v.4, n.3, p. 278-295, dez. 1997

MODELAGEM POR MEIO DE TEORIA DE FILAS DO *TRADEOFF* ENTRE INVESTIR EM CANAIS DE ATENDIMENTO E SATISFAZER O NÍVEL DE SERVIÇO EM PROVEDORES *INTERNET*

Gisele Castro Fontanella
Universidade Federal de São Carlos
Departamento de Engenharia de Produção
13565-905 - São Carlos - SP
E-mail: pgcf@iris.ufscar.br

Reinaldo Morabito
Universidade Federal de São Carlos
Departamento de Engenharia de Produção
13565-905 - São Carlos - SP
E-mail: morabito@power.ufscar.br

Resumo

A conexão de computadores à INTERNET é feita por meio de empresas denominadas provedores de acesso à INTERNET, que atendem vários tipos de usuários oferecendo diferentes formas de conexão física. A mais simples é aquela em que o usuário se conecta a um canal de atendimento do provedor por meio de uma linha telefônica comum. Conseguir um canal disponível pode não ser uma tarefa fácil, principalmente nos horários de pico, o que resulta num problema enfrentado pelos provedores, que é o de determinar a relação ideal entre o número de usuários e o número de canais de atendimento. Para resolvê-lo, é preciso analisar o tradeoff entre investir em capacidade (número de canais) e satisfazer o nível de serviço desejado aos usuários (probabilidade de acesso). O objetivo deste trabalho é modelar este tradeoff por meio de teoria de filas. A metodologia envolve basicamente três passos: (i) analisar os processos de chegada (chamadas) e serviço (atendimento) dos usuários em certos períodos, (ii) selecionar um modelo de filas apropriado, sob algumas hipóteses simplificadoras, e (iii) construir curvas de tradeoff entre medidas de desempenho do sistema, em particular, da probabilidade de acesso em função do número de canais de atendimento, ou da taxa média de usuários. Para ilustrar a aplicação da metodologia, são apresentados os resultados de um estudo de caso realizado num provedor no interior do estado de São Paulo, com processo de chegada Poisson mas com tempo de atendimento não exponencialmente distribuído. Algumas perspectivas para pesquisa futura são apontadas, como agrupar os usuários em diferentes classes, em função do comportamento da chegada e do serviço, e analisar o particionamento da capacidade em função destas classes.

Palavras-chave: provedor INTERNET, teoria de filas, nível de serviço.

1. Introdução

A *INTERNET*, maior rede de computadores e informações do mundo, começou nas universidades, depois migrou para as empresas, e hoje invade os ambientes domésticos. A conexão de computadores à *INTERNET* é feita por meio de empresas denominadas *provedores de acesso* à *INTERNET* (*INTERNET access providers*), ou simplesmente *provedores INTERNET*. Qualquer usuário, seja ele um usuário doméstico ou uma empresa, que tenha acesso a um computador com um *modem*, uma linha telefônica (ou uma linha dedicada) e um *software* de comunicação, pode acessar a *INTERNET*, mas para isso é necessário que ele esteja conectado à um provedor (CHARLAB, 1996).

Para que um usuário doméstico possa conectar-se a um provedor, ele precisa conseguir um *canal de atendimento* disponível neste provedor. O canal de atendimento compreende o seguinte conjunto de recursos: linha telefônica, *modem*, porta de comunicação e espaço em memória no servidor de comunicação. Em grande parte dos provedores, devido à limitação de investimentos em canais de atendimento, obter um canal disponível nos horários de pico nem sempre é uma tarefa fácil (RAMOS, 1996). Em muitos casos, o usuário precisa tentar várias vezes até conseguir um canal livre (quando não desiste antes), o que pode deixá-lo insatisfeito.

Este fato retrata um dos problemas enfrentados hoje pelos provedores de acesso, que é o de determinar a relação ideal entre o número de usuários (clientes) e o número de canais de atendimento. Se esta relação for muito alta, o usuário provavelmente terá dificuldade em conseguir um canal disponível, especialmente nos horários de pico. Por outro lado, se esta relação for muito baixa, haverá ociosidade nos canais de atendimento, indicando que a empresa investiu desnecessariamente em canais. Para determinar a relação ideal, é preciso analisar o *tradeoff* entre investir em *capacidade* e satisfazer o *nível de serviço* que a empresa deseja oferecer aos seus usuários. Para o propósito deste trabalho, entende-se capacidade simplesmente como sendo o número de canais de atendimento, e nível de serviço como sendo a probabilidade de acesso, isto é, a probabilidade de um usuário encontrar um canal de atendimento disponível num certo instante.

Dado que, nestes últimos anos, o número de usuários da *INTERNET* vem crescendo de forma surpreendente, e a previsão é de que continue crescendo (ANDRIES, 1997), muitos provedores encontram-se hoje com mais usuários do que suas infra-estruturas são capazes de suportar. Esse excesso tem como conseqüência imediata uma redução do nível de serviço oferecido aos usuários. Provedores que não investirem no planejamento e configuração de seus sistemas provavelmente estarão deteriorando seus níveis de serviço, o que poderá levar seus usuários a buscar outros provedores com melhores níveis de serviço.

O objetivo deste trabalho é analisar, por meio de modelos da teoria de filas, o *tradeoff* entre investir em capacidade e satisfazer o nível de serviço desejado aos usuários. A metodologia envolve basicamente três passos: (i) analisar os processos de chegada (chamadas telefônicas) e serviço (tempos de conexão) dos usuários em certos períodos, (ii) selecionar um modelo de filas apropriado, sob algumas hipóteses simplificadoras, e (iii) construir curvas de *tradeoff* entre medidas de desempenho do sistema, em particular, da probabilidade de acesso em função do número de canais de atendimento, ou da taxa média de chegada de usuários. Convém salientar que não foram encontrados trabalhos na literatura analisando este problema

por meio de modelos de filas. A pesquisa bibliográfica foi realizada em base de dados como *ArticleFirst* e INSPEC na área de teoria de filas e *INTERNET*.

A próxima seção apresenta uma breve discussão sobre provedores *INTERNET* e teoria de filas. A seção 3 analisa os modelos de filas mais adequados para representar o problema e as hipóteses envolvidas, e a seção 4 detalha os passos a serem seguidos na metodologia mencionada acima, conforme FONTANELLA (1997). A seção 5 apresenta os resultados obtidos ao aplicar esta metodologia num estudo de caso realizado num provedor de médio porte (58 canais de atendimento, 800 usuários), localizado no interior do estado de São Paulo. Finalmente, a seção 6 apresenta as conclusões e perspectivas para pesquisa futura.

2. Provedores *INTERNET* e Teoria de Filas

Conforme foi mencionado, a conexão de computadores à *INTERNET* ocorre por meio de provedores *INTERNET*. Primeiramente é estabelecida a conexão com os computadores do provedor de acesso que, por sua vez, estão conectados a um provedor de serviço de *backbone* (p.e., Embratel, Telesp) e, a partir deste, à *INTERNET*. Os provedores podem atender diversos tipos de usuários, como usuários domésticos, pequenas empresas, grandes corporações, e para isso disponibilizam diferentes formas de conexão física. A mais simples é aquela em que a ligação entre o computador do usuário e do provedor é feita por meio de uma linha telefônica, o que caracteriza um acesso discado. Outra forma de conexão ao provedor é mediante o acesso dedicado, no qual existe uma conexão permanente que nunca é desligada. Este tipo de acesso é indicado para empresas nas quais a freqüência de acesso é alta e o tempo de duração da conexão é longo. Neste caso, não é necessário fazer uma chamada telefônica para o provedor, pois a empresa está sempre conectada à *INTERNET*.

O acesso discado, também chamado de *dial-up*, é indicado para usuários com baixas freqüências de acesso e tempos de duração de conexão reduzidos (CHARLAB, 1995). Neste tipo de acesso, o usuário disca para um número de telefone do provedor e a conexão só existe enquanto a ligação telefônica estiver ativa. Para que se estabeleça a conexão, é necessário, por parte do usuário, um computador, um *modem*, uma linha telefônica e um *software* de comunicação, e, por parte do provedor, um canal de atendimento, que compreende o conjunto de recursos descrito na seção 1. Este trabalho aborda apenas o problema do acesso discado (usuários domésticos), no qual é necessário que o usuário consiga um canal de atendimento disponível para que possa estabelecer conexão.

Em geral, entende-se nível de serviço como sendo uma medida da satisfação dos usuários com relação à qualidade e à quantidade dos serviços oferecidos pelo provedor. Para o propósito deste trabalho, entende-se nível de serviço como sendo a probabilidade de acesso, ou seja, a probabilidade do usuário encontrar um canal de atendimento disponível (conforme descrito na seção anterior). Outras medidas poderiam ter sido utilizadas, tais como a velocidade do canal de comunicação, preços, suporte técnico, entre outros. Como este trabalho preocupa-se exclusivamente com o sistema de acesso aos canais de atendimento do provedor, entendemos que a probabilidade de acesso seja uma medida representativa do nível de serviço aos usuários.

Teoria de Filas

A teoria de filas tem como objeto de estudo os sistemas geradores de espera, também chamados sistemas de filas. O que existe de comum nesses sistemas é o fluxo de usuários em busca de serviço e algum tipo de restrição no serviço a ser provido. Esta restrição pode se dar com relação ao número máximo de usuários que podem ser servidos simultaneamente, que é o caso dos provedores de acesso à *INTERNET*, ou quando o serviço só está disponível por um período de tempo limitado, como por exemplo no caso de um semáforo (COX & SMITH, 1974).

Estudar um sistema em congestão tem como objetivo entendê-lo e, se possível melhorá-lo, mudando-o de alguma forma. O fenômeno da congestão pode ser descrito de várias formas, por exemplo, com relação ao número médio de usuários na fila, a proporção de tempo em que todos os servidores estão ocupados, o número médio de usuários no sistema, entre outras, que correspondem às medidas de desempenho do sistema. A teoria de filas, pela análise matemática detalhada, procura calcular essas medidas, com a intenção de melhor entender o comportamento do sistema (GROSS & HARRIS, 1974).

Um sistema de fila pode ser descrito por um processo de chegada a uma instalação de serviço, que pode consistir de um ou mais servidores, e um processo de atendimento, que pode ou não causar fila de espera. Esses sistemas são caracterizados por três elementos básicos: o processo de chegada, o processo de serviço e a disciplina de atendimento. Outros componentes podem ser acrescentados como o número de servidores (ou canais de atendimento), a capacidade de armazenagem do sistema e o tamanho da população de usuários (KLEINROCK, 1975).

Para um provedor *INTERNET*, o processo de chegada corresponde aos intervalos de tempo entre as chamadas de usuários. Assume-se que as chamadas ocorrem individualmente e que, num certo período de tempo, os intervalos entre elas sejam variáveis aleatórias independentes e identicamente distribuídas. O processo de serviço corresponde aos tempos de conexão. Cada canal de atendimento atende no máximo um usuário por vez. Também assume-se que os tempos de serviço sejam variáveis aleatórias independentes e identicamente distribuídos, além disso, que sejam independentes do processo de chegada. A capacidade de armazenagem do sistema é igual ao número de canais de atendimento, pois, o sistema não dispõe de uma "sala de espera" onde os usuários aguardam em fila para serem atendidos.

Podemos definir diversas medidas de desempenho de interesse para provedores *INTERNET*. Dentre elas, podemos destacar: a probabilidade de perda de usuários no sistema, a probabilidade de acesso, o número médio de usuários no sistema, o índice de congestionamento médio do sistema, a probabilidade de encontrar o sistema vazio e a proporção de tempo em que todos os servidores estão ocupados. Todas estas medidas são estimadas sob condição de equilíbrio, ou seja, assume-se que o sistema esteja em regime.

3. Modelagem por Meio de Teoria de Filas

Processos Markovianos com espaço de estados discreto e tempo contínuo têm sido amplamente empregados para analisar sistemas de filas. As possíveis transições entre os estados do problema podem ocorrer em qualquer instante de tempo. A propriedade Markoviana garante que a condição futura do processo depende

apenas do estado atual, ou seja, a forma como a história passada influencia na previsão do comportamento futuro do processo deve estar completamente representada no estado atual. Esta propriedade pode ser descrita por:

$$\text{Prob}\{ X(t+\Delta t) = i \mid X(t) = j, X(s) = k_s, 0 \le s < t \} = \text{Prob}\{ X(t+\Delta t) = i \mid X(t) = j \}$$

onde $\Delta t > 0$ e $X(t)$ é o estado do sistema no instante t. Os possíveis valores de $X(t)$ pertencem a um conjunto discreto. Note que podemos ter $k_s = j$ para valores de $s<t$.

A propriedade Markoviana impõe uma grande limitação na distribuição do tempo T em que o processo pode permanecer num certo estado — este tempo tem de ser exponencialmente distribuído, dado que a distribuição exponencial é a única função densidade de probabilidade "sem memória", isto é,

$$\text{Prob}\{ T > t+\Delta t \mid T > t\} = \text{Prob}\{ T > \Delta t \}$$

Se isto ocorrer, o estado do sistema pode ser simplesmente descrito pelo número de usuários presentes no sistema, e nenhuma informação adicional se faz necessária. A descrição do estado é unidimensional e contável, o que facilita a análise matemática (KLEINROCK, 1975).

Modelo G/G/c/c/N

O modelo de filas em princípio mais apropriado para representar o sistema de acesso aos canais de atendimento de um provedor *INTERNET* é o modelo G/G/c/c/N. Essa notação significa que ambos os processos de chegada e de serviço são genéricos (i.e., G/G), o sistema possui apenas c canais de atendimento (i.e., c servidores iguais e em paralelo), a capacidade do sistema é igual a c usuários (i.e., não há sala de espera), e a população de usuários é limitada em N, $N \ge c$ (i.e., o número de usuários cadastrados no provedor). Se um usuário chegar e todos os c servidores estiverem ocupados, ele voltará a fazer parte da população de N-c usuários fora do sistema, pois, não é permitido a formação de filas. Para maiores detalhes dessa e das outras notações usadas nesta seção, veja por exemplo KLEINROCK (1975) ou GROSS & HARRIS (1974).

O modelo G/G/c/c/N, com espaço de estados descrito pelo número de usuários no sistema, não possui a propriedade Markoviana devido aos processos de chegada e serviço poderem ser genéricos, e isto traz grandes dificuldades para sua análise exata. Em geral, recorre-se a aproximações (TIJMS, 1986; WHITT, 1993), ou à simulação (SHANNON, 1975; PEGDEN *et al*, 1995). Entretanto, conforme é visto a seguir, nos casos reais, algumas hipóteses adicionais podem ser validadas, e por meio delas, o problema original pode ser representado por modelos de filas mais simples, com soluções exatas fáceis de serem obtidas.

Modelo M/M/c/c/N

Por exemplo, se o processo de chegada for Poisson (i.e., os intervalos de tempo entre chegadas têm distribuição exponencial) e o processo de serviço for exponencialmente distribuído, então o modelo de filas M/M/c/c/N passa a representar adequada-

mente o sistema de acesso dos provedores (a notação M/M indica que os processos de chegada e serviço não têm memória). Isto traz enormes simplificações conforme visto a seguir, pois, trata-se de um modelo Markoviano com espaço de estados descrito apenas pelo número de usuários presentes no sistema.

Sejam λ_k, λ_k' e μ_k as taxas médias de chegada, entrada e serviço, respectivamente, no estado k, isto é, com k usuários. Como o espaço de estados é limitado em c usuários (devido à ausência de sala de espera), as taxas λ_k e λ_k' nem sempre são iguais - note que elas só coincidem quando $c \geq N$. A taxa média de chegada é uma função de N e k, por exemplo, $\lambda_k=\gamma(N-k)$, onde γN é a taxa média de chegada quando o sistema encontra-se vazio (i.e., k=0). A taxa média de entrada é definida por $\lambda_k'=\lambda_k(1-p_c)$ onde p_c é a probabilidade de perda do sistema (veja expressão (1) abaixo). E a taxa média de serviço é dada por: $\mu_k=k\mu$, onde μ é a taxa média de serviço por canal.

Dados γ, μ, c e N, a distribuição de equilíbrio para o sistema M/M/c/c/N é facilmente obtida por (KLEINROCK, 1975):

$$p_k = \frac{\binom{N}{k}\left(\frac{\gamma}{\mu}\right)^k}{\sum_{i=0}^{c}\binom{N}{i}\left(\frac{\gamma}{\mu}\right)^i}, \qquad k = 0, 1, ..., c \qquad \text{(M/M/c/c/N)} \qquad (1)$$

onde cada p_k indica a probabilidade de k usuários presentes no sistema. Em particular, quando k=c, tem-se a probabilidade de perda do sistema, p_c, ou seja, a probabilidade de um usuário, ao chegar, encontrar todos os servidores ocupados e ser impedido de entrar no sistema. Como a chegada é Poisson, essa probabilidade é igual a probabilidade de que todos os servidores estejam ocupados num certo instante (PASTA - Poisson Arrivals See Time Average). Portanto, p_c descreve também a fração de tempo em que todos os servidores estão ocupados.

Outras medidas de desempenho podem ser facilmente obtidas a partir de p_k, por exemplo, a probabilidade do sistema vazio p_0, ou do número médio de usuários no sistema, dado por:

$$E[L] = \sum_{k=0}^{c} k p_k = \sum_{k=0}^{c} k \frac{\binom{N}{k}\left(\frac{\gamma}{\mu}\right)^k}{\sum_{i=0}^{c}\binom{N}{i}\left(\frac{\gamma}{\mu}\right)^i} \qquad (2)$$

ou o índice de congestionamento do sistema, ρ, que pode ser calculado por E[L]/c.

Modelo M/M/c/c

No modelo M/M/c/c/N, a taxa média de chegada, λ_k, em cada estado k, é função do número de usuários no sistema. Se N for suficientemente grande em relação a c, tal que a taxa média de chegada seja aproximadamente independente do estado k (i.e., $\lambda_k \approx \lambda$), o sistema pode ser representado por um modelo de filas ainda mais simples: o modelo M/M/c/c. Neste caso, o tamanho da população é considerado, por simplicidade, infinito.

A figura 1 a seguir ilustra os números médios de usuários E[L] obtidos pelo modelo M/M/c/c/N para diferentes valores de N.

No caso, utilizamos λ = 1,2014 usuários/minuto, μ = 0,0359 usuários/minuto e c = 58. Note na figura que, à medida que N cresce, a aproximação pelo modelo M/M/c/c tende a ser exata. Em particular, para N = 800 usuários, obtemos E[L] = 33,43 (M/M/c/c) e E[L] = 32,09 (M/M/c/c/N), um erro de menos de 4,2% em relação ao valor exato.

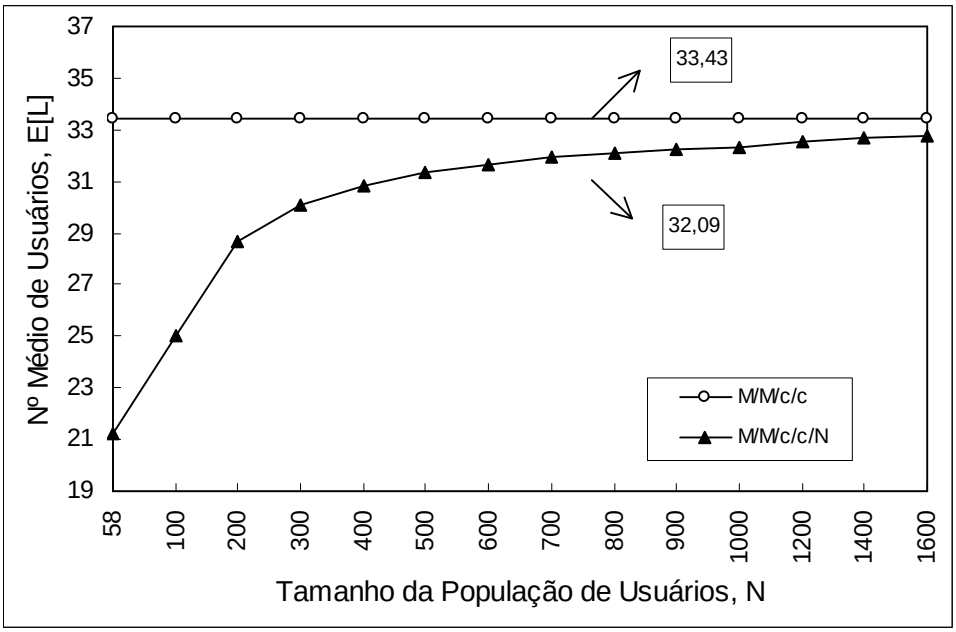

Figura 1: Relação entre o tamanho da população de usuários e o número médio de usuários no sistema, para os modelos M/M/c/c e M/M/c/c/N.

O modelo M/M/c/c também é um sistema Markoviano com taxa média de chegada $\lambda_k=\lambda$, taxa média de entrada $\lambda_k'=\lambda'=\lambda(1-p_c)$, e taxa média de serviço $\mu_k=k\mu$, para k=0,1,...,c. Dados λ, μ e c, as probabilidades de equilíbrio para este sistema são (KLEINROCK, 1975):

$$p_k = \frac{(\lambda/\mu)^k / k!}{\sum_{i=0}^{c} (\lambda/\mu)^i / i!}, \qquad k = 0, 1, ..., c \qquad \text{(M/M/c/c)} \tag{3}$$

Assim como no sistema M/M/c/c/N, medidas de desempenho, tais como a probabilidade de perda p_c, a probabilidade de encontrar o sistema vazio p_0, a probabilidade de acesso $1-p_c$, e o número médio de usuários no sistema E[L], também podem ser facilmente calculadas. Por exemplo:

$$E[L] = \sum_{k=0}^{c} k p_k = \sum_{k=0}^{c} k \frac{(\lambda/\mu)^k / k!}{\sum_{i=0}^{c} (\lambda/\mu)^i / i!} \tag{4}$$

e o índice de congestionamento do sistema ρ é dado por (lei de Little):

$$\rho = \lambda'/c\mu = \lambda(1-p_c)/c\mu \quad (5)$$

Como em grande parte das empresas provedoras de acesso à *INTERNET* o número de usuários cadastrados N é bem maior que o número de canais de atendimento disponíveis c (PARODI, 1996), isso sugere que supor que N seja suficientemente grande parece ser razoável.

Modelo M/G/c/c

Os modelos M/M/c/c/N e M/M/c/c assumem que o processo de chegada seja Poisson e os tempos de serviço sejam exponencialmente distribuídos. Nos sistemas reais de acesso dos provedores, os processos de chegada em geral podem ser razoavelmente bem descritos pelo processo de Poisson, entretanto, os tempos de serviço não são exponencialmente distribuídos (veja seção 5). Portanto, se o tamanho da população for suficientemente grande, o modelo M/G/c/c parece ser o mais indicado para representar o sistema de acesso dos provedores.

O modelo M/G/c/c, com espaço de estados igual ao número de usuários no sistema, não possui a propriedade Markoviana. Entretanto, conforme provado em GROSS & HARRIS (1974), os modelos M/G/c/c e M/M/c/c curiosamente têm a mesma distribuição de equilíbrio p_k definida em (3). Este importante e surpreendente resultado mostra que p_k em (3) é válida, mesmo que os tempos de serviço não sejam exponencialmente distribuídos. Note que isso amplia consideravelmente o universo de provedores que podem ser analisados de forma exata por meio de (3). Note também que p_k em (3) envolve apenas a taxa média de serviço μ por canal, ou seja, não é preciso conhecer os demais momentos da distribuição do processo de serviço. As medidas de desempenho para o modelo M/G/c/c são facilmente calculadas, da mesma forma que para o modelo M/M/c/c.

4. Metodologia

Os provedores de acesso à *INTERNET* precisam determinar qual o mínimo número de canais de atendimento necessário para oferecer um certo nível de serviço desejado aos usuários. Ou, inversamente, qual o nível de serviço alcançado com um dado número de canais de atendimento. Conforme discutido na seção 2, neste trabalho foi escolhida a probabilidade de acesso, ou a probabilidade de um usuário encontrar um canal disponível, para representar o nível de serviço.

Inicialmente escolhe-se o período do dia no qual o estudo será realizado pois, como será visto mais adiante, o comportamento dos usuários ao longo do dia é bem variável. Definido o período de estudo, analisa-se os processos de chegada e serviço no período. Após essa análise, procede-se à escolha de um modelo de filas apropriado. Nesta etapa verifica-se as hipóteses simplificadoras discutidas na seção anterior que, se satisfeitas, permitem a utilização de modelos de filas com soluções mais simples de serem obtidas. Após a modelagem, medidas de desempenho são estimadas e relacionadas em curvas de *tradeoff*, que permitem entender melhor o comportamento atual e futuro do sistema. Tais curvas descrevem de forma quantitativa os *tradeoffs* entre as medidas de desempenho do sistema (BITRAN & MORABITO, 1996).

Nos provedores são registrados os instantes de conexão e desconexão dos usuários para calcular os tempos de conexão e fazer a cobrança. Para decidir o período de estudo, os instantes de conexão e desconexão, juntamente com o número de canais de atendimento ocupados nesses instantes, são plotados em gráficos (ao longo do dia). Como cada canal ocupado corresponde a um usuário presente no sistema, esses gráficos descrevem o número de usuários presentes no sistema ao longo do dia (figura 2). É recomendável que seja traçado um número de gráficos tal que, após a eliminação dos dias atípicos (por exemplo, relacionados com uma queda no sistema), exista um número de gráficos ainda suficiente para garantir uma amostra significativa (por exemplo, 10 a 30 gráficos). Os dias escolhidos devem ser tais que alterações que ocorram no tamanho da população de usuários não sejam significativas a ponto de interferirem na taxa média de chegada dos mesmos. Assume-se que os dias sejam independentes entre si.

Conforme pode ser observado na figura 2, o comportamento dos usuários varia bastante ao longo do dia. A escolha do período (ou períodos) de análise deve limitar-se a um intervalo no qual o comportamento dos usuários possa ser considerado aproximadamente constante, para garantir que as taxas médias de chegada e serviço possam ser consideradas aproximadamente estacionárias. Em geral escolhe-se período(s) de pico - a escolha do(s) período(s) é feita pela comparação visual entre todos os gráficos da amostra.

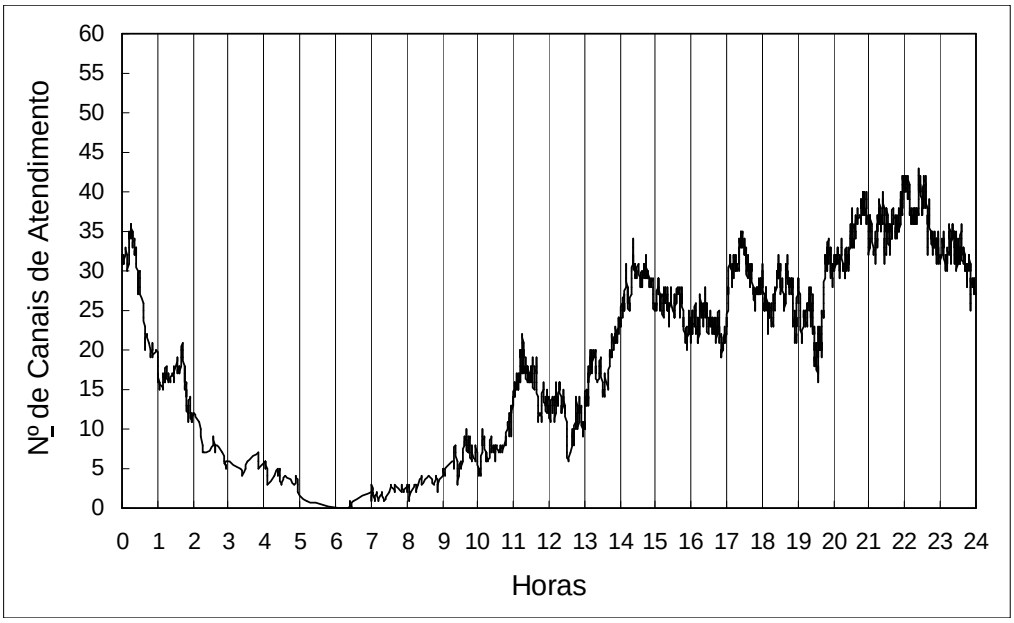

Figura 2: Número de canais de atendimento ocupados ao longo do dia.

Escolhido o período de análise, os próximos passos constituem em analisar os processos de chegada e serviço dos usuários. Para análise do processo de chegada, é feito o levantamento dos intervalos de tempo entre chegadas (ou chamadas), e para os tempos de serviço, o levantamento dos tempos de conexão. A título de ilustração, as figuras 3 e 4 mostram os processos de chegada e serviço correspondentes ao período das 20h às 23h do gráfico da figura 2.

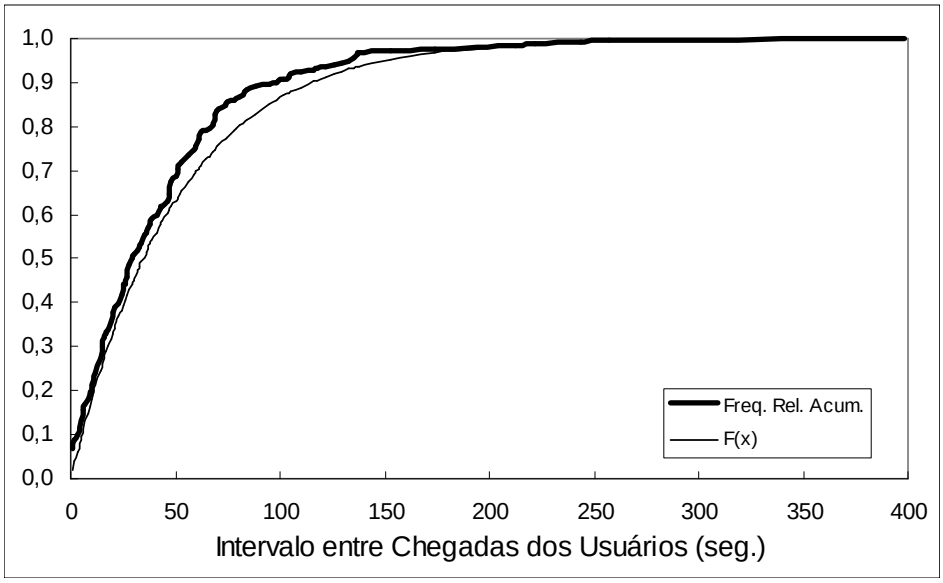

Figura 3: Freqüência relativa acumulada amostral e respectiva Distribuição Exponencial Acumulada, F(x), associada para os intervalos de tempos entre chegadas dos usuários do período das 20h às 23h, correspondente ao gráfico da figura 1.

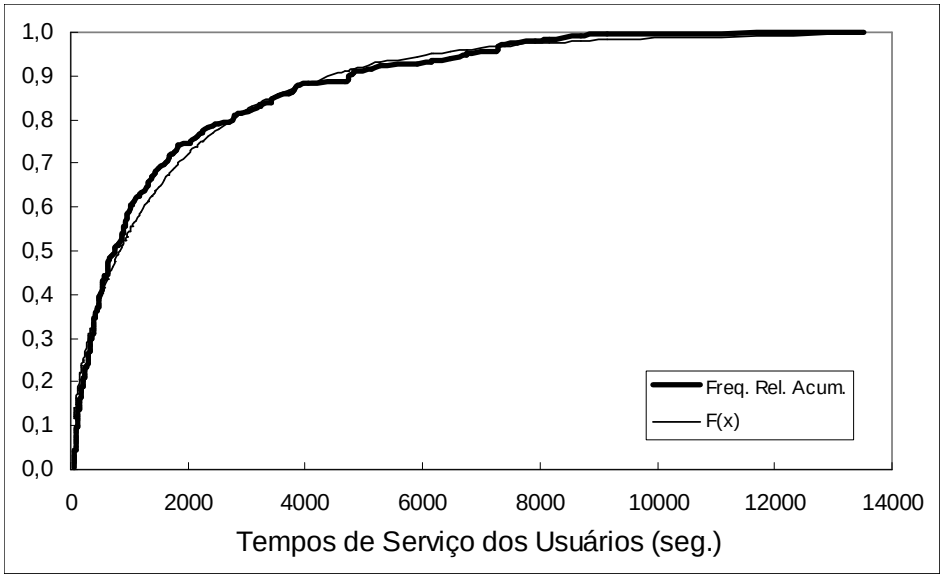

Figura 4: Freqüência relativa acumulada amostral e respectiva Distribuição Gama Acumulada, F(x), associada para tempos de serviço dos usuários do período das 20h às 23h, correspondente ao gráfico da figura 1.

Durante o período de análise, se o número de usuários que não consegue se conectar é bem pequeno, isto é, se a perda de usuários é praticamente nula, então a taxa média de entrada λ', é bem próxima da taxa média de chegada λ. Neste caso, a taxa média de chegada é estimada simplesmente pelo inverso da média dos intervalos de tempo entre conexões ocorridas durante o período. A taxa média de entrada é então obtida por $\lambda' = \lambda(1-p_c)$, onde p_c é a probabilidade de perda calculada conforme (3). A taxa média de serviço μ é obtida pelo inverso da média dos tempos de conexão levantados durante o período.

Se a perda de usuários no período é significativa (i.e., o sistema fica bem con-

gestionado no período), então é necessário, além dos intervalos de tempo entre conexões, levantar também os instantes em que os usuários tentam e não conseguem conexão. Dado que no provedor são registrados somente os instantes em que ocorrem conexões, uma alternativa é alocar um dos canais de atendimento exclusivamente para registrar aqueles instantes em que ocorrem perda de usuários. Por meio desse procedimento, todos os intervalos de tempo entre chegadas podem ser então levantados. Após a obtenção da taxa média de chegada por meio do inverso da média de todos os intervalos de tempo entre chegadas, estima-se a taxa média de entrada conforme anteriormente.

Assume-se que os intervalos de tempo entre chegadas e os tempos de serviço são independentes e identicamente distribuídos durante o período. Para cada um dos dias da amostra coletada, testes de aderência (p.e., teste de Kolmogorov-Smirnov) são realizados para escolher as distribuições de probabilidade que melhor se ajustam aos dados dos intervalos de tempo entre chegadas e dos tempos de serviço. Após a escolha das duas distribuições e seus respectivos parâmetros, estima-se os parâmetros médios dos dias da amostra por meio de média aritmética simples. Intervalos de confiança também são determinados para os parâmetros médios estimados, para analisar a sensibilidade da estimação. Para validar a hipótese de que os dias da amostra são identicamente distribuídos, pode-se utilizar o teste bilateral de Smirnov para n amostras (CONOVER, 1971).

Na aplicação de modelos de filas para projeto e dimensionamento de provedores, o interesse maior é no comportamento em equilíbrio destes sistemas, quando os efeitos das condições iniciais podem ser desconsiderados. Portanto, é necessário verificar se o sistema atinge equilíbrio durante o período de análise, o que pode ser feito por meio de simulação. É possível, então, relacionar graficamente o número de usuários presentes no sistema (ou o número de canais de atendimento ocupados) com o tempo total simulado e, desta forma, verificar graficamente se é razoável assumir que o sistema entra em regime durante o período de análise. Por segurança, a simulação do sistema pode ser feita partindo-se da situação mais desfavorável possível, ou seja, do sistema vazio, o que em geral não ocorre na prática.

Com as hipóteses de equilíbrio e de igualdade das distribuições validadas, e com os parâmetros médios dos processos de chegada e serviço, o número de canais de atendimento, e o tamanho da população de usuários, o modelo de filas escolhido é utilizado para estimar medidas de desempenho do sistema, como a probabilidade de perda, a probabilidade de acesso, o número médio de usuários, o índice de congestão do sistema. Estas medidas são, então, relacionadas por meio de curvas de *tradeoff*, que permitem fazer análise de sensibilidade para prever o comportamento do sistema sob diferentes configurações. Por exemplo, avaliar como o nível de serviço varia em função de diferentes números de canais de atendimento, ou em relação à diferentes tempos médios de serviço de usuários. Para maiores detalhes desta metodologia, veja FONTANELLA (1997).

5. Estudo de Caso e Análise dos Resultados

Um estudo de caso foi realizado em um provedor *INTERNET* localizado no interior do estado de São Paulo. Na época em que a coleta de dados foi realizada (outubro e novembro de 1996), a empresa contava com uma população de aproximadamente N=800 usuários e possuía c=58 canais de atendimento. Foram coletados os dados de 45 dias, dentre os quais, escolheu-se 30 dias considerados mais

representativos. Gráficos como o da figura 2, relacionando o número de canais de atendimento ocupados e os instantes de conexão e desconexão, foram traçados para escolha do período de análise. No caso, escolheu-se o período de maior congestionamento dos canais, isto é, o período de pico, que correspondeu ao intervalo das 20h às 23h.

Para o levantamento dos tempos de serviço dos usuários, foram coletados os tempos de conexão dos mesmos nos 30 dias, durante o período definido. Curiosamente a perda de usuários foi praticamente nula durante esse período, isto é, o número de usuários no sistema quase nunca atingiu a capacidade do mesmo (veja p.e. a figura 2). Portanto, para obter a taxa média de chegada, levantou-se apenas os intervalos de tempo entre conexões, não sendo necessário alocar um *modem* especialmente para coletar os instantes em que ocorreram perda de usuários.

Em seguida, procedeu-se à análise estatística dos processos de chegada e serviço. Foram traçados os histogramas de freqüência para os 30 dias e, por simples análise gráfica, escolheu-se algumas distribuições de probabilidade, que poderiam melhor representar os dados. As distribuições candidatas, tanto para os intervalos de tempo entre chegadas quanto para os tempos de conexão, foram: exponencial, Weibull, gamma e lognormal. Após alguns testes iniciais, a distribuição exponencial pôde ser logo descartada para os tempos de conexão. Aplicou-se o teste de aderência de Kolmogorov-Smirnov em cada um dos 30 dias, para escolher a distribuição mais ajustada aos dados. Para isso, utilizou-se o módulo estatístico do *software* de simulação Arena, versão 1.24 (PEGDEN *et al*, 1995).

Não se pôde rejeitar, com nível de significância 5%, a hipótese de que os intervalos de tempo entre chegadas fossem exponencialmente distribuídos, e portanto, assumiu-se um processo de chegada Poisson. O intervalo de tempo médio entre chegadas $1/\lambda$ resultou em 49,94 segundos, ou 0,8323 minutos. Esse valor foi obtido simplesmente pela média dos intervalos médios estimados em cada um dos 30 dias da amostra. A taxa média de chegada de usuários λ resultou em 0,0200 usuários/segundo, ou 1,2014 usuários/minuto. Os intervalos de 99% de confiança para $1/\lambda$ e λ foram [0,7663; 0,8983] minutos e [1,1131; 1,3049] usuários/minuto, respectivamente. Note que os intervalos são pequenos.

O mesmo procedimento foi realizado para o processo de serviço. Não se pôde rejeitar, com 5% de significância, a hipótese de que os tempos de serviço fossem distribuídos conforme uma distribuição gamma. Os parâmetros de escala β e de forma α foram obtidos pela média dos parâmetros estimados em cada um dos 30 dias. Esses valores resultaram em 2968,67 segundos (ou 49,4778 minutos) e 0,56, respectivamente, o que resulta num tempo médio de conexão de $\beta\alpha$=1669,63 segundos, ou 27,8272 minutos. A taxa média de serviço $\mu=1/\alpha\beta$ resultou em 0,0006 usuários/segundo, ou 0,0359 usuários/minuto. Os intervalos de 99% de confiança para $\alpha\beta$ e μ foram [26,1459; 29,5084] minutos e [0,0339; 0,0382] usuários/minuto, respectivamente. Note que os intervalos também são pequenos.

A análise estatística acima assume que os intervalos de tempo entre chegadas e os tempos de serviço sejam independentes e identicamente distribuídos nos 30 dias. Pelas características do sistema modelado, parece ser razoável assumir que essas variáveis sejam independentes nos 30 dias, mas para verificar se elas são de fato identicamente distribuídas nos 30 dias, utilizou-se o teste bilateral de Smirnov com n=30 amostras (CONOVER, 1971). Em ambos os casos, o teste não rejeitou essa hipótese para um nível de significância de 1% e, portanto, os intervalos de tempos entre chegadas e os tempos de conexão foram considerados

identicamente distribuídos nos 30 dias (para detalhes desse teste, veja FONTANELLA, 1997).

Assim, o modelo de filas M/G/c/c discutido na seção 3, com: (i) intervalo de tempo entre chegadas exponencialmente distribuído com parâmetro $1/\lambda$=0,8323 minutos, (ii) tempo de conexão com distribuição gama com parâmetros β=49,4778 minutos e α=0,56, (iii) c=58 canais de atendimento, e (iv) tamanho N da população considerado suficientemente grande, parece ser adequado para analisar o período das 20h às 23h no provedor em estudo. Para verificar se a hipótese de que o sistema atinge equilíbrio é razoável, o sistema foi simulado durante estas 3 horas, partindo da situação mais desfavorável possível, ou seja, do sistema vazio. O *software* utilizado para a simulação foi o Arena, referido anteriormente. Por meio de um gráfico relacionando o número de usuários presentes no sistema e o tempo simulado, foi possível verificar que, mesmo partindo do estado vazio, o sistema atingiu equilíbrio após cerca de 1,5 horas, o que sugere que é razoável utilizarmos medidas de equilíbrio para analisá-lo.

Por meio deste modelo, obtivemos as seguintes medidas de desempenho: número médio de usuários no sistema E(L)=33,46, conforme expressão (2), índice de congestionamento ρ=0,5764, conforme (5) (o que indica que, em média, os servidores estavam ocupados 58% do tempo), e probabilidade de perda do sistema $p_c \approx 0$, conforme (3) (o que eqüivale a um nível de serviço praticamente igual a 1). Os resultados indicaram uma alta ociosidade dos canais de atendimento, o que implica num alto nível de serviço oferecido aos usuários naquele período.

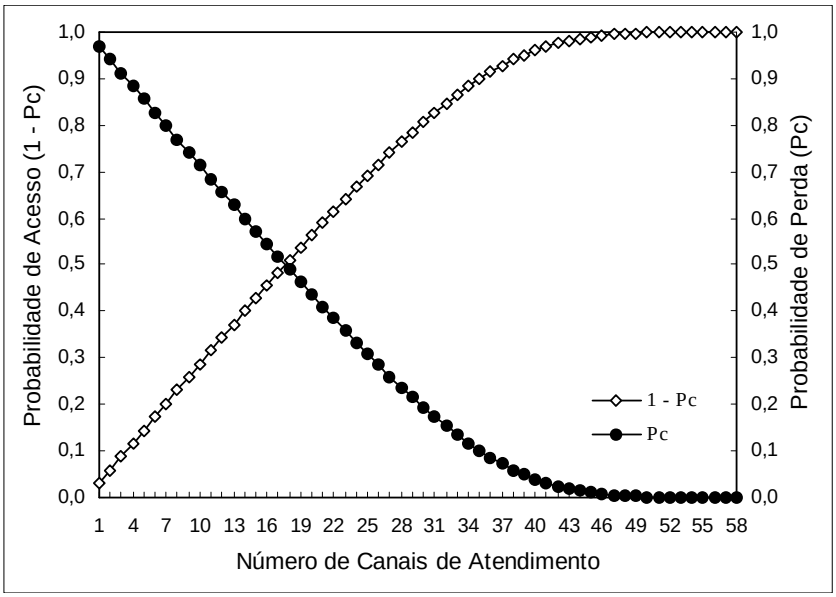

Figura 5: Curva de *tradeoff* entre a capacidade do sistema e o nível de serviço oferecido, e entre a capacidade do sistema e a probabilidade de perda de usuários.

Com a finalidade de avaliar melhor o comportamento presente e futuro do sistema, algumas curvas de *tradeoff* e relações entre as diversas medidas foram traçadas. A figura 5 ilustra como o nível de serviço varia em função do número de canais, para λ=1,2014 usuários/minuto e μ=0,0359 usuários/minuto. Note que, para 58 canais de atendimento, a probabilidade de perda do sistema é praticamente igual a zero, o que corresponde a um nível de serviço praticamente igual a 1. Por meio deste gráfico, é possível prever o

que acontece com o nível de serviço caso o provedor decida, por exemplo, alugar alguns canais para outros provedores. Note que, se esse número passar de 58 para 40, o nível de serviço ainda continua alto, o que indica que alguns canais de atendimento podem ser alugados, sem praticamente nenhum prejuízo ao nível de serviço oferecido aos usuários.

Na figura 6 podemos observar que, para 58 canais de atendimento, o índice de congestionamento do sistema é 0,58. Se o número de canais for reduzido para 40, por exemplo, o índice de congestionamento aumenta para aproximadamente 0,80, o que indica que os canais passam de uma média de ocupação de aproximadamente 60% para 80%. Isto implica em um melhor aproveitamento dos mesmos. Para 58 canais, o número médio de usuários no sistema é de 33,46, e para 40, é de 32,16, supostos λ=1,2014 e μ=0,0359 usuários/minuto.

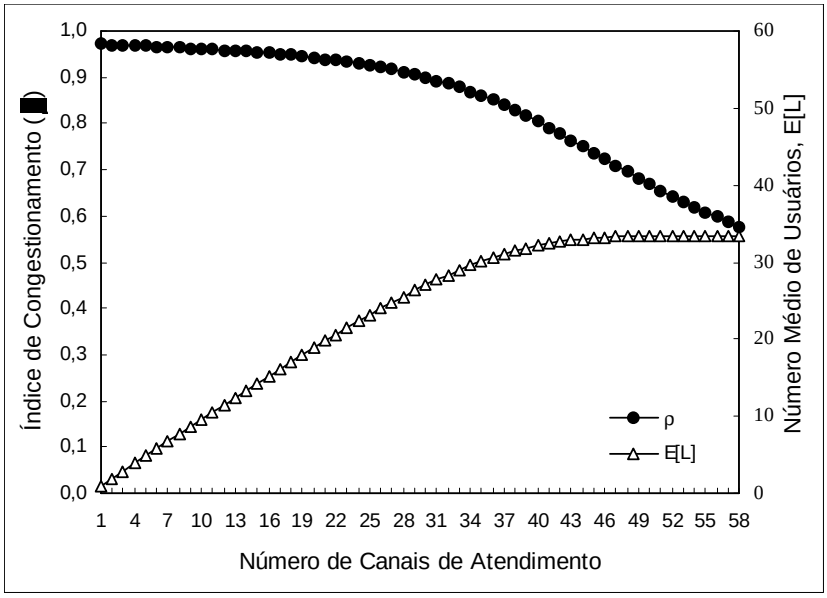

Figura 6: Relação entre a capacidade do sistema e o índice de congestionamento do sistema, e entre a capacidade dos sistema e o número médio de usuários.

Uma segunda análise que pode ser feita é em relação à taxa média de chegada de usuários, λ, conforme ilustra a figura 7. De acordo com a análise anterior, sabe-se que a ociosidade do sistema é alta e, portanto, algumas alternativas podem ser exploradas para diminuir essa ociosidade. Por exemplo, pode ser feita uma campanha publicitária agressiva para aumentar a população de usuários e, conseqüentemente, λ. Mantidos os valores de c=58 canais e μ=0,0359 usuários/minuto, mesmo que a taxa média de chegada dobre, note na figura 7 que, o nível de serviço do sistema diminui de 1 para aproximadamente 0,81, o que implica em uma probabilidade de perda de apenas 0,19. Por meio dessa análise, é possível concluir que um aumento razoável na taxa de chegada faz com que a ociosidade dos canais de atendimento seja reduzida, sem prejudicar sensivelmente o nível de serviço oferecido aos usuários. Vale a pena observar que, apesar deste fato ter ocorrido durante o período analisado no provedor em questão, acreditamos que na maior parte dos provedores isto não deva se verificar, principalmente nos períodos de pico.

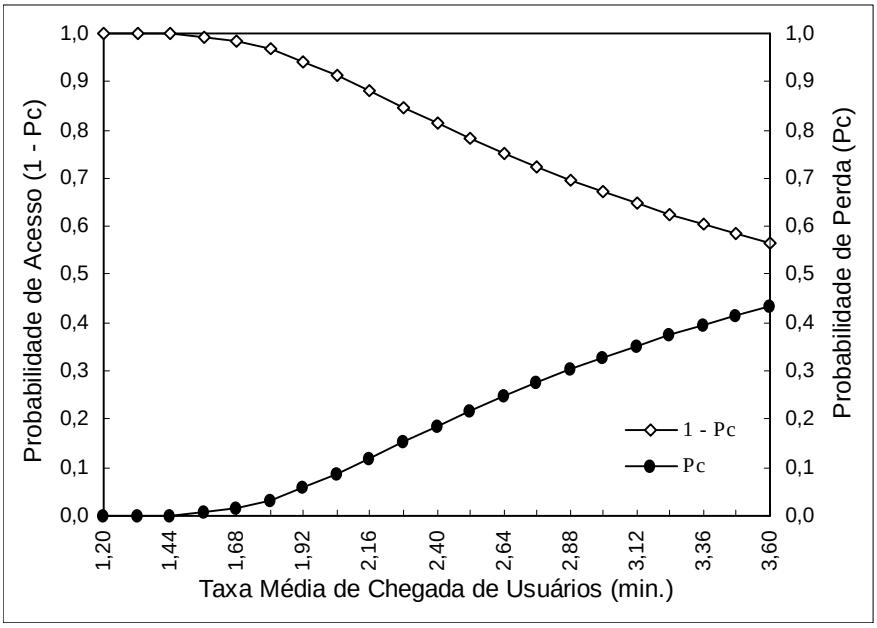

Figura 7: Relação entre a taxa média de chegada de usuários e o nível de serviço oferecido, e entre a taxa média de chegada de usuários e a probabilidade de perda de usuários.

Uma terceira análise foi feita em relação ao tempo médio de serviço, $1/\mu$. Para λ =1,2014 usuários/minuto e c = 58 canais, tem-se que μ = 0,0359 usuários/minuto, o que corresponde a um tempo médio de serviço de 27,8272 minutos. O provedor pode, por exemplo, considerar reduzir as tarifas para aumentar o tempo médio de conexão. Se, por exemplo, este tempo aumentar em aproximadamente 150%, isto é, passar de 27,8272 minutos para 69,46 minutos (i.e., μ = 0,0144), o índice de congestionamento do sistema passa de 0,58 para 0,97,

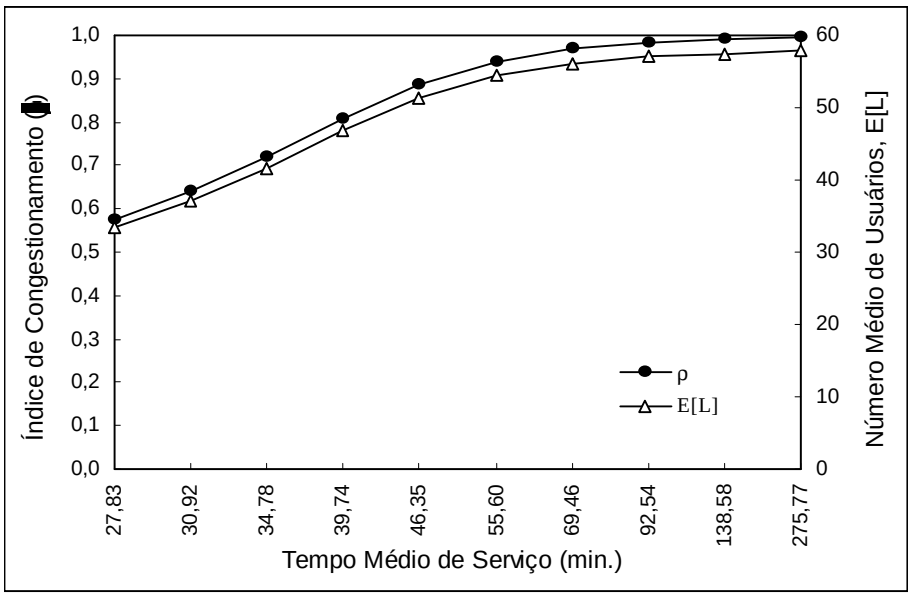

Figura 8: Relação entre o tempo médio de serviço de usuários e o índice de congestionamento do sistema, e entre o tempo médio de serviço e o número médio de usuários.

o que resulta num aumento significativo da utilização dos canais de atendimento, conforme pode ser verificado na figura 8. Note que o número médio de usuários no sistema passa de 33,46 para 56,12, para os mesmos 58 canais de atendimento.

Convém salientar que diversas questões podem ser respondidas por meio desta metodologia, fornecendo melhores *insights* sobre o comportamento atual e futuro do sistema, tais como: Qual o número médio de usuários no sistema? Qual o impacto no nível de serviço do sistema caso ocorra um aumento na taxa de chegada de usuários? Qual a proporção de tempo em que o sistema opera completamente congestionado? Qual o índice de congestionamento do sistema? Para maiores detalhes sobre o presente estudo de caso, veja FONTANELLA (1997).

6. Conclusões e Perspectivas

A teoria de filas pode ser uma ferramenta útil para modelar o sistema de acesso aos canais de atendimento de um provedor *INTERNET*. A modelagem por meio desta teoria pode auxiliar no projeto e configuração de provedores *INTERNET*, especialmente nas decisões que envolvam o *tradeoff* entre investir em capacidade e satisfazer o nível de serviço pretendido aos usuários. A metodologia utilizada é simples de ser implementada, e parece ser flexível o suficiente para ser aplicada em qualquer provedor de acesso à *INTERNET*. Para exemplificar a utilização da metodologia, foram apresentados os resultados de um estudo de caso realizado num provedor de médio porte no interior de São Paulo. O presente trabalho ilustra uma nova aplicação para a teoria de filas, pois, até onde temos conhecimento, nenhum outro trabalho do gênero foi encontrado na literatura especializada.

Perspectivas

Uma perspectiva deste trabalho é reanalisar o estudo de caso considerando os clientes particionados em múltiplas classes. As classes poderiam representar os diferentes comportamentos dos usuários quanto aos tempos de conexão, ou quanto aos intervalos de tempo entre chegadas. O particionamento pode ser uma boa estratégia em situações em que o coeficiente de variação é bem maior que 1, sugerindo que as distribuições dos intervalos de tempo entre chegadas ou dos tempos de conexão poderiam ser aproximadas por distribuições hiperexponenciais. Uma possível divisão de classes poderia ser usuários de *e-mail*, usuários de WWW, usuários de IRC, e demais usuários. Uma questão a ser respondida é quanto se ganha em termos de precisão de análise ao tratar os usuários em múltiplas classes, ao invés de apenas uma classe agregada.

Além do particionamento das classes de usuários, o particionamento dos canais de atendimento em função dessas classes é outra perspectiva a ser explorada. Um possível particionamento poderia ser, por exemplo, canais para usuários *vips* (especiais) e para usuários comuns. Essa prática é comum em balcões de atendimento de companhias aéreas e em supermercados com caixas rápido e caixas comuns. Exemplos de questões que poderiam ser úteis são: Quanto seria a tarifa cobrada dos usuários *vips*? Quantos canais de atendimento deveriam ser alocados para os *vips*? Qual o nível de serviço a ser oferecido aos usuários *vips*?

Outra perspectiva deste trabalho é realizar um estudo de caso em um provedor que o fenômeno da congestão esteja bem evidenciado, ou seja, exista pouca ociosidade dos canais de atendimento (isto é, as taxas de chegada e entrada de usuários sejam bem diferentes). Uma questão a ser respondida neste caso é se o processo de chegada continua sendo aproximadamente um processo de Poisson (homogêneo). Outra questão é quanto se ganha em precisão de análise ao aproximá-lo por um processo de Poisson heterogêneo, isto é, com taxa média de chegada não estacionária.

Os provedores estão preocupados com a relação ideal entre o número de usuários cadastrados na empresa e o número de canais de atendimento disponíveis. No estudo de caso realizado analisamos como o nível de serviço do sistema varia em função da taxa média de chegada de usuários ou do número de canais. Porém, não analisamos como o tamanho da população (N) de usuários está relacionado com a taxa média de chegada (λ). Uma outra perspectiva deste trabalho é derivar uma função que represente o relacionamento entre N e λ. Isto poderia ser feito, por exemplo, por meio de regressão não-linear. Com isso, definido o nível de serviço do sistema, poderíamos estimar qual a relação ótima entre N e c.

Agradecimentos

Os autores agradecem à Prof. Maria Cecília Mendes Barreto do DEs-UFSCar e aos três revisores pelos úteis comentários e sugestões, e ao Estevam Varga Junior e Waldemar Scudeller Jr. da Widesoft Sistemas Ltda, pelo apoio durante o desenvolvimento deste trabalho.

Referências Bibliográficas:

ANDRIES, E.: "O ano em que a Internet decolou no Brasil", *Internet World* 2(18), pp.66-70, fev. 1997.

BITRAN, G.R. & MORABITO, R.: "An overview of tradeoff curve analysis in the design of manufacturing systems", *Gestão e Produção* 3(2), p. 108-134, ago. 1996.

CHARLAB, S.: *Você e a Internet no Brasil*, Objetiva, Rio de Janeiro, 1995.

CHARLAB, S.: "Chegou a sua vez", *Internet World* 1(12), pp. 34-40, 42, ago. 1996

CONOVER, W.J.: *Practical Nonparametric Statistics*, J. Wiley, New York, 1971.

COX, D.R. & SMITH, W.L.: *Queues: monographs on applied probability and statistics*, Chapman and Hall, London, 1974.

FONTANELLA, G.C.: Uma metodologia baseada em teoria de filas para auxiliar no dimensionamento de canais de atendimento de provedores INTERNET, Dissertação de Mestrado, Departamento de Engenharia de Produção, UFSCar, São Carlos, 1997.

GROSS, D. & HARRIS, C.M.: *Fundamentals of Queueing Theory*, J. Wiley, New York, 1974.

KLEINROCK, L.: *Queueing systems: theory*. v.1, J. Wiley, New York, 1975.

PARODI, B.: "Guia de provedores de acesso. *Internet Brasil*", encarte especial 2(16), pp. 25-35, dez.1996.

PEGDEN, C.D.; SHANNON, R.E. & SADOWSKI, R.P.: *Introduction to simulation using siman*, McGraw-Hill, New York, 1995.

RAMOS, T.O.: "Upgrade da rede brasileira pode resolver congestionamento", *O Estado de São Paulo*, Caderno Informática, 1/07/1996.

SHANNON, R.E.: *System simulation: the art and science*, Prentice-Hall, Inc., New Jersey, 1975.

TIJMS, H.C.: *Stocastic modelling and analysis: a computational approach*, J. Wiley, 1986.

WHITT, W.: "Approximations for the GI/G/m Queue", *Production and Operations Management* 2(2), pp.114-161, 1993.

A QUEUEING MODEL TO ANALYSE THE TRADEOFF BETWEEN INVESTING IN ATTENDING CHANNELS AND SATISFYING SERVICE LEVEL IN INTERNET PROVIDERS

Abstract

The computer connection to INTERNET is provided by firms known as INTERNET access providers, which can handle several kinds of users by offering different forms to of physical connection. The simplest one is when the user connects to a provider's attending channel by an ordinary telephone line. Finding an available channel may not be an easy task, especially during peak hours. This results in a problem for the providers, which is to determine the optimal relationship between the number of users and the number of attending channels. To solve this problem, a provider needs to analyze the tradeoff between investing in capacity (number of channels) and satisfying the targeted service level to the users (probability of finding an available channel). The objective of this paper is to model this tradeoff by means of queueing theory. The methodology involves basically three steps: (i) analyze user arrival (calling) and service (attending) processes in chosen periods, (ii) select an appropriate queueing model under some simplifying assumptions, and (iii) generate tradeoff curves among system performance measures, in particular, between the probability of finding an available channel and the number of channels or the mean user arrival rate. To illustrate the application of this methodology, some results deriving from a case study performed in an access provider of São Paulo state, with Poisson arrival process but non-exponential service times, are presented. Some perspectives for future research are pointed out, such as grouping users into different classes as a function of their arrival and service behavior, and analyzing capacity partitioning as a function of these classes.

Key words: INTERNET provider, queueing theory, service level.