ANÁLISE DE UM ALGORITMO DE OTIMIZAÇÃO DE REDES DE FILAS FINITAS

Helinton A. L. Barbosa
Departamento de Estatística – ICEx – UFMG
Av. Antônio Carlos, 6627, 31270-901 – Belo Horizonte – MG
e-mail: helinton@ufmg.br

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

Resumo

Neste trabalho abordamos o problema de determinação ótima das áreas de espera em redes de filas finitas gerais com servidores múltiplos. Um método de aproximação é utilizado para estimar o desempenho da rede de filas e um algoritmo iterativo de busca é empregado para a alocação ótima da área de espera. Examinamos então algumas configurações e os resultados mostram que a alocação faz sentido e é estável. Além disso, em grande parte dos casos testados, os resultados analíticos aproximam-se dos intervalos de 95% de confiança, estimados através de simulações. Foi confirmado que o coeficiente de variação do tempo de serviço influencia significativamente na alocação das áreas de espera. Em relação ao método de aproximação proposto, um experimento planejado apresentou resultados sobre o desempenho do algoritmo em várias topologias de filas.

Palavras-Chaves: Redes finitas; probabilidade de bloqueio; alocação de área de espera; planejamento de experimentos.

Abstract

The topological network design of general service, finite waiting room, multi-server queueing networks is addressed. Several topologies are examined using an approximation method to estimate the performance of the queueing networks and an iterative search method is developed to find the optimal buffer allocation within the network. Extensive computational results are included to illustrate the methodology and to show that the buffer allocations derived by the algorithms make sense. The results were quite satisfactory as in the cases tested the approximate analytical results were within the 95% confidence intervals estimated by simulation. Additionally, quite different topologies may result in a similar performance, which may bring flexibility to the planner. Finally, it was shown that the coefficient of variation of the service times is significant in the buffer allocation for both uniform and bottlenecked server networks.

Keywords: Finite networks; blocking probabilities; buffer allocation; design of experiments.

1. INTRODUÇÃO

A alocação de recursos para processamento de um fluxo de bens resulta em uma rede de filas finitas, quando há uma incerteza sobre os fluxos e sobre os tempos de processamento

destes produtos nos nós da rede. A alocação de recursos que nos interessa aqui inclui as áreas de espera, a ordem dos servidores e a interação entre estes dois fatores. Uma pergunta relevante é como podemos modelar de maneira eficaz e predizer com precisão suas medidas de desempenho, além projetamos corretamente estes sistemas estocásticos?

Neste trabalho, procuramos caracterizar e otimizar a topologia de um sistema de redes de filas finitas. Buscamos propriedades que nos permitam modelar e otimizar tais sistemas, além de construir algoritmos para sua solução. Este texto é uma extensão de trabalhos anteriores em que redes de filas finitas com um único servidor foram consideradas (Smith & Cruz, 2005). Em sistemas com servidores múltiplos, precisamos identificar como os servidores afetam a alocação ótima da área de espera e também como as várias topologias e variações sistemáticas no coeficiente de variação do tempo de serviço geral influenciam o sistema.

Assumimos como dada uma rede finita *G(N,A)* em uma topologia especificada, com o conjunto de nós *N* e de arcos *A*, sendo uma distribuição geral do tempo de serviço nos nós e probabilidade de roteamento nos arcos conhecida. Procuramos determinar as medidas de desempenho ótimas de tais redes, como o número de atendimentos por unidade de tempo (do inglês *throughput)*, o trabalho em processo, a utilização e otimização de custos ou lucros. Uma vez que a rede tem capacidade finita, poderá haver bloqueio, que acarretará características na forma não-produto, que dificultam muito a determinação do número de usuários da rede. Assim, somos forçados a procurar maneiras eficazes de decomposição do problema para avaliar adequadamente as suas medidas de desempenho.

O texto está organizado da seguinte forma. Na seção 2, descrevemos as origens do problema e trabalhos anteriores relacionados a ele. Na seção 3, descrevemos os modelos matemáticos apropriados para nossa aproximação e, na seção 4, os algoritmos empregados na análise. Na seção 5, descrevemos resultados experimentais e, na seção 6, analizaremos o desempenho do algoritmo em diferentes redes de filas através de um experimento planejado. Por fim na seção 7, apresentamos conclusões e observações finais, além de levantarmos tópicos para possíveis trabalhos na área.

2. ORIGEM DO PROBLEMA

Como mencionado na seção 1, o problema é desafiador e tem recebido grande atenção de pesquisadores, com várias publicações a seu respeito. Abordagens exatas foram limitadas às suposições de distribuições exponenciais, mas mesmo estas aproximações da cadeia de Markov de tempo contínuo (MCTC) são limitadas a modelar redes de pequeno tamanho, uma vez que o espaço do estados explode e freqüentemente há relacionamentos probabilísticos complexos, os quais não são compreendidos facilmente. Tempos de serviços não exponenciais nas redes podem ser difíceis de analisar de maneira exata, uma vez que a propriedade da falta de memória de distribuições exponenciais deixa de valer. Conseqüentemente, o uso de aproximações é bem razoável e prático.

Abordagens a dois momentos foram muito bem sucedidas no passado e utilizaremos aqui também esta aproximação, pois ela apresenta uma metodologia poderosa para aproximar a probabilidade de bloqueio para redes com serviços gerais. Metodologias de aproximação da probabilidade de bloqueio em sistemas de *M/G/1/K* e *M/G/c/K* têm uma longa e detalhada história. Seguindo a notação de Kendall, *M* indica que o tempo entre chegadas segue uma distribuição markoviana (exponencial), *G* representa a distribuição geral do tempo de serviço, 1 e *c* representam o número de servidores e *K* se refere à capacidade total do sistema. Métodos exatos não são praticáveis para *c* e *K* grandes, pois a propriedade de falta de

memória da distribuição exponencial não vale. As aproximações começam essencialmente com o trabalho de Gelenbe, com uma aproximação baseada em técnicas de difusão (Gelembe,1975). Também, fórmulas baseadas nas distribuições estacionárias de sistemas infinitos aparecem propostas por Schweitzer & Konheim (1978), Tijms (1987) e Sakasegawa et al. (1993) e foram muito difundidas. Finalmente surgem as aproximações a dois momentos de Tijms (1992, 1994), Kimura (1996a,b) e Smith (2003).

O problema de alocação da área de espera possui também uma história longa e detalhada, devido à grande riqueza de autores que o abordaram. Entre alguns tipos de abordagens, incluem-se aquelas baseadas em programação dinâmica (Yamashita & Onvural, 1994), em métodos da busca (Smith & Cruz, 2005), em metaheuristica (Spinellis et al., 2000) e métodos baseados em simulação (Harris & Powell, 1999). Como o problema de alocação de área de espera é um problema estocástico inteiro com uma função objetivo não-linear, as abordagens heurísticas tendem a ser melhores que algoritmos de otimização. Para este problema então, necessitamos de um procedimento robusto e acurado de otimização, junto a uma maneira eficaz para medir o desempenho do sistema. Isto é o que mostraremos neste trabalho.

3. MODELOS MATEMÁTICOS

3.1. NOTAÇÃO

Esta seção apresenta alguma notação, necessária ao bom entendimento do trabalho:

- λ_j : taxa de chegada poisson ao nó *j*;
- μ_j : taxa média de serviço exponencial do nó *j*;
- c : número dos servidores;
- $\rho = \lambda/(\mu c)$: intensidade do 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*;
- $s^2 = Var(T_s)/E(T_s)$: quadrado do coeficiente de variação do tempo de serviço, T_s;
- $\Theta(x)$: taxa média de atendimento.

3.2. FORMULAÇÃO MATEMÁTICA PROBABILIDADE DE BLOQUEIO EM FILAS ÚNICAS

Neste trabalho, consideraremos o seguinte tipo de problema de otimização, que foi também o objetivo central usado por Smith & Cruz (2005) em seu artigo:

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

Sujeito a:

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

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

que minimiza a alocação da área de espera $\sum_{\forall i} x_i$, restrito a garantir uma 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 alocação total na *i*-ésima fila. Ficaremos restritos a processos de chegadas markovianos porque resultados exatos podem ser derivados destes sistemas. Adicionalmente, resultados para chegadas gerais são escassos e limitados a servidores simples (veja, por exemplo, o artigo de Kim & Chae (2003)).

3.3. PROBABILIDADE DE BLOQUEIO EM FILAS ÚNICAS

A probabilidade de bloqueio para um sistema de *M/M/1/K* com $\rho < 1$ é bem conhecida de qualquer livro texto de processos estocásticos (Gross & Harris,1985):

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

Se relaxarmos as restrições de integralidade de *K*, podemos expressar *K* em função de ρ e p_K e chegar a uma expressão em forma fechada para o tamanho da alocação ótima da área de espera, que é o menor inteiro não inferior a:

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

Em artigos anteriores Smith (2003); Smith & Cruz (2005), foi mostrado que uma vez que exista a expressão fechada para a área total *B*=K*-1* em um sistema *M/M/1/K*, pode-se usar um esquema de aproximação a dois momentos baseado no trabalho de Tijms e Kimura (Kimura, 1996a,b; Tijms, 1987, 1992, 1994) para desenvolver a expressão da alocação ótima *B** em sistemas com serviço geral. Para c=1 e s^2, uma aproximação para a área de espera ótima *B** para sistemas de *M/G/1/K* é:

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

Se s^2=1, então a expressão se reduz àquela de sistemas *M/M/c/K*, c=1, quando subtraímos o espaço do servidor. Podemos continuar este processo e desenvolver p_K e *B** para diferentes valores de *c* obtendo formas fechadas aproximadas para a área de espera e probabilidades de bloqueio em sistemas *M/G/c/K* (Smith et al. 2006).

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

$$\Theta(x) = \lambda(1-p_K)$$

3.4. PROBABILIDADE DE BLOQUEIO EM FILAS CONFIGURADAS EM REDES

O método da expansão generalizado (MEG) é uma técnica robusta e bastante eficaz de aproximação desenvolvida por Lerbache & Smith (1987) para determinação de medidas de desempenho de redes de filas finitas. Como descrito em artigos anteriores, este método é caracterizado como uma combinação de tentativas repetidas e decomposição nó a nó. Metodologias para cálculo de medidas de desempenho para uma rede de filas finitas utilizam principalmente dois tipos de bloqueio a seguir:

• Tipo I : O nó *i* é bloqueado quando um atendimento for concluído, pois o usuário não pode avançar devido à fila no nó *j*, o qual está cheio. Isto é chamado às vezes como bloqueio após o serviço (do inglês *Blocking After Service* (BAS)) (Onvural,1990).

• Tipo II: O nó é bloqueado quando está na fase saturada e o atendimento é suspenso, estando terminado ou não. Isto é chamado às vezes de bloqueio antes do serviço (do inglês *Blocking Before Service* (BBS)) (Onvural, 1990).

O MEG usa o bloqueio tipo I, que prevalece em sistemas de produção, manufaturas, transporte e outros sistemas similares. Considere um único nó com capacidade finita *K* (incluindo serviço). Este nó oscila essencialmente entre dois estados - a fase saturada e a fase não-saturada. Na fase não-saturada, o nó *j* tem no máximo $K - 1$ usuários (em serviço ou na fila). Por outro lado, quando o nó está saturado nenhum usuário pode juntar a fila. Veja na Figura 1 uma representação gráfica dos dois cenários.

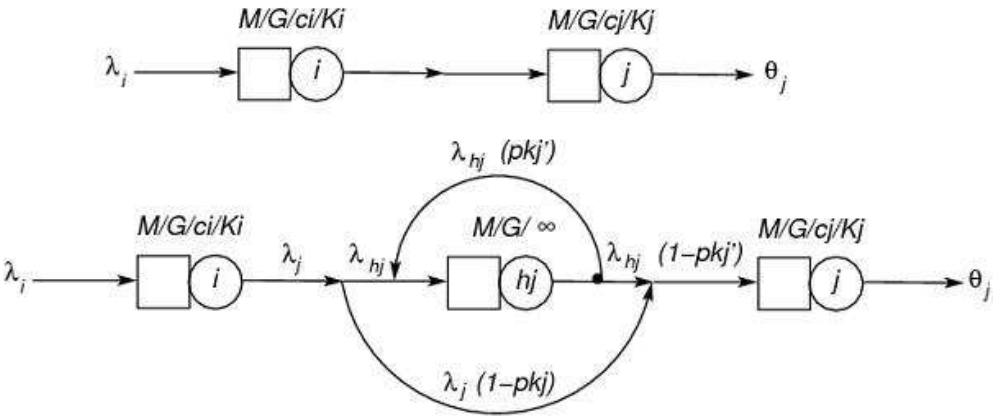

Figura 1: Método da Expansão Generalizado.

O MEG possui os seguintes estágios:

• Estágio I: re-configuração da rede;
• Estágio II: estimação do parâmetro;
• Estágio III: eliminação da realimentação.

Os detalhes do MEG não serão dados aqui e podem ser encontrados no artigo de Kerbache & Smith (1987). O objetivo final do MEG é fornecer um esquema de aproximação

para atualizar a taxa de serviço em nós que possuem outros nós com capacidade finita à sua frente, de forma a levar em consideração o bloqueio entre nós finitos adjacentes, isto é:

$$\tilde{\mu}_i^{-1} = \mu_i^{-1} + p_{K_j}\left(\mu_{h_j}\right)^{-1}$$

Recapitulando, a rede é expandida, seguido pela aproximação das probabilidades de roteamento devido ao bloqueio e serviço no nó h*j* . Em seguida, os parâmetros do nó de espera são estimados e finalmente a realimentação é eliminada. Uma vez que estes três estágios estão completos, temos uma rede expandida, que pode então ser usada para calcular as medidas de desempenho para a rede original. Como uma técnica de decomposição, esta aproximação permite a adição sucessiva de nós de espera para cada nó finito, a estimação dos parâmetros e a eliminação subseqüente do nó de espera. Um ponto importante sobre este processo é que não modificamos fisicamente as redes, mas somente representamos a expansão como um artifício para implementação computacional do método de aproximação.

4. ALGORITMOS

O problema de otimização de sistemas *M/M/c/K* e *M/G/c/K* aqui examinado á dado pelas equações (1)-(3). Um modo de incorporar a restrição da taxa de serviço, Eq.(2), é através de uma função de penalidade, tal como a relaxação lagrangeana (veja um tutorial recentemente publicado em Lemaréchal (2003)). Assim, definindo uma variável dual α e relaxando-se a restrição (2), o seguinte problema penalizado é dado:

$$Z_\alpha = \min\left[\sum_{i=1}^{N} \quad + \alpha\underbrace{\left(\Theta^\tau - \Theta(x)\right)}_{\leq 0}\right]$$

Sujeito a: $x_i \in \{1,2,\mathrm{K}\}, \forall i$

$$\alpha \geq 0$$

Note que Θ^τ pode ser pré-especificado e servir como taxa de entrada λ de um algoritmo aproximado para determinação de medidas de desempenho como o MEG (Kerbache & Smith, 1987), que fornecerá uma taxa de saída correspondente. Então, o termo $\alpha\left(\Theta^\tau - \Theta(x)\right)$ será sempre não-positivo para todo *x* viável e será uma penalidade da função objetivo, relacionada à diferença entre taxa de atendimento pré-especificado, $\lambda = \Theta^\tau$ e a taxa de serviço alcançada, $\Theta(x)$. Assim, segue que $Z_\alpha \leq Z$, onde Z_α é um limite inferior para Z, a solução ótima do problema dado pelas equações (1)-(3). A relaxação lagrangeana do problema primal, Z_α, acrescida de uma relaxação adicional na integridade das restrições para x_i, torna-se um problema clássico de otimização irrestrita. Nesta formulação em particular do problema, as variáveis x_i são variáveis de decisão para controle da otimização. Apesar de serem essencialmente variáveis inteiras, elas podem razoavelmente ser aproximadas por arredondamento de soluções provenientes de um algoritmo de programação não-linear. A fim acoplar o problema do otimização com o MEG, o algoritmo de Powell será usado para encontrar o(s) vetor(es) ótimo(s) de alocação de área de espera, enquanto o MEG calcula a taxa de saída resultante. O método de Powell (Himmelblau, 1972), encontra o mínimo *f(x)* de

uma função não-linear por sucessivas buscas unidimensionais a partir de um ponto inicial $x_{(0)}$, via um conjunto de direções conjugadas. Estas direções conjugadas são geradas dentro do próprio procedimento. O método de Powell é baseado na idéia que um mínimo de *f(x)* não-linear seja encontrado ao longo de *p* direções conjugadas em um estágio da busca, com um passo adequado em cada direção. Há relatos na literatura de grande sucesso do trabalho conjunto entre algoritmo de Powell e do MEG (Smith & Cruz, 2005) e por isso serão utilizados aqui.

5. RESULTADOS EXPERIMENTAIS

Nesta seção apresentaremos resultados experimentais de nossa metodologia de planejamento de redes de filas com servidores múltiplos. Mostraremos alguns resultados encontrados para redes de dois nós.

5.1. REDES COM DOIS NÓS E TRÊS SERVIDORES

O modelo mais simples de rede é uma configuração com dois nós e três servidores. Uma topologia organizada em série envolvendo os servidores é apresentada na Figura 2. Gostaríamos de testar se uma topologia domina a outra, isto é, verificar se existe uma configuração mais eficiente que a outra, baseado apenas na ordem dos servidores.

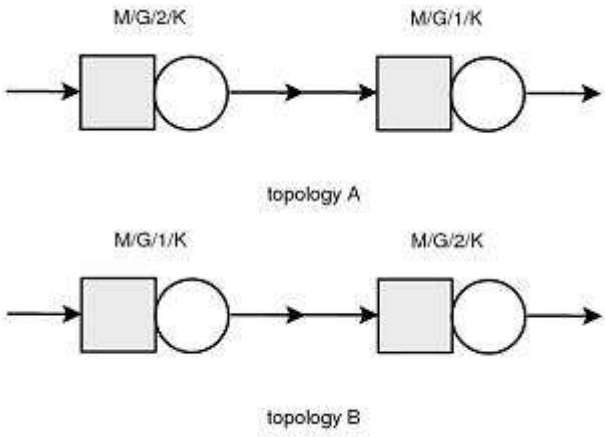

Figura 2: Topologias de redes com dois nós e três servidores.

No nosso primeiro experimento, cujos resultados apresentamos na Tabela 1, fixamos a taxa de chegada no sistema em λ=1 e λ = 2. A taxa de serviço foi fixada como μ = 4 e μ = 8, iguais em cada servidor. Utilizamos diversos valores para o coeficiente de variação do tempo de serviço s^2, para analisar se as alocações de áreas de espera são afetadas por ele. Queremos de examinar quais são as alocações de áreas espera necessárias para estas diferentes topologias. Com o objetivo de analisar a precisão dos resultados analíticos obtidos no algoritmo de otimização, foram feitas simulações, com 20 replicações, para determinação de intervalos de confiança, adotando um período de estabilização (do inglês *burn-in*) com 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, 2}, utilizamos a distribuição Gama com parâmetros α e β adequados. O computador usado neste estudo foi um CyberPower PC com processador AMD Athlon XP 1800+ GHz e 512 MB de memória RAM com Windows XP.

Os resultados de um modo geral são bastante encorajadores, como visto na Tabela 1. A 9_a coluna de resultados na tabela se refere à semi-amplitude dos intervalos de 95%,de

confiança. Em praticamente todos os casos a taxa de saída analítica $\Theta(x)$ foi igual à taxa de saída simulada intervalar. Um ponto que merece destaque é que a mudanças no coeficiente de variação do tempo de serviço provocam variações na alocação da área de espera, principalmente em taxas de atendimento menores. Na alocação da área de espera entre as topologias A e B, existe uma pequena diferença na solução ótima Z_α e Z_α^s. Devido à simetria, é difícil dizer que uma topologia é melhor do que outra, simplesmente devido à otimização da alocação da área de espera. Assim podemos afirmar que não existe o domínio de uma topologia sobre a outra. Em uma outra experiência com redes de dois nós, queremos examinar se tempos de serviço diferentes (gargalos) afetam as topologias.

Tabela 1: Resultados da rede com dois nós e três servidores.

							Simulação		
λ	μ	s^2	c	x	Θ(x)	Z_α	$\Theta(x)^s$	δ	Z_α^s
1,0	(4,4)	0,5	(2,1)	(3,4)	0,999	8,000	0,997	0,001	9,710
			(1,2)	(4,3)	0,999	8,000	0,998	0,001	8,780
		1,0	(2,1)	(3,4)	0,998	9,000	0,997	0,001	9,670
			(1,2)	(4,3)	0,998	9,000	0,997	0,001	10,260
		2,0	(2,1)	(4,5)	0,999	10,000	0,999	0,001	10,380
			(1,2)	(3,4)	0,999	10,000	0,997	0,001	12,010
	(8,8)	0,5	(2,1)	(5,4)	1,000	5,000	0,993	0,001	12,180
			(1,2)	(2,3)	1,000	5,000	0,999	0,001	5,650
		1,0	(2,1)	(3,2)	0,999	6,000	0,993	0,001	12,350
			(1,2)	(3,2)	0,999	6,000	0,998	0,001	7,050
		2,0	(2,1)	(2,3)	0,999	6,000	0,993	0,001	12,280
			(1,2)	(3,2)	0,999	6,000	0,996	0,001	8,810
2,0	(4,4)	0,5	(2,1)	(7,7)	1,997	17,000	2,001	0,015	13,400
			(1,2)	(7,7)	1,997	17,000	1,997	0,001	16,800
		1,0	(2,1)	(9,9)	1,997	21,000	2,000	0,002	17,600
			(1,2)	(9,9)	1,997	21,000	1,999	0,002	18,900
		2,0	(2,1)	(9,11)	1,996	24,000	2,000	0,001	20,200
			(1,2)	(11,9)	1,996	24,000	1,997	0,001	23,500
	(8,8)	0,5	(2,1)	(4,4)	1,999	9,000	2,001	0,002	7,500
			(1,2)	(4,4)	1,999	9,000	1,998	0,001	10,400
		1,0	(2,1)	(4,5)	1,999	10,000	2,000	0,002	8,900
			(1,2)	(5,4)	1,999	10,000	2,000	0,002	8,800
		2,0	(2,1)	(4,5)	1,997	12,000	1,999	0,002	10,500
			(1,2)	(5,4)	1,997	12,000	1,994	0,001	14,900

Tabela 2: Resultados serviços heterogêneos.

							Simulação		
λ	μ	s^2	c	x	Θ(x)	Z_α	$\Theta(x)^s$	δ	Z_α^s
1,0	(4,8)	0,5	(2,1)	(3,3)	0,999	7,000	0,997	0,001	8,670
	(8,4)		(1,2)	(3,3)	0,999	7,000	0,999	0,001	6,720
	(4,8)	1,0	(2,1)	(3,3)	0,999	7,000	0,997	0,001	8,870
	(8,4)		(1,2)	(3,3)	0,999	7,000	0,998	0,001	8,130
	(4,8)	2,0	(2,1)	(4,3)	0,999	8,000	0,999	0,001	8,320
	(8,4)		(1,2)	(3,4)	0,999	8,000	0,996	0,001	10,930
2,0	(4,8)	0,5	(2,1)	(8,5)	1,998	15,000	1,999	0,002	14,100
	(8,4)		(1,2)	(5,8)	1,998	15,000	1,999	0,002	14,400
	(4,8)	1,0	(2,1)	(9,6)	1,998	17,000	2,001	0,002	14,300
	(8,4)		(1,2)	(6,9)	1,998	17,000	2,000	0,001	15,200
	(4,8)	2,0	(2,1)	(9,5)	1,997	17,000	2,000	0,001	14,300

(8,4)	(1,2)	(5,9)	1,997	17,000	1,995	0,002	19,200

Para isso, usamos tempos de serviço μ = 4 e μ = 8 alternadamente em cada servidor. Usamos também taxas de chegada λ=1 e λ= 2. Os resultados são apresentados na Tabela 2. Os resultados experimentais na Tabela 2 nos levam a algumas conclusões importantes. Eles indicam que o desempenho independe do tipo de topologia, ou seja, havendo ou não o mesmo tempo de serviço nos servidores. Os resultados analíticos e de simulação são concordantes, como mostram as taxas de saída analíticas $\Theta(x)$ e simuladas $\Theta(x)^s$. Novamente mudanças no coeficiente de variação do tempo de serviço provocaram variações na alocação de áreas de espera. Um ponto que merece destaque é que um maior número de alocações de área de espera são designadas ao gargalo, isto é, servidores com menor capacidade de atendimento. E por fim, percebemos que os resultados do algoritmo de otimização Z_α e a simulação Z_α^s são estáveis, ou seja, existe uma pequena variação destas medidas quando se modifica a configuração da rede.

6. ANÁLISE DE DESEMPENHO

Após algumas simulações, desejamos verificar o desempenho do algoritmo proposto. Como exemplo de bom desempenho, é esperado que em redes de filas maiores o tempo de convergência do algoritmo seja maior. Também é importante investigar se existe relação entre o coeficiente de variação do tempo de serviço e o tempo de convergência, pois verificamos que este coeficiente influencia na alocação ótima da área de espera. Tais hipóteses poderão ser testadas através de técnicas de planejamento de experimentos. Na seção a seguir apresentaremos esta abordagem.

6.1. REDES COM DOIS NÓS E TRÊS SERVIDORES

O experimento será realizado adotando-se três tipos de redes de filas conforme visto na Figura 3. Nestas redes de filas adotam-se taxas de chegada λ iguais a 1, 2 e 3. A taxa de atendimento μ é 4, para todos os servidores. São apresentadas redes com 4, 8 e 16 servidores, configurados em nós dois a dois. Para o coeficiente de variação do tempo de serviço s^2, serão usados valores entre 0,5 e 2. Os dados obtidos com a realização do experimento podem ser vistos no Anexo A. Serão analisados os tempos (em segundos) de convergência do algoritmo. O computador usado neste estudo é mesmo já descrito anteriormente. A ordem que os experimentos foram executados foi aleatorizada, o mesmo acontecendo com a taxa de chegada λ.

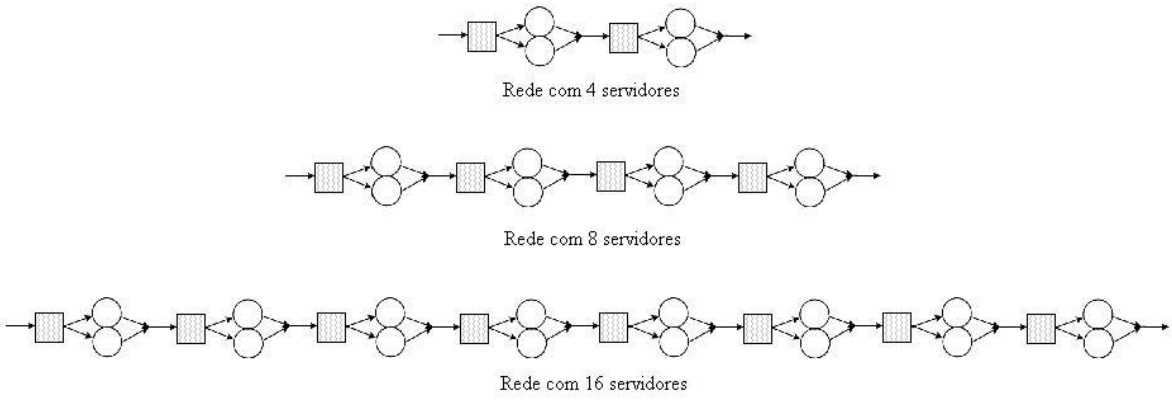

Figura 3: Topologias de redes com vários nós e servidores.

6.2. MODELO PROPOSTO

O modelo proposto para essa situação é um modelo fatorial (Montgomery, 1991), configurado em dois fatores (A e B) e um bloco, sendo os fatores e bloco fixos. Como estamos interessados em saber se redes mais complexas aumentam o tempo de convergência, o número de servidores (c) será considerado o fator A. O outro fator de interesse é o coeficiente de variação do tempo de serviço (s^2) chamado de fator B, sobre o qual queremos investigar a influência no tempo de convergência. Uma possível interação entre os fatores A e B também será investigada. A taxa de chegada (λ) será considerada bloco, pois não desejamos neste momento, investigar sua influência no algoritmo. O modelo proposto é dado por:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \delta_k + \varepsilon_{ijk} \quad (4)$$

para $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ e $k = 1, 2, \ldots, n$ onde:

- Y_{ijk} é a observação coletada sob o i-ésimo nível do fator A, j-ésimo nível do fator B e no k-ésimo bloco;
- μ é a média global;
- τ_i é o efeito do i-ésimo nível do fator A, sujeito a restrição $\sum \tau_i = 0$;
- β_j é o efeito do j-ésimo nível do fator B, sujeito a $\sum \beta_i = 0$;
- $(\tau\beta)_{ij}$ é o efeito da interação entre o i-ésimo nível do fator A e o j-ésimo nível do fator B, sujeito a restrição $\sum\sum (\tau\beta)_{ij} = 0$;
- λ_k é o efeito do k-ésimo bloco, sujeito a restrição $\sum \delta_k = 0$;
- ε_{ijk} é a componente de erro aleatório associado à observação Y_{ijk}.

Temos ainda a suposição de que os componentes de erro ε são variáveis aleatórias independentes e identicamente distribuídas com distribuição normal de média zero e variância σ^2, ou seja, $\varepsilon_{ijk} \sim iidN(0, \sigma^2)$.

Tabela 3: ANOVA para análise de desempenho.

Fonte de Variação	GL	SQ	QM	F	Valor-p
c	2	28,778	14,389	472,84	<0,000
s^2	2	0,1779	0,0889	2,92	0,083
c x s^2	4	0,0447	0,112	0,37	0,828
λ	2	2,9746	1,4873	48,87	<0,000
Erro	16	0,4869	0,0304		

Total	26	32,4621

6.3. ANÁLISE DOS RESULTADOS

Usamos neste experimento a transformação logarítmica para os tempos de convergência do algoritmo. Esta transformação foi necessária para satisfazer às suposições iniciais do modelo (4), tais como normalidade e homocedasticidade. Todas as demais suposições associadas ao modelo ajustado foram respeitadas (veja detahes no Anexo A). Na Tabela 3 apresentamos os resultados do ajuste do modelo obtido usando o Minitab (Minitab, 2006). Se adotarmos um nível de significância $\alpha = 0{,}05$, pela coluna de valores-*p* notamos que o fator A é significativo, o mesmo não acontecendo com o fator B, ou seja, o coeficiente de variação do tempo de serviço. Perceba que não existe interação entre os fatores A e B. Em relação ao fator A, aplicando um método de comparações múltiplas (Hsu, 1996) entre seus níveis, apresentado na Figura 4, notamos que os intervalos de confiança construídos não possuem o valor zero, indicando que existe diferença entre eles. A rede com 16 servidores possui um tempo de convergência maior que as redes com 8 e 4 servidores, além destes últimos também serem diferentes entre si, sendo o tempo de convergência na rede com 8 servidores maior que na com 4 servidores.

Intervalos Silmutâneos de Tukey com 95% de Confiança
Comparações entre os níveis de c

Nível c = 4 em comparação com:

```
c     inf.   central   sup.     ---------+---------+---------+-------
 8   1,071    1,283   1,495     (----*---)
16   2,317    2,529   2,741                          (----*---)
                                ---------+---------+---------+-------
                                      1,50      2,00      2,50
```

Nível c = 8 em comparação com:

```
c     inf.   central   sup.     ---------+---------+---------+-------
16   1,033    1,246   1,458     (---*---)
                                ---------+---------+---------+-------
                                      1,50      2,00      2,50
```

Figura 4: Comparações múltiplas entre os níveis de *c*.

7. CONCLUSÕES E OBSERVAÇÕES FINAIS

Apresentamos detalhadamente o problema de alocação de áreas de espera em redes de filas finitas com o serviço geral e servidores múltiplos. Descrevemos as fórmulas de probabilidade de bloqueio usadas nas experiências e a metodologia de otimização. Várias experiências ilustram o desempenho e as limitações da abordagem. Em geral, a alocação da área de espera obtida pelos algoritmos é simétrica para os casos testados e faz sentido. Os resultados foram satisfatórios, como resultados analíticos estavam dentro dos intervalos da confiança de 95%, estimados pela simulação. Um outro resultado interessante é que topologias completamente diferentes (por exemplo, topologias A e B nas redes de dois nós e três servidores) podem resultar em um desempenho similar, com alocação da área de espera ótima. Assim é difícil encontrarem-se regras heurísticas, tais como "o servidor múltiplo ocupa o primeiro lugar na topologia", antes de aplicar um procedimento de otimização para dizer

qual topologia é melhor. Sabe-se que uma topologia é direcionada geralmente pela aplicação, mas tal resultado pode trazer alguma flexibilidade para aqueles casos em que topologias alternativas estão competindo. Mostrou-se que o coeficiente de variação dos tempos de serviço é significativo na alocação da área de espera para redes uniformes e redes com gargalos. Com a realização do experimento planejado, concluímos que o algoritmo proposto apresenta resultados coerentes e que fazem sentido. Situações mais complexas devem ser investigadas, principalmente envolvendo a influência do coeficiente de variação do tempo de serviço no tempo de convergência do método de aproximação. Possivelmente resultados semelhantes devem ser alcançados. Esperamos que o leitor sinta o poder desta aproximação e sua capacidade em lidar com problemas complexos de planejamento de redes de filas finitas.

7.1. QUESTÕES EM ABERTO PARA TRABALHOS FUTUROS

Sobre as possíveis direções que esta pesquisa pode tomar, podemos citar a aplicação do algoritmo a problemas na área de manufatura e montagem, planejamento de plantas, bem como a problemas de planejamento de redes de telecomunicações e sistemas de computação.Não foram examinadas situações em que o número de servidores, c, fosse considerado uma variável de decisão. Isto poderá requerer uma reestruturação da abordagem e decidimos não tratar este aspecto neste momento. Um ponto 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. Uma outra abordagem é estudar outros tipos de redes de filas finitas, por exemplo, as filas *M/G/c/c* e também redes de filas infinitas tais como M/G/1. Outra possibilidade é incluir estudos sobre redes com laços de realimentação, muito encontrados em manufaturas e no setor de serviços. Os laços de alimentação causam grande dependência entre as chegadas e precisam de cuidadosa consideração. Estas são apenas algumas idéias interessantes para futuros trabalhos nesta área.

8. ANEXO A

A Tabela 4 apresenta os dados referentes ao experimento realizado, perceba nesta tabela que os resultados do experimento estão apresentados de forma completa, inclusive com os valores da função objetivo Z_α e o valor de $\Theta(x)$. Na Figura 5, nas suposições iniciais do modelo, pode ser notado que os dados transformados seguem a distribuição normal e não há violação de variabilidade constante entre os fatores e o bloco. A Figura 6 apresenta a análise residual do modelo ajustado, perceba que não há nenhuma violação quanto à normalidade, homocedasticidade e independência dos resíduos, indicando por isso, a validade do modelo (4) apresentado e dos resultados e conclusões obtidas.

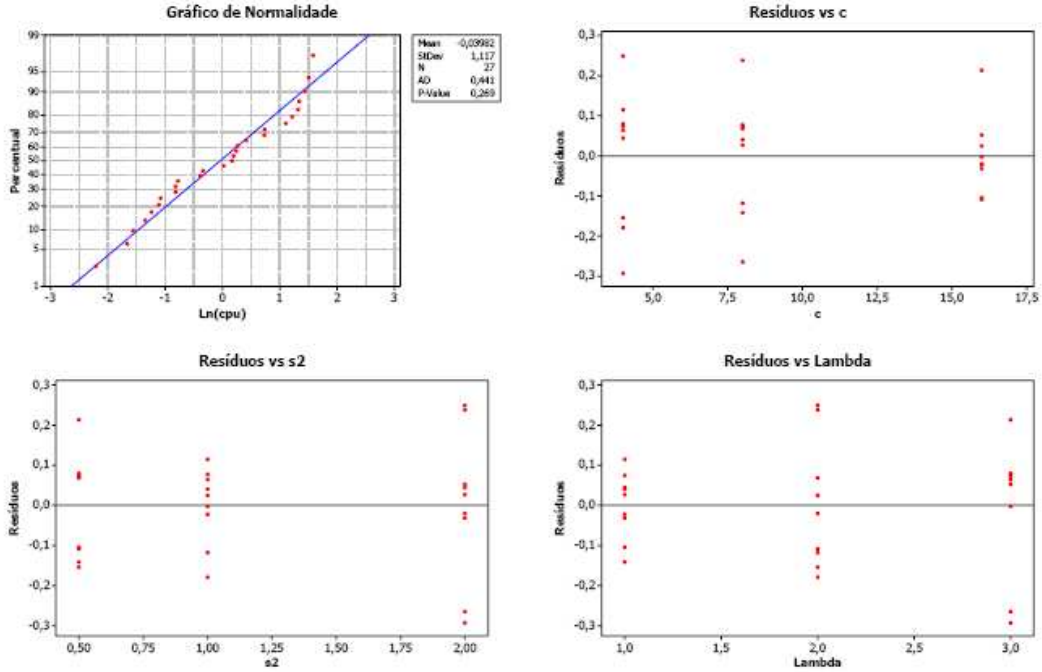

Figura 5: Suposições iniciais do modelo.

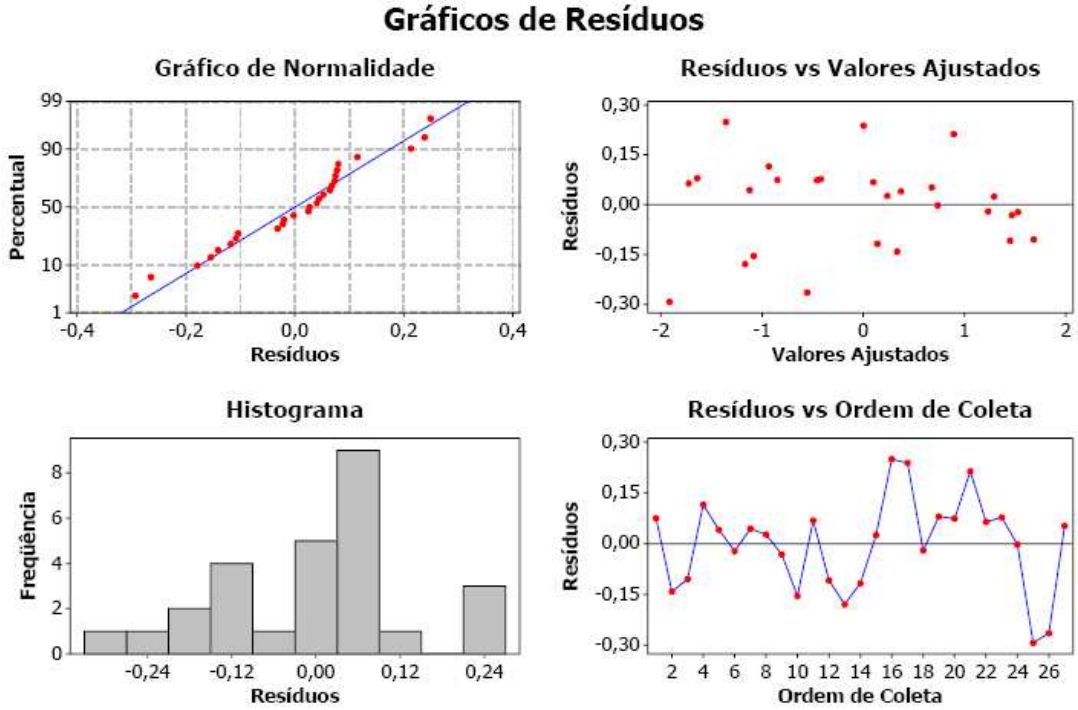

Figura 6: Análise residual do modelo ajustado.

Tabela 4: Resultados do experimento planejado.

λ	μ	s^2	c	x	Θ(x)	Z_α	CPU(s)
1,0	(4,4)	0,5	(2,2)	(3,3)	0,998	8,00	0,461
	(4,4,4,4)		(2,2,2,2)	(3,3,3,3)	0,100	15,00	1,212
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(3,3,3,3,3,3,3,3)	0,993	31,00	4,857
	(4,4)	1,0	(2,2)	(3,3)	0,998	8,00	0,441

	(4,4,4,4)		(2,2,2,2)	(3,3,3,3)	0,995	17,00	1,512
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(3,3,3,3,3,3,3,3)	0,991	33,00	4,506
	(4,4)	2,0	(2,2)	(3,3)	0,999	9,00	0,340
	(4,4,4,4)		(2,2,2,2)	(3,3,3,3)	0,998	18,00	1,302
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(3,3,3,3,3,3,3,3)	0,995	37,00	4,206
2,0	(4,4)	0,5	(2,2)	(7,7)	1,997	17,00	0,290
	(4,4,4,4)		(2,2,2,2)	(7,7,7,7)	1,994	34,00	1,182
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(7,7,7,7,7,7,7,7)	1,989	67,00	3,826
	(4,4)	1,0	(2,2)	(9,9)	1,997	21,00	0,260
	(4,4,4,4)		(2,2,2,2)	(9,9,9,9)	1,995	41,00	1,021
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(8,8,8,8,8,8,8,8)	1,990	74,00	3,735
	(4,4)	2,0	(2,2)	(3,3)	1,996	22,00	0,330
	(4,4,4,4)		(2,2,2,2)	(3,3,3,3)	1,992	44,00	1,272
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(9,9,9,9,9,9,9,9)	1,985	87,00	3,365
3,0	(4,4)	0,5	(2,2)	(15,2)	2,993	37,00	0,210
	(4,4,4,4)		(2,2,2,2)	(15,15,15,15)	2,987	73,00	0,681
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(15,15,15,15,15,15,15,15)	2,975	145,00	3,024
	(4,4)	1,0	(2,2)	(17,2)	2,993	41,00	0,190
	(4,4,4,4)		(2,2,2,2)	(17,17,17,17)	2,986	82,00	0,711
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(17,17,17,17,17,17,17,17)	2,973	163,00	2,083
	(4,4)	2,0	(2,2)	(21,2)	2,992	50,00	0,110
	(4,4,4,4)		(2,2,2,2)	(20,20,20,20)	2,980	100,00	0,441
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(20,20,20,20,20,20,20,20)	2,963	197,00	2,073

9. AGRADECIMENTOS

Durante o desenvolvimento deste trabalho, Helinton A. L . Barbosa era bolsista de iniciação científica pela Fundação de Amparo à Pesquisa de Minas Gerais, FAPEMIG.

A pesquisa de Frederico R. B. Cruz contou com o apoio parcial do CNPq, processos 201046/1994-6, 301809/1996-8, 307702/2004-9 e 472066/2004-8, da FAPEMIG, processos CEX-289/98 e CEX-855/98, e da PRPq da UFMG, processo 4081-UFMG/RTR/FUNDO/PRPQ/99.

10. REFERÊNCIAS BIBLIOGRÁFICAS

Gelenbe, E. (1975), ‘On approximate computer system models’, *Journal of the ACM* 22(2), 261–269.

Gross, D. & Harris, C. (1985), *Fundamentals of Queueing Theory*, J. Wiley & Sons, NewYork.

Harris, J. H. & Powell, S. G. (1999), ‘An algorithm for optimal buffer placement in reliable serial lines’, *IIE Transactions* 31, 287–302.

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

Hsu, J. C. (1996), Multiple Comparisons, Theory and methods, Chapman & Hall.

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

Kim, N. K. & 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.

Kimura, T. (1996a), 'Optimal buffer design of an M/S/s queue with finite capacity', *Communications in Statistics - Stochastic Models* 12(1), 165–180.

Kimura, T. (1996b), 'A transform-free approximation for the finite capacity M/G/s queue', *Operations Research* 44(6), 984–988.

Lemaréchal, C. (2003), 'The omnipresence of Lagrange', 4OR 1, 7–25.

Minitab (2006), Minitab user's guide 1 and 2, *in* 'url:http//www.minitab.com'.

Montgomery, D. (1991), *Design and Analysis of Experiments*, 3rd edn, John Wiley & Sons, New York.

Onvural, R. O. (1990), 'Survey of closed queueing networks with blocking', *ACM Computing Surveys* 22(2), 83–121.

Sakasegawa, H., Miyazawa, M. & Yamazaki, G. (1993), 'Evaluating the overflow probability using the infinite queue', Management Science 39(10), 1238–1245.

Schweitzer, P. J. & Konheim, A. G. (1978), 'Buffer overflow calculations using an infinitecapacity model', *Stochastic Processes and their Applications* 06(3), 267–276.

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

Smith, J. M. & Cruz, F. R. B. (2005), 'The buffer allocation problem for general finite buffer queueing networks', *IIE Transactions on Design & Manufacturing* 37(4), 343–365.14

Smith, J. M., Cruz, F. R. B. & van Woensel, T. (2006), 'Topological network design of general, finite, multi-server queueing networks', *manuscript submitted for publication* .

Spinellis, D., Papadoulos, C. T. & Smith, J. M. (2000), 'Large production line optimization using simulated annealing', *International Journal of Production Research* 38(3), 509–541.

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

Tijms, H. C. (1992), 'Heuristics for finite-buffer queues', *Probability in the Engineering and Informational Sciences* 06, 267–276.

Tijms, H. C. (1994), *Stochastic Modelling*: An Algorithmic Approach, JohnWiley & Sons, New York.

Yamashita, H. & Onvural, R. (1994), 'Allocation of buffer capacities in queueing networks with arbitrary topologies', *Annals of Operations Research* 48, 313–332.