Analysis of the Impact of Team-Based Organizations in Call Center Management

Oualid Jouini, Yves Dallery, Rabie Nait-Abdallah

Analysis of the Impact of Team-Based Organizations in Call Centers Management

Oualid Jouini[†] **Yves Dallery**[†] **Rabie Nait-Abdallah**[‡]

[†] Laboratoire Génie Industriel
Ecole Centrale Paris
Grande Voie des Vignes
92295 Chatenay-Malabry Cedex, France

walid.jouini@ecp.fr, yves.dallery@ecp.fr

[‡] Bouygues Telecom
20 Quai du Point du Jour
92640 Boulogne Billancourt Cedex, France

rnait@bouyguestelecom.fr

Abstract

We investigate the benefits of migrating from a call center where all agents are pooled and customers are treated indifferently by any agent, towards a call center where customers are grouped into clusters with dedicated teams of agents. Each cluster is referred to as a portfolio. Customers of the same portfolio are always served by an agent of the corresponding team. There is no specialization involved in this organization in the sense that all customer portfolios as well as all agent teams have (statistically) identical behaviors. The reason for moving to this organization is that dealing with teams of limited size allows a much better workforce management compared to the situation usually encountered in large call centers.

The purpose of this paper is to examine how the benefits of moving to this new organization can outweigh its drawback. The drawback comes from the fact that there is less pooling effect in the new organization than in the original one. The benefit comes from the better human resource management that results in a higher efficiency of the agents, both in terms of speed and in terms of the quality of the answer they provide to customers. Our analysis is supported by the use of some simple queueing models and provides some interesting insights. In particular, it appears that for some reasonable ranges of parameters, the new organization is attractive in the sense that it can outperform the original organization. We then extend the analysis to the case where in addition to the identified customer portfolios, there is an additional flow of calls called out-portfolio flow. It is shown that this feature makes the new organization even more efficient.

Keywords: call centers management, quality of service, pooling, human resource management, customer portfolio management, team-based organization, queueing models

November 09, 2004; revised June 03, 2005 and January 26, 2006.
To appear in *Management Science*.

1 Introduction

Recently, the number of call centers has been increasing drastically. Call centers are a means of customer relationship management and their role is growing. A call center is a complicated service system, in which managers must take into account the behavior of both customers and agents, see Gans et al. (2003). Many issues are related to call centers: an issue of corporate image which is related to the quality of service perceived by customers, a financial issue due to the cost of labor, and finally a human resource issue related to the social conditions of work. The purpose of this paper is to provide some insights into the impact of internal organization of call centers on their performances. It is the result of a collaboration with *Bouygues Telecom*, a French mobile phone company. *Bouygues Telecom* call center handles an average of 100,000 phone calls daily. Some of the calls are treated by an automated operator. Agents, also called customer representatives, deal with about 60% of these contacts. There are also about one million contacts per year handled by mail, e-mail and fax. Here, we investigate the adequacy of migrating from a call center where all agents are pooled and customers are treated indifferently, towards a call center where customers are grouped into clusters with dedicated agents. In our terminology, each cluster will be called a portfolio. The customers that do not fit into a precise portfolio generate the so-called out-portfolio flow, and must wait in a lower priority out-portfolio flow queue. Managers of *Bouygues Telecom* believe that the challenge is not only to answer quickly but also to answer customers correctly. In this sector (mobile telephony), it is not rare to see customers switching from one company to another as a consequence of low quality responses provided by customer representatives. Agents are the interface between the company and the customers; hence, customer satisfaction is closely linked to agents. Managers need to motivate their employees so that the assistance they provide to customers is efficient, both in terms of speed and quality of answers. On the other hand, employees need to feel strongly supported by the company so that the turnover is as low as possible. In fact, turnover means training new employees, and it implies more costs.

The aim of *Bouygues Telecom* through migrating into customer portfolio management is to better manage their employees and as a consequence to satisfy customers more efficiently. This management approach makes agents more responsible towards their own customers. Moreover, partitioning agents into groups creates competition, which increases agents' motivation. These

factors result in overall agents' efficiency improvement, both quantitatively and qualitatively. By quantitative efficiency, we mean the speed (processing time) in providing assistance to the customers. By qualitative efficiency, we mean the quality provided by the agents when addressing the customers' request. In this paper, we argue that these advantages may outweigh the variability that results from the loss in economy of scale originally associated with the pooled system. In addition, in the proposed organization, all portfolios and corresponding sets of dedicated agents are identical (statistically). Therefore, issues such as training and forecasting can be done in a homogeneous manner. Also, having homogeneous teams yields a more efficient human resource management. In fact, it allows the call center manager to compare the teams performances, which results in a "global competition".

Such a managerial approach has been widely and successfully used in industry and is also likely to be of interest in service activities such as call center operations. It is, indeed, one of the key success factors of the so-called World Class Manufacturing. For example, Schonberger (1986) refers to it as cellular manufacturing and describes its benefits as follows: "Cells create responsibility centers where non existed before. The cell leader and the work group may be charged with making improvements in quality, cost, delays, etc."

The remainder of this paper is structured as follows. In Section 2, we review two kinds of literature close to our work. The first one is on pooling, and the second is on integrating human factors in queueing systems, and in particular in call centers. In Section 3, we give a comprehensive presentation of the problem we study in this paper. In Section 4, we develop a simple queueing model that is then used to address the issue of benefits versus costs of migrating from the pooled organization to the dedicated organization when there is no out-portfolio flow. We provide some interesting insights on the tradeoff between reduction of the pooling effect and agents' efficiency improvement, both quantitatively and qualitatively. In Section 5, we extend this analysis to the situation where there is an out-portfolio flow. To do that, we first develop some approximate queueing models of the call center operating with a mix of portfolio and out-portfolio flows. One additional insight is that the drawback of not having a totally pooled system is less important in this context. Finally, we conclude and propose some directions for future research.

2 Literature Review

There is an extensive and growing literature on call centers, and in general on contact centers. We refer the reader to Gans et al. (2003) and Whitt (2002) for an overview. Our work is related to two streams of literature, one dealing with pooling and the other with human factors in queueing systems. The literature dealing with pooling falls mainly into two categories: pool queues or pool servers, see Mandelbaum and Reiman (1998). Kleinrock (1976) is one of the first researchers who gave a depiction of these alternative structures of pooling. He began by considering a collection of m identical $G/G/1$ queues, each of which has a single server with service rate μ and faces a job stream at rate λ. Pooling only queues would change this collection into a $G/G/m$ queue, which has m servers, each server with service rate μ and a job stream at rate $m\lambda$. Pooling only servers would change the last $G/G/m$ queue into a $G/G/1$ queue with arrival rate $m\lambda$ and service rate $m\mu$. In this paper we only deal with queue pooling issues.

There are two important aspects in the design of call centers. The first one deals with the issue of skills: are all agents cross-trained with all skills (full-flexible call centers) or are the agents only trained for a subset of skills? In the later case, what are the subsets of skills that will be considered and how many agents will have each subset of skills? A typical example of such multi-skill call centers is an international call center where incoming calls are in different languages, see Gans et al. (2003). Related studies include those by Garnett and Mandelbaum (2001), Akşin and Karaesmen (2002) and references therein. The second aspect deals with the issue of the level of pooling in call centers, i.e., are the agents all gathered into a single large team or are they partitioned into a set of independent teams? This issue is encountered in general in multi-skill call centers but in particular in full-flexible call centers, i.e., call centers in which all agents have all skills (all agents are flexible enough to answer all requirements of service). Our concern in this paper is a full-flexible call center. It is a plausible assumption for many real cases, especially for unilingual call centers where the complete flexibility is not as difficult as in multilingual call centers. Several papers discuss the effectiveness of pooling in multi-server queues, see for example Tekin et al. (2004). Beyond service systems, pooling effect problems arise in various applications, such as manufacturing, and computer network systems. The standard argument for combining queues is due to the economies of scale advantage, which absorbs stochastic variability (Borst et al. (2004)).

While it is easy to see that pooled systems are more effective than independent ones, this intuition was for a long time based on experience and numerical data rather than rigorous mathematical proof. Smith and Whitt (1981) were the first to formally prove this result, when combining systems with identical service time distributions. They applied analytic methods for the $M/M/s/s$ loss systems (Erlang-B, no waiting room) and the $M/M/s$ delay systems (Erlang-C, infinite waiting room). By using sample-path methods, they also showed that efficiency increases through combining queues in systems with general arrival processes and general service time distributions. Benjaafar (1995) extended theses results by providing performance bounds on the effectiveness of several pooling scenarios. When we allow service rates in separate systems to become different, combining queues can be counterproductive (Smith and Whitt (1981), Benjaafar (1995)). van Dijk and van der Sluis (2006) present a case-study simulation supporting this outcome. Using approximations for $M/G/s$ performance measures, Whitt (1999) explored the tradeoff between economies of scale associated with larger systems and the benefit of having customers with shorter service times separated from customers with longer service times.

All of the above results do, in no way, take into account the human element. This takes us to the second area of literature close to our work. Human element is the main characteristic of call centers and contact centers. Both customers and agents are people. Even though it is natural to focus on understanding human behavior, few papers integrate this aspect to analyze call centers and, in general, queueing systems. We refer the reader to the survey of Gans et al. (2003) where we find some references examining queueing models of call centers that incorporate customers' behaviors, such as, abandonments and retrials. Some other models include the link between agents' and customers' experiences. In 1987, two papers have launched discussions about human factors in queueing systems. The first is Larson's (1987) paper which goes beyond the classical interest on delays and points out the psychological experiences of people in queues. He argues the importance of perceptions of fairness, and shows for instance how the violation of the first come, first served order may contribute to customers' dissatisfaction. The second is Rothkopf and Rech's (1987) paper, which deals with the question of combining queues. They discuss the tradeoff between pooled and separated systems by including customers' reactions and jockeying between separate queues (a customer can change to one queue while he was waiting in another). Moreover, they show how separate systems may lead to servers that are

more responsible towards their own customers. It may also allow for a faster service due to the degree of specialization gained through experience. To our knowledge, they were the first to emphasize this issue.

Fischer et al. (1999) conclude that call center management requires a mix of disciplines that are not typically found in organizations. The review of Boudreau et al. (2003) follows through this new area. They propose a framework which is a fertile source of research opportunities. They justify by real examples that operations management itself, without human resource management, can not well analyze systems such as those we are dealing with, and vice versa. In others words, there is a mutual impact between the two fields, and taking into account this fact yields to more realistic and precise insights. In particular, Boudreau et al. (2003) consider that more realistic operations management models need to integrate human factors, such as; turnover, motivation and team structure. In fact, a team setting allows for better communication, and may allow for more responsible and motivated agents. In a recent paper, Boudreau (2004) underlines once again the significant opportunities for fruitful research at the boundaries between the traditional topics of operations management and human resource management. In this paper, we address this issue in a call center context. We explore how managing agents by creating separate pools might lead the agents performing more efficiently.

3 Problem Setting

In this section, we present the general problem under consideration in this paper. Consider a company operating a fairly large call center. The call center provides assistance to the customers of the company. Customers call the company whenever they need assistance and their request is addressed by a set of agents (or customer representatives). In the setting of this paper, we assume that the call center is operated in such a way that all agents have the same skill. Therefore assistance to the customers can be provided by any agent. In other words, all agents are totally identical (statistically) in the sense that they can answer all questions coming from the customers with the same efficiency, both quantitatively and qualitatively.

3.1 Current Organization Mode

Let us describe the behavior of the call center under the current organization mode. The call center is operated in such a way that at any time, any call can be addressed by any agent.

So, whenever a call arrives, it is addressed by one of the available agents, if any. If not, the call is placed into a queue and will be addressed as soon as possible. There is a single queue and waiting calls are answered on a first come, first served (FCFS) basis. For simplicity, we assume that the queue has no capacity constraint and that customers do not abandon while waiting. Under this organization, the agents have a given efficiency. The quantitative efficiency is measured by the distribution of the processing times, which represents the time it takes for an agent to answer a call. Note that the randomness of the processing times comes in particular from the variety of questions asked by the customers. The qualitative efficiency is measured by the probability of successfully answering the question of the customer. We assume that if the call has not been addressed in an adequate manner, the customer will call back to get assistance from another agent. This concept of call resolution probability was argued by de Véricourt and Zhou (2005) in a call routing problem. As for the global efficiency of the call center under the current organization, its positive side comes from the pooling effect. Its negative side is in terms of human resource (HR) management, given that, it is usually very difficult to have an efficient management of a large set of agents in a large call center.

3.2 New Organization Mode

Let us describe the following new organization mode. The set of agents is split into a set of independent teams. The teams are homogeneous in the sense that they have the same number of agents and that all agents have the same skills. In other words, there is no specialization. Let n be the number of independent teams.

In the new organization, in addition to the partitioning of the total number of agents in a set of autonomous teams of agents, there is also a partitioning of the customers into a set of n customer portfolios. Again, this partitioning is done in such a way that the portfolios are homogeneous. In other words, the overall request coming from the different customer portfolios are statistically identical. So, whenever a call arrives from a customer of a given portfolio, it is routed to the corresponding team. The behavior at the team level is then exactly identical to that described above for the original large call center. This new organization is equivalent to operating independently n smaller call centers with each call center having its own customers portfolio. Again, it is important to emphasize that under this new organization, all teams and

customer portfolios have the same behavior (statistically).

In the research study we performed with *Bouygues Telecom*, the size of the original call centers (total number of agents) was in the order of 2000, and they were considering team sizes ranging from 40 to 100 agents. Because all agents are not always present, this would mean that the number of agents simultaneously present in the call center would be in the order of 1000 and the corresponding number of agents present in each team would be ranging from 20 to 50. The reason advocated for moving to this new organization was along the line of the World Class Manufacturing literature. Namely, that the human resource management could be performed in a much better way at a small team level rather than at the global call center level. Agents' motivation and responsibility would increase. Performance measures, both quantitative (processing times) and qualitative (rate of calls successfully addressed), could be examined more appropriately and could be used for internal team management. Due to the team/portfolio one-to-one link, a customer not satisfied with the answer he got from the agent would call back and the additional burden would fall on the same team. Also, the fact that all teams are homogeneous would allow for performance comparisons between the different teams resulting in a "global competition". Incentives could be given to agents based on the global performance of the team.

3.3 Research Objectives

In this research project, our goal is to study the tradeoff between the pros and cons of moving from the original organization to the team-based organization, also referred to as the portfolio organization. To do that, we consider a simple stochastic model of the original pooled organization. This model captures the original behavior of the call center when all agents are pooled. Under this situation, the call center has a nominal behavior in terms of efficiency (quantitative and qualitative efficiencies). It achieves a given quality of service (QoS). We actually consider two different QoS measures: the average waiting time and the 80/20 rule, which is an industry standard for telephone service, see Gans et al. (2003). Under the 80/20 rule, at least 80% of customers must wait no longer than 20 sec.

In this work, we study the increase of efficiency required so that the team-based organization achieves the same QoS as the pooled system with the same total number of agents, i.e., no

cost increase. We successively consider the case where the improvement comes only from the increase of the quantitative efficiency (decrease of the average processing time) and then the case where the improvement comes only from the increase of the qualitative efficiency (increase of the average rate of successful answer). To perform the different analyzes described above, we consider a generic model that captures the important features needed for the comparison between the pooled organization and the team-based organization. As it is often the case in call center modeling, our analysis is based on the use of a stationary queueing model (Gans et al. (2003)). We use standard assumptions on the nature of the underlying processes: Poisson arrival processes, and exponential service time. These assumptions are plausible for *Bouygues Telecom* as for many other cases of call centers, especially for the arrival processes. In addition, we assume that calls not satisfied successfully occur randomly and therefore the splitting of the output flow follows a Bernoulli process. Moreover, delays before customers call back are assumed to be i.i.d. random variables. This allows us to use simple results from standard queueing theory. The generic model under consideration (for both the pooled organization and the team-based organization) is illustrated on Figure 1. The results of our study are presented in Section 4.

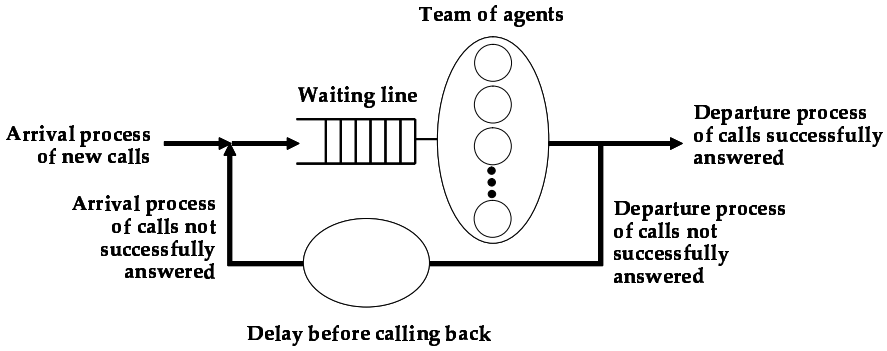

Figure 1: The generic model

3.4 Out-Portfolio Flow

There was another important feature in the *Bouygues Telecom* call center. Not all the calls received at the call center could be identified as belonging to a given portfolio of customers. Therefore, the actual situation was that in addition to the flow of calls coming in for each portfolio, there was an additional flow of independent calls, referred to as out-portfolio calls. In

the original pooled organization, the calls were treated with lower priority. For the team-based organization, all out-portfolio calls are sent to a single queue. These calls can be served by any agent of any team, but portfolio calls have (non-preemptive) priority over out-portfolio calls. This means that when an agent becomes available, he deals with a call from his portfolio first (the first call in line). If the queue is empty, this agent provides service to a call from the out-portfolio queue (the first in line). Under this more general setting, we also investigate the improvement in either quantitative or qualitative efficiency that would be required to counterbalance the unpooling effect. The results of our study are presented in Section 5.

4 Analysis of the Efficiency of the Team-Based Organization

In this section, we restrict our attention to the situation where there is no out-portfolio flow. In Section 4.1, we present the relevant queueing models and determine the performance measures of interest. In Sections 4.2 and 4.3, we analyze the required increase in terms of quantitative or qualitative efficiency that must be achieved in order to counterbalance the reduction of the pooling effect. Finally, in Section 4.4, we provide some further insights on the advantages of the team-based organization.

4.1 Modeling and Performance Analysis

Consider first the queueing model of the original call center. The model consists of a single infinite queue and a set of s identical servers representing the agents. Service times are assumed to be exponentially distributed with rate μ. The arrival process of first-attempt calls (primary calls) is assumed to be Poisson with an arrival rate of λ^a. There is a probability α that the customer is not satisfied with the answer he got and therefore will call again. Thus, $(1-\alpha)$ represents the probability that a call is successfully answered. Delays before customers call back are assumed to be i.i.d. random variables with a general distribution. For tractability purposes, we assume independence between successive calls, both in terms of service times and probability of success. Let λ be the overall arrival rate to the queue, i.e., the sum of the primary calls and the feed-back calls. Under stability conditions, $\lambda = \lambda^a/(1-\alpha)$. This simple model falls into the class of product-form networks analyzed by Baskett et al. (1975). As a result, the stationary behavior of this queueing model does not depend on the distribution of the call-back delays. They can thus be ignored. The resulting model is shown on Figure 2. It is equivalent

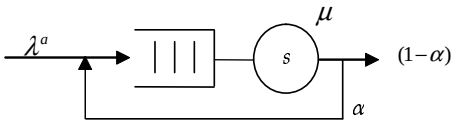

Figure 2: Pooled System

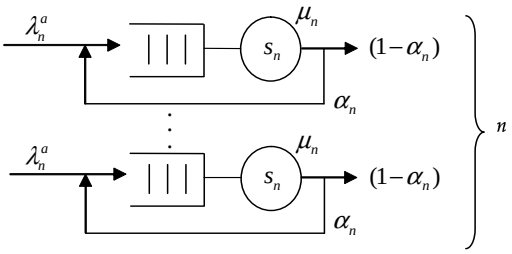

Figure 3: Dedicated System

to a simple $M/M/C$ queue with $C = s$ servers, a Poisson arrival rate $\lambda = \lambda^a/(1-\alpha)$ and a service rate μ. This model will be referred to as the Pooled System.

Consider now the modeling of the team-based organization with a partitioning of the original call center into n autonomous teams, each one being associated with a customer portfolio. It is assumed that n is such that s is a multiple of n. Recall that all teams and all customer portfolios are statistically identical. Under this organization, the call center can be modelled as a set of n independent and identical queueing models. In the following, we focus our attention on the generic model of any portfolio/team subsystem. The assumptions are similar to those above. The model consists of a single infinite queue and a set of s_n identical servers. Service times are assumed to be exponentially distributed with rate μ_n. The arrival process of first-attempt calls (primary calls) is Poisson with an arrival rate of $\lambda_n^a = \lambda^a/n$. The probability that a call is not successfully answered is given by α_n. The resulting model is shown on Figure 3. It is equivalent to multiple simple $M/M/C$ queues, with $C = s_n$ servers, a Poisson arrival rate $\lambda_n = \lambda_n^a/(1-\alpha_n)$ and a service rate μ_n. This model will be referred to as the Dedicated System. Note that the Pooled System can be viewed as the particular case of the Dedicated System for $n = 1$. In the *Bouygues Telecom* call center, like in most call centers, the arrival rate of calls varies over time. Therefore, we use queueing models to estimate stationary system performance of half-hour intervals. We assume constant number of agents, and constant arrival and service rates, as well as a system that achieves a steady-state quickly within each half-hour interval of time, see Gans et al. (2003).

Consider the Dedicated System. Let $W_n(t)$ be the Cumulative Distribution Function of the waiting time in the queue. Let $r_n = \lambda_n/\mu_n = \lambda_n^a/(1-\alpha_n)\mu_n$ be the traffic intensity, and $\rho_n = r_n/s_n$ the server utilization (proportion of time each server is busy). Note that the

condition for existence of a steady-state solution is $\rho_n < 1$; that is, the mean total arrival rate must be less than the mean maximal service rate of the system. As in Gross and Harris (1998), Equation (1) gives the probability that the waiting time in the queue is less than t.

$$W_n(t) = 1 - \frac{r_n^{s_n}\, p_n^0}{s_n!\,(1-\rho_n)} e^{-(s_n\mu_n - \lambda_n)\,t}. \tag{1}$$

Equation (2) gives the expression of the average waiting time in the queue.

$$W_n = \left(\frac{r_n^{s_n}}{s_n!\,(s_n\mu_n)\,(1-\rho_n)^2} \right) p_n^0. \tag{2}$$

Equation (3) gives the expression of p_n^0, which is the stationary probability of finding no customers in the system.

$$p_n^0 = \left(\sum_{i=0}^{s_n-1} \frac{r_n^i}{i!} + \frac{r_n^{s_n}}{s_n!\,(1-\rho_n)} \right)^{-1}. \tag{3}$$

The above equations give performance measures of the Dedicated System with n teams. The Pooled System is a special case of the Dedicated System. Performance measures of the Pooled System are obtained by using $n = 1$, $s_1 = s$, $\mu_1 = \mu$, and $\lambda_1^a = \lambda^a$ in Equations (1), (2) and (3).

Due to the variability effect in the arrival and service processes, the comparison between the Pooled System and the Dedicated System will always show that, for any positive integer $n \geq 2$, the Pooled System outperforms the Dedicated System under the same situation, i.e., $s_n = s/n$, $\mu_n = \mu$, $\lambda_n^a = \lambda^a/n$, and $\alpha_n = \alpha$. Under these conditions, it is intuitively clear that the Dedicated System is less efficient because a call may wait for one server (of one team) while another server (of another team) is idle; such a situation does not occur in the Pooled System.

In this section as in the next one, we perform the study of the quantities of interest as a function of the number of dedicated pools, n. Alternatively, we could have chosen to perform this study according to the size of the dedicated pools, s/n. However, because the total staffing level s is fixed in our study, the two analyzes are totally equivalent and the conclusions drawn in terms of n can readily be interpreted in terms of s/n.

4.2 Evaluation of Service Rate Percentage Increase

We start from a Pooled System with a given QoS ($W(t)$ or W), and our purpose is to evaluate the required service rate in a Dedicated System with n pools in order to ensure the same QoS ($W_n(t) = W(t)$ or $W_n = W$). The total staffing level, the total arrival rate of first-attempt calls, and the call back proportion are all held constant.

In the Pooled System, the arrival rate of first-attempt calls is $\lambda^a = 177.36$ calls per min, the call back proportion is $\alpha = 10\%$, the service rate is $\mu = 0.2$ calls per min, and the number of agents is $s = 1000$. In this system, 80% of customers wait no more than 20 sec, and the corresponding average waiting time is $W = 0.18$ min. In the Dedicated System, each call center has a staffing level $s_n = s/n$, an arrival rate of first-attempt calls of $\lambda_n^a = \lambda^a/n$, and a call back proportion of $\alpha_n = \alpha = 10\%$. We vary n from 1 to 50. For each number n of separated call centers, we calculate the service rate μ_n, so that, the average waiting time is $W_n = 0.18$ min. We repeat the same analysis for the QoS in terms of the 80/20 rule. The results are presented in Table 1.

	$W_n = 0.18$		$W_n(20sec) = 80\%$	
n	μ_n	ρ_n	μ_n	ρ_n
1	0.200	98.53%	0.200	98.53%
2	0.202	97.49%	0.202	97.51%
4	0.206	95.80%	0.205	95.91%
5	0.207	95.07%	0.207	95.24%
8	0.212	93.13%	0.211	93.51%
10	0.214	91.99%	0.213	92.51%
20	0.226	87.30%	0.222	88.57%
25	0.231	85.35%	0.227	86.98%
40	0.245	80.40%	0.237	82.99%
50	0.254	77.60%	0.244	80.76%

Table 1: Required service rates in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

Figure 4 shows the curves of the required percentage of service rate increase, calculated as $100 \times (\mu_n - \mu)/\mu$, according to the number of pools n in order to reach $W_n = 0.18$ min and $W_n(20sec) = 80\%$, respectively. Figure 4 shows that, for both types of QoS, the required increase of service rate does not grow in a dramatic fashion. We notice that for a Dedicated System with $n = 20$ separate teams, the required mean service time is about 4 min and 25 sec in order to reach $W_n = 0.18$ min, and it is about 4 min and 30 sec in order to reach

$W_n(20sec) = 80\%$. In a Dedicated System with $n = 10$ separate call centers, the required mean service time is only about 4 min and 40 sec for both types of QoS. All these values are not too far from the actual mean service time (5 min). In conclusion, it is possible to even up the performances of a Pooled System by slightly increasing the service rate. In practice, an increase in service rate efficiency in the order of 10% seems very reasonable to achieve because of the competition created by the team-based organization.

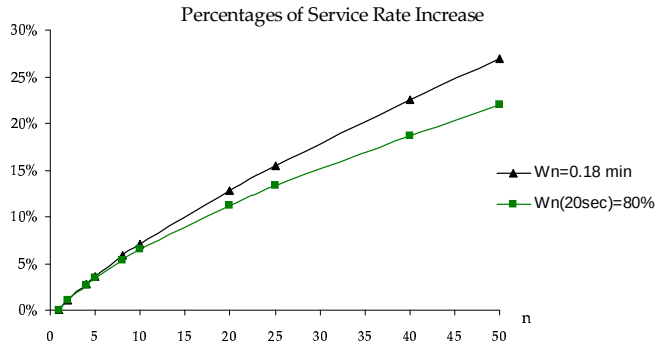

Figure 4: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

4.3 Evaluation of Percentage of Call Back proportion Decrease

Now, we focus on evaluating the required decrease of the call back proportion in a Dedicated System with n separated call centers, in order to ensure the same QoS ($W_n(t) = W_n(t)$ or $W_n = W$) as in the corresponding pooled configuration.

We consider again the same example of the Pooled System (with the same parameters s, λ^a, α and μ) as in the previous subsection. Each one of the n separated call centers of the Dedicated System has an arrival rate of first-attempt calls of $\lambda_n^a = \lambda^a/n$, a service rate $\mu_n = \mu = 0.2$ calls per min, and a staffing level $s_n = s/n$. For each n, we calculate the required call back proportion α_n, so that, the average waiting time is $W_n = 0.18$ min. We repeat the same analysis for the QoS in terms of the 80/20 rule. The corresponding results are presented in Table 2. We choose to vary n only from 1 to 10, so that, α_n stays positive.

In Figure 5, we plot the required percentages of the call back proportion decrease, calculated as $100 \times (\alpha - \alpha_n)/\alpha$, versus the number of pools n in order to reach $W_n = 0.18$ min and $W_n(20sec) = 80\%$, respectively. Once again, we see from Figure 5 that the required percentage

	$W_n = 0.18$		$W_n(20sec) = 80\%$	
n	α_n	ρ_n	α_n	ρ_n
1	10.00%	98.53%	10.00%	98.53%
2	9.02%	97.47%	9.04%	97.49%
4	7.38%	95.74%	7.50%	95.87%
5	6.64%	94.98%	6.83%	95.18%
8	4.61%	92.96%	5.06%	93.40%
10	3.36%	91.76%	4.00%	92.38%

Table 2: Required call back proportions in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

decrease of the call back proportion grows with the number of pools n in a not so strong way, and that the curves for each type of QoS are similar. For example, in a Dedicated System with 10 pools, we have to decrease the call back proportion α_n by about 60% with regard to $\alpha = 10\%$ in order to reach $W_n(20sec) = 80\%$. Note that it is possible to achieve this required decrease in practice, especially when the quality of response within the pooled configuration is quite poor. The reason of the improvement is that the agents in the team-based organization are more responsible for their own customers than in the case of the pooled organization. Agents will try to provide answers that are as good as possible, in order to diminish the call back flow, and as a consequence, improve the performance of their team.

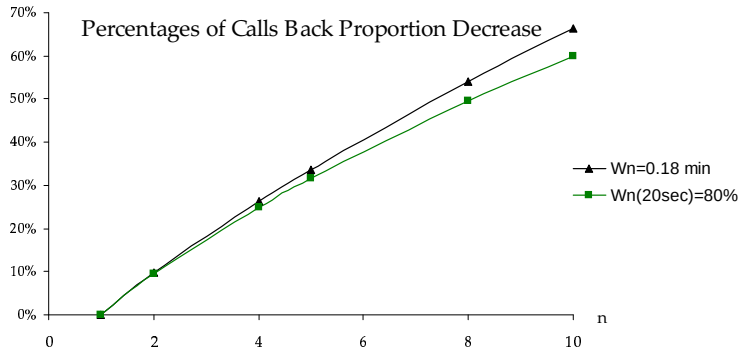

Figure 5: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

4.4 Synthesis

The results of the previous sections have shown that migrating towards separated call centers may not be as bad an idea as it seems. In addition to the analysis reported above, we have

performed a more systematic analysis to confirm the robustness of our conclusions. This analysis is reported in Appendix A. It shows that the qualitative results discussed above are valid for a large range of parameters typical of those that would be encountered in real situations.

In addition, it would be realistic to assume that the better team management enabled by the new organization implies an improvement of both parameters, i.e., an increase of the service rate and a decrease of the call back proportion. To see the combined improvement of the two factors, we perform the same analysis as that of Section 4.2 by also incorporating an improvement on the call back proportion. We consider two cases corresponding to two values of α: $\alpha = 8\%$ and $\alpha = 5\%$, corresponding to a 20% and 50% improvement in the call back proportion with respect to the initial value of $\alpha = 10\%$, respectively. The results are provided in Tables 3 and 4 for the range of values of n of interest, i.e., n from 20 to 50. The results show that by having an improvement on both efficiencies, the required performance improvement on each one is not as high as when focusing on each one separately.

	$W_n = 0.18$			$W_n(20sec) = 80\%$		
n	μ_n	% of improvement	ρ_n	μ_n	% of improvement	ρ_n
20	0.221	10.53%	87.20%	0.218	8.90%	88.52%
25	0.226	13.08%	85.24%	0.222	10.90%	86.92%
40	0.240	20.09%	80.26%	0.232	16.23%	82.93%
50	0.249	24.45%	77.45%	0.239	19.44%	80.70%

Table 3: Assuming an improvement of the call back proportion by 20% ($\alpha = 8\%$)

	$W_n = 0.18$			$W_n(20sec) = 80\%$		
n	μ_n	% of improvement	ρ_n	μ_n	% of improvement	ρ_n
20	0.214	7.22%	87.06%	0.211	5.55%	88.44%
25	0.219	9.71%	85.08%	0.215	7.50%	86.83%
40	0.233	16.58%	80.07%	0.225	12.69%	82.84%
50	0.242	20.85%	77.24%	0.232	15.81%	80.60%

Table 4: Assuming an improvement of the call back proportion by 50% ($\alpha = 5\%$)

Let us now focus on the mix of efficiency improvements for a given value of n. Table 5 presents the results pertaining to the case $n = 10$. When we migrate to a Dedicated System with $n = 10$ separated call centers, we need either to increase the service rate μ_n by about 7.11% with regard to $\mu = 0.2$ call per min, or to decrease the call back proportion α_n by about 66% with regard to the initial $\alpha = 10\%$ in order to achieve $W_n = 0.18$ min. Another solution is

to increase μ_n by 3% and decrease α_n by about 37% at the same time. In such a case, it should come as no surprise that we improve the performances in the dedicated systems rather than deteriorate them. Team management effects may change both parameters and may go beyond the simple fact of outweighing the increase of variability.

μ_n Percentage Increase	α_n Percentage Decrease
0.00%	66.43%
1.00%	56.51%
2.00%	46.79%
3.00%	37.26%
4.00%	27.92%
5.00%	18.76%
6.00%	9.78%
7.00%	0.97%
7.11%	0.00%

Table 5: Percentages of call back proportion decrease according to percentages of service rate increase in a Dedicated System with $n = 10$ in order to achieve $W_n = 0.18$ min

Figure 6 shows the variation of the percentage decrease of α_n according to the percentage increase of μ_n. The graph suggests that improving α_n is linear according to improving μ_n. However it is not the case in general. In fact, let us take the particular case of a Dedicated System with a collection of n separated $M/M/1$ queues ($s_n = 1$). In this case, the average waiting time is given by

$$W_n = \frac{\rho_n}{\mu_n(1-\rho_n)}, \tag{4}$$

where ρ_n is the server utilization, $\rho_n = \lambda_n/\mu_n$. Since $\lambda_n = \lambda_n^a/(1-\alpha_n)$, we deduce from Equation (4) that:

$$\alpha_n = 1 - \frac{\mu_n W_n + 1}{\mu_n^2 W_n}\,\lambda_n^a. \tag{5}$$

If W_n and λ_n^a are held constant, Equation (5) shows that α_n is not linear according to μ_n. However, we obtain an almost linear behavior when the number of pools n is not very large and therefore the number of servers per pool s_n is not very small, which is the case in our call center. It is an interesting result, since, if we can assume linearity between these two parameters, we are able to approximate them through simple formulas.

Another advantage of the team-based organization is its robustness with respect to errors in the estimation of the arrival rate of primary calls. For example, consider again the Pooled

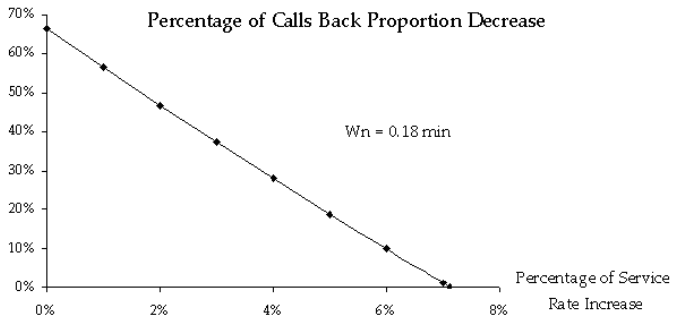

Figure 6: Percentages of call back proportion decrease according to percentages of service rate increase in a Dedicated System with $n = 10$ in order to achieve $W_n = 0.18$ min

System, $n = 1$, described in Section 4.2 and the Dedicated System, $n = 10$, obtained by increasing the service rate in order to ensure the same QoS as in the Pooled System. We denote by $\lambda^{a,real}$ the real first-attempt arrival rate. Figure 7 plots the average waiting time W_n versus $\lambda^{a,real}$ for the Pooled System, $n = 1$, and for the Dedicated System, $n = 10$. We observe that the QoS of the Pooled System is much more affected than the one of the Dedicated System by an underestimation of the first-attempt calls arrival rate. Let us give an explanation. Under the original expected first-attempt arrival rate, the server utilization in the Pooled System, 98.53%, is much closer to 1 than the one in the Dedicated System, 91.99%. If the first-attempt calls arrival rate is underestimated, the deterioration of the quality of service is increasing faster when the server utilization is closer to 1, since the queue becomes less and less stable. For example, assume that we underestimate the total arrival rate of first-attempt calls (which is now $\lambda^a = 177.36$ calls per min for both systems) by only 1.41%. Then the real server utilization of the Pooled System becomes 99.92% and the one of the Dedicated System becomes 93.28%. As a consequence, the average waiting time of the first system goes beyond 5 min and the one of the second system is only 0.27 min. This shows that the team-based organization is more robust than the original pooled organization. This is a very attractive feature that gives another strong argument in favor of the team-based organization.

5 Call Center Models with Out-Portfolio Flow

In this section, we address the same issues as in Section 4 by considering two new models (a pooled model and a dedicated model) of call centers. They differ from the above models by

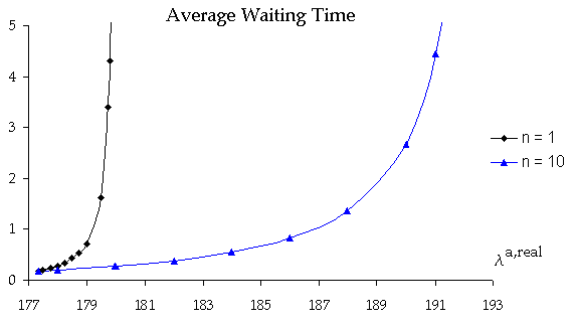

Figure 7: Average waiting time of a Pooled System ($n = 1$) and a Dedicated System ($n = 10$) according to the total arrival rate of first-attempt calls

taking an anonymous flow (out-portfolio flow) of calls into account. The latter consists of calls for which one cannot associate a portfolio when they enter the call center. An anonymous call can be a call of a customer of *Bouygues Telecom* who does not communicate his phone number to the Computer-Telephone Integration (CTI), a person who is not a customer of *Bouygues Telecom*, etc. In Section 5.1, we present the models and we develop approximations to estimate the performance measures of interest. In Sections 5.2 and 5.3, we analyze the required improvement in terms of quantitative or qualitative efficiency that must be achieved in order to outweigh the loss in economy of scale. Finally, in Section 5.4, we further investigate the consequence of having an out-portfolio flow on the behavior of the Dedicated System.

5.1 Modeling and Performance Analysis

Consider first the queueing model of the original large call center with two types of customers: identified customers (portfolio or PTF customers) and non-identified (anonymous) customers (out-portfolio or OPTF customers). PTF customers have priority over OPTF customers in the sense that agents are providing assistance to PTF customers first. The priority rule is non-preemptive, which simply means that an agent currently serving an OPTF customer while a PTF customer joins the waiting queue will complete this service before turning to the PTF customer. The model consists of two infinite queues and a set of s identical servers representing the set of agents. All agents are able to answer all types of customers. Each type of customer has its own queue. Service times are assumed to be exponentially distributed and independent of each other with rate μ for both types of customers. The arrival process of first-attempt calls

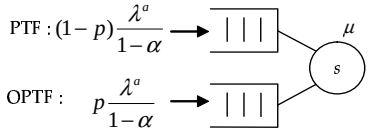

Figure 8: Portfolio Pooled System

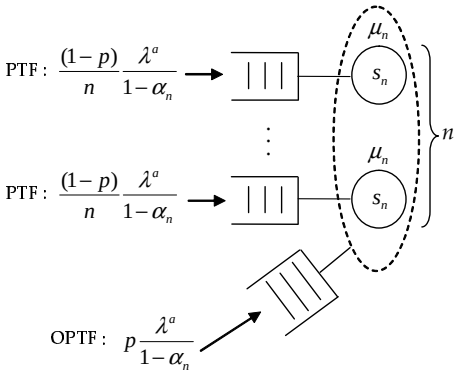

Figure 9: Portfolio Dedicated System

(primary calls) is assumed to be Poisson with a total arrival rate of λ^a. The proportion of OPTF first-attempt calls is p. So, the total arrival rate of first-attempt PTF calls is $\lambda^{a,PTF} = (1-p)\lambda^a$, and that of the OPTF calls is $\lambda^{a,OPTF} = p\lambda^a$. There is a probability α that the customer is not satisfied with the answer he got and therefore will call again. We assume that α is the same for both types of customers. We make the same detailed assumptions as those presented in Section 4.1. Following similar arguments, the behavior of this call center can be approximated by a simple $M/M/C$ queue with two classes of customers (PTF and OPTF), $C = s$ servers, a Poisson arrival rate of PTF customers $\lambda^{PTF} = (1-p)\lambda^a/(1-\alpha)$, a Poisson arrival rate of OPTF customers $\lambda^{OPTF} = p\lambda^a/(1-\alpha)$ and a service rate μ. PTF customers have non-preemptive (head-of-line) priority over OPTF customers. Within each queue, the discipline is FCFS. This model, referred to as the Portfolio Pooled System, is illustrated on Figure 8. Note however that in this more general situation, this model is only an approximation of the behavior of the Portfolio Pooled System. This is due to the fact that the model does no longer belong to the class of product-form networks analyzed by Baskett et al. (1975), because of the priority of the PTF customers over the OPTF ones. Note also that the Portfolio Pooled System reduces to the Pooled System studied in Section 4 when $p = 0\%$.

Consider now the modeling of the team-based organization with a partitioning of the original call center (with out-portfolio customers) into n autonomous teams, each one being associated with a customer portfolio. It is assumed that n is such that s is a multiple of n. All teams and all customer portfolios are statistically identical. Each team has s_n identical servers, and has its own infinite queue for its own PTF customers. There is a single infinite queue for all

OPTF customers. An OPTF customer is served only when at least one of the agents (of any team) is idle and no PTF customers are waiting in the corresponding portfolio queue. The assumptions are similar to those above. The arrival process of PTF first-attempt calls to each PTF queue is Poisson with an arrival rate of $\lambda_n^{a,PTF} = (1-p)\lambda^a/n$. The arrival process of OPTF first-attempt calls to the OPTF queue is Poisson with an arrival rate $\lambda^{a,OPTF} = p\lambda^a$.

The behavior of this call center can be approximated by a set of n identical parallel $M/M/C$ systems with $C = s_n$ servers. Each $M/M/C$ system has its own arrival process corresponding to its PTF customers. This arrival process is Poisson with a rate $\lambda_n^{PTF} = \lambda_n^{a,PTF}/(1-\alpha_n) = ((1-p)/n)(\lambda^a/(1-\alpha_n))$. In addition, there is an additional queue for the OPTF customers. The arrival process to this queue is Poisson with a rate $\lambda^{OPTF} = \lambda^{a,OPTF}/(1-\alpha_n) = p(\lambda^a/(1-\alpha_n))$. The OPTF customers can be served by any server of any of the parallel $M/M/C$ queues. However, PTF customers have non-preemptive (head-of-line) priority over OPTF customers. The resulting model is shown on Figure 9. This model will be referred to as the Portfolio Dedicated System. Note that the Portfolio Dedicated System reduces to the Portfolio Pooled System when $n = 1$, and to the Dedicated System when $p = 0\%$.

We now focus on evaluating the stationary performances of the two above models. Exact performance measures of the Portfolio Pooled System (viewed as a non-preemptive priority $M/M/C$ queue) in terms of the waiting time distribution can be found in Kella and Yechiali (1985). However, the exact quantitative analysis of the Portfolio Dedicated System is complicated. For that, we developed a set of models which allow us to approximate its performance measures.

In the following, we present two approximations for the Portfolio Dedicated System. The first enables us to calculate a pessimistic estimate (upper bound) of the waiting times of the PTF customers, while the second one enables us to calculate a pessimistic estimate (upper bound) of the waiting times of the OPTF customers. The reason for choosing pessimistic estimates is such that the improvements in efficiency that will follow from our analysis can be viewed as a lower bound on the improvements in efficiency that will actually be required in the exact analysis. These approximate models are of the same nature as that of the Portfolio Pooled Model. Their performance measures can then be exactly calculated in the same way. Validations of these approximations are presented in Appendix B.

Pessimistic Model for PTF customers The pessimistic model for PTF customers is obtained from the Portfolio Dedicated model by splitting the flow of OPTF into a set of n independent flows. The resulting model consists of a set of n independent and identical $M/M/C$ systems with $C = s_n$ servers and a service rate μ_n. As in the original model, the arrival process of PTF customers is a Poisson process with rate $\lambda_n^{PTF} = ((1-p)/n)(\lambda^a/(1-\alpha_n))$. The arrival process of OPTF customers is a Poisson process with rate $\lambda_n^{PTF} = (p/n)(\lambda^a/(1-\alpha_n))$. The OPTF customers can be served by any of the s_n servers from the corresponding team. However, PTF customers have non-preemptive (head-of-line) priority over OPTF customers. The model is shown on Figure 10.

In this model, there is not a single OPTF queue as opposed to the Portfolio Dedicated System. The OPTF flow is equally divided and assigned to each one of the separate OPTF queues. Thus, it may happen that an OPTF customer is delayed to access a server, while a server of another team is available. This delay may later on delay the access of a PTF customer to a server because the OPTF customer will now be served and the priority is non-preemptive. Another way to look at this approximate model is to see it as an unpooling of the OPTF flow, which indirectly causes additional delays to the PTF customers. Thus this model provides a pessimistic estimate of the waiting times of the PTF customers.

Pessimistic Model for OPTF customers The pessimistic model for OPTF customers is obtained from the Portfolio Dedicated System by merging the flow of PTF into a single flow and simultaneously merging all the servers into a single pool of ns_n servers. This model is similar to the Portfolio Pooled System. The arrival process to the single PTF queue is Poisson with rate $\lambda_n^{PTF} = (1-p)(\lambda^a/(1-\alpha_n))$. The arrival process to the single OPTF queue is a Poisson process with rate $\lambda_n^{PTF} = p(\lambda^a/(1-\alpha_n))$. PTF customers have non-preemptive priorities over OPTF customers. The model is shown on Figure 11.

Because of the pooling effect of the PTF customers, the average waiting time of OPTF customers will be higher. In the Portfolio Dedicated System, it may happen that an OPTF customer gets access to a server while a PTF customer is waiting in a queue (not served by this server), which reduces the waiting time of the OPTF customer. In the pessimistic model, this will never occur due to the pooling of all PTF queues and all servers. Thus this model provides

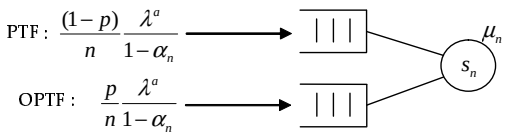

Figure 10: Pessimistic model for the PTF customers

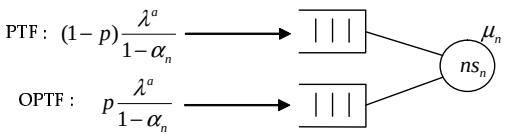

Figure 11: Pessimistic model for the OPTF customers

a pessimistic estimate of the waiting time of OPTF customers.

In the following subsections, we evaluate the impact of migrating from a Portfolio Pooled System towards a Portfolio Dedicated System. We concentrate on the evaluation of efficiency improvement (both qualitative and quantitative) required to counterbalance the reduction of the pooling effect of the team-based organization. We define a global quality of service QoS^{global} for all types of customers as $QoS^{global} = (1-p)\,QoS^{PTF} + p\,QoS^{OPTF}$. In what follows, we only focus on the QoS measured in terms of the average waiting time. Similar results could be obtained for the 80/20 rule.

5.2 Evaluation of Service Rate Percentage Increase

We start from a Portfolio Pooled System with a given quality of service W^{global}, and our purpose is to evaluate the required service rate μ_n in a Portfolio Dedicated System with n identical teams in order to ensure the same global average waiting time $W_n^{global} = W^{global}$. The total staffing level, the total arrival rate of first-attempt calls, and the call back proportion are all held constant.

In the Portfolio Pooled System, the arrival rate of first-attempt calls is $\lambda^a = 177.36$ calls per min, the call back proportion is $\alpha = 10\%$, the service rate is $\mu = 0.2$ calls per min, and the number of servers is $s = 1000$. The server utilization is then $\rho = 98.53\%$. We choose the same parameters as in the Pooled System that we studied in Section 4. In addition, we vary the proportion of OPTF customers, $p = 0\%$, $p = 5\%$, $p = 10\%$, or $p = 20\%$. As expected, W^{global} does not depend on p. In fact, the Portfolio Pooled System is a workconserving system. It is not the case that one server is idle while a customer (PTF or OPTF) is waiting for service. If we vary p, the order of service of a given customer may change, but the overall average waiting time remains unchanged. We give the mathematical explanation in Appendix

C. Here, $W^{global} = 0.18$ min for all values of p. For each value of p, we now consider the corresponding Portfolio Dedicated System. We vary n from 1 to 50. For each n, we evaluate the required service rate μ_n (using pessimistic models), so that, the global average waiting time is $W_n^{global} = W^{global} = 0.18$ min. We present the results in Table 6.

	$p = 0\%$		$p = 5\%$		$p = 10\%$		$p = 20\%$	
n	μ_n	ρ_n	μ_n	ρ_n	μ_n	ρ_n	μ_n	ρ_n
1	0.2	98.53%	0.2	98.53%	0.2	98.53%	0.2	98.53%
2	0.202	97.49%	0.201	98.23%	0.2	98.36%	0.2	98.39%
4	0.206	95.80%	0.202	97.38%	0.201	97.95%	0.201	98.23%
5	0.207	95.07%	0.203	96.84%	0.202	97.67%	0.201	98.13%
8	0.212	93.13%	0.207	95.08%	0.204	96.54%	0.202	97.74%
10	0.214	91.99%	0.21	93.92%	0.206	95.60%	0.202	97.39%
20	0.226	87.30%	0.221	89.06%	0.217	90.80%	0.209	94.16%
25	0.231	85.35%	0.226	87.05%	0.222	88.74%	0.214	92.13%
40	0.245	80.40%	0.24	81.96%	0.236	83.54%	0.227	86.81%
50	0.254	77.60%	0.249	79.09%	0.244	80.62%	0.235	83.81%

Table 6: Required service rate increase in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

Moreover, we calculate by simulation the exact values of μ_{20} for $p = 5\%$, 10% and 20%, in order to get some indications of their deviations regarding the values given by the proposed models. Recall that for $p = 0\%$, the Portfolio Dedicated System behaves like separate Erlang-C models. So, the corresponding μ_{20} is obtained by an exact numerical result. For $p = 5\%$, the required service rate given by simulation is $\mu_{20} = 0.217$ instead of 0.221 given by our approximation models, for $p = 10\%$ it is 0.210 instead of 0.217, and for $p = 20\%$ it is 0.202 instead of 0.209. This shows us that the costs of partitioning given by our models are not too far from those given by simulation, at least for the reasonable parameters of the Portfolio Dedicated System we have chosen. In Figure 12, we plot for each value of p, the curve of the percentage of required service rate increase, calculated as $100 \times (\mu_n - \mu)/\mu$, in the Portfolio Dedicated System according to the number of pools n, in order to reach a global quality of service of $W_n^{global} = 0.18$ min. We notice that for a given p, we have the same qualitative results as in Section 4.2. The required increase of the service rate is not very important and it is feasible to reach in practice. This is due to the competition element of the team-based organization.

The new interesting insight here is that the necessary increase of the service rate to com-

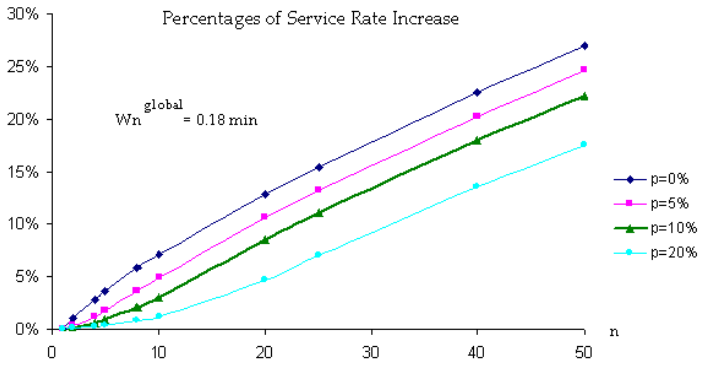

Figure 12: Percentages of required service rate increase according to number of pools n in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

pensate the loss of pooling effect decreases when the proportion of OPTF customers increases. Particulary, migrating towards a Portfolio Dedicated System (with any $p > 0\%$) is always less costly than migrating towards a Dedicated System ($p = 0\%$). For example, consider a Portfolio Dedicated System with $n = 10$. If the OPTF proportion is $p = 5\%$, we need to increase the service rate by 4.91%. However, with a proportion $p = 20\%$ we only need to increase the service rate by 1.17%. We explain this advantage by the fact that the OPTF flow is used to reduce idle periods of servers while customers are waiting in the Portfolio Dedicated System. Idle times would not exist in the case of $p = 100\%$ (Pooled System). This will be explained with more details in Section 5.4.

5.3 Evaluation of Percentage of Call Back proportion Decrease

Let us again start from the Portfolio Pooled System of Section 5.2. We aim to evaluate the call back proportion α_n for a Portfolio Dedicated System with n identical teams, in order to get $W_n^{global} = W^{global} = 0.18$ min as in the Portfolio Pooled System. We again vary p ($p = 0\%$, $p = 5\%$, $p = 10\%$, or $p = 20\%$). For each p, we vary n from 1 to 10. We choose to vary n only from 1 to 10, so that, α_n stays positive. We present the results in Table 7.

In Figure 13, we plot for each value of p, the curve of the required percentage of call back proportion decrease, calculated as $100 \times (\alpha - \alpha_n)/\alpha$, in the Portfolio Dedicated System according to n, in order to reach a global quality of service of $W_n^{global} = 0.18$ min. Again, we get the same qualitative results as in Section 4.3. In addition, we notice that the cost in terms of the required decrease of call back proportion is decreasing according to p. It is due again to the OPTF flow.

	$p = 0\%$		$p = 5\%$		$p = 10\%$		$p = 20\%$	
n	α_n	ρ_n	α_n	ρ_n	α_n	ρ_n	α_n	ρ_n
1	10.00%	98.53%	10.00%	98.53%	10.00%	98.53%	10.00%	98.53%
2	9.02%	97.47%	9.72%	98.23%	9.86%	98.37%	9.93%	98.46%
4	7.38%	95.74%	8.91%	97.35%	9.46%	97.94%	9.77%	98.28%
5	6.64%	94.98%	8.39%	96.80%	9.20%	97.66%	9.67%	98.17%
8	4.61%	92.96%	6.61%	94.95%	8.09%	96.48%	9.29%	97.76%
10	3.36%	91.76%	5.39%	93.73%	7.12%	95.47%	8.94%	97.39%

Table 7: Required call back proportion decrease in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

When p increases, the variability in the Portfolio Dedicated System decreases. For instance, consider a Portfolio Dedicated System with $n = 10$. If the OPTF proportion is $p = 5\%$, we need to decrease the call back proportion by 46.07%. However, with a proportion of $p = 20\%$ we need to decrease the call back proportion by only 10.57%.

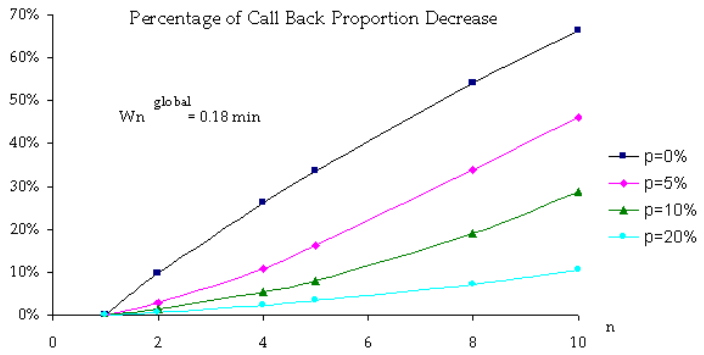

Figure 13: Percentages of call back proportion decrease according to number of pools n in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

5.4 Synthesis

In the previous sections, we showed that the reduction of pooling effect when migrating from a Pooled System to a Dedicated System could be outweighed by team management benefits. The discussion and insights presented in Section 4.4 are still valid in the more general setting of a call center with an out-portfolio flow. However, there is a new important insight. It clearly appears that having an out-portfolio flow may reduce the drawback of migrating from the Pooled System to the Dedicated System. This is due to the fact that as opposed to the PTF flows, the OPTF flow is not decomposed into several independent flows, each one being

associated with a specific team. Thus the OPTF flow maintains the benefits of the pooling effect. The out-portfolio flow can then be seen as an "idle time killer" in a Portfolio Dedicated System: an out-portfolio call is distributed only when an agent is idle and no customer of his portfolio is waiting. As out-portfolio calls have less priority, this allows reducing idle time without significantly penalizing the QoS of portfolio calls.

The Portfolio Dedicated System can thus be considered as a particular case of partial pooling. We call this configuration "partial calls pooling" because a proportion p of incoming calls (pooled calls or out-portfolio calls in the *Bouygues Telecom* case) can be served by any agent while the remaining calls $(1-p)$ are dedicated to specific agents. It clearly appears from the graphs presented in Sections 5.2 and 5.3 that this improvement in efficiency required to counterbalance the reduction of pooling effect decreases as p increases. Now, one additional very attractive feature is that this decrease is not linear in p. Let us, for instance, consider the case of quantitative efficiency improvement (service rate increase) discussed in Section 5.2 (a similar analysis could be done for the qualitative efficiency improvement discussed in Section 5.3).

In order to illustrate this behavior, let us again consider the same basic example of the Portfolio Pooled System (1000 servers, $\lambda^a = 177.36$ calls per minute, $\mu = 0.2$ calls per min, and 10% of call back proportion). In Figure 14, we plot the required percentage of service rate increase in the Portfolio Dedicated System to reach the same performance as the Portfolio Pooled System, as a function of the proportion of out-portfolio flow p (p ranges from 0% to 100%). There are three graphs corresponding to three Portfolio Dedicated System configurations: $n = 10$, 20, and 40. The graphs confirm the non linear shape of the curve. This means that with a fairly small percentage of out-portfolio flow, the required efficiency improvement is much smaller than that of the system without out-portfolio flow (corresponding to the case where $p = 0\%$). In other words, a rather small out-portfolio flow significantly reduces the drawback of the unpooling effect of the PTF customers. Recall also that since we are using pessimistic approximations, the actual curves would be stiffer.

For this case, we also performed an extensive numerical study to validate that the conclusions discussed above remain valid for a large set of parameters, thereby confirming the robustness of our analysis.

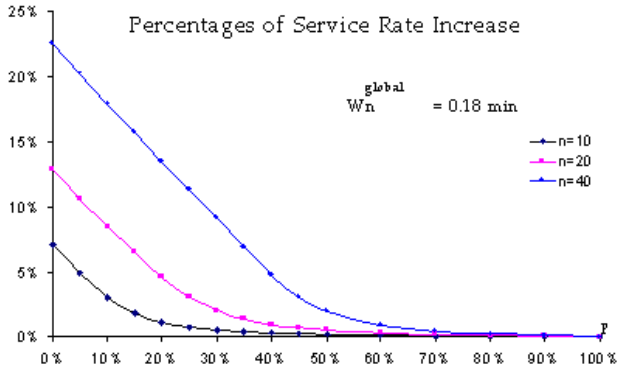

Figure 14: Percentages of required service rate increase according to OPTF proportion p in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

6 Conclusion

We focused on a fundamental problem in the design and management of stochastic service systems. We investigated the impact of team-based organizations in call centers management. Agents of call centers are the interface between the company and the customers. Thus, managers have to support and motivate their employees, so that, the assistance they provide to the customers is efficient. Partitioning agents into groups creates competition and makes agents more responsible, which motivates them to provide both rapid and improved responses.

In this paper, we argued how team management benefits, that come from the portfolio/team one-to-one link, may outweigh the economy of scale associated with the pooled organization. First, we study partitioning of a large call center into identical and separated call centers, where agents of a same team are dedicated to one portfolio of customers. Queueing models involved in this part of the study are simple. They give us important insights and help us understand the behavior of more complicated systems. We show that the costs of migrating towards a Dedicated System are not as important as it may appear. In practice, combining the benefits of the team-based organization in terms of both improved service rate efficiency and reduced call back proportion can easily outweigh the loss of the economy of scale. We also present further insights, such as robustness of the Dedicated System regarding errors in the estimation of the arrival rate.

In the second part of the study, we extend our analysis to the more general situation with an additional out-portfolio flow. We develop a set of models that give us lower bounds of

performance measures. We verify the same qualitative results as in the first part. In addition, we present an interesting insight, that is, a small proportion of out-portfolio calls may be sufficient to approximately attain the same performances as in the Pooled System. This property fits into a general idea in queueing theory. It is like saying that with a small amount of flexibility, an SBR call center may yield most of the benefits of full-flexibility (Chevalier et al. (2004)).

The application of customer portfolio management had very significant effects in the *Bouygues Telecom* call center. The quality of answers has been improved reducing callbacks by 25%. The proportion of disconnected calls (because of a full queue) was divided by 2 (in our paper we assumed an infinite queue for simplicity). And no supplementary agents were hired in spite of the increase of the total number of customers by 15%. This provides an experimental confirmation of the results and insights presented in this paper.

In a future study, we will extend our models by considering abandonments and limited waiting lines and more general service time distributions. We will also try to improve the approximation models discussed here to get more accurate analyzes. Finally, a more ambitious extension would be to investigate the introduction of team-based organization in an SBR call center where agents have specific skills.

Acknowledgements

This research was supported by *Bouygues Telecom*. The authors would like to express their gratitude to two anonymous referees for their several useful suggestions to improve this paper, as well as John Buzacott and Zeynep Aksin Karaesmen for their helpful discussions.

References

Akşin, O. Z. and Karaesmen, F. (2002). Designing Flexibility: Characterizing the value of Cross-Training Practices. Working paper, INSEAD, Fontainebleau, France.

Baskett, F., Chandy, K., Muntz, R., and Palacios-Gomez, F. (1975). Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22:248–260.

Benjaafar, S. (1995). Performance Bounds for the Effectiveness of Pooling in Multi-Processing Systems. *European Journal of Operational Research*, 87:375–388.

Borst, S., Mandelbaum, A., and Reiman, M. (2004). Dimensioning Large Call Centers. *Opera-*

tions Research, 52:17–34.

Boudreau, J. (2004). Organizational Behavior, Strategy, Performance, and Design in Management Science. *Management Science*, 50:1463–1476.

Boudreau, J., Hopp, W., McClain, J., and Thomas, L. (2003). On the Interface between Operations and Human Ressources Management. *Manufacturing & Service Operations Management*, 5:179–202.

Chevalier, P., Shumsky, R., and Tabordon, N. (2004). Routing and Staffing in Large Call Centers with Specialized and Fully Flexible Servers. Université catholique de Louvain, University of Rochester and Belgacom Mobile/Proximus. Working paper.

de Véricourt, F. and Zhou, Y.-P. (2005). Managing Response Time in a Call Routing Problem with Service Failure. *Operations Research*, 53:968–981.

Fischer, M., Garbin, D., Gharakhanian, A., and Masi, D. (1999). Traffic Engineering of Distributed Call Centers: Not as Straight Forward as it May Seem. Mitretek Systems.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5:73–141.

Garnett, O. and Mandelbaum, A. (2001). An Introduction to Skills-Based Routing and its Operational Complexities. Teaching notes, Technion.

Gross, D. and Harris, C. (1998). *Fundamentals of Queueing Theory*. Wiley series in probability and mathematical statistics. 3rd edition.

Kella, O. and Yechiali, U. (1985). Waiting Times in the Non-Preemptive Priority M/M/c Queue. *Stochastic Models*, 1:257–262.

Kleinrock, L. (1976). *Queueing Systems, Computer Applications*, volume II. A Wiley-Interscience Publication.

Larson, C. (1987). Perspectives on Queues: Social Justice and the Psychology of Queueing. *Operations Research*, 35:895–905.

Mandelbaum, A. and Reiman, M. (1998). On Pooling in Queueing Networks. *Management Science*, 44:971–981.

Rothkopf, M. and Rech, P. (1987). Perspectives on Queues: Combining Queues is not Always Beneficial. *Operations Research*, 35:906–909.

Schonberger, R. (1986). *World Class Manufacturing: The Lessons of Simplicity Applied*. Free Press, New York. 10-11.

Smith, D. and Whitt, W. (1981). Resource Sharing for Efficiency in Traffic Systems. *The Bell System Technical Journal*, 60:39–55.

Tekin, E., Hopp, W., and varOyen, M. (2004). Pooling Strategies for Call Center Agent Cross-Training. Submitted for publication.

van Dijk, N. and van der Sluis, E. (2006). Check-in Computation and Optimization by Simulation and IP in Combination. *European Journal of Operational Research*, 171:1152–1168.

Whitt, W. (1999). Partitioning Customers into Service Groups. *Management Science*, 45:1579–1592.

Whitt, W. (2002). Stochastic Models for the Design and Management of Customer Contact Centers: Some Research Directions. Working paper, Columbia University.

Appendix A: Extension of the quantitative analysis

The numerical study in Section 4 was based on a set of basic data for the initial Pooled System: $\mu = 0.2$, $\alpha = 10\%$, $W(20sec) = 80\%$, and $s = 1000$. These basic data are representative of typical parameters encountered in the *Bouygues Telecom* call center. However, to make sure that the conclusions drawn from this set of data are robust, we have performed a large set of experiments, some of which are reported in this appendix. The study is divided into four steps. In each step, we first vary one parameter (μ, α, $W(20sec)$ or s), then we deduce λ^a and λ to get different initial pooled systems which cover many realistic call center cases. Next, we consider each case and we compute the required increase in the service rate or decrease in the call back proportion, in order to reach the same performance as in the fully pooled system, for different numbers of separated teams in the corresponding dedicated systems.

Varying the Service Rate μ

We consider four pooled systems: $s = 1000$, $\alpha = 10\%$, $W(20sec) = 80\%$, and $\mu = 0.1$, 0.2, 0.5, and 1, respectively. Then, $\lambda^a = 88.14$, 177.35, 446.52, and 896.12, and $\lambda = 97.93$, 197.06, 496.13, and 995.69, respectively. In Figure 15, we plot the curves of the required service rate increase versus the number of pools in the dedicated systems. In Figure 16, we plot the curves of the required call back proportion decrease versus the number of pools. We vary n only from 1 to 10, so that, α_n stays positive.

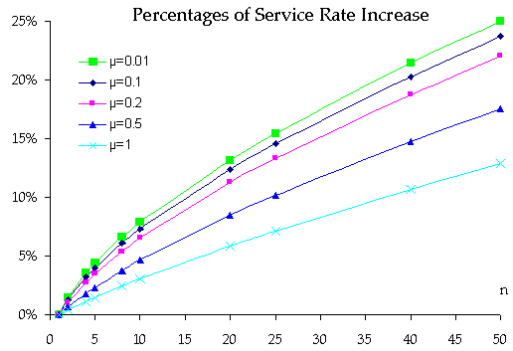

Figure 15: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial service rates

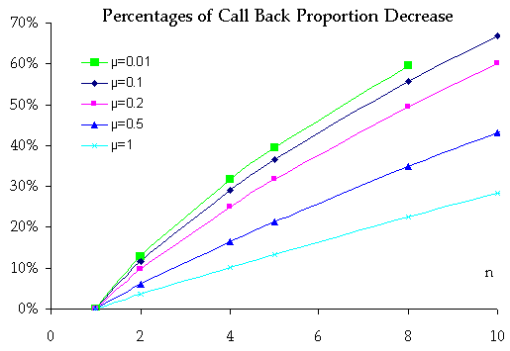

Figure 16: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial service rates

For every value of μ, the results are pretty much of the same quality as in Section 4. Furthermore, an additional insight is that the costs of migrating to the team-based organization ($(\mu_n - \mu)/\mu$ or $(\alpha - \alpha_n)/\alpha$) are decreasing as the initial service rate is increasing. One intuitive explanation is as follows. Consider two pooled systems. The parameters of the first are s, $\lambda^{(1)}$, $\mu^{(1)}$, and $W(t)$. The parameters of the second are s, $\lambda^{(2)}$, $\mu^{(2)}$, and $W(t)$. We assume that $\mu^{(1)} < \mu^{(2)}$, then $\lambda^{(1)}$ must be less than $\lambda^{(2)}$ in order to match the same performance $W(t)$ in the two systems. Besides, since the servers are slower in the first system, the server utilization of the last is less than the one in the second system, else $W(t)$ will be higher in the second system. Hence, the second system has more pooling effect than the first one. Now, let us divide each system to n identical unpooled systems, so that, s is a multiple of n. The parameters of one of the first unpooled models are s/n, $\lambda^{(1)}/n$, and the service rate is $\mu_n^{(1)}$ such that $W_n(t) = W(t)$. The parameters of one of the second unpooled models are s/n, $\lambda^{(2)}/n$, and $\mu_n^{(2)}$ such that $W_n(t) = W(t)$. Thanks to the pooling effect that is more present in the second pooled system than in the first one, the second unpooled system will need an increase in the service rate regarding μ^2 being less than the one regarding $\mu^{(1)}$ in the first unpooled system, $(\mu_n^{(2)} - \mu^{(2)})/\mu^2 < (\mu_n^{(1)} - \mu^{(1)})/\mu^{(1)}$. An additional insight is that it appears that when μ decreases, the set of curves (for different values of μ) converges towards an asymptotic curve. Indeed, we have checked that the curves for $\mu = 0.001$ almost coincide with those for $\mu = 0.01$ in Figures 15 and 16.

Varying the Call Back Proportion α

Here, we vary the call back proportion with regard to the Pooled System of Section 4. We consider four pooled systems: $s = 1000$, $\mu = 0.2$, $W(20sec) = 80\%$, and $\alpha = 5\%$, 10%, 15%, and 20%, respectively. Then, $\lambda^a = 187.21$, 177.36, 167.5, and 157.65, respectively, and $\lambda = 197.06$, for the four systems. Figures 17 and 18 show, respectively, the required quantitative (service times) and qualitative (rate of calls successfully addressed) improvements according to the number of pools.

Again, we have the same qualitative results as in Section 4. In Figure 17, the curves are identical. The explanation is as follows. Let us consider two pooled systems with the same parameters except for the dissatisfaction probability α. The required total arrival rates, to meet

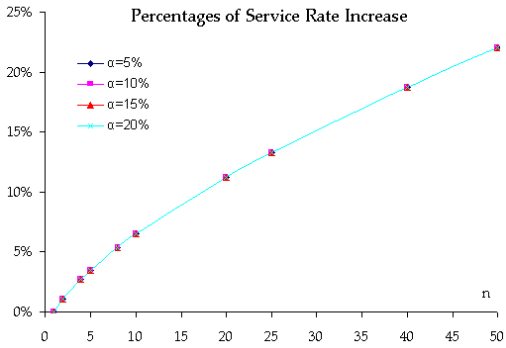

Figure 17: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial call back proportions

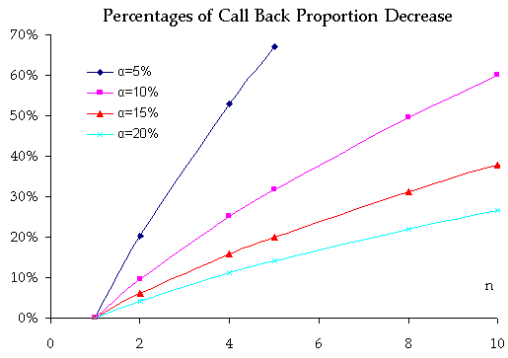

Figure 18: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial call back proportions

a given QoS, are identical in the two systems because they do not depend on α. Therefore, the two systems are equivalent to the same Erlang-C model. The required service rate μ_n and increase in the service rate $(\mu_n - \mu)/\mu$, for the corresponding dedicated systems, do not change for a fixed number of pools n.

Figure 18 shows that the required improvement in the dissatisfaction probability $(\alpha - \alpha_n)/\alpha$ is decreasing with the initial call back proportion α. The proof of this result is as follows. Consider two pooled systems with the same parameters s, λ, μ, and $W(t)$. The arrival rate of first-attempt calls and the call back proportion for the first system are $\lambda^{a,1}$ and $\alpha^{(1)}$, respectively. The ones for the second system are $\lambda^{a,2}$ and $\alpha^{(2)}$, respectively. We assume that $\alpha^{(1)} < \alpha^{(2)}$. Now, let us divide each pooled system to n identical unpooled systems, while leaving unchanged the total number of servers s, the service rate μ, and the quality of service $W_n(t) = W(t)$. So, the total arrival rate, the number of servers, and the service rate for each type of unpooled system are λ/n, s/n, and μ, respectively. The arrival rate of first-attempt calls and the call back proportion for the first unpooled systems are $\lambda_n^{a,1} = \lambda^{a,1}/n$ and $\alpha_n^{(1)}$, respectively. The ones for the second unpooled systems are $\lambda_n^{a,2} = \lambda^{a,2}/n$ and $\alpha_n^{(2)}$, respectively. Clearly, we have $\alpha_n^{(1)} < \alpha^{(1)}$ and $\alpha_n^{(2)} < \alpha^{(2)}$ because of the loss of the pooling effect. From the pooled systems, we deduce that $\lambda = \lambda^{a,1}/(1-\alpha^{(1)}) = \lambda^{a,2}/(1-\alpha^{(2)})$, and from the unpooled systems, we deduce

that $\lambda/n = \lambda_n^{a,1}/(1-\alpha_n^{(1)}) = \lambda_n^{a,2}/(1-\alpha_n^{(2)})$. The two last relations give Equation (6) below.

$$\frac{1-\alpha_n^{(1)}}{1-\alpha^{(1)}} = \frac{1-\alpha_n^{(2)}}{1-\alpha^{(2)}}. \tag{6}$$

Since $\alpha^{(1)} < \alpha^{(2)}$, then $\lambda^{a,1} > \lambda^{a,2}$, and equivalently $\lambda_n^{a,1} > \lambda_n^{a,2}$, so $\alpha_n^{(1)} < \alpha_n^{(2)}$. Moreover, $\alpha_n^{(1)} < \alpha^{(1)}$, we deduce then from Equation (6) that $\alpha_n^{(1)}/\alpha^{(1)} < \alpha_n^{(2)}/\alpha^{(2)}$. Hence, $1-(\alpha_n^{(1)}/\alpha^{(1)}) > 1-(\alpha_n^{(2)}/\alpha^{(2)})$, and finally $(\alpha^{(1)}-\alpha_n^{(1)})/\alpha^{(1)} > (\alpha^{(2)}-\alpha_n^{(2)})/\alpha^{(2)}$.

Varying the Quality of Service $W(20sec)$

Now, we vary the quality of service $W(20sec)$ with regard to the Pooled System of Section 4. We consider five pooled systems: $s = 1000$, $\mu = 0.2$, $\alpha = 10\%$, and $W(20sec) = 60\%$, 80%, 90%, 95%, and 99%, respectively. Then, $\lambda^a = 178.47$, 177.36, 176.27, 175.22, and 172.90, and $\lambda = 198.30$, 197.06, 195.86, 194.69, and 192.11, respectively. Figures 19 and 20 show, respectively, the required quantitative and qualitative improvements according to the number of pools.

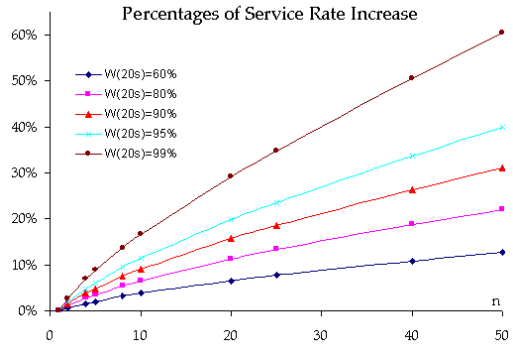

Figure 19: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve the same $W_n(20sec)$ as in the Pooled System, for a different values of $W_n(20sec)$

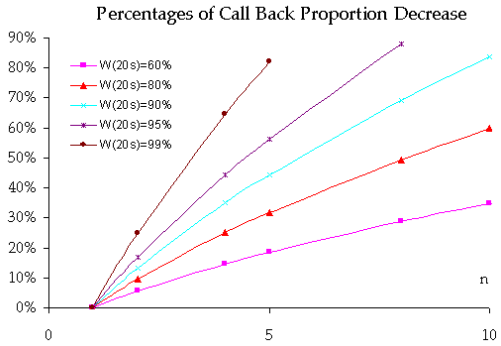

Figure 20: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve the same $W_n(20sec)$ as in the Pooled System, for a different values of $W_n(20sec)$

Once again, we underline the qualitative similarity of the results as in Section 4. The additional insight here is that the costs of partitioning the big call center $((\mu_n-\mu)/\mu$ or $(\alpha-\alpha_n)/\alpha)$ are increasing as the chosen quality of service is increasing. For instance, let us partition two pooled systems into n identical unpooled systems. The two pooled systems have

the same number of servers and the same service rate. However, the first pooled system has a quality of service lower than the one in the second. Both of the unpooled systems will need an increase in the service rate, because of the loss of the pooling effect. Moreover, since we have to reach a higher QoS in the second unpooled systems, then we will need for them a higher increase in the service rate. We notice again from the curves that the costs of migrating do not roughly increase with the chosen quality of service.

Varying the Number of Servers s

Up to now, all our analyzes were performed as a function of n, the number of dedicated pools. As stated in the paper, as long as the total number of servers s is fixed, the results obtained can alternatively be reinterpreted in terms of s/n. Indeed, specifying n is equivalent to specifying s/n. In this section however, we want to perform our analyzes for different values of s. In that case, it seems more consistent to compare configurations having the same number of servers in each pool. Therefore, the analyzes will be performed as a function of the number of dedicated servers in each pool, s/n, for different values of the total number of servers, s.

Consider now five pooled systems: $\mu = 0.2$, $\alpha = 10\%$, $W(20sec) = 80\%$, and $s = 100$, 200, 500, 1000, and 5000, respectively. Then, $\lambda^a = 16.63$, 34.26, 87.74, 177.36, and 896.60, and $\lambda = 18.48$, 38.07, 97.49, 197.06, and 996.22, respectively. In Figure 21, we plot the curves of the required service rate improvement, when we partition the pooled systems chosen here, according to the size of the generated teams s/n. We notice from Figure 21 that the costs, for a fixed size of pools, are increasing with the initial number of servers. One explanation may be as follows. Consider once again two pooled systems. The parameters of the first are $s^{(1)}$, $\lambda^{(1)}$, μ, and $W(t)$. Those of the second are $s^{(2)}$, $\lambda^{(2)}$, μ, and $W(t)$. We assume that $s^{(1)} < s^{(2)}$. Let us now migrate to the corresponding unpooled systems such that the size of each type of unpooled system is $s^{(p)}$, $s^{(1)} = n_1 s^{(p)}$ and $s^{(2)} = n_2 s^{(p)}$. It goes without saying that $n_1 < n_2$. The parameters of the first unpooled systems are $s^{(p)}$, $\lambda^{(1)}/n_1$, and $\mu_n^{(1)}$ such that the quality of service is $W_n(t) = W(t)$. The ones of the second unpooled systems are $s^{(p)}$, $\lambda^{(2)}/n_2$, and $\mu_n^{(2)}$ such that the quality of service is $W_n(t) = W(t)$ too. Due to the pooling effect that is more present in the second pooled system than in the first, $\lambda^{(2)}/n_2$ is larger than $\lambda^{(1)}/n_1$. Else, the first pooled system will match a quality of service that is lower than the second. To do a

summary for the unpooled systems, we have the same number of servers $s^{(p)}$, the same quality of service $W(t)$, and a larger arrival rate for the second unpooled systems. Thus, we easily deduce that the servers in the latter cases must be faster so as to match the same performance in both types of dedicated systems, $\mu_n^{(1)} < \mu_n^{(2)}$. Finally, $(\mu_n^{(1)} - \mu)/\mu < (\mu_n^{(2)} - \mu)/\mu$.

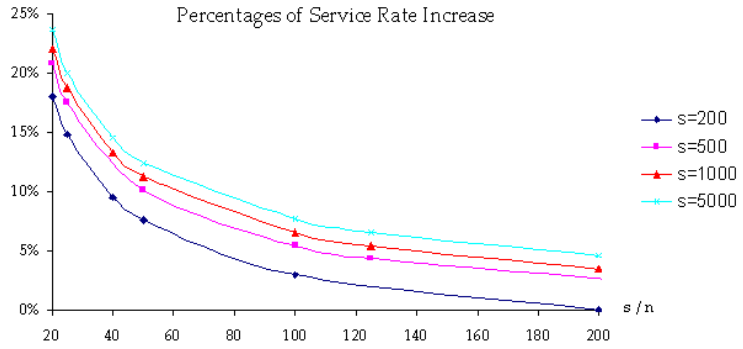

Figure 21: Percentages of service rate increase according to size of pools s/n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial number of servers

In addition, we see from Figure 21 that the gap between the curves is decreasing when s increases. Then, we can deduce that the unpooling of two pooled systems with different large number of servers, namely greater than 500, will need quite the same increase in the service rates. This is due to the fact that a "large" Pooled System does not gain too much in pooling effect by adding more servers.

Appendix B: Validation of the Approximation Models

The analysis of the Portfolio Dedicated System is to be used to design our call center; calculating staffing level, required total arrival rate (or required call back proportion), or required service rate in order to achieve a given QoS. To examine the accuracy of the approximation models, we propose two different formulations:

- QoS, $s \Rightarrow \lambda$: formulation 1 consists of calculating the required total arrival rate λ given a fixed QoS and a fixed staffing level s.
- QoS, $\lambda \Rightarrow s$: formulation 2 consists of calculating the required staffing level s given a fixed QoS and a fixed total arrival rate λ.

We compare performances given from the pessimistic models with those from simulation. We simulated 30 cases: the number of pools is $n = 10$, the OPTF customers proportion is $p = 5\%$ or 10%, and for each p, $n\,s_n = 250$, 350 or 500, and for each p and s we chose 5 values of λ^a (in order to vary the server utilization). The mean service time, and the call back proportion are kept constant ($1/\mu_n = 5$ min, and $\alpha_n = 10\%$).

Deviations between performance measures given by pessimistic models and those given by simulation are presented in Table 8. For each pessimistic model (PTF or OPTF), deviations for one parameter are calculated as $\frac{performance(model) - performance(simulation)}{performance(simulation)}$.

	Total Arrival Rate, $\lambda^a/(1-\alpha_n)$		Total Staffing Level, $n\ s_n$	
	$W_n(20sec)$	W_n	$W_n(20sec)$	W_n
PTF Pessimistic Model	-2.84%	-2.92%	4.00%	4.00%
OPTF Pessimistic Model	-5.65%	-5.61%	4.43%	4.43%

Table 8: Deviations between pessimistic models and simulation

Appendix C: Proof of the Result: W^{global} does not depend on p

First, consider an $M/M/C$ queue with a single queue. The arrival rate is λ, the number of servers s, and the service rate is μ. The service discipline is FCFS. Then the average waiting time in queue is given by the equation below.

$$W = \frac{P_D}{s\mu - \lambda}, \tag{7}$$

where P_D is the probability of delay, that is, the probability that an incoming customer waits for service. Second, consider a non-preemptive priority $M/M/C$ queue with two types of customers, say A and B. Type A customers have priority over type B ones. The total arrival rate is λ, the number of servers is s, and the service rate to handle any type of customers is μ, as in the first $M/M/C$ model. The arrival rate of type A customers is λ^A, and the one of type B customers is λ^B, $\lambda^A + \lambda^B = \lambda$. Let p be the proportion of type B customers. Then, $\lambda^A = (1-p)\lambda$. As in Kella and Yechiali (1985), the average waiting times in queue of customers A (W^A) and

customers B (W^B) are given as follows.

$$W^A = \frac{P_D}{s\mu - \lambda^A}, \text{ and } W^B = \frac{P_D}{(s\mu - \lambda^A)(1 - \frac{\lambda}{s\mu})}, \tag{8}$$

where the probability of delay P_D is identical to the one in the first model. Now, let us show that $W^{global} = W$, for any proportion p in [0,1].

$$\begin{aligned} W^{global} &= (1-p)\, W^A + p\, W^B \\ &= (1-p)\, \frac{P_D}{s\mu - \lambda^A} + p\, \frac{P_D}{(s\mu - \lambda^A)(1 - \frac{\lambda}{s\mu})} \\ &= \frac{P_D}{s\mu - (1-p)\lambda} \times \frac{s\mu - (1-p)\lambda}{s\mu - \lambda} \\ &= \frac{P_D}{s\mu - \lambda} = W, \end{aligned} \tag{9}$$

which completes the proof. □