Queueing Models of Call Center

Abstract: Telephone call centers are fundamental parts of numerous organizations, and their financial part is critical and developing. They are additionally interesting socio-specialized frameworks in which the conduct of clients and representatives is nearly entwined with physical execution measures. In these situations customary operational models are of incredible esteem – and in the meantime on a very basic level restricted – in their capacity to characterized system performance. This is an overview of some scholarly research on telephone call centers. The overviewed inquire about has its starting point in, or is identified with, lining hypothesis. To be sure, the "queueing-see" of call centers is both regular and helpful. In like manner, queueing models have filled in as pervasive standard help instruments for call focus administration. Notwithstanding, the advanced call focus is a complex socio-specialized framework. It in this manner appreciates focal highlights that test existing lining hypothesis as far as possible, and past.

Introduction: Call centers, or their contemporary successors contact centers, are the favored and pervasive path for some, organizations to speak with their clients. The call focus industry is in this manner tremendous, and quickly extending as far as both workforce and financial extension. For instance, it is evaluated that 3.8% of the U.S. what's more, U.K. workforce is included with call centers, the call focus industry appreciates a yearly development rate of 23% and, in general, the greater part of the business exchanges are directed over the phone.

There as of now exist a few academic surveys on call centers. (There are various overview papers in the business writing, which are not tended to here.) We know about five such articles: Pinedo et al. [56], giving nuts and bolts of call focus administration, including some investigative models; Anupindi and Smythe [7], which depicts the innovation that empowers current and conceivably future call centers; Grossman et al. [36] and Mehrotra [53], which are both short reviews of a few OR difficulties in call focus research and hone; lastly Anton [6], who gives an administrative overview of the past, present and fate of client get to (contact) centers.

Call Center and queuing systems: A call center is a centralised office which is used for receiving a large volume of valuable information by the means of telecommunication. Call center can be called as a backbone of a company. This area of study was started almost 100 year ago by the pioneering work of Danish telephone engineer Agner Krarup Erlang. An inbound call center (deal with incoming call) is operated to administer incoming product support from consumers. Outbound call (initiate call to the customers) centers are operated for marketing through phone, order taking, financial transaction, market research etc. Call centers enables customers to obtain a fast and exact response from the organizations. Call center provides a link between the customer and the service provider. According to recent survey call center industry is the fastest growing industry in the world. The reason for rapid development

of the call center is that both customer and company is profited with the remote service. Due to rapid development of the technology, the call centers also have other channels like email, fax, instant messaging and so on. Because of this the agents in a call center are more busier and the agents have to master many skills. Call center have significant general management challenges in human resources, MIS (Multi-user multiple site database) training and quality. Due to limited resources and unpredictable demand not all the call can be answered immediately. Information related to the delays is having a special importance in service systems with invisible queues, as in call centers. In this systems, the uncertainty involved in waiting is higher than that in visible queues, and it does not decrease over time. Customers have no means to estimate queue lengths or progress rate. So a feelings of frustration and anxiety increase during the waiting. By experiment it has been found that the delay information would avoid such these kind of situation . Zakay interpreted that waiting information may distract customer's attention from the passage of time. Hence, they may perceive the length of the wait as short. To overcome this number of model has been proposed by many mathematician . Some famous model are :- Erlang Model A , Erlang Model B , Erlang Model C , Poisson Model ,Markov Model , Reneging Model and many more are there . A queueing system is a stochastic system having a service facility at which a population arrives for service, and whenever there are more customer in the system than the service facility can handle simultaneously a queue develops. Queueing theory is a branch of applied probability theory that studies service system prone to congestion. In queueing system the input is an arriving population that enters the system in order to receive the service provided by the company. The output is the same population that leaves the system before or after receiving the service. As a consequence of a queueing model defines the interacting process and the nature of their interaction which determines the characteristics of the general process.

//

Models of the system

Model 1

Assumption made:-

1) There are two types of customers available namely call 1 and call 2. The two types of calls arrive according to a Poisson process with rates $\lambda_1$and λ_2, respectively. There are two queues, Queue 1 and Queue 2, which consist of customers of Call 1 and Call 2, respectively. We assume that the calls through the call center's selection system can be accurately classified

2) An assumption is made that customers may leave the queue due to impatience. The impatience time is exponentially distributed with the means θ

3) There are two categories of servers, Group 1 with N1 servers and Group 2 with N2 servers. Group 1 is of specialized servers who can only serve customers of Call 1, while Group 2 of flexible servers who can serve customers of both Call 1 and Call 2. The service times of servers in Group 1 and 2 are all exponentially distributed with means μ_1 and $\mu_2$respectively.

4) The routing policy of the model is based on skills and the importance of the two different types of calls. It is assumed that Call 1 is important than Call 2. In other words, Call 1 has non-primitive priority that Call 2. When a server in Group 2 completes his (her) service, if there are customers of Call 1 waited

in Queue 1 this server will service a customer waited in Queue 1, and if there is no customers of Call 1 waited in Queue 1 this server will serve a customer waited in Queue 2. When a server in Group 1 completes his service, if there is a customer waited in Queue 1 he/she will serves a customers of Call 1, otherwise he/she will be free.

5) There are infinite waiting spaces for both queues. For the same type calls, they are served in First-come First- serviced discipline. These queues are independent of each other.

Calculation for steady state :-

There are 7 state sets in the system, let Si,(i=1,2,…,7)denote the specific state set, define S1 is the state set that the agents in Group 1 are the idle state (n1<N1), and the agents in Group 2 are the idle state (n2<N2) too. S2 is the agents in Group 1 are the idle state (n1<N1), but the agents in Group 2 are the just full state (n2=N2). S3 is the agents in Group 1 are the idle state (n1<N1), but the agents in Group 2 are the busy state (n2>N2). S4 is the agents in Group 1 are the just full state (n1=N1), but the agents in Group 2 are the idle state (n2<N2). S5 is the agents in Group 1 are the just full state (n1=N1), and the agents in Group 2 are the just full state (n2=N2) too. S6 is the agents in Group 1 are the just full state (n1=N1), but the agents in Group 2 are the busy state (n2>N2). S7 is the agents in Group 1 are the busy state (n1>N1), and the agents in Group 2 are the busy state (n2>N2)

(1)The change of states due to arrivals of calls

Mechanism :-

For n1≤N1-2, if a call arrives at the system then the set S1 will not be changed. For n1=N1-1, if a call arrives at the system then set of states will be changed from set S1to set S4. The trigger of the transfer from set S1 to set S4 is due to arrivals of Call 1, and the arrive rate is ⍰⍰. Thus, we can obtain the transfer rate from set S1 to set S4

$$q(s_1 - s_4) = \lim_{\Delta t \to 0} \frac{P_{s_1,s_4}(\Delta t)}{\Delta t}$$

$$\lambda_1 \times P(n_1 = N_1 - 1)$$

$$P(n_1 = N_1 - 1) = , \frac{1}{(N_1 - 1)!}\left(\frac{\lambda_1}{\mu_1}\right) N_1 - 1 \Big/ \sum_{j=0}^{N_1} \frac{\left(\frac{\lambda_1}{\mu_1}\right)^j}{j!}$$

A similar analysis can also get the rest of the seven state transfer rate caused by call arriving, as follows:-

$q(S_1\text{-}S_2) = P(n_2=N_2\text{-}1)\lambda_2$; $q(S_2\text{-}S_3) = \lambda_2$;

$q(S_2\text{-}S_5) = P(n_1=N_1\text{-}1)\lambda_1$;$q(S_3\text{-}S_6)=P(n_1\text{-}N_1\text{-}1) \times \lambda_1$;

（2）The change of states due to leaves of calls:

Mechanism :-

One leaves of calls due to the service completed, the state of the agents will from the just full state to the idle state. Consider the state-transfer of agents in Group 2 for set S_2 . In the set S_2 that $n_2=N_2$, if a call completed the service then set of states will be changed from set S_2 to set S_1. The trigger of the

transfer from set S_2 to set S_1 is due to service completed of call 2, and the service rate is $N_2\,\mu_2$. Thus, we can obtain the transfer rate from set S_2 to set S_1 as follows:

$$q(s_1 - s_4) = \lim_{\Delta t \to 0} \frac{P_{s_1,s_4}(\Delta t)}{\Delta t}$$

$$\lim_{\Delta t \to 0} \frac{P(n_1 = N_1 - 1 \cap there\ is\ arrival\ of\ calll\ in\ time\ Delta\ t\ and\ has\ not\ been\ fimished)}{\Delta t}$$

$$\lim_{\Delta t \to 0} \frac{(\lambda_1 \Delta t + \sigma(\Delta t)) \times 1 P(n_1 = N_1 - 1)}{\Delta t}$$

$$\lambda_1 \times P(n_1 = N_1 - 1)$$

One leaves of calls due to the service completed, the state of the agents will from the just full state to the idle state. If a call completed the service then set of states will be changed from set S_2 to set S_1. The trigger of the transfer from set S_2 to set S_1 is due to service completed of call 2, and the service rate is $N_2\mu_2$ Thus, we can obtain the transfer rate from set S_2 to set S_1 as follows:

$$\mathrm{q(S2 - S1)} = N_2\mu_2$$

Model 2:

Model Description

The queueing model of a call center with two classes of customers; valuable customers type A, and less valuable ones type B. The model consists of two infinite priority queues type A and B, and a set of s parallel, identical servers representing the set of agents. All agents are able to answer all types of customers. The call center is operated in such a way that at any time, any call can be addressed by any agent. The scheduling policy of service assigns customers A (B) to queue A (B). Customers in queue A have priority over customers in queue B in the sense that agents are providing assistance to customers belonging to queue A first. The priority rule is non-preemptive, which simply means that an agent currently serving a customer pulled from queue B, while a new arrival joins queue A, will complete this service before turning to queue A customer. Within each queue, customers are served in FCFS manner. Arrival processes of type A and B customers follow a Poisson process with rates λA and λB, respectively. Let λT be the total arrival rate, λT = λA +λB. Successive service times are assumed to follow a common exponential distribution with rate μ for both types of customers. Then, the server utilization ρ (proportion of time each server is busy) is ρ = λT/sμ. The condition for stability is ρ < 1, that is to say that the mean total arrival rate must be less than the mean maximal service rate of the system. The resulting model, referred to as Model 1, is shown in Figure 1. There are two reasons for considering common distributions for service times. The first one relates to the types of call centers that motivate our analysis. We are considering call centers where customers are segmented into different groups based on their value to the firm. This segmentation can be based on life time value or profitability. The call center then provides different levels of service to these groups. This type of service level differentiation is widely used in financial service and telecommunication call centres. In the presence of this type of segmentation, the difference between customer types is not related to

the statistical behaviour of customers but to their importance for the company, which we capture through priorities. In concrete terms, we assume for our models that customer queries do not differ from one type of customer to another. The second reason is due to the complexity of the analysis when assuming different behaviours in the statistical sense.

Predicting and Announcing Virtual Delays

When a new arrival call. There are two possibilities: either at least one server is idle, or all servers are busy. In the former case, the customer enters service immediately without having to wait. So, the service provider does not announce any information to the customer. In the second case, he has to wait in queue for service to begin. In the following we give the distribution of the waiting time of a new arrival. This analysis will be used by the service provider afterwards, in order to inform customers about their delays.

If the latter is larger than or equal to the number of servers s, then all servers are busy and the new arrival has to wait in queue. Let n_A be the number of type A customers in queue A seen by the new arrival, and nB that of customers B in queue B, $n_A, n_B \geq 0$. Finally, let n_T be the total number of customers in queues seen by our new customer,

$n_T = n_A + n_B$

Finitely Patient Customers

The analysis of a call center with a single group of identical agents, serving two classes of impatient customers, high and low priority classes. The model is identical to that described above .However in addition we allow customers to be impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time he will renege and is considered to be lost. Times before reneging for both types are assumed to be exponentially distributed with a common rate γ for both customer types.

Call Center Modeling with Announcement

Moving from the call center described above to a call center with delay announcement. On the contrary to a call center with infinitely patient customers, there is a modeling complexity when we provide delay information to customers, due to possible changes in their behavior. In this section, we investigate the impact of announcing delays on the customer abandonment experience. When we inform a customer about his anticipated delay, he will decide from the beginning, either to hang up immediately because he estimates that his delay is too long, or to start waiting in queue. In the latter case, there are two further possibilities. The first is that customers never abandon thereafter. The second possibility is that the customer patience will change, i.e., customers may abandon even if they had chosen to start waiting. It is easy to see that customers would have a patience behavior different from that in the original system (without announcement), depending on the information we provide to them. We refer the reader to Armony et al. and Guo and Zipkin for further details on the subject. Several forms of delays information are possible. The best is that we give to a new customer his actual delay, which cannot be known in advance because it is random. The most natural in practice is that the service provider gives a certain percentile β of the virtual delay distribution to each new arrival. The virtual delay is the time it takes for a server to become free for the customer of interest. In other words, it is the time until all higher priority customers ahead of the arrival leave the queue plus the duration of a service completion. Whitt has considered a similar problem for a single class call center. He proposed a model incorporating announcement by assuming that a new customer who finds all

servers busy balks with a given probability. Once a customer elects to wait in queue, he would never abandon thereafter. We assume that each new arrival comes with its own deadline of time patience, and paralleling to the model of Whitt we stipulate that a new customer elects to join the queue with the probability that a server becomes free for him (his virtual waiting time) before he would renege. This is exact only if we assume that the customer acts as if the delay information was his actual delay, which is not the case. We do not let customers renege once they join the waiting line. This may be reasonable for high values of β, since the estimation of the anticipated delay should be fairly accurate in that case, so that ignoring reneging would be valid. Assume that a new arrival finds n_A waiting type A customers in queue A, and n_B waiting type B customers in queue B. Note that implicitly we are focusing on new arrivals finding all servers busy. If the number seen by an arrival is less than s, then the new arrival never balks and enters service immediately. Let us come back to a new arrival finding all servers busy. It should be clear that the probability of balking for a type A new arrival depends only on n_A (due to the priority rule), say $p_A\, b_k(n_A)$. However, the probability of balking for a new type B arrival depends on the couple (n_A,n_B), say $p_B\, b_k(n_A,n_B)$. Furthermore, we should not fall in the confusion of only considering it as a function of $n_T = n_A+n_B$. Having different values of n_A and n_B, so that $n_T = n_A+n_B$ is held constant, would affect the virtual delay distribution of the customer of interest. The reason is that with delays information, the arrival rate of type A customers, seen by our new type B customer, is state of queue A dependent. As a consequence, not considering the couple (n_A,n_B) to compute the balking probability of that customer would lead to a wrong result.

Model 3:

The basis for proper selection of a mathematical model to describe a specific call-centre in practice represents the knowledge of the probability density functions of inter-arrival times (for example times between two successive incoming calls) and service times (generally calls length). These functions can be procured if accurate and complete data about the call-centre operation are available. Since most of modern call centres use modern technology, which enables automatic logging of all the events in the call centre, the data needed for the mathematical analysis are usually provided. However, the lack of expert knowledge in practice prevents the companies from efficient use of them. The field data of the call-centres operation will be used to analyse the arrival and service arrangements. On the basis of this analysis an appropriate theoretical queueing model will be selected to describe the call centre taken into consideration.

A typical queueing model consists of one or more service units (like the servers), arrivals of customers demanding the services, and the services process. When all the customers cannot be served at once, queues are formed. This brings costs (or losses) due to waiting which increase with the number of customers in the queue. To decrease the waiting costs and increase the service level guaranteeing better system performance different improvements can be followed. However, any kind of improvement often is linked to a certain investment leading to higher costs of the queueing system.

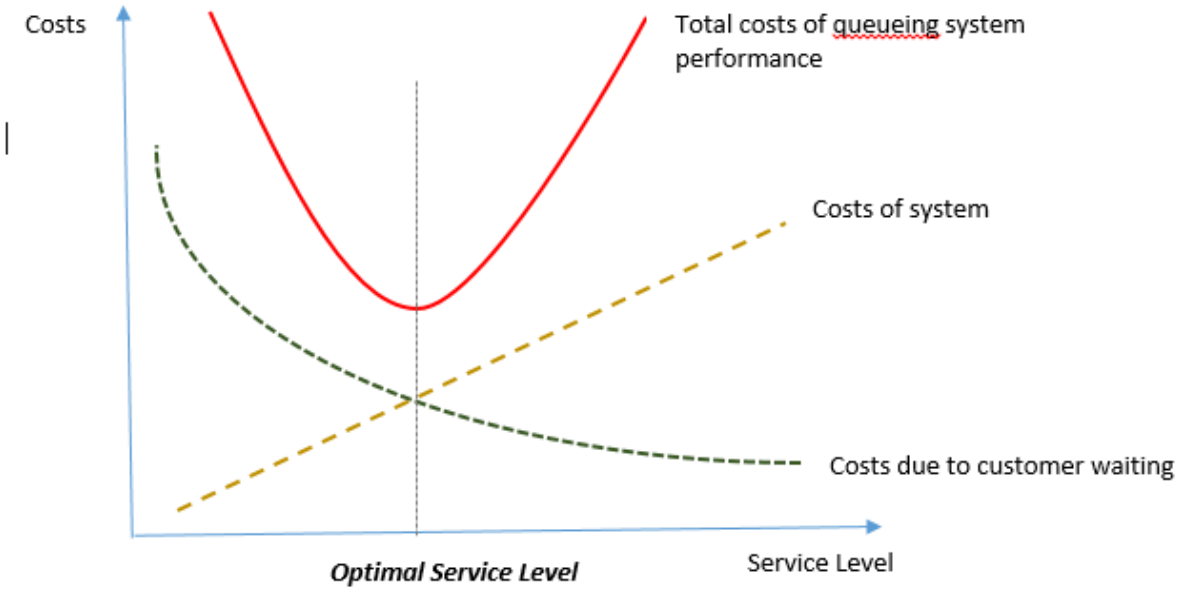

Figure shows that it is always possible to establish the optimal service level which ensures the minimum total costs of the queueing system performance.

To determine the optimal service level of a queueing system, different quantitative characteristics, like performance measures, can be used. The values of these measures can be calculated using an appropriate mathematical model. A suitable selection of the mathematical model is based on the following elements of the queueing system:

- Arrival process: Population of customers can be considered either limited (closed systems) or unlimited (open systems). Most mathematical models assume individual arrivals of customers and independent identically distributed inter arrival times.
- Service mechanism is determined with the system capacity, availability and probability density function of service times. Most mathematical models assume that service times are independent identically distributed random variables.
- Queueing discipline represents the way the queue is organised (First-In-First-Out (FIFO), Last-In-First-Out (LIFO), random selection of customers or selection based on customer priorities, for examples emergencies first).

Simple queueing models use the standard remark for describing the probability density function of inter-arrivals and service times: M – a Poisson process of the number of occurrences (i.e. customer arrivals or end of services); and an exponential density function of times between two successive events. G – A general distribution of times between two successive events (with a known mean and variance; for example, a normal density function). D – A deterministic situation; which means times between two successive events are constant. Notation M/M/c {infinity/infinity/FIFO} as a result describes the queueing system with c parallel serving channels, unlimited population, unlimited queue (no restriction for the maximum number of customers allowed to join the queue), and First-In-First-Out queueing discipline, meanwhile both of the inter-arrival and the service times are distributed according to the exponential density function. For many types of simple queueing models there exists

closed formulations for most system performance measures. Assuming that we have the M/M/c {infinity/infinity/FIFO} queueing model the closed form of all four performance.

The expected waiting time can be calculated based on the following equation:

$$E(Wq) = \frac{1}{s}\frac{(c\rho)^2}{c!\,(1-\rho)^2 c\mu} \qquad [1]$$

The expected number of waiting customers can be found according to the expression:

$$E(Wq) = \frac{1}{s}\frac{(c\rho)^2\rho}{c!\,(1-\rho)^2} \qquad [2]$$

The probability that one customer is going to wait because all agents are busy can be calculated as follows:

$$P_{wait} = \frac{1}{s}\frac{(c\rho)^2}{c!\,(1-\rho)} \qquad [3]$$

In the literature, the equation (3) is also known as the Erlang C formula and plays a determinant role in the performance of the telephone systems. The service level is the most frequent measure of quality of the call centres service. It is determined by a given percentile of the waiting time distribution that is given by the following expression:

$$SL(t_0) = P\left[W_q \le t_0\right] = 1 - \frac{1}{s}\frac{(c\rho)^2}{c!\,(1-\rho)}\exp(-(1-\rho)c\mu t_0 \qquad [4]$$

The equation (4) gives the long-term fraction of customers whose waiting time Wq in the queue is no larger than a given limit. The symbols used in equations (1), (2), (3) and (4) stand for:

c – number of serving channels

λ – arrival rate; 1/λ is the expected time between two successive arrivals

μ – service rate; 1/μ is the expected service time

ρ – traffic intensity calculated as ρ = λ /c μ

S – the sum which can be calculated by the following expression:

$$S = 1 + c\rho + \frac{c\rho^2}{2!} + \cdots + \frac{c\rho^{c-1}}{(c-1)!} + \frac{c\rho^c}{c!}\frac{1}{1-\rho} \qquad [5]$$

Equations (1), (2), (3) and (4) make sense when S $<\infty$. This condition stands if ρ <1 . The condition ρ <1 ensures that the steady state distribution exists. In this case the infinite queues are not formed and the queueing system still operates after a long run. The minimum number of servers cmin needed to satisfy the regular state condition is the lowest integer that fulfil the equation. c > λ / μ

Model 4:

POISSON PROCESS

In probability theory, a *Poisson process* is a stochastic process which counts the number of events and the time that these events occur in a given time interval. The time between each pair of consecutive events has an exponential distribution with parameter **λ** and each of these inter-arrival times is assumed to be independent of other inter-arrival times. The process is named after the French mathematician Siméon-Denis Poisson and is a good model of radioactive decay, telephone calls and requests for a particular document on a web server, among many other phenomena.

The Poisson process is a continuous-time process; the sum of a Bernoulli process can be thought of as its discrete-time counterpart. A Poisson process is a pure-birth process, the simplest example of a birth-death process. It is also a point process on the real half-line.

The basic form of Poisson process, often referred to as the *Poisson process*, is a continuous-time **counting process** $\{N(t), t \geq 0\}$ that possesses the following properties:

- $N(0) = 0$

- Independent increments (the numbers of occurrences counted in disjoint intervals are independent of each other)
- Stationary increments (the probability distribution of the number of occurrences counted in any time interval only depends on the length of the interval)
- No counted occurrences are simultaneous.

Consequences of this definition include:

- The probability distribution of $N(t)$ is a Poisson distribution.
- The probability distribution of the waiting time until the next occurrence is an exponential distribution.
- The occurrences are distributed uniformly on any interval of time. (Note that $N(t)$, the total number of occurrences, has a Poisson distribution over $(0, t]$, whereas the location of and individual occurrence on $t \in (a, b]$ is uniform.)

The *homogeneous* Poisson process is one of the most well-known Lévy processes. This process is characterized by a rate parameter λ, also known as *intensity*, such that the number of events in the time interval $(t, t + \tau]$ follows a Poisson distribution with associated parameter $\lambda\tau$. This relation is given as

[5]

where $N(t+\tau) - N(t) = k$ is the number of events in time interval $(t, t + \tau]$.

Just as a Poisson random variable is characterized by its scalar parameter λ, a homogeneous Poisson process is characterized by its rate parameter λ, which is the expected number of *events* or *arrivals* that occur per unit time. $N(t)$ is a sample homogeneous Poisson process, not to be confused with a density or distribution function.

In general, the rate parameter may change over time; such a process is called a *non-homogeneous Poisson process or inhomogeneous Poisson process.* In this case, the generalized rate function is given as $\lambda(t)$. Now the expected number of events between time *a* and time *b* is

$$\lambda a, b = \int_a^b \lambda(t)dt$$

Thus, the number of arrivals in the time interval $(a, b]$, given as $N(b) - N(a)$, follows a Poisson distribution with associated parameter $\lambda_{a,b}$.

$$P\big[\big(N(b) - N(a)\big) = k\big] = \frac{e^{\lambda_{(a,b)}}(\lambda_{(a,b)})^k}{k!} \qquad [6]$$

A homogeneous Poisson process may be viewed as a special case when $\lambda(t) = \lambda$, a constant rate.

Call focuses can be seen, normally and helpfully, as queuing frameworks. Which is an operational plan of a straightforward call focus. In a queuing model of a call focus, the clients are guests, servers (assets) are phone specialists (administrators) or correspondence hardware, and tele-queues comprise of guests that anticipate benefit by a framework asset. The least complex and most-generally utilized such model is the M/M/s queue, additionally known in call focus hovers as Erlang C. For most applications, be that as it may, Erlang C is an over-disentanglement: for illustration, it accept out occupied signs, clients anxiety and administrations traversed over various visits

Call focuses can be seen, normally and helpfully, as queuing frameworks. Which is an operational plan of a straightforward call focus. In a queuing model of a call focus, the clients are guests, servers (assets) are phone specialists (administrators) or correspondence hardware, and tele-queues comprise of guests that anticipate benefit by a framework asset. The least complex and most-generally utilized such model is the M/M/s queue, additionally known in call focus hovers as Erlang C. For most applications, be that as it may, Erlang C is an over-disentanglement: for illustration, it accept out occupied signs, clients anxiety and administrations traversed over various visits

//

Models

The multi-server queue M/M/m with impatient customers is called **Erlang-A** model, “A”

“Abandonment”, in contrast with the well-known **Erlang-B** model, M/M/m/m,

Erlang-C model, M/M/m with only patient customers.

Queuing model

The basic framework of our model is an M/M/K/J queue in the Kendall notation, where K, the number of servers, represents the total number of operators working in the centre and J, the maximum number of customers accommodated in the system, stands for the number of incoming telephone lines. In addition, as often with a queuing model of a call centre, customers in the waiting room may depart before getting service (abandonment). A new feature of our model is that each server (operator) must spend some amount of time for post-processing work after finishing the service with a customer. This corresponds to the after-call work (ACW). The customer leaves the system as soon as his service with an operator has finished. However, during the ACW the server cannot give service to another customer. Note that this feature make sour model different from the one assuming that each customer has effectively a service time consisting of two exponentially distributed phases. Unlike usual queuing models, we do not necessarily assume that $J \geq K$, because servers may be working on ACW while some customers are present in the waiting room.

Parameters:

Waiting customers are impatient such that they may leave the system before getting service at rate _, i.e. the patience time of each waiting customer is exponentially distributed with mean 1/_. If all customers are patient, i.e. they never leave the system once accepted until
Service completion, our model reduces to the one studied by Harris. We assume that the maximum number of customers accommodated in the total system is limited to J and the total number of servers is assumed to be K. Customers (calls) arrive in a Poisson process with rate _. The service time, i.e. the time each customer is processed by a server, is assumed to be exponentially distributed with mean 1/μ. After completion of service, the customer leaves
the system, while the server does not become free but starts the ACW of which the duration is assumed to be exponentially distributed with mean 1/_. After the completion of ACW, the server accepts a new customer, if any, from the waiting room. If there are no customers waiting, the server becomes idle. When a customer (call) arrives, if there are already J customers in the system, he is blocked and lost immediately and forever (we do not consider retrials of customer call). Otherwise, he is accepted

Transition of system state

Let N(t) be the number of customers present in the system, and let A(t) be the number of servers working for ACW at time t. Then the two-dimensional process $\{(N(t),A(t)), t \geq 0\}$ is a continuous-time Markov process with a finite state space $0 \leq N(t) \leq J$ and $0 \leq A(t) \leq K$.
We consider the steady-state distribution
$P_{jk} := \lim$
$t!1$
$P\{N(t) = j,A(t) = k\}$ $0 \leq j \leq J, 0 \leq k \leq K$.
Let us denote by (j, k) the state $\{N(t) = j,A(t) = k\}$. In this sate, the number of operators serving customers is $\min\{j,K - k\}$, and the number of waiting customers is $j - \min\{j,K - k\} = \max\{0, j - K + k\}$. The state transition rates to and from state (j, k) are shown in Figure 3. Four kinds of events to happen in the system state are the arrival of a new customer, the abandonment of a customer in the waiting room, the end of service, and the end of ACW.

Performance of the model

- Probability that an arriving customer is blocked (blocking probability): P_b,
- Probability that an arriving customer is accepted but waits (probability of wait): P_W,
- Mean number of customers present in the waiting room at an arbitrary time: E[L],
- Mean waiting time for both customers who are served and who abandon: E[W],
- Probability that an waiting customer abandons before getting service (probability of abandonment): $P\{Ab\}$,
- Fraction of time that each operator is either serving a customer or working for ACW (server utilization): U, and

Formuleas:

Blocking Probability

$$Pb=\sum_{k=0}^{K} \boldsymbol{Pjk}$$

Probability of Wait

$$Pjk=\frac{\boldsymbol{Pjk}}{\boldsymbol{1-Pb}} \qquad 0\leq j \leq J-1,\ 0 \leq k \leq K.$$

the probability of wait of an accepted customer is given by

$$P_w = \sum_{k=0}^{K} \sum_{j=K-k}^{J-1} Pjk \quad \text{if} \quad J>=K,$$

$$P_w = \sum_{j=K-k}^{j-1} \sum_{k=0}^{K} Pjk \quad \text{if} \quad J<K$$

Erlang A model, a queuing model often applied to analyze call center performance. While not a new model, Erlang A is becoming a popular alternative to the widely used Erlang C model. In this paper we analyze the accuracy of Erlang A predictions in high traffic environments, a situation where the Erlang C model is not applicable. Our findings indicate that in this high traffic region the Erlang A model is subject to a moderate to high level of error that has a strong pessimistic bias; that is the system tends to perform better than predicted. This is in sharp contrast to lower volume scenarios where the model tends to be optimistically biased. We find that in addition to utilization, the model is most sensitive to arrival rate uncertainty and balking.

In this section we present a revised model of a call center, relaxing several key assumptions discussed previously. In our model calls arrive at a call center according to a Poisson process. Calls are forecasted to arrive at an average rate of $\lambda\hat{}$. The realized arrival rate is λ , where λ is a normally distributed random variable with mean $\lambda\hat{}$ and standard deviation $\sigma\lambda$. The time required to process a call by an average agent is a lognormally distributed random variable with mean 1 μ – and standard deviation σ μ . Arriving calls are routed to the agent who has been idle for the longest time if one is available. If all agents are busy the call is place in a FCFS queue. When placed in queue a proportion of callers will balk; i.e. immediately hang up. Callers who join the queue have a patience time that follows a Weibull distribution with parameters α and β. If wait time exceeds their patience time the caller will abandon.

Calls are serviced by agents who have variable relative productivity i r . An agent with a relative productivity level of 1 serves calls at the 1792 Robbins average rate. An agent with a relative productivity level of 1.5 serves calls at 1.5 times the average rate. Agent productivity is assumed to be a normally distributed random variable with a mean of 1 and a standard deviation of σ r.

Our call center model is evaluated using a straightforward discrete event simulation model coded in Visual Basic. The purpose of the model is to predict the long term, steady state behavior of the queuing system. The model generates random numbers using a combined multiple recursive generator (CMRG) based on the Mrg32k3a generator described in (L'Ecuyer 1999). Common random numbers are used across design points to reduce output variance. To reduce any start up bias we use a warm up period of 5,000 calls, after which all statistics are reset. The model is then run for an evaluation period of 25,000 calls and summary statistics are collected. For each design point we repeat this process for 500 replications and report the average value across replications.

References:

[1] B. Cleveland and D. Harne (editors), Call Center Operations Management:Handbook and Study Guide, Version 2.1, ICMI Press, Colorado Springs(2004).

[2] M. J. Fischer, D. A. Garbin and A. Gharakhanian, Performance modelingof distributed automatic call distribution systems, TelecommunicationSystems, 9, No.2 (1998), 133–152, doi: 10.1023/A:1019139721840.

[3] N. Gans, G. Koole and A. Mandelbaum, Telephone call centers: tutorial,review, and research prospects, Manufacturing & Service Operations Man-agement, 5, No.2 (2003), 79–141, doi: 10.1287/msom.5.2.79.16071.

[4] C. M. Harris, K. L. Hoffman and P. B. Saunders, Modeling the IRS telephonetaxpayer information system, Operations Research, 35, No.4 (1987),504–523, doi: 10.1287/opre.35.4.504.

[5] R. J. Harris and M. J. Phillips, Extensions to a model for postcall activity in ACD systems, Preprint, December 1989. http://wwwist.massey.ac.nz/rharris/publicationfiles/pre2000 /acdsystems-foatrs.pdf