Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

ROUTING HEURISTICS FOR MULTI-SKILL CALL CENTERS

Ger Koole, Auke Pot

Department of Mathematics
Vrije Universiteit
De Boelelaan 1081a, 1081 HV Amsterdam
the Netherlands

Jérôme Talim

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
K1S 5B6 Canada

ABSTRACT

We give an approximation method for analyzing the performance of call centers with skill-based routing, for both blocking and delay systems. We use this method to determine optimal skill sets for call center employees.

1 INTRODUCTION

Planning issues in call centers are a fruitful research area, both from an economical and a mathematical point of view. See Gans, Koole, and Mandelbaum (2003) for an overview. One of the main challenges in the planning of call centers is dealing with multiple skills. This problem arises from the fact that telephone calls requesting different types of services arrive in different queues at the call center's telephone switch (on different numbers, or through an interactive process such as *Interactive Voice Response*). This allows call center management to differentiate between different call types, and to treat them separately, by separate employees (which are often called *agents*). This avoids *cross-training* agents, which saves money and time. On the other hand, with only specialized agents we cannot profit from the economies of scale that arise when we have only cross-trained agents. Can we choose a mix of specialized and cross-trained agents such that we have few cross-trained agents and still profit from economies of scale?

We show by an example that this is indeed the case. Consider a call center with two types of calls, with each a load of 100 Erlang. Customers that find no free agent with the right skill abandon (a motivation for this choice is given below). The objective of this call center is to have less than 5% abandonments. In Table 1 we give the minimum numbers of agents needed for different condigurantions. Our conclusions are as follows: the economies of scale are substantial, and can be obtained by cross-training only a minor fraction of the agents.

There are a number of questions of interested related to skill-based routing. An important issue is the way in which skill-based routing is implemented. We discuss this in the next section. Based upon this choice, the mathematical problems have to be solved: how to route incoming calls (an online control problem) and which agents to schedule (a workforce planning problem). These problems are both practically very relevant and difficult to solve, as standard methods fail due to the curse of dimensionality.

# agents with skill			
1	2	1&2	% abandonments
105	105	0	4.8
90	90	25	4.8
0	0	202	4.8

Table 1: Performance of a two-skill call center

In this paper we consider routing and scheduling problems under the assumption that there is no queueing: we assume that queued calls are blocked, or, equivalently, that queued calls abandon right away. This might appear an arbitrary choice, certainly given the fact that it is current practice to assume no blocking or abandonments at all (e.g., many single and multi-skill call centers use the Erlang C formula to determine occupancy levels). We think that assuming no abandonments is equally arbitrary as assuming no waiting; in fact it even is much less realistic in situations with a high load or overload. Assuming no queueing has the advantage that it simplifies the problem considerably. In Section 4 we also pay attention to systems with delay. The preferred way is tthe intermediate case where abandonments are modeled explicitly (Gans, Koole, and Mandelbaum 2003); we plan to extend our methods also for this situation.

2 TYPES OF SKILL-BASED ROUTING

The typical situation in a call center is that we have a number of skill groups (say G), where all s_i agents within group i have the same skills $C_i \in \{1, \ldots, K\}$,

with K the number of call types, and λ_k the arrival rate of type k. We also assume that $C_i \neq C_j$ for all $i \neq j$. Agents in a group i with $|C_i| = 1$ are called *specialists*, agents in the group i with $|C_i| = K$ (if it exists) are called *generalists*, and agents in a group i with $|C_i| > 1$ are said to be *cross-trained*. We assume that call holding times are exponentially distributed, with parameter μ_k for call type k.

Under some aditional assumptions (time homogeneity, known arrival rates, etc.) it is in theory possible to compute the optimal assignment rule. This would be a function that specifies in each possible configuration (state) for each call type if and to which skill group a potentially arriving call should be assigned. A state is characterized by the number of calls that are in process for each type at each for that skill possible skill group. Therefore the state space has dimension $\sum_{i=1}^{G} |C_i|$. This has two implications: standard solution methods from Markov decision theory suffer from the curse of dimensionality and therefore it is numerically infeasible to compute the optimal policy; it is practically impossible and undesirable to implement a policy that has no obvious structure and that can differ in any two states. This is what we call dynamic routing.

Because of the implementation issue we encounter most often in call centers a call routing structure that allows less flexibility in routing policies but that is much easier to characterize. This structure consists of a list of skill groups for each call types that are, upon call arrival, successively searched for the presence of available agents. If no agent is available in any of the groups in the list then the call is queued; under our assumption concerning abandonments the call leaves the system immediately. This is equivalently to calls overflowing from skill groups without available agents; we will call this type of routing therefore *overflow routing*. The class of overflow routing policies is smaller than the class of dynamic routing policies, and also the optimal overflow routing policy performs in general worse than the optimal dynamic routing policy. We illustrate this and some other issues with the simple example of Figure 1. In the figure we depicted that type-2 calls are first routed to skill group 1 and then to skill group 2. We analyzed three different routing policies: overflow routing as depicted in the figure, a probabilistic routing policy that assigns type-2 calls with equal probability to skill group 1 or 2, and the optimal dynamic routing policy that assign type-2 calls to the skill group with the most available agents. The optimal routing policies is strictly better than the other two, with overflow routing the worst, but the difference are small in this example. Because of its practical importance we restrict ourselves in this paper to overflow routing.

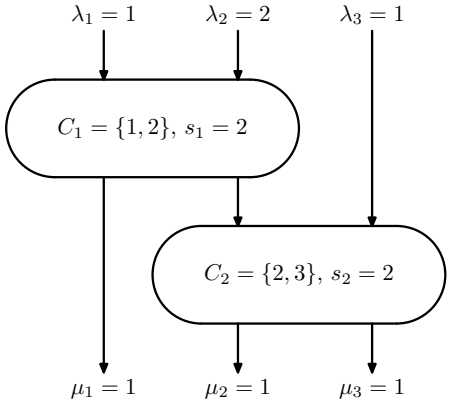

Figure 1: A three-class call center.

3 APPROXIMATION AND VALIDATION OF OVERFLOW ROUTING

Consider again Figure 1. The flow of type-2 calls from skill group 1 to 2 is a renewal process, but not a Poisson process. In more complex situations overflow processes are not even renewal processes anymore. This makes it very hard to analyse analytically call centers with more than a handful of agents. We introduced in Koole and Talim (2000) an approximation based on replacing all overflow processes by Poisson processes and by assuming that the stations are stochastically independent. To present this approximation we first have to introduce $B(s, a)$, the blocking probability of the Erlang B model with s agents and load a. For $B(s, a)$ we have the following well-known formula:

$$B(s, a) = \frac{\frac{a^s}{s!}}{1 + \cdots + \frac{a^s}{s!}}.$$

Note that Erlang B is the correct model for a skill group assuming that input is Poisson. We assume that skill groups are used for all their skills, i.e., if $k \in C_i$, then i occurs in the routing table for k. Then for a certain skill group i, call of type k arrive at rate $\gamma_k(i)$, with $\gamma_k(i) = \lambda_k$ if group i is the first in the routing table for k. We define $\gamma_k(0)$ as the loss rate of type k.

The load a_i to skill group i can now be defined:

$$a_i = \sum_{k \in C_i} \frac{\gamma_k(i)}{\mu_i}.$$

Let $n_k(i)$ be the successor skill group i for call type i, define $n_k(i) = 0$ if overflow of type k is lost after group i. Then we get the following set of equations for the rates $\gamma_k(i)$:

$$\gamma_k(i) B(s_i, a_i) = \gamma_k(n_k(i)),$$

for all k and i. Using an iterative method the rates $\gamma_k(i)$ can easily be solved. Note that when the graph with arcs corresponding to overflow of one of the call types has no cycles, then the rates can be directly computed. We illustrate this with the simple example of Table 1. For a more thorough investigation, see Koole and Talim (2000).

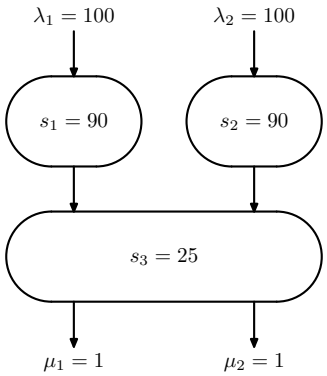

Figure 2: A two-class call center.

In Figure 2 we depicted the overflow network of Table 1. The abandonment percentage is 4.8; this was calculated using simulation. The abandonment percentage based on the exponential approximation is only 3.4. This is calculated as follows: $B(90, 100) = 14.6\%$, thus the input to the cross-trained skill group is 2×14.6, and $B(25, 29.2) = 23\%$. The overflow probability approximation for arriving calls is $0.146 \times 0.23 = 0.034$. This is considerably lower than the real 0.048, our experience using simulations shows however that the approximation is better for more complicated call centers. Both the fact that the exponential approximation is a lower bound and that the approximation gets better when more classes are involved are observations that still lack a theoretical underpinning: only for the case of one call type and two groups (both with skill set $\{1\}$) it is known that the exponential approximation results in a lower bound (Hordijk and Ridder 1987).

Next we discuss a more elaborate model with 5 types, 5 groups of specialist, 4 groups of agents with 2 skills, with skill sets $\{1, 2\}$, $\{2, 3\}$, $\{3, 4\}$, and $\{4, 5\}$. Finally there is a group of generalists. Each skill group has five agents (except for the group with generalists), the load of each class is 10 Erlang. In Table 2 our results are displayed, for varying numbers of total agents (note that 10 generalists corresponds to a total of 55 agents). It is interesting to note for example that $B(55, 50) = 0.054$, the lowest blocking probability that can be obtained for load $a = 55$.

The results of Table 2 are a little disappointing, certainly compared to the numerical results of Koole and Talim (2000). This is explained by the fact that most of the results in Koole and Talim (2000) are for highly loaded systems; there the approximation methods works best.

Situation	Simulation	Approximation
53	0.11	0.084
55	0.085	0.057
57	0.063	0.035
59	0.044	0.019

Table 2: Approximations of blocking probabilities for a 5-class system.

4 APPROXIMATION AND VALIDATION OF DELAY SYSTEMS

In Cooper (1981) (page 92) a relation is given between the blocking probability $B(s, a)$ in the Erlang B system and the delay probability $C(s, a)$ of the Erlang C system:

$$C(s, a) = \frac{sB(s, a)}{s - a(1 - B(s, a))}.$$

Although this formula is only valid for standard single-skill models, we use it to construct an approximation of C for the multi-skill situation, by taking B in the formula equal to the approximation based on Poisson overflow processes. This gives the results of Table 3 for the 5-class model of the previous section. In each row we give the approximation based on a simulation of the delay system, the approximation of C based on the Poisson approximation of B, and the approximation of C based on a simulation of the blocking system. Thus the last two columns correspond to the last columns of Table 2, but in reversed order.

# of agents	Sim. delay	Poisson appr.	Sim. blocking
53	0.64	0.62	0.69
55	0.45	0.40	0.51
57	0.30	0.23	0.35
59	0.19	0.11	0.23

Table 3: Approximations of delay probabilities for a 5-class system.

The approximation of the delay probability is considerably better than the approximation of the blocking probability, although also this approximation slowly deteriorates as the number of agents increases.

5 OPTIMIZATION

The results presented so far are of limited use in the approximation of service levels. The approximations

are very helpful in optimizing call centers with multiple skills. In these optimizattion procedures it is necessary to evaluate the performance of many different configurations. The speed of the Poisson approximation makes it possible to do this in limited time.

We analyzed the following situation. Consider the load and possible skill groups of the previous sections. We fixed the total number of agents to 57, and aimed at finding the best configuration where the probability of exceeding a delay of 20 seconds is smaller than 0.20, thus a service level (SL) of 0.8. We started with only generalists, and iteratively moved agents to skill sets with less skill as to minimize $\sum_{i=1}^{G} s_i|C_i|$, the total number of skills that has to be acquired. Our local search algorithm changes in each configuration that agent that induces the smallest reduction in service level. A problem is the fact that the number of skills might reduce by more than 1 at a time. How to choose which agent to replace? We solved this by considering the decrease in SL divided by the number of skills. The final configuration is as follows. The skill groups with skill sets $\{2\}$, $\{3\}$, $\{4\}$, $\{2,3\}$, and $\{3,4\}$ each get 5 agents, the skill groups with skills $\{1\}$ and $\{5\}$ each get 7 agents, and skill groups $\{1,2\}$, $\{4,5\}$, and $\{1,2,3,4,5\}$ each got 6 agents. In Table 4 the results are summarized.

Scenario	Sim. delay	Poisson appr.	SL
initial	0.30	0.23	0.82
optimized	0.34	0.27	0.80
only gen.	0.75	0.75	0.98

Table 4: Delay probabilities for an optimized 5-class system.

The approximation appeared to be very useful while optimizing, and we are convinced that a (nearly) optimal solution has been found. On the other hand, the inaccuracy of the approximation obliged us to run simulations now and then to see whether the SL constraint was still satisfied.

6 FUTURE WORK

Currently we are studying improvements of the exponential approximation of the blocking system. By modeling the burstiness of the overflow processes we hope to find better approximations. We also consider the case of abandonments.

REFERENCES

Cooper, R. 1981. *Introduction to queueing theory*. 2nd ed. North Holland.

Gans, N., G. Koole, and A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2). To appear.

Hordijk, A., and A. Ridder. 1987. Stochastic inequalities for an overflow model. *Journal of Applied Probability* 24:696–708.

Koole, G., and J. Talim. 2000. Exponential approximation of multi-skill call centers architecture. In *Proceedings of QNETs 2000*, 23/1–10.

AUTHOR BIOGRAPHIES

GER KOOLE is a professor in the Department of Mathematics at the Vrije Universiteit Amsterdam. He graduated in 1992 from Leiden University, and held postdoc position at CWI (Amsterdam) and INRIA (Sophia Antipolis). He is interested in the theory and applications of controlled queueing systems. His e-mail address is koole@cs.vu.nl, and his web page is http://www.cs.vu.nl/~koole.

AUKE POT is a PhD student at the Vrije Universiteit Amsterdam. He is working on call center planning, especially skill-based routing. His e-mail address is sapot@cs.vu.nl, and his web page is http://www.cs.vu.nl/~sapot.

JEROME TALIM is an assistant professor at the Department of Systems and Computer Engineering of Carleton University. He graduated in 1998 at INRIA (Sophia Antipolis), and held postdoc position at the University of Cambridge, at the Vrije Universiteit Amsterdam, and at the University of Saskatchewan. He is interested in call center performance evaluation, network traffic processing, and genes decoding. His e-mail address is jtalim@sce.carleton.ca, and his web page is http://www.sce.carleton.ca/faculty/talim.html.