Cache na WWW: Limitações e Potencial

Cristina Duarte Murta[1] Virgilio A. F. Almeida[2]

[1] Departamento de Informática
Universidade Federal do Paraná
Curitiba, PR, Brasil
cristina@inf.ufpr.br
[2] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brasil
virgilio@dcc.ufmg.br

Resumo—

WWW caching systems are essential for Web performance and scalability. Web cache workloads and goals show important differences when compared with traditional caching systems, such as memory caches. This article points out and discusses these differences. In order to illustrate our discussion, we present characterization of ten workloads from Web caches. The analysis focuses on the parameters response size and access patterns. We discuss the influence of the characterized parameters in cache performance according to relevant metrics for the WWW and we show directions for the design of Web cache replacement policies.

Keywords— **WWW, Caching, Internet performance.**

I. Introdução

Caches tradicionais tais como os encontrados em hierarquias de memória [PJH 96], e caches da WWW são fundamentados no princípio da localidade. Em ambos os casos, o objetivo principal é reduzir o tempo de acesso aos dados. No entanto, a utilização de caches na WWW traz vantagens adicionais importantes, a saber, redução da carga na rede Internet, obtida pela eliminação de transmissões redundantes, redução da carga nos servidores originais dos documentos, e aumento da disponibilidade dos documentos na Web.

Devido à sobrecarga na Internet, a vantagem dos caches da WWW de reduzir a carga na rede passa a ser um objetivo destes caches. Para aumentar a escalabilidade da WWW e reduzir a possibilidade de congestionamento, a ordem é maximizar a utilização da infra-estrutura existente na Internet, eliminando usuários e tráfego redundante da rede através da utilização de servidores cache. Reduzir o tráfego é especialmente interessante para os administradores da rede e para provedores de acesso e de serviço, que pagam por banda de acesso. Por outro lado, reduzir o tempo de acesso é interessante para o usuário que quer ver a página pedida o mais rápido possível. Portanto, a ação dos caches WWW passa a ter pelo menos dois objetivos: economia no uso da rede e melhoria da latência.

Outra diferença entre os caches tradicionais e da WWW é em relação à carga de trabalho que cada um experimenta. A WWW é composta por milhões de objetos, usuários e servidores. Para seus sistemas de cache, isto significa que há um grande número de objetos candidatos ao armazenamento. É observada também uma variabilidade extrema nas medidas da carga, por exemplo, nos tamanhos dos documentos que circulam na Web [MTA 98]. Há documentos que são várias ordens de grandeza maiores do que outros. Enquanto os caches tradicionais existentes nos sistemas de computação tratam de objetos de tamanho único (ex.: bloco ou linha), os sistemas de cache na Web operam com objetos de tamanhos extremamente variáveis.

Nos esquemas tradicionais de cache, o número de bytes servidos pelo cache dividido pelo número de bytes requisitados é exatamente a fração de *hits*. O mesmo não ocorre nos caches da Web porque os documentos são transmitidos e armazenados na sua forma integral; não há o conceito de bloco. Como consequência, precisamos trabalhar com duas métricas para medir o desempenho dos sistemas de cache na Web: a fração das requisições que é atendida pelo cache, (*Hit Ratio*, HR) e a fração dos bytes requisitados que é atendida pelo cache, (*Byte Hit Ratio*, BHR).

A construção de soluções eficientes para o gerenciamento do espaço de caches requer um conhecimento detalhado das propriedades da carga. Este conhecimento ajuda no dimensionamento do cache e na definição de uma política de substituição adequada ao padrão de acesso identificado.

Este artigo apresenta caracterização da carga de dez caches reais da WWW e identifica as propriedades da carga que afetam as métricas HR e BHR. São reveladas características comuns, que são invariantes para o conjunto de cargas e que são relevantes para o desempenho dos caches em termos das métricas em avaliação. A análise focaliza os parâmetros tamanho das requisições e dos arquivos únicos, e o padrão de acesso identificado nas cargas. A variabilidade nos tamanhos e suas implicações são mostradas. O entendimento do padrão de acesso revela o potencial e as limitações dos caches na WWW.

TABELA I
IDENTIFICAÇÃO, ORIGEM E DATA DAS CARGAS CARACTERIZADAS.

Carga	Instituição	Data
BL	VTech	set a out 1995
BR	VTech	set a out 1995
U	VTech	abril a out 1995
NASA	Nasa	04-31 ago 1995
POP98	POP-MG	24 out a 4 nov 1997
Portugal	Esotérica	01-07 mar 1998
POP99-hug	POP-MG	12-15 jan 1999
POP99-zez	POP-MG	12-14 jan 1999
NLANR-uc	NLANR	14 jan 1999
NLANR-bo1	NLANR	14 jan 1999

II. CARGAS DE TRABALHO ANALISADAS

A carga de um cache consiste do conjunto de requisições recebidas durante um período de observação. Com o objetivo de generalizar os resultados deste trabalho, foram analisadas cargas de dez caches da WWW. Estas cargas foram obtidas a partir do registro em ordem cronológica (*trace*) das requisições feitas a sistemas de caches reais. O conjunto apresenta grande amplitude em vários aspectos. Os registros foram feitos ao longo de cinco anos (1995 a 1999), em caches de cliente, de servidor e de rede, em países de três continentes, em estágios diferentes de implantação da Internet. As instituições são pioneiras em alguns aspectos. Esoterica [ESO 98], POP-MG [POP 98] e NLANR [NLA 99] implementam os maiores caches de rede em seus respectivos países. VTech [VTE 95] foi uma das primeiras instituições a registrar e analisar tráfego da WWW. A tabela I apresenta a identificação das cargas, sua origem e data de registro. As cargas BR e NASA são de caches de servidor, a carga BL é de cache de clientes e as demais são de cache de rede.

Todas as cargas utilizadas neste trabalho são resultado de pré-processamento aplicado aos *traces* originais. O objetivo do pré-processamento é separar as requisições que são pedidos de arquivos ao cache e que alteram o estado de seu sistema de armazenamento ou das listas de metadados, de outras requisições como, por exemplo, o pedido de verificação de presença ou de validade de documentos. As cargas correspondem às requisições TCP com método GET, que foram transmitidas com sucesso (código 200). Requisições para objetos dinâmicos foram excluídas.

III. CARACTERIZAÇÃO DOS TAMANHOS

A caracterização dos tamanhos é feita em dois conjuntos: o conjunto de arquivos transmitidos em resposta às requisições dos clientes, denominado "requisições", e o conjunto de arquivos únicos transmitidos, denominado "arquivos". O conjunto de requisições é composto por todos os URLs requisitados pelos usuários. Este conjunto pode conter várias vezes o mesmo URL. A repetição ocorre porque um usuário requisita o mesmo arquivo mais de uma vez ou porque usuários diferentes requisitam o mesmo arquivo. Muitas destas requisições repetidas resultam em *hits* no cache, o que significa que elas são satisfeitas sem gerar tráfego no trecho da rede entre o cache e o servidor.

O conjunto de arquivos contém exatamente uma entrada para cada arquivo distinto, independente do número de vezes que ele foi requisitado. Este é, portanto, um subconjunto do conjunto de requisições. O tamanho associado a cada URL é o tamanho da resposta, ou seja, do conteúdo do URL, e não o tamanho da requisição propriamente dita.

A seqüência de requisições contém informação sobre localidade. O conjunto de arquivos é o conjunto de documentos que são armazenados no cache. Sua caracterização é fundamental para a análise de desempenho dos Web-caches. As duas seções seguintes apresentam a caracterização dos dois conjuntos definidos para todas as cargas.

A. Tamanhos das Requisições

A tabela II apresenta a caracterização básica de cada carga, a saber, o número de requisições, o total de bytes transmitidos, o tamanho médio das requisições e a mediana. No total são analisadas 8746315 requisições que compreendem mais de 110 gigabytes de dados transmitidos.

TABELA II
CARACTERIZAÇÃO DAS REQUISIÇÕES: NÚMERO, TOTAL DE BYTES TRANSMITIDOS (EM MBYTES) E ÍNDICES DE DISPERSÃO PARA O TAMANHO: MÉDIA E MEDIANA EM BYTES, MAIOR DOCUMENTO EM MBYTES.

Carga	Requis.	Bytes	Média	Mediana	Maior
BL	53399	672	12600	2654	7
BR	179600	10070	56074	1999	10
U	173597	2070	11927	2268	18
NASA	1385259	26766	19323	4179	3
POP98	2111766	19142	9065	3235	20
Portugal	1193404	10780	9033	2779	17
POP99hug	1121747	11505	10257	2929	27
POP99zez	1120830	17062	15223	2905	66
NLANR-uc	800534	11291	14105	3461	53
NLANR-bo1	606179	9028	14894	3683	49

A observação dos dados revela evidências de variabilidade. A mediana está entre 2 e 3.7 kbytes. A média está entre 9 e 19 kbytes, com exceção para a carga BR. A carga BR contém requisições repetidas para arquivos de áudio, que são muito grandes, e elevam a média para 56 kbytes. O fato de a mediana ser menor do que a média indica que os dados seguem uma distribuição com cauda à direita. Os maiores documentos são da ordem de 10^7 bytes, seis ordens de grandeza maiores do que os menores documentos. Estes índices de dispersão põem em evidência a grande variabilidade existente nos tamanhos das requisições. Pelos dados da tabela II podemos calcular que milhares de documentos de tamanho igual à mediana podem ocupar o mesmo lugar de apenas um

dos maiores arquivos.

A figura 1 mostra a distribuição de probabilidade acumulada dos tamanhos de cada conjunto de requisições. A idéia não é identificar cada curva em particular, mas mostrar que o comportamento das distribuições é bastante similar. Para todas as cargas, as requisições de tamanho pequeno são a maioria. Há também um conjunto pequeno de requisições muito grandes. O eixo x está em escala logarítmica. As cargas NASA e BR se diferenciam das demais cargas. BR apresenta um número maior de requisições pequenas, menores do que 1 kbyte. Ambas apresentam um número maior de requisições grandes do que as outras cargas. Este comportamento é consistente com os valores das médias e medianas apresentados na tabela II.

A distribuição de Pareto é apontada em [MAW 96, PBM 98] como a que melhor representa os tamanhos dos objetos da WWW, em especial a cauda da distribuição dos tamanhos. As distribuições de cauda pesada, como a de Pareto, apresentam propriedades que são qualitativamente diferentes das distribuições identificadas mais comumente, tais como a exponencial e a normal. A função de densidade de probabilidade de uma distribuição de cauda pesada apresenta uma cauda que prolonga-se à direita. Distribuições de cauda pesada são caracterizadas por uma variabilidade extrema, que cresce muito à medida que o parâmetro da variabilidade α decresce, variando no intervalo $0 < \alpha < 2$.

Para mostrar a variabilidade, estamos interessados na distribuição que representa a cauda. A figura 2 mostra em um gráfico log-log o complemento da distribuição de probabilidade para os conjuntos de requisições e para as funções exponencial e Pareto. Neste gráfico, a inclinação da curva indica o valor do parâmetro α para a distribuição de cauda pesada correspondente. A reta corresponde à distribuição de Pareto com $\alpha = 1.3$. O gráfico demonstra que as caudas de todas as cargas têm decaimento similar ao apresentado pela distribuição de Pareto para tamanhos entre 10^4 e 10^6. As cargas NASA, BR e BL apresentam decaimento rápido para valores de tamanho maiores que 10^6 (extremo da cauda) enquanto as cargas restantes continuam seguindo Pareto.

O método *scaling* [MCT 99] foi utilizado para estimar os valores de α para todas as cargas. Os valores de α, entre 0.95 e 1.50, confirmam as evidências de que as distribuições são de cauda pesada, exibindo, portanto, grande variabilidade.

Pelos valores das medianas e dos tamanhos das maiores requisições, mostrados na tabela II, podemos observar um crescimento do tamanho dos arquivos ao longo do tempo. As cargas de cache de rede registradas em 1998 e 1999 apresentam medianas maiores do que as medianas das cargas registradas anteriormente. Esta informação pode ser reforçada pela observação de que as cargas de 1999 também registram os maiores arquivos transmitidos e são uma evidência de que a variabilidade nos tamanhos tende a aumentar. Paralelamente, não foi registrado nenhum aumento no tamanho do menor documento.

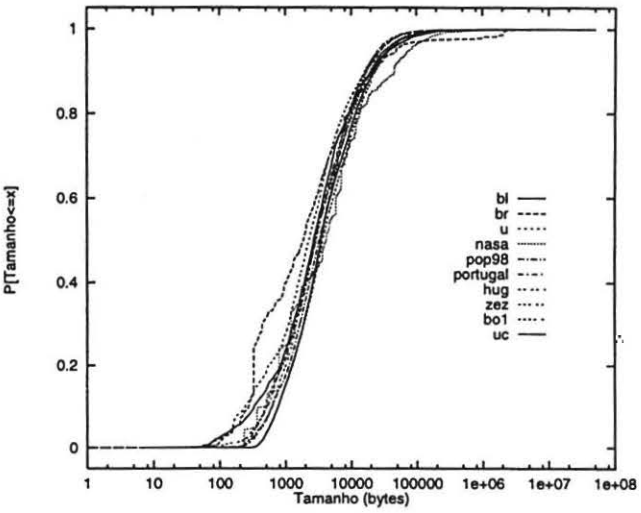

Fig. 1. Distribuição de probabilidade acumulada empírica dos tamanhos do conjunto de requisições para todas as cargas.

B. Tamanhos dos Arquivos Únicos

As estatísticas dos conjuntos de arquivos são apresentadas na tabela III. O número de arquivos e o total de bytes está representado como uma fração do número de requisições e do número de bytes requisitados, respectivamente, apresentados na tabela II. O conjunto de arquivos representa uma porcentagem significativa do conjunto de requisições, entre 36% e 67% (com exceção para BR e NASA), o que corresponde a um total de bytes únicos que está entre 46% e 74% dos bytes transmitidos. As cargas BR e NASA apresentam uma fração muito menor de arquivos e bytes distintos, o que é esperado pois seu domínio é restrito apenas aos objetos armazenados no servidor.

TABELA III
CARACTERIZAÇÃO DOS ARQUIVOS.

Carga	Arquivos	Bytes	Média	Mediana
BL	50.13	63.39	15929	2752
BR	5.21	2.07	22187	1519
U	44.97	61.77	16319	2748
NASA	0.332	0.798	46463	6297
POP98	36.06	46.33	10755	3563
Portugal	46.49	62.26	12052	3389
POP99-hug	40.85	52.41	12743	3284
POP99-zez	60.59	73.61	18339	3512
NLANR-uc	66.59	69.57	14354	3660
NLANR-bo1	66.77	67.67	14801	3777

As distribuições de probabilidade acumulada dos tamanhos para os conjuntos de arquivos têm um perfil ainda mais similar entre si do que as curvas dos conjuntos de requisições

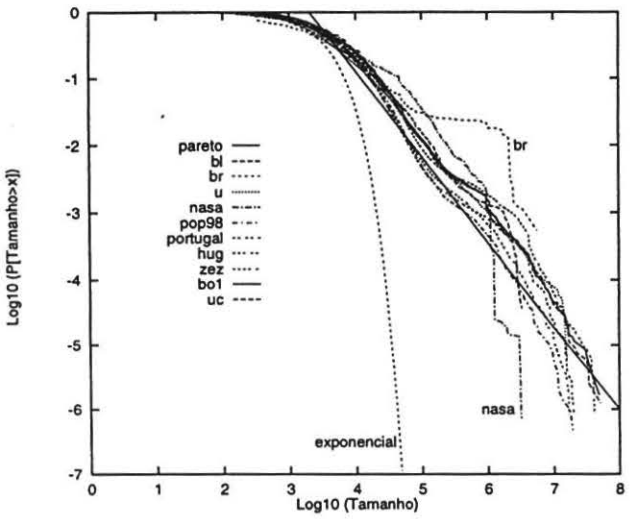

Fig. 2. Cauda da distribuição de probabilidade dos tamanhos do conjunto de requisições para todas as cargas e para as funções exponencial e Pareto.

(figura 1). Os valores estimados para α estão entre 0.89 e 1.08. Comparando os dois conjuntos, temos que o conjunto de arquivos apresenta maior variabilidade (menor α) do que o respectivo conjunto de requisições para a maioria das cargas.

A análise dos tamanhos do conjunto de arquivos e do conjunto de requisições mostra que as cargas apresentam indubitavelmente uma grande variabilidade nos seus tamanhos e que o conjunto de arquivos, que forma a carga de armazenamento do cache, exibe ainda maior variabilidade do que o conjunto de requisições.

IV. Caracterização do Padrão de Acesso

Padrão de acesso é um conjunto específico de características de uma seqüência de requisições que define a localidade e o limite de desempenho que um sistema de caches pode alcançar quando submetido à referida carga. Esta seção examina o padrão de acesso presente nas cargas apresentando diversas medidas para a concentração de referências. A próxima seção examina a localidade temporal.

A. Valores Máximos para HR e BHR

As frações de arquivos e bytes distintos, apresentadas na tabela III são uma medida da quantidade de documentos e bytes novos que ocorre em cada seqüência. A partir destes dados podemos calcular os valores máximos de HR e BHR que um sistema de cache pode alcançar quando submetido a estas cargas. HR_{max} e BHR_{max} são a fração de requisições repetidas e a fração de bytes que compõem as repetições. São, portanto, o complemento de 100% das frações da tabela III. Estes valores poderão obtidos se o cache tem tamanho "infinito", isto é, tamanho suficiente para armazenar todos os documentos requisitados. Neste caso, nenhuma retirada será necessária. Isto elimina a influência das políticas de reposição no desempenho do cache, que fica limitado apenas pela fração de requisições novas feitas pelos clientes ao cache.

As frações HR_{max} e BHR_{max} dão uma medida da localidade para o período de tempo que compreende toda a carga. Os valores encontrados para HR_{max}, entre 33% e 64%, e BHR_{max}, entre 26% e 54%, são comuns para caches de rede. Valores maiores, como os das cargas BR e NASA, são encontrados em caches de servidores. Para todas as cargas exceto BR, os valores de HR_{max} são maiores do que os de BHR_{max}, o que indica que há mais *hits* para objetos pequenos. Os valores reais de HR e BHR são função do tamanho do cache e da política de gerenciamento do espaço.

B. Impacto dos Acessos Únicos

A tabela IV apresenta a fração dos arquivos com um único acesso e a concentração de requisições feitas aos arquivos mais acessados. A distribuição dos arquivos por número de acessos apresenta valores bem parecidos para todas as cargas, exceto NASA e BR. Entre 65% e 88% dos arquivos são requisitados uma única vez. NASA tem 19% dos arquivos com uma única requisição enquanto BR tem 42%. Isto significa que entre 25% e 59% (exceto NASA e BR) das requisições são para objetos que não foram requisitados anteriormente e não serão requisitados novamente.

Seja *hit-docs* o conjunto de documentos com dois ou mais acessos e *hit-bytes* o total de bytes destes documentos. A tabela IV mostra a fração de *hit-docs* (coluna *docs*) e de *hit-bytes*, ou seja, a fração dos bytes únicos que pertence aos *hit-docs*, presente em cada carga. A coluna *hit-req* mostra a porcentagem das requisições que são para os *hit-docs*. Por exemplo, para a carga POP98, somente 30% dos arquivos distintos são requisitados mais de uma vez. Estes 30% dos arquivos ocupam 31% dos bytes únicos e respondem por 75% das requisições. Para a carga NASA, 98% dos acessos são para 58% dos arquivos que compreendem 84% dos bytes únicos. Finalmente, para as cargas do NLANR, 41% das requisições são para 12% dos arquivos. Dada a grande quantidade de documentos com apenas uma requisição, seria interessante utilizar uma política de substituição que discriminasse estes documentos.

V. Localidade de Referência Temporal

Localidade temporal refere-se à noção de que o mesmo documento é requisitado sucessivas vezes dentro de curtos intervalos de tempo. Uma medida simples e interessante para localidade é o gráfico de localidade. Neste gráfico, cada URL recebe uma identificação distinta, de acordo com a ordem em que aparece na carga. As referências também são numeradas e indicam o tempo decorrido em termos do número

TABELA IV
FRAÇÃO DOS ARQUIVOS COM UM ÚNICO ACESSO (FR1) E CONCENTRAÇÃO DE REFERÊNCIAS.

		HITS		
Carga	FR1	docs	bytes	req
BL	0.75	0.25	0.18	0.62
BR	0.42	0.58	0.84	0.98
U	0.76	0.24	0.21	0.66
NASA	0.19	0.81	0.91	0.99
POP98	0.70	0.30	0.31	0.75
Portugal	0.77	0.23	0.38	0.64
POP99-hug	0.65	0.35	0.32	0.73
POP99-zez	0.79	0.21	0.18	0.52
NLANR-uc	0.88	0.12	0.12	0.41
NLANR-bo1	0.88	0.12	0.13	0.41

de requisições feitas ao cache. O número da referência e o número da URL correspondente são plotados no gráfico. Cada URL aparece apenas a primeira vez em que é referenciada.

A figura 3 apresenta este gráfico para todas as cargas. Quanto maior a localidade, mais perto do eixo x estarão as linhas do gráfico, indicando que referências novas ocorrem infrequentemente. Por outro lado, quanto mais frequentes são as novas referências, mais inclinada é a curva. No pior caso, os dados seguem a diagonal, o que significa que cada requisição é para uma URL distinta. O valor para a localidade é a inclinação, dada por 1 - (número de documentos únicos / número de requisições).

As cargas com maior localidade são as dos servidores NASA e BR. Dentre as cargas de cache de rede, U, POP98 e POP99-hug são as que apresentam maior localidade, enquanto POP99-zez é a que apresenta menor localidade. Estes resultados são consistentes com os valores de HR_{max}.

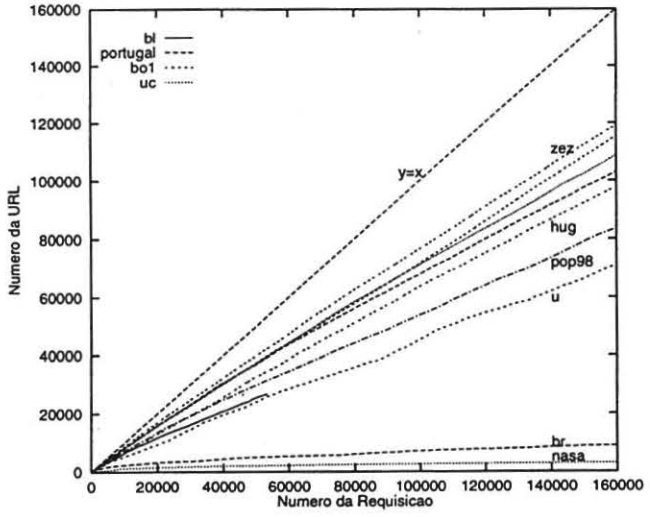

Fig. 3. Evidência de localidade temporal para todas as cargas.

A efetividade dos esquemas de cache está na presença da localidade temporal nas seqüências de requisições da WWW e no uso de políticas apropriadas de gerenciamento. Como mostrado nesta seção, o grau de localidade observado nos caches da WWW pode ser tão bom quanto o encontrado em caches tradicionais (acima de 90% para as cargas BR e NASA) ou não. A localidade existente nas cargas representa, ao mesmo tempo, o potencial que os sistemas de cache devem procurar explorar e o limite de desempenho máximo para o cache.

VI. PADRÃO DE ACESSO EM CLASSES DE REQUISIÇÕES

A caracterização do padrão de acesso para a carga completa fornece uma idéia ampla da localidade presente na carga. No entanto, a grande variabilidade dos tamanhos nos impede de avaliar a influência de conjuntos de requisições de tamanho similar.

Para estudar a distribuição relativa dos bytes, dos arquivos e das requisições classificamos as requisições de cada carga de acordo com seus tamanhos. As classes foram definidas da seguinte forma:

classe 1	tamanho $< 10^3$ bytes
classe 2	$10^3 \leq$ tamanho $< 10^4$ bytes
classe 3	$10^4 \leq$ tamanho $< 10^5$ bytes
classe 4	$10^5 \leq$ tamanho $< 10^6$ bytes
classe 5	tamanho $> 10^6$ bytes

Foram caracterizados os conjuntos de requisições e de arquivos por classes de tamanho para todas as cargas. Os dados mostram que há uma concentração de referências e bytes:

- 80% das requisições são para objetos pequenos, das classes 1 e 2;
- 98% das requisições são para objetos das classes 1, 2 e 3;
- entre 75% e 85% dos bytes requisitados são de objetos das classes 3, 4 e 5; estas classes compõem a cauda da distribuição.
- a classe 5, com cerca de apenas 0.2% das requisições, engloba de 16% (POP98) a 42% dos bytes requisitados (POP99-zez). A classe 5 de BR compreende 2% das requisições e 85% dos bytes requisitados.

A fração do espaço do cache "infinito" necessária para armazenar cada classe foi calculada. Cerca de 0.7% do tamanho do cache é o espaço suficiente para armazenar todos os documentos da classe 1, que corresponde a um número de arquivos entre 13% e 38% do total de arquivos de cada carga. Os arquivos da classe 1 respondem de 15% a 40% de todas as requisições da carga.

Entre 40% e 63% dos arquivos únicos pertence à classe 2. Para armazená-los seria necessário um espaço equivalente a valores entre 3% e 21% do tamanho do cache "infinito". Este espaço responderia por um número enorme de requisições, entre 42% e 62%. As duas últimas classes, embora armazenem menos de 2% dos arquivos, ocupam a maior parte do

cache, entre 32% e 74%. No entanto, estes arquivos respondem pela maioria de bytes requisitados.

Este enorme diferença na concentração de referências e bytes nas diversas classes sugere que a localidade seja diferente nas diversas classes e que a carga completa tenha uma localidade que seja um compromisso das localidades das classes.

VII. Considerações para Projeto de Políticas de Substituição para Caches WWW

Os resultados da caracterização mostram que há uma localidade temporal a ser explorada, embora esta não seja tão grande quanto a dos caches tradicionais (90% dos acessos vão para 10% dos objetos).

A grande variabilidade nos tamanhos dos arquivos tem duas consequências importantes. A primeira é que ela torna contraditória a otimização das métricas HR e BHR. Através da caracterização das classes vimos que 80% das requisições são para objetos das classes 1 e 2 e 20% para as demais classes. Já a concentração de bytes é inversa: apenas cerca de 20% dos bytes requisitados estão nas classes 1 e 2. Há dois casos comuns cuja otimização é contraditória. Para otimizar HR, por exemplo, basta armazenar os arquivos menores, que são os mais populares. Além disto, o armazenamento preferencial de arquivos pequenos permite manter um número maior de arquivos no cache. Os arquivos pequenos são responsáveis por um número grande de *hits* mas constituem uma fração pequena dos bytes servidos. Por outro lado, os arquivos maiores são responsáveis por um número pequeno de *hits* que constituem uma grande fração dos bytes requisitados. O aumento do BHR pode levar a uma diminuição do HR porque implica no armazenamento de documentos maiores, que podem ocupar o lugar de vários documentos pequenos, porém mais populares.

A segunda consequência da variabilidade nos tamanhos dos objetos é a ocorrência de uma alteração significativa no estado do cache (*cache disruption*), que acontece quando a chegada de um documento grande implica na retirada de centenas ou milhares de documentos menores. Esta alteração ocorre em caches gerenciados por políticas que classificam os objetos armazenados no cache utilizando uma regra única de comparação para todos os objetos. A política LRU é um exemplo.

A caracterização por classes baseadas em tamanho mostrou enormes diferenças entre as classes em termos de número de requisições, número de bytes requisitados e concentração e localidade de referências. Estas diferenças nos levam a crer que o estabelecimento de uma única regra não é suficiente para (*i*) alcançar objetivos contraditórios e (*ii*) tratar a diversidade e a variabilidade, observados na caracterização das cargas, tendo em vista a preservação da localidade.

A variabilidade nos tamanhos das requisições da WWW tem influência significativa no desempenho de sistemas de caches destes objetos. Uma alternativa para as políticas atuais, que leva em consideração os problemas gerados pela variabilidade nos tamanhos dos arquivos, é o particionamento do cache em função do tamanho dos objetos. Esta é uma estratégia que deve ser pesquisada. O particionamento permite explorar características específicas de cada classe e permite um maior controle sobre o desempenho do cache em relação ao cache não particionado. Por exemplo, a análise sugere que reservando frações muito pequenas do cache para classes de arquivos pequenos podemos garantir que a maioria das requisições será atendida com frações elevadas de HR e BHR, com consequente aumento do HR e BHR global.

VIII. Conclusões

As características da carga dos caches da WWW e os objetivos destes caches tornam estes sistemas bastante diferentes dos sistemas tradicionais de cache. O projeto de políticas de reposição para caches da WWW deve levar em consideração os aspectos de larga escala e grande variabilidade nos tamanhos dos arquivos. Uma direção para esta consideração é o particionamento do cache em função do tamanho dos objetos, que será considerado em nossos trabalhos futuros.

Agradecimento

Os autores agradecem aos administradores dos caches cujas cargas foram utilizadas neste trabalho por disponibilizarem seus registros de acesso.

Referências

[ESO 98] Esoterica. "http://www.esoterica.pt", 1998.

[MAW 96] Martin F. Arlitt and Carey L. Williamson. Web Server Workload Characterization: The Search for Invariants. **Proceedings of the 1996 ACM Sigmetrics Conference**, pp 126-137, 1996.

[MCT 99] Mark E. Crovella and Murad S. Taqqu. Estimating the Heavy Tail Index from Scaling Properties. **Methodology and Computing in Applied Probability**, V. 1, N. 1, 1999.

[MTA 98] Mark E. Crovella, Murad S. Taqqu and Azer Bestravos. **A Practical Guide to Heavy Tails**. *in Heavy-Tailed Probability Distributions in the World Wide Web*, chapter 1, pp 3-26, Chapman & Hall, New York, 1998.

[NLA 99] NLANR. "http://ircache.nlanr.net", 1999.

[PBM 98] Paul Barford and Mark E. Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. **Proceedings of Performance '98/ACM SIGMETRICS '98**, pp. 151-160, 1998.

[PJH 96] D. A. Patterson and J. L. Hennessy. **Computer Arquitecture: A Quantitative Approach, 2nd Edition**. Morgan Kaufmann Publishers, California, 1996.

[POP 98] POP-MG. Ponto de Presença da RNP em Minas Gerais. "http://www.po-mg.rnp.br/index.html", 1998.

[VTE 95] Virginia Tech. Traces from Virginia Polytechnic Institute. "http://www.cs.vtech.edu", 1995.