BUBA: Uma Ferramenta para Análise do Comportamento de Usuários de Internet Banda Larga

Pedro H. Calais Guerra , Elisa Tuler de Albergaria , Leonardo C. Dutra Rocha , Humberto Marques Neto , Jussara M. Almeida , Wagner Meira Jr. , Virgílio Almeida

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - Pampulha - CEP 31270-010
Belo Horizonte - Minas Gerais - Brasil

{pcalais,elisa,lcrocha,hmarques,jussara,meira,virgilio}@dcc.ufmg.br

Abstract. *Understanding the nature and characteristics of Internet broadband user behavior is an important step to improve the quality of the service offered to these costumers. Characterizing the behavior of these users may lead to a better comprehension of the interaction between the users and the Internet Service Providers (ISPs). It also may help in designing systems with better QoS metrics, such as performance, availability, security and cost. In this paper we present* ***BUBA*** *(*Broadband User Behavior Analysis*), a tool designed to create an environment to understand, analyze, and model the traffic data and the request pattern of the users of the ISP. This information allow ISPs to perform, in a more systematic and safer way, the capacity planning of the infrastructure needed to offer this kind of service.*

Resumo. *Compreender a natureza e as características do comportamento de usuários de Internet de banda larga é um passo importante para melhorar a qualidade do serviço oferecido aos usuários desses ambientes. A caracterização do comportamento desses usuários pode conduzir a uma melhor compreensão da interação entre usuários e fornecedores desse serviço. Pode também ajudar no projeto dos sistemas com melhores métricas de QoS, tais como o desempenho, a disponibilidade, a segurança e o custo. Neste artigo apresentamos a ferramenta* ***BUBA*** *(*Broadband User Behavior Analysis*), cujo o propósito é criar um ambiente adequado para entender, analisar e modelar os dados de tráfego de um provedor de banda larga, o que criaria condições para estes planejarem, de uma forma mais sistemática e segura, a capacidade da infra-estrutura necessária à prestação desse tipo de serviço.*

1. Introdução

A crescente difusão de conexões de banda larga, particularmente de conexões utilizando tecnologias de TV a cabo e ADSL, tem criado novas possibilidades de uso dos recursos da Internet por parte dos usuários de pequenas instituições e usuários domésticos. A disponibilidade contínua de conexão aliada ao aumento da qualidade de serviço de acesso conduz o usuário ao uso mais intenso dos diferentes serviços disponíveis na Internet atual, em particular, aplicações de comércio eletrônico, *Internet banking* e sistemas *Peer-to-Peer* (P2P) para compartilhamento de recursos, principalmente, arquivos de vídeo e áudio.

Em outras palavras, a maior disponibilidade provida pela conexão de banda larga leva a um comportamento típico de usuário bastante complexo e potencialmente diferente

daquele de um usuário de acesso discado. Na verdade, um estudo recente da Pew Internet & American Life [Pew Internet, 2004] mostrou que, se comparado com usuários de acesso discado, usuários de banda larga se envolvem com mais atividades diferentes, têm uma tendência maior de criar e gerenciar conteúdo on-line e buscam informações com muito mais frequência.

Esse comportamento mais complexo merece uma atenção especial dos provedores de acesso (ISPs). Em particular, pode-se citar o seu papel central no planejamento de capacidade da infra-estrutura necessária para atendimento da demanda de seus usuários e na proposição de políticas de gerenciamento e tarifação personalizadas, com boa relação custo-benefício tanto para o usuário quanto para o provedor.

Este artigo apresenta a ferramenta BUBA - *Broadband User Behavior Analysis*, projetada para auxiliar o provedor na análise do comportamento de seus usuários e que visa prover subsídios para novas políticas de planejamento e gerenciamento de recursos e serviços. O BUBA utiliza como dados de entrada os logs do servidor de autenticação, como o servidor RADIUS [Rigney, 2000] e os logs do tráfego gerado pelos clientes, coletados pela ferramenta Netflow [Netflow, 2005]. Além disso, é utilizada uma tabela de categorias de usuários, usualmente classificados junto aos provedores em duas categorias, a saber residencial e comercial. A partir desses dados, o BUBA provê um conjunto de funcionalidades para análise de (1) consumo de recursos (em particular largura de banda), (2) popularidade de diferentes serviços da Internet e (3) padrões comportamentais típicos de diferentes usuários. A ferramenta permite a análise do comportamento de usuários individuais, categorias de usuários (residencial ou comercial) e mesmo a descoberta de sub-categorias de usuários que se diferenciam do restante por apresentar um comportamento peculiar.

O restante deste artigo está organizado como a seguir. A seção 2 discute trabalhos relacionados. As principais funcionalidades providas pelo BUBA são discutidas na seção 3. A seção 4 apresenta alguns detalhes de implementação da ferramenta, e a seção 5 apresenta um estudo de caso. Por fim, as conclusões são apresentadas na seção 6.

2. Trabalhos Relacionados

A análise do consumo de recursos de um ISP, tais como, quantidade de conexões simultâneas e a largura de banda, pode ser feita através de ferramentas que utilizam os logs das sessões dos usuários de um ISP, normalmente gerados por servidores de autenticação. Algumas ferramentas são capazes de interpretar diversos formatos de log, onde se destaca o padrão *Livingston* utilizados pelos servidores de autenticação RADIUS [RadRep, 2005, RadiusContext, 2005, ICRadius, 2005]. Contudo, tais ferramentas não permitem a análise dos tipos de serviço requisitados em cada sessão.

Informações acerca dos serviços Internet requisitados podem ser obtidas através da análise da tupla protocolo/porta (como TCP/80 para o HTTP), o que requer a análise dos pacotes enviados e recebidos pelos usuários em suas respectivas sessões. O Netflow captura as tuplas protocolo/porta, além dos IPs de origem e destino dos fluxos de pacotes de um roteador CISCO do ISP. Ferramentas como o *Cflowd*, *FlowScan*, *RRDTool*, *Flow-Tools* e *Cisco Netflow Performance Analysis* [Cflowd, 2005, FlowScan, 2005, RRDtool, 2005, M. Fullmer, 2000, CNPA, 2005] são utilizadas para analisar logs gerados pelo Netflow. Entretanto, como um determinado IP pode ser utilizado por usuários distintos em sessões diferentes, não é possível estabelecer diretamente o relacionamento da sessão de um determinado usuário com os dados obtidos pelo Netflow.

A partir da integração dos logs de um servidor de autenticação, mais especificamente o RADIUS, com logs dos serviços requisitados, como por exemplo os gerados pelo Netflow, o BUBA é capaz de mapear os serviços utilizados em cada sessão de usuário. Essa integração viabiliza a análise tanto do consumo de recursos quanto da popularidade de serviços utilizados pelos usuários do ISP, bem como dos padrões comportamentais típicos destes usuários.

3. Funcionalidades

Nesta seção apresentamos quais são as funcionalidades oferecidas pelo sistema BUBA. Para cada uma delas, o usuário do sistema pode determinar o período do log a ser analisado. Isso permite análises de um dia específico, de um fim-de-semana ou de um mês inteiro, por exemplo. As análises são divididas em três conjuntos:

- **Consumo de Recursos:** neste conjunto é avaliada a utilização de recursos por parte dos usuários, em termos das características gerais de suas sessões;
- **Serviços Acessados:** o objetivo deste conjunto é analisar o uso de serviços da Internet pelos usuários através da avaliação da popularidade dos serviços, em três aspectos diferentes: a presença desses serviços em cada sessão, o número de requisições a esses serviços e os bytes transferidos através de cada um;
- **Comportamento dos Usuários:** por este conjunto é possível avaliar o comportamento de um usuário específico ou de um IP ao longo do tempo, além de analisar os padrões de comportamento dos usuários.

O objetivo dessas funcionalidades é tornar possível uma análise da atividade dos usuários enquanto os mesmos estão conectados à Internet, quantificando e qualificando a carga gerada por eles. Nas seções seguintes serão detalhados cada um dos três conjuntos.

3.1. Consumo de Recursos

Essa parte da ferramenta oferece funcionalidades que permitem averiguar o consumo de recursos por parte dos usuários do ISP no que se refere ao consumo de banda, à frequência de conexões e à duração das mesmas. Com esses dados os provedores podem realizar um melhor planejamento na capacidade da infra-estrutura, melhorando conseqüentemente o atendimento da demanda dos seus usuários. Os seguintes critérios de análise são providos:

- **sessões ativas:** mostra como o número de sessões ativas varia durante o período analisado.
- **início e término de sessões:** mostra o número de sessões iniciadas e finalizadas ao longo do dia.
- **número de sessões por cliente:** mostra a distribuição do número de sessões estabelecidas por um cliente durante o período analisado.
- **duração das sessões:** permite avaliar o tempo durante o qual os usuários do ISP permanecem conectados em cada uma de suas sessões, através de um histograma em que é possível verificar quantas sessões duraram um certo número de horas.
- **bytes transferidos:** mostra um histograma do número de bytes enviados e recebidos nas sessões dos usuários.

3.2. Popularidade de Serviços

Nesse conjunto de funcionalidades, o objetivo é prover um entendimento das atividades que os usuários desempenham em suas sessões, ou seja, que serviços são mais utilizados por eles e qual o consumo de largura de banda esses serviços demandam. Dessa forma a ferramenta disponibiliza quatro critérios de análise:

- **número de serviços por sessão:** mostra um histograma do número de serviços utilizados em cada uma das sessões dos usuários.
- **número de requisições:** mostra a distribuição do número de requisições por serviços.
- **presença de serviços:** mostra os serviços mais comuns entre as sessões dos usuários, considerando apenas o fato do serviço ter sido requisitado ou não em cada sessão, ao invés da intensidade desse uso.
- **tráfego por serviço:** possibilita analisar quais os serviços que transmitem grande volume de dados pela rede, sendo especialmente útil para investigar o impacto causado pelas aplicações de P2P.

3.3. Comportamento dos Usuários

O objetivo desse conjunto de funcionalidades é prover um melhor entendimento acerca do comportamento dos usuários que utilizam a rede do ISP. Dessa forma, temos:

- **histórico do Usuário/IP**: esta funcionalidade mostra quando um determinado usuário se conectou ao ISP, por quanto tempo permaneceu conectado e qual IP ele utilizou em cada sessão, além dos bytes que ele enviou e recebeu em cada uma dessas sessões. Essa consulta pode ser feita pelo *login* do usuário ou por um IP.
- **CBMGs:** é uma forma de representar e modelar o comportamento dos usuários da Internet. Ele consiste em um grafo direcionado onde os vértices são os diversos serviços oferecidos na rede (como HTTP, FTP, POP3 etc) e as arestas representam as probabilidades de um usuário deixar de utilizar um serviço para utilizar outro. O CBMG (*Customer Behavior Model Graph*) permite identificar padrões de comportamento de usuários, o que pode ser muito útil em diversos aspectos, discutidos na seção 5.

4. Implementação

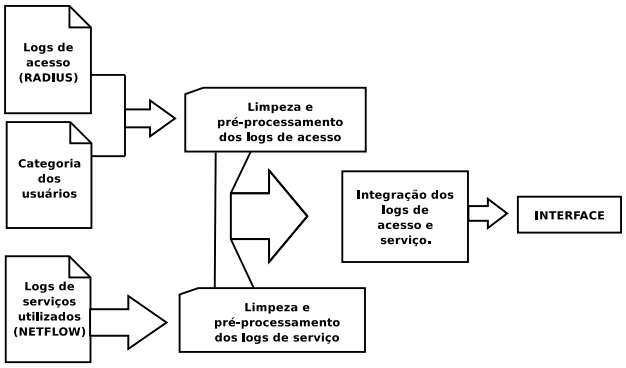


Figura 1. Arquitetura da Implementação do BUBA

A Figura 1 ilustra a arquitetura do sistema desenvolvido. No BUBA são utilizadas três fontes de dados: um log de autenticação, um banco de dados dos usuários e um log com o tráfego de dados. O log de autenticação é compatível com o protocolo RADIUS. Para cada entrada de uma nova sessão de usuário, o arquivo contém as seguintes informações: data/hora do início e fim da sessão, duração, número de bytes enviados e recebidos e o IP dinamicamente atribuído ao usuário. O banco de dados dos usuários consiste em uma tabela que armazena a informação da categoria do usuário, que pode ser, por

exemplo, residencial ou comercial. O terceiro log é coletado utilizando o Netflow. O Netflow divide o tráfego em fluxos e cada fluxo é caracterizado por um *timestamp* que indica quando o mesmo foi registrado no log, o IP de origem e de destino dos dados, a porta de destino utilizada, o protocolo utilizado (TCP ou UDP) e o volume de bytes transferidos.

A partir dos logs acima apresentados, a ferramenta realiza um pré-processamento dos mesmos para extrair as informações necessárias de forma mais simplificada e eficiente. Nessa etapa, o BUBA lê os logs do Netflow e gera um arquivo para cada IP utilizado pelos usuários do provedor. Cada um desses arquivos contém informações sobre todos os pacotes que trafegaram na rede através de um IP. O log do RADIUS, por sua vez, informa qual usuário estava utilizando cada IP em cada momento. Dessa forma, combinando os dois logs é possível extrair todas as informações necessárias para caracterizar a atividade dos usuários em suas sessões.

O BUBA conta com um mecanismo de *cache* que permite que resultados já gerados anteriormente fiquem armazenados e, quando solicitados novamente, dispensem qualquer novo processamento.

Os CBMGs [Menascé and Almeida, 2000] são obtidos através do *k-means* [Hartigan, 1975], um algoritmo de agrupamento. Esse tipo de algoritmo é uma técnica bem conhecida para segmentar um grupo heterogêneo de entidades (no caso, sessões de usuários de Internet) em grupos (ou *clusters*) que possuem um comportamento mais semelhante e homogêneo. No caso, queremos determinar grupos de usuários que acessem os mesmos serviços *Web* e com uma frequência semelhante. Os passos para geração dos CBMGs estão descritos no manual da ferramenta.

5. Estudo de Caso

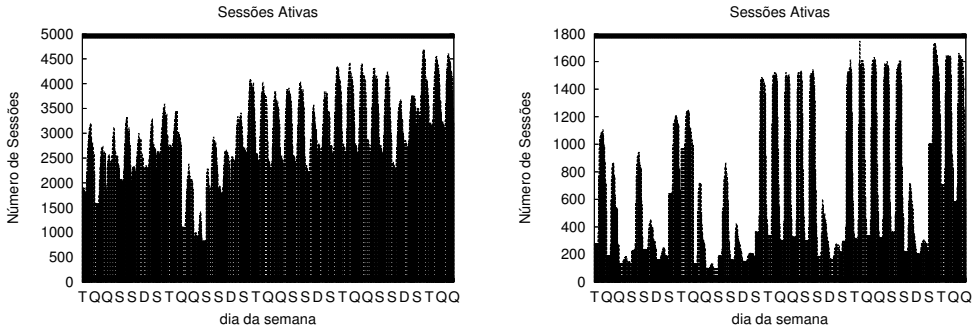
Nesta seção apresentamos um estudo de caso para ilustrar o funcionamento da ferramenta e a aplicabilidade dos resultados fornecidos. Os logs utilizados nesse estudo representam um período de 28 dias de funcionamento de um provedor brasileiro de banda larga. Por questões de privacidade, a identificação dos usuários foi anonimizada, através de uma opção da própria ferramenta.

Através da funcionalidade de Recursos é possível sumarizar os dados relacionados ao consumo de recursos pelos usuários, como mostra a Tabela 1. Aproximadamente 80% das sessões são de usuários residenciais, e essa proporção se mantém no que se refere aos bytes enviados e recebidos (embora os usuários comerciais recebam proporcionalmente um volume um pouco maior de dados, possivelmente devido ao fato de que as organizações comerciais costumam ter muitas máquinas ligadas em rede compartilhando a conexão à Internet). Notamos também que a frequência com que os usuários residenciais estabelecem sessões é um pouco maior que a dos usuários comerciais, porém esses últimos tendem a ter sessões um pouco mais longas.

Ainda com relação ao consumo de recursos, um estudo mais aprofundado com relação às sessões pode ser feito. Na Figura 2 apresentamos os gráficos de sessões ativas extraídos do BUBA. É interessante notar a grande queda no número de usuários comerciais utilizando a rede durante os fins-de-semana. Essa queda também é notada no comportamento dos usuários residenciais, apesar de ser bem mais branda. As duas primeiras semanas do log abrangem o período de fim de ano, justificando o menor volume de conexões nesses dias. A partir desses resultados, os provedores podem se planejar para realizar reparos e *upgrades* no sistema em horários em que o menor número de usuários

Tabela 1. Resumo dos dados fornecido pelo BUBA (CV = co-variância)

classe de usuários	residencial	comercial
# de sessões completadas	162741 (81%)	37880 (19%)
total de bytes recebidos (GB)	5387 (75%)	1791 (25%)
total de bytes enviados (GB)	2063 (82%)	452 (18%)
média de sessões completadas por usuário	29,60	22,70
duração média das sessões (horas) e CV	6,43 (2,32)	8,59 (2,26)
bytes recebidos por sessão (MB) e CV	33,90 (7,83)	48,42 (8,08)
bytes enviados por sessão (MB) e CV	12,98 (4,57)	12,23 (3,46)



(a) usuários residenciais (b) usuários comerciais

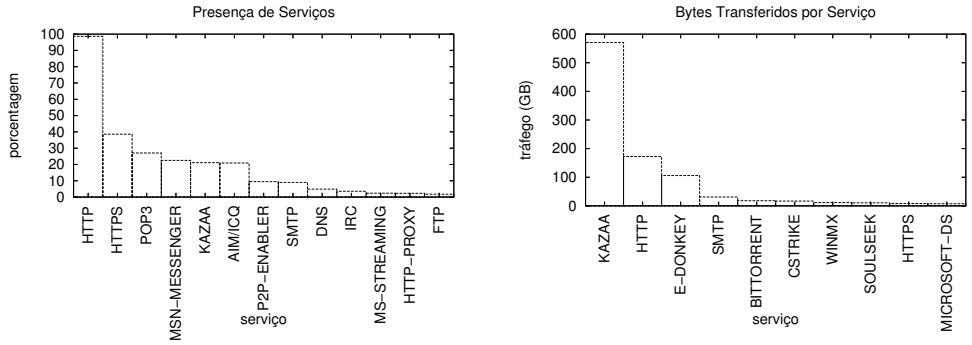
Figura 2. Sessões Ativas

estejam conectados. Além disso, o número de conexões simultâneas criadas por usuários de um ISP de banda larga deve ser bem administrado, devido à limitação do número de IPs válidos distribuídos para cada ISP.

Como mencionado anteriormente, o BUBA também permite um estudo relacionado à utilização de serviços pelos usuários. Os gráficos da Figura 3 foram extraídos dessa funcionalidade da ferramenta. observando o gráfico 3(a) temos que 99% das sessões utilizam o serviço HTTP. Uma forma de otimizar o uso do *backbone* do provedor seria através da implantação de um servidor de *cache*, como por exemplo, o *Squid* [SQUID, 2005]. Entretanto, o gráfico 3(b) indica que o HTTP não é o serviço que mais consome banda. Apesar de não estarem presentes na maioria das sessões dos usuários de Internet de banda larga, os serviços de *Peer-to-Peer* (P2P) são os que mais transferem bytes, como *Kazaa*, *e-Donkey*, *BitTorrent*, *SoulSeek* e *WinMX*. Isso reforça a importância de se desenvolver mecanismos de *cache* para este tipo de serviço ou até mesmo promover a cooperação entre os usuários do provedor.

Outra funcionalidade disponível é a de Comportamento. Se observarmos a Tabela 1, temos que os CV's (co-variância) referentes aos dados são muito grandes, o que demonstra uma hetereogeneidade no comportamento desses grupos, fazendo com que a busca por grupos mais homogêneos e representativos se faça necessária.

Através do BUBA, encontramos três grupos de sessões de usuários comerciais e três grupos residenciais bastante distintos. A co-variância da distribuição dos valores de duração da sessões e bytes transferidos é menor na análise feita em cada grupo, o que mostra que, de alguma forma, grupos mais homogêneos e representativos foram encontrados. Os dados de cada grupo estão resumidos nas Tabelas 2 e 3. A Figura 4 representa



(a) presença de serviços (b) tráfego por serviço

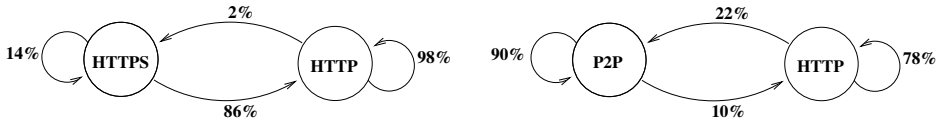
Figura 3. Popularidade de Serviços

graficamente dois desses grupos, através de CBMGs. O grafo 4(a), que representa os usuários que se limitam a navegar em sites *Web* a maior parte do tempo, corresponde à classe 2 da Tabela 3. O grafo 4(b) modela os usuários que utilizam aplicações P2P, que correspondem àqueles que estão na classe 3 da Tabela 2. Analisando cada CBMG em conjunto com suas respectivas tabelas é possível verificar que a carga que estes exercem no provedor é significativamente superior aos outros grupos de usuários.

Identificando padrões de acesso como os mostrados nesse estudo de caso, os provedores de acesso podem disponibilizar para seus usuários serviços mais personalizados, mais voltados para o tipo de uso que cada um faz de sua conexão larga, além de criar novas categorias menos genéricas que residencial e comercial, com tarifações diferenciadas.

Tabela 2. Grupos de usuários residenciais (CV = co-variância)

grupos de usuários	**grupo 1**	**grupo 2**	**grupo 3**
requisições aos serviços	HTTP (36%) MS-DS (30%) EPMAP (22%) MSN (12%)	HTTP (97%) MSN (3%)	KAZAA (64%) HTTP (36%)
# de sessões completadas	2775 (4%)	55048 (80%)	10858 (16%)
total de bytes recebidos (GB)	143,39 (6%)	894,12 (35%)	1502,49 (59%)
total de bytes enviados (GB)	70,47 (8%)	141,62 (15%)	722,71 (77%)
duração média das sessões (horas) e CV	9,94 (2,30)	5,23 (1,86)	15,68 (1,86)
bytes recebidos por sessão (MB) e CV	52,91 (3,71)	16,63 (4,03)	141,70 (2,64)
bytes enviados por sessão (MB) e CV	26.01 (4,87)	2,63 (3,69)	68,16 (3,69)



(a) (b)

Figura 4. CBMGs típicos

Tabela 3. Grupos de usuários comerciais (CV = co-variância)

grupos de usuários	grupo 1	grupo 2	grupo 3
requisições aos serviços	HTTP (73%) POP3 (27%)	HTTP (96%) HTTPS (4%)	HTTP (54%) KAZAA (46%)
# de sessões completadas	483 (8%)	4770 (77%)	911 (15%)
total de bytes recebidos (GB)	17,93 (9%)	65,16 (31%)	125,40 (60%)
total de bytes enviados (GB)	5,92 (7%)	11,10 (14%)	63,86 (79%)
duração média das sessões (horas) e CV	8,22 (1,06)	5,32 (4,07)	14,09 (1,59)
bytes recebidos por sessão (MB) e CV	38,02 (3,51)	13,99 (4,03)	71,78 (2,79)
bytes enviados por sessão (MB) e CV	12,55 (4,00)	2,38 (3,18)	14,09 (3,41)

6. Conclusões e Trabalhos Futuros

Neste artigo apresentamos o BUBA, uma ferramenta que, baseada em informações obtidas através de um log de autenticação, um log de serviços e uma tabela de categorias de usuários, permite aos provedores de acesso entenderem melhor o comportamento de seus clientes. Manipulando esses logs em conjunto, o BUBA provê uma visão completa do comportamento de cada usuário, no que se refere ao consumo de recursos do ISP, aos serviços *Web* utilizados por eles e à determinação de grupos de usuários que possuam comportamento semelhante. Essas informações auxiliam o provedor no planejamento do serviço que será oferecido aos seus clientes, seja na avaliação de sua infra-estrutura ou na personalização de serviços.

Como trabalho futuro, pretendemos deixar a ferramenta mais eficiente, além de torná-la mais genérica de forma que a mesma possa ser útil em outros contextos de caracterização, em especial no estudo da carga de trabalho em *Web Services*.

Referências

Cflowd (2005). Traffic Flow Analysis Tool. www.caida.org/tools/measurement/cflowd.

CNPA (2005). CISCO Netflow Performance Analysis. www.cisco.com/warp/public/cc/pd/iosw/prodlit/ntfo_wp.pdf.

FlowScan (2005). Network Traffic Flow Visualization and Reporting Tool. www.caida.org/tools/utilities/flowscan.

Hartigan, J. (1975). Clustering Algorithms. *John Wiley and Sons, Inc.*

ICRadius (2005). www.icradius.org.

M. Fullmer, S. R. (2000). The OSU Flow-tools Package and Cisco Netflow Logs.

Menascé, D. and Almeida, V. (2000). *Scaling for e-Business: Technologies, Models, Performance and Capacity Planning*. Prentice Hall, Upper Saddle River - NJ.

Netflow (2005). www.cisco.com/warp/public/732/Tech/netflow.

Pew Internet (2004). The Broadband Difference. www.pewinternet.org.

RadiusContext (2005). www.tummy.com/Community/software/radiuscontext.

RadRep (2005). Radius Reporting & Billing Software. www.radiusreporting.com.

Rigney, C. (2000). Remote Authentication Dial in User Service (RADIUS) RFC 2866.

RRDtool (2005). Round Robin Database Tool. www.caida.org/tools/utilities/rrdtool.

SQUID (2005). http://www.squid-cache.org/.