Karlsruhe Institute of Technology

Ideas on Customer-oriented Queuing in Service Incident Management

by Peter Hottum[1], Melanie Reuter[1]

KIT SCIENTIFIC WORKING PAPERS 24

[1] Karlsruhe Service Research Institute (KSRI)

Results of this study have been presented at the International Conference on Operations Research 2014 (Aachen, Germany, September 2-5, 2014).

Karlsruhe Service Research Institute (KSRI)
Englerstr. 11
D-76131 Karlsruhe, Germany
www.ksri.kit.edu

Impressum

Karlsruhe Institute of Technology (KIT)
www.kit.edu

2014

ISSN: 2194-1629

1 Introduction

The provision of services hinges considerably on the contribution of the provider and the customer and – if present – on their involved networks. In this working paper we focus on incident management – a service domain that is highly relevant for all kinds of industries and is described from a provider internal perspective in the ITIL documentation (Steinberg, 2011).

By understanding the influence of a customer's contribution to a service, the provider should be able to improve the interaction quality in general. Furthermore the provider should be able to determine and control his effort based on the expected customer's contribution.

In incident management, tickets can arrive per call, email or web interface. For this research we just assume tickets to arrive by web interface as done by many big companies.

This has two implications: On the one hand side, tickets have a predefined structure, such as a predefined content in general, and on the other hand side, the interactions between the customer and the provider are asynchronous – therefore it is possible to collect tickets for some time and then assign them using the knowledge about the other tickets in the queue. This results in an online problem with lookahead or in the extreme case even in an offline problem if we collect all tickets that arrive within a certain period of time (e.g. one day) and schedule them the next period (e.g. the next day). It also means that the content of the tickets can be analyzed and the tickets can therefore be categorized. In contrast, in a regular call center tickets often have to be assigned right away. In addition, no incident ticket would quit the queue for new tickets before scheduling, whereas waiting customers would do, if their processing lasts too long.

In previously conducted studies, we have derived result influencing factor classes and instantiated a framework based on qualitatively and textual analyzed service incident tickets from a worldwide operating IT service provider. We have proven the customer induced contribution to the service generation and aggregated a customer contribution factor (ccf). By complementing these provider-centric service processes with that factor, we are able to use information about the customer's ability to contribute, that was not able to process before. In addition, we can now classify the tickets in more detail than just to use to the severity level of a ticket that is defined by the customer and therefore reorder and prioritize.

The aim is to build a decision support tool in the end that assigns tickets to servers based on a set of rules depending on the underlying objectives and including ticket characteristics as well as the customer contribution factor.

In the working paper at hand, we address the question: How can the customer's potential to contribute be used to organize the queuing in service incident management in a customer-oriented way? We present a mathematical formulation for assigning tickets to servers and discuss first results of a discrete event simulation. We use this simulation to test basic assignment rules based on the ticket complexity and the servers' level of experience. We also study the impact of the ccf in a small example.

2 Problem Formulation and Solution Approach

Service providers in incident management have to handle different topics on several levels of complexity. For our model we choose a process that is based on the incident management process by the ITIL 2011 standard (Steinberg, 2011). In that ITIL standard, the incident handling and the interaction between the provider and the customer are formalized in different process steps. In Fig. 1 we draw out a simplified process view in which we aggregate all the internal escalation steps in a so called ”black box” and focus on the transition, where the assignment of service incident tickets to dedicated service agents is processed.

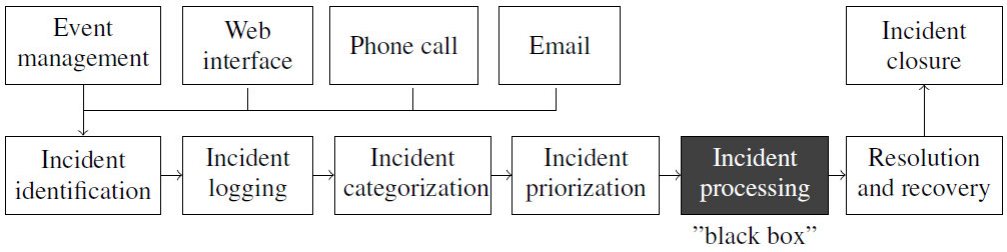

Figure 1: Simplified ITIL process for incident management

For each incident ticket the providers have to determine the topic and the expertise needed to solve the incident. We represent that in the following by giving each incoming ticket a set of attributes and each service agent a specific skill level for each topic he or she is working on.

Fig. 2 visualizes the problem of assigning tickets to agents. If new incident tickets are reported to the incident management web interface, it has to be decided which agent should work on it. This depends on different aspects: which agent is currently available and fits best to the present topic and necessary expert level? Agents are allowed to work on an incident with a complexity equal or less compared to their expertise level, but not higher.

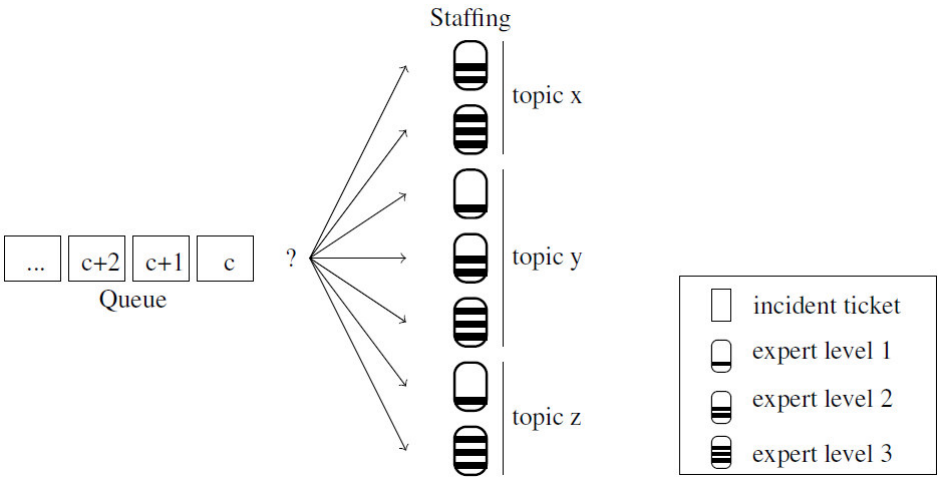

Figure 2: Examined scenario of ticket scheduling

For the deterministic formulation that we need later in our research for comparison purposes we assume a set of incoming service incident tickets C that are known but that cannot be handled before their release date, i.e. arrival time e_c for all $c \in C$. Each ticket $c \in C$ then has a processing time d_c. The set of topics is represented by T with $|T|$ being the number of different topics.

As described, tickets that arrive have a different level of complexity, which we represent be the set of levels L, again $|L|$ being the number of levels. The binary parameter g_{ctl} is equal to 1 if ticket $c \in C$ has topic $t \in T$ with complexity level $l \in L$ whereas for each ticket $c \in C$ only one parameter is equal to 1, i.e. $\sum_{t \in T} \sum_{l \in L} g_{ctl} = 1 \; \forall c \in C$.

The tickets are handled by a set of agents S, where $|S|$ stands for the number of agents that are part of our model. Each agent can only serve a defined subset of topics $t \in T$ and for each topic he has a certain knowledge level that matches with the levels of complexity $l \in L$. f_{stl} is equal to 1 if agent $s \in S$ can solve a ticket with topic $t \in T$ at level $l \in L$ and 0 else. Due to work regulations and as agents are the most valuable resource (especially those with the highest knowledge level $|L|$), their workload should not exceed $\propto$ percent of the daily working time W. By P we denote the number of consecutive days we are looking at, i.e. the length of the considered period. We currently assume that it is possible to schedule all tickets within the planning horizon and that each agent is only able to work on one ticket at a time. By M we denote a sufficiently large number.

In addition, we introduce the following decision variables:

$$x_{bcs} = \begin{cases} 1 \; if \; agent \; s \in S \; solves \; ticket \; c \in C \; after \; ticket \; b \in C \\ 0 \; else \end{cases}$$

$y_c \geq 0$ starting time for solving ticket $c \in C$

The formulation then looks as follows:

$$min \quad \sum_{b,c \in C, b \neq c} \sum_{s: f_{st|L|}=1} x_{bcs} \, d_c \qquad (1)$$

$$s.t. \quad y_c \geq e_c \qquad \forall c \in C \qquad (2)$$

$$y_c \geq y_b + d_b - M(1 - x_{bcs}) \qquad \forall b, c \in C, s \in S \qquad (3)$$

$$y_c + d_c \leq P \cdot W \qquad \forall c \in C \qquad (4)$$

$$x_{bcs} \leq \sum_{t \in T} \sum_{l \in L} (g_{ctl} f_{stl}) \qquad \forall c \in C \qquad (5)$$

$$\sum_{b \in C+0, b \neq c} \sum_{s \in S} x_{bcs} = 1 \qquad \forall c \in C + 0 \qquad (6)$$

$$\sum_{c \in C+0, c \neq b} \sum_{s \in S} x_{bcs} = 1 \qquad \forall b \in C + 0 \qquad (7)$$

$$\sum_{a \in C+0, b \neq a} x_{abs} - \sum_{c \in C+0, b \neq c} x_{bcs} = 0 \qquad \forall b \in C, s \in S \qquad (8)$$

$$\sum_{b,c \in C, b \neq c} x_{bcs} \, d_c \frac{1}{p} \leq \alpha W \qquad \forall s \in S \qquad (9)$$

$$x_{bcs} \in \{0,1\} \qquad \forall b, c \in C, s \in S \qquad (10)$$

$$y_c \geq 0 \qquad \forall s \in C \qquad (11)$$

The objective function (1) minimizes the workload for the agents with the highest skill levels. Constraints (2) assure that the service of a ticket cannot start before the release date, constraints (3) that an agent only starts a new ticket when the last one is finished. By (4) all tickets must be finished within the planning horizon. Of course an agent can only serve a ticket with the right topic and level that he or she is able to solve as expressed in constraints (5). Constraints (6), (7) and (8) make sure that we start and end a schedule for each agent once, that each ticket is served and that the same agent starts and ends serving a ticket. Agents shall not work more than $\propto\%$ of the daily working hours in average throughout the considered period as expressed in (9). (10) and (11) are the domain constraints.

Based on already examined studies in that domain (Giurgiu et al., 2014; Reynolds, 2010; Mazzuchi and Wallace, 2004; Mehrotra and Fama, 2003) we assume the following conditions for an example scenario that we want to study in a discrete event simulation:

We examine the incident management of a medium-sized company. Seven employees with different levels of expertise are working on their day-to-day operations and additionally have to solve incidents that are reported by customers via the company's incident management web interface. We assume an equally distribution of these two kinds of tasks. Furthermore we assume an average availability of each expert of less than 70% of the working time (a so called "shrinkage" with over 30%), which results in a maximum workload of 35% per expert for incident management tasks in general. Each expert could gain a level of expertise from low (1) to medium (2) to high (3) for each topic. In our model there are tickets in the domains of 3 different topics (topic x, topic y, and topic z). Each ticket has a complexity of low (1) or medium (2) or high (3). The agents work on the tickets on maximum five days per week for eight hours. The incidents, reported via the incident management web interface, are Poisson distributed with a lambda of 50 minutes. The customer contribution is rate-able for each ticket as ccf from low (0) to medium (1) to high (2). The time to resolve an incident is calculated by $20min + max(complexity - ccf; 0) \cdot t$ where t is normally distributed with a mean of 60 minutes and a standard deviation of 10 minutes. An incident ticket has always be scheduled to the available agent with the lowest expert level. This is important to give the highly educated (and therefore higher paid) experts more time for solving issues in their day-to-day operations. Every incident ticket in the queue is scheduled by the first-come-first-serve principle.

3 Computational Results

Based on the above described model, we simulated the scheduling of tickets and the utilization of corresponding agents with AnyLogic to also study the impact of the percentage of tickets with a high customer contribution factor. Therefore we used ten base seeds each to reduce variations for different shares of tickets with a high customer contribution - from 0 to 1 in steps by 1%. The remaining share of tickets with a low and a medium customer contribution have been divided equally.

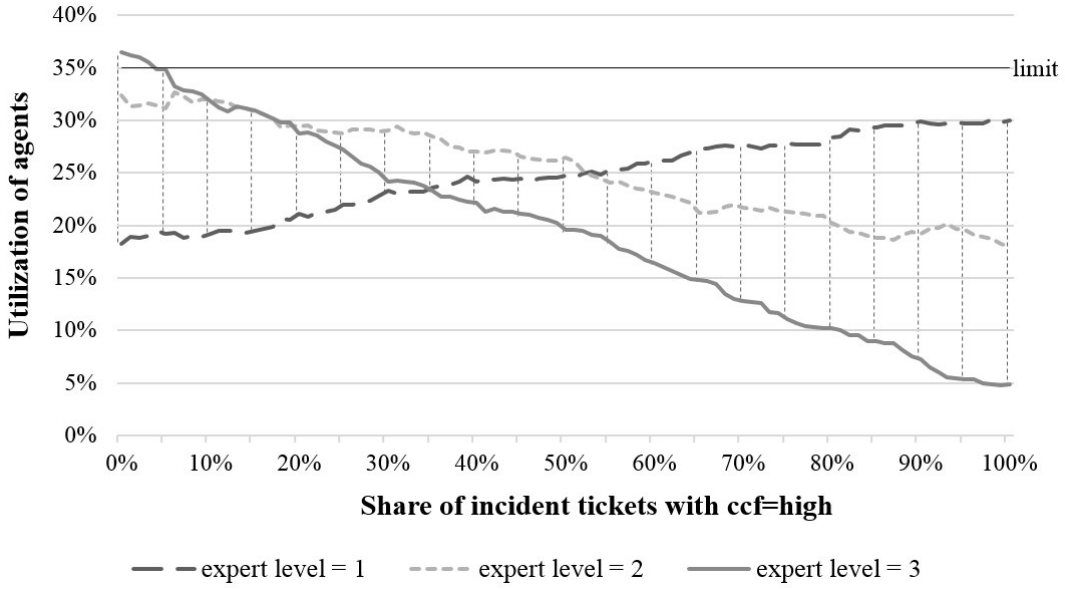

Figure 3: Utilization of service agents relatively to the customer contribution

In Fig. 3 the effects of different shares of tickets with a high ccf on the utilization of the agents is presented. Given a maximum utilization limit of 35% per agent, it gets obvious, that with an increase of ccf the utilization of agents with higher expert levels reduces. The calculated curves are specific for each provider's setting and process handling and serve as an indicator for each provider's sensitivity concerning the spectrum of ccf.

4 Conclusions and Recommendations for Further Research

In this working paper it could be shown that the customer contribution factor ccf can help to reduce the unbalanced utilization of service agents by assigning tickets to agents that are able to handle them properly. By applying information about their customers, providers could be able to save resources and time internally and – at the same time – serve their customers more individual, faster and with no more effort.

The first results already raise mainly two implications for service providers: First they may use the knowledge about the ccf operationally – for providing the service in a customer-individualized way (e.g. skip unnecessary process steps of information gathering and involve agents with the according level of expertise more quickly). Second, providers may use the results in a strategic way: by understanding the effects of the ccf on their own service setup – their provider-specific sensitivity – they can plan actions to qualify their customers or redesign their incident management web interface towards the customer to raise the share of high ccf tickets.

Within this working paper we were not able to apply our approach to the real world case, where we took our motivation and initial set up from. As the exact cause effect relationships of the ccf are estimated in the starting model, the next step in our research is to prove these effects with the real interaction data, captured with our application partner. From a mathematical point of view, we will use queuing theory to further study waiting times, business of agents and the time a ticket stays in the system. In a future stage of the research we also want to implement and solve the

deterministic problem in CPLEX with the same data used in the simulation for comparison purposes.

References

Giurgiu, I., Bogojeska, J., Nikolaiev, S., Stark, G., Wiesmann, D. (2014): Analysis of Labor Efforts and Their Impact Factors to Solve Server Incidents in Datacenters. Proceedings of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 424-433.

Mazzuchi, T.A., Wallace, R.B. (2004): Analyzing skill-based routing call centers using discrete-event simulation and design experiment. Proceedings of the 2004 Winter Simulation Conference, pp. 1812-1820.

Mehrotra, V., Fama, J. (2003): Call center simulation modeling - methods, challenges, and opportunities. Proceedings of the 2003 Winter Simulation Conference, pp. 135-143.

Reynolds, P. (2010): Call Center Metrics - Fundamentals of Call Center Staffing and Technologies. NAQC Issue Paper. Phoenix.

Steinberg, R.A. (2011): ITIL Service Operation - 2011 Edition. 2nd edition. TSO, Belfast.

KIT Scientific Working Papers
ISSN 2194-1629

www.kit.edu