ISSN 0101-7438

MODELO HIPERCUBO: ANÁLISE E RESULTADOS PARA O CASO DE SERVIDORES NÃO-HOMOGÊNEOS

Fernando Chiyoshi
Roberto D. Galvão
Programa de Engenharia de Produção / COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro – RJ

Reinaldo Morabito
Departamento de Engenharia de Produção
Universidade Federal de São Carlos
São Carlos – SP

Recebido em 04/2001, aceito em 11/2001 após 1 revisão

Resumo

O objetivo do presente trabalho é analisar o uso e a solução do modelo hipercubo para o caso de servidores não homogêneos (servidores com diferentes tempos médios de serviço). Sistemas com servidores não homogêneos podem ser encontrados em diversas situações reais, como por exemplo nos Serviços de Atendimento Médico de Emergência (SAMU) de algumas cidades brasileiras. A importância de se considerar explicitamente a não homogeneidade dos servidores no modelo hipercubo é mostrada inicialmente através de um exemplo ilustrativo. É em seguida demonstrado que a solução para o caso não homogêneo pode ser obtida vantajosamente pelo método de Gauss-Siedel. Este método foi testado para uma rede de 55 vértices disponível na literatura, em modelos com 10 a 17 servidores, variando-se a taxa de ocupação do sistema de 0,1 a 0,9. Finalmente, propõe-se um modelo de regressão para estimar o tempo de processamento necessário para resolver um dado problema.

Palavras-chave: modelo hipercubo, servidores não-homogêneos, método de Gauss-Siedel.

Abstract

The objective of the present paper is to analyze the use and solution of the hypercube model for the case of non-homogeneous servers (servers with different mean service times). Systems with non-homogeneous servers can be found in several real world applications, such as for example in the provision of Emergency Medical Services (EMS) in some Brazilian cities. The importance of explicitly considering non-homogeneous servers in the hypercube model is initially demonstrated through an illustrative example. It is then shown that the solution for the non-homogeneous case can be advantageously obtained by the method of Gauss-Siedel. This method was tested for a network of 55 nodes, in models with between 10 and 17 servers, with the total system workload varying between 0.1 and 0.9. Finally, a regression model is proposed to estimate the computing time required to solve a specific problem.

Keywords: hypercube model, non-homogeneous servers, method of Gauss-Siedel.

1. Introdução

A partir dos anos 60, diversos estudos são encontrados na literatura de Pesquisa Operacional tratando problemas de localização probabilísticos dos serviços urbanos de emergência, tais como sistemas de atendimento médico, patrulhamento policial e combate a incêndios (bombeiros). Tais problemas envolvem decisões com respeito ao número e à localização das unidades de emergência (ambulâncias, viaturas de polícia, carros de bombeiros), além de políticas de despacho destas unidades para atender, o mais rápido possível, as chamadas de emergência. Uma importante ferramenta para analisar tais problemas é o modelo hipercubo, proposto por Larson (1974, 1975) e estudado por diversos autores (Swersey, 1994). Apesar de descritivo, o modelo pode ser utilizado em métodos de solução para problemas de localização probabilísticos (Batta *et al.*, 1989; Chiyoshi *et al.*, 2000). Ele é adequado para analisar sistemas coordenados, onde o usuário telefona para uma central de atendimento solicitando serviço e o administrador do sistema despacha a unidade disponível mais próxima do local da chamada para atendê-la. Caso nenhuma unidade esteja disponível a solicitação entra em uma fila de espera, para ser atendida assim que alguma unidade ficar disponível, ou é transferida para outro sistema de atendimento.

Segundo Swersey (1994), o modelo hipercubo tem por base principalmente as seguintes hipóteses: (i) a região de serviço é dividida em sub-regiões ou "átomos"; (ii) dentro de cada átomo as chamadas por serviço ocorrem de acordo com um processo de Poisson; (iii) os tempos de serviço de cada unidade são exponencialmente distribuídos; (iv) exatamente uma unidade é despachada para cada chamada; (v) associada a cada átomo há uma lista ordenada de unidades preferenciais; (vi) a unidade, quando disponível, está localizada em cada átomo de acordo com probabilidades especificadas; (vii) os tempos de viagem entre átomos são conhecidos ou estimados. Embora algumas destas hipóteses possam parecer muito restritivas, elas são satisfeitas em diversos sistemas reais.

A solução do modelo hipercubo envolve resolver um sistema de equações lineares que fornece as probabilidades de equilíbrio dos possíveis estados do sistema (*probabilidades de estado*). Essas probabilidades permitem estimar diversas medidas de desempenho interessantes para o gerenciamento do sistema, tais como cargas de trabalho (*workloads*) das unidades, tempo médio de resposta do sistema ou de cada unidade, freqüência de atendimento de cada unidade a cada região, entre outras.

Alguns exemplos de aplicação do modelo hipercubo nos Estados Unidos são a localização de ambulâncias em Boston (Brandeau & Larson, 1986), o patrulhamento policial em Orlando (Sacks & Grief, 1994), e um programa de visitas do serviço social (Larson & Odoni, 1981).

No Brasil, alguns exemplos são o atendimento a interrupções na distribuição de energia elétrica em Santa Catarina (Albino, 1994), a localização de ambulâncias em um trecho da BR-111 (Gonçalves *et al.*, 1994, 1995), o balanceamento da carga de trabalho de ambulâncias no sistema "Anjos do Asfalto" da Via Dutra (Mendonça, 1999; Mendonça & Morabito, 2000), e a configuração do Serviço de Atendimento Médico de Urgência (SAMU) da prefeitura de Campinas, SP (Takeda, 2000; Takeda *et al.*, 2000). Extensões do modelo hipercubo foram consideradas em Halpern (1977) e Burwell *et al.* (1993). Outras referências podem ser encontradas em Swersey (1994) e Chiyoshi *et al.* (2000).

Nosso interesse no presente trabalho é enfatizar a importância de se considerar, em algumas aplicações, servidores *não homogêneos* (servidores com diferentes tempos médios de serviço) no modelo hipercubo e estudar métodos de solução adequados para este caso. É importante notar que o método aproximado, desenvolvido por Larson (1975) para a solução

do modelo hipercubo, parte da hipótese da homogeneidade dos servidores, não podendo portanto ser aplicado ao caso não homogêneo. Quando se considera servidores não homogêneos, a solução do modelo hipercubo requer necessariamente a solução de um sistema de equações lineares. Pode-se, para tanto, usar o método de eliminação de Gauss. Trata-se de um método direto que, a menos de erros de arredondamento, produz resultados exatos em um número finito de etapas. Este método no entanto torna-se rapidamente inviável do ponto de vista computacional à medida que o porte do modelo aumenta.

O sistema de equações lineares associado ao modelo hipercubo envolve uma matriz de coeficientes que torna-se esparsa para problemas de maior porte, sugerindo que métodos iterativos podem ser mais indicados para resolvê-lo. Pelas características da matriz de coeficientes as indicações são de que, dos métodos iterativos simples mais conhecidos, o método de Gauss-Siedel é o mais apropriado para resolver o sistema de equações associado ao modelo hipercubo. Não obstante tais indicações, não é imediatamente óbvio que esse sistema satisfaz a condição de convergência do método de Gauss-Siedel. Ao invés de aprofundar a análise teórica do problema de convergência optou-se pela via experimental, com a aplicação do método a um conjunto de problemas teste com características passíveis de serem encontradas em situações práticas.

Deste trabalho experimental resultou uma base empírica capaz de evidenciar que o método de Gauss-Siedel, formulado conforme descrito abaixo, é um método adequado para resolver o sistema de equações lineares associado ao modelo hipercubo. Adequado no sentido que, mesmo usando recursos modestos de computação, problemas de porte relativamente elevado podem ser resolvidos através do método. A formulação do método de Gauss-Siedel para resolver o sistema de equações lineares associado ao modelo hipercubo, e a apresentação da evidência empírica de sua utilidade, são as contribuições mais importantes de nosso trabalho.

Sistemas com servidores não homogêneos podem ser encontrados em diversas situações reais. Por exemplo, nos Serviços de Atendimento Médico de Urgência (SAMU) de algumas cidades brasileiras, as ambulâncias podem ser veículos de suporte avançado (VSA ou UTI móveis), suporte básico (VSB), remoção simples (VRS), ou psiquiátrico (PSQ) (Takeda *et al.*, 2000). Estes servidores, mesmo compartilhando certa carga de trabalho, podem possuir tempos médios de serviço bastante diferentes em função do tipo de serviço oferecido.

Podemos por outro lado encontrar sistemas com servidores iguais do ponto de vista do tipo de veículo, equipe e equipamentos, porém com tempos médios de serviço diferentes em função de suas localizações. Por exemplo, nos Serviços de Atendimento aos Usuários (SAU) de algumas rodovias brasileiras, os tempos médios de serviço dos veículos podem variar devido a tempos médios de viagem diferentes, em função da localização das bases dos veículos ao longo da rodovia (Mendonça & Morabito, 2000). Nestes casos os tempos de viagem (deslocamento da base até o local do acidente, do local do acidente até o hospital e do hospital até a base) podem representar uma parcela razoável do tempo total de serviço. Além disso, as cargas de trabalho dos veículos também podem ser diferentes.

Em ambas as aplicações descritas acima (Takeda *et al.*, 2000; Mendonça & Morabito, 2000) foi utilizado o modelo hipercubo com servidores não homogêneos. Devido a um número de servidores não elevado nessas duas aplicações foi utilizado, em ambos os casos, o método de eliminação de Gauss para resolver o sistema de equações lineares.

O presente artigo está organizado da seguinte maneira. Na Seção 2 é apresentado um exemplo hipotético para ilustrar como as soluções do modelo hipercubo com servidores homogêneos e não homogêneos podem variar. Resultados computacionais de diversas medidas de

desempenho para este exemplo são analisados e comparados. Dado que a solução do modelo com servidores homogêneos não resulta numa boa aproximação para o modelo com servidores não homogêneos, e considerando as dificuldades computacionais para resolver o segundo caso com um número de servidores relativamente grande, na Seção 3 são estudados métodos mais adequados para resolver este caso. Aspectos de convergência e detalhes das necessidades computacionais dos métodos são discutidos. Na Seção 4 são mostrados resultados computacionais obtidos ao aplicar um dos métodos da Seção 3 para resolver um conjunto de problemas teste. Finalmente, na Seção 5 apresenta-se as conclusões do artigo.

2. Um Exemplo Ilustrativo

A importância de se considerar explicitamente a não homogeneidade dos servidores no contexto de uma aplicação do modelo hipercubo é demonstrada através de um exemplo ilustrativo. Consideremos um sistema de emergência operando em uma região representada por três átomos. Os átomos estão conectados por vias de mão única, conforme mostrado na Figura 1 abaixo.

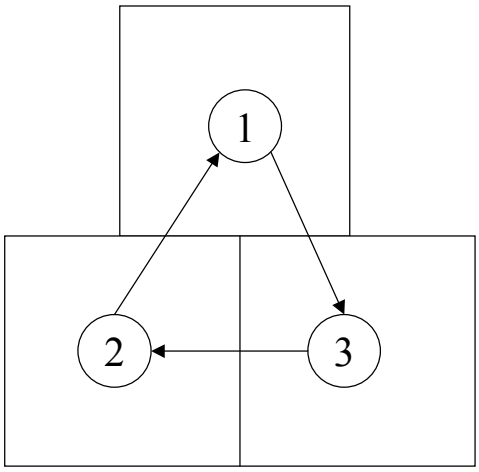

Figura 1 – Mapa da região de três átomos

Parte-se da hipótese que os centróides dos átomos estão localizados nos vértices de um triângulo eqüilátero com lados de comprimento unitário; as distâncias inter-átomos são mostradas na Tabela 1.

Tabela 1 – Distâncias inter-átomos

Do Átomo	**Ao Átomo**		
	1	**2**	**3**
1	0	2	1
2	1	0	2
3	2	1	0

Supõe-se, por outro lado, a localização de uma unidade de serviço em cada átomo, e que é utilizada uma política de despachos de preferência fixa, com as preferências alocadas com base na distância mínima de viagem. A matriz de despachos é mostrada na Tabela 2.

Tabela 2 – Matriz de Despachos

Átomo	**Preferências**		
	1ª.	**2ª.**	**3ª.**
1	1	2	3
2	2	3	1
3	3	1	2

É interessante notar que embora a matriz de distâncias não seja simétrica, a matriz de despachos é bem estruturada, com cada servidor aparecendo apenas uma vez como primeira, segunda ou terceira preferência de despacho para cada átomo. Esta estrutura foi na realidade incorporada ao modelo, para que os efeitos da não homogeneidade dos servidores pudessem ser mais facilmente observados.

Sejam 000, 001, 010, ..., 111 os $2^3 = 8$ possíveis estados do sistema, onde os 0's e 1's indicam se os servidores estão livres ou ocupados. Por exemplo, o estado 110 representa o estado em que o servidor 1 está livre e os servidores 2 e 3 estão ocupados (note que 110 descreve o estado dos servidores da direita para a esquerda). Para um sistema não saturado, com capacidade de fila infinita, as probabilidades de estado do modelo hipercubo são calculadas a partir de 7 equações de balanço de fluxo, construídas a partir dos 8 possíveis estados descritos acima, e de uma equação que corresponde à condição de normalização das probabilidades. Estas equações são as seguintes:

$\{000\}$ $\lambda p_{000} = \mu_1 p_{001} + \mu_2 p_{010} + \mu_3 p_{100}$

$\{001\}$ $(\lambda+\mu_1) p_{001} = \lambda_1 p_{000} + \mu_2 p_{011} + \mu_3 p_{101}$

$\{010\}$ $(\lambda+\mu_2) p_{010} = \lambda_2 p_{000} + \mu_1 p_{011} + \mu_3 p_{110}$

$\{100\}$ $(\lambda+\mu_3) p_{100} = \lambda_3 p_{000} + \mu_1 p_{101} + \mu_2 p_{110}$

$\{011\}$ $(\lambda+\mu_2+\mu_1) p_{011} = (\lambda_1+\lambda_2) p_{001} + \lambda_1 p_{010} + \mu_3 p_{111}$

$\{101\}$ $(\lambda+\mu_3+\mu_1) p_{101} = \lambda_3 p_{001} + (\lambda_1+\lambda_3)\, p_{100} + \mu_2 p_{111}$

$\{110\}$ $(\lambda+\mu_2+\mu_3) p_{110} = (\lambda_3+\lambda_2) p_{010} + \lambda_2 p_{100} + \mu_1 p_{111}$

$\{norm\}$ $p_{000}+p_{001}+p_{010}+\ldots+p_{111}/(1-\rho) = 1,$

onde:

λ_i é a taxa de chegada de chamadas no átomo *i*;

μ_j é a taxa de serviço do servidor *j*;

$\lambda = \lambda_1+\lambda_2+\lambda_3$ é a taxa de chegada total; e

ρ é a taxa de ocupação do sistema.

2.1 A não homogeneidade dos servidores

É necessário estabelecer, independentemente do modelo hipercubo, a maneira pela qual a não homogeneidade dos servidores será medida. Uma vez que a não homogeneidade está relacionada à dispersão da distribuição das taxas de serviço, o desvio padrão da distribuição seria a escolha mais óbvia. No caso de três servidores, no entanto, medidas de dispersão associadas com os valores extremos da distribuição, tal como por exemplo o intervalo ($\mu_{max} - \mu_{min}$), ou a "relação máximo/mínimo" (μ_{max}/μ_{min}), são tão boas quanto o desvio padrão desde que a distribuição seja considerada simétrica. Nossa escolha recaiu sobre a "relação máximo/mínimo" porque a mesma tem um apelo intuitivo e pode ser facilmente acoplada ao modelo computacional como um dado de entrada.

Na prática pode ser imaginado que, dos três servidores, um é rápido (digamos, o servidor 1), outro é lento (servidor 3) e o outro é padrão (servidor 2), e que a não homogeneidade dos servidores é medida através da relação entre as taxas de serviço dos servidores rápido e lento, digamos através da "relação rápido/lento" (RRL). Convém notar que nos casos reais a taxa de serviço μ_j pode não depender apenas das características do servidor j, mas também da configuração do sistema, como por exemplo da localização dos servidores e da política de despacho (que alteram a maneira pela qual a carga de trabalho é compartilhada entre os servidores e os tempos médios de viagem dos servidores). Nestes casos admitimos que as taxas de serviço dos servidores já foram calibradas, conforme discussão em Chiyoshi *et al.* (2000).

Objetivando utilizar nossa medida de não homogeneidade como um dado de entrada, arbitramos o tempo médio de serviço igual à unidade de tempo, de modo que $(\mu_1+\mu_2+\mu_3)/3 = 1$. Dado que consideramos a distribuição das taxas de serviço simétrica, a taxa de serviço do servidor padrão é considerada igual à unidade ($\mu_2 = 1$), de modo que, para uma dada RRL, temos $\mu_1 = 2(1+1/RRL)^{-1}$ e $\mu_3 = 2(1+RRL)^{-1}$.

2.2 Cálculo de medidas de desempenho para servidores não homogêneos

Os efeitos da não homogeneidade dos servidores são analisadas em termos de quatro medidas específicas relacionadas aos mesmos:

a) *taxa de ocupação*, que é o fator de utilização do servidor;

b) *freqüência de despacho*, que reflete a fração de chamadas atendidas alocadas ao servidor;

c) *distância média viajada por despacho*, que é a distância média viajada pelo servidor para atender uma chamada; e

d) *distância média viajada por chamada*, que é a distância média viajada pelo servidor para cada chamada atendida *pelo sistema* (note que esta medida é o produto das medidas (b) e (c)). Ou seja, a distância média (c) é simplesmente a distância média (d) normalizada pela freqüência (b).

Em geral as medidas de desempenho do modelo hipercubo podem ser calculadas, para servidores não homogêneos, através de fórmulas bem conhecidas para o caso de servidores homogêneos. As exceções referem-se a medidas associadas a chamadas em fila: freqüência de despacho e distância média viajada no sistema. Para calcular estas medidas para o caso não homogêneo é necessário calcular a probabilidade que o servidor *j* seja o primeiro servidor a ficar desocupado, dado que todos os servidores estão ocupados. Para tempos de serviço

exponencialmente distribuídos esta probabilidade, devido a uma das propriedades básicas desta distribuição, é proporcional à taxa de serviço do servidor *j*, sendo dada por $\mu_j / \sum_j \mu_j$.

Deve também ser enfatizado que, para o caso não homogêneo, a taxa de ocupação do sistema não é simplesmente a média aritmética das taxas de ocupação de cada um dos servidores. Nesse caso a taxa de ocupação do servidor *j* é o tempo de ocupação do servidor por unidade de tempo, dada por $\mu_j \rho_j$ (chamadas atendidas por unidade de tempo). A taxa de ocupação do sistema é obtida relacionando o somatório desta medida para todos os servidores à capacidade total do sistema: $\rho = \sum_j \mu_j \rho_j / \sum_j \mu_j$.

2.3 Resultados obtidos

Adicionalmente à não homogeneidade dos servidores, as medidas de desempenho para este caso dependem da taxa de ocupação do sistema e do padrão da demanda. Para uma taxa de ocupação de ρ = 0,5 e uma demanda uniformemente distribuída entre os átomos (i.e., $\lambda_1 = \lambda_2 = \lambda_3$), as medidas de desempenho variam suavemente com a não homogeneidade dos servidores. A Figura 2 mostra o comportamento das taxas de ocupação dos servidores, expressas como desvio percentual das taxas calculadas para o caso homogêneo; essas taxas são mostradas em função da não homogeneidade dos servidores, dada por RRL. Para efeitos de análise das outras medidas de desempenho descritas na seção anterior, três graus de não homogeneidade dos servidores foram testados, representados por valores de RRL iguais respectivamente a 1,2; 1,5 e 2,0 (correspondendo ao servidor rápido ser respectivamente 20%, 50% e 100% mais rápido que o servidor lento). Os resultados são mostrados na Tabela 3.

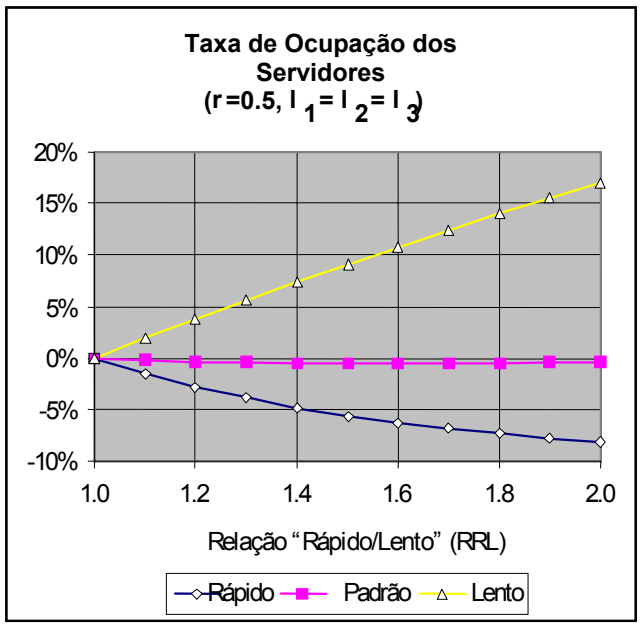

Figura 2 – Variação da Taxa de Ocupação dos Servidores

Tabela 3 – Efeitos da Não Homogeneidade dos Servidores nas Medidas de Desempenho para $\rho = 0,5$ e $\lambda_1 = \lambda_2 = \lambda_3$

Medida de Desempenho	Servidor	Relação Rápido/Lento (RRL)		
		1,2	**1,5**	**2,0**
Taxa de Ocupação	*Rápido*	-2,9%	-5,6%	-8,2%
	Padrão	-0,3%	-0,5%	-0,4%
	Lento	3,8%	9,1%	16,9%
Freqüência de Despacho	*Rápido*	3,8%	8,4%	14,3%
	Padrão	-0,3%	-0,5%	-0,4%
	Lento	-3,5%	-7,9%	-13,9%
Distância Média Viajada por Despacho	*Rápido*	4,5%	10,3%	18,2%
	Padrão	-0,5%	-0,5%	0,4%
	Lento	-3,4%	-7,3%	-11,6%
Distância Média Viajada por Chamada	*Rápido*	8,4%	19,6%	35,1%
	Padrão	-0,8%	-1,0%	0,0%
	Lento	-6,8%	-14,6%	-23,8%

O padrão que emerge da Tabela 3 é que os valores das medidas de desempenho associadas ao servidor mais rápido são ou uma função decrescente (taxa de ocupação) ou uma função crescente (outras medidas de desempenho) da não homogeneidade dos servidores. Para o servidor lento, os efeitos da não homogeneidade dos servidores atuam em direção oposta ao caso do servidor rápido. Para o servidor padrão as medidas de desempenho são praticamente invariantes em relação à não homogeneidade dos servidores.

As taxas de ocupação e freqüências de despacho caem nos intervalos –3,5% a +3,8%, –7,9% a +9,1% e –13,9% a +16,9% para valores de RRL iguais a 1,2, 1,5 e 2,0, respectivamente. Os resultados sugerem que, quando o servidor rápido é 20% mais rápido que o servidor lento, os resultados proporcionados pelo modelo para servidores homogêneos são aproximações aceitáveis para as medidas de desempenho do caso não homogêneo. Por outro lado, se o servidor rápido é 100% mais rápido que o lento, os resultados do caso homogêneo não podem ser usados para o caso não homogêneo.

As medidas de desempenho relacionadas a distâncias viajadas devem ser analisadas com mais cuidado. De certo modo todas as medidas de desempenho dependem da topologia da rede, dado que as preferências de despacho são funções das distâncias inter-átomos. É importante ressaltar, no entanto, que enquanto esta dependência das distâncias é medida em uma escala ordinal (na qual se considera apenas a ordem dos valores) para as taxas de ocupação e freqüências de despacho, as medidas de desempenho relacionadas a distâncias são dependentes das distâncias em uma escala proporcional (na qual se considera o valor real das distâncias), o que resulta em uma dependência mais forte destas últimas em relação à topologia da rede.

Com uma RRL de valor 1,2, a distância média viajada por chamada é 8,4% maior, para o servidor rápido, que a distância correspondente para o caso homogêneo; já a distância média viajada por chamada pelo servidor lento é 6,8% menor que a medida correspondente para o caso homogêneo. Assim, dependendo da topologia da rede, mesmo quando o servidor rápido é apenas 20% mais rápido que o lento, o modelo para servidores homogêneos pode não produzir previsões confiáveis para o tempo real de viagem por chamada para o caso não homogêneo. Se o servidor rápido é 50% mais rápido que o lento, os desvios acima passam a

ser de +19,6% e –14,6%, respectivamente; neste caso o uso dos resultados do modelo homogêneo para o caso não homogêneo não seria apropriado.

Conforme já observado, a distância média viajada por chamada é obtida multiplicando a freqüência de despachos por chamada pela distância média viajada por despacho. A contribuição desses dois fatores é igualmente importante no cálculo dos desvios associados à distância média viajada por chamada.

A análise acima refere-se aos efeitos da não homogeneidade dos servidores nas medidas de desempenho para uma taxa de ocupação do sistema específica ($\rho = 0,5$) e para uma demanda uniformemente distribuída entre os átomos ($\lambda_1 = \lambda_2 = \lambda_3$). Parece interessante analisar como as medidas de desempenho são influenciadas pela taxa de ocupação do sistema e pela distribuição da demanda entre os átomos. Na Tabela 4 as medidas de desempenho relativas ao *servidor rápido* para $\rho = 0,5$ são comparadas com outros dois níveis da taxa de ocupação do sistema ($\rho = 0,3$ e 0,9), mantendo-se uma distribuição uniforme da demanda entre os átomos.

Tabela 4 – Efeitos da Não Homogeneidade dos Servidores no *Servidor Rápido*, para Diferentes Taxas de Ocupação do Sistema, com $\lambda_1 = \lambda_2 = \lambda_3$

Medida de Desempenho	**Taxa de Ocupação do Sistema**	**Relação Rápido/Lento (RRL)**		
		1,2	**1,5**	**2,0**
Taxa de Ocupação	*0,3*	-4,6%	-9,0%	-13,1%
	0,5	-2,9%	-5,6%	-8,2%
	0,9	-0,5%	-0,9%	-1,3%
Freqüência de Despacho	*0,3*	3,5%	7,8%	13,4%
	0,5	3,8%	8,4%	14,3%
	0,9	1,2%	2,5%	4,2%
Distância Média Viajada por Despacho	*0,3*	5,3%	12,5%	22,8%
	0,5	4,5%	10,3%	18,2%
	0,9	4,3%	9,7%	16,5%
Distância Média Viajada por Chamada	*0,3*	9,0%	21,3%	39,2%
	0,5	8,4%	19,6%	35,1%
	0,9	5,5%	12,5%	21,4%

Como seria de se esperar, como com maiores taxas de ocupação do sistema há maior compartilhamento da carga de trabalho entre os servidores, os efeitos da não homogeneidade dos servidores nas medidas de desempenho tende a ser menor para valores maiores de ρ. No entanto, mesmo para $\rho = 0,9$, as distâncias médias viajadas por chamada para o servidor rápido são, respectivamente, 5,5%, 12,5% e 21,4% superiores aos valores correspondentes para o caso homogêneo, para valores de RRL iguais respectivamente a 1,2, 1,5 e 2,0. Aparentemente o compartilhamento da carga de trabalho entre os servidores não é suficiente para tornar confiável o uso do modelo para servidores homogêneos para prever o valor desta medida de desempenho para o caso não homogêneo.

Finalmente, com a finalidade de se observar o efeito da variação da demanda entre os átomos nas medidas de desempenho, são mostrados alguns resultados na Tabela 5. Com a finalidade de introduzir o desbalanceamento da demanda de forma controlada para uma dada taxa de ocupação do sistema, definimos λ_1/λ_3 como "relação de demanda alta/baixa" (ABR) e

arbitramos $\lambda_2=\rho$. λ_1 e λ_3 podem ser então expressos em termos de ABR e ρ da seguinte forma: $\lambda_1 = 2\rho(1+1/ABR)^{-1}$ e $\lambda_3 = 2\rho(1+ABR)^{-1}$.

Na Tabela 5 a taxa de ocupação do sistema foi fixada em $\rho = 0,5$ e as medidas de desempenho para o servidor rápido foram calculadas para três níveis de desbalanceamento da demanda, correspondendo a valores de ABR iguais respectivamente a 1,0, 1,5 e 2,0. Aparentemente o desbalanceamento da demanda não afeta de forma significativa os valores das medidas de desempenho para servidores não homogêneos, quando comparadas com os valores respectivos para o caso homogêneo.

Tabela 5 – Efeitos da Não Homogeneidade dos Servidores no *Servidor Rápido*, para Diferentes Valores de ABR (Desbalanceamento da Demanda), com $\rho = 0,5$

Medida de Desempenho	**Desbalanceamento da Demanda**	**Relação Rápido/Lento (RRL)**		
		1,2	**1,5**	**2,0**
Taxa de Ocupação	*1,0*	-2,9%	-5,6%	-8,2%
	1,5	-3,0%	-6,1%	-8,9%
	2,0	-3,1%	-6,3%	-9,3%
Freqüência de Despacho	*1,0*	3,8%	8,4%	14,3%
	1,5	3,7%	8,1%	13,6%
	2,0	3,6%	7,9%	13,2%
Distância Média Viajada por Despacho	*1,0*	4,5%	10,3%	18,2%
	1,5	5,1%	11,8%	21,1%
	2,0	5,5%	12,7%	22,8%
Distância Média Viajada por Chamada	*1,0*	8,4%	19,6%	35,1%
	1,5	9,0%	20,9%	37,6%
	2,0	9,3%	21,6%	39,0%

3. Métodos de Solução para o Modelo Hipercubo

Os resultados relativos ao exemplo ilustrativo indicam que, dependendo do grau de não homogeneidade das taxas de atendimento dos servidores, a solução com base na hipótese da homogeneidade das taxas de serviço pode não produzir boas aproximações para as medidas de desempenho do caso não homogêneo.

O método exato de solução, isto é, o método com base no sistema de equações lineares relativas ao equilíbrio dos fluxos em torno dos estados do sistema, foi tratado por Larson (1974) e pode ser aplicado tanto ao caso de taxas de atendimento homogêneas como não homogêneas. Tal sistema de equações pode ser resolvido, por exemplo, pelo método de eliminação de Gauss; no entanto, como o sistema tem dimensão 2^N (onde N é o número de servidores), dificuldades computacionais rapidamente aparecem a medida que N cresce. Na proposta apresentada por Larson, o sistema de equações foi resolvido pelo método iterativo de Gauss-Siedel, com formulação que pressupõe homogeneidade das taxas de serviço dos servidores, da forma descrita a seguir.

Supondo que os estados do sistema sejam descritos apenas pelo número agregado de usuários presentes no sistema, a probabilidade de transição do estado S_1 (i.e., apenas um usuário no sistema) para o estado S_0 (sistema vazio) em um intervalo (pequeno) de tempo h é dada por:

$$\text{Prob}\{S_1 \rightarrow S_0 \text{ em h}\} = \mu_1 p_{001} h + \mu_2 p_{010} h + \mu_3 p_{100} h.$$

Se os servidores forem homogêneos, isto é, $\mu_1 = \mu_2 = \mu_3 = k$, temos

$$k\text{Prob}(S_1)h = kp_{001}h+kp_{010}h+kp_{100}h, \text{ ou } \text{Prob}(S_1) = p_{001}+p_{010}+p_{100}.$$

A probabilidade Prob$\{S_1\}$ pode ser obtida através do modelo M/M/N. Equações similares podem ser construídas para os demais estados agregados. Estas equações, chamadas de equações de hiperplanos, são usadas por Larson em adição às equações de equilíbrio para aumentar a velocidade de convergência do método de Gauss-Siedel.

O segundo método de solução do modelo hipercubo é o método aproximado, também devido a Larson (1975). Neste método o sistema de 2^N equações lineares é substituído por um sistema de N equações não lineares. Na construção do método aproximado a hipótese da homogeneidade dos servidores também é crítica; como se verá nos próximos parágrafos, a hipótese da homogeneidade dos servidores é indispensável à derivação deste método e portanto o método aproximado não se aplica ao caso de servidores com taxas de atendimento não homogêneas.

No método aproximado, as equações em ρ_j são construídas com base nas taxas totais de despacho dos servidores. Usando o exemplo ilustrativo da Seção 2, quando se considera a taxa total de despacho do servidor 1, um componente a ser considerado é a taxa de despacho, sem atraso, da unidade 1 para o átomo 2. Sendo [2 3 1] o vetor de despachos associado ao átomo 2 (ver Tabela 2), esta taxa é dada pela demanda do átomo 2 multiplicada pela probabilidade de que os servidores 2 e 3 estejam ocupados e o servidor 1 esteja livre, isto é, $\lambda_2\text{Prob}\{B_2B_3F_1\}$, onde B_j refere-se ao servidor *j* ocupado e F_k ao servidor *k* livre. Se os servidores fossem independentes, a probabilidade do evento $\{B_2B_3F_1\}$ seria $\rho_2\rho_3(1-\rho_1)$. Para levar em consideração a não independência dos servidores, Larson propõe o uso de um fator de correção C tal que $\text{Prob}\{B_2B_3F_1\} \cong C.\rho_2.\rho_3.(1-\rho_1)$.

Para derivar uma expressão para o fator de correção, Larson recorre ao modelo M/M/N de filas, com servidores homogêneos por definição, definindo o evento $\{B_1B_2F_3\}_{(M/M/N)}$ em que, na amostragem aleatória sem reposição de servidores, os dois primeiros servidores estão ocupados e o terceiro está livre, obtendo $\text{Prob}\{B_1B_2F_3\}_{(M/M/N)} = Q.\rho^2.(1-\rho)$. É através da comparação destas probabilidades que Larson deriva o fator de correção Q do seu método aproximado (para maiores detalhes ver Chiyoshi *et al.*, 2000).

Assim, dado que os métodos de solução do modelo hipercubo descritos na literatura tomam por base a hipótese da homogeneidade dos servidores, parece oportuno dedicar alguma atenção ao caso dos servidores não homogêneos, especialmente tendo em vista os resultados apresentados no exemplo ilustrativo. Como o método aproximado de Larson não se aplica ao caso de servidores não homogêneos, é fundamental examinar a eficiência computacional dos métodos disponíveis para resolver o sistema de equações lineares do modelo hipercubo. É disto que nos ocupamos a seguir.

3.1 Métodos de solução de sistemas de equações lineares

São apresentados a seguir, com razoável nível de detalhamento, os métodos da eliminação de Gauss e o método Gauss-Siedel, aplicados à solução do sistema de equações lineares do modelo hipercubo.

O método da eliminação de Gauss

Grassman *et al.* (1985) propõem um algoritmo para a determinação da distribuição estacionária de uma cadeia de Markov com base no método da eliminação de Gauss, em que exploram a particularidade de que a matriz de coeficientes A do sistema de equações deriva de uma matriz de probabilidades de transição. O assim chamado algoritmo GHT, considerado econômico por Seneta (1998), requer $3.2^{3N}/4$ operações para resolver o sistema linear associado a um modelo hipercubo de N servidores, trabalhando com a matriz A cheia.

Alguma redução no número de operações pode ser obtida quando os elementos não nulos da matriz de coeficientes localizam-se dentro de uma banda em torno da diagonal principal, de modo que $a_{ij} = 0$ se $j > i + g$ ou $j < i – g$ para algum inteiro positivo g. Nesse caso o número de operações do método GTH é $2.2^N g^2$. Quando os estados do hipercubo são ordenados em ordem crescente dos valores decimais correspondentes aos vetores binários que representam os estados, tem-se uma matriz de banda com $g = 2^{N–1}$. Neste caso o número de operações do algoritmo para resolver um sistema de N servidores é $2^{3N}/2$.

O número de operações acima representa um custo de processamento explosivo em função do número de servidores, tornando-se rapidamente inviável. Para se ter uma idéia desse comportamento, em um equipamento capaz de realizar 20.2^{20} operações de ponto flutuante por segundo (20 megaflops), seriam necessários 205 segundos para resolver um modelo com 11 servidores e 233 horas para resolver um com 15 servidores.

Além do tempo de processamento, há que se considerar a memória necessária para armazenar a matriz de coeficientes. Se forem utilizados 8 bytes para armazenar um coeficiente, a necessidade de memória em Mbytes é dada por $8.2^{2N–20}$ para a matriz associada a um problema de N servidores. A necessidade de memória é também de comportamento explosivo: as necessidades de memória para armazenar a matriz de coeficientes são de 32 e 128 Mbytes, respectivamente, para problema de 11 e 12 servidores, de modo que, por exemplo, com disponibilidade de 64 Mbytes de memória, a solução de um problema de 12 servidores estaria fora de cogitação.

Vê-se assim que o método da eliminação mostra-se inviável, em termos de tempo de processamento e necessidade de memória, para resolver o sistema linear associado ao modelo hipercubo, mesmo para problemas de porte modesto. Deve-se portanto recorrer a métodos iterativos de solução de sistemas lineares.

O método de Gauss-Siedel

Os métodos iterativos apresentam a vantagem de não alterar a matriz de coeficientes ao longo do processo e utilizar apenas seus elementos não nulos. Isto tem implicação importante na necessidade de memória para sua implementação.

Parte-se da hipótese que os elementos da matriz são armazenados em forma de tríade (linha, coluna, valor), com 2 inteiros de 4 bytes e um real de 8 bytes, totalizando 16 bytes por cada elemento da matriz. Como cada linha da matriz tem apenas N+1 elementos não nulos, a necessidade de memória seria de $16(N+1)2^{N–20}$ Mbytes. Um problema com 17 servidores, por exemplo, que requer 36 Mbytes de memória para armazenar seus coeficientes, poderia ser trabalhado em 64 Mbytes de memória.

A dificuldade da previsão do tempo de processamento dos métodos iterativos deriva do fato de que não se conhece o número de iterações necessárias para a convergência do método,

para uma dada tolerância ou critério de convergência. De qualquer forma, mesmo que fossem necessárias 200 iterações, o número de operações seria de $200(N+1)2^N$, número que pode ser comparado favoravelmente com as $2^{3N}/2$ operações do método da eliminação de Gauss: seriam necessários apenas 5 segundos de um processador de 20 Megaflops para resolver um sistema de 15 servidores, em comparação com as 233 horas do método da eliminação.

Existem dois métodos iterativos simples para resolver sistemas de equações lineares: Gauss-Jacobi e Gauss-Siedel. Na análise desses métodos, Rice (1983) usa a decomposição da matriz A em sub-matrizes D (diagonal), U (parte triangular superior) e L (parte triangular inferior), concluindo que o método de Gauss-Jacobi supõe que D é uma boa aproximação de A, enquanto que o de Gauss-Siedel supõe que D+L é uma boa aproximação de A.

A Tabela 6 mostra a matriz de coeficientes do sistema linear associado a um modelo hipercubo 3x3 sem fila de espera. Constata-se que os elementos abaixo da diagonal principal são taxas de transição de um estado para outro com maior número de servidores ocupados. Como qualquer chegada provoca este tipo de transição, desde que o sistema não esteja saturado, vê-se que com exceção da última coluna (relativa ao estado saturado {111}), a soma das colunas de L totaliza $\lambda = \lambda_1+\lambda_2+\lambda_3$. Se adotarmos a soma dos valores absolutos de seus elementos para medir o "peso" de uma matriz, vemos que o peso de L é dado por $w(L) = (2^N-1)\lambda$.

Tabela 6 – Matriz de Coeficientes do Sistema Linear Associado a um Modelo Hipercubo 3x3 sem Fila de Espera

	000	**001**	**010**	**011**	**100**	**101**	**110**	**111**
000	$-\lambda$	μ_1	μ_2		μ_3			
001	λ_1	$-(\lambda+\mu_1)$		μ_2		μ_3		
010	λ_2		$-(\lambda+\mu_2)$	μ_1			μ_3	
011		$(\lambda_1+\lambda_2)$	λ_1	$-(\lambda+\mu_1+\mu_2)$				μ_3
100	λ_3				$-(\lambda+\mu_3)$	μ_1	μ_2	
101		λ_3			$(\lambda_1+\lambda_3)$	$-(\lambda+\mu_1+\mu_3)$		μ_2
110			$(\lambda_3+\lambda_2)$		λ_2		$-(\lambda+\mu_2+\mu_3)$	μ_1
111				λ		λ	λ	$-\mu$

A parte triangular superior de A é formada por taxas de transição para estados do sistema com menor número de servidores ocupados, de modo que seus elementos não nulos são formados por alguma combinação dos μ_j's. Na verdade, considerando um μ_j particular, vê-se que ele ocorre em todas as colunas associadas a estados com o servidor *j* ocupado, isto é, em 2^{N-1} colunas. Como isto ocorre para cada servidor, o peso da parte triangular superior de A é dado por $w(U) = 2^{N-1}\mu$, onde $\mu = \mu_1 + \mu_2 + \mu_3$.

Quanto à diagonal D, como a matriz A tem estrutura especial em que a soma dos elementos de cada coluna é zero, segue que $w(D) = w(U) + w(L)$ e $w(D+L) = w(U) + 2w(L)$. Desta forma, utilizando-se a soma dos valores absolutos dos elementos da matriz como critério de comparação, a matriz (D+L) seria melhor "aproximação" de A do que a matriz diagonal D, sendo a "aproximação" tanto melhor quanto maior w(L) em relação a w(U). Esta constatação é indicativa de que o *método de Gauss-Siedel* é em princípio mais apropriado do que o *método de Gauss-Jacobi* para resolver o sistema de equações associado ao modelo hipercubo.

É importante notar que, para uma dada capacidade de atendimento do sistema μ, quanto menor a demanda λ, menor será w(L); conseqüentemente (D+L) aproxima-se menos de A. A matriz A, no entanto, conforme mostrado na Tabela 6, pressupõe que os estados do hipercubo são ordenados em ordem crescente dos valores decimais correspondentes aos vetores binários que representam os estados. Se a ordenação fosse na ordem inversa, a disposição dos elementos de A seria alterada, com as taxas de chegada compondo os elementos da matriz U e as taxas de serviço os elementos de L, de modo que teríamos nesse caso $w(L) = 2^{N-1}\mu$ e $w(U) = (2^N-1)\lambda$, D mantendo-se inalterada. Assim, dependendo da relação demanda/capacidade do problema em análise, sempre é possível escolher a ordenação mais adequada dos estados do modelo.

Na prática os efeitos da inversão da ordenação dos estados pode ser obtida utilizando-se a mesma matriz A, com uma simples inversão da ordem em que as variáveis são atualizadas: em vez de processar-se na seqüência X_1->X_n, processa-se na seqüência X_n->X_1, com $n = 2^N$.

4. Resultados Computacionais

Os resultados computacionais apresentados a seguir referem-se ao método Gauss-Siedel. Após o detalhamento do método explica-se a geração dos problemas-teste, o que é seguido pelos resultados obtidos, apresentados em forma de gráficos. É mostrado finalmente um modelo de regressão que estima o número de iterações do método e o tempo computacional respectivo.

4.1 Detalhes do método

Na implementação do método de Gauss-Siedel, optou-se por incluir todas as 2^N equações de equilíbrio, mesmo que uma seja redundante, e realizar a normalização das probabilidades a cada iteração. Esta opção teve por base duas considerações. Em primeiro lugar, qualquer equação de equilíbrio estabelece uma relação precisa que deve existir entre as probabilidades associadas a N+1 estados específicos do hipercubo; por conseguinte, todas são importantes. Em segundo lugar, a normalização é um processo que deve afetar o conjunto das probabilidades e não apenas uma, o que ocorreria se uma equação de equilíbrio fosse descartada e utilizássemos a condição de normalização em seu lugar. As equações dos hiperplanos não foram incorporadas ao sistema de equações para evitar que o mesmo fique particularizado para o caso de servidores homogêneos.

Quando os servidores são homogêneos, os estados agregados do modelo hipercubo têm a distribuição dada por um modelo M/M/N; assim, para o estado de um servidor ocupado em um modelo hipercubo de três servidores tem-se $\text{Prob}(S_1)_{M/M/3} = p_{001}+p_{010}+p_{100}$.

No estabelecimento dos valores iniciais das probabilidades para o método de Gauss-Siedel, considerou-se que, embora esta relação não seja válida para sistemas com servidores não homogêneos, uma boa forma de se determinar os valores iniciais das probabilidades seria fazer-se $p_{001} = p_{010} = p_{100} = \text{Prob}\{S_1\}_{M/M/3}/3$. Em geral, para um sistema com N servidores, o valor inicial da probabilidade de um estado pertencente ao hiperplano com k servidores ocupados pode ser dado por $\text{Prob}\{S_k\}_{M/M/N}/C(N,k)$, onde C(N,k) é a combinação de N elementos k a k.

Para definir um critério de convergência é necessário considerar que o problema em análise consiste na determinação de 2^N probabilidades, cuja ordem de grandeza é 2^{-N}. Para utilização de forma uniforme em problemas com diferentes valores de N, o erro relativo é um critério

mais apropriado do que o erro absoluto para definir a convergência do método. Considerou-se 10^{-6} um valor suficientemente rigoroso para testar o método, de modo que definiu-se como critério de convergência a condição de que os erros relativos de todas as variáveis, em módulo, sejam menores que 10^{-6}.

O método foi codificado para modelos sem fila de espera. Este fato não é restritivo porque os resultados relativos a um sistema sem fila podem ser ajustados para sistemas com previsão de fila de espera, conforme mostrado abaixo.

Na construção de seu algoritmo, Grassman *et al.* (1985) mostram que existe uma relação de recursão $p_j = f(p_{j-1})$, j=1...n, entre as probabilidades de equilíbrio de uma cadeia de Markov. Essa recursão não pode ser usada diretamente porque p_0 é desconhecido no início do processo. O algoritmo de Grassman determina inicialmente uma seqüência auxiliar r_j usando a recursão f a partir de $r_0 = 1$, utilizando então a normalização para converter os termos da seqüência r_j nas probabilidades desejadas.

Se o algoritmo for aplicado a um modelo hipercubo sem fila de espera e desejar-se ajustá-lo para um modelo com espaço ilimitado para fila, basta incorporar o termo adicional $[\rho r_n/(1-\rho)]$ à seqüência r_j para representar a fila. Na verdade, este ajuste pode ser feito também diretamente sobre as probabilidades p_j's através de:

$$p_j^{(\infty)} = \frac{p_j^{(0)}}{1 + \frac{\rho\, p_n^{(0)}}{1-\rho}},$$

em que os índices superiores indicam o modelo a que se referem as probabilidades, sem fila ou com capacidade infinita. Esta relação reflete o fato de que a distribuição de probabilidades dos estados do hipercubo, na versão sem fila de espera, deve ser expandida em um termo representando a cauda do modelo com fila, e que a série resultante deve ser normalizada para somar 1, para termos uma distribuição de probabilidades.

4.2 Geração dos problemas teste

Os problemas teste para o método de Gauss-Siedel foram gerados a partir da rede de 55 nós usada por Batta *et al.* (1989), utilizando-se um gerador que tem como dados de entrada a taxa de utilização (ρ) e o número de servidores (N). Os dados de saída são o vetor de demandas (λ_i), o vetor de taxas de serviço (μ_j) e a matriz de despachos. O gerador supõe um modelo hipercubo com taxa média de serviço unitária, de modo que $\sum\mu_j = N$ e $\sum\lambda_i = \rho N$.

As demandas são geradas proporcionalmente aos pesos dos nós de modo a totalizar ρN. As taxas de serviço são geradas com distribuição uniforme entre 0,5 e 1,5 e depois ajustadas para gerar média igual à unidade. Para gerar a matriz de despachos, sorteia-se aleatoriamente os nós onde localizar os servidores e determina-se a seguir as prioridades de despacho usando o critério de menor distância.

Os testes foram feitos em modelos com N = 10 a 17 servidores, variando-se a taxa de ocupação de $\rho = 0{,}1$ a $\rho = 0{,}9$, com incrementos de 0,1 para cada N. Foram gerados 10 problemas teste para cada par (N, ρ), o que dá um total de 720 problemas. Cada problema foi resolvido para as seqüências $X_1 \to X_n$ e $X_n \to X_1$ de atualização das variáveis, totalizando 1440 aplicações do método.

4.3 Resultados obtidos

A primeira constatação de importância sobre os resultados obtidos é que houve convergência em todos os 1440 casos, com o número de iterações situando-se entre 22 e 114. Isto parece indicar que o método de Gauss-Siedel, na implementação descrita acima, permite resolver um modelo hipercubo de dimensão razoável sem recursos computacionais sofisticados.

Como era de se esperar, o número de iterações necessárias para resolver o sistema de equações cresce com o número de servidores de forma uniforme para todos os valores da taxa de ocupação do sistema, como mostra a Figura 3 (note-se que na Figura 3 cada ponto representa a média de 10 observações). Quanto ao comportamento do número de iterações em função da taxa média de ocupação do sistema, pela análise acima, com base no peso relativo da matriz D+L, seria de se esperar que tal número fosse uma função monotônica crescente de ρ. Não é o que mostram os resultados do método para valores extremos de ρ.

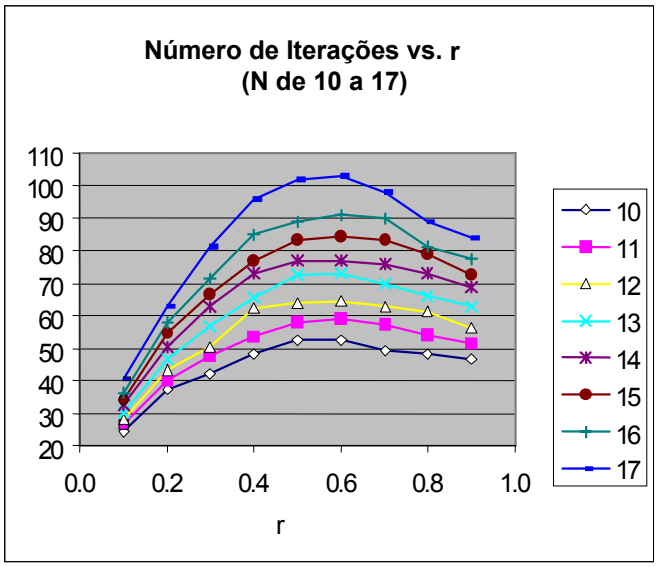

Figura 3 – Número de Iterações em Função de ρ
e do Número de Servidores ($10 \leq N \leq 17$)

O recurso de alterar a seqüência de atualização das variáveis também não atua na forma esperada. Na verdade, para valores pequenos de ρ, o método com a seqüência direta $X_1 \rightarrow X_n$, embora com menor w(D+L), requer menos iterações do que com a seqüência inversa $X_n \rightarrow X_1$ (ver Figura 4). Neste caso o fator de assimetria subjacente parece ser de identificação mais fácil. Quando a taxa de ocupação do sistema é baixa, a distribuição de probabilidades se concentra nos estados com menor número de servidores ocupados. Como o critério de convergência se baseia no erro relativo, deve-se esperar que as probabilidades destes estados sejam mais estáveis por serem maiores e convirjam mais rapidamente. Os resultados indicam que a melhor seqüência de atualizar as variáveis é aquela que prioriza as variáveis de maior valor.

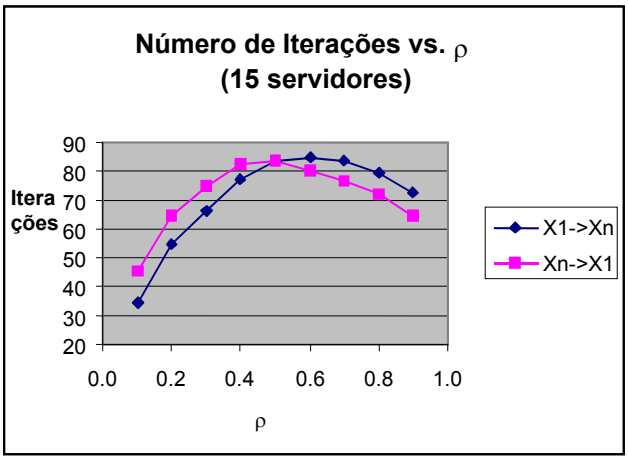

Figura 4 – Número de Iterações vs. ρ, Seqüências Direta e Inversa (N = 15)

Para valores grandes de ρ, o método com base na seqüência X_n->X_1 de atualizar as variáveis mostra convergência melhor que o método que usa a seqüência X_1->X_n. Este comportamento corrobora a conjectura de que as variáveis de maior valor devem ser priorizadas no processo de atualização das mesmas.

4.4 Modelo de regressão para estimar o número de iterações

Para obter uma equação que permita estimar o número de iterações necessárias para resolver um dado problema, realizou-se uma análise de regressão partindo de três variáveis explicativas óbvias:

- ρ – taxa de ocupação do sistema;
- N – número de servidores;
- Z – variável binária tal que Z=0 para a seqüência ascendente e Z=1 para a seqüência descendente de processamento das variáveis.

Os dados sugerem que a relação entre a taxa de ocupação e o número de iterações poderia ser descrita por uma função quadrática, o que requer a inclusão do termo ρ^2 no modelo. Além disso, a referida relação depende da seqüência de processamento das variáveis informada em Z. Para modelar esta dependência devem ser incluídas as interações Z.ρ e Z.ρ^2. O modelo assim definido mostrou boa aderência aos dados, com coeficiente de determinação da ordem de 93%. A função ajustada para Y, número médio de iterações necessárias para resolver um problema, é dada por:

$$Y = -59{,}2 + 200{,}2\rho - 162{,}6\rho^2 + 14{,}2Z + 5{,}5N - 35{,}3Z.\rho + 12{,}7Z.\rho^2,$$

com erro padrão da estimativa de ± 4,9 iterações.

A partir da estimativa do número de iterações é possível estimar o tempo de processamento: para um sistema de N servidores, o número de operações por iteração é $(N+1)2^N$, de modo que para realizar Y iterações um computador com capacidade de C megaflops requer aproximadamente $Y.(N+1).2^{N-20}/C$ segundos de processamento.

Quanto ao tempo total de processamento, tem-se que se considerar o tempo necessário para gerar os coeficientes das equações. Nos resultados obtidos neste artigo o tempo para gerar os coeficientes foi em média maior do que o tempo para resolver as equações, na proporção aproximada de 14/10.

5. Conclusões

No presente trabalho foi enfatizada a importância de se considerar, em algumas aplicações, servidores não homogêneos no modelo hipercubo e estudar métodos de solução adequados para este caso. Um exemplo ilustrativo demonstrou inicialmente como as soluções do modelo hipercubo podem variar quando se compara os resultados para servidores homogêneos e não homogêneos. Dependendo do grau de não homogeneidade das taxas de atendimento dos servidores, a solução com base na hipótese da homogeneidade pode produzir resultados que se afastam das características operacionais do modelo quando as taxas de atendimento são não homogêneas.

O método exato de solução do modelo hipercubo pode ser aplicado tanto ao caso de taxas de atendimento homogêneas como não homogêneas. O sistema de equações lineares envolvido pode ser resolvido, por exemplo, pelo método de eliminação de Gauss; no entanto, como este sistema tem dimensão 2^N, dificuldades computacionais rapidamente aparecem a medida que N cresce. Na proposta apresentada por Larson o sistema de equações foi resolvido pelo método iterativo de Gauss-Siedel, com formulação que pressupõe homogeneidade das taxas de serviço dos servidores. Foi demonstrado neste artigo que o mesmo procedimento, com as adaptações necessárias, é também o mais indicado para o caso de servidores não-homogêneos.

O método de Gauss-Siedel para servidores com taxas de serviço não-homogêneas foi testado utilizando-se a rede de 55 nós usada por Batta *et al.* (1989). Os testes foram feitos em modelos com 10 a 17 servidores, variando-se a taxa de ocupação de 0,1 a 0,9, com incrementos de 0,1 para cada N. Foram gerados 10 problemas teste para cada par (N, ρ) e cada problema foi resolvido para as seqüências $X_1 \rightarrow X_n$ e $X_n \rightarrow X_1$ de atualização das variáveis, totalizando 1440 aplicações do método.

Conforme esperado, o número de iterações necessárias para resolver o sistema de equações cresce com o número de servidores de forma uniforme para todos os valores da taxa de ocupação do sistema. Quanto ao comportamento do número de iterações em função da taxa média de ocupação do sistema, trata-se de uma função unimodal, com o valor máximo ocorrendo em torno de $\rho = 0,5$. Por outro lado, para valores grandes de ρ o método com base na seqüência $X_n \rightarrow X_1$ de atualizar as variáveis mostra convergência melhor que o método que usa a seqüência $X_1 \rightarrow X_n$.

O número de iterações necessárias e o tempo de processamento para resolver um dado problema podem ser estimados através de um modelo de regressão. É interessante notar que, nos resultados obtidos neste artigo, o tempo para gerar os coeficientes do modelo hipercubo foi em média maior do que o tempo para resolver as equações, na proporção aproximada de 14/10.

Referências Bibliográficas

(1) Albino, J.C.C. (1994). Quantificação e locação de unidades móveis de atendimento de emergência e interrupções em redes de distribuição de energia elétrica: aplicação do modelo hipercubo. Tese de Mestrado, Depto. de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

(2) Batta, R.; Dolan, J.M. & Krishnamurthy, N.N. (1989). The maximal expected covering location problem: Revisited. *Transportation Science*, **23**, 277-287.

(3) Brandeau, M.L. & Larson, R.C. (1986). Extending and applying the hypercube model to deploy ambulances in Boston. **In**: *Delivery of Urban Services – TIMS Studies in the Management Sciences*, *Vol. 22* [edited by A. Swersey e E. Ignall], Elsevier Science, B.V., 121-153.

(4) Burwell, T.H.; Jarvis, J.P. & McKnew, M.A. (1993). Modeling co-located servers and dispatch ties in the hypercube model. *Computers and Operations Research*, **20**, 113-119.

(5) Chiyoshi, F.Y.; Galvão, R.D. & Morabito, R. (2000). O uso do modelo hipercubo na solução de problemas de localização probabilísticos. *Gestão & Produção*, **7**, 146-174.

(6) Gonçalves, M.B.; Novaes, A.G.N. & Albino, J.C.C. (1994). Modelos para localização de serviços emergenciais em rodovias. *Anais do XXVI Simpósio Brasileiro de Pesquisa Operacional*, **1**, 591-596.

(7) Gonçalves, M.B.; Novaes, A.G.N. & Schmitz, R. (1995). Um modelo de otimização para localizar unidades de serviços emergenciais em rodovias. *Anais do IX ANPET – Congresso de Pesquisa e Ensino em Transportes*, 962-972.

(8) Grassman, W.K.; Taksar, M.L. & Heyman, D.P. (1985). Regenerative analysis and steady state distributions of Markov Chains. *Operations Research*, **33**, 1107-1116.

(9) Halpern, J. (1977). Accuracy of estimates for the performance criteria in certain emergency service queueing systems. *Transportation Science*, **11**, 223-242.

(10) Larson, R.C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, **1**, 67-95.

(11) Larson, R.C. (1975). Approximating the performance of urban emergency service systems. *Operations Research*, **23**, 845-868.

(12) Larson, R.C. & Odoni, A.R. (1981). *Urban Operations Research*, Prentice Hall, Inc., N.J.

(13) Mendonça, F.C. (1999). Aplicação do modelo hipercubo, baseado em teoria de filas, para análise de um sistema médico-emergencial em rodovia. Tese de Mestrado, Depto. de Engenharia de Produção, Universidade Federal de São Carlos.

(14) Mendonça, F.C. & Morabito, R. (2000). Aplicação do modelo hipercubo para análise de um sistema médico-emergencial em rodovia. *Gestão & Produção*, **7**, 73-91.

(15) Rice, J.R. (1983). *Numerical Methods, Software, and Analysis,* McGraw-Hill Book, New York.

(16) Sacks, S.R. & Grief, S. (1994). Orlando Police Department uses OR/MS methodology, new software to design patrol districts. *OR/MS Today*, Feb. 1994, 30-32.

(17) Seneta, E. (1998). Complementation in stochastic matrices and GTH algorithm. *Siam J. Matrix Anal. Appl.*, **19**, 556-563.

(18) Swersey, A.J. (1994). The deployment of police, fire and emergency medical units. **In:** *Handbooks in OR & MS, Vol. 6* [edited by S.M. Pollock *et al.*], Elsevier Science B.V., 151-200.

(19) Takeda, R.A. (2000). Uma contribuição para avaliar o desempenho de sistemas de transporte emergencial de saúde. Tese de Doutorado, Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, USP.

(20) Takeda, R.A.; Widmer, J.A.. & Morabito, R. (2000). Uma proposta alternativa para avaliação do desempenho de sistemas de transporte emergencial de saúde brasileiros. Aceito para publicação na revista *Transportes*.