Distribuição da demanda telefônica de um call center através da criação e priorização de filas inteligentes

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
O crescimento na automação das empresas e os novos serviços criados que permitem a orientação e captação de Clientes via telefone contribuíram para o surgimento das centrais de atendimento. Este mercado 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 das centrais, que têm como meta atingir os resultados esperados com recursos, muitas vezes, escassos, tornando esta área cada vez mais importante neste segmento corporativo. Este trabalho mostra a bem sucedida experiência da implantação de uma disciplina de fila diferente das mais freqüentemente usadas. Esta disciplina que foi apelidada de Fila Inteligente, foi implantada em um Call Center de uma empresa de telecomunicações e, além de ter gerado uma economia de quase R$ 780.000,00/ano, aumentou a satisfação dos seus Clientes.
Palavras-chaves: Telefonia, Teoria das filas, Call Center.

1. Introdução

A empresa em que o projeto foi implantado, atua no segmento de Telecomunicações, com sede em Recife.

A Diretoria de Relacionamento com o Cliente - CRM (Customer Relationship Management) é responsável pelo gerenciamento das interações dos Clientes com a empresa.

Dentre as principais atribuições desta diretoria, está a de elaboração de estratégias voltadas ao atendimento e antecipação das necessidades dos Clientes e *prospects* desta empresa.

Nas estratégias de atendimento às necessidades dos Clientes, se faz necessária a criação de estruturas de Grupos de atendimento e filas inteligentes no PABX de maneira que, com recursos finitos (agentes), se consiga maximizar os resultados de performance (tempo de espera em fila, nível de serviço observado, etc...) e, consequentemente, aumentar a satisfação dos Clientes que ligam para esta Central de atendimento.

Além da diminuição da fila de espera, as estruturas de filas inteligentes do Call Center têm a função de diferenciar os diversos tipos de Clientes, segmentando-os, por exemplo, de acordo com o perfil de consumo (segmentação de marketing), tipo de conta (jurídica ou pessoa física) e adimplência, dando, assim, um tratamento diferenciado para Clientes com perfil diferentes com problemas de diferentes tipos.

2. Filas

Independentemente da sua complexidade, as filas de espera são caracterizadas por Mecanismo de chegadas, Mecanismo do serviço e Disciplina da fila.

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 poderão ser utilizadas as taxas de serviço (μ) e da distribuição, o número de postos de serviço (número de agentes, no nosso caso).

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 para a disciplina da fila é a FIFO (*first in, first out*), na qual o primeiro Cliente a chegar ao inicio da fila é o primeiro a ser atendido. Outras regras existem, desde o também comum LIFO (*last in, first out*) a outras mais complexas baseadas na definição de prioridade e que serão abordadas nesse trabalho.

A teoria das filas tem como objetivo principal o desenvolvimento de modelos matemáticos que nos permitam prever o comportamento de sistemas de prestação de serviços. Para tanto, é preciso manter o sistema em conformidade probabilística, observando cuidadosamente a ordem de chegada e de saída dos Clientes.

As filas estão presentes nos serviços, não fugindo à regra da produção de produtos, tendo o papel de "gargalos". No caso de um call center, as filas poderiam ser representadas pelo tempo de espera do Cliente para ser atendido. Na concepção de Marques e Philippi (2001) é essencial saber quando os momentos de pico vão ocorrer e até que ponto pode ocasionar a espera do Cliente.

Contudo, a natureza dos serviços e sua produção são mais complexas e menos previsíveis que a produção de bens. Segundo Marques e Philippi (2001) há várias formas pelas quais a incerteza estatística ou a variabilidade pode afetar um processo de serviço, podem influenciar tanto a oferta do processo ou a sua demanda. Em um Call Center, por exemplo, poderíamos exemplificar as variabilidades que poderiam distorcer as previsões, do lado da oferta, quando um funcionário tira uma licença médica (ocasiona um número maior de ligações recebidas por agentes - um congestionamento - um tempo maior de espera) e do lado da demanda quando são lançadas promoções de marketing que tendem a elevar a demanda média esperada, problemas técnicos na rede e erros nas contas enviadas para os Clientes que geram uma demanda adicional. 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. Desta forma é muito importante não desconsiderar acontecimentos de certa forma imprevisíveis.

Fildes (2002) reforça a tese de que o desconhecimento de dados relativos a campanhas promocionais, aumenta a dificuldade de previsão do tráfego telefônico a ser recebido no Call Center.

Marques e Philippi (2001) dizem que "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".

3. Componentes da Central de Atendimento

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 CallCenter/Computador. É um dispositivo composto por portas, que, após ser acessado pelo

Cliente irá fornecer automaticamente informações ao mesmo, configurando o que se chama "Auto-Atendimento". Neste dispositivo, normalmente, são dadas, além de informações aos Clientes e opções de saída para que este possa falar com o atendimento pessoal (os agentes). Cada opção de saída deve estar associada a um *VDN*, que pode ser compartilhado com mais de uma opção ou ser exclusivo de cada uma.

- O *Skill* é um grupo virtual ao qual o atendente está conectado. É para estes grupos que as chamadas são roteadas e nela fica enfileirado para posterior atendimento pelos agentes.

- O *VDN* do inglês *Vector Directory Number,* é um ramal virtual (não-físico) utilizado para o roteamento das chamadas. Toda chamada proveniente do PABX está associada a um *VDN* que, por sua vez, está sempre associado a um vetor. O *VDN* funciona como referência na vetorização.

- 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 associadas a este VDN sigam a regra de roteamento presente neste vetor (regra também chamada de vetorização). Pode-se citar como exemplos de comandos utilizados no vetor o *Queue-to Skill 10 pri m*(serve para enfileirar a chamada em um determinado Skill, neste caso, de número 10 e com prioridade média), o *Goto Vec 120* (serve para rotear a chamada para um outro vetor, neste caso de número 120) e o condicional *If Calls Queued in Skill 10 Pri L >10 disconnect after announcement 1254* (é um condicional que serve, neste caso, para checar se existe mais de 10 chamadas com prioridade média enfileiradas no Skill número 10. Se existir, o PABX deve desconectar a chamada atual após tocar o anúncio número 1254). Abaixo se encontra um exemplo de um vetor simples.

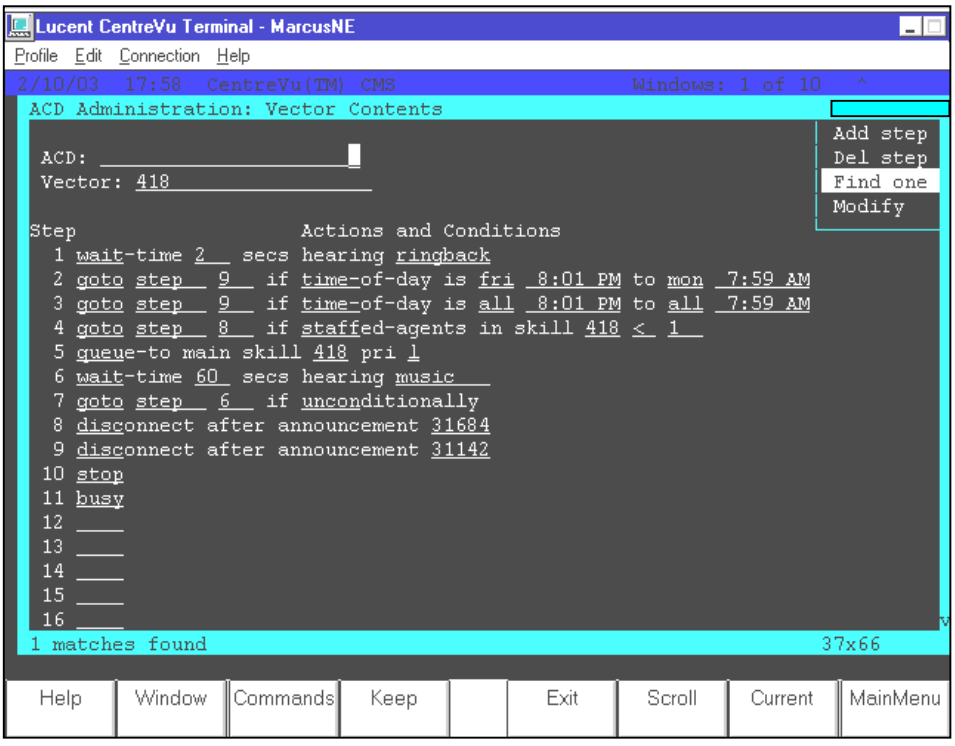

Figura 1 – Tela de exemplo de um vetor

No processo de roteamento das chamadas para um determinado Skill, a prioridade da chamada deve ser atribuída de acordo com a sua importância que será definida na regra de negócio do setor de atendimento.

A priorização do atendimento das chamadas é definida de acordo com o Algoritmo de Distribuição presente no PABX que considera algumas regras. Se as chamadas forem colocadas na fila (nenhum agente livre) e um agente ficar livre, as chamadas serão atendidas nesta ordem:

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

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

4. Criação e priorização das filas

Em um Call Center, a escassez de recursos e a constante cobrança por resultados fazem parte do dia a dia da área responsável pelo Planejamento e Controle da Produção. O Call Center onde este trabalho foi desenvolvido não fugiu à regra.

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 obter-se a satisfação dos Clientes atendidos nesta central. Para que esta satisfação se mantivesse em níveis aceitáveis, se fazia necessária a manutenção da força de trabalho proporcional à demanda recebida, de modo que não existissem filas desnecessárias.

Com o passar do tempo e conseqüente envelhecimento da mão de obra (agentes), um problema crônico começou a aparecer nesta central : As LER/DORTs. Estas doenças de acordo com Araújo, Melo e Andrade (2002), são, respectivamente, 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 : pressão excessiva para os resultados, ambiente excessivamente tenso, rigidez excessiva no sistema de trabalho, estresse emocional, repouso inadequado, o 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 indicadores de desempenho da central, pois, apesar do afastamento dos funcionários por licença, todos continuavam na folha de pagamento da empresa, o que inviabilizava sua substituição.

O dimensionamento da Central de Atendimento, em termos de quantidade de agentes, era feito em função de três variáveis :

a) A demanda prevista em Erlangs que é obtida com a previsão de quantidade de ligações/mês e do tempo médio de atendimento;

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

c) A disciplina/modelo de fila utilizado.

Como a variável de 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 sofreu grandes mudanças e, por isso, inviabilizou um novo dimensionamento, foi o absenteísmo gerado pelas LER/DORTs. Além de não se poderem contratar novos agentes em substituição aos afastados, devido ao número máximo de agentes previsto em orçamento ter

sido atingido, um novo dimensionamento com a premissa de absenteísmo alterada resultava em números irreais que também não seriam aprovados pela diretoria. Um problema estava, então, criado : Como manter (e, se possível, melhorar) os indicadores de desempenho da central dentro de padrões aceitáveis com uma mão de obra reduzida ? Dos conceitos utilizados na Engenharia de Produção veio a idéia : Modificar o modelo de fila utilizado de modo que, criando-se uma fila inteligente onde as chamadas mais rápidas fossem atendidas na frente das mais lentas, conseguíssemos manter os indicadores dentro dos limites esperados, mesmo estando com recursos reduzidos.

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, se fez necessário o mapeamento dos diversos VDNs utilizados na central de atendimento, considerando os tipos de ligações, suas respectivas demandas e tempos médios de duração.

Abaixo se observa uma estrutura de fila com prioridades diferentes.

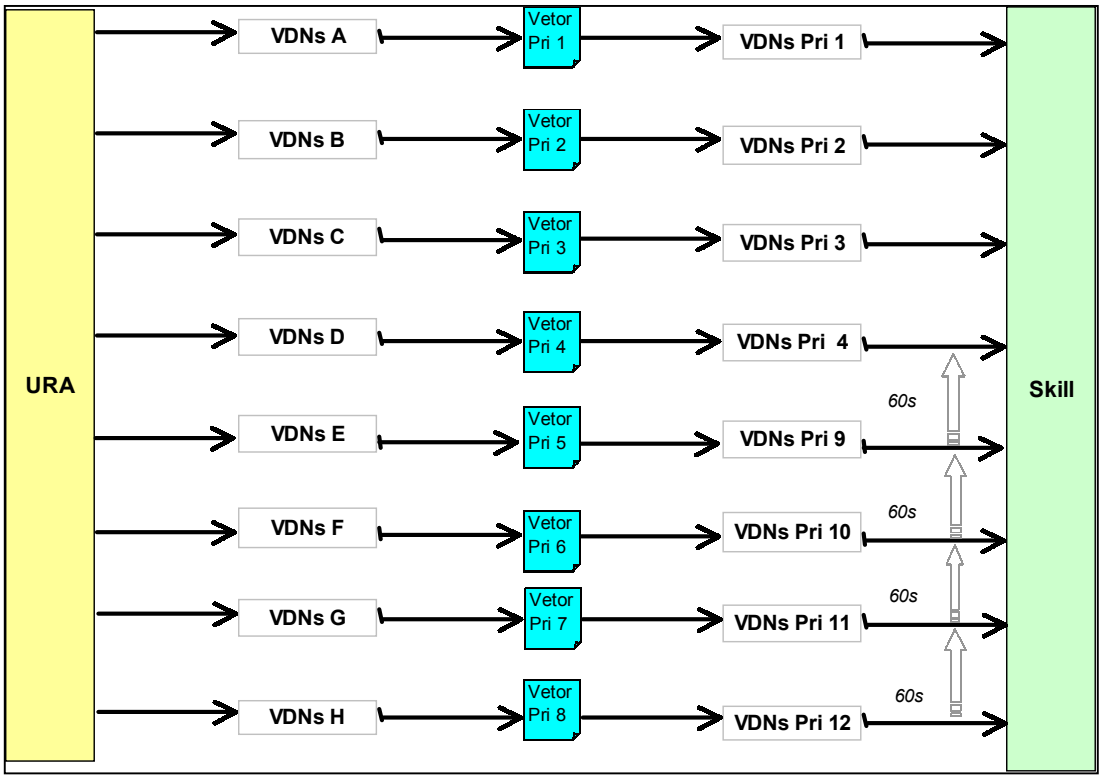

Figura 2– Esquema de roteamento por prioridade em um Call Center

Em uma tabela, coloca-se nas duas primeiras colunas os VDNs e os tempos médios de duração de suas chamadas. Os VDNs estão classificados em ordem crescente de tempo médio, já indicando a associação que deve ser feita : 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.

Na terceira coluna desta tabela é calculado o coeficiente de variação do tempo médio de cada VDN. Este número serve para indicar altas variabilidades nos tempos dos VDNs, auxiliando na escolha da sua prioridade (VDNs com um grande coeficiente de variação, devem estar sendo 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.

Na quarta e na quinta colunas, são calculadas, respectivamente, a quantidade de ligações que cada VDN recebeu no período analisado e a participação percentual destas quantidades em relação ao somatório total das ligações.

Com estes valores em mãos, pode-se calcular as demandas totais e sua participação percentual gerados por estes VDNs para que, na escolha da prioridade de cada um, não haja o risco de sobrecarregarmos as maiores prioridades com grandes demandas, o que poderia uma espera desnecessária para as chamadas associadas aos VDNs com prioridades mais baixas.

A próxima etapa consiste análise e o agrupamento dos VDNs por prioridade e deve levar em consideração a ordem crescente de tempos médios de duração, o coeficiente de variação e a demanda média de cada VDN.

Agrupando-se os VDNs em prioridades, além das regras básicas de priorização descritas anteriormente, também é levada em consideração, em caráter de exceção, a importância de cada VDN. Pode-se exemplificar este tipo de exceção com o VDN que é utilizado pela opção da URA de Suspensão por Perda ou Roubo. As chamadas deste VDN, apesar de, historicamente, apresentarem um tempo médio de duração relativamente alto, têm que ter uma prioridade alta na fila devido à urgência do assunto a ser tratado. Todas as pessoas que escolhem essa opção na URA têm que ter os seus aparelhos suspensos urgentemente para evitar que alguém faça ligações indevidas do mesmo.

Respeitadas as premissas já citadas, quanto menor a variabilidade dentro de cada prioridade melhor. Assim garantimos que em uma mesma prioridade, não estão sendo deixados VDNs com tempos médios muito diferentes, o que infringiria, dentro da prioridade, a lei de “as menores na frente” definida anteriormente. Nesta etapa, são analisados os desvios padrões dos tempos médios dos VDNs de cada prioridade, os agrupando de dois em dois, três em três, quatro em quatro e cinco em cinco. Assim pode-se escolher qual a melhor quantidade de VDNs que se deve ter em cada prioridade: dois, três, quatro, cinco ou mais.

Uma última regra incluída nos vetores foi a criação de um caminho alternativo para as ligações de prioridades baixas que esperavam na fila mais de um determinado tempo limite. Com este caminho alternativo, sempre que uma ligação ultrapassava um tempo de espera em fila considerado alto, esta ganhava uma prioridade mais alta para que fosse atendida de imediato.

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

5. Considerações finais

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 e, por definição, é o percentual das ligações atendidas dentro de um determinado tempo, em relação ao total de ligações atendidas, 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.

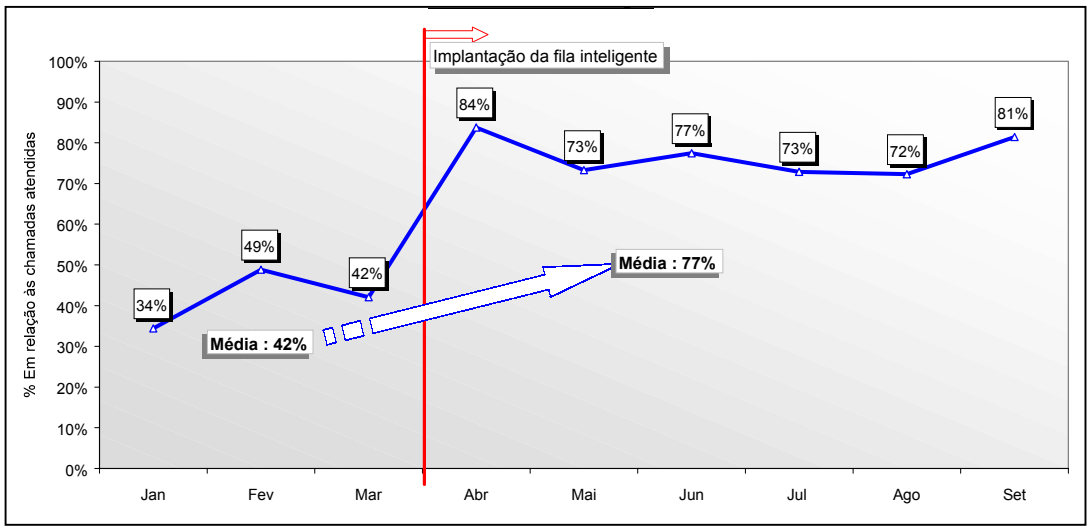

Figura 3 – Gráfico do nível de serviço observado

Um outro surpreendente resultado foi a estabilidade adquirida por este indicador em momentos de picos de demanda. Por mais que tivéssemos 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 no Skill (pelos agentes), 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.

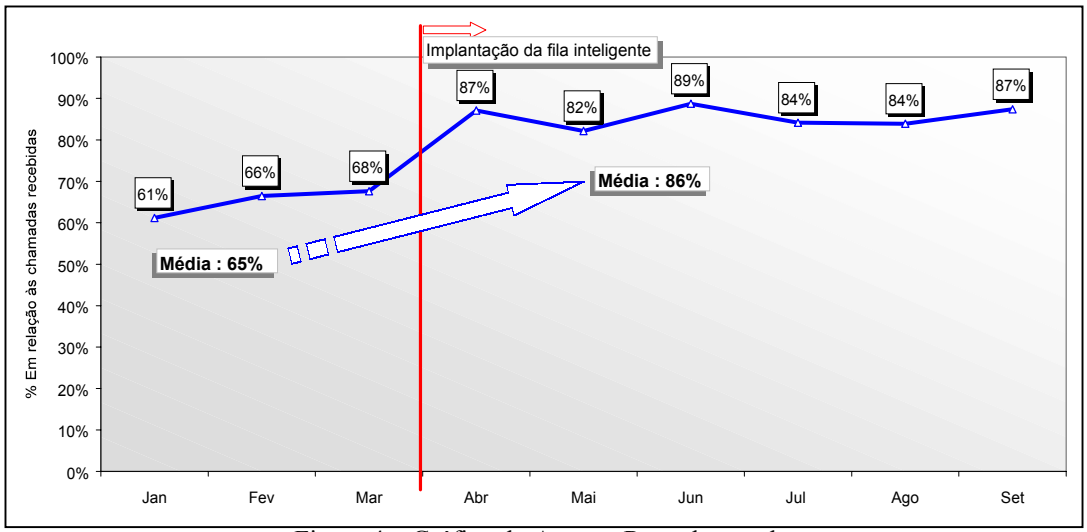

Figura 4 – Gráfico do Answer Rate observado

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%.

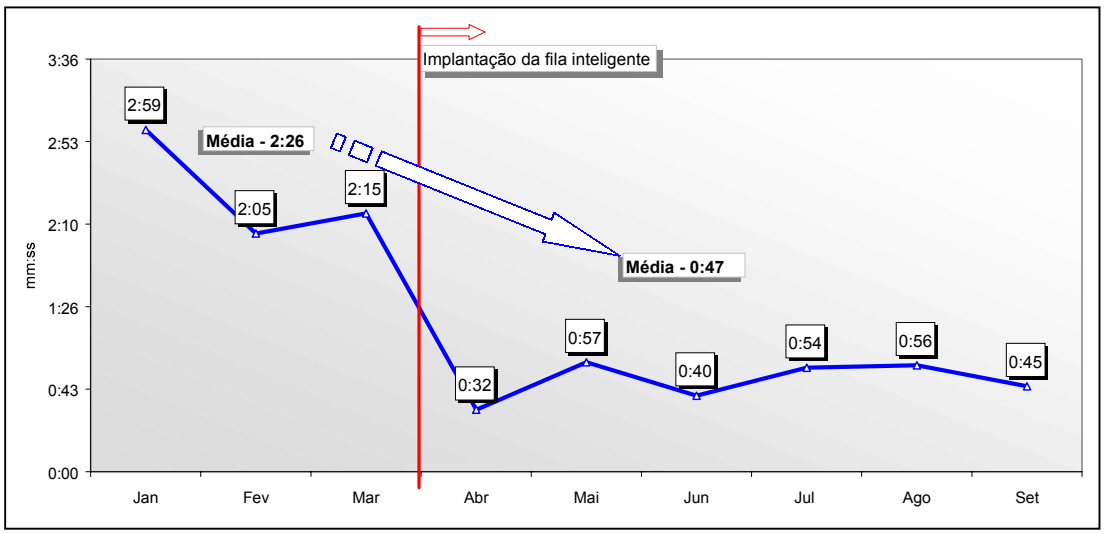

Figura 5 – Gráfico do TME observado

Para os mesmos resultados fossem obtidos, seria necessário aumentar a força de trabalho, em, aproximadamente, 40 agentes. Deixando de fazê-lo, foi economizado R$ 780.000,00/ano.

Conclui-se, então que, as disciplinas das filas, quando bem administradas, podem trazer ganhos significativos em relação ao tempo de atendimento dos Clientes que por elas passam. Já a regra utilizada na fila inteligente, traz ganhos imediatos para a operação, tornando o processo estável e aumentando a imunidade dos indicadores aos picos de demanda.

Para que se garanta a satisfação dos Clientes enquanto nas filas de espera, se faz necessário criar um caminho extra de modo que, depois de um determinado tempo, se aumente a prioridade das ligações com prioridades baixas.

Recomenda-se a criação de uma estrutura de aumento de prioridade em degraus, garantindo que as prioridades planejadas funcionem enquanto houver filas que a justifiquem, para possível migração automática para uma disciplina de fila simples (FIFO). Além disso, juntamente com a disciplina de filas criada, pode-se trabalhar com a disciplina de Skill Based Routing, onde, além das durações médias das chamadas, é levado em consideração os motivos das ligações. Recomenda-se ainda que sejam incluídos de avisos sobre o tempo médio que o Cliente deve esperar em fila.

Referências

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

DAROS, E.; MAY, F.; CAULKINS, K.; OLIVEIRA, R.; ALVES, R. (2000) - Teoria das filas aplicada à rede de Supermercados Angeloni. Florianópolis.

ERDMANN, R. (2000) - Administração da produção: planejamento, programação e Controle. Florianópolis.

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.

MIRANDA, A.; PINTO W.; AMARAL P. (2002) - Como gerenciar as expectativas na prestação de serviços. Rio de janeiro.

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