Module 2 – The Driving Forces of Contact Centers

2.1 Three Driving Forces of Contact Centers

Key Points

- **There are three driving forces of contact centers:**
  - **Workload Arrival Types: Smooth, Random or Peaked**
  - **The Visible or Invisible Queue**
  - **The Seven Factors of Customer Tolerance**

Explanation

Contact centers operate in a unique environment. The workload does change from moment to moment. And when customers don't know how long the queue is, they often become impatient much more quickly than in settings where they can "see" the line and the progress they are making.

In any center that handles contacts initiated by customers (versus only outbound), three major driving forces are at work:

- Workload arrival types: smooth, random or peaked
- The visible or invisible queue
- The seven factors of customer tolerance

These driving forces help explain why the contact center environment is so unique.

Understanding them is a prerequisite to making good decisions on everything from staffing and scheduling requirements to establishing the right performance objectives.

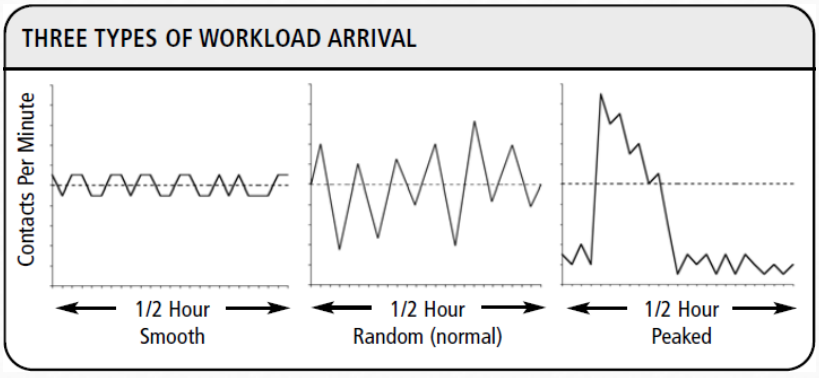
2.2 Workload Arrival Types: Random, Smooth and Peaked

Key Points

- **There are three types of workload traffic arrival that represent how contacts arrive moment by moment:**
  - ***Random traffic* represents the moment-by-moment ebb and flow of contacts typical in inbound contact centers.**
  - ***Smooth traffic* is when contacts arrive in an even flow, which is virtually nonexistent in incoming contact centers, but can apply in outbound environments.**
  - ***Peaked traffic* is a surge of contacts beyond random variation within a half-hour.**
- **The traffic arrival type determines the formulas used for staffing calculations, impacts performance standards and dictates which real-time strategies make the most sense.**

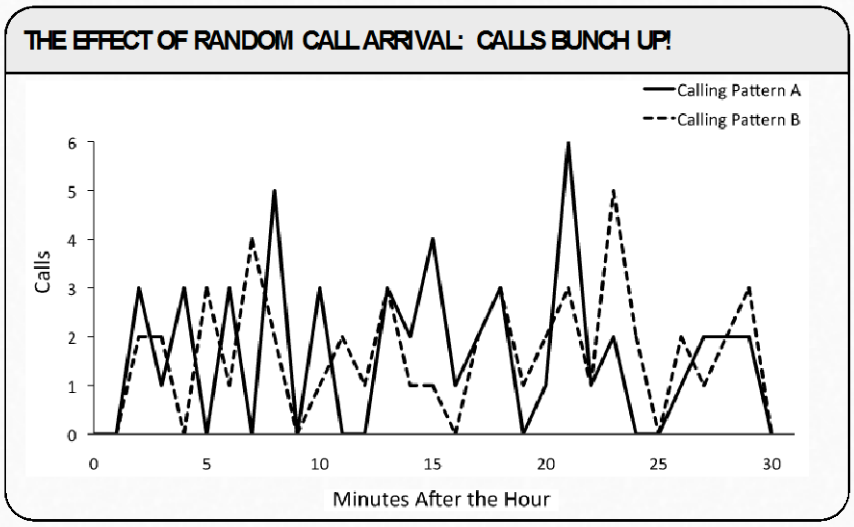
Explanation

To correctly calculate staffing needs, management must determine whether the type of traffic arrival will be random, smooth or peaked. An understanding of contact arrival is also essential to setting the right performance standards and managing effectively in real-time. Telecommunications traffic engineers have assigned statistical "variance-to-mean" ratios to designate each type of traffic, but, essentially, the patterns look like this:



Random (Normal) Arrival

Customer contacts arrive randomly in most contact centers most of the time. Take a look at a monitor or a readerboard on the wall. Watch the dynamics. In comes a call. Then one, two more ... there's another. And two, three, four more ... Exactly when calls arrive from moment to moment is the result of decisions made by customers who are motivated by a myriad of individual needs and conditions. Put another way, calls bunch up!



There is an important distinction between random call arrival and predictable call arrival patterns. Virtually all centers — even those of the more volatile type, such as emergency services centers — have distinctive calling patterns, which are usually detectable down to at least a half- hour. For example, it can be predicted that the center will receive around 240 calls next Tuesday between 11:00 a.m. and 11:30 a.m. What can't be predicted with any precision is how many of those calls are going to arrive in the first minute, the second minute and so forth. There are several important implications to random call arrival:

- First, staffing must be calculated by using either a queuing formula that takes random arrival into account or a computer simulation program that accurately models this phenomenon. Other approaches almost always lead to inaccurate staffing calculations. And unfortunately, it's not just staffing that will be off. Because staffing affects the load the network and systems must carry, miscalculated staff inherently leads to miscalculated system and network resources.

- Second, centers that handle customer-initiated (inbound) contacts operate in a "demand-chasing" environment. At any given time, there are either more calls than staff to handle them or more staff than calls. That means contact centers must augment good forecasting and

staffing plans with real-time management. A solid understanding of random arrival is necessary to avoid overreacting to normal variation in traffic arrival and underreacting to bona fide trends.

- Third, performance objectives and standards must take random workload arrival into account. For example, a standard of "N widgets per day" makes no sense in an environment where the workload arrives randomly. Unless the queue is always backed up and service is lousy, your agents will spend a portion of their day just waiting for calls to arrive.

Smooth Traffic

Smooth traffic is virtually nonexistent in centers handling incoming contacts, but can apply in outbound environments. For example, a group of people may be assigned to make outbound calls (e.g., for surveys, political interests, non-profit donations, etc.), one after another, for the duration of their shift. In that case, staffing requirements can be based on a units-of-output approach common in many manufacturing and service settings, and the number of trunks (telecommunications circuits) required will be equal to the number of agents placing the calls.

Peaked Traffic

Another type of call arrival — peaked traffic — is a reality in some centers. The term "peak," in a general sense, can refer to contact center workload: What's your peak time of year? Peak day of the week? Peak time of day? But the term "peaked traffic" specifically refers to a surge of traffic beyond random variation. It is a spike within a short period of time.

Television and radio ads will often generate peaked traffic. For example, QVC, Inc. gets a surge of calls when new products are advertised on its home-shopping channel. Disaster relief and humanitarian organizations, such as the Red Cross and World Vision, get peaked traffic when their television or radio ads are aired, as do mobile phone providers when they send SMS messages to a large portion of their subscribers. And because of the proliferation of always-on Internet services, email promotions that are sent in batches can also generate an initial surge of contacts (as well as traffic that will arrive in subsequent hours and days). The (typically) large centers that handle peaked traffic can go from zero to hundreds of contacts a minute, almost instantly.

It is important to correctly distinguish between random and peaked traffic. When catalog companies send out thousands of new catalogs by mail or send email offers to their lists, they begin getting calls associated with the promotion. But that's not peaked call arrival. It's random arrival, but at a much higher level than recent history. Similarly, a utility that has a power outage will get a lot of calls until the problem is fixed. But other than the few minutes following the outage, calls will arrive randomly, albeit at a much higher level than usual.

The key question is this: Is there a surge of calls that come and go within less than a half hour? For staffing purposes, if the surge lasts longer than a half hour, it's probably random call arrival.

The implications of peaked call arrival on resource requirements include:

- Staffing must be calculated at a smaller interval than half-hour, such as 10 minutes or even five minutes.
- For a given service level, peaked traffic requires more staff than random traffic.
- How concentrated peaked traffic is within a brief period of time will dramatically impact service level.

2.3 The Visible or Invisible Queue

Key Points

- **How much callers know about their place and progress in a queue impacts their perceptions and behaviors.**
- **Callers who wait longer than they feel is appropriate often voice their complaints to agents; this takes time, drives up average handling time and further backs up the queue.**
- **Contact centers can offer visible queues to their customers through the use of predictive wait announcements that announce expected wait times to individual callers. These predictions are generally accurate in straightforward queues, especially large agent groups.**
- **Callers who abandon a visible queue usually do so immediately after hearing the prediction; those who decide to remain generally do not abandon before reaching an agent.**

Explanation

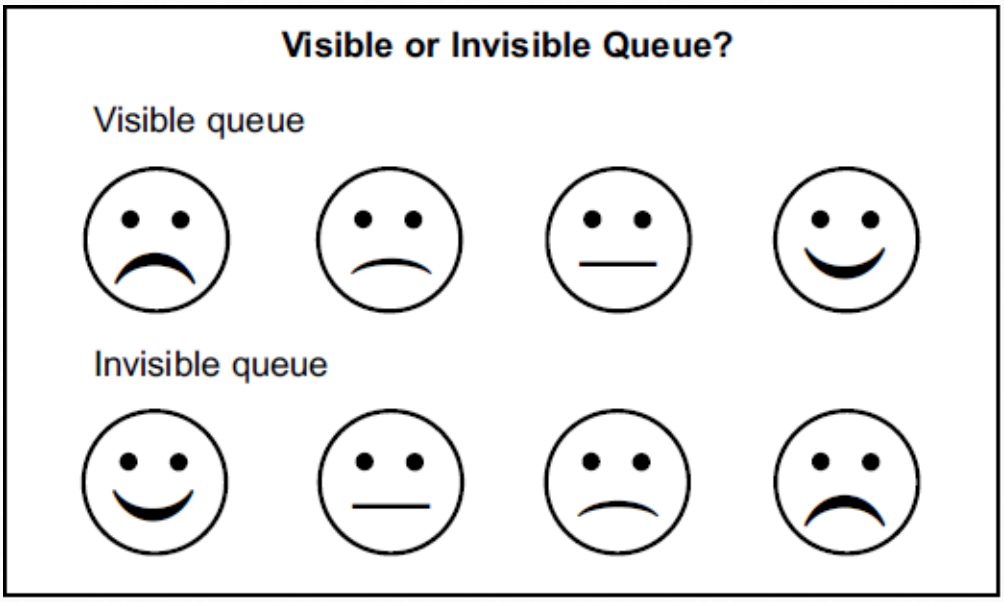
Queue comes from the word cue, a term from Old French that means "line of waiting people." It is common in everyday British English (less so in North America, where "line" is typical) and appears frequently in contact center terminology.

Queues are a fact of life in contact centers. After all, answering every call at once would be about as practical for many centers as it would be for airlines to check in every passenger at one time. But an important difference between a contact center and the lines at an airline counter, grocery store or sports arena is that callers usually can't see how long the queue is and the progress they are making in it.

The top row of faces in the figure Visible or Invisible Queue, reflects a queue that the customers can see. Few would choose to wait in line so, as they enter the queue, the first face represents them. As

they move forward, the subsequent faces illustrate their progress. The final face reflects the fact that they "made it." They are at the counter, hearing the sweet words, How may I help you?

The second row of faces represents a setting where customers are ignorant of the queue they are entering. "Ignorance is bliss," and expectations are initially high. But after some amount of waiting, say 10 or 15 seconds of ringing, they begin to doubt that they are going to get right through (second face). The third face illustrates the transition from doubt to mild frustration. By now, they have probably heard the first delay announcement and it confirms that they are in a queue.



The fourth face represents callers who, from their perspective, have waited too long. Often, the first thing they do when they reach an agent is tell him or her about the miserable experience they just had. That's a bad situation because it lengthens call-handling time, which will back up the queue even more (and may cause even more callers to unload on agents once their calls are answered).

Callers who have waited a long time in queue also tend to "dig in their heels" as they attempt to squeeze all the value out of the call that they can. Customers seeking technical support are a classic example: "Geesh, I waited this long to get through, I'd better go over a few more things while I have you."

Turning the invisible queue into a visible queue by providing predictive wait time announcements may cause some callers to hang up immediately after hearing the message, if wait times are long, but those who stay in queue are more likely to remain until they get an answer. Estimated wait time calculations can be very accurate under the following conditions:

- The agent group has more than 20 agents assigned to it. Smaller agent groups result in a wider diversity of wait times experienced by callers.

- The agent group design isn't too complex. Environments that use skills-based routing, for example, have routing plans that are too complicated to give specific callers accurate wait times.
- The handling time for calls in the group is clustered closely around the average handling time for the group. For example, if a technical support center handles calls that vary widely in length from 3 minutes to 90 minutes, estimated wait time calculations will not be very accurate.

Visible Queue

Software company WordPerfect (now part of Corel Corporation) pioneered the "visible queue" in the mid-1980s. They set up their system to enable live "queue jockeys" to make announcements of expected hold times to incoming callers. They could also play music and deliver announcements to keep callers entertained and informed while they waited in queue:

Thank you for calling WordPerfect. If you're calling for assistance with Version 2, there are nine of you in queue, and if you just joined us, it looks like the wait is just over three minutes. If you are calling for Version 3, there are 18 callers in queue. But we've got more staff there this morning, and it looks like your wait will be about two minutes. Now, here's Kenny G, from his latest album ...

WordPerfect discovered that callers who abandon a visible queue do so at the beginning. Callers who decide to wait generally do so until they reach an agent.

Many contact center managers keep a diligent eye on how many customers abandon (give up on a phone call, chat, or similar, while waiting for an agent). But when callers abandon is an important consideration, as well. If they are abandoning early on because they are making an informed choice, that's a different story than waiting for what seems like forever in an invisible queue, then hanging up in frustration.

What the queue jockey never said, but what was implicit in the message, was something like:

Thanks for calling. If you're going to abandon, would you kindly do so now, before you get frustrated, drive up our costs and clog up the queue only to abandon before we get to you anyway?

Other software companies, including Microsoft, soon followed WordPerfect's lead. The feedback from callers was overwhelmingly positive. But having real-time, live queue jockeys is impractical for most organizations. Accordingly, ACDs that could "tell time" began to appear in the early 1990s. With this technology, the ACD can analyze real-time variables, make predictions and announce expected wait times to callers as they arrive. (Today's virtual queue capabilities go a step further, enabling callers to hang up and receive a callback when an agent becomes available or at a later time.)

It's a great feature, but there's a catch. These systems provide fairly accurate predictions in reasonably straightforward environments, especially in large agent groups. However, if you are using any form of complex, contingency-based routing, the system can outsmart itself. Some callers have found themselves actually moving backward in queue as arriving priority callers are moved to the front of the line. If predictions aren't accurate in your environment, you won't be able to use them. This is a challenge that system designers have yet to fully conquer.

There has been quite a bit of debate and study over the years around the question of how callers react to visible versus invisible queues. But I think that the discussion sometimes misses the real point: Given the choice, customers want to know what's happening, and they want alternatives (e.g., to continue to wait or receive a callback). There will come a time when we'll look back on the days when we, as callers, entered queues we knew nothing about. Predicted wait times might even be graphically displayed on mobile apps — and we'll be given choices when the queue is busy.

Excerpt from *Call Center Management on Fast Forward* Third Edition by Brad Cleveland, ICMI, 2012.

2.4 Factors Affecting Customer Tolerance

Key Points

- **There are seven primary factors that affect customer tolerance:**
  1. **Degree of motivation**
  2. **Availability of substitutes**
  3. **Competition's service level**
  4. **Level of expectations**
  5. **Time available**
  6. **Who's paying for the call**
  7. **Human behavior**
- **These factors influence such things as:**
  - **How long callers will wait in queue**
  - **How many callers will abandon**
  - **How many will retry when they get busy signals**
  - **How they will react to automation, such as IVR or Web services**
  - **How they perceive the service the contact center is providing**

Explanation

The seven factors of customer tolerance include:

1. Degree of motivation: How motivated are customers? For example, those experiencing a power outage will usually wait longer to reach their utility than catalog customers placing an order for merchandise.

2. Availability of substitutes: Are there substitutes customers can use (e.g., Web or IVR services) if they cannot get through to the initial number? If they are highly motivated and have no substitutes, they will retry many times if they get busies and generally will wait a long time in queue, if necessary.

3. Competition's service level: If it's easier for callers to use competitive services, they may go elsewhere.

4. Level of expectations: The experiences callers have had with the contact center and the reputation that the organization or industry has for service (or the level of service being promoted) have a direct impact on tolerance. A 10-minute wait for tax help during filing season may be perceived as quite acceptable — but a similar wait for a shipping company that otherwise has a reputation for speedy service would be an unpleasant surprise.

5. Time available: How much time do customers have at the time of the contact? Doctors who call insurance providers have a well-deserved reputation for not tolerating even a modest wait, while retirees calling the same companies may have more time or inclination to talk. Further, the widespread use of mobile phones has created many small windows of time customers have to reach the center, such as before boarding a flight or in between meetings, when long waits are unworkable and frustrating.

6. Who's paying for the call: In general, customers are more tolerant of a queue when toll-free service is available. They are less tolerant of a queue when they are paying toll charges.

7. Human behavior: The weather, the customer's mood, the time of day and other human behavior factors all have a bearing on customer tolerance.

These seven factors are not static. They are constantly changing, making abandonment extremely difficult to predict. Even so, it is important to have a general understanding of the factors affecting your customers' tolerance. Important questions to consider include:

- How motivated are your customers?
- How much does that vary based on the reason for the contact?
- What are their expectations?
- Do they vary based on the reason for the contact?
- What type of customer is least motivated?
- What type of customer is most motivated?
- What alternatives to contacting you do they have?
- Which alternatives would you want/not want them to use?
- Will they already have spent time seeking help (e.g., self-service, search, customer community, etc.)?
- What level of service are others in the industry providing?
- How do customers tend to rate your service at different service levels?
- What impact does meeting/not meeting expectations have on your brand?
- What is their tendency to share good or bad experiences through social channels?

The answers to these questions will enable you to better understand customer behavior and establish services that meet their needs and expectations.

2.5 The Planning and Management Process

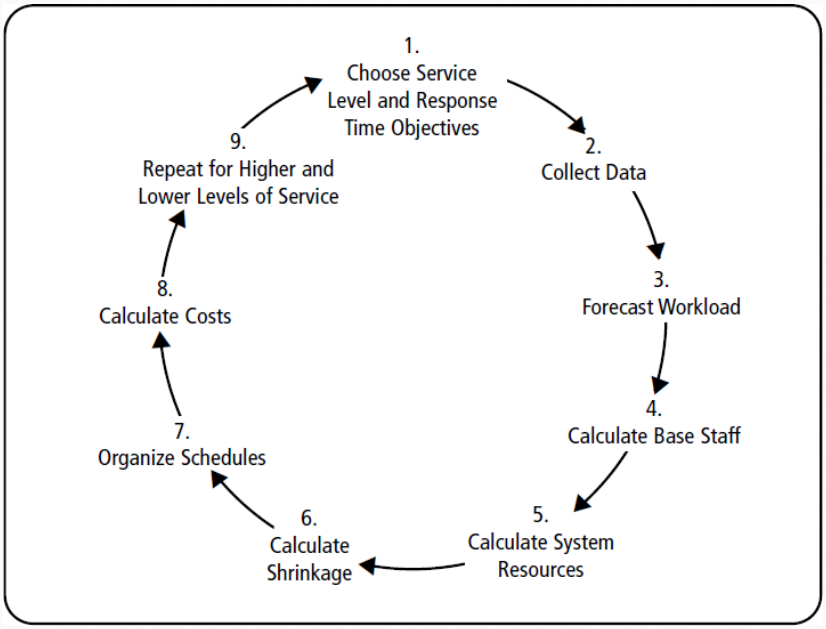
Key Points

- **An effective planning and management process includes these basic steps:**
  1. **Choose service level and response time objectives**
  2. **Collect data**
  3. **Forecast the workload**
  4. **Calculate base staff**
  5. **Calculate system resources**
  6. **Calculate shrinkage**
  7. **Organize schedules**
  8. **Calculate costs**
  9. **Repeat for higher and lower levels of service**

- **A collaborative effort among contact center personnel and managers in other departments is necessary to ensure accurate planning.**

Explanation

To deliver consistent, high levels of service, a contact center requires a systematic planning and management process. This process can be summarized in nine steps:



Step 1: Choose Service Level and Response Time Objectives

Service level and response time objectives (accessibility objectives) are at the heart of effective contact center management. These objectives are essential in defining staffing and network requirements, and associated costs.

Step 2: Collect Data

Planning and managing a contact center requires information from many places. Today's contact center systems are important sources of planning data that reveal much about the interactions being handled. But much of the information required comes from beyond the contact center's walls — e.g., what marketing is doing, how customer preferences are changing, conversations and trends in social channels, competitive activity that may have an impact on the workload, and relevant developments in the economy.

Step 3: Forecast the Workload

Workload forecasts must include each of the components of customer contacts: average talk time, average after-call work (wrap-up), and volume. (Or just handling time and volume for contacts that don't have separate talk and wrap-up requirements, such as social or email). A good forecast predicts these components accurately for future time periods, usually down to a half hour. Forecasts should encompass all types of contacts — phone, email, chat, social interactions, SMS, et al., and plans should also account for any related work that will require contact center resources.

Step 4: Calculate Base Staff

Most developed contact centers use Erlang C or variations of it to calculate staffing requirements. Erlang C is the base formula used in virtually all workforce management systems. But capabilities such as skills-based routing and complex network environments present challenges that may require computer simulation and modeling.

Step 5: Calculate System Resources

Staffing and system resource issues are inextricably associated and must be calculated together.

Steps 6 and 7: Calculate Shrinkage and Organize Schedules

Rostered staff factor and shrinkage take into account breaks, absenteeism, training, work not directly part of handling customer interactions, and all of the other considerations that occupy agent time.

Schedules are essentially forecasts of who needs to be where and when, and plans of action for agents and supervisors. They should lead to getting the right people in the right places at the right times.

Steps 8 and 9: Calculate Costs, and Repeat for Higher and Lower Levels of Service

These final steps in the process involve projecting costs for the resources required and preparing budgets. Preparing budgets around different levels of service provides an understanding of cost trade-offs, which is invaluable in budgeting decisions.