Efeitos de Sobrecargas Transientes em Servidores Web

Paulo Roberto Farah[1], **Cristina Duarte Murta**[2]

[1] Faculdade Atual da Amazônia (FAA)
Rua Y, nº 308 – 69313792 – Boa Vista – RR – Brasil

[2]Departamento de Informática – Universidade Federal do Paraná (UFPR)
Caixa Postal 19.081 – 81531-990 – Curitiba – PR – Brasil

farah@faculdadeatual.edu.br, cristina@inf.ufpr.br

Resumo. *Sobrecarga transiente é o aumento súbito, inesperado e de curta duração da taxa de chegada de requisições legítimas em um servidor Web. Ocorrências de sobrecarga transiente em servidores Web vêm sendo observadas com maior freqüência à medida que o número de internautas aumenta e os serviços Web se tornam mais populares. Este artigo apresenta resultados de experimentos para caracterizar os efeitos de eventos de sobrecarga transiente no servidor Apache. Os resultados mostram que a sobrecarga transiente, mesmo em pequena intensidade, prejudica significativamente o desempenho do servidor, diminuindo sua taxa máxima de serviço e registrando tempos de resposta inaceitáveis. Os resultados contribuem para fundamentar novos projetos de servidores que respondam de forma adequada à sobrecarga transiente.*

Abstract. *Transient overload is a short term event characterized by a sudden and unexpected rise of the arrival rate in a Web server. Transient overload events in Web servers have been observed more frequently as the number of Internet users grows and the Web services become more popular. This paper presents an evaluation of the Apache Web server under transient overload events. The results shows that even a low intensity transient overload disrupts the server's performance, reducing its service rate and resulting in unacceptable response times. The results can be applied to the design of new servers.*

1. Introdução

Sobrecarga transiente é o aumento súbito e de curta duração da carga de um sistema, para valores acima de sua capacidade. Eventos de sobrecarga transiente em servidores Web são conhecidos como *flash crowds*, *hotspots* ou efeito *slashdot*, e se tornam cada vez mais comuns à medida que a utilização e o número de usuários da WWW aumentam. Em maior ou menor escala, picos de sobrecarga transiente ocorrem freqüentemente em servidores devido a diversas razões, sempre direcionadas pelos interesses dos usuários, tais como alterações no clima ou nas ações da bolsa de valores, feriados, eleições, guerras, desastres, ações terroristas, eventos culturais, anúncios de lançamentos de produtos, datas comemorativas, dentre outras. As requisições são legítimas, isto é, não configuram um ataque ao sistema, e devem ser tratadas da melhor forma possível pelo servidor.

Este trabalho apresenta uma avaliação dos efeitos de eventos de sobrecarga transiente no desempenho do servidor Web Apache. Várias condições de sobrecarga com tipos de carga diferentes são analisadas. Os resultados indicam que o desempenho do

servidor, de forma geral, é bastante comprometido em eventos de sobrecarga. A análise do comportamento de servidores em tais condições permite gerar conclusões e contribuições que podem ser utilizadas para aperfeiçoar o projeto e a melhorar a qualidade desses servidores.

Este artigo está organizado da seguinte forma. Na Seção 2 discutimos os trabalhos relacionados a esta pesquisa. A Seção 3 descreve o gerador de carga, o ambiente de teste e a metodologia de avaliação. A Seção 4 apresenta os resultados obtidos com as avaliações empíricas e é seguida pela conclusão.

2. Sobrecarga Transiente

Eventos de sobrecarga transiente são caracterizados por períodos curtos de carga intensa que se alternam com períodos longos de carga leve. A carga do servidor corresponde à taxa de chegada de requisições. Estes eventos podem ser recorrentes em servidores Web, devido às suas características de sistema aberto e às características da carga como, por exemplo, tamanho dos arquivos seguindo distribuições de cauda pesada [Almeida et al. 1996, Arlitt and Williamson 1996, Crovella et al. 1999], auto-similaridade da carga [Crovella and Bestavros 1996, Crovella and Bestavros 1997, Arlitt and Williamson 1996], dentre outras.

As características das sobrecargas transientes foram definidas em [Pan et al. 2004]. Os autores sugerem que eventos de grande interesse ocasionam esse tipo de carga, e que o volume de requisições de objetos populares pode aumentar dramaticamente, chegando atingir taxas dez a cem vezes maiores do que a taxa média observada no servidor. Esses eventos são de curta duração, o que indica que o superdimensionamento não é uma solução razoável, tendo em vista que o servidor será subutilizado na maior parte do tempo. O crescimento do volume de carga é instantâneo, o que torna difícil a detecção da sobrecarga antes que ela ocorra. Além disso os autores demonstram que recursos de CPU e largura de banda de rede são os primeiros a serem sobrecarregados. Até 60% dos objetos são acessados durante as súbitas rajadas de carga, o que indica que um servidor cache pode não conter todos os objetos durante a ocorrência de sobrecarga transiente. O uso de políticas de escalonamento que visam melhorar o desempenho de servidores Web com sobrecarga transiente foi proposto em [Harchol-Balter et al. 2003, Schroeder and Harchol-Balter 2003].

Alguns estudos caracterizam o comportamento das rajadas súbitas das sobrecargas transientes (*flash crowds*). A caracterização desse tipo de carga e a diferenciação entre sobrecarga transiente e ataque de negação de serviço (DOS) é apresentada em [Jung et al. 2002]. *Flash crowds* se caracterizaram pelo grande crescimento do número de clientes acessando os servidores, mantendo o perfil da popularidade dos arquivos. Os ataques DOS apresentam taxa de crescimento menor do número de clientes e o perfil de acesso aos arquivos não segue o padrão normal. A sobrecarga transiente foi caracterizada como uma das anomalias de fluxos de tráfego de redes por Crovella et al. em [Lakhina et al. 2004].

Analiticamente, a definição de sobrecarga transiente considera que o servidor se mantém estável durante o período de observação. Duas intensidades de carga são definidas: carga leve (*low*), com $\rho_l < 1$, observada durante um período t_l, e carga intensa (*high*), observada durante um período t_h no qual $\rho_h >= 1$. A intensidade de tráfego

média recebida pelo servidor pode ser calculada utilizando a equação 1 e a carga média deve ser, via de regra, menor do que 1 ($\rho < 1$) [Bansal and Harchol-Balter 2001].

$$\rho = \frac{t_h}{t_h + t_l}\rho_h + \frac{t_l}{t_h + t_l}\rho_l \quad (1)$$

Esta equação foi utilizada nesse trabalho para definir os valores das cargas intensa e leve, bem como os tempos de duração de cada tipo de carga.

Nosso trabalho difere das pesquisas citadas porque (i) caracterizamos o comportamento do servidor ao processar picos de sobrecarga com diferentes durações, (ii) utilizamos um gerador de sobrecarga transiente preciso para avaliar o comportamento do servidor, ao passo que nas pesquisas acima foram utilizados modelos e traces e (iii) utilizamos duas cargas sintéticas que nos permitiram isolar efeitos da alta variabilidade do tamanho dos arquivos de cargas Web reais. Não encontramos pesquisas com as características enumeradas acima, o que demonstra a originalidade e a contribuição deste trabalho.

3. O Gerador de Sobrecarga TORÓ

A geração de sobrecarga transiente foi realizada pelo gerador TORÓ. O TORÓ é um gerador escalável e robusto para geração efetiva de sobrecarga de servidores Web, que permite criar eventos de *flash crowd* de maneira controlada com alta precisão. Apresentamos o TORÓ e demonstramos sua precisão com resultados detalhados em [Farah and Murta 2006]. Nesta seção apresentamos uma descrição resumida do TORÓ. Em seguida, detalhamos a configuração do ambiente dos experimentos e a metodologia utilizada para caracterizar a sobrecarga transiente.

3.1. Descrição do TORÓ

O TORÓ foi projetado para suportar a criação massiva de requisições Web concorrentes e ser configurável para produzir efeitos de sobrecarga transiente de maneira prática e fácil. A especificação dos níveis de sobrecarga a serem gerados pelo TORÓ é feita em função da capacidade máxima útil do sistema a ser avaliado. Essa capacidade pode ser especificada pelo número de requisições por segundo ou pela taxa de transferência de dados do servidor para seus clientes, dada em bits por segundo. Esse valor define a carga a ser gerada pelo TORÓ para atingir a capacidade útil do sistema, e equivale a uma carga $\rho = 1$. Assim, os eventos de sobrecarga são especificados em função de ρ , por exemplo, $\rho = 1,10$ significa que o sistema receberá uma carga 10% acima de sua capacidade útil.

Geradores de carga e servidores Web são sistemas interdependentes. Para produzir sobrecarga nos servidores, os geradores precisam gerar continuamente novas requisições, independente das respostas às requisições anteriores. Se o gerador for implementado segundo o modelo fechado [Lazowska et al. 1984], utilizando, por exemplo, um *pool* de *threads*, as *threads* do conjunto podem ser esgotadas e novas requisições aguardarão na fila o término da execução de alguma requisição anterior, que, por sua vez, fica aguardando a resposta do servidor. Dessa forma, o gerador passa a depender da resposta do servidor e não consegue gerar carga acima da capacidade do servidor. Para evitar este problema, o TORÓ é projetado em modelo aberto, criando novas *threads* sempre

que necessário, com operações de entrada e saída não bloqueantes. Modelos de sistemas abertos e fechados são discutidos em [Lazowska et al. 1984, Schroeder et al. 2006].

O TORÓ utiliza um conjunto pré-definido de requisições (URLs) como carga, especificadas de acordo com o padrão *Common Log Format* [Bloom 2002]. Este conjunto pode ser gerado a partir de modelos analíticos definidos pelo usuário, possibilitando que o gerador utilize carga sintética, ou pode ser uma carga real, proveniente de um registro de acesso feito em servidores Web utilizados em ambientes de produção.

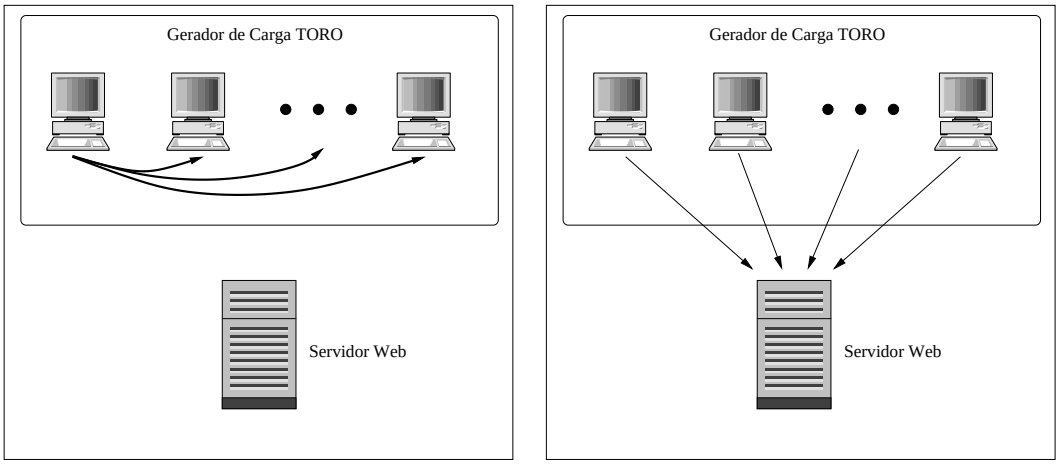
Geradores de carga tradicionais medem a intensidade da carga em requisições por segundo ou em número de clientes simultaneamente ativos. No entanto, devido a grande variabilidade observada no tamanho das respostas [Arlitt and Williamson 1996, Crovella et al. 1999, Crovella and Bestavros 1997], o servidor pode experimentar sobrecarga em recursos distintos. Por exemplo, um número pequeno de requisições para arquivos muito grandes pode facilmente encontrar limite na capacidade da rede, e não no servidor, enquanto um número grande de requisições para arquivos pequenos pode ser limitado pelos recursos internos do servidor. Assim, diferentes tipos de carga podem ocasionar sobrecarga em diferentes recursos de software e hardware.

Para tratar este problema e possibilitar a geração de carga precisa, o TORÓ utiliza duas métricas para definir a intensidade da carga gerada: número de requisições por segundo feitas ao servidor, e taxa de dados solicitada ao servidor, em bits por segundo. Os padrões de carga a serem gerados são definidos com base em uma destas métricas, sendo necessário escolher inicialmente a métrica que se deseja trabalhar.

As etapas para execução do gerador envolvem a configuração dos parâmetros da carga, o cálculo das requisições a serem efetuadas, a distribuição da carga para todos os nodos geradores de carga, o agendamento do horário de início da produção da carga, a realização das requisições e o registro de parâmetros de avaliação obtidos. Definido o padrão de carga desejado, a máquina mestre estabelece conexões com os demais nodos do gerador e envia a configuração da carga para estes nodos, conforme representado na Figura 1(a). Dados sobre as intensidades e suas respectivas durações, duração total do experimento e a lista de requisições são distribuídos para todos os nodos do gerador. Cada nodo calcula, então, a taxa de requisições que deve gerar para que o gerador produza as intensidades pré-definidas.

O próximo passo é agendar o horário de início da geração de carga em todos os nodos. Para gerar carga com precisão em intensidade e tempo, o gerador deve controlar o momento exato de envio das requisições. Assim, os relógios de todos os computadores envolvidos no processo de geração de carga devem estar sincronizados. Para sincronizar as máquinas utilizamos o protocolo NTP (Network Time Protocol) [NTP 2005], que é capaz de sincronizar o horário de computadores com precisão de milissegundos em redes locais, o que é plenamente suficiente para esta finalidade.

No horário determinado durante a fase de preparação, todos os nodos geradores de carga iniciam o envio de requisições ao servidor Web, conforme ilustra a Figura 1(b). Durante o período definido no experimento, o TORÓ produz a carga configurada, reproduzindo a seqüência de intensidades estabelecida até que a duração total do experimento tenha sido atingida. Para todas as requisições enviadas e respondidas, o gerador registra o horário em que a requisição foi efetuada e o horário em que a resposta enviada pelo



(a) Fase de preparação do gerador de carga. (b) Fase de produção da carga.

Figura 1. A fase de preparação da carga envolve a distribuição da configuração e o agendamento do horário de início da geração da carga. Durante a fase de produção, todas as máquinas produzem a carga de forma independente.

servidor foi recebida. O gerador também registra erros de conexão que tenham ocorrido e outras informações como o número identificador da requisição.

3.2. Vantagens do TORÓ

O gerador de carga TORÓ possui várias características desejáveis para a geração de sobrecarga transiente, descritas a seguir.

- **Escalabilidade:** Seu projeto escalável, *multithreaded* e com entrada e saída não bloqueante permite distribuir a geração de carga em várias máquinas e sobrecarregar efetivamente os servidores.
- **Flexibilidade:** A sobrecarga pode ser medida com duas métricas, requisições por segundo e bits por segundo, o que possibilita a análise de sobrecargas em diferentes recursos do servidor avaliado.
- **Praticidade:** Diversos tipos de carga e padrões de sobrecarga podem ser definidos, por exemplo, carga incremental e sobrecarga transiente, com parâmetros diversos.
- **Precisão:** O TORÓ foi desenvolvido para registrar com precisão parâmetros de desempenho e falhas monitorados. Essa sua característica é fundamental para que a análise de eventos de sobrecarga transiente seja realizada de maneira confiável.

3.3. Configuração do Ambiente

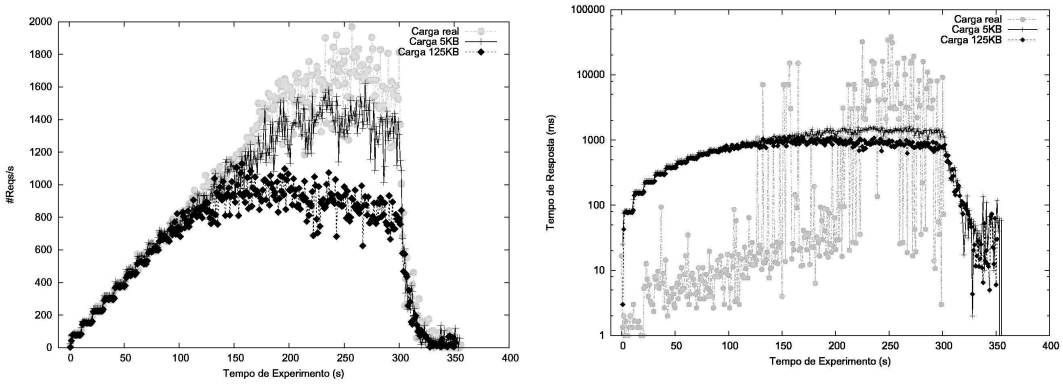
O TORÓ foi instalado em 16 microcomputadores Athlon XP 2200+, com 256MB de memória principal e sistema operacional Mandrake Linux 10.2. Todas estas máquinas operaram como geradoras de carga. As máquinas foram conectadas em uma rede local de 100 Mb/s, onde os experimentos foram realizados. A rede possui um comutador 3Com 3C16476 SuperStack 3 Baseline 10/100 Switch Plus Gigabit.

O hardware do servidor Web é um Athlon XP 1800+, com 512MB de memória principal, onde foi instalado o sistema operacional Debian Linux, kernel versão 2.4. O servidor Web Apache 2.0 módulo *worker* foi utilizado para a avaliação dos efeitos da sobrecarga transiente. O programa *sysstat/sar* foi instalado para monitorar o desempenho do servidor.

3.4. Metodologia de Avaliação

Inicialmente a capacidade máxima útil do servidor foi medida, aumentando-se progressivamente a taxa de requisições enviadas ao servidor (carga incremental). Em seguida, avaliamos o comportamento do servidor Apache com dois padrões de sobrecarga transiente denominados *impulsos* e *picos longos de sobrecarga*. Os impulsos de sobrecarga são inserções de sobrecarga com pequena duração, equivalente a 5% do tempo dos experimentos. Os picos longos de sobrecarga duram o equivalente a 20% do intervalo de tempo dos experimentos.

Uma carga real e duas cargas sintéticas foram utilizadas nos testes. A carga real foi obtida do servidor Web da copa do mundo de futebol de 1998 [Arlitt 2004]. A primeira carga sintética foi composta por arquivos de 5 KB e a segunda por arquivos de 125 KB. Foi importante utilizar as cargas sintéticas para isolarmos o efeito da variabilidade do tamanho dos arquivos da carga real.



(a) Número de requisições respondidas por segundo pelo servidor Apache.

(b) Tempos médios de resposta (ms).

Figura 2. Throughput e tempo de resposta para a carga incremental em função do tempo do experimento.

Para identificar os limites de capacidade do sistema avaliado, aplicamos o padrão de carga incremental, que aumenta progressivamente a taxa de chegada de requisições no servidor. A Figura 2(a) apresenta a taxa de serviço do servidor Web durante a aplicação desse padrão, para todas as cargas. Observamos que, no início da execução, as curvas apresentam degraus que correspondem exatamente ao padrão de carga incremental. Isto significa que o servidor consegue responder imediatamente as requisições feitas. Este padrão é seguido até o tempo de cerca de 100 segundos, quando, para a carga com o maior tamanho médio (125 KB), a taxa de respostas começa a se desviar da linha planejada.

Observamos também oscilações nas taxas de serviço para todas as cargas, conseqüência da carga mais elevada no sistema.

Valores máximos distintos da taxa de serviço, em requisições por segundo, podem ser observados para cada carga testada. A carga real, que apresenta o menor tamanho médio e um número grande de arquivos bem pequenos, consegue alcançar as maiores taxas de serviço (em req/s). Em seguida, temos a carga sintética cujo tamanho médio é 5 KB. Finalmente, a carga sintética com arquivos maiores alcança o menor limite superior para a taxa de serviço.

Os tempos médios de resposta do servidor Web para as cargas testadas são apresentados na Figura 2(b). O tempo médio de resposta ultrapassa e se mantém maior do que um segundo nos momentos em que o sistema fica sobrecarregado. Os tempos médios de resposta são proporcionais aos tamanhos médios da carga, isto é, a carga cujo tamanho médio é 125 KB obtém o maior tempo médio, seguida pela carga sintética de tamanho médio 5 KB, e pela carga real, que possui o menor tamanho médio. No entanto, pode-se observar que a carga real apresenta a maior variabilidade para o tempo de resposta. Esse comportamento ocorre em função da grande variabilidade dos tamanhos dos arquivos dessa carga.

Para definir um valor para a capacidade máxima útil do sistema, procuramos encontrar a taxa máxima de serviço que mantinha o sistema ainda operacional, com tempos de resposta aceitáveis. Assim, analisamos em conjunto os gráficos da Figura 2, e definimos os valores de 1000 reqs/s para a carga sintética de arquivos de 5 KB, 800 reqs/s para a carga composta de arquivos de 125 KB, e 1200 req/s para a carga real. Após essas respectivas taxas serem atingidas, o servidor apresenta alta variabilidade em seu *throughput*, indicando estar sobrecarregado.

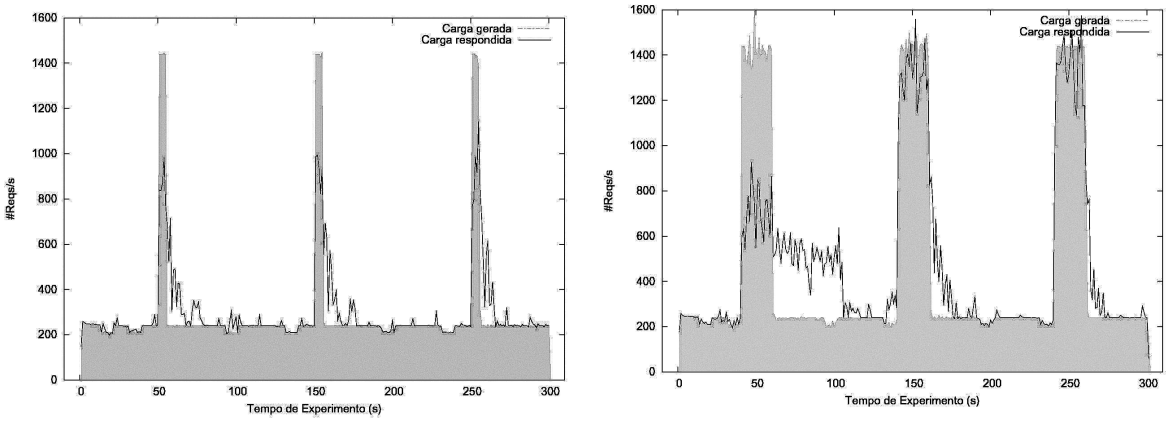
Para os testes com sobrecarga transiente, duas intensidades de carga foram definidas, uma denominada carga leve, com $\rho = 0,2$ e a outra, carga intensa, com $\rho = 1,2$. Estas intensidades referem-se à carga máxima, para cada tipo de carga testado. Assim, o servidor recebe um total de 240 reqs/s e 1440 reqs/s, para momentos de cargas leves e intensas, respectivamente, para a carga real. A carga de arquivos de 5KB, as taxas são de 200 reqs/s e 1000 reqs/s, respectivamente. E para a carga de arquivos com 125KB, as taxas são 160 reqs/s e 960 reqs/s, para períodos de carga leve e intensa, respectivamente. Foram realizados 27 medições com cada carga em cada padrão de sobrecarga transiente.

4. Caracterização da Sobrecarga Transiente

Nesta seção apresentamos detalhadamente os resultados. Demonstramos como diferentes padrões de carga e requisições a diversos tamanhos de arquivos afetam o desempenho do servidor Web ao processar sobrecargas transientes.

4.1. Efeitos no Throughput

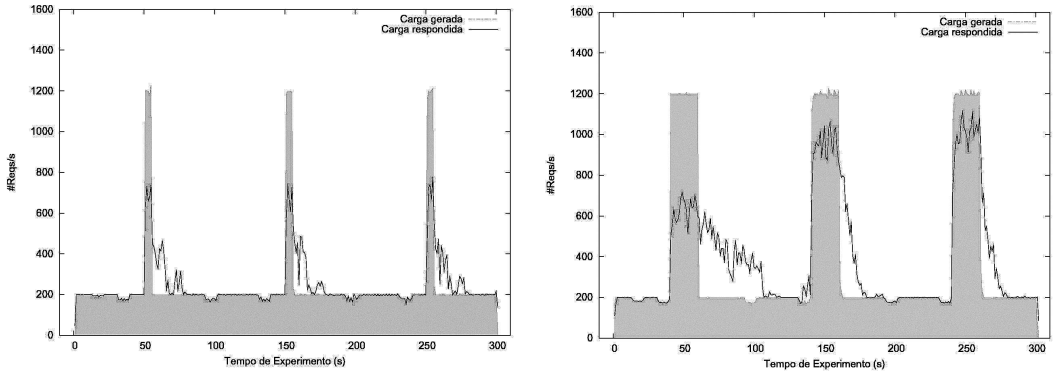
Iniciamos examinando a taxa de serviço do servidor para a carga real e em seguida para as cargas sintéticas. Nestes experimentos avaliamos a influência da duração do tempo de sobrecarga transiente no desempenho do Apache. A Figura 3 mostra a taxa de serviço medida para os padrões impulsos e picos longos de sobrecarga com a carga real, em comparação com a carga gerada. Pode-se observar que, ao responder a impulsos de sobrecarga, o servidor atinge uma taxa de serviço menor do que ao tratar os picos longos.



(a) Impulsos. (b) Picos longos.

Figura 3. Taxa de Serviço com a Carga Real.

Para o padrão de picos longos, o primeiro pico de sobrecarga diminui significativamente o desempenho do sistema mas nos próximos picos o desempenho melhora consideravelmente. Nos impulsos, o comportamento dos três picos segue o mesmo padrão. As taxas de serviço alcançadas durante os impulsos são muito inferiores às taxas dos picos longos de sobrecarga. Este comportamento do servidor foi verificado também para as demais cargas, de arquivos de 5KB e de 125KB, como é possível observar nas Figuras 4 e 5.

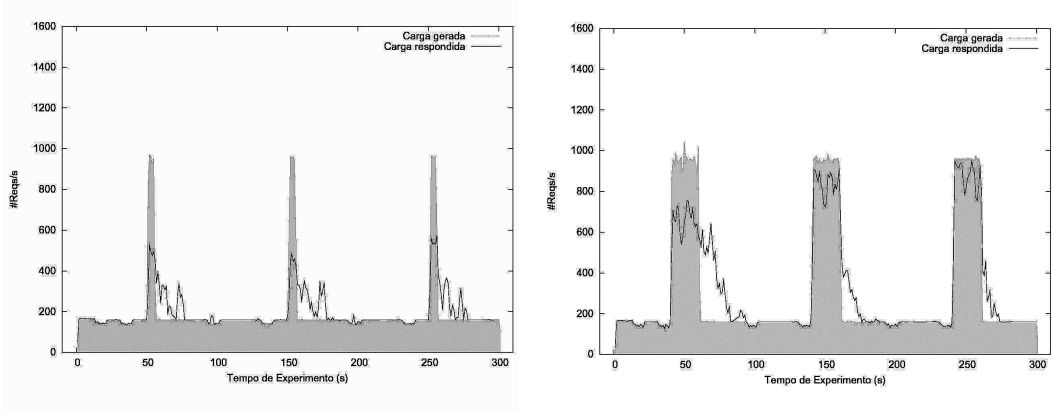


(a) Impulsos. (b) Picos longos.

Figura 4. Taxa de Serviço com a Carga 5KB.

Com exceção do primeiro pico do padrão de picos longos de sobrecarga, os impulsos apresentam maior variabilidade da taxa de serviço após o pico. No caso dos impulsos, observamos a variabilidade da taxa de serviço mesmo após o período de sobrecarga, enquanto no padrão picos longos a taxa de serviço é reduzida de modo mais gradativo. A maior variabilidade das taxas de serviço, além de valores menores, indicam maior dificul-

dade do servidor em responder a impulsos de sobrecarga.

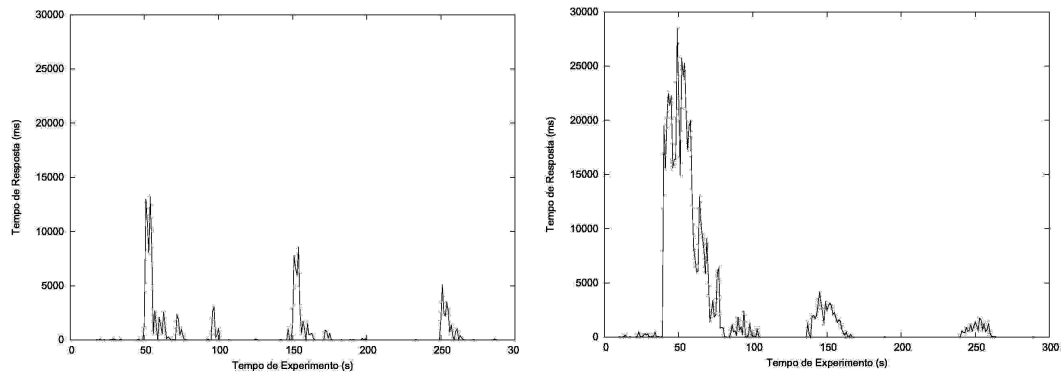


(a) Impulsos. (b) Picos longos.

Figura 5. Taxa de Serviço com a Carga 125KB.

4.2. Efeitos no Tempo de Resposta

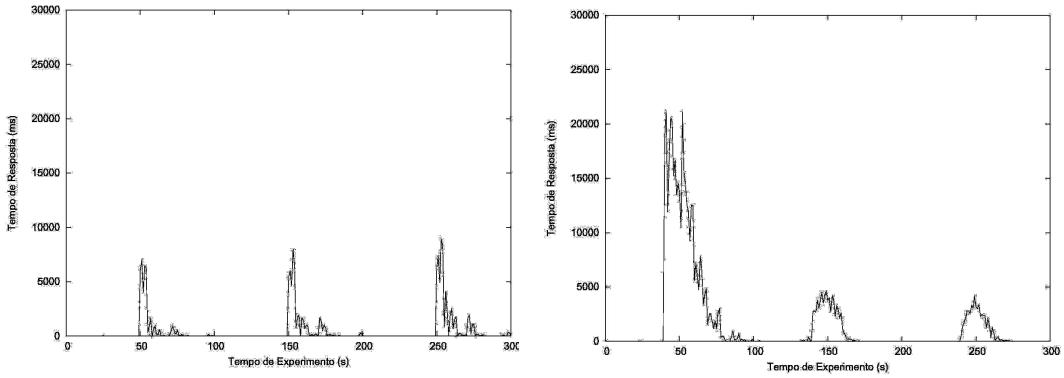
O tempo de resposta, ou latência, é uma métrica importante para a caracterização do desempenho de servidores Web. Examinamos o tempo de resposta do servidor ao receber as cargas de teste. O tempo de resposta foi medido desde o momento em que o gerador de carga inicia a conexão com o servidor Web até o momento em que o último byte é transmitido para a máquina geradora de carga. Nesse tempo de resposta estão incluídos os tempos de transmissão dos dados pelo meio físico, o tempo de processamento do protocolo de rede do cliente e do servidor e o tempo de processamento do servidor. As Figuras 6, 7 e 8 apresentam os tempos de resposta dos experimentos com as cargas real, de arquivos de 5KB e de 125KB, respectivamente.



(a) Impulsos. (b) Picos longos.

Figura 6. Tempos Médios de resposta com a Carga Real.

Os tempos de resposta alcançam valores muito elevados em todos os momentos de sobrecarga do sistema. Os maiores tempos de resposta são registrados nos primeiros picos longos de sobrecarga. Pode-se observar que o tempo de resposta do servidor para a carga real, composta em sua maior parte por arquivos pequenos, é maior do que o tempo de resposta da carga de arquivos com 125 KB. Isso ocorre porque, para a carga de arquivos menores, o servidor Web gerencia um número significativamente maior de requisições. As estruturas internas do servidor são sobrecarregadas e ocasionam esse aumento evidente do tempo de resposta.



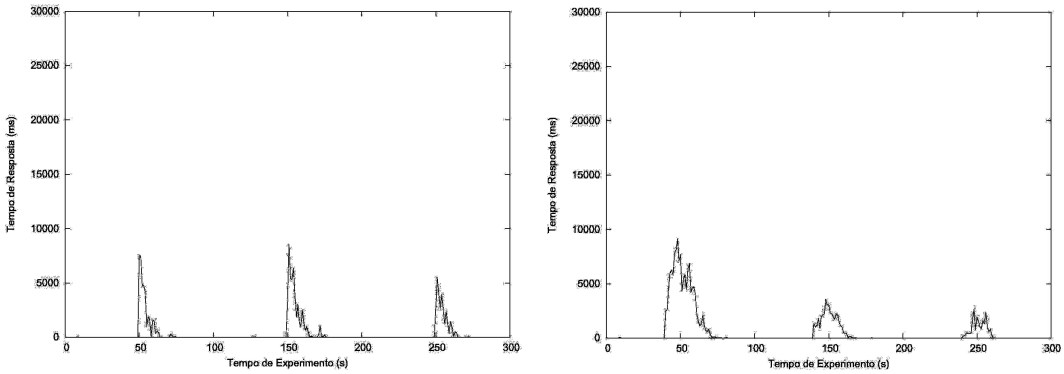
(a) Impulsos. (b) Picos longos.

Figura 7. Tempos Médios de resposta com a Carga 5KB.

4.3. Efeitos na Taxa de Admissão de Requisições

Outra técnica empregada para comparação do comportamento do servidor frente a várias cargas foi a correlação entre a quantidade de requisições solicitadas pelo gerador e respondidas pelo servidor Web. Para demonstrar esta correlação, os dados são representados por pares ordenados (x, y) onde x é a variável independente e y é a variável dependente. A correlação pode ser observada em mapas de dispersão ou pelo coeficiente de correlação.

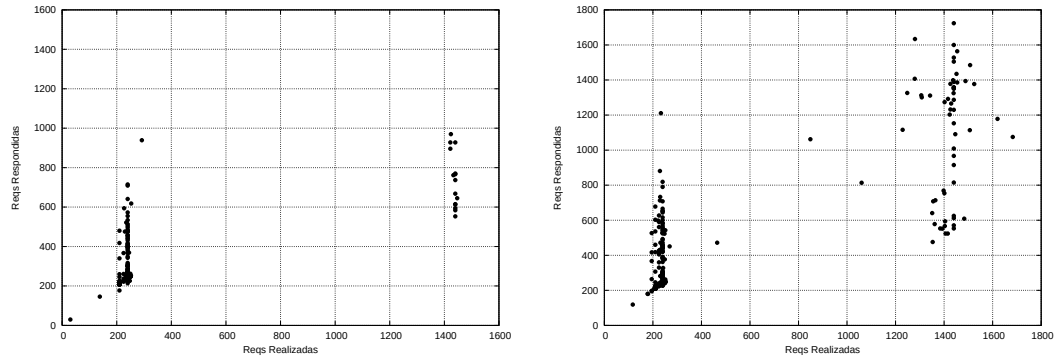
Os gráficos apresentados na Figura 9 indicam a correlação entre a quantidade de requisições realizadas pelo TORÓ e respondidas pelo servidor Web, para experimentos realizados com impulsos e picos longos de sobrecarga transiente, com a carga real. Esses gráficos correlacionam o comportamento do gerador de carga e do servidor Web em cada momento. Cada ponto do gráfico representa uma unidade de tempo do experimento, medida em segundos. Os pontos cujos eixos x e y possuem valores aproximados ou iguais indicam que o servidor respondeu aproximadamente ou exatamente a mesma quantidade de requisições realizadas. Em uma situação ideal, os experimentos realizados com impulsos e picos de sobrecarga para a carga real devem registrar uma grande quantidade de pontos no valor 240 para os eixos x e y. Esse valor representa o número de requisições geradas para a intensidade de ρ igual a $0,2$ das cargas de trabalho aplicadas ao servidor. Durante os momentos de carga intensa, os valores mostrados pelo gráfico devem ser de 1440 para os eixos x e y. Esse comportamento ideal indica que o gerador requisitou essa quantidade de requisições a cada segundo e o servidor respondeu a essa taxa no mesmo



(a) Impulsos. (b) Picos longos.

Figura 8. Tempos Médios de resposta com a Carga 125KB.

período. Quando um ponto do gráfico possui valores diferentes desses valores em relação ao eixo `x`, isso indica que o gerador não conseguiu estabelecer conexão com o servidor e gerar a carga planejada. Se a variação ocorrer em relação ao eixo `y`, o servidor Web não foi capaz de responder o número de requisições solicitadas.



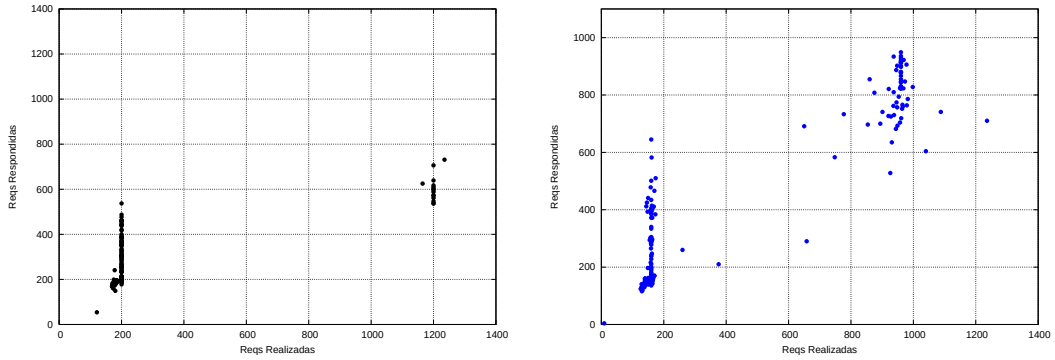
(a) Impulsos. (b) Picos longos.

Figura 9. Mapa de Dispersão da correlação existente entre o número de requisições realizadas pelo Toró e respondidas pelo Apache com a Carga Real.

Pode-se observar que existem vários pontos relativos à carga leve em que a quantidade de requisições respondidas é muito maior do que a solicitada. Nesses momentos, o servidor ainda está tratando requisições solicitadas durante os momentos de pico. Esses casos ocorrem com maior freqüência do que os momentos em que o servidor respondeu corretamente a mesma quantidade de solicitações recebidas durante a sobrecarga. Para os impulsos de sobrecarga, o servidor não conseguiu responder a taxa de requisições solicitada. Observe que a taxa máxima de requisições respondidas pelo servidor é inferior a 1000 req/s, enquanto os impulsos equivalem a 1440 req/s.

Para o padrão de picos longos de sobrecarga, há pontos que variam em relação

aos valores 240 e 1440 no eixo x, como podemos observar na Figura 9(b). Isto indica a dificuldade do servidor em receber novas conexões do gerador de carga com o padrão de picos longos. Contudo, existem vários pontos com alta correlação. Em cerca de 12% do tempo do experimento com esse padrão, o servidor apresentou dificuldade para tratar novas conexões provenientes do Toró.



(a) Carga de 5 KB. (b) Carga de 125 KB.

Figura 10. Mapa de dispersão da correlação existente entre o número de requisições realizadas pelo TORÓ e respondidas pelo Apache para o padrão picos longos com as cargas de 5 KB e 125 KB.

Analisamos a influência que o tamanho dos arquivos exerce sobre o desempenho do servidor Web avaliado. A Figura 10 ilustra a correlação entre as requisições efetuadas pelo Toró e respondidas pelo Apache para as cargas sintéticas de arquivos de 5 KB e 125 KB para o padrão de picos longos de sobrecarga. Pode-se observar que, para a carga de arquivos de 125 KB, o mapa de dispersão apresenta maior variação das requisições realizadas (eixo x) do que o mapa de dispersão da carga de 5 KB. Esse comportamento foi semelhante para os impulsos de sobrecarga, mas em menor escala. Esse comportamento sugere que o servidor Web apresenta dificuldade maior para processar novas requisições que chegam ao sistema quando está sobrecarregado com requisições de arquivos grandes.

4.4. Efeitos na Resposta das Requisições

A Tabela 1 apresenta valores médios de métricas medidas no servidor durante os momentos de atraso gerados pelos padrões de sobrecarga impulsos e picos longos, para todos os experimentos realizados. Os impulsos de sobrecarga apresentam maior percentual de requisições realizadas durante a carga intensa e respondida após esse momento. Dividimos o tempo total do atraso para o servidor responder todas as requisições efetuadas durante os impulsos e picos de sobrecarga pelo número de requisições solicitadas nesses períodos. Analisando os resultados para as cargas sintéticas, pode-se observar que esse tempo de atraso proporcional por requisição não respondida no momento de alta intensidade é maior para arquivos com tamanho maior. Além disso, foi possível constatar que o tempo proporcional para os impulsos é maior do que para os picos longos. Esse comportamento indica evidências de que o desempenho do servidor pode ser pior ao receber vários impulsos de sobrecarga em curto espaço de tempo do que ao receber picos longos de sobrecarga.

Tabela 1. Número médio de requisições, percentual de requisições e tempo médio de atraso para requisições respondidas após os momentos de sobrecarga.

	Impulsos				Picos Longos			
Carga	#Reqs	%	Tempo	tempo/req	#Reqs	%	Tempo	tempo/req
Real	3532	49,14	27,27s	0,0077s	7729	26,83	64,05s	0,0082s
5 KB	3173	52,87	31,27s	0,0098s	7437	31,40	64,89s	0,0087s
125 KB	2175	45,31	32,88s	0,015s	3992	20,78	41,92s	0,010s

Analisando os resultados obtidos com os experimentos que aplicaram impulsos e picos longos de sobrecarga com a carga real, o percentual de requisições realizadas nos momentos de sobrecarga é maior nos impulsos do que nos picos longos. Por outro lado, o tempo proporcional para respostas das requisições atrasadas é maior para os picos longos de sobrecarga. Esse tempo é maior porque o gerador requisita arquivos com tamanho grande durante os momentos de pico e não o faz para os impulsos.

5. Conclusão

Este artigo apresentou uma caracterização de desempenho do servidor Web Apache ao processar diferentes tipos de sobrecarga transiente, com padrões de impulsos e picos longos de sobrecarga. Constatamos que o servidor Web apresentou melhor desempenho quando processou picos longos de sobrecarga do que quando processou impulsos de sobrecarga. Mapas de dispersão da correlação do número de requisições realizadas e respondidas indicaram redução na taxa de chegada de requisições ao servidor. O tempo de atraso por requisição foi maior à medida que o tamanho dos arquivos aumentou.

Os resultados indicam que quanto mais variável é a carga, maior a variabilidade no desempenho do servidor. Isto demonstra que o servidor precisa tratar melhor a variabilidade do conjunto de tarefas, e isto pode ser feito com políticas de escalonamento baseadas no tamanho do arquivo a ser servido, que é conhecido a priori. Outro aspecto refere-se à identificação do recurso que experimenta sobrecarga; no caso da carga de arquivos maiores, é a interface de rede, e no caso da carga de arquivos pequenos, são os recursos de software tais como processos, threads e descritores de arquivos. O conhecimento destes aspectos ajuda no dimensionamento do servidor.

Vários aspectos do desempenho de servidores Web em condições de sobrecarga transiente podem ser explorados. Estamos estendendo esta pesquisa para analisar diferentes intensidades de impulsos e picos longos de sobrecarga e avaliar o impacto dessas alterações no desempenho de servidores Web.

Referências

Almeida, V., Bestavros, A., Crovella, M., and Oliveira, A. (1996). Characterizing Reference Locality in the WWW. In *Proc. IEEE/ACM Int. Conf. on Parallel and Distributed System*.

Arlitt, M. (2004). 1998 World Cup Web Site Access Logs. Disponível em: <http://ita.ee.lbl.gov/html/contrib/WorldCup.html> Acesso em Setembro de 2004.

Arlitt, M. F. and Williamson, C. L. (1996). Web Server Workload Characterization: The Search for Invariants. In *Proc. ACM SIGMETRICS Conference on Measurement of Computer Systems*, Philadelphia, PA.

Bansal, N. and Harchol-Balter, M. (2001). Scheduling Solutions for Coping with Transient Overload. Technical Report CMU-CS-01-134, Carnegie Melon University.

Bloom, R. B. (2002). *APACHE Server 2.0: The Complete Reference*. McGraw-Hill.

Crovella, M., Barford, P., Bestavros, A., and Bradley, A. (1999). Changes in Web Client Access Patterns: Characteristics and Caching Implications. *World Wide Web, Special Issue on Characterization and Performance Evaluation*, 2:15–28.

Crovella, M. and Bestavros, A. (1996). Explaining World Wide Web Traffic Self-Similarity. In *Proc. ACM SIGMETRICS Conference*, Philadelphia, PA.

Crovella, M. E. and Bestavros, A. (1997). Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Trans. on Networking*, 5(6):835–846.

Farah, P. R. and Murta, C. D. (2006). TORÓ: Um Gerador de Carga para Servidores Web. In *V Workshop de Desempenho de Sistemas Computacionais e de Comunicação (Wperformance06)*, Campo Grande, MS, Brasil.

Harchol-Balter, M., Schroeder, B., Bansal, N., and Agrawal, M. (2003). Size-based Scheduling to Improve Web Performance. *ACM Transactions on Computer Systems*, 21(2):1–27.

Jung, J., Krishnamurthy, B., and Rabinovich, M. (2002). Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites. In *The 11th Int. World Wide Web Conf.*, Honolulu, Hawaii, EUA.

Lakhina, A., Crovella, M. E., and Diot, C. (2004). Characterization of Network-Wide Anomalies in Traffic Flows. In *IMC'04*, Sicília, Itália.

Lazowska, E. D., Zahorjan, J., Graham, G. S., and Sevcik, K. C. (1984). *Quantitative System Performance: Computer System Analysis using Queueing Network Models*. Prentice Hall, Inc., Upper Saddle River, NJ, USA.

NTP (2005). NTP: The Network Time Protocol. Disponível em: <http://www.ntp.org> Acesso em Janeiro de 2005.

Pan, C., Atajanov, M., Shimokawa, T., and Yoshida, N. (2004). Flash Crowds Alleviation via Dynamic Adaptive Network. In *Proc. Internet Conference 2004*, pages 21–28.

Schroeder, B. and Harchol-Balter, M. (2003). Web Servers under Overload: How Scheduling can Help. In *18th International Teletraffic Congress*, Berlin, Germany.

Schroeder, B., Wierman, A., and Harchol-Balter, M. (2006). Open Versus Closed: A Cautionary Tale. In USENIX, editor, *Proc. NSDI '06: 3rd Symposium on Networked Systems Design and Implementation*.