Proposta de Dissertação

Análise do Impacto do uso de *proxies* em redes *Peer-to-Peer*

Tiago Alves Macambira *
Orientador: Dorgival Olavo Guedes Neto
{tmacam,dorgival,}@dcc..ufmg.br

30 de janeiro de 2003

1 Introdução

Existe uma constante pesquisa no meio acadêmico e comercial por formas de economizar os recursos de rede de universidades, instituições, ISPs[1], órgãos governamentais, etc. Tradicionalmente, o foco desta pesquisa é direcionado para a aplicação e/ou formas de acesso tidas como responsáveis pela maior parte do tráfego gerado e consumido na internet no momento em questão.

Até meados da década de 90, antes do surgimento e da *Web*, grande parte das pesquisas com este foco não se concentravam nas aplicações, mas nos próprios mecanismos de transporte e roteamento da internet[4, 12]. Naquela época *pré-web*, a maior parte do tráfego era devido aos protocolos SMTP e FTP que, em conjunto, eram responsáveis por aproximadamente 80% de tudo que transitava no *backbone* da NSFNet [13, 9].

O surgimento da Web, no entanto, modificou esse quadro profundamente. Enquanto em 1995 calculava-se que 21% de todo o tráfego na internet era devido a tráfego HTTP, em 1997 esse valor já oscilava entre 60% e 80% [18].

Essa mudança nos padrões do tráfego na rede mundial motivaram trabalhos de análise, caracterização, controle e redução desse tráfego [6, 5, 3, 1]. Neste cenário, o uso de *proxy caching* aparece como uma ferramenta importante na diminuição da carga *Web* na rede.

O aparecimento de aplicações *peer-to-peer* (P2P) de compartilhamento de arquivos acarretou uma nova mudança nos padrões de tráfego da internet. Em 2000, 23% do tráfego de saída da Universidade de Wisconsin era tráfego P2P, ao passo que o tráfego web representava então apenas 20% do total [11]. De maneira semelhante ao que aconteceu com a Web, o tráfego P2P apenas aumentou daquela época para os dias atuais, já chegando a ocupar 45% de todo o tráfego daquela universidade. Comportamentos similares são relatados na literatura e acredita-se que ele seja um reflexo do que esteja acontecendo em toda a internet.[8].

Assim sendo, a maior parte do tráfego da internet atual é devia a aplicações P2P de troca de arquivos. Desta forma, similarmente ao que foi feito quando da explosão do tráfego Web, faz-se necessário analisar, caracterizar, achar formas de controlar e diminuir esse tráfego. Não somente, analisar o impacto e a eficiência destas propostas tanto na dinâmica como no tráfego P2P. É exatamente nestes últimos tópicos que pretendemos focar o nosso trabalho.

O restante deste documento é organizado da seguinte forma. A seção 2 delimita o escopo desse trabalho, apresentando explicitamente o objeto do nosso estudo. A seção 3 apresenta trabalhos relacionados. Na seqüência, as seções 4, 5 e 6 apresentam os objetivos traçados para a realização de nosso trabalho, a forma como pre-

*Menção à bolsa do CNPq

[1] *Internet Service Providers* - Provedores de Acesso à Internet

tendemos realizá-los e um cronograma de como os faremos. Finalmente, na seção 7 apresentaremos a bibliografia referenciada.

2 Definição do Problema

As características do tráfego *Web* já são conhecidas e documentadas na literatura. O uso de *proxy caches* são sabidamente a forma mais eficiente de promover a conservação de recursos de rede devido a esse tráfego. Tais ferramentas melhoram a percepção que o usuário tem da rede, na medida que mascaram eventuais perdas de conectividade, diminuem a latência percebida pelo usuário e reduzem o consumo de largura de banda da instituição.[3, 5]

Já existem propostas na literatura para o uso de soluções como *proxy caches* pra sistemas P2P de troca de arquivo [10]. A esperança é que, dadas as características do tráfego P2P, existam formas de conseguir a conservação da largura de banda usada por essas aplicações, da mesma forma a conseguida no tráfego *web* através de *proxies*.

Contudo, como apontado por Gummadi[8], tráfego *Web* é diferente do tráfego P2P. Isto faz com que todas as soluções propostas para *web-proxies* tenham que ser repensadas para o cenário particular de aplicações *peer-to-peer*.

Apesar de toda essa motivação para o uso de *proxies* para sistemas P2P, não existem trabalhos nem estudos divulgados na literatura que:

- comprovem empiricamente os resultados obtidos com o uso de tal abordagem;
- analisem o impacto que essa solução provocaria na dinâmica das aplicações P2P;
- caracterizem como tal solução modifica o tráfego P2P de uma instituição pra o mundo exterior e
- especifiquem como melhor modelar e construir tais *proxies* de formar a maximizar os ganhos possíveis.

Há ainda o fato de que caracterizações existentes da carga de sistemas de P2P de troca de arquivo são por vezes contraditórias e ainda em reduzido número. Isto nos motiva a:

- validar as pesquisa já efetuadas, procurando por comportamentos similares em outros cenários ou em cenários mais abrangentes ou
- observar novos comportamentos e métricas que não foram observadas ou que não foram observadas com o rigor necessário.

3 Trabalhos Relacionados

Como já dito, a idéia de *proxies* para sistemas P2P não é nova. Krishnamurthy e Zhang propuseram um sistema que não somente faria *caching* de pesquisas entre vários sistemas distribuídos de troca de arquivo bem como dos dados trocados entre clientes destes sistemas[10]. Nosso foco não é na implementação de um *proxy* nem na proposta de um mecanismo que sirva de ponte entre vários sistemas P2P, mas nas observação da aplicabilidade de um *proxy* pra sistemas P2P, não necessariamente nestes moldes, bem como na na observação de como um *proxy* alteraria a dinâmica de um ou de vários desses sistemas.

Vários artigos recentes relatam métricas de sistemas P2P, focando nas redes de troca de arquivo mais populares. Como exemplo de alguns destes *papers* temos [7, 11, 15]. Esses artigos, no entanto, obtêm seus dado de uma forma intrusiva, colocando um nó nessa rede que interage com os outros nós, inferindo e obtendo métricas. Como apontado por Sen[16], esse tipo de estudo é limitado pela próprio consumo de banda e recursos que gera, não sendo capaz de analisar redes com um grande número de nós eficientemente. Outros, como o do próprio Sen e um artigo de Gummadi[8], utilizam-se de monitoramento passivo de tráfego para obter seus dados e estatísticas. Essa abordagem é menos intrusiva, mais escalável e mais confiável, haja vista que toda a atividade dos nós de uma rede pode ser detectada e monitorada.

Nosso trabalho diferencia-se de trabalhos recentes sobre redes P2P pois não propomos uma nova rede P2P, não propomos formas de melhorar buscas nestas redes, nem uma formas de torná-las mais escaláveis. [2, 17, 14]. Não abordando o uso de técnicas de roteamento com conhecimento de localidade (location-aware routing). Tampouco propomos a modificação dos protocolos existentes para que esses tirem proveito da localidade de referência existentes dentro de uma **instituição** a fim de evitar alguns meandros legais que envolvem o uso de caches e o armazenamento (temporário) de arquivos ilegais.

4 Objetivos

Planejamos implementar uma arquitetura de monitoração de tráfego P2P não-intrusiva e colocá-la em um ambiente real. Após um período de monitoração no qual observaremos a escalabilidade de nossa arquitetura, planejamos estendê-la, acrescentando a ela a capacidade de recuper e armazenar dados e arquivos trocados no ambiente em questão. Nossa idéia é colocar toda essa arquitetura nas fronteiras da rede de uma instituição parceira, onde poderimos efetivamente monitorar e realizar *cache* de todo o tráfego P2P desta instituição com o mundo. Finalmente, seguindo o a mesma estratégia não-intrusiva, colocaríamos um nodo na rede que interagiria de forma não intrusiva e não-disruptiva com a rede P2P, oferecendo seu cache para os nós de dentro da instituição. . Observaremos como a existência desse nodo-cache afeta a dinâmica da rede P2P desse ambiente sob vários aspectos e métricas:

- Média de sucesso de downloads;
- tempo médio de conclusão de downloads;
- velocidade média de downloads e uploads;
- variação do tráfego externo de entrada e saída;
- localidade de referência dos arquivos (e seus fragmentos);
- et cetera

4.1 Resultados esperados

Com essas ações e objetivos pretendemos:

1. ajudar a suprir a inexistência de estudos que comprovem empiricamente o impacto que a existência de um cache causaria na dinâmicas de redes P2P;
2. comprovar o potencial que existe para *caching* de dados em redes P2P;
3. obter caracterizações mais precisas do tráfego P2P;
4. obter mais informações que possam nortear o desenvolvimento de novas redes P2P e de soluções para controle de tráfego gerado por esse elas;
5. Publicação dos resultados obtidos.

5 Metodologia

5.1 Atividades

As seguintes atividades são previstas para o desenvolvimento do nosso trabalho:

Levantamento Bibliográfico fase onde pesquisas e estudos da bibliografia existente relacionados a área de sistemas P2P, web proxies e proxies no geral, monitoração de tráfego em alta-velocidade, et cetera, serão realizados

Implementação da Infra-estrutura de Monitoração Pesquisa e análise das ferramentas existentes para a sua elaboração. Implementação e implantação desta infra-estrutura em um cenário real

Coleta de Dados Obtenção e análise dos dados coletados através da infra-estrutuda de Monitoração

Implementação do Proxy (nodo-cache) Pesquisa e análise das ferramenta disponíveis para a elaboração do proxy e ajuste deste mediante dados obtidos na etapa de coleta de dados

Coleta de Dados 2 Obtenção, análise e comparação dos dados coletados através da Infraestrutura de Monitoração na presença do Proxy

Análise Observação, averigüação dos dados obtidos na 2a. coleta. Verificação da aplicabilidade e do sucesso da abordagem adotada para implementação do proxy.

Refinamento Re-estruturação das partes do projeto para obtenção de dados e resultados mais precisos e/ou satisfatórios

Escrita da dissertação

6 Cronograma

As atividades propostas em 5.1 seguem o cronograma descrito na a Tabela 1.

Referências

[1] Martin Arlitt and Tai Jin. A workload characterization study of the 1998 world cup web sit. In *Network*. IEEE, May 2000.

[2] Hari Balakrishnan, M. Frans Kaashoek, David Karger, Robert Morris, and Ion Stoica. Looking up data in p2p systems. In *Communications of the ACM*, Feb 2003.

[3] G. Barish and K. Obraczka. World wide web caching: Trends and techniques. In *IEEE Communications Magazine - Internet Technology Series*, May 2000.

[4] Lawrence S. Brakmo, Sean W. O'Malley, and Larry L. Peterson. TCP vegas: New techniques for congestion detection and avoidance. In *SIGCOMM*, pages 24–35, 1994.

[5] R. Caceres, F. Douglis, A. Feldmann, G. Glass, and M. Rabinovich. Web proxy caching: the devil is in the details, 1998.

[6] Anawat Chankhunthod, Peter B. Danzig, Chuck Neerdaels, Michael F. Schwartz, and Kurt J. Worrell. A hierarchical internet object cache. In *USENIX Annual Technical Conference*, pages 153–164, 1996.

[7] J. Chu, K. Labonte, and B. Levine. Availability and locality measurements of peer-to-peer file systems, 2002. abordagem intrusiva.

[8] Krishna P. Gummadi, Richard J. Dunn, Stefan Saroiu, Steven D. Gribble, Henry M. Levy, and John Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 314–329. ACM Press, 2003.

[9] Steven A. Heimlich. Traffic characterization of the nsfnet national backbone. In *Proceedings of the 1990 ACM SIGMETRICS conference on Measurement and modeling of computer systems*, pages 257–258. ACM Press, 1990.

[10] Balachander Krishnamurthy and Yin Zhang. P4P: Proxies for p2p systems, 2002.

[11] Evangelos P. Markatos. Tracing a large-scale peer to peer system: an hour in the life of gnutella. In *2nd IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2002. Tem um TechnicalReport desse paper.

[12] J. Moy. RFC 1247: OSPF version 2, July 1991.

[13] Ramon (ramon@tenet.berkeley.edu). Caceres. Measurements of wide area internet traffic. Technical report, Berkeley, 1989.

Tarefa	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Levantamento	X	X	X									
Elaboração	X	X										
Coleta 1		X	X									
Analise 1		X	X	X								
Proxy			X	X	X	X						
Coleta 2			X	X	X	X						
Analise 2								X	X	X		
Refinamento					X	X	X	X	X	X	X	
Escrita		X	X	X	X	X	X	X	X	X	X	X

Tabela 1: Cronograma de atividades - por mês

[14] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Schenker. A scalable content-addressable network. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 161–172. ACM Press, 2001.

[15] Stefan Saroiu, P. Krishna Gummadi, and Steven D. Gribble. A measurement study of peer-to-peer file sharing systems. In *Proceedings of Multimedia Computing and Networking 2002 (MMCN '02)*, San Jose, CA, USA, January 2002. gnutella/napster, abordagem intrusiva.

[16] Subhabrata Sen and Jia Wong. Analyzing peer-to-peer traffic across large networks. In *Second Annual ACM Internet Measurement Workshop*, November 2002.

[17] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 149–160. ACM Press, 2001.

[18] K. Thompson, G.J. Miller, and R. Wilder. Wide-area internet traffic patterns and characteristics. In *IEEE Network*, volume 11, pages 20–23, Nov/Dec 1997.