Revista Produção On Line
Universidade Federal de Santa Catarina
www.producaoonline.inf.br
ISSN 1676 - 1901 / Vol. 4/ Num. 3/ Agosto de 2004

Modelo para segmentação da demanda de um Call Center em múltiplas prioridades: Estudo da implantação em um Call Center de Telecomunicações

Marcus Augusto Vasconcelos Araújo (UPE) marcus-recife@uol.com.br
Francisco José Costa Araújo (UPE) paco51@terra.com.br
Paulo José Adissi (UFPB) adissi@producao.ct.ufpb.br

Resumo

A crescente necessidade de relacionar-se de forma eficiente com seus Clientes levou as empresas à criação de um serviço que permite a sua orientação e captação pelo telefone: O Call Center. Este segmento corporativo tem se tornado cada vez mais competitivo e só sobrevivem as empresas que conseguem, com operações enxutas, obter bons resultados. Neste cenário, as disciplinas das filas, quando bem administradas, são fortes aliadas da área de planejamento e controle da produção dos Call Centers, que tem como meta atingir os resultados esperados com recursos, muitas vezes, escassos, tornando esta área cada vez mais importante nestas empresas. Este trabalho mostra a bem sucedida experiência da implantação de uma disciplina de fila diferente daquelas mais freqüentemente usadas. Esta disciplina foi implantada em um Call Center de uma empresa de telecomunicações e, além de ter gerado uma economia de quase R$ 1.200.000,00/ano, aumentou a satisfação dos seus Clientes.

Palavras-chaves*: Telefonia, Teoria das filas, Call Center.*

Abstract

The increasing necessity in maintaining an effective relationship with the costumers created a service that allows the orientation and capitation of the costumers using the telephone: The Call Center. This business area has become more competitive and only the companies that achieve the best results with efficient operations will survey. In this reality, the queue disciplines, when well managed, are strong allies of Call Centers' production planning and control area, which goal is to achieve good results with limited resources, increasing this area's importance in Call Centers'

Data de Submissão: Jul/03 Data de Aprovação: Jul/04

operations. This article shows the successful experience in using a different queue discipline. This queue discipline that was implanted in a telecommunications' company Call Center and generated an economy of R$ 1.200.000,00 per year, beyond increased the Customers satisfaction
***Key-words**: Telephony, Queuing theory, Call Center.*

1. Introdução

A empresa em que o projeto foi implantado, atua no segmento de Telecomunicações e sua Diretoria de Relacionamento com o Cliente é responsável pelo gerenciamento das interações dos Clientes com a empresa.

As principais responsabilidades desta diretoria são as de elaboração e operacionalização de estratégias voltadas ao atendimento e antecipação das necessidades dos seus Clientes e *prospects*.

Dentre as várias estratégias com foco no atendimento às necessidades dos Clientes, sem dúvida, algumas das mais importantes são as voltadas para o gerenciamento eficaz da Performance do Call Center, onde, com recursos finitos (atendentes), consegue-se maximizar os resultados de performance (tempo de espera em fila, nível de serviço observado, etc...) e, conseqüentemente, aumentar a satisfação dos Clientes.

Em todo Call Center se observa que a diminuição da fila de espera sempre é um objetivo almejado e, para tanto, comumente, são utilizadas apenas a previsão de demanda com posterior contratação/realocação de mão-de-obra para atendê-la. Este gerenciamento de Demanda e Capacidade de atendimento do Call Center é feito pela área de Planejamento e Tráfego. Porém, apesar da constante necessidade de aumento de mão-de-obra em conseqüência do aumento da demanda que, normalmente, é gerada pelo aumento da quantidade de usuários das companhias, os custos associados a este aumento nem sempre são aprovados, pois os recursos muitas vezes não são previstos em orçamento na sua totalidade. Por este motivo, torna-se freqüente a situação de subdimensionamento de recursos frente à demanda recebida pelo Call Center, aumentando o grau de dificuldade da atividade de gerenciamento da performance pela área de Planejamento.

Este trabalho apresenta uma solução alternativa encontrada para o problema de subdimensionamento onde foram criadas estruturas de filas diferentes das que normalmente são utilizadas. Para tanto, no lugar de filas simples não-priorizadas, são feitas priorizações para o atendimento à demanda considerando algumas variáveis associadas às ligações, como tipo da ligação e sua duração média esperada.

Será mostrado que as filas inteligentes, assim chamadas, têm a função de diferenciar os diversos tipos de Clientes, segmentando-os de acordo com o tipo da sua dúvida e criticidade da ligação, dando, assim, um tratamento diferenciado para Clientes com perfis diferentes e problemas de diferentes tipos, além de melhorar os indicadores de Performance da Central.

2.1. Filas

As filas estão presentes tanto nos serviços quanto na produção de produtos e, em ambos os casos, têm o papel de "gargalos". No caso de um call center, as filas são notadas mais claramente quando um Cliente espera para ser atendido “ouvindo música”. Neste contexto, a teoria das filas tem como objetivo principal o desenvolvimento de modelos matemáticos que permitam prever o comportamento de sistemas com fila de espera.

Na concepção de Marques, Philippi e Nascimento (2001) é essencial saber quando os momentos de pico vão ocorrer e até que ponto podem ocasionar a espera do Cliente. Contudo, a natureza dos serviços e sua produção são mais complexas e menos previsíveis que a produção de bens. Ainda de acordo com Marques, Philippi e Nascimento (2001), há várias formas pelas quais a incerteza estatística ou a variabilidade podem afetar um processo de serviço de modo a influenciar tanto a oferta do processo quanto à sua demanda. Em um Call Center, as variabilidades que distorcem as previsões podem ser exemplificadas, do lado da capacidade, quando um funcionário tira uma licença médica e, do lado da demanda, quando são lançadas promoções de marketing que tendam a elevar a demanda média esperada, tese esta reforçada por Fildes (2002), sinalizando que o desconhecimento de dados relativos a campanhas promocionais aumenta a dificuldade de previsão do tráfego telefônico a ser recebido no Call Center. Podem ser citadas, ainda, como variabilidades que distorcem as previsões, a ocorrência de problemas técnicos na rede e os erros nas

contas enviadas para os Clientes que sempre geram uma demanda adicional não prevista. Quanto maior for a variabilidade na demanda ou na oferta do processo e a incerteza estatística utilizada na projeção, maior é a probabilidade de ocorrência de um gargalo (capacidade de atendimento inferior à demanda oferecida). Desta forma é muito importante não desconsiderar acontecimentos, de certa forma, imprevisíveis. Neste cenário, Marques, Philippi e Nascimento (2001) apontam um dos principais desafios dos administradores de Call Centers: "os prestadores de serviço podem aumentar a capacidade do processo pela simples descoberta de formas de administrar a variabilidade na demanda ou na oferta (estreitar a variância) à qual o processo está sujeito, sem adicionar equipamentos ou mão de obra".

Independentemente da sua complexidade, as filas de espera são caracterizadas pelos Mecanismos de chegadas, de serviço e pela Disciplina utilizada. De acordo com Prado (1999), o Mecanismo de chegada descreve a forma como os Clientes chegam ao sistema. Estas chegadas podem ser caracterizadas pela taxa de chegadas λ (nº de chegadas por unidade de tempo) e uma distribuição (um exemplo típico é considerar que as chegadas seguem uma distribuição de Poisson). Para caracterizar o Mecanismo do serviço são utilizadas as taxas de serviço (μ) e a distribuição, que é o número de postos de serviço (número de atendentes). Já a disciplina da fila refere-se às regras de escolha do Cliente seguinte a ser servido. A regra mais comum e que, normalmente, é utilizada em Call Centers é a FIFO (*first in, first out*), na qual o primeiro Cliente a chegar ao inicio da fila é o primeiro a ser atendido. Além dela, podem ser citadas a também comum LIFO (*last in, first out*) e outras mais complexas, cuja ordem de atendimento pode ser baseada na definição de várias prioridades diferentes (modelo abordado neste trabalho).

2.2 O Dimensionamento de Call Centers

Analisando-se superficialmente, a atividade de projetar um Call Center parece ser muito simples: Desde que se conheça a demanda de ligações, deve-se calcular o número de atendentes que seriam suficientes para atender a tal demanda de modo que não haja fila.

Ao ampliar-se a análise do projeto de uma central de modo a considerar os seus custos, conclui-se que esse ponto de vista é ingênuo, pois, no ambiente corporativo,

os recursos são limitados e a conseqüência deste tipo de dimensionamento seria uma capacidade de atendimento superior à demanda oferecida em todos os horários, o que implicaria em recursos gastos desnecessariamente. De acordo com Moreira (1998), operar com uma capacidade acima ou abaixo das necessidades aumenta inutilmente os custos operacionais. A situação de superdimensionamento normalmente gera um grande desconforto, pois, diferentemente das áreas de vendas, o ganho obtido com um Call Center é difícil de ser medido e o retorno financeiro desta área nem sempre é mensurado. Dessa forma, por ser visto, basicamente, como uma fonte de despesas e não de receita, é de interesse das empresas manter este serviço sob um forte controle de custos, o que cria a necessidade de que essas operações atinjam os resultados esperados de Satisfação dos Clientes e Qualidade no atendimento com estruturas otimizadas. Além disso, conforme já foi observado, deve-se atentar para os fatores imprevisíveis que podem provocar um pico inesperado de demanda a ponto de ultrapassar a capacidade de atendimento projetada da central, estabelecendo-se a situação de fila.

Na prática, no dimensionamento de um Call Center se faz necessário achar um número de atendentes tal de modo a garantir que a probabilidade de haver um excesso de demanda com conseqüente fila não seja maior do que um valor considerado razoável.

O matemático dinarmaquês A. K. Erlang, responsável pelos primeiros estudos teóricos das redes telefônicas feitos no inicio do século XX, descobriu que as chamadas recebidas por um grupo de troncos e, ampliando-se para o caso estudado, em um Call Center, poderiam ser aproximadas por uma distribuição de probabilidade de Poisson onde devem ser consideradas as seguintes variáveis:

- Número de atendentes **a** que devem estar à disposição dos usuários do Call Center;
- A demanda da central **d**, que é o tempo relativo ao total de ligações atendidas na central. Para seu cálculo se considera o somatório dos tempos de ocupação das chamadas recebidas pela central em uma hora. A sua unidade de medida é o Erlang;

- Probabilidade de ocorrência de fila na central **c**, que reflete as chances de uma ligação, ao chegar na central, não ser atendida de imediato por não haver atendentes livres;

Número de atendentes = a
Demanda na central = d
Probabilidade de fila = c

Tabela 1 – Variáveis para cálculo da quantidade atendentes

Dessa forma, Erlang chegou a uma fórmula para cálculo da probabilidade de fila em um Call Center, chamada de formula B de Erlang:

$$c(a,d) = \frac{\frac{d^{\,a}}{a!}}{1 + \frac{d}{1!} + \frac{d^2}{2!} + \frac{d^3}{3!} + \ldots + \frac{d^{\,a}}{a!}}$$

Figura 1 – Fórmula para cálculo da probabilidade de fila em um Call Center

Conclui-se, então, que, para o dimensionamento simples de um Call Center, é necessário expressar o número **a** de atendentes em termos da demanda **d** a ser atendida e da Probabilidade de filas **c** a qual se está disposto a aceitar. Assim o problema básico da telefonia é achar a função a (c , d).

Porém, como a variável a ser calculada é o número de atendentes **a** e não a probabilidade de fila **c** que, normalmente, é escolhida, conclui-se que a fórmula B de Erlang não possibilita o cálculo direto do valor de **a**, a não ser que ela seja tratada como uma equação na incógnita **a** ou seja utilizada uma tabela de valores de **a** para uma grande quantidade de possibilidades de **d** e **c**, com posterior uso de interpolação matemática.

O resultado obtido ao se utilizar a fórmula B de Erlang para dimensionar um Call Center, por si só, não garante o tempo de espera a que será submetida uma certa

quantidade de Clientes enquanto em fila. Por isso, além da probabilidade de haver filas, deve-se levar em consideração o tempo médio de espera a que o Cliente será submetido nessa fila ou o Nível de Serviço desejado, que é o percentual de ligações atendidas dentro de um determinado tempo, em relação ao total de ligações recebidas (usualmente é considerado um tempo de 20 segundos como aceitável). Devido à impossibilidade de ser usada para esta situação, a partir da fórmula B de Erlang, foi desenvolvida a fórmula C de Erlang:

$$NS(\%) = 1 - c(a,d)e^{-(a-d)\frac{AWT}{\beta}}$$

Figura 2 – Fórmula para cálculo do nível de serviço

Onde:

- NS(%) é o nível de serviço percentual;
- c(a,d) é a probabilidade de que haja fila no Call Center;
- a é a quantidade de atendentes;
- d é a demanda oferecida em Erlangs;
- AWT é o tempo de espera aceitável;
- β é o tempo médio de duração das ligações.

Com a fórmula C de Erlang, pode-se estimar o nível de serviço dadas a demanda e a quantidade de atendentes. Convém lembrar que, analogamente à formula B de Erlang, o objetivo real em um dimensionamento de Call Center é determinar a quantidade de atendentes necessários baseando-se na demanda oferecida e no nível de serviço desejado.

2.2. Componentes de um Call Center

Para um melhor entendimento do processo de priorização de filas em um Call Center, faz-se necessário o conhecimento das funções de alguns componentes da central:

- URA é uma interface entre o sistema telefônico e o banco de dados do Call Center. É um dispositivo composto por canais de conversação, que, após ser acessado pelo Cliente disponibiliza informações de acordo com as opções escolhidas, configurando o "Auto-Atendimento". Neste dispositivo existem opções com conteúdos explicativos e opções de saída para que o Cliente possa falar com o atendimento pessoal. Para o desenvolvimento do modelo de filas priorizadas, cada opção de saída deve estar associada a um *VDN* e cada *VDN* deve estar associado a apenas uma opção de saída.

- O *VDN* (do inglês *Vector Directory Number*)é um ramal virtual (não-físico) utilizado para o roteamento das chamadas. Toda chamada se associa um *VDN* que, por sua vez, está sempre associado a um vetor.

- O Vetor é o ambiente onde, efetivamente, são escritas as regras de roteamento às quais as chamadas devem ser submetidas. Associar um VDN a um determinado vetor faz com que todas as ligações deste VDN sigam a regra de roteamento presente no vetor (regra também chamada de vetorização).

- O *Skill* é o grupo de atendimento ao qual o atendente está conectado. É para estes grupos que as chamadas são roteadas e, neles, ficam enfileiradas para posterior atendimento.

Na figura seguinte, é mostrado um exemplo de um vetor simples que enfileira com prioridade Baixa (L) no Skill 418.

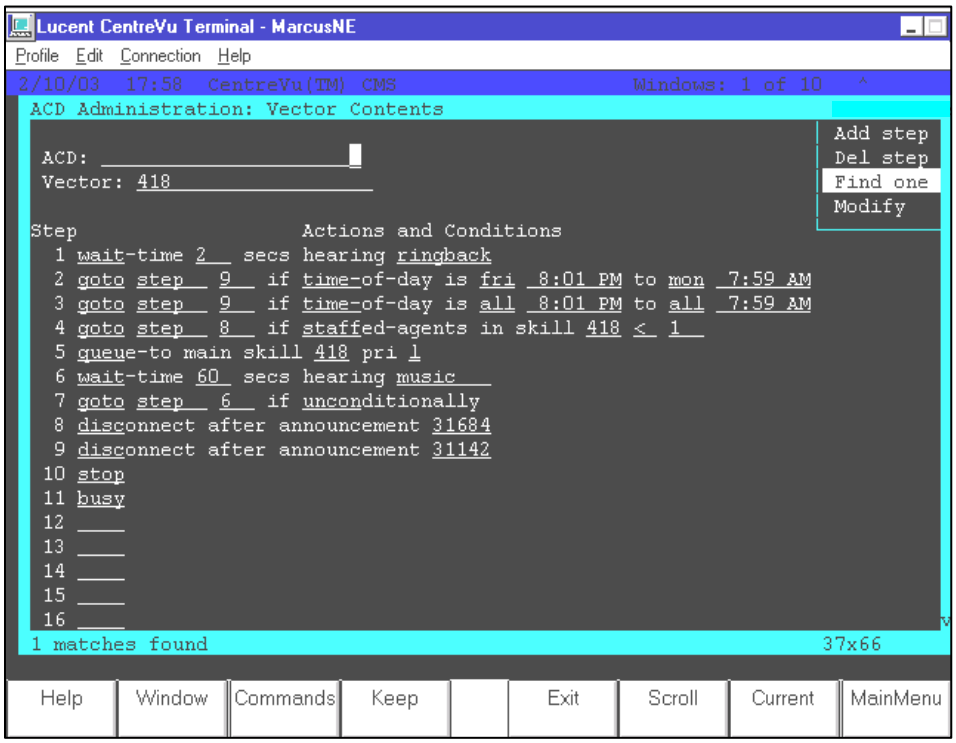

Figura 3 – Tela de exemplo de um vetor

No processo de roteamento das chamadas para um determinado Skill de atendimento, a prioridade das chamadas é atribuída no vetor, como também os possíveis transbordos para prioridades superiores.

A ordem para o atendimento das chamadas é definida baseada em um Algoritmo de Distribuição que considera algumas regras básicas: Quando uma chamada é enfileirada (nenhum atendente livre), ela passa aguardar que um atendente fique livre. A partir do momento que o primeiro atendente fica livre, as chamadas são atendidas na seguinte ordem:

1- A chamada em espera com a maior prioridade na fila é sempre atendida antes das chamadas com menor prioridade nesta mesma fila (4 prioridades disponíveis: Máxima, Alta, Média e Baixa).

2- Dentre as chamadas de mesma prioridade na fila, a chamada que estiver esperando há mais tempo será atendida prioritariamente.

Para ilustrar o funcionamento de um Call Center, a seguir, é mostrado o macro-fluxo do processo de tratamento de uma ligação feita por um Cliente para um Call Center.

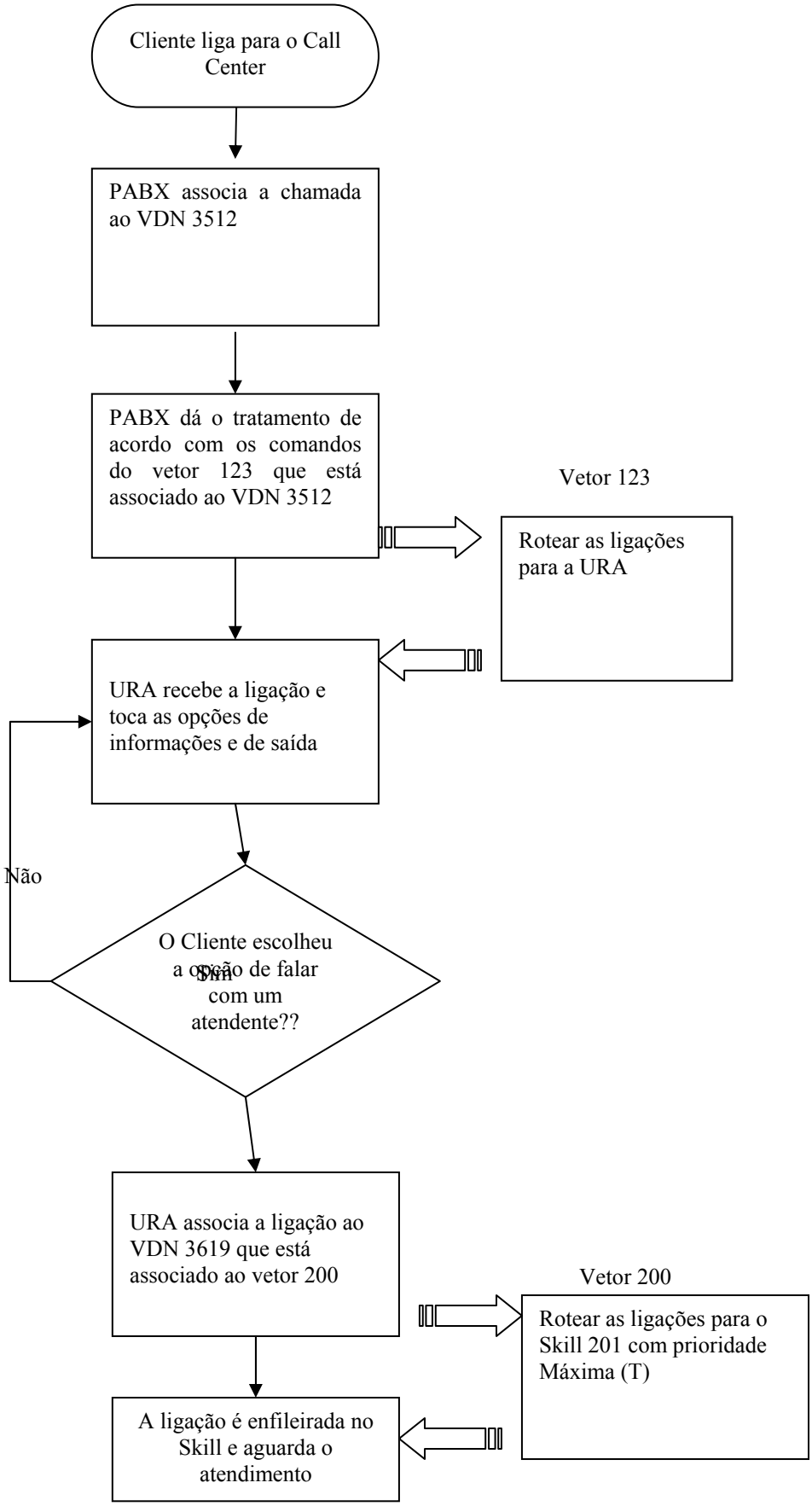

Figura 4 – Processo simplificado de tratamento de uma ligação no Call Center

3. Criação e priorização das filas

Juntamente com a qualidade do atendimento prestado e a eficiência na resolução de problemas, o nível de serviço e a velocidade de atendimento das ligações sempre foram grandes contribuintes para a obtenção da satisfação dos Clientes atendidos no Call Center desta empresa. Apesar disto, notava-se que, mesmo prestando um atendimento com qualidade e eficiente, a ocorrência de longas esperas em fila sempre geravam uma grande insatisfação dos Clientes. Essa situação é descrita por Zeithaml e Bitner (2000): "Para a maior parte das organizações de serviço, os Clientes em espera são inevitáveis em algum momento" e complementam afirmando que as organizações que fazem os Clientes esperar em demasia, correm o risco de perder negócios ou, pelo menos de deixá-los insatisfeitos. Para evitar tal insatisfação, se fazia necessária a manutenção da força de trabalho de forma a atender a demanda recebida sem que fossem geradas grandes filas de espera.

Com o passar do tempo e conseqüente envelhecimento da mão de obra , um problema crônico começou a aparecer nesta central: As LER e DORTs. Estas doenças, de acordo com Araújo, Melo e Andrade (2002), são lesões por Esforços Repetitivos causadas em pessoas que executam tarefas nas quais movimentos continuados ou repetitivos são realizados constantemente e são causadas, muitas das vezes, pela combinação de problemas de postura com a pressão excessiva para os resultados, um ambiente excessivamente tenso, rigidez excessiva no sistema de trabalho, estresse emocional, repouso inadequado, fator cognitivo, entre outros.

Com este problema, imediatamente, o percentual de absenteísmo aumentou significativamente e, consequentemente, a capacidade produtiva da central foi reduzida, ficando menor do que aquela que seria necessária para a manutenção dos seus indicadores de desempenho. Isso aconteceu pelo fato da substituição dos atendentes ausentes por novos atendentes não poder ser feita mesmo com o afastamento dos funcionários por licença, pois todos os afastados continuavam na folha de pagamento da empresa, o que inviabilizava tal ação.

Foi, então, criada uma situação em que, durante boa parte do dia, a demanda era superior à capacidade de atendimento (**d**>**a**) e, neste tipo de situação, utilizando-se a disciplina de filas FIFO, o nível de serviço tornava-se zero. Esta situação pode ser verificada pela fórmula C de Erlang.

O dimensionamento do Call Center, em termos de quantidade de atendentes, era feito em função de três variáveis:

a) A demanda prevista em Erlangs;

b) As premissas utilizadas, como, por exemplo, a improdutividade dos atendentes, o seu absenteísmo e o percentual do tempo de descanso;

c) A disciplina/modelo de fila utilizado.

Como a demanda prevista sempre tinha um alto grau de acerto e o modelo de fila que, no caso, era o *FIFO (First-In-First-Out)*, era sempre o mesmo, a única variável que tinha sofrido uma alteração significativa e, por isso, inviabilizava o novo dimensionamento, era o absenteísmo gerado pelas LER e DORTs. Além de inviabilizar a contratação de novos atendentes em substituição aos afastados, quando era gerado um novo dimensionamento com a premissa real de absenteísmo, os números obtidos eram proibitivos e não eram aprovados pela Diretoria.

O problema estava claro: Como manter (e melhorar) os indicadores de desempenho da central dentro dos padrões esperados com uma capacidade de atendimento reduzida?

Dos conceitos de Engenharia de Produção surgiu a idéia: Modificar o modelo de fila utilizado de modo que, criando-se uma fila inteligente onde as chamadas mais curtas fossem atendidas na frente das mais longas, fosse possível manter os indicadores dentro dos limites esperados, mesmo estando com recursos reduzidos. O raciocínio era simples: era melhor atender 10 ligações com duração prevista de 1 minuto cada, antes de atender uma só ligação de 10 minutos de duração. Dessa forma além da melhoria no indicador de Nível de Serviço, a maior parte dos Clientes seria atendida mais rapidamente.

A tarefa consistia, então, em escrever as regras de priorização em vetores onde todas as ligações que tivessem um tempo médio de duração pequeno, fossem atendidas

antes das que tinham um tempo médio de duração maior. Para tanto, fez-se necessário o mapeamento dos diversos VDNs utilizados no Call Center, considerando os tipos de ligações, suas respectivas demandas e tempos médios de duração.

Abaixo se observa a estrutura de fila com prioridades diferentes.

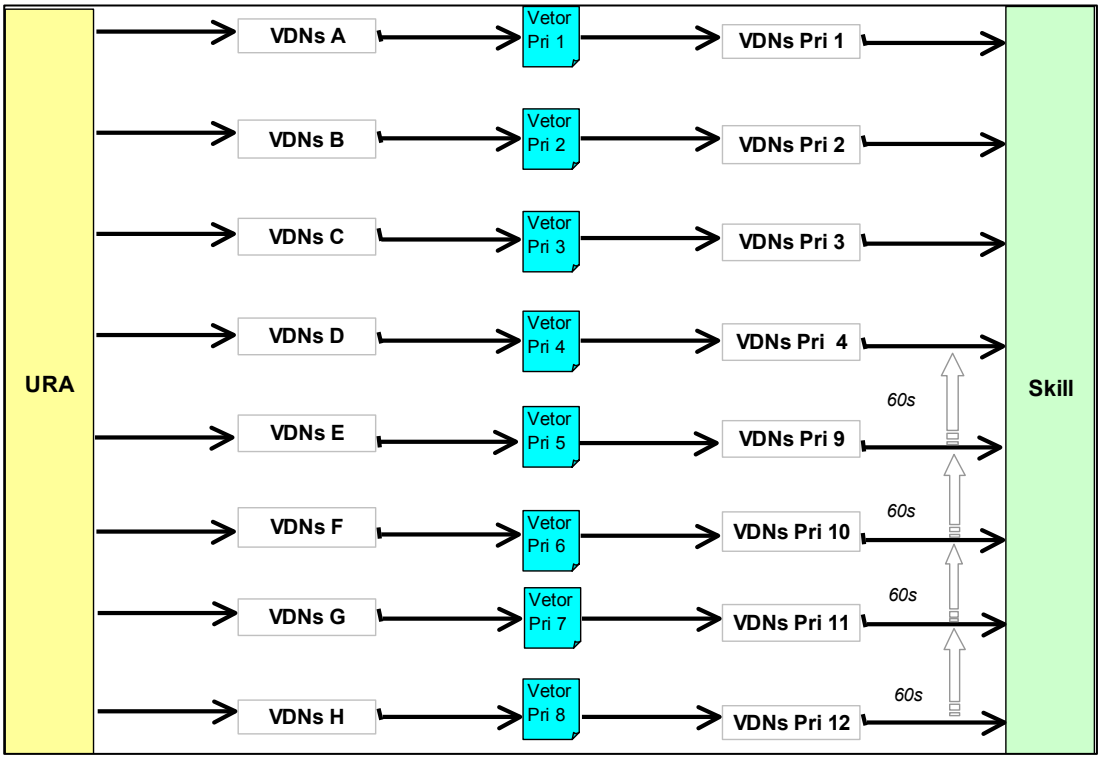

Figura 5– Esquema de roteamento por prioridade em um Call Center

Basicamente, o que permitia segmentar as ligações era a variedade de perfis e tipos de dúvidas dos Clientes. Só com a confirmação da alta variabilidade entre as durações médias das ligações associadas a cada opção da URA foi possível iniciar a análise da Demanda. Após criteriosa análise das ligações associadas a cada opção, pôde-se notar que as ligações provenientes de algumas opções em especial geravam questionamentos mais complexos por parte dos Clientes, que poderiam durar mais de dez minutos. Por outro lado, existiam as ligações de resolução simples, cujo tempo necessário para o completo atendimento, muitas vezes, não ultrapassava um minuto. Concluiu-se, então, que esta situação se configurava em um cenário perfeito para a implantação do novo modelo de fila.

Para a operacionalização do agrupamento em prioridades, sugere-se utilizar uma tabela onde são colocados todos os VDNs e os respectivos tempos médios previstos de duração de suas chamadas. Como o interesse é priorizar as ligações curtas, faz-se necessário ordená-los em ordem crescente de tempo médio. Dessa forma, os primeiros VDNs se associam aos vetores que enfileiram com as maiores prioridades na fila ou, em outras palavras, as chamadas destes VDNs serão atendidas prioritariamente.

Duração média							*Demanda Total*		
VDN	*Total*	Coef. De Variação	Chamadas (%)	Chamadas (#)	Ocupação	Ocupação (%)	VDN	*Total*	Demanda (%)
35556	00:55	0,23	1,0%	161	149:01:50	0,3%	32107	1.019	6,2%
32027	01:19	**0,78**	0,0%	1	1:19:07	0,0%	32260	871	5,3%
32803	01:35	0,27	1,4%	233	371:00:38	0,7%	32106	687	4,2%
32201	01:40	0,23	0,4%	67	112:33:08	0,2%	32598	642	3,9%
32260	01:46	0,10	5,3%	871	1543:22:50	2,7%	32264	552	3,3%
32024	01:50	**0,71**	0,0%	3	4:41:07	0,0%	32552	478	2,9%
32807	02:01	0,20	0,6%	91	184:23:59	0,3%	32102	395	2,4%
32028	02:10	0,41	0,0%	3	7:05:32	0,0%	32284	381	2,3%
32019	02:11	0,17	0,9%	155	339:23:38	0,6%	35466	346	2,1%
32288	02:17	0,18	0,4%	73	167:14:10	0,3%	32294	302	1,8%
32274	02:19	0,28	0,2%	29	66:32:40	0,1%	32105	291	1,8%
32026	02:25	**0,67**	0,0%	4	9:51:15	0,0%	32602	280	1,7%
32015	02:26	0,12	0,3%	53	128:26:29	0,2%	32103	242	1,5%
32806	02:27	0,22	0,4%	66	162:12:48	0,3%	32803	233	1,4%
32071	02:32	0,37	0,1%	14	36:18:21	0,1%	32108	206	1,2%
32278	02:32	**0,59**	0,0%	6	16:20:08	0,0%	32290	200	1,2%
32065	02:32	0,24	0,1%	14	35:48:56	0,1%	35556	161	1,0%
32280	02:34	0,27	0,1%	12	32:06:41	0,1%	32019	155	0,9%
32282	02:41	0,13	0,6%	97	259:23:32	0,5%	32021	143	0,9%
32064	02:45	0,38	0,1%	13	35:57:00	0,1%	32068	142	0,9%
32600	02:45	0,20	0,4%	63	173:26:36	0,3%	32601	130	0,8%
32242	02:46	0,05	7,3%	1.193	3306:25:20	5,9%	32266	109	0,7%
32286	02:46	0,10	0,5%	88	244:46:25	0,4%	32262	104	0,6%
32266	02:49	0,12	0,7%	109	306:18:06	0,5%	32282	97	0,6%
32284	02:49	0,07	2,3%	381	1074:57:02	1,9%	32807	91	0,6%
32022	02:52	0,19	0,2%	32	91:46:23	0,2%	32286	88	0,5%
32264	02:57	0,07	3,4%	552	1630:22:04	2,9%	32292	78	0,5%
32061	03:00	0,27	0,2%	25	74:50:34	0,1%	32605	75	0,5%
32270	03:02	0,36	0,1%	9	27:57:07	0,0%	32288	73	0,4%
32025	03:03	0,42	0,1%	12	35:42:04	0,1%	32062	70	0,4%
32106	03:09	0,05	4,2%	687	2167:08:57	3,8%	32201	67	0,4%
32272	03:12	0,15	0,2%	39	126:35:49	0,2%	32806	66	0,4%
32601	03:13	0,13	0,8%	130	420:01:39	0,7%	32095	63	0,4%
32062	03:14	0,15	0,4%	70	226:28:07	0,4%	32066	63	0,4%
32021	03:15	0,13	0,9%	143	466:55:50	0,8%	32600	63	0,4%
32605	03:16	0,24	0,5%	75	245:25:57	0,4%	35552	56	0,3%

Figura 6 – Tabela para organização dos dados dos VDNs

Ainda nesta tabela, sugere-se incluir o coeficiente de variação do tempo médio de cada VDN, pois este número serve para indicar as altas variabilidades nos tempos dos VDNs. Os VDNs com um grande coeficiente de variação devem ser constantemente observados, principalmente se estiverem com uma prioridade relativamente alta na fila. O coeficiente de variação é calculado pela fórmula $C = \sigma \setminus TMA$, onde TMA é o tempo médio de duração das chamadas e σ o desvio padrão dos tempos de duração das chamadas.

Com os valores em mãos, é possível calcular as demandas totais e sua participação percentual geradas pelos VDNs para que, na escolha da prioridade de cada um, não haja o risco de que as maiores prioridades recebam uma grande concentração de demanda, o que poderia gerar um tempo espera muito alto para as chamadas associadas aos VDNs com prioridades mais baixas.

Além das regras básicas de priorização descritas, no agrupamento dos VDNs em prioridades, também deve ser levada em consideração, em caráter de exceção, a importância de cada VDN no lugar da duração de suas chamadas. Um exemplo deste tipo de exceção é o VDN associado à opção da URA de Suspensão por Perda ou Roubo. As chamadas deste VDN apesar de, historicamente, apresentarem um tempo médio de duração relativamente alto, precisam ter uma prioridade alta na fila devido à urgência do assunto a ser tratado. Todas as pessoas que escolhem essa opção na URA precisam ter os seus aparelhos suspensos rapidamente para evitar que alguém faça ligações indevidas a partir dele.

Respeitadas as premissas já citadas, quanto menor for a variabilidade dentro de cada prioridade melhor. Sugere-se que seja feita uma análise de *Clusters* para garantir que, em uma mesma prioridade, não sejam alocados VDNs com tempos médios muito diferentes, o que infringiria, dentro da prioridade, a lei de “as menores na frente”.

No modelo criado, foi incluída uma última regra que criava um caminho alternativo para as ligações de prioridades baixas que esperavam na fila por mais de um tempo limite. Com este caminho alternativo, sempre que uma ligação ultrapassava um tempo de espera em fila considerado alto, esta ganharia prioridades mais altas até que fosse atendida.

Por fim, ao associar-se os VDNs aos seus devidos vetores, era colocada em produção a estrutura de fila inteligente, onde grandes resultados foram obtidos, como pode ser visto a seguir.

3. Resultados obtidos

Pode-se observar no gráfico abaixo que, já no primeiro mês de implantação, o nível de serviço, que é o principal indicador de performance da central, aumentou de 42% para 84%, mantendo-se estável pelos demais meses. Analisando os meses anteriores à mudança e os imediatamente posteriores, observa-se o aumento médio de 35 pontos percentuais.

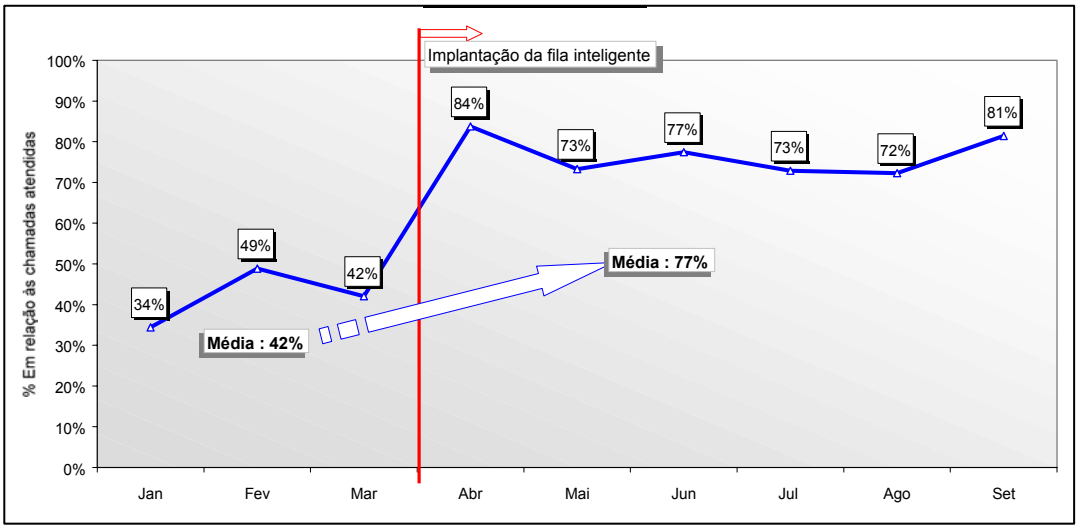

Figura 7 – Gráfico do nível de serviço observado

Um outro surpreendente resultado foi a estabilidade adquirida por este indicador em momentos de picos de demanda. Por mais que aparecessem picos inesperados de demanda, o modelo de fila implantado garantiu a estabilidade do nível de serviço, fazendo com que a maior parte dos Clientes não sentissem diferença no tempo médio de espera, diferentemente do que aconteceria se fossem utilizadas filas do tipo FIFO que são muito mais frágeis para problemas deste tipo.

Já o *Answer Rate*, que é, por definição, o percentual das ligações atendidas em relação ao total de ligações recebidas, passou de 65% para 86%, em média. Isso representa um ganho real na quantidade de ligações atendidas dos Clientes.

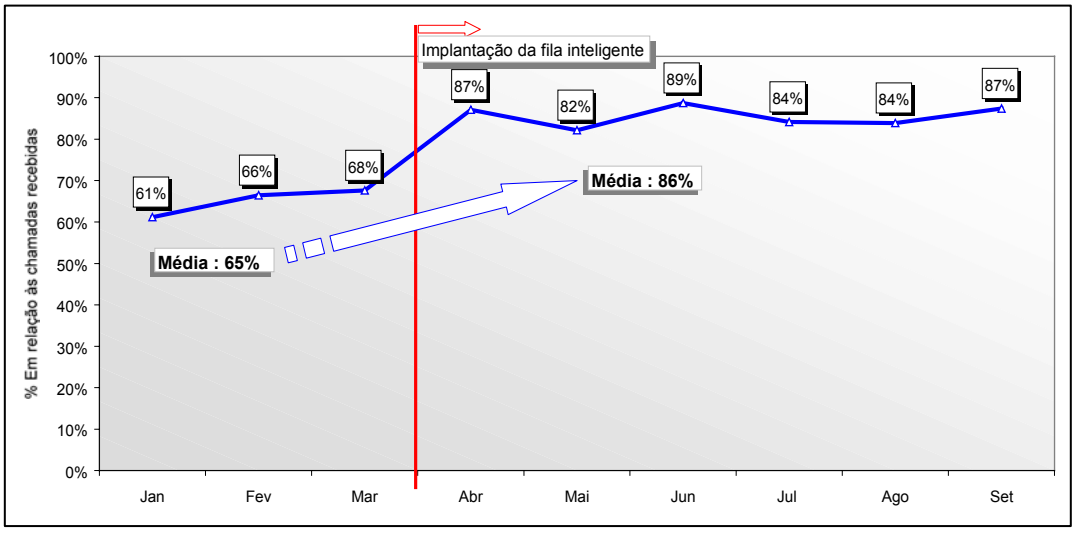

Figura 8 – Gráfico do Answer Rate observado

O Tempo médio de espera em fila diminuiu de 2 minutos e 26 segundos para 47 segundos, em média. Uma diminuição de 67% beneficiando diretamente os Clientes, que passaram a esperar menos em fila. Já o Tempo médio de atendimento, que influi diretamente na força de trabalho necessária para atender a demanda oferecida, diminuiu, em média, de 3 minutos e 18 segundos para 2 minutos e 42 segundos, uma diminuição de 18%.

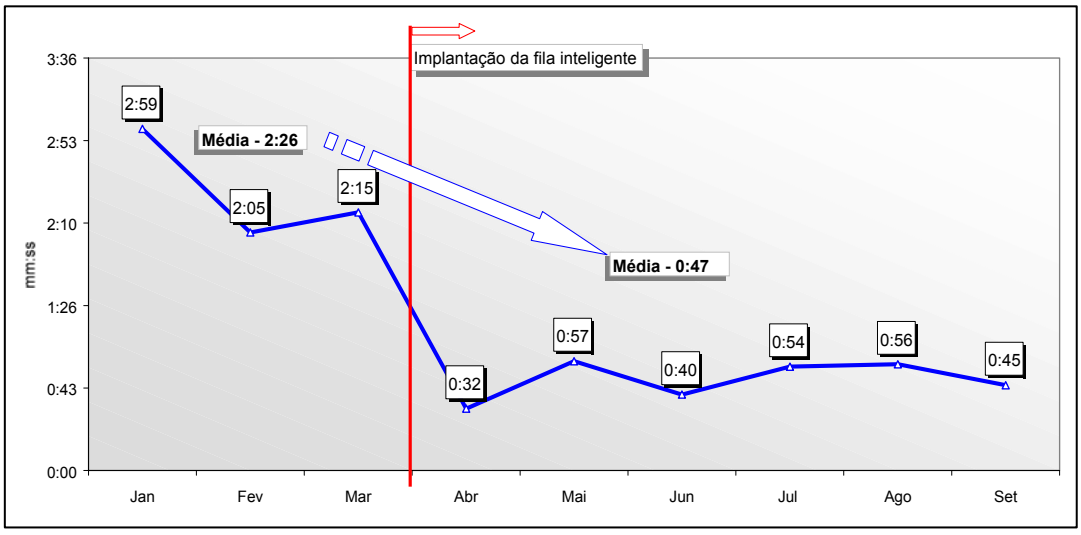

Figura 9 – Gráfico do TME observado

Para que os mesmos resultados fossem obtidos, seria necessário aumentar a força de trabalho, em, aproximadamente, 40 atendentes. Deixando de fazê-lo, foi economizado R$ 1.200.000,00/ano.

4. Conclusões

- Conclui-se que as disciplinas das filas, quando bem administradas, podem trazer reduções significativas nos tempos de espera dos Clientes que por elas passam.
- Foi observado, ainda, que a regra utilizada na fila inteligente para priorizar o atendimento das ligações curtas, traz ganhos imediatos para a operação, tornando o processo estável através do aumento da estabilidade dos indicadores aos picos de demanda.
- Para que a satisfação dos Clientes enquanto nas filas de espera seja garantida, faz-se necessário criar um caminho extra de modo que, depois de um determinado tempo, seja aumentada a prioridade das ligações que esperam há mais tempo em fila.

5. Recomendações

- Recomenda-se criar uma lógica para o início do processo de priorização de modo que este só funcione enquanto a fila estiver em um tamanho tal que o justifique (tempo de espera da ligação mais lenta superior a 20s). Caso contrário, a disciplina da Fila Inteligente migra automaticamente para uma disciplina de fila simples (FIFO).
- Recomenda-se, ainda, que sejam incluídos avisos sobre o tempo médio que o Cliente deve esperar em fila.
- Por fim, juntamente com a disciplina de filas priorizadas criada, sugere-se trabalhar com o modelo chamado *Skill Based Routing*, onde, além das durações médias das chamadas, é levado em consideração o perfil de cada uma e é feito o roteamento diferenciado para os Skills de atendentes. Dessa forma, são criadas estruturas de atendimento compartilhadas, chamadas de multi-skill, que adquirem um maior nível de especialização para cada tipo de assunto e leva à redução do tempo médio de atendimento de todas as ligações. Abaixo pode ser observado o esquema de funcionamento do *Skill Based Routing.*

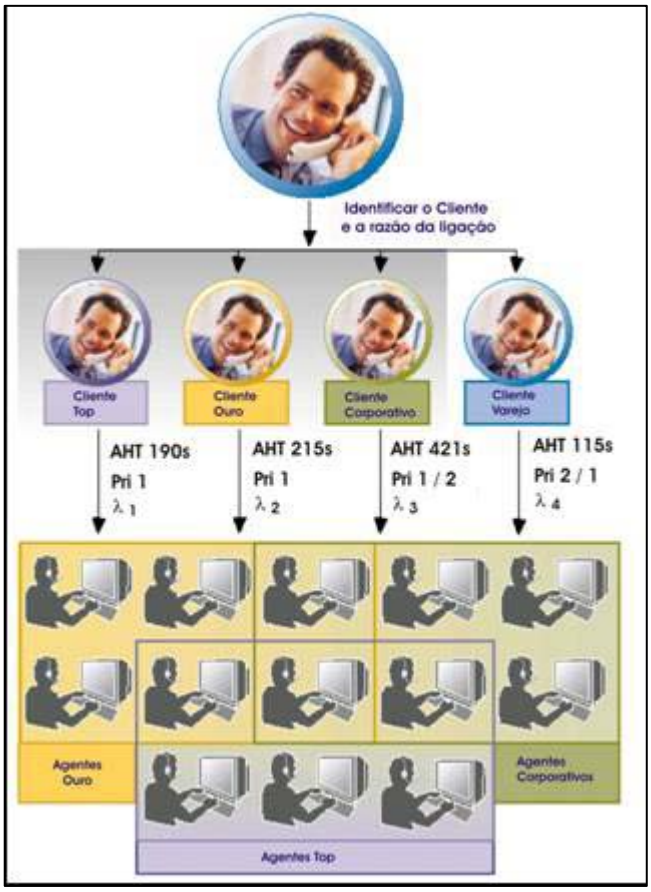

Figura 10 – Esquema de atendimento do *Skill Based Routing*

6. Referências bibliográficas

ARAÚJO, M.; MELO, L.; ANDRADE, T.(2002) - Análise da incidência e prevenção de LER/DORT em centrais de atendimento. Recife.

FILDES, R. (2002) - Telecommunications demand forecasting – a review. Lancaster.

MARQUES M.; PHILIPPI D.; NASCIMENTO G.(2001) - Dimensionamento de Posições de Atendimento para Call Centers. Florianópolis.

MOREIRA, D. (1998) - Administração da Produção e Operações. Ed. Pioneira. Rio de Janeiro.

PRADO, D. (1999) – Teoria das Filas e da simulação. Ed. de Desenvolvimento Gerencial. Belo Horizonte.

ZEITHAML, V.; BITNER, M; (2000) – Marketing de Serviços: A empresa com foco no Cliente. Ed. Bookman. São Paulo.