

Um Modelo Analítico de Desempenho para Servidores Web Virtualizados com Xen

Ricardo J. Pfitscher¹, Rafael R. Obelheiro¹, Maurício A. Pillon¹

¹Departamento de Ciência da Computação (DCC)
Universidade do Estado de Santa Catarina (UDESC)

ricardo.pfitscher@gmail.com, {rro,mpillon}@joinville.udesc.br

Resumo. A virtualização é uma ferramenta chave para consolidação de servidores, pois permite que cada servidor execute em uma máquina virtual isolada das demais máquinas virtuais no mesmo hardware, proporcionando a redução de custos. Por outro lado, o desempenho dos servidores virtualizados pode não satisfazer os requisitos de usuários ou aplicações, devido ao compartilhamento de recursos entre MVs. Atualmente, o desempenho de servidores em ambientes de virtualização precisa ser avaliado experimentalmente, ou seja, é preciso implantar o servidor em uma máquina virtual para depois verificar se o seu desempenho é adequado. Visando preencher esta lacuna, este trabalho propõe um modelo analítico de desempenho para servidores web virtualizados no ambiente Xen. Os resultados obtidos na validação do modelo demonstram que ele pode ser usado com sucesso para fins de planejamento de capacidade.

Abstract. Virtualization is a key technology for server consolidation, which allows each server to run inside a virtual machine that is isolated from other virtual machines in the same hardware, thereby reducing costs. On the other hand, the performance of virtualized servers may not satisfy user or application requirements, due to resource sharing across VMs. Currently, server performance in virtualized environments has to be evaluated experimentally, i.e., first we deploy a server in a virtual machine and then determine whether its performance is adequate. Aiming to bridge this gap, we propose an analytic performance model of virtualized web servers using the Xen hypervisor. The results obtained during model validation indicate that it can be successfully used for capacity planning purposes.

1. Introdução

Uma tendência recente no gerenciamento de servidores de rede, particularmente na hospedagem de servidores web, tem sido a consolidação de múltiplos servidores em um único hardware, com o objetivo de reduzir custos operacionais e atingir maior eficiência na utilização dos recursos computacionais disponíveis [Tickoo et al. 2010]. Uma tecnologia chave para a consolidação de servidores é a virtualização, que permite que cada servidor execute em uma máquina virtual isolada das demais máquinas virtuais no mesmo hardware, facilitando assim o gerenciamento de cada servidor.

Muito embora a virtualização seja uma solução eficaz para a consolidação de vários servidores em um único hardware, em determinados cenários o desempenho obtido com os servidores virtualizados pode ficar aquém das necessidades de usuários e aplicações [Kotsovinos 2010]. Isso ocorre porque a carga de trabalho aplicada sobre esses servidores pode exceder a capacidade do hardware subjacente. Muitas vezes, essa inadequação só é percebida depois da implantação dos servidores virtualizados, o que acarreta transtornos e despesas adicionais. Isso demonstra a importância de se ter ferramentas

que permitam estimar o desempenho de servidores virtualizados, sem que seja necessário realizar uma implementação para a condução de experimentos visando a avaliação de desempenho. Ferramentas particularmente úteis nesse contexto são os modelos analíticos de desempenho [Lazowska et al. 1984, Jain 1991, Menascé e Almeida 2003].

A modelagem analítica de ambientes de virtualização é uma área pouco explorada na literatura. Em [Menascé 2005] é apresentado um modelo analítico caracterizado pela divisão proporcional dos recursos computacionais entre as MVs executantes, sendo que o *overhead* do ambiente de virtualização é considerado como um fator multiplicativo constante sobre os tempos de resposta de cada recurso (processador e disco são usados no artigo). Uma dificuldade no uso desse modelo é a necessidade de implantação do ambiente virtualizado, uma vez que os parâmetros de entrada dependem de dados que só podem ser mensurados no ambiente a ser avaliado, o que se contrapõe à ideia de estimar o desempenho sem precisar da implementação.

O foco deste trabalho consiste em propor um modelo analítico que possibilite estimar o desempenho de servidores web virtualizados, com base em parâmetros mensurados no hardware que irá hospedar esses servidores. A principal contribuição se dá na possibilidade de estimar o desempenho do ambiente virtualizado usando dados do ambiente conhecido, sem que haja a necessidade de implantar virtualização *a priori*.

O restante deste artigo está organizado da seguinte forma. A seção 2 apresenta conceitos de virtualização e da ferramenta Xen. A seção 3 descreve um modelo analítico para servidores web proposto em [Menascé e Almeida 2003], e introduz o modelo analítico para servidores web virtualizados com Xen, baseado no primeiro, que é o foco deste trabalho. Na seção 4 os resultados fornecidos por esses modelos são comparados com resultados obtidos através de experimentação, e a seção 5 conclui o artigo.

2. Virtualização

O conceito original de virtualização, conhecido atualmente como virtualização em nível de hardware, consiste na inserção de uma camada de software entre o sistema operacional (SO) e o hardware que fornece ao usuário (no caso o SO) uma abstração dos recursos físicos da máquina, chamada de máquina virtual [Rosenblum 2004]. A camada de software que cria essa abstração é chamada de ambiente de virtualização, hipervisor ou monitor de máquina virtual (MMV), e é responsável por intermediar os acessos do SO que executa em uma máquina virtual ao hardware subjacente, fornecendo isolamento e proteção entre as várias MVs – que podem, inclusive, executar SOs distintos. Um exemplo do uso de virtualização na consolidação de servidores é ilustrado na figura 1(a). Pela figura, é possível perceber que o desempenho das aplicações que executam em uma máquina virtual é influenciado pela camada adicional de software do MMV e pelo compartilhamento de recursos de hardware com as outras MVs.

O ambiente de virtualização usado neste trabalho é a ferramenta de código aberto Xen [Barham et al. 2003], uma das mais adotadas na consolidação de servidores. O Xen adota uma técnica de paravirtualização, onde o sistema operacional convidado (que executa em uma MV) deve ser modificado para gerenciar recursos de hardware e executar operações de E/S em colaboração com o ambiente de virtualização (em outras palavras, o SO convidado é ciente do hipervisor). No esquema de paravirtualização do Xen, ilustrado na figura 1(b), uma máquina virtual privilegiada (chamada de Dom0) é usada para ge-

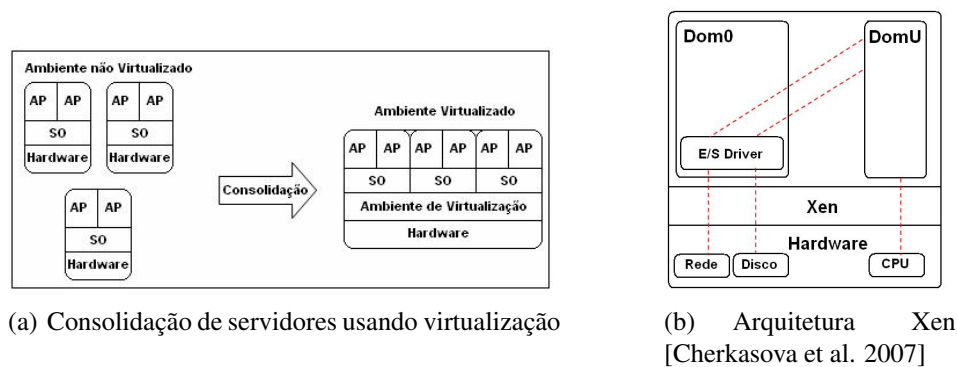


Figura 1. Consolidação de servidores e arquitetura do Xen

reenciar o hipervisor, e possui também os *drivers* de dispositivo necessários para E/S. As demais máquinas virtuais, chamadas de DomU, realizam as suas operações de E/S usando os *drivers* em Dom0.

3. Modelos Analíticos para Servidores Web

Nesta seção são descritos dois modelos analíticos para servidores web, o definido por [Menascé e Almeida 2003] e o modelo elaborado para virtualização. O objetivo é tomar por base o modelo existente, no qual os recursos utilizados pelo servidor web já estão caracterizados, para elaborar um novo modelo que contemple também a virtualização.

3.1. Modelo definido por Menascé

Em [Menascé e Almeida 2003], diferentes modelos analíticos para servidores web são apresentados, com variados graus de detalhe. Dentre estes, utilizou-se o modelo baseado em componentes para servidores web com conteúdo estático. Esse modelo, mostrado na figura 2, representa o relacionamento entre os diferentes recursos usados pelo servidor web (CPU, disco e rede) através de uma rede de filas. Uma requisição web chega ao servidor através do enlace de entrada e é encaminhada à fila da CPU, onde fica aguardando processamento. Ao ser processada, a página requisitada é identificada, e o arquivo correspondente pode ser solicitado ao subsistema de disco. Depois que o conteúdo do arquivo relativo à página é lido, uma nova iteração na CPU é necessária para que a resposta HTTP contendo a página seja montada, e em seguida retornada ao cliente através do enlace de saída. Os recursos considerados no modelo são CPU, disco e rede, e o desempenho desses recursos é descrito pelas seguintes variáveis:

- X : Taxa de processamento do servidor (req/s)
- U_i : Utilização do recurso i (%)
- D_i : Demanda de serviço sobre um determinado recurso i (s)
- R'_i : Tempo de residência da requisição no recurso i (s)
- R_0 : Tempo de resposta final (s)

Dentre essas variáveis, as taxas de processamento (X) e utilização (U_i) são tipicamente determinadas de forma experimental, e as demais são derivadas a partir de relações matemáticas entre os parâmetros experimentais [Menascé e Almeida 2003], mostradas nas equações (1) e (2). O tempo de residência na rede, definido pela equação (3), é aproximado pela razão entre o volume de dados trafegados e a largura de banda disponível. O

tempo de resposta é dado pela soma das residências em cada recurso (CPU, disco, rede), conforme a equação (4).

$$D_i = \frac{U_i}{X} \quad (i = \text{CPU, disco}) \quad (1) \quad R'_{rede} = \frac{\text{TamArquivo} + \text{Overhead}}{\text{Larg. Banda}} \quad (3)$$

$$R'_i = \frac{D_i}{1 - U_i} \quad (2) \quad R_0 = R'_{cpu} + R'_{disco} + R'_{rede} \quad (4)$$

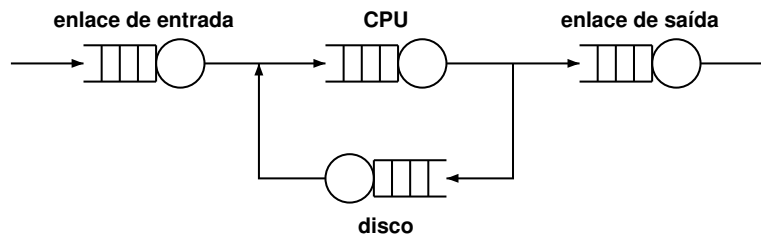


Figura 2. Modelo de rede de filas para um servidor web

3.2. Modelo elaborado para virtualização

Um modelo analítico para servidores web virtualizados permite estimar o desempenho dos servidores em máquinas virtuais com base em dados obtidos em um servidor não virtualizado, eliminando a necessidade de criar todas as MVs para obter dados experimentais de desempenho. Para desenvolver um modelo que considere a virtualização com base no modelo apresentado na seção anterior, foi necessário analisar de que forma o ambiente de virtualização influencia o desempenho dos recursos considerados pelo modelo, como mostrado na seção 3.2.1. A partir dessa análise, foram feitas modificações no modelo original para contemplar o uso de servidores virtualizados, descritas na seção 3.2.2.

3.2.1. Caracterização do impacto da virtualização

Para identificar os fatores que impactam o uso de CPU, decidiu-se por avaliar, de forma isolada, cada um dos recursos relacionados ao servidor web (rede, disco e CPU). Dentro de uma máquina virtual, foram usadas ferramentas para exercitar cada um dos recursos separadamente, causando o mínimo de interferência nos demais. O sistema sob teste foi monitorado usando a ferramenta *xentop*¹, que permite medir a utilização de CPU em cada MV e também em Dom0. As ferramentas usadas para gerar carga em cada recurso foram as seguintes:

- Rede: *netperf*², com o teste *TCP_STREAM*;
- Disco: *bonnie*³, com o teste *Writing Intelligently*;
- CPU: *NAS Parallel Benchmark*⁴, com o teste *Embarrassingly Parallel (EP)*.

Com os testes realizados, percebeu-se que a saturação dos processadores virtuais não influencia diretamente o consumo da CPU em Dom0, ou seja, mesmo que a utilização

¹<http://linux.die.net/man/1/xentop>

²<http://www.netperf.org/>

³<http://www.coker.com.au/bonnie++/>

⁴www.nas.nasa.gov/Software/NPB/

de CPU em máquina virtual estivesse em 100%, essa utilização se manteve no valor médio de 0,5% em Dom0. No entanto, ao avaliar os resultados dos testes sobre os recursos de rede e disco, verificou-se que uma carga sobre o processador é gerada pelas operações de E/S, em virtude do gerenciamento dos *drivers* das MVs exercido por Dom0. Ou seja, cada acesso ao recurso físico de rede ou disco realizado por uma MV (DomU) deve ser gerenciado pelo sistema operacional executado em Dom0, controlado pela ferramenta Xen [Cherkasova et al. 2007], como ilustra a figura 1(b).

Além da avaliação dos fatores que influenciam sobre o desempenho de CPU, fez-se necessário avaliar o quanto o ambiente de virtualização afeta o desempenho do disco e da rede. Para avaliar o desempenho do disco, utilizou-se de um programa desenvolvido na linguagem C, que mede o tempo necessário para ler um arquivo, simulando então a ação do servidor web. Esta ferramenta foi usada para medir a demanda de disco no ambiente nativo e no virtualizado, e verificou-se que o tempo necessário para ler o arquivo em ambos os ambientes se mantinha com valores comparáveis. Quanto à rede, foram feitas medições de vazão usando a ferramenta *Iperf*⁵, e se constatou que a redução de vazão decorrente do uso do ambiente de virtualização era inferior a 2%. Portanto, considera-se que a camada de virtualização não interfere significativamente sobre o desempenho isolado do disco e da rede em si.

3.2.2. Impactos sobre o modelo analítico

Com base na análise descrita na seção anterior constatou-se, primeiramente, que o uso do Xen não afeta significativamente o desempenho de rede e disco. Portanto, o tempo de rede continua sendo descrito pela equação (3), uma vez que as transmissões das várias MVs ocorrem em paralelo enquanto a largura de banda consumida for inferior à capacidade disponível. Por sua vez, o tempo de disco corresponde ao total do tempo de disco de todas as MVs. Considerando uma situação de carga homogênea, onde todas as MVs geram a mesma demanda de disco, essa situação é descrita pela equação (5), onde $R'_{disco_{mv}}$ é a residência em disco em cada MV, dada pela equação (2), e N é o número de MVs.

$$R'_{disco} = R'_{disco_{mv}} \times N \quad (5)$$

Com relação ao processador, a monitoração revelou que operações de E/S dentro de uma MV no Xen geram demanda na CPU, relacionada aos *drivers* de dispositivo que executam em Dom0. Portanto, a utilização do processador é afetada não só pelo compartilhamento de CPU entre as MVs como também pelo gerenciamento realizado pelo hipervisor. Isso se reflete, nas equações do modelo, da seguinte forma:

1. A taxa de processamento (*throughput*) em cada MV é igual à taxa de processamento do servidor nativo dividida pelo número de máquinas virtuais, o que se deve ao uso de cargas iguais entre as MVs e à característica da distribuição equilibrada de recursos do Xen. Obtém-se assim a equação (6), onde X_{mv} é a taxa de processamento em uma MV.

$$X_{mv} = \frac{X}{N} \quad (6)$$

⁵<http://iperf.sourceforge.net/>

2. As seguintes variáveis descrevem a utilização e o tempo de residência em CPU:

U_{Xen} :	utilização de CPU pelo Xen	R'_{cpu0} :	residência na CPU em Dom0
U_{cpu0} :	utilização de CPU em Dom0	R'_{cpumv} :	residência na CPU em máq. virtual
U_{cpunat} :	utilização de CPU na máq. nativa	R'_{cpu} :	residência final em CPU
U_{cpumv} :	utilização de CPU em máq. virtual		

3. Para determinar a utilização de CPU nas máquinas virtuais, é necessário determinar se há saturação da CPU física. Isso pode ser capturado por um parâmetro Ω , definido pela equação (7), que contabiliza a utilização total de CPU nas máquinas virtuais (dada pela utilização no servidor nativo multiplicada pelo número de MVs) somada a U_{Xen} :

- Quando $\Omega \geq 1$, há saturação de CPU, e a capacidade de CPU remanescente (descontando-se U_{Xen}) é dividida igualmente entre as máquinas virtuais;
- Quando $\Omega < 1$, não há saturação de CPU, e a utilização de CPU por máquina virtual U_{cpumv} é dada pela utilização no servidor nativo U_{cpunat} multiplicada por um fator de *overhead*.

A equação (8) mostra a utilização de CPU por máquina virtual.

$$\Omega = U_{cpunat} \times N + U_{Xen} \quad (7)$$

$$U_{cpumv} = \begin{cases} \frac{1-U_{Xen}}{N} & , \text{ se } \Omega \geq 1 \\ U_{cpunat} \times (1 + U_{Xen}) & , \text{ se } \Omega < 1 \end{cases} \quad (8)$$

4. A residência final em CPU a ser considerada deve ser igual à residência no ambiente de virtualização, somada ao tempo de residência em máquina virtual, este multiplicado pelo número de máquinas virtuais, como definido na equação (9).

$$R'_{cpu} = R'_{cpu0} + (R'_{cpumv} \times N) \quad (9)$$

O tempo de resposta final continua sendo dado pela soma entre as residências de CPU, rede e disco, expresso pela equação (4).

4. Avaliação Experimental

4.1. Descrição dos ambientes

Os testes foram realizados com o ambiente de virtualização Xen versão 3.2.1 instalado sobre um sistema operacional Debian Lenny, Linux 2.6.26, com servidor web Apache 1.3. As máquinas virtuais testadas foram criadas com 256 MB de memória RAM e 128 MB de *swap*. Foram usados dois computadores como servidores, possuindo as seguintes características:

- C1. AMD Sempron 2300+ 1,5 GHz, 1256 MB RAM, disco 40 GB IDE
- C2. AMD Phenom II X4 B93 2,8 GHz (*quad-core*), 4 GB RAM, disco 500 GB SATA

Como cliente, foi usada uma máquina com configuração idêntica à de C2. Cliente e servidor estavam ligados em uma rede Ethernet 100 Mbps através de um *switch*, não havendo tráfego concorrente de outras máquinas durante a realização dos testes.

A experimentação tomou como base quatro cenários distintos, três usando o computador C1 com diferentes tamanhos de arquivo (45 bytes, 45 KB e 450 KB) e um usando o computador C2 com o tamanho de arquivo intermediário (45 KB). Inicialmente, foram realizados experimentos em um servidor não virtualizado, com vistas a obter os parâmetros de entrada do modelo e assim permitir a estimativa do desempenho dos servidores

virtualizados, conforme será discutido na seção 4.2. A tabela 1 mostra esses dados de desempenho de base em cada um dos cenários. A carga de trabalho aplicada consistiu no envio de 10.000 requisições HTTP para arquivos com o tamanho considerado, usando o ApacheBench⁶. A taxa de processamento X foi fornecida pelo próprio ApacheBench. As utilizações de disco e CPU ($U_{disco_{nat}}$ e $U_{cpu_{nat}}$) foram medidas com o comando `vmstat` do Linux. A demanda de disco $D_{disco_{nat}}$ foi mensurada diretamente, usando o programa em C descrito na seção 3.2.1. Os dados da tabela correspondem à média de cinco execuções, sendo que o desvio padrão foi sempre inferior a 0,5% da média.

cenário		dados de desempenho (mensurados)			
computador	tam. arquivo	X (req/s)	$D_{disco_{nat}}$	$U_{disco_{nat}}$	$U_{cpu_{nat}}$
C1	45 B	1100,00	0,194 ms	0,11%	40%
C1	45 KB	238,09	1,452 ms	0,05%	6%
C1	450 KB	28,52	11,770 ms	0,11%	5%
C2	45 KB	208,33	0,0932 ms	0,07%	6%

Tabela 1. Dados de desempenho mensurados em servidores não virtualizados

4.2. Análise sobre modelo elaborado

Para a validação do modelo elaborado foram analisados os tempos de resposta de requisições HTTP obtidos experimentalmente em dois computadores, C1 e C2, usando o ApacheBench. Esses tempos de resposta serão comparados com os valores estimados pelo modelo introduzido na seção 3.2.2. Os parâmetros usados para o modelo são aqueles apresentados na tabela 1. A utilização de CPU para o gerenciamento do Xen (seção 3.2.2) foi mensurada no computador C2, e o valor obtido foi $U_{Xen} = 23\%$. Os dados experimentais correspondem à média de cinco execuções, sendo que o desvio padrão, em todos os casos, ficou abaixo de 4% da média.

Os resultados para o computador C1 são mostrados nas figuras 3(a) a 3(c). Foram usados três tamanhos de arquivo, 45 bytes, 45 KB e 450 KB, para analisar a sensibilidade do modelo em relação a esse parâmetro. O número de máquinas virtuais foi definido entre um e quatro, que é o máximo que pode ser criado com a memória física disponível. A curva tracejada apresenta o tempo de resposta medido, e a curva sólida o tempo de resposta obtido usando o modelo. Os números juntos a cada ponto representam a diferença porcentual entre os valores medidos e modelados. Com o arquivo menor (figura 3(a)), o modelo é pessimista (estima um tempo de resposta maior do que o mensurado) com qualquer número de máquinas virtuais, com diferenças entre 7 e 33%. Com os arquivos maiores (figuras 3(b) e 3(c)), o modelo apresenta resultados pessimistas com uma máquina virtual, e otimistas (estima um tempo de resposta menor do que o medido) a partir de duas máquinas virtuais. As diferenças nesses dois cenários variam entre 2,5 e 37,7%. Destaca-se que, em ambos os cenários, a menor diferença entre os resultados medidos e estimados é verificada com duas máquinas virtuais, sendo inferior a 6,7%. Analisando o conjunto dos gráficos, é possível observar que o comportamento para os arquivos maiores é comparável entre si, mas difere do comportamento encontrado para o arquivo menor.

A figura 3(d) mostra os resultados obtidos para o computador C2. Neste experimento, foram usados apenas arquivos de 45 KB. O número de máquina virtuais foi

⁶<http://httpd.apache.org/docs/2.0/programs/ab.html>

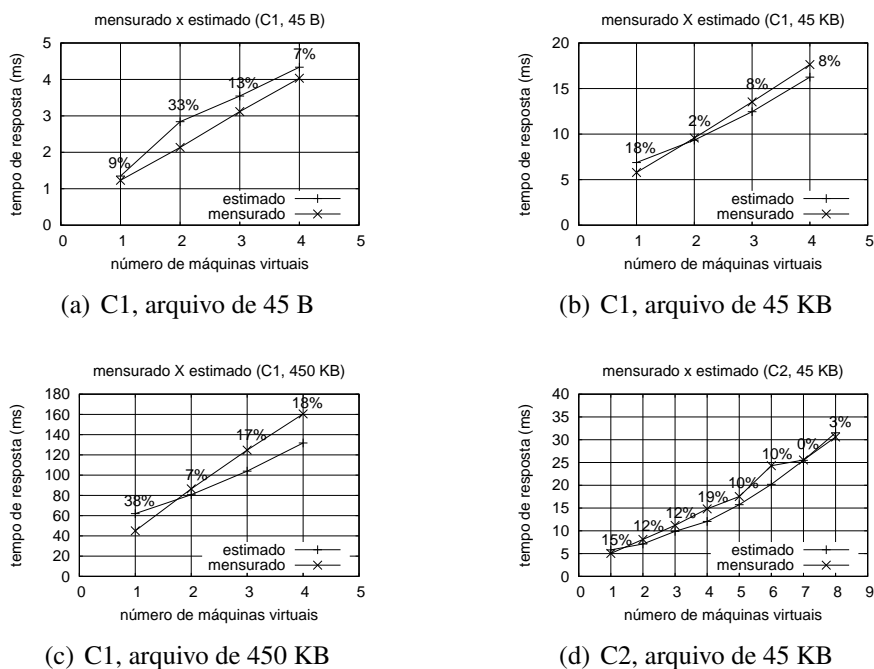


Figura 3. Comparação de resultados analíticos e experimentais

definido entre um e oito; esse número é maior do que o definido para o computador C1 porque C2 possui mais memória física. É possível observar que a diferença entre os valores mensurados e modelados ficou entre 0,3% (com sete máquinas virtuais) e 18,6% (com quatro MVs). Dos oito pontos observados, em dois deles (uma e oito máquinas virtuais) o modelo estima um tempo de resposta superior ao mensurado, e nos demais a estimativa é inferior ao tempo medido.

Discussão dos resultados. Analisando os resultados obtidos, é possível perceber que o modelo apresenta resultados mais precisos no quarto cenário (figura 3(d)), referente ao computador C2, que é o mesmo no qual foi obtido experimentalmente o valor adotado para U_{Xen} . O uso do U_{Xen} obtido em C2 nos experimentos realizados em C1 (resultados da figura 3) teve o propósito de averiguar se a mensuração desse parâmetro no hardware onde se pretende implementar a virtualização seria imprescindível para a utilização do modelo. Percebe-se que isso não é estritamente necessário, pois os resultados fornecidos pelo modelo ainda ficam próximos aos resultados experimentais. Entretanto, deve-se observar que os resultados do modelo seriam mais precisos se o valor de U_{Xen} fosse determinado experimentalmente no hardware a ser usado.

Margens de diferença de 20 a 30% são consideradas aceitáveis para fins de planejamento de capacidade [Lazowska et al. 1984, Menascé e Almeida 2003]. Dos 20 cenários avaliados, a diferença entre as medições e o modelo esteve acima de 30% em apenas dois deles, o que permite concluir que o modelo proposto oferece uma estimativa útil do desempenho de servidores web virtualizados com base em dados do servidor físico.

5. Conclusão

A consolidação de servidores usando ferramentas de virtualização é uma tendência crescente. Um fator importante na adoção desse tipo de solução é que o desempe-

nho dos servidores virtualizados se mantenha dentro do aceitável para seus usuários, não sendo excessivamente prejudicado pelo compartilhamento dos recursos do hardware subjacente. Atualmente, tal análise requer que primeiro se implante a virtualização, para depois verificar o seu desempenho.

A contribuição deste trabalho é a proposição de um modelo analítico que tem por objetivo fornecer estimativas de desempenho de servidores web virtualizados com Xen considerando as especificidades desse ambiente de virtualização, e se baseando em parâmetros que dispensem uma implantação prévia de virtualização. Em 18 dos 20 cenários estudados, o modelo forneceu resultados com uma margem de erro aceitável para fins de planejamento de capacidade, em comparação com dados experimentais.

Como continuação deste trabalho, pretende-se inicialmente realizar uma análise sobre a utilização de CPU associada a operações de E/S no Xen usando um conjunto amplo de máquinas distintas, com o intuito de determinar se o parâmetro U_{Xen} deve ou não ser recalibrado para cada hardware. Outra perspectiva é usar o modelo desenvolvido como uma ferramenta preditiva no gerenciamento de ambientes virtualizados ou computação em nuvem, permitindo que se escolha o servidor físico onde será instanciada uma máquina virtual de modo a satisfazer parâmetros de desempenho especificados.

Referências

- Barham, P., Dragovic, B., Fraser, K., et al. (2003). Xen and the art of virtualization. In *Proc. ACM Symp. on Operating Systems Principles (SOSP'03)*, pp. 164–177.
- Cherkasova, L., Gupta, D. e Vahdat, A. (2007). Comparison of the three CPU schedulers in Xen. *SIGMETRICS Perform. Eval. Rev.*, 35(2):42–51.
- Jain, R. (1991). *The Art of Computer Systems Performance Analysis*. Wiley.
- Kotsovinos, E. (2010). Virtualization: Blessing or curse? *Queue*, 8(11):40–46.
- Lazowska, E. D., Zahorjan, J., Graham, G. S. e Sevcik, K. C. (1984). *Quantitative System Performance: Computer System Analysis Using Queueing Network Models*. Prentice Hall.
- Menascé, D. A. (2005). Virtualization: Concepts, applications, and performance modeling. In *Proc. of the Computer Measurement Group Conference*, Orlando, FL.
- Menascé, D. A. e Almeida, V. A. F. (2003). *Planejamento de Capacidade para Serviços na Web: Métricas, modelos e métodos*. Editora Campus.
- Rosenblum, M. (2004). The reincarnation of virtual machines. *Queue*, 2(5):34–40.
- Tickoo, O., Iyer, R., Illikkal, R. e Newell, D. (2010). Modeling virtual machine performance: Challenges and approaches. *Performance Evaluation Review*, 37(3):55–60.

