O PLANEJAMENTO ESTATÍSTICO DE CAPACIDADE NA ÁREA DE INFORMÁTICA

Renato de Paula Freitas Pereira - renato.de.paula@bol.com.br

Leydervan de Souza Xavier - xavierls@cefet-rj.br

José Luiz Fernandes - jlfernandes@cefet-rj.br

CEFET/RJ – CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA, Departamento de Ensino Superior.
Av. Maracanã, 229 – Bloco E – Sala 506
CEP 20271-110 – Rio de Janeiro - RJ

Resumo: *No planejamento da capacidade dos recursos necessários à produção na área de informática utiliza-se a projeção do recurso mais consumido no sistema, como por exemplo, o processador, sendo os demais recursos expandidos na mesma proporção. Dessa forma, o planejamento pode resultar em excessos, ou insuficiências, de capacidades nos recursos expandidos por falta de uma modelagem de performance do ambiente computacional mais confiável do que intuitivo. No mercado de software, a modelagem de performance é restrita a poucos produtos que são fornecidos em pacotes fechados, sem indicação do modelo teórico usado ou do grau de incerteza nas predições. Sendo assim, o ingresso no meio acadêmico favorece a estruturação e o aprendizado de ferramentas aos profissionais que trabalham na área computacional. Na universidade identificou-se dois modelos para solucionar o problema acima descrito: o primeiro, com base na teoria das filas, apresentado em 1984 e um desdobramento desta teoria utilizando-se cadeia de Markov, desenvolvido em 1994, que foi a técnica adotada. Sendo assim estudou-se as limitações deste modelo Markoviano, aplicado em estudos de caso em computadores de grande porte Unisys, de forma a verificar sua adequação.*

Palavras-chave: *Planejamento, Capacidade, Informática, Markov.*

1. INTRODUÇÃO

O planejamento é a parte do processo de melhoria em que são elaborados os métodos para alcançar objetivos e permitir atuar de forma pró-ativa por meio de ações preventivas. Na área computacional, o planejamento de capacidade traduz as necessidades futuras dos clientes em recursos computacionais, a fim de assegurar o contínuo atendimento aos níveis de serviços contratados, sem investimentos antecipados. Os investimentos devem ser postergados ao máximo possível para que não haja custos desnecessários, porém este tempo deve ser cuidadosamente analisado para que as implementações de recursos adicionais sejam transparentes aos usuários.

Quando o serviço oferecido é um sistema interativo (“on-line”), a qualidade do serviço é percebida em tempo real, medida pelo tempo de resposta do sistema para o usuário final. Nesse caso, na medida em que se tenha milhares de usuários simultaneamente conectados a nível nacional, o impacto da insatisfação do usuário gera custos adicionais significativos, além de uma imagem negativa e de difícil recuperação. Dessa forma, o processo de planejamento

da capacidade deve ser periódico, validando as predições feitas para pontos mais distantes através de verificações em períodos menores, reduzindo assim, a incerteza nas predições iniciais. Devido a este controle, o tempo hábil para a tomada de decisão, contratações e efetivações de ajustes na capacidade necessários se torna mais preciso.

De acordo com o exposto anteriormente, adotou-se a metodologia de planejamento de capacidade proposta por MENASCÉ *et al.* (1994), que trata os problemas de fila de espera formulados como cadeias de Markov, SHAMBLIN e STEVENS (1979) para a modelagem de performance e análise de agrupamento ("clustering"), seguida de análise de regressão linear no tratamento da carga, com ajustes de periodicidade.

No objeto de estudo deste trabalho a carga inicial é a de um sistema de grande porte que atende em horário comercial a maior parte do mercado de mutuários no sistema de financiamento da casa própria, e as projeções de seu crescimento são referentes à absorção de parcelas pequenas do mercado restante. As características da carga descrita serão monitoradas pela coleta de médias amostrais calculadas sobre períodos de poucos minutos, proporcionando dados de entrada em grande quantidade e normalidade, conforme o "Teorema Central do Limite", MONTGOMERY e RUNGER (2003), na sua distribuição de freqüências, o que viabiliza a aplicação de outras técnicas estatísticas, como por exemplo, de análise multivariada, eventualmente necessárias.

2. A METODOLOGIA DE PLANEJAMENTO DA CAPACIDADE

Nos dias atuais dificilmente uma empresa opera sem o apoio na informática. Em especial, as organizações prestadoras de serviços na área da tecnologia de informação têm grande dependência dos seus recursos de informática e, nesse caso, de tal forma que o planejamento de capacidade passa a ser o desdobramento das estratégias corporativas, que suporta a execução dos seus planos de negócios e o cumprimento das suas metas. Com este objetivo, o seguimento da metodologia de planejamento da capacidade, apresentada na Figura 1, possibilita a disponibilidade dos recursos conforme a demanda. Qualquer que seja o ramo de atividade da empresa, a atividade de planejamento da capacidade dos seus recursos computacionais deve responder a quatro questões básicas:

→Qual é a capacidade do sistema atualmente instalado?

→Quais são os crescimentos em serviços atuais e em novos serviços?

→Quais são os níveis de serviço contratados com os clientes?

→Qual é a configuração de menor custo capaz de processar os serviços atuais e futuros dentro dos níveis de serviço requeridos?

Essas respostas devem levar em consideração os planos corporativos, os sistemas de informação, os produtos de hardware e software em uso e os oferecidos no mercado, de forma a gerar configurações alternativas para realimentar periodicamente o planejamento estratégico da empresa.

A metodologia de planejamento apresentada pode ser decomposta em seis etapas:

1) *Conhecimento do ambiente de tecnologia de informação (TI)*: a metodologia de planejamento de capacidade começa pelo conhecimento global do ambiente computacional e das diretrizes corporativas da empresa;

2) *Caracterização da carga*: parte de uma visão geral das características principais para definir os componentes e parâmetros da carga atual e, em seguida, é feita a modelagem da carga usando os dados coletados com as ferramentas de monitoração do sistema;

3) *Calibração do modelo base*: a carga modelada é usada para parametrizar o modelo de performance base que é então resolvido analiticamente e cuja solução é comparada com as medidas feitas com as informações monitoradas no próprio sistema. A calibração do modelo normalmente requer várias modificações para ajustar iterativamente a carga e o modelo ao comportamento observado do sistema;

4) *Previsão da carga futura*: a demanda de carga futura é então projetada e aplicada no modelo de performance já calibrado, resultando no modelo de performance futura;

5) *Predição da performance futura*: neste passo é predita a performance futura do sistema computacional sob diferentes cenários considerando, por exemplo, expansões da capacidade em partes do sistema, crescimentos da carga e novos níveis de serviço. Desta predição obtém-se várias configurações possíveis de atender às exigências de carga, de tempo de serviço, etc.

6) *Configurações alternativas*: são identificadas para a escolha daquela que proporcione a melhor relação custo/benefício, em relação ao atendimento ao cliente.

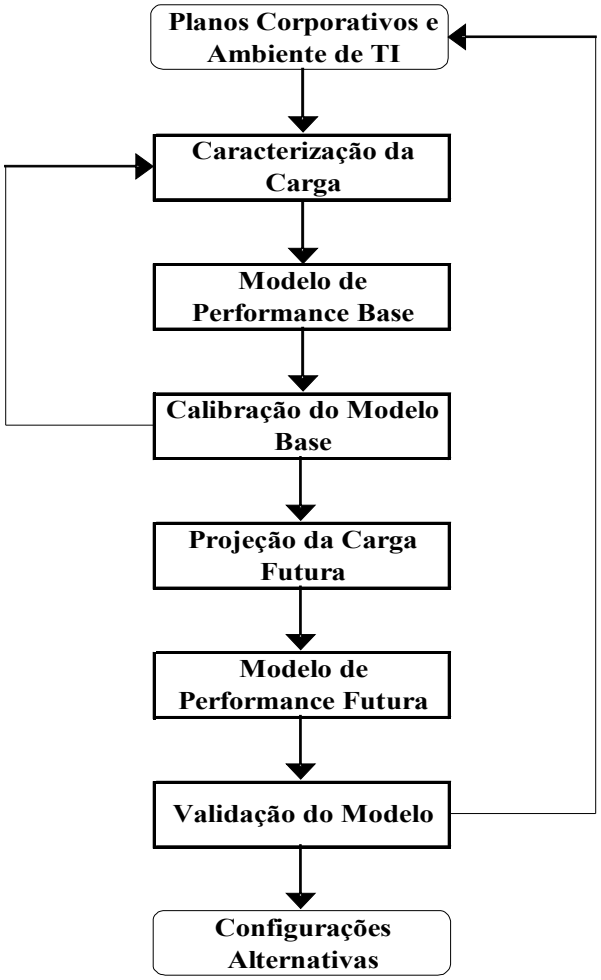

Figura 1 – Metodologia de planejamento da capacidade

O conjunto carga projetada e modelo de performance futura é gradualmente validado ao longo do tempo, utilizando periodicamente a metodologia acima descrita. Com isto, verifica-se o quanto as mudanças previstas na carga se realizaram e os respectivos resultados de performance ocorreram.

3. CARACTERIZAÇÃO DA CARGA

A caracterização da carga leva em conta o propósito a que se destina. Por exemplo, uma caracterização puramente funcional da carga, descreve o trabalho executado por um serviço ou uma transação sem conter qualquer informação quantitativa. Para fins de planejamento de capacidade, a caracterização deve ser em nível físico, descrevendo os consumos de recursos que têm impacto significante na performance do sistema, com informação quantitativa para aplicar em um modelo analítico.

De acordo com MENASCÉ *et al*. (1994) o termo "carga de trabalho" significa todas as requisições de processamento submetidas ao sistema pela comunidade de usuários durante algum período de tempo dado. Para a sua caracterização, MENASCÉ *et al*. (1994) define um processo cujo primeiro passo é a identificação dos componentes básicos da carga no sistema de computação.

Um componente básico é uma unidade genérica de trabalho que chega no sistema vinda de uma fonte externa. A natureza do serviço prestado pelo sistema determina o tipo de componente básico, isto é, um serviço ("job"), uma transação, um comando ou uma requisição, sendo os dois primeiros os mais comuns em sistemas de grande porte ("mainframes") com grandes cargas comerciais. Uma vez que os componentes básicos sejam

identificados, é necessário definir quais parâmetros caracterizam esses componentes. Esses parâmetros se dividem em dois grupos:

→Parâmetros de intensidade da carga: taxa de chegada de usuários, o número de terminais com seu tempo de resposta e o número de programas simultâneos em execução;

→Parâmetros relacionados com as demandas de serviços que cada componente básico requer de cada recurso do sistema. Estes conjuntos de parâmetros são representados por (D_{i1}, D_{i2}, ..., D_{ik}) onde D_{ij} significa o serviço demandado pelo componente *i* no recurso *j*.

O próximo passo é a definição das variáveis para a coleta dos dados em intervalos de tempo onde estes são consolidados nas medidas que serão armazenadas. Este passo inclui também a parametrização dos softwares monitores e a atribuição dos valores resultantes a cada parâmetro de caracterização adequado a cada componente da carga.

Para aumentar a representatividade da caracterização e a capacidade de predição da modelagem é feita pela subdivisão da carga em grupos de componentes homogêneos. Vários atributos podem ser usados como medida de similaridade para agrupar os componentes da carga. Estes atributos podem ser os seguintes:

→Tipo: este atributo pode ser *interativo*, ou terminal em linha, com tempos de consulta e respostas característicos; *transacional*, que chega no sistema com determinadas taxas, solicita transações específicas e deixa o sistema quando atendido; e *processamento em lote* ("batch");

→Aplicação, ou serviço utilizado: folha de pagamento; atualização de estoque; ponto de venda etc;

→Nível de consumo de recursos: leve; médio, e pesado.

Neste trabalho os componentes serão classificados por tipo, aplicação e nível de consumo. Utiliza-se a técnica estatística de agrupamento ("clustering") multivariado pelo método *K-médias*, JOHNSON e WICHERN (2002), que possibilita a classificação pelo consumo no espaço K-dimensional, a determinação prévia da quantidade de classes e ainda o cálculo dos valores médios dos parâmetros de demanda de recursos, para a caracterização de cada grupo. Uma análise estatística inicial também é feita para verificar as distribuições dos parâmetros, e as necessidades de transformações logarítmicas, de padronização e de tratamentos de valores discrepantes.

4. PROJEÇÃO DA CARGA DE TRABALHO FUTURA

Comumente, três problemas típicos acompanham a etapa do planejamento de capacidade para projeção da carga de trabalho futuro. O primeiro advém da dificuldade de obter informações confiáveis dos usuários devido a falta de conhecimento dos termos técnicos usados no detalhamento do consumo de recursos dos dispositivos do sistema. Assim sendo, é necessário ter o domínio da linguagem utilizada por usuários para se traduzir em consumo de recursos computacionais, as necessidades de crescimento e ajuste do uso atual da configuração instalada. Em segundo lugar vem a estimativa de recursos a serem consumidos por novos serviços ainda não totalmente desenvolvidos. Por último, a demanda reprimida em sistemas computacionais que estejam operando saturados, isto é, além dos limites de utilização.

Quando se expande um sistema que opera saturado, acompanhando o aumento da capacidade vem a melhoria nos níveis de serviço que desperta nos usuários a disposição de submeter muito mais serviços, que anteriormente eram programados para outros períodos, sujeitos a grandes filas de espera ou com tempo de resposta intolerável. Assim sendo, a demanda reprimida deve ser tratada com cuidado especial, pois pode ser de porte suficientemente grande para causar impacto desastroso nas previsões feitas.

De todo modo, ter entrevistas com os usuários, desenvolvedores de novos serviços e gerentes de contas, periodicamente, permite confirmar, ou melhor dimensionar, mudanças que dependem de acontecimentos futuros na medida que eles estejam mais próximos. Conseqüentemente, essas entrevistas são normalmente o melhor caminho para se obter as informações necessárias na estimativa das cargas futuras, inclusive quanto à demanda reprimida.

Segundo MENASCÉ e ALMEIDA (1998), a literatura descreve várias técnicas para projeção sendo que a escolha entre elas leva em consideração aspectos tais como: disponibilidade e confiabilidade de dados históricos; grau de incerteza; horizonte do planejamento; e algum padrão que se identifique nos dados históricos, além das condições das comunicações de dados. Para o caso de um padrão estacionário, por exemplo, focando uma aplicação de grande carga comercial com previsão de novos pequenos clientes, longo histórico confiável e horizonte distante a regressão linear é uma das técnicas adotadas. A análise de regressão é uma coleção de ferramentas estatísticas para encontrar as estimativas dos parâmetros no modelo de regressão. Então, essa equação (ou modelo) ajustada de regressão é tipicamente usada na previsão de observações futuras de *Y* ou para estimar a resposta média em um nível particular de *x*, MONTGOMERY e RUNGER (2003).

Mais especificamente, a aplicação da técnica de análise de regressão linear é usada para investigar as relações entre quantidade de usuários e consumo de recursos em um aplicativo de grande carga, de modo a possibilitar a projeção das cargas de pequenos novos clientes que venham a se somar às grandes cargas em serviço.

5. MODELAGEM DA PERFORMANCE

Um modelo é a uma forma de representar artificialmente um sistema real. No caso de computadores, na representação do comportamento em termos de performance, são usados modelos que permitam prever os valores de saída, que são as respostas dos sistemas computacionais, a partir dos parâmetros de entrada que descrevem a carga, o software básico e o hardware. As saídas são as medidas de performance que demonstram a reação do sistema quando submetido aos parâmetros de entrada.

Freqüentemente, os modelos de performance são construídos para analisar o efeito das variações nos parâmetros de entrada nas suas saídas. Assim, para estudar as mudanças de comportamento é construído um modelo inicial, ou o modelo de performance base, que precisa ser calibrado para garantir a representação do sistema com precisão aceitável antes de provocar variações de parâmetros na entrada do modelo.

A calibração do modelo base é feita comparando as medidas de performance geradas nele com as medidas feitas no próprio sistema, a partir dos mesmos parâmetros de entrada. Utilizando-se dados históricos, são aplicados dados conhecidos na entrada do modelo e as suas saídas são comparadas com os resultados também conhecidos para que sejam feitos os ajustes necessários. Verifica-se a cada ajuste o erro médio quadrático para então ser adotado o modelo que apresente o menor desvio da realidade.

Uma vez calibrado o modelo base, nele são alimentados os parâmetros da carga projetada tornando-o de performance futura, cujas saídas representam o comportamento futuro do sistema.

Nos modelos de performance analíticos são especificadas as interações entre os parâmetros de entrada e as medidas de saída por meio de fórmulas como, por exemplo:

$$T = \frac{S}{(1-\lambda S)} \quad . \tag{1}$$

A qual estabelece uma relação entre o tempo médio de resposta *T(s)*, o tempo médio de serviço *S(s)* e a taxa de chegada de clientes na fila $\lambda(s^{-1})$.

O desenvolvimento dos modelos de performance segue também uma metodologia conforme mostrado na Figura 2. A construção do modelo se inicia, necessariamente pela obtenção dos parâmetros e estabelecimento de premissas básicas. Nesse caso, optou-se pela exata de modelos analíticos aproximados, para os quais são obtidos, por exemplo, os tempos médios de serviço dos componentes nos vários recursos.

Após a construção e parametrização do modelo os parâmetros de entrada são computados de forma a resultar em medidas de performance, como utilização, rendimento ou tempo de resposta na saída e a calibração do modelo deve ser verificada com os ajustes necessários feitos e novas soluções do modelo ajustado. Dessa forma, as medidas de performance encontradas pela solução do modelo são comparadas com as medidas feitas diretamente no

sistema que está sendo modelado. Do resultado da comparação de valores o modelo pode ser considerado aceitável, ou não, em função da quantidade de erros.

O nível de erros aceitável é determinado a critério do pesquisador, ou definido à priori em conjunto com o contratante do serviço de planejamento. Segundo MENASCÉ *et al*. (1994) valores aceitáveis são de até 10% na utilização de dispositivos, 10% no rendimento do sistema e 20% nos tempos de resposta.

A partir do modelo base calibrado, o primeiro passo na construção do modelo de projeção é a alteração nos valores dos parâmetros que reflitam as modificações que se deseja estudar. Basicamente, uma expansão de capacidade em um dispositivo é representada pela mudança no respectivo tempo de serviço. Contudo, é necessário verificar cuidadosamente os efeitos que cada mudança possa ter em outros dispositivos e, se necessário, alterar seus parâmetros.

Em seguida, o modelo de projeção é resolvido gerando nas suas saídas as medidas de performance que representam o comportamento futuro das configurações definidas, de carga e de sistema. Várias configurações conjuntas de carga e sistema são projetadas e as alternativas com melhores relações custo/benefício, que se alinhem com o direcionamento global da organização são escolhidas.

O passo final se refere à validação gradual das projeções na medida que rotineiramente as mudanças projetadas na carga se concretizem, proporcionando novos ajustes, conforme toda a metodologia seja praticada periodicamente.

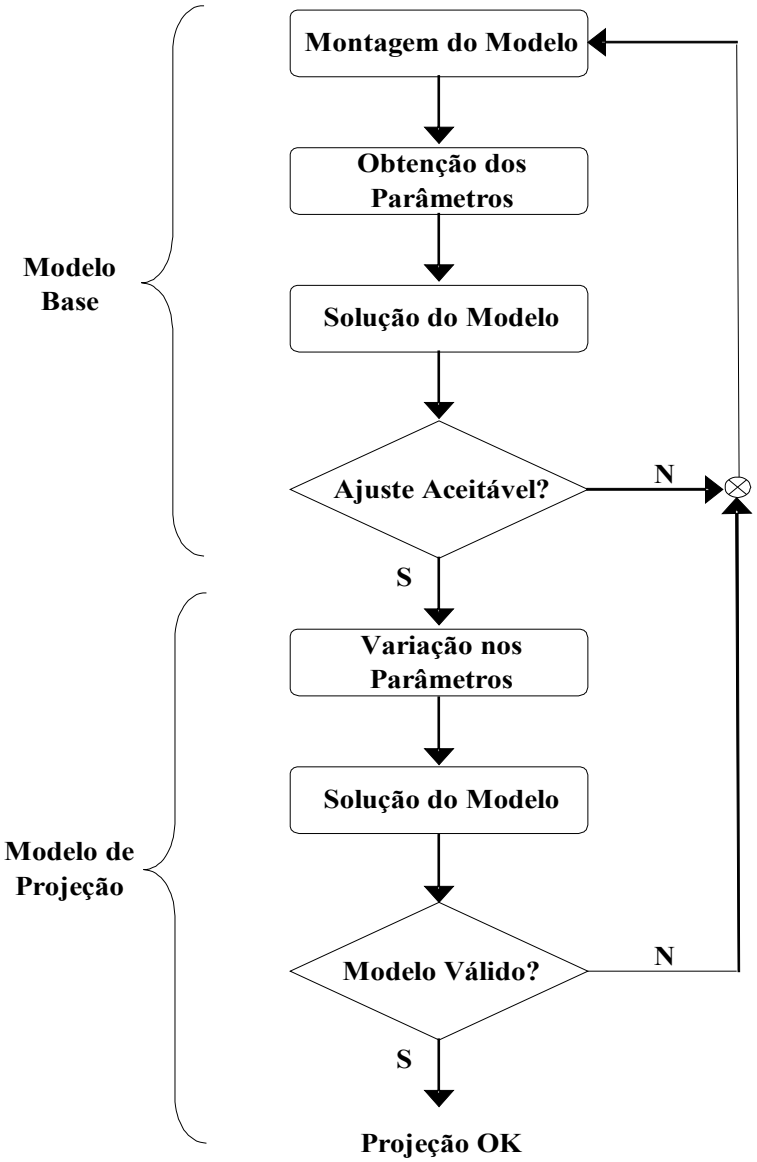

Figura 2 – Modelagem de performance

Na etapa específica da solução do modelo, o primeiro passo é construir um diagrama que descreva os estados em que pode ser encontrado o sistema, de tal forma, que uma vez paralisado, seja possível reiniciar o sistema exatamente do estado em que foi paralisado. Esta tarefa significaria registrar quantos clientes estão em cada dispositivo, quanto tempo os clientes que estão sendo atendidos terão ainda de serviço e todos os outros parâmetros que descrevam o comportamento do sistema. No entanto, a tarefa é reduzida por serem assumidas as seguintes premissas simplificadoras:

→Todos os clientes são iguais: não importa qual cliente está presente, mas sim apenas o número de clientes presentes. Esta é a premissa da carga homogênea;

→A história é irrelevante: não importa como o sistema chegou ao estado em que se encontra, ou seja, não interessa o tempo que cada cliente que está sendo atendido já foi servido. Se o sistema for observado antes de um cliente partir e, depois da sua partida, quando outro tiver chegado, não haverá distinção entre estes dois estados. Segundo CLARK e DISNEY (1979), esta é a premissa da falta de memória, que baseia a análise Markoviana. Em uma seqüência dependente de Markov, o conhecimento do presente torna o futuro independente do passado.

Montado o diagrama de estado, como mostrado na Figura 3, as equações de fluxo balanceado são determinadas a partir da condição alcançada quando, após longo tempo o sistema entra em estado estacionário levando em consideração que em cada estado, o fluxo total é nulo, isto é, o fluxo de entrada no estado é igual ao fluxo de saída deste mesmo estado.

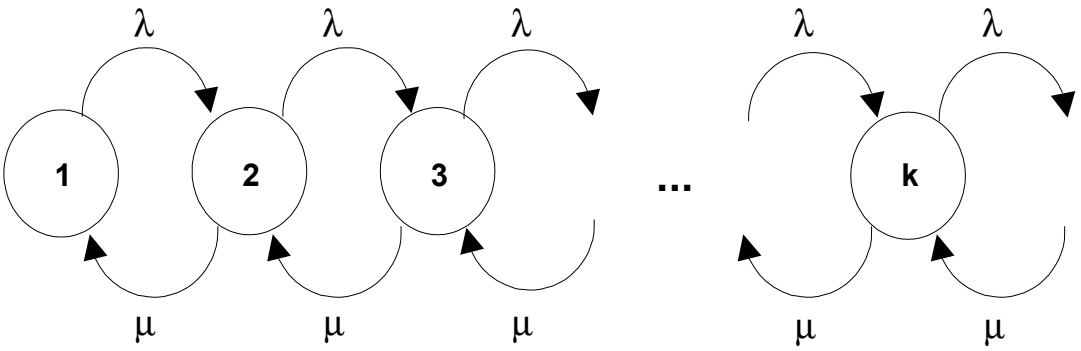

Figura 3 – Diagrama de estado para **k** clientes.

A equação geral de fluxo balanceado é:

$$\lambda P_{K-1} + \mu P_{K+1} = \lambda P_K + \mu P_K \quad (s^{-1}) \qquad (2)$$

Sendo λ e μ , respectivamente, as taxas de chegada e de atendimento aos clientes, **k** o número de clientes no sistema e P_K a probabilidade do sistema se encontrar no estado **k.**

Da equação geral dos fluxos balanceados, que expressa o equilíbrio dos fluxos em estado estacionário, e da equação de conservação da probabilidade total, que é a constatação de que em qualquer instante a soma das probabilidades do sistema ter todos os possíveis estados é sempre um, são deduzidas as expressões das medidas de performance desejadas. A conservação da probabilidade total é expressa pela seguinte equação:

$$P_0 + P_1 + P_2 + \ldots = 1 \qquad (3)$$

O modelo resultante de performance é constituído do conjunto de equações (4) a (7), cujas variáveis independentes são os parâmetros da carga, do software básico e do hardware e as variáveis dependentes são as medidas de performance.

$$\text{Utilização} = \frac{\lambda}{\mu} \qquad (4)$$

$$\text{Rendimento (“Throughput”)} = \mu * \frac{\lambda}{\mu} = \lambda \quad (s^{-1}) \tag{5}$$

$$\text{Comprimento da fila} = \frac{\lambda}{\mu - \lambda} \tag{6}$$

$$\text{Tempo de resposta} = \frac{1}{\mu - \lambda} \quad (s) \tag{7}$$

Por conseguinte, para a resolução do modelo de performance inicialmente é resolvido o sistema de equações (2) e (3) para então solucionar o conjunto de equações (4) a (7), deduzidas das medidas de performance.

6. CONCLUSÃO

A metodologia de planejamento proposta por MENASCÉ *et al.* (1994), apoiada nas técnicas estatísticas adotadas, permite desenvolver uma ferramenta com base em modelo analítico para o planejamento da capacidade na área de informática. A utilização dessa ferramenta com periodicidade possibilita, além da verificação da incerteza inerente ao modelo base, a validação progressiva das projeções de comportamentos futuros do sistema computacional modelado. Para cada cenário futuro, as características da carga e do sistema são traduzidas em um conjunto de equações específico.

Os resultados da solução desse conjunto de equações são os valores das medidas de performance futura para os quais o planejamento estatístico de capacidade na área de informática foi desenvolvido.

REFERÊNCIAS BIBLIOGRÁFICAS

CLARK, A.BRUCE; DISNEY, RALPH L. **Probabilidade e Processos Estocásticos**. Rio de Janeiro: Ed LTC, 1979.

JOHNSON, RICHARD A.; WICHERN, DEAN W. **Applied Multivariate Statistical Analysis**. Englewood Cliffs, New Jersey: Prentice-Hall, 2002.

LAZOWSKA, EDWARD D.; ZAHORIAN, JOHN; GRAHAM, G. SCOTT & SEVCIK KENNETH C. **Quantitative System Performance: Computer System Analysis Using Queueing Network Models**. Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

MENASCÉ, DANIEL A.; ALMEIDA, VIRGILIO A.F. **Capacity Planning for Web Performance: Metrics, Models, & Method*s***. Upper Saddle River, New Jersey: Prentice-Hall, 1998.

MENASCÉ, DANIEL A.; ALMEIDA, VIRGILIO A.F. & DOWDY, LARRY W **Capacity planning and performance modeling; from mainframes to client-server systems**. Englewood Cliffs, New Jersey: Prentice-Hall, 1994.

MONTGOMERY, DOUGLAS C. & RUNGER, GEORGE C **Estatística Aplicada e Probabilidade para Engenheiros**. Rio de Janeiro: EdLTC, 2003.

SHAMBLIN, JAMES E.; STEVENS, G.T. JR. **Pesquisa Operacional: Uma abordagem básica**. São Paulo: EdATLAS, 1979..

THE STATISTICAL CAPACITY PLANNING IN THE COMPUTER SCIENCE AREA

Abstract: *In the capacity planning of necessary resources to the production in the computer science area, the projection of the more consumed resource in the system, for instance, the processor, is being used to expand the other resources in the same proportion. In that way, the planning can result in excesses, or inadequacies, of capacities in the expanded resources by lack of a performance modeling of the computational environment more reliable than intuitive. In the software market, the performance modeling is restricted the few products that have been supplied in closed packages, without indication of the used theoretical model or of the uncertainty degree in the predictions. In this way, the academic medium favors structuring and learning tools to the professionals that work in the computational area. In the university two models to solve the problem described were identified: the first, with base in the Queueing theory presented by Lazowska et al. (1984) and an unfolding of this theory using Markov's chain developed by Menascé et al. (1994). In this work it was explored the limitations of the Menascé model for application in a case study with mainframes Unisys.*

Key-words: *Capacity planning, Computer science, Markov's chain.*