PLANEJAMENTO DE TOPOLOGIA DE REDES DE FILAS FINITAS GERAIS COM SERVIDORES MÚLTIPLOS

Frederico R. B. Cruz
Departamento de Estatística – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 – 31270-901 – Belo Horizonte – MG
e-mail: fcruz@ufmg.br

Milene S. Castro
Departamento de Engenharia de Produção – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 – 31270-901 – Belo Horizonte – MG
e-mail: mscastro@ufmg.br

Resumo

Este artigo aborda um dos problemas de planejamento mais desafiadores, que é o problema de otimização de redes de filas finitas gerais com servidores múltiplos. Diversos problemas, configurados em topologias diversas (série, divisão e fusão), são analisadas através de um método aproximado de estimação de medidas de desempenho e de um algoritmo iterativo para encontrar a alocação ótima das áreas de espera na rede. O coeficiente de variação do serviço desempenhou um papel significativo no espaço alocado em redes uniformes e com gargalos. Resultados computacionais ilustram a simetria nos padrões de alocação que surgem das redes em topologia série, fusão e junção.

Palavras-Chaves: Servidores múltiplos; Redes finitas; Probabilidades de bloqueio; Alocação de área de espera.

Abstract

This paper deals with one of the most complex topological network design problems, which is the optimization of general service, finite waiting room, multi-server queueing networks. Topologies in series, merge, and split are examined by means of an approximation method to estimate the performance measures of these queueing networks and an iterative search methodology to find the optimal buffer allocation within the network. The coefficient of variation of the service is shown to be a significant factor in the buffer allocation for multiple servers in uniform and bottleneck server networks. Computational results are included to illustrate the symmetries in the buffer patterns which emerge from the series, merge, and splitting topologies.

Keywords: Multi-server; Finite networks; Blocking probabilities; Buffer allocation.

1. INTRODUÇÃO

Redes de filas finitas, gerais e com servidores múltiplos ocorrem em muitos sistemas físicos de interesse, como por exemplo, sistemas de fabricação, de telecomunicações e de transportes. Sempre que há fluxo de produtos e incerteza sobre o processamento destes produtos, a alocação de recursos para processamento deste fluxo resulta em um sistema de redes de filas finitas.

O tipo de alocação que nos interessa aqui inclui as áreas de espera, a ordem dos servidores e a iteração destes dois fatores. A questão colocada para esta pesquisa é como podemos modelar e predizer com precisão suas medidas de desempenho e planejar adequadamente estes sistemas?

Neste artigo procuramos caracterizar e otimizar a topologia de um sistema de filas

finitas. Procuramos propriedades que nos permitam modelar e construir algoritmos para e otimizar tais sistemas. Este artigo estende trabalhos anteriores, nos quais somente redes de filas finitas com um único servidor foram consideradas [17]. Assim, para sistemas com servidores múltiplos, precisamos compreender como estes servidores podem afetar a alocação de áreas de espera e como as várias topologias e variações sistemáticas no coeficiente de variação do tempo de serviço influenciam a configuração ótima do sistema.

Assumimos fornecida uma rede finita $G=(N,A)$ em uma determinada topologia, com o conjunto de nós N e o conjunto de arcos A, com distribuição geral do tempo de serviço nos nós e probabilidades de roteamento nos arcos conhecidas. Procuramos otimizar medidas de desempenho desta rede, tais como o número de atendimentos por unidade de tempo (do inglês, *throughput*), o trabalho em processo (do inglês, *work-in-process*), a utilização ou os custos e lucros. Uma vez que a rede tem capacidade finita, poderá haver bloqueio, o que acarretará características em forma não-produto que dificultam muito a determinação das distribuições de probabilidade do número de usuários na rede. Assim, somos forçados a procurar formas eficazes de decomposição do problema de modo a aproximar adequadamente as medidas de desempenho.

Este artigo está organizado como se segue. Na seção 2 do artigo descrevemos as origens do problema e algum trabalho correlato. Na seção 3 descrevemos os modelos matemáticos necessários para nossa abordagem e na seção 4, os algoritmos empregados para sua resolução. Na seção 5 apresentamos resultados experimentais para diversas topologias. Finalmente, na seção 6 apresentamos conclusões, observações finais e tópicos para futuras pesquisas na área.

2. ORIGEM DO PROBLEMA

O problema de alocação de áreas de espera em filas finitas configuradas em redes é bastante desafiador e tem recebido uma atenção constante dos pesquisadores. Abordagens exatas têm sido limitadas aos casos de distribuições exponenciais, mas mesmo nestes casos o tratamento é limitado a redes de pequeno tamanho, uma vez que o espaço de estados nas cadeias de Markov envolvidas explode, além de que freqüentemente o inter-relacionamento probabilístico entre os filas fica muito complexo e de difícil compreensão. Serviços não-exponenciais nas redes não tornam os problemas mais fáceis, uma vez que a propriedade de falta de memória da distribuição exponencial deixa de valer. Assim, o desenvolvimento de aproximações tem sido considerado uma estratégia razoável e prática.

Com relação ao problema de determinação de medidas de desempenho, em particular da probabilidade de bloqueio, p_K, aproximações a dois momentos têm sido bem sucedidas. Esta deverá ser a abordagem aqui empregada, por possibilitar uma metodologia poderosa para aproximar as probabilidades de bloqueio nestas redes de filas finitas gerais. Métodos aproximados para obtenção da probabilidade de bloqueio em filas finitas gerais do tipo $M/G/1/k$ e $M/G/c/K$ têm uma longa e detalhada história (em que, adotando-se a conhecida notação de Kendall, o M indica que a chegada é um processo markoviano, o G representa um tempo de serviço com distribuição geral, 1 e c se referem ao número de servidores e, por fim, a capacidade do sistema é restrita a K usuários, incluindo aqueles nos servidores). Métodos exatos não são possíveis para grandes valores de c ou K, pela perda da propriedade de falta de memória da distribuição exponencial.

Aproximações começaram a surgir essencialmente a partir do trabalho de Gelenbe [3], com seu método baseado em técnicas de difusão. Também expressões baseadas em distribuições estacionárias de sistemas infinitos surgem, com destaque para os trabalhos de Scheweiter & Konheim [16], Tijms [21] e Sakasegawa et al. [15]. Finalmente surgem as aproximações a dois momentos de Tijms [22,23], Kimura [10,11] e Smith [18].

Do seu lado, os problemas de alocação de áreas de espera também possuem uma longa e detalhada história, com uma grande riqueza de autores que abordaram o tema. Vários algoritmos baseados em programação dinâmica, em metaheurísitcas, em simulação ou em

métodos de busca surgiram, com destaque para as abordagens por programação dinâmica de Kubat & Sumita [12] e de Yamashita & Onvural [24], o algoritmo de simulated annealing de Spinellis et al. [20], os métodos baseados em simulação de Soyester et al. [19], Baker et al. [2] e Haris & Powel [5], e, finalmente, os algoritmos baseados em busca de Altiok & Stidham [1] e, mais recentemente, o método de Smith & Cruz [17]. Uma vez que a alocação ótima de áreas de espera é um problema de programação inteira estocástica, as abordagens heurísticas tendem a ser mais bem sucedida que os algoritmos que garantem a otimalidade. Para este problema, portanto, necessita-se de um procedimento de otimização robusto e acurado, acoplado a uma forma efetiva de determinação de medidas de desempenho do sistema. Isto é que buscamos neste artigo.

3. MODELOS MATEMÁTICOS E COMPUTACIONAIS

3.1. NOTAÇÃO

Esta seção apresenta um pouco da notação necessária para o artigo:

- λ_j : taxa de chegada ao nó *j*;
- μ_j : taxa média de serviço do *j*;
- c: número de servidores;
- $\rho = \lambda/(\mu c)$: intensidade de tráfego;
- B_j : capacidade da área de espera do nó *j* excluindo-se aqueles em serviço;
- K_j : capacidade total do nó *j* incluindo-se aqueles em serviço;
- p_K : probabilidade de bloqueio da fila finita de capacidade *K*;
- $s^2 = \mathrm{Var}(T_s)/\mathrm{E}(T_s)$: quadrado do coeficiente de variação do tempo de serviço, T_s ;
- $\Theta(\mathbf{x})$: taxa de atendimento.

3.2. FORMULAÇÃO MATEMÁTICA

Neste artigo, consideraremos a seguinte formulação para o problema de otimização, que também foi o objetivo central no trabalho de Smith & Cruz [17]:

$$Z = \min\left(f(\mathbf{x}) = \sum_{\forall i} x_i \right), \tag{1}$$

sujeito a:

$$\Theta(\mathbf{x}) \geq \Theta^{\tau}, \tag{2}$$

$$x_i \in \{1,2,\ldots\}, \forall i, \tag{3}$$

que minimize a alocação total de áreas de espera, $\sum_{\forall i} x_i$, restrito a garantir um taxa mínima de atendimento, Θ^{τ} .

Nesta formulação, Θ^{τ} é a taxa de atendimento limiar e $x_i \equiv K_i$ é a variável de decisão, que é a área de alocação total na *i*-ésima fila, incluindo-se aqueles em serviço. Ficaremos restritos a processos de chegada markovianos, uma vez que representam uma boa aproximação para inúmeras situações práticas e também por ser possível a derivação de resultados exatos para alguns casos particulares destes sistemas. Resultados para chegadas gerais são escassos e limitados a servidores simples [9].

Notemos que a restrição (2) deve ser satisfeita. Assim, precisamos de métodos para estimar a taxa de atendimento total $\Theta(\mathbf{x})$. Apresentamos em seguida de estimação da taxa de atendimento para filas únicas, após o que generalizamos para filas finitas configuradas em redes.

3.3. PROBABILIDADE DE BLOQUEIO EM FILAS ÚNICAS

A probabilidade de bloqueio para sistemas *M/M/1/K* com ρ<1 é bem conhecida de

qualquer livro básico de processos estocásticos [4]:

$$p_K = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}.$$

Se relaxamos as restrições de integralidade de K, podemos expressar K em função de ρ e p_K e obter uma expressão fechada para a alocação ótima de áreas de espera, que é o menor inteiro não inferior a:

$$K = \frac{\ln\left(\frac{p_K}{1-\rho+p_K\rho}\right)}{\ln(\rho)}.$$

Smith [18] e Smith & Cruz [17] mostraram que uma vez que haja disponível uma expressão fechada para a área total $B^*=K^*$-1 em um sistema *M/M/1/K*, é possível utilizar um esquema de aproximação a dois momentos baseado nos trabalhos de Kimura e Tijms [10,11,23] para desenvolver uma expressão para a alocação ótima B^*, em sistemas com serviço geral. A aproximação a dois momentos, desenvolvida por Smith [18] é baseada em uma combinação ponderada da expressão para a alocação ótima de sistemas markovianos, denotada por K^{M}:

$$K_{\varepsilon}^{\mathrm{Smith}}(s^2) = K^{\mathrm{M}} + \frac{(s^2-1)}{2}\sqrt{\rho}K^{\mathrm{M}} = \underbrace{\left(\frac{\ln\left(\frac{p_K}{1-\rho+p_K\rho}\right)}{\ln(\rho)}\right)}_{K^{\mathrm{M}}} + \frac{(s^2-1)}{2}\sqrt{\rho}\underbrace{\left(\frac{\ln\left(\frac{p_K}{1-\rho+p_K\rho}\right)}{\ln(\rho)}\right)}_{K^{\mathrm{M}}}.$$

É importante ressaltar que a expressão de Kimura estima a área de espera excluindo-se aqueles usuários em serviço. Para c=1 e s^2, temos uma aproximação para a área de espera ótima B^* para sistemas *M/G/1/K*:

$$B^* = \frac{\left[\ln\left(\frac{p_K}{1-\rho+p_K\rho}\right)+\ln(\rho)\right]\left(2+\sqrt{\rho}s^2-\sqrt{\rho}\right)}{2\ln(\rho)}.$$

Se s^2=1, a expressão se reduz àquela de sistemas *M/M/c/K*, c=1, a menos do espaço do servidor. Como esperado, podemos prosseguir com este processo de desenvolvimento de p_K para diferentes valores de c e então obter formas fechadas aproximadas para a área de espera ótima em sistemas *M/G/c/K*.

A seguinte ligação, existente entre a probabilidade de bloqueio p_K e a taxa de atendimento $\Theta(\mathbf{x})$, explicita a relação entre a expressão desenvolvida para o B^* e a solução do problema de programação matemática das Eq. (1)-(3):

$$\Theta(\mathbf{x}) = \lambda(1-p_K).$$

Desta forma, notamos que definir uma alocação ótima que atenda à restrição (2), $\Theta(\mathbf{x}) \geq \Theta^{\tau}$, é equivalente a garantir um B^* que atenda a $p_K \leq \varepsilon$.

3.4. PROBABILIDADE DE BLOQUEIO EM FILAS CONFIGURADAS EM REDES

O problema de determinação das probabilidades de bloqueio, e, conseqüentemente, da taxa de atendimento, fica bem mais complexo em filas finitas configuradas em redes. O Método da Expansão Generalizado (MEG) é uma forma robusta e eficaz para determinação aproximada de medidas de desempenho em redes de filas finitas. O MEG foi desenvolvido por Kerbache & Smith [8] e tem uma longa história de aplicações bem sucedidas a diversas situações similares. Descrito em detalhes em vários artigos, o método é uma combinação de

tentativas repetidas e decomposição nó-a-nó. Considere uma rede única, com capacidade finita *K* (incluindo os servidores). Este nó essencialmente oscila entre dois estados – fase saturada e não-saturada. Na fase não-saturada, o nó *j* tem no máximo *K*-1 usuários (em serviço ou na área de espera). Por outro lado, quando o nó está na fase saturada, nem mais um único usuário pode juntar-se à fila. A Fig. 1 representa graficamente estes dois cenários.

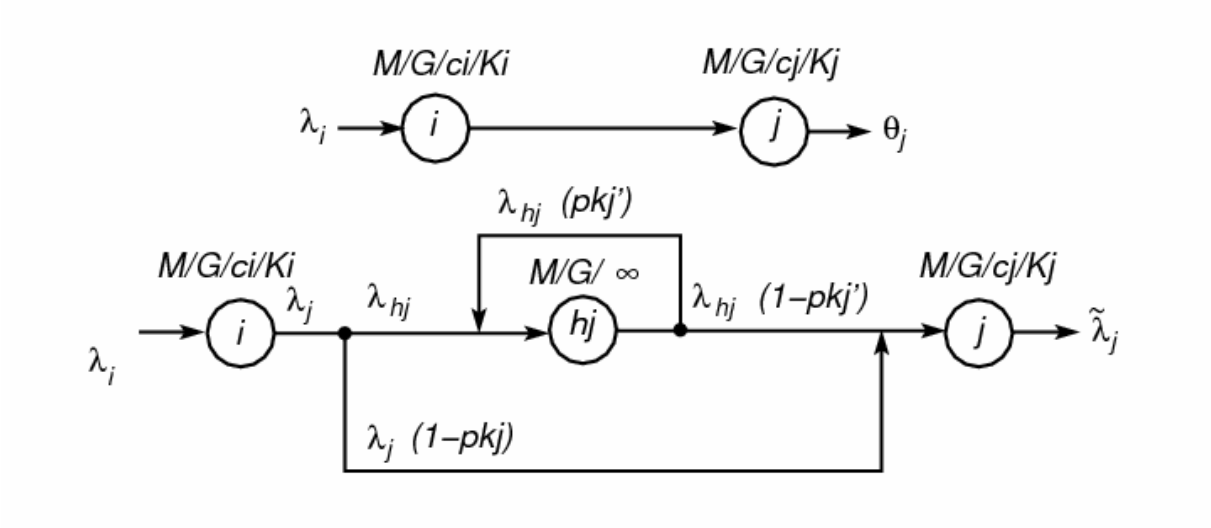

Figura 1: Representação gráfica do Método da Expansão Generalizado

O MEG possui os seguintes estágios:

- Estágio I: re-configuração da rede;
- Estágio II: estimação de parâmetros;
- Estágio III: eliminação da realimentação.

Não daremos aqui maiores detalhes do MEG, que podem ser encontrados no artigo de Kerbache & Smith [8]. Diremos apenas que o objetivo final do método é prover um esquema aproximado para atualizar a taxa de serviço média dos nós que possuem outros nós com capacidade finita à sua frente (por exemplo, o nós *i*, na Fig. 1), de forma a levar em consideração o bloqueio entre nós finitos adjacentes, isto é:

$$\tilde{\mu}_i^{-1} = \mu_i^{-1} + p_{K_j}\,\mu_{h_j}^{-1} \tag{4}$$

Em outras palavras, primeiramente a rede é expandida, acrescentando-se antes de cada nó finito um nó artificial de espera h_j, de capacidade ilimitada, para receber as entidades que forem bloqueadas. Em seguida, os parâmetros do nó de espera são estimados . Finalmente, a realimentação é eliminada, a taxa de serviço é atualizada de acordo com a Eq. (4) e o arco de realimentação e o nó de espera artificial h_j são eliminados. Completado estes três estágios, tem-se uma rede expandida que pode ser utilizada para o cálculo das medidas de desempenho para a rede original. Como uma técnica de decomposição, este método permite sucessivas adições de nós de espera para cada nó finito, estimação de parâmetros e subseqüente eliminação do nó de espera. Uma observação importante sobre o processo é que não modificamos fisicamente as redes, mas tão somente utilizamos a expansão como um artifício para implementação computacional do método de aproximação. O resultado final deste processo sofisticado é a obtenção de uma aproximação para $\Theta(\mathbf{x})$ bastante acurada [8].

4. ALGORITMOS

O problema aqui examinado é o da determinação ótima de áreas de espera em sistemas de filas *M/M/c/K* e *M/G/c/K* configuradas em redes, dado pelas Eq. (1)-(3). Uma

forma de incorporar a restrição de taxa de serviço mínima é através de uma função penalidade, tal como a relaxação lagrangeana, sobre a qual um tutorial recentemente publicado pode ser encontrado em Lemaréchal [13]. Assim, definindo-se uma variável dual α e relaxando-se a restrição (2), o seguinte problema penalizado surge:

$$Z_\alpha = \min\left[\sum_{\forall i} x_i + \alpha\underbrace{\left(\Theta^\tau - \Theta(\mathbf{x})\right)}_{\leq 0}\right],$$

sujeito a:

$$x_i \in \{1,2,\ldots\}, \forall i,$$
$$\alpha \geq 0.$$

Note que Θ^τ pode ser previamente especificado e servir como taxa de chegada λ para um algoritmo aproximado para determinação de medidas de desempenho como o MEG [8], que fornecerá o taxa de saída resultante $\Theta(\mathbf{x})$. Assim, o termo $\alpha\left(\Theta^\tau - \Theta(\mathbf{x})\right)$ será não-positivo para qualquer **x** viável e será uma penalidade para a função objetivo relacionada com a diferença entre a taxa de atendimento pré-determinada $\lambda = \Theta^\tau$ e a taxa de serviço efetivamente alcançada $\Theta(\mathbf{x})$. Desta forma, segue que $Z_\alpha \leq Z$, sendo Z_α um limite inferior para a solução ótima do problema, Z.

A relaxação lagrangeana do problema primal, Z_α, acrescida de uma relaxação adicional na integralidade das restrições para x_i, torna-se um problema clássico de otimização irrestrita. Neste formulação em particular, as variáveis $x_i, \forall i$, são as variáveis de decisão. Mesmo sendo essencialmente variáveis inteiras, elas podem ser razoavelmente aproximadas por arredondamento de soluções provenientes de um algoritmo de otimização não-linear.

Para acoplar o problema de otimização com o MEG, o algoritmo de Powell [6] será utilizado para a busca pelo vetor ótimo, enquanto calcula-se pelo MEG a taxa de saída resultante. O método de Powell localiza o mínimo de uma função não linear $f(\mathbf{x})$ por meio de sucessivas buscas unidimensionais, a partir de um ponto inicial $\mathbf{x}(0)$, via um conjunto de direções conjugadas. Estas direções conjugadas são geradas dentro do próprio procedimento de Powell, que se baseia na idéia de o mínimo de uma função não linear $f(\mathbf{x})$ poder ser encontrado ao longo de p direções conjugadas, com um passo adequado em cada direção. Temos visto relatos de grande sucesso do algoritmo de Powell e do MEG [17] e por essa razão o utilizaremos aqui.

5. RESULTADOS EXPERIMENTAIS

Nesta seção do artigo apresentaremos resultados experimentais da nossa metodologia de planejamento de redes de filas gerais com servidores múltiplos. Adotaremos a tática de apresentar resultados para redes pequenas, mas configuradas em topologias interessantes. O leitor precisa ter em mente que o número de experimentos possíveis é exponencial e selecionamos apenas uma amostra para apresentar.

Examinemos uma rede que agrega uma combinação interessante de topologias série, fusão e divisão. Esta combinação será denominada rede primal e é apresentada na Fig. 2. Na rede primal, os nós 1 e 6 são gargalos, uma vez que o bloqueio será mais severo nestes nós e uma área maior de espera deverá ser alocada a eles.

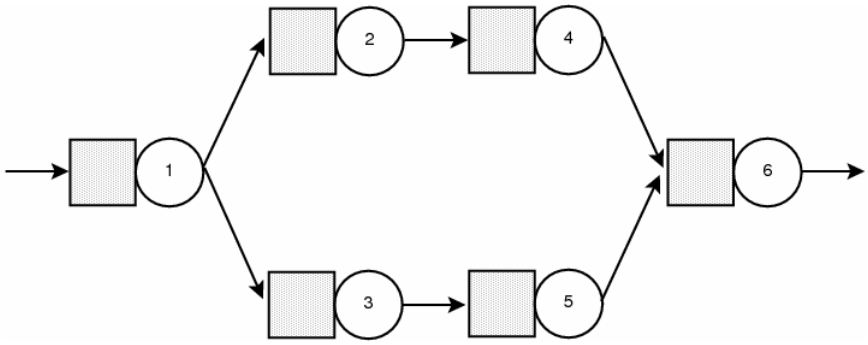

Figura 2: Topologia de rede primal

A rede dual, derivada da grafo dual da rede primal, representa uma topologia em árvore, conforme pode ser observado na Fig. 3. Na rede dual, os nós 3 e 4 é que são os gargalos. Assim como na rede primal, a alocação de áreas deverá ser direcionada a estes nós.

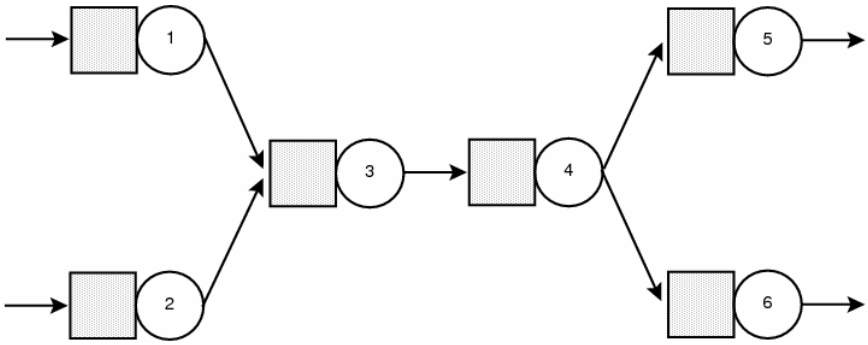

Figura 3: Topologia de rede dual

Para ambas as redes, seis experimentos foram conduzidos, para diversos valores de s^2. Em todos os experimentos escolhemos λ=7, como taxa de chegada, e μ=10, como taxa de serviço média. Para avaliar a precisão dos resultados analíticos obtidos pelo algoritmo de otimização, simulações foram conduzidas no programa Arena [7], com 20 replicações, para determinação dos intervalos de confiança, com um período de estabilização (do inglês, *burn-in*) igual a 20.000 unidades de tempo e um tempo total de simulação igual a 100.000 unidades de tempo. Para simular os tempos de serviços gerais com s^2={1/2,3/2}, utilizamos a distribuição gama. Os experimentos foram conduzidos em um PC AMD AthlonXP 1800+ 1.53 GHz, com 512 MB de RAM e Windows XP.

Tabela 1: Resultados para a rede primal

					Simulação				
s^2	**c**	$\Theta(\mathbf{x})$	Z_α	**x**	$\Theta(\mathbf{x})^s$	δ	Z_α^s	$\Delta\%\Theta(\mathbf{x})$	$\Delta\%Z_\alpha$
1/2	(1,1,1,1,1,1)	6,993	59,3	(14,6,6,6,6,14)	6,988	0,004	64,5	0,07	-8,14
1	(2,1,1,1,1,2)	6,989	69,4	(17,6,6,6,6,17)	6,990	0,003	67,8	-0,01	2,37
3/2	(4,1,1,1,1,4)	6,989	75,2	(18,7,7,7,7,18)	6,992	0,003	72,3	-0,04	4,07

A alocação ótima **x**, obtida pelo algoritmo, é apresentada nas Tab. 1 e 2. Notamos

que os resultados analíticos e de simulação são próximos, conforme mostram as colunas δ, que apresentam a metade do intervalos de 95% de confiança obtidos via simulação. Além disso, os valores para Z, analíticos e de simulação, divergem no máximo de 8% (coluna $\Delta\%Z_\alpha$). A alocação é simétrica para a rede primal e dual, exatamente conforme esperado, haja vista os mecanismos de gargalo das duas redes.

Tabela 2: Resultados para a rede dual

					Simulação				
s^2	**c**	$\Theta(\mathbf{x})$	Z_α	**x**	$\Theta(\mathbf{x})^s$	δ	Z_α^s	$\Delta\%\Theta(\mathbf{x})$	$\Delta\%Z_\alpha$
1/2	(1,1,1,1,1,1)	6,993	59,3	(6,6,14,14,6,6)	6,993	0,004	55,4	0,00	6,95
1	(1,1,2,2,1,1)	6,989	69,4	(6,6,17,17,6,6)	6,989	0,003	68,7	0,00	1,03
3/2	(1,1,4,4,1,1)	6,989	75,2	(7,7,18,18,7,7)	6,990	0,004	74,0	-0,01	1,68

Nas comparações, os resultados analíticos e de simulação para ambas as redes são surpreendentemente próximos, o que é bastante encorajador. Também incluído nestas tabelas é a desvio percentual entre os resultados analíticos e de simulação da taxa de atendimento, $\Delta\%\theta(\mathbf{x})$. Note-se que os desvios são muito próximos de zero.

Neste ponto, poderíamos continuar rodando mais experimentos com redes maiores e mais complexas. Escolhemos parar aqui, pois este é um processo sem fim e, talvez, não revele muito mais do que já foi mostrado.

6. CONCLUSÕES E OBERVAÇÕES FINAIS

Apresentamos uma abordagem para o problema de alocação de áreas de espera em filas configuradas em redes com distribuição geral do tempo de serviço e servidores múltiplos. Descrevemos as expressões desenvolvidas para as probabilidades de bloqueio, bem como a metodologia de otimização empregada. Vários experimentos foram apresentados e detalhadamente discutidos, para ilustrar o escopo e as limitações da abordagem. Esperamos que o leitor sinta a força desta nova abordagem na sua habilidade em atacar complexos problemas de planejamento de redes de filas finitas.

6.1. QUESTÕES EM ABERTO E TRABALHOS FUTUROS

Dentre as possíveis direções que esta pesquisa pode tomar, podemos citar a aplicação do algoritmo a problemas da área de manufatura e montagem, planejamento de plantas e leiautes, bem como a problemas de planejamento de redes de telecomunicações e sistemas de computação.

Também não examinamos situações nas quais o número de servidores, **c**, fosse uma variável de decisão. Isto poderá requerer uma reestruturação da abordagem e decidimos não tratar este aspecto neste momento da pesquisa.

Um aspecto desanimador com respeito ao número de servidores é que ele não parece ser tão crítico quanto a área de espera, conforme evidenciado pelos nossos resultados. Outros pesquisadores chegaram a resultados similares a respeito da importância da área de espera.

Outra possibilidade é incluir nos estudos experimentais exemplos de redes com laços de realimentação, bastante encontrados em manufatura e no setor de serviços. Laços de realimentação causam forte dependência entre as chegadas e necessitam cuidadosa consideração. Estas são apenas algumas direções interessantes para futuros trabalhos na área.

7. AGRADECIMENTOS

O trabalho de pesquisa de Frederico Cruz foi realizado com apoio parcial do CNPq, processos 201046/1994-6, 301809/1996-8, 307702/2004-9 e 472066/2004-8, da Fundação de Ampara à Pesquisa do Estado de Minas Gerais, FAPEMIG, processos CEX-289/98 e CEX-855/98, e da PRPq da UFMG, processo 4081-UFMG/RTR/FUNDO/PRPQ/99. Durante a elaboração deste trabalho, Milene S. Castro era aluna do Curso de Pós-Graduação em Engenharia de Produção da UFMG, apoiada pela FAPEMIG.

8. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Altiok, T. & Stidham, S., 1983, The allocation of interstage buffer capacities in production lines, *IIE Transactions* **15**, 251-261.

[2] Baker, K. R., Powell, S. & Pyke, D., 1990, Buffered and unbuffered assembly systems with variable processing times, *Journal of Manufacturing and Operations Management* **3**, 200-223.

[3] Gelenbe, E., 1975, On approximate computer system models, *Journal of the ACM* **22**(2), 261-269.

[4] Gross, D. & Harris, C., 1985, *Fundamentals of Queueing Theory*, John Wiley & Sons, New York.

[5] Harris, J. H. & Powell, S. G., 1999, An algorithm for optimal buffer placement in reliable serial lines, *IIE Transactions* **31**, 287-302.

[6] Himmelblau, D. M., 1972, *Applied Nonlinear Programming*, McGraw-Hill Book Company, New York.

[7] Kelton, D., Sadowski, R. P. & Sadowski, D. A., 2001, *Simulation with Arena*, MacGraw-Hill College Div, New York.

[8] Kerbache, L. & Smith, J. MacGregor, 1987, The generalized expansion method for open finite queueing networks, *European Journal of Operational Research* **32**, 448-461.

[9] Kim, N. K. and Chae, K. C., 2003, Transform-free analysis of the *GI/G/1/K* queue through the decomposed Little's formula, *Computers & Operations Research* **30**(3), 353-365.

[10] Kimura, T., 1996a, Optimal buffer design of an *M/G/s* queue with finite capacity, *Communications in Statistics - Stochastic Models* **12**(1), 165-180.

[11] Kimura, T., 1996b, A transform-free approximation for the finite capacity *M/G/s* queue, *Operations Research* **44**(6), 984-988.

[12] Kubat, P. & Sumita, U., 1985, Buffers and backup machines in automatic transfer lines, *International Journal of Production Research* **23**(6), 1259-1280.

[13] Lemaréchal, C., 2003, The omnipresence of Lagrange, *4OR* **1**, 7-25.

[14] Onvural, R. O., 1990, Survey of closed queueing networks with blocking, *ACM Computing Surveys* **22**(2), 83-121.

[15] Sakasegawa, H., Miyazawa, M. & Yamazaki, G., 1993, Evaluating the overflow probability using the infinite queue, *Management Science* **39**(10), 1238-1245.

[16] Schweitzer, P. J. & Konheim, A. G., 1978, Buffer overflow calculations using an infinite-capacity model, *Stochastic Processes and their Applications* **6**(3), 267-276.

[17] Smith, J. MacGregor & Cruz, F. R. B., 2005, The buffer allocation problem for general finite buffer queueing networks, *IIE Transactions on Design & Manufacturing* **37**(4), 343-365.

[18] Smith, J. MacGregor, 2003, *M/G/c/k* blocking probability models and system performance, *Performance Evaluation*, **52**(4), 237-267.

[19] Soyster, A. L., Schmidt, J. W. & Rohrer, M. W., 1979, Allocation of buffer capacities for a class of fixed cycle production lines, *AIIE Transactions* **11**(2), 140-146.

[20] Spinellis, D., Papadopoulos, C. T. & Smith, J. MacGregor, 2000, Large production line optimization using simulated annealing, *International Journal of Production Research*, **38**(3), 509-541.

[21] Tijms, H. C., 1987, Stochastic Modelling and Analysis: A Computational Approach, John Wiley & Sons, New York.

[22] Tijms, H. C., 1992, Heuristics for finite-buffer queues, *Probability in the Engineering and Informational Sciences* **6**, 267-276.

[23] Tijms, H. C., 1994, *Stochastic Models: An Algorithmic Approach*, John Wiley & Sons, New York.

[24] Yamashita, H. & Onvural, R., 1994, Allocation of buffer capacities in queueing networks with arbitrary topologies, *Annals of Operations Research*, **48**, 313-332.