



Service Level Notes with Brad Cleveland

Understanding How Queues Behave

Mastering basic staffing, trunking and service level principles is the key to success with more sophisticated processes.

Do you want to set realistic performance objectives? Or create accurate plans and schedules? Or better understand caller behavior? How about respond appropriately to real-time developments? Or put together realistic budgets?

Having a good understanding of how queues behave will not inherently accomplish these things for you, but it's an important prerequisite. You simply can't manage a call center – at least not very well – without knowing how staff, trunks, caller delay and other variables are related. These dynamics are taking place moment by moment, day after day – they are constantly at work.

One of the best ways to learn about queues is to use a basic staffing calculator to explore variables and run different scenarios. (Alternatively, you can use more powerful tools – e.g., a workforce management system or a computer simulation program. But a simple staffing calculator based on the Erlang C and Erlang B formulas is often the easiest way to become familiar with basic dynamics and trade-offs.)

Input Variables

Since half-hour increments provide an adequate level of detail and accuracy for most call centers, half-hours are the usual basis for calculations. There are four basic variables required to run a staffing scenario:

Average talk time, in seconds: Input the projected average for the future half-hour you are analyzing.

Average after-call work, in seconds: Input the projected average for the future half-hour you are analyzing.

Number of calls: Input the projected volume for the future half-hour you are analyzing.

Service level objective in seconds: If your service level objective is to answer 90 percent of calls in 20 seconds, input 20 seconds. If it's 80 percent in 15 seconds, plug in 15 seconds. In other words, the formula needs the Y seconds in the service level definition, "X percent of calls answered in Y seconds."

Output

Output is shown in Figure 1, below. (Note: In this example, we are using an Erlang C program provided by Incoming Calls Management Institute. Similar Erlang C and Erlang B calculator programs are available from ACD vendors and software companies.)

Here's what the column headings stand for:

Agents: Number of agents required

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

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

ASA (average speed of answer): With 34 agents handling calls, ASA will be 13 seconds. ASA is the average delay of all calls, including the ones that aren't delayed at all. (Note: While 13 seconds is the correct mathematical average, it's anything but a "typical experience," as most callers get through quicker than that and some wait far longer. For this reason, ASA is often misinterpreted without further knowledge of what goes into the average.)

DLYDLY (average delay of delayed calls): This is the average delay only of those calls that are delayed – 43 seconds

Figure 1: Staff Module

Average talk time in seconds: 180 Average after call work in seconds: 30 Calls per half hour: 250 Service level in seconds: 30								
Agents	P(0)	ASA	DLYDLY	Q1	Q2	SL	OCC	TKLD
30	82.8%	208.7	252.0	29.0	35.0	23.5%	97%	54.0
31	65.2%	74.7	114.5	10.4	15.9	45.2%	94%	35.4
32	50.7%	37.6	74.1	5.2	10.3	61.3%	91%	30.2
33	38.8%	21.3	54.8	3.0	7.6	73.0%	88%	28.0
34	29.3%	12.7	43.4	1.8	6.0	81.5%	86%	26.8
35	21.8%	7.8	36.0	1.1	5.0	87.5%	83%	26.1
36	15.9%	4.9	30.7	0.7	4.3	91.7%	81%	25.7
37	11.4%	3.1	26.8	0.4	3.7	94.6%	79%	25.4
38	8.1%	1.9	23.8	0.3	3.3	96.5%	77%	25.3
39	5.6%	1.2	21.4	0.2	3.0	97.8%	75%	25.2
40	3.8%	0.7	19.4	0.1	2.7	98.6%	73%	25.1
41	2.6%	0.5	17.7	0.1	2.5	99.2%	71%	25.1
42	1.7%	0.3	16.4	0.0	2.3	99.5%	69%	25.0

Source: QueueView Staffing Calculator, by ICMI, Inc. www.incoming.com

in this example.

Q1: Average number of calls in queue at any time, including times when there is no queue. The label is somewhat of a misnomer, because Q1 incorporates all calls into the calculation, including those that don't end up in queue.

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

SL (service level): The percentage of calls that will be answered in the number of seconds you specify.

OCC (percent agent occupancy): The percentage of time agents will spend handling calls, including talk time and after-call work. The rest of the time, they are available and waiting for calls. In the example, occupancy will be 86 percent. Notice the trade-off: When service level goes up, occupancy goes down.

TKLD: This column is the hours (erlangs) of trunk traffic, which is the product of [talk time plus average speed of answer] multiplied by number of calls in an hour. Since Erlang B and other alternatives used for calculating trunks often require input in hours, these numbers can be readily used as is. The actual traffic carried by trunks in a half-hour will, in each row, be half of what is given.

(Note: If you are putting schedules together, you will need to calculate base staff for each half-hour of the day and for every unique group of agents – sales, customer service and other types of groups you have. Obviously, a WFM system that runs the calculations for many half-hours is a great time-saver. Our purpose here is to understand queue dynamics.)

Notice some of the dynamics that become evident from these calculations:

- The more agents you have handling calls, the higher service level will be.
- The more agents you have handling calls, the lower trunk load will be.
- The more agents you have handling calls, the lower occupancy will be.

Figure 2: Delay Module

Average talk time in seconds: 180

Average after call work in seconds: 30

Calls per half hour: 250

Service level in seconds: 30

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

Source: QueueView Staffing Calculator, by ICMI, Inc. www.incoming.com

Understanding Delay

A good question to ask for any service level is, “What happens to the calls that aren’t answered in Y seconds?” Some Erlang C and computer simulation software programs will calculate the answers to that and related questions.

In the example, above, 34 agents will result in a service level of 82 percent of calls answered in 20 seconds. Sixty-five callers will wait five seconds or longer. In the next five seconds, seven of those callers reach agents, so only 58 callers are waiting 10 seconds or longer. In the next five seconds, six more callers will reach agents, leaving only 52 callers waiting 15 seconds or more. At this service level, one caller is still waiting at three minutes.

Note two important implications of the principle of delay:

1. Because of random call arrival, different callers have different experiences even though they called during the same half-hour, and even though the call center may be hitting its target service level.

2. Some call centers attempt to set two service levels for the same queue; e.g., to handle 80 percent of calls in 20 seconds and the rest within 60 seconds. Obviously, that is not possible; 80/20 and 100/60 are distinctly different service levels.

Are the Rules Changing?

Demonstrating these basic trade-offs is useful in educating managers outside the call center – e.g., those involved in approving budgets. “What would our service level be if we had five fewer people? What would we need to achieve 90 percent answer in 20 seconds? How many callers would be in queue beyond 90 seconds? How come we can’t have a service level of 80/20 and an occupancy of 95 percent?” These questions are inevitable; be ready for them. (And no, you don’t have to go into the finer details of Q1, Q2 and P(O) with the CFO. But be prepared to illustrate basic trade-offs.)

I’ve heard some people suggest that skills-based routing, CTI-enabled routing, networked call centers and caller prioritization alternatives have changed the rules, and that basic staffing, trunking and service level trade-offs no longer apply. In fact, they apply as much as ever. Like the law of gravity in aeronautics, these principles are always at work. Mastering the fundamentals is key to knowing how to successfully establish more sophisticated applications. CCMReview

Brad Cleveland

Brad Cleveland is president of Incoming Calls Management Institute (ICMI) and publisher of Call Center Management Review. He can be reached at 410-267-0700 (ext. 958), or bradc@incoming.com.