

Performance Analysis Of Call Center Using Queuing Model And Simulation

G.D. Mishra

Professor, Department of Applied Sciences, Amity University, Noida

E-mail : gd.mishra@rediffmail.com

Vijiya Chauhan

Lecturer, Department of Applied Sciences, College of Science & Technology, Jhansi

E-mail: vsvijayachauhan@gmail.com

Abstract

A group of agents which deals a large number of incoming calls for various services is known as call center. It consist telephone trunk line, a switching machine called as automatic call distributor (ACD), a voice recorded unit (VRU) and telephone agents. A customer dials a number provided by the call center. In this paper we check the performance of call center by using quality of service and efficiency of call center agents. We also find the waiting time of a customer in ACD unit, total system response time, agents utilization, average number of customer waiting in ACD, blocking probability and total cost.

Key Factors

M/M/c and M/M/c/N queuing models, queuing model simulator

1. Introduction

A call center may be defined as a service unit where a group of agents handles a large volume of incoming telephone calls for the purpose of sales, services, or other specialized transactions. Typically, a call center consists of telephone trunk lines, a switching machine known as the automatic call distributor (ACD), a voice recording unit (VRU), and telephone sales agents.

Customers usually dial a special number provided by the call center. If a trunk line is free the customer seizes it, otherwise the call is lost. A fraction of calls that do not receive service become retrials that attempt to reenter service. Once the trunk line is seized, the caller is instructed to choose among several options provided by the call center via the VRU. After completing the instructions at the VRU, the call is routed to an available agent. If all agents are busy, the call is queued at the ACD until one agent is free. Once the trunk line is seized and until the caller leaves the system, any other customer cannot use the seized trunk line. Besides, an agent can service one caller at a time. Moreover, a caller remains in the system until it gets the requested service from an agent.

Consider a call center having three types of agents. The agents in each type have the ability to provide one, two, or three types of services. The call center is assumed to provide Banking (B), Insurance (I), and Travel (T) services. Each one of the three services has its own specialized software and database. The agents who provide more than one service need extra time to switch from one database to the other. The database switching time is assumed exactly 30 seconds. The calls arrive to the center with exponentially distributed inter-arrival time with mean $1/\lambda$. There are k trunk lines available. The VRU service time is assumed to be exponentially distributed with mean $1/\mu$.

2. Performance Measures

We check the performance of call center on the following basis

1. The quality of service is determined by the probability of blocking a customer. The average waiting time of a customer after dealing with the VRU till he talks to an agent. The low blocking probability of a customer is known as a good quality of service and a small average waiting time.
2. The efficiency of the system is measured by the utilization of agents. It is also measured by the total cost of all the agents working in the call center based on their salaries. The cross training cost is a one time cost, hence it is not considered.

For this we have calculated the following things

- Average number of the customers waiting in the ACD units
- Waiting time of the customers in the ACD units
- The percentage of the time when agent is busy that is agents utilization
- The total responding time
- The total cost depending on salary of the agents
- The probability which is given by the number of customer rejected divided the total number of the customers arrived to the system is known as blocking probability.

3. Model

The queuing system for the call centre is shown in Figure 4.1. The trunk lines and the VRU unit can be modelled as a loose multi-server system (with no waiting room) with k servers, and independent and exponentially distributed service times. K acts as an upper bound on the number of calls that can be in the system, either waiting or being served, at one time.

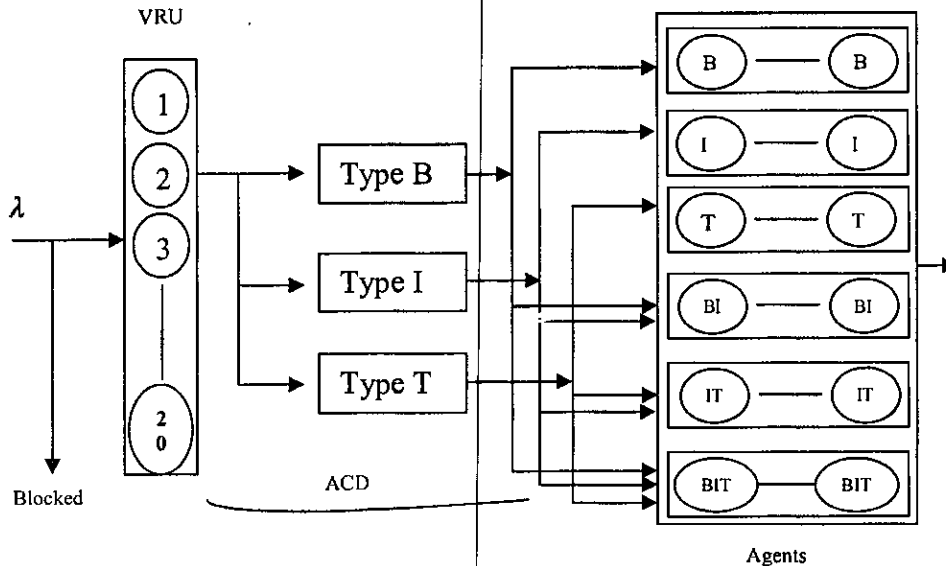


Figure 1.1: Queuing model for the call center.

The ACD is modelled with three queues, one for each type of service requested. We assume that customers do not hang up (leave the system when they get impatient). All customers that wait will eventually be served. The service discipline is FIFO per customer type. The agents are modelled as servers, which are divided into six categories depending on their skill level. The service time of an agent (the talk time to an agent) is independent and exponentially distributed. The number of agents $N \leq k$ provides an upper bound on the number of calls that can be in service simultaneously. The type of service required is divided probabilistically with probabilities of 0.3, 0.25, and 0.45 for B, I, and T services respectively. Finally, served customers may also return to the system. A state is characterized by the number of calls that are in process for each possible type/skill-group combination. The queuing system parameters of the simulated call centre are summarized in Table 1.1.

There are number of questions of interest related to skill-based routing. An important issue is the way in which skill-based routing is implemented. One policy is the probabilistic routing policy that assigns all different groups of agent with same probability that the next customer will be routed to it.

Parameter	Value	Description
N	15	Total number of agents
K	20	Number of trunk lines
$1/\lambda$	20 seconds	Average inter-arrival time
$1/\mu$	60 seconds	VRU average service
$1/\lambda_B$	5 minutes	Banking average service time
$1/\lambda_I$	6 minutes	Insurance average service time
$1/\lambda_T$	9 minutes	Travel average service time

Table 1.1: Parameters of the call centre queuing model.

Another policy is to route the call to the specialized agent group first. In this work, it is assumed that the first group to be considered when a customer requests a specific service is the group of specialized agents. If all of agents in the group are busy, the call will be routed to a group with two-skill agents that can provide the requested service. If all of those agents are also

busy then and only then the call will be routed to a three skill group of agents. If all the three-skill agents are busy, the call has to wait in the queue. The assumption is plausible because the performance of the specialized agents is considerably better than that of the multi-skill agents, since they do not require database-switching time. In addition, the use of multi-skill agents in the system provides the system with the flexibility to cope with load fluctuation in one service or the other, using them first will make them unavailable for the other services when overload in the other services happens.

4. QMS Implementation

The queuing model of Figure 1.1 is implemented using QMS modeller. Figure 1.2 shows the model used to simulate the call centre under study.

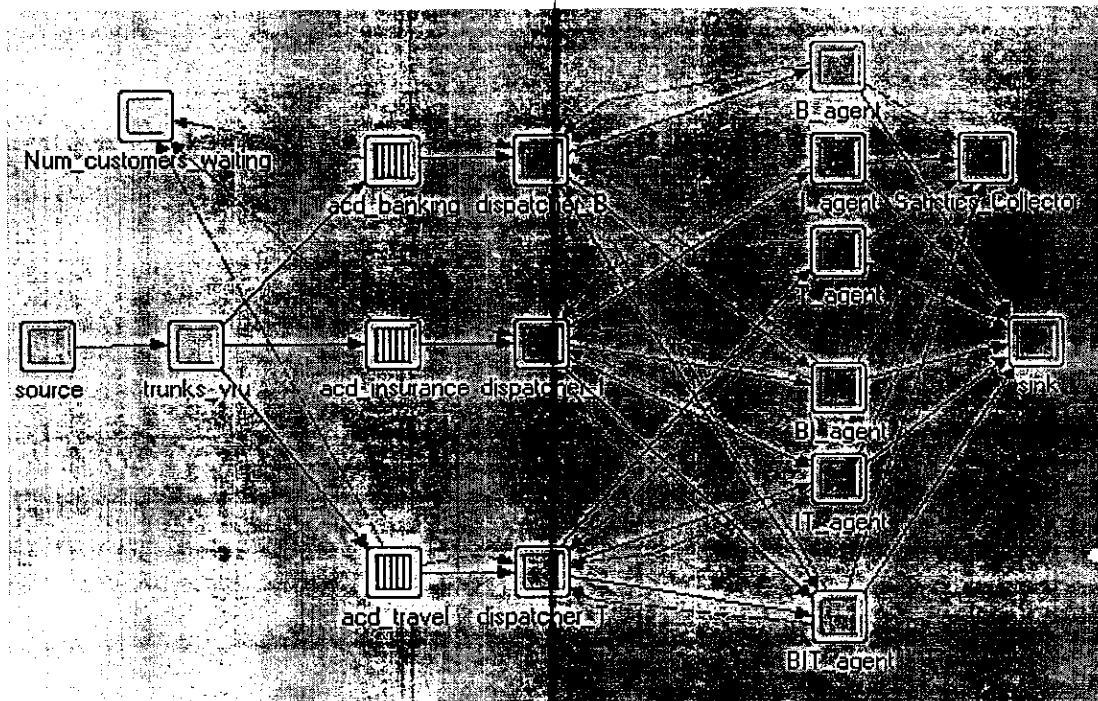


Figure 1.2: Model of Call Centre.

A source is to simulate the arrival of customers to the system. The attribute of this source can be easily modified to incorporate different arriving scenarios. A single multi server process with no waiting space and k servers implements the trunks and VRU.

An arriving call is simulated by arrival to the system. Three passive queues are used to provide waiting space for the customers depending on the three types of services. Three processors “dispatchers” are used to direct the customers (starting from the head of the queue) to the appropriate idle agents in order to be served. These dispatchers are designed in such a way to keep the strict FIFO policy throughout the system. After a customer is served, it will be sent to a sink node to be destroyed (to simulate departure from the system) and the occupied resources will be freed again.

5. Statistics And Calculations

The following statistics are collected during the simulation:

1. Waiting time of a customer in ACD unit.
2. Total system response time (end-to-end delay).
3. Agent’s utilization: percentage of time an agent is busy.
4. Average number of customer waiting in ACD.
5. Blocking probability: provided by the system as the number of customer rejected divided by the total number of customers arrived to the system.
6. Cost: calculated depending on the agent’s salaries.

6. System Verification

WE solve the $M/M/c$ and $M/M/c/N$ queuing systems analytically. It provides a practical way to compare the simulation results with that obtained analytically for verification purposes. Table 1.2 shows some of these cases we tested. One of the models used for validation is the $M/M/15$, which is implemented by making the VRU service time 0 and the number of trunks very large to prevent customers from being blocked. There is only one type of service in this scenario, and only banking agents (node no. 5 to 19) in some customers get blocked. Table 1.2 shows how close the observed results can be to the calculated ones. As shown, the error rate is less than 1% at most. Hence, we concluded that the simulator behaviour is close enough to the assumed system behaviour.

Performance Analysis of Call Center Using Queuing Model and Simulation

PASION Queuing Model Generator - file: C:\Users\WARESH\Desktop\qms\work\Call Centre\MM15.blq
 File Edit/view structure Define parameters View parameters Simulate Help Animator Quit

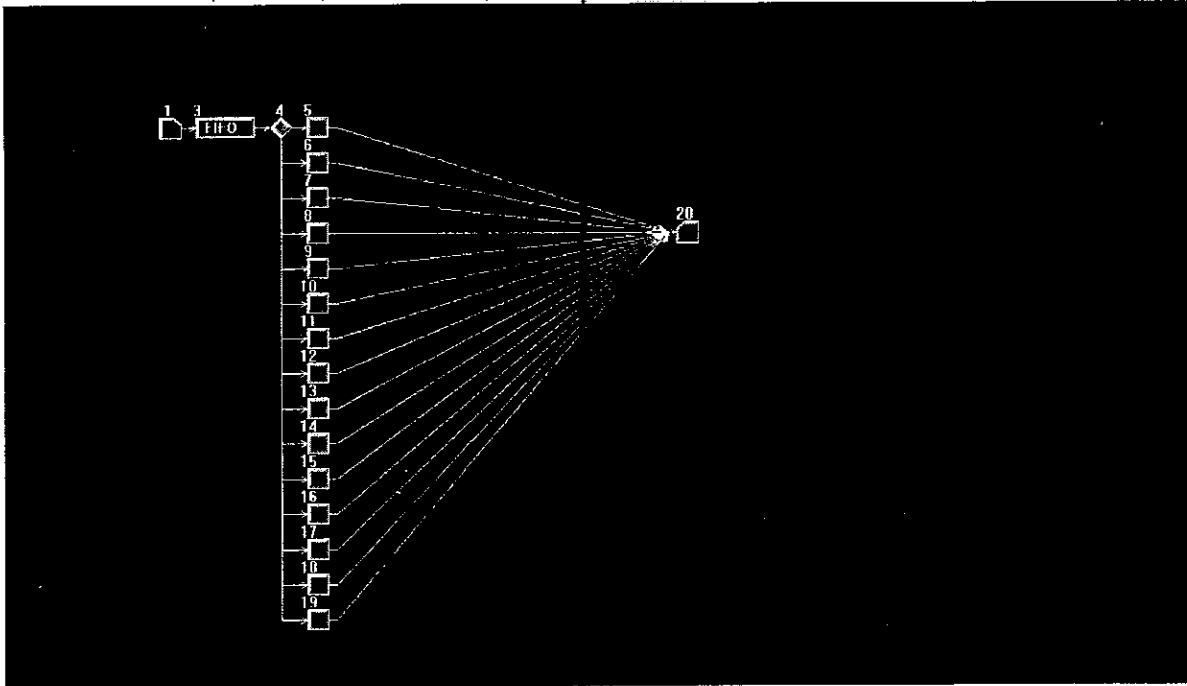


Figure 1.3: QMS Model for M/M/15.

Scenario	Parameter	Analytic	Simulation	Error%
M/M/15 1/λ = 30 sec, 1/μ = 300 sec	L _Q	0.204	0.205	0.49
	W _Q (Sec)	6.122	6.15	0.46
M/M/15/20 1/λ = 15 sec, 1/μ = 300 sec	L _Q	2.877	2.9	0.8
	W _Q (Sec)	58.74	59	0.44
	P _{Blocking}	0.265	0.266	0.37

Table 1.2: Verification results

7. Simulation Results

Table 1.3 defines different scenarios depending on the number of agents in each category. The symbols B, I, and T refer to the number of agents in B, I, and T groups respectively. BI, IT and BIT refer to cross-trained agents of two or all of these services.

8. Comparison And Interpretation

Specialized agents cost less in the sense of wages, training requirements, management becomes easier in certain aspects, and they provide scalability. On the other hand, multi skill agents cost more, need more training, and are less efficient in each individual skill, but they provide more flexibility in dealing with different types of services required. Having only generalist (e.g., scenario 1 in Table 1.3) is only good for very small call centres. The advantage is having more flexibility that leads to a small waiting time. The disadvantage is that higher cost has to be paid. On the other hand, having only specialized agents will have the worst waiting time and the largest number of waiting calls. This means that more customers have to wait for longer time. However, it results in the lowest salary cost. An intermediate solution with specialists and generalists and some form of flexible task assignment is usually the best solution (e.g., scenario 4). From a customer perspective, having only multi-skill agents (e.g., scenario 1) is the best as it results in the lowest waiting time (0.248 seconds in this example). This is because the multi-skill agents could deal with all different service types requested so no customer will wait for a special kind of agent. From the call centre manager's perspective having only specialized agents (e.g., scenario 2) is the best as it costs the lowest (15p). Note that the blocking probability does not change tremendously when switching between these extreme scenarios. This is because the blocking probability is more associated with the total number of trunks and agents irrespective of their type in the cases we tested.

The average number of waiting calls in scenario 2 is more than five even though the largest number of calls that can enter the system at a time is 20, and the number of agents in the system is 15. The reason is that in scenario 2 the system has only specialized agents whom some

Performance Analysis of Call Center Using Queuing Model and Simulation

of them will be idle. Those idle agents cannot serve customers who require a service different from the service that the idle agents provide. This can result in many idle agents even when some queues are full of customers. This is where multi-skill agents become handy. They can move from one service type to the other to adapt the system to the dynamic load variation. This is why the other scenarios did not experience this phenomenon. In addition, a small amount of multiple skills (e.g., scenario 4) produces almost the same performance as if all agents had all skills (scenario 1) with much lower cost. Therefore, the conclusion is that the economies of scale could be obtained by cross training only a minor fraction of the agents.

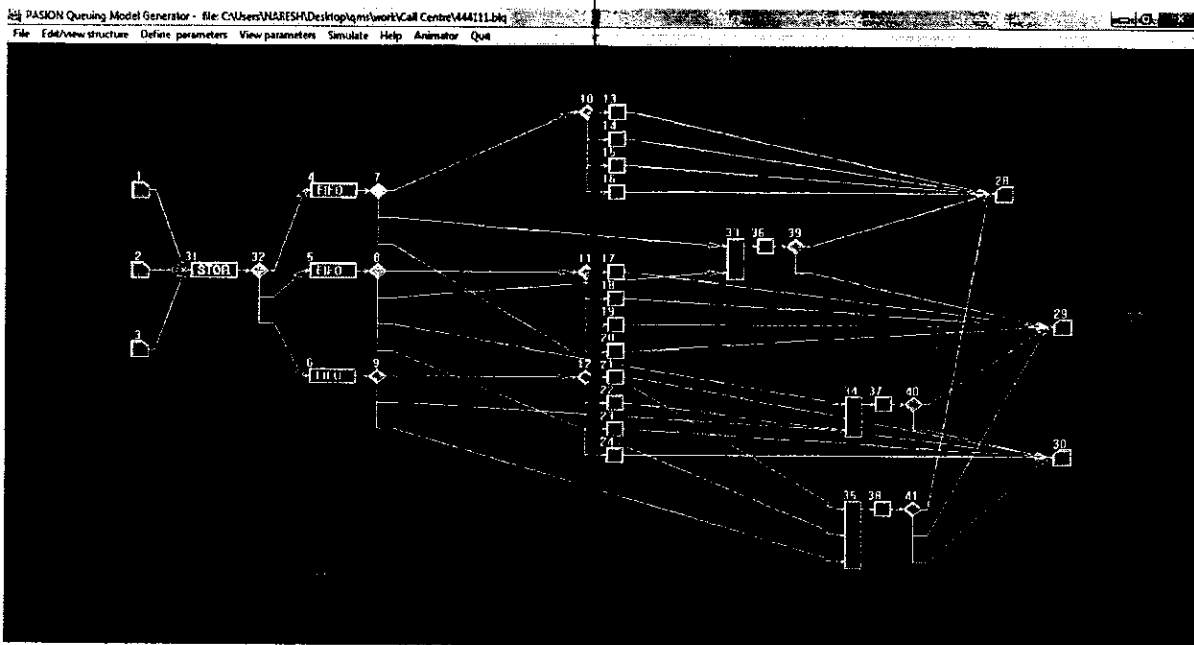


Figure 1.4: QMS Model for B,I,T,BI,IT ,BIT (4,4,4,1,1,1).

Scenario	No. of agents in B,I,T,BI,IT and BIT	Average waiting time (sec)	Agent utilizat- ion	Average number of waiting calls	Blocking Prob.	Cost
1	0,0,0,0,15	0.248	0.80	0.506	0.445	30 p
2	5,3,7,0,0,0	211.59	0.77	5.36	0.451	15 p
3	1,1,1,3,3,6	0.70	0.655	0.51	0.471	24 p
4	4,4,4,1,1,1	1.44	0.40	0.525	0.70	17 p
5	2,2,2,1,1,7	0.60	0.65	0.51	0.51	23 p
6	3,3,3,0,0,6	0.52	0.58	0.51	0.57	21 p
7	0,0,0,7,8,0	103.5	0.78	3.60	0.315	22 p

Table 1.3: Results collected from the simulator for different scenarios.

9. Conclusion

The simulator has been verified and a maximum error of 0.8% was obtained. The tested scenarios show that the use of only specialized agents results in more waiting calls and very large average waiting time; however, it costs less in terms of salaries. The quality of service deteriorates and more agents will be needed to overcome this deterioration. On the other hand, the use of all multi-skill agents enhances the overall service quality and increases the agent's utilization. However, the overall cost also increases.

Better performance can be achieved by using a combination of specialized and multi-skill agents. Using only a small fraction of multi-skill agents can do the trick. This means that the call centre management can plan for growth by only cross training a minor fraction of their agents. The initial and long term cost of the call centre expansion will be less in this sense.

References

1. Aksin, Z., Armony, M., Mehrotra, V., (2007). "The modern call-center: A multi-disciplinary perspective on operations management research." *Production and Operations Management* 16 (6), 665–668.
2. Banks, J., J.S. Carson II, B.L. Nelson, and D.M. Nicol, 2000. *Discrete Event System Simulation*, 3rd Ed., Printice-Hall.
3. Easton, F. F., J. C. Goodale. (2005). "Schedule recovery: Unplanned absences in service operations." *Decision Sciences* 36(3), 459–488.
4. G. Koole and A. Mandelbaum (2002). "Queueing models of call centers: An introduction." *Annals of Operations Research*, 113:41-59.
5. Gans, N., Koole, G., Mandelbaum, A., (2003). "Telephone call centers: Tutorial, review, and research prospects." *Manufacturing and Service Operations Management* 5 (2), 79–141.
6. Gross & Harris , *Fundamentals of Queueing Theory*, 3rd Ed., John Wiley & Sons, Inc.
7. Hur, D., V. A. Mabert, K. M. Bretthauer. (2004). "Real-time work schedule adjustment decisions: An investigation and evaluation." *Production and Operations Management*, 13(4) 322–339.
8. Koole, G & Talim, J., "Routing heuristics for multiskill call centers," *Proceedings of the 2003 Winter Simulation Conference*, 2003.
9. Koole, G., A. Pot. (2006). *An overview of routing and staffing algorithms in multi-skill customer contact centers*. Technical Report, Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands.

10. Koole, G., Mandelbaum, A. (2002). "Queueing models of call centers: An introduction." *Annals of Operations Research* 113(1– 4) 41–59.
11. L'Ecuyer, P. (2006). "Modeling and optimization problems in contact centers." in *Proceedings of the Third International Conference on the Quantitative Evaluation of Systems (QEST 2006)*, University of California, Riversdale, IEEE Computing Society, 145-154.
12. Law, A.M. and W.D. Kelton, 2000. *Simulation Modeling and Analysis*, 3rd Ed., McGraw-Hill.
13. Mehrotra, V., O. Ozluk, R. Saltzman (2006). "Intelligent procedures for intra-day updating of call center agent schedules." Working Paper, Department of Decision Sciences, San Francisco State University. Easton, F. F., J. C. Goodale. Schedule recovery: Unplanned absences in service operations. *Decision Sciences* 36(3) 459–488.
14. Van Dijk, N.M., 1999. On Hybrid Combination of Queueing and Simulation, proceeding of the 2000 Winter Simulation Conference.
15. Whitt, W. (1989). "Planning Queueing Simulations." *Management Science* 35 (11): 1341–1366.