Markov Models for Multi-Skill Call Centers

Manfred Schneps-Schneppe\textsuperscript{1,*,} Janis Sedols\textsuperscript{2}

\textsuperscript{1}Ventspils University College, Ventspils International Radio Astronomy Centre Inzenieru st 101, Ventspils, LV-3601, Latvia
\textsuperscript{2}Institute of Mathematics and Computer Science, University of Latvia, Raina boulevard 29, Riga, LV-1459, Latvia

manfreds.sneps@gmail.com, janis.sedols@lumii.lv

Abstract We use parallels between the older telephone switches and the multi-skill call centers. We compare different call center routing policies at low and high call rates $\lambda$. By numerical results it is shown that a call center with equally distributed skills is preferable compared to traditional grading-type design. The strong proof is given by expansion of call loss probabilities in powers of $\lambda$ and of $1/\lambda$, respectively. The proof draws on one excellent V. Benes’s paper (from Bell Labs). The main conclusion leads to a new principle for call center design: from throughput point of view a multi-skill call center with equally distributed skills is preferable compared to traditional grading-type design.

Keywords Call Center, Skill Based Routing, Teletraffic, Limited Availability, Asymptotic Expansion for Loss Probability

1. Introduction

In a typical call center, the arriving calls are classified in different types, according to the required technical skill to answer the call, the language, importance of the call, etc. Agents are also classified in skill groups according to the subset of call types they can handle. Calls arrive at random according to some stochastic process. When a call arrives, it may be assigned immediately to an agent that can handle it (if there is one available) or it may be put in a queue (usually one queue per call type). When an agent becomes available, the agent may be assigned a call from one of the queues, or may remain idle (e.g., waiting for more important calls). All these assignments are made according to some routing policy that often incorporates priority rules for the calls and agents.

Calls waiting in a queue may abandon after a random patience time. Those subscribers who abandon waiting may call again later, although those retrials are rarely modelled in practice, usually because of lack of sufficient data. Callers who received service may also call again for a number of reasons; these are called returns.

In the (degenerate) special case where each agent has a single skill, we have several single queues in parallel. If each agent has all skills, then we have a single skill set and a single queue. The system is obviously easier to analyse in these extreme cases. With all agents having all skills, the system is also more efficient (smaller waiting times, fewer abandonment). Agents with more skills are also more expensive; their salaries depend on their skill sets. Thus, for large volume of call types, it makes sense to select a number of single-skill agents (specialists) to handle most of the load. A small number of agents with two or more skills can cover the fluctuations in the proportion of calls of each type in the arriving load.

The multi-skill call center is far too complex to be available for strong mathematical analysis. The state of research into solving skill-based routing problems is still in its infancy. A special case where two classes of calls are served by a single pool of cross-trained agents is studied in [1] and multi-skill centers in [2]. For a literature survey on asymptotic heavy-traffic regimes we refer to [3,4]. The excellent survey of multi-skill call centers contains paper[5]. As initial point for our study we use paper[6]. There the authors consider routing and scheduling problems under the assumption that there is no queuing, assuming that queued calls are blocked, or, equivalently, that queued calls abandon waiting right away, which result in the so called blocked calls cleared discipline.

The paper is organized as follows. Section 2 contains numerical examples for simple multi-skill call center models. These models take into account the structure of the call center and method used for routing of calls. The results have shown that a call center with equally distributed skills is preferable comparing to the traditional grading-type design. Section 3 contains a short version of mathematical proof on call loss probabilities at low and high call rates. The proof uses one excellent V. Benes’s paper[7] from Bell Labs. In Section 4 we apply these results to compare the structures of call centres at low and high call rates. Section 5 contains a discussion on the basis of numerical examples looking for future research.

The paper is a version of our previous paper[8] extended basically for tutorial purposes, namely: the simple numeri-
cal examples are given in Section 2, numerical results – for unsolved issues in Section 5, and some additional detail – proving the theorems.

2. Numerical Examples

Calls from different skill classes are offered to the call center according to a Poisson process with rate $\lambda$. The agents in the center are grouped according to their heterogeneous skill sets that determine the classes of calls that they can serve. Each agent group serves calls with independent and equally exponentially distributed service times equal to one (for simplicity). We consider a call center with no buffers in the system, so that every arriving call either has to be routed immediately or has to be blocked and is lost. The objective in the system is to calculate the call lost probability.

Example 1. Markov process for the simplest two-skill call center

![Figure 1. Two-skill call center with 3 agents](image1)

Fig 1 displays the simplest two-skill center model: 2 skills and 3 agents, two of them are individualists (single skill) available for calls as the first choice and one generalist (2-skill agent) available for calls as the second choice. To calculate the call lost probability we build Markov process having 8 states as shown in the state diagram (Fig 2). (Busy agents are given in black.)

![Figure 2. State diagram and Markov process intensities](image2)

Markov process stationary state probabilities $p_i, i = 0, \ldots, 7$ are defined by the equations:

\[
\begin{align*}
(\lambda+2)p_0 &= \lambda p_1 + \lambda p_2 + p_3 \\
(2\lambda+1)p_1 &= \lambda p_0 + p_4 + p_6 \\
(2\lambda+1)p_2 &= \lambda p_0 + p_5 + p_6 \\
(2\lambda+1)p_3 &= p_4 + p_5 \\
(\lambda+2)p_4 &= \lambda p_4 + \lambda p_5 + 2\lambda p_6 \\
p_0 + p_1 + \ldots + p_7 &= 1
\end{align*}
\]

By solving the system we get the call loss probability $\pi$ (pictured in Fig 3):

\[
\pi = \frac{p_4 + p_5}{2} + p_7 = \frac{3\lambda^2 + 16\lambda^3 + 20\lambda^4 + \lambda^5}{6 + 26\lambda + 50\lambda^2 + 54\lambda^3 + 32\lambda^4 + 8\lambda^5}
\]

![Figure 3. Two-skill call center loss probability](image3)

Discussion. Let us compare now the just above studied call center model (Fig 1 and Fig 4a) with two others: a similar one (Fig 4b) but by opposite choice of agents – generalist is considered as the first choice and one more in which each agent has both skills (Fig 4c). As Fig 5 shows, the minimum call loss is in the case of full availability system (Fig 4c), and the corresponding loss probability is defined by Erlang formula

\[
E(3, 2\lambda) = \frac{(2\lambda)^3 / 6}{1 + 2\lambda + (2\lambda)^2 / 2 + (2\lambda)^3 / 6}
\]

The above studied scheme (Fig 1) is in the middle and the scheme with generalist as the first choice (Fig 4b) gives the largest call loss.

![Figure 4. Three simple call center models](image4)

![Figure 5. Loss probability; three simple call center models](image5)
Example 2. Grading-type center vs center with equally distributed skills

Let us start with a historical remark. Up to 50 years ago one of the authors (the first one) in his young student age had a chance to deliver the numerical results for the two schemes shown in Fig. 6 at the Kolmogorov’s applied probability seminar at Moscow State University. These schemes relate to so called step-by-step telephone exchanges. The results have surprised Andrey Kolmogorov, namely: in his opinion, in the case of limited access of switch outlets the optimal scheme should be in the form of grading (Fig. 6a). But...this statement, as it turned out, is true only for low call rate. In the case of high call rate, the optimal one is the scheme with equally distributed outlets as in Fig. 6b. In terms of call center, both these schemes are 4-skill centers with 6 agents, and each skill call rate is equal to λ.

Traditionally, there are two types of agents in a call center: individualists (handling calls of one type) and generalists (handling calls of any type). Figure 6a corresponds to the traditional scheme: there are 4 agents who handle calls of one type (from different 4 skills) and 2 agents who are generalists. Every callflow has access to 3 agents, and the calls are looking for idle agent from below (as arrays show). We show that it is advantageous to reject the traditional scheme and switch to a scheme with the same number of different skills for any agent (as is shown in Fig. 6b). These results were included in our paper[[9].

Figure 7 depicts the loss probability curves for these two schemes. And what is surprising? Beginning with a loss probability as low as 0.25 (less than 1%), it is advantageous to use the equally distributed scheme where any agent is available to some call flow (i.e. there is less than one call in the system) and the agents are busy only 0.73/6 = 12% of the time. Therefore, the traditional grading-type call centers could be recommended when call rates are very low.

3. Expansion of Loss Probability in Powers of λ and 1/λ

As we have shown above by numerical analysis, the classical grading-type call centers are preferable for low call flows only. At call loss around 1%, the loss probability curves are cross, and more preferable are the schemes with uniformly distributed skills among the agents. Now we go to the strong mathematical analysis of this phenomenon.

![Figure 6](image1.png)

Figure 6. Two multi-skill call center schemes: a) grading-type, b) with equally distributed skills

![Figure 7](image2.png)

Figure 7. Comparison of call center routing policy: grading (a) is preferable for low rates only, but as load grows the scheme (b) becomes preferable

For simplicity of formulas, we consider rectangular-type call centers with parameters: n – skills (call flows), d – number of available agents for each call flow, v – total number of agents. Therefore, the call center has n*d skill-points divided into v groups (among v agents). The call center serves n Poisson call flows (each of rate λ), the holding time is exponentially distributed (with parameter equal to one, for simplicity of formulas). If all d agents available to some call flow are busy, the call is lost.

We describe the call processing by means of Markov process with a state set S of 2^n states. If x is a state, the notation |x| will denote the number of calls in progress in state x – the number of busy lines in state x.

We define the levels

\[ L_k = \{ x \in S : |x| = k \}, \quad k = 0,1,...,\max_{x \in S}|x|, \]

as the sets of states in which a specified number of calls is in progress. The \( L_k \) form a partition of S,

\[ \bigcup_k L_k = S, \quad L_k \cap L_j = 0, \quad k \neq j. \]

The “neighbours” of a state x are just those states which can be reached from x by adding or removing one call. These neighbours y of x can be divided into two sets according as y > x or y < x; so we are led to define

\[ A_x = \text{set of neighbours above x} \]

\[ B_x = \text{set of neighbours below x} \]

\[ = \text{set of states accessible from x by adding a call} \]

\[ = \text{set of states accessible from x by removing one call (see Fig 8).} \]

To describe how routes are assigned to calls, we introduce a routing matrix \( R = (r_{xy}) \), where \( r_{xy} \) – number of the call flows (each flow of separate skill) which move the system from state x to state y, s(x) – number of call flows having at least one idle agent in state x. Obviously,

\[ s(x) = \sum_{y \in A_x} r_{xy}, \]

\( r_x \) – element of matrix R in \( |x| \) degree having indexes (0, x).

Obviously, \( r_x \) means the number of paths to move from state 0 to state x by means of incoming calls only.
For the purpose of defining a Markov stochastic process it is convenient and customary to collect the assumptions introduced above in a matrix of transition rates

\[
q_{xy} = \begin{cases} 
1 & y \in B_x \\
\lambda r_{xy} & y \in A_x \\
-|x| - \lambda s(x) & y = x \\
0 & \text{otherwise}
\end{cases}
\]

We get the state probabilities \( p_x \) as the solution of linear algebraic system

\[
\sum_{y \in S} \left[ p_y + \lambda \sum_{y \in B_x} p_y r_{yx} \right] p_x = \sum_{y \in S} [n - s(x)] p_x
\]

and probability of call loss is equal

\[
\pi = \sum_{x \in S} \left[ n - s(x) \right] p_x
\]

In order to study the optimal call center design principles we apply to the expansion of call loss probability in powers of \( \lambda \) (Theorem 1) and in powers of \( 1/\lambda \) (Theorem 2). As mentioned above, we consider rectangular-type call center with parameters: \( n \) – skills (call flows), \( d \) – number of available agents for each call flow, \( v \) – total number of agents.

**Theorem 1** [9]. Expansion of loss probability in powers of \( \lambda \)

If the sequence \( \{c_m(x), m \geq 0, x \in S\} \) is defined by

\[
c_m(0) = \delta_{m0}, \\
c_m(x) = 0 & 0 \leq m < x > 0,
\]

\[
c_m(x) = \frac{r_x}{|x|^m},
\]

\[
|y|c_m(x) + s(x)c_{m-1}(x) = \sum_{y \in A_x} c_m(y) + \sum_{y \in B_x} c_{m-1}(y)r_{yx}, \quad m > |x|, \quad x > 0
\]

and

\[
\lambda < \min_{x \in S; |x| = 0} \left( \lim_{m \to \infty} [c_m(x)]^{1/m} \right)^{-1}
\]

then the asymptotic expansion of call loss probability in powers of \( \lambda \) is available at point of \( \lambda = 0 \), and the members not equal to 0 are of power not less than \( \lambda^d \).

According to this theorem for \( |x| > 0 \) we have expansion

\[p_x = p_0 \sum_{i=0}^{\infty} \lambda^i c_i(x)\]

and the call loss probability is equal to

\[
b = \frac{\lambda^d \sum_{i=0}^{\infty} \lambda^i \sum_{x \in S} c_{d+i}(x)[n-s(x)]}{n \sum_{i=0}^{\infty} \lambda^i \sum_{x \in S} c_i(x)} = \frac{\lambda^d B(\lambda)}{A(\lambda)}
\]

**Theorem 2** [10]. Expansion of loss probability in powers of \( 1/\lambda \)

If the sequence \( \{d_m(x), m \geq 0, x \in S\} \) is defined by

\[
d_m(x) = \delta_{mx},
\]

\[
d_m(x) = 0, \quad 0 \leq m < v - |x|,
\]

\[
d_{v-|x|}(x) = \frac{1}{s(x)} \sum_{y \in A_x} d_{v-|y|},
\]

\[
|x|d_m(x) + s(x)d_{m+1}(x) = \sum_{y \in A_x} d_m(y) + \sum_{y \in B_x} d_{m+1}(y)r_{yx},
\]

\[
m > v - |x|, \quad |x| \neq v
\]

and

\[
\lambda > \max_{x \in S; |x| = 0} \left( \lim_{m \to \infty} |d_m(x)|^{1/m} \right)
\]

then the asymptotic expansion of call loss probability in powers of \( \lambda^{-1} \) is available at point of \( \lambda = \infty \).

According to this theorem we have the expansion

\[p_x = p_0 \sum_{i=0}^{\infty} \left( \frac{1}{\lambda^i} \right) d_i(x), \quad |x| \neq v\]

Call loss probability is equal to

\[
\pi = \frac{\sum_{i=0}^{v-1} \left( \frac{1}{\lambda^i} \right) \sum_{y \in S} d_i(y)[n-s(x)]}{\sum_{i=0}^{v-1} \left( \frac{1}{\lambda^i} \right) \sum_{y \in S} d_i(y) C(\lambda)} = D(\frac{1}{\lambda})
\]

and according to Taylor series

\[
\pi = D(0) + \frac{1}{\lambda} C(0)D(0) - C(0)D(0) + o(1/\lambda)
\]

The four first expansion terms are defined by

\[
\lambda = 1 - \frac{1}{n} + \frac{L}{n} + \left( \frac{L}{n} \right)^2 + 2T + M - 2^2 \left( \frac{1}{\lambda} \right)^3
\]

\[
L = \sum_{i=1}^{n} \frac{l_i}{l_i}, \quad T = \sum_{i,j} \frac{1}{l_i + l_j}, \quad M = \sum_{i,j} S_{ij} \left( \frac{1}{l_i} + \frac{1}{l_j} \right) - \frac{v-1}{l_i}
\]

where

\[
l_i \text{ – the total number of agents with } i\text{-skills},
\]

\[
s_{ij} \text{ – the number of skills for which at least one agent is available in state when only agents } i \text{ and } j \text{ are free},
\]

\[
\xi_{ij} \text{ – the number of skills (part of } s_{ij} \text{) from which calls are coming to agent } j \text{ when only agents } i \text{ and } j \text{ are free}.
\]
4. On New Principle of Call Center Design

Let us continue with one more historical remark. To a large extent, our paper is based on the results of a distinguished mathematician of Bell Laboratories V.E. Benes[7, 11] who carried out a research on so-called crossbar telephone switches. Theorem 3 given below follows Benes’s paper[7] from 1963. This paper contains formulas for expansion of loss probability in powers of $\lambda$. A little later, in 1965 we developed the same type results for step-by-step telephone switches[10] using expansion of loss probability in powers of $\lambda$. Benes paper[11] from 1966 contains a similar expansion in powers of $1/\lambda$. And in this context it is worthy to mention Benes words[11]. He has written: “The question has arisen whether there are examples of pairs of networks, with the same number of crosspoints, the first of which is better than the second at one value of $\lambda$, while the second is better than the first at another value of $\lambda'$. The same type of problem was a goal of our studies[10]—only for another telephone exchange generation—step-by-step switches.

Theorem 3[10]. On optimality of grading-type call center at $\lambda \to 0$

At $\lambda \to 0$ and given rectangular-type call center with parameters: $n$ – skills (call flows), $v$ – total number of agents, $d$ – number of available agents for each call flow, the optimal call center design should follow the principles:

1). The skill-field $(n, d, v)$ is divided (as possible) in skill-sets with 1 and $n$ skill-points (it means that each agent has 1 or $n$ skills), and individualists are available earlier than generalists,

2). If the above requirement could not be fulfilled, then skill-sets with another skill-points are available after individualists and before generalists.

The idea for proof. The proof is based on Theorem 1. The coefficients $c_{n}(x)$ are related to the number of calls, moving system from state 0 to state $x$. For example, $c_{1,1}(x) = r_{x}/|x|$ (means the number of paths to move from state 0 to state $x$ by incoming calls only) and the first term in the expansion at $\lambda^{d}$ is equal to

$$B(0)/A(0) = \frac{1}{n} \sum_{|x|=d} r_{x} / (n-s(x))$$

Therefore, the first choice agents should be individualists. The proof follows from the analysis of the next terms.

Theorem 4[10]. On optimality of equally distributed skill-points at $\lambda \to \infty$

At $\lambda \to \infty$ and given rectangular-type call center with parameters: $n$ – skills (call flows), $v$ – total number of agents, $d$ – number of available agents for each call flow, the optimal call center design should follow the principles:

1). The skill-field $(n, d, v)$ is divided between agents with $r$ or $r+1$ skills, where $r=[nd/v]$ ([] denotes the whole part),

2). Agents with $r$ skills are available earlier than agents with $r+1$ skills,

3). Each agent has equal (as possible) number of different skills with each of the other agents.

Idea for proof. The proof is based on Theorem 3. The first two terms in expansion of loss probability in powers of $\lambda^{-1}$ don’t give any information on the scheme structure. The third term at power $\lambda^{-2}$, equal to $L/n$, gets minimum value when

$$L = \sum_{i=1}^{v} 1/l_{i}$$

takes minimum value. According to definition

$$\sum_{i=1}^{v} l_{i} = nd,$$

where $l_{i}$ are integer numbers, therefore they should be (as possible) equal i.e. equal to $[nd/v]$ or $[nd/v]+1$. From that the requirement 1 of Theorem 4 follows.

Let’s prove the requirement 3 for particular case when $nd/v$ is an integer, i.e. equal $r$. Then

$$M = \sum_{i=1}^{v} \frac{L_{i}}{r} = 0$$

and we need to minimize the first sum. If we denote the number of common inlets for lines $i$ and $j$ by $q_{ij}$ then

$$\sum_{i,j=1}^{v} s_{ij} = \sum_{i,j=1}^{v} (l_{i}+l_{j}-q_{ij}) = (v-1) \sum_{i=1}^{v} l_{i} - \sum_{i,j=1}^{v} q_{ij} = (v-1)nd - n \frac{d(d-1)}{2} = \text{const}$$

This sum takes minimum value when all $s_{ij}$ are equal (as possible). It means that each skill-set has equal (as possible) number of different skills with each of the other skill-sets, and the requirement 3 for particular case is proved necessary.

5. Discussion

To make clear the advantage of equally distributed skills solution, we shall give some additional numerical examples as a basis for future study.

Example 3. Universal agents influence

In Fig. 9, the two previous schemes are shown but with one more agent-generalist added as the last choice. It shows strong influence: the loss curves are crossing at total traffic of only 0.105 and when the loss probability is $0.23 \times 10^{-9}$,
i.e., almost zero. This means that it is reasonable to have more universal agents in reserve, but apply the equally distributed scheme.

**Example 4. Agents number doubling**

The other possible variant of modification is the doubling of the agents number in each of the six groups that were shown in Fig. 6. This way we will get two schemes with 12 agents in each (Fig. 10). The solution of the system of $2^{12}$ equations shows that the pattern almost does not change: the loss curves are crossind if the losses are 0.0032 (instead of 0.0025 in Fig. 6). However, this result may be achieved under total traffic of 3.65, which constitutes 31% of the agent’s workload, i.e., less than a third of the time. Generally, the overall picture shows no changes: the scheme with equally distributed skills among the agents is practically preferable at any load.

![Figure 10. Doubled agent groups (in comparing to Fig 6a and 6b)](image)

**Example 5. Waiting space influence**

Obviously, use of waiting places always reduces call losses: due to waiting places the loss curves are moved to the right. What is more interesting: due to waiting places the preference of equally distributed agent skills grows.

![Figure 11. Unexplained effect of call flow correlation.](image)

**Table 1.** The crosspoint of loss curves for schemes 6a and 6c with waiting places

<table>
<thead>
<tr>
<th>Number of waiting places</th>
<th>Total load</th>
<th>Loss probability crosspoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.72855</td>
<td>0.2542 x 10^-2</td>
</tr>
<tr>
<td>1</td>
<td>0.6427</td>
<td>0.9424 x 10^-4</td>
</tr>
<tr>
<td>2</td>
<td>0.616</td>
<td>0.4091 x 10^-3</td>
</tr>
</tbody>
</table>

The calculation shows that the waiting spaces increase the advantage of equally distributed schemes (Table 1). The absence of waiting spaces corresponds to the initial case in Fig. 6; i.e. the curves cross when the losses are 0.25%. Adding one waiting space for every call flow moves the concurrence point to 0.94 x 10^{-7}, and, if there are two waiting spaces, the curves cross even earlier when the losses are 0.41 x 10^{-9} only.

**Example 6. One unsolved problem**

See examples (Fig. 11) of three almost same call centers but with different agent access policy. The above studied scheme 11a and two similar schemes 11b and 11c are depicted, but there is a slight difference in the calls assignment, which should provide, in our opinion, more even distribution and higher performance. However, this change worsens the scheme’s traffic capacity (Table 2): the initial scheme 11a remains the most preferable, scheme 11c is the second, and scheme 11b is the third. This is hard to explain, and additional research is required. Seems it depends on the correlation between call flows.

![Figure 11. Unexplained effect of call flow correlation.](image)

**Table 2.** Comparison of schemes with equally distributed skill-points but different access policy

<table>
<thead>
<tr>
<th>Total traffic</th>
<th>Loss probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fig. 11a</td>
</tr>
<tr>
<td>4</td>
<td>0.67315</td>
</tr>
<tr>
<td>1</td>
<td>0.18903</td>
</tr>
<tr>
<td>0.25</td>
<td>0.6143 x 10^{-2}</td>
</tr>
<tr>
<td>0.0625</td>
<td>0.92525 x 10^{-4}</td>
</tr>
</tbody>
</table>

6. **Conclusions**

Numerical analysis of simple multi-skill call centers proposes a new principle for call center design: from throughput point of view, a multi-skill call center with equally distributed skills is preferable compared to traditional grading-type design. The strong proof is given by expansion of call loss probabilities in powers of call rate $\lambda$ and of $1/\lambda$, respectively.

**REFERENCES**


