/M/m - Poisson arrivals/Poisson departures/m servers
  - $\lambda_k = \lambda$, $k = 0, 1, 2, \ldots$
  - $\mu_k = k\mu, \forall k = 0, 1, 2, \ldots, m$, $\mu_k = m\mu, \forall k > m$
- M/M/1/K - Poisson arrivals/Poisson departures/1 server / System size K
  - $\lambda_k = \lambda \forall k < K$, $\lambda_k = 0 \forall k \geq K$
  - $\mu_k = k\mu, \forall k = 0, 1, 2, \ldots, K$
- M/M/m/m - Poisson arrivals/Poisson departures/m server / System size m
  - $\lambda_k = \lambda \forall k < m$, $\lambda_k = 0 \forall k \geq m$
  - $\mu_k = k\mu, \forall k = 0, 1, 2, \ldots, m$
- M/M/1//M - Poisson arrivals/Poisson departures/1 server / No system constraint / customers M
  - $\lambda_k = \lambda(M - k) \forall k \leq M$, $\lambda_k = 0 \forall k > M$
  - $\mu_k = \mu$
- M/M/∞//M - Poisson arrivals/Poisson departures/∞ servers / No system constraint / customers M
  - $\lambda_k = \lambda(M - k) \forall k \leq M$, $\lambda_k = 0 \forall k > M$
  - $\mu_k = k\mu$
- M/M/m/K/M - Poisson arrivals/Poisson departures/m servers / System size K / customers M
  - $\lambda_k = \lambda(M - k) \forall k \leq M$, $\lambda_k = 0 \forall k > M$
  - $\mu_k = k\mu, \forall k = 0, 1, 2, \ldots, m$, $\mu_k = m\mu, \forall k > m$

For all these systems, the metrics can be derived from just two equations

$$p_k = \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots \mu_k} p_0$$

with

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \Pi_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}$$

4 Markovian Queues in Equilibrium

The Equilibrium Equations

If we venture outside of the usual *birth death* processes and allow for more general transitions, the system does become complex. However the *flow* method can still be used to obtain a set of simultaneous equations, whose solution gives limiting probabilities of the system in various states. There are two specific quantities that are generally of interest :

- p_k : The equilibrium probability that the system contains k customers
- r_k : The equilibrium probability that an arriving customer finds k customers in the system

Intuitively one might think that these probabilities are always equal. However a simple D/D/1 case shows that they are not equal. However there is a class of processes for which they are equal. If the arrival process is Poisson then

$$P_k(t) = R_k(t)$$

where $P_k(t)$ is the probability that the system is in state E_k at time t and $R_k(t)$ is the probability that a customer arriving at time t finds the system in state E_k

The Method of Stages - Erlangian Distribution E_r

For systems more general than *birth-death* systems, the Markovian structure of the system is lost. It was Erlang again who recognized the extreme simplicity of exponential distribution and its great power in solving Markovian queueing systems. What Erlang conceived was the notion of decomposing the service time distribution into a collection of structured exponential distributions. This method was conceived at a time when *imbedded Markov chain* was not developed.

The basic idea behind using Erlangian distribution is :

> If one has measured a service-time operation and had sufficient data to give acceptable estimates of its mean and variance only, then one could select a member of Erlangian two parameter family such that $1/\mu$ matched the mean and $1/(r\mu^2)$ matched the variance. One could then analyze the system as M/E_r/1

The Queue $M/E_r/1$

To work with this system, one must specify the number of customers in the system as well as specify the number of stages remaining in the service facility in the system. The two dim state vector can be crunched in to one dim by taking the state variable as the total number of service stages yet to be completed by all customers in the system at the time the state is described. If we consider the state at a time when the system has k customers and when the i the stage of service contains the customer in service, then the total number of stages contained in the system is

$$j \triangleq (k-1)r + (r-i+1) = rk - i + 1$$

Let p_k denote the equilibrium probability for the number of customers in the system, then

$$P_j \triangleq P[j \text{ stages in system}]$$

The relationship between customers and stages is

$$p_k = \sum_{j=(k-1)r+1}^{kr} P_j$$

The beauty of Erlang's approach is in drawing an appropriate state transition rate diagram, which helps in quickly writing down the equilibrium conditions

$$\begin{aligned} \lambda P_0 &= r\mu P_1 \\ (\lambda + r\mu)P_j &= \lambda P_{j-r} + r\mu P_{j+1} \end{aligned}$$

Using z transform method, one can obtain

$$P(z) = \frac{r\mu(1-\rho)(1-z)}{r\mu + \lambda z^{r+1} - (\lambda + r\mu)z}$$

Using partial fractions, one can obtain the equilibrium probabilities.

The Queue $E_r/M/1$

The analysis is similar to $M/E_r/1$. In this type of system, every arriving customer goes through r stage facility and once he completes these stages, he is said to have "arrived" to the queueing system. State description should contain the number of stages in the system. If there are k customers in the system and when the arriving customer is in the ith stage of arrival, then the total number of stages contained in the system is

$$j \triangleq rk + i - 1$$

Let p_k denote the equilibrium probability for the number of customers in the system, then

$$P_j \triangleq P[j \text{ stages in system}]$$

The relationship between customers and stages is

$$p_k = \sum_{j=(k+1)r-1}^{rk} P_j$$

By inspection method, one can write down the equations as

$$\begin{aligned} r\lambda P_0 &= r\mu P_r \\ r\lambda P_j &= r\lambda P_{j-1} + \mu P_{j+r}, \quad 1 \le j \le r-1 \\ (\mu + r\lambda)P_j &= r\lambda P_{j-1} + \mu P_{j+r}, \quad r \le j \end{aligned}$$

The solution to these equations is again via z transform method.

Bulk Arrival Systems

This is a system where a bulk arrival with r customers arrive at each "customer" arrival. This system is exactly similar to $M/E_r/1$ and hence the generating function of the P_j in the $M/E_r/1$ will be same as the generating function for the number of customers in the Bulk arrival system. If we want to develop a more generic system where we allow other than fixed arrivals, i.e.

$$g_i \triangleq P[\text{ bulk size is } i]$$

The equilibrium equations by inspection method are

$$(\lambda + \mu)p_k = \mu p_{k+1} + \sum_{i=0}^{k-1} p_i \lambda g_{k-i}$$

$$\lambda p_0 = \mu p_1$$

The solution to the z transform of the distribution of number of customers is in terms of z transform of the distribution of bulk size.

$$P(z) = \frac{\mu(1-\rho)(1-z)}{\mu(1-z) - \lambda z(1-G(z))}$$

Bulk Service Systems

If the server takes a fixed number of customers for service everytime, then this system is similar to $E_r/M/1$. The solution to the distribution of the number of customers in the bulk service system must be given by the distributions of stages in $E_r/M/1$ system as stages correspond to customers in the Bulk service system. The author generalizes this scenario and gives the generating function for the distribution of the number of customers in the system.

Series-Parallel Stages : Generalizations

The key step in the method of stages is the selection of parameters μ and r. This selection necessitates an acceptance of coefficient of variation less than that of exponential. If one wants to fit in a bigger variance in the system, then a parallel arrangement of the system does the job. This section explores *hyper exponential distribution* and its usage in coming up queueing systems that mirror the real world service time pdf moments.

5 The Queue M/G/1

If you start thinking about elementary queueing systems, and ask yourself,"What is the appealing aspect of the system, that makes it analytically tractable?", you will come to the conclusion that it is *simplicity of state description*. For Markovian systems all you need is the number of customers present in the system. Using this one piece of information, the future behavior of a pure Markovian system can be elegantly summarized. You do not need the history of the system at all.

This chapter deals with queueing systems that are driven by non-Markovian stochastic processes. What are the ways to handle these non-Markovian systems ?

1. *imbedded Markov chain*
2. *method of stages* that requires inter arrival time and service time pdf's to have Laplace transforms that are rational. The basic disadvantage of this system is that this is a procedure for carrying out the solution. It does not show the solution as an explicit expression
3. solving *Lindley's integral equation*
4. method of supplementary variables
5. using *random-walk* approach
6. *method of Green's function*

The M/G/1 system

This is a single-server system where inter arrivals are exponential but the service time is an arbitrary distribution denoted by $B(x)$. If we think about the state description for the system, we realize that we need to specify two things

1. N_t Number of customers present at time t
2. $X_0(t)$ The service time already received by the customer in service at time t

We do not have to keep track of the service time elapsed since the last arrival of a customer in to the system as the arrival process is memoryless. Hence $N(t)$ is a non-Markovian process but $[N(t), X_0(t)]$ is a Markov process and is an appropriate state vector for M/G/1 system.

In an elementary queueing system, the state variable is $N(t)$ and we have a discrete-state space, where the states themselves are finite or countable. In the case of M/G/1, the state space has a time element in it, making it a continuous-state Markov chain. Indeed this complicates the analysis. One can work with this continuous state using *method of supplementary variables*. This chapter proceeds by using *imbedded Markov chain* method.

The Paradox of Residual Life

The section begins by describing *inspection paradox*

> Assume that a hippie arrives at a roadside cafe at an arbitrary instant in time and begins hitchhiking. Assume further that automobiles arrive at this cafe according to a Poisson process at an average rate of λ cars per minute. How long must the hippie wait, on the average, until the next car comes along?

There are two ways to think about it. Since the average time between inter arrivals is $1/\lambda$, the time the hippie waits is $0.5/\lambda$. Another way to think about is : whatever instant the hippie lands up, the time for

the next arrival is $1/\lambda$. Also the time between the previous arrival of the car and the time at which the hippie lands is also $1/\lambda$. Hence the average time between the last car and next car to arrive is $1/\lambda$, twice the average inter arrival interval. Which is the correct way to think about ? As an aside, why is it called *inspection paradox* ? When you "inspect" a process, you are likely to find that things take longer than their (uninspected) average.The correct answer is $2/\lambda$ and the author introduces renewal theory to understand this paradox.

The time instants at which automobiles arrive at the cafe form a renewal process. Let's say the hippie lands at a time instant t. Recasting the problem in terms of *machine failures*, the time until the next automobile arrives is a random variable that measures the residual life of a component. The time since last arrival of the car is a random variable that measures the age of the component. The resolution of the paradox lies in discovering that the distribution of the random time interval when the hippie lands has a different distribution than the interarrival distribution of the automobiles. If we denote the selected lifetime X has a pdf $f_X(x)$ and PDF as $F_X(x)$,then it can be seen that

$$f_X(x) = \frac{xf(x)}{m_1}, \quad m_1 = E[\tau_k - \tau_{k-1}]$$

If we denote the residual lifetime as Y has a pdf $\hat{f}_(y)$ and PDF as $\hat{F}_Y(y)$,then it can be seen that

$$\hat{f}(t) = \frac{1 - F(y)}{m_1}$$

The above gives the density of residual life in terms of the common distribution of interval length and its mean. Using transforms, it is shown that

$$r_1 = \frac{m_1}{2} + \frac{\sigma^2}{2m_1}$$

Only if the variance of inter arrivals is 0, does the average waiting time become $0.5/\lambda$. In the case of Poisson it works out to be $1/\lambda$.

The author also introduces some basic terminology from *Renewal theory*.

- The age-dependent failure rate $r(x)$ is the instantaneous rate at which a component will gail given that is has already attained an age of x. This conditional density is given by
- The renewal function $H(x)$ is defined as

$$H(x) = E\{ \text{ Number of renewals in an interval of length } x\}$$

- The renewal density $h(x)$ is defined as

$$h(x) = \frac{dH(x)}{dx}$$

- Renewal theorem

$$\lim_{h\to\infty} h(x) = \frac{1}{m_1}$$

  In the limit one cannot identify when the renewal process began, and so the rate at which components are renewed is equal to the inverse of the average time between renewals $(m_1(t))$
- Renewal equation

$$h(x) = f(x) + \int_0^x h(x-t)f(t)dt$$

In transform form , the above equation is

$$H^*(x) = \frac{F^*(x)}{1 - F^*(x)}$$

The Imbedded Markov Chain

The basic idea behind this technique is the reduction of a dimension of the state space. By considering the number in the system at each of the customer's departure points, the state space becomes a semi-markovian system. At customer departure instants, an imbedded Markov chain represents the number of customers present in the system present in the system. The transitions take place only at the imbedded points and form a discrete-state space. The behavior of the chain at these imbedded points is completely describable Markov process. The solution at these imbedded Markov points happens also to provide the solution for all points in time. This can be proven by *method of supplementary variables*. This section ends with two conclusions:

1. For Poisson arrivals, it is always true that

$$P[N(t) = k] = P[\text{arrival at time } t \text{ finds } k \text{ in system}]$$

2. In in any non-Markovian system, $N(t)$ makes only discontinuous changes of size one, then if either one of the following limiting distributions exists, so does the other

$$r_k = \lim_{t\infty} P[\text{arrival at time } t \text{ finds } k \text{ in system}]$$

$$d_k = \lim_{t\infty} P[\text{departure at time } t \text{ finds } k \text{ in system}]$$

$$r_k = d_k$$

The Transition Probabilities

This section introduces two new random variables, q_n, the number of customers left behind the departure of C_n from service and v_n as the number of customers arriving during the service of C_n. The distribution for q_n gives us the transition probabilities for the imbedded Markov chain.

$$p_{ij} = P[q_{n+1} = j | q_n = i]$$

The transition probability matrix for the imbedded market chain is a upper triangular matrix. It is easy to see that

$$\alpha_k = P[v_n = k]$$

can be easily computed as

$$\alpha_k = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x) dx$$

The stationary probabilities of imbedded markov chain can be computed via

$$\pi = \pi \mathbf{P}$$

The Mean Queue Length

The main equation that connects the number in the system between two successive departures is

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1}$$

Taking expectations, one gets

$$E[\tilde{v}] = \rho = \text{ average number of arrivals in a service time}$$

If we square the main equation and take expectations, one obtains

$$E[\tilde{q}] = \rho + \frac{E[\tilde{v}^2] - E[\tilde{v}]}{2(1-\rho)}$$

The only unknown in the above equation is $E[\tilde{v}^2]$. The author derives an important relationship between z-transform of the probability distribution of the random variable $\tilde{v}$ and the Laplace transform of the pdf of the random variable $\tilde{x}$, evaluated at the critical point $\lambda - \lambda z$.

$$V(z) = B^*(\lambda - \lambda z)$$

Using the above equation, the author derives the *Pollaczek-Khinchin mean-value formula*. This gives the average number of customers in a M/G/1 system

$$\overline{q} = \rho + \rho^2 \frac{1 + C_b^2}{2(1-\rho)}$$

The above can be used to obtain the average number in the system for elementary Markovian systems. This equation can be used to answer quick questions like, Does M/D/1 system have more customers than M/M/1 ? Just plugging in to the formula shows the M/D/1 has fewer customers. Now one can use the powerful Little's law to get to the average time spent by the customer in the system. Since

$$\overline{N} = \lambda T$$

we get

$$T = \frac{1}{\lambda}\overline{N} = \overline{x} + \rho\overline{x}\frac{1 + C_b^2}{2(1-\rho)}$$

One also can infer the average waiting time in the system as

$$W = \rho\overline{x}\frac{1 + C_b^2}{2(1-\rho)}$$

Distribution of Number in System

The previous section dealt with average number of customers in the system, average time spent by a customer in the system and average waiting time of a customer in the system. Moments for q_n can be obtained by working with the equation

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1}$$

The key idea is to use the above equation and the z-transform of the limiting random variable $\tilde{q}$ and obtain Pollaczek-Khinchin transform equation - 1

$$Q(z) = B^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{B^*(\lambda - \lambda z) - z}$$

The above formula is immensely convenient given the service distribution $B(x)$. All you have to do is to take its Laplace transform,evaluate a the critical point $\lambda - \lambda z$ and then generate the entire probability distribution from the z-transform of $\tilde{q}$. The solution at the imbedded Markov points gives the solution for all points in time.

Distribution of Waiting time

The learning so far from Kleinrock has been absolutely terrific. For the first time, I am seeing the interplay between the Laplace transform of the service time distribution and all the important random variables in queueing theory such as distribution of customers in the system, waiting time in the system, busy period etc. Many months ago, when I read in a paper that even the simplest queueing system such as M/G/1 has a representation in terms of the LT of its service time distributions, I was completely clueless. Now thanks to this beautiful book, I understand it completely.

There are five quantities that are interrelated and understanding the equations describing the relationships between these quantities will totally demystify the M/G/1 system behavior.

- $V(z)$ - the z transform of $\tilde{v}$, the number of customers arriving during the service of a typical customer
- $B(z)$ - the z transform of $\tilde{q}$, the number of customers in the system at the departure of a typical customer
- $B^*(s)$ - the Laplace transform of the service time distribution
- $S^*(s)$ - the Laplace transform of the total system time distribution
- $W^*(s)$ - the Laplace transform of the total waiting time distribution

The relations among the above quantities are

$$\begin{aligned} V(z) &= B^*(\lambda - \lambda z) \\ Q(z) &= V(z)\frac{(1-\rho)(1-z)}{V(z) - z} \\ S^*(\lambda - \lambda z) &= Q(z) \\ S^*(s) &= W^*(s)B^*(s) \end{aligned}$$

The last three of the above equations can be recast so that each of the quantities $Q(z), S^*(s), W^*(s)$ can be represented in terms of $B^*(s)$. These three equations are called Pollaczek-Khinchin transform equations

The Busy Period and its duration

In a M/G/1 system, the length of busy time and length of idle time are important random processes. Denote $U(t)$ as the unfinished work in the system at time t. This is also called *virtual waiting time* as this measures the waiting time of a customer who enters the system at time t. If we trace the system, the random process $U(t)$ is a continuous state Markov process subject to discontinuous jumps. The surprising thing about $U(t)$ is that it is independent of the order of service. As long sA the server is work conserving, the queue discipline

is *inconsequential*. This property can be exploited in many ways. Denote the idle period distribution as

$$F(y) \triangleq P(I_n \leq y)$$

and busy period distribution as

$$G(y) \triangleq P(Y_n \leq y)$$

The key idea behind the system is this : Assume that the service is LCFS and each customer who enters in to the service generates his own busy period and these busy periods are statistically similar. Let $\tilde{n}$ be the number of customers that arrive when the first customer is undergoing service. Let x_1 denote the service time of the first customer. Then the busy period can be denoted by

$$Y = x_1 + X_{\tilde{v}+1} + X_{\tilde{v}} + \ldots + X_3 + X_2$$

Since $G^*(s) = E(e^{-sY})$, one can write the equation connecting the Laplace transform of the $G(y)$ and LT of the service time distribution

$$G^*(s) = B^*[s + \lambda - \lambda G^*(s)]$$

An interesting relationship between busy period and the utilization factor flows from

$$\lim_{s \to 0} G^*(s) = \lim_{s \to 0} \int_0^\infty e^{-sy} dG(y)$$

This is nothing but the probability that the busy period ends

$$P[\text{ busy period ends}] = G^*(0)$$

Number served in a Busy period

Using conditioning logic, and the fact that easy sub-busy period is statistically independent, one arrives at

$$F(z) = zB^*[\lambda - \lambda F(z)]$$

The key takeaway from the chapter is that one needs to look at any queueing system and think about getting a grip on the following variables and stochastic processes

- busy period distribution
- number served in a busy period
- average number of customers in a system
- mean queue length
- average time in the system
- distribution of the number of customers in the system
- distribution of the time spent by the customers in the system
- distribution of waiting time of the customers

6 The Method of Collective Marks

Analyzing any queueing system involves two steps. First involves coming up with a probabilistic argument like the equation

$$q_{n+1} = q_n - 1 + v_n$$

This requires careful thought process. The second step often involves solving the problem using some z transform or Laplace transform.

One can give a probabilistic interpretation to a generating function. Assume that each customer arriving in to a system is marked with probability $1 - z$ and not marked with a probability z. Denote $q(z, t)$ as the probability that no marked customers arrive in $(0, t)$. The author shows that $q(z, t)$ is nothing but the generating function of a Poisson process . Superb way of thinking about generating functions. I have never come across something which gave me a probabilistic reasoning behind a generating function. The chapter also gives a superb probabilistic interpretion of Laplace transform using catastrophe process. Probability that an event with $f(t)$ pdf occurs before a Poisson catastrophe process(rate s) is given by

$$P[\text{event occurs before catastrophe}] = \int_0^\infty e^{-st} f(t) dt = F^*(s)$$