FATORES INFLUENTES NA ALOCAÇÃO CONJUNTA DE SERVIDORES E ÁREAS DE ESPERA EM REDES DE FILAS FINITAS

Helgem S. R. Martins
Departamento de Estatística, Universidade Federal de Ouro Preto
35400-000 – Ouro Preto – MG, Brazil
helgem.souza@gmail.com

Frederico R. B. Cruz
Departamento de Estatística, Universidade Federal de Minas Gerais
31270-901 – Belo Horizonte – MG, Brazil
fcruz@est.ufmg.br

Anderson R. Duarte, Fernando L. P. Oliveira
Departamento de Estatística, Universidade Federal de Ouro Preto
35400-000 – Ouro Preto – MG, Brazil
anderson@iceb.ufop.br,fernandoluiz@iceb.ufop.br

RESUMO

O problema de alocação conjunta de servidores e áreas de espera (BCAP) é um problema de otimização não-linear inteira que visa obter uma configuração ótima em redes de filas que garanta um limiar mínimo de desempenho pré-estabelecido. Este artigo enfoca uma metodologia para solução do BCAP em redes de filas finitas markovianas, ou redes de filas *M*/*M*/*c*/*K*, na notação de Kendall, que consiste em combinar métodos aproximados com o algoritmo de Powell, um algoritmo de otimização livre de derivadas. Tal metodologia foi aplicada a redes nas topologias básicas série, divisão e fusão. Os resultados apresentados indicam robustez e homogeneidade das soluções. Foram analisados fatores influentes nos padrões de alocação.

PALAVRAS CHAVE. Alocação. Filas finitas. Redes de filas

Tópicos. MP – Modelos Probabilísticos. IND – PO na Indústria

ABSTRACT

The joint buffer and server optimization problem (BCAP) is a non-linear integer optimization problem that aims at obtaining an optimal configuration of a queueing network such that the resulting throughput is greater than a pre-defined threshold. This paper focuses on a methodology designed to solve the BCAP, for networks of finite Markovian queues, or in Kendall notation, networks of *M*/*M*/*c*/*K* queues, which consists in a combination of approximate methods and Powell algorithm, a derivative-free optimization algorithm. The methodology was applied to networks of queues in series, split, and merge basic topologies. The results produced robust and homogeneous solutions. Factors influencing the allocation patterns were investigated.

KEYWORDS. Allocation. Finite queues. Queueing networks.

Paper topics. MP – Probabilistic models. IND – OR in industry

1. Introdução

Um dos grandes interesses ao modelar sistemas por filas é encontrar uma configuração ótima, de servidores e áreas de espera, que atenda aos requisitos de desempenho e minimize o custo envolvido. Dado que a grande maioria dos sistemas possui limitação de recursos, tanto de servidores quanto de áreas de espera, que geram impactos financeiros direto nos custos do processo, é importante seu uso racional.

Neste contexto, podemos destacar o *problema de alocação conjunta de servidores e áreas de espera*, conhecido como BCAP (do inglês ***B**uffer and Server (**C**) **A**llocation **P**roblem*). Para redes de filas finitas markovianas, ou redes de filas *M*/*M*/*c*/*K*, na notação de Kendall [Kendall, 1959], é um problema desafiador de otimização em redes de filas [van Woensel et al., 2010; Cruz & van Woensel, 2014a,b], cuja metodologia de solução pode ser aplicada a diversas situações reais modeladas na forma de filas ou redes de filas, tais como linhas de produção, serviços telefônicos, filas de estabelecimentos (bancos, supermercados), sistemas computadorizados, *webshops*, dentre outras aplicações análogas. Sua principal característica é a relação intrínseca existente entre o custo dos servidores e áreas de espera. Conhecer uma configuração ótima desta relação de custo permite o desenvolvimento de redes de filas cujo serviço seja eficiente e o desempenho máximo possa ser obtido.

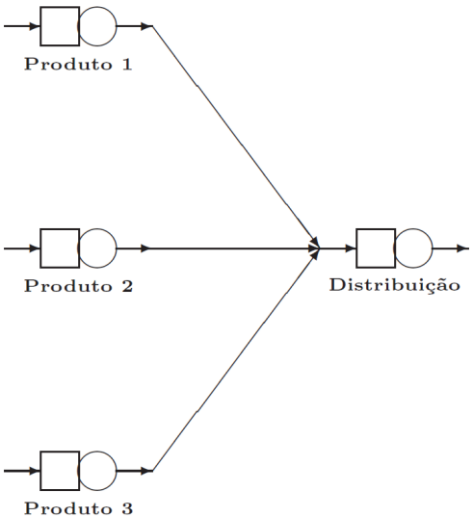


Figura 1: Exemplo de rede de filas na produção de alimentos congelados

Exemplo 1: *Uma empresa de gêneros alimentícios irá iniciar a produção de três variedades de alimentos congelados (Fig. 1). Para tanto, ela receberá as matérias primas processadas e resfriadas, que serão armazenadas em refrigeradores até o momento da produção. Após a montagem do produto, eles serão então congelados e encaminhados até a distribuição para os clientes. Para armazenar as matérias primas e os produtos finais serão necessárias câmaras frigoríficas, cujos custos de fabricação e de funcionamento aumentam conforme seu tamanho. A distribuição também é realizada por caminhões frigoríficos, que assim como as câmaras frigoríficas, são bastante caros. Conhecidas as taxas de entrada das matérias primas, qual seria a configuração de tamanho de câmara frigorífica e o número de caminhões que otimizaria a produção e o custo envolvido?*

No problema apresentado, Fig. 1, uma proposta inadequada de dimensionamento das áreas de espera (câmaras frigoríficas) ou do número de servidores (linhas de montagem e caminhões frigoríficos) poderia levar à perda significativa da eficiência e rentabilidade do processo, dados os altos custos envolvidos. Em casos como estes, a utilização de métodos que otimizem simultaneamente o número de servidores e a capacidade das áreas de espera se faz necessária.

O *problema de alocação conjunta de servidores e áreas de espera* (BCAP) é um problema de difícil solução do ponto de vista computacional, por se tratar de um problema de programação não-linear, cuja função-objetivo não apresenta forma fechada e consequentemente requer métodos aproximados em sua solução. A solução deste problema se torna ainda mais complexa quando o objeto de interesse deixa de ser um sistema de filas simples e se torna uma rede de filas complexa, como no caso do exemplo apresentado.

Na seção 2 é apresentado o modelo de programação matemática inteira do *problema de alocação conjunta de servidores e áreas de espera* (BCAP). Na seção 3 são descritos brevemente os algoritmos empregados na resolução do BCAP, com destaque para o *Método da Expansão Generalizado* (GEM), para estimação das medidas de desempenho, e o método de otimização

irrestrita de Powell. Na seção 4 são apresentados resultados obtidos em redes de filas nas topologias básicas *série*, *divisão* e *fusão*. São também analisados os padrões de alocação para cada topologia e os fatores que impactam na alocação ótima de servidores e de áreas de espera. Para concluir, na seção 5 serão discutidos os resultados obtidos e apresentadas propostas para continuação dos estudos na área de otimização em redes de filas finitas.

2 O Problema de Alocação Conjunta de Servidores e Áreas de Espera (BCAP)

O Exemplo 1 apresenta uma cenário real simplificado que ilustra uma situação em que não basta efetuar a otimização isolada do número de servidores ou do tamanho das áreas de espera, para que seja garantido um funcionamento ótimo da produção. Deve-se considerar conjuntamente a melhor configuração entre servidores e áreas de espera para garantir um bom funcionamento do sistema. De situações semelhantes, emerge a necessidade de solucionar o *problema de alocação conjunta de servidores e áreas de espera* (BCAP) [van Woensel et al., 2010; Cruz & van Woensel, 2014a,b]. De maneira simplificada, o BCAP pode ser descrito como o problema de determinação de quantos servidores e qual capacidade de áreas de espera serão tornadas disponíveis, de modo a garantir que o desempenho do sistema, em termos da taxa de atendimento, θ, seja superior a um limiar mínimo de aceitação pré-estabelecido, $\theta_{\min}$. Deverá, para isso, ser considerada a relação existente entre os custos dos servidores e das áreas de espera, representada pela razão C_c/C_B.

2.1 Formulação Matemática

Defina uma rede de filas $M/M/c/K$ como um grafo $G = (N,A,\mathbf{p})$, em que N é o conjunto de todos os nós (cada fila $M/M/c/K$) que compõe a rede, A é o conjunto de arcos que interconectam pares de nós (duas filas $M/M/c/K$) e $\mathbf{p}$ é o vetor das respectivas probabilidades de roteamento nos arcos, conforme apresentado na Fig. 2.

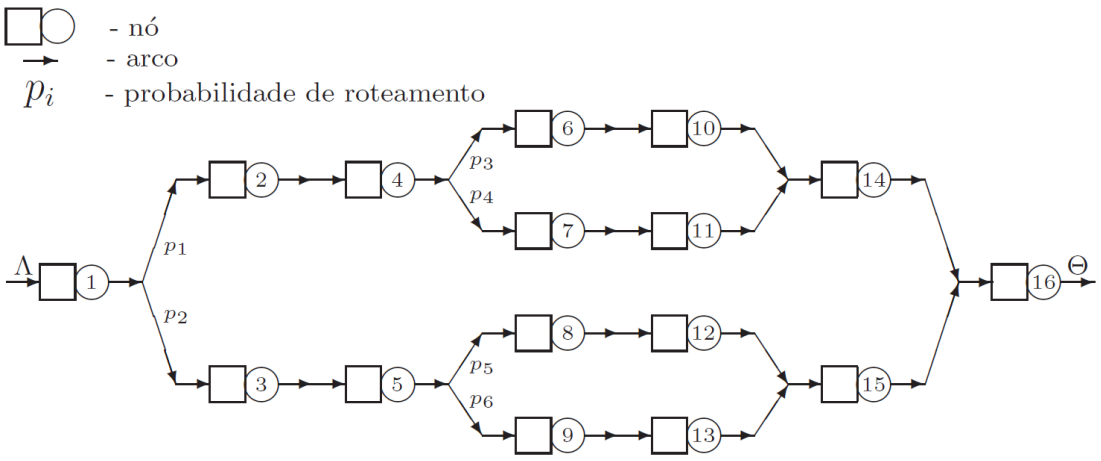


Figura 2: Grafo representativo de uma rede de filas $M/M/c/K$

O BCAP busca minimizar o tamanho das áreas de espera e o número de servidores, de forma que a taxa de atendimento resultante, θ, seja maior que um limiar predefinido. O BCAP pode ser definido matematicamente pela seguinte formulação de programação matemática inteira não-linear [van Woensel et al., 2010].
BCAP:

$$Z = \min\left[\sum_{\forall i \in N} \varpi_i c_i + \sum_{\forall i \in N} (1-\varpi_i) B_i\right], \tag{2.1}$$

sujeito a

$$\theta(\mathbf{c},\mathbf{B}) \geq \theta_{\min}, \tag{2.2}$$

$$c_i \in \{1,2,\ldots\}, \forall i \cup N, \tag{2.3}$$

$$B_i \in \{0,1,\ldots\}, \forall i \cup N, \tag{2.4}$$

em que c_i é o número de servidores no nó i, B_i representa o tamanho das áreas de espera, ou seja, é o capacidade total da fila, K, excluídos os itens que se encontram em atendimento (isto é, para o nó i, a área de espera é dada por $B_i = K_i - c_i$), $\theta(\mathbf{c},\mathbf{B})$ é a taxa de saída resultante, que é dada em função dos servidores (**c**) e das áreas de espera (**B**), $\theta_{\min}$ é o limiar mínimo aceitável para a taxa de atendimento resultante e ω_i é uma variável que representa o custo relativo entre servidores e áreas de espera ($0 \leq \omega_i \leq 1$).

O valor de ω_i pode ser alterado para refletir o custo de servidores em comparação ao custo das áreas de espera. Quando o valor de ω_i diminui, o custo dos servidores se torna relativamente menor se comparado ao custo das áreas de espera. Por outro lado, se o valor de ω_i sofre acréscimos, os servidores se tornam relativamente mais caros se comparados às áreas de espera. Variando-se os valores possíveis de ω_i, é possível verificar as alterações resultantes nas alocações de servidores e nas áreas de espera. Outra possibilidade que emerge da relação de custos é a verificação de quais valores de ω_i tornam significativas as alterações em alocações de servidores e áreas de espera. Se definirmos $\omega_i = 0$, $\forall\ i\ \in N$, o BCAP se reduz ao *problema de alocação de áreas de espera* (ou BAP, do inglês ***B**uffer **A**llocation **P**roblem* [MacGregor & Cruz, 2005; Cruz et al., 2008]). Por outro lado, se definirmos $\omega_i = 1$, $\forall\ i\ \in N$, obtém-se o *problema de alocação de servidores* (ou CAP, do inglês *Server (**C**) **A**llocation **P**roblem* [MacGregor Smith et al., 2010]).

É importante observar a possibilidade de existência de soluções que não apresentem áreas de espera (conhecidos, em inglês, como sistemas *zero-buffer* ou *bufferless*), ou seja, em determinados nós pode ocorrer $B_i = 0$. Redes de filas sem áreas de espera são observadas em diversos sistemas de produção reais e em determinados ambientes eles são estritamente necessários, devido a características do processo em si ou simplesmente pela ausência de capacidade de armazenamento entre duas etapas consecutivas de operação em um processo produtivo [Andriansyah et al., 2010]. Também é importante ressaltar que, além das restrições (2.2)–(2.4), uma restrição adicional é necessária para garantir a existência de uma solução ótima finita:

$$\rho_i \equiv \frac{\lambda_i}{c_i \mu_i} < 1, \tag{2.5}$$

em que λ_i é a taxa de chegada no nó i e μ_i é a sua taxa de serviço.

2.2 Relaxação Lagrangeana

O BCAP, definido pelas expressões (2.1)–(2.5), apresenta restrições complexas, sobretudo a representada pela expressão (2.2). A incorporação das restrições mais complexas na função objetivo via relaxação lagrangeana tem sido aplicada com sucesso em problemas de alocação, como o BAP [MacGregor & Cruz, 2005; Cruz et al., 2008] e o CAP [MacGregor Smith et al., 2010]. Assim, fica natural a escolha da relaxação langreana como ferramenta para a resolução do BCAP. Uma descrição aprofundada da relaxação lagrangeana não será apresentada aqui, uma vez que pode ser encontrada com facilidade na literatura [Lemaréchal, 2007].

De forma simplificada, a relaxação lagrangeana é uma técnica que consiste em incorporar as restrições complexas do problema de otimização diretamente na função objetivo, na forma de uma penalidade. No BCAP, a restrição complexa (2.2) pode ser relaxada em termos da variável lagrangeana $\alpha > 0$, resultando no BCAP relaxado, ou RBCAP (do inglês ***R**elaxed **B**uffer and Server (**C**) **A**llocation **P**roblem*), cuja formulação é apresentada a seguir.
(RBCAP):

$$Z_\alpha = \min\left[\sum_{\forall i \in N} \varpi_i c_i + \sum_{\forall i \in N} (1-\omega_i) B_i + \alpha(\theta_{\min} - \theta(\mathbf{c},\mathbf{B}))\right], \tag{2.6}$$

sujeito às restrições (2.3)-(2.5).

É importante notar que, na formulação RBCAP, o termo $\alpha(\theta_{\min} - \theta(\mathbf{c},\mathbf{B}))$ é sempre não-positivo, para qualquer solução factível da formulação original do BCAP. Ou seja, se as restrições (2.2)–(2.5) forem satisfeitas, elas devem garantir que $\alpha(\theta_{\min} - \theta(\mathbf{c},\mathbf{B})) \leq 0$, e assim, temos $Z_\alpha \leq Z$. Consequentemente, Z_α será utilizado como uma cota inferior para o valor ótimo de Z. Espera-se que tal cota inferior seja próxima de Z tanto quanto possível. Vale salientar que α fornece uma penalidade para os casos em que a restrição não seja atendida. Experimentalmente, chega-se a um valor adequado da variável lagrangeana α, que no caso aqui tratado pode ser fixada em 10^3 [Cruz et al., 2008].

Definido o algoritmo de resolução do RBCAP e estabelecido algum limiar mínimo $\theta_{\min}$, (por exemplo, um percentual da taxa de chegada), alguma ferramenta para análise de desempenho da rede de filas deve ser empregada. Será usado aqui o *Método da Expansão Generalizado* (GEM, do inglês *Generalized Expansion Method* [Kerbache e MacGregor Smith, 1987, 1988]), uma aproximação consolidada para as probabilidades de bloqueio, p_K, necessárias à estimação da taxa de saída, $\theta(\mathbf{c},\mathbf{B})$, para uma particular configuração de servidores e áreas de espera.

3 Algoritmos para Análise de Desempenho e Otimização em Redes de Filas Finitas

A taxa de atendimento (ou, em inglês, *throughput*) é uma das medidas mais utilizadas na análise de desempenho de filas. Quando se trata de uma única fila finita, tal taxa está diretamente relacionada com a taxa de entrada da fila e da sua probabilidade de bloqueio, p_K, sendo o bloqueio o momento em que o número de entidades presentes na fila preenche toda sua capacidade K. O desempenho da fila, em termos de taxa de atendimento, é dado por:

$$\theta = \lambda(1 - p_K), \tag{3.1}$$

que indica que, conhecendo a taxa de entrada da fila, a obtenção da taxa de atendimento fica condicionada à determinação do valor da sua probabilidade de bloqueio p_K. Assim, definindo-se um limiar mínimo para a taxa de saída, $\theta_{\min}$, é possível expressar a probabilidade de bloqueio p_K a partir da seguinte expressão:

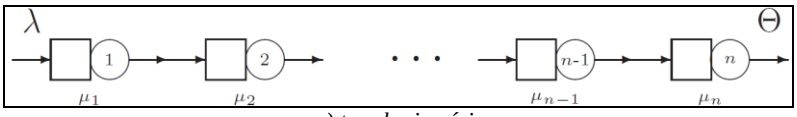
$$p_K \leq 1 - \frac{\theta_{\min}}{\lambda}. \tag{3.2}$$

Como a taxa de atendimento $\theta_{\min}$ é necessariamente não superior à taxa de entrada λ ($\theta_{\min} \leq \lambda$), analisando a expressão (3.2), pode-se concluir que, quando $\theta_{\min} \to 0$, então $p_K \to 1$ e $K \to 1$. Por outro lado, quando $\theta_{\min} \to \lambda$, temos $p_K \to 0$ e $K \to +\infty$. Considerando tais fatos, para obter um desempenho ótimo da fila, devem-se buscar configurações de servidores e de áreas de espera que garantam um taxa mínima de atendimento $\theta_{\min}$, suficientemente próxima de λ. Assim é minimizada a proabilidade de que clientes sofram bloqueio durante o processo de atendimento na fila.

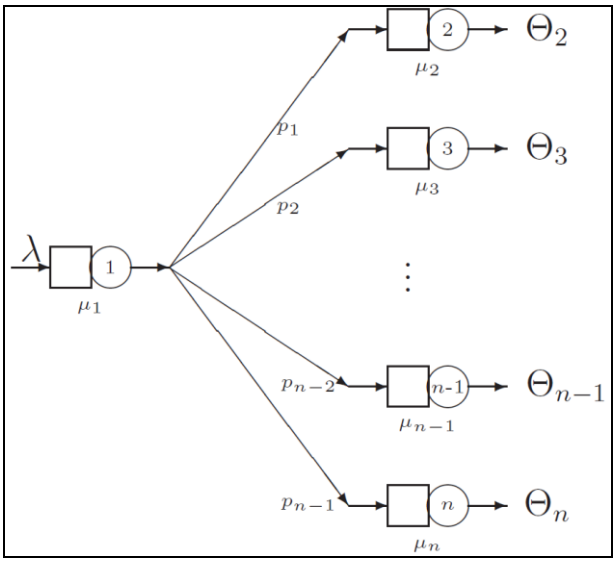
Para otimizar o RBCAP, definido pela função objetivo (2.6) e pelas restrições (2.3) a (2.5), será utilizado o método de Powell, um algoritmo clássico de otimização não-linear que não utiliza derivadas [Himmelblau, 1972], em conjunto com o GEM [Kerbache e MacGregor Smith, 1987, 1988], que determina aproximadamente as medidas de desempenho, ou, mais especificamente no caso aqui tratado, a taxa de atendimento θ. O método de Powell, conjuntamente com o GEM, tem sido bem sucedido na resolução de diversos problemas de alocação em redes de filas finitas [MacGregor, 2003,2004].

4 Resultados e Discussão

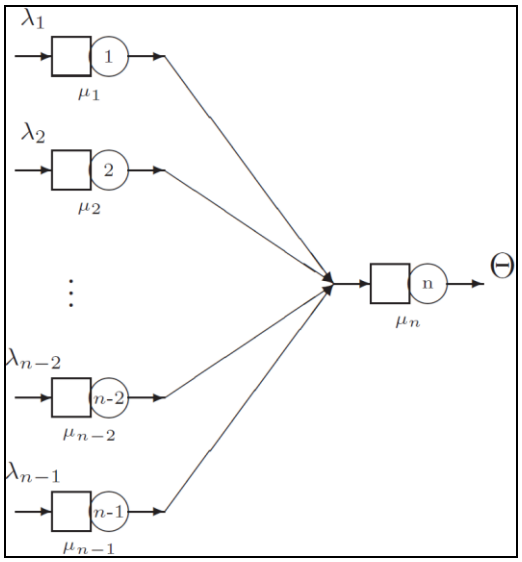
O algoritmo de otimização conjunta foi testado em redes nas topologias *série*, *divisão* e *fusão*, conforme apresentado na Fig. 3, com o objetivo de identificar possíveis padrões de alocação, em função de alguns parâmetros de interesse. Todos os algoritmos utilizados foram implementados em FORTRAN e estão disponíveis a pedido, para fins de pesquisa. Apresentam-se a seguir os resultados obtidos.



a) topologia série



b) topologia divisão



c) topologia fusão

Figura 3: Redes de filas básicas testadas

A Fig. 3-a apresenta um exemplo de rede de filas na topologia *série*. Esta topologia de rede implica que o cliente, após sair de um nó, não possui possibilidade de seleção do próximo nó, ou seja, o serviço será recebido de forma serial. Supondo um sistema com três nós em série, o cliente, ao sair do nó 1, automaticamente seria destinado ao nó 2 e, consequentemente, ao nó 3. Para os experimentos na topologia *série*, foram consideradas redes de filas com o número de nós N = 2, 3, 4, 5, taxa de entrada na rede $\lambda = 5$ e servidores com taxas de atendimento $\mu_j = 10$, $\forall\, j \in N$.

A Fig. 3-b representa a topologia *divisão*. Nesta topologia, diferentemente da topologia *série*, apresenta ao término do atendimento em um nó a possibilidade de seleção do próximo nó. Supondo que após o término do atendimento o cliente possua n possíveis nós subsequentes, ele pode selecionar o próximo nó de destino com probabilidade p_i, i = 1, 2,...,n−1. As redes de filas na topologia *divisão*, utilizadas nos experimentos computacionais, contam com um número de nós fixo N = 3, em que um nó inicial 1 recebe as entidades a serem atendidas e, após o atendimento, elas são encaminhadas para os nós 2 e 3, com probabilidades de roteamento p_1 e p_2, respectivamente. Foram configuradas redes com probabilidades de roteamento variando dentre as seguintes opções: $p_1 = 0{,}1$ e $p_2 = 0{,}9$; $p_1 = 0{,}2$ e $p_2 = 0{,}8$; $p_1 = 0{,}3$ e $p_2 = 0{,}7$; $p_1 = 0{,}4$ e $p_2 = 0{,}6$; e a situação de equilíbrio, $p_1 = p_2 = 0{,}5$. Assim como nas redes de filas em série, foram consideradas taxas de entrada na rede $\lambda = 5$ e servidores com taxas de atendimento $\mu_j = 10$, $\forall\, j \in N$.

Finalmente, na Fig. 3-c está apresentada uma rede de filas na topologia *fusão*. Uma rede de filas na topologia *fusão* caracteriza-se pela união de clientes atendidos em servidores distintos em um único servidor subsequente, ou seja, supondo uma série de filas paralelas, todos os clientes destas ao término do atendimento serão direcionados para uma única fila, independentemente do atendimento anterior. Nas redes na topologia *fusão*, também foram considerados 3 nós, nos quais as entidades a ser atendidas ingressam na rede pelos nós 1 e 2 e, após atendimento, são necessariamente destinadas ao nó 3. Foi mantida a taxa total de entrada no sistema, $\lambda = 5$. Porém, tal taxa foi dividida entre os nós iniciais nas seguintes proporções variáveis λ_1 (taxa de entrada no nó 1) e λ_2 (taxa de entrada no nó 2): $\lambda_1 = 0{,}5$ e $\lambda_2 = 4{,}5$; $\lambda_1 = 1{,}0$ e

$\lambda_2 = 4{,}0$; $\lambda_1 = 1{,}5$ e $\lambda_2 = 3{,}5$; $\lambda_1 = 2{,}0$ e $\lambda_2 = 3{,}0$; e $\lambda_1 = \lambda_2 = 2{,}5$. A taxa de atendimento em cada servidor é $\mu_j = 10$, $\forall\, j \in N$.

Para as três topologias básicas, *série*, *divisão* e *fusão*, três valores foram considerados para os quadrados dos coeficientes de variação dos tempos de serviço, s^2, quais sejam, sistemas *hipoexponenciais* (isto é, com $s^2 = 0{,}5$), *markovianos* ($s^2 = 1{,}0$) e *hiperexponenciais* ($s^2 = 1{,}5$).

Na Fig. 4 tem-se uma representação gráfica dos padrões de alocação, para as diferentes configurações de números de nós e s^2, em que se observaram os seguintes aspectos. As alocações, tanto de servidores quanto das áreas de espera (ou *buffers*), são homogêneas, conforme esperado. Não poderia ser diferente, pois as filas possuem servidores com a mesma taxa de serviço ($\mu_j = 10$, $\forall\, j \in N$). É interessante observar que as configurações subótimas não possuem área de espera ($\Sigma_{\forall i}\, B_i = 0$) até uma relação de custos C_c/C_B superior a 2:1 (ou seja, $\omega_i\,/(1-\omega_i) > 2$). Infelizmente, esse limiar é bastante imprevisível, dependente da configuração da rede e é, até o presente momento, determinável apenas pela aplicação do algoritmo. Em outras palavras, não há ainda uma regra que preveja essa mudança de estado (sem *buffers* → com *buffers*).

No que diz respeito ao s^2, um aumento no seu valor (isto é, na variabilidade) conduz a uma maior alocação (tanto de servidores, quanto de áreas de espera). É notável o comportamento da rede de filas no sistema hipoexponencial, que apresenta queda no número de servidores e aumento brusco do número das áreas de espera na transição $C_c/C_B = 4{:}1 \rightarrow C_c/C_B = 8{:}1$. Para os sistemas markovianos, houve estabilidade tanto para servidores quanto para áreas de espera e para o caso hiperexponencial ($s^2 = 1{,}5$), observa-se a manutenção do número de servidores e menor aumento no número de áreas de espera. Reforça-se aqui a convicção, já mencionada na literatura da área [van Woensel et al., 2010], que os servidores são mais eficazes para lidar com a variabilidade que as áreas de espera, possivelmente pelo fato de que eles têm a função dupla de reter o cliente e ao mesmo tempo atendê-lo. Finalmente, em relação à topologia *série*, notou-se que o aumento do número de filas na linha não altera o padrão homogêneo da alocação nem a transição sem/com *buffers*, o que demonstra a robustez da solução com a configuração do sistema de filas a ser otimizado.

Os resultados da topologia *divisão* podem ser vistos na Fig. 5. Como observado na topologia série, as alocações são sem áreas de espera até certa proporção de custos C_c/C_B (na maioria dos casos observa-se ausência de áreas de espera para proporções iguais ou inferiores a 2:1), independente do s^2). Também interessante, é que a proporção de clientes destinados para cada braço da divisão após o atendimento inicial não se verifica na alocação de recursos. Por exemplo, a razão das probabilidades de roteamento 0,1:0,9 não resulta em uma divisão de recursos nesta mesma proporção, mas sim de 2:3. Entretanto, na medida em que as probabilidades de roteamento tornam-se próximas (i.e., $p_1 \rightarrow p_2$), a alocação torna-se homogênea.

Também digno de nota, é que a alocação na fila antes da divisão, em geral, *não* é a soma das alocações nos ramos após a divisão, tanto no que diz respeito a c quanto a K. Neste aspecto, observa-se que quanto maior o custo dos servidores, maior será a diferença entre o número de áreas de espera do ramo inicial e a soma dos ramos que receberão as entidades atendidas pelo primeiro nó. Quando é necessário, para valores de $s^2 = 0{,}5$, observa-se uma alocação superior de áreas de espera nos ramos finais da rede, enquanto que, para valores de $s^2 = 1{,}0$ e $1{,}5$, o maior contingente de áreas de espera é destinado ao nó inicial da rede. Finalmente, nota-se que, com o aumento do número de nós na rede (resultados não apresentados), a sequência de filas após a divisão comporta-se como uma topologia série (isto é, com alocação homogênea), como seria previsível.

A Fig. 6 apresenta os resultados de alocação obtidos para redes em topologia *fusão*. Nota-se uma simetria entre a topologia fusão e a anterior (topologia *divisão*), conforme seria esperado. É interessante notar que, apenas no caso hiperexponencial ($s^2 = 1{,}5$), com $\lambda_1 = 0{,}5$ e $\lambda_2 = 4{,}5$, houve um acréscimo de um servidor no nó que recebe o maior número de clientes e no nó final. O acréscimo destes dois servidores impactou e reduziu as áreas de espera, em comparação com as demais configurações.

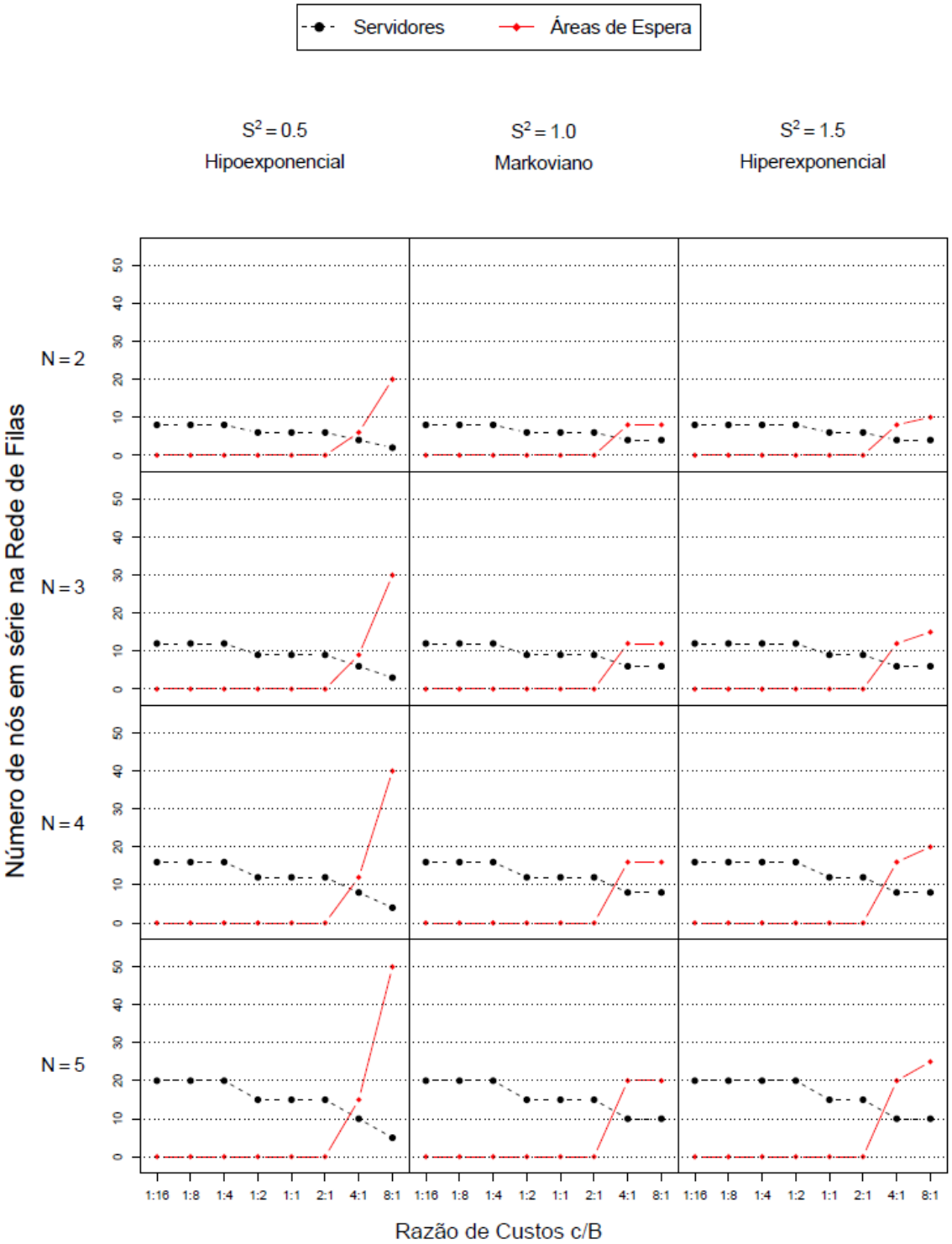


Figura 4: Alocação de servidores e áreas de espera para topologia série

Também aqui, o nó após a divisão necessita de uma alocação que em geral não é a soma das alocações nos ramos antes da divisão. Verifica-se, portanto, um efeito do tipo economia de escala neste nó compartilhado. Assim como ocorrido nas demais topologias, as oscilações provocadas no sistema, neste caso o desequilíbrio entre as taxas de chegada λ_1 e λ_2 não afetaram o padrão homogêneo das alocações nem a transição da rede do estado sem áreas de espera para o estado com áreas de espera.

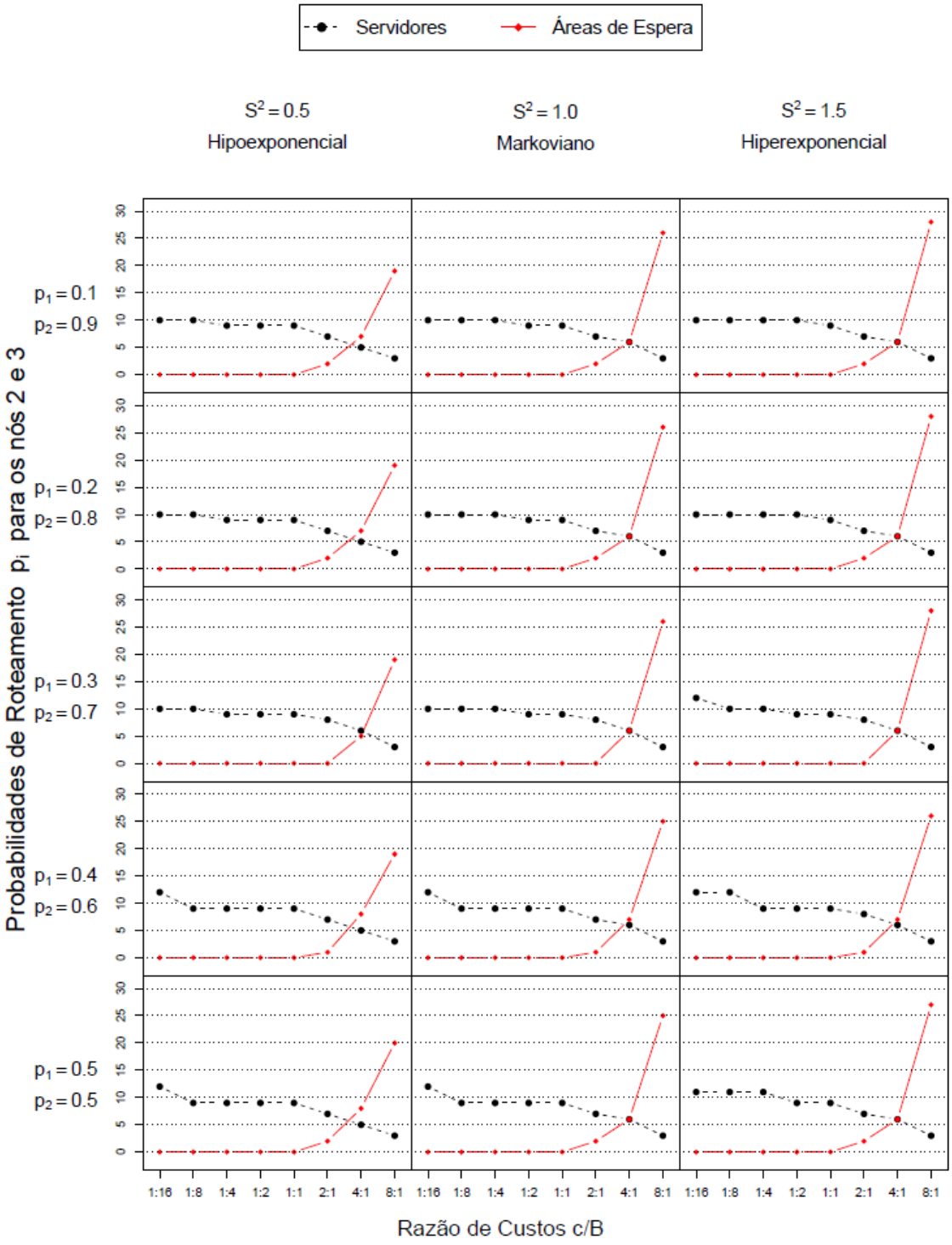


Figura 5: Alocação de servidores e áreas de espera para topologia divisão

Até este ponto, foram observados padrões que indicaram a existência de homogeneidade na alocação ótima de servidores e áreas de espera, para valores de s^2 = 0,5, 1,0 e 1,5. A fim de verificar a possível influência de alguns fatores em tais padrões de alocação foi construído o diagrama de dispersão apresentado na Fig. 7.

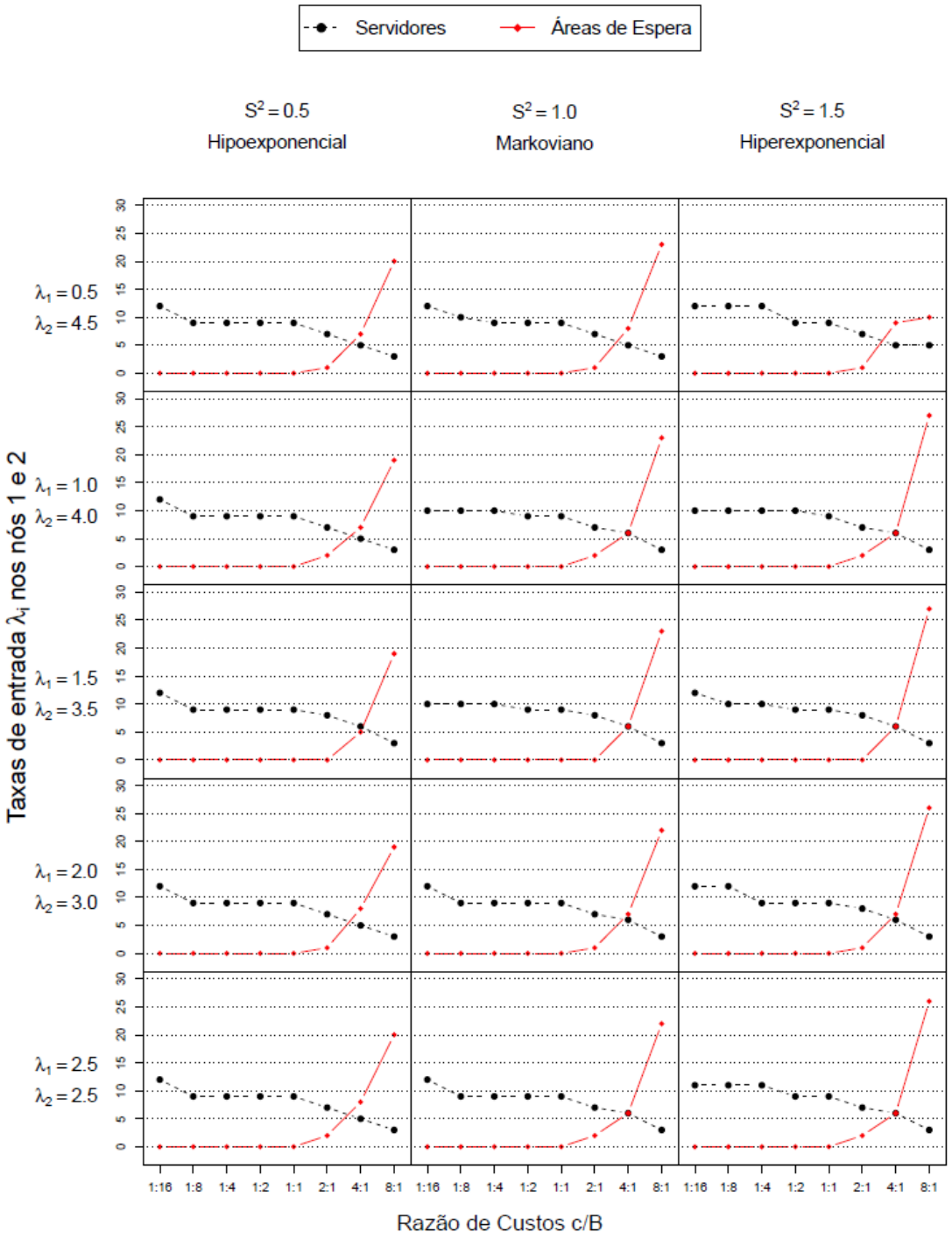


Figura 6: Alocação de servidores e áreas de espera para topologia fusão

Na construção do diagrama apresentado na Fig. 7, para cada topologia básica testada, as soluções foram estratificadas em $s^2 = 0{,}5$ (sistemas hipoexponenciais) e $s^2 \geq 1{,}0$ (sistemas markovianos e hiperexponenciais). Também foram estratificadas considerando-se as relações de custo C_c/C_B como inferiores a 4:1 ou iguais ou superiores a esta relação (isto é, 4:1 ou 8:1).

Nota-se que, quando se tem uma relação de custos entre servidores e áreas de espera C_c/C_B igual ou superior a 4:1, a alocação ótima total (considerando servidores e áreas de espera conjuntamente) tende a ser mais alta em relação à alocação ótima de servidores. Ademais, para relação de custos C_c/C_B iguais ou inferiores a 2:1, resultam-se soluções sem áreas de espera (ou, do inglês, *zero-buffer sytems*), na forma de pares ordenados exatamente sobre a bissetriz do

primeiro quadrante do plano cartesiano. O quadrado do coeficiente do tempo de serviço, s^2, parece não ter influência no padrão de alocação, uma vez que as soluções estratificadas por este fator não se concentram em nenhuma sub-região específica do primeiro quadrante. Resultados obtidos para outras topologias e diferentes taxas de chegada e probabilidades de bloqueio (não apresentados) conduzem a conclusões similares.

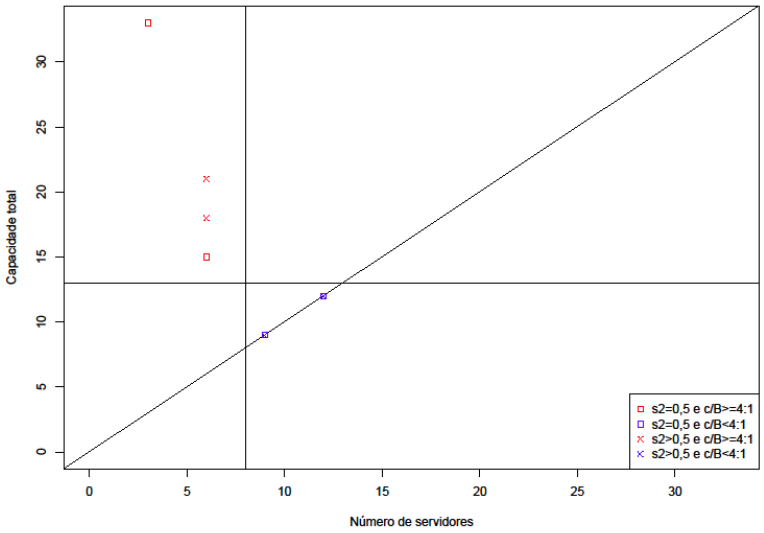


Figura 7: Alocação total em função de s^2 e C_c/C_B para topologia série

5 Conclusões e Observações Finais

Neste artigo, foi apresentada uma análise do *problema de alocação conjunta de servidores e áreas de espera* (BCAP) que estende trabalho anteriormente publicado [van Woensel et al., 2010]. A principal contribuição deste artigo é apresentar um estudo e a identificação de padrões de alocação de servidores em redes de filas configuradas nas topologias *série*, *divisão* e *fusão*, apontando para alocações satisfatórias em termos de robustez. Observa-se que apesar das diferentes topologias, a alocação de servidores e áreas de espera apresenta resultados homogêneos para todos os casos. É indicada ausência de áreas de espera para situações nas quais os seus custos sejam superiores aos (ou não tão distantes dos) custos de servidores. Por outro lado, ocorre crescimento da alocação de tais áreas de espera quanto o custo do servidor é elevado. Tal constatação pode colaborar para a resolução do problema para redes de filas mais gerais, dado que se podem construir sistemas complexos baseados na junção de tais topologias básicas. Pode também auxiliar na compreensão de problemas reais modelados por redes de filas.

Como trabalhos futuros, podem-se citar (i) a análise dos padrões obtidos em redes de filas complexas que combinam as três topologias básicas; (ii) a verificação da homogeneidade das soluções, com perturbações nas taxas de serviço μ_j; (iii) a investigação das soluções do BCAP para filas com tempos de chegada gerais e independentes, ou seja, filas *GI*/*M*/*c*/*K*, na notação de Kendall; (iv) a alocação ótima em redes de filas com ciclos, para modelar realimentação e/ou retrabalho, e assim por diante. Estes são apenas alguns tópicos para trabalhos futuros nesta instigante linha de pesquisa.

Agradecimentos

Os autores agradecem ao CNPq (processos n° 304671/2014-2 e 300825/2016-1) e à FAPEMIG (processos n° CEX-PPM-00564-17 e CEX-PPM-00427-17), pelo financiamento parcial a esta pesquisa.

Referências

Andriansyah, R., van Woensel, T., Duczmal, L. H. e Cruz, F. R. B. (2010). Performance Optimization of Open Zero-Buffer Multi-Server Queueing Networks. *Computers & Operations Research* 37(8): 1472–1487.

Cruz, F. R. B. e van Woensel, T. (2014a). Buffers and servers allocation in general finite queueing networks. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, v. 2, p. 0101081–0101086.

Cruz, F. R. B. e van Woensel, T. (2014b). Joint buffer and server allocation in general finite queueing networks. *XLVI Simpósio Brasileiro de Pesquisa Operacional - XLVI SBPO [CD-ROM]*, Salvador, Brasil, p. 2213–2220.

Cruz, F. R. B., Duarte, A. R. e van Woensel, T. (2008). Buffer allocation in general single-server queueing networks.*Computers & Operations Research* 35(11): 3581–3598.

Himmelblau, D. M. (1972). Applied Nonlinear Programming, McGraw-Hill Book Company, New York, NY.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Annals of Mathematical Statistics* 24: 338–354.

Kerbache, L. e MacGregor Smith, J. (1987). The generalized expansion method for open finite queueing networks. *European Journal of Operational Research* 32: 448–461.

Kerbache, L. e MacGregor Smith, J. (1988). Asymptotic behavior of the expansion method for open finite queueing networks, *Computers & Operations Research*, 15(2): 157–169.

Lemaréchal, C. (2007). The omnipresence of Lagrange. *Annals of Operations Research* 153(1): 9–27.

MacGregor Smith, J. (2003). *M*/*G*/*c*/*K* blocking probability models and system performance. *Performance Evaluation* 52(4): 237–267.

MacGregor Smith, J. (2004). Optimal design and performance modelling of *M*/*G*/1/*K* queueing systems. *Mathematical and Computer Modelling* 39(9-10): 1049–1081.

MacGregor Smith, J. e Cruz, F. R. B. (2005). The buffer allocation problem for general finite buffer queueing networks. *IIE Transactions* 37(4): 343–365.

MacGregor Smith, J., Cruz, F. R. B. e van Woensel, T. (2010). Optimal server allocation in general, finite, multi-server queueing networks. *Applied Stochastic Models in Business & Industry* 26(6): 705–736.

van Woensel, T., Andriansyah, R., Cruz, F. R. B., MacGregor Smith, J. e Kerbache, L. (2010). Buffer and server allocation in general multi-server queueing networks. *International Transactions in Operational Research* 17(2): 257–286.