

From the pages of  
Call Center Management Review



## CCMReview ARTICLE BUNDLE

# How Incoming Call Centers Behave: Back to the Basics

## How Incoming Call Centers Behave: Back to the Basics (Parts 1, 2 and 3)

By Brad Cleveland



Post Office Box 6177 • Annapolis, Maryland 21401

800-672-6177 • 410-267-0700

[www.incoming.com](http://www.incoming.com)

AB00002

# How Incoming Call Centers Behave: Back to the Basics, Part 1 of 3

by Brad Cleveland

*ICMI is dedicated to periodically reviewing the “basics,” the core knowledge call center managers must have in order to manage effectively. This three-part series will review the fundamental, underlying principles that shape incoming call center behavior. It will address such questions as: How are service level and occupancy inter-related? How does agent group size affect productivity? How does staffing affect network costs? These principles will be discussed in the context of recent industry developments.*

Just as the laws of physics define the parameters for air travel, there are fundamental, underlying principles that apply to the call center environment. Inappropriate group configuration, inefficient network designs and inherently-unfair standards for telephone agents are signals that one or more of these principles has been either misunderstood or ignored. This article reviews random call arrival and caller tolerance — key driving forces in incoming call centers — and it identifies important immutable “laws” that shape what occurs within a call center and how they impact staffing and service levels.

## Random Call Arrival

Theorem One of incoming call center management is that “Calls Bunch Up.” For virtually any call center, calls arrive in predictable repeating patterns, by time of day, day of week and season of year. With appropriate interdepartmental coordination and judgment — i.e., what’s the marketing department doing? how is the call mix changing? — fore-

casts for the aggregate workload can be quite accurate, down to specific half-hour increments.

But even with near-perfect forecasts, the actual moment-to-moment arrival of calls within the half hour is a random phenomenon — the luck of the draw. It is ultimately the result of countless individual decisions made by callers, based on a myriad of individual habits and motivations.

**Simulation is the call center equivalent of a flight simulator -- you can learn how to program an intelligent ACD without suffering the consequences of bad design.**

company engineer, assumes random call arrival and that calls will go into queue if a server (agent) is not immediately available.

More specifically, Erlang C assumes “infinite sources of traffic” (infinite trunking capacity — i.e., no busies) and that “lost calls are delayed” (infinite caller patience — i.e., no abandoned calls). The formula has been criticized for its tendency to lead to overstaffing because, Erlang’s findings notwithstanding, busies and lost calls are a reality in many call centers. Fortunately this has become less of a problem with definitions of “acceptable” service levels, which

minimize abandonments and busies.

Erlang C is currently the basis for staffing calculations in virtually all of the workforce management software packages on the market. However, computer simulation has the potential to provide greater flexibility, and to account for issues such as busies, abandonments, skill based routing, overflow, etc. Simulation is the call center equivalent of a flight simulator — you can learn how to program an intelligent ACD without suffering the consequences of bad design.

Currently, a small but growing percentage of call centers is utilizing simulation. While general-purpose simulator software has been available in a variety of formats for years, a handful of vendors provide packages designed specifically for call center modeling, and more are under development.

Either Erlang C or simulation is a stark contrast to a common “wrong” approach to calculating staff — taking the average calls reps can now handle in a half hour or hour and dividing that into projected future traffic volumes.

**Because supply and demand are rarely equal, demand must be “chased” with the supply of answering capabilities.**

This approach ignores both random call arrival and service level. Poor staffing results in either idle reps or a long queue for callers, with the associated repercussions. Further, because staffing impacts the

load the network and system must carry, miscalculated staff can also mean miscalculated system and network resources. In short, calculating staff correctly is essential.

Even with good forecasts and solid staffing calculations, random call arrival

means that call center systems operate in a “demand-chasing” mode — there are either more calls to be answered than resources available, or more resources than calls. Because supply and demand are rarely equal, demand must be “chased” with the supply of answering capabilities. Systems and networks can help by making useful historical and real-time data available on trends and queue dynamics; providing flexibility to respond to changing traffic loads; and enhancing call answering capabilities. Likewise, flexibility in staffing — getting the right people in the right place at the right times — is imperative.

### Seven Factors of Caller Tolerance

Random call arrival, the fact that calls “bunch up,” makes answering every call immediately highly impractical for most call centers. Accordingly, caller tolerance is another important driving force in incoming call centers. How long are callers willing to wait in queue? How many will abandon? Further, caller tolerance affects such things as how callers will react to automation, such as a voice response unit (VRU), and how favorably they perceive the service they get. Caller tolerance is influenced by seven major factors that are not static. They include:

**1. Degree of motivation:** Callers to airlines wait longer during special price promotions than at other times. Callers with power-outages will wait longer to reach their utility than those with billing questions.

**2. Availability of substitutes:** The

World Wide Web, fax, mail, other numbers and other selections in the VRU are examples of potential substitutes that callers try to reach the primary group. If a primary queue backs up, callers may dial other numbers available, choose incorrect routing selections in an automated attendant (press one for this, two for that...)

or even call the company’s main number (switchboard). If callers are highly motivated and have no workable substitutes, they will retry many times if they get busies, and will generally wait a long time in queue.

**3. Competition’s service level:** This factor applies when callers have the alternative of using a competitor’s services.

**4. Level of expectations:** The reputation that an organization or industry has for service (or the level of service being promoted) has a direct bearing on tolerance. For example, callers to catalog companies generally expect comparatively high levels of service, and are much less tolerant of queues than, say, callers to utilities or software support centers.

**5. Time available:** Doctors who call insurance providers are infamous for being intolerant of even modest waits. Retirees calling the same companies may have time to chat.

**6. Who’s paying for the call:** In general, callers are more tolerant of a queue when the call is free to them.

**7. Human behavior:** The weather, the caller’s mood and the day’s news all have some bearing on caller tolerance.

Some call centers have discovered that telling callers how long the queue is can lower abandonment, generally if the

queue is no more than two or three minutes. When the wait is unknown, a caller may wait 90 seconds in queue and hang up in frustration, when he or she would have been cut through to an agent in 95 seconds. In the mid-80s, Word Perfect pioneered the use of “queue jockeys” to announce queue times to incoming callers, play music, deliver informational announcements, and generally entertain callers on hold. A handful of others companies have followed Word Perfect’s example.

Although this approach is impractical for most, a small but growing number of call centers are using an automated approach, whereby the ACD looks at real time data, calculates the expected wait and relays that information to callers via intelligent system announcements. You need an ACD that has this capability. You also need a reasonably straightforward environment; if you’re using some form of complex contingency-based routing (such as skills-based routing), the ACD generally cannot predict wait times accurately.

In the future, visible queues will be common. In fact, we’ll probably look back on these pioneering ‘70’s, ‘80’s and ‘90’s and smile at how little we knew as callers. With multimedia technologies, the queue will most likely be represented graphically on the caller’s computer or video screen.

These driving forces behind us, the next articles in this series will cover such issues as staffing, service level, occupancy, group size, system configuration and the fundamentals of getting the right people and supporting resources in the right places at the right times. 

*This article was originally published in April 1997.*

# How Incoming Call Centers Behave: Back to the Basics, Part 2 of 3

by Brad Cleveland

In any inbound call center, there are predictable, fundamental laws at work. And just as you would need to understand such principles as gravity, lift and velocity in order to design or operate an aircraft, it's imperative to understand the laws that shape how call centers behave in order to manage effectively. Part 1 of this series reviewed random call arrival and caller tolerance, important "driving forces" in incoming call centers. This article will introduce key "laws of call center nature" and discuss their importance in planning and management decisions.

## Occupancy, Service Level and Group Size

Service level is expressed as "X percent of all calls answered in Y seconds." Occupancy is the percent of time during a half-hour that agents who are on the phones are either in talk time or after-call work (wrap-up). The inverse of occupancy is the time agents spend waiting for calls, plugged in and available.

Tables 1 and 2, which are based on Erlang C calculations, illustrate the relationship between service level and occupancy: for a given call load, occupancy goes up when service level goes down. In Table 1, for service level of 80/20, occupancy is 78 percent; if service level drops to 14/20, occupancy goes up to 97 percent. What's the reason for this inverse relationship? If occupancy is high, it's because the agents on the phones are taking call after call after call, with little or no wait between calls. In other words, calls are backed up in queue and service level is low. In the worst scenario, occupancy is 100 percent for a long stretch of

time because service level is so low that all callers spend at least some time in queue.

**When adherence to schedule improves (goes up), service level will get better and occupancy will drop.**

The size of the agent group also affects occupancy. At comparable levels of service, a large airline reservation center will have higher occupancy than a small call center serving a regional insurance company. Table

1 shows that 15 agents are required to handle 100 calls at a service level of 80/20; agent occupancy is at 78 percent. In Table 2, 40 agents are required to handle a call load three times as large and at the same service level, with occupancy at 88 percent.

Occupancy cannot be directly controlled. That can be a tough case to make in the budgeting process. The reality, though, is that the time agents spend waiting for calls — nine seconds here, 18

seconds there, two seconds there — is a necessary part of a good service level and is driven by how calls are arriving. Sure, even with very good forecasts and schedules, there will be times when you either have too many or too few agents; it only makes sense for them to do other activities when there is time to do so. But don't expect them to get other work done and still meet your service level objective if, according to Erlang C or computer simulation, you have no more than the minimum staff required to handle the call load.

At the individual level, standards on number of calls handled are usually inherently unworkable or unfair. Agents can't control occupancy, and those assigned to larger groups, busier shifts or shifts that have lower service levels, will naturally have the opportunity to handle more calls. On the other hand, agents should be responsible for their, "adherence factor," or how well they adhere to their schedules. The terms occupancy

Table 1

| Talk Time: 180 sec; Work Time: 30 sec; Calls: 100 |                    |     |      |               |
|---------------------------------------------------|--------------------|-----|------|---------------|
| Agents                                            | SL %<br>In 20 Sec. | ASA | Occ. | Trunk<br>Load |
| 12                                                | 14%                | 561 | 97%  | 41.2          |
| 13                                                | 46%                | 97  | 90%  | 15.4          |
| 14                                                | 67%                | 37  | 83%  | 12.1          |
| 15                                                | 80%                | 17  | 78%  | 10.9          |
| 16                                                | 89%                | 8   | 73%  | 10.5          |
| 17                                                | 94%                | 4   | 69%  | 10.2          |
| 18                                                | 97%                | 2   | 65%  | 10.1          |
| 19                                                | 98%                | 1   | 61%  | 10.1          |
| 20                                                | 99%                | 0   | 58%  | 10.0          |

Table 2

| Talk Time: 180 sec; Work Time: 30 sec; Calls: 300 |                    |     |      |               |
|---------------------------------------------------|--------------------|-----|------|---------------|
| Agents                                            | SL %<br>In 20 Sec. | ASA | Occ. | Trunk<br>Load |
| 36                                                | 26%                | 171 | 97%  | 58.4          |
| 37                                                | 46%                | 68  | 95%  | 41.4          |
| 38                                                | 61%                | 36  | 92%  | 36.0          |
| 39                                                | 72%                | 21  | 90%  | 33.6          |
| 40                                                | 81%                | 13  | 88%  | 32.2          |
| 41                                                | 86%                | 8   | 85%  | 31.4          |
| 42                                                | 91%                | 5   | 83%  | 30.9          |
| 43                                                | 94%                | 4   | 81%  | 30.6          |
| 44                                                | 96%                | 2   | 80%  | 30.4          |
| 45                                                | 97%                | 1   | 78%  | 30.2          |
| 46                                                | 98%                | 1   | 76%  | 30.2          |

and adherence factor are often incorrectly used interchangeably. But they have very different meanings and actually have an inverse relationship — when adherence to schedule improves (goes up), service level will get better and occupancy will drop.

As anyone who's handled calls knows, extended periods of high occupancy are stressful. Studies generally conclude that agents begin to burn out when occupancy is higher than around 90 to 92 percent. Unfortunately, occupancy tends to feed on itself. When it's high to begin with, agents need and will often increasingly take breaks from the action. They may sign-off or, more subtly, stretch out talk time and/or after-call work time, which will cause

service level to drop and occupancy to go higher still. (If you are running into this problem, see the suggestions in the third column on this page.)

### The Law of Diminishing Returns

The law of diminishing returns says, "When successive individual telephone agents are assigned to a given call load, marginal improvements in service level that can be attributed to each additional agent will eventually decline." For example, Table 2 shows that 36 agents will provide a service level

of 26 percent answer in 20 seconds with an ASA (average speed of answer) of 171 seconds. With just one more agent, service level jumps to 46/20 and ASA drops

to 68 seconds — a quantum improvement. Adding one more person yields another big improvement. But keep adding staff, and the returns begin to diminish. At some point, the cost of adding additional staff outweighs the small improvements in service that they would bring.

Call centers that struggle with a low service level ought to like this law — it often doesn't take a lot of resources to improve things dramatically. On the other hand, those who want to be the "best of the best" in terms of service level find it takes a real commitment in staffing. That is why many call centers have target service levels such as 80/20 or 90/20 versus 100/20 or 100/0.

If you find yourself short-staffed, you'll notice that delay grows exponentially and occupancy quickly becomes high. Here are three proven strategies for avoiding these problems:

- Ensure that your staffing calculations are as accurate as possible. If service level is volatile throughout the day and week, or frequently below your objective, the fix may go to the fundamentals of managing a call center — a good forecast and schedules that better match staff to the workload.

- Make every agent aware of how much they contribute, even if they are tempted to feel like just "one of many." Explain service level and show them the call load patterns so they know how important schedule adherence is.

- Provide real-time queue information to supervisors and agents so they can adjust their activities according to real-time conditions.

In Part 3 of this series, we'll pick up with these and other immutable laws and further discuss their implications on planning and management. CCMReview

*This article was originally published in May 1997.*

**At some point, the cost of  
adding additional staff  
outweighs the small  
improvements in service that  
they would bring.**

# How Incoming Call Centers Behave: Back to the Basics, Part 3 of 3

by Brad Cleveland

There are predictable, fundamental laws at work in an incoming center. In order to manage effectively, it's imperative to understand how these laws shape how call centers behave. Part 1 of this series reviewed random call arrival and caller tolerance, important "driving forces" in incoming call centers. Part 2 discussed the relationship between service level and agent occupancy, and the impact of staffing levels on service level and agent occupancy. In Part 3, the last in this series, we'll look at how staffing levels impact the load on the trunks (and therefore, network costs), and discuss the management decisions surrounding the "powerful pooling principle."

## Agents Impact Trunk Load

When more telephone agents are assigned to handle a given call load, the load that the trunks (telephone lines) must handle goes down. The converse is also true: when fewer agents are available to handle a given call load, trunk load goes up because delay increases (see Figure 1). Consider checking in for a flight — fewer agents at the counter

means longer lines. In a call center, each person waiting in queue requires a trunk. You can see this relationship in the last column of Table 1, which gives the load to be carried by the trunks, expressed in erlangs (hours) — (talk time + average speed of answer) X number of calls in an hour (calculations are based on Erlang C). To determine ACD system capacity required, the following inputs are necessary:

- Calling load, which includes the number of calls the ACD will be required to handle at peak capacity, average talk time, and average after call wrap-up. Note, since wrap-up does not show up on trunking reports, trunk reports alone are insufficient to predict ACD capacity requirements.

- Service level objective, or "X percent of calls to be answered in Y seconds."

- Agent scheduling factors that accommodate absenteeism, lunch, breaks, training and other realities that will keep agents from the phones. Call center managers should know these numbers — often factors of 1.2 to 1.5 (for exam-

ple, if Monday's 9:00 am factor is 1.3 and 20 agents need to be "plugged in," 26 agents will need to be scheduled, or  $20 \times 1.3$ ). For more information, see *Service Level Notes*, November 1995.

Once the above inputs are used to calculate staff, trunks can be configured to carry the load (plus any VRU load not reflected in delay, etc.). Since delay is a function of staffing, staffing must be calculated before trunking. Further, the costs on the network are directly related to staffing levels (for information on calculating these costs, see *Service Level Notes*, February 1996). The major message from this immutable law is that staffing, system, and network resources must be planned and calculated in sync because they are inextricably associated. (For more information on calculating trunks, see *SLN's Q&A* column, December 1992).

## The Powerful Pooling Principle

The pooling principle is a mathematical fact, based on the laws of probability. It states: any movement in the direction of consolidation of resources will result in improved traffic-carrying efficiency. Conversely, any movement away from consolidation of resources will result in reduced traffic-carrying efficiency. Note the efficiencies illustrated in Table 2; for example, one combined group of 15 agents can handle the same call load at the same service level as two groups of nine agents.

To the degree you can combine smaller groups of agents into larger groups without increasing average handling time, you can A) handle more calls, at the same service level, with the same number of agents, B) handle same number of calls, at the same service level, with fewer agents, or C) handle the same number of calls, at a better service level,

Figure 1

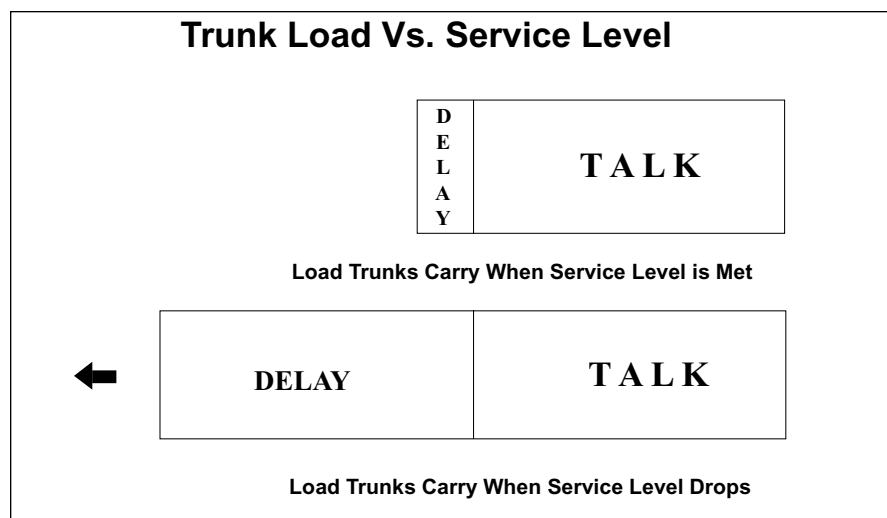


Table 1

| Talk Time: 180 sec; Work Time: 30 sec; Calls: 300 |                    |     |      |               |
|---------------------------------------------------|--------------------|-----|------|---------------|
| Agents                                            | SL %<br>In 20 Sec. | ASA | Occ. | Trunk<br>Load |
| 36                                                | 26%                | 171 | 97%  | 58.4          |
| 37                                                | 46%                | 68  | 95%  | 41.4          |
| 38                                                | 61%                | 36  | 92%  | 36.0          |
| 39                                                | 72%                | 21  | 90%  | 33.6          |
| 40                                                | 81%                | 13  | 88%  | 32.2          |
| 41                                                | 86%                | 8   | 85%  | 31.4          |
| 42                                                | 91%                | 5   | 83%  | 30.9          |
| 43                                                | 94%                | 4   | 81%  | 30.6          |
| 44                                                | 96%                | 2   | 80%  | 30.4          |
| 45                                                | 97%                | 1   | 78%  | 30.2          |
| 46                                                | 98%                | 1   | 76%  | 30.2          |

Table 2

| Calls | SL             | Agents Req. | Occ. |
|-------|----------------|-------------|------|
| 25    | 80% in 20 sec. | 5           | 58%  |
| 50    | 80% in 20 sec. | 9           | 65%  |
| 100   | 80% in 20 sec. | 15          | 78%  |
| 500   | 80% in 20 sec. | 65          | 90%  |

\*Assumption: Calls last 3.5 minutes.

with the same number of agents.

The pooling principle is a consideration at the highest levels of strategic planning (i.e., call center consolidation or networking multiple sites) down to more specific decisions about how far to proceed with cross-training. In one sense, pooling resources is at the heart of what ACDs do. In fact, when ACDs first came into the market in the early 1970s, the big challenge was to get users to abandon the “cliente” approach, or the need to reach specific individuals.

Further, geography no longer matters. In some cases, agents can work out of their homes if the proper telecommuting environment is established. And, networked-ACDs are virtually a “must-do” for organizations with multiple call centers.

Intelligent call processing capabilities in modern ACDs provide the means to bring diverse resources and skills together at just the right time. For example, skill-based routing enables the skills of each agent (i.e., knowledge of products or services or languages spoken) to be

defined and identified to the ACD. Then, specific types of incoming calls can be matched with specific skills — assuming a good planning and management process is in place so the right agents are in fact available at the right times.

But it’s important to remember that the most efficient environment would be one where any call could be handled by any agent. Further, if capabilities such as skill-based routing are implemented poorly, the number of contingencies can multiply beyond the call center’s ability to manage them, reducing efficiencies and causing poor service. One thing is certain — as real and pervasive as the pooling principle is, it is not an all-or-nothing proposition. There is a continuum between pooling and specialization. Call centers should specialize when it clearly is necessary — i.e., for different languages or significantly different product lines — but should also look for cross-training opportunities wherever practical.

### In a Nutshell

This series reviewed random call arrival and caller tolerance, important “driving forces” in incoming call centers. It then summarized important immutable laws:

- When service level goes up, occupancy goes down (at a given call load).
- With more staff, delay goes down and therefore trunk load goes down (at a given call load).
- The law of “diminishing returns.”
- The “powerful pooling principle.”
- Larger groups have higher occupancy (at a given service level).

These immutable laws shape the way call centers behave, and understanding them is a prerequisite to understanding the call center environment and managing it effectively. CCMReview

*This article was originally published in June 1997.*

## About the Author



Brad Cleveland is President and CEO of Incoming Calls Management Institute (ICMI), and Publisher of Call Center Management Review. Recognized for his pioneering work in management and leadership, he has delivered keynotes, seminars and consulting services in over 25 countries. Brad is coauthor of the popular book, Call Center Management on Fast Forward, and is coeditor of ICMI's influential handbook/study guide series of publications.

# About Incoming Calls Management Institute

Incoming Calls Management Institute (ICMI), based in Annapolis, Maryland, offers the most comprehensive educational resources available for call center (contact center, interaction center, help desk) management professionals. ICMI's focus is helping individuals and organizations understand the dynamics of today's customer contact environment in order to improve performance and achieve superior business results. From the world's first seminar on incoming call center management, to the first conference on call center/Internet integration and subsequent research on multichannel integration, ICMI is a recognized global leader. Quality, usability and value have become trademarks of ICMI's award-winning services. ICMI is independent and is not associated with, owned or subsidized by any industry supplier; ICMI's only source of funding is from those who use its services.

## **ICMI's services include:**

- Public and onsite (private) seminars
- Web seminars and e-learning courses
- Certification review seminars and study guides
- Industry studies and research papers
- Consulting services
- Software tools for scheduling and analysis
- Books (including the industry's best-selling book, *Call Center Management on Fast Forward*)
- *QueueTips*, the popular (and free) monthly e-newsletter
- Membership in Incoming Calls Management Institute
- *Call Center Management Review*, the authoritative monthly journal for ICMI members

For more information and to join a network of call center leaders, see [www.incoming.com](http://www.incoming.com).



Incoming Calls Management Institute  
Post Office Box 6177  
Annapolis, Maryland 21401  
410-267-0700 • 800-672-6177  
[icmi@incoming.com](mailto:icmi@incoming.com)  
[www.incoming.com](http://www.incoming.com)