



QUEUEING THEORY FUNDAMENTALS



Contents

1	Introduction to Queueing Theory	2
1.1	Kendall's Notation	2
1.2	Queueing Models	3
1.2.1	Single-Server Queue	3
1.2.2	Multiple-Servers Queue	3
1.2.3	Infinitely Many Servers	4
2	Queueing Models in Call Centers	5
2.1	Erlang C and Erlang B Models	5
2.2	Erlang C Model	6
2.3	Erlang B Model	7
3	Data Requirements for Developing a Queueing Model	8
3.1	Data Analysis and Forecasting	8
3.2	Call Center Data	8
3.2.1	Using the Data	9
3.2.2	Modeling Primitives	9
3.2.3	Performance Management	9
3.3	Operational Regimes	10
3.4	Other Models	10
4	Call Center Applications	11
4.1	Workforce Management	11
4.2	Staffing Models	11
4.2.1	Square Root Safety Staffing	11
4.2.2	Single Skill Call Center	11
4.2.3	Multi-Skill Call Center	12
4.3	Skill Based Routing	12
4.4	Load Balancing in Geographically Dispersed Call Centers	13
4.5	Call Blending	13
5	Limitations of Queueing Theory	15
6	References	16

1 Introduction to Queueing Theory

Queueing theory is a mathematical discipline that studies the “phenomena of standing, waiting, and serving”, according to Leonard Kleinrock. In essence, any [system where a queue is formed and is served](#) can be modeled and studied using queueing theory. Practical applications of queueing theory in the field of telecommunications started in the early 20th century when Agner Erlang, the Danish telecommunication engineer, started applying principles of queueing theory in the area of telecommunications.

Queueing theory is fundamentally based on three things – a **queueing model** which is a mathematical representation of the characteristics and constraints of the queue, a **real world system** such as a telephone network or a call center that you are trying to model, and a **mapping** between the two. While the queueing model can be a simplified representation, the real world system can never be mapped completely to it, as there are several uncontrollable factors in a real world system. Practitioners of queueing theory strive to obtain as close a model as possible to map the real world system in order to predict the behavior of the real world system and consequently improve the performance of the real world system.

Queueing models are mainly used to determine steady state performance measures of the queue such as the average length of the queue, average time spent in the queue, and the probability that the queue is in a particular state at a given point in time – i.e. the queue is full, the queue is empty and so on.

From the perspective of call centers, these performance measures can be a measure of customer satisfaction as well as agent utilization and can help to arrive at the number of agents required to serve callers, as well as to route the calls effectively.

1.1 Kendall’s Notation

Queueing models are usually represented using notation which was initially comprised of three factors, A/B/C, suggested by D.G.Kendall in 1953. Later K, D and N were also included in the model to make it A/B/C/K/N/D. In most models, K, N and D are skipped and assumed to be ∞ , ∞ , and FIFO (First In, First Out) respectively. The notation of the queueing model is detailed below:

A: The arrival time distribution

B: The service time distribution

C: The number of servers or service channels. In the case of call centers, it could be the number of agents available.

K: The system capacity or the number of customers in the system at any given point in time, including those being served, as well as those waiting in queue. This is usually $C + \kappa$ where κ is the

buffer capacity. In the call center context, callers beyond K would get a busy signal. If all callers are placed in the queue, then K is assumed to be ∞ .

N : The calling population.

D : The queue discipline. It is usually assumed to be FIFO, where customers are served in the order in which they come in. Other possibilities include LIFO (Last in First Out), SIRO (Service in Random Order) and PNP (Priority Service). While most call centers follow a FIFO model, others have experimented with LIFO and Priority Service as well.

Some of the typical notations for A and B are given below:

- M – Markovian or Poisson distribution, denoting exponential service time
- $E\kappa$ – Erlang distribution with κ phases
- D – Degenerate distribution with fixed service time
- G – General distribution with arbitrary service time
- PH – Phase type distribution

A queueing model is usually representative of the steady state or the long run average state of the queueing system. As a result, most queueing models are stochastic in nature and are based on the probability that the system would be in a particular configuration.

In the next section, we will look at some of the most common types of queues encountered in real life.

1.2 Queueing Models

1.2.1 Single-Server Queue

This is one of the most common queueing situations in real life, where a single server is available to service several arrivals. Some common examples are a bank teller queue or an industrial production line. A single server queue with infinite capacity and infinite calling population, and where both the inter-arrival time and service time follows an exponential distribution, can be represented by $M/M/1$. A variation of this queueing model is the $M/G/1$ model where the service time may follow any general statistical distribution and not necessarily an exponential distribution.

1.2.2 Multiple-Servers Queue

A multiple server queue is usually the queue situation in a call center where the [incoming calls can be distributed](#) to any of the free agents. This is different from a network of single server queues where each customer needs to be serviced by a single agent and will wait until that specific agent is available even if other agents are free. A network of single server queues can be found in a banking system, where each employee handles a specific product and the skills of employees are not

interchangeable. One key insight that queueing theory has provided is that a multi-server queue performs better than a network of single server queues. In addition, a large pool of servers will perform better than a set of smaller pools even if the total number of servers in both systems remains the same. What this means to a call center is that it is always advantageous to have multi-skilled agents who can handle all types of calls rather than small pools of agents who can only handle a narrow portfolio of calls and then use skill based routing to direct the calls.

1.2.3 Infinitely Many Servers

This queueing model, represented by $M/M/\infty$, is not typically found in real life; however, it is useful to model situations where there are storage or delay aspects such as in a warehouse or a parking lot. In these models, there is really no queue as customers receive the service as soon as they arrive.

2 Queueing Models in Call Centers

The call center industry is rapidly growing all over the world, both in terms of workforce as well as revenue generated. It is estimated that nearly 3% of the global workforce is involved with call centers and it has an annual growth rate of 20%. In a fairly large call center, there will typically be several hundreds of agents who handle thousands of calls every hour. Agent utilization rates average between 90%-95%, the waiting time for customers and call abandon rates are closely monitored, and the industry aims to optimize these parameters without cost escalations.

Call centers can be easily modeled as queueing systems. In a call center queueing model, the customers are the callers and the agents or telecommunications equipment or IVR systems become the servers. The simplest representation of a call center is the M/M/c queue or the Erlang C model which we will discuss in detail in the next section. The model is an oversimplification for most practical situations. It cannot accommodate busy signals, customer dropouts, or services spanning over multiple calls. In addition, modern call centers also have a much more complicated network consisting of multiple queues. The queues are comprised of multi-skilled agents and IVR systems dispersed geographically over different interconnected call centers, where loads and peak times would vary.

2.1 Erlang C and Erlang B Models

In this section, we will look at the most popular queueing model used in call centers – the Erlang C Model (M/M/c/∞) and a variation of it known as the Erlang B Model (M/M/c/c).

Both models assume the existence of a finite number of servers, c , which operate in a system with a certain number of slots, r . r is always greater than or equal to c and r can also be infinite. Slots where servers are present are known as server slots; and waiting slots, denoted by $r-c$ are those where a server is not present.

Thus, each slot may have 0 or 1 server and 0 or 1 customer or caller. A caller may arrive in the system and leave immediately or stay for a certain duration. A customer may either occupy a server slot or a waiting slot.

Customer behavior in such a system can be characterized as follows:

When a customer arrives, if there is a server slot with no customers, then he occupies that slot, receives service, and leaves the system. If not, he will search for a waiting slot, and if available will wait in a waiting slot to be serviced.

A server behaves similarly in the system. The server will serve the customer and then look at the waiting positions that have customers in them. The server will then serve the longest waiting customer. If no customer is waiting, then the server remains idle. It is also assumed that customers arrive at random times with an arrival rate of $\lambda > 0$. At any point in time, t , the probability that at

least one customer will arrive in the system during the time period t to $t + \Delta t$ (where Δt is a small time duration), is approximately $\lambda \Delta t$.

Service time or average handle time (AHT) is measured as the duration of time when a customer is served by a server. The service rate μ is calculated as $1 / \text{AHT}$. The fraction of service time that is less than a given value t is denoted by $f = 1 - e^{-\mu t}$

The average number of simultaneous services that are trying to take place is known as the offered traffic load and is calculated as arrival rate/service rate. i.e $x = \lambda / \mu$. This is denoted in erlangs. Thus, if arrival rate in the queue is 20/sec and service rate is 2/second, then the traffic load is calculated as 10 erlangs.

2.2 Erlang C Model

The Erlang C model is denoted as M/M/n as it is assumed that customers arrive at a queueing system having n servers and an infinite capacity. Since the capacity is infinite, customers are not turned away and every customer will be served at some point in time. However, if the traffic load x (in erlangs) is equal to or larger than the number of servers n , then this system becomes unstable as the number of customers who are waiting keeps increasing over time. This model assumes a FIFO service discipline. The Erlang C queueing model is often used to model an inbound call center.

Given below are some of the key formulas for Erlang C based on steady state operations:

Assuming an arrival rate, λ and a service rate, μ :

- Probability of Delay, $C = C(n, x) = n * B(n, x) / (n - x * (1 - B(n, x)))$, where $B(n, x)$ is the Erlang B function.
- Average Handling Time, $AHT = 1 / \mu$
- Average Wait for Callers (including those with 0 wait time), $AWA = C / (\mu * (n - x)) = C * AHT / (n - x)$
- Average Wait for Delayed customers $AWD = AHT / (n - x)$
- Average time in system $T = AWA + AHT$
- Average Number of Waiting customers $WC = C * x / (n - x)$
- Average Number of Busy servers $BS = x$
- Average Number of customers in System $SC = BS + WC = x + (C * x / (n - x))$
- Utilization fraction (occupancy, average fraction of the time that each server is busy) $UF = x / n$

- Probability that the wait time of a customer will be less than or equal to t , $P(t) = 1 - C * e^{-(n-x)*t/AHT}$
- This is the same as the fraction of callers whose wait time is less than or equal to t .

2.3 Erlang B Model

The Erlang B model is similar to the Erlang C model with one key difference – there are no waiting positions. Thus, if a customer finds that there are no servers available, then the customer experiences a blockage and is lost. This model is also known as a **loss system**. While the Erlang B model is useful to include abandonment as a factor for analysis, the limitation of the model is the assumption that a lost customer does not return for a retrial. If a large number of customers try to call again, then the assumption of a Poisson arrival fails, and subsequently, the model fails.

Given below are some of the key formulas for Erlang B based on steady state operations:

Assuming an arrival rate, λ and a service rate, μ :

- Probability of Blockage, $B = B(n,x)$
- Average Handling Time, $AHT = 1/\mu$
- Average Number of Busy servers $BS = (1 - B)*x$
- Average Number of customers in System $SC = BS = (1 - B)*x$
- Utilization fraction (occupancy, average fraction of the time that each server is busy) $UF = BS/n = (1 - B)*x$

3 Data Requirements for Developing a Queueing Model

In the initial days, call center management was reactive in nature where performance metrics were measured and fine-tuned based on a trial and error method. Today a more scientific proactive method using software tools is used for planning and managing call centers. For example, when the wait time increases, instead of adding more agents, a change in the routing rules is attempted to reduce the wait time without increasing the number of agents. Quantitative analytic and simulation models are used for designing a call center and controlling its operations. Simulation models are often used increasingly when compared to analytical models because of their ease of use. Ideally, a call center should use both analytical and simulation models: analytical models to identify the characteristics of the call center and calibrate the technology and rules settings of the ACD, and simulation models should then be used to fine tune the model on an ongoing basis.

3.1 Data Analysis and Forecasting

Before developing a queueing model of a call center, it is necessary to do a careful data analysis. Even to develop a simple Erlang C model of a call center, it is necessary to estimate the call rate and the average wait time and service time. It is also important to keep in mind that the performance of call centers in peak traffic is very sensitive to even small changes in underlying parameters. Thus, it is essential to accurately estimate the parameters in order to develop a model that ensures consistent service levels.

3.2 Call Center Data

Call center data can be broadly classified into three types – operational data, marketing data, and psychological data. Operational data is usually collected by the ACD hardware. Marketing data is gathered by the CTI software, and psychological data is collected typically from post call surveys of callers as well as agent surveys to understand perceptions about service levels, service quality, and the working environment.

Operational data collected in the ACD include the arrival time of each call, the waiting time and the service time. Typically, ACD data is aggregated into totals and averages over a daily, weekly, or yearly format, and in some cases, supervisors use the data aggregated over a 15-minute period or an hour to monitor performance and fine tune the system. Individual transaction data is not stored, due to the large amount of storage space required. However, this can be used for data mining to identify profiles that can then be incorporated into a CRM system to further improve customer service.

The data collected by the CTI system include the caller ID. This can be used to obtain the caller's prior transaction history, which pops up on the agent screen as soon as the call is routed to the agent.

Apart from these three sources of data, some call centers record individual calls for legal or training reasons. However, these may be too subjective in nature to be analyzed effectively for developing a queueing model.

3.2.1 Using the Data

The dynamic nature of a call center where the agent groups and routing rules keep changing makes it challenging for use of historical data. Ideally, a call center should maintain several months of transaction data at an individual call level along with the utilization levels for each agent. It should also include the full trace of the calls, the period of availability of each agent, the break times, reasons for unavailability and so on. Aggregate data should also be available and in case of storage limitations, the aggregation of older data could be at a higher interval.

In short, as more and more call centers opt for a scientific queueing theory based model, the importance of data cannot be overstressed.

3.2.2 Modeling Primitives

Queueing models are built from parameters such as the arrival process, the service times, and the number of agents. In order to apply a queueing model, the call center must first estimate these parameters based on historical data. However, the biggest challenge for call centers is the question of whether the assumptions are valid in the first place. Arrivals may not follow a Poisson distribution, and service times may not be exponential. This will be a key area of research in future – to be able to identify and validate practically viable models for the parameters as well as performance measures of the call centers.

3.2.3 Performance Management

The fundamental tradeoff in call center management is between service quality and operational efficiency. Queueing theory based performance analysis of key metrics is used to aid this tradeoff by arriving at service levels and agent utilization as functions of the traffic load and available servers (agents). Research related to this area includes studies on caller patience and abandonment rates, retrial patterns, time-sensitiveness of operations and [skill based routing](#) among others. However, most theoretical models only support some key characteristics of the modern day call center and no analytical model accommodates all the complexities of a practical operation.

Most call centers rely on the Erlang C model or the M/M/c model, which describes a single type single-skill call center with c agents and which have a constant and random arrival rate with constant staffing and service rates. These assumptions could prove to be wrong in real life scenarios such as a new product introduction. The model also does not accommodate busy signals, abandonment, retrials, or time-variations (peak and off-peak loads). However, it is often used because it offers closed form expressions for most performance measures. These could be highly inaccurate as real life scenarios often violate the underlying assumptions in the Erlang C model. On

the other hand, the model $M/G/c$, which models a call center with non-exponential service times, is an analytically intractable model.

Call centers are often modeled on heavy traffic queueing systems that are usually present in call centers during peak time of high agent utilization.

3.3 Operational Regimes

Call centers can operate in one of the three possible regimes – 1. A quality driven operation where agent utilization levels are low, customer wait time is near to nil and priority is given for service levels than cost optimization. 2. An efficiency driven operation, where wait time is typically high and agent utilization is close to 100%. 3. A rationalized regime in which most call centers operate where quality and efficiency are balanced based on economies of scale. A typical characteristic of a rationalized regime is the fraction of delayed customers that will be neither close to 0 nor to 1.

3.4 Other Models

One way to model busy signals is to have the number of lines equal to the number of agents available, so that there are no delays. If all agents are busy, then customers are lost. This can be modeled using the Erlang B model discussed earlier, but this would result in a high rate of busy signals. Today, the practice in call centers is just the opposite. The number of lines is kept very high or IVR systems are used in order to ensure that busy signals rarely ever occur. However, the negative side of this is that customers are forced into long delays. Customers may prefer to get a busy signal rather than wait inordinately without being serviced, only to finally abandon the call. Some studies show that a good tradeoff between busy signals and abandonment can be obtained by having only 10% of lines in excess of total agents. Other models such as $M/M/c/B+M$ are used to model such call centers and patience is assumed exponential.

4 Call Center Applications

4.1 Workforce Management

Queueing theory based models help call centers in making decisions at both the tactical as well as strategic levels. For example, decisions on whether costly technology such as speech recognition or a voice response unit (VRU) should be deployed, the various skills groups to be employed, whether agents should be offered flexible work hours and so on, are better measured via quantitative analysis. There are models currently available which can aid in identifying the needs of the organization around recruitment and training based on operational characteristics and call center performance goals. However, care should be taken not to base staffing decisions on operational parameters alone. Otherwise, it could lead to over-utilization of agents, which in turn leads to burnout and higher staff turnover. This can be addressed by offering good working conditions and incentives.

4.2 Staffing Models

4.2.1 Square Root Safety Staffing

One of the most common formulas used in call centers to arrive at the number of agents required is known as the **square-root safety-staffing principle**. Here, the number of servers (agents) s , is calculated as

- $s = R + D = R + \beta \sqrt{R}, -\infty < \beta < \infty,$

where $R = \lambda / \mu$ is the offered load (λ = arrival rate, μ = service rate) and β represents service grade. The actual value of β depends on the particular model and performance criterion used. For example, in an Erlang C model, β could be taken as a function of the ratio between staffing costs and delay costs. The square-root principle is essentially asymptotically optimal for large heavy traffic call centers ($\lambda \uparrow \infty, s \uparrow \infty$), and it is mostly used for a rationalized call center which aims to achieve a balance between service levels and operational costs. The model itself is quite robust and can be applied for queueing models other than the Erlang C. For an M/M/s model with abandonment, β can even take negative values, as there will be stability at all staffing levels. This formula can also be used for skill based routing models and even for time-varying models.

4.2.2 Single Skill Call Center

Queueing theory is used to arrive at staffing models in call centers. In case of a single skill call center the problem is twofold – the first step is to determine the shifts and the next step is to assign agents to the shifts. Different approaches may be used to arrive at optimal shifts. Some researchers have used a heuristic approach for the problem while others have used linear programming techniques. Along with arriving at shifts, one need to also determine when to schedule break time for agents within each shift. Integer programming has been widely suggested for the preparation of agent rosters. However, practical staff scheduling in a call center involves several additional

constraints such as employee leave and variable call volumes. As a result, software that aims at staff scheduling uses optimization techniques such as constraint satisfaction and local search. In fact splitting the problem into two parts gives a sub-optimal solution, as the availability of agents influences the types of shifts needed. Heuristic models in which shift determination and staff assignment are integrated are also used in call center scenarios. Additional factors that influence the staffing model are the desired occupancy levels for agents and the desired service levels for customers. The maximum shift length as per local regulations is also a consideration in arriving at the staffing model.

4.2.3 Multi-Skill Call Center

Staffing in a multi-skill call center is a complicated task as different agent combinations can be used to meet service requirements. For example, in a two-skill scenario, you may need a certain skill at some shifts while the other skills may be more in demand during the other shift. It may be difficult to arrive at the number of multi-skill agents that you need as opposed to the number of single skill agents. Many different combinations may be able to satisfy the service levels needed and external factors such as skill availability, agent salaries, and labor norms would play a role in deciding the optimal staffing model.

The problem is further complicated if multiple service channels are considered. From a scheduling perspective, the key difference is that low-service level channels such as faxes and e-mails can be responded to by the agents when they are relatively free from the high demand channels such as telephone. While a multi-channel service model reduces costs significantly, the scheduling problem by itself becomes more complex.

4.3 Skill Based Routing

Queueing theory finds many practical applications in skills-based routing in call centers with multi-skilled agents. If each skill requirement in a call center were handled by a dedicated set of agents, then the call center is essentially a group of independent single skill call centers and does not require any modeling. However, this does not take advantage of economies of scale due to the flexibility that will be available if there are multi-skilled agents. The other extreme is to have all agents trained in all skills, but this may have several practical problems. Most call centers would be configured in such a way that a subgroup of agents would be trained on a specific skill and these agent sub groups would overlap. This means that any given agent may possess more than one skill and for any given skill, there will be more than one agent – but not all agents will be trained on all skills. This leads to operational challenges such as how to determine the number of agents needed for each skill, how many permanent and temporary agents should be there and how to schedule them in shifts. While these are planning challenges, there are real time challenges as well, which include how to determine which agent should cater to an incoming call.

Queueing theory-based models have been used to arrive at two selection rules – agent selection, which determines where an incoming call will be routed to; and call selection, which determines which among the waiting calls an idle agent would choose to answer.

Some of the practical ways in which this is done is discussed next:

The PABX or ACD will contain a list of agent groups in decreasing order of priority for each skill. An incoming call would then be assigned to the agent group with an available agent. Typically, a call will flow from one agent group to the next only when all agents in the former group are occupied. On the other hand, when an agent becomes available, some priority rules will be applied to select which among the waiting calls should be serviced.

Another configuration that can be used is to assign the call to an agent group only if there is at least a certain threshold of idle agents in the group. This is especially useful if there are varying priorities of incoming calls, so that agents can be kept idle to wait for a higher priority call while low priority calls wait in the queue to be serviced. However, this is not an optimal solution from a theoretical perspective. You may encounter a situation where the last agent in an agent group with skills A and B is handling a skill B call, whereas there are several agents idle with skill B and skill C in the system. When the next call requiring skill A comes in, the caller would be placed in a queue. In order to avoid this and achieve optimal routing, one must develop a model that considers available agents in all groups before routing the call. This is an intractable dynamic programming problem, as the number of possible configurations is exponential to the number of agent groups. Call centers have resolved this issue using simple structures such as assigning an aging factor to waiting customers and servicing the customer with the largest aging factor.

4.4 Load Balancing in Geographically Dispersed Call Centers

Another area where principles of queueing theory can be put in practice is in the dynamic load balancing between interconnected but geographically dispersed call centers so as to exploit economies of scale. Network ACDs are used to route the calls between the call centers. Two of the strategies used are 1. Centralized FIFO and 2. Distributing a call to the call center with the least expected delay of service. FIFO is a more complex solution and studies have shown that it does not perform optimally if there is a significant delay to switch calls between the centers.

4.5 Call Blending

Queueing theory can be effectively used for call blending in call centers that handle both inbound and outbound calls. Studies show that call centers making use of scientific call blending models are able to achieve high levels of agent productivity as well as customer service levels. This is done by assigning agents to outbound calls every time the number of available agents exceeds a certain threshold.

Call blending is achieved by assuming the call center to be a queueing system with two types of jobs. Inbound calls form the first type of job and have a constraint on performance – the average

waiting time for callers has to be below a specified level. The second type of job is the outbound calls that are available in an infinite amount (theoretically) for which the objective is to maximize the number of calls made i.e. to serve as many customers as possible. For the purpose of arriving at the model, inbound calls are assumed to be a Poisson distribution and the service times of both jobs are assumed to be exponentially distributed and independent of each other. It is also assumed that agents are multi-skilled and can be interchanged between jobs. The model should ideally schedule the agents (known as servers in queueing theory) in such a way as to maximize the throughput of the second type of jobs(outbound calls) while meeting the waiting time constraint on the first type of job (inbound calls). Since one cannot schedule the arrival of an inbound call, the key question is to identify whether an idle agent should be deployed for an outbound call, or kept waiting to handle a potential inbound call. This would primarily depend on how many other agents are free at that point in time. Typically, in a call center that does not use queueing theory-based scheduling, different set of agents are assigned to handle inbound and outbound calls. This affects the productivity of agents, as the number of agents assigned to handle inbound calls would be based on peak volumes and these agents would remain relatively free during non-peak hours. Call blending on the other hand helps to dynamically assign agents to either handle inbound or outbound calls depending on the volume of inbound traffic. If call blending is incorporated in workforce management software for call centers, it will not only help cut costs by improving agent productivity, but will also help to enhance revenue by maximizing the number of outgoing calls and improving customer satisfaction, as no caller would be left waiting beyond the threshold wait time.

Another application area for call blending is call centers that offer multiple service channels such as telephone, fax, website and so on. Each of these channels would need different response times and since responses to emails and faxes can be preempted by telephone calls and then resumed later, a mathematical asymptotic framework of Markovian Service Networks can be used to model this scenario. Customers would be served based on a preemptive-resume priority rule. The primitives of such a network would be time-varying and can easily accommodate abandonment and retrials. This framework can thus be used for large multi-channel call centers and can be used for arriving at performance analysis parameters as well. However, the disadvantage of such a model is that it cannot accommodate non-preemptive priority disciplines or busy signals.

[Brandt and Brandt], have proposed a birth and death queueing model to map a call center with abandonment and an integrated IVR system. In the model, callers who wait are transferred to an IVR queue, once they have been waiting online beyond a threshold. If there are no customers in the waiting line and the number of idle agents crosses a defined threshold, then the callers in the IVR queue are transferred to a live agent. Other studies have established an asymptotic equilibrium for such a model if customers rationally choose between balking, abandoning, or retrial.

5 Limitations of Queueing Theory

While there is no doubt about the advantages of using queueing theory to model the behavior of call centers and to optimize performance and service levels, one cannot rely solely on the assumptions of queueing theory to arrive at optimal solutions for planning call center operations. This is because classical queueing theory assumptions may be too restrictive to accurately model real life scenarios. For example, mathematical models often assume an infinite number of callers, an infinite server capacity, and no restrictions on service times; however, in practice all of these have bounds.

Some of the assumptions in the Erlang C model, which is widely used in call centers, do not hold true for a typical call center. For example, there would typically be a limitation on the number of waiting slots in a call center and thus only a finite number of customers can be kept on hold. However, if the number of slots is quite large or if an IVR, system is used, then there will never be a situation when all slots are busy and customer gets a busy signal. Therefore, the Erlang C model can be a fair approximation.

Call arrival rates also do not follow a Poisson distribution, and queueing theory may not be able to accommodate sudden peaks in call volumes. Service time for each call would also vary, and most queueing models are not robust enough to model calls that cannot be resolved at the first level.

The other limitation of using a queueing model is that it is difficult to predict caller behavior, which does not follow any specific pattern.

The good news, though, is that despite these bounds, the difference between theory and real life is not very significant and most queueing models are robust enough to provide close enough approximations to optimize call centers. Still, call centers often rely on simulation tools to help analyze and optimize performance based on dynamic queueing line behavior. Data analytics of past call data is also heavily relied upon before strategic decisions such as technology upgrades are undertaken.

The world of call centers is an interesting area for the application of queueing models, and simple models are already widely used in workforce management and staff scheduling software. Applications continue to grow, and the use of more complex models to solve the challenges of the modern day call center need to be formulated, analyzed and incorporated into the latest software in order to derive the full benefits.

6 References

1. Reed, J.. "Queueing models for large scale call centers". Ph.D. diss., Georgia Institute of Technology, 2007. In ABI/INFORM Global [database on-line]; available from <http://www.proquest.com> (publication number AAT 3271583; accessed March 28, 2012).
2. Koole, Ger, Mandelbaum Avishai. "Queueing Models of Call Centers: An Introduction." *Annals of Operations Research (Springer Netherlands)* 113, no. 1 (07 2002): 41-59.
3. Whitt, W. (1999). Predicting queueing delays. *Management Science*, 45(6), 870-888. <http://search.proquest.com/docview/213170597?accountid=10559>
4. Mandelbaum, M., & Hlynka, M. (2008). Examples of applications of queueing theory in canada. *INFOR*, 46(4), 247-263. <http://search.proquest.com/docview/228519383?accountid=10559>
5. U, N. B. (1969). Sixty years of queueing theory. *Management Science (Pre-1986)*, 15(6), B280-B280. <http://search.proquest.com/docview/205839819?accountid=10559>
6. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., & al, e. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469), 36-50. <http://search.proquest.com/docview/274789078?accountid=10559>