SIMULAÇÃO DO DESEMPENHO DE REDES DE FILAS FINITAS GERAIS SEM ÁREA DE ESPERA EM CONFIGURAÇOES PARETO-EFICIENTES

Karina A. Rodrigues, Priscila B. Reis, Frederico R. B. Cruz

UFMG

karinazevedor@gmail.com, pribreis@gmail.com, fcruz@est.ufmg.br

RESUMO

Nesse artigo descrevemos resultados de uma experiência com o Arena®, versão estudante, para simulação do desempenho de redes de filas finitas gerais sem área de espera. Este é um problema relevante para cuja solução esforços justificam-se, pela segurança que traz a resultados recentemente obtidos para otimização de redes de filas finitas gerais. A medida de desempenho aqui analisada é a taxa de atendimento na unidade de tempo, mas outras medidas de desempenho igualmente importantes poderiam ser obtidas de maneira similar. Os resultados obtidos demonstram consistência, ora subestimando resultados analíticos aproximados consagrados, ora superestimando-os.

PALAVRAS-CHAVE: Simulação, redes de filas, filas finitas, avaliação de desempenho.

1. INTRODUÇÃO

Uma rede de filas finitas configuradas em série é apresentada na Fig. 1. A ocorrência de um ambiente sem área de espera pode dar-se tanto devido a uma limitação própria da tecnologia do produto em questão quanto por conta da simples ausência de área de espera entre duas operações consecutivas de um processo. Para maiores detalhes, ver Fransoo & Rutten [2].

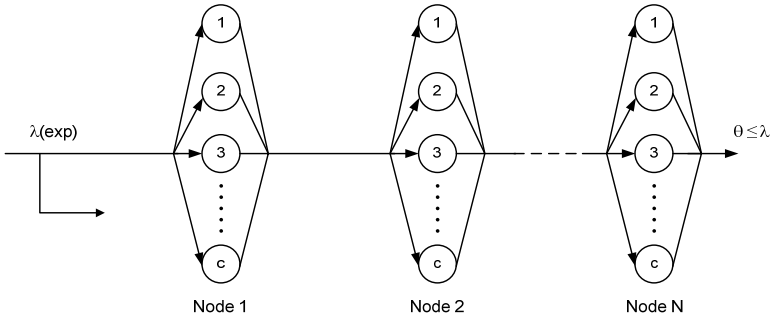


Figura 1: Uma rede de filas finitas configurada na topologia série

Poderíamos exemplificar muitos sistemas reais como modelos aproximados de filas finitas sem área de espera, configuradas em redes. Um processo de produção de aço é descrito por

Hall & Sriskandarajah [3]. O aço flui por diversas operações que vão do lingotamento, retirada dos moldes, reaquecimento, têmpera e laminação preliminar. Neste processo de produção, o aço precisa passar de uma operação para a seguinte continuamente, sem nenhuma espera ou armazenagem de trabalho em processo, uma vez que tal espera poderia resultar em um inconveniente resfriamento para uma temperatura abaixo da qual não fosse aceitável para a próxima fase do processo. Portanto, a etapa precisa ser concluída e o trabalho passado imediatamente à fase seguinte, ou caso contrário deverá ser segurado na fase atual, mesmo após conclusão, até que seja possível seu recebimento na etapa seguinte. De maneira similar, em ambientes de processamento de alimentos, nenhuma área de espera deve ser permitida entre a operação de cozimento e a embalagem, o que é devido à restrição de que o produto ainda esteja fresco quando enlatado. Problemas semelhantes podem ser encontrados na produção de sucos e cerveja. Nesses exemplos citados, restrições na tecnologia de processamento e suas características criam um sistema de produção sem área de espera. A natureza de produtos, e.g., como condimentos tais como maionese e vários outros tipos de molhos para saladas, por vezes dita as normas de higiene como o fator crítico na sua produção. Conforme estudado por Ramudhin & Ratliff [12], algumas vezes simplesmente não há espaço para trabalho em processo e o produto não pode nunca esperar entre duas operações. Como último exemplo, sistemas de comunicação móvel de terceira geração podem ser caracterizados como modelos de filas finitas multisevidor sem área de espera [16]. Nestes sistemas, as chegadas representam as requisições por áudio, dados e mensagens de vídeo, nos quais o tempo de serviço (geral) é o tempo de transmissão e para os quais a ausência de capacidade de armazenamento conduz a um sistema sem área de espera. Entretanto, apesar da grande relevância de redes de filas sem áreas de espera tanto na indústria (principalmente nas indústrias de processo e semiprocesso), como nas áreas acima mencionadas, somente pode ser encontrada uma literatura reduzida que foque em redes de filas finitas gerais sem áreas de espera.

O objetivo aqui é determinar medidas de desempenho para redes de filas finitas sem área de espera baseado em simulação a eventos discretos. Mais especificamente, o método é implementado no Arena® [5], versão estudante, um programa de acesso fácil e custo reduzido. Em geral, a simulação é reconhecida como uma boa ferramenta para validação de métodos analíticos aproximados (veja mais detalhes em Kelton et al. [5]). Como explicaremos adiante em mais detalhes, a análise exata do desempenho de filas finitas configuradas em redes apresenta uma séria de dificuldades matemáticas ainda intransponíveis e técnicas

aproximadas cuja validação depende, p.e., da simulação, parecem ser as únicas soluções possíveis, ainda que para redes de interesse prático de tamanho reduzido.

O artigo está organizado da seguinte forma. Inicialmente apresentaremos na Seção 2 o algoritmo para otimização da taxa de saída (que considera como variáveis de decisão o número de servidores em cada um dos nós da rede), bem como apresentaremos a técnica utilizada para estimação aproximada do desempenho de redes de filas finitas sem área de espera. Na Seção 3 apresentaremos resultados de simulação obtidos para as configurações identificadas na literatura como subótimas, pelos algoritmos de otimização e de análise aproximada de desempenho. Conclusões e observações finais apresentadas na Seção 4 encerram o artigo.

2. ALGORITMOS PARA OTIMIZAÇÃO

O modelo de otimização cujas soluções aferiremos aqui é o problema multiobjetivo de minimização do número de servidores (que é frequentemente um recurso muito caro), em uma rede de filas finitas sem áreas de espera, e de maximização da taxa de saída (que usualmente guarda relação direta com o lucro produzido pela rede). A única restrição é que o número de servidores seja um inteiro positivo. Este problema, de formulação muito simples, é abordado por van Wonsel & Cruz [17] por meio de um algoritmo genético multiobjetivo (AGMO) que se encontra associado, como ferramenta de análise aproximada de desempenho, ao método de expansão generalizado (MEG), que foi desenvolvido por Kerbache & Smith [6, 7, 8] e tem sido usado com sucesso em diversos tipos similares de redes de filas finitas [4, 14, 15].

2.1 Algoritmo Genético Multiobjetivo (AGMO)

A versão do AGMO utilizada por van Wonsel & Cruz [17] é apresentada na Fig. 2, a seguir. Há grandes vantagens da utilização do AGMO, que incluem a facilidade com que uma aproximação do conjunto de Pareto é gerada em uma única rodada (obviamente durante várias iterações), sem mencionar a simplicidade para tratar funções-objetivo que não possuem forma fechada (como será o caso aqui, conforme esclareceremos em breve), mas outras abordagens poderiam ser utilizadas. Não forneceremos maiores detalhes, mas um AGMO é capaz de encontrar o conjunto de soluções não-dominadas, conhecido como conjunto de soluções eficientes. A curva de Pareto aproximada, resultante de conjunto de soluções eficientes, permite ao pesquisador escolher dentre diversas soluções eficientes aquele que melhor lhe atende e também avaliar o que pode ganhar em um objetivo se aceitar perder um pouco no outro (o ***trade-off*** entre os objetivos).

```
algorithm
   read graph, arrival, service rates, G(V, A), λ_v, μ_v, ∀ v ∈ V
   /* generate initial population */
   P_1 ← GetInitPopulation(popSize)
   for i = 1 until numGen do
      /* generate offspring by crossover and mutation */
      Q_i ← MakeNewPopulation(popSize, P_i)
      /* combine parent and offspring */
      R_t ← P_t ∪ Q_t
      /* select new population */
      P_{t+1} ← SelectNewPopulation(popSize, R_t)
   end for
   write P_{numGen+1}
end algorithm
```

Figura 2: Algoritmo genético multiobjectivo (AGMO)

2.2 Método de Expansão Generalizado (MEG)

O MEG surgiu da necessidade de uma ferramenta para avaliação aproximada de desempenho de uma ***rede*** de filas finitas, conhecidos sua topologia (nós em série, em junção, fusão etc.), a probabilidades associadas às rotas alternativas (se houver), taxas de chegadas externas (λ), taxas de serviço (μ) e número de servidores (c). Quando temos uma ***fila única***, o problema se reduz à aplicação da expressão a seguir, para o caso de a medida de desempenho de interesse ser a taxa de saída θ:

$$\theta = \lambda(1 - p_c) = \lambda\left(1 - \frac{(\lambda/\tilde{\mu})^c / c!}{\sum_{i=0}^{c} (\lambda/\tilde{\mu})^c / c!}\right),$$

em que a probabilidade de bloqueio p_c vem diretamente da conhecida expressão de perda de Erlang e $\tilde{\mu} = \mu$, no caso de fila única.

Quando temos filas finitas configuradas em redes, um nó pode ser afetado por eventos em nós remotos, que podem lhe causar bloqueios e falta de serviço (***starvation*** [11]). Assim, precisaremos lançar mão de métodos aproximados para determinar $\tilde{\mu}$, como o MEG, que é uma técnica robusta desenvolvida por Kerbache & Smith [8], para esta finalidade. O método é uma combinação de tentativas repetidas e decomposição nó-a-nó e utiliza bloqueio após serviço, que felizmente é o que prevalece em muitos dos sistemas de manufatura, transporte e similares (veja detalhes em Buzacott & Shanthikumar [1]). O MEG é composto por três estágios: reconfiguração da rede, estimação dos parâmetros e eliminação do laço de realimentação. A seguir, detalhamos cada um desses estágios.

Reconfiguração da Rede

O primeiro estágio do MEG inclui a reconfiguração da rede pela inclusão de um nós artificial para cada nó que seja sucedido por uma fila finita, para registrar as entidades que ficarem bloqueadas, conforme visto na Fig. 3.

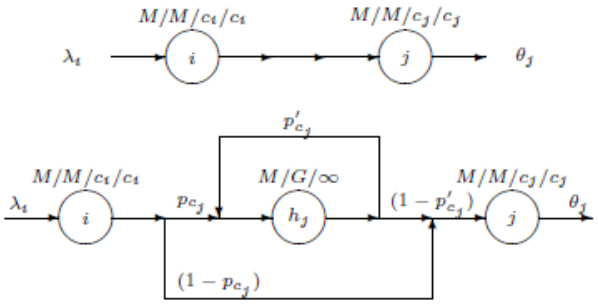


Figura 3: Método da expansão generalizado

Estimação dos Parâmetros

No segundo estágio, estimamos inicialmente a probabilidade de bloqueio, que, conforme já anteriormente mencionado aqui, vem da fórmula de perda de Erlang, válida para qualquer fila sem área de espera, i.e., com *c* servidores em paralelo e espaço total *c*, incluindo os itens em serviço. O segundo parâmetro a ser estimado é a probabilidade de que uma entidade seja forçada a uma segunda visita ao nó de espera h_j, dado que foi rejeitada na tentativa anterior, p_c', a qual pode ser aproximada pelo método de Labetoulle & Pujolle [10]. Finalmente, o terceiro parâmetro é a taxa de serviço do nó de espera, que pode ser obtido via teoria de renovação [9], como:

$$\mu_h = \frac{2\mu_j}{1+\sigma_j^2\mu_j^2}.$$

Eliminação do Laço de Realimentação

As visitas repetidas ao nó de espera criam dependências no processo de chegada. Para sua retirada, as entidades têm seu tempo de serviço convenientemente dilatado, que se torna

$$\mu_h' = (1 - p_c')\mu_h .$$

Em resumo, o objetivo do MEG é prover um método para atualizar a taxa de serviço de um nó levando em consideração o possível bloqueio após serviço que pode ser causado por sobrecarga da fila finita à sua frente, ou seja

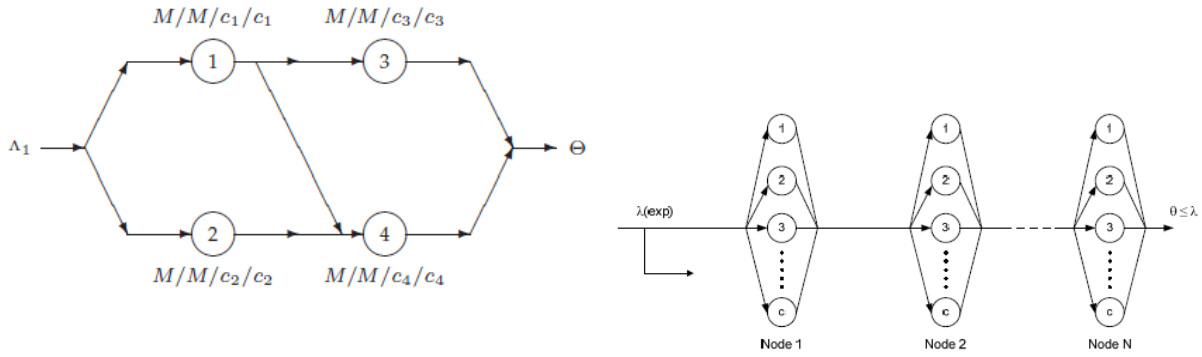
$$\widetilde{\mu}_i^{-1} = \mu_i^{-1} + p_{c_j}(\mu_{h_j}')^{-1},$$

através do qual podem ser estimadas as probabilidades de bloqueio p_c, via fórmula de perda de Erlang e, consequentemente, as taxas de saída:

$$\theta = \lambda(1 - p_c).$$

3. RESULTADOS EXPERIMENTAIS

O algoritmo de otimização foi aplicado a diversas redes configuradas em várias topologias. Para nossa conveniência, escolhemos mostrar resultados para redes mistas e em série, Fig. 4. Foram testados vários tamanhos de redes, número de servidores e taxas de chegada, $N \in \{3, 5\}$, $c \in \{2, 4, 10\}$, e $\lambda \in \{2, 4, 8, 16\}$, respectivamente.



a) rede mista b) rede série

Figura 4: Topologias examinadas

Para avaliar a qualidade das aproximações fornecidas pelo MEG, conduzimos experimentos no Arena® [5]. A modelagem de redes de filas finitas sem área de espera envolve apenas o uso de seus objetos comuns. A tela para um modelo de três nós ($N = 3$) em série pode ser vista na Fig. 5.

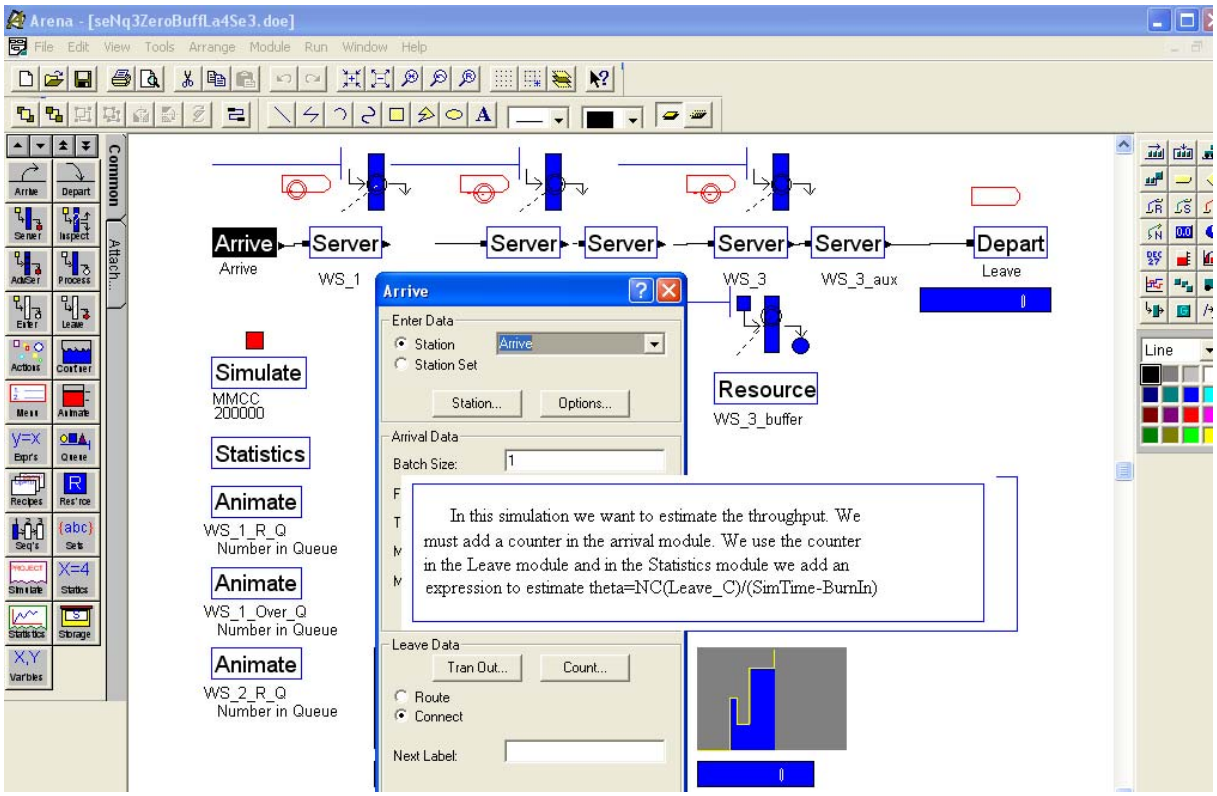


Figura 5: Modelo desenvolvido no Arena®

Nas simulações, utilizamos um período de observação de 200000 unidades de tempo e um período de inicialização de 2000 unidades de tempo (mais sobre a escolha períodos de inicialização em Robinson [13]). Os resultados obtidos para a topologia mista podem ser visualizados na Tabela 1, onde temos os valores do método analítico, Θ, e simulado, Θs, bem como o semi-intervalo de 95% de confiança, δ.

Tabela 1: Resultados para redes em topologia mistas

			Analítico	Simulação			Erro
λ	c	μ	Θ	Θs	δ	CPU(mim)	Δ%
2	(2,2,2,2)	10	1,9995	1,991	0,002	1,6	0,42
	(4,4,4,4)	10	2,0000	2,001	0,002	1,6	-0,03
	(10,10,10,10)	10	2,0000	2,000	0,002	1,6	0,03
4	(2,2,2,2)	10	3,9934	3,933	0,002	3,1	1,53
	(4,4,4,4)	10	3,9999	4,000	0,002	3,1	0,00
	(10,10,10,10)	10	4,0000	4,000	0,002	3,1	0,01
8	(2,2,2,2)	10	7,9119	7,541	0,003	6,0	4,92
	(4,4,4,4)	10	7,9994	7,992	0,003	6,2	0,10
	(10,10,10,10)	10	8,0000	7,999	0,003	6,3	0,01
16	(2,2,2,2)	10	14,966	13,237	0,003	11,0	13,1
	(4,4,4,4)	10	15,975	15,871	0,003	12,4	0,66
	(10,10,10,10)	10	16,000	15,998	0,004	12,6	0,01

Notamos na Tabela 1 erros tipicamente inferiores a 5%, exceto quando a rede encontra-se bastante congestionada, quando os erros podem ser superiores a 10%. Essa é uma importante avaliação, que ainda não havia sido feita. Também comparamos pontos do conjunto de Pareto aproximada com resultados de simulação, para redes em topologia série com $N \in \{3, 5\}$, conforme visto na Tabela 2.

Tabela 2: Resultados para redes em topologia série

				Analítico		Simulação		Erro
N	λ	μ	c	Θ	□	Θ(s)	δ	Δ%
3	4	10	(1,1,1)	3,258		2,715	0,001	20,0
			(2,1,1)	3,654		3,544	0,001	3,12
			(2,2,1)	3,869		3,745	0,002	3,31
			(2,2,2)	3,949		3,776	0,002	4,59
			(3,2,2)	3,991		3,966	0,002	0,63
			(3,3,2)	3,995		3,970	0,002	0,64
			(4,2,3)	3,999		3,997	0,002	0,06
			(5,2,3)	4,000		3,999	0,002	0,02
			(5,3,3)	4,000		3,998	0,002	0,04
			(5,3,4)	4,000		3,998	0,002	0,05
			(6,3,4)	4,000		3,999	0,002	0,02
			(6,4,4)	4,000		3,999	0,002	0,03
			(7,4,4)	4,000		3,999	0,002	0,04
5	16	10	(1,1,1,1,1)	4,984		4,893	0,001	1,86
			(1,1,2,1,1)	5,806		5,501	0,001	5,54
			(1,2,1,2,1)	6,851		6,092	0,001	12,46
			(2,1,2,1,2)	7,660		7,471	0,002	2,54
			(2,2,2,1,2)	8,759		9,051	0,002	-3,23
			(2,2,2,2,2)	11,52		10,55	0,002	9,20
			(3,2,2,2,2)	12,53		12,05	0,002	3,92
			(3,2,3,2,2)	13,35		12,83	0,002	4,09
			(3,2,3,2,3)	13,93		12,99	0,002	7,28
			(4,2,3,2,3)	14,46		13,94	0,002	3,79
			(4,2,3,3,3)	14,91		14,16	0,002	5,34
			(4,3,3,3,3)	15,43		15,12	0,002	2,05
			(4,3,4,3,3)	15,60		15,24	0,003	2,37
			(5,3,4,3,3)	15,76		15,69	0,003	0,45
			(5,3,4,3,4)	15,85		15,70	0,003	0,99
			(6,3,4,3,4)	15,90		15,86	0,003	0,24
			(6,3,5,3,4)	15,94		15,89	0,003	0,34

É importante salientar que quando aplicamos o AGMO para encontrar os valores **c** do conjunto de Pareto mostrados na Tabela 2, alguns casos se situam na condição de tráfego congestionado (por exemplo, para um total de 3 servidores, a taxa de saída é tão baixa quanto 3,256 enquanto a taxa de chega é 4,0).O MEG notadamente deteriora seu desempenho sob tráfego pesado. De fato, comparando os resultados analíticos aproximados com os resultados de simulação, notamos uma discordância maior entre esses dois valores, a medida que o

tráfego vai ficando mais pesado. Em conclusão, a metodologia pode não encontrar as melhores soluções sob tráfego pesado, pois a taxa de saída real não é tão bem estimada pelo MEG nesses casos. Quando a taxa de saída alvo é consideravelmente menor que a taxa de chegada (i.e., quando a rede de filas estiver operando sob tráfego pesado), um número menor de servidores que o realmente necessário será indicado pelo algoritmo de otimização para alcançar a taxa de saída especificada.

4. CONCLUSÕES E OBSERVAÇÕEOS FINAIS

O método da expansão generalizado (MEG) foi utilizado aqui como ferramenta de avaliação aproximada de desempenho de redes de filas finitas gerais sem áreas de espera. Mostramos que o MEG tipicamente produz resultados dentro de 5% de erro, mas esse erro pode ser significantemente mais alto, chegando a cerca de 20% em configurações sujeitas a tráfego pesado. Esse resultado tem um impacto significativo quando se pensa em utilizar o MEG como ferramenta auxiliar em algoritmos de otimização de redes de filas, principalmente em algoritmos mutiobjetivos. Esses algoritmos geram uma aproximação para o conjunto de Pareto que pode ser tanto pior quanto mais congestionamento na rede uma determinada solução eficiente representar. Tópicos para futura investigação incluem uma análise mais detalhada de tais situações.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] J. Buzacott, J. & J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice-Hall; 1993.

[2] J. C. Fransoo & W. G. M. M. Rutten, "A typology of production control situations in process industries", *International Journal of Operations and Production Management*. Vol. 14, n. 12, p. 47-57, 1994.

[3] N. G. Hall & C. Sriskandarajah, "A survey of machine scheduling problems with blocking and no-wait in process, *Operations Research*, Vol. 44, p. 510-525, 1996.

[4] S. Jain & J. M. Smith, "Open finite queueing networks with M/M/C/K parallel servers", *Computers & Operations Research*, Vol. 21, n. 3, p. 297-317, 1994.

[5] D. Kelton, R. P. Sadowski & D. A. Sadowski, *Simulation with Arena*, McGraw Hill College Div., New York; 2001.

[6] L. Kerbache. & J. M. Smith, "Assymptotic behavior of the expansion method for open finite queueing networks", *Computers & Operations Research*, Vol. 15 , n. 2, p. 157-169, 1988.

[7] L. Kerbache & J. M. Smith, "Multi-objective routing within large scale facilities using open finite queueing networks", *European Journal of Operational Research*, Vol.121, p. 105-123, 2000.

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

[9] L. Kleinrock, *Queueing Systems*, Vol. I: Theory, John Wiley & Sons, New York; 1975.

[10] J. Labetoulle & G. Pujolle, "Isolation method in a network of queues", *IEEE Transactions on Software Engineering*, Vol. SE-6, n. 4, p. 373-381, 1980.

[11] H. G. Perros, *Queueing Networks with Blocking*, Oxford University Press; 1994.

[12] A. Ramudhin & H. D. Ratliff, "Generating daily production schedules in process industries", *IIE Transactions*, Vol. 27, p. 646-656, 1995.

[13] S. Robinson, "A statistical process control approach to selecting a warm-up period for a discrete-event simulation", *European Journal of Operational Research*, Vol. 176, n. 1, p. 332-346, 2007.

[14] J. M. Smith & F. R. B. Cruz, "The buffer allocation problem for general finite buffer queueing networks", *IIE Transactions*, Vol. 37, n. 4, p. 343-365, 2005.

[15] D. Spinellis, C. Papodopoulos, J. M. Smith, "Large production line optimisation using simulated annealing", *International Journal of Production Research*, Vol 38, n. 3, p. 509-541, 2000.

[16] B. Tsybakov, "Optimum discarding in a bufferless system", *Queueing Systems*, Vol. 41, p. 165-197, 2002.

[17] T. van Wonsel & F. R. B. Cruz, "Multi-objective optimization of networks of open zero-buffer multi-server queues", XLI SBPO, Porto Seguro – BA, 2009.