

Estatística, Computação e Engenharia se Encontram: Avanços em Otimização em Redes Estocásticas

Frederico R. B. Cruz

Departamento de Estatística
Universidade Federal de Minas Gerais
E-mail: fcruz@est.ufmg.br



20 de maio de 2014



Organização

- 1 Introdução
- 2 Análise de Desempenho das Redes de Filas Finitas
- 3 Alguns Problemas Abordados
- 4 Trabalhos em Progresso
- 5 Observações Finais



Introdução

Foco





Elementos de Filas

Teoria de Filas

- Tem lugar de destaque entre as modernas técnicas analíticas da Pesquisa Operacional;
- Não está diretamente relacionada com o propósito da PO: tomada de decisão ótima;
- Entretanto, produz informação sobre o comportamento do sistema;

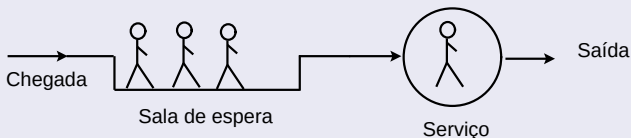
Quando as filas ocorrem?

O fenômeno das filas ocorre quando a demanda por determinado serviço é maior do que a sua capacidade de atendimento, seja pelo alto custo ou até mesmo pela impossibilidade física.



Elementos de Filas

Filas: elementos



Notação de Kendall: $A/S/m/K/N/Q$

- A - Distribuição do tempo entre chegadas (M - markoviana);
- S - Distribuição do tempo de serviço (G - geral);
- m - Número total de servidores;
- K - Capacidade do sistema (*incluindo* os itens em serviço);
- N - Tamanho da população (omitida, quando é ∞);
- Q - Disciplina de atendimento utilizada (omitida quando é *FIFO*, a mais usual).



Elementos de Filas

Filas $M/M/1$ (mais simples):



Distribuição de probabilidade:

$$p(n) \equiv P[N = n] = (1 - \rho)\rho^n, \text{ em que } \rho = \lambda/\mu < 1.$$

Número médio de usuários no sistema:

$$L \equiv E[N] = \sum_{n=0}^{\infty} np(n) = \frac{\lambda}{\mu - \lambda}.$$



Elementos de Filas

Número médio de usuários na fila:

$$L_q = \sum_{n=1}^{\infty} (n-1)p(n) = \frac{\lambda^2}{\mu(\mu - \lambda)}.$$

Tempo médio no sistema (Lei de Little):

$$W = L/\lambda = \frac{1}{\mu - \lambda}.$$

Tempo médio na fila:

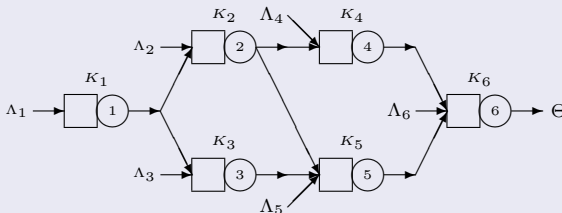
$$W_q = L_q/\lambda = \frac{\rho}{\mu - \lambda}.$$



Motivação

- Redes de filas ocorrem em muitos sistemas reais:

Rede de filas em topologia arbitrária



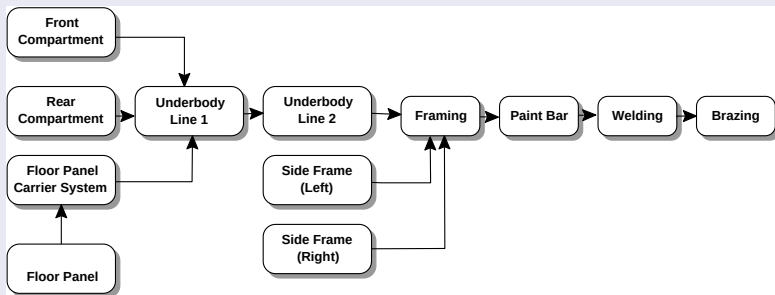
- Não limitados a sistemas de manufatura (Tan & Gershwin, 2009);
- ou *call centers* (Jouini et al., 2009);
- mas também a outras áreas interessantes e inovadoras, incluindo sistemas de saúde (Osorio & Bierlaire, 2009).



Motivação

- O contínuo crescimento dessas áreas, em importância e complexidade, requerer modelos mais sofisticados:

Redes de filas de uma montadora (Spieckermann et al., 2000).

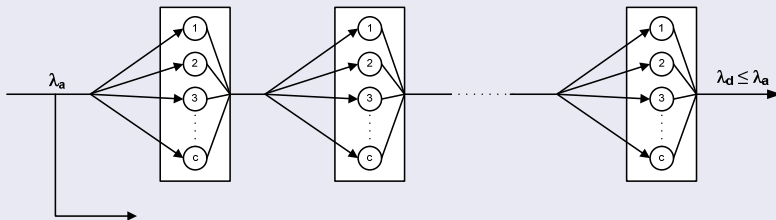




Motivação

- Redes de filas *sem áreas de espera* são também comuns em muitos sistemas reais:

Uma rede de filas sem área de espera configurada na topologia série





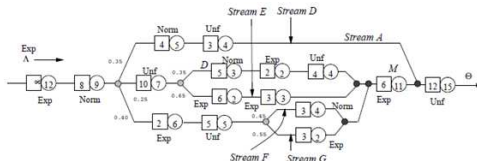
Motivação

- Ambientes sem área de espera ocorrem devido:
 - a uma limitação própria da tecnologia do produto;
 - à ausência pura e simples de espaço entre duas operações consecutivas de um processo.
- Alguns exemplos:
 - Em um processo de produção de aço, o produto passar por uma série de operações:
 - lingotamento, retirada dos moldes, reaquecimento, têmpera e laminação preliminar, sem nenhuma espera entre operações (sem *buffering*) ou sem trabalho em processo (*work-in-process*);
 - Na indústria alimentícia, áreas de espera não são permitidas, p.e., entre as operações de cozimento e o enlatamento:
 - é requisito que o produto esteja ainda fresco quando colocado em latas.



Objetivos

- Determinar configurações que otimizem alguma medida de desempenho:
 - minimizar o custo total da planta;
 - maximizar a taxa de saída (*throughput*);
 - minimizar o trabalho em processo (WIP, *work-in-process*);
 - minimizar o tempo de processo (*cycle time*);
 - etc.;
- Por meio de ajuste das variáveis de decisão:
 - número de servidores c ;
 - tamanho das áreas de espera K ;
 - taxa serviço μ ;
 - topologia.

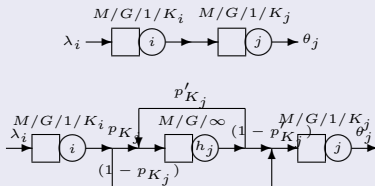




Análise de Desempenho das Redes de Filas Finitas

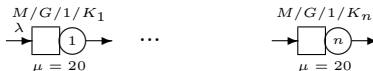
Método da Expansão Generalizado (MEG)

- A literatura mostra que o MEG produz estimativas *aceitáveis* conhecidos $\Theta(\mathbf{K}, \boldsymbol{\mu})$ e as taxas de chegada λ ;
- É uma combinação de decomposição nó-a-nó e tentativas repetidas;
- Três estágios, i) *reconfiguração da rede*, ii) *estimação de parâmetros* e iii) *eliminação da retroalimentação*.





Análise de Desempenho (cont.)



Comparações com Simulações

N	$\mathbf{K}$	λ	Análítico	Simulação			$\Delta\% \theta$
			θ	θ^s	δ	CPU(min)	
3	(2,2,2)	2	2.000	1.991	0.002	1.9	0.45
		4	3.993	3.930	0.001	3.8	1.61
		8	7.900	7.473	0.002	8.2	5.71
		16	14.52	12.54	0.003	14	15.8
5	(4,4,4,4,4)	2	2.000	1.999	0.001	3.0	0.04
		4	4.000	3.999	0.002	5.8	0.03
		8	7.999	7.993	0.003	12	0.08
		16	15.98	15.88	0.003	24	0.63
9	(10,...,10)	2	2.000	1.999	0.001	5.3	0.04
		4	4.000	4.000	0.003	10	0.05
		8	8.000	7.998	0.003	20	0.03
		16	16.00	16.00	0.005	45	0.00



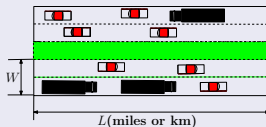
Alguns Problemas Abordados

- Análise de filas simples;
- Análise de redes de filas;
- Otimização de rotas de evacuação;
- Alocação de áreas de espera.

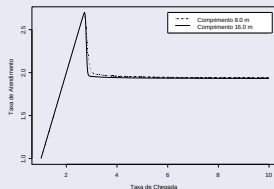
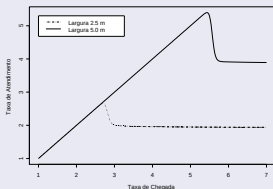


Análise de Filas Simples

Fila finita criada por um segmento de rodovia com largura e comprimento finitos conjuntamente com o tamanho dos veículos

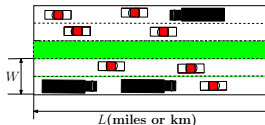


Variação da taxa de chegada

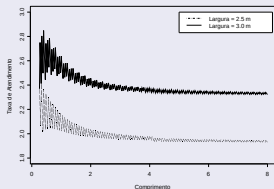




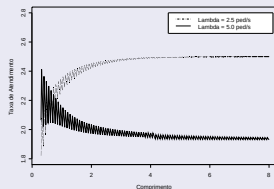
Análise de Filas Simples (cont.)



Variação do Comprimento



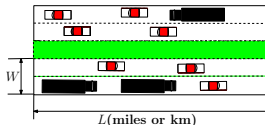
$$\lambda = 5,0 \text{ ped/s}$$



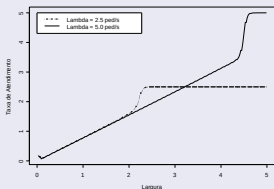
Largura fixa em 2,5 m



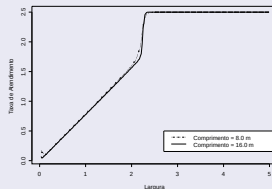
Análise de Filas Simples (cont.)



Variação da Largura



Comprimento 8,0 m

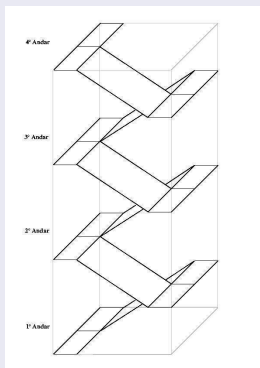


$\lambda = 2,5 \text{ ped/s}$

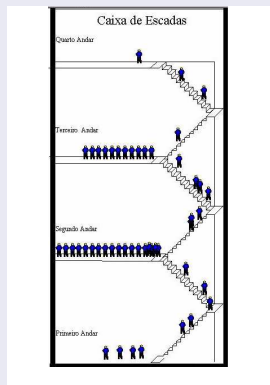


Análise de Redes de Filas

Análise de redes de filas dependentes do estado ($M/G/c/c$)



Caixa de escadas

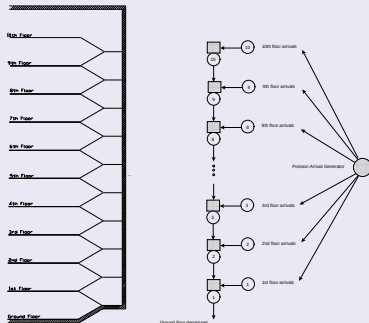


Simulação no Arena



Análise de Redes de Filas (cont.)

Um configuração real maior e o modelo de filas

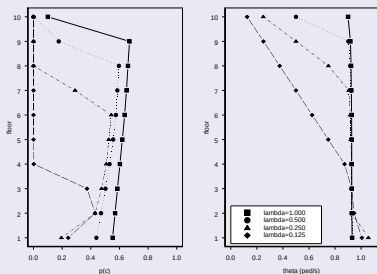




Análise de Redes de Filas (cont.)

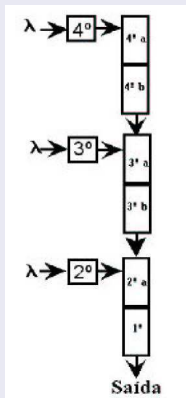
- Análise via modelo de simulação a eventos discretos (Cruz et al., 2005):

Probabilidade de bloqueio e taxa de saída

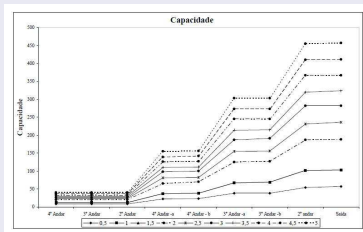




Otimização de Rotas de Evacuação



Estrutura

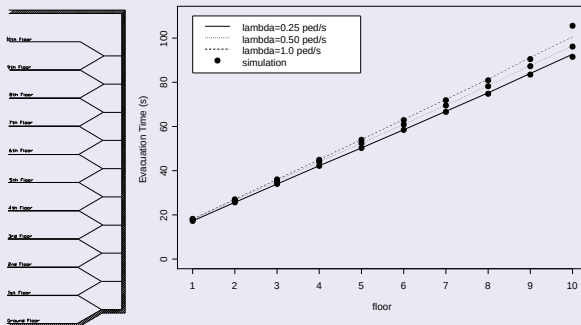


Alocação de capacidades



Otimização de Rotas de Evacuação (cont.)

Otimização de redes de filas dependentes do estado ($M/G/c/c$), via técnicas de busca elementares





Alocação de Áreas de Espera

- O problema é determinar quanto espaço deve ser alocado a fim de se obter um certo nível de qualidade (máxima taxa de saída, mínimo tempo de permanência no sistema, mínimo trabalho em processo);
- Foco em filas configuradas em redes com tempos de serviço gerais;
- Formulação clássica:

$$Z = \min \left(\sum_i x_i \right), \quad (1)$$

s.a.:

$$\Theta(\mathbf{x}) \geq \Theta^{\min}, \quad (2)$$

$$x_i \in \mathbb{N}, \forall i, \quad (3)$$

que minimiza a área total de espera, $\sum_i x_i$, $x_i \equiv K_i$, sujeito ao provimento de uma taxa de saída mínima $\Theta^{\min}$.



Alocação de Áreas de Espera (cont.)

- Resolvido via técnica de relaxação lagrangeana (Cruz, Duarte, van Woensel, 2008):

$$L(\alpha) = \min \left[\sum_{i=1}^N x_i + \alpha \underbrace{(\Theta^{\min} - \Theta(\mathbf{x}))}_{\leq 0} \right] \quad (4)$$

s.a:

$$x_i \in \mathbb{N}, \forall i,$$

$$\alpha \geq 0.$$

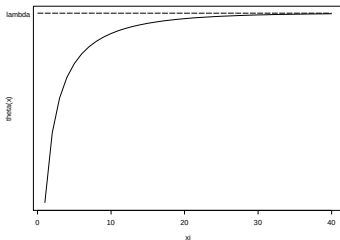
Teorema

Se $\Theta^{\min}$ é exatamente igual à taxa de saída total Λ , então o maior limite superior, $L(\alpha^) = \max_{\alpha \geq 0} L(\alpha)$, é obtido para $\alpha^* \rightarrow \infty$.*

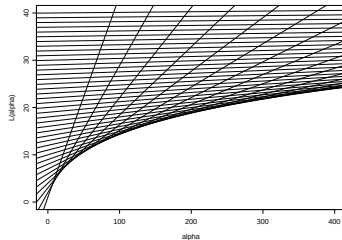


Alocação de Áreas de Espera (cont.)

- Uma prova segue de:



Taxa de saída *versus* área de espera



Função lagrangeana $L(\alpha)$.

- um valor prático para o multiplicador de Lagrange α pode ser obtido por $\alpha(\Theta^{\min} - \Theta(\mathbf{x})) \leq 1$,
- por exemplo, assumindo $(\Theta^{\min} - \Theta(\mathbf{x})) \leq 10^{-3}$, tem-se $\alpha_{\epsilon} = 10^3$.



Alocação de Áreas de Espera (cont.)

- Um formulação multi-objectivo alternativa:

$$\text{optimize } Z = \{f_1(\mathbf{x}), f_2(\mathbf{x})\} \quad (5)$$

s.a.:

$$x_i \in \{1, 2, \dots\}, \forall i, \quad (6)$$

em que

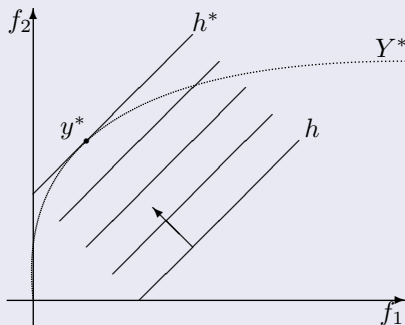
$$\begin{aligned} f_1(\mathbf{x}) &= \sum_i x_i, \text{ área de espera,} \\ f_2(\mathbf{x}) &= \Theta(\mathbf{x}), \text{ taxa de saída,} \end{aligned}$$

- A avaliação da perda em um dos objetivos (p.e., taxa de saída) comparada à melhoria simultânea no outro objetivo (p.e., a área de espera, $x_i \equiv K_i$) resulta no conjunto de Pareto ótimo;



Alocação de Áreas de Espera (cont.)

- Conjunto de Pareto e uma solução aproximada obtida por meio de um algoritmo mono-objetivo:





Alocação de Áreas de Espera (cont.)

Algoritmo Genético

- Abordagem eficiente para problemas multi-objetivos;
- Baseia-se na informação obtida da avaliação de diversos pontos no espaço de busca;
- Convergência baseada nos operadores genéticos clássicos:
 - mutação;
 - cruzamento;
 - seleção;
 - elitismo;
- Faz ainda uso de:
 - atribuição de aptidão (*fitness*) a cada membro da população;
 - preservação da diversidade;



Alocação de Áreas de Espera (cont.)

Algoritmo NSGA-II

algorithm

read graph, arrival, service rates, $G(V, A), \lambda_v, \mu_v, \forall v \in V$

$P_1 \leftarrow \text{GenerateInitialPopulation}(\text{popSize})$

for $i = 1$ **until** numGen **do**

/ generate offspring by crossover and mutation */*

$Q_i \leftarrow \text{MakeNewPop}(P_i)$

/ combine parent and offspring */*

$R_t \leftarrow P_t \cup Q_t$

/ find non-dominated fronts $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \dots)$ */*

$\mathcal{F} \leftarrow \text{FastNonDominatedSort}(R_t)$

/ find new population by the crowding-distance-assignment */*

$P_{t+1} \leftarrow \text{GenerateNewPopulation}(R_t)$

end for

$P_{\text{numGen}+1} \leftarrow \text{ExtractParetoSet}(P_{\text{numGen}})$

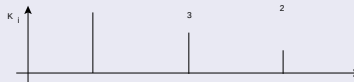
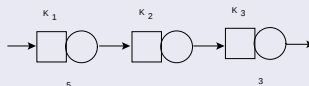
write $P_{\text{numGen}+1}$

end algorithm



Alocação de Áreas de Espera (cont.)

Cruzamento

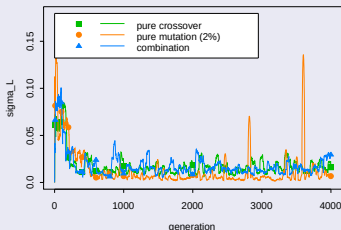




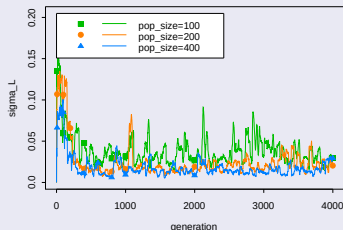
Alocação de Áreas de Espera (cont.)

Parâmetros relevantes do NSGA-II (i.e., tamanho da população, número máximo de gerações, taxa de mutação etc.) foram ajustados, para assegurar a convergência:

Ajuste do Algoritmo



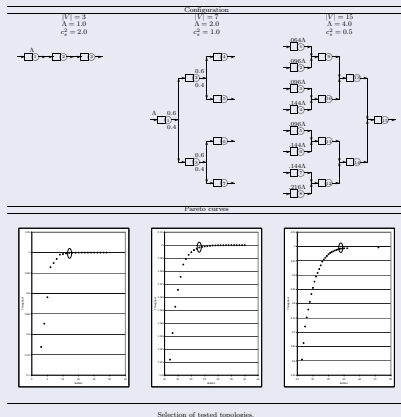
a) cruzamento e mutação



b) tamanho da população



Alocação de Áreas de Espera (cont.)





Trabalhos em Progresso

- Colaborações?



Otimização Multiobjetivo de K e μ

Formulação Matemática

$$\text{minimize } F(\mathbf{K}, \boldsymbol{\mu}), \quad (7)$$

s.a.:

$$K_i \in \{1, 2, \dots\}, \forall i \in N, \quad (8)$$

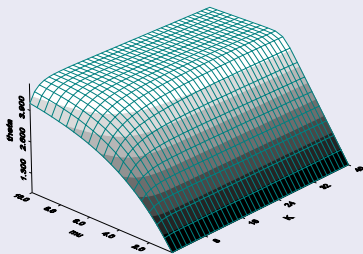
$$\mu_i \geq 0, \forall i \in N, \quad (9)$$

em que as variáveis de decisão K_i e μ_i indicam a capacidade total e a taxa de serviço total para a i -ésima fila $M/G/1/K$. As funções-objetivo, $F(\mathbf{K}, \boldsymbol{\mu}) \equiv (f_1(\mathbf{K}), f_2(\boldsymbol{\mu}), -f_3(\mathbf{K}, \boldsymbol{\mu}))$, são a capacidade total alocada, $f_1(\mathbf{K}) = \sum_{i \in N} K_i$, a taxa de serviço total alocada, $f_2(\boldsymbol{\mu}) = \sum_{i \in N} \mu_i$, e a taxa de serviço global, $f_3(\mathbf{K}, \boldsymbol{\mu}) = \Theta(\mathbf{K}, \boldsymbol{\mu})$.

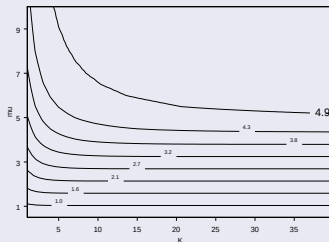


Otimização Multiobjetivo de K e μ (cont.)

Resultados para uma fila única fila $M/G/1/K$, para $\lambda = 5$



a) Θ versus μ e K

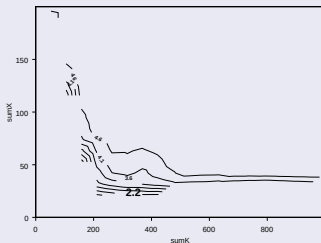
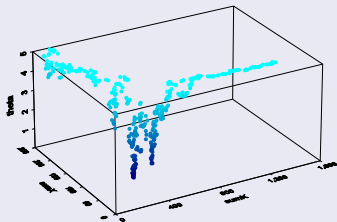


b) curvas de nível



Otimização Multiobjetivo de K e μ (cont.)

Resultados para redes de filas (preliminar)

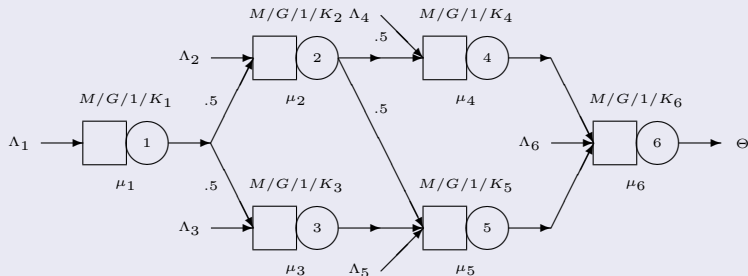


a) superfície final para $c_s^2 = 0,5$ b) curvas de nível para $c_s^2 = 0,5$



Otimização Multiobjetivo de K e μ (cont.)

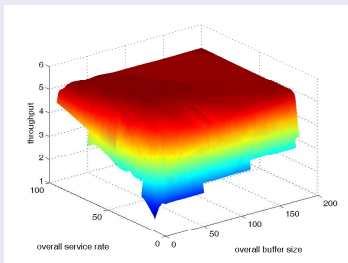
Rede analisada



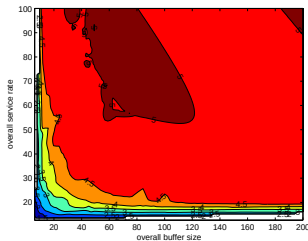


Otimização Multiobjetivo de K e μ (cont.)

Resultados



superfície final para $c_s^2 = 0,5$

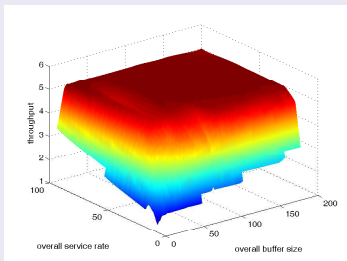


curvas de nível para $c_s^2 = 0,5$

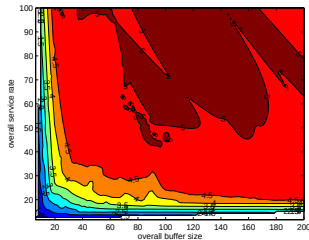


Otimização Multiobjetivo de K e μ (cont.)

Resultados (cont.)



superfície final para $c_s^2 = 1,5$



curvas de nível para $c_s^2 = 1,5$



Otimização Multiobjetivo de K e μ (cont.)

Soluções de Pareto selecionadas

c_s^2	$f_1(\mathbf{K}) = \sum_{\forall i \in N} K_i$	$\Delta\%$	$f_2(\boldsymbol{\mu}) = \sum_{\forall i \in N} \mu_i$	$\Delta\%$	$f_3(\mathbf{K}, \boldsymbol{\mu}) = \Theta(\mathbf{K}, \boldsymbol{\mu})$	$\Delta\%$
0,5	316		20,3		4,850	
	480	52	20,3	0	4,935	2
	340		20,5		4,899	
	343	1	26,0	27	5,000	2
1,5	118		23,6		4,656	
	378	220	23,6	0	4,956	6
	520		20,5		4,914	
	520	0	22,7	11	4,991	2



Roteamento em Filas Finitas

Formulação Matemática

$$\max \Theta(\boldsymbol{\alpha}), \quad (10)$$

s.a.:

$$0 \leq \alpha_{i,j} \leq 1, \quad \forall (i,j) \in A, \quad (11)$$

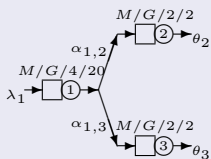
$$\sum_{j \in \delta(i)} \alpha_{i,j} = 1, \quad \forall i \in V, \quad (12)$$

em que $\Theta(\boldsymbol{\alpha})$ é a taxa de saída, que é o objetivo a ser minimizado, $\boldsymbol{\alpha}$ é a matriz de probabilidades de roteamento, $\boldsymbol{\alpha} \equiv [\alpha_{i,j}]$, i.e. a matriz que maximiza a função-objetivo desta rede predefinida, e $\delta(i)$ é o conjunto de nós sucessores ao nó i , ou seja, $\delta(i) \equiv \{j \mid (i,j) \in A\}$.

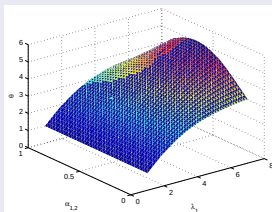


Roteamento em Filas Finitas

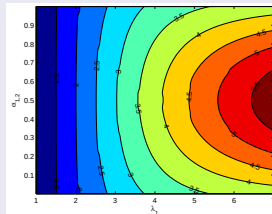
Forma de função objetivo



Rede básica na
topologia divisão



a) efeito do λ_1 versus $\alpha_{1,2}$
na taxa de atendimento

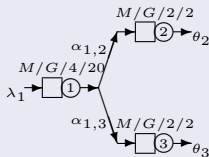


b) curvas de nível para λ_1
versus $\alpha_{1,2}$

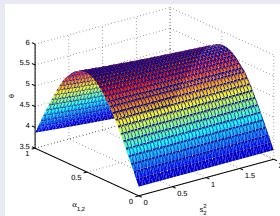


Roteamento em Filas Finitas

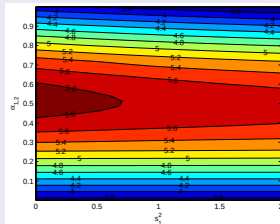
Forma de função objetivo (cont.)



Rede básica na
topologia divisão



a) efeito do $c_{s_2}^2$ versus $\alpha_{1,2}$
na taxa de atendimento



b) curvas de nível para $c_{s_2}^2$
versus $\alpha_{1,2}$



Observações Finais

Sumário

- Simulações comprovam que o método da expansão generalizado produz estimativas tipicamente dentro de 5% de erro, para topologias básicas de pequeno porte (até 9 nós);
- Contudo, o erro máximo pode ser bem maior, de cerca de 20%, principalmente para configurações sob tráfego pesado;
- Esses resultados têm um impacto na forma como se pode utilizar o método em algoritmos multiobjetivos, quando configurações congestionadas precisam ser analisadas.



Observações Finais (cont.)

Tópicos para trabalhos futuros

- Tópicos futuros de pesquisa incluem a extensão para redes gerais, com ciclo, para modelar retrabalho e sistemar com realimentação;
- Adaptação do MEG para utilização em outros tipos de filas finitas ($GI/G/c/K$, $GI^{[X]}/G/c/K$ etc.);
- Investigação de problemas de redes de filas finitas do mundo real;
- ?



Bibliografia I

-  Cruz, F. R. B., Duarte A. R. & van Wonsel, T. (2008). Buffer Allocation in General Single-Server Queueing Networks. *Computers & Operations Research*, 35(11):3581-3598.
-  Cruz, F. R. B., Oliveira, P.C.; Duczmal, L. H. (2010). State-dependent stochastic mobility model in mobile communication networks. *Simulation Modelling Practice and Theory*, 18(3):348-365.
-  Cruz, F. R. B., Smith J. M., Medeiros R. O. (2005). An $M/G/c/c$ State Dependent Queueing Network Simulation Model. *Computers & Operations Research*, 32:919-941.
-  Cruz, F. R. B., Smith J. M., Queiroz D. C. (2005b). Service and Capacity Allocation in $M/G/c/c$ State Dependent Queueing Networks. *Computers & Operations Research*, 32:1545-1563, 2005.



Bibliografia II

-  Jouini, O., Dallery, Y. & Aksin, Z. (2009). Queueing Models for Full-flexible Multi-class Call Centers with Real-time Anticipated Delays. *International Journal of Production Economics*, 120(2):389-399.
-  Osorio, C. & Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, *European Journal of Operational Research*, 196(3):996-1007.
-  Spieckermann, S., Gutenschwager, K., Heinzl, H. Voß, S. (2000). Simulation-based Optimization in the Automotive Industry - A Case Study on Body Shop Design. *Simulation*, 75(5):276-286.
-  Tan, B. & Gershwin, S. B. (2009). Analysis of a general Markovian two-stage continuous-flow production system with a finite buffer. *International Journal of Production Economics*, 120:327-339.



Fim

Questões? Envie-me um e-mail (fcruz@est.ufmg.br).

