Modelos de Filas com Desistências

Henry Kiyoshi Oyagawa

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração : **Probabilidade Aplicada**
Orientador : **Prof. Dr. Marcos Nascimento Magalhães**

– São Paulo, dezembro de 2004 –

Modelos de filas com desistências

Este exemplar corresponde à redação
final da dissertação devidamente
corrigida e defendida por
Henry Kiyoshi Oyagawa
e aprovada pela comissão julgadora.

São Paulo, dezembro de 2004.

Banca examinadora:

- Prof. Dr. Marcos Nascimento Magalhães (Orientador) - IME - USP
- Profa. Dra. Elizabeti Kira - IME - USP
- Profa. Dra. Graça Bressan - POLI - USP

Aos meus pais,

Tiyoko e Mitsuo

e às minhas irmãs,

Eliane, Tatiana e Inês.

Agradecimentos

Antes de qualquer comentário, devo enorme agradecimento ao meu amigo e orientador, Prof. Dr. Marcos Nascimento Magalhães. Sem ele nem uma frase desta dissertação seria escrita. Agradeço-lhe imensamente por sua dedicação e paciência em orientar meus passos nessa jornada.

À minha colega de pesquisa, Elisa, e aos Profs. Drs. Graça e Rui pelo auxílio com as simulações em Arena e à Prof. Dra. Elizabeti Kira pelas valiosas críticas e sugestões. Ao Eduardo pela colaboração na formatação dos gráficos e à Fernanda Cerdeira pela ajuda na versão em inglês do resumo. Aos meus amigos Alexandre, Edward, Daniel Dantas e Rodnei pelas dicas de Latex.

Agradeço à Carla, Patrícia, Flávia, Takao, Masaishi e Victor pelo companheirismo e apoio nos momentos difíceis. Ao casal, Elisa e Stefan, que mesmo à distância, foram grandes incentivadores. À Carla, Edson, Fabio, Octavio, Orlando e Wellington por compreenderem que eu precisei de todo o tempo disponível para me dedicar aos estudos.

A todos que contribuiram para que eu pudesse concluir este sonho.

Muito obrigado!

Resumo

O objetivo deste trabalho é estudar possíveis melhoras no atendimento de usuários em um sistema de filas, através de informações disponibilizadas aos clientes. Em geral, na visão dos usuários, a qualidade do atendimento está fortemente ligada ao tempo necessário à conclusão do serviço. Entretanto, nem sempre é possível atender prontamente a todas as solicitações. Nesse caso, uma possível contribuição à satisfação do cliente seria informá-lo, no instante de sua chegada, sobre a previsão de tempo de espera até que seja devidamente atendido. A partir dessa informação o usuário pode concluir ser melhor desistir de esperar e sair do sistema. Buscando modelar esses sistemas de atendimento, estuda-se um modelo de filas com desistências. O modelo considera três peculiaridades: impedimento (*blocking*), recusa (*balking*) e abandono (*reneging*), além das características usuais de uma fila. Apresenta-se a estrutura matemática envolvida e, através de simulação, avalia-se o comportamento do sistema sujeito a perturbações. O interesse concentra-se nos efeitos das diferentes características da tolerância dos usuários nas medidas de performance.

Abstract

The main objective of this work is to study possible improving services in a queue system for customers information. Customers tend to associate service time to quality. However, it is not always possible to answer immediately to all requests. In this case, one possible contribution to customers satisfaction is to inform them immediately upon arrival the estimate of the waiting time. However, customers can decide to leave the system. So, this work study a model of queue with abandonment. The model has three features: (1) blocking, (2) balking, and (3) reneging, besides the usual queue characteristics. The mathematical structure involved is presented, and performance is evaluated by submitting the system to several perturbations. The main concern is to focus on the effects that different customers tolerance have on system performance.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O objetivo desta dissertação é estudar o impacto causado na performance de um sistema de atendimento quando, no momento da chegada, a quantidade de usuários existentes no sistema é informada aos clientes. Tal informação será utilizada para estimar o tempo de espera que terão que aguardar na fila.

Partindo de um modelo de filas com sala de espera limitada, será estudado o comportamento de dois modelos semelhantes em que, um deles fornece o estado do sistema para os usuários que chegam enquanto o outro não disponibiliza essa informação. O desenvolvimento e a análise desses modelos foram feitos em [Whitt, 1999b] e serão reapresentados de forma resumida nesta dissertação.

Nesses modelos são consideradas três peculiaridades: *impedimento* (blocking), *recusa* (balking) e *abandono* (reneging). Sempre que a capacidade do sistema estiver esgotada, as chegadas serão obrigatoriamente descartadas, fato designado por *impedimento*. Se todos os servidores estiverem ocupados, mas ainda existirem posições vagas na fila, a nova chegada poderá decidir não permanecer no sistema. A essa ocorrência dá-se o nome de *recusa*. Por outro lado, caso o cliente tenha aceitado aguardar na fila, o tempo de espera pode extrapolar o seu limite aceitável. Nessa situação, o usuário retira-se do sistema configurando o chamado *abandono*. Apesar de representarem saídas do sistema, a principal diferença entre a *recusa* e o *abandono* diz respeito ao momento em que ocorrem. A *recusa* acontece no instante da chegada, enquanto o *abandono* é

efetuado em um estágio posterior, ou seja, durante a espera pelo atendimento.

1.1 Contribuições desta dissertação

O modelo desenvolvido em [Whitt, 1999b] é apresentado e, como extensão desse estudo, serão implementadas simulações para observar o efeito sobre o desempenho dos modelos quando ocorrerem perturbações nos seus parâmetros. O intuito principal é analisar qual é o comportamento do sistema quando diferentes tipos de usuários solicitam serviço.

1.2 Organização do trabalho

Inicialmente, apresentou-se os objetivos que motivaram este trabalho além da descrição das técnicas a serem utilizadas. Nos próximos capítulos serão mostrados os procedimentos adotados e os resultados obtidos. No Capítulo 2 serão descritos conceitos necessários ao entendimento dos modelos. No Capítulo 3 são apresentados os modelos de filas com desistências. A simulação dos modelos é apresentada no Capítulo 4, juntamente com a discussão dos efeitos de variações no modelo original. As conclusões e possíveis extensões são apresentadas no Capítulo 5.

Capítulo 2

Conceitos preliminares

Neste capítulo serão apresentados alguns conceitos que auxiliam na compreensão dos estudos desenvolvidos ao longo desta dissertação.

A abordagem aqui adotada é focada no estudo que será desenvolvido mais à frente e, portanto, será mais restrita do que aquelas usualmente mostradas na literatura.

2.1 Processos de Nascimento e Morte

Uma cadeia de Markov com espaço de estados $V = \{0, 1, \dots\}$ e transições em saltos unitários, ou seja, $q_{ij} = 0$ sempre que $|i - j| > 1$ sendo q_{ij} a probabilidade de transição do estado i para o j, é conhecida como um *Processo de Nascimento e Morte.*

Inicialmente utilizado no estudo do crescimento de populações (o que justifica seu nome) o processo de nascimento e morte representa um importante papel na Teoria das Filas. Tais processos são úteis na modelagem de fenômenos cujas variações de tamanho são unitárias, justamente o que ocorre em muitos modelos de filas.

Sempre que o estado sofre o acréscimo de uma unidade é dito que houve um *nascimento*, ao passo que ao decréscimo é dado o nome de *morte*. Assim, sejam λ_k a *taxa de nascimento* e μ_k a *taxa de morte* do processo quando o sistema está no estado k. Ambas constituem as taxas em que o tamanho do sistema é alterado para mais ou para menos, respectivamente.

Assim, quando há k usuários no sistema,

$$\lambda_k = q_{k,k+1} \qquad e \qquad \mu_k = q_{k,k-1}.$$

Uma vez que $\sum_j q_{kj} = 0$ decorre que

$$q_{kk} = -(\lambda_k + \mu_k), \quad \forall k \in V.$$

2.1.1 Distribuição estacionária dos processos de nascimento e morte

Seja $\{N(t), t > 0\}$ um processo de nascimento e morte com taxas λ_i de nascimento e μ_i de morte, $i \in V$.

Sejam $\pi_n(t) = P(N(t) = n)$ sua distribuição de probabilidade no instante t e π_n sua distribuição estacionária. Seja, ainda, Λ o gerador infinitesimal do processo $N(t)$:

$$\Lambda = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \dots & \dots & \dots \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & 0 & \dots & \dots \\ 0 & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & 0 & \dots \\ 0 & 0 & \mu_3 & -(\mu_3+\lambda_3) & \lambda_3 & \dots \\ 0 & 0 & 0 & \ddots & \ddots & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Segundo a teoria dos processos de Markov, que inclui os processos de nascimento e morte, a distribuição estacionária será calculada a partir da solução do sistema:

$$\begin{cases} \Pi\Lambda = 0, \\ \Pi e = 1, \end{cases} \tag{2.1}$$

sendo:

$$e = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix} \quad \text{e} \quad \Pi = (\pi_0, \pi_1, \ldots).$$

Assim, das equações (2.1), seguem:

$$\begin{aligned} 0 &= -(\lambda_j + \mu_j)\pi_j + \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1}, \qquad (j \geq 1) \\ 0 &= -\lambda_0\pi_0 + \mu_1\pi_1. \end{aligned} \tag{2.2}$$

Reescrevendo esse sistema, obtém-se as *equações de balanço global*:

$$\begin{aligned} \lambda_j\pi_j + \mu_j\pi_j &= \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1}, \qquad (j \geq 1); \\ \lambda_0\pi_0 &= \mu_1\pi_1. \end{aligned} \tag{2.3}$$

Resolvendo as equações (2.3) para $n = 0, 1, 2, \ldots$, segue:

$$\begin{aligned} \pi_1 &= \frac{\lambda_0}{\mu_1}\pi_0 \\ \pi_2 &= \frac{\lambda_1\lambda_0}{\mu_2\mu_1}\pi_0 \\ \pi_3 &= \frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}\pi_0 \\ &\vdots \end{aligned}$$

Utilizando o princípio da indução finita, vem que

$$\pi_n = \prod_{j=1}^{n} \frac{\lambda_{j-1}}{\mu_j}\pi_0, \qquad n \geq 1. \tag{2.4}$$

Sabendo-se que $\sum_{n=0}^{\infty} \pi_n = 1$ então:

$$\pi_0 = \left(1 + \sum_{n=1}^{\infty}\prod_{j=1}^{n} \frac{\lambda_{j-1}}{\mu_j}\right)^{-1}. \tag{2.5}$$

Percebe-se, pela equação (2.5), que uma condição necessária e suficiente para a existência da solução estacionária é a convergência da série:

$$\sum_{n=1}^{\infty} \prod_{j=1}^{n} \frac{\lambda_{j-1}}{\mu_j}.$$

As equações (2.4) e (2.5) são muito úteis no estudo de filas. Em geral, não é simples encontrar a distribuição estacionária para qualquer sistema. Na Seção 2.3 serão apresentados alguns casos especiais.

Antes de passar à próxima seção, será apresentada brevemente a fila M/M/1. Por ser um dos modelos mais simples, serve como base para outros mais elaborados.

A fila M/M/1 apresenta chegadas de acordo com um processo de Poisson com parâmetro λ, um servidor com atendimento exponencial de média $1/\mu$ e sala de espera ilimitada. Assume-se que as chegadas e serviços são processos independentes. A disciplina de atendimento segue a ordem de chegada (*FCFS*, do inglês: First Come, First Served). Sendo $N(t)$ o número de clientes no sistema no instante t, $N(t)$ é um processo de nascimento e morte e pode-se aplicar os resultados acima apresentados. Seja $\rho = \lambda/\mu$ a intensidade de tráfego do sistema. Para $0 < \rho < 1$, a existência da distribuição estacionária para a fila M/M/1 está garantida e é dada por:

$$\pi_0 = 1 - \rho \qquad \text{e} \qquad \pi_n = \rho^n(1-\rho), \quad n \geq 0.$$

2.2 Alguns resultados de filas

Esta Seção descreve de forma sucinta alguns dos principais resultados de filas.

2.2.1 PASTA

Um resultado importante conhecido como *PASTA*, do inglês *Poisson Arrivals See Time Average*, diz respeito à distribuição da quantidade de usuários no instante imediatamente anterior a uma

chegada.

Esse resultado, apesar de bastante geral, não se aplica para todos os tipos de filas (ver [Magalhães, 1996]). Aqui será demonstrada sua validade para a fila M/M/1 (mais detalhes sobre outros modelos de filas serão apresentados na próxima seção).

Sejam $\pi_n^a(t)$ a probabilidade de uma chegada encontrar n usuários no sistema e π_n^a sua distribuição estacionária. Note que $\pi_n^a(t)$ considera o usuário que está chegando na contagem dos n presentes no sistema.

Teorema 2.1 (PASTA) *Em regime estacionário, para a fila M/M/1, segue que:*

$$\pi_n = \pi_n^a,$$

ou seja, a distribuição do número de usuários, nos instantes de chegada, é igual à distribuição em tempo contínuo.

Demonstração:

Seja o evento $A(t, t+\Delta t)$, a ocorrência de uma chegada no intervalo de tempo $(t, t+\Delta t)$. Percebe-se que tal evento é independente do número de usuários no sistema no instante t. Assim,

$$\begin{aligned}
\pi_n^a(t) &= \lim_{\Delta t \to 0} P(N(t) = n | A(t+\Delta t)) \\
&= \lim_{\Delta t \to 0} \frac{P(N(t) = n, A(t+\Delta t))}{P(A(t+\Delta t))} \\
&= \lim_{\Delta t \to 0} \frac{P(A(t+\Delta t)|N(t) = n)P(N(t) = n)}{P(A(t+\Delta t))} \\
&= \lim_{\Delta t \to 0} \frac{P(A(t+\Delta t))P(N(t) = n)}{P(A(t+\Delta t))} \\
&= \lim_{\Delta t \to 0} P(N(t) = n) \\
&= \pi_n(t).
\end{aligned}$$

□

2.2.2 Fórmulas de Little

Para o estudo de desempenho de uma fila, há uma ferramenta de grande ajuda: as *fórmulas de Little.*

Esse resultado relaciona os tempos médios de espera com as quantidades médias de usuários. Sua validade abrange modelos mais gerais, não se restringindo apenas aos modelos markovianos.

Sejam L o número médio de usuários no sistema e L_q o número médio de usuários na fila. Sejam W o tempo médio total de espera de um usuário no sistema e W_q o tempo médio de espera na fila. Também, deve-se lembrar que a disciplina de atendimento é FCFS.

Teorema 2.2 (Fórmulas de Little)

$$L = \lambda W \tag{2.6}$$

e

$$L_q = \lambda W_q. \tag{2.7}$$

Demonstração:

Será feita a demonstração para a fila M/M/1. Para outras filas, ver [Cooper, 1981].

Dado que o atendimento obedece a ordem de chegada, pode-se argumentar que, em equilíbrio, a distribuição de n usuários no sistema, no instante do evento determinado por uma saída, é igual à probabilidade de haver n chegadas durante o tempo total de permanência de um usuário no sistema. Assim, pela lei de probabilidade total,

$$\pi_n^d = \int_0^\infty P(n \text{ chegadas durante } \theta_{tot} | \theta_{tot} = t) dF_w(t), \qquad n \geq 0,$$

sendo, π_n^d a distribuição estacionária do número de usuários nos instantes de saída ($\Pi^d = (\pi_0^d, \pi_1^d, \ldots)$), θ_{tot} o tempo total de um usuário no sistema e F_w sua função distribuição de probabilidade.

Multiplicando por n ambos os lados da equação e somando para valores de n variando de 1

ao infinito, segue:

$$\sum_{n=1}^{\infty} \pi_n^d = \sum_{n=1}^{\infty} \int_0^{\infty} P(n \text{ chegadas durante } \theta_{tot} | \theta_{tot} = t) dF_w(t), \qquad n \geq 0.$$

Sabendo que é válida a igualdade das distribuições, imersa nos instantes de saída e em tempo contínuo, pelo Teorema 2.1 (PASTA), o valor esperado do número no sistema em relação à Π^d será igual a L, pois:

$$\sum_{n=1}^{\infty} n\pi_n^d = \sum_{n=1}^{\infty} n\pi_n = L.$$

Então, admitindo ser possível a troca entre os sinais de soma e de integração, vem que:

$$\begin{aligned} L &= \sum_{n=1}^{\infty} n \int_0^{\infty} \frac{(\lambda t)^n e^{-\lambda t}}{n!} dF_w(t) = \int_0^{\infty} \lambda t e^{-\lambda t} \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} dF_w(t) \\ &= \int_0^{\infty} \lambda t e^{-\lambda t} e^{\lambda t} dF_w(t) = \lambda \int_0^{\infty} t dF_w(t) = \lambda E(\theta_{tot}) = \lambda W. \end{aligned}$$

Com isso, está demonstrada a equação (2.6).

Para (2.7) deve-se lembrar que, pode-se repartir o tempo no sistema entre o tempo na fila mais o de serviço, logo $W = E(\theta_q) + E(Z) = W_q + 1/\mu$, sendo θ_q o tempo na fila e Z o tempo de serviço. Também, $L_q = E(N_q) = L - (1 - \pi_0)$ onde N_q o número de usuários na fila.

Assim, ressaltando que $\pi_0 = 1 - \lambda/\mu$, prova-se que $L_q = \lambda W_q$.

□

2.3 Modelos clássicos de filas

O objetivo desta seção é apresentar alguns modelos de filas que serão úteis no próximo capítulo.

2.3.1 Fila M/M/s

Uma fila M/M/s apresenta chegadas Poisson e tempos de serviço exponenciais, além de s servidores em paralelo e sala de espera infinita. A disciplina de atendimento é FCFS.

Seja $N(t)$ a quantidade de usuários no sistema, no instante t, observada em estado estacionário. Pelas características das chegadas e serviços, trata-se de um processo de nascimento e morte cujas taxas são dadas por:

$$\lambda_k = \lambda, \qquad k \geq 0, \tag{2.8}$$

e

$$\mu_k = \begin{cases} k\mu, & 0 \leq k \leq s-1, \\ s\mu, & k \geq s. \end{cases} \tag{2.9}$$

As taxas de saída são diferenciadas para alguns estados devido ao fato de haver s servidores. Quando há menos de s usuários no sistema, ou seja, $k < s$, apenas k servidores estarão ocupados. Tal fato ocasiona uma saída média à taxa de $k\mu$. No caso de haver s ou mais usuários, todos os postos de atendimento permanecerão ocupados significando a taxa de $s\mu$ saídas.

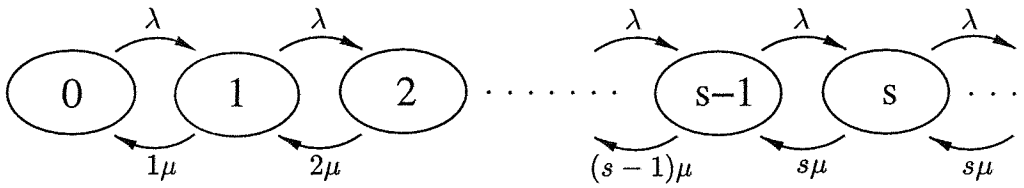

Figura 2.1: Esquema de transições em uma fila M/M/s.

Para encontrar a distribuição estacionária de $N(t)$ utilizam-se as taxas dadas pelas equações

(2.8) e (2.9) na equação genérica descrita por (2.4). Assim, chega-se a:

$$\pi_k = \begin{cases} \frac{\lambda^k}{k!\mu^k}\pi_0, & 1 \leq k \leq s-1, \\ \frac{\lambda^k}{s^{k-s}s!\mu^k}\pi_0, & k \geq s. \end{cases} \tag{2.10}$$

A fim de se encontrar a distribuição estacionária para este modelo, utiliza-se novamente o fato de que a soma das probabilidades é 1.

Deve-se, contudo, observar se as condições de existência da distribuição estacionária estão garantidas. Analogamente para este modelo, define-se por $\rho = \lambda/s\mu$ a intensidade de tráfego. A condição que possibilita os cálculos é $\rho < 1$, em outras palavras, a taxa total de serviço disponível deve ser maior que a taxa média de chegada. Assim,

$$\pi_0 = \left(\sum_{k=0}^{s-1} \frac{\lambda^k}{k!\mu^k} + \frac{\lambda^s}{s!\mu^s}\left(\frac{s\mu}{s\mu+\lambda}\right) \right)^{-1}, \qquad \rho < 1. \tag{2.11}$$

Iniciando o estudo das medidas de desempenho, determina-se o comprimento médio da fila (L_q).

$$\begin{aligned} L_q &= \sum_{n=s+1}^{\infty} (n-s)\pi_n = \sum_{n=s+1}^{\infty} (n-s)\frac{\lambda^n}{s^{n-s}s!\mu^n}\pi_0 \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0 \sum_{n=s+1}^{\infty} (n-s)\frac{\lambda^{n-s}}{s^{n-s}\mu^{n-s}} = \frac{\lambda^s}{s!\mu^s}\pi_0 \sum_{k=1}^{\infty} k\rho^k \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0\rho \sum_{k=1}^{\infty} k\rho^{k-1} = \frac{\lambda^s}{s!\mu^s}\pi_0 \sum_{k=1}^{\infty} \frac{\partial}{\partial\rho}\rho^k \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0\rho\frac{\partial}{\partial\rho}\sum_{k=1}^{\infty}\rho^k = \frac{\lambda^s}{s!\mu^s}\pi_0\rho\frac{\partial}{\partial\rho}\left(\frac{\rho}{1-\rho}\right) \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0\rho\left(\frac{1}{1-\rho}\right)^2, \end{aligned}$$

logo,

$$L_q = \left(\frac{(\lambda/\mu)^s\lambda\mu}{(s-1)!(s\mu-\lambda)^2} \right)\pi_0. \tag{2.12}$$

Dessa forma, conhecendo-se as fórmulas de Little (2.6) e (2.7), a partir de L_q chega-se às expressões do tempo médio na fila (W_q), o tempo médio no sistema (W) e do tamanho médio do sistema (L). Os resultados são os seguintes:

$$W_q = \frac{L_q}{\lambda} = \left(\frac{(\lambda/\mu)^s \mu}{(s-1)!(s\mu - \lambda)^2} \right) \pi_0, \tag{2.13}$$

$$W = \frac{1}{\mu} + W_q = \frac{1}{\mu} + \left(\frac{(\lambda/\mu)^s \mu}{(s-1)!(s\mu - \lambda)^2} \right) \pi_0, \tag{2.14}$$

$$L = \lambda W = \frac{\lambda}{\mu} + \left(\frac{(\lambda/\mu)^s \lambda\mu}{(s-1)!(s\mu - \lambda)^2} \right) \pi_0. \tag{2.15}$$

2.3.2 Fila M/M/s/r

Como ocorre na fila M/M/s, a M/M/s/r apresenta chegadas Poisson e serviços exponenciais com s servidores em paralelo e disciplina de atendimento FCFS. Porém, possui uma sala de espera finita com r vagas. Assim, no momento em que todas as $s + r$ posições de atendimento e de espera do sistema estiverem ocupadas, as novas chegadas serão perdidas.

Seguindo a linha desenvolvida para a fila M/M/s, considere o processo de nascimento e morte $N(t)$ indicando a quantidade de usuários no sistema no instante t. Suas taxas são dadas por:

$$\lambda_k = \begin{cases} \lambda, & 0 \leq k \leq s + r - 1, \\ 0, & k \geq s + r, \end{cases} \tag{2.16}$$

e

$$\mu_k = \begin{cases} k\mu, & 0 \leq k \leq s - 1, \\ s\mu, & s \leq k \leq s + r, \\ 0, & k > s + r. \end{cases} \tag{2.17}$$

Uma vez que os servidores estão distribuídos paralelamente, as taxas de saída seguem o mesmo padrão existente na fila M/M/s, porém, limitadas a $s + r$ usuários no sistema.

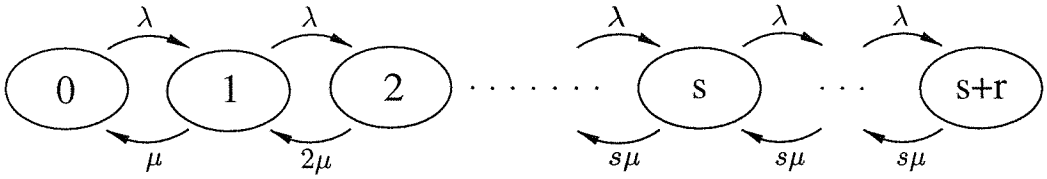

Figura 2.2: Esquema de transições em uma fila M/M/s/r.

Os passos para encontrar a distribuição estacionária são análogos aos adotados para a fila M/M/s. Logo,

$$\pi_k = \begin{cases} \frac{\lambda^k}{k!\mu^k}\pi_0, & 1 \leq k \leq s-1, \\ \frac{\lambda^k}{s^{k-s}s!\mu^k}\pi_0, & s \leq k \leq s+r. \end{cases} \tag{2.18}$$

Novamente, utiliza-se a propriedade de que a soma das probabilidades é 1 para se obter π_0. O espaço de estados é limitado ao conjunto $\{0, 1, 2, \ldots, s+r\}$, logo vem que:

$$\sum_{n=0}^{s+r} \pi_n = \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n}\pi_0 + \sum_{n=s}^{s+r} \frac{\lambda^n}{s^{n-s}s!\mu^n}\pi_0. \tag{2.19}$$

Considerando-se que $\rho = \lambda/s\mu$, chega-se a dois resultados diferentes para π_0, caso ρ seja igual ou diferente de 1. Por tratar-se de duas somas finitas não é necessário que ρ seja menor que 1 para que haja a convergência das séries. Portanto,

$$\pi_0 = \begin{cases} \left(\sum_{k=0}^{s-1} \frac{\lambda^k}{k!\mu^k} + \frac{\lambda^s}{s!\mu^s}(r+1) \right)^{-1}, & \rho = 1, \\ \left(\sum_{k=0}^{s-1} \frac{\lambda^k}{k!\mu^k} + \frac{\lambda^s}{s!\mu^s}\left(\frac{1-(\lambda/s\mu)^{r+1}}{1-(\lambda/s\mu)}\right) \right)^{-1}, & \rho \neq 1. \end{cases} \tag{2.20}$$

Agora, é possível prosseguir com o estudo das medidas de desempenho. Inicialmente, obtém-

se o número médio de usuários na fila. Para o caso em que $\rho \neq 1$, segue

$$\begin{aligned} L_q &= \sum_{n=s+1}^{s+r} (n-s)\pi_n = \sum_{n=s+1}^{s+r} (n-s)\frac{\lambda^n}{s^{n-s}s!\mu^n}\pi_0 \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0 \sum_{n=s+1}^{s+r} (n-s)\frac{\lambda^{n-s}}{s^{n-s}\mu^{n-s}} = \frac{\lambda^s}{s!\mu^s}\pi_0 \sum_{k=1}^{r} k\rho^k \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0\rho \sum_{k=1}^{r} k\rho^{k-1} = \frac{\lambda^s}{s!\mu^s}\pi_0\rho \sum_{k=1}^{r} \frac{\partial}{\partial\rho}\rho^k \\ &= \frac{\lambda^s}{s!\mu^s}\pi_0\rho\frac{\partial}{\partial\rho}\left(\sum_{k=1}^{r}\rho^k\right) = \frac{\lambda^s}{s!\mu^s}\pi_0\rho\frac{\partial}{\partial\rho}\left(\frac{\rho(1-\rho^r)}{1-\rho}\right), \end{aligned}$$

resultando em:

$$L_q = \frac{\lambda^{s+1}}{s!s\mu^{s+1}}\pi_0\left(\frac{1-\rho^r-r\rho^r(1-\rho)}{(1-\rho)^2}\right). \tag{2.21}$$

No caso de $\rho = 1$ deve-se recorrer a outro método para o cálculo de π_0 [Gross and Harris, 1998].

Para as demais medidas de desempenho é necessário determinar a taxa efetiva de entrada no sistema. Uma vez que o tamanho da fila é limitado, ocorre o fenômeno chamado de *overflow*. Isto significa que algumas chegadas podem ser perdidas por não encontrarem espaço ao chegar.

A taxa efetiva de chegada é dada por $\lambda' = \lambda(1-\pi_{s+r})$ pois a probabilidade do sistema permitir entradas ou, a probabilidade de ainda haver espaço no sistema, é $1-\pi_{s+r}$ (ver [Magalhães, 1996]). Assim, as demais medidas de desempenho podem ser obtidas a partir de (2.21) apenas substituindo λ por λ' nas fórmulas de Little. Dessa forma, seguem:

$$L = L_q + \frac{\lambda}{\mu}(1-\pi_{s+r}), \tag{2.22}$$

$$W = \frac{L}{\lambda(1-\pi_{s+r})}, \tag{2.23}$$

$$W_q = \frac{L_q}{\lambda(1-\pi_{s+r})}. \tag{2.24}$$

2.3.3 Fila M/G/1

Na fila M/G/1 a distribuição do serviço é geral, mantendo a independência entre sucessivos serviços e entre serviços e chegadas. Chegadas obedecem uma distribuição de Poisson de parâmetro λ enquanto os tempos de serviço Z seguem uma distribuição geral B. A disciplina de atendimento é FCFS.

Seja, como antes, $N(t)$ a quantidade de usuários no sistema no instante t. Percebe-se, agora, que $N(t)$ não é um processo de Markov já que o tempo para completar um serviço é indeterminado (não é possível garantir a propriedade de perda de memória, uma vez que o serviço segue uma distribuição geral B).

Uma abordagem para compreender o comportamento do sistema é estudar o *processo imerso nos instantes de saída.*

Sejam $T_0, T_1, T_2, \ldots$ os sucessivos instantes de fim de serviço ou, saídas da fila, e X_n o número de usuários na fila imediatamente após o evento T_n, ou seja, $X_n = N(T_n^+)$. Pode-se ver que o processo $(X,T) = \{(X_n, T_n); n \geq 0\}$ é um processo de renovação markoviano (ver [Magalhães, 1996]). Assim, calcula-se o núcleo de transição desse processo, representado pela matriz $Q(t)$, cujos elementos são:

$$Q_{ij}(t) = P(X_n = j, T_n - T_{n-1} \leq t | X_{n-1} = i), \quad t \geq 0, \quad i,j = 0,1,2,\ldots$$

O núcleo deve ser analisado em duas etapas:

- transição $0 \to j,\ j \geq 0$

  Considere o evento: (primeira chegada) + (j chegadas antes do fim de serviço) $\leq$ t.

  Assim,

$$Q_{0j}(t) = \int_0^t \int_0^{t-u} \lambda e^{-\lambda u} \frac{(\lambda y)^j e^{-\lambda y}}{j!} dB(y)du, \qquad j \geq 0.$$

- transição $i \to j,\ j \geq i-1,\ i > 0$

  Considere o evento: (tempo para $j - i + 1$ chegadas antes do fim de serviço)$\leq$ t.

Logo,

$$Q_{ij}(t) = \int_0^t \frac{(\lambda y)^{j-i+1} e^{-\lambda y}}{(j-i+1)!} dB(y), \qquad i > 0, \quad j \geq i-1.$$

Com o uso de funções auxiliares para a representação do núcleo:

$$g_n(t) = \int_0^t \frac{(\lambda y)^n e^{-\lambda y}}{n!} dB(y), \qquad n \geq 0,$$

e

$$h_n(t) = \int_0^t \lambda e^{-\lambda y} g_n(t-y) dy, \qquad n \geq 0.$$

Segue,

$$Q(t) = \begin{pmatrix} h_0(t) & h_1(t) & h_2(t) & \dots \\ g_0(t) & g_1(t) & g_2(t) & \dots \\ 0 & g_0(t) & g_1(t) & \dots \\ 0 & 0 & g_0(t) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Tomando-se o limite de $Q(t)$, quando t vai ao infinito, chega-se à matriz de transição da cadeia imersa:

$$P = \begin{pmatrix} h_0 & h_1 & h_2 & \dots \\ g_0 & g_1 & g_2 & \dots \\ 0 & g_0 & g_1 & \dots \\ 0 & 0 & g_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

com

$$g_n = \int_0^\infty \frac{(\lambda y)^n e^{-\lambda y}}{n!} dB(y), \qquad n \geq 0,$$

nota-se que h_n e g_n tornam-se iguais nessas condições.

Seja, como antes, π_n^d a distribuição estacionária do número de usuários nos instantes de saída e $\Pi^d = (\pi_0^d, \pi_1^d, \dots)$. Assim, pode-se obter a distribuição estacionária da cadeia imersa nos instantes de partida resolvendo-se as equações $(\Pi^d = \Pi^d P)$ e $(\Pi^d e = 1)$, segue:

$$\pi_i^d = \pi_0^d g_i + \sum_{j=1}^{i+1} \pi_j^d g_{i-j+1}, \quad i = 0, 1, 2, \dots, \tag{2.25}$$

$$\sum_{i=0}^{\infty} \pi_i^d = 1. \tag{2.26}$$

Para o cálculo de algumas medidas de desempenho, utiliza-se a *função geradora de probabilidade* da distribuição estacionária:

$$\Phi(z) = \sum_{i=0}^{\infty} \pi_i^d z^i, \qquad |z| \leq 1.$$

Outra função geradora, referente à probabilidade de i chegadas durante um serviço, também é utilizada para auxiliar os cálculos:

$$\Upsilon(z) = \sum_{i=0}^{\infty} g_i z^i, \qquad |z| \leq 1.$$

Multiplicando-se a expressão (2.25) por z^i e somando para $i \geq 0$ segue:

$$\sum_{i=0}^{\infty} \pi_i^d z_i = \sum_{i=0}^{\infty} \pi_0^d g_i z_i + \sum_{i=0}^{\infty} z^i \sum_{j=1}^{i+1} \pi_j^d g_{i-j+1},$$

assim, usando as funções geradoras,segue:

$$\begin{aligned}
\Phi(z) &= \pi_0^d \Upsilon(z) + \sum_{j=1}^{\infty} \sum_{i=j-1}^{\infty} \pi_j^d g_{i-j+1} z^i \\
&= \pi_0^d \Upsilon(z) + \frac{1}{z} \sum_{j=1}^{\infty} \pi_j^d z^j \sum_{i=j-1}^{\infty} g_{i-j+1} z^{i-j+1} \\
&= \pi_0^d \Upsilon(z) + \frac{1}{z} \big(\Phi(z) - \pi_0^d\big) \Upsilon(z),
\end{aligned}$$

então,

$$\Phi(z) = \frac{\pi_0^d \Upsilon(z)(1-z)}{\Upsilon(z) - z}. \tag{2.27}$$

Para obter π_0^d, nota-se que $\Phi(1) = 1$ e $\Upsilon(1) = 1$. Logo, aplicando a regra de L'Hospital em (2.27) para $z = 1$, segue:

$$1 = \left. \frac{\pi_0^d [\Upsilon'(z)(1-z) - \Upsilon(z)]}{\Upsilon'(z) - 1} \right|_{z=1}, \tag{2.28}$$

sendo, $\Upsilon'(z) = \sum_{i=0}^{\infty} i g_i z^{i-1}$, resulta:

$$\begin{aligned} \Upsilon'(1) &= \sum_{i=0}^{\infty} \int_0^{\infty} \frac{(\lambda y)^i e^{-\lambda y}}{i!} dB(y) = \int_0^{\infty} \sum_{i=0}^{\infty} \frac{(\lambda y)^i e^{-\lambda y}}{i!} dB(y) \\ &= \int_0^{\infty} \lambda y dB(y) = \lambda \int_0^{\infty} y dB(y) = \lambda E(Z). \end{aligned}$$

Definindo a intensidade de tráfego por $\rho = \lambda E(Z)$ e substituindo o valor de $\Upsilon'(1)$ em (2.28), segue que:

$$\pi_0^d = 1 - \rho,$$

o que caracteriza a existência da distribuição estacionária se, e somente se, $\rho < 1$. Assim,

$$\Phi(z) = \frac{(1 - \lambda E(Z))\Upsilon(z)(1-z)}{\Upsilon(z) - z}. \tag{2.29}$$

As medidas de desempenho podem ser obtidas através dessa equação e das fórmulas de Little. São dadas por:

$$L = \rho + \frac{\rho^2 + \lambda^2 \sigma_Z^2}{2(1-\rho)}, \tag{2.30}$$

$$W = E(Z) + \frac{\rho E(Z) + \lambda \sigma_Z^2}{2(1-\rho)}, \tag{2.31}$$

$$W_q = \frac{\rho E(Z) + \lambda \sigma_Z^2}{2(1-\rho)}, \tag{2.32}$$

$$L_q = \frac{\rho^2 + \lambda^2 \sigma_Z^2}{2(1-\rho)}, \tag{2.33}$$

onde σ_Z^2 é a variância do tempo de serviço Z.

2.3.4 Fila M/G/∞

Em uma fila com infinitos servidores não há interação entre os usuários pois, uma vez que todas as chegadas são prontamente atendidas, os tempos de serviço e entre chegadas não influenciam os diferentes atendimentos.

Aqui será apresentado um resultado para a fila M/G/∞ referente ao número de usuários no sistema. Estudos mais aprofundados podem ser encontrados em [Takács, 1962], [Ross, 1982] e [Whitt, 1993] e demais referências lá citadas.

Define-se os instantes de chegada por $\tau_1, \tau_2, \ldots, \tau_n, \ldots$ de forma que os intervalos entre chegadas, $\eta_n = \tau_n - \tau_{n-1}$, sejam independentes e identicamente distribuídos segundo a distribuição:

$$P(\eta_n = \tau_n - \tau_{n-1} \leq x) = F(x), \qquad n \in \mathbf{N}.$$

Uma vez que há infinitos servidores, todos os usuários entram em serviço imediatamente após as chegadas, ou seja, não há tempo de espera.

Seja χ_n o tempo de serviço do n-ésimo cliente. Supõem-se que os tempos de serviço de diferentes usuários sejam independentes entre si e igualmente distribuídos, assim, $\chi_1, \chi_2, \ldots, \chi_n, \ldots$ são independentes e identicamente distribuídos de acordo com a função de distribuição:

$$P(\chi_n \leq x) = H(x), \qquad n \in \mathbf{N},$$

e, além disso, são independentes das chegadas, $\{\tau_n\}$.

Seja como antes, $N(t)$, a quantidade de usuários no sistema no instante t. Aqui, $N(t)$ também representa os usuários sendo atendidos. Em uma fila $M/G/\infty$ as chegadas são Poisson. Assim,

sejam $P(N(t) = k) = P_k(t)$ e $\alpha = \int_0^\infty x dH(x)$ o tempo médio de serviço. No próximo teorema obtém-se o valor de $P_k(t)$.

Teorema 2.3 *Se* $N(0) = 0$, *então,*

$$P_k(t) = \frac{\left[\lambda \int_0^t \big((1 - H(x)\big)dx\right]^k}{k!} e^{-\lambda \int_0^t (1-H(x))dx}, \qquad k = (0, 1, 2, \dots), \tag{2.34}$$

e, se $\alpha < \infty$ *então* $\lim_{t\to\infty} P_k(t) = P_k^*$ *existe e*

$$P_k^* = e^{-\lambda\alpha}\frac{(\lambda\alpha)^k}{k!}. \tag{2.35}$$

Prova:

No intervalo $(0, t]$ pode haver n chegadas ($n = 0, 1, 2, \dots$). Assim, se $n \geq k$, para que o evento $N(t) = k$ ocorra deve haver k serviços ativos no instante t. A probabilidade de um usuário que chegou no instante $x \in (0, t]$ ainda estar em serviço no instante t é $1 - H(t - x)$.

Um resultado útil (ver Teorema 2.3 em [Ross, 1970]) diz que, dado que houve n chegadas em $(0, t]$, os instantes das n chegadas $t_1, t_2, \dots, t_n$ têm a mesma distribuição que os estatísticos de ordem de n variáveis aleatórias independentes e uniformemente distribuidas em $(0, t]$. E assim, dado que houve n chegadas em $(0, t]$, a probabilidade de uma chegada arbitrária continuar presente no sistema no instante t é dada por:

$$p_t = \int_0^t \big(1 - H(t - x)\big) f(x)dx = \int_0^t \big(1 - H(t - x)\big)\frac{1}{t}dx = \frac{1}{t}\int_0^t \big(1 - H(t)\big)dx,$$

onde f é a função densidade de probabilidade de uma distribuição Uniforme$(0, t)$.

Logo, a distribuição condicional de $N(t)$, dado que houve n chegadas em $(0, t]$, segue uma distribuição Binomial com parâmetros n e p_t.

Seja C_n o evento correspondente a n chegadas no intervalo $(0, t]$. Então, pela regra da

probabilidade total, pode-se escrever que:

$$\begin{aligned}
P_k(t) &= \sum_{n=k}^{\infty} P\left[(N(t)=k)\cap C_n\right] \\
&= \sum_{n=k}^{\infty} P\left[(N(t)=k)|C_n\right] P(C_n) \\
&= \sum_{n=k}^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!}\binom{n}{k}\left[\frac{1}{t}\int_0^t (1-H(x))dx\right]^k \left[\frac{1}{t}\int_0^t H(x)dx\right]^{n-k} \\
&= \frac{1}{k!}\left[\int_0^t (1-H(x))dx\right]^k e^{-\lambda t}\sum_{n=k}^{\infty}\frac{\lambda^n}{(n-k)!}\left[\int_0^t H(x)dx\right]^{n-k},
\end{aligned}$$

após uma mudança de variáveis no somatório, resulta:

$$\begin{aligned}
P_k(t) &= \frac{1}{k!}\left[\int_0^t \big(1-H(x)\big)dx\right]^k e^{-\lambda t}\sum_{i=0}^{\infty}\frac{\lambda^{i+k}}{i!}\left[\int_0^t H(x)dx\right]^i \\
&= \frac{1}{k!}\left[\lambda\int_0^t \big(1-H(x)\big)dx\right]^k e^{-\lambda t}\sum_{i=0}^{\infty}\frac{\left[\lambda\int_0^t H(x)dx\right]^i}{i!} \\
&= \frac{1}{k!}\left[\lambda\int_0^t \big(1-H(x)\big)dx\right]^k e^{-\lambda t}e^{\lambda\left[\int_0^t H(x)dx\right]} \\
&= \frac{1}{k!}\left[\lambda\int_0^t \big(1-H(x)\big)dx\right]^k e^{-\lambda\int_0^t dx}e^{\lambda\left[\int_0^t H(x)dx\right]} \\
&= \frac{\left[\lambda\int_0^t \big(1-H(x)\big)dx\right]^k}{k!}e^{-\lambda\int_0^t \left(1-H(x)\right)dx}.
\end{aligned}$$

Pela proposição 4.6 em [Magalhães, 2004], segue:

$$\lim_{t\to\infty}\int_0^t \big(1-H(x)\big)dx = \alpha,$$

então,

$$\lim_{t\to\infty} P_k(t) = P_k^* = \lim_{t\to\infty}\left\{\frac{\left(\lambda\int_0^t \big(1-H(x)\big)dx\right)^k}{k!}e^{-\lambda\int_0^t \left(1-H(x)\right)dx}\right\} = e^{-\lambda\alpha}\frac{(\lambda\alpha)^k}{k!}.$$

□

Capítulo 3

Modelos de filas com desistências

3.1 Introdução

Neste Capítulo serão apresentados dois modelos para filas com desistências, um sem e outro com informação ao usuário. Ambos os casos estão baseados em um processo de nascimento e morte e serão consideradas as três características já mencionadas: *recusa*, *abandono* e *impedimento*. A apresentação segue as linhas desenvolvidas no trabalho de [Whitt, 1999b].

3.2 Modelo sem informação ao usuário

No modelo apresentado aqui não há qualquer tipo de informação sobre o estado do sistema aos usuários que chegam. Suas características são: as chegadas seguem um processo de Poisson com taxa λ, há s servidores e uma fila de tamanho máximo r. Assim, o espaço de estados desse sistema está restrito ao conjunto $\{0, 1, 2, \ldots, s + r\}$. A disciplina de atendimento é FCFS. Os tempos de serviço são exponenciais de taxa μ, independentes e identicamente distribuídos (iid).

Quando ocorre uma chegada, caso a capacidade total do sistema esteja sendo utilizada ($s+r$ usuários), o novo cliente não poderá entrar no sistema e será rejeitado (*impedimento*). Se existirem servidores disponíveis, o cliente será atendido imediatamente por um dos servidores vagos. No caso de todos estarem ocupados, mas ainda houver espaço na sala de espera, restam

duas alternativas: permanecer em fila e aguardar pelo atendimento ou sair do sistema (*recusa*). Uma vez que o usuário juntou-se à fila ainda existe a possibilidade de que este abandone o sistema se o tempo de espera exceder seu limite de aceitação. Se antes desse prazo nenhum servidor estiver livre então ocorrerá o *abandono* por parte desse usuário. Ao tempo máximo que cada usuário aceita esperar dá-se o nome de *tolerância.*

A *tolerância* pode ser nula com probabilidade β. Dessa maneira, a *recusa* ocorre com probabilidade β, representando os usuários que não estão dispostos a esperar em fila. Caso a *tolerância* seja positiva, será representada por T e assume-se que sua distribuição esteja de acordo com uma Exponencial de média $1/\alpha$.

Assim, temos as taxas de nascimento e morte para esse modelo:

$$\lambda_k = \begin{cases} \lambda, & 0 \leq k \leq s-1, \\ \lambda(1-\beta), & s \leq k \leq s+r-1, \\ 0, & k \geq s+r, \end{cases} \tag{3.1}$$

e

$$\mu_k = \begin{cases} k\mu, & 0 \leq k \leq s, \\ s\mu + (k-s)\alpha, & s+1 \leq k \leq s+r, \\ 0, & k > s+r. \end{cases} \tag{3.2}$$

Nota-se que a ocorrência de *recusa* e *abandono*, na situação em que todos os servidores estão ocupados, acrescentam às taxas termos diferentes daqueles encontrados em modelos $M/M/s/r$.

Conhecendo-se as propriedades da distribuição exponencial, percebe-se que quando há k usuários no sistema o próximo evento terá distribuição exponencial com parâmetro $(\lambda_k + \mu_k)$. Da mesma forma, a probabilidade de que o próximo evento seja uma chegada é $\frac{\lambda_k}{\lambda_k+\mu_k}$.

O modelo apresentado nesta seção será, de agora em diante, denominado de Modelo 1.

3.3 Modelo com informação ao usuário

Aqui será apresentado um modelo mais amplo que o visto na seção anterior. Supõem-se que, no momento da chegada, o usuário é avisado sobre o estado do sistema. Esta informação representa um aspecto importante para o usuário, uma vez que auxilia na decisão de aguardar por atendimento. Se o usuário encontrar todos os servidores ocupados, sabendo o tamanho da fila, é possível estimar o tempo de espera até ser atendido. Assim, torna-se mais clara a opção pela *recusa* imediata ou a permanência na fila. Dessa forma, seria natural esperar que, neste modelo, parte ou a totalidade da intensidade de *abandono* seja incorporada pela *recusa*.

Inicialmente, será apresentado um modelo genérico com o intuito de caracterizar qual a reação dos usuários à informação adicional acerca do estado do sistema. Nesse modelo o *abandono* ainda será considerado para permitir maior flexibilidade e incorporar a variabilidade inerente aos tempos de atendimento. Mais adiante será definido o Modelo 2 como um caso particular em que o *abandono* é completamente substituído pela *recusa* na chegada. Um dos objetivos deste trabalho é comparar o Modelo 1, visto na seção anterior, com o Modelo 2.

Assim como foi feito anteriormente, as chegadas seguem um processo de Poisson com taxa λ, os tempos de serviço têm distribuição exponencial de taxa μ nos s servidores, a sala de espera possui tamanho máximo igual a r e a disciplina de atendimento é FCFS. Da mesma forma que na seção anterior, o cliente possui uma probabilidade β de não aceitar a fila. Além disso, conhecendo o estado do sistema, no momento da chegada, o usuário também pode decidir por recusá-lo.

No instante da chegada, após conhecer o estado do sistema o usuário deve decidir por sair imediatamente ou permanecer em fila junto aos demais. O critério que irá adotar sugere uma comparação entre a *tolerância* e o tempo estimado para receber atendimento, ou seja, se o tempo para que um servidor esteja disponível para essa chegada for menor que sua *tolerância* então a decisão será por esperar na fila. Caso contrário, ocorrerá a *recusa*.

Seja S_k o tempo necessário para que uma chegada, encontrando $s+k$ usuários à sua frente no sistema, inicie o atendimento em um dos s servidores (note que S_k corresponde à soma de $k+1$ exponenciais de média $\frac{1}{s\mu}$). A probabilidade de uma nova chegada juntar-se à fila, sendo

que há $s+k$ usuários no sistema (desconsiderando a própria chegada), será dada por:

$$q_k \equiv P(T > S_k), \qquad 0 \leq k \leq r-1. \tag{3.3}$$

A fim de se calcular explicitamente a probabilidade q_k, assume-se que os usuários serão informados apenas sobre o estado do sistema na chegada. Assim, pode-se considerar que (3.3) será parte da *recusa* desses usuários.

Proposição 3.1 *Sejam S_k e T variáveis aleatórias independentes como definidas acima. Então, a probabilidade de uma chegada juntar-se à fila é:*

$$q_k = E\left(e^{-\alpha S_k}\right) = \left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1}, \qquad 0 \leq k \leq r-1. \tag{3.4}$$

Prova:

Seja g_k a densidade de probabilidade de S_k, então:

$$\begin{aligned} q_k \equiv P(T > S_k) &= \int_0^\infty P(T > S_k | S_k = s) g_k(s) ds = \int_0^\infty P(T > s | S_k = s) g_k(s) ds \\ &= \int_0^\infty P(T > s) g_k(s) ds = \int_0^\infty e^{-s\alpha} g_k(s) ds \\ &= E\left(e^{-\alpha S_k}\right) = \left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1}. \end{aligned}$$

A última igualdade acima resulta do fato de $E\left(e^{-\alpha S_k}\right)$ representar a transformada de Laplace-Stieltjes de S_k. Logo, consultando uma tabela de transformadas (por exemplo, [Kleinrock, 1975]), chega-se ao último termo da equação. □

Caso o gerenciador do sistema informe o tempo esperado quando há $s+k$ usuários no sistema, a equação (3.4) torna-se:

$$q_k^* \equiv P(T > E(S_k)) = e^{-\frac{\alpha(k+1)}{s\mu}}, \qquad k \geq 0. \tag{3.5}$$

Para verificar a validade de (3.5), observe que S_k possui distribuição Gama por ser a soma

de Exponenciais independentes, ou seja, $S_k \sim Gama(k+1, s\mu)$ (ver [Hoel et al., 1971]). Assim,

$$P(T > E(S_k)) = P\left(T > \frac{k+1}{s\mu}\right) = e^{-\frac{\alpha(k+1)}{s\mu}}.$$

Agora, assim como na Seção 3.2, utiliza-se o processo de nascimento e morte para a situação em que o usuário recebe informações acerca do estado do sistema.

Primeiramente, seja δ_j' a taxa de *abandono* do j-ésimo usuário na fila. Assim, a taxa total de *abandono*, quando há $s+k$ usuários no sistema é:

$$\delta_k = \sum_{j=1}^{k} \delta_j' \tag{3.6}$$

Então, seguem as taxas de nascimento e morte para um modelo genérico:

$$\lambda_k = \begin{cases} \lambda, & 0 \leq k \leq s-1, \\ \lambda(1-\beta)q_{k-s} & s \leq k \leq s+r-1, \\ 0, & k \geq s+r, \end{cases} \tag{3.7}$$

e

$$\mu_k = \begin{cases} k\mu, & 1 \leq k \leq s, \\ s\mu + \delta_{k-s}, & s+1 \leq k \leq s+r, \\ 0, & k > s+r. \end{cases} \tag{3.8}$$

Uma vez que este modelo contempla desistências dos usuários, sua distribuição estacionária não é a mesma de uma fila M/M/s/r, como foi visto na Seção 2.3.2. A seguir, será indicado como calcular, numericamente, a distribuição estacionária π_n do processo de nascimento e morte no modelo genérico. Por conveniência, a distribuição estacionária será obtida recursivamente a partir do estado s, pois, é esperado que as probabilidades se concentrem em torno desse ponto. Um algoritmo para obter essas probabilidades é descrito em seguida.

Sejam $x_i, 0 \leq i \leq s+r$, variáveis auxiliares e seja fixado $x_s = 1$,

$$x_{s+j+1} = \frac{\lambda_{s+j} x_{s+j}}{\mu_{s+j+1}} = \frac{\lambda(1-\beta) q_j x_{s+j}}{s\mu + \delta_{j+1}}, \quad 0 \leq j \leq r-1 \tag{3.9}$$

e

$$x_{j-1} = \frac{\mu_j x_j}{\lambda_{j-1}} = \frac{j\mu x_j}{\lambda}, \quad 1 \leq j \leq s. \tag{3.10}$$

Então, seja

$$y = \sum_{j=0}^{s+r} x_j, \tag{3.11}$$

logo

$$\pi_k = \frac{x_k}{y}, \quad 0 \leq k \leq s+r. \tag{3.12}$$

Agora, serão calculados a probabilidade para uma chegada completar o serviço, seus momentos e a distribuição do tempo restante, dado que o serviço foi completado. Da mesma forma, serão estudados o abandono, os momentos e a distribuição do tempo para abandono, dado que houve o abandono.

Inicialmente, observa-se os estados vistos pelas chegadas para obter as medidas em tempo contínuo. O que permite esse enfoque é o fato das chegadas constituirem um processo de Poisson pois, conforme estudado na Seção 2.2.1, o resultado *PASTA* mostra que a distribuição da quantidade de usuários no sistema, nos instantes de chegada, é igual à distribuição em tempo contínuo.

Após a chegada de um determinado usuário, que observa $s+k$ usuários no sistema, pode-se considerar o sistema como um processo de morte puro (sem nascimentos) que ignora as chegadas posteriores, iniciando no estado $s+k+1$. Um resultado importante desse tipo de processo diz que os tempos entre sucessivas mortes distribuem-se exponencialmente com parâmetros dependentes do estado. Dessa forma, estuda-se o tempo restante na fila desse usuário específico.

Sejam γ_k a probabilidade de que o k-ésimo usuário na fila abandone o sistema no próximo evento de saída e m_k o tempo médio até o próximo evento de saída, considerando apenas os $s+k$ usuários no sistema.

Então, sabendo que tanto os serviços quanto os *abandonos* são exponenciais, seguem:

$$\gamma_k = \frac{\delta'}{s\mu + \delta_k} \quad \text{e} \quad m_k = \frac{1}{s\mu + \delta_k}. \tag{3.13}$$

Assim, a probabilidade de que o usuário na posição $s + k$ venha a ser atendido corresponde à probabilidade de que não abandone em nenhum momento, ou seja, que ele não abandone em qualquer uma das k posições que irá ocupar na fila até que chegue a um servidor,

$$\Gamma_k = (1 - \gamma_k)(1 - \gamma_{k-1}) \cdots (1 - \gamma_1), \tag{3.14}$$

em que $(1 - \gamma_j)$ é a probabilidade de que esse usuário, quando na j-ésima posição da fila, não abandone o sistema.

Para obter a probabilidade de um usuário completar o serviço é necessário observar as condições do sistema desde sua chegada. Caso haja algum servidor livre no momento da chegada ele será atendido imediatamente. Entretanto, se houver fila de k usuários à sua frente, deve-se considerar a probabilidade dele não recusar, $(1 - \beta)q_k$, e a probabilidade de não abandonar durante toda sua permanência na fila, Γ_{k+1}.

Assim, sendo S o evento que corresponde a uma chegada vir a completar o serviço, temos:

$$P(S) = \left(\sum_{k=0}^{s-1} \pi_k\right) + \sum_{k=0}^{r-1} \pi_{s+k}(1 - \beta)q_k\Gamma_{k+1}. \tag{3.15}$$

Seja C o tempo que um usuário leva desde a chegada até completar o serviço. Considera-se $C = 0$ se o serviço não for completado. Então, sendo N o número de usuários no sistema,

$$E(C) = E\big(E(C|N)\big) = \left(\sum_{k=0}^{s-1} \pi_k\right)\frac{1}{\mu} + \sum_{k=0}^{r-1} \pi_{s+k}(1 - \beta)q_k\Gamma_{k+1}\left(\frac{1}{\mu} + \sum_{j=1}^{k+1} m_j\right) \tag{3.16}$$

e

$$E(C^2) = E\big(E(C^2|N)\big) = \left(\sum_{k=0}^{s-1} \pi_k\right)\frac{2}{\mu^2} + \sum_{k=0}^{r-1} \pi_{s+k}(1 - \beta)q_k\Gamma_{k+1}\big(V_{k+1} + M_{k+1}^2\big) \tag{3.17}$$

em que

$$V_{k+1} = \frac{1}{\mu^2} + \sum_{j=1}^{k+1} m_j^2 \tag{3.18}$$

e

$$M_{k+1} = \frac{1}{\mu} + \sum_{j=1}^{k+1} m_j. \tag{3.19}$$

Nas duas parcelas de (3.16) encontra-se o termo $1/\mu$ que representa o tempo médio que o novo usuário permanecerá no servidor quando atingir a condição de atendimento.

O primeiro e segundo momentos do tempo condicional para completar o serviço, dado que o serviço foi concluído, são

$$E(C|S) = \frac{EC}{P(S)} \quad e \quad E(C^2|S) = \frac{EC^2}{P(S)}, \tag{3.20}$$

pois, sendo $\bar{S}$ o evento complementar de S, ou seja, a chegada não completou o serviço, temos,

$$\begin{aligned} E(C) &= E(C|S)P(S) + E(C|\bar{S})P(\bar{S}) \\ &= E(C|S)P(S), \end{aligned}$$

pois convencionou-se que $C = 0$ quando o serviço não é completado. Logo,

$$E(C|S) = \frac{E(C)}{P(S)},$$

e cálculos análogos resultam no segundo momento apresentado em (3.20).

Paralelamente às equações (3.20), seguem a variância condicional:

$$Var(C|S) = E(C^2|S) - (E(C|S))^2 \tag{3.21}$$

e o desvio padrão condicional:

$$SD(C|S) = \sqrt{Var(C|S)}. \tag{3.22}$$

Seja $\hat{c} \equiv Ee^{-zC}$ a transformada de Laplace-Stieltjes de C, segue

$$\hat{c}(z) = \left(\sum_{k=0}^{s-1} \pi_k\right)\left(\frac{\mu}{\mu+z}\right) + \sum_{k=0}^{r-1} \pi_{s+k}(1-\beta)q_k\Gamma_{k+1}\hat{d}_{k+1}(z). \tag{3.23}$$

em que,

$$\hat{d}_{k+1}(z) = \left(\frac{\mu}{\mu+z}\right)\prod_{j=1}^{k+1}\left(\frac{m_j^{-1}}{m_j^{-1}+z}\right). \tag{3.24}$$

Assim, pode-se calcular $P(C > t)$ invertendo sua transformada $(1-\hat{c}(z))/z$. Logo, segue a distribuição condicional:

$$P(C > t|S) = \frac{P(C > t)}{P(S)}. \tag{3.25}$$

Agora, será estudada a distribuição de *abandono* e do tempo para *abandono*, dado que o cliente abandone.

Define-se R como o evento em que uma chegada venha a abandonar o sistema. A probabilidade de um usuário abandonar deve considerar que não houve a recusa no momento da chegada. Assim,

$$P(R) = \sum_{k=0}^{r-1} \pi_{s+k}(1-\beta)q_k(1-\Gamma_{k+1}). \tag{3.26}$$

Sejam A o tempo que uma chegada leva para abandonar (assume-se que $A = 0$ se não houver *abandono*) e A_k o tempo para abandono de um usuário que começou na *k-ésima* posição da fila. Condicionando em N_f, o número de usuários na fila, segue

$$E(A) = E\big(E(A|N_f)\big) = \sum_{k=0}^{r-1} \pi_{s+k}(1-\beta)q_k E(A_{k+1}) \tag{3.27}$$

e

$$E(A^2) = E\big(E(A^2|N_f)\big) = \sum_{k=0}^{r-1} \pi_{s+k}(1-\beta)q_k E(A_{k+1}^2). \tag{3.28}$$

Para obter $E(A_k)$ e $E(A_k^2)$, leva-se em conta a probabilidade desse usuário abandonar em cada posição que ele venha a ocupar na fila e o tempo decorrido até que chegue nessa posição,

assim,

$$\begin{aligned} E(A_k) &= \gamma_k m_k + (1-\gamma_k)\gamma_{k-1}(m_k + m_{k-1}) + \\ &+ (1-\gamma_k)(1-\gamma_{k-1})\gamma_{k-2}(m_k + m_{k-1} + m_{k-2}) + \\ &\vdots \\ &+ (1-\gamma_k)\ldots(1-\gamma_2)\gamma_1(m_k + \ldots + m_1) \end{aligned} \tag{3.29}$$

e

$$\begin{aligned} E(A_k^2) &= \gamma_k m_k^2 + \\ &+ (1-\gamma_k)\gamma_{k-1}[m_k^2 + m_{k-1}^2 + (m_k + m_{k-1})^2] + \\ &\vdots \\ &+ (1-\gamma_k)(1-\gamma_{k-1})\ldots(1-\gamma_2)\gamma_1 \times \\ &\times [m_k^2 + \ldots + m_1^2 + (m_k + \ldots + m_1)^2]. \end{aligned} \tag{3.30}$$

Os momentos condicionais do tempo para abandonar dado que houve *abandono* são:

$$E(A|R) = \frac{E(A)}{P(R)} \qquad \text{e} \qquad E(A^2|R) = \frac{E(A^2)}{P(R)}. \tag{3.31}$$

Consequentemente, para a variância condicional,

$$Var(A|R) = E(A^2|R) - (E(A|R))^2 \tag{3.32}$$

e para o desvio padrão condicional,

$$SD(A|R) = \sqrt{Var(A|R)}. \tag{3.33}$$

Seja $\hat{a} \equiv Ee^{-zA}$ a transformada de Laplace-Stieltjes de A, da mesma forma que (3.23),

$$\hat{a}(z) = \sum_{k=0}^{r-1} \pi_{s+k}(1-\beta)q_k \hat{a}_{k+1}(z). \tag{3.34}$$

em que,

$$\hat{a}_k(z) = \left(\frac{m_k^{-1}}{m_k^{-1}+z}\right)\sum_{j=0}^{k-1}\gamma_{k-j}\prod_{l=1}^{j}\left[(1-\gamma_{k-l+1})\left(\frac{m_k^{-1}}{m_k^{-1}+z}\right)\right]. \tag{3.35}$$

Analogamente ao desenvolvido em (3.25), pode-se calcular $P(A > t)$ invertendo a transformada $(1-\hat{a}(z))/z$. Logo, é possível encontrar

$$P(A > t|R) = \frac{P(A > t)}{P(R)}. \tag{3.36}$$

Portanto, sendo π_{s+r} a probabilidade de *impedimento*, segue a probabilidade de *recusa*:

$$P(recusa) = 1 - P(S) - P(R) - \pi_{s+r}. \tag{3.37}$$

A partir do modelo genérico, define-se o Modelo 2 como o caso especial de (3.7) e (3.8) quando $\delta_k = 0$. Este modelo considera que o *abandono* é totalmente incorporado pela *recusa* na chegada.

Seguem as taxas para o processo de nascimento e morte do Modelo 2,

$$\lambda_k = \begin{cases} \lambda, & 0 \leq k \leq s-1, \\ \lambda(1-\beta)q_{k-s} & s \leq k \leq s+r-1, \\ 0, & k \geq s+r, \end{cases} \tag{3.38}$$

e

$$\mu_k = \begin{cases} k\mu, & 1 \leq k \leq s, \\ s\mu, & s+1 \leq k \leq s+r, \\ 0, & k > s+r. \end{cases} \tag{3.39}$$

Nas Equações (3.38) é possível observar que o *abandono* não é considerado e o parâmetro de *tolerância*, α, é absorvido pelo termo q_{k-s}.

A partir do modelo genérico percebe-se, também, que o Modelo 1 é um caso particular de

(3.7) e (3.8) quando $q_{k-s} = 1$ e $\delta_{k-s} = (k-s)\alpha$.

Algumas simulações envolvendo os Modelos 1 e 2 serão apresentadas no Capítulo 4. Também podem ser encontradas discussões sobre comparações estocásticas desses modelos no trabalho de [Whitt, 1999b], mas que não serão abordadas nesta dissertação.

Capítulo 4

Simulações numéricas

A fim de ilustrar os conceitos apresentados no Capítulo 3 e discutir possíveis aperfeiçoamentos nos modelos, foram elaboradas simulações numéricas dos sistemas apresentados.

Neste capítulo, primeiro será apresentada a implementação do modelo de simulação. A seguir, veremos como foi feita a validação desse modelo e, por último, serão exibidos os resultados obtidos através das simulações.

4.1 Descrição do modelo de simulação

As simulações foram implementadas com o auxílio do *software Arena*, versão 7.0. Tal programa foi escolhido por facilitar a construção de estruturas que representam modelos de filas. Ele trabalha com conceitos gráficos para compor cada componente de um sistema e, assim, torna-se possível acompanhar visualmente as características do modelo e também do fluxo de usuários.

Como usual em *softwares* de simulação, o *Arena* mantém constante a semente do gerador de números aleatórios e, portanto, os resultados sempre são idênticos. Para contornar esse detalhe, utilizou-se de replicações de simulações, ou seja, foram executadas seguidamente algumas simulações do modelo. Dessa forma, ao término de cada execução o último número aleatório gerado é utilizado como semente da simulação seguinte. No estudo apresentado foram definidas, empiricamente, o uso de 10 replicações para cada etapa das simulações. Para uma quantidade

maior de replicações não houve alterações significativas nas medidas de desempenho do sistema.

Outro recurso utilizado foi o uso de períodos de *warm up*, ou seja, a partir do sistema vazio espera-se um determinado intervalo de tempo até que o sistema passe a operar em regime estacionário. Desse momento em diante inicia-se a contagem das estatísticas. Nas simulações apresentadas foi utilizado um período de 5 horas, avaliado como suficiente para buscar garantir que as medidas fossem tomadas com o sistema em estado estacionário.

O término de cada simulação foi definido de acordo com a carga a que o sistema foi exigido. Quando a taxa média de chegadas foi de 4 por minuto adotou-se um limite de 5.000 chegadas para concluir a simulação. Para 20 chegadas por minuto o limite foi 20.000 e, para 40 chegadas por minuto, utilizou-se 50.000 chegadas como limite em cada simulação.

As medidas de performance foram geradas automaticamente pelo programa durante as simulações. Como já mencionado, os valores referem-se apenas às simulações após os intervalos de *warm up* pois, nesses períodos, as estatísticas coletadas são desconsideradas.

O objetivo foi implementar simulações dos modelos propostos por [Whitt, 1999b]. Dessa forma, obedeceram-se as definições discutidas nas Seções 3.2 e 3.3. Ainda de acordo com [Whitt, 1999b], é conveniente utilizar uma *n-upla* de parâmetros comuns $(\lambda, \mu, \alpha, \beta, s, r)$ para comparar os modelos.

Assim, os modelos foram construídos com chegadas seguindo uma distribuição de Poisson com taxa λ chegadas por minuto, s servidores com serviço Exponencial de média $1/\mu$ minutos, sala de espera de tamanho r, *tolerância* de espera Exponencial de média α^{-1} e probabilidade de *recusa* igual a β. Os parâmetros foram alterados para cada conjunto de simulações (ver Figura 4.1), conforme a necessidade de cada estudo.

Para definir o *abandono* utilizou-se, para o Modelo 1, um processo paralelo de controle dos tempos de permanência dos usuários no sistema. No instante de chegada o usuário era marcado com sua *tolerância*, ou seja, o tempo máximo que poderia suportar até que fosse atendido. Assim, a cada segundo os tempos de permanência dos usuários na fila eram verificados. Se algum deles ultrapassasse a *tolerância*, era automaticamente retirado do sistema, o que se caracterizava como *abandono*.

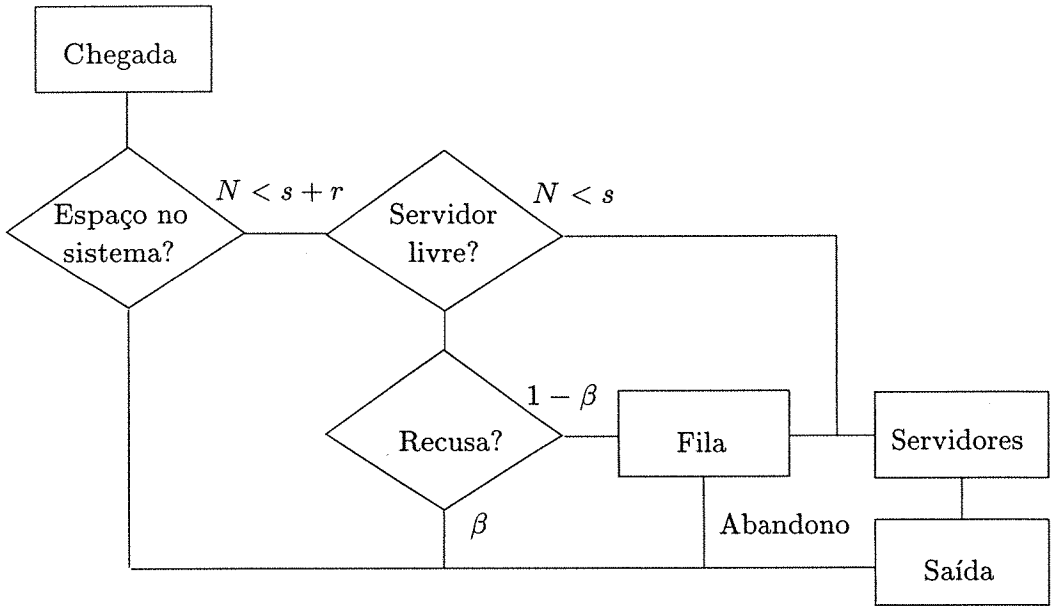

Figura 4.1: Diagrama do Modelo 1.

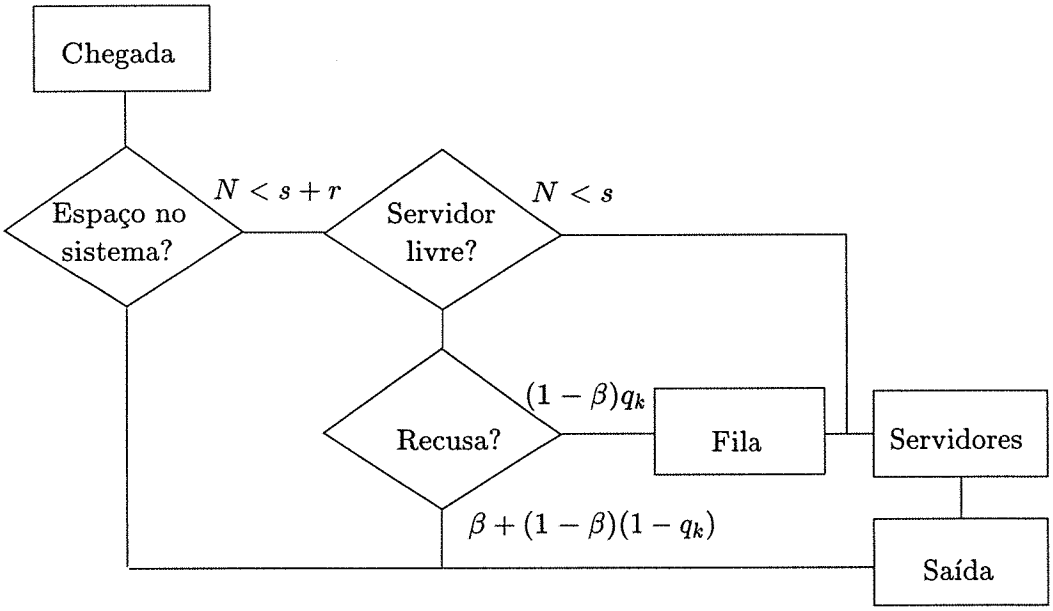

Figura 4.2: Diagrama do Modelo 2.

Conforme apresentado no capítulo anterior, o Modelo 2 difere do Modelo 1 por informar ao usuário, no momento da chegada, qual o estado do sistema. De posse dessa informação o usuário

avalia se o tempo que irá aguardar na fila é maior do que ele está disposto a dispender. Na simulação essa característica foi implementada através de mais um controle do fluxo de usuários representando a avaliação do tempo de espera em relação à *tolerância*. Na Figura 4.2 pode-se notar que há dois blocos decisórios que conjuntamente representam o processo de *recusa*. Dessa forma, para o Modelo 2, desconsiderou-se o bloco de *abandono* já que, conforme visto no Capítulo 3, o conhecimento do número no sistema permitiria a entrada somente daqueles usuários dispostos a aguardar tempo suficiente até serem atendidos.

4.2 Validação dos modelos

Antes de buscar conclusões através da simulação, foi necessário validar os modelos de simulação construídos e para tanto, utilizou-se os valores, simulados, obtidos por [Whitt, 1999b] como referências. Originalmente tais resultados foram apresentados para comparar o desempenho dos modelos. Aqui, também serviram para aferir a implementação.

O conjunto de testes foi composto de duas etapas. A primeira consistia em uma situação em que os modelos foram submetidos a variações de escala de serviço. Na segunda, observou-se a reação dos modelos a cargas maiores de demanda do sistema.

Nas Tabelas 4.1 e 4.2 são apresentados os resultados da primeira etapa da simulação. A variação foi obtida pela diferença relativa entre os valores obtidos pelas simulações e suas referências. Na primeira tabela utilizou-se um número de servidores $s = 4$ enquanto na segunda usou-se $s = 40$. Em ambos os casos, o tamanho da sala de espera (r) foi devidamente dimensionado de forma a não haver saídas do sistema por motivo de *impedimento*, ou seja, foi definido um tamanho suficientemente grande para que sempre houvesse espaço na fila para um nova entrada. Tal escolha foi necessária para adequar-se aos parâmetros adotados por [Whitt, 1999b].

As medidas de performance utilizadas para a validação dos modelos foram:

- $P(N \geq s)$: Probabilidade de haver fila;
- $E(N-s)^+$: Número médio de usuários na fila;
- EN: Número médio de usuários no sistema;

- $SD(N)$: Desvio padrão do número no sistema;
- $P(abandono)$: Probabilidade de um usuário na fila abandonar o sistema;
- $P(atendimento)$: Probabilidade de uma chegada ser atendida;
- $E(C|S)$: Tempo médio dos usuários que completaram o serviço;
- $SD(C|S)$: Desvio padrão do tempo de permanência dos usuários que completaram o serviço;
- $E(A|R)$: Tempo médio de permanência dos usuários que abandonaram o sistema.

Considerou-se válidas as medidas de performance cujas variações permaneceram dentro de um limite relativo de 5%. Observa-se na Tabela 4.1 que, para os dois modelos, 6 das 9 medidas foram aceitas, ou seja, ficaram abaixo da variação máxima de 5%. Já na Tabela 4.2, 6 medidas foram aceitas para o Modelo 1 e 7 para o Modelo 2.

Medida de Performance	**Modelo 1**			**Modelo 2**		
	Simulação	**Referência**	**Variação(%)**	**Simulação**	**Referência**	**Variação(%)**
$P(N \geq s)$	0,564	0,501	12,6	0,419	0,493	-15,0
$P(abandono)$	0,142	0,124	14,5	–	–	–
$P(atendimento)$	0,795	0,775	2,6	0,711	0,772	-7,9
$E(N-s)^+$	0,580	0,498	16,5	0,427	0,445	-4,0
EN	3,740	3,600	3,9	3,290	3,530	-6,8
$SD(N)$	1,790	1,740	2,9	1,740	1,670	4,2
$E(C\|S)$	1,129	1,115	1,3	1,159	1,144	1,3
$SD(C\|S)$	1,013	1,026	-1,3	1,038	1,046	-0,8
$E(A\|R)$	0,279	0,282	-1,1	–	–	–

Tabela 4.1: Comparação dos Modelos 1 e 2 em função do tamanho do sistema ($s = 4$, $\lambda = 4$, $\mu = 1$, $\alpha = 1$, $\beta = 0,2$ e $r = 50$).

A seguir, nas Tabelas 4.3 e 4.4, encontram-se os resultados da segunta etapa de simulações. Aqui, optou-se por testar os modelos a cargas maiores do sistema. Tal efeito foi conseguido ao elevar-se a taxa de chegadas ($\lambda = 20$ e $\lambda = 40$ nas duas simulações).

Medida de Performance	Modelo 1			Modelo 2		
	Simulação	Referência	Variação(%)	Simulação	Referência	Variação(%)
$P(N \geq s)$	0,383	0,335	14,3	0,347	0,333	4,2
$P(abandono)$	0,020	0,020	0,0	–	–	–
$P(atendimento)$	0,913	0,913	0,0	0,911	0,912	-0,1
$E(N-s)^+$	0,974	0,816	19,4	0,840	0,796	5,5
EN	37,670	37,300	1,0	37,110	37,300	-0,5
$SD(N)$	5,200	4,860	7,0	5,080	4,830	5,2
$E(C\|S)$	1,024	1,021	0,3	1,021	1,022	-0,1
$SD(C\|S)$	0,997	1,001	-0,4	1,000	1,001	-0,1
$E(A\|R)$	0,070	0,069	1,4	–	–	–

Tabela 4.2: Comparação dos Modelos 1 e 2 em função do tamanho do sistema ($s = 40$, $\lambda = 40$, $\mu = 1$, $\alpha = 1$, $\beta = 0,2$ e $r = 80$).

Medida de Performance	Modelo 1			Modelo 2		
	Simulação	Referência	Variação(%)	Simulação	Referência	Variação(%)
$P(N \geq s)$	0,975	0,970	0,5	0,961	0,958	0,3
$P(abandono)$	0,315	0,308	2,3	–	–	–
$P(atendimento)$	0,492	0,498	-1,3	0,497	0,497	0,0
$E(N-s)^+$	6,300	6,170	2,1	4,660	4,660	0,0
EN	16,260	16,100	1,0	14,590	14,600	-0,1
$SD(N)$	3,880	3,900	-0,5	3,020	3,090	-2,3
$E(C\|S)$	1,450	1,440	0,8	1,470	1,470	0,0
$SD(C\|S)$	1,040	1,040	0,0	1,067	1,065	0,2
$E(A\|R)$	0,294	0,293	0,3	–	–	–

Tabela 4.3: Comparação dos Modelos 1 e 2 sob altas cargas do sistema ($s = 10$, $\lambda = 20$, $\mu = 1$, $\alpha = 1$, $\beta = 0,2$ e $r = 50$).

Medida de Performance	Modelo 1			Modelo 2		
	Simulação	Referência	Variação(%)	Simulação	Referência	Variação(%)
$P(N \geq s)$	0,999	0,998	0,1	0,998	0,996	0,3
$P(abandono)$	0,247	0,251	-1,6	–	–	–
$P(atendimento)$	0,252	0,250	0,8	0,250	0,250	0,0
$E(N-s)^+$	10,000	10,040	-0,4	6,750	6,840	-1,3
EN	20,010	20,000	0,1	16,740	16,800	-0,4
$SD(N)$	4,460	4,430	0,7	3,020	3,160	-4,4
$E(C\|S)$	1,650	1,650	-0,2	1,660	1,680	-1,2
$SD(C\|S)$	1,040	1,045	-0,5	1,060	1,080	-1,9
$E(A\|R)$	0,354	0,356	-0,6	–	–	–

Tabela 4.4: Comparação dos Modelos 1 e 2 sob altas cargas do sistema ($s = 10$, $\lambda = 40$, $\mu = 1$, $\alpha = 1$, $\beta = 0,5$ e $r = 50$).

Nesta etapa das simulações obteve-se maior aderência às referências já que todas as medidas puderam ser consideradas aceitas pois tiveram variação abaixo de 5% em relação às referências.

Em resumo, para os resultados apresentados, das 72 medidas observadas nas duas etapas e para os dois modelos, 60 foram consideradas aceitas, ou seja, cerca de 83% de aceitação. Conclui-se que os modelos de simulação apresentados estão validados.

4.3 Perturbação dos modelos

Inicialmente, utilizou-se os Modelos 1 e 2 submetidos a uma escala maior de serviço. A seguir os mesmos modelos foram estudados quando o sistema era solicitado por uma demanda maior. Por fim, verificou-se o comportamento do Modelo 2 quando as taxas de *recusa* foram modificadas.

Nos estudos de maiores escala de serviço e demanda, há quatro conjuntos de perturbações dos modelos que serão utilizados. O intuito foi estudar o comportamento do sistema com diferentes características dos usuários em relação aos tempos que se sujeitam a aguardar por atendimento.

O primeiro grupo representou aqueles usuários com baixa aceitação à espera por atendimento. Para tanto, foi utilizada uma distribuição Gama(0,5;0,25) caracterizada por uma cauda mais pesada à esquerda, ou seja, concentrando a *tolerância* dos usuários em níveis baixos.

A seguir, foi definida uma distribuição Exponencial de média 2 para a *tolerância* de forma a representar usuários menos exigentes.

No terceiro grupo enquadrou-se um conjunto de usuários que não apresentam um padrão definido de suscetibilidade à fila. Essa característica foi modelada com o uso de uma distribuição dos tempos de tolerância de acordo com uma Uniforme(0;4). Com isso, esperou-se retratar um público heterogêneo em que cada tipo de *tolerância* é igualmente provável.

No último grupo, pretendeu-se estudar o sistema com usuários que tendem a permanecer o tempo necessário até que sejam atendidos. Pode-se modelar tal *tolerância* por meio de uma outra distribuição Gama com parâmetros de forma e escala iguais a 4 e 2, respectivamente. Com isso, manteve-se a concentração dos tempos em níveis mais elevados.

Pode-se notar que em todos os casos manteve-se intencionalmente a média do tempo de

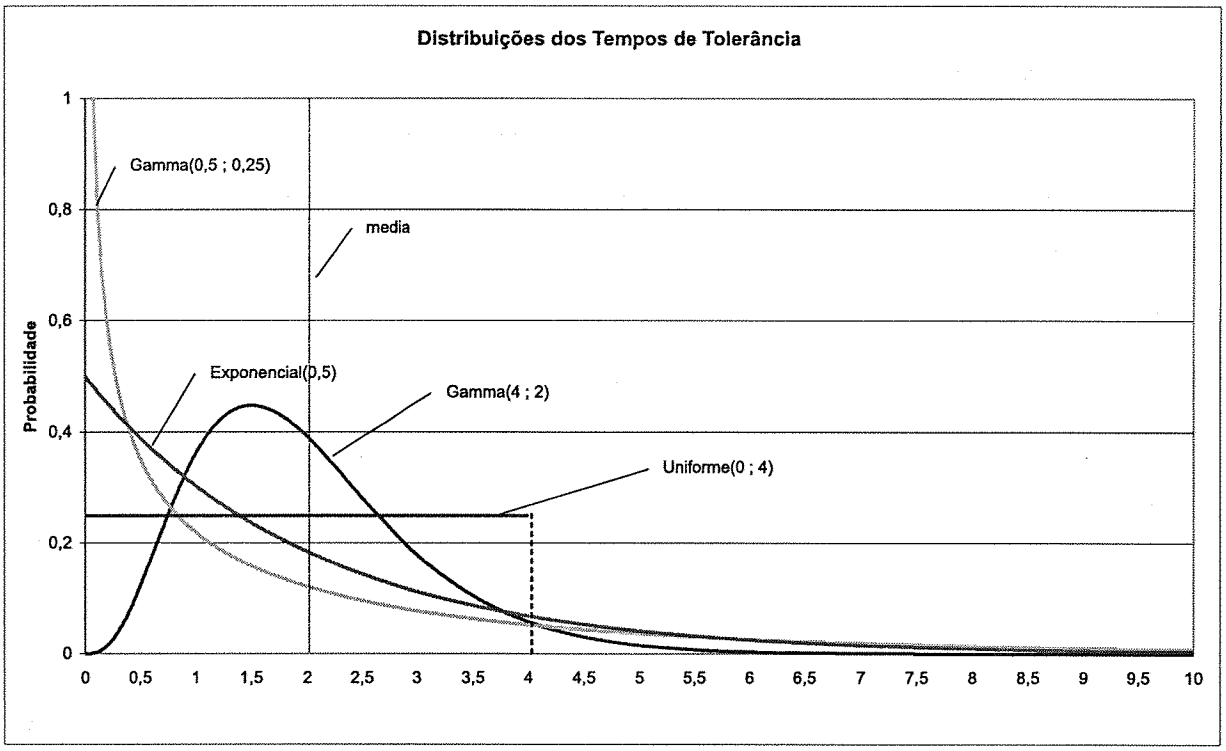

Figura 4.3: Distribuições dos tempos de *tolerância*.

tolerância igual a 2 minutos (ver Figura 4.3). Tal medida foi escolhida para que fosse possível comparar, de uma maneira geral, os resultados obtidos em cada situação.

Na apresentação dos resultados optou-se por utilizar duas casas decimais para as medidas de desempenho. Acredita-se que tal quantidade é suficiente para representar adequadamente os valores obtidos. Na seção anterior o uso de três casas decimais deveu-se aos dados originais de referência estarem com essa precisão.

4.3.1 Sistema com maior escala de serviço

Para simular maior escala utilizou-se os seguintes parâmetros: $\lambda = 400$, $\mu = 1$, $\beta = 0,2$, $s = 400$ e $r = 1000$.

As medidas de desvio padrão do número de usuários no sistema, $SD(N)$, e do tempo para completar o serviço condicionado que o serviço foi completado, $SD(C|S)$, não foram apresentadas

por possuirem valores similares para as quatro distribuições da *tolerância*. O mesmo aconteceu com as probabilidades de *abandono*, praticamente nulas, e as probabilidades de atendimento, que oscilaram entre 0,96 e 0,97. A seguir são apresentados os resultados obtidos:

Quando o sistema atinge maiores dimensões ($s \geq 400$), as perturbações na *tolerância* dos usuários não são tão relevantes nos dois modelos estudados (ver Tabelas 4.5 e 4.6). As únicas diferenças palpáveis dizem respeito ao tamanho médio da fila e à probabilidade de haver fila no Modelo 1. Essas duas medidas refletem a natureza dos usuários. Quando a *tolerância* segue uma distribuição Gama(0,5;0,25) encontram-se os menores valores de tamanho médio de fila e probabilidade de haver fila (ver Tabela 4.5). Tal resultado era esperado devido à escolha de uma distribuição que representasse usuários mais exigentes que não tolerassem espera demasiada na fila.

No caso da *tolerância* distribuir-se de acordo com uma Gama(4;2) o tamanho médio da fila aumenta em relação ao das demais distribuições, da mesma forma que o número médio no sistema. Para a *tolerância* distribuída segundo uma Exponencial(0,5) e Uniforme(0,4) encontram-se valores intermediários aos das distribuições Gama. Na Uniforme não há um padrão claro dos usuários, que podem representar igualmente clientes muito pacientes ou extremamente impacientes. No Modelo 2 o impacto de diferentes distribuições da *tolerância* é praticamente desprezível (ver Tabela 4.6).

Medida de Performance	**Modelo 1**			
	Gama(0,5;0,25)	**Exponencial(0,5)**	**Uniforme(0;4)**	**Gama(4;2)**
$P(N \geq s)$	0,15	0,19	0,18	1,00
$E(N-s)^+$	0,32	0,74	0,78	0,81
EN	386,75	387,60	387,60	387,79

Tabela 4.5: Comparação de diferentes distribuições do tempo de *tolerância* em sistema com maior escala (Modelo 1).

Medida de Performance	**Modelo 2**			
	Gama(0,5;0,25)	**Exponencial(0,5)**	**Uniforme(0;4)**	**Gama(4;2)**
$P(N \geq s)$	0,07	0,07	0,08	0,07
$E(N-s)^+$	0,04	0,04	0,04	0,04
EN	385,15	387,11	385,20	384,92

Tabela 4.6: Comparação de diferentes distribuições do tempo de *tolerância* em sistema com maior escala (Modelo 2).

Percebe-se, portanto, que quando a escala supera os 400 servidores, não são relevantes os perfis dos usuários em relação ao tempo de *tolerância* para o desempenho do sistema. Dessa forma, caso a performance não seja satisfatória, deve-se procurar outros parâmetros para enquadrar o sistema no desempenho esperado ao invés de tentar adaptá-lo à tolerância dos usuários.

4.3.2 Sistema com maior demanda

Nesta Seção os Modelos 1 e 2 foram estudados quando o sistema era solicitado por uma demanda maior, representada pelos parâmetros: $\lambda = 40$, $\mu = 1$, $\beta = 0,5$, $s = 10$ e $r = 50$.

Nas Tabelas 4.7 e 4.8 a seguir são apresentados os valores dessas perturbações. As probabilidades de atendimento foram excluídas por serem todas iguais a 0,25, independentemente da distribuição escolhida do tempo de *tolerância*.

Medida de Performance	**Modelo 1**			
	Gama(0,5;0,25)	**Exponencial(0,5)**	**Uniforme(0;4)**	**Gama(4;2)**
$P(N \geq s)$	0,94	1,00	1,00	1,00
$P(abandono)$	0,37	0,25	0,26	0,05
$E(N-s)^+$	1,50	20,16	29,60	48,54
EN	11,44	30,16	39,69	58,54
$SD(N)$	1,41	6,25	6,78	1,81
$E(C\|S)$	1,07	2,35	2,92	5,19
$E(A\|R)$	0,05	0,66	1,05	3,26

Tabela 4.7: Comparação de diferentes distribuições do tempo de *tolerância* em sistema com alta carga (Modelo 1).

Medida de Performance	**Modelo 2**			
	Gama(0,5;0,25)	**Exponencial(0,5)**	**Uniforme(0;4)**	**Gama(4;2)**
$P(N \geq s)$	0,94	1,00	0,95	0,97
$E(N-s)^+$	0,78	1,02	1,03	1,60
EN	10,70	10,96	10,97	11,57
$SD(N)$	0,65	0,81	0,80	1,02
$E(C\|S)$	1,07	1,10	1,10	1,16

Tabela 4.8: Comparação de diferentes distribuições do tempo de *tolerância* em sistema com alta carga (Modelo 2).

Nas duas tabelas pode-se notar claramente que há distinção das medidas para cada distribuição do tempo de *tolerância* contrariamente ao que foi visto nos casos de maior escala de serviço.

Conforme o perfil dos usuários torna-se menos exigente (concentração da *tolerância* em tempos maiores) percebe-se que os tamanhos médios da fila e do sistema crescem; no Modelo 1 também é observado um decréscimo das probabilidades de abandono e aumento dos tempos condicionais de serviço e abandono quando os usuários completam o serviço ou abandonam, respectivamente. O desvio padrão do tempo para completar o serviço condicionado que o serviço foi completado, não foi apresentado nas tabelas desta seção por não representarem uma medida relevante para comparação, já que em cada perfil de *tolerância* os valores foram muito próximos. Em todos os casos não há grandes alterações das probabilidades de atendimento, pois a disciplina dos servidores não foi alterada.

Em vista dos resultados apresentados nas Tabelas 4.7, para o Modelo 1, se houver forte demanda do sistema, é importante conhecer o comportamento dos usuários em relação ao *abandono*. Percebe-se, também para esse modelo que a variabilidade do número de usuários no sistema é alta para a maioria das distribuições da *tolerância*. Isso indica que, em determinadas ocasiões, há um grande prejuízo na qualidade do atendimento devido à grande quantidade de usuários, enquanto que em outros momentos pode haver ociosidade do sistema. No caso do Modelo 2 as diferenças são menores, mas nem sempre podem ser desprezadas.

4.3.3 Taxa de recusa

Nesta seção estudou-se o comportamento do Modelo 1 para diferentes taxas de *recusa*. Utilizou-se uma configuração do sistema submetido a uma carga elevada. Conforme visto no Capítulo 3 a probabilidade de *recusa* foi indicada por β. Assim, mantendo-se constantes os demais parâmetros ($s = 10$, $\lambda = 40$, $\mu = 1$, $\alpha = 1$ e $r = 50$), variou-se β entre $1,0$ a $0,0$ com decrementos de $0,2$. Esses dois extremos representam, respectivamente, um modelo M/M/s/0 (sem espaço na fila de espera) e um M/M/s/r sem *recusa*.

A apresentação dos dados foi dividida entre três gráficos, cada um com um grupo de medidas de mesma natureza. A primeira figura mostra algumas probabilidades indicativas do desempenho do sistema. Na segunda, constam os valores de tempos médios enquanto a última apresenta as informações de tamanhos médios. Há também uma tabela agrupando os valores obtidos.

Na Figura 4.4 observa-se que, conforme diminui a proporção de *recusa*, as probabilidades de haver fila e de abandono aumentam. Por outro lado a probabilidade de atendimento permanece praticamente constante.

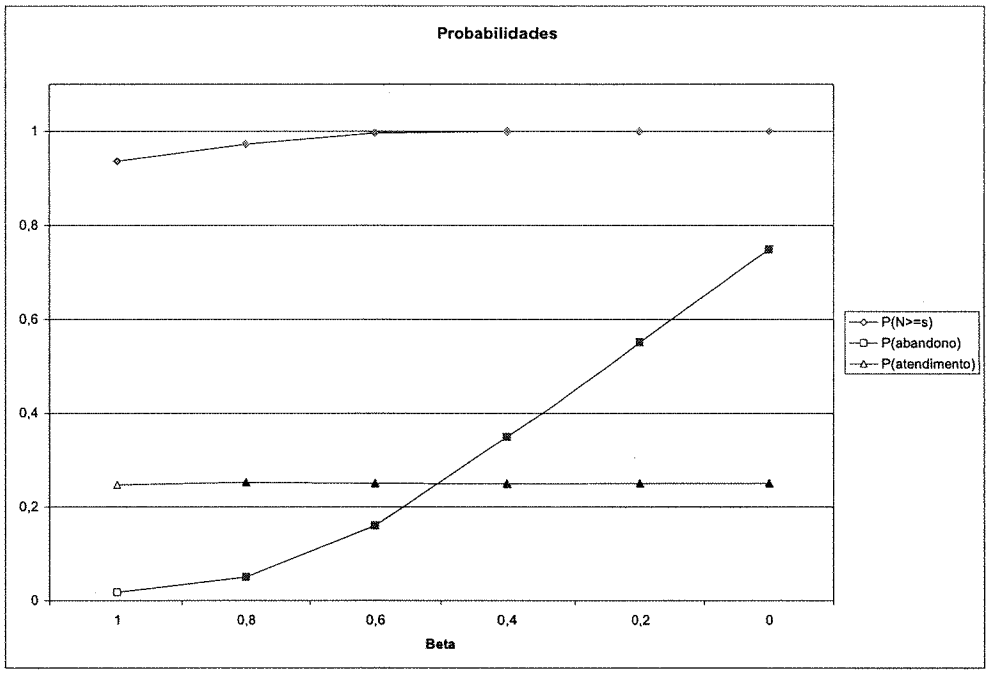

Figura 4.4: Evolução das probabilidades para diferentes taxas de *recusa*.

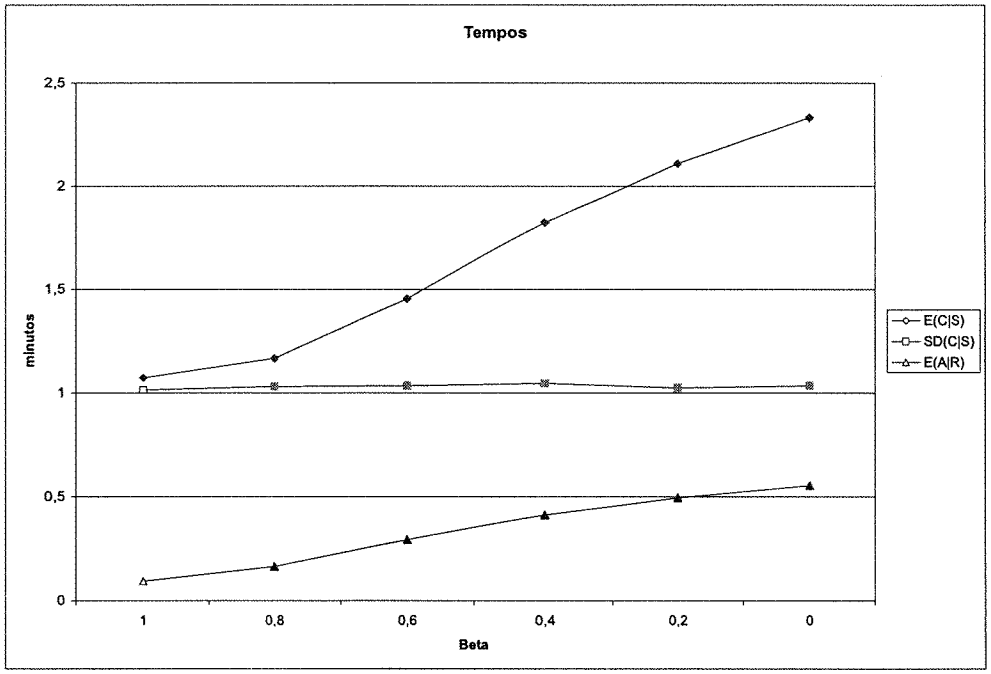

Figura 4.5: Evolução dos tempos para diferentes taxas de *recusa*.

Pela Figura 4.5 é possível notar que os tempos, para completar o serviço condicionado que foi completado e de abandono condicionado que houve o abandono, aumentam conforme diminui-se a probabilidade de *recusa*. Importante notar que o desvio padrão dos tempos para que uma chegada venha a completar o serviço, dado que o serviço foi completado, tem pouca alteração para todos os parâmetros.

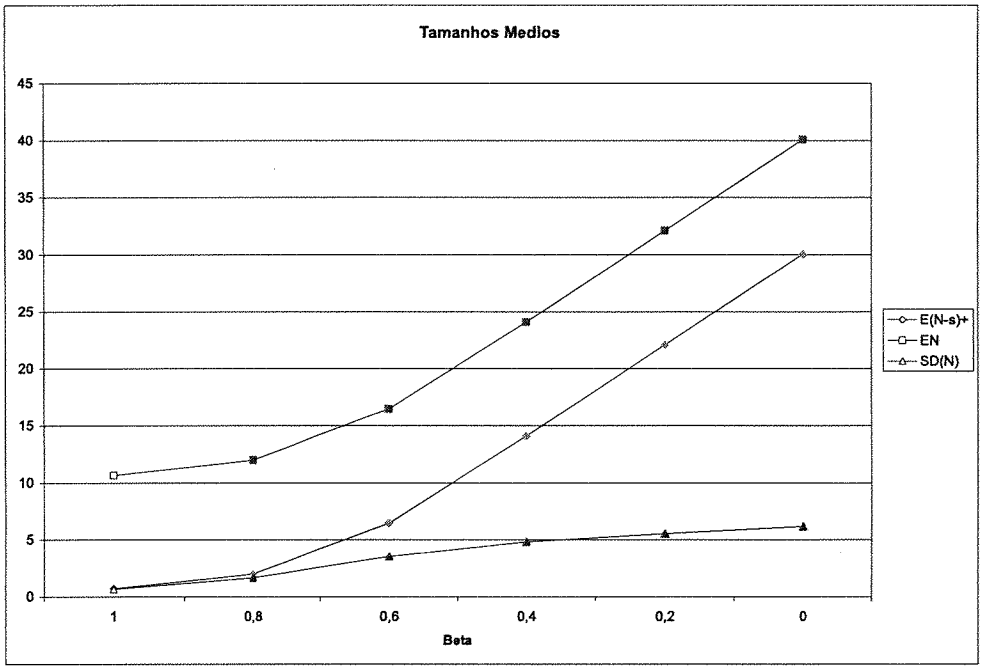

Figura 4.6: Evolução dos tamanhos para diferentes taxas de *recusa*.

Por fim, na Figura 4.6, estão apresentadas as curvas dos tamanhos médios da fila e do sistema, além do desvio padrão do tamanho do sistema. Todas elas crescem de acordo com a redução da proporção de *recusa*.

Percebe-se um resultado natural na evolução do tamanho médio da fila e do sistema de acordo com as alterações da taxa de *recusa*. Uma vez que aumenta-se a predisposição dos usuários em permanecer em fila também eleva-se a quantidade de usuários no sistema. De qualquer forma, a taxa de *recusa* representa um parâmetro de forte impacto para o desempenho do sistema.

Na Tabela 4.9 encontram-se os valores das medidas de desempenho para cada probabilidade de *recusa* utilizada. Tanto por essa tabela quanto pelos gráficos nota-se que são praticamente

insensíveis às diferentes taxas de recusa, como por exemplo, a probabilidade de haver fila e a probabilidade de atendimento. Tal constatação justifica-se pelo uso de parâmetros representando um sistema sob alta demanda. Talvez essa configuração possa ocultar outros resultados pertinentes à variação do parâmetro β.

Medida de Performance	**Probabilidade de recusa (β)**					
	1,0	**0,8**	**0,6**	**0,4**	**0,2**	**0,0**
$P(N \geq s)$	0,94	0,97	1,00	1,00	1,00	1,00
$P(abandono)$	0,09	0,05	0,16	0,35	0,55	0,75
$P(atendimento)$	0,25	0,25	0,25	0,25	0,25	0,25
$E(N-s)^+$	0,74	2,00	6,47	14,08	22,09	30,06
EN	10,66	11,97	16,47	24,09	32,11	40,09
$SD(N)$	0,66	1,66	3,55	4,82	5,56	6,19
$E(C\|S)$	1,07	1,17	1,46	1,82	2,11	2,33
$E(A\|R)$	0,09	0,16	0,29	0,41	0,50	0,55

Tabela 4.9: Modelo 2 submetido a diferentes taxas de *recusa*.

Capítulo 5

Conclusões

Este trabalho estudou um modelo de filas com desistências. Suas principais características foram apresentadas e um modelo de simulação foi construído para avaliar algumas alterações nos parâmetros iniciais do modelo.

As perturbações do sistema foram aplicadas sobre os parâmetros que dizem respeito ao comportamento dos usuários. À primeira vista essas características aparentavam possuir fortes ligações com os processos de *recusa* e *abandono*. A partir dos resultados apresentados no Capítulo 4 observa-se que é realmente importante conhecer o comportamento dos usuários do sistema.

Nos dois modelos discutidos, as características de *recusa* e *abandono* têm efeitos diferentes, ficando mais evidentes quando o sistema é submetido a maiores cargas de demanda.

Observando a escala em que atua o sistema, verificou-se que, para uma grande quantidade de servidores, a característica dos usuários em termos do tempo de *tolerância* não representa um fator de grande alteração no desempenho do sistema. Ou seja, em situações de maior escala como as apresentadas no Capítulo 4, a *tolerância* dos usuários não é a principal característica a ser considerada para uma melhoria do sistema.

Para outra situação estudada, aquela em que o sistema é submetido a uma grande demanda, observou-se, para o Modelo 1, que o impacto da distribuição da *tolerância* é bastante significativo para o desempenho do sistema. Além disso, nessas condições a variabilidade do número de usuários no sistema é alta. Porém, o Modelo 2 é mais estável quando solicitado por maiores

demandas.

No caso de perturbações na taxa de *recusa*, o fato de que as simulações foram feitas com base em um sistema sob alta demanda pode ter afetado a visibilidade de outros efeitos da variação na *recusa*. Indiferentemente a isso, os resultados confirmaram a expectativa de que quanto maior a predisposição dos usuários à espera em fila, aumentam a quantidade de usuários na fila e no sistema, a probabilidade de abandono e também o tempo médio no sistema.

Neste trabalho adotou-se várias hipóteses de forma a facilitar a modelagem. Entretanto, a não-linearidade do mundo real exige alterações nos modelos que contemplem outros tipos de solicitações que não foram previstas. Assim, é possível citar algumas frentes de pesquisa que podem ajudar a aperfeiçoar os modelos aqui estudados.

Talvez contribuísse para o aperfeiçoamento do modelo algumas perturbações em parâmetros específicos do sistema em lugar de aplicá-las sobre características dos usuários. Nesse sentido, uma maior variação no tamanho da fila de espera, talvez possa causar impacto no desempenho do sistema. Uma vez reduzido o espaço para a fila haveria a incidência do *impedimento*. Independentemente das características dos usuários, o impacto de chegadas perdidas poderia ser relevante para as medidas de desempenho do sistema. Dessa forma, todos os resultados obtidos nas simulações anteriores seriam passíveis de alterações, caso fossem originados a partir de sistemas que contemplassem o *impedimento*.

Outro possível ponto a ser explorado, em futuros estudos, seria uma tentativa de análise mais realista sob a óptica do processo de chegada. Neste trabalho a taxa de chegada permaneceu constante em todas as ocasiões. Poderia-se definir a taxa de chegada através de uma função dependente do tempo, $\lambda(t)$, a fim de se considerar a não estacionariedade das chegadas. Tal técnica tem sido adotada por vários autores para simular ciclos periódicos de demanda, por exemplo, [Jennigs et al., 1996] e [Green et al., 1991]. Também seria proveitoso mesclar a característica de funções cíclicas das chegadas com os benefícios de informar o estado do sistema aos usuários, conforme discutido em [Whitt, 1999a].

Referências Bibliográficas

[Cooper, 1981] Cooper, R. B. (1981). *Introduction to queueing theory.* North Holland.

[Green et al., 1991] Green, L., Kolesar, P., and Svoronos, A. (1991). Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research*, 39:502–511.

[Gross and Harris, 1998] Gross, D. and Harris, C. M. (1998). *Fundamentals of queueing theory.* John Wiley & Sons, 3ª edição.

[Hoel et al., 1971] Hoel, P. G., Port, S. C., and Stone, C. J. (1971). *Introduction to probability theory.* Houghton Mifflin Company.

[Jennigs et al., 1996] Jennigs, O. B., Mandelbaun, A., Massey, W. A., and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science*, 42:1383–1394.

[Kleinrock, 1975] Kleinrock, L. (1975). *Queueing systems*, volume I. John Wiley & Sons.

[Magalhães, 1996] Magalhães, M. N. (1996). *Introdução à rede de filas.* Associação Brasileira de Estatística.

[Magalhães, 2004] Magalhães, M. N. (2004). *Probabilidade e Variáveis Aleatórias.* Instituto de Matemática e Estatística - USP.

[Ross, 1970] Ross, S. M. (1970). *Applied probability models with optimization applications.* Holden-Day.

[Ross, 1982] Ross, S. M. (1982). *Stochastic processes.* John Wiley & Sons.

[Takács, 1962] Takács, L. (1962). *Introduction to the theory of queues.* Oxford University Press, NY.

[Whitt, 1993] Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742.

[Whitt, 1999a] Whitt, W. (1999a). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212.

[Whitt, 1999b] Whitt, W. (1999b). Improving service by informing customers about anticipated delays. *Management Science*, 45:192–207.