Performance evaluation of a multiserver queue with dynamic joining probability and load dependent service rates

Aman Madaan G. V. Sandeep

October 7, 2013

Abstract

We will try to model the performance of a system which has multiple servers and multiple queues, but wherein the queue joining probabilities and the service rates shall be a function of the instantaneous queue length. It was observed that an arriving customer makes use of common sense(join the smaller queue) and experience (the knowledge of expected waiting time for each task) to select a queue to join. A visit to the IIT Powai branch of Canara bank motivated the idea and we propose the same as a case in point for this project.

Background

Queuing is a typical phenomenon in organizations which have a direct public interface. Such institutions usually have a number of employees (servers) who deal with the arriving customers. We observe that the rate of change of queue length depends on the common sense and the experience of the arriving customers. The service rates also depend on the instantaneous queue length. In situations where the number of customers in the queue is large, the server tends to hurry up. In cases when there are only a few customers in the queue, the server tends to work at leisure. Since the probability of joining queue and the service rates are ever changing in such a system, an analytical approach towards obtaining the answers to interesting questions about the system is unlikely to work. In addition, the server is not continuously serving. Occassionaly, there are events that demand the server's attention(this can be a phone call, or a quick tea break in our case). A simulation thus will be an ideal way to answer questions about such the system in consideration.

Motivation

Lets assume that a person with a strong mathematical background has been posted as the Branch Manager of a Bank. He is motivated to bring about more customer satisfaction by increasing the efficient usage of resources at his disposal. For this purpose, he is interested in knowing the current performance metrics which would help him in taking informed decisions. He decides to model the Bank as a queuing system with the help of two queuing system specialists. This situation that a newly appointed Bank Manager faces is the main motivation behind undertaking this queuing system simulation which would try to address all the situations and the special cases that the Manager might consider as significantly affecting the interested performance metrics of the Bank.

Some situations are mentioned here and the special cases are mentioned in a separate section.

A person is visiting a bank to perform three tasks, say

$$T_1,\ T_2\ and\ T_3$$

. There are 2 possibilities :

1. The order in which the tasks must be performed is predetermined. For example,

$$T_1 -> T_2 -> T_3$$

   Formally, the probability of joining any queue is a **constant**.

2. The person decides the order in which he/she should join a queue. This decision involes 2 factors :

   (a) **Common sense** :
   (i) Join the queue which is smaller.
   (ii) Join the queue which is being served more quickly.

   (b) **Experience** :
   A frequent visitor would know which queue will be better based on his past experience. This essentially means that the customer would have some a priori information of which server has a better service rate (*"The server on counter 3 is very slow"*).

   This implies that the order in which the queue grows is a function of it's current length! and thus makes the situation quite interesting. The probability of joining any queue keeps changing.

Assumptions

1. No preemption at any counter of the bank.
2. The maximum number of customers allowed in a queue is assumed to be finite and same for all the counters.

Parameters

Parameter	Definition
λ	Arrival rate of each customer
$\mu_{counter}$	Service rate of each counter
$I_{counter}$	Interrupt rate of each counter
T	Distribution of type of service requests of customers
$P_{counter}$	Probability of joining a counter
m	Maximum number of customers allowed in a queue for each counter
E	Distribution of experience of customers

Metrics

Notation	Definition
W	Average waiting time(in queue) of a customer between service times.
S_{peak}	Average service time of each of the counters during peak hours.
S_{lean}	Average service time of each of the counters during lean hours(off-peak hours).
R_{peak}	Average response time of a counter during peak hours.
R_{lean}	Average response time of a counter during lean hours.
D	Average departure rate of customers from the system.
$N_{counter}$	Average number of customers in a queue
$Q_{counter}$	Current number of customers in a queue of a counter

Proposed Methodology

The idea is a result of a visit to the IIT Powai branch of Canara bank. We plan to model the same in this project. The data for simulating the system would be randomly generated based on the situations and special cases considered from the perspective of the Branch Manager and also based on the peak and off-peak working hours of the branch. We'll try to be as close to reality as possible while doing this task.

The data shall include :

1. The customer arrival rate, along with the list of tasks for which they visit the bank.
2. The average service time at each of the counters.
3. The interrupt rate, maximum number of customers allowed in a queue and many other metrics of interest which have been mentioned in the Metrics section will also be randomly generated.

Special cases to be considered

1. For physically challenged / elderly people, there will be a separate queue.
2. Distribution of customer arrival will vary. It will depend on :
   (a) Time of the day
   (b) Day of the week
   (c) Special days : Start of the month, days after/before holidays or festivals.
3. Since most of the servers are humans, there will be a certain level of fatigue which will bring down the service rates. Along with this, there will be few automated servers also (for example, ATM machines or Automated Cheque Processing Systems inside a bank). Modelling these will be interesting as well.
4. There are limited number of waiting chairs in any bank. So as and when an arriving customer sees heavy rush in the bank(i.e with people standing in the queues and also waiting in chairs), he has two choices. Either he has to wait for a long period of time to avail the service or to leave the bank and arrive on some other day at the off-peak hours. This phenomenon is also dependent on the service the customer requires. Human instinct prefers the second option if the work to be done is not urgent. This is special case where people are arriving into the system but are leaving the system without getting the required service.
5. When a counter closes at the end of the day(during closing time of the bank), there might still be some customers left in the queue who are not being offered service on that day and they will have to arrive on some other day to avail the service. This is also similar to the case described above.
6. There might be cases where customers visiting a bank for the first time unknowingly enter a wrong queue and wait for a long time. When he arrives at the counter, he is asked to join some other queue that offers the service he requires. This also increases the average waiting time of the customers as a whole.

Some Interesting questions

Apart from obtaining the standard system metrics, we will try to find out answers to the following questions :

1. In some cases, the ordering of tasks in not under the control of the customer but they have to be done in a predefined order. (Ex: Opening an account in a bank is mandatory before depositing money in the Bank). How will the system react to this new scheme?

2. How will the essential metrics change if every customer uses just their common sense(join the smaller queue), and not experience?

3. Waiting time prediction : We plan to create an application which given the time and day of the visit, can predict the time that the customer might have to spend in the bank. A customer may then reschedule the visit or plan accordingly.

4. There may be services which are more popular than the others, we can find out which are the services that need more people.

Proposed Outcome of our Simulation Project

The outcome of our simulation project will be two fold. We are planning to make it beneficial both for the customers as well as for the bank.

1 From Customer's perspective

We are planning to develop a **GUI based tool** that will be useful for the customers while planning their next visit to bank. In this tool, the customer will provide the various tasks he/she is planning to during his/her visit to the bank and also the planned date and time. The tool will consider all possible factors that we mentioned in this document and will give an estimate of the **expected waiting time**. *It must be noted that there are several tasks which are inter-dependent and has to be performed in sequential order. Also sutitable time gap must be maintained (which may span upto days) between any two such tasks. The tool will also make the user aware of this and will help him/her to plan the visit(s) accordingly.*

We are also planning to add a feature in this tool which will show the customer a *Expected Waiting Time Variation* graph w.r.t time of a day. Using this, the customer will be able to choose the time to visit the bank without providing any detailed information.

2 From Bank's perspective

The bank will also get lots of information from the simulation results.

1. The bank managers will have an estimate of the average throghput of each employee at the counters. This may directly reflect upon their performance appraisals.

2. The results will also be useful while deciding about granting leaves to employees. The managers will come to know when in a month, there will be a huge rush and for which particular service(s) *(For example, there can be huge rush for withdrawing pension money at the beginning of the month.)* He may not want to grant leaves to many employees during that time.

A prototype of the Performance Predictor tool we are trying to develop looks like this :