

Jacada White Paper

The Impact of Business Process Optimization in Contact Centers on the Service Level and Staffing

Andre Knüpling
Assistant Vice President Solution Engineering

Today's contact centers must provide the best possible customer service at the lowest possible cost per customer contact. There are several factors that are used to rate the quality of customer service. Typically contact center managers try to improve the first call resolution, the average waiting time or queue length, or the service level, in the sense of percentage of answered calls in a certain period of time. Customer satisfaction and customer retention as a result of high quality service is correlated to personnel costs. It is obvious that more customer service representatives can answer more calls and that a longer (and more expensive and hopefully thorough) training results in better customer interaction. Unfortunately personnel costs are about 70-80% of overall costs per customer contact. That's why operational personnel planning is a key for the economic performance of a customer center. To make a service center effective is a question of finding the best balance between quality and costs while meeting the contact center's goals. This research will focus on the service level as a quality indicator and the number of agents for the expense. Among other approaches, e.g. training intensification or changed call routing, the reduction of the service time has an effect on the overall contact center performance. In fact a comparatively small time saving can have a significant impact on the operational staffing.

THE MODEL

We use some typical measurements to define a model of a customer service center - the average number of customer contacts per hour, the average service time and the aimed service level (percentage of calls that are answered in the first X seconds). We then suppose that in a process optimization initiative we identified inefficiencies in the agent's productivity and that we are able to reduce the service time by a certain amount.

Unfortunately customers are undisciplined. They don't call in constant time periods and service times may also extremely vary (see appendix a1). To get a good estimation of the results from the process optimization we need some mathematical assistance.

ERLANG-C

Queueing theory is, similar to mathematical statistics, a special field of the probability theory. It is used to investigate processes like waiting lines or calls in a contact center. Conceived by A. K. Erlang, a danish mathematician, and at the beginning of the 20th century it became one of the central themes of Operations Research.

Most widely used in contact centers is the Erlang-C model, also called M|M|n system (a2). It allows the calculation of the theoretical distribution of the duration an arriving customer has to wait in queue before being served. In "contact center speak" - the service level. Additionally it gives the average length of the waiting queue. For our service center model this is the number of callers on hold in the telephone system. In turn, using Erlang-C we can calculate the number of agents needed to meet a given service level.



THE NUMBER OF AGENTS NEEDED BEFORE PROCESS OPTIMIZATION

Factors that describe our contact center model in its current situation are:

Arrival rate (a3): $\lambda = 1000 \frac{\text{calls}}{\text{hour}}$; Service rate (a4): $\mu = 12 \frac{\text{calls}}{\text{hour}}$;

Service time (a5): $\bar{t}_B = \frac{1 \text{ hour}}{12 \text{ call}} = 300 \text{ seconds}$; Service level (a6): $P(t \leq 20s) = 80\%$

Using this basic information and the Erlang-C formula (see appendix a6) it is now possible to calculate the number of agents needed for the given service level of 80/20. The result is 91 agents. The manager of our virtual contact center hopefully knows that already – it is the unchanged current situation. Even if he does not use Erlang-C for staffing, his experiences will lead him to a similar number. Image 1 shows the results for different service levels.

Contact Center managers rely on reports – so let's see what else we can get from Erlang-C. The average waiting time (a7) for example: how long must a customer in average wait before he is served? For our model it is 13 seconds before optimization.

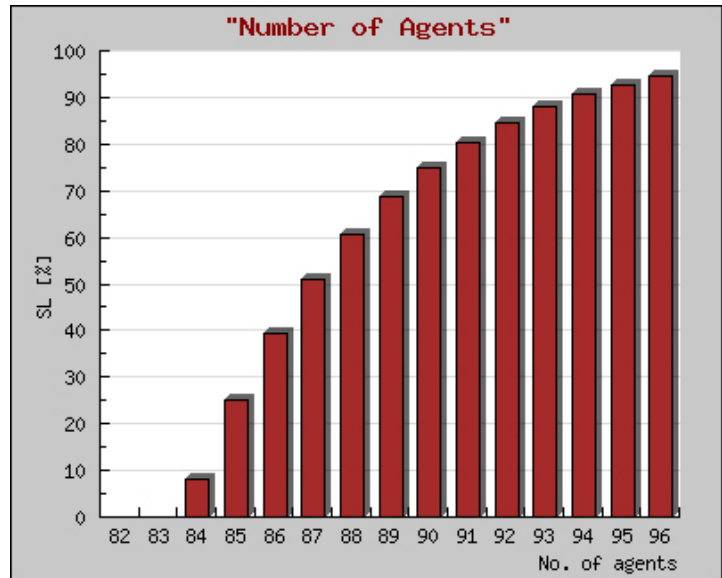


Image 1: Number of agents

AFTER PROCESS OPTIMIZATION

Is it worth starting a project for the optimization of the service time when the result is a time saving of merely 30 seconds?

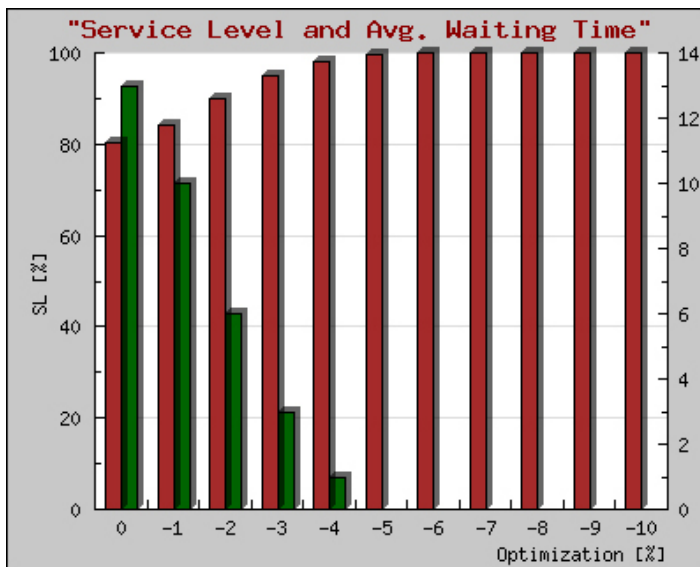


Image 2: service level and avg. waiting time

Using the Erlang-C formula again, but with a reduced service time of 270 seconds instead of 300 seconds we get astonishing results. The result of the calculation of the service level that could be achieved if we keep all **91 agents** is **98% in 20 seconds!** Image 2 also shows that a 5% optimization results in an **average waiting time of only 1 second.**

This sounds like a stress free life. Of course, in reality the contact center manager has to reduce the number of agents to save costs. For the initial service level of 80/20 after the process optimization are **now 83 agents necessary instead of 91.**

As you can see from Image 3 there is a big potential for the operational staffing if contact centers optimize the agent's productivity in the sense of average service time. Additionally, a relatively small time saving allows big flexibility in the service level definitions.

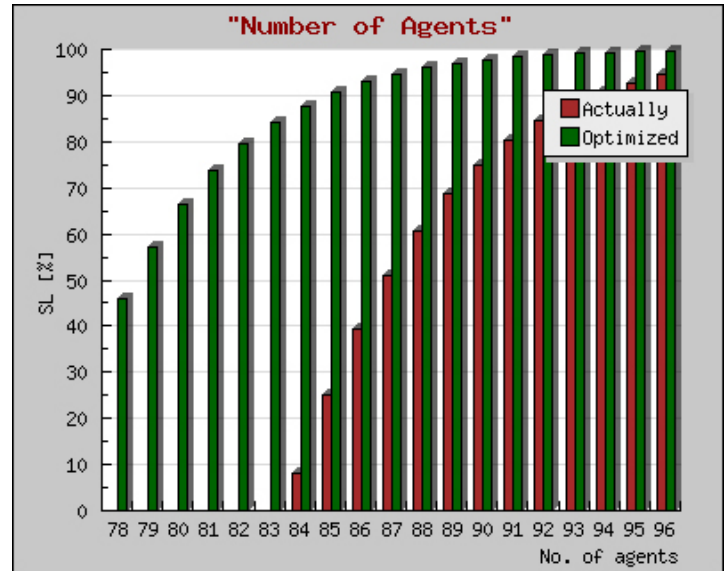


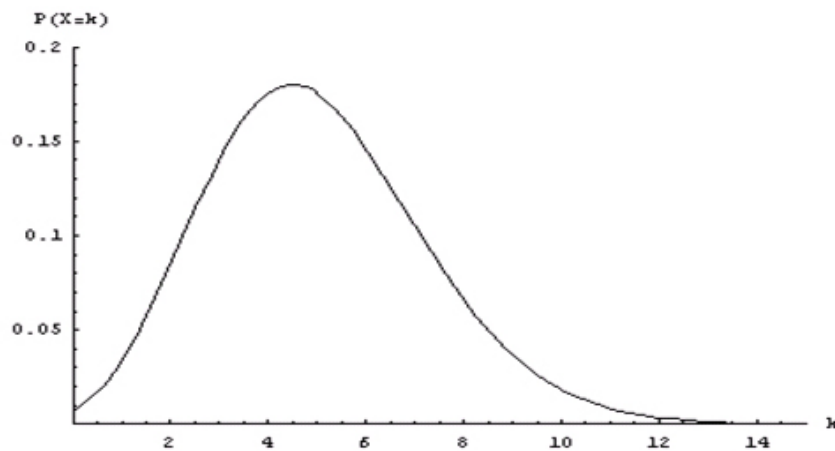
Image 3: Number of agents after optimization

a1: POISSON PROCESSES

A stochastic process $X(k)$ is a (time-homogeneous, one-dimensional) Poisson process if the probability of the number of events in the interval $[0, k]$ is given by
$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

where the positive number λ is a fixed parameter, known as the rate parameter.

The following graph shows the Poisson distribution for $\lambda=5$.



In our scenario the Poisson distribution describes the probability that the service time for a certain call is k minutes and how many customers call per hour.

a2: CLASSIFICATION PATTERN FOR WAITING QUEUE SYSTEMS

Based on D. G. Kendall (1951) waiting queue systems are described using the following notation: **A/B/s** with

- A: distribution of the arrival times
- B: distribution of the service times
- s: number of servers



Shortcuts for distributions

- M: Markov distribution (exponential distributed times)
- G: general distribution (times not specified)
- D: deterministic distribution (constant times)

a3: ARRIVAL RATE

The arrival rate λ is defined as

$$\lambda = \lim_{t \rightarrow \infty} \bar{A}(t), \text{ with } \bar{A}(t) = \frac{A(t)}{t}$$

$A(t)$ is the number of arrivals in the period $[0, t]$, $\bar{A}(t)$ the average number of arrivals in $[0, t]$ and λ the long-term mean.

a4: SERVICE RATE

The service rate μ is defined as

$$\mu = \lim_{t \rightarrow \infty} \bar{D}(t), \text{ with } \bar{D}(t) = \frac{D(t)}{t}$$

$D(t)$ is the number of calls served in the period $[0, t]$, $\bar{D}(t)$ the average number of served calls in $[0, t]$ and μ the long-term mean.

a5: SERVICE TIME

The service time is the average time needed to complete a call: $\bar{t}_B = \frac{1}{\mu}$

a6: ERLANG-C

$P(t=0)$ is the probability that an arriving customer doesn't have to wait and is served immediately. It results from the conditions of a balanced system. The interested reader should consult specialized queueing theory literature for details.

$$P(t=0) = \frac{a^s}{(s-1)! \cdot (s-a)} \cdot \left[\sum_{j=0}^{s-1} \frac{a^j}{j!} + \frac{a^s}{(s-1)! \cdot (s-a)} \right]^{-1}$$

with s : number of agents

$a = \frac{\lambda}{\mu}$. It is called "load", it's unit is *Erlang*.

The service level $P(t \leq t_W)$ for the acceptable waiting time t_W is:

$$P(t \leq t_W) = 1 - P(t=0) \cdot e^{-\frac{(s-a)t_W}{\mu}}$$

a7: AVERAGE WAITING TIME

The average waiting: time t_Q :

$$t_Q = \frac{P(t=0) \cdot \bar{t}_B}{s-a}$$

