Basics of Queueing Theory

Queuing theory is a branch of mathematics that studies and models the act of waiting in lines.

This module defines the building blocks of – and derives – **basic queuing systems**.

Contents

Introduction

Queueing Theory Terminology

- Poisson Distribution
- Exponential Distribution
- Erlang Distribution
- Input Process
- Output Process

Queueing Notation

Little's Queueing Formula

The $M/M/1$ Queueing System

- M/M/1 With Limited Capacity

The $M/M/c$ Queueing System

Introduction

Queueing theory boils down to answering simple questions:

- How likely is it that things will queue up and wait in line?
- How long will the line be?
- How long will the wait be?
- How busy will the server/person/system servicing the line be?
- How much capacity is needed to meet an expected level of demand?
- etc.

Introduction

Knowing how to think about these kinds of questions will help you **anticipate bottlenecks**.

As a result, you'll build your systems and teams to be more efficient, to have higher performance and lower cost, and ultimately provide better service both for yourself and for your customers.

Introduction

Let's take a simple example. Suppose a grocery store has a single checkout line and a single cashier. Suppose an average of one shopper arrives at the line to pay for their groceries every 5 minutes. Scanning, bagging and paying takes 4.5 minutes on average.

Will there be queueing and waiting? Intuition says no, there won't.

But that's not what really happens. In reality, there will be lots of shoppers waiting in line and they'll have to wait a long time!

Introduction

Fundamentally, queueing happens due to 3 phenomena:

- **irregular arrivals** – shoppers don't arrive at the checkout line on a regular schedule;
- **irregular job sizes** – shoppers are not all processed in 45 seconds. Some of them will take much longer;
- **waste** – lost time can never be regained. Shoppers overlap because earlier shoppers didn't have time to finish in their "allotted" 45 secs.

Introduction

Queueing gets worse when the following holds:

- **high utilization** – the busier the cashier is, the longer it takes to recover from wasted time;
- **high variability** – the more variability in arrivals or job sizes, the more waste and the more overlap (queueing) occurs;
- **few servers** – fewer cashiers means less capacity to absorb spikes of arrivals, leading to more wasted time and higher utilization.

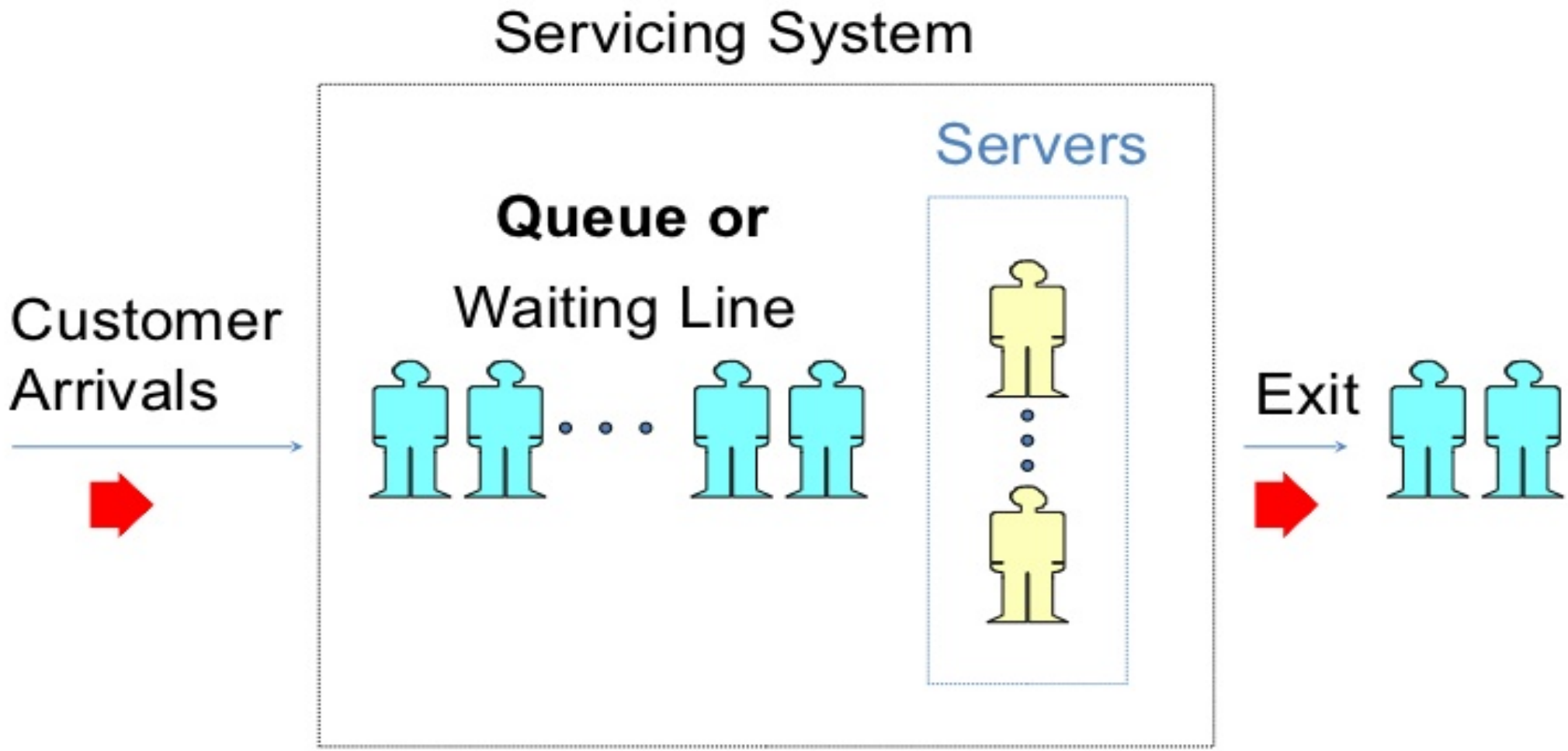
Introduction

All discussions of queueing theory analyze systems and processes in terms of three key concepts:

- **customers** are the units of work that the system serves (a customer can be a real person, a web request, a database query, a part to be milled by a machine, etc.);
- **servers** are the things that do the work (this might be the cashier at the grocery store, the web server or database server, or the milling machine, etc.);
- **queues** are where the units of work wait if the server is busy and can't do the work of processing them.

Queueing Theory Terminology

Components of the Queuing System



Queueing Theory Terminology

To begin understanding and describing queues, we must first have some knowledge of

- some useful probability distributions,
- an input process and
- an output process.

Poisson and Exponential Distributions

Both the Poisson and Exponential distributions play a prominent role in queuing theory:

- the Poisson distribution **counts the number of discrete events** in a fixed time period;
- the exponential distribution **measures the time between arrivals** of the events.

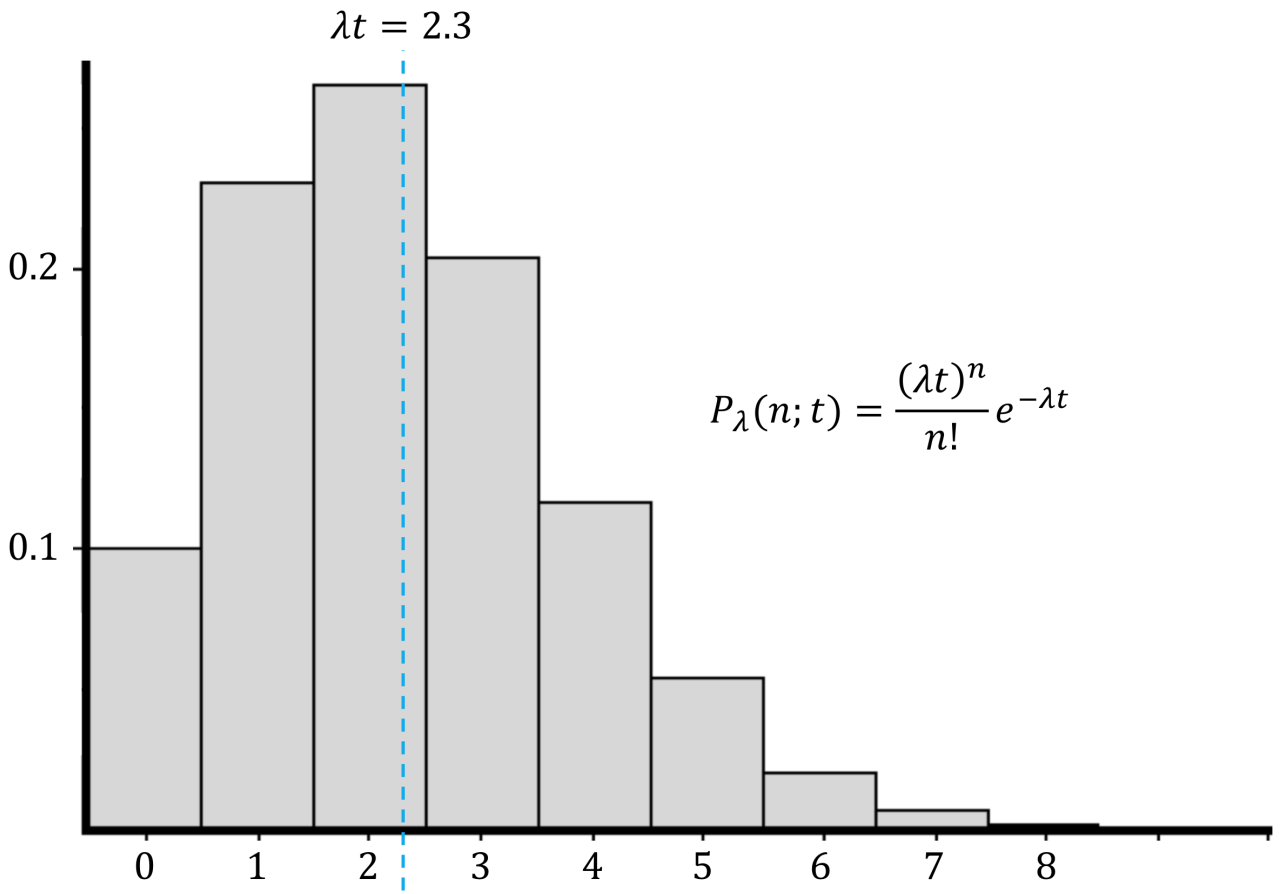
Poisson Distribution

Given an average arrival rate λ (in seconds, minutes, hours, days, etc.), the probability that n arrivals will be observed in a time interval of length t is:

$$P(n,t) = \frac{(\lambda t)^n}{n!}\, e^{-\lambda t}$$

where $n = 0,1,2, \ldots$

Poisson Distribution



Poisson Distribution

On average, 50 customers arrive in a coffee shop every hour.

What is the probability that exactly 20 customers will arrive in a 30-minute period, if the arrivals follow a Poisson distribution?

Poisson Distribution

Given $\lambda = 50$ customers per hour, $t = 30$ minutes $= 0.5$ hour, and $n = 20$, we have

$$P(20,0.5) = \frac{(50 \times 0.5)^{20}}{20!} e^{-(50 \times 0.5)} \approx 5\%.$$

Exponential Distribution

The time between successive arrivals is the **inter-arrival time**.

When the number of arrivals in a given time interval has Poisson distribution, inter-arrival times can be shown to follow an exponential distribution

$$f(t) = \mu e^{-\mu t}.$$

Hence, the probability that no more than t time periods are required in order to serve a customer is

$$P(W \leq t) = 1 - e^{-\mu t}.$$

Exponential Distribution

A manager of a fast food restaurant observes that an average of 9 customers are served by a waiter in a one-hour time period.

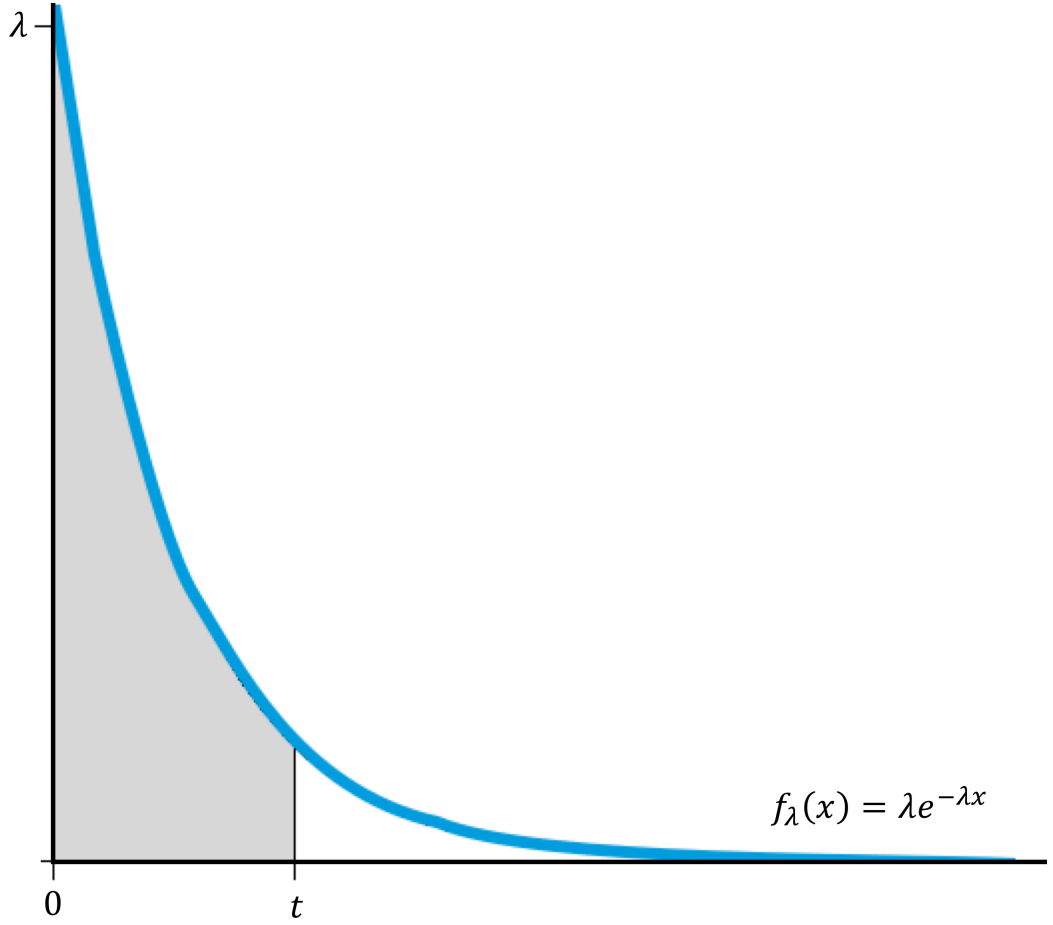
Assuming that the service time follows an exponential distribution, what is the probability that a customer will be served within 15 minutes?

Exponential Distribution

Let w be the average waiting time. Given $\mu = 9$ customers per hour, $t = 15$ minutes $= 0.25$ hour, we have

$$P(w \leq 15 \text{ minutes}) = 1 - e^{-9 \times 0.25} \approx 89\%.$$

Exponential Distribution



Memory-Less Property

$$P(X > t + h \mid X > h) = P(X > t), \qquad \forall h$$

The **memory-less property** of the exponential distribution implies that the probability distribution of the time until the next arrival is independent of the time since the last arrival …

(is that how it works for buses?)

Memory-Less Property

The time w a customer spends waiting in a bank queue is exponentially distributed with mean 10 minutes, say.

Then

$$P(w > 15 | w > 10)$$
$$= P(w > 5) = e^{-5/10}$$
$$\approx 61\%$$

If they've already waited 10 minutes, there is a 61% chance they'll wait more than 15 minutes in total.

Erlang Distribution

The exponential distribution is not always an appropriate model of inter-arrival times; wait times are not always memory-less, for instance (see bus example).

An alternative approach uses the **Erlang** distribution $\mathcal{E}(R,k)$, a random variable with 2 parameters $R > 0, k \in \mathbb{Z}^+$, whose p.d.f. is:

$$f_{R,k}(t) = \frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!}, t > 0.$$

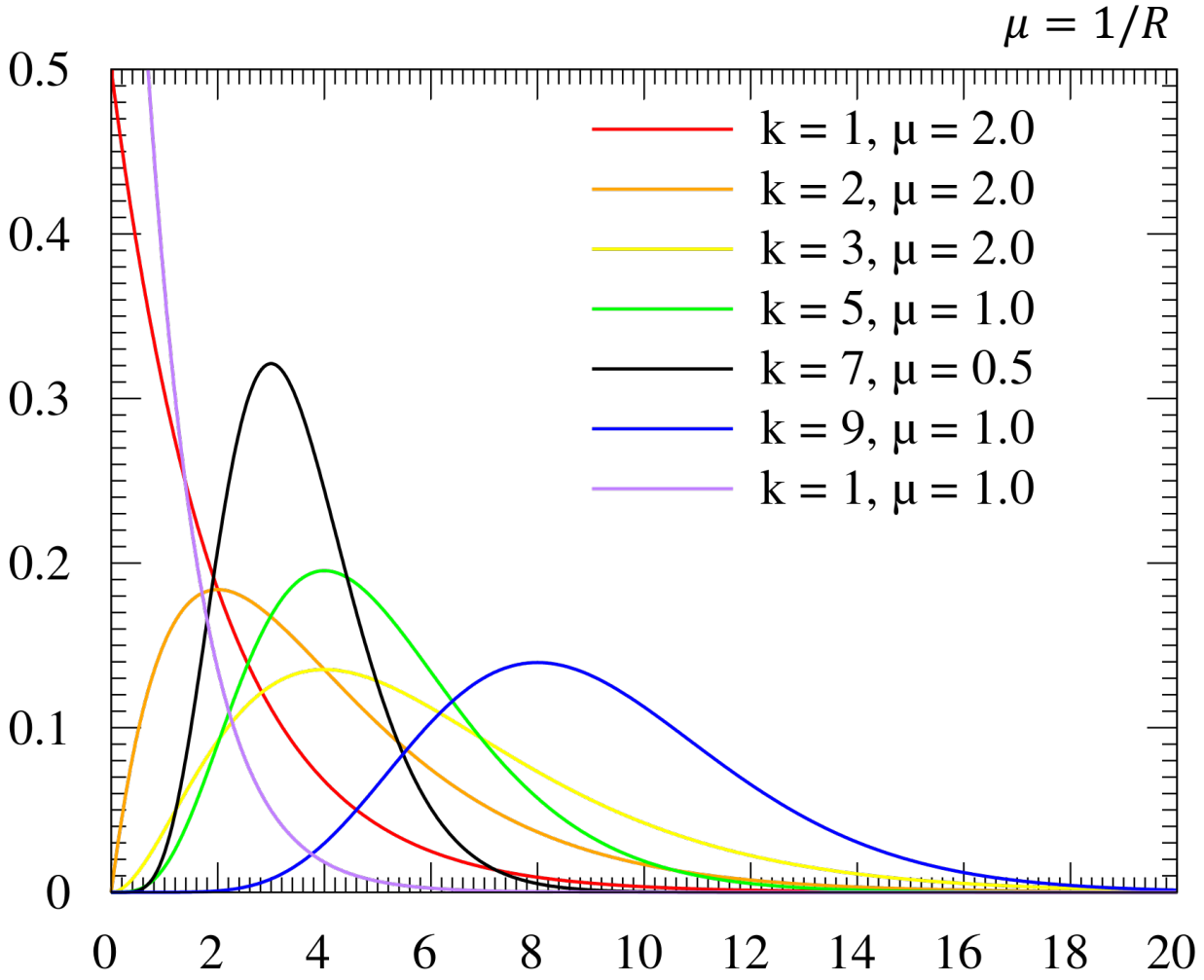
Erlang Distribution

When $k = 1$, $\mathcal{E}(R, 1) = \mathrm{Exp}(R)$. In general, we write $R = k\lambda$, and we obtain a decomposition into k independent exponentials:

$$\mathcal{E}(k\lambda, k) = \mathrm{Exp}(k\lambda) + \cdots + \mathrm{Exp}(k\lambda) = \sum_{i=1}^{k} \mathrm{Exp}(k\lambda)$$

If inter-arrival times follow an Erlang $\mathcal{E}(k\lambda, k)$, we are assuming that go through k memory-less phases before being served.

Erlang Distribution



Input or Arrival Process

Usually, we assume that the arrival process is **unaffected by the number of customers present in the system**.

In the context of a bank, this would imply that whether there are 500 or 5 people at the bank, the process governing arrivals remains unchanged.

Poisson arrival process are often assumed as many real-world arrival processes can be modeled using a **Poisson process** (but other processes can be used as well).

Output or Service Process

How long does it take to service a job or customer?

In most cases, we assume that the service time distribution is **independent of the number of customers present in the system**.

- is this realistic?

Servers in parallel or servers in series?

Exponential service times often assumed.

- works well for maintenance or unscheduled service situations.

Examples of Queueing Models

Situation	Input Process	Output Process
Bank	Customers arrive at Bank	Tellers serve the customers
Pizza parlor	Requests for pizza delivery are received	Pizza parlor sends out truck to deliver pizzas
Hospital blood bank	Pints of blood arrive	Patients use up pints of blood
Naval shipyard	Ships at sea break down and are sent to Shipyard for repairs	Ships are repaired and sent back to sea

Queueing Notation

Queueing (Kendall) Notation

Queueing systems can be described by six characteristics:

$$1/2/3/4/5/6$$

The 1-characteristic specifies the nature of the **arrival process**.

The following standard abbreviations are used:

- M = inter-arrival times are independent, identically distributed (iid) random variables having an exponential distribution;
- D = inter-arrival times are iid and deterministic;
- E_k = inter-arrival times are $\mathcal{E}(R, k)$, iid
- G = inter-arrival times follow a general distribution, iid

Queueing (Kendall) Notation

The 2-characteristic specifies the nature of the **service times**:

- M = service times are i.i.d. and exponentially distributed;
- D = service times are i.i.d. and deterministic.

The 3-characteristic represents the number of **parallel servers**.

The 4-characteristic describes the **queue discipline**:

- FCFS = first come, first served;
- LCFS = last come, first served;
- SIRO = service in random order;
- GD = general queue discipline.

Queueing (Kendall) Notation

The 5-characteristic specifies the **maximum allowable number of customers in the system**.

The 6-characteristic gives the **size of the population from which customers are drawn**.

In many important models $4/5/6$ is $GD/\infty/\infty$. When that is the case, $4/5/6$ is usually omitted.

Queueing (Kendall) Notation

Name (Kendall Notation)	Example
Simple system $(M/M/1)$	Customer service desk in a store
Multi-server system $(M/M/c)$	Airline ticket counter
Constant service $(M/D/1)$	Automated car wash
General service $(M/G/1)$	Auto repair shop
Limited capacity $(M/M/1/N)$	Barber shop with N waiting seats

Little's Queuing Formula

Little's Queuing Formula

λ = average number of arrivals entering the system per unit time

L = average number of customers present in the queuing system

L_q = average number of customers waiting in line

L_s = average number of customers in service

W = average time a customer spends in the system

W_q = average time a customer spends in line

W_s = average time a customer spends in service

Little's Queuing Formula

For any queuing system in which a **steady-state distribution exists**, the following relations hold:

- $L = \lambda W$
- $L_q = \lambda W_q$
- $L_s = \lambda W_s$

Example: if $\lambda = 46$ clients arrive at a restaurant every hour, on average, and if they spend $W = 10$ minutes before being served, on average, then there will be $L = 46 \times 1/6 \approx 7.7$ clients waiting to be served at all times, on average.

The $M/M/1$ Queuing System

The $M/M/1$ Queuing System

An $M/M/1$ system has exponential interarrival times with rate λ, exponential service times with rate μ, and one server.

Let $\rho = \lambda/\mu$ be the **traffic intensity of the queuing system**. Assuming $\rho \leq 1$, the probability of exactly n customers in the system is

$$\rho^n (1 - \rho), \qquad n = 0,1,2, \ldots$$

The probability of exactly no customers in the system is thus

$$(1 - \rho)$$

The $M/M/1$ Queuing System

Average number in service: $L_s = \rho$

Average number waiting in line: $L_q = \frac{\rho^2}{1-\rho}$

Average number waiting in the system: $L = L_q + L_s = \frac{\lambda}{\mu - \lambda}$

Average waiting time in service: $W_s = \frac{1}{\mu}$

Average waiting time in the line: $W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)}$

Average waiting time in the system: $W = W_q + W_s = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$

The $M/M/1$ Queuing System

Intuitively, if $\rho \geq 1$, then it must be that $\mu \geq \lambda$, and if the arrival rate is greater than the service rate, then the state of the system will grow without end.

Notice that (as expected) as ρ approaches 1, both W and W_q become very large.

For ρ near zero, W_q approaches zero, but for small ρ, W approaches $1/\mu$, the **mean service time**.

Single-Pump Gas Station

Suppose that all car owners fill up when their tanks are exactly half full.

At the present time, an average of 7.5 customers per hour arrive at a single-pump gas station.

It takes an average of 4 minutes to service a car. Assume that inter-arrival times and service times are both exponential.

[Erickson, W., 1973]

Single-Pump Gas Station Example

By assumption the single-pump gas station is a $M/M/1$ queueing system with $\lambda = 7.5$ arrivals per hour and the capacity to serve $\mu = 60/4 = 15$ vehicles per hour.

Then the

- traffic intensity is $\rho = \frac{\lambda}{\mu} = 0.5$;
- average number of customers waiting in this system is $L = \frac{\lambda}{\mu - \lambda} = 1$;
- average waiting time in the system $W = \frac{L}{\lambda} = \frac{1}{7.5} = 0.13$ hour $= 6\frac{2}{3}$ mins.

Single-Pump Gas Station

Suppose now that all car owners purchase gas when their tanks are exactly three-quarters full due to a gas shortage and panic buying takes place.

Assume that the average service time has been reduced to $3\ \frac{1}{3}$ minutes.

How has panic buying affected L and W?

[Erickson, W., 1973]

Single-Pump Gas Station

With these new assumptions, we have $\lambda = 2(7.5)=15$ arrivals per hour and the capacity to serve $\mu = 18$ cars per hour.

Then the

- traffic intensity is $\rho = \frac{5}{6}$;
- average number of customers waiting in this system is $L = \frac{\lambda}{\mu - \lambda} = 5$;
- average waiting time in the system $W = \frac{L}{\lambda} = \frac{5}{15} = 0.33$ hour $= 20$ mins.

Thus, panic buying has caused longer lines.

Single-Pump Gas Station

The previous example illustrates the fact that as ρ approaches 1, L (and therefore W) increase rapidly.

ρ	L for $M/M/1$ queueing model
0.30	0.43
0.60	1.50
0.80	4.00
0.90	9.00
0.95	19.00
0.99	99.00

$M/M/1$ With Limited Capacity

In real cases, queues never become infinite, but are limited due to space, time or service operating policy.

Examples: parking of vehicles in a supermarket is restricted to the spaces in the parking area; limited seating arrangement in a restaurant.

The probability of exactly no customers in such a system is

$$p_0 = \frac{1-\rho}{1-\rho^{N+1}}$$

where N is the maximum number allowable in the system.

$M/M/1$ With Limited Capacity

The probability of exactly n customers in the system is

$$p_n = \begin{cases} \rho^n \, p_0, & n = 1,2, \ldots, N \\ 0, & n = N+1, N+2, \ldots \end{cases}$$

The average number waiting in the system is

$$L = \frac{\rho \left[1 - (N+1)\rho^N + N \, \rho^{N+1}\right]}{(1-\rho)(1-\rho^{N+1})}$$

and $L_s = 1 - p_0$, $L_q = L - L_s$.

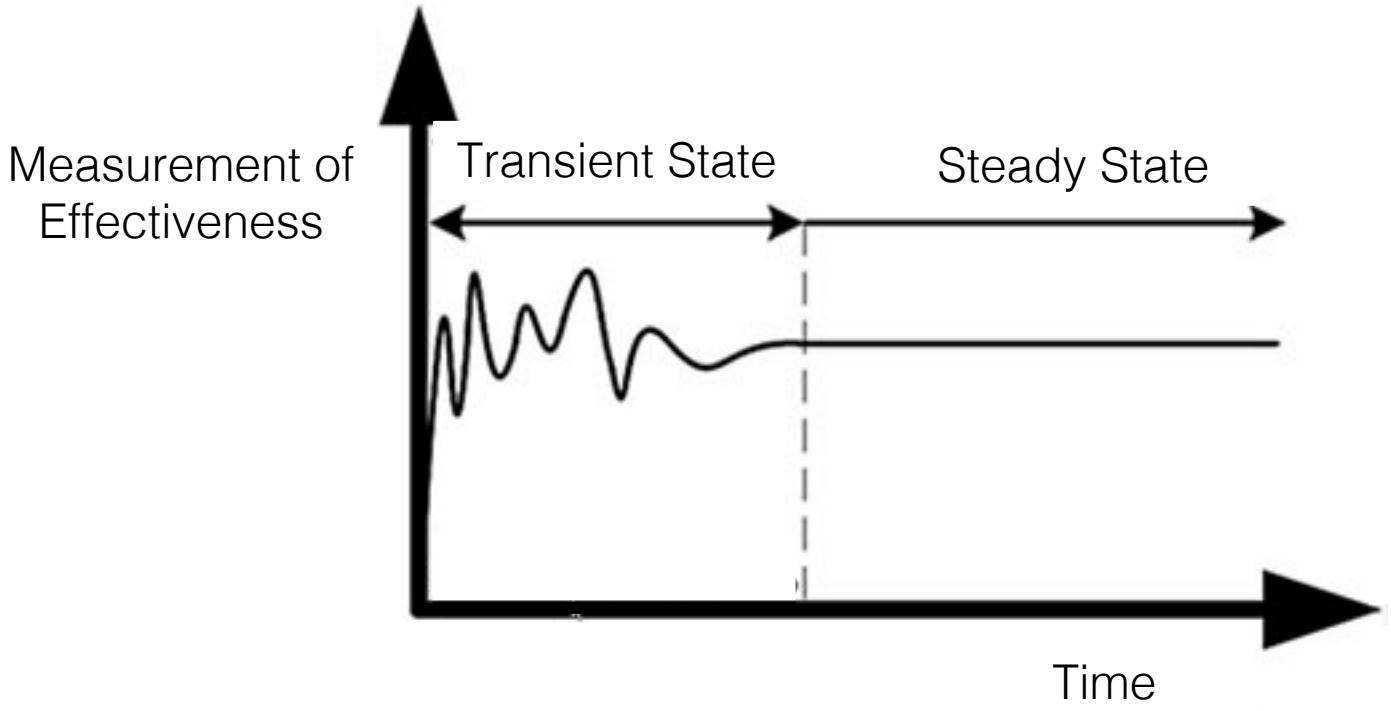
M/M/1 With Limited Capacity

Note that $\lambda - \lambda p_N$ arrivals per unit time actually enter the system on average due to the capacity limit. With this fact, we can show that:

$$W = \frac{L}{\lambda(1 - p_N)}, \qquad W_q = \frac{L_q}{\lambda(1 - p_N)}.$$

As a consequence of this restriction, a **steady state** always exists, because even if $\lambda \geq \mu$, there is never more than N customers in the system.

Steady-State (Queue Equilibrium)



Barber Shop

A 1-man barber shop has a total of 10 waiting seats.

Inter-arrival times are exponentially distributed, and an average of 20 prospective customers arrive each hour at the shop.

The barber takes an average of 12 minutes to cut each customer's hair (haircut times are exponentially distributed).

On average, how much time does an arriving customer spend in the barber shop?

Barber Shop Example

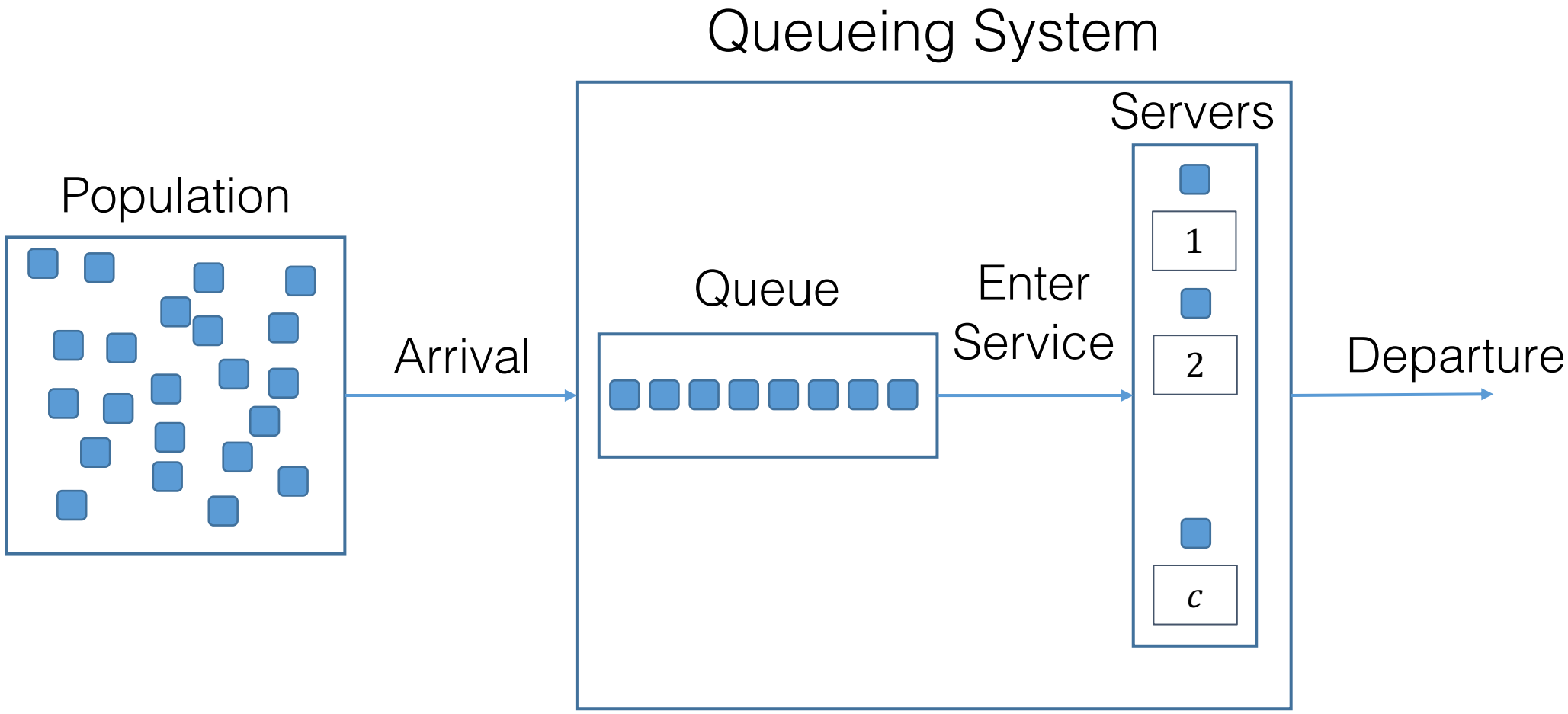
From the statement of the problem, $N = 10$, $\lambda = 20$ customers per hour, and $\mu = 60/12 = 5$ customers per hour. Then the traffic intensity in the system is $\rho = 20/5 = 4$, and we have

$$L = \frac{4\,[1 - (11)4^{10} + (10)\,4^{11}]}{(1-4)(1-4^{11})} = 9.67,$$

so that $W = \frac{L}{\lambda} = 1.93$ hours.

This shop is crowded, and the barber would be well advised to hire at least one more barber – what effect would that have on L and W?

The $M/M/c$ Queuing System


Queueing System
Servers
Population
1
Queue
Enter
Service
Arrival
2
Departure
c

The $M/M/c$ Queuing System

Same assumptions as $M/M/1$ except that the system now has c servers able to serve from a **single line of customers**, like one could find in a bank.

If each server completes service at rate μ, the system rate is $c\mu$.

The traffic intensity is $\rho = \frac{\lambda}{c\mu}$, and we again assume that $\rho \leq 1$.

If $\rho \geq 1$, no steady state exists. In other words, if the arrival rate λ is at least as large as the maximum possible service rate $c\mu$, the system "blows up" and the queue never empties.

The $M/M/c$ Queuing System

It can be shown that the **steady-state** (long-run) probability that all servers are busy is given by:

$$P(n \geq c) = \frac{(c\rho)^c}{c!\,(1-\rho)}\, p_0$$

where p_0 is the probability that there is no customer in the system (its formula is omitted for simplicity). We thus have

- $L_q = \frac{\rho}{1-\rho} P(n \geq c)$, $W_q = L_q/\lambda$
- $L = \frac{\lambda}{\mu} + L_q$, $W = \frac{1}{\mu} + W_q$.

The $M/M/c$ Queuing System

$P(n \geq c)$ for a variety of situations.

ρ	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$	$c = 7$
.10	.02	.00	.00	.00	.00	.00
.30	.14	.07	.04	.02	.01	.00
.50	.33	.24	.17	.13	.10	.08
.70	.57	.51	.43	.38	.34	.30
.80	.71	.65	.60	.55	.52	.49
.90	.85	.83	.79	.76	.74	.72
.95	.92	.91	.89	.88	.87	.85

Bank Tellers

A bank has two tellers.

An average of 80 customers per hour arrive at the bank and wait in a single line for an idle teller.

The average time to serve a customer is 1.2 minutes.

What is the expected number of customers present in the bank queue?

What is the expected length of time a customer spends in the bank queue?

Bank Tellers

We are dealing with an $M/M/2$ model with $\lambda = 80$ customers per hour and $\mu = 50$ customers per hour, whence $\rho = \frac{\lambda}{2\mu} = 0.8$. From the table, we have $P(n \geq 2) = 0.71$.

Then

- $L_q = \frac{0.8}{1-0.8}(0.71) = 2.84$ customers per hour
- $L = \frac{80}{50} + 2.84 = 4.44$ customers per hour
- $W = \frac{L}{\lambda} = 0.055 \text{ hours} = 3.3 \text{ minutes}$