Excerpts from:

Contact Center Management on Fast Forward: Succeeding in the New Era of Customer Experience.

4th Edition

By Brad Cleveland

These excerpts, Chapters 6, 7 and 10, are for ICMI conference attendees. Please do not distribute or duplicate.

Contact Center Management on Fast Forward

Succeeding in the New Era of Customer Experience

Brad Cleveland

Fourth Edition

The #1 best seller on contact center management — trusted by thousands of organizations — now with new information on customer experience, omnichannel, quality, performance metrics, and much more!

Table of Contents

CHAPTER 6:

Forecasting Customer Workload

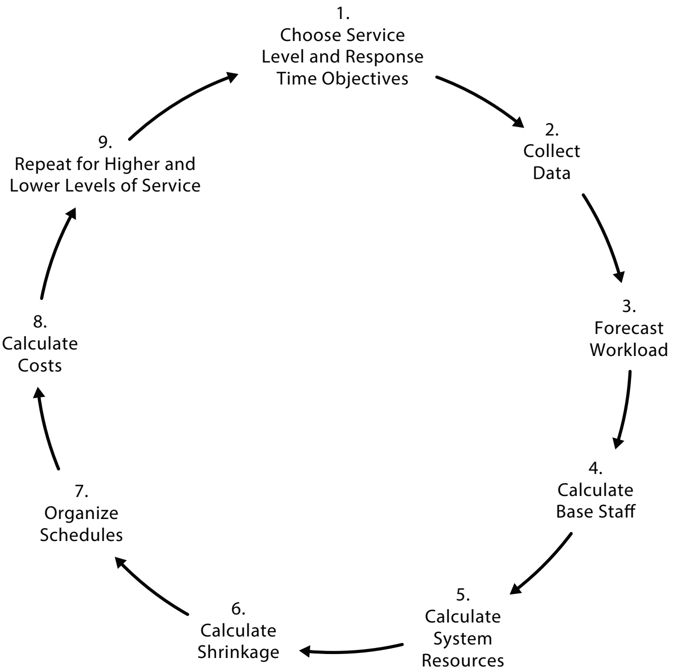
"My interest is in the future because I am going to spend the rest of my life there."
C.F. KETTERING

Matching resources with the workload is a critical step in managing a contact center effectively. This responsibility goes to the heart of contact center management, as defined in the first chapter: "having the right number of properly skilled people and supporting resources in place at the right times to *handle an accurately forecasted workload*, at service level and with quality."

Here's the scoop: If the forecast–step 3 in the planning and management process–is not reasonably accurate, the rest of the planning process will be off the mark. The forecast is the basis for determining staffing needs as well as system capacity, supervisors, analysts, workstations, and other resources. It provides the foundation for:

- Calculating base staff required to meet your service level and response time objectives
- Calculating trunking and system requirements

- Minimizing abandoned and blocked contacts
- Creating accurate, workable schedules
- Predicting staffing and network costs
- Meeting customer expectations
- Establishing an environment in which quality service can be provided



Key Terms and Essential Data

Forecasting is a mix of art and science. It begins with predicting how many contacts you are going to get in a future period, usually 12 months. To do that, you first look at historical data to determine patterns that reflect when people contact you, and you consider possible trends that will affect patterns going forward.

You then take that information and break it into contacts by months, weeks of the month, days of the week, and half hours of the day–or even five minutes of the half hour if you are forecasting peaked traffic. Next, you factor in the handling times of the interactions. Finally, you modify results based on conditions that are not reflected in the historical data.

Weekly, daily, and intraday forecasts are short-term tactical forecasts used to tighten up schedules and adjust priorities around current conditions and near-term events. Shorter-term forecasts project workload for the upcoming three months. They are necessary for organizing and adjusting scheduling requirements, anticipating seasonal staffing needs, planning for holidays, and determining imminent hiring requirements.

Longer-term forecasts project workload for a year and beyond. They are used to estimate future annual budgets, establish long-term hiring plans, and define future system needs.

How far out you forecast will depend on the purpose of the forecast. Regardless, the basic terms and concepts are similar.

Agent Group

Agent groups are the building blocks of a contact center. An agent group (also called a split, gate, queue, or skills group) shares a common set of skills and knowledge, handles a specified mix of contacts (e.g., service, sales, technical support) and/or channels (e.g., phone, chat, social media, email). and may comprise a handful of agents or hundreds of agents across multiple sites. Supervisory groups and teams are often subsets of agent groups.

Your forecasts will be built around agent groups. If you have one group of 100 agents handling all contacts, you'll have one forecast and one set of schedules. If you have 10 groups of around 10 agents, you'll need 10 forecasts and schedules–one for each unique agent group. In other words, planning must be specific enough to ensure that you get the right number of properly skilled people and supporting resources in place at the right times, for each agent group.

Workload

The basic historical data you need to forecast for an agent group includes how many contacts you have received in the past, when they arrived and how long they took to handle. Four key terms reflect this activity:

- **TALK TIME** is everything from “hello” to “goodbye.” In other words, it’s the time customers are connected with agents. Anything that happens during talk time–such as putting the customer on hold to confer with a supervisor, research an issue or make an outbound call–should be included in this measurement.

- **AFTER-CALL WORK** (also referred to as ACW, wrap-up or not ready) is the time agents spend completing contacts after saying goodbye to customers. Legitimate after-call work should immediately follow talk time.

- **AVERAGE HANDLING TIME (AHT)** is average talk time plus average after-call work.

- **WORKLOAD** (also referred to as call load) is the volume of contacts coupled with how long they last. The formula is: volume × average handling time, for a given period of time.

While these terms are typical and make sense for calls, they can vary for other channels. For example, “talk time” and “after-call work” are fine for video. They don’t make sense for email or text, where “handling time” is a better fit. In an omnichannel environment, the channels you handle will dictate the terms that best describe workload components.

> *In an omnichannel environment, the channels you handle will dictate the terms that best describe workload components.*

For chat or social media, the series of exchanges with a customer over a short period of time (the back-and-forth that takes place in many of these interactions) is referred to as a session (chat) or conversation (social media). You'll need to anticipate the number of customer sessions or conversations, the time it takes to handle them and the number of customers an agent can simultaneously handle. We'll discuss variations as we consider other channels, in this chapter and in Chapter 7.

Clean the Data!

It's important to "clean" the data you use to forecast workload. Don't make the mistake of taking existing data from systems and using it "as is" without giving it a second thought. What if the IVR was down for an hour? What if a reporting system malfunctioned for part of a morning? What if a breaking news story impacted workload?

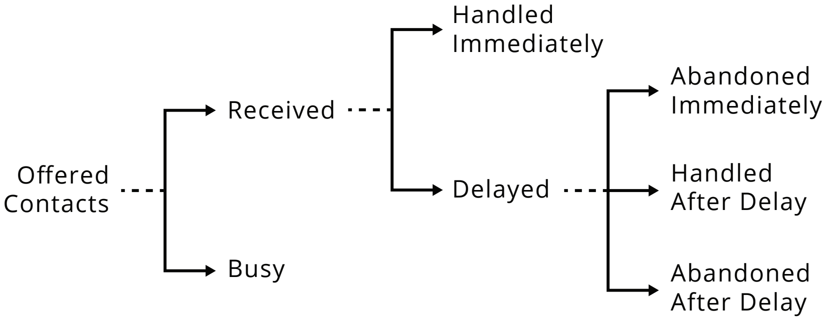
I remember working with a large contact center in Australia during the Melbourne Cup—the thoroughbred horse race that is one of the most-watched sporting events of the year. Just before the race, contacts plummeted, then resumed in a matter of minutes after its conclusion. It was fascinating to see the race on one monitor and follow real-time contact center traffic on another.

What about unique variables in your environment that will, or will not, repeat and that need to be either included in or backed out of the numbers? Making adjustments that reflect what's likely to continue will ensure that you are building forecasts on a solid foundation.

Offered Contacts

Offered contacts include all of the attempts your customers make to reach you. There are three possibilities for offered contacts: they can get a busy signal or message (rare in most contact centers); they can be received by the system but abandon before reaching an agent; or they can successfully reach an agent. In forecasting, what you really want to know is the number of *individuals* attempting to reach you.

Acquiring data on abandoned contacts is usually straightforward. Most routing systems provide reports on abandonments down to specific increments of time.



Many managers count abandoned contacts "one-for-one." But to the degree that customers who abandon try again later and get through to agents, they will be counted more than once. Consequently, you may want to "normalize" or discount them. If available, it's helpful to have hard data, which can come from automatic number identification (ANI), calling line identification (CLI) reports, or other systems.

Without good data, you run the risk of discounting too deeply, which will lead to forecasts that underestimate demand. I generally recommend that you include most (70 percent or more) abandoned contacts in the data, unless you have solid information that tells you otherwise. This may lead to forecasts that overstate demand; however, forecasts that underestimate the workload will result in insufficient staffing and abandoned contacts, which perpetuate the problem.

Similarly, busy signals will need to be discounted in the data you use for forecasting. Busies are far less common today than they once were (and I rarely recommend intentionally using them). But they still can and do happen. When they do, they wreak havoc on reports. The age-old question is, for every 100 busy signals, was that 100 people who tried to reach you once, or one persistent soul who tried 100 times? Of course, the answer usually is somewhere in between.

If your ACD can dynamically generate busy signals based on real-time circumstances, it will likely provide a report on how many customers received busies. Telecommunications network carriers can also provide reports that can help solve the retry mystery. Alternatives to retry reports can include customer surveys, answering all contacts for a short period (even if by voicemail) to determine true demand, and judgment (guessing). Naturally, it's best to have hard data.

In sum, your forecast should, as accurately as possible, reflect the *number of individuals* attempting to reach you. If you count every abandoned contact or busy signal in the data you use, the forecast will overestimate true demand. If you ignore busy signals and abandoned contacts, the forecast will underestimate demand.

IVRs and Routing Contingencies

Many organizations use interactive voice response units (IVRs) to provide customers with self-service options and to help route contacts ("press or say one" or "tell us the reason for your call"). Additionally, contingency-based routing alternatives can mean that contacts start out in one place and end up in another depending on real-time circumstances.

I always feel most comfortable when I know how the systems are configured and how contacts reach their destinations. (I'll often have someone walk me through a step-by-step flowchart so that I understand.) You'll need a forecast specific to each agent group for staffing purposes. In other words, count and forecast the contacts intended for each individual agent group so that your forecasts and staff predictions are accurate going forward.

Proportions

The fundamental information you need for forecasting includes the three components of workload: talk time, after-call work, and volume (or their variations for other contact channels). From this data, proportions can be derived. If you received 1,000 contacts for the day, and 60 came in between 10:00 and 10:30, that half hour's proportion would be 6 percent or .06 (60/1,000). Proportions are used to project patterns into the future.

This is information your ACD and/or workforce management system should be collecting now and forever. Building on this essential half-hour data, you will accumulate necessary daily, weekly, and monthly data. Whatever you do, don't throw data away. You'll never know when you will need information from several years ago. "Hey, how did customers react that time we ..."

	Contacts	Prop.	Average Talk Time	Average Work Time	Average Hndl.Time
08:00-08:30					
08:30-09:00					
09;00-09:30					
09:30-10:00					
10:00-10::30					
10:30-11:00					
11:00-11:30					
11:30-12:00					
12:00-12:30					
12:30-13:00					
13:00-13:30					
13:30-14:00					
14:00-14:30					
14:30-15:00					
15:00-15:30					
15:30-16:00					
16:00-16:30					
16:30-17:00					
17:00-17:30					
17:30-18:00					
Totals/Avgs					

Basic Quantitative Information Necessary for Forecasting

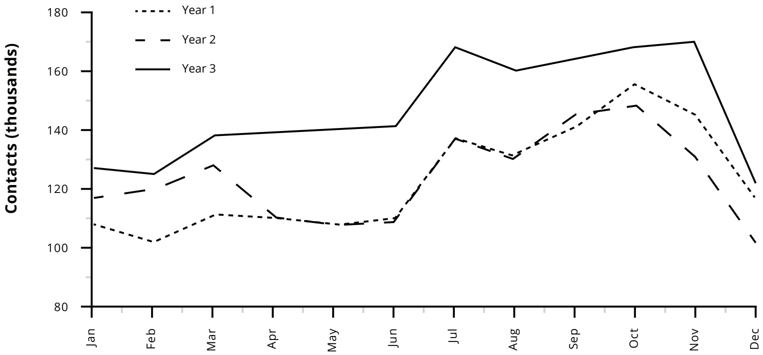
Repeating Patterns

Virtually all centers handling customer-initiated contacts notice at least three dominant patterns.

MONTH OF YEAR OR SEASONALITY.

The graph, Monthly Contacts Offered, illustrates data from a financial services company. Notice that the most recent year is at a higher plane, but looks similar to the patterns in previous years. Even if your organization is going through dramatic changes, you usually will detect seasonality in your contact arrival patterns. Three years of data will provide a good reading on these patterns; if you have additional history, even better. If you don't have three years of data, use what you have; even one year will often reflect seasonality that is likely to continue.

Monthly Contacts Offered

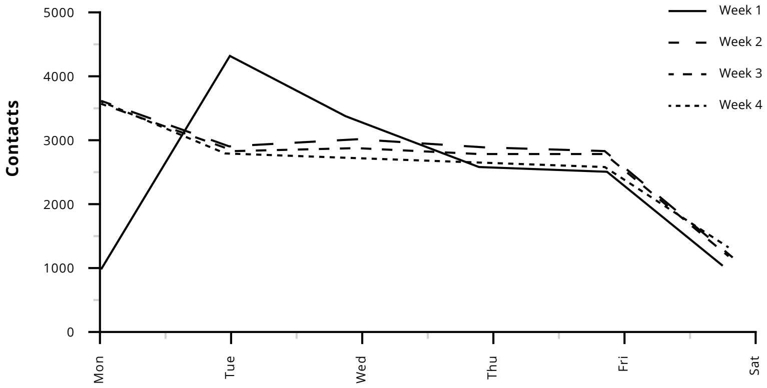


DAY OF WEEK.

The graph, Contacts by Day of Week, is from a communications company (telecommunications and Internet). The first week reflects a holiday on a Monday. The contact center was open, but, of course, customers were behaving differently than usual. Consequently, Tuesday gets more contacts than normal, illustrating the "pent-up demand" that is common after holidays.

Otherwise, the pattern is highly predictable from one week to the next. (Even holiday weeks are predictable, if you have some history of similar holidays.) As the example shows, as few as four or five weeks' worth of history can reveal this pattern.

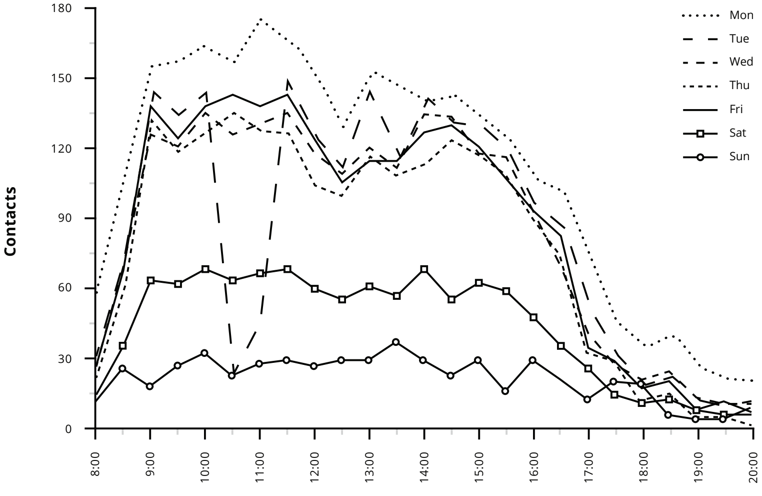
Contacts By Day of Week



HALF HOUR OF DAY.

The data for the graph Half-Hourly Contacts Offered is from a bank. Notice the system outage? That kind of exception from the norm tends to really stick out. And it raises an important point: Exceptions need to be adjusted (smoothed over or normalized) or they will throw off predictions. Data for just a week or two is often enough to identify this pattern.

Half-Hourly: Contacts Offered



You may see other patterns. For example, if you send out statements to your customers on the 5th and 20th of each month, you'll notice day-of-month patterns. And marketing campaigns will create their own patterns.

Individuals contact your organization for myriad reasons, but (and I find this fascinating) they become part of highly predictable patterns. It's pretty amazing, actually.

So, one of the most essential steps in forecasting is to look at your data and identify the patterns that exist. Even if you are using forecasting software, it is still important to graph the "raw" patterns so you can identify exceptions.

Breaking Down a Forecast

Okay, grab that double tall latte, and let's go through a basic approach that illustrates how to break down a forecast. This example starts with longer-term patterns and works its way down to specific half-hour increments. The steps involved include:

1. Obtain the number of contacts received in the past 12 months (720,000 in this example).

2. Multiply the year's contacts by 1.12 to reflect 12 percent expected growth. Factoring in growth at this level assumes that contacts will increase proportionally to previous years' patterns. If growth will instead be concentrated around marketing campaigns or other events that don't necessarily happen at the same time from year to year, you should factor it in at a more specific level, such as monthly or weekly.

3. Multiply the estimated contacts in the year you are forecasting by January's proportion, 7.1 percent. This percentage comes from historical data and is the typical proportion of the year's contacts received in January.

Breaking Down a Forecast

720,000	Current year's contacts
× 1.12	add 12% (proportion)
806,400	Forecasted annual contacts
× .071	January proportion
57,254	January contacts
÷ 31	Operation days - January
1,847	Average contacts per day
× 1.47	Monday's index factor
2,715	Monday's contacts
× .055	10:00 to 10:30 proportion
149	Forecasted contacts 10:00-10:30

Notes:

1. Determine operation days by counting the days the contact center will be open.

2. Calculate day-of-week index factors by multiplying day-of-week proportion by days open.

JANUARY

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Example:	DOW Prop.		Days Open		Index Factor
Monday	.210	×	7	=	1.47
Tuesday	.170	×	7	=	1.19
Wednesday	.165	×	7	=	1.155
Thursday	.165	×	7	=	1.155
Friday	.150	×	7	=	1.05
Saturday	.095	×	7	=	.665
Sunday	.045	×	7	=	.315

4. Divide the number of operation days in the month into the estimated monthly contacts. This yields average contacts per day. In this example, the center is open every day of the month.

5. Adjust average contacts per day, using the appropriate daily index factor. The first column in the index factor calculation (lower right corner) gives the proportion of the week's contacts that typically arrive each day. For example, Monday normally gets 21 percent of the week's traffic; Tuesday gets 17 percent, and so forth.

The next column reflects the number of days in a week that the contact center is operating. If you're open seven days a week, use 7. If five days, use 5. Even if you're only open part of a day (e.g., a half day on Saturday), count that day.

The final column is the result of multiplying the first column by the second column. These index factors are then multiplied against the average contacts per day to estimate traffic by the specific day of the week. In this example, Monday's index factor, 1.47, is multiplied against 1,847.

6. The final step is to multiply the predicted contacts for each day of the week by each half hour's proportion. In this example, the half hour 10:00 to 10:30 will get a projected 149 contacts.

This process must take a lot of time, right? Actually, once you establish an approach, it won't take nearly as much time as you may think. You will get better at it with practice. And forecasting software, or even simple spreadsheets, can take much of the labor out of it. And the time you do invest is well worth it. Remember that forecasting is one of the most high-leverage activities in the planning process. You'll spend a lot more time "putting out fires" later on if you don't have a good forecast.

This is a basic approach and there are many possible refinements—to account for calendar variations, intra-month trends, or other variables—that may improve accuracy. But if you are pulling out the exceptions and working with good data, going through this process provides a good foundation on which to build. You will still need to blend in judgment, coordinate with marketing, etc. (see upcoming discussions). After all, past history doesn't always reflect what's going to happen in the future.

You may also need to incorporate other patterns into the forecast. For example, if you send out billing statements twice a month, that activity will generate traffic when the bills begin to arrive. But the percent increase caused by these events will also fall into predictable patterns, and you can adjust accordingly. You may need to calculate day-of-month index factors, a process similar to deriving day-of-week index factors.

Holiday Weeks

Holiday weeks will require their own index factors. But the pattern for one week with a holiday on a Monday will often be similar to another week in the year with a holiday on a Monday. Holidays that fall on various days of the week are another reason to hang on to your historical data.

Examples of Calculating Day-of-Week Index Factors for Week With A Holiday

	Prop.	Days in Week	Index Factor
Monday	0	0	0
Tuesday	.290	6	1.74
Wednesday	.240	6	1.44
Thursday	.175	6	1.05
Friday	.155	6	0.93
Saturday	.095	6	0.57
Sunday	.045	6	0.27

Intraday Forecasts

Intraday or intraweek forecasts are quick and easy to produce, and are often quite accurate. Typically, short-term forecasts are more accurate than long-term forecasts.

The approach works like this: at some point in the morning, say just after 10:30 a.m., you begin to realize that this is not a typical day. Your reports indicate that you have received 402 contacts so far, which may be more or fewer than originally expected. Either way, you divide the usual proportion of the day's contacts that you would expect by 10:30–18 percent in this case–into 402 (18 percent came from looking at traffic patterns on previous days and calculating half-hourly proportions). Bingo, you now know that if the trend continues, you can expect to receive 2,233 contacts for the day.

Next, you can break down the revised daily forecast into the remaining half hours by multiplying historical half-hourly proportions by 2,233. For

Intraday Forecasting

402	Contacts received by 10:30 a.m.
÷ .18	Usual proportion of contacts by 10:30 a.m.
2,233	Revised forecast for day
× .066	3:30 - 4:00 p.m. proportion
147	Intraday forecast for 3:30 - 4:00 p.m.

example, since you would normally expect to get 6.6 percent (.066) of a day's contacts between 3:30 and 4:00 p.m., you can expect 147 calls during that half hour.

The assumption behind intraday forecasting is that the morning will set the tone for the afternoon. However, if you are a utility getting swamped with contacts in the morning due to a major power outage, this is a bad assumption. When the outage is fixed, the contacts will go away. In many cases, though, intraday forecasting is a useful and accurate tool. You can use similar logic to create an intraweek forecast.

Intraweek Forecasting

3,050	Contacts received on Monday
÷ .23	Usual proportion of contacts by Monday
13,261	Revised contacts forecast for week
× .17	Friday's proportion
2,254	Intraweek forecast for Friday

Direct Marketing Campaigns

Most organizations that run direct marketing campaigns will look at typical response rates to help forecast workloads. Here, you work with your marketing colleagues to gauge the size of the target audience and expected response rates that involve the contact center. (Some orders may be all self-service, some through retail stores—what you're looking for is the response rate that will involve contact center agents.)

Usually there is a taper-down effect, where volume is relatively high in the initial days of a campaign and then decreases over time. One of the things that makes this tricky is that there are often overlapping campaigns going on at any given time. Another is deciding what constitutes an order—is it a single contact from a customer or each item ordered? So, you and your marketing team will need to decide on definitions and stick to them so that you have a solid baseline to work from.

Direct Marketing

A. Target Audience Size ____

B. Overall Response Rate (orders ÷ target audience) ___________________

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14
C. Percent Orders by Day	__	__	__	__	__	__	__	__	__	__	__	__	__	__
D. Projected Orders (A × B × C)	__	__	__	__	__	__	__	__	__	__	__	__	__	__
E. Conversion Factor 1 ÷ (orders ÷ contacts)	__	__	__	__	__	__	__	__	__	__	__	__	__	__
F. Number of Contacts (D × E)	__	__	__	__	__	__	__	__	__	__	__	__	__	__

Peaked Arrival

In many ways, underlying forecasting principles are similar for both random and peaked traffic. But one big difference is in the level of detail required in your reports. While half-hour reports are sufficient for random workload arrival, you will need historical reports down to more specific increments of time (five- or 10-minute segments) in order to adequately forecast and staff for peaked traffic.

The other big difference is that the specific targeted promotions you are delivering–television or radio ads, for example–will dramatically influence response. That's true for any kind of marketing, of course, but because of concentrated workload arrivals, the impact on contact center resources is that much greater. The audience you are reaching (numbers and demographics), the products and services being offered, the channels you are making available, and the effectiveness of the ads themselves will drive how many contacts you get, the nature of the interactions, and how concentrated they are within a specific timeframe. Identifying primary drivers and correlating them with response is key.

You'll also want to take the steps you can to influence when the work will come in so that you're ready to handle it. You may have a choice between running television ads at precise times, or receiving discounts to run them anytime within a larger block of time (as the network's programming needs dictate). Ad-buy savings can quickly evaporate when you weigh the costs of having staff standing by. These decisions should be made collaboratively (including marketing, the contact center and the media partner), with the total picture in mind.

Social Media

Social media monitoring tools—which range from free to high-end and can be standalone or integrated into existing contact center systems—enable you to make sense of what's being said about your company, products, and services. They can help you identify customer sentiment, influencers, where conversations are taking place, and which are having the most impact on the perception of your brand. And some social media posts are clearly customer service contacts—individuals who need and expect a response.

Armed with this information, you can establish criteria for when and how to engage and determine what needs to be included in the workload that is delivered to your agents. As part of forecasting, you'll want to divide the workload into categories, including one-to-many responses where anyone can see (and potentially benefit from) the post, versus direct messages to individuals. Further categories might include known and unknown issues—any grouping has distinct handling time requirements.

If you're launching a new channel and don't have as much to go by, think more along the lines of forecasting the weather: Partly cloudy this morning, with a warming trend this afternoon. You won't always get it right. Look out as far as possible, think through as many variables as feasible, and observe patterns and how they are developing. In the end, dress for any kind of weather (meaning, build flexibility and scalability into your staffing plans).

The good news is that underlying patterns almost always exist. Yes, social media trending topics and posts that quickly multiply can create unique staffing challenges. But being responsive early can help head off what would be repetitive contacts. The same basic principles apply: look for patterns, consider the variables, and use what you're seeing to project future workload. (See Pioneer Forecasts, below).

PIONEER FORECASTS

New products, marketing campaigns, support channels or accounts may not have much, or any, history you can use for forecasting. AI and self-service initiatives can change patterns that existed in the past. Even adjusting hours of operation can leave you with unknowns.

So, where do you begin? Todd Hixson—who has led workforce management initiatives for Hulu, Intuit, Travelocity and other organizations that have gone through explosive growth and change—offers sound advice. He says, "These are pioneer forecasts, and in every case, you need a volume and average handling time set." Here are some examples:

Data from scratch. Sometimes, thinking out of the box yields surprisingly accurate data. For several weeks, a bank launching its first contact center had tellers manually track interactions with hash marks on simple paper forms divided by time of day. They also had them use a simple stopwatch to record samples of handling times. They looked at market size and clues to seasonality in volume and AHT. Finally, they pieced these variables together into a surprisingly accurate workload forecast for the new contact center.

Rapid growth model. In some cases, rapid growth and product changes will negate the value of historical patterns. Because new customers contact your organization more often, one option is to look at contact rates by customer tenure, combined with anticipated growth. Begin with the current month, broken down by the tenure and contact rate. Next, look at projected growth, and work with your marketing team to predict tenure—down to future months or (even better) weeks. "You'll also need to estimate the impact of new initiatives, such as product improvements," recommends Hixson. (See rate adjustment column.)

Deflection. "Eliminating or deflecting customer contacts through AI, self-help, live communities or product improvement can create havoc in

Rapid Growth Forecast

May, 3 Year Forecast

	May-'17			May-'18			May-'19			
Tenure	Customer Base	Contact Rate	Contacts	Customer Base	Contact Rate	Contacts	Customer Base	Contact Rate	Rate Adjustment	Contacts
0-1 month	10,000	12.0%	1,200	14,000	12.0%	1,680	22,500	12.0%	-2.0%	2,250
1-2 month	7,900	7.0%	553	9,900	7.0%	693	15,500	7.0%	-1.5%	853
2-3 month	6,900	4.0%	276	7,400	4.0%	296	12,600	4.0%	-0.9%	391
4-6 month	15,000	3.3%	495	15,800	3.3%	521	14,200	3.3%	-0.7%	369
7-9 month	12,480	2.9%	362	13,750	2.9%	399	16,800	2.9%	-0.5%	403
10-12 month	11,500	2.7%	311	22,500	2.7%	608	14,900	2.7%	-0.3%	358
1 to 2 year	40,000	1.8%	720	42,000	1.8%	756	61,000	1.8%	-0.5%	793
> 2 year	84,000	0.8%	672	96,000	0.8%	768	108,000	0.8%	-0.2%	654
Total Forecast			**4,588**			**5,721**				**6,070**

historical trends," says Hixson. "Be sure to partner with those working on these initiatives and build deflection estimates into your forecasts. And don't forget that automation tends to offload the easiest contacts, leaving you with a higher AHT for those that remain. You'll need to adjust both data sets."

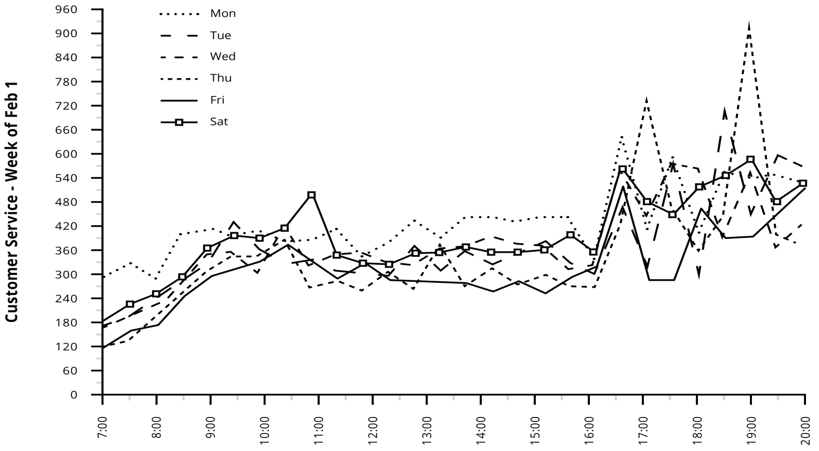
For any pioneer forecast, "create the data sets with a scientific approach in mind," adds Hixson. "Get an agent's point of view, then launch, learn and tune until you achieve a realistic historical sample. Then standard forecasting methodology can take over."

Average Handling Time

Many of us have a habit of referring to the volume of contacts as the only criterion in the workload: "How many contacts did you handle last year? How about this morning?" Equally important, though, is average handling time, which, when coupled with volume, makes up workload. It is workload that matters. Volume alone is relatively meaningless.

As with contact volume, average handling time will fall into predictable, repeating patterns. Similarly, the basic forecasting approach involves utilizing historical reports along with a measure of good judgment. You begin by looking at the average handling time for a recent week, broken down by increments. If the week is "typical," the data represented by this pattern is what will likely continue.

Average Handling Time, Week of Feb 6



The graph of average handling time is from a mobile phone company. Their average handling time went up in the evenings, and was far more variable, for several reasons. First, they let agents bid on shifts based on seniority. Most agents, when given the choice, prefer to start and end earlier in the day, so they had a higher concentration of new agents (and, probably, less experienced supervisors) assigned to the evening shift. That's not necessarily a bad approach, but it will impact average handling time and must be reflected in the forecast.

Second, they did not have a good definition of or process for after-call work, and much of it was getting postponed until late in the day. Third, the mix of contact types changed throughout the day and contacts got relatively longer in the evening.

Average handling time, like call volume, must be incorporated into planning by interval (e.g., half hour). Assuming the same average handling time all day for forecasting purposes will not reflect the environment accurately.

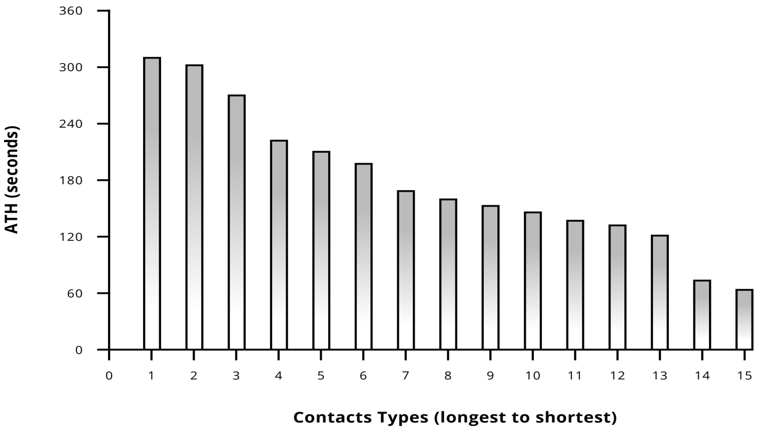
Some relatively simple analysis can go a long way toward tightening up your projections. Here are a few important tips for getting this part of your forecast right:

1. LOOK FOR PATTERNS.
For each agent group, identify how average talk time and average after-call work vary. You may also discover patterns by day of the week, season of the year, billing cycles or marketing campaigns. For contact channels that have both, make separate graphs for average talk time and average after-call work. This will reveal the patterns for each.

2. TRAIN YOUR AGENTS TO USE WORK MODES CONSISTENTLY.
Each agent has an impact on the components of handling time (talk time and after-call work) and on the data that will be used in forecasting and planning for future workloads. When the queue is building, it can be tempting to postpone some after-call work that should be done at the time of the contact. This skews reports, causes planning problems, and may lead to increased errors. An important and ongoing training issue is to define ahead of time what type of work should follow contacts and what type of work can wait.

3. IDENTIFY THE AVERAGE HANDLING TIME FOR DIFFERENT CONTACT TYPES.
This assumes that you have defined and categorized contacts by type, that you are accurately tracking contacts based on the categories, and that you have the reporting capability to link average handling time to the categories. A Pareto chart is often the best way to represent this data.

You can use this information in a number of ways. For example, when you are forecasting an increase or decrease of a specific type of contact, you will be able to project the impact on average handling time. A marketing campaign will generate certain types of contacts. Launching a new web-based service, AI capabilities or mobile app will likely reduce some types of contacts agents handle and may increase others. In each case, you'll be equipped to estimate average handling time.



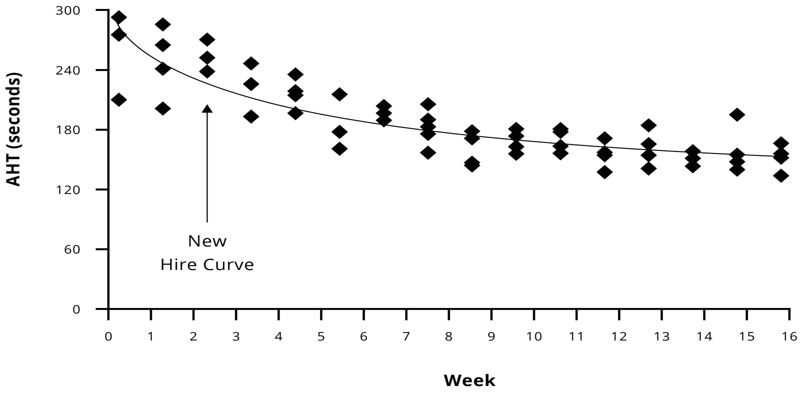
Average Handling Time: By Type of Contact

4. ASSESS THE IMPACT OF NEW AGENTS, LANGUAGES HANDLED, AND PROCESS CHANGES.

Less-experienced agents almost always require more time to handle contacts as they learn how to deal with processes, systems, cultures, and customers. Process changes may either add to or decrease average handling time, depending on the specifics. Additionally, some languages require more time than others (for example, French takes somewhat longer than German, and Spanish requires more time than English).

Compare average handling time with the experience levels of your agents, languages handled, etc. Doing this will enable you to estimate the impact of these variables on AHT, and will be useful in establishing realistic expectations.

Average Handling Time: By New Agents



These steps are recommended, regardless of contact channels. For example, look for patterns in the social media contacts you are handling, ensure agents are using work modes consistently, identify the handling time for different kinds of contacts, and assess the impact of new agents and process changes on handling time.

Beyond the Basics

The major categories of quantitative forecasting include “time-series” and “explanatory” approaches. The examples in this chapter are fairly basic, and more advanced alternatives within each category exist. I’ll introduce a few of them here to give you an idea of the breadth of possibilities.

Time-series forecasting methods include simple or “naive” rules (e.g., the forecast equals last year’s same month, plus 12 percent), decomposition, simple time-series, and advanced time-series methods. The governing assumption behind time-series forecasting is that past data reflects trends that will continue into the future. Time-series methodologies are common in workforce management software. Most time-series forecasts are reasonably accurate when projecting out three months or less.

Explanatory forecasting methods include simple regression analysis, multiple regression analysis, econometric models, and multivariate methods. Explanatory forecasting essentially attempts to reveal a connection between two or more variables. For example, if you manage an ice cream shop, you could statistically correlate the weather (e.g., outside temperature) to ice cream sales. In a contact center, you might correlate a price increase with the impact on contact volumes.

Advanced time-series and explanatory forecasting methods go beyond the scope of this book. In fact, you can spend a couple of college semesters–make that a career–learning about forecasting. For more information, there are many courses (some of them contact center-specific), books, software tools and other resources available.

SOFTWARE AND SERVICES FOR ADVANCED FORECASTING

If you're in an organization or industry going through significant changes, or if your center handles contacts generated by television commercials or overlapping direct marketing campaigns, commonly-used time-series and explanatory techniques might not cut it. Fortunately, there are alternatives.

Increasingly, today's more advanced workforce optimization systems—as well as standalone business forecasting packages from sources such as SAS Forecast Server, Forecast Pro, and others—include a wide range of forecasting methodologies. These programs enable you to build in multiple events and variables and produce forecasts based on different models.

By comparing forecasts, you can identify the best methodology for your environment. Because one approach might work better than others for specific types of events, you can change models as circumstances dictate. You may need an analyst to spearhead this effort, but more accurate forecasts (and therefore, better staffing plans and schedules) often provide a solid return on the investment.

You may also want to explore outside services. For example, forecasting services offered by consultants can provide forecasts based on advanced methodologies for centers that need transitional or ongoing help. Similar services exist in other fields (e.g., financial and sales forecasting), and this alternative has become common and accessible in the contact center environment.

Blending in Judgment

So far, we've looked at quantitative forecasting–in other words, how to use hard data in your forecasting process. Judgmental forecasting goes beyond purely statistical techniques and encompasses what people *think* is going to happen. It is in the realm of intuition, interdepartmental committees, market research and executive opinion.

Many things, from organizational politics to personal agendas, can influence judgmental forecasting. However, some judgment is inherent in virtually all forms of forecasting. And a degree of good judgment can significantly improve accuracy. The trick is to combine quantitative and judgmental approaches effectively, and to be aware of the limitations of each.

The worksheet “Blending in Judgment” illustrates one way of applying common sense and a logical approach to judgmental forecasting. In a customer service environment, the number of contacts is often primarily a function of the total number of customers in the organization's universe. It is possible to project contacts based on historical data, utilizing the relationships between contact volume and total customers (contacts per customer). To the degree that the future echoes the past, this forecast will be accurate.

Part D of the form is where judgment plays a significant role. In this section, you customize the forecast by adding or reducing contacts, based on information you develop from your own and others' input. For some of these factors, you may have some hard data that you can use. For others, you'll be making more of an “educated guess.”

The factors in Part D are only examples, and you will want to create your own list specific to your environment. For example, in a support center

Blending in Judgment

	May 1	May 8	May 15	May 22	May 29	June 12	June 19
A. Projected Customers	____	____	____	____	____	____	____
B. Contacts per Customer	____	____	____	____	____	____	____
C. Base Contacts (A × B)	____	____	____	____	____	____	____
D. Activity Level Change	____	____	____	____	____	____	____
1. New Customers	____	____	____	____	____	____	____
2. Media Attention	____	____	____	____	____	____	____
3. Advertising	____	____	____	____	____	____	____
4. New Rate Structure	____	____	____	____	____	____	____
5. New Terms & Conditions	____	____	**Contacts (+ or -)**				____
6. New Service Procedures	____	____	____	____	____	____	____
7. New Information Required	____	____	____	____	____	____	____
8. New Product Introduction	____	____	____	____	____	____	____
9. General Activity Level	____	____	____	____	____	____	____
10. Product Performance	____	____	____	____	____	____	____
11. Competitors' Actions	____	____	____	____	____	____	____
12. Other	____	____	____	____	____	____	____
E. Total (add 1 through 12)	____	____	____	____	____	____	____
F. Projected Contacts (C + E)	____	____	____	____	____	____	____

for broken-down or stranded vehicles, weather would be a key influence on workload.

You will need a routine mechanism or forum for blending judgment into the forecast. A fairly common approach in contact centers is a weekly forecasting meeting. These meetings typically include members of the scheduling department and a representation of supervisors and managers from the contact center and other departments.

The meeting will typically last only 30 or 45 minutes. It often works like this:

- The person in charge of the meeting prepares an agenda of items to be discussed.
- The scheduling person (or team) prepares the quantitative forecast before the meeting.
- During the meeting, the attendees discuss issues that may influence the forecast, such as those in Part D of the worksheet. Each participant brings a unique perspective to the process.
- As each issue is discussed, the forecast is adjusted up or down, based on what the group believes will happen.

The collaborative approach is most effective when key team members who are accountable for staffing take an active role in forecasting (in large contact centers, they can be rotated through this process). The forecast not only improves as a result of their perspective, but these key team members gain an understanding of the factors that contribute to staffing. As a result, they lead their teams more effectively.

Measuring Accuracy—Strive for Five

How accurate should your forecast be? Large agent groups (100 or more agents) generally see relatively stable workload patterns and should strive for plus or minus 5 percent (or better) of workload down to specific intervals. Small groups (15 or fewer agents) often have more volatile patterns and should shoot for plus or minus 10 percent. Those in between should strive for something close to 5 percent.

This is not to suggest you can't do better. On the other hand, if you're just getting started with, say, a small group handling social media interactions, being even remotely close might be a pretty good start! Don't give up. Make this aspect of planning a priority. Be relentless in your focus until you begin to get your arms around it. Forecasting impacts everything to follow–staffing calculations, schedules and budgets, and, ultimately, the services you deliver.

Measuring Forecast Accuracy

Contact Volume

		Forecast	Actual	Difference	Percent*
8:30	9:00	342	291	51	17.5%
9:00	9:30	399	343	56	16.3%
9:30	10:00	461	499	-38	-7.6%
10:00	10:30	511	582	-71	-12.2%
10:30	11:00	576	649	-73	-11.2%
11:00	11:30	605	578	27	4.7%
11:30	12:00	572	513	59	11.5%
12:00	12:30	505	412	93	22.6%
12:30	1:00	456	540	-84	-15.6%
		4427	**4407**	**20**	**0.5%**

The accuracy of forecasting must be measured here ...

Not here!

* Variance of forecast to actual

Note: This example illustrates volume only; for an even more accurate assessment, apply this same approach to workload (volume × AHT).

When measuring accuracy, there's a running debate over whether to use the formula (forecast - actual ÷ forecast) or (forecast - actual ÷ actual). The important thing, though, is to describe what you calculate: actuals were X percent under forecast, or the forecast was Y percent over actuals.

It's essential to look at intervals rather than an average over a day or more. One good alternative is to use mean absolute percent error (MAPE), which is the average of the percentage error regardless of whether it was over or under. The advantage to using MAPE is that it measures how far off the forecast was. As seen in the table, the day's forecast was only off by half of a percent (.5%, or 4427 forecasted contacts versus 4407 received). The MAPE is 13.2% (the average of errors by increment with positive and negative signs removed).

Another alternative is to summarize the percentage of intervals that fall within various ranges of accuracy (see example). You can summarize a week, month, year or more in this way, and still provide meaningful data.

Forecast Accuracy by Interval

Accurate Within:	Percent of Intervals
5% or less	11%
5.1% to 10%	11%
10.1% to 15%	33%
15.1% to 20%	33%
Over 20%	11%

Common Problems (That You Can Avoid!)

The ICMI team often investigates why some contact centers have accurate forecasts and others don't. Ten common problems consistently emerge, and they are summarized here (in no specific order). In centers with inaccurate forecasts, usually two or three of these issues are most prevalent.

The good news? You can avoid these problems, and the remedies in most cases are fairly obvious.

1. NO SYSTEMATIC PROCESS IN PLACE.
There are often two erroneous beliefs that some managers use to justify the absence of a systematic forecasting process. Some say, "Our environment is too unpredictable. We're growing; we're going omnichannel; we're introducing new services; you can't predict social media interactions … there is no way we can produce an accurate forecast." Just know, there are many centers in similar situations that are enjoying the benefits of respectably accurate forecasts. Others aren't convinced that forecasting is

worth the time. Yep, it takes time–but not nearly as much as some imagine. Further, a good forecast will save a lot of time later on.

2. AN ASSUMPTION THAT "THE FORECASTING SOFTWARE KNOWS BEST."

If you have forecasting software or a workforce management system, don't relinquish decisions to the program, assuming that it knows best. The software doesn't know what the marketing department is about to do, or that average handling time will be affected by changes you are making to processes or systems.

Further, it is important to understand the assumptions your forecasting software is making. Some methods are user-definable. For example, you can program the system to give more weight to recent historical data, or you can tell it to ignore data that varies beyond X percent of the norm. It's a great idea to have the supplier provide a flow chart of the methodology the system is using and decision points where your input is necessary.

3. NOT FORECASTING AT THE AGENT GROUP LEVEL.

Even a perfect forecast of the aggregate workload will be of limited use if you route contacts to specialized groups. If you have a group of Mandarin-speaking agents handling services A, B and C, you will need to forecast contacts from Mandarin-speaking customers who need help with those services.

4. THE FORECAST IS TAKEN LIGHTLY.

If the forecast has been wildly inaccurate in the past or if no one understands the assumptions used in the process, it will not be given the prominence it needs in the planning steps to follow.

5. EVENTS THAT SHOULD BE EXCEPTIONS BECOME A PART OF THE FORECAST.

Utilities tend to get lots of contacts when storms knock out power, the financial industry gets swamped when confusing tax changes are implemented, and many centers have, on at least one occasion, dealt with contacts from an uncoordinated marketing campaign. (Have your agents ever had to sheepishly ask a customer, "Um, what does the promotion say we are offering?")

Those preparing the forecast have to be aware of the root causes of contacts. That will enable better judgment on what is likely to continue (and therefore should be built into the forecast) versus the exceptions.

6. ONGOING COMMUNICATION WITH OTHER DEPARTMENTS DOESN'T EXIST.
Most of what happens in a contact center is caused by something going on outside the center. The forecast is doomed if strong ties with other departments don't exist.

7. PLANNING IS DONE AROUND GOALS, NOT REALITY.
If staffing is based on a handling time of four minutes when actual handling time is more like seven minutes, the resulting schedules will be based on a pipe dream. Maybe improved training, streamlined procedures and better systems would move things in that direction. But ignoring reality in the planning process is no way to achieve better results or build confidence in the forecast.

8. NO ONE IS ACCOUNTABLE.
As vital as a good forecast is, often there is no one who spearheads the effort. Someone needs to be responsible for bringing the various types of input together, ensuring that it is integrated into the forecast, and investigating which assumptions were off when the forecast is not accurate.

9. AGENTS ARE MIXING FLEXIBLE ACTIVITIES INTO WORK MODES.
If agents are not using work modes consistently, especially after-call work, then accurate forecasting will be elusive.

10. NOT MAKING THE CONNECTION WITH STAFFING.
Forecasts mean nothing unless they are tied to staff and system resources required. That is the subject of Chapter 7.

ONE DAY, A LONG TIME AGO

As a new manager many years ago, I recall the “pain” of having to deal with a “rogue” product manager who routinely bombarded my contact center with direct marketing campaigns, without giving my team (or me) any advance warning. The product manager, who we’ll call Theresa, had dropped three such campaigns on us the previous month and as you might expect, the results were not pretty: Stressed out agents, angry customers, and a nightmare for our workforce management team, who were valiantly trying to improve our forecasts.

When the fourth “mystery” campaign hit the following week, I had Bob, our technology manager, reroute the 1-800 number to Theresa’s direct extension. In less than 10 minutes, I got a call from a distraught Theresa: “What the heck is going on with the 1-800 campaign? How am I supposed to handle these calls?” My reply: “Now you know how we feel!”

Things got better quickly. I explained that launching campaigns without properly preparing the contact center was not only stressful for our agents, it also reduced the success of the campaigns and potentially damaged our brand in the eyes of our customers.

I’m older and wiser now, and NO, I don’t recommend trying this approach in your own center. It can get you fired! It was, admittedly, the result of both my inexperience and Theresa’s in this area. But it’s a story I tell today because it reinforces the importance something I’ve since stressed for many years, that of developing strong relationships across the organization. Educate and engage. We’re all in this together.

By Gina Szabo, Senior Certified Associate, ICMI

Look Back and Adjust

Organizations that produce accurate forecasts are not necessarily those that have the most stable environments. Rather, they have a group of people (or an individual) who have made accurate forecasting a priority. They have taken responsibility, established good ties across departments, pulled in the data required, and established a forecasting process they are continually improving. They set accuracy goals and monitor progress. And they continue to work on and improve the assumptions they are making when looking ahead. In short, they consider accurate forecasting to be mission-critical.

Forecasting takes practice. You will never learn all there is to know about it—but you'll get better at it. One of the most important steps you can take to improve accuracy is to compare your forecasts with actual results and then ask, "Why?"

Points to Remember

- Forecasting is a blend of art and science, and it incorporates both quantitative and judgmental approaches.
- The forecast should accurately predict volume and average handling time. Volume alone is meaningless.
- The forecast should reflect adjusted offered calls (the individuals who try to reach you).
- There are numerous quantitative forecasting methodologies. Time-series forecasting provides a good foundation for many contact centers.
- You need mechanisms, such as a collaborative weekly meeting, to blend judgment into your forecasts.
- You need a workload forecast for each agent group.
- Accurate forecasts provide a solid foundation for the planning steps that follow.

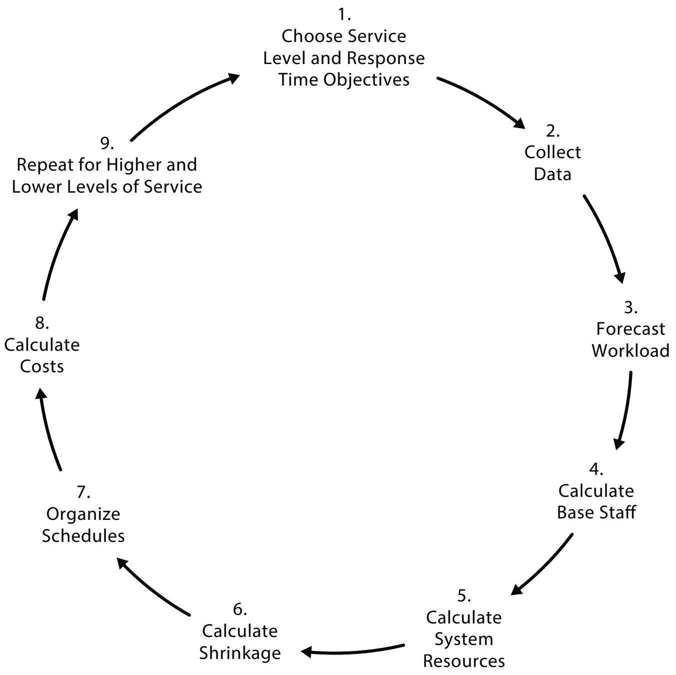
CHAPTER 7:

Determining Base Staff and System Resources

"Having the right number of
properly skilled people and
supporting resources in place
at the right times ..."
ICMI

The key to achieving service level and response time objectives ultimately comes down to having the right people in the right places at the right times, supported by sufficient system resources. With a reasonably accurate forecast, base staff calculations are usually straightforward.

In this chapter, we'll go through steps 5 and 6 of the planning process. We'll dispel common myths about staffing and system resource capacity. We'll go through the mechanics of correct calculations and explore how chat, social media, long calls, and other types of contacts impact requirements. And we'll review important definitions and measurements related to this aspect of planning.



Wrong Ways to Calculate Staff

Let's look at staffing for service level contacts—those that must be handled as they arrive. To calculate how many agents you need to staff for an increment (say, half hour), why not use this formula? Multiply the number of contacts forecasted by average handling time. Then, divide the result by 1,800 seconds (the total seconds in a half hour). You may even build in extra time, such as an added 10 percent or 20 percent, assuming agents will actually need a breather now and then.

Or what about this formula? Determine the actual average contacts per agent in a group. Then, divide that into the number of contacts forecasted. Or use target objectives, as in "our agents ought to be able to handle N contacts per half hour; therefore ..."

These methods may sound logical, and some managers use them. Unfortunately, they are dead wrong. They do not relate the outcome to a target

Wrong Ways to Calculate Staff

250 Contacts
× 210 Seconds Each
52,500 seconds ÷ 1,800 seconds
= 29.17 or:
29 Agents

Average contacts per agent, per half hour: 6.5
Contacts forecasted next month for time of day: 250
Therefore:
250 Contacts
÷ 6.5 Contacts Per Agent
= 38.46 or:
38 Agents

service level. Further, they are based on moving targets. The average group productivity (contacts that the group can handle) is not a constant factor. Instead, it is continually fluctuating because it's heavily influenced by vacillating workloads and the service level objective. But the biggest problem is that these approaches ignore a fundamental driving force in centers that handle customer-initiated interactions: contacts bunch up! (See Chapter 3.)

The following figure, Simulation of a Queue, illustrates an example queuing situation. (It's not as complicated as it first looks!)

In this scenario, 10 contacts arrive in a half hour, and each contact is assumed to last three minutes. The second column shows when each of the 10 contacts arrives. The third column gives the time each contact is answered, and the fourth column is the waiting time (the difference between when a contact arrives and when it is answered).

For example, contact number two arrives 4.4 minutes into the half hour, but has to wait 2.9 minutes before being answered because the first contact is still in progress. With one agent, the waiting times build throughout the half hour and beyond, and service is poor. With two agents, it's a different story; service is much better and waiting times are minimal.

Simulation of a Queue

Arrival		One Agent		Two Agents	
Arrival number	Time of arrival	Contact reaches agent	Waiting time (min.)	Contact reaches agent	Waiting time (min.)
1	0:04.3	0:04.3	0	0:04.3	0
2	0:04.4	0:07.3	2.9	0:04.4	0
3	0:15.7	0:15.7	0	0:15.7	0
4	0:17.3	0:18.7	1.4	0:17.3	0
5	0:21.1	0:21.7	0.6	0:21.1	0
6	0:22.1	0:24.7	2.6	0:22.1	0
7	0:25.4	0:27.7	2.3	0:25.4	0
8	0:26.3	0:30.7	4.4	0:26.3	0
9	0:27.4	0:33.7	6.3	0:28.4	1.0
10	0:27.5	0:36.7	9.2	0:29.3	1.8
		Average Delay:	**2.97**		**.28**

If sorting out staffing for random call arrival is this involved with two agents, imagine a scenario with 15 agents. Or 115! The point is, to determine staffing correctly, you need the right tools. You need a method that takes the usual randomness of contact arrival into consideration. That means using the Erlang C formula (or a variation of it) or computer simulation.

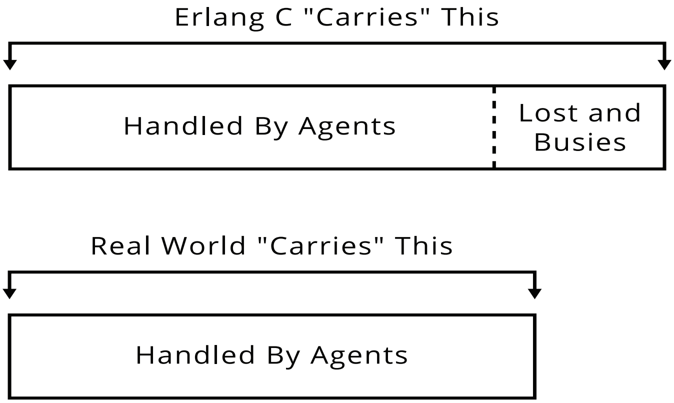
Staffing the Right Way

As introduced in the first chapter, the widely used Erlang C formula was developed in 1917 by A.K. Erlang, a Danish engineer with the Copenhagen Telephone Company. Erlang C can be used to determine resources in just about any situation where people might wait in queue for service—whether it is at a ticket counter, a bank of elevators, or a concessions stand in a stadium. Erlang C (or a variation of it) is currently built into virtually all of the commercially available workforce management software packages.

(If you go to Copenhagen, Denmark, you can visit an interesting display on

A.K. Erlang's life and work, in the building in which he used to work. I've been there. At the time, I was the only one there ... Okay, so it's not the most popular attraction just yet.)

Erlang C calculates predicted waiting times (delay) based on three things: the number of servers (agents); the number of people waiting to be served (customers); and the average amount of time it takes to serve each person. It can also predict the resources required to keep waiting times within targeted limits, and that's why it is useful for contact centers.



As with any mathematical formula, Erlang C has built-in assumptions that don't perfectly reflect real-world circumstances. For one, it assumes that "lost calls are delayed." In plain English, that means that the formula assumes that customers are queued when no agent is available. No problem with that. The problem is, it assumes that customers have infinite patience–they will wait as long as necessary to reach an agent and nobody will abandon. *Oops!*

Erlang C also assumes that you have infinite trunking and system capacity or that nobody will ever get a busy signal. But busies, rare as they may be in many contact centers, can and do happen. *Oops again!*

The result, in a nutshell, is that Erlang C may overestimate the staff you really need. If some of your customers abandon or get busy signals, your agents won't have to handle all of the work Erlang C is including in its calculations. Erlang C also assumes that you have the same level of staff handling the workload the entire half hour. In reality, if service level starts taking a nosedive, you may be able to add reinforcements on short notice.

ADVANTAGES OF ERLANG C

- Assumes random arrival and that contacts go into queue if an agent is not immediately available
- Is accurate at good service levels, where abandoned contacts and busy signals are minimal
- Is easy and quick to use, and is available in software programs from a wide variety of sources
- Illustrates resource tradeoffs well (e.g., when service level goes up, occupancy goes down)
- Is the basis for staffing calculations in most workforce management programs

DISADVANTAGES OF ERLANG C

- Assumes no abandoned contacts or busy signals
- Assumes "steady-state" arrival, or that traffic does not increase or decrease beyond random fluctuation within the time period
- Assumes you have a fixed number of staff handling the work throughout the time period
- Assumes that all agents within a group can handle the contacts presented to the group
- Calculations assume no contacts in queue from the prior half hour (unless the user or workforce management system makes this adjustment)

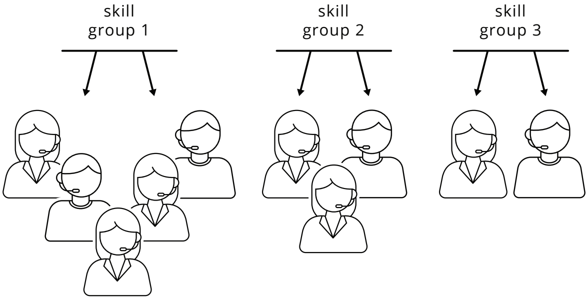
So then why is Erlang C so popular? As you might guess, there are defensible reasons to use it. For one, it's a *planning* tool, and most contact centers are planning to have good service levels. Erlang C is fairly accurate for good service levels—and when service level is decent, you should theoretically have little in the way of lost contacts or busy signals. If you do have a lot of contacts abandoning or getting busy signals, it's probably because you don't have enough staff to handle the load. In that case, who's worried about overstaffing? As your staffing more accurately reflects the workload demand, Erlang C will inherently become more accurate.

Additionally, if you adjust for abandoned contacts and busy signals (which

Erlang C does not do) and retry rates turn out to be higher than you estimate, you could end up underestimating staff. (And frankly, there's a little industry secret … *shhh* … some managers have decided that a little over-calculation as a safety net isn't such a bad thing. They figure that they don't get full use of their already authorized headcount, due to staff turnover and the time it takes to hire and train replacements.)

Finally, many software calculators and workforce management systems use modified forms of Erlang C. Alternatives with names such as Merlang, Erlang X and others enable you to adjust for abandoned contacts, busy signals, or variations in agent groups. Just be certain to have the supplier review with you the assumptions being made—your CFO will probably ask you the same thing.

Traditional Agent Groups



Erlang C is designed for straightforward environments, like sales contacts going here and customer support contacts going there. But the realities of today's contact centers are not so straightforward. You may have complex routing contingencies in place, such as agent groups that overlap, skills-based routing and complex network interflow.

Enter computer simulation. What staffing simulation does for contact centers is comparable to what flight simulators do for aircraft manufacturers. Aircraft manufacturers spend a lot of time simulating new designs. And they do fail—on computer. By the time the real thing is produced, they know the ins and outs of good design.

ADVANTAGES OF COMPUTER SIMULATION

- Can be programmed to assume a wide variety of variables, including overlapping agent groups and skills-based routing
- The assumptions can include abandoned contacts and busy signals
- Variables can often be labeled to use the terminology of your systems
- Results may cover a wide range of outcomes and include additional analysis (such as impact on costs)
- Results may be more credible to decision makers who are not familiar with alternative queuing formulas

DISADVANTAGES OF COMPUTER SIMULATION

- Takes time to set up and use
- Requires a relatively advanced user
- Is often a standalone tool that is not integrated with forecasting and staffing modules
- Does not tell you what to do (it instead illustrates what will happen based on variables you input)
- Is more expensive than standalone Erlang C programs or entry-level staffing and scheduling packages

Similarly, you can use simulation to zero in on the resources your contact center needs without making too many mistakes in your live environment. There are various simulation packages available (some designed specifically for contact centers and some more generic), and workforce management systems are increasingly providing simulation modules within their applications.

But computer simulation has some downsides. For one thing, simulation by itself is designed for modeling, design, and verification, and is generally not meant to be a forecasting and scheduling tool. As a result, if you want the time-saving benefits of software, you will still need a forecasting and scheduling system.

Second, simulation software takes more time and expertise to set up and use than Erlang C. Like a flight simulator, you have to run it over and over to identify potential results. That is a phenomenon of its added flexibility, and the time spent will be time saved if you have a complex environment that requires a simulator's perspective. But it takes more effort and know-how to enter and test variables and interpret the results.

So, what should you use? For fairly straightforward environments with good service levels, Erlang C or variations of it will likely be sufficient. And even if you have a more complex environment, there is something to be said for a combination of Erlang C, intuition, and experience. But if you really want to understand requirements in the most complex settings—skills-based routing or an omnichannel environment with many contact handling variables in play—no formula will ever beat simulation.

Just remember, no method can perfectly predict outcomes, at least not consistently. As much science as may be involved, it's inexact and must be augmented with common sense and resource plans that are at least somewhat flexible.

Base Staff

Recall the two major categories of contacts defined in Chapter 4: those that must be handled when they arrive (service level contacts) and those that can be handled at a later time (response time contacts). Let's look first at calculating base staff for service level contacts, using Erlang C.

Service Level

One of the advantages of using Erlang C is that it is a great educational tool, and it illustrates queue dynamics and resource tradeoffs well. So, that's what I'll be using here to outline basic staffing requirements and tradeoffs for service level, in a contact center setting.

For most of us, Erlang C in its raw form is unwieldy at best and totally unusable at worst. For all the complexities of a modern contact center, we can count our blessings for much better tools!

Erlang C

$$P(>0) = \frac{\dfrac{A^N}{N!} \quad \dfrac{N}{N-A}}{\sum_{x=0}^{N-1} \dfrac{A^X}{x!} + \dfrac{A^N}{N!} \; \dfrac{N}{N-A}}$$

Where A = total traffic offered in erlangs
N = number of servers in a full availability group
P(>0) = probability of delay greater than 0
P = probability of loss — Poisson formula

(Note: in the examples that follow, I am using QueueView, a low-cost program provided by ICMI. Other free and low-cost staffing calculators are available from a wide variety of sources–online, as apps, as stand-alone programs, as spreadsheet add-ins, or as modules within workforce management programs.)

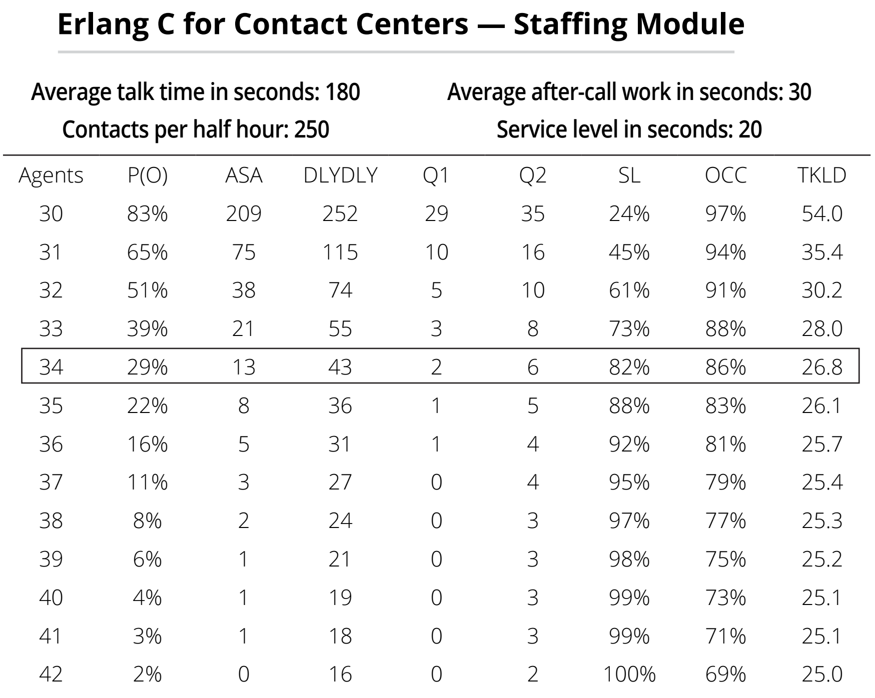
Erlang C requires you to input four variables:

- **AVERAGE TALK TIME IN SECONDS.** Input the projected average for the future half hour you are analyzing.
- **AVERAGE AFTER-CALL WORK IN SECONDS.** Input the projected average for the future half hour you are analyzing.
- **CONTACTS PER HALF HOUR.** Input the projected volume for the future half hour you are analyzing.
- **SERVICE LEVEL IN SECONDS.** If your service level objective is to answer 90 percent of contacts in 20 seconds, you will input 20 seconds. If it's 80 percent in 15 seconds, plug in 15 seconds. In other words, the program needs the Y seconds in the definition, "X percent of contacts answered in Y seconds."

Input the numbers and *voilà*! The output provides a wealth of information and insight into the dynamics of contact center queues (see Erlang C for

Contact Centers–Staffing Module).

Probably the first column you'll look at is labeled "SL," which is service level. That's the X percent to be answered in the Y seconds you input (meaning X percent of contacts reach agents within Y seconds). In the first row, the number 24 means that 24 percent of the contacts reach agents within 20 seconds. The next row is 45 percent, meaning 45 percent reach agents within 20 seconds, and so forth.



Erlang C for Contact Centers — Staffing Module

Average talk time in seconds: 180 **Average after-call work in seconds: 30**
Contacts per half hour: 250 **Service level in seconds: 20**

Agents	P(O)	ASA	DLYDLY	Q1	Q2	SL	OCC	TKLD
30	83%	209	252	29	35	24%	97%	54.0
31	65%	75	115	10	16	45%	94%	35.4
32	51%	38	74	5	10	61%	91%	30.2
33	39%	21	55	3	8	73%	88%	28.0
34	29%	13	43	2	6	82%	86%	26.8
35	22%	8	36	1	5	88%	83%	26.1
36	16%	5	31	1	4	92%	81%	25.7
37	11%	3	27	0	4	95%	79%	25.4
38	8%	2	24	0	3	97%	77%	25.3
39	6%	1	21	0	3	98%	75%	25.2
40	4%	1	19	0	3	99%	73%	25.1
41	3%	1	18	0	3	99%	71%	25.1
42	2%	0	16	0	2	100%	69%	25.0

Source: ICMI's QueueView Staffing Calculator

Let's say your objective is to answer 80 percent of contacts in 20 seconds. Keep going down the rows and … hey, where's 80 percent? The answers go from 73 percent to 82 percent, but where's 80 percent? You guessed it–the program is calculating staff required, and people come in "whole numbers," so some rounding is involved. Because 82 percent meets your standard, that's the row on which you would concentrate.

Next, glancing across that row, you can see that you need 34 agents (first column), average speed of answer will be 13 seconds (third column), etc. In other words, each column provides insight and information into the service level you choose.

Here's what the column headings stand for:

AGENTS: number of agents required to be plugged in and available to handle contacts. In this example, 34 agents will achieve a service level of 82 percent answered in 20 seconds.

P(0): probability of a delay greater than zero seconds. In other words, the probability of not getting an immediate answer. In this example, about 29 percent of contacts will be delayed. That means 71 percent of customers won't be delayed, but instead will go right to an agent.

ASA: average speed of answer. With 34 agents handling calls, ASA will be 13 seconds. ASA is the average delay of all contacts, including the ones that aren't delayed at all. In this example, 250 contacts are included in the calculation. (See discussion on why ASA is often misinterpreted, Chapter 4.)

DLYDLY: average delay of delayed contacts. This is the average delay of those contacts that are delayed—43 seconds, in this example. DLYDLY is a better reflection than ASA of what's actually happening to the customers that end up in queue. But keep in mind, it's still an average. Some wait five seconds and others may wait several minutes. If customers end up in queue any amount of time, they will be included in the calculation.

Q1: average number of contacts in queue at any time, including times when there is no queue. The label is somewhat of a misnomer, because Q1 incorporates all contacts into the calculation, including those that don't end up in queue. However, this column makes a useful contrast with the next, Q2.

Q2: average number of contacts in queue when all agents are busy or when there is a queue. In the example, an average of six contacts are in queue (when there is a queue). Again, this is an average, and there will sometimes be more than six contacts in queue, sometimes fewer. But this figure can provide useful guidance for what to look for when monitoring real-time information, and it can also be useful for determining overflow parameters.

SL: service level, the percentage of contacts that will be answered (meaning, that will reach agents) within the number of seconds you specify (e.g., 82 percent in 20 seconds).

OCC: percent agent occupancy. The percentage of time agents will spend handling contacts, including talk time and after-call work. The rest of the time they are available and waiting for contacts. In the example, occupancy will be 86 percent. Notice the tradeoff: when service level goes up, occupancy goes down. We will discuss this dynamic in Chapter 9.

TKLD: This column is the hours (erlangs) of trunk traffic, which is the product of (talk time + average speed of answer) × number of contacts in an hour. Because Erlang B, bandwidth calculators and other alternatives used for determining trunk and related system capacity often require input in hours, these numbers can be readily used as is. The actual traffic carried by trunks in a half hour will, in each row, be half of what is given. (See discussion of trunks and system resources in this chapter.)

The mechanics of staffing are easy enough. Plug in your numbers and you get some answers. Great! However, the interpretation takes a bit of thought and application.

A good question to ask for any service level is, “What happens to the customers that don’t get answered in Y seconds?” Programs that calculate delay can be very useful in answering this question. (The table, Erlang C for Contact Centers—Customer Delay, is also part of ICMI’s QueueView program.)

As you can see, 34 agents will result in a service level of 82 percent of contacts answered in 20 seconds. But here we get additional insight into what happens to individual customers. Sixty-five customers will wait five seconds or longer. In the next five seconds, seven of those customers reach

Erlang C for Contact Centers — Customer Delay

Average talk time in seconds: 180 **Average after-call work in seconds: 30**

Contacts per half hour: 250 **Service level in seconds: 20**

		Number of customers waiting longer than x seconds											
Agents	SL%	5	10	15	20	30	40	50	60	90	120	180	240
30	24	203	199	195	191	184	177	170	163	145	129	101	80
31	45	156	149	143	137	126	115	105	97	74	57	34	20
32	61	118	111	104	97	85	74	65	56	38	25	11	5
33	73	89	81	74	67	56	47	39	32	19	11	4	1
34	82	65	58	52	46	37	29	23	18	9	5	1	0
35	88	47	41	36	31	24	18	14	10	4	2	0	0
36	92	34	29	24	21	15	11	8	6	2	1	0	0
37	95	24	20	16	14	9	6	4	3	1	0	0	0
38	97	16	13	11	9	6	4	2	2	0	0	0	0
39	98	11	9	7	5	3	2	1	1	0	0	0	0
40	99	7	6	4	3	2	1	1	0	0	0	0	0
41	99	5	4	3	2	1	1	0	0	0	0	0	0
42	100	3	2	2	1	1	0	0	0	0	0	0	0

Source: ICMI's QueueView Staffing Calculator

agents, so only 58 customers are waiting 10 seconds or longer. In the next five seconds, six more customers will reach agents, leaving only 52 callers waiting 15 seconds or more. At this service level, one customer is still waiting three minutes or more (Murphy's law: that's the chairman of the board, testing your service).

Note an important implication of delay: because of random contact arrival, different customers have different experiences even though they contacted the organization during the same half hour and even though the contact center may be hitting its target service level. What's the worst case that your organization is willing to tolerate? This becomes a key question when exploring these tradeoffs.

(Some centers attempt to set two service levels for the same queue: 80 percent of contacts handled within 20 seconds and the rest within 60 seconds. As you can see, that is not possible—80/20 and 100/60 are distinctly different service levels, and therefore require different staffing levels to achieve.)

If you have never used an Erlang C program, we recommend that you get one and experiment with it. You will learn a lot about staffing dynamics and tradeoffs very quickly by using your own numbers and playing a bit of "what if." *What if handling time goes up? What if the number of contacts goes down? What happens to service level if we have fewer agents than required?* And so forth—have at it!

So far, you have calculated staff required to handle a specified mix of inbound contacts that must be handled when they arrive, for one half hour of the day. You will also need to calculate base staff for each half hour of the day and for every unique group of agents—sales, customer service, or other agent groups you have. In step 6 of the planning and management process, we will discuss how to factor in breaks, absenteeism and other variables that keep agents from the work, so that the schedule (step 7) reflects the total staff you need.

Response Time

Calculating staff requirements for response time contacts—workload that does not have to be handled at the time it arrives—is typically based on the centuries-old "units of output" approach. Here's the logic, and let's use customer email as an example: if you get 60 email messages that have an average handling time of four minutes, that's four hours of work. One agent working non-stop could handle the load in four hours. If you need to complete the interactions within two hours, you will need a minimum of two agents working over a two-hour period.

In other words, the workload and your response time objective dictate staff requirements. The basic formula is as shown.

Volume is the quantity of interactions you must handle, AHT is the average

Basic Response Time Formula

$$\frac{\text{Volume}}{(\text{RT} \div \text{AHT})} = \text{Agents}$$

Volume = Number of contacts to be handled
RT = Response time
AHT = Average handling time

amount of time it takes agents to handle them, and response time is the time you have to respond to customers after receiving their messages. Using the formula, you could handle the 60 messages previously mentioned in two hours with two agents: 60 ÷ (120 ÷ 4) = 2.

Here are some things to keep in mind:

- You can slice and dice base staff schedules many ways to achieve your objectives. In fact, in the example, you could have 60 agents rush in and handle all 60 interactions just before the promised response time and still meet your objective. What you are really doing is looking for an efficient way to distribute the workload across your schedules within the promised response time.

- The basic response time formula assumes a "static" amount of work to be completed–in other words, you have a defined amount of work that has already arrived and is waiting to be processed. However, contacts that can be deferred arrive throughout the day in patterns that are often similar to service level-type traffic. With 24-hour response time objectives, projected workload can simply be built into the following day's staffing requirements. But if you have more aggressive response time objectives, you'll need to look at both on-hand workload and projections by interval to determine staffing requirements.

- When response time objectives are less than an hour, use Erlang C or computer simulation to calculate base staff. This would be a queuing and service level scenario, like inbound calls.

- Breaks, absenteeism and other activities that keep agents from the work need to be added to base staff calculations (a step we'll cover in Chapter 8).
- An "efficiency factor" acknowledges that agents cannot handle one interaction after another with no "breathing" time in between. For example, if you want to build in an efficiency factor with a ceiling of 90 percent, divide base staff calculations by .9 to determine if additional agents are required.

In short, meeting response time objectives requires:

- Setting response time objectives
- Forecasting these interactions, within timeframes specific enough to calculate base staff required
- Calculating base staff needed
- Factoring in breaks and other activities that will take staff away from the work (step 6 covered in Chapter 8)
- Factoring these staffing requirements into overall schedules

Variations to Base Staffing

Today's contact centers are often characterized by a variety of contact channels and routing alternatives. You may be utilizing skills-based routing, sophisticated network environments or other configurations that go beyond simple agent groups. Your environment will dictate the staffing methodology that will yield the best results.

Outbound Contacts

There are three general types of outbound contacts. Each has implications for staffing requirements:

1. OUTBOUND THAT IS A PART OF THE INBOUND WORKLOAD.
For example, contacts to an emergency roadside service from stranded customers often involve outbound contacts to arrange towing or repair services. Whether these outbound contacts happen during contacts from customers or immediately following the contacts as part of after-call work, they should be considered part of handling time. As such, they should be included in base staff requirements for the inbound workload.

2. OUTBOUND CONTACTS THAT ARE SCHEDULED.
In many cases, outbound contacts to customers or prospects can be scheduled based on factors such as convenience to customers, highest probability of making successful connections, etc. Within these blocks of time, outbound contacts can be generated and handled one after another. Base staff requirements can be calculated using traditional response time calculations; for example, a minimum of five agents would be required to handle 20 hours' worth of contacts in four hours' time. As with other response time calculations, a reasonable efficiency factor should also be included in the assumptions.

3. OUTBOUND CONTACTS THAT ARE NEITHER A PART OF RANDOMLY ARRIVING INBOUND WORKLOAD NOR SCHEDULED.
In every contact center, there are at least some outbound contacts to colleagues or customers that are neither part of the inbound load nor specifically scheduled. If significant enough to merit consideration in resource requirements, they can be reflected in step 6, covered in the next chapter.

Skills-Based Routing

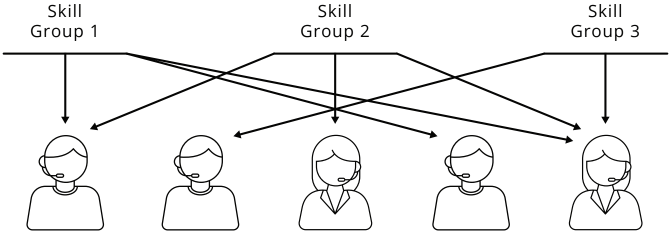
Available in ACD routing systems since the early 1990s, skills-based routing is a powerful capability designed to match each customer with the agent who has the skill set best suited to handle the contact on a real-time basis. It has been a boon to the efficiency and quality of services provided by centers that, by nature, have overlapping groups or complex routing contingencies.

But to be effective, skills-based routing must be managed well. Erlang C's assumption of traditional agent groups doesn't fit well, but computer simulation can help fill the gap.

The basic requirements for skills-based routing include:

- Identify and define the skills required for each contact type.
- Identify and define individual agent skills.
- Prioritize agent skills, based on individual competency levels.
- Devise and program an appropriate routing plan into the ACD.

Skills-Based Routing



Although specific programming approaches vary by system, you will essentially create two "maps" when you set up your ACD for skills-based routing. One will specify the types of contacts to be handled and the other will identify the skills available by agent. As an example, the maps for a technical support center handling calls across Europe might look like this:

MAP 1

English-speaking customers who need assistance with tablets
English-speaking customers who need assistance with laptops
English-speaking customers who need assistance with printers
English-speaking customers who need assistance with storage devices

French-speaking customers who need assistance with tablets
French-speaking customers who need assistance with laptops
French-speaking customers who need assistance with printers
French-speaking customers who need assistance with storage devices

German-speaking customers who need assistance with tablets
German-speaking customers who need assistance with laptops
German-speaking customers who need assistance with printers
German-speaking customers who need assistance with storage devices

And so on. The second map might look like this:

MAP 2

TOM—Speaks English, Dutch, and French. Trained on tablets and printers.

ANGELIQUE—Speaks French and Italian. Trained on laptops, printers, and storage devices.

ERIK—Speaks Swedish, French and English. Trained on tablets and laptops.

MARIA—Speaks Spanish, Italian and French. Trained on printers and storage devices.

Consider a simple case that illustrates the basic steps in staffing for skills-based routing. Assume you have two languages to handle—English and Spanish. And let's say that you have four contact types to handle—orders and technical support in each language. The agent skills can be illustrated as shown.

Agent Skills

Contact Types	Agent Type 1	Agent Type 2	Agent Type 3	Agent Type 4	Agent Type 5	Agent Type 6	Agent Type 7	Agent Type 8	Agent Type 9
Orders – English	X	X						X	X
Orders – Spanish				X		X		X	X
Tech Support – English		X	X		X				X
Tech Support – Spanish			X			X	X		X

Next, let's assume that your plan is to route customers to the least-skilled agent who can handle the contact because you want to keep more experienced or skilled agents available for less common or more complex contacts. Consequently, the routing plan would appear as shown.

Routing Plan

Routing Hierarchy	Order-English	Order-Spanish	Tech Support English	Tech Support Spanish
Skill Choice 1	Agent Type 1	Agent Type 4	Agent Type 5	Agent Type 7
Skill Choice 2	Agent Type 2	Agent Type 8	Agent Type 2	Agent Type 6
Skill Choice 3	Agent Type 8	Agent Type 6	Agent Type 3	Agent Type 3
Skill Choice 4	Agent Type 9	Agent Type 9	Agent Type 9	Agent Type 9

You set up the simulator the same way you program the maps into your ACD. You tell it what types of contacts you are going to get and the skills of your group. You also plug in the volumes of each type of contact you expect and corresponding handling time estimates.

We used this data to run three different scenarios, all with the same workload and service level objective:

- Conventional ACD groups (one group for each contact type)
- Skills-based routing
- Universal agents (a fully cross-trained group)

As the results in the table indicate, skills-based routing is more efficient than separate, segmented groups. Also note that universal agents, where each agent is fully cross-trained and speaks both languages, is the most efficient arrangement.

In general, skills-based routing works best in environments that have small groups where multiple skills are required. It has the potential to improve efficiency by matching customers with "just the right agent." Skills-based routing can also help to integrate new agents in "stages"—for example, they may handle billing calls first, then get additional training on tech support before handling those contacts.

Results of Each Routing Scenario

Time Period	Separate Groups by Contact Type	Skills-Based Routing Scenario	Universal Agents
9:00-9:30	30	27	24
9:30-10:00	43	41	39
10:00-10:30	64	62	59
10:30-11:00	58	56	52
11:00-11:30	44	41	40
11:30-12:00	31	28	27

Skills-based routing does have some disadvantages. It could be that the agent with just the right skill is on break at the wrong time. Mapping out skills and programming routing scenarios is one thing; getting people in the right place at the right times can be quite another. Small, specialized groups are tough to manage, and they can eliminate the efficiencies of pooling (a principle I'll cover in Chapter 9) common to more straightforward agent groups.

Further, routing and resource planning become more complex. Be prepared to run enough simulations to learn what's workable in your environment. You also need to develop contingency plans for when the workload of a specific contact type is greater than expected, or when you don't have the specialized staff you planned for (e.g., because of unplanned absences). And get used to an irony—the most skilled agent may be the most idle, given the usual intent to route contacts to agents with the minimum qualifying skills to handle them (as in the example).

Skills-based routing is a powerful capability. But it must be managed well. That means going through the planning process diligently. You'll need a good forecast and solid staff calculations. Also, remember to work toward pooled groups, to the degree that your circumstances allow. All things being equal, an environment with proficient, cross-trained agents will be the most efficient.

TROUBLESHOOTING SKILLS-BASED ROUTING

Skills-based routing … Intelligent. Flexible. Real-time. The perfect answer to that perennial contact center challenge of getting the right contact to the right place at the right time. At least, that's the way it's supposed to work. But in too many cases, skills-based routing also creates difficult new challenges that have wiped out potential benefits. Here are five of the most common problems:

STAFF SHRINKAGE.
Breaks, lunch, meetings, projects, research, training, absenteeism … you know the story. These things are particularly vexing in a skills-based routing environment where you are trying to get contacts to just the right agents. There's no substitute for realistically planning and budgeting for the things that keep agents from handling contacts (a subject we'll cover in Chapter 8).

INACCURATE FORECASTS.
The inability to forecast accurately for specific types of skill requirements is the Achilles' heel of the powerful simulation tools available—and of skills-based routing in general. To anticipate staffing needs, you first need to know how many Spanish-speaking customers you're going to get between 10:00 and 10:30, how the contact mix will change throughout the day for the expert group handling call types A, B and C, and when your Mandarin Chinese-speaking agents will go on break. Accurate forecasting at this level of specificity takes time and effort. If you're struggling with the detail, see if it's possible to combine skills (through hiring and training) to form more manageable agent groups.

INACCURATE BASE STAFF CALCULATIONS.
Whatever staffing method you use (Erlang C, a variation of it, or simulation), a certain amount of trial and error and a healthy dose of intuition and experience are necessary to accurately model the environment. You will need to run through quite a few (sometimes dozens of) "what-if" scenarios to get it right.

POOR ASSUMPTIONS AND RATIONALE.
Skills-based routing works best in environments that require many skills and have many possible combinations of skill sets. It can also help to quickly integrate new agents by initially routing only simple contacts or those of a predefined nature to them. What it can't do is compensate

for poor planning, inadequate training or poorly designed information systems. Keep it as simple as possible, and remember that skills-based routing depends on—rather than compensates for the lack of—accurate planning and good processes.

NO ROUTING MANAGER/COORDINATOR.
If all of this sounds time-consuming, that's because it is. Even relatively small contact centers have learned through tough, practical experience that it often takes the equivalent of a full-time person to keep skills-based routing running smoothly. Projecting requirements, assessing current capabilities, updating system programming and adjusting staffing plans and schedules to accommodate evolving circumstances are ongoing activities.

Network Environments

Network environments, like skills-based routing, can introduce complex contingencies into staffing calculations. The method you use for calculating staff will depend on the type of network environment you have.

Your contacts can be routed in different ways. Some examples include:

VIRTUAL CONTACT CENTER. In a true virtual environment, each contact is routed to the first available agent (or longest-waiting agent). Other routing and queuing contingencies (such as skills-based routing) notwithstanding, this environment represents a traditional agent group regardless of where agents are located (at home, across campus, or on the other side of the world), and Erlang C will generally produce accurate calculations.

NETWORK INTERFLOW. Contacts initially presented to one site can be simultaneously queued at other sites or sent to other sites based on parameters you define. Though not an all-in virtual center, it is similar in that you don't let contacts languish in queue too long before having the network look for other agents. The criteria that determine how contacts are interflowed can run the gamut, from availability at each site to the types of contacts you are handling. For example, you might immediately send high-priority contacts to available agents in any site, but queue all other contacts longer for the intended agent group. Simulation can help to model and test the environment under different conditions.

PERCENT ALLOCATION. With this traditional approach–less common now than more sophisticated alternatives–routing is set up to allocate contacts among sites. For example, you may program 40 percent of contacts to be routed to one site, 35 percent to a second, and the balance, 25 percent, to a third site. Anyone with network administration access can change allocations as needed. Erlang C will generally provide good results in this environment. As with agent groups in a single site, you will forecast the workload you anticipate and run Erlang C calculations for agent groups in each site.

There are many variations, and (you'll find this no surprise by now) I always feel more comfortable when someone shows me the specific parameters that determine how contacts are routed. I remember making a spine-tingling discovery one time with the help of another consultant. By mapping out network routing parameters step-by-step, we isolated why, on rare occasions, contacts to a crisis line were ending up in voice mail. It was *never supposed to happen*, but by going through call flow, step-by-step, we were able to isolate the problem. It was a good day.

The underlying principle is simple: staff appropriately for how and where you intend contacts to be handled. And, as with skills-based routing and other alternatives, keep it simple enough so that you and your team can manage it.

Long Contacts

Contacts that take a relatively long time to handle pose another staffing challenge. Thirty-minute reporting periods provide an adequate level of detail and accuracy for most contact centers. However, some centers, particularly those in technical support environments, handle contacts with long average handling times–defined as those that exceed 20 minutes.

When long contacts are not distributed as Erlang C anticipates, they may violate the assumptions of the formula. Compounding the problem is the fact that some ACDs count contacts in the period in which they begin, but report average handling time in the period in which they end. Consequently, reported averages can be skewed.

If your AHT approaches or exceeds 20 minutes, you may need to adjust your default reporting interval to an hour. Most Erlang C programs will allow you to define the interval you want to use. Alternatively, you can program a simulator to model the mix of contacts you are handling. If, on the other hand, long contacts are not common, but they do occasionally occur, you will need to adjust your statistics (remove them from assumptions) before using your historical data.

You will also need to consider how you manage long contacts. Most technical support environments have a second tier of support to handle complex issues. You will need to manage the service level for both tiers, or it will suffer in both. But when managed well, this approach can ensure that longer contacts don't tie up the primary group and cause erratic service levels.

Peaked Traffic

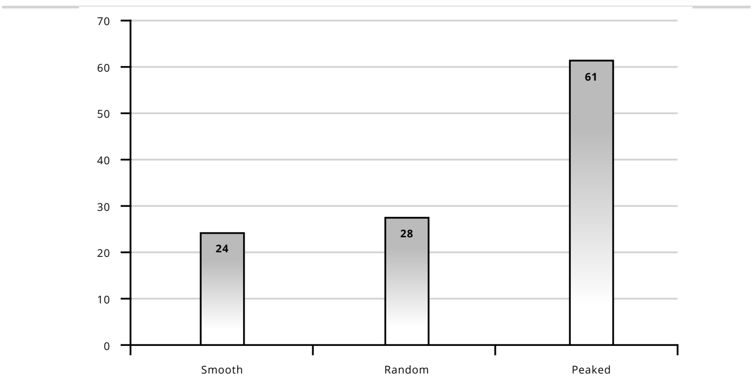
Peaked traffic, as discussed in Chapter 3, is a surge beyond random variation within a half hour, which poses a unique staffing challenge. For the purposes of this discussion, there are two types of peaked traffic—the type you can plan for and those incidents that are impossible to predict.

The contacts that utilities get just after a major power outage will surge far beyond normal random variation. Similarly, if a national news program unexpectedly provides your contact information to the viewing audience as part of a story, you will get unannounced peaked traffic—and it will be quite an experience!

The problem is, you can't predict these events, and you're probably not willing to staff up for them just in case they happen. So staffing for unexpected peaks falls more in the category of real-time management or disaster recovery planning.

On the other hand, peaked traffic that you *are* expecting belongs squarely in the realm of planning. Forecasting, staffing, and scheduling to meet a specified service level still apply. However, planning must happen at much more detailed periods of time, often in 5- or 10-minute increments. For a given service level, peaked traffic requires more staff than random traffic, and agents will have a lower occupancy over a half-hour period.

Comparison of Relative Staffing Levels Required for Each of the Three Contact Arrival Patters (Equal Traffic)



Most centers use Erlang C to calculate base staff for predicted peaks. If you expect 200 contacts in a 5-minute span, that's the equivalent of 1,200 contacts in a half hour. If you use an Erlang C program based on half-hour data, you will assume 1,200 calls for the calculations. Alternatively, some staffing programs allow you to specify the timeframe you choose, and can accommodate short intervals.

But common sense is required. If you have 75 people staffed to handle the 200 contacts, and the contacts come in at virtually the same time, you know that the first 75 are going to be answered immediately. The next 75 are going to have to wait, with an average wait that approximates the average handling time of the first 75 contacts. The last 50 customers will have to wait something like two times the average handling time of the contacts.

The situation is similar to a bus dropping people off in front of a sports arena. Those reaching the ticket booths first get quick service. For others, wait times can be dramatically different, depending on where they end up in the line. Consequently, how peaked the traffic is (how concentrated it is within a small period of time) and the sequence in which specific customers arrive will dramatically impact service level. Use common sense as you think through these scenarios, and take steps to ensure your customers don't spend too long in queue.

Chat

Let's turn to interactions with multiple back-and-forth messages between a customer and agent. Chat is the most common, but the same considerations apply to channels with similar characteristics (e.g., some text and social media interactions).

Chat is a service-level-oriented contact, and the essential planning steps apply. You can determine the most conservative (highest) estimate of agents you'll need by assuming that each agent can interact with only one customer at a time, then using Erlang C or simulation to calculate staff requirements based on the usual input—number of contacts, average session transaction time (the equivalent of AHT), and your service level objective. You'll then adjust for (divide by) the average number of concurrent sessions your agents handle. As we'll see, though, workflow-routing and handling options lead to different staffing requirements.

TERMS

Let's review some important terms. The following are definitions that we use and recommend. (Note that, depending on the application, the roles may be reversed—the agent may make an initial request to a customer browsing the organization's website.)

SESSION: The whole of the interaction, from hello to goodbye

EXCHANGE: A part of a session that begins with an inquiry from the customer and concludes with a response from the agent

SESSION RESPONSE TIME: The time it takes the organization to respond to the initial request for a session from the customer

EXCHANGE RESPONSE TIME: The time that elapses between the customer sending a question or comment and the delivery of the agent's response

CUSTOMER RESPONSE TIME: The time it takes the customer to read an agent's reply and send a response

EXCHANGE HANDLE TIME: The time it takes for the agent to prepare and deliver a response during an exchange

SESSION HANDLE TIME: The cumulative total of the exchange handle times for the session

SESSION TRANSACTION TIME: The time elapsed from the beginning of the initial exchange to close-out

CLOSE-OUT: The moment in time when the session is considered complete

KEY DECISIONS

There are some important decisions you'll need to make as you set up and manage chat interactions. They include:

WHEN WILL YOU NEED FORMAL PLANNING? While some organizations use chat extensively, it makes up a relatively small portion of the contact workload for many others. If you're just starting out, you'll need to answer a fundamental question: When do you move from "educated guessing" to staffing approaches that are more scientific? After all, if you only need one or two agents handling chat, advanced mathematical approaches won't yield any more accuracy than common sense.

ICMI recommends that a sensible threshold is five—when you need five or more agents handling chat at any one time, a more disciplined approach will begin to pay off. (This is also a sensible threshold for social media and other types of interactions.)

HOW MANY CONCURRENT SESSIONS ARE FEASIBLE? Another key decision is around the number of simultaneous sessions you allow agents to handle. Systems can be configured to enable 16 or more simultaneous sessions per agent—which, of course, is impractical from a human standpoint. The number of maximum concurrent sessions you allow will impact response times, average handling times, customer satisfaction, accuracy, and employee morale.

To determine how far beyond one session at a time you can move, basic math comes in handy. Assume that you set the maximum number of concurrent sessions at five (too many for most, but let's go with it for this example). It's simple and valuable to develop worst-case estimates.

Here's the formula: Multiply the maximum number of concurrent sessions you expect by the average exchange handle time. The result will give you an idea of what could happen (worst case) to customer wait times. For example, if five customers initiate an exchange at the same time, and the average exchange handle time is 1.25 minutes, the last customer in line will have to wait 6.25 minutes for a response (5 × 1.25). This scenario won't happen often. But if and when it does, the delay would be well beyond the expectations of most customers. So, five concurrent sessions would be too high for an organization focused on delivering high levels of service.

My advice to those just starting out: Go with no more than one or two until you get a better read on what's possible and get the kinks worked out of the system. Overstating the number of concurrent sessions will leave you short-staffed. Consultant Jay Minnucci adds another important consideration: some organizations will only do one chat at a time, for fear of an agent accidentally providing one customer's information to another.

WHEN DO AGENTS RECEIVE A SESSION? Another decision you must make is when an agent will receive a session. If a customer's initial request is immediately delivered to an agent, you can send an automated, personalized greeting from that agent to the customer. If you decide to delay routing to the agent, you will need to deliver either a blank chat box or one with a generic greeting.

Here's the staffing tradeoff: If you provide the more personalized approach, you will need to live with the chance that you may be tying up an agent too early–some customers will request a chat session but then never initiate the exchange, and the agent will be left waiting for a question that never comes. Given this possibility, you will probably want to allow relatively more concurrent sessions per agent than in a scenario where an agent is selected only after an exchange is initiated.

WHEN DOES A SESSION END? You will also need to define when a session ends. Often, the point of close-out is clear–but sometimes it's not.

For example, customers may get what they need and ignore further attempts at communication; they may step away from their computer or device; or they might head off to a competitor's site. (Chat is often perceived to be less personal than calls, and customers may apply different rules of courtesy.) While your agent waits for a response, the session is considered active. So you'll need to decide on procedures to try to re-engage the customer, and when the agent can, in effect, "give up" and close the session.

Staffing implication: The longer the threshold until close-out, the more time the agent will spend waiting for an exchange that may never occur; accordingly, a long threshold would suggest you can allow a relatively higher number of concurrent sessions per agent.

In short, staffing for chat revolves more around questions of workflow and technology application than on mathematical calculations. As volumes rise, you'll need to make decisions in each of these areas that are right for your customers.

CHAT REPORTS

Although concurrent chat sessions can improve productivity, they can also make reporting more difficult. Consider again the example of five concurrent sessions, where all five customers initiate an exchange at the same time. The last customer served will have to wait 6.25 minutes for a response–but most of that time was spent on other exchanges with other customers. The reporting challenge is accounting for these variables.

For example, when a customer initiates an exchange, the reporting system must note how many concurrent exchanges are already in queue for that agent in order to determine the exchange handle time. The customer who is fourth in line will wait a total of five minutes for a response. Divide that wait time by the number of exchanges in queue, and you'll come up with the exchange handle time of 1.25 minutes (5 ÷ 4). But you can see that reporting must account for many variables.

In short, these are issues you'll need to review with your technology provider. How does the system make these calculations and what do the reports produced really mean?

Social Media and Text

For staffing purposes, there are different types of social media interactions, each requiring a specific approach to resource planning. Text, direct messages through social platforms, or in-app messaging have similar considerations. Here are some common variations:

REAL-TIME INTERACTIONS, WITH SINGLE RESPONSE. In this setting, the organization handles interactions through social media channels as they occur, with one response generally being sufficient. Typical examples include responding to customers with numbers they can contact, specific email addresses, or links to online resources that provide necessary information. These are service level-type interactions, and the staffing approach is like that for inbound calls.

REAL-TIME INTERACTIONS, WITH MULTIPLE EXCHANGES. In this case, the organization strives to handle interactions when they are initiated, and the dialogue often involves multiple back-and-forth messages. Once engaged, customers may continue to ask questions or seek clarification. These are service level-type contacts with staffing considerations like those of chat (see discussion on chat, above).

INTERACTIONS THAT CAN BE DEFERRED. This approach involves addressing inquires or issues that do not require an immediate response. Common examples include responding to general inquiries posted through social media sites, or sending responses, FAQ documents, or relevant links that address questions posted in forums. In this scenario, staffing is response time-oriented, like that for email or scheduled outbound contacts.

INTERNAL INTERACTIONS. The impact of internal collaboration tools on staffing requirements must be considered in context. If internal communication is necessitated by (and happens while) handling customer interactions, the time required should be reflected realistically in the average handling time associated with those contacts. On the other hand, internal communication that is not directly associated with handling customer contacts (e.g., for internal projects or, simply, the everyday communication that is part of a normal work environment) should at least be accounted for in overall schedule requirements (see Chapter 8).

Note that different systems will deliver social media interactions to agents in different ways. For example, some present social interactions in email-like format. That's fine. Just remember, it's not email in the usual sense, and staffing requirements should be driven by whether the work needs to happen at the time and whether it involves multiple exchanges.

Considerations for text messages (those that involve agents) are similar—do they occur at the time of initial inquiry (requiring a service level approach) or can they happen later (response time)? And, are they generally handled in one response (like a call) or do they require some back-and-forth (like chat)?

Practice!

Whew! All the variables in staffing are a lot to think about.

Omnichannel environments, depending on how you route and queue contacts, can include any combination of channels and contact types. You'll want to consider both customer experience and the ability of your agents to handle different channels and workload.

Given the variables and increasing variety in today's workloads, you and your team have a lot to think about. But take heart: Getting staffing right is something you get better at with practice. It gets easier. And you can always make adjustments as you go along.

Staffing goes to the heart of what contact centers have always excelled at: matching up supply and demand in real time. Remember that when all is said and done, success is more important than perfection.

Trunks and System Resources

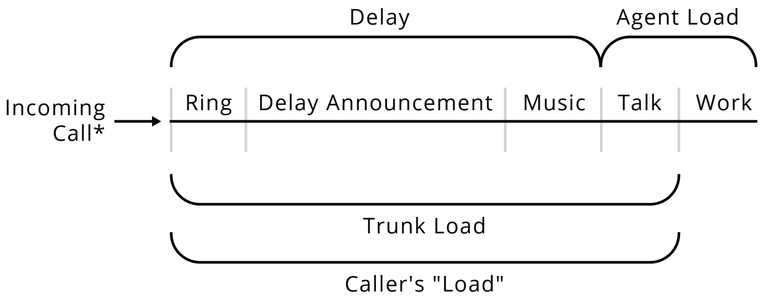
Along with staffing, you'll need sufficient trunks and system resources. And there's an important relationship to understand: trunking (the lines or bandwidth capacity you need) must be calculated with an understanding of staffing requirements. Staffing impacts delay, which affects the load that systems and networks must carry.

If you are thanking your lucky stars that you have an IT/telecom department to worry about system issues, and, frankly, never wanted to know about trunks and certainly not any relationship between staff and trunks … I do understand. But stick with me on this. Understanding the underlying dynamics is important and will serve you well, even if you're not the one doing the calculations.

Calls

Let's look at a few basic definitions. We'll consider them in a traditional context (e.g., a trunk as a line required to carry a conversation). When interpreting the diagram, Progression of a Customer Call, assume inbound calls entering a "straight-in" environment, where customers dial a number and are routed directly to the agent group. For our simple example, let's assume no IVR ("Press or say one for ...") is involved.

Progression of a Customer Call



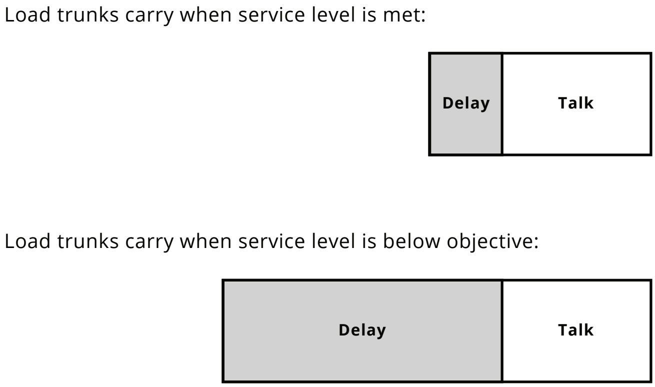
DELAY: Delay is everything from when the trunk is seized to the point at which the customer is connected to an agent.

AGENT LOAD: Agent load includes the two components of handling time–talk time and after-call work.

TRUNK LOAD: Trunk load includes all aspects of the interaction other than after-call work, which does not require a trunk. The "caller's load" is the same as the trunk load, other than the short time it takes for the network to route the call to the contact center.

Notice that agent load and trunk load both include talk time. However, trunk load carries the delay, which is not a direct part of the agent load. And the agents handle after-call work, which is not carried by the trunks. The more staff handling a given call load, the less delay callers will experience. In other words, staffing impacts delay; therefore, it directly impacts the trunking capacity that is required.

The Impact of Service Level on Trunk Load


Load trunks carry when service level is met:
Delay
Talk
Load trunks carry when service level is below objective:
Delay
Talk

Calculations

As you consider trunks and related system resources, let's start with the assumption that you need a trunk (a line, a path) for each simultaneous conversation (we'll then look at how that might not be the case). More specifically, you need enough trunks to carry the delay that customers experience (the time from the moment contacts arrive at the routing system until they reach agents) and the conversation time (talk time), for the period you are analyzing.

The general method for calculating trunks is as follows:

1. Forecast the workload to be handled for the busiest half hour in the foreseeable future.

2. Compute the number of agents required to handle the forecasted workload at your service level objective.

3. Determine the trunk load according to the workload you will be handling and the service level you can realistically achieve. The trunk load represents "erlangs"—hours of traffic in an hour (how much time, in hours, customers are in queue or connected to agents over an hour).

4. Determine the number of trunks required to handle the calculated trunk load, using (typically) Erlang B.

Erlang B or variations of it (yep, also the work of our beloved Danish engineer) is widely used and is often available in workforce management programs or in resource calculators (including QueueView, the program that I am using for examples). With any of the usual formulas used for calculating trunks, you will need to specify the probability of busy signals you can live with; if you input zero, you'll need as many trunks as there are calls. But if you can tolerate even a small probability of busy signals, then the number of trunks required becomes much more realistic. In the staffing sample on page 143, 38 trunks are required to handle 26.8 erlangs of traffic with a 1 percent probability of busy signals.

Erlang B

$$P = \frac{\frac{A^N}{N!}}{\sum_{x=0}^{N} \frac{A^x}{x!}}$$

Where A = total traffic in erlangs
N = number of trunks
P = grade of service

SIP Trunking, IVRs, Chatbots

Today, physical circuits have largely been replaced by SIP (Session Initiation Protocol) services. SIP is a standard for initiating and managing connections. SIP can manage sessions across a variety of media in an omnichannel contact center. SIP trunking can boost efficiency by dynamically allocating bandwidth to make use of network capacity.

> *SIP is a standard for initiating and managing connections. SIP can manage sessions across a variety of media in an omnichannel contact center.*

Enter bandwidth calculators. They consider the load you need to carry, channels you're handling, the level of "voice compression" you can live with, and other factors to help ensure you've got sufficient bandwidth in place. They typically use Erlang B to determine simultaneous connections you'll have with customers, and calculations of bandwidth requirements to ensure those connections are clear and of good quality.

You will probably have other variables to consider. For example, most contact centers have an IVR system that customers go through before they reach an agent (e.g., to enter their account number, route themselves, authenticate through voice biometrics, and other applications). If so, these requirements will need to be factored into the calculations.

It is also a consideration if trunk capacity is shared across agent groups or the broader organization. When all is said and done, you'll need adequate bandwidth to support the busiest foreseeable increments in your "universe."

Workload demands will also drive other system requirements: IVR capacity, chatbots, how you engineer your network, the resources you'll need for backup in case of disaster, etc. The underlying engineering principles are the same.

There are many scenarios for system and trunking requirements that go beyond the scope of this book. I highly recommend that you get the help of a competent IT/telecommunications professional to engineer your system. But *be sure* they understand the relationship between trunks and staff, and what your workload projections look like. And make sure that staffing and trunking are coordinated activities, both in calculations and in your budgets.

Points to Remember

- The Erlang C formula (or variations of it) is commonly used for calculating base staff; it is easy to use and widely available. Computer simulation is more difficult to use than Erlang C, but can more accurately model complex environments.
- Social media interactions, email, chat, text, outbound calls, and other kinds of contacts require staffing approaches appropriate to the unique characteristics of each.
- No staffing methodology is perfect, and it's important to understand the assumptions each makes and to blend in a good dose of common sense.
- Staffing and trunking are inextricably related. The fewer people you have staffed for a given workload, the more network and system capacity you'll need.
- SIP trunking is the standard for managing contacts in an omnichannel contact center.

CHAPTER 10:

Communicating Requirements to Senior Management

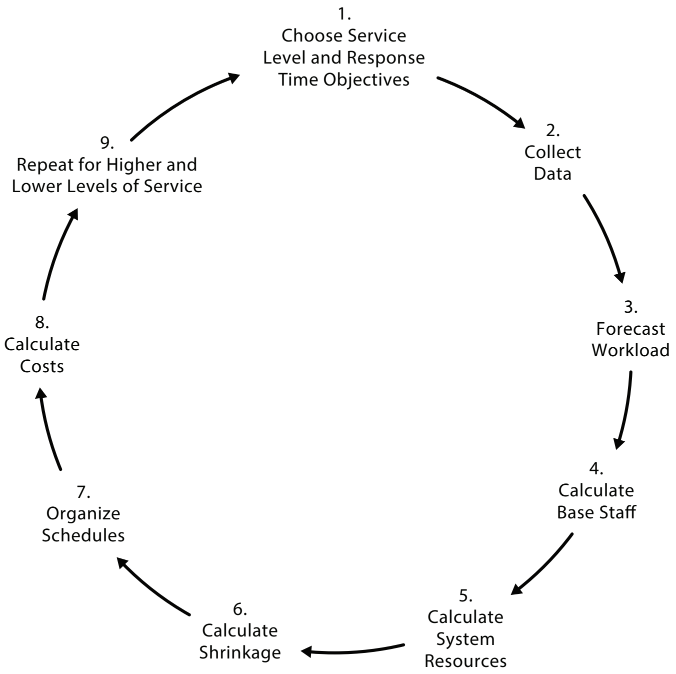
"Price is what you pay, value is what you get."

WARREN BUFFETT

Contact center managers have the responsibility to succinctly (yet adequately) convey contact center resource requirements to senior-level management. That can be quite a balancing act. There's a lot going on in most contact centers, and simplified budgetary requests and summary reports can gloss over important details. Complex budgets and reports filled with pages of numbers may provide ample information, but senior managers may not have the time, inclination, or expertise to make sense of them.

In short, conveying requirements effectively is critical to success. Just as important as the information itself is establishing good lines of communication, and fostering an understanding of how contact centers operate and how they support the organization's overall mission.

In this chapter, we finish the planning process (steps 8 and 9). We'll summarize what senior-level managers need to know about contact centers. We'll then identify essential principles of budgeting, and how (step by step) to determine and communicate long-term staffing requirements.



What Senior-Level Managers Should Know About Contact Centers

To fulfill their potential, contact centers need commitment and involvement from the top. A first step to getting necessary support is ensuring that senior-level managers understand the unique contact center environment–what they do and how they operate. Here's a list of 10 "must knows" that I believe are a good starting point for understanding the nature of contact centers. I encourage you to take stock of these and look for ways to boost your senior management's understanding.

1. CONTACT CENTERS ARE INCREASINGLY IMPORTANT TO THE ORGANIZATION'S SUCCESS.

They are strategic assets, not clerical/administrative/backroom operations. They are hubs of communication–vital to understanding and serving diverse customers, capturing marketplace intelligence, and harnessing the voice of the customer to improve products and services.

2. CONTACTS "BUNCH UP."

In any center that handles at least some inbound work, the workflow dynamics are unique (see Chapter 3). Customers decide when and how they will contact the organization, and the resulting work will not arrive in a nice, even flow. Staffing and productivity issues must be considered in that context (see Chapters 7 and 14).

3. THERE'S GENERALLY NO INDUSTRY STANDARD FOR ACCESSIBILITY.

No single service level or response time objective makes sense for every contact center. Different organizations will have different costs, customers and brand objectives. However, there are objectives that will make sense for your organization–that fit your customers' needs and your organization's brand (Chapter 4).

4. THERE'S A DIRECT LINK BETWEEN RESOURCES AND RESULTS.

You may need 36 people handling contacts to achieve a service level of 90 percent answer in 20 seconds, given your customer workload. It's not going to work if you have only 25 people and are told to hit a 90/20 service level. And scrimping on staffing is expensive, leading to high agent occupancy, burnout and turnover, unhappy customers, poor word of mouth, and other direct and indirect costs (Chapters 7, 9 and 13).

5. WHEN SERVICE LEVEL IMPROVES, "PRODUCTIVITY" DECLINES.

Productivity is often measured as contacts handled or occupancy. (This is a perspective I hope to help change; see Chapter 14.) As discussed in Chapter 9, when service level goes up, occupancy goes down, as does the average number of contacts handled per agent. Translation: in any center that is achieving a good service level, agents will be waiting (idle) some of the time, given the nature of random contact arrival (Chapter 9).

6. YOU WILL NEED TO SCHEDULE MORE STAFF THAN BASE STAFF REQUIRED.

Schedules should realistically reflect the many things that can keep agents from handling contacts, such as training, breaks, holidays, collateral or ancillary work and other diversions (see Chapter 8). In many organizations, these factors are becoming more prevalent, as the increasingly complex environment requires more time for training, research, and other activities.

7. SUMMARY REPORTS DON'T GIVE AN ACCURATE PICTURE OF WHAT'S REALLY HAPPENING.
Reports that show averages of activity may suggest that performance is just fine, yet they may be concealing serious problem areas. Those producing and interpreting data must know what they're really looking at (see discussions in this chapter and Chapters 4 and 12).

8. QUALITY AND SERVICE LEVEL WORK TOGETHER.
Though they are sometimes presented as tradeoffs, service level is inextricably tied to getting contacts into the center and handled in a quality fashion. And better quality is the key to a better service level, by upping first-contact resolution, reducing repeat contacts, and picking up intelligence ("knowledge") that helps improve processes, products, and services across the organization (Chapters 12, 13).

9. CONTACT CENTERS ARE BECOMING MORE COMPLEX.
Traditional transaction-oriented centers have evolved into more dynamic and holistic operations that contribute to and require the support of departments across the organization. Social media, omnichannel, multi-generational customers, competition, AI-driven self-serve technologies that handle more routine activities, and other trends are raising the bar (see Chapters 2 and 15).

10. TO FULFILL THEIR POTENTIAL, CONTACT CENTERS NEED SUPPORT FROM THE TOP.
They need commitment and involvement from senior management to ensure that they get the support and resources they need, and in turn they deliver maximum strategic value.

I am convinced that the only way to really understand the unique customer contact environment is to spend some time in it. These are issues you'll need to continually reinforce–but they tend to come to life when experienced firsthand. Senior-level executives who have made the effort to understand contact center issues and processes invariably come away with better insight into evolving customer requirements and interdependencies across the larger organization.

HANDS-ON LEADERSHIP

One of the keys to high levels of employee engagement in contact centers—and the strong performance that follows—is hands-on involvement from the top. For example:

- Among other endeavors, Dan Gilbert is the founder and chairman of Quicken Loans, the giant Detroit-based mortgage lender. Gilbert speaks often of Quicken's core values, saying that they "drive every decision, every action, behavior, and prioritization." As he grew the company, Gilbert made a practice of spending an entire day with groups of new customer service employees (a responsibility now passed to others in top management). As of this writing, J.D. Power's Customer Satisfaction Study has listed Quicken as the highest-ranked mortgage servicer for five consecutive years.
- When Mary Barra took over as CEO and chairman of General Motors (GM), she inherited an extreme challenge. An ignition switch fault had led to more than 100 deaths and the recall of more than 2.6 million vehicles. She took accountability, met with affected families, set up a compensation fund, and communicated directly and honestly with employees. "Employees saw Barra in call centers taking calls and listening in, and speaking with employees," recounts Jeanne Bliss in her insightful book, *Would You Do That to Your Mother*. "Barra's courage gave her entire company the values they are to uphold, when unity and solidarity in her organization mattered the most." (See Chapter 15, GM Leverages AI in Social Customer Care.)

Principles of Effective Budgeting

Ensuring that you're getting necessary resources is an important part of enabling the contact center's strategic potential. That, of course, requires an effective budget–and a clear understanding of what the returns on those investments should be.

A budget is simply a summary of proposed or agreed-upon expenditures (costs) for a given period of time, for specified purposes. Sounds easy enough. But the process of putting together a budget is often seen as tedious, time-consuming and, some say, a distraction from "more important management responsibilities." However, don't forget the outcome of this much-maligned process: the funding the contact center needs to accomplish its mission and potential.

Here are the essential principles I've uncovered in analyzing and working with contact centers that consistently get the right amount of funding, at the right times, for the right things:

VIEW THE BUDGET AS MUCH MORE THAN A DOCUMENT.
Those who picture rows and columns of line items and figures when they think "budget" are missing the point. It's really a communication process that presents a larger opportunity to learn about the business and make a case that's a win for everyone (employees, customers, and the business). I've seen managers spend many hours—make that many days—putting the details together, only to have their priorities swept away or diluted in a matter of minutes in the CFO's office. I've also seen powerful (and positive) budgetary agreements happen over coffee, literally on the back of a napkin. Remember, it's the effectiveness of your case, not the detail of your analysis, that matters most.

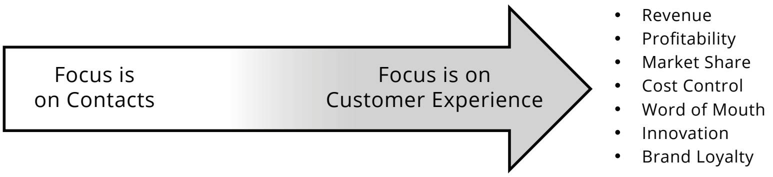
When you see the budget as an ongoing dialogue, and not just a document, you spend more of your time and talent on opening channels of communication, educating decision makers and highlighting key priorities and tradeoffs. In short, you focus on ensuring that the effort leads to the right results.

ANSWER THE BIG QUESTIONS.
Anticipate and be ready for the big questions. Why are we spending this money? Why does the contact center exist? Why are we spending more (or less) than last year? These questions form the backdrop of the budgetary process. The answers are sometimes addressed in the communication that

takes place during the process, and also may be summarized in budgetary documents. Regardless, those who are involved in preparing and approving the budget need a shared understanding of the value the center contributes to the organization.

REMEMBER TO FOCUS ON RESULTS.
Handling 1.7 million calls, achieving 90 percent first-call resolution, or hitting service level targets are not the results decision makers are looking for. They are only means to an end. As your center's objectives and focus mature from handling interactions efficiently to delivering great customer experiences, you will have a greater impact on business results–including revenues, profitability, market share and word of mouth (see figure). Illustrating this connection focuses budgetary discussion on the things that matter most (see Chapters 12 and 13).

Strategic Impact



BASE THE BUDGET ON A CLEAR STRATEGY.
A necessary first step for a successful budgeting process is agreement on the contact center's direction and priorities. Your customer access strategy is the framework that defines how customers will interact with your organization (see Chapter 2). The customer access strategy is the de facto blueprint for the budget–defining who your customers are; when and how they want to reach you; the means by which you will identify, route, handle and track those contacts; and how you will leverage the information that comes from those contacts. Without this foundation, budgetary decisions

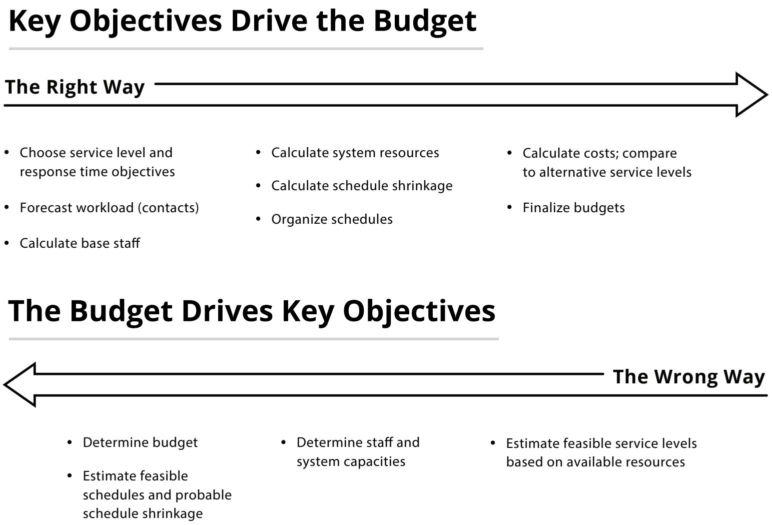
are likely to head off in many unrelated directions and may be at odds with your organization's broader objectives.

ENSURE THAT BUDGETING IS AN EXTENSION OF RESOURCE PLANNING.
In well-run contact centers, forecasting, staffing, scheduling, and cost analysis are ongoing responsibilities. These activities should take much of the work out of the budget process, because the budget should ultimately be based on the already-established forecasting and planning steps.

> *Forecasting, staffing, scheduling, and cost analysis are ongoing responsibilities. These activities should take much of the work out of the budget process, because the budget should ultimately be based on the already-established forecasting and planning steps.*

There's an important principle at work here. Objectives should drive the budget, not the other way around. If your budget is based solely on precedent (last year's numbers), arbitrary decisions, or anything other than the objectives identified in your customer access strategy and workload predictions, you are at a disadvantage from the start. If that's the case, you've got a great opportunity to reshape assumptions (see figure, Key Objectives Drive the Budget).

IDENTIFY KEY TRADEOFFS.
What happens if the forecast is high? Low? What happens if you provide better levels of service? Worse levels of service? How much would you save/spend if ...? Once the budget for the expected workload and recommended resources is established, it is fairly straightforward to rerun scenarios for both different workload assumptions and alternative service levels (step 9 of the planning process). These scenarios will contribute to good budgeting decisions and will improve the understanding others have of contact center dynamics.



Key Objectives Drive the Budget

The Right Way

- Choose service level and response time objectives
- Forecast workload (contacts)
- Calculate base staff
- Calculate system resources
- Calculate schedule shrinkage
- Organize schedules
- Calculate costs; compare to alternative service levels
- Finalize budgets

The Budget Drives Key Objectives

The Wrong Way

- Determine budget
- Estimate feasible schedules and probable schedule shrinkage
- Determine staff and system capacities
- Estimate feasible service levels based on available resources

LOOK FOR OPPORTUNITIES TO MAXIMIZE CROSS-FUNCTIONAL RESOURCES.

Often, an organization's overall results can be improved by investing more in one specific area. Rather than focus on expenditures in a departmental vacuum, effective budgetary strategy maximizes cross-functional resources.

For example, marketing managers might be willing to provide the contact center with budget to capture and analyze information on consumer trends and expectations. Legal departments are increasingly helping the center make the case for investments that will improve tracking and consistency in handling customer contacts. And product development budget may be directed to the contact center for improved analysis on customer suggestions and input. These possibilities become evident to the degree that relationships exist and collaboration is in place among functional areas.

HIGHLIGHT INVESTMENT OPPORTUNITIES.

As with organizations in general, most contact centers are consistently searching for ways to do more with less. But there's also a place for making some high-leverage investments in sensible and practical areas, including:

- Planning and process improvements
- Selective technology investments
- Management-level education
- Cross-sell and upsell programs
- Focused agent and supervisor coaching initiatives
- Research and development

The key is being selective–to focus on those areas that are most likely to yield a high return on investment.

PRESENT THE BUDGET FORMALLY.
This recommendation may seem to be a contradiction, given the emphasis on collaboration and communication. But a formal presentation can be an important part of the process. To start, it can be the catalyst for getting all decision makers together at one time. (How many times did you answer the same questions for different people last year?) All in attendance will hear the questions and comments of the others, saving time and raising the general level of understanding more quickly.

KEEP THE PRESENTATION SHORT AND UNCLUTTERED.
Use graphs and illustrations where possible. Provide backup material as necessary, such as actual system reports (but not as a part of the main presentation). And sprinkle the conversation with real examples–for instance, “Sarah Johnson, a small-business owner in Seattle and a seven-year customer, was one of the 1,200,000 customers we helped last year. She contacted us because she was concerned that …” Examples bring realities to life. And service tradeoffs become much more relevant when the loyalty and positive word-of-mouth from Sarah and 1,199,999 other customers are at stake.

ANTICIPATE AND PREPARE FOR THE “USUAL QUESTIONS.”
They have come up a jillion (give or take) times before, and they will come up a jillion times in the future:

- What did we spend on the center in total last year?
- Did it accomplish what we intended it to?
- What was our ROI on these investments?
- What's the contact per customer ratio? Sales per customer?
- Is growth in some channels (e.g., chat, social media, self-service) changing the workload for agents? (Reducing? Increasing? Altering?)
- What's our cost per contact? Is it going down or up?
- What are you doing to reduce unnecessary contacts?
- Can we use the resources we have now to handle the expected workload?

There are others, and you probably know what they are in your situation. Be ready. These questions are your opportunity to shine. Some may be quite relevant, some less so–but having a complete grasp of the facts will provide you with credibility throughout the process.

ENSURE THE BUDGETING PROCESS IS HONEST AND RESPONSIBLE. You should be realistic and candid about the recent past and whether or not the contact center has been meeting its objectives. The budget must put that in context with customer satisfaction, agent performance, and the objectives and funding being proposed. It must support the mission of the organization and dovetail with the roles and requirements of other areas. And it must be transparent about opportunities and challenges.

Yes, effective budgeting requires some number crunching and analysis. But above all, it requires a clear direction, good communication, and a solid understanding of the contact center's needs and strategic contributions. This is a process that will bring your leadership, communication skills, cultural savvy and professional expertise to bear. Don't treat it as a once-a-year event. It should be part of a continuous effort. Revisit it often and, as with other aspects of planning, make adjustments as necessary.

"I know we didn't accomplish anything, but that's what meetings are for."

Reprinted with permission.

Growth or Contraction—Plan Accordingly

Larger workloads remain one of the biggest challenges facing many customer contact centers. Yes, even with the growth of social communities and dramatic advancements in self-service capabilities, many centers continue to grow. (Why? One reason is the Econ 101 principle of elasticity. When you improve service, customers will use more of it!) Senior-level management will need to know why these services require more budget, may represent a greater percentage of the organization's expenses, and where the money is going.

An important principle in managing growth is to do an analysis of its likely impact in advance. The objective is to avoid surprises as you go into the budget process.

Accurate growth projections often take the form of a document that illustrates projected costs and timeframes, such as 5 percent growth in workload, 10 percent growth, 20 percent growth, and so on (up to at least double the current size if you're growing quickly). Your analysis should consider each major contact center component and answer important questions such as: When will you need additional ACD and IVR capacity? More space? Additional supervisors or analysts? What is the ideal lead time for each increment of growth? How long does it take to recruit, hire and train agents?

Contraction is also a planning challenge. Even as many grapple with growth, long-established contact centers in some industries have closed or reduced in size. For example, hotels and airlines have successfully encouraged a large portion of customers to use self-service systems for inquiries, bookings, check-in, and upgrades. In these cases, plans and budgets must anticipate how contact centers can be scaled down as workload drops. React too slowly, and expensive and unnecessary resources drive up costs. Cut too quickly, and service will be poor.

Because the document is a projection, it won't precisely predict required resources. But it will illustrate required lead times and key decision points necessary to align resources with workload. Your goal is to help your organization avoid costly surprises.

Long-Term Staffing Requirements

For most contact centers, staffing makes up between 65 and 75 percent of the budget. These figures can vary greatly depending on salaries, technology investments, cost differences by region, and other factors. But it's safe to conclude that this one slice of the budgetary pie usually exceeds all other costs combined. Getting it right is a make-or-break factor in the center's efficiency and effectiveness.

Let's take a look at the basics of longer-term staff planning. As discussed in Chapter 8, effective scheduling depends on both longer-term budgets and short-term execution. You'll need a big enough bucket of resources to work with—in other words, the right number of staff on payroll (or through contracts) to put together schedules that match workload requirements. You'll also need to manage schedule adherence, a subject we'll discuss in Chapter 14.

A long-term staffing plan (sometimes called budgetary staffing plan) generally represents staffing requirements at a monthly level for the next 12 months. The goal is to accurately predict the paid hours required to handle the workload at your target service level and response time objectives. The best long-term plans are set up so that they are easily adjusted and clearly demonstrate the "whys" behind budget requirements.

Projections should be based on your workload forecast and required staff, accounting for "availability factors" that keep agents from handling the work. Staff availability can be grouped into three categories:

PRESENCE. Is the agent working today (i.e., is he or she in the building or connected remotely)?

UTILIZATION. Is the agent scheduled to handle customer contacts?

RANDOM. Is the agent actually handling a contact?

Let's walk through a staffing example that accounts for each of these categories and leads to the number of full-time equivalents (FTEs) required to handle the workload. Here, we'll look at a month, which will provide the template for projections you'll normally be making, which go out 12 months or more.

Note that throughout the example, rounding variations can produce slightly different totals and results. Also note that we'll consider the two types of work: service level and response time. You can apply the model to the mix of the channels you handle (e.g., if you need to build budgets for different divisions or agent groups).

Begin with Workload

The workload forecast is the primary driver of staffing needs. Workload includes the projected volume of contacts multiplied by average handling time. The result is then converted into staff hours required. Let's say your July projections for service level-type contacts are as follows (we'll factor in response time contacts later):

July Workload—Contact Load

Item	Projection	Rationale
Contact volume	89,857	Based on forecasts
AHT (sec.)	210	Based on forecasts. Typical since last system upgrade.
Workload (hours)	5,242	Calculation: (volume × AHT) ÷ 3,600 (sec. in an hour)

So, you have a projected 5,242 hours of workload to handle in July. (If you're remembering from Chapter 6 that average handling time often varies increment by increment–you're right. This estimate is a broad brushstroke used for longer-term staffing calculations and is based on the number you're most likely to see on average over the month. It's okay to use it this way for longer-term budgeting purposes–just don't try to base half-hour-by-half-hour staffing calculations and schedule requirements on an average!)

Identify Availability Factors

Next, you'll calculate agent availability factors, beginning with presence. The most typical variables that will keep agents from working are vacations, absenteeism, leaves of absence, disability, and holidays. They might be as shown in the table, Availability: Presence, for the month of July.

According to the calculations, you'll lose an estimated 16.52 percent of paid hours to these factors. Agents will be at work 33.39 hours out of the 40-hour workweek (83.48 percent).

Availability: Presence

	Projection			
Item	In units at left	As % of paid time	Rationale	Percent of paid time calculation
Holidays (days per agent for the month)	1.0	4.35%	As per holiday calendar	1 ÷ 23 (number of paid days in the month)
Disibility (days per agent for the month)	1.20	5.22%	Based on past history for this month	1.20 ÷ 23
Vacation (days per agent for the month)	1.60	6.96%	Based on past history and vacation policy	1.60 ÷ 23
Total absence	3.80	16.52%	Sum of presence factors. Presence will be 83.48% (100% - 16.52%)	3.80 ÷ 23

Next, you will project utilization, which includes all of the things that keep your agents from handling contacts even though they are at work: breaks, meetings, training, and various projects. These variables are illustrated in the following table.

Note that lunch is missing from the list. Because (in our illustration) it is not paid time, it is not included in this model. Also, the factor used for breaks is adjusted for "presence" (you shouldn't count breaks for agents not at work).

Consequently, if agents are not at work 16.52 percent of the time (meaning they are at work 83.48 percent of the time), the factor would be 30 minutes (time on breaks) divided by 480 minutes (minutes in a workday), multiplied by 83.48 percent. Breaks as a percentage of paid time is therefore 5.22 percent and not the usual 6.25 percent many managers associate with breaks.

(Note: Training and coaching percentages are not adjusted by the presence factor, because these activities will be rescheduled when missed due to absence.)

Availability: Utilization

	Projection			
Item	In units at left	As % of paid time	Rationale	Percent of paid time calculation
Breaks (minutes per day)	30.00	5.22%	Per work rules	(30 minutes ÷ 480 paid minutes per day) × presence factor of 83.48%
Meetings (hours per month)	3.00	1.36%	Three one-hour meetings per month	3 ÷ 184 (number of paid hours in the month) × presence factor of 83.48%
Training (hours per month)	2.50	1.36%	Required for new application training	2.5 ÷ 184
Coaching (hours per month)	2.00	1.09%	Two one-hour coaching sessions per month	2 ÷ 184
Committee (hours per month)	2.00	0.91%	As requested by VP of Customer Service	(2 ÷ 184) x presence factor of 83.48%
Total non-contact-related utilization (hours per week)	3.97	9.93%	Out of a 40-hour week, an agent will be utilized for non-contact tasks an average 3.97 hours (40 × .0993)	The sum of all utilization categories

So, you're down another 9.93 percent in total payroll hours to account for variables that keep agents from handling contacts. Added together, presence and utilization factors total 26.45 percent. Put another way, your projections show that agents will be scheduled to handle contacts 73.55 percent of the time (100% - 26.45%).

But you're not there yet. A third category of factors, which can be termed "random," also needs to be included. Don't let the term trip you up—schedule adherence, which is in the example below, isn't random in a mathematical sense like random contact arrival, as you can cause a positive impact on schedule adherence (see Chapter 14). But while you can accurately predict the total amount of time that will go to these factors, they are random because you cannot predict the minute-to-minute impact. This inability to control the timing of these events is what separates them from activities like breaks, meetings and training.

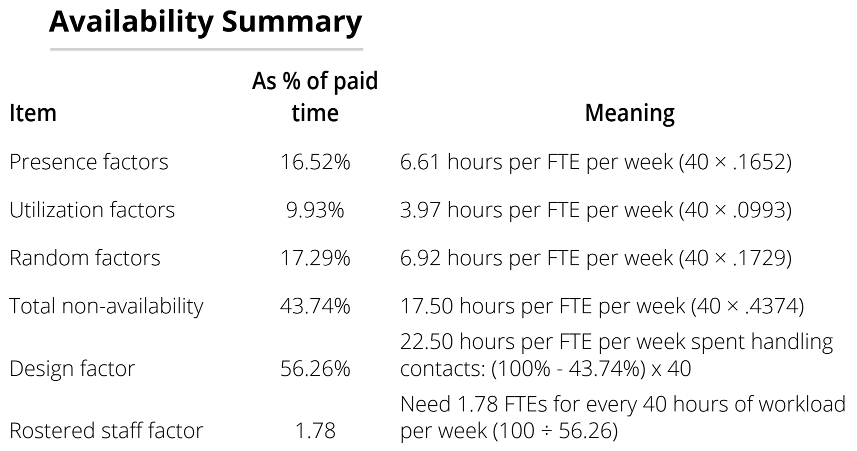
Availability: Random Factors

	Projection			
Item	In units at left	As % of paid time	Rationale	Percent of paid time calculation
Adjustment for adherence (90% of scheduled time handling contacts)	10%	7.36%	As per objectives and past history	10% × scheduled rate of 73.55%
Adjustment for occupancy (85% of time in AHT)	15%	9.93%	85% occupancy is from Erlang C calculations based on volume, handle time and service level objectives	15% × manned percent of 66.20%
Total random	6.91	17.29%	Out of a 40-hour week, agents will lose an average 6.91 hours to random factors	The sum of all random categories

In the example, you subtract adherence time from your scheduled rate of 73.55 percent because you do not want to double-count time for agents not handling contacts. Following this logic, you'll also remove adherence time from scheduled time when calculating occupancy so that you do not include hours adjusted for schedule adherence in the occupancy rate.

The expected occupancy rate is determined by running enough Erlang C calculations (based on expected volume, average handling time and service level scenarios) to feel comfortable you've identified a "typical" occupancy rate. (If you're recalling from Chapters 7 and 9 that occupancy varies increment by increment, you're right! As with average handling time, this estimate is a broad brushstroke used for longer-term staffing calculations and is based on the number you're most likely to see over the course of the month.)

You have now identified all of the factors keeping your agents from han-



Availability Summary

Item	As % of paid time	Meaning
Presence factors	16.52%	6.61 hours per FTE per week (40 × .1652)
Utilization factors	9.93%	3.97 hours per FTE per week (40 × .0993)
Random factors	17.29%	6.92 hours per FTE per week (40 × .1729)
Total non-availability	43.74%	17.50 hours per FTE per week (40 × .4374)
Design factor	56.26%	22.50 hours per FTE per week spent handling contacts: (100% - 43.74%) x 40
Rostered staff factor	1.78	Need 1.78 FTEs for every 40 hours of workload per week (100 ÷ 56.26)

dling the workload. Next, you can convert that into a rostered staff factor, as illustrated.

All of the factors keeping your agents from handling the workload total 43.74 percent. Consequently, agents are projected to spend 56.26 percent of their time (100% - 43.74%) actually handling contacts. This is converted into a longer-term rostered staff factor of 1.78, which is the ratio of staff needed on schedule divided by staff needed to handle the workload (100% ÷ 56.26%).

Convert to Full-Time Equivalents (FTEs)

Using full-time equivalents (FTEs) instead of headcount will allow you to accurately account for part-timers; so the final step in determining required staff is to convert these figures into FTEs. If a full workweek is 40 hours, one full-time employee working 40 hours is one FTE. Two part-time employees, working 20 hours each, would equal one FTE, as would four employees who work 10 hours each.

To convert the workload to FTEs, multiply the workload hours by the RSF and divide by the number of hours per month worked by a full-time employee. For example:

Service Level FTEs Required

Item	Amount
Workload hours	5,242
Staff ratio	1.78
Staff hours required (5,242 × 1.78)	9,331
Staff hours per FTE for the month	184
FTEs required (9,331 ÷ 184)	50.71

Going through a similar process for non-real-time (response time) work might produce the following:

Response Time FTEs Required

Item	Amount
Workload hours	1,365
Staff ratio	1.51
Staff hours required (1,365 × 1.55)	2,061
Staff hours per FTE for the month	184
FTEs required (2,061 ÷ 184)	11.20

Adding phone and email FTE requirements yields a total of 61.91 FTEs, as shown.

Total FTEs Required

Item	Amount
Service level FTEs required	50.71
Response time FTEs required	11.20
Total FTEs required	61.91

The end result of this part of the model is calculating the number of agents required on payroll to handle your planned workload and achieve your service level and response time objectives. Since this number often does not match the current staffing in the center, I recommend going one step further and incorporating a staff planning component that illustrates gaps between the required and the current headcount.

The staff planning section includes current staff, turnover and new-hire information. It also factors in part-time employees and shows how close your current staff comes to your required staff. Hiring plans are often produced months in advance (perhaps by someone outside the contact center) around general business trends. This section allows you to assess and adjust hiring plans a so they match workload needs as precisely as possible. For example, before going through this final step, your hiring activity might produce the sample comparison of required FTEs versus planned FTEs.

Hiring Plan (Before)

	July	August	Sept	Oct	Nov	Dec
FTEs required	61.91	55.00	62.50	64.10	65.05	61.00
Planned staff						
Starting FTEs	52.00	68.92	56.56	57.87	59.13	62.36
Attrition %	4%	4%	3%	3%	3%	3%
Net FTEs	49.92	56.56	54.87	56.13	57.36	60.49
New-hire FTEs*	9	0	3	3	5	4
Planned FTEs	58.92	56.56	57.87	59.13	62.36	64.49
FTE +/-	-2.99	1.56	-4.63	-4.97	-2.69	3.49

* Agents released from training; hire date (e.g., 60 days prior) must allow for training and nesting.

In the example, there are two months (September and October) where you will be understaffed by three or more FTEs, and one (December) where you are overstaffed by more than three. Since your goal is to keep your actual staff numbers as close as possible to required numbers, you can adjust staffing plans (represented by the new-hire FTEs in the next table) to reduce the over/under. The results might be as shown.

Hiring Plan (After)

	July	August	Sept	Oct	Nov	Dec
FTEs required	61.91	55.00	62.50	64.10	65.05	61.00
Planned staff						
Starting FTEs	52.00	58.92	56.56	60.87	65.08	63.13
Attrition %	4%	4%	3%	3%	3%	3%
Net FTEs	49.92	56.56	54.87	59.04	63.13	61.24
New-hire FTEs*	9	0	6	6	0	0
Planned FTEs	58.92	56.56	60.87	65.08	63.13	61.24
FTE +/-	-2.99	1.56	-1.63	0.98	-1.92	0.24

* Agents released from training; hire date (e.g., 60 days prior) must allow for training and nesting.

The new plan keeps every month within three FTEs of requirements. It also reduces total hiring during the six months shown. All in all, it is a better fit to requirements.

Once the plan is created, you are set to make a case for your staffing needs. You have created a model that is fully adjustable at the workload, staffing factor and staff planning levels. It illustrates staffing needs while allowing all stakeholders to quickly see the results of changes in any variable.

I've found that talking through the process line-by-line helps those who are involved understand and participate in the assumptions. As a result, they usually feel much more comfortable about how you reached your requirements. There may be spirited discussion along the way about specific issues. But with line-by-line agreement (and changes that may be merited), one plus one plus three should add up to five—not four or six.

Points to Remember

- Senior management needs (and deserves) a basic knowledge of the contact center; the summary of 10 key principles covered here is good starting point.

- Anticipating the impact of growth (or reduction) of the center’s workload is critical, and it should be a part of the communication and budgeting process.

- The budgeting process should build credibility and clearly demonstrate important tradeoffs and decision points.

- An effective staffing budget is fully adjustable, clearly demonstrates the “whys” behind budget requirements, and enables all stakeholders to easily see the results of changes in any variable.

- In essence, making a case for the resources the contact center needs is communication. It happens best as part of an approach that ensures the right information is presented and understood by all who are part of the process.