A flexible architecture for call centers with skill-based routing

Benjamin Legros, Oualid Jouini, Yves Dallery

A Flexible Architecture for Call Centers with Skill-Based Routing

Benjamin Legros • Oualid Jouini • Yves Dallery
Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, 92290 Châtenay-Malabry, France

benjamin.legros@centraliens.net • oualid.jouini@ecp.fr • yves.dallery@ecp.fr

International Journal of Production Economics. To appear, 2014

Abstract

We focus on architectures with limited flexibility for multi-skill call centers. The context is that of call centers with asymmetric parameters: unbalanced workload, different service requirements, a predominant customer type, unbalanced abandonments and high costs of cross-training. The most well-known architectures with limited flexibility such as chaining fail against such asymmetry. We propose a new architecture referred to as single pooling with only two skills per agent and we demonstrate its efficiency. Using simulation, we conduct a comprehensive comparison between this architecture and chaining. As a function of the various system parameters, we delimit the regions where either chaining or single pooling is the best. Single pooling leads to a better performance than chaining while being less costly under various situations of asymmetry: asymmetry in the number of arrivals, in the service durations, in the variability of service times, or in the service level requirements. It is also shown that these observations are more apparent for situations with a large number of skills, or for those with a large call center size.

Keywords: Call centers; queueing models, skill-based routing; flexibility; performance measures; chaining; simulation; asymmetric parameters.

1 Introduction

Context and Motivation. The concept of flexibility is related to the ability of a company to efficiently match its capacity to an uncertain demand with multiple types. The need for flexibility arises in a wide range of manufacturing systems. It also extends to service systems, such as call centers, where different types of customers ask for a quasi-instantaneous processing. Resource flexibility in call centers reduces to cross-training agents, which allows to improve both the utilization and the performance. Since cross-training agents is achieved with higher operating costs, resource flexibility could result in a trade-off between performance and cost. The performance is measured through operational indicators such as the expected waiting time, the probability of delay, or the waiting time distribution.

We consider flexibility questions in the context of queueing models for call centers. A wide literature has focused on contrasting two extreme situations. The *full flexible* architecture (FF) versus the the *full dedicated* (FD) one. In the FF model, each agent is fully cross-trained for all call types. In most situations in which call types have similar service duration requirements, FF would require less agents than any other architecture, in order to reach a given predefined service level. The reason is that it benefits from the economies of scale, which absorb stochastic variability (Borst et al. (2004)). However the agents in FF are too costly and even sometimes impossible to find. As commented by Marengo (2004), the multilingual Compaq call center certainly could not find or train agents to speak eleven languages! In the other extreme situation of the FD model, an agent is only trained to handle a single call type. Agents are then less costly, but FD would require a larger staffing level to reach the same service level as in FF or any other architecture.

Full flexibility and full dedication, however, are only two extreme situations. A well-known intermediate configuration is *chaining*, first pointed out by Jordan and Graves (1995). Under chaining, each call type can be assigned to one of two adjacent agent teams, and each agent can handle calls from two adjacent types. Sheikhzadeh et al. (1998), Gurumurthi and Benjaafar (2004), and Jordan et al. (2004) prove that chaining, with an appropriate linkage between demand and resource types, behaves just as well as full flexibility. In the context of Constant Work in Process (CONWIP) serial production lines, Hopp et al. (2004) showed that the impact of forming a complete chain of skill sets can be substantial in increasing throughput. Wallace and Whitt (2005) consider the problem of routing and staffing in multi-skill call centers. They again confirm the principal that a little flexibility has the potential to achieve the performance of total flexibility. Using simulation they demonstrate that the performance, with an appropriate and limited cross-training of agents (two skills per agent) such as in chaining, is almost as good as when each agent has all skills.

Developing intelligent configurations such as chaining is very interesting for practitioners. They allow to capture the benefits of pooling by only having a limited flexibility. However, the robustness of chaining fails in the case of asymmetric demand (Sheikhzadeh et al. (1998)). By asymmetric demand, we mean different workload intensities and service time requirements, and also different variabilities in inter-arrival and service times. For such cases in practice, it is important to develop new architectures that allows from on one hand to account for demand asymmetry, and on the other hand to capture the benefits of pooling with only a limited flexibility.

In this paper, we consider skill-based routing (SBR) call centers with two particular features: demand asymmetry and costly/difficult agent training. The typical example is that of an European multilingual call center where customers call from several countries. It is difficult for managers to find agents speaking more than two languages. For instance, in the call center of an European

Airline company, each agent speaks two languages: her own native language and English. Note that this call center is more interested in agents speaking two languages rather than those speaking three or more languages. The reason is that the latter often feel themselves over-qualified. They are therefore likely to leave the company faster than the others, which increases the turnover. The workload is also unbalanced ranging from only some few calls from a given country to several thousand of calls from another country. Another example is post-sales service call centers of major retailers that are, at the same time, distributors of white goods, telecommunications products, information technology, but also internet services, photo services or travel services. We also give the example of retail banking call centers where questions are with regard to savings or stock exchange for examples. The main characteristics in the previous examples are (i) the demand is unbalanced, (ii) the required agent skills can be very different which make difficult or too costly the agent training, and (iii) one may find a predominant and "easy" type of questions that could be handled by most of the agents without any particular training, for example the English task in a multilingual call center, account information and simple bank tasks in banking, order tracking and payment for retailers, etc.

Main findings. Motivated by this prevalence in practice, we propose a new organizational model, referred to as *single pooling* (SP), where we dedicate a team of agents to each difficult type of calls, and the easy type of calls have access to all agents from all teams. Balancing the workload among the agents in this way captures the benefits of pooling without requiring every agent to process every call type. We do not claim that our model is better than chaining in all cases, but only in the particular situations of the call center examples above. The value of our architecture is that it has a low degree of flexibility (each agent handles one difficult type and the easy task) while behaving in terms of performance as a fully flexible call center. This is important in practice since additional flexibility often comes at the cost of high operating overhead.

Using simulation, we conduct a comprehensive comparison between single pooling and chaining. As a function of the various system parameters, we delimit the regions where either chaining or single pooling is the best. Few of our key findings are highlighted next. Single pooling leads to better performance while being less costly than chaining under various situations of asymmetry between the customer types: asymmetry in the number of arrivals, in the service and abandonment times, in the variability of service times, or in the service level requirements. Moreover, we conclude that these observations are more apparent for situations with a large number of skills, or for those with a large call center size.

The rest of the paper is organized as follows. In Section 2, we review some of the literature related to this paper. In Section 3, we describe chaining and single pooling models, and provide

the comparison framework. Under some particular assumptions, we develop in Section 4 two numerical approximate methods for the analysis of single pooling and chaining. In Section 5, we use simulation to compare between the two models under various situations of asymmetry on the parameters. Section 6 concludes the paper and highlights some future research.

2 Literature Review

There is an extensive and growing literature on call centers. We refer the reader to Akşin et al. (2007) for an overview. We review in what follows some of the literature related to this work.

Impact of Pooling. The value of pooling comes from the creation of flexibility. The general known intuition is that pooled systems are more effective than independent ones. The impact of pooling has been first studied in Smith and Whitt (1981). They show that pooling always leads to better performance in terms of the expected delay in the queue. Akşin and Karaesmen (2007) investigate the impact of the call center size on the opportunity to add flexibility. They demonstrate that a small call center will benefit more from adding flexibility than a larger one.

Benjaafar (1995) studies the impact of pooling for a variety of manufacturing, telecommunication and computer systems. He considers a multi-processing system consisting of several facilities and shows that in some situations of heterogeneity in the workloads, increasing flexibility can deteriorate performance. Mandelbaum and Reiman (1998) consider stochastic service systems modeled as queueing networks. They show that adding flexibility does not automatically improve performance. They point out that adding a partial flexibility could be devastating. Recently, van Dijk and van der Sluis (2008) show in the context of SBR call centers that without any clever routing rules and under a high variability in the call types and the resources, pooling could deteriorate the performance in terms of the average waiting time. Inspired by the results of Smith and Whitt (1981), Tekin et al. (2009) show that pooling teams could be counterproductive if service time means are very different from one customer type to another (for example when one is about six times higher than the other ones).

Flexible Architectures. The most fundamental work on flexibility is that by Jordan and Graves (1995) for the automobile assembly plants, but it can be also applied to broader manufacturing system settings. They conclude that "a little flexibility can achieve almost all the benefits of total flexibility" under a configuration referred to as chaining, with two product types per plant. Similar results in the context of cellular manufacturing systems have been found by Garavelli (2001, 2003). We also refer the reader to Albino and Garavelli (1999); Nomden and van der Zee (2008).

For queueing systems, Gurumurthi and Benjaafar (2004) compare different scenarios of adding flexibility under different routing policies. They prove that the value of chaining decreases for an

asymmetric demand. Hopp and van Oyen (2004) consider the question of how to cross-train a worker to two skills in the context of serial production lines. They conclude that a novel strategy called skill chaining strategy is more robust against variability than a cherry-picking strategy (a team is full flexible) when demand is symmetric. The cherry-picking strategy in a serial production line can be seen as similar to single pooling, where the customers are the machines, and the bottleneck machine represents the easy type of calls. Tomlin and Wang (2005) consider the context of unreliable supply chains that produce multiple products. They study four canonical supply chain design strategies, where one of them, referred to as dual-source flexible, has been already proposed by Chevalier et al. (2004) in the context of call centers. They refine the prevailing intuition that a flexible network is preferable to a dedicated network by proving that this intuition is valid if either the resource investments are perfectly reliable or the firm is risk neutral. In a similar setting to ours, Robbins and Harrison (2010) introduce an SBR call center queueing model with two customer types, referred to as partial pooling. They consider two dedicated agent teams for each customer type, and one cross-trained team for both types. They show that cross-training a small number of agents can deliver a substantial benefit.

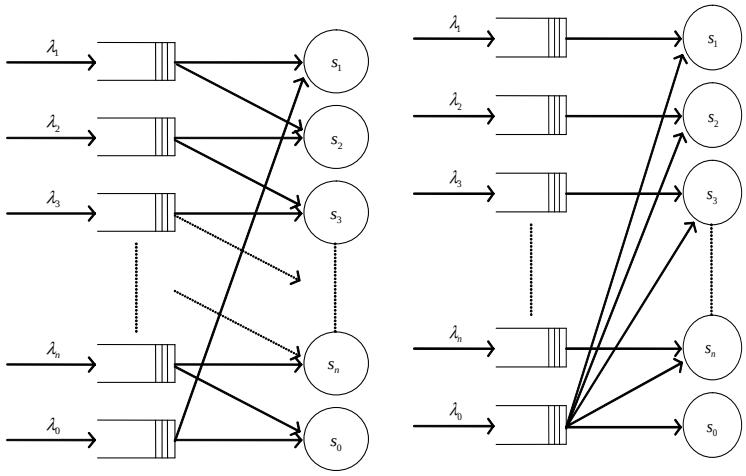
Garnett and Mandelbaum (2001) argue on the importance of adapting the system architecture to the asymmetry in the customer arrival rates. In summary, chaining is robust according to its ability to support variability. It however fails when the demand is asymmetric. It can be also too expensive to train agents on various combinations of two skills. For these situations, we propose and analyze in this paper a new efficient configuration of a queueing call center model.

3 Problem Setting

We consider call center models with $n+1$ call types (types 0, 1, ..., n). Customer types 1, 2, ..., n, referred to as also regular types are those requiring specific agent skills 1, 2, ..., n, respectively, while customers 0 can be handled by any agent without a particular "sophisticated" training as required for the regular types. In other words, skill 0 is an easy skill. The mean arrival, service and abandonment rates of customers type i are λ_i, μ_i and γ_i, respectively ($i = 0, 1, ..., n$). The agents are organized in homogeneous teams, i.e., all agents from a given team have the same set of skills. We only consider agent teams with at most two skills per agent. We define an economic framework as follows. We assume that skill 0 costs 1, and that skill i costs $1+t_i$ (for $i = 1, \cdots, n$). For two skills i and j, the cost is $1+t_{i,j}$ (for $i, j \in \{0, \cdots, n\}$). Since skill 0 is the easy skill, we assume that $t_{i,0} \leq t_{i,j}$ (for $i, j \in \{0, \cdots, n\}$).

We focus on the performance in terms of the steady-state expected waiting time in the queue of each customer type i taken in service, denoted by W_i, for $i = 0$, 1, ..., n. We denote the objective

service level for a type i by W_i^*, for $i = 0, 1, ..., n$. In what follows, we describe the two models that we compare in this paper: chaining and single pooling. They are shown in Figures 1(a) and 1(b), respectively.



(a) Chaining (b) Single pooling

Figure 1: Call center configurations

Under chaining, a customer upon her arrival has access to agents from two teams. If at least an agent is available in one of them, then the customer is routed to the team with the higher proportion of idle agents (number of idle agents in a team over the total number of agents in that team). If this proportion is the same for the two teams, then she is equiprobably routed to one of the two teams. Otherwise if all agents from the two teams are busy upon her arrival, the customer waits in her queue (each customer type has its own infinite queue). An agent can handle customers from two queues. Within each queue, the discipline of service is FCFS. When an agent becomes idle, she selects to service one of the customers that are waiting in the two queues, if any. The priority is given to the customer with the longest waiting time.

For single pooling, the discipline of service in each one of the $n + 1$ infinite queues is FCFS. A customer type i ($i = 1, ...n$) can be served by only an agent from its associated team. A customer 0 however can be served by any agent from any one of the $n + 1$ teams. Upon arrival, a customer 0 is in priority handled by an idle agent from team 0, if any. If not, she is handled by an idle agent from one of the teams of the regular types, if any. If more than one team have at least one idle agent, then customer 0 is routed to the team with the higher proportion of idle agents. If many teams have the same highest proportion, then customer 0 is equiprobably routed to one of these teams. If all agents of all teams are busy, then customer 0 is placed in her queue. When an agent from one of the teams of the regular customers becomes free, it can serve either a regular customer or a customer 0. However a regular customer has a non-preemptive priority over a customer 0.

We compare between the two models chaining and single pooling through simulations. In order to have a coherent comparison we optimize their total staffing cost under the constraints $W_i \leq W_i^*$, for $i = 0, 1, ..., n$. We use greedy heuristics for the simulation based optimization step. We refer the reader to the details in Section 1 of the online supplement. For the staffing optimization of SP, we use an increasing greedy algorithm. Starting from an under-staffed situation (a full dedicated model with customers 0), we increase step by step the arrival rate of customers 0. In each iteration, we increment the number of agents in the various teams such that we strictly reach the service level constraints. For chaining, we develop a decreasing greedy algorithm. The algorithm starts with an over-staffed situation using a full dedicated model, which is the worst for chaining since it ignores the links between the teams. We then use the approach suggested by Wallace and Whitt (2005) in order to correct the staffing levels to the chaining setting.

4 Approximate Numerical Comparison

We numerically compute approximate expected waiting times for single pooling and chaining. For tractability, we consider Markovian assumptions for inter-arrival and services times, and customer abandonment is ignored. The objective of this analysis is to obtain some sense on the effect of the parameters asymmetry on the comparison between the two architectures. A more comprehensive analysis is then conducted in Section 5 using simulation. We employ a Markov chain method for the performance analysis of each design. We first compute the steady-state system probabilities, from which we deduce the expected waiting time for customers type i, $i = 0, 1, ..., n$.

Single Pooling. Consider a single pooling model with $n + 1$ skills and $n + 1$ teams, $n \geq 1$. Let us define the stochastic process $\{x(t), t \geq 0\}$ as

$$\{x(t), t \geq 0\} = \{(x_0(t), x_1(t), \cdots, x_n(t), x_{0,1}(t), \cdots, x_{0,n}(t), q_0(t), q_1(t), \cdots, q_n(t)), t \geq 0\},$$

where for an instant $t \geq 0$, $x_i(t)$ denotes the number of agents in team i that are busy with a customer i, for $i = 0, 1, \cdots n$; $x_{0,i}(t)$ denotes the number of agents in team i that are busy with a customer 0, for $i = 1, \cdots n$; and $q_i(t)$ denotes the number of customers in queue i, for $i = 0, 1, \cdots n$. Since inter-arrival and service times are Markovian, $\{x(t), t \geq 0\}$ is a Markov chain with $3n + 2$ dimensions. Let us denote the system steady-state probabilities by π_x,

$$\pi_x = \pi_{x_0,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_n},$$

with x_i, $x_{0,i} \in [0, s_i]$ for $i = 0, 1, \cdots n$, and $q_i \in \mathbb{N}$ for $i = 0, 1, \cdots n$. The computation method of these probabilities is given in the appendix. The expected waiting time for customers type i may be then written as

$$W_i = \frac{1}{\lambda_i} \left(\sum_{x_0=0}^{s_0} \sum_{x_1+x_{0,1}=0}^{s_1} \cdots \sum_{x_n+x_{0,n}=0}^{s_n} \sum_{y_0,\cdots,y_n=0}^{D} q_i \pi_x \right),$$

for $i = 0, 1, \cdots n$, where D is a truncation point as explained in the appendix.

Chaining. Consider a chaining model with $n+1$ skills and $n+1$ teams. Because of the routing mechanism in chaining, a standard Markov chain modeling is not appropriate. Once an agent completes a service, shes chooses next to service the oldest customer among those in the head of two queues, if any. A standard modeling only based on the number of customers in the queues can not take this decision into account. We thus propose to discretize the waiting time of the first in line in each queue instead of using the number of agents in each queue. The modeling of the first in line as a tool for analyzing a queueing system was proposed by Koole et al. (2012). Let us define the stochastic process $\{x(t), t \geq 0\}$ by the tuple

$$\{(x_{0,1}(t), x_{1,1}(t), x_{1,2}(t), x_{2,2}(t) \cdots, x_{n-1,n}(t), x_{n,n}(t), x_{n,0}(t), x_{0,0}(t), q_0(t), q_1(t), \cdots, q_n(t)), t \geq 0\},$$

where for an instant $t \geq 0$, $x_{i,j}(t)$ denotes the number of agents in team j that are busy with a customer i, for $i, j = 0, 1, \cdots n$; and $q_i(t)$ denotes the stage of the waiting time of the first in line in queue i, for $i = 0, 1, \cdots n$. We consider an exponential elapsing of time with parameter ζ. Recall from Koole et al. (2012) that when the first customer in line leaves queue i from a given stage of the waiting duration k ($k > 0$), the weight of the transitions from this state k to a state $k - h$ for $k > 0$ and $0 \leq h \leq k$, $p_{k,k-h}$ are

$$p_{k,k-h} = \begin{cases} 1 - \sum_{h=0}^{k-1} \left(\frac{\lambda_i}{\lambda_i+\zeta}\right) \left(\frac{\zeta}{\lambda_i+\zeta}\right)^h = \left(\frac{\zeta}{\lambda_i+\zeta}\right)^k, \text{ for } k = h \\ \left(\frac{\lambda_i}{\lambda_i+\zeta}\right) \left(\frac{\zeta}{\lambda_i+\zeta}\right)^h, \text{ for } 0 \leq h < k. \end{cases} \tag{1}$$

Since inter-arrival, service and elapsing times are exponentially distributed, $\{x(t), t \geq 0\}$ is a Markov chain with $3n + 3$ dimensions. Let us denote the system steady-state probabilities by π_x,

$$\pi_x = \pi_{x_{0,1}, x_{1,1}, x_{1,2}, x_{2,2} \cdots, x_{n-1,n}, x_{n,n}, x_{n,0}, x_{0,0}, q_0, q_1, \cdots, q_n},$$

with $x_{i,j} \in [0, s_j]$ for $i, j = 0, 1, \cdots n$; and $q_i \in \mathbb{N}$ for $i = 0, 1, \cdots n$. The computation method of these probabilities is given in the appendix. The expected waiting time for customers type i, for $i = 0, 1, \cdots n$, is then given by

$$W_i = \frac{1}{\lambda_i} \sum_{x_{0,1}=0}^{s_1} \sum_{x_{1,1}=0}^{s_1} \sum_{x_{1,2}=0}^{s_2} \cdots \sum_{x_{n,n}=0}^{s_n} \sum_{x_{n,0}=0}^{s_0} \sum_{x_{0,0}=0}^{s_0} \sum_{q_0,\cdots,q_n=0}^{D}$$
$$\frac{q_i}{\zeta} \left(\mathbf{1}_{(q_i > q_{i-1})} \mathbf{1}_{(q_i > q_{i+1})} + 0.5(\mathbf{1}_{(q_i = q_{i-1})} \mathbf{1}_{(q_i > q_{i+1})} + \mathbf{1}_{(q_i > q_{i-1})} \mathbf{1}_{(q_i = q_{i+1})}) \right)$$
$$(x_{i-1,i}\mu_{i-1} + (x_{i,i} + x_{i,i+1})\mu_i + x_{i+1,i+1}\mu_{i+1})\pi_x,$$

where $\mathbf{1}_{(x \in A)}$ is the indicator function of a subset A, and again D is a truncation point as explained in the appendix.

Numerical Illustration. An exact analytical comparison between single pooling and chaining is too complex. The two architectures are SBR queueing models with complex routing mechanisms and general settings for the parameters. Even in the case of Markovian assumptions, the analysis is very difficult, and there are no existing exact results for them in the literature. Using the approximate analysis above, we numerically illustrate the comparison. We first consider a real multi-language call center setting. We then generate various other settings of asymmetry to cover other call center settings. The real example consists of an airline company call center, located in Australia and handling 4 types of customers: Japanese (type 1), Korean (type 2), Bahasa (type 3) and English (type 0) speaking customers. Customer types are identical in their requests (flight booking and modification, claims, etc.). The expected service time is the same for all types, $\frac{1}{\mu_i}$=6.8 minutes for $i = 0, ..., 3$. An example of the daily arrival rates is given in Figure 2. For the numerical illustration, we consider a given time interval with the parameters $\lambda_0 = 4.6$, $\lambda_1 = 7.7$, $\lambda_2 = 10.1$ and $\lambda_3 = 1.5$. Note that we ignore here several features such as abandonment, retrial, rejection, agent reservation routing rules, back-office tasks, etc.

This call center uses the SP architecture, where an agent from a given team has skill 0 and skill i, for $i = 0, 1, 2, 3$. Let us compare the costs of using SP and chaining. We know from this call center that the salary per hour of an agent with the easy skill (English) and 1 regular skill (one of the other languages), is 20% higher than that of an agent with only the easy skill. Also, the salary of an agent with the easy skill and 2 regular skills, is 16% higher than that of an agent with the easy skill and 1 regular skill. We then consider that the salary of an agent in SP is 1.2 and that in chaining is either 1.2 or 1.4 according to her set of skills. Under a service level constraint ($W_i^* = 0.2$ for $i = 0, ..., 3$), the total staffing costs are 230.2 and 210 for chaining and

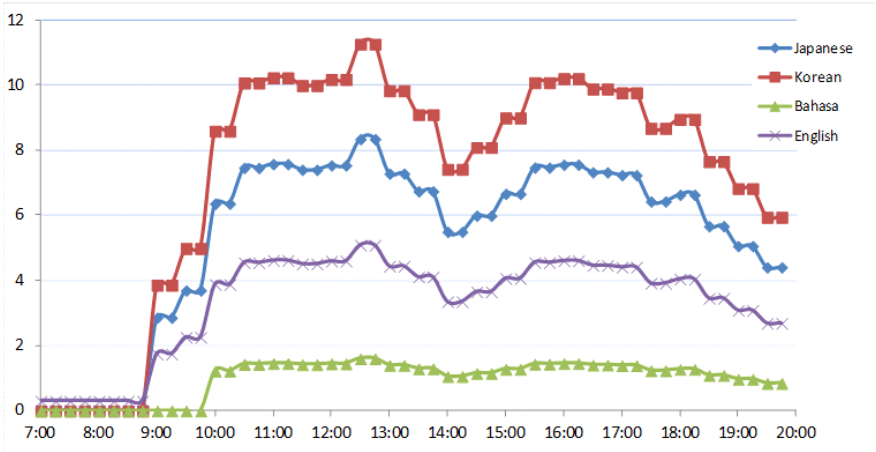


Figure 2: Customer arrival rates

SP, respectively. SP behaves better in this example because of the asymmetry in the arrival rates and also the agents salary structure. Using the same cost structure, we next compare between SP and chaining by generating 9 scenarios with different levels of asymmetry in the arrival and service rates. We compute the overall staffing costs for SP and chaining, and also their relative difference. A positive relative difference corresponds to a higher cost for SP, and viceversa. The results are shown in Table 1.

Table 1: Comparison between SP and chaining ($n = 4$, $\zeta = 30$, $D = 200$, $W_i^* = 0.2$ for $i = 0, ..., 3$)

	λ_0	λ_1	λ_2	λ_3	μ_0	μ_1	μ_2	μ_3	Cost SP	Cost Chaining	Relative difference
Sc 1	1	1	1	1	0.2	0.2	0.2	0.2	36	35	2.86%
Sc 2	1	2	3	4	0.2	0.2	0.2	0.2	75.6	78	-3.08%
Sc 3	4	3	2	1	0.2	0.2	0.2	0.2	68.4	78	-12.31%
Sc 4	1	1	1	1	0.1	0.2	0.3	0.4	42	38	10.53%
Sc 5	1	1	1	1	0.4	0.3	0.2	0.1	37.2	38	-2.11%
Sc 6	1	2	3	4	0.4	0.1	0.2	0.3	78	78.6	-0.76%
Sc 7	4	1	2	3	0.4	0.1	0.2	0.3	54	61.2	-11.76%
Sc 8	1	2	3	4	0.1	0.2	0.3	0.4	62.4	61.2	1.96%
Sc 9	4	1	2	3	0.1	0.2	0.3	0.4	85.2	72.8	17.03%

We observe that the asymmetry in the parameters has an effect on the comparison between SP and chaining. For a symmetric case in arrival and service rates (scenario 1), we observe that chaining is the best. For an asymmetric case in the arrival rates (scenarios 2 and 3), SP is the best. When the asymmetry is in service rates (scenarios 4 and 5), we observe that chaining is the best for the case of slowly served customers 0, and viceversa. In scenarios 6-9, we observe that the mix of asymmetry in both arrival and service rates makes SP prevail in some situations, whereas chaining does in others. While the observation related to the benefits of pooling for customers 0 in SP is evident (scenario 3), others are not clear and require a deeper analysis. Note that the approximations used here

do not account for customer abandonment or non Markovian distributions, where the asymmetry may have an important effect as we show later. In order to obtain a comprehensive understanding of the comparison, we resort to simulation in the next section. In using simulation for call center operations management, we are following longstanding practice, see for example Wallace and Whitt (2005).

5 Effect of Parameter Asymmetry

We describe here the simulation results of the comparison between chaining and single pooling. We simplify the cost model such that the SP cost is upper bounded and that of chaining is lower bounded. The cost of single pooling is $\sum_{i=0}^{n}(1+t_{i,0})s_i$. This is upper bounded by $(\sum_{i=0}^{n} s_i)\max_i(1+t_{i,0})$. The cost of chaining is $(1+t_{0,1})s_0+(1+t_{1,2})s_1+\cdots(1+t_{n,0})s_n$ and is lower bounded by $(1+t_{0,1})s_0+(1+\min_{i,j}(1+t_{i,j}))(\sum_{i=1}^{n-1} s_i)+(1+t_{n,0})s_n$. Let us now simplify the problem as follows. An agent with skills 0 and i $(i=1,...,n)$ costs 1. An agent with skills i and j $(i,j=1,...,n$ and $i\neq j)$ costs $1+t$, $t\geq 0$. In this simplification, we have $\max_i(1+t_{i,0})=1$ and $\min_{i,j}(1+t_{i,j})=1+t$ $(i,j=1,...,n$ and $i\neq j)$. The parameter t is then the incremental cost of an agent with two regular skills compared to that with a regular skill and skill 0. All the numerical comparisons are based on the lower and upper bounds values. This makes the results pessimistic for SP and optimistic for chaining, though the bounds are likely to be tight in practice.

Design of Experiments. As we are interested in the effect of asymmetry of the parameters on performance, we propose various forms of asymmetry. For customers 0, we define the parameters p and p' to measure the relative importance in arrivals and service durations, respectively. They are given by $p=\frac{\lambda_0}{\sum_{i=0}^{n}\lambda_i}$ and $p'=\frac{\frac{1}{\mu_0}}{\sum_{i=0}^{n}\frac{1}{\mu_i}}$. We measure the asymmetry between the arrival rates of regular customers by $V=\frac{\lambda_1}{\lambda_2}=\frac{\lambda_2}{\lambda_3}=\cdots=\frac{\lambda_{n-1}}{\lambda_n}$, and that between service durations by $U=\frac{1/\mu_1}{1/\mu_2}=\frac{1/\mu_2}{1/\mu_3}=\cdots=\frac{1/\mu_{n-1}}{1/\mu_n}$. We also consider for customers 0 the asymmetry in the variability of service times, measured by the coefficient of variation of its distribution and denoted by cv_s. We consider other forms of asymmetry in terms of the required service level and also the time to abandon for customers 0 relatively to those for the regular customers. These effects are studied in the settings of small and large call centers, and also in the settings of small and large number of skills. Although the considered forms of asymmetries do not cover all possibilities, they allow to obtain the main useful conclusions.

The approach to conduct the simulation experiments is as follows. Due to the high number of parameters, we first run experiments by separately treating one parameter at a time. In a systematic way, we vary one parameter while holding all the others constant. Second, to assess the possible interaction effects, we simultaneously vary the values of more than one of them at a time.

For the values of the parameters, we choose wide ranges that allow to cover most of call center situations in practice. For the rest of the paper, inter-arrival times are assumed to be Markovian. Service times are also assumed to be Markovian, except in Section 5.2.1. The abandonment rates are assumed to be equal to zero, except for Section 5.4.

5.1 Asymmetry in Arrival Rates

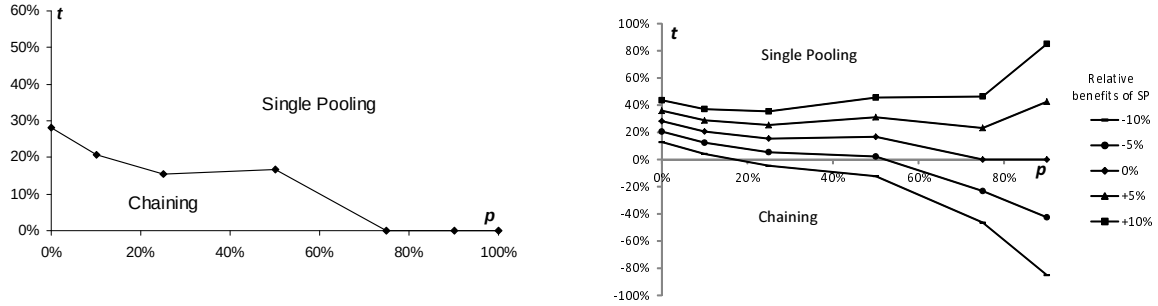
We first construct the asymmetry only on the arrival rate of customers 0. We then construct it by differentiating between all the arrival rates of all customer types.

5.1.1 Asymmetry on Customers 0

To isolate the impact of $p = \frac{\lambda_0}{\sum_{i=0}^{n} \lambda_i}$, we assume that all customer types have the same expected service time, and all the arrival rates of the regular customers are the same, $\lambda_i = \lambda$ for $i = 1, ..., n$ ($V = 1$). In particular, we are interested to know, for the different ranges of p, which one of the models would be preferred to the other. We choose call center examples with $n = 4$, i.e., 5 agent teams and 5 skills including skill 0. The results are shown in Table 2 and Figures 3(a) and 3(b).

Table 2: Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = V = 1$, $p' = 20\%$, $n = 4$)

	Chaining						SP	Crossing value
p	t=0%	t=5%	t=10%	t=25%	t=50%	t=100%		(Chaining = SP)
0%	49	50.95	52.9	58.75	68.5	88	60	t=28.21%
10%	49	50.7	52.4	57.5	66	83	56	t=20.58%
25%	48	49.3	50.6	54.5	61	74	52	t=15.38%
50%	49	49.9	50.8	53.5	58	67	52	t=16.67%
75%	51	51.55	52.1	53.75	56.5	62	51	t=0%
90%	51	51.3	51.6	52.5	54	57	51	t=0%
100%	47	47	47	47	47	47	47	t=0%



(a) Preference zone (b) Relative benefits

Figure 3: Comparing single pooling and chaining ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = V = 1$, $p' = 20\%$, $n = 4$)

Since any agent in SP has skills 0 and i (i.e., costs 1), the staffing cost of SP does not depend

on t. In Table 2, the column *Crossing value* gives the value of t for which the two models chaining and SP are equivalent. Below this threshold chaining is better than SP and viceversa (see Figure 3(a)). Consider small values of t. Table 2 reveals that chaining performs well for small values of p. The best situation for chaining is reached in the symmetric case (identical arrival rates). The performance of SP improves as p increases. For small values of p, SP approaches FD which has the worst performance. For high values of p, customers 0 are first preponderant and second benefit from pooling, which highly improves the performance of SP. With $t = 0$, SP and chaining become equivalent for values of $p \geq 75\%$.

For higher values of t, SP goes ahead of chaining. The reason is related to the increase of the costs of the agents with two skills i and j $(i, j = 1, ..., 4)$. It suffices to have $t = 15.38\%$ to outperform the best performance of chaining (the symmetric case). Note that for the real-life airline company example of Section 4, t is about 16%. For any t beyond 30%, SP is systematically better than chaining whatever is p. We also measure the relative benefits between SP and chaining. Figure 3(b) provides, for various values of the relative benefits, the associated curve of t as a function of p. We observe that the sensitivity of the relative benefit as a function of t decreases in p. The reason is that the number of customers 0 increases in p, which decreases the number of agents with two regular skills in chaining (i.e., decreases the cost sensitivity in t). The main conclusion here is that SP can be better than chaining when the demand for skill 0 is important and/or when skill 0 is less costly than the other ones.

5.1.2 Asymmetry on the other Arrival Rates

We focus on the asymmetry between regular customer types, measured by $V = \frac{\lambda_1}{\lambda_2} = \frac{\lambda_2}{\lambda_3} = \frac{\lambda_3}{\lambda_4}$. The simulation results for the cases $V = 2$ and 5 are shown in Table 3. The experiments for the case $V = 1$ reduces to those given in Table 2.

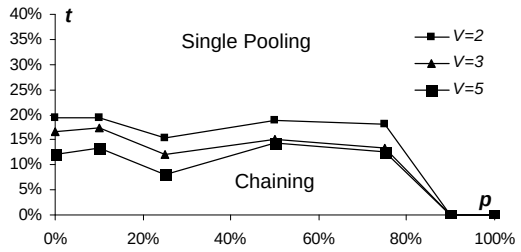


Figure 4: Preference zone ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = 1$, $p' = 20\%$, $n = 4$)

Table 3: Impact of V ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = 1$, $p' = 20\%$, $n = 4$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$V = 2$	0%	50	51.8	53.6	59	68	57	t=19.44%
	10%	50	51.55	53.1	57.75	65.5	56	t=19.35%
	25%	49	50.3	51.6	55.5	62	53	t=15.38%
	50%	48	48.8	49.6	52	56	51	t=18.75%
	75%	50	50.55	51.1	52.75	55.5	52	t=18.18%
	90%	52	52.3	52.6	53.5	55	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$V = 3$	0%	50	51.8	53.6	59	68	56	t=16.67%
	10%	50	51.45	52.9	57.25	64.5	55	t=17.24%
	25%	49	50.25	51.5	55.25	61.5	52	t=12.00%
	50%	49	50	51	54	59	52	t=15.00%
	75%	50	50.75	51.5	53.75	57.5	52	t=13.33%
	90%	52	52.2	52.4	53	54	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$V = 5$	0%	49	51.05	53.1	59.25	69.5	54	t=12.20%
	10%	50	51.5	53	57.5	65	54	t=13.33%
	25%	50	51.25	52.5	56.25	62.5	52	t=8.00%
	50%	50	50.7	51.4	53.5	57	52	t=14.29%
	75%	51	51.4	51.8	53	55	52	t=12.50%
	90%	52	52.25	52.5	53.25	54.5	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%

Table 3 and Figure 4 reveal that the performance of SP increases in V. An intuitive explanation is as follows. Remark that the team size $s_i = s(\lambda_i)$ is increasing and concave in λ_i, for $i = 1, ..., n$. Applying then the Jensen inequality leads to $\sum_{i=1}^{n} s(\lambda_i) \leq n \cdot s\left(\frac{\sum_{i=1}^{n} \lambda_i}{n}\right)$. Note that the equality may happen because of the discrete nature of the staffing levels. In the inequality, the left hand side corresponds to the overall staffing level for an arbitrary situation, i.e., with arbitrary values of λ_is. As for the right hand side, it gives the overall staffing level for a symmetric situation, i.e., all the λ_is are identical. We also observe from Table 3 that the performance of chaining is however relatively insensitive to V. Note that we change each time the configuration of chaining such that the large teams are close to each others in order to create more pooling effect. This is better than having small teams each of which connected to a large team.

5.2 Asymmetry in Service Rates

We first define the asymmetry only on customers 0, and second on all customer types.

5.2.1 Asymmetry on Customers 0

We measure the asymmetry on customers 0 by $p' = \frac{\frac{1}{\mu_0}}{\sum_{i=0}^{n} \frac{1}{\mu_i}}$. The asymmetry here is defined by the difference between the value of the mean service time of customers 0 and that of the regular types. The results are shown in Table 4 and Figure 5.

From Table 4, we observe that the performance of both models chaining and SP improves in p'

Table 4: Impact of p' ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)

p'	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
0%	60	62.45	64.9	72.25	84.5	72	t=24.49%
10%	59	60.95	62.9	68.75	78.5	67	t=20.51%
25%	58	59.65	61.3	66.25	74.5	62	t= 12.12%
50%	60	61.05	62.1	65.25	70.5	65	t=23.81%
75%	61	61.6	62.2	64	67	68	t=58.33%
90%	65	65.25	65.5	66.25	67.5	69	t=80.00%

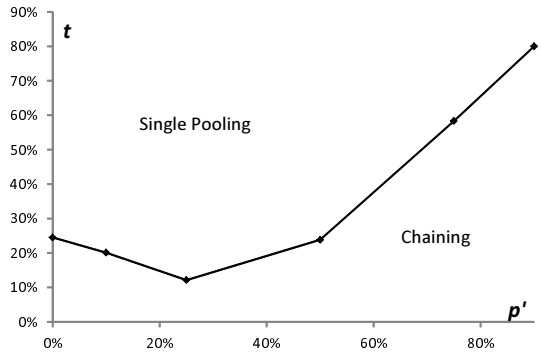
(from 0 until the symmetric case for $p' = 25\%$). The reason is that for chaining we are approaching the symmetric case where it behaves well, and for SP we are profiting better from the pooling effect when all service times are statistically identical. However the performance of the two models deteriorates in p' (for p' above 25%), and no model performs well for a high asymmetry in service times. The explanation is related to a phenomenon referred to as the *blocking effect*. The blocking effect is the situation where the agents are excessively blocked by customers 0 (who are in need of large service times) which deteriorates the waiting time of the regular customers. This phenomenon is more apparent for single pooling since in the latter customers 0 have access to all teams, whereas in chaining they do only have access to two teams.

In what follows, we go further by defining the asymmetry on the variability of customers 0 service times. We choose to measure this variability by the coefficient of variation (ratio of standard deviation over expected value), denoted by cv_s. We consider a log-normal distribution for the service times of customers 0 (inter-arrival times of all types, and service times of all regular types are Markovian). The choice of the log-normal distribution is based on the call center statistical analysis in Brown et al. (2005). The results are shown in Table 5 and Figure 5(c). We draw the same conclusions as those for service rates. Due to the blocking effect, both models do not behave well as the variability increases. Figure 5(d) reveals that the relative benefit as a function of t is not sensitive to the variation of cv_s. Contrary to the cases for p and p', the arrival and service rates of regular types do not vary here.

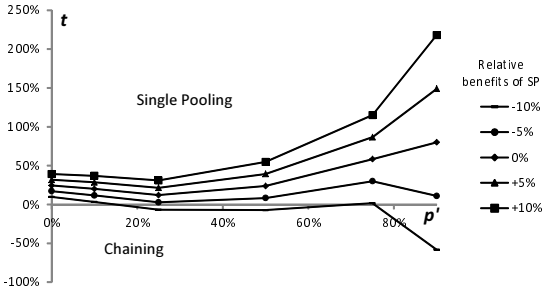
5.2.2 Asymmetry on the Other Service Rates

The service times can be now different from one regular customer to another. Recall that the ratio U is defined by $U = \frac{1/\mu_1}{1/\mu_2} = \frac{1/\mu_2}{1/\mu_3} = \frac{1/\mu_3}{1/\mu_4}$. We also consider cases with a high proportion of customers 0, $p = 50\%$. The simulation results are shown in Table 6, and Figures 6(a) and 6(b).

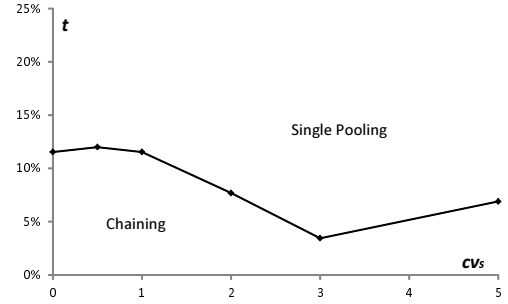
From the numerical results we observe that SP is preferred to chaining for a wide range of parameters. The performance of SP is quite insensitive to the asymmetry defined by U. The reason



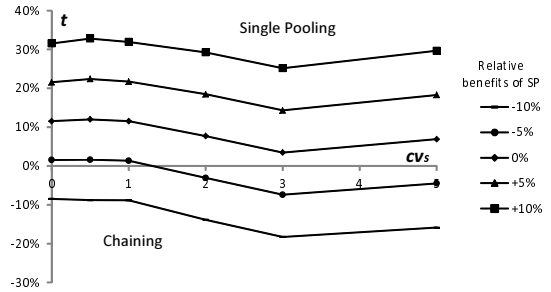
(a) Preference zone ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)



(b) Relative benefits ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)



(c) Effect of variability in service times ($\lambda_0 = 2$, $\lambda_i = 1.5$, $\mu_0 = \mu_i = 0.2$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)



(d) Relative benefits ($\lambda_0 = 2$, $\lambda_i = 1.5$, $\mu_0 = \mu_i = 0.2$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)

Figure 5: Preference zone

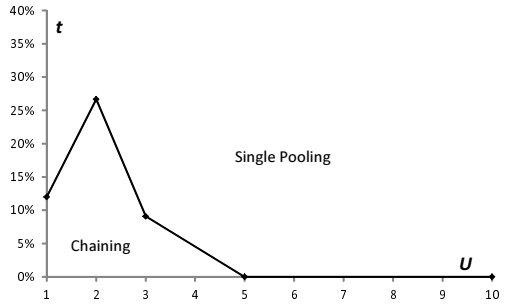
is that whatever U is, the agent teams in SP are divided into to two types. One first type with two teams where customers 0 are served faster than regular customers (positive effect), and a second type with two teams where customer 0 are served slower than regular customers (negative effect of blocking). The performance of chaining is however decreasing in asymmetry. In chaining, each team receives two customer types with different service times, which creates a negative blocking effect in all teams and deteriorates as a consequence the performance. In general for both single pooling and chaining with $U \neq 1$, regular customers require different mean service times. We then have regular customers that are served faster than others. The slowly served ones block the teams

Table 5: Impact of variability in service times ($\mu_i = \mu_0 = 0.2$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = V = 1$, $\sum_{i=0}^{4} \lambda_i = 8$, $n = 4$)

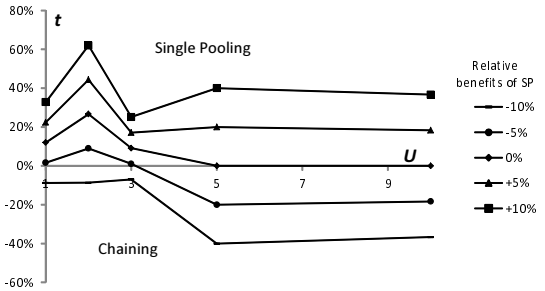
	Chaining					SP	Crossing value
cv_s	0%	5%	10%	25%	50%		(Chaining = SP)
0	49	50.3	51.6	55.5	62	52	11.54%
0.5	49	50.25	51.5	55.25	61.5	52	12.00%
1	50	51.3	52.6	56.5	63	53	11.54%
2	54	55.3	56.6	60.5	67	56	7.69%
3	62	63.45	64.9	69.25	76.5	63	3.45%
5	64	65.45	66.9	71.25	78.5	66	6.90%

Table 6: Impact of U ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $V = 1$, $n = 4$)

U	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
1	49	50.25	51.5	55.25	61.5	52	t=12.00%
2	49	49.75	50.5	52.75	56.5	53	t=26.67%
3	50	51.65	52.3	54.25	57.5	53	t=9.09%
5	52	52.65	53.3	55.25	58.5	52	t=0.00%
10	55	55.75	56.5	58.75	62.5	55	t=0.00%



(a) Preference zone



(b) Relative benefits

Figure 6: Preference zone ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$ for $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $V = 1$, $n = 4$)

in which they are routed to. This is more apparent in chaining because regular customers are routed to two teams (and to only one in SP). We also measure the relative benefits between SP and chaining. Figure 6(b) reveals that this benefit as a function of t is not sensitive to the variation of U. The reason is that although the service rates of regular types do vary, the total staffing level for the regular types do not.

Remark. We have so far compared SP and chaining based on the staffing costs. In what follows, we instead compare between their expected waiting times for a given same total staffing level. The results are shown in Table 7. Note that we optimize the staffing of the various teams in the two models for the case $t = 0$, i.e., no incremental cost for regular skills. Under this framework, we again observe the same qualitative conclusions as those derived under the staffing costs comparison.

5.3 Asymmetry in the Service Level Constraints

We define the asymmetry on the service level of customers 0, W_0^*. The results are shown in Table 8 and Figure 7.

We observe as expected that SP behaves better than chaining in the case of a high asymmetry in the service levels. Chaining is requiring higher staffing levels than needed for some customer

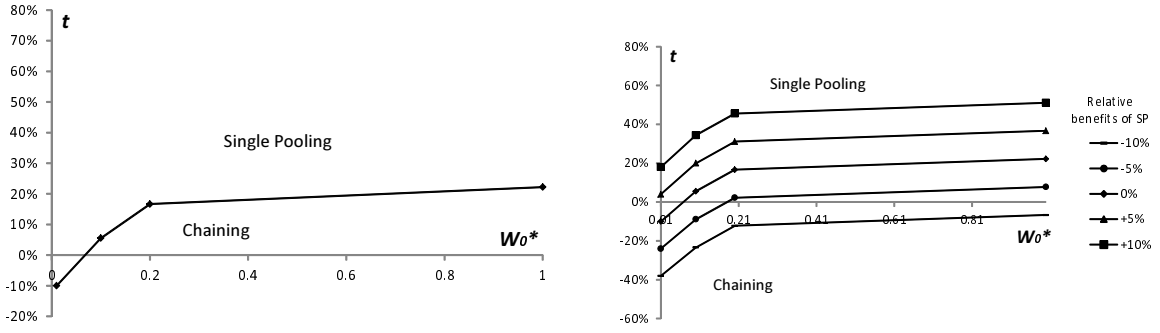
Table 7: Performance measures of SP and chaining

Impact of p			Impact of p'			Impact of V			Impact of U		
p	SP	Chaining	p'	SP	Chaining	V	SP	Chaining	U	SP	Chaining
0%	3.41	0.86	0%	0.78	0.09	1	0.91	0.67	1	0.33	0.17
10%	1.47	0.77	10%	0.30	0.09	2	0.84	0.76	2	0.37	0.19
25%	0.91	0.67	25%	0.25	0.13	3	0.77	0.73	3	0.39	0.34
50%	0.74	0.74	50%	0.97	0.76	5	0.68	0.66	5	0.39	0.38
75%	0.71	0.89	75%	5.11	4.10						
90%	0.66	0.69	90%	139	102						
100%	0.54	0.54									

Table 8: Impact of W_0^* ($\lambda_0 = 4$, $\lambda_i = 1$, $\mu_i = \mu_0 = 0.2$ and $W_i^* = 0.2$ for $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)

W_0^*	Chaining t=0%	t=5%	t=10%	t=25%	SP	Crossing value (Chaining = SP)
0.01	58	59	60	63	56	t=-10.00%
0.1	51	51.9	52.8	55.5	52	t=5.56%
0.2	49	49.9	50.8	53.5	52	t=16.67%
1	48	48.9	49.8	52.5	52	t=22.22%

types. The agent teams are less correlated in SP than in chaining. This gives more flexibility under SP to adjust the size of the teams as required. However, the strong link between the chains in chaining forces the size of the teams to be adjusted with regard to the high requirement of some customer types while it is not needed for other types. As for the relative benefits between SP and chaining, we observe from Figure 7(b) that it is not sensitive to the variation of W_0^*. Since the parameters related to the regular types do not vary, the associated staffing levels do not change also.



(a) Preference zone (b) Relative benefits

Figure 7: Preference zone ($\lambda_0 = 4$, $\lambda_i = 1$, $\mu_i = \mu_0 = 0.2$ and $W_i^* = 0.2$ for $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)

5.4 Asymmetry in Abandonments

We allow in this section customers to abandon. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time she will abandon

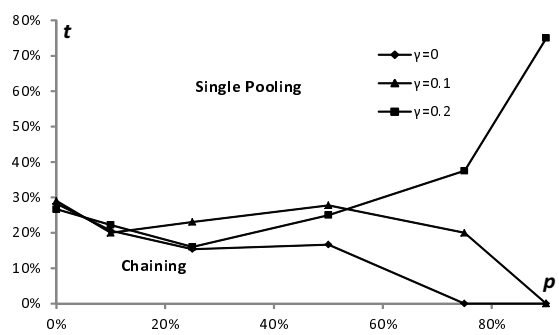
and be lost. We first investigate the impact of abandonment on the performance of single pooling and chaining. We then investigate the effect of the asymmetry in the abandonment rate of customers 0. Recall that that γ_i denotes the abandon rate of customers i, for $i = 0, \cdots n$. In the experiments below, times before abandonment are assumed to be exponentially distributed. Note that with customer abandonment, new performance measures do appear for waiting times. Since the sojourn time in queue may end up with a start of service or an abandonment, we distinguish the conditional waiting time given service, that given abandonment, and the unconditional one. We focus here on the conditional waiting time given service.

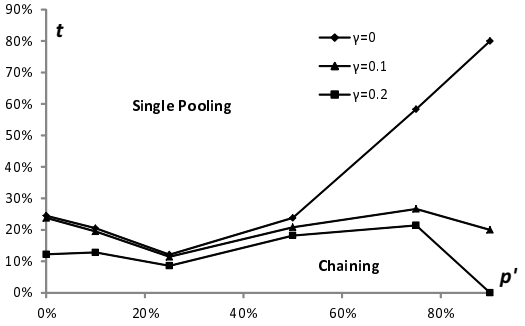
Impact of Abandonment. We investigate the impact of abandonment on the performance of SP and chaining in various situations of asymmetries. We consider homogeneous abandonments for all customer types, $\gamma_i = \gamma$ for $i = 0, \cdots n$. The results are shown in Figures 8(a)-8(d). Further results are also given in Tables 6-9 in Section 3 of the online supplement. An important observation here is that the effect of the parameters asymmetry changes in the presence of abandonment. For example, to the contrary to the results with no abandonment, the performance of SP deteriorates in p, but improves in p'. The reason is that the abandonment of customers reduces the arrivals to service, which in turn reduces the asymmetry. This can be seen from Table 9, where the the probability to abandon of customers 0 increases in p.

Table 9: Probability of abandonment ($\mu_i = \mu_0 = 0.2$, $\sum_{i=0}^{2} \lambda_i = 8$, $\lambda_i = \lambda_j$, $W_0^* = W_i^* = 0.2$, $\gamma_i = \gamma_0 = \gamma$ for $i, j = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

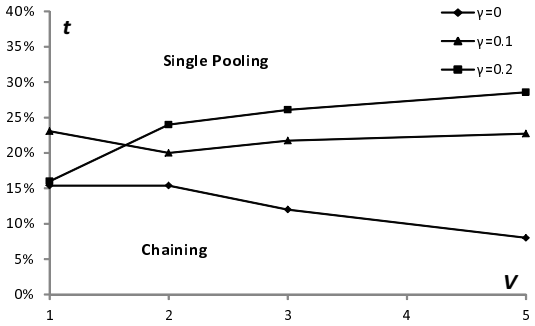
	$\gamma = 0.1$				$\gamma = 0.2$			
	Single Pooling		Chaining		Single Pooling		Chaining	
p	Type i	Type 0	Type i	Type 0	Type i	Type 0	Type i	Type 0
0%	3.04%		1.68%		7.00%		3.20%	
10%	2.05%	0.00%	2.24%	1.95%	5.00%	0.03%	3.58%	3.13%
25%	1.84%	0.01%	1.52%	1.79%	4.30%	0.07%	4.44%	4.24%
50%	1.73%	0.08%	1.69%	2.11%	4.18%	0.40%	3.82%	4.13%
75%	1.70%	0.53%	1.09%	1.27%	4.12%	1.45%	3.80%	3.47%
90%	1.68%	1.14%	0.29%	1.13%	4.10%	2.62%	4.63%	3.33%
100%		1.67%		1.67%		4.25%		4.25%

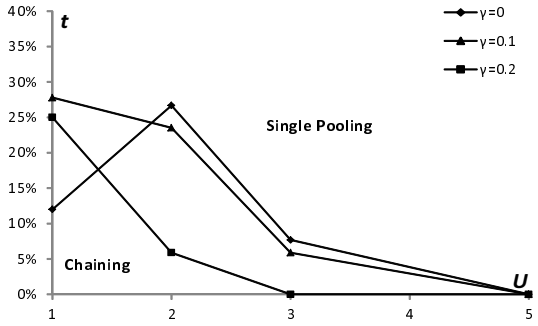
Asymmetry in Abandonment. Consider the asymmetry in the abandonment rates measured by the relative difference between the abandonment rate of customers 0 compared to those of the regular customers. The results are shown in Figures 9(a)-9(d). Further results are also given in Tables 10-13 in Section 3 of the online supplement. We again observe an important impact of the abandonment on the performance of SP and chaining. This impact mainly depends on how the abandonment affects the asymmetry. For example, we observe from Figure 9(a) that when regular customers have higher abandonment rates than customers 0, the asymmetry in terms of





(a) Impact of p ($\mu_i = \mu_0 = 0.2$, $\sum_{i=0}^{2} \lambda_i = 8$, $\lambda_i = \lambda_j$, (b) Impact of p' ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_0^* = W_i^* = 0.2$, for $i, j = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)
$n = 4$)





(c) Impact of V ($\lambda_0 = 2$, $\mu_0 = \mu_i = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, (d) Impact of U ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$, $W_0 = W_i^* = W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$, 0.2, $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $n = 4$)
$V = 1$, $n = 4$)

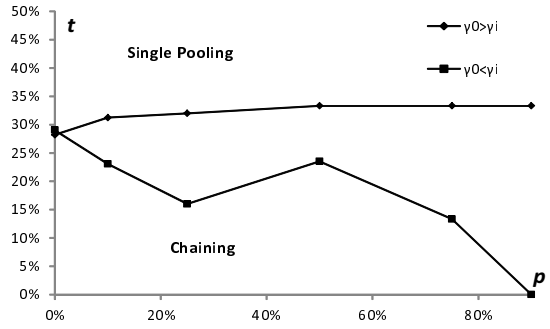
Figure 8: Impact of abandonment

p is accentuated (which further improves SP performance). In the opposite case however, the asymmetry in p reduces because of the abandonment of customers 0.

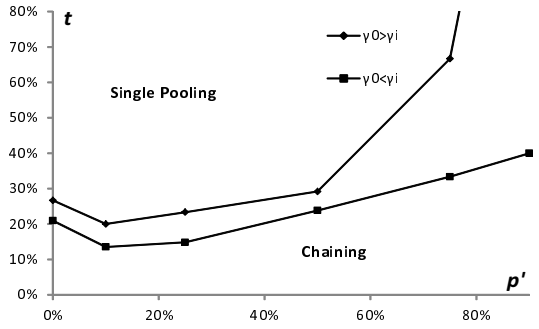
5.5 Impact of the Call Center Size

Akşin and Karaesmen (2007) showed that a small call center benefits more from a flexible architecture than a larger one. From the simulation experiments conducted here, we confirm this conclusion. The results are shown in Table 10 and Figure 10. In Table 11, we provide the achieved expected waiting times for the optimal staffing levels.

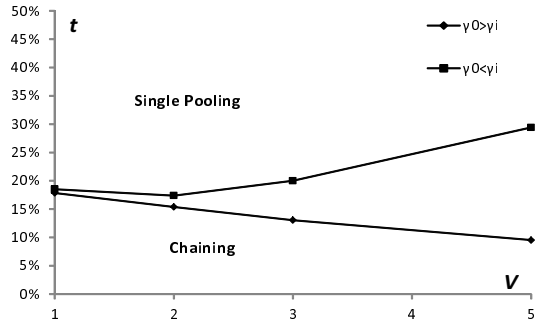
Because of the small teams, the lack of the pooling effect in small call centers makes the threshold values of t higher than those in large call centers. However in large call centers, the team sizes are quite large in the sense that we have a less need to the chains. This makes SP better than chaining even under the symmetric case of arrival rates. From Table 11 we observe that contrary to small call centers, the service level constraints are saturated for large call centers. Because of the discrete nature of staffing levels, the impact of adding or removing an agent on performance is higher in small call centers. For the same reason, the staffing levels of the regular teams do not vary much in small call centers. This makes the relative benefits between SP and chaining not sensitive to



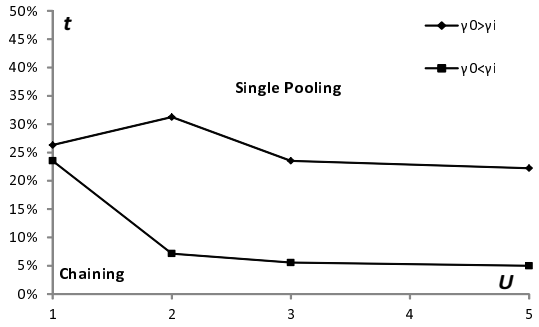
(a) Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)



(b) Impact of p' ($W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $i = 1, ..., 4$, $p = 50\%$, $U = V = 1$, $n = 4$)



(c) Impact of V ($\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\mu_i = \mu_0 = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$, $n = 4$)



(d) Impact of U ($\lambda_0 = 4$, $W_0^* = W_i^* = 0.2$, $\mu_0 = 0.2$, $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $V = 1$, $n = 4$)

Figure 9: Impact of the asymmetry in abandonment

the variation of p in small call centers, while the opposite is true for large call centers (see Figures 10(b) and 10(d)).

5.6 Impact of the Number of Skills

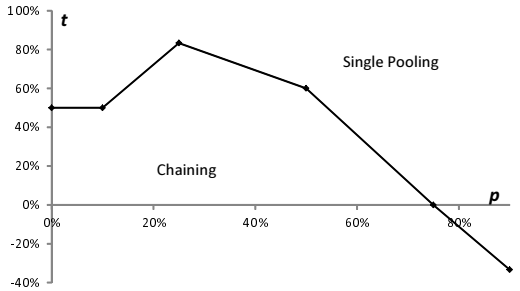
In this section, we investigate the effect of the number of skills (denoted by $N = n + 1$). For two cases with different number of skills, it is not possible to keep at the same time a constant workload on each team and a constant overall workload. We choose to separately treat each situation.

Table 10: Impact of the Call Center Size ($\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$ for $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

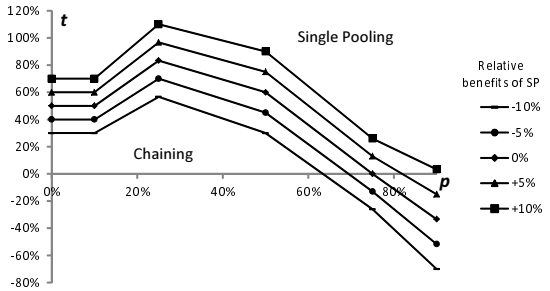
	Small Call Center ($\sum_{i=0}^{4} \lambda_i = 1$)					Large Call Center ($\sum_{i=0}^{4} \lambda_i = 100$)				
	Chaining			SP	Crossing value	Chaining			SP	Crossing value
p	$t = 0\%$	$t = 5\%$	$t = 10\%$		(Chaining = SP)	$t = 0\%$	$t = 5\%$	$t = 10\%$		(Chaining = SP)
0%	12	12.4	12.8	16	$t = 50\%$	513	534.6	556.2	536	$t = 5.32\%$
10%	12	12.4	12.8	16	$t = 50\%$	513	531.15	549.3	518	$t = 1.38\%$
25%	11	11.3	11.6	16	$t = 83.33\%$	513	527.2	541.4	513	$t = 0\%$
50%	12	12.25	12.5	15	$t = 60\%$	513	522.1	531.2	513	$t = 0\%$
75%	13	13.25	13.5	13	$t = 0\%$	515	519.45	523.9	513	$t = -2.25\%$
90%	12	12.15	12.3	11	$t = -33.33\%$	517	519	521	513	$t = -10.00\%$
100%	9	9	9	9	$t = 0\%$	513	513	513	513	$t = 0\%$

Table 11: Expected waiting times ($\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$ for $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

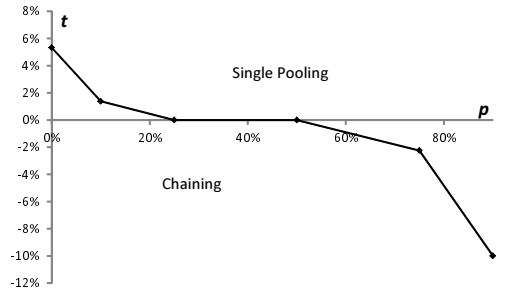
	Small Call Center ($\sum_{i=0}^{4} \lambda_i = 1$)				Large Call Center ($\sum_{i=0}^{4} \lambda_i = 100$)			
	Single Pooling		Chaining		Single Pooling		Chaining	
p	W_i	W_0	W_i	W_0	W_i	W_0	W_i	W_0
0%	0.08		0.06		0.18		0.20	
10%	0.07	0.00	0.06	0.06	0.15	0.20	0.19	0.19
25%	0.05	0.00	0.09	0.08	0.08	0.19	0.19	0.20
50%	0.04	0.00	0.07	0.05	0.05	0.17	0.18	0.20
75%	0.19	0.01	0.07	0.02	0.04	0.20	0.17	0.19
90%	0.19	0.03	0.06	0.05	0.02	0.19	0.15	0.18
100%		0.10		0.10		0.17		0.17



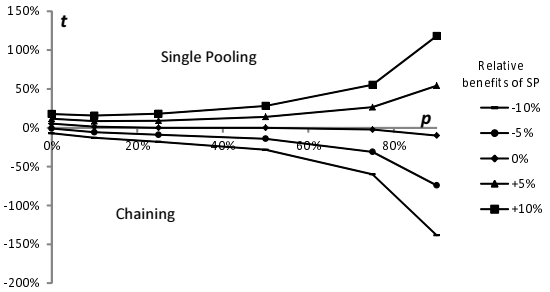
(a) Small Call Center ($\sum_{i=0}^{4} \lambda_i = 1$)



(b) Relative benefits ($\sum_{i=0}^{4} \lambda_i = 1$)



(c) Large Call Center ($\sum_{i=0}^{4} \lambda_i = 100$)



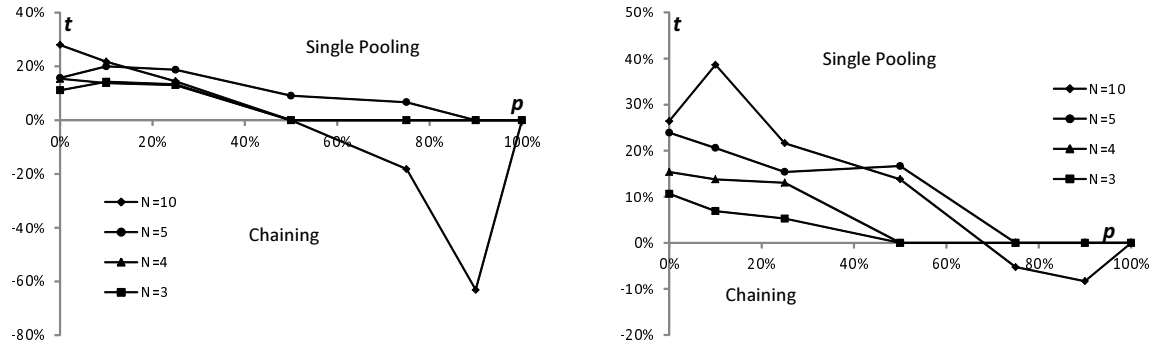
(d) Relative benefits ($\sum_{i=0}^{4} \lambda_i = 100$)

Figure 10: Preference zone ($\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$ for $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

Constant Workload per Team. We consider identical service rates for all customer types. In the experiments below, the ratio $\frac{\sum_{i=0}^{n} \lambda_i}{N}$ is then hold constant. The results are given in Figure 11(a), and in Table 14 of the online supplement. We observe that SP behaves much better than chaining as the number of skills increases. Figure 11(a) shows that for $N = 10$, the crossing value of t should be negative for high values of p (this means that SP is better in all cases). Single pooling behaves much better than chaining for the following two reasons. First as N increases, the flexibility in chaining decreases. A customer type in the chaining configuration has access to a fewer proportion of agent as N increases (the gap with the full flexible model increases). The second reason is related to the impact of the constant ratio $\frac{\sum_{i=0}^{n} \lambda_i}{N}$, which increases the overall size

of the call center as N increases. Having large call centers makes SP more efficient (see Section 5.5).

Constant Overall Workload. We again consider identical service rates for all customer types. The summation $\sum_{i=0}^{n} \lambda_i$ is then hold constant. The results are given in Figure 11(b), and in Table 15 of the online supplement. We distinguish two effects depending on p. For small values of p, the preference zone for SP reduces. The opposite is true for large values of p. The reason is related to the decreasing of the size of each team as N increases. Since we keep constant the overall workload, increasing the number of skills implies a lower demand per skill, which requires less agents per team. This makes the effect of pooling predominant. For the case of large p, the large number of customers 0 benefits from pooling under SP. For the case of small p, the system contains more regular customers, each of which benefits in chaining from the pooling of two adjacent teams.



(a) Preference zone ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{n} \lambda_i/N = 2$, $i = 1, ..., 4$, $U = V = 1$)

(b) Preference zone ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{n} \lambda_i = 8$, $i = 1, ..., 4$, $U = V = 1$)

Figure 11: Preference zone

5.7 More than One Easy Skill

We want to understand the impact of having more than one easy skill, as it would be the case in more complex call centers. Let us consider the case of a call center with Markovian assumptions for inter-arrival and service times. There are two easy skills, denoted by 0 et 0', with arrival rates λ_0 and $\lambda_{0'}$ and service rates μ_0 and $\mu_{0'}$, respectively. We consider the same cost modeling, i.e., an agent with a regular skill and two easy skills costs 1, and an agent with two regular skills costs $1 + t$. All easy customers arrive at queue 0 and are served in the order of their arrival regardless of their type (0 or 0'). One may see that the conclusions drawn so far still hold for the case $\mu_0 = \mu_{0'}$. There is no distinction between the two easy skills that would change the results. However, one may expect that a difference in the service rates of the easy skills may have an impact on the conclusions. Some numerical illustrations are given next.

Since the arrival processes of customers 0 and 0' are Poisson, the probability that a new easy

arrival is type 0 (0') is $\frac{\lambda_0}{\lambda_0+\lambda_{0'}}$ ($\frac{\lambda_{0'}}{\lambda_0+\lambda_{0'}}$). Moreover, easy customers are served under FCFS, then, the service time of an arbitrary easy customer follows a hyperexponential distribution (an exponential distribution with rate μ_0 with probability $\frac{\lambda_0}{\lambda_0+\lambda_{0'}}$, and an exponential distribution with rate with $\mu_{0'}$ with probability $\frac{\lambda_{0'}}{\lambda_0+\lambda_{0'}}$). We thus measure the difference between μ_0 and $\mu_{0'}$ by the coefficient of variation of the hyperexponential distribution, denoted by cv.

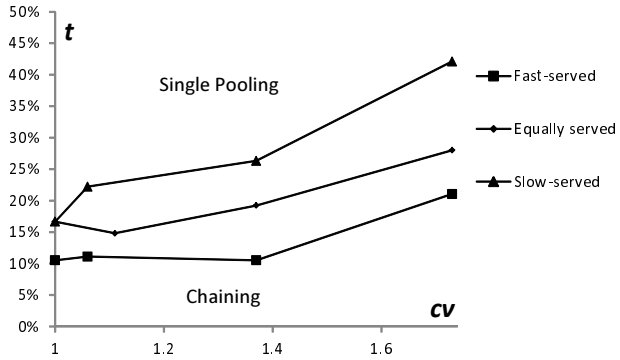


Figure 12: Preference zone ($\lambda_0 = \lambda_{0'} = 2$, $\mu_i = 0.2$, $\lambda_i = 1$, $W_0^* = W_{0'}^* = W_i^* = 0.2$, for $i = 1, ..., 4$, $U = 1$, $V = 1$, $n = 4$)

Table 12: Cost comparison with two easy skills ($\lambda_0 = \lambda_{0'} = 2$, $\mu_i = 0.2$, $\lambda_i = 1$, $W_0^* = W_{0'}^* = W_i^* = 0.2$, for $i = 1, ..., 4$, $U = 1$, $V = 1$, $n = 4$)

			Chaining				SP	Crossing value
μ_0	μ_0'	cv	t=0%	t=5%	t=10%	t=25%		(Chaining = SP)
1	1	1	33	33.95	34.9	37.75	35	t =10.53%
0.8	1.33	1.06	33	33.9	34.8	37.5	35	t =11.11%
0.6	3	1.37	33	33.95	34.9	37.75	35	t =10.53%
0.5	∞	1.73	32	32.95	33.9	36.75	36	t =21.05%
0.2	0.2	1	49	49.9	50.8	53.5	52	t =16.67%
0.15	0.3	1.11	49	50.35	51.7	55.75	53	t =14.81%
0.12	0.6	1.37	49	50.3	51.6	55.5	54	t =19.23%
0.1	∞	1.73	49	50.25	51.5	55.25	56	t =28.00%
0.1	0.1	1	73	73.9	74.8	77.5	76	t =16.67%
0.08	0.13	1.06	73	73.9	74.8	77.5	77	t =22.22%
0.06	0.3	1.37	73	73.95	74.9	77.75	78	t =26.32%
0.05	∞	1.73	72	72.95	73.9	76.75	80	t =42.11%

In Figure 12 and Table 12, we compare between SP and chaining under various scenarios with two easy skills. We consider a call center case with 6 customer types, 2 easy and 4 regular skills. In the experiments, we vary the cv of the service time distribution of an easy customer (or equivalently the difference between μ_0 and $\mu_{0'}$). For a coherent comparison, the quantity $\frac{1}{\mu_0} + \frac{1}{\mu_{0'}}$ is kept constant. There are three parts in the experiments. In the first part, an arbitrary easy customer is in average served faster than a regular one. In the second part, the average service durations for easy and regular customers are the same. In the last part, an arbitrary easy customer is in average served slower than a regular one.

We observe that SP is not preferred for large differences between the service rates of the easy skills (high cv). This is especially apparent when the easy customers are served slower than the regular ones. The explanations of these observations are again related to the blocking effect, as explained for the analysis of the parameters p' and cv_s in Section 5.2.1. One may then conclude that an asymmetry in the service of the easy customers deteriorates the performance of SP.

5.8 Impact of the Agent Costs

We change the cost framework such that the cost of regular agents are no longer identical. This allows to have situations where some agents may be much more expensive than others. We examine then the impact of an asymmetry in the costs on the comparison between the staffing costs of SP and chaining. The detailed experiments are given in Section 5 of the online supplement.

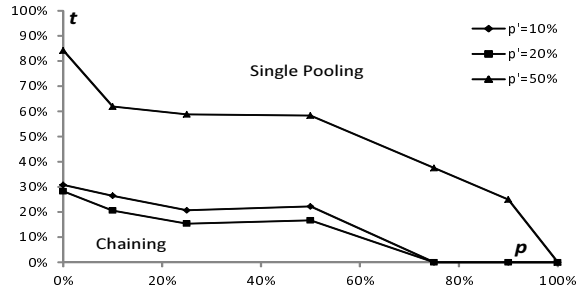
The main conclusions from this study is that the asymmetry in the agent costs has not a significant impact on the team staffing levels and costs in SP. Because of the lack of pooling for the regular customers, SP does not have enough flexibility to act on the team staffing levels so as to reduce the overall costs. However, the impact of the agent costs on the performance of chaining is important. Under asymmetrical situations of arrival or service rates, chaining prefers an asymmetrical cost framework. This allows to have large and cheap teams. The opposite is true under symmetrical situations of arrival or service rates.

5.9 Mix of Asymmetry

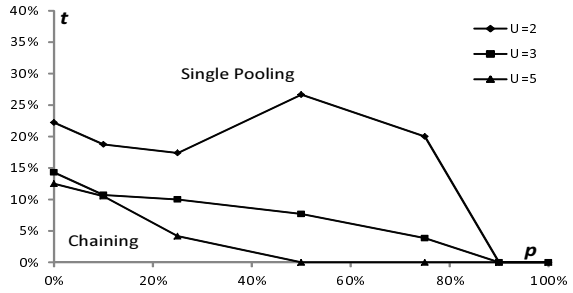
In this section we mix the effects of more than a parameter at a time. We propose to interact the effects of p and p', U and p, U and p', U and V, and also all of them. The results are presented in Tables 13-17 and Figures 13(a)-13(d).

From the numerical results, we observe that the individual effects are still present, but they may accumulate or make up for one another. One important observation is that two asymmetries may lead to a bad performance for SP. For example SP behaves well in each one of the asymmetric situations ($U = 2$ and $V = 1$) and ($U = 1$ and $V = 1/3$) in isolation. However, it does not behave well for the mixed situation ($U = 2$ and $V = 1/3$). In such a situation, the customers types with large arrival rates are the faster to be served, and viceversa. Therefore, the different customer types workloads are likely to be symmetric. For the same reason, SP behaves well in the situation ($U = 2$ and $V = 3$) because the mix of asymmetries further accentuates the asymmetry in workloads.

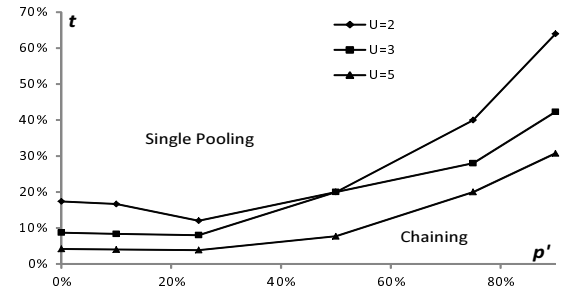
Tables 14 and 15 reveal also that the most predominant effects are those of p (because of pooling) and p' (because of blocking). Various scenarios of mixed asymmetries are considered in Table 17. We find again that SP behaves well in large call centers (the first four scenarios). Scenarios 3 and 7 are similar in terms of the values of p and p' (high values for the two parameters). This means that



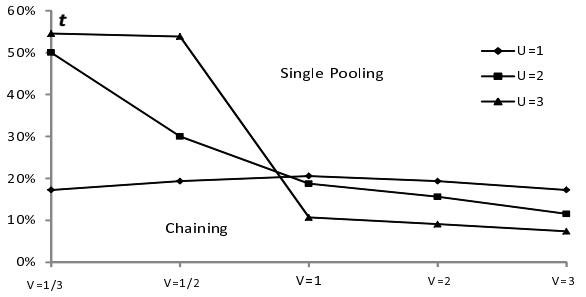
(a) Impact of p and p' ($\mu_i = \mu_j$ and $\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$,$\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $U = V = 1$)



(b) Impact of U and p ($\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$,$\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p' = 20\%$, $V = 1$)



(c) Impact of U and p' ($\lambda_i = \lambda_j = 1.5$ for $i, j = 1, ..., 4$, $\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p = 25\%$, $V = 1$)



(d) Impact of U and V ($\lambda_0 = 0.8$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ and $\sum_{i=0}^{4} \lambda_i = 8$, $p = 10\%$, $p' = 20\%$)

Figure 13: Preference zone

Table 13: Impact of p and p' ($\mu_i = \mu_j$ and $\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$,$\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $U = V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$p' = 10\%$	0%	52	53.95	55.9	61.75	71.5	64	t=30.77%
	10%	51	52.7	54.4	59.5	68	60	t=26.47%
	25%	46	47.45	48.9	53.25	60.5	52	t=20.69%
	50%	39	39.9	40.8	43.5	48	43	t=22.22%
	75%	35	35.6	36.2	38	41	35	t=0.00%
	90%	31	31.3	31.6	32.5	34	31	t=0.00%
	100%	24	24	24	24	24	24	t=0.00%
$p' = 20\%$	0%	49	50.95	52.9	58.75	68.5	60	t=28.21%
	10%	49	50.7	52.4	57.5	66	56	t=20.58%
	25%	48	49.3	50.6	54.5	61	52	t=15.38%
	50%	49	49.9	50.8	53.5	58	52	t=16.67%
	75%	51	51.55	52.1	53.75	56.5	51	t=0.00%
	90%	51	51.3	51.6	52.5	54	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$p' = 50\%$	0%	24	24.95	25.9	28.75	33.5	40	t=84.21%
	10%	40	41.05	42.1	45.25	50.5	53	t=61.90%
	25%	53	53.85	54.7	57.25	61.5	63	t=58.82%
	50%	75	75.6	76.2	78	81	82	t=58.33%
	75%	97	97.4	97.8	99	101	100	t=37.50%
	90%	111	111.2	111.4	112	113	112	t=25.00%
	100%	112	112	112	112	112	112	t=0.00%

the effect of pooling and blocking are highly present in both scenarios. An important observation here is that scenario 3 is the best among scenarios 1-4, while scenario 7 is the worst among scenarios

Table 14: Impact of p and U ($\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$,$\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p' = 20\%$, $V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$U = 2$	0%	49	50.8	52.6	58	67	57	t=22.22%
	10%	49	50.6	52.2	57	65	55	t=18.75%
	25%	49	50.15	51.3	54.75	60.5	53	t=17.39%
	50%	49	49.75	50.5	52.75	56.5	53	t=26.67%
	75%	51	51.5	52	53.5	56	53	t=20.00%
	90%	53	53.35	53.7	54.75	56.5	53	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$U = 3$	0%	51	52.4	53.8	58	65	55	t=14.29%
	10%	50	51.4	52.8	57	64	53	t=10.71%
	25%	50	51.5	53	57.5	65	53	t=10.00%
	50%	50	51.3	52.6	56.5	63	52	t=7.69%
	75%	51	52.3	53.6	57.5	64	52	t=3.85%
	90%	52	53.35	54.7	58.75	65.5	52	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$U = 5$	0%	52	53.2	54.4	58	64	55	t=12.50%
	10%	51	51.95	52.9	55.75	60.5	53	t=10.53%
	25%	52	53.2	54.4	58	64	53	t=4.17%
	50%	52	52.05	52.1	52.25	52.5	52	t=0.00%
	75%	52	52.05	52.1	52.25	52.5	52	t=0.00%
	90%	52	52.15	52.3	52.75	53.5	52	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%

5-8. This gives an indication on the direct competition between the effects of p and p'. In large call centers, the pooling effect created by customers 0 is predominant over the blocking effect, and the opposite is true in small call centers.

6 Concluding Remarks

We focused on a fundamental problem in the design and management of SBR call centers, for which it is important to choose a flexible architecture. We considered the context of call centers with unbalanced workload, different service requirements, a predominant customer type and high costs of cross-training. With these asymmetry in the parameters, the well-known existing architectures such as chaining lose their robustness. We proposed the new call center architecture single pooling and demonstrated its efficiency. SP allows to balance the workload among the agents in a way that captures the benefits of pooling, without requiring every agent to process every type of call.

The numerical analysis showed that single pooling performs better than chaining for various cases of asymmetry. In the case of a predominance of customers 0 and/or an important asymmetry in the arrival rates of the regular types (captured by V), SP is more robust than chaining even for small differences between the costs of a regular skill and that of skill 0. Because of the blocking effect, the performance of both chaining and SP deteriorates in the asymmetry defined by the service time duration of customers 0 relatively to that of regular customers. This is more apparent in single pooling because customers 0 have access to all teams, while in chaining they do only

Table 15: Impact of p' and U ($\lambda_i = \lambda_j = 1.5$ for $i, j = 1, ..., 4$, $\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p = 25\%$, $V = 1$)

		Chaining					SP	Crossing value
	p'	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$U = 2$	0%	51	52.15	53.3	56.75	62.5	55	t=17.39%
	10%	50	51.2	52.4	56	62	54	t=16.67%
	25%	51	52.25	53.5	57.25	63.5	54	t=12.00%
	50%	51	52.25	53.5	57.25	63.5	56	t=20.00%
	75%	52	53.25	54.5	58.25	64.5	62	t=40.00%
	90%	52	53.25	54.5	58.25	64.5	68	t=64.00%
$U = 3$	0%	52	53.15	54.3	57.75	63.5	54	t=8.70%
	10%	51	52.2	53.4	57	63	53	t=8.33%
	25%	51	52.25	53.5	57.25	63.5	53	t=8.00%
	50%	51	52.25	53.5	57.25	63.5	56	t=20.00%
	75%	54	55.25	56.5	60.25	66.5	61	t=28.00%
	90%	56	57.3	58.6	62.5	69	67	t=42.31%
$U = 5$	0%	52	53.2	54.4	58	64	53	t=4.17%
	10%	51	52.25	53.5	57.25	63.5	52	t=4.00%
	25%	51	52.3	53.6	57.5	64	52	t=3.85%
	50%	52	53.3	54.6	58.5	65	54	t=7.69%
	75%	55	56.25	57.5	61.25	67.5	60	t=20.00%
	90%	58	59.3	60.6	64.5	71	66	t=30.77%

have access to two teams. We have also observed that SP is more robust than chaining against an increasing asymmetry between the service times of regular types. Since the teams under SP are less inter-dependent than under chaining, SP is again preferred in the case of an asymmetry between the objective service levels. We therefore avoid over-staffing situations that may happen in chaining.

From this study one may summarize the recommendations and guidelines to call center managers as follows. The manager choice of a flexible call center design should be single pooling under situations of asymmetry in arrival and service rates. This holds even for small differences between the skill costs. This choice more apparently prevails for large call centers and/or in the case of a high number of skills. However, the choice of the design is highly impacted in the context of call centers with customer abandonment. Abandonments may affect the system by either increasing or decreasing the asymmetry of the parameters. In the first case, the preference remains for single pooling, while it is for chaining in the second case.

In a future research, it would be useful to extend the numerical approximations, of the performance of SP and chaining, in the case of customer abandonment or non-Markovian assumptions. Another interesting work is to generalize the functioning of single pooling in order to avoid the blocking effect in the case of long service times for customers 0.

Table 16: Impact of U and V ($\lambda_0 = 0.8$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ and $\sum_{i=0}^{4} \lambda_i = 8$, $p = 10\%$, $p' = 20\%$)

U	V	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
1	1/3	50	51.45	52.9	57.25	64.5	55	t=17.24%
	1/2	50	51.55	53.1	57.75	65.5	56	t=19.35%
	1	49	50.7	52.4	57.5	66	56	t=20.58%
	2	50	51.55	53.1	57.75	65.5	56	t=19.35%
	3	50	51.45	52.9	57.25	64.5	55	t=17.24%
2	1/3	26	26.8	27.6	30	34	34	t=50.00%
	1/2	31	32	33	36	41	37	t=30.00%
	1	49	50.6	52.2	57	65	55	t=18.75%
	2	71	72.6	74.2	79	87	76	t=15.63%
	3	81	82.3	83.6	87.5	94	84	t=11.54%
3	1/3	20	20.55	21.1	22.75	25.5	26	t=54.55%
	1/2	24	24.65	25.3	27.25	30.5	31	t=53.85%
	1	50	51.4	52.8	57	64	53	t=10.71%
	2	79	80.65	82.3	87.25	95.5	82	t=9.09%
	3	93	94.35	95.7	99.75	106.5	95	t=7.41%

Table 17: Impact of p, p', U and V ($W_i^* = 0.2$ for $i = 0, \cdots 4$)

	Scenarios λ_1	λ_2	λ_3	λ_4	λ_0	μ_1	μ_2	μ_3	μ_4	μ_0	Chaining $t = 0\%$	$t = 10\%$	$t = 20\%$	SP	Crossing value (Chaining=SP)
Sc 1	1	2	3	4	5	0.05	0.1	0.2	0.5	1	78	83.5	89	87	16.36%
Sc 2	2	3	4	5	1	0.05	0.1	0.2	0.5	1	115	122.6	130.2	127	15.79%
Sc 3	1	2	3	4	5	1	0.5	0.2	0.1	0.05	179	184.9	190.8	184	8.47%
Sc 4	2	3	4	5	1	1	0.5	0.2	0.1	0.05	111	116.9	122.8	119	13.56%
Sc 5	0.1	0.2	0.3	0.4	0.5	0.05	0.1	0.2	0.5	1	14	15	16	18	40.00%
Sc 6	0.2	0.3	0.4	0.5	0.1	0.05	0.1	0.2	0.5	1	18	19.4	20.8	24	42.86%
Sc 7	0.1	0.2	0.3	0.4	0.5	1	0.5	0.2	0.1	0.05	26	26.7	27.4	30	57.14%
Sc 8	0.2	0.3	0.4	0.5	0.1	1	0.5	0.2	0.1	0.05	18	18.7	19.4	21	42.86%

References

Akşin, O., Armony, M., and Mehrotra, V. (2007). The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16:665–688.

Akşin, O. Z. and Karaesmen, F. (2007). Characterizing the performance of process flexibility structures. *Operations Research Letters*, 35:477–484.

Albino, V. and Garavelli, A. (1999). Limited flexibility in cellular manufacturing systems: A simulation study. *International Journal of Production Economics*, 60-61:447–455.

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. *Operations Research*, 52:17–34.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association*, 100:36–50.

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

Garavelli, A. (2001). Performance analysis of a batch production system with limited flexibility. *International Journal of Production Economics*, 69(1):39 – 48.

Garavelli, A. (2003). Flexibility configurations for the supply chain management. *International Journal of Production Economics*, 85(2):141 – 153.

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

Gurumurthi, S. and Benjaafar, S. (2004). Modeling and Analysis of Flexible Queueing Systems. *Naval Research Logistics*, 51:755–782.

Hopp, W., Tekin, E., and van Oyen, M. (2004). Benefits of Skill Chaining in Production Lines with Cross-Trained Workers. *Management Science*, 50:83–98.

Hopp, W. and van Oyen, M. (2004). Agile Workforce Evaluation: A Framework for Cross-Training and Coordination. *IIE Transactions*, 36:919–940.

Jordan, W. and Graves, S. (1995). Principles on the Benefits of Manufacturing Process Flexibility. *Management Science*, 41:577–594.

Jordan, W., Inman, R., and Blumenfeld, D. (2004). Chained Cross-Training of Workers for Robust Performance. *IIE Transactions*, 36:953–967.

Koole, G., Nielson, B., and Nielson, T. (2012). First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60:1258–1266.

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

Marengo, W. (2004). Skill based routing in multi-skill call center. Working Paper. Vrije universiteit, The Netherlands.

Nomden, G. and van der Zee, D. (2008). Virtual cellular manufacturing: Configuring routing flexibility. *International Journal of Production Economics*, 112(1):439–451.

Robbins, T. and Harrison, T. (2010). Cross Training in Call Centers with Uncertain Arrivals and Global Service Level Agreements. *International Journal of Operations and Quantitative Management*, 16:307–329.

Sheikhzadeh, M., Benjaafar, S., and Gupta, D. (1998). Machine Sharing in Manufacturing Systems: Total Flexibility versus Chaining. *International Journal of Flexible Manufacturing Systems*, 10:351–378.

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 van Oyen, M. (2009). Pooling Strategies for Call Center Agent Cross-Training. *IIE Transactions*, 41:546–561.

Tomlin, B. and Wang, Y. (2005). On the Value of Mix Flexibility and Dual Sourcing in Unreliable Newsvendor Networks. *Manufacturing & Service Operations Management*, 7:37–57.

van Dijk, N. and van der Sluis, E. (2008). To Pull or not to Pull in Call Centers. *Production and Operations Management*, 17:1–10.

Wallace, R. and Whitt, W. (2005). A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*, 7:276–294.

Appendix

This appendix is related to Section 4. We provide the method to compute the steady-state probabilities in single pooling and chaining.

Single Pooling. The equilibrium equation related to the tuple x is

$$
\begin{aligned}
&\left(\sum_{k=0}^{n}\lambda_k + x_k\mu_k + x_{0,k}\mu_0\right)\pi_{x_0,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_n} \qquad (2)\\
&= \sum_{k=1}^{n}\lambda_k \mathbf{1}_{(0\le x_k+x_{0,k}-1<s_k)}\pi_{x_0,x_1,x_2,\cdots,x_{k-1},x_k-1,x_{k+1},\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=1}^{n}\lambda_k \mathbf{1}_{(x_k+x_{0,k}=s_k,q_k-1\ge 0)}\pi_{x_0,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_{k-1},q_k-1,q_{k+1},\cdots q_n}\\
&+ \lambda_0 \mathbf{1}_{(0\le x_0-1<s_0)}\pi_{x_0-1,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_n}\\
&+ \lambda_0 \mathbf{1}_{(x_0=s_0,x_1+x_{0,1}=s_1,\cdots x_n+x_{0,n}=s_n,q_0-1\ge 0)}\pi_{x_0,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0-1,q_1,\cdots,q_n}\\
&+ \lambda_0 \mathbf{1}_{(x_0=s_0)}\sum_{k=1}^{n}\mathbf{1}_{\left(\frac{x_k+x_{0,k}-1}{s_k}<1\right)}\mathbf{1}_{\left(\frac{x_k+x_{0,k}-1}{s_k}<\frac{x_j+x_{0,j}}{s_j}\ \text{for}\ j\ne k\right)}\pi_{x_0,\cdots,x_n,x_{0,1},\cdots,x_{0,k-1},x_{0,k}-1,x_{0,k+1},\cdots,x_{0,n},q_0,\cdots,q_n}\\
&+ \cdots + \frac{\lambda_0}{n}\mathbf{1}_{(x_0=s_0)}\sum_{k=1}^{n}\mathbf{1}_{\left(\frac{x_k+x_{0,k}-1}{s_k}<1\right)}\mathbf{1}_{\left(\frac{x_k+x_{0,k}-1}{s_k}=\frac{x_j+x_{0,j}}{s_j}\ \text{for}\ j\ne k\right)}\pi_{x_0,\cdots,x_n,x_{0,1},\cdots,x_{0,k-1},x_{0,k}-1,x_{0,k+1},\cdots,x_{0,n},q_0,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\mathbf{1}_{(q_k=0)}\mu_k(x_k+1)\pi_{x_0,x_1,x_2,\cdots,x_{k-1},x_k+1,x_{k+1},\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=1}^{n}\mathbf{1}_{(q_k=0)}\mu_0(x_{0,k}+1)\pi_{x_0,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,k-1},x_{0,k}+1,x_{0,k+1},\cdots,x_{0,n},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\mathbf{1}_{(x_k+x_{0,k}=s_k)}(\mu_k x_k+\mu_0 x_{0,k})\pi_{x_0,x_1,x_2,\cdots,x_n,x_{0,1},x_{0,2},\cdots,x_{0,n},q_0,q_1,\cdots,q_{k-1},q_k+1,q_{k+1},\cdots q_n},
\end{aligned}
$$

with the convention $x_{0,0} = 0$. We next add the normalization condition and numerically solve the obtained system of equations relating the steady-state probabilities. We use a finite state space approximation in order to obtain a system with a finite number of equations. This consists of truncating the number of states by assuming that each queue has a finite capacity D. In the numerical experiments, we choose the smallest value of D such that beyond this value the expected waiting time does not vary with a sufficiently high precision (six digits beyond the decimal point).

Chaining. The equilibrium equation related to the vector $(x_{0,1}, \cdots, x_{n,0}, x_{0,0}, q_0, q_1, \cdots, q_n)$ is

$$
\begin{aligned}
&\left(\sum_{k=0}^{n}(\lambda_k + \mu_k(x_{k,k+1} + x_{k,k})) + \zeta\left(1 - \prod_{k=0}^{n}\mathbf{1}_{(q_k=0)}\right)\right)\pi_{x_{0,1},\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n} \qquad (3)\\
&= \sum_{k=0}^{n}\lambda_k \mathbf{1}_{(0\le x_{k,k}+x_{k-1,k}-1<s_k)}\mathbf{1}_{(x_{k,k}+x_{k-1,k}-1<x_{k+1,k+1}+x_{k,k+1})}\pi_{x_{0,1},\cdots,x_{k,k}-1,\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\lambda_k \mathbf{1}_{(0\le x_{k+1,k+1}+x_{k,k+1}-1<s_{k+1})}\mathbf{1}_{(x_{k+1,k+1}+x_{k,k+1}-1<x_{k,k}+x_{k-1,k})}\pi_{x_{0,1},\cdots,x_{k,k+1}-1,\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\frac{\lambda_k}{2} \mathbf{1}_{(0\le x_{k,k}+x_{k-1,k}-1<s_k)}\mathbf{1}_{(x_{k,k}+x_{k-1,k}-1=x_{k+1,k+1}+x_{k,k+1})}\pi_{x_{0,1},\cdots,x_{k,k}-1,\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\frac{\lambda_k}{2} \mathbf{1}_{(0\le x_{k+1,k+1}+x_{k,k+1}-1<s_{k+1})}\mathbf{1}_{(x_{k+1,k+1}+x_{k,k+1}-1=x_{k,k}+x_{k-1,k})}\pi_{x_{0,1},\cdots,x_{k,k+1}-1,\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\lambda_k \mathbf{1}_{(q_k-1=0)}\mathbf{1}_{(x_{k,k}+x_{k-1,k}=s_k)}\mathbf{1}_{(x_{k+1,k+1}+x_{k,k+1}=s_{k+1})}\pi_{x_{0,1},\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_k-1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\zeta \mathbf{1}_{(q_k-1>0)}\mathbf{1}_{\left(q_j=0 \text{ for } j\neq k\right)}\pi_{x_{0,1},\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_k-1,\cdots,q_n}\\
&+ \cdots + \zeta\prod_{k=0}^{n}\left(\mathbf{1}_{(q_k-1>0)}\right)\pi_{x_{0,1},\cdots,x_{n,0},x_{0,0},q_0-1,q_1-1,\cdots,q_n-1}\\
&+ \sum_{k=0}^{n}\mathbf{1}_{(q_k=0)}\mu_k(x_{k,k}+1)\pi_{x_{0,1},\cdots,x_{k,k}+1,\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\mathbf{1}_{(q_k=0)}\mu_k(x_{k,k+1}+1)\pi_{x_{0,1},\cdots,x_{k,k+1}+1,\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_n}\\
&+ \sum_{k=0}^{n}\mathbf{1}_{(x_{k-1,k}+x_{k,k}=s_k)}\mathbf{1}_{(x_{k,k+1}+x_{k+1,k+1}=s_{k+1})}\left(\mu_k x_{k,k} + \mu_{k-1,k}x_{k-1,k}\right)\\
&\times \sum_{i,j=0}^{\infty}\mathbf{1}_{(q_k+j>q_{k-1}+i)}p_{q_k+j,q_k}\pi_{x_{0,1},\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_k+j,\cdots,q_n}\\
&+ \frac{1}{2}\sum_{k=0}^{n}\mathbf{1}_{(x_{k-1,k}+x_{k,k}=s_k)}\mathbf{1}_{(x_{k,k+1}+x_{k+1,k+1}=s_{k+1})}\left(\mu_k x_{k,k} + \mu_{k-1,k}x_{k-1,k}\right)\\
&\times \sum_{i,j=0}^{\infty}\mathbf{1}_{(q_k+j=q_{k-1}+i)}p_{q_k+j,q_k}\pi_{x_{0,1},\cdots,x_{n,0},x_{0,0},q_0,q_1,\cdots,q_k+j,\cdots,q_n},
\end{aligned}
$$

with the convention $x_{-1,0} = x_{n,n+1} = x_{n,0}$. Again, we add the normalization condition and numerically solve the obtained system of equations using truncation with parameter D. The value of ζ has a significant impact on the approximation. Increasing it allows to better model the continuous elapsing time. However, this increases the number of used states in the Markov chain, which would require to increase the truncation threshold D. Again in the numerical experiments, we choose the values of ζ and D such that for better combinations of values (with higher computational efforts),

the expected waiting time does not vary with a sufficiently high precision (six digits beyond the decimal point).