A New Flexible Architecture for Call Centers with Skill-Based Routing

Benjamin Legros, Oualid Jouini, Yves Dallery

9^{th} *International Conference of Modeling, Optimization and Simulation - MOSIM'12*
June 6-8, 2012 - Bordeaux - France
"Performance, interoperability and safety for sustainable development"

A New Flexible Architecture for Call Centers with Skill-Based Routing

B. LEGROS, O. JOUINI, Y. DALLERY
Laboratoire de 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

ABSTRACT: *We consider flexible architectures with multi-skills agents, revisiting the reference model of Chaining we propose a new architecture : the Single Pooling model. Using simulation and developing analytical models we show that this model can be almost as performing as the chaining model but also less expensive. The purpose of the article is to present insights for the manager in its strategy of designing a call center. The impact of most of the parameters implied in a call center like arrival rate, service rate, variation, number of teams, size of the call center, workload or quality of service will be studied.*

KEYWORDS: *Call centers; skill-based routing; flexibility; optimization; performance measures; staffing*

1 INTRODUCTION

Telephone call centers are an integral part of many businesses. Their economic role is significant and growing. Flexibility of the resources is a way to reduce global costs. Organizing an efficient distribution of the skills in a flexible architecture of call center is an important issue for managers. In this paper, we identify the key characteristics that enable agility through cross-trained agents, which allows to improve the call center operations management. We develop an innovative organizational architecture. We also conduct a comprehensive comparative study in order to prove its efficiency. The full cross-training of every agent for every call types is the most efficient flexible architecture but also the most costly. To obtain a given quality of service the Full Flexible (FF) model will require less agents than any other architectures, but these agents will be too costly and sometimes impossible to find. The Full Dedicated (FD) model in which every agent has only one skill will require more agents to achieve a high quality of service than any other architectures. In this architecture, the agents are less costly but they will be less efficient and vacant during longer periods of time. Then the FD model might not be the best proposition to achieve a good quality of service. The literature proposes different architectures between the FF and the FD models like chaining or pairing. Jordan and al. (1995) studies showed that chaining, where each call can be routed to one of two adjacent servers and each server can process calls from two adjacent classes (see figure 2), has the potential to achieve most of the benefits of pooling with respect to performance measures such as the expected time spent in the system and throughput. We propose in this article to present a new flexible architecture of call center.

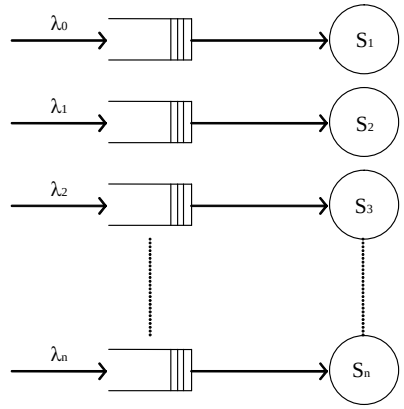
2 RESEARCH OBJECTIVE

In this paper we study call centers that can treat $n+1$ different call types, namely call types $0, 1, 2, \cdots n$. For $i = 1, 2, ..., n$, calls i are named regular calls. Except in section 2.2 calls arrive randomly with an exponential inter-arrival time, the arrival rate of arriving calls is named λ_i for $i = 0, 1, ..., n$. The service will also follow an exponential law (except in section2.2). We suppose that the service rate only depend on the call type (not on the agent) then calls of type i will be serve with a service rate of μ_i for $i = 0, 1, ..., n$. We suppose in our models that there is no abandonment and that the queues are infinite. We also use the parameter $\rho_i = \frac{\lambda_i}{\mu_i}$ to calculate the minimal number of agents $(E(\rho_i) + 1)$ needed to have stability on call types i. The overall arrival rate is noted Λ $(\Lambda = \lambda_0 + \lambda_1 + \cdots + \lambda_n)$. The call centers are organized in homogenous teams. A team k has S_k agents who perform in a limited number of skills. We use the letter S for the overall number of agents in the call center and N for the number of teams. The average waiting time of calls i is noted W_i and P_i is the waiting probability. We choose to over line a element to suggest an average parameter like $\overline{W}$ for the average waiting time in the call center or $\overline{\mu}$ for the average service rate. The letter Ω represents the ratio $\frac{\Lambda}{\overline{\mu}}$. A

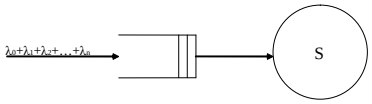
required quality of service (QoS) is noted with a star. For example $W_i^* = 0.2$ means that we try to reach a quality of service of an average waiting time below 0.2 for calls i. The costs are supposed to be proportional to the number of agents. Because the teams are homogenous, every agent has the same cost C_k in the team k. The global cost of the call center is then supposed to be $\sum_{k=1}^{N} C_k \times S_k$. We suppose that the more skills an agent has the more costly she will be and that every skill doesn't have the same cost. We suppose that the skill 0 is the less expensive one.

2.1 Motivation

Figure 1(a) represents the FD model and figure 1(b) represents the FF one for a call center that receives $n+1$ different call types. Our problem is how can



(a) The FD model



(b) The FF model

Figure 1: The references

we create an architecture of call center that can have almost as few agents as FF model and the costs per agent closed to the FD model ? First, in order to limit individual costs we choose to limit as much as possible the number of skills per agent. Jordan and al. (1995) showed that with only two skills per agent in a configuration called chaining (see figure 2) we can reach the performances of the full flexible model. The chaining architecture is a reference not only for call centers' architecture but also in production lines. Could we propose a cheaper architecture? As Gurumurthy and al. (2004) studied a symmetric architecture is performing when the demand is symmetric. When the asymmetry in demand increases the chaining performances decreases. That is why we present a new model; the Single Pooling (Figure 3). In this model calls 0 benefit from a complete pooling. Every team of agent has skill 0 and a regular skill i for $i = 1, 2, ...n$. What is the added value of this architecture ? First, this architecture could support an important asymmetry in favor of calls 0 because they benefit from pooling. Second, if a skill is easy to find then usually this skill is less costly than the others. A cross trained agent to skill 0 and i is less costly than an agent trained into two regular skills i and j. This architecture could be adaptable in many cases. For example, in a airplane company, when the skills are languages, the English language is more common and then easier to find. We compare the Single Pooling model with the Chaining model which is the main reference in architectures. In a call center the parameters are numerous. We explore the impact of all of them; the arrival rates, the service rates, the variation, the asymmetry, the workload, the quality of service, the size of the call center, the number of teams and the costs. The comparison are made trough simulations and simplified analytical models. To go further into the comparison, we build easy usable tools to help the manager in the decision of creating a call center architecture. Figure 2 presents the chaining architecture and Figure 3 presents the Single Pooling architecture. The

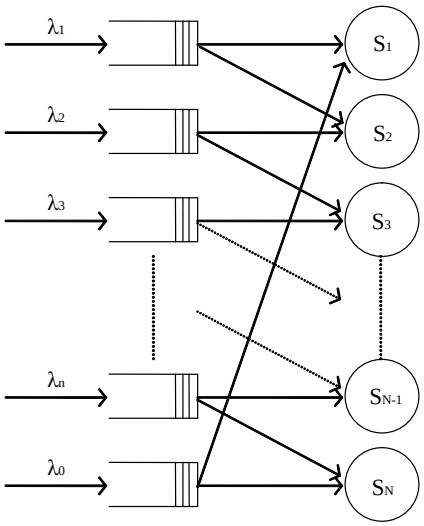


Figure 2: The Chaining architecture

routing rules in the chaining architecture treats calls in equivalence. In deed, when a call is entering the system then she is routed in priority to the team that has the bigger proportion of idle agents, if all the agents able to serve this call are busy then the call has to wait until an agent get free. When an agent has finished a service, priority is given to the call that has waited the longest time in the queue. The routing rules for Single Pooling are inspired by Borst and

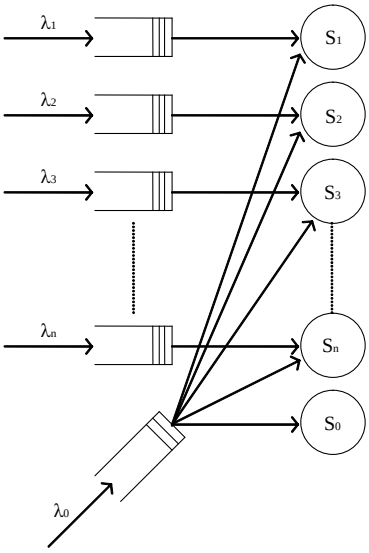


Figure 3: The Single Pooling architecture

al. (2004). A call type i has a strict priority over call type 0 because this call does not benefit from any pooling effect. A call 0 is routed in priority to team 0 because of the specialist first principle. If team 0 is busy, call 0 is routed to the team that has the bigger proportion of idle agents.

2.2 How to compare the models ?

We will compare the models in term of performance, Which performance will we choose ? There are different way to measure performance in a call center. The performance measures are always bound to an objective. As Zohar and al. (2002) have seen the operational performance measures are mostly inter correlated and the measure of one can inform on the others. We choose to limit the measure of performance to the costs and the service. How many agents are necessary to achieve a quality of service of an average waiting time lower than W_i^* per calls type ? The optimization problem is minimizing $\sum C_k \times S_k$ while respecting the constraints $W_i < W_i^*$ for $i = 0, 1, ..., n$. This formalization avoid the problem of the average waiting time ($\overline{W}$) that disadvantages the small group of calls or the problem of multi objective optimization, because average waiting time, abandon rate and waiting probability are positively correlated. In the perspective of building an architecture, the first question for a manager is how many agents do I need ? or how much will it cost ? The waiting time is seen as a quality of service to achieve.

3 Effect of Asymmetry of the Parameters

In this section, we compare between the two models, Chaining and SP. We investigate on the impact of the parameters on the performance of the two models. In Sections 3.1, 3.2 and 2.2, we focus on the effect of the asymmetry in arrival rates, the asymmetry in service rates, as well as the variability in arrival and service rates on the performance of the two models under consideration, respectively.

3.1 Asymmetry of Arrival Rates

We want to understand the impact of the asymmetry in demand. We separate the study in two steps. First, we construct the asymmetry only on the arrival rate of calls 0. Second, we construct it by differentiating between all the arrival rates of all call types.

3.1.1 Asymmetry on Calls 0

In this section, we study the impact of the parameter p on the comparison between Chaining and SP. Recall that p is the proportion of calls 0 among all arriving calls. To isolate the impact of p, we assume that all call types have the same expected service time, and all the arrival rates of the regular calls are the same, $\lambda_i = \lambda$ for $i = 1, ..., n$. In particular, we are interested to know, for the different ranges of p, which one of the models would be preferred to the other. To do so, we conduct simulation experiments and draw some conclusions on the impact of p on the comparison between Chaining and SP. Using simple models, we also analytically confirm these conclusions.

We choose call center examples with $N = 5$ teams and $n = 4$, i.e., 5 types including type 0. Recall that an agent with skills 0 and i ($i = 1, ..., n$) costs 1, and an agent with skills i and j ($i, j = 1, ..., n$) costs 1+t. We consider various sets of parameters (for Chaining and SP) by varying p and t. For coherency, we keep the overall arrival rate, $\sum_{i=0}^{n} \lambda_i$, constant. For each set of parameters, we optimize the call center size under the constraints $W_i \leq W_i^*$ for $i = 0, 1, ..., n$. The results are shown in Table 1 and Figure 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 1, the column *Crossing value* gives the value of t for which the two models Chaining and SP are equivalent. Below (beyond) this value, Chaining is better (worse) than SP.

Consider small values of t. Table 1 reveals that chaining performs well for small values of p. The best sit-

Table 1: Impact of p and t on the Costs ($\mu_i = \mu_0 = 0.2$ for $i = 1, ..., 4$, $\sum_{i=0}^{n} \lambda_i = 8$, $W_i^* = 0.2$)

	Chaining			SP	Crossing
p	t=0%	t=10%	t=25%		value
0%	49	52.9	58.75	60	t=28.21%
10%	49	52.4	57.5	56	t=20.58%
25%	48	50.6	54.5	52	t=15.38%
50%	49	50.8	53.5	52	t=16.67%
75%	51	52.1	53.75	51	t=0%
90%	51	51.6	52.5	51	t=0%
100%	47	47	47	47	t=0%

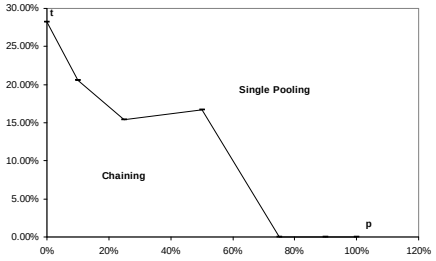


Figure 4: Preference zone ($\mu_i = \mu_0 = 0.2$ for $i = 1, ..., 4$, $\sum_{i=0}^{n} \lambda_i = 8$, $W_i^* = 0.2$)

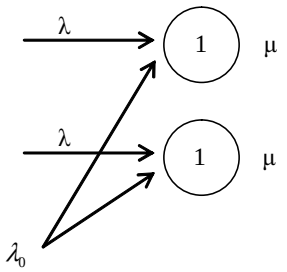
uation for Chaining is reached in the symmetric case (identical arrival rates). The performance of SP improves as p increases. For small values of p, SP approaches FD which has the worst performance. For high values of p, calls 0 are first preponderant and second benefit from pooling, which highly improves 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 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). For any t beyond 30%, SP is systematically better than Chaining whatever in p.

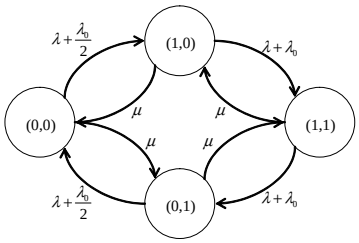
The main conclusion here is that SP can be better than Chaining when the demand for skill 0 is important and/or when skill 0 is less costly than the other ones. In what follows, we analytically retrieve the above conclusions using simple models with no queues.

Understanding with Analytical Simple Models: Consider the simplified model for SP as shown in Figure 5(a): 3 skills (0, 1 and 2) with identical service rates for all skills and with the same arrival rate λ for skills 1 and 2; 2 teams with a single server in each team; no queues, i.e., an arriving call is immediately served or rejected. We want to analyze the impact of p on the performance of SP. We focus on the performance in terms of the system throughput.

We define the process $\{(a(t), b(t)), t \geq 0\}$, where $a(t)$ and $b(t)$ denote the status of the agent 1 and 2, respectively. We use the number 0 for an idle agent and the number 1 for a busy one. Since inter-arrival times and service times are Markovian, $\{(a(t), b(t)), t \geq 0\}$ is a Markov chain, see Figure 5(b). We calcu-



(a) Simplified model



(b) Markov Chain

Figure 5: Understanding Single Pooling

late the stationary probabilities denoted by $\pi_{i,j}$ (for $i, j \in \{0, 1\}$) of this Markov chain as a function of the proportion of calls 0, p, and the global workload, Ω. We have $\Omega\pi_{0,0} = \pi_{1,0} + \pi_{0,1}$, $\pi_{1,0} = \pi_{0,1}$, and $\pi_{1,1} = \frac{(1+p)}{2}\Omega\pi_{0,1}$. Therefore, $\pi_{1,0} = \pi_{0,1} = \frac{\Omega}{2}\pi_{0,0}$. Since the sum of the stationary probabilities equals to one, we obtain $\pi_{0,0} = \frac{1}{1+\Omega+(1+p)\frac{\Omega^2}{4}}$. Because the service rates are identical for all call types, the expected number of calls in the system, say $E(Q)$, is proportional to the throughput. We have $E(Q) = \pi_{1,0} + \pi_{0,1} + 2\pi_{1,1} = \frac{\Omega+\frac{(1+p)}{2}\Omega^2}{1+\Omega+(1+p)\frac{\Omega^2}{4}}$. Then the derivative of $E(Q)$ in p is $\frac{\partial E(Q)}{\partial p} = \frac{4\Omega^2(\Omega+2)}{(4+4\Omega+(1+p)\Omega^2)^2}$. Since $\Omega > 0$, $\frac{\partial E(Q)}{\partial p} > 0$. As a consequence, the performance of Single Pooling increases in p. This agrees with our observation above from simulation. The observation that Chaining performs well in the case of

symmetric arrival rates has been already confirmed in previous work (see for example Benjaafar and al. (1995)).

3.1.2 Asymmetry on the other Arrival Rates

Note that the Chaining model does not make any difference between calls 0 and the regular ones. Increasing the asymmetry on the regular calls has the same effect as increasing the asymmetry on calls 0 (performance decreases). A remaining question is how does SP react with the asymmetry defined on the other arrival rates? The parameter p, used in the previous section, is one way of measuring asymmetry in arrivals. However, this would not allow us to vary the asymmetry between regular call types. In addition to the parameter p, we define here a ratio between the arrival rates that would allow us to vary the asymmetry in arrivals. We denote by V the ratio of the different arrival rates, $V = \frac{\lambda_1}{\lambda_2} = \frac{\lambda_2}{\lambda_3} = \frac{\lambda_3}{\lambda_4}$. We want to study the impact of the parameter V on the performance of Chaining and SP. In order to isolate the impact of V, we assume that all calls require the same service times. We consider the same call centers examples as in Section 3.1.1. The experiments for of the case $V = 1$ reduces to those given in Section 3.1.1 in Table 1. The simulation results for the cases $V = 2$ and $V = 5$ are shown in Table 2 with the same individual costs for every agent.

Table 2: Simulation results ($\mu_i = \mu_0 = 0.2$ for $i = 1, ..., 4$, $\sum_{i=0}^{n} \lambda_i = 8$, $W_i^* = 0.2$)

	$V = 2$		$V = 5$	
p	*SP*	*Chaining*	*SP*	*Chaining*
0%	57	50	54	52
10%	56	49	54	55
25%	53	48	52	52
50%	51	49	52	50
75%	52	49	52	51
90%	51	52	51	52
100%	47	47	47	47

We observe that the performance of SP increases in V and chaining deteriorates when V is important and p is small. This simulation is made with the same individual costs for every agent. Including costs the Single Pooling model is much better than the Chaining one when the assymetry on the regular calls is important.

How can we explain that SP performs better when the asymmetry on the regular calls is important? The Jensen inequality implies that if f is a convex function and X a random variable then $E(f(X)) \geq f(E(X))$. This inequality is useful to understand the impact of asymmetry in SP. Each team i can treat skills i and 0. When all service rates are equal the number of agents in team i only depends on λ_i and λ_0. This number, denoted by N_i, is a concave and increasing function f of λ_i and λ_0 ($N_i = f(\lambda_i; \lambda_0)$). In deed, this function is increasing because increasing the demand requires more agents to achieve a quality of service. Because large teams are more efficient than small ones, this function is concave. The need of an additional agent occurs less frequently when the demand increases. The concavity of the first variable of N and the Jensen inequality leads to $\frac{1}{n}(f(\lambda_1; \lambda_0) + f(\lambda_2; \lambda_0) + \cdots + f(\lambda_n; \lambda_0)) \leq f(\frac{\lambda_1+\lambda_2+\cdots+\lambda_n}{n}; \lambda_0)$. Then, $f(\lambda_1; \lambda_0) + f(\lambda_2; \lambda_0) + \cdots + f(\lambda_n; \lambda_0) \leq n \times f(\frac{\lambda_1+\lambda_2+\cdots+\lambda_n}{n}; \lambda_0)$, which implies that any asymmetric configuration is better than the symmetric one for SP.

The main insight of the previous section is that our proposition, the Single Pooling architecture, can be less costly and as performing as the Chaining one. The asymmetry, usually met in demand, impact on the preference between the two models. When asymmetry is important the Single Pooling is better than Chaining.

3.2 Asymmetry in Service Rates

In section 3.2 we study the impact of the asymmetry of the service rates on the performance of SP and Chaining. Is the asymmetry in service rates equivalent to the asymmetry in arrival rates ? We would like it to be, but in fact it is not. The decreasing of a service rate is more impactive on the optimal staffing than the increasing of an arrival rate. For an M/M/S queue the expression

$W = \frac{1}{\sum_{k=0}^{N-1} \frac{\rho^k}{k!} + \frac{\rho^N}{N!} \frac{1}{1-\rho/N}} \times \frac{\rho^N}{(N-1)!(N-\rho)^2\mu}$ of the average waiting time shows that the ratio $\rho = \frac{\lambda}{\mu}$ does not permit to avoid the knowledge of μ or λ. The expression contains an isolated μ which break the idea of an equivalence between λ and μ inside the ratio $\rho = \frac{\lambda}{\mu}$. Because the handling of the service times is not equivalent to the one of the arrival rates, the parameter p could not directly be extended to $p' = \frac{\rho_0}{\sum \rho_i}$ including the service rates in its definition. The behavior of the architectures (SP and chaining) to different level of the ratio $\frac{\rho_0}{\sum \rho_i}$ has to be studied through simulation and analytical models.

Simulations: We want to understand the impact of p'. We choose the same parameters as in Section 3.1.1 but instead of having the overall arrival rate, $\sum_{i=0}^{n} \lambda_i$, constant we choose the overall workload, $\sum_{i=0}^{n} \rho_i$, to be constant. We assume that all agents have the same cost in order to isolate the impact of p' without per-

turbations. The parameter p' may not have the same impact depending on whether the service rates or the arrival rates are asymmetric. In order to isolate the impact of μ and λ we propose two series of simulations. In the first series we keep the same service rate, $\mu_i = \mu_0 = 0.2$, for every agent. In the second series we keep the same arrival rate, $\lambda_i = \lambda_0 = 2$, for every call type, see Table 3 and 4.

Table 3: Simulation results ($\mu_i = \mu_0 = 0.2$ for i=1,...,4, $\sum_{i=0}^{4} \rho_i = 50$, $W_i^* = 0.2$)

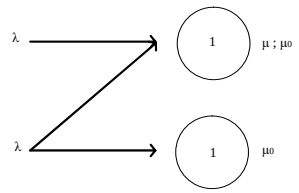
p'	SP	Chaining
0%	68	60
10%	67	59
25%	64	58
50%	61	59
75%	61	61
90%	61	61
100%	57	57

Table 4: Simulation results($\lambda_i = \lambda_0 = 2$ for i=1,...,4, $\sum_{i=0}^{4} \rho_i = 50$, $W_i^* = 0.2$)

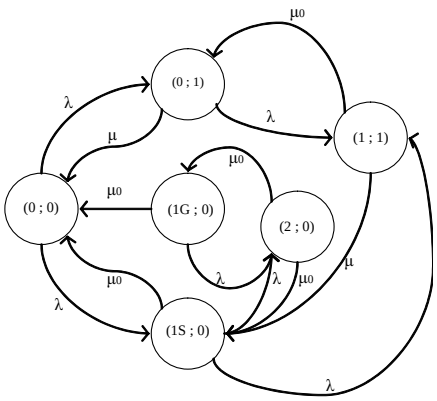
p'	SP	Chaining
0%	72	60
10%	67	59
25%	62	58
50%	65	60
75%	68	61
90%	69	65
100%	62	62

We observe that pooling benefits decrease when service times are different. This is more apparent for SP since calls 0 can go to all teams. However calls 0 have an access to two teams.

Simplified Model Figure 6(a) presents a simplified model with only two calls type, a specialized agent and a generalist one. The priority is given to the specialized agent. We define the process $\{(a(t), b(t)), t \geq 0\}$ where $a(t)$ and $b(t)$ denotes the number of calls 0 and i in the system respectively. When there is only one call 0 and no call i we specified the position of the call 0 by $1G$ if the call 0 is with the generalist agent and $1S$ if the call 0 is with the specialist one. The process $\{(a(t), b(t)), t \geq 0\}$ is still a Markov chain, see Figure 6(b). Table 5 presents calculated values of the throughput as a function of p' and Ω. The worst situation is not always for the highest values of p'. Indeed, if Ω is small then the workload is light and the best situation is when p' is big. When p' is big then calls i are fast-served, calls 0 then see almost two idle agents and benefit from Pooling. When Ω is important it implies that in average the calls are slow-served. In that configuration there will be important rejection. If the calls 0 would be slower served (when p' is big) they would block the system to other calls. This simple model permit to understand two phenomenon in competition:



(a) Simplified Model



(b) Markov Chain

Figure 6: Simplified model of SP with a specialist and a generalist agent

Table 5: Impact of p' and Ω on the throughput ($\lambda = 1$)

p'	$\Omega = 0.5$	$\Omega = 1$	$\Omega = 2$	$\Omega = 3$
0%	1.6667	1.5	1.3333	1.25
10%	1.673	1.4788	1.2418	1.0915
20%	1.681	1.4647	1.1805	0.9964
50%	1.7143	1.4545	1.0909	0.8649
80%	1.7594	1.4749	1.0648	0.8179
100%	1.7949	1.5	1.0667	0.8088

- **The impact of Pooling**: Calls 0 benefit from more agents so they could spend more time being served. This phenomenon is preponderant when the average service time is small or when the workload is light.
- **The impact of Blocking**: Calls 0 can be served by every agent so if they are served in too much time they can block the agents who can not treat the regular calls. This phenomenon is preponderant when the average service time is important or when the workload is heavy.

In a managerial purpose this observation can be turned into an insight. If the workload becomes important during a period of time, it is essential to serve calls who benefits from Pooling as fast as possible in order to avoid a Blocking Effect and rejection.

We also studied asymmetry on the other service rates by simulation and analytical models. The main insight on the impact of the asymmetry of the service rates is about the opportunity of having a flexible architecture. The general advice for the manager would be "avoid flexibility for calls who require much longer service times than the others". SP can suffer from an asymmetry impacting a long service time for calls 0. The Blocking effect impacts every team in SP whereas it would be limited to only two teams in Chaining. SP is less sensitive to the asymmetry on the regular service rates than Chaining. A slow served regular call impacts only one team in SP instead of two in Chaining.

3.3 Asymmetry in Variability

We want to understand the impact of the variability on the performance of the architectures. We evaluate the impact of the variability in the arrival process and in time distributions. The use the coefficient of variation $cv = \frac{\sigma[X]}{E[X]}$ to measure the variability of a distribution X. An exponential distribution has a coefficient of variation of 1. We need to use another distribution to change the variability. Inspired by Brown and al. (2005) we choose a log normal distribution. The log normal distribution is practical to create important values of cv.

3.3.1 Variability in the Arrival Process of Calls 0

We consider the same call centers examples as in section 3.1. We choose a log-normal distribution for the inter-arrival times of calls 0. In order to isolate the impact of the variability in the arrival of calls 0 we keep the other random variables (service times and interarrival of regular calls) Markovian. The simulation results are shown in table 6. We note that increasing cv_0 decreases the performance of the two models. When cv_0 is between 0 and 1 Chaining is as performing as the FF model and better than SP. When cv_0 is between 0 and 2 Chaining is better than SP. When cv_0 is higher than 2 SP is better than Chaining. So the impact of cv_0 is equivalent to the impact of p. Increasing the variability of demand 0 is almost like increasing arrival rate 0.

Table 6: Simulation results ($\mu_i = \mu_0 = 0.2$,$\lambda_i = \lambda_0 = 5$ for $i = 1, ..., 4$)

cv_0	SP	Chaining
0	134	134
0.5	136	134
1	138	134
2	142	141
3	146	149
5	152	153
10	155	156

3.3.2 Variability in the Service Time Distribution of Calls 0

We are interested now on the impact of the variability of the service 0. All the random variables are still Markovian except the service time of calls 0 which follows a log normal distribution. The results are shown in table 7. Increasing cv_0 is equivalent to increasing the average service time of calls 0. First, increasing the cv_0 of the service rate 0 is a good thing in the comparison for SP, but after a limit of $cv_0 > 1$, the Blocking effect seems to reduce the performance much more in the Single Pooling model than in the Chaining model.

Table 7: Simulation results ($\lambda_0 = 2$,$\lambda_i = 1.5$,$\mu_0 = \mu_i = 0.2$ for i=1,...,4)

cv_0	SP	Chaining
0	50	49
0.5	52	50
1	52	49
2	54	54
3	60	62
5	66	64
10	82	75

Conclusion: The variability in the arrival process or in service time distribution is equivalent to the asymmetry of the arrival rate or the service rate if we admit a Markovian distribution for inter-arrival and service times.

4 IMPACT OF THE WORKLOAD

In this section, we focus on the impact of the workload on the performance of the two models. The quality of service, the size of the call center and the workload are correlated. The purpose of this section is to explore the link between those elements.

4.1 Quality of Service constraints

In section 4.1 we investigate on the impact of the standard in quality of service in correlation with the workload. A higher demand in quality of service implies usually a bigger number of agents in the call

center. Then a higher constraint in quality of service goes with a lighter workload. Large call centers are more efficient than the small ones. Then, a good quality of service can be achieved with a heavy workload in a large call center. The size of the call center then need to be studied separately in Section 4.2. To avoid the question of the size in Section 4.1, we choose to compare between the two models for an overall workload, $\sum_{i=0}^{n} \rho_i$ that is constant. We differentiate the study on the quality of service between two asymmetric configurations, an asymmetric configuration in the arrival rates (section 4.1.1) and an asymmetric one in the service rates (section 4.1.2).

4.1.1 Asymmetry in the Arrival Rates

In this section we present simulated results revealing the impact of the demand in quality of service in the two models. We consider various levels of demand by varying the constraints W_i^* for $i = 0, 1, \cdots, n$. We assume that $W_0^* = W_i^*$ for $i = 1, 2, \cdots, n$ in order to have a symmetric constraint in demand. The results for $W_i^* = 1$ and $W_i^* = 0.05$ are shown in table 8.

Table 8: Simulation results ($\mu_i = \mu_0 = 0.2$ for i=1,...,4, $\sum_{i=0}^{4} \lambda_i = 8$)

$W_i^* = 1$	SP	Chaining	Deviation
$p = 0\%$	52	44	8
$p = 10\%$	48	44	4
$p = 25\%$	44	44	0
$p = 50\%$	44	44	0
$p = 75\%$	44	45	-1
$p = 90\%$	43	45	-2
$p = 100\%$	43	43	0
$W_i^* = 0.05$	SP	Chaining	Deviation
$p = 0\%$	64	53	11
$p = 10\%$	64	53	11
$p = 25\%$	60	53	7
$p = 50\%$	59	53	6
$p = 75\%$	59	54	5
$p = 90\%$	59	54	5
$p = 100\%$	50	50	0

Increasing the quality of service (or decreasing the workload) make a preference for Chaining even when p is important because Chaining is more flexible. When the demand in quality of service is low, or when the workload is heavy, the FF model and the FD model get closer in required number of agents, then Chaining and SP get closer because their performances are between the ones of FF and FD. When the performance is measured by the needed number of agent -which is an integer- a very low quality of service conduct to make Chaining and Single Pooling equivalent in number of agents, but if we include costs in the comparison the preference would be for Single Pooling because this architecture is cheaper.

4.1.2 Asymmetry in the Service Rates

In this section we increase the workload by increasing the service times of the calls while keeping the arrival rates constant. We observe that increasing the workload (or decreasing the demand in quality of service) gets the two models closer. When the asymmetry of the service rates is too important (with slow served calls 0) the appearance of the Blocking effect make a preference for Chaining even when the workload is heavy. This observation is similar to the one in the section 3.2.

Conclusion: The main insight for the manager in this Section is that the Chaining architecture is more efficient for a high constraint in quality of service or a moderate workload. In other words, if the demand in quality of service is light, the less flexible architecture can be the cheapest because the different architectures converge in number of agents.

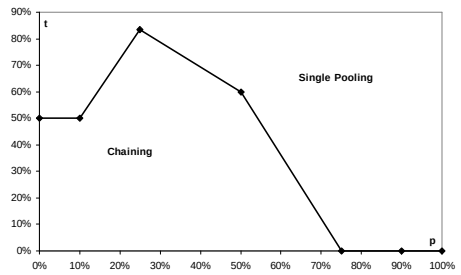
4.2 Workload and Size of the Call Center

In section 4.2 we investigate on the impact of the size of the call center one the two models. We could guess that Chaining would be more efficient than SP for small call centers because its architecture is more flexible. We confirm this intuition by simulations. We choose the same parameters as in section 3.1. We consider a small call center with an overall arrival rate of 1 and a large one with an overall arrival rate of 100. Medium size call centers have already been simulated in section 3. The cost of agents training is different in the two models. Figure 7(a) and 7(b) illustrate a main insight, a Single Pooling architecture is usually less costly than a Chaining one in a large call center even if the demand on skill 0 is not important (small values of p).

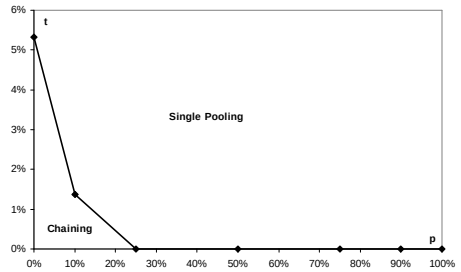
Conclusion: The smaller a call center is, the more benefit it will take from flexibility. That is why the Chaining architecture must be preferred in a small call center and the Single Pooling one in large call center.

4.3 Asymmetry on quality of service constraints

How will chaining and Single Pooling react if the quality of service is different for every calls types ? In our simulations we choose the same parameters as in the previous section. We choose the same demand in quality of service for the regular calls, $W_i^* = 0.2$ for $i = 1, ..., 4$. We consider various sets of parameters by varying the demand W_0^*. The results are shown in table 9.



(a) Small Call Center



(b) Big Call Center

Figure 7: Preference zone in costs between Chaining and SP

Table 9: Simulation results ($\lambda_0 = 4$, $\lambda_i = 1$, $\mu_i = \mu_0 = 0.2$ and $W_i^* = 0.2$ for $i = 1, 2, 3, 4$

W_0^*	0.01	0.1	0.2	1
Single Pooling	56	52	52	52
Chaining	58	51	49	48

Conclusion: Thanks to the priority on regular calls and the Pooling effect, the Single Pooling architecture resists more to an important asymmetry in the quality of service constraints than Chaining.

5 CONCLUSION

In the conclusion we present the main insights of our study for the manager.

- **Costs of regular skills**: When regular skills are more costly than skill 0 then the preference is for SP.
- **Arrival rate 0**: When the arrival rate 0 represent more than 50% of the overall arrival rate then the SP architecture can be better.
- **Arrival rate of regular calls**: An important asymmetry on arrival rates improves the performances of SP and deteriorates the ones of Chaining.
- **Service rate 0**: An important service time for calls 0 can improve the performances of SP more than the ones of Chaining until the appearance of the Blocking effect.
- **Service rate of regular calls**: The blocking effect bound to regular calls is more impactive in Chaining. The preference can be for SP when regular calls have very different service rates.
- **Variability**: The conclusions on variability follows the ones of arrival and service rates
- **Workload**: Increasing workload favorites Single Pooling
- **Quality of service**: Increasing the quality of service favorites Chaining
- **Size of the call center**: Single Pooling is a more efficient architecture in a large call center
- **Number of team**: The gap between the two models increases when the number of teams is increasing

REFERENCES

Avramidis A., Chan W. and L'Ecuyer P., 2009. Staffing Multi-Skill Call Centers via Search Methods and a Performance Approximation. *IIE Transactions*, vol. 41, p. 483-497.

Bassambo A., Randhawa R., and Van Mieghem J., 2010. Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing. *Management Science*,vol. 56, p.1285-1303.

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

Borst S., Mandelbaum A., and Reiman M., 2004. Dimensioning Large Call Centers. *Operations Research*, vol. 52, p.17-34.

Brown L., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S., Zhao L., 2005. Statistical Analysis of a Telephone Call Center. *Journal of the American Statistical Association*, vol. 100, p.36-50.

Garves S. and Tomlin B., 2003. Process Flexibility in Supply Chains. *Management Science*, vol. 49, p. 907-919.

Gurumurthy S. and Benjafaar S., 2004. Modeling and Analysis of Flexible Queueing Systems. *Naval Research Logistics*, vol. 51, p. 755-782.

Iravani S., Kolfal B. and Van Oyen M., 2007. Call-Center Labor Cross-Training: It's a Small World After All. *Management Science*, vol. 53, p. 1102-1112.

Jordan W. and Graves S. 1995. Principles on the Benefits of Manufacturing Process Flexibility. *Management Science*, vol. 41, p. 577-594.

Van Dijk N. and Van Der Sluis E., 2008. To Pull or not to Pull in Call Centers. *Production and Operations Management*, vol. 17, p. 1-10.

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

Zohar E., Mandelbaum A., and Shimkin N., 2002. Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support. *Management Science*, vol. 48, p. 566-583.