Exponential Approximation of Multi-Skill Call Centers Architecture

Ger Koole and Jérôme Talim
Vrije Universiteit - Division of Mathematics and Computer Science
De Boelelaan 1081 a - 1081 HV Amsterdam - The Netherlands
koole@few.vu.nl - jtalim@few.vu.nl

In: *Proceedings of QNETs 2000*, Ilkley (UK), pages 23/1-10, 2000

Abstract

We model a multi-skill call center as a network of queues : Calls are considered as customers requesting service, agents as servers. A customer that finds all servers busy at a queue may be routed to another queue (if any) or is lost (otherwise). In order to evaluate the losses of such a network, we approximate each queue as an $M/M/r$ loss system. Based on simulations, we illustrate the efficiency of the exponential approximation, and its application to the design of a call center architecture.

Key words: Overflow Process - $M/M/r$ Loss System - Erlang Formula

1 Introduction

A call center is a more and more common service brought by companies such as banks, insurance or software companies to their customers. Via phone calls, faxes, e-mails, or web applications, it can advertise new products, or provide its customers with real time services such as bank account consulting, financial information desk, or technical assistance. A call center is made of telecommunications resources and human resources (namely the agents and their skills). One of the task of its management consists in determining the number of agents required, for each period of activity, in order to cope with the customers demand and to minimize their waiting time, or the rate of abandoned calls. The efficiency in achieving this goal lies in a detailed description of the architecture of a call center. A customer call requests one Type of Service (ToS) that an agent with the adequate skill can cope with. We will use equivalently the terms skill and ToS. The description will include then:

1. **The statistical characteristics of calls** (inter-arrival time and service time) which depend only on the ToS requested

2. **The** *skill set (SS)* **of each agent**: Based on it, we can make a partition of the set of all agents in G different groups; each group $j = 1...G$ is characterized by its skill set, $SS(j)$. A partial order between the groups can then be defined. the group j is "more generalist" that k if $SS(k) \subset SS(j)$.

3. **The call processing rules** which define the path of an arriving call within the center: Among all groups with agents that can serve the incoming call, the system should define the routing order. It is forwarded first to group j_1; if all agents of j_1 are busy, then this call is an overflow call from j_1 and it is forwarded to the next possible group j_2 and so on, until the system finds a group with at least one available agent. Possibly, if agents of all

relevant groups are busy, then the call may be stored (when there is any waiting room) or it is lost (otherwise). In this paper, we consider no waiting room and we are interested in evaluating the loss probabilities and the overflow processes. Limiting the number of lost calls in a loss system (i.e. without waiting room) would contribute, in practice, in limiting the waiting time in a queue with buffer. Furthermore, overflows from a loss system are simpler to evaluate than those from a system with a waiting room.

The quality of service provided to customers is related to the time spent in the system (waiting time and service time) or to the rate of lost calls due to the full activity within the center. So far, focus has been on functionality and less on the efficiency. Few research work has been done on this topic. One can refer to [1] for an approach of a call center design based on quality of service (in terms of waiting time) provided to customers. Garnett and al., in [2], introduce the concept of impatient customers in their model. Koole and van der Sluis, in [3], focus on the manpower scheduling aspect. In this paper, we investigate a calls processing approximation to evaluate the overflow process and the loss rate. We consider a multi-skill center, without any waiting room, which processes only incoming calls (or inbound calls) and that can be seen as a network of queues. The center is assumed to process S different and independent ToSs. Incoming calls from customers and requesting ToS i are modelled as a Poisson process, with rate λ_i. We also assume that calls processes requesting ToSs i and j are mutually independent. And from now on, we use equivalently the expression "calls of type i' and "calls requesting ToS i". The duration of a call of type i is modelled as an exponential distribution, with rate μ_i whatever the group the agent that will service the call belongs to. Finally, we introduce the ratio: $\rho_i := \lambda_i/\mu_i$. As mentionned previously, there are G groups in the center. Each one has R_j agents and is characterized by its skill set $SS(j)$. When there are S ToSs, the number of all possible groups is equal to $2^S - 1$. For any ToS $i = 1, ..., S$, let us denote with E_i the subset of groups which can serve calls of type i:

$$E_i := \{j = 1, ..., G | i \in SS(j)\}$$

A forwarding rule for calls of Type i is a permutation of elements of E_i, defining the order according to which overflow calls are routed. In practice, the center can define

- a deterministic forwarding rule, i.e., a unique predefined permutation,
- a probabilistic forwarding rule. In this case, the call routing path is chosen randomly (according to a predefined distribution) among all possible permutations of elements of E_i.
- a dynamic forwarding rule where the choice of a permutation is not predefined and depends on the activity of the groups in the system.

In this paper, we consider only deterministic forwarding rules, in order to keep the computation as simple as possible and to focus on the overflow approximation. Moreover, deterministic forwarding rules are quite common in practice, even if dynamic forwarding is implemented in some centers. For any ToS i, we will denote with n_i the number of elements in E_i and τ_i its forwarding rule. $\tau_i(1)$ is the first group towards which a call of type i will be forwarded. The overflow from $\tau_i(1)$ is sent to $\tau_i(2)$ and so on, until the group $\tau_i(n_i)$. If all agents of this latter are busy, the call is rejected, since we do not consider any waiting room. From now on, we use the terms of *exogeneous calls* for arrivals at the group $\tau_i(1)$, *overflow (calls)* from group j for calls that cannot be served by agents of j and *lost calls* for the overflows from group $\tau_i(n_i)$.

Due to the complexity of such a system, even with exponentially distributed inter arrival times and service times, there is no analytical formula dealing with the behavior of such a network of queues, in particular to evaluate the loss probabilities. From the previous description,

two important issues have to be solved :
(i) the overflow calls from a group characteristics (mean of inter overflow time, or further statistical description);
(ii) the service time, though exponentially distributed, depends on the ToS requested and not on the servicing agent.

Concerning the first issue, one can refer to [6] for the single queue system with an arbitrary waiting room size, and with exponentially distributed inter arrival time and service time. Riordan in [5] focuses on the overflow from the $GI/M/s$ loss system, while Halfin in [4] looked at the $GI/G/1$ loss system overflow. When considering the specific case of Markov-Modulated Poisson Process (MMPP) arrivals, one can refer to [8] for the general analysis of the $MMPP/G/1$ queue and to [9] for the networks of $MMPP/M/s/r$ queues analysis. The second issue deals with services of different types of customers, within a server. A general presentation of queues with priorities and with Poisson arrivals and exponential service time can be found in [11]. Gross and Harris ([7] - pp 195-205) analyzed an $M/M/1$ queueing system, with two types of customers competing for service at different rates. Back to the present problem, the approximation model for which we can evaluate the performance should be simple enough to be applied at any queue of the whole system. The remainder of the paper is organized as follows. In Section 2.1, we define our approximation scheme. We illustrate its efficiency through some numerical experiments in Section 2.2. In Section 3, we conclude with final remarks and present some further planned work.

2 Exponential Approximation

In this section, we approximate the inter overflow time process and the service times of each group with exponential distributions. Generally speaking, each group (with its R_j agents) can be viewed as a $G/G/R_j$ loss system, where the arrival process is the superposition of possibly exogeneous calls (which inter arrival time is exponentially distributed) and of overflows from other groups, and where the service time distribution depends on the ToS requested. Riordan [5] stated that the inter overflow time from an $M/M/r$ loss system has an hyperexponential distribution of order $r+1$ with a general form:

$$F(t) = \sum_{i=1}^{r+1} a_i(1 - exp(-b_i t)), \text{ with } \sum_i a_i = 1 \tag{1}$$

where the parameters a_i and b_i can be computed numerically. Meier-Hellstern [9] analyzed the overflow from an $MMPP/M/r/b$ system and proposed to approximate it with an hyperexponential distribution of order 2, or equivalently an Interrupted Poisson Process. Hordijk and Ridder, in [12], focus on inequalities for overflow approximations. We propose to approximate the inter overflow time process at each node in the network as an exponential distribution. Equivalently, the overflow process is approximated by a Poisson process.

Most of the work dealing with overflow process considers an exponentially distributed service time. For a more general distribution, one can refer to [4] for the $G/G/1$ loss system. In the present multi-skill call center model, the service time at a group of agents depends on the ToS requested. There is no homogeneous distribution. We propose to use an exponentially distributed service time, in order to keep the approximation model simple.

Each node in this network is modelled then as an $M/M/r$ loss system for which we can apply the Erlang Formula (See [5]-pp.39) giving the Loss Probability:

$$G(\rho, r) := \frac{\rho^r}{r!\left(1 + \rho + \cdots + \frac{\rho^r}{r!}\right)} \tag{2}$$

where λ is the arrival rate, μ the service rate and $\rho := \lambda/\mu$. The overflow rate from such a system is equal to: $\lambda\ G(\rho, r)$. One can notice that the Erlang formula is insensitive of the service time distribution, as long as the inter arrival times are exponentially distributed. For general service time distributions, we can use $\lambda\ G(\rho, r)$ as an approximation of the overflow rate. In the next section, we present the computation scheme of the approximation model

2.1 Definition of the Model and computation

For any node (or group) $j = 1...G$ in the network, we will denote with γ_j (resp. ν_j) the approximated arrival rate (resp. service rate). We introduce also the variables $\gamma_{i,j}$ the approximated type i arrival rate at the entry of the group j. Remember that the permutation τ_i is the routing rule of type i calls. Then, $\gamma_{i,\tau_i(1)}$ is (exactly) λ_i (the exogenous calls rate); $\gamma_{i,\tau_i(2)}$ the overflow rate from the group $\tau_i(1)$ or equivalently the arrival rate of type i calls at group $\tau_i(2)$, and so on. The approximation method consists in computing γ_j and ν_j and $\gamma_{i,j}$ at each node:

1. Approximated arrival rate at the entry of the group j:

   Based on the exponential approximation of the overflow process, the arrival rate of the group j is the sum of all relevant types calls rate:

   $$\gamma_j := \sum_{i \in SS(j)} \gamma_{i,j} \tag{3}$$

2. Approximated service rate of the group j:

   The mean service time of j is a weighted sum of the service times of each type i; the weights are proportional to the arrival rates $\gamma_{i,j}$:

   $$\frac{1}{\nu_j} := \sum_{i \in SS(j)} \frac{\gamma_{i,j}}{\gamma_j} \frac{1}{\mu_i} = \frac{1}{\gamma_j} \sum_{i \in SS(j)} \frac{\gamma_{i,j}}{\mu_i} \tag{4}$$

   The approximated service time distribution is exponential with rate ν_j.

3. Overflow rate:

   For each type of call i, the routing rule is defined by the permutation $\tau_i(k), k = 1, ..., n_i$.
   - For $k = 1$, we have:

   $$\gamma_{i,\tau_i(1)} := \lambda_i \tag{5}$$

   - For any $k \geq 1$, the total arrival rate at the node $\tau_i(k)$ (resp. service rate) is equal to $\gamma_{\tau_i(k)}$ (resp. $\nu_{\tau_i(k)}$). The loss probability is then given by: $G(\gamma_{\tau_i(k)}/\nu_{\tau_i(k)}, R_{\tau_i(k)})$ which yields the overflow rate from node $\tau_i(k), k = 2, ..., n_i$:

   $$\begin{aligned} \gamma_{i,\tau_i(k)} &:= \gamma_{i,\tau_i(k-1)}\ G\left(\frac{\gamma_{\tau_i(k-1)}}{\nu_{\tau_i(k-1)}},\ R_{\tau_i(k-1)}\right) \\ &= \gamma_{i,\tau_i(k-1)}\ G\left(\sum_{l \in SS(\tau_i(k-1))} \frac{\gamma_{l,\tau_i(k-1)}}{\mu_{\tau_i(k-1)}},\ R_{\tau_i(k-1)}\right) \end{aligned} \tag{6}$$

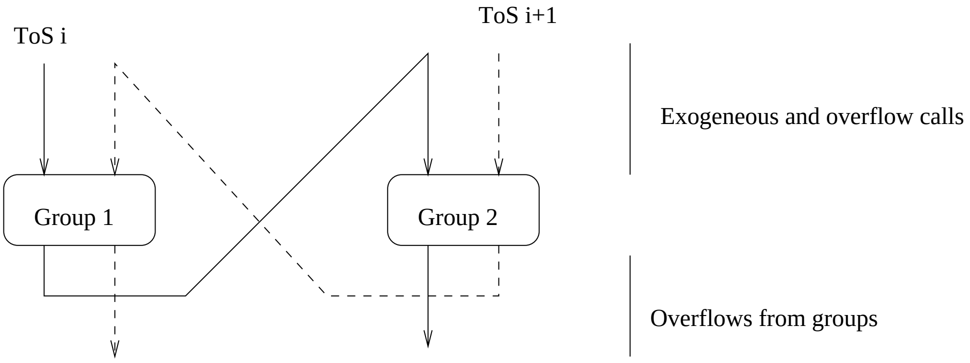

Figure 1: Crossed routing rules

We have completely defined the computation of the approximated rates. In order to evaluate at each node the overflow rate, we have to consider the role of the routing rule $\tau_i, i = 1, ..., S$, which can introduce supplementary steps in the method. Let us consider the example of a crossed routing rules, shown in Figure 1. Overflow calls from two different groups are symmetrically routed to the other. From the computation scheme presented previously, we have:

$$(P) \begin{cases} \gamma_1 = \lambda_i + \gamma_{i+1,1} \\ \\ \gamma_2 = \lambda_{i+1} + \gamma_{i,2} \\ \\ \gamma_{i,2} = \lambda_i \; G(\lambda_i/\mu_i + \gamma_{i+1,1}/\mu_{i+1}, R_1) \\ \\ \gamma_{i+1,1} = \lambda_{i+1} \; G(\lambda_{i+1}/\mu_{i+1} + \gamma_{i,2}/\mu_i, R_2) \end{cases} \tag{7}$$

where γ_1, γ_2, $\gamma_{i+1,1}$ and $\gamma_{i,2}$ have to be determined. This problem can be solved by introducing two sequences $\{U_n\}$ and $\{V_n\}$ associated respectively to $\gamma_{i+1,1}$ and $\gamma_{i,2}$ defined as:

$$\begin{cases} U_0 = 0 \text{ and } U_n := \lambda_{i+1} \; G(\lambda_{i+1}/\mu_{i+1} + V_{n-1}/\mu_i, R_2) \text{ for } n \geq 1 \\ \\ V_0 = 0 \text{ and } V_n := \lambda_i \; G(\lambda_i/\mu_i + U_{n-1}/\mu_{i+1}, R_1) \text{ for } n \geq 1 \end{cases} \tag{8}$$

If the two sequences have finite limits U_∞ and V_∞ which satisfy:

$$\begin{cases} U_\infty := \lambda_{i+1} \; G(\lambda_{i+1}/\mu_{i+1} + V_\infty/\mu_i, R_2) \\ \\ V_\infty := \lambda_i \; G(\lambda_i/\mu_i + U_\infty/\mu_{i+1}, R_1) \end{cases} \tag{9}$$

then, these limits, that can be computed by iteration, give the solutions to the initial problem (P).

The proof of the existence of the finite limits U_∞ and V_∞ uses the following Lemma:

Lemma 1 *The function $G(\rho, r)$ is increasing in respect to ρ for any fixed $r \geq 1$.*

Proof 1 *The increasingness of $G(\rho, 1)$ is trivial. We will consider from now the case where $r > 1$. Let us derive $G(\rho, r)$:*

$$\frac{\partial}{\partial \rho} G(\rho, r) \quad = \quad \frac{1}{\left(\sum_{i=0}^{r} \frac{\rho^i}{i!}\right)^2} \frac{\rho^{r-1}}{(r-1)!} \left(\sum_{i=0}^{r} \frac{\rho^i}{i!} \; - \; \frac{\rho}{r} \sum_{i=0}^{r-1} \frac{\rho^i}{i!} \right)$$

and develop the term in the parentheses:

$$1+\rho+\cdots+\frac{\rho^i}{i!}\cdots+\frac{\rho^r}{r!}-\left(\frac{\rho}{r}+\frac{\rho^2}{r}+\cdots\frac{\rho^i}{(i-1)!\;r}\cdots+\frac{\rho^r}{r!}\right)$$

This quantity is obviously (strictly) positive. This shows the increasingness of $G(\rho, r)$. ■

Using the monotonicity of $G(\rho, r)$ in respect to ρ, one can show by induction that the sequences $\{U_n\}$ and $\{V_n\}$ are increasing and are obviously upper bounded. So, they have finite limits that can be computed numerically.

This iterative method will be used to compute the rates of all groups within the whole network, for general routing rules $\tau_i, i = 1, ..., S$. We implemented the exponential approximation according to the computation scheme shown in Table 1. The next section deals with the analysis of some numerical examples.

1. Equations of the system:

   Write the equations (5) and (6) for all ToSs $i = 1, ..., S$:

   - $\gamma_{i,\tau_i(1)} := \lambda_i$
   - $\gamma_{i,\tau_i(k)} := \gamma_{i,\tau_i(k-1)}\; G\left(\sum_{l\in SS(\tau_i(k-1))} \frac{\gamma_{l,\tau_i(k-1)}}{\mu_{\tau_i(k-1)}},\; R_{\tau_i(k-1)}\right)$ for all $k = 2, ..., n_i$.

2. Iterative computation:

   For each $\gamma_{i,j}$, we define the sequence $\left\{\gamma_{i,j}^{(n)}\right\}_{n\geq 0}$ such that:

   - $\gamma_{i,j}^{(0)} = \begin{cases} \lambda_i & \text{if } j = \tau_i(1) \\ 0 & \text{otherwise} \end{cases}$
   - $\gamma_{i,\tau_i(k)}^{(n+1)} = \begin{cases} \lambda_i & \text{if } k = 1 \\ \gamma_{i,\tau_i(k-1)}^{(n)}\; G\left(\sum_{l\in SS(\tau_i(k-1))} \frac{\gamma_{l,\tau_i(k-1)}^{(n)}}{\mu_{\tau_i(k-1)}},\; R_{\tau_i(k-1)}\right) & \text{if } 2 \leq k \leq n_i \end{cases}$

3. Solutions of the system:

   Compute numerically the sequences $\left\{\gamma_{i,j}^{(n)}\right\}_n$ for all i, until:

   $$|\gamma_{i,j}^{(n+1)} - \gamma_{i,j}^{(n)}| < \epsilon, \forall j \in \{\tau_i(1), ..., \tau_i(n_i)\}$$

   where ϵ is the expected accuracy. For instance, we can use: $\epsilon = 10^{-10}$.

Table 1: Computation scheme

2.2 Numerical Experiments

We will apply the computation to the Example of the Figure 2 and which description is given in the Table 2. We consider a call center which can process 5 different ToSs. It has 12 groups with different skill sets. For different configurations (number of agents allocations, and routing rules), we applied the approximation model, on one hand and simulated the call center model on the other hand. We are interested in the lost calls rate in both systems and for all ToSs and the relative difference between both models. The results are displayed in the Table 3. Based on these results, we can notice:

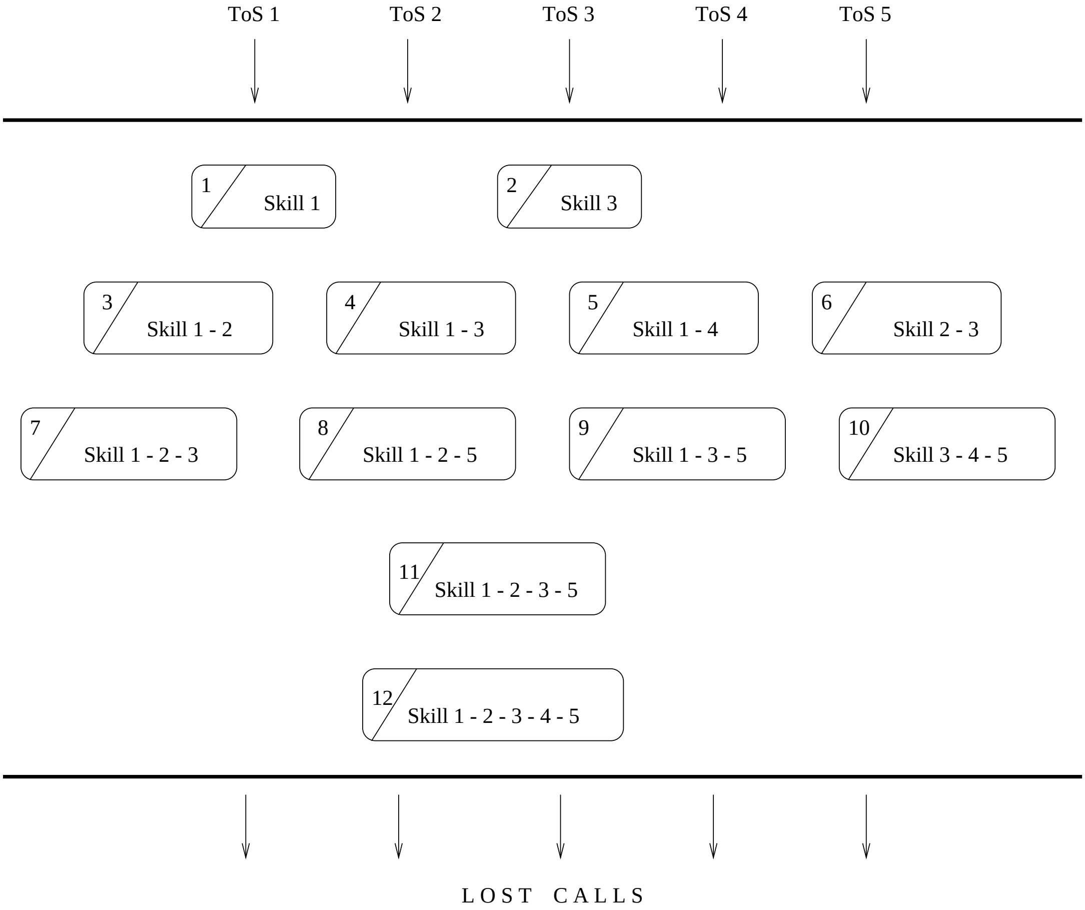

Figure 2: Example of a call center architecture

ToS	1	2	3	4	5
λ_i	0.604	0.590	0.620	0.717	0.019
μ_i	0.023	0.180	0.030	0.195	0.018

Group	1	2	3	4	5	6	7	8	9	10	11	12
Number of agents (i)	3	2	4	5	3	2	2	3	4	10	1	7
Number of agents (ii)	6	4	8	10	6	4	4	6	8	20	2	14
Number of agents (iii)	0	0	5	6	4	4	2	3	4	10	1	7
Number of agents (iv)	5	4	4	1	3	2	2	3	4	10	1	7

Table 2: Example description

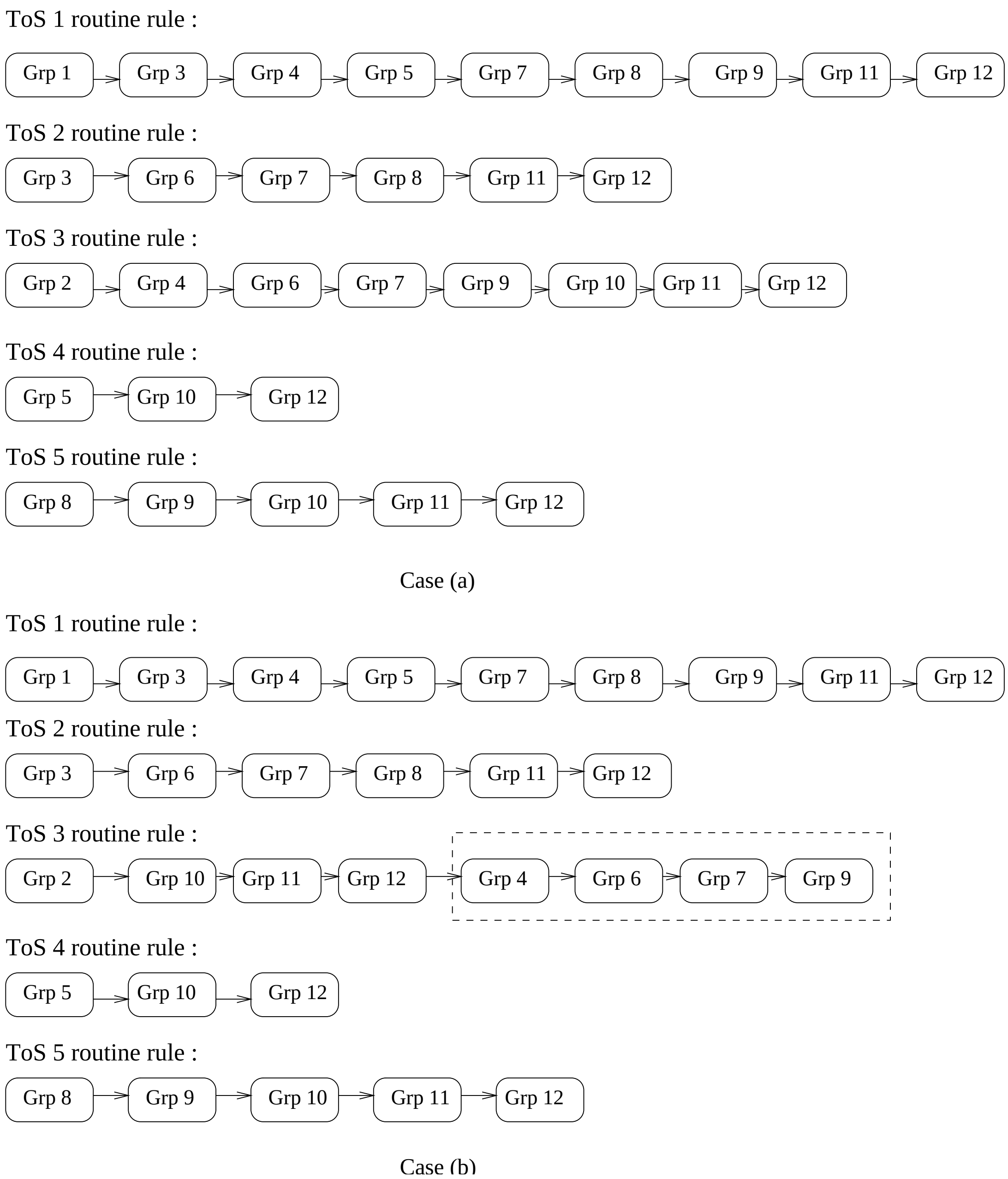

Figure 3: Routing rules

Examples		Simulation (1)	Approximation (2)	$\frac{\|(1)-(2)\|}{(1)}$
1 -	routing rule (a) number agents (i)	0.649	0.657	0.012
2 -	routing rule (b) number agents (i)	0.710	0.709	0.001
3 -	routing rule (a) number agents (ii)	7.10e-5	1.29e-25	1
4 -	routing rule (a) number agents (iii)	0.553	0.544	0.017
5 -	routing rule (a) number agents (iv)	0.660	0.651	0.014

Table 3: Lost calls rates

- The efficiency of the approximation:

  In most cases, the results illustrate the efficiency of the approximation model. However, the Case 3 where the difference between simulation and approximation models points out the weakness of our method for some specific instances. In this Case 3 (in comparison to case 1), we doubled the number of agents allocated to each group. Thus, (much) less calls were lost. Remember that the approximation consists in multiplying the arrival rate at each node by the loss probability. For a ToS $i = 1, ..., S$, the loss rate equals:

  $$\lambda_i \cdot P_{\tau_i(1)} \cdot P_{\tau_i(2)} \cdots P_{\tau_i(n_i)} \tag{10}$$

  where P_j is the loss probability at group j. When there are few overflow calls or when P_j is "small", the difference between simulation and approximation is emphasized at each downward node. In such cases, agents are likely to experience some low activity periods, waiting for calls. In practice, one of the staff management aim is to minimize the idle periods or not to over estimate the required number of agents.

- The routing rule impact:

  The comparison of Cases 1 and 2 illustrates the impact of the routing rule in performance of the system. Roughly speaking, a ToS i calls process is more "greedy" than another ToS j calls process, when the ratio $\rho_i = \lambda_i/\mu_i$ is larger than ρ_j. In the Case 2, we changed the routing rule of ToS 3 calls, which process is more greedy that ToSs 2, 4 and 5 calls processes. Due to this change, ToS 3 calls are more likely to use agents in the system and to deprive the others of getting service.

- The use of the approximation to design a call center:

  The Cases 1, 4 and 5 are different in the number of agents allocation. In any case, the sum of all agents used in the system is constant. The analysis of these cases suggests that the use of experts (one skill agents) for greedy ToS calls process does not improve the performance of the system.

3 Conclusion and Future Research

We have proposed to model a multi-skill call center as a network of $M/M/r$ loss systems, for which we can evaluate at each node the overflow process. Based on the simulation of the system, we demonstrate the efficiency of the approximation model, in all practical cases, up to a performance difference of 1 or 2%. Therefore, this model can be used in the performance evaluation of call centers (to evaluate the loss rate or to evaluate the activity rate of agents). It can also be used to design a center or to allocate agents to each group, in order to minimize the loss rate, using the statistical characteristics of calls (inter arrival time and service time). And finally, the exponential approximation may be used for further investigation of the (dynamic) forwarding rules.

References

[1] S. Borst and P. Serri Robust Algorithms for Sharing Agents with Multiples skills Preprint, sem@cwi.nl

[2] O. Garnett, A. Mandelbaum and M. Reiman Designing a Call Center with Impatient Customers October 8, 1999, Preprint

[3] G. Koole and E. van der Sluis An optimal local search procedure for manpower scheduling in call centers Technical Report WS-501, Vrije Universiteit Amsterdam, 1998, http://www.math.vu.nl/obp/callcenters

[4] S. Halfin Distribution of the interoverflow time for the GI/G/1 loss system Mathematics of Oper. Res., Vol 6, No. 4. November 1981, 563-570

[5] J. Riordan Stochastic Service Systems John Wiley and Sons, New York, 1962

[6] E. A. van Doorn *On the Overflow Process from a Finite Markovian Queue*, Perf. Eval. 4 (1984), 233-240.

[7] D. Gross and C. M. Harris *Fundamentals of Queueing Theory, Second Edition*, John Wiley & Sons, 1985

[8] W. Fisher and K. Meier-Hellstern *The Markov-modulated Poisson process (MMPP) cookbook*, Perf. Eval. 18 (1992), 149-171.

[9] K. Meier-Hellstern *The Analysis of a Queue Arising in Overflow Models*, IEEE Trans. on Comm. vol 37, No 4, April 1989, 367-372

[10] D. Medhi *Some Results for Renewal Arrival to A communication link*, In Probability Models and Statistics: A J. Medhi Festschrift edited by A. C. Borthakur and H. Choudhury, New Age International Limited, New Delhi, India, pp. 85-107, 1996 http://www.cstp.umkc.edu/~dmedhi

[11] N.K. Jaiswal Priority queues Mathematics in Science and Engineering, Vol. 50 Academic Press, New York-London 1968

[12] A. Hordijk and A. Ridder Stochastic Inequalities for an Overflow Model J. Appl. Prob. 24, 696-708 (1987)