Caracterização de Tráfego de Armazenamento em Nuvem

Rodrigo Costa Duarte[1], Alex Borges Vieira[1], Ana Paula Couto da Silva[1]

[1]Departamento de Ciência da Computação – Instituto de Ciências Exatas
Universidade Federal de Juiz de Fora(UFJF) – Juiz de Fora, MG – Brazil

{rodrigo.duarte, alex.borges, ana.coutosilva}@ufjf.edu.br

Área de pesquisa: Redes de computadores

Ingresso no PGCC: março de 2012

Resumo *Serviços na Internet baseados em nuvem podem ser considerados um dos maiores contribuidores para a geração de tráfego na Internet. Com o uso massivo destes serviços e o impacto da quantidade do fluxo de informação gerado nos canais da rede faz-se necessário a caracterização do mesmo. O principal objetivo é compreender o tipo de tráfego, bem como auxiliar no dimensionamento de novas redes e serviços. Este trabalho visa analisar e caracterizar o tráfego gerado pelo serviço de armazenamento em nuvem em diferentes ambientes acadêmicos.*

Palavras-chave: Internet, Armazenamento em nuvem, Caracterização.

1. Introdução

Nos últimos anos, a utilização de variados serviços computacionais em nuvem tem se tornado cada vez mais comum. Estes serviços oferecem aos usuários finais a possibilidade de adquirir equipamentos e infra-estrutura com custos reduzidos. Entre os diversos serviços oferecidos, existe uma forte demanda pelos modelos de armazenamento de dados em nuvem. Este serviço atende desde usuários com demanda de armazenamento de informações do cotidiano, como fotos e vídeos, até grandes empresas que desejam guardar bancos de dados importantes para posterior manipulação dos mesmos.

Dado o grande volume de tráfego gerado por serviços de armazenamento em nuvem, caracterizar este tipo de serviço é importante para dimensionar novas redes e prover qualidade de serviço aos usuários que contratam estes serviços. Desta forma, este trabalho visa caracterizar e analisar este tipo de tráfego, considerando diferentes perfis de usuários.

2. Caracterização do problema

Com a popularização da Internet, a utilização de serviços que dependem e fazem uso dos recursos de redes de comunicação se tornam cada vez mais evidentes. Um dos serviços de maior crescimento é o armazenamento remoto de informação, através da estrutura de computação em nuvem.

O aumento da adesão a este tipo de serviço gera grande volume de dados na infra-estrutura de redes existente. Em alguns casos, a qualidade de serviço pode diminuir caso um grande volume de dados seja enviado constantemente a estes servidores remotos. Por outro lado, analisar o perfil dos usuários destes serviços pode ser importante para a definição de serviços mais eficientes, no que tange a segurança e melhor utilização dos recursos computacionais e de rede.

O problema a ser tratado neste trabalho é a caracterização do tráfego gerado por serviços de armazenamento em nuvem. Através de coleta de dados de diferentes pontos na rede, como ambientes acadêmicos, será analisada a quantidade de tráfego gerado, os tipos de arquivos que são armazenados remotamente, bem como a presença de redes sociais neste tipo de serviço.

3. Metodologia

O trabalho proposto se baseia na coleta de dados em redes reais, bem como da análise dos mesmos. Desta forma, metodologias deverão ser propostas ou poderão ser utilizadas técnicas encontradas na literatura e que auxiliam na obtenção dos resultados esperados.

Um vasto conjunto de artigos que descrevem metodologias para caracterização é encontrado na literatura. Em [Claffy et al. 1993] foram descrita metodologias de amostragem de tráfego de rede.

3.1 Ambiente dos experimentos

Para fins comparativos foram capturadas várias amostras de tráfego de dados no ambiente da Universidade Federal de Juiz de Fora, entre outubro e novembro de 2012,

possibilitando através da análise das diferentes amostras, a definição entre outros fatores, da melhor localização para o posicionamento da estação de coleta, tamanho e fragmentação dos conjuntos de dados coletados e dimensionamento do equipamento necessário para a coleta.

Comparando-se os resultados das capturas definiu-se a instalação do ponto de coleta no link entre a entrada do roteador Internet e o *Firewall* corporativo da instituição. Este posicionamento permitiu que os dados capturados mantivessem a integridade de fluxo, sem ser afetados por regras de filtragem aplicadas internamente pela instituição.

Ao se analisar o tamanho das amostras obtidas, foi verificado que com os recursos disponíveis, seria praticamente inviável realizar as coletas com algumas ferramentas disponíveis que não permitem a redução, já na fase de coleta, do volume dos dados a serem armazenados. A escolha da ferramenta de coleta a ser utilizada recaiu sobre o aplicativo TSTAT por sua adesividade ao ambiente acadêmico, promovendo a possibilidade de descartar na própria coleta dados não pertinentes ao trabalho e consequentemente preservando a confidencialidade das informações coletadas já em sua fonte.

Verificando o desempenho dos equipamentos nas atividades de coleta, foi definido que a estação coletora terá *hardware* baseado em dois processadores Intel Xeon com 3.40 Ghz, memória RAM de 4 Gb e um *Hard Disk* SCSI de 300 GB. A estação estará ligada à rede através de duas interfaces de rede *Gigabit Ethernet,* sendo uma delas dedicada exclusivamente ao serviço de captura do tráfego. O sistema operacional utilizado é o Linux com *Kernel* versão 3.2.0

3.1.1 Descrição das coletas

Para a coleta de tráfego foi utilizada a ferramenta TSTAT, desenvolvida pelo grupo de telecomunicações em redes da *Politecnico di Torino*, que a partir de bibliotecas de software padrão oferece informações importantes sobre índices de desempenho e dados estatísticos sobre o tráfego da Internet [TSTAT 2008].

A ferramenta analisa os fluxos de entrada e saída de pacotes, correlacionando-os para inferir índices avançados de medições. Como exemplo, dados TCP e segmentos de ACK podem analisados conjuntamente e os status de cada conexão TCP podem ser reconstruídos. Esta bi-direcionalidade da análise de fluxo do TCP permite a derivação de novas estatísticas, como tamanho de janelas de congestionamento, segmentos fora de sequência e segmentos duplicados, que são coletadas distinguindo entre clientes e servidores das conexões e também identificando nós internos e externos em relação ao ponto de captura.

3.1.2 Dados Obtidos e métricas

Os dados obtidos nas coletas fornecem informações sobre os fluxos dados de cada conexão observada, sendo ela TCP ou UDP além de dados gerais do período de coleta. Com a automação provida por parte do TSTAT foi possível alcançar algumas métricas de real interesse para nosso estudo, como:

- *Bitrate* da aplicação TCP
- Endereços IP de origem e destino dos fluxos TCP

- Tempo total de fluxo TCP
- Fluxos TCP interrompidos prematuramente
- Máximo, mínimo, média e desvio-padrão do tempo de ida e volta (RTT) dos fluxos TCP.
- Comprimentos dos fluxos TCP em pacotes transmitidos

Informações como estas permitem caracterizar o tráfego gerado pelas aplicações de armazenamento em nuvem, destacando pontos chaves como desempenho, padrões de utilização e pontos de melhoria possíveis de serem implementados.

4. Caracterização da contribuição

O artigo proposto visa contribuir para a comunidade acadêmica com a caracterização do tráfego de rede gerado por aplicações pessoais de armazenamento em nuvem em ambientes acadêmicos no Brasil. Um adicional pretendido pelo trabalho, até então não explorado, é a conexão entre a utilização destas aplicações e a utilização de redes sociais pelos seus usuários.

5. Estado atual do trabalho

Para o trabalho foram realizadas várias coletas de dados chegando-se ao volume de aproximadamente 3 Terabytes de dados coletados no período de outubro a novembro de 2012. Estes dados correspondem a coletas que variam em parametrizações de filtragem de dados coletados, localização de coletor, duração da amostra entre outros fatores.

Através destas coletas experimentais foram alcançados aos resultados obtidos até o momento, que são as definições de parametrização e ambiente de coleta já citadas no item 3.1.

No atual momento a coleta definitiva foi iniciada, e estão sendo realizadas análises parciais dos dados, para as observações iniciais sobre o tráfego estudado. Para a próxima fase a ser executada pretende-se concluir um período de amostragem e trabalhar de forma definitiva para gerar os resultados estatísticos sobre o conjunto de dados desta amostra.

6. Comparação com trabalhos relacionados

Nos últimos 3 anos alguns trabalhos com enfoque em armazenamento em nuvem foram publicados. [Hu et al. 2012] focou sua pesquisa em comparações entre diferenças arquiteturais entre os provedores do serviço, verificar a performance e avaliar as garantias, privacidade e segurança oferecidas. [Drago et al. 2012] realizou um estudo direcionado ao sistema Dropbox, considerado o sistema de *cloud storage* mais amplamente utilizado no período de sua pesquisa, caracterizando a carga de trabalho em diferentes ambientes e seu reflexo no tráfego de rede e mostrou possíveis estrangulamentos de desempenhos causados pela arquitetura do sistema e pelo protocolo de armazenamento.

Os resultados esperados para este trabalho diferem em relação aos demais encontrados na literatura no escopo e na profundidade da análise dos dados bem como no perfil geográfico dos estudos: o interesse é analisar os dados capturados em instituições localizadas geograficamente no Brasil, região para a qual não existem

estudos com este foco realizados até o momento, além de trazer resultados estatísticos precisos sobre o objeto estudado, observar suas relações com outras aplicações típicas da Internet, como as redes sociais.

Referências

Hu, W., Yang, T. and Matthews, J. (2010) “The good, the bad and the ugly of consumer cloud storage”. *SIGOPS Oper. Syst. Rev.* 44, 3 (August 2010), 110-115. DOI=10.1145/1842733.1842751 http://doi.acm.org/10.1145/1842733.1842751

Claffy, K., Polyzos, G., and Braun, H. (1993). “Application of sampling methodologies to network traffic characterization”. In *Conference proceedings on Communications architectures, protocols and applications* (SIGCOMM '93). ACM, New York, NY, USA, 194-203. http://doi.acm.org/10.1145/166237.166256

Zang, Q., Cheng, L. and Boutaba, R. (2010) “Cloud Computing: State-of-the-art and Research Challenges”. Journal of Internet Services and Applications, 1:7-18, 2010.

Ranum, M., Landfield, K., Stolarchuk, M., Sienkiewicz, M., Lambeth, A., Wall, E. “Implementing a generalized tool for network monitoring”, Information Security Technical Report, Volume 3, Issue 4, 1998, Pages 53-64, ISSN 1363-4127, 10.1016/S1363-4127(98)80038-1. http://dx.doi.org/10.1016/S1363-4127(98)80038-1

Drago, I., Mellia, M., Munafò, M., Sperotto, A., Sadre, R. and Pras, A. (2012) “Inside Dropbox: Understanding Personal Cloud Storage Services”, IMC’12, November 14–16, 2012, Boston, Massachusetts, USA.

Zink, M., Suh, K., Gu, Y. and Kurose, J. (2008) "Watch global, cache local: YouTube network traffic at a campus network: measurements and implications", Proc. SPIE 6818, Multimedia Computing and Networking 2008, 681805 (January 28, 2008); doi:10.1117/12.774903;http://dx.doi.org/10.1117/12.774903

Veloso, E., Almeida, V., Meira, W., Bestavros, A. and Jin, S. (2002). “A hierarchical characterization of a live streaming media workload”. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement* (IMW '02). ACM, New York, NY, USA, 117-130. DOI=10.1145/637201.637220 http://doi.acm.org/10.1145/637201.637220

Finamore, A., Mellia, M., Munafò, M., Rossi, D. (2007). “10-year Experience of Internet Traffic Monitoring with Tstat ”. , ICT Call 1 FP7-ICT-2007-1. http://www.telematica.polito.it/oldsite/mellia/papers/tstat-IEEENET.pdf

Ziviani, A. and Duarte, O. (2011). “Metrologia na Internet”. Minicursos do XXIII Simpósio Brasileiro de Redes de Computadores, SBRC, pp. 285–329, 2005.

Tstat - TCP *STatistic and Analysis Tool. (2008)* Disponível em: http:// http://tstat.polito.it/index.shtml . Último acesso: 03/12/2012.