Production, v. 24, n. 1, p. 13-25, jan./mar. 2014
doi: http://dx.doi.org/10.1590/S0103-65132013005000043

Proposta de análise de desempenho de algoritmos para otimização de redes de filas M/G/c/K baseada em DOE

Helinton André Lopes Barbosa[a], Gabriel Bahia Caldas[b], Frederico Rodrigues Borges da Cruz[c]

[a]helinton@ufmg.br, UFMG, Brasil
[b]gabrielbc@ufmg.br, UFMG, Brasil
[c]*fcruz@ufmg.br, UFMG, Brasil

Resumo

Neste artigo são apresentados resultados da análise empírica de um algoritmo proposto na literatura para alocação de áreas de espera em redes de filas finitas, abertas e acíclicas, com serviços gerais e servidores múltiplos. Dos resultados computacionais é concluído que o tempo de processamento do algoritmo depende do número de servidores da rede, como era de se esperar, mas independe do quadrado do coeficiente de variação do tempo de serviço. Conclui-se também que as alocações obtidas são robustas e que, em geral, o desempenho global previsto para a rede é acurado, conforme atestado por simulações. Finalmente, chega-se à conclusão de que não é fácil encontrar regras heurísticas para o posicionamento dos servidores múltiplos na rede de filas sem aplicar um algoritmo de alocação de áreas de espera para determinar qual configuração é a melhor.

Palavras-chave

Otimização. Avaliação de desempenho. Processos estocásticos. Delineamento de experimentos.

1. Introdução

Modelos baseados em redes de filas são muito úteis para representar sistemas de manufatura discretos em geral (BITRAN; MORABITO, 1995, 1996; SELLITTO; BORCHARDT; PEREIRA, 2008). Existem, entretanto, outras aplicações, como, e.g., na modelagem de caixas de supermercados (MORABITO; LIMA, 2000), de sistemas de atendimento médico de urgência (TAKEDA; WIDMER; MORABITO, 2004; IANNONI; MORABITO, 2006, 2008) e de serviços de correio eletrônico (DOY et al., 2006). Em particular, modelos de filas podem ser usados também para representar sistemas *job-shop* (SILVA; MORABITO, 2007a, b), que são um tipo de processo em que é produzido um elevado número de artigos diferentes, normalmente em pequenas quantidades, e, frequentemente, de acordo com determinadas especificações do cliente. No caso de *job-shops*, os nós dessas redes de filas representam as estações de trabalho (*shops*) e os produtos (*jobs*) representam os usuários com demanda por serviço nessas estações de trabalho. Os arcos que conectam os nós da rede correspondem às rotas dos produtos.

Há vários tipos de redes de filas e uma descrição detalhada dos tipos mais populares pode ser encontrada na literatura (SILVA; MORABITO, 2007a). Nesse contexto, há o interesse por um tipo particular de filas, as filas finitas (isto é, com uma capacidade limitada) e com tempos de serviço gerais. Na conhecida notação de Kendall (1953) são as redes compostas por filas do tipo M/G/c/K, em que M corresponde a um processo de chegada markoviano (modelado pela distribuição exponencial), o G, a um tempo de serviço com distribuição geral, c é o número de servidores em paralelo e, finalmente, K é o número máximo de usuários no sistema incluindo aqueles em serviço (isto é, $K = c + x$, em que x é o tamanho da área de espera; do inglês *buffer*).

Uma das razões do interesse pelas filas M/G/c/K é a sua flexibilidade em modelar áreas de espera

*UFMG, Belo Horizonte, MG, Brasil
Recebido 10/08/2011; Aceito 19/03/2012

finitas e taxas de serviço gerais, que são hipóteses bastante convenientes em aplicações reais (SMITH; CRUZ, 2005). Se por um lado as redes de filas M/G/c/K têm tal flexibilidade, por outro a capacidade finita de áreas de espera abre a possibilidade de ocorrência do fenômeno de bloqueio, que é quando um usuário não pode seguir à fila seguinte, quando ela tem esgotada a sua capacidade máxima *K*. O bloqueio, agravado pela consideração de tempos de serviço gerais, acarreta características na forma não produto. Formas não produto dificultam a determinação de medidas de desempenho de cada fila M/G/c/K individualmente (SMITH, 2003) e tornam-se um problema ainda maior quando essas filas estão configuradas em redes (SMITH; CRUZ; VAN WOENSEL, 2010).

O presente artigo é uma complementação de um artigo recentemente publicado (SMITH; CRUZ; VAN WOENSEL, 2010) em que, segundo seus autores, foi proposto o primeiro algoritmo para alocação ótima de áreas de espera em redes de filas M/G/c/K abertas e acíclicas. Desde que foi publicado esse algoritmo, algumas questões permaneceram em aberto a respeito do seu desempenho. Este artigo pretende responder algumas delas. Dos resultados computacionais, conclui-se que o tempo de processamento do algoritmo depende do número de servidores da rede, como era de se esperar, mas independe do quadrado do coeficiente de variação do tempo de serviço. Confirma-se também que as alocações obtidas são robustas e que, em geral, o desempenho global previsto para a rede é acurado, conforme atestado por simulações, em configurações não anteriormente testadas. Finalmente, chega-se à conclusão de que não é fácil encontrar regras heurísticas para o posicionamento dos servidores múltiplos na rede de filas sem aplicar um algoritmo de alocação de áreas de espera para determinar qual configuração é a melhor.

Na próxima seção são descritas as origens do problema e os trabalhos anteriores relacionados a ele, bem como apresentados os modelos matemáticos apropriados à análise das redes de filas e os algoritmos empregados para sua otimização. Na seção 3 são apresentados resultados experimentais obtidos para diferentes topologias de redes de filas, através de um experimento planejado. Na seção 4 são discutidos os resultados obtidos. Por fim, na seção 5 são apresentadas conclusões e observações finais, além de levantados tópicos para possíveis trabalhos futuros na área.

2. Materiais e métodos

A exemplo de vários trabalhos publicados na área de engenharia de produção (YANASSE; BECCENERI; SOMA, 2007; ARGOUD; GONÇALVES FILHO; TIBERTI, 2008), este artigo apresenta resultados de uma pesquisa de natureza aplicada com caráter experimental, de acordo com classificação apresentada em Miguel (2010). Em seguida, o problema é formalizado matematicamente, pois o algoritmo de resolução é derivado diretamente da sua formulação. Passa-se inicialmente à definição da notação utilizada.

2.1. *Notação*

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

- $G = (N, A)$, grafo direcionado, em que N é o conjunto de nós da rede (filas do tipo M/G/c/K) e A é o conjunto de arcos da rede (ou pares de nós conectados);
- $p = (..., p_{ij}, ...)$, vetor das probabilidades de roteamento nos arcos $(i,j) \in A$;
- λ_i, taxa de chegada Poisson (markoviana) na fila $i \in N$;
- μ_i, taxa de serviço (com distribuição geral G) na fila $i \in N$;
- cv_i^2, quadrado do coeficiente de variação do tempo de serviço na fila $i \in N$, definido pela razão entre a variância e o quadrado do valor esperado do tempo de serviço T_S, isto é, $V(T_S)/E(T_S)^2$;
- c_i, número de servidores em paralelo na fila $i \in N$;
- $\rho_i = \lambda_i/(c_i\mu_i)$, intensidade de tráfego na fila $i \in N$;
- K_i, capacidade total da fila $i \in N$, *incluindo* os itens em serviço;
- p_{Ki}, probabilidade de bloqueio, i.e., probabilidade de um item encontrar a fila i cheia;
- $x_i = K_i - c_i$, capacidade da área de espera da fila $i \in N$;
- $\Theta(x)$, taxa de atendimento (do inglês, *throughput*) global da rede, em função do vetor de alocação de áreas de espera, $x = (x_1, x_2, ..., x_n)$, em que n é a cardinalidade do conjunto N;
- Θ^τ, taxa de atendimento global mínima requerida.

2.2. *Formulação matemática*

Um modelo de programação matemática inteira para o problema de alocação de áreas de espera em redes de filas M/G/c/K (SMITH; CRUZ; VAN WOENSEL, 2010), definido sobre o grafo direcionado $G = (N, A)$, é apresentado a seguir.

Modelo (M):

$$Z = min \sum_{\forall i \in N} x_i \quad (1)$$

sujeito a:

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

$$x_i \in \{0,1,\ldots\}, \forall i \in N \tag{3}$$

Note-se que o conjunto N (nós da rede), o conjunto A (arcos da rede) e a taxa de atendimento global mínima requerida (Θ^{τ}) são parâmetros do modelo (M). Por sua vez, os x_i, com $i \in N$, são as variáveis do modelo (M). Finalmente, a taxa $\Theta(x)$ é determinada algoritmicamente a partir dos parâmetros e das variáveis do modelo (M), conforme será explicitado na subseção 2.4.

Note-se também que, apesar de a função objetivo ser linear nas variáveis de decisão x_i, esse é um problema de otimização não linear, por causa da restrição (2). Além disso, o modelo (M) envolve variáveis de decisão x_i inteiras. Finalmente, é importante ressaltar que a medida de desempenho aqui considerada, $\Theta(x)$, não é a única possibilidade. De fato, pode-se encontrar na literatura o exame de problemas de alocação em redes de filas que consideram diferentes medidas de desempenho, tais como, e.g., o *work-in-process* e o *lead time* (BITRAN; MORABITO, 1995; SILVA; MORABITO, 2007b) ou, ainda, várias medidas de desempenho conflitantes (análise de *trade-off*) tomadas simultaneamente (BITRAN; MORABITO, 1996; CRUZ; VAN WOENSEL; SMITH, 2010).

2.3. *Análise de desempenho em filas únicas*

Quando se trata com uma fila finita única, a taxa de atendimento $\Theta(x)$ relaciona-se diretamente com a taxa de chegada λ e a probabilidade de bloqueio p_K, que é a probabilidade de um item encontrar o sistema cheio (isto é, o número de itens no sistema j iguala-se à sua capacidade total K):

$$\Theta(x) = \lambda(1 - p_K) \tag{4}$$

Assim, o problema de determinação da medida de desempenho $\Theta(x)$ fica condicionado apenas ao conhecimento da taxa de chegada λ e à determinação da probabilidade de bloqueio p_K.

Para sistemas finitos markovianos puros, com servidor único, isto é, filas M/M/1/K, com $\rho < 1$, de acordo com Gross et al. (2009), a probabilidade de bloqueio pode ser escrita como:

$$p_K = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} \tag{5}$$

Se for relaxada a sua restrição de integralidade de K, chega-se a uma expressão em forma fechada para o tamanho ótimo da capacidade total da fila, em função da intensidade de tráfego ρ e da probabilidade de bloqueio p_K:

$$K_M = \left\lceil \frac{ln\left(\frac{p_K}{1-\rho+p_K\rho}\right)}{ln(\rho)} \right\rceil \tag{6}$$

em que $\lceil x \rceil$ é o menor inteiro não inferior a x. Por conseguinte, está determinada a alocação ótima da área de espera para filas M/M/1/K:

$$x_M = K_M - 1 \tag{7}$$

Para filas M/G/c/K a determinação da probabilidade de bloqueio torna-se um problema bem mais complicado e parece improvável a existência de um método exato geral. Entretanto, em artigos anteriores (SMITH; CRUZ, 2005; SMITH; CRUZ; VAN WOENSEL, 2010) foi mostrado que o esquema de aproximação a dois momentos de Kimura (1996), baseado na expressão markoviana, Equação 7, produz resultados satisfatórios:

$$x_{\varepsilon,\text{Kimura}}\left(cv^2\right) = x_M + \text{INT}\left[\frac{\left(cv^2-1\right)\sqrt{\rho}}{2} x_M\right] \tag{8}$$

em que INT(x) representa a parte inteira de x.

Para filas M/G/1/K, e.g., com uma intensidade de tráfego ρ e um dado quadrado do coeficiente de variação do tempo de serviço (geral), cv^2, uma aproximação para a área de espera ótima é:

$$x_{\varepsilon,\text{Kimura}} = \frac{\left(ln\left(\frac{p_K}{1-\rho+p_K\rho}\right)+ln(\rho)\right)\left(2+\sqrt{\rho}cv^2-\sqrt{\rho}\right)}{2\,ln(\rho)} \tag{9}$$

Por conseguinte, pode-se explicitar p_K e determinar uma expressão em forma fechada para a probabilidade de bloqueio para uma fila M/G/1/K, em função de K (para filas M/G/1/K, note-se que $K = 1 + x_{\varepsilon,\text{Kimura}}$):

$$p_K = \frac{(1-\rho)\rho^{\left(\frac{2+\sqrt{\rho}cv^2-\sqrt{\rho}+2(K-1)}{2+\sqrt{\rho}cv^2-\sqrt{\rho}}\right)}}{1-\rho^{\left(2\frac{2+\sqrt{\rho}cv^2-\sqrt{\rho}+(K-1)}{2+\sqrt{\rho}cv^2-\sqrt{\rho}}\right)}} \tag{10}$$

Pode-se continuar esse processo de desenvolvimento de p_K, para diferentes valores de c, obtendo-se formas fechadas aproximadas para a

probabilidade de bloqueio (SMITH, 2003), em sistemas M/G/c/K, para $c = 2, 3, ...$, e, consequentemente, sua taxa de atendimento, $\Theta(x)$, pela Equação 4.

2.4. *Análise de desempenho em redes de filas*

O problema de análise de desempenho em filas finitas torna-se muito mais complexo quando elas estão configuradas em redes. O método da expansão generalizado (GEM, do inglês, *generalized expansion method*) é uma técnica robusta e bastante eficaz de aproximação de medidas de desempenho de redes de filas finitas (KERBACHE; SMITH, 1987). O método é caracterizado por uma combinação de tentativas repetidas e decomposição nó a nó, para cada fila *i* que for sucedida por uma fila finita *j*, conforme apresentado na Figura 1.

O GEM possui três estágios, descritos a seguir, após a definição de uma notação adicional:

- h_j, nó artificial, adicionado pelo GEM, antecedendo cada fila finita encontrada na rede;
- $\tilde{\lambda}_j$, taxa de chegada efetiva à fila *j* (descontados os itens que são bloqueados);
- $\tilde{\mu}_i$, taxa de serviço efetiva na fila *i* (devido ao bloqueio que sofreu da fila subsequente *j*);
- p_{Kj}', probabilidade de bloqueio no laço de retroalimentação no GEM.

Estágio I – Reconfiguração da rede

Usando o princípio das duas fases da fila finita *j* (saturada ou insaturada), uma fila artifical de espera h_j, infinita, com um número infinito de servidores, do tipo M/G/∞, é adicionada para cada fila finita na rede. A finalidade da fila de espera é registrar os itens bloqueados (ver Figura 1). Essa fila modela o atraso adicional, causado àqueles clientes que tentam entrar na fila *j* e a encontram cheia, o que ocorre com probabilidade p_{Kj}. Os itens são bem-sucedidos na tentativa de entra na fila *j*, com uma probabilidade $(1-p_{Kj})$. Com essa fila artificial também são incluídos novos arcos na rede, com probabilidades de roteamento p_{Kj}', caso o item continue bloqueado para um segundo período de atraso, e $(1-p_{Kj}')$, caso possa prosseguir para a fila finita *j*. Esse processo continua até que se encontre um espaço na fila finita *j*. Um arco de retroalimentação é utilizado para modelar esses repetidos atrasos. A fila artificial de espera é modelada como uma fila do tipo M/G/∞ porque é usada simplesmente para dar ao item bloqueado um tempo extra de atraso, sem enfrentar filas.

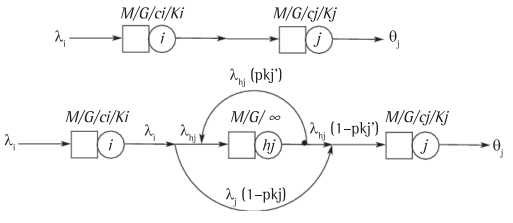


Figura 1. O método da expansão generalizado para duas filas adjacentes, *i* e *j*.

Estágio II – Estimação de parâmetros

Nesse estágio, estima-se aproximadamente os parâmetros p_{Kj}, p_{Kj}' e m_{hj}, via resultados conhecidos para filas M/G/c/K, conforme descrito a seguir. Por simplicidade, será omitido o subíndice *j*, referente à *j*-ésima fila finita.

p_K: as probabilidades de bloqueio podem ser obtidas pela utilização de resultados analíticos aproximados (neste artigo, será via aproximação a dois momentos de Kimura), como, e.g., para filas M/G/1/K, repetida a seguir por clareza,

$$p_K = \frac{(1-\rho)\rho^{\left(\frac{2+\sqrt{\rho}cv^2-\sqrt{\rho}+2(K-1)}{2+\sqrt{\rho}cv^2-\sqrt{\rho}}\right)}}{1-\rho^{\left(2\frac{2+\sqrt{\rho}cv^2-\sqrt{\rho}+(K-1)}{2+\sqrt{\rho}cv^2-\sqrt{\rho}}\right)}} \quad (10)$$

e, de forma similar, expressões para filas M/G/c/K, para $c = 2, 3, ... 10, ...$, podem ser incluídas aqui, de forma a ter-se um conjunto completo de probabilidades de bloqueio;

p_K': não há uma forma fechada para essa probabilidade (probabilidade de um segundo bloqueio) e utiliza-se a seguinte aproximação, obtida por técnicas de difusão (LABETOULLE; PUJOLLE, 1980),

$$p_K' = \left\{\frac{\mu_j+\mu_h}{\mu_h} - \frac{\lambda\left[\left(r_2^K - r_1^K\right) - \left(r_2^{K-1} - r_1^{K-1}\right)\right]}{\mu_h\left[\left(r_2^{K+1} - r_1^{K+1}\right) - \left(r_2^K - r_1^K\right)\right]}\right\}^{-1} \quad (11)$$

em que r_1 e r_2 são raízes do polinômio

$$\lambda - \left(\lambda + \mu_h + \mu_j\right)x + \mu_h x^2 = 0 \quad (12)$$

em que $\lambda = \lambda_j - \lambda_h(1-p_K')$ e λ_j e λ_h são taxas de chegadas efetivas à fila finita *j* e à fila artificial *h*, respectivamente;

μ_h: a distribuição do tempo de atraso causado por bloqueio na fila *j* é assumida ser a mesma da fila *j* e, por meio da teoria da renovação, é possível mostrar que o tempo de serviço na fila de espera possui média

$$\mu_h = \frac{2\mu_j}{1+\sigma_j^2\mu_j^2} \quad (13)$$

em que σ_j^2 é a variância do tempo de serviço (KLEINROCK, 1975).

Estágio III – Eliminação da retroalimentação

Devido ao laço de retroalimentação em torno da fila de espera, haverá uma grande dependência no processo de chegada à fila *j*. A eliminação dessas dependências requer a reconfiguração da fila de espera, o que pode ser feito por um ajuste no seu tempo de serviço, dado por:

$$\mu_h' = (1 - p_K')\mu_h \tag{14}$$

As probabilidades de a fila *j* estar em uma das duas fases (saturada ou não saturada) são p_K e $(1-p_K)$, respectivamente. Assim, o tempo de serviço médio na fila *i*, que precede uma fila finita *j*, é μ_i^{-1}, quando na fase não saturada, e $[\mu_i^{-1} + (\mu_h')^{-1}]$, na fase saturada. Portanto, o tempo médio de serviço no *i* é dado por:

$$\tilde{\mu}_i^{-1} = \mu_i^{-1} + p_K \mu_h^{-1} \tag{15}$$

Equações similares podem ser estabelecidas para cada um das filas finitas, não somente em configurações em série mas também em outras configurações acíclicas mais gerais.

2.5. *Algoritmo de otimização*

Uma forma eficiente de resolver o problema de alocação de áreas de espera aqui examinado, definido pelas Equações 1-3, é pela incorporação da restrição (2) na função objetivo (SMITH; CRUZ; VAN WOENSEL, 2010), através de uma função de penalidade, tal como a relaxação lagrangeana (LEMARÉCHAL, 2007). Assim, definindo-se uma variável dual α e relaxando-se a restrição (2), o problema relaxado a seguir é obtido.

Modelo (MR):

$$Z_\alpha = min\left[\sum_{\forall i \in N} x_i + \underbrace{\alpha\left(\Theta^\tau - \Theta(x)\right)}_{\leq 0}\right] \tag{16}$$

sujeito a:

$$x_i \in \{0,1,\ldots\}, \forall i \in N \tag{17}$$

$$\alpha \geq 0 \tag{18}$$

A taxa de atendimento global mínima requerida Θ^τ pode ser pré-especificada e servir como taxa de entrada λ de um algoritmo aproximado para determinação de medidas de desempenho, como o GEM, que fornecerá uma taxa de saída correspondente. Nesse caso, para um vetor x *viável* para o problema (1)-(3), o termo $\alpha(\Theta^\tau - \Theta(x))$ será sempre não positivo (a taxa de entrada, Θ^τ, nunca poderá exceder a taxa de saída, $\Theta(x)$) e será uma penalidade da função objetivo. Segue assim que $Z_\alpha \leq Z$, isto é, Z_α será um limite inferior para *Z*, que é o valor ótimo da função objetivo do problema (1)-(3).

O melhor limite inferior será dado pela solução ótima do problema a seguir, conhecido como dual lagrangeano.

Modelo (DL):

$$max\ Z_\alpha \tag{19}$$

sujeito a:

$$x_i \in \{0,1,\ldots\}, \forall i \in N \tag{20}$$

$$\alpha \geq 0 \tag{21}$$

É possível perceber que se a taxa de atendimento global mínima requerida Θ^τ for exatamente igual à taxa de chegada externa λ, o melhor (maior) limite inferior dado pelo modelo (DL) será alcançado quando $\alpha \to \infty$, o que não é prático, pois exigiria que $(\Theta^\tau - \Theta(x)) = 0$ e, por conseguinte, que $x_i \to \infty$. Por outro lado, se uma "pequena" diferença, diga-se $(\Theta^\tau - \Theta(x)) = \varepsilon$, for aceitável, será necessário verificar que $\alpha(\Theta^\tau - \Theta(x)) \leq 1$, pois caso contrário teria sido melhor utilizar uma unidade extra de área de espera, isto é, x_i+1, em alguma fila $i \in N$, para aumentar $\Theta(x)$ (lembre-se que $\Theta(x)$ é uma função não decrescente de x). Dessa forma, é possível definir um α_ε correspondente, como se segue:

$$\alpha_\varepsilon \leq \frac{1}{\left(\Theta^\tau - \Theta(x)\right)} \tag{22}$$

o qual, assumindo-se, e.g., $(\Theta^\tau - \Theta(x)) \leq 10^{-3}$, resultará em $\alpha_\varepsilon = 10^3$ (será o valor aqui adotado).

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. Assim, a fim de resolver aproximadamente o problema (1)-(3), o GEM será acoplado a um clássico algoritmo de busca, o algoritmo de Powell.

O método de Powell, apresentando esquematicamente na Figura 2, encontra o mínimo de uma função não linear $f(x)$ por meio de sucessivas buscas unidimensionais, a partir de um ponto inicial $x^{(0)}$, via um conjunto de direções conjugadas, que são geradas dentro do próprio procedimento. Ele é baseado na idéia de que um mínimo de uma função não linear $f(x)$ pode ser encontrado ao longo de *n* (dimensão do problema) direções conjugadas em um estágio da busca, com um passo adequado em cada

```
algoritmo
  leia G(N, A), p, λ, μ, x^(0)
  /* escolha n direções de busca linearmente independentes */
  escolha d^(1), ..., d^(n)
  x^(opt) ← x^(0)
  repita
    x^(1) ← x^(opt)
    para i =1 até n faça
      /* executar uma busca unidimensional */
      /* f(•) calculado com a ajuda do GEM */
      x^(i+1) ← {x*|f(x*) = min_{γ∈ℜ} f(x^(i) + γd^(i))}
    fim para
    x^(n+2) ← 2x^(n+1) ← x^(1)
    se f(x^(n+2)) ≥ f(x^(1)) então
      x^(opt) ← x^(n+1)
    senão
      x^(opt) ← {x*| f(x*) = min_{γ∈ℜ} f(x^(n+1) + γ(x^(n+1) − x^(1)))}
      escolher novas d^(1), ..., d^(n)
    fim se
  até ||x^(opt) − x^(1)|| < ε
  imprima x^(opt)
fim algoritmo
```

Figura 2. Método de Powell.

direção. Maiores detalhes sobre o algoritmo de Powell podem ser encontrados na literatura (BAZARAA; SHERALI; SHETTY, 2006).

2.6. *Análise do desempenho do algoritmo de otimização*

É de interesse prático verificar como o algoritmo de otimização se comporta em termos de tempo de processamento até a convergência, em função de vários parâmetros da rede de filas finitas. Técnicas de planejamento de experimentos (DOE) são utilizadas para essa avaliação de desempenho do algoritmo. Em especial há interesse na influência que o número de servidores, *c*, exerce sobre o tempo até convergência. Também é importante investigar se existe relação entre o quadrado do coeficiente de variação do tempo de serviço, cv^2, e o tempo até convergência, pois se verifica que, em princípio, o cv^2 influencia na alocação ótima das áreas de espera.

O delineamento probabilístico proposto para essa situação é um modelo fatorial (MONTGOMERY, 2008), configurado em dois fatores (*A* e *B*) e em um bloco, sendo fixos tanto os fatores quanto o bloco. Como há interesse em saber se redes mais complexas aumentam o tempo de convergência, o número total de servidores na rede ($C = \Sigma_{\forall i \in N} c_i$) será considerado o fator *A*. O outro fator de interesse é o quadrado do coeficiente de variação do tempo de serviço (cv^2), chamado fator *B*. Uma possível interação entre os fatores *A* e *B* também será investigada. Note-se que a taxa de chegada (λ), outro parâmetro importante na área alocada, será considerada como bloco, pois não se deseja, nesse momento, investigar sua influência no tempo de convergência do algoritmo. O modelo proposto é dado por

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \gamma_k + \varepsilon_{ijk} \quad (23)$$

em que:

- Y_{ijk} é a observação coletada sob o *i*-ésimo nível do fator *A*, o *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 à restrição $\Sigma_i\tau_i = 0$;
- β_j é o efeito do *j*-ésimo nível do fator *B*, sujeito à restrição $\Sigma_j\beta_j = 0$;
- $(\tau\beta)_{ij}$ é o efeito da interação entre o *i*-ésimo nível do fator *A* e o *j*-ésimo nível do fator *B*, sujeito à restrição $\Sigma_i\Sigma_j(\tau\beta)_{ij} = 0$;
- γ_k é o efeito do *k*-ésimo bloco, sujeito à restrição $\Sigma_k\gamma_k = 0$;
- ε_{ijk} é a componente de erro aleatório associado à observação Y_{ijk}.

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

3. Resultados experimentais

Todos os algoritmos descritos foram codificados em FORTRAN, pela reconhecida eficiência e exatidão de suas sub-rotinas numéricas. Os códigos estão disponíveis a pedido, para fins educacionais e de pesquisa, diretamente com os autores. Inicialmente foi feita uma análise de desempenho do algoritmo de alocação de áreas de espera, em termos de tempo de processamento até a convergência. Em seguida, o algoritmo foi aplicado a algumas configurações simples, mas que permitiram conclusões interessantes a respeito do problema de alocação de áreas de espera.

3.1. *Análise de desempenho do algoritmo*

No que diz respeito aos níveis dos fatores e do bloco, o experimento foi realizado adotando-se três redes de filas, com $N \in \{2; 4; 8\}$, com dois servidores em cada fila, perfazendo-se o total de 4; 8; e 16 servidores, respectivamente, conforme visto na Figura 3. Nessas redes de filas serão adotadas taxas de chegada $\lambda \in \{1; 2; 3\}$, para uma taxa de atendimento única $\mu = 4$, para todos os servidores. Finalmente, para o quadrado do coeficiente de variação do tempo de serviço será considerado $cv^2 \in \{0{,}5; 1{,}0; 2{,}0\}$. A variável de interesse é o tempo (em segundos)

até a convergência do algoritmo. A ordem em que os experimentos foram executados foi aleatorizada, o mesmo acontecendo com a taxa de chegada λ.

Para essa análise foi utilizado um computador pessoal com o sistema operacional Windows 7. Os dados obtidos com a realização do experimento podem ser vistos na Tabela 1, que apresenta os dados referentes ao experimento realizado para a análise de desempenho do algoritmo. Além dos tempos até a convergência, em segundos, coluna CPU(s), resultados também disponibilizados são as alocações ótimas, x, as respectivas taxas de saída, Θ(x), e os valores da função objetivo, Z_α.

Usaram-se inicialmente, para a análise, os tempos de execução até a convergência do algoritmo na escala original (isto é, em segundos). Entretanto, foram violadas as suposições de normalidade e de homocedasticidade (i.e., variância constante dos erros) do modelo (23). Assim, usou-se a transformação logarítmica para os tempos de execução e todas as suposições associadas ao modelo ajustado foram respeitadas (MONTGOMERY, 2008), conforme pode

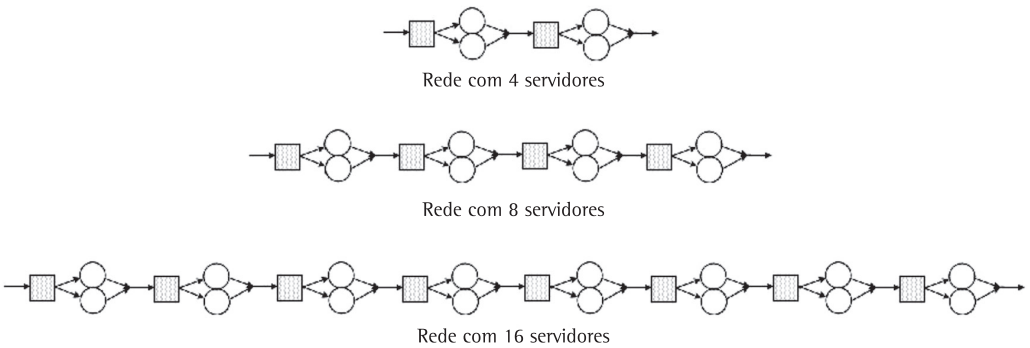


Figura 3. Redes de filas de teste na topologia série.

Tabela 1. Resultados para a análise de desempenho do algoritmo.

λ	μ	cv^2	c	x	Θ(x)	Z_α	CPU(s)
1,0	(4,4)	0,5	(2,2)	(3, 3)	0,998	07,66	0,109
	(4,4,4,4)		(2,2,2,2)	(3, 3, 3, 3)	0,997	15,30	0,265
	(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	30,60	1,014
	(4,4)	1,0	(2,2)	(3, 3)	0,998	08,34	0,172
	(4,4,4,4)		(2,2,2,2)	(3, 3, 3, 3)	0,995	16,60	0,296
	(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	33,20	1,123
	(4,4)	2,0	(2,2)	(4, 4)	0,999	09,21	0,109
	(4,4,4,4)		(2,2,2,2)	(4, 4, 4, 4)	0,998	18,40	0,312
	(4,4,4,4,4,4,4,4)		(2,2,2,2,2,2,2,2)	(4, 4, 4, 4, 4, 4, 4, 4)	0,995	36,80	0,858
2,0	(4,4)	0,5	(2,2)	(7, 7)	1,997	16,90	0,140
	(4,4,4,4)		(2,2,2,2)	(7, 7, 7, 7)	1,994	33,80	0,218
	(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	67,40	0,811
	(4,4)	1,0	(2,2)	(8, 8)	1,997	18,60	0,078
	(4,4,4,4)		(2,2,2,2)	(8, 8, 8, 8)	1,995	37,20	0,203
	(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	74,20	0,655
	(4,4)	2,0	(2,2)	(9, 9)	1,996	21,90	0,078
	(4,4,4,4)		(2,2,2,2)	(9, 9, 9, 9)	1,992	43,70	0,281
	(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	87,20	0,577
3,0	(4,4)	0,5	(2,2)	(15, 15)	2,993	036,70	0,047
	(4,4,4,4)		(2,2,2,2)	(15, 15, 15, 15)	2,987	073,10	0,125
	(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	0,437
	(4,4)	1,0	(2,2)	(17, 17)	2,993	041,20	0,062
	(4,4,4,4)		(2,2,2,2)	(17, 17, 17, 17)	2,986	082,10	0,140
	(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	0,499
	(4,4)	2,0	(2,2)	(20, 20)	2,992	050,00	0,047
	(4,4,4,4)		(2,2,2,2)	(20, 20, 20, 20)	2,980	099,60	0,187
	(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	0,390

ser conferido na Figura 4, que apresenta a análise residual do modelo ajustado, equação (23), e também no Anexo A, que verifica as suas suposições iniciais. Note-se que não há nenhuma violação quanto à normalidade, homocedasticidade e independência dos resíduos, indicando a validade do modelo, bem como dos resultados e conclusões obtidos a partir dele. Pela Figura 4, nota-se também a presença de um *outlier* (fora da faixa ± 2), o que geralmente ocorre quando se aplica o DOE a metamodelos (equações simuladas por modelos probabilísticos associados às suas variáveis aleatórias), como é o caso de simulações Monte Carlo.

Na Tabela 2 são apresentados os resultados do ajuste do modelo (23). Na Figura 5 são apresentados os resultados da comparação múltipla (HSU, 1996), entre as médias dos tempos até convergência, para os diferentes níveis do fator *A* (número total de servidores, *C*). Foi escolhido o fator *A* por ter-se mostrado significativo na análise de variâncias, para um nível de significância de 5% ($\alpha = 0{,}05$). Todos os resultados foram obtidos por meio do pacote estatístico MINITAB (2006).

Note-se que está fora de escopo um estudo de análise de variância para as outras saídas importantes do algoritmo (p.e., $\Theta(x)$ e $Z\alpha$), por se tratar de um estudo do algoritmo de análise de desempenho (o GEM) e não do algoritmo de otimização em si, que é o principal objeto de estudo deste trabalho (para resultados da análise de desempenho do GEM, sugere-se consultar, p.e., ANDRIANSYAH et al., 2010; CRUZ; VAN WOENSEL; SMITH, 2010; SMITH; CRUZ; VAN WOENSEL, 2010).

3.2. *Análise das alocações obtidas*

Para uma análise das alocações ótimas fornecidas pelo algoritmo, foi utilizada uma das topologias mais simples de rede de filas, que é uma configuração

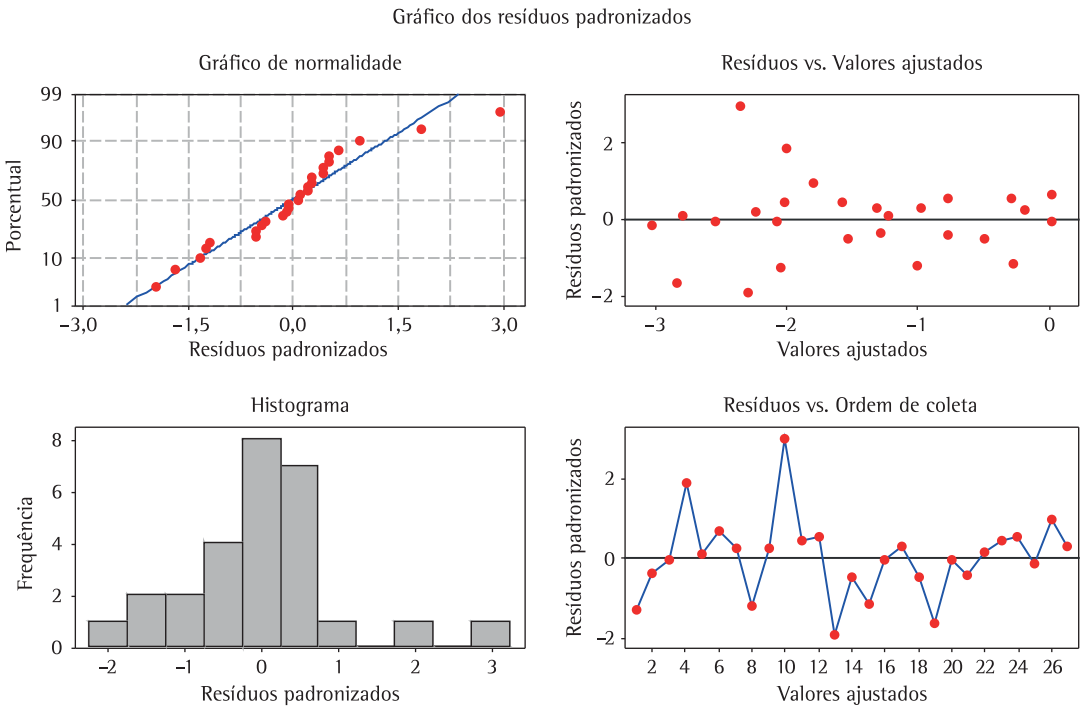
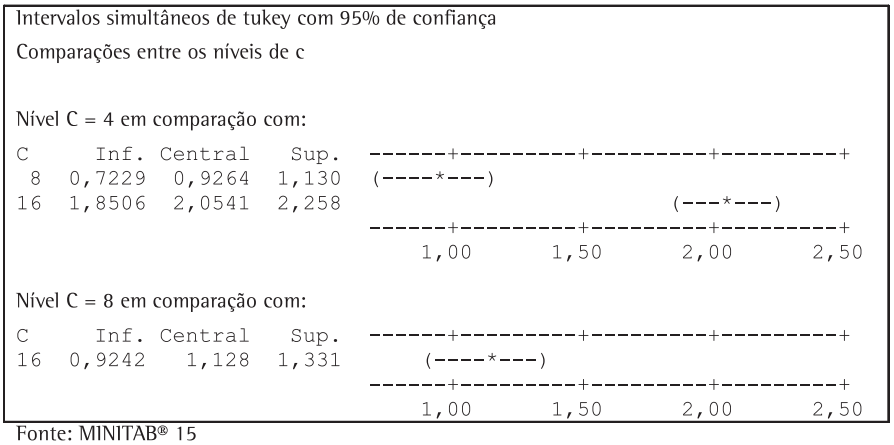


Figura 4. Análise de resíduos do modelo ajustado. Fonte: MINITAB (2006).

Tabela 2. Análise de variância para o tempo até convergência do algoritmo

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F	Valor-*p*
c	02	19,0475	9,5237	340,44	0,0002
cv^2	02	00,0279	0,0140	000,50	0,6160
$c \times cv^2$	04	00,2878	0,0719	002,57	0,0780
λ	02	02,9006	1,4503	051,84	0,0000
Erro	16	00,4476	0,0280		
Total	26	22,7114			

Fonte: MINITAB (2006).



Fonte: MINITAB® 15

Figura 5. Comparações múltiplas entre os níveis do número de servidores. Fonte: MINITAB (2006).

com duas filas em série e três servidores. As duas possibilidades para essa configuração são apresentadas na Figura 6. A questão que se coloca aqui é se uma configuração domina a outra. Isto é, deseja-se verificar se existe uma configuração mais eficiente, baseado apenas na ordem dos servidores.

O primeiro grupo de experimentos foi realizado considerando-se duas taxas de chegada diferentes, $\lambda \in \{1; 2\}$, dois tempos médios de serviço, $\mu = \{4; 8\}$, que foram iguais para todos os servidores (servidores homogêneos), e três valores para o quadrado do coeficiente de variação da taxa de serviço, $cv^2 \in \{0{,}5; 1{,}0; 2{,}0\}$. Os resultados podem ser vistos na Tabela 3, onde são apresentadas as alocações ótimas (x) as taxas de atendimento alcançadas ($\Theta(x)$) e os valores da função objetivo penalizada (Z_α). Com o objetivo de avaliar a exatidão das aproximações analíticas, são apresentados também os resultados de simulações, em que a coluna δ é a semiamplitude dos intervalos de 95% de confiança (i.e., os valores que precisam ser subtraídos dos valores médios obtidos via simulação ($\Theta(x)^s$), para obtenção dos limites inferiores dos intervalos de 95% de confiança, e depois a esses valores médios somados, para obtenção dos limites superiores). Essas simulações foram feitas no *software* Arena, com 20 replicações, para determinação do δ, adotando-se um período de estabilização (do inglês, *burn-in*) de 20 mil unidades de tempo e um tempo total de simulação de 100 mil unidades de tempo. Para simular os tempos de serviço gerais com $cv^2 \in \{0{,}5; 2{,}0\}$ foi utilizada a distribuição gama, com parâmetros α e β adequados.

No segundo grupo de experimentos, com configuração bastante semelhante à do primeiro grupo, foi considerado dessa vez que os servidores eram heterogêneos, com taxas de serviço $\mu = 4$ e $\mu = 8$, alternadamente em cada servidor, sempre com a taxa menor para a fila com o maior número ($c = 2$) de servidores. Os resultados podem ser vistos na Tabela 4.

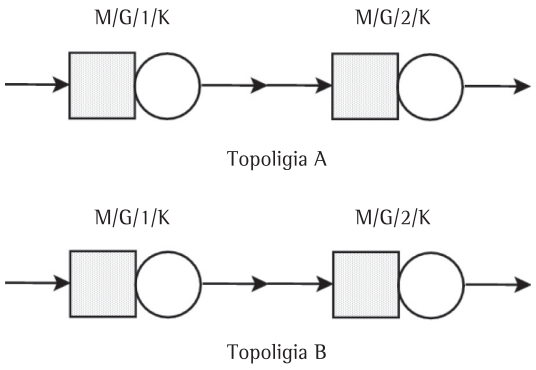


Figura 6. Redes em topologia série com duas filas e três servidores.

4. Discussão

Com relação à análise de desempenho do algoritmo, nota-se pela coluna de valores-*p* da Tabela 2 que o fator *A* (número total de servidores, *C*) pode ser considerado significativo, adotando-se um nível de significância 5% ($\alpha = 0{,}05$). O mesmo não aconteceu com o fator *B* (quadrado do coeficiente de variação do tempo de serviço, cv^2). Percebe-se também que não existiu interação entre os fatores *A* e *B* ao nível de 5%.

Ainda com relação à análise de desempenho do algoritmo, nas comparações múltiplas da Figura 5 nota-se que os intervalos de confiança construídos não possuem o valor zero. Isso indica que existem diferenças entre os valores médios para esses níveis do fator *A*. A rede com 16 servidores possui um tempo

Tabela 3. Resultados para a rede de duas filas em série com serviços homogêneos.

							Simulação		
λ	μ	cv^2	c	x	Θ(x)	Z_α	Θ(x)^s	δ	Z_α^s
1,0	(4,4)	0,5	(2, 1)	(3, 4)	0,999	8,16	0,997	0,001	9,710
			(1, 2)	(4, 3)	0,999	8,16	0,998	0,001	8,780
		1,0	(2, 1)	(3, 4)	0,998	8,90	0,997	0,001	9,670
			(1, 2)	(4, 3)	0,998	8,90	0,997	0,001	10,260
		2,0	(2, 1)	(4, 5)	0,999	10,3	0,999	0,001	10,380
			(1, 2)	(5, 5)	0,999	11,3	0,997	0,001	12,010
	(8,8)	0,5	(2, 1)	(2, 3)	1,000	5,42	0,993	0,001	12,180
			(1, 2)	(3, 2)	1,000	5,42	0,999	0,001	5,650
		1,0	(2, 1)	(2, 3)	0,999	5,59	0,993	0,001	12,350
			(1, 2)	(3, 2)	0,999	5,59	0,998	0,001	7,050
		2,0	(2, 1)	(2, 3)	0,999	6,11	0,993	0,001	12,280
			(1, 2)	(3, 2)	0,999	6,11	0,996	0,001	8,810
2,0	(4,4)	0,5	(2, 1)	(7, 7)	1,997	16,8	2,001	0,015	13,400
			(1, 2)	(7, 7)	1,997	16,8	1,997	0,001	16,800
		1,0	(2, 1)	(8, 8)	1,997	19,2	2,000	0,002	17,600
			(1, 2)	(8, 8)	1,997	19,2	1,999	0,002	18,900
		2,0	(2, 1)	(9, 11)	1,996	23,7	2,000	0,001	20,200
			(1, 2)	(11, 9)	1,996	23,7	1,997	0,001	23,500
	(8,8)	0,5	(2, 1)	(4, 4)	1,999	9,03	2,001	0,002	7,500
			(1, 2)	(4, 4)	1,999	9,03	1,998	0,001	10,400
		1,0	(2, 1)	(4, 5)	1,999	9,95	2,000	0,002	8,900
			(1, 2)	(5, 5)	1,999	11,0	2,000	0,002	8,800
		2,0	(2, 1)	(4, 5)	1,997	11,7	1,999	0,002	10,500
			(1, 2)	(5, 4)	1,997	11,7	1,994	0,001	14,900

Tabela 4. Resultados para a rede de duas filas em série com serviços heterogêneos.

							Simulação		
λ	μ	cv^2	c	x	Θ(x)	Z_α	Θ(x)^s	δ	Z_α^s
1,0	(4,8)	0,5	(2, 1)	(3,3)	0,999	6,95	0,997	0,001	8,670
	(8,4)		(1, 2)	(3, 3)	0,999	6,95	0,999	0,001	6,720
	(4,8)	1,0	(2, 1)	(3, 3)	0,999	7,39	0,997	0,001	8,870
	(8,4)		(1, 2)	(3, 3)	0,999	7,39	0,998	0,001	8,130
	(4,8)	2,0	(2, 1)	(4, 3)	0,999	8,15	0,999	0,001	8,320
	(8,4)		(1, 2)	(3, 4)	0,999	8,15	0,996	0,001	10,930
2,0	(4,8)	0,5	(2, 1)	(7, 4)	1,998	13,1	1,999	0,002	14,100
	(8,4)		(1, 2)	(4, 7)	1,998	13,1	1,999	0,002	14,400
	(4,8)	1,0	(2, 1)	(8, 5)	1,998	14,7	2,001	0,002	14,300
	(8,4)		(1, 2)	(5, 9)	1,998	15,7	2,000	0,001	15,200
	(4,8)	2,0	(2, 1)	(9, 5)	1,997	17,4	2,000	0,001	14,300
	(8,4)		(1, 2)	(5, 9)	1,997	17,4	1,995	0,002	19,200

médio de convergência significativamente maior que as redes com 8 e 4 servidores. Além disso, os tempos médios de convergência para os nível 4 e 8 também apresentam diferenças significativas entre si. O tempo médio para a rede com 8 servidores é maior do que o para a rede com 4.

No que diz respeito às alocações ótimas produzidas pelo algoritmo de otimização (ver Tabelas 3 e 4), de um modo geral os valores encontrados foram bastante encorajadores. As alocações foram bastante estáveis, ou seja, com pequenas mudanças nos parâmetros da rede têm-se mudanças também pequenas na alocação ótima. Um ponto que merece destaque é a influência que exerceu o quadrado do coeficiente de variação do tempo de serviço, cv^2, na área de espera alocada ótima, reforçando-se a importância de desenvolverem-se metodologias para tratar filas com tempos de serviço gerais.

Quanto à qualidade das soluções analíticas aproximadas, os resultados mostraram-se mais modestos. Dos 24 experimentos realizados com redes homogêneas (Tabela 3), 15 deles tiveram seus

valores analíticos confirmados pelos intervalos de confiança de 95% (com 6 valores analíticos fora dos intervalos de confiança). Em 12 experimentos realizados com redes heterogêneas (Tabela 4), a metade dos resultados analíticos aproximados foi confirmada por simulação. Também, em alguns casos, as diferenças entre os valores das soluções analíticas e simuladas, Z_α e Z_α^s, foram relativamente grandes (maiores que 50%). Isso é explicado em parte pelo valor alto utilizado para a variável dual α (α = 1000). Resultados obtidos para um nível de 90% de confiança (não apresentados) não possuem uma cobertura mais próxima da cobertura nominal esperada de 90%. Esses resultados dão uma ideia da dificuldade que é a determinação de medidas de desempenho para filas finitas configuradas em redes.

Comparando-se a alocação das áreas de espera para as topologias A e B, Figura 6, é difícil dizer que uma topologia supera a outra em termos de valor de função objetivo, apesar de existir uma pequena diferença nas soluções ótimas, Z_α. Assim, não se pode afirmar que existe o domínio de uma topologia sobre a outra. Fica difícil, portanto, estabelecer regras que prevejam qual topologia é a melhor em função apenas do posicionamento dos servidores.

O conjunto de experimentos com redes heterogêneas (i.e., diferentes taxas de serviço), Tabela 4, também leva a algumas conclusões importantes. Eles indicam que o desempenho pode ser independente do tipo de topologia se for utilizada uma combinação adequada entre o número de servidores e a taxa de serviço. De fato, as taxas de saída foram similares, tanto para os casos em que a fila mais lenta estava no início da rede (μ = 4), quanto para quando estava a mais rápida (μ = 8). Outro ponto que merece destaque é que, conforme esperado, áreas de espera maiores foram designadas para as filas com menor taxa de serviço. Em outras palavras, os servidores com menor capacidade de atendimento têm uma tendência a receber uma maior área de espera, para compensar.

5. Conclusões e observações finais

Neste artigo foi apresentado em detalhes um algoritmo recentemente proposto na literatura (SMITH; CRUZ; VAN WOENSEL, 2010) para alocação de áreas de espera em redes de filas M/G/c/K abertas e acíclicas. Por meio de um experimento planejado inédito, concluiu-se que o tempo de processamento do algoritmo depende do número de servidores dessas redes. Além disso, o quadrado do coeficiente de variação do tempo de serviço não interfere significativamente no tempo de execução do algoritmo, apesar de influenciar na alocação ótima, o que é um resultado surpreendente.

Experimentos em configurações que ainda não haviam sido testadas indicaram que também nesses casos a alocação obtida pelo algoritmo é robusta e faz sentido. Além disso, a aproximação para a medida de desempenho de interesse (a taxa de saída) também se confirmou satisfatória, pois em grande parte dos casos ficaram dentro dos intervalos de confiança de 95%, que foram estimados por simulação.

Outro resultado interessante obtido foi que topologias diferentes podem resultar em um desempenho similar se as áreas de espera são as ótimas. Dessa forma não pareceu ser fácil a obtenção de regras heurísticas para a alocação dos servidores dentro da topologia de interesse antes de aplicar-se um procedimento de otimização para dizer qual topologia é melhor. Sabe-se que a topologia é direcionada geralmente pela aplicação, mas tal resultado pode trazer alguma flexibilidade para aqueles casos em que topologias alternativas estejam competindo.

Sobre as possíveis direções que esta pesquisa pode tomar, pode-se citar a aplicação do algoritmo a problemas reais na área de manufatura e montagem, que podem apresentar redes de tamanho da ordem de centenas de nós (SPINELLIS; PAPADOULOS; SMITH, 2000). Não foi feita uma análise da ordem de complexidade do algoritmo de alocação, pois se queria apenas assegurar que os resultados fossem acurados. Entretanto, pelos experimentos realizados, observou-se que os tempos de processamento não cresceram dramaticamente com o aumento do número de nós da rede. Assim, é possível que problemas reais bem grandes sejam resolvidos pelo algoritmo. De fato, problemas de alocação de servidores em redes de filas finitas sem áreas de espera foram resolvidos por método similar para mais de uma centena de nós (ANDRIANSYAH et al., 2010).

Para problemas muito grandes, quando o tempo de processamento ficar proibitivo, podem-se empregar como último recurso técnicas de agregação. Essas são técnicas comumente utilizadas para reduzir o tamanho de redes em problemas reais, quando são retidos apenas os nós mais importantes da rede.

Outra possibilidade é incluir estudos sobre redes com laços de realimentação, muito encontrados em sistemas de manufatura, com fluxos reversos e retrabalho. Os laços de realimentação causam grande dependência entre as chegadas e precisam de cuidadosa consideração. Essas são apenas algumas possíveis ideias para futuros trabalhos nesta área.

Referências

ANDRIANSYAH, R. et al. Performance optimization of open zero-buffer multi-server queueing networks. *Computers & Operations Research*, v. 37, n. 8, p. 1472-1487, 2010. http://dx.doi.org/10.1016/j.cor.2009.11.004

ARGOUD, A. R. T. T.; GONÇALVES FILHO, E. V.; TIBERTI, A. J. Algoritmo genético de agrupamento para formação de módulos de arranjo físico. *Gestão & Produção*, v. 15, n. 2, p. 393-405, 2008. http://dx.doi.org/10.1590/S0104-530X2008000200014

BAZARAA, M. S.; SHERALI, H. D.; SHETTY, C. M. *Nonlinear Programming*: Theory and Algorithms. 3rd ed. New York: Wiley-Interscience, 2006. p. 872.

BITRAN, G. R.; MORABITO, R. An overview of tradeoff curve analysis in the design of manufacturing systems. *Gestão & Produção*, v. 3, n. 2, p. 108-134, 1996. http://dx.doi.org/10.1590/S0104-530X1996000200001

BITRAN, G. R.; MORABITO, R. Um exame dos modelos de redes de filas abertas aplicados a sistemas de manufatura discretos: Parte II. *Gestão & Produção*, v. 2, n. 3, p. 297-321, 1995. http://dx.doi.org/10.1590/S0104-530X1995000300005

CRUZ, F. R. B.; VAN WOENSEL, T.; SMITH, J. M. Buffer and throughput trade-offs in *M*/*G*/1/*K* queueing networks: A bi-criteria approach. *International Journal of Production Economics*, v. 125, n. 2, p. 224-234, 2010. http://dx.doi.org/10.1016/j.ijpe.2010.02.017

DOY, F. E. et al. Simulação do serviço de correio eletrônico através de um modelo de filas. *Pesquisa Operacional*, v. 26, n. 2, p. 241-253, 2006.

GROSS, D. et al. *Fundamentals of queueing theory*. 4th ed. New York: Wiley-Interscience, 2009. p. 600.

HSU, J. *Multiple comparisons*: Theory and methods. Boca Raton: Chapman and Hall/CRC, 1996. p. 296.

IANNONI, A. P.; MORABITO, R. Modelo de fila hipercubo com múltiplo despacho e backup parcial para análise de sistemas de atendimento médico emergenciais em rodovias. *Pesquisa Operacional*, v. 26, n. 3, p. 493-519, 2006. http://dx.doi.org/10.1590/S0101-74382006000300004

IANNONI, A. P.; MORABITO, R. Otimização da localização das bases de ambulâncias e do dimensionamento das suas regiões de cobertura em rodovias. *Produção*, v. 18, n. 1, p. 47-63, 2008. http://dx.doi.org/10.1590/S0103-65132008000100004

KENDALL, D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Annals of Mathematical Statistics*, v. 24, p. 338-354, 1953. http://dx.doi.org/10.1214/aoms/1177728975

KERBACHE, L.; SMITH, J. M. The generalized expansion method for open finite queueing networks. *European Journal of Operational Research*, v. 32, p. 448-461, 1987. http://dx.doi.org/10.1016/S0377-2217(87)80012-7

KIMURA, T. A transform-free approximation for the finite capacity *M*/*G*/*s* queue. *Operations Research*, v. 44, n. 6, p. 984-988, 1996. http://dx.doi.org/10.1287/opre.44.6.984

KLEINROCK, L. *Queueing Systems*. New York: John Wiley & Sons, 1975. v. I: Theory, p. 417.

LABETOULLE, J.; PUJOLLE, G. Isolation method in a network of queues. *IEEE Transactions on Software Engineering*, v. 6, n. 4, p. 373-380, 1980. http://dx.doi.org/10.1109/TSE.1980.234493

LEMARÉCHAL, C. The omnipresence of Lagrange. *Annals of Operations Research*, v. 153, n. 1, p. 9-27, 2007. http://dx.doi.org/10.1007/s10479-007-0169-1

MIGUEL, P. A. C. (Org.). *Metodologia de pesquisa em engenharia de produção e gestão de operações*. Rio de Janeiro: Elsevier, 2010. p. 226.

MINITAB INC. *Minitab Statistical Software*, Release 15 for Windows. Pennsylvania: State College, 2006. Minitab® is a registered trademark of Minitab Inc.

MONTGOMERY, D. C. *Design and Analysis of Experiments*. 7. ed. New York: John Wiley & Sons, 2008. p. 680.

MORABITO, R.; LIMA, F. C. R. Um modelo para analisar o problema de filas em caixas de supermercados: um estudo de caso. *Pesquisa Operacional*, v. 20, n. 1, p. 59-71, 2000.

SELLITTO, M. A.; BORCHARDT, M.; PEREIRA, G. M. Medição de tempo de atravessamento e inventário em processo em manufatura controlada por ordens de fabricação. *Produção*, v. 18, n. 3, p. 493-507, 2008.

SILVA, C. R. N.; MORABITO, R. Análise de problemas de partição de instalações em sistemas job-shops por meio de modelos de redes de filas. *Pesquisa Operacional*, v. 27, n. 2, p. 333-356, 2007a. http://dx.doi.org/10.1590/S0101-74382007000200008

SILVA, C. R. N.; MORABITO, R. Aplicação de modelos de redes de filas abertas no planejamento do sistema job-shop de uma planta metal-mecânica. *Gestão & Produção*, v. 14, n. 2, p. 393-410, 2007b.

SMITH, J. M. *M*/*G*/*c*/*K* blocking probability models and system performance. *Performance Evaluation*, v. 52, n. 4, p. 237-267, 2003. http://dx.doi.org/10.1016/S0166-5316(02)00190-6

SMITH, J. M.; CRUZ, F. R. B. The buffer allocation problem for general finite buffer queueing networks. *IIE Transactions on Design & Manufacturing*, v. 37, n. 4, p. 343-365, 2005.

SMITH, J. M.; CRUZ, F. R. B.; VAN WOENSEL, T. Topological network design of general, finite, multi-server queueing networks. *European Journal of Operational Research*, v. 201, n. 2, p. 427-441, 2010. http://dx.doi.org/10.1016/j.ejor.2009.03.012

SPINELLIS, D.; PAPADOULOS, C. T.; SMITH, J. M. Large production line optimization using simulated annealing. *International Journal of Production Research*, v. 38, n. 3, p. 509-541, 2000. http://dx.doi.org/10.1080/002075400189284

TAKEDA, R. A.; WIDMER, J. A.; MORABITO, R. Aplicação do modelo hipercubo de filas para avaliar a descentralização de ambulâncias em um sistema urbano de atendimento médico de urgência. *Pesquisa Operacional*, v. 24, n. 1, p. 39-71, 2004.

YANASSE, H. H.; BECCENERI, J. C.; SOMA, N. Y. Um algoritmo exato com ordenamento parcial para solução de um problema de programação da produção: experimentos computacionais. *Gestão & Produção*, v. 14, n. 2, p. 353-361, 2007.

Agradecimentos

Esta pesquisa foi parcialmente financiada pelo CNPq (projetos 201046/1994-6, 301809/1996-8, 307702/2004-9, 472066/2004-8, 304944/2007-6, 561259/2008-9, 553019/2009-0, 550207/2010-4, 501532/2010-2, 303388/2010-2), pela Capes (projeto BEX-0522/07-4) e pela Fapemig (projetos CEX-289/98, CEX-855/98, TEC-875/07, CEX-PPM-00401/08, CEX-PPM-00390-10). Os autores gostariam de deixar expressos aqui agradecimentos aos professores Anderson Duarte, Luiz Duczmal e Roberto Quinino e a dois revisores anônimos pelas valiosas críticas e sugestões.

A DOE-based proposal for performance analysis of M/G/c/K queueing network optimization algoritms

Abstract

This paper presents the results of an empirical analysis of a previously proposed algorithm for buffer allocation in finite open acyclic general-service multi-server queuing networks. Based on the computational results, it is concluded that the processing time of the algorithm depends on the number of network servers (as expected) but is independent of the squared coefficient of variation of service time. It is also concluded that the obtained allocations are robust and that the approximations for the performance measures are accurate, as verified by simulation. Finally, it is found that it is not straightforward to develop heuristic rules to allocate multiple servers in the topology without applying a buffer allocation algorithm to determine the optimal configuration.

Keywords

Optimization. Performance evaluation. Stochastic process. Design of experiments.

Anexo A

Na Figura 7 são verificadas as suposições iniciais do modelo proposto, Equação 23. Nota-se que os dados transformados seguem a distribuição normal e não há violação de variabilidade constante entre os fatores e o bloco.

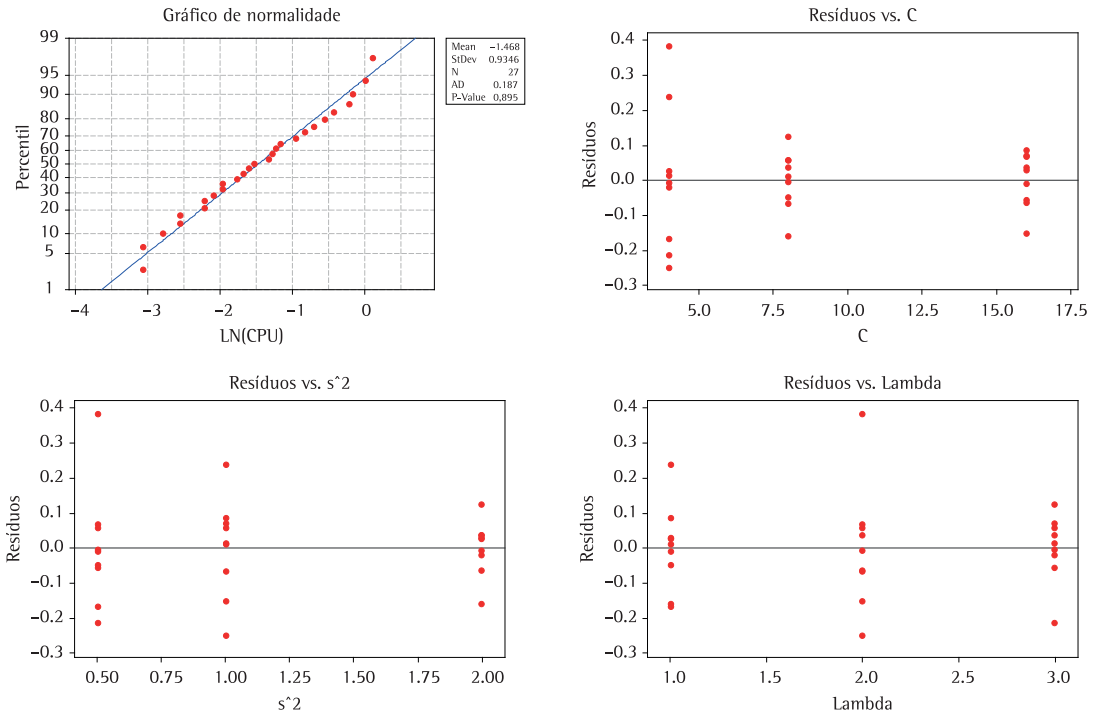


Figura 7. Suposições iniciais do modelo proposto para análise de desempenho. Fonte: MINITAB (2006).