1. When Service Level Goes Up, Occupancy Goes Down

As discussed in Chapter 4, service level is expressed as “X percent of contacts answered in Y seconds.” Occupancy is the percentage of time during an interval that agents who are signed in and handling the workload are actually handling contacts. The inverse of occupancy is the time that agents are waiting to handle contacts.

As the following table illustrates, a service level of 80 percent of contacts answered in 20 seconds (82/20 in this scenario) equates to an occupancy of 86 percent for that workload. If service level drops to 24 percent answered in 20 seconds, occupancy goes up to 97 percent.

The relationship between occupancy and service level is often misunderstood. The incorrect logic goes something like, “If agents really dig in, service level will go up and so will their occupancy.” In reality, if occupancy is high, it is because agents are working on one contact after another, with little or no wait in between. Contacts are stacked up in queue and service level is low. In the worst scenario, occupancy is 100 percent because all customers spend at least some time in queue and agents have no breathing room between contacts.

When service level goes up, occupancy goes down (see figure, Contacts Per Agents Versus Service Level). Therefore, the average contacts handled per agent will also go down. Setting standards on number of contacts handled is not recommended, because agents can’t directly control occupancy. Doing so would also conflict with an important objective: ensure that enough agents are available to handle contacts so that your service level objectives are achieved. (We will discuss individual performance standards in Chapter 14.)

Occupancy is driven by random contact arrival and is heavily influenced by service level and group size (see the third immutable law). Managers don’t exactly love this principle—no one wants “unproductive” time baked into the process. However, the time that agents spend waiting for contacts is sliced into 12 seconds here, two seconds there, and so on, and is a factor of how contacts are arriving.

In many centers, agents handle other types of work when the inbound workload slows down. Blended environments (where agents handle different types of work based on workload requirements) make sense—no one has a perfect forecast, so schedules don’t always match staff to the call load. (As a rule, successful blended environments don’t switch agents from one type of work to another minute-by-minute; that’s simply too hard for any person to juggle and quite honestly inefficient. These blended environments do, however, make the switch for larger blocks of time: a few hours of this and a few hours of that.)

But understand what is really happening. When other work is getting done, either: a) you have more agents scheduled than necessary to handle the workload at your service level goal; or b) service level is being sacrificed. Don’t try to force occupancy higher through non-contact work than what base staff calculations predict it should be.

When Is Occupancy Too High?

As any agent knows, periods of high occupancy are stressful. Studies suggest that agents

begin to burn out around 90 percent occupancy if the condition lasts for an extended time, such as several half hours in a row (some studies set the threshold as low as 88 percent, others as high as 92 percent). Taking breaks is a natural reaction to high occupancy, but it compounds the problem.

Consider this scenario. Jen, Ben, and Mary are three of 32 agents plugged in and taking calls. Staffing calculations (see table above) predict the average occupancy for the half hour for 32 agents will be 91 percent and service level will be just above 60 percent answered in 20 seconds. This is what these three agents may be thinking:

Jen: Whew! It’s contact after contact this morning. I need a breather. I don’t have a scheduled break for 45 minutes, so I’m just going to grab some water for a couple of minutes.

OOPS. Now there are only 31 agents handling the work. If traffic keeps arriving at the same clip, service level will drop and occupancy will go up.
Ben: Things are busy today, one contact after another. This customer sure is friendly. I wonder if she's getting that storm I've been hearing about…
So Ben takes longer on the contact, essentially taking a breather during handling time.
Service level drops another notch, and occupancy increases. Mary really begins to feel the load...
Mary: This contact doesn't really require wrap-up, but…
This is the start of a downward spiral. If contacts are chronically backed up, service level will consistently be low and occupancy will be high. The real fix, of course, goes to the fundamentals of managing a contact center—a good forecast, accurate staffing calculations, and schedules that match people to the workload. It's also important that each individual understands their impact (see The Power of One).
Occupancy and Adherence to Schedule
Notice an important distinction that this law reveals. When adherence to schedule improves, occupancy goes down. Why? Because when agents are available to handle more contacts, service level will go up. And when service level goes up, occupancy goes down. This means that if your agents adhere to their schedules, they don't have to work as hard. This is an important concept for everyone to understand.
When adherence to schedule improves (goes up), occupancy goes down.

The terms adherence to schedule and occupancy are often incorrectly used interchangeably. They not only mean different things, they move in opposite directions. And as we will discuss at in Chapter 14, while adherence to schedule is within the control of individuals, occupancy is determined by factors outside of an individual's control.
2. The Law of Diminishing Returns

Economists identified the law of diminishing returns as it applies to manufacturing many years ago, but it also can have significant impact on other environments, including contact centers. It can be defined this way: when successive individual agents are assigned to a given workload, the incremental improvements in service level that can be attributed to each additional agent will eventually decline.

The Law of Diminishing Returns figure is based on Erlang C data from the staffing table at the beginning of this chapter. It shows that 30 agents at the given call load will provide a service level of 24 percent in 20 seconds. Keep in mind, these numbers will not be exact—at that low of a service level, many of the contacts will abandon (not get answered at all). But the exact results aside, service level will be poor.
With 31 agents, things improve dramatically, as service level jumps to 45 percent. Adding one more person yields another big improvement. In fact, adding only four or five people takes service level from the depths to something respectable. That means an associated drop in average speed of answer (ASA) and trunk load.
The same principle is true for larger groups, as the next table shows. Each person has a significant positive impact on the queue when service level is low, even in groups that are much larger.

Contact center managers who struggle with a low service level are fond of this immutable law because it often doesn't take many people to improve things significantly. Those managers who want to be the "best of the best" in terms of service level find that the relationship between varying levels of resources and service level must be clearly outlined in the budgeting process.
Viewed from a different angle, if you have the right number of people handling contacts to begin with, but just a few of them unplug or go unavailable at an inopportune moment, contacts begin to back up. Think of what a stalled car blocking just one lane can quickly do to a busy expressway.

The Power of One

The power of one is among the most important principles to introduce to new hires and reinforce with experienced agents. While we know the impact each agent has on individual customers and the subsequent publicity (good or bad) that can come from those experiences, the power of one refers more specifically to queues and wait times.

The central theme that shapes contact center operations is that they are dynamic; workloads arrive randomly in any center that handles customer-initiated contacts. This, coupled with the reality of how queues behave, means that agents who are helping manage the workload affect service level—in a good way—far more than they may realize.

(The concept of the power of one has been used frequently across the business world, in charitable fundraising, and in the contact center profession, including as the title of an excellent booklet by author Penny Reynolds. However, it was first popularized by Australian author Bryce Courtenay, who used it as the title of his 1989 book about a young boy growing up in South Africa.)

Take a look again at the impact of different staffing levels (this is the same table shown at the beginning of the chapter).

Service level is bad with 30 agents—just 24 percent answered in 20 seconds. With one additional agent, things improve dramatically. Service level jumps to 45 percent (still not great, but almost twice as good). Average speed of answer drops from 209 to 75 seconds. Occupancy goes down, from 97 percent to 94 percent (that might not sound like a big drop, but it feels a lot better!).

Yes, one person makes that much of a difference for customers and the rest of the team!

Adding one more person yields another big improvement. As you can see, if you let your eyes follow the rows down the table, there's a point at which adding agents doesn't help much, because service is already good. You get into the law of diminishing returns.

Next, look at what happens to customers at different staffing levels (see figure, Customer Delay). If you have 34 agents handling contacts, 65 customers are waiting five seconds or longer. Seven customers reach agents in the next five seconds, so 58 are still waiting ten seconds or longer. Another six customers reach agents in the next five seconds, leaving 52 waiting 15

seconds or longer, and so forth. There's still one customer waiting 180 seconds, and no customer waits more than four minutes. It's a very different story, however, if there are only 30 agents handling calls. Dozens of customers are waiting four minutes or longer. The results look far better when just one additional agent is added.

Erlang C for Contact Centers: Customer Delay

Just remember, when queues back up, everybody makes a big difference. Each person has a significant positive impact on customer wait times—which goes far beyond the customers they serve directly. Knowing about these dynamics helps agents understand why schedules are a big deal and why schedule adherence matters.

Experienced customer service employees will correctly point out that the power of one also has a qualitative aspect. Just consider the ripple effect of customer reviews or publicity (good or bad) that can come from any interaction. An American Express study found that consumers tell 21 others on average about a poor service experience. We've all seen videos of bad experiences that went viral. And positive word of mouth builds powerful brands. When you give a good customer service experience, you're creating a powerful marketing force for your company.

The power of one principle is as important as ever, given ever-expanding contact channels and heightened customer expectations for quick and easy service. My encouragement is to keep it front and center with your team!

Here are some of the steps organizations are taking to reinforce the power of one. Think through how you could approach them with your team.

■

Educate each person on how much impact he or she has on the queue—and incorporate these or similar scenarios into training.

■

Develop reasonable expectations for adherence to schedules, and explain the reasoning behind those expectations.

■

Educate everyone on the core steps involved in forecasting and resource planning, so that they know how schedules are produced and where they come from.

■

Provide real-time queue information to agents on readerboards, desktops or phones. (We'll discuss how to use this information in Chapter 11.)

■

Develop appropriate priorities for the full range of tasks that your team handles and guidelines for how to respond to evolving conditions.

***

UNDERSTANDING HOW CONTACT CENTERS BEHAVE

The central reality that shapes contact center operations is that they are dynamic in the truest sense. Because of the rand
arriving nature of customer contacts, each person has a big impact on the organization's responsiveness. That, in turn, is
important enabler to delivering great customer experiences, boosting loyalty and contributing to successful business resu
A notable trend among the most effective contact centers is to educate their entire teams (agents, supervisors, managers
analysts, as well as colleagues from across their organizations) on contact center dynamics, the power of one, and the va
organization delivers when strong cross-functional support is in place. American Express, USAA, and FedEx are just a fe
examples of highly rated companies that have made this an ongoing priority.

Providing an understanding of how contact centers behave is a gift to those who work in or support them. It just makes go
sense.

***

3. Larger Groups Are More Efficient

Average group productivity (contacts that a group handles) is not a constant factor. Instead, it's constantly fluctuating as the workload ebbs and flows. Even when you maintain a consistent service level through good planning and on-target scheduling; you'll find that average productivity is relatively lower at lower volumes and relatively higher at higher workloads. Because the number of contacts is changing throughout the day, so is average group productivity. Why? Mathematically, larger groups of agents are more efficient than smaller groups, at the same service level. Therefore, larger groups assigned to heavy mid-morning traffic will be more

efficient than smaller groups handling the lighter evening load. So, calculating staff the wrong way—assuming fixed productivity—will be highly inaccurate.

The Impact of Group Size

This is yet another reason why setting standards on the number of contacts that agents handle is an unfair way to measure productivity. Attempting to compare groups or sites in a multi-site environment may also be misleading (the exception would be a network that finds the longest-waiting agent, a true virtual contact center).

Despite mathematical efficiencies, there is a point where groups become so large that occupancy becomes too high for agents. Some managers believe that the number of agents in a single group should be limited to 125 to 150 people. However, plenty of centers have much larger agent groups (the U.S. Social Security Administration, Centrelink Australia, China Mobile and others have hundreds or even thousands of agents in a single agent group).

Rather than establishing a strict limit to group size, a better approach is to watch occupancy

and take appropriate measures when it climbs above 90 percent. For example, in the scenario presented under law #2, scheduling so that 129 agents or more are handling the work is recommended, even though the required service level may be exceeded. Customers sure won't mind, and your staff will be able to function efficiently throughout their shift.

4. The Powerful Pooling Principle

The powerful pooling principle is a mathematical fact that goes like this: any movement in the direction of consolidation of resources will result in improved traffic-carrying efficiency. Conversely, any movement away from consolidation of resources will result in reduced traffic-carrying efficiency. Put more simply, if you take several small, specialized agent groups, and effectively cross-train them and put them into a single group, you'll have a more efficient environment.

***

The Powerful Pooling Principle

■ Handle more contacts, at the same service level, with the same number of agents

■ Handle the same number of contacts, at the same service level, with fewer agents

■ Handle the same number of contacts, at a better service level, with the same number of agents

***

Note, again, "The Impact of Group Size" table, which compares service level to group size. Fifteen agents are required to provide a service level of 80/20. But only 124 agents, not 150, are necessary to handle a load 10 times as large.

The pooling principle should be a consideration from the highest levels of strategic planning (How many centers should you have? How should agent groups be designed?), down to more tactical decisions related to real-time adjustments or how to best invest training time and resources.

In one sense, pooling resources is at the heart of what ACDs and networks do. A clear trend in recent years, though, is the recognition that customers often have different needs and expectations, and that different agents with a mix of aptitudes and skills are required. Powerful capabilities, such as skills-based routing, give us the means to route and handle contacts based on myriad criteria (see Chapter 7).

Can you have specialization without forgoing the benefits of the powerful pooling principle? It depends. Skills-based routing can boost efficiency by getting contacts to the agents best suited to handle them. But when not managed well, or when overused, the number of contingencies can multiply beyond your management team's ability to understand and manage them. The interplay can become stupefying. And the whole notion of agent groups and pooling begins to erode. When skills and routing priorities become too complex, related dangers begin to emerge. Doug Casterton, Head of Global Workforce Planning and Scheduling for Trip Advisor in Oxford, U.K., warns of the "eversinking queue." When skill priorities have been set a different level, "it's possible for the higher priority contacts to jump the queue, and if you have not correctly staffed, the lower priority contacts may never actually reach an agent."

As real and pervasive as the pooling principle is, it is not an all-or-nothing proposition. There is a continuum between pooling and specialization—think of a variable thermostat rather than an on/off switch. Your objective should be to get as close to the pooled end of the spectrum as

circumstances allow. Examples of supporting steps would include:

■ Improve training and coaching, to enable agents to handle more contact types.

- Hire multilingual agents to better cover all supported languages.
- Integrate new channels—such as social media or chat—into existing agent groups as much as feasible.
- Improve knowledge management systems so that agents have the information needed to handle a broad a range of contacts.
- Work on reducing turnover and improving agent tenure (their experience levels).

You get the idea. These and other steps you can take to effectively broaden the work types agents handle will, by definition, boost efficiencies.

5. Add Staff and ASA Goes Down

Anyone who has ever waited in line for anything knows that if there were a few more tollbooths, open check-out aisles, or people behind the counter, the line wouldn't be so long! And when someone behind the counter gets reassigned to another task or goes on break, the wait increases (this happens anytime I enter a physical line—perhaps it's a cosmic joke on those of us who study queues).

The same principle is at work in contact centers. When more agents are plugged in and handling contacts, assuming they are trained to do so proficiently, the queue will be shorter. Fewer agents means a longer queue. This principle leads to the next immutable law.

6. Add Staff and Trunk Load Goes Down

When more agents are assigned to a given workload, trunk load (the load on the network that handles voice and data) goes down. The converse is also true: when fewer agents are available to handle a given workload, trunk load goes up because the delay increases (see discussion on trunks, Chapter 7).

Each customer connected to your system is part of the workload, whether they are talking to an agent or waiting in queue. If you have toll-free service (or any other service that charges a usage fee), you are paying for this time. Telecommunications costs are inextricably wrapped in staffing issues. If service level is continually low, the costs of network services will escalate.

The following example illustrates the tradeoffs between staffing levels and service level, average speed of answer, occupancy, and trunk load. Recall from Chapter 7 that trunk load represents how much time (in hours) customers are queued for and/or talking to agents in a particular group over the equivalent of an hour. Staff is calculated for a half hour's traffic, but the trunk load is converted to an hour's traffic ("erlangs") simply because telecom managers universally use hour increments for engineering and management purposes.

Understanding Trunk Load

Want to see where these numbers come from? (Not to worry; it doesn't take long, and you'll get the idea). Using the scenario in the table, assume that you have 46 agents handling contacts and, therefore, will be able to achieve a service level of 80/20. Here's how the calculations produced an estimated 37.6 hours on the trunks:

- First, you can see that customers will be queued for agents an average of 13 seconds (ASA) and will be connected to agents an average of 180 seconds (average talk time), for a total of 193 seconds. The 180 seconds represents the forecast for what average talk time will likely be; the 13 seconds ASA comes from the Erlang C calculation.
- Because the table provides volume for a half hour, double 350 contacts to assume 700 contacts in an hour.
- Because the 700 calls spend an average 193 seconds queuing for and connected to agents, the trunk load in seconds is 135,100 seconds (700 contacts × 193 seconds).

■

Finally, because trunk load is presented in erlangs (hours of traffic over the course of an hour), divide 135,100 by 3,600 (the number of seconds in an hour) and you come up with 37.6 hours. To use the correct telecom lingo, you'll have the equivalent of 37.6 erlangs of traffic on the trunks for this agent group during this time period. (Note: this example does not include the time customers may spend in the IVR before arriving at the agent group, which would need to be

added.)

The big variable is the time customers spend in queue before they get connected. It goes up (gets worse) with fewer agents, and goes down (gets better) with more agents. Glance through the table, and you'll see that if 50 agents are handling calls, ASA will be a projected 3 seconds. If 42 agents are handling calls, ASA will be a projected 144 seconds. In short, the number of agents handling work determines the average delay, which is a key variable in trunk load and, accordingly, in what you pay for toll-free services.

The Impact of Staff on Toll-Free Costs

As you decide on service levels and how to allocate budgets, you should think about network costs, too. All other things being equal, if your service level is low, adding an agent will often bring total costs down because network costs will drop dramatically.

The staff versus toll-free costs tradeoff used to be much more dramatic. In many parts of the world, toll-free costs are just pennies (or less) a minute. Toll-free service used to be much more expensive to organizations (15 cents per minute or higher). So, improving service level meant huge drops in network costs, often producing savings that far surpassed the cost of adding staff. With today's far lower network costs, the tradeoff is less significant.

However, you still have expense for carrying the traffic, and there are also costs related to ports, IVR capacity, maintenance, taxes, and other budgetary line items. Delay takes resources and it is not free. Assessing the impact on network costs underscores the importance of considering both agent and network costs together. Improving service level will save money on network services; these savings should be factored into predictions of overall costs.

The Cost of Delay

The direct expense of putting customers in queue is called the "cost of delay." It is expressed in terms of how much you pay for toll-free service each day (or month, hour, or half hour) just for customers to wait in queue until they reach an agent.

You may want to plot the cost of delay. It's simple. First, take the total delay for the day, as reported by your ACD, and convert that into minutes or hours. Next, multiply the minutes or hours of delay by the average per-minute or per-hour cost of your toll-free service. Then add that figure to a graph that illustrates these costs (see example).

This graph will be a reminder that poor service is not cheap. And it will catch the interest of senior managers, who will look at the graph and wonder aloud, "You mean that's what we're paying just for customers to wait? Why, we could use that money for…"

***

Points to Remember

■ There are immutable laws at work in any center that handles inbound contacts.

■ A common theme runs through these laws: do a good job of matching staff with the workload or both customers and agents will suffer the consequences.

■ The burden doesn't fall solely on those who do the planning and scheduling. Designing and managing a contact center requires a big-picture perspective and the collaborative effort of all involved.

■ A good understanding of these immutable laws is necessary for developing an accurate planning process, setting fair objectives and standards, developing a good strategy, and just about every other aspect of effective management.

■ It's important that agents, senior-level managers, and others who work in or support the

contact center are aware of these principles.

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 s employees (a responsibility now passed to others in top management). As of this writing, J.D. Power’s Customer Satisfac 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 accountabi

met with affected families, set up a compensation fund, and communicated directly and honestly with employees. “Emplo
saw Barra in call centers taking calls and listening in, and speaking with employees,” recounts Jeanne Bliss in her insight
book, Would You Do That to Your Mother. “Barra’s courage gave her entire company the values they are to uphold, wher
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).

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.

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.

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

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 below, for the month of July.

Availability: Presence

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).

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

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

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

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

Going through a similar process for non-real-time (response time) work might produce the following:
Response Time FTEs Required
Adding phone and email FTE requirements yields a total of 61.91 FTEs, as shown.
Total FTEs Required
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)

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)

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.

Chapter 11: Real-Time Management

"The pessimist complains about the wind; the optimist expects it to change; the realist adjusts the sails."

WILLIAM ARTHUR WARD

Okay, the planning is done, the schedules are in place and the contacts are pouring in—far more, or perhaps far fewer, than expected. Uh oh, now what?

Even with good forecasts and accurate schedules, the random arrival of customer contacts means that contact centers operate in a "demand and supply" mode. Demand must be followed by the supply of agents and supporting resources able to handle the workload. If you think about it, any given moment, there are almost always either a) more contacts to be handled than resources available, or b) more resources than contacts. Those rare times that they are in perfect balance (with no waiting agents and no waiting customers) last a fleeting moment or two, before the balance tips in one direction or the other.

Up and down, ebb and flow. Contact centers are dynamic, ever-changing environments.

And then—adding to challenge—there are those times that planning goes off the rails. The forecast misses the mark. Customers behave differently than expected (e.g., that new app was supposed to reduce workload, not increase it!). Marketing sends a social media blast without telling you. The flu is taking a toll on your team. Oh, and IT is working on a desktop upgrade and you are finding that moving … from … screen … to … screen … is … sluggish. Just a little more time, every contact, every agent, all day long.

So, even the most accurate contact center planning must be augmented by effective real-time management: monitoring events as they happen and making adjustments as necessary. Real-time management should complement planning. After the planning is done, it's the moment-by-moment decisions and actions that enable you to maintain an appropriate service level and response time. Effective real-time management includes:

- Empowering your team
- Building a cultural foundation
- Monitoring real-time developments
- Implementing a workable escalation plan

Empowering Your Team

Real-time management is often viewed as a matter of responding to workload quantities. What's often missed in these discussions is how to respond to the nature of the workload. If it's heavy, there usually are underlying issues driving it. You can throw all the resources at it that you want, but if your organization isn't prepared to handle the content of the work, it's going to be a struggle—for customers and employees.

This is an area where empowerment is so important (we'll discuss empowerment further in Chapters 14 and 16). Many organizations want to do the right thing for customers, but too often put a multi-layered, time-eroding approval process in place to get there. By then, the customer is gone, or the loyalty that could arise from their experience has dwindled. That's not effective empowerment; in fact, it's not empowerment at all. Your agents must be able to take action as circumstances unfold. You can't expect them to be effective unless they have the authority and means to make decisions.

For years, The Ritz-Carlton has given staff $2,000 of discretion, per employee and per guest, to resolve problems as the employee feels is appropriate. As a senior manager explains, "Sometimes the most delightful 'wow' moments happen in the blink of an eye. If employees are not empowered and need to cross layers of approval, these moments could be lost forever." (I have seen this in action—while staying at a Ritz-Carlton, I once had a lunch immediately comped when it was delivered later than promised.)

Many executives, understandably, are initially concerned with empowering employees to the extent that The Ritz-Carlton does. But empowerment is actually cost-effective. Employees appreciate the trust and want to make decisions that are right for customers and the organization. And because it's happening on the spot, you are saving resources and aggravation by minimizing the need for managers to review and approve decisions.

The key is to have clear standards and guidance on how to make good decisions. Here are the kinds of questions each employee should be equipped to answer:

- What's the right thing to do?
- What would resolve the problem for this customer?
- What decision best aligns with our values and mission?
- If absolutely necessary, how and to whom do I escalate this issue?
- How should I best capture information and learnings about this issue so that we are equipped as an organization to make improvements going forward?

Training and coaching should be focused on these key questions. Create scenarios and role plays that strengthen judgment and decision-making skills. You won't be able to anticipate and train on every situation that comes along, but you can provide a foundation that leads to good decisions no matter the circumstance.

***

HANDLING TOUGH CONTACTS

How equipped are your agents for those situations when something is going wrong? When customers are clearly upset? are some time-tested tips:

■ If your organization messed up, acknowledge it in a sincere way—and in plain language. (How often as a customer do yo see or hear the scripted words, “We regret any inconvenience this may have caused”?) As writing coach and trainer Lesli O’Flahavan soundly advises, if you wouldn’t say something to a customer face-to-face, don’t use it. (I prefer something lik “Thanks for letting us know we let you down, and for giving us a chance to make it right.”)

■ Some things shouldn’t play out in a public forum, even if that’s where they begin (e.g., through a social media post). Get the customer’s okay to move the discussion elsewhere.

■ Take ownership and resolve the issue—if at all possible, fix it! What would it take to earn back the loyalty of this customer? If the customer is asking for the impossible, at least be prepared to give options.

■ How you frame things matters. When speaking with a customer whose flight was hampered by bad weather, there’s a big difference between “I can’t get you on a flight until tomorrow” and “I can get you out first thing in the morning—would you for me to secure one of those seats?”

■ Document what happened and the circumstances that drove it. Problems can (will!) continue to occur until a root cause is identified and resolved.

How your agents respond to tough situations, as much as anything else, shows the true character of your organization.

***

Building a Cultural Foundation

Along with empowerment, there are principles that, when better understood by your team, will help you build a strong cultural foundation. These include understanding the relationship between service level and quality, the impact of each person, consistency, and others.

Service Level and Quality

One very important principle is the complementary (hand-in-hand) relationship between service level and quality (covered in Chapters 4 and 13). Although service level and quality seem to be at odds in the short term, poor quality will force a negative impact on service level in the longer term—by contributing to repeat and escalated contacts, duplicate or parallel contacts, and other forms of waste and rework. So, the emphasis should be on handling each contact cleanly and correctly, regardless of how backed up the queue is.

Your agents might believe they are getting mixed signals from management: “Hey, you train us to do a quality job, but then you put a lot of emphasis on achieving a service level and response time objectives. You put queue displays all over the place and get unhappy when service level and response times drop. What do you really want?”

“I have no time to hurry.”
IGOR STRAVINSKY

This is not a matter of either/or. Look at the contacts in queue, make sure that you are plugged in (and in the right mode), and do what’s possible to arrange flexible activities around the workload. But remember that each contact must be handled with quality, regardless of what’s happening with service level.

The Impact of Each Person

Remember the law of diminishing returns and the power of one (covered in Chapter 9). The message, as it relates to real-time management, is clear: When the queue is backed up, each person makes a big difference!

This issue sheds light on the importance of training agents on how a queue behaves (including how quickly it can spin out of control) and arming them with real-time information so they can adjust priorities as necessary. Real-time information can be delivered via:

- Queue information on desktop displays
- Wall- or ceiling-mounted readerboards
- Displays on phones
- Mobile apps
- Supervisor monitors
- Flip charts (which sounds so last century, but can help bring focus)

Queue information must be complemented with appropriate training so that agents know what to look for and how to react. Remember that there are only two things directly within the control of agents: being in the right place at the right times, and doing the right things (schedule adherence and quality). A backed-up queue does not mean you should modify the process for handling contacts with quality. Real-time queue information must be interpreted in that context, a subject we'll look at further in this chapter, and in Chapters 12 and 14.

It's important to establish clear expectations around schedule adherence, a subject we'll cover in Chapter 14. Too often, adherence is viewed only as an issue of "how much" (the total time the agent is plugged in) when it's equally an issue of "when" (are they available at the right times?). A key responsibility of supervisors and team leaders is to remove schedule roadblocks so that agents can be in the right places at the right times.

Auto Available and Auto Wrap-Up

Most ACD systems can be programmed for either "auto available" or "auto wrap-up." Auto available automatically puts agents into the available mode after they complete a contact. Auto wrap-up automatically puts them into the after-call work mode.

It usually makes sense to program your system to put agents into the mode they will most often need to be in initially. This can save seconds (which add up over the course of a year's contacts) and minimize the need for agents to manually put themselves into one mode or another. But also give them control over how much time they spend in wrap-up, as the needs of individual contacts dictate.

Some managers program their ACD to put agents into the after-call work mode for a predetermined amount of time. This is a bad idea, as there is no way you can anticipate how much time after-call work will take for any individual contact. Averages don't work here. If your objective is to give agents "breathers" by adding time to after-call work, you are by default adding more time to each contact, further backing up the queue (and defeating your purpose).

Another alternative you have when programming your ACD is to use a feature generally referred to as "contact forcing" (or "call forcing"). This may strike you as an autocratic-sounding term, but it's actually a valuable and agent-friendly capability. With this feature, contacts are automatically connected to agents who are available and ready (thereby eliminating the need for them to manually answer contacts). Agents are notified that a contact has arrived by a gentle "zip tone" (a beep).

Studies indicate that contact forcing can cut four to six seconds from average handling time. And agents almost always like the feature once they get used to it. It removes what would be an extra step. Importantly, they remain in full control. If they aren't ready for a contact, they simply stay out of the available mode.

***

PERPETUATING tHE PROBLEM

Real-time management, though essential, has a serious downside. Real-time tactics that increase supply or curb demand also undermine the organization's ability to create accurate resource plans. These approaches—overflowing contacts to s

groups, postponing breaks and training, or reassigning agents to unplanned work—create skewed activity reports. They c
(delay) essential work or training. And they often complicate future workload and schedule predictions.
In short, real-time management tends to perpetuate the imbalances that created the need for reactionary measures. This
doesn't mean that real-time tactics shouldn't be used, but you should employ them only as needed and be alert to their

implications on planning and management.
***

Consistency
Another important step in building a good foundation is to ensure that agents maintain a consistent approach to handling contacts, regardless of queue conditions. Each agent has an impact on the components of workload, and therefore 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 work that should happen as part of wrap-up. This skews reports, causes planning problems, and can lead to more errors.
The solution is to define which types of work should immediately follow contact handling and which types can be completed later. Then, train agents and supervisors accordingly.
Accurate Resource Planning
Real-time management can never make up for inadequate planning. The nine-step planning process covered in preceding chapters should be as accurate as possible. This includes:
■ Establishing service level and response time objectives that everyone understands
■ Accurately forecasting the workload associated with all types of contacts
■ Calculating staffing requirements
■ Planning for and managing activities not related to handling customer contacts
■ Building schedules that match staff to the workload as closely as possible
Monitoring Real-Time Developments
The third major principle in effective real-time management is to monitor developments and identify trends as early as possible. The key is to react appropriately to evolving conditions. Random contact arrival means that, at times, it will look like you are falling behind even though you are staffed appropriately. But if you are experiencing a genuine trend, you need to move quickly. Time is of the essence.

Interpreting Queue Status
Service level is "rolling" history. The ACD has to look back some amount of time or at some number of contacts to make the calculation. Consequently, even though service level is a primary focus in planning, it is not a sensitive real-time report. (In the strictest definition, it's not a real-time report at all.)
With many ACDs, you can define how far back the system looks to provide real-time service level status. You may need to experiment some. You'll need enough of a sample that reports aren't all over the place, but to be valuable, the sample also must be fairly recent. Also note that "screen refresh" does not correlate to the timeframe used for calculations. Your monitors may display updated information every few seconds, but that has nothing to do with how much data your ACD uses for the calculations that require rolling history.
Service level will tell you what has already happened, given recent unique contact volume, random arrival, average handling time and staff availability patterns. But it's important to realize that what is being reported is not necessarily an indication of what is about to happen.
However, the number of contacts presently in queue is a real-time report, as is "longest

current wait" and "current agent status." Understanding the distinction between reports that are genuinely real-time and those that must incorporate some history explains apparent contradictions.

For example, service level may indicate 65 percent of contacts are answered in 20 seconds, even though there are no contacts in queue at the moment. Keep watching the monitor, though, and service level will begin to climb. Alternatively, service level may look high at the moment, even though an enormous volume of contacts recently entered the queue. Give it a few minutes and, unless circumstances change, service level will begin to drop like a rock.
There will be a delay of at least several minutes before service level begins to reflect the impact of what is happening. So for service level to have meaning, it must be interpreted in light of the recent past, contacts in queue and current longest wait. If you focus on service level alone, you could badly misread the situation.
Unless conditions change, the number of contacts in queue signals where service level is about to go. So—this piece of data should be a primary focus, along with longest current wait. That information should drive you to assess the mode agents are in—signed off, auxiliary, handling contacts, etc.—and make appropriate adjustments.
In summary, focus on real-time reports in this order:
1. NUMBER OF CONTACTS IN QUEUE. This is the real-time report most sensitive to changes and trends. Look at this first.
2. LONGEST CURRENT WAIT (OLDEST CONTACT). This is a real-time report, but it behaves like a historical report (e.g., many contacts can enter the queue, but longest current wait will take some time to reflect the problem). This report gives context to the number of contacts in queue. For example, if there are far more contacts in queue than normal, but longest current wait is modest, you are at the beginning of a downward trend. Now is the time to react.
3. SERVICE LEVEL, AVERAGE SPEED OF ANSWER, AVERAGE TIME TO ABANDONMENT AND OTHER MEASURES OF THE QUEUE AND CUSTOMER BEHAVIOR. These reports provide additional context to the number of contacts in queue and longest current wait. For example, if service level is low, but there are few or no contacts in queue, you have addressed the problem and service level will begin to climb. Don't sweat it.
4. AGENT STATUS. This real-time report indicates how many agents are available and what modes they are in. I generally place agent status after other reports because it can be difficult to interpret unless you know something about the queue. For instance—it doesn't matter if few agents are handling contacts if there aren't many contacts to handle. In that case, you would want agents to be working on other tasks.
In the end, you should monitor and interpret these reports together. With the right training on what real-time information means and the activity it is conveying, experienced agents and supervisors can scan and make sense of the updates quickly.
***

Monitoring Real-Time Reports

1. Number of contacts in queue
2. Longest current wait (oldest contact)
3. Service level/average speed of answer
4. Agent status
5. Escalation plan...
***

Display Thresholds
Most ACD and wall display systems allow you to establish various priority thresholds. For example, you can color-code information yellow when the queue begins to back up and red when it's in really bad shape.

The problem is that the thresholds are often set arbitrarily. Additionally, agents often don't understand what is expected of them at different threshold levels. If that's the case, real-time information will raise everyone's stress level. And your agents might feel as if it's their fault that they can't clear up the queue.

To avoid this, proper programming and training are necessary:

- Generally, the first threshold should be set for one contact in queue. Agents should proceed normally, and no tactical adjustments are required.
- The second threshold should indicate that there are more contacts in queue than the average expected for the desired service level (see "Q2" in Chapter 7). Routine adjustments, such as postponing flexible work, should be made to get the contacts answered.
- The next threshold should indicate that there are more contacts in queue than the agents can realistically handle. In this case, more extensive real-time tactics (e.g., bringing in reinforcements or triaging the work) are required.

You can program many of today's systems to adjust thresholds dynamically as workloads and staffing levels change (10 contacts in queue may be no problem during a fully staffed shift, but would be a nightmare for two agents handling work at 3 a.m.). The most important thing is understanding where the information comes from and what it takes into account, so that those interpreting it can make good decisions.

Considering All Channels

Ever had this happen? Someone looks at the real-time reports and everything is fine, so he or she decides to unplug and begin work on a project or take a break. Unfortunately, others have the same idea at the same time, resulting in queues that quickly spin out of control. Interpreting real-time information when there is no queue is almost as tricky as identifying trends when there is a queue!

Most large or multi-site centers have a designated "traffic control" person (or small team) coordinating activities. The traffic controller's authority can range from making informed suggestions on priorities to flatly dictating what can and can't happen at any given time. This responsibility should include the full range of channels.

There's a balance here, though. If you pull agents from response time work (e.g., email, social media messages that can be deferred) to handle service level (e.g., calls, chat), customers who sent email or social media messages might begin calling. If you don't have coordination across systems, you could be handling the same customer inquiry twice (or more!). In the worst case, the answers given to customers could be inconsistent, creating confusion and further work (and, frankly, making the organization look a bit silly).

Whatever the size of your contact center, you'll need someone with a view of the whole landscape in order to effectively monitor changing conditions. This person may also produce and interpret intraday forecasts (see Chapter 6) and make adjustments to system or network thresholds.

Implementing a Workable Escalation Plan

Regardless of what channels you are supporting, you will need to make appropriate tactical adjustments as conditions change. Examples of real-time tactics include:

- "Everybody help a customer"
- Adjust breaks
- Assist agents stuck in wrap-up
- Postpone or move up flexible work

■
Triage channels and contacts by type
■
Position appropriate announcements
■
Bring in secondary groups
■
Adjust overflow or network parameters
■
Reassign agents to groups that need help
■
Use supervisors wisely
■
Bring in agents who are on call
■
Send contacts to outsource partner

■
Mobilize the SWAT team
■
Adjust contact-routing priorities
■
Take messages for callback
■
Generate controlled busy signals
■
Offer voluntary time off (when slow)
■
Work on projects (when slow)
■
Go through training modules (when slow)
■
Help with back-office work (when slow)
■
Others

An important principle in effective real-time management is to outline a workable escalation plan that is in place before a crisis. Most contact centers use a tiered approach. This list is not exhaustive and not every option will be available. I encourage you to explore the options available to your center. Then prioritize them into tiers for busy times and tiers for slow times and define when you could/would use each.

Level 1

The first level of action involves routine, common-sense adjustments that enable you to get the contacts handled. Agent status becomes the focus, and many use a variation of the time-honored phrase “Everybody help a customer!” This is generally directed toward people on the floor who are not currently handling contacts.

At this level, agents make routine adjustments to work priorities. Flexible tasks are postponed. If you have agents handling contacts that are not as time-sensitive as service level contacts—such as email, outbound calls, or data entry—they can be temporarily re-assigned. You might also automatically overflow contacts to agents in other groups (who are, of course, trained to handle those specific types of contacts).

Make sure that your agents handling calls understand that speeding up their rate of speech

will not help. Customers can usually sense they are being rushed, and will often dig in their heels to slow things down. However, agents shouldn't go beyond what is necessary to completely satisfy the customer's stated objectives and handle the contact with quality. There's a line somewhere, and common sense applies.

Level 2 and Beyond

If the workload still outpaces the staff required to handle it, the contact center can move on to more significant real-time alternatives. For example, it may be reasonable to reassign agents from one group to another.

In omnichannel environments with well-integrated channels, it may make sense to parse out and prioritize channels (e.g., phone or chat over email). This typically happens through automated system changes, based on criteria that you determine and program ahead of time. A related tactic is to prioritize high-value or urgent contacts—though caution is in order to ensure other contacts don't languish too long (see Chapter 9).

In omnichannel environments with well-integrated channels, it may make sense to parse out and prioritize channels.

Another possible Level 2 activity is to change system announcements so that they offload what would otherwise be routine contacts. Utilities use messages such as, "We are aware of the power outage in the Bay Ridge area. We hope to have power restored by 11 a.m. We apologize for the inconvenience. If you need further assistance, please remain on the line and one of our representatives will be with you momentarily."

More routinely, contacts can be directed elsewhere: "If you would like current arrival and departure information, press or say one…" Some centers also give customers the ability to check

the status of an order, find specific product information, or hear answers to commonly asked questions while they wait, without losing their place in the queue. (And, increasingly, centers are encouraging customers to use self-service alternatives.)

Sometimes, you can foster understanding with system announcements: "Due to the snowstorm hitting the East Coast, we are operating with fewer of our associates than normal. We apologize for the delay and will be with you just as soon as possible. Thank you for your patience." This tactic will backfire if it's overused or stretches the truth. (And yes, the message "we are experiencing unusually heavy call volumes …" has been abused and overused.)

You might also be able to improve circumstances by changing routing thresholds between groups or sites. Most of today's routing systems use sophisticated programming logic, based on thresholds you determine in advance. But remember: those thresholds will need to be adjusted as circumstances dictate.

(As with any of these changes, be sure agents are reassigned back to the "default" structure. You'll need to determine how agents are notified of changes. Those responsible for workforce management will also need to understand the impact so that they can adjust data used in forecasts.)

It also may make sense for supervisors to help handle contacts. This can be an effective tactic if used judiciously. But it must be well coordinated, because if no supervisors are available when agents need help, the situation could deteriorate further. (And in some cases, they may not be licensed or fully proficient in the work that is queued.) Additionally, some union agreements restrict supervisors and managers from handling contacts.

Some centers take messages for a later callback, a capability that is greatly facilitated by virtual queue technologies (see Chapter 4). However, this approach doesn't work well in all cases. Potential challenges include: How do you ensure that the callbacks are timely if you're busy now? What is your policy when you reach the customer's voicemail? You may have to experiment to find out whether it's effective in your environment.

Other Level-2 tactics include calling in a SWAT team or bringing in agents who are on reserve (see Chapter 8), routing some contacts to outsource partners, or adjusting the placement

and timing of system announcements (see the next section, The Psychology of Announcements).
In summary, establishing an effective escalation plan involves:

- Identifying feasible real-time tactics (ahead of time)
- Determining the conditions in which each tactic should be implemented (ahead of time)
- Monitoring conditions (real time)
- Deciding on necessary adjustments (real time)
- Coordinating and communicating changes to all involved (real time)
- Implementing the tactics (real time)
- Assessing how well the escalation plan worked (after the fact)

It is wise to “react in advance” (a term I first heard from industry executive Tim Montgomery)—that is, make adjustments ahead of time. For example, if you know by looking at workload trends that your schedules aren’t matching requirements, you can make adjustments to prevent the consequences of that mismatch.
In the example, you have more agents than needed early tomorrow morning and fewer than required later in the morning. By planning ahead, you can schedule elearning modules and adjust some breaks to address the over-staffing in the earlier increments. By rescheduling agents assigned to email, you can cover the shortage later in the day.
Schedule Variance, Tomorrow

This kind of approach makes sense. It’s more limited than a perfect forecast and schedule assembled weeks in advance. But, given the reality that plans sometimes get out of sync with workloads, it’s far better than reacting as the crunch happens. And it underscores an important principle: do everything possible, as far in advance as possible, to align schedules with workload requirements. Advance planning beats in-the-moment reaction every time.
Planning to React in Advance

When you’ve gotten through the crunch, there’s an important (but sometimes neglected) aspect of real-time management: analyzing what happened so that you can prevent recurring problems. How well did your escalation plan work? Were the right tactics deployed? What could you have done differently? This analysis will help you to fine-tune your escalation plan and improve the planning process. It’s especially important if you are responding to crises on a regular basis.
***

COMMUNICATION REQUIREMENTS
Here are key communication requirements that ICMI recommends be part of your real-time management plan:

- Communicating to agents and supervisors their roles in real-time management
- Communicating to agents and supervisors what you expect them to do in various real-time scenarios
- Communicating how often you had to pull various levers
- Communicating why you may not have resumed “normal operations” as soon as the queue cleared
- Communicating with your social media team—giving them a “heads up”; what are they hearing when the contact center i

backed up?

- Keeping a strong line of communication between the floor and the real-time management team

- Communicating root causes of unexpected workload

***

The Psychology of Announcements

Don't forget about delay announcements. Yep, they are still part of most contact centers. The first announcement recognizes customers, provides reassurance they are in the right place, and promises that the contacts will be answered. Many also include something like, "Your call will be answered in the order in which it was received …"

The typical behavior of customers who abandon can provide insight into the use of delay announcements. Customers who abandon when they hear the first delay announcement are referred to as "fast clear-downs." The customer may have dialed the wrong number, changed their mind about calling, or simply decided to bail out when they didn't get right through to an agent.

Some centers have found that repositioning the first delay announcement can lower abandonment. For example, if the delay announcement is normally set to play just after a contact enters a queue, moving the threshold out to provide additional rings can buy your agent group additional seconds to get to customers before they become fast clear-downs. Also, because most customers don't feel that they are in a queue until they hear the announcement, they may wait longer.

Will this help? Maybe. Does it solve serious staffing imbalances? Nope. You'll have to experiment to see if it has a positive impact. And keep in mind, this technique will actually worsen average speed of answer and reduce service level. But you have a higher value in mind: get to as many customers as possible before they give up. For that, you'll want to consider every alternative possible.

You may also be able to reduce abandonment by adjusting the position of a second announcement. For example, if average time to abandonment is 50 seconds and the second delay announcement is set for 60 seconds, you might hang on to more customers by adjusting it to play earlier. This is psychology, pure and simple, and the purpose of the second delay announcement is to give customers who are about to abandon renewed hope that you will get to them: "We apologize for the delay; thank you for continuing to hold…"

The first and second delay announcements are valuable, but you'll need to use good judgment with repeating announcements, or they can make things worse. Keep in mind that customers often have you on speakerphone or are using a headset while they work on other things. (Once shunned by the general public, headsets are becoming de rigueur in the smartphone era.) Remember, they have to mentally tune in every time the announcement is played. Repeating announcements are repeating interruptions. "Yeah, I know this call is important and that you'll be with me momentarily. You've told me eight times so far."

That said, if subsequent announcements are interesting and valuable, they can be helpful in minimizing abandonment. I recall calling an airline to sort out some changes to a complex itinerary. They were backed up due to serious weather-related disruptions in their flight network, and various announcements covered such topics as: baggage policies, what can be carried on, vacation packages, the airline's mobile app, updates to the frequent flyer program, etc. The announcements were all different, and (even though I was familiar with the information) were interesting and well-timed. For this airline in this situation, it was a good approach. When considering a similar approach, let common sense and your company's personality and brand be your guide. (See the discussion on customer access strategy, Chapter 2.)

Extreme Alternatives

What if things are really rough? What if you get more contacts than expected and have exhausted all other alternatives? Should you continue to let customers into the queue? The choices, neither ideal, are to give callers busy signals (or limit "queue depth," as some put it) or let customers enter a long queue, only to abandon anyway.

Some centers could never consider using busy signals. Emergency services are a notable example. But for others, busy signals may occasionally be acceptable. Customers who get busy signals are more likely to make immediate and repeated attempts to reach you than those who abandon.

Busy signals are far less common in today's contact centers, but they remain enticing to some managers because they make reports look better and take the pressure off of agents. Some centers depend on them as a crutch for inadequate staffing. It probably goes without saying, but that's poor customer service, and it defeats the mission of the contact center. This tactic should only be used in extreme or short-lived situations.

Finally, there is the option to end all options: closing the center. During precipitous declines, times of national or world crises, and other such conditions, stock exchanges will suspend trading (the very thing they're there to enable). In the world of contact centers, Zappos (which has built a brand on delivering outstanding service) once temporarily stopped all calls and resorted to email in order to handle the workload deluge caused by a hacking incident that forced the company to change customer passwords.

Yes, this last-resort alternative is completely out of alignment with the contact center's mission to provide access. But a situation that forces you to consider such a drastic approach can happen, no matter how well you've planned. The message here is: do everything you can to prepare for the workloads you will need to handle. Then, if things go awry, deploy the options you have, scaled for the severity of the situation.

Planning and Practice

Contact centers that do a great job of real-time management have some important things in common:

- They plan the escalation procedure ahead of time and define the thresholds that will determine when each tactic should be implemented.
- They continually review and refine their escalation plans.
- When the dust settles, they take the time to go back and analyze what happened: what worked and what didn't.
- They continually improve their planning process so that they are not leaving those on the floor with an impossible mission.

Real-time management takes planning, coordination, and practice. It's also gratifying. One of the rewards of working in a contact center is to step up to a challenge, then be able to look back at the results…and smile.

***

Points to Remember

- Empower your team and establish a strong cultural understanding of key principles ahead of time, so that you are not creating the crises to which you are responding.
- Provide real-time information to agents and supervisors, and train them on how to interpret the information.
- Plan the escalation procedure ahead of time and define the thresholds that will determine

when alternatives are deployed.

- Establish a person or a team to coordinate real-time tactics.
- Review and refine your escalation plan on an ongoing basis.
- Continually improve your planning process. Real-time management will never be an effective substitute for accurate resource planning.