1. Publicação nº
INPE-2767-PRE/340

2. Versão

3. Data
Junho, 1983

5. Distribuição
☐ Interna ☒ Externa
☐ Restrita

4. Origem *DIN/DEP* Programa *NAS*

6. Palavras chaves - selecionadas pelo(s) autor(es)
FILA M/G/1
CONTROLE DE FILAS
PROCESSO DE MARKOV

7. C.D.U.: *519.872.7*

8. Título *INPE-2767-PRE/340*

CONTROLE ÓTIMO DO TEMPO DE ESPERA NUMA FILA M/G/1

10. Páginas: *14*

11. Última página: *13*

12. Revisada por

Horácio Hideki Yanasse

13. Autorizada por

Nelson de Jesus Parada
Diretor

9. Autoria *Antonio Fernando Branco Costa*
Paulo Renato de Morais

Assinatura responsável

14. Resumo/Notas

Este trabalho considera uma fila M/G/1 com duas taxas de serviço disponíveis e com entrada controlada. Em qualquer instante de tempo, uma decisão deve ser tomada sobre a taxa de serviço a ser empregada e sobre o fechamento ou não do sistema a novas chegadas. Esta decisão é baseada na comparação do trabalho acumulado (tempo de espera virtual), nesse instante, com valores críticos. Existem custos de espera de clientes, de ociosidade do servidor, de troca de taxas de serviço, de perdas de clientes e uma recompensa dependente da taxa de serviço empregada. É obtida a distribuição estacionária da quantidade de serviço acumulado, bem como a expressão do retorno médio por unidade de tempo. Resultados numéricos são apresentados para a fila M/M/1, onde são determinados os valores críticos ótimos que maximizam o retorno médio por unidade de tempo.

15. Observações
Trabalho aceito para apresentação no XVI Simpósio Brasileiro de Pesquisa Operacional, Florianópolis, Outubro de 1983.

CONTROLE ÓTIMO DE TEMPO DE ESPERA NUMA FILA M/G/1

ANTONIO FERNANDO BRANCO COSTA
Faculdade de Engenharia de Guaratinguetá

PAULO RENATO DE MORAIS*
Instituto de Pesquisas Espaciais
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

RESUMO

Este trabalho considera uma fila M/G/1 com duas taxas de serviço disponíveis e com entrada controlada. Em qualquer instante de tempo, uma decisão deve ser tomada sobre a taxa de serviço a ser empregada e sobre o fechamento ou não do sistema a novas chegadas. Esta decisão é baseada na comparação do trabalho acumulado (tempo de espera virtual), nesse instante, com valores críticos. Existem custos de espera de clientes, de ociosidade do servidor, de troca de taxas de serviço, de perdas de clientes e uma recompensa dependente da taxa de serviço empregada. É obtida a distribuição estacionária da quantidade de serviço acumulado, bem como a expressão do retorno médio por unidade de tempo. Resultados numéricos são apresentados para a fila M/M/1, onde são determinados os valores críticos ótimos que maximizam o retorno médio por unidade de tempo.

ABSTRACT

This paper is concerned with an M/G/1 queue with two available service rates and a controlled input. A decision has to be made, at each instant of time, about which service rate to use and whether the entrance to the system should be closed or not for new arrivals. This decision is based on the residual workload (virtual waiting time), at this time, compared with critical numbers. There are customer's waiting costs, server's iddleness costs, service rate switching costs, penalty costs due to lost customers, and a reward dependent upon the service rate being used. The stationary distribution of the residual workload and the average return per unit time are obtained. Numerical results are given for the M/M/1 queue, where the optimal critical values are determined so as to maximize the average return per unit time.

1. INTRODUÇÃO

O problema de controle de filas já foi abordado por diversos autores (Crabill et alii, 1977). Em geral, o controle atua somente no processo de chegadas (Doshi, 1977) ou somente no processo de atendimentos (Thatcher, 1968; Cohen, 1976b; Tijms, 1977; Doshi, 1978) ou ainda em ambos (Tijms and van der Duyn ' Schouten, 1977; Marins, 1981).

O controle em todos esses trabalhos atua continuamente no tempo, baseado na quantidade de serviço acumulado no sistema (tempo virtual de espera), o qual deve ser processado pelo servidor, e não no comprimento da fila.

Este trabalho baseia-se principalmente em Cohen (1976b). Este autor considera o controle do processo de atendimentos de uma fila M/G/1 com duas taxas de serviço disponíveis através de uma política de controle que atua sobre a estação de serviços, aumentando a taxa de atendimento sempre que a quantidade de serviço acumulado no sistema no instante t, X_t, exceder a um nível crítico k. Isto visa a uma redução no tempo de espera e, conseqüentemente, nos seus custos.

Outra maneira de diminuir os custos de espera é fechar o sistema para novas entradas sempre que X_t exceder a um certo nível crítico.

Assim sendo, aqui se estende o trabalho de Cohen (1976b), considerando-se agora uma política de controle de dois níveis críticos k_1 e k_2, com $k_1 < k_2$, que atua sobre a estação de serviços aumentando a taxa de atendimento sempre que X_t exceder ao nível crítico k_1, bem como sobre o acesso à estação de serviços fechando-o sempre que X_t exceder ao nível crítico k_2.

Associados ao funcionamento do sistema existem prêmios relacionados com o processo de atendimentos e quatro custos distintos: custo de espera de clientes, custo de ociosidade do servidor, custo de troca de taxa de atendimento e custo de perda de clientes devido ao fechamento do sistema.

Na Seção 2 formula-se o modelo de controle. Na Seção 3 determina-se a distribuição estacionária da quantidade de serviço acumulado no sistema. Na Seção 4 determina-se a expressão do retorno médio por unidade de tempo. Para o caso particular em que $k_1 = k_2 = k$, na Seção 5 determina-se o valor ótimo de k que maximiza o retorno médio por unidade de tempo. Finalmente, na Seção 6 faz-se uma aplicação numérica à fila M/M/1, considerando três casos particulares de controles e comparando os resultados obtidos.

2. DESCRIÇÃO DO MODELO DE CONTROLE

Considera-se uma fila M/G/1 onde as chegadas ocorrem de acordo com um processo de Poisson a uma taxa $\lambda > 0$. Se o sistema está aberto, os clientes juntam-se à fila; em caso contrário eles são perdidos. Denota-se por $\{Y_n\}$ a sequência formada pelas quantidades de trabalho demandadas por chegadas sucessivas e admite-se que $\{Y_n\}$ é formada por variáveis aleatórias independentes, com função de distribuição de probabilidades $B(\cdot)$ comum a todas, a qual satisfaz:

a) $B_Y(\cdot)$ é contínua em $[0,\infty)$, com $B_Y(0) = 0$;

b) $E(Y) = \beta$;

c) $E(Y^2) = \beta_2 < \infty$ e $E(Y^3) = \beta_3 < \infty$;

onde Y é uma variável aleatória com função de distribuição de probabilidades $B(\cdot)$.

A disciplina de atendimento será de acordo com "Primeiro a Entrar, Primeiro a Sair (PEPS)". O trabalho existente no sistema será processado de acordo com uma taxa $\sigma \in \{\sigma_1, \sigma_2\}$, $\sigma_2 > \sigma_1$. Definem-se:

$$a_1 = \frac{\lambda\beta}{\sigma_1} \quad \text{e} \quad a_2 = \frac{\lambda\beta}{\sigma_2}$$

e supõe-se que $a_2 < 1$.

Associado a este modelo de filas, existe o processo estocástico $\{X_t;\ t \geq 0\}$, onde X_t é o nível de trabalho acumulado no sistema no instante t. Se $X_t \leq k_1$, o controlador deve manter o sistema aberto e usar σ_1. Se $k_1 < X_t \leq k_2$, o controlador deve manter o sistema aberto e usar σ_2. Finalmente, se $X_t > k_2$, deve fechar o sistema para novas entradas e usar σ_2.

Admite-se a seguinte estrutura de prêmios e custos:

a) retornos a taxa $h_i > 0$ enquanto estiver em uso a taxa de atendimento σ_i $(i = 1, 2)$;

b) custos de espera avaliados por $c_1 X_t$ se $X_t \leq k_1$, por $c_2 X_t$ se $k_1 < X_t \leq k_2$, e por $c_3 X_t$ se $X_t > k_2$;

c) custo de ociosidade $h_3 > 0$ por unidade de tempo, cobrado enquanto o servidor permanecer ocioso;

d) custo de perda de clientes $P > 0$ por unidade de tempo, cobrado enquanto o acesso à estação de serviços estiver fechado;

e) custo de troca de taxa de atendimento avaliado por $f_1 > 0$ se a troca for de σ_1 para σ_2 e por $f_2 > 0$ se for de σ_2 para σ_1.

3. DETERMINAÇÃO DA DISTRIBUIÇÃO ESTACIONÁRIA V(·) DA QUANTIDADE DE SERVIÇO ACUMULADO NO SISTEMA

A partir das Equações 3.2 e 4.2 de Gaver and Miller (1962), determina-se que a função de distribuição $V(x,t) = \Pr\{X_t \leq x\}$ da variável aleatória X_t satisfaz as equações integro-diferenciais:

$$\frac{\partial}{\partial t} V(x,t) = \sigma_1 \frac{\partial}{\partial x} V(x,t) - \lambda V(x,t) + \lambda \int_0^x B(x-y) d_y V(y,t)$$

para $0 \leq x < k_1$;

$$\frac{\partial}{\partial t} V(x,t) = \sigma_2 \frac{\partial}{\partial x} V(x,t) - \lambda V(x,t) + \lambda \int_0^x B(x-y) d_y V(y,t)$$

para $k_1 \leq x < k_2$;

$$\frac{\partial}{\partial t} V(x,t) = \sigma_2 \frac{\partial}{\partial x} V(x,t) - \lambda V(x,t) + \lambda \int_0^{k_2} B(x-y) d_y V(y,t) + $$
$$+ \lambda \int_{k_2}^x d_y V(y,t) \quad \text{para } x \geq k_2;$$

onde todas as derivadas que aparecem são derivadas ou à direita para $x \geq 0$ ou à esquerda para $x > 0$.

Trabalhando com estas equações, determinam-se duas expressões para a transformada de Laplace-Stieltjes da distribuição estacionária da quantidade de serviço acumulado no sistema:

$$\phi(\rho) = \int_{0^-}^{\infty} e^{-\rho x}\, dV(x), \quad \text{ou seja:}$$

$$\phi(\rho) = \frac{V(0) + (1-\frac{a_1}{a_2}) \int_{k_1}^{k_2} e^{-\rho x}\, dV(x) + \left[1-\frac{a_1}{a_2} - a_1 (\frac{1-\beta(\rho)}{\beta\rho})\right] \int_{k_2}^{\infty} e^{-\rho x}\, dV(x)}{1 - a_1(\frac{1-\beta(\rho)}{\beta\rho})},$$

$$\phi(\rho) = \left[1 + a_2(\frac{1-\beta(\rho)}{\beta\rho})\right] V(0) + a_2(\frac{1-\beta(\rho)}{\beta\rho}) \int_{k_1}^{k_2} e^{-\rho x}\, dV(x) +$$

$$+ \left[1 - \frac{a_2}{a_1} + a_2(\frac{1-\beta(\rho)}{\beta\rho})\right] \int_{0}^{k_1} e^{-\rho x}\, dV(x),$$

onde:

$$\beta(\rho) = \int_{0^-}^{\infty} e^{-\rho x}\, dB(x) \qquad \text{e}$$

$V(0) = \lim_{t\to\infty} V(0,t) = Pr\ \{X_t = 0\}$, isto é, $V(0)$ é a porcentagem de tempo em que o servidor permanece ocioso.

Uma vez conhecida a distribuição de probabilidades $B(\cdot)$, pode-se determinar a distribuição estacionária $V(\cdot)$ usando a metodologia desenvolvida por Cohen (1969), obtendo-se o seguinte resultado (Costa, 1983):

$V(x) = V(0) \cdot W_1(x)$ para $x \leq k_1$, com:

$$\int_{0^-}^{\infty} e^{-\rho x}\, dW_1(x) = \left[1 - a_1(\frac{1-\beta(\rho)}{\beta\rho})\right]^{-1},$$

$V(x) = V(0) \cdot W_2(x)$ para $k_1 < x \leq k_2$, com:

$$\int_{0^-}^{\infty} e^{-\rho x}\, dW_2(x) = \frac{1 + (1 - \frac{a_2}{a_1}) \int_0^{k_1} e^{-\rho x}\, dW_1(x)}{1 - a_2(\frac{1-\beta(\rho)}{\beta\rho})},$$

$$V(0) = \left[\frac{a_2}{a_1} + a_2(W_2(k_2)) + (1 - \frac{a_2}{a_1})\, W_1(k_1)\right]^{-1}.$$

4. DETERMINAÇÃO DA EXPRESSÃO DO RETORNO MÉDIO POR UNIDADE DE TEMPO

Seja C_k a duração de um ciclo de renovações, isto é, o tempo entre dois instantes sucessivos t_1 e t_2 com $X_{t_i^-} = 0$ e $X_{t_i^+} > 0$, $i = 1, 2$. Segue-se facilmente que:

$$E(C_k) = (\lambda\, V(0))^{-1}.$$

Da teoria dos processos regenerativos segue-se para qualquer função mensurável de Borel $g: R \to R$ que:

$$\int_{0^-}^{\infty} g(x)\, dV(x) = \frac{1}{E(C_k)}\, E\left\{\int_0^{C_k} g(x_t)\, dt\right\},$$

se a primeira integral existe (Feller, 1971).

Desta maneira, o custo médio de espera por unidade de tempo será:

$$I(k_1,k_2) = c_1 \int_0^{k_1} x\, dV(x) + c_2 \int_{k_1}^{k_2} x\, dV(x) + c_3 \int_{k_2}^{\infty} x\, dV(x).$$

O custo médio de ociosidade por unidade de tempo será:

$$T(k_1,k_2) = h_3\, V(0).$$

O custo médio por manter o sistema fechado por unidade de tempo será:

$$P(k_1,k_2) = P\,[1 - V(0)\, W_2(k_2)].$$

O ganho médio por unidade de tempo, pelo uso das taxas de atendimento, será:

$$R(k_1,k_2) = h_1 \left[V(0)\ W_1(k_1) - V(0)\right] + h_2 \left[1 - V(0)\ W_1(k_1)\right].$$

Seja $S(k_1,k_2)$ o custo médio por unidade de tempo de troca de taxas de serviço de σ_1 para σ_2 e vice-versa. Para determinar $S(k_1,k_2)$, é necessário obter o número médio de cruzamentos do processo estocástico $\{X_t;\ t \geq 0\}$ com o nível k_1 durante um ciclo ocupado. Sabe-se que o número de cruzamentos do nível k_1 durante um ciclo ocupado, a partir de valores inferiores a esse nível, é, com probabilidade um, igual ao número de cruzamentos do nível k_1, a partir de valores superiores a esse nível, durante um ciclo ocupado. Seja D_{k_1} o número de cruzamentos do processo $\{X_t;\ t \geq 0\}$ com o nível $k_1 > 0$, durante um ciclo ocupado, a partir de valores superiores a esse nível.

Cada instante no qual ocorre tal cruzamento do nível k_1 é um instante regenerativo do processo $\{X_t;\ t \geq 0\}$; segue-se imediatamente que a distribuição de D_{k_1} para o presente modelo é a mesma distribuição do número de cruzamentos, a partir de valores superiores ao nível k, do processo $\{X_t;\ t \geq 0\}$ com o nível k durante um ciclo ocupado de um sistema de filas M/G/1 com intensidade de tráfego $a_1 \leq 1$ (Cohen, 1976b).

Esta distribuição foi estudada em Cohen (1969) e, dos resultados lá obtidos [ver Capítulo III.5, Equação 5.93], segue-se que para o presente modelo:

$$E\{D_{k_1}\} = W_1(k_1) - B(k_1) * W_1(k_1) = \left(\frac{\beta}{a_1}\right) \frac{dW_1(k_1)}{dk_1}.$$

Logo:

$$S(k_1;k_2) = \left[(f_1 + f_2)\ \sigma_1\right] V(0) \frac{dW_1(k_1)}{dk_1}.$$

Finalmente, tem-se que o retorno médio por unidade de tempo, $G(k_1,k_2)$, é dado por:

$$G(k_1,k_2) = R(k_1,k_2) - I(k_1,k_2) - T(k_1,k_2) - P(k_1,k_2) - S(k_1,k_2)$$

ou:

$$G(k_1,k_2) = h_1 \left[V(0)\, W_1(k_1) - V(0)\right] + h_2 \left[1 - V(0)\, W_1(k_1)\right] -$$

$$- \left[c_1 \int_0^{k_1} x\, dV(x) + c_2 \int_{k_1}^{k_2} x\, dV(x) + c_3 \int_{k_2}^{\infty} x\, dV(x)\right] - h_3 V(0)$$

$$- P\left[1 - V(0)\, W_2(k_2)\right] - \left[(f_1 + f_2)\, \sigma_1\right] V(0)\, \frac{dW_1(k_1)}{dk_1} . \qquad (1)$$

5. DETERMINAÇÃO DO VALOR ÓTIMO DE k ($k_1 = k_2 = k$) QUE MAXIMIZA O RETORNO MÉDIO POR UNIDADE DE TEMPO

Para o caso particular em que $k_1 = k_2 = k$ tem-se a seguinte expressão para o retorno médio por unidade de tempo (Costa, 1983):

$$G(k) = h_2 - P + V(0) \left\{ - (c_1 + c_2\,(a_2 - \frac{a_2}{a_1}))\int_0^k x\, dW_1(x) - c_2\, a_2 (\frac{\beta_2}{2\beta})\, W_1(k) + \right.$$

$$\left. + h_1 (W_1(k) - 1) - (h_2 - P)\, W_1(k) - h_3 - (f_1 + f_2)\, \sigma_1 \frac{dW_1(k)}{dk} \right\} . \qquad (2)$$

Na análise de G(k) o termo que envolve os custos de troca de taxas exige uma investigação da derivada $dW_1(k)/dk$ que nem sempre existe e, quando existe, pode ter um comportamento razoavelmente irregular (Cohen, 1976b).

Porém, para $a_1 = 1$ a função $W_1(\cdot)$ é uma função de renovação tal que (Cohen, 1976a):

$$W_1(x) \to \frac{2\beta}{\beta_2}\, x + \frac{2\beta_3\beta}{3\beta_2^2} , \qquad (3)$$

quando $x \to \infty$.

Esta relação é exata se $B(x) = 1 - e^{-x/\beta}$, $x \geq 0$. Contudo, para uma grande classe de distribuições $B(\cdot)$, a Relação 2 é, também para valores moderados de x, uma aproximação muito boa para a expressão exata de $W_1(\cdot)$.

No presente caso tem-se (Costa, 1983):

$$\frac{dG(k)}{dk} = V(0)^2 \left[Ak^2 + Bk + C \right],$$

onde:

$$A = -2\,c_1 \left(\frac{\beta}{\beta_2}\right)^2,$$

$$B = -2\,c_1 \left[a_2(\beta/\beta_2) + 2/3\ (\beta/\beta_2)^2\ (\beta_3/\beta_2) \right],$$

$$C = \Big\{ h_1\, a_2 + h_1 + h_3 + P\, a_2 + (f_1 + f_2)\,(\sigma_1)\,(2\beta/\beta_2) -$$

$$- h_2\, a_2 - c_2\ a_2^2\ (\beta_2/2\beta) \Big\}\ (2\beta/\beta_2).$$

O valor ótimo de k, k*, será portanto a raiz positiva da expressão:

$$Ak^2 + Bk + C = 0,$$

e a condição para a existência do valor ótimo, k*, é que $C \geqslant 0$.

6. UMA APLICAÇÃO À FILA M/M/1

O modelo de controle considerado nesta seção é o mesmo utilizado para a fila M/G/1, admitindo-se agora que $B(x) = 1 - e^{-x/\beta},\ x \geqslant 0$.

Neste caso, não é necessário fazer $a_1 = 1$, uma vez que a derivada $dW_1(k)/dk$ existe e pode ser obtida da Relação 2. Para o caso particular em que $k_1 = k_2 = k$ a derivada da expressão do retorno médio por unidade de tempo será dada por (Costa, 1983):

$$\frac{dG(k)}{dk} = V(0)^2 \Bigg[-B\left(\frac{a_2}{a_1}\right) k - B\left(1 + a_2 - \frac{a_2}{a_1}\right) \int_0^k W_1(x)\ dx +$$

$$+ C\left(\frac{a_2}{a_1}\right) + D\left(1 + a_2 - \frac{a_2}{a_1}\right) + \frac{E}{\beta} \Bigg] \frac{dW_1(k)}{dk},$$

onde:

$$B = c_1 + c_2(a_2 - a_2/a_1),$$

$$C = h_1 - h_2 + P - c_2\ a_2\beta,$$

$D = h_1 + h_3$,

$E = (f_1 + f_2)\,\sigma_1$.

O valor de k, k*, que maximiza o retorno médio por unidade de tempo, é o valor correspondente à única raiz positiva da equação implícita:

$$k + A_1 \int_0^k W_1(x)\, dx = A_2 \ , \tag{3}$$

onde:

$A_1 = a_1/a_2 + a_1 - 1$,

$A_2 = -C/\beta + (D/\beta)\left[(a_1/a_2) + a_1 - 1\right] + (E/\beta)\left(\dfrac{a_1}{a_2\beta}\right)$,

desde que:

$B < 0$ e $C(a_2/a_1) - D(1 + a_2 - a_2/a_1) - E(1/\beta) \geq 0$.

Para efeito de comparação numérica, a seguir consideram-se três tipos diferentes de controles que atuam em uma fila M/M/1, a saber:

<u>CONTROLE A</u>: é um controle que atua sobre a estação de serviços, aumentando a taxa de atendimento sempre que a quantidade de serviço acumulado no sistema exceder a um nível crítico k.

<u>CONTROLE B</u>: é um controle que atua sobre a estação de serviços, aumentando a taxa de atendimento, e sobre o acesso à estação de serviços, fechando-o, sempre que a quantidade de serviço acumulado no sistema exceder a um nível crítico k.

<u>CONTROLE C</u>: é um controle que utiliza dois níveis críticos, k_1 e k_2, com $k_1 < k_2$, e atua sobre a estação de serviços, aumentando a taxa de atendimento sempre que a quantidade de serviço acumulado no sistema exceder a um nível crítico k_1, e sobre o acesso à estação de serviços, fechando-o quando a quantidade de serviço acumulado no sistema exceder ao nível crítico k_2.

Utilizam-se os seguintes dados:

$\lambda = 12,0$; $1/\beta = 3,0$; $c_1 = c_2 = c_3 = 8,0$; $P = 1,0$;

$(f_1 + f_2) = 0,5$; $h_3 = 1,0$ e $\sigma_1 = 5,0$.

Os valores de h_1 e h_2 serão dados pela seguinte expressão:

$$h_i = -0,1\,\sigma_i^2 + 2,0\,\sigma_i\,, \quad i = 1, 2.$$

Desta forma procura-se encontrar não só os valores ótimos de k, k_1 e k_2, como também a melhor relação $A = \sigma_2/\sigma_1$, $A \geq 1,0$, de forma a maximizar $G(k)$ ou $G(k_1,k_2)$.

Os seguintes resultados foram obtidos:

CONTROLE A: $k^* = 0,875$, $A^* = 2,20$ e $G(k^*) = 0,798$;

CONTROLE B: $k^* = 0,891$, $A^* = 2,20$ e $G(k^*) = 0,814$;

CONTROLE C: $k_1^* = 0,852$, $A^* = 2,10$ e $G(k_1^*,k_2^*) = 0,821$; $k_2^* = 1,015$.

O resultado para o Controle A foi obtído a partir de equações derivadas por Cohen (1976b) para a fila M/G/1 e o resultado para o Controle B foi obtido a partir da Equação 3. Para a obtenção do resultado para o Controle C a partir da Equação 1, foi utilizado um algoritmo de programação não-linear (método de Fletcher-Powell).

Conclui-se então nesse exemplo particular que, para valores moderados de P e h_3, o controle que atua simultaneamente nos processos de atendimentos e chegadas (Controle B) promove um melhor retorno que aquele que atua apenas no processo de atendimentos (Controle A), e que a política de controle de dois níveis críticos (Controle C) promove um melhor retorno que a de um único nível crítico (Controle B).

REFERÊNCIAS BIBLIOGRÁFICAS

CRABILL, T.B.; GROSS, D.; MAGAZINE, M. A classified bibliography of research on optimal design and control of queues. *Operations Research*, 25(2): 219-312, 1977.

COHEN, J.W. *The Single Server Queue*. Amsterdam, North-Holland, 1969.

———. On regenerative processes in queueing theory. *Lecture Notes in Economics and Mathematical Systems*, 121, Berlin, Springer, 1976a.

———. On the optimal switching level for an M/G/1 queueing system. *Stochastic Processes and Their Applications*, 4: 297-316, 1976b.

COSTA, A.F.B. *Controle ótimo do tempo de espera numa fila M/G/1*. Tese de Mestrado em Pesquisa Operacional. São José dos Campos, Instituto Tecnológico de Aeronáutica, 1983.

DOSHI, B.T. Continuous time control of the arrival process in an M/G/1 queue. *Stochastic Processes and Their Applications*, 5: 265-285, 1977.

———. Optimal control of the service rate in an M/G/1 queueing system. *Advances in Applied Probability*, 10: 682-701, 1978.

FELLER, W. *An Introduction to Probability Theory and its Applications*. New York, Wiley, 1971. v. II, 2.ed.

GAVER, D.P.; MILLER, R.G. Limiting distributions for some storage problems. In: ARROW, K.T.; KARLIN, S.; SCARF, H., ed. *Studies in Applied Probability and Management Science*. Stanford, Stanford University Press, 1962.

MARINS, F.A.S. *Controle ótimo do processo de chegadas e da taxa de atendimento de um sistema de filas M/G/1*. Tese de Mestrado em Pesquisa Operacional. São José dos Campos, Instituto Tecnológico de Aeronáutica, 1981.

THATCHER, R.M. *Optimal single-channel service policies for stochastic arrivals*. Berkeley, University of California, 1968. (ORC-68-16, Operations Research Center).

TIJMS, H.C. On a switch-over policy for controlling the workload in a queueing system with two constant service rates and fixed switch-over costs. *Zeitschrift für Operations Research*, 21: 19-32, 1977.

TIJMS, H.C.; van der DUYN SCHOUTEN, F.A. Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate. *Stochastic Processes and Their Applications*, 6: 213-222, 1977.