SISTEMAS & GESTÃO
Revista Eletrônica

GESTÃO DE FILAS ATENDIDAS POR DOIS SERVIDORES COM TAXAS DE ATENDIMENTO DIFERENTES

Fabio Favaretto
fabio.favaretto@unifei.edu.br
Universidade Federal de Itajubá – UNIFEI, Itajubá, Minas Gerais, Brasil

RESUMO

A literatura de Gestão de Filas trata de filas com mais de um servidor em que todos possuem a mesma eficiência e os indicadores utilizados são de difícil obtenção na prática. Em grande parte das situações reais não se observa esta situação, visto que é natural que as pessoas que desempenham o papel de servidores atuem com diferentes rendimentos. O objetivo deste trabalho é apresentar uma forma de gestão de um sistema com dois servidores, que possuem eficiências distintas e os indicadores utilizados são obtidos de forma prática. Foi proposto um método de gerenciamento visual para um sistema de serviços no qual o gestor precisa decidir o momento de abertura de um segundo servidor para atendimento de uma fila e o recurso que irá utilizar, considerando sua eficiência. Os resultados permitem uma gestão prática desta situação, com indicadores de fácil obtenção e um maior controle sobre a necessidade ou não de abertura de novos servidores.

Palavras-chave: Gestão de Filas; Gestão de Serviços; Simulação.

PROPPI / LATEC
DOI: 10.20985/1980-5160.2018.v13n1.825

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

1. INTRODUÇÃO

A gestão da prestação de serviços possui um processo importante e diretamente associado à satisfação dos clientes, que é a gestão de filas (Fitzsimmons *et* Fitzsimmons, 2006). Uma fila se forma em um sistema de prestação de serviços sempre que um ponto de atendimento (chamado de servidor) está atendendo um cliente e existe um ou mais clientes para serem atendidos. Caso se deseje que não ocorra espera ou que esta seja a menor possível, a solução óbvia é a abertura de outro servidor. Por outro lado, um servidor que atenda poucos clientes ou que fique ocioso parte do tempo onera o sistema de prestação de serviços. Assim, é necessário encontrar um equilíbrio entre o rápido atendimento e o custo de operação do sistema.

A Teoria das Filas é um campo de pesquisa quantitativo que propõe a relação entre diversos indicadores. Os indicadores usualmente associados ao nível de satisfação dos clientes são o tempo de espera médio, o tamanho médio da fila e a probabilidade de haver *n* clientes na fila em determinado momento (Morabito *et* Lima, 2000; Chwif *et* Medina, 2006; Bouzada, 2009; Camelo *et al.*, 2010). A obtenção destes indicadores pressupõe uma coleta de dados constante no sistema produtivo, sendo que este processo pode ser complexo e requerer recursos dedicados a este propósito. Em situações práticas, o uso destes indicadores pode ser impedido devido a estas limitações, restando ao gestor a decisão baseada principalmente em seu *feeling* ou em regras desprovidas de embasamento conceitual e/ou quantitativo.

O problema a ser tratado nesta pesquisa é a definição do momento em que um sistema de prestação de serviços, atendido por um único servidor, deve realizar o atendimento de uma fila com um segundo servidor. Mais ainda, entre os recursos humanos disponíveis para operação deste servidor, qual deles deve ser escolhido. Espera-se com estas definições um aumento na qualidade do serviço prestado, mediante a diminuição dos tempos de espera dos clientes e do tamanho das filas.

O objetivo principal deste trabalho é apresentar uma forma de gestão de um sistema com dois servidores que possuem eficiências distintas e os indicadores utilizados são obtidos de forma prática. Para isso, serão propostos alguns indicadores que permitirão ao gestor decidir a conveniência ou não de abertura de um segundo servidor em diferentes condições de atendimento e demanda. As diferentes taxas representam eficiências diferentes entre os recursos humanos disponíveis.

O resultado esperado deste trabalho é um conjunto de implicações gerenciais práticas para os gestores de sistemas com filas. Este trabalho apresenta uma contribuição ao tornar práticos o processo de coleta de dados para a gestão das filas e o uso do ferramental matemático. A estrita aplicação dos métodos quantitativos requer dados como o tempo médio de permanência na fila, e para sua obtenção é necessário acompanhar individualmente os clientes, registrar os momentos de entrada e saída na fila de cada cliente e calcular a média resultante. Assim, a obtenção deste dado em tempo real, principalmente em sistemas de prestação de serviços não informatizados, pode inviabilizar decisões rápidas e eficazes.

Este artigo está estruturado da seguinte forma: após esta introdução, aborda-se um breve referencial conceitual sobre a gestão de filas. Na seção seguinte é feita uma análise dos indicadores utilizados neste tipo de gestão. Na sequência, apresenta-se uma descrição do sistema real que serviu de base para este estudo, assim como a simulação realizada. Por fim, são apresentados os resultados obtidos na simulação, as implicações gerenciais propostas e a conclusão deste trabalho.

2. GESTÃO DE FILAS

A Gestão de Filas é um processo da Gestão de Operações de Serviços. Uma das atividades deste processo é a gestão do indicador do tempo de espera dos clientes (Fitzsimmons *et* Fitzsimmons, 2006). Este indicador influencia a satisfação geral dos clientes de um sistema de serviço. Hwang *et* Lambert (2009) argumentam que em um sistema perfeito não haveria esperas, mas, na prática, é comum os clientes esperarem em uma fila até o recurso estar disponível. Geralmente, as ações dos administradores são concentradas na diminuição do tempo de espera quando o sistema está sobrecarregado com grandes filas, e uma ação comum, neste sentido, é a utilização de servidores adicionais, de acordo com Stolletz *et* Manitz (2013).

De acordo com Jones *et* Peppiat (1996), a primeira opção dos gestores de sistemas de prestação de serviços é projetar uma operação de tal forma que o tempo de espera seja o menor possível, até o ponto em que o aumento de custo é maior que o valor adicionado por esperas pequenas. Outro tipo de ação usada é dividir os clientes em classes diferentes, com filas específicas para atendimento (Alotaibi *et* Liu, 2013).

Houston *et al.* (1998) apresentam uma série de considerações sobre a percepção do tempo de espera em uma fila e a avaliação do serviço feita pelo consumidor. Uma destas considerações é a de que um longo tempo de espera pode influenciar mais em uma avaliação negativa que um mau atendimento. Outra consideração feita pelos autores, colocada para os gestores de serviços, é de que diminuir o tempo de atendimento e a *percepção* do tempo de espera (através de atividades paralelas ou de itens para distração)

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

causa menos impacto que a diminuição de fato do tempo de espera. Quando uma espera longa em fila é necessária, uma explicação sobre o motivo ou alguma forma de pedido de desculpas diminuem a avaliação negativa. Ainda, para estes autores, a percepção de uma espera *desnecessária*, motivada pela visão de funcionários ou recursos parados ou em atividades que não contribuem para o atendimento impacta fortemente na avaliação negativa do serviço prestado.

3. INDICADORES USADOS PARA GESTÃO DE FILAS NESTE TRABALHO

Para atender o objetivo deste trabalho se faz necessário o uso de indicadores para gestão das filas. Pretende-se fornecer aos gestores de sistemas de serviço onde ocorre a formação de filas um conjunto de observações aplicáveis, e, para isso, é necessário que estes indicadores sejam significativos e práticos. Significativos para que com poucos indicadores o gestor tenha um panorama da situação e das possibilidades das filas. Práticos para permitirem sua obtenção facilmente e sem necessidade de grande esforço para coleta e tratamento dos dados.

O estudo quantitativo das filas é baseado nas relações entre alguns parâmetros, principalmente a taxa de chegada ao sistema (λ) e a taxa de atendimento (μ). A relação da taxa de chegada dividida pela taxa de atendimento mostra a ocupação do sistema, representada por ρ (Chwif *et* Medina, 2006). Uma taxa de chegada maior que a taxa de atendimento, mostra claramente que o sistema não terá capacidade de atender as chegadas, gerando filas que nunca diminuem enquanto mantidas as taxas. Diz-se que um sistema está estável ou em estado permanente quando a ocupação do sistema está entre 0 e 0,8 (Chwif *et* Medina, 2006). Um sistema estável pode possuir filas, porém estas permanecerão do mesmo tamanho na pior das situações. No caso de sistemas com mais de um servidor, o quantitativo deles deve ser considerado. A Equação 1 (Chwif *et* Medina, 2006; Fitzsimmons *et* Fitzsimmons, 2006) calcula a ocupação (ρ) do sistema, em que λ é a *taxa média de chegada* (número de chegadas por intervalo de tempo), μ é a *taxa média de atendimento* (número de atendimentos por intervalo de tempo) e *c* é o número de servidores do sistema. Os distintos servidores são considerados todos com a mesma eficiência ou taxa média de atendimento.

$$\rho = \frac{\lambda}{c\mu} \qquad (1)$$

A probabilidade de haver *n* clientes no sistema (P_n) é outro indicador utilizado por Chwif *et* Medina (2006) e Fitzsimmons *et* Fitzsimmons (2006). Este indicador considera todos os clientes no sistema, tanto os em atendimento como os que estão em fila. Como o foco deste trabalho é o gerenciamento das filas, será utilizado um indicador que mostra a probabilidade de não haver fila no momento da chegada de um novo cliente, que equivale às *chegadas atendidas sem espera*. Para que isso aconteça em um sistema com dois servidores, pelo menos um deles deve estar vago. Justifica-se o uso deste indicador por ser de particular interesse do novo cliente, que sempre deseja ser atendido sem espera.

Outro indicador utilizado neste trabalho é o número de clientes na fila no momento da chegada. Chwif *et* Medina (2006) e Fitzsimmons *et* Fitzsimmons (2006) apresentam o indicador *tamanho médio da fila* (L_q), considerado como uma média da quantidade de pessoas na fila em todo o período considerado. A obtenção do indicador L_q a partir de observações do sistema real é pouco prática, pois depende de contagens constantes. O indicador proposto – *tamanho médio da fila no momento da chegada* (L_{qa}) pode ser obtido de forma relativamente mais prática, fazendo contagens do tamanho da fila no momento em que cada novo cliente chega ao sistema. Além disso, a fila no momento da chegada é de particular interesse para o cliente que chega, pois define sua percepção da espera.

A simulação utilizada procura reproduzir um sistema de prestação de serviços onde normalmente existe formação de fila para fazer o pagamento do serviço utilizado em um caixa. Sempre existe um caixa aberto (chamado de primeiro servidor), e é possível a abertura de um segundo caixa. O primeiro servidor é operado por um atendente padrão, treinado para a função. O segundo posto pode ser operado por outro atendente treinado, mas nem sempre existe a disponibilidade de tal recurso. Caso não exista um atendente treinado especificamente para esta função e o gestor do sistema deseje que o segundo posto seja aberto, será utilizado o operador disponível, mesmo que com eficiência inferior ao padrão. A variação na eficiência equivale a um segundo servidor mais ou menos capacitado à execução da tarefa. Um segundo servidor com eficiência 0,5 (ou 50%) equivale a um recurso com menos treinamento ou aptidão, e que gastaria o dobro do tempo em comparação ao servidor de referência exercendo a mesma tarefa. Assim, a disciplina da fila implica que sempre o primeiro servidor é preferencial para atendimento e o segundo servidor só é utilizado quando o primeiro estiver ocupado. Caso os dois servidores estejam livres, a prioridade é do primeiro. Caso os dois servidores estejam ocupados haverá necessariamente uma espera e consequentemente a formação de uma fila.

Pelas considerações feitas anteriormente, será utilizado o indicador de *chegadas atendidas sem espera*, equivalente à probabilidade de um cliente chegar ao sistema e não encontrar fila – com a notação P_0. Este indicador é voltado ao ges-

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

tor para definir um nível de serviço ou nível de atendimento, através do atendimento imediato dos clientes. A formação deste indicador se dá pela proporção (em porcentagem) dos clientes atendidos sem espera pelo total de clientes, conforme visto na Equação 2. É proposto o uso deste indicador por não se tratar de uma média, facilitando sua obtenção.

$$P_0 = \frac{Número\_de\_clientes\_atendidos\_sem\_espera}{Número\_total\_de\_clientes\_atendidos} \quad (2)$$

4. AMBIENTE SIMULADO

A situação simulada retrata o sistema de atendimento de um restaurante universitário. Neste sistema foram coletados dados relativos aos tempos de atendimento e aos intervalos entre as chegadas, que permitiram o cálculo das taxas médias de chegada e de atendimento. Estes dados foram coletados através de observações, cronometragens, fotografias e filmagens, durante os anos de 2013 e 2014. Os valores individuais permitiram analisar a distribuição de probabilidade com melhor aderência e, para ambos os casos, a distribuição sugerida (Morabito *et* Lima, 2000) foi a exponencial de probabilidade. As observações foram estratificadas para cobrirem diversos dias da semana, períodos do mês e meses diferentes. Isso permitiu obter médias representativas de todo o período.

O tempo médio de atendimento medido foi de 30 segundos, o que equivale a uma taxa média de atendimento (μ) de 0,033 clientes/segundo. Este tempo não apresentou variações significativas em situações distintas. O intervalo entre as chegadas apresenta grandes variações ao longo do dia, relacionadas aos intervalos entre aulas, períodos de refeições e início e término de expediente de funcionários da universidade. Analisando-se os dados, podem-se distinguir períodos com diferente nível de movimentação, resultando em três situações: pequeno, médio e grande movimento do restaurante. Os intervalos entre as chegadas e as taxas médias de chegada são apresentados na Tabela 1, na qual também é apresentada a ocupação do sistema (ρ) para cada situação, considerando-se apenas um servidor ativo e a taxa média de atendimento apresentada acima. Estes dados foram obtidos em observações presenciais e com o uso de contagens e cronometragens. O período analisado compreende os meses entre abril e outubro de 2014. Além das observações presenciais, o autor obteve dados observando os vídeos diários do circuito de segurança, disponibilizados pela administração.

Na Tabela 1 são definidas as situações de movimento utilizadas neste trabalho. A situação de grande movimento corresponde aos horários de pico, correspondente ao horário de almoço e dos intervalos de aulas. Nestes horários ocorrem as maiores filas, sendo que o intervalo médio de chegadas é de 30 segundos. Além disso, a ocupação do sistema, caso seja utilizado apenas um servidor para atendimento, é de 1 (um), o que significa um sistema instável.

Nos horários de menor movimento raramente ocorrem filas, pois o intervalo de 120 segundos entre as chegadas praticamente garante o atendimento imediato do sistema. A ocupação do sistema em 0,25 retrata esta situação.

5. SIMULAÇÃO

Foram criados três cenários de simulação, um para cada uma das situações de movimento apresentadas na Tabela 1. Para cada cenário, foram feitas simulações considerando-se um ou dois servidores, e nas situações com dois servidores a eficiência será variada. A variação na eficiência do segundo servidor foi de zero (equivalente a um único servidor) a 1 (equivalente a um segundo servidor com a mesma eficiência do servidor de referência), com incremento de 0,1.

A *ocupação* (ρ) do sistema foi calculada através da Equação 3, em que λ é a *taxa média de chagada* (número de chegadas por intervalo de tempo) e μ é a *taxa média de atendimento* (número de chegadas por intervalo de tempo). O denominador é uma adaptação da Equação 1 apresentada anteriormente, na qual a eficiência do primeiro servidor é igual a 1, por se tratar da referência, e a eficiência do segundo servidor (em porcentagem) é representada por E_2.

$$\rho = \frac{\lambda}{(1+E_2)\mu} \quad (3)$$

Tabela 1. Dados do sistema analisado

Situação de movimento	Intervalo entre as chegadas (segundos)	Taxa média de chegada (chegadas/segundos)	Ocupação do sistema com um servidor (ρ)
Grande	30	0,033	1
Médio	50	0,02	0,6
Pequeno	120	0,00833	0,25

Fonte: O próprio autor

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

Cada cenário foi simulado com 5000 replicações, utilizando-se o software Crystal Ball, e os resultados estão apresentados a seguir. Cada replicação mostra um período seguido de uma hora em cada uma das situações de movimento e considera que a taxa média de chegada segue a Distribuição Exponencial de Probabilidade. A variável independente utilizada foi a eficiência do segundo servidor. Em cada cenário foi feita uma variação da eficiência do segundo servidor, com valores entre zero (quando somente o primeiro servidor está disponível) e 1 (quando o segundo servidor é igual ao padrão), com variação de 0,1.

As variáveis dependentes analisadas são apresentadas na sequência. A primeira variável analisada é a espera média em segundos, que mostra a média dos tempos de espera de todos os usuários. Esta variável é utilizada apenas como referência, principalmente para o gestor do sistema. A segunda variável utilizada é a probabilidade de o usuário ser atendido imediatamente após sua chegada ao sistema ou encontrar uma fila de tamanho zero (P_0), conforme a Equação 2. A última variável dependente analisada é o tamanho médio da fila no momento da chegada, que considera uma média do tamanho da fila existente no momento da chegada de cada usuário.

6. RESULTADOS

A seguir, são apresentados os resultados obtidos com as simulações. Em todas elas se considerou a taxa de atendimento (μ) do primeiro servidor constante, com valor de 0,033 clientes/segundo.

Na situação de movimento pequeno, a taxa média de chegadas (λ) é uma variável aleatória que segue a Distribuição Exponencial com taxa 0,0083 chegadas/segundo. Os resultados obtidos para este cenário são apresentados na Tabela 2.

Tabela 2. Resultados da simulação para o cenário de movimento pequeno.

Número de servidores	Eficiência do segundo servidor	Espera média (segundos)	P0	Tamanho médio da fila na chegada
1	0	9,8	0,75	0,08
2	0,1	3,56	0,9	0,03
2	0,2	2,17	0,93	0,02
2	0,3	1,6	0,94	0,01
2	0,4	1,26	0,95	0,01
2	0,5	1,04	0,96	0,01
2	0,6	0,85	0,96	0,01
2	0,7	0,72	0,97	0,01
2	0,8	0,63	0,97	0,01
2	0,9	0,55	0,97	0,01
2	1	0,48	0,97	0,01

Fonte: O próprio autor

Com o atendimento de apenas um servidor, a espera média foi de 9,8 segundos e a probabilidade de não haver fila no momento da chegada de um novo usuário (P_0) é de 0,75 ou 75%. Com a adição de um segundo servidor, mesmo que com somente 0,1 ou 10% da eficiência do primeiro, existe uma redução significativa do tempo médio de espera que passa a 3,56 segundos. A P_0 é aumentada para 0,9 ou 90%. A espera média em função da eficiência do segundo servidor pode ser observada na Figura 1. A espera média é reduzida sensivelmente com o aumento da eficiência do segundo servidor até valores próximos a 0,5 ou 50%. A partir deste ponto, o aumento da eficiência do segundo servidor permite uma pequena redução no tempo de espera média.

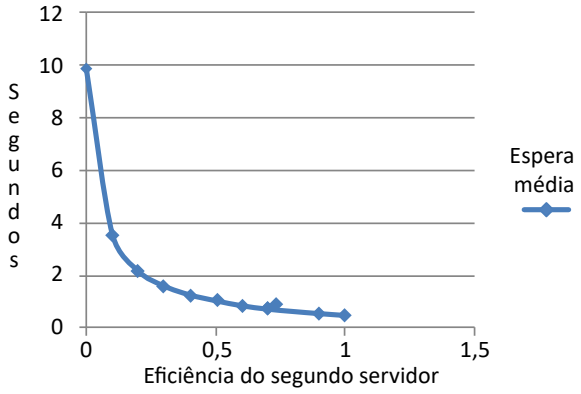

Figura 1. Relação entre tempo médio de espera e eficiência do segundo servidor para o caso de pequeno movimento.
Fonte: O próprio autor

Em uma situação de pequeno movimento, a espera para ser atendido com apenas um servidor não é significativamente alta, sendo na ordem de 10 segundos. Mesmo assim, a abertura de um segundo servidor com metade da eficiência do primeiro servidor pode levar a tempos de espera médios de aproximadamente 1 segundo e a 96% de chance de não haver fila no momento da chegada de um novo usuário.

Na situação de movimento médio, a taxa média de chegadas (λ) é uma variável aleatória que segue a Distribuição Exponencial com taxa 0,02 chegadas/segundo. Os resultados obtidos na simulação podem ser vistos na Tabela 3.

Com o atendimento de apenas um servidor, a espera média foi de 40,44 segundos e a probabilidade de não haver fila na chegada de um novo usuário (P_0) é de 0,42 ou 42%. Com a adição de um segundo servidor com 0,1 ou 10% da eficiência do primeiro, existe uma redução do tempo médio de espera que passa a 25,44 segundos. A P_0 é aumentada para 0,57 ou 57%. A espera média em função da eficiência do segundo servidor pode ser observada na Figura 2.

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

Tabela 3. Resultados da simulação de movimento médio.

Número de servidores	Eficiência do segundo servidor	Espera média (segundos)	P0	Tamanho médio da fila na chegada
1	0	40,44	0,42	0,7
2	0,1	25,24	0,57	0,49
2	0,2	16,9	0,66	0,33
2	0,3	12,14	0,71	0,24
2	0,4	9,18	0,75	0,18
2	0,5	7,19	0,78	0,14
2	0,6	5,7	0,81	0,11
2	0,7	4,76	0,82	0,09
2	0,8	4,05	0,84	0,08
2	0,9	3,42	0,85	0,07
2	1	3	0,86	0,06

Fonte: O próprio autor

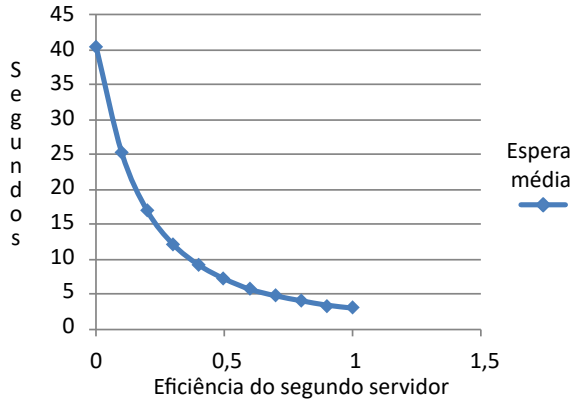

Figura 2. Relação entre tempo médio de espera e eficiência do segundo servidor para o caso de médio movimento.
Fonte: O próprio autor

Na situação de grande movimento, a taxa média de chegadas (λ) é uma variável aleatória que segue a Distribuição Exponencial com taxa 0,033 chegadas/segundo, correspondendo à chegada, em média, de um cliente a cada 30 segundos. Como o tempo médio de atendimento é da mesma ordem, tem-se que a ocupação do sistema (ρ) é igual a 1, o que indica um sistema instável com os tamanhos de fila tendendo ao infinito. Esta é a situação que requer maior atenção. Em algumas situações observadas no sistema real, chegou-se a filas de mais de 40 pessoas com tempos de espera superiores a 15 minutos. Os resultados obtidos na simulação deste cenário são apresentados na Tabela 4.

Tabela 4. Resultados da simulação de movimento grande.

Número de servidores	Eficiência do segundo servidor	Espera média (segundos)	P0	Tamanho médio da fila na chegada
1	0	198,7	0,11	5,35
2	0,1	128,6	0,19	3,88
2	0,2	81,35	0,28	2,54
2	0,3	54,34	0,36	1,73
2	0,4	39,18	0,43	1,26
2	0,5	28,68	0,49	0,93
2	0,6	22	0,54	0,71
2	0,7	17,55	0,58	0,57
2	0,8	14,21	0,62	0,47
2	0,9	11,39	0,65	0,37
2	1	9,49	0,67	0,31

Fonte: O próprio autor

Com o atendimento de apenas um servidor, a espera média foi de 198,7 segundos e a probabilidade de não haver fila na chegada de um novo usuário (P_0) é de 0,11 ou 11%. Com a adição de um segundo servidor com 0,1 ou 10% da eficiência do primeiro, existe uma redução do tempo médio de espera que passa a 128,6 segundos. A P_0 é aumentada para 0,19 ou 19%. A espera média em função da eficiência do segundo servidor pode ser observada na Figura 3.

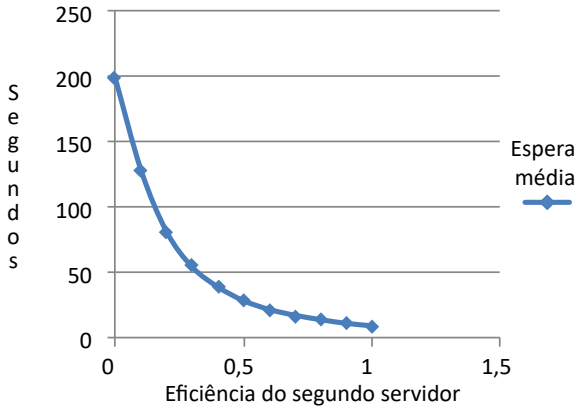

Figura 3. Relação entre tempo médio de espera e eficiência do segundo servidor para o caso de grande movimento.

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

7. IMPLICAÇÕES GERENCIAIS

Conforme observado na Tabela 4, a situação mais crítica é no caso de grande movimento, em que somente existe um servidor em atendimento. Neste caso, espera-se uma fila média de 5,35 clientes que, para efeitos práticos, é aproximada para 6 clientes. Assim, quando a fila *média* atinge esta marca e o sistema está no período de grande movimento, a ocupação do sistema (ρ) é igual a 1 e isso significa que o sistema está instável. Assim, a existência de 6 clientes, em média, na fila marca a necessidade de abertura de um novo servidor. Para isso, pode ser feita uma delimitação da fila utilizando-se postes e correntes que forçam a formação de uma fila em linha reta onde necessariamente um cliente fica atrás do cliente que chegou imediatamente antes dele. Com uma avaliação da distância ocupada pelos clientes nesta fila, pode ser feita uma marca no chão que deve estar localizada entre o sexto e o sétimo cliente da fila. Enquanto a fila estiver antes do final da marca, significa que existem seis ou menos clientes. Quando a fila ultrapassa a marca é porque está na fila o sétimo cliente. Por se tratar de uma decisão baseada no tamanho médio da fila, ela pode eventualmente ultrapassar este limite em situações com $\rho < 1$. Por este motivo, o gestor precisa observar a fila durante algum tempo ou por algumas vezes. Caso se confirme uma fila com mais de 7 clientes, é necessária a abertura de um novo servidor.

A abertura de um segundo servidor pode ser feita utilizando-se o recurso humano mais eficiente disponível. A menor eficiência observada entre os recursos humanos disponíveis no sistema analisado é da ordem de 20% (0,2) da eficiência do servidor padrão. Neste caso, observando-se a Tabela 4, verifica-se que o tamanho médio da fila esperado é de 2,54 clientes, sendo utilizado o valor 3 clientes. Deve ser feita outra marcação na fila entre o terceiro e quarto cliente. Quando o sistema opera com dois servidores e esta marca é ultrapassada, significa que é necessário um servidor com maior eficiência. Neste caso, o gestor deve substituir o segundo recurso humano por um com maior eficiência.

8. CONCLUSÕES

Este trabalho apresentou um cálculo da taxa de ocupação (ρ) de um sistema de prestação de serviços com dois servidores onde a eficiência destes não é necessariamente igual. Esta é uma contribuição para a área, visto que o problema com dois servidores é amplamente tratado considerando os servidores com a mesma eficiência. Ao estudar e compreender um sistema real, foi criado um modelo de simulação computacional para reproduzir casos de pequeno, médio e grande movimento. Os resultados obtidos permitiram compreender a situação na qual é necessária a abertura de um segundo servidor em situações de grande movimento. Foi apresentada uma regra prática para que o gestor do sistema identifique facilmente esta situação. Também foi apresentada uma regra prática para identificar se é necessário substituir o segundo servidor por um com maior eficiência. Conclui-se que é possível o desenvolvimento de ações gerenciais práticas eficazes ao se utilizar a aplicação do conhecimento de Teoria das Filas, garantindo a agilidade nas decisões necessária ao gestor do sistema.

AGRADECIMENTOS

O autor agradece os apoios recebidos do CNPq e da FAPEMIG.

REFERÊNCIAS

Alotaibi, Y., Liu, F. (2013), “Average waiting time of customers in a new queue system with different classes”, Business Process Management Journal, Vol. 19, No. 1, pp. 146-68.

Bouzada, M. A. C. (2009) “Dimensionamento de um call center: simulação ou Teoria das Filas?”, Anais... SIMPOI 2009: Simpósio de Administração da Produção, São Paulo, SP.

Camelo, G. R., Coelho, A. S. et al. (2010) “Teoria das filas e da simulação aplicada ao embarque de minério de ferro e manganês no terminal marítimo de ponta da madeira”, Cadernos do IME, Vol. 29, disponível em: http://www.e-publicacoes.uerj.br/index.php/cadest/article/view/15733/11904 (acesso em 18 jan. 2018).

Chwif, L., Medina, A. C. (2006) “Uma análise crítica da Lei Municipal 13.948 ou ‘Lei das Filas’ sob a ótica da Pesquisa Operacional: conclusões derivadas de modelos de simulação de eventos discretos”, Anais... XXVI ENEGEP – Encontro Nacional de Engenharia de Produção, Fortaleza, CE, 2006.

Fitzsimmons, J. A., Fitzsimmons, M. J. (2006), Service management: Operations, strategy, and information technology, 5th ed., McGraw Hill, New York.

Hwang, J., Lambert, C. U. (2009), “The use of acceptable customer waiting times for capacity management in a multistage restaurant”, Journal of Hospitality & Tourism Research, Vol. 33, No. 4, pp. 547-61.

Houston, M. B., Bettencourt, L. A. et al. (1998), “The relationship between waiting time in a service queue and evaluations of service quality: a field theory perspective”, Psychology & Marketing, Vol. 15, No. 8, pp. 735-753.

Revista Eletrônica Sistemas & Gestão
Volume 13, Número 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

Jones, P., Peppiat, E. (1996), "Managing perceptions of waiting times in service queues", International Journal of Service Industry Management, Vol. 7, No. 5, pp. 47-61.

Morabito, R., Lima, F. C. R. (2000) "Um modelo para analisar o problema de filas em caixas de supermercados: um estudo de caso", Pesquisa Operacional, Vol. 20, No. 1, pp. 59-71.

Stolletz, R., Manitz, M. (2013), "The impact of a waiting--time threshold in overflow systems with impatient customers", Omega, Vol. 41, No. 2, pp. 280-86.

Recebido: 10 set. 2014

Aprovado: 15 jan. 2018

DOI: 10.20985/1980-5160.2018.v13n1.825

Como citar: Favaretto, F. (2018), "Gestão de filas atendidas com taxas de atendimento diferentes", Sistemas & Gestão, Vol. 13, No. 1, pp. 2-9, disponível em: http://www.revistasg.uff.br/index.php/sg/article/view/825 (acesso dia mês abreviado. ano).